Session 7: GWAS, fine-mapping and PRS in diverse-genetic-ancestry and admixed samples

DCEG Statistical Genetics Workshop

David V. Conti, Ph.D.

University of Southern California

Topics for Workshop

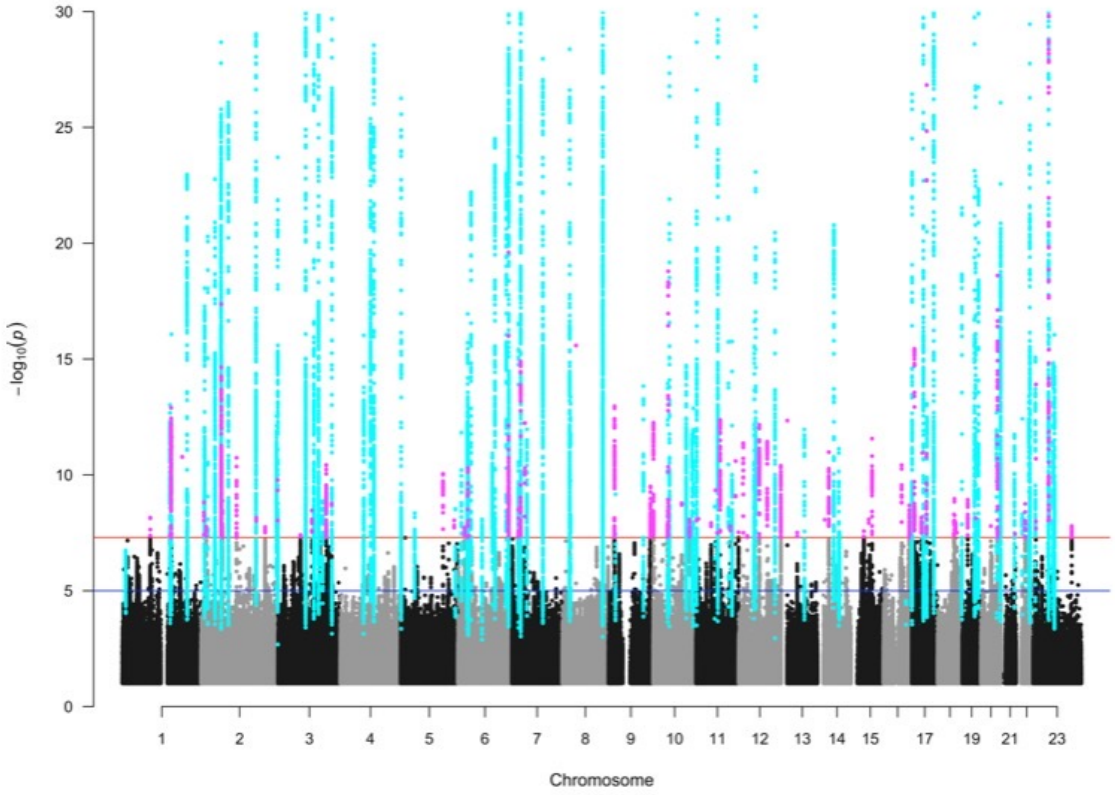
- Population-specific and trans-ancestry GWAS
- Multi-population fine-mapping
- PRS construction
- PRS evaluation across populations
- Absolute risk across populations

Goals of Genetic Association Studies

- **Identify the causal variants for a given disease.**

- **Discover** regions of the genome harboring risk loci.
- **Rank** SNPs as putative causal variants via fine-mapping.
- **Characterize** the impact of these risk loci:
  - Impact on familial relative risk (FRR) and heritability.
  - Biological impact.
- **Translation/Prediction**: using genetic risk scores for screening and individual risk prediction.

GWAS


30
25
20
15
10
5
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
23
Chromosome

Genetic Risk Score (GRS):

Weighted sum of # risk alleles carried by each participant

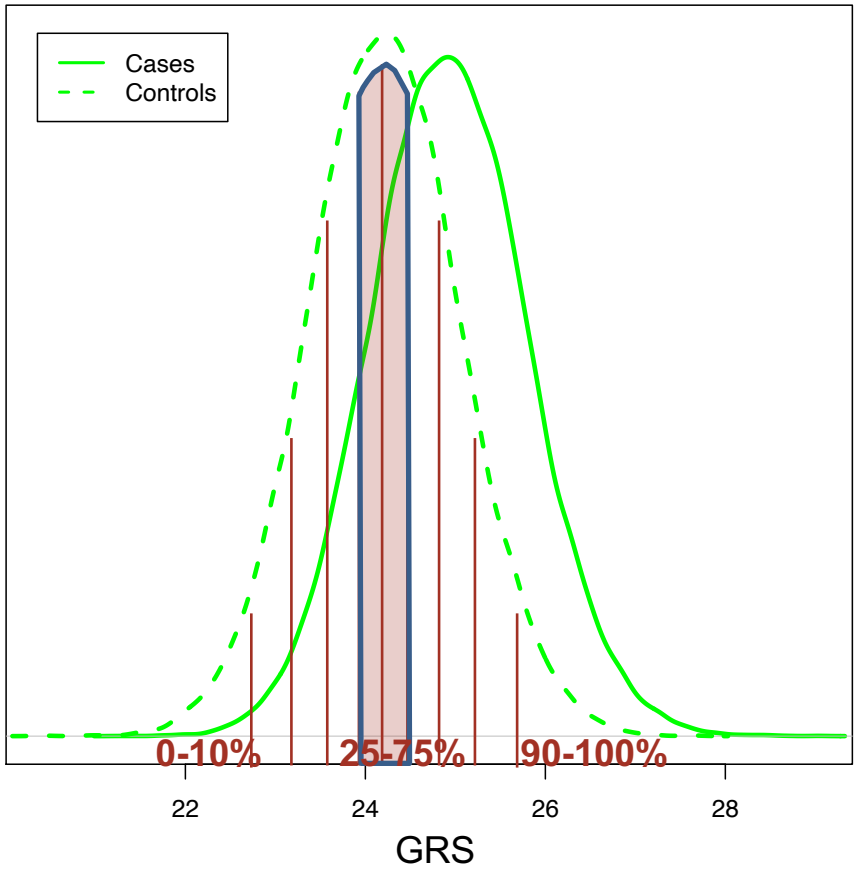
Count of risk alleles for variant *m* for individual *i*

What SNPs?

$$GRS_i = \sum_{m=1}^{M} w_m G_{im}$$

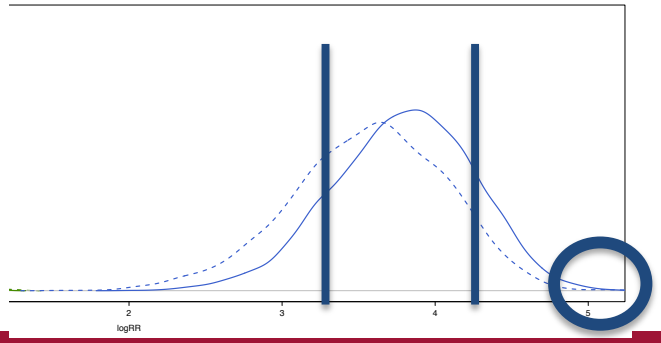
What weight?

How to evaluate?



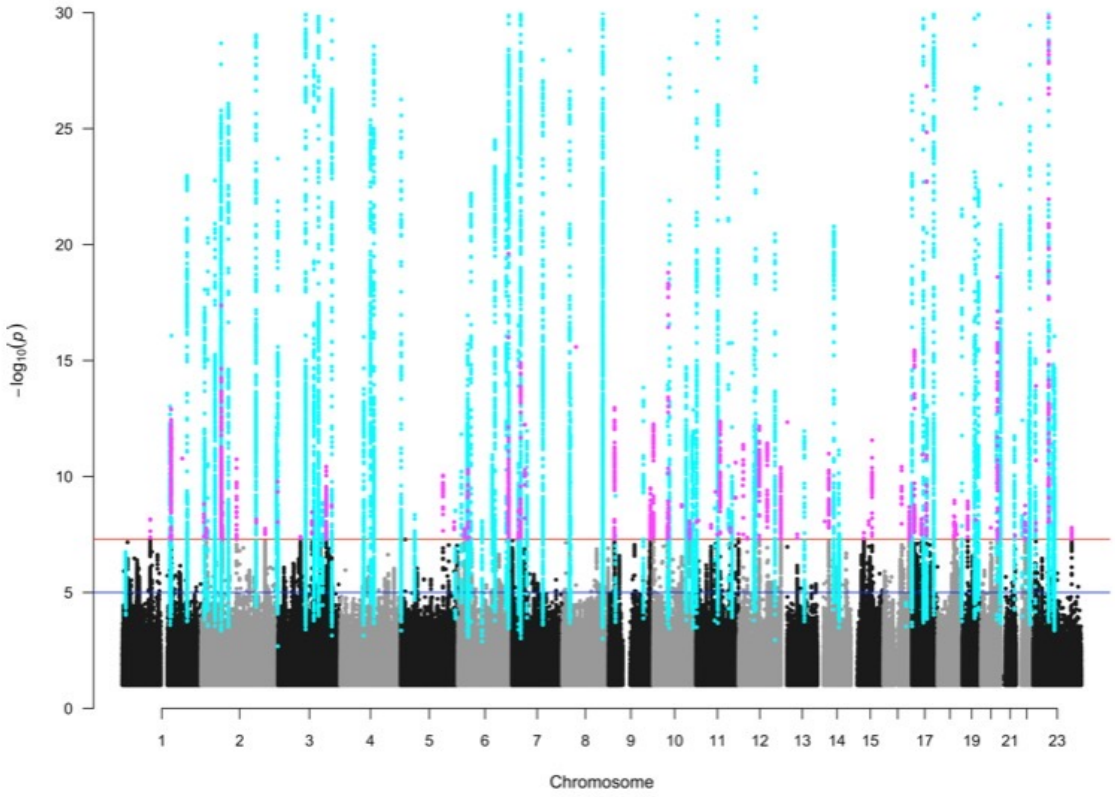
ELLIPSE: EUR Prostate Cancer

Risk Categories	OR (95%CI)
<1%	0.19 (0.13-0.27)
1-10%	0.31 (0.28-0.35)
10-25%	0.52 (0.48-0.55)
25-75%	-
75-90%	1.78 (1.68-1.88)
90-99%	2.93 (2.75-3.12)
≥99%	5.65 (4.83-6.62)



- Previous GWAS identified over 100 known variants.
- Discovery with 79,194 cases/61,112 controls (EUR) in over 50 studies:
  - **Identified 62 novel variants.**
- All loci capture **28.4%** of the FRR (~2.0).
- GRS identifies over **10%** of the population with an odds ratio that exceeds 2.0.

GWAS


30
25
20
15
10
5
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
23
Chromosome

Meta-Analysis

- What studies to include?
  - Ideally, one would include as many studies as possible to increase the sample size.
  - But the effect size of the association may vary between studies.
  - This is referred to as between-study heterogeneity.
  - Sources of heterogeneity:
    - LD differences between populations in different studies.
    - Environmental factors different between studies.

Meta-Analysis Notation

- X_1 , …, X_C are the effect-size estimates (e.g. log odds ratios) in *C* studies.
- $SE(X_1)$ = standard error of the estimate
- $V_i = SE(X_1)^2$ and $W_i = V_i^{-1}$ is the inverse-variance estimate.
- τ^2 is the between-study variance.

Likelihood of Fixed-Effects

- Null:

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{X_i^2}{2V_i}\right)$$

- Alternative:

$$L_1 = \prod_i \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{(X_i - \mu)^2}{2V_i}\right)$$

Likelihood of Random-Effects

- Null:

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi(V_i + \hat{\tau}^2)}} \exp\left(-\frac{X_i^2}{2(V_i + \hat{\tau}^2)}\right)$$

Heterogeneity is assumed in the null.
Why?

- Alternative:

$$L_1 = \prod_i \frac{1}{\sqrt{2\pi(V_i + \hat{\tau}^2)_i}} \exp\left(-\frac{(X_i - \mu)^2}{2(V_i + \hat{\tau}^2)}\right)$$

New RE Test (Han and Eskin 2011)

- Null:

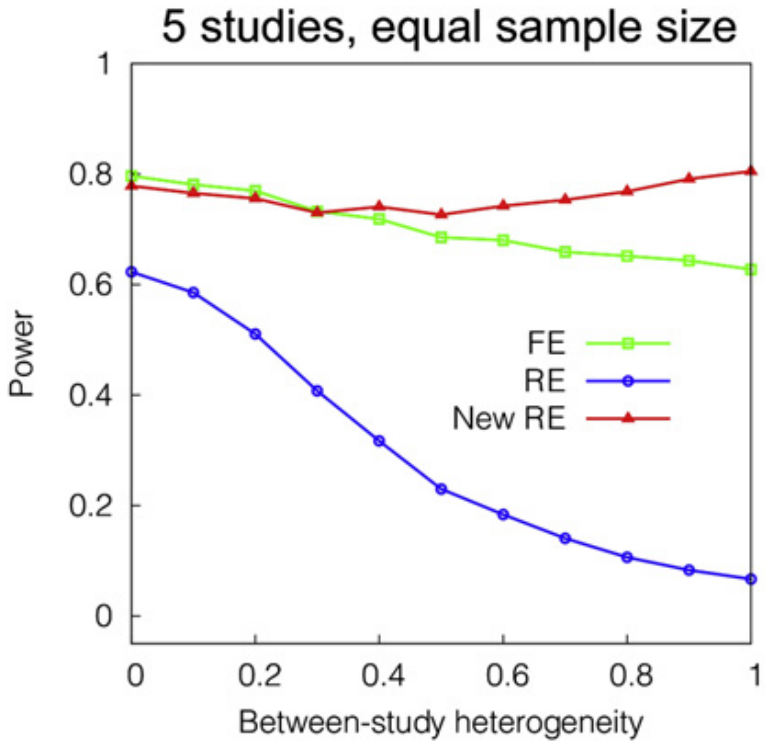
$$L_0 = \prod_i \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{X_i^2}{2V_i}\right)$$

No heterogeneity in the null.
Is this always appropriate?

- Alternative:

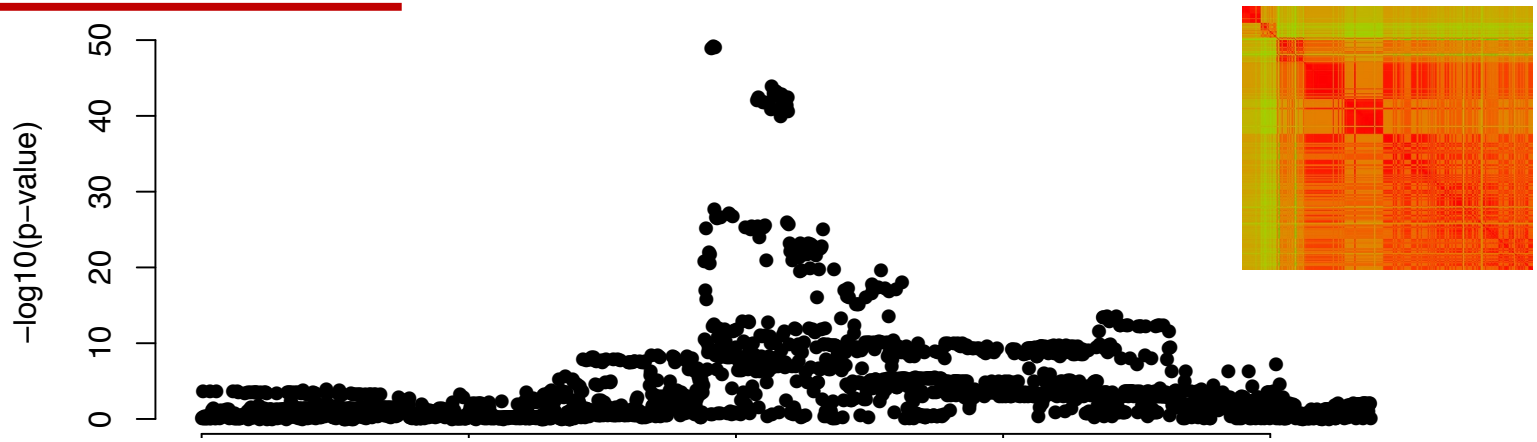
$$L_1 = \prod_i \frac{1}{\sqrt{2\pi(V_i + \hat{\tau}^2)_i}} \exp\left(-\frac{(X_i - \mu)^2}{2(V_i + \hat{\tau}^2)}\right)$$

New RE Test



Han B, Eskin E. *Am J Hum Genet*. 2011;88:586–598.

Challenge of Linkage Disequilibrium



1. Thresholding and Pruning (T + P):
2. Fine-mapping to identify putative 'causal' variants:
   - Requires selection with a conditional model.
   - What if you only have access to summary statistics?

Prostate Cancer Fine-Mapping

- PRACTICAL/ELLIPSE consortium
- **54 studies:** 82,591 cases; 61,123 controls (European ancestry)
- Individual-level data available on some, but not all studies.
- **~70 million SNPs** available.

Joint Analysis of Marginal SNP Effects

- Association of trait (Y) with Genotype (G) using individual-level data:
- Estimation of joint effects in linear regression:
- Substitute reference data and summary statistics:

$$y \sim N(G\beta, \sigma^2 I_N)$$

$$\hat{\beta} = (\boldsymbol{G'G})^{-1}\boldsymbol{G'}y$$

$$z := G'y$$

$$\hat{\beta} = (\boldsymbol{G_R}'\boldsymbol{G_R})^{-1}\boldsymbol{z}$$

Joint Analysis of Marginal SNP Effects

- Uses marginal summary statistics (e.g. log odds ratios) to perform conditional regression analysis.
- Incorporates external reference data for correlation structure (i.e. LD between SNPs).
- Via Bayesian selection yields posterior inclusion probability (PIP) for each SNP and inference for the number of signals within a region.

Bayesian Model Selection

- Marginal model likelihood:

JAM normal | G-prior | Inverse-Gamma

$$p(z_L^* | M_\gamma) = \int p(z_L^* | \beta_\gamma, \sigma^2, M_\gamma) \cdot p(\beta_\gamma | \sigma^2, M_\gamma) \cdot p(\sigma^2 | M_\gamma) d\beta_\gamma d\sigma^2$$

Analytical

- G-prior:

$$p(\boldsymbol{\beta_\gamma} | \sigma^2, M_\gamma) = MVN(0, \sigma^2 \tau (\boldsymbol{L_\gamma' L_\gamma})^{-1})$$

$$(\boldsymbol{L_\gamma' L_\gamma})^{-1} = (\boldsymbol{G_\gamma' G_\gamma})^{-1}$$

Prior structure supports SNP effects inversely proportional to their co-variances in the observed genotype matrix *G*.

