Polygenic risk scores

Alicia Martin, PhD

Stanley Center Global Plenary 2018

September 12, 2018

armartin@broadinstitute.org

@genetisaur

Outline

- What are polygenic risk scores?
- How to compute them
- Methods, interpretations, and uses
- Ancestry, health disparities, and ongoing/future directions

The New York Times

Opinion

GRAY MATTER

How Genetics Is Changing Our Understanding of 'Race'

By David Reich

March 23, 2018

900

The New York Times

Opinion

LETTERS

Race, Genetics and a Controversy

April 2, 2018

SCIENCE

An Enormous Study of the Genes Related to Staying in School

Researchers have found 1,271 gene variants associated with years of formal education. That's important, but not for the obvious reasons.

ED YONG JUL 23, 2018

The New York Times

Opinion

Why Progressives Should Embrace the Genetics of Education

By Kathryn Paige Harden

Dr. Harden is a psychologist who studies how genetic factors shape adolescent development.

July 24, 2018

365

Why We Shouldn't Embrace the Genetics of Education

It's a trap!

By John Warner // July 26, 2018

43 COMMENTS

WANT

COLLEGE PAC

MIT Technology Review

Forecasts of genetic fate just got a lot more accurate

by Antonio Regalado February 21, 2018

The New York Times

Clues to Your Health Are Hidden at 6.6 Million Spots in Your DNA

By Gina Kolata

Aug. 13, 2018

With a sophisticated new algorithm, scientists have found a way to forecast an individual's risks for five deadly diseases.

Vox

How scientists are learning to predict your future with your genes

But what are the limits?

By Brian Resnick | @B_resnick | brian@vox.com | Updated Aug 25, 2018, 9:35am EDT

Insight & Intelligence

August 22, 2018

Why Do Polygenic Risk Scores Get So Much Hype?

GWAS for Common Disease Variants Gains Prominence

Julianna LeMieux, Ph.D.

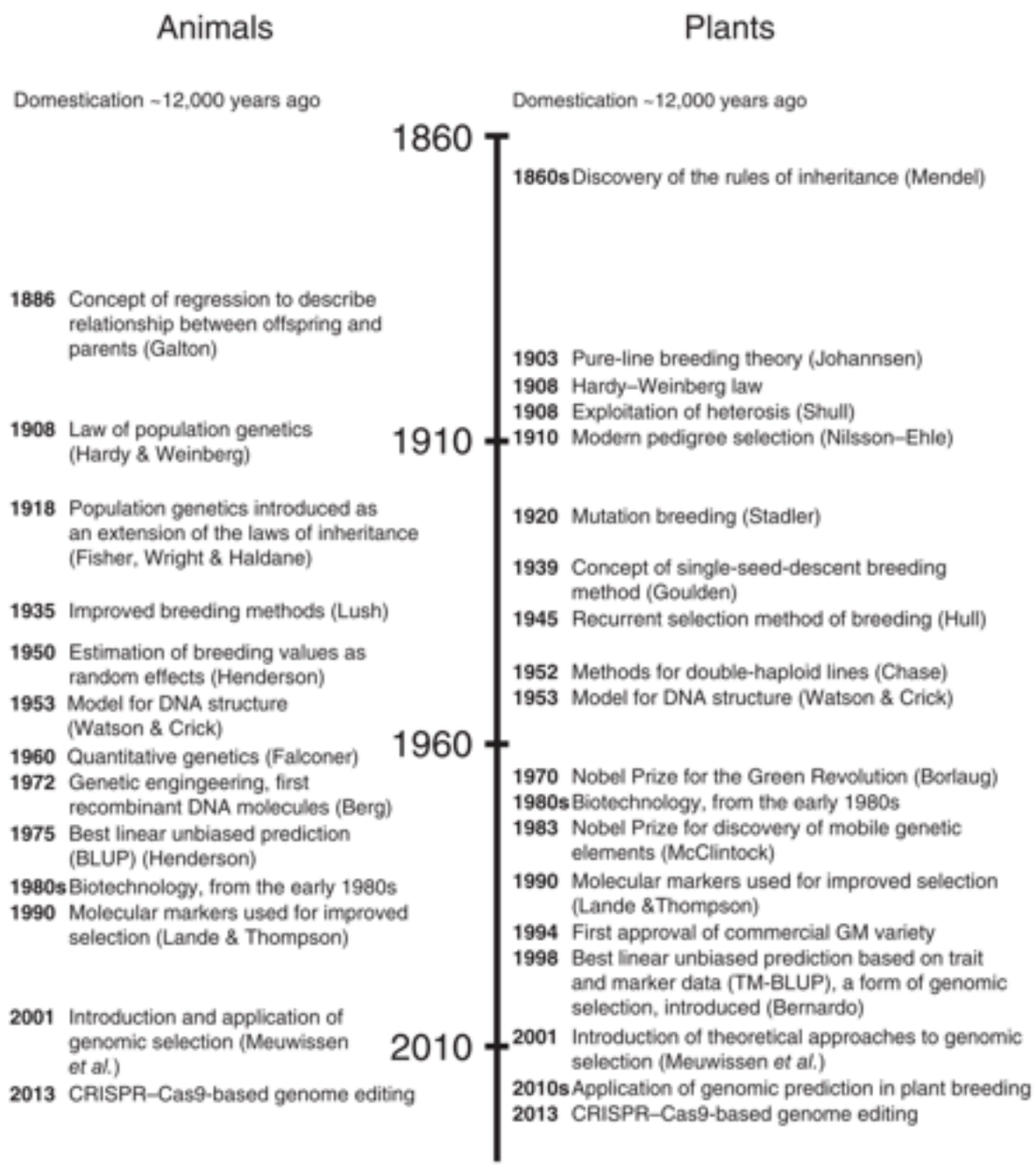
The rise of the polygenic risk score

"We propose the time has come to incorporate genetic risk scores into clinical practice"

- Previous criticisms: limited sample size
- Cheap test for insights into many diseases
- Integrate with other clinical factors for therapeutic decision-making

Knowles JW, Ashley EA (2018) Cardiovascular disease: The rise of the genetic risk score. PLoS Med 15(3): e1002546.

A long shared history between PRS and breeding values



Hickey, J.M., et al. (2017). Nat. Genet. 49, 1297–1303.

nature Vol 460|6 August 2009|doi:10.1038/nature08185

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

- The dark days of low-powered GWAS
- PRS show value of GWAS even in the absence of genome-wide significant loci

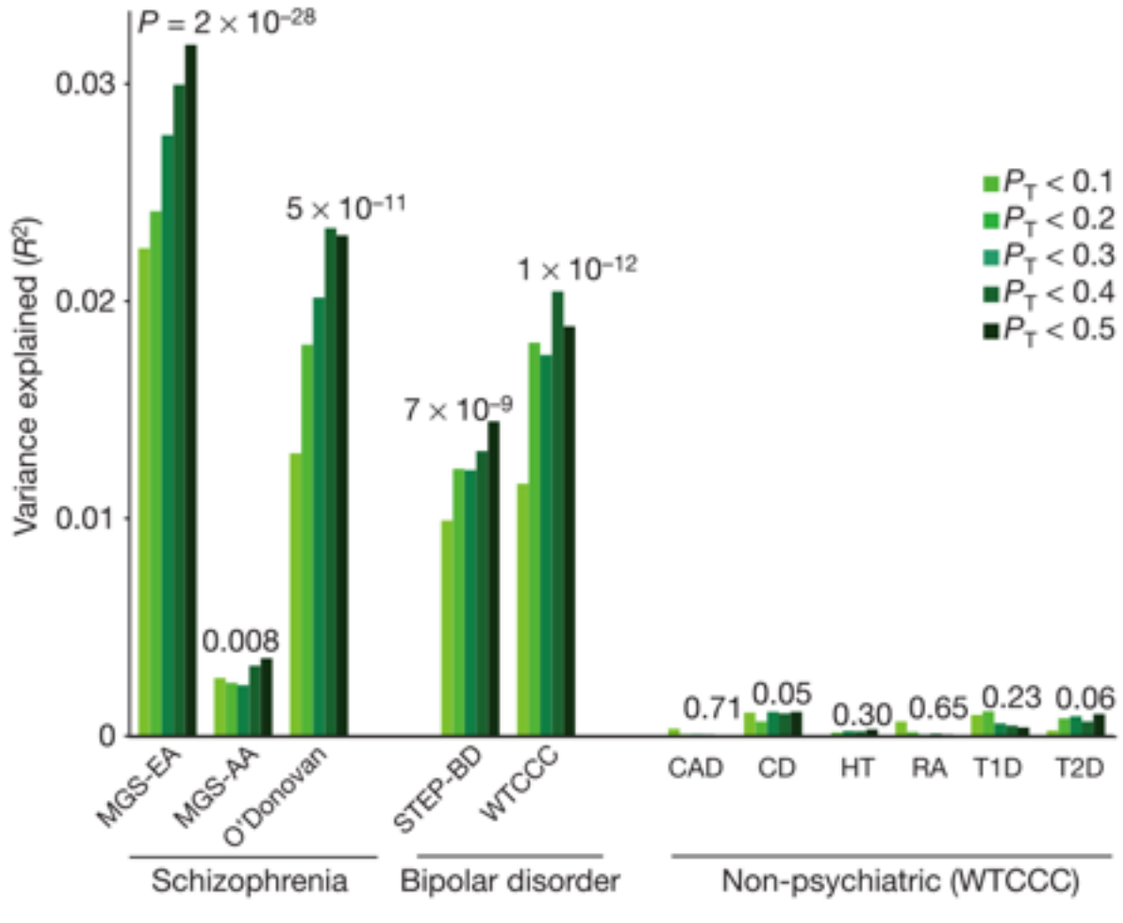
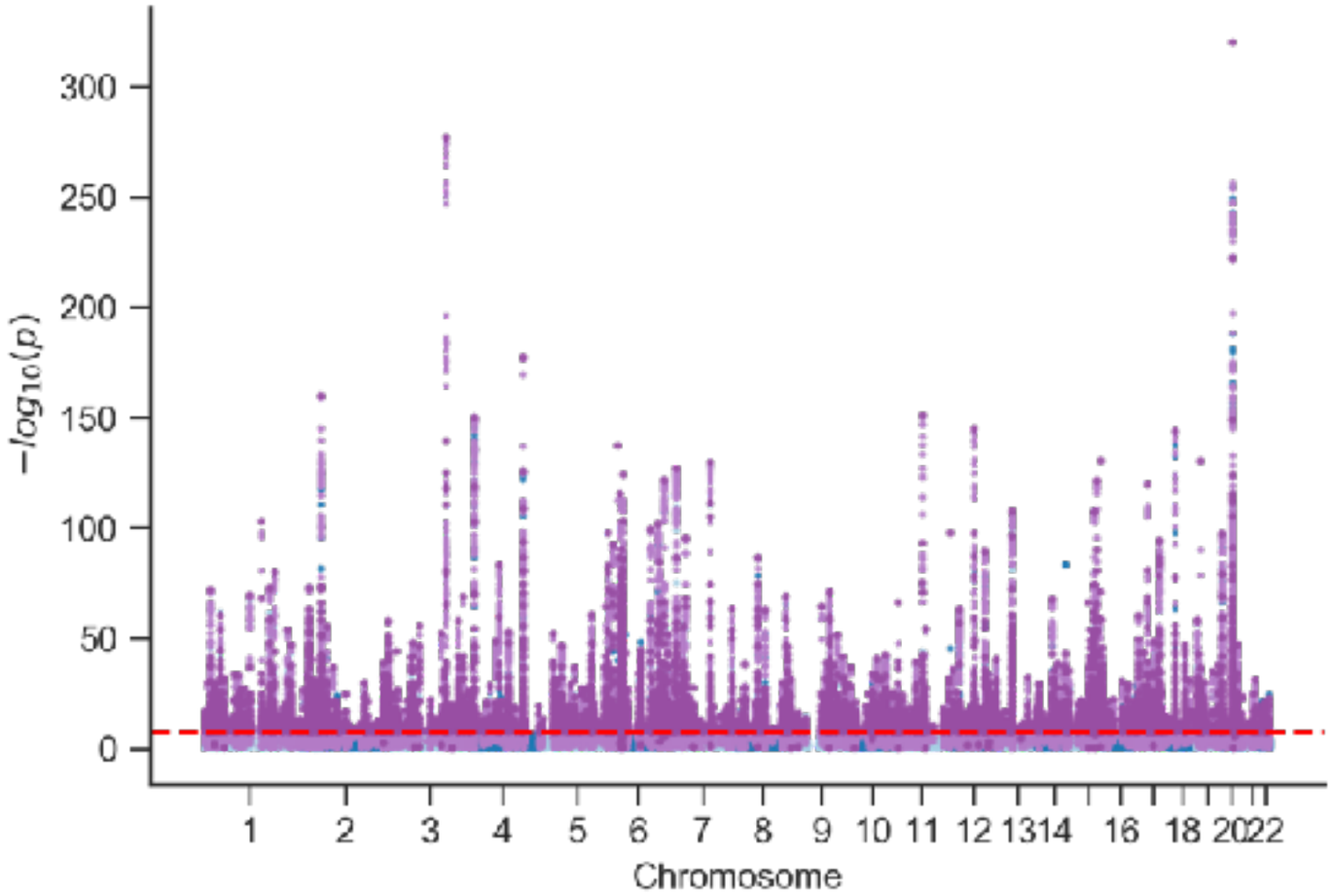


Figure 2 | Replication of the ISC-derived polygenic component in independent schizophrenia and bipolar disorder samples. Variance

What is a polygenic risk score?

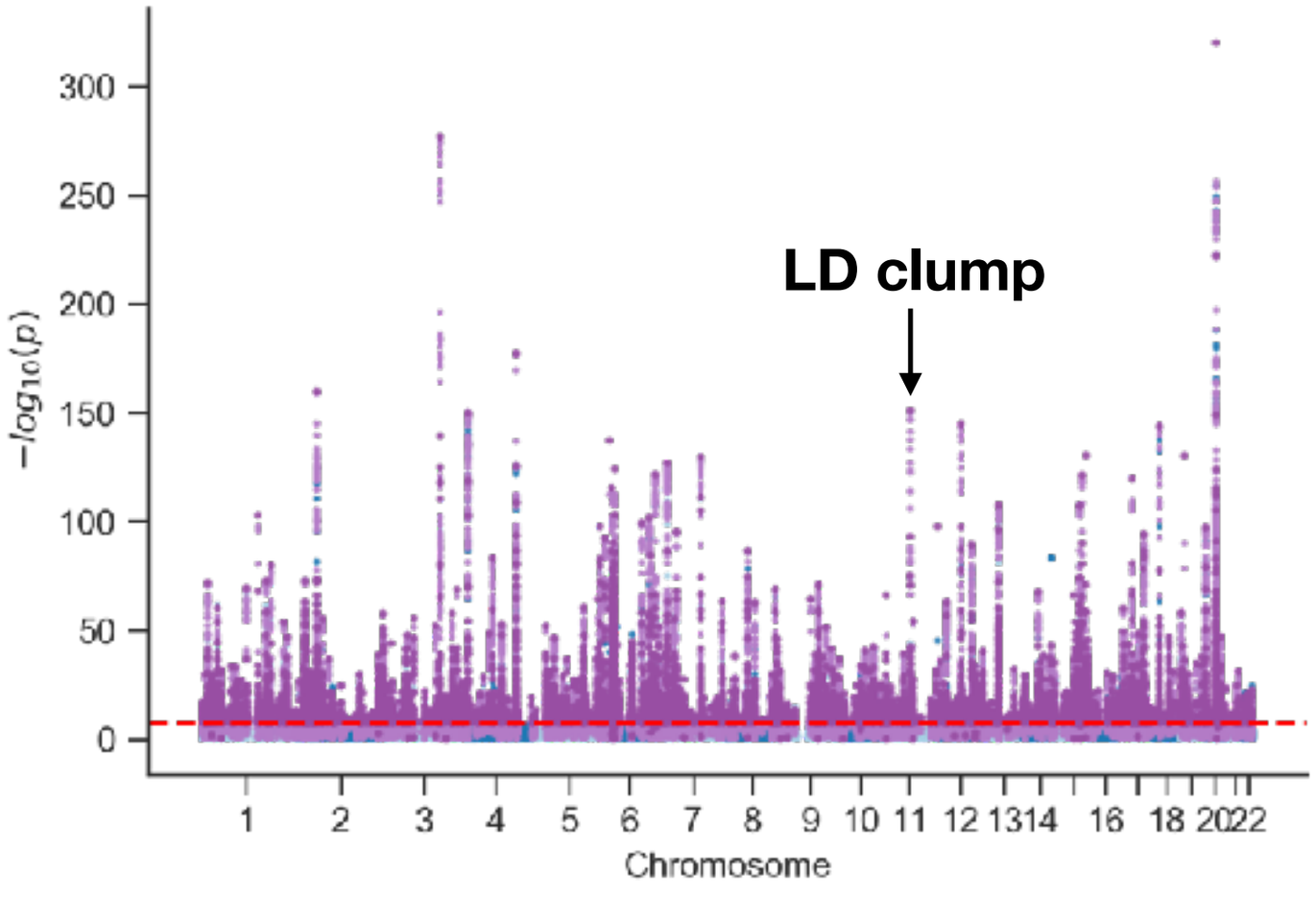


Genetic prediction of an individual's phenotype

$$Y = \sum_{j=1}^{m} g_j \beta_j$$

Sum the products of genotypes × effect size estimates from a GWAS across the genome

What is a polygenic risk score?



