

# *Scores* de risco poligênico

Prof. Michel Naslavsky

# O que determina o fenótipo?



Copyright © 2005 Pearson Prentice Hall, Inc.

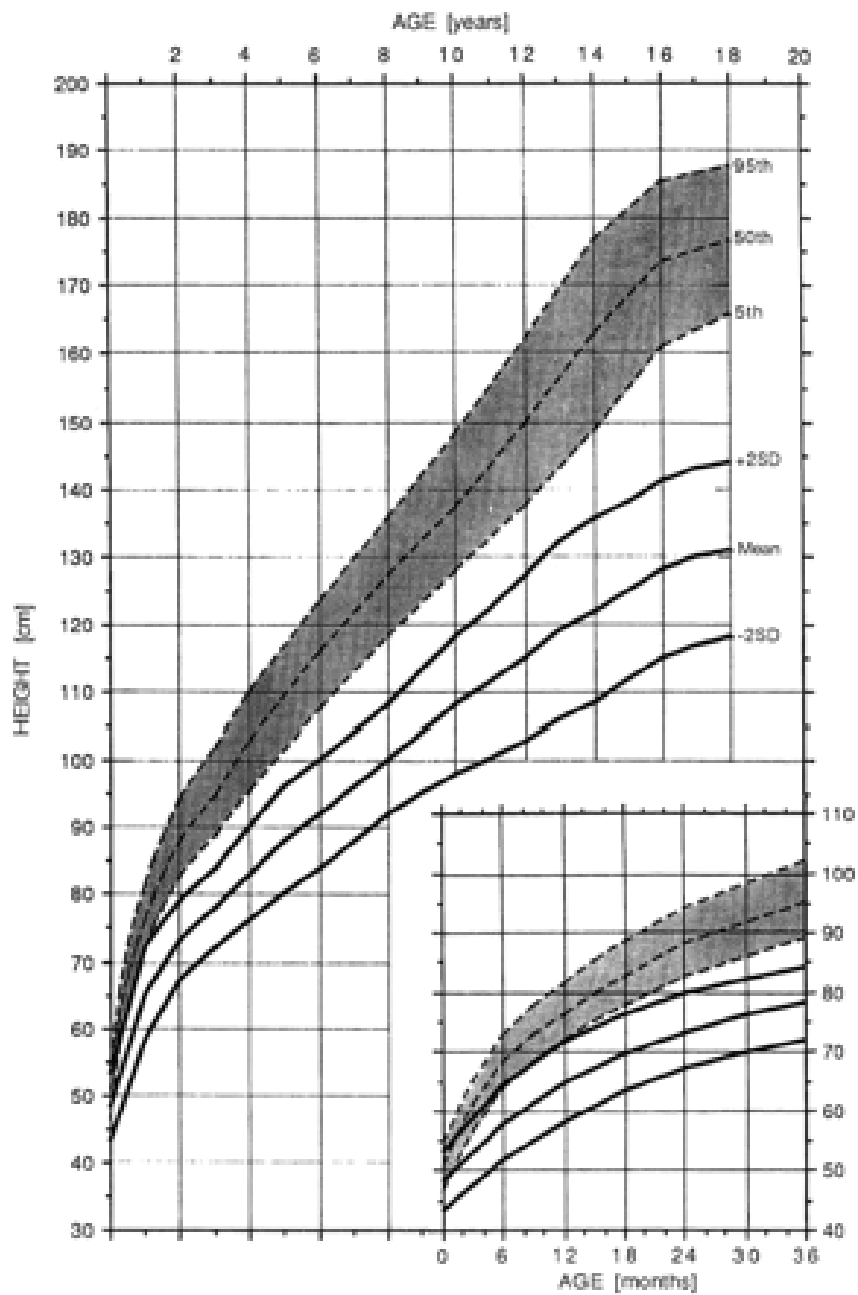




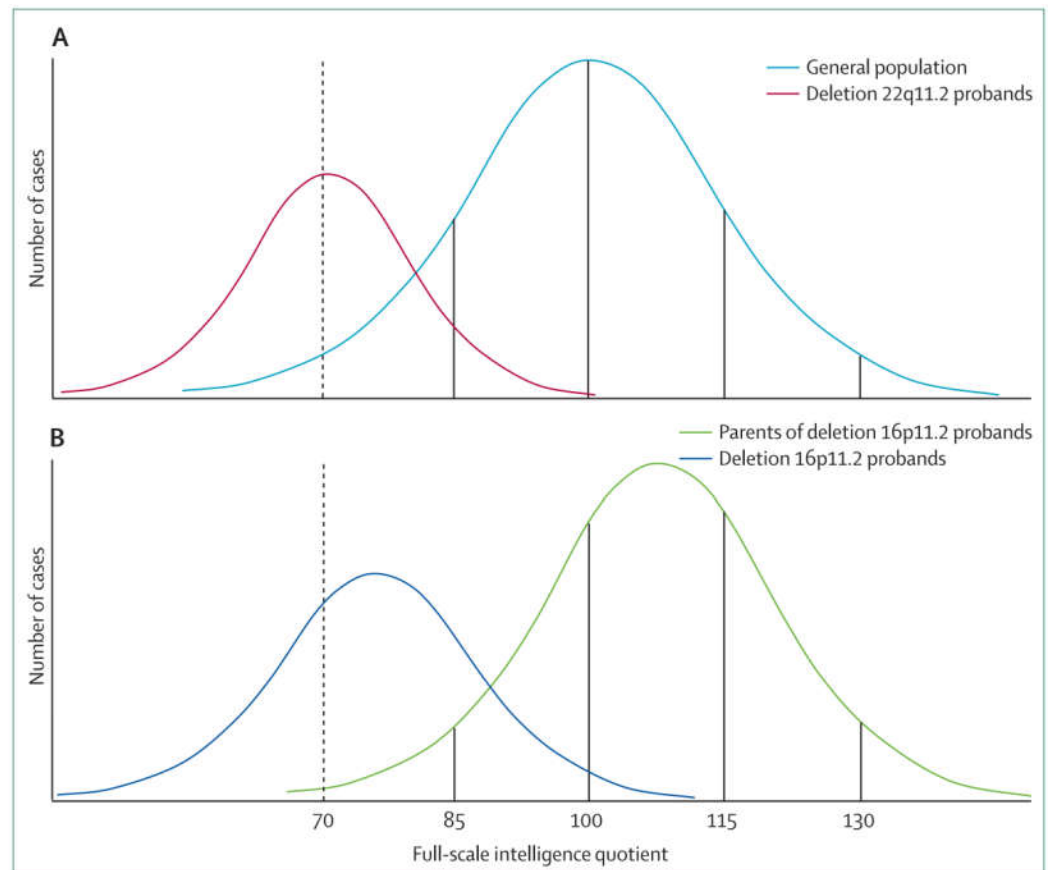
Chr	bp	Size (Mb)	Number of gen
1	115887358	115.89 Mb	2661
2	114274878	114.27 Mb	1736
3	101945329	101.95 Mb	1497
4	75099208	75.1 Mb	1028
5	81932842	81.93 Mb	1241
6	77372128	77.37 Mb	1401
7	79672333	79.67 Mb	1279
8	65453440	65.45 Mb	983
9	53729873	53.73 Mb	1082
10	69198901	69.2 Mb	1066
11	64180007	64.18 Mb	1633
12	65647084	65.65 Mb	1338
13	40819826	40.82 Mb	609
14	39562559	39.56 Mb	894
15	45545726	45.55 Mb	941
16	38659571	38.66 Mb	1094
17	46773318	46.77 Mb	1520
18	32188280	32.19 Mb	409
19	31689818	31.69 Mb	1743
20	30178018	30.18 Mb	752
21	14992224	14.99 Mb	344
22	21403934	21.4 Mb	616
X	52468705	52.47 Mb	1110
Y	3528703	3.53 Mb	113

Cariograma:  
Representação do cariótipo em imagens

National Human Genome Research Institute



*J Pediatr.* 1978;93:435-438



Lancet Neurol 2013; 12: 406–14

Fenótipo

=

Genótipos

Quanto?

