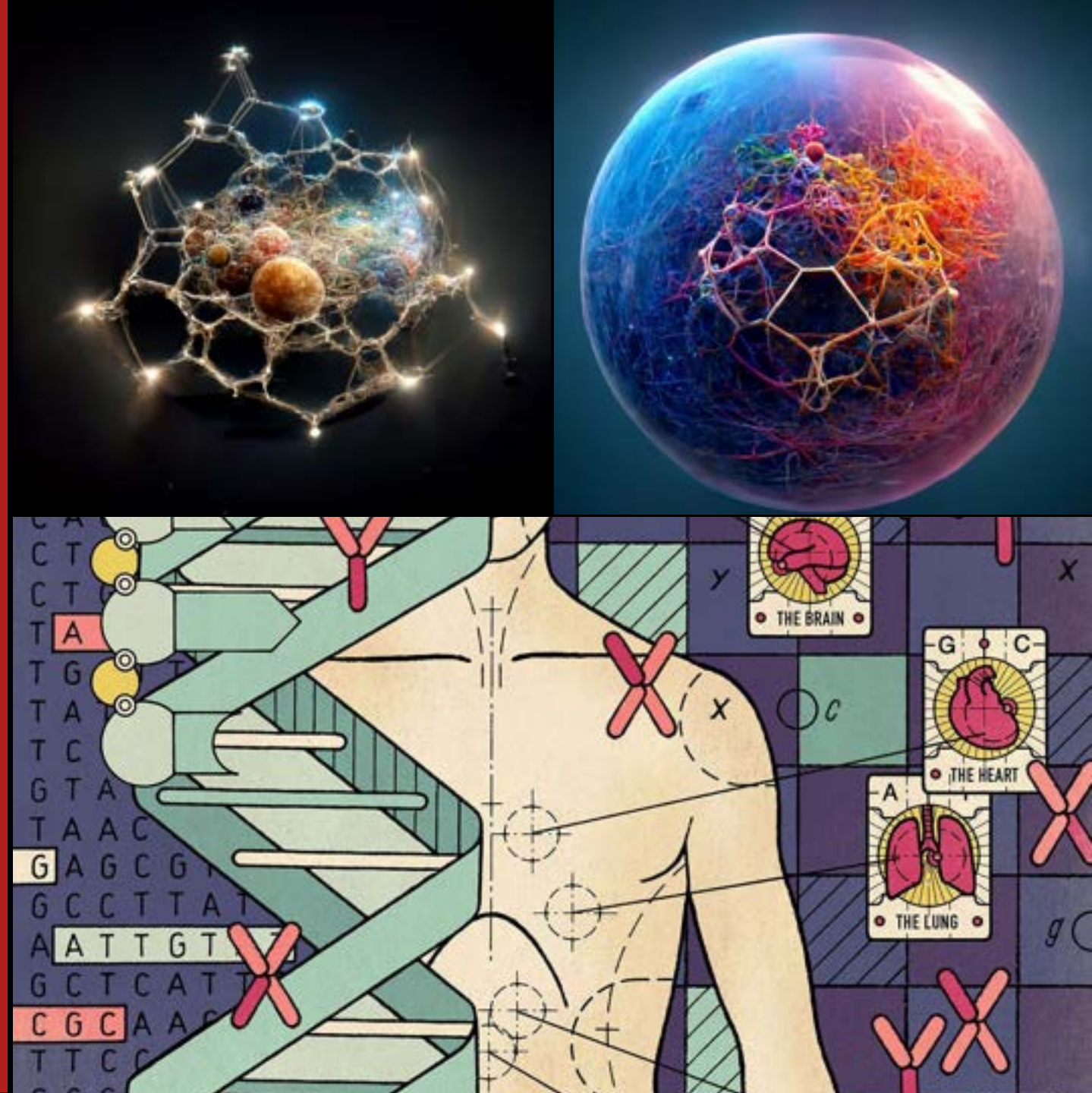


Annotation-aware genetic analysis

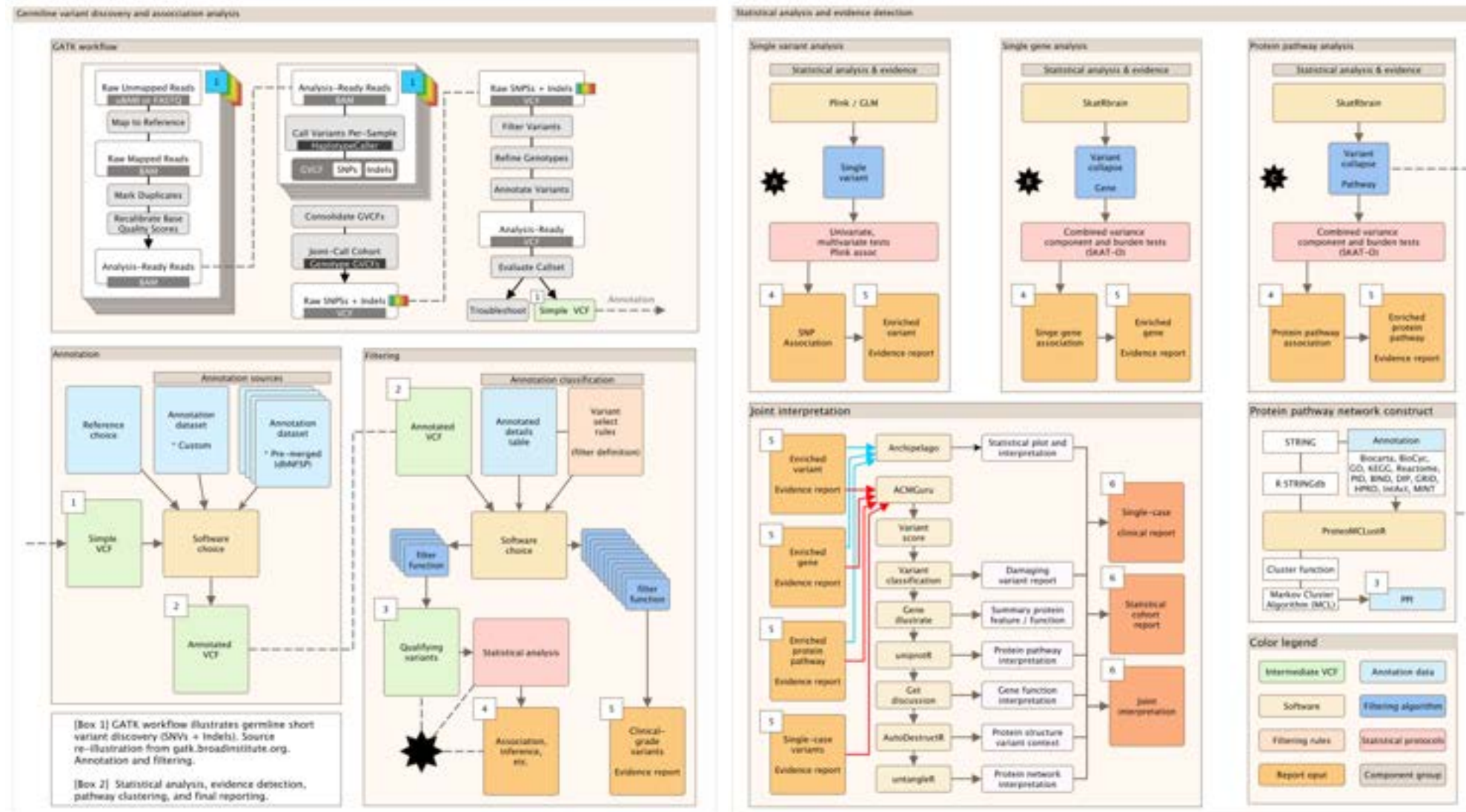


Cauchy–Lorentz distribution

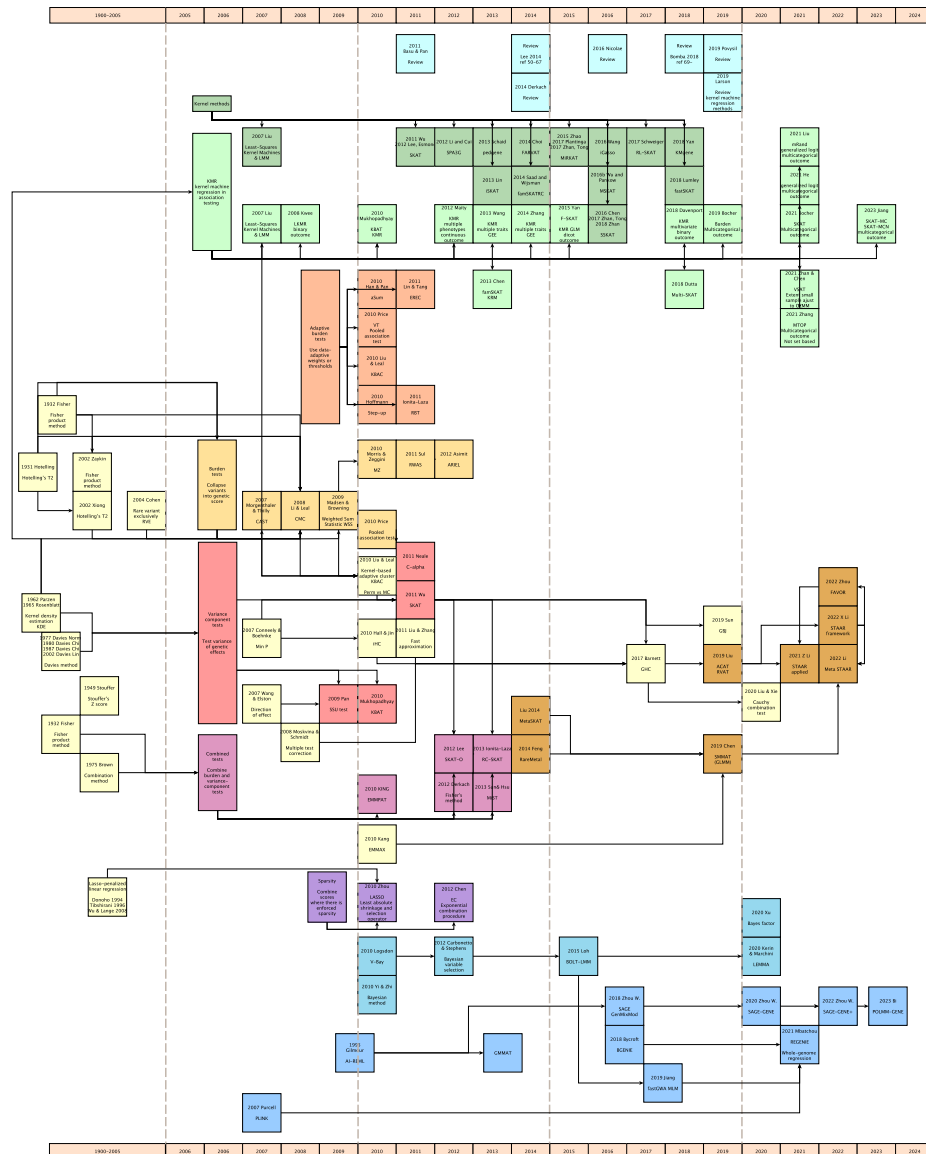
The Witch
of Agnesi



Model application

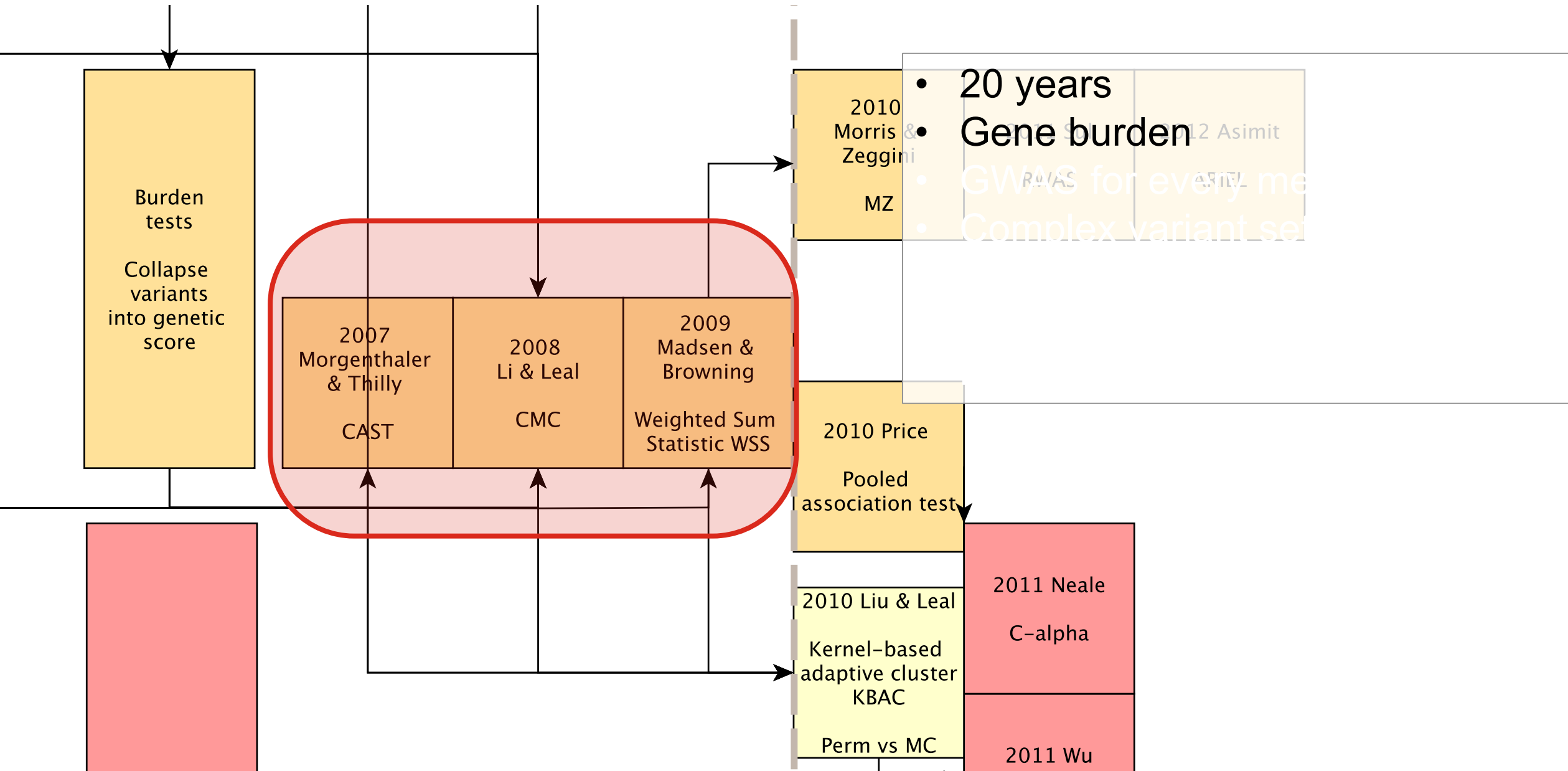


Variant set analysis

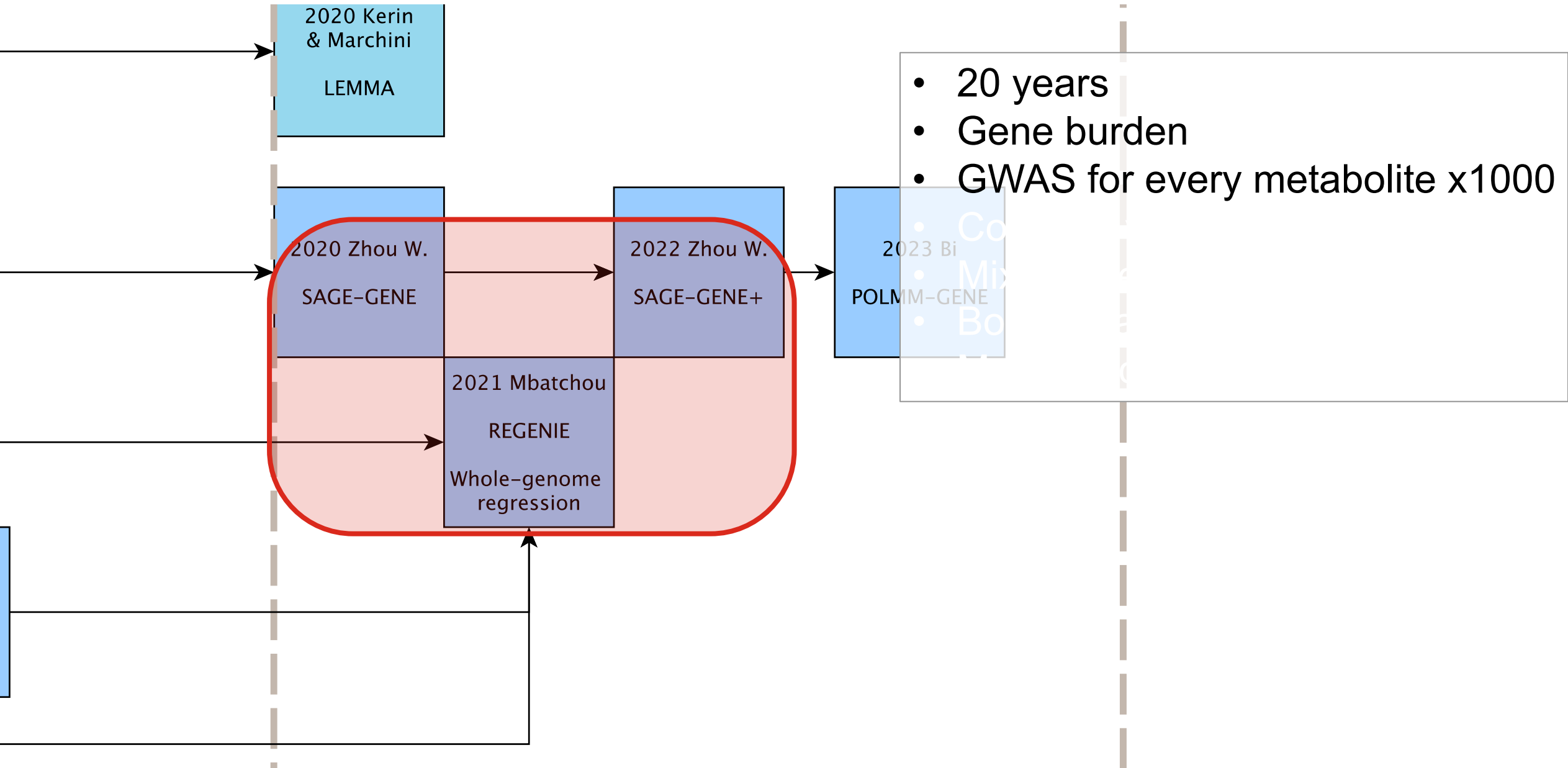


- 20 years

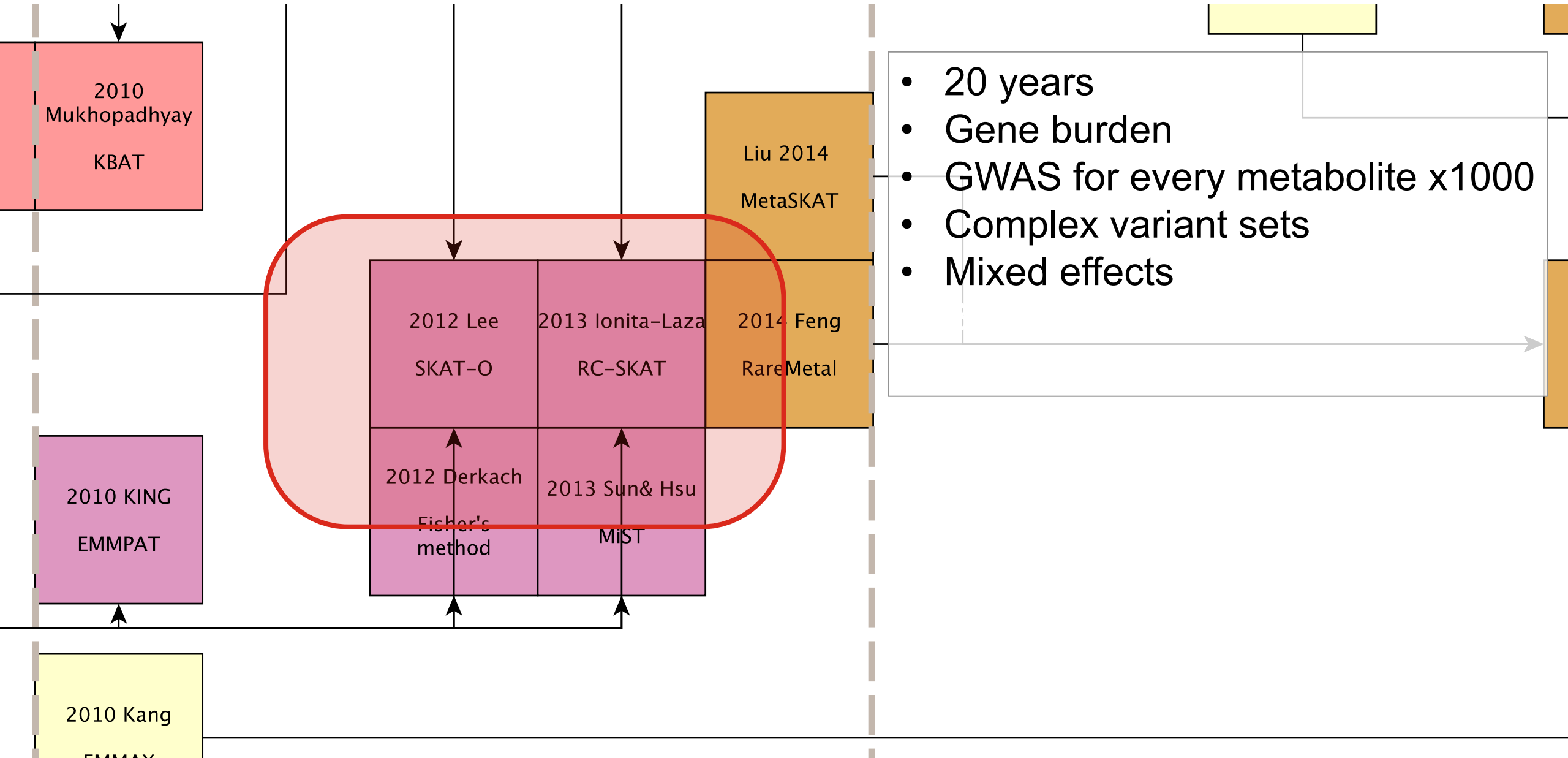
Variant set analysis



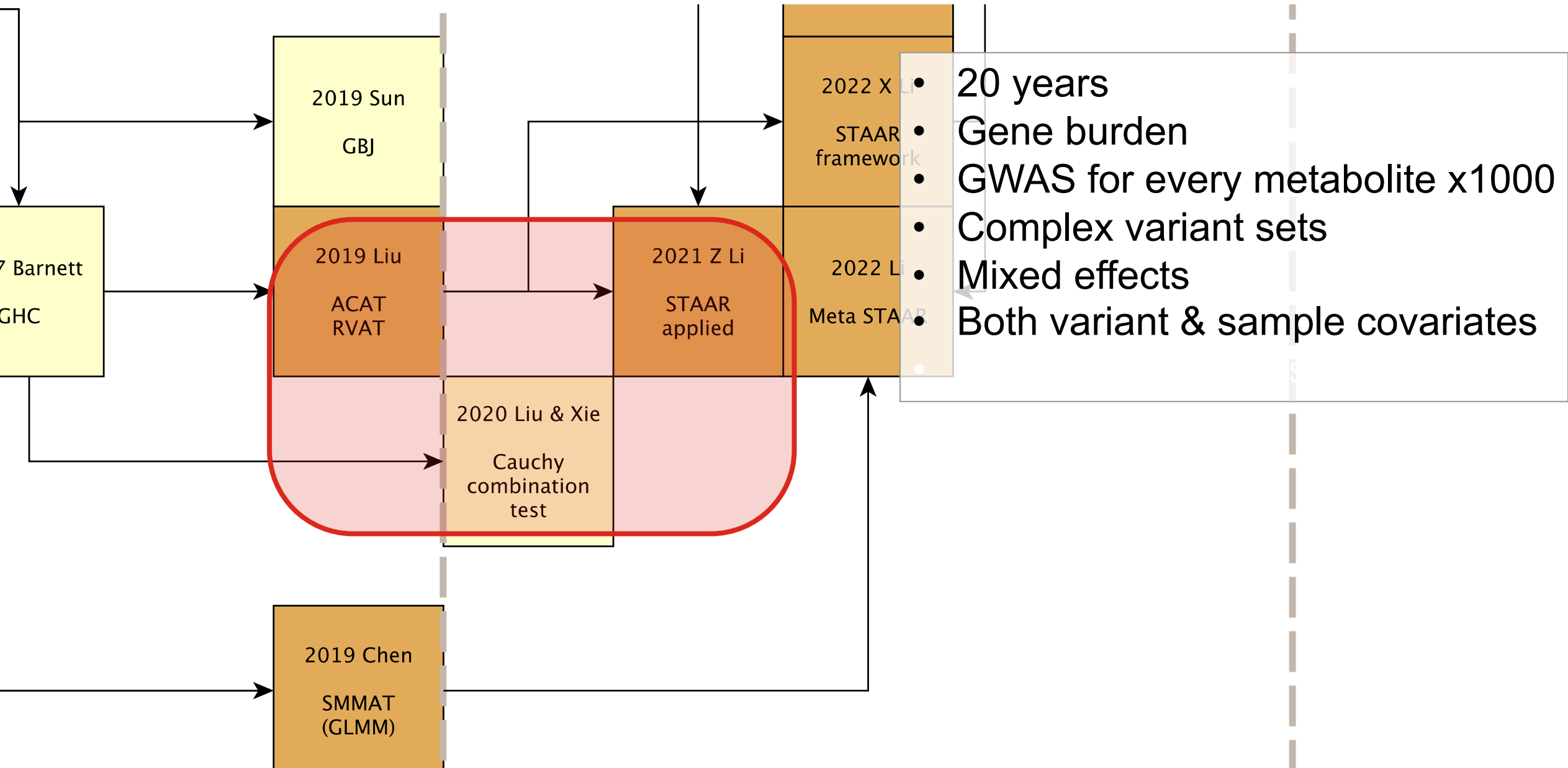
Variant set analysis



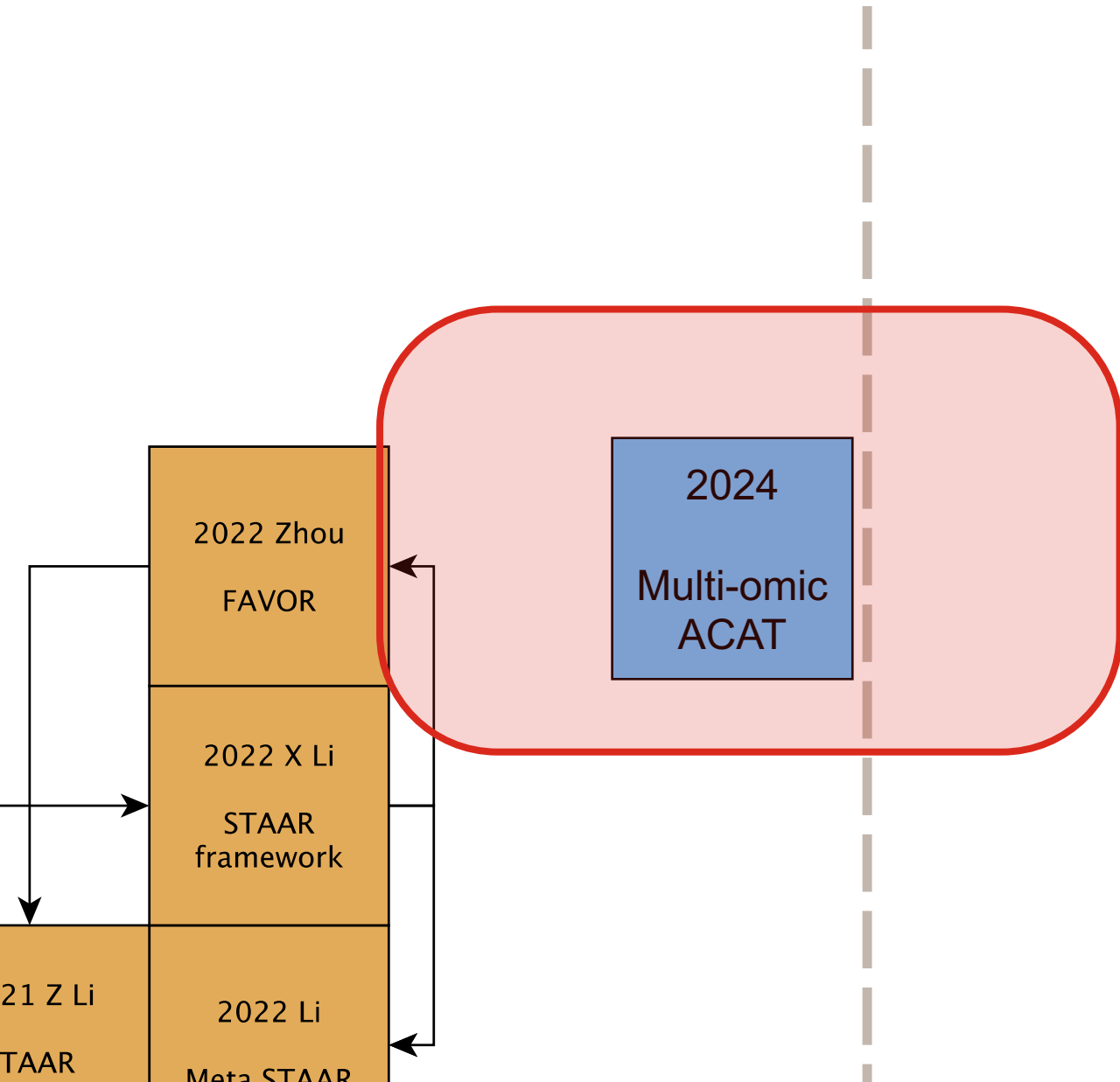
Variant set analysis



Variant set analysis

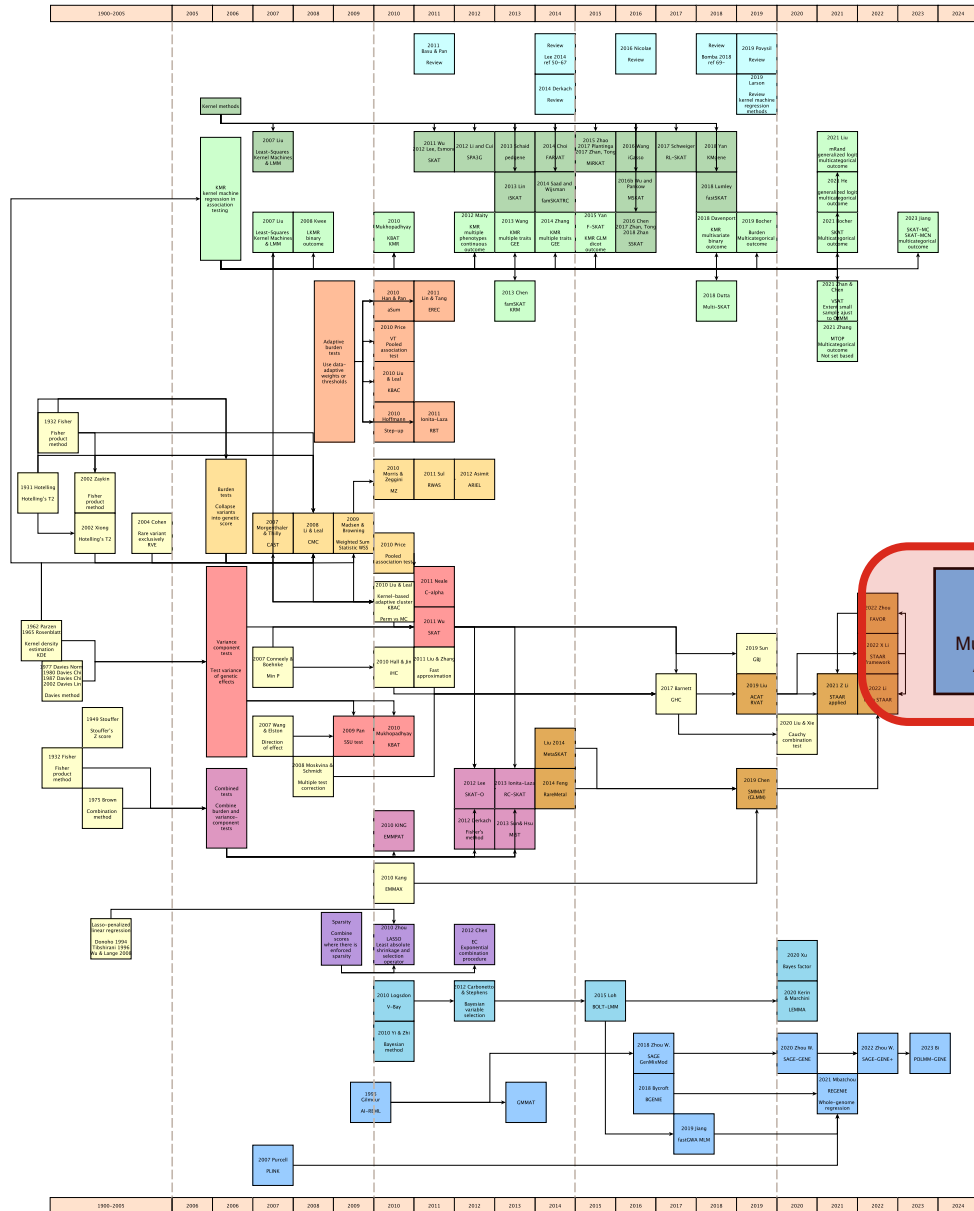


Variant set analysis



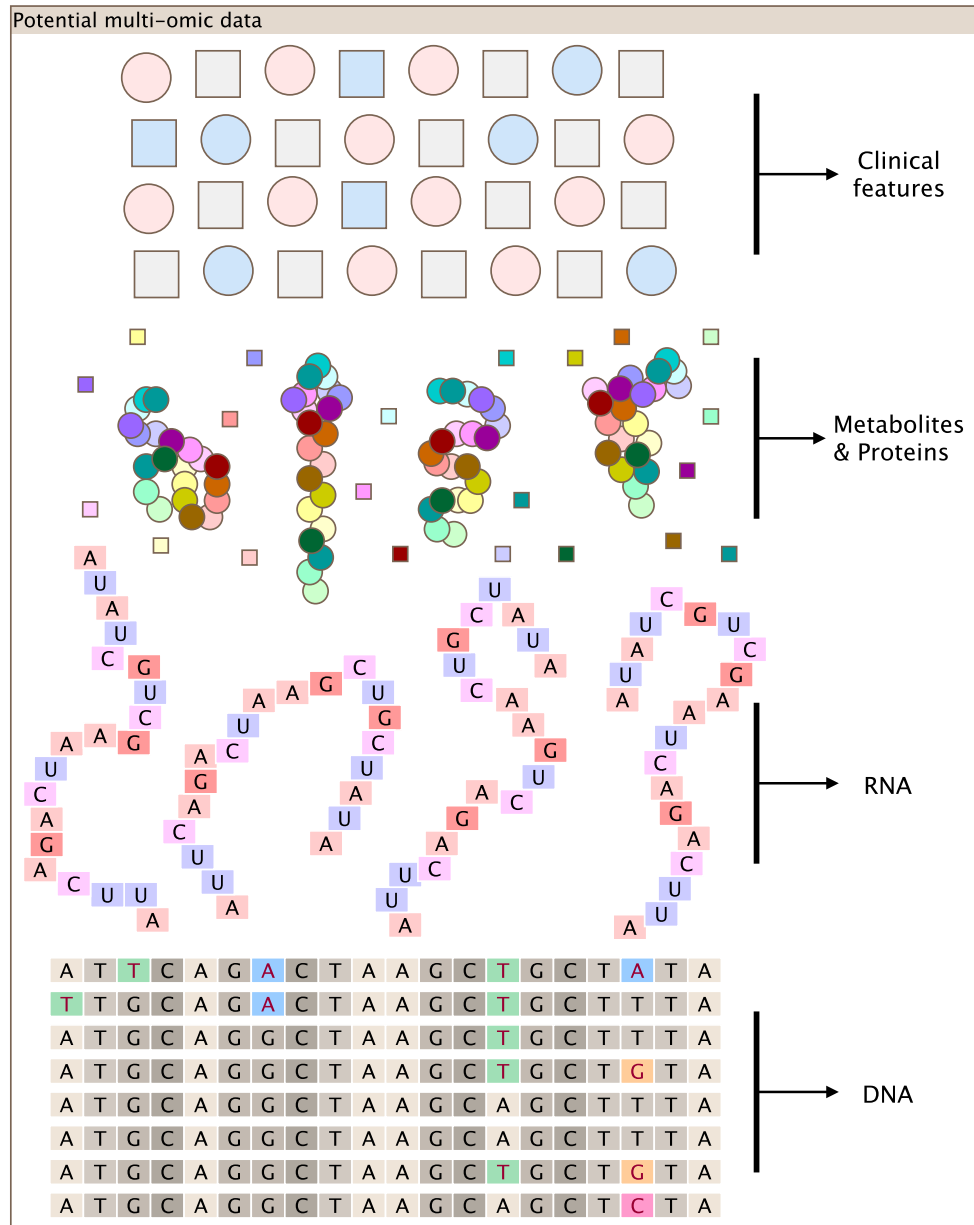
- 20 years
- Gene burden
- GWAS for every metabolite x1000
- Complex variant sets
- Mixed effects
- Both variant & sample covariates
- Multi-omic variables

Variant set analysis



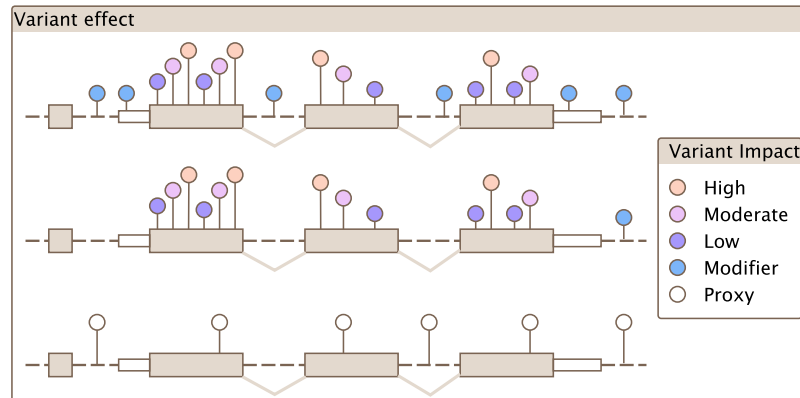
- 20 years
- Gene burden
- GWAS for every metabolite x1000
- Complex variant sets
- Mixed effects
- Both variant & sample covariates
