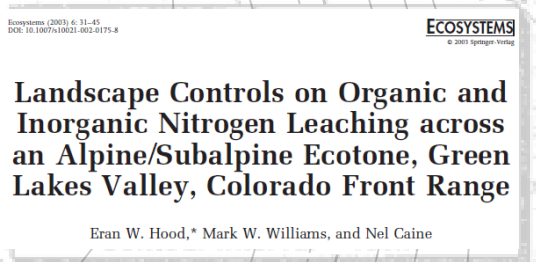
Using genotypic data in our models

Sarah Medland

Boulder 2024

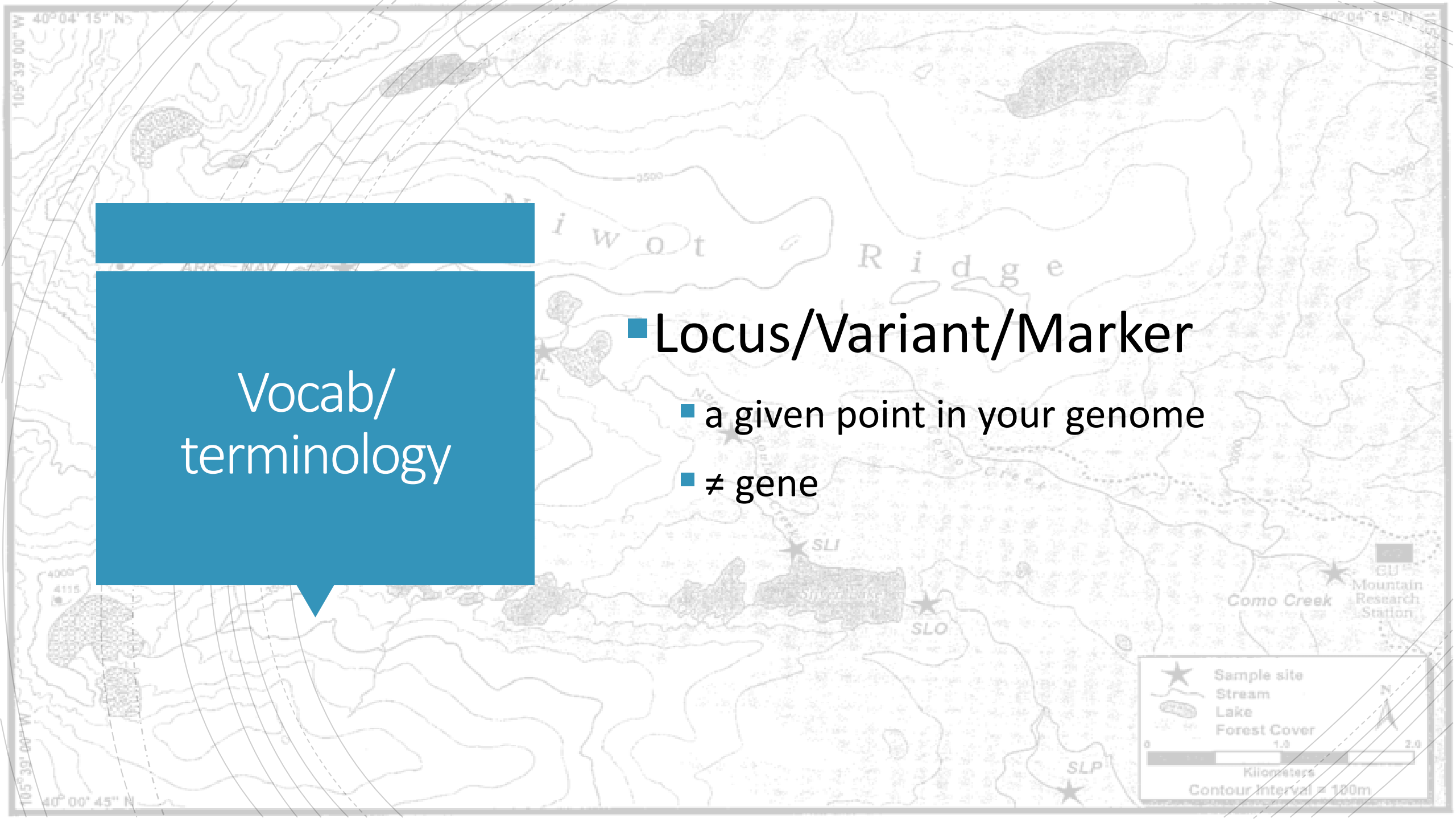


Overview

- Today I'll talk a bit about
  - How we get genotypic data
  - What we do to get it ready to use
  - Estimating relatedness
  - GWAS
  - PRS

Vocab/ terminology

- Locus/Variant/Marker
  - a given point in your genome
  - ≠ gene

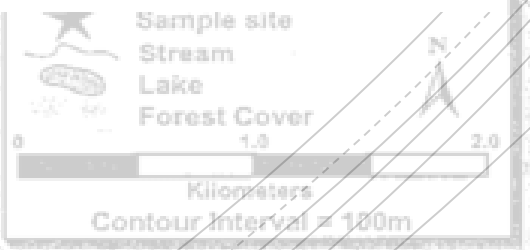


Vocab/ terminology

- Allele
  - 1 of the 2 copies of a variant at a given locus
  - related concept – ambiguous snp

Etymology [edit]

The word "allele" is a short form of "allelomorph" ("other form", a word coined by British geneticists William Bateson and Edith Rebecca Saunders) in the 1900s,[7][8] which was used in the early days of genetics to describe variant forms of a gene detected as different phenotypes. It derives from the Greek prefix ἀλληλο-, *allelo-*, meaning "mutual", "reciprocal", or "each other", which itself is related to the Greek adjective ἄλλος, *allos* (cognate with Latin *alius*), meaning "other".



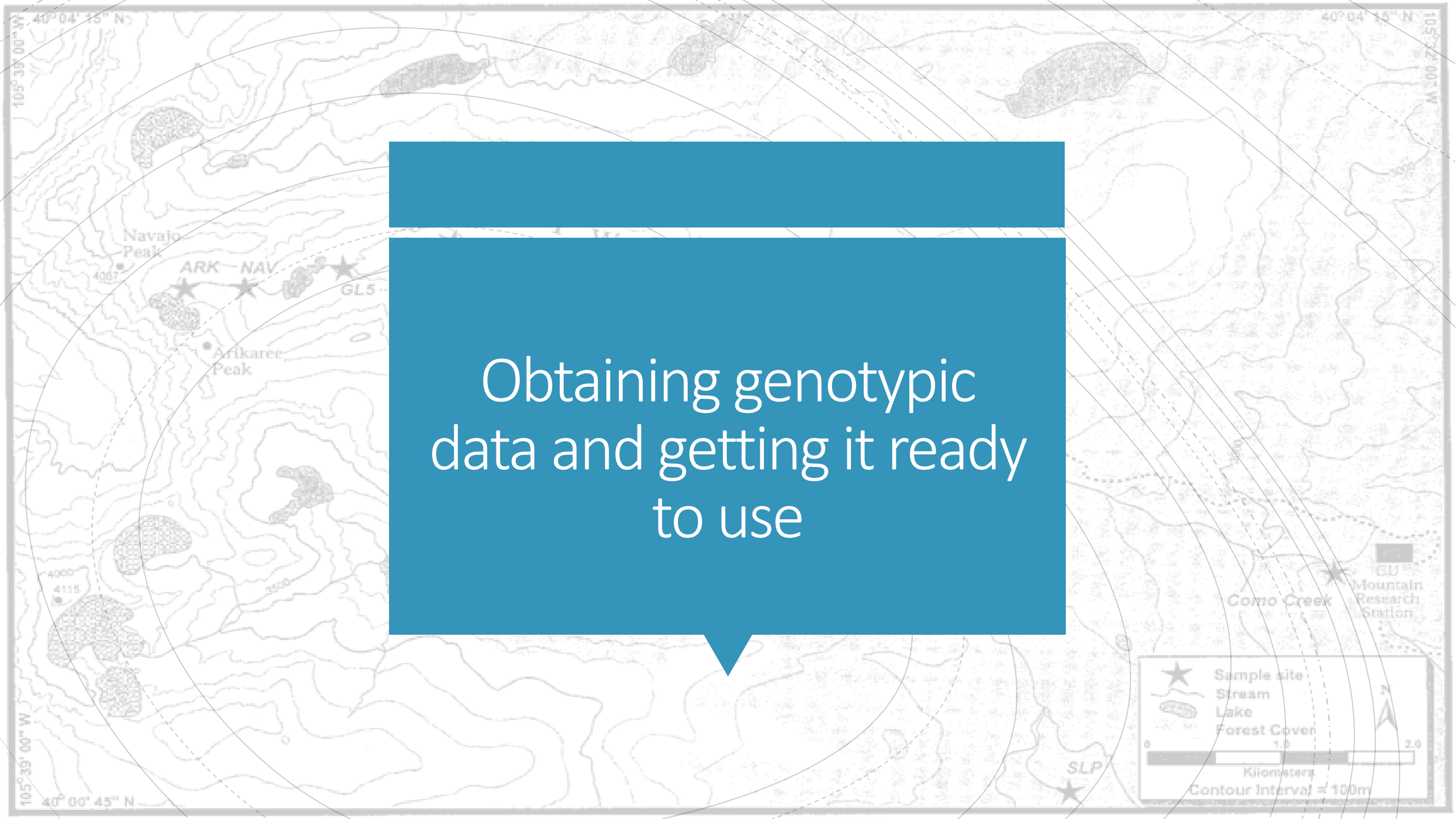
Vocab/ terminology

- Linkage disequilibrium/ LD
  - measure of whether an allele at one locus tends to be found more often with an allele at another locus.
  - related concept – LD block

Vocab/ terminology

- Terminology about locations within the genome
  - Base pair location/BP
  - Build
  - Centimorgan
  - strand

Obtaining genotypic data and getting it ready to use



How do we get genetic data?

1. Recruit a large sample
   - From clinics
   - From the public
   - From an existing twin/cohort sample
   - Pay to access an existing sample
     - UK Biobank…

How do we get genetic data?

2. Collect information from the participants

Survey Completion

0% 100%

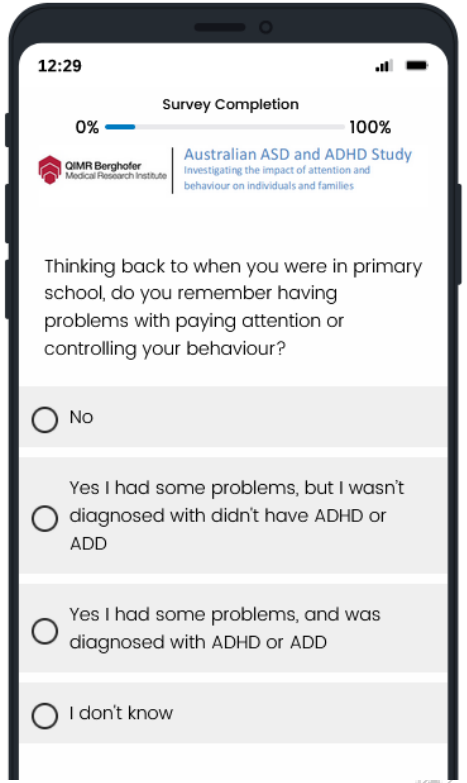
QIMR Berghofer Medical Research Institute

Australian ASD and ADHD Study

Investigating the impact of attention and behaviour on individuals and families

Thinking back to when you were in primary school, do you remember having problems with paying attention or controlling your behaviour?

- No
- Yes I had some problems, but I wasn't diagnosed with didn't have ADHD or ADD
- Yes I had some problems, and was diagnosed with ADHD or ADD
- I don't know



How do we get genetic data?

3. Collect a DNA sample

How do we get genetic data?

4. Extract the DNA

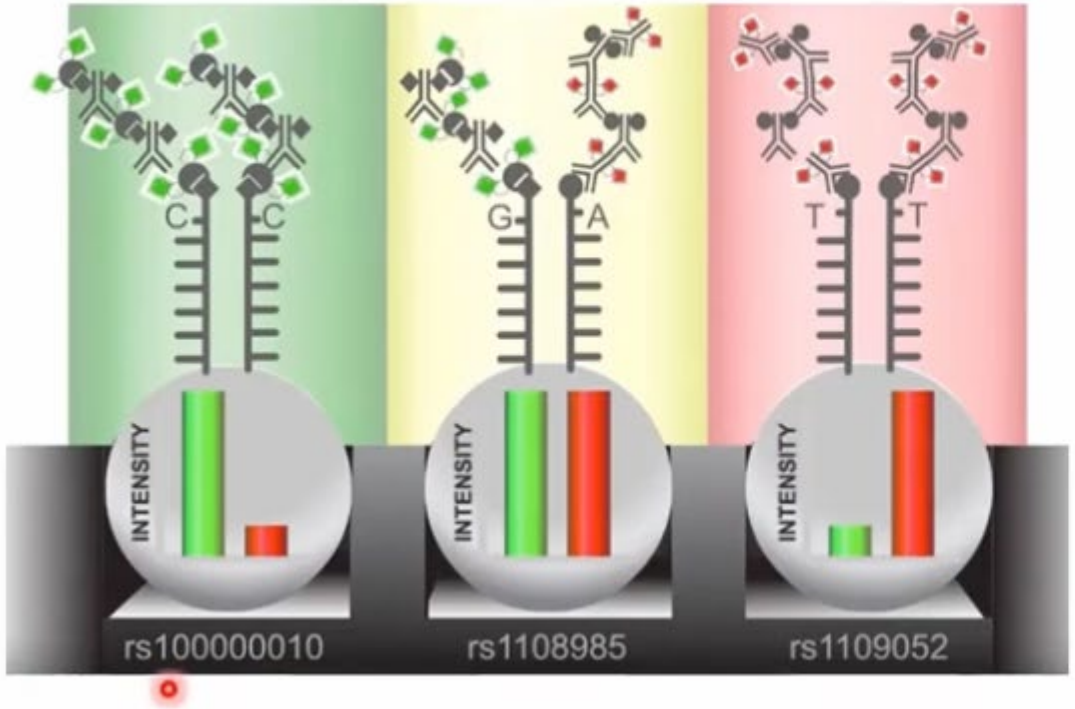
For non lab trained folk this webpage gives a lay overview that you might find helpful

https://learn.genetics.utah.edu/content/labs/extraction/howto/

How do we get genetic data?

