
بسم الله الرحمن الرحیم

جمهوری اسلامی ایران
وزارت بهداشت، درمان و آموزش پزشکی

دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهید بهشتی

پژوهشکده علوم غدد درون ریز و متابولیسم

ژمیران

ژنوم مرجع ایرانیان

Iranian Reference Genome Project (IRGP)

Research Institute for Endocrine Sciences, SBMU

The 14TH International Congress of Endocrine Disorders (ICED14)

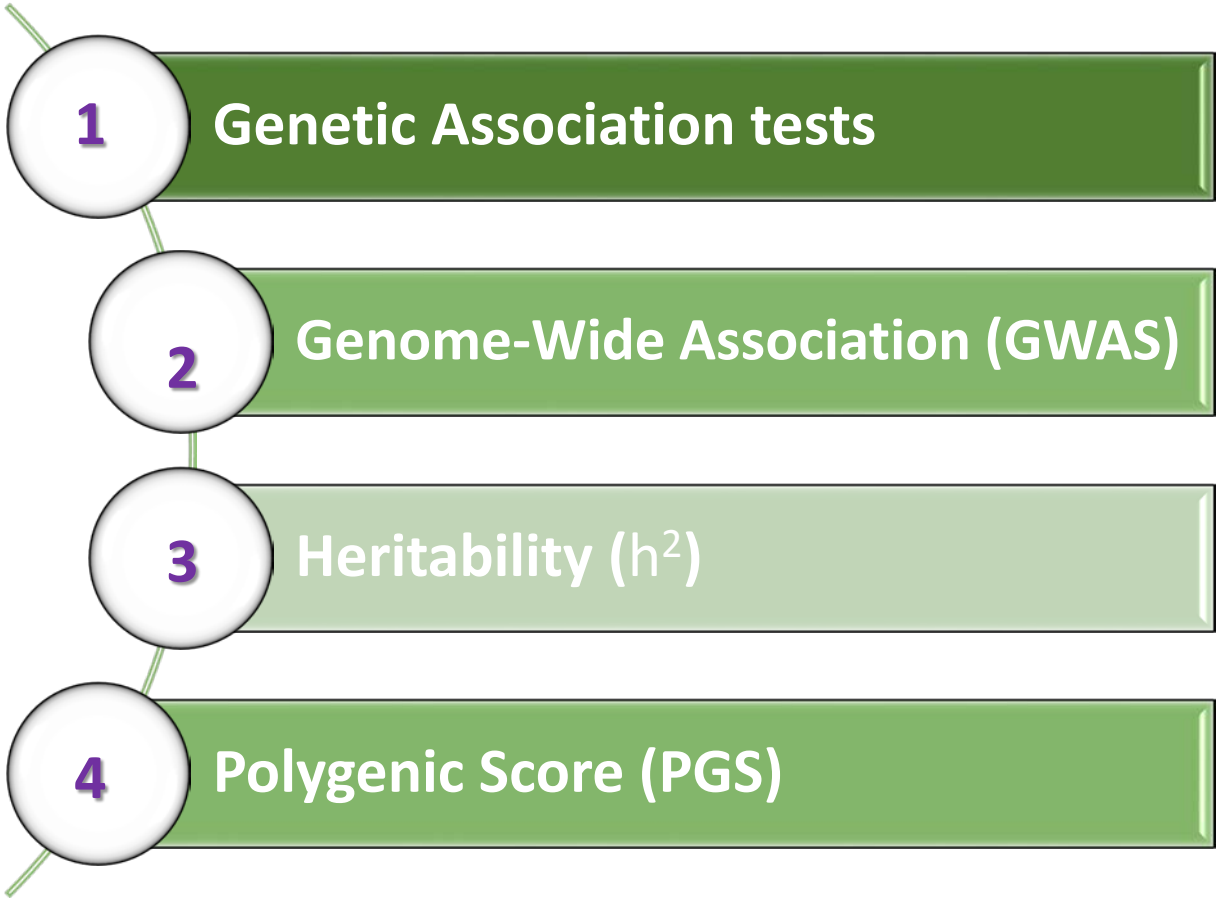
Symposium 9: Precision Medicine

Genome-Wide Association (GWAS) and Polygenic Score (PGS) on Type 2 Diabetes in the TCGS

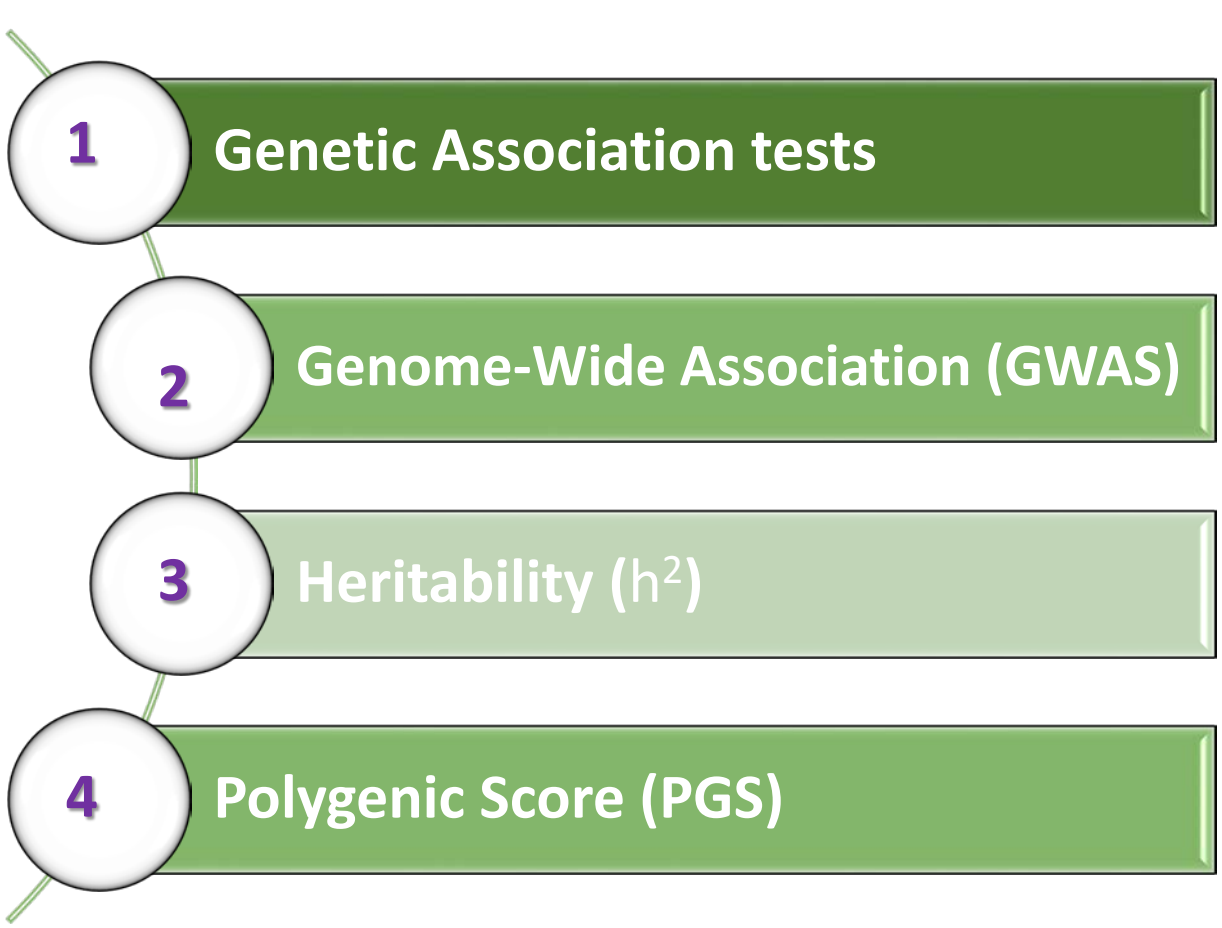
Mahdi Akbarzadeh

Assistant Professor, Biostatistics (Statistical Genetics)
Shahid Beheshti University of Medical Sciences (SBMU)
Research Institute for Endocrine Sciences (RIES)
Cellular and Molecular Endocrine Research Center (CMER)

Outline


1
Genetic Association tests
2
Genome-Wide Association (GWAS)
3
Heritability (h^2)
4
Polygenic Score (PGS)

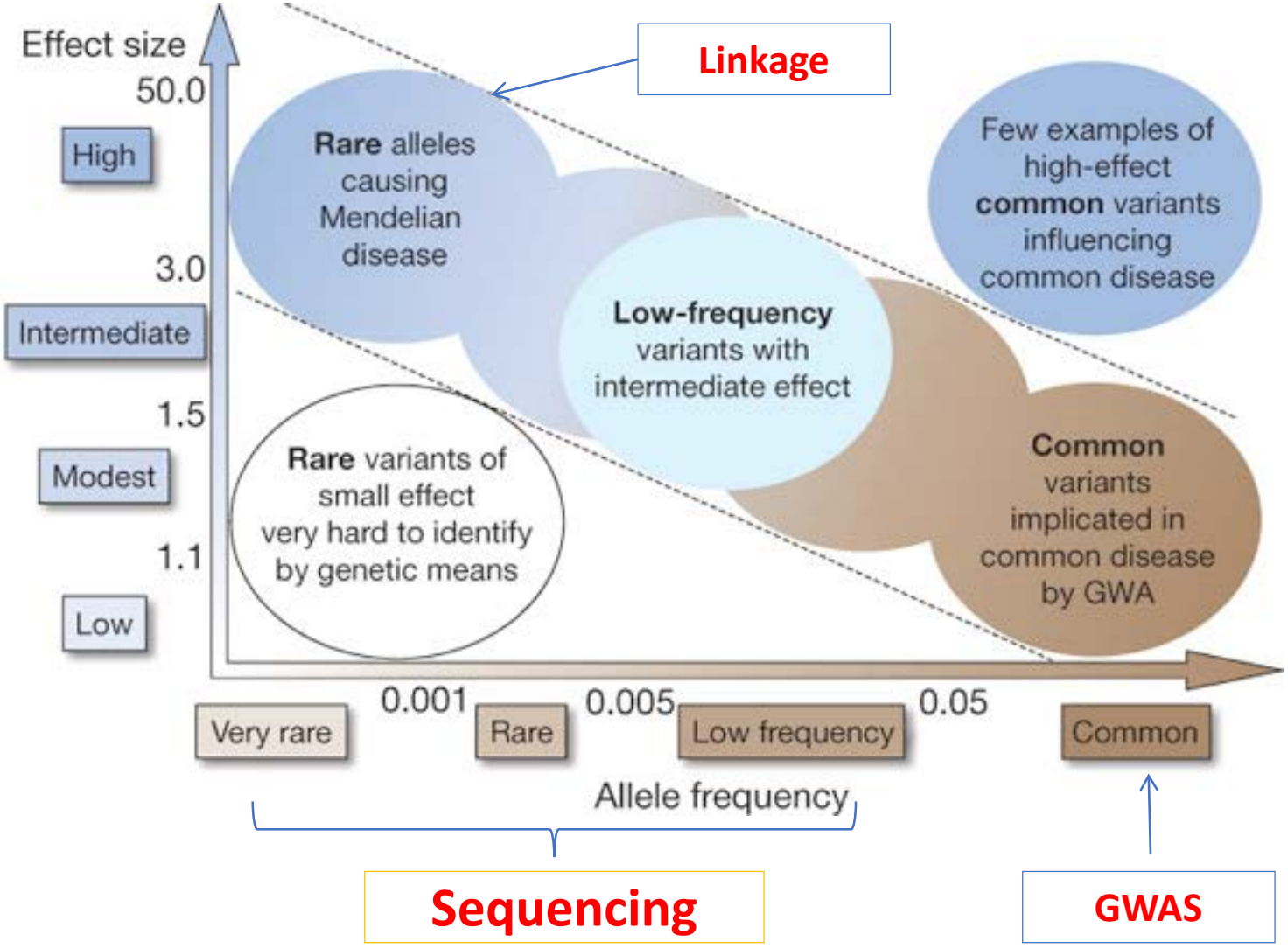
Outline



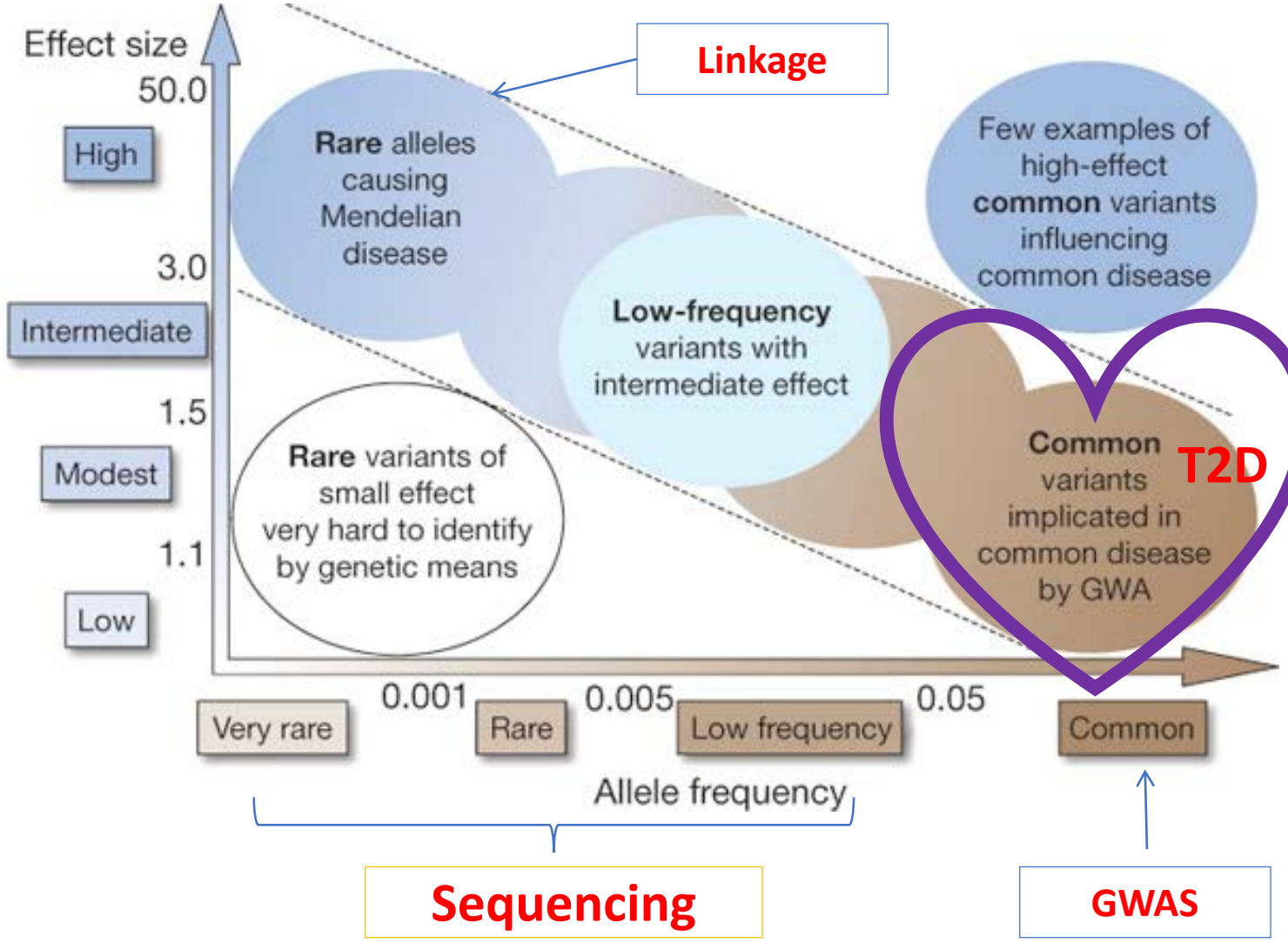
Defining T2D in TCGS

Results of these methods in in Iranians

T2D is as a multifactorial disorder

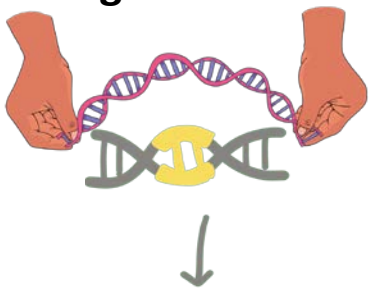


T2D is as a multifactorial disorder

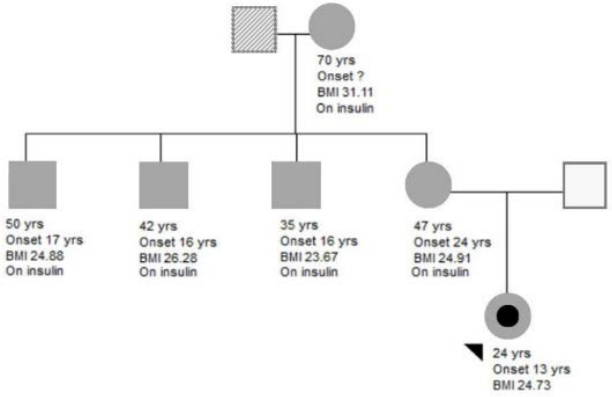


Preface

Monogenic disorders

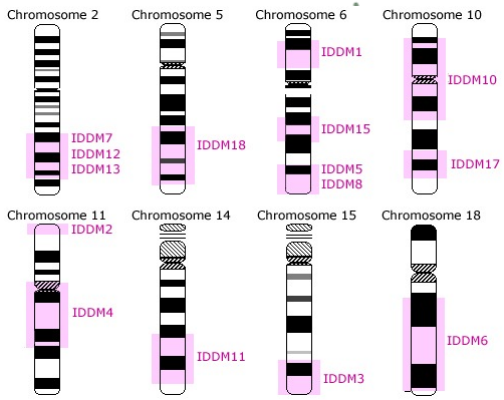


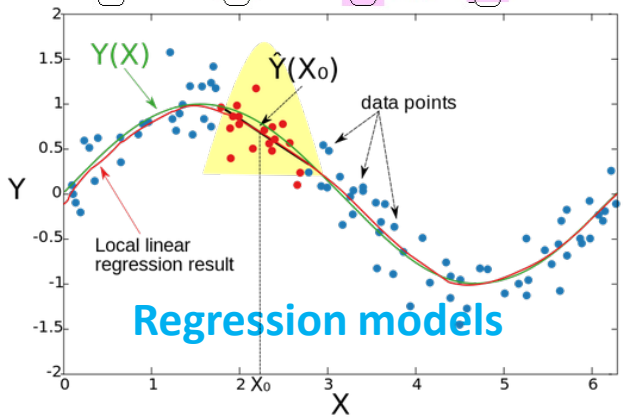
MODY



Simple statistical methods

Oligogenic disorders

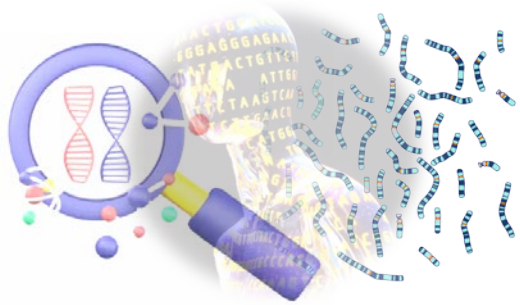


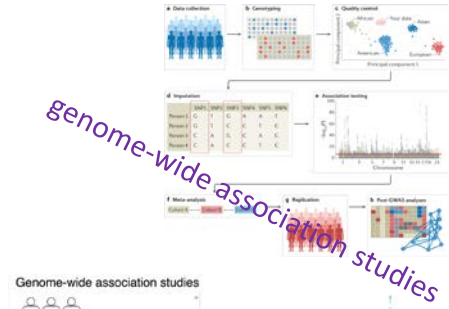


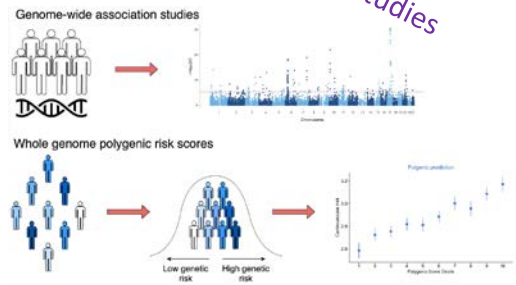
Regression models

Common statistical methods

Polygenic disorders







Modern statistical methods

Genetic Association tests

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

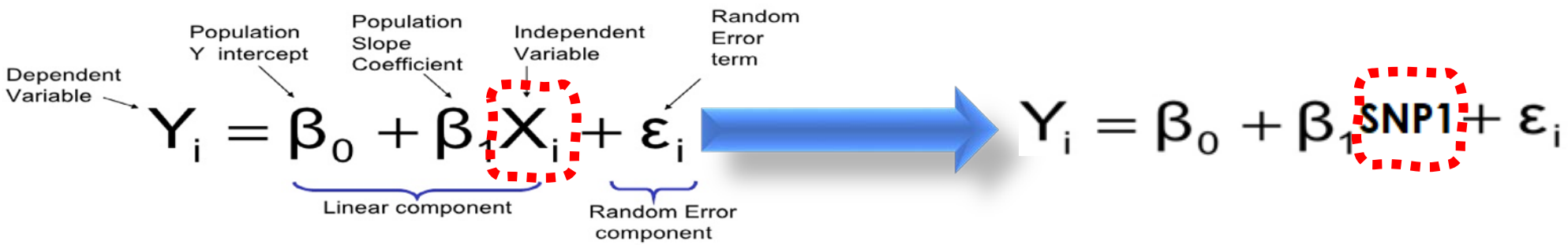
Linear component

Random Error component

Genetic Association tests

	Chrom.	DNA sequence	Single Nucleotide Polymorphisms (SNPs) SNP1	SNP2
Person 1	Mat	GTAACTTGGGATCTAGACCAGATAGAT	A A	G G
	Pat	GTAACTTGGGATCTAGACCAGATAGAT		
Person 2	Mat	GTAACTTGGGATCTAGACCAGATAGAT	A C	G G
	Pat	GTAACTTGGGATCTCGACCAGATAGAT		
Person 3	Mat	GTAACTTGGGATCTCGACCAGATAGAT	C C	G T
	Pat	GTAACTTGGGATCTCGACCATATAGAT		

SNP1 SNP2



Genome-Wide Association Studies (GWAS)

Introduction to GWAS

Genome-wide association studies (GWAS) aim to identify associations of genotypes with phenotypes by testing for differences in the allele frequency of genetic variants between individuals who are ancestrally similar but differ phenotypically.

