Novel strategy for disease risk prediction incorporating predicted gene expression and DNA methylation: a multi-phased study of prostate cancer

Chong Wu
Assistant Professor
Department of Statistics

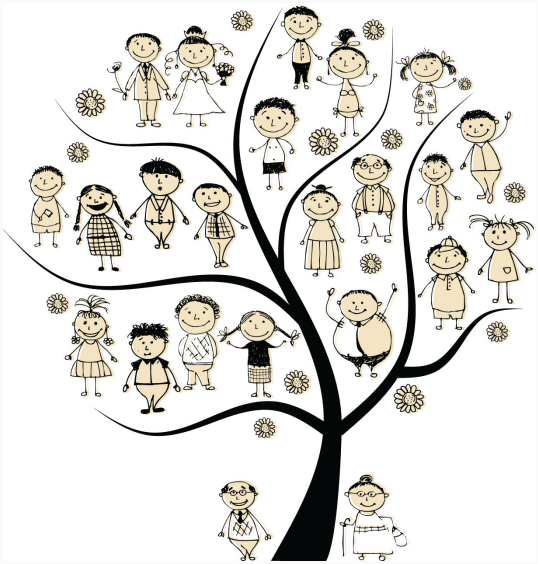
IISA 2021

May 22, 2021

Outline

- Why studying polygenic risk score?
- Existing methods
- Novel method
- Study design and Results
- Discussion

Why studying polygenic risk score?



copyright @ John Gray Center

- When we visit a new doctor, they will ask about family history
- Family history helps doctors get a sense of our risk of developing those diseases
- Family history does not tell everything about your genetic risk

Why studying polygenic risk score?

- Sometimes, mutation in a single gene can greatly increase you risk for a given disease
- Breast cancer: *BRCA1* and *BRCA2*
- Late onset Alzheimer's disease: *APOE*

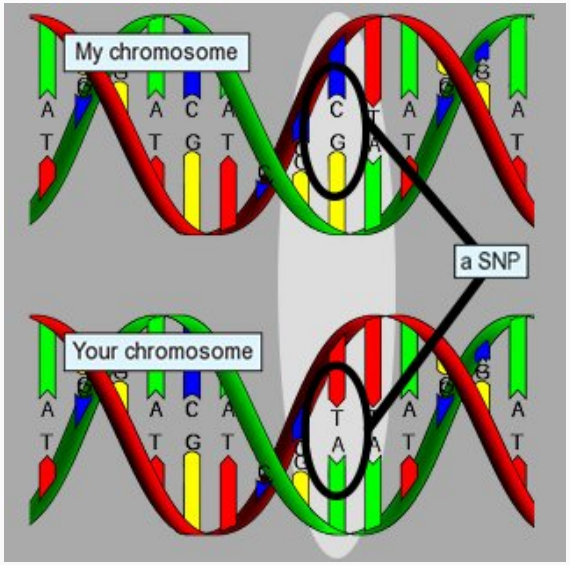
Why studying polygenic risk score?

- Infinitesimal model: a large number of small-effect common variants across the entire allele frequency spectrum
- The more you inherited, the greater you risk
- Can not be fully explained by family history information

Polygenic risk score:

Aggregate genetic information across all the genome

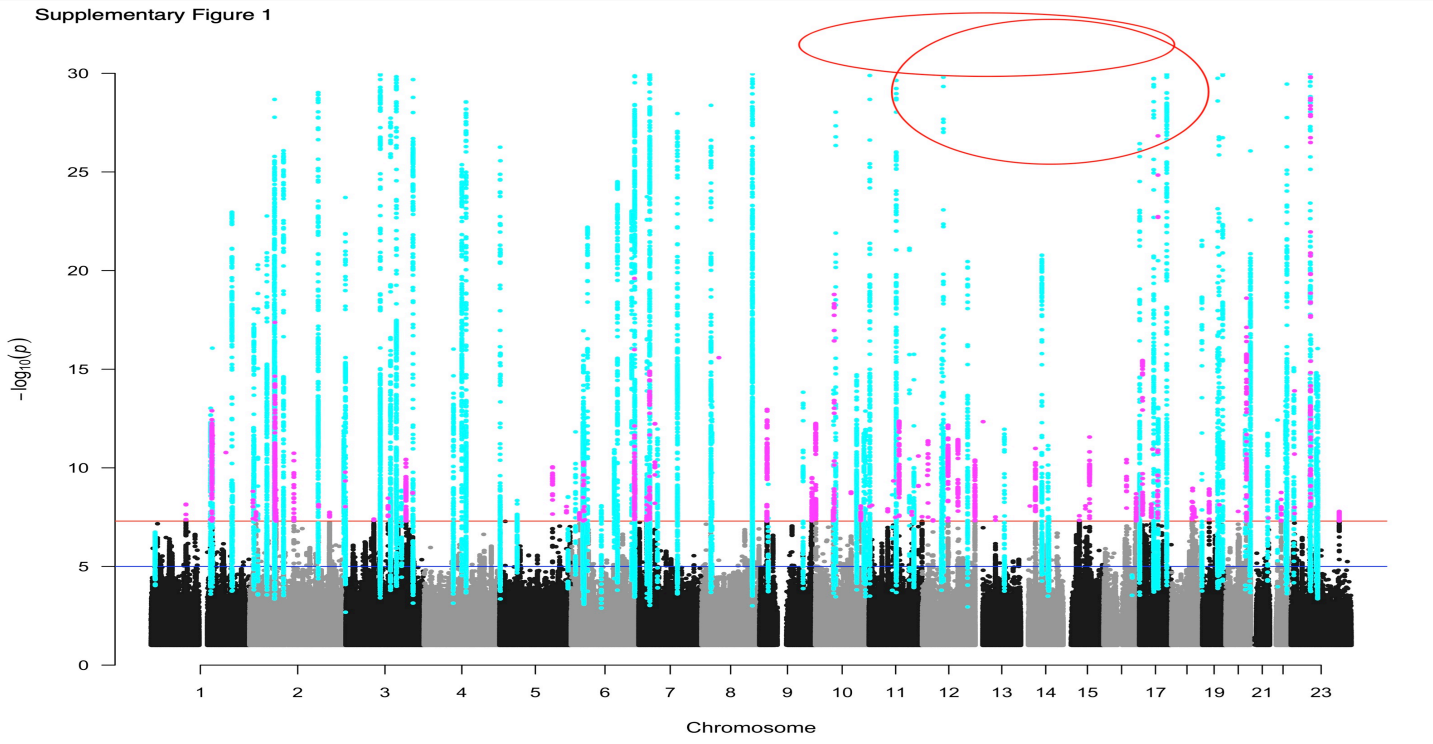
Genome-wide association study (GWAS)



