VarGen:
An R package for disease-associated variant discovery and annotation

A data integration use case

All credit to: Dr Corentin Molitor

www.cranfield.ac.uk

What is VarGen ?

- An R package that access different databases to discover and annotate variants associated with a certain disease (based on an OMIM ID).

rsid	chr	pos	ensembl_gene_id	hgnc_symbol
rs17817449	chr16	53779455	ENSG00000140718	FTO
rs17817469	chr3	12368850	ENSG00000132170	PPARG
rs17847043	chr6	131872658	ENSG00000197594	ENPP1
rs17847044	chr6	131873208	ENSG00000197594	ENPP1
rs17847045	chr6	131878802	ENSG00000197594	ENPP1
rs17847045	chr6	131878802	ENSG00000197594	ENPP1
rs17847047	chr6	131890297	ENSG00000197594	ENPP1
rs17847050	chr6	131860393	ENSG00000197594	ENPP1

e.g.: Alzheimer's (104300)

FANTOM FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

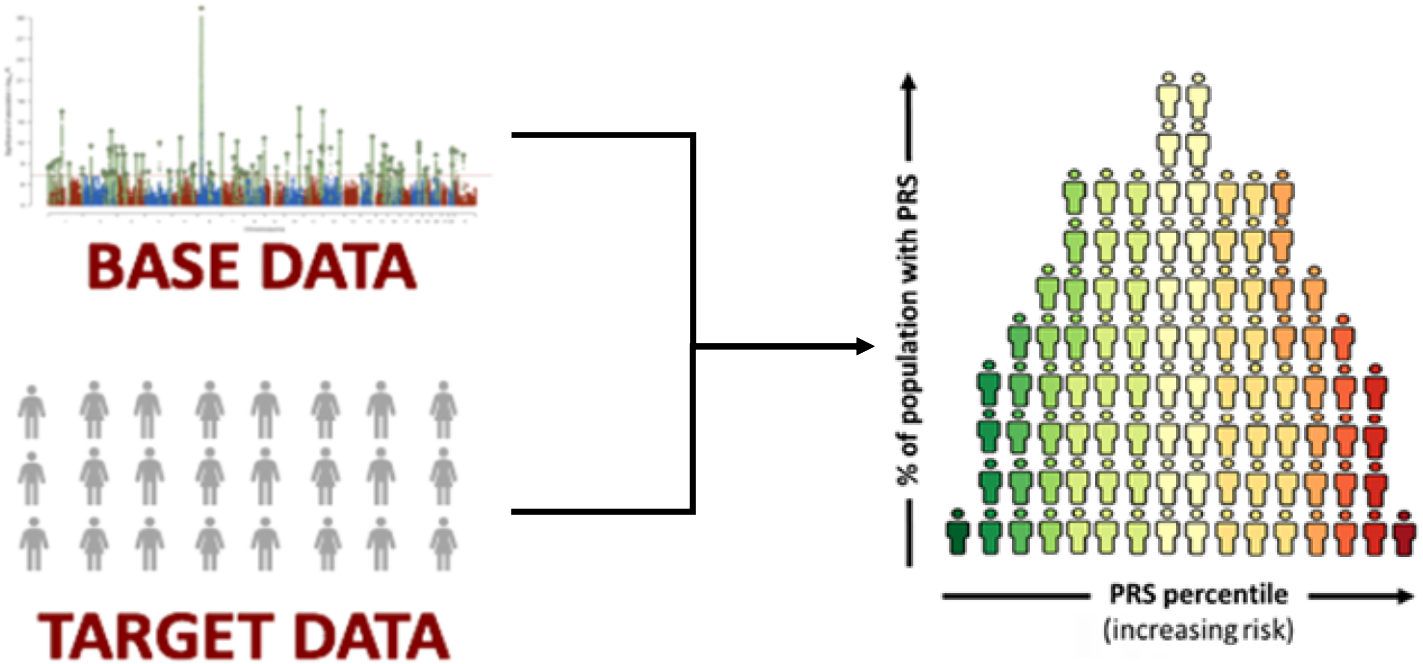
NCBI

GWAS Catalog

biomart

Why VarGen?

- Complex diseases are due to **the accumulation of a lot of variants**, **each having a small impact** on the disease.
- Knowing the variants that are linked to a disease is crucial to the understanding and treatment of the disease. (e.g.: precision medicine)



A. V. Khera *et al.*, '*Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations*', *Nat Genet*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018, doi: 10.1038/s41588-018-0183-z

The growing amount of data in public repositories

7,324

Phenotypes with known molecular basis in

1,053,623,523

Variants (rsids) in dbSNP for *"homo sapiens"*

470,406

Variant-trait associations in the GWAS catalog

GWAS catalog diagram

The challenges of genomic data integration

The challenges of genomic data integration

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38

The challenges of genomic data integration

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38

- **Different means** of **accessing** each database (API, local files…)

The challenges of genomic data integration

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38

- **Different means** of **accessing** each database (API, local files…)

- **Different kind** of data (genes, variants, association, annotation...)

How to solve it?

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38

- **Different means** of **accessing** each database (API, local files…)

- **Different kind** of data (genes, variants, association, annotation...)

How to solve it:

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38
  LiftOver: VarGen reports all coordinates on GRch38

- **Different means** of **accessing** each database (API, local files…)

- **Different kind** of data (genes, variants, association, annotation...)

How to solve it:

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38
  LiftOver: VarGen reports all coordinates on GRch38

- **Different means** of **accessing** each database (API, local files…)
  VarGen automatically downloads and access the necessary files / API

- **Different kind** of data (genes, variants, association, annotation…)

How to solve it:

- **Different versions** of the human genome: GRch37 *(2014!)* vs GRch38
  LiftOver: VarGen reports all coordinates on GRch38

- **Different means** of **accessing** each database (API, local files…)
  VarGen automatically downloads and access the necessary files / API

- **Different kind** of data (genes, variants, association, annotation…)
  For simplicity: output all the information into a single data.frame
  (*SQLite* was also considered as an alternative)

Output example:

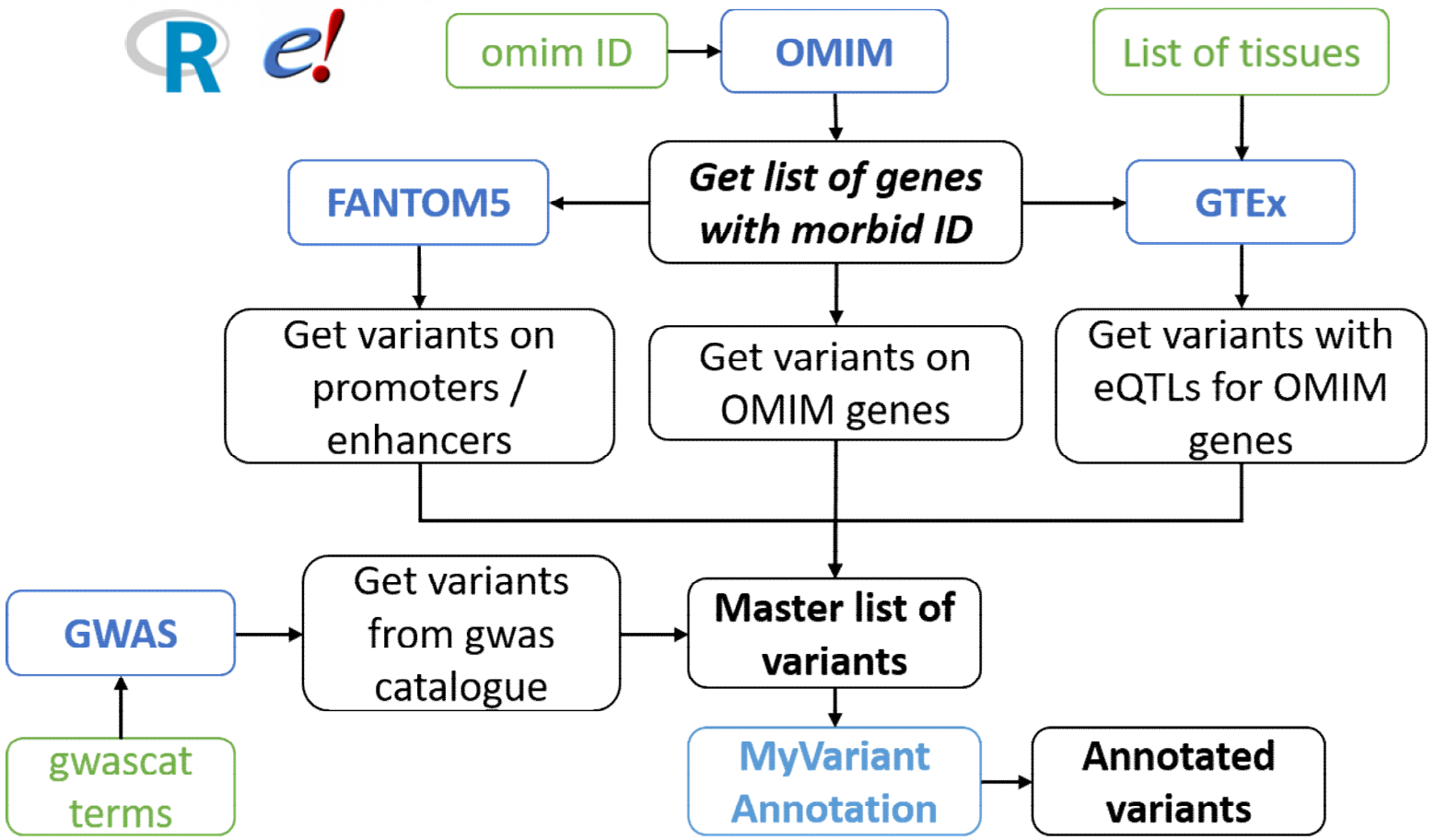
rsid	chr	position	HGNC
rs1043543911	chr8	37965344	ADRB3
rs1043811762	chr11	74006927	UCP3
rs1044022939	chr18	60371575	MC4R
rs1044454193	chr10	93567045	FFAR4
rs1044498	chr6	131851228	ENPP1
rs1044548	chr6	131890623	ENPP1
rs1044737470	chr5	96412463	PCSK1
rs1045328134	chr8	37966251	ADRB3
rs1045550142	chr16	54040160	FTO
rs1046537667	chr10	93587390	FFAR4
rs1046595093	chr5	71719400	CARTPT
rs10467147	chr12	40373560	LRRK2
rs10468281	chr16	53892854	FTO
rs1047063799	chr11	74005885	UCP3
rs1048093688	chr6	131858703	ENPP1
rs1048624393	chr5	96394998	PCSK1
rs104894319	chr11	74005844	UCP3
rs1049269132	chr5	96432931	PCSK1
rs10492872	chr16	54086418	FTO
rs1049385311	chr18	60371460	MC4R
rs1049385311	chr18	60371460	MC4R

Output example: after the annotation (credits to Matt Brember):

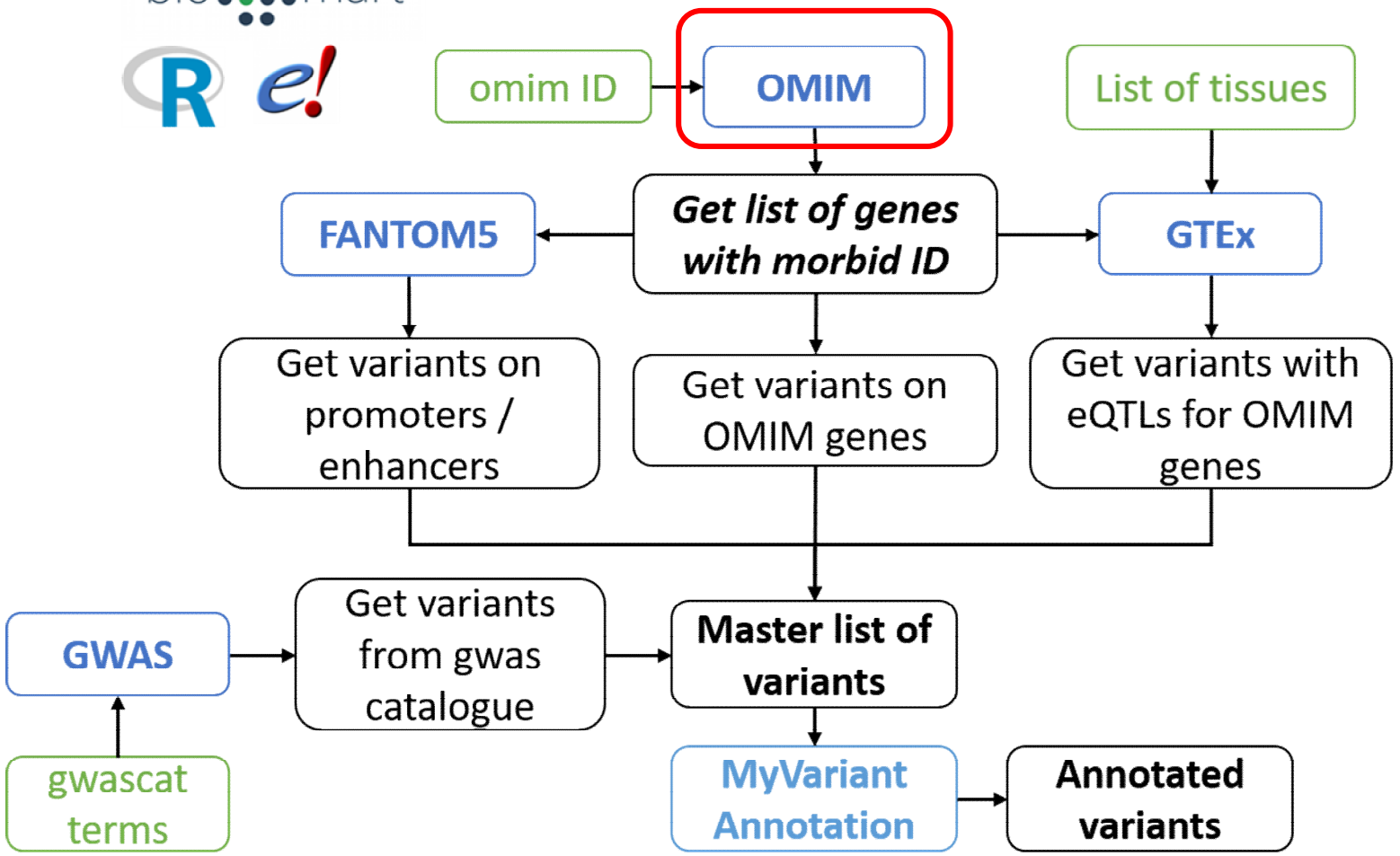
rsid	chr	position	HGNC	CADD	FATHMM	consequence	clinVar sign.	snpeff_ann
rs1043543911	chr8	37965344	ADRB3	13.35	0.084341	NON_SYNONYMOUS		MODERATE
rs1043811762	chr11	74006927	UCP3	32	0.89594	NON_SYNONYMOUS		MODERATE
rs1044022939	chr18	60371575	MC4R	28.6	0.574921	NON_SYNONYMOUS		MODERATE
rs1044454193	chr10	93567045	FFAR4	17.85	0.13112	NON_SYNONYMOUS		MODERATE
rs1044498	chr6	131851228	ENPP1	12.06	0.172496	NON_SYNONYMOUS		MODERATE
rs1044548	chr6	131890623	ENPP1	10.33	NA	3PRIME_UTR	Benign;Benign	
rs1044737470	chr5	96412463	PCSK1	32	0.952236	NON_SYNONYMOUS		MODERATE
rs1045328134	chr8	37966251	ADRB3	28	0.754983	NON_SYNONYMOUS		MODERATE
rs1045550142	chr16	54040160	FTO	13.23	NA	NONCODING_CHANGE		MODIFIER
rs1046537667	chr10	93587390	FFAR4	14.89	NA	SYNONYMOUS		LOW;LOW;
rs1046595093	chr5	71719400	CARTPT	28.5	0.674017	NON_SYNONYMOUS		MODERATE
rs10467147	chr12	40373560	LRRK2	10.39	NA	DOWNSTREAM		MODIFIER
rs10468281	chr16	53892854	FTO	10.89	NA	INTRONIC		MODIFIER
rs1047063799	chr11	74005885	UCP3	23	0.914394	NON_SYNONYMOUS		MODERATE
rs1048093688	chr6	131858703	ENPP1	27.2	0.865041	NON_SYNONYMOUS		MODERATE
rs1048624393	chr5	96394998	PCSK1	18.08	0.806494	NON_SYNONYMOUS		MODERATE
rs104894319	chr11	74005844	UCP3	38	0.277287	STOP_GAINED	Pathogenic	HIGH;HIGH
rs1049269132	chr5	96432931	PCSK1	25.3	0.892915	NON_SYNONYMOUS		MODERATE
rs10492872	chr16	54086418	FTO	16.38	NA	INTRONIC		
rs1049385311	chr18	60371460	MC4R	24.8	0.92239	NON_SYNONYMOUS		MODERATE
rs1049385311	chr18	60371460	MC4R	26	0.922059	NON_SYNONYMOUS	Uncertain sign.	MODERATE