Genome-wide association studies (GWAS):

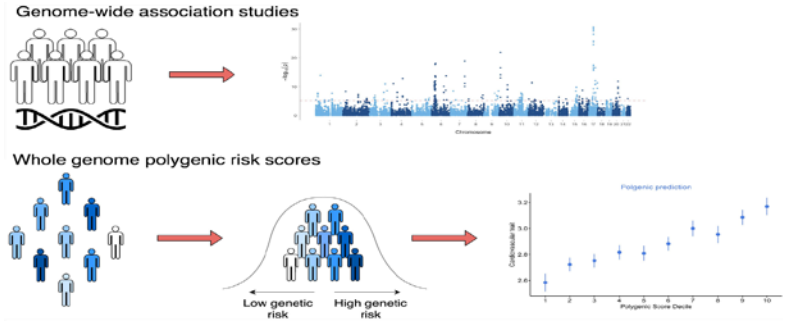
1- Test hundreds of thousands of genetic variants across many genomes to find those statistically associated with a specific trait or disease.

2- Has generated a myriad of robust associations for a range of traits and diseases.

3- Number of associated variants is expected to grow steadily as GWAS sample sizes increase.

4- GWAS results have a range of applications

- Gaining insight into a phenotype's underlying biology
- Estimating its heritability,
- Calculating genetic correlations
- Making clinical risk predictions
- Informing drug development programmes
- Inferring potential causal relationships between risk factors and health outcomes.



Introduction to GWAS

Poor quality data → **false positives / negatives**

Steps in QC (plink commands)

1. **Sex-check (chr X heterozygosity) (--check-sex)**
2. **Genotyping Call Rate (SNPs missing individuals) (--geno)**
3. **Hardy-Weinberg Equilibrium (--hwe)**
4. **Minor Allele Frequency (--maf)**
5. **Sample Call Rate (individuals missing genotypes) (--mind)**
6. **Proportion of Heterozygosity (--het)**
7. **Relatedness (--genome)**
8. **Population Structure / Stratification (--cluster-mds-plot)**

*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

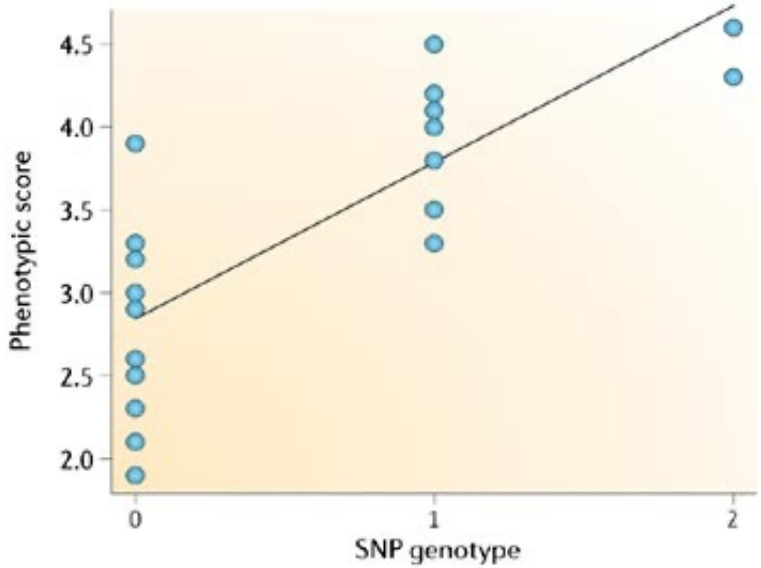
Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Marees et al. *Int J Methods Psychiatr Res*. 2018 Jun; 27(2): e1608.

Introduction to GWAS: Quantitative Trait

Linear Regression

$$\hat{Y} = \alpha + \beta X + \varepsilon$$



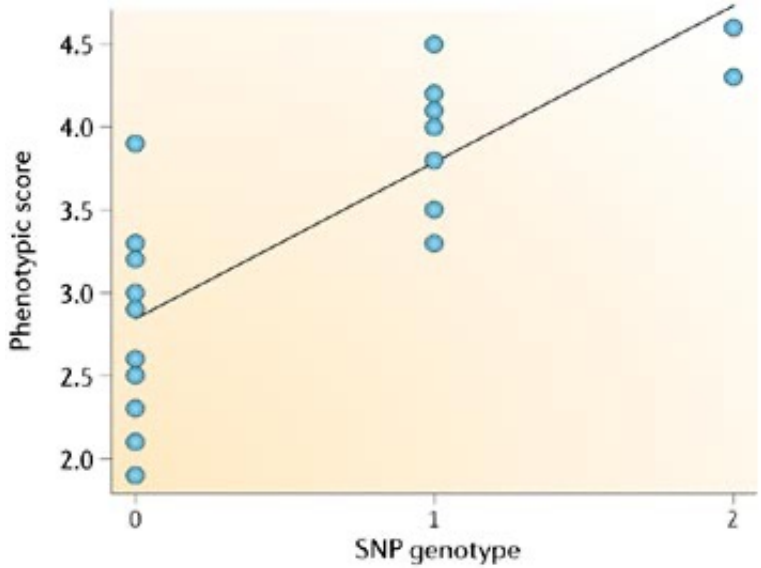
Introduction to GWAS: Quantitative Trait

Linear Regression

$$\hat{Y} = \alpha + \beta X + \varepsilon$$

$\hat{Y}$ = score on phenotype

X = 0, 1 or 2 copies of allele (“G”)



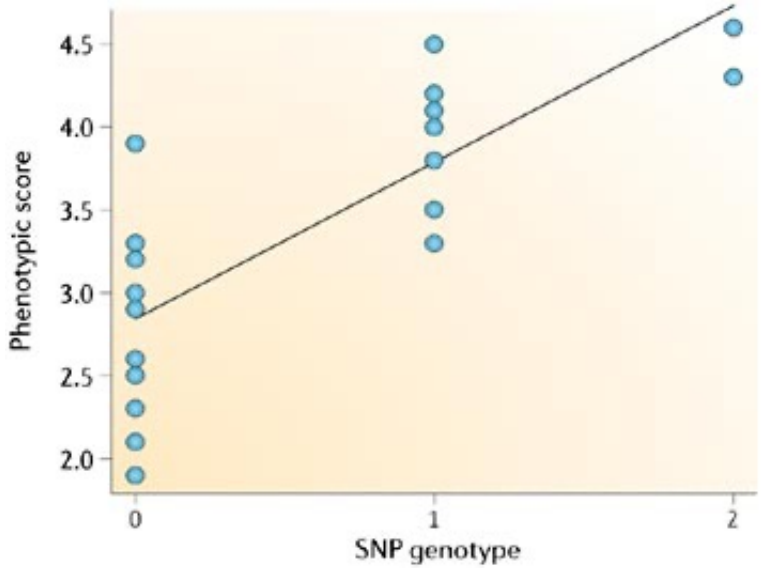
Introduction to GWAS: Quantitative Trait

Linear Regression

$$\hat{Y} = \alpha + \beta X + \varepsilon$$

$\hat{Y}$ = score on phenotype

X = 0, 1 or 2 copies of allele (“G”)



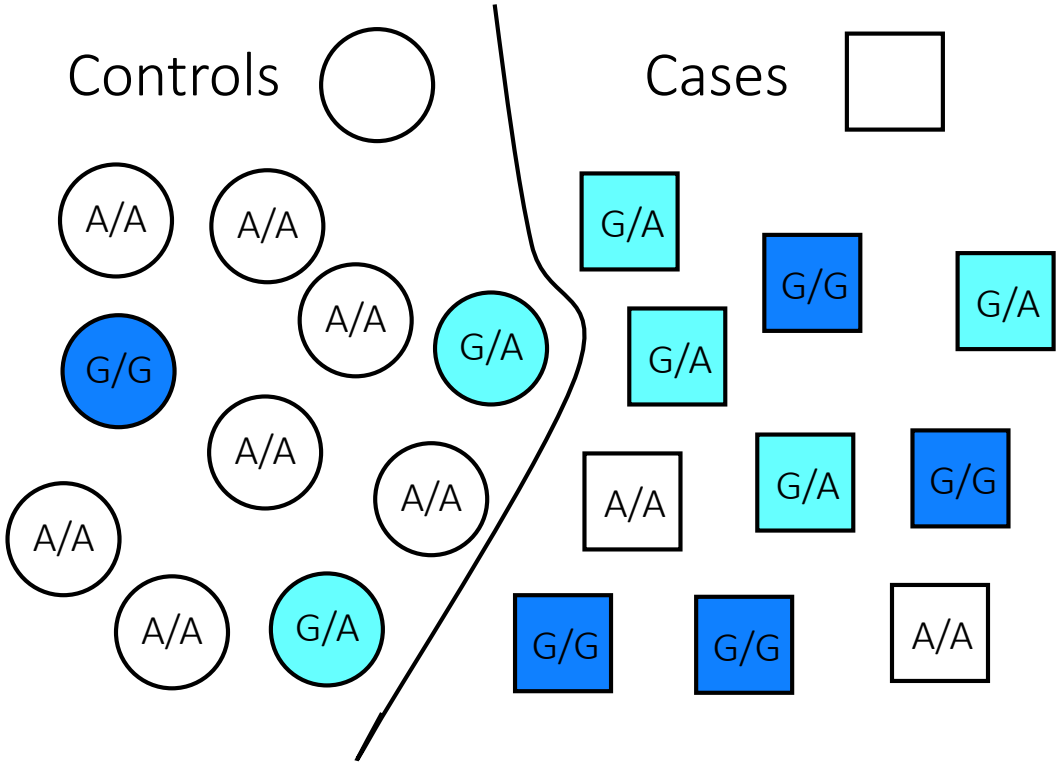
$\beta = 0$ no association

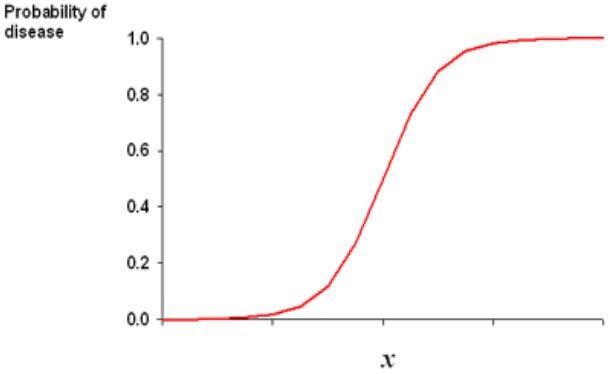
$\beta > 0$ G allele associated with higher score on trait

$\beta < 0$ G allele associated with lower score on trait

Introduction to GWAS: Case-Control

Logistic Regression



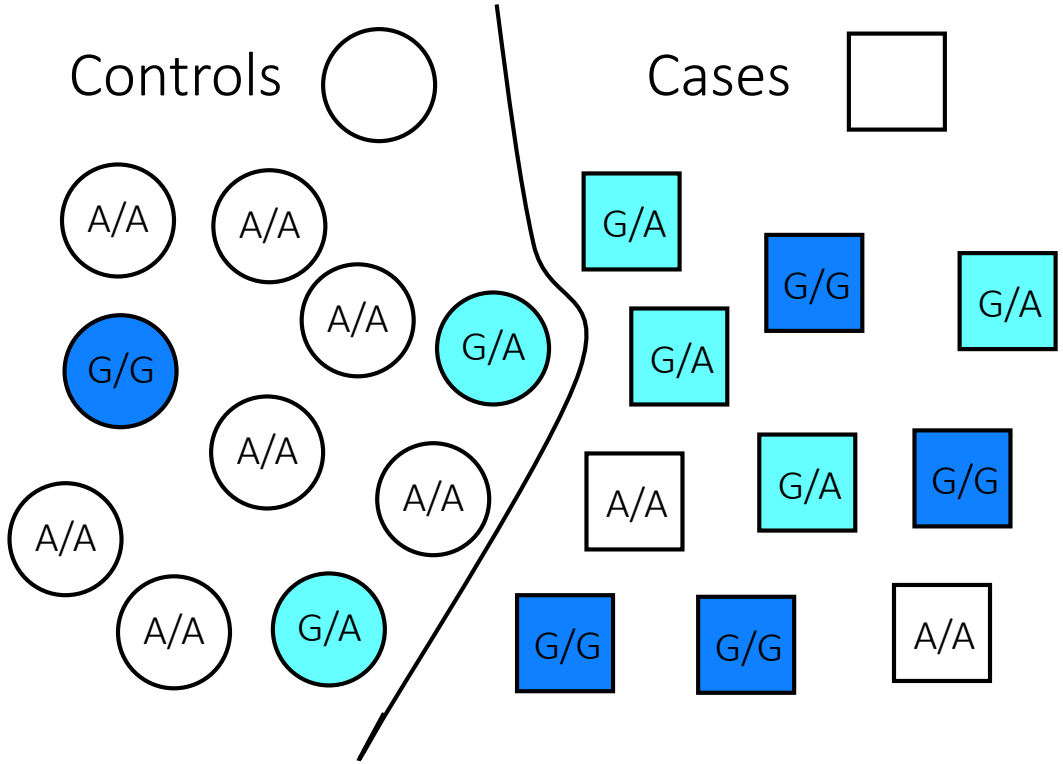


The G allele is associated with disease

Balding. *Nat Rev Genet* (2006)

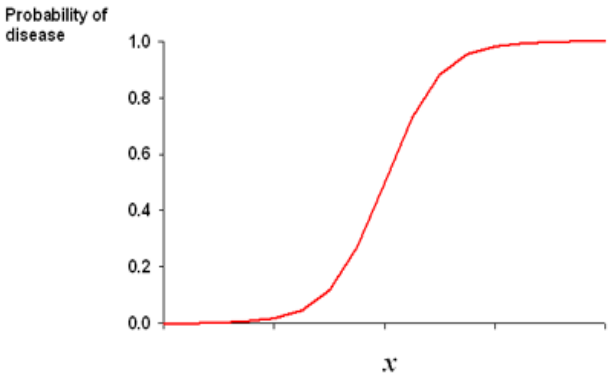
Introduction to GWAS: Case-Control

Logistic Regression



The G allele is associated with disease

Balding. *Nat Rev Genet* (2006)



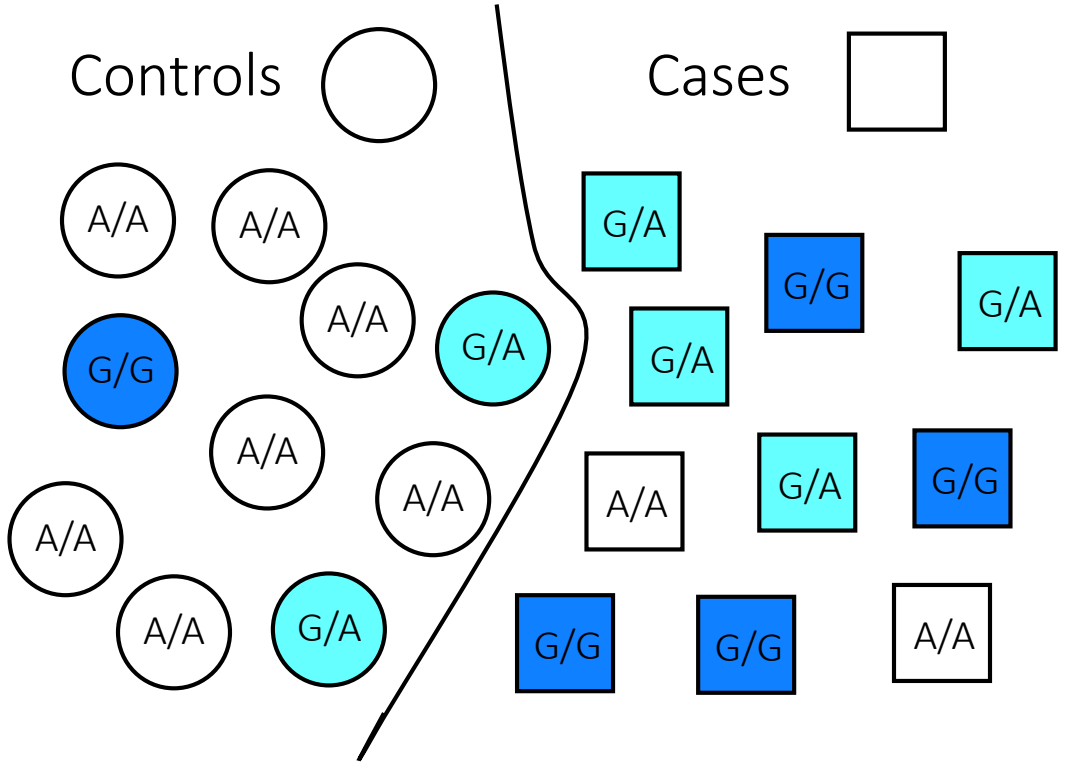
$\ln(P/1-P) = \alpha + \beta X + \varepsilon$

β = difference in log odds for cases vs. controls

$e^{(\beta)}$ = difference in odds
= Odd Ratio (OR)

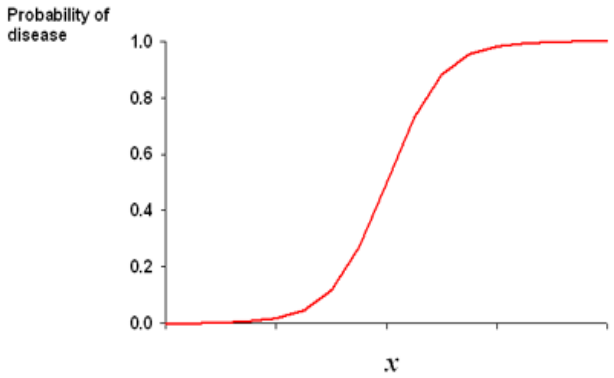
Introduction to GWAS: Case-Control

Logistic Regression



The G allele is associated with disease

Balding. *Nat Rev Genet* (2006)



$\ln(P/1-P) = \alpha + \beta X + \varepsilon$

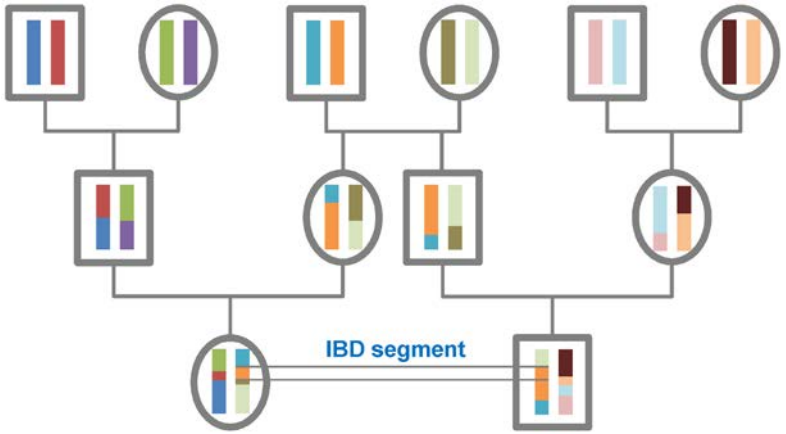
β = difference in log odds for cases vs. controls

$e^{(\beta)}$ = difference in odds
= Odd Ratio (OR)

Allelic effect is an OR:
OR > 1 increased risk
OR < 1 decreased risk

Introduction to GWAS: Relatedness

- Only a few in the total sample = drop



Source: Wikipedia

Introduction to GWAS: Relatedness

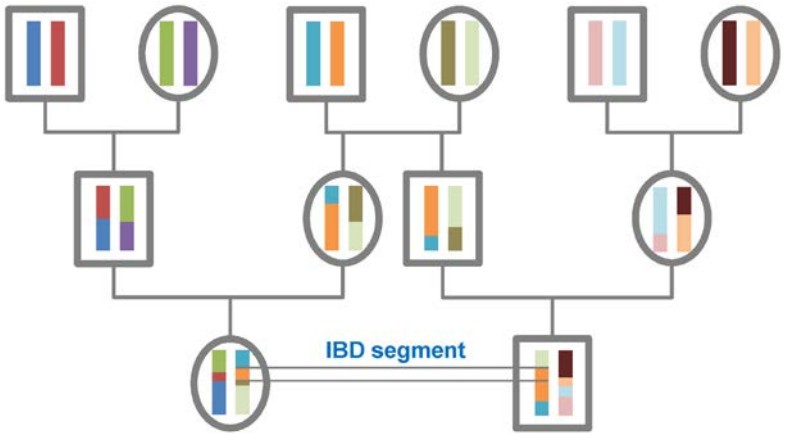
- Only a few in the total sample = drop

- Random Effects Model

$$\hat{Y} = \alpha + \beta X + G + \varepsilon$$

β = fixed effect of the allele

G = genetic relationship random effect



Source: Wikipedia

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

Introduction to GWAS: Relatedness

- Only a few in the total sample = drop

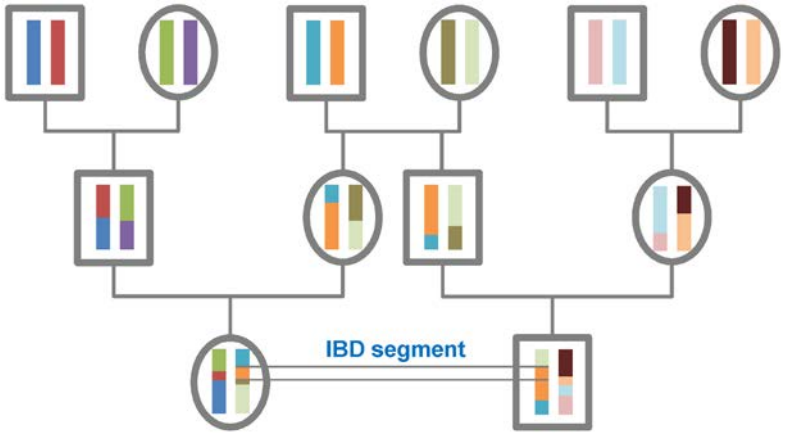
- Random Effects Model

$$\hat{Y} = \alpha + \beta X + G + \varepsilon$$

β = fixed effect of the allele

G = genetic relationship random effect

- Genetic Relationship Matrix (GRM)
  - Sub-sample of SNPs
  - Leave One Chromosome Out (LOCO)