5. Genotype the samples

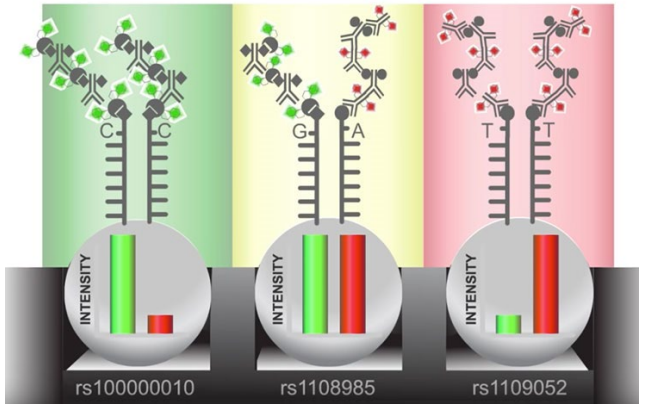
5. Genotype the samples

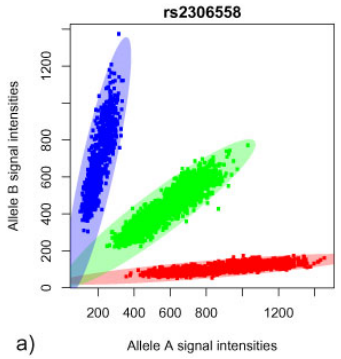


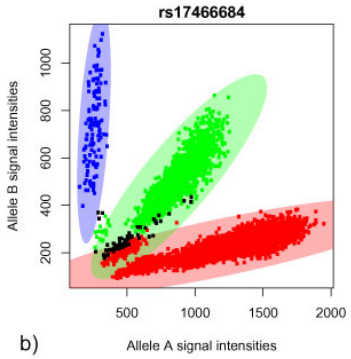
- As DNA fragments pass over the BeadChip
- Each probe binds to a complementary sequence in the sample DNA, stopping one base before the locus of interest
- Single base extension that incorporates one of four labeled nucleotides
- When excited by a laser, the nucleotide label emits a signal
- The intensity of that signal conveys information about the allelic ratio at that locus

How do we get genetic data?

5. Genotype the samples







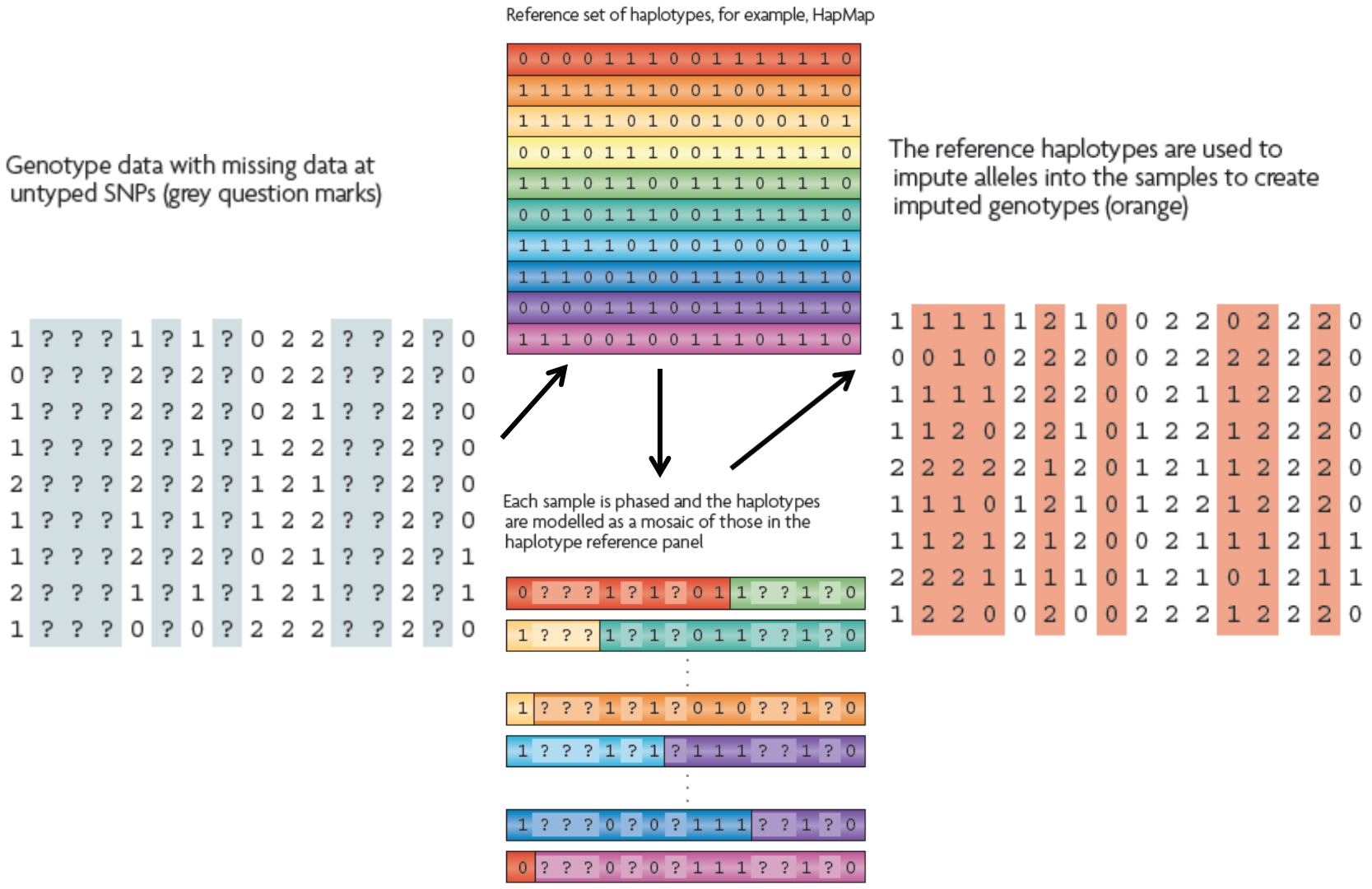
What do we do to get it ready to use?

- Quality Control
  - Genotype level – allele frequency, missingness, Hardy-Weinberg Equilibrium (distribution of alleles)
  - Sample level – missingness, heterozygosity, chromosomal distributions

Come back next year to learn how to do this

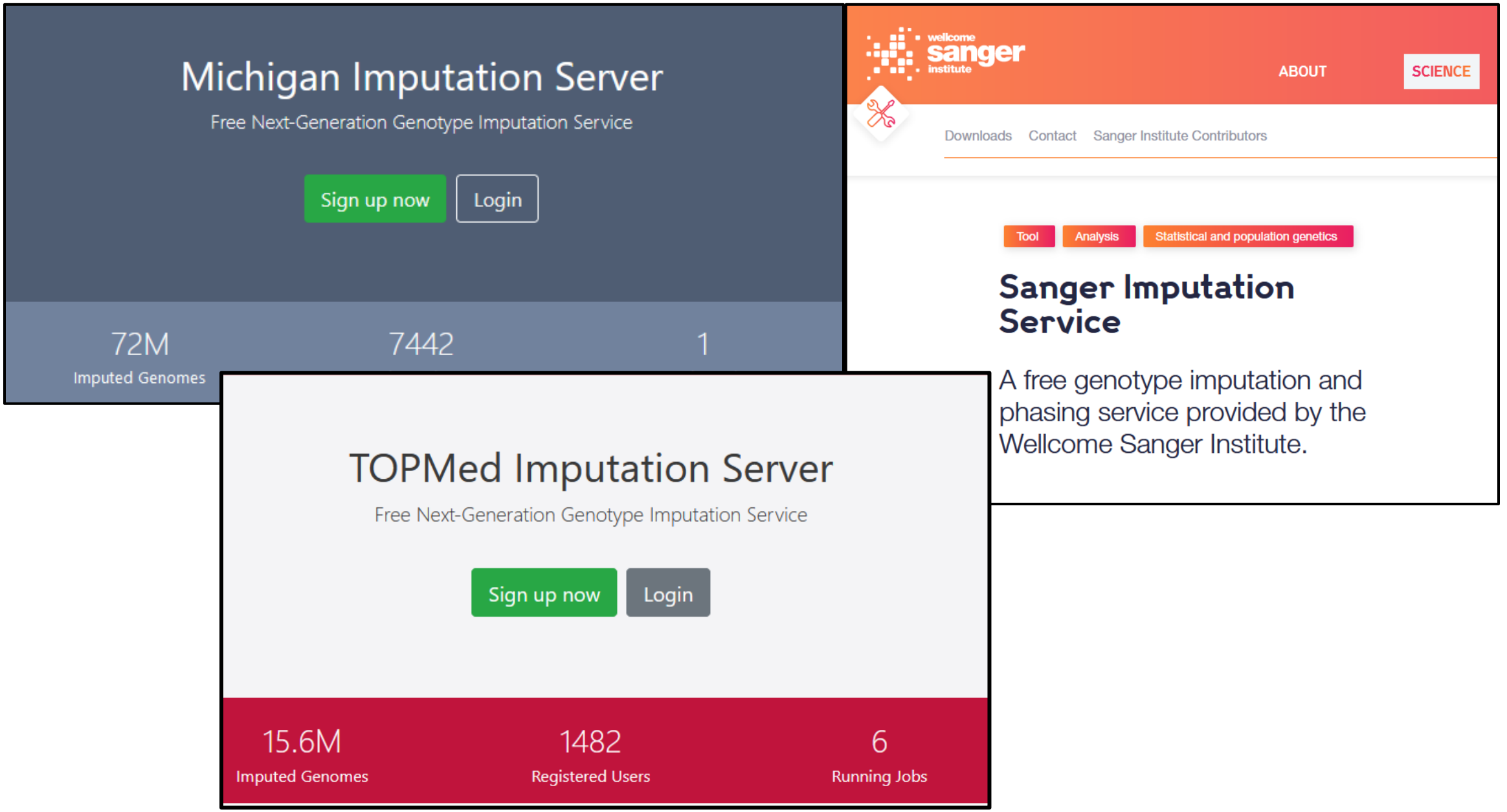
What do we do to get it ready to use?

- Imputation
  - We pay to genotype ~.4-1M markers
  - Through imputation we can get data for ~9M extra markers for free*


Reference set of haplotypes, for example, HapMap
Genotype data with missing data at untyped SNPs (grey question marks)
The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)
Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

Imputation reference sets

- Publicly Available References
  - HapMap
  - 1KGP – phase 3 version v5
- References only available via custom imputation servers
  - HRC - 64,976 haplotypes 39,235,157 SNPs
  - CAPPA – African American/Caribbean
  - Multi-ethnic HLA
  - Genome Asia Pilot - GAsP
  - TopMed - 97,256 haplotypes 308,107,085 SNPs (b38)


Michigan Imputation Server
Free Next-Generation Genotype Imputation Service
Sign up now
Login
72M
Imputed Genomes
7442
1
wellcome sanger institute
ABOUT
SCIENCE
Downloads
Contact
Sanger Institute Contributors
Tool
Analysis
Statistical and population genetics
Sanger Imputation Service
A free genotype imputation and phasing service provided by the Wellcome Sanger Institute.
TOPMed Imputation Server
Free Next-Generation Genotype Imputation Service
Sign up now
Login
15.6M
Imputed Genomes
1482
Registered Users
6
Running Jobs

- DIY – Use a cookbook!

  http://genome.sph.umich.edu/wiki/Minimac3_Imputation_Cookbook OR
  http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook

- UMich Imputation Server
  - https://imputationserver.sph.umich.edu/
- Sanger Imputation Server
  - https://imputation.sanger.ac.uk/
- TOPMed Imputation Server
  - https://imputation.biodatacatalyst.nhlbi.nih.gov/

- 2 commonly used genotype output formats
  - Hard call or best guess
  - Dosage data (most common – 1 number per SNP, 1-2)

```
##fileformat=VCFv4.1
##filedate=2015.7.12
##source=Minimac3
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1 ">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  A0001_A0001     A0003_A0003     A0004_A0004     A0007_A0007     A0008_A0008     A0009_A0009     A0010_
10      27754636        10:27754636     C       G       .       PASS    MAF=0.00032;R2=0.81788  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754678        10:27754678     G       A       .       PASS    MAF=0.00042;R2=0.77190  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754849        10:27754849     C       G       .       PASS    MAF=0.00001;R2=0.00262  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754857        10:27754857     T       C       .       PASS    MAF=0.00120;R2=0.72916  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754954        10:27754954     T       C       .       PASS    MAF=0.11410;R2=0.97841  GT:DS:GP        1/1:2.000:0.000,0.000,1.000     1/1:2.000:0.000,0.000,1.000     1/1:2.000:0.000,0.000,1.000
10      27755014        10:27755014     G       T       .       PASS    MAF=0.00000;R2=0.00082  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755016        10:27755016     C       T       .       PASS    MAF=0.00003;R2=0.01909  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755047        10:27755047     T       C       .       PASS    MAF=0.02255;R2=0.87665  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755175        10:27755175     C       T       .       PASS    MAF=0.00004;R2=0.13821  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755281        10:27755281     C       T       .       PASS    MAF=0.00061;R2=0.86168  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755330        10:27755330     A       G       .       PASS    MAF=0.00273;R2=0.90295  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755439        10:27755439     A       C       .       PASS    MAF=0.00000;R2=0.00138  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755489        10:27755489     C       A       .       PASS    MAF=0.00003;R2=0.39172  GT:DS:GP        0/0:0.000:1.000.0.000.0.000     0/0:0.000:1.000.0.000.0.000     0/0:0.000:1.000.0.000.0.000
```

Not all markers are well imputed

Imputation quality evaluation

Minimac hides each of the genotyped SNPs in turn and then calculates 3 statistics:

- looRSQ - this is the estimated rsq for that SNP (as if SNP weren't typed).
- empR - this is the empirical correlation between true and imputed genotypes for the SNP. If this is negative, the SNP alleles are probably flipped.
- empRSQ - this is the actual R2 value, comparing imputed and true genotypes.