Data sources:

VarGen Workflow



VarGen Workflow



Data sources: OMIM

Online Mendelian Inheritance in Man

- Database of **gene – phenotype** interactions
  - 7,324 phenotypes
  - 16,656 genes

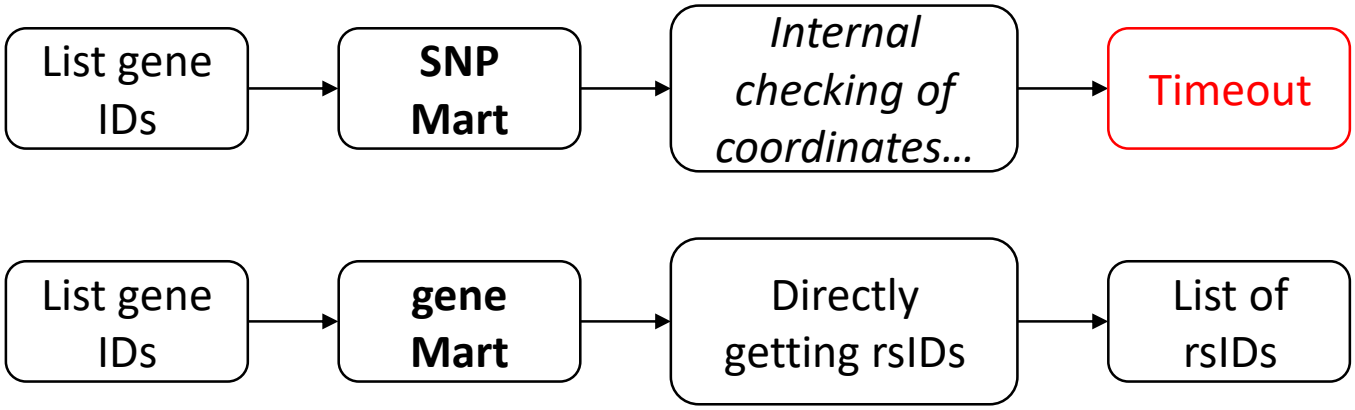
- Based on manual selection and review of the biomedical literature, by curators

- VarGen use the OMIM ID as the starting point of the pipeline, to get a list of genes related to a phenotype.

https://omim.org/

biomaRt timeout

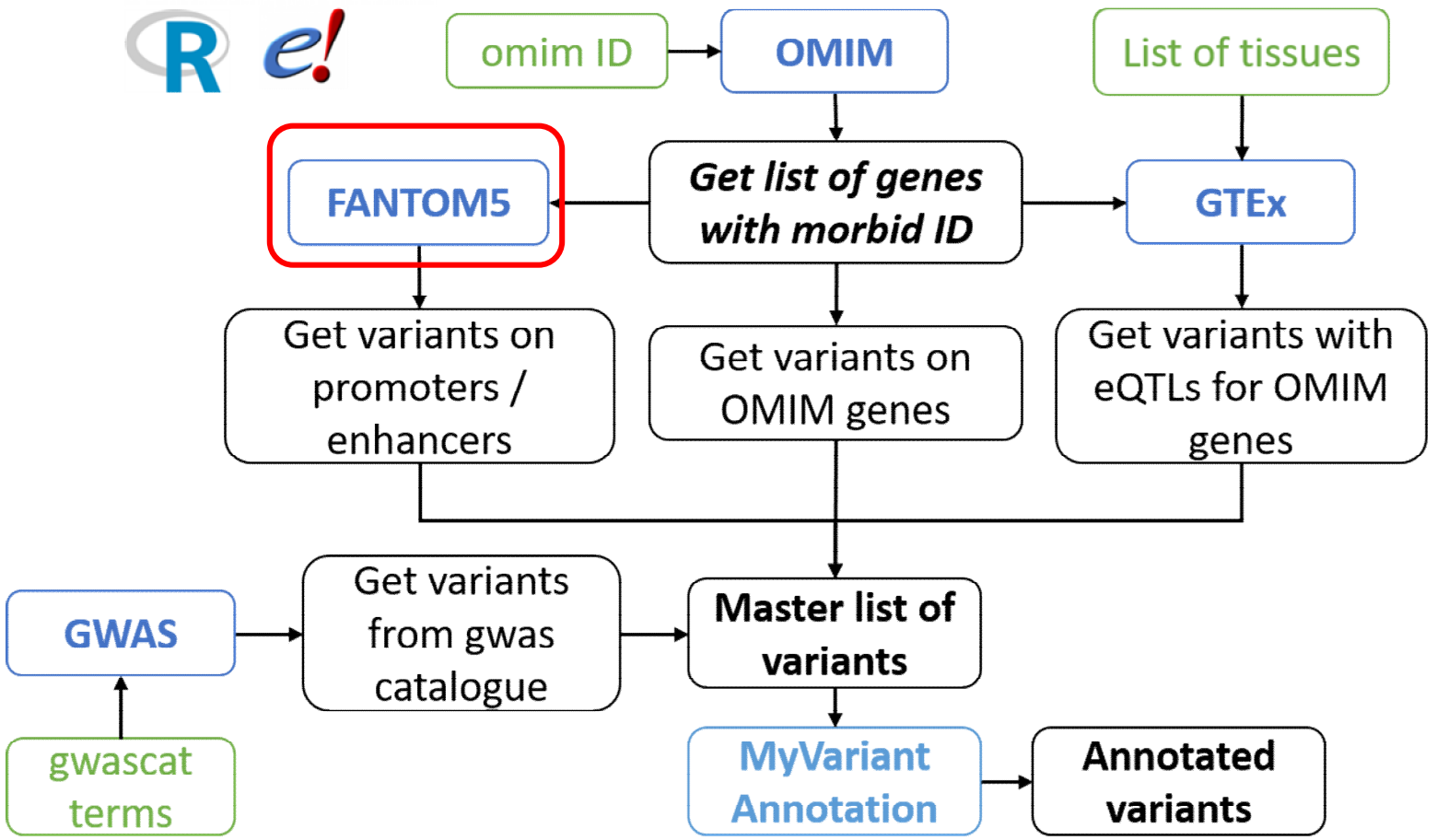
- Goal is to get list of SNPs rsIDs from gene ensembl IDs



⇒ If query is gene centric then it is better to use *gene mart* than *snp mart*

⇒ Inversely, if the query is snp centric (annotation etc…) then it is faster to use *snp mart*

VarGen Workflow



Data sources: FANTOM5

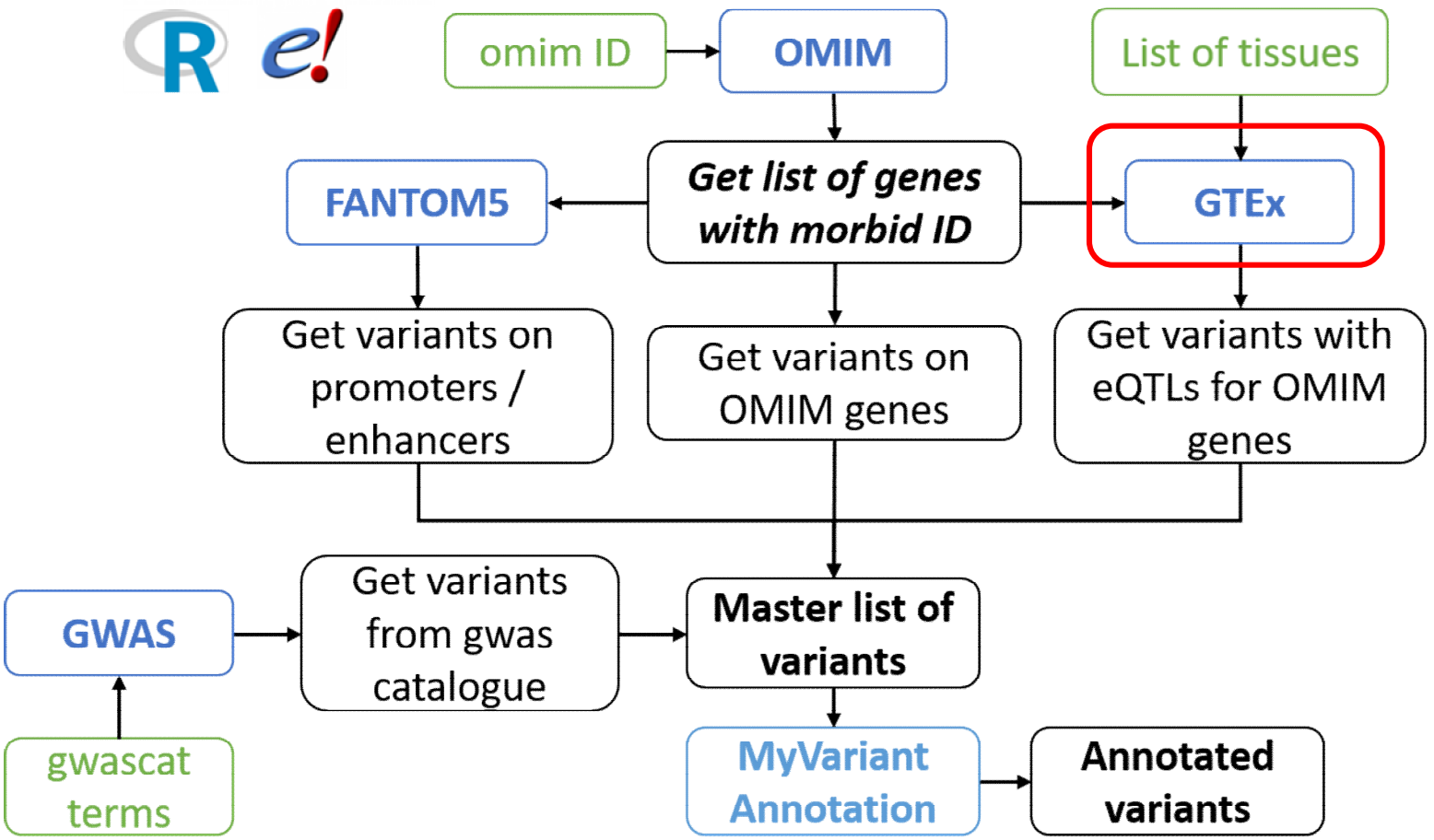
Functional ANnoTation Of the Mammalian genome 5

- Library of enhancers and promoters, based on Cap Analysis of Gene Expression (CAGE) data

- VarGen use FANTOM5 to get variants on enhancers of the OMIM genes (variants which could affect gene expression).

A promoter-level mammalian expression atlas
https://www.nature.com/articles/nature13182

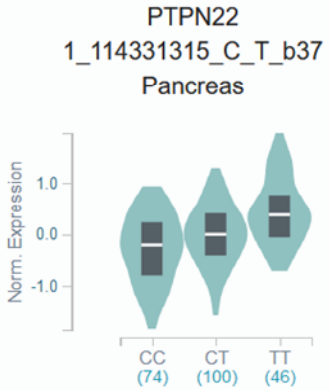
VarGen Workflow

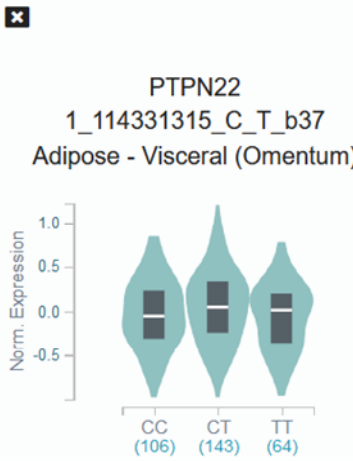


Data sources: Genotype-Tissue Expression

- Public resource to study tissue-specific gene expression and regulation.

