HELMHOLTZ
MUNICH

Rare variant association tests

Ozvan Bocher
November 28th, 2023

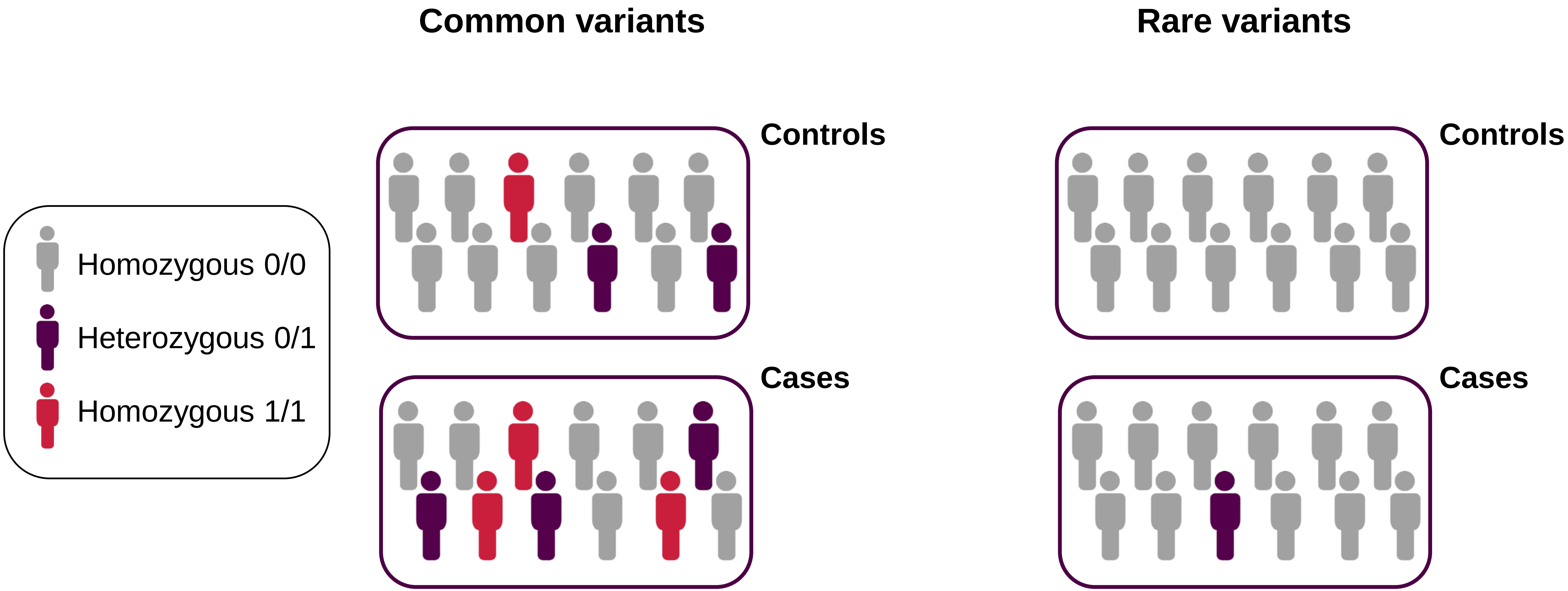
Rare variant association tests (RVAT)