- Multi-omic variables

Genetic association with disease



- Genetic variant $\rightarrow$ disease
- Include other biological info in model

Genetic association with disease



Sequence dataset

S1	A	T	T	C	A	G	A	C	T	A	A	G	C	T	G	C	T	A	T	A
S2	T	T	G	C	A	G	A	C	T	A	A	G	C	T	G	C	T	T	T	A
S3	A	T	G	C	A	G	G	C	T	A	A	G	C	T	G	C	T	T	T	A
...	A	T	G	C	A	G	G	C	T	A	A	G	C	T	G	C	T	T	T	A
...	A	T	G	C	A	G	G	C	T	A	A	G	C	A	G	C	T	T	T	A
...	A	T	G	C	A	G	G	C	T	A	A	G	C	A	G	C	T	T	T	A
...	A	T	G	C	A	G	G	C	T	A	A	G	C	T	G	C	T	G	T	A
Sn	A	T	G	C	A	G	G	C	T	A	A	G	C	A	G	C	T	C	T	A

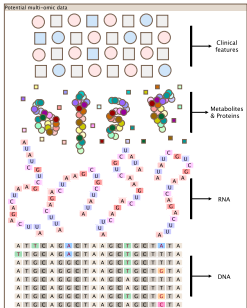
Very rare Rare shared Common Multiple rare

Decomposed genotype matrix

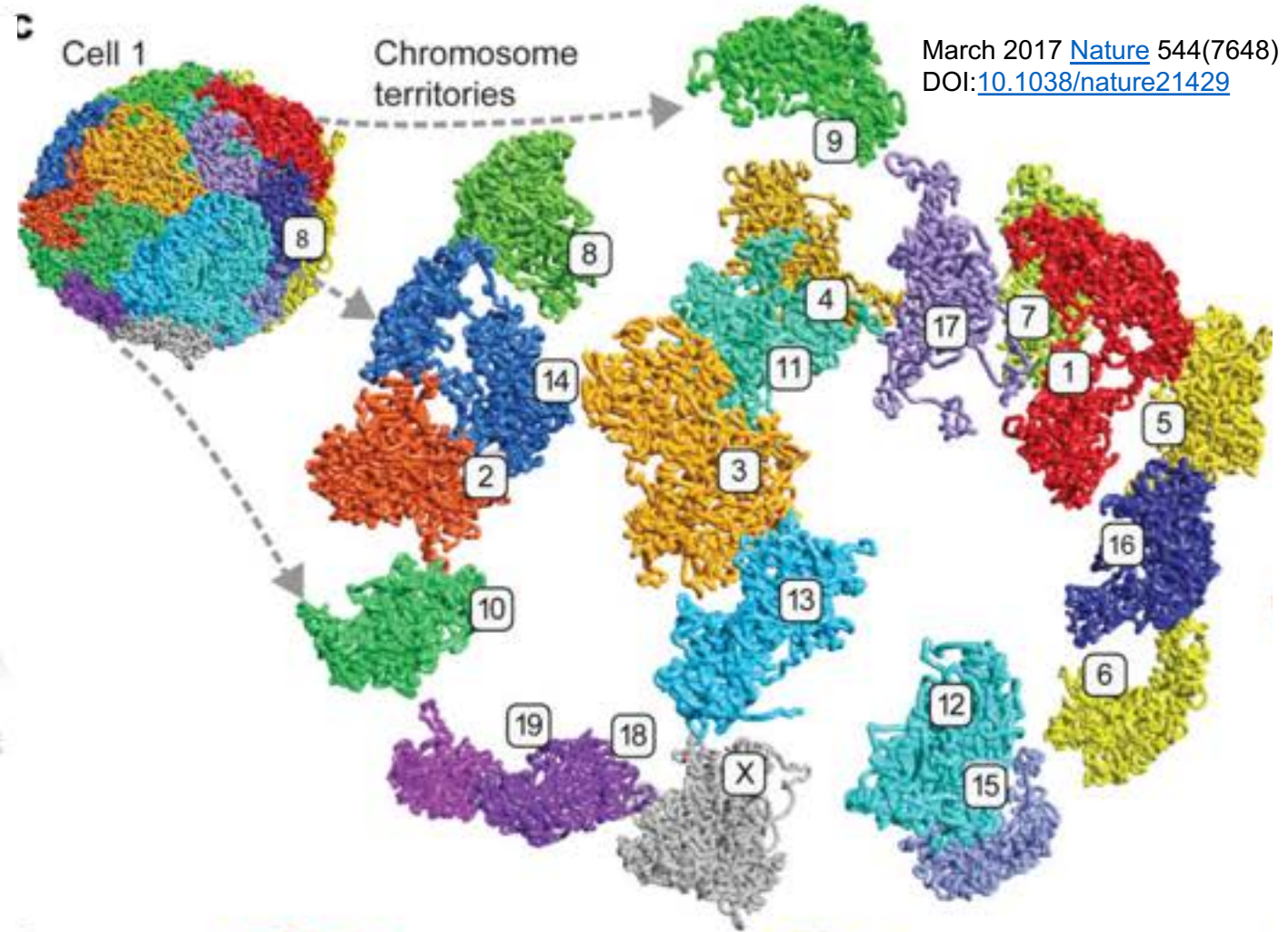
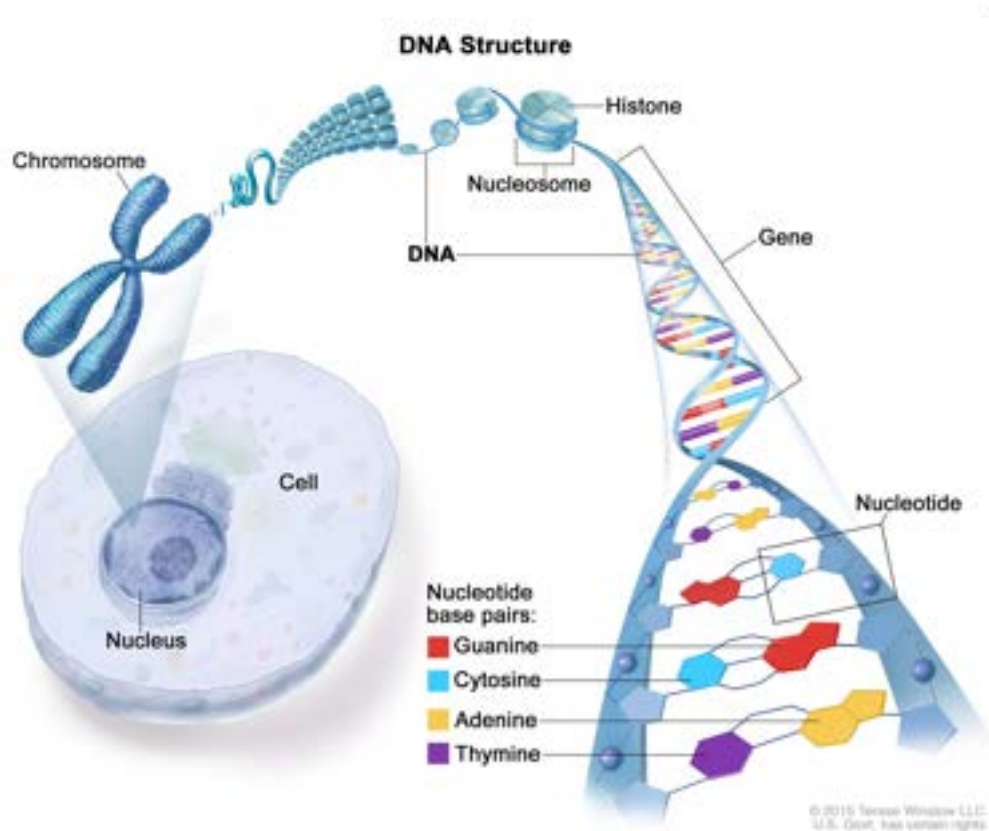
S1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
S2	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
S3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Sn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Very rare Rare shared Common Decomposed multi-allele*