These statistics can be found in the *.info file

Be aware that, unfortunately, imputation quality statistics are not directly comparable between different imputation programs (MaCH/minimac vs. Impute vs. Beagle etc.).

SNP	Al1	Al2	Freq1	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose1	Dose2
1:10583	G	A	0.79288	0.20712	0.79288	-0.00000	-	-	-	-	-	-
1:10611	C	G	0.97889	0.02111	0.97889	0.00000	-	-	-	-	-	
1:13302	C	T	0.86280	0.13720	0.86280	-0.00000	-	-	-	-	-	-
1:13327	G	C	0.96042	0.03958	0.96042	-0.00000	-	-	-	-	-	-
1:95207182	T	C	0.99547	0.00453	0.99547	0.10108	-	-	-	-	-	-
1:95207382	T	T	1.00000	0.00000	1.00000	0.00000	-	-	-	-	-	-
1:95207442	C	T	0.62754	0.37246	0.99999	1.00507	Genotyped	0.98810	0.99822	0.99645	0.99484	0.00421
1:95207524	G	A	0.78061	0.21939	1.00000	1.00511	Genotyped	1.00059	1.00000	1.00000	0.99924	0.00083
1:95207532:TG_T	R	D	0.78620	0.21380	0.99441	0.97729	-	-	-	-	-	-
1:95207558	C	T	0.99399	0.00601	0.99399	0.05165	-	-	-	-	-	-
1:95207633	A	C	0.93366	0.06634	0.99998	1.00482	Genotyped	0.94847	0.99901	0.99802	0.99621	0.00372
1:95207846	G	T	0.98937	0.01063	0.98942	0.31316	-	-	-	-	-	-

- Imputation accuracy is calculated differently by the two main imputation programs
- But is highly correlated and conceptually the same

The MACH $\hat{r}^2$ measure

This is the ratio of the empirically observed variance of the allele dosage to the expected binomial variance at Hardy-Weinberg equilibrium. At the jth SNP this is defined as

$$\hat{r}_j^2 = \begin{cases} \dfrac{\frac{\sum_{i=1}^N e_{ij}^2}{N} - \left(\frac{\sum_{i=1}^N e_{ij}}{N}\right)^2}{2\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases} \quad (1)$$

When all the genotypes are predicted with high certainty this ratio will be close to 1, although it can go above 1 (Figure 1). As the amount of uncertainty increases the allele dosages will tend to 2θ, the empirical variance will tend to 0 and so $\hat{r}^2$ tends to 0.

The IMPUTE info measure I_A

This is based on measuring the relative statistical information about the population allele frequency, θ_j. If the G_{ij}'s were observed then the full data likelihood is given by

$$L(\theta_j) = \prod_{i=1}^N \theta_j^{G_{ij}} (1-\theta_j)^{2-G_{ij}} \quad (10)$$

For this likelihood the score and information are given by

$$U(\theta_j) = \frac{d \log L(\theta_j)}{d\theta_j} = \frac{X - 2N\theta_j}{\theta_j(1-\theta_j)} \quad (11)$$

$$I(\theta_j) = \frac{-d^2 \log L(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N - X}{(1-\theta_j)^2} \quad (12)$$

The IMPUTE info measure is based on the same idea used to calculate the SNPTEST information measure i.e. the ratio of the observed and complete information.

$$I_A = \frac{\mathbb{E}_{G.j}[I(\hat{\theta})] - V_G[U(\hat{\theta})]}{\mathbb{E}_{G.j}[I(\hat{\theta})]} \quad (13)$$

where the expectations are taken over the imputed genotype distribution and evaluated at the allele frequency estimate, $\hat{\theta}_j$. The exact terms are given by

$$\mathbb{E}_{G.j}[I(\hat{\theta})] = \frac{2N}{\hat{\theta}(1-\hat{\theta})} \quad (14)$$

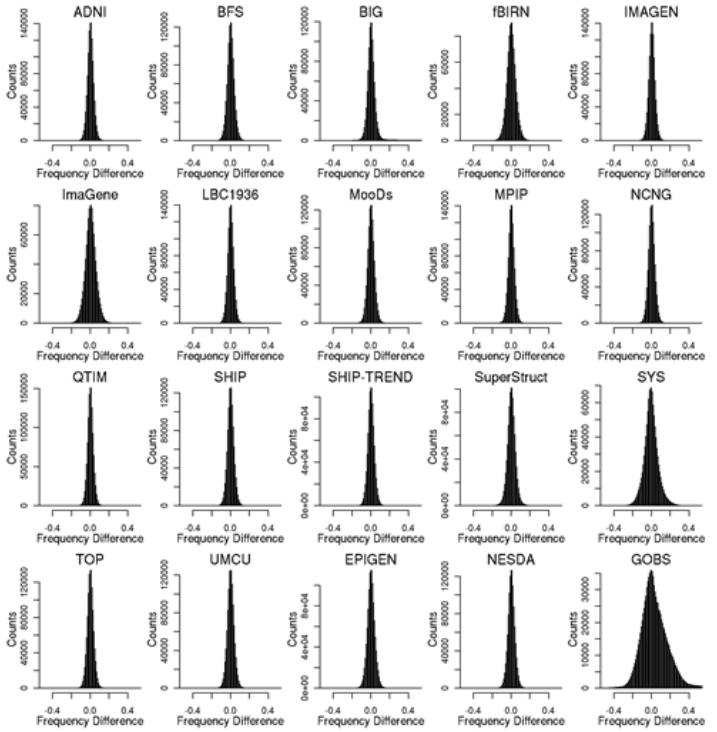
$$V_G[U(\hat{\theta})] = \frac{\sum_{i=1}^N (f_{ij} - e_{ij}^2)}{\hat{\theta}^2(1-\hat{\theta})^2} \quad (15)$$

so that

$$I_A = \begin{cases} 1 - \frac{\sum_{i=1}^N (f_{ij} - e_{ij}^2)}{2N\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1. \end{cases} \quad (16)$$

So I_A is bounded above at 1 and will equal 0 when the sample mean variance of the imputed genotypes equals the variance you would expect if alleles where sampled with frequency $\hat{\theta}$.

- After imputation you need to check that it worked and the data look ok
- Things to check
  - Plot r^2 across each chromosome look to see where it drops off
  - Plot MAF-reference MAF



Estimating relatedness

Slide Credit: Loci Yengo

Genetic relationship matrices

- Genetic relationship matrices (GRM) are important tools for estimating heritability
- Yesterday we used family level GRM based on expected relatedness

MZ		DZ	
1	1	1	.5
1	1	.5	1

Genetic relationship matrices

- We can also calculate a GRM using SNP data.
- There are many ways to calculate a GRM using SNP data
- Common to use the standard estimator implemented in the software **GCTA** (but often calculated using Plink)

$$\hat{\pi}_{jk} = \frac{1}{m}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

where, x_{ij} and x_{ik} are the minor allele count (x_{ij}, x_{ik} = 0,1 or 2) at SNP i for individuals j and k respectively, p_i the minor allele frequency (MAF) of SNP I and m the number of SNPs used to calculate the GRM.

Example of GRM between *N*=3 individuals m=1000 SNPs

```
[$bash] zless myGRM.grm.gz
1 1 1000  0.99
1 2 1000 -0.01
1 3 1000  0.01
2 2 1000  1.03
2 3 1000  0.03
3 3 1000  1.01
```

Am J Hum Genet. 2011 Jan 7; 88(1): 76–82.
doi: 10.1016/j.ajhg.2010.11.011

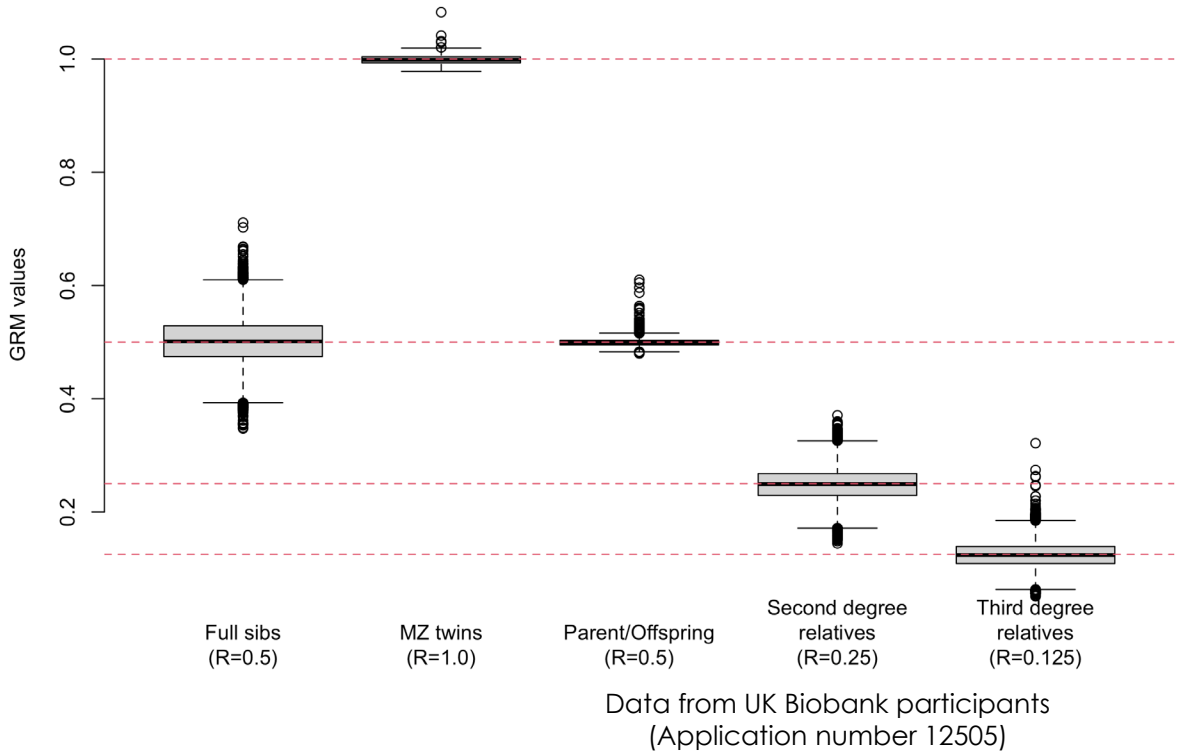
PMCID: PMC3014363
PMID: 21167468

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

The expectation (over a large sample of relatives) of the $\hat{\pi}_{jk}$ = the expected relatedness

Observed relatedness may be still vary within a type of pedigree relationship.



