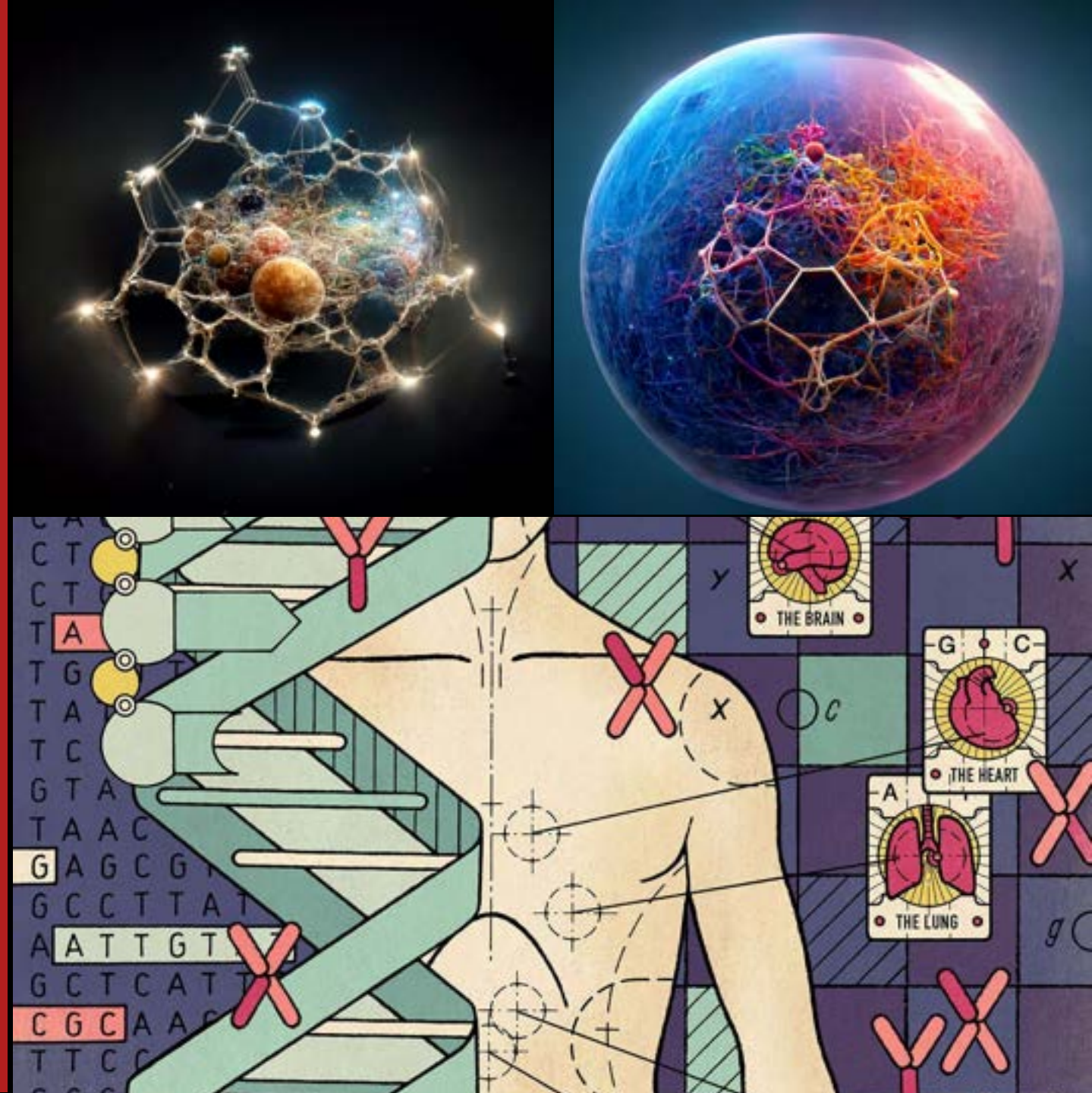
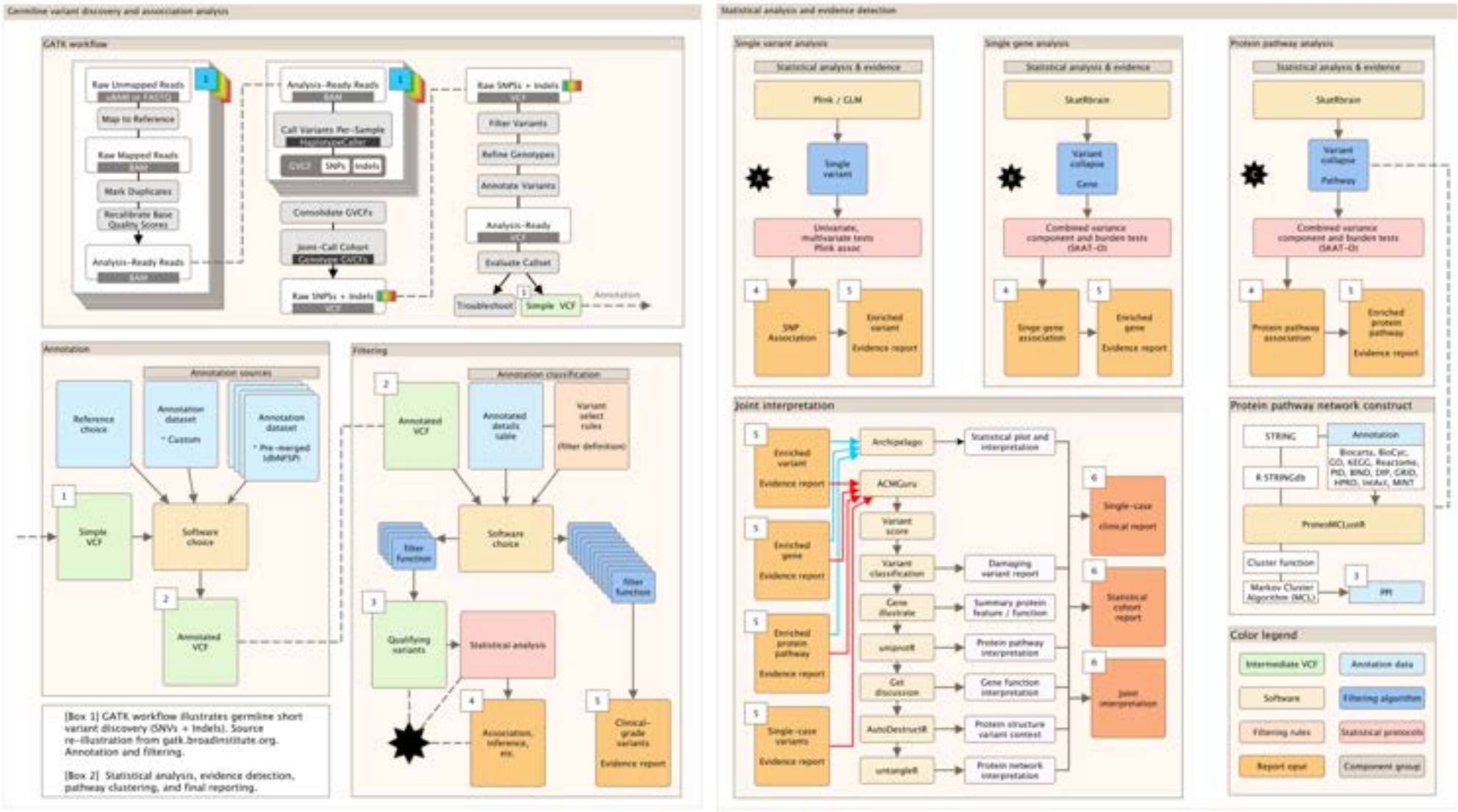
Annotation-aware genetic analysis