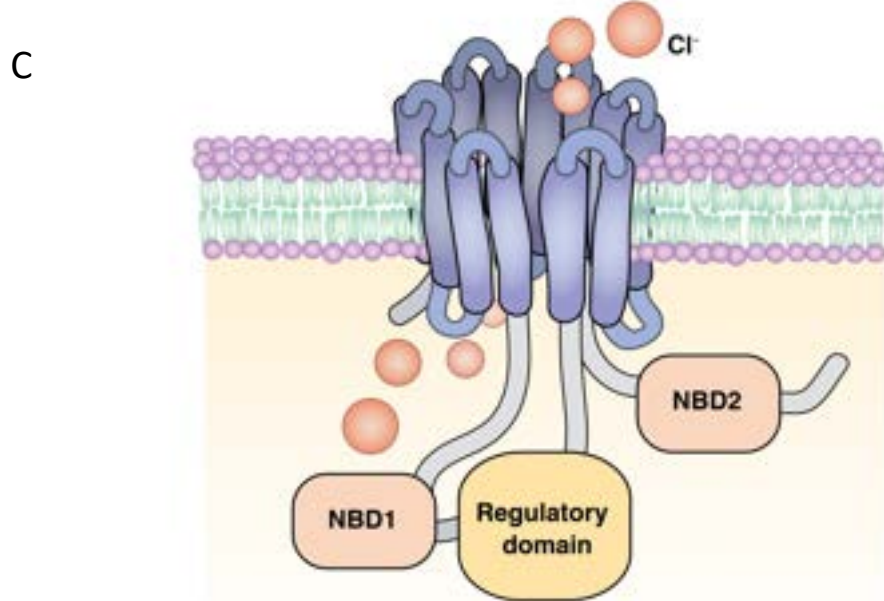
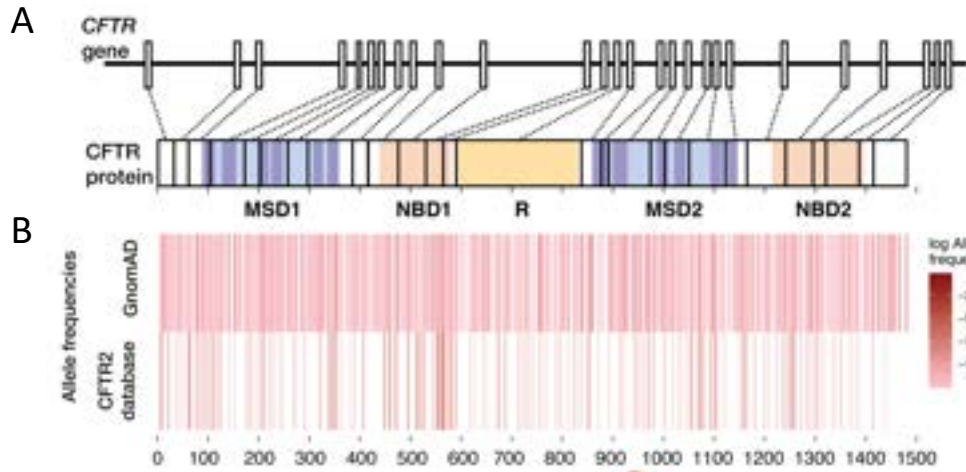
- Genetic variant → disease
- Include other biological info in model
- Consider variant effect



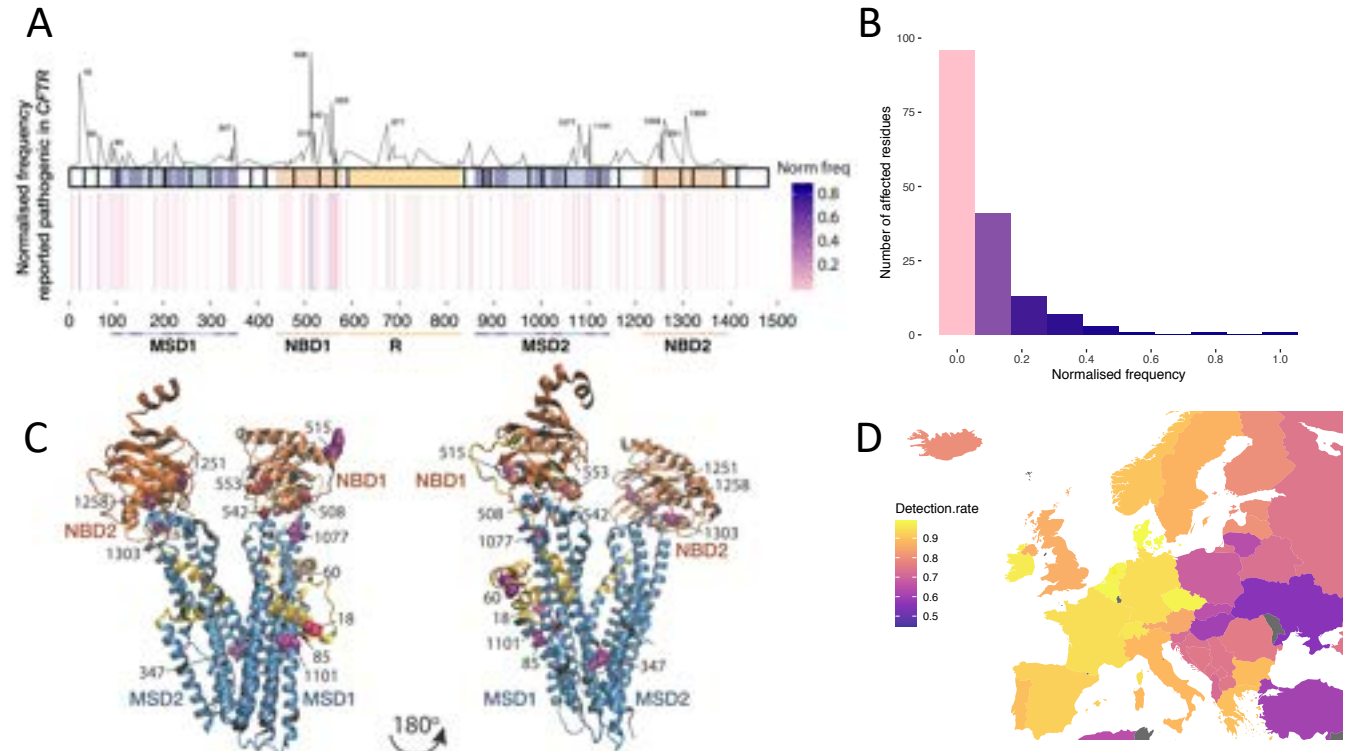
Genomic variants



Example: Cystic fibrosis (*CFTR*)

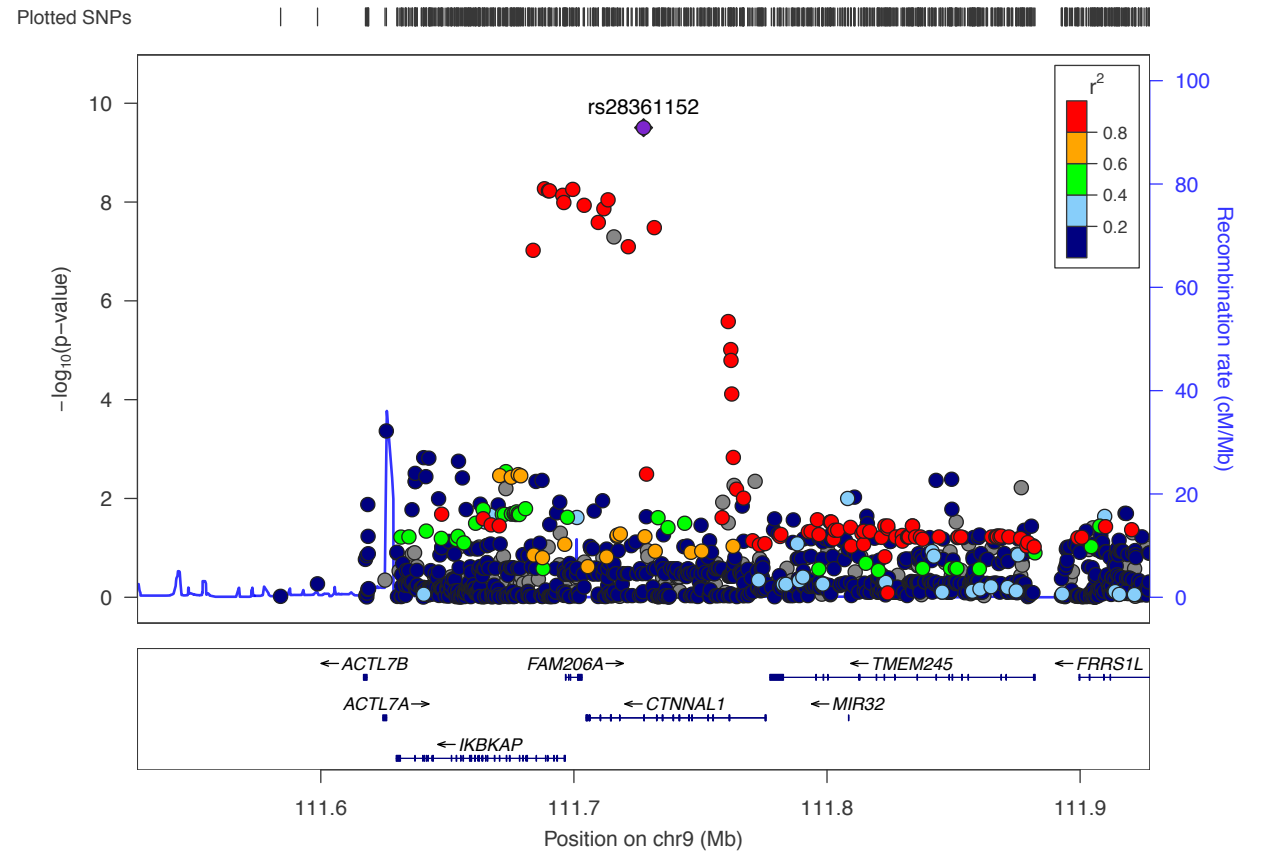
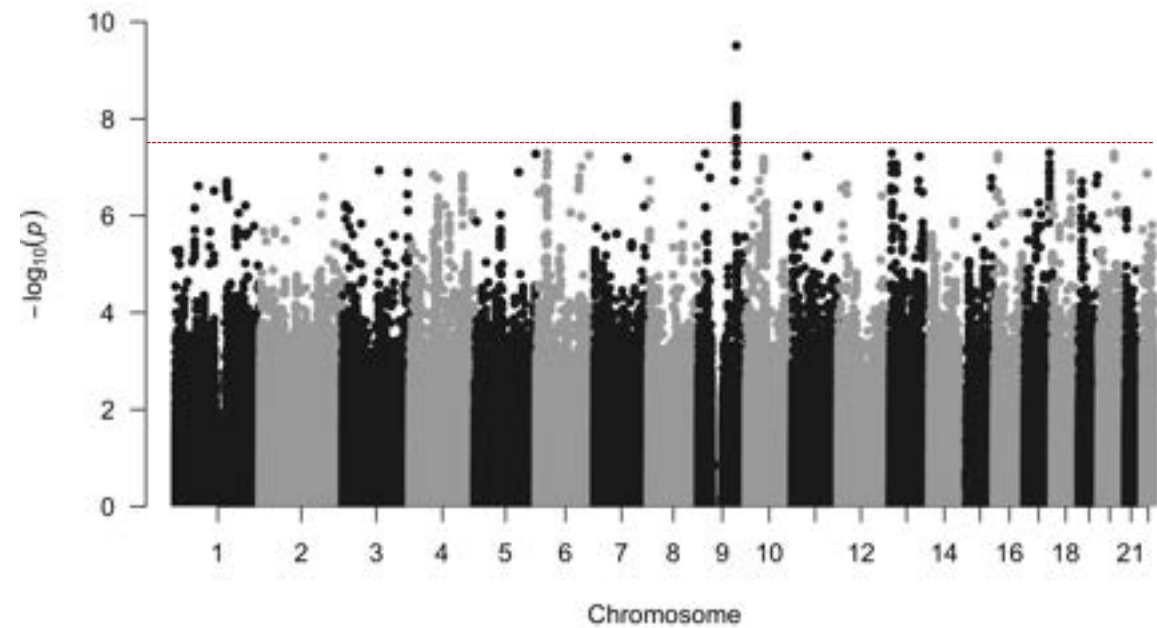


- Known mechanism quantified



Example: Sepsis susceptibility

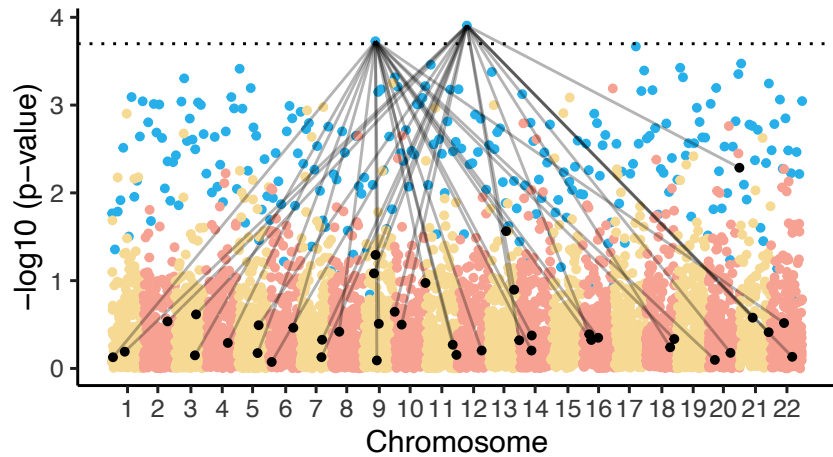
- Purely statistical
- Unknown mechanism



Pathway analysis

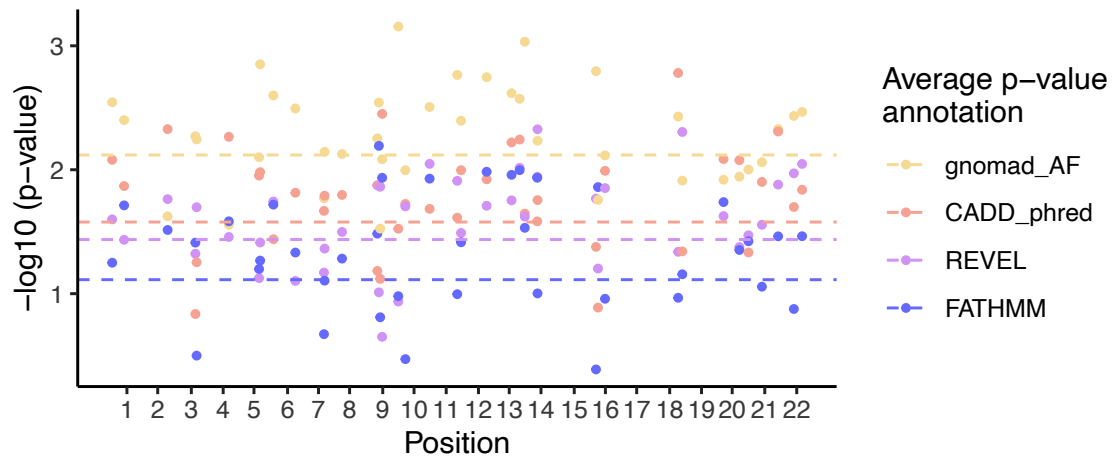
Archipelago Plot

Variant Set Association Test (blue)
with individual variant contributions (alternating yellow/orange)
and contributions for significant VSAT (black)



- Purely statistical
- Unknown mechanism
- Known biology

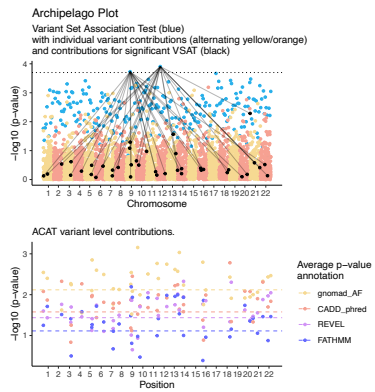
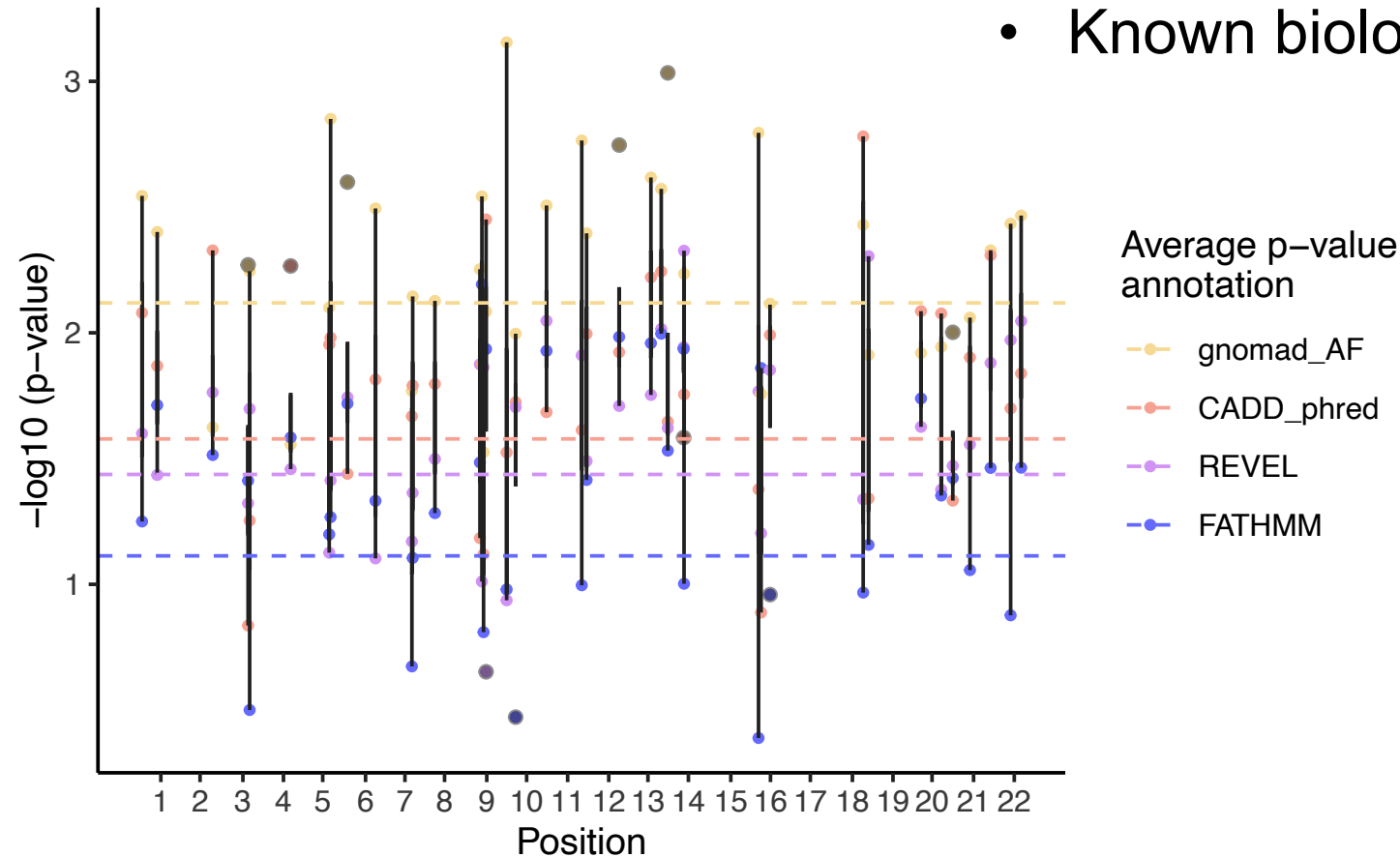
ACAT variant level contributions.



Pathway analysis

- Purely statistical
- Unknown mechanism
- Known biology

ACAT variant level contributions.



