Polygenic Risk Score

in Practice

Delnaz Roshandel, MD, PhD

The Centre for Applied Genomics (TCAG)

The Hospital for Sick Children, Toronto

Delnaz.Roshandel@sickkids.ca

Overview

1. **PRS calculation process:** base & target data

2. **Important considerations to choose the base data:** population structure & sample overlap

3. **Methods to calculate PRS:** clumping-thresholding vs. shrinkage methods

4. **Optimizing the parameters:** SNP filtering, clumping parameters & p-value thresholds

5. **Base data checks:** direction of effect & genome build

6. **Target data checks:** QC, finding SNPs & mismatching alleles

7. **Other considerations:** PRS in multiple studies & control group

8. **PRS calculation**

PRS Calculation Process

- **Base data:** GWAS/meta-GWAS summary statistics (e.g. effect sizes or p-values) for the trait of interest.
  - Has PRS been calculated & evaluated?
  - If not, target data can be used for both evaluation and analysis (e.g. subsampling, leave one out method).

- **Target data:** Genotype data (e.g. in PLINK format) of individuals in whom PRS is calculated.

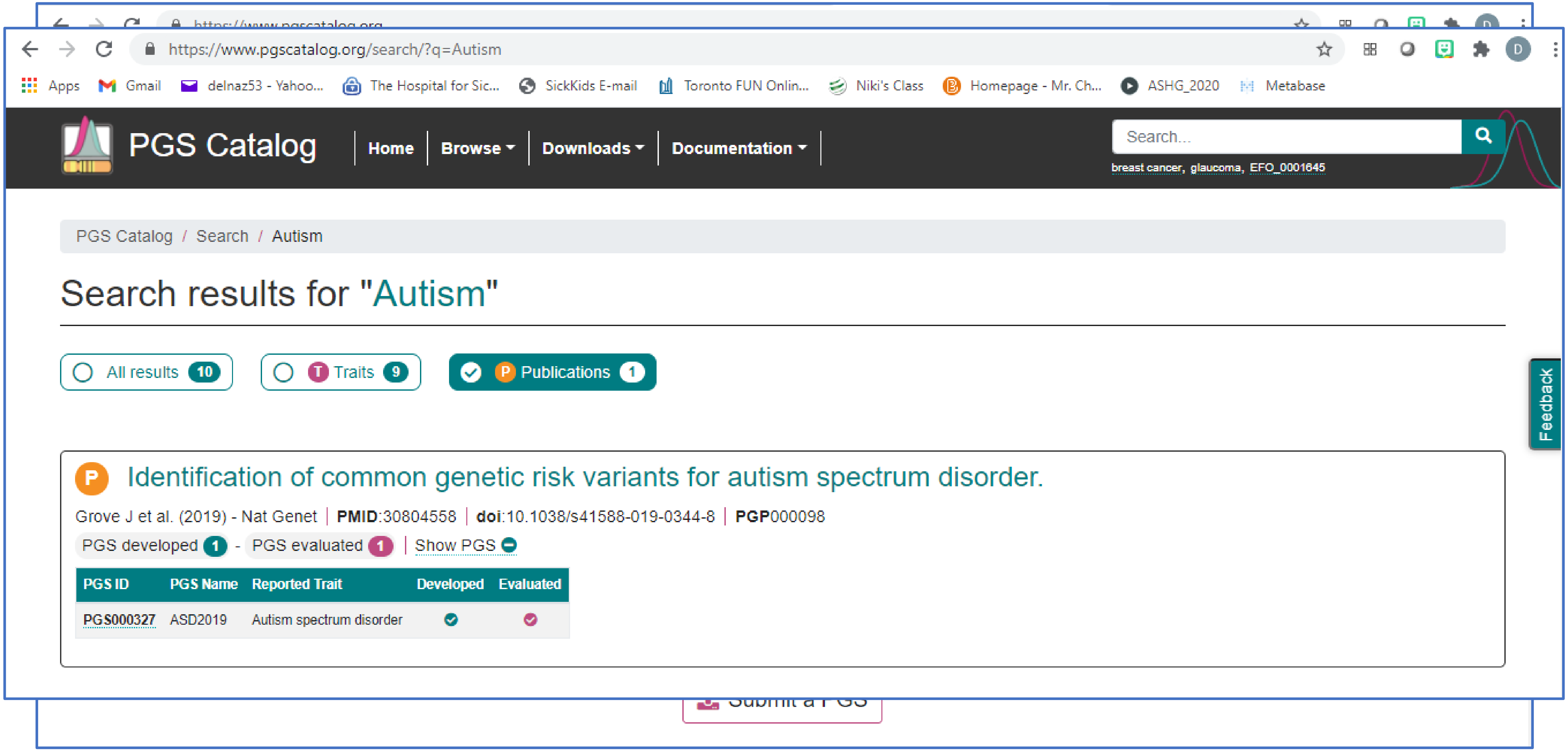
- **PRS:**
  - An estimate of an individual's genetic liability to a trait
  - Calculated by computing the sum of risk alleles that an individual has, weighted by the risk allele effect size estimates derived from GWAS summary stats
  - Often only common biallelic SNPs are included.
  - Rare or other types of variations can be included.

Find the Largest GWAS/Meta-GWAS of the Trait of Interest

GWAS Catalogue
https://www.ebi.ac.uk/gwas/

The Polygenic Score (PGS) Catalog
https://www.pgscatalog.org/

PGS Catalogue



PGS Catalog / Search / Autism

Search results for "Autism"

