Rare Variant Analysis

Xihao Li

Department of Biostatistics and Department of Genetics

UNC-Chapel Hill

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

UNC
SCHOOL OF MEDICINE

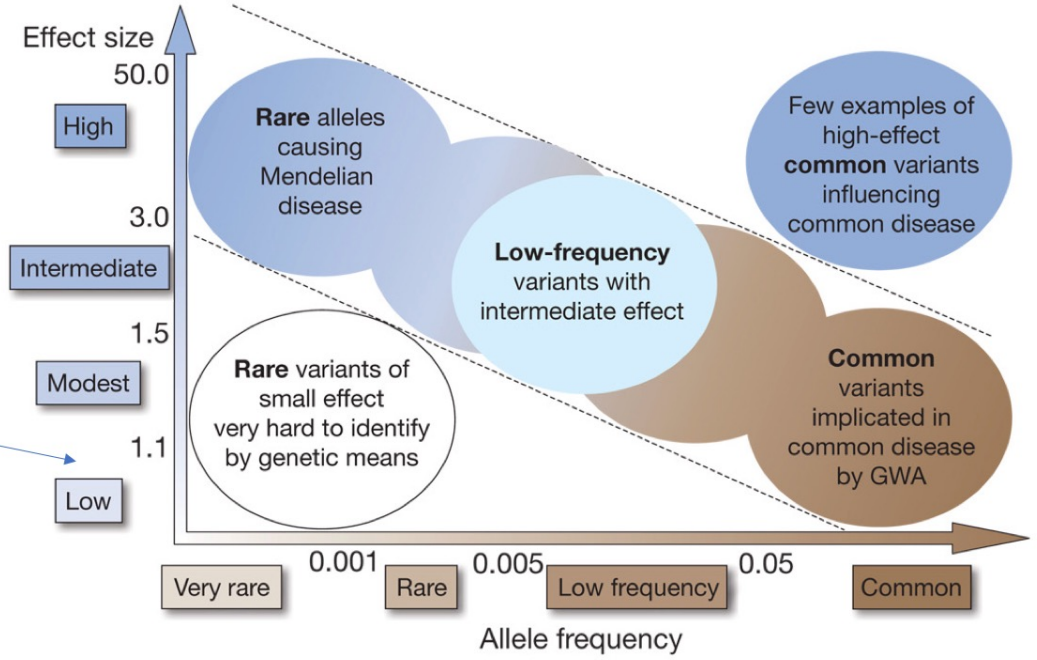
Acknowledgements

Part of the slides were adapted from Dr. Alkes Price
and Dr. Xihong Lin

Recall Week #2: GWAS are used for common variation

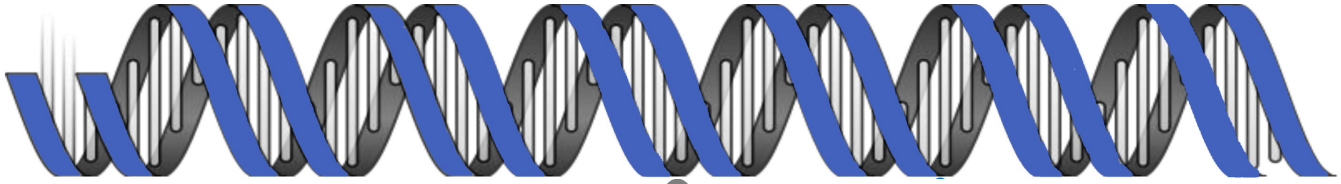
GWAS Power Calculations are a function of effect size, allele frequency and sample size.

Common variants have smaller effect sizes which means large sample sizes are necessary to detect them!



Manolio et al. 2009 Nature

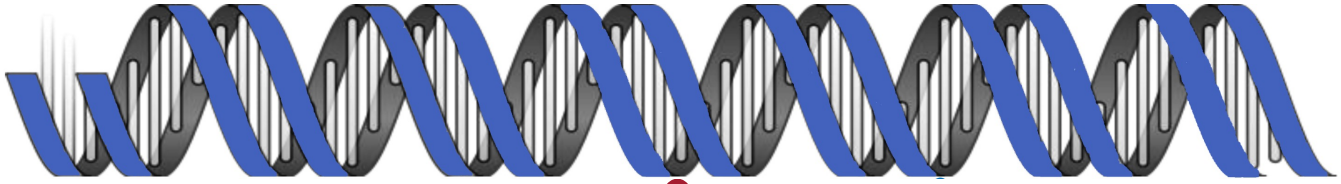
What are rare variants?



Chromosome 19 DNA

Curated phenotypes

What are rare variants?



Chromosome 19 DNA

		G						Y:LDL
	CCGATCCAAGTCCATATATACCGATTTGACCGAA	0	0	0	1	0	1	141
	CCGATCCAAGTCCATATATACCAATTTAACCGAA	0	0	0	1	1	0	151
	CCGATCCAAGTCCATACATACCGATTTAACCGAA	0	0	0	0	0	0	126
	CCGATCCAAGTCCATACATACCGATTTAACCGAA	0	0	0	0	0	0	116
	CCAATCCAAGTCCATATATACCGATTTAACCGAA	1	0	0	1	0	0	178
	CCGATCTAAGTCCATACATACCGATTTAACCGAA	0	1	0	0	0	0	139
	CCGATCCAAGTTCATACATACCGATTTAACCGAA	0	0	1	0	0	0	113

Are rare and low-frequency variants important?

Common Disease/Common Variant hypothesis

"For common diseases, there will be one or a few predominating disease alleles with relatively high frequencies at each of the major underlying disease loci"

Five Years of GWAS Discovery

Peter M. Visscher,[1,2,*] Matthew A. Brown,[1] Mark I. McCarthy,[3,4] and Jian Yang[5]

Introduction: Have GWASs Been a Failure?

From McCLellan and King, *Cell* 2010[1]: If common alleles influenced common diseases, many would have been found by now. The issue is not how to develop still larger studies, or how to parse the data still further, but rather whether the common disease–common variant hypothesis has now been tested and found not to apply to most complex human diseases."

10 Years of GWAS Discovery: Biology, Function, and Translation

Peter M. Visscher,[1,2,*] Naomi R. Wray,[1,2] Qian Zhang,[1] Pamela Sklar,[3] Mark I. McCarthy,[4,5,6] Matthew A. Brown,[7] and Jian Yang[1,2]

Furthermore, using WGS data for association analysis of rare variants has the potential to boost power through the combination of alleles of similar impact (e.g., via burden tests across a gene) under the assumption of multiple independent causative variants in a gene region.

15 years of GWAS discovery: Realizing the promise

Abdel Abdellaoui,[1,*] Loic Yengo,[2] Karin J.H. Verweij,[1] and Peter M. Visscher[2]

Rare variants can explain a substantial fraction of the heritability and have different properties than common variants, at least in part as a result of natural selection. They generally have larger effects and behave differently in their relationship with ancestry,[58] geography,[59] and therefore potentially also with respect to their association with environmental effects. The vast majority of human variants are rare; among 400 million detected variants in

Reich & Lander 2001 Trends Genet
Visscher et al. 2012 Am J Hum Genet
Visscher et al. 2017 Am J Hum Genet
Abdellaoui et al. 2023 Am J Hum Genet

The 1000 Genomes (1000G) Project

Sequence the entire genomes of 1,092 individuals:
379 of European ancestry (Europe and USA)
286 of East Asian ancestry (Asia)
246 of African ancestry (Africa and USA)
181 of Latino ancestry (Latin America and USA)

Use next-generation sequencing technologies (~4x coverage):
e.g. Illumina, 454, SOLiD (read lengths 25-400bp)

1000 Genomes Project Consortium 2012 Nature

1000G Project: Summary of main results

- 38 million SNPs discovered and successfully genotyped.
  - Most of these are rare and low-frequency variants.

- The 38 million SNPs include
  - 99.7% of all SNPs with minor allele frequency 5%
  - 98% of all SNPs with minor allele frequency 1% ***
  - 50% of all SNPs with minor allele frequency 0.1%
  - based on an independent (~2,500) UK European sample (the Wellcome Trust-funded UK10K project)

***: stated goal to identify >95% of SNPs with frequency 1% was successfully achieved.

1000 Genomes Project Consortium 2012 Nature

1000G Project: Phase 3

Sequence the entire genomes of 2,504 (unrelated) individuals:
503 of European ancestry (Europe and USA)
504 of East Asian ancestry (Asia)
661 of African ancestry (Africa and USA)
347 of Latino ancestry (Latin America and USA)
489 of South Asian ancestry (South Asia and USA)

Use next-generation sequencing technologies (~7x coverage):
Illumina only (read lengths 70-400bp only)
85 million SNPs, of which 64 million have MAF<0.5%

1000 Genomes Project Consortium 2015 Nature

1000G Project: High coverage phase

Sequence the entire genomes of 3,202 individuals:
633 of European ancestry (Europe and USA)
601 of East Asian ancestry (Asia)
893 of African ancestry (Africa and USA)
490 of Latino ancestry (Latin America and USA)
585 of South Asian ancestry (South Asia and USA)
2,504 unrelated samples + 698 related samples (602 parent-child trios)

Use high coverage whole-generation sequencing technologies (~30x coverage):
Illumina NovaSeq 6000 instruments
111 million SNVs, 14.4 million indels
47.9% singletons, 37.5% rare (MAF < 1%)

Byrska-Bishop et al. 2022 Cell

What about rare variants?

- The 1000G project has identified most low-frequency variants
  - (minor allele frequency 1% - 5%). These variants can be placed on genotyping arrays or imputed

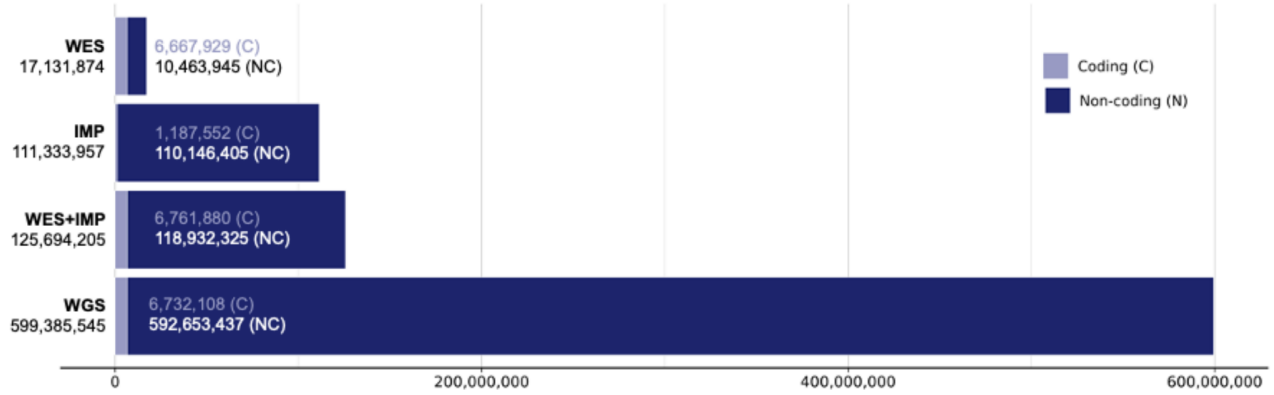
- Rare variants: most have not been identified by 1000 Genomes!
  - Must sequence disease samples directly. Or impute?
  - Past focus has been mostly on exome sequencing, but
  - now shifting to whole-genome sequencing.

Kiezun et al. 2012 Nat Genet, Tennessen et al. 2012 Science
Karczewski et al. 2022 Cell Genom; Szustakowski et al. 2021 Nat Genet

WES vs IMP vs WES + IMP vs WGS

UK Biobank Dataset: Genomes, Exomes and Imputation

A



B

Consequence	WGS (% Singleton)	WES+IMP (% Singleton)	Intersection	WGS only	WES+IMP only	% WGS only	% WES+IMP only
Coding variants	6,732,108 (48%)	6,761,880 (48%)	6,544,263	187,845	217,617	2.7	3.1
Non-coding variants	592,653,437 (47%)	118,932,325 (5%)	111,394,188	481,259,249	7,538,137	80.1	1.3
All variants	599,385,545 (47%)	125,694,205 (7%)	117,938,451	481,447,094	7,755,754	79.3	1.3

Gaynor et al. 2023 medRxiv

Larger WGS reference panels

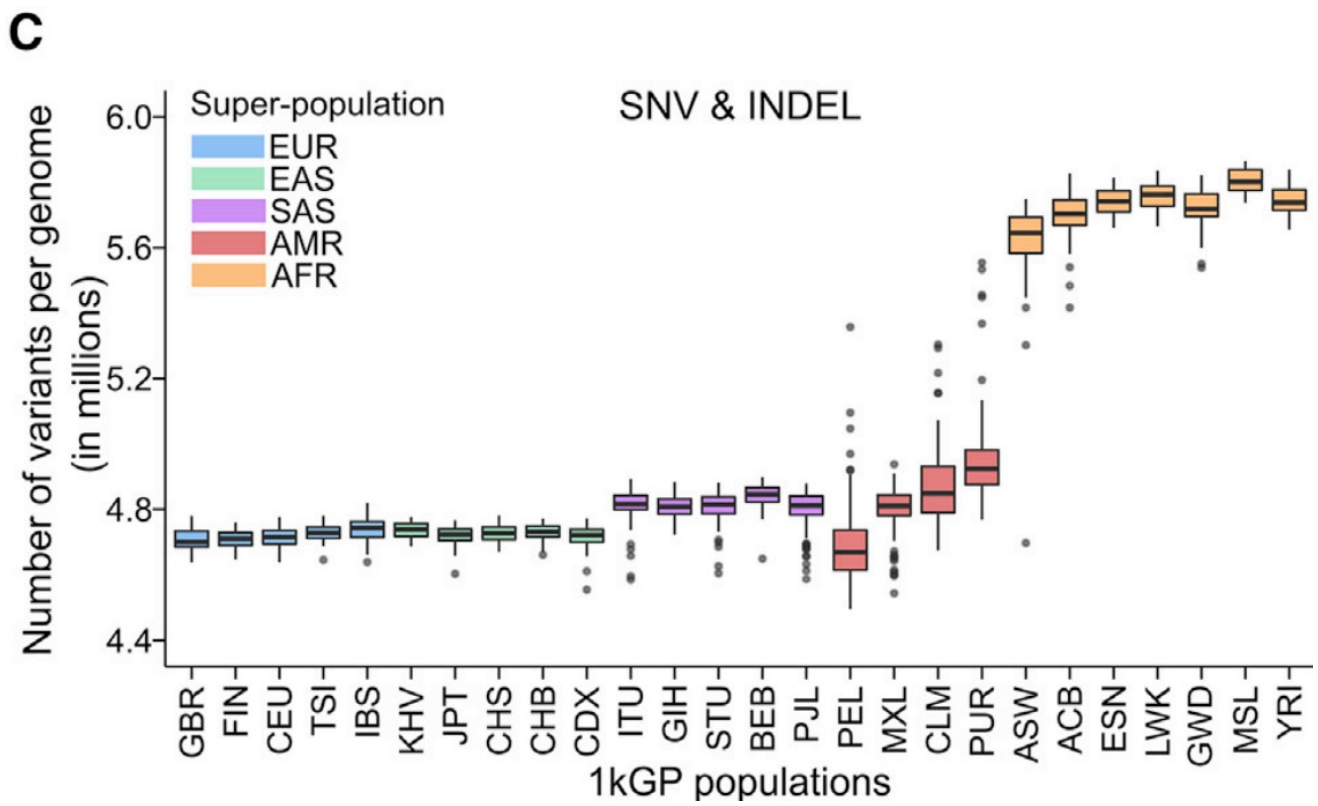
Haplotype Reference Consortium (McCarthy et al. 2016 Nat Genet):
4-8x WGS of 32,488 mostly European samples from 20 studies

deCODE Genetics WGS data set (Gudbjartsson et al. 2015 Nat Genet):
20x WGS of 2,636 Icelanders
Accurate imputation down to 0.1% MAF in the Icelandic population