Bayesian Model Selection

- Marginal model likelihood:

JAM normal | G-prior | Inverse-Gamma

$$p(z_L^* | M_\gamma) = \int p(z_L^* | \beta_\gamma, \sigma^2, M_\gamma) \cdot p(\beta_\gamma | \sigma^2, M_\gamma) \cdot p(\sigma^2 | M_\gamma) d\beta_\gamma d\sigma^2$$

Analytical

- Stochastically search across models using reversible jump MCMC.

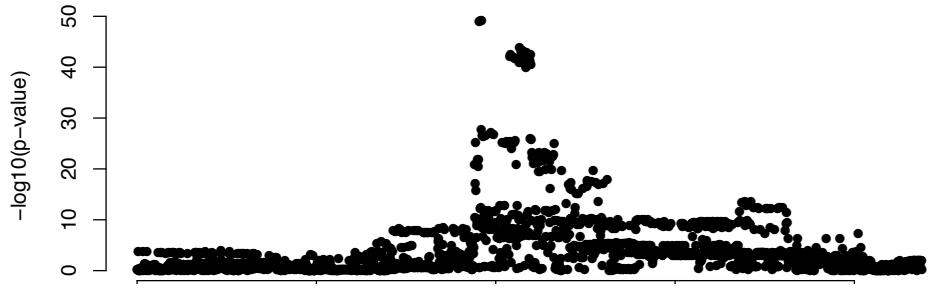
$$\gamma_m \sim Binomial(\omega)$$
$$\omega \sim Beta(a, b)$$

Prostate Cancer Fine-Mapping

- Applied JAM to finemap **84** previously known regions (excludes 8q24 region) with over 40,000 SNPs.
  - **99** independent signals.
  - All 95% credible sets: **343** variants
  - Evidence for multiple signals in **12** regions (up to **5** signals)
- Impact:
  - For the 84 previous signals: replacement variants increased proportion of FRR explained from **23.2% to 26.5%.**
  - Including all **99** independent signals: **30.3%** of FRR explained.

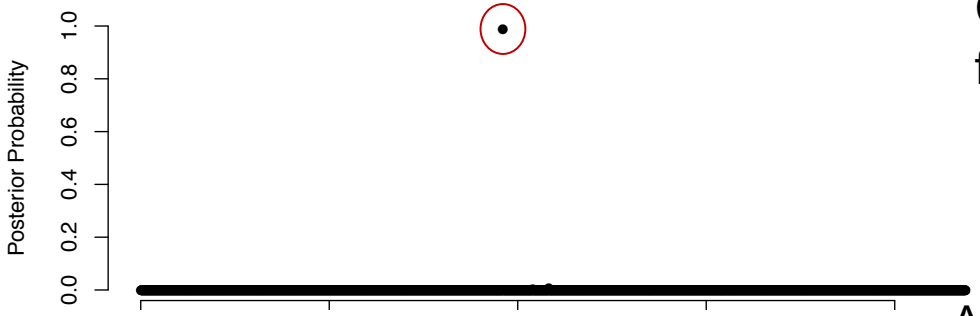
TET2 (Chr 4): Single signal

Meta-analysis one-at-a-time p-values



438 variants

JAM posterior probabilities

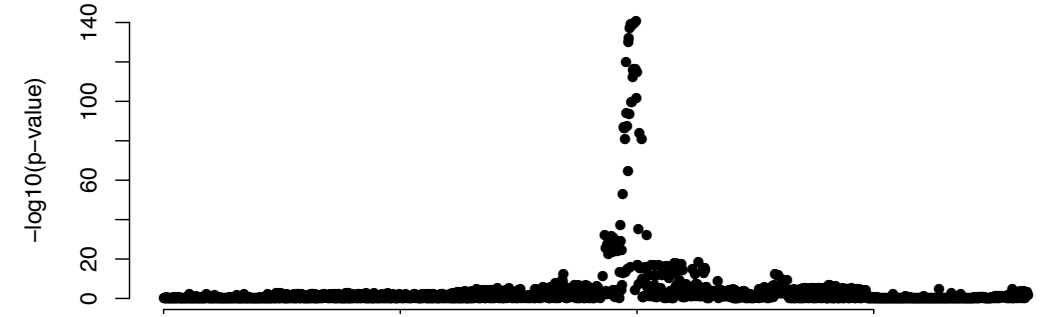


Clear evidence for **1 signal**

Analysis by Tokhir Dadaev

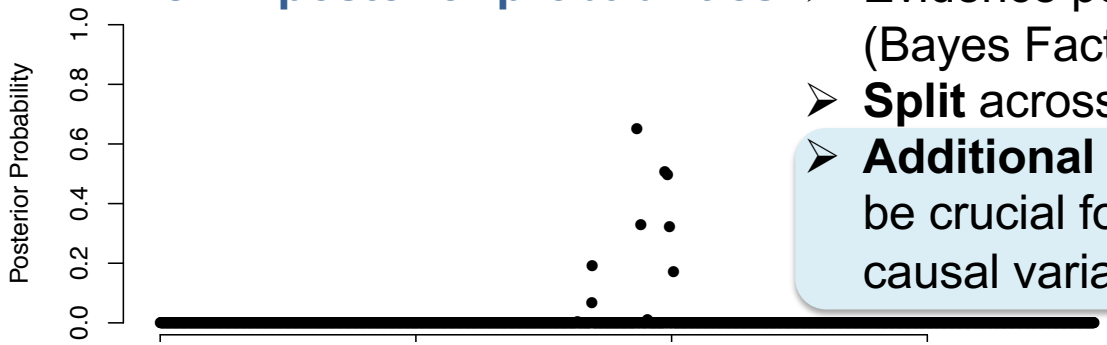
HNF1B (Chr 17): Multiple signals

Meta-analysis one-at-a-time p-values



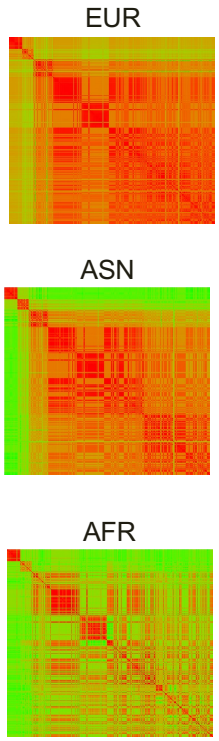
444 variants

JAM posterior probabilities



- Evidence points to **4 signals** (Bayes Factor >10)
- **Split** across multiple SNPs
- **Additional information** will be crucial for identifying causal variant(s)

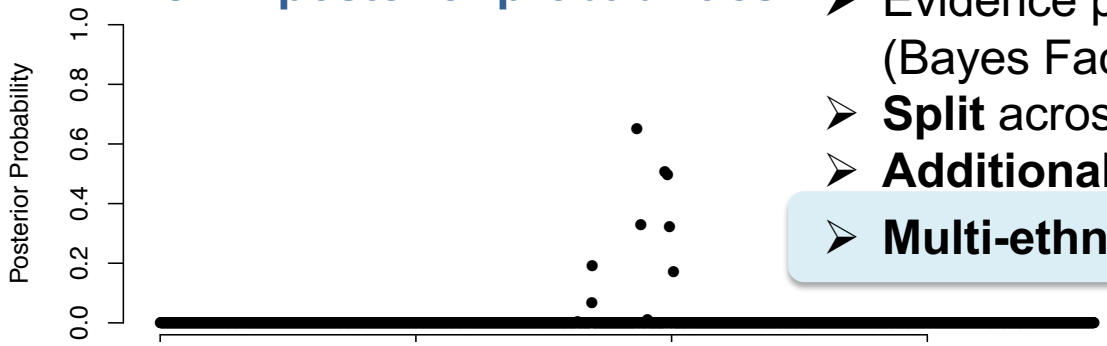
HNF1B (Chr 17): Multiple signals



Meta-analysis one-at-a-time p-values

444 variants

JAM posterior probabilities



- Evidence points to **4 signals** (Bayes Factor >10)
- **Split** across multiple SNPs
- **Additional Information**
- **Multi-ethnic** fine-mapping

Multiethnic Fine-Mapping

- Localization success rate (LSR)

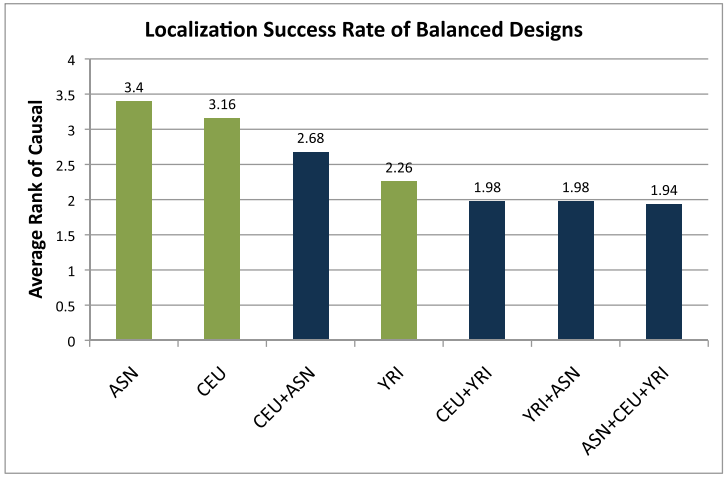


Figure 1. The Average Rank of the Causal Variant in 10,000 Simulated Loci, with 3000 Cases, 3000 Controls, and $\gamma = 1.4$ for Seven Different Study Designs

Designs over multiple populations, such as the CEU+YRI, split individuals evenly among them. Using multiple populations reduces the number of functional assays expected before the causal variant is identified.

Multiethnic Fine-Mapping

- Localization success rate (LSR)

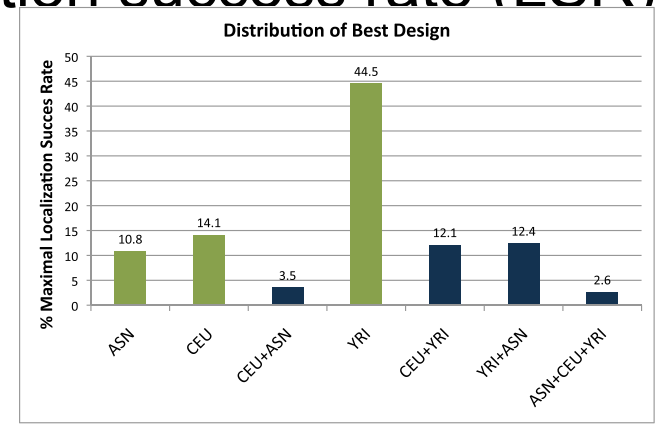
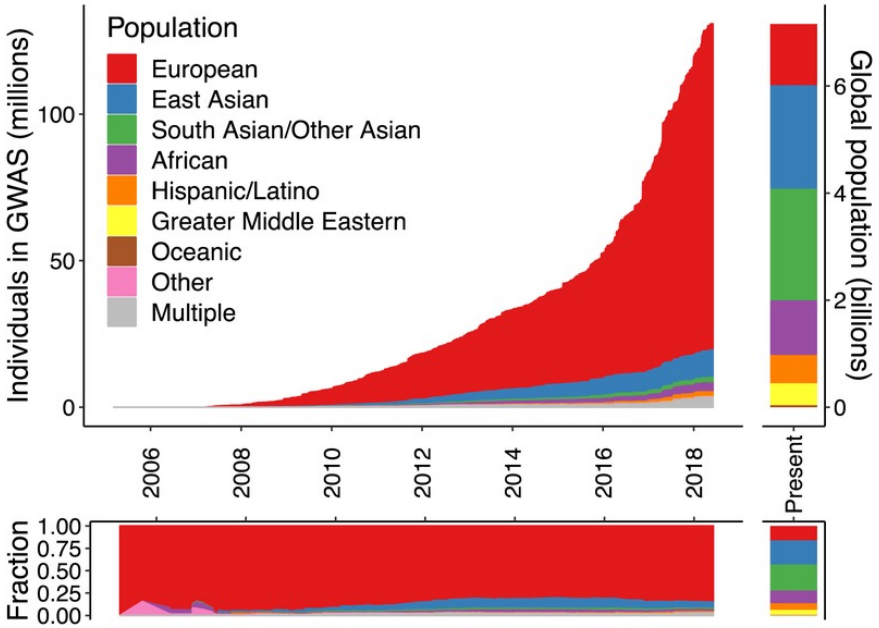
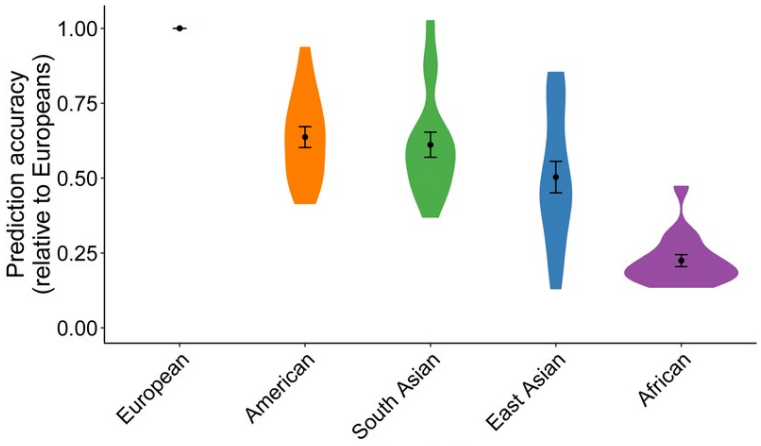


Figure 2. The Fraction of Times that a Design Achieves the Maximal LSR for Each of the Study Designs
The statistics are based on 10,000 simulated loci, with 3000 cases, 3000 controls, and $\gamma = 1.4$. As expected, the YRI population is most often the best choice for study design. However, it is the top choice only 44% of the time. The combination of all three populations is almost never the best study design, accounting for only 2.6% of the 10,000 designs. Interestingly, it maximizes the average LSR, suggesting, first, that it protects against the variance of different local LD structures and, second, that tailoring study designs to the loci in the follow-up study is beneficial.

Lack of Diversity in GWAS Contributes to Health Disparities

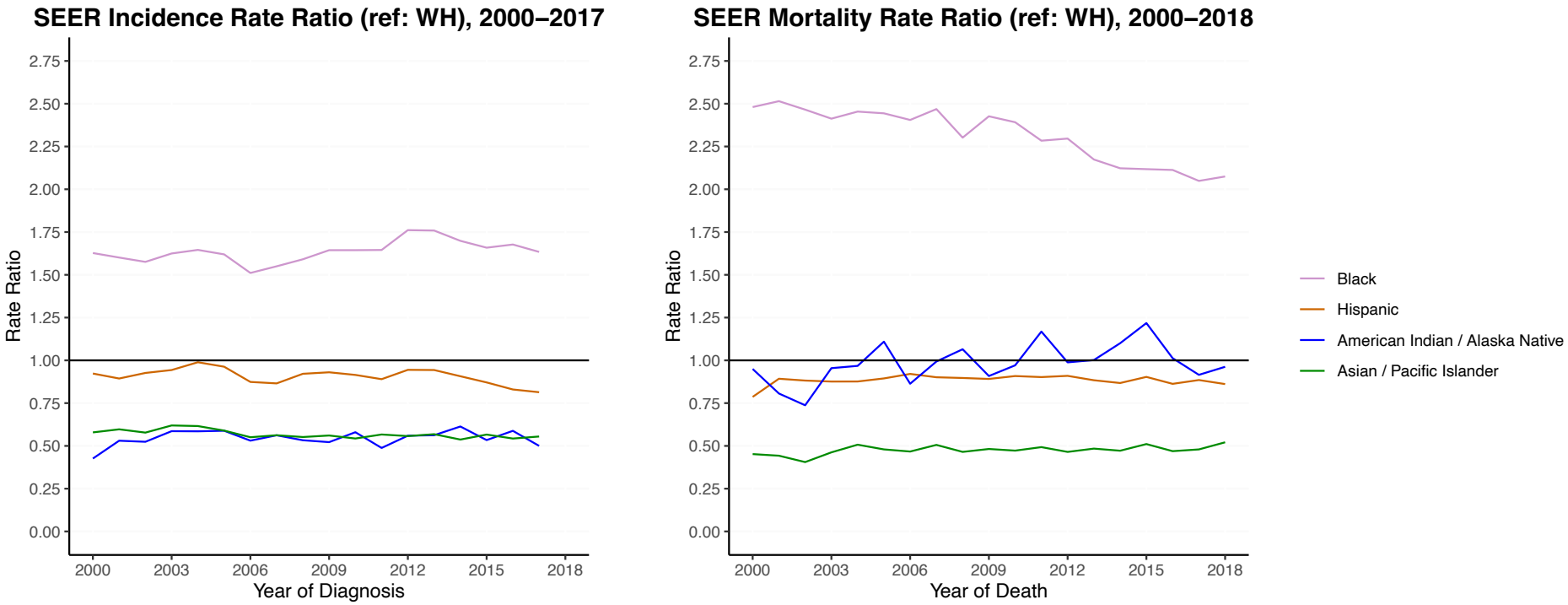


Ancestry of GWAS participants over time relative to the global population



Polygenic prediction accuracy relative to European ancestry individuals across 17 quantitative traits

Prostate Cancer Differences by Race/Ethnicity



Population-specific GWAS Discovery

Risk Categories	European Ancestry OR (95%CI)	African Ancestry OR (95%CI)
<1%	0.19 (0.13-0.27)	0.18 (0.10-0.33)
1-10%	0.31 (0.28-0.35)	0.36 (0.30-0.43)
10-25%	0.52 (0.48-0.55)	0.55 (0.48-0.63)
25-75%	-	-
75-90%	1.78 (1.68-1.88)	1.68 (1.47-1.92)
90-99%	2.93 (2.75-3.12)	3.02 (2.52-3.63)
≥99%	5.65 (4.83-6.62)	4.23 (2.39-7.50)

- Prostate Cancer incidence is **1.6-fold** higher in **African Americans**
- Discovery: 10,202 cases/10,810 controls (African ancestry).
- Of the 100 reported known risk loci:
  - 94 variants are polymorphic (MAF>0.05).
  - 81 are directionally consistent.
  - 47 are nominally statistically significant.

- **Identified 2 novel signals with risk-associated alleles found only in men of African ancestry.**

Trans-Ancestry GWAS Meta-Analysis of Prostate Cancer

Goal: Combine GWAS data across diverse populations to identify novel variants and stronger markers of risk in known regions

PRACTICAL

	Number of Samples		
Population	**Cases**	**Controls**	**Total**
African	10,368	10,986	21,354
East Asian	8,611	18,809	27,420
European	85,554	91,972	177,526
Hispanic	2,714	5,239	7,953
Total	**107,247**	**127,006**	**234,253**

Sub-study Name	Sub-study Abbreviation	Population	No. of Cases in the analysis	No. of Controls in the analysis	Individual or Summary Level Data
African Ancestry Studies					
Multiethnic Cohort (MEC)	MEC	African	1784	1669	Individual
Southern Community Cohort Study	SCCS	African	250	513	Individual
The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	PLCO	African	231	240	Individual
Ghana Prostate Study	GPS	African	640	634	Summary
Kaiser	ProHealth	African	601	1,650	Summary

Analysis: Use meta-analysis techniques (i.e. fixed effects) to obtain marginal summary statistics.