○ All results 10 ○ T Traits 9 ✓ P Publications 1

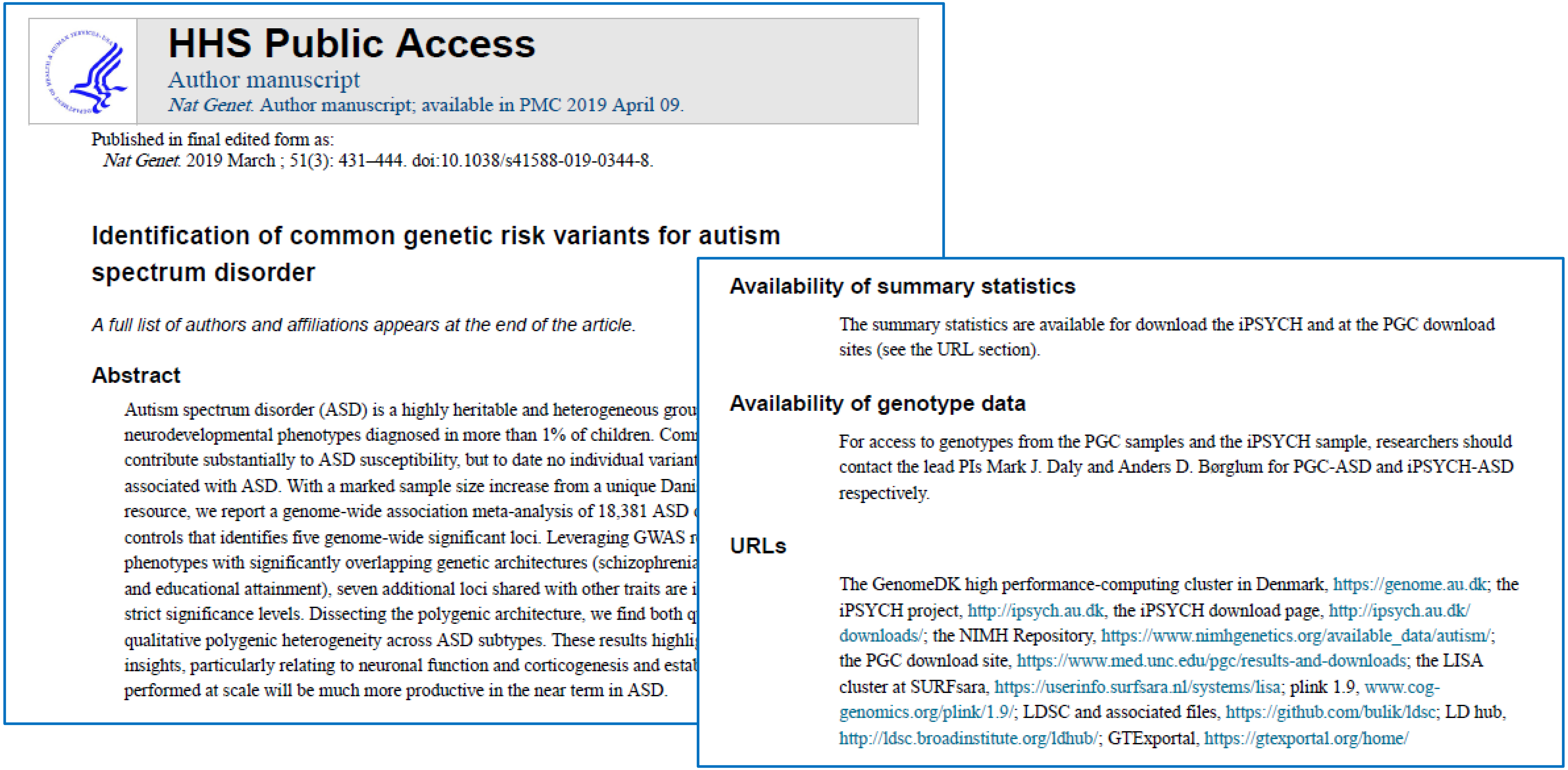
P Identification of common genetic risk variants for autism spectrum disorder.

Grove J et al. (2019) - Nat Genet | **PMID**:30804558 | **doi**:10.1038/s41588-019-0344-8 | **PGP**000098

PGS developed 1 - PGS evaluated 1 | Show PGS

PGS ID	PGS Name	Reported Trait	Developed	Evaluated
PGS000327	ASD2019	Autism spectrum disorder	✓	✓

Largest Meta-GWAS, Autism Spectrum Disorder (ASD)



Published in final edited form as:
Nat Genet. 2019 March ; 51(3): 431–444. doi:10.1038/s41588-019-0344-8.

Identification of common genetic risk variants for autism spectrum disorder

A full list of authors and affiliations appears at the end of the article.

Abstract

Autism spectrum disorder (ASD) is a highly heritable and heterogeneous grou
neurodevelopmental phenotypes diagnosed in more than 1% of children. Com
contribute substantially to ASD susceptibility, but to date no individual variant
associated with ASD. With a marked sample size increase from a unique Dani
resource, we report a genome-wide association meta-analysis of 18,381 ASD c
controls that identifies five genome-wide significant loci. Leveraging GWAS r
phenotypes with significantly overlapping genetic architectures (schizophrenia
and educational attainment), seven additional loci shared with other traits are i
strict significance levels. Dissecting the polygenic architecture, we find both q
qualitative polygenic heterogeneity across ASD subtypes. These results highli
insights, particularly relating to neuronal function and corticogenesis and estab
performed at scale will be much more productive in the near term in ASD.

Availability of summary statistics

The summary statistics are available for download the iPSYCH and at the PGC download sites (see the URL section).

Availability of genotype data

For access to genotypes from the PGC samples and the iPSYCH sample, researchers should contact the lead PIs Mark J. Daly and Anders D. Børglum for PGC-ASD and iPSYCH-ASD respectively.

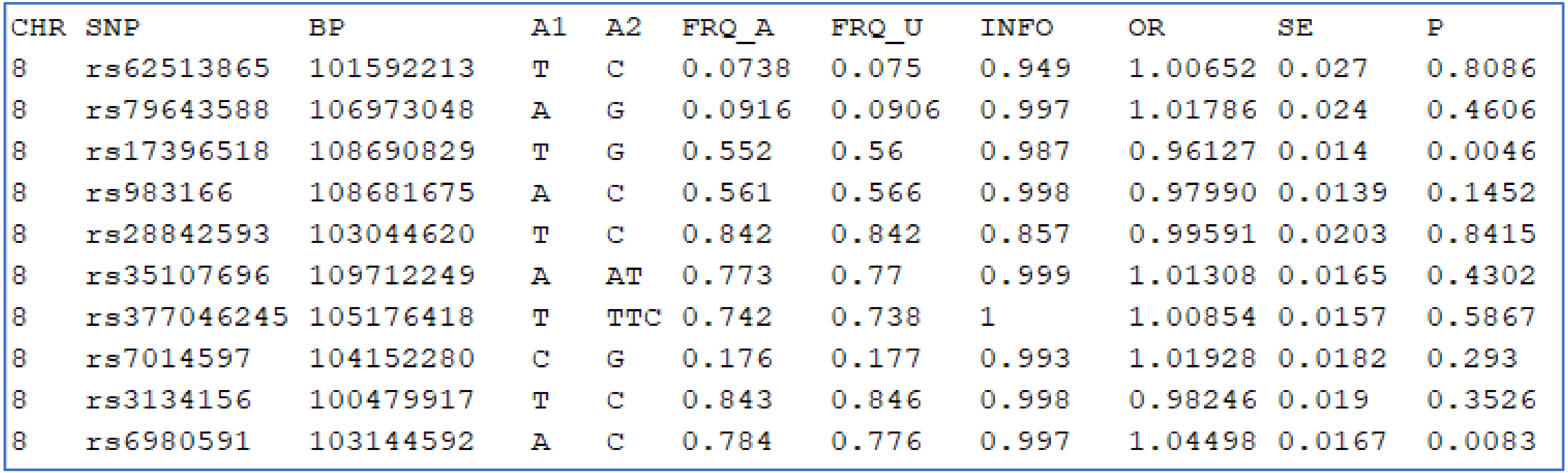
URLs

The GenomeDK high performance-computing cluster in Denmark, https://genome.au.dk; the iPSYCH project, http://ipsych.au.dk, the iPSYCH download page, http://ipsych.au.dk/downloads/; the NIMH Repository, https://www.nimhgenetics.org/available_data/autism/; the PGC download site, https://www.med.unc.edu/pgc/results-and-downloads; the LISA cluster at SURFsara, https://userinfo.surfsara.nl/systems/lisa; plink 1.9, www.cog-genomics.org/plink/1.9/; LDSC and associated files, https://github.com/bulik/ldsc; LD hub, http://ldsc.broadinstitute.org/ldhub/; GTExportal, https://gtexportal.org/home/