Source: Wikipedia

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

Introduction to GWAS

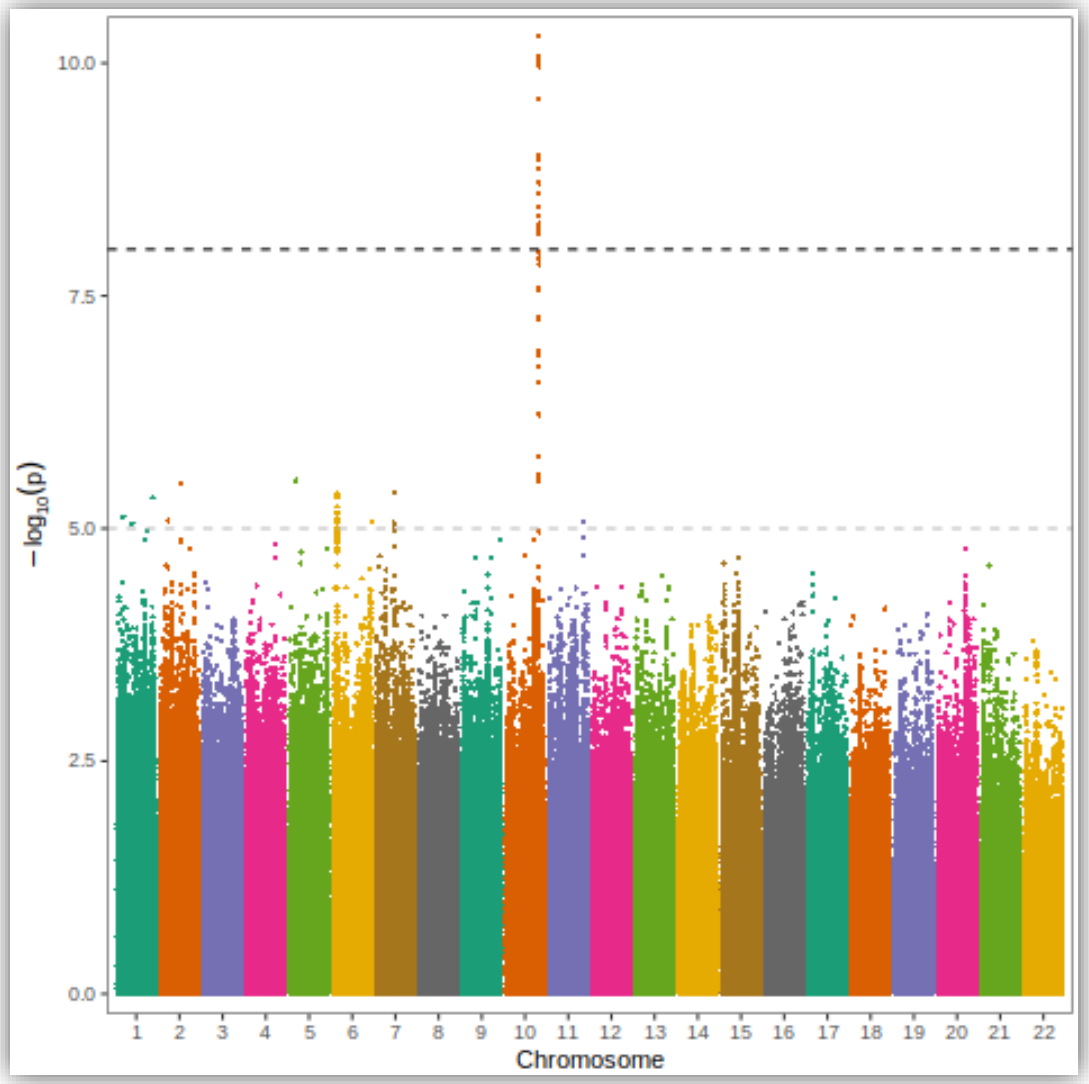
Statistical modeling in GWAS:

1. Linear regression
   I. Fixed effect model
   II. Random effect mode with GRM (relatedness adjustment)
2. Logistic regression
   I. Fixed effect model
   II. Random effect mode with GRM (relatedness adjustment)

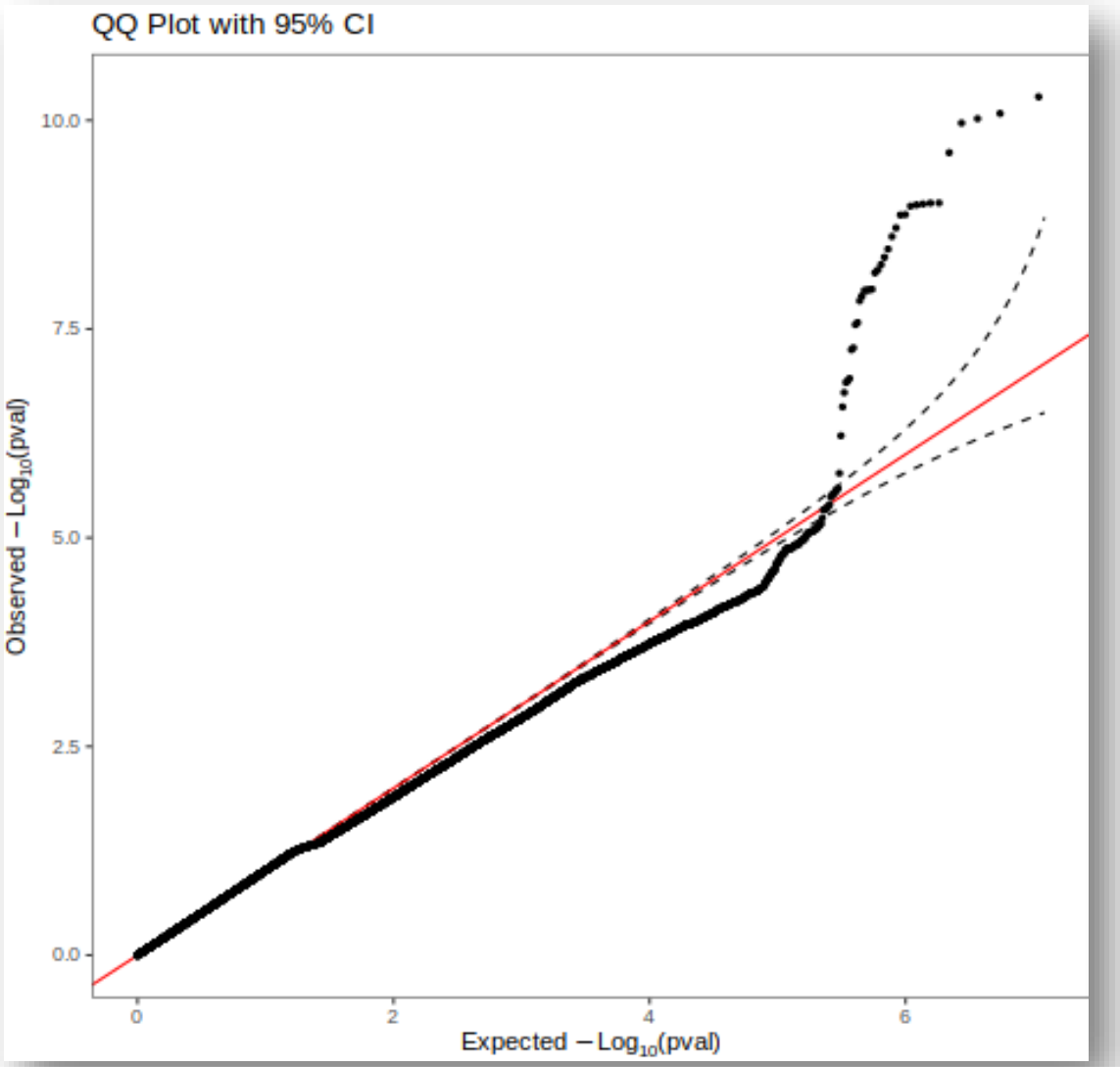
Advanced modeling: Using Machin Learning

Visualizing the GWAS results

Manhattan plot



QQ-plot

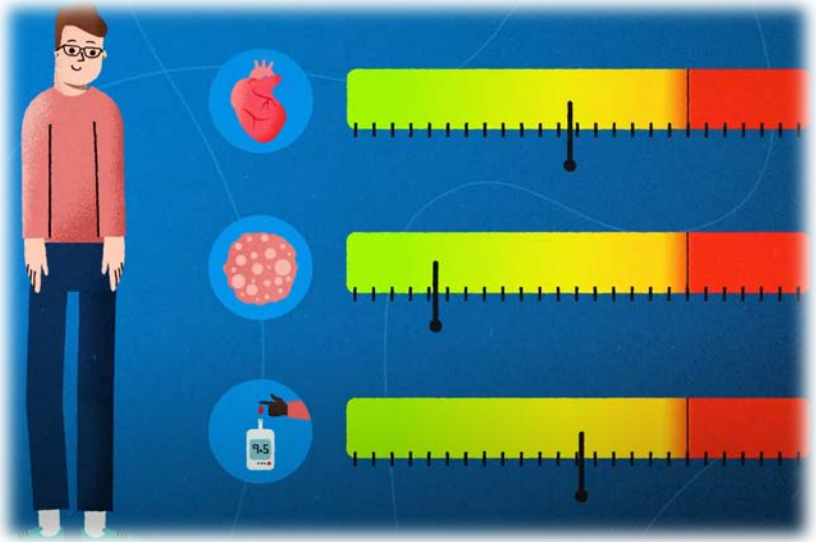


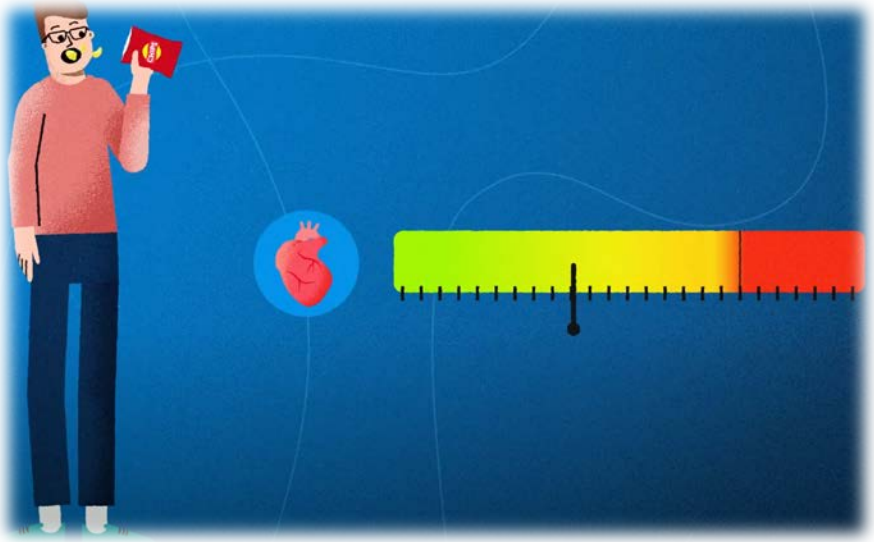
What is the

Polygenic Risk Scores

Introduction to PRS

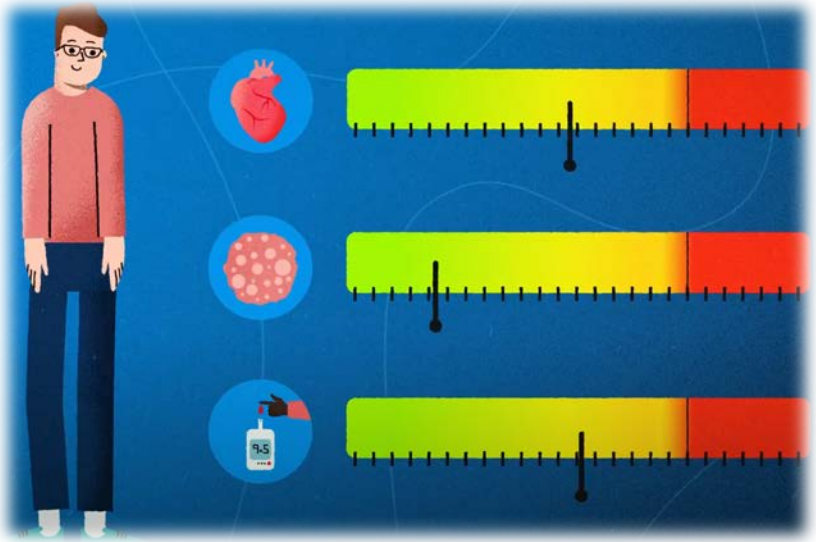
- Knowing our risk of developing diseases
- Change diet, lifestyle, or medication use
- But how could you know that you're at a high risk ?

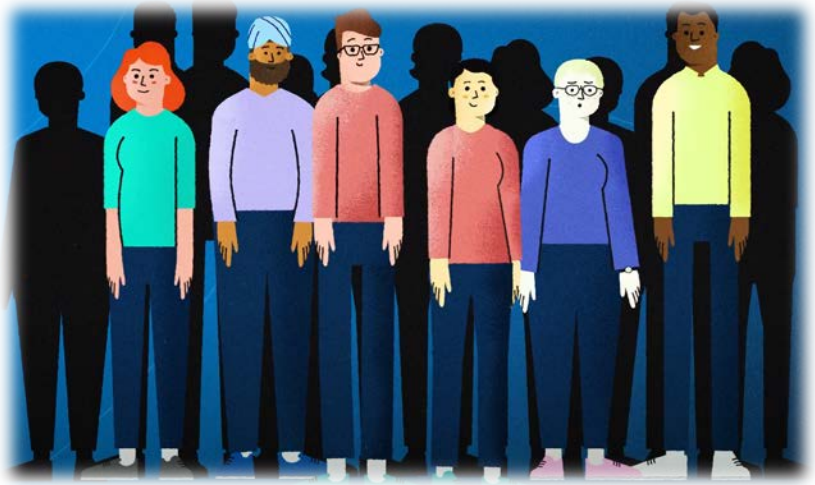




Introduction to PRS

- A major risk factor for common diseases such as heart disease, cancer, and diabetes is our own genetic makeup.
- New studies show that we can now analyze an individual's genes and actually measure that risk using something called polygenic risk score.
- Our genes vary from person to person and it's why we're not all the same, but some of these genetic differences can contribute to our risk of complex diseases.





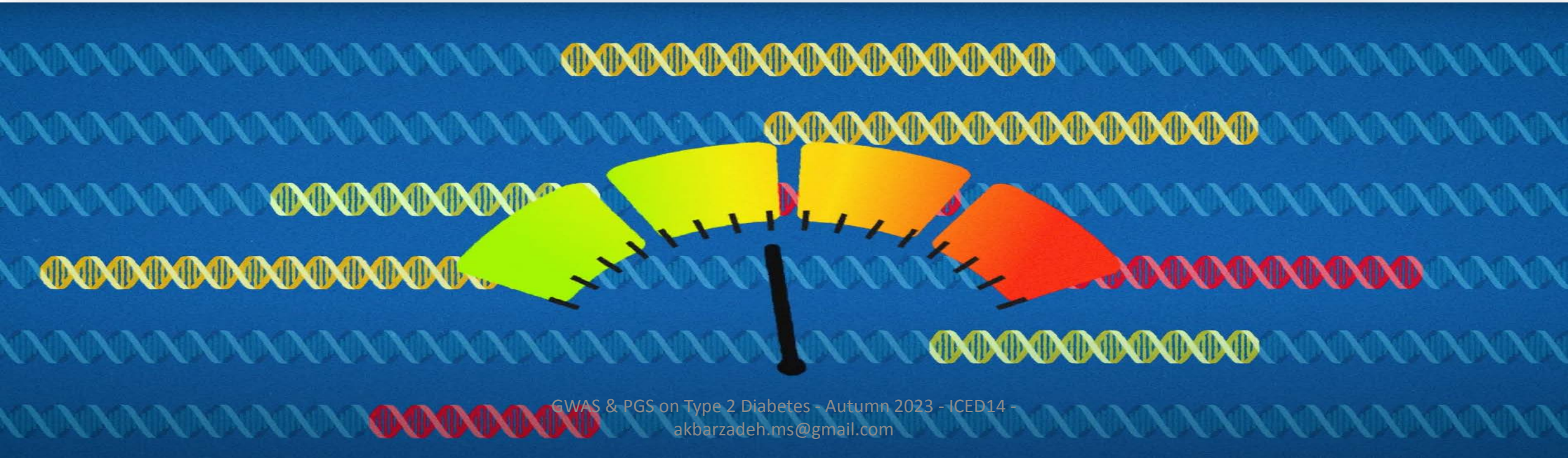
Introduction to PRS

- For the most common diseases such as heart disease and T2D, it's often not just one or two of these genetic changes that are important.
- There is many of them, each having a small effect on the polygenic risk.

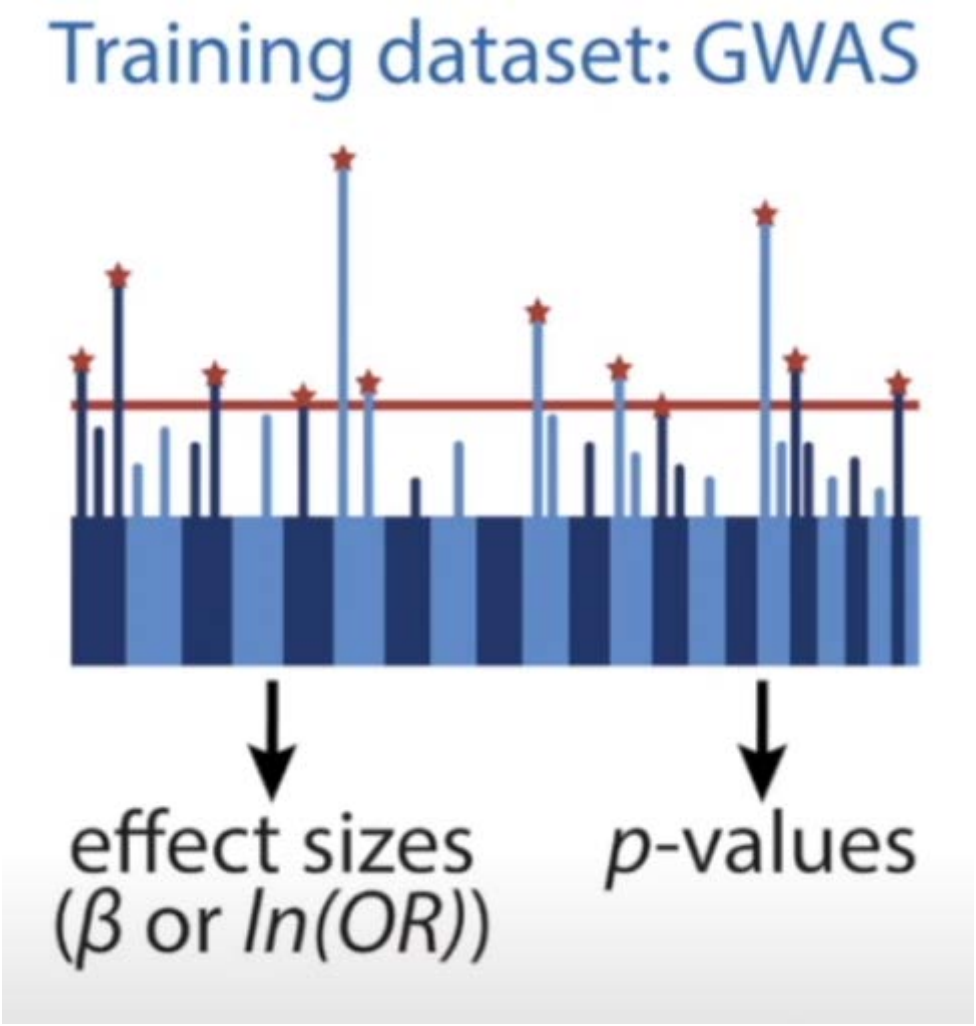
Introduction to PRS

- For the most common diseases such as heart disease and T2D, it's often not just one or two of these genetic changes that are important.
- There is many of them, each having a small effect on the polygenic risk.

POLY many **GENIC** to do with genes **RISK SCORES** scoring a risk.