Multi variable combination

Aggregated Cauchy Association Test (ACAT)

$$\text{Burden} = \sum_j^p w_j U_j$$

Powerful if SNV effects on LDL in the same direction

$$\text{SKAT} = \sum_j^p w_j^2 U_j^2$$

Powerful if SNV effects on LDL in different directions

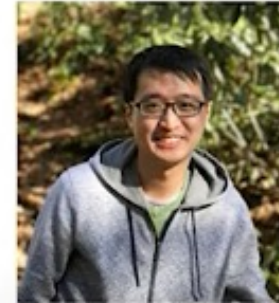
U_j = assoc score stat of SNV j
 p = # of SNPs in APOE enhancer

Powerful
correlated
p-value
combination
methods

Transform p-value to Cauchy

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

Weights



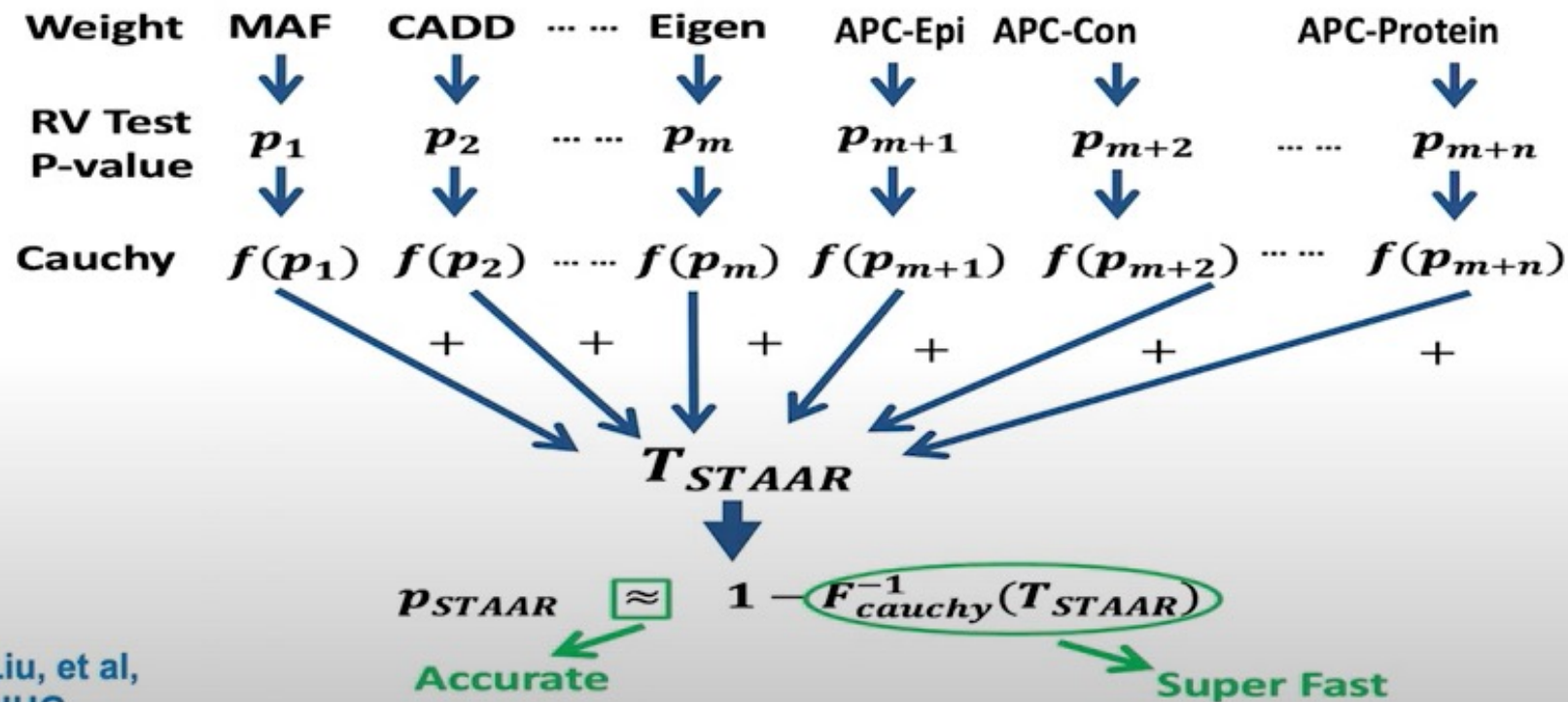
Yaowu Liu
(Liu, et al, AJHG, 2019)

ACAT can be used to (1) combine SNV-level p-values for a SNV-set test and (2) combine p-values of different SNV-set tests using different weights.

Variant-set in silico analysis

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

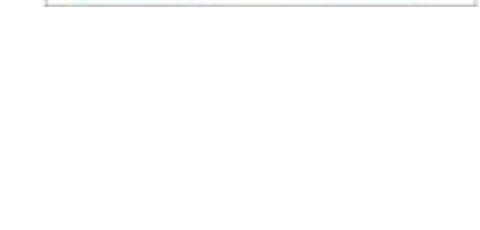
For a given SNV mask, e.g promoter of APOE



ACAT: Liu, et al,
2019, AJHG



--	--

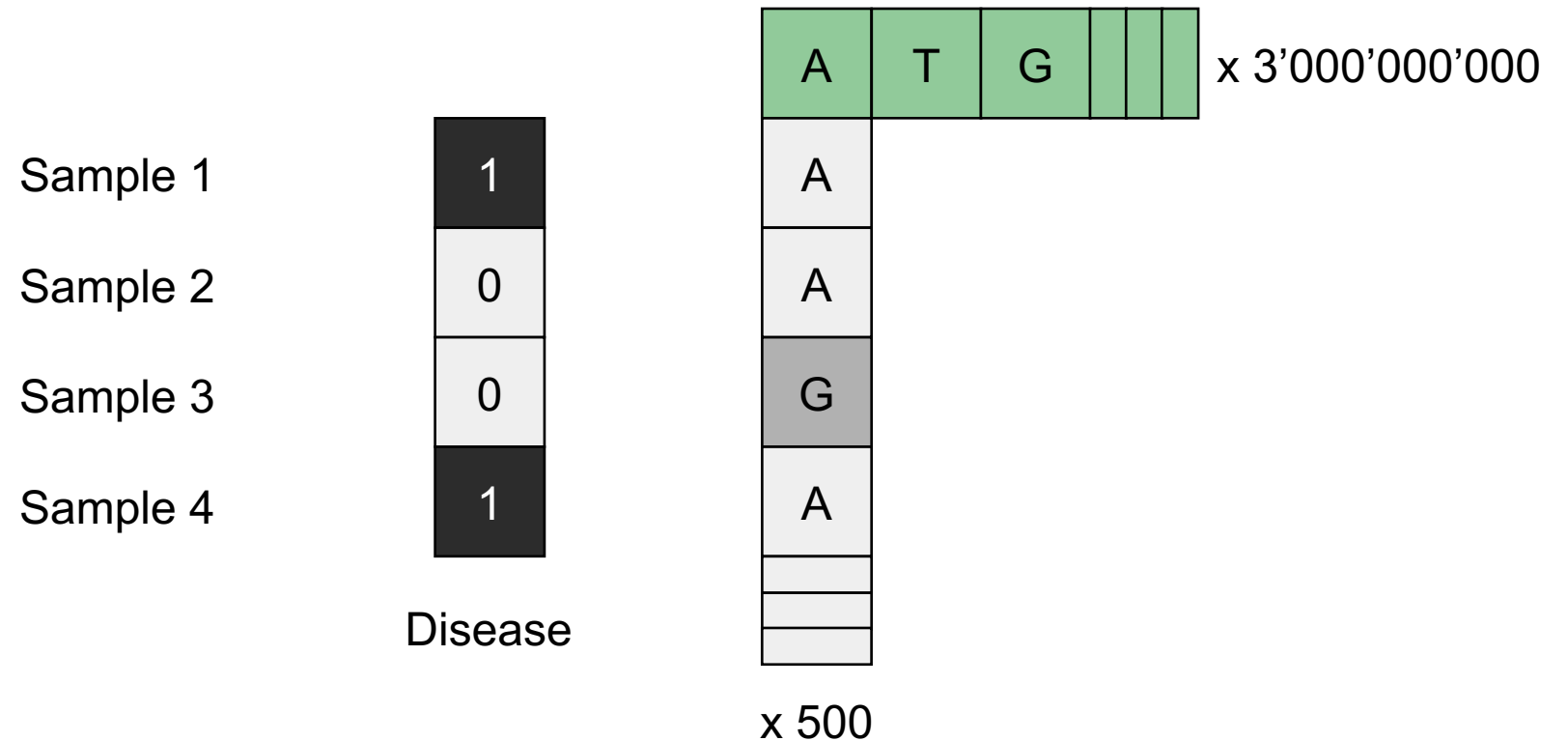


	Reference genome	A	T	G	C	T	A	C	
Sample 1	Variant 1	A							
Sample 2	Variant 1	A							
Sample 3	Variant 1	G							
Sample 4	Variant 1	A							

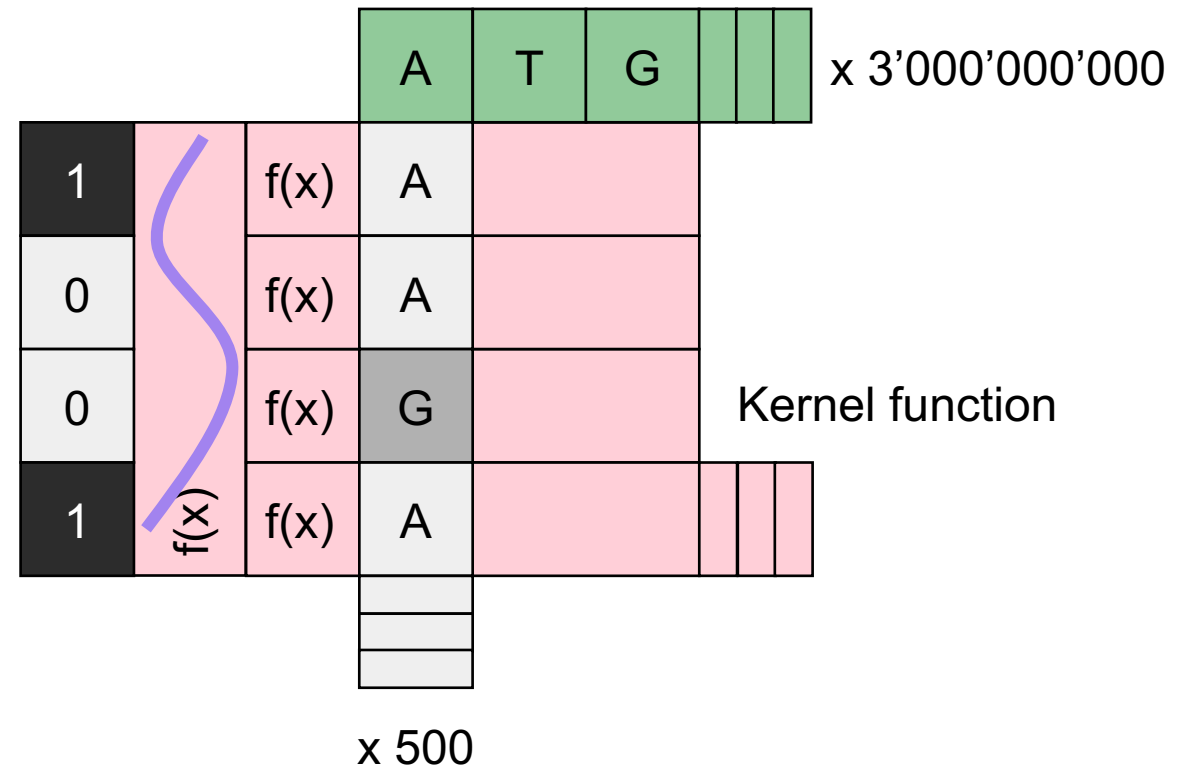
x 3'000'000'000

x 500

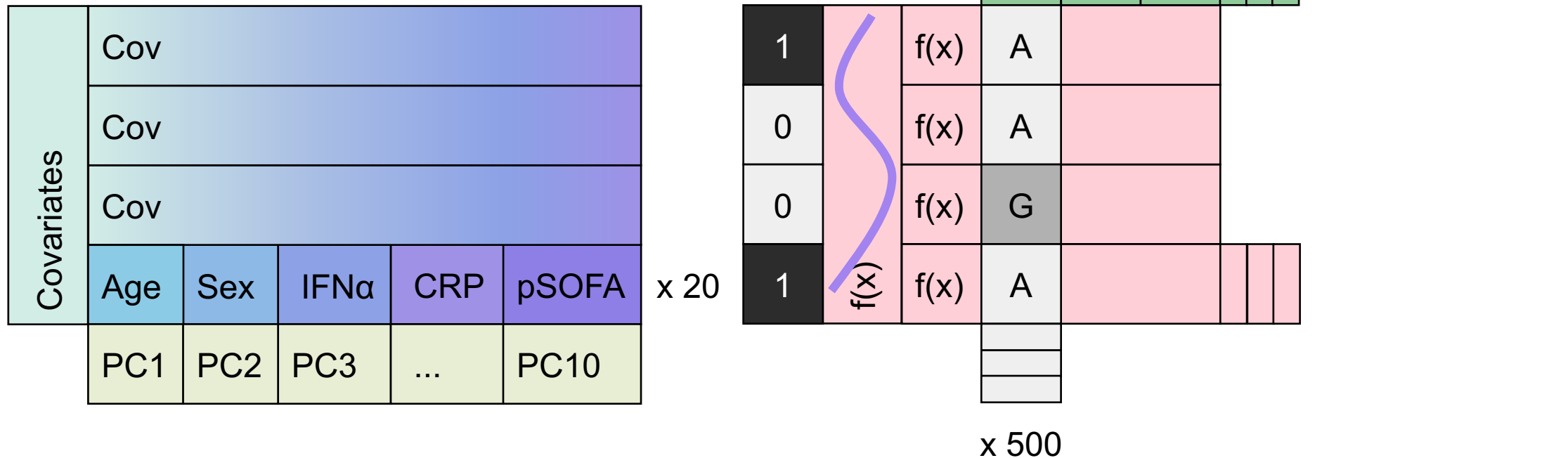
Model components



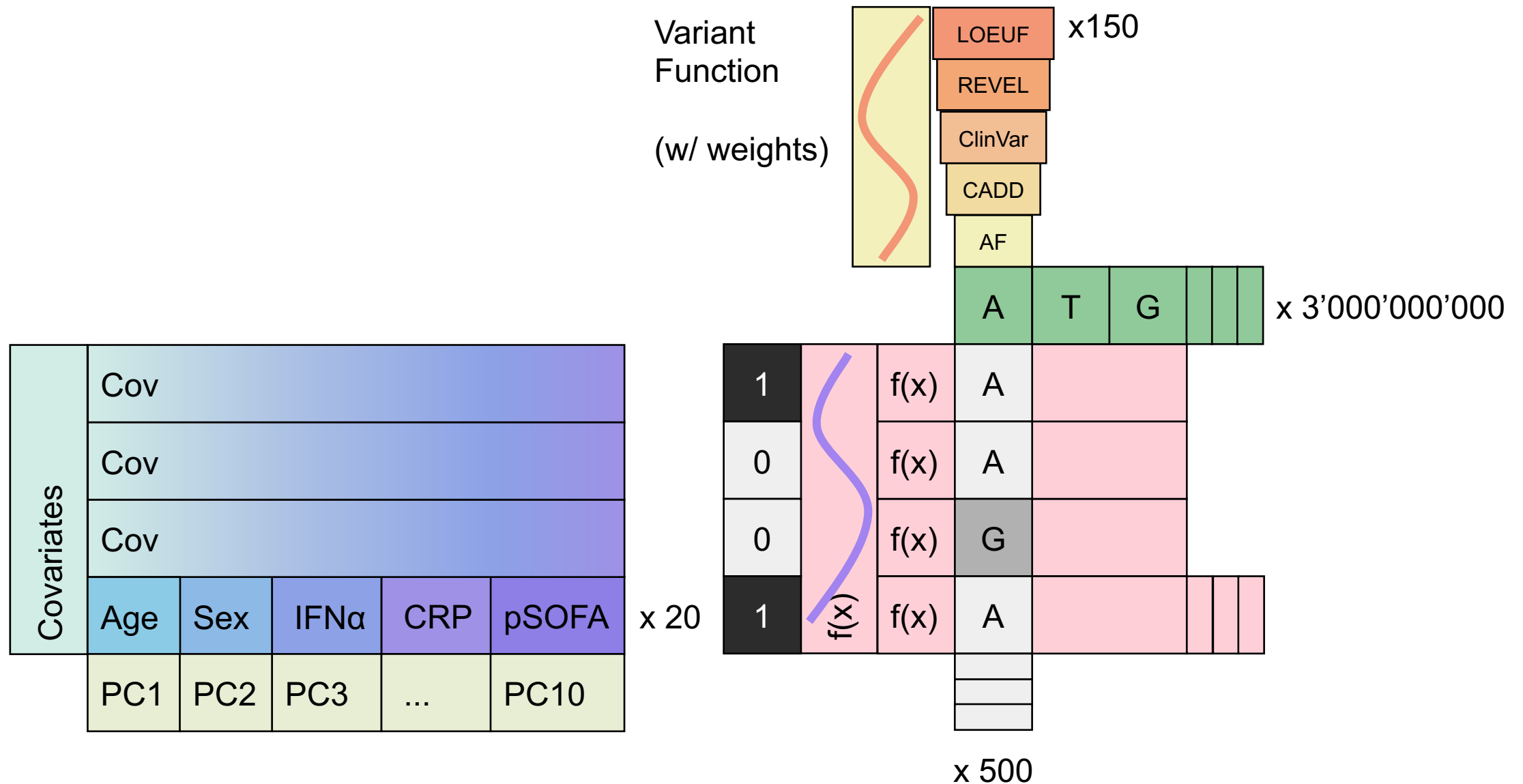
Model components



Covariates
(Clinical features, etc.)

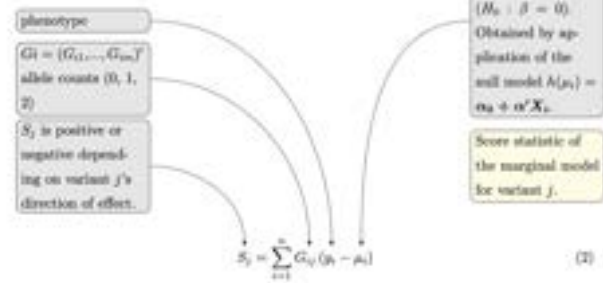


Model components

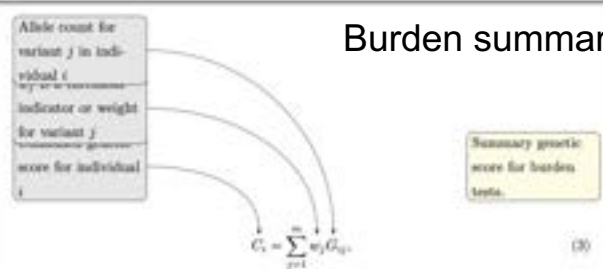


Model algorithm

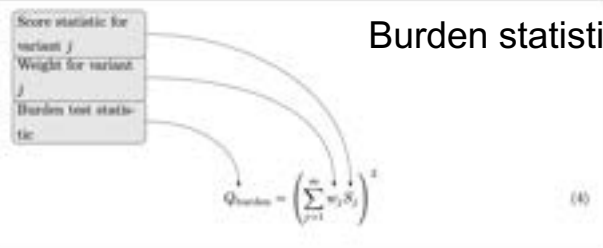
Score statistic



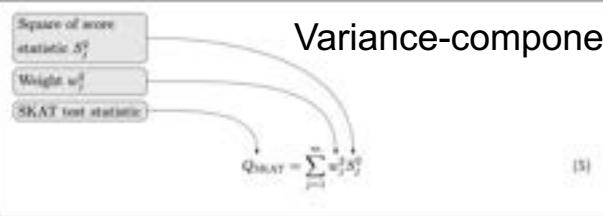
Burden summary



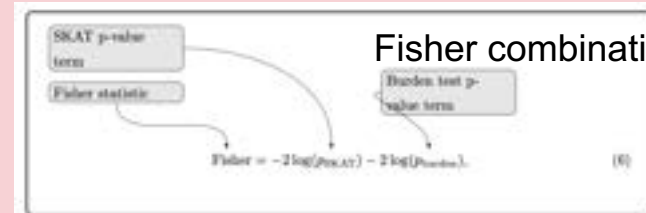
Burden statistic



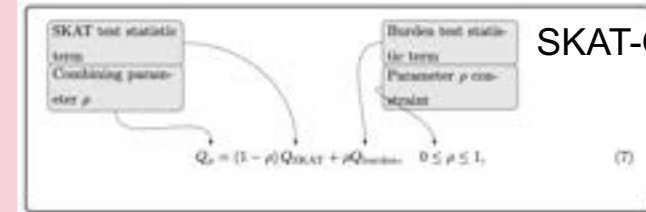
Variance-component



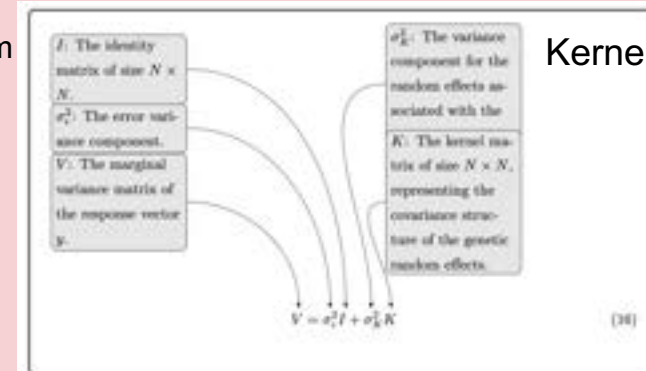
Fisher combination



SKAT-O



Kernel



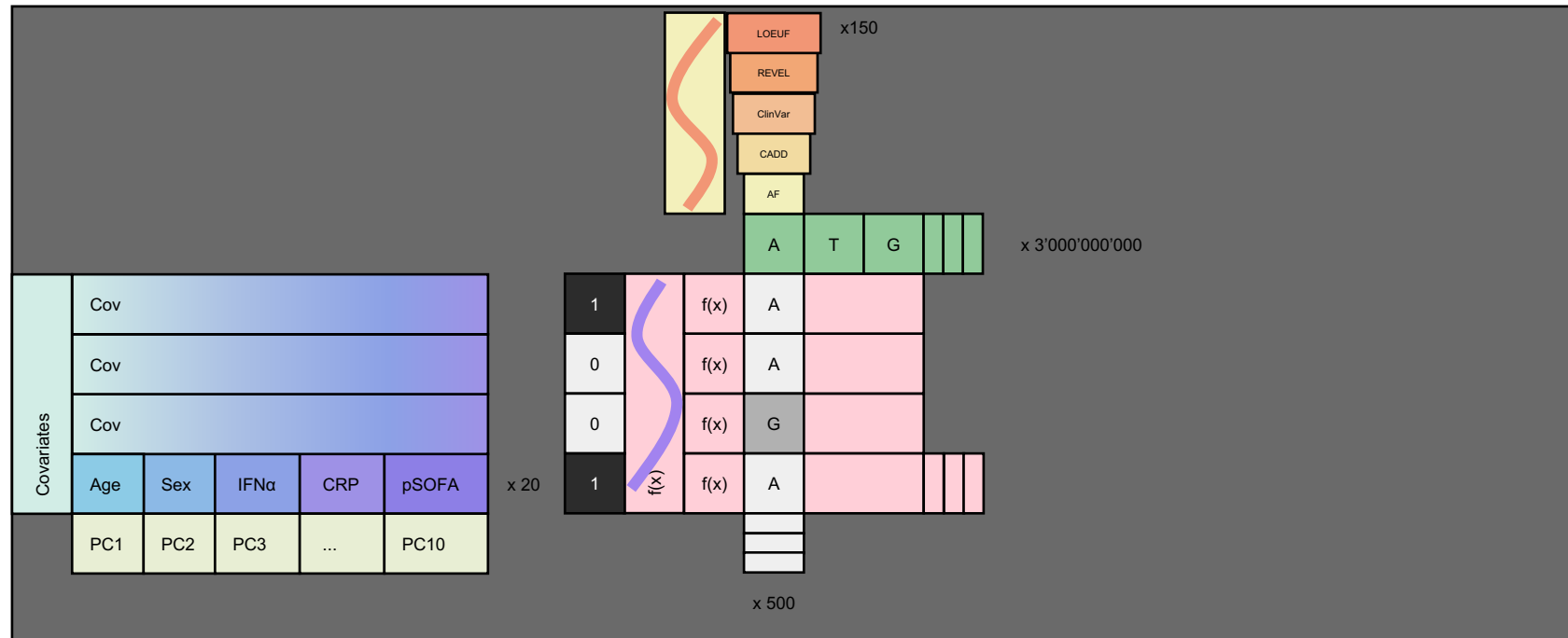
$$f \sim MVN(0, \sigma_K^2 K)$$

(9)

$$y_i = \alpha_0 + X_i' \alpha + f(G_i) + \epsilon_i$$

(10)