Base Data, GWAS Summary Statistics

Which build?

Log transformed



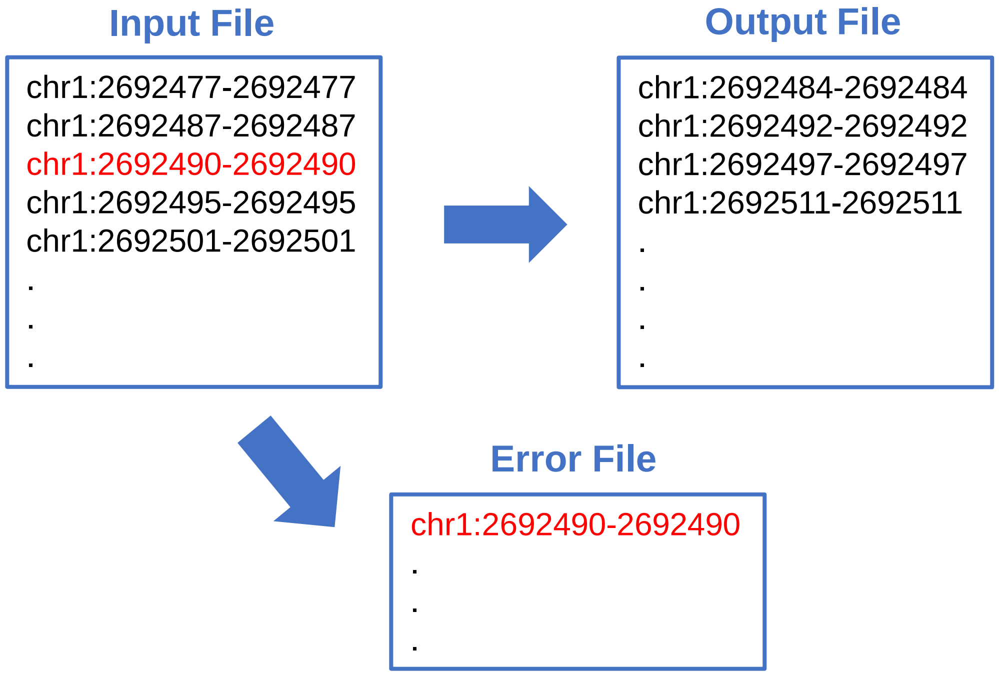
CHR	SNP	BP	A1	A2	FRQ_A	FRQ_U	INFO	OR	SE	P
8	rs62513865	101592213	T	C	0.0738	0.075	0.949	1.00652	0.027	0.8086
8	rs79643588	106973048	A	G	0.0916	0.0906	0.997	1.01786	0.024	0.4606
8	rs17396518	108690829	T	G	0.552	0.56	0.987	0.96127	0.014	0.0046
8	rs983166	108681675	A	C	0.561	0.566	0.998	0.97990	0.0139	0.1452
8	rs28842593	103044620	T	C	0.842	0.842	0.857	0.99591	0.0203	0.8415
8	rs35107696	109712249	A	AT	0.773	0.77	0.999	1.01308	0.0165	0.4302
8	rs377046245	105176418	T	TTC	0.742	0.738	1	1.00854	0.0157	0.5867
8	rs7014597	104152280	C	G	0.176	0.177	0.993	1.01928	0.0182	0.293
8	rs3134156	100479917	T	C	0.843	0.846	0.998	0.98246	0.019	0.3526
8	rs6980591	103144592	A	C	0.784	0.776	0.997	1.04498	0.0167	0.0083

Base Data Checks

- Which allele is the effect allele? → To make sure that the effect of the PRS in the target data is in the correct direction

- Genome build → HG19 or HG38

  - LiftOver Tool: https://genome.ucsc.edu/cgi-bin/hgLiftOver



Input File

chr1:2692477-2692477
chr1:2692487-2692487
chr1:2692490-2692490
chr1:2692495-2692495
chr1:2692501-2692501
.
.
.

Output File

chr1:2692484-2692484
chr1:2692492-2692492
chr1:2692497-2692497
chr1:2692511-2692511
.
.
.
.

Error File

chr1:2692490-2692490
.
.
.

Population Structure

Match base and target data for ethnicity

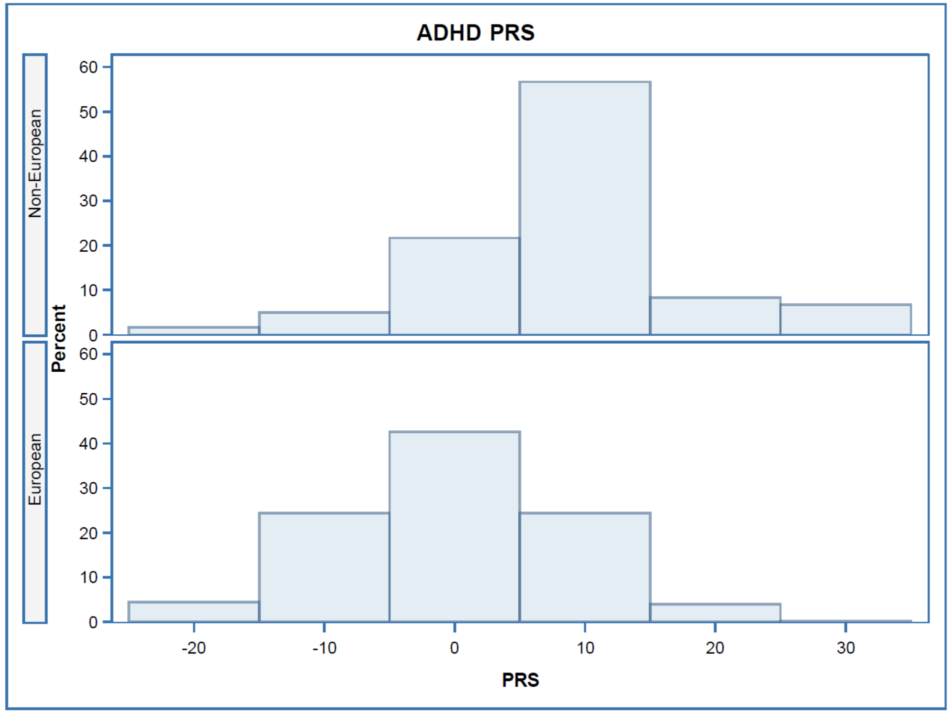
ADHD PRS in BIOJUME

- ➢ Largest meta-GWAS of ADHD: Demontis et al 2019
  - 19,099 cases
  - 34,194 controls
  - All European