TOPMed (Taliun et al. 2021 Nature)
38x WGS of 53,831 diverse samples (increasing to ~180K)
40% European, 29% African American, 19% Latino, 8% Asian

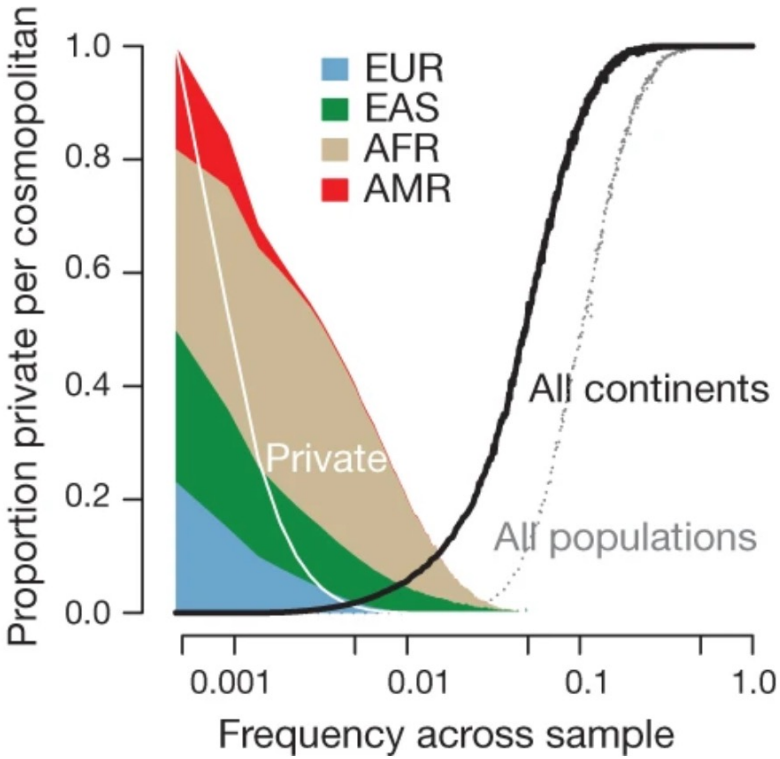
UK Biobank (Halldorsson et al. 2022 Nature)
24x WGS of 150,119 mostly European samples (increasing to 500K)

African populations have more genetic diversity



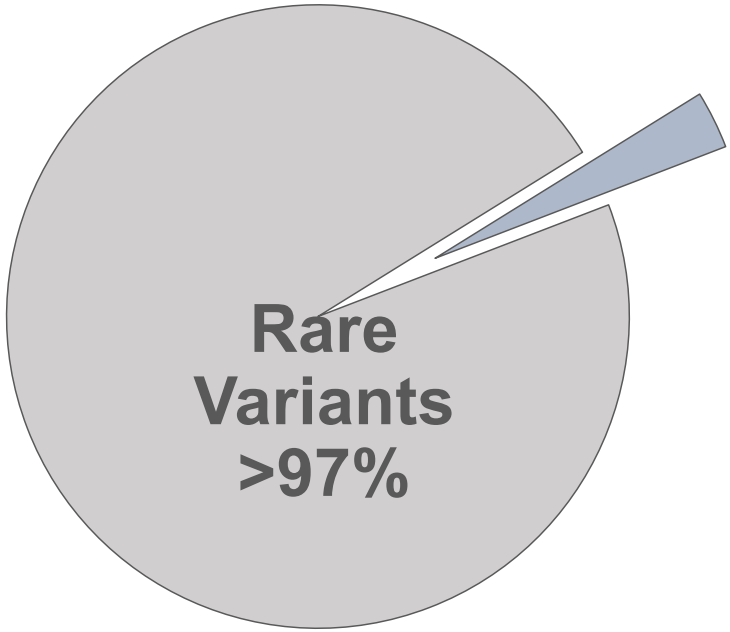
Byrska-Bishop et al. 2022 Cell

Common variants are shared across populations, but rare variants are often population specific



1000 Genomes Project Consortium 2012 Nature

WGS covers 100% of the genome (common and rare; coding and noncoding)

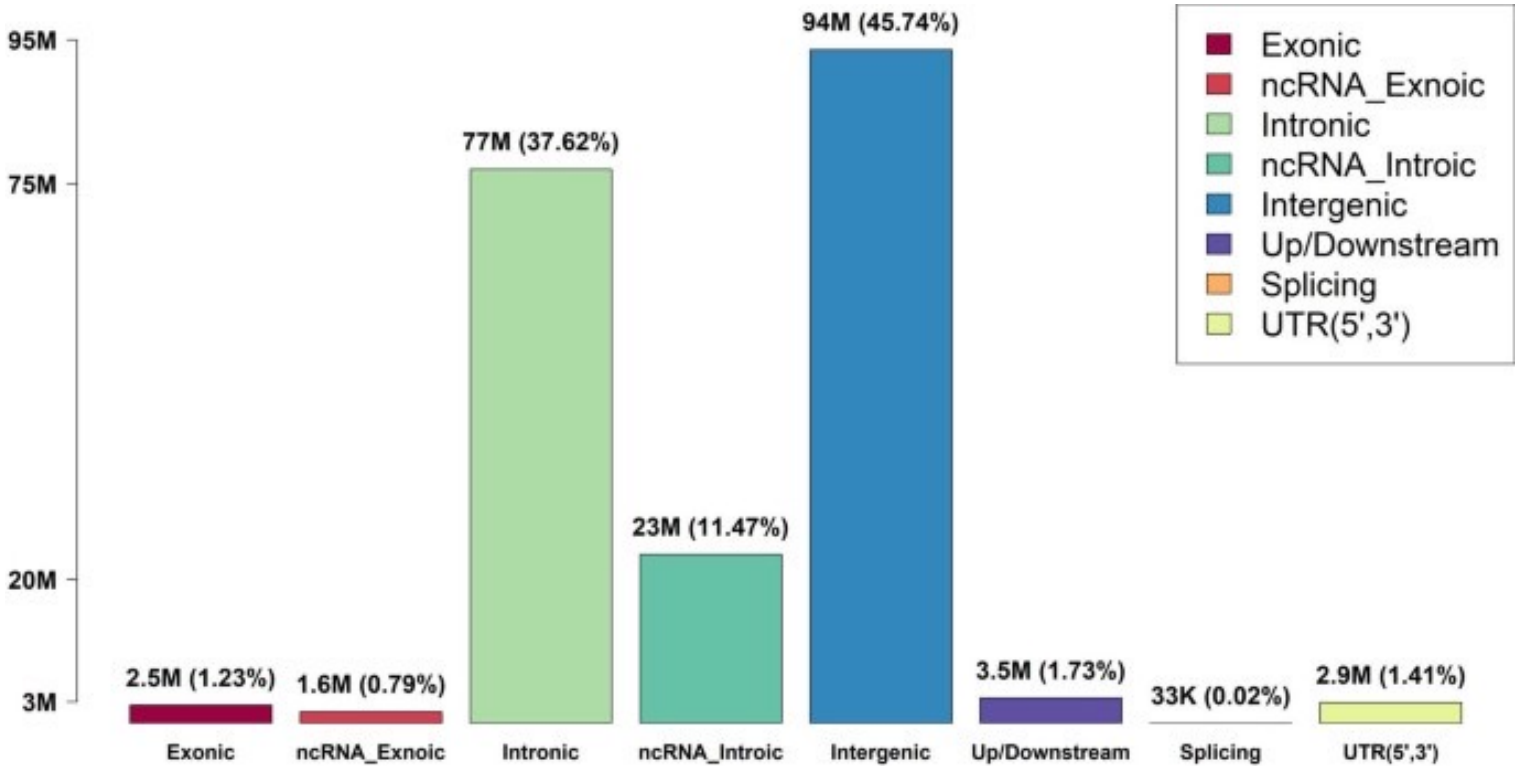


GWAS Common Variants < 3%

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

Cirulli & Goldstein 2010 Nat Rev Genet

WGS covers 100% of the genome (coding and noncoding)



Li et al. 2022 Nat Methods

Massive rare variants in WGS data

Trans-Omics for Precision Medicine (TOPMed) Freeze 5b (n = 54,499) and Freeze 8 (n = 140,306) data

Minor Allele Frequency	# Genetic Variants (%) TOPMed Freeze 5b	# Genetic Variants (%) TOPMed Freeze 8
Singletons	211.4M (47%)	322.4M (46%)
Doubletons	70.5M (16%)	107.5M (15%)
Doubletons - .1%	144.9M (32%)	240.9M (35%)
.1% - 1%	13.8M (3.0%)	14.0M (2.0%)
> 1%	**12.0M (2.7%)**	**12.9M (1.8%)**
Total	453M (100%)	698M (100%)

Large Scale WGS Timeline

2008

1000 Genomes
N=1000

2015

TOPMed
(NHLBI)
N=185,000

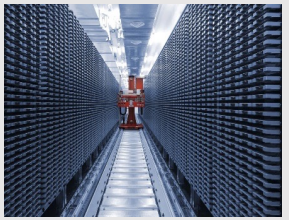
AllofUs
(NIH)
N=250,000

2016

GSP(NHGRI)
N=200,000

2018

Biobanks
(e.g. UKB)
N=500,000



Current status of biobank-scale WGS studies

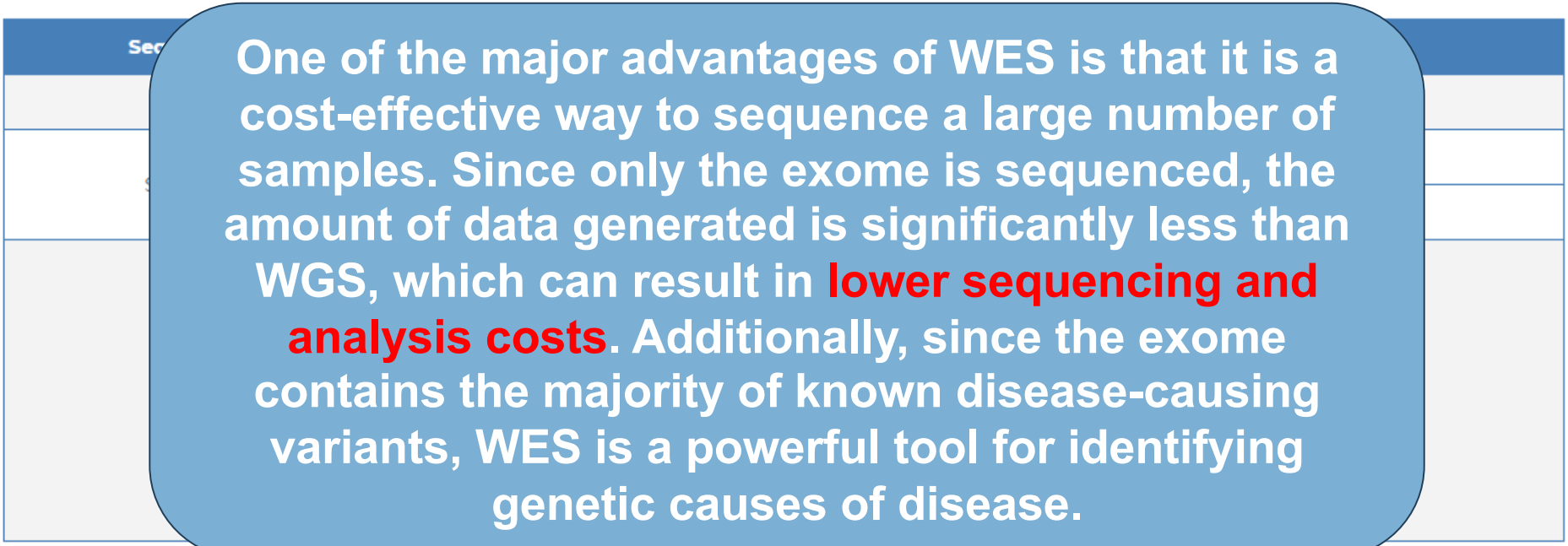
- **TOPMed (Trans-Omics Precision Medicine)**
  - Finished n=185,000 WGS (Freeze 10)
  - >1 billion genetic variants
  - Rich -omics resources
- **UK Biobank**
  - Finished n=500,000 WGS/WES
  - EHRs, imaging, activity monitors
- **All of Us**
  - Finished n=245,350 WGS
  - Survey responses, EHRs, physical measurements

Advantages of Whole-Exome Sequencing

Sequencing Platform	Illumina NovaSeq 6000
Read Length	Paired-end 150 bp
Sequencing Depth	For Mendelian disorder/rare disease: effective sequencing depth above 50× (6G)
	For tumor sample: effective sequencing depth above 100× (12G)
Data Analysis	• Data quality control • Alignment to a reference genome • SNP and InDel calling • Somatic SNP/InDel/CNV mutation detection (tumor-normal paired samples)

Source: Novogene (https://www.novogene.com/us-en/resources/blog/wgs-vs-wes-which-genetic-sequencing-method-is-right-for-you/)

Advantages of Whole-Exome Sequencing



One of the major advantages of WES is that it is a cost-effective way to sequence a large number of samples. Since only the exome is sequenced, the amount of data generated is significantly less than WGS, which can result in **lower sequencing and analysis costs**. Additionally, since the exome contains the majority of known disease-causing variants, WES is a powerful tool for identifying genetic causes of disease.

Source: Novogene (https://www.novogene.com/us-en/resources/blog/wgs-vs-wes-which-genetic-sequencing-method-is-right-for-you/)

Advantages of Whole-Genome Sequencing

Platform Type	Illumina NovaSeq 6000	PacBio Sequel II/IIe	Nanopore PromethION
Read Length	Paired-end 150 bp	> 15 kb for Sequel II (Average)	> 17 kb (Average)
Sequencing Depth	For rare diseases: 30-50×	For genetic diseases: 10-20×	For genetic diseases: 10-20×
	For tumor tissues: 50×; For adjacent normal tissues and blood: 30×	For tumor tissues: ≥20×	For tumor tissues: ≥20×
Standard Data Analysis	• Data quality control • Alignment with reference genome • SNP/InDel/SV/CNV detection • Somatic SNP/InDel/SV/CNV detection (For tumor-normal paired samples)	• Data quality control • Sequence alignment • Structural variant (SV) detection • Variation annotation	

Source: Novogene (https://www.novogene.com/us-en/resources/blog/wgs-vs-wes-which-genetic-sequencing-method-is-right-for-you/)

Advantages of Whole-Genome Sequencing

Platform Type	Illumina NovaSeq 6000	PacBio Sequel II/IIe	Nanopore PromethION
Rea			age)
Seque			eases:
			ues:
Sta	detection (For tumor-normal paired samples)	• Variation annotation	

One of the major advantages of WGS is that it provides a more comprehensive view of an individual's genetic makeup. WGS can identify variants that are not present in the exome, including those in non-coding regions and structural variants. This can be particularly useful for identifying rare or novel variants that may be missed by WES. Additionally, WGS can provide information about ancestry and population genetics.