- Samples from :
  - 54 non-diseased tissue sites
  - 1,000 individuals
  - Primarily for molecular assays including WGS, WES and RNA-Seq.





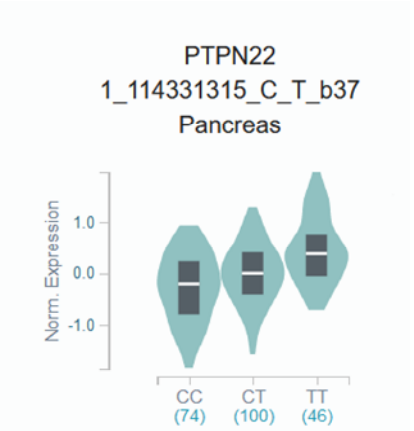
GTEx Consortium, 'The GTEx Consortium atlas of genetic regulatory effects across human tissues', *Science*, vol. 369, no. 6509, pp. 1318–1330, Sep. 2020, doi: 10.1126/science.aaz1776.

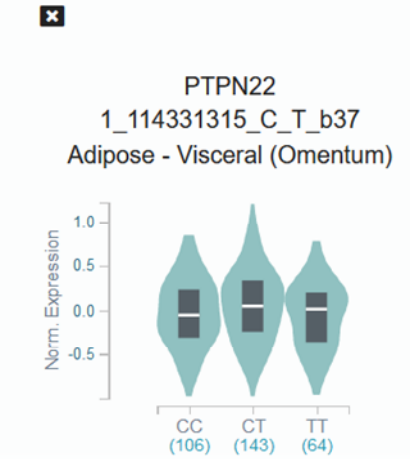
Data sources: Genotype-Tissue Expression

- Public resource to study tissue-specific gene expression and regulation.

- Samples from :
  - 54 non-diseased tissue sites
  - 1,000 individuals
  - Primarily for molecular assays including WGS, WES and RNA-Seq.

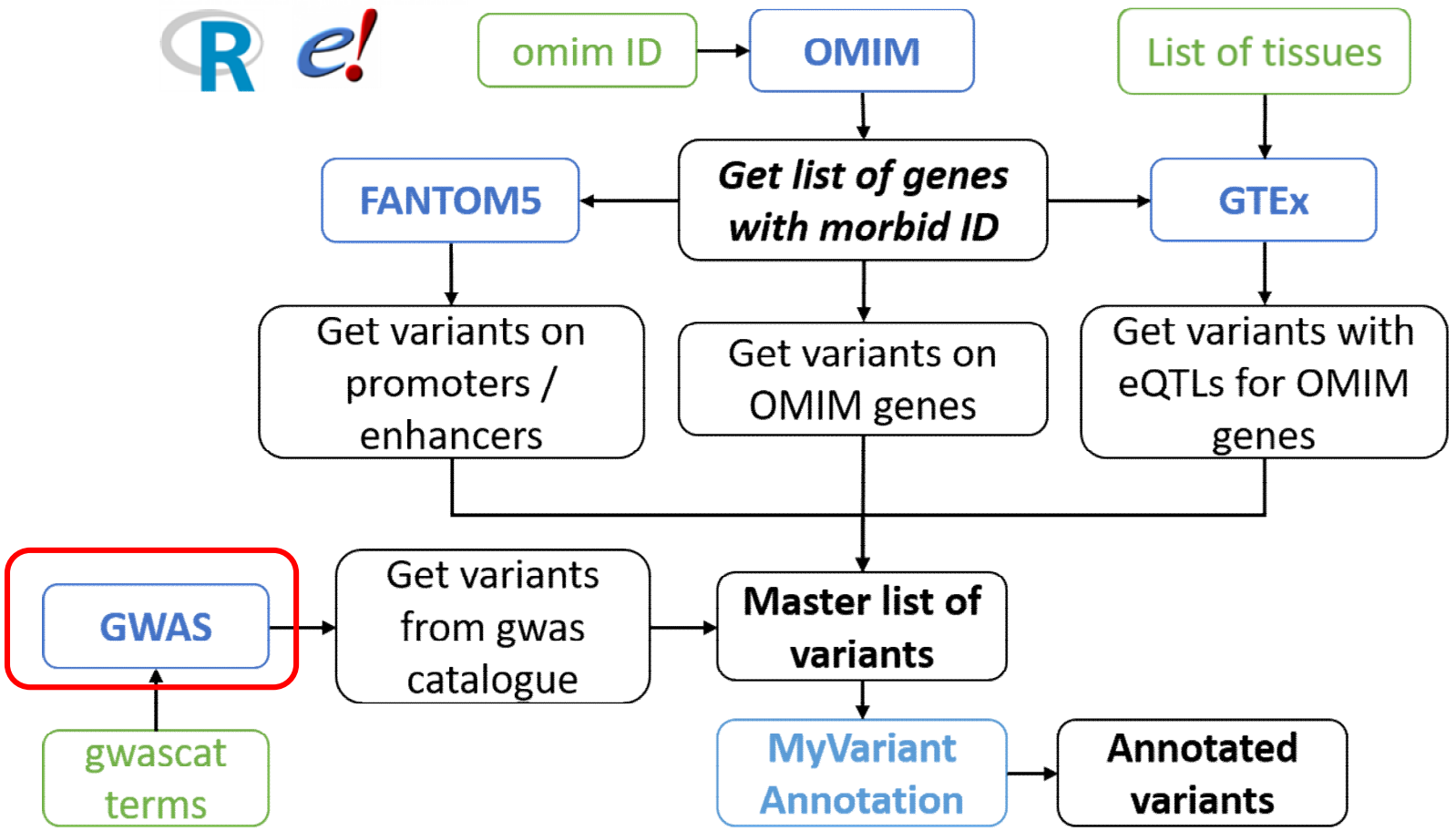
- GTEx v7 is on GRCh37
- GTEx v8 is on GRCh38



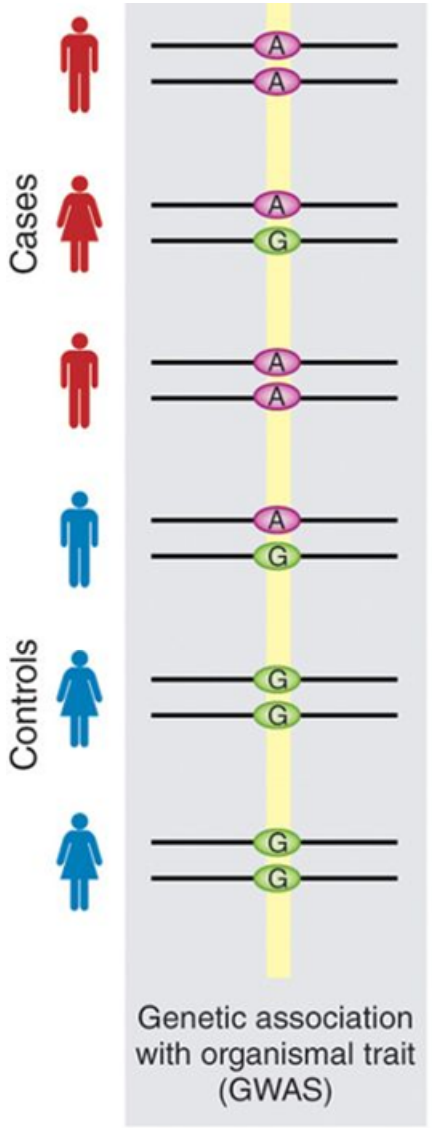


GTEx Consortium, 'The GTEx Consortium atlas of genetic regulatory effects across human tissues', *Science*, vol. 369, no. 6509, pp. 1318–1330, Sep. 2020, doi: 10.1126/science.aaz1776.

VarGen Workflow

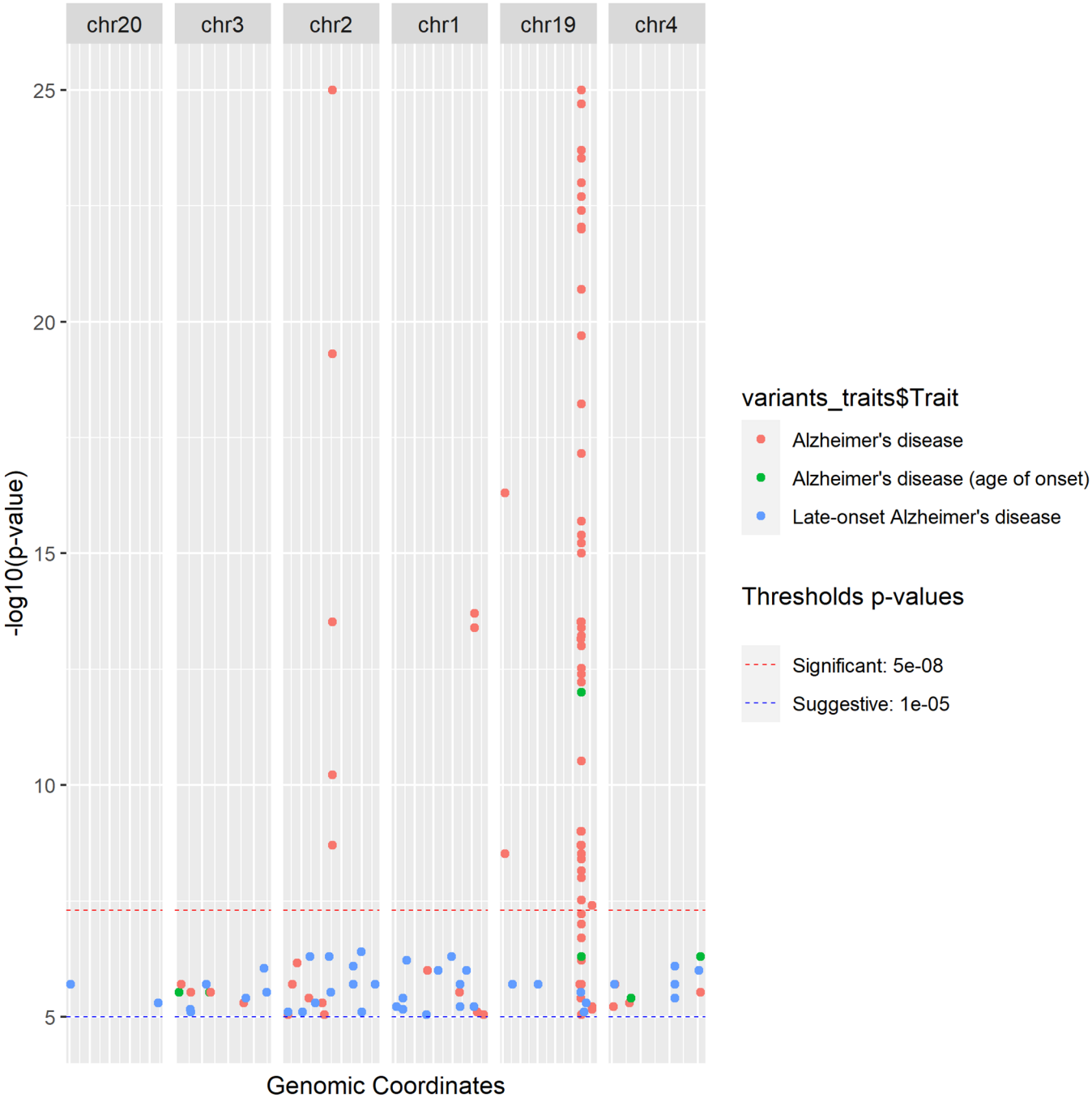


Data sources: GWAS



From: Ward, L. D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, *30*(11), 1095–1106. https://doi.org/10.1038/nbt.2422

Data sources: GWAS

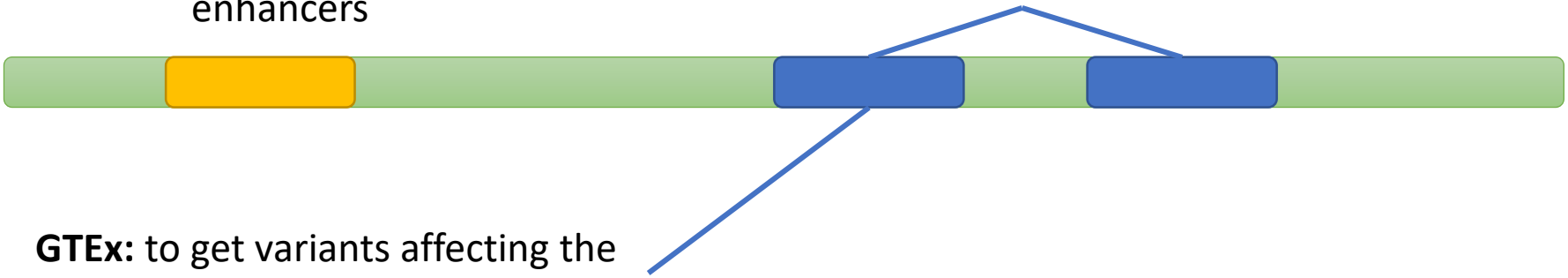


GWAS Catalog

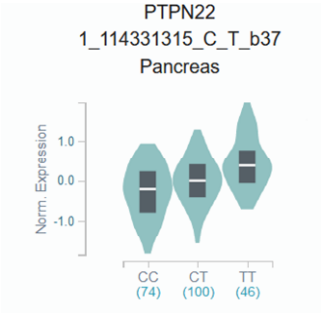
Data sources: summary

FANTOM5: to get the genes enhancers

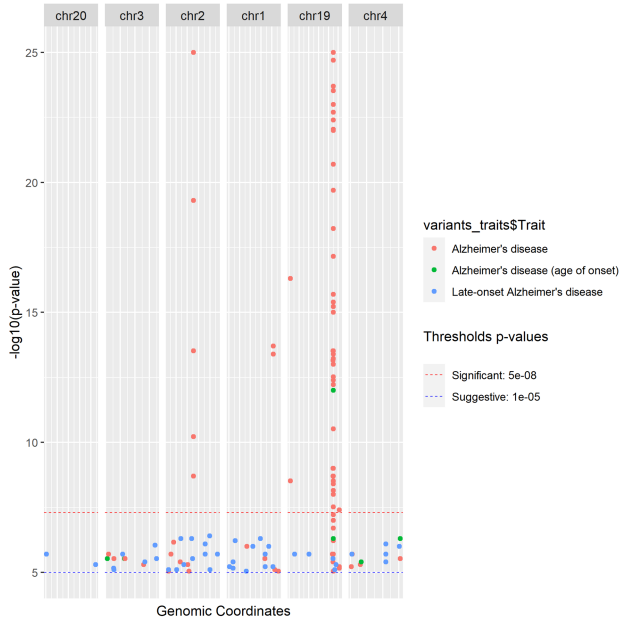
OMIM: to get the genes linked to the phenotype