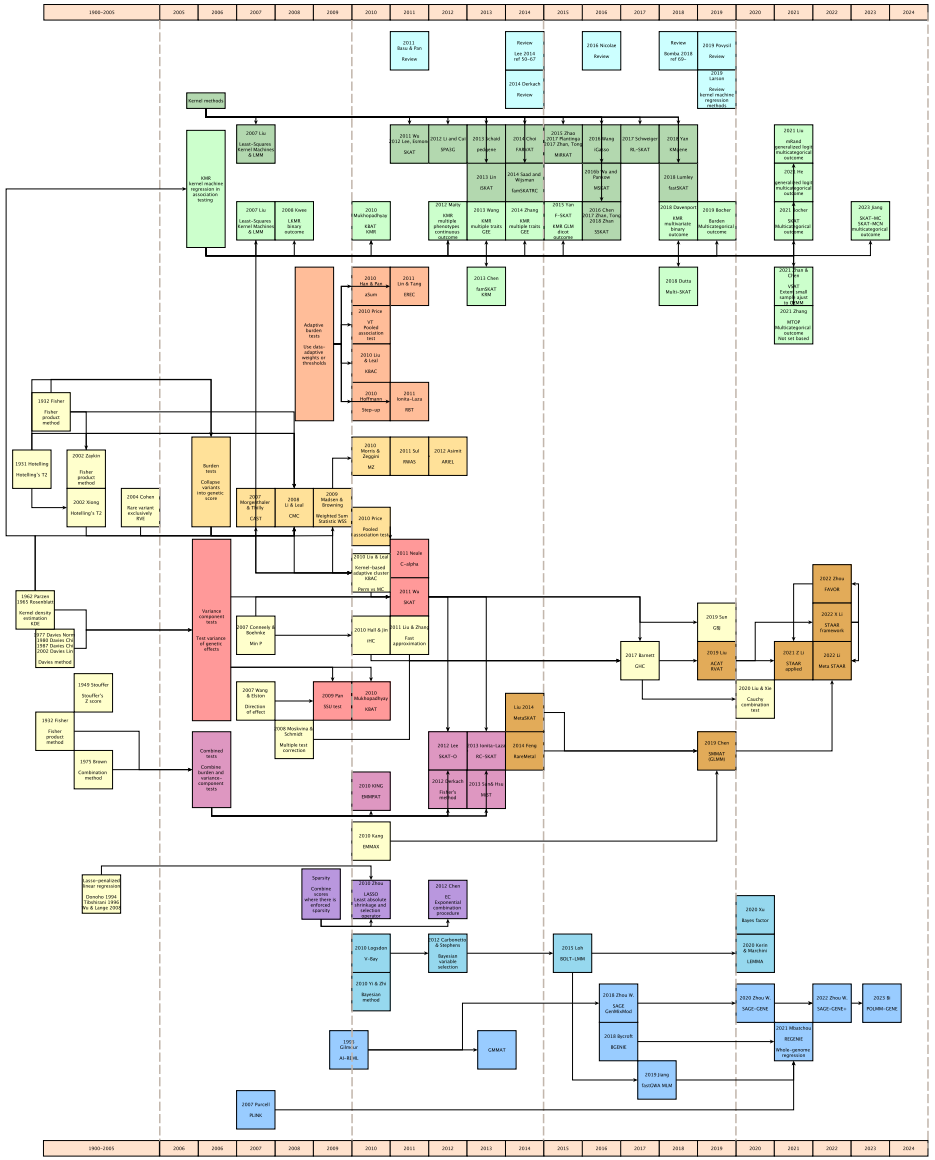
Cauchy–Lorentz distribution

The Witch of Agnesi

Model application

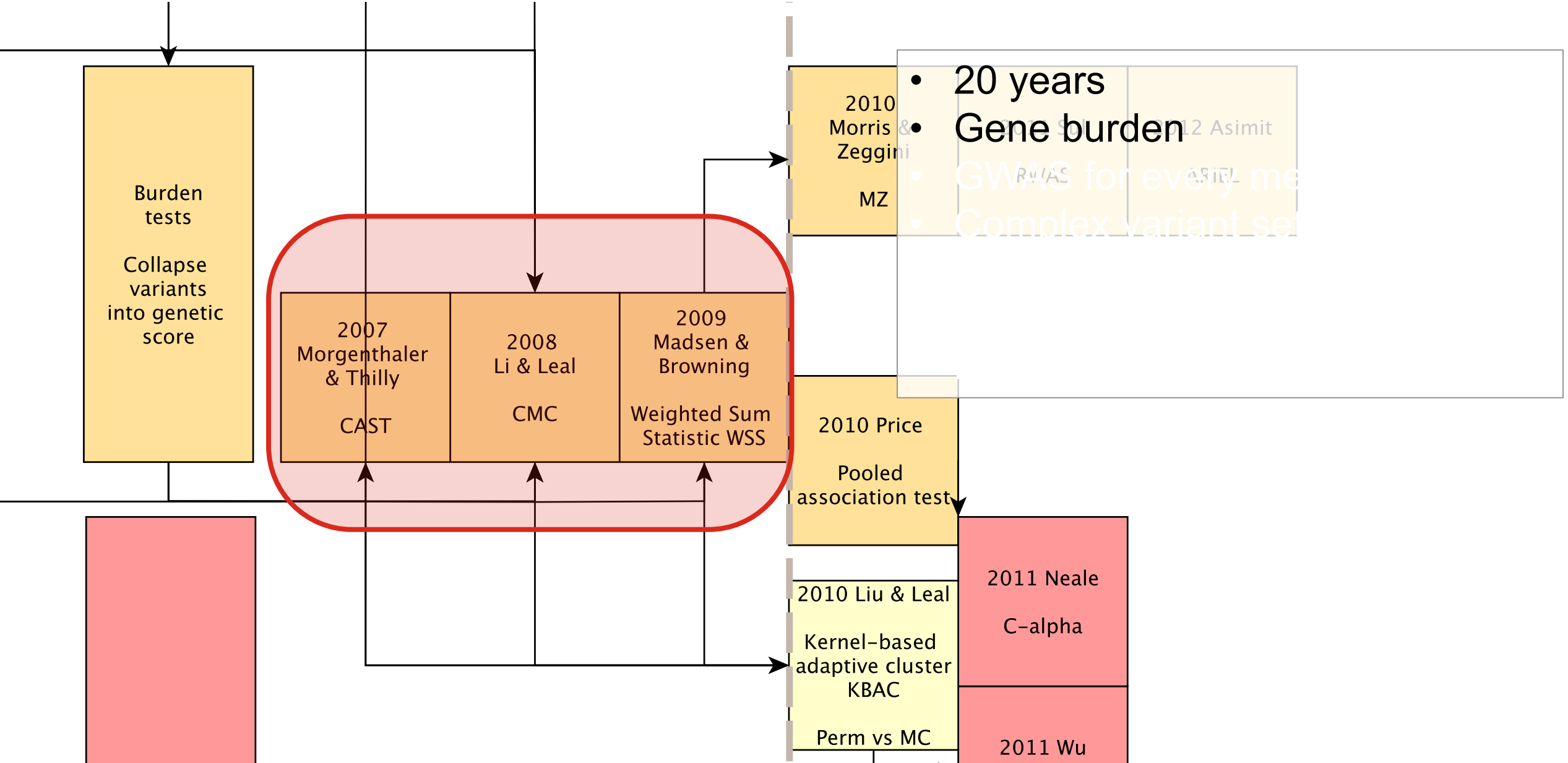


Variant set analysis

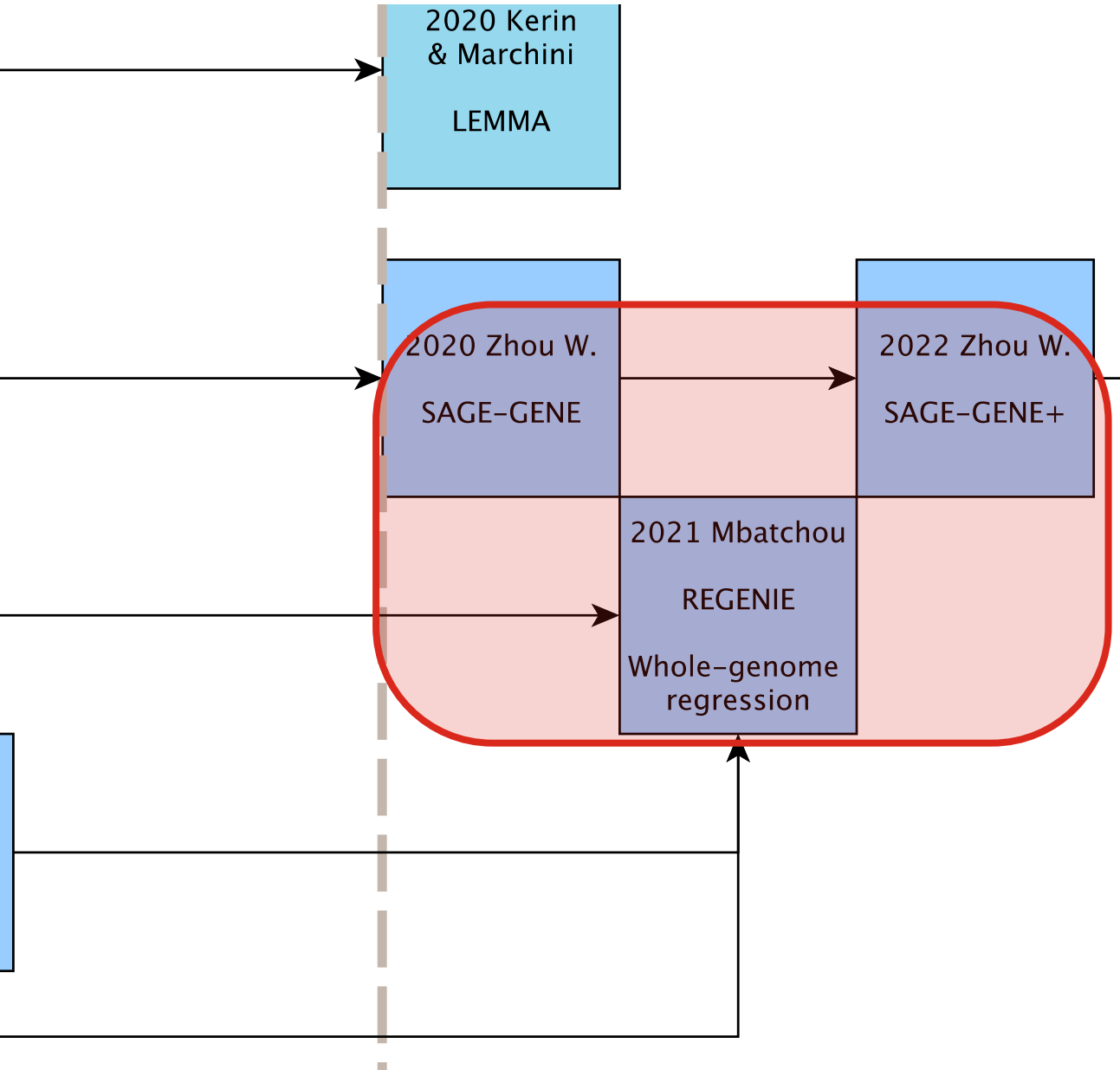


- 20 years

Variant set analysis

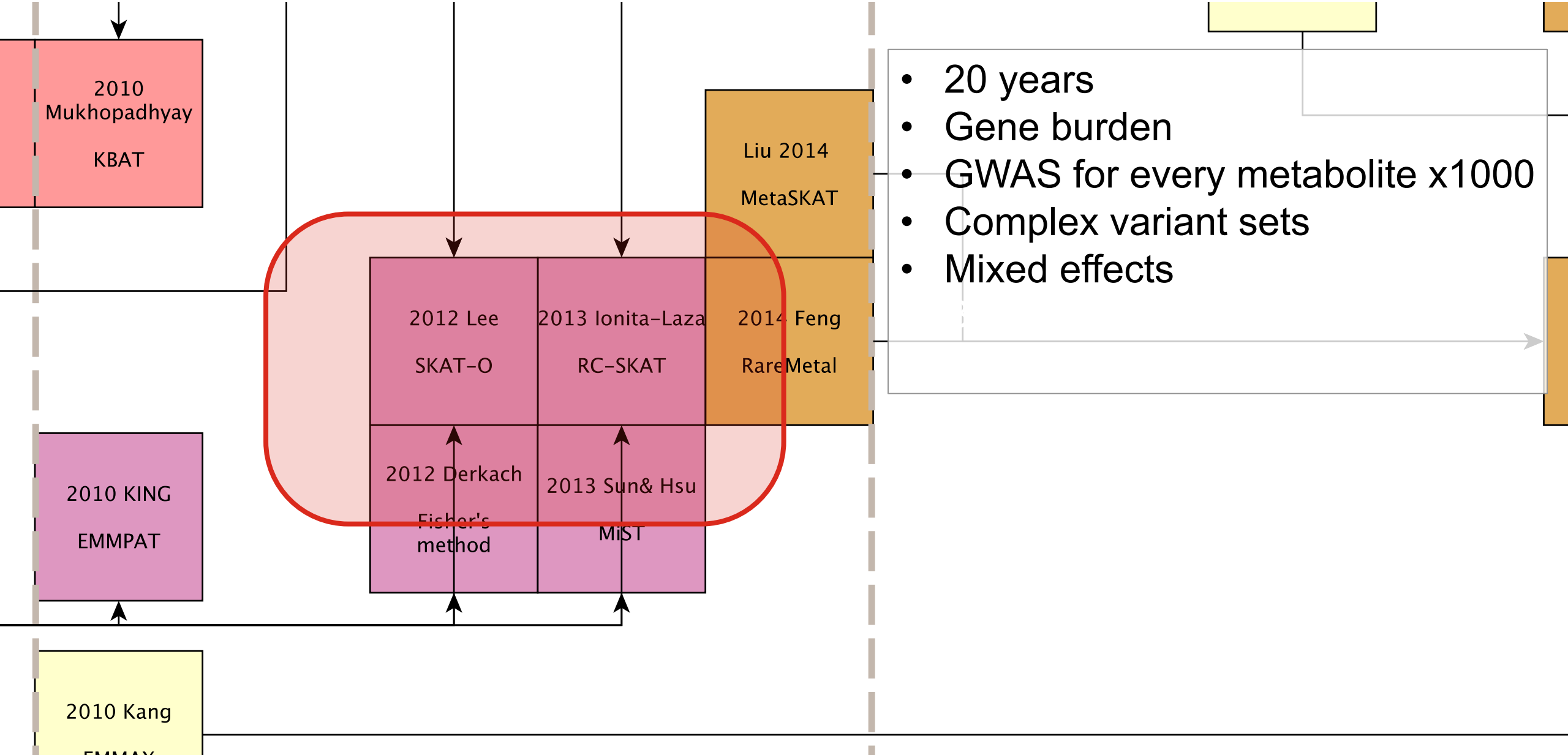


Variant set analysis

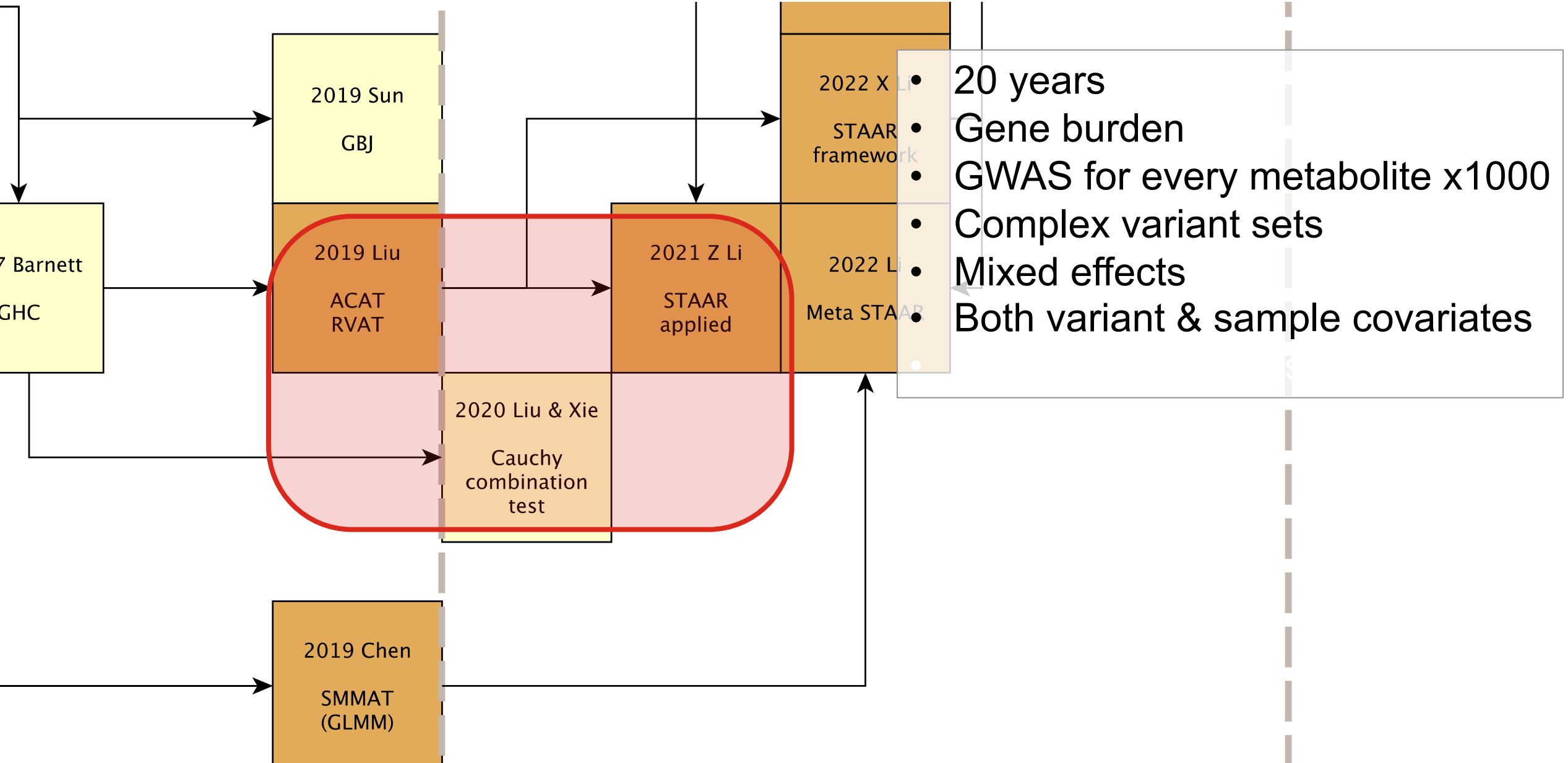


- 20 years
- Gene burden
- GWAS for every metabolite x1000

Variant set analysis

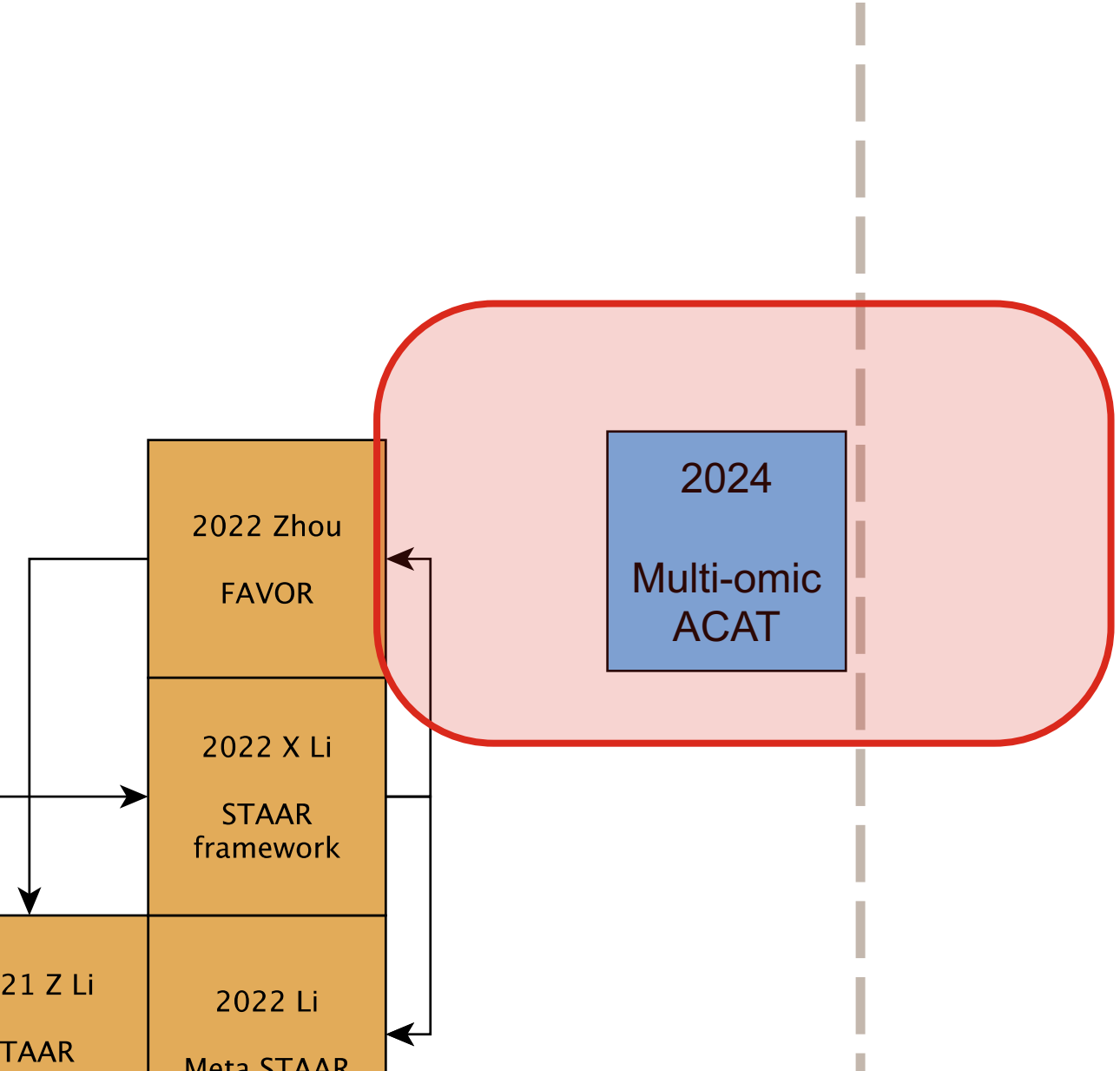


Variant set analysis



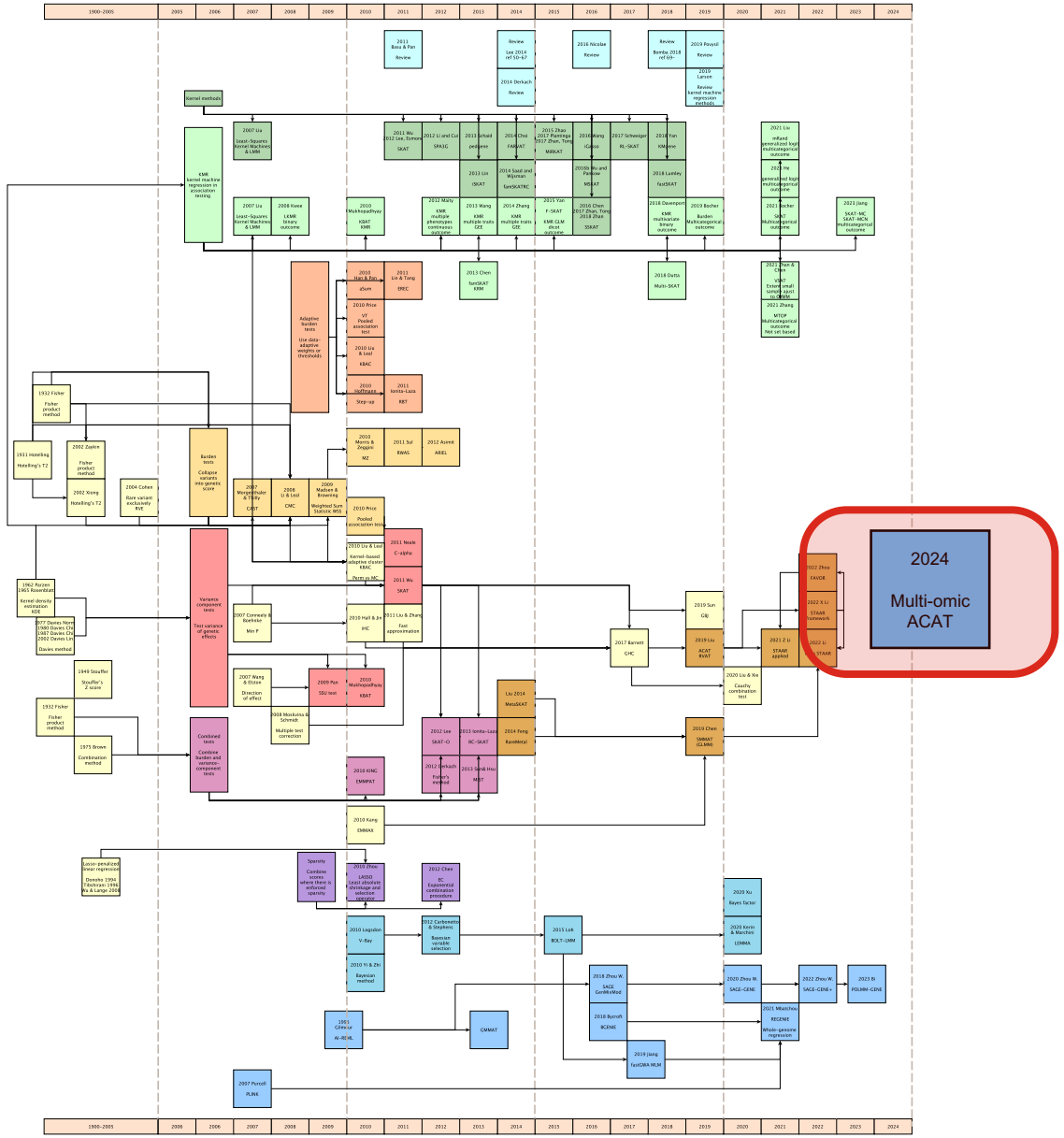
- 20 years
- Gene burden
- GWAS for every metabolite x1000
- Complex variant sets
- Mixed effects
- Both variant & sample covariates

Variant set analysis



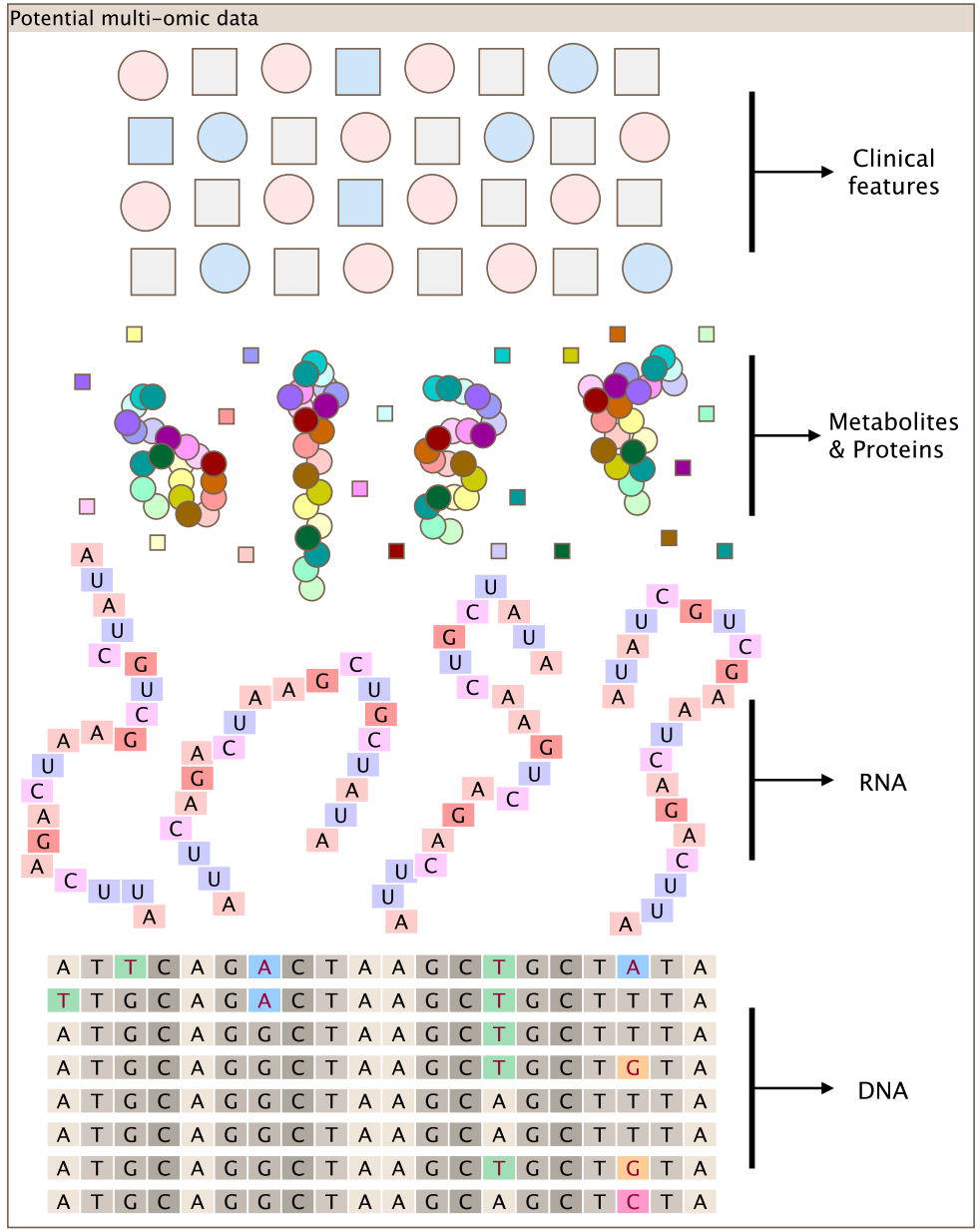
- 20 years
- Gene burden
- GWAS for every metabolite x1000
- Complex variant sets
- Mixed effects
- Both variant & sample covariates
- Multi-omic variables

Variant set analysis



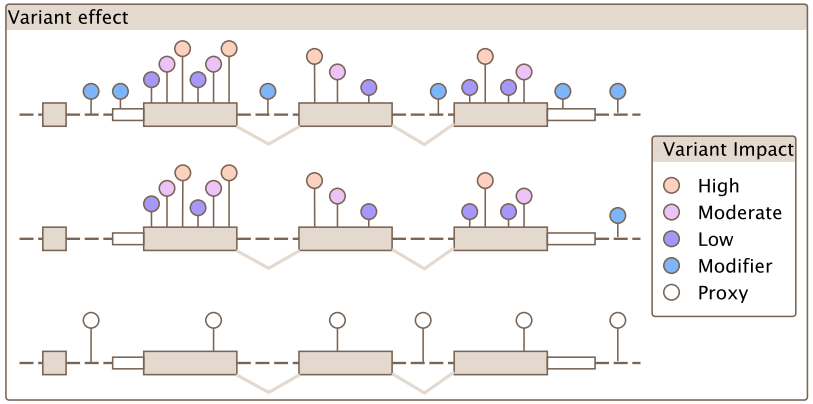
- 20 years
- Gene burden
- GWAS for every metabolite x1000
- Complex variant sets
- Mixed effects
- Both variant & sample covariates
- Multi-omic variables

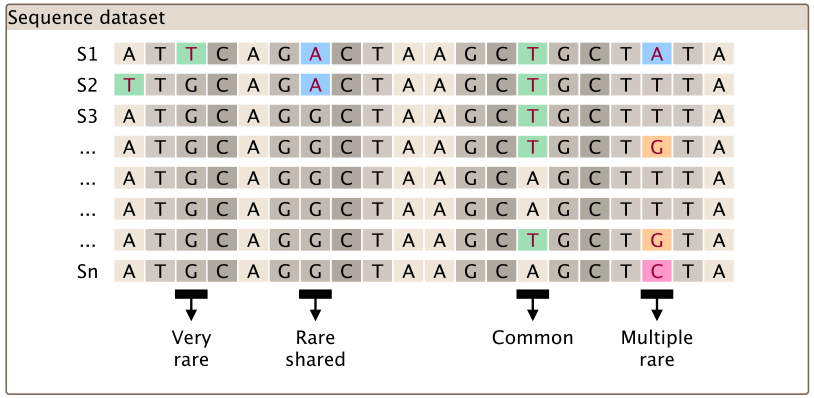
Genetic association with disease

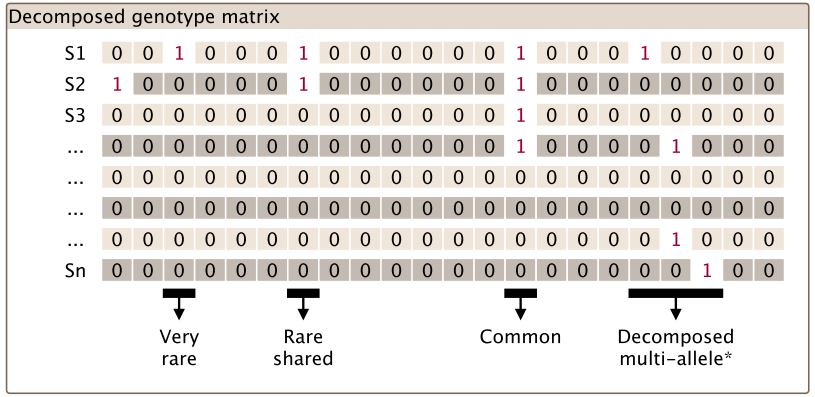


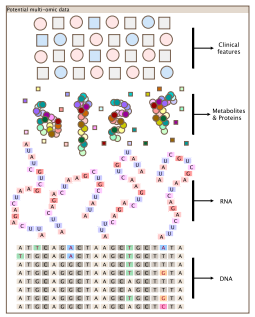
- Genetic variant → disease
- Include other biological info in model

Genetic association with disease



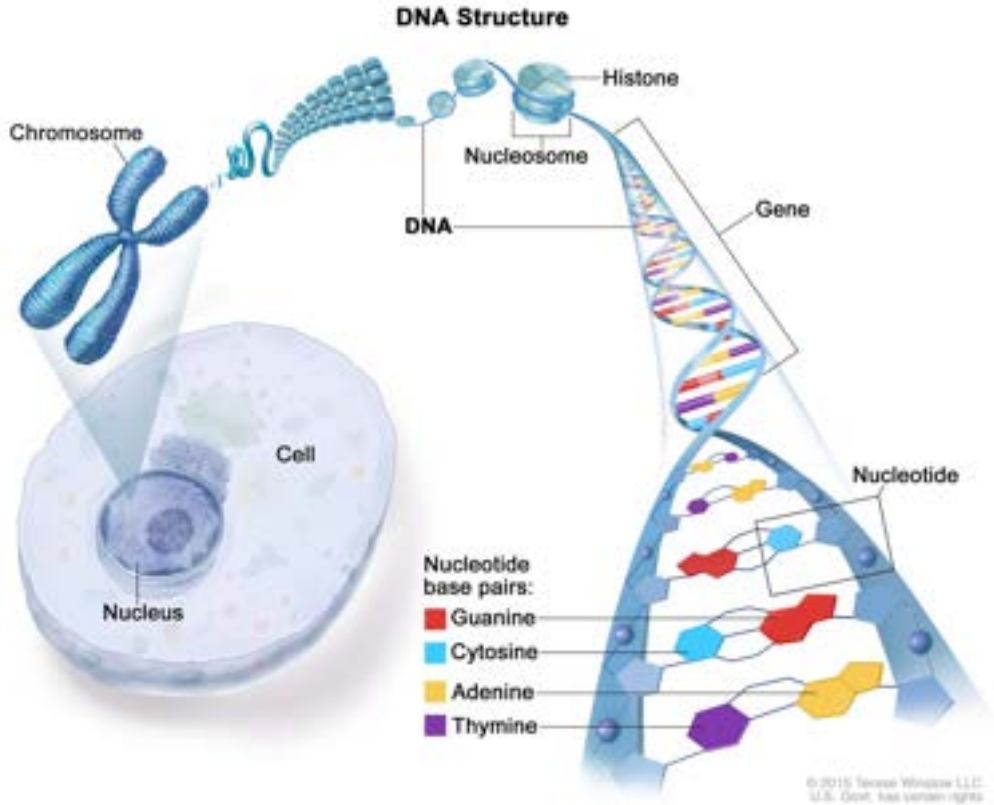


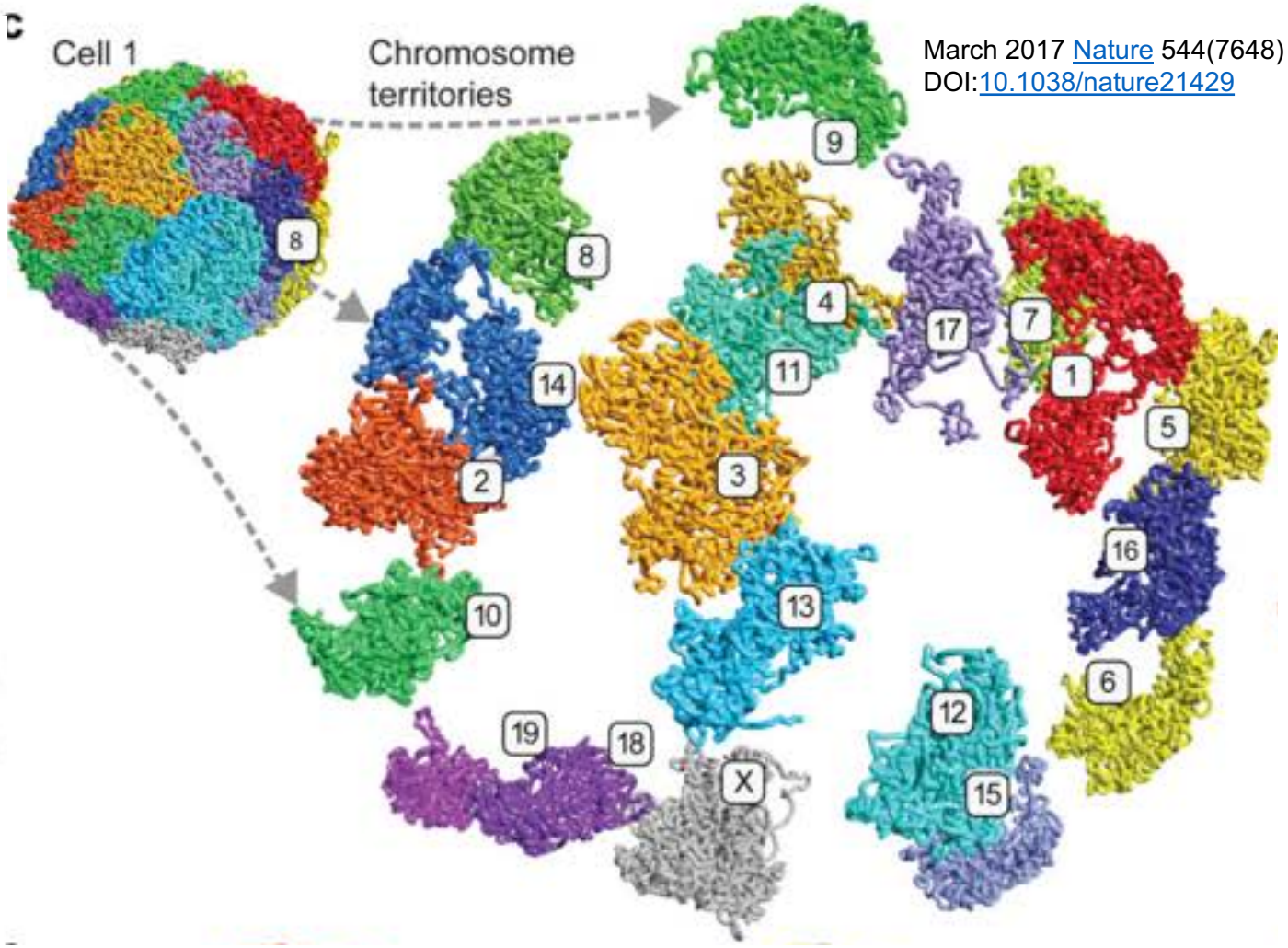




- Genetic variant → disease
- Include other biological info in model
- Consider variant effect