Transethnic Meta-Analysis

- b_1 , …, b_C are the effect-size estimates (e.g. log odds ratios) in *N* studies/populations.
- s_i = standard error of the estimate
- K unknown clusters
- Ψ is the cluster specific effect
- T_{ik} = indicator of if population *i* is assigned to cluster *k*.

Morris AP. *Genet Epidemiol*. 2011;35:809–822.

Likelihoods for Transethnic Meta

- Cluster-specific:

$$L_1 \propto \frac{1}{s_i} \exp\left[-\frac{\left(b_i - \sum_{k=1}^{K} T_{ik}\psi_k \right)^2}{2s_i^2} \right]$$

- K=1: no heterogeneity between population-specific effects, Bayesian fixed-effects meta-analysis.
- K=N: each population is assigned to a different cluster, Bayesian random-effects meta-analysis.

Transethnic Meta-Analysis

TABLE I. Transethnic meta-analysis of five association studies of T2D at 19 variants in established susceptibility loci

Locus	SNP	Chromosome	Position (bp)	Effect allele frequencies	*K* unconstrained		*K* = 1 (fixed effect)
					$\log_{10}$ BF	*P*(heterogeneity)	$\log_{10}$ BF
NOTCH2	rs10923931	1	120,319,482	0.02–0.29	0.1	21.8%	0.0
THADA	rs7578597	2	43,586,327	0.75–0.99	0.8	25.4%	0.8
PPARG	rs1801282	3	12,368,125	0.89–0.97	0.8	55.2%	0.2
ADAMTS9	rs4607103	3	64,686,944	0.61–0.73	−0.3	9.2%	−0.3
IGF2BP2	rs4402960	3	186,994,381	0.27–0.49	3.3	24.6%	3.3
WFS1	rs10010131	4	6,343,816	0.59–0.98	2.0	70.1%	1.6
CDKAL1	rs7754840	6	20,769,229	0.29–0.55	11.0	99.2%	8.9
JAZF1	rs864745	7	28,147,081	0.51–0.77	7.4	22.7%	7.3
SLC30A8	rs13266634	8	118,253,964	0.60–0.89	3.7	11.0%	3.8
CDKN2A/B	rs2383208	9	22,122,076	0.56–0.85	5.0	15.6%	5.3
HHEX	rs1111875	10	12,368,016	0.28–0.74	0.4	32.2%	0.1
TCF7L2	rs7903146	10	94,452,862	0.04–0.28	17.0	21.9%	16.9
CDC123	rs12779790	10	114,748,339	0.14–0.18	1.3	16.1%	1.1
KCNQ1	rs2237895	11	2,813,770	0.20–0.42	1.7	13.3%	1.8
KCNQ1	rs2237897	11	2,815,122	0.62–0.95	3.9	13.7%	3.8
KCNJ11	rs5219	11	17,366,148	0.09–0.37	4.0	20.1%	3.8
TSPAN8	rs7961581	12	69,949,369	0.21–0.29	−0.3	13.2%	−0.4
FTO	rs8050136	16	52,373,776	0.20–0.43	−0.3	10.0%	−0.3
HNF1B	rs4430796	17	33,172,153	0.31–0.65	0.4	48.0%	0.1

Two MANTRA analyses are performed at each variant: (i) with an unconstrained number of clusters, *K*, of populations; and (ii) with a single cluster (*K* = 1, i.e. fixed-effects). For each analysis, the log10 Bayes' factor (BF) in favor of association is presented. For the analysis with *K* unconstrained, the posterior probability of heterogeneity in allelic effects, *P*(heterogeneity), is also presented. T2D, type 2 diabetes.

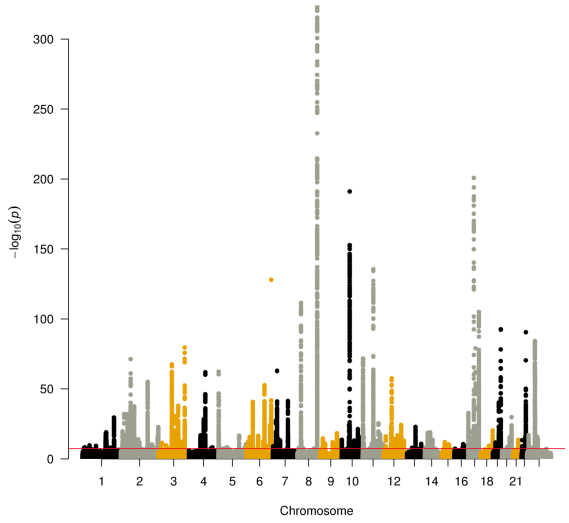
Across Populations

- How does discovery, characterization, and translation/prediction differ across populations?
    - Different linkage disequilibrium.
    - Different allele frequencies: common vs. population-specific.
    - Differential coverage between populations due to genotyping arrays.
    - Different imputation quality between populations.
    - Different missingness patterns in real data.
    - Different sample sizes.
    - Different allelic *risk* (i.e. interactions).
    - Different variation in the trait.

Trans-Ancestry GWAS Meta-Analysis of Prostate Cancer

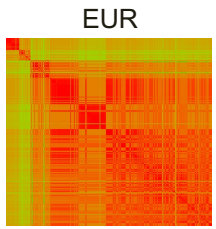
Goal: Combine GWAS data across diverse populations to identify novel variants and stronger markers of risk in known regions

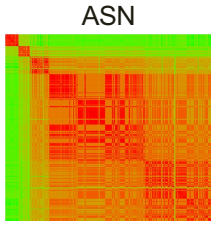
	Number of Samples		
Population	**Cases**	**Controls**	**Total**
African	10,368	10,986	21,354
East Asian	8,611	18,809	27,420
European	85,554	91,972	177,526
Hispanic	2,714	5,239	7,953
Total	**107,247**	**127,006**	**234,253**

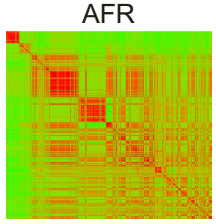


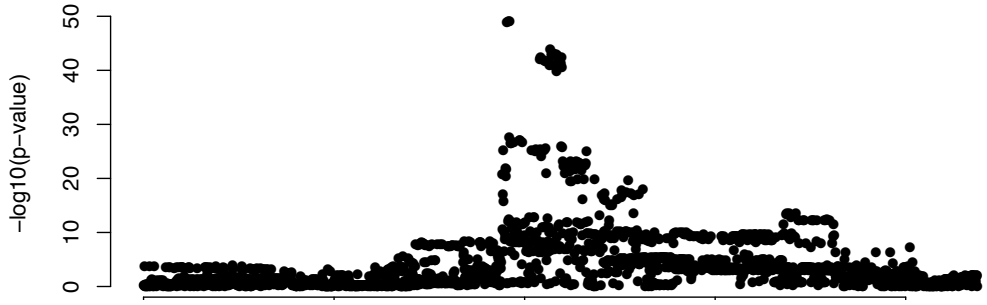
How do we determine the SNP(s) driving an association within a region?

Linkage Disequilibrium Across Populations









1. Thresholding and Pruning (T + P):
   - How to prune across multiple ancestries?
2. Fine-mapping to identify putative 'causal' variants:
   - Requires selection with a conditional model for summary statistics by ancestry.
   - How do we incorporate ancestry specific reference panels?

Multipopulation JAM (*mJAM*)

- Conceptually:
  - Stage 1: Fit a JAM model in each population.
  - Stage 2: Meta-analyze the joint estimates with a meta-regression.

Population *i* $z_L^{(i)} \sim MVN_P(L^{(i)}\beta, \sigma^2 I_P)$

$$\begin{pmatrix} \hat{\beta}^{(1)} \\ \hat{\beta}^{(2)} \\ \hat{\beta}^{(3)} \end{pmatrix} = \begin{pmatrix} I_P \\ I_P \\ I_P \end{pmatrix} \pi + \delta$$

π is the overall multiethnic meta-analysis effect estimate

Multipopulation JAM (*mJAM*)

$$\boldsymbol{y}^{(1)} = \boldsymbol{G}^{(1)}\widehat{\boldsymbol{\beta}}^{(1)}$$ Linear model for population (1)

$$\boldsymbol{y}^{(2)} = \boldsymbol{G}^{(2)}\widehat{\boldsymbol{\beta}}^{(2)}$$ Linear model for population (2)

$$\boldsymbol{y}^{(3)} = \boldsymbol{G}^{(3)}\widehat{\boldsymbol{\beta}}^{(3)}$$ Linear model for population (3)

Model multiple populations jointly under a fixed-effect framework

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}^{(1)}\\ \widehat{\boldsymbol{\beta}}^{(2)}\\ \widehat{\boldsymbol{\beta}}^{(3)}\end{pmatrix}_{3p\times 1} = \begin{pmatrix}\boldsymbol{I}_{\mathrm{p}} & 0 & 0\\ 0 & \boldsymbol{I}_{\mathrm{p}} & 0\\ 0 & 0 & \boldsymbol{I}_{\mathrm{p}}\end{pmatrix}\boldsymbol{\beta}_{fixed} + \boldsymbol{\delta}$$

$$\begin{pmatrix}\boldsymbol{y}^{(1)}\\ \boldsymbol{y}^{(2)}\\ \boldsymbol{y}^{(3)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{G}^{(1)} & 0 & 0\\ 0 & \boldsymbol{G}^{(2)} & 0\\ 0 & 0 & \boldsymbol{G}^{(3)}\end{pmatrix}\begin{pmatrix}\boldsymbol{I}_{\mathrm{p}} & 0 & 0\\ 0 & \boldsymbol{I}_{\mathrm{p}} & 0\\ 0 & 0 & \boldsymbol{I}_{\mathrm{p}}\end{pmatrix}\boldsymbol{\beta}_{fixed} + \boldsymbol{\epsilon}'$$ **"mJAM"**

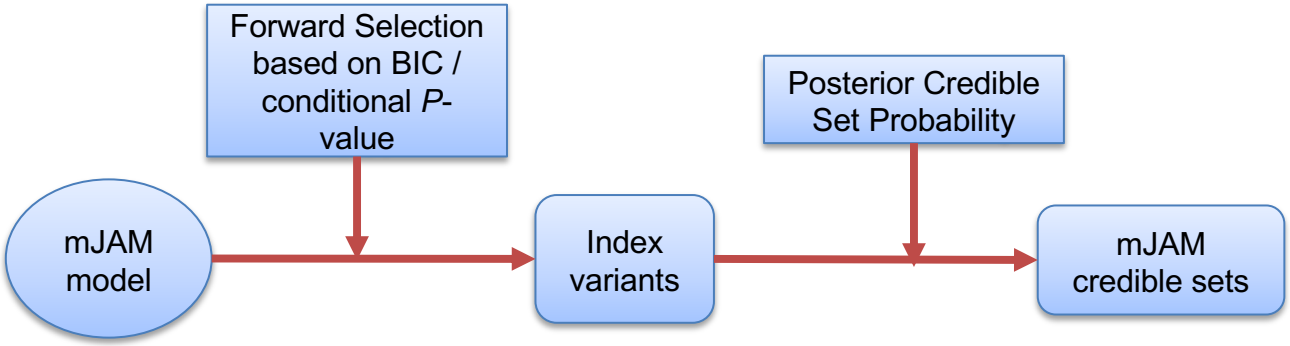
Allows for population-specific missing SNPs

Considerations for Using Summary Statistics

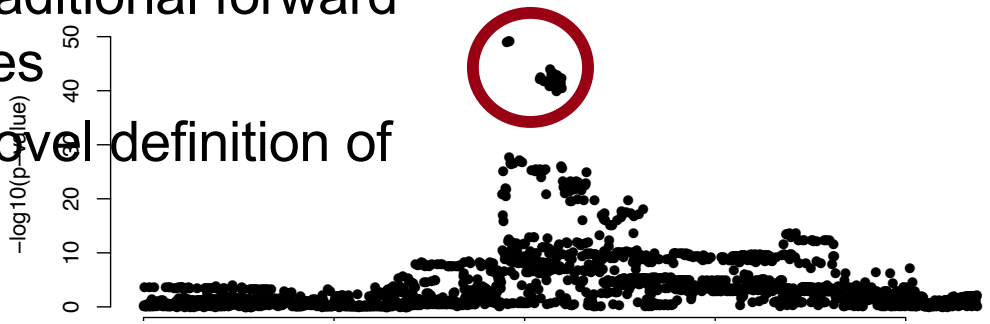
$$\hat{\beta} = (\boldsymbol{G_R}'\boldsymbol{G_R})^{-1}\boldsymbol{z}$$

- Does the reference correlation (LD) represent the same correlation that exists in the individual-level data?
  - Is the reference sample large enough to estimate correlations?
  - Do all SNPs exist in the reference data?
  - What about across multiple ancestries?

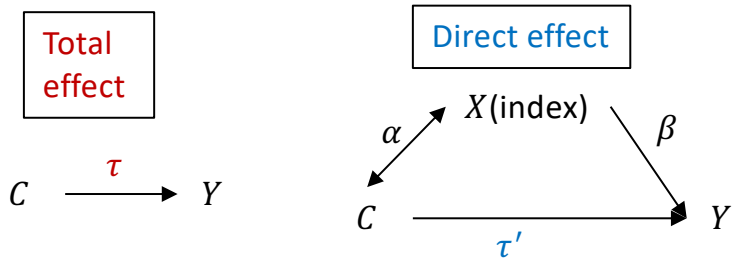
Implementation: mJAM-Forward



I. Selection of index SNPs using traditional forward selection with conditional *P*-values
II. Creation of credible sets using novel definition of posterior credible set probability.



Posterior Credible Set Probability



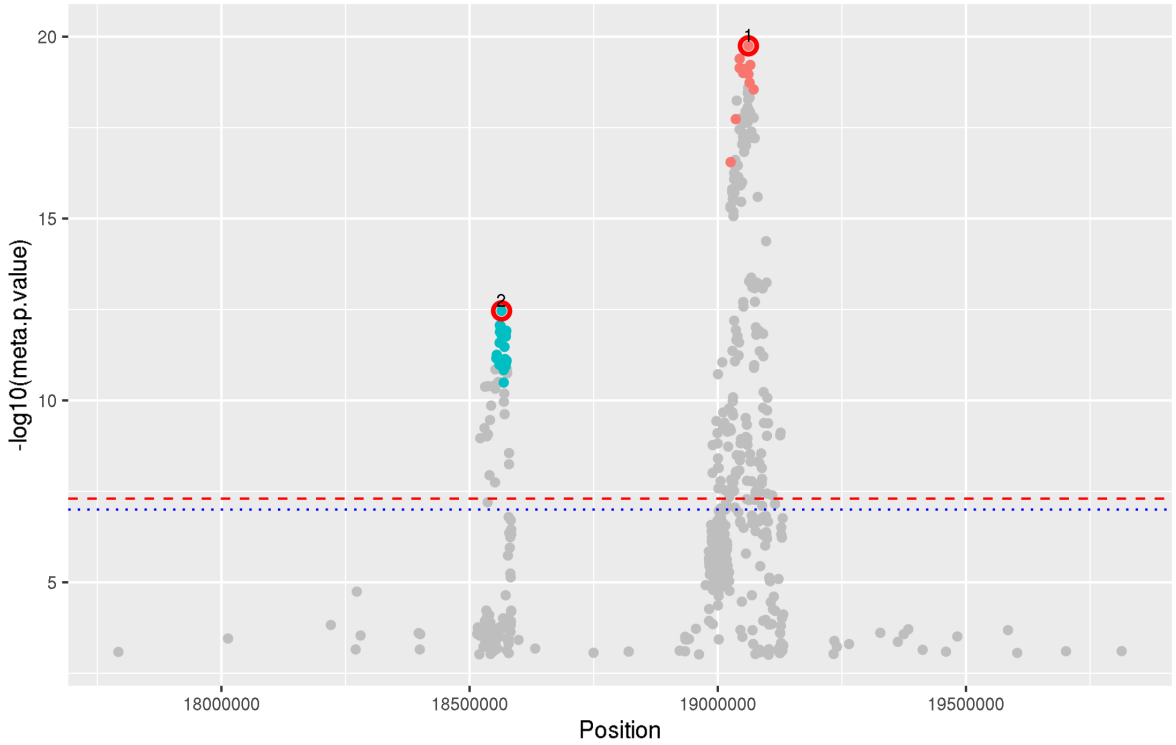
X: selected index SNP
C: a SNP in its credible set
Y: outcome

Criteria for credible set SNPs:

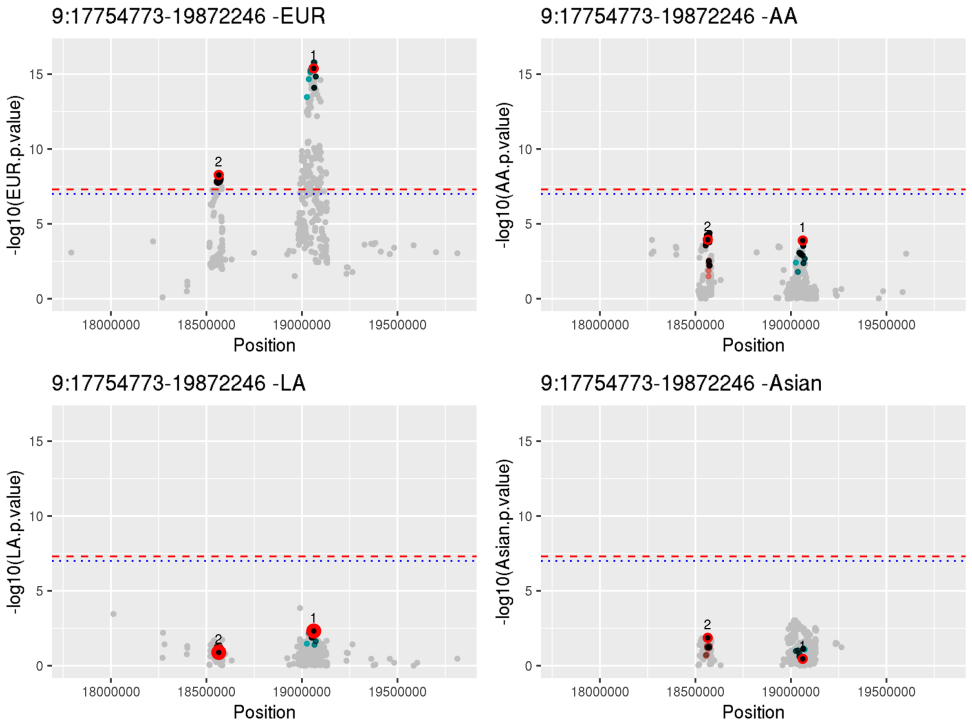
1. C is significantly associated with Y (Posterior Inclusion Probability)
2. By adding X, the effect of C on Y changes (Posterior Mediation Probability)

$$p_C := \Pr(Model|\ Data) \cdot \Pr(Mediation|Data)$$

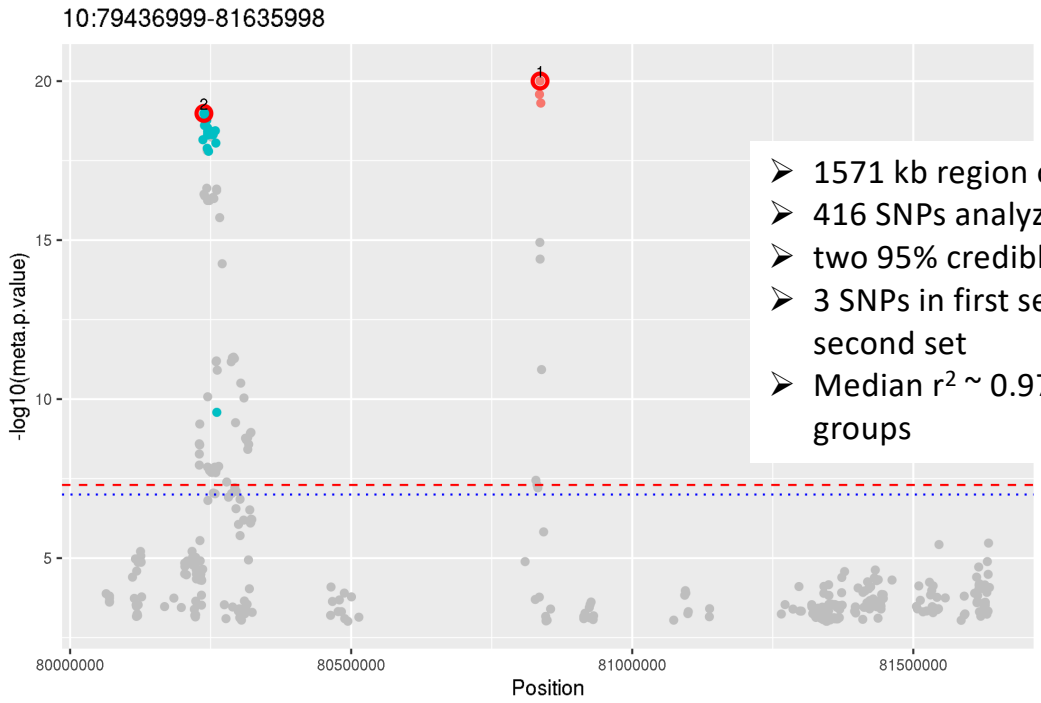
Multi-Ancestry Analysis: Region 1



Ancestry-Specific: Region 1

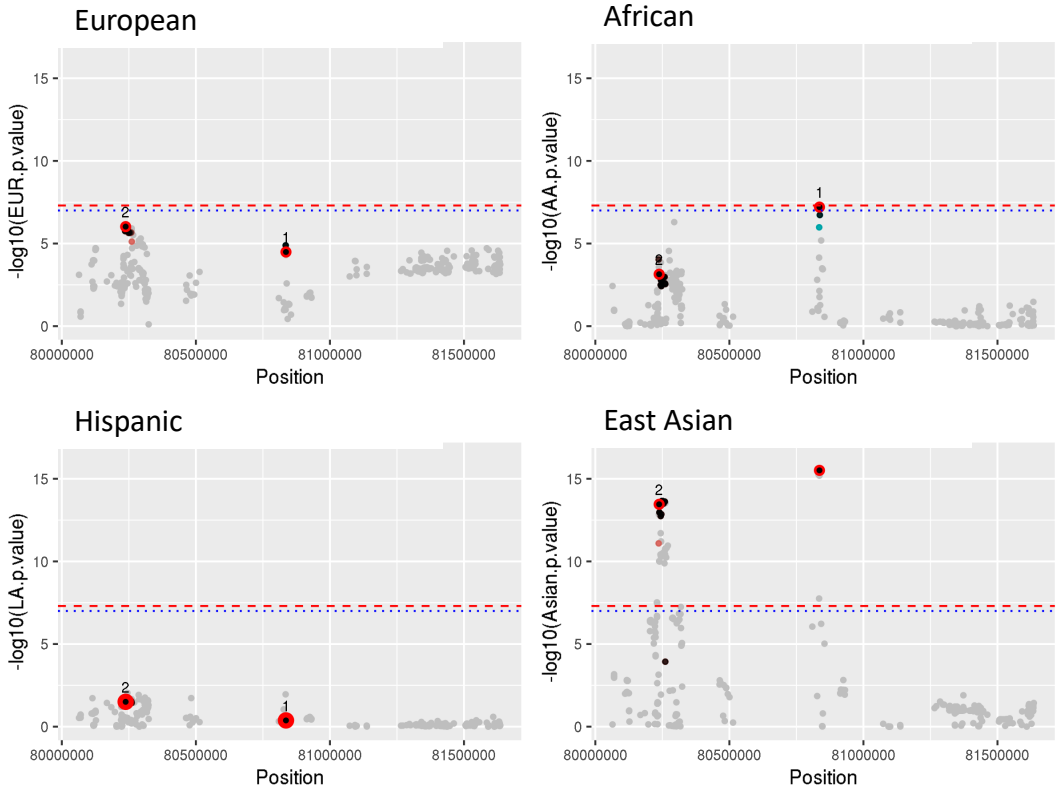


Multi-Ancestry Analysis: Region 2



- 1571 kb region on chr 10
- 416 SNPs analyzed
- two 95% credible sets identified
- 3 SNPs in first set and 45 SNPs in the second set
- Median r^2 ~ 0.97 across all ancestry groups

Ancestry-Specific: Region 2



SuSiE as the fine-mapping tool

$$\begin{pmatrix} \boldsymbol{y}^{(1)} \\ \boldsymbol{y}^{(2)} \\ \boldsymbol{y}^{(3)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}^{(1)} & 0 & 0 \\ 0 & \boldsymbol{G}^{(2)} & 0 \\ 0 & 0 & \boldsymbol{G}^{(3)} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{\mathrm{p}} & 0 & 0 \\ 0 & \boldsymbol{I}_{\mathrm{p}} & 0 \\ 0 & 0 & \boldsymbol{I}_{\mathrm{p}} \end{pmatrix} \boldsymbol{\beta}_{fixed} + \boldsymbol{\epsilon}'$$

Which one is the causal SNP?

SuSiE: "Sum of Single Effects" model[1]

$$\underline{\beta}_{fixed} = \sum_{l=1}^{L} \beta_l = \sum_{l=1}^{L} \beta_l \gamma_l$$

$$\gamma_l \sim Multi(1, \tau) \text{ and } \beta_l \sim N_1(0, \sigma_{0l}^2)$$

Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). *Journal of the Royal Statistical Society, Series B* **82**, 1273--1300.

SuSiE as the fine-mapping tool

$$\begin{pmatrix} \boldsymbol{y}^{(1)} \\ \boldsymbol{y}^{(2)} \\ \boldsymbol{y}^{(3)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}^{(1)} & 0 & 0 \\ 0 & \boldsymbol{G}^{(2)} & 0 \\ 0 & 0 & \boldsymbol{G}^{(3)} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{\mathrm{p}} & 0 & 0 \\ 0 & \boldsymbol{I}_{\mathrm{p}} & 0 \\ 0 & 0 & \boldsymbol{I}_{\mathrm{p}} \end{pmatrix} \boldsymbol{\beta}_{fixed} + \boldsymbol{\epsilon}'$$

Which one is the causal SNP?

SuSiE: "Sum of Single Effects" model[1]

- Computationally fast

  Iterative Bayesian Stepwise Selection — a Bayesian analogue of stagewise selection.

- Easily interpretable

  Outputs credible set(s), each has a high probability of containing a causal SNP from a set of correlated SNPs.

Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). *Journal of the Royal Statistical Society, Series B* **82**, 1273--1300.

mJAM Summary

- mJAM model
  - Within-population LD explicitly modeled in mJAM.
  - Able to incorporate population-specific missing SNPs.
- mJAM-SuSiE
  - Good performance in modestly significant regions, but breaks down in highly significant regions
- mJAM-Forward
  - Good performance and easy to implement and interpret
- mJAM credible set construction
  - Given any index SNPs, construct credible sets using mediation techniques
  - Provide index-specific credible sets

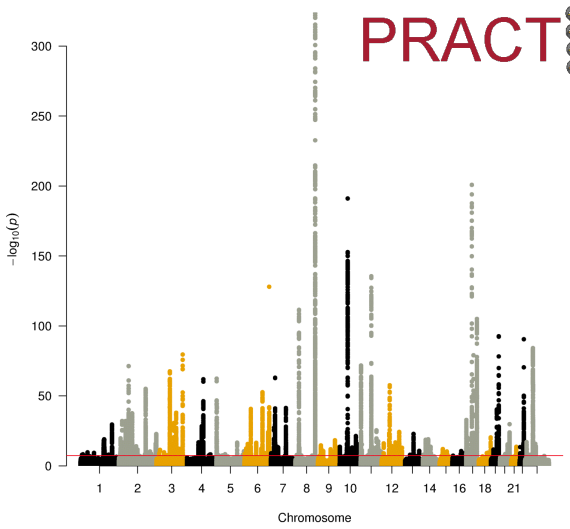
⁎ genome-wide mJAM implementation and incorporate *xtune*.

Trans-Ancestry GWAS Meta-Analysis of Prostate Cancer

Goal: Combine GWAS data across diverse populations to identify novel variants and stronger markers of risk in known regions

PRACTICAL

	Number of Samples		
Population	**Cases**	**Controls**	**Total**
African	10,368	10,986	21,354
East Asian	8,611	18,809	27,420
European	85,554	91,972	177,526
Hispanic	2,714	5,239	7,953
Total	**107,247**	**127,006**	**234,253**



"Fine-mapping" via JAM with forward stepwise selection using ancestry-specific summary stats

Primary Findings

86 novel risk variants

62 known risk regions → identified stronger markers identified in fine-mapping

+121 known variants remain lead signal in regions

269 total prostate cancer variants *Includes ethnic-specific variants

Construction of a Genetic Risk Score (GRS):

Weighted sum of # risk alleles carried by each participant

Count of risk alleles for variant *m* for individual *i*

SNPs=model selection

$$GRS_i = \sum_{m=1}^{M} w_m G_{im}$$

Construction of a Genetic Risk Score (GRS):

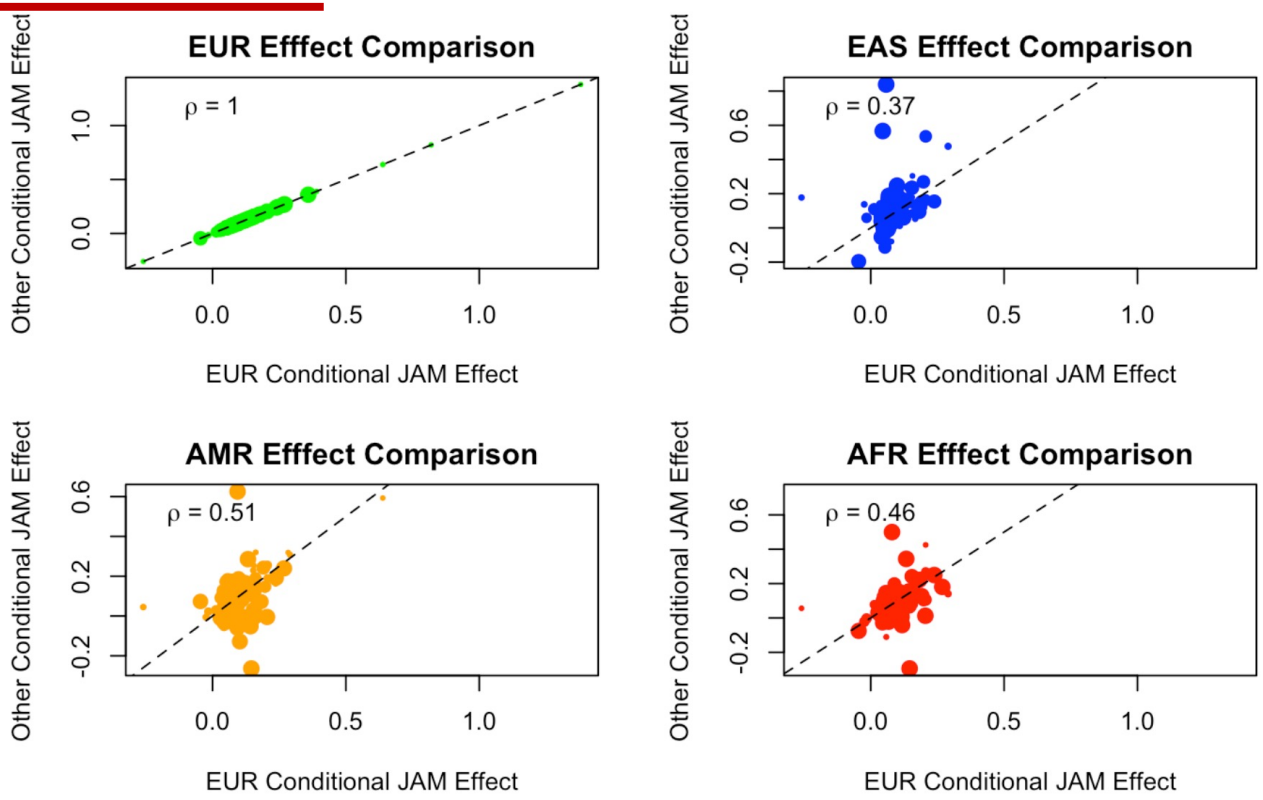
Weighted sum of # risk alleles carried by each participant

Count of risk alleles for variant *m* for individual *i*

$$GRS_i = \sum_{m=1}^{M} w_m G_{im}$$

What weight?

GRS Weights: Multi-Ancestry or Ancestry-Specific?



Construction of a Genetic Risk Score (GRS):

Weighted sum of # risk alleles carried by each participant

Count of risk alleles for variant *m* for individual *i*

$$GRS_i = \sum_{m=1}^{M} w_m G_{im}$$

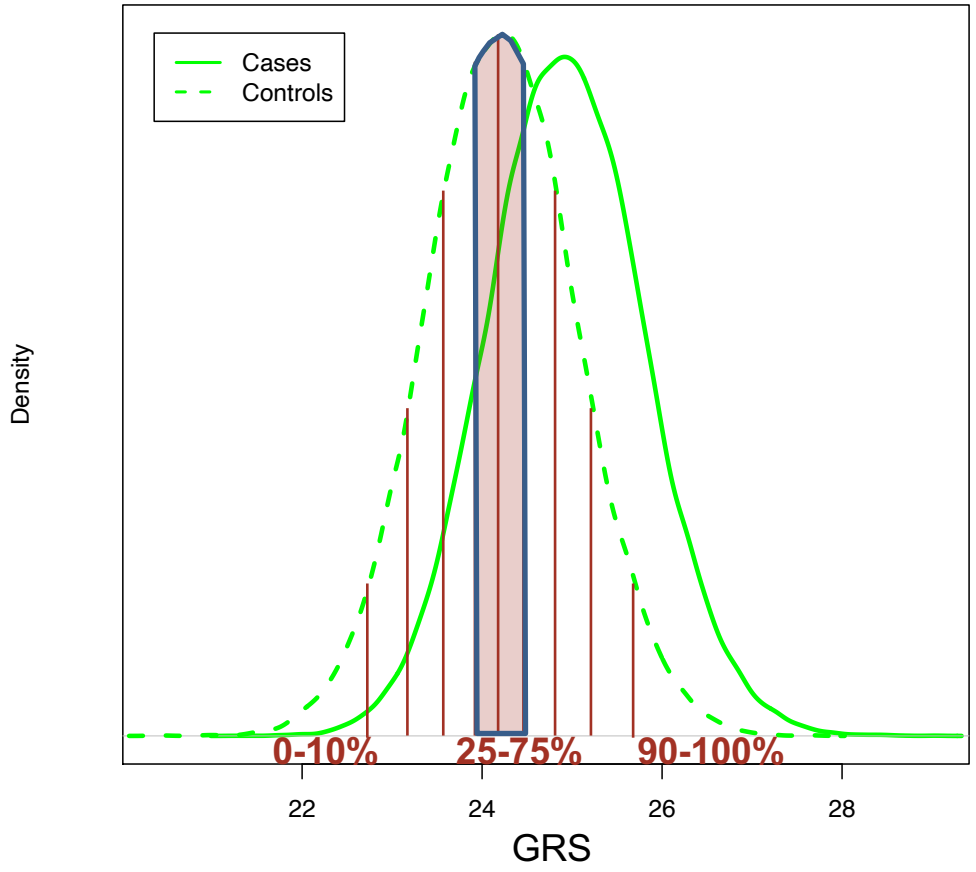
Multi-ancestry weight for variant *m* (from mJAM meta-analysis of population-specific conditional effects)

Evaluation of a Genetic Risk Score (GRS):

Weighted sum of # risk alleles carried by each participant

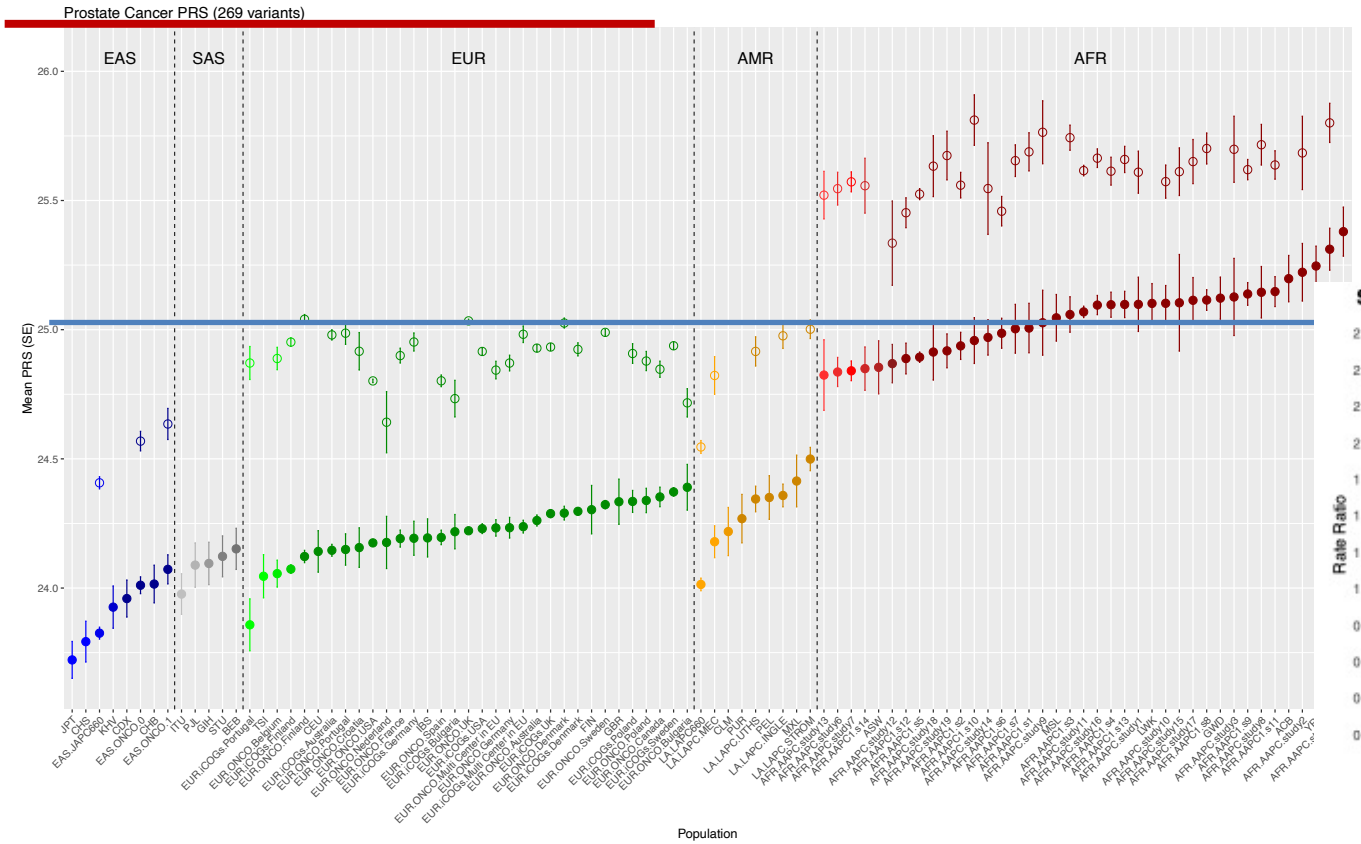
Count of risk alleles for variant *m* for individual *i*

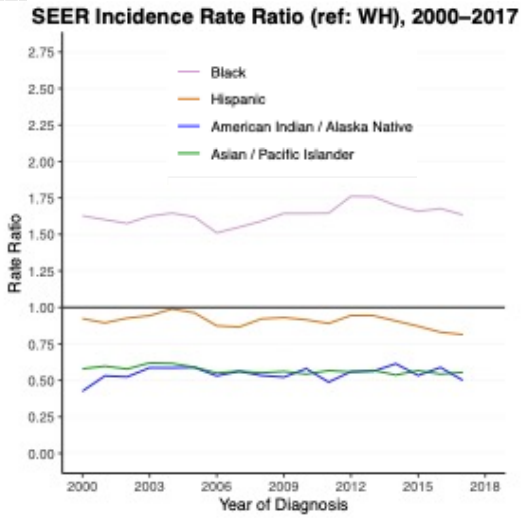
$$GRS_i = \sum_{m=1}^{M} w_m G_{im}$$



How to evaluate within and across populations?