+

Ambiente

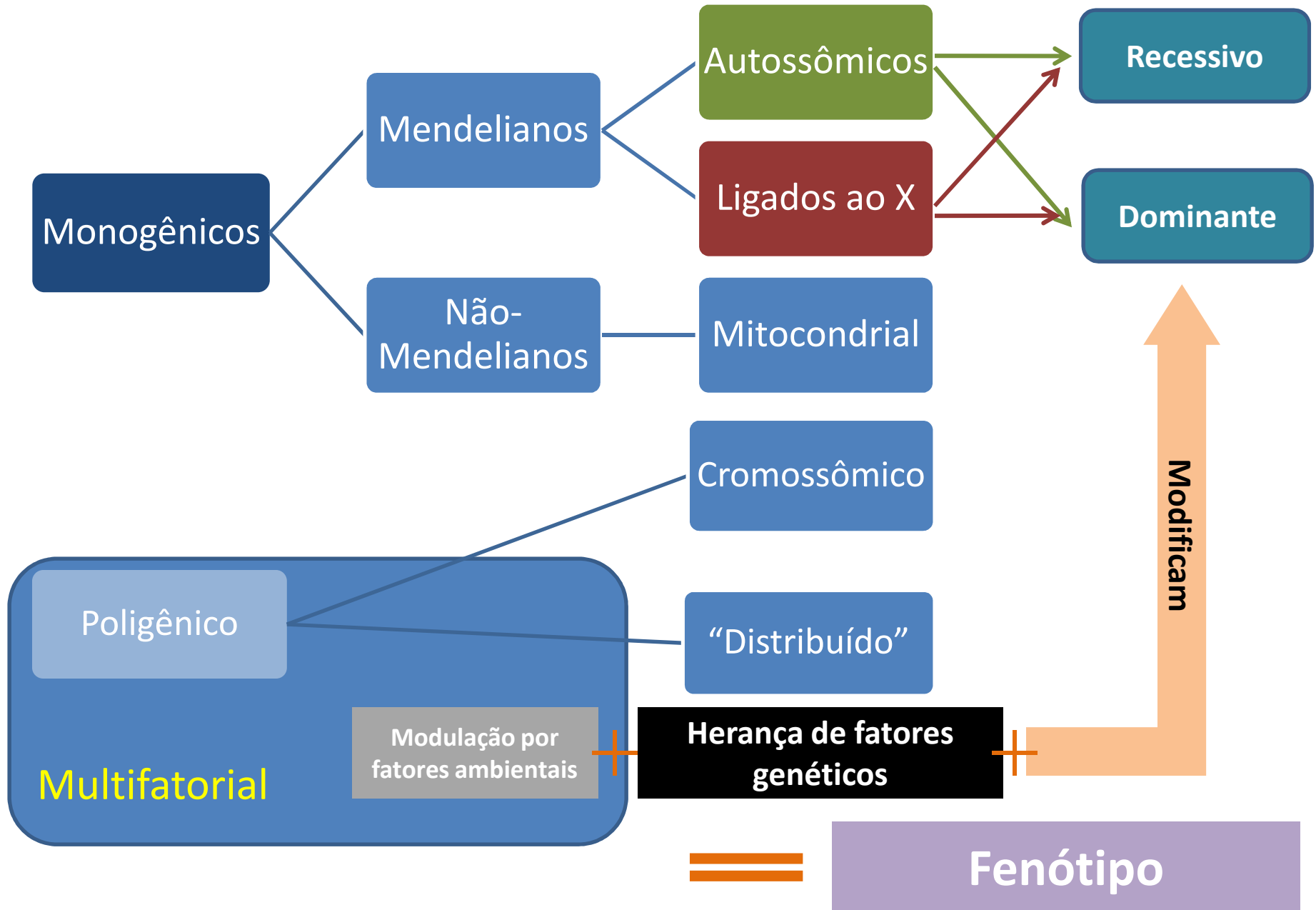
Quanto?

+

Interação Genótipo-Ambiente

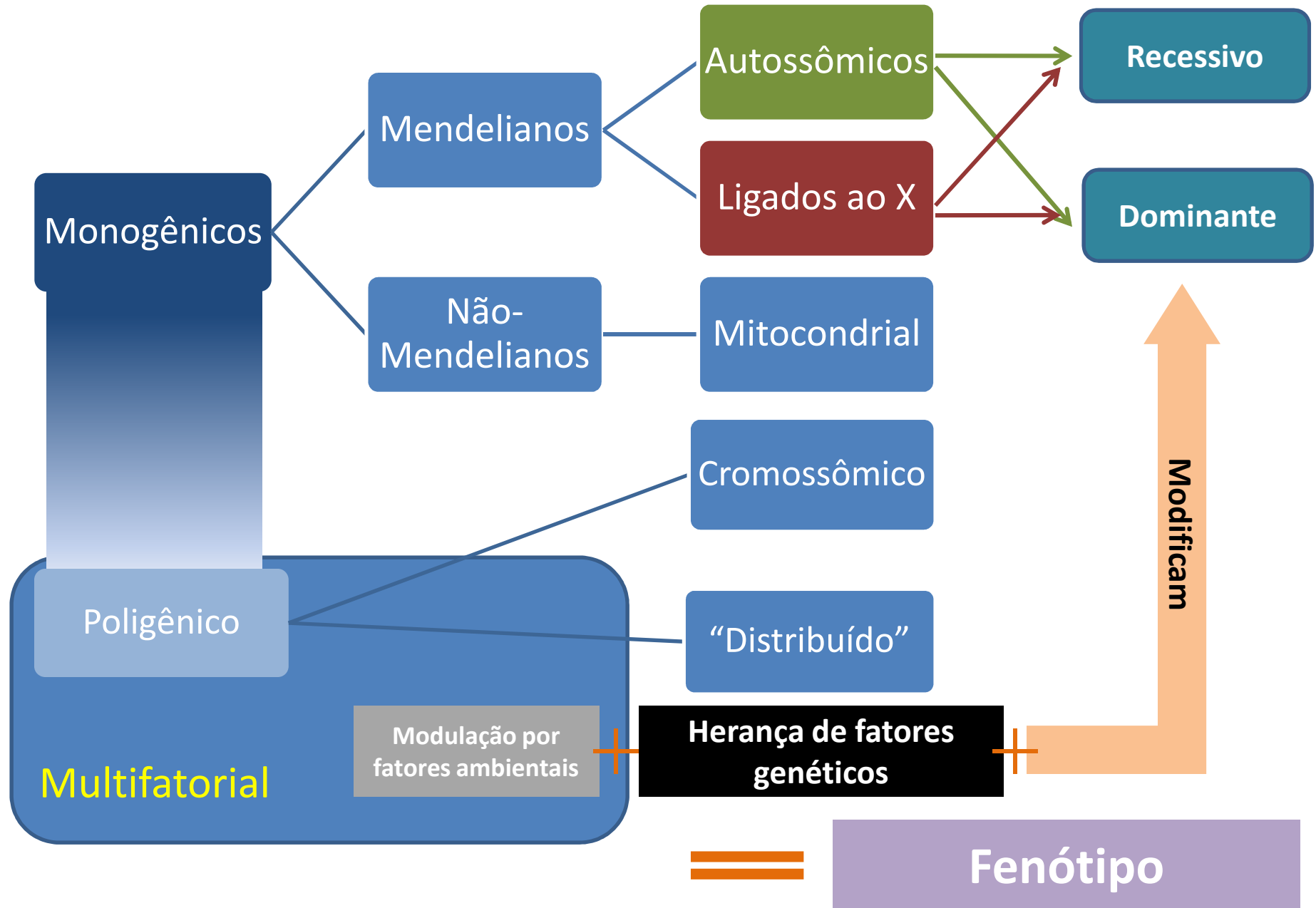
Quanto?