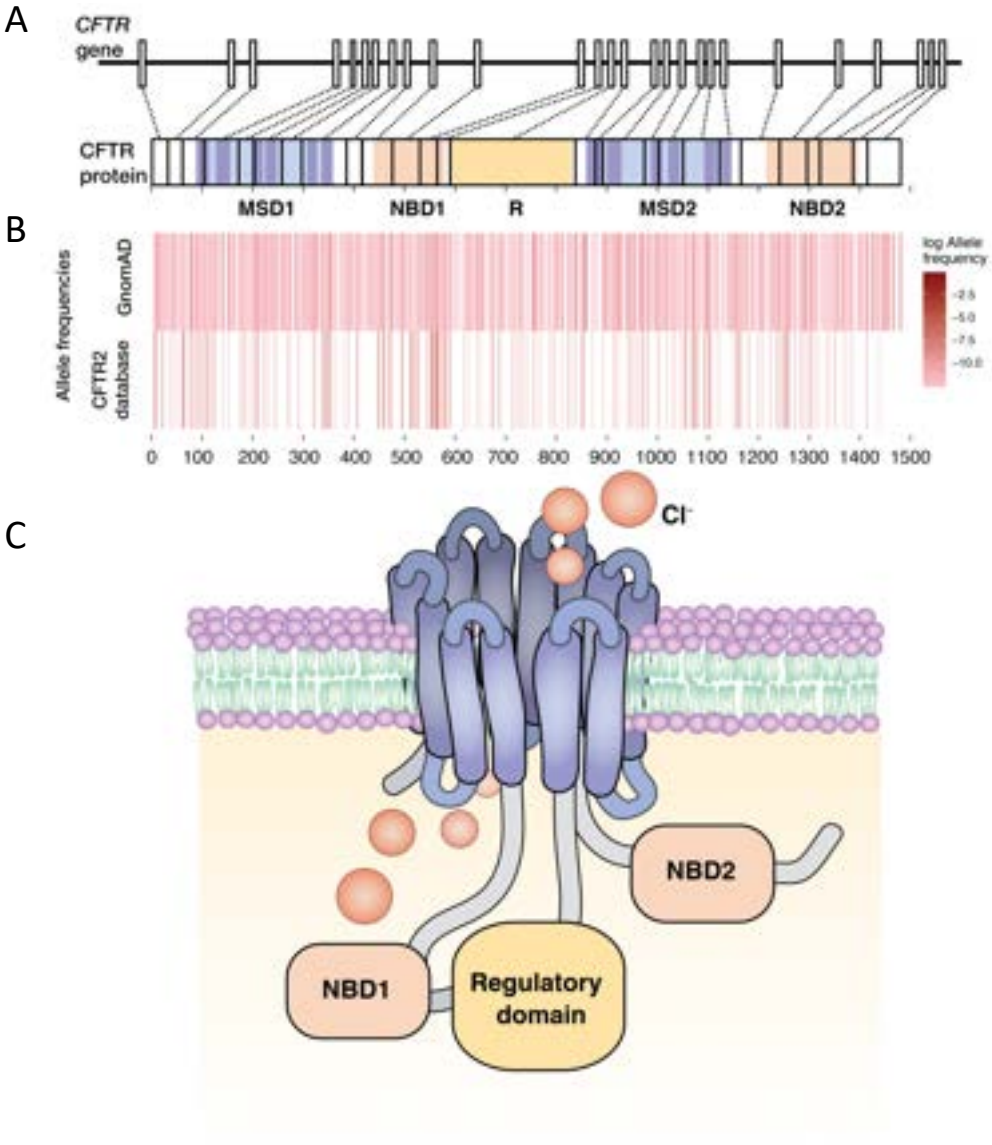
Genomic variants



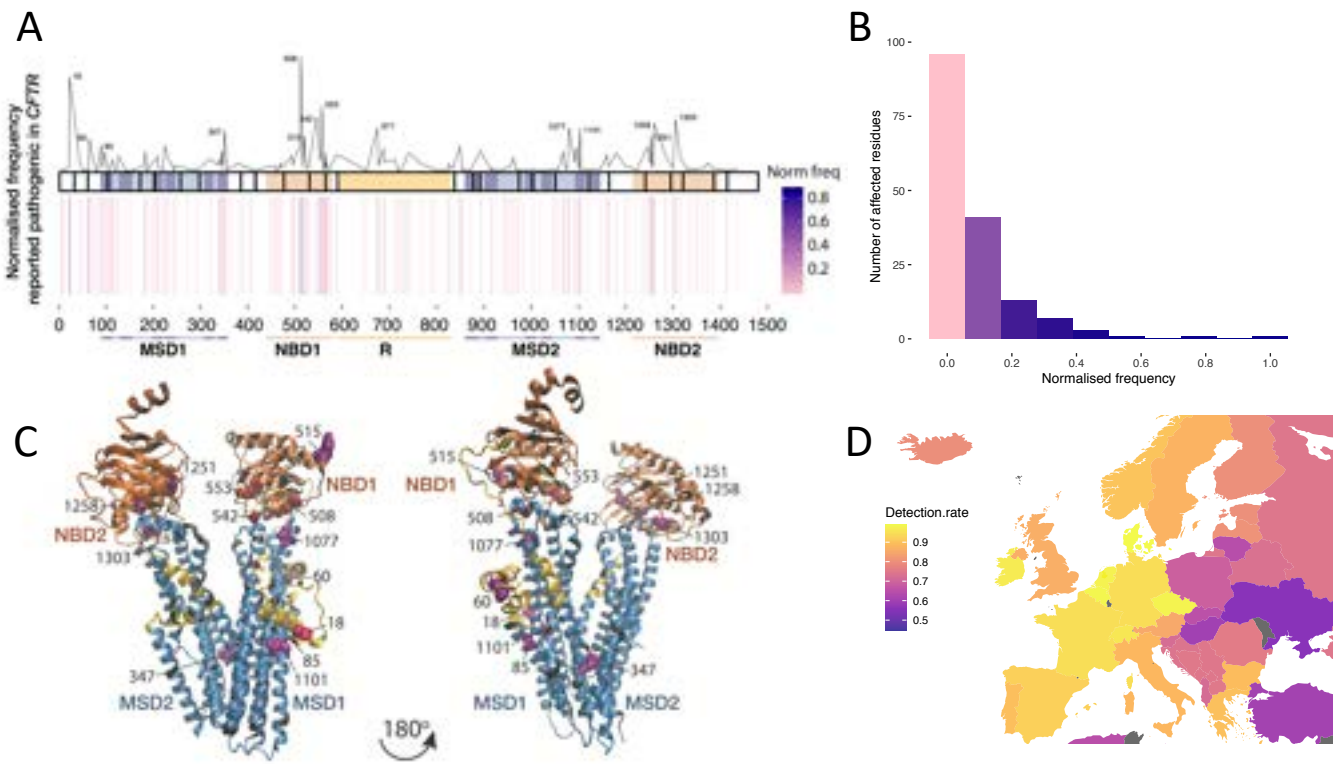


March 2017 Nature 544(7648)
DOI:10.1038/nature21429

Example: Cystic fibrosis (*CFTR*)



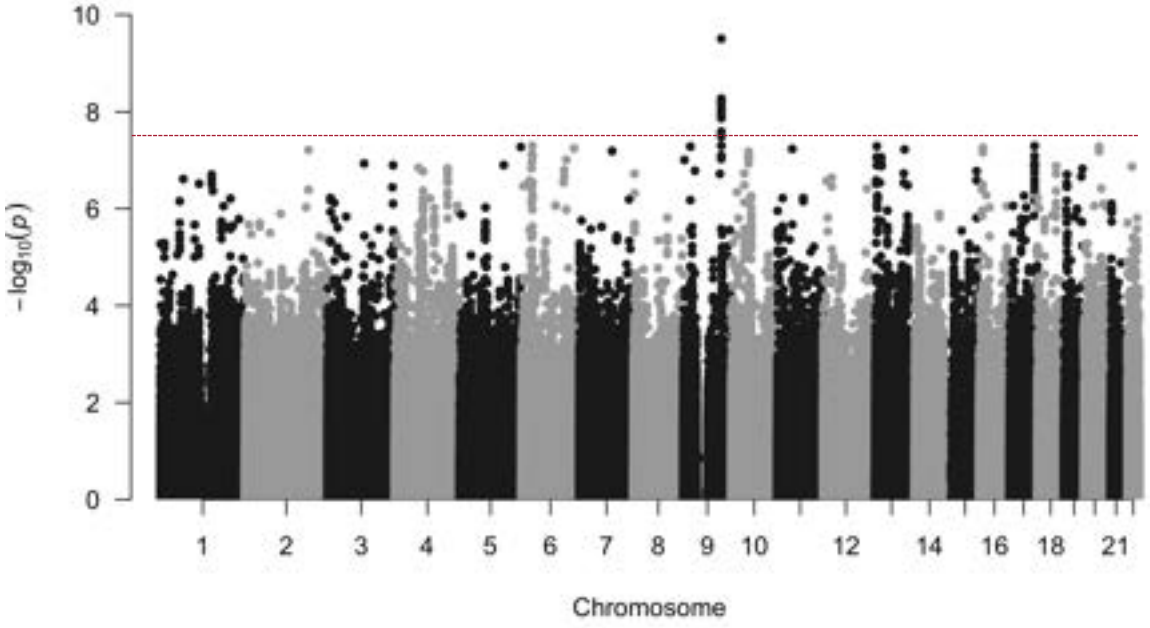
- Known mechanism quantified

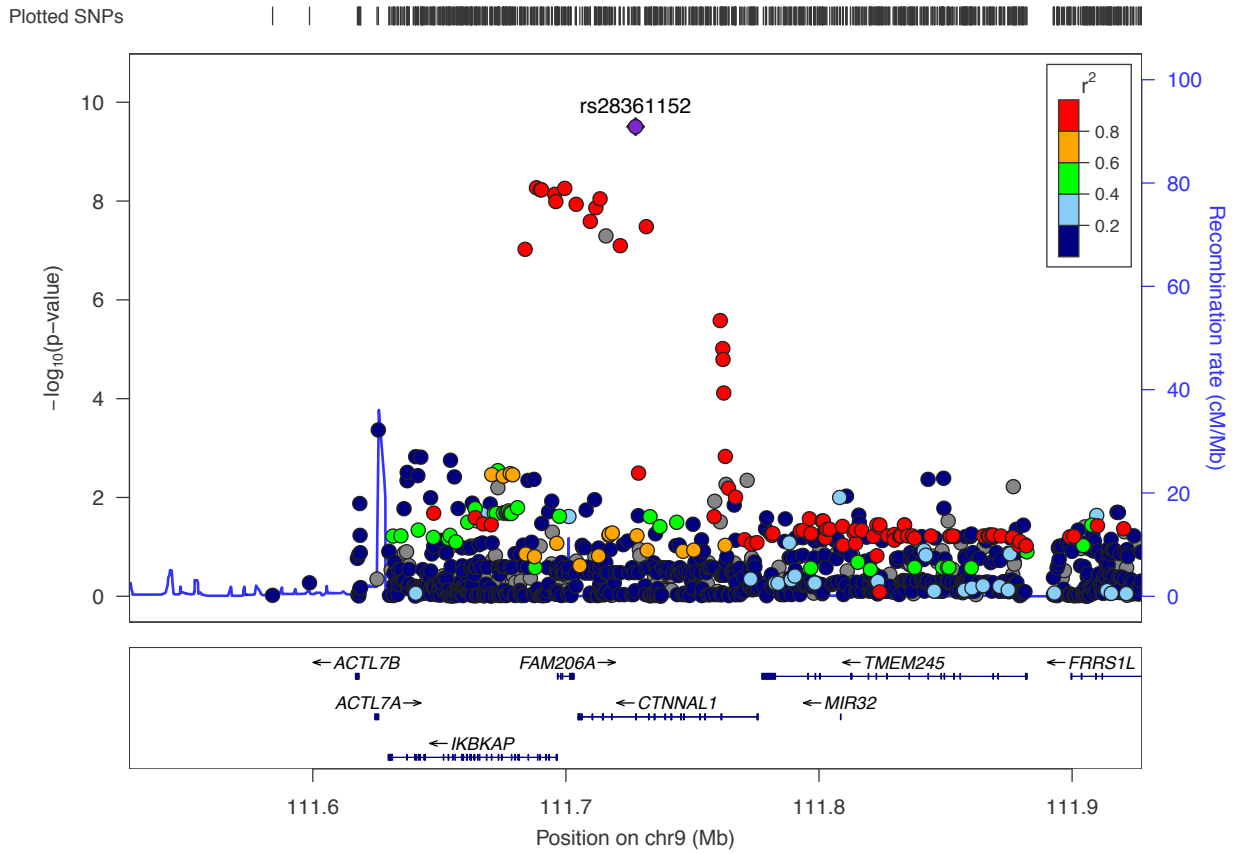


2023 *JACI* doi.org/10.1016/j.jaci.2023.01.035

Example: Sepsis susceptibility

- Purely statistical
- Unknown mechanism



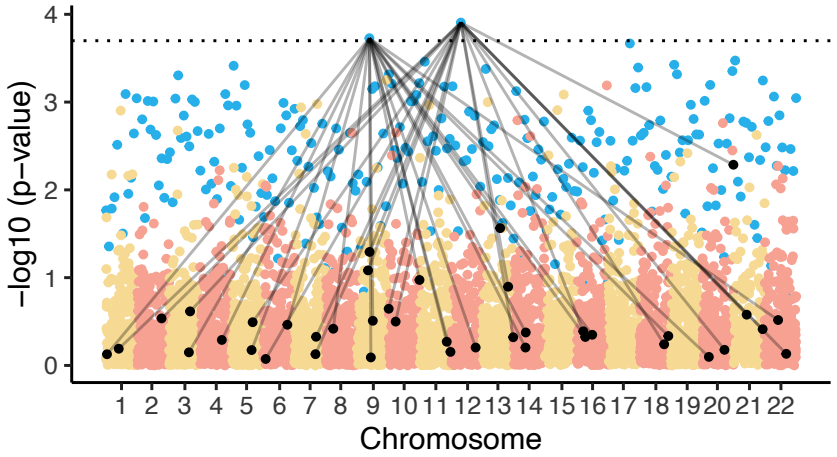


Pre-print 2023

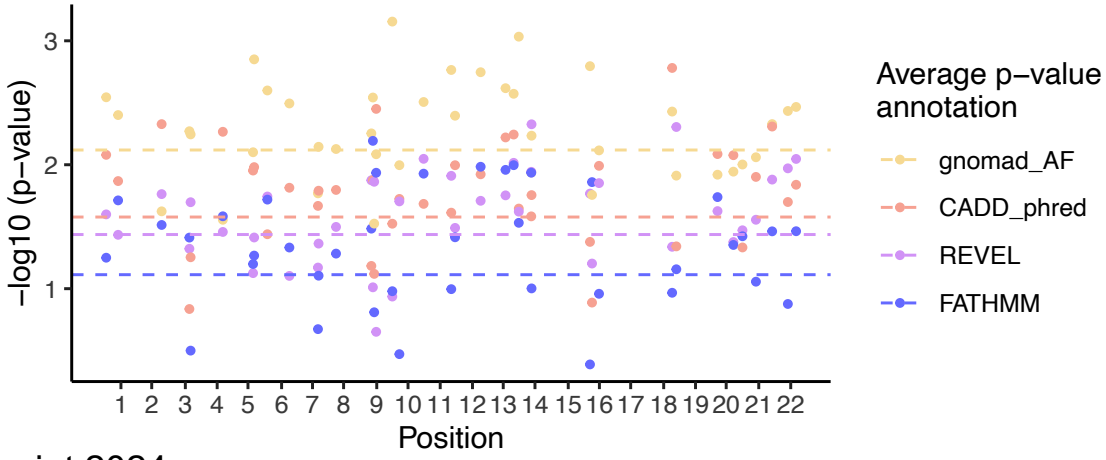
Pathway analysis

Archipelago Plot

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



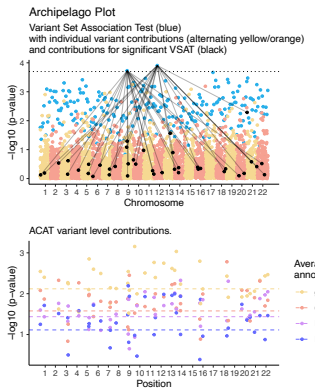
ACAT variant level contributions.

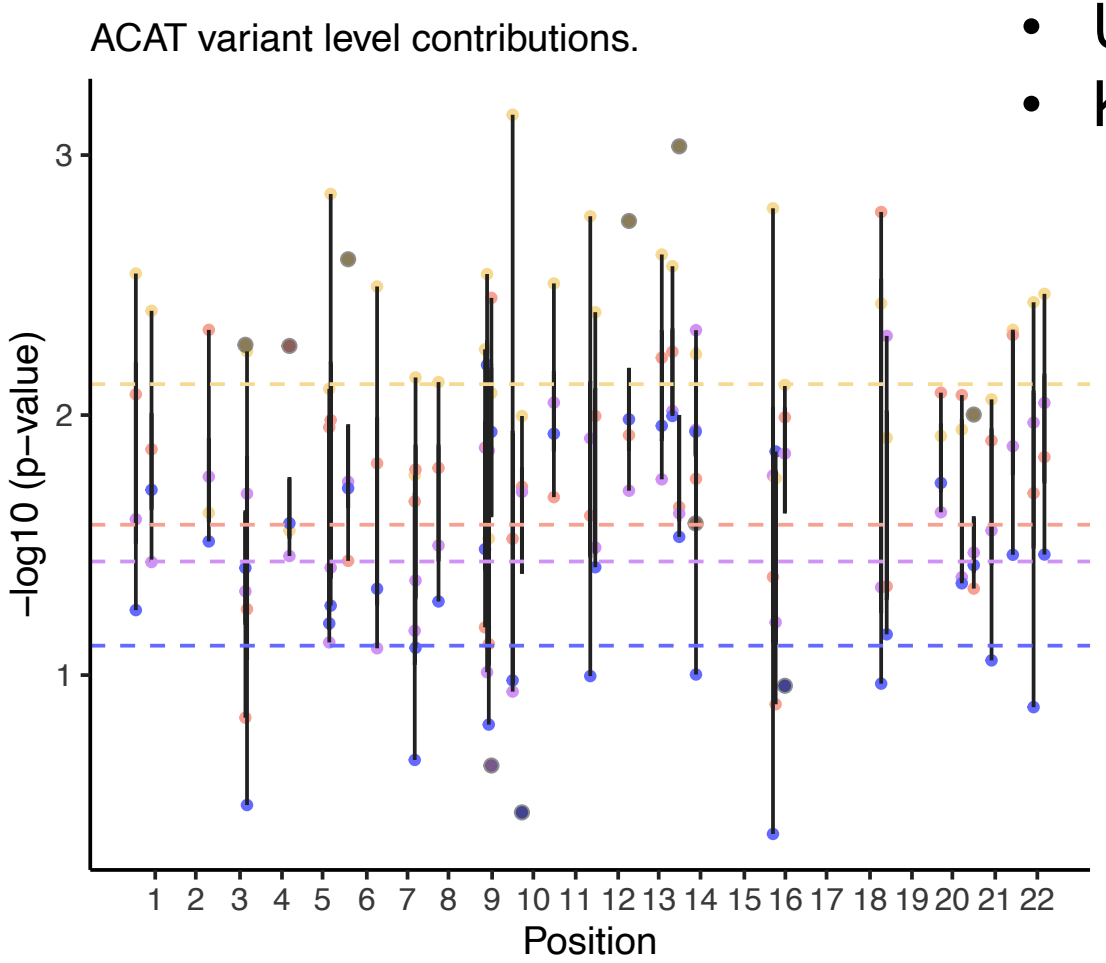


- Purely statistical
- Unknown mechanism
- Known biology

Pre-print 2024

Pathway analysis





- Purely statistical
- Unknown mechanism
- Known biology

Pre-print 2024

Multi variable combination

Aggregated Cauchy Association Test (ACAT)

Burden= $\sum_j^p w_j U_j$

Powerful if SNV effects on LDL in the same direction

SKAT= $\sum_j^p w_j^2 U_j^2$

Powerful if SNV effects on LDL in different directions

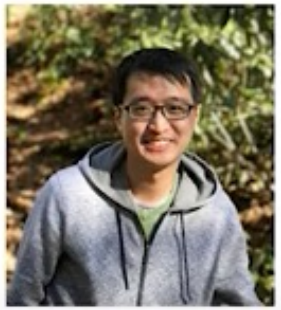
U_j=assoc score stat of SNV j
p=# of SNPs in APOE enhancer

Powerful correlated p-value combination methods

Transform p-value to Cauchy

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

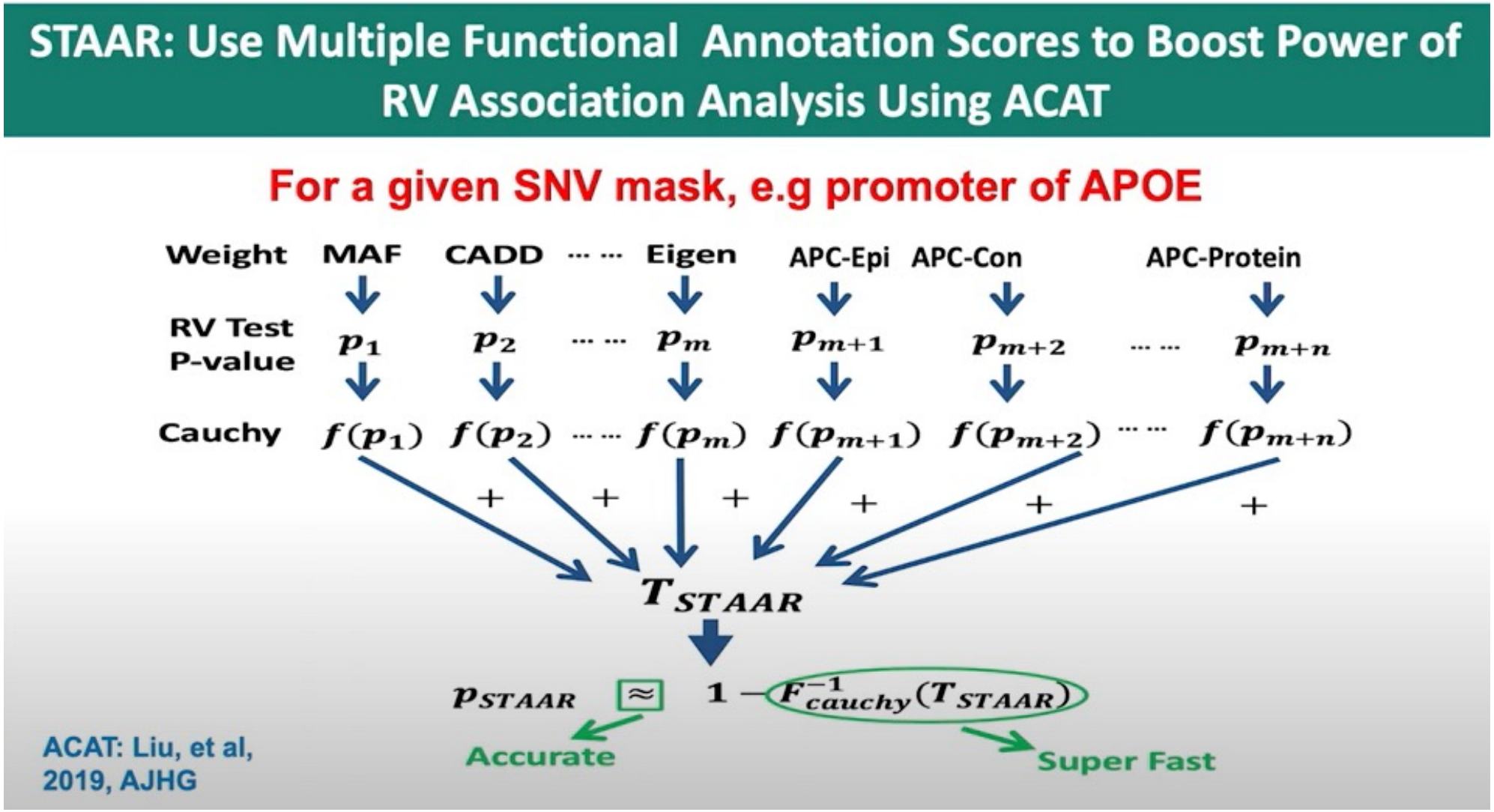
Weights



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

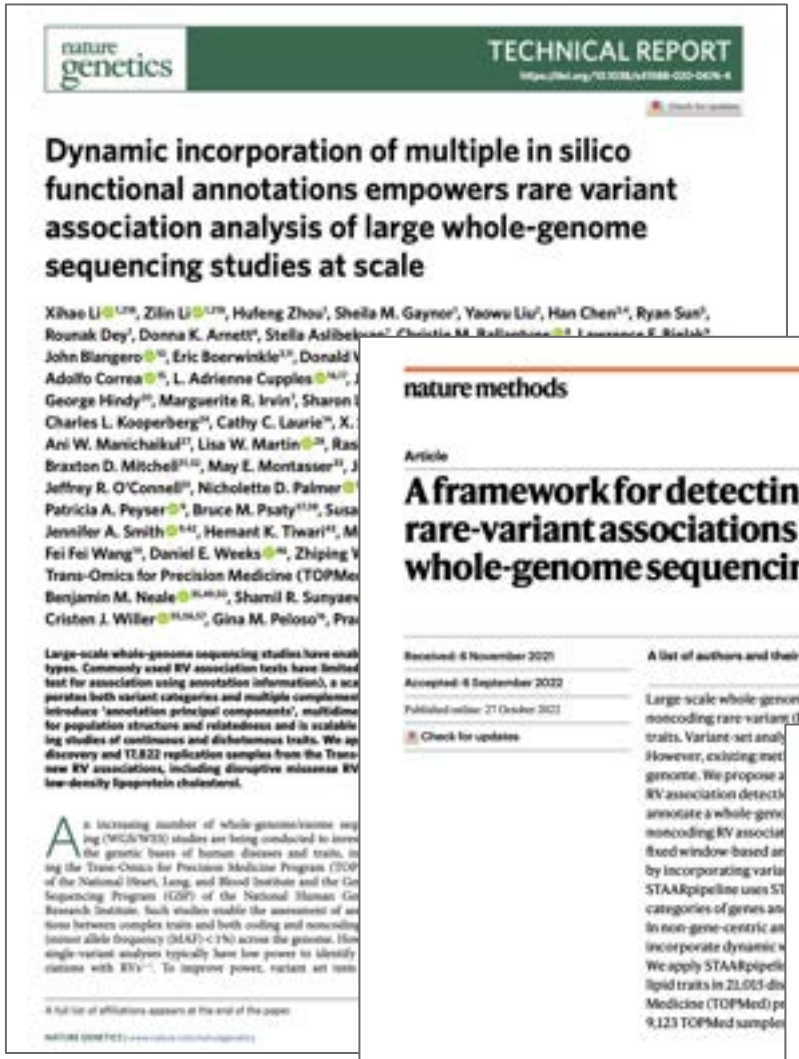
ACAT can be used to (1) combine SNV-level p-values for a SNV-set test and (2) combine p-values of different SNV-set tests using different weights.