Source: Novogene (https://www.novogene.com/us-en/resources/blog/wgs-vs-wes-which-genetic-sequencing-method-is-right-for-you/)

Goal of WGS analysis

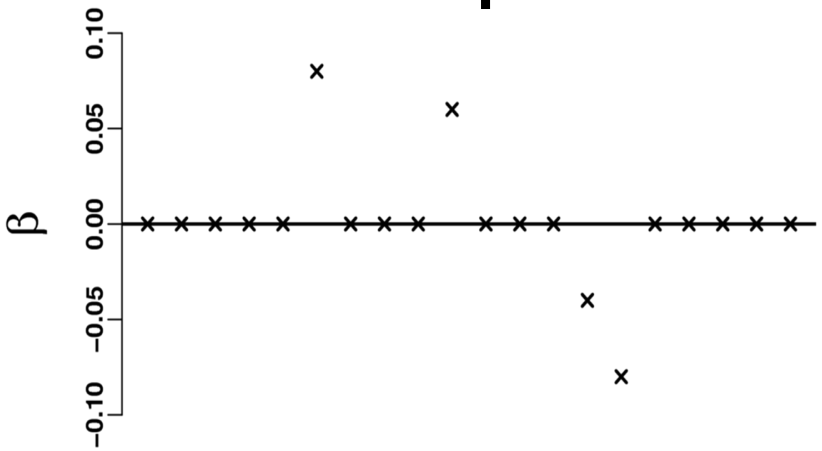
Signal Detection

Scan the whole genome to identify genetic regions associated with diseases/traits

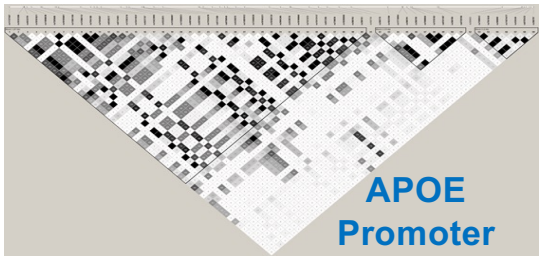
Challenges in rare variant analysis of WGS data

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

$$LDL = G\beta + e$$



Genomic Location



WGS association analysis workflow

Common Variants Single Variant Analysis

Rare Variants Variant-Set Analysis

- **Gene-centric analysis** (gene-centric genomic elements): pLOFs, missense, synonymous, promoters, enhancers
- **Non-gene-centric analysis**: sliding windows, dynamic windows (SCANG)

Gene-Centric Analysis

Non-Gene-Centric Analysis

Povysil et al. 2019 Nat Rev Genet
Li, X. et al. 2020 Nat Genet
Li, Z. et al. 2022 Nat Methods

Basic model and problem

- Y_i = phenotype (outcome) $(i = 1, \cdots, n)$
- $X_i = q$ covariates
- G_i = p genetic variants/SNVs (AA, AG, GG = 0, 1, 2) in a set.

- **Model**

$$g(\mu_i) = \boldsymbol{X_i'\alpha} + \boldsymbol{G_i'\beta}$$

- **Hypothesis of no variant-set effect**:

$$H_0: \boldsymbol{\beta} = 0 \quad vs\ H_1: \boldsymbol{\beta} \neq 0$$

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 $U_j = \sum_{i=1}^{n} G_{ij}(Y_i - \hat{\mu}_{i0})$
  - Scalability: Fit the same null model $g(\mu_i) = \boldsymbol{X_i'\alpha}$ only once when scanning the genome

Conventional rare variant tests

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

$$Q_{Burden} = \left(\sum_{j=1}^{p} w_j U_j \right)^2$$

Li & Leal 2008 AJHG
Madsen & Browning 2009 PLoS Genet
Price et al. 2010 AJHG

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

$$Q_{SKAT} = \sum_{j=1}^{p} w_j U_j^2$$

Wu et al. 2010 & 2011 AJHG

- **SKAT-O** (combined burden and SKAT)

$$Q_{SKAT-O} = \rho Q_{Burden} + (1-\rho) Q_{SKAT}$$

Lee et al. 2012 AJHG & Biostatistics

Conventional rare variant tests

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

$$Q_{Burden} = \left(\sum_{} w_i U_i \right)^2$$

Li & Leal 2008 AJHG
Madsen & Browning 2009 PLoS Genet

Remark: for both burden test and SKAT, in-sample LD is required to obtain correct null distributions and avoid false-positive associations (cannot use LD from population reference panel)

- **SK** … different dir…

$$Q_{SKAT} = \sum_{j=1} w_j U_j^2$$

Wu et al. 2010 & 2011 AJHG

- **SKAT-O** (combined burden and SKAT)

$$Q_{SKAT-O} = \rho Q_{Burden} + (1-\rho) Q_{SKAT}$$

Lee et al. 2012 AJHG & Biostatistics

Emerging needs in powerful and scalable rare variant analysis

- High-dimensional hypothesis testing:

  Signals might be dense or sparse under H_1

- Cryptic relatedness

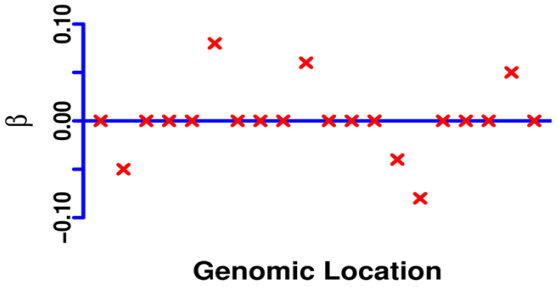
- Boost power by dynamically incorporating functional annotations

- Meta-analysis of rare variant associations across different studies

Scalable rare variant tests for dense & sparse alternatives

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

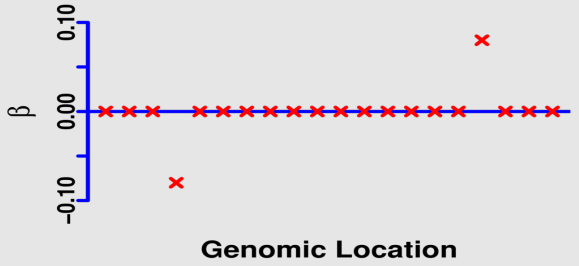
Dense Alternative



Sequencing Kernel Association Test (SKAT)

- Wu et al. 2010 & 2011, AJHG

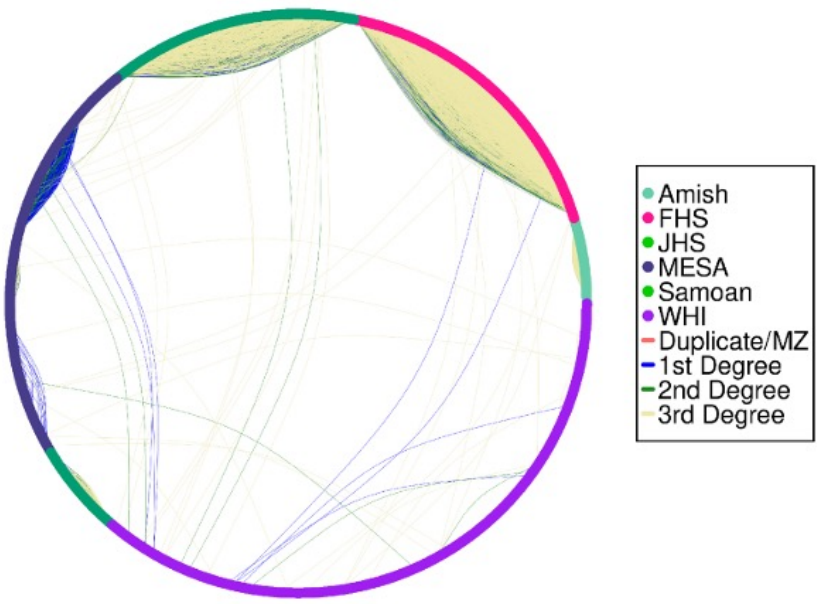
Sparse Alternative

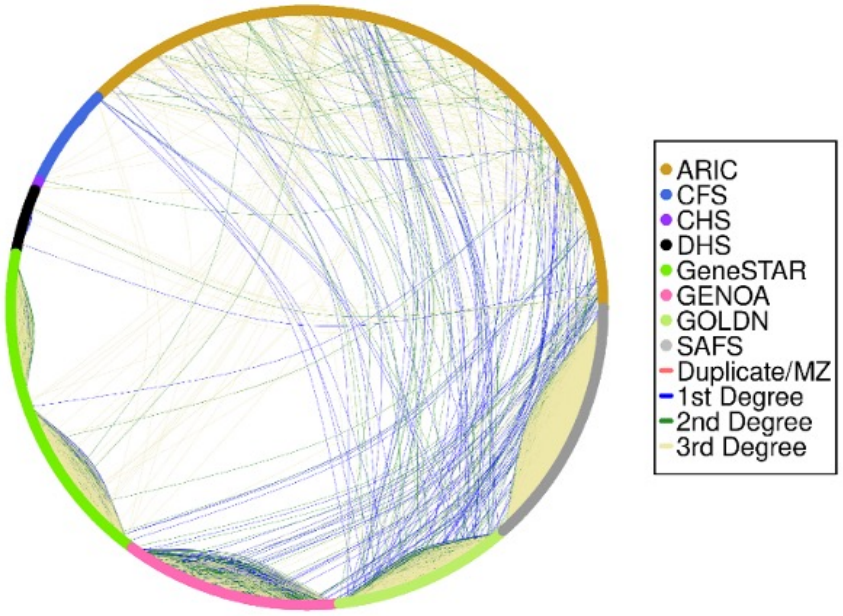


Generalized Higher Criticism (GHC) /Generalized Berk-Jones (GBJ)/ACAT

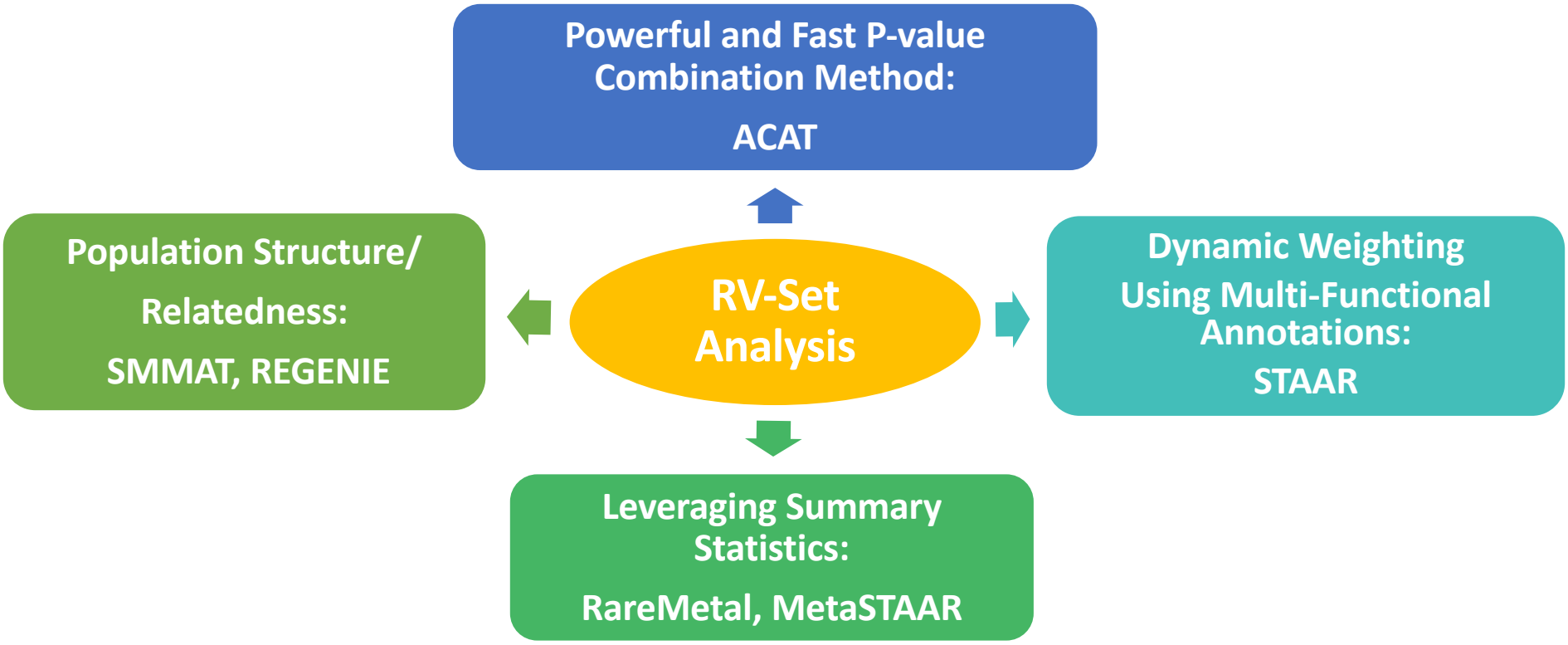
- Barnett et al (GHC) 2017, JASA
- Sun et al. (GBJ) 2019, PLoS Genet & 2020, JASA
- Liu et al (ACAT) 2019, AJHG

Population stratification and relatedness (TOPMed Freeze 5)





Moving beyond: new rare variant set-based tests



SMMAT: variant-Set Mixed Model Association Tests

Key features: Control for population structure and relatedness

- Only fit a linear/logistic mixed model once for whole genome analysis
- Tests:
  - SMMAT-Burden (SMMAT-B)
  - SMMAT-SKAT (SMMAT-S)
  - SMMAT-O (SKAT-O type)
  - SMMAT-E (efficiently combines SMMAT-B and SMMAT-S using matrix projection)