# Padrões de herança

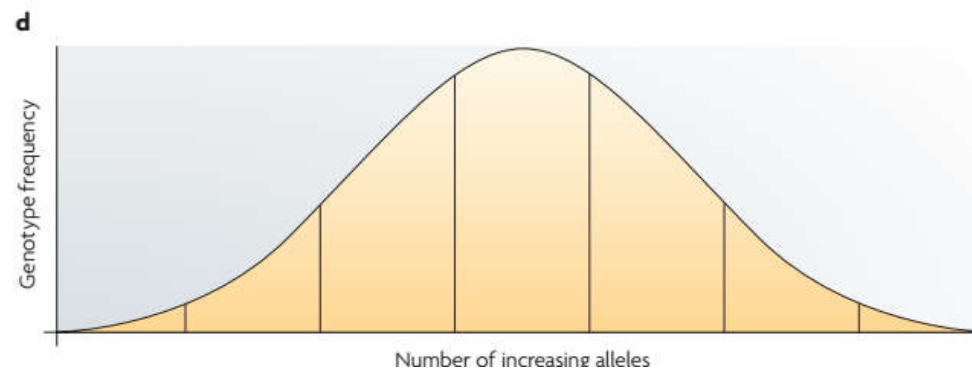
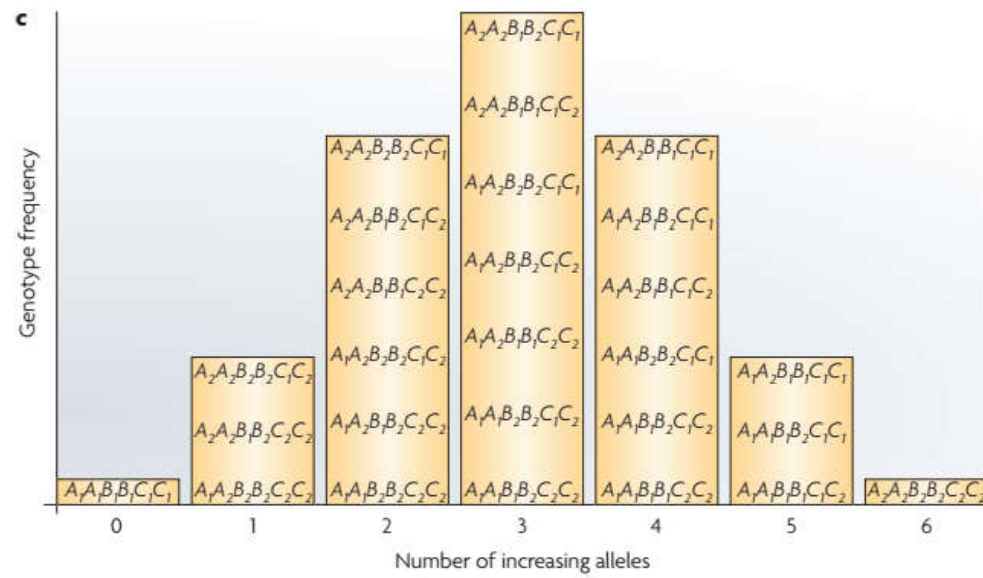
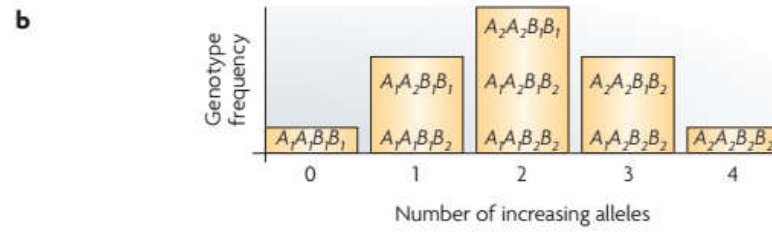