- ➢ BIOJUME → Juvenile Myoclonic Epilepsy
  - 627 Europeans
  - 60 Non-Europeans

ADHD: Attention deficit hyperactivity disorder
BIOJUME: Biology of Juvenile Myoclonic Epilepsy

ADHD PRS		
	Mean	SD
Europeans	0.08	8.58
Non-Europeans	7.82	8.93



Sample Overlap

The target sample or part of it within the meta-GWAS.

↓

Over-fitting & Inflation for association of PRS with the trait in target sample

ASD PRS in MSSNG/SSC

- Base data → Grove et al, 2019

  - iPSYCH → Danish population-based case-cohort → 23 GWAS regarding each batch → Meta-GWAS

  - PGC → 5 family-based trio studies → Meta-GWAS with iPSYCH
    1. ACE (Geschwind Autism Center of Excellence)
    2. AGP (Autism Genome Project) → Overlap with MSSNG
    3. AGRE (Autism Genetic Resource Exchange)
    4. MONBOS (NIMH Repository, Montreal/Boston Collection)
    5. SSC (Simons Simplex Collection)

- Target data: MSSNG & SSC → Large studies of ASD

ASD PRS in MSSNG

ASD PRS Using **iPSYCH + PGC** Meta-GWAS Summary Stats				
	Not in PGC (N = 1,892)		**In PGC (N = 372)**	
	Mean	**SD**	**Mean**	**SD**
Probands	0.37	5.22	5.96	4.76
ASD PRS Using **iPSYCH Only** Meta-GWAS Summary Stats				
Probands	0.15	5.49	0.72	5.29

Has PRS been calculated in the base data & the parameters have been optimized?

ASD PRS, Grove et al

- 18,381 cases & 27,969 controls
  - iPHYCH → 13,076 cases & 22,664 controls
  - PGC → N = 5,305
- Divided the iPSYCH sample in 5 sub-samples of roughly equal size
- Ran 5 GWAS leaving out one sub-sample in turn
- Meta analyzed each of these GWASs with the PGC results
- Produced a set of PRS for each of the five sub-samples trained on their complement
- Evaluated the predictive power of PRS in each group & on the whole sample combined → using Nagelkerke's R^2

PRS Calculation Methods

- GWAS
  - Association tests → Performed one SNP at a time
  - SNPs correlated due to LD
  - Independent genetic effects required for PRS calculation

- Methods available

1. Clumping + thresholding (C+T) → Classic method
   - Clumping → Pruning with prioritizing SNPs with the smallest p-value
   - Thresholding → Keeping SNPs with a p-value less than a certain value

2. Shrinkage techniques → Including all SNPs accounting for the LD between them

ASD PRS, Grove et al

- ➢Classic method → C+T

- ➢SNP filtering:

  - Minor allele frequency < 0.05 → Arbitrary 0.01 or 0.05 considering sample size

  - Low imputation quality → INFO < 0.9 → Arbitrary INFO < 0.8 or $R^2 < 0.5$

  - Complementary SNPs (A > T or C > G)
    - Target and base data → Different genotyping platforms/imputation
    - The used strand (+/-) not clear

  - Non-autosomal SNPs → Sex chromosomes should be modeled separately

  - HLA region
    - High LD
    - Highly variable
    - Major locus for autoimmune diseases

Type 1 Diabetes

- ➢ DR3/DR4 Haplotype → Major genetic risk factor

- ➢ DR3/DR4 haplotype is perfectly tagged by:

  - rs2187668 (chr6:32,605,884; C>T)
  - rs7454108 (chr6:32,681,483; T>C)

Haplotype Genotype	rs2187668	rs7454108
DR3/DR3	TT	TT
DR3/DR4	TC	CT
DR3/X	TC	TT
DR4/DR4	CC	CC
X/DR4	CC	CT
X/X	CC	TT

SNP	Chr	BP (HG19)	Gene	Effect Allele	OR	Weight
DR3/DR3	6				21.12	3.05
DR3/DR4	6				48.18	3.87
DR3/X	6				4.53	1.51
DR4/DR4	6				21.98	3.09
X/DR4	6				7.03	1.95
X/X	6				1	0
rs1264813	6	29,939,900	HLA_A_24	T	1.54	0.43
rs2395029	6	31,431,780	HLA_B_5701	T	2.5	0.92
rs3129889	6	32,413,545	HLA_DRB1_15	A	14.88	2.7
rs2476601	1	114,377,568	PTPN22	A	1.96	0.67
rs689	11	2,182,224	INS	T	1.75	0.56
rs12722495	10	6,097,283	IL2RA	T	1.58	0.46
rs2292239	12	56,482,180	ERBB3	T	1.35	0.3
rs10509540	10	90,023,033	C10orf59	T	1.33	0.29
rs4948088	7	51,027,194	COBL	C	1.3	0.26
rs7202877	16	75,247,245		G	1.28	0.25
rs12708716	16	11,179,873	CLEC16A	A	1.23	0.21
rs3087243	2	204,738,919	CTLA4	G	1.22	0.2
rs1893217	18	12,809,340	PTPN2	G	1.2	0.18
rs11594656	10	6,122,009	IL2RA	T	1.19	0.17
rs3024505	1	206,939,904	IL10	G	1.19	0.17
rs9388489	6	126,698,719	C6orf173	G	1.17	0.16
rs1465788	14	69,263,599		C	1.16	0.15
rs1990760	2	163,124,051	IFIH1	T	1.16	0.15
rs3825932	15	79,235,446	CTSH	C	1.16	0.15
rs425105	19	47,208,481		T	1.16	0.15
rs763361	18	67,531,642	CD226	T	1.16	0.15
rs4788084	16	28,539,848	IL27	C	1.16	0.15
rs17574546	15	38,902,476		C	1.14	0.13
rs11755527	6	90,958,231	BACH2	G	1.13	0.12
rs3788013	21	43,841,328	UBASH3A	A	1.13	0.12
rs2069762	4	123,377,980	IL2	A	1.12	0.11
rs2281808	20	1,610,551		C	1.11	0.1
rs5753037	22	30,581,722		T	1.1	0.1