GRS Distribution Across Populations





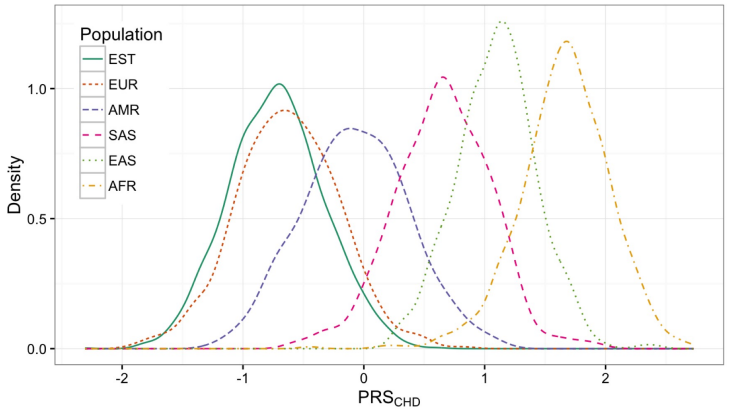


Fig 1. PRS$_{CHD}$ distributions in different populations.

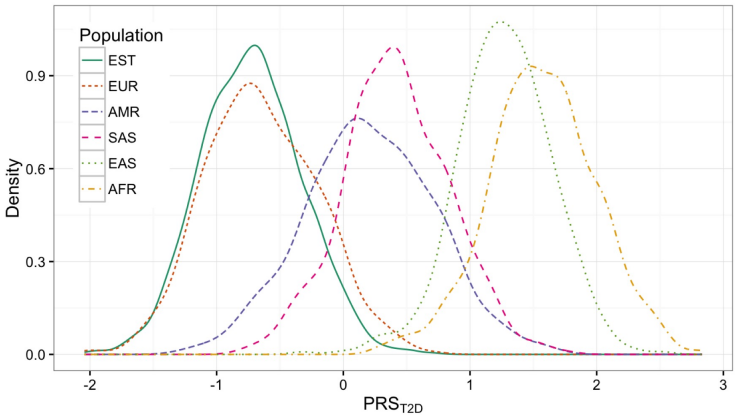


Fig 2. PRS$_{T2D}$ distributions in different populations.

PLOS ONE

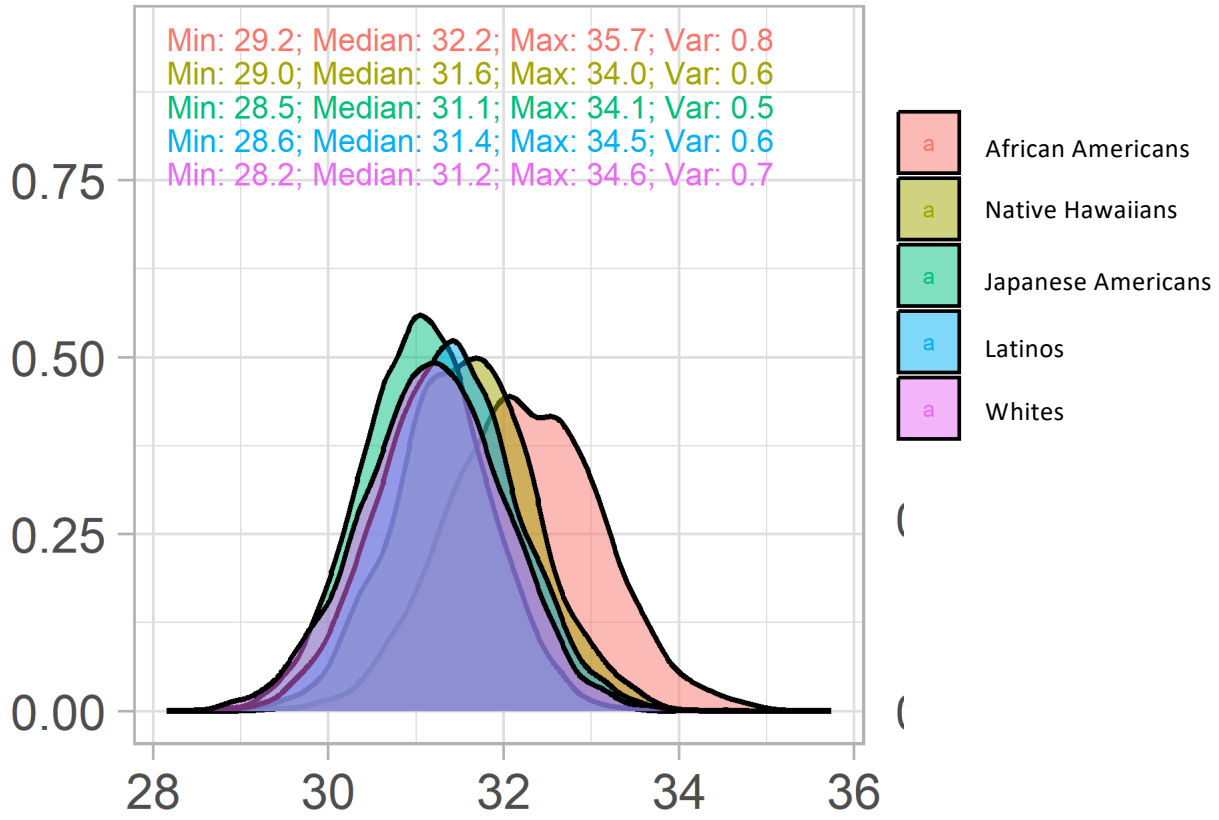
RESEARCH ARTICLE

Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations

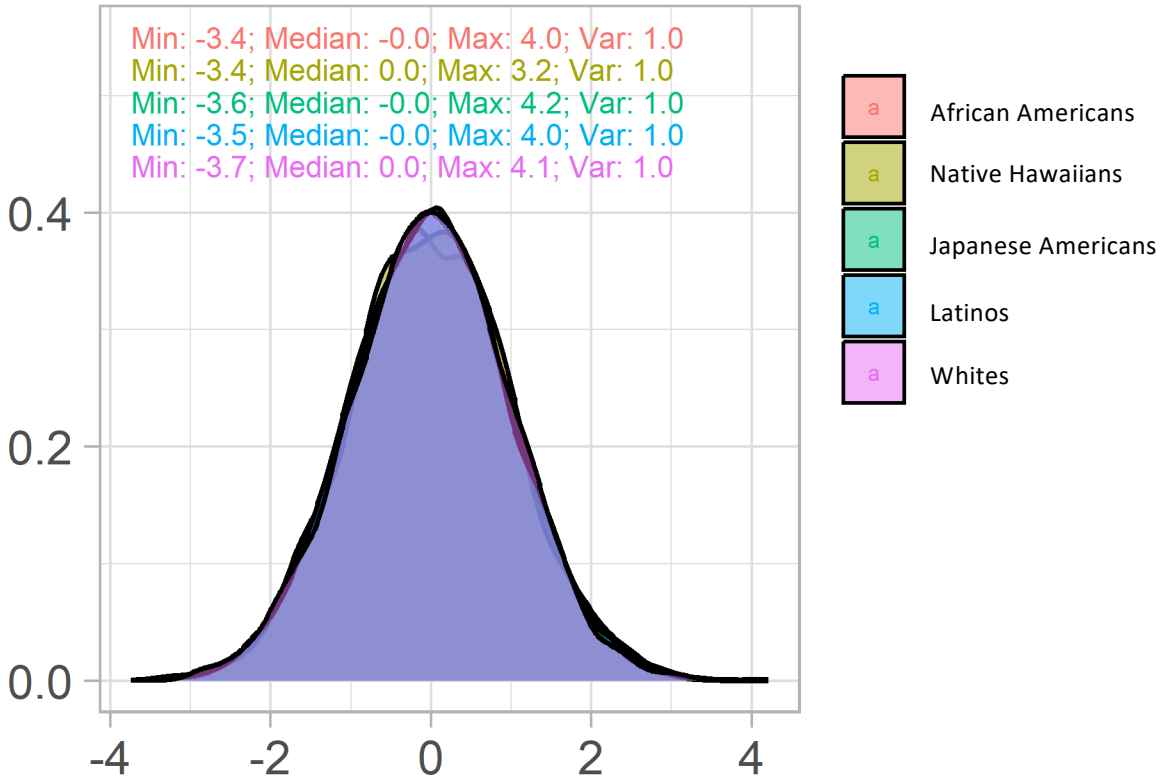
Sulev Reisberg[1,2,3]*, Tatjana Iljasenko[1], Kristi Läll[4,5], Krista Fischer[5☯], Jaak Vilo[1,2,3☯]

- "…as the linkage disequilibrium between causal and associated SNPs varies in different populations, the effect size of the disease-associated variant tends to be over-estimated in non-European ancestries for approximately a quarter of SNPs"
- "…SNPs in our models tend to have higher effect allele frequencies in African populations compared to Europeans, consequently leading to relatively higher scores."

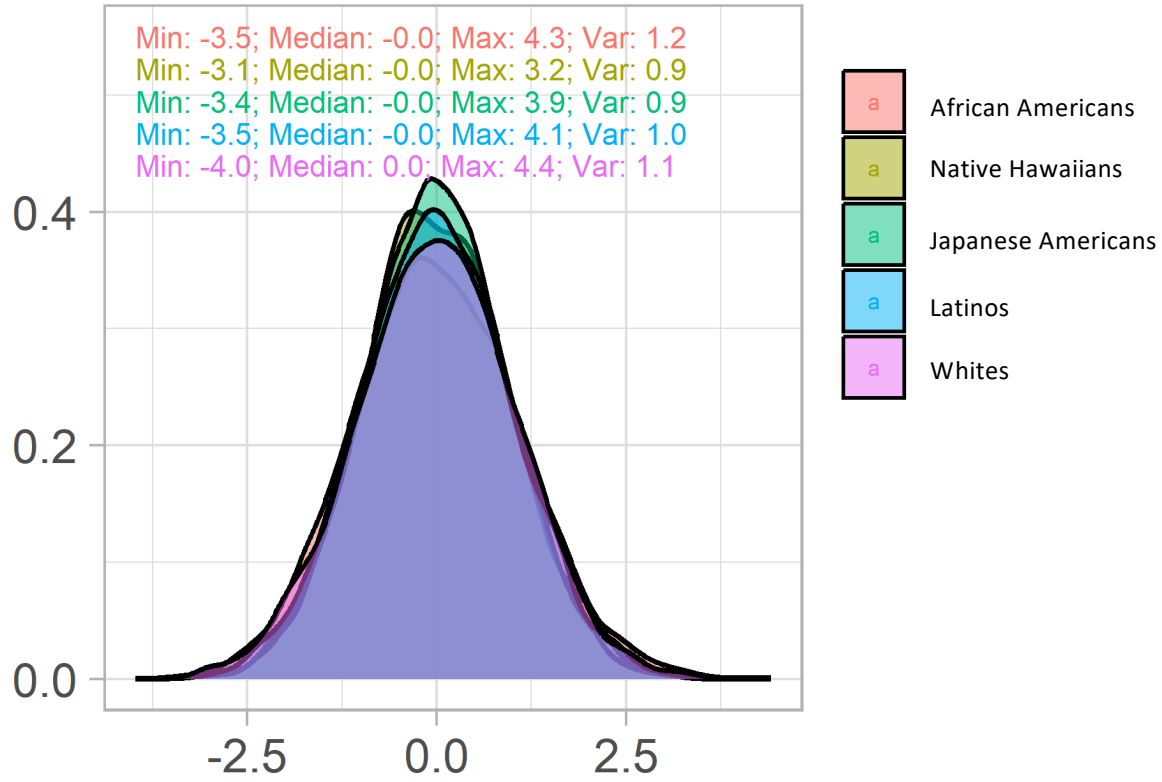
GRS Distribution by Populations



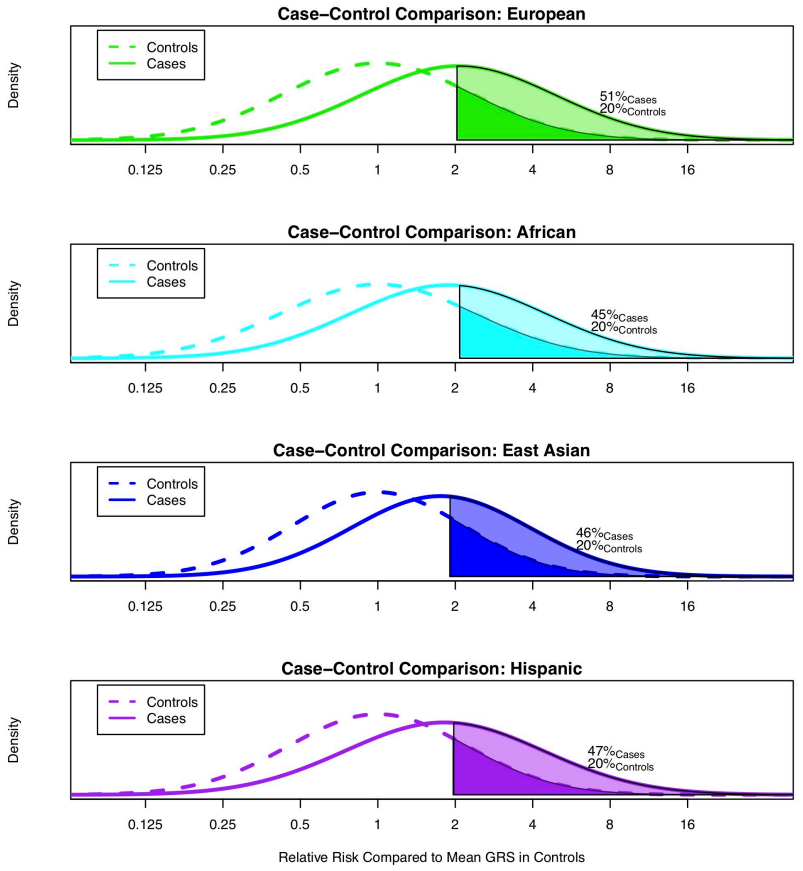
Standardized (via Population)



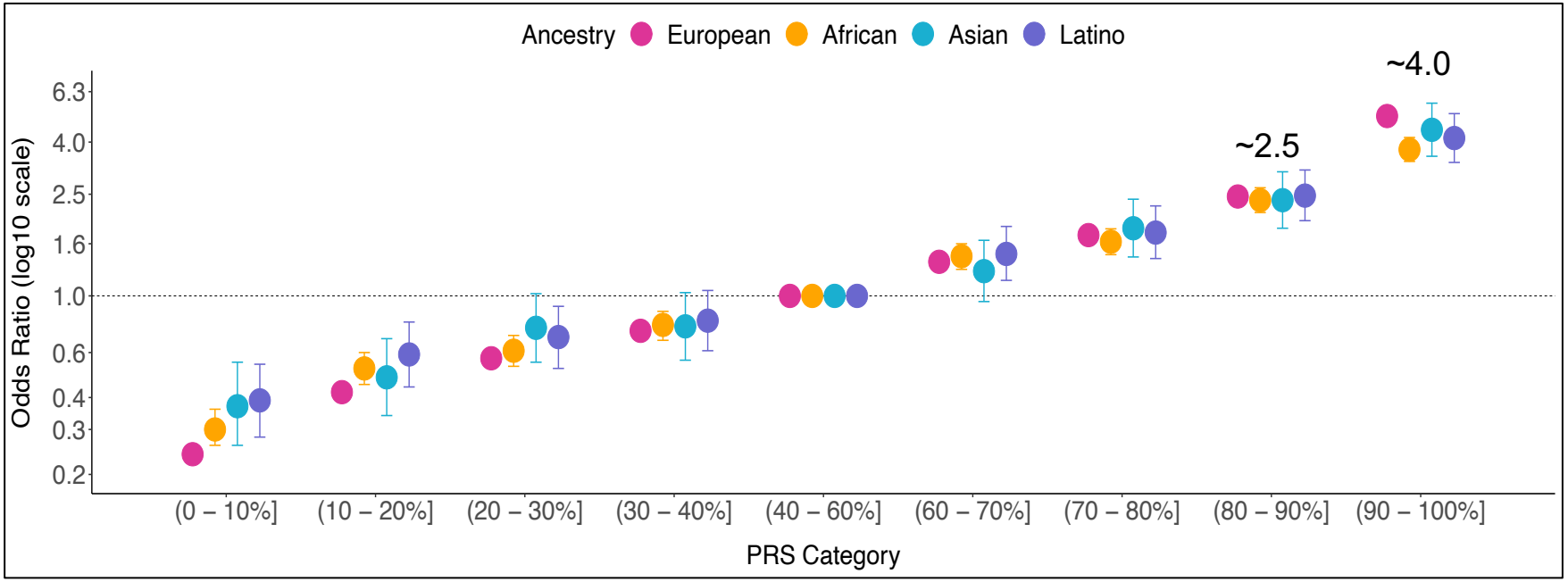
Standardized (via PCs)