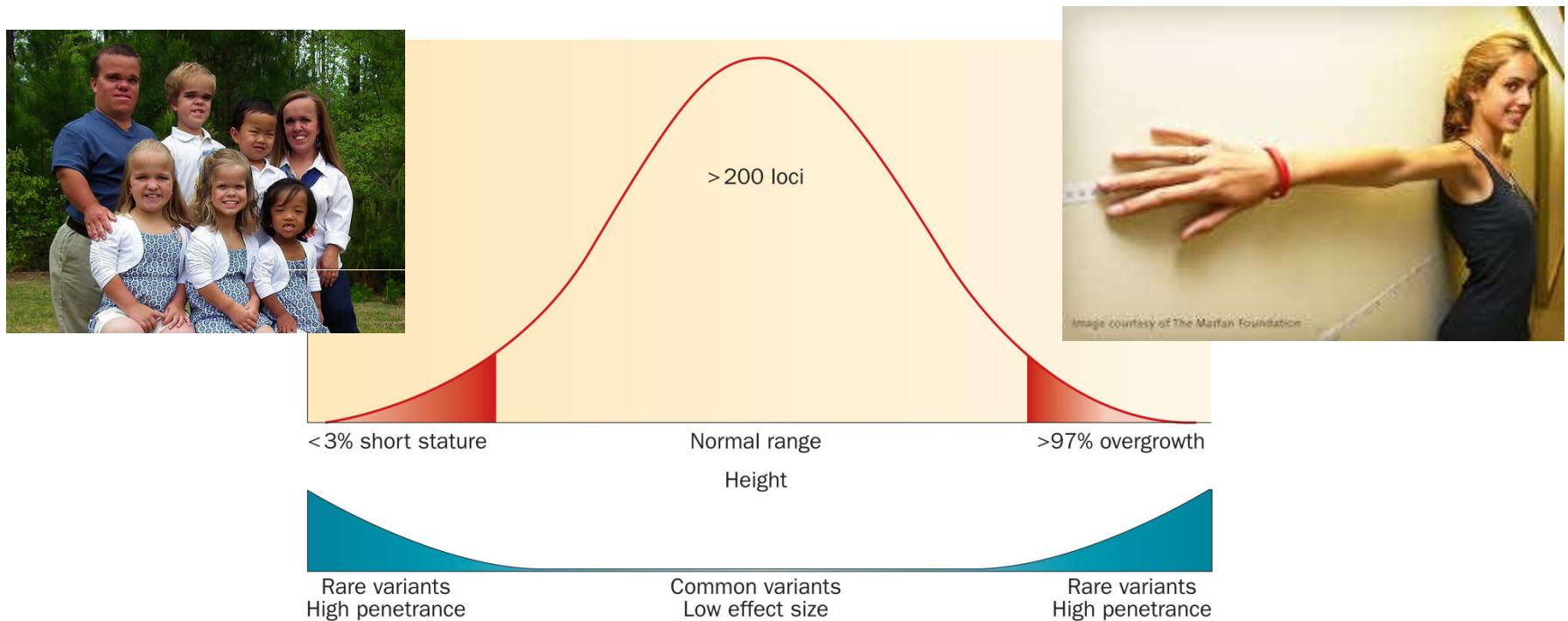
# Padrões de herança



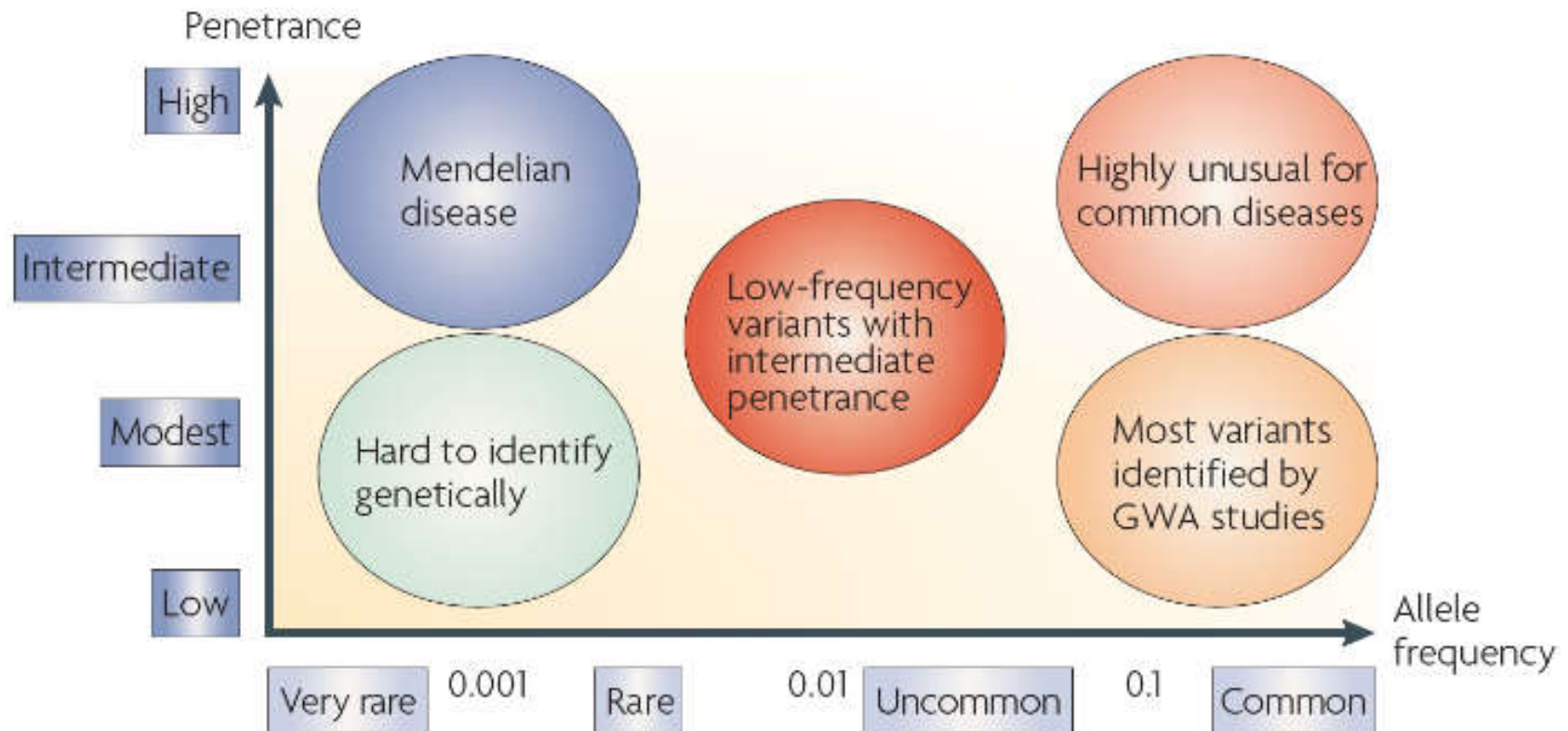




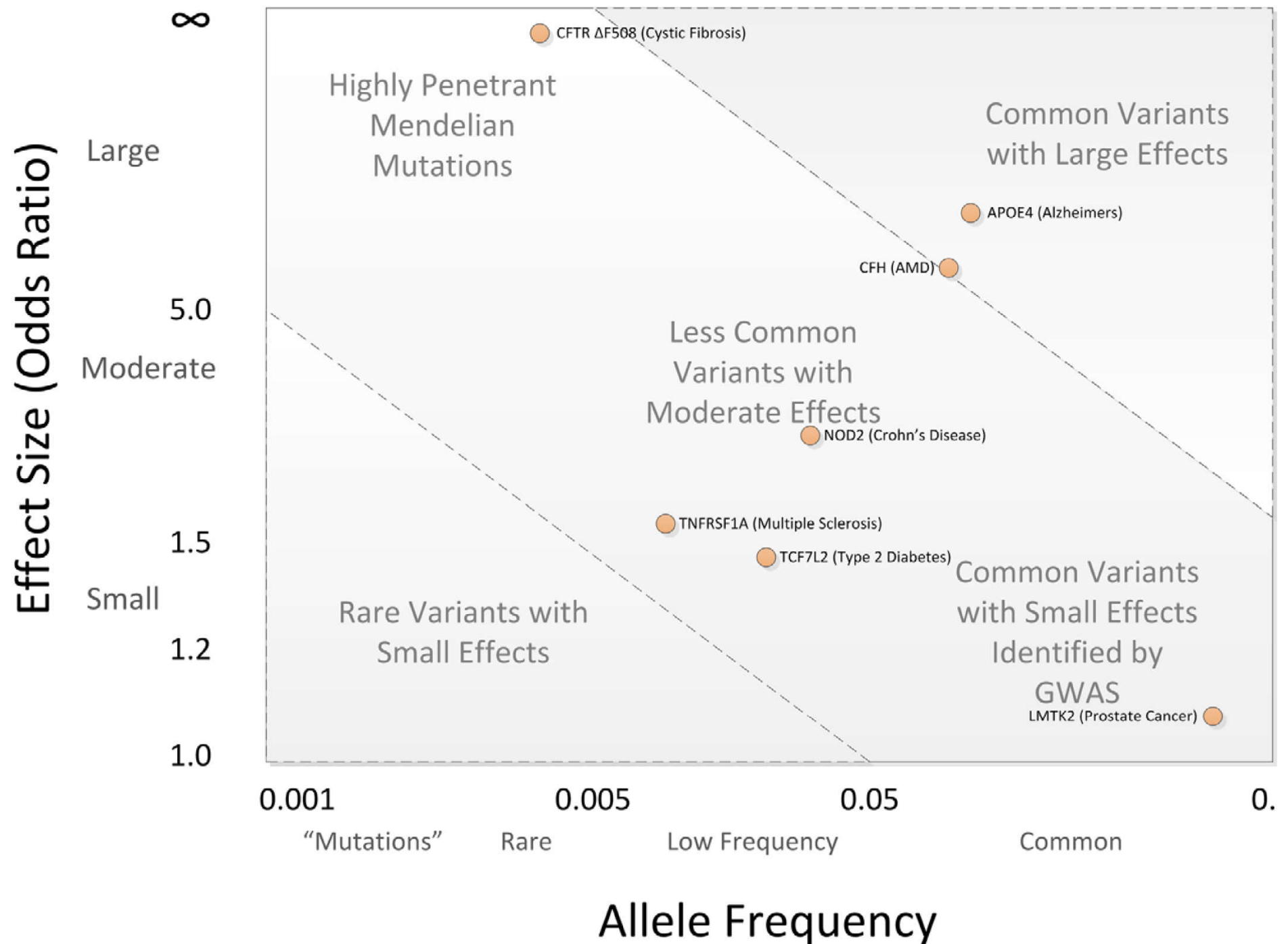
# Penetrância = Tamanho do efeito

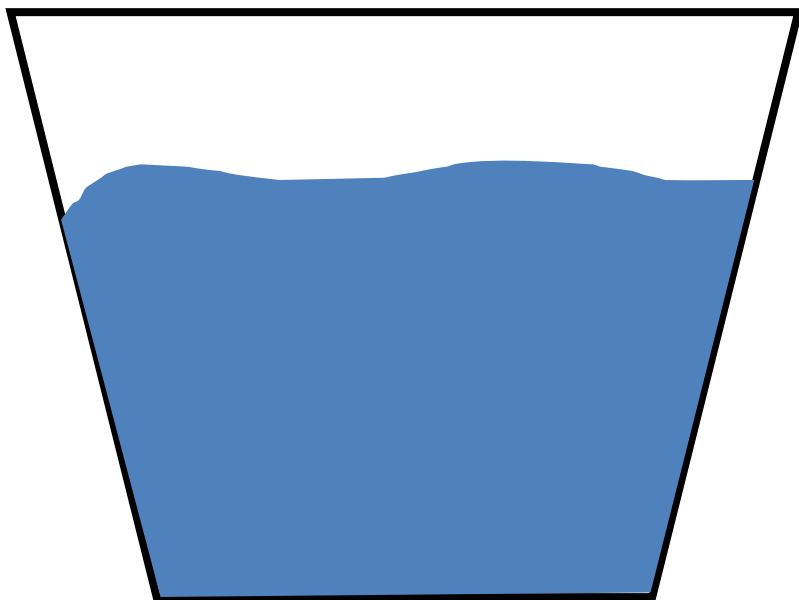


- Durand & Rappold, Height matters-from monogenic disorders to normal variations. Nature Reviews Endocrinology, 2013



- McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 367 (2009)







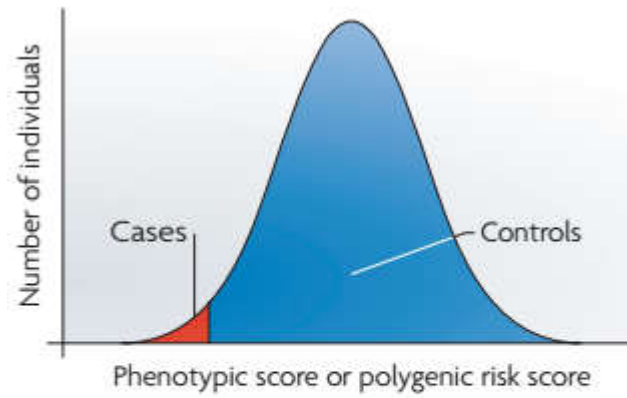
- Variante 1
- Variante 2
- Variante 3
- .....  
• Variante 1 milhão

# GWAS como ferramenta

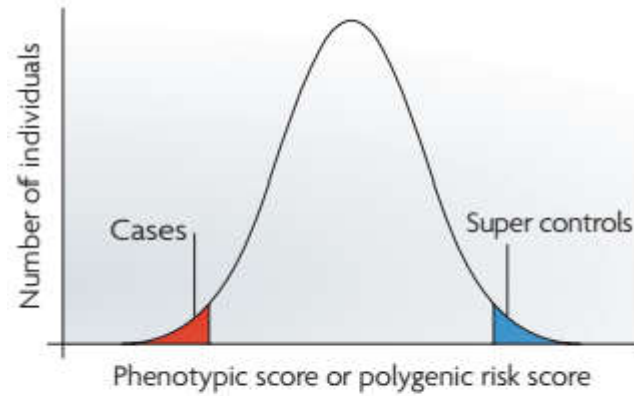
- Princípio 1: indivíduos não aparentados (ou “pouco aparentados”) com variância no fenótipo estudado
  - Caso-controle: fenótipos discretos/qualitativos
  - Curva de distribuição contínua: fenótipos contínuos/quantitativos
- \*Fenotipagem precisa e padronizada é essencial