Genetic prediction of an individual's phenotype

$$Y = \sum_{j=1}^{m} g_j \beta_j$$

Sum the products of genotypes × effect size estimates from a GWAS across the genome

Fundamental choices:

- Which SNPs to include
- What weights to apply

Considerations:

- LD
- P-value thresholds

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

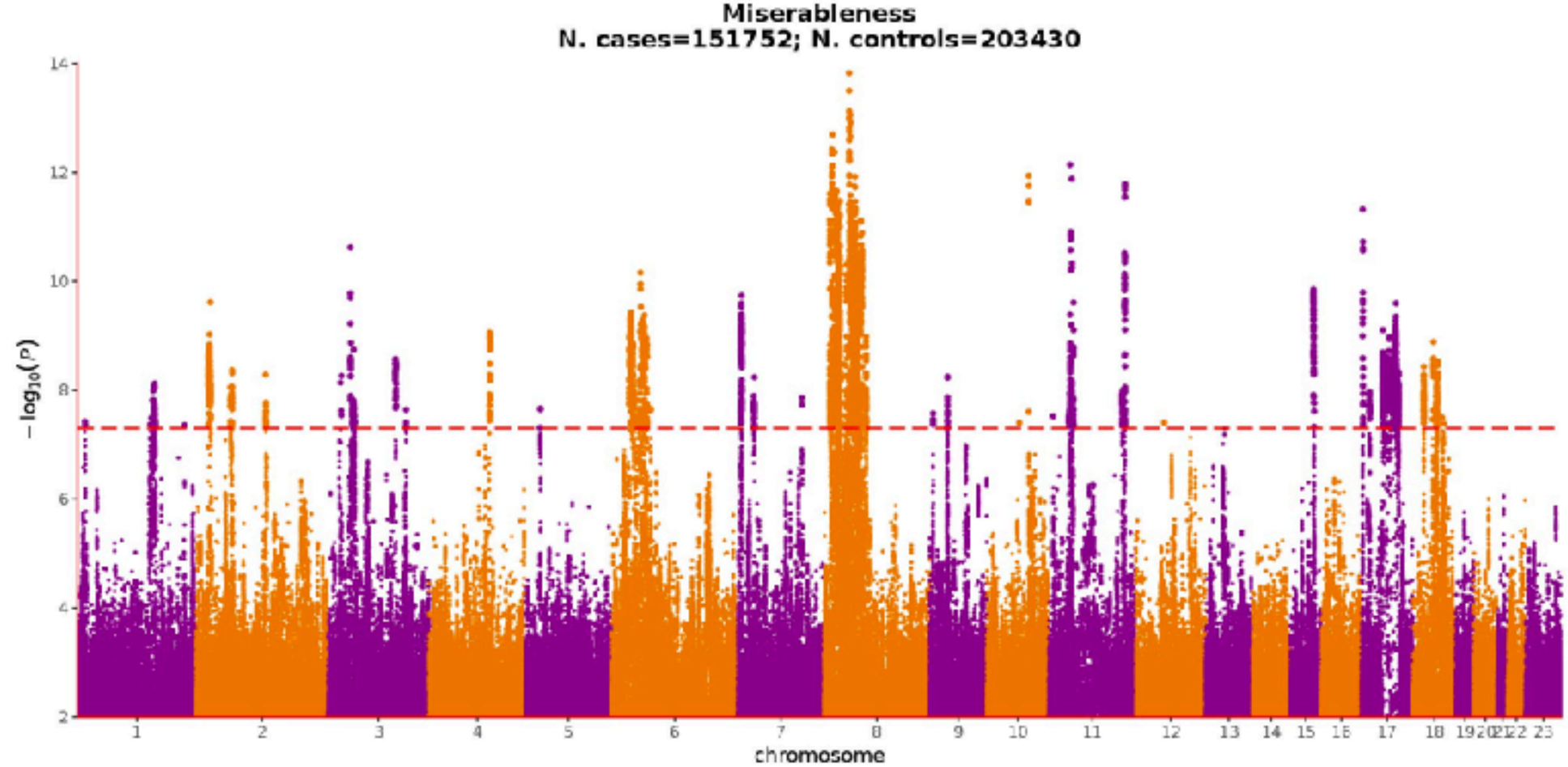
Content shamelessly borrowed and modified from Matthew Keller

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

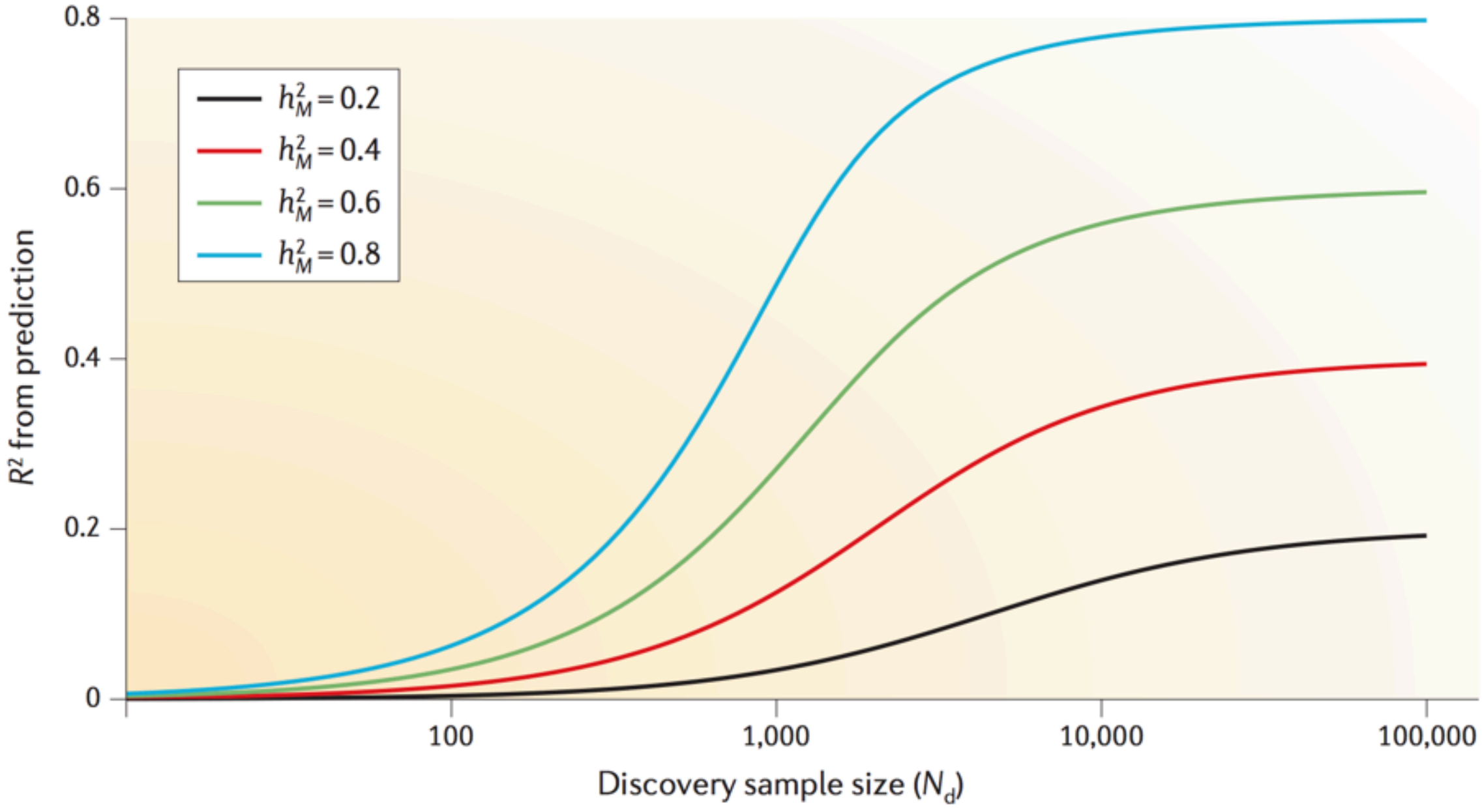
Content shamelessly borrowed and modified from Matthew Keller

1. Obtain large GWAS



Trait info: http://www.ukbiobank.ac.uk/data-showcase/
All things UK Biobank GWAS: http://www.nealelab.is/uk-biobank/

More powerful GWAS = more accurate predictor

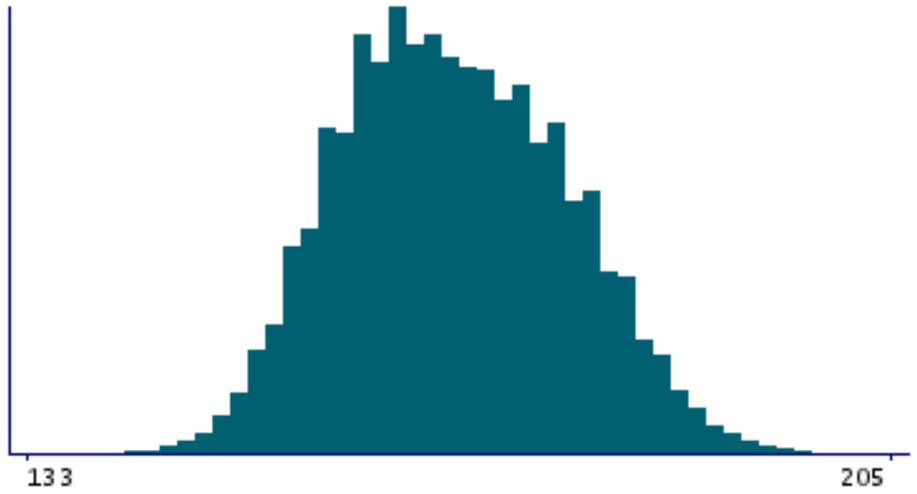


Wray, N.R., et al. (2013). Nat. Rev. Genet. 14, 507–515.

What do GWAS summary statistics contain?

Minimal useful info: variant ID, chromosome, position, risk and protective allele, sample size, p-value, effect size, standard error

Example: standing height



variant	minor_allele	minor_AF	low_confidence_variant	n_complete_samples	AC	ytx	beta	se	tstat	pval
1:15791:C:T	T	5.44E-09	TRUE	360388	3.92E-03	3.47E-03	1.80E+01	1.78E+02	1.01E-01	9.19E-01
1:69487:G:A	A	5.76E-06	TRUE	360388	4.15E+00	-8.75E-02	-4.13E-02	3.5E-01	-1.18E-01	9.06E-01
1:69569:T:C	C	1.88E-04	TRUE	360388	1.36E+02	-2.08E+00	-4.70E-02	6.27E-02	-7.5E-01	4.54E-01
1:139853:C:T	T	5.67E-06	TRUE	360388	4.09E+00	-1.06E-01	-4.21E-02	3.5E-01	-1.2E-01	9.04E-01
1:692794:CA:C	C	1.11E-01	FALSE	360388	7.97E+04	1.02E+02	7.97E-04	2.90E-03	2.75E-01	7.83E-01
1:693731:A:G	G	1.16E-01	FALSE	360388	8.35E+04	-6.93E+01	-1.44E-03	2.74E-03	-5.24E-01	6.00E-01
1:707522:G:C	C	9.73E-02	FALSE	360388	7.01E+04	-7.86E+00	2.47E-04	3.08E-03	8.02E-02	9.36E-01
1:717587:G:A	A	1.57E-02	FALSE	360388	1.13E+04	5.47E+00	1.13E-03	7.35E-03	1.54E-01	8.77E-01
1:723329:A:T	T	1.73E-03	FALSE	360388	1.25E+03	3.87E+01	2.22E-02	2.17E-02	1.02E+00	3.06E-01

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

Content shamelessly borrowed and modified from Matthew Keller

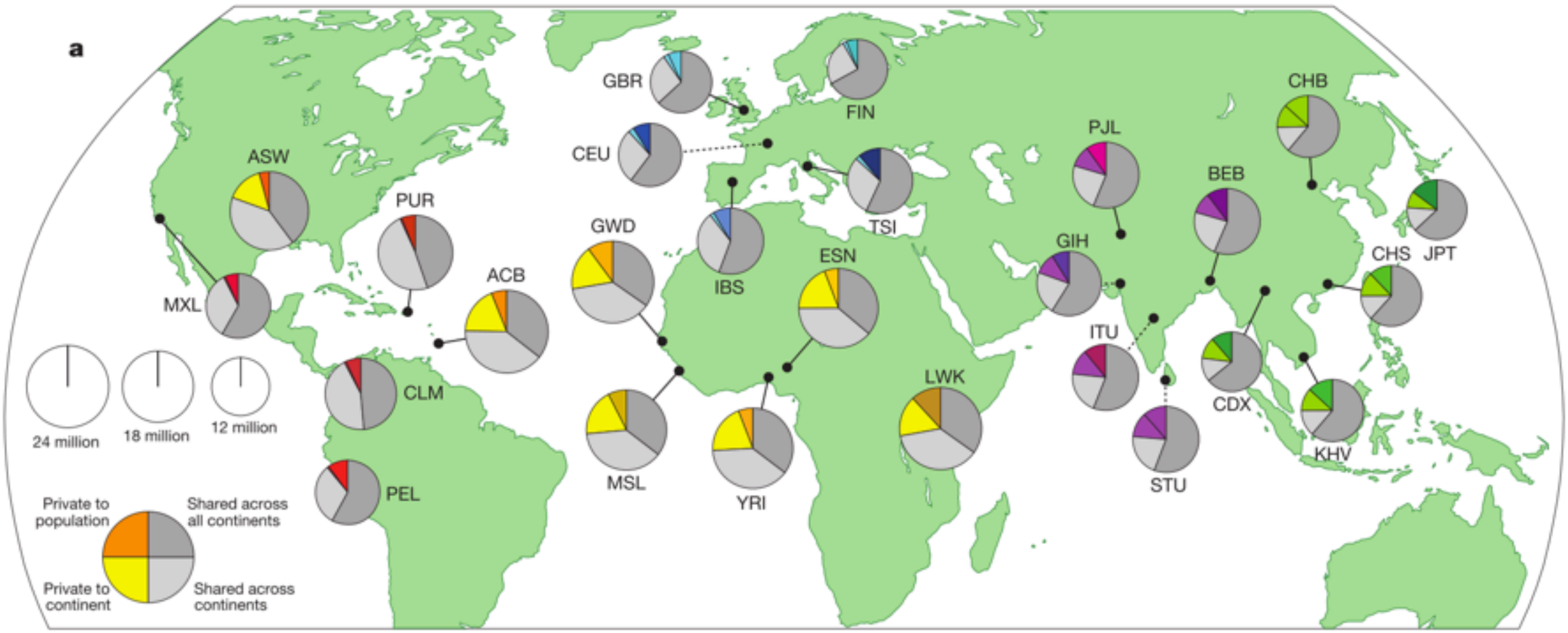
2. Independent target cohort must be independent

Prediction “accuracy” measures will be overestimated if discovery and target are not independent. This can arise if:

- The same people are in both cohorts
- There are close relatives between the two
- SNPs are selected from meta-analysis of discovery + target

Choose your favorite dataset

Most people like phenotypes, but...



The 1000 Genomes Project Consortium. (2015). Nature 526, 68–74.

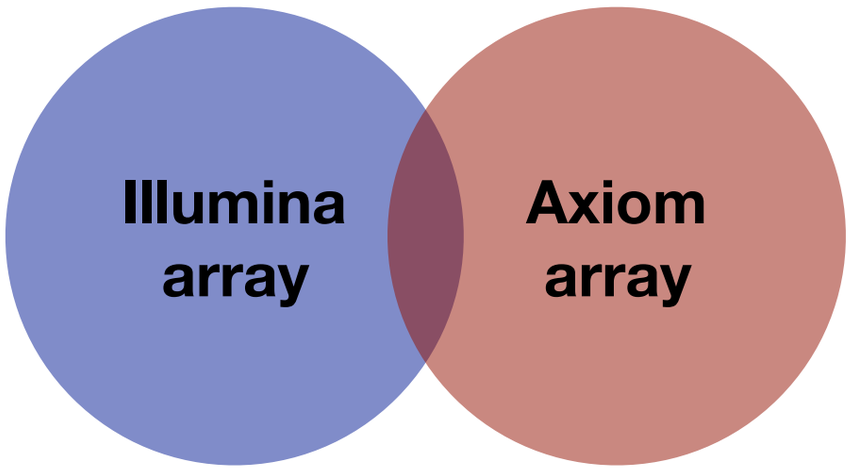
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

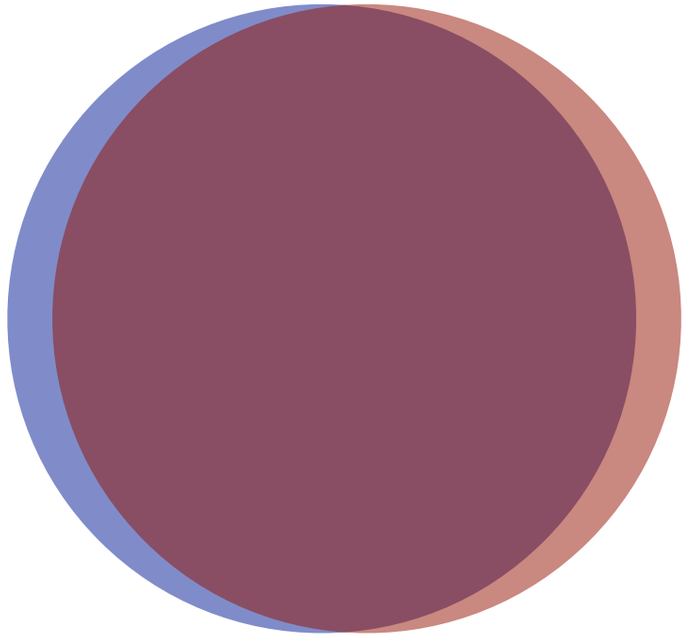
Content shamelessly borrowed and modified from Matthew Keller