Introduction to PRS



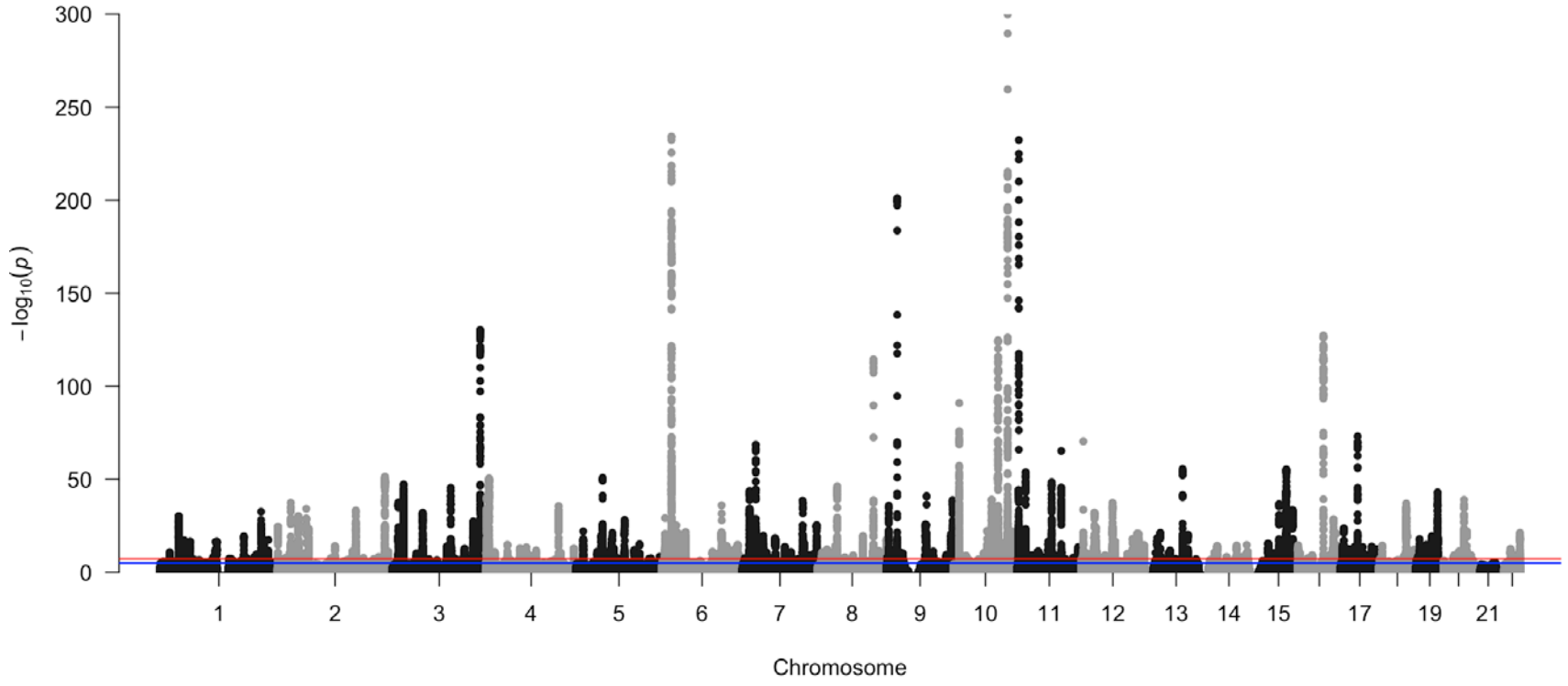
DIAGRAM

Home
About
DIAGRAM Publications
Data Download

Welcome to DIAGRAM

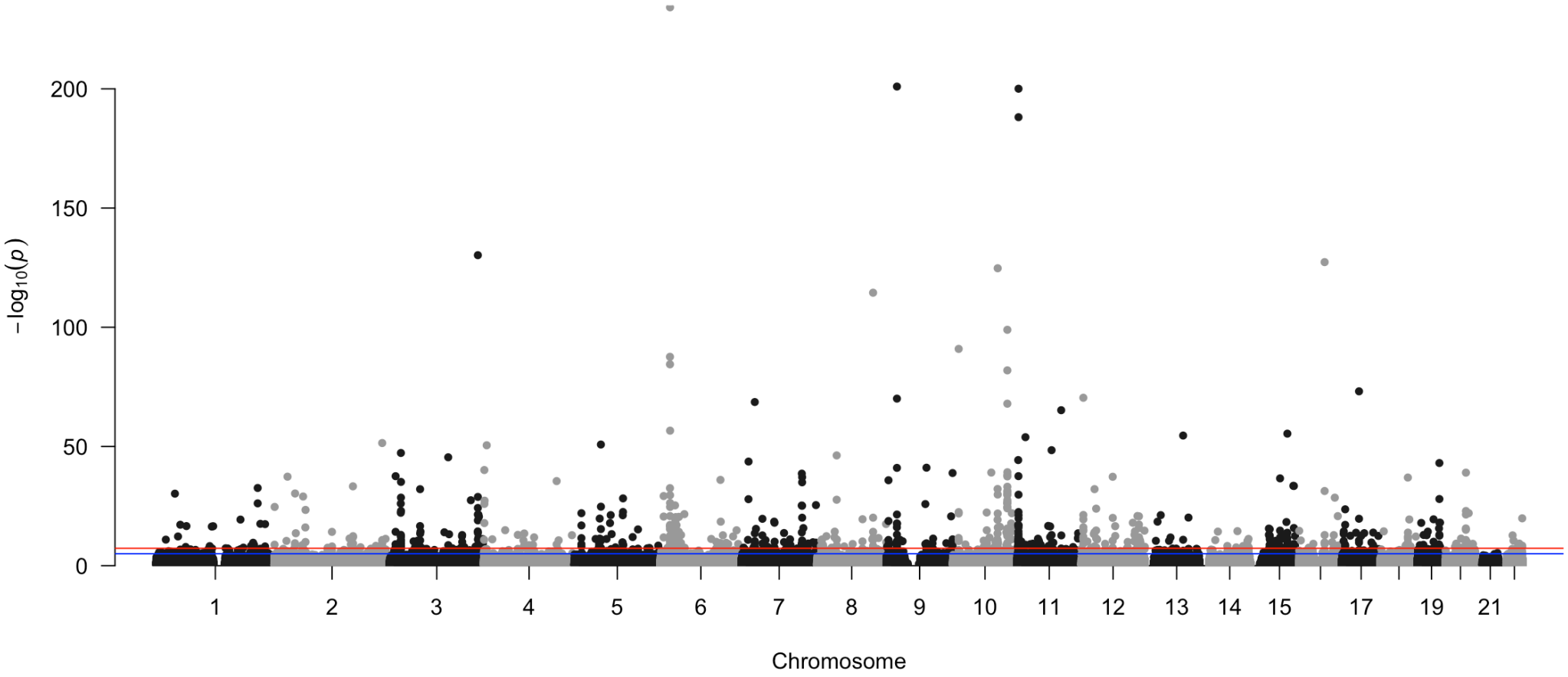
We currently provide a list of DIAGRAM publications. Summary data is available for Morris et al (2012): Stage 1 GWAS & Stage 2 Metabochip. Also available are summary statistics for the trans-ethnic T2D GWAS meta-analysis as published in Mahajan et al. (2014). For access to these, please follow the links to the left.

Polygenic Risk Score :
The first challenge is “Correlated SNPs”



DS: Discovery Sample

Polygenic Risk Score : Manhattan plot: Clumped SNPs



DS: Discovery Sample

Polygenic Risk Score: The second challenge is: How many SNPs do we need? P+T method (Traditional Method)

The discovery set in this study:

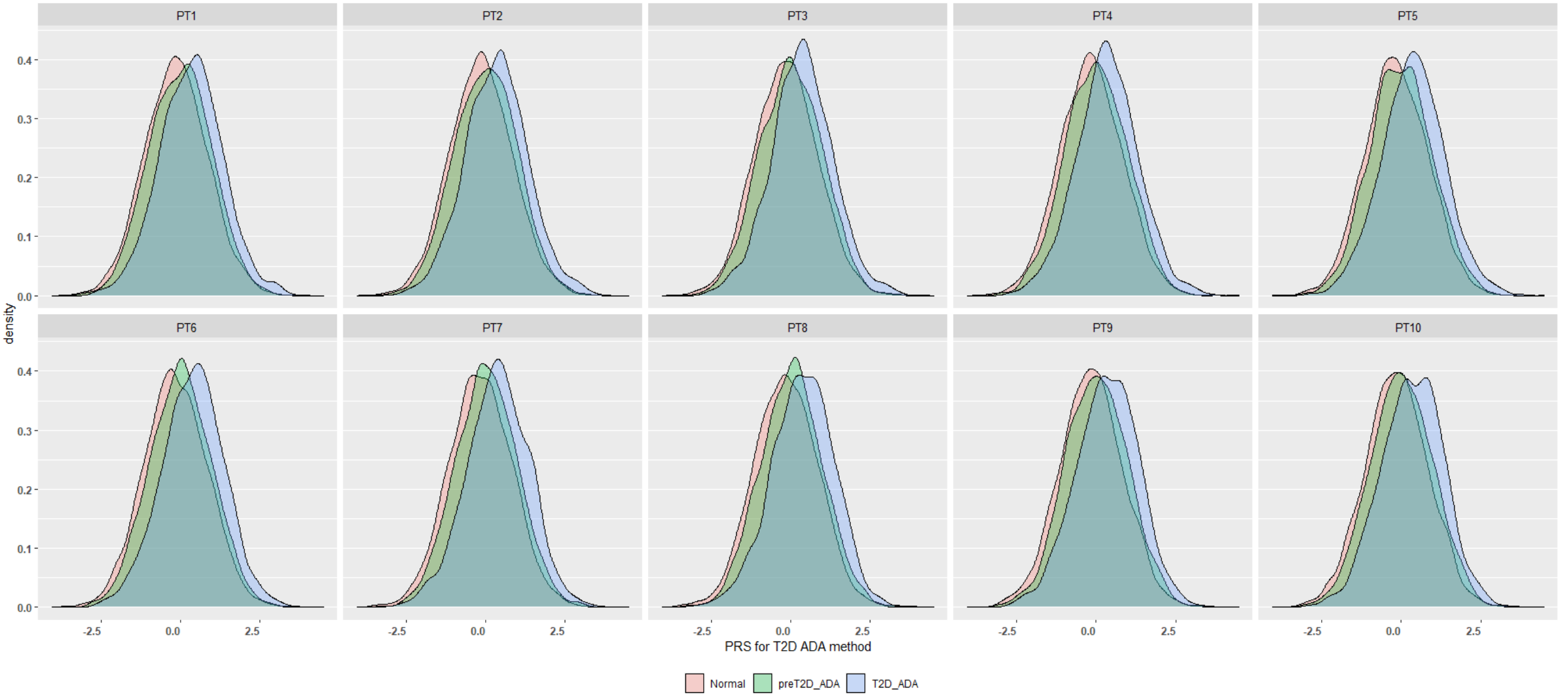
Discovery sample:
180,834 affected individuals and 1,159,055 controls (48.9% non-European descent)
Effective Sample size: 79,074
#SNP: 10,454,875

In TCGS (target data) after QC:
#SNP: 555,835
#common SNPs: 538,505 (96.88%)

After Clumping:
#SNP: 94,532

Pvalue sets
S1: 0.00 , 1
S2: 0.00 ,0.5
S3: 0.00 , 0.2
S4: 0.00 , 0.1
S5: 0.00 , 0.05
S6: 0.00 , 0.01
S7: 0.00 , 0.001
S8: 0.00 , 0.0001
S9: 0.00 , 0.000001
S10: 0.00 , 0.00000005

Polygenic Risk Score: How many SNPs do we need?



Polygenic Risk Score: How many SNPs do we need?

PT	Threshold	AUC	N.SNPs
PT10	P<5E-08	0.6454	473
PT9	P<1E-06	0.6495	660
PT8	P<1E-04	0.6574	1528
PT7	P<0.001	0.6632	3026
PT6	P<0.01	0.6621	8439
PT5	P<0.05	0.6549	20120
PT4	P<0.1	0.6527	29861
PT3	P<0.2	0.6503	44054
PT2	P<0.5	0.6442	70489
PT1	P<1	0.6396	94517

Introduction to PRS

nature genetics

Explore content | About the journal | Publish with us | Subscribe

nature > nature genetics > articles > article

Article | Published: 12 May 2022

Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation

Anubha Mahajan, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima, Grace Z. Yu, Sina Rüeger, Leo Speidel, Young Jin Kim, Momoko Horikoshi, Josep M. Mercader, Daniel Taliun, Sanghoon Moon, Soo-Heon Kwak, Neil R. Robertson, Nigel W. Rayner, Marie Loh, Bong-Jo Kim, Joshua Chiou, Irene Miguel-Escalada, Pietro della Briotta Parolo, Kuang Lin, Fiona Bragg, FinnGen, eMERGE Consortium, ... Andrew P. Morris + Show authors

Nature Genetics **54**, 560–572 (2022) | Cite this article

12k Accesses | **21** Citations | **366** Altmetric | Metrics

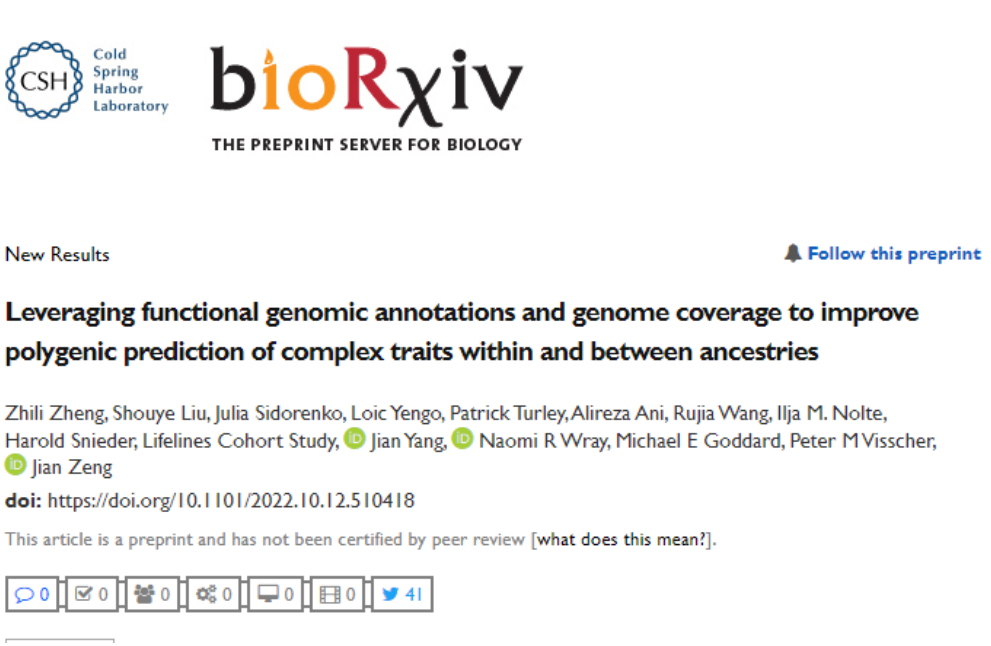
DIAGRAM

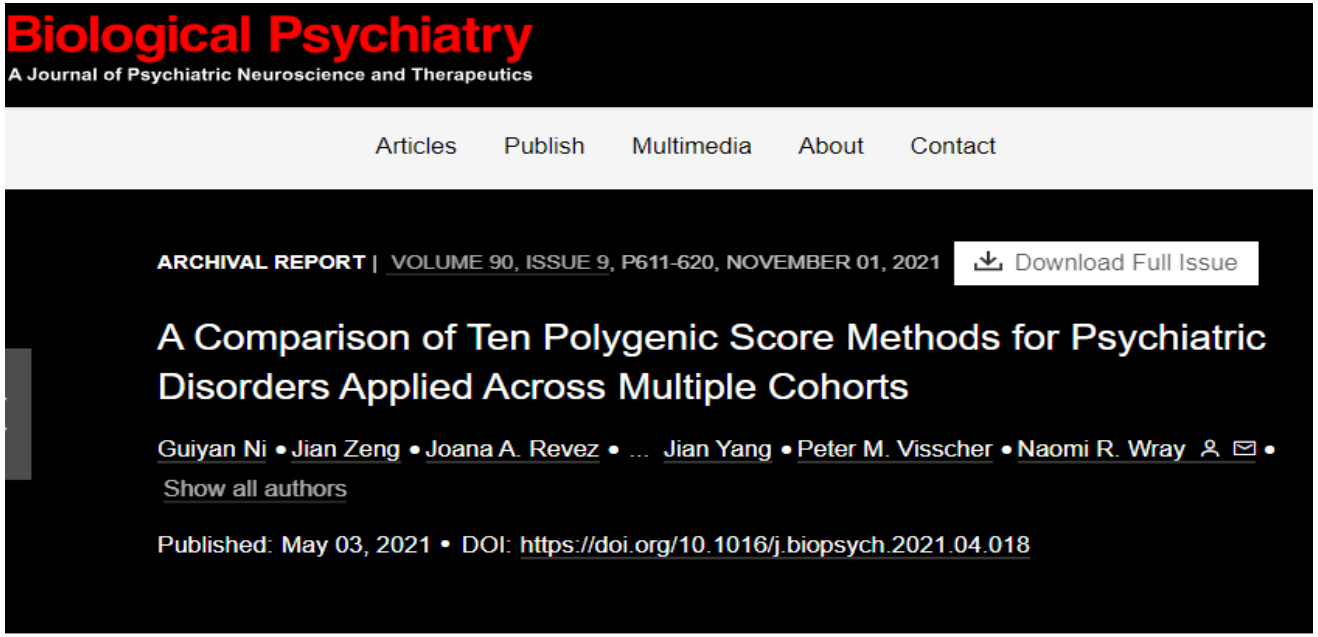
Home
About
DIAGRAM Publications
Data Download

Welcome to DIAGRAM

We currently provide a list of DIAGRAM publications. Summary data is available for Morris et al (2012): Stage 1 GWAS & Stage 2 Metabochip. Also available are summary statistics for the trans-ethnic T2D GWAS meta-analysis as published in Mahajan et al. (2014). For access to these, please follow the links to the left.

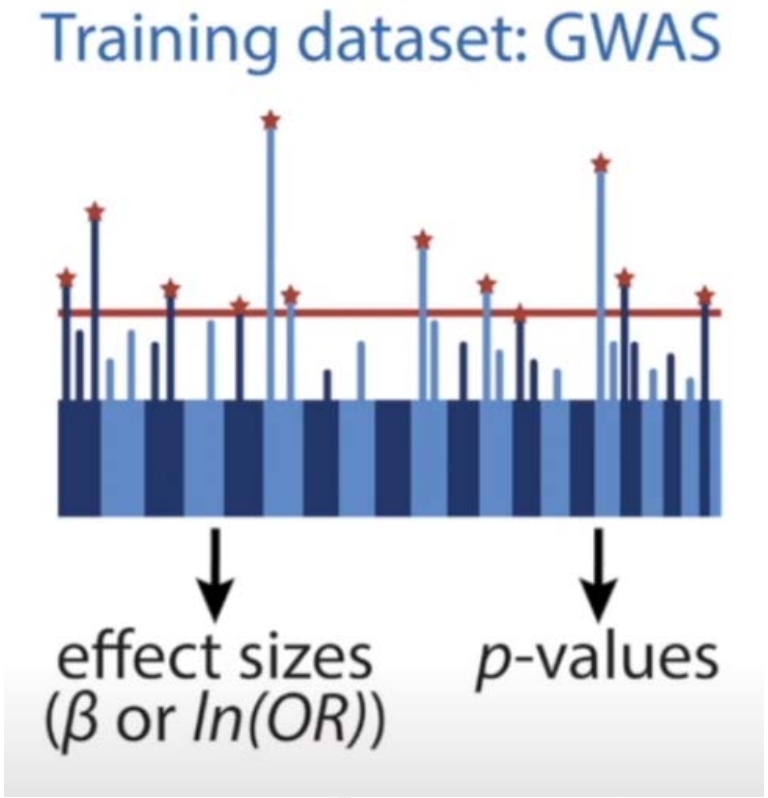
Introduction to PRS





Calculate LD corrected weights by LDpred, PRSice and SBayesR, and SBayesRC.