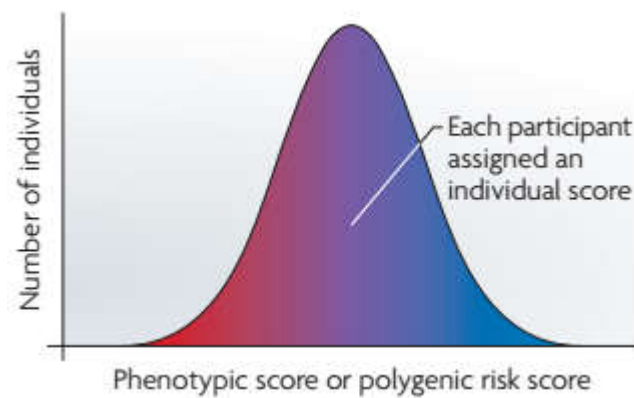
**a Case-control**



**b Extreme selection**



**c Quantitative measurement**

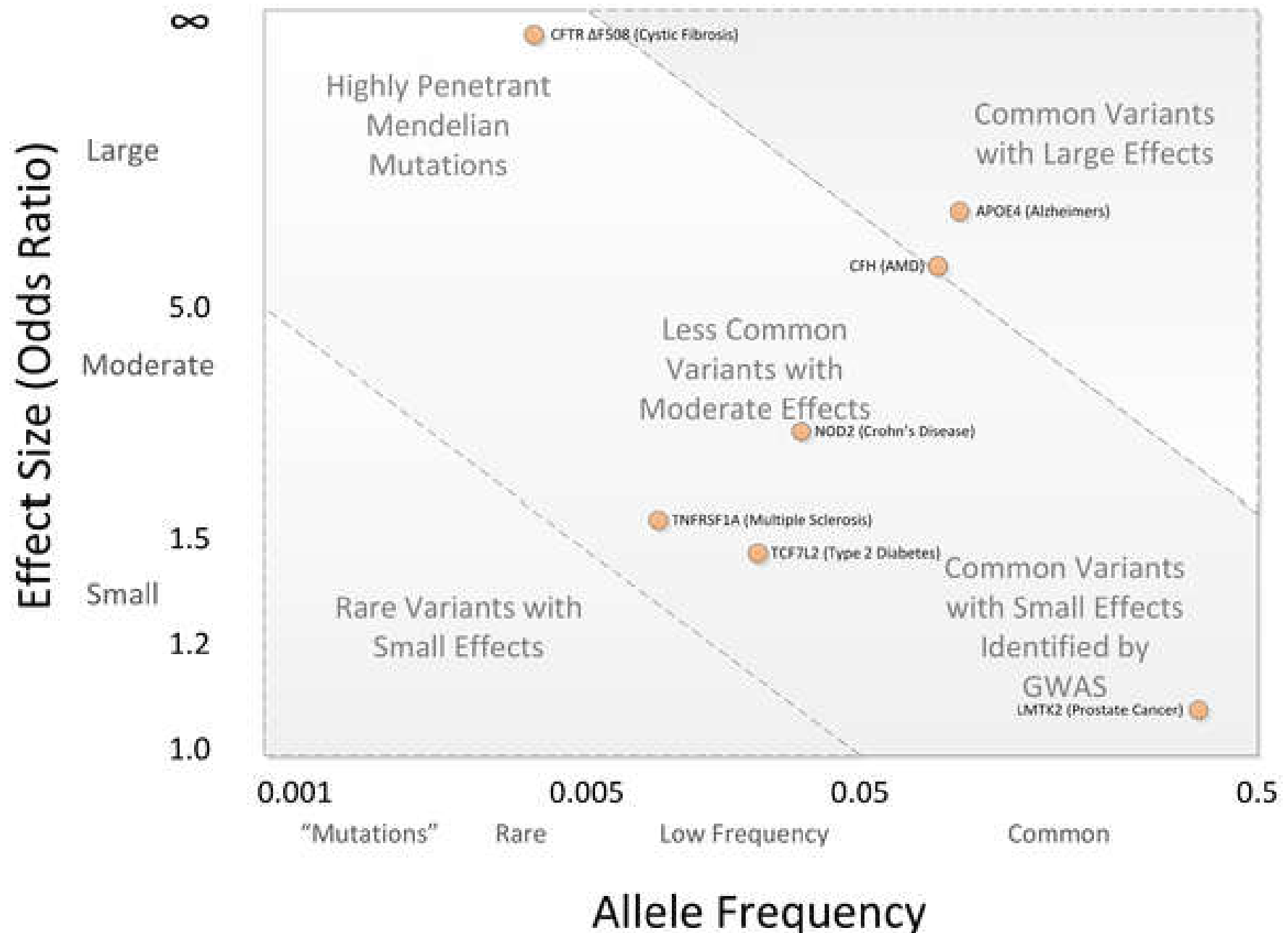


# GWAS como ferramenta

- Princípio 2: uso de variantes comuns

( $q > 0.5\%$ / $q > 1\%$ )  $\rightarrow$  vantagem de um  $m$  (número de marcadores) elevado ( $>100k$ )

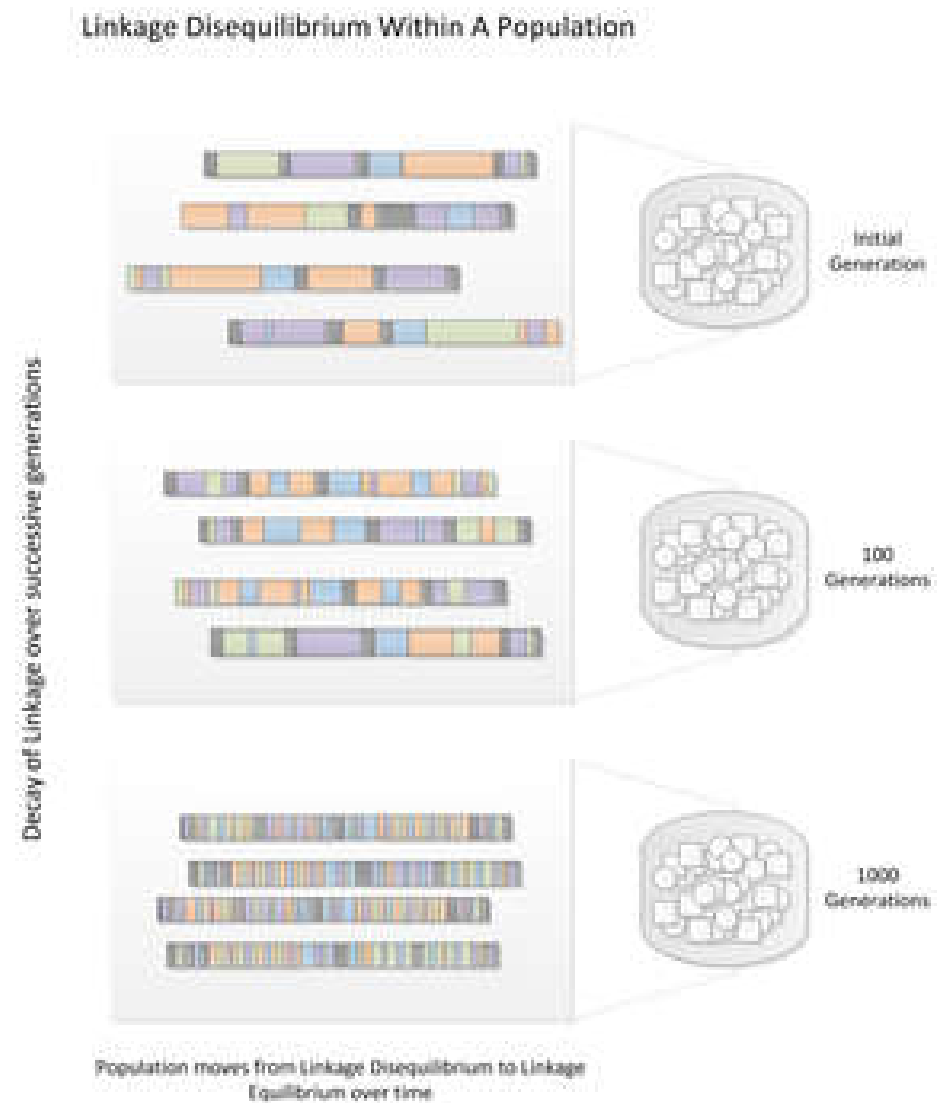
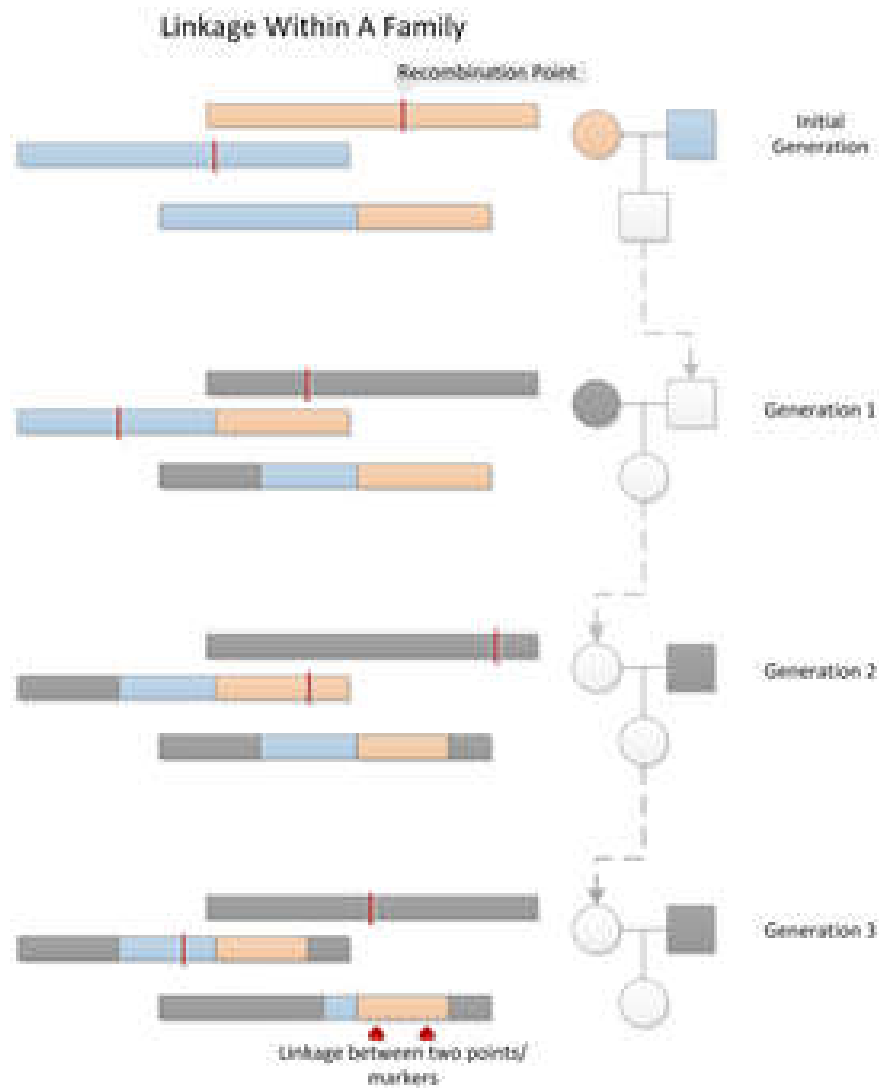
- Dependência de um efeito moderado para a variante individual, parcialmente compensado com o  $N$