3. Use SNPs in common

Array data



Imputed data



Phase and impute data to help maximize overlap

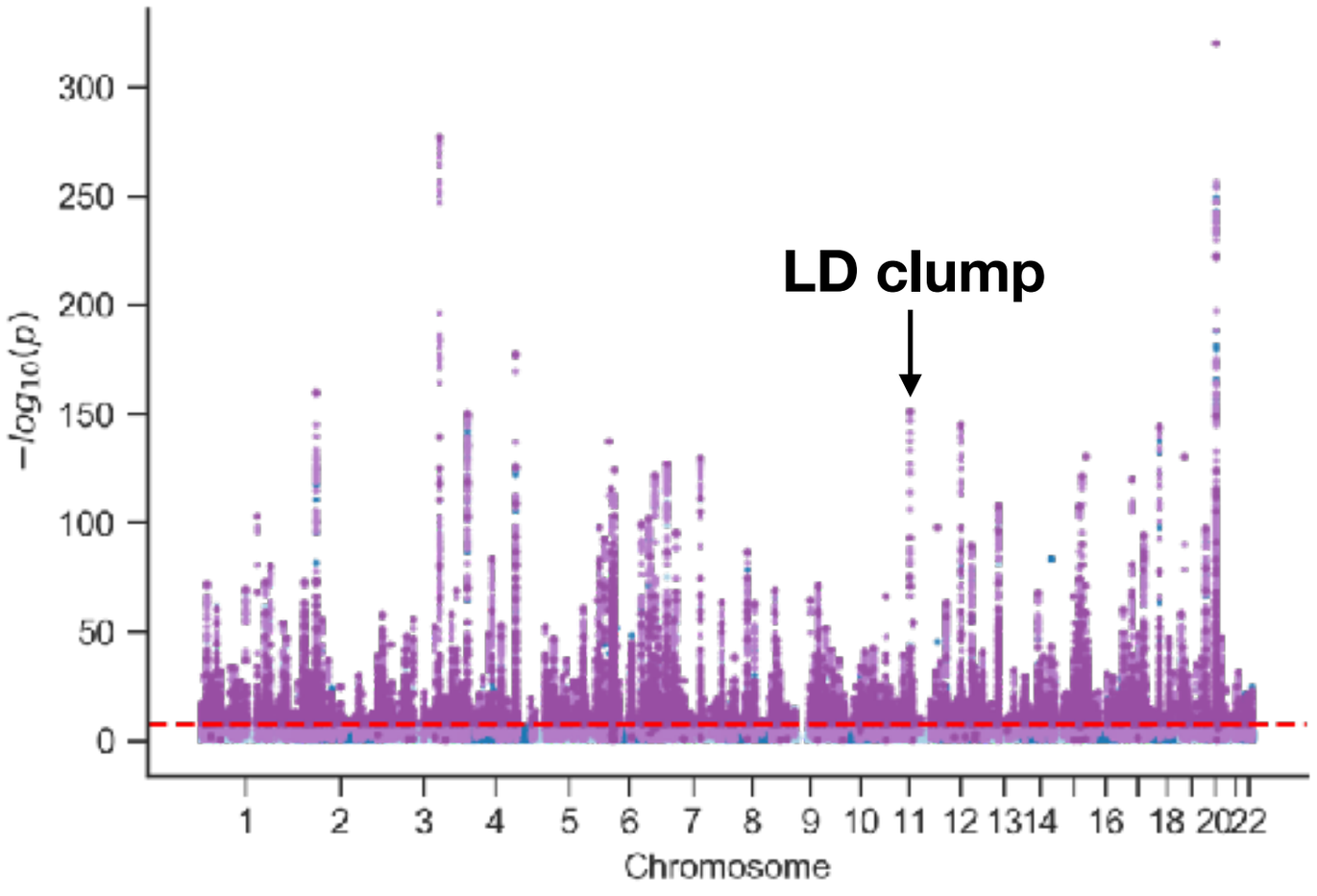
Content shamelessly borrowed and modified from Matthew Keller

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

Content shamelessly borrowed and modified from Matthew Keller

4. Account for LD



Two primary approaches:

- LD clumping (heuristic, less good)
  - In PLINK, --clump
- Model LD! LDPred (better, but harder to run)

Clumping with PLINK

Example:

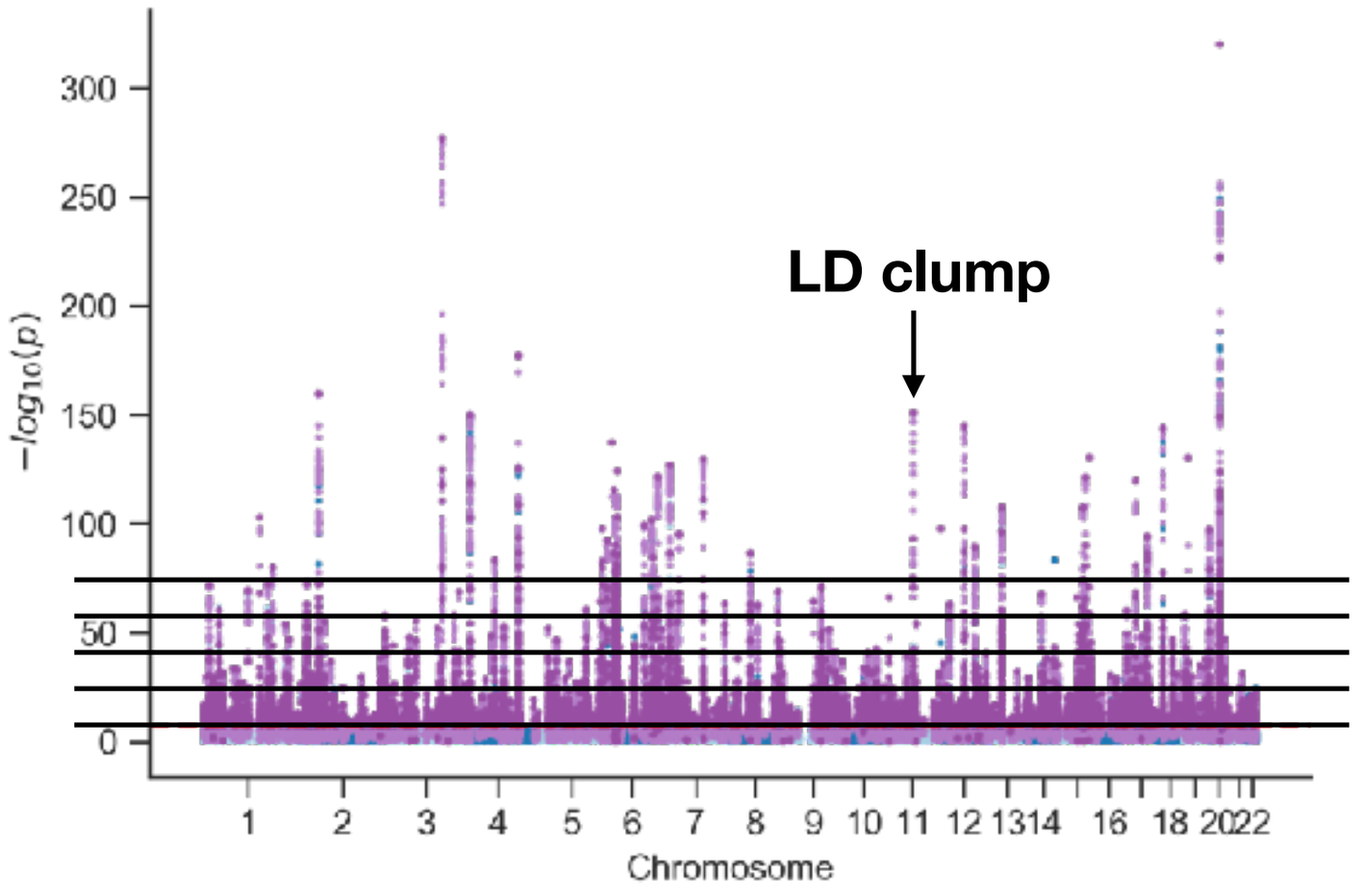
```
plink --bfile [reference LD panel] \
--clump [summary statistics] \
--clump-field [summary statistics p-value column name] \
--clump-snp-field [summary statistics snp column name] \
--clump-p1 1 \
--clump-p2 1 \
--clump-r2 0.5 \
--clump-kb 250 \
--out [output filename]
```

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

Content shamelessly borrowed and modified from Matthew Keller

5. Use various p thresholds



Use p-thresholds from 5e-8, 1-e7,...0.05...1
Report results from all thresholds

For PLINK

Create a file with multiple thresholds, for example:

[Threshold name] [lower bound] [upper bound]

```
s1      0       0.00000005
s2      0       0.000001
s3      0       0.0001
s4      0       0.001
s5      0       0.01
s6      0       0.05
s7      0       0.1
s8      0       0.2
s9      0       0.5
s10     0       1
```

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

Content shamelessly borrowed and modified from Matthew Keller

6. Calculate PRS

- $PRS_j = \Sigma [\beta_{i,discovery} * SNP_{ij}]$
  - $\beta_{i,discovery}$ = effect size in discovery sample from
    - linear regression (continuous trait)
    - logistic regression (binary trait; β = log(OR))
- SNP_{ij} = # alleles (0,1,2) for SNP i of person j in target sample
- In PLINK, --score.

In PLINK

Example:

```
plink --bfile [best guess genotypes] \
--extract [clumped snps] \
--q-score-range [range file] [summary stats] [variant ID
column #] [p-value column #] [header] \
--score [summary stats] [variant ID column #] [allele column
#] [effect size column #] \
--out [output file]
```

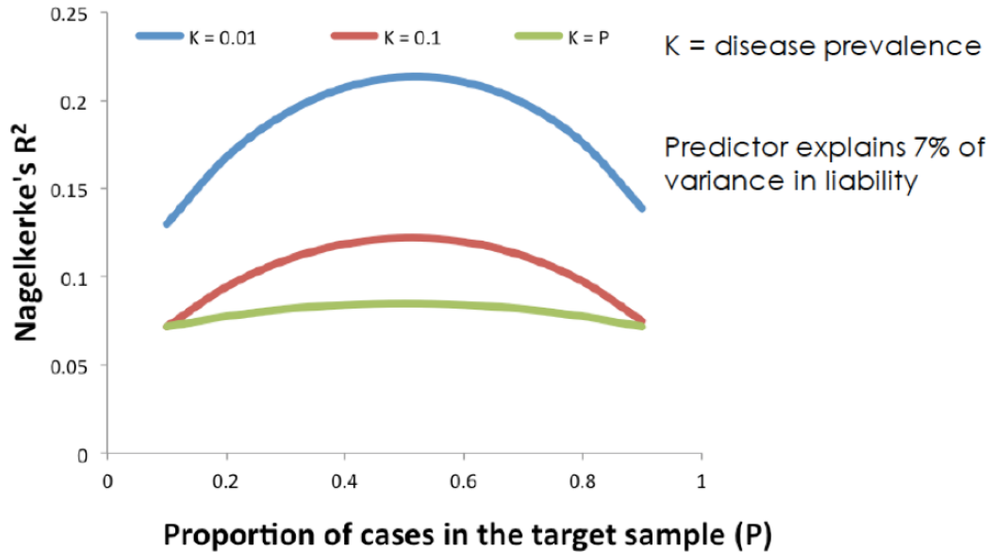
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with **p < various thresholds** (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by β from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. R^2)

Content shamelessly borrowed and modified from Matthew Keller

7. Evaluate PRS accuracy

- For continuous traits, this is simply the R^2 from regressing trait ~ PRS in target + covariates
- Trickier for binary (e.g., case-control) data due to ascertainment
- Often Nagelkerke's R^2 is reported. Unfortunate, because this depends on prevalence and case:control ratio.



Content shamelessly borrowed and modified from Matthew Keller

7. Evaluate PRS accuracy

- For continuous traits, this is simply the R^2 from regressing trait ~ PRS in target + covariates
- Trickier for binary (e.g., case-control) data due to ascertainment
  - Often Nagelkerke's R^2 is reported. Unfortunate, because this depends on prevalence + case:control ratio.
  - Better: liability-scale R^2

Content shamelessly borrowed and modified from Matthew Keller

7. Please report comparable R^2 !

(thrilling stuff, I know)

TABLE I. Brief description of R^2 measures used in this study and their theoretical expectation

Brief description	Notation and formula	Expectation
R^2 on the observed scale	$R^2_O = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i^N (y_i - \bar{y})^2}$	$h^2_l \frac{z^2}{K(1-K)}$
Cox and Snell's R^2 on the observed scale	$R^2_{C\&S} = 1 - \left\{ \frac{\text{Likelihood}_{\text{null}}}{\text{Likelihood}_{\text{full}}} \right\}^{2/N}$	$h^2_l \frac{z^2}{K(1-K)}$
Nagelkerke's R^2 on the observed scale (hard to compare)	$R^2_N = \frac{R^2_{C\&S}}{1-(\text{Likelihood}_{\text{null}})^{2/N}}$	$\frac{R^2_{C\&S}}{1-K^{2K} \cdot (1-K)^{2(1-K)}}$
R^2 on the liability scale	$R^2_l = R^2_O \frac{K(1-K)}{z^2}$	h^2_l
R^2 on the probit liability scale (easy to compare!)	$R^2_{\text{probit}} = \frac{\text{var}(\hat{b}_{\text{probit}} g_i)}{\text{var}(\hat{b}_{\text{probit}} g_i)+1}$	h^2_l
R^2 on the logit liability scale	$R^2_{\text{logit}} = \frac{\text{var}(\hat{b}_{\text{logit}} g_i)}{\text{var}(\hat{b}_{\text{logit}} g_i)+3.29}$	h^2_l
R^2 on the liability scale using AUC	$R^2_{\text{AUC}} = \frac{2Q^2}{(m_2-m)^2+Q^2 m(m-t)+m_2(m_2-t)}$	h^2_l
R^2 on the liability scale when using ascertained case-control studies	$R^2_{l_{cc}} = \frac{R^2_{o_{cc}} C}{1+R^2_{o_{cc}} \theta C}$	h^2_l

y, observations that are 0 or 1 for unaffected and affected individuals; h^2_l, heritability on the liability scale, in this context the proportion of variance on the liability scale explained by the genetic profile; K, population prevalence; z, the height of a normal density curve at the point according to K; g, the sum of all additive genetic factors in the estimated genetic predictor; b, regression coefficient from generalized linear model; m, the mean liability for cases; m_2, the mean liability for controls; t, the threshold on the normal distribution that truncates the proportion of disease prevalence K; Q, the inverse of the cumulative density function of the normal distribution up to values of AUC; C and θ, correcting factors for ascertainment.

Lee, S.H., et al. (2012). Genet. Epidemiol. *36*, 214–224.

So now you have a PRS...

- What are polygenic risk scores?
- How to compute them
- Methods, interpretations, and uses
- Ancestry, health disparities, and ongoing/future directions

The rise of the polygenic risk score

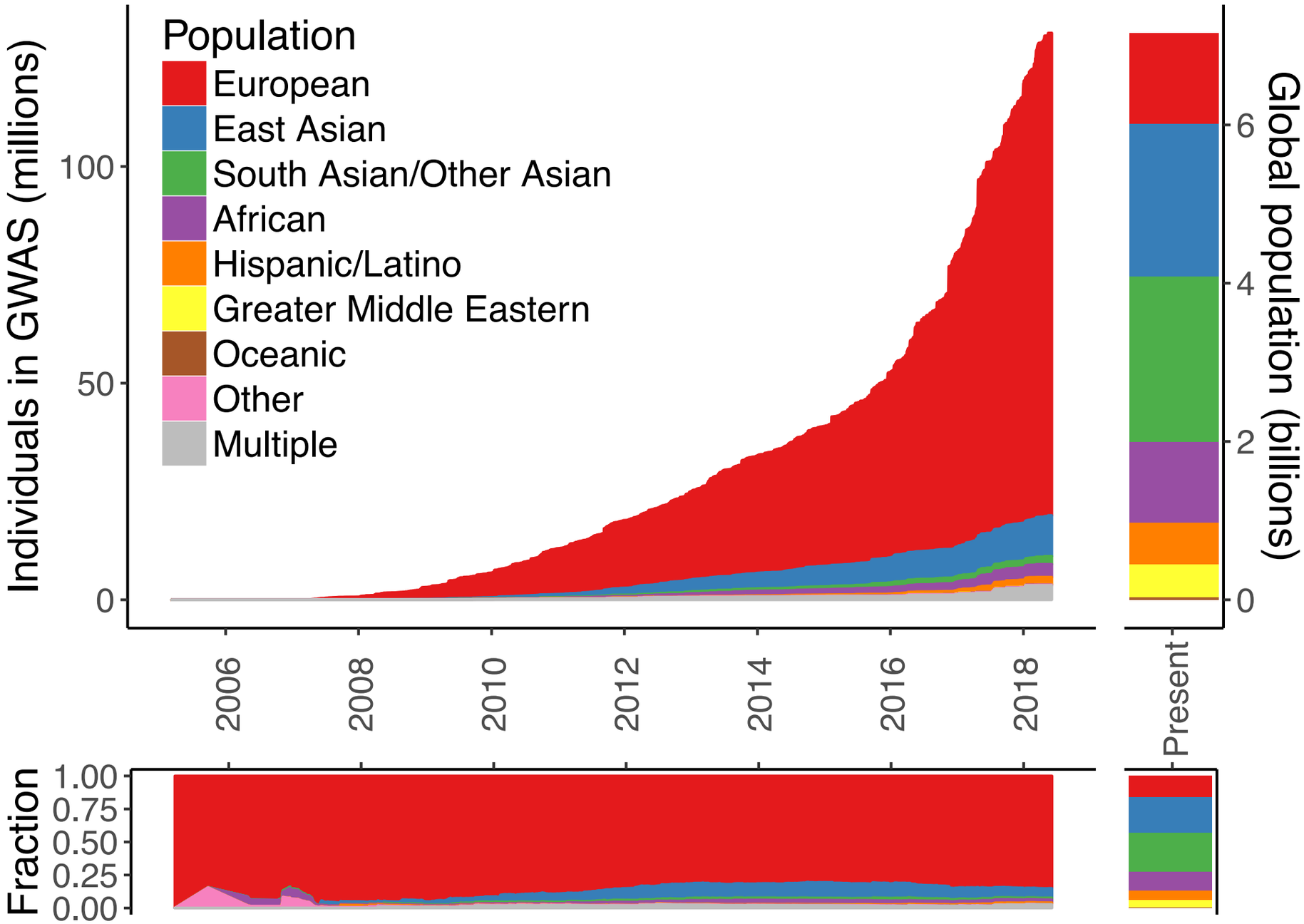
No discussion of ancestry!