Model algorithm



Algorithm

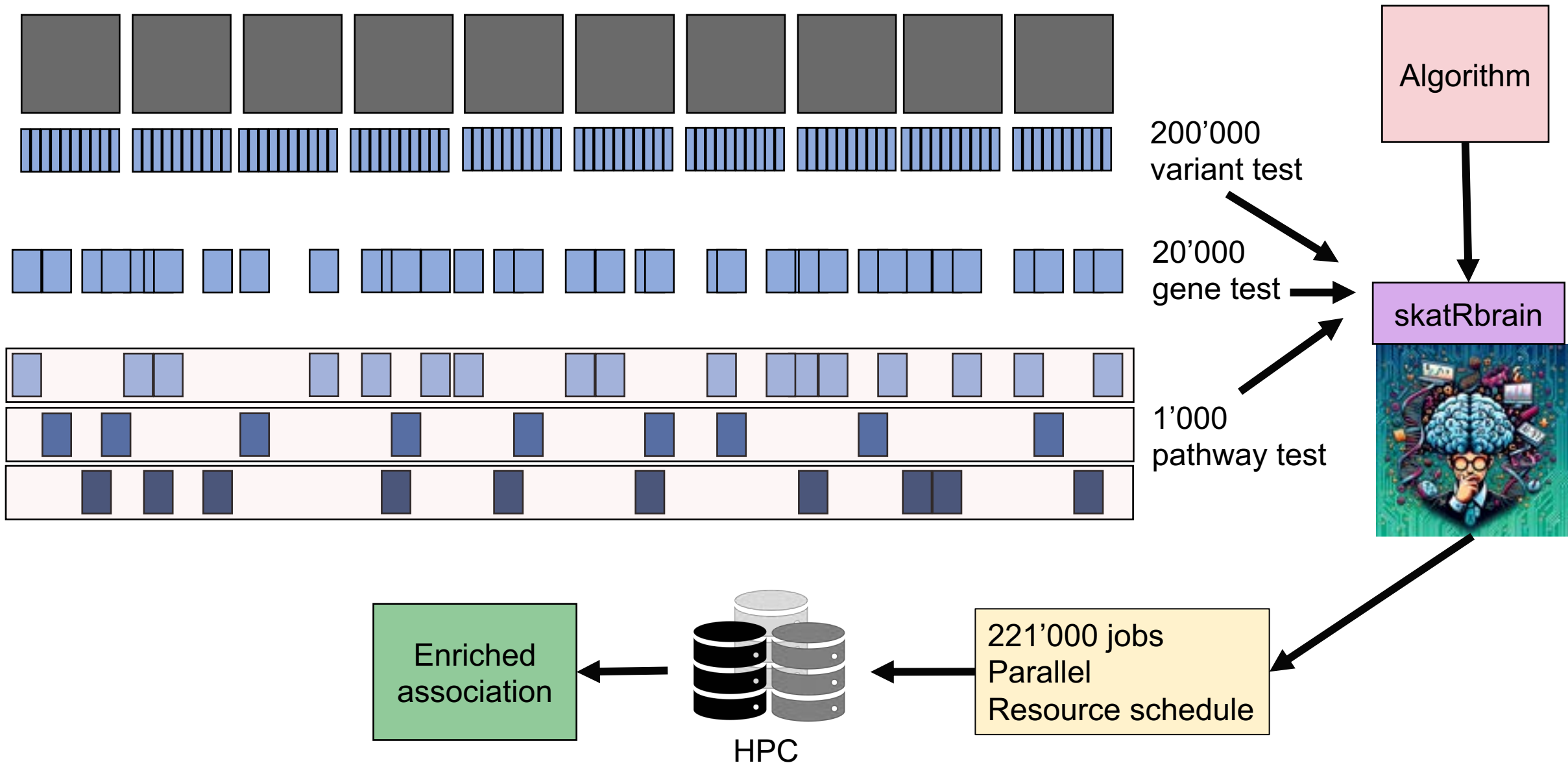
skatRbrain



Chunk data



Model application



Detour: Genetic association test

Subject	Phenotype (y_i)	Covariates (X_i)	Genotypes (G_i)
1	y_1	$X_{11}, X_{12}, \dots, X_{1m}$	$G_{11}, G_{12}, \dots, G_{1p}$
2	y_2	$X_{21}, X_{22}, \dots, X_{2m}$	$G_{21}, G_{22}, \dots, G_{2p}$
...	...	...	...
i	y_i	$X_{i1}, X_{i2}, \dots, X_{im}$	$G_{i1}, G_{i2}, \dots, G_{ip}$
...	...	...	...
n	y_n	$X_{n1}, X_{n2}, \dots, X_{nm}$	$G_{n1}, G_{n2}, \dots, G_{np}$

Subject	Phenotype (y_i)	Covariates (X_i)	Genotypes (G_i)
1	High BP	Male, 45, PC1	0, 1, 2, ..., 2
2	Normal BP	Female, 30, PC1	2, 0, 1, ..., 0
3	High BP	Male, 55, PC1	1, 1, 0, ..., 1
4	Normal BP	Female, 40, PC1	0, 2, 2, ..., 1

$$y_i = \alpha_0 + \alpha' X_i + \beta' G_i + \epsilon_i$$

Detour: Genetic association test

Subject	Phenotype (y_i)	Covariates (X_i)	Genotypes (G_i)
1	y_1	$X_{11}, X_{12}, \dots, X_{1m}$	$G_{11}, G_{12}, \dots, G_{1p}$
2	y_2	$X_{21}, X_{22}, \dots, X_{2m}$	$G_{21}, G_{22}, \dots, G_{2p}$
...	...	...	...
i	y_i	$X_{i1}, X_{i2}, \dots, X_{im}$	$G_{i1}, G_{i2}, \dots, G_{ip}$
...	...	...	...
n	y_n	$X_{n1}, X_{n2}, \dots, X_{nm}$	$G_{n1}, G_{n2}, \dots, G_{np}$

Subject	Phenotype (y_i)	Covariates (X_i)	Genotypes (G_i)
1	High BP	Male, 45, PC1	0, 1, 2, ..., 2
2	Normal BP	Female, 30, PC1	2, 0, 1, ..., 0
3	High BP	Male, 55, PC1	1, 1, 0, ..., 1
4	Normal BP	Female, 40, PC1	0, 2, 2, ..., 1

$$y_i = \alpha_0 + \alpha' X_i + \beta' G_i + \epsilon_i$$

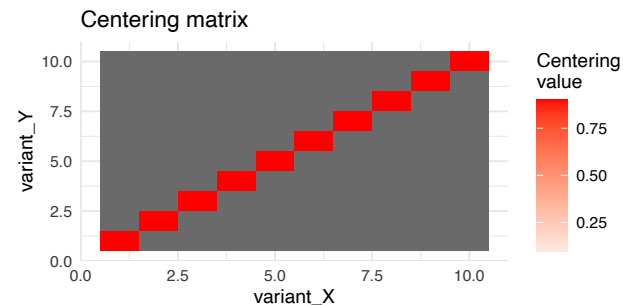
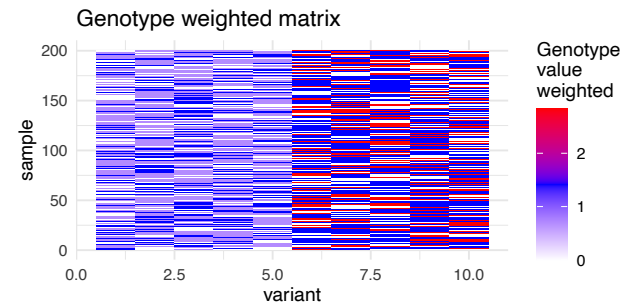
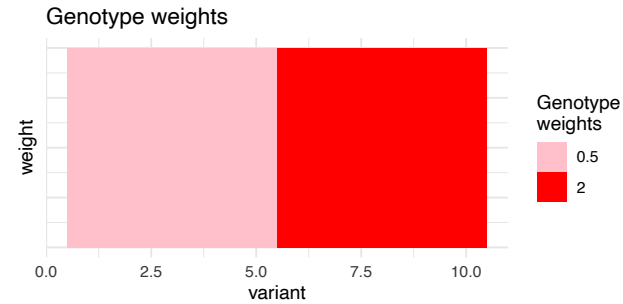
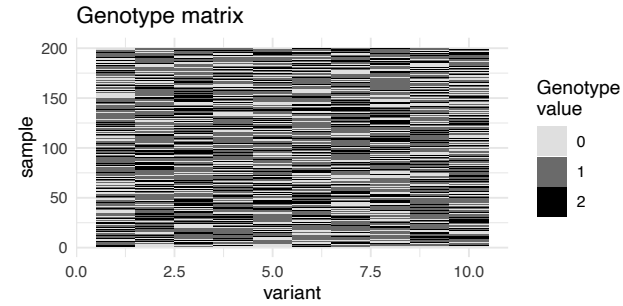
G Matrix			
0	1	2	...
2	0	1	...
1	1	0	...
0	2	2	...

W Matrix			
1	0	0	...
0	0.5	0	...
0	0	2	...
0	0	0	...

Detour: Genetic association test

G Matrix			
0	1	2	...
2	0	1	...
1	1	0	...
0	2	2	...

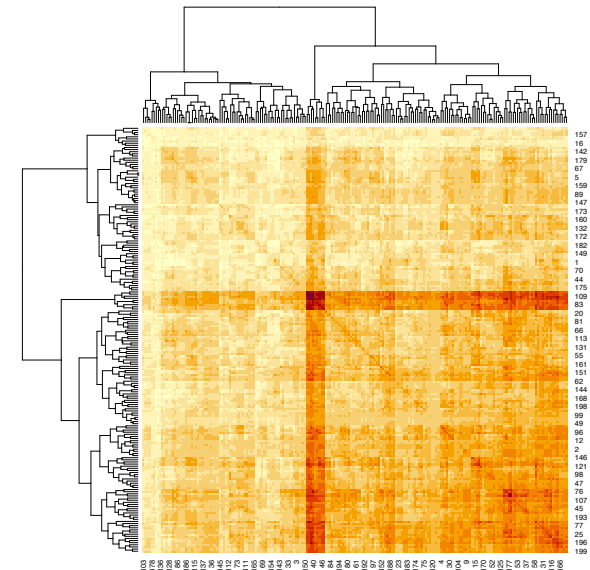
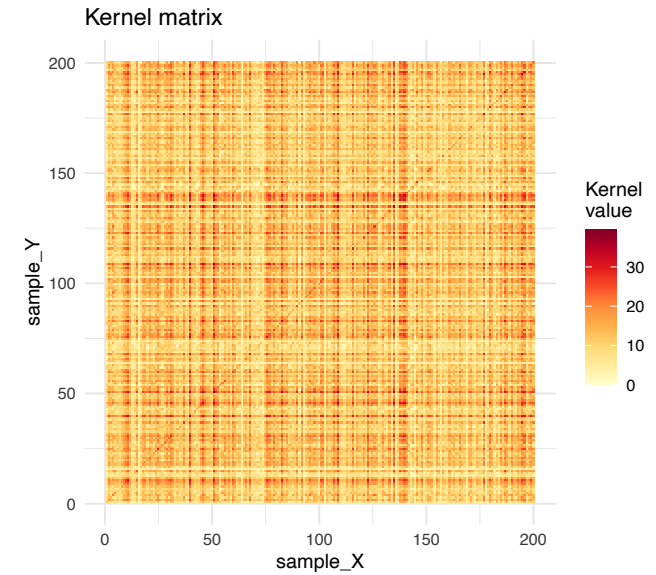
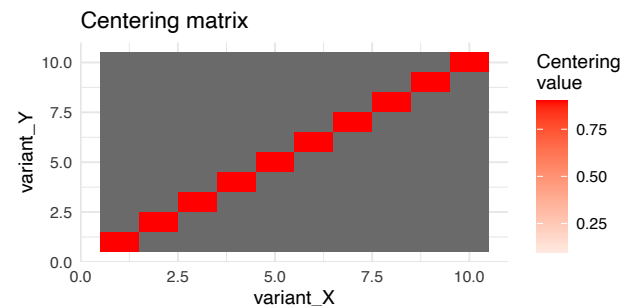
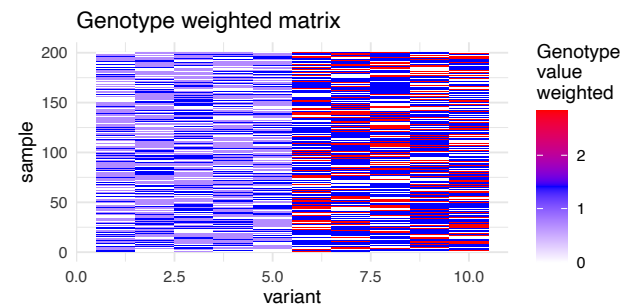
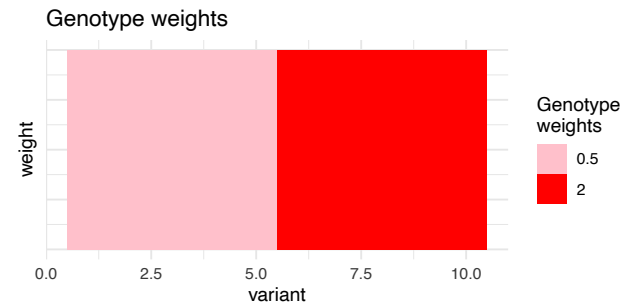
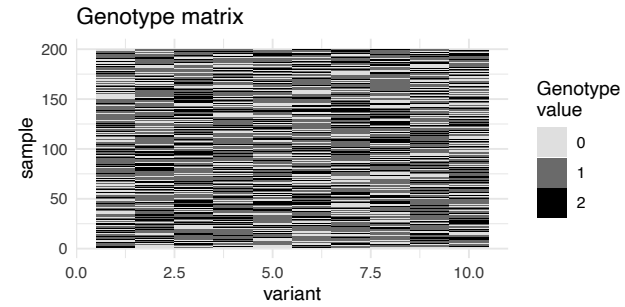
W Matrix			
1	0	0	...
0	0.5	0	...
0	0	2	...
0	0	0	...



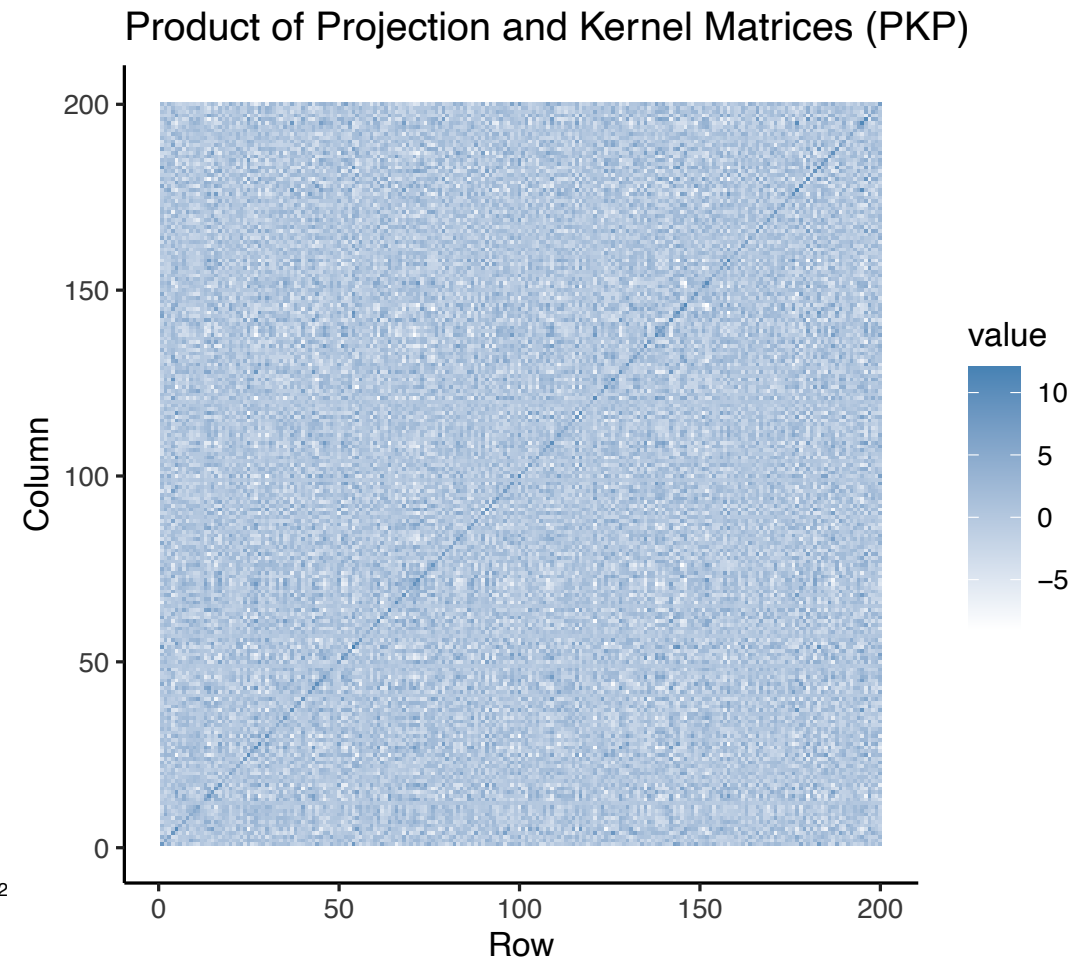
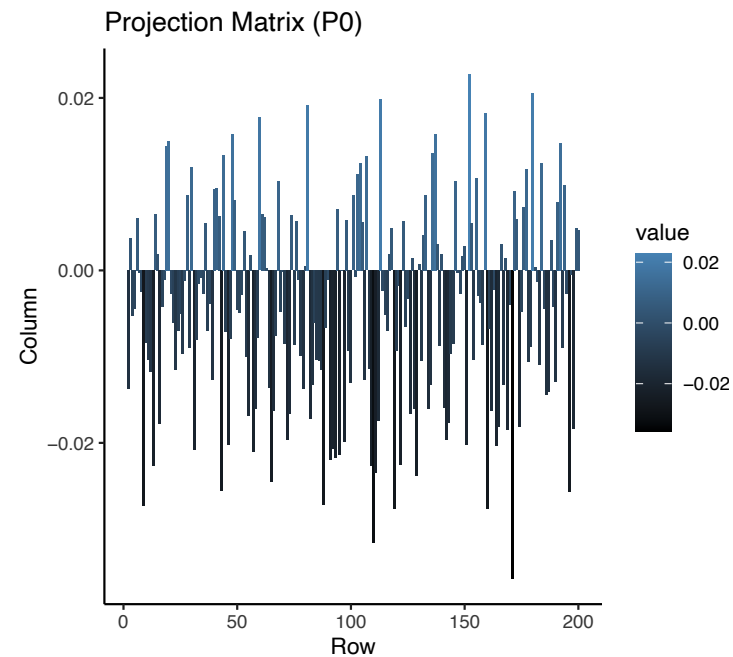
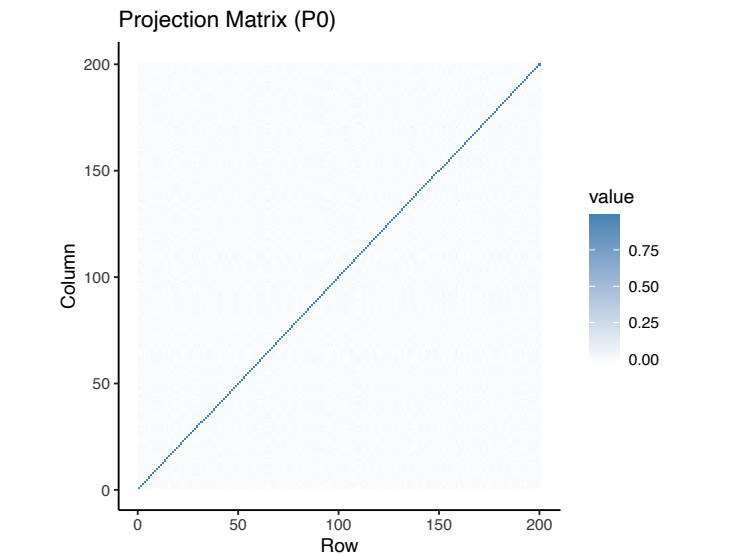
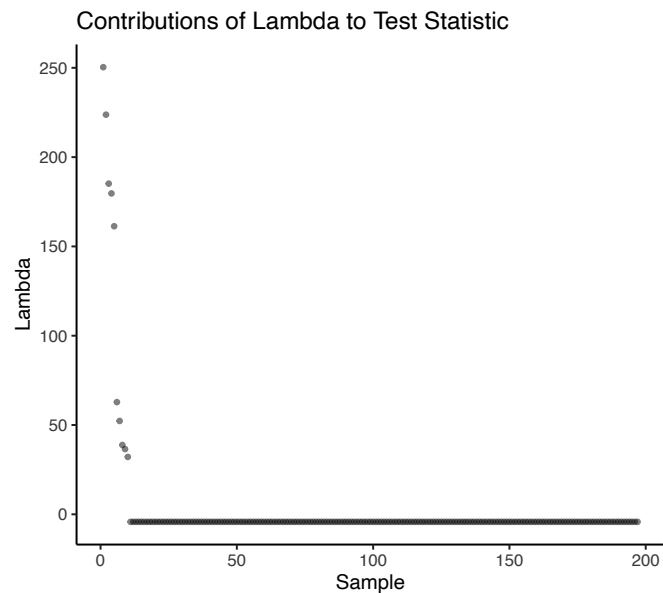
Detour: Genetic association test

G Matrix			
0	1	2	...
2	0	1	...
1	1	0	...
0	2	2	...

