Polygenic risk scores

Adrian Campos

Adrian.Campos@qimrberghofer.edu.au

Thanks to

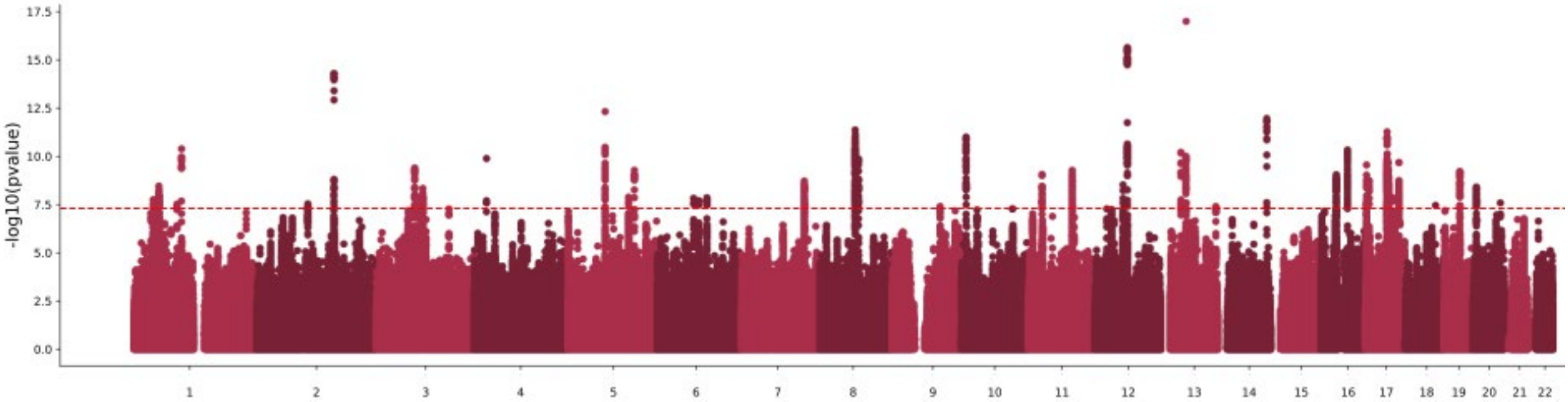
Sarah Medland, Lucia Colodro Conde & Baptiste Couvy Douchesne

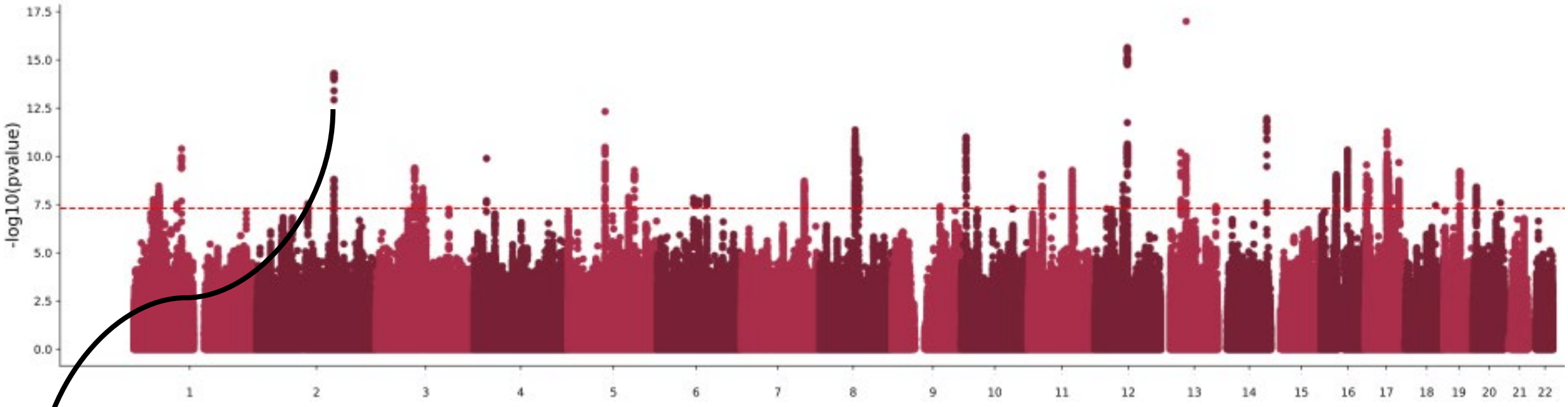
Layout

- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary

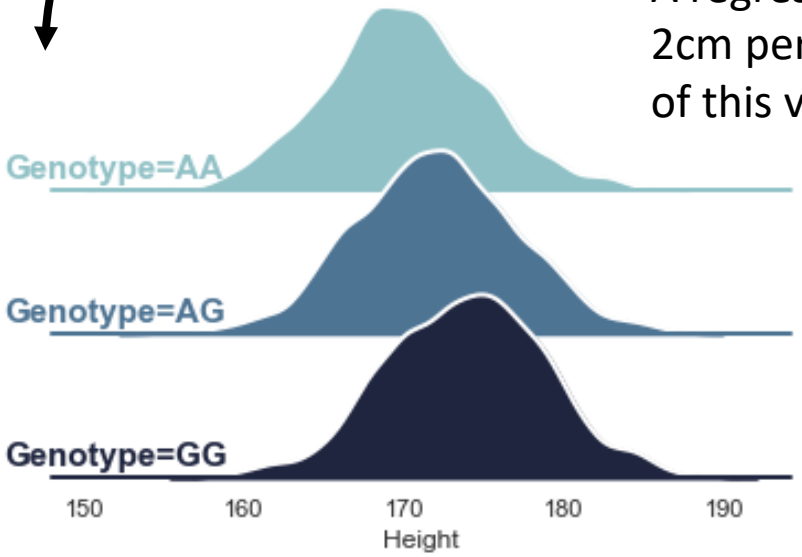
Layout

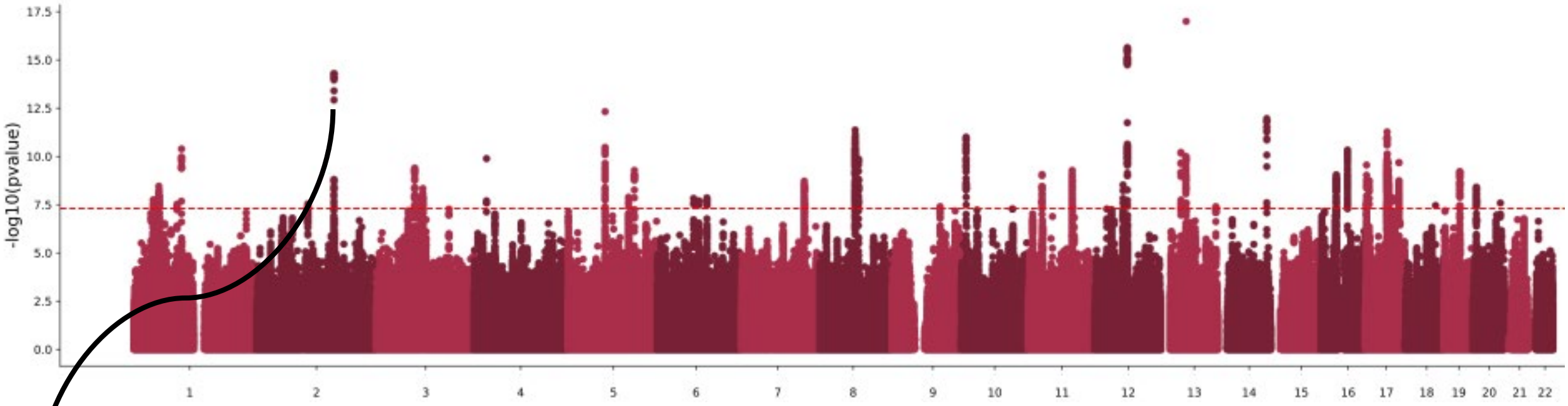
- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary

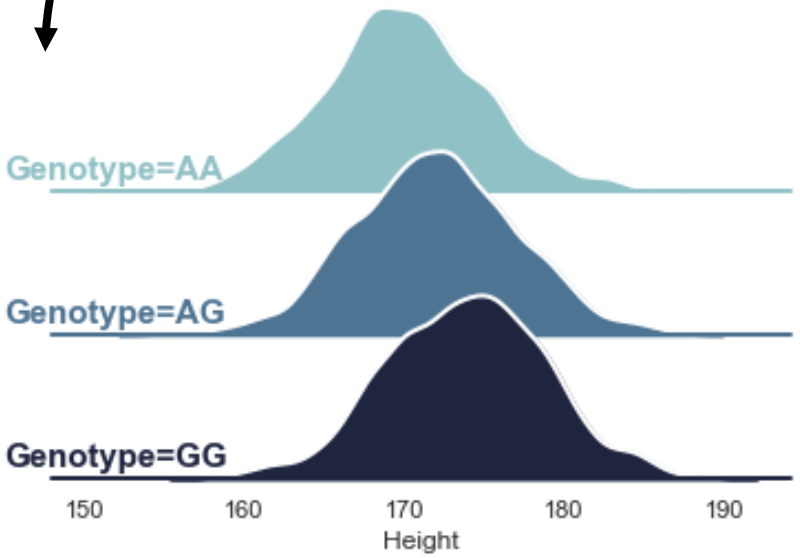

-log10(pvalue)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22



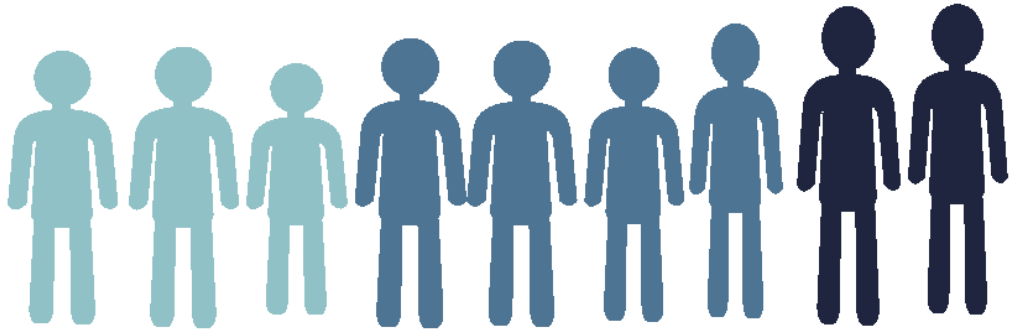
- A regression would show an average increase of 2cm per copy of the G allele. So the effect size of this variant would be approximately 2.