> "We propose the time has come to incorporate genetic risk scores into clinical practice"

- Previous criticisms: limited sample size
- Cheap test for insights into many diseases
- Integrate with other clinical factors for therapeutic decision-making

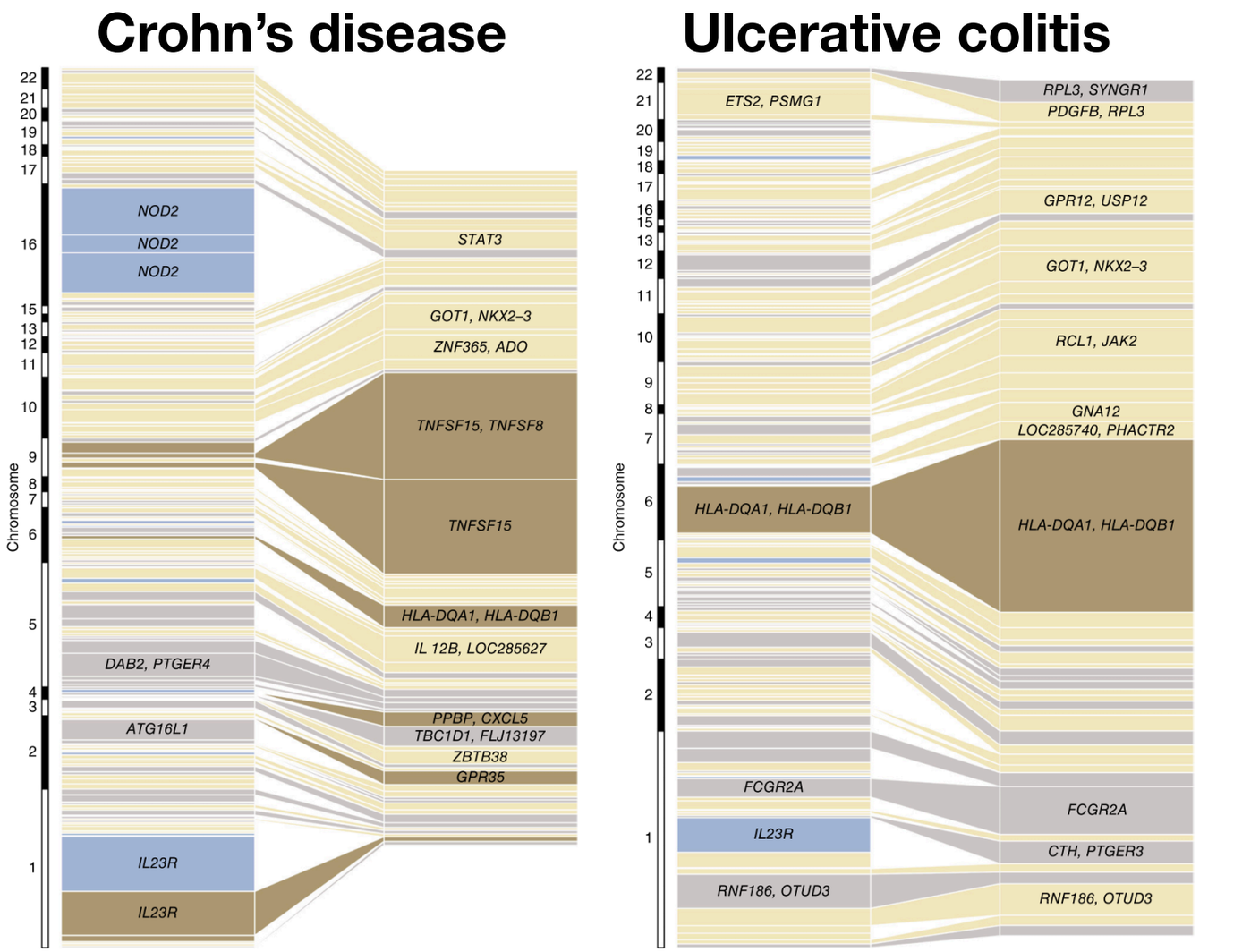
Knowles JW, Ashley EA (2018) Cardiovascular disease: The rise of the genetic risk score. PLoS Med 15(3): e1002546.

Genomics has a diversity problem



Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

Causal effects are mostly shared across populations



Monomorphic in non–Europeans

Similar MAF and OR

Different MAF

Different OR

Different MAF and different OR

Liu et al (2015) Nat Gen

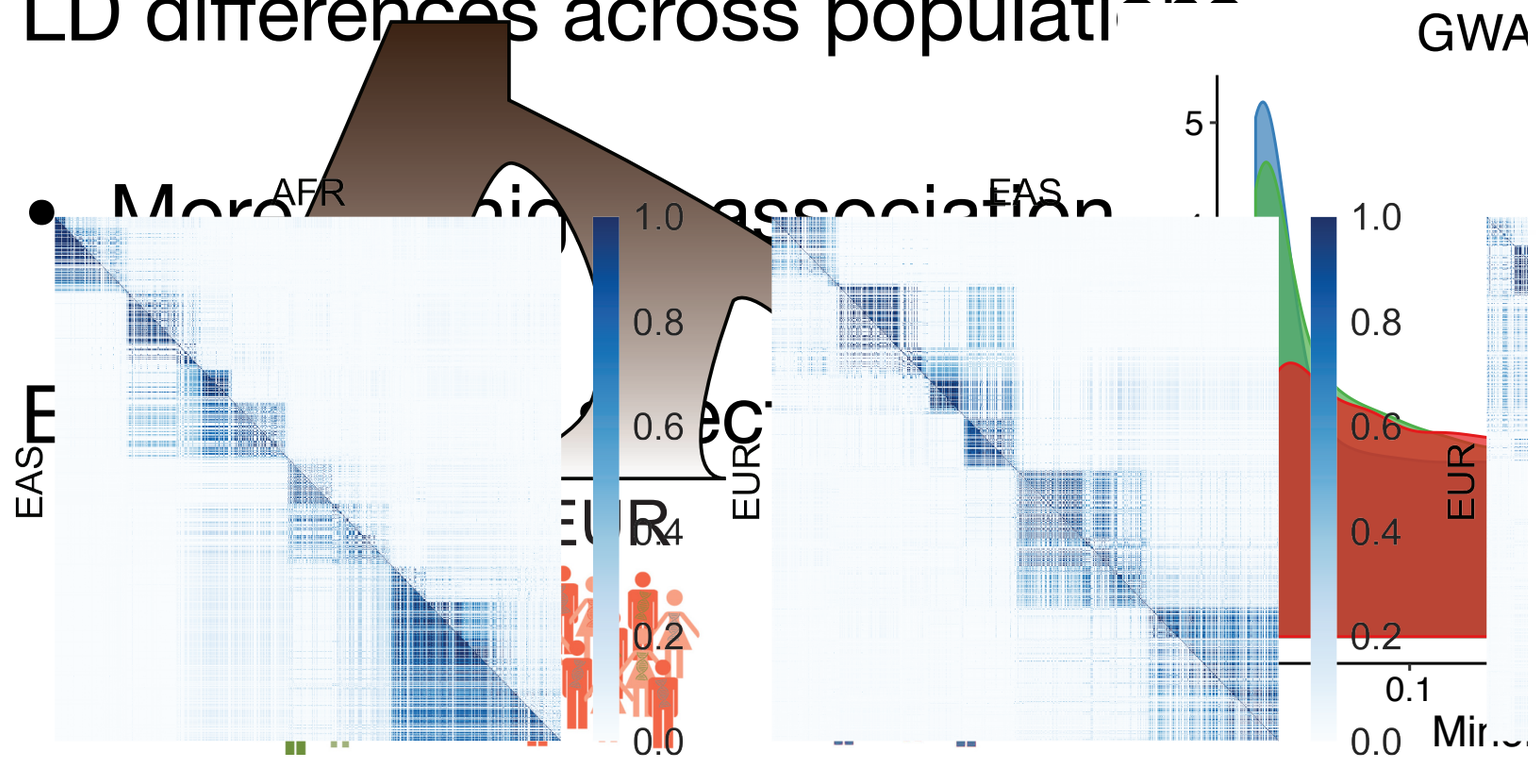
...but what about other effects?

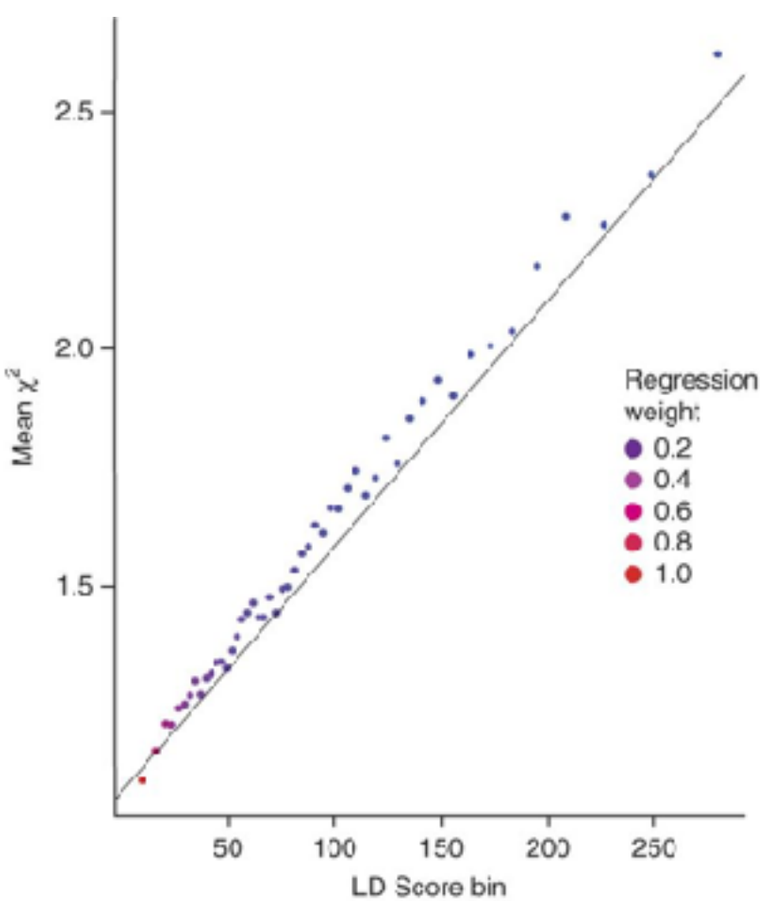
Predictable basis of PRS disparities

Prediction accuracy decays with F_{ST}

Why?

- GWAS best-powered to discover common variants
- LD differences across populations
  - More [illegible] association
  - [illegible]





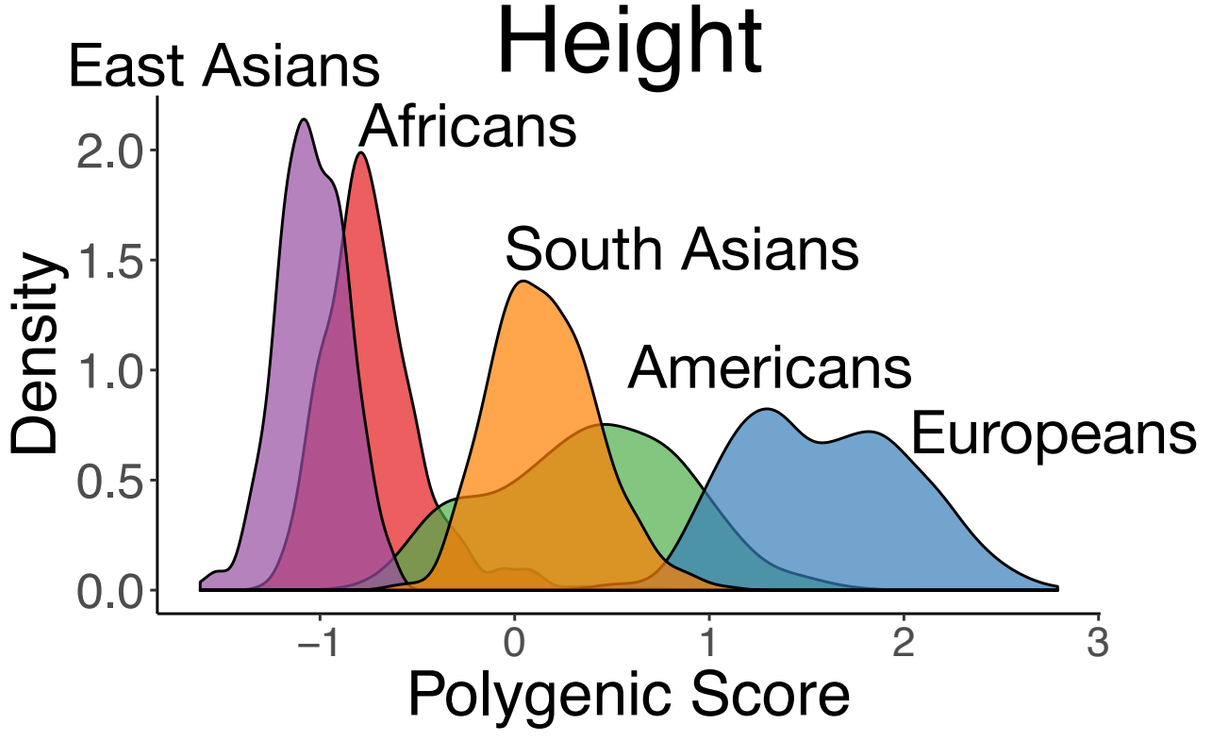
Please cite this article in press as: Martin et al., Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations, The American Journal of Human Genetics (2017), http://dx.doi.org/10.1016/j.ajhg.2017.03.004

ARTICLE

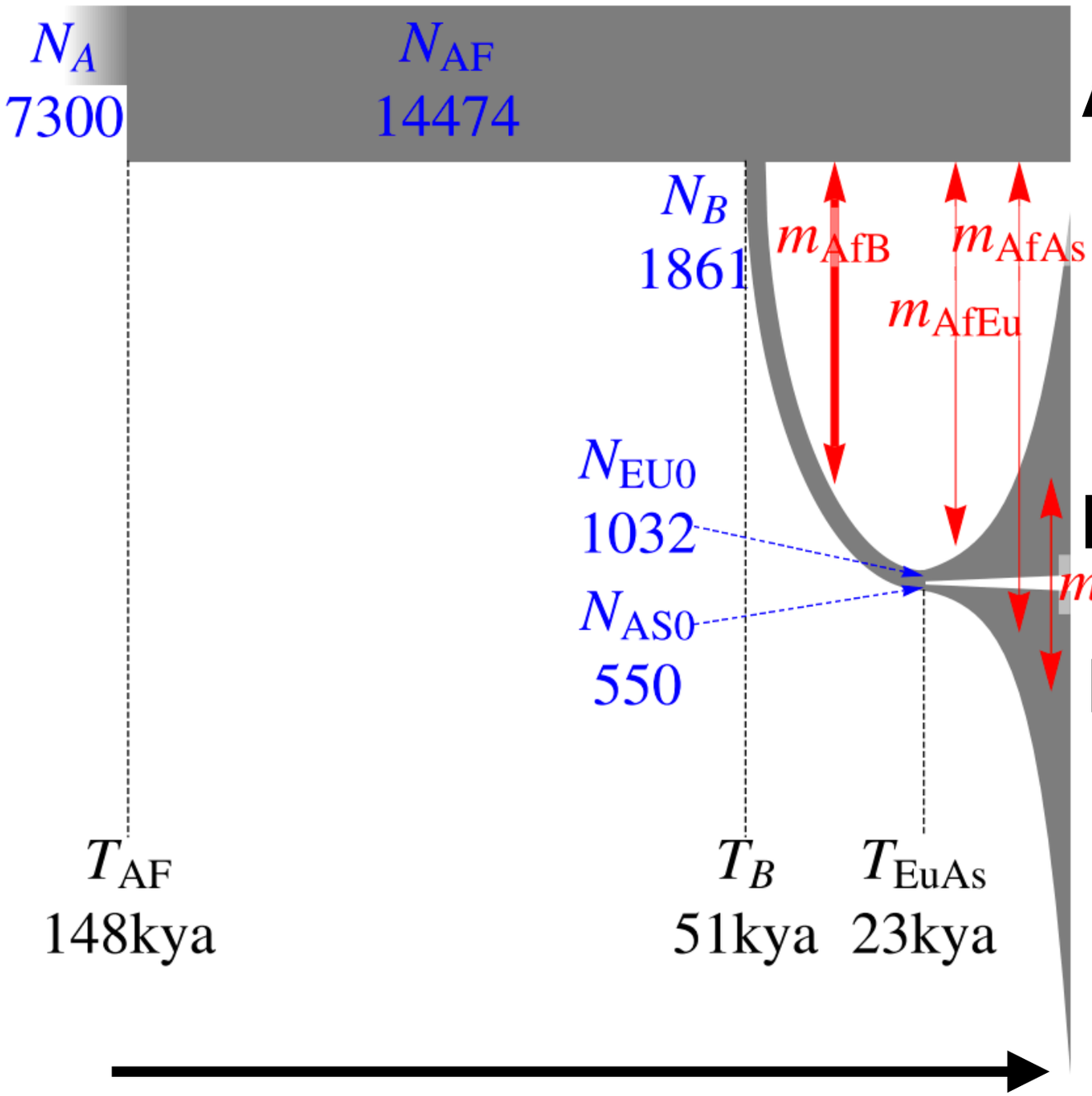
Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin,[1,2,3,4] Christopher R. Gignoux,[4] Raymond K. Walters,[1,2,3] Genevieve L. Wojcik,[4] Benjamin M. Neale,[1,2,3] Simon Gravel,[5,6] Mark J. Daly,[1,2,3] Carlos D. Bustamante,[4] and Eimear E. Kenny[7,8,9,10,*]

- Polygenic height scores are substantially different across populations
- These differences are not meaningful



Coalescent model for simulation framework



Africans

Europeans

East Asians

Model <u>parameters</u>

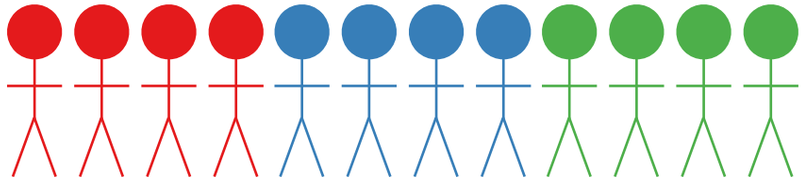
- N_e: population size
- m: migration rates
- T: time
- r: growth

Demographic model: Gravel, S., et al. (2011). Proc. Natl. Acad. Sci. U. S. A. *108*, 11983–11988.

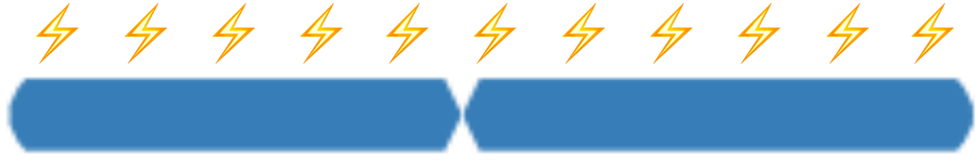
msprime: Kelleher, J., Etheridge, A.M., and Mcvean, G. (2016). PLoS Comput Biol 1–22.