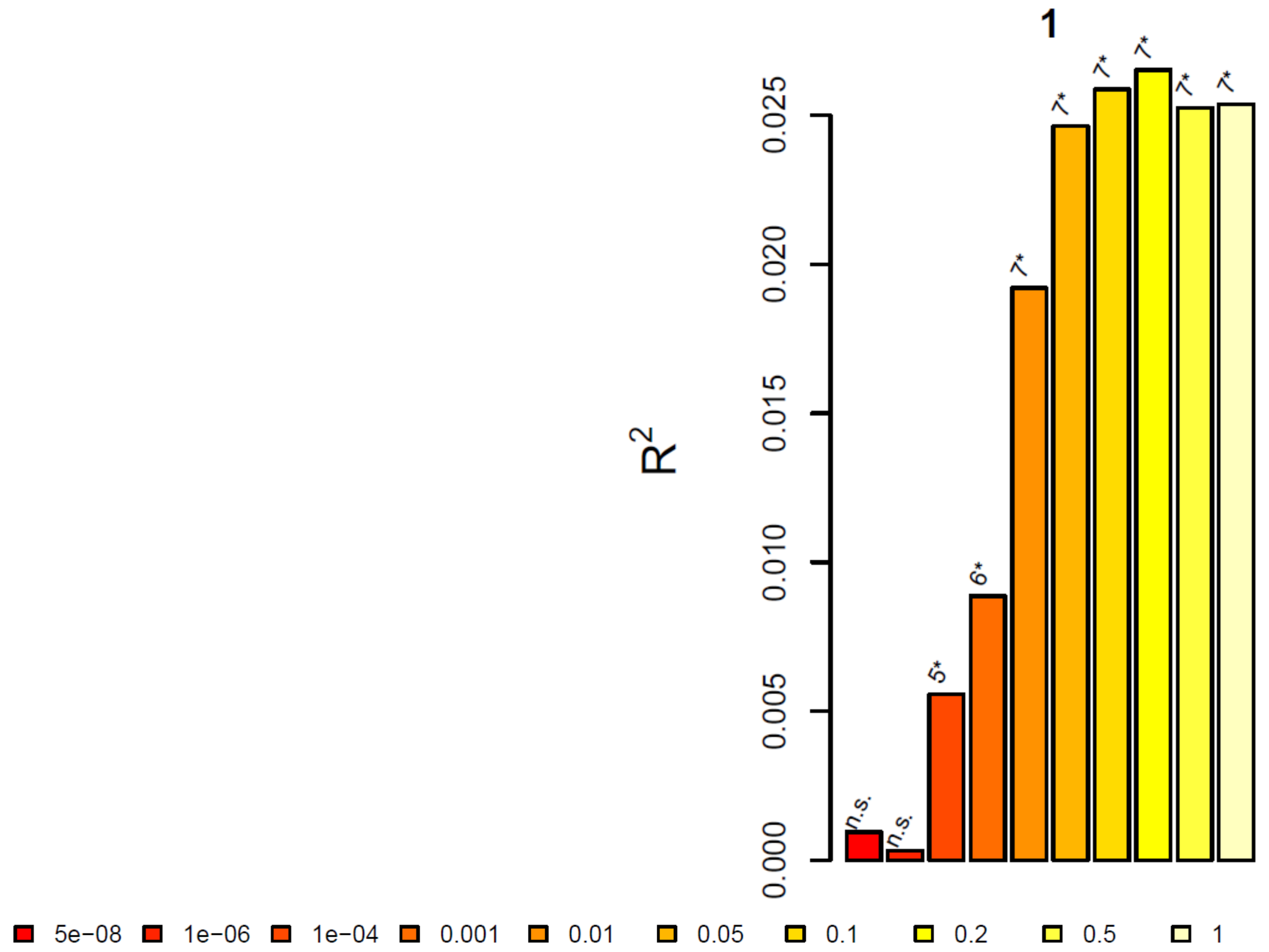
Weight = log(OR), X: Non-DR3 non-DR4

ASD PRS, Grove et al

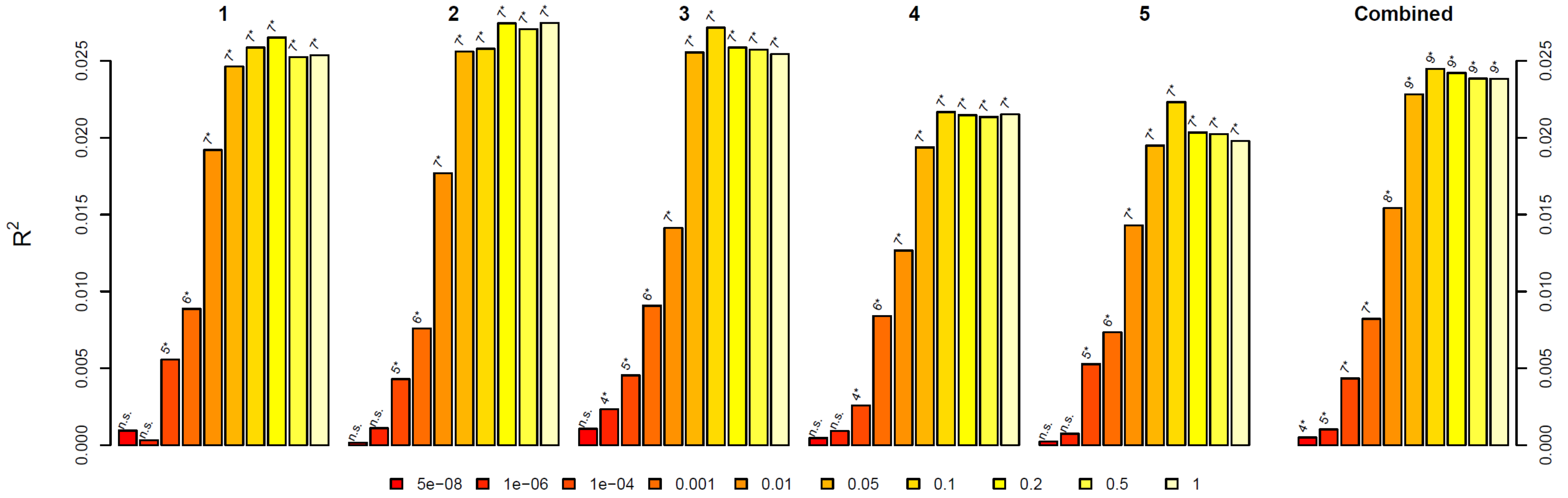
- Clumping
  - $r^2 < 0.1$ → Arbitrary
  - Radius = 500 kb → Arbitrary