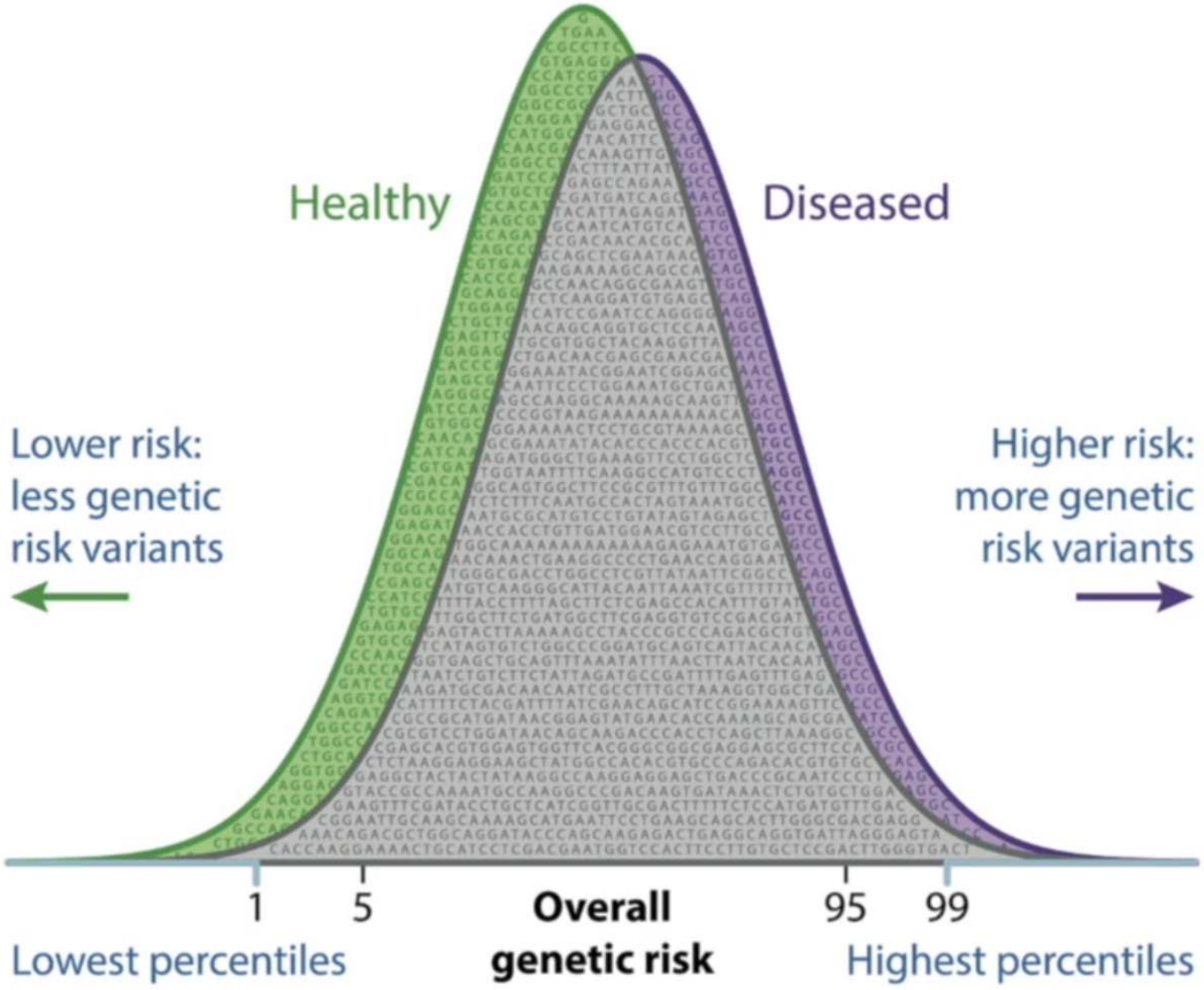
Introduction to PRS



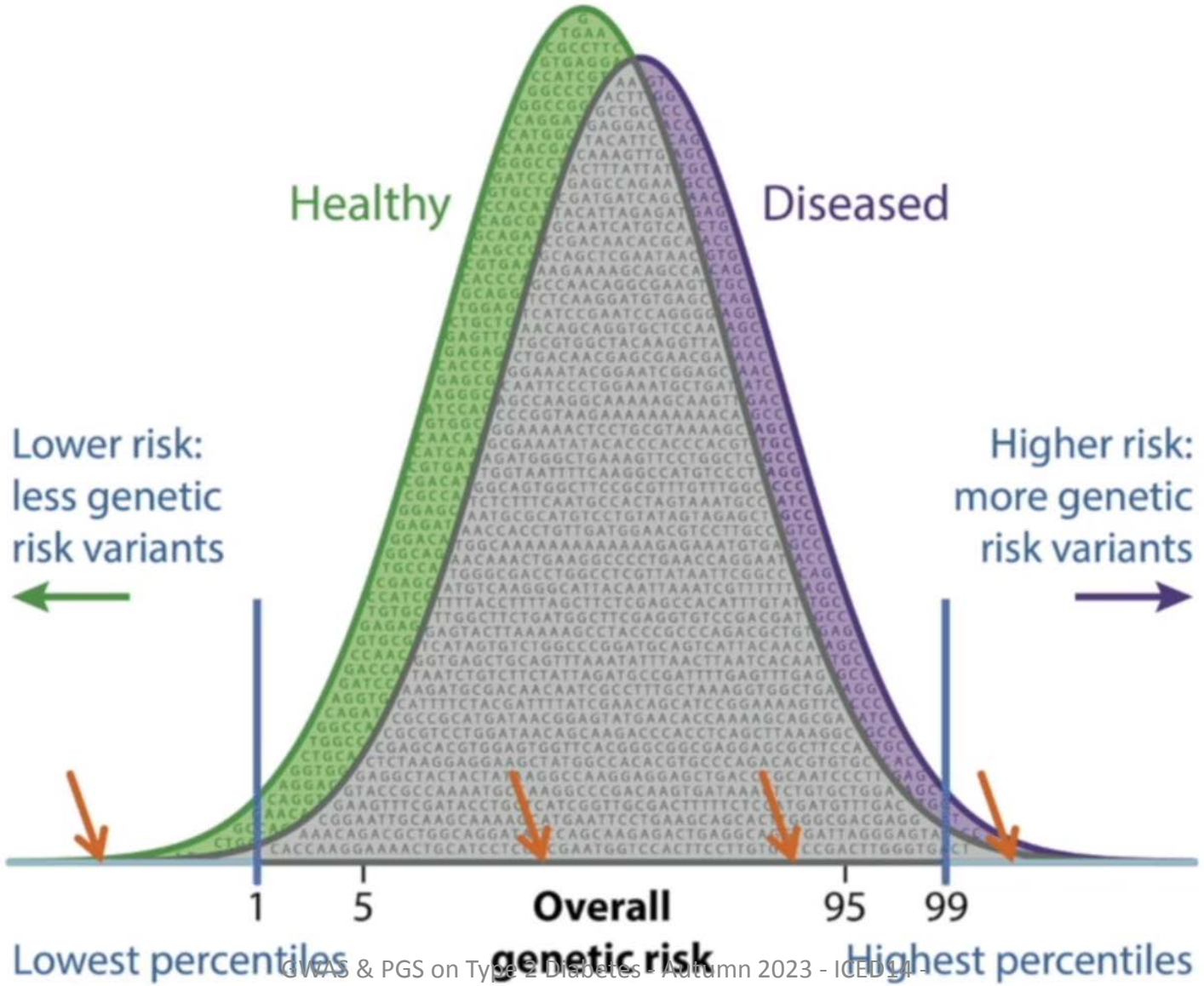
Test dataset: PGS

$$\sum_{p\text{-val}} \sum_{\text{SNVs}} (\text{allele count})_{\text{test}} \times (\text{effect size})_{\text{training}}$$

Introduction to PRS



Introduction to PRS

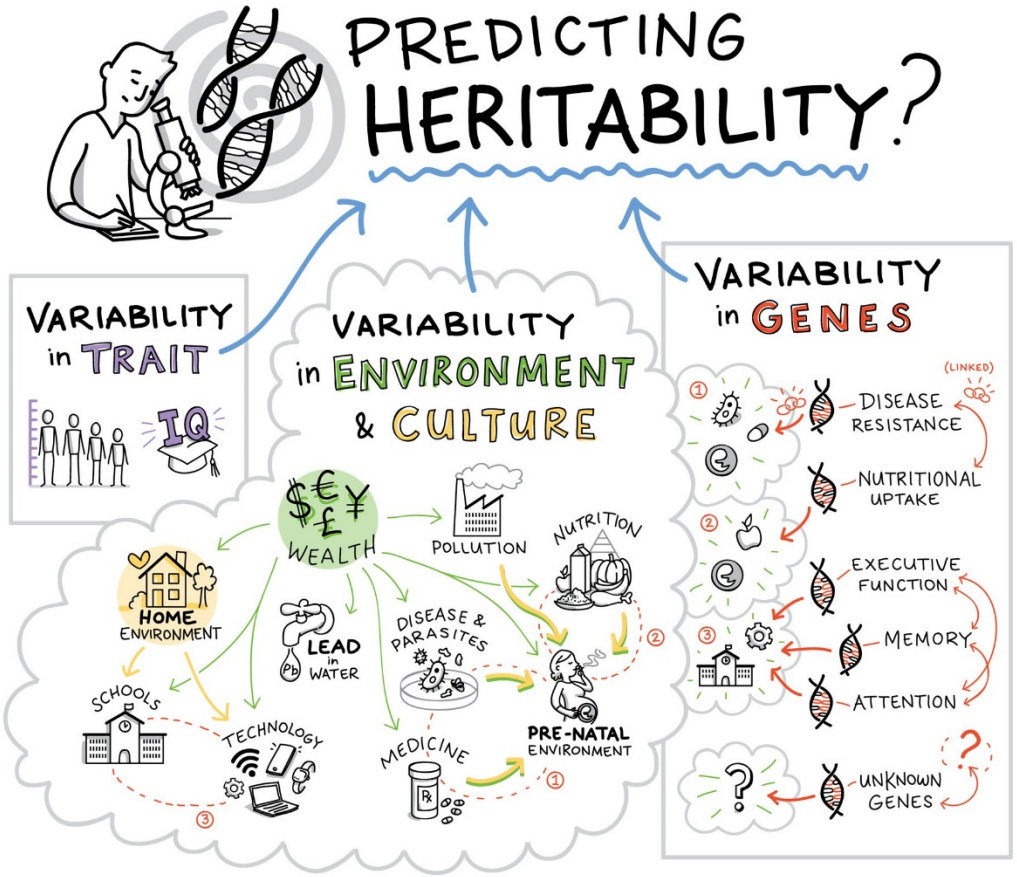


akbarzadeh.ms@gmail.com

What is the Heritability

Introduction to Heritability

Heritability (h^2) quantifies the degree to which inter-individual differences and resemblance in the population are due to genetic factors.



Introduction to Heritability

Heritability (h^2) quantifies the degree to which inter-individual differences and resemblance in the population are due to genetic factors.

If the value, Y, of trait (=phenotype) can be modelled as

$Y = G + E$

Nice definition but not very useful unless we can observe G!

Then h^2 = var(G) / var(Y), i.e. proportion of trait variance explained by genetic factors.

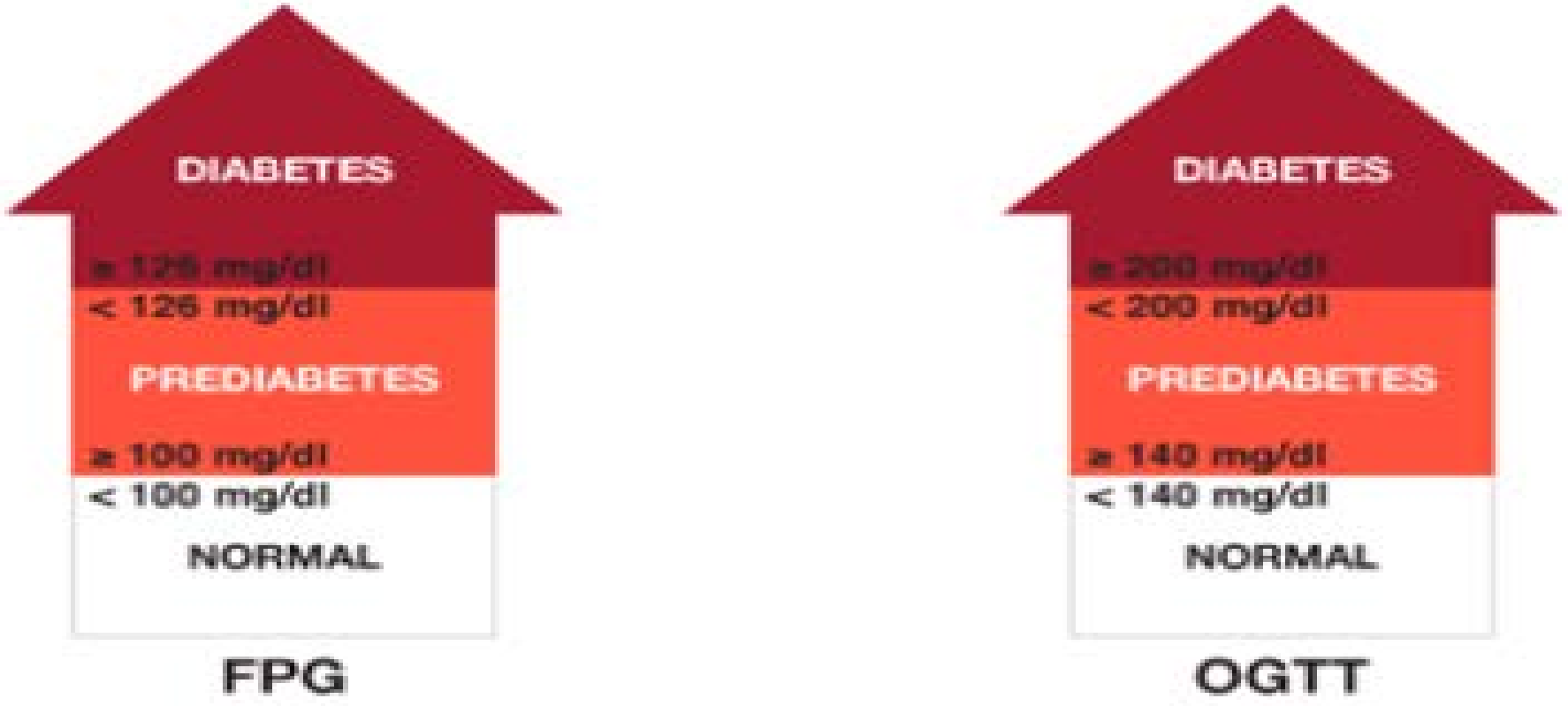
Any question?

Definition of T2D in TCGS

and

Results

American Diabetes Association (ADA) definition



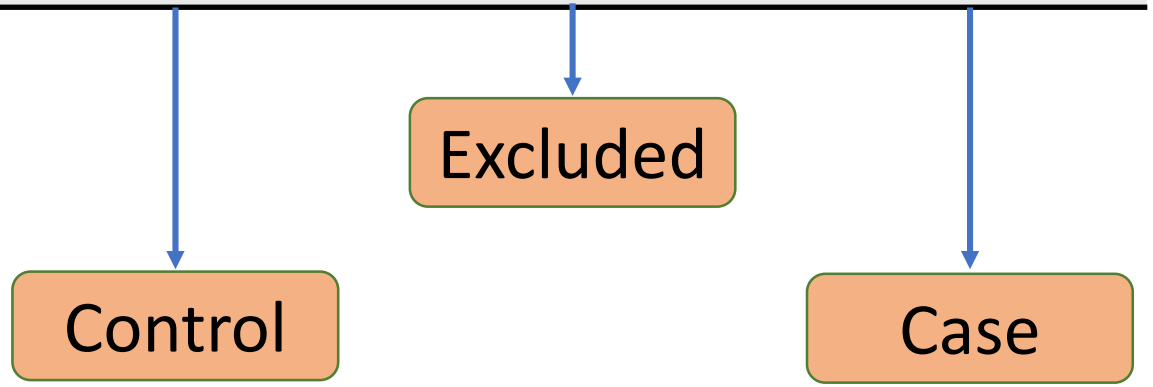
source: www.diabetes.org

Two general types of definition: ADA definition

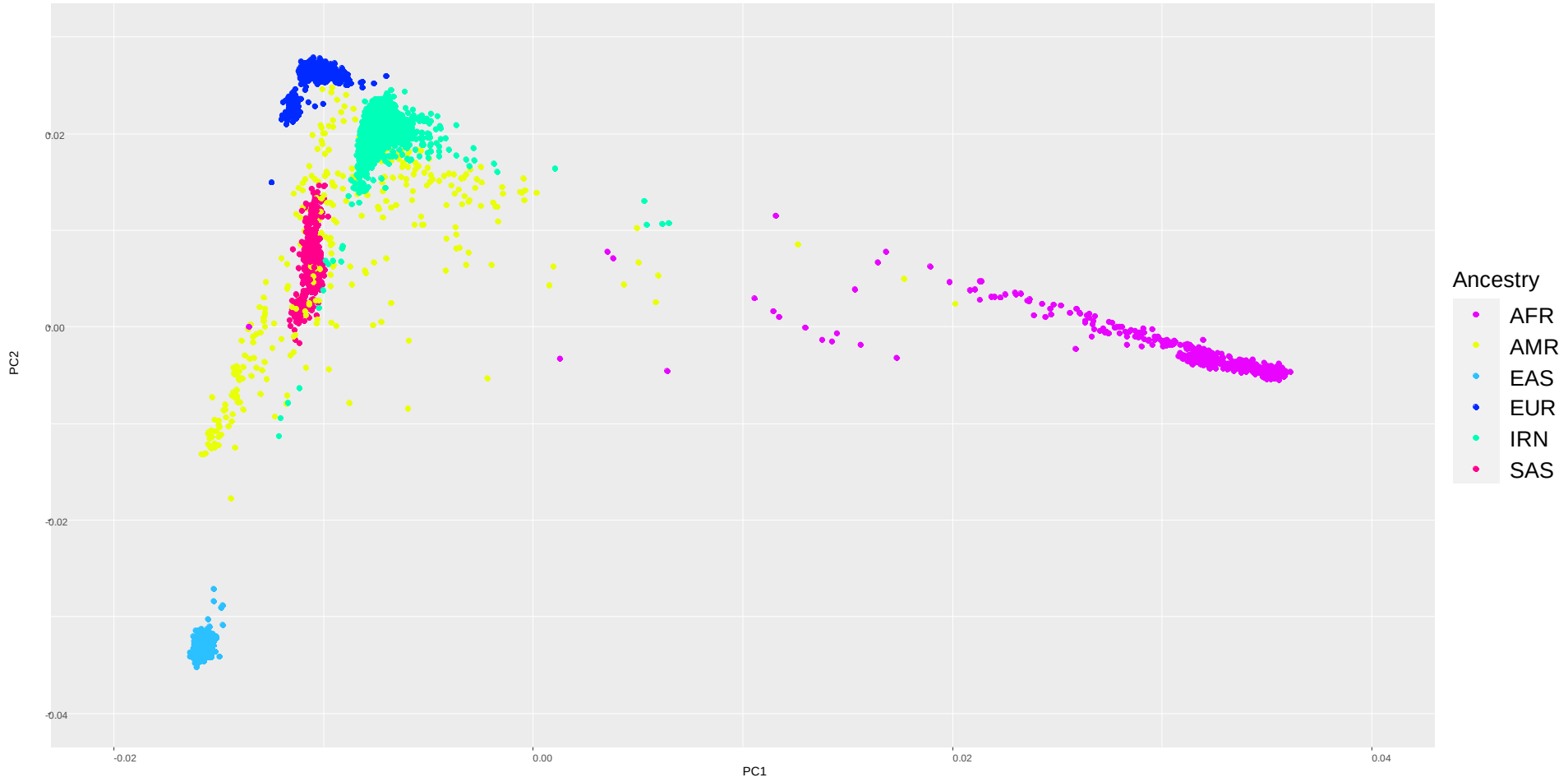
TLGS Phase	Normal	Prediabetes	T2D	Unclear
Phase 1	6,701	1,062	319	2,812
Phase 2	6,481	1,131	466	2,962
Phase 3	7,224	1,057	596	2,284
Phase 4	6,680	1,976	904	1,819
Phase 5	6,027	2,195	1,142	2,153
Phase 6	5,121	1,837	1,305	3,358
ADA	**5,780**	4,130	**2,167**	1

American Diabetes Association (ADA) definition

TLGS Phase	Normal	Prediabetes	T2D
ADA	5,780	4,130	2,167



1000 genome projection: Principal Components



Quality control before GWAS (imputed data)

Imputed data: 61,999,570 (~62M)

Imputation Information > 0.50 ➔ 58,124,808

HWE assumption: 10e-5 ➔ 55,223,222

Missing per SNP 0.05 ➔ 49,746,903

MAC 5 & MAF 0.05 ➔ 12,419,966

Remove Duplicate ➔ 12,046,290

Remove Reference allele from Multi-Allelic variants ➔ 10,945,256 (**~11M**)

Quality control before GWAS (imputed data)

Imputed data: 61,999,570 (~62M)

Imputation Information > 0.50 ➔ 58,124,808

HWE assumption: 10e-5 ➔ 55,223,222

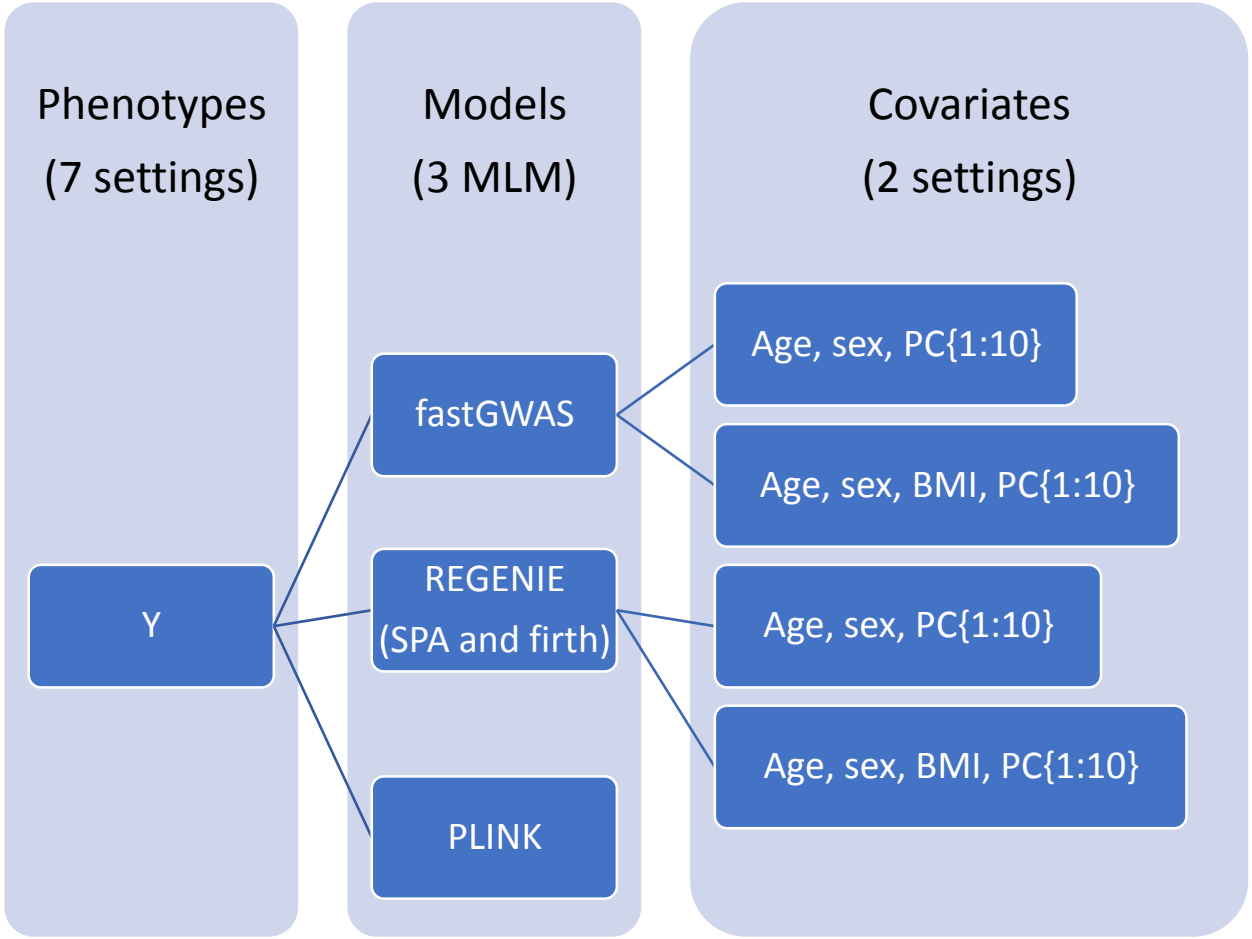
Missing per SNP 0.05 ➔ 49,746,903

MAC 5 & MAF 0.05 ➔ 12,419,966

Remove Duplicate ➔ 12,046,290

Remove Reference allele from Multi-Allelic variants ➔ 10,945,256 (**~11M**)

GWAS Scenarios:

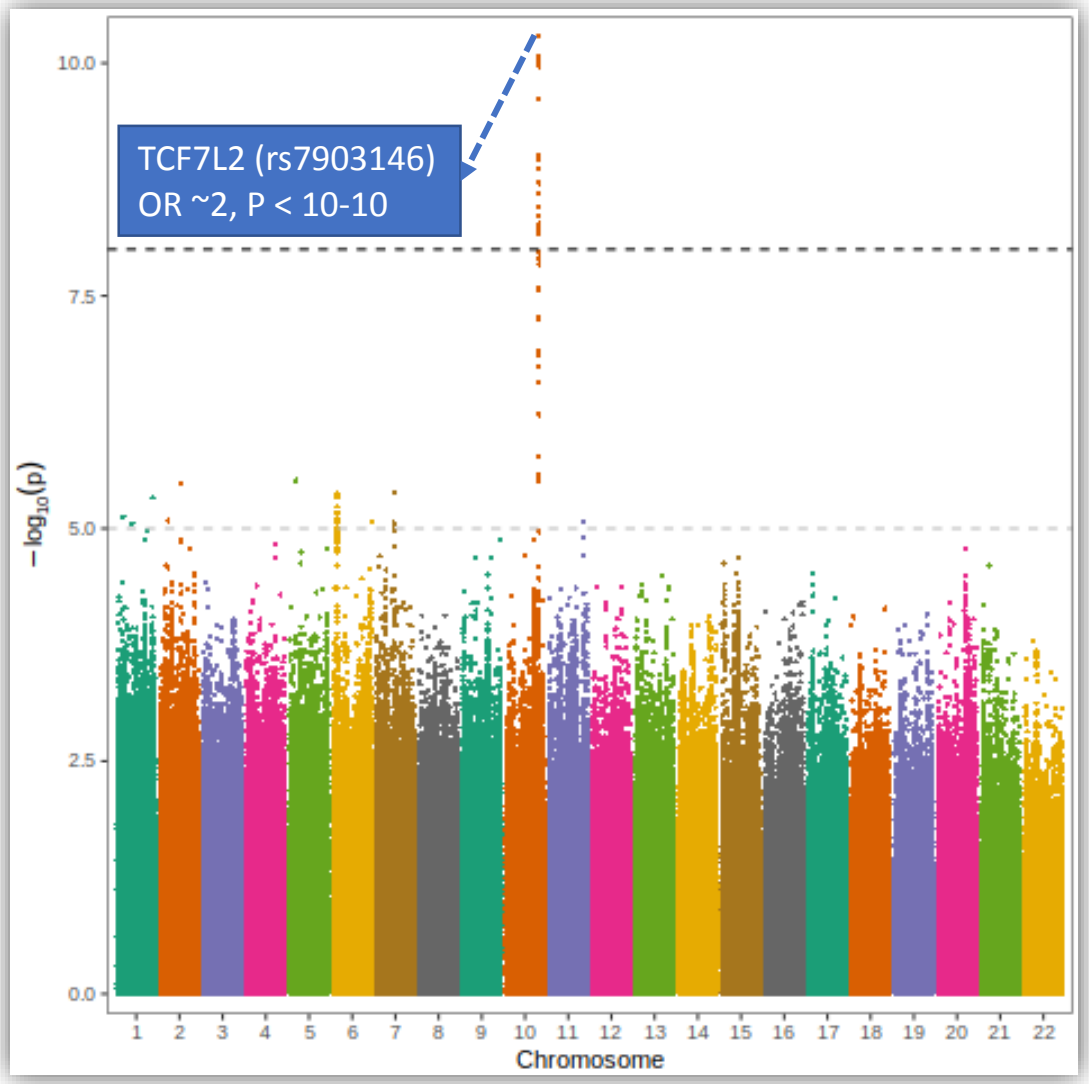


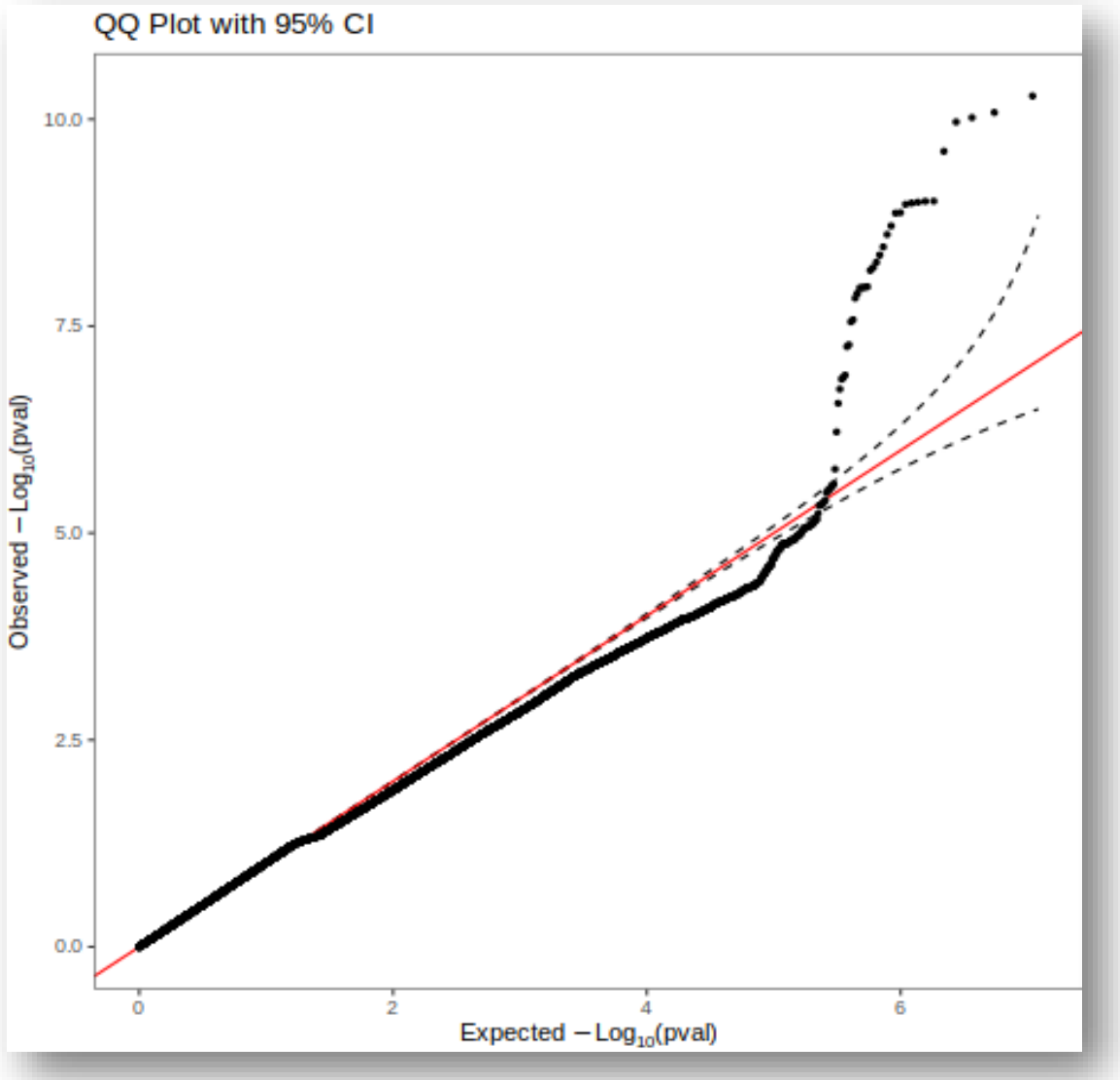
> 40 GWAS Models

fastGWA mixed linear model,
LOCO, COJO, mtCOJO

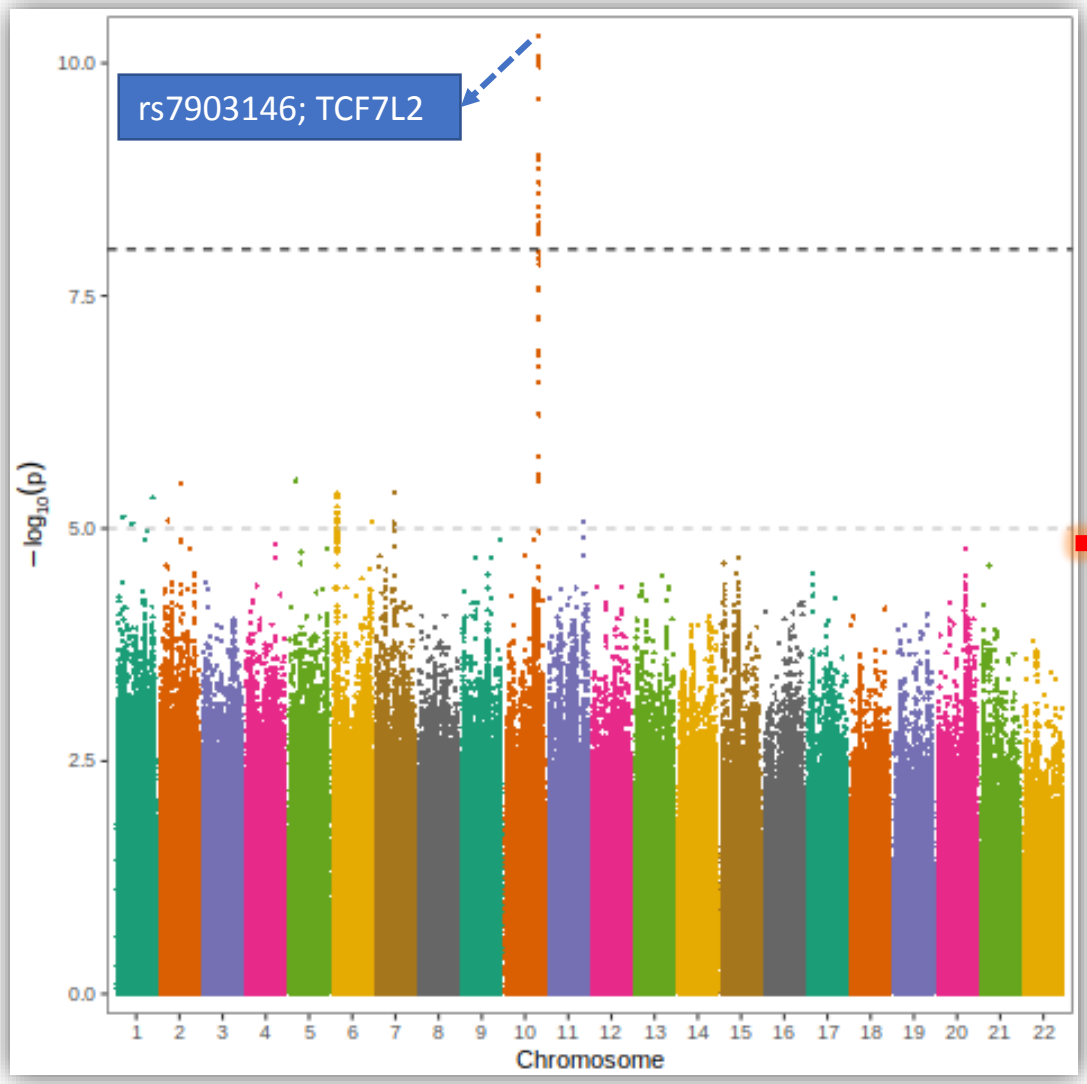
Manhattan plot of the T2D GWAS

(Adjusted for age – sex – 10PCs)

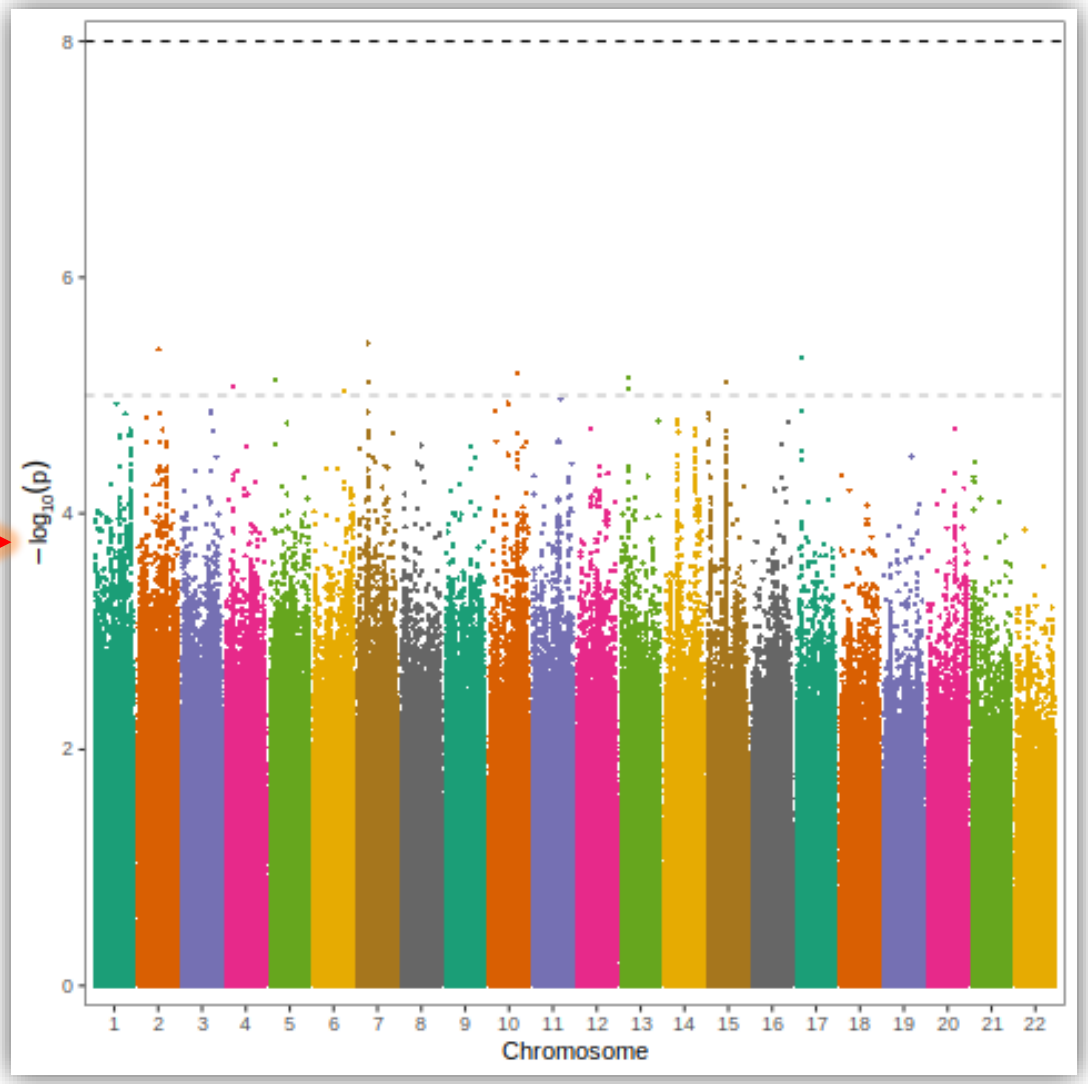




Manhattan plot of the T2D GWAS
(Adjusted for age – sex – 10PCs)



Manhattan plot of the T2D GWAS adjusted for
(Adjusted for age + sex + 10PCs + PRS**)**



Heritability – T2D

Familial and SNP-based heritability are reported based on two relatedness threshold, 0.025 and 0.05, and adjusted by age, sex, 10 PCs, and with/without BMI.

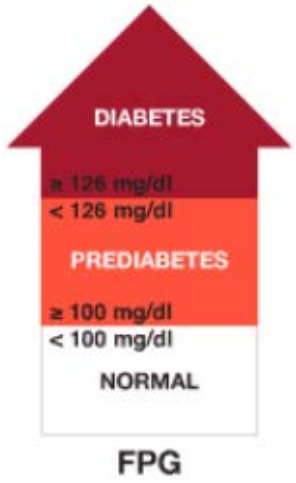
Without adjusted for BMI

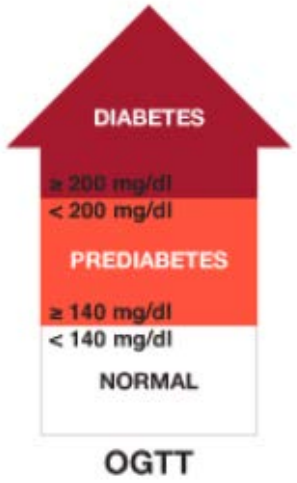
Method		Threshold	Heritability	SE	P-value
Single GRM		0.025	0.391	0.0144	2.56e-3
		0.05	0.280	0.0120	8.72e-3
bigK/smallK	pedigree	0.025	0.367	0.045	3.8e-4
		0.05	0.364	0.0457	2.77e-4
	unrelated	0.025	0.227	0.068	3.8e-4
		0.05	0.232	0.068	2.77e-4

It needs to say that there is 8% - 13% still-missing heritability.

Polygenic Risk Score (PRS) Analysis

Polygenic Risk Score (PRS) Analysis





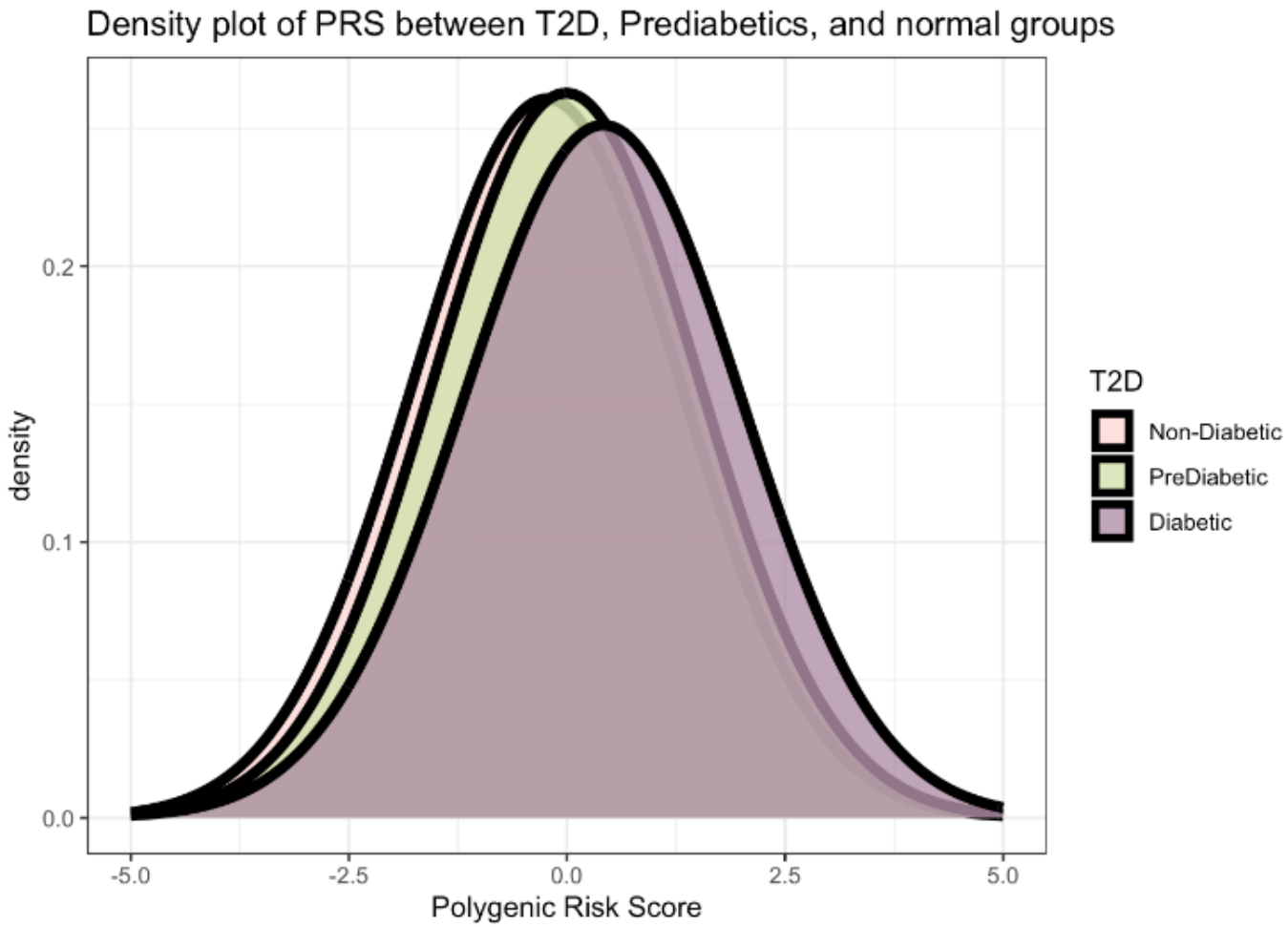
Total Sample size in each group

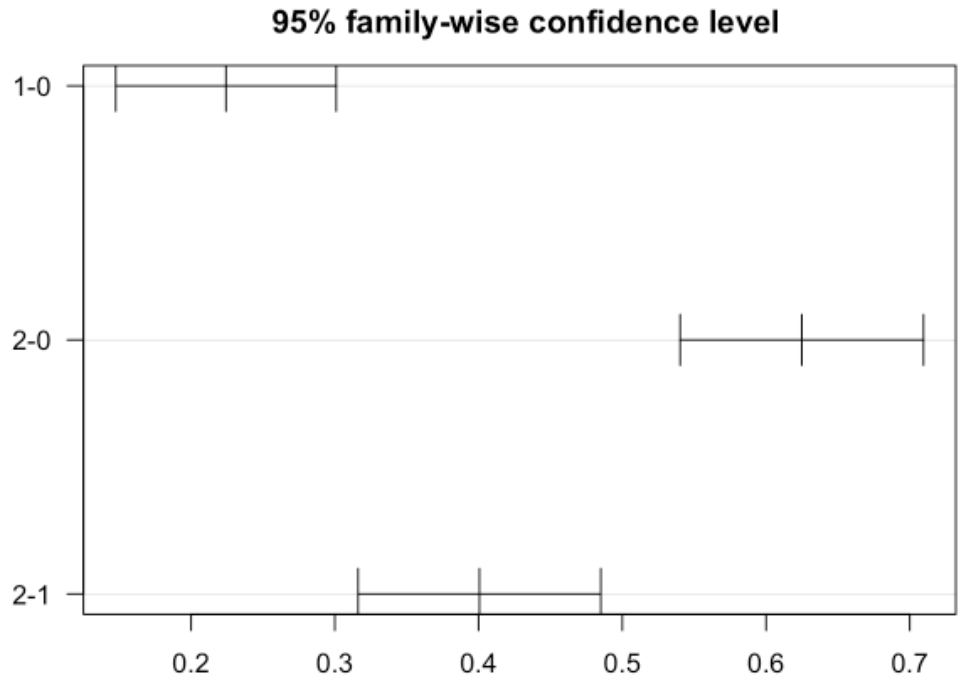
Normal	Prediabetes	T2D
۵,۷۴۱	۴,۱۱۱	۲,۱۶۵

Unrelated Sample size in each group (grm0.05)