GTEx: to get variants affecting the genes expression in certain tissues

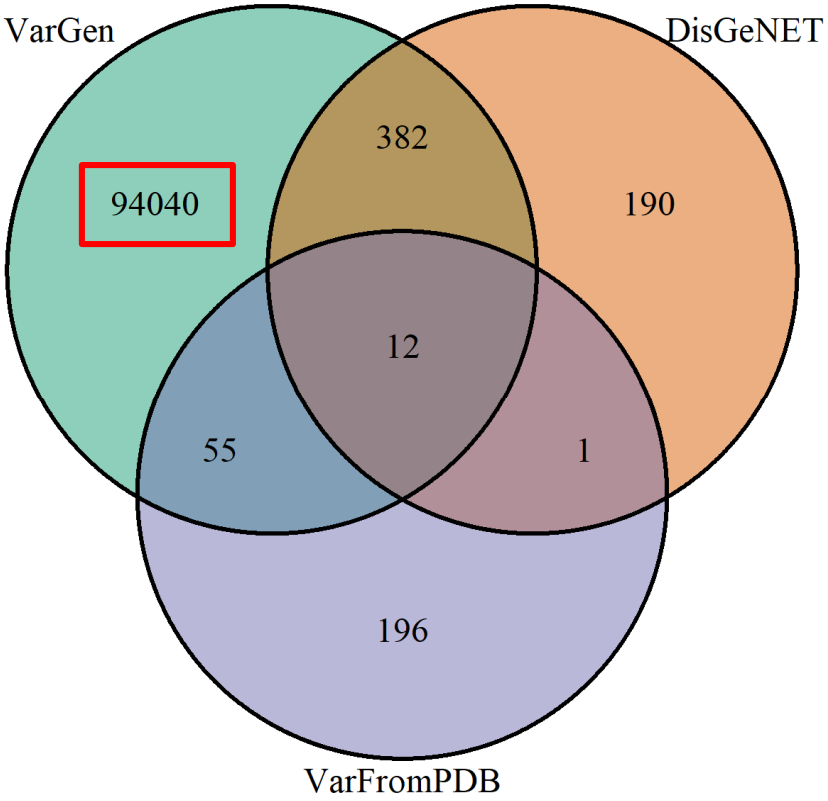


GWAS: variants from large association studies, not necessarily linked to a gene

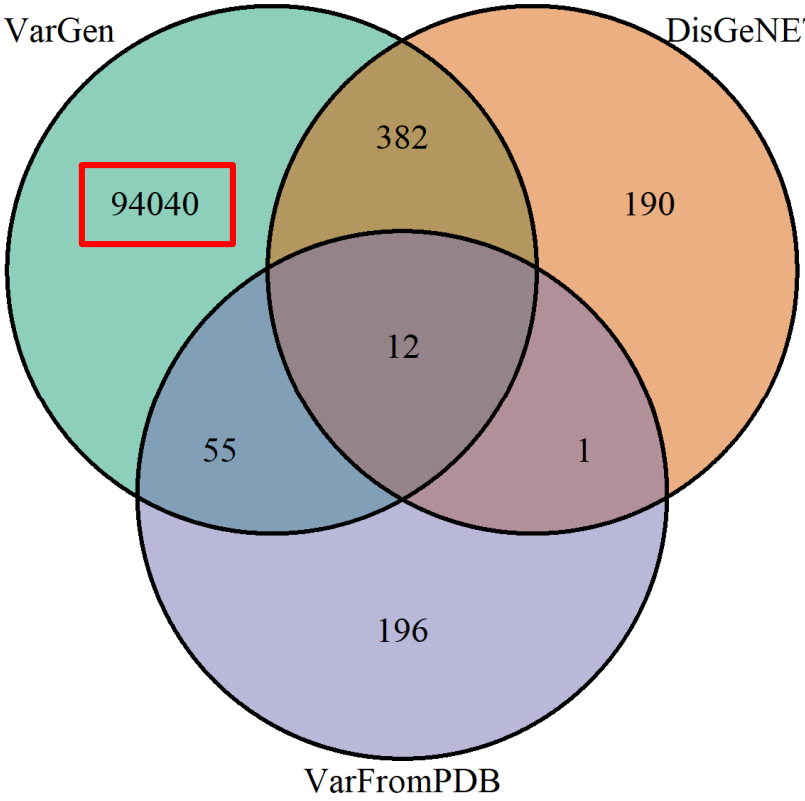


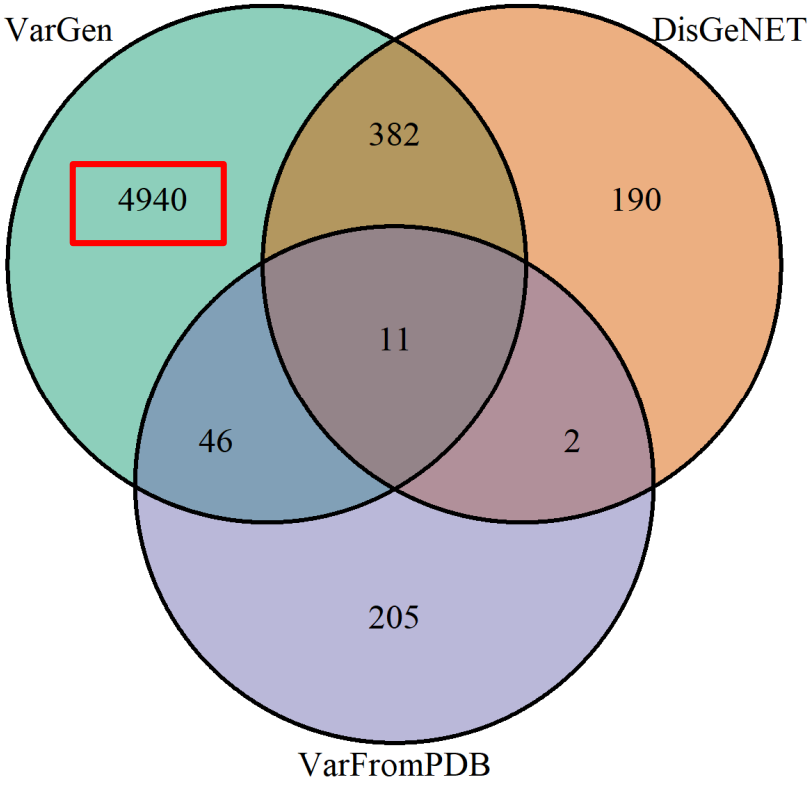
Limitations

Limitation: potential False Positives



First solution: filtering by annotation



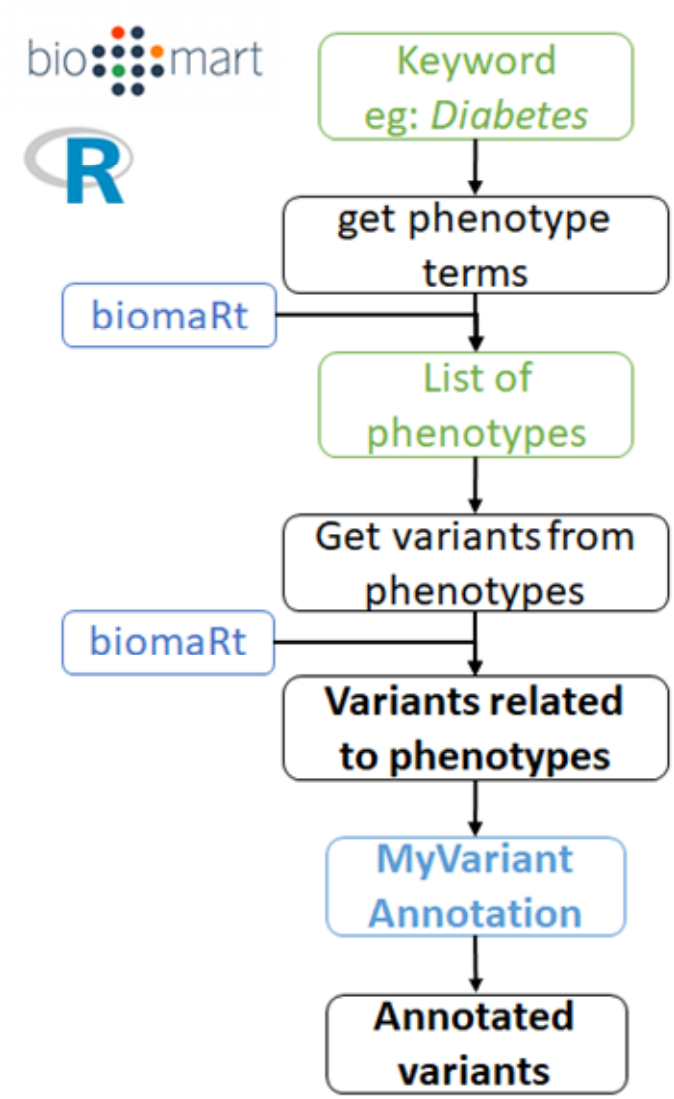


CADD score > 20

With ClinVar significance

All GWAS hits

Second solution: using VarPhen

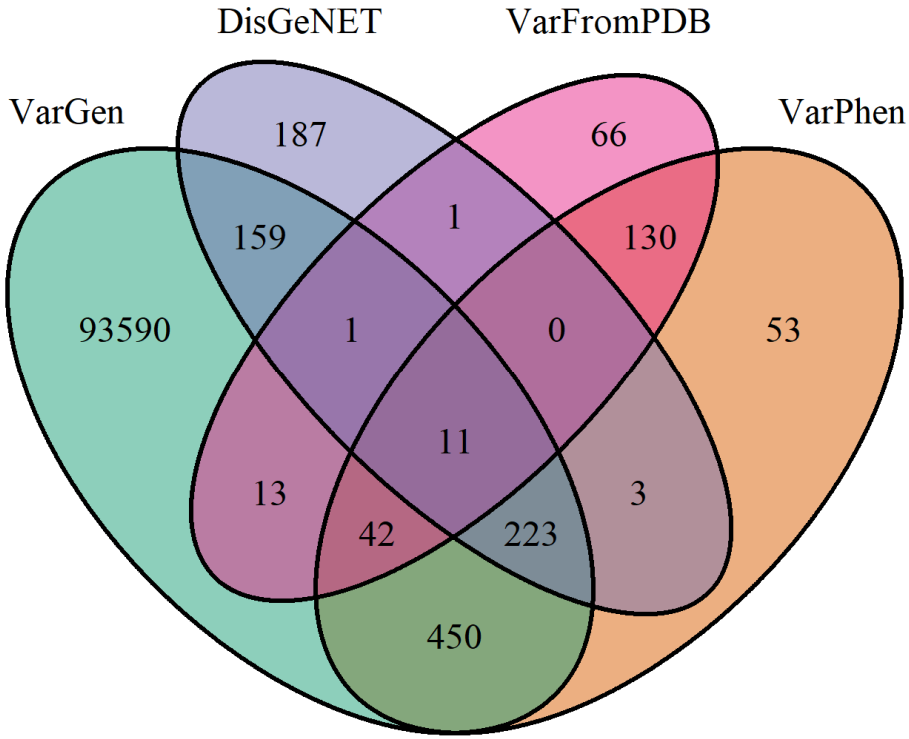


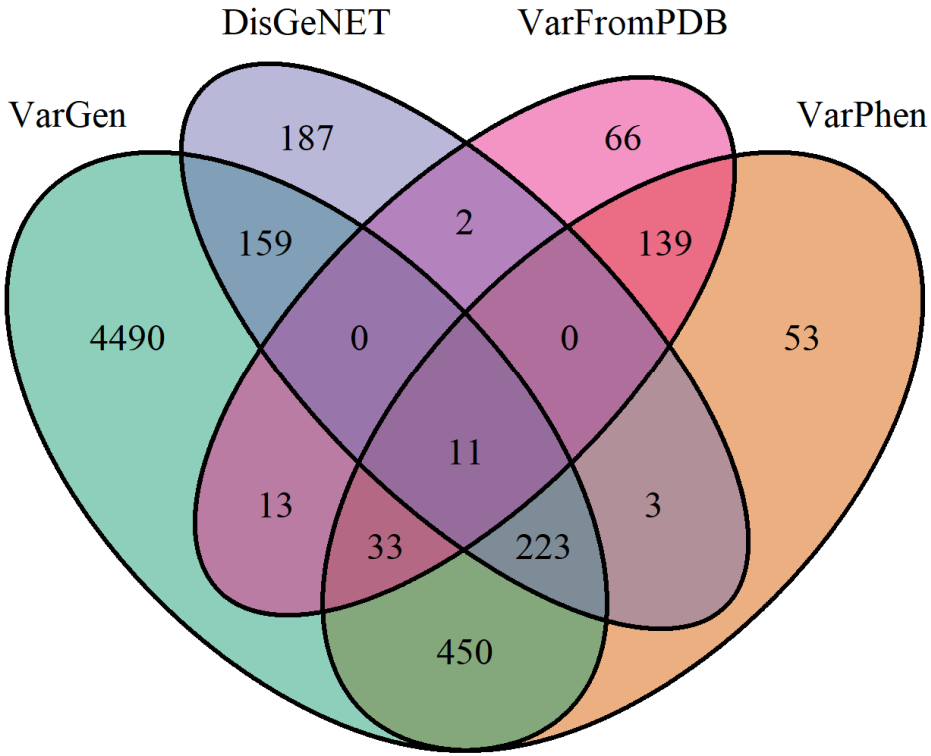
Phenotypes associated with SNPs and other variants come from multiple sources:

- COSMIC (Catalogue Of Somatic Mutations In Cancer)
- ClinVar (Variants of clinical significance from ClinVar)
- dbGaP (The database of Genotypes and Phenotypes)
- EGA (European Genome-phenome Archive)
- GIANT (Genetic Investigation of ANthropometric Traits)
- HGMD-Public (The Human Gene Mutation Database)
- MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium)
- NHGRI-EBI GWAS catalog

Source: https://www.ensembl.org/Help/View?id=299

Second solution: using VarPhen





CADD score > 20

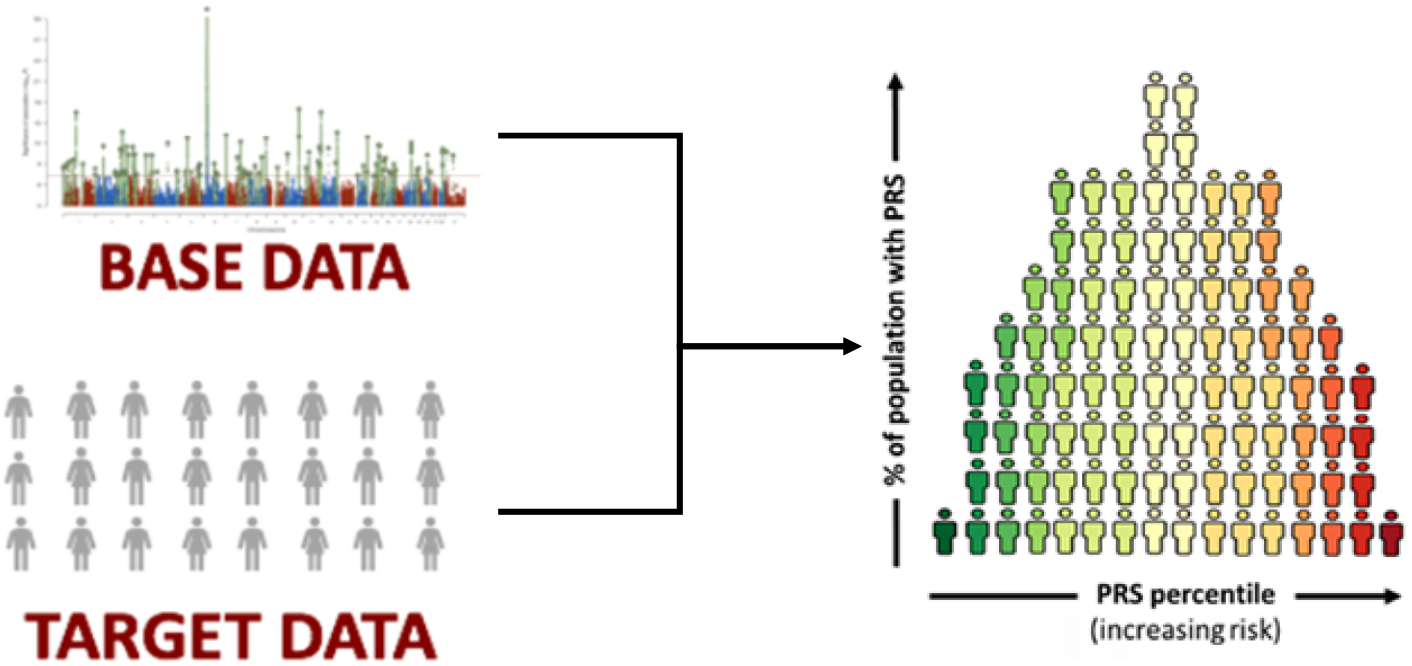
With ClinVar significance

All GWAS hits

Application to a Polygenic Risk Score analysis

What is a PRS model?

- Complex diseases are due to **the accumulation of a lot of variants**, **each having a small impact** on the disease.
- Knowing the variants that are linked to a disease is crucial to the understanding and treatment of the disease. (e.g.: precision medicine)



A. V. Khera *et al.*, '*Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations*', *Nat Genet*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018, doi: 10.1038/s41588-018-0183-z

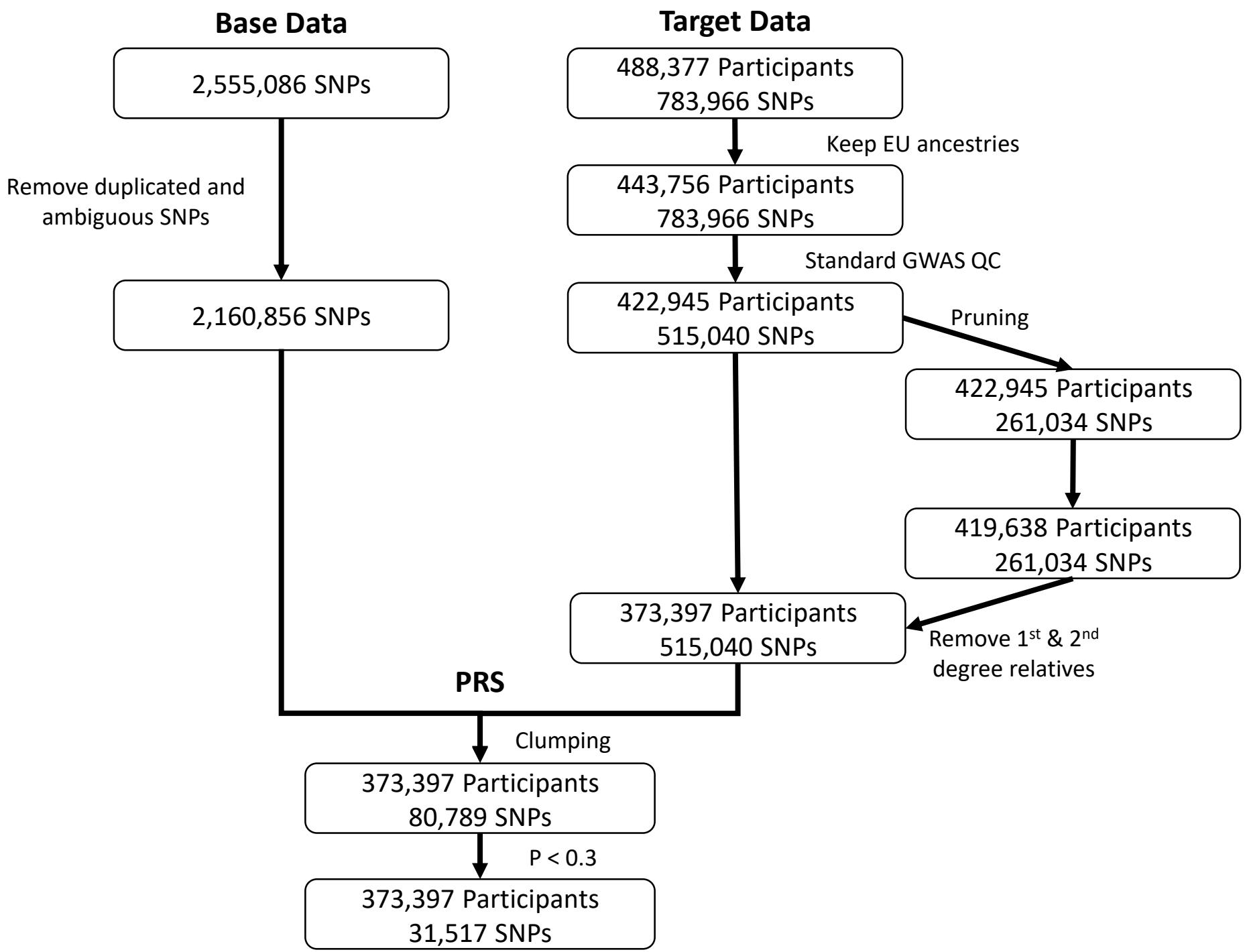
PRS for Body Mass Index

- *Base set:*
  - meta-analysis of 82 GWAS and 43 Metabochip studies of BMI, performed by Locke et al. (GWAS catalog ID: GCST002783)

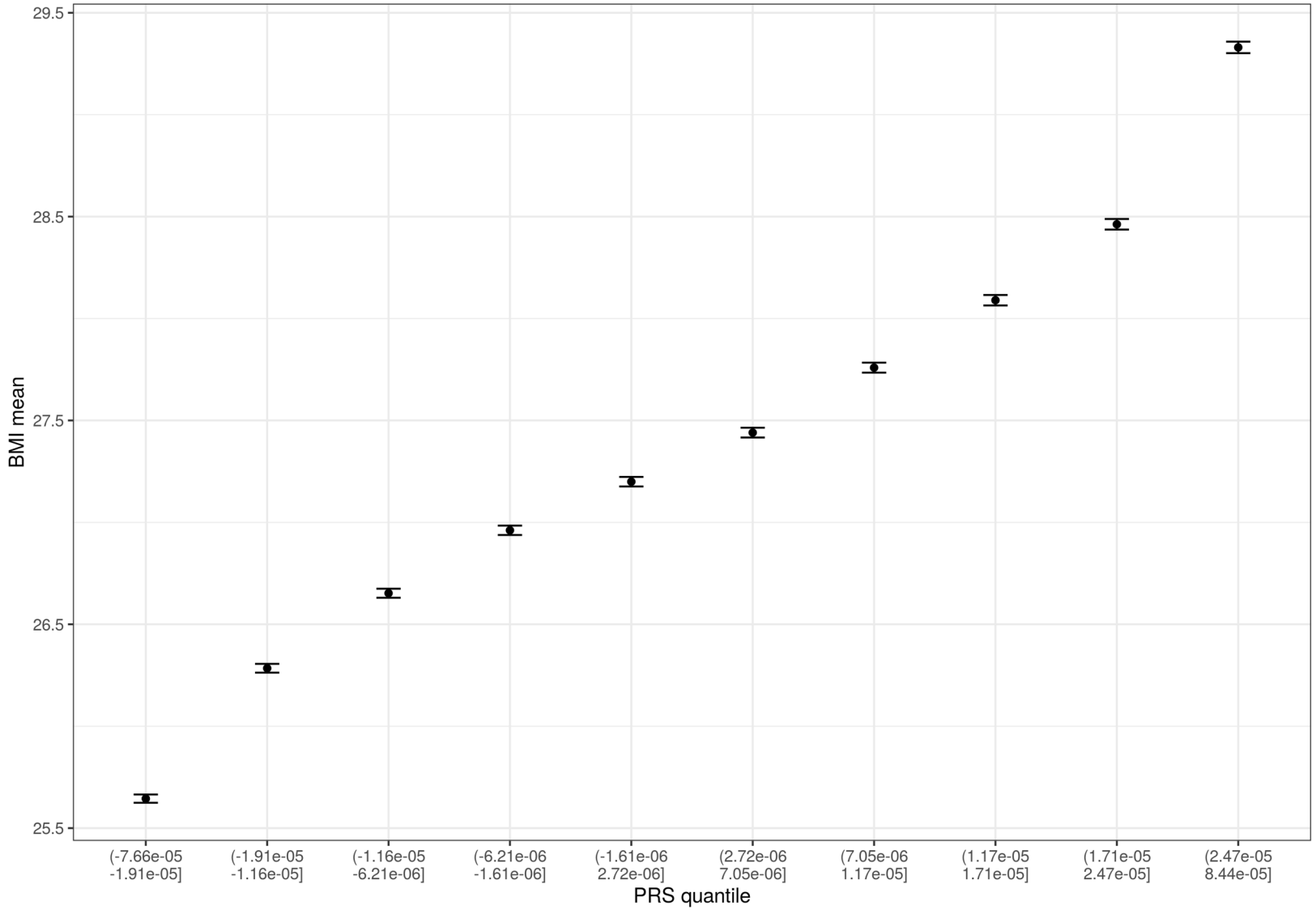
A. E. Locke *et al.*, 'Genetic studies of body mass index yield new insights for obesity biology', *Nature*, vol. 518, no. 7538, Art. no. 7538, Feb. 2015, http://doi.org/10.1038/nature14177

- *Target set*:
  - Filtered UK Biobank

Number of individuals	373,397	
Age at recruitment (years)	Median IQR (Q1-Q3)	58 12 (51 – 63)
Gender	Male	46% (n = 172,264)
BMI	Median IQR (Q1-Q3)	26.7 5.7 (24.1 - 29.8)


Base Data
Target Data
2,555,086 SNPs
488,377 Participants
783,966 SNPs
Keep EU ancestries
443,756 Participants
783,966 SNPs
Remove duplicated and ambiguous SNPs
Standard GWAS QC
2,160,856 SNPs
422,945 Participants
515,040 SNPs
Pruning
422,945 Participants
261,034 SNPs
419,638 Participants
261,034 SNPs
373,397 Participants
515,040 SNPs
Remove 1st & 2nd degree relatives
PRS
Clumping
373,397 Participants
80,789 SNPs
P < 0.3
373,397 Participants
31,517 SNPs