1. Introduction
2. Types of RVAT
3. Grouping of variants
4. Selection of variants
5. RVAT in practice

1

Introduction

GWAS – rare variants



Homozygous 0/0

Heterozygous 0/1

Homozygous 1/1

Common variants

Rare variants

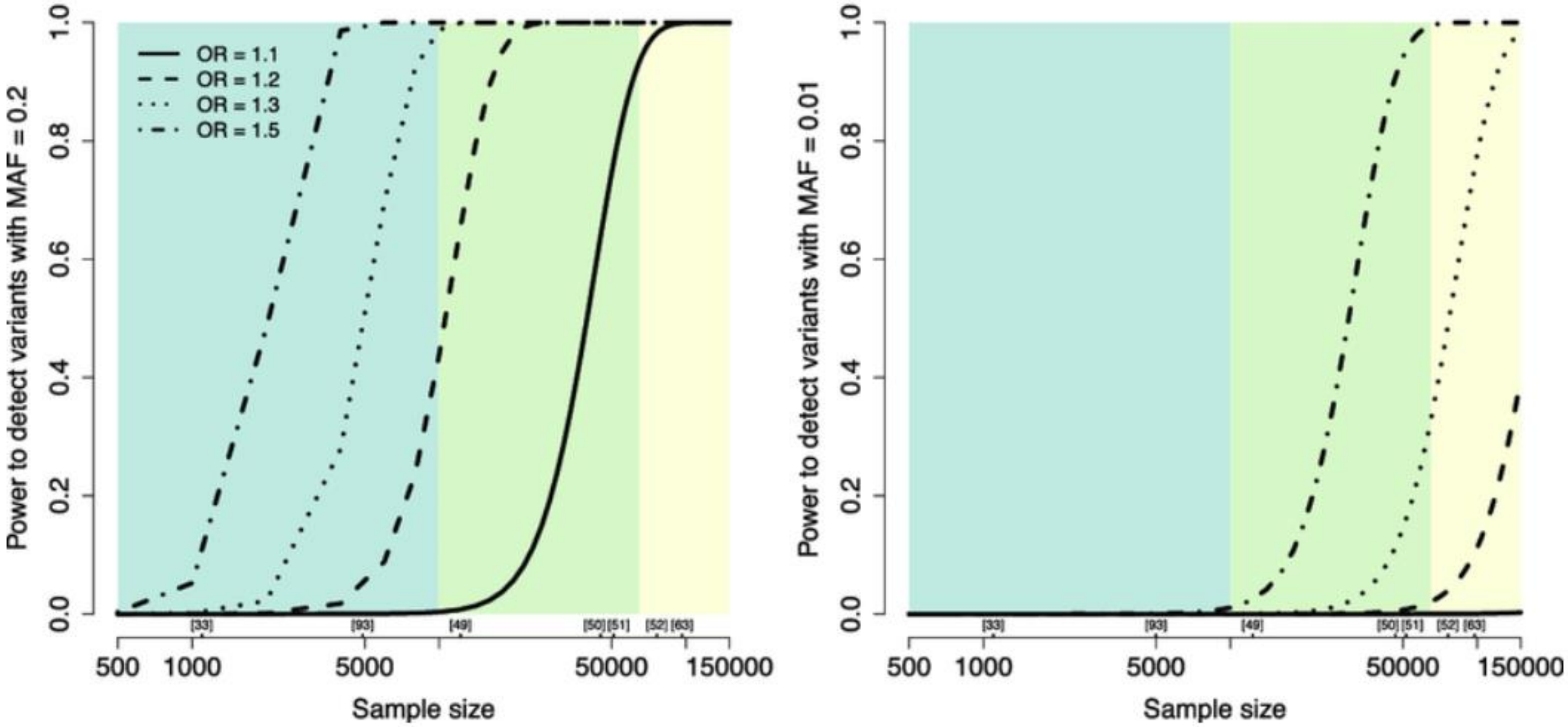
Controls

Cases

Controls

Cases

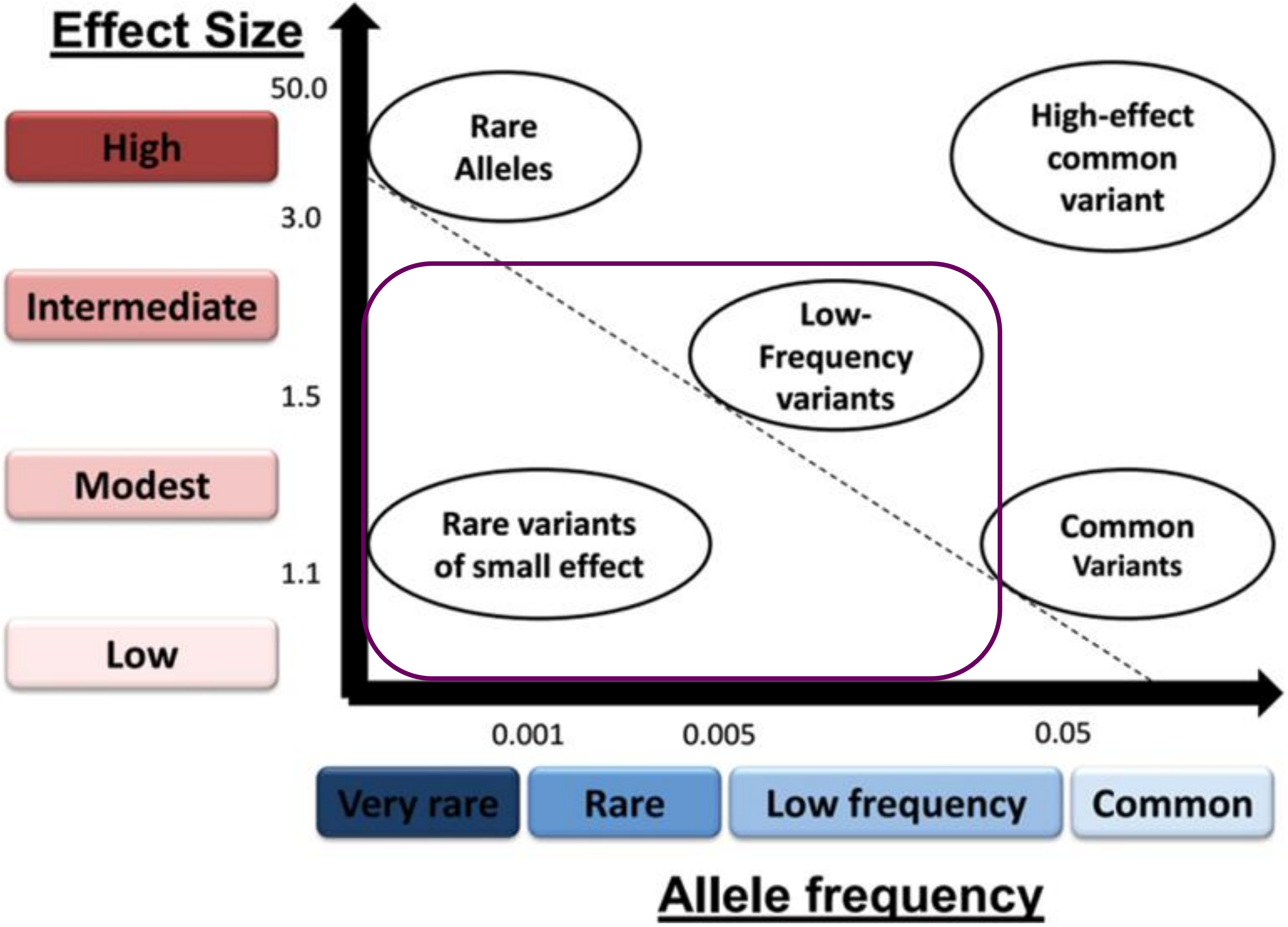
GWAS – statistical power



Power to detect associations of different effect size (odds ratio, OR) are compared for common variants (MAF ¼ 0.2, panel A) and rare variants (MAF ¼ 0.01, panel B). Effective sample sizes of several key studies are indicated along the x-axis, to reflect the power of the GWAS studies (blue), meta-analyses (green) and Immunochip-based studies (yellow) [93].

De Lange and Barrett, 2015, Journal of Autoimmunity

GWAS – rare variants



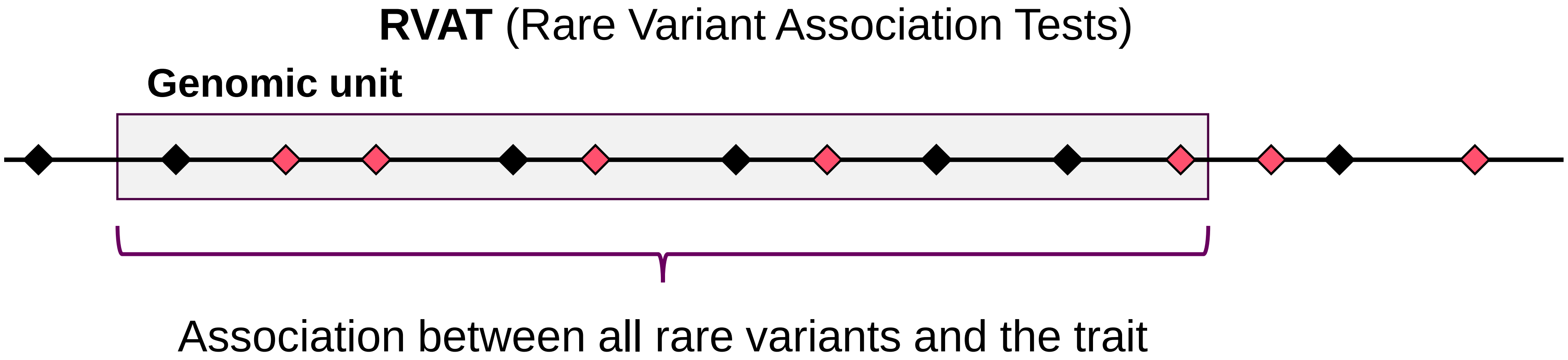
‘Missing heritability’

HELMHOLTZ MUNICH

Paulo et al., 2017

Rare variant association tests (RVAT)

- Issue : Lack of power to detect rare variants with low and intermediate effect sizes
- Hypothesis : Different rare variants can be observed in different individuals but with a similar effect on a given genomic region



RVAT (Rare Variant Association Tests)

Genomic unit

Association between all rare variants and the trait

◆ Rare variants

◇ Common variants: single-point associations in GWAS

2

Rare Variant Association Tests

(RVAT)

Burden tests

Hypotheses:

- All RV in the genomic region influence the phenotype
- The effect is similar between the variants

Computation of a burden score for each individual in each region

Comparison of the score’s distributions with regression models:

$$Y = \beta_{Cov} X_{Cov} + \beta_G X_G$$

X_G: *matrix of genetic score*

X_{Cov}: *matrix of covariates*

Y can be binary, continuous or categorial

Burden tests – types of scores

ID	Phenotype	RV1	RV2	RV3	RV4	RV5	CAST	X_G
1	Case	0	0	**1**	0	**1**	1	2
2	Case	**1**	**2**	0	**1**	0	1	4
3	Case	**1**	**1**	0	**1**	**1**	1	4
4	Control	**1**	0	**1**	0	0	1	2
5	Control	0	0	**1**	0	0	1	1
6	Control	0	0	0	0	0	0	0

- CAST score (*Morgenthaler et Thilly, 2007, Mutat Res*):
  - Binary score: present of at least one rare allele

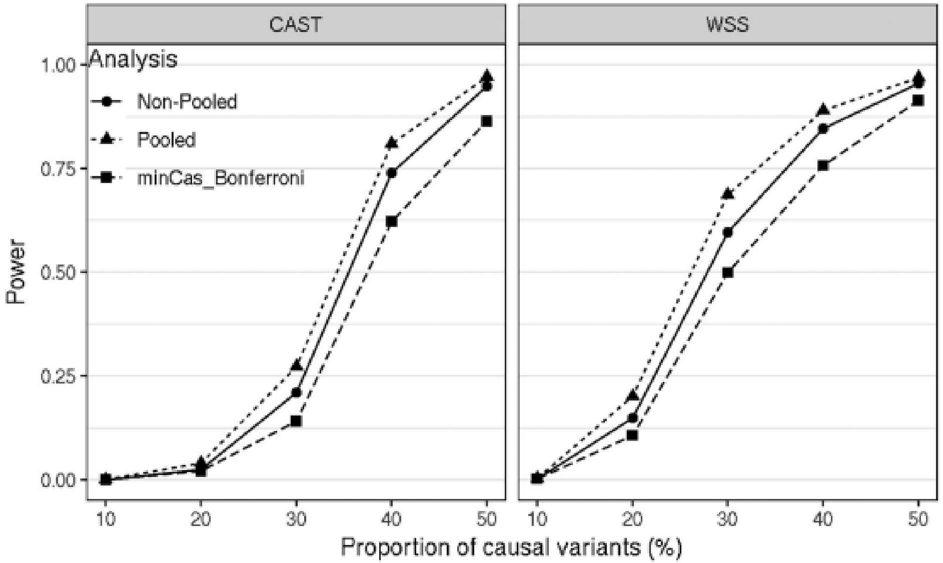
- Sum of alleles (here X_G) *(Li and Leal, 2008, AJHG)*

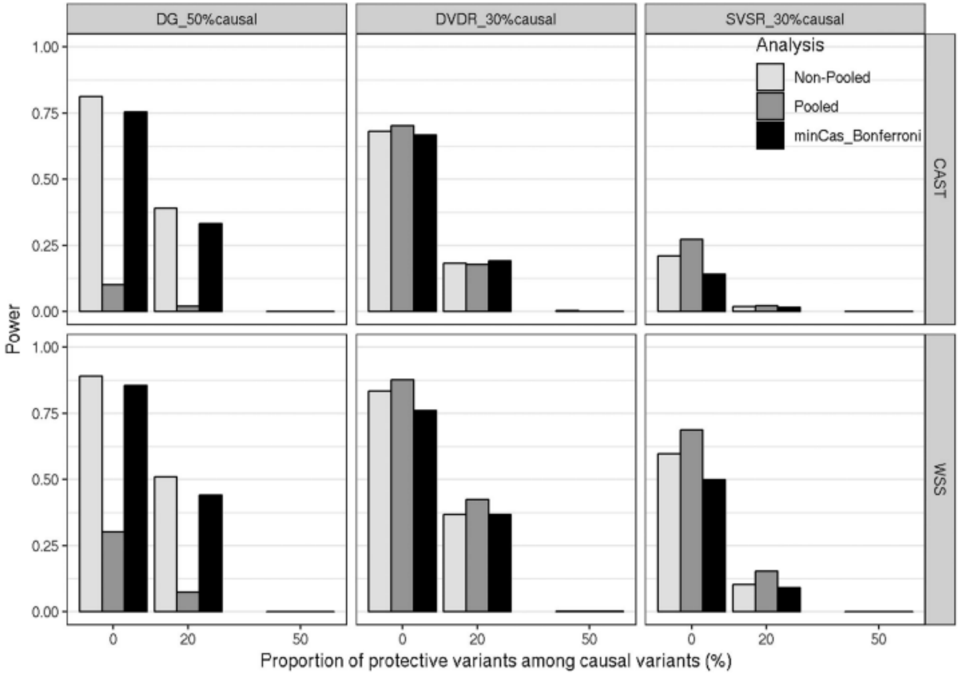
- WSS score (*Madsen et Browning, 2009, Plos Genetics*):
  - Hypothesis that rarer variants have stronger effects