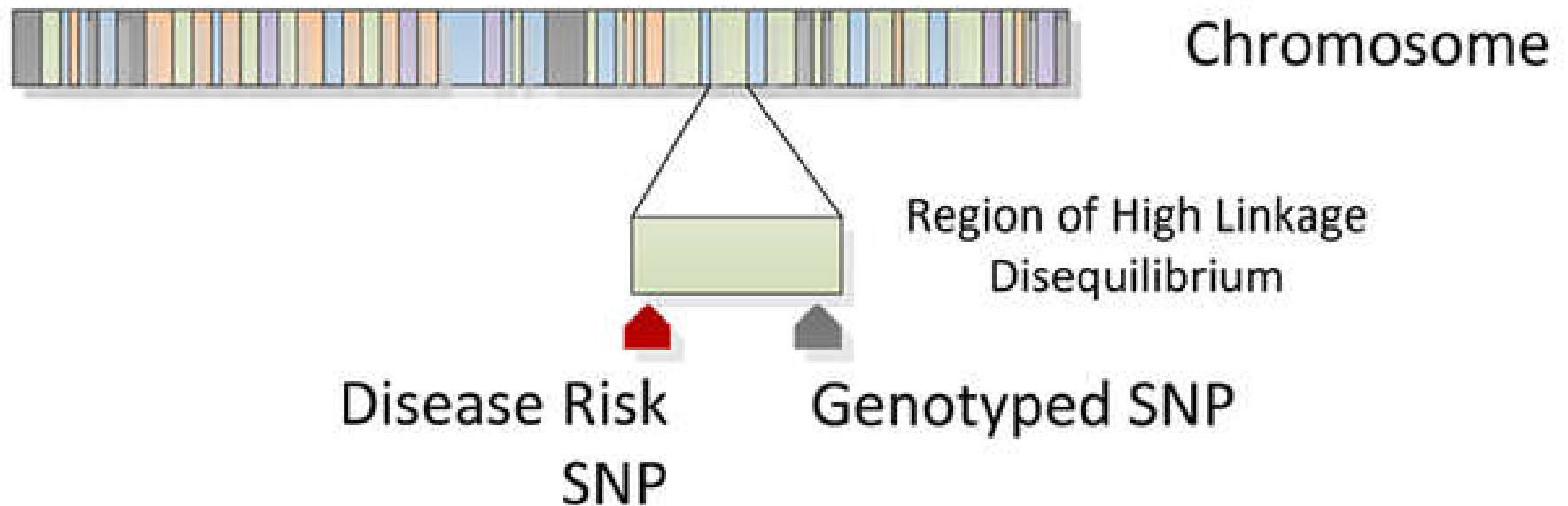
Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology 8(12): e1002822.  
<https://doi.org/10.1371/journal.pcbi.1002822>  
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822>

# GWAS como ferramenta

- Princípio 3: estrutura de desequilíbrio de ligação entre um polimorfismo (ou alguns) com um grande sinal de associação e a variação causal (caso não seja o próprio);
- Este princípio também baseia os estudos de famílias, mas há uma mudança de escala (geracional e, por consequência, física).

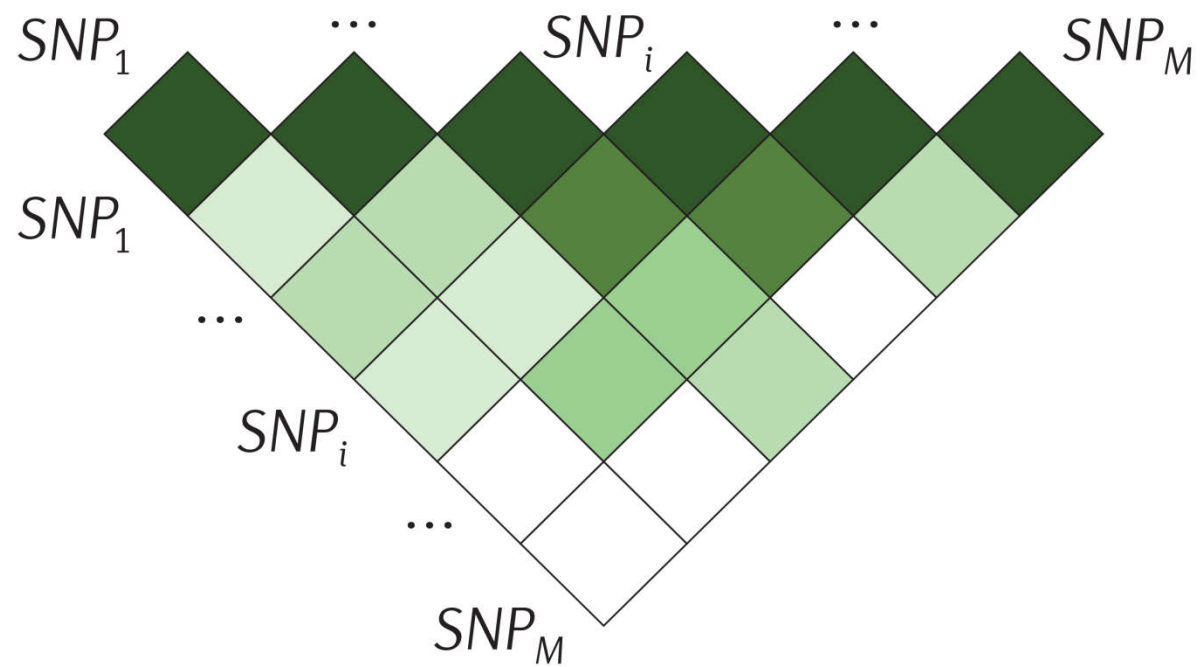


# Indirect Association



Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology 8(12): e1002822.  
<https://doi.org/10.1371/journal.pcbi.1002822>  
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822>

SNP LD (V)