copyright @ John Fouts (2016)

- Genome: the set of genetic information encoded in 23 chromosome pairs
- SNP: Variation in a single base pair
- Genetic score (additive) for each SNP and a person:
  AA = 0, AB = 1, BB = 2
- Associated SNPs are not necessarily causal

Genome-wide association study (GWAS)



Nature Genetics volume 50, pages 928–936 (2018)

- Prostate cancer; more than 140,000 men
- Run marginal regression for each SNP

Outline

- Why studying polygenic risk score?
- Existing methods
- Novel method
- Study design and Results
- Discussion

Challenges

- Effect size of associated SNPs is really small
- They are correlated

Genetic prediction: polygenic risk score (PRS)

- $\text{PRS}_k = \sum_j \beta_j x_{kj}$ (summation of the effects from all GWAS SNPs), where
  - PRS_k: PRS for sample k;
  - β_j: effect size for SNP j;
  - x_{kj}: genotype for SNP j, sample k
- Because of much larger studies and improved algorithms, PRS shows very promising results (will show later)

PRS: P + T method

- Pruning plus thresholding
- $\text{PRS}_k = \sum_i \hat{\beta}_i x_{ik}$
- Frist, $\hat{\beta}_i$ may be very noisy. Use hard threshold. For example, only use $\hat{\beta}_i$ with $p < 0.01$.
- Second, SNPs from a similar location are highly correlated. Use pruning to preserve independent signals.

PRS: LassoSum & LDpred

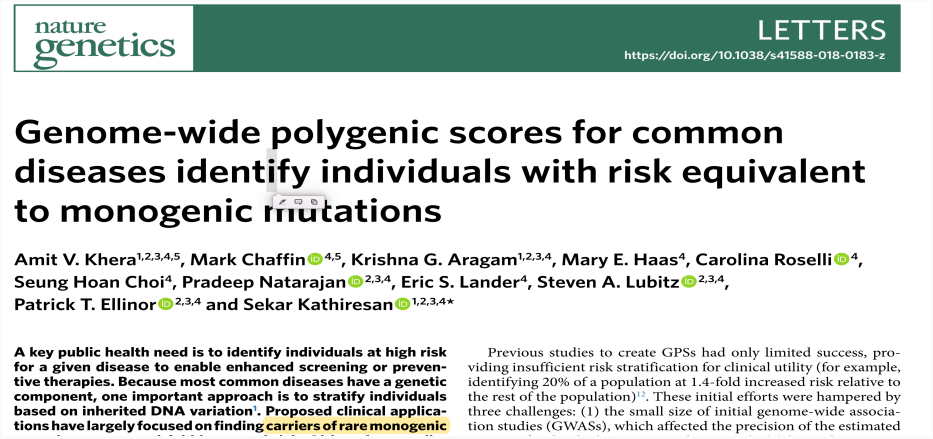
- P + T does not take correlation among genetic variants into account;
- LassoSum: penalized regression with LASSO
- LDpred: a Bayesian way to deal with correlations

PRS: Incorporating functional information

- AnnoPred
- LDpred-funct
- EBPRS

There is debate for the utility of PRS

- Many studies show that PRS are extremely useful.



nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0183-z

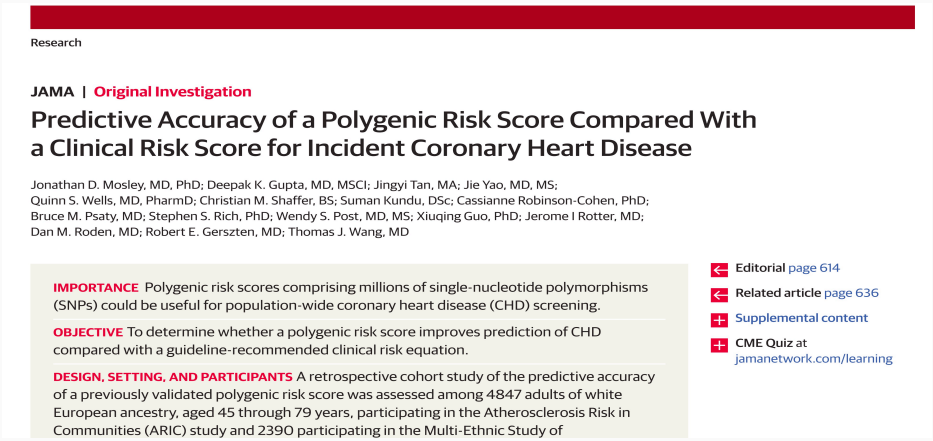
Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4]*

A key public health need is to identify individuals at high risk for a given disease to enable enhanced screening or preventive therapies. Because most common diseases have a genetic component, one important approach is to stratify individuals based on inherited DNA variation[1]. Proposed clinical applications have largely focused on finding carriers of rare monogenic

Previous studies to create GPSs had only limited success, providing insufficient risk stratification for clinical utility (for example, identifying 20% of a population at 1.4-fold increased risk relative to the rest of the population)[12]. These initial efforts were hampered by three challenges: (1) the small size of initial genome-wide association studies (GWASs), which affected the precision of the estimated

- Some show that the utility of PRS may be minimial



Research

JAMA | Original Investigation

Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease

Jonathan D. Mosley, MD, PhD; Deepak K. Gupta, MD, MSCI; Jingyi Tan, MA; Jie Yao, MD, MS; Quinn S. Wells, MD, PharmD; Christian M. Shaffer, BS; Suman Kundu, DSc; Cassianne Robinson-Cohen, PhD; Bruce M. Psaty, MD; Stephen S. Rich, PhD; Wendy S. Post, MD, MS; Xiuqing Guo, PhD; Jerome I Rotter, MD; Dan M. Roden, MD; Robert E. Gerszten, MD; Thomas J. Wang, MD

IMPORTANCE Polygenic risk scores comprising millions of single-nucleotide polymorphisms (SNPs) could be useful for population-wide coronary heart disease (CHD) screening.

OBJECTIVE To determine whether a polygenic risk score improves prediction of CHD compared with a guideline-recommended clinical risk equation.

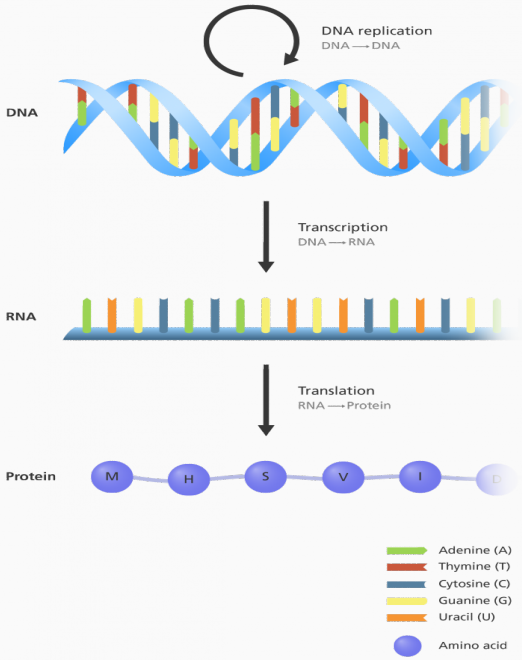
DESIGN, SETTING, AND PARTICIPANTS A retrospective cohort study of the predictive accuracy of a previously validated polygenic risk score was assessed among 4847 adults of white European ancestry, aged 45 through 79 years, participating in the Atherosclerosis Risk in Communities (ARIC) study and 2390 participating in the Multi-Ethnic Study of

Editorial page 614
Related article page 636
Supplemental content
CME Quiz at jamanetwork.com/learning

Outline

- Why studying polygenic risk score?
- Existing methods
- **Novel method**
- Study design and Results
- Discussion

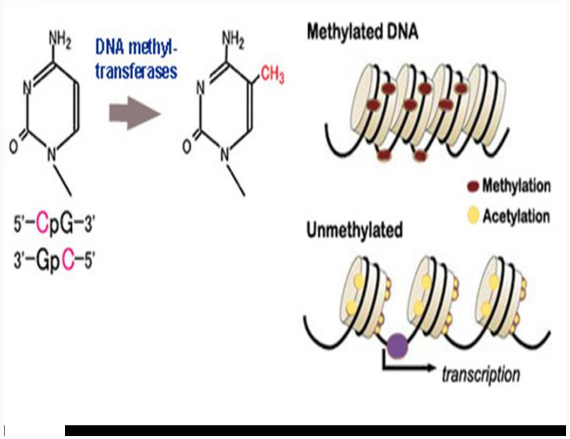
Novel method



copyright @ yourgenome.org

- Central dogma
- **Our idea:** Instead of using SNPs as predictors, we can use gene expression levels as predictors
- Gene expression heritability is **everywhere** (GTEx 2017 *Nature*)