    $w_i = {}^{1}/_{\sqrt{n_i \cdot q_i \cdot (1 - q_i)}}$ with q_i representing the frequency

Burden tests – limitations

- Lack of power when variants with different directions and non-causal variants are present





→ Variance-component RVAT to model mixed effects of RV

Bocher et al., 2019, Genetic Epidemiology

Variance-based RVAT: SKAT

SKAT: Sequence Kernel Association Tests

Hypothesis: RV in a genomic region have a mixture of null, protective and deleterious effects

→ Statistics based on the dispersion of the genetic effects

$$Y = \beta_{Cov}X_{Cov} + Zu$$

Z: matrix of weighted genotypes

$$u \sim MVN(0, \tau I)$$

- **H0: $\boldsymbol{\tau = 0}$** → All RV have a null genetic effect

Wu et al., 2011, AJHG

Variance-based RVAT: SKAT

ID	Phenotype	RV1	RV2	RV3	RV4	RV5
1	Case	1	0	0	0	1
2	Case	0	0	1	1	0
3	Case	0	2	0	1	0
4	Case	1	0	0	0	1
5	Control	0	**1**	0	1	**1**
6	Control	1	0	0	**1**	0
7	Control	0	**1**	1	0	0
8	Control	1	0	**1**	0	0
		0	0	-1	0	1

Variance=0.5

ID	Phenotype	RV1	RV2	RV3	RV4	RV5
1	Case	1	**1**	0	0	0
2	Case	0	0	1	1	0
3	Case	0	2	0	1	0
4	Case	1	**1**	0	**1**	0
5	Control	0	0	0	1	0
6	Control	1	0	0	0	1
7	Control	0	0	1	0	1
8	Control	1	0	1	0	1
		0	4	-1	2	-3

Variance=2.5

RVAT

Burden tests

- Easy to interpret
- OR estimates
- Sensitive to variant selection ++

Variance-component tests

- Can incorporate mixed-effects variants
- Harder to interpret

→ SKAT-Optimal

$$S_{SKAT-O} = \boldsymbol{\rho} \cdot S_{SKAT} + (1 - \boldsymbol{\rho}) \cdot S_{Burden}$$

- Finds optimal $\boldsymbol{\rho}$ that minimizes the p-value
- Needs permutations to compute the p-value
- Can also be hard to interpret

Lee et al., 2012, AJHG

3

Aggregation of rare variants

WES vs WGS

- WES: access to coding regions of the genome

→ Positions of genes often used to define genomic regions

→ Approximately 20,000 RVAT performed

- WGS: access to the whole-genome

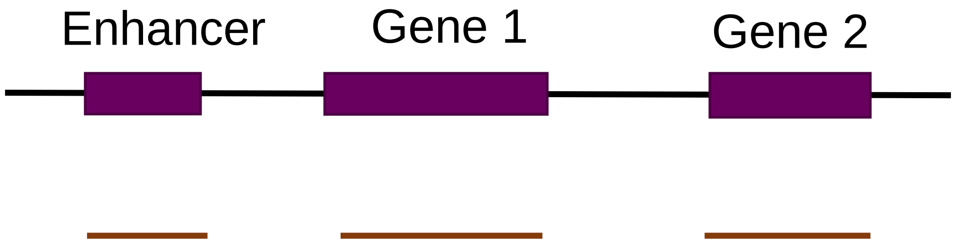
→ RVAT very often limited to coding parts of the genome

RVAT outside the coding genome

- Use of defined elements (TADs, enhancers, silencers, …)

→ Can be of limited size and gather a limited number of variants

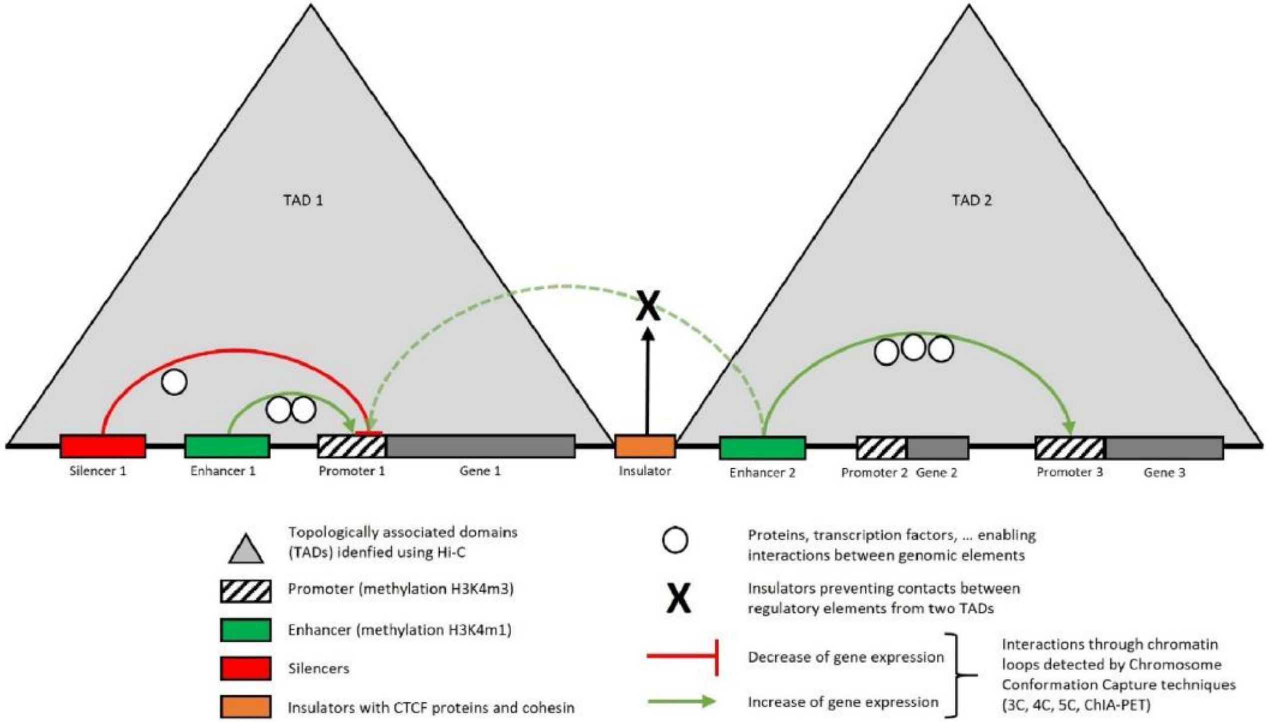
→ Not covering the whole-genome



Enhancer

Gene 1

Gene 2


TAD 1
TAD 2
X
Silencer 1
Enhancer 1
Promoter 1
Gene 1
Insulator
Enhancer 2
Promoter 2
Gene 2
Promoter 3
Gene 3
Topologically associated domains (TADs) idenfied using Hi-C
Promoter (methylation H3K4m3)
Enhancer (methylation H3K4m1)
Silencers
Insulators with CTCF proteins and cohesin
Proteins, transcription factors, … enabling interactions between genomic elements
Insulators preventing contacts between regulatory elements from two TADs
Decrease of gene expression
Increase of gene expression
Interactions through chromatin loops detected by Chromosome Conformation Capture techniques (3C, 4C, 5C, ChIA-PET)

HELMHOLTZ MUNICH

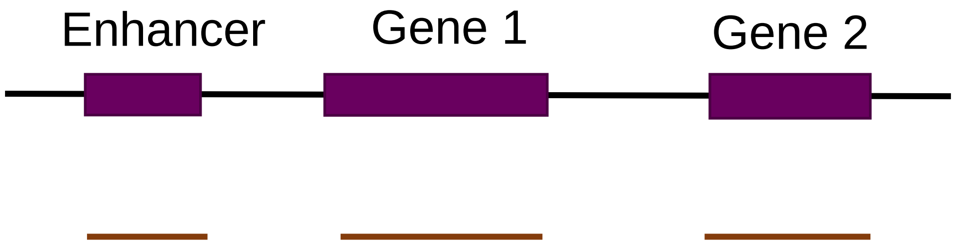
Bocher et al. 2020, HMG

RVAT outside the coding genome

- Use of defined elements (TADs, enhancers, silencers, …)