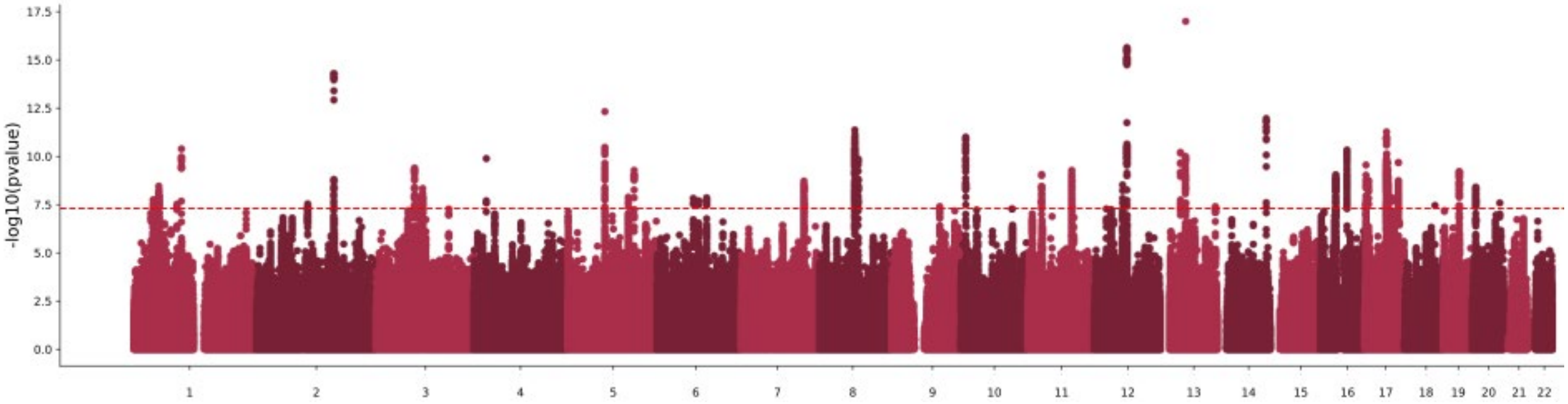






In a new sample we would expect AG individuals to be on average 2cm taller than AA and 2cm shorter than GG





Complex traits are highly polygenic!

From above we can see there are many more genetic variants that contribute to the phenotype

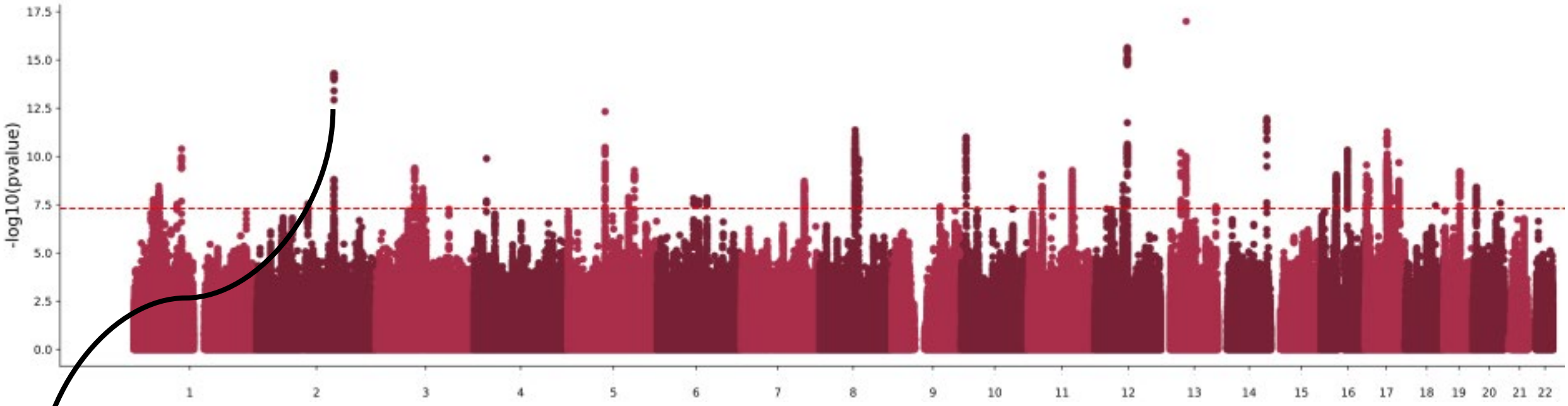
Common variants typically have a small effect size (our example is an exaggeration for a common variant!). This would cause single-loci based prediction useless

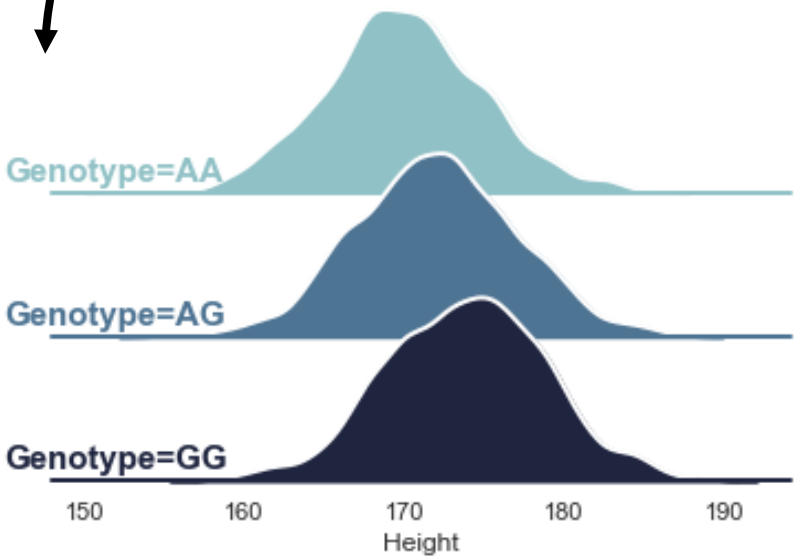
We can combine the information we gain from several genetic variants to estimate an overall score and gain a better estimate of the trait. This is essentially what a PRS does

Layout

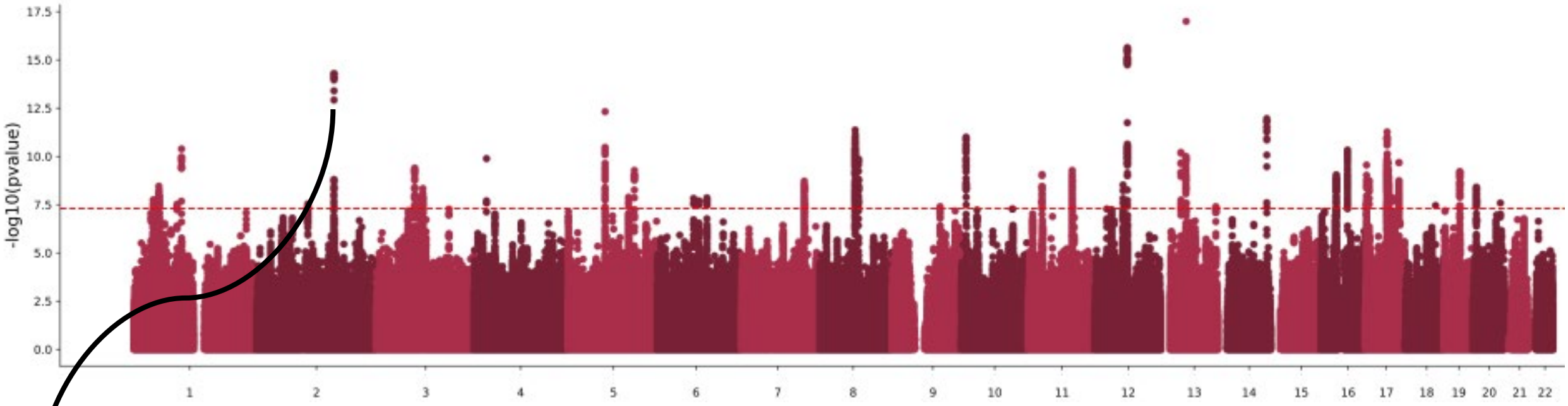
- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary

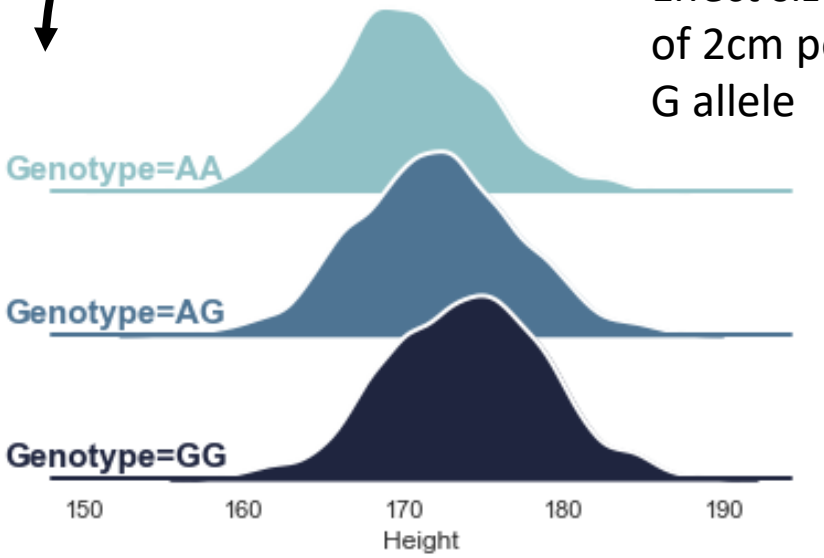
PRS overview



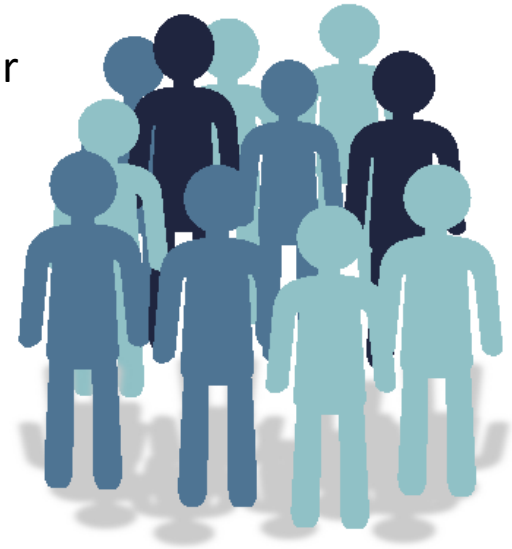


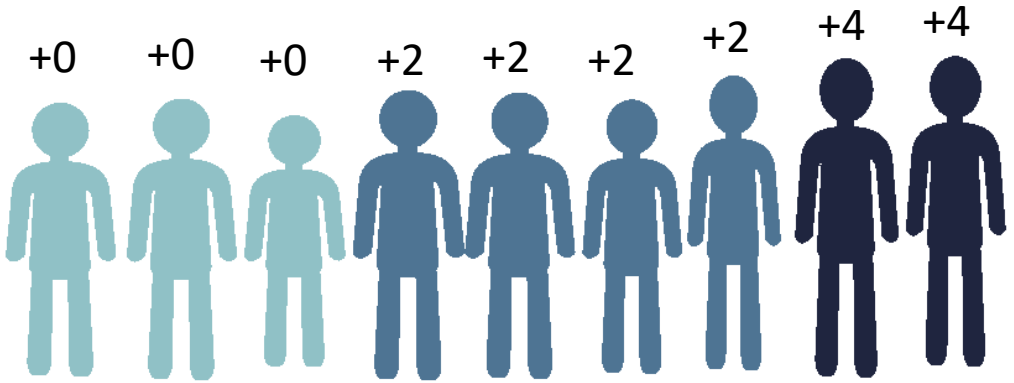
PRS overview

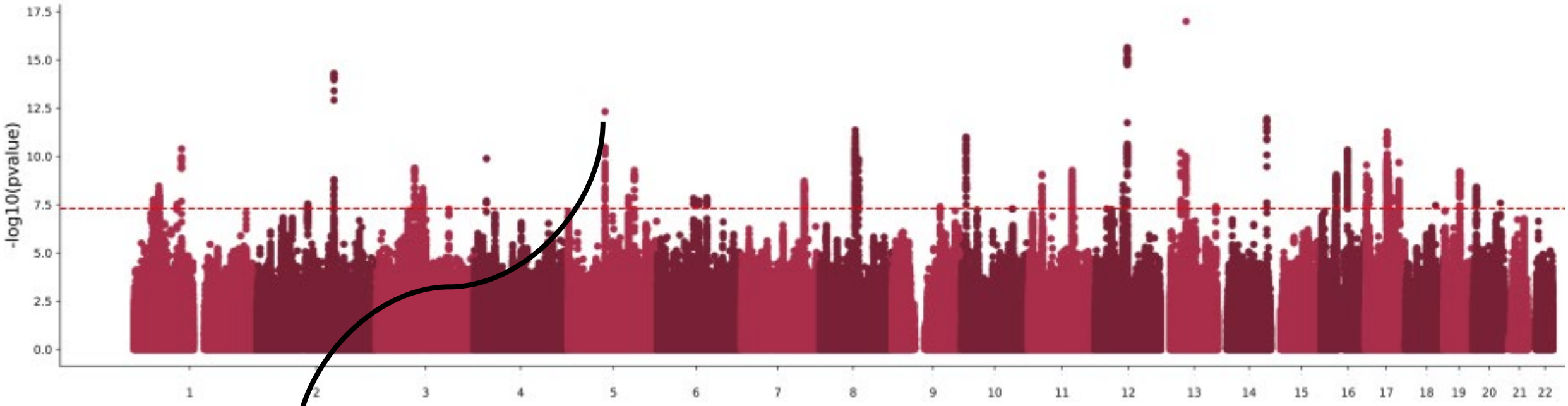




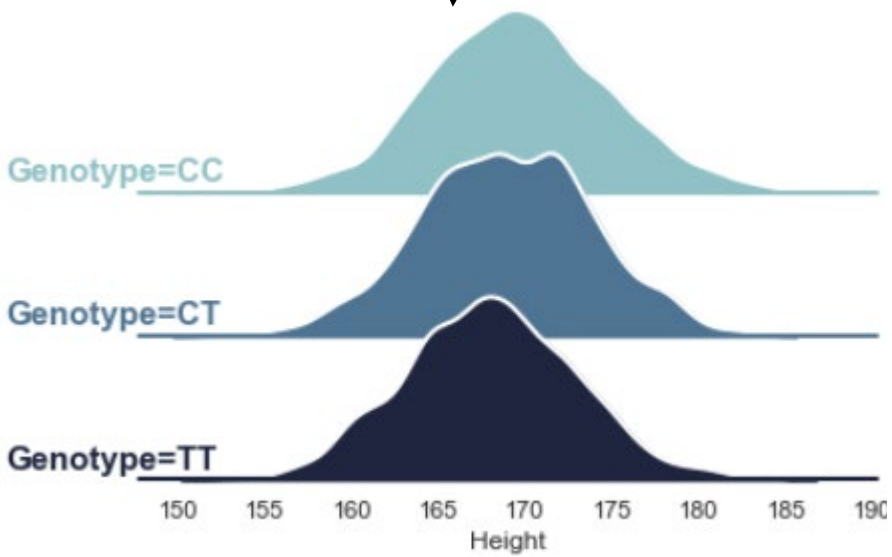
Effect size of 2cm per G allele

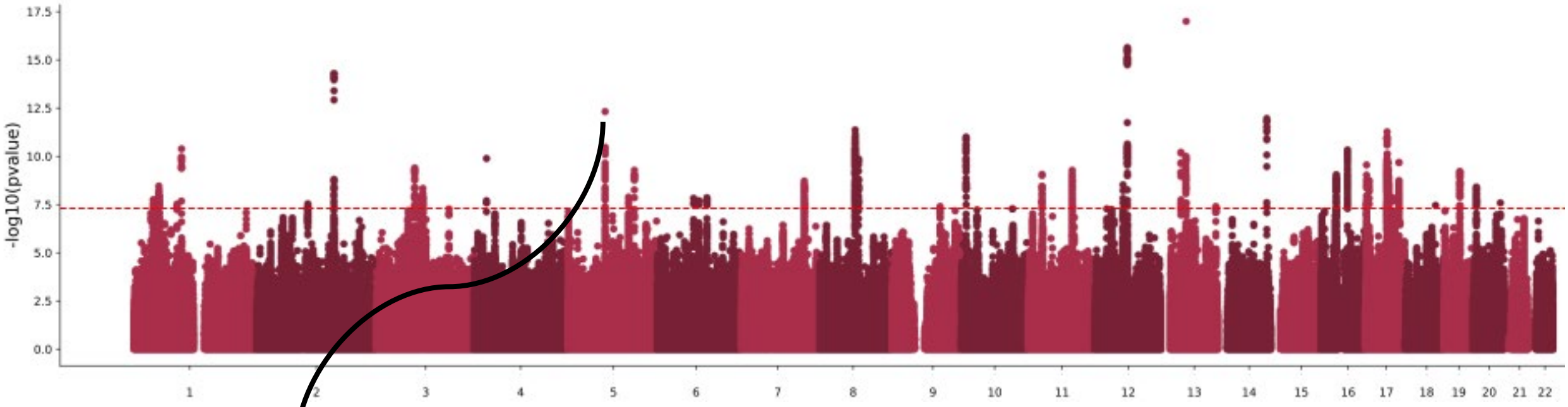




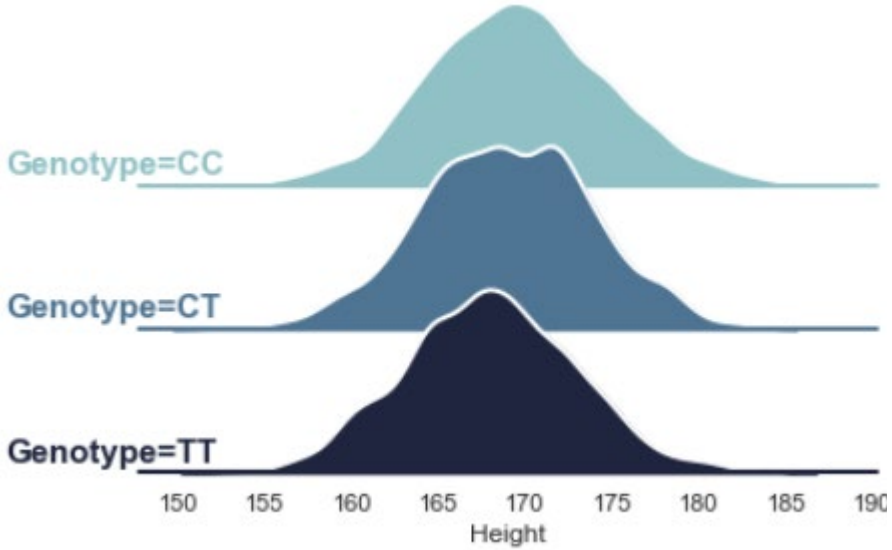


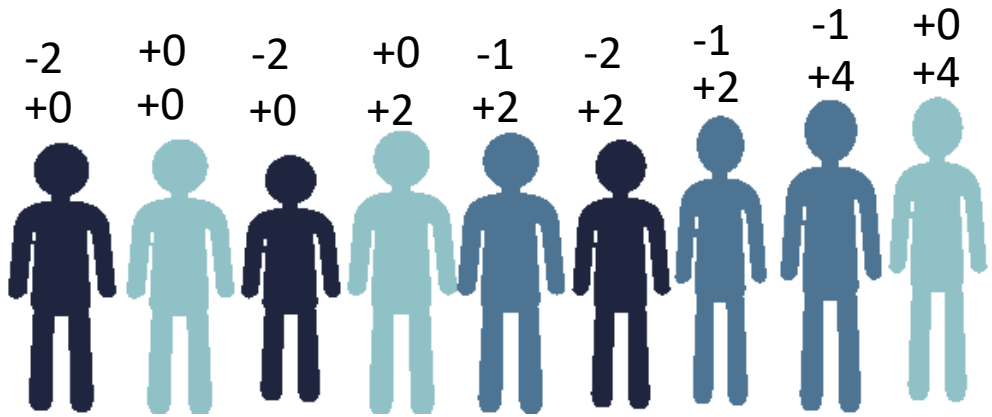
Effect size of -1 per T allele

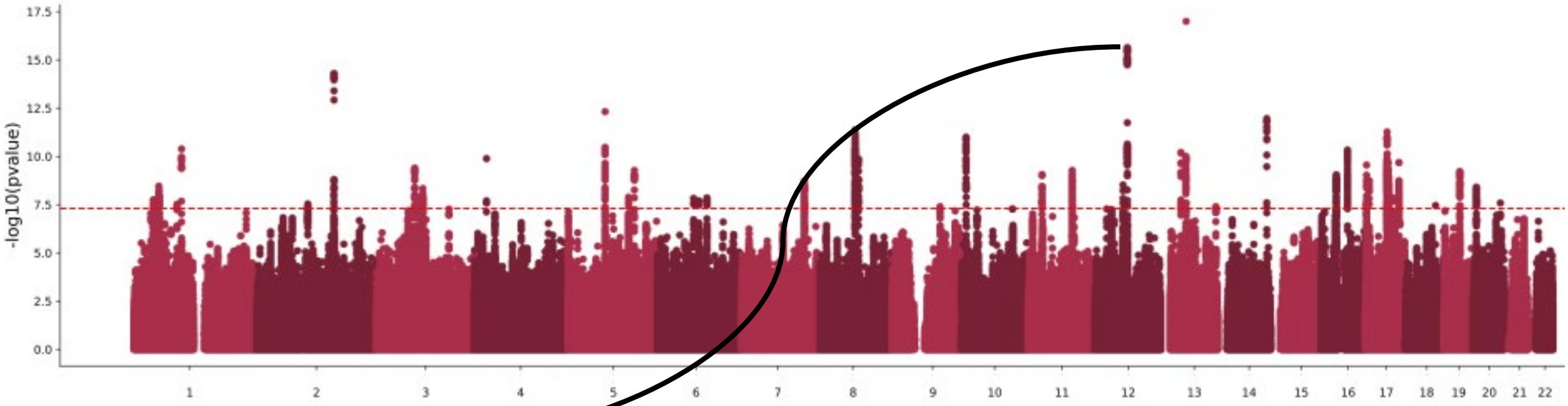




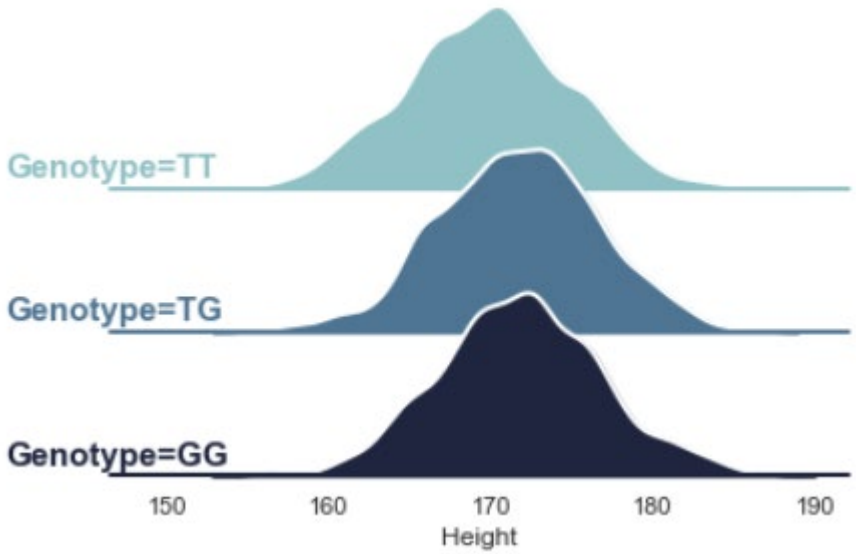
Effect size of -1 per T allele

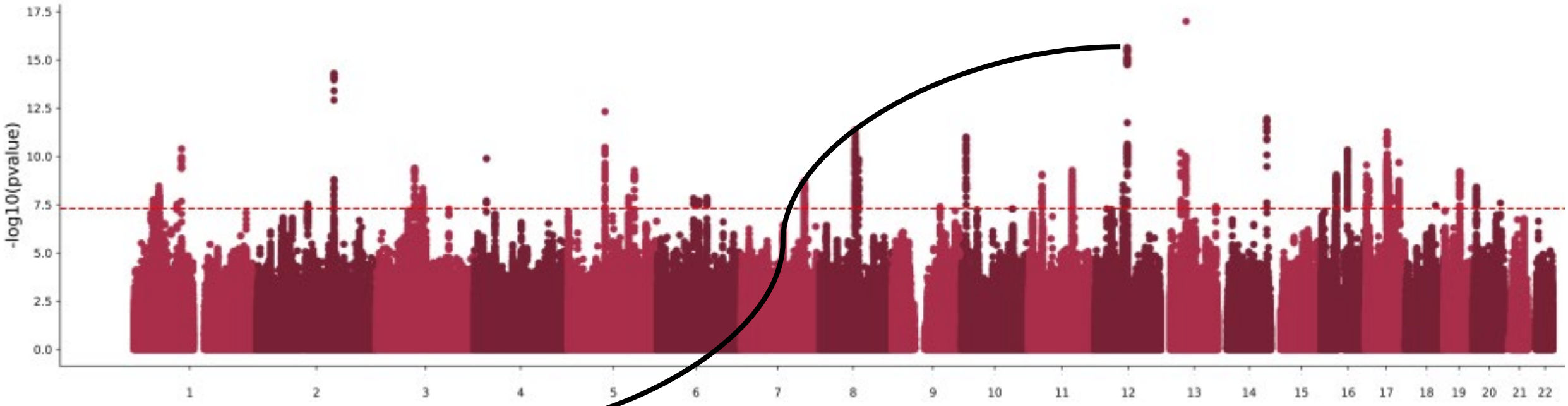




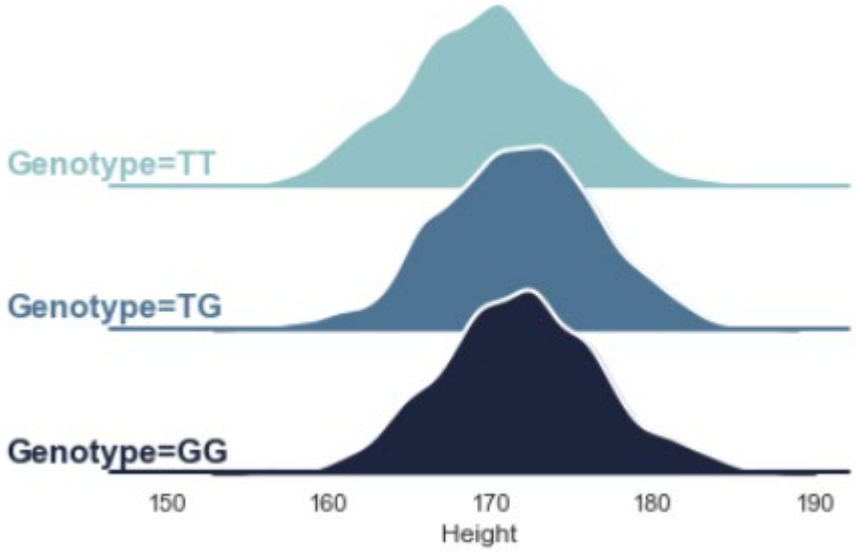