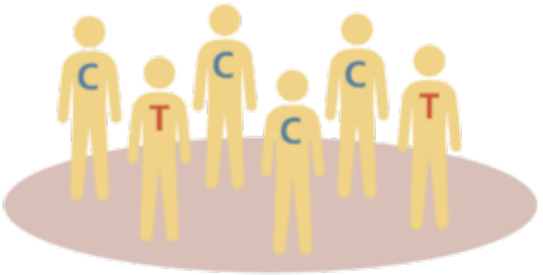
Data from UK Biobank participants
(Application number 12505)

Genetic relationship matrices

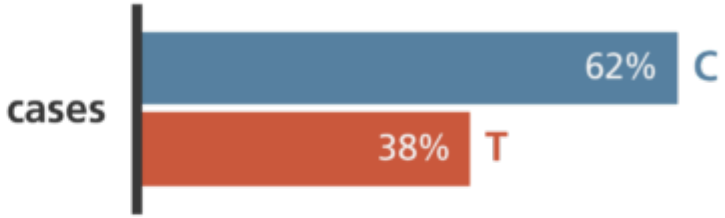
- Later today we will use relatedness calculated from SNPs in an OpenMx model to estimate heritability
- On Thursday we will use relatedness calculated from SNPs to run Trio-GCTA

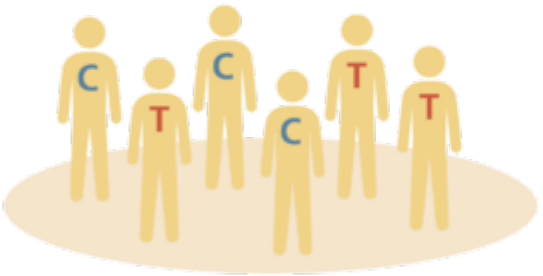
Analysing our genotypic data…

Association analyses



cases (n=1,000)
people with heart disease

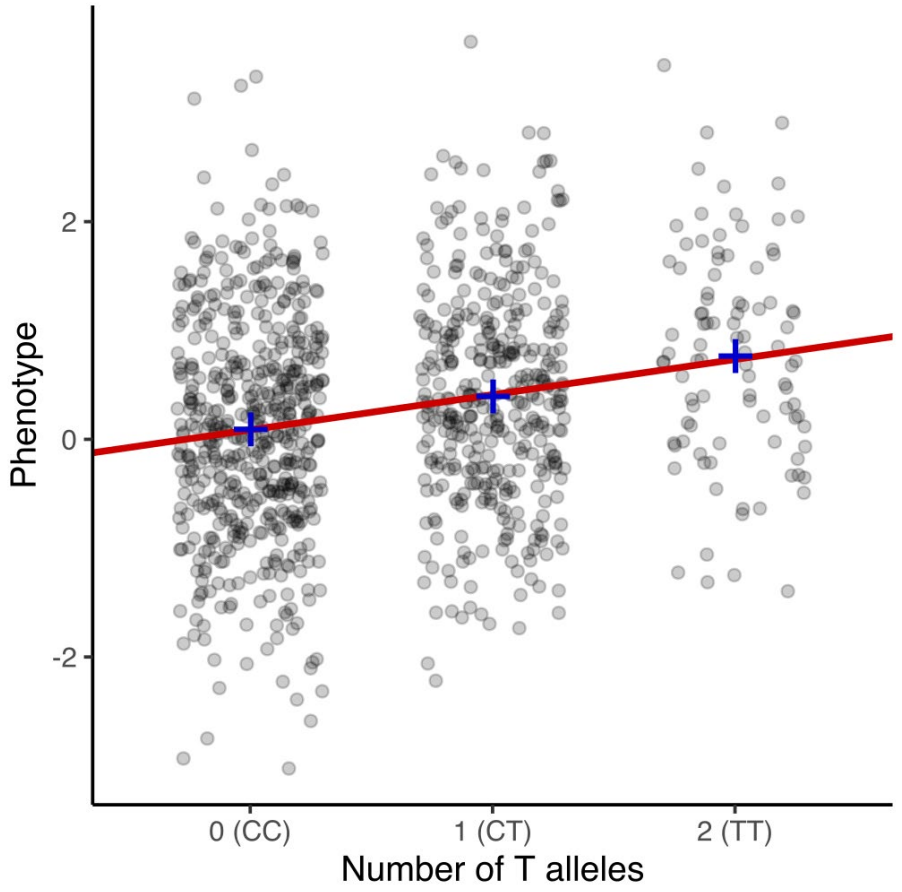




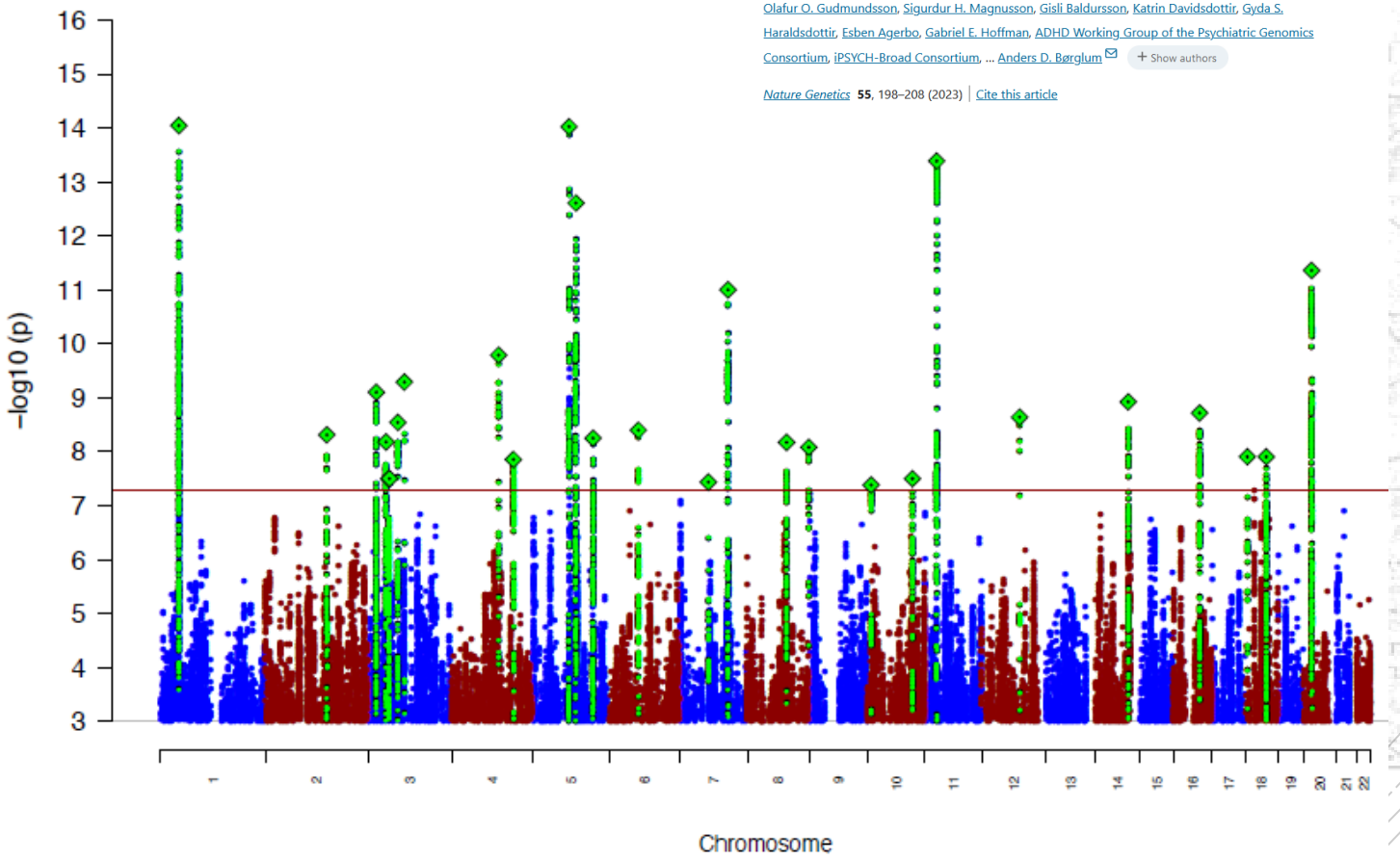
controls (n=1,000)
people without heart disease

https://www.yourgenome.org/theme/genome-wide-association-studies/

Association analyses



Genome-wide Association Study (GWAS)



Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains

Ditte Demontis, G. Bragi Walters, Georgios Athanasiadis, Raymond Walters, Karen Therrien, Trine Tollerup Nielsen, Leila Farajzadeh, Georgios Voloudakis, Jaroslav Bendl, Biau Zeng, Wen Zhang, Jakob Grove, Thomas D. Als, Jinjie Duan, F. Kyle Satterstrom, Jonas Bybjerg-Grauholm, Marie Bækved-Hansen, Olafur O. Gudmundsson, Sigurdur H. Magnusson, Gisli Baldursson, Katrin Davidsdottir, Gyda S. Haraldsdottir, Esben Agerbo, Gabriel E. Hoffman, ADHD Working Group of the Psychiatric Genomics Consortium, iPSYCH-Broad Consortium, ... Anders D. Børglum + Show authors

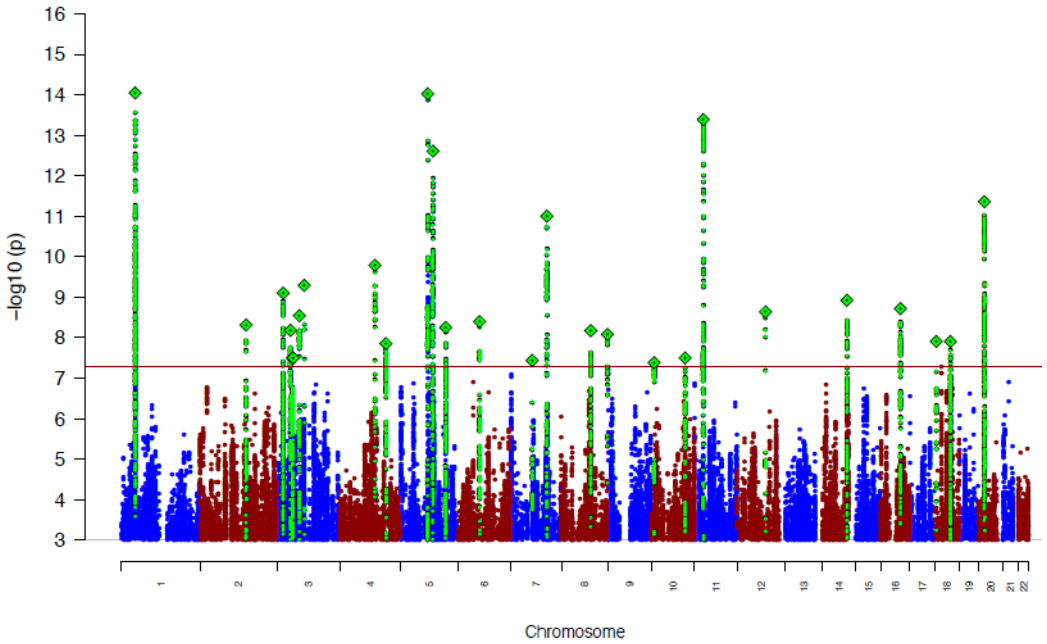
Nature Genetics **55**, 198–208 (2023) | Cite this article

Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains

Ditte Demontis, G. Bragi Walters, Georgios Athanasiadis, Raymond Walters, Karen Therrien, Trine Tollerup Nielsen, Leila Farajzadeh, Georgios Voloudakis, Jaroslav Bendl, Biau Zeng, Wen Zhang, Jakob Grove, Thomas D. Als, Jinjie Duan, F. Kyle Satterstrom, Jonas Bybjerg-Grauholm, Marie Bækved-Hansen, Olafur O. Gudmundsson, Sigurdur H. Magnusson, Gisli Baldursson, Katrin Davidsdottir, Gyda S. Haraldsdottir, Esben Agerbo, Gabriel E. Hoffman, ADHD Working Group of the Psychiatric Genomics Consortium, iPSYCH-Broad Consortium, ... Anders D. Børglum + Show authors

Nature Genetics **55**, 198–208 (2023) | Cite this article

38,899 people living with ADHD, 186,843 without

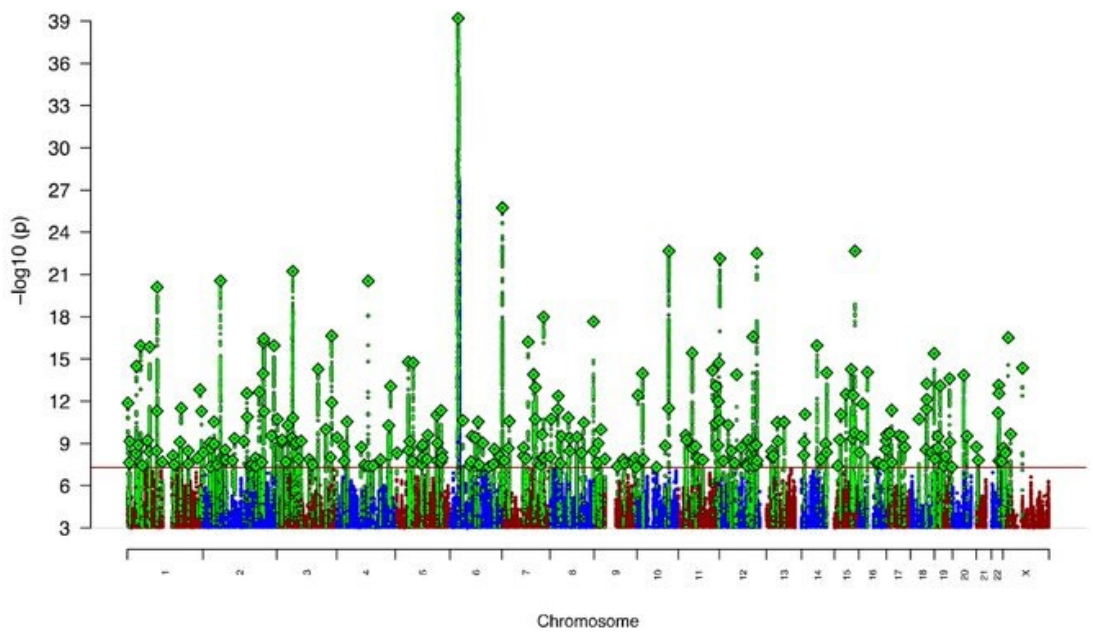


Mapping genomic loci implicates genes and synaptic biology in schizophrenia

Vassily Trubetskoy, Antonio F. Pardiñas, Ting Qi, Georgia Panagiotaropoulou, Swapnil Awasthi, Tim B. Bigdeli, Julien Bryois, Chia-Yen Chen, Charlotte A. Dennison, Lynsey S. Hall, Max Lam, Kyoko Watanabe, Oleksandr Frei, Tian Ge, Janet C. Harwood, Frank Koopmans, Sigurdur Magnusson, Alexander L. Richards, Julia Sidorenko, Yang Wu, Jian Zeng, Jakob Grove, Minsoo Kim, Zhiqiang Li, Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, The SynGO Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium + Show authors

Nature **604**, 502–508 (2022) | Cite this article

74,776 people living with Schizophrenia, 101,023 without



- GWAS typically uses a significance threshold of $5x10^{-8}$
- This is based on the approximate number of independent tests conducted
  - Caveat: Many/most follow-up analyses use full distribution of effect estimates and don't restrict to significant loci