→ Can be of limited size and gather a limited number of variants

→ Not covering the whole-genome



- Use of sliding windows (WGScan[2], ScanG[3])

→ Covers the whole genome

→ No biological information used

→ Results in a high number of tests, needs permutations

- Alternative method: RAVA-FIRST[4]

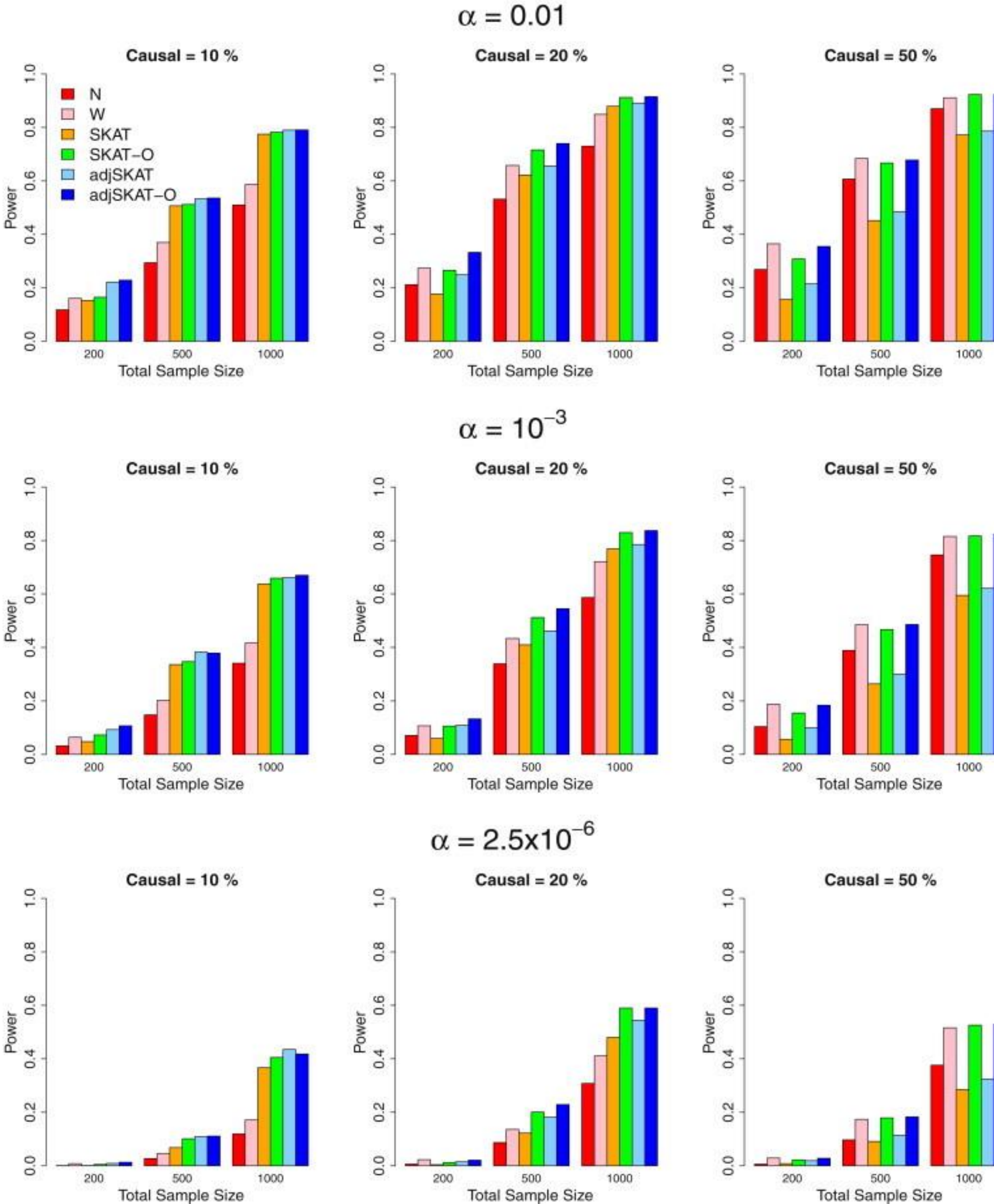
→ Definition of regions in the non-coding genome based on genomic constraint

[1] Bocher et al. 2020, HMG; [2] He et al., 2019. Nature Communications; [3] Li et al., 2019, AJHG; [4] Bocher et al. 2022, PlosG

4

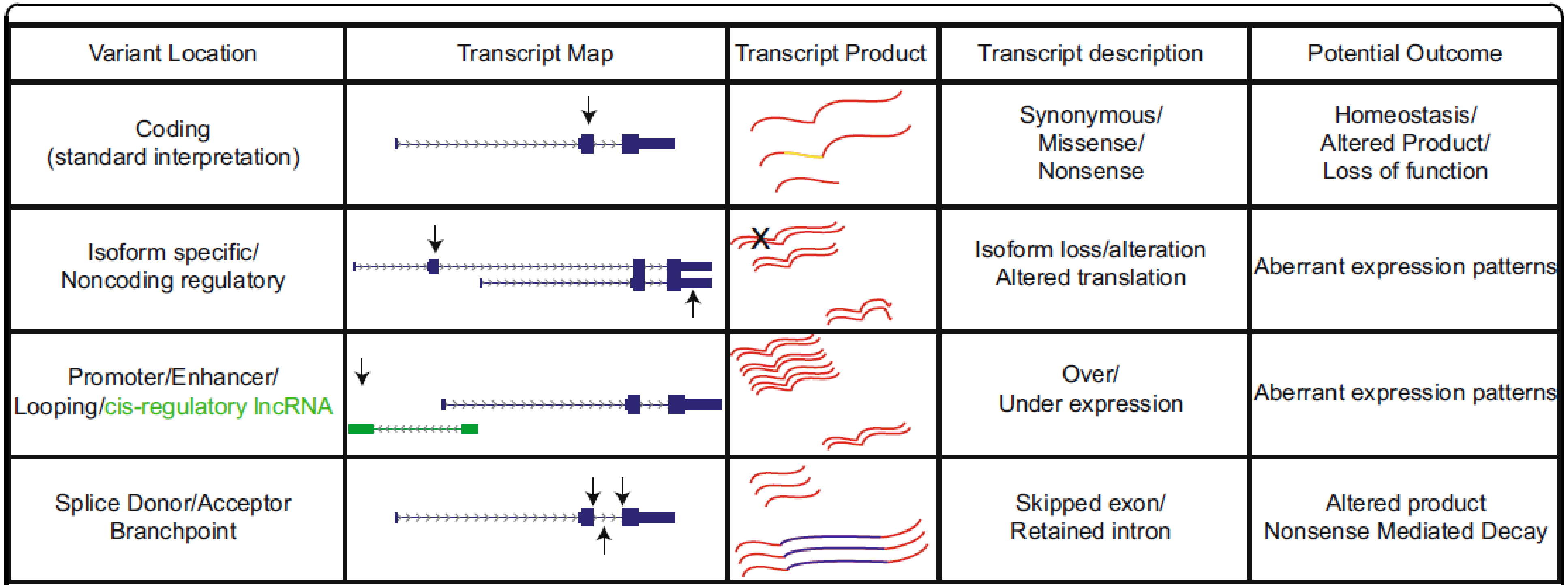
Selection of qualifying RV

Importance of selecting variants



Lee et al. 2012, AJHG

WES vs WGS



Variant Location	Transcript Map	Transcript Product	Transcript description	Potential Outcome
Coding (standard interpretation)			Synonymous/ Missense/ Nonsense	Homeostasis/ Altered Product/ Loss of function
Isoform specific/ Noncoding regulatory		X	Isoform loss/alteration Altered translation	Aberrant expression patterns
Promoter/Enhancer/ Looping/cis-regulatory lncRNA			Over/ Under expression	Aberrant expression patterns
Splice Donor/Acceptor Branchpoint			Skipped exon/ Retained intron	Altered product Nonsense Mediated Decay