Novel method



copyright @ Helicase

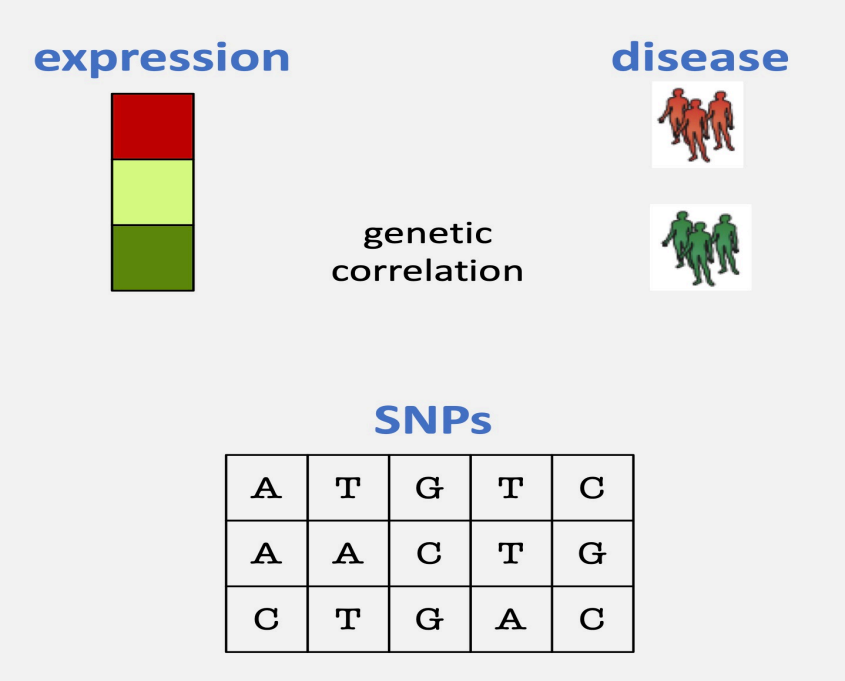
- DNA methylation: methyl groups are added to the DNA molecule; epigenetic mechanisms
- Modify the function of genes and affect gene expression
- Have a very strong prediction power: smoking

Novel method

- Idea: because DNA methylation and gene expression play a vital role in the etiology of a disease, we can use DNA methylation and gene expression as predictors
- Challenge 1: unlike GWAS with a very large sample size, the sample size for gene expression and DNA methylation data is very small
- Challenge 2: gene expression and DNA methylation are affected by environmental factors; very hard to adjust all confounding factors
- Solution: we can use genetically imputed gene expression/DNA methylation as predictors

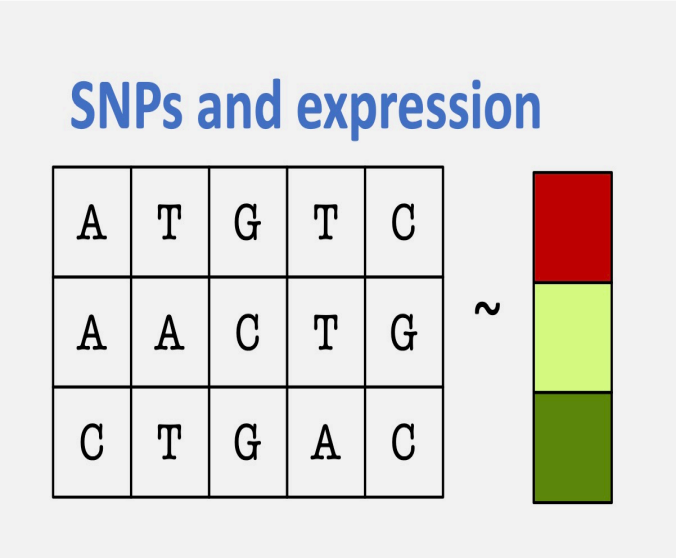
TWAS/PrediXcan idea review

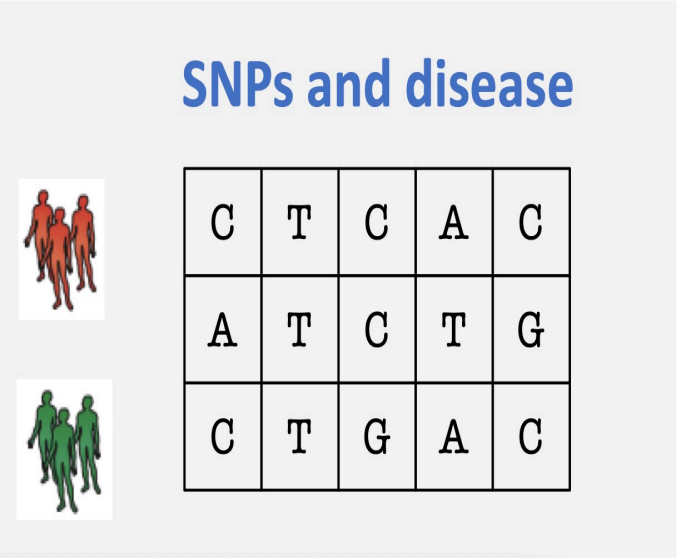
We want: test the association between gene expression and disease