Variant-set in silico analysis



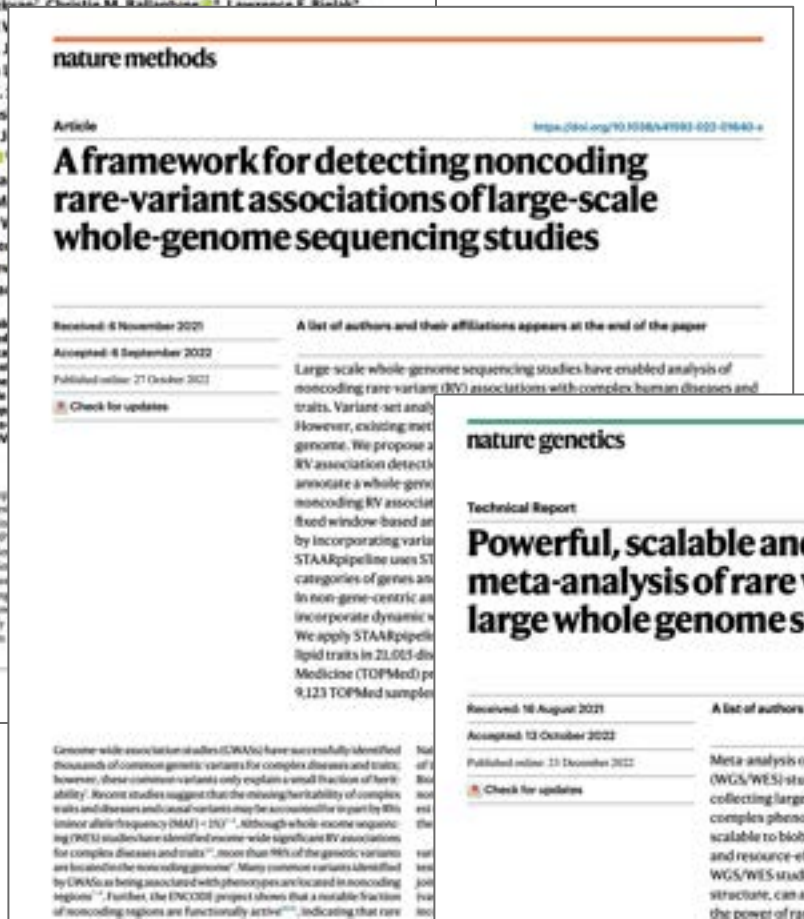
Dr Xihong Lin

Pathway analysis


nature genetics

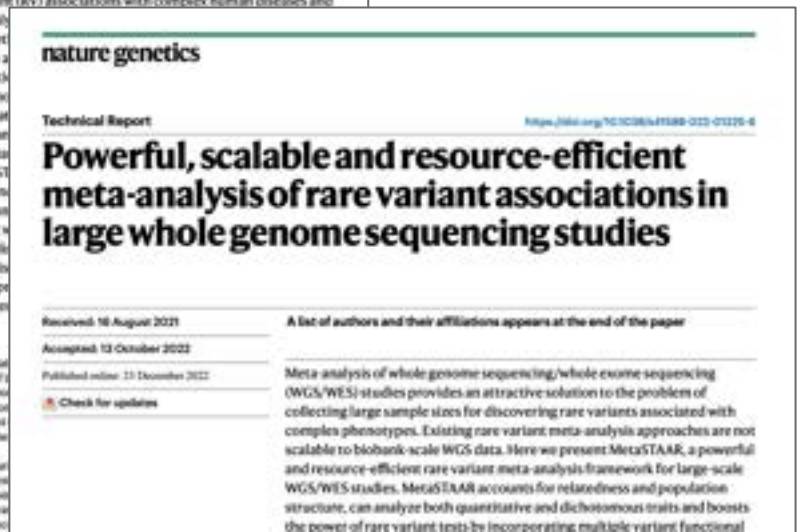
TECHNICAL REPORT

Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale


nature methods

Article

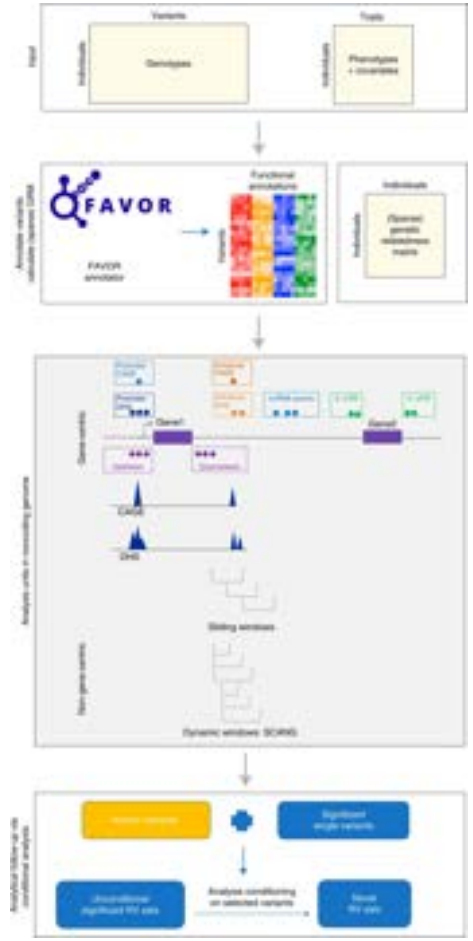
A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies


nature genetics

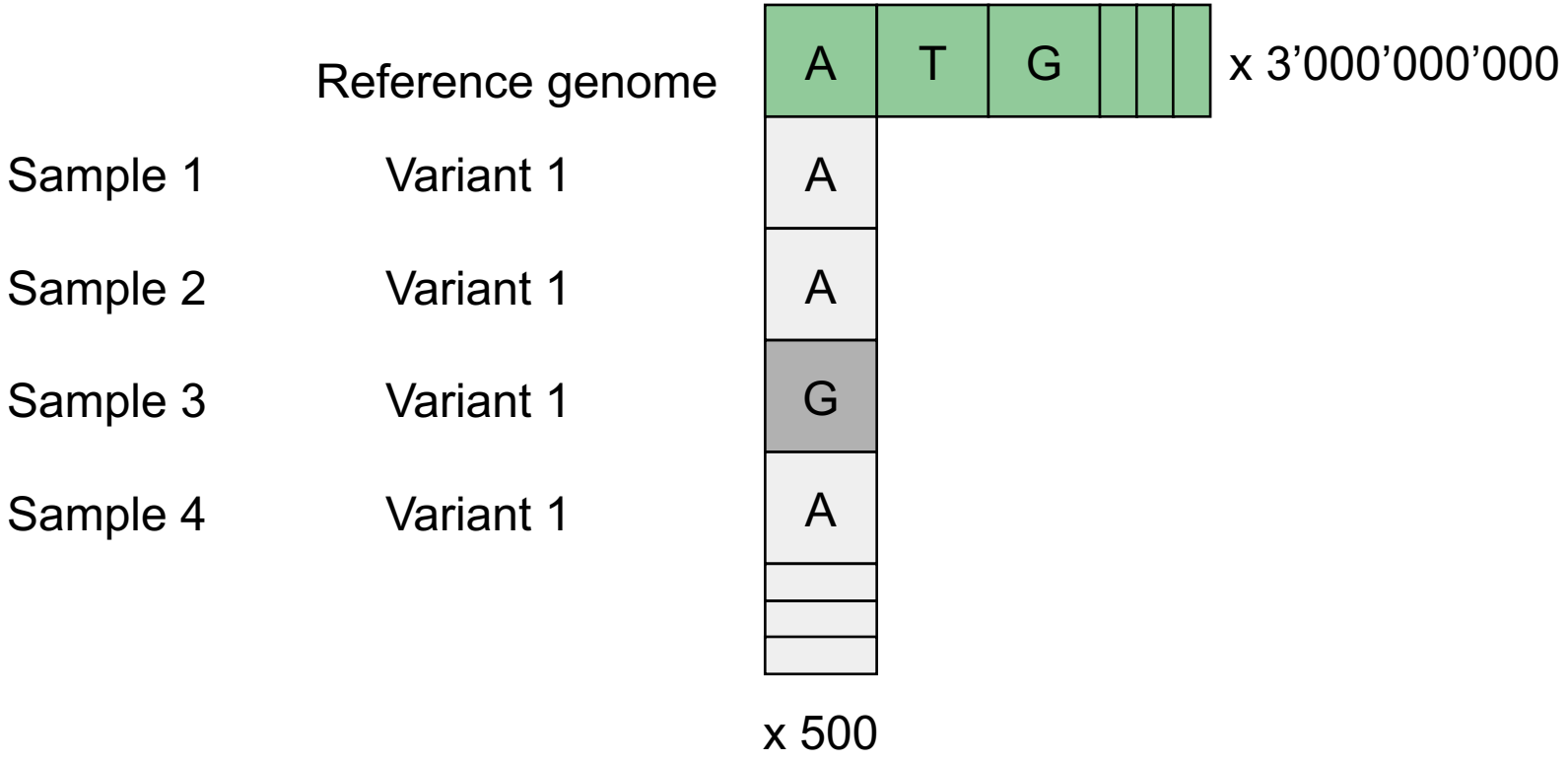
Technical Report

Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies

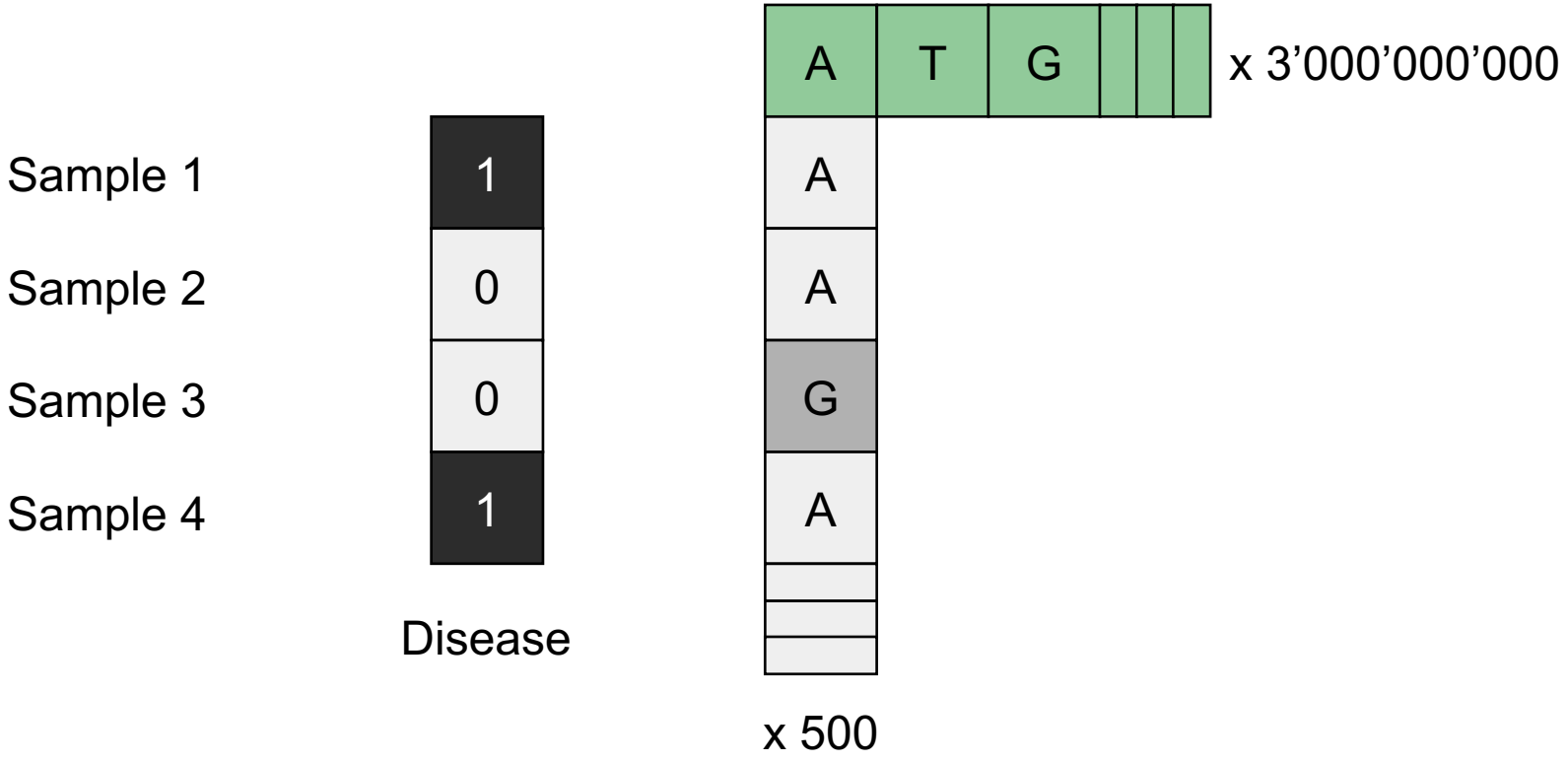
2020 STAAR applied
2022 STAAR framework
2023 MetaSTAAR