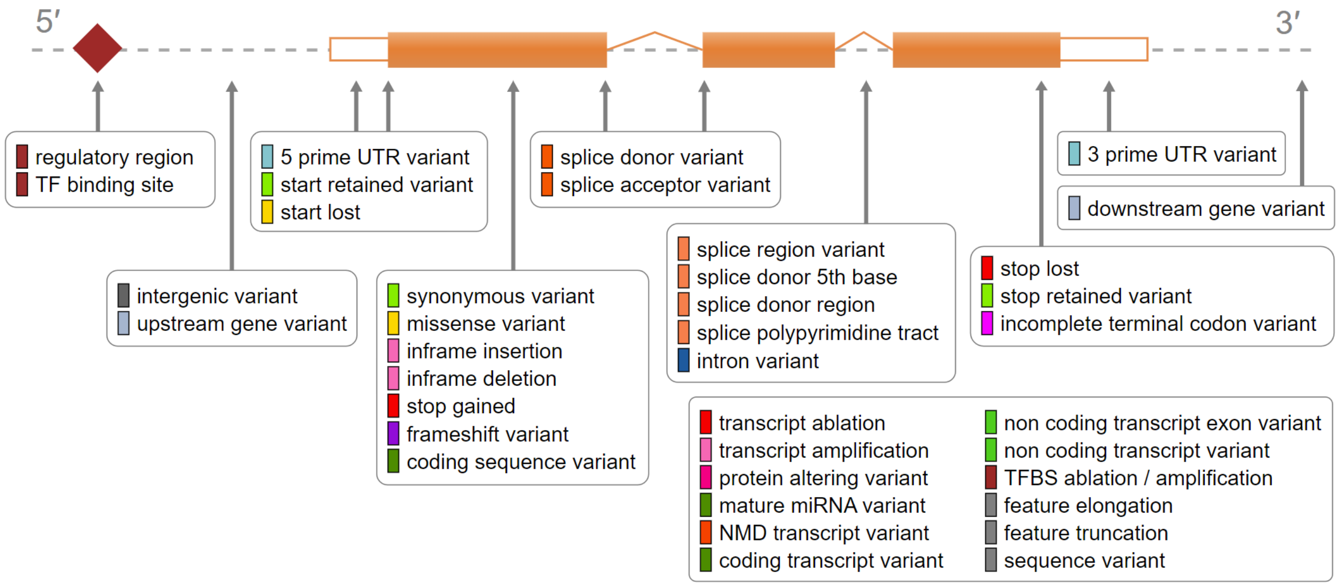
Fig. 2 How coding and noncoding variation can impact gene function. Variants (arrows) at a hypothetical locus are shown along with potential functional impacts

HELMHOLTZ MUNICH

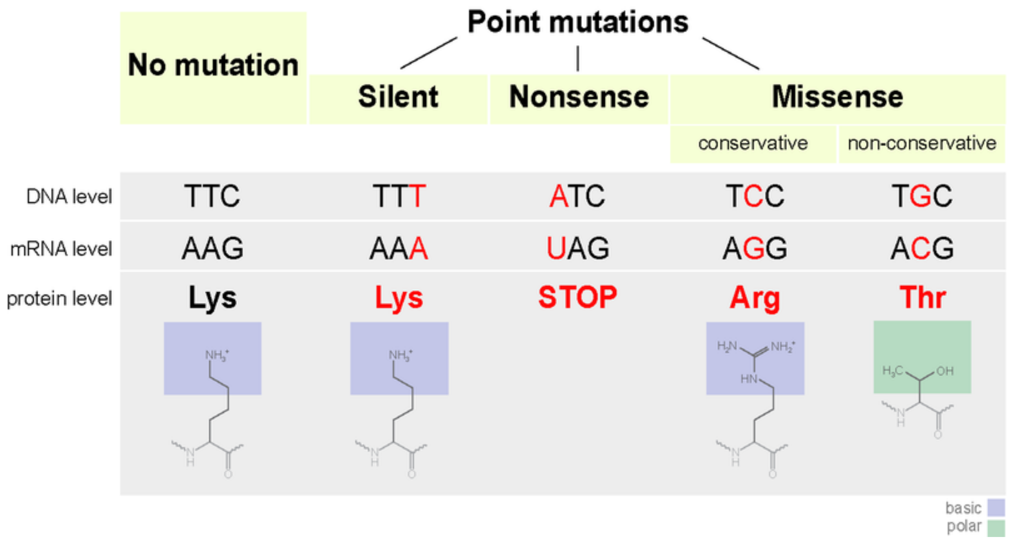
Gloss and Dinger 2018, EMM

Selection in the coding genome – Annotation tools

In the coding genome, focus on the consequences on the proteins



- regulatory region
- TF binding site
- 5 prime UTR variant
- start retained variant
- start lost
- splice donor variant
- splice acceptor variant
- 3 prime UTR variant
- downstream gene variant
- intergenic variant
- upstream gene variant
- synonymous variant
- missense variant
- inframe insertion
- inframe deletion
- stop gained
- frameshift variant
- coding sequence variant
- splice region variant
- splice donor 5th base
- splice donor region
- splice polypyrimidine tract
- intron variant
- stop lost
- stop retained variant
- incomplete terminal codon variant
- transcript ablation
- transcript amplification
- protein altering variant
- mature miRNA variant
- NMD transcript variant
- coding transcript variant
- non coding transcript exon variant
- non coding transcript variant
- TFBS ablation / amplification
- feature elongation
- feature truncation
- sequence variant



	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr

Commonly applied filtering = variants with a consequence of at least mis-sense

→ Impact on the protein

Available tools: VEP

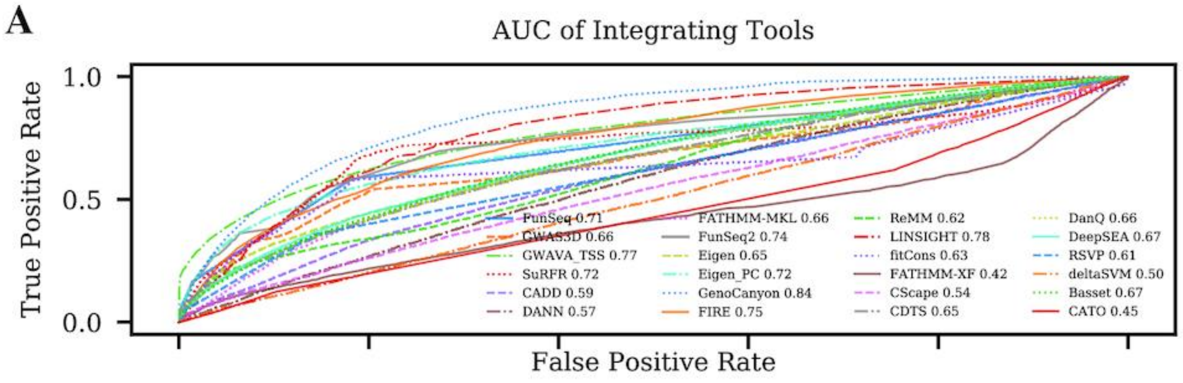
McLaren et al. 2016, Genome Biology

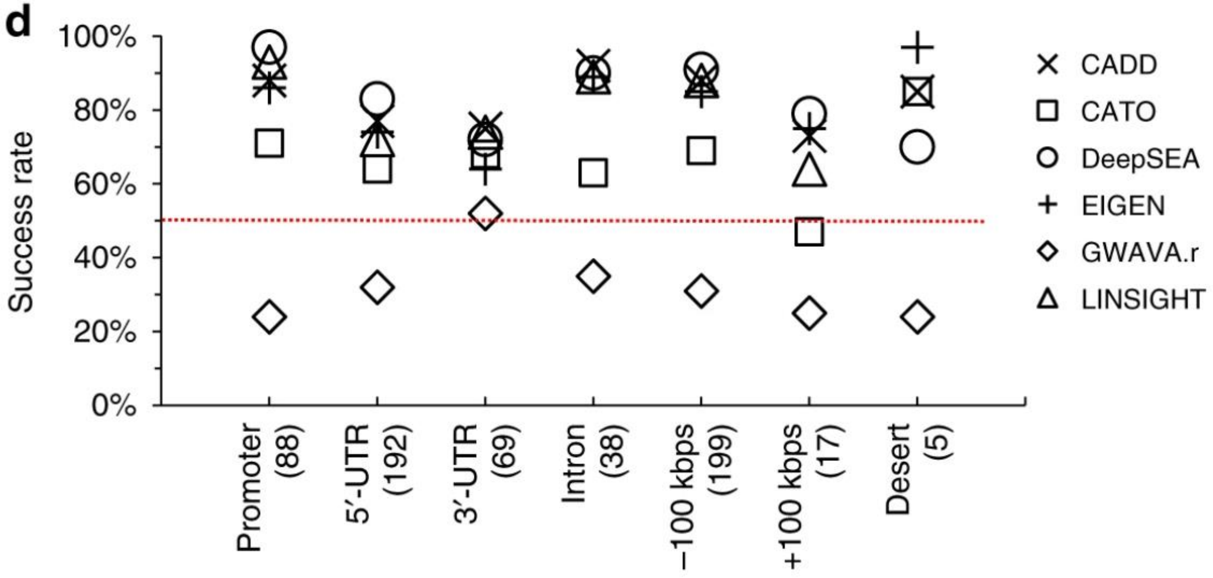
Selection in the non-coding genome – scores

- In the non-coding genome: no direct consequence on the proteins
- Variants that regulate gene expression → harder to class them in categories of variants
- Development of pathogenic scores