Simulation overview

1. Simulate genotypes (AFR, EUR, EAS)



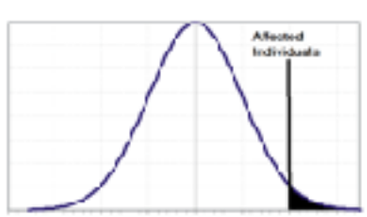
2. Assign evenly spaced causal variants



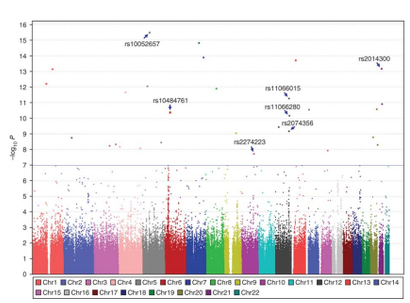
3. Compute PRS_{TRUE}

$$X = \sum_{i=1}^{m} g_i \beta_i$$

4. Define EUR cases, controls (10k each)

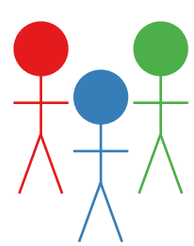


5. Run a EUR GWAS



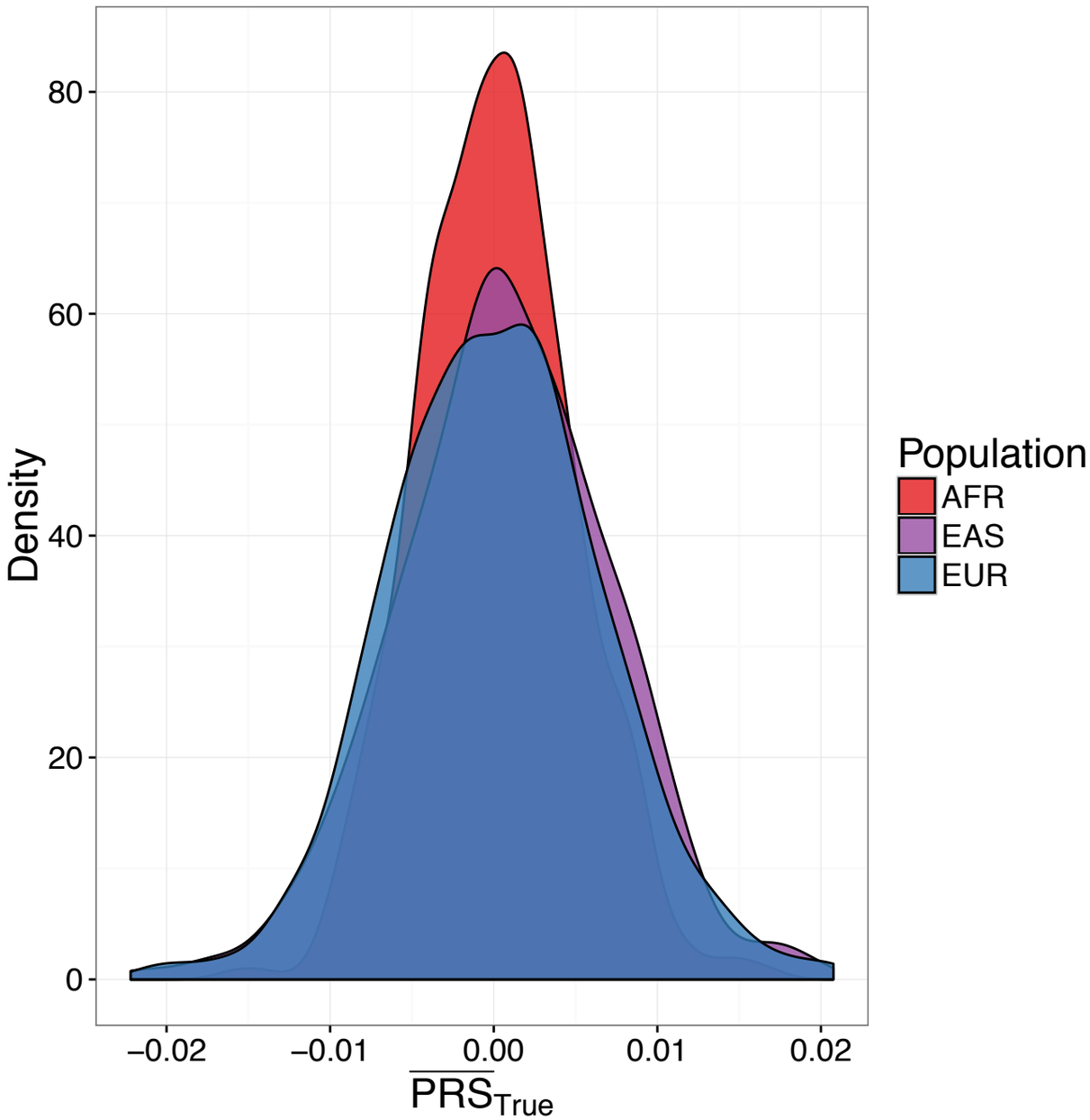
6. Compute PRS_{INFER} across populations

$$X = \sum_{i=1}^{m} g_i \beta_i \longrightarrow$$



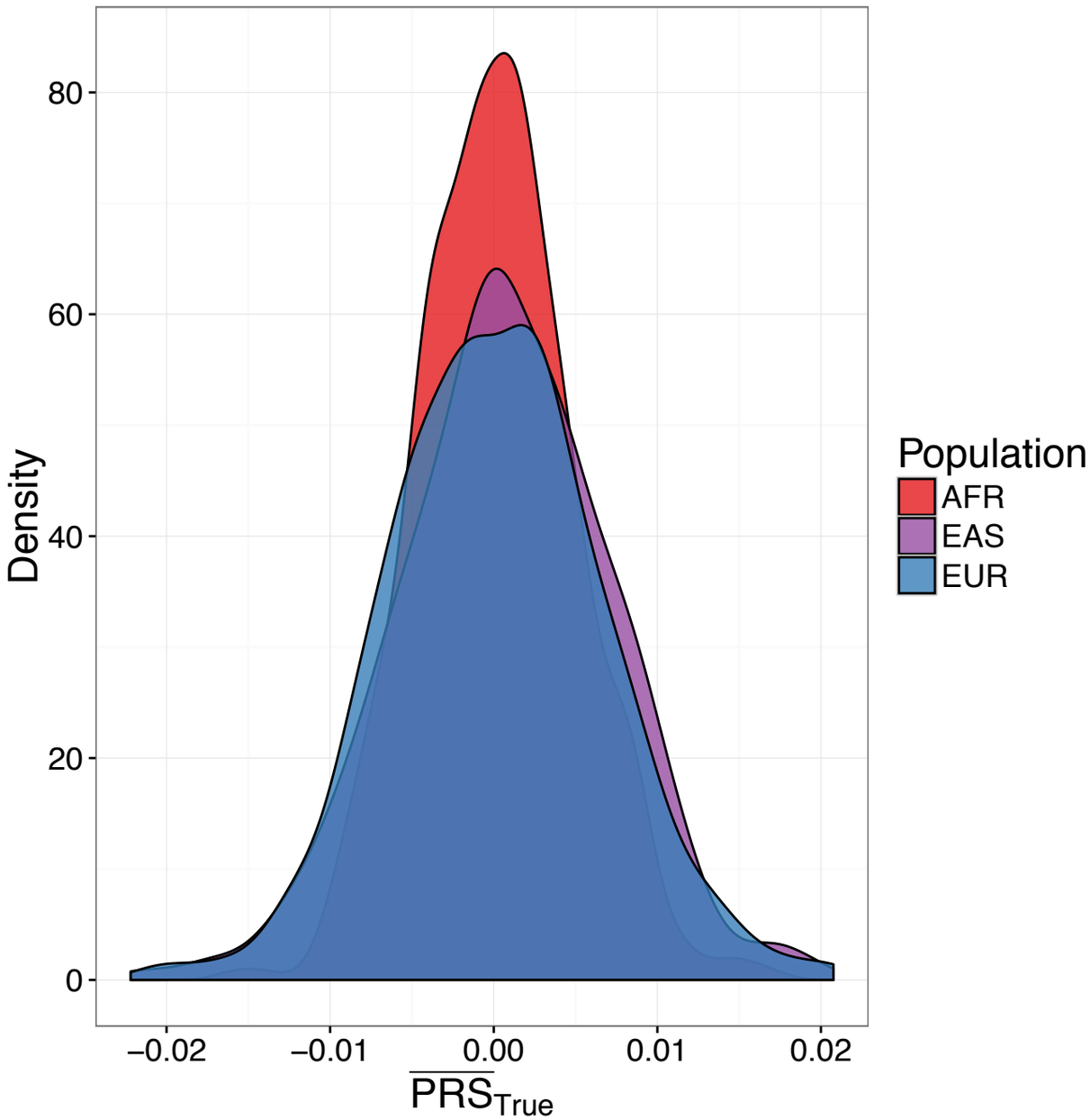
PRS_{TRUE} is not significantly different across populations

True causal variants

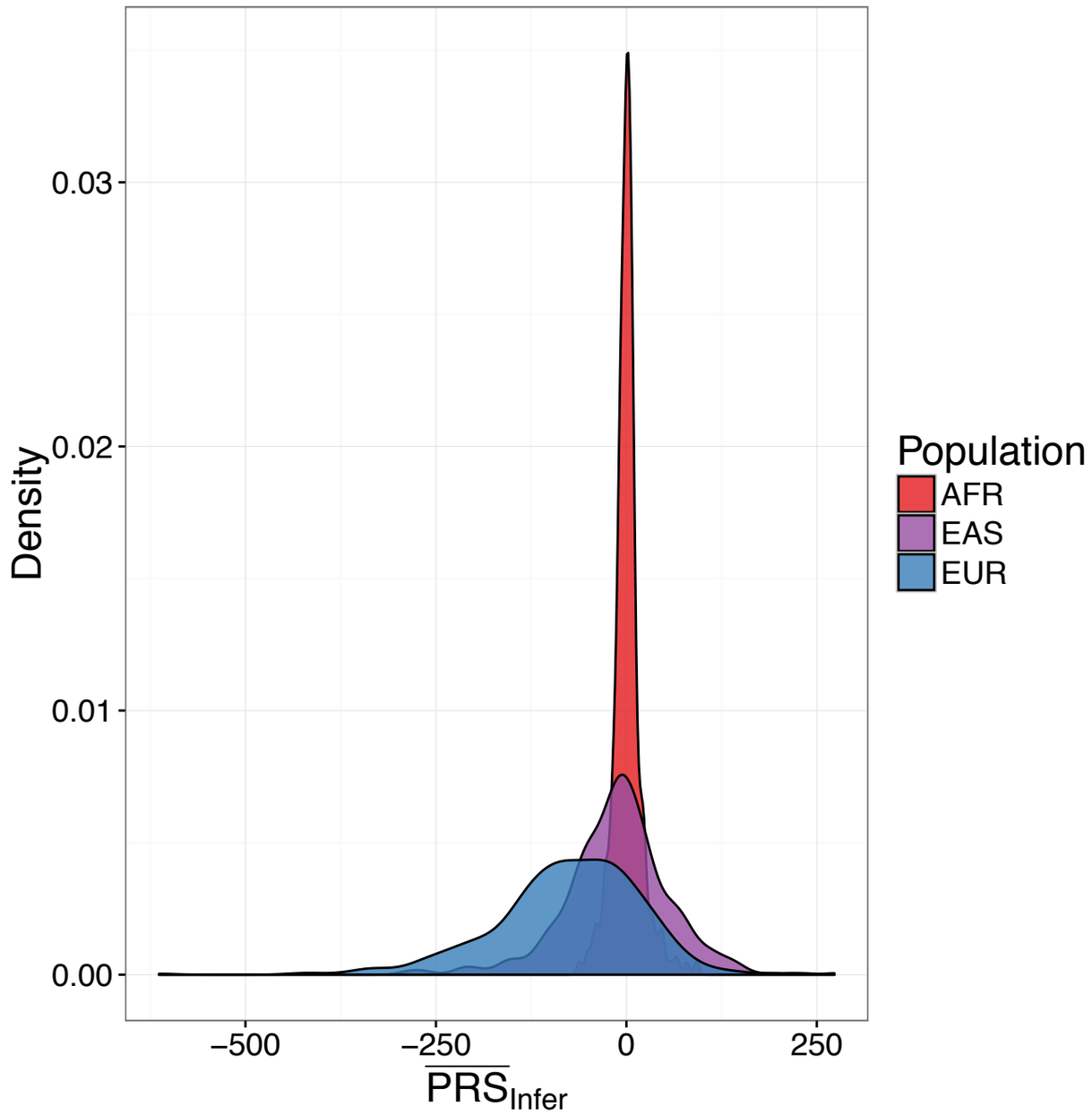


PRS_{INFER} is highly stratified across populations

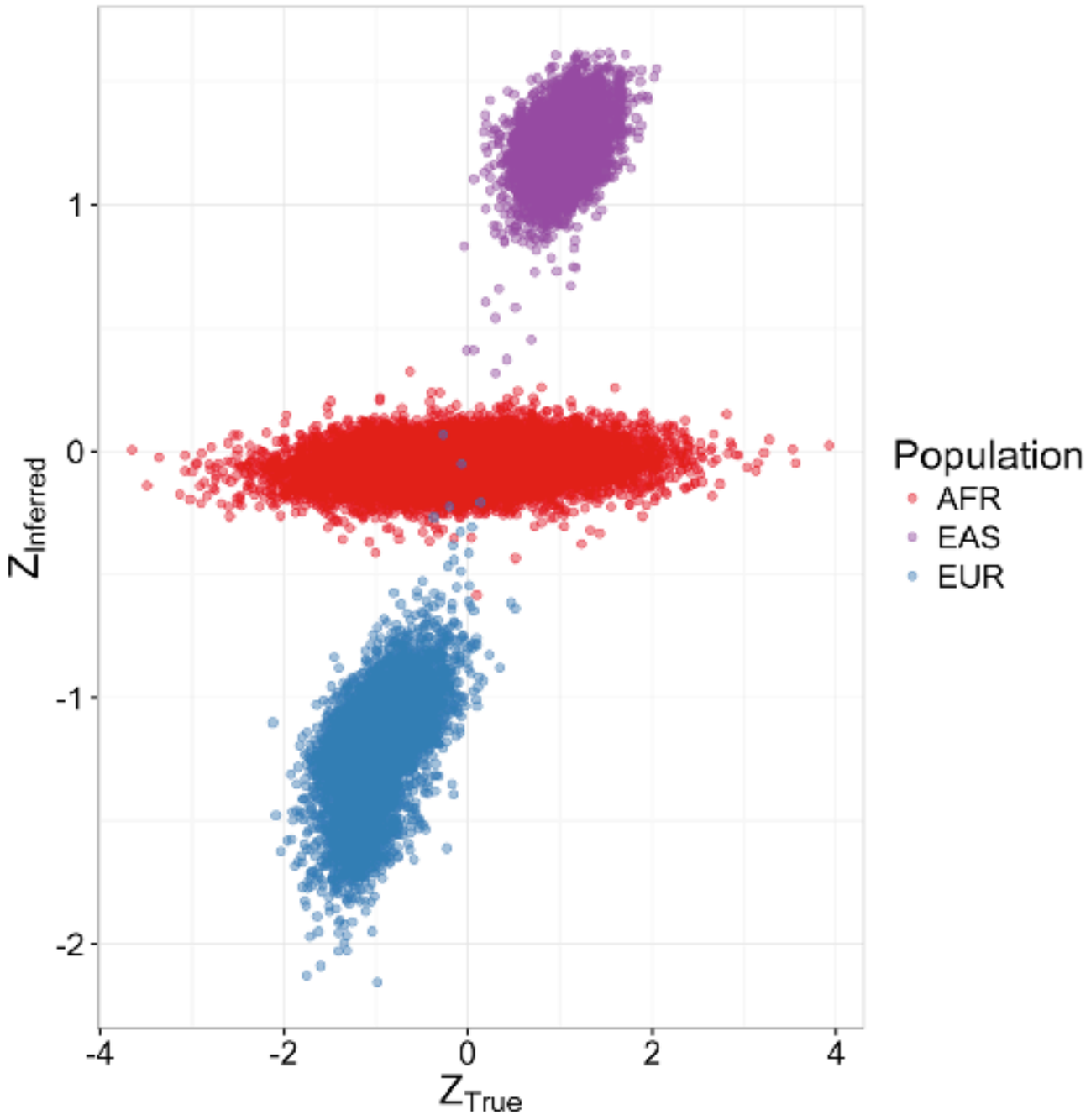
True causal variants



GWAS inferred variants



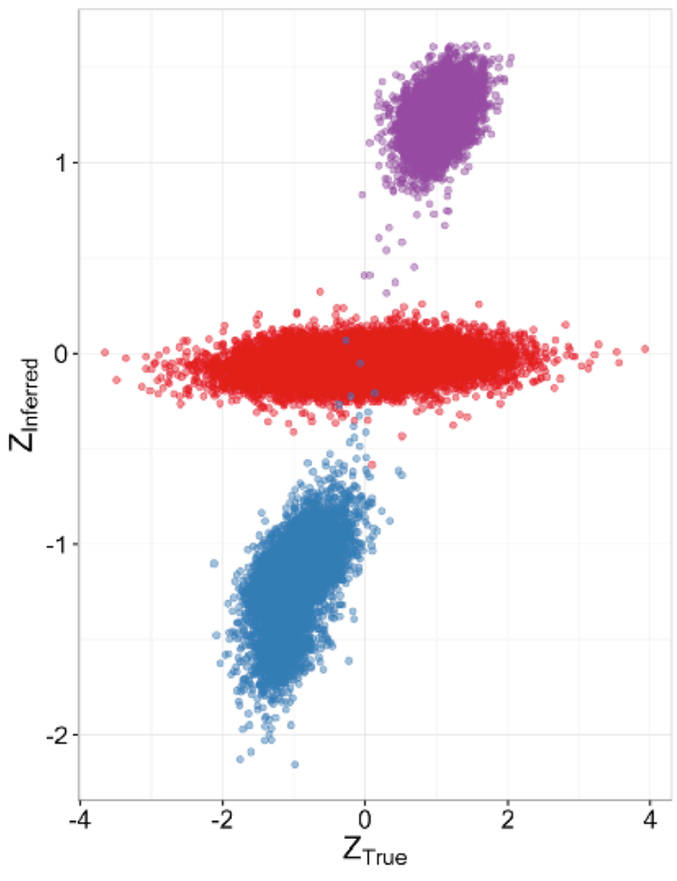
Unpredictable PRS biases across populations