- P-value thresholds:
  1. P ≤ 5E-8
  2. P ≤ 1E-6
  3. P ≤ 1E-4
  4. P ≤ 0.001
  5. P ≤ 0.01
  6. P ≤ 0.05
  7. **P ≤ 0.1**
  8. P ≤ 0.2
  9. P ≤ 0.5
  10. P ≤ 1

Nagelkerke R^2 of PRS in the first sub-sample



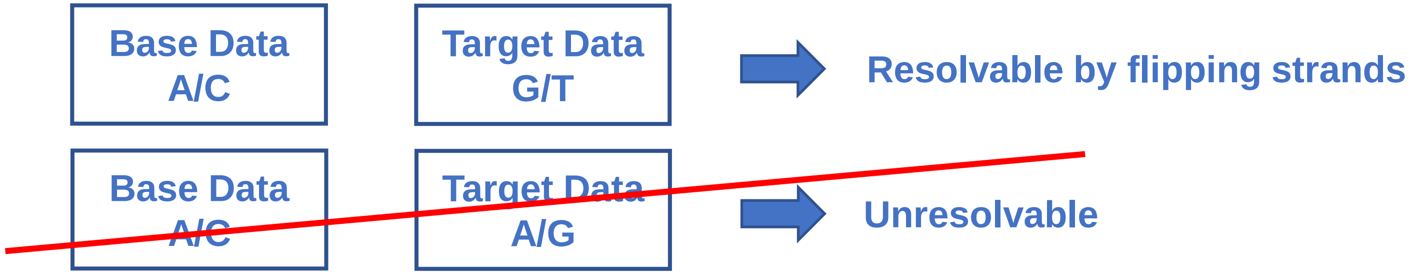
ASD PRS, Grove et al



Supplementary Figure 89: Nagelkerke R^2 of PRS trained internally on leave-one-group-out and the PGC ASD shown here when estimated on each of the five groups left out when training as well as on the combined sample (cases/controls in groups 1: 2 624/3 694, 2: 2 622/5 432, 3: 2 611/4 666, 4: 2 583/4 360, 5: 2 636/4 512, and in total 13 076/22 664). Colouring is as shown in the legend signifying the 10 different p-value cut-off in the training set.

Target Data Checks

- Standard GWAS/sequencing QC
  - GWAS → high imputation quality (e.g. INFO > 0.8, R^2> 0.5)
  - Sequencing → Filter flag = PASS
- Find the SNPs from base data according to their position and alleles
- Exclude SNPs with mismatching alleles in base and target data not due to strand-flipping



Base Data A/C | Target Data G/T → Resolvable by flipping strands

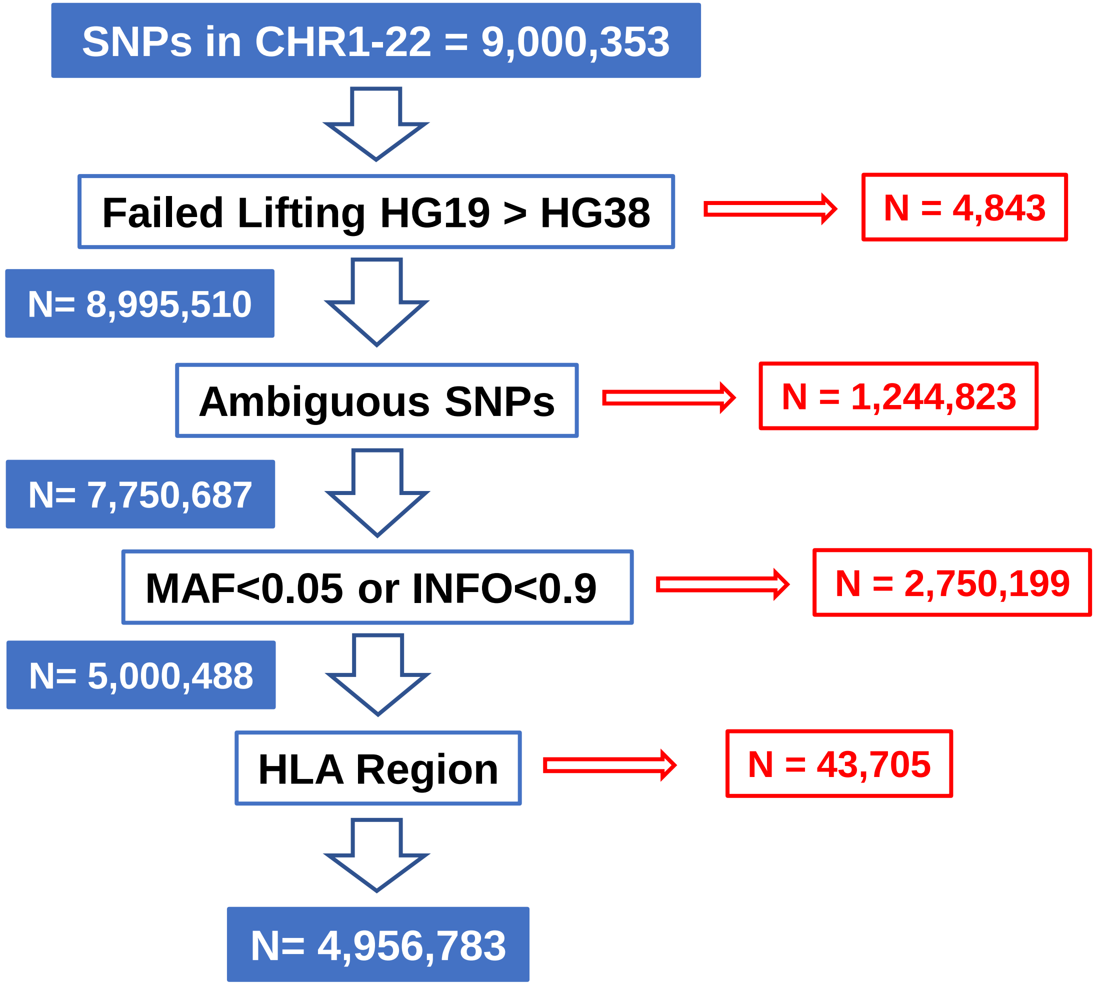
Base Data A/C | Target Data A/G → Unresolvable

Target Data Checks

- Calculating PRS in multiple studies → Need PRS to be comparable
    - Merge data before clumping
    - Keep only common SNPs

- No independent control sample
    - Find an independent data from the same ethnicity as representative of normal population → e.g. 1000 Genomes
    - Merge target data with the independent control sample before clumping
    - Keep only common SNPs