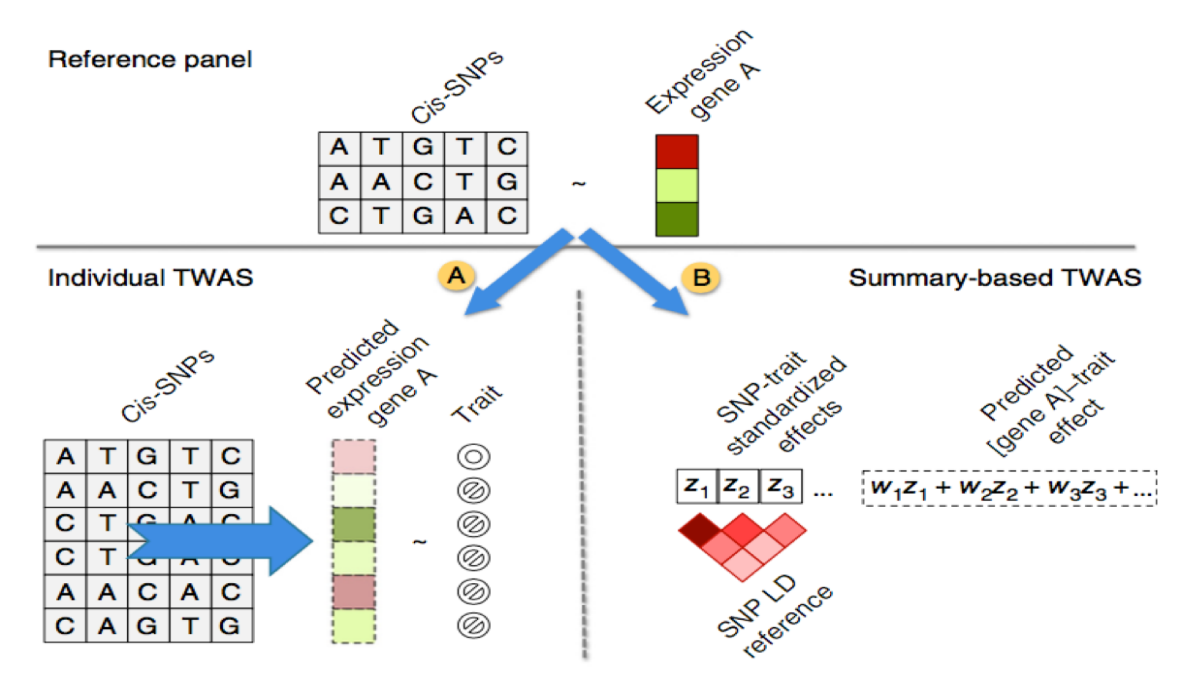
TWAS/PrediXcan idea review

We have



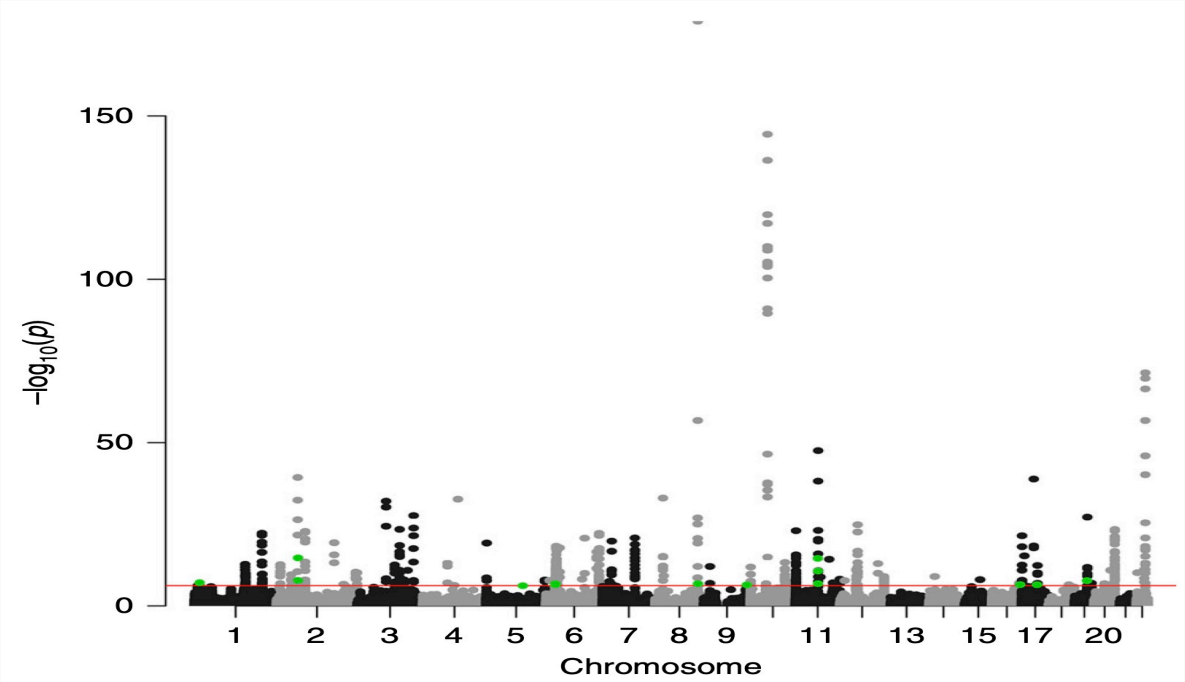


TWAS/PrediXcan idea review



- Step 1: Build gene expression prediction models by using a reference panel
- Step 2: Test the association between predicted expression levels and trait