W Matrix			
1	0	0	...
0	0.5	0	...
0	0	2	...
0	0	0	...



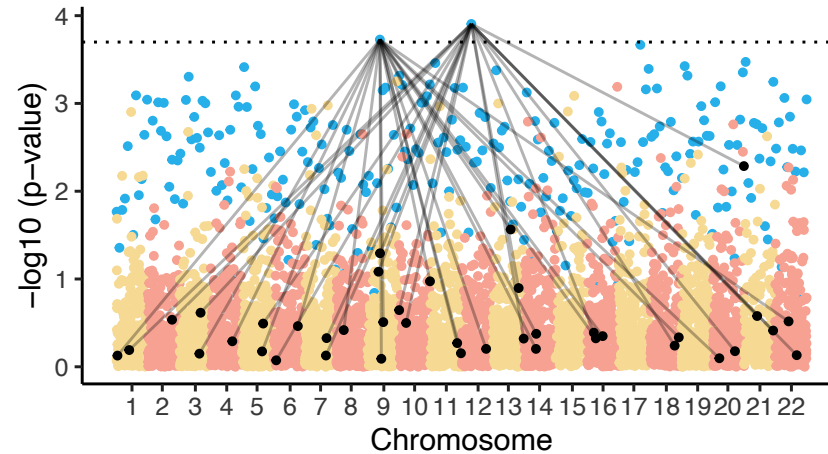
Detour: Genetic association test



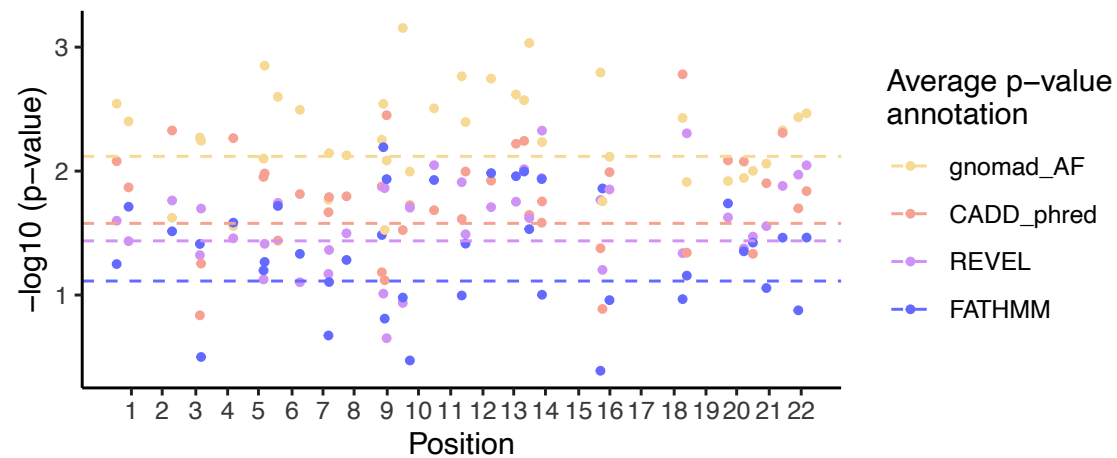
Detour: Genetic association test

Archipelago Plot

Variant Set Association Test (blue)
with individual variant contributions (alternating yellow/orange)
and contributions for significant VSAT (black)

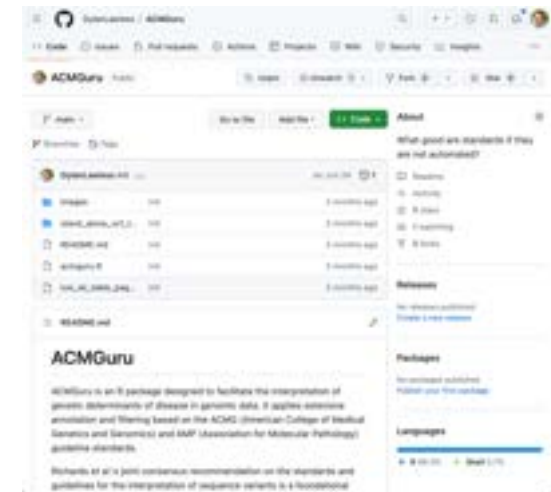


ACAT variant level contributions.

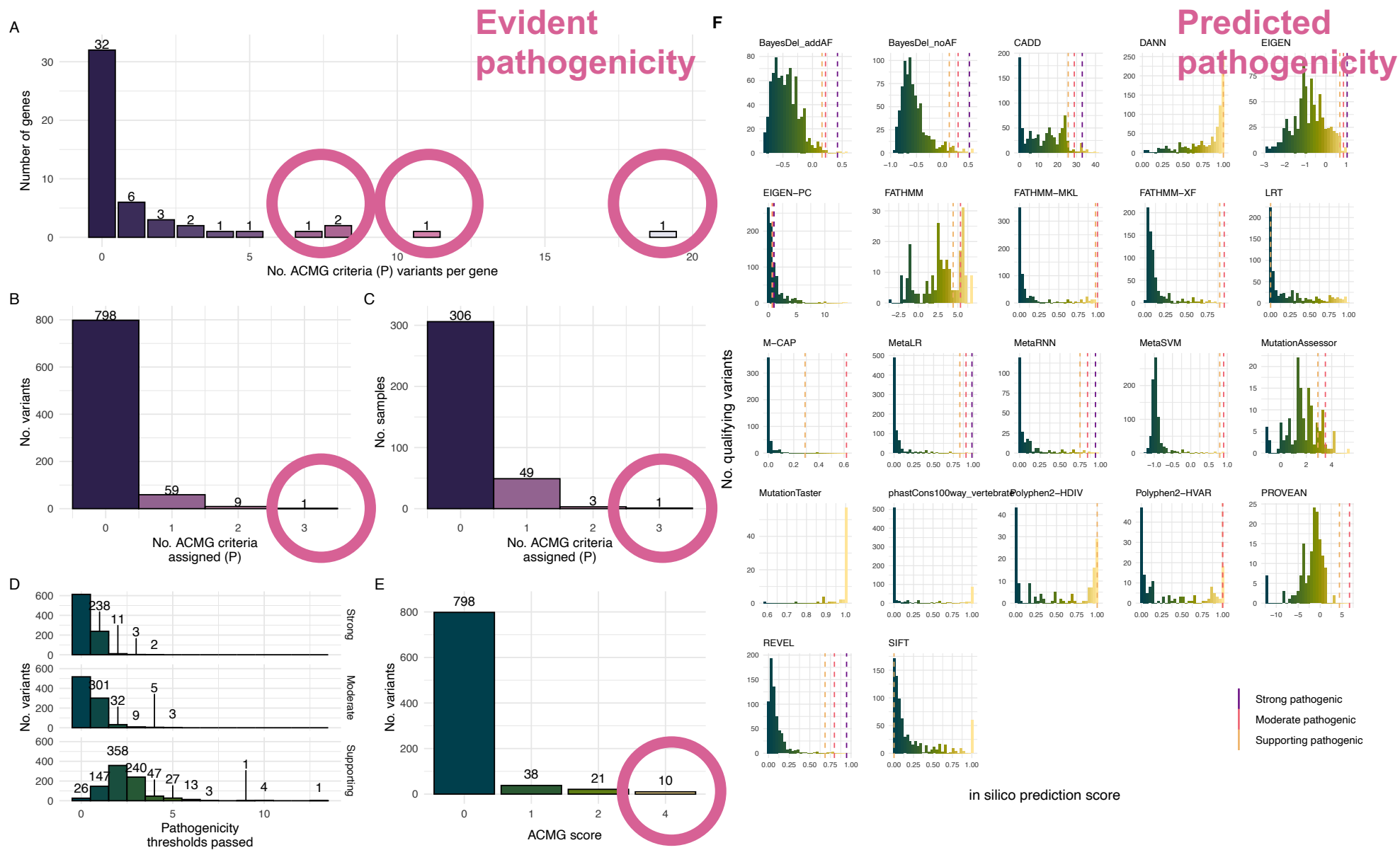


Detour #2: Variant interpretation

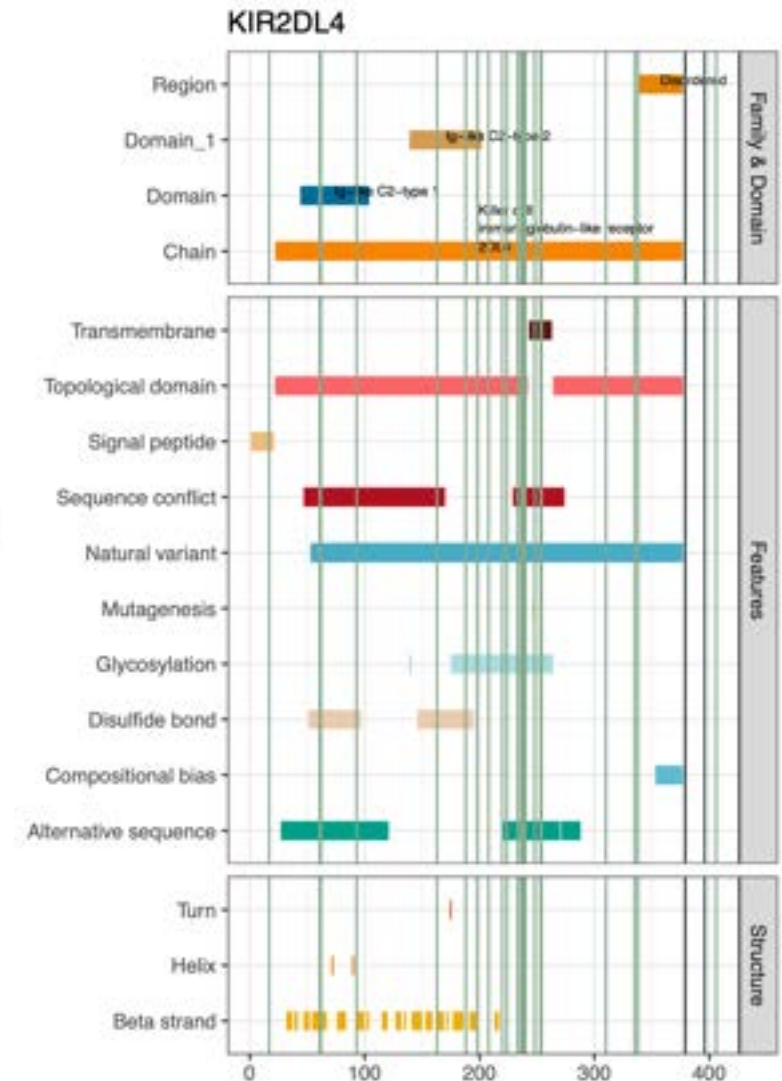
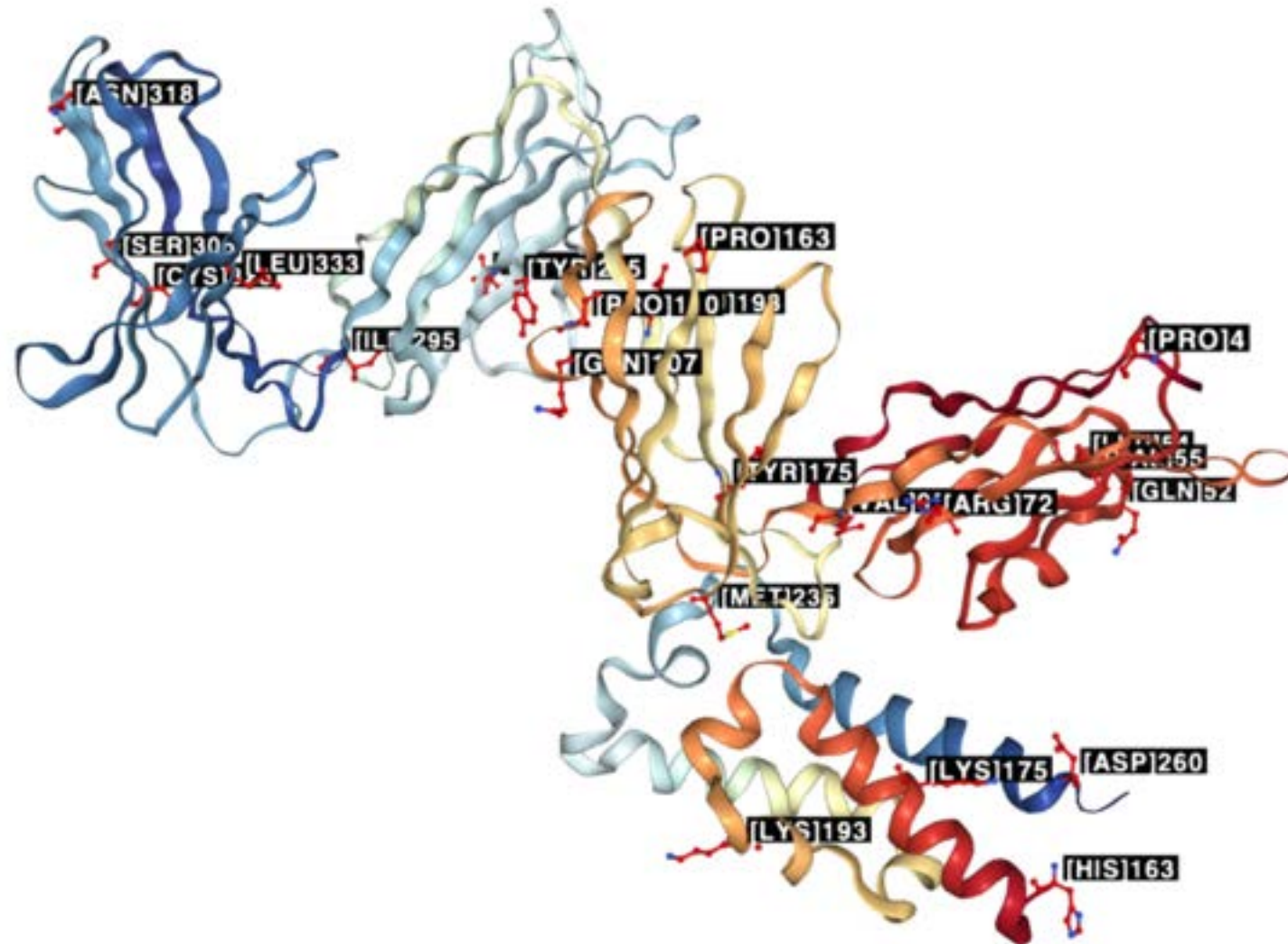
- Joint consensus recommendation on the standards and guidelines for the interpretation of sequence variants (Richards 2015).
- Standards and guidelines for the interpretation and reporting of sequence variants in cancer (Li 2017).
- Technical standards for the interpretation and reporting of constitutional copy-number variants (Riggs 2020).
- Effective variant filtering and expected candidate variant yield in studies of rare human disease (Pedersen 2021).
- ...



Detour #2: Variant interpretation



Detour #2: Variant interpretation



Detour #2: Variant interpretation

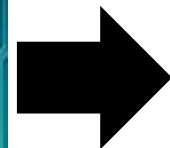
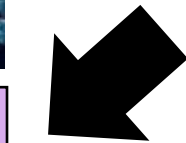
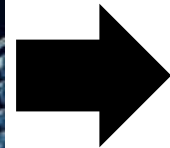
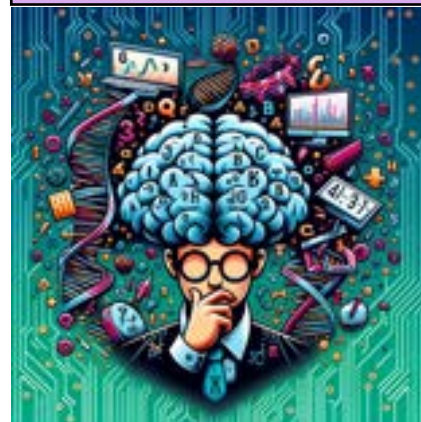
Variant annotation															
Variant	Set	Sample	Genotype	Age	Sex	Cohort AF	CADD phred	REVEL	ClinVar	FATHMM	OMIM	PANTHER	Gene ontology	...	GnomAD AF
1	A	S1	0	18	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	...	0	21	1	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	Sn	1	45	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
2	A	S1	1	18	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	...	0	21	1	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	Sn	0	45	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
3	B	S1	1	18	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	...	0	21	1	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	Sn	0	45	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
4	B	S1	0	18	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	...	0	21	1	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	Sn	0	45	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
↓ Set level		↓ Sample level		↓ Variant level		↓ Gene level		↓ Ontology level							

Detour #2: Combined

ACMGuru



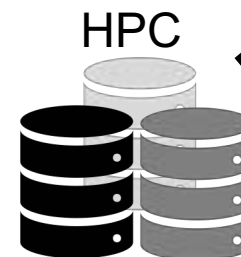
skatRbrain



Variant annotation															
Variant	Set	Sample	Genotype	Age	Sex	Cohort AF	CADD phred	REVEL	ClinVar	FATHMM	OMIM	PANTHER	Gene ontology	...	GnomAD AF
2	A	S1	1	18	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
3	B	S1	1	18	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	...	0	21	1	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	Sn	0	45	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
4	B	Sn	0	45	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05

SKAT-O
Sequence kernel
association test - optimised

2011 Wu AJHG
2012 Lee BioStat
2012 Lee & Esmond AJHG
2013 Chen GenEpi
2013 Ionita-Laza AJHG



Test: Single variant
200'000

Test: Single genes
20'000

Test: protein pathways
1'000

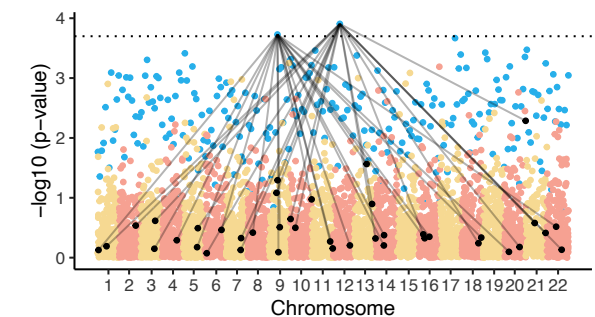
Result 1

Result 2

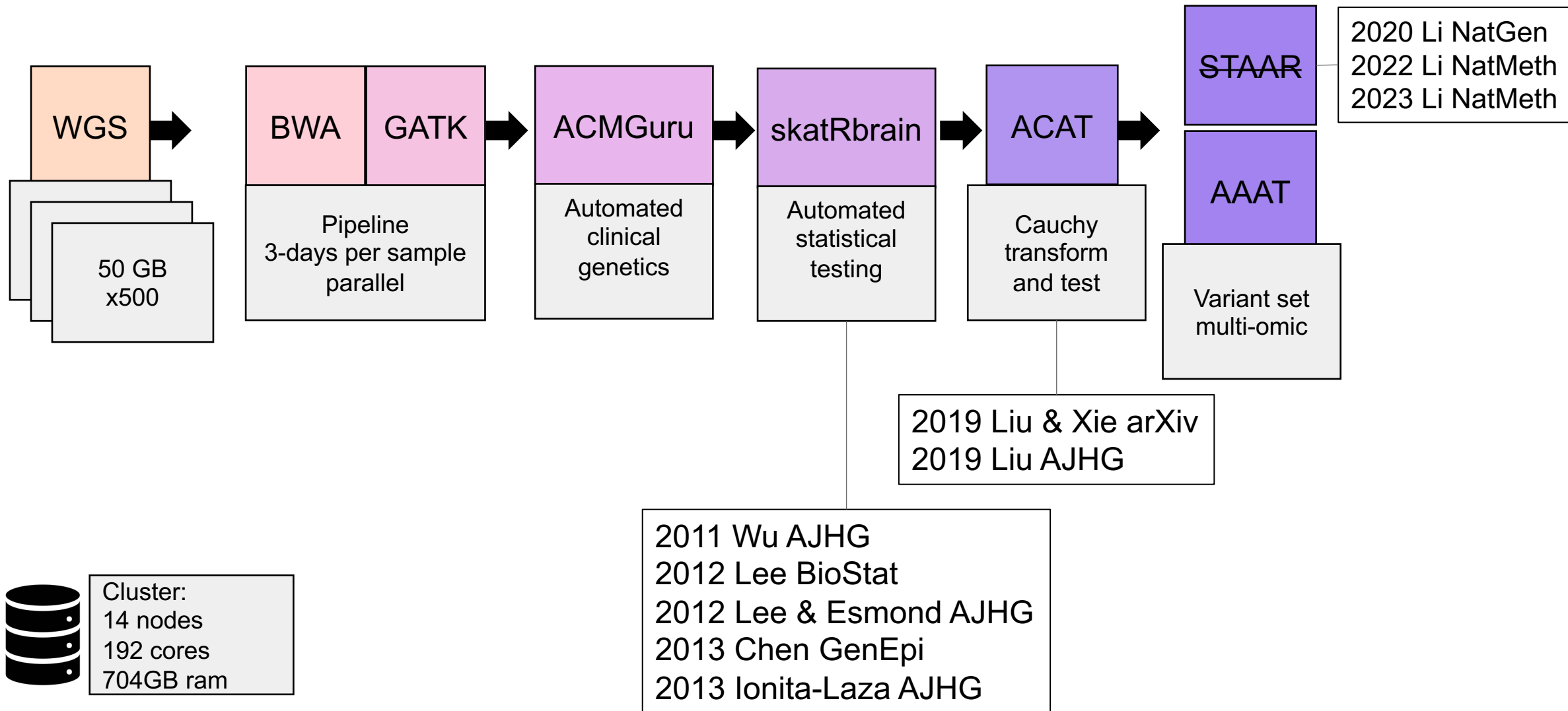
Result 3

Result ...