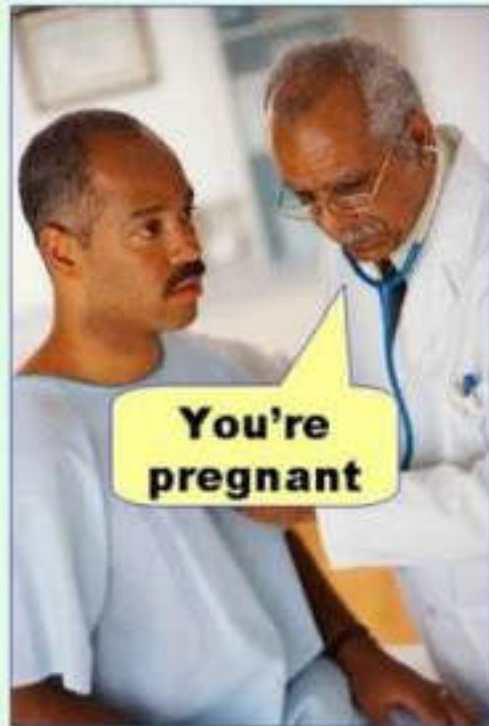




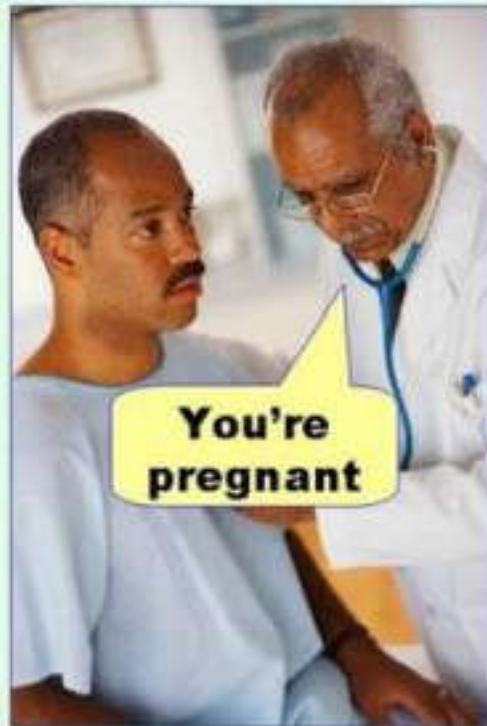
# Odds ratio e tabela de contingência

Table of error types		Null hypothesis ( $H_0$ ) is	
		True	False
Decision about null hypothesis ( $H_0$ )	Fail to reject	<b>Correct inference (true negative)(probability = <math>1 - \alpha</math>)</b>	<b>Type II error (false negative) (probability = <math>\beta</math>)</b>
	Reject	<b>Type I error (false positive)(probability = <math>\alpha</math>)</b>	<b>Correct inference (true positive) (probability = <math>1 - \beta</math>)</b>

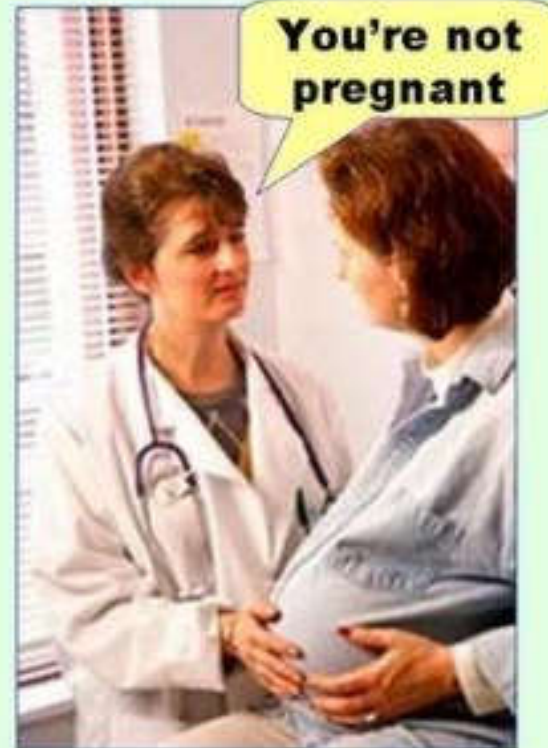
**Type I error**  
(false positive)



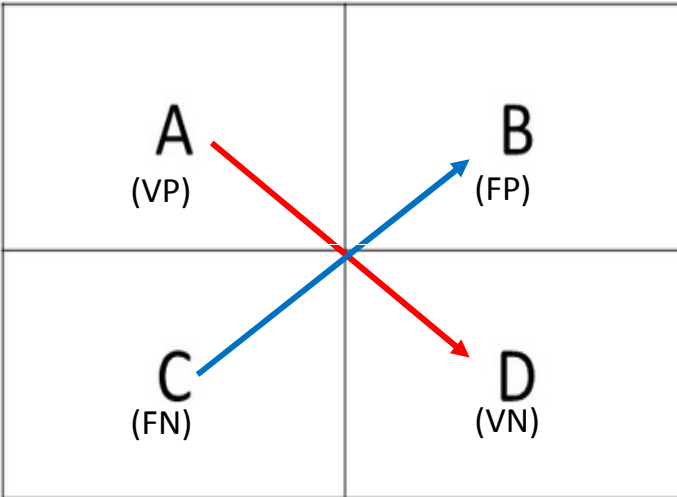
**Type I error**  
(false positive)



**Type II error**  
(false negative)



		Outcome	
		Yes	No
Predictor	Yes	A (VP)	B (FP)
	No	C (FN)	D (VN)



The diagram shows a 2x2 confusion matrix. A red arrow points from the top-left cell (A, VP) to the bottom-right cell (D, VN). A blue arrow points from the bottom-left cell (C, FN) to the top-right cell (B, FP).

$$OR = \frac{(A * D)}{(B * C)}$$

	Câncer de pulmão	Sem cancer de pulmão	Total
Tabagistas			
Não tabagistas			
Total			

# Caso controle

	Câncer de pulmão	Sem cancer de pulmão	Total
Tabagistas			
Não tabagistas			
Total	100	100	200

# Caso controle

	Câncer de pulmão	Sem cancer de pulmão	Total
Tabagistas	80		
Não tabagistas	20		
Total	100	100	200



# Caso controle

	Câncer de pulmão	Sem cancer de pulmão	Total
Tabagistas	80	30	
Não tabagistas	20	70	
Total	100	100	200

# Caso controle

	Câncer de pulmão	Sem cancer de pulmão	Total
Tabagistas	80	30	110
Não tabagistas	20	70	90
Total	100	100	200

# Caso controle

	Câncer de pulmão	Sem cancer de pulmão	Total
Tabagistas	80	30	110
Não tabagistas	20	70	90
Total	100	100	200

Quantas vezes mais fumantes tiveram cancer de pulmão em relação aos que não tiveram:

$$\frac{80}{30} = 2,67$$

Quantas vezes menos não fumantes desenvolveram câncer de pulmão em relação aos que não desenvolveram:

$$\frac{20}{70} = 0.29$$

Quantas vezes é **mais provável** ter **cancer de pulmão sendo tabagista** em relação a **ser tabagista** mas não ter o cancer?

$$\text{Odds Ratio (OR)} = (80 \cdot 70) / (20 \cdot 30) = 9,33$$

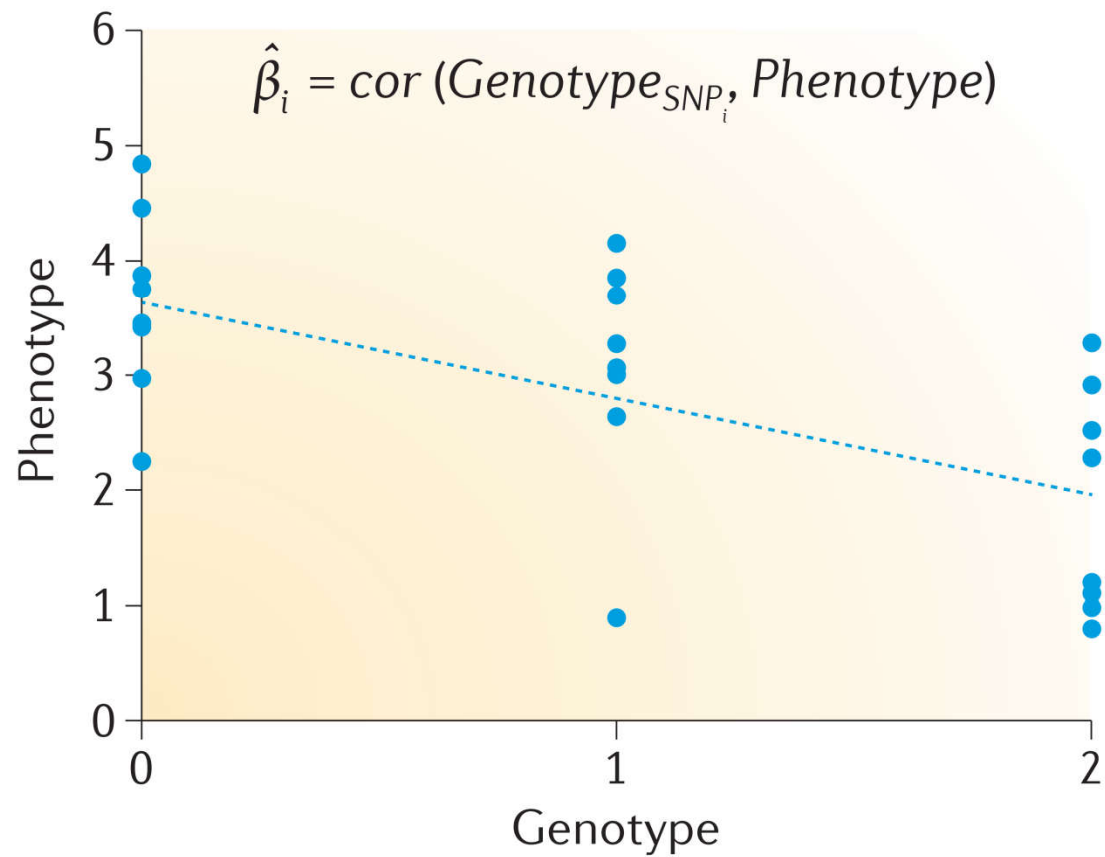
# Exemplo: fator de risco individual vs. doença