Chen et al. 2019 AJHG

SMMAT: variant-Set Mixed Model Association Tests

Covariates

Genotypes

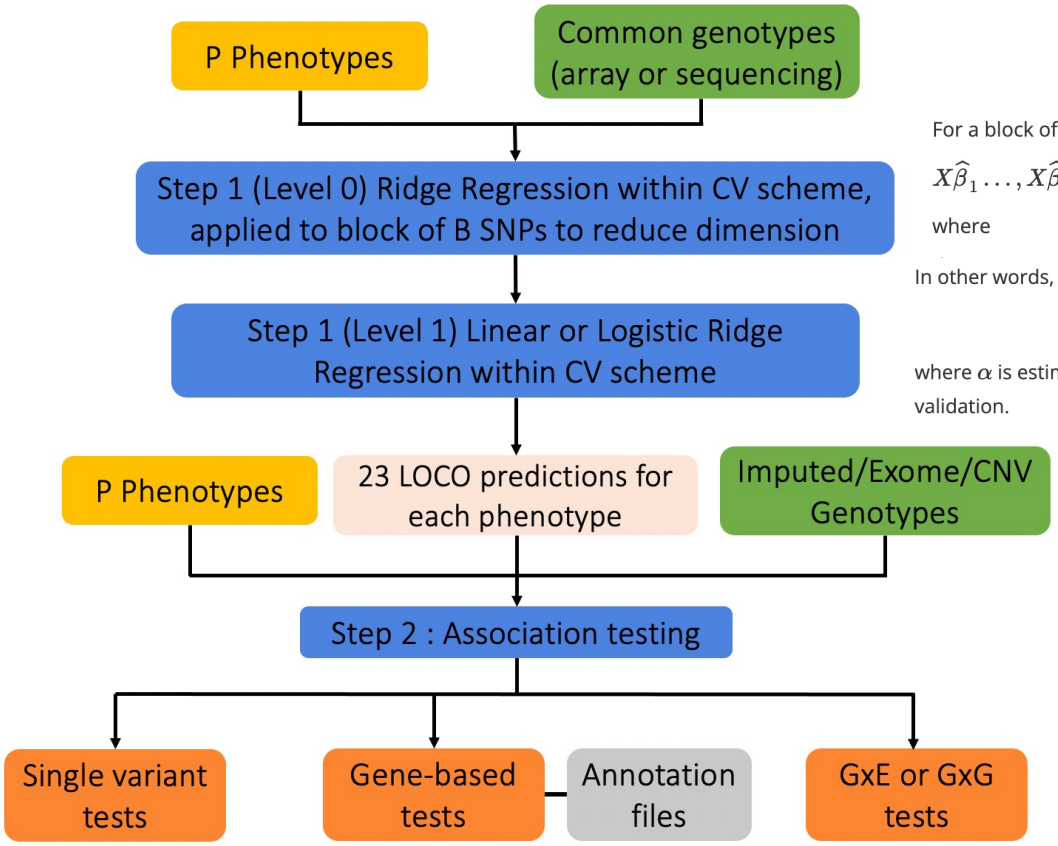
- Full GLMM: $g(\mu_i) = \boldsymbol{X}_i'\alpha + \boldsymbol{G}_i'\beta + b_i$, → Random effects for relatedness

where $b_i \sim N(0, \theta\boldsymbol{\Phi})$ and Φ is the Whole Genome Genetic Related Matrix (GRM) estimated using the whole genome data

- Fit the null GLMM only once : $g(\mu_{0i}) = \boldsymbol{X}_i\boldsymbol{\alpha} + b_i$
- Scores: $U_j = \sum_i G_{ij}(Y_i - \hat{\mu}_{i0})$ → Burden, SKAT, omnibus

Chen et al. 2016 & 2019 AJHG

REGENIE: Whole-Genome Regression



For a block of SNPs in a $N \times B$ matrix X and $N \times 1$ phenotype vector Y we calculate J predictors

$X\widehat{\beta}_1 \ldots, X\widehat{\beta}_J$

where

In other words, we fit the model

$$Y = W\alpha + \epsilon$$

where α is estimated as $\widehat{\alpha} = (W^T W + \phi I)^{-1} W^T Y$ and the parameter ϕ is chosen via K-fold cross-validation.

Mbatchou et al. 2021 Nat Genet

REGENIE: Whole-Genome Regression (Cont'd)

- *Why doesn't **regenie** need a genetic relatedness matrix (GRM)?*

regenie performs whole genome regression using the following model

$$Y = X\beta + \epsilon$$

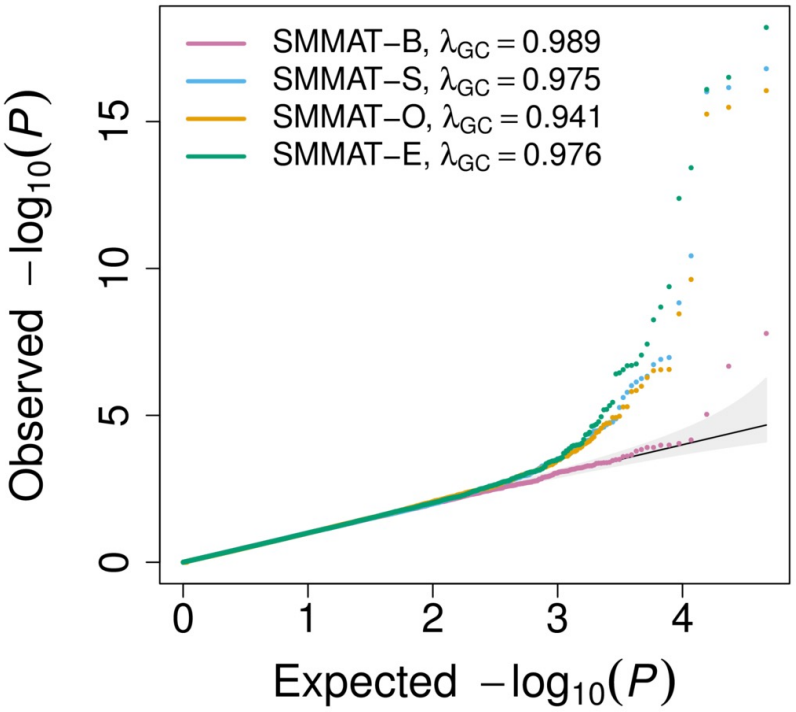
where $Y_{N\times 1}$ is a phenotype, $X_{N\times M}$ is a genotype matrix, and $\epsilon_i \sim N(0, \sigma^2)$. This model has close ties to a linear mixed model (LMM) based on an infinitesimal model

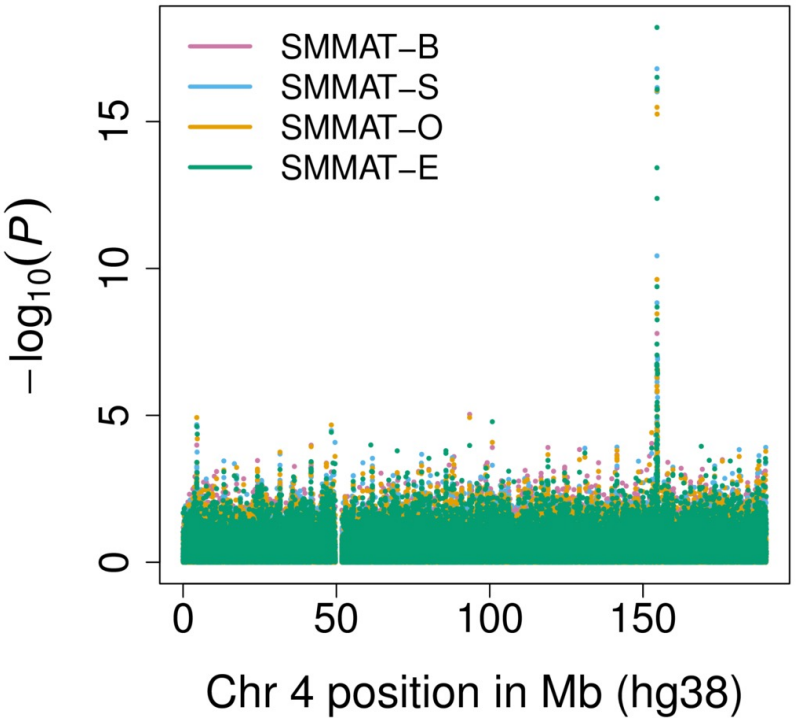
$$Y = u + \epsilon$$

where $u \sim N(0, \sigma_u^2 K)$ with $K_{N\times N} = XX^T/M$ is referred to as the genetic relatedness matrix (GRM). In the LMM, the polygenic effects have been integrated out so that model only involves the GRM K through a variance component in the covariance matrix of the trait.

In **regenie**, we directly estimate the polygenic effects parameter β by using ridge regression, which corresponds to fitting a linear regression model with a L2 penalty to impose shrinkage. Hence, we bypass having to use the GRM K and use the polygenic effect estimates $X\hat{\beta}$ to control for population structure when testing variants for association.

Analysis of TOPMed fibrinogen WGS data (n=23,763)





TOPMed WGS Fibrinogen Analysis

Known association rare variants:

- rs6054 (154,568,456)
- rs201909029 (154,567,636)

Start (kb)	End (kb)	No. of Variants	SMMAT-B	SMMAT-S	SMMAT-O	SMMAT-E
154,562	154,566	326	0.76	1.5×10^{-9}	3.5×10^{-9}	4.2×10^{-10}
154,566	154,570	309	1.6×10^{-8}	9.7×10^{-17}	3.3×10^{-16}	3.1×10^{-17}
154,570	154,574	332	0.030	1.9×10^{-7}	5.2×10^{-7}	8.9×10^{-8}

ACAT: Aggregated Cauchy Association Test

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.

Liu et al. 2019 AJHG

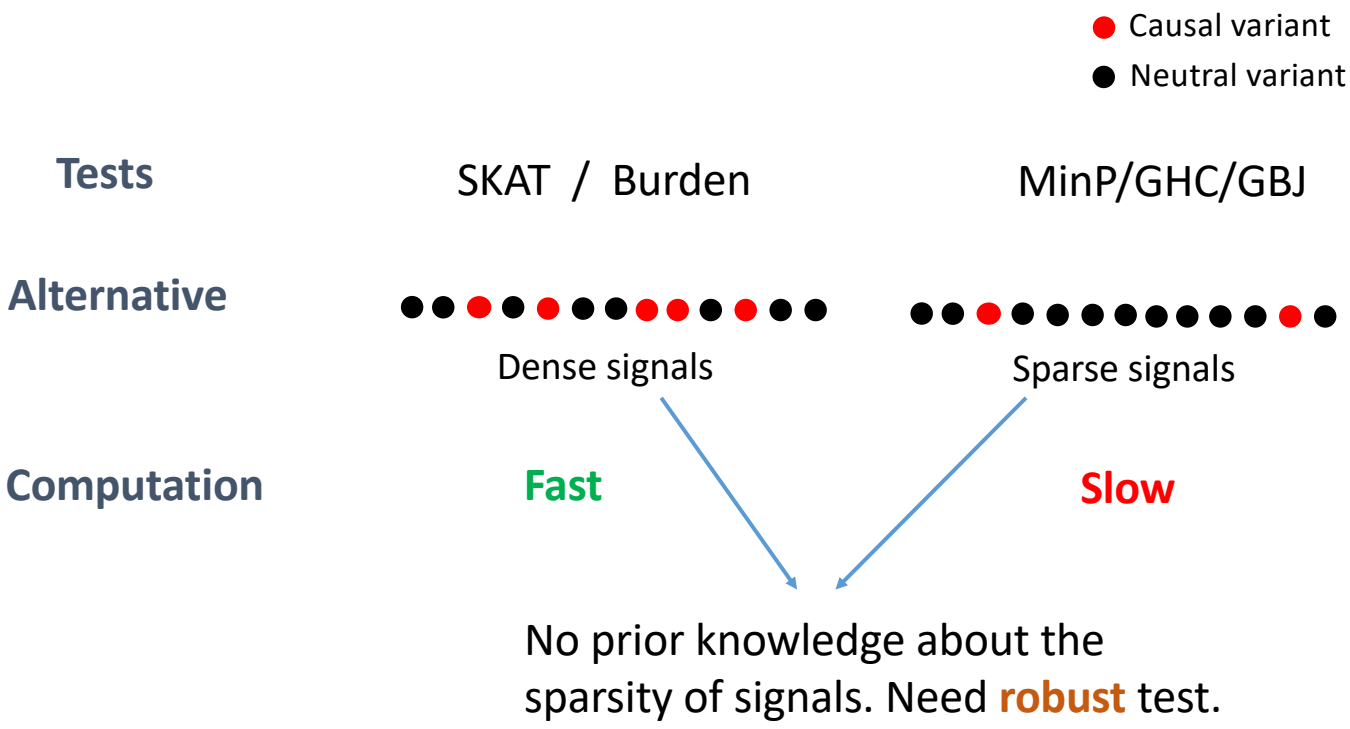
Aggregated Cauchy Association Test (ACAT)

Transform p-value to Cauchy

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

Weights

Existing variant-set tests

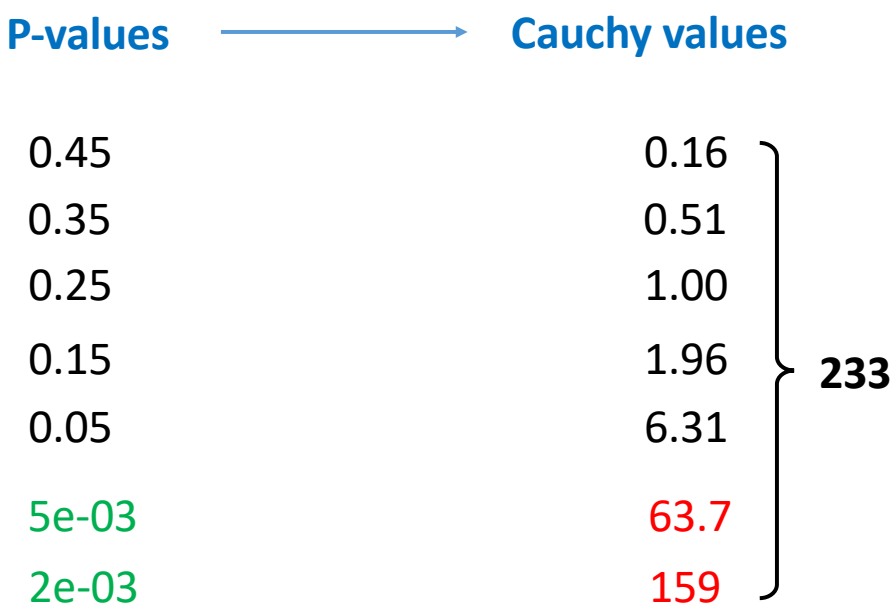


Why using Cauchy distribution – some insights

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)	≈Cauchy(0,1)
$X_i \sim$ Normal(0,1)	$\bar{X} \sim$ N(0,1)	$\bar{X} \sim$ N(0,1/d)	