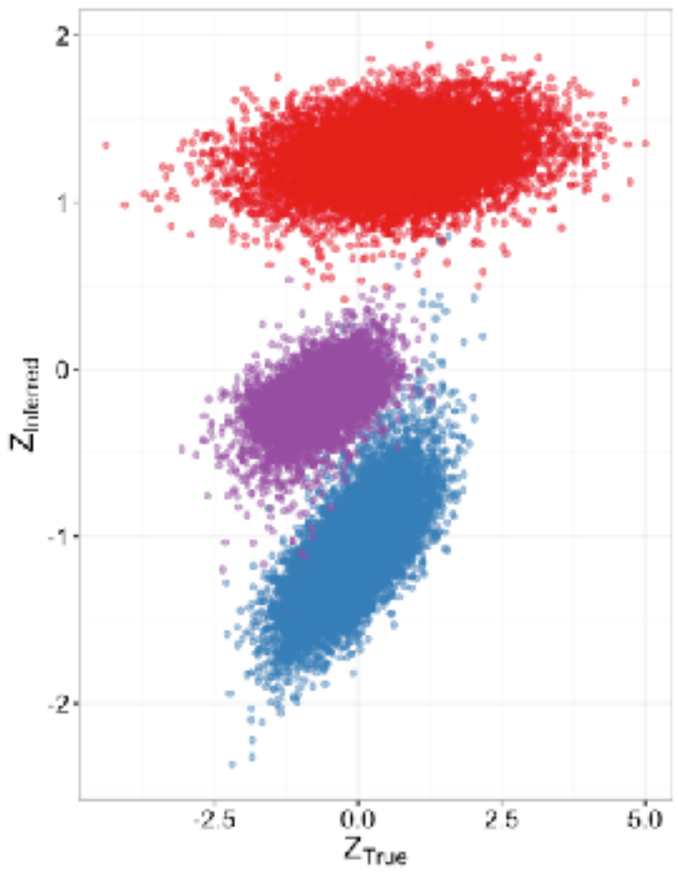
Unpredictable PRS biases across populations

Analogous to different traits:

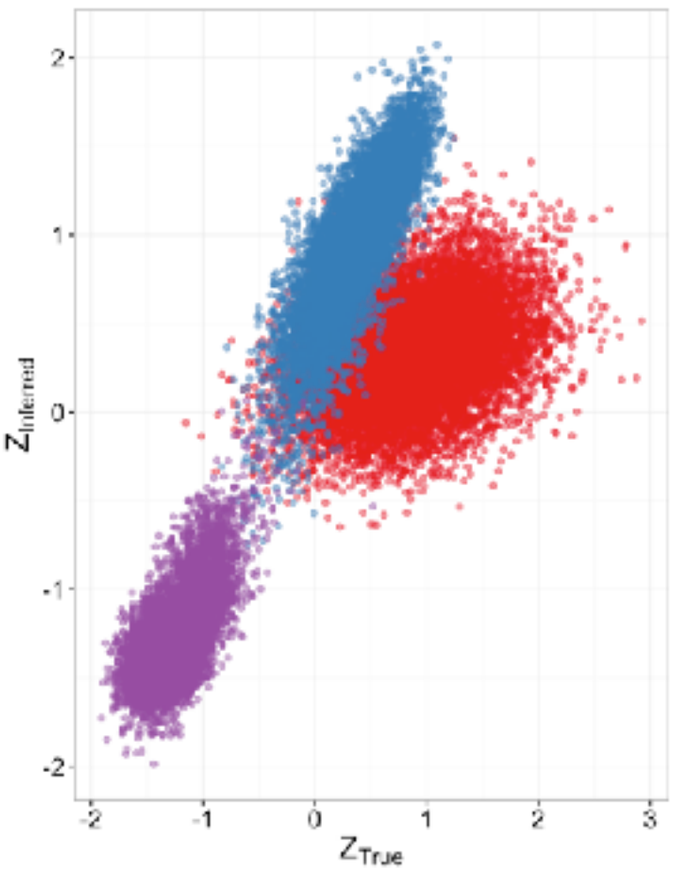
Height



Schizophrenia

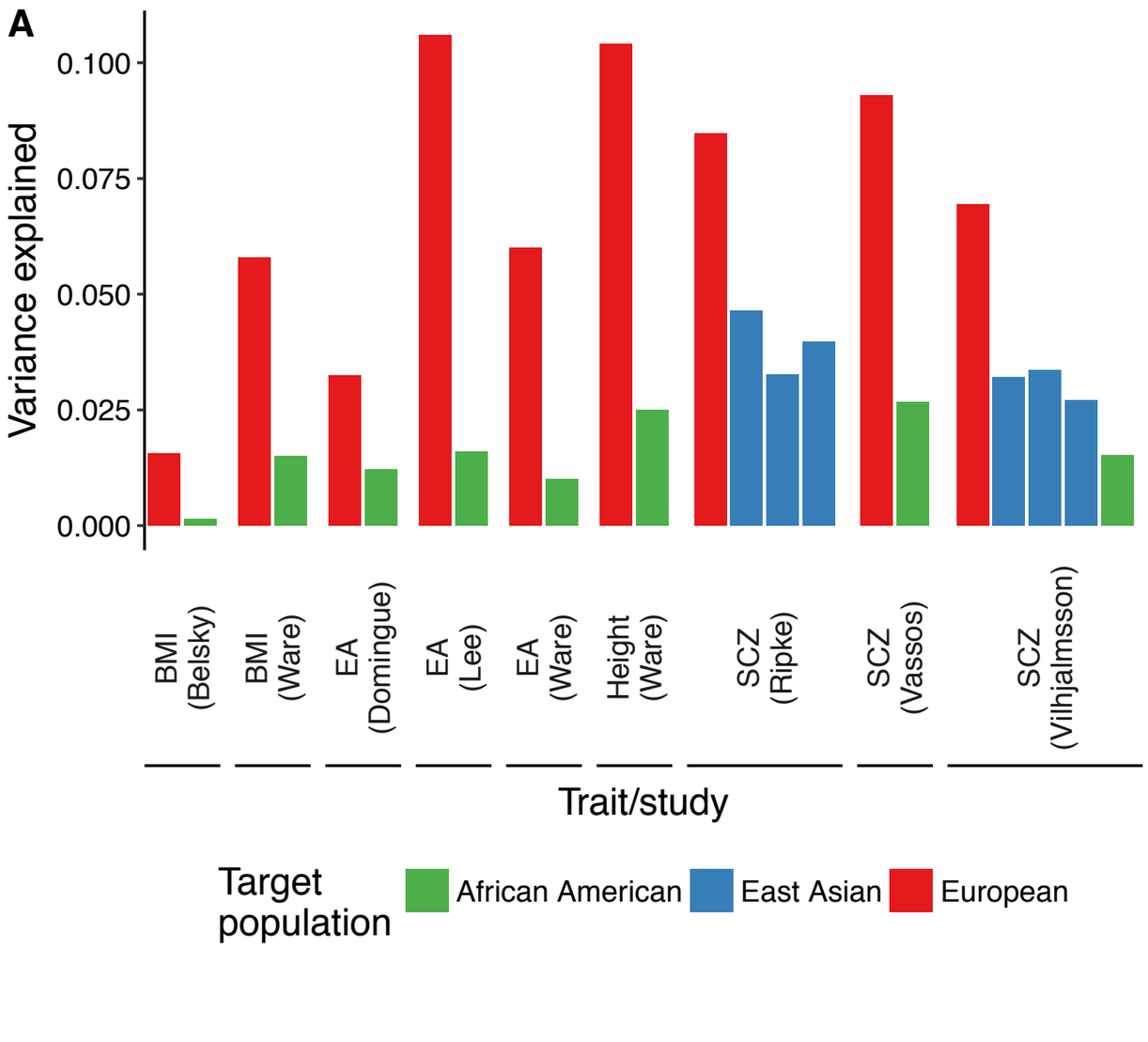


T2D



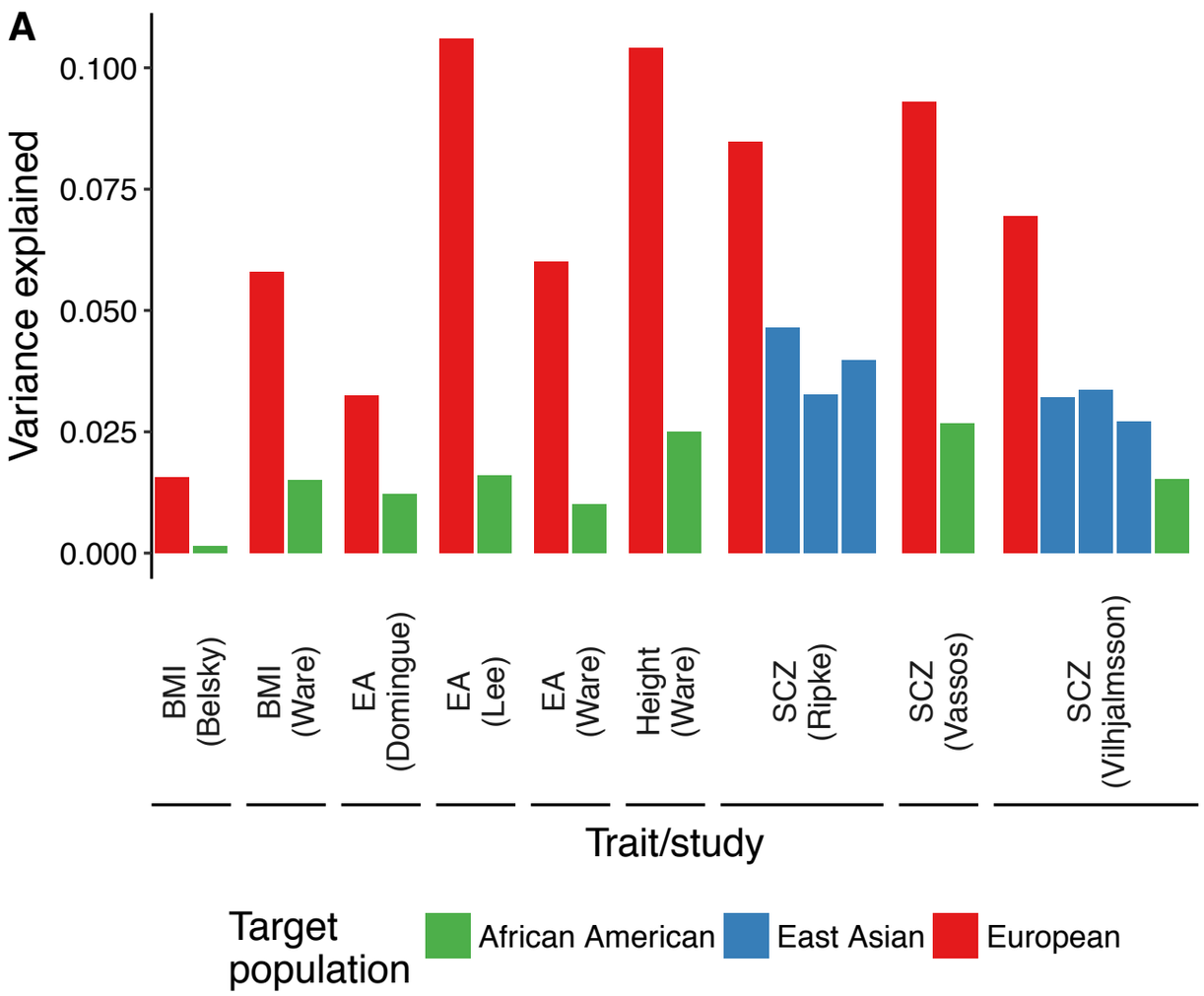
<u>For a given trait, impossible to predict *a priori* which population will have highest inferred risk!</u>

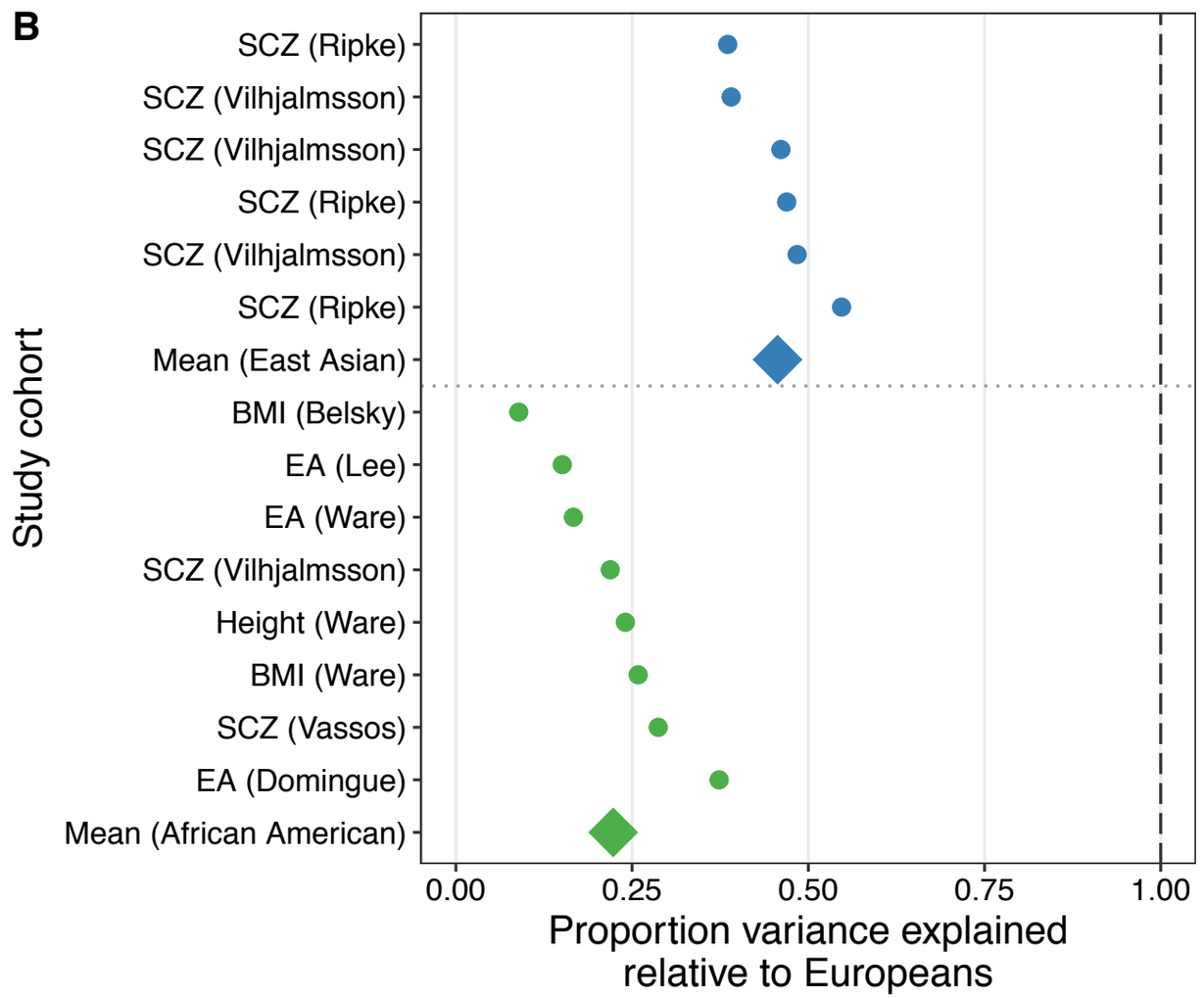
Staggering PRS disparities across populations



Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

Staggering PRS disparities across populations





Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

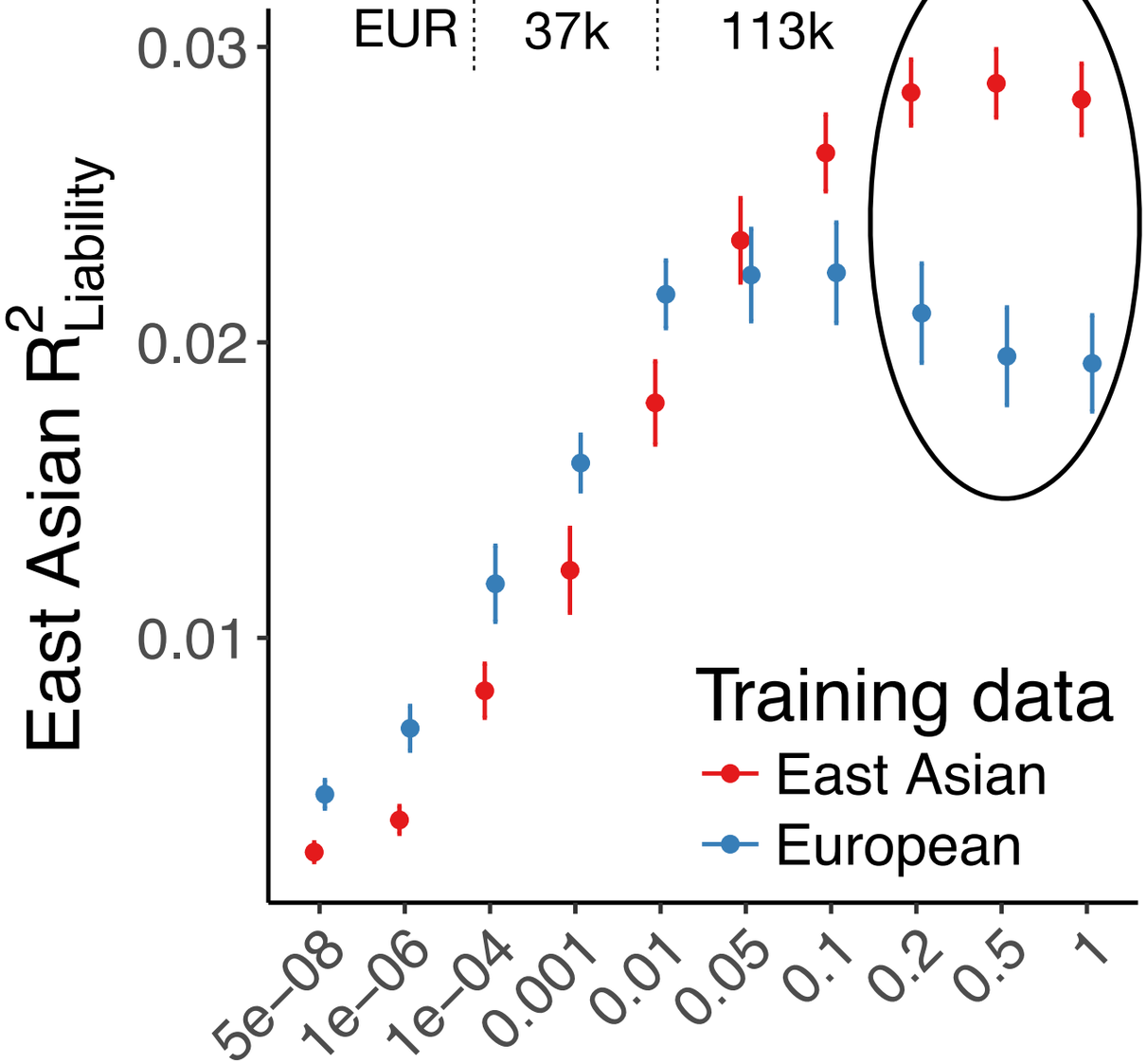
Consistent promise from diversifying efforts