$OR = 1 \rightarrow$  a exposição ao preditor/fator de risco não interfere na chance de ocorrência da doença

$OR < 1 \rightarrow$  a exposição ao preditor/fator de risco interfere diminuindo a chance de ocorrência da doença

$OR > 1 \rightarrow$  a exposição ao preditor/fator de risco interfere aumentando a chance de ocorrência da doença

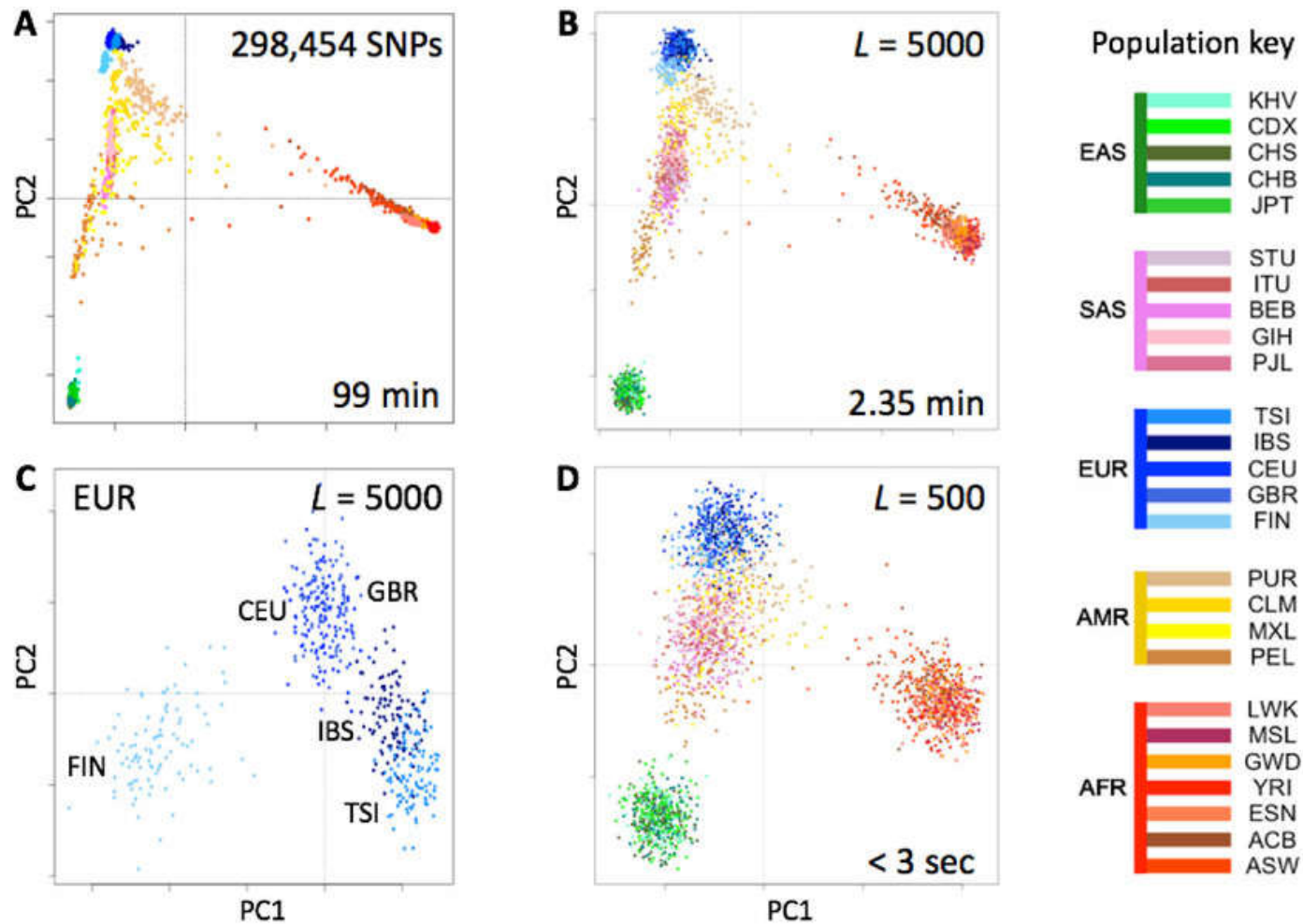
Effect size at  $\text{SNP}_i$



z-scores

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}, \dots, \frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)}, \dots, \frac{\hat{\beta}_M}{s.e(\hat{\beta}_M)} \sim \text{MVN}(0, V)$$

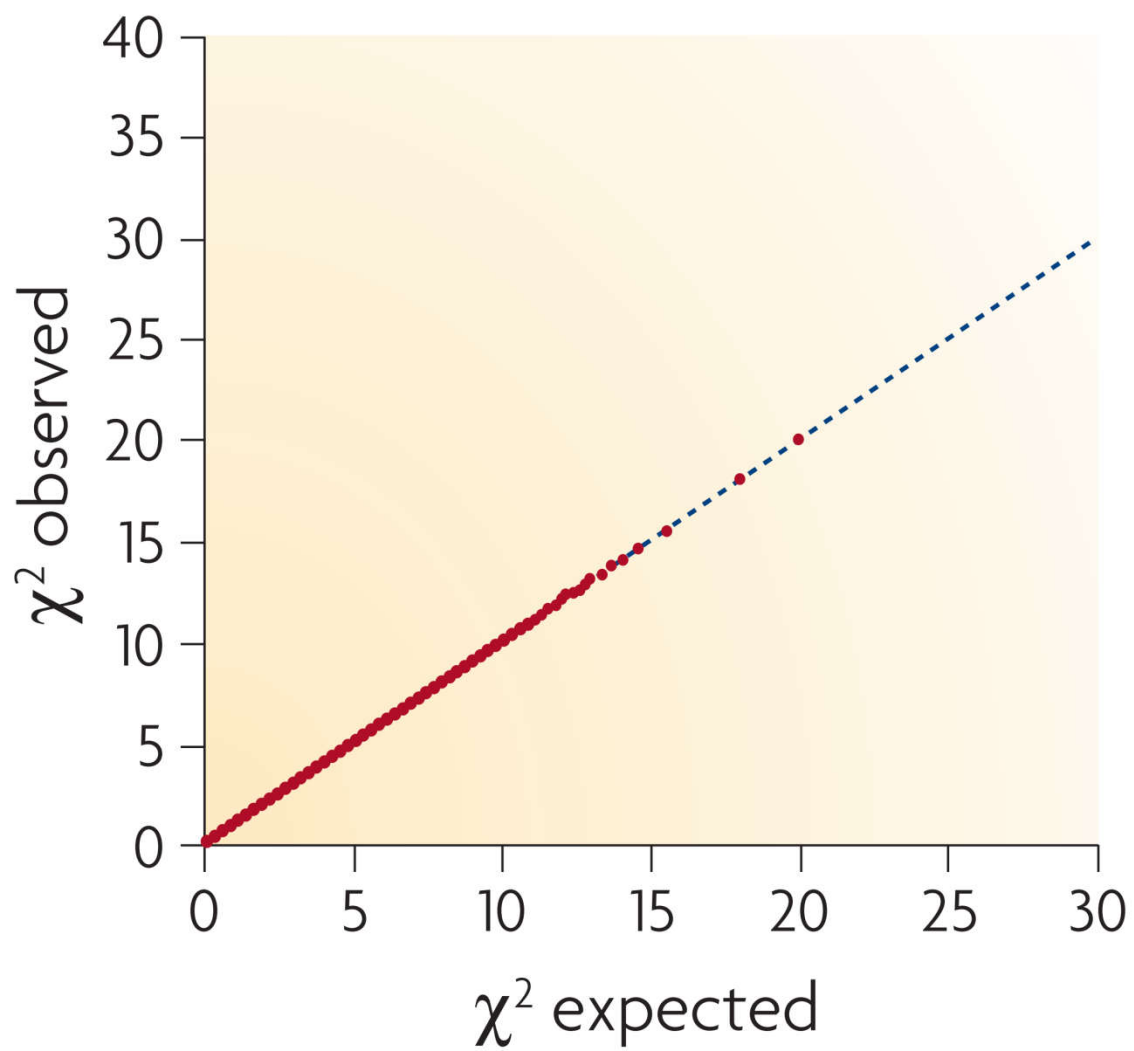
# Efeitos de estrutura populacional