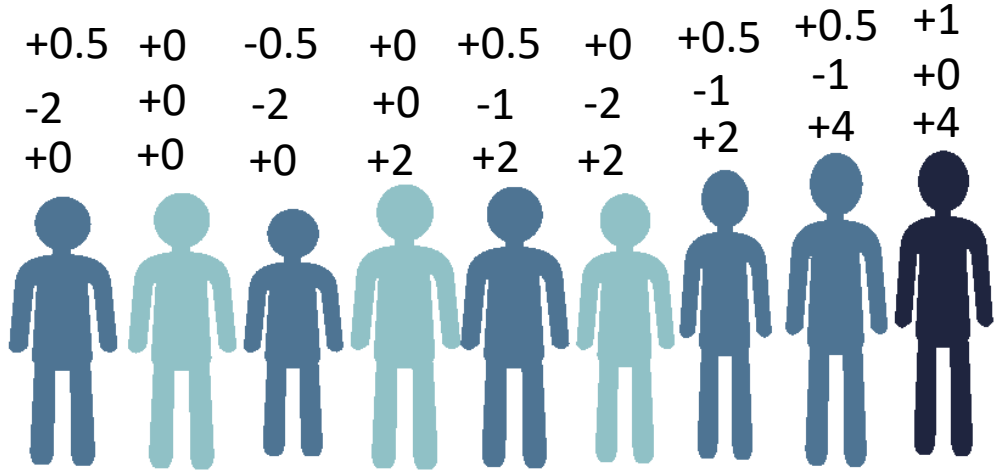
Effect size of +0.5 per G allele



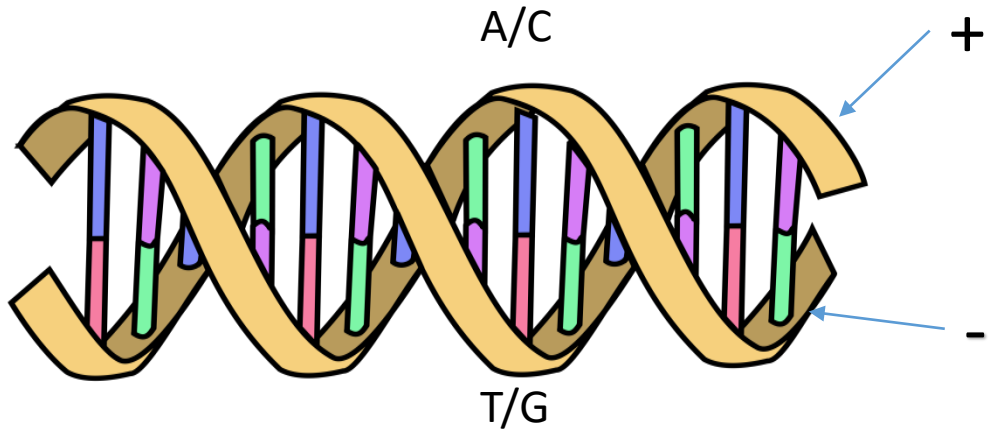


Effect size of +0.5 per G allele



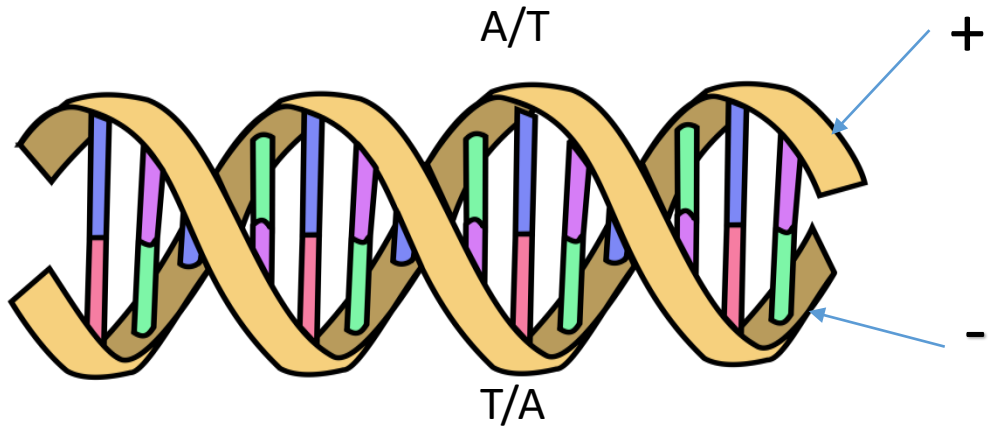


Note on ambiguous variants



rsxxy	A	C
	MAF	
rsxxy	T	G
	MAF	

This variant is not ambiguous

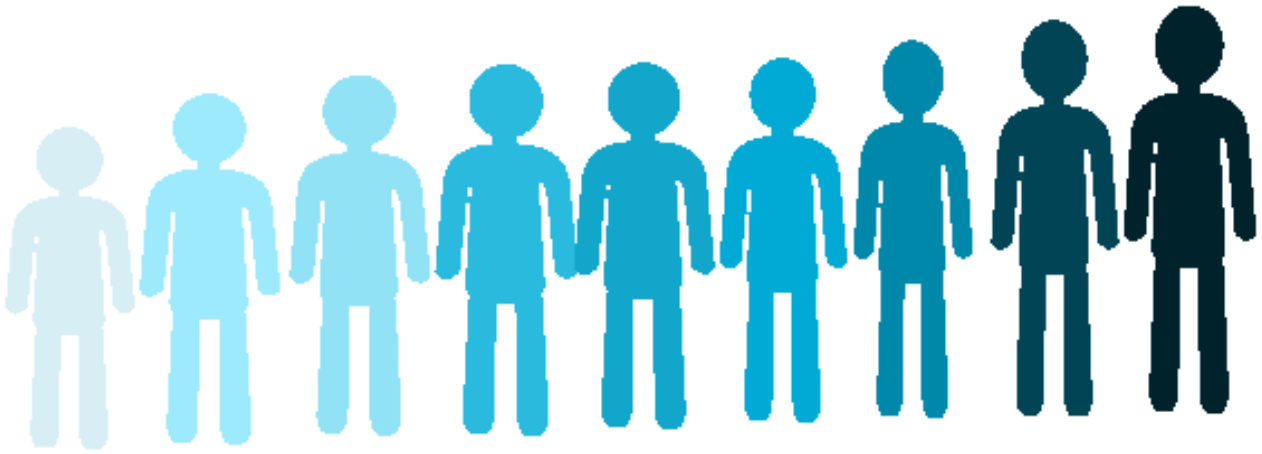


rsxxx	A	T
	MAF	
rsxxx	T	A
	1-MAF	

This variant **is** ambiguous

Note that one can usually solve ambiguity with information on allele frequency, but it gets tricky if its close to 0.5 (it is easy to drop them; as non-ambiguous SNPs will still tag variance thanks to LD)