Hailiang Huang

Chia-Yen Chen

Psychiatric
Genomics
Consortium

	Cases	Controls
EAS	13k	16k
EUR	37k	113k



Despite 3X larger sample sizes in Europeans, prediction in East Asians performs best with matched training data

Other examples:
BMI (Akiyama et al, 2018 Nat Gen)
SCZ (Li et al, 2017 Nat Gen)

Consistent promise from diversifying efforts

Masahiro Kanai

Goal: Compare PRS accuracy for 17 traits in UKBB and BBJ

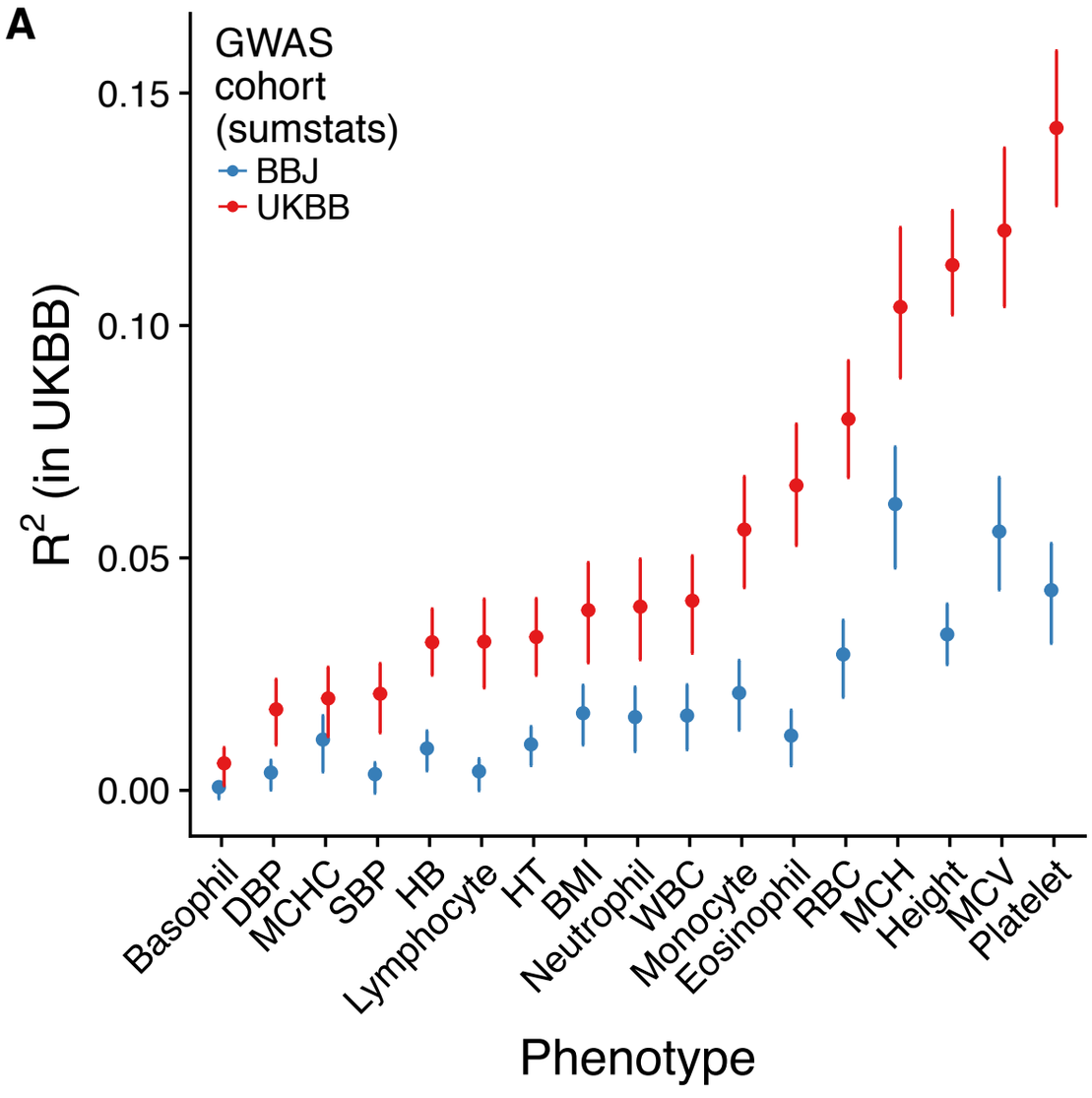
- Randomly set aside 5,000 individuals from each biobank
    - Match BBJ proportion with disease ascertainment
- Run GWAS on all other BBJ individuals. Match numbers in UKBB.

Do we see symmetric, comparable PRS accuracy?

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

Consistent promise from diversifying efforts

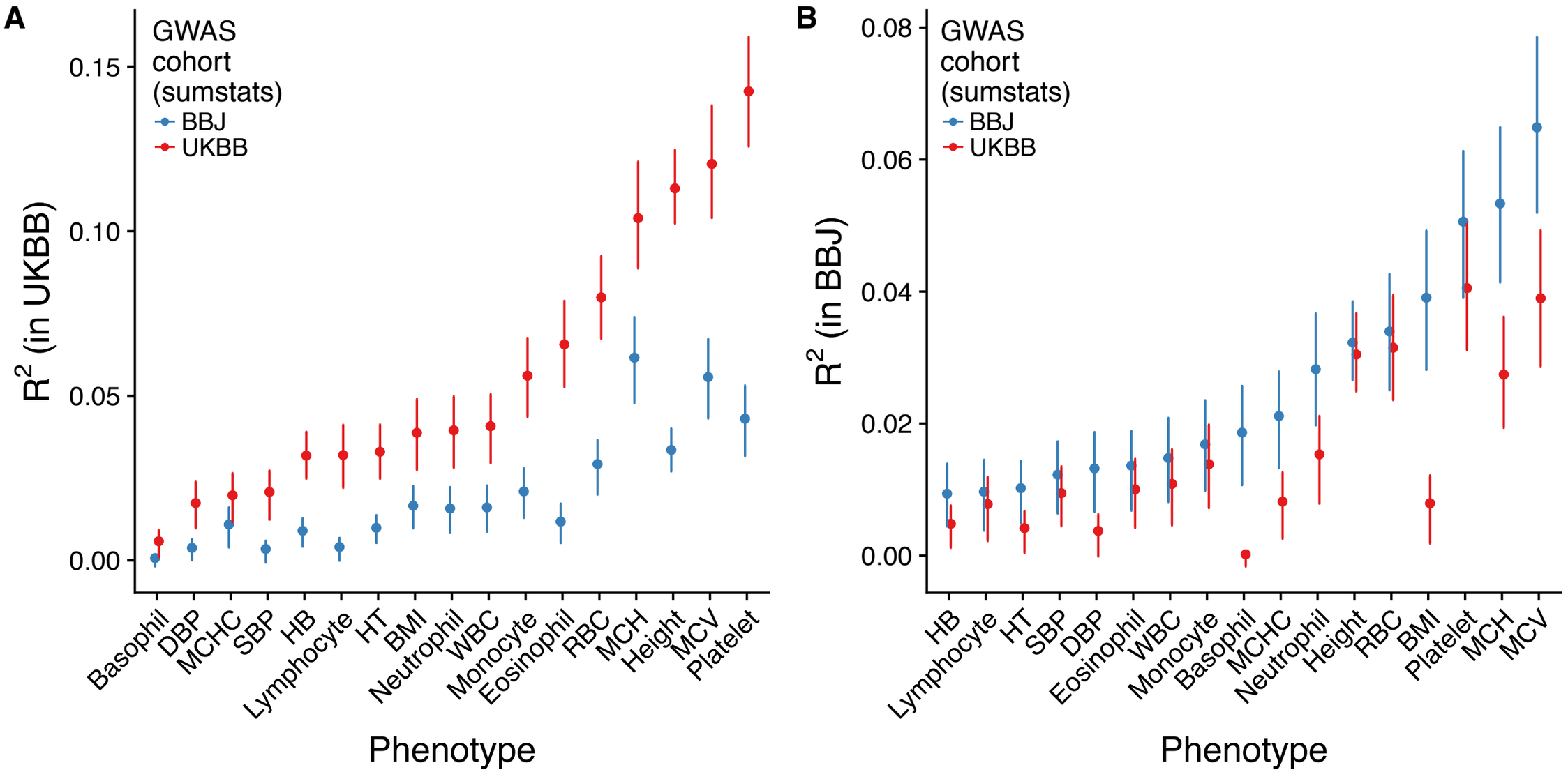
Masahiro Kanai



Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

Consistent promise from diversifying efforts

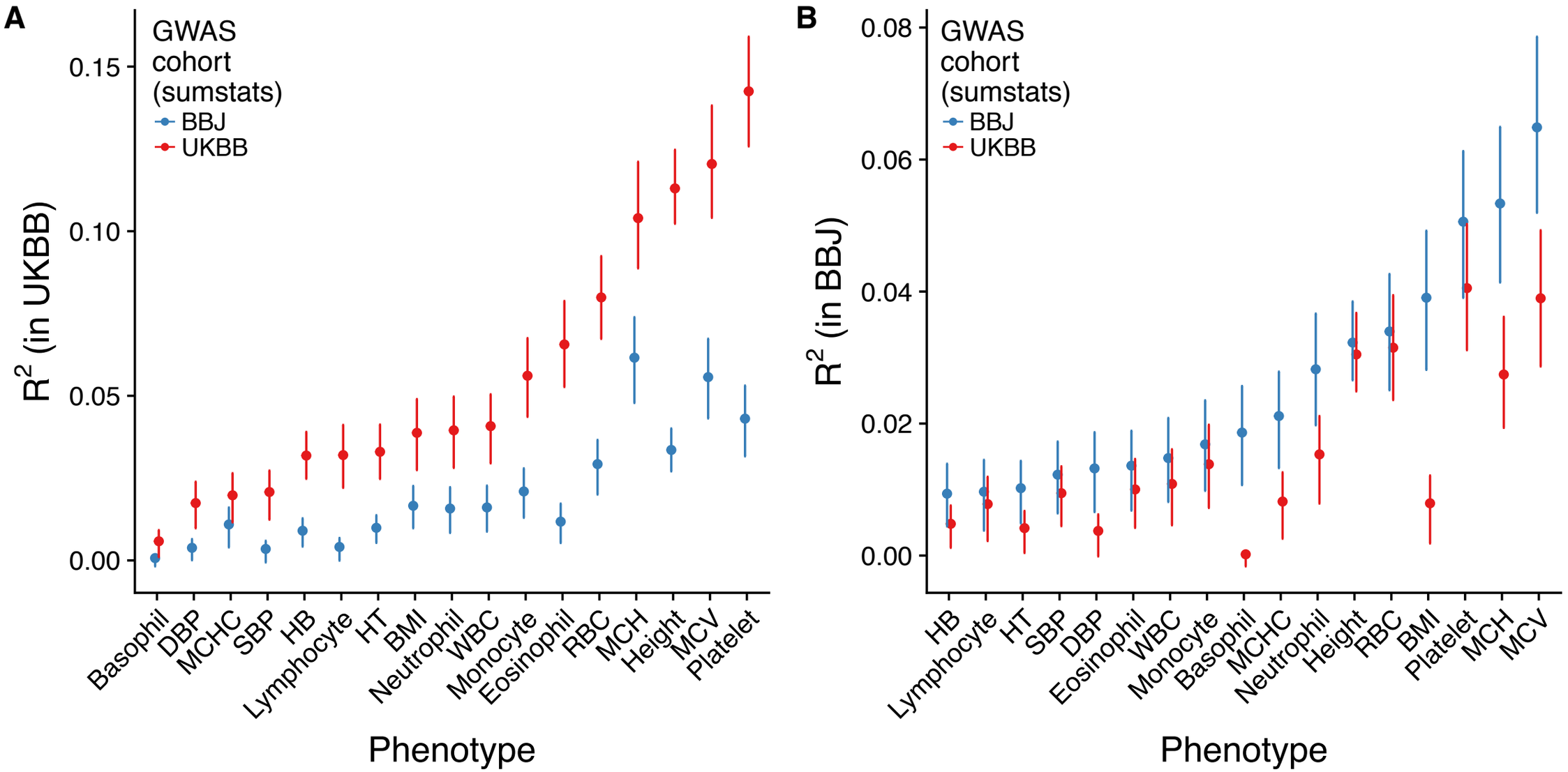
Masahiro Kanai



Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

Consistent promise from diversifying efforts

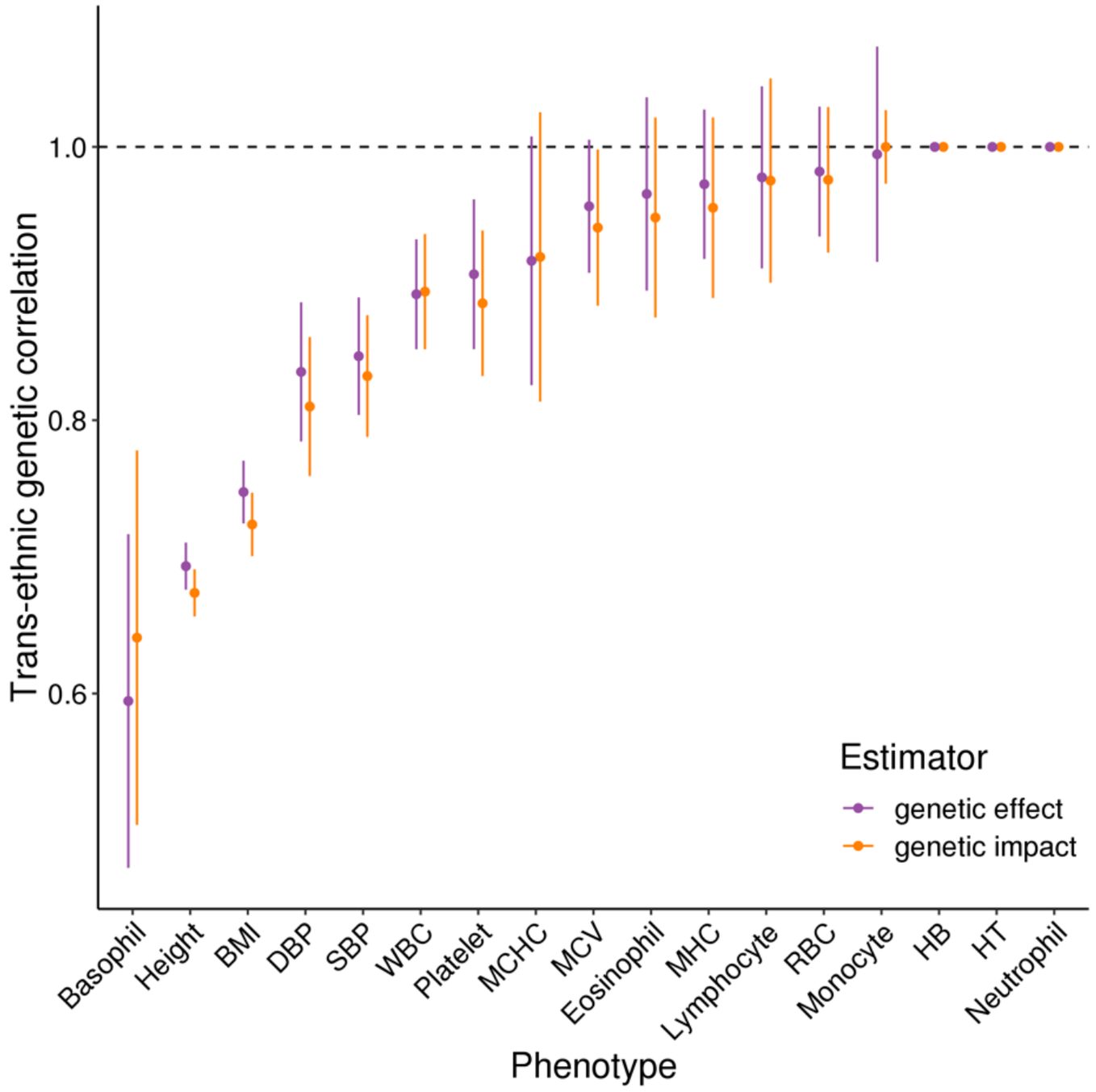
Masahiro Kanai



Note: differing axes... why?

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

Trans-ethnic genetic correlation is quite high



Cohort definition matters!