GRS Within Population Comparisons



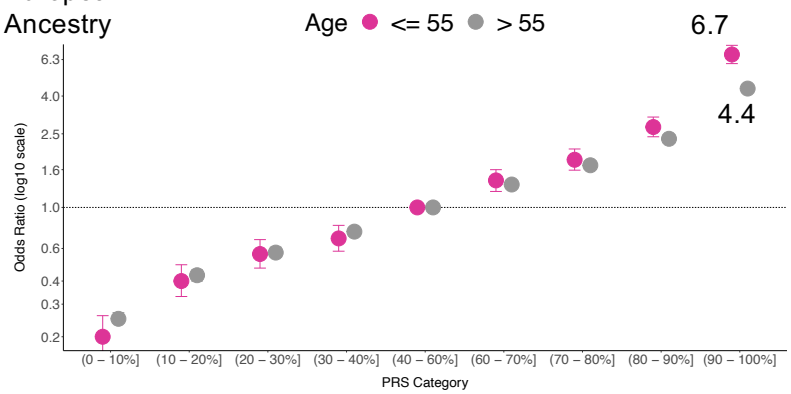
Genetic Risk Score (GRS) for Prostate Cancer, 269 variants

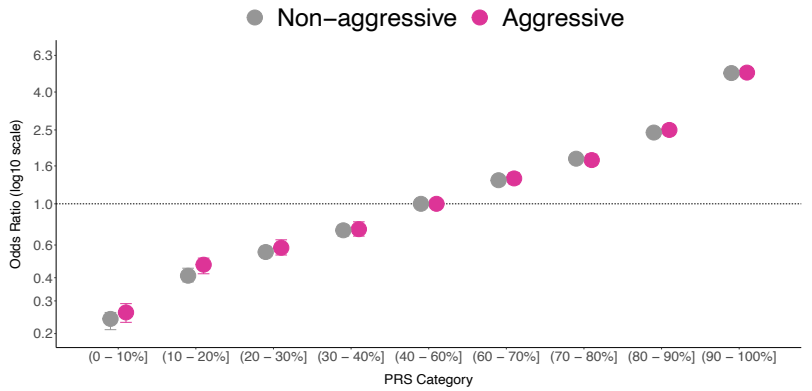


Conti et al Nat Genet 2021

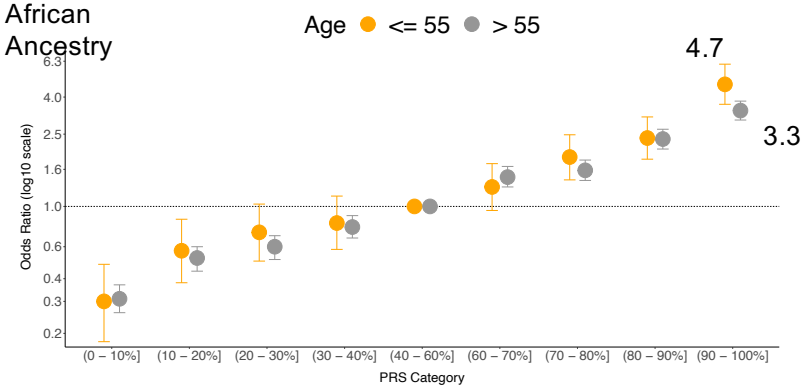
Characteristics of the Genetic Risk Score

European Ancestry

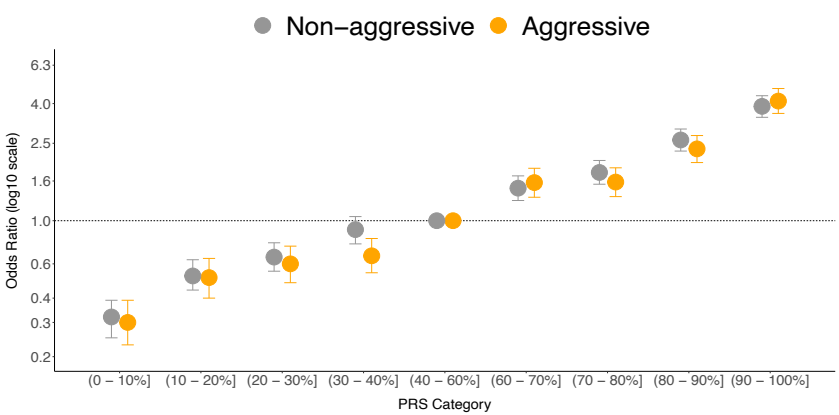




African Ancestry



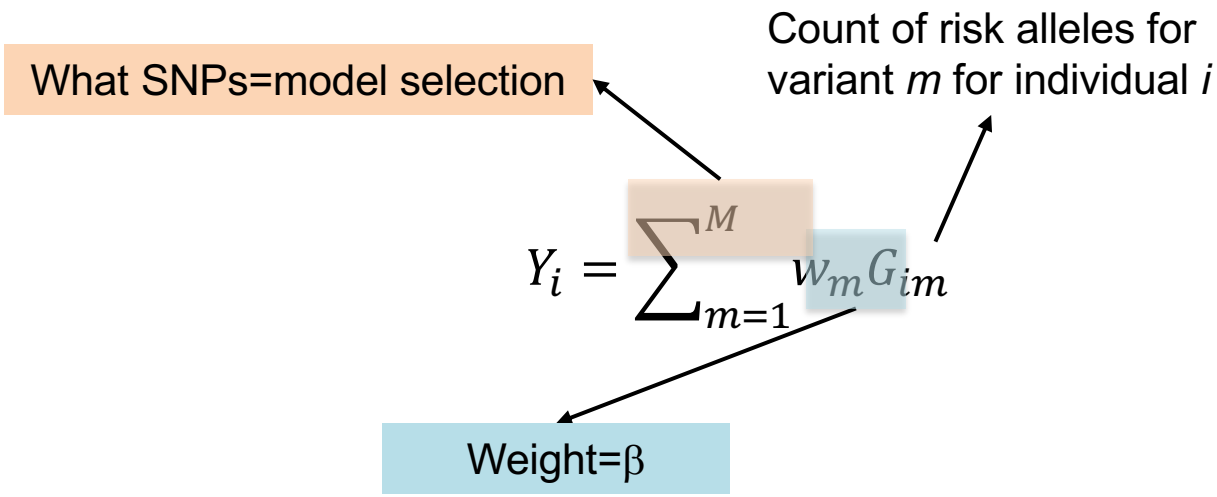
Greater impact on prostate cancer risk for early versus late onset



Does not discriminate risk of aggressive vs non-aggressive disease

Improving the Genetic Risk Score: Genomewide Scores

Weighted sum of # risk alleles carried by each participant

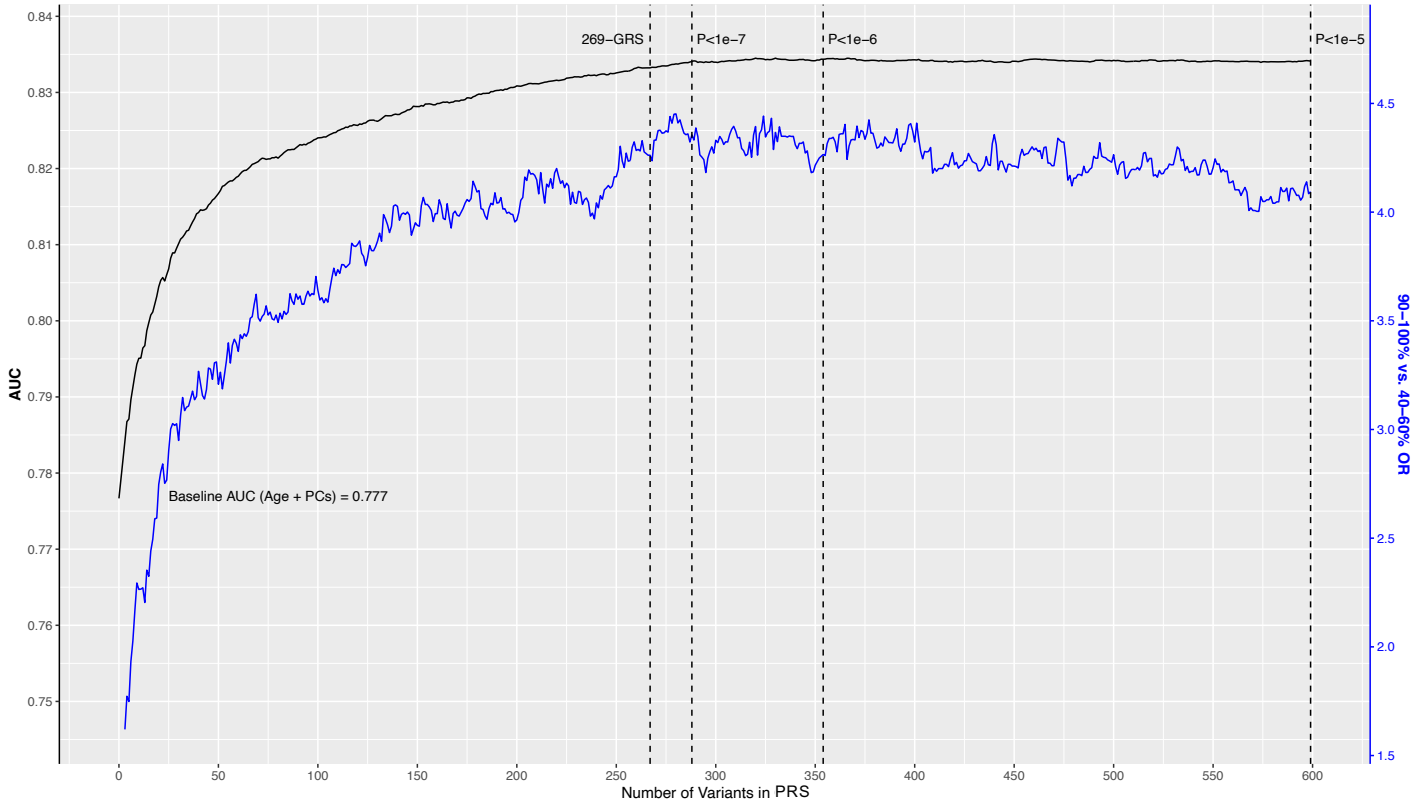


SNP Selection and Effect Estimation

When the number of SNPs is large, estimation of effects can be unstable.

1. Selection of SNPs:
   a) Selection using statistical significance of each SNP (e.g. P+T/forward selection).

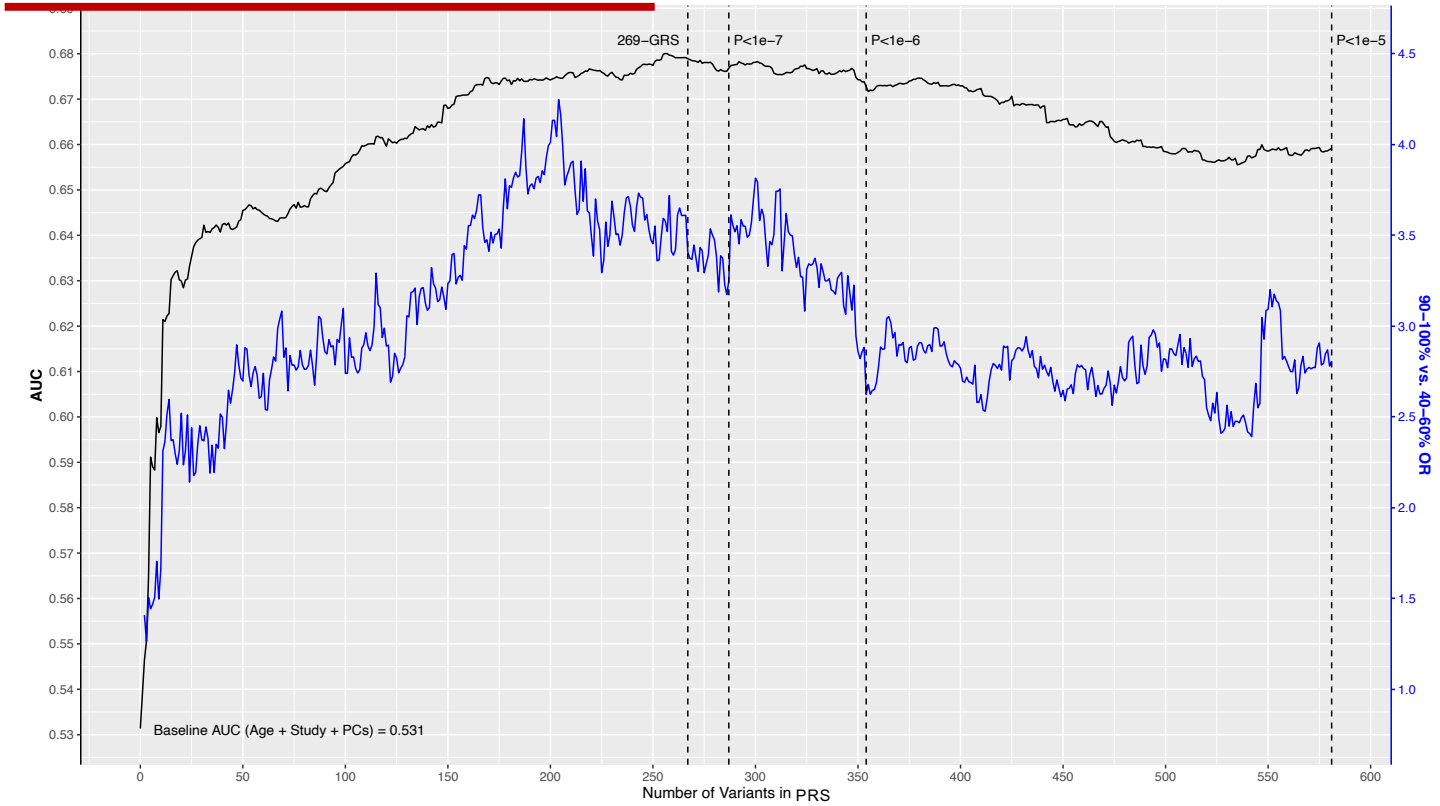
Genome-wide Genetic Risk Score in European Ancestry Men



<u>UK Biobank</u>
6,852 cases
193,117 controls

We iteratively evaluated the area under the curve (AUC) and OR (90%-100% vs. 40-60% GRS categories) of a *genome-wide* GRS by adding one variant to the model at a time including variants with multi-ancestry P-value<1e-5

Genome-wide Genetic Risk Score in African Ancestry Men



CA UG Study
1,586 cases
1,047 controls

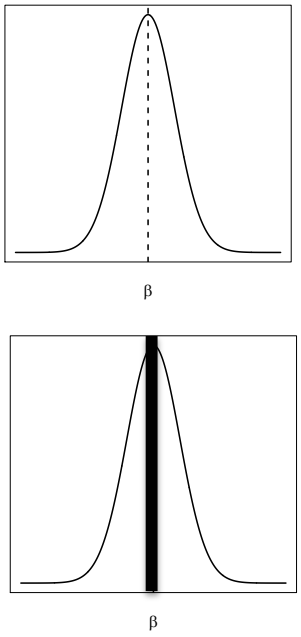
We iteratively evaluated the area under the curve (AUC) and OR (90%-100% vs. 40-60% GRS categories) of a *genome-wide* GRS by adding one variant to the model at a time including variants with multiancestry P-value<1e-5

Genome-wide GRS may not improve prediction of prostate cancer risk

SNP Selection and Effect Estimation

When the number of SNPs is large, estimation of effects can be unstable.

1. Selection of SNPs:
   a) Selection using statistical significance of each SNP (e.g. P+T/forward selection).
2. Penalization to stabilize estimation and/or select SNPs.
   a) Regularized regression (e.g. ridge and lasso).
   b) Bayesian modeling and selection.
3. How to perform these analyses with multiple populations?