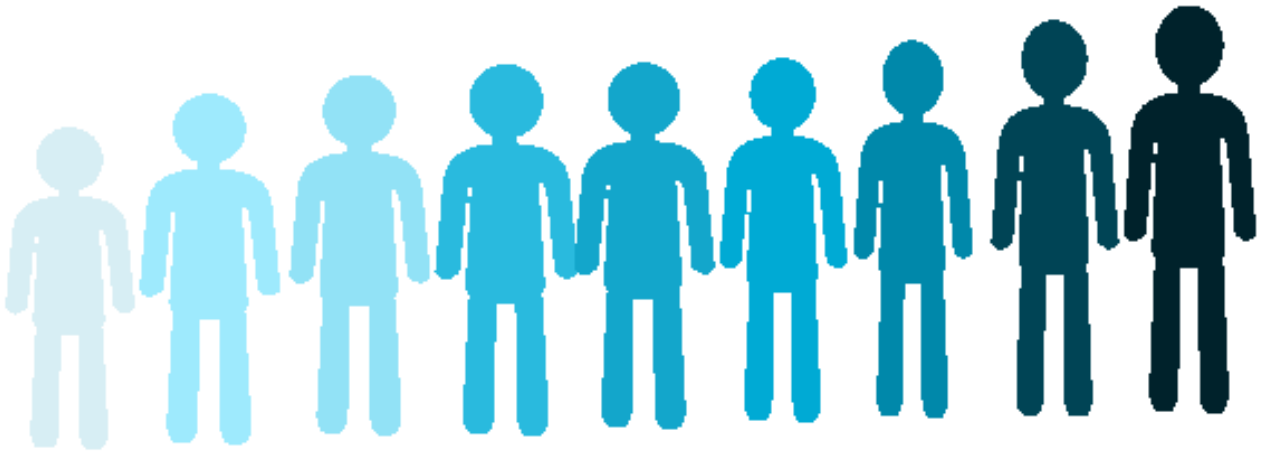
Repeat including the other variants and sum across all loci



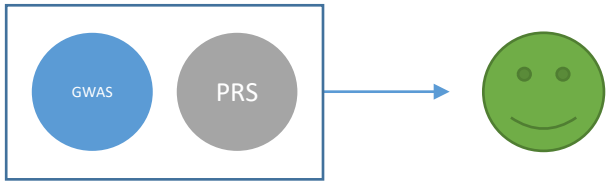
Will give you an estimate of their polygenic risk for the trait of interest

Polygenic risk score – Weighted sum of alleles which quantify the effect of several genetic variants on an individual's phenotype.

Repeat including the other variants and sum across all loci



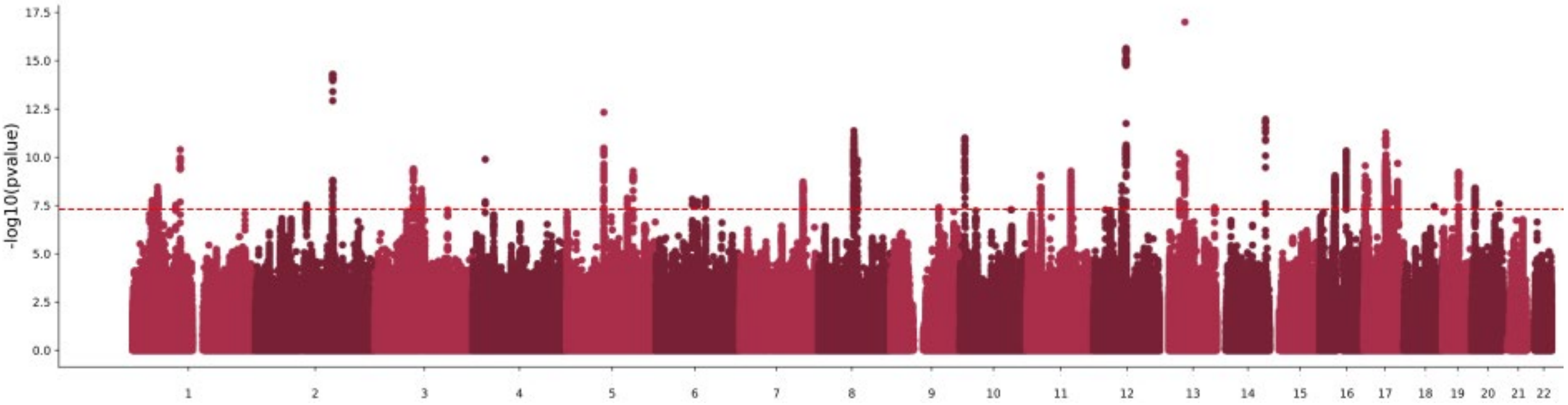
Caution! The sample for which PRS will be calculated should be independent from that of the discovery GWAS. Sample overlap will bias your results.



Layout

- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary

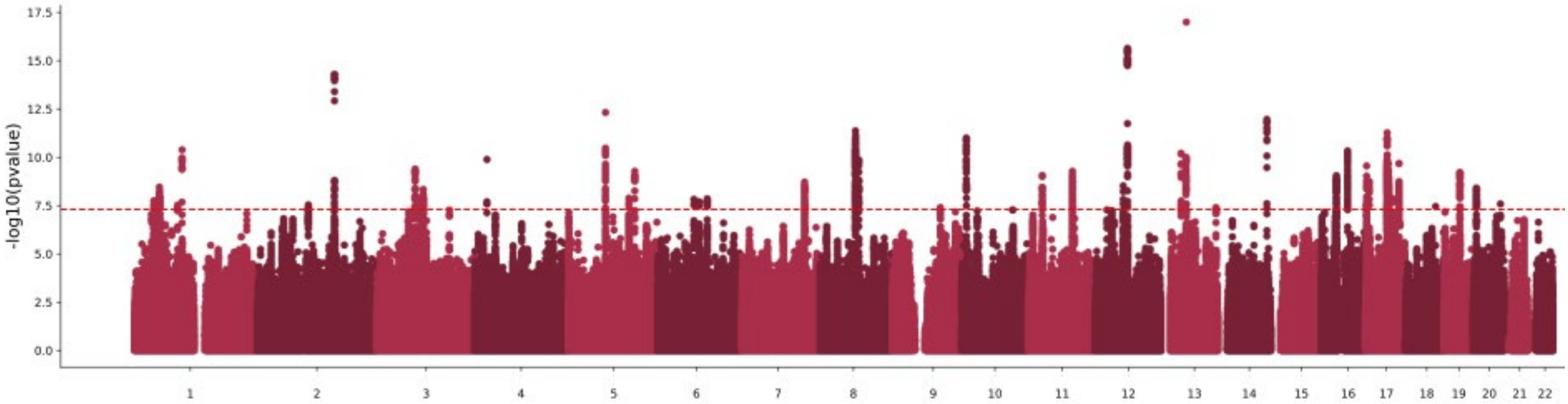
Repeat including the other variants and sum across all loci



Things to consider:
We know many GWAS are underpowered (there's many more true associations than those discovered)

Linkage-disequilibrium creates a correlation structure within the variants. Its important to use independent SNPs (or account for their correlation somehow)

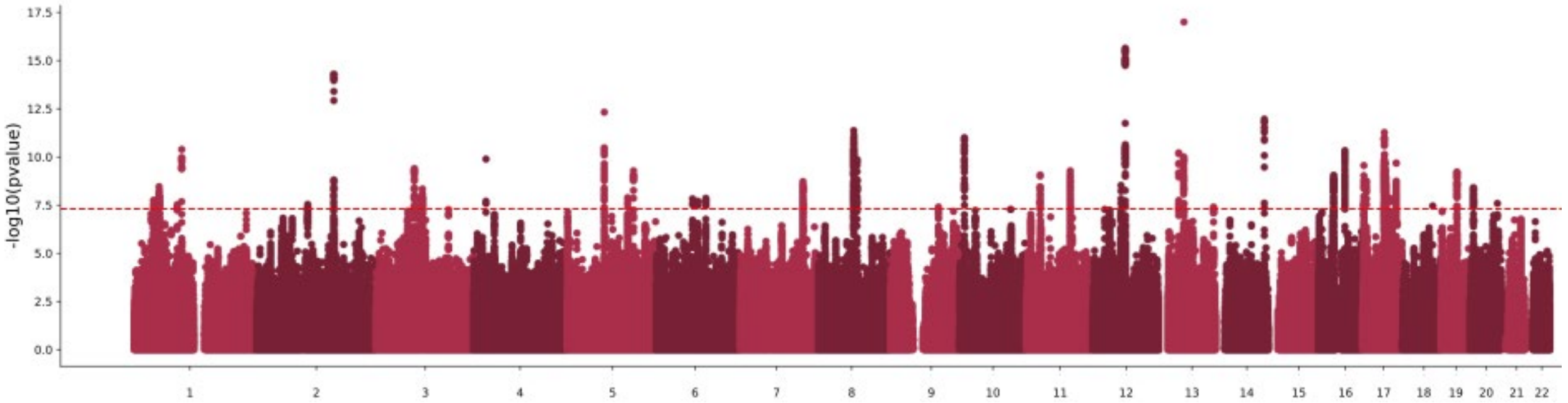
Clumping



Select all SNPs that are significant at a certain p-value threshold (*p1* parameter, set to 1 for traditional approach)
Form *clumps* of SNPs within a certain distance (*kb* param) to the index SNP if they are in LD with the index SNP (*r2* param)

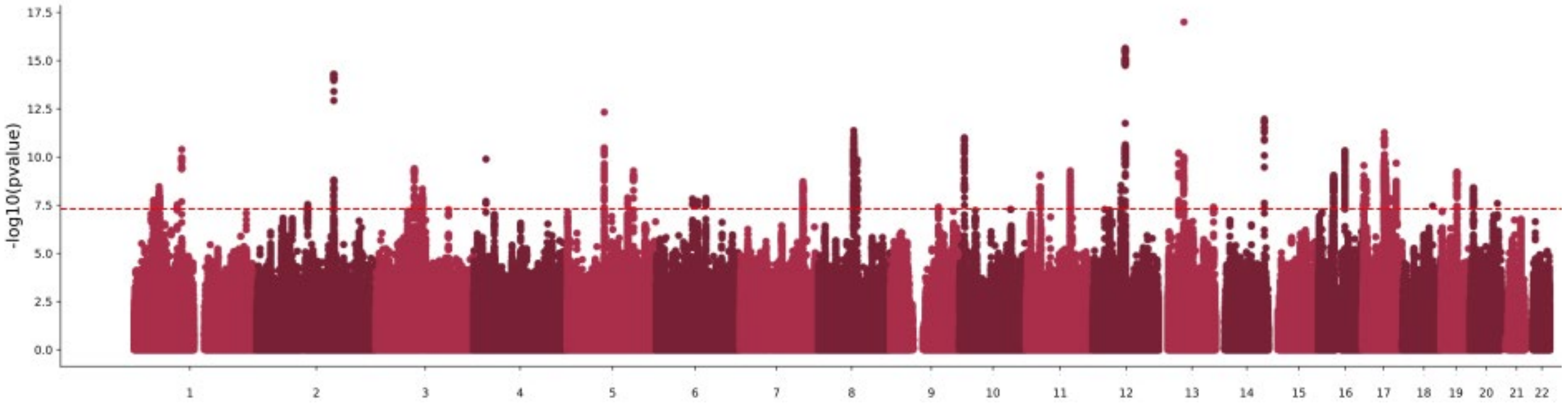
CHR	F	SNP	BP	P	TOTAL	NSIG	S05	S01	S001	S0001	SP2			
14	1	rs2614488	99751588	1.51e-07		79	33	6	13	15	12	rs7148256(1),rs1257434(1),rs941521(1),rs11
14	1	rs8012767	47242425	1.82e-07		648	212	131	218	55	32	rs10133371(1),rs8008347(1),rs4506807(1),rs
14	1	rs7141420	79899454	5.73e-07		234	93	5	55	64	17	rs4903841(1),rs8011958(1),rs4356397(1),rs6
14	1	rs2774042	37361555	1.14e-06		122	66	11	26	2	17	rs848048(1),rs848047(1),rs848046(1),rs8480
14	1	rs1257437	99675579	2.59e-06		101	38	5	15	10	33	rs807569(1),rs807732(1),rs17637919(1),rs74
14	1	rs2300861	33294781	2.92e-06		161	105	8	22	5	21	rs10131246(1),rs11156763(1),rs2383377(1),r
14	1	rs111505678	23580645	3.71e-06		108	54	19	33	1	1	rs117677916(1),rs11543947(1),rs72684308(1)