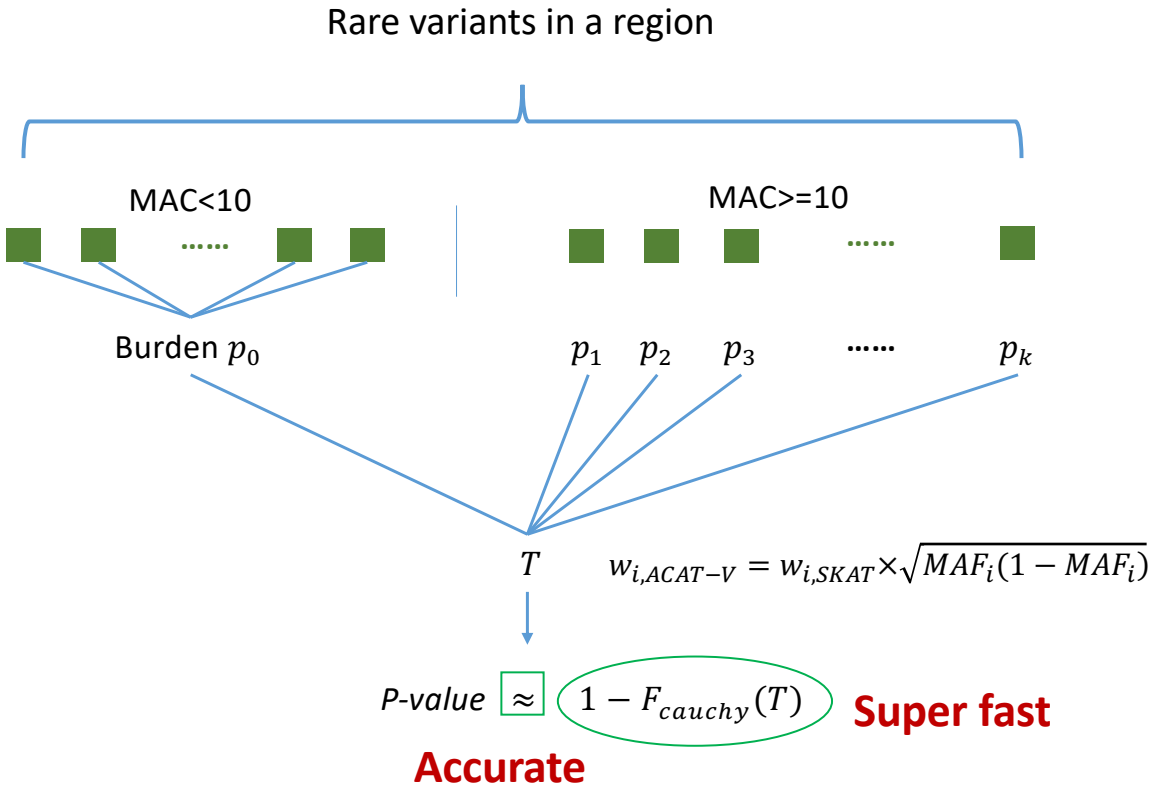
Heavy tail makes Cauchy distribution insensitive to correlation

ACAT is powerful against sparse alternatives



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

ACAT-V for testing a variant-set



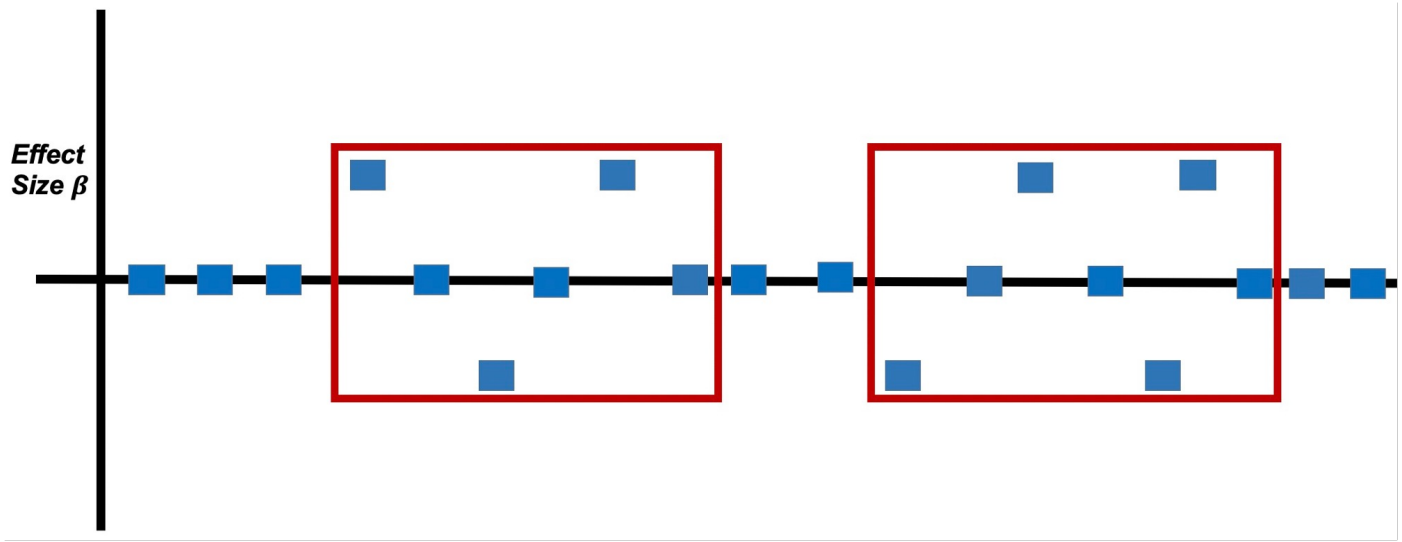
STAAR: variant-Set Test for Association using Annotation infoRmation

Key features:

- Boost RV analysis power by optimally combining statistical evidence of MAFs (default in burden/SKAT/ACAT-V) and multiple **functional annotations**
- Computationally scalable
- Applicable to any given variant-set

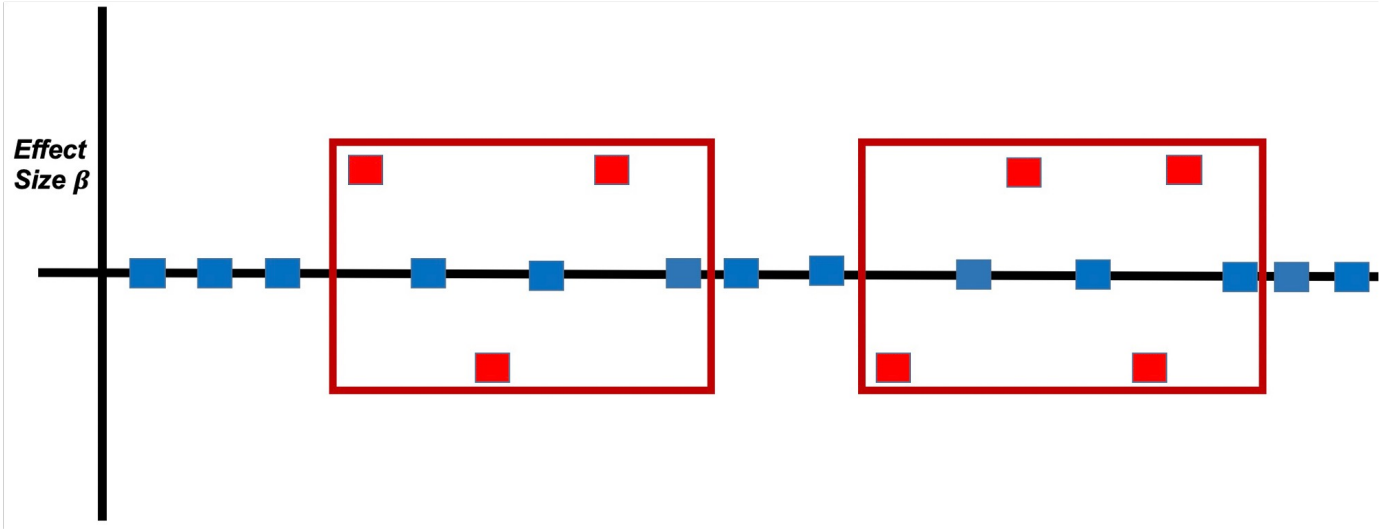
Li et al. 2020 Nat Genet

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



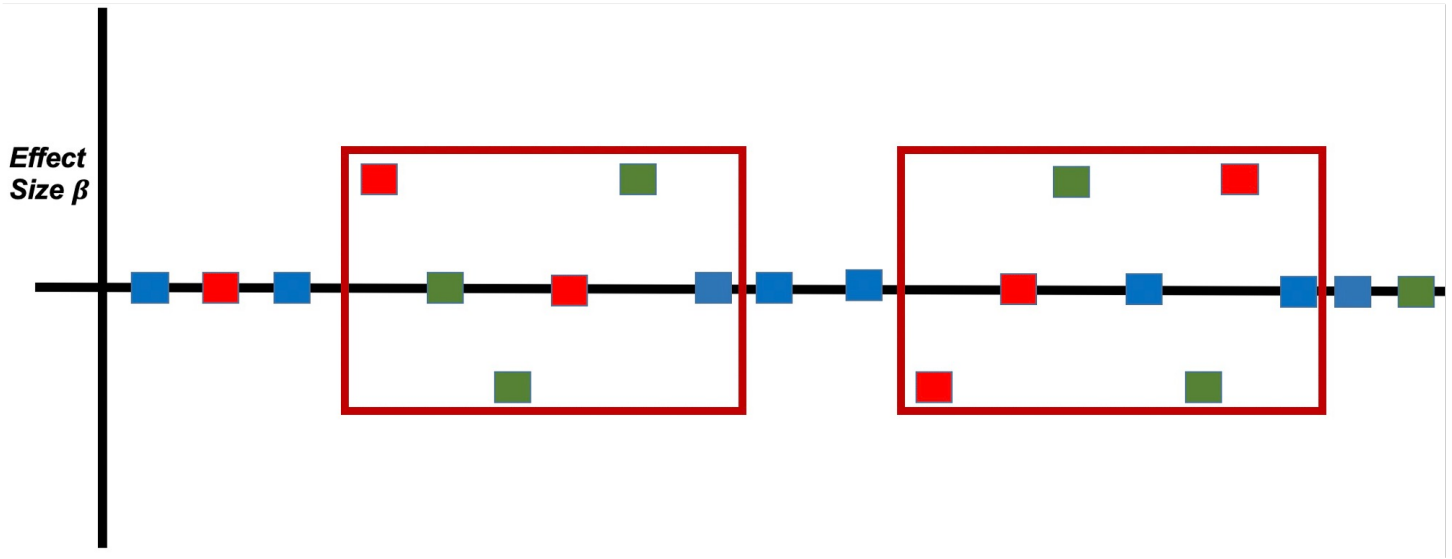
Optimal weighting: True effect sizes (unknown)

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

Use Functional Annotations to Prioritize Variants in a Variant-Set

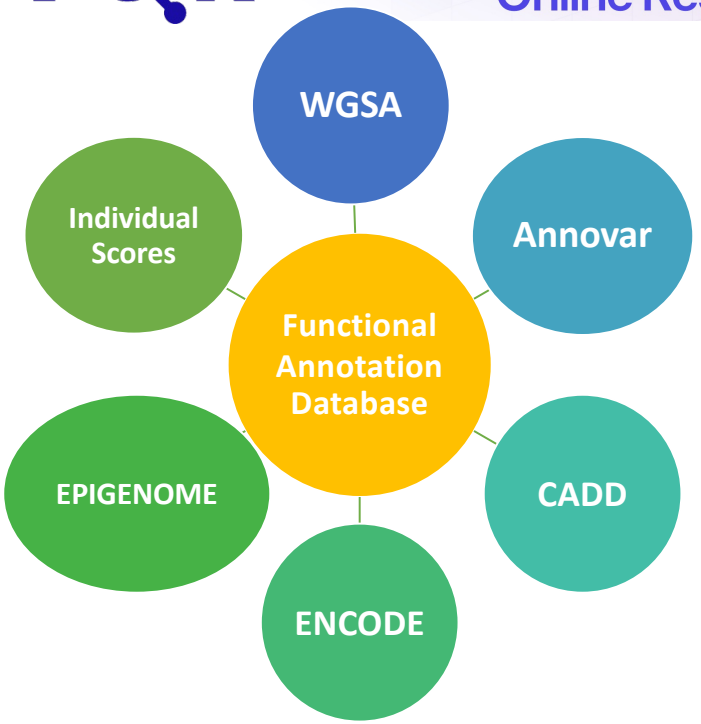


Using multi-dimensional functional annotations to **approximate** the optimal weights (= true effect sizes)

Use Functional Annotations to Prioritize Variants in the Genome

FAVOR

Functional Annotation of Variants - Online Resource



>260 annotations

80% built on hg38

FAVOR Database

- Stable version: https://favor.genohub.org
- New Beta version: https://beta.genohub.org

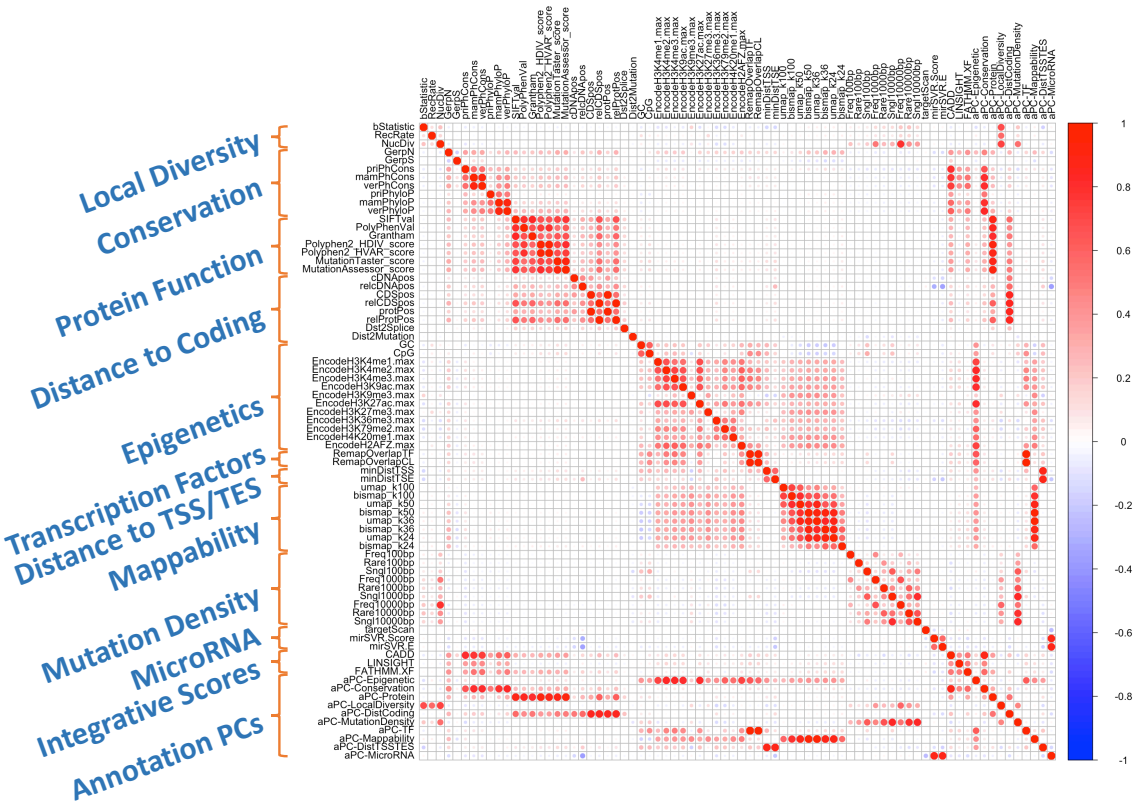
15 Types of Annotations

- 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

Zhou et al. 2023 Nucleic Acids Res

Different annotation scores capture diverse aspects of variant function

Correlation Heatmap with Annotation PCs (TOPMed Freeze 5, hg38)



Li et al. 2020 Nat Genet

10 aPCs: APC-Local Diversity, aPC-conservation, aPC-ProteinFunction, etc.