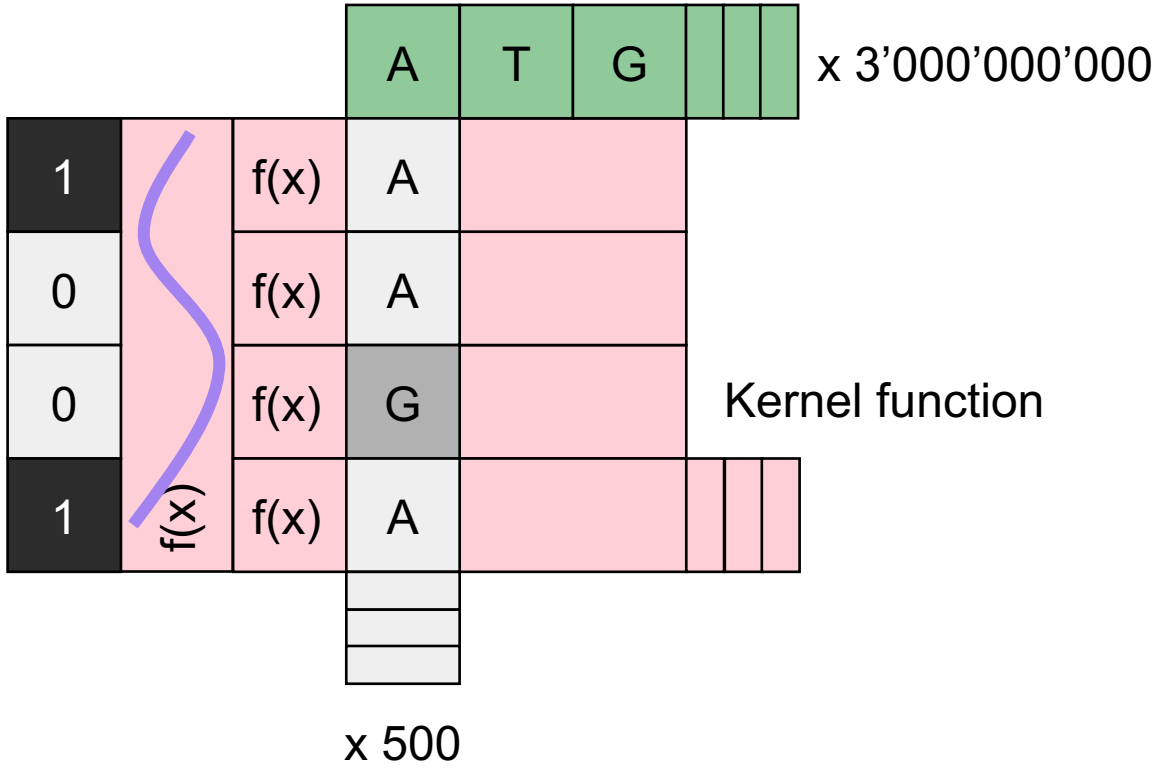
Model components



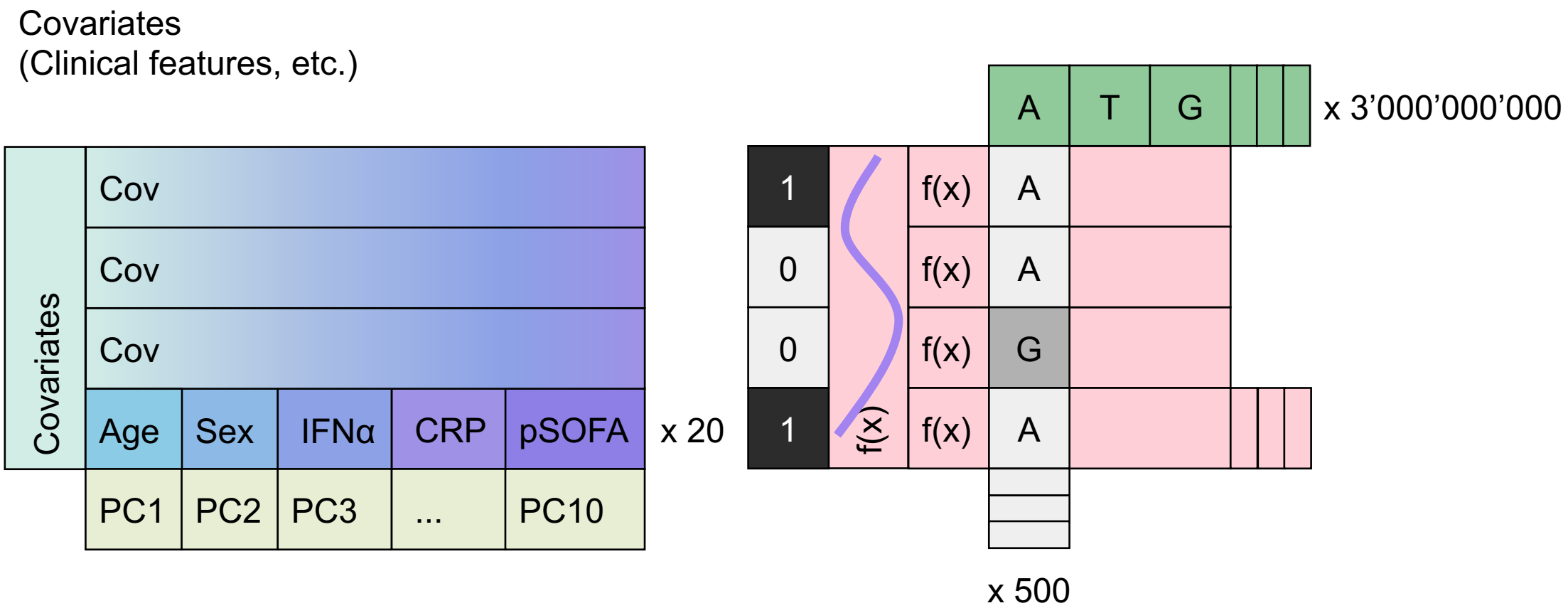
Model components



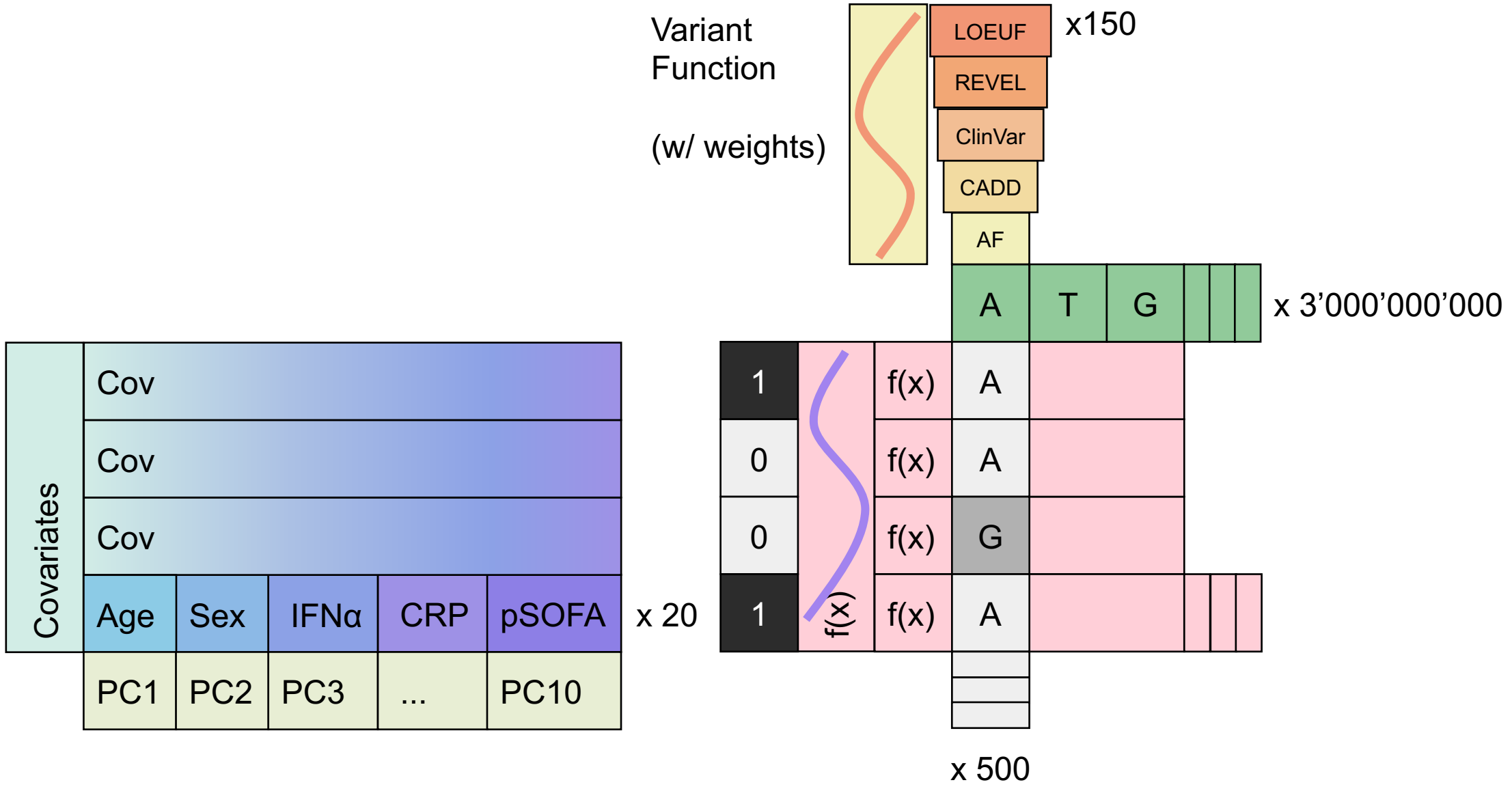
Model components



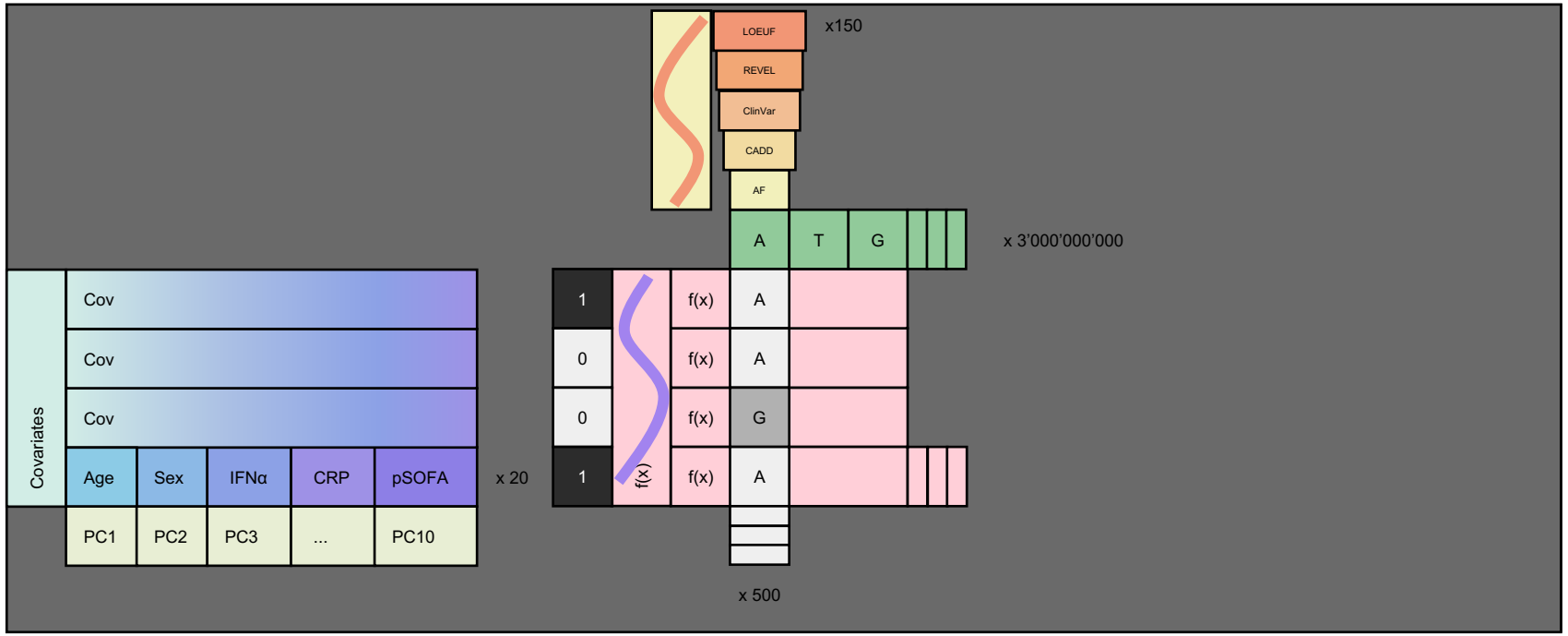
Model components

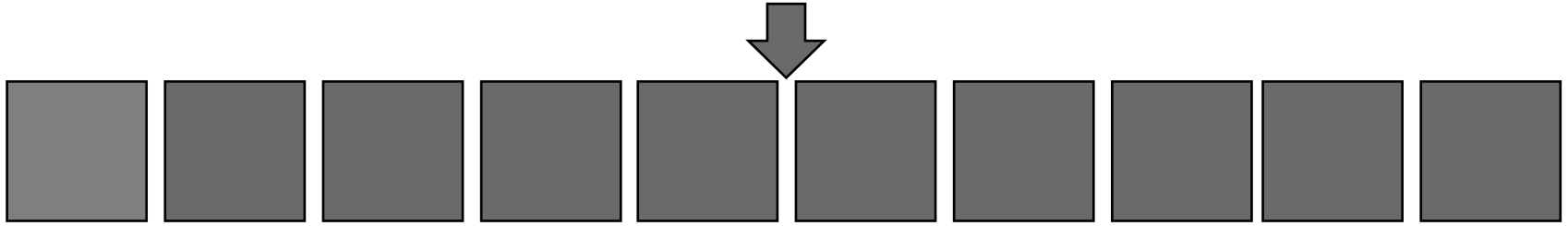


Model components



Model algorithm

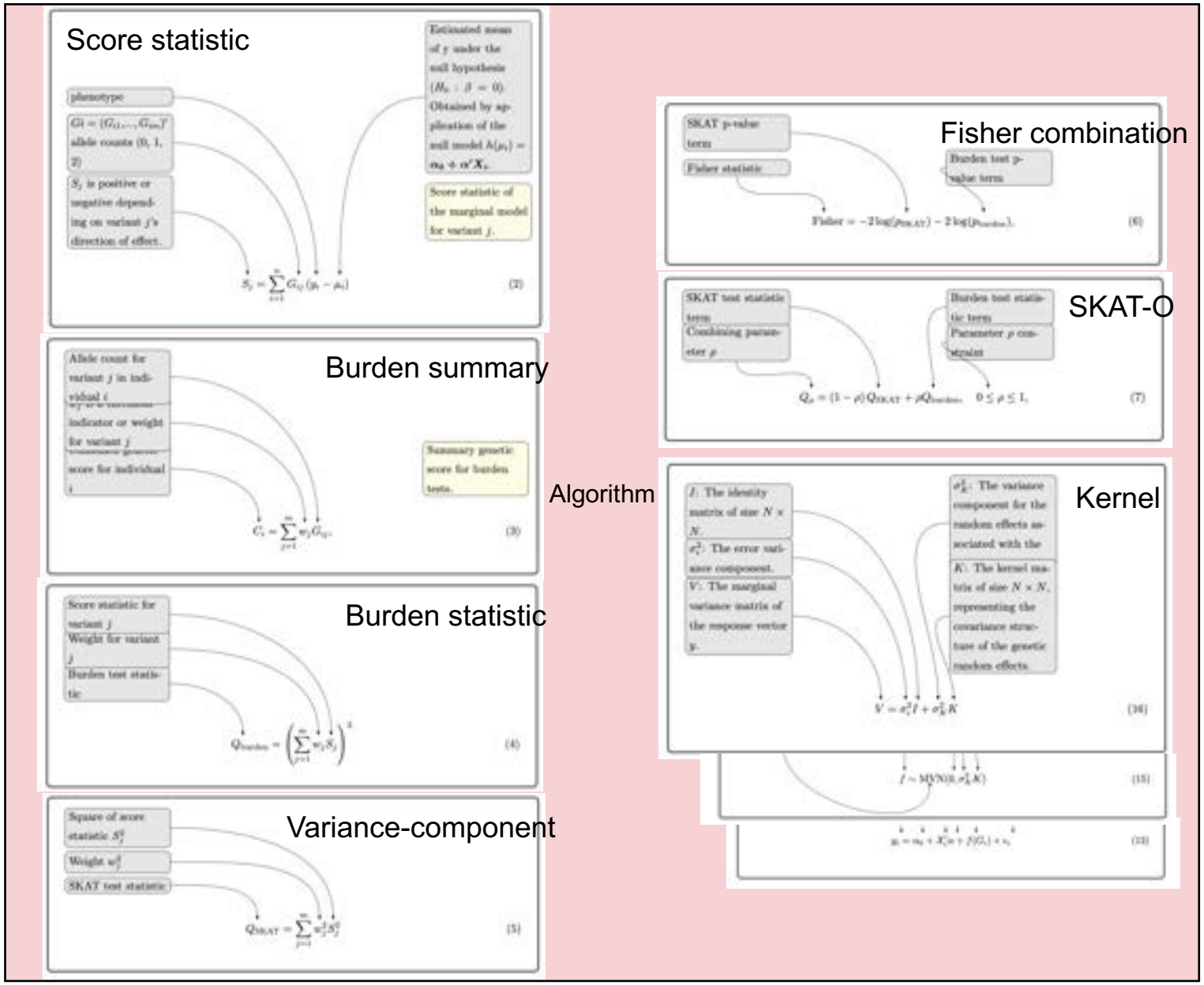




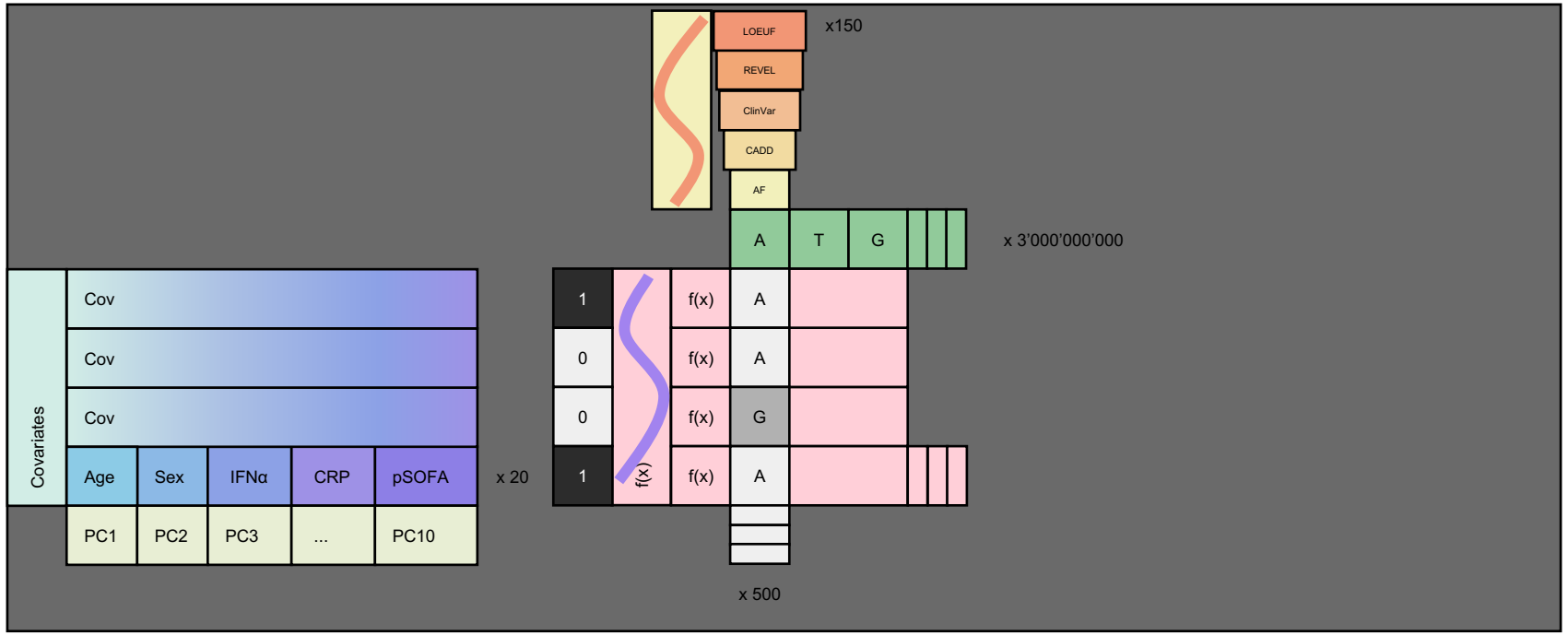
Chunk data

Model algorithm

Chunk data



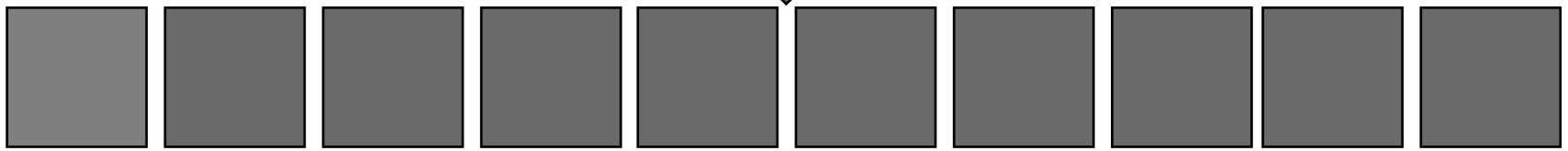
Model algorithm



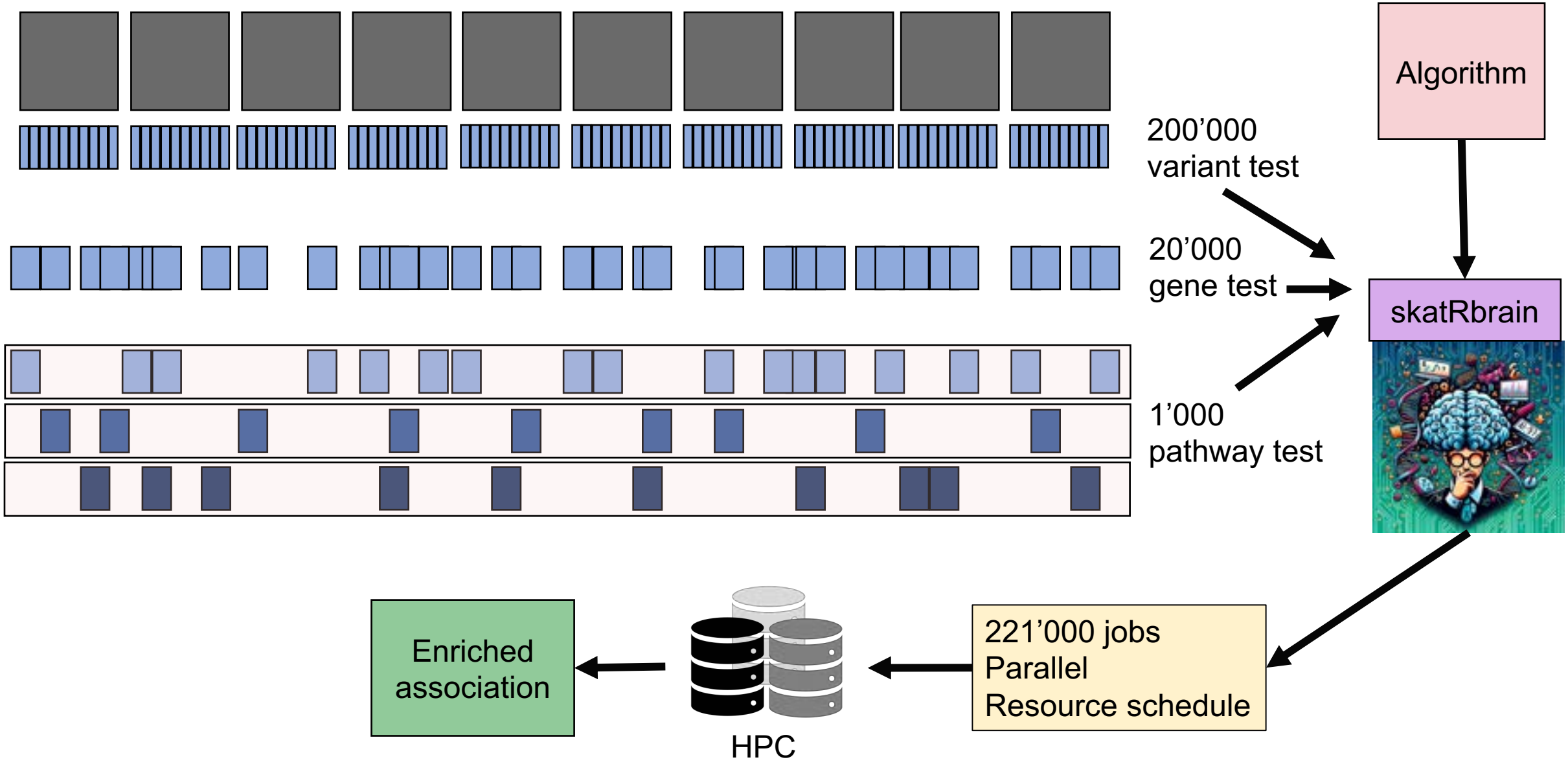
Algorithm

skatRbrain

Chunk data



Model application



Detour: Genetic association test

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

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	$X_{11}, X_{12}, ..., X_{1m}$	$G_{11}, G_{12}, ..., G_{1p}$
2	y_2	$X_{21}, X_{22}, ..., X_{2m}$	$G_{21}, G_{22}, ..., G_{2p}$
...	...	...	...
i	y_i	$X_{i1}, X_{i2}, ..., X_{im}$	$G_{i1}, G_{i2}, ..., G_{ip}$
...	...	...	...
n	y_n	$X_{n1}, X_{n2}, ..., X_{nm}$	$G_{n1}, G_{n2}, ..., G_{np}$

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

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

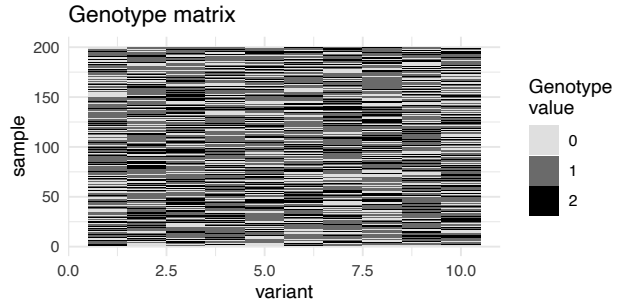
G Matrix			
0	1	2	...
2	0	1	...
1	1	0	...
0	2	2	...

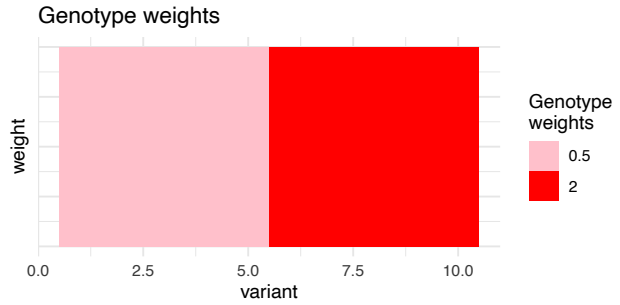
W Matrix			
1	0	0	...
0	0.5	0	...
0	0	2	...
0	0	0	...

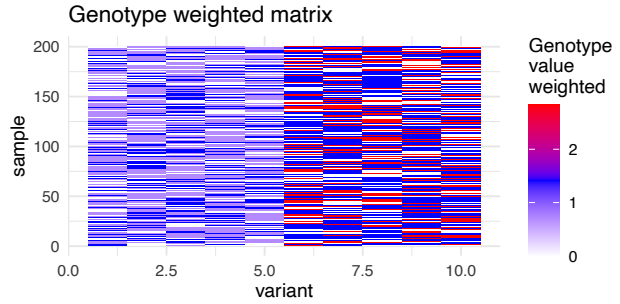
Detour: Genetic association test

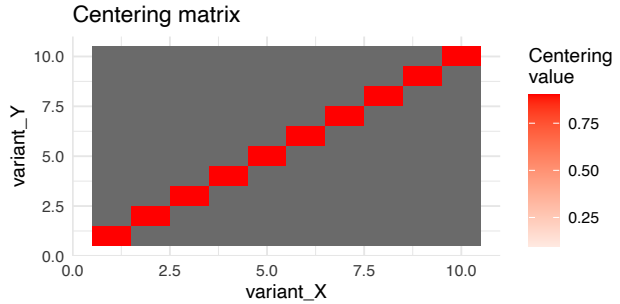
G Matrix			
0	1	2	...
2	0	1	...
1	1	0	...
0	2	2	...

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





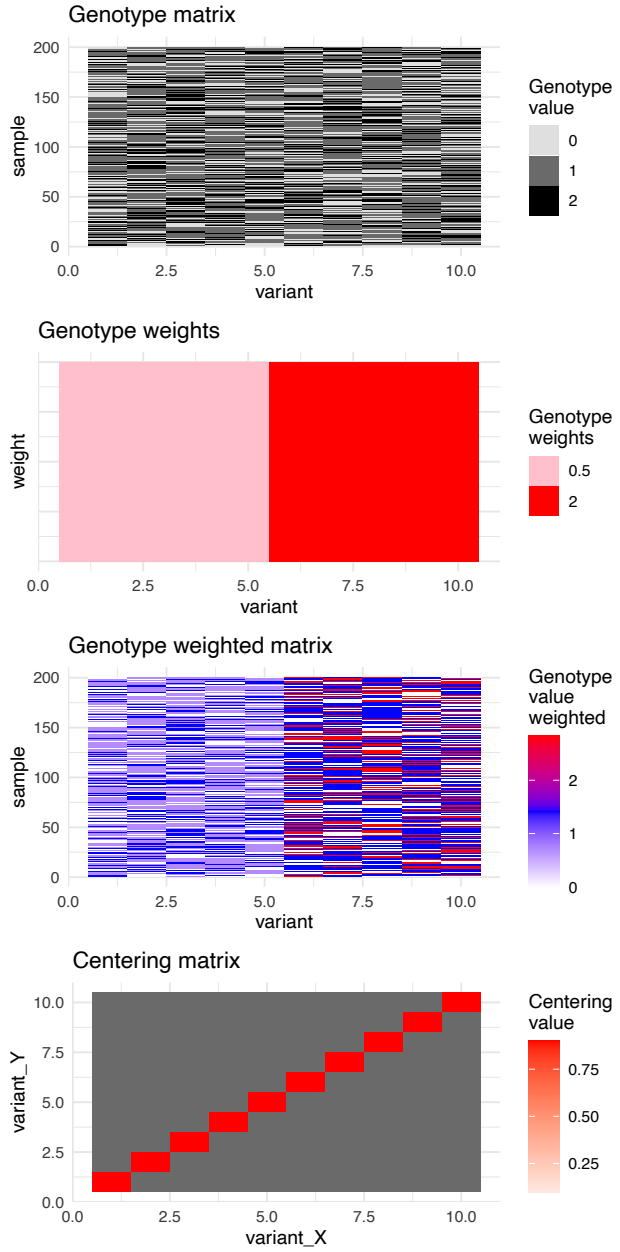


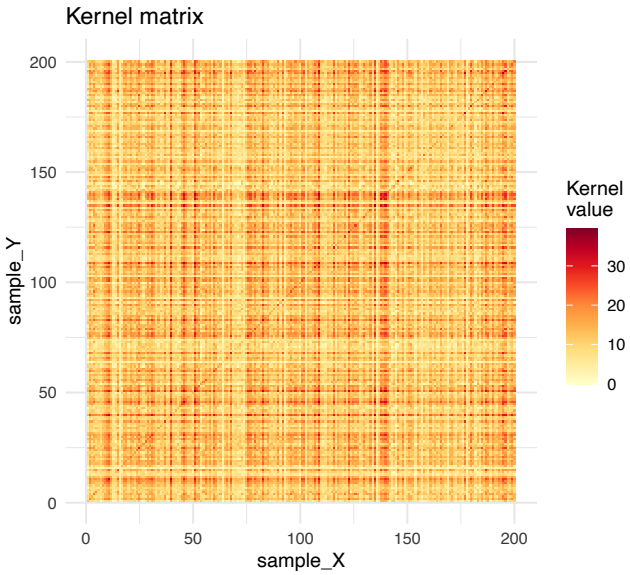


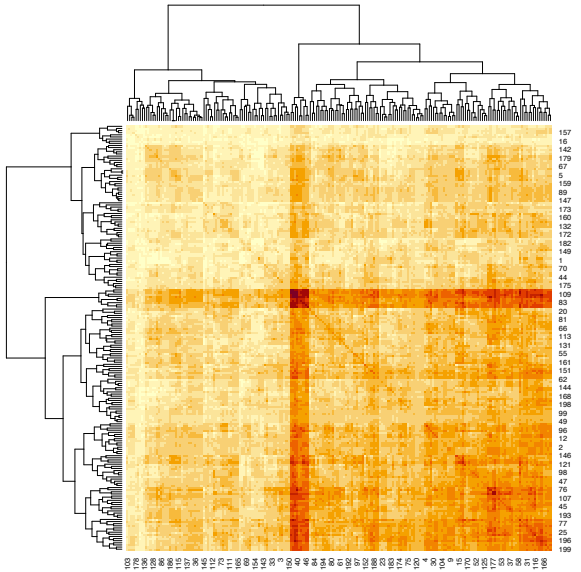
Detour: Genetic association test

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

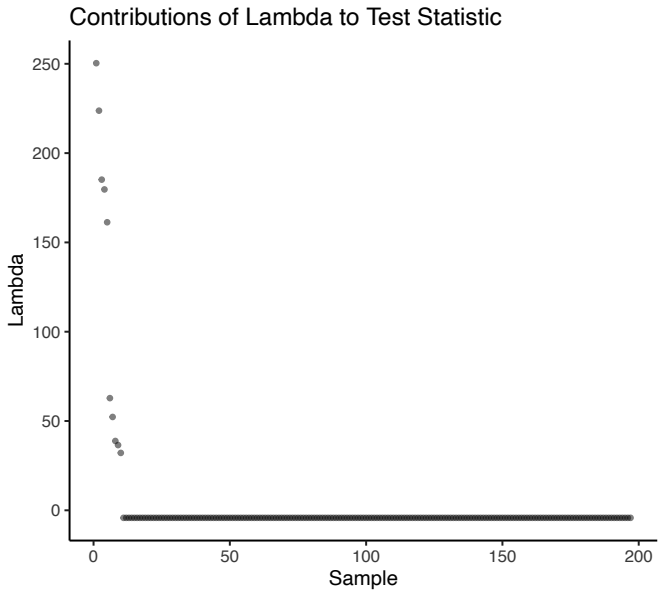
W Matrix			
1	0	0	...
0	0.5	0	...
0	0	2	...
0	0	0	...