To account for multiple testing in genome-wide association studies (GWAS), a fixed *P*-value threshold of 5×10^{-8} is widely used to identify association between a common genetic variant and a trait of interest. Risch and Merikangas (1996) suggested this strict *P*-value threshold for studying the genetics of complex diseases due to the many false positive discoveries reported by candidate gene studies at that time. Later, the International HapMap Consortium (Altshuler and Donnelly 2005), Dudbridge and Gusnanto (2008), and Pe'er *et al.* (2008) independently suggested near-identical thresholds for common variant (minor allele frequency [MAF] >5%) GWAS. Each group of investigators sought to control the family-wise error rate

Zhongsheng Chen, Michael Boehnke, Xiaoquan Wen, Bhramar Mukherjee, Revisiting the genome-wide significance threshold for common variant GWAS, *G3 Genes|Genomes|Genetics*, Volume 11, Issue 2, February 2021, jkaa056, https://doi.org/10.1093/g3journal/jkaa056

Finding variants influencing the trait is important BUT

- The next steps to determine function and mechanism are $$$$$
- Better to have *robust* findings that will replicate than find *more* variants

Publicly available summary statistics

- Most GWAS efforts make their results publicly available
  - http://www.nealelab.is/uk-biobank
  - https://pgc.unc.edu/for-researchers/download-results/
  - https://enigma.ini.usc.edu/research/download-enigma-gwas-results/

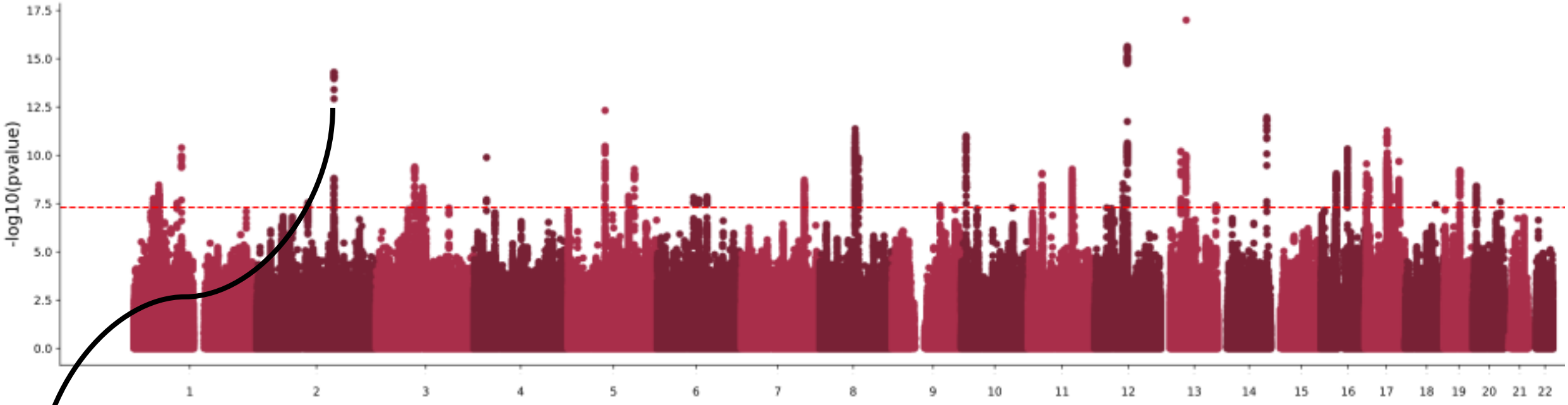
GWAS output = PRS input
GWAS output = MR input
GWAS output = LDscore input
GWAS output = SNP h^2 input
…

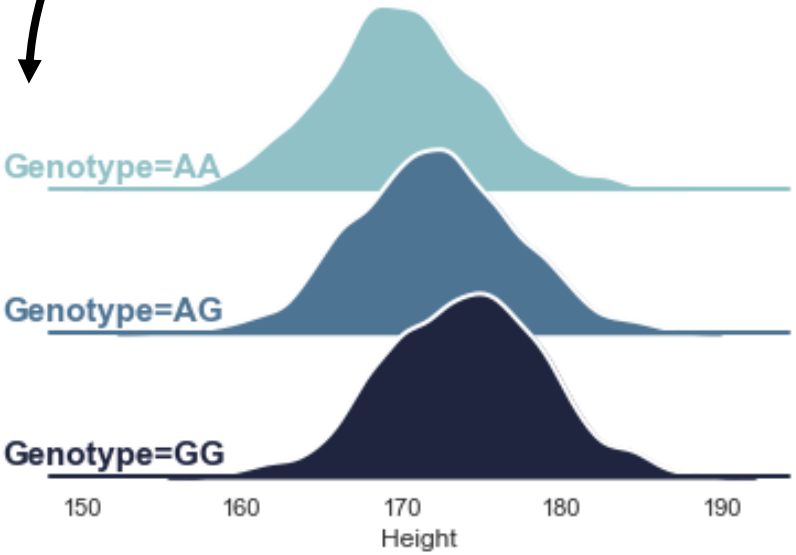
Polygenic Scores

- Perline will be talking more about this tomorrow morning
  - 8am

Polygenic Scores

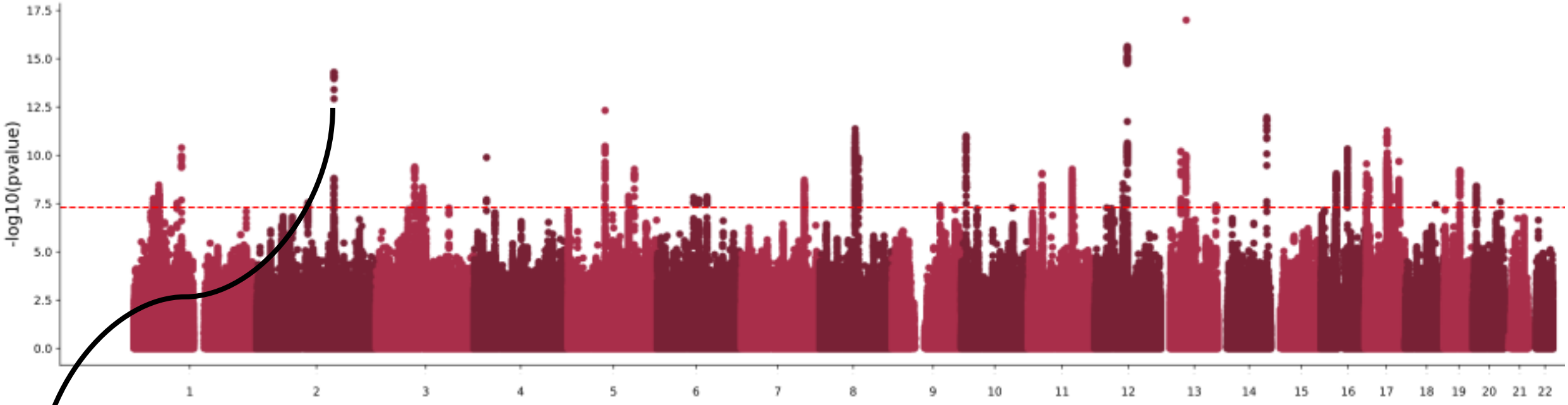
- Many names same concept
  - Polygenic Risk Scores (PRS)
  - Polygenic Scores (PGS)
  - Allelic Scores
  - Polygenic Index (PCI
- Polygenic risk score – Weighted sum of alleles which quantify the effect of several genetic variants on an individual's phenotype.

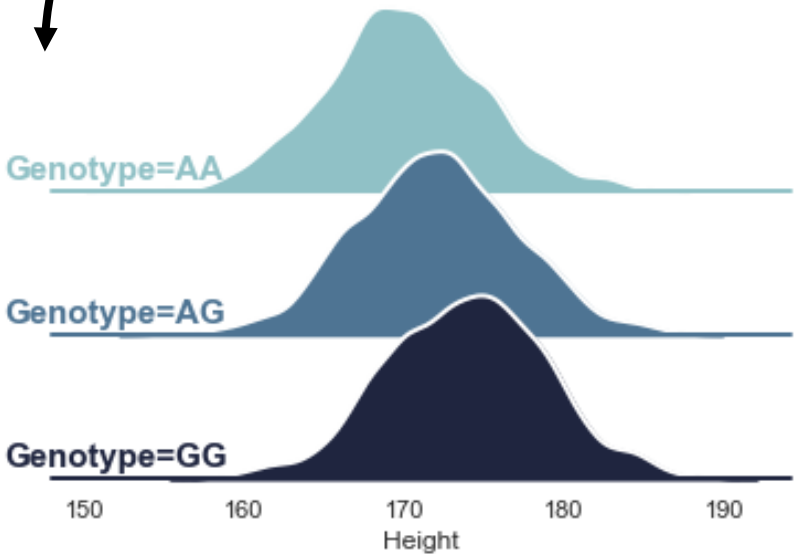




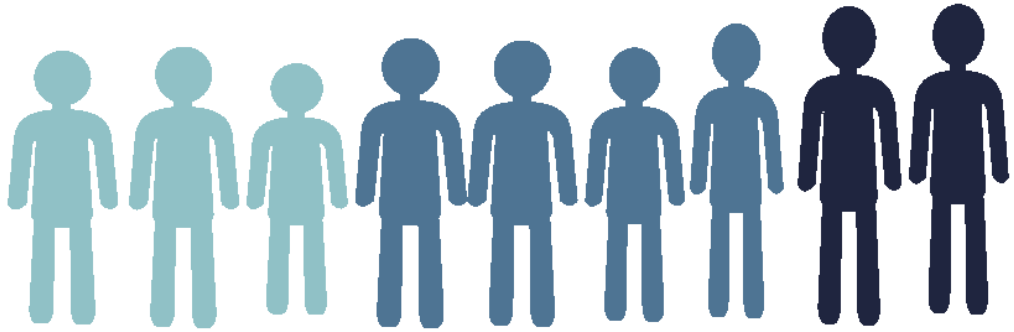
- Average increase of 2cm per copy of the G allele. So the effect size of this variant would be approximately 2.

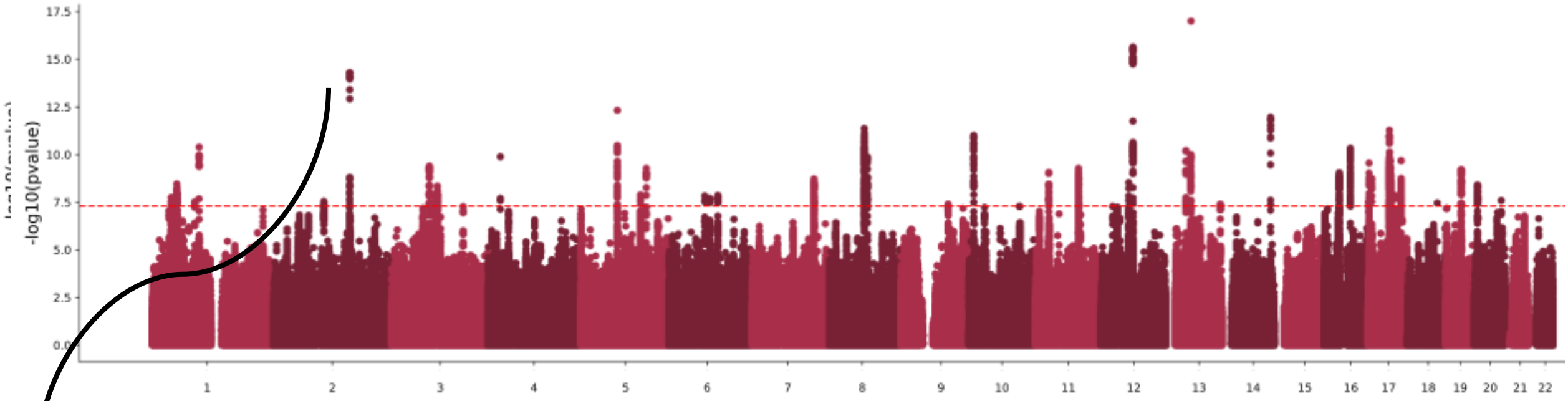
Slide credit: Adrian Campos

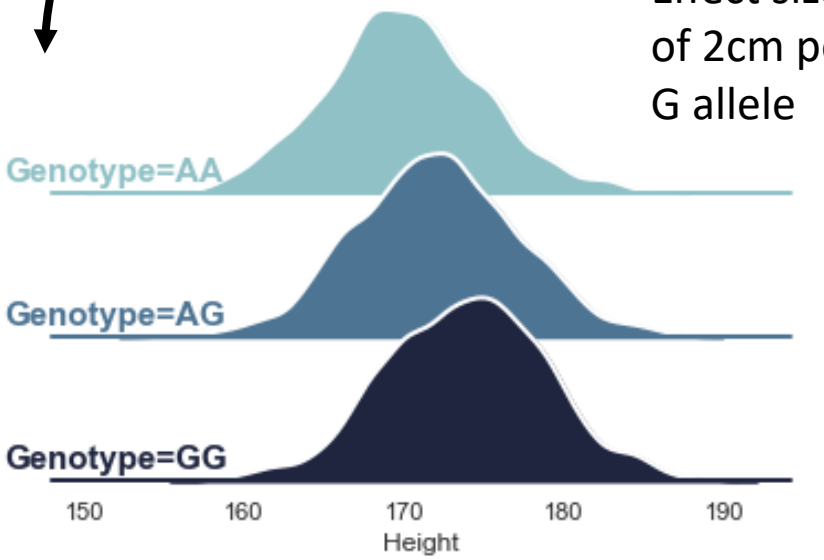




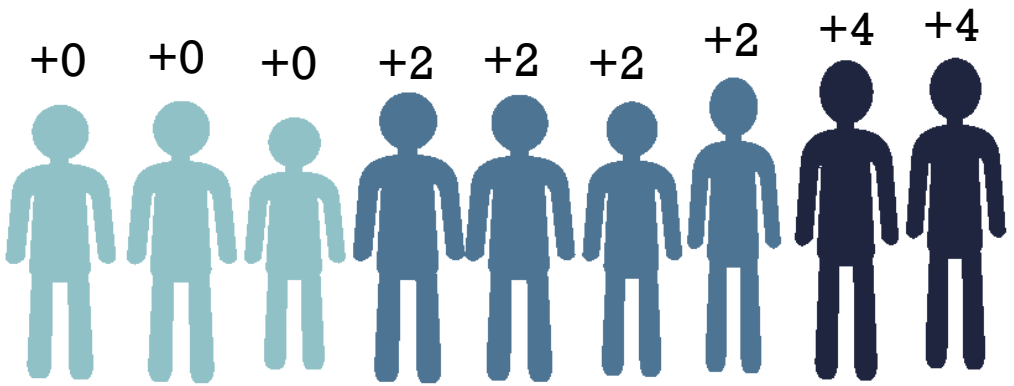
In a new sample we would expect AG individuals to be on average 2cm taller than AA and 2cm shorter than GG