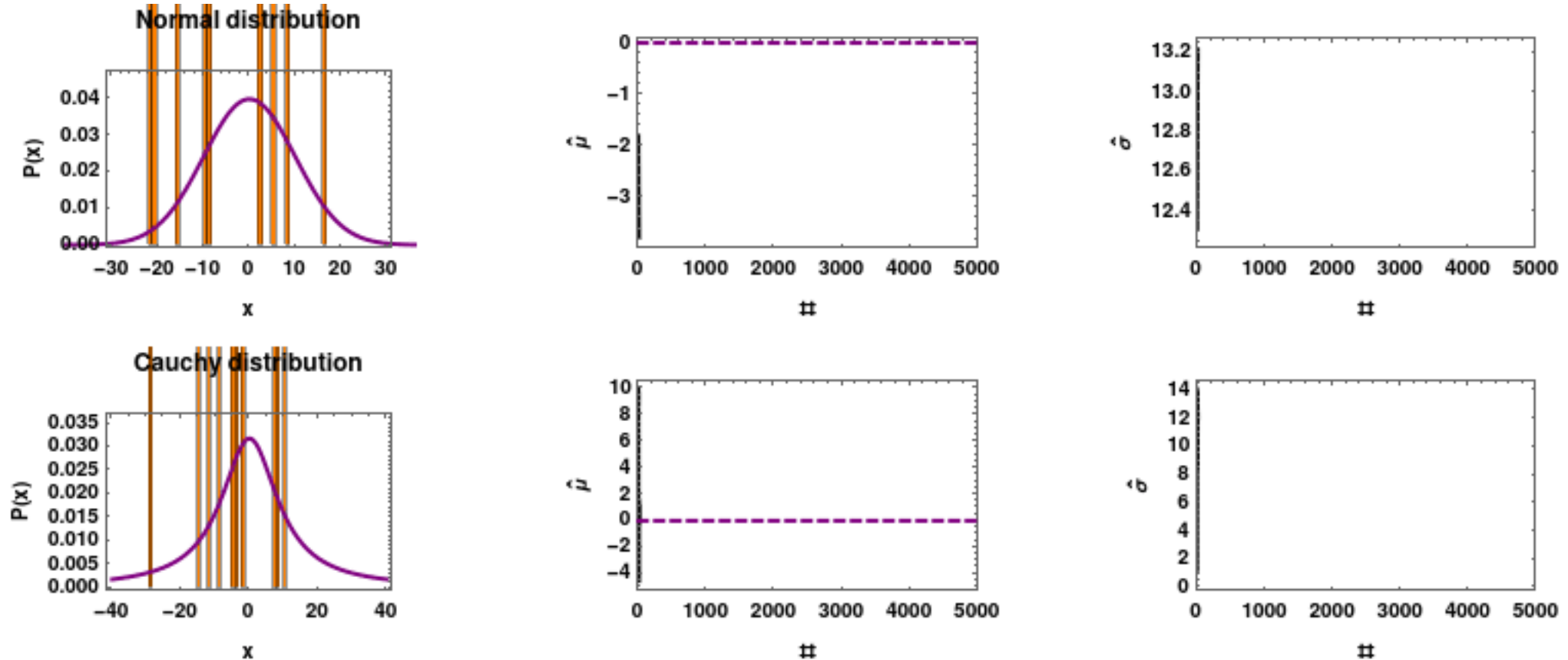
Result 219'913



Analysis chain



Cauchy distribution in variant analysis



Cauchy distribution, the sample mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) fluctuate widely even as the sample size increases, illustrating the lack of consistency in mean estimators for distributions without a defined mean or variance.

ACAT main steps

Multiple
annotation-aware
tests per variant

Variant annotation															
Variant	Set	Sample	Genotype	Age	Sex	Cohort AF	CADD phred	REVEL	ClinVar	FATHMM	OMIM	PANTHER	Gene ontology	...	GnomAD AF
1	A	S1	0	18	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	...	0	21	1	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
1	A	Sn	1	45	0	.3	25.90	.902	PL	.7	179615	11539	0002331	...	3E-05
2	A	S1	1	18	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	...	0	21	1	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
2	A	Sn	0	45	0	.01	29.3	.783	P	.9	179615	11539	0002331	...	7E-06
3	B	S1	1	18	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	...	0	21	1	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
3	B	Sn	0	45	0	.5	25.9	.888	PL	NA	147660	10960	0002331	...	3E-05
4	B	S1	0	18	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	...	0	21	1	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
4	B	Sn	0	45	0	.02	12.1	NA	NA	NA	147660	10960	0002331	...	3E-05
↓ Set level		↓ Sample level			↓ Variant level					↓ Gene level		↓ Ontology level			

ACAT main steps

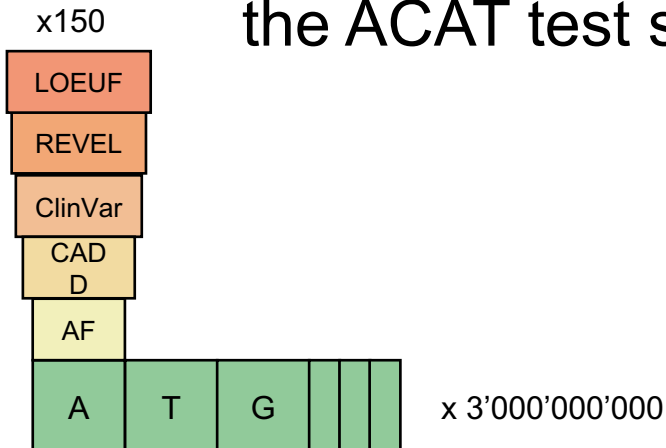
- ACAT test statistic:
$$T_{ACAT} = \sum_{i=1}^k w_i \tan [(0.5 - p_i)\pi]$$

- p_i are the p-values
- w_i are non-negative weights.

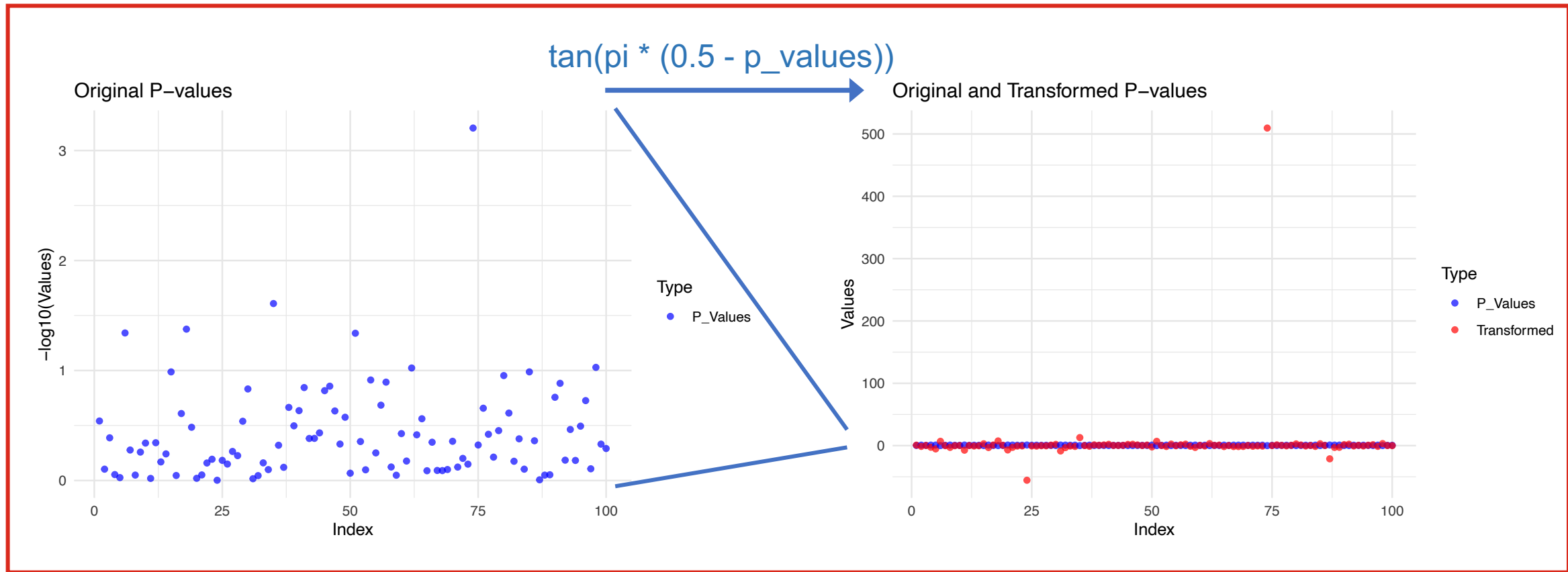
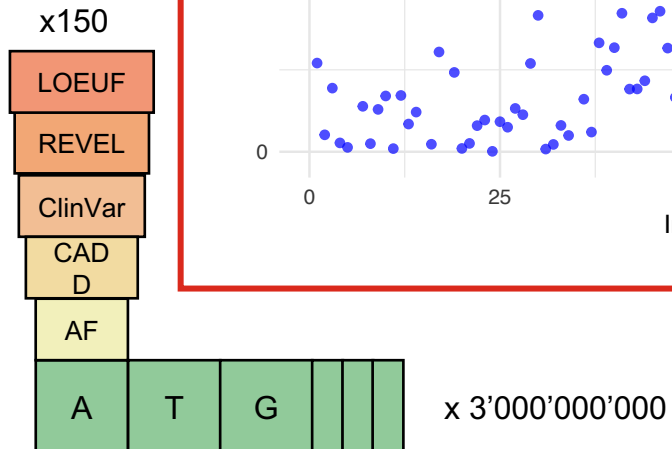
- P-value calculation for the ACAT test statistic:

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

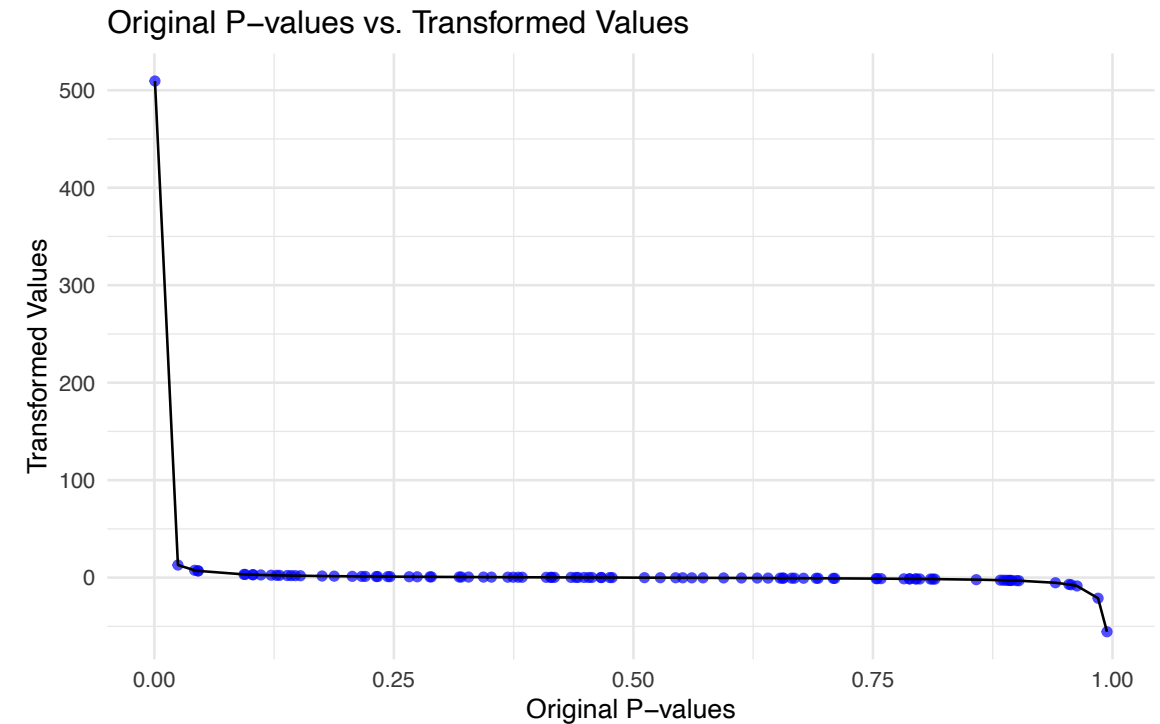
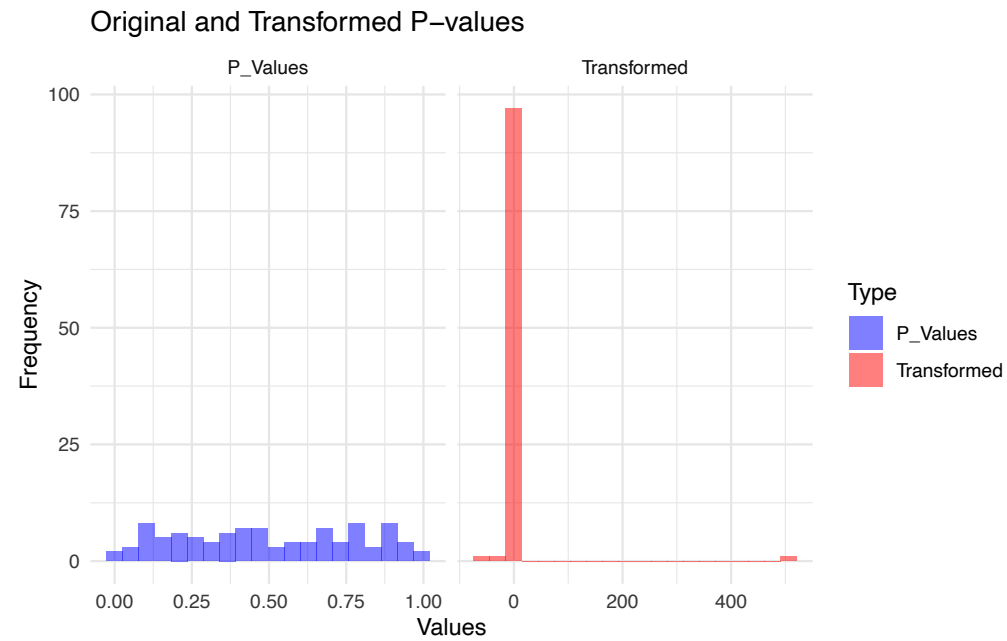
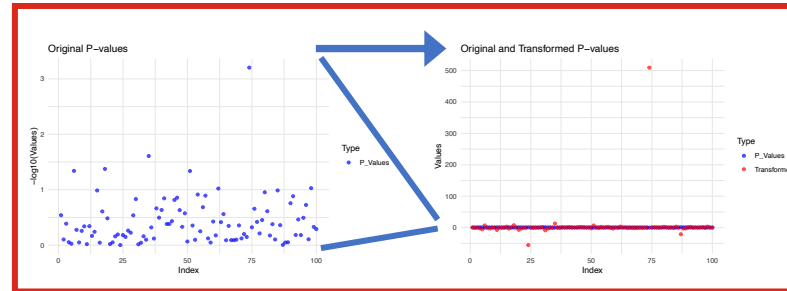
$$w = \sum_{i=1}^k w_i.$$



Pathway analysis



Pathway analysis



x150

LOEUF

REVEL

ClinVar

CAD

D

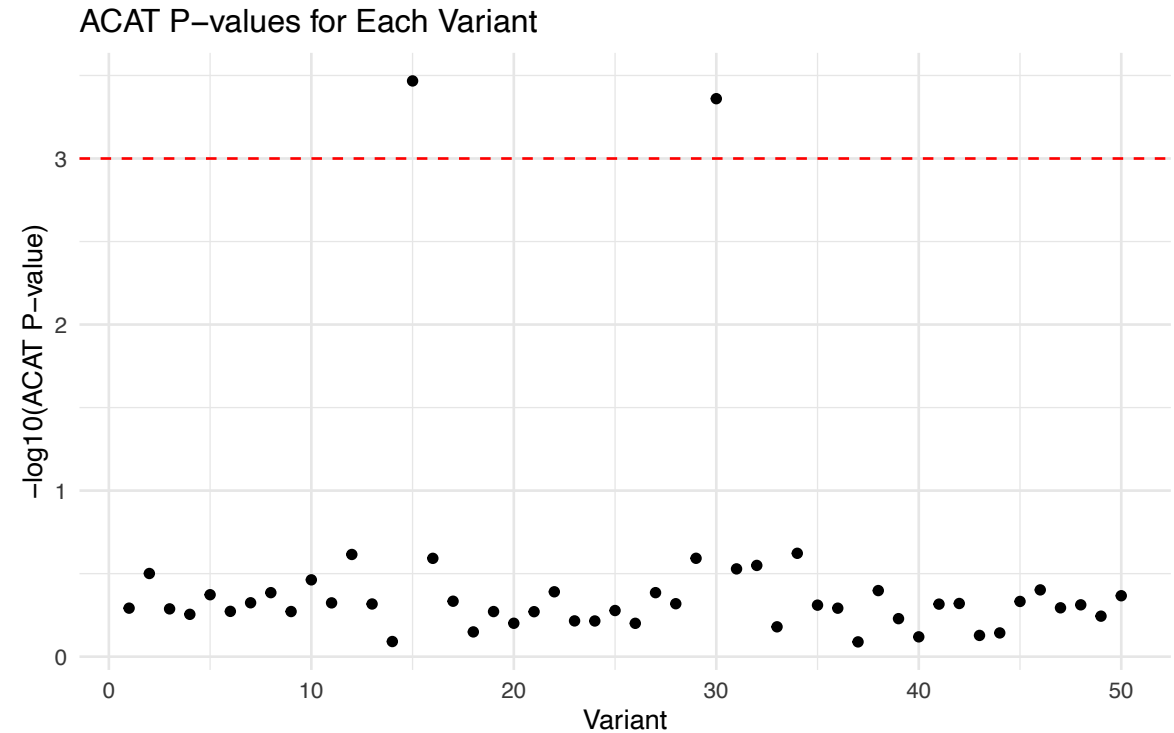
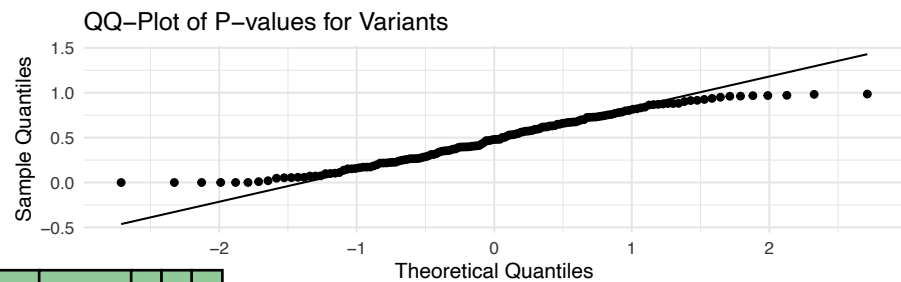
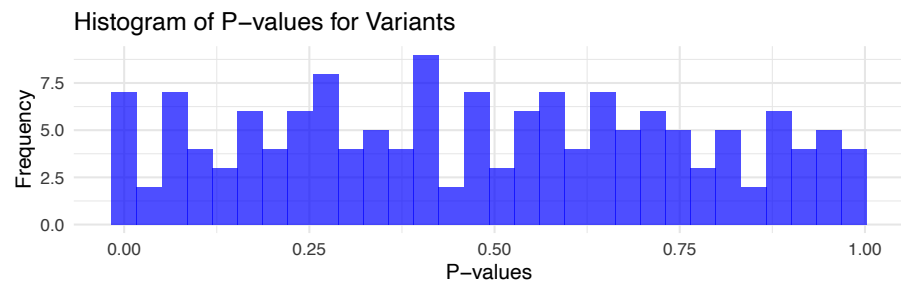
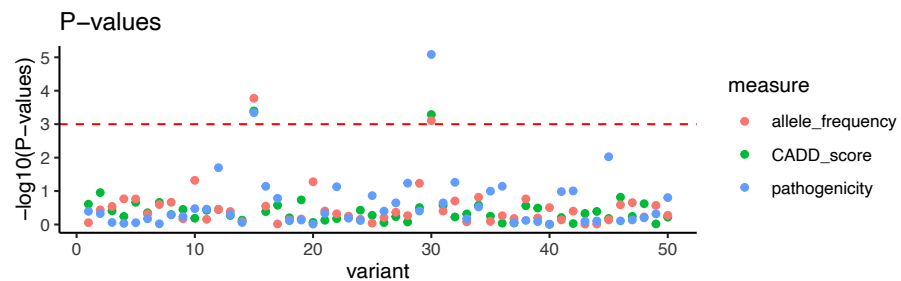
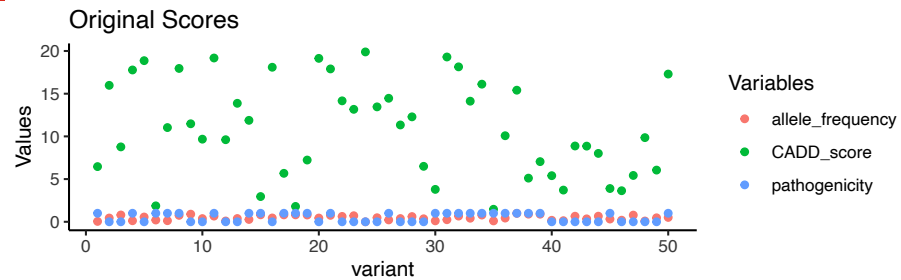
AF