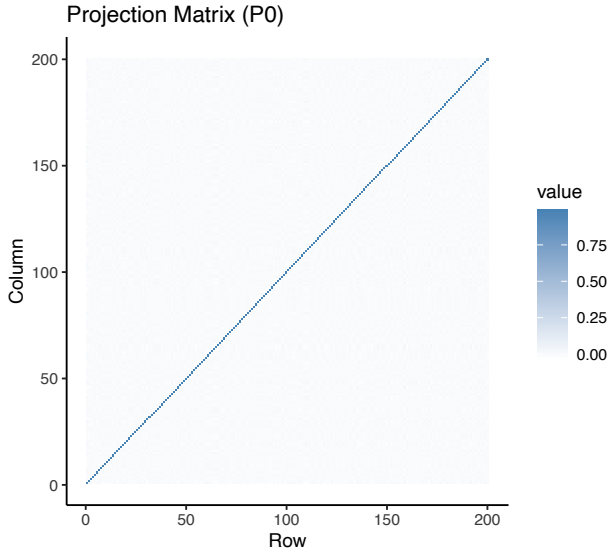


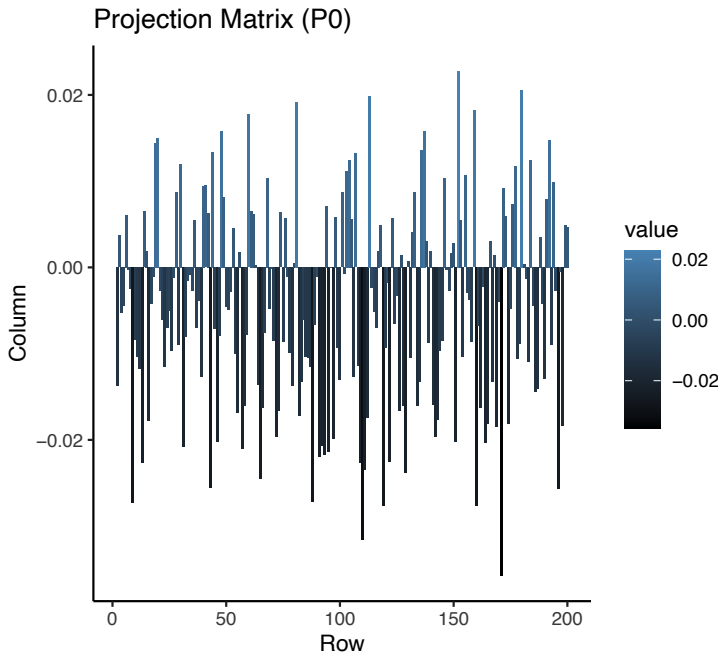


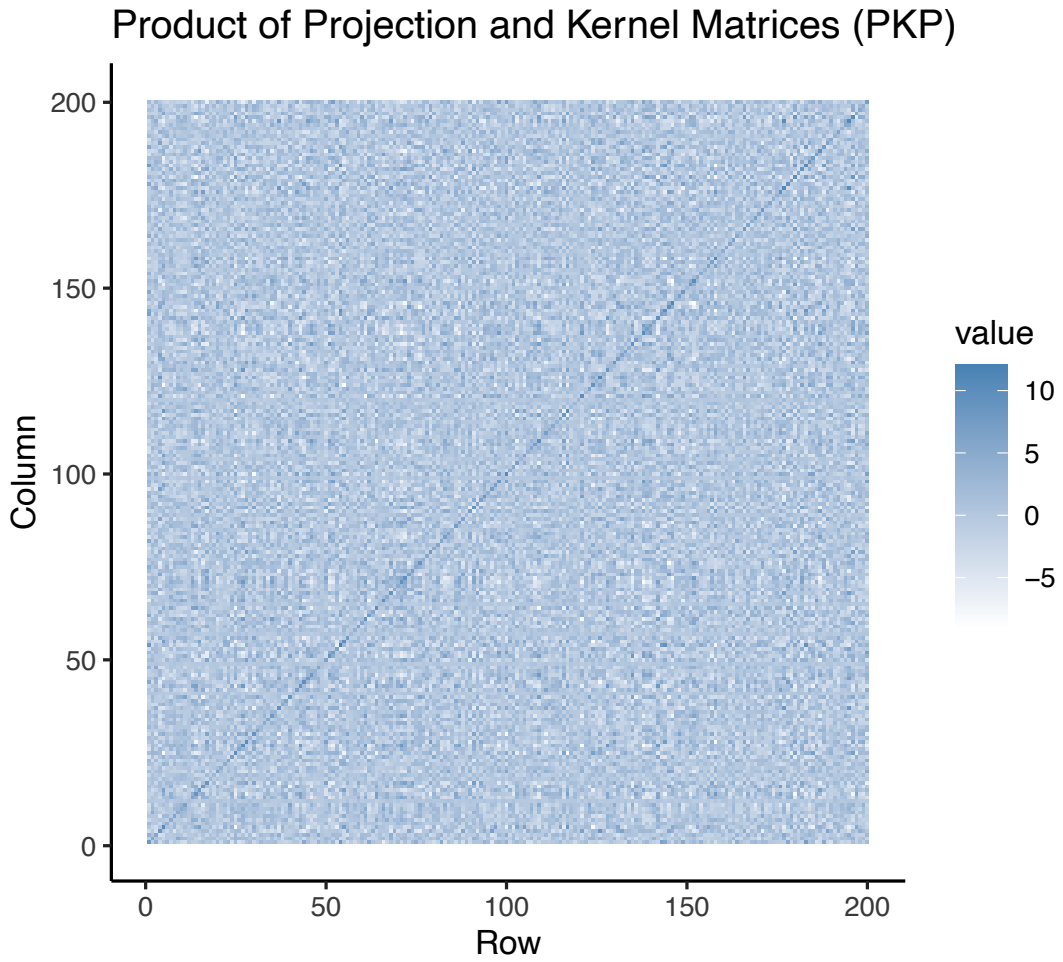


Detour: Genetic association test

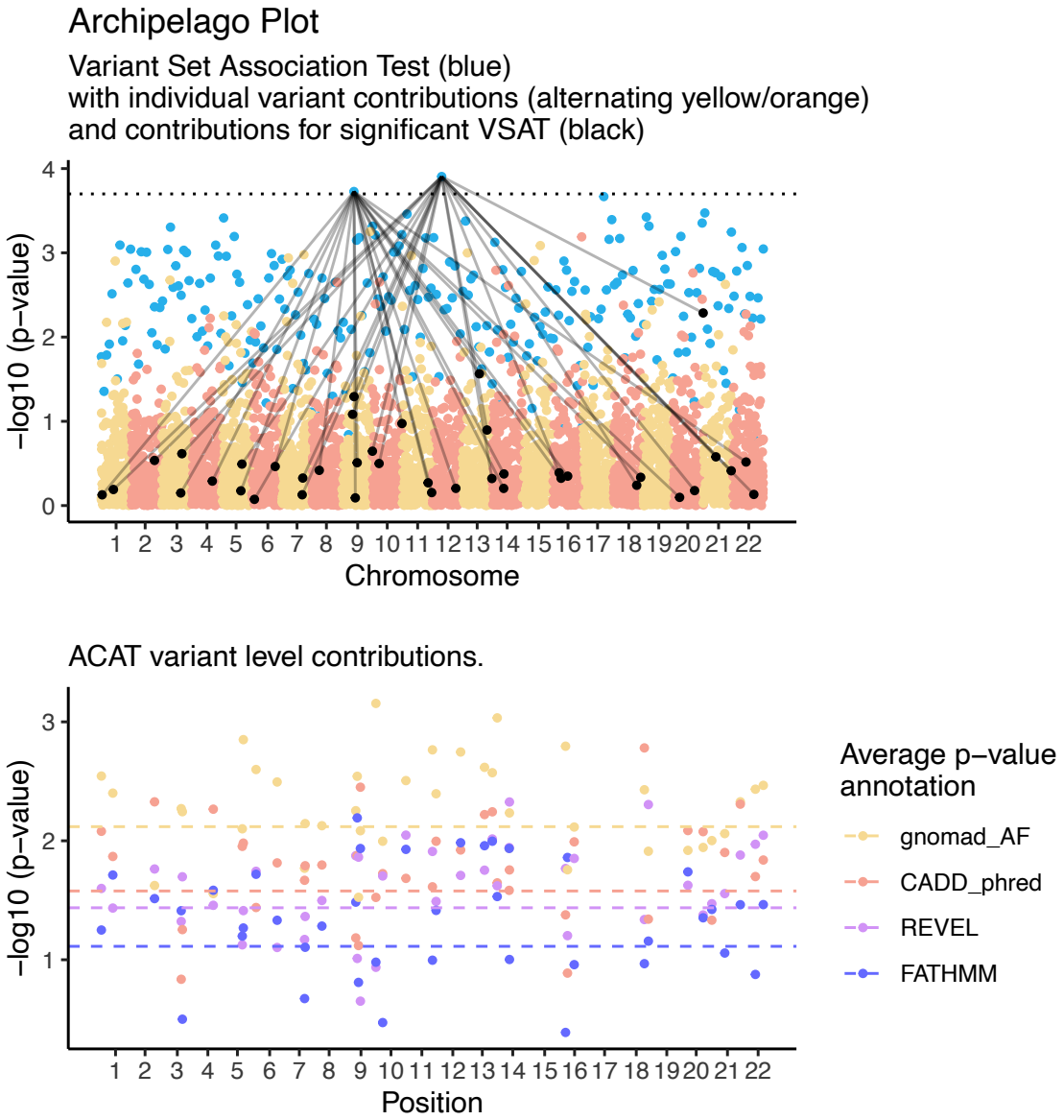








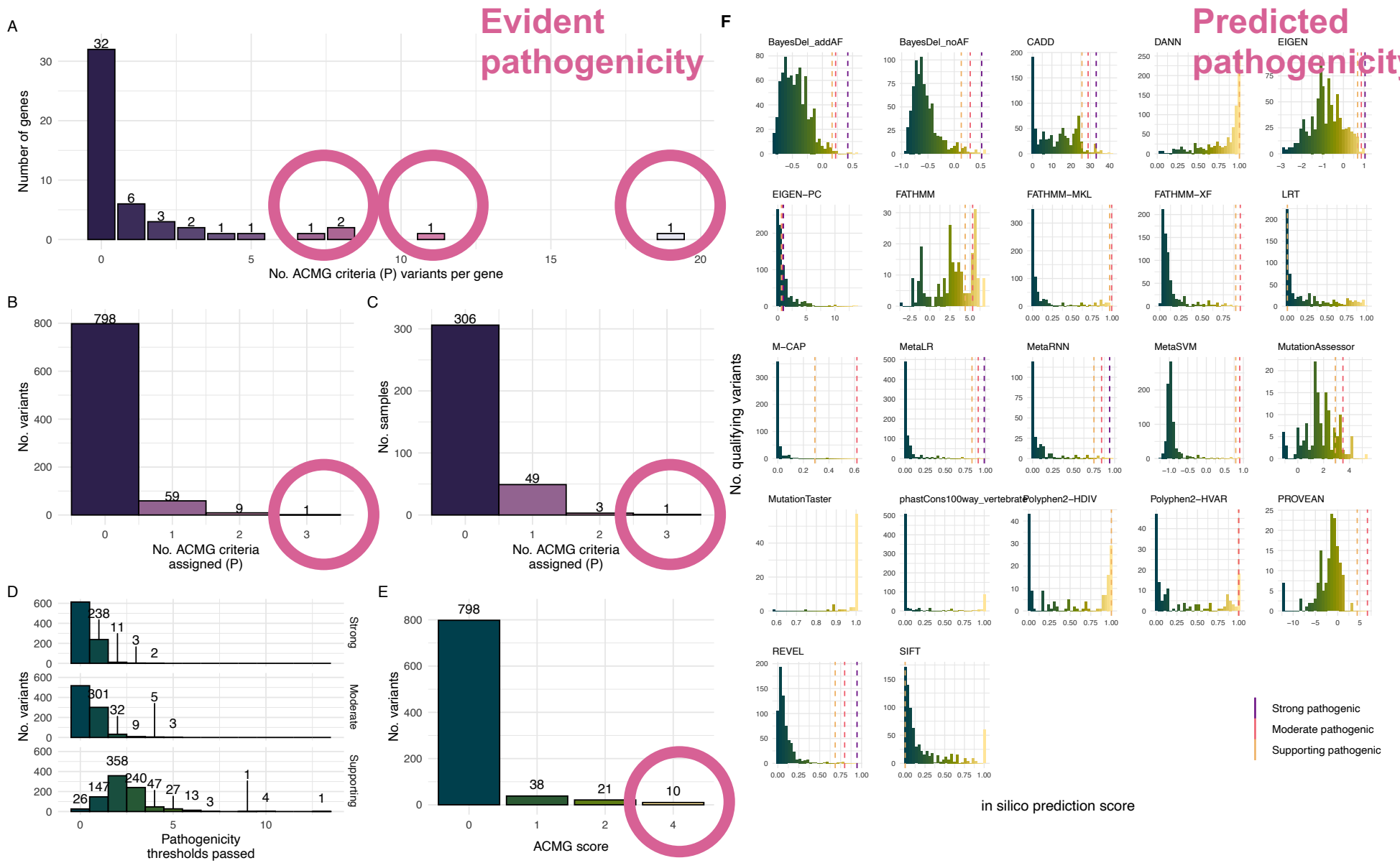
Detour: Genetic association test



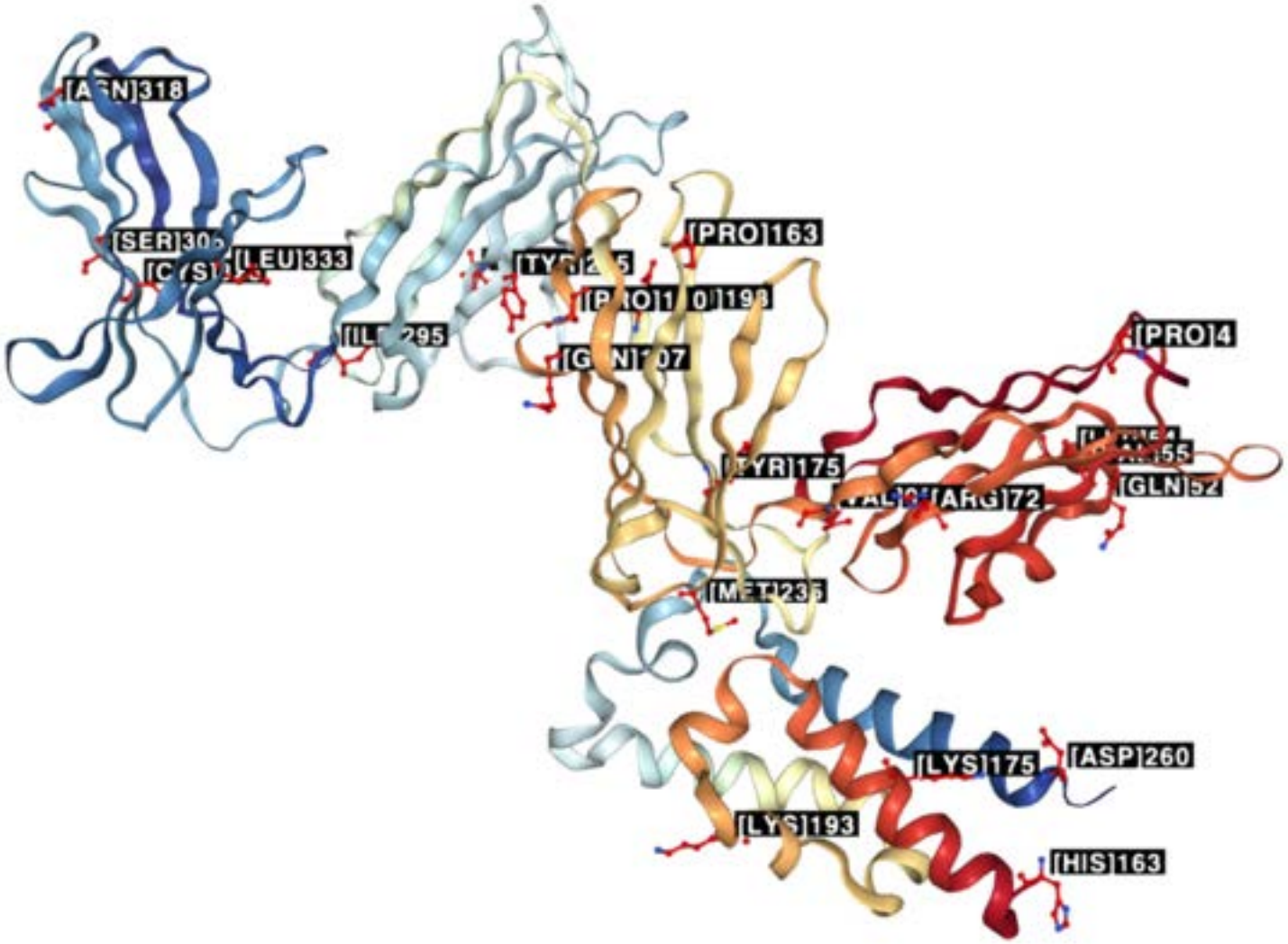
Detour #2: Variant interpretation

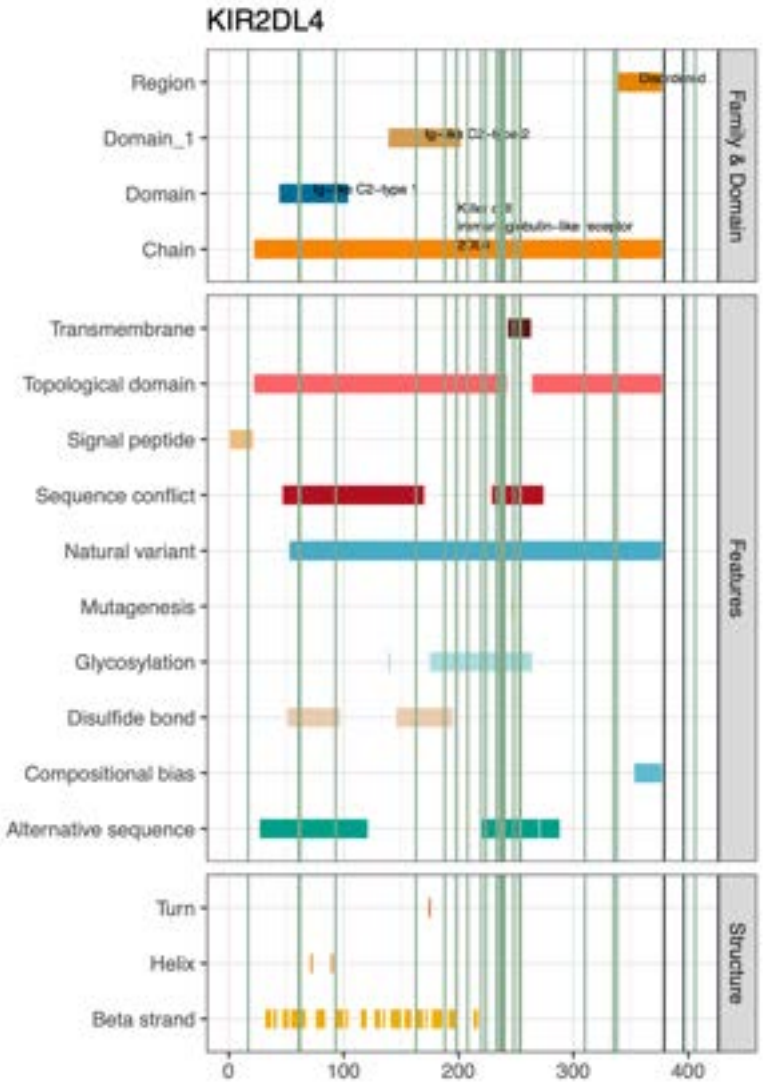
- Joint consensus recommendation on the standards and guidelines for the interpretation of sequence variants (Richards 2015).
- Standards and guidelines for the interpretation and reporting of sequence variants in cancer (Li 2017).
- Technical standards for the interpretation and reporting of constitutional 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