ASD PRS, Base Data, SNP Filtering



SNPs in CHR1-22 = 9,000,353

Failed Lifting HG19 > HG38

N = 4,843

N= 8,995,510

Ambiguous SNPs

N = 1,244,823

N= 7,750,687

MAF<0.05 or INFO<0.9

N = 2,750,199

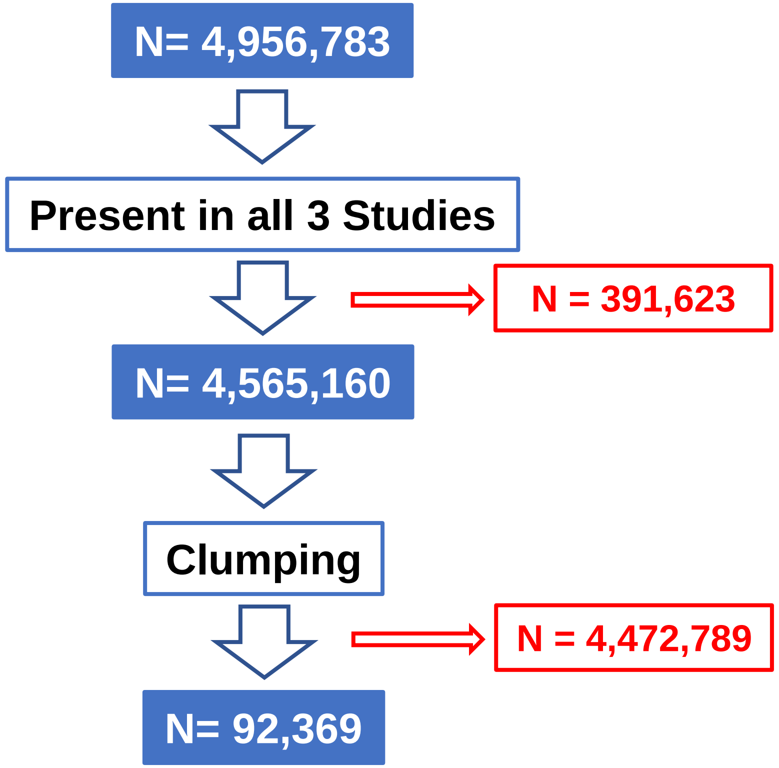
N= 5,000,488

HLA Region

N = 43,705

N= 4,956,783

ASD PRS, MSSNG/SSC/1000 Genomes, Pruning & Thresholding



N= 4,956,783

Present in all 3 Studies

N = 391,623

N= 4,565,160

Clumping

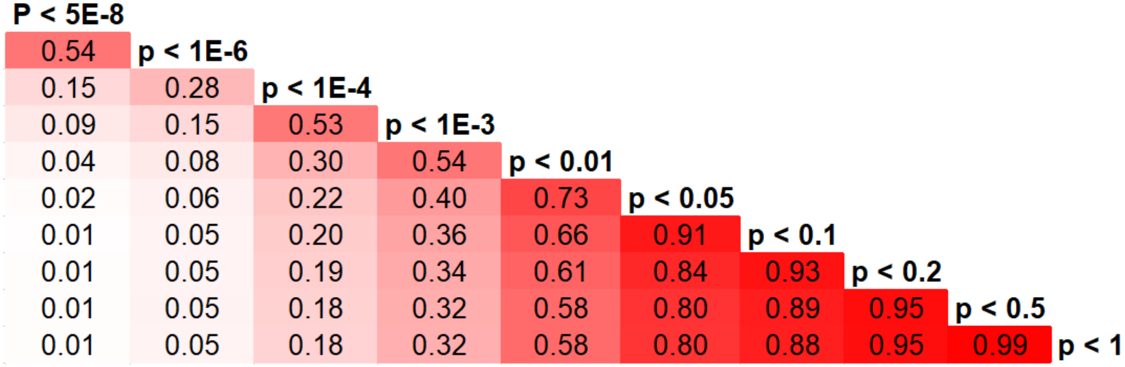
N = 4,472,789

N= 92,369

P-value Threshold	N of SNPs
5E-8	2
1E-6	9
1E-4	175
1E-3	875
0.01	4,997
0.05	15,960
0.1	**25,837**
0.2	40,968
0.5	70,743
1	92,369

ASD PRS, MSSNG/SSC/1000 Genomes

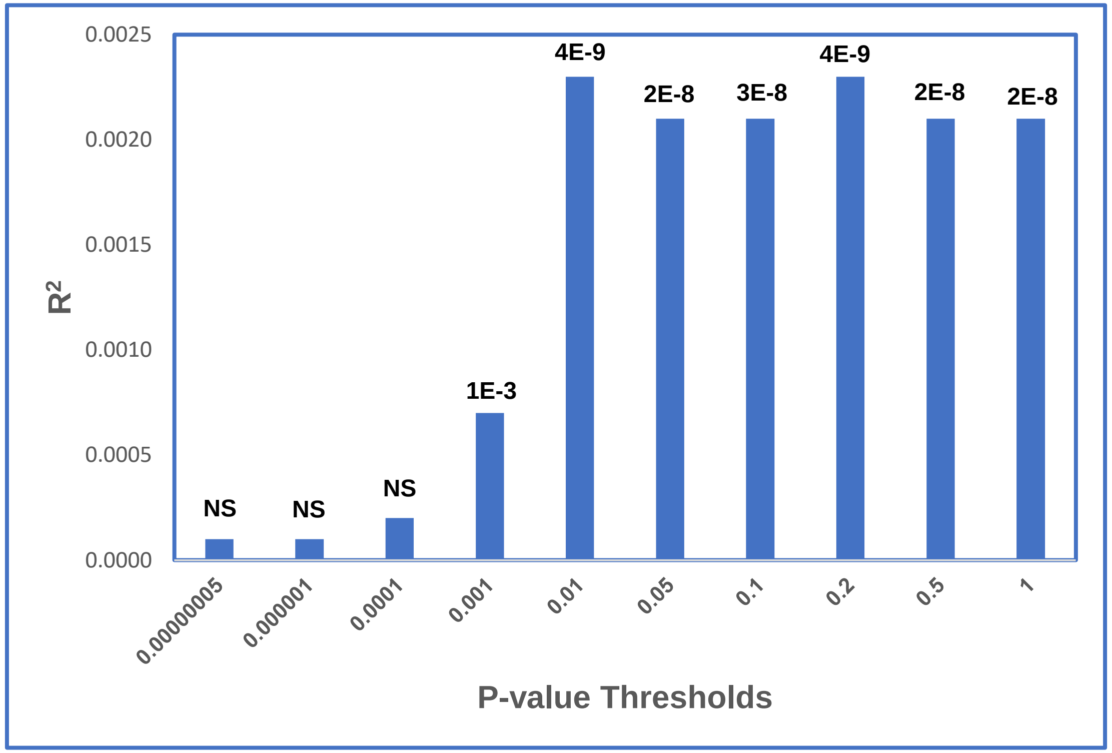
Correlation between PRSs with different p-value thresholds