Normal	Prediabetes	T2D
1,746	1,764	1,228

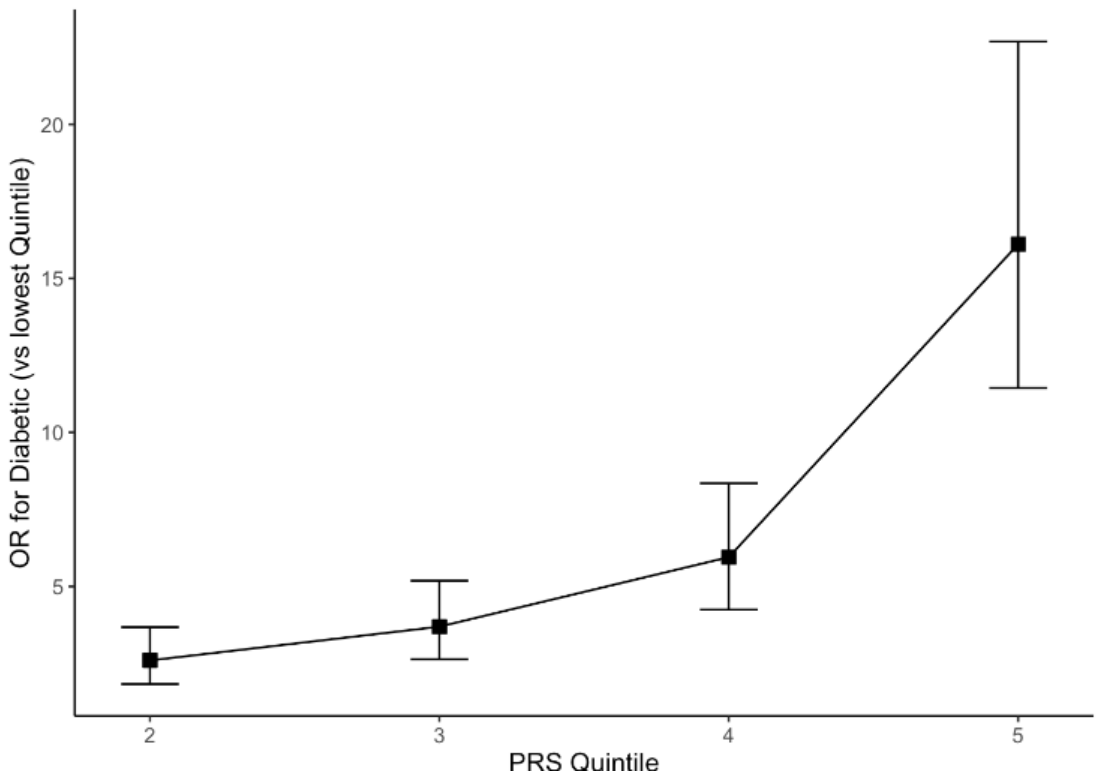
Polygenic Risk Score (PRS) Analysis



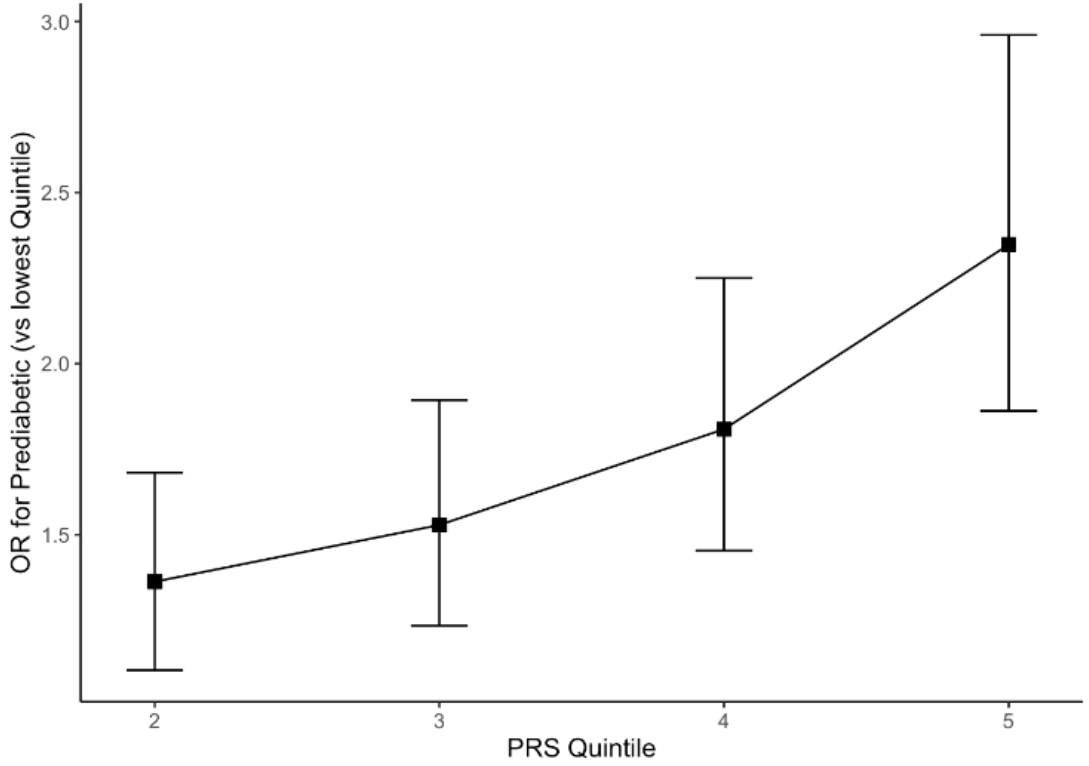


Polygenic Risk Score (PRS) Analysis

T2D vs Control



PreT2D vs Control



Interaction effect on T2D

(BMI × PGS & FH × PGS)

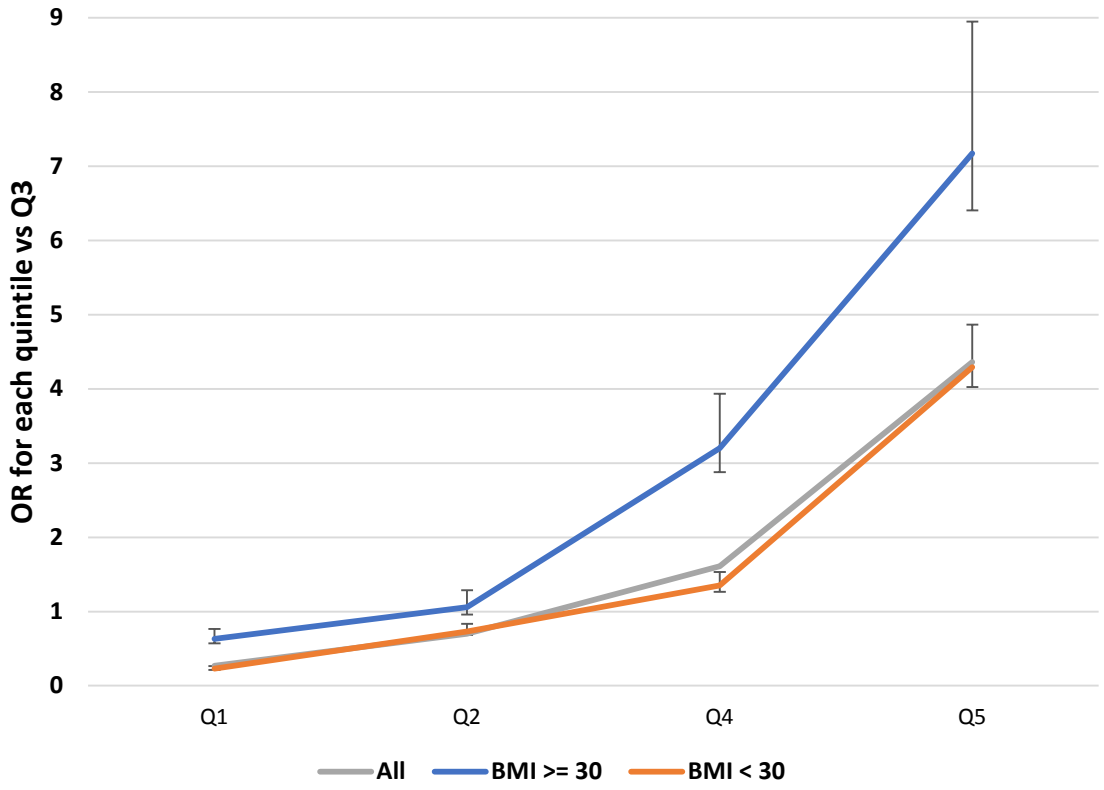
Body Mass Index (BMI) and Family History (FH)

Investigating on the interaction (BMI × PGS) and (FH × PGS)

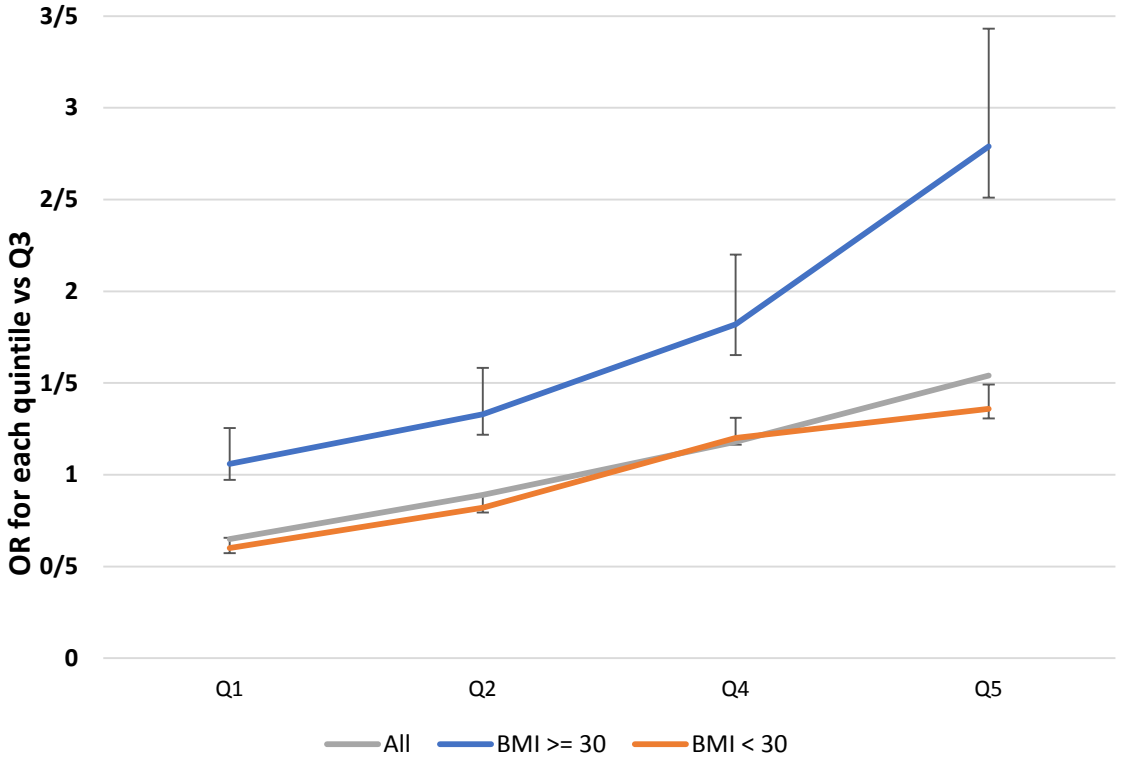
- ✓ **BMI: BMI ≥30, BMI < 30**
- ✓ **FH: Family history of T2D (related to the first degree of individual, GRM based selection)**
- ✓ **A multinomial logistic regression was employed to evaluate the genetic susceptibility risk for T2D within BMI categories.**
- ✓ **The model was adjusted for covariates, including age, sex, PRS, 10PCSs, BMI×PRS and FH×PRS.**

Investigating on the interaction (BMI × PGS)

T2D vs Normal



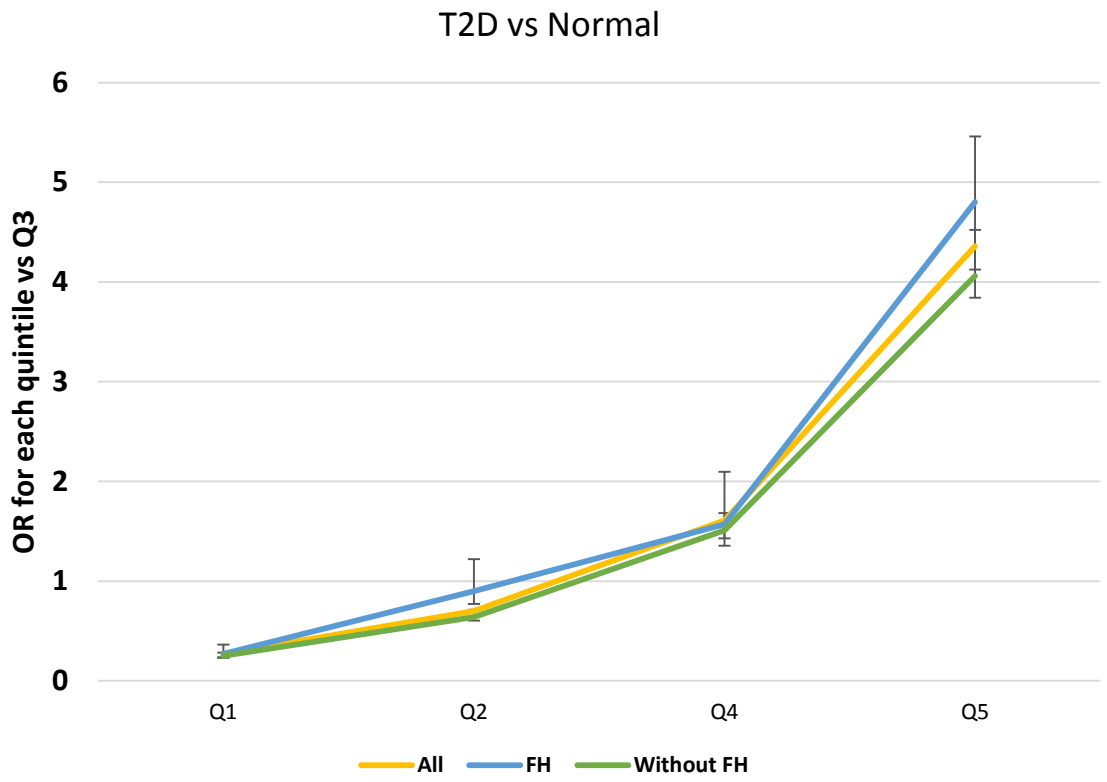
PreT2D vs Normal



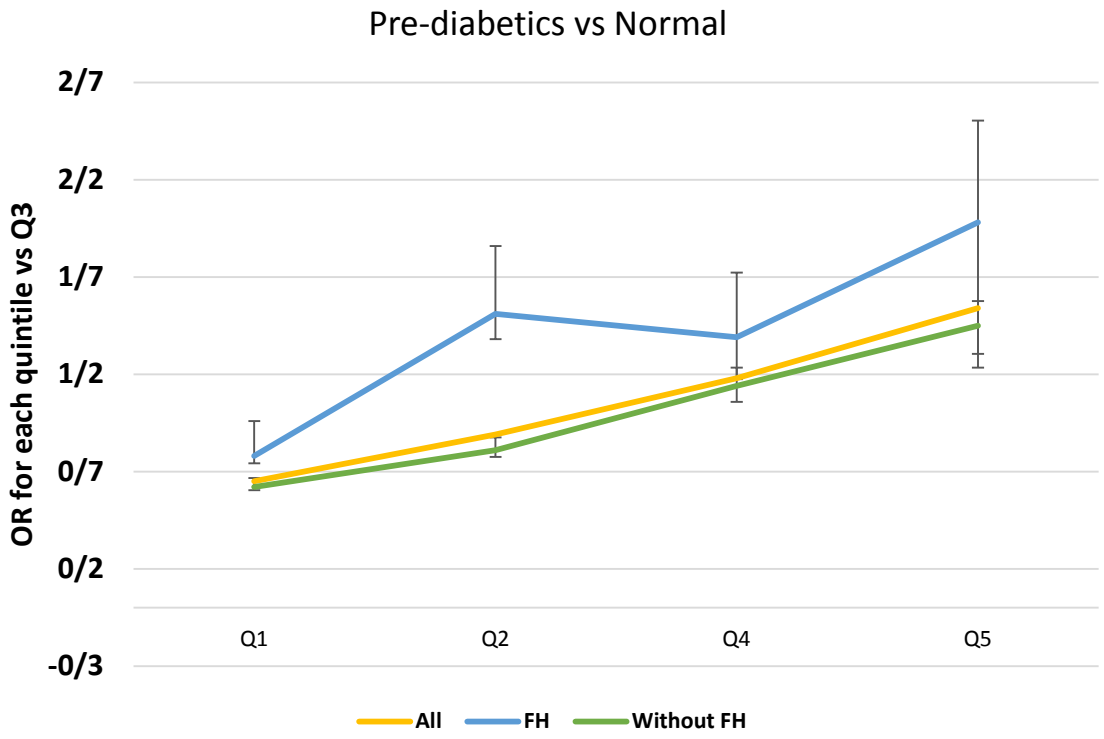
P-value of the interaction term in the model was not statistically significant.

Investigating on the interaction (FH × PGS)

T2D versus Normal people



PreT2D vs Normal people



P-value of the interaction term in the model was not statistically significant.

Regression to normoglycemia
and
Progression to T2D
among
Prediabetics

Regression to normoglycemia and Progression to T2D among Prediabetics

TLGS Phase	Normal	Prediabetes	T2D
ADA	5,780	4,130	2,167

Control

Excluded

Case

Unrelated Sample size in each group (grm0.05)

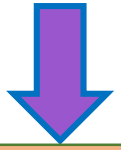
Normal	Prediabetes	T2D
1,746	1,764	1,228

Followed up for ~10yrs (Phase 3 to Phase6

Regression to normoglycemia and Progression to T2D among Prediabetics

Unrelated Sample size in each group (grm0.05)