TWAS can be applied to DNA methylation data

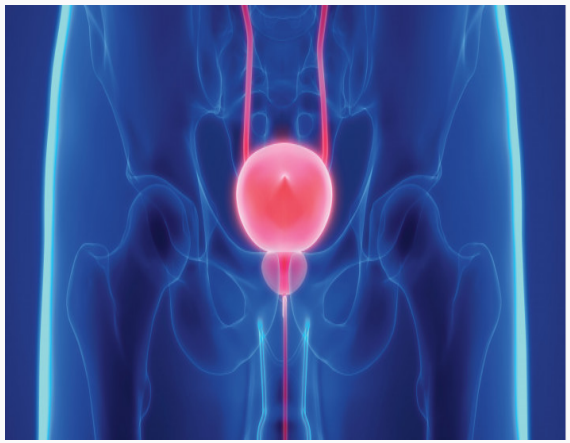


A Manhattan plot of the association results from the prostate cancer methylome-wide association study using S-PrediXcan

Outline

- Why studying polygenic risk score?
- Existing methods
- Novel method
- **Study design and Results**
- Discussion

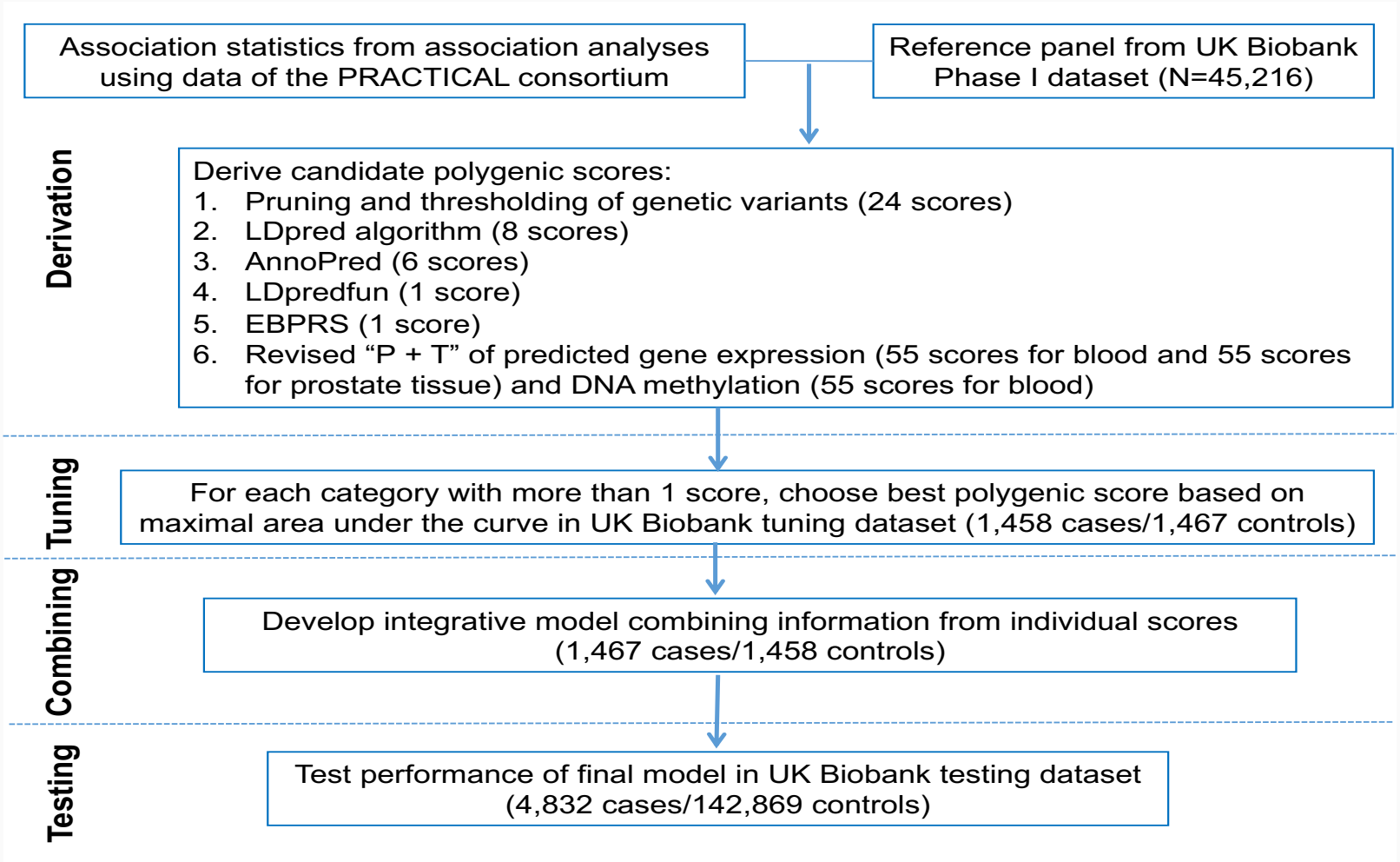
Prostate cancer



copyright @ Cancer health

- The second most commonly diagnosed malignancy in men worldwide
- Prostate-specific antigen (PSA) has been used widely for PrCa screening
- PSA screening is controversial
- Prostate cancer is highly heritable