PRS quantiles on BMI

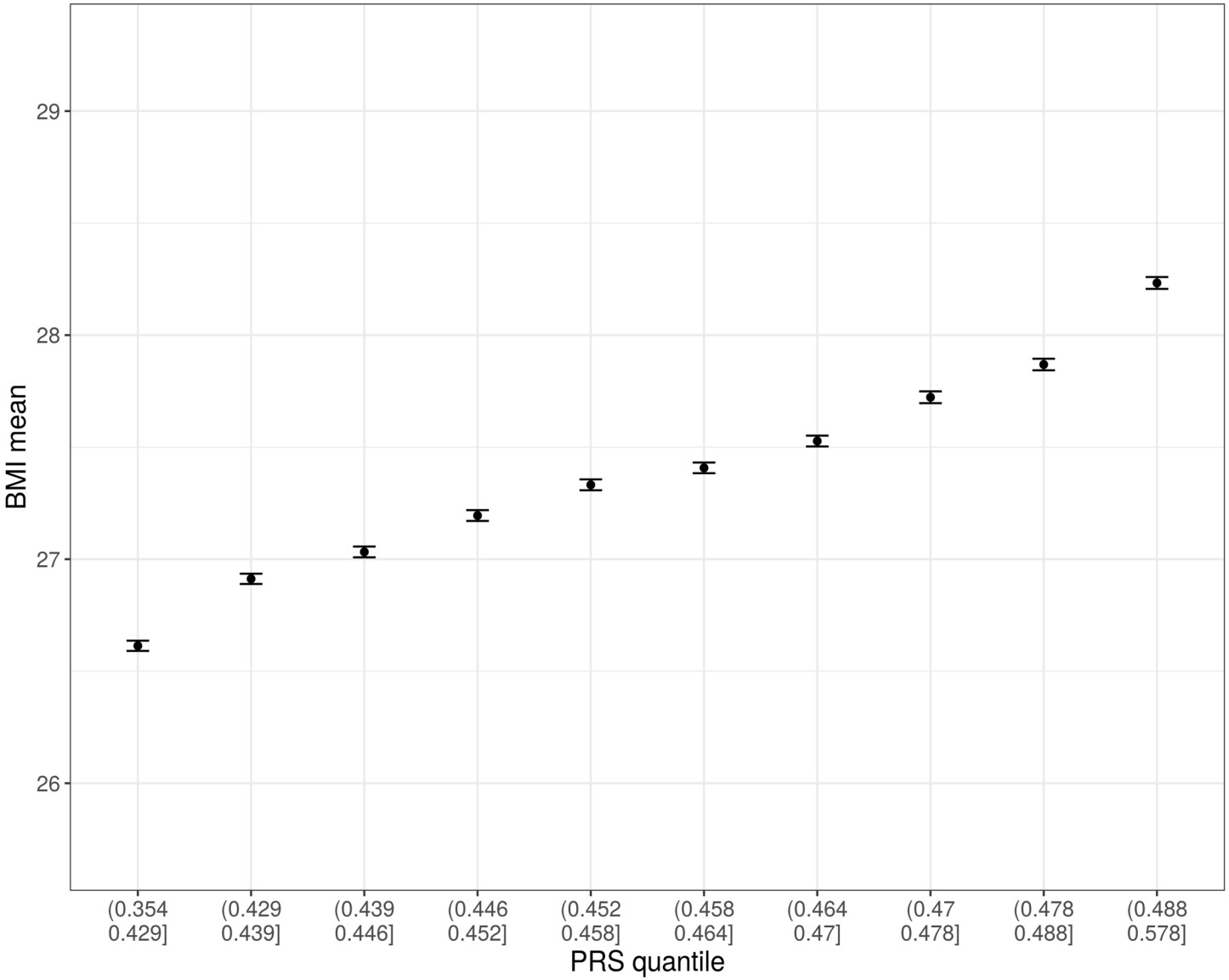


The PRS is informative but only based on GWAS variants

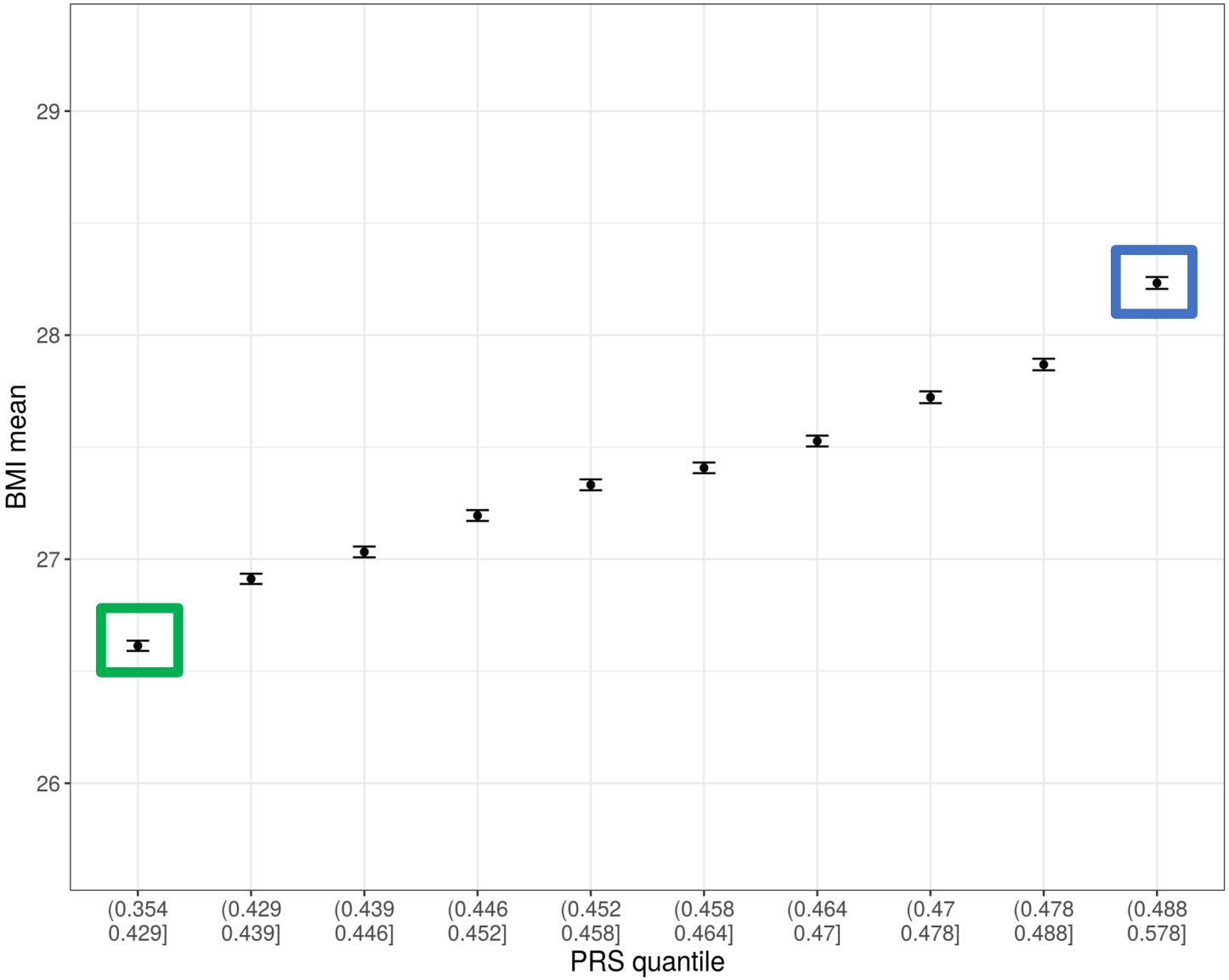
- By design, variants found through GWAS analyses are common in the population.

- But, rarer variants can have a high impact:
  - The idea is to use VarPhen to get these rarer variants and adjust the PRS risk

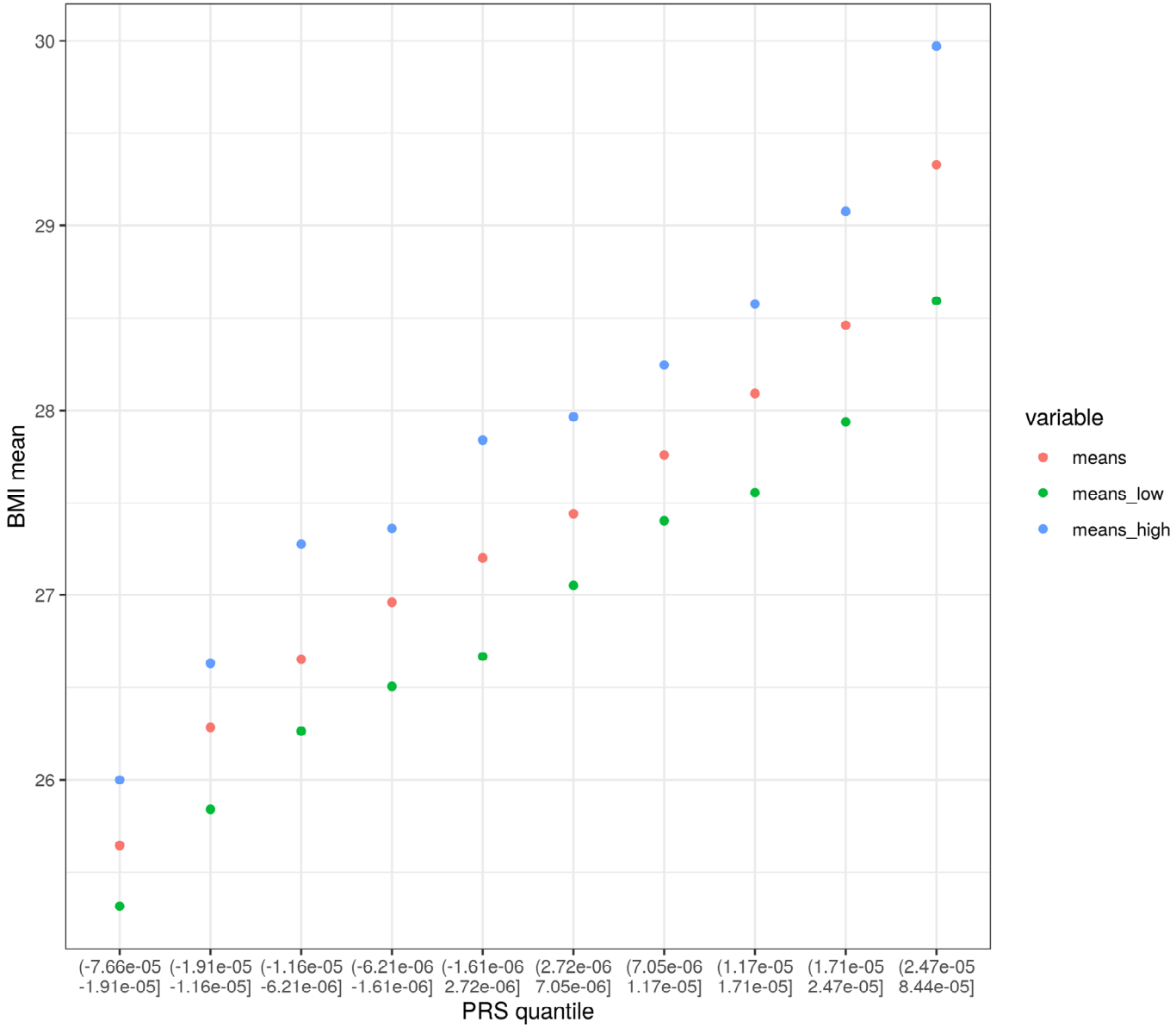
2nd PRS based on VarPhen (287 variants)



Selected the individuals from the highest and lowest deciles: assigned them to *high* and *low* groups respectively (so ~30.000 individuals / group)



Optimisation results



Further reading

- F. W. Albert and L. Kruglyak, ‘The role of regulatory variation in complex traits and disease’, *Nature Reviews Genetics*, vol. 16, no. 4, Art. no. 4, Apr. 2015, https://doi.org/10.1038/nrg3891

- A. V. Khera *et al.*, ‘Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations’, *Nat Genet*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018, https://doi.org/10.1038/s41588-018-0183-z

- Ferrero E. Using regulatory genomics data to interpret the function of disease variants and prioritise genes from expression studies [version 2; peer review: 2 approved]. *F1000Research* 2018, **7**:121 (https://doi.org/10.12688/f1000research.13577.2)

- Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, Laura I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Research*, Volume 45, Issue D1, January 2017, Pages D833–D839, https://doi.org/10.1093/nar/gkw943

- GTEx Consortium, ‘The GTEx Consortium atlas of genetic regulatory effects across human tissues’, *Science*, vol. 369, no. 6509, pp. 1318–1330, Sep. 2020, https://doi.org/10.1126/science.aaz1776

Any Questions ?