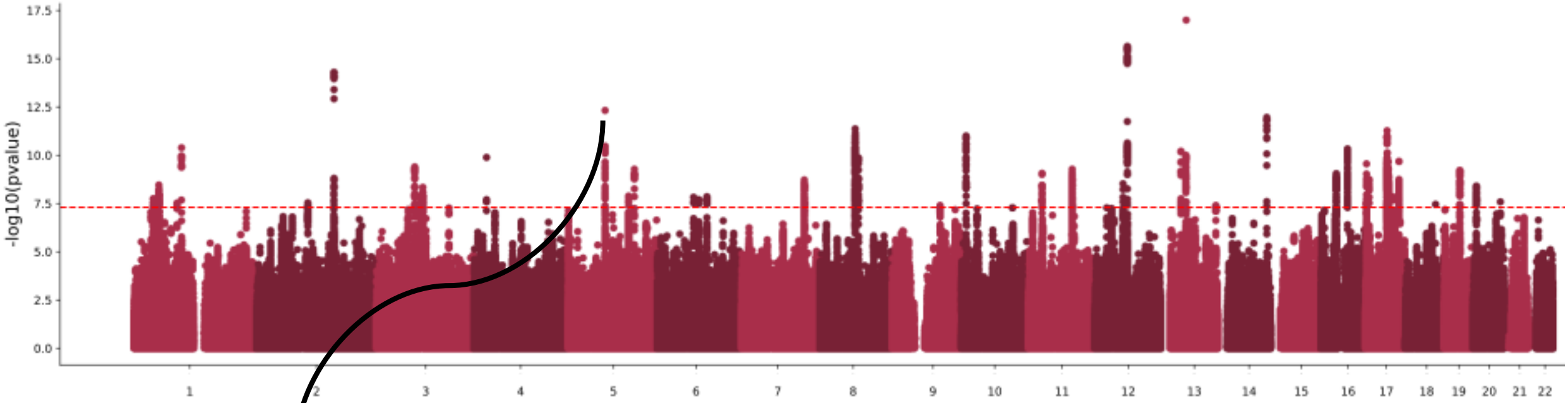




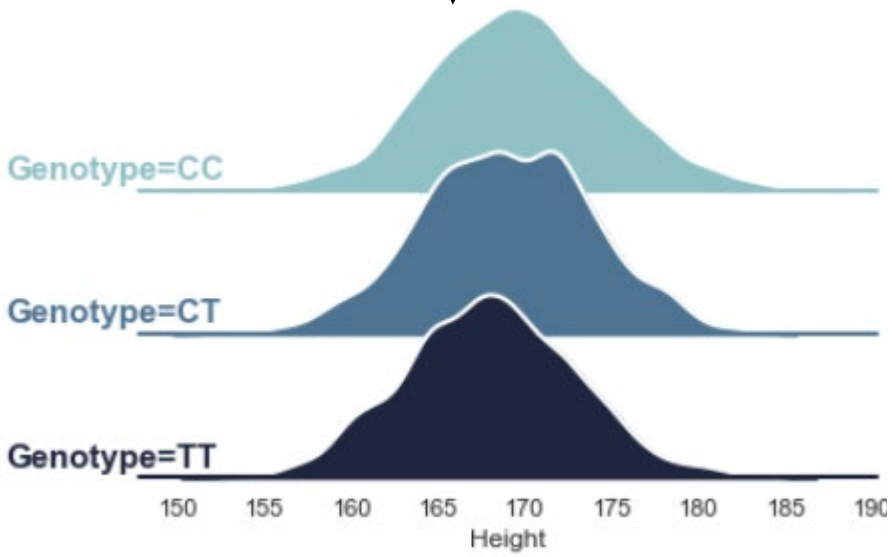


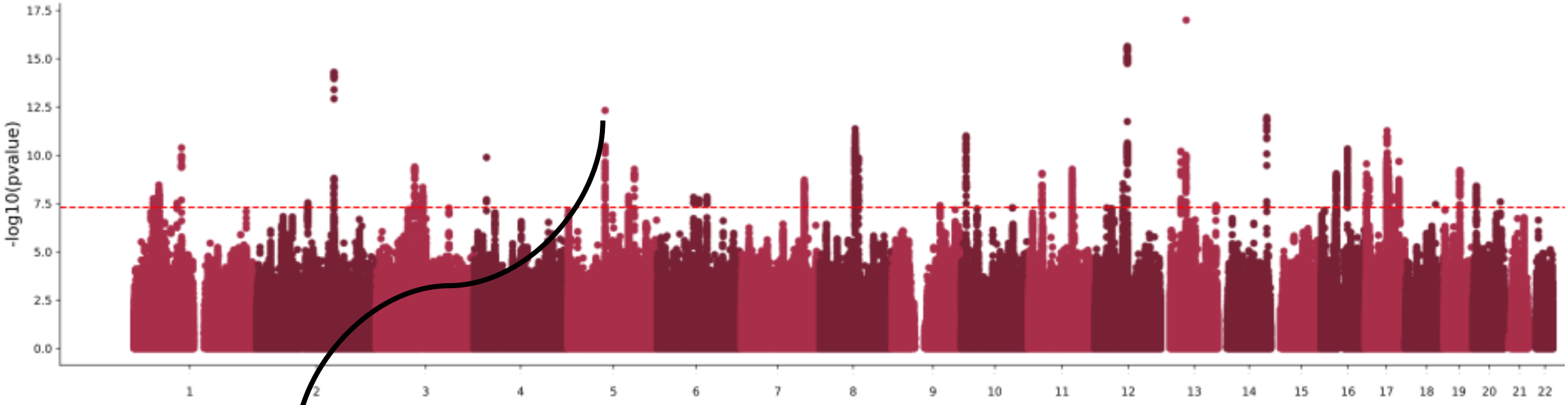
Effect size of 2cm per G allele



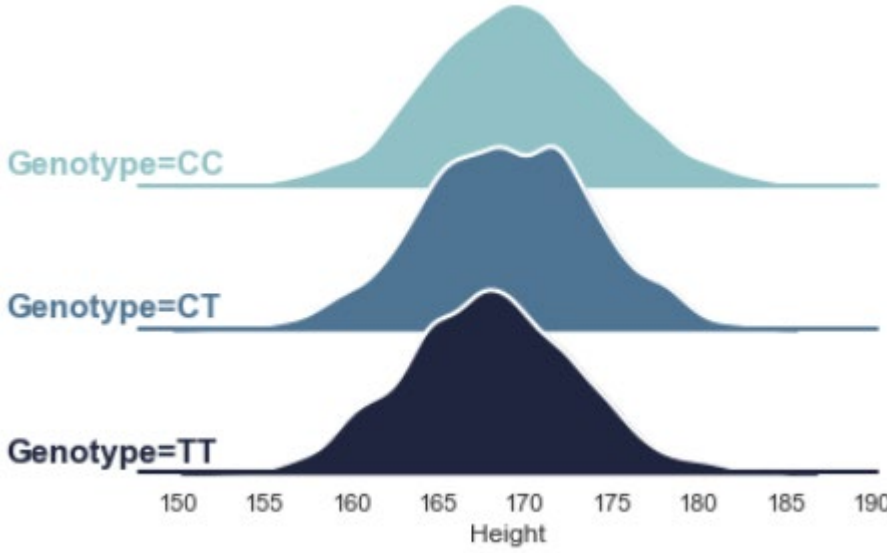


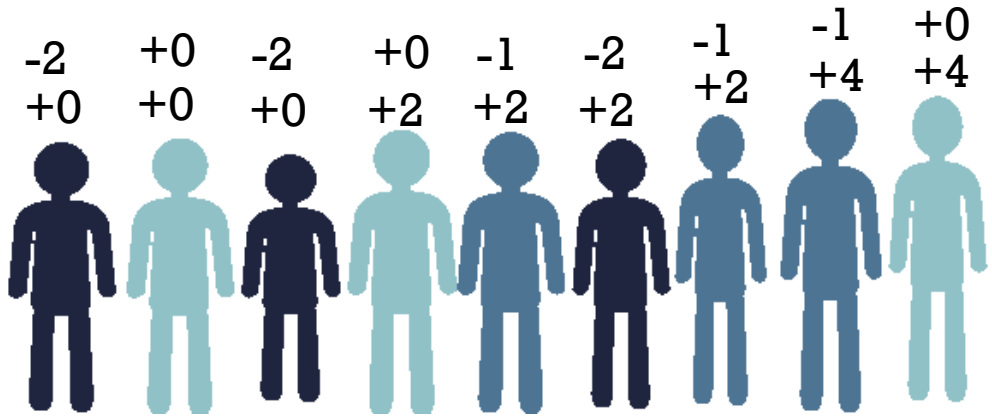
Effect size of -1 per T allele

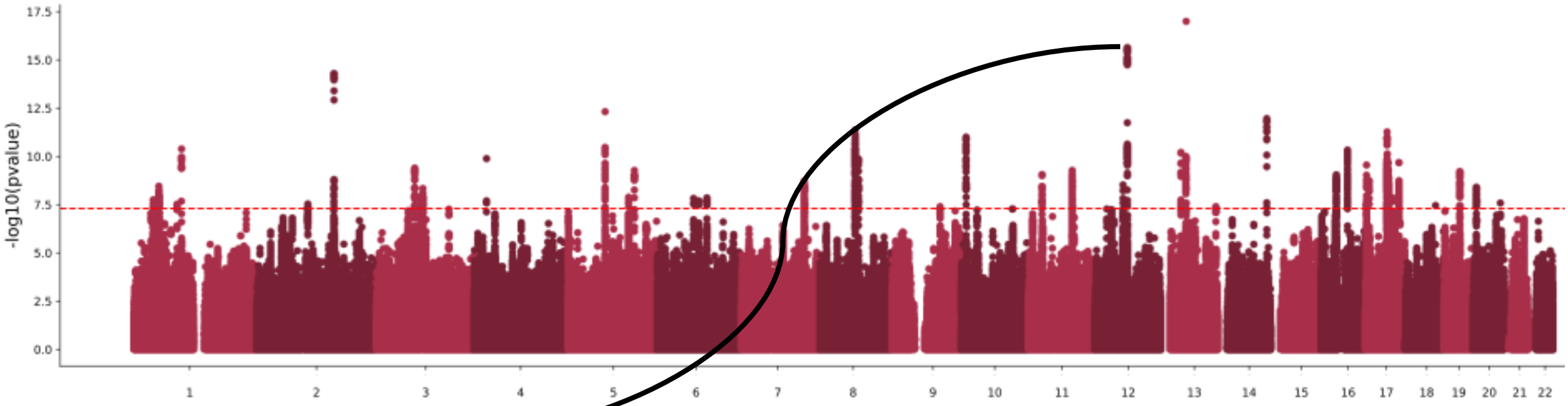




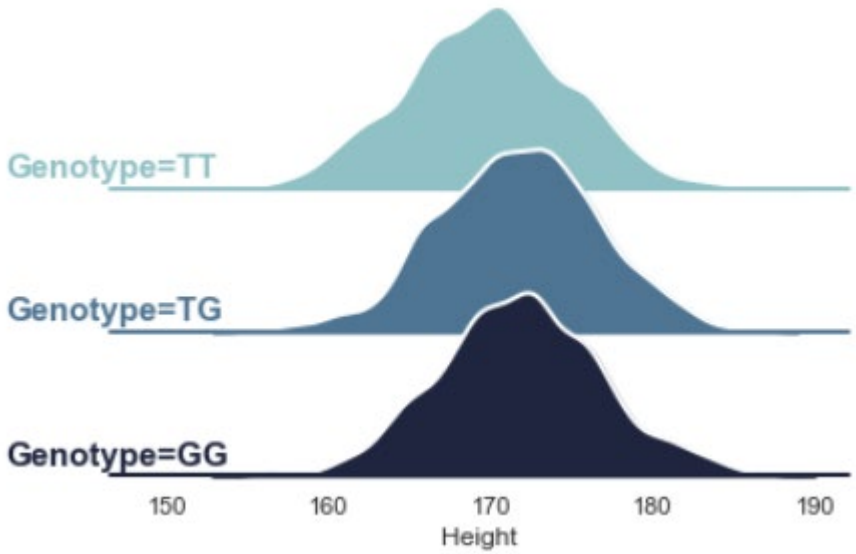
Effect size of -1 per T allele

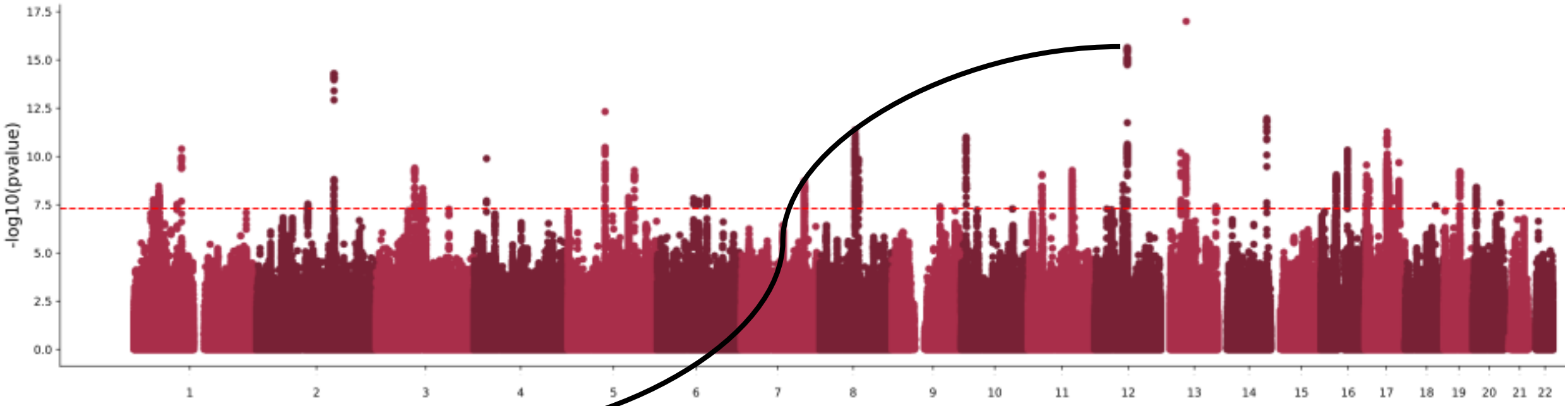




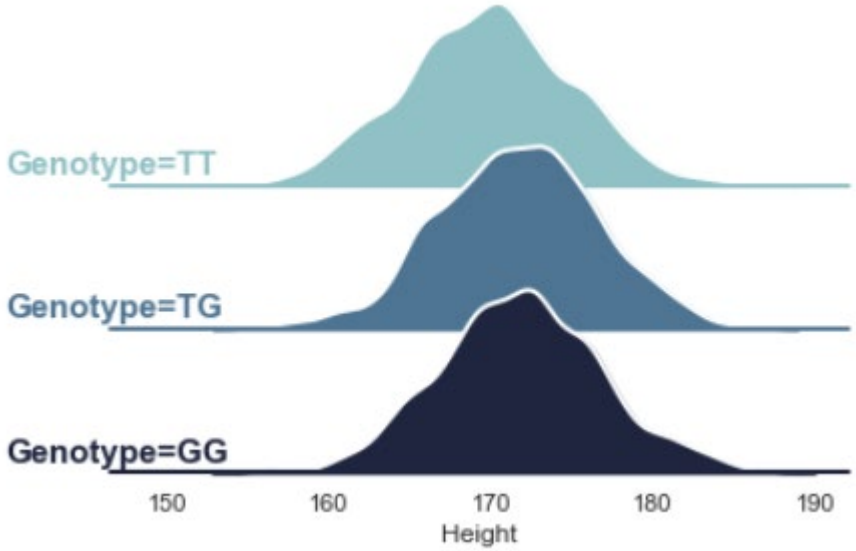


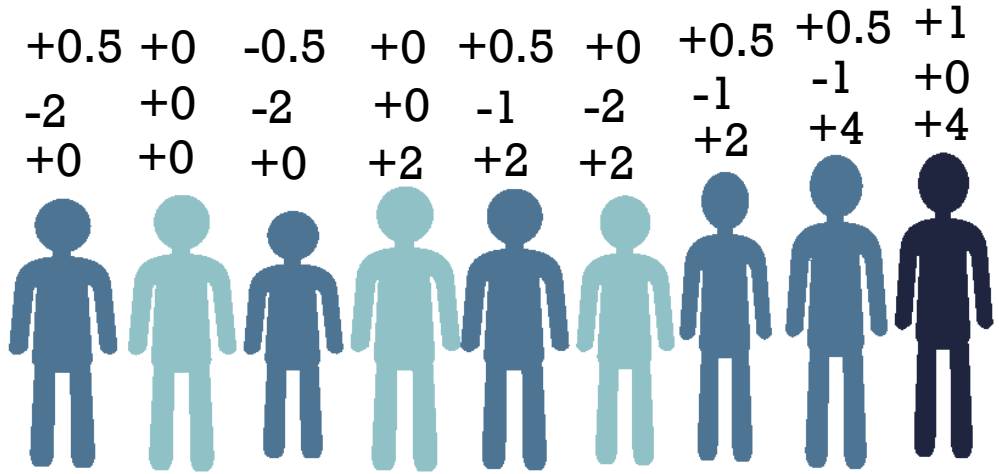
Effect size of +0.5 per G allele

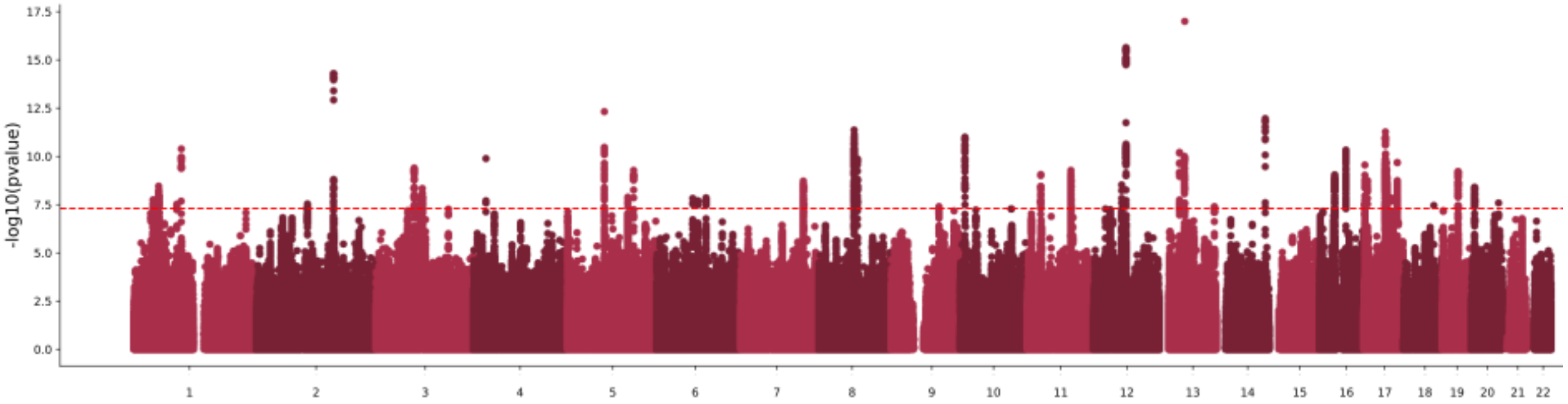




Effect size of +0.5 per G allele







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

Polygenic Scores

- A couple of important gotchas