P < 5E-8									
0.54	p < 1E-6								
0.15	0.28	p < 1E-4							
0.09	0.15	0.53	p < 1E-3						
0.04	0.08	0.30	0.54	p < 0.01					
0.02	0.06	0.22	0.40	0.73	p < 0.05				
0.01	0.05	0.20	0.36	0.66	0.91	p < 0.1			
0.01	0.05	0.19	0.34	0.61	0.84	0.93	p < 0.2		
0.01	0.05	0.18	0.32	0.58	0.80	0.89	0.95	p < 0.5	
0.01	0.05	0.18	0.32	0.58	0.80	0.88	0.95	0.99	p < 1

Association of PRS with ASD

5,010 cases & 9,839 controls

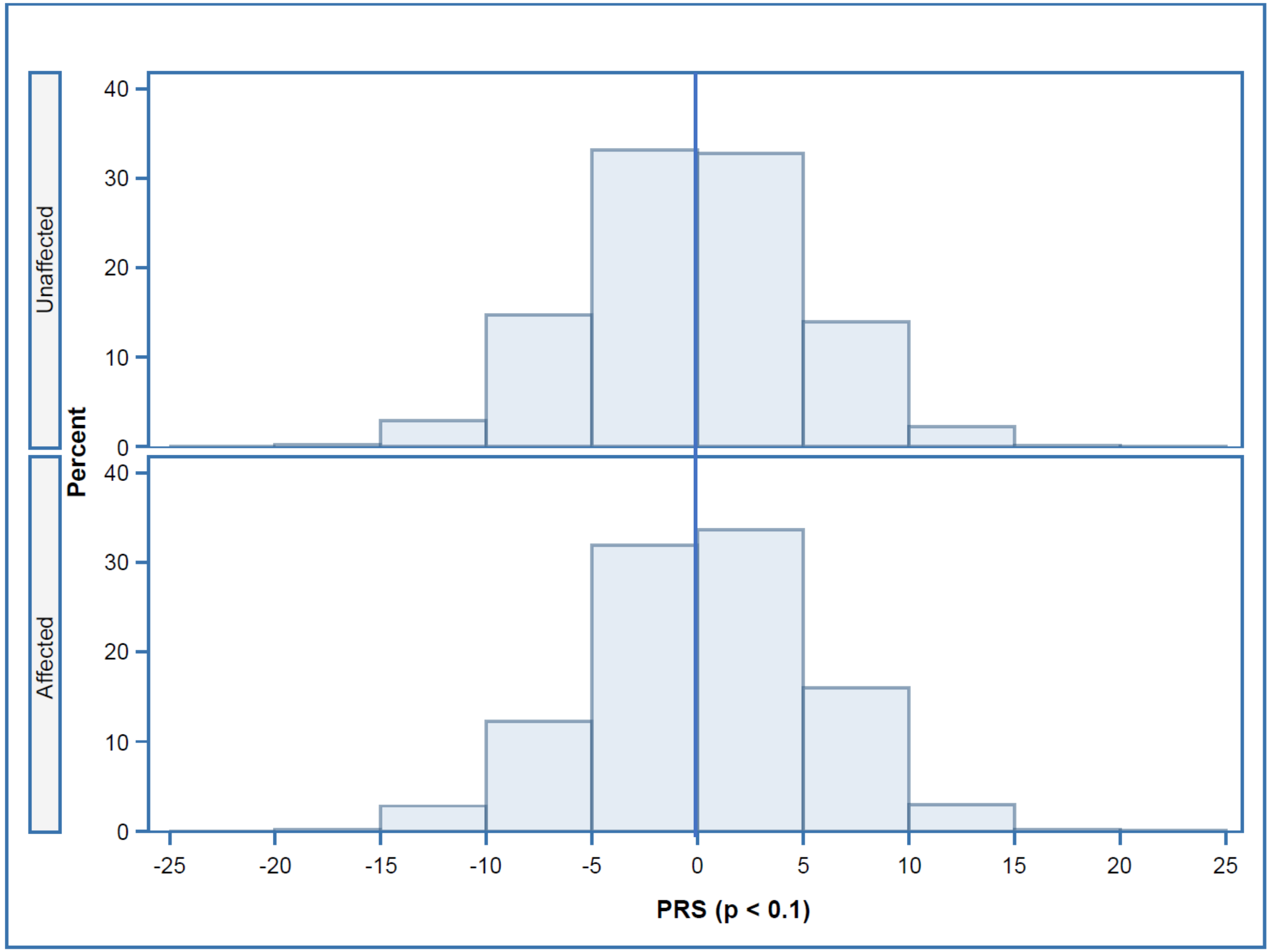


X axis: p-value thresholds

Y axis: psuedo-R^2

The values on top of the bars show p-values.

ASD PRS, MSSNG/SSC/1000 Genomes



PRS is centred to mean.

ASD PRS, MSSNG/SSC/1000 Genomes

	MSSNG			SSC			1000 Genomes		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Affected Siblings	639	0.11	5.25	10	-1.95	4.01	-	-	-
Child	330	0.31	5.43	-	-	-	-	-	-
Father	1948	-0.08	5.15	1938	-0.32	5.25	-	-	-
Mother	1899	-0.01	5.16	1925	-0.10	5.20	-	-	-
Proband	2264	0.36	5.40	1869	0.32	5.05	-	-	-
Unaffected Sibling	77	0.06	4.86	1519	-0.25	5.19	-	-	-
None	-	-	-	-	-	-	516	-0.40	5.28

P = 3.12E-3

P = 6.29E-3

Summary

- Two data are required:

  1. Base → Summary stats from the largest meta-GWAS available
  2. Target → Genotype data of individuals in whom PRS is calculated

- Base and target sample should be independent.

- Base and target sample should be from the same ethnic group.

- Base data checks:
  - Effect allele
  - Genome build

- SNP filtering → MAF, imputation quality, complementary SNPs, Chr X & HLA region

Summary

- ➢ PRS calculation methods → C + T

    - Clumping parameters → r^2 & radius
    - Different p-value thresholds

- ➢ Evaluate PRS → PRS with best predictive power → Variance explained by PRS (R^2)

- ➢ Target data checks:
    - Standard genotyping/sequencing QC
    - Finding SNPs based on position & alleles
    - Mismatching alleles

- ➢ Merge data from multiple target datasets keeping common SNPs before clumping → Comparable PRS

- ➢ Control dataset from the same ethnic group → 1000 Genomes