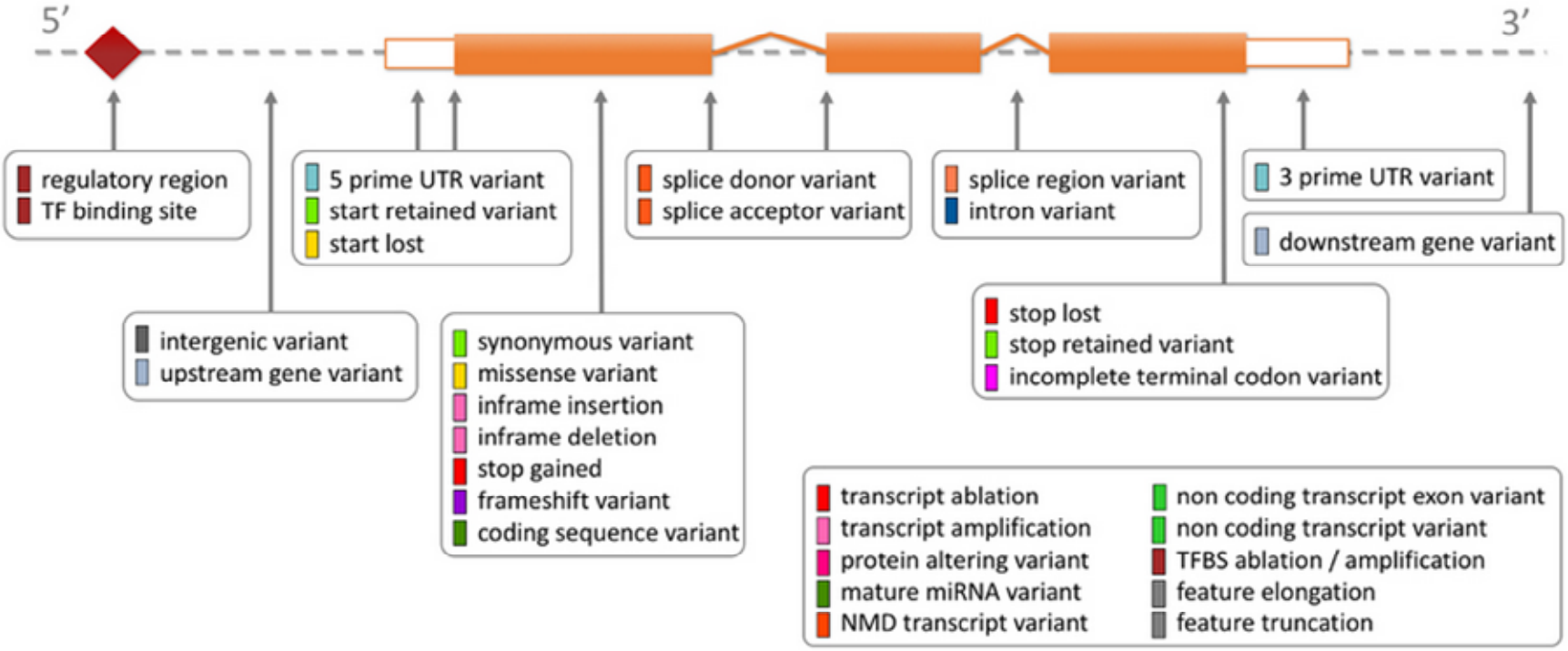
rsid	chr	pos	hgnc_symbol	cadd_phred	fathmm_xf_score	annot_type	consequence	clinical_significance	snpeff_ann
rs17821714	chr16	53921044	FTO	16.870	NA	Transcript	INTRONIC		
rs17823912	chr16	53987891	FTO	11.630	NA	Transcript	INTRONIC		
rs17847050	chr6	131860393	ENPP1	25.900	0.944065	CodingTranscript	NON_SYNONYMOUS	Uncertain significance...	MODERATE
rs17847261	chr20	56249663	MC3R	25.900	0.938632	CodingTranscript	NON_SYNONYMOUS		MODERATE
rs17848368	chr11	74006298	UCP3	33.000	0.725808	CodingTranscript	NON_SYNONYMOUS	Pathogenic	MODERATE;MODERATE
rs17848372	chr11	74005915	UCP3	35.000	0.927409	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs17854409	chr20	62860142	TCFL5	19.450	0.109146	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs17854409	chr20	62860142	TCFL5	24.900	0.148471	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs1799904	chr5	96429259	PCSK1	25.900	0.746210	CodingTranscript;T...	NON_SYNONYMOUS;...		MODERATE;MODERATE;MODI...
rs1799904	chr5	96429259	PCSK1	25.600	0.487390	CodingTranscript;T...	NON_SYNONYMOUS;...		MODERATE;MODERATE;MODI...
rs17085675	chr5	96391960	PCSK1	1.702	NA	RegulatoryFeature...	REGULATORY;3PRIME_...	Likely benign;Likely be...	MODIFIER;MODIFIER;MODIFIER
rs17085675	chr5	96391960	PCSK1	1.702	NA	RegulatoryFeature...	REGULATORY;3PRIME_...	Likely benign;Likely be...	MODIFIER;MODIFIER;MODIFIER
rs17782313	chr18	60183864	RNU4-17P - A...	2.607	NA	Intergenic	INTERGENIC	drug response	MODIFIER
rs17847050	chr6	131860393	ENPP1	25.900	0.944065	CodingTranscript	NON_SYNONYMOUS	Uncertain significance...	MODERATE
rs17848368	chr11	74006298	UCP3	33.000	0.725808	CodingTranscript	NON_SYNONYMOUS	Pathogenic	MODERATE;MODERATE
rs17011478	chr2	75565017	EVA1A	13.030	NA	Transcript	INTRONIC		MODIFIER;MODIFIER
rs1701930	chr19	50916290	KLK4 - PPIAP59	0.333	NA	Intergenic	INTERGENIC		MODIFIER

Corentin Molitor, Matt Brember, Fady Mohareb, VarGen: An R package for disease-associated variant discovery and annotation, Bioinformatics, https://doi.org/10.1093/bioinformatics/btz930

https://github.com/MCorentin/VarGen

Supplementary Slides

VEP consequences



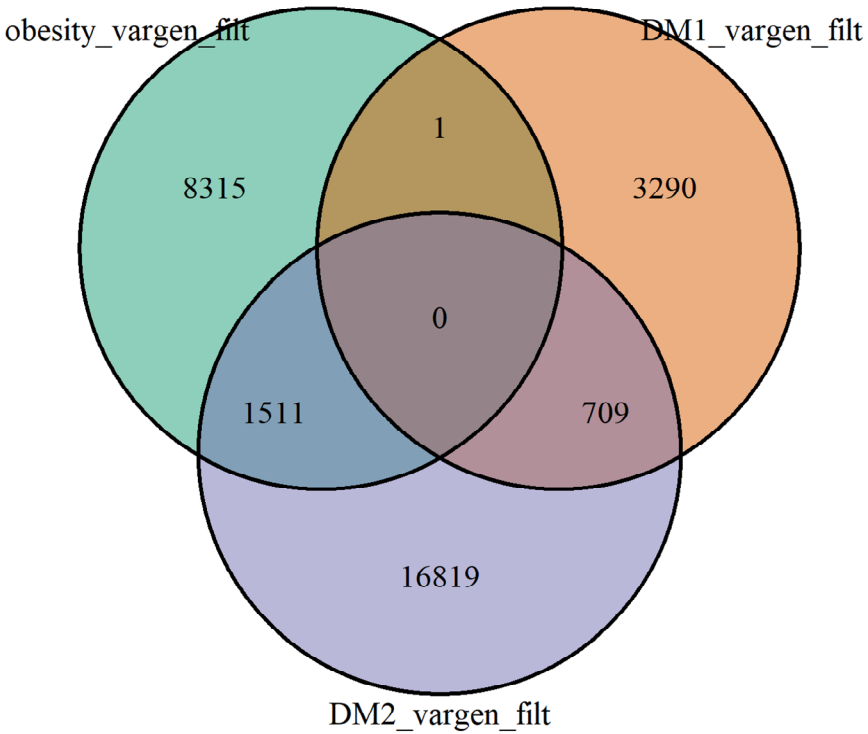
https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

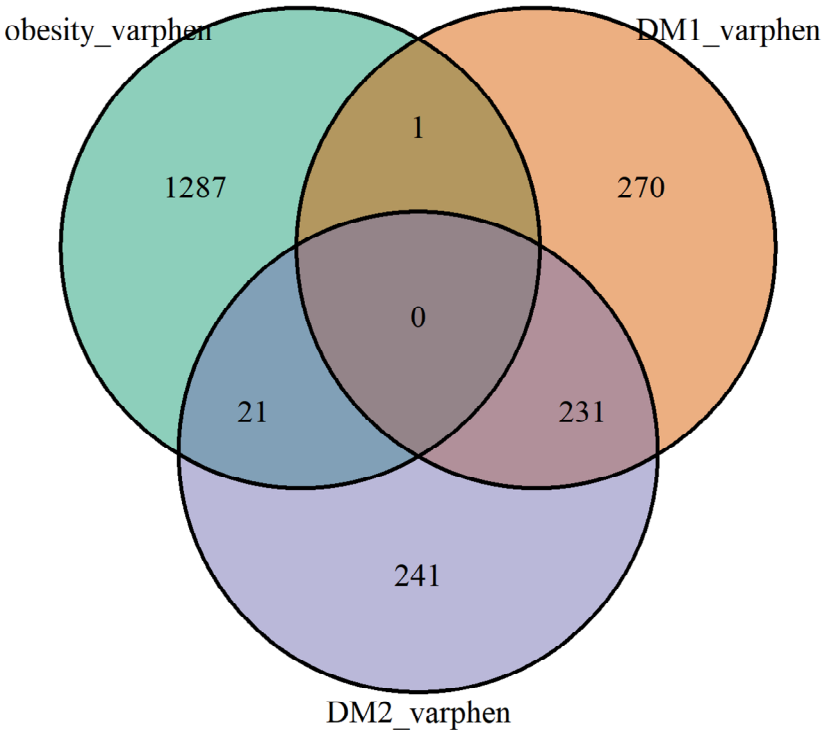
GWAS threshold: Significant vs Suggestive

- **Significant threshold:**
  - P-value < 5*e-08
  - Comes from Bonferroni Correction: “p-value / # comparisons” (here snps) (https://xkcd.com/882/)

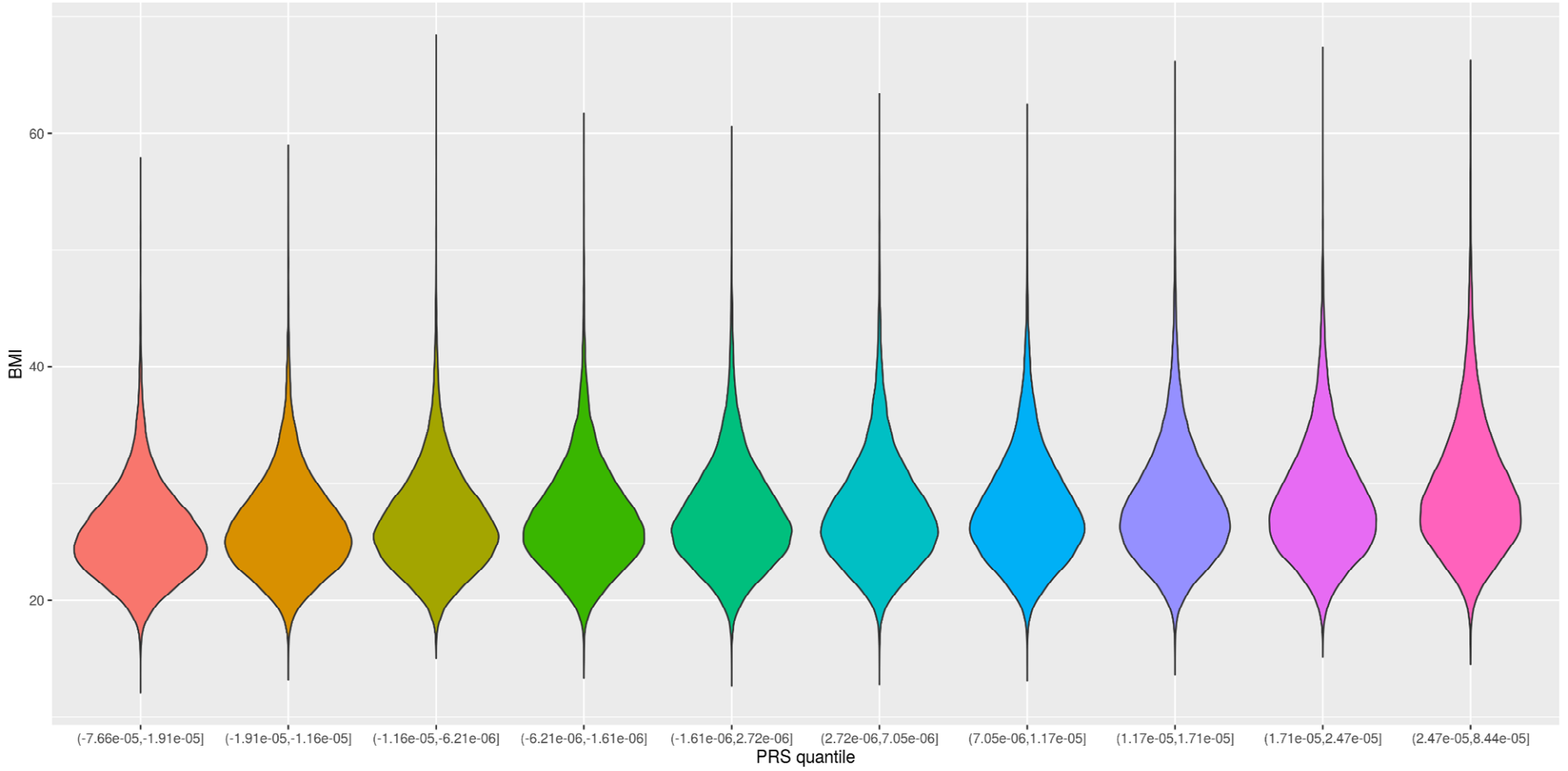
- **Suggestive threshold:**
  - P-value < 1*e-05
  - First proposed by Lander and Kruglyak: represents the threshold where one false positive is expected per genome scan (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5307927/) (https://www.ncbi.nlm.nih.gov/pubmed/7581446)

An interesting result





PRS – BMI distribution for each PRS decile (~37.000 individuals per decile)



Limitations

- The base set was performed mostly on individuals of European ancestry.
  - Causal variants are unlikely to be directly genotyped, as GWAS only identify variants that are in LD with them. This means that the PRS is not generalisable to other ethnicities (as LD patterns are population specific).

- The sex chromosomes were ignored in this analysis:
  - They are particularly challenging to analyse and might lead to misinterpretations
  - The GWAS meta-analysis used as the base data only contained six variants on the X chromosome.