→ Can also be used in the coding genome

→ Performance highly dependent on training set and type of variants

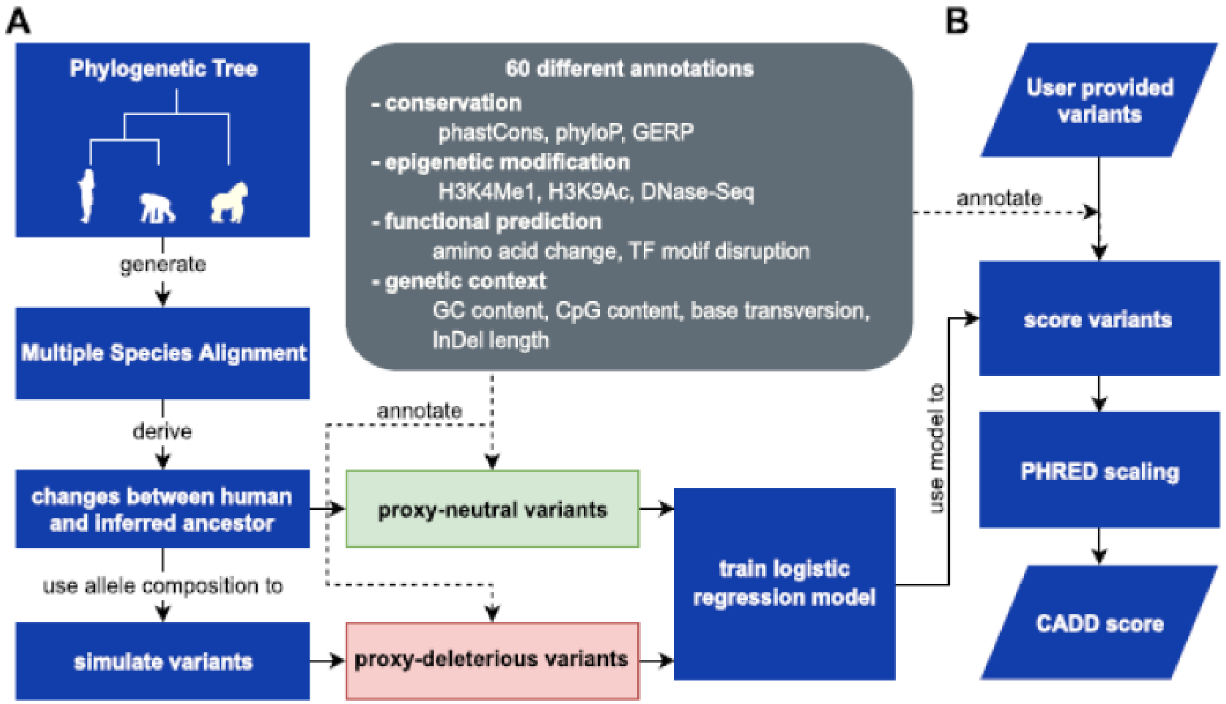




Zhang et al. 2019, NAR; Liu et al. 2019, Nat. Comm

Selection in the non-coding genome – CADD scores

- CADD - Combined Annotation Dependent Depletion
- Define for every position and allele for SNPs and possible INDEL annotation
- Commonly used in RVAT



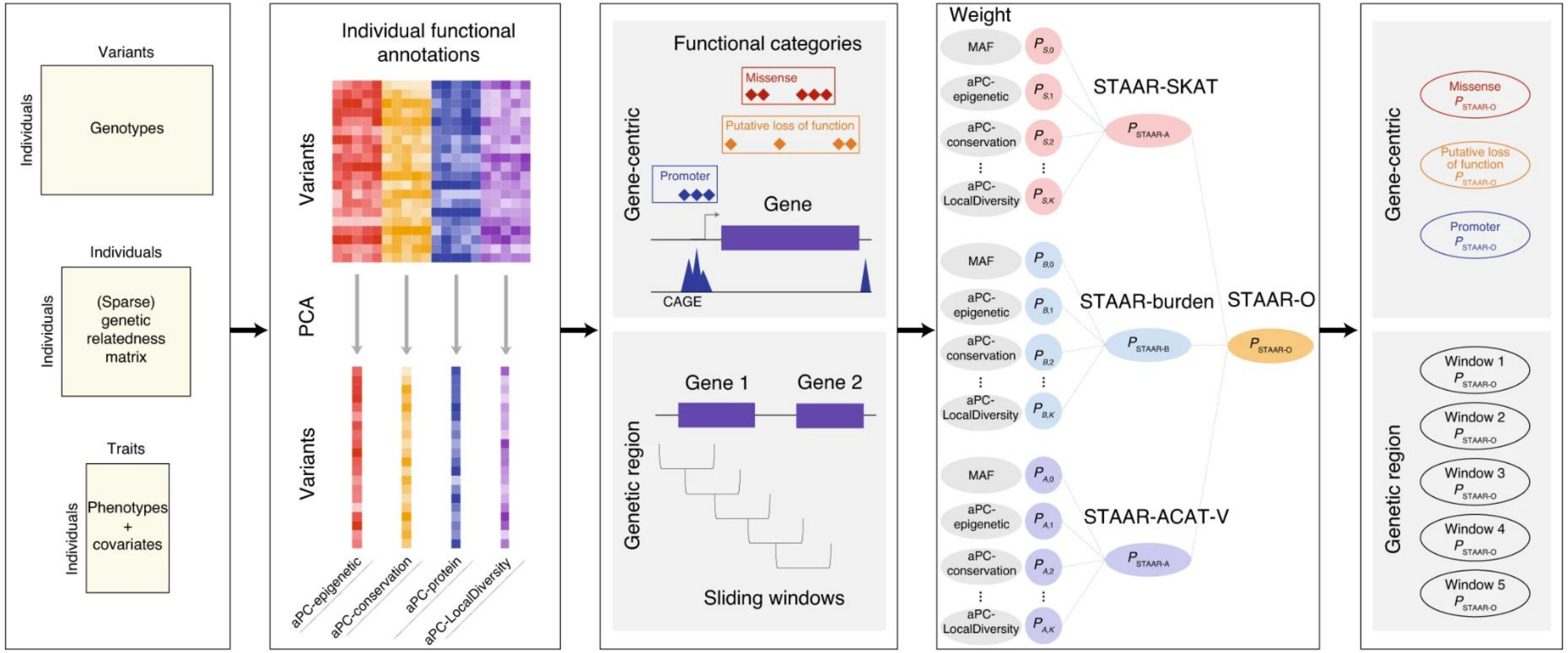
A
B
Phylogenetic Tree
generate
Multiple Species Alignment
derive
changes between human and inferred ancestor
use allele composition to
simulate variants
60 different annotations
- conservation
phastCons, phyloP, GERP
- epigenetic modification
H3K4Me1, H3K9Ac, DNase-Seq
- functional prediction
amino acid change, TF motif disruption
- genetic context
GC content, CpG content, base transversion, InDel length
annotate
proxy-neutral variants
proxy-deleterious variants
train logistic regression model
use model to
User provided variants
annotate
score variants
PHRED scaling
CADD score

Rentz et al. 2018, NAR

Alternative of hard filtering

- Include all variants and **weight** them according to the scores

- Include **multiple annotation scores** and select the best combination:
  - Annotation scores capture complementary information
  - Examples: STAAR (*Li et al. 2020, Nat. Gen*.), FunSPU (*Ma et al. 2019*)
  - Bayesian methods: DoEstRare (*Persyn 2017, Plos One*), BeviMed (*Greene 2017, AJHG*)