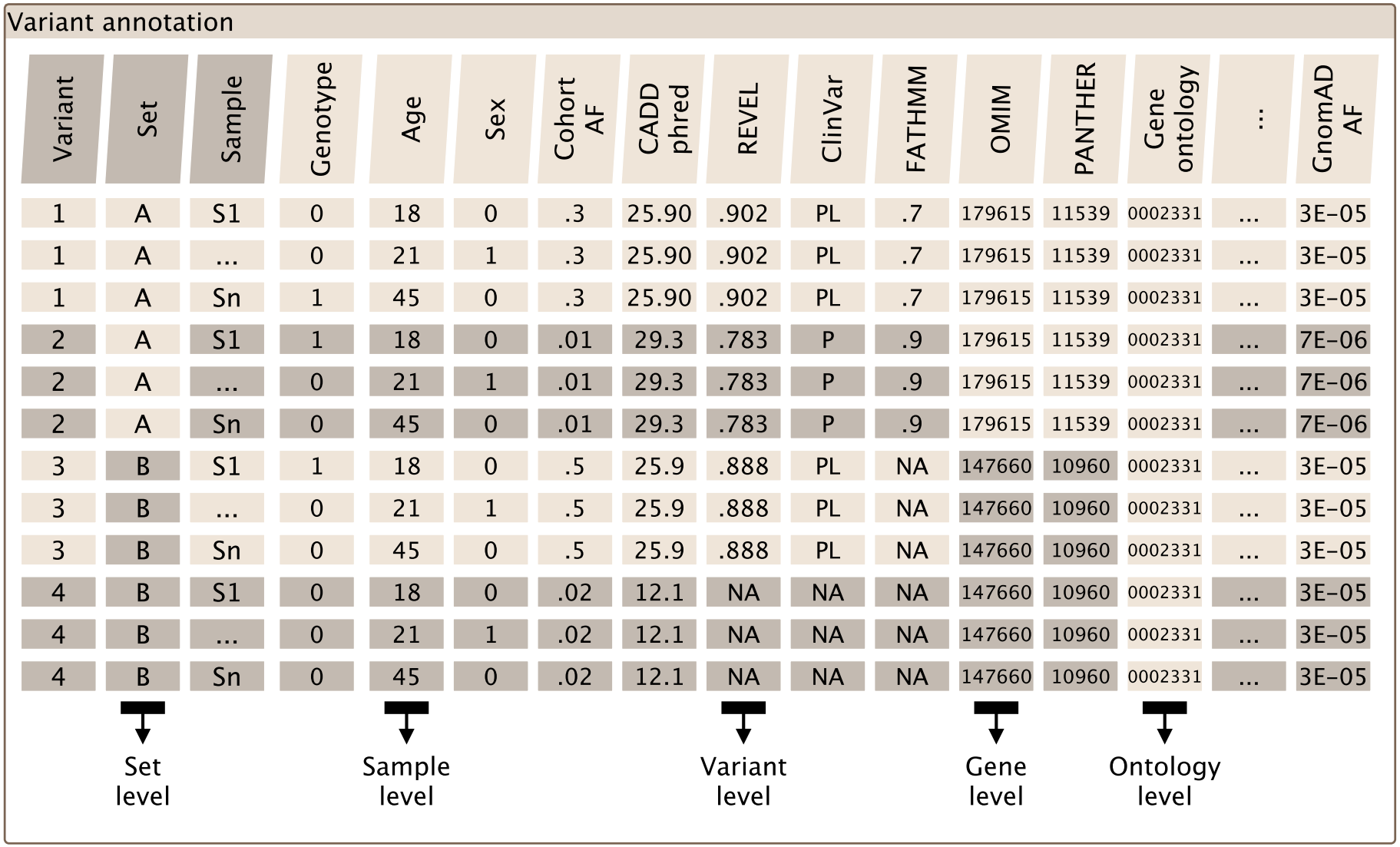


Detour #2: Variant interpretation





Detour #2: Variant interpretation



Detour #2: Combined

ACMGuru

skatRbrain

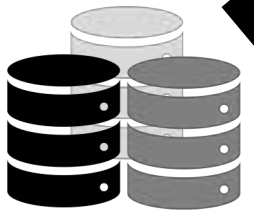
Variant annotation

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

HPC

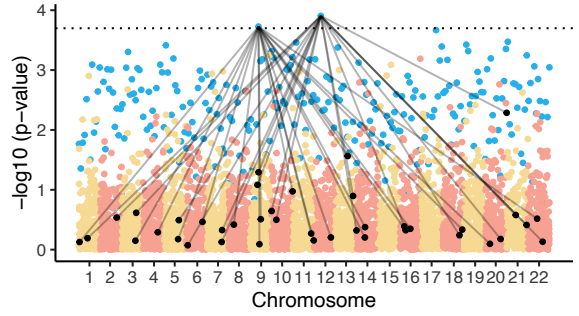


Test: Single variant 200’000

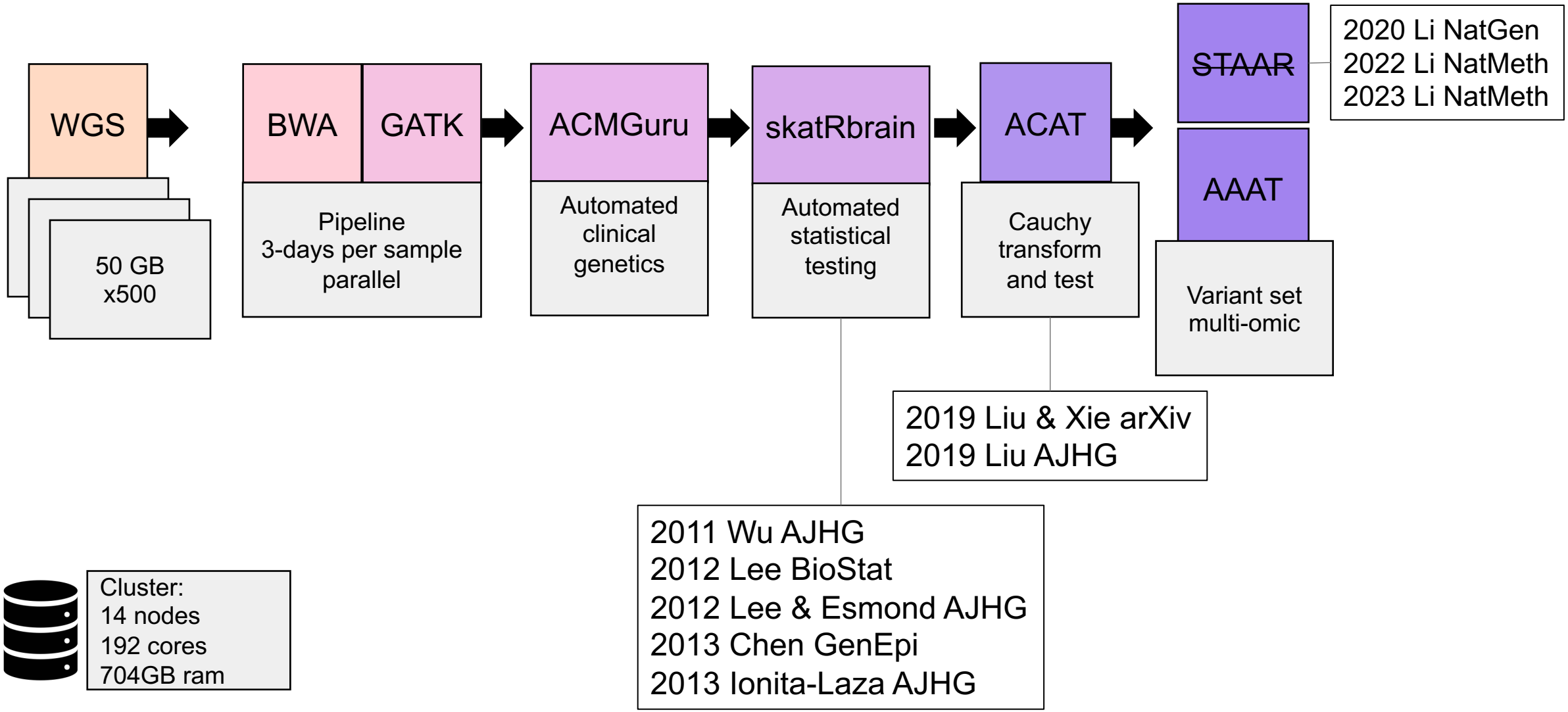
Test: Single genes 20’000

Test: protein pathways 1’000

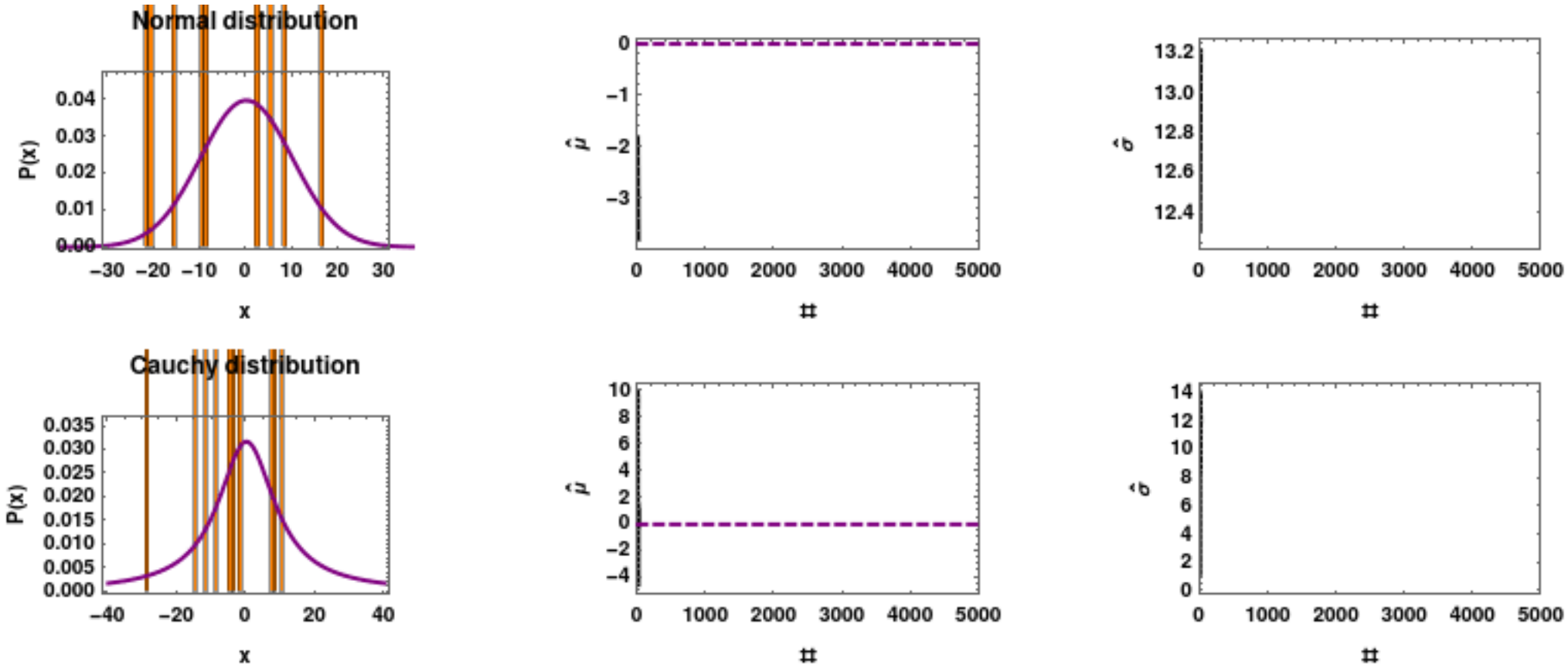
Result 1
Result 2
Result 3
Result ...
Result 219'913



Analysis chain



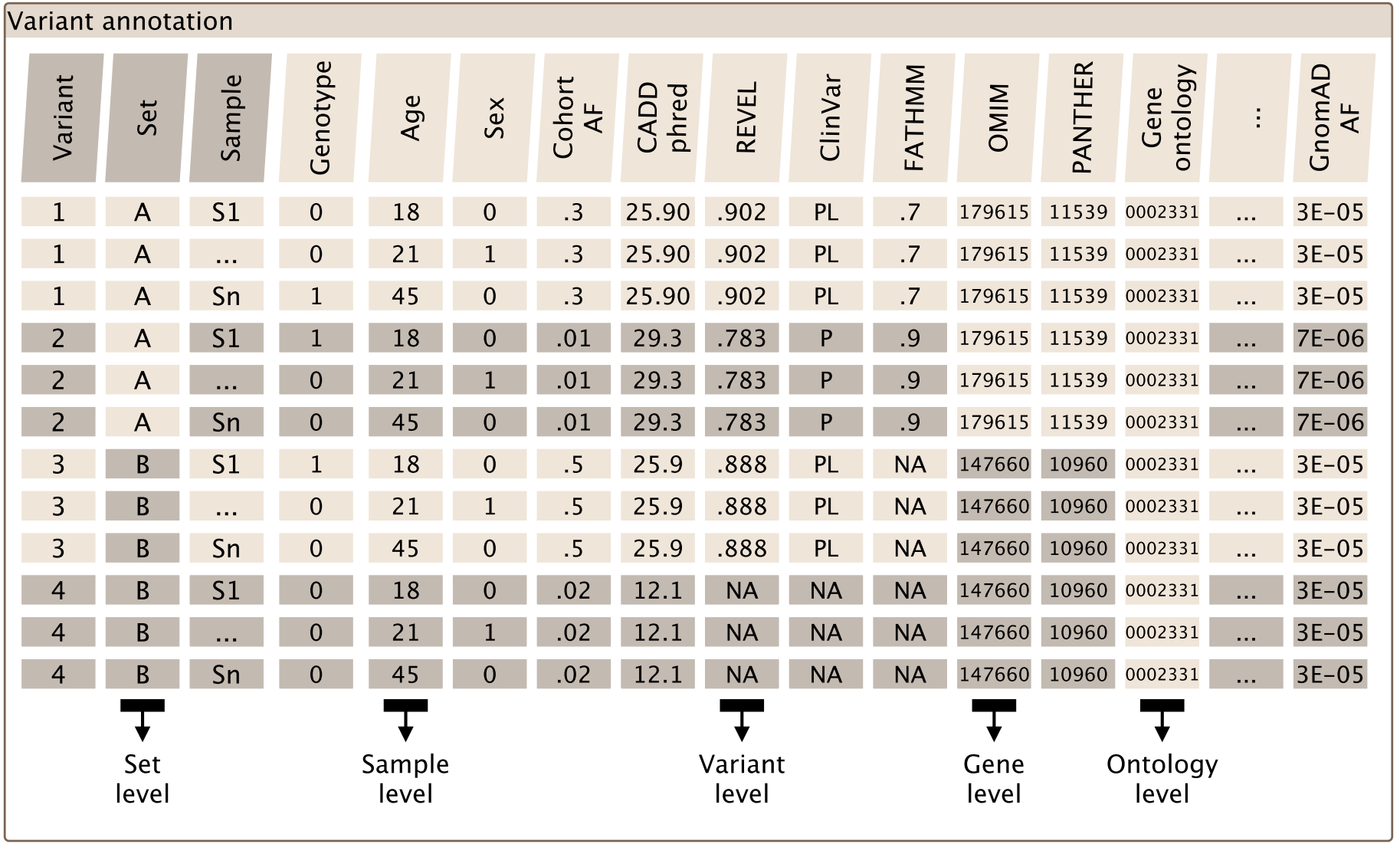
Cauchy distribution in variant analysis



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



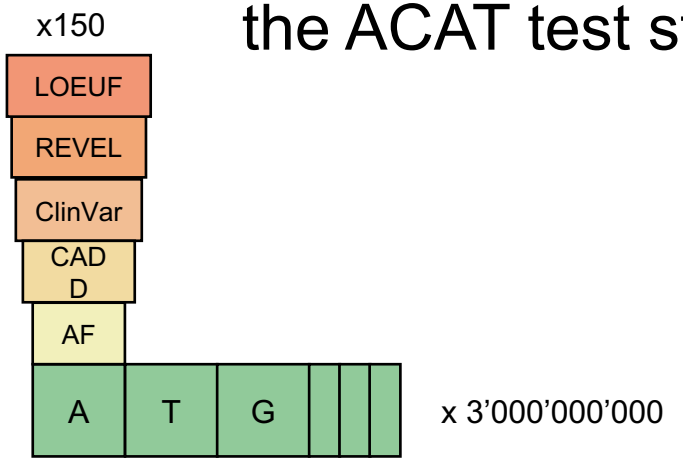
ACAT main steps

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

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

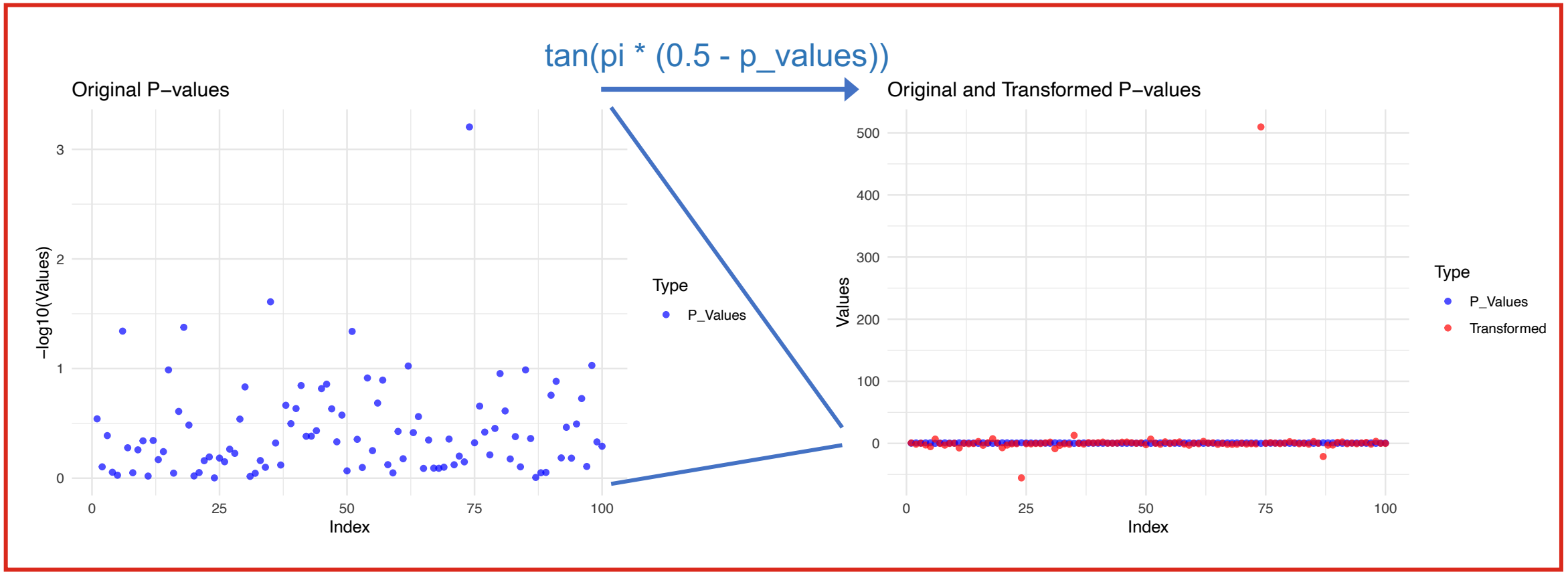
- P-value calculation for the ACAT test statistic: $$p\text{-value} \approx 1 - \frac{1}{2} + \frac{\arctan\left(T_{ACAT}/w\right)}{\pi}$$