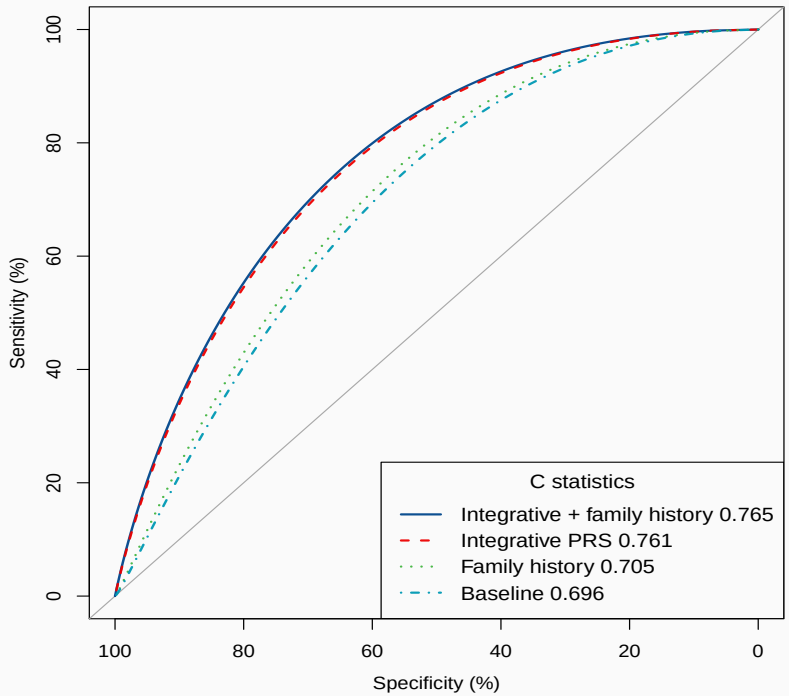
Study design



Setup

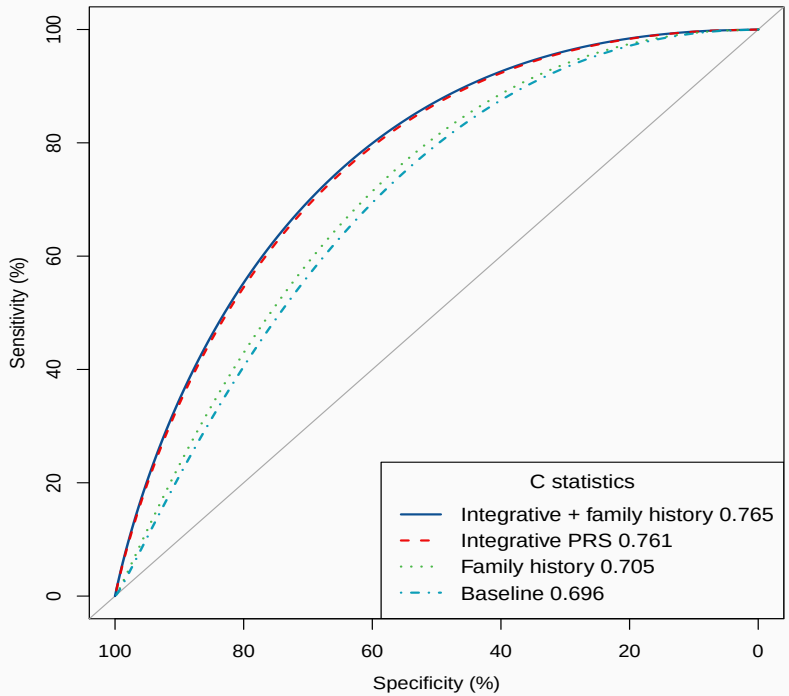
- Our primary analyses focused on incident PCa events
- Use Cox proportional hazards model
- Baseline model: Age + top 4 PCs of genotype matrix (approximate population structure) + genotype array

Results



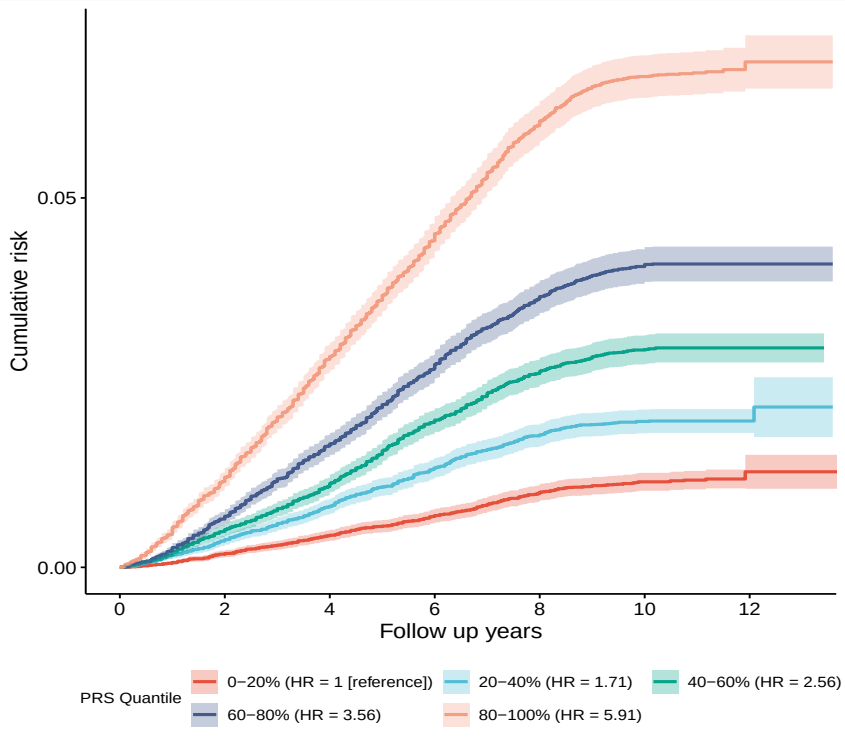
- C statistic is a rank-order statistic for predictions against true outcomes; from 0.5 (no discrimination) to 1.0 (perfect discrimination)