Mean Variance

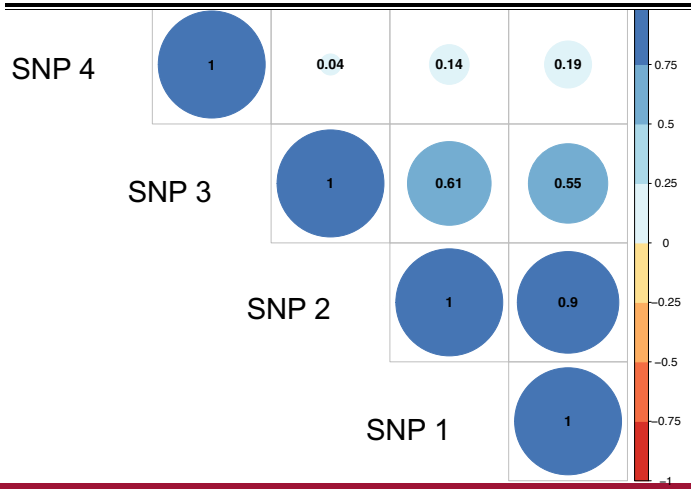
$$\beta_m = \begin{cases} N(0, \tau^2) & \gamma_m = 1 \\ 0 & \gamma_m = 0 \end{cases}$$

PRS Methods

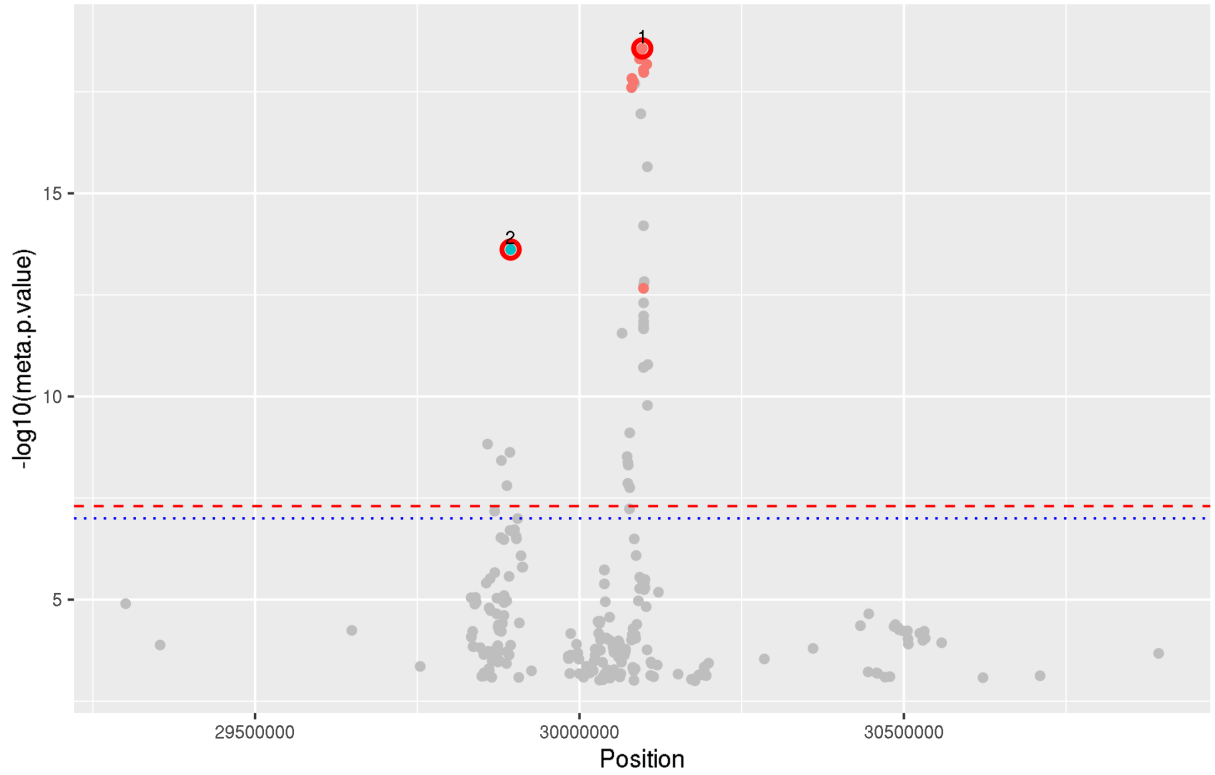
Point-normal	$\beta_j \sim \pi N(0, \sigma_\beta^2) + (1-\pi)\delta_{\{0\}}$	LDpred/ BVSR
Normal mixture	$\beta_j \sim \pi N(0, g\tau^2) + (1-\pi)N(0, \tau^2)$	BayesC/ SSVS
Normal mixture	$\beta_j \sim \pi N(0, \sigma_a^2 + \sigma_b^2) + (1-\pi)N(0, \sigma_b^2)$	BSLMM
Point-normal	$\beta_j \sim \pi N(0, [2p_j(1-p_j)]^S \sigma_\beta^2) + (1-\pi)\delta_{\{0\}}$	BayesS
Point-t	$\beta_j \sim \pi t_\nu(\tau^2) + (1-\pi)\delta_{\{0\}}$	BayesB/ BayesDπ
Normal mixture	$\beta_j \sim \pi_0\delta_{\{0\}} + \pi_1 N(0, 10^{-4}\times\sigma_\beta^2)$ $+\pi_2 N(0, 10^{-3}\times\sigma_\beta^2) + \pi_3 N(0, 10^{-2}\times\sigma_\beta^2)$	(S)BayesR
Infinite normal mixture	$\beta_j \sim \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_k^2),\ \pi_k = \nu_k \prod_{l=1}^{k-1}(1-\nu_l),\ \nu_k \sim \text{Beta}(1, \alpha)$	(S)DPR

Philosophy of Approaches

	Univariate		Regularized Regression			Bayesian selection	
Protein	Estimate	P-value	Ridge	Lasso	Elastic Net	Estimate	Pr(β≠0)
SNP 1	0.93	1.81E-186	0.27	0.63	0.48	0.89	1.00
SNP 2	0.83	6.94E-142	0.21	0.02	0.18	0.01	0.05
SNP 3	0.04	2.76E-01	-0.02	-	-	-0.17	1.00
SNP 4	0.61	2.28E-64	0.11	-	0.01	0.10	0.87



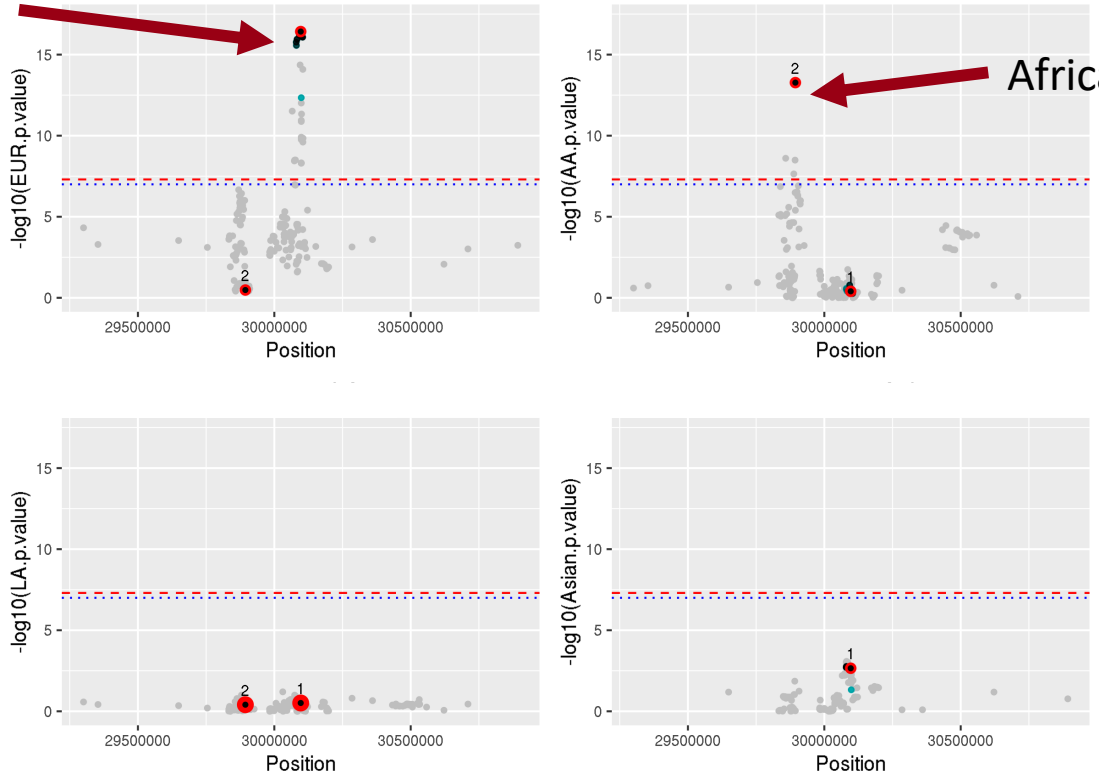
Multi-Ancestry Analysis: Region 3



Ancestry-Specific: Region 3

EUR/ASN effect

African-specific effect



PRS-CSx

A separate effect/weight for each SNP *j*
And each population *k*.

$$\beta_{jk} \sim N\left(0, \frac{\sigma_k^2}{N_k}\psi_j\right), \psi_j \sim \text{Gamma}\left(a, \delta_j\right), \delta_j \sim \text{Gamma}\left(b, \phi\right),$$

For each SNP *j*, across populations

For each population *k*, across all SNPs

Comparison to Genome-Wide PRS Approaches

Training Data

Population	Cases	Controls	Total
African	10,368	10,986	21,354
East Asian	8,611	18,809	27,420
European	85,554	91,972	177,526
Hispanic	2,714	5,239	7,953
Total	**107,247**	**127,006**	**234,253**

Multi-ancestry & pop-specific weights
Conti, Darst et al., *Nat Gen* 2021

PRACTICAL

Genome-wide PRS Construction

Six GW-PRS approaches were evaluated:
1) LDpred2; 2) PRS-CSx; 3) EB-PRS;
4) DBSLMM; 5) XPASS+; 6) multi-PRS

GWAS Summary Stats: Conti, Darst et al., *Nat Gen* 2021
Variants Included: HapMap3 Panel (1.1M)
LD Reference: 1000 Genomes Project

Testing Data

Population	Cases	Controls	Total
CA/UG African	1,586	1,047	2,633
UKB European	8,046	191,825	199,871
Total	**9,632**	**192,872**	**202,504**

Validation Data

Population	Cases	Controls	Total
MVP African	6,353	53,362	59,715
MVP European	13,643	210,214	223,857
Total	**19,996**	**263,576**	**283,572**

PRS Evaluation

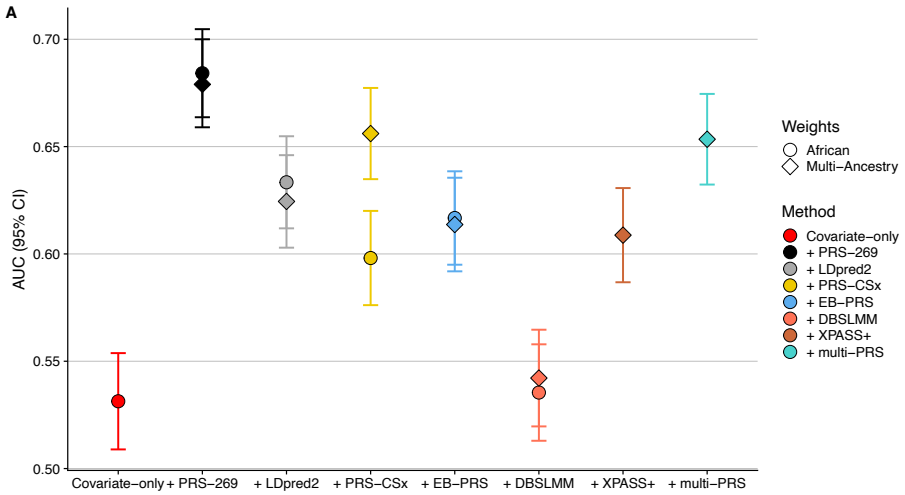
Area under the curve (AUC) & PRS OR (per SD)
PC ~ PRS + Age + PCs

**Analyses performed separately by pop*

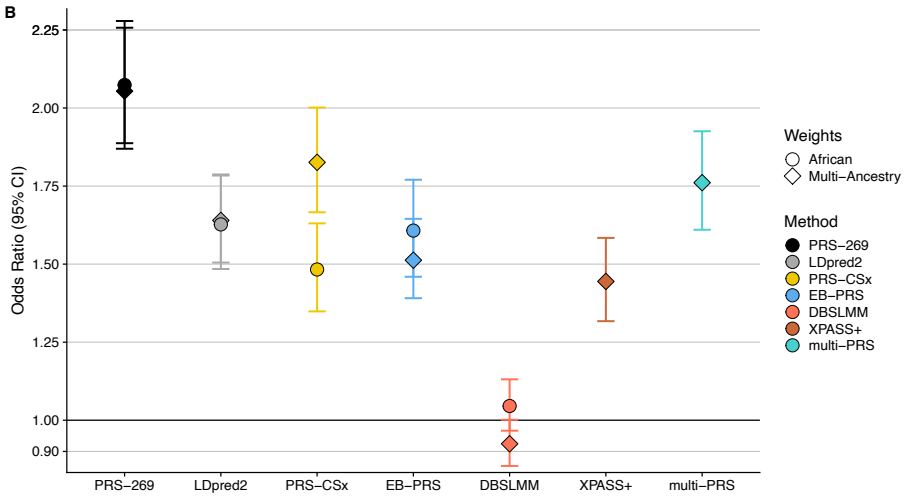
Testing Results: Men of African Ancestry

CA/UG Study: 1,586 Cases & 1,047 Controls

Area Under the Curve (AUC)



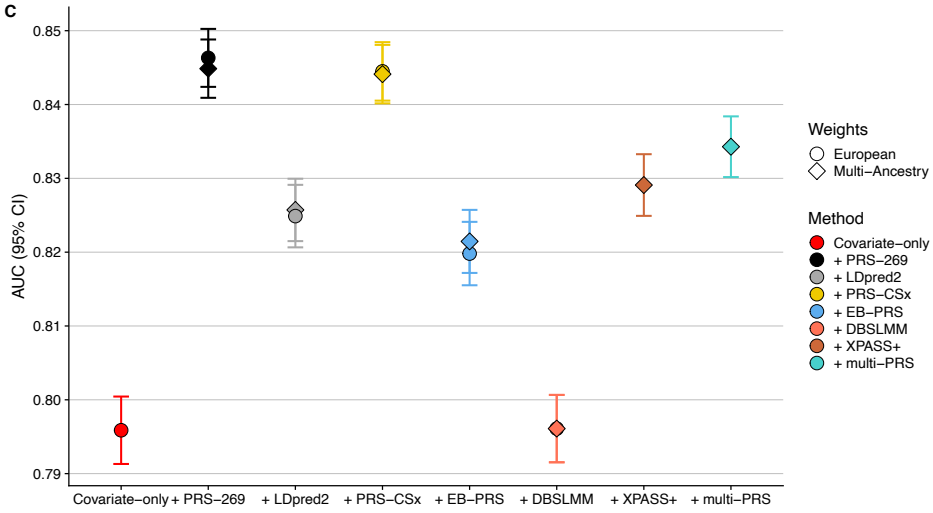
Odds of Prostate Cancer (per PRS SD)



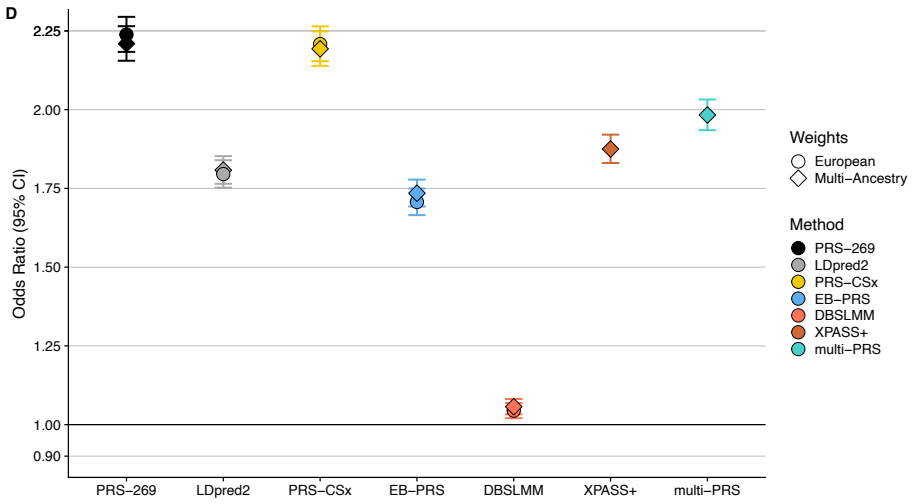
Testing Results: Men of European Ancestry

UK Biobank: 8,046 Cases & 191,825 Controls

Area Under the Curve (AUC)



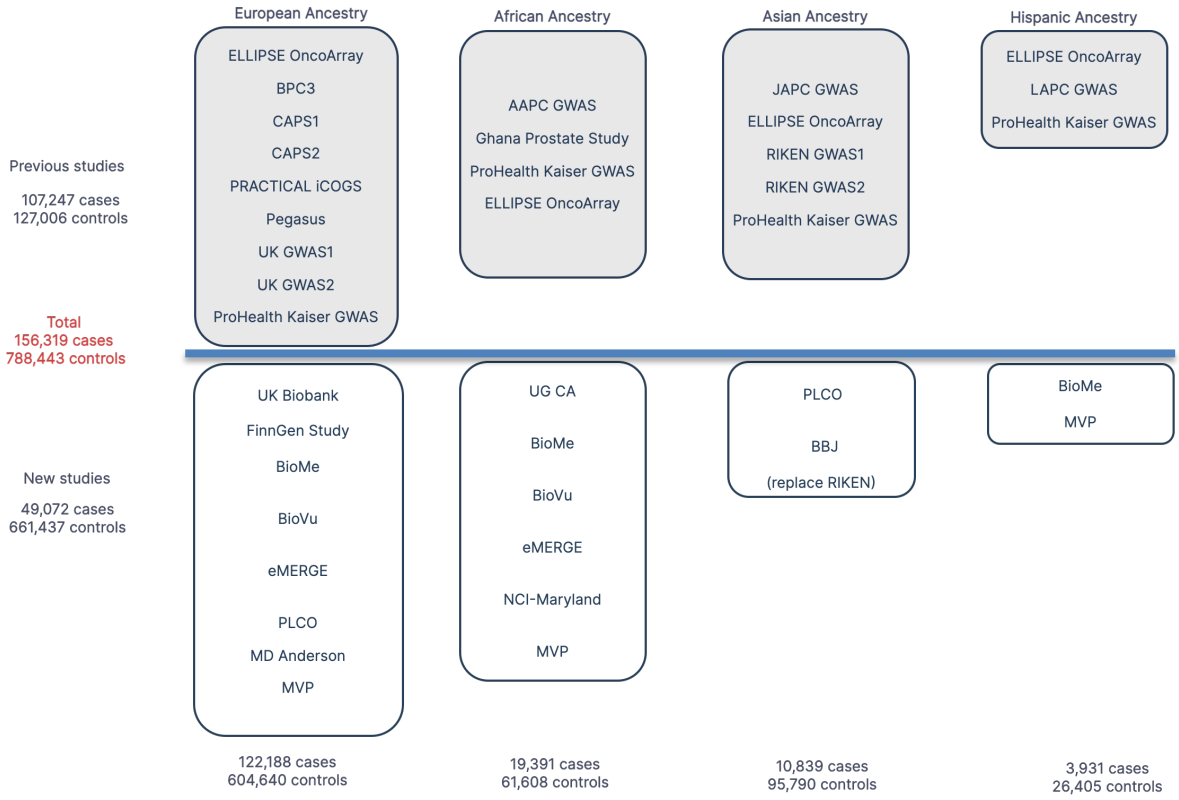
Odds of Prostate Cancer (per PRS SD)