Normal	Prediabetes	T2D
1,746	1,764	1,228



Followed up for ~10yrs (Phase 3 to Phase6

Prediabetes

1,764

Regressed to normal

120

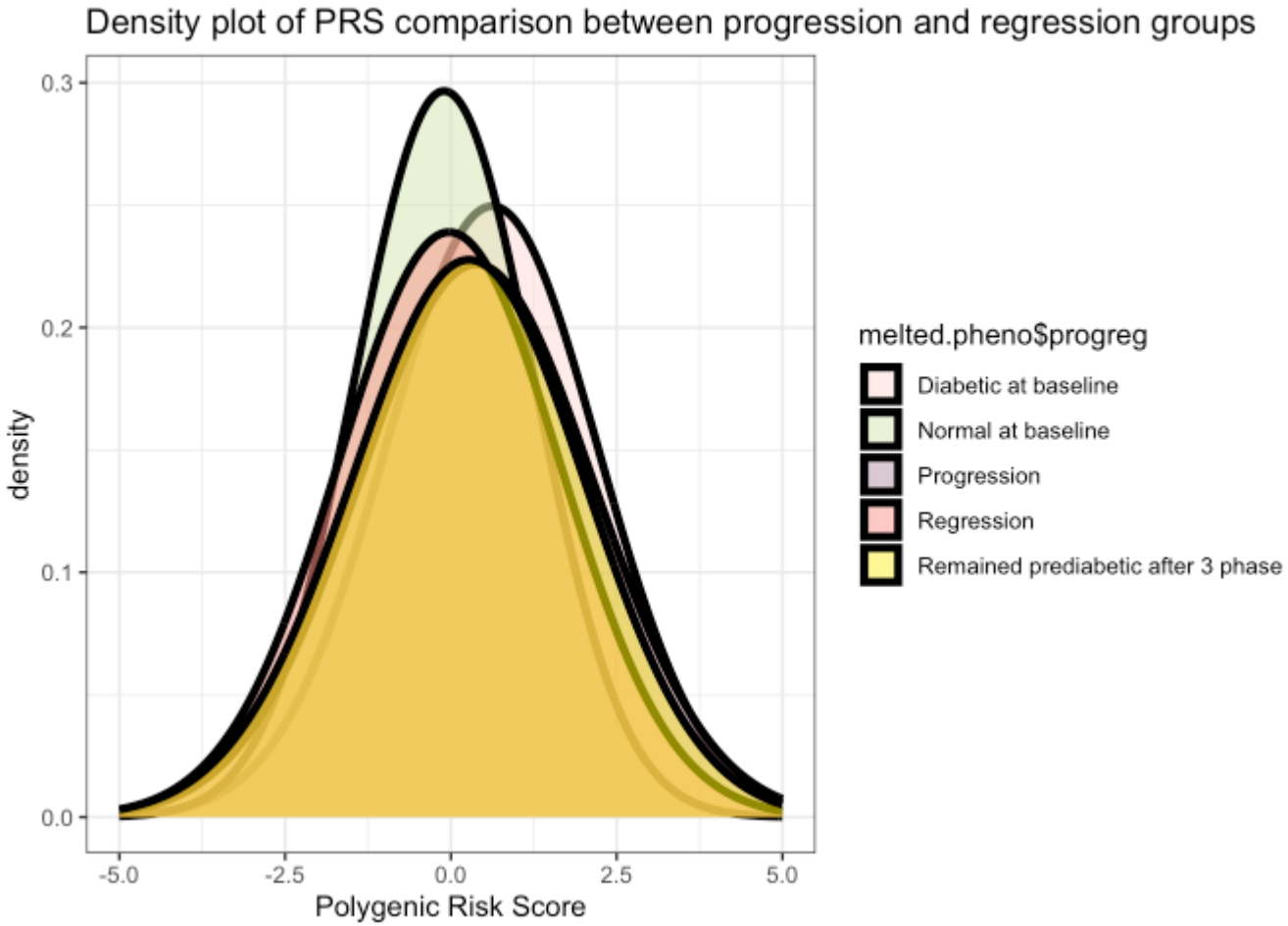
Remained Pre-T2D

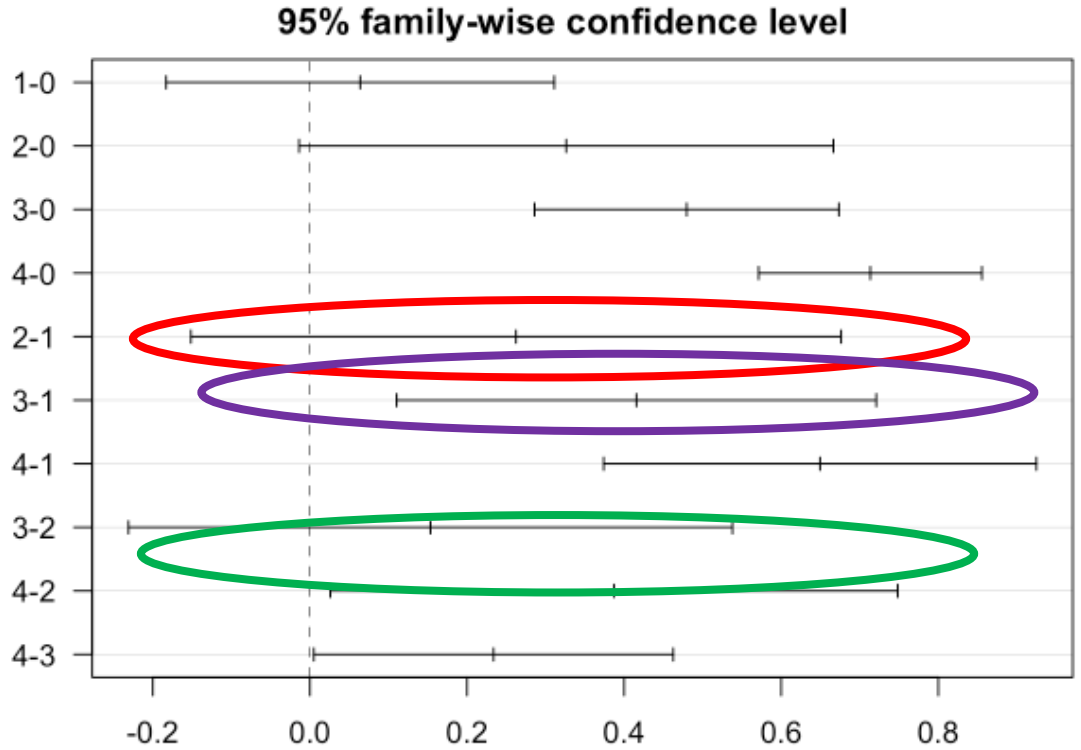
62

Progressed to T2D

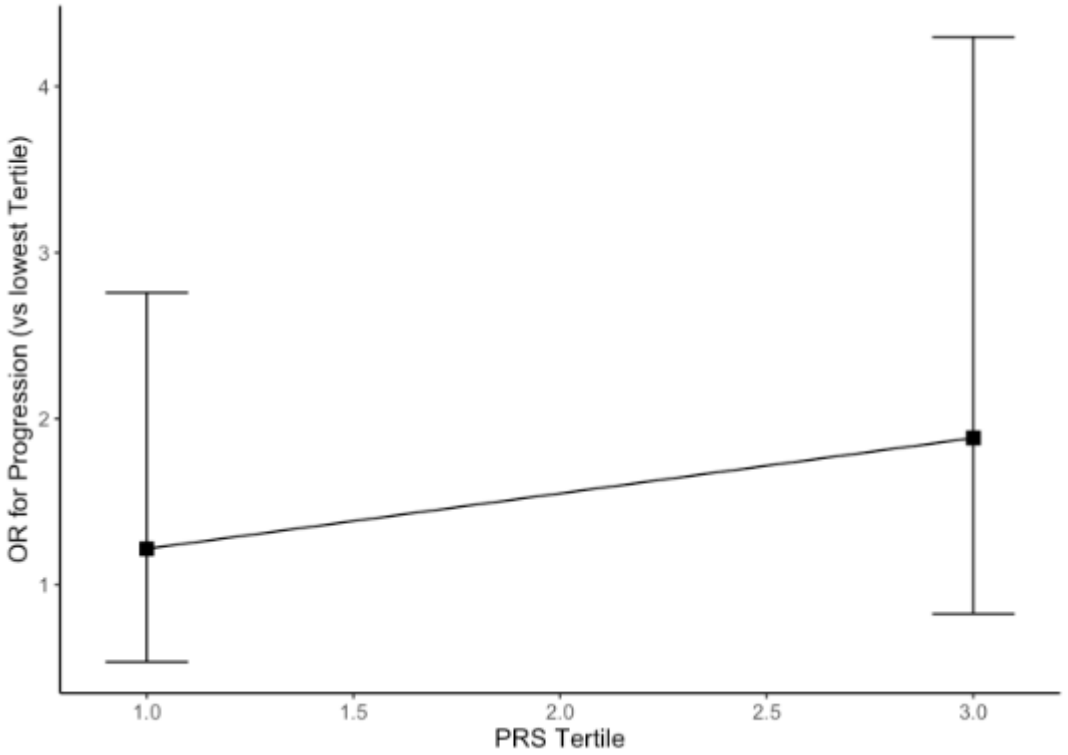
201

We followed the prediabetes participants through phase 3 to phase 6

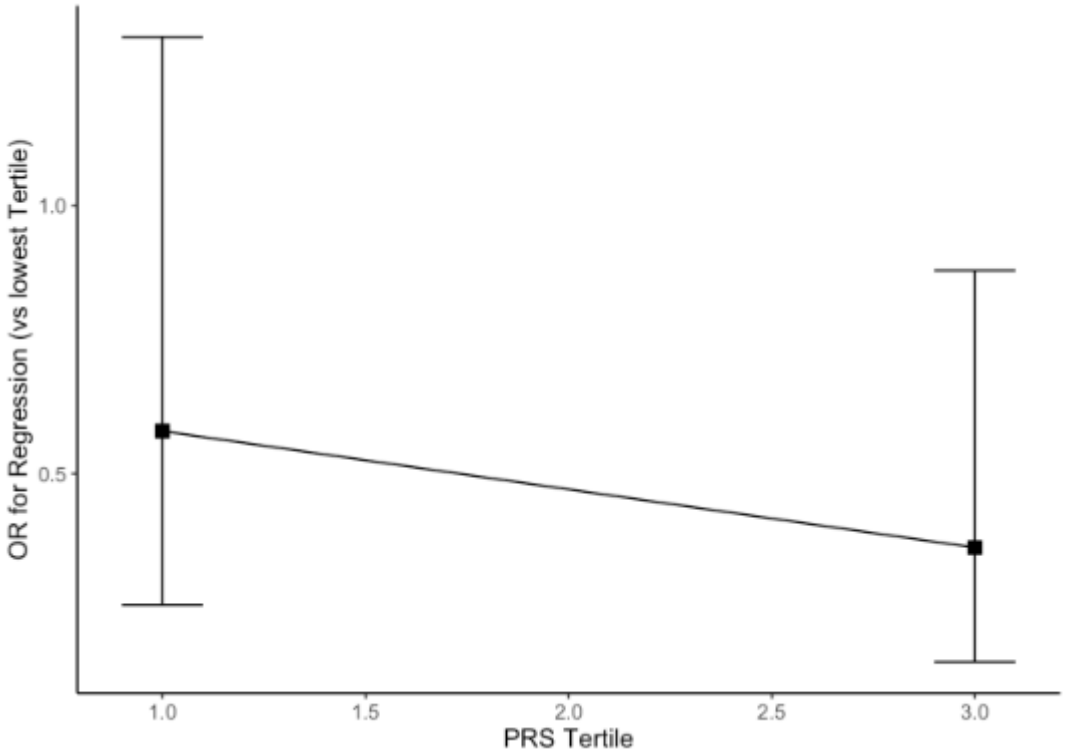




Progression vs remained prediabetics



Regression vs remained prediabetics

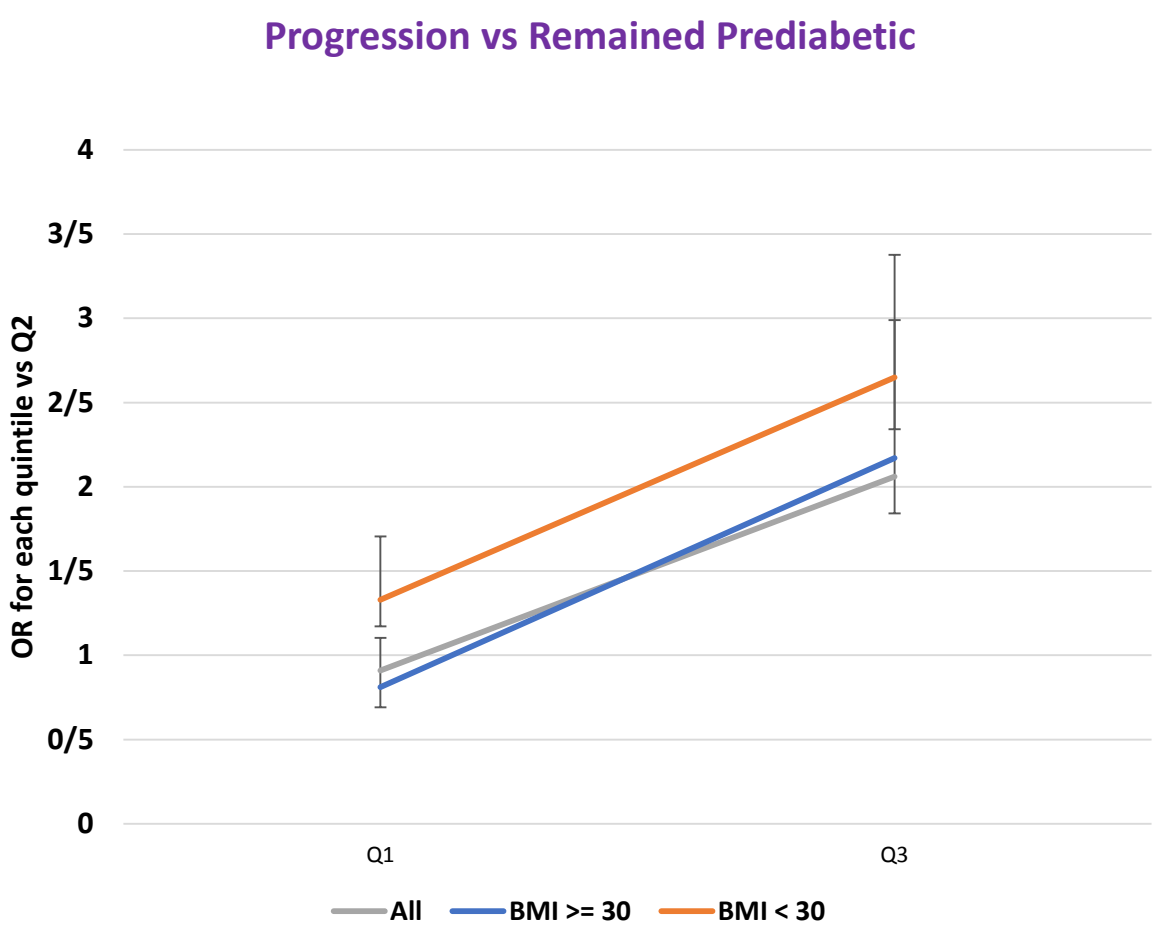


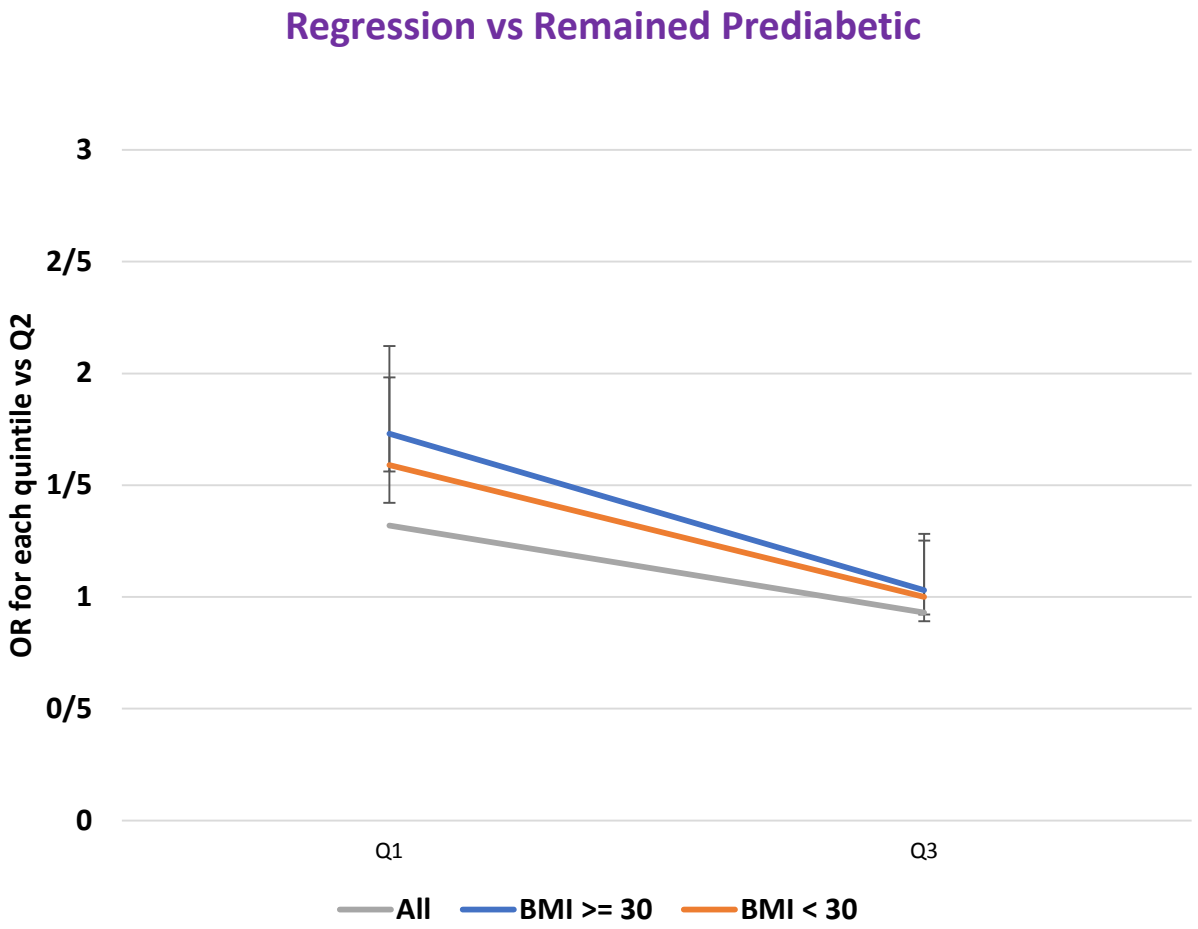
Interaction effect on Progression & Regression

(BMI × PGS & FH × PGS)

Body Mass Index (BMI) and Family History (FH)

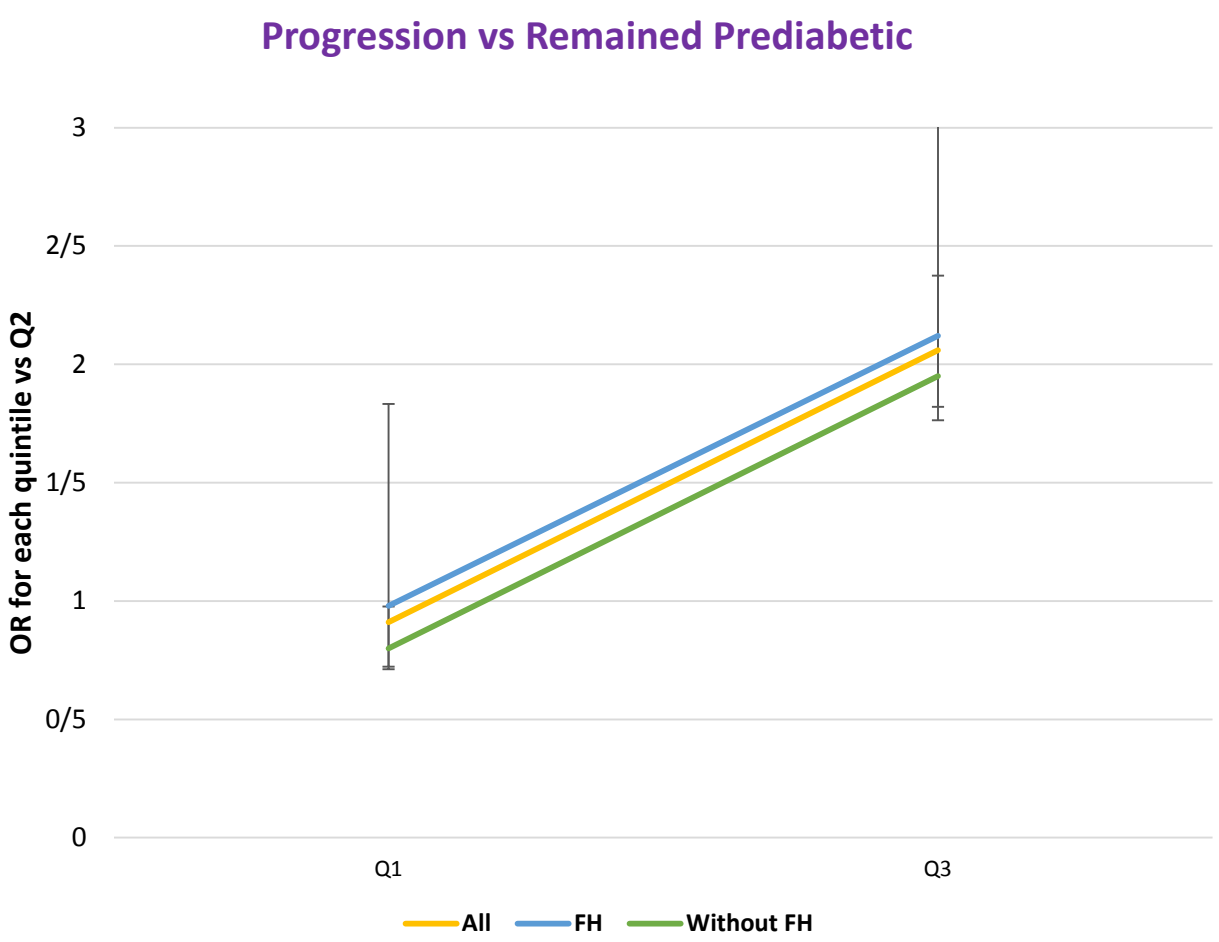
Investigating on the interaction (BMI × PGS)

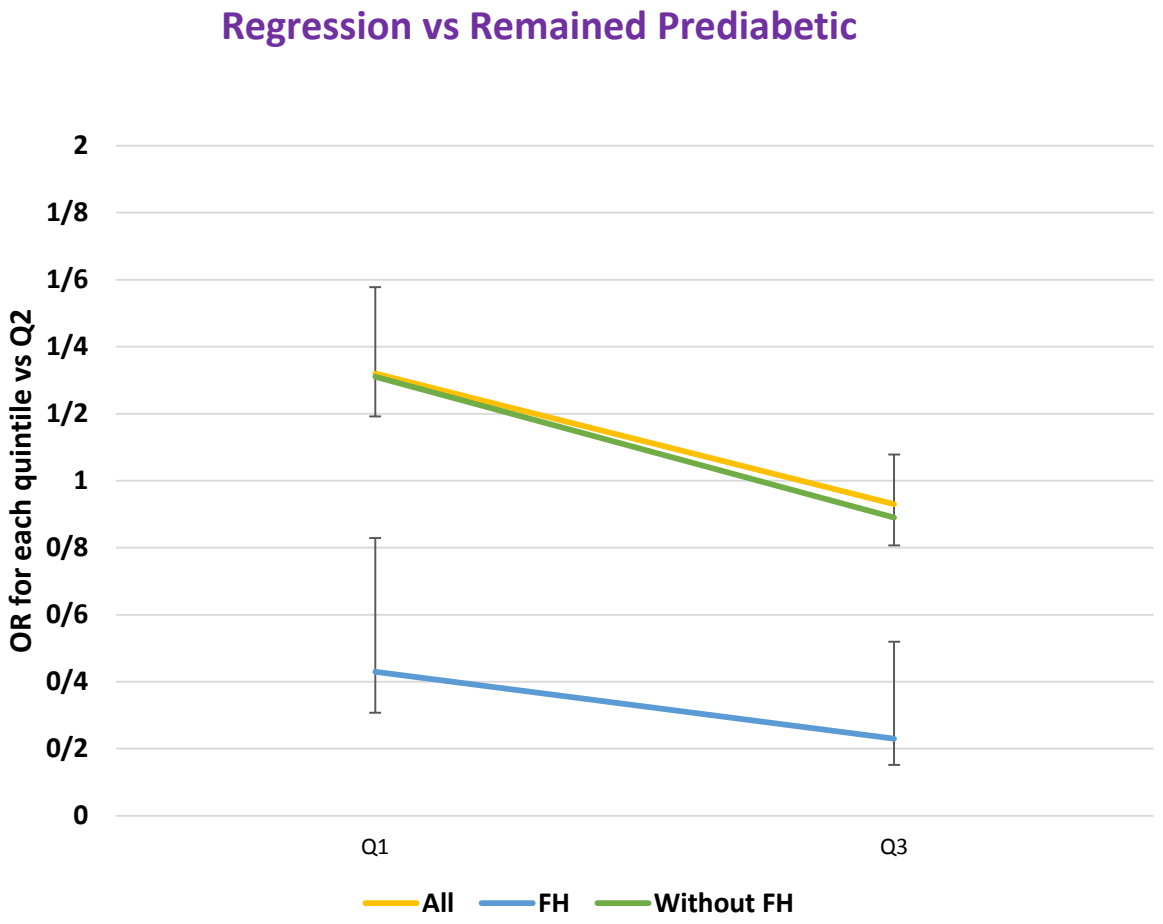




P-value of the interaction term in the model was not statistically significant.

Investigating on the interaction (FH × PGS)





P-value of the interaction term in the model was not statistically significant.

Conclusion

I. I need to add that the AUC of the prediction models by adding PRS were statistically higher than the models with only other covariates. So, adding a PRS to the models as a new risk factor can improve the prediction power.

II. As a result of this study, it is claimed that researchers or clinicians can independently rely on the PRS information, in addition the use of FH and/or BMI information.

GEMIRAN team

- Queensland University.
- Oxford University.
- University of Pennsylvania.
- University of Michigan.

AhTnk you!

Mahdi Akbarzadeh
akbarzadeh.ms@gmail.com

ThAnk you!

Mahdi Akbarzadeh
akbarzadeh.ms@gmail.com

Any Questions?

Mahdi Akbarzadeh
akbarzadeh.ms@gmail.com