5

RVAT in practice

Available R packages

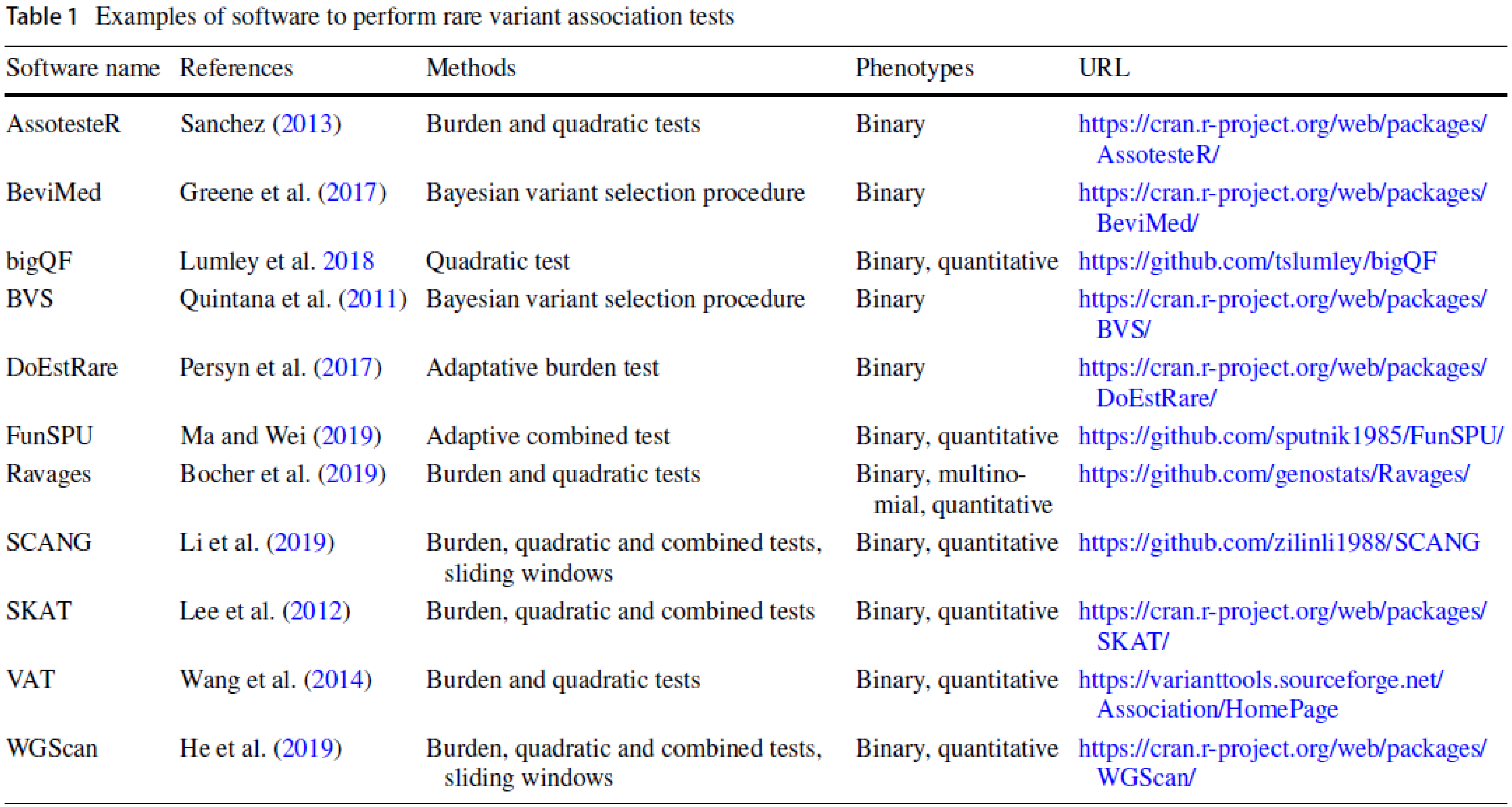


Table 1 Examples of software to perform rare variant association tests

Software name	References	Methods	Phenotypes	URL
AssotesteR	Sanchez (2013)	Burden and quadratic tests	Binary	https://cran.r-project.org/web/packages/AssotesteR/
BeviMed	Greene et al. (2017)	Bayesian variant selection procedure	Binary	https://cran.r-project.org/web/packages/BeviMed/
bigQF	Lumley et al. 2018	Quadratic test	Binary, quantitative	https://github.com/tslumley/bigQF
BVS	Quintana et al. (2011)	Bayesian variant selection procedure	Binary	https://cran.r-project.org/web/packages/BVS/
DoEstRare	Persyn et al. (2017)	Adaptative burden test	Binary	https://cran.r-project.org/web/packages/DoEstRare/
FunSPU	Ma and Wei (2019)	Adaptive combined test	Binary, quantitative	https://github.com/sputnik1985/FunSPU/
Ravages	Bocher et al. (2019)	Burden and quadratic tests	Binary, multinomial, quantitative	https://github.com/genostats/Ravages/
SCANG	Li et al. (2019)	Burden, quadratic and combined tests, sliding windows	Binary, quantitative	https://github.com/zilinli1988/SCANG
SKAT	Lee et al. (2012)	Burden, quadratic and combined tests	Binary, quantitative	https://cran.r-project.org/web/packages/SKAT/
VAT	Wang et al. (2014)	Burden and quadratic tests	Binary, quantitative	https://varianttools.sourceforge.net/Association/HomePage
WGScan	He et al. (2019)	Burden, quadratic and combined tests, sliding windows	Binary, quantitative	https://cran.r-project.org/web/packages/WGScan/

Bocher et al. 2020, HMG

Other tools

- RVTests (*Zhan et al. 2016, Bioinformatics*)
- SEQSpark (*Zhang et al. 2017, AJHG*)
- REGENIE (*Mbatchou et al. 2021, Nat. Genet*.)

→ Can deal with large sample sizes such as Biobank data but restricted to classical methods

- **Which tool to choose?**

→ Depends on the desired statistical test

→ Depends on how to group and filter the variants

→ Depends on the size of data to analyze

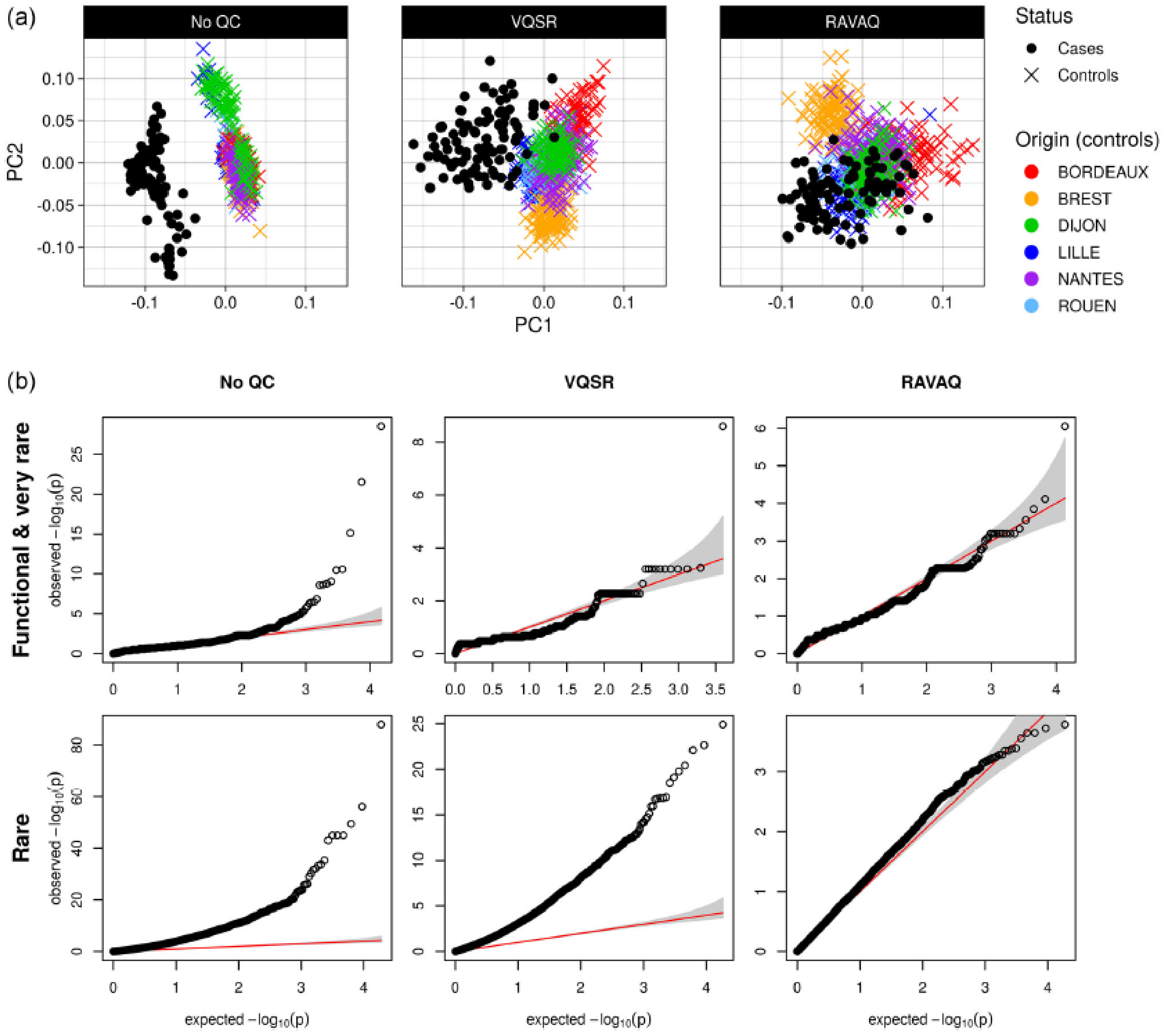
Further considerations

- **Quality control**
  - Measures like HWE not adapted for RV
  - Problem of stratification more important and at finer scale that common variants as RV have appeared more recently and are often specific to population
    - PCA is not capturing well stratification from RV
    - Including PCs as covariates do no always well correct
    - Still no clear answer on how to use PCA for RVAT
  - Dedicated pipeline for RVAT is useful
    - Example: RAVAQ