Variant-Set Tests Weighted by Functional Annotations

- Functionally-informed Burden(B)

$$Q_{Burden} = \left(\sum_{j=1}^{p} \hat{\pi}_{lj} w_j U_j \right)^2$$

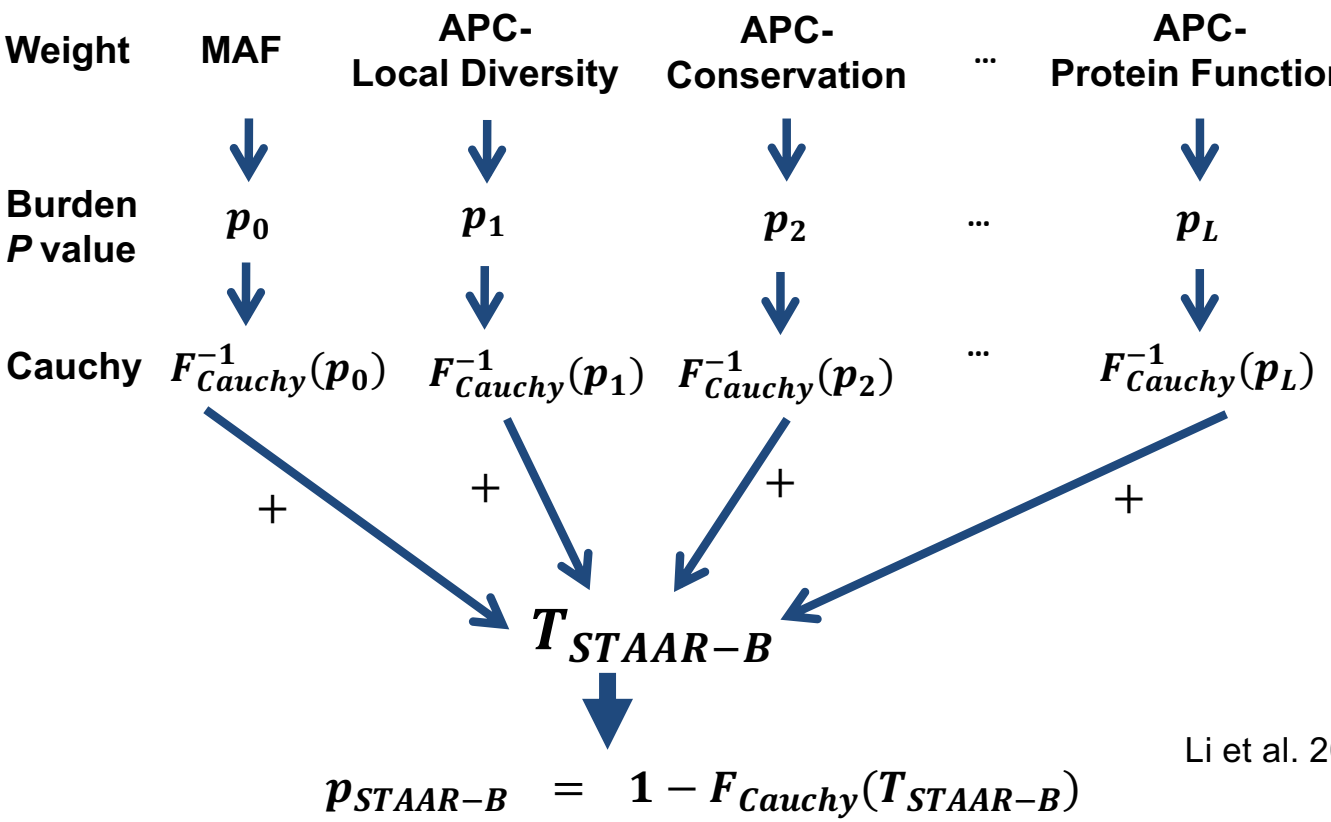
- Functionally-informed SKAT(S)

$$Q_{SKAT} = \sum_{j=1}^{p} \hat{\pi}_{lj} w_j U_j^2$$

- Functionally-informed ACAT-V (A)
- $\hat{\pi}_{lj}$ is the estimated probability of jth variant being causal using the lth annotation

$$\hat{\pi}_{lj} = ECDF(A_{lj}) \propto rank(A_{lj})$$

STAAR: Incorporate Multiple Functional Scores to Empower RV Association Analysis

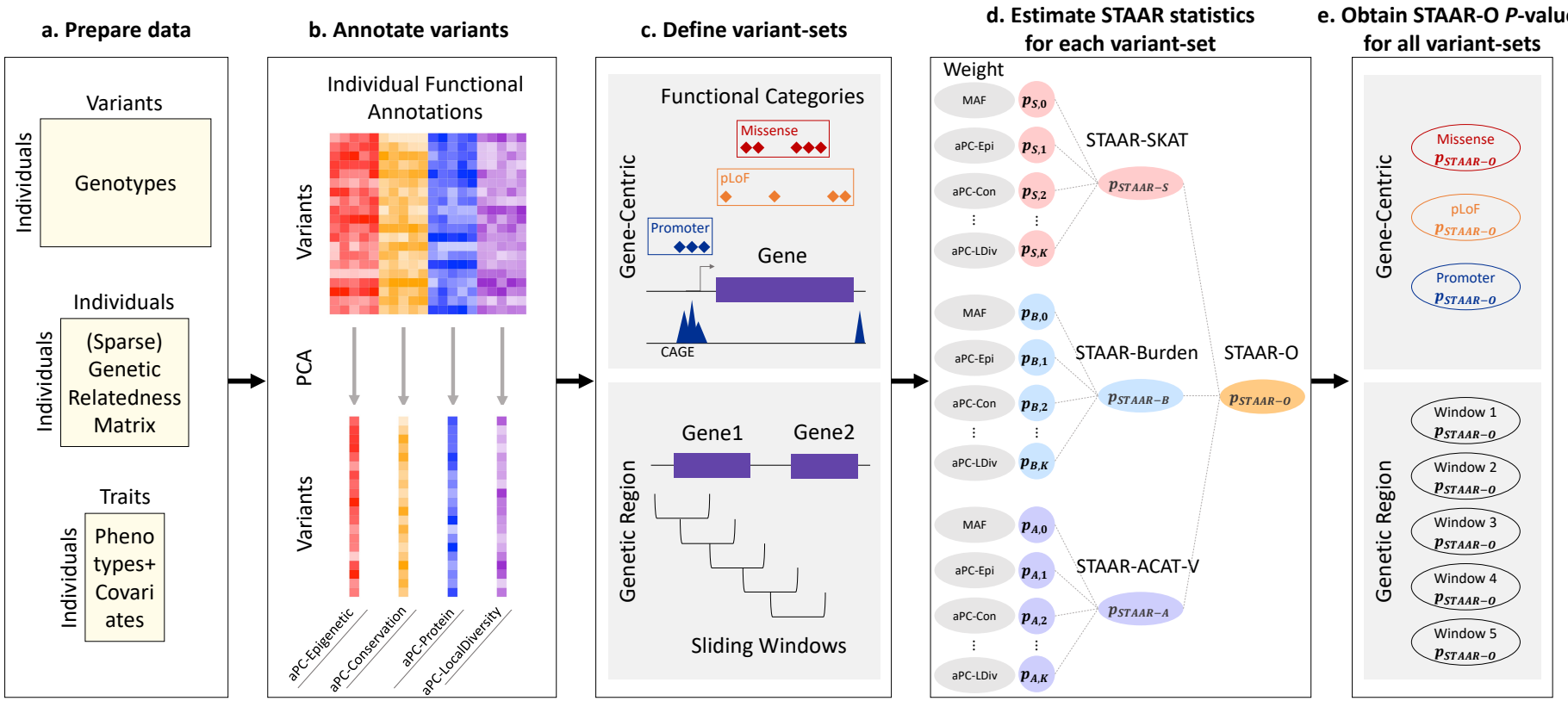


Li et al. 2020 Nat Genet

STAAR: Incorporate Multiple Functional Scores to Empower RV Association Analysis

- STAAR-O is an omnibus test that combines Burden, SKAT and ACAT-V using multiple annotation weights
- By incorporating **different tests**, STAAR-O is robust to the sparsity of causal variants and their directionality of effects
- By incorporating **multiple functional annotations**, STAAR-O is powerful when any of these functional annotations can pinpoint (upweight) causal variants

STAAR workflow



Li et al. 2020 Nat Genet

STAAR: Gene-Centric Genomic Element Analysis of TOPMed WGS data of Lipids Discovery (n=12,316), Replication (n=17,822) and Pooled (n=30,138) (Freeze 5)

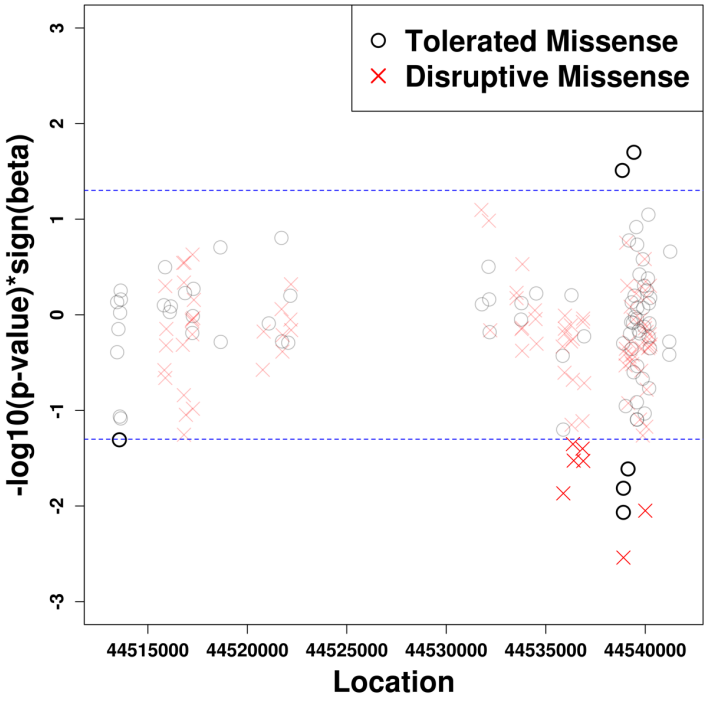
	pLoF	Missense	Synonymous	Promoter	Enhancer
LDL-C	*PCSK9* (***) *APOB* (***)	*PCSK9* (***) *NPC1L1* (**) *LDLR* (***) *APOE* (***)	*SMARCA4* (**)	*QTRT1* (*) *LDLR* (***)	*QTRT1* (*) *LDLR* (***)
HDL-C	*APOC3* (***)	*ABCA1* (**)			
Triglyceride	*APOC3* (***)	*APOC3* (**)			
Total cholesterol	*PCSK9* (***) *APOB* (***)	*PCSK9* (**) *LIPG* (*) *APOE* (*)	*SMARCA4* (**)	*QTRT1* (*) *LDLR* (**)	*QTRT1* (*) *LDLR* (**)

- Genome-wide significant in discovery phase ($p < 2.5\times10^{-6} = 0.05/20{,}000$ genes)
- Replication p-value $\leq 10^{-3}$
- Pooled p-value: “***” $< 10^{-12}$ “**” $< 10^{-9}$ “*” $< 2.5\times10^{-6}$
- rDNase promoters and enhancers

Association Between Disruptive Missense RVs of NPC1L1 and LDL-C in TOPMed Freeze 5 was driven by Aggregated RV Effects, replicated and remains significant in Conditional Analysis

STAAR Disruptive missense

$p = 3.1 \times 10^{-9}$



Adjust for 5 Common Variants in Ference's JACC Paper

Phase	# SNV	Analysis Type	STAAR
Discovery (n=12,316)	94	Unconditional	3.1E-09
		Conditional	1.3E-08
Replication (n=17,822)	129	Unconditional	1.5E-04
		Conditional	2.5E-04
Pooled (n=30,138)	173	Unconditional	8.0E-12
		Conditional	4.5E-11

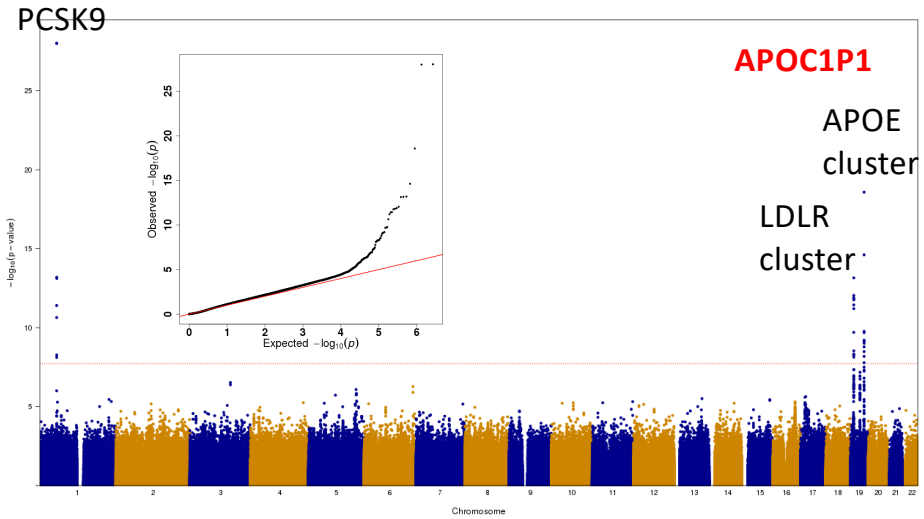
STAAR TOPMed WGS RV Analysis of LDL-C and TG (n=12,316) Manhattan/QQ Plots of 2 kb Sliding Windows

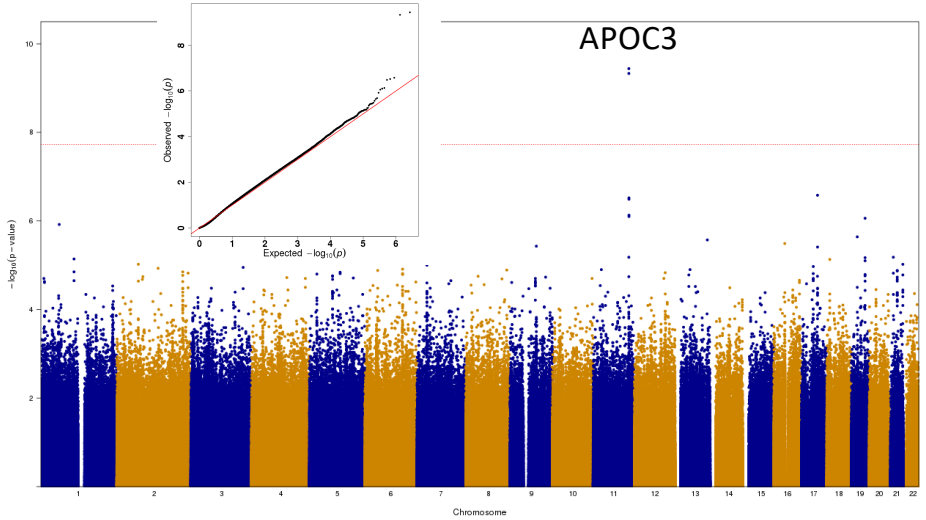
LDL-C

# of Significant ($\alpha = 1.88\times10^{-8}$)	SKAT	Burden	ACAT-V	STAAR-O
2 kb Sliding window (2.6 M)	21	0	31	33

TG

# of Significant ($\alpha = 1.88\times10^{-8}$)	SKAT	Burden	ACAT-V	STAAR-O
2 kb Sliding window (2.6 M)	0	0	2	2





Scalability: Whole Genome RVAS Computational Speed

STAAR-O	n=30,000*	n=500,000**
Gene Centric	0.6 hrs	1 hrs
Genetic Region	15 hrs	25 hrs