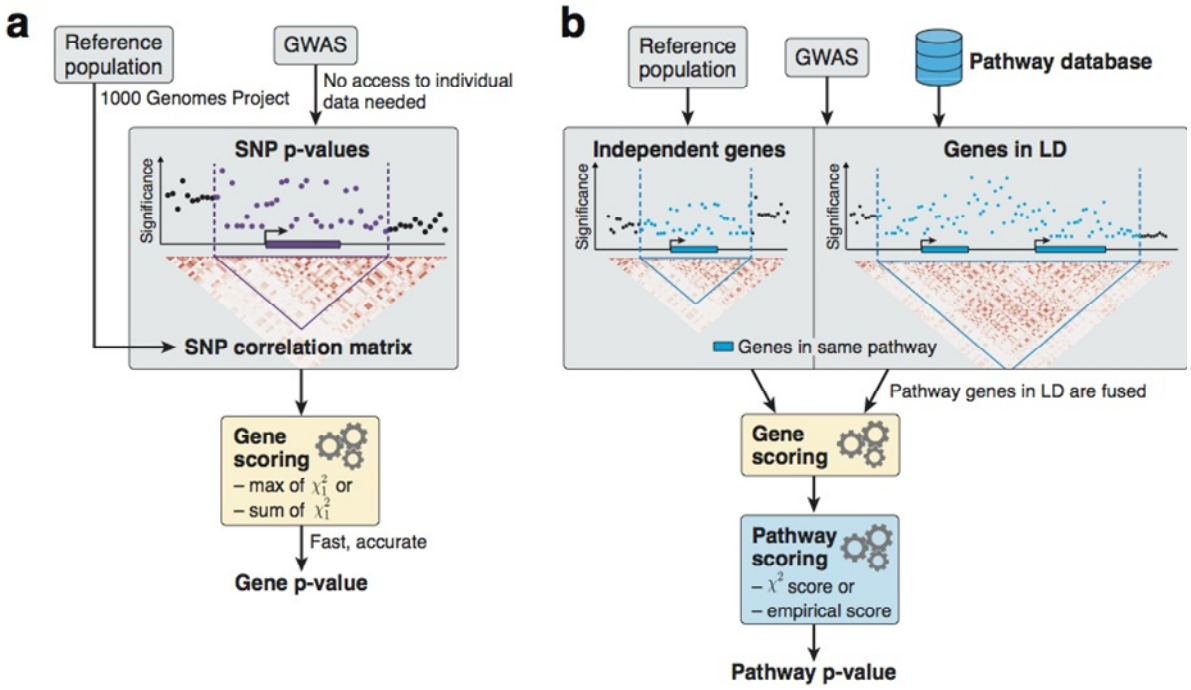
Output example, after the annotation:

(some columns removed for clarity)

rsid	chr	position	HGNC	CADD	FATHMM	consequence	clinVar sign.	snpeff_ann
rs1043543911	chr8	37965344	ADRB3	13.35	0.084341	NON_SYNONYMOUS		MODERATE
rs1043811762	chr11	74006927	UCP3	32	0.89594	NON_SYNONYMOUS		MODERATE
rs1044022939	chr18	60371575	MC4R	28.6	0.574921	NON_SYNONYMOUS		MODERATE
rs1044454193	chr10	93567045	FFAR4	17.85	0.13112	NON_SYNONYMOUS		MODERATE
rs1044498	chr6	131851228	ENPP1	12.06	0.172496	NON_SYNONYMOUS		MODERATE
rs1044548	chr6	131890623	ENPP1	10.33	NA	3PRIME_UTR	Benign;Benign	
rs1044737470	chr5	96412463	PCSK1	32	0.952236	NON_SYNONYMOUS		MODERATE
rs1045328134	chr8	37966251	ADRB3	28	0.754983	NON_SYNONYMOUS		MODERATE
rs1045550142	chr16	54040160	FTO	13.23	NA	NONCODING_CHANGE		MODIFIER
rs1046537667	chr10	93587390	FFAR4	14.89	NA	SYNONYMOUS		LOW;LOW;
rs1046595093	chr5	71719400	CARTPT	28.5	0.674017	NON_SYNONYMOUS		MODERATE
rs10467147	chr12	40373560	LRRK2	10.39	NA	DOWNSTREAM		MODIFIER
rs10468281	chr16	53892854	FTO	10.89	NA	INTRONIC		MODIFIER
rs1047063799	chr11	74005885	UCP3	23	0.914394	NON_SYNONYMOUS		MODERATE
rs1048093688	chr6	131858703	ENPP1	27.2	0.865041	NON_SYNONYMOUS		MODERATE
rs1048624393	chr5	96394998	PCSK1	18.08	0.806494	NON_SYNONYMOUS		MODERATE
rs104894319	chr11	74005844	UCP3	38	0.277287	STOP_GAINED	Pathogenic	HIGH;HIGH
rs1049269132	chr5	96432931	PCSK1	25.3	0.892915	NON_SYNONYMOUS		MODERATE
rs10492872	chr16	54086418	FTO	16.38	NA	INTRONIC		
rs1049385311	chr18	60371460	MC4R	24.8	0.92239	NON_SYNONYMOUS		MODERATE
rs1049385311	chr18	60371460	MC4R	26	0.922059	NON_SYNONYMOUS	Uncertain sign.	MODERATE

Visualisation per gene (Matt Brember)

VarGen + Pascal



- Pascal (Pathway scoring algorithm) is a tool for gene scoring and pathway analysis from GWAS results.

- Here, we adapted the CADD score from VarGen to proxy the p-value (0.1 was chosen to obtain values in a range similar to p-values obtained from GWAS):

$$pascal\ score = \frac{0.1}{CADD\ score}$$

Top 15 pathways for Diabetes type 1

Pathway Name
Biocarta Mitochondria Pathway
Kegg Vascular smooth muscle contraction
Kegg Gap junction
Kegg Long term potentiation
Kegg Long term depression
Kegg Taste transduction
Kegg GNRH signaling pathway
Kegg Alzheimers disease
Reactome DAG and IP3 signaling
Reactome antigen activates B cell receptor leading to generation of second messengers
Reactome Opioid Signalling
Reactome PLC beta mediated events
Reactome Elevation of cytosolic Ca2+ levels
Reactome Regulation of insulin secretion by glucagon-like Peptide-1
Reactome Platelet homeostasis

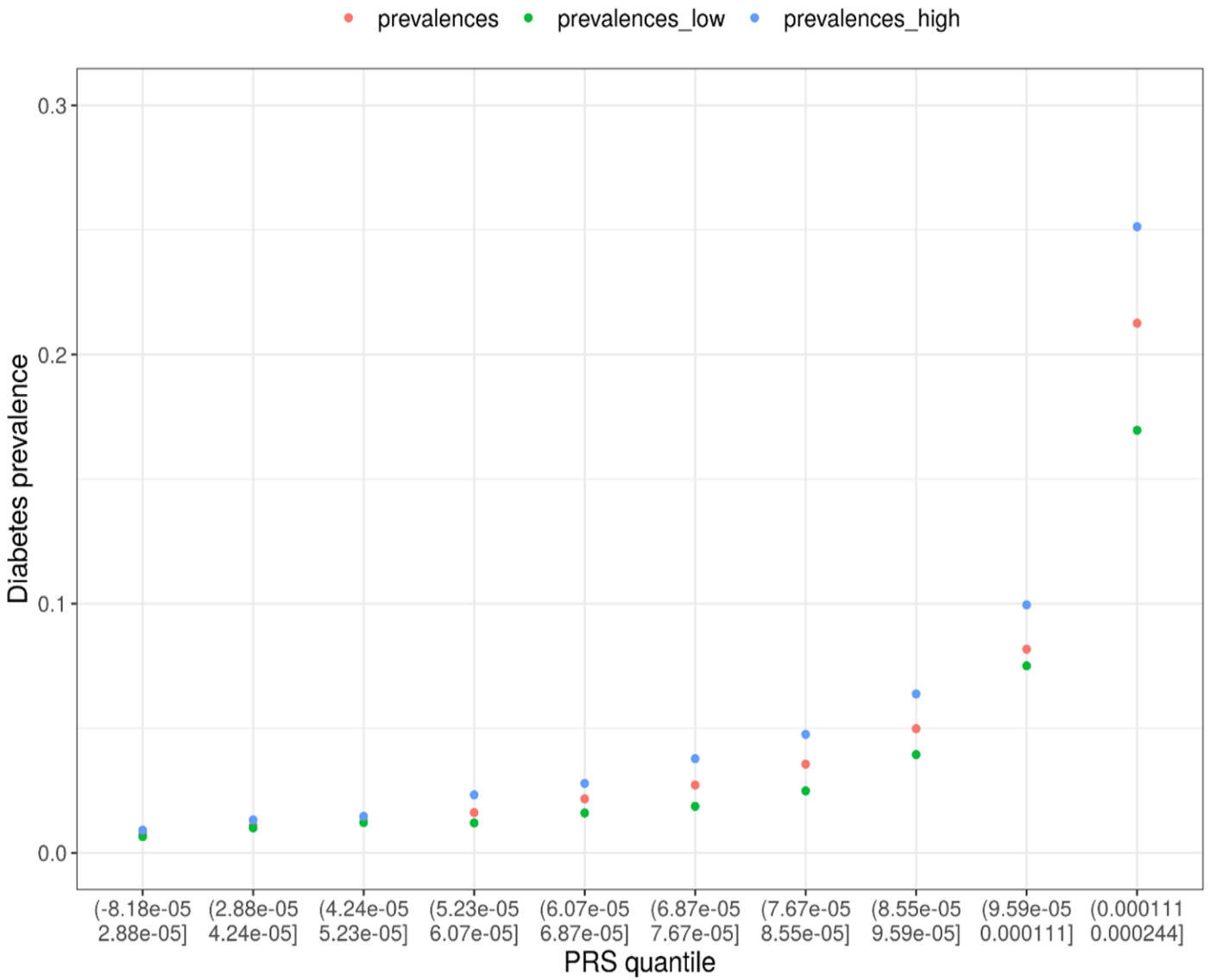
Top 15 pathways for Diabetes type 2

Pathway Name
Reactome Regulation of beta-cell development
Kegg Insulin signaling pathway
Kegg Arrhythmogenic right ventricular cardiomyopathy
Reactome Regulation of gene expression in beta cells
Kegg type II diabetes mellitus
Kegg Melanogenesis
Kegg Thyroid cancer
Kegg Focal adhesion
Kegg Colorectal cancer
Kegg Adherens junction
Reactome Developmental Biology
Reactome Neuronal System
Reactome Nuclear Receptor transcription pathway
Kegg Starch and sucrose metabolism
Kegg Renal cell carcinoma

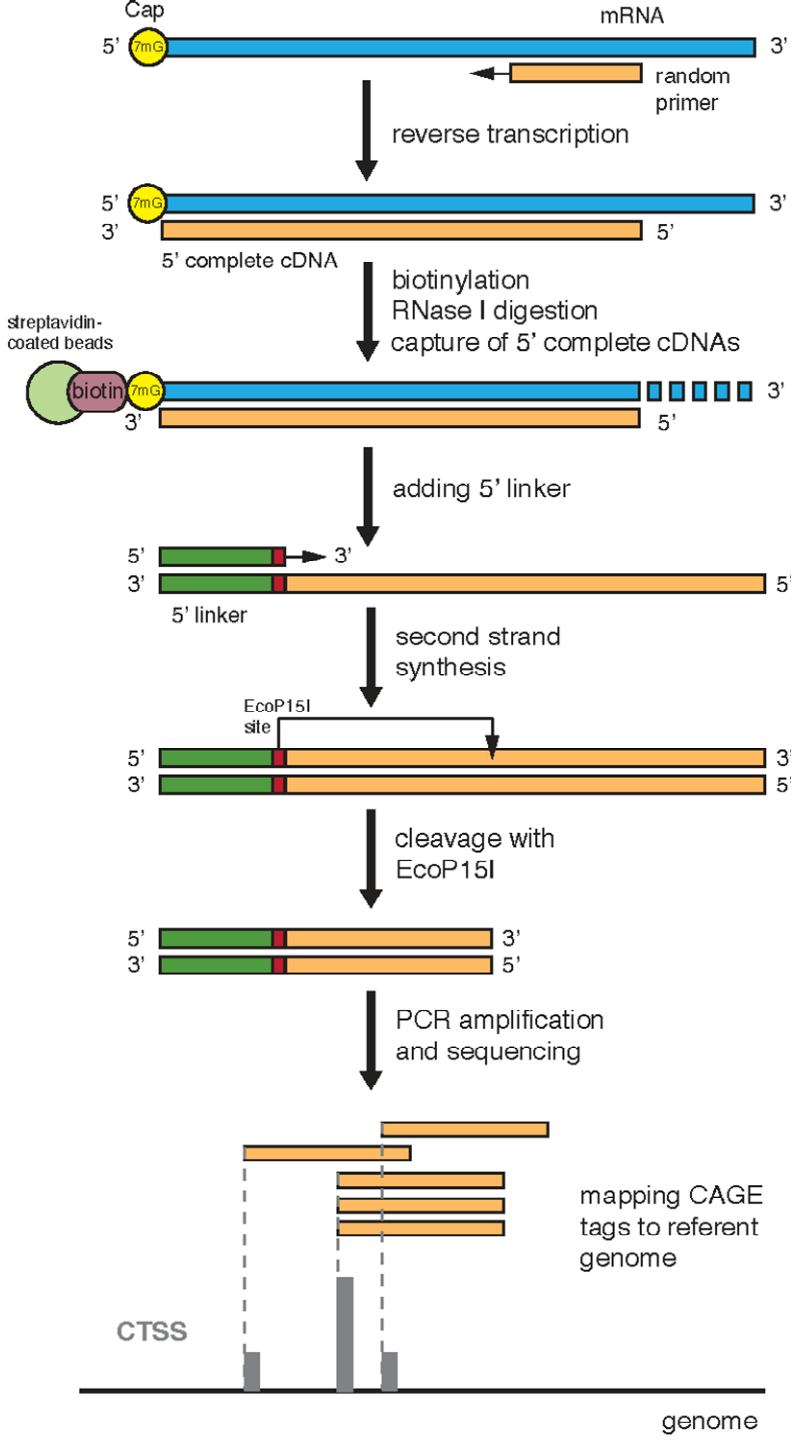
Top 15 pathways for obesity

Pathway Name
Reactome Mitotic Prometaphase
Reactome DNA Replication
Reactome Nuclear Receptor transcription pathway
Reactome Transport to the Golgi and subsequent modification
Reactome Antigen Presentation: Folding assembly and peptide loading of class I MHC
Reactome mitotic MM G1 phases
Reactome Asparagine N-linked glycosylation
Biocarta Nuclearrs pathway
Reactome Synthesis secretion and deacylation of Ghrelin
Reactome Generic Transcription Pathway
Reactome MHC class II antigen presentation
Kegg ECM receptor interaction
Reactome activation of chaperone genes by XBP1S
Reactome Unfolded Protein Response
Reactome class I MHC mediated antigen processing & presentation

Methods was validated with diabetes mellitus



cap analysis of gene expression (CAGE)



GWAS study - concept