Improvement of the Genetic Risk Score: Larger Sample Size And Diversity

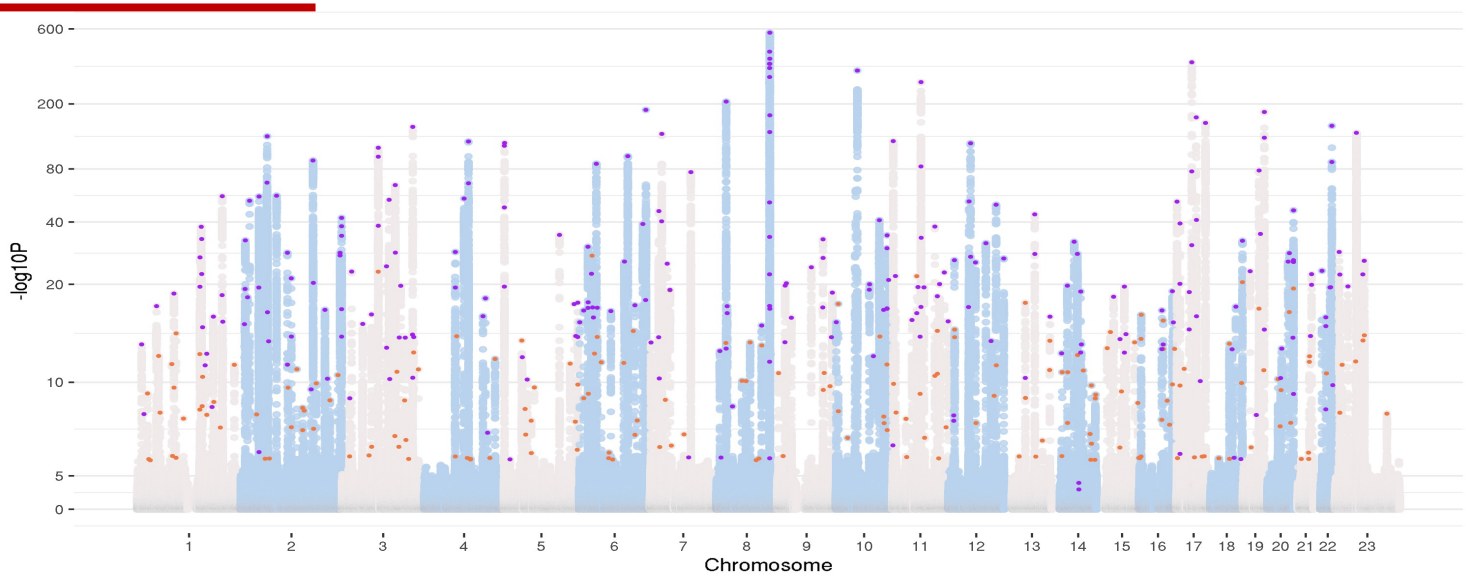
Ancestry-specific and cross-ancestry GWAS in prostate cancer

Studies and sample sizes

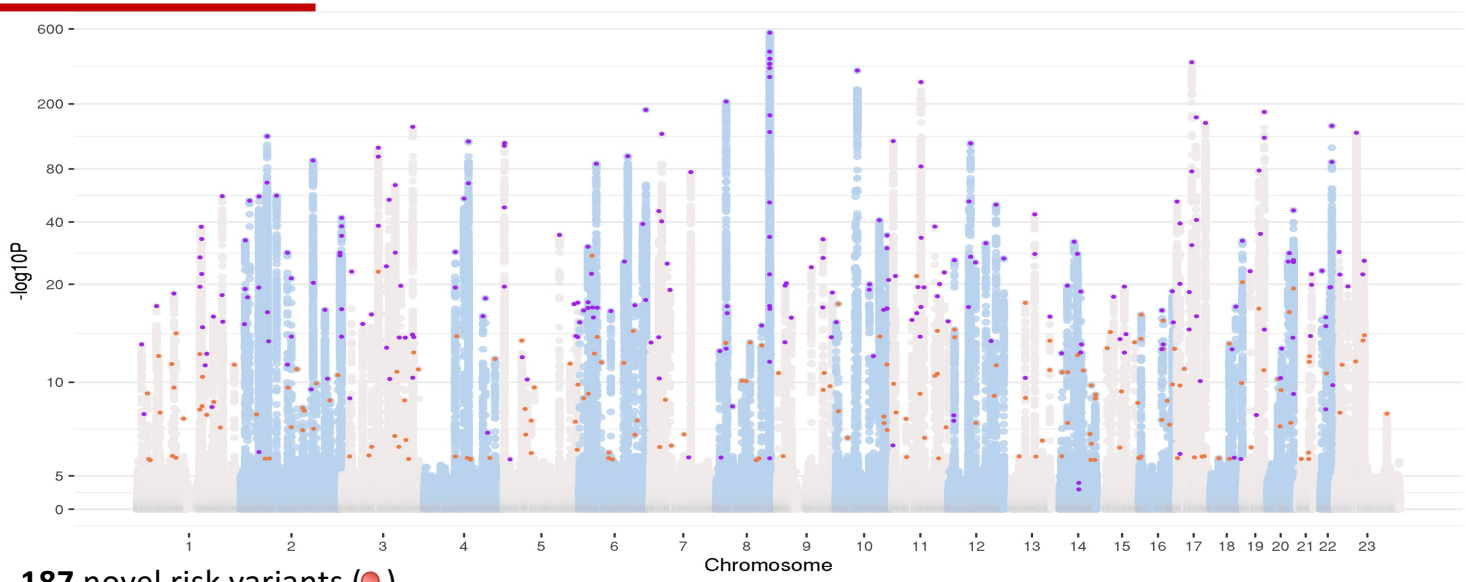


- **11** new studies
- **57%** increase in non-Eur case samples
- **≥128%** effective sample size increase in each population

Latest GWAS Results



451 Prostate Cancer Risk Variants

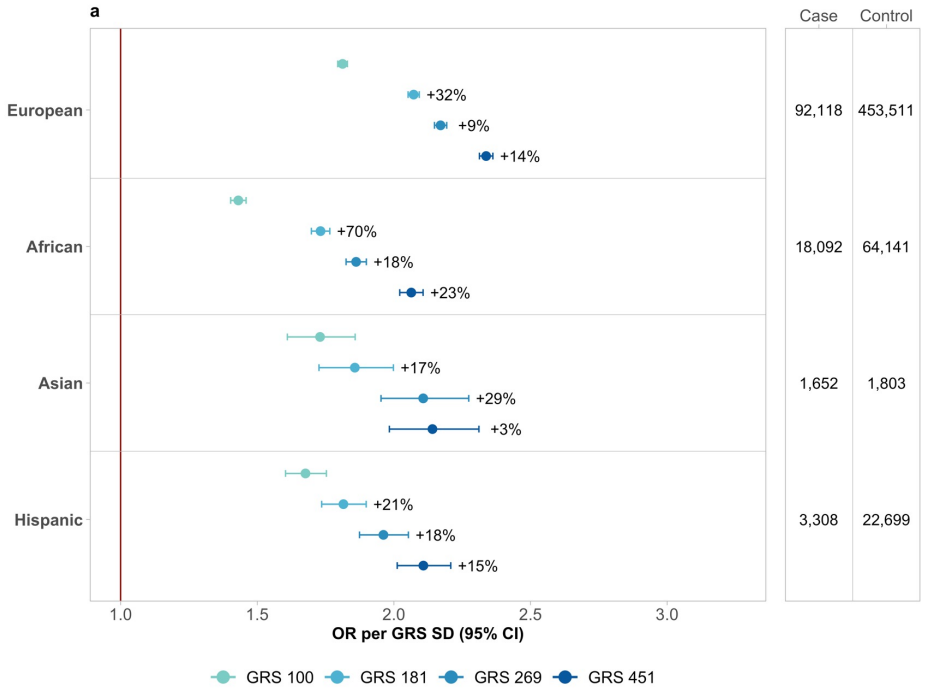


- **187** novel risk variants ( )
  - 61 variants within 800kb of known variants
- **264** known risk variants ( )
  - 150 were replaced by a stronger marker
  - 114 remained as the lead variants in the region
- **18** previously reported variants were dropped

Genetic Risk Score (GRS)
GRS performance improvement with additional risk variants

M	# case/ # control	Ancestry	Source(s)
451	**156,319 / 788,443**	**Multi-ancestry**	**Current meta**
269	107,247 / 127,006	Multi-ancestry	Conti*, Darst* et al., 2021
181	79,194/61,112 10,202 / 10,810 3,000 / 4,394	European African Asian	Schumacher et al. 2018; Conti et al., 2017 Wang et al., 2015
100	43,303 / 43,737	Multi-ancestry	Amin Al Olama et al., 2014

GWAS Discovery samples



Absolute Risk

$$Risk_{Exposed} = Risk_{Baseline} * RR$$

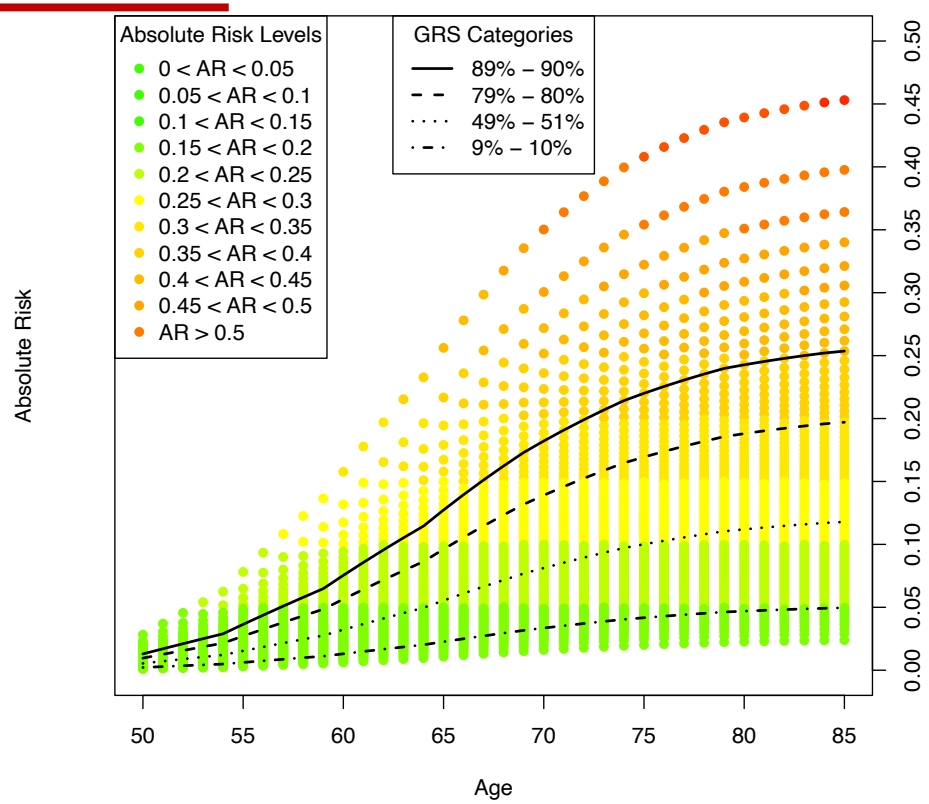
To get the baseline we use the following:

1. Ethnic-Specific Incidence Rates (SEER)
2. Ethnic-Specific Mortality Rates (SEER)

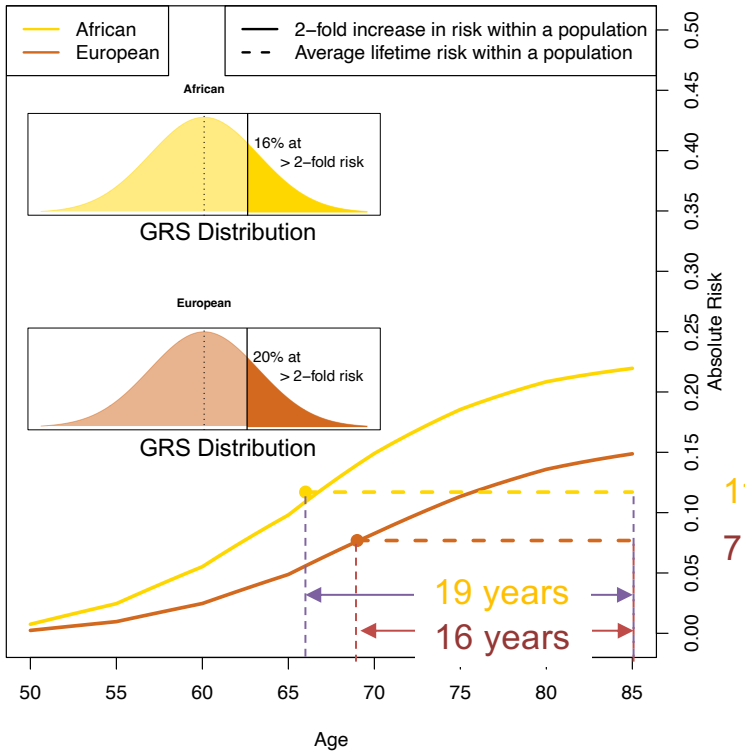
For RR we use PRS risk estimates from the GWAS:

1. Ancestry-specific PRS estimates.
2. Or a single, multi-ancestry RR for PRS (~2.0 per 1SD for prostate cancer).

Cumulative Age-Specific Absolute Risk

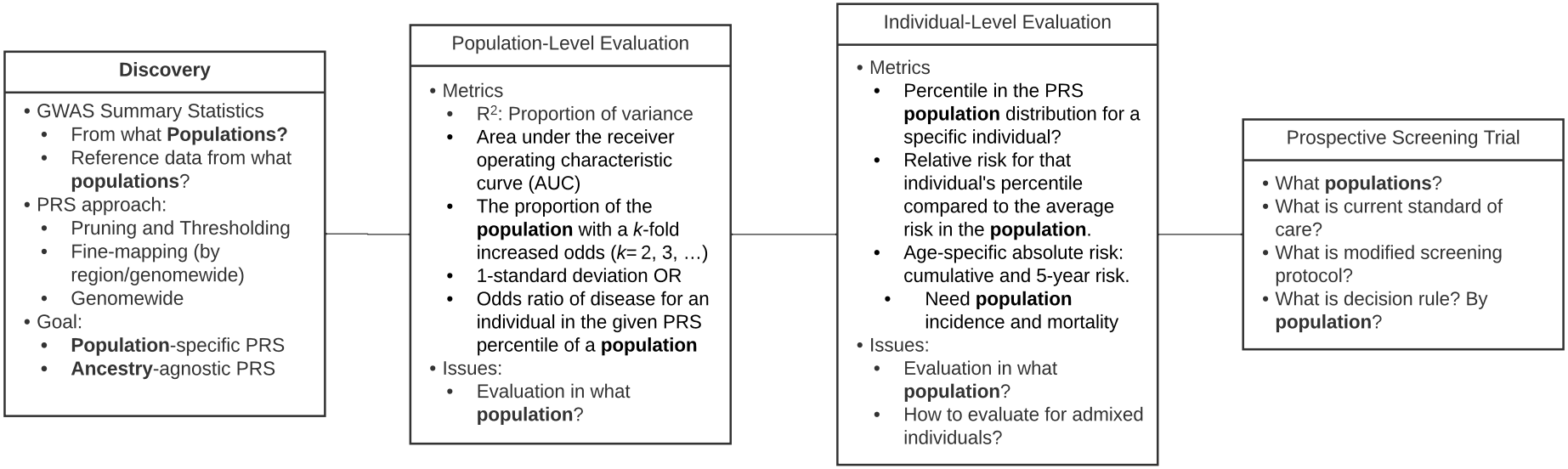


Absolute Risk Assessment



What is the absolute risk for an admixed individual?

Process of Developing a PRS for Screening



How do we define and use “populations” throughout this process?

PRIMED consortium

About Sites and Centers Research Get Involved

What

The NIH-funded Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium is developing and evaluating methods to improve the use of polygenic risk scores (PRS) to predict disease and health outcomes in diverse ancestry populations.

The PRIMED Consortium has the following goals:

1. Gather Diverse Datasets

Bring together large datasets with genomic and health measures from diverse ancestry populations

2. Develop New Methods

Develop new methods to improve genetic risk prediction across diverse populations for a broad range of health and disease outcomes

3. Enable Collaboration

Enable collaborative analysis by sharing PRS-related data, software, and other resources with the scientific community

4. Improve Health

Leverage existing precision medicine partner programs to develop, test, and refine PRS in diverse populations to improve health outcomes

PRIMED Cancer:
Leveraging Diversity in Cancer Epidemiology Cohorts and Novel Methods to Improve Polygenic Risk Scores

- Aim 1: Develop methods to construct polygenic risk scores (PRS) for cancer across multiethnic populations.

- Aim 2: Evaluate polygenic risk scores (PRS) for cancer across multiethnic populations.

- Aim 3: Estimate absolute and excess relative risk of cancer jointly with established risk factors and PRS in multiethnic populations.

PRIMED-Cancer Data

- Prospective cancer cohorts:
  - **Multiethnic Cohort (MEC)**
  - **Kaiser Genetic Epidemiology Research on Adult Health and Aging Cohort (GERA)**
  - **Women's Health Initiative (WHI)**
  - **Nurses Health Studies (NHS)**
  - **Health Professionals Follow-Up Study (HPFS)**
  - **Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)**

- 300,000 individuals (100,000 non-European ancestry).
- 91,000 incident cancer cases (24,000 non-European ancestry).
- Individual-level data from cohorts and cancer site specific consortiums.
- Summary statistics from cancer site specific consortiums.

Acknowledgements

Burcu Darst
Fred Hutch

David Bogumil
Post-Doctoral Trainee

Chris Haiman
Grace Sheng
Lucy Xia
Loreall Pooler
Peggy Wan
Alisha Chou
Raymond Hughley

PRACTICAL

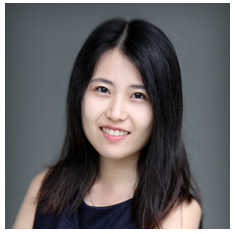
MVP MILLION VETERAN PROGRAM

Amy Justice
Mike Gaziano
Ravi Madduri
Alexis Rodriguez
Janet Tate
Chris Rentsch
Jennifer Huffman
Kelly Cho
MVP Data Core
and others..

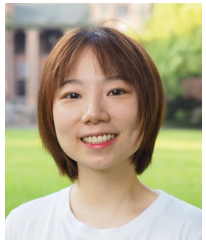
African Ancestry Prostate Cancer Consortium (AAPC)

Bill Blot
Stephen Chanock
Sue Ingles
Sonja Berndt
Janet Stanford
Rick Kittles
William Isaacs
Edward Yeboah
Wei Zheng
John Carpten
Adam Kibel
Timothy Rebbeck
Benjamin Rybicki
Eric Klein
John Witte
Jeannette Bensen
Esther John
Stephen Van Den Eeden
Jay Fowke
Stephen Watya
Luc Multigner
Marie-Elise Parent
Florence Menegaux
Geraldine Cancel-Tassin
Laurent Brureau

Fei Chen
Asst. Prof.



Anqi Wang
PhD Candidate



Sylvia Shen
PhD Candidate

Ros Eeles
Zsofia Kote-Jarai
Ed Saunders
Tokhir Dadaev
Mark Brooks
Hidewaki Nakagawa
Fredrik Wiklund
Graham Giles
Robert MacInnis
>200 collaborators

Prostate Cancer Foundation
Curing Together.

MADCaP NETWORK

biobank uk
Enabling scientific discoveries that improve human health

Funding:
PCF Challenge Award (20CHAS03); RESPOND, NCI/NIMHD, Cancer Moonshot; U19 CA214253, NCI U19 GAME-ON/ELLIPSE, U19 CA148537; AAPC/Others: R01 CA196931, R01 CA165862, U01 CA164973, RC2 CA148085, U01 CA1326792, U01 HG004726, R01 CA063464

USCIMAGE
Integrative Methods of Analysis for Genetic Epidemiology

RESPOND
African American Prostate Cancer Study

USC Center for Genetic Epidemiology