Clumping and thresholding approach



The variants left are approximately independent, but there is still the question of how significant the association needs to be for inclusion in the PRS calculation

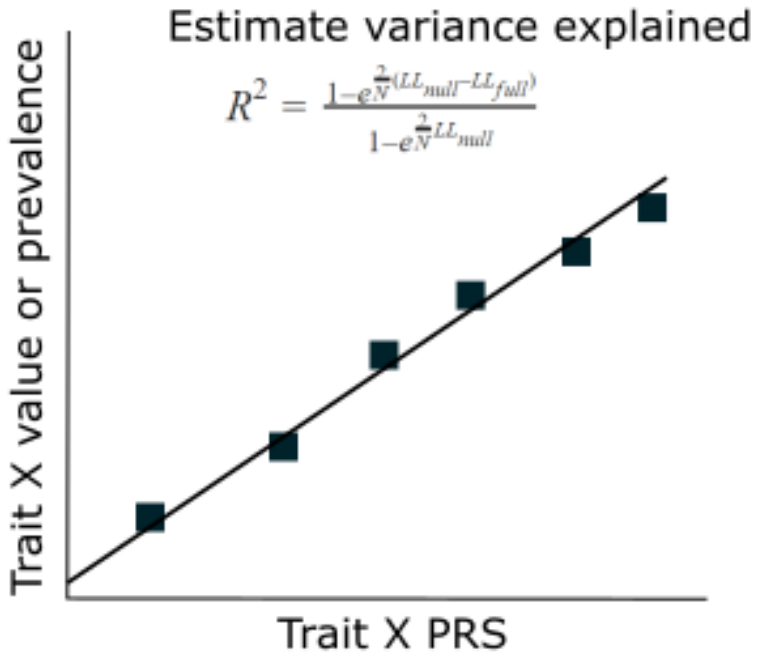
Clumping and thresholding approach



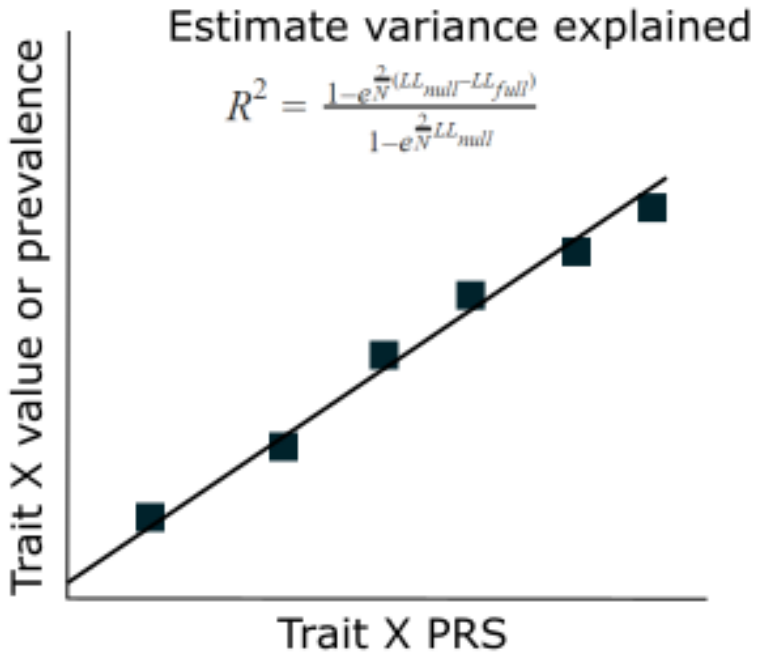
Solution: Calculate many PRS including more and more variants (reducing the p-value threshold used to filter them)
Example 8 p-value thresholds:

Number of independent variants included in PRS calculation							
p<5e-8	p<1e-5	p<0.001	p<0.01	p<0.05	p<0.1	p<0.5	p<1
723	2310	10473	30201	73120	110168	285410	393492

PRS – trait association



PRS – trait association



Think about your sample:

> Is it a family based sample? ! Adjust for relatedness e.g. LMM

> Is it homogeneous in terms of ancestry?

-Always a good idea to adjust for genetic PCs

>Does it match the GWAS ancestry?

Think about your trait:

> Is it continuous – linear regression

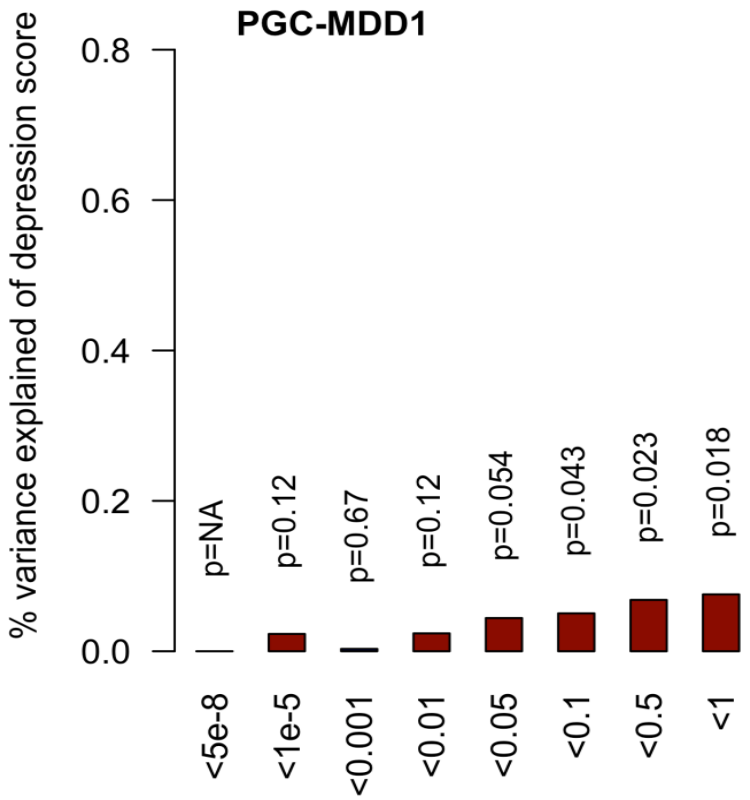
> Binary – logistic or probit regression

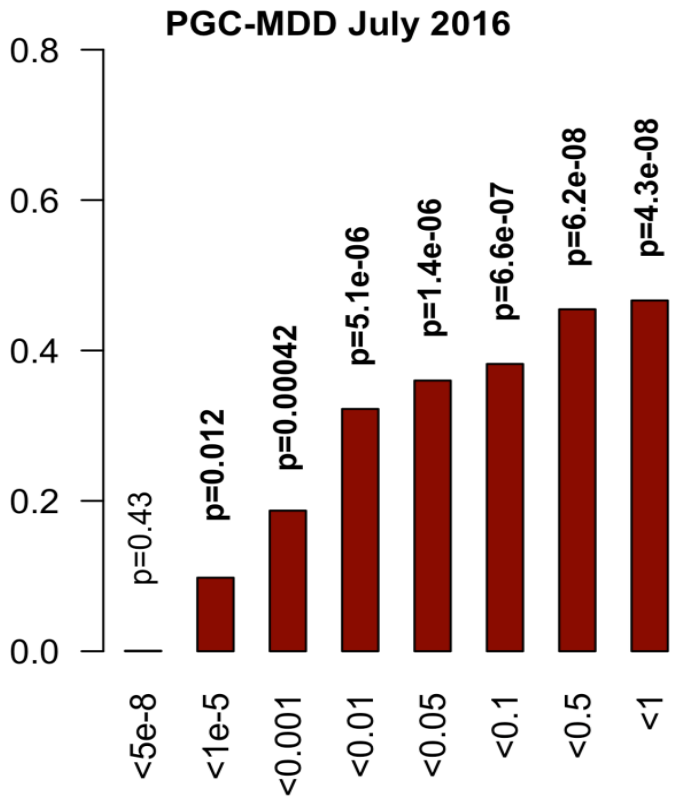
> Ordinal – cumulative linked mixed models