  - You need to make sure the weights are being applied to the right allele (ambiguous snps)
  - The individuals you are calculating the PRS for needs to be completely independent from the individuals in the GWAS

Polygenic Scores

- Numerous methods
  - Clumping and thresholding
  - Bayesian approaches

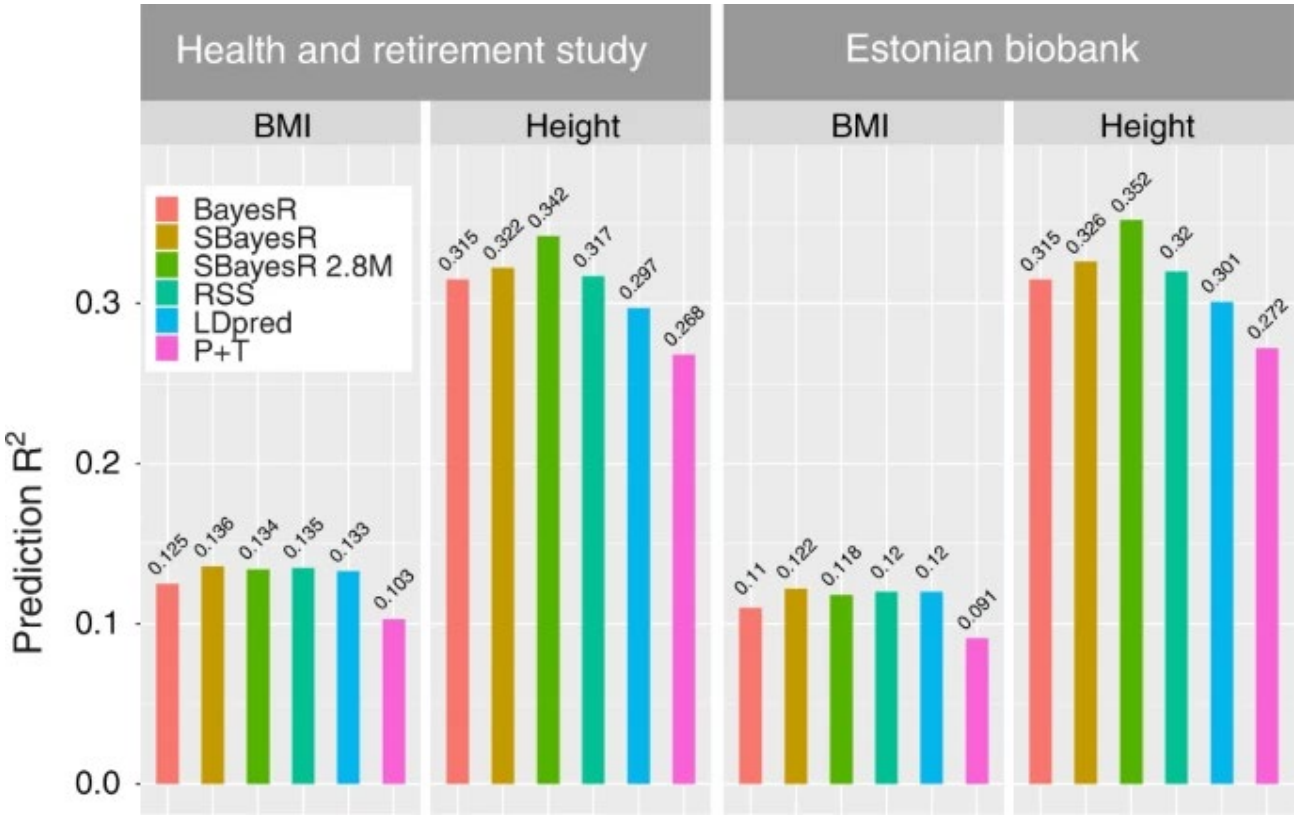
Article | Open Access | Published: 08 November 2019

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

- 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



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

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

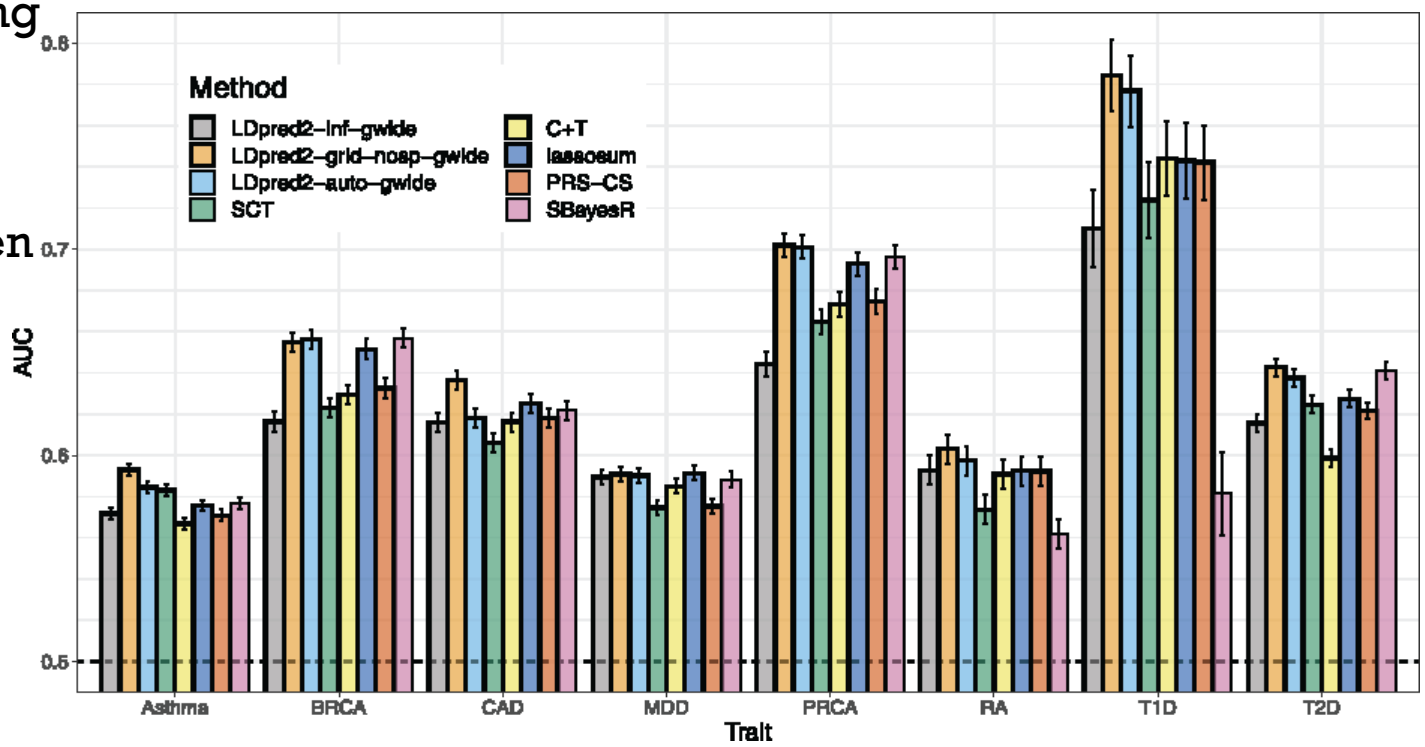
$$\widehat{\gamma}_{\text{joint}} = \boldsymbol{S}^{-1}\boldsymbol{R}^{-1}\boldsymbol{S}\widehat{\gamma}_{\text{marg}}$$

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).