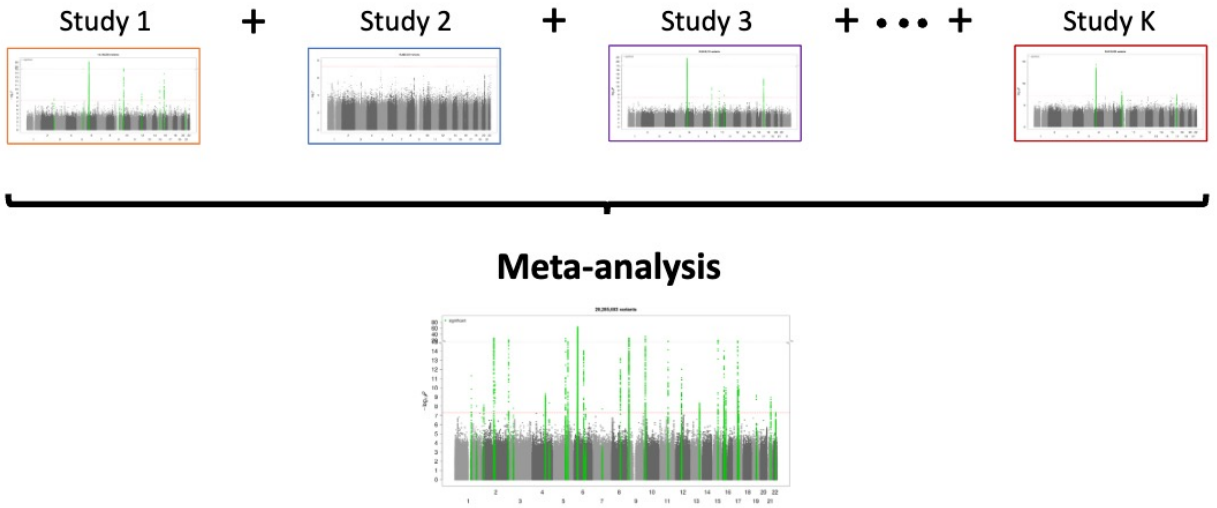
* Lipid Data (TOPMed Freeze 5) ** Simulated Data and > 1 billion variants

- **Use Sparse GRM to fit the null GLMM and scan the genome**
- **Computation time using 100 CPUs**
- **20,000 genes x 5 categories in gene centric analysis**
- **2.5 million sliding windows in genetic region analysis**

From One Study to Multiple Studies

Benefits of meta-analysis in genetic association studies :

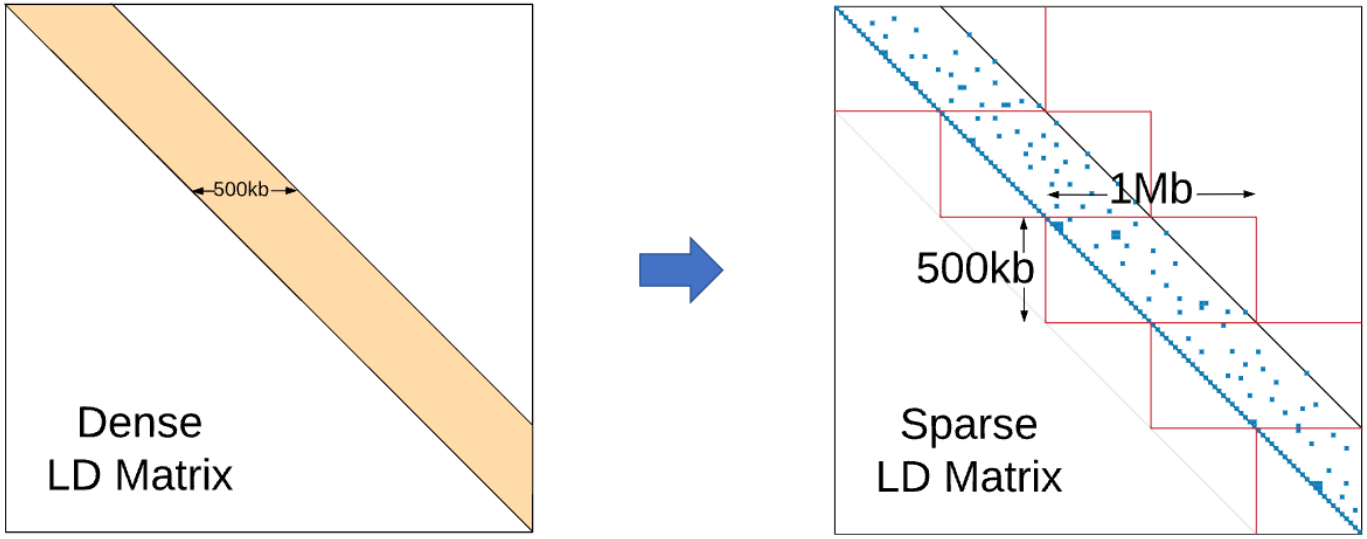
- Boost statistical power for association analysis ;
- Increase sample sizes, especially for underrepresented populations ;
- Protect data privacy by only sharing summary statistics



Zhou et al. 2022 Cell Genomics

Meta-Analysis of Rare Variant Association Tests

- RareMetal (Feng et al. 2014, Liu et al. 2014), MetaSKAT (Lee et al. 2013), SMMAT-meta (Chen et al. 2019)
- **Main challenge :** store and share summary-level data (score statistics and LD matrices) of rare variants

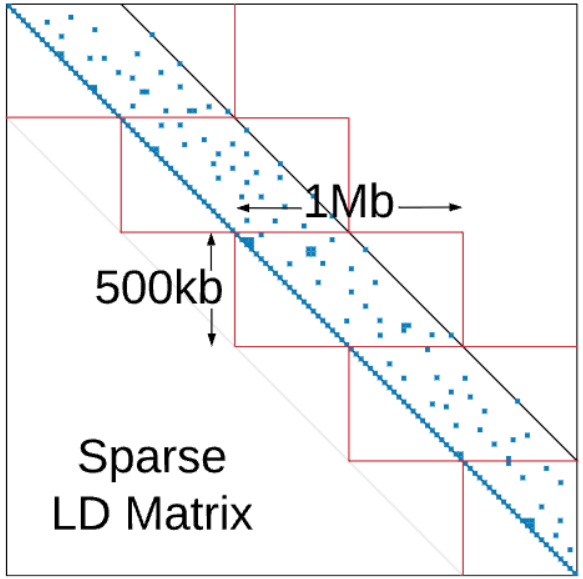


Existing methods

Our proposed method (MetaSTAAR)

Efficient Storage of Rare Variant Summary Statistics: Shared Data for Each Study using MetaSTAARWorker

Chr	Pos	Ref	Alt	Score	Var	X_1	X_2	X_3	...
1	10498	G	T	0.03	5.7E-04	1.9E-04	-1.5E-07	-3.1E-05	...
1	10534	A	G	0.03	5.8E-04	1.9E-04	-1.5E-05	2.0E-05	...
1	13054	C	T	-0.03	6.1E-04	1.4E-03	2.1E-05	2.2E-04	...
1	13213	C	G	0.02	5.8E-04	1.9E-04	-4.5E-05	3.3E-05	...
1	13216	C	A	-0.02	1.1E-03	3.8E-04	-4.2E-05	-5.7E-05	...
1	13324	C	A	0.08	2.3E-03	3.9E-04	-8.6E-05	-1.9E-04	...
...	...	...	...	...	...	...	...	...	...



Single-variant meta-analysis summary statistics

$G^T\widehat{\Sigma}^{-1}X\left(X^T\widehat{\Sigma}^{-1}X\right)^{-1/2}$

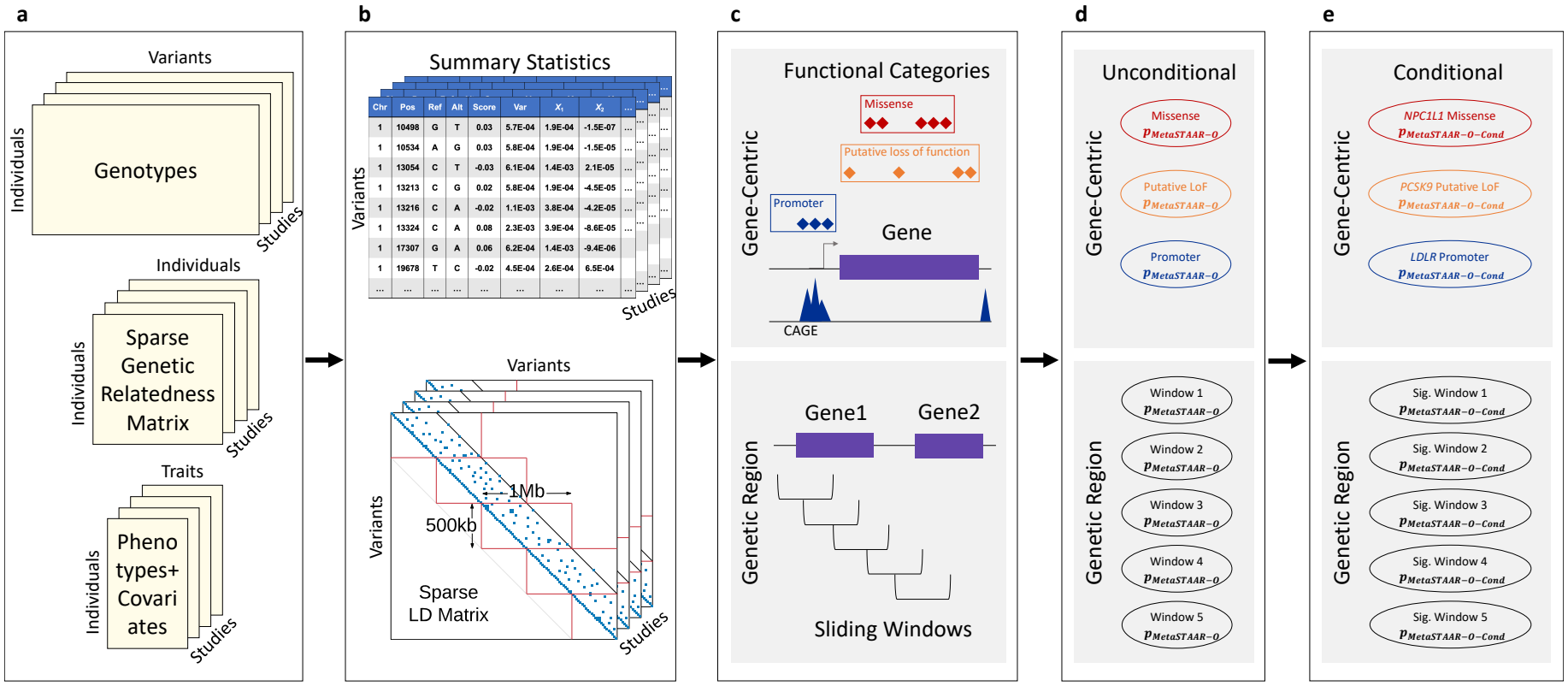
$G^T\widehat{\Sigma}^{-1}G$ stored as banded sparse matrix

Storage and Computational Time of MetaSTAARWorker are 100x Smaller Than RareMetalWorker using TOPMed Data

Type	Trait: TC	RareMetal-Worker	MetaSTAAR-Worker	Compression
Storage (GB)	500kb banded region**	10.14	0.03	340
	Whole Genome	> 58,437***	570	> 100
Time (CPU hrs)	500kb banded region**	69.94	0.21	330
	Whole Genome	> 418,870***	4,200	> 100

- * A collection of 14 multi-ethnic studies ($n = 30,138$) with a total of $\sim$ 250 million variants, using Total Cholesterol (TC) phenotype for illustration
- ** 500kb banded region in chromosome 6 : 160Mb - 161Mb
- *** Conservatively estimated storage and computation time for the whole genome

MetaSTAAR WGS Rare Variant Analysis Workflow



Li et al. 2023 Nat Genet

MetaSTAAR-O Showed Consistent Unconditional and Conditional Results of TC as Pooled Analysis using STAAR-O

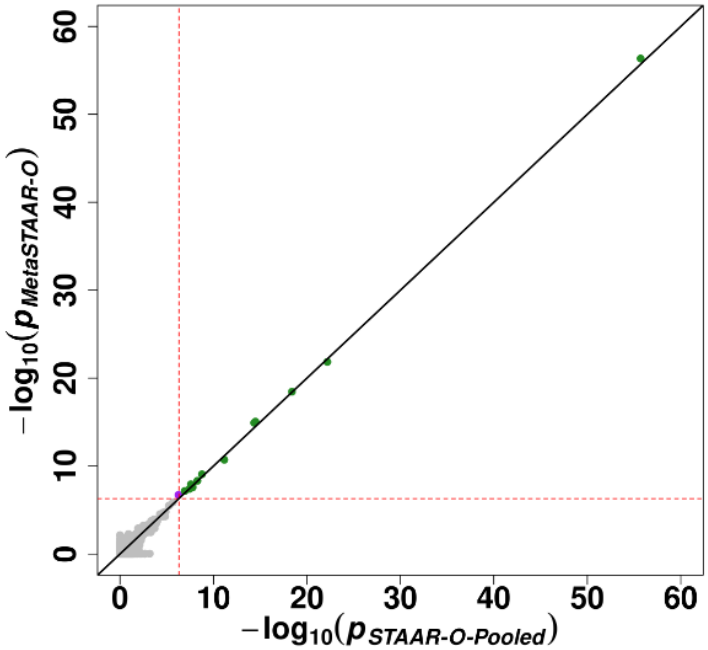


Figure – Gene-centric unconditional results of total cholesterol ($r^2 > 0.99$).