> Always remember potential confounders of the trait and of the discovery GWAS

Power of PRS analysis increases with GWAS sample size

PGC-MDD1: N=18k max variance explained = 0.08%, p=0.018

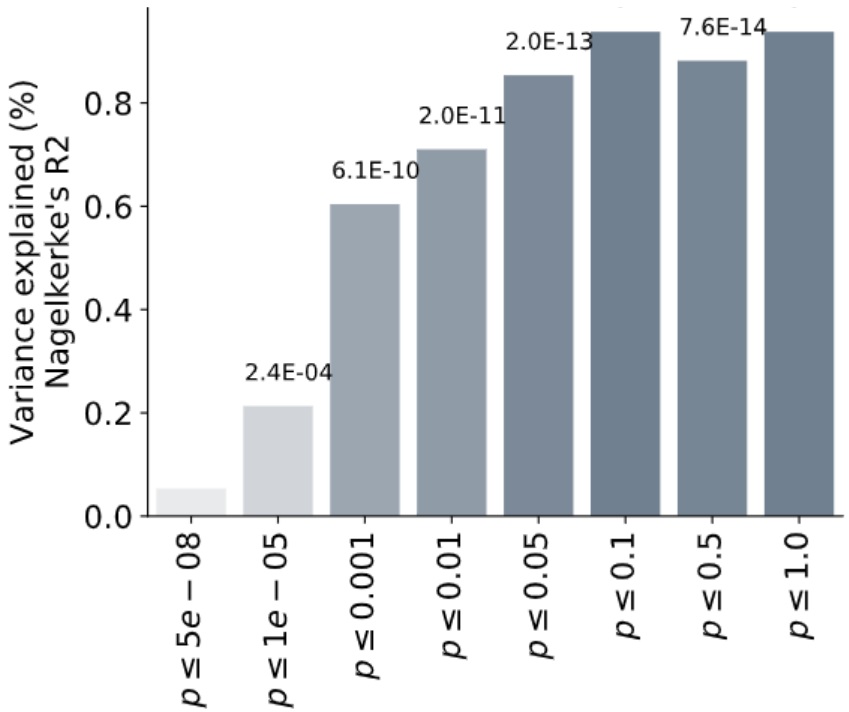


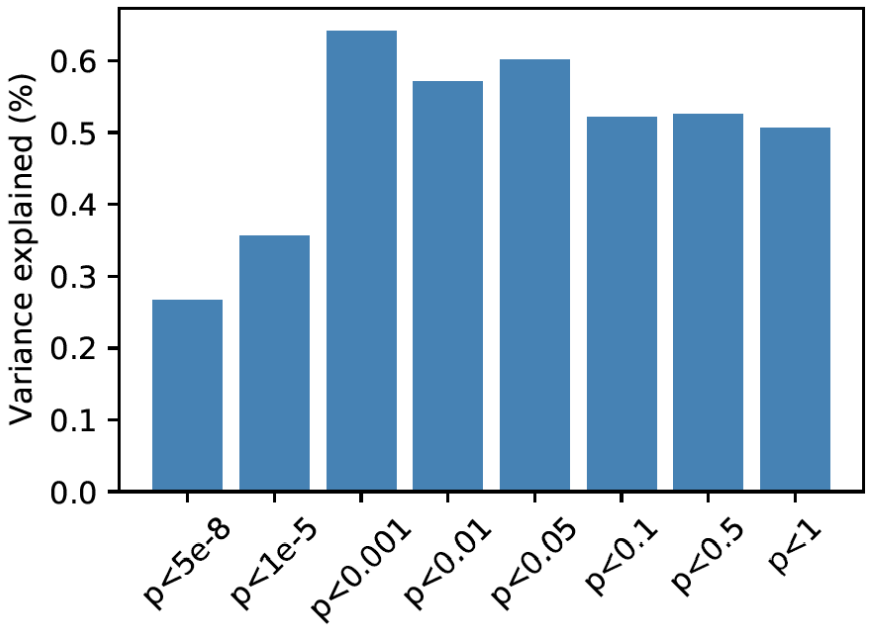


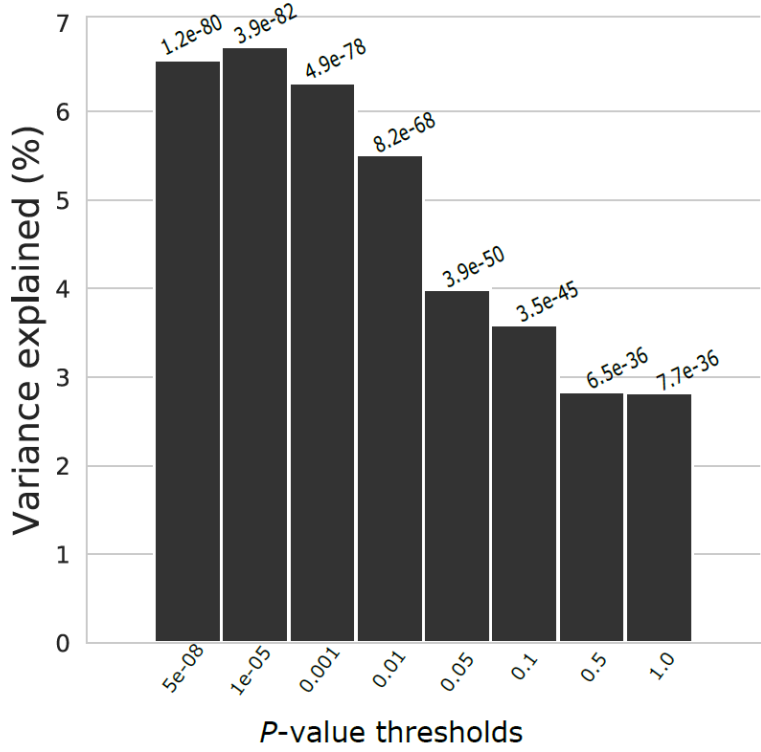
PGC-MDD2: N=163k max variance explained =0.46%, p= 5.01e-08

Colodro-Conde L, Couvy-Duchesne B, et al, (2017) *Molecular Psychiatry*

C+T also allows us to explore the pattern of variance explained







Variance explained = partial R^2 for quantitative traits. Different ways of estimating it for binary traits

Layout

- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary

- Test for GWAS association and quantify variance explained
- Risk stratification (i.e. identifying people to later test for specific disease)
- Aid in clinical diagnosis
- Test for genetic overlap between traits (e.g. does a Depression PRS predict cardiovascular disease?)
- Trait imputation when not measured (obviously imperfect and dependent on heritability)
- Personalized treatment (GWAS on treatment response are gaining power)
- Any hypothesis where you rely on a risk or liability (e.g. GxE interactions)

Layout

- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary

Beyond clumping and thresholding

C+T (your options):

- PLINK
- PRSice2
- bigsnpR (R library)

Other types of PRS:

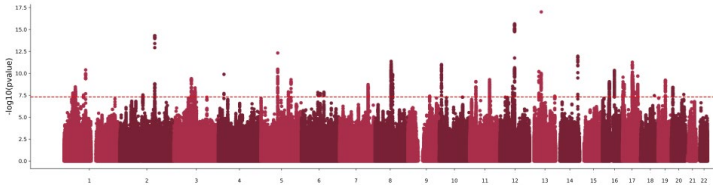
- LDpred2 – Implemented in bigsnpR
- SBayesR – Implemented in GCTB
- Lassosum (and lassosum2) – Implemented in bigsnpR
- PRS-CS
- JAMPred

Commonality across these approaches

- If our sample size and computational power was big enough we could run a multiple linear regression model, and use the joint effect sizes (also called sometimes conditional) for PRS

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Because we can't, what we do is to run *m* regressions (one for each SNP) thus obtaining their marginal effect sizes. The lack of adjustment for correlation is obvious from the Manhattan plot "skyscrapers"



- To solve this problem we need to find a method to approximate the multiple linear regression results based on the GWAS summary statistics

Beyond clumping and thresholding

Approaches for fancier PRS:

- LDpred2 – Implemented in bigsnpR
  - Gibbs sampler to estimate joint SNP effects (replacing clumping)
- SBayesR – Implemented in GCTB
  - Estimates joint SNP effects using Bayesian multiple regression
- Lassosum (and lassosum2) – Implemented in bigsnpR
  - Penalized (LASSO) regression (complementary to LDpred2 for MHC)
- PRS-CS
  - Joint SNP effects using Bayesian regression with continuous shrinkage priors
- JAMPred
  - Two step Bayesian regression framework

SBayesR

- Combines a likelihood connecting the joint effects with GWAS summary statistics and a finite mixture of normal distribution priors for marker effects.

- Models the SNP effect sizes as a mixture of normal distributions with mean zero and different variances.

- Requires GWAS summary statistics with FREQ, BETA, SE and N; and an LD reference matrix

Article | Open Access | Published: 08 November 2019

Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R. Wray, Michael E. Goddard, Jian Yang & Peter M. Visscher

Nature Communications **10**, Article number: 5086 (2019) | Cite this article

$$\beta_j|\pi,\sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0,\gamma_2\sigma_\beta^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0,\gamma_C\sigma_\beta^2) & \text{with probability } 1-\sum_{c=1}^{C-1}\pi_c, \end{cases}$$

Typically uses four normal distributions with mean zero and variances = $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)' = (0, 0.0001, 0.001, 0.01)$

Then performs a Markov chain Monte Carlo Gibbs sampling for the model parameters: $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\pi}', \sigma_\beta^2, \sigma_\varepsilon^2)'$

$$p(\boldsymbol{\beta}|\mathbf{b},\mathbf{D},\mathbf{B}) \propto p(\mathbf{b}|\boldsymbol{\beta},\mathbf{D},\mathbf{B})p(\boldsymbol{\beta}|\mathbf{D},\mathbf{B}).$$

SBayesR

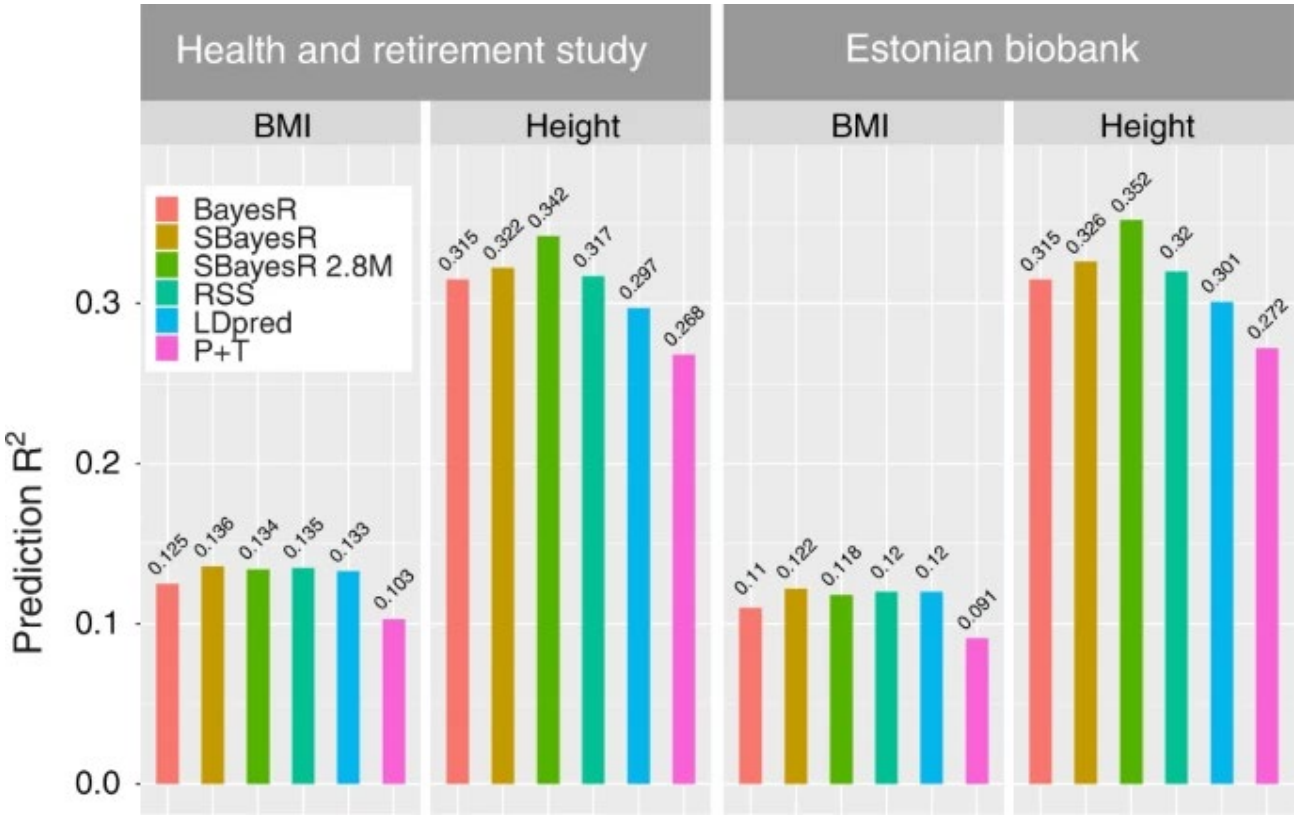
- Combines a likelihood connecting the joint effects with GWAS summary statistics and a finite mixture of normal distribution priors for marker effects.
- Models the SNP effect sizes as a mixture of normal distributions with mean zero and different variances.
- Requires GWAS summary statistics with FREQ, BETA, SE and N; and an LD reference matrix

Article | Open Access | Published: 08 November 2019

Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R. Wray, Michael E. Goddard, Jian Yang & Peter M. Visscher

Nature Communications **10**, Article number: 5086 (2019) | Cite this article



Lloyd-Jones, Jian Zeng, et al (2019)

LDpred2

LDpred2: better, faster, stronger

Florian Privé ✉, Julyan Arbel, Bjarni J Vilhjálmsson ✉

Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5424–5431,
https://doi.org/10.1093/bioinformatics/btaa1029
Published: 16 December 2020 **Article history** ▾

Addressed instability issues in LDpred providing a more stable workflow. Models long range LD such as that found near the HLA region.