A

T

G

x 3'000'000'000



x150

LOEUF

REVEL

ClinVar

CAD

AF

A

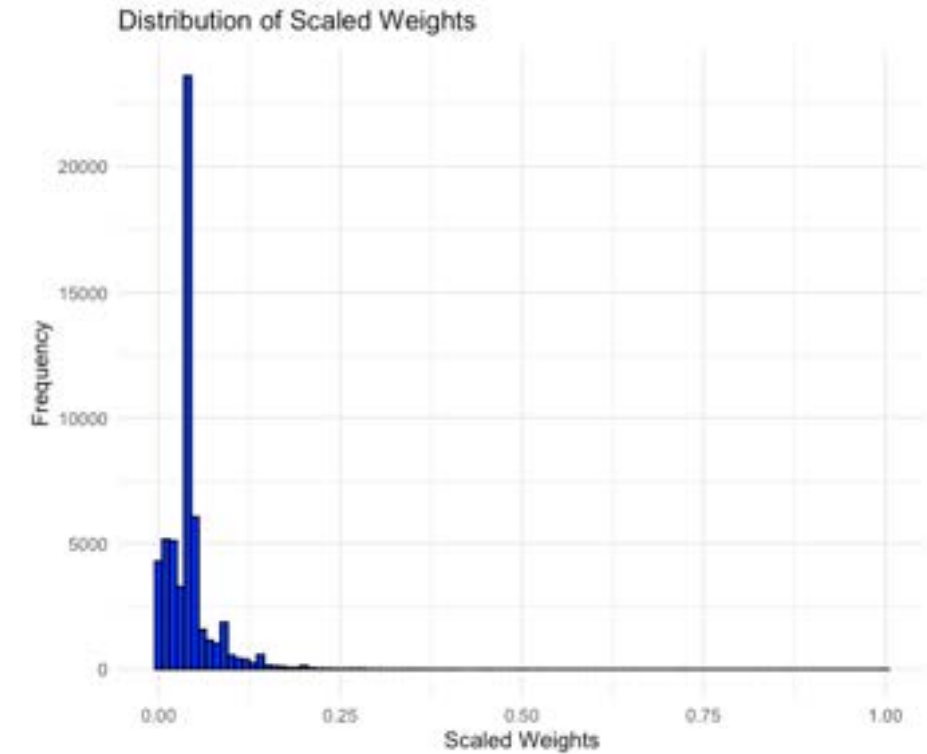
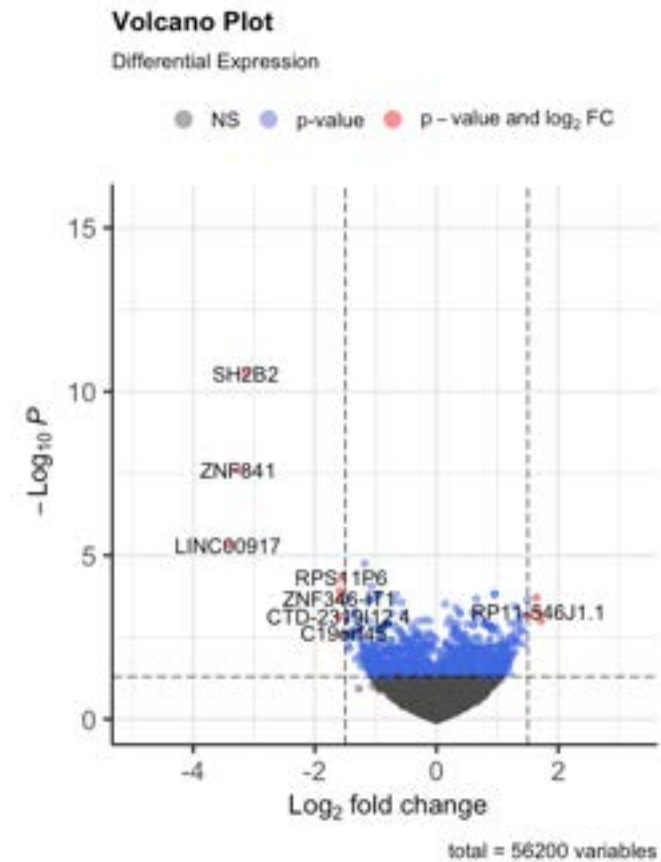
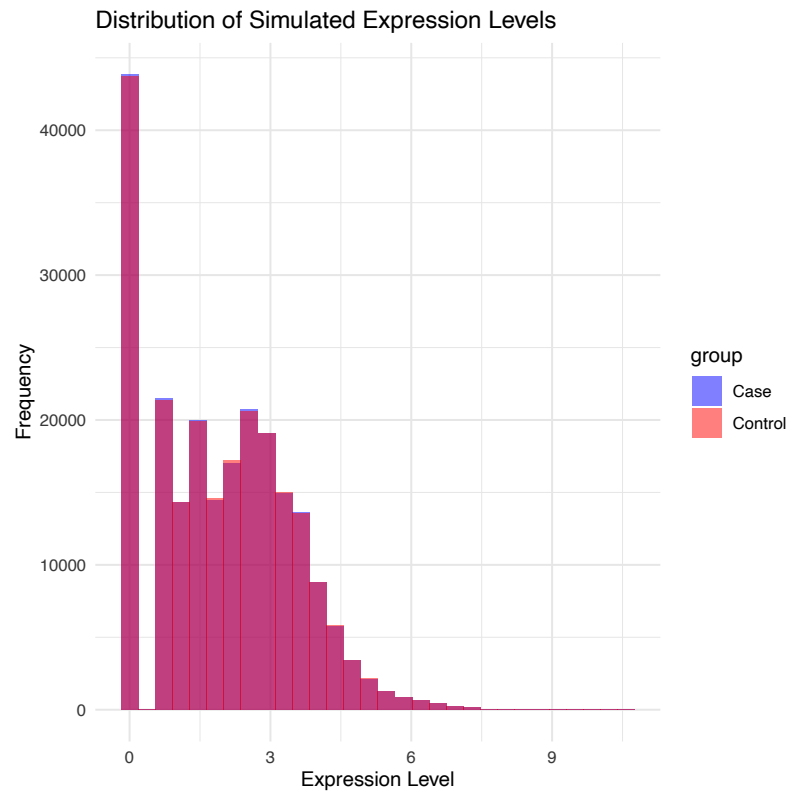
T

G

x 3'000'000'000

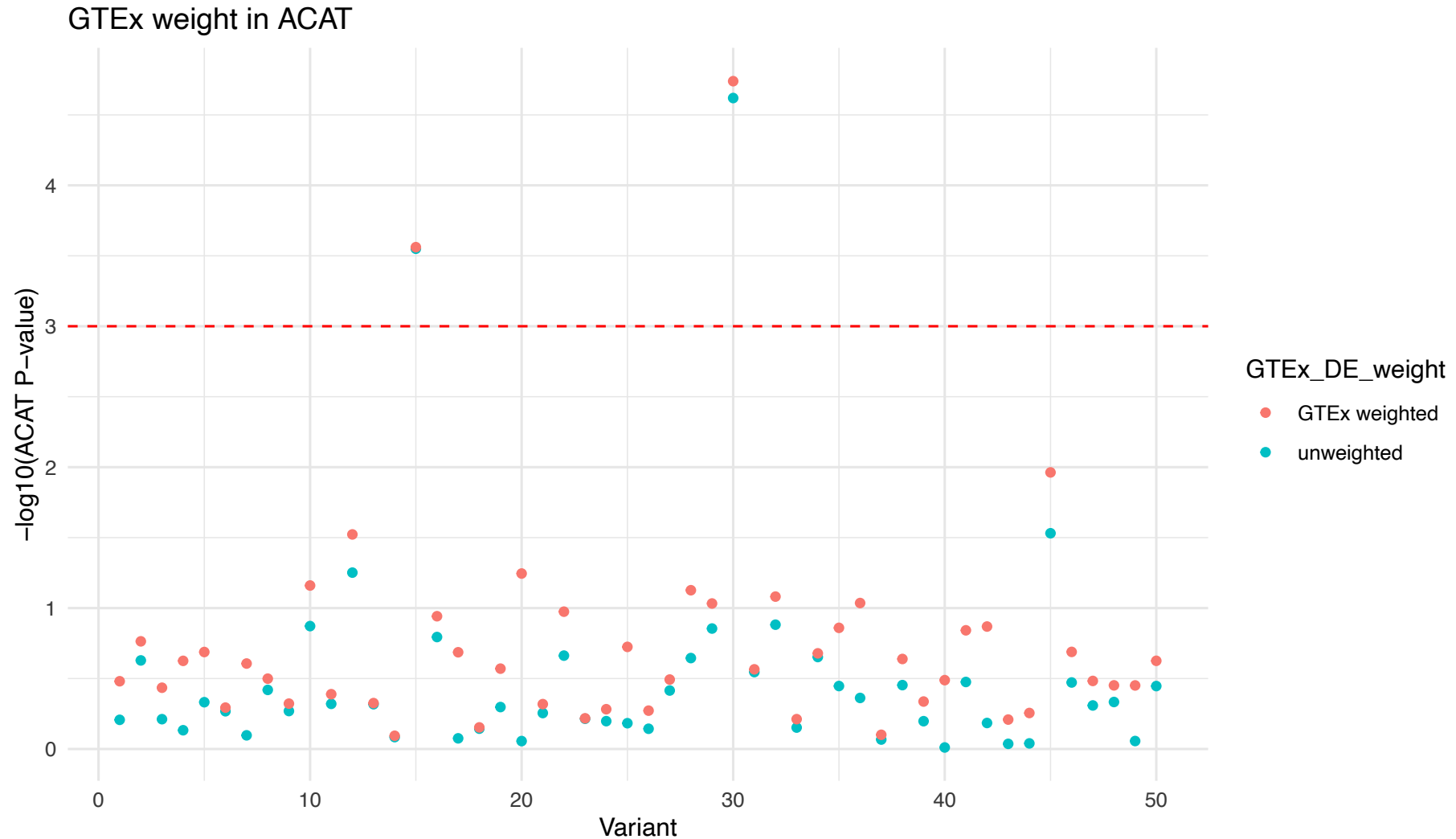
Tests

- **RNAseq** GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct

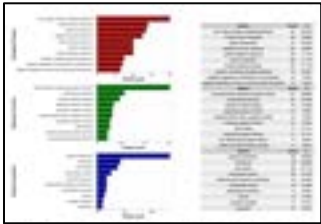
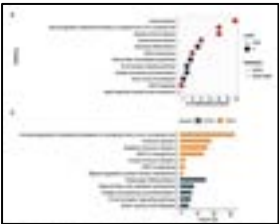


Tests

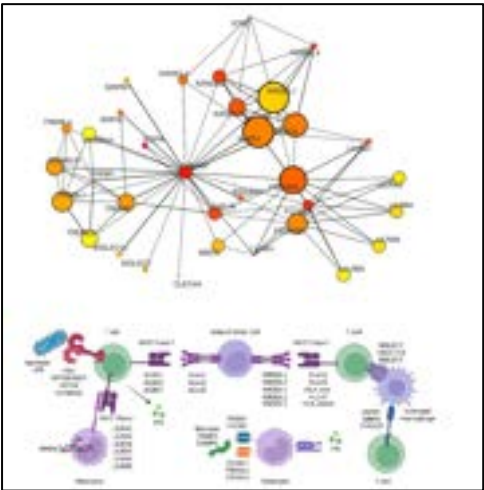
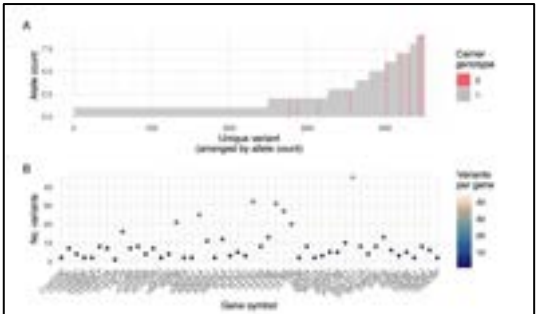
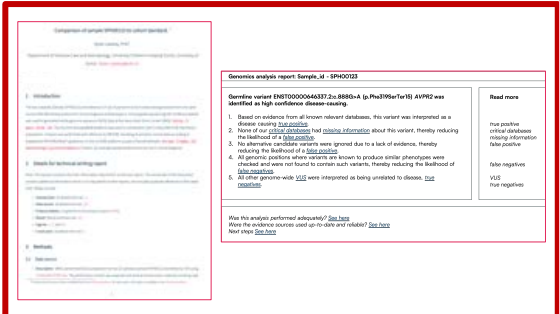
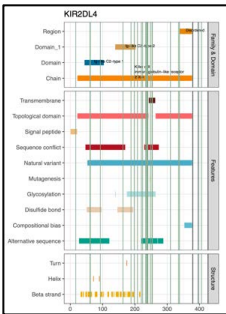
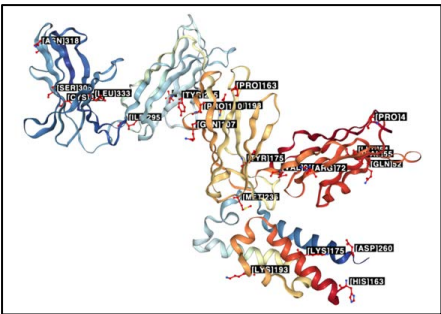
- **RNAseq** GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct



Final result

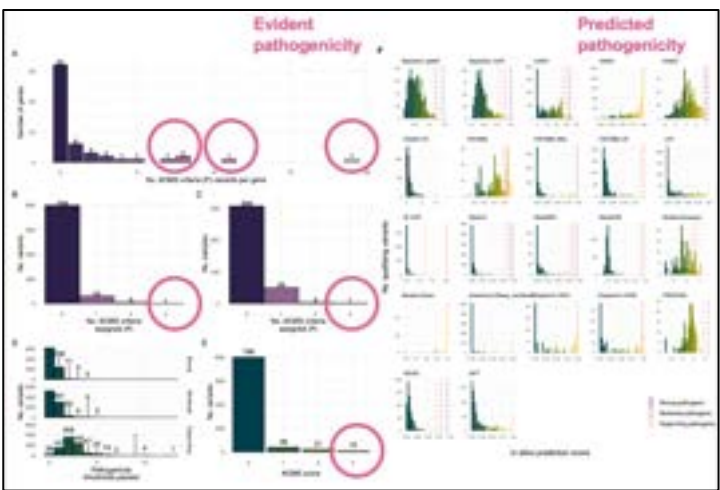


SYMBOL	HGVSp	Consequence	sample.id	group.type
KIR2DL4	ENSP00000279580.3:p.Ala179Val	missense, variant	G0076	1
KIR2DL4	ENSP00000279580.3:p.Asp118LysSer1	frameshift, variant	G0076	1
LILRA1	ENSP00000251372.3:p.Glu436Asp	missense, variant	P001	1
LILRA1	ENSP00000251372.3:p.Glu436Gly	missense, variant	P001	1
LILRA1	ENSP00000251372.3:p.Leu277Pro	missense, variant	ZH114	1
LILRA1	ENSP00000251372.3:p.Ser434Pro	missense, variant	P001	1
LILRA1	NA	splice acceptor, variant	P001	1
LILRA1	NA	splice acceptor, variant	ZH114	1
SIRPG	ENSP00000305529.3:p.Arg348Cys	missense, variant	LAU176	1
SIRPG	ENSP00000305529.3:p.Pro136CdeThr28	frameshift, variant	LAU176	1



Gene function
KIR2DL4, KIR3DL3, KLRD1: Regulate NK cell activity, central to innate immune response.
LILRA1, LILRB1, LILRB4: MCH class I
TREML1,2,4: Pathogen recognition
SIGLEC1: Pathogen recognition, macrophage activation
SIRPG: immunomodulation; can lead to sepsis' hallmark overactive response.

Key Processes
Antigen presentation
Signaling receptors
Interactions between lymphoid & non-lymphoid cells
Adaptive & innate immune responses



- **ACMGuru**
 - uniprotR
 - AutoDestructR
 - gene-illustrate
 - untangleR
 - getDiscussion
 - interpretation
- **Clinical genetics report**
- **Technical quality report**
- **Association report**
- **Workflow code**
- **Logs**

SPHN RDF Schema: Sequencing assay

Comparison of sample SPH00123 to variant standard:

SPH00123, P01

Reported by: SPH00123, P01

SPH00123, P01

Genomics analysis report: Sample_Id - SPH00123

Germine variant ENST00000646337.2:c.888G>A (p.Phe319SerTer15) AVPR2 was identified as high confidence disease-causing.

- Based on evidence from all known relevant databases, this variant was interpreted as a disease causing **true positive**.
- None of our **critical databases** had **missing information** about this variant, thereby reducing the likelihood of a **false positive**.
- No alternative candidate variants were ignored due to a lack of evidence, thereby reducing the likelihood of a **false positive**.
- All genomic positions where variants are known to produce similar phenotypes were checked and were not found to contain such variants, thereby reducing the likelihood of **false negatives**.
- All other genome-wide **VUS** were interpreted as being unrelated to disease, **true negatives**.

Read more

true positive
critical databases
missing information
false positive
false negatives
VUS
true negatives

Was this analysis performed adequately? [See here](#)
Were the evidence sources used up-to-date and reliable? [See here](#)
Next steps [See here](#)

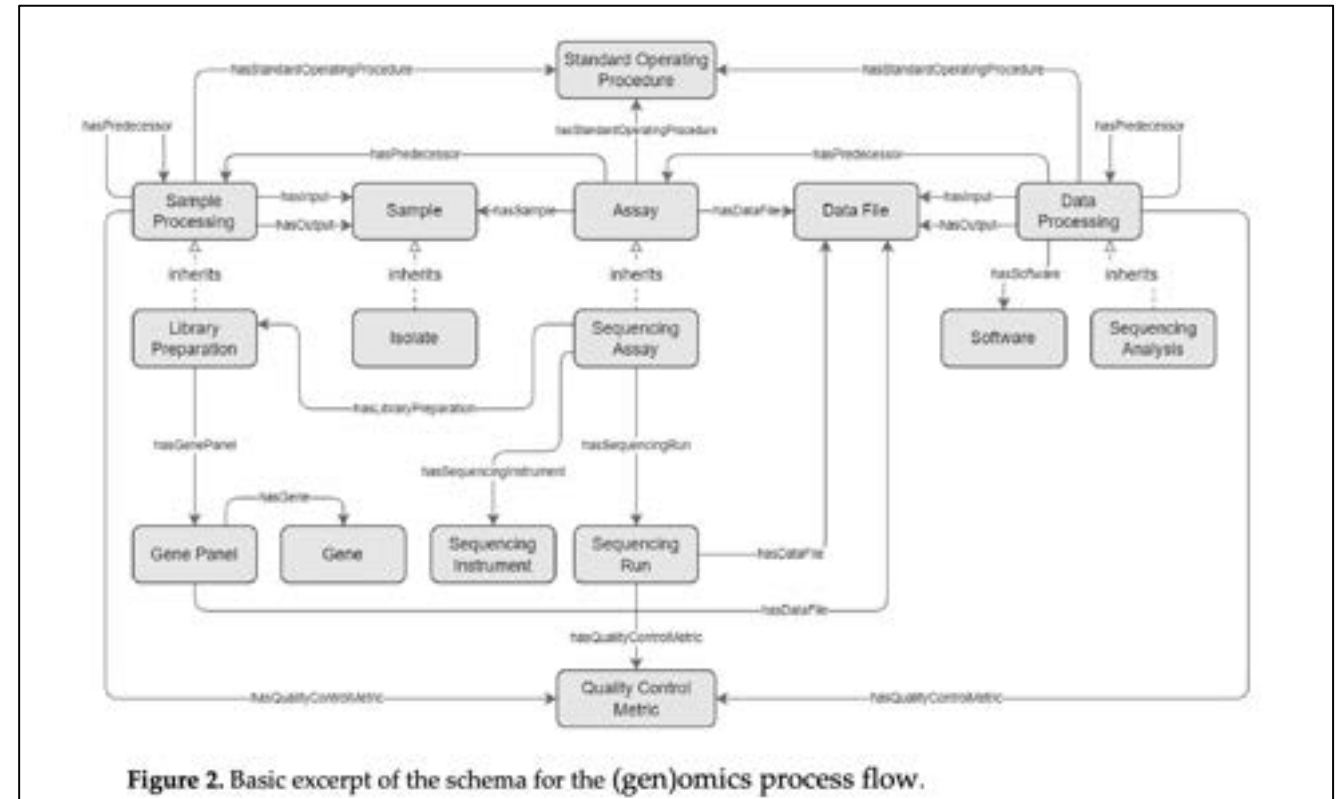
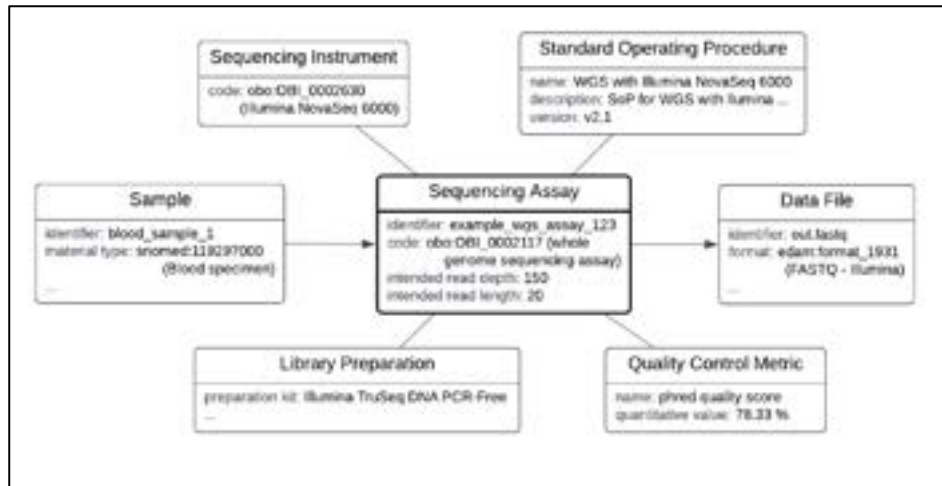


Figure 2. Basic excerpt of the schema for the (gen)omics process flow.



Future