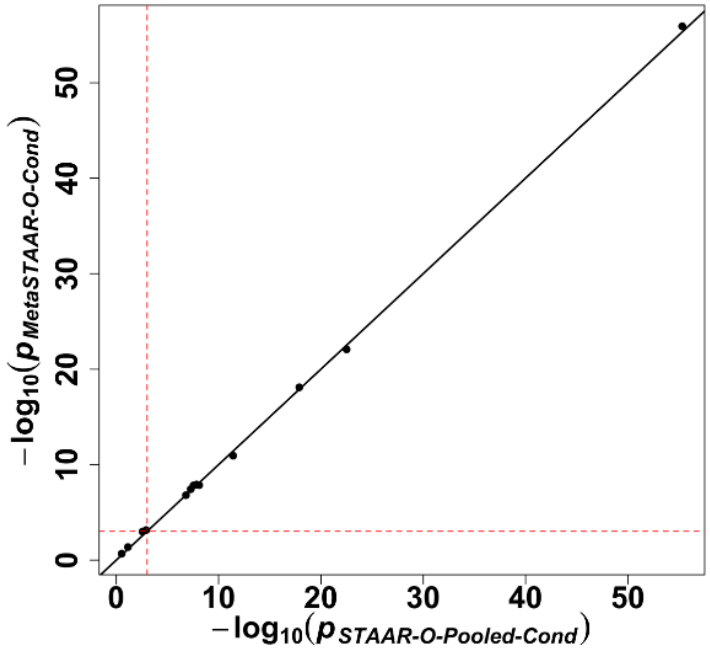


Figure – Gene-centric conditional results of total cholesterol ($r^2 > 0.99$). Significant associations in unconditional analysis were used in the comparison.

Application of MetaSTAAR to TOPMed and UK Biobank Data

- **TOPMed and UK Biobank Meta-Analysis**
- Phenotypes : Total Cholesterol (TC)
- Sample sizes :
  - **TOPMed Freeze 5 : considered as one study ($n = 30,138$)**
  - **UK Biobank WES ($n = 190,415$)**
- Rare variant (MAF < 1%) association meta-analysis :
  - Gene-centric meta-analysis :
    - 3 coding categories : pLoF, missense, synonymous

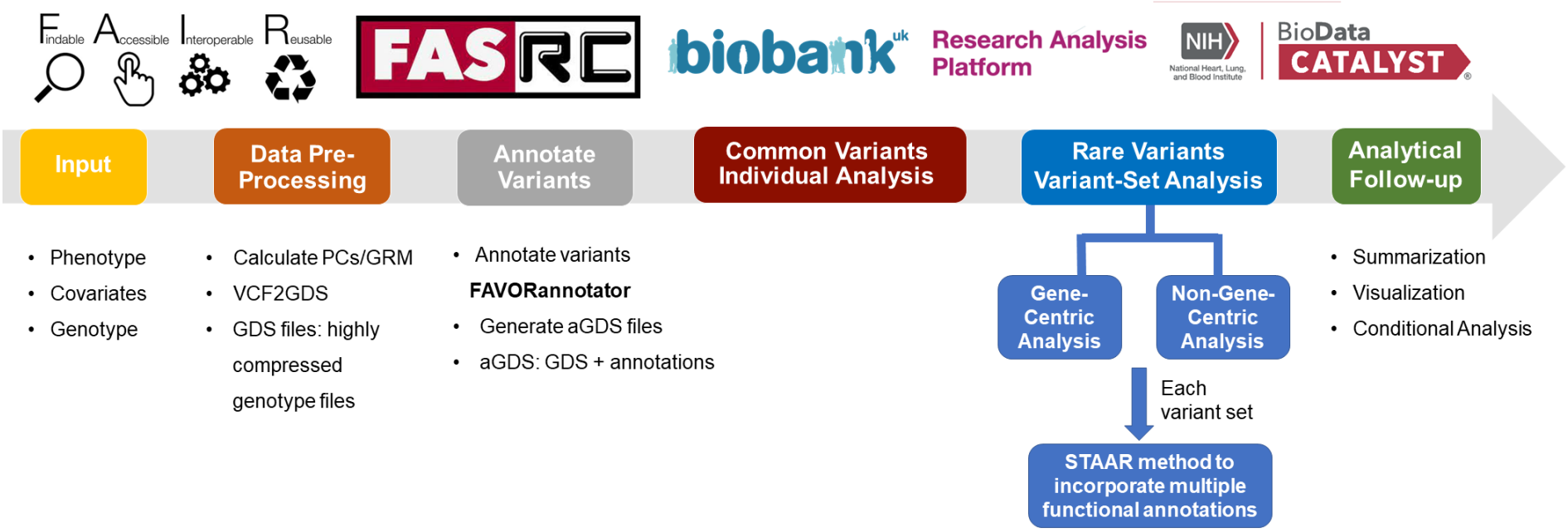
Trait	Gene	Chromosome no.	Category	TOPMed meta-analysis			TOPMed and UK Biobank meta-analysis		
				No. of SNVs	MetaSTAAR-O (unconditional)	MetaSTAAR-O (conditional)	No. of SNVs	MetaSTAAR-O (unconditional)	MetaSTAAR-O (conditional)
TC	*PCSK9*	1	Putative loss of function	9	4.46E-57	1.23E-56	27	4.45E-124	7.97E-125
	APOB	2	Putative loss of function	16	3.52E-19	7.85E-19	66	2.80E-138	4.63E-137
	PCSK9	1	Missense	169	1.94E-11	1.15E-11	414	3.55E-72	6.27E-71
	ABCG5	2	Missense	157	4.74E-09	1.21E-08	400	7.91E-27	1.50E-24
	NPC1L1	7	Missense	301	5.19E-08	2.08E-07	676	2.98E-13	1.69E-12
	ABCA1	9	Missense	346	6.89E-08	3.72E-08	1042	4.14E-33	5.85E-34
	LIPG	18	Missense	101	2.69E-08	1.39E-08	293	4.04E-171	2.07E-171
	LDLR	19	Missense	200	1.41E-22	8.33E-23	488	7.98E-81	1.16E-77
	APOE	19	Missense	90	1.18E-08	1.46E-08	213	9.15E-17	1.47E-12

MetaSTAAR Analyzes Biobank Data at Scale (n > 220, 000)

Total Cholesterol	UK Biobank Storage		MetaSTAAR-O	
Analysis Type	SumStat	Sparse LD	Time	Memory
Fitting Null Model	-	-	3 min	12 GB
MetaSTAARWorker	1.6 GB	1.2 GB*	3 hrs**	12 GB
Association Analysis	-	-	3 hrs***	12 GB

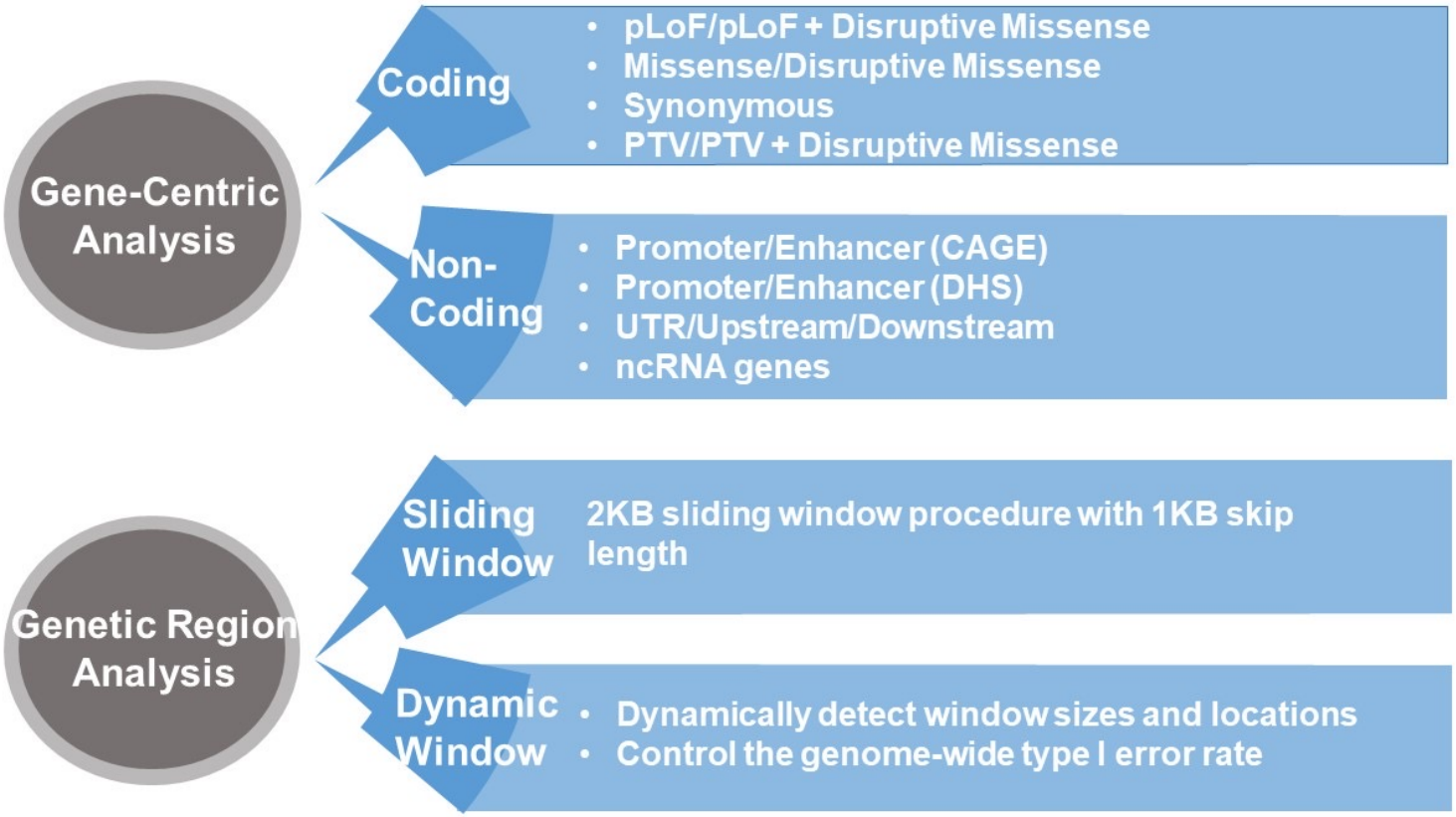
- * Sparse LD matrices storage is **even less** than the summary statistics files
- ** Full computation time for association analysis based on 100 CPUs
- *** Computation was benchmarked on TOPMed and UKB meta-analysis
- Computation time and memory also depend on the sparsity of GRM

STAARpipeline: Functionally-Informed WGS Analysis Pipeline



Li et al. 2022 Nat Methods

STAARpipeline: Functionally-Informed WGS Analysis Pipeline



Li et al. 2022 Nat Methods

Other available softwares and tools for rare variant analysis

- Single-trait rare variant analysis:
  - BOLT-LMM (Loh et al. 2015, 2018 Nat Genet)
  - fastGWA (Jiang et al. 2019, 2021 Nat Genet)
  - GENESIS (Gogarten et al. 2019 Bioinformatics)
  - REGENIE >= v3.0 (Mbatchou et al. 2021 Nat Genet)
  - SAIGE-GENE+ (Zhou et al. 2022 Nat Genet)
- Multi-trait rare variant analysis:
  - GAMuT (Broadaway et al. 2016 AJHG)
  - Multi-SKAT (Dutta et al. 2019 Genet Epidemiol)
  - MultiSTAAR (Li et al. 2023 bioRxiv)

Open questions for rare variants: missing heritability

nature genetics

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature genetics › articles › article

Article | Published: 07 March 2022

Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data

Pierrick Wainschtein ✉, Deepti Jain, Zhili Zheng, TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, L. Adrienne Cupples, Aladdin H. Shadyab, Barbara McKnight, Benjamin M. Shoemaker, Braxton D. Mitchell, Bruce M. Psaty, Charles Kooperberg, Ching-Ti Liu, Christine M. Albert, Dan Roden, Daniel I. Chasman, Dawood Darbar, Donald M. Lloyd-Jones, Donna K. Arnett, Elizabeth A. Regan, Eric Boerwinkle, Jerome I. Rotter, Jeffrey R. O'Connell, Lisa R. Yanek, ... Peter M. Visscher ✉ + Show authors

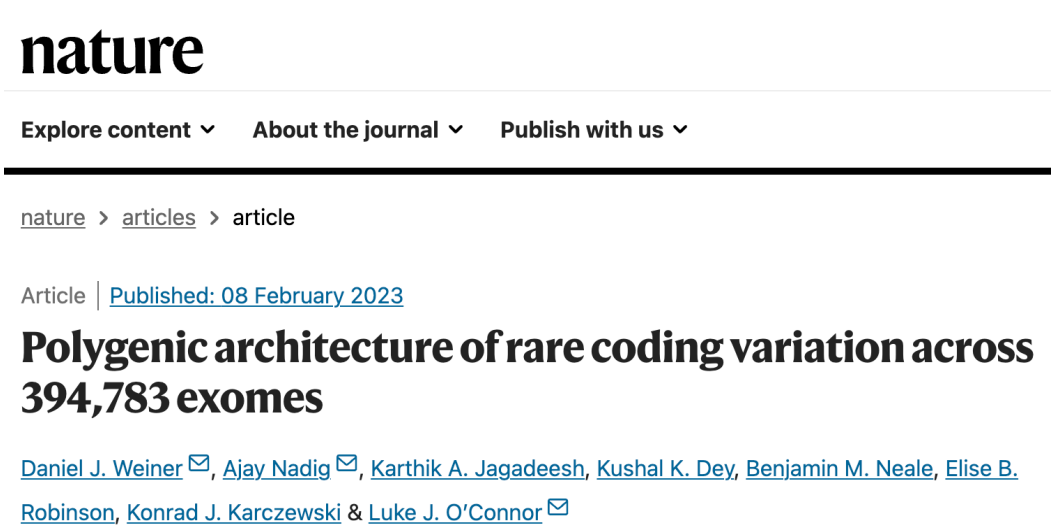
Wainschtein et al. 2022 Nat Genet

Here we estimated heritability for height and body mass index (BMI) from whole-genome sequence data on 25,465 unrelated individuals of European ancestry. The estimated heritability was **0.68** (standard error 0.10) for height and **0.30** (standard error 0.10) for body mass index.

Low minor allele frequency variants in low linkage disequilibrium (LD) with neighboring variants were enriched for heritability, to a greater extent for protein-altering variants, consistent with negative selection.

Our results imply that **rare variants, in particular those in regions of low linkage disequilibrium, are a major source of the still missing heritability of complex traits and disease**.

Open questions for rare variant analysis: missing heritability



nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature › articles › article

Article | Published: 08 February 2023

Polygenic architecture of rare coding variation across 394,783 exomes

Daniel J. Weiner ✉, Ajay Nadig ✉, Karthik A. Jagadeesh, Kushal K. Dey, Benjamin M. Neale, Elise B. Robinson, Konrad J. Karczewski & Luke J. O'Connor ✉

Weiner et al. 2023 Nature

Rare coding variants (allele frequency < 1 × 10−3) explain **1.3%** (s.e. = 0.03%) of phenotypic variance on average—much less than common variants—and most burden heritability is explained by ultrarare loss-of-function variants (allele frequency < 1 × 10−5).

We find that burden heritability for schizophrenia and bipolar disorder is approximately **2%**.

Our results indicate that **rare coding variants will implicate a tractable number of large-effect genes, that common and rare associations are mechanistically convergent, and that rare coding variants will contribute only modestly to missing heritability and population risk stratification**.