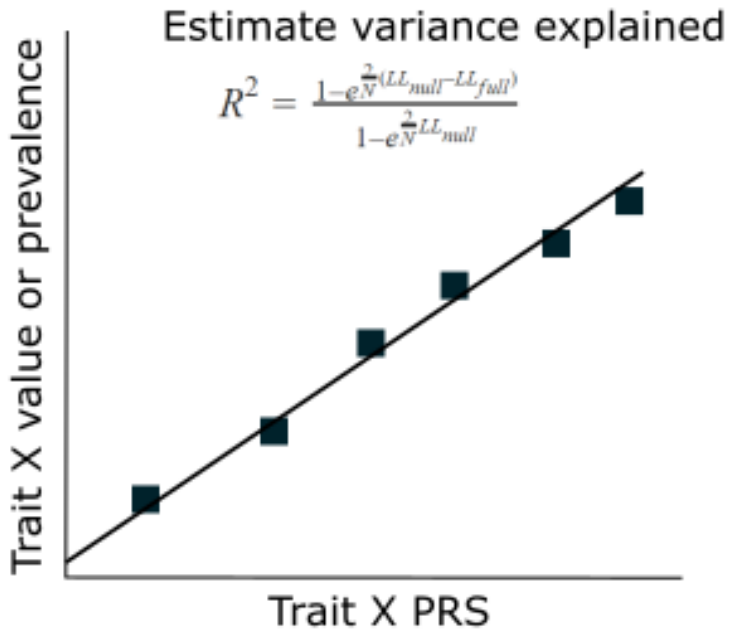


Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5424–5431

Polygenic Scores

- Working across samples
  - Replicating the finding
    - Requires the same snps, same weights etc
    - Restrict the pool of possible SNPs to those available across all cohorts
  - Replicating the concept
    - Different clumping and/or weights

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:

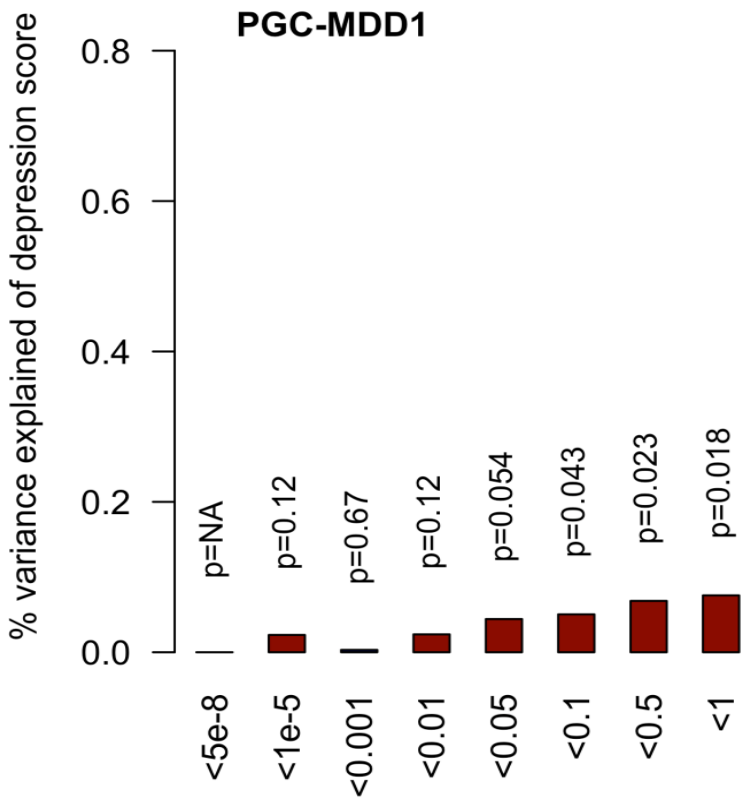
> Is it continuous – linear regression

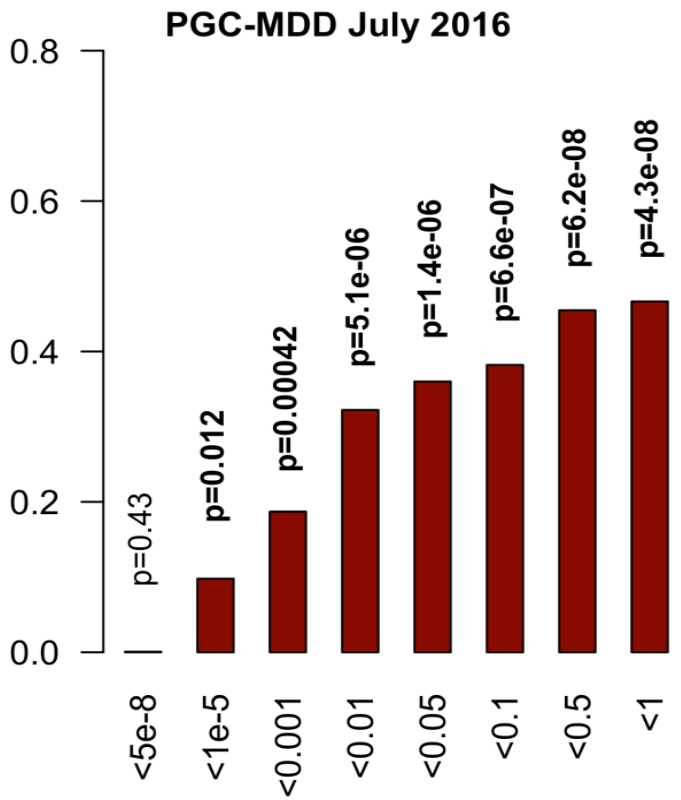
> Binary – logistic or probit regression

> Ordinal – cumulative linked mixed models

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

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

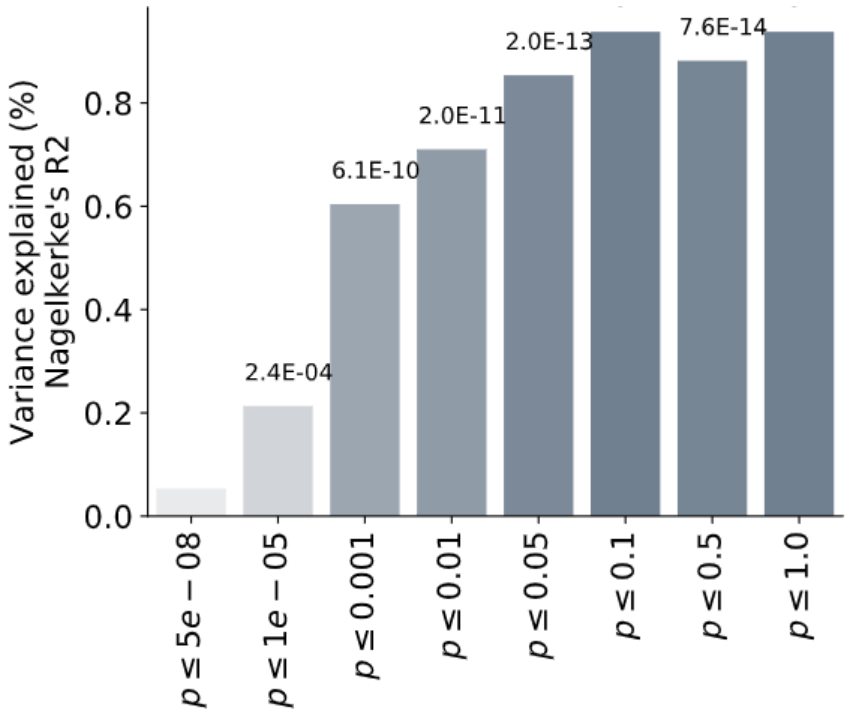


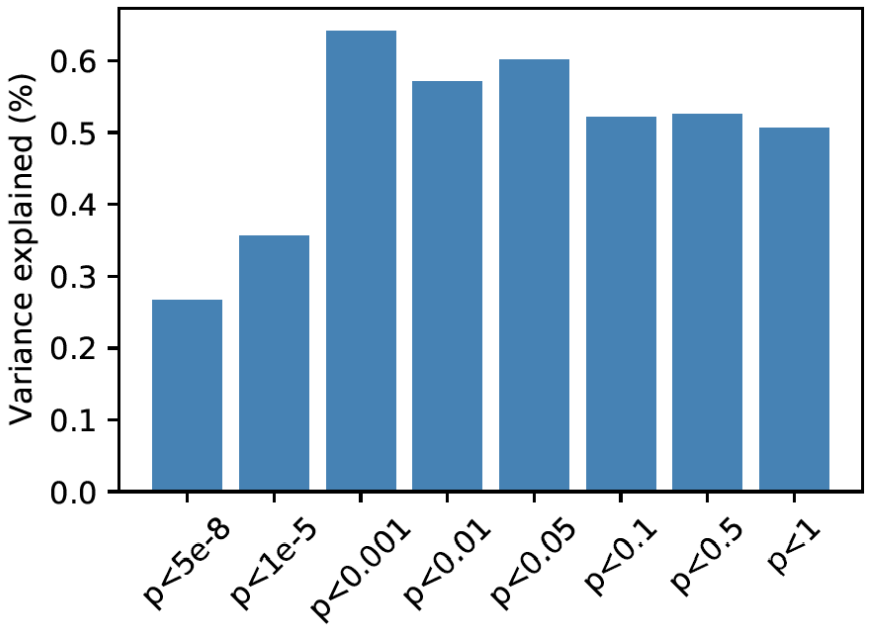


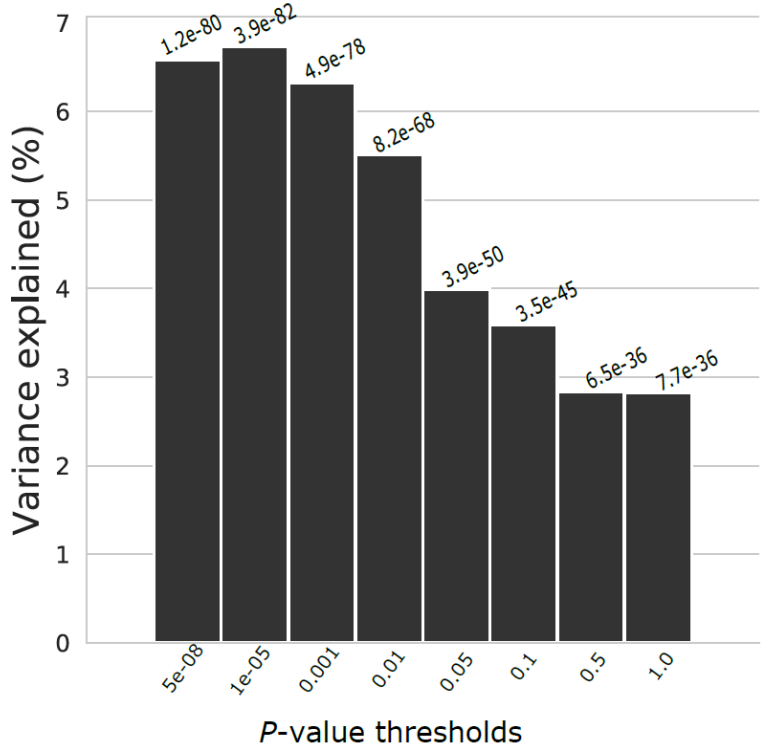
PGC-MDD2:
N=163k
max variance explained =0.46%,
p= 5.01e-08

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

Applications

- Quantify variance explained & explore architecture
- 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)

Additional considerations in mental health

- When people present it is not usually a question of if someone has a future risk
  - Individuals or their families typically seek help
- Individuals often present with symptoms that might fit more than one diagnostic criterion
  - So the question is usually one of differential diagnosis

How is mental health different?

- Comorbidity (co-occurrence) is the rule than the exception
- Presentations and diagnoses are expected to change over time
  - This is ≠ misdiagnosis
  - PRS is static across the lifespan

Further complications…

- Diagnoses are used for more than just treatment planning
  - Medico-legal contexts
    - Criminal, Civil and Family proceedings
    - Compensation
  - Access to support
    - Financial, Educational, Housing, Social

Further complications…

- Because of this, diagnostic processes are ideally
  - Static
  - Reproducible
  - Easy to implement
  - Useable in low resource settings
  - Measurement invariant

- 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)


Navajo Peak
ARK
NAV
GL5
Arikaree Peak
Como Creek
GU Mountain Research Station
SLP
Sample site
Stream
Lake
Forest Cover
Kilometers
Contour Interval = 100m

Questions?