Also derives an expectation of joint effects given marginal effects and correlation between SNPs

$$\widehat{\gamma}_{\text{joint}} = \boldsymbol{S}^{-1}\boldsymbol{R}^{-1}\boldsymbol{S}\widehat{\gamma}_{\text{marg}}$$

Assumes:

$$\beta_j = S_{j,j}\gamma_j \sim \begin{cases} \mathcal{N}\left(0, \frac{h^2}{Mp}\right) & \text{with probability p,} \\ 0 & \text{otherwise,} \end{cases}$$

With p= proportion of causal variants and *h2* estimated using Ldscore regression. Grid for p:

p (1, 0.3, 0.1, 0.03, 0.01, 0.003 and 0.001).

Estimated effect sizes from a Gibbs sampler (also MCMC)

It also adds two new models to the traditional LDpred:

1. Estimate *p* and *h2* from the model instead of testing several values and LD-score regression (LDpred2-auto). Thus no intermediate validation dataset is needed to tune these parameters.

2. LDpred2-sparse allows for effect sizes to be exactly 0 (similar to the first mixture component of SBayesR)

LDpred2

LDpred2: better, faster, stronger

Florian Privé ✉, Julyan Arbel, Bjarni J Vilhjálmsson ✉

Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5424–5431,
https://doi.org/10.1093/bioinformatics/btaa1029
Published: 16 December 2020 **Article history** ▾

Addressed instability issues in LDpred providing a more stable workflow. Models long range LD such as that found near the HLA region.

Also derives an expectation of joint effects given marginal effects and correlation between SNPs

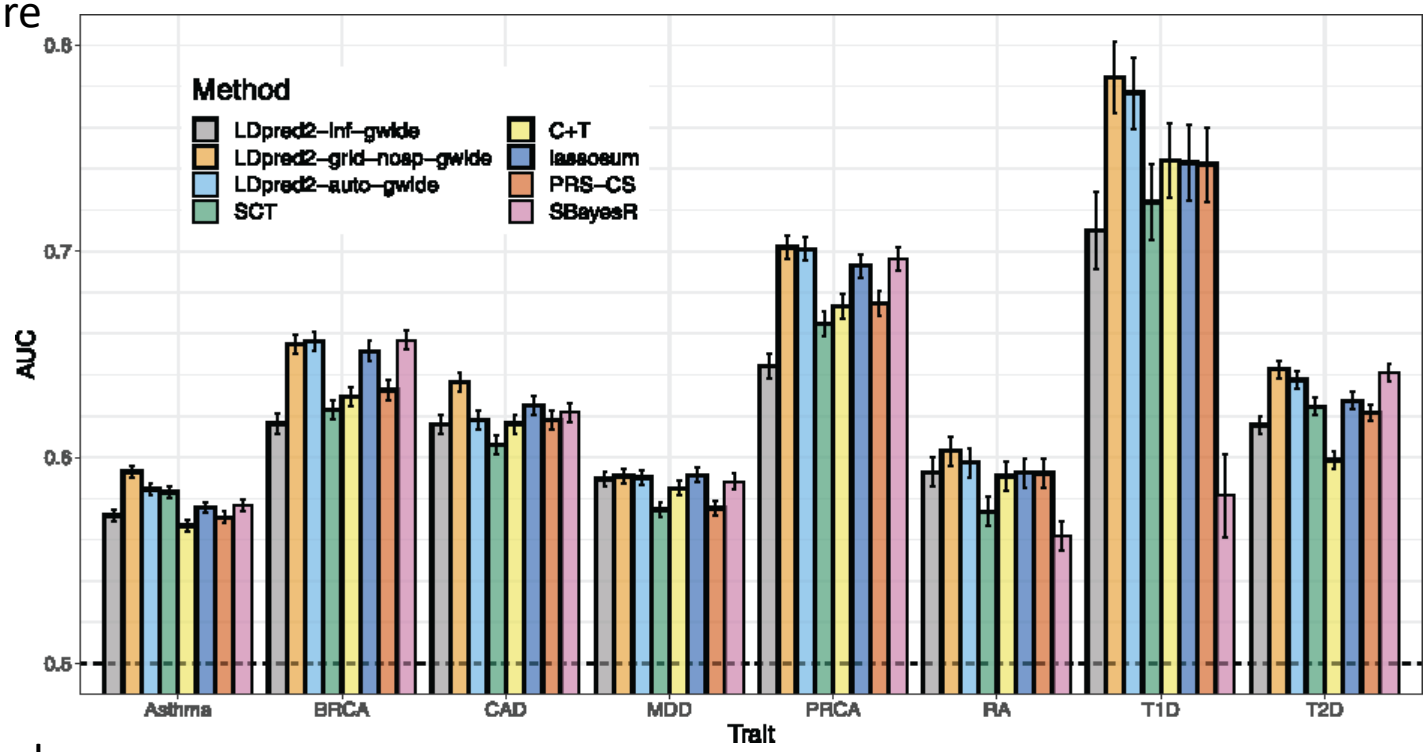
$$\widehat{\gamma}_{\text{joint}} = \boldsymbol{S}^{-1}\boldsymbol{R}^{-1}\boldsymbol{S}\widehat{\gamma}_{\text{marg}}$$

Assumes:

$$\beta_j = S_{j,j}\gamma_j \sim \begin{cases} \mathcal{N}\left(0, \frac{h^2}{Mp}\right) & \text{with probability p,} \\ 0 & \text{otherwise,} \end{cases}$$

With p= proportion of causal variants and *h2* estimated using Ldscore regression. Grid for p:

p (1, 0.3, 0.1, 0.03, 0.01, 0.003 and 0.001).



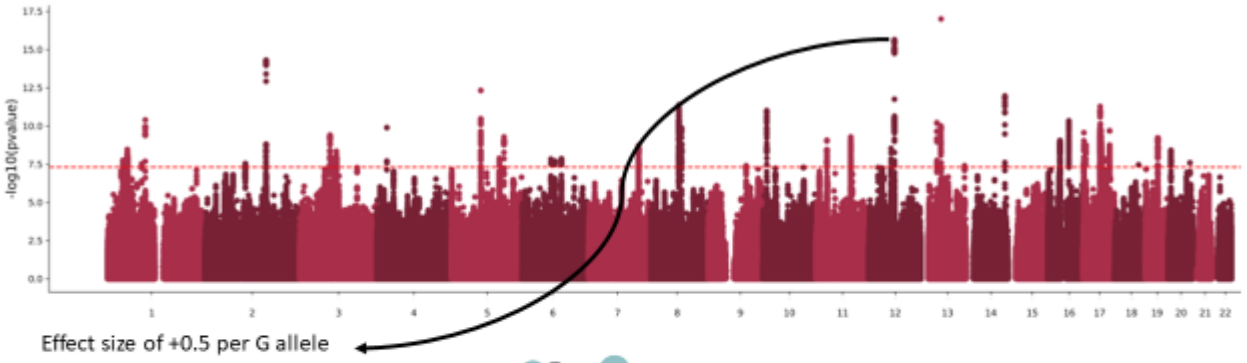
Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5424–5431

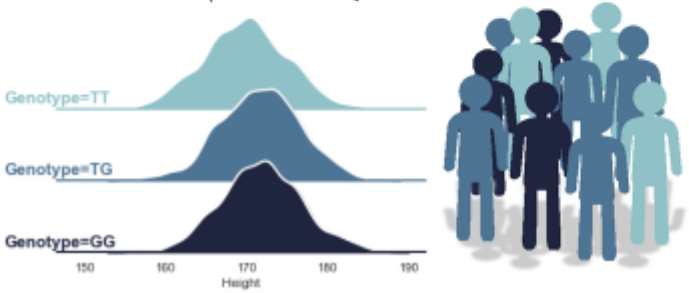
Beyond clumping and thresholding

- These approaches usually perform better than (or at least as well as) C+T
  - When they don’t, maybe raise an eyebrow (sometimes the models don’t converge and they might fail silently)
- Still an area of active research and a clear battle between complexity and power vs scalability and ease of use
- There’s many publications comparing them, read them and pick the one that better fits your needs

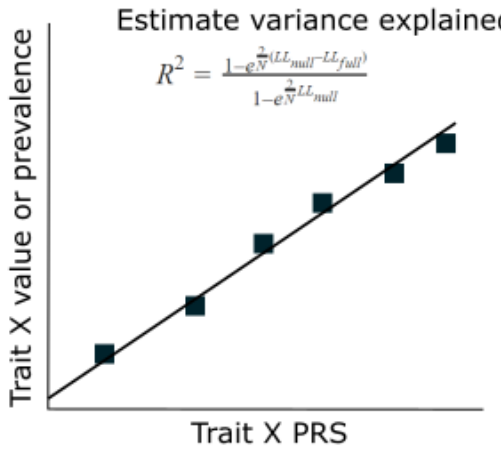
Layout

- Introduction – recapitulating GWAS and allele effect sizes
- PRS overview – graphical summary of what a PRS is
- Which variants to include and accounting for LD
  - Traditional ‘clumping and thresholding’
- Applications for PRS
- Other methods for PRS
- Summary





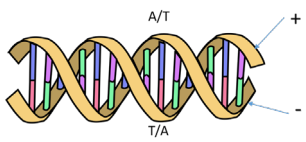
Number of independent variants included in PRS calculation							
p<5e-8	p<1e-5	p<0.001	p<0.01	p<0.05	p<0.1	p<0.5	p<1
723	2310	10473	30201	73120	110168	285410	393492



PRS- Weighted sum of alleles. A tool for estimating the genetic liability or risk to traits

Essential:

- **QC GWAS data (discovery)**
- **QC Genotype data (target)**
- **SNP identifiers need to be matched**
- **Independent discovery and target samples**
- **Consider statistical power**



When using PRS:

- Beware of related individuals in the sample
- Adjust for population stratification
- Ancestry consideration (portability issues)
- Be wary of jumping too fast to conclusions consider potential biases in the discovery GWAS and the target sample.

References for PRS

- Wray NR, Goddard, ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research. 2007; 7(10):1520-28.
- Evans DM, Visscher PM., Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Human Molecular Genetics. 2009; 18(18): 3525-3531.
- International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P . Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748-52
- Evans DM, Brion MJ, Paternoster L, Kemp JP, McMahon G, Munafò M, Whitfield JB, Medland SE, Montgomery GW; GIANT Consortium; CRP Consortium; TAG Consortium, Timpson NJ, St Pourcain B, Lawlor DA, Martin NG, Dehghan A, Hirschhorn J, Smith GD. Mining the human phenome using allelic scores that index biological intermediates. PLoS Genet. 2013,9(10):e1003919.
- Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013 Mar;9(3):e1003348. Epub 2013 Mar 21. Erratum in: PLoS Genet. 2013;9(4). (Important discussion of power)
- Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, Middeldorp CM. Research review: Polygenic methods and their application to psychiatric traits. J Child Psychol Psychiatry. 2014;55(10):1068-87. (Very good concrete description of the traditional methods).
- Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507-15. (Very good discussion of the complexities of interpretation).
- Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. Nat Rev Genet. 2014;15(11):765-76. (Important in the understanding of the effects of ascertainment on PRS work).
- Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, Harris SE, Liewald D, Henders AK, Mendelson MM, Liu C, Joehanes R, Liang L; BIOS Consortium, Levy D, Martin NG, Starr JM, Wijmenga C, Wray NR, Yang J, Montgomery GW, Franke L, Deary IJ, Visscher PM. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. Am J Hum Genet. 2015; 97(1):75-85. (Important for the conceptualization of polygenicity)