- UKBB has a “healthy volunteer” bias (healthier than average population)

- BBJ cohort is ascertained for 47 diseases (sicker than average population)

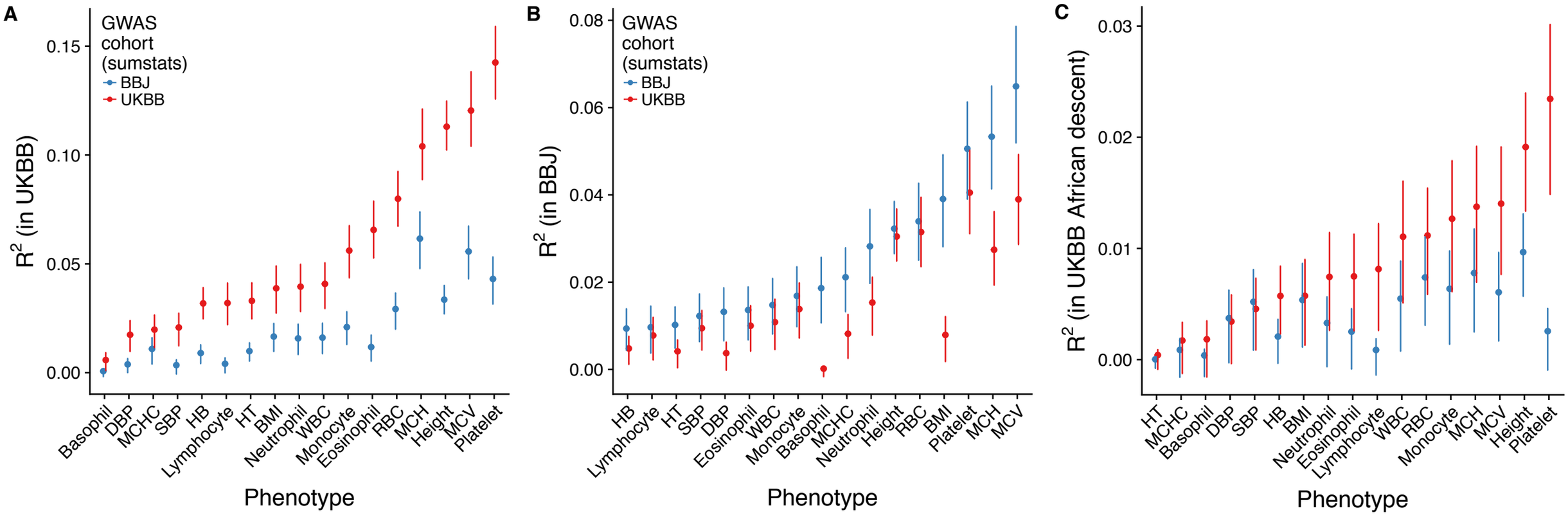
- Manually transcribe patients’ data from medical records in each hospital, read through and re-enter into BBJ’s electronic database

Cohort definition matters!

Trait	Observed h^2 (BBJ)	Observed h^2 (UKBB)	SE (BBJ)	SE (UKBB)
Basophil	0.0441	0.0213	0.0121	0.0050
BMI	0.1361	0.1955	0.0087	0.0090
DBP	0.0430	0.0984	0.0051	0.0068
Eosinophil	0.0586	0.1354	0.0093	0.0167
Hb	0.0452	0.1054	0.0053	0.0107
Height	0.3059	0.3675	0.0187	0.0208
Ht	0.0457	0.0942	0.0056	0.0093
Lymphocyte	0.0516	0.1318	0.0073	0.0118
MCH	0.1309	0.1942	0.0184	0.0210
MCHC	0.0481	0.0402	0.0080	0.0052
MCV	0.1447	0.1994	0.0178	0.0201
Monocyte	0.0448	0.1331	0.0090	0.0177
Neutrophil	0.0758	0.1153	0.0097	0.0131
Platelet	0.1260	0.2012	0.0148	0.0179
RBC	0.0818	0.1586	0.0093	0.0141
SBP	0.0574	0.1041	0.0063	0.0070
WBC	0.0778	0.1286	0.0074	0.0114

...but a lot of room for growth

Masahiro Kanai

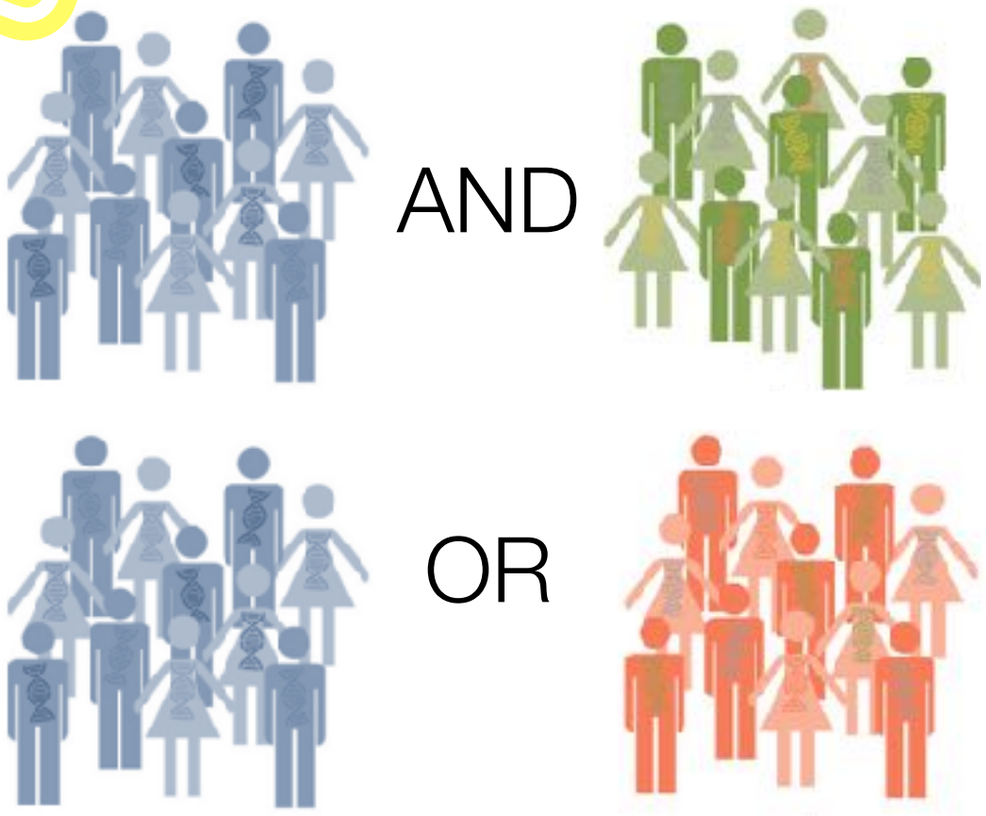
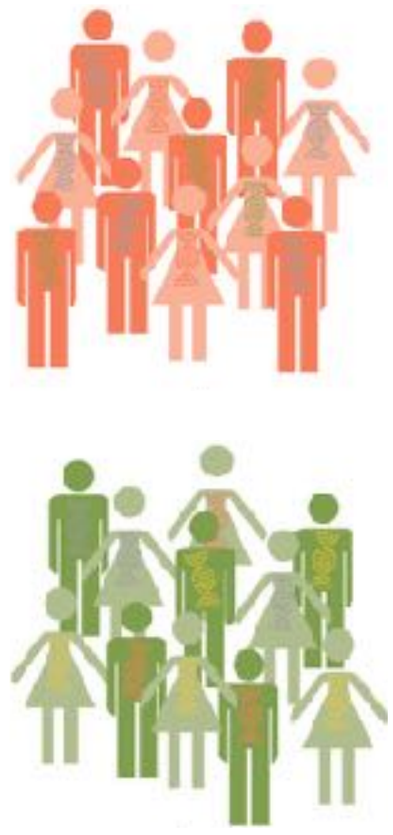
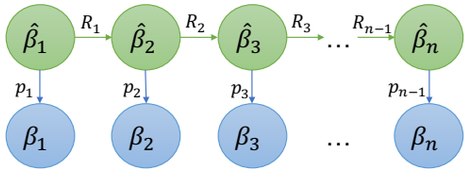
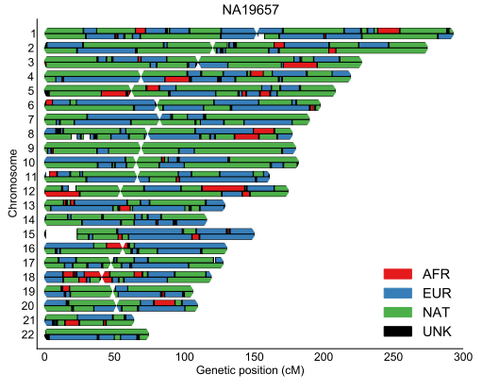


Note: differing axes

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)

New statistical approaches for genetic prediction

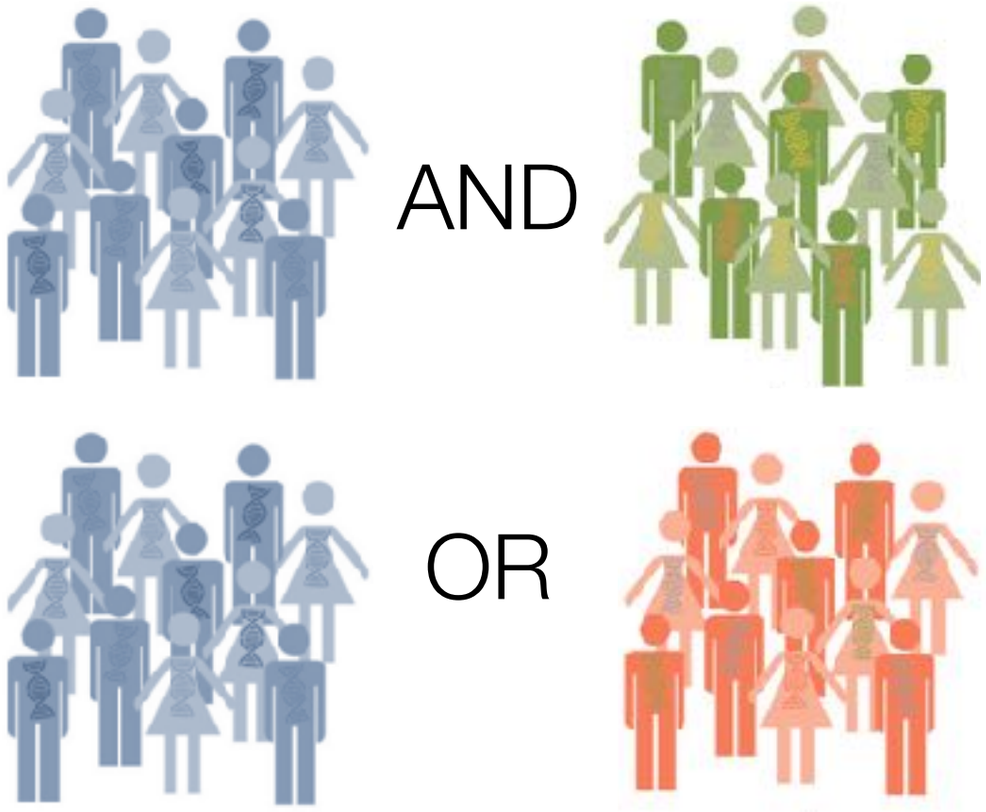
Under construction

Study	Multi-population	Single population mismatch	Recently admixed population
Target			
Approach	**Multi-ancestry meta-analysis (MAMA)**	**Kalman filter** 	**Personalized LD Panel** 

Genetic prediction with GWAS from multiple populations

Study

Multi-population



Approach **Target**

Multi-ancestry meta-analysis (MAMA)

- **Approach**: Consider cross-population LD to recalibrate effect sizes in each population
- **Related methods**: LD score regression, MTAG
- **Status**: Implementing across global biobanks

Patrick Turley

Hui Li

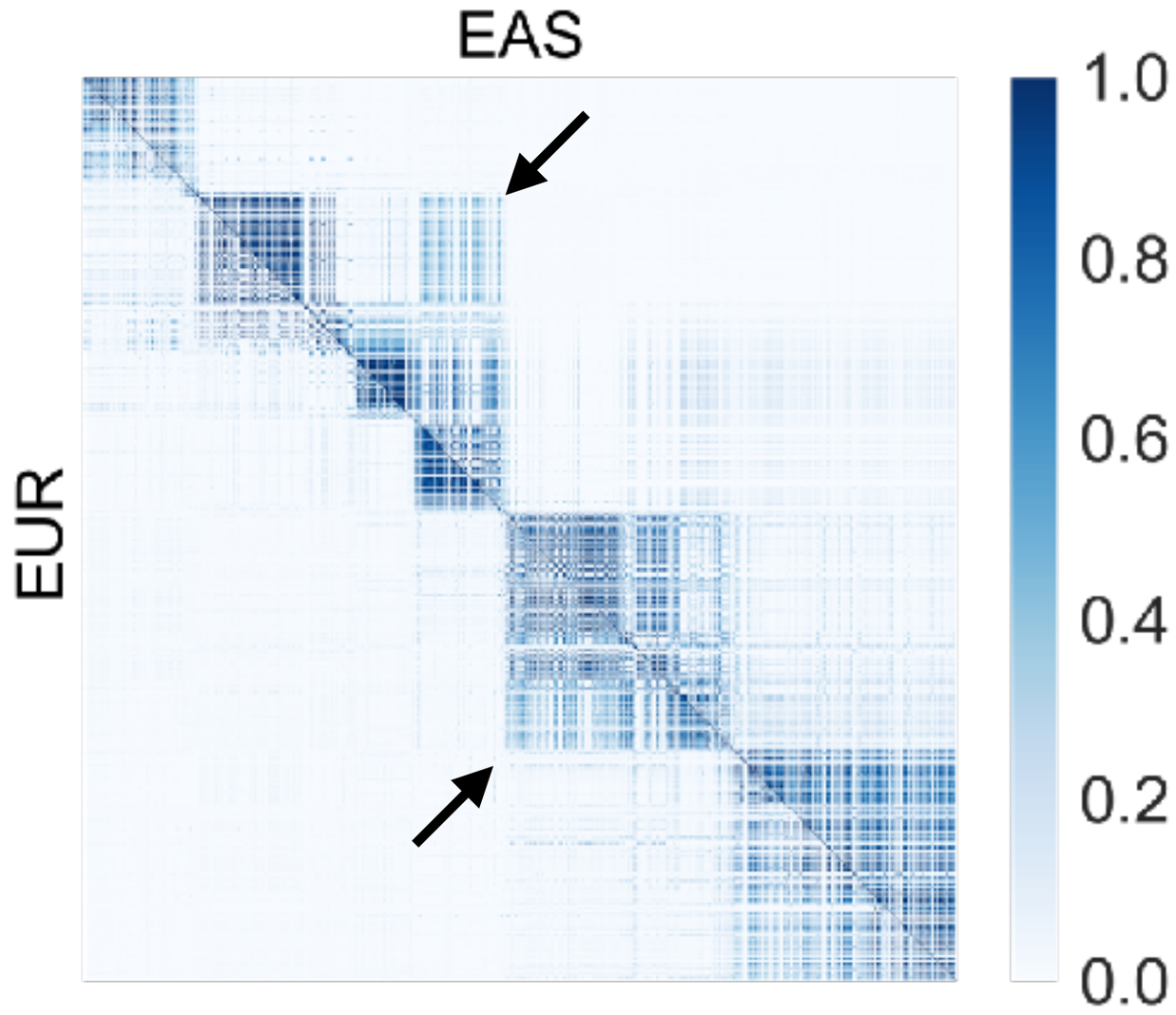
Raymond Walters

GWAS stats differ across populations due to LD

$$\hat{\beta}_j = \sum_{k=1}^{M} r_{j,k} b + \epsilon$$

- Causal effect sizes tend to be the same...
- ... but effect size **estimates** vary with LD differences across populations

$$\hat{\beta}_{A,j} \not\equiv \hat{\beta}_{B,j}$$



Key elements of MAMA

Ω_j

- Variance-covariance of genetic component

- More information shared when LD patterns and conditional effects are similar

Σ_j

- Variance-covariance of error and biases

- Less information shared when estimates are noisy or biased

Applications in real data

- **Psychiatric disorders**

Phenotype	Population/ Location	N Cases	N Controls	Source
Schizophrenia	Europe	34,989	113,075	PGC
Schizophrenia	East Asia	13,305	16,244	PGC
Schizophrenia	African Americans	6,981	2,564	PGC
Bipolar/Schizophrenia	Hispanic/Latinos	3,982	4,553	PGC
Schizophrenia	Africa	~18,000	~18,000	NeuroGAP
PTSD	U.S. minorities	21,845	58,769	PGC/CVB

- **Anthropometric traits (height, BMI, blood panels, etc)**

Biobank	Code	Sample sizes
UK Biobank	UKBB	~500k
Finnish biobank	Finrisk	~50k
BioBank Japan Project	BBJ	~162k
China Kadoorie Biobank	CKB	~100k
PAGE (US minorities)	PAGE	~50k

Lots of nice resources!

Some nice reviews:

- Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.
- Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet.
- Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14, 507–515.

Coming soon:

- Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
- Martin, Daly, Robinson, Hyman, & Neale. Predicting polygenic risk of psychiatric disorders (in revision)

Conclusions

- Polygenic risk scores have the potential to improve clinical models, but are currently likely to **increase health disparities** due to Eurocentric GWAS biases

- We need more **diverse GWAS studies** and **new methods** to address these major issues

- We are developing new methods that can use biobank-scale data from diverse populations to improve the generalizability of genetic prediction across populations

Future directions

- How will we use PRS in the future?
    - Biomarker for: behavioral interventions? differential diagnosis? personalized drug therapies? reducing cost of clinical trials?
- Tricky issues to resolve:
    - Pleiotropy
    - Healthcare economics: $ and life disparities by ethnicity?

Acknowledgments

MGH/Broad
- Mark Daly
- Ben Neale
- Patrick Turley
- Raymond Walters
- Duncan Palmer
- Hailiang Huang
- Chia-Yen Chen
- Masahiro Kanai
- Elizabeth Atkinson
- Caitlin Carey
- Alex Bloemendal
- Karestan Koenen
- Giulio Genovese
- Elise Robinson
- Steve Hyman

Stanford
- Carlos Bustamante
- Genevieve Wojcik

Mt. Sinai
- Eimear Kenny

UC Denver
- Chris Gignoux

McGill
- Simon Gravel

FIMM
- Samuli Ripatti
- Juuka Koskela
- Juulia Partanen
- Sini Kerminen
- Matti Pirinen
- Aarno Palotie

RIKEN
- Y. Okada
- Yoichiro Kamatani

Consortia

1000 Genomes Project

NeuroGAP-Psychosis

PGC-SCZ
- Max Lam

SSGAC
- Hui Li
- Dan Benjamin
- David Cesarini
- Meghan Zacher

Funding
- K99MH117229

Questions/comments?

armartin@broadinstitute.org

@genetisaur