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



lawlessgenomics.com/topic/acat

Pathway analysis



x150

LOEUF

REVEL

ClinVar

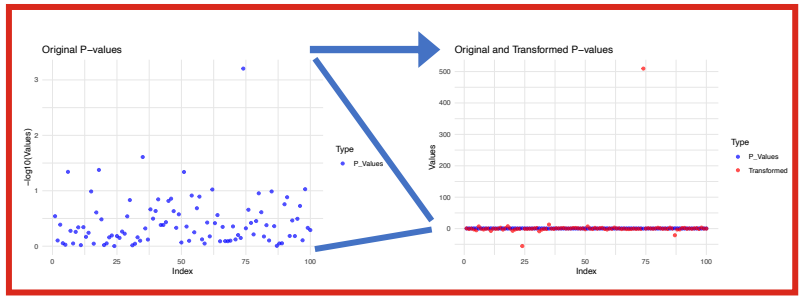
CAD D

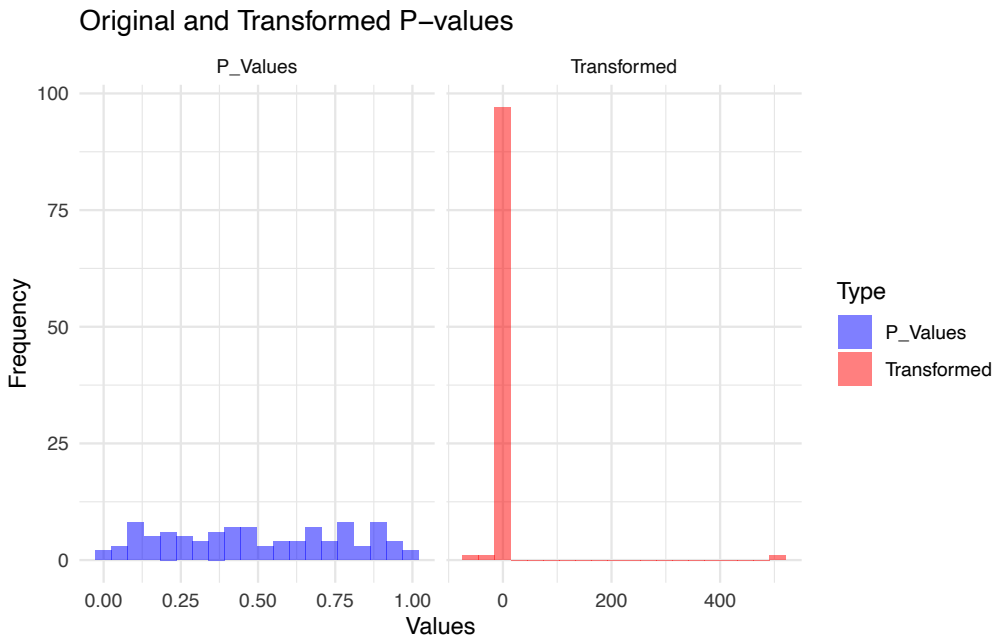
AF

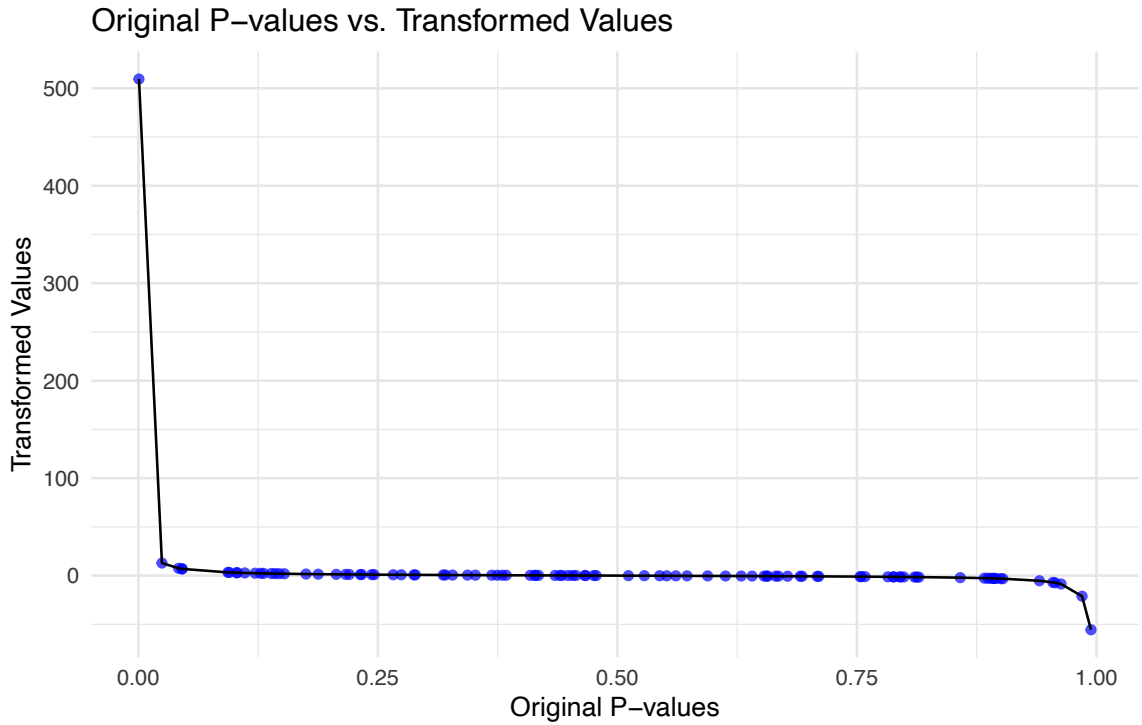
A T G

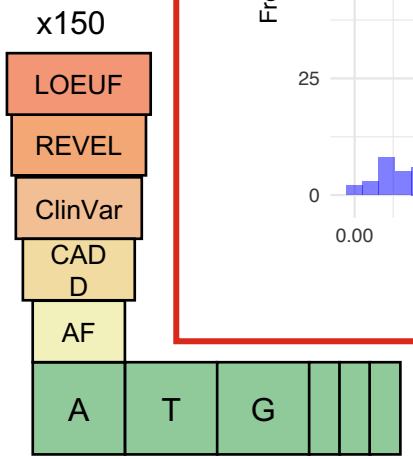
x 3’000’000’000

Pathway analysis



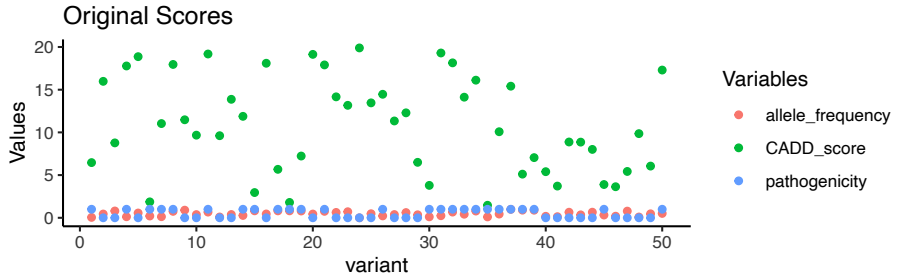


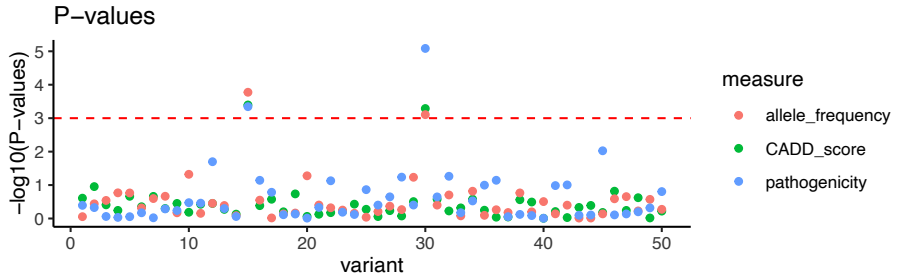


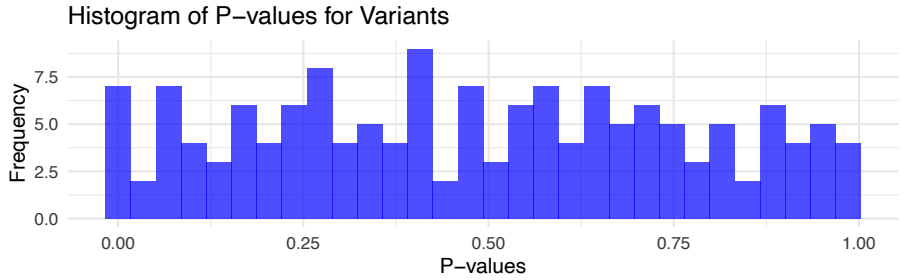


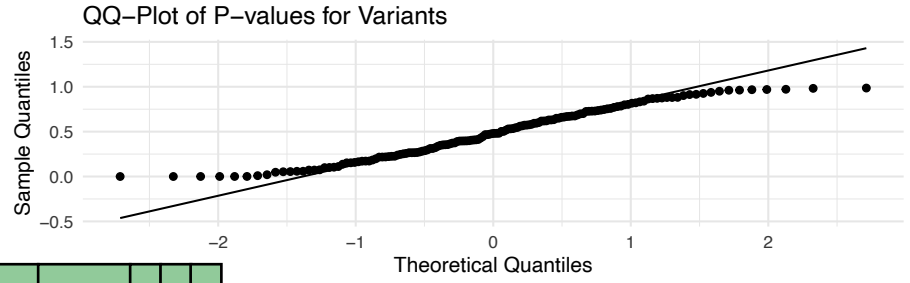
x 3’000’000’000

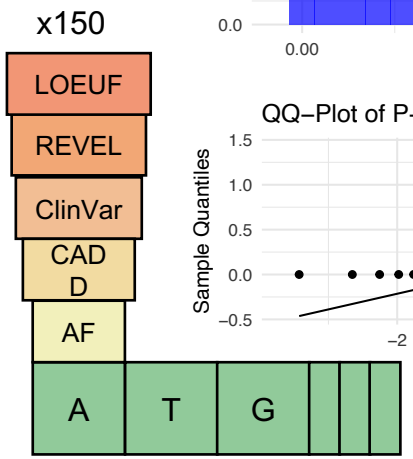
ACAT for multi-variable association



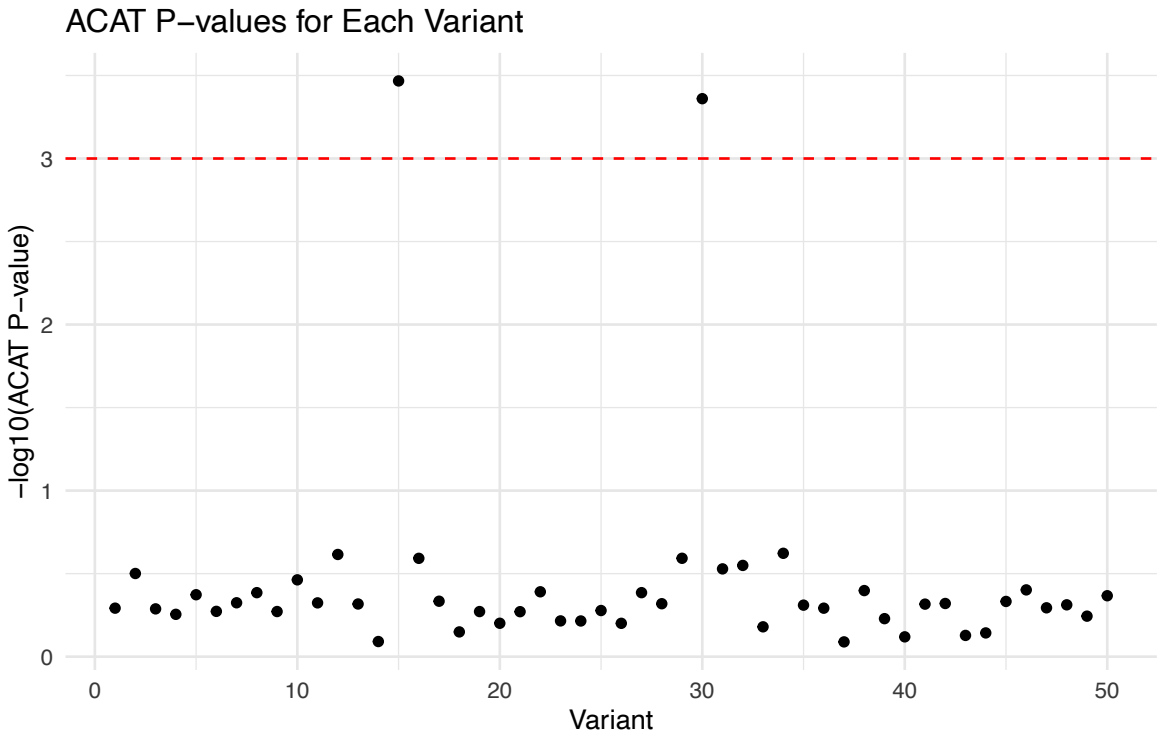






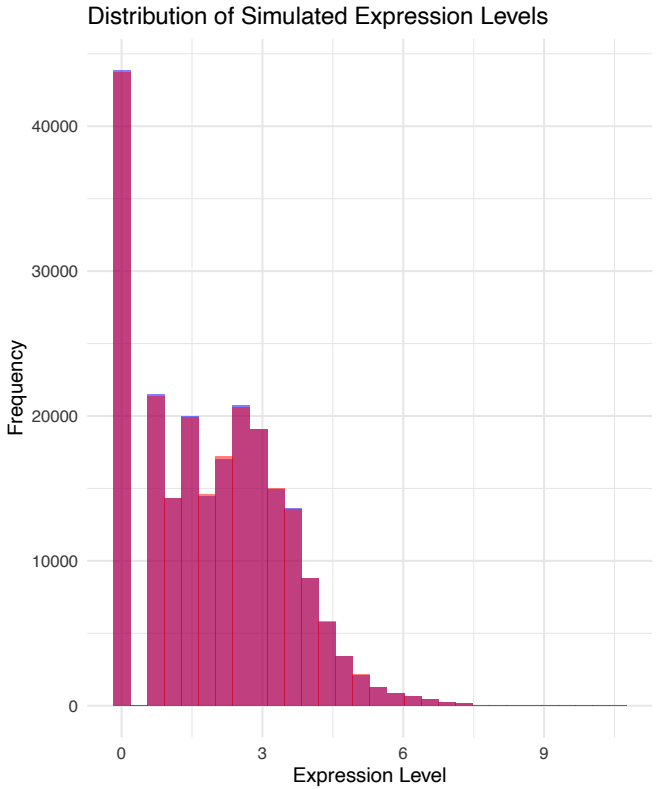


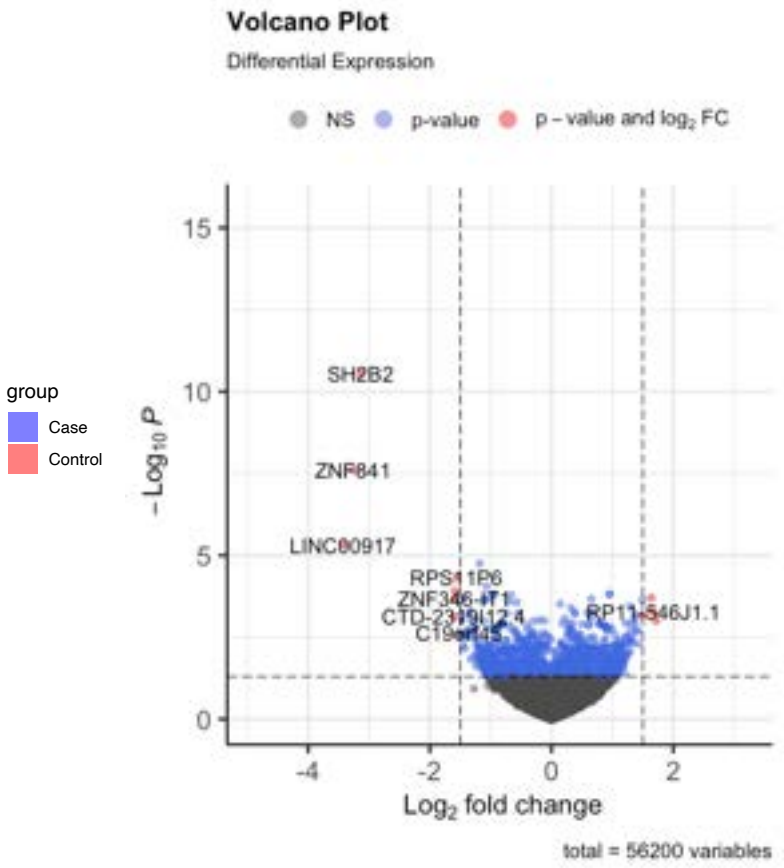
x 3’000’000’000

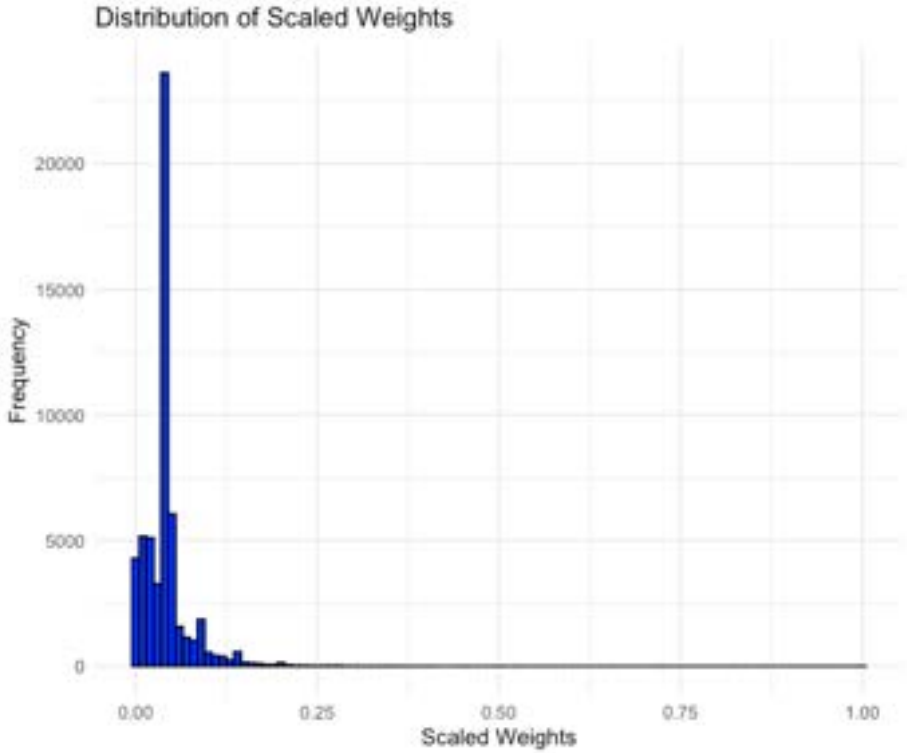


Tests

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

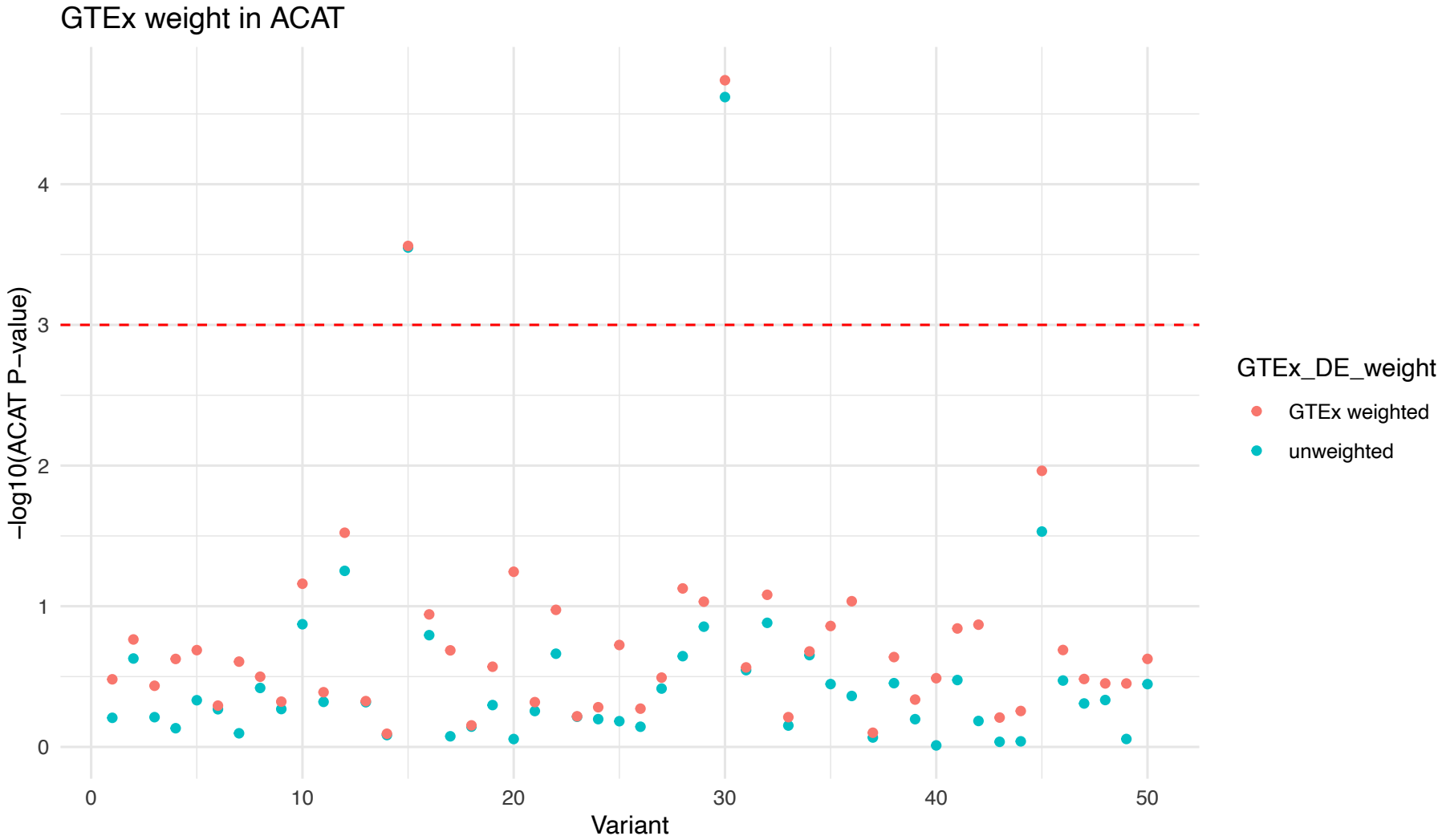




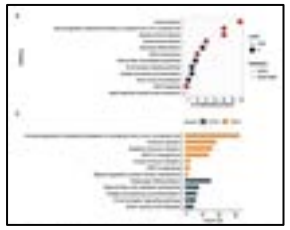


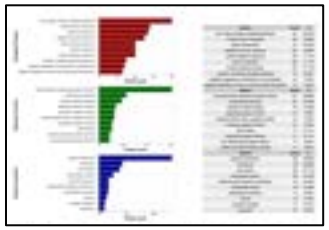
Tests

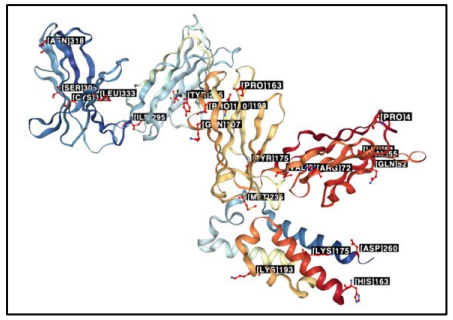
- **RNAseq** GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct

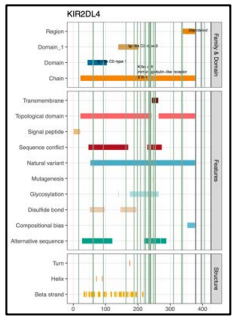


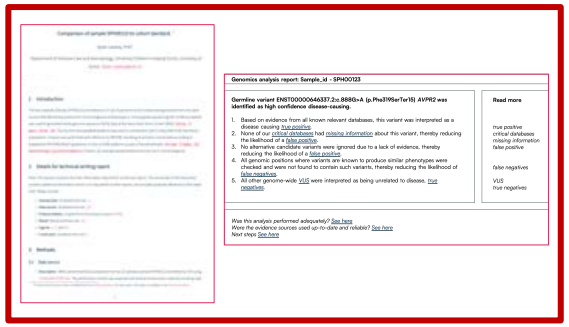
Final result

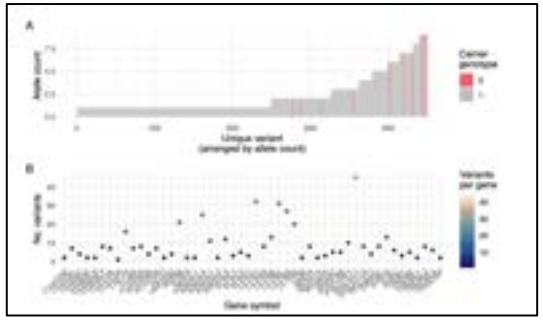


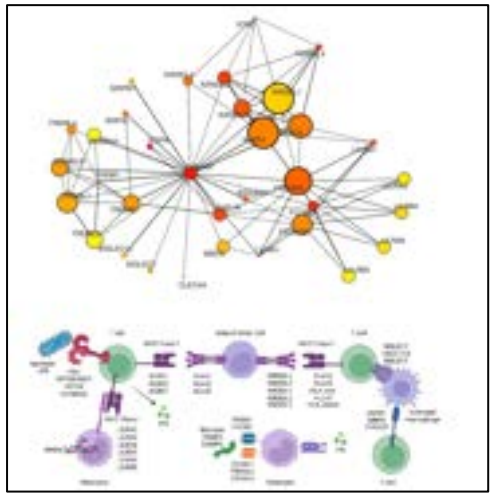






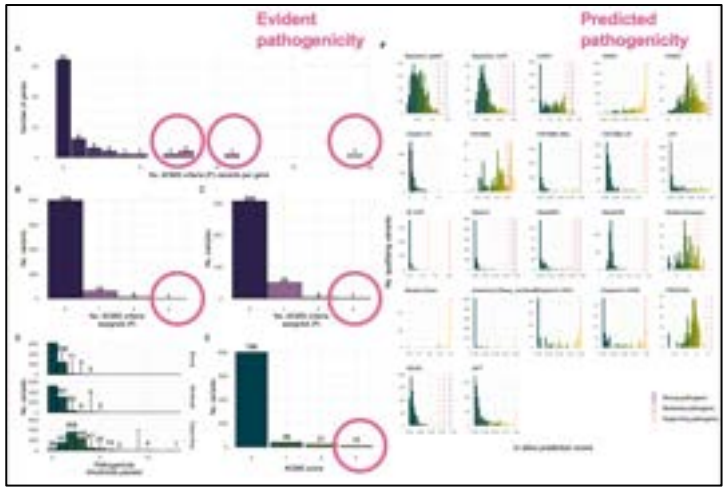






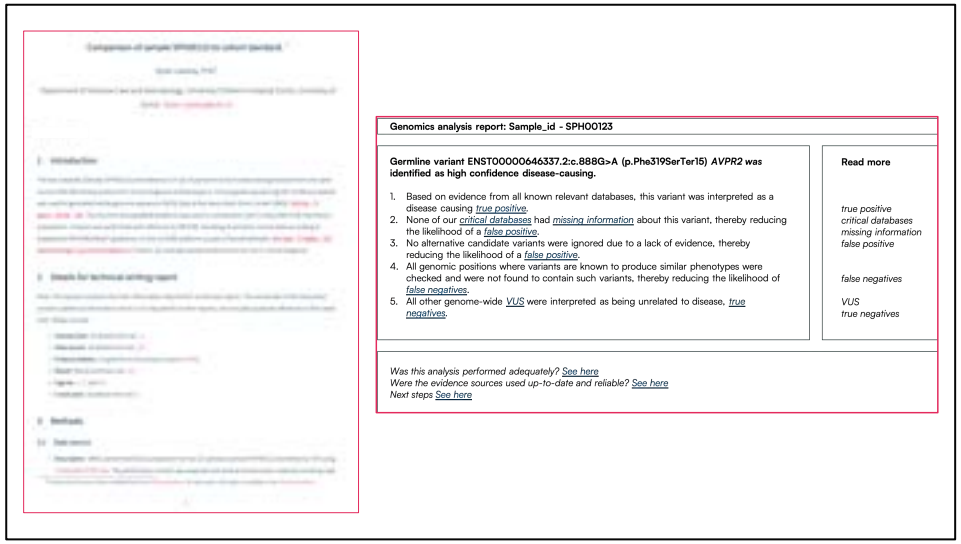
Gene function
KIR2DL4, KIR3DL3, KLRD1: Regulate NK cell activity, central to innate immune response.
LILRA1, LILRB1, LILRB4: MCH class I
TREML1,2,4: Pathogen recognition
SIGLEC1: Pathogen recognition, macrophage activation
SIRPG: immunomodulation; can lead to sepsis' hallmark overactive response.

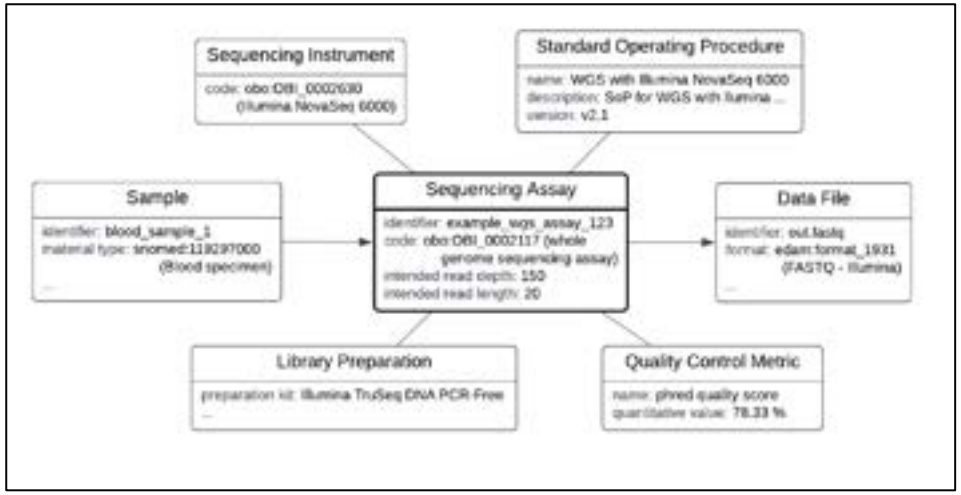
Key Processes
Antigen presentation
Signaling receptors
Interactions between lymphoid & non-lymphoid cells
Adaptive & innate immune responses



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

SPHN RDF Schema: Sequencing assay





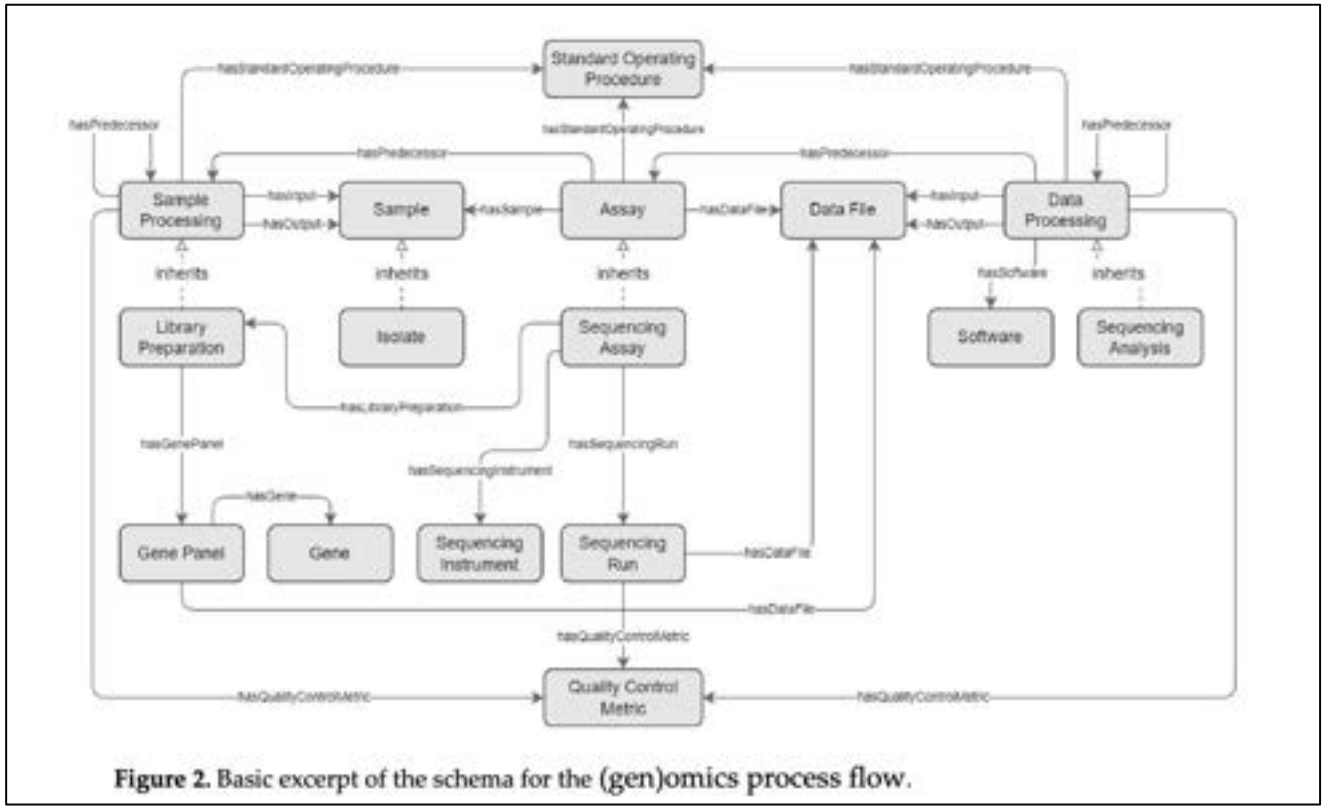


Figure 2. Basic excerpt of the schema for the (gen)omics process flow.

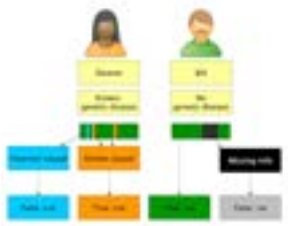
https://doi.org/10.20944/preprints202312.0373.v1

Future

Quantify integrity

Quantify evidence

Quantify a priori combination



Bayesian confidence diagnostic rate