Marenne et al. 2022, Genetic Epidemiology

Further considerations

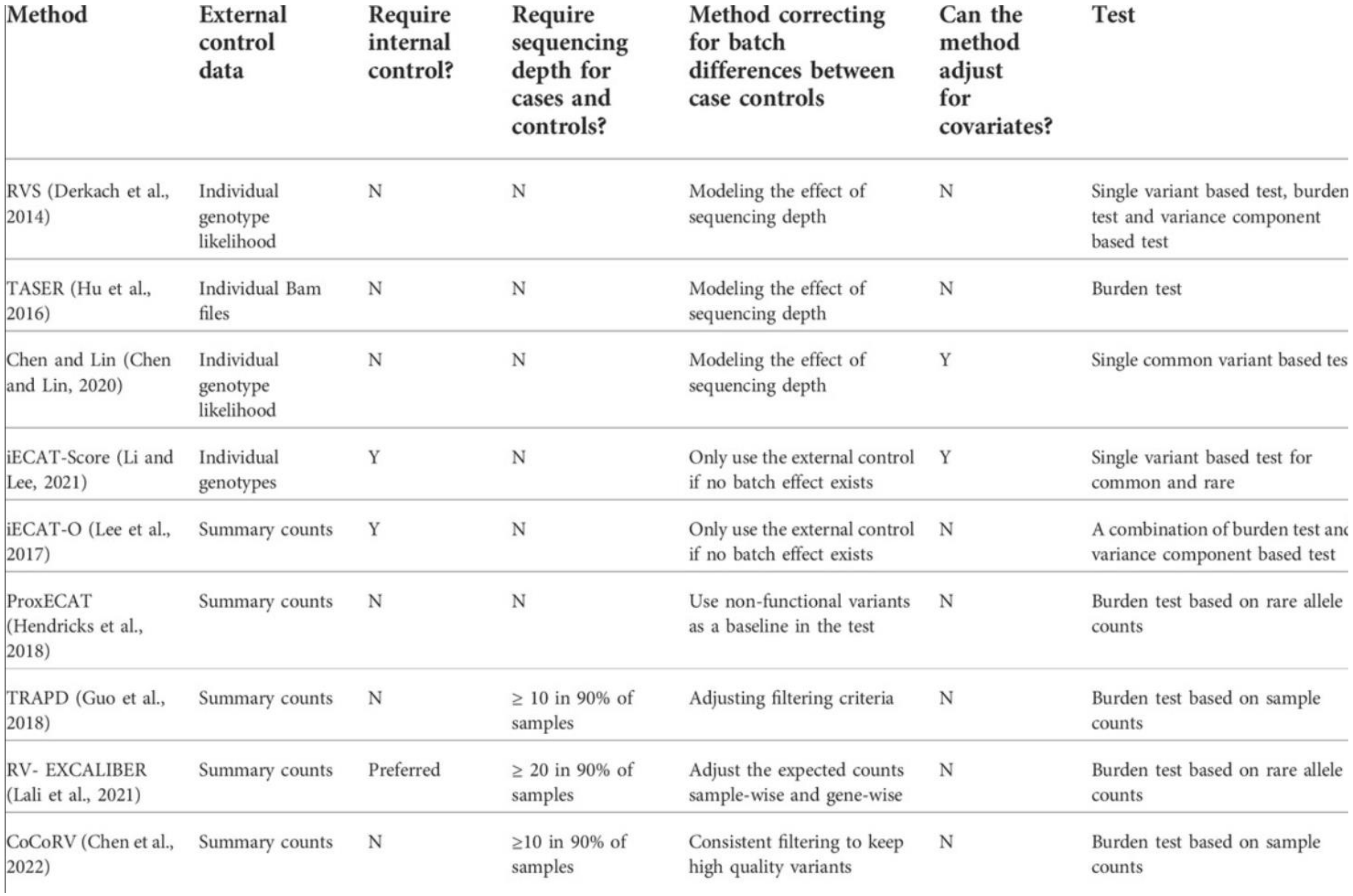
- **Quality control**



Marenne et al. 2022, Genetic Epidemiology

Further considerations

- **Selection of controls**



Method	External control data	Require internal control?	Require sequencing depth for cases and controls?	Method correcting for batch differences between case controls	Can the method adjust for covariates?	Test
RVS (Derkach et al., 2014)	Individual genotype likelihood	N	N	Modeling the effect of sequencing depth	N	Single variant based test, burden test and variance component based test
TASER (Hu et al., 2016)	Individual Bam files	N	N	Modeling the effect of sequencing depth	N	Burden test
Chen and Lin (Chen and Lin, 2020)	Individual genotype likelihood	N	N	Modeling the effect of sequencing depth	Y	Single common variant based tes
iECAT-Score (Li and Lee, 2021)	Individual genotypes	Y	N	Only use the external control if no batch effect exists	Y	Single variant based test for common and rare
iECAT-O (Lee et al., 2017)	Summary counts	Y	N	Only use the external control if no batch effect exists	N	A combination of burden test anc variance component based test
ProxECAT (Hendricks et al., 2018)	Summary counts	N	N	Use non-functional variants as a baseline in the test	N	Burden test based on rare allele counts
TRAPD (Guo et al., 2018)	Summary counts	N	≥ 10 in 90% of samples	Adjusting filtering criteria	N	Burden test based on sample counts
RV- EXCALIBER (Lali et al., 2021)	Summary counts	Preferred	≥ 20 in 90% of samples	Adjust the expected counts sample-wise and gene-wise	N	Burden test based on rare allele counts
CoCoRV (Chen et al., 2022)	Summary counts	N	≥10 in 90% of samples	Consistent filtering to keep high quality variants	N	Burden test based on sample counts

- **Tissue-specific** annotations could empower RVAT by better predicting variant consequences

Chen et al. 2022, Frontiers in Genetics

Conclusions

- Most studies are based on burden or variance-component tests

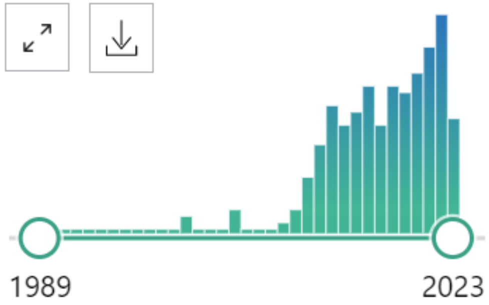
  → Burden tests: easier to interpret

  → Variance-component tests: less sensitive to selection of variants

- Definition of regions and qualifying variants

  → Active area of research, especially in the non-coding genome

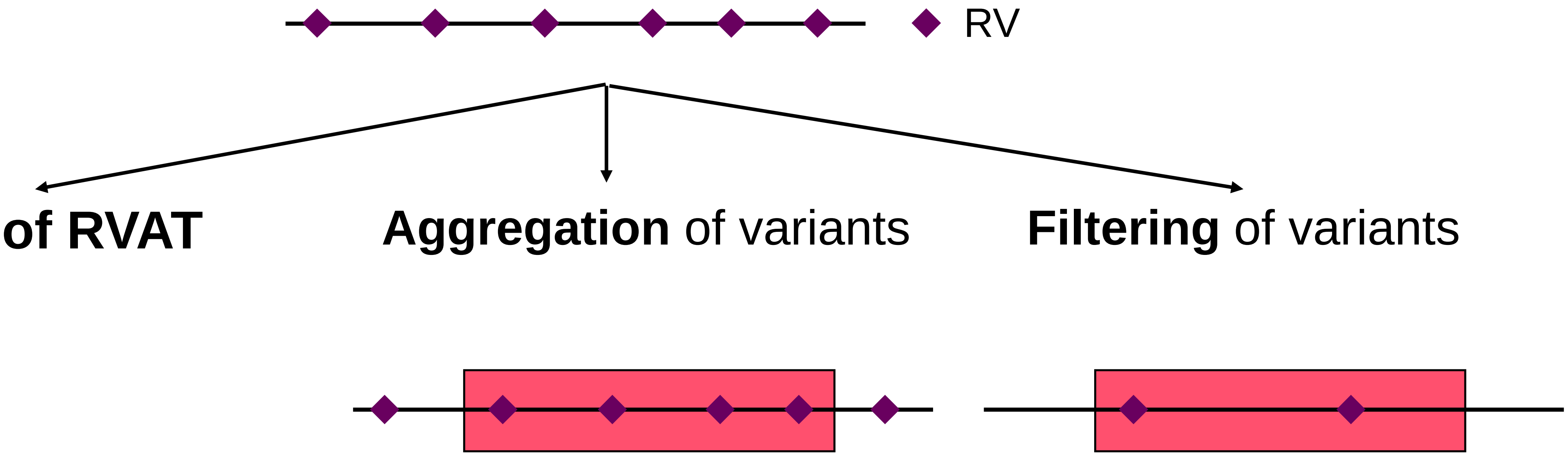
  → Most of the WGS data are not currently used in the RV context



Rare variant aggregation tests

- Meta-analysis possible but more challenging (aggregation of variants + stratification)

Conclusions



Type of RVAT

Aggregation of variants

Filtering of variants

Thank you.