Results: Methods comparison

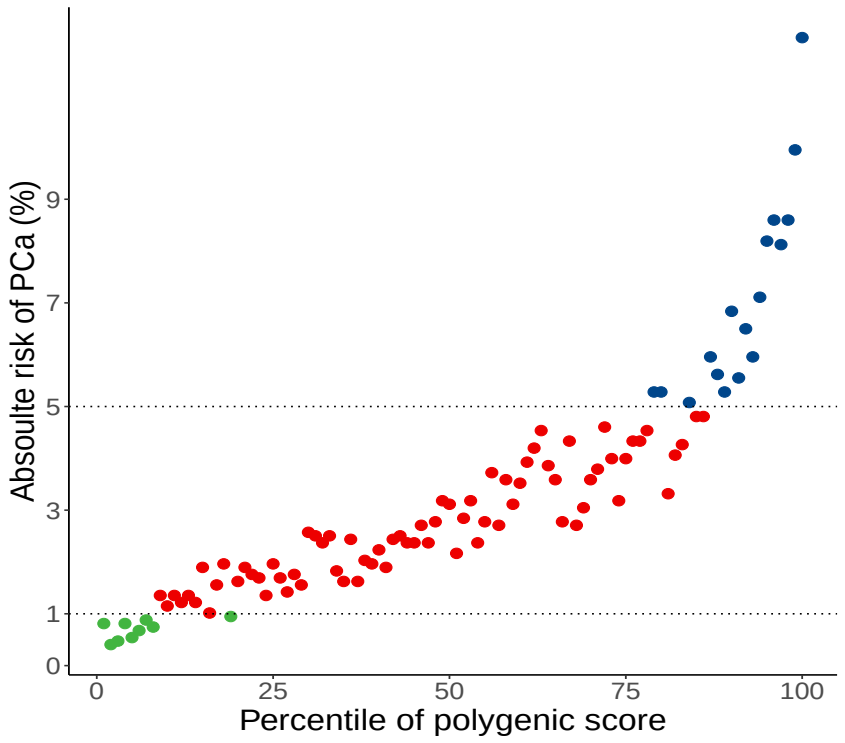


- C statistic is a rank-order statistic for predictions against true outcomes; from 0.5 (no discrimination) to 1.0 (perfect discrimination)

Results: Survival analysis



Results: PRS can identify individuals at risk



- Absolute risk: 0.6% in the lowest percentile to 8.8% in the highest percentile

Results: compare with family history

- predicted ten-year risk changed by more than 1% for 44.5% of participants, and changed by 5% or more for 6.4% of participants
- The overall net reclassification improvement (NRI) was 69.0% (95% CI, 64.9% to 70.5%)
- In comparison, when family history was added to the baseline model, predicted ten-year risk changed by more than 1% for 5.3% of participants, and changed by 5% or more of 0.15% participants.
- The increase in risk difference between cases and noncases (overall NRI) was 12.5% (95% CI, 11.3% to 16.5%)

Outline

- Why studying polygenic risk score?
- Existing methods
- Novel method
- Study design and Results
- Discussion

Discussion

- While appealing, some studies raised concerns for the clinical utility of such PRS; CAD
- Our newly developed PRS can have significantly higher risk assessment power than family history

Discussion

- Our developed PRS could potentially bring multiple opportunities for reducing the public health burden of PCa
  - For males with a PRS within the bottom 50% range, their absolute risk is lower than 1.8%
  - men with a PRS within the top 5% range has an absolute risk of 6.7%

Limitations

- UK Biobank are known to be healthier than the general population
- Focus on European population
- We only focus on incorporating imputed gene expression and DNA methylation in blood

Acknowledgement

- Another corresponding author: Lang Wu (at Hawaii)
- Co-authors: Jingjing Zhu, Xiaoran Tong, Qing Lu, Jong Y Park, Liang Wang, Guimin Gao, Hong-Wen Deng, Yaohua Yang, Karen E Knudsen, Timothy R Rebbeck, Jirong Long, Wei Zheng, Wei Pan, David V Conti, Christopher A Haiman
- UK Biobank recourse (application numbers 48240 and 53866)
- The PRACTICAL, CRUK, BPC3, CAPS, PEGASUS consortia for making the prostate cancer GWAS summary statistics publicly available
- many grants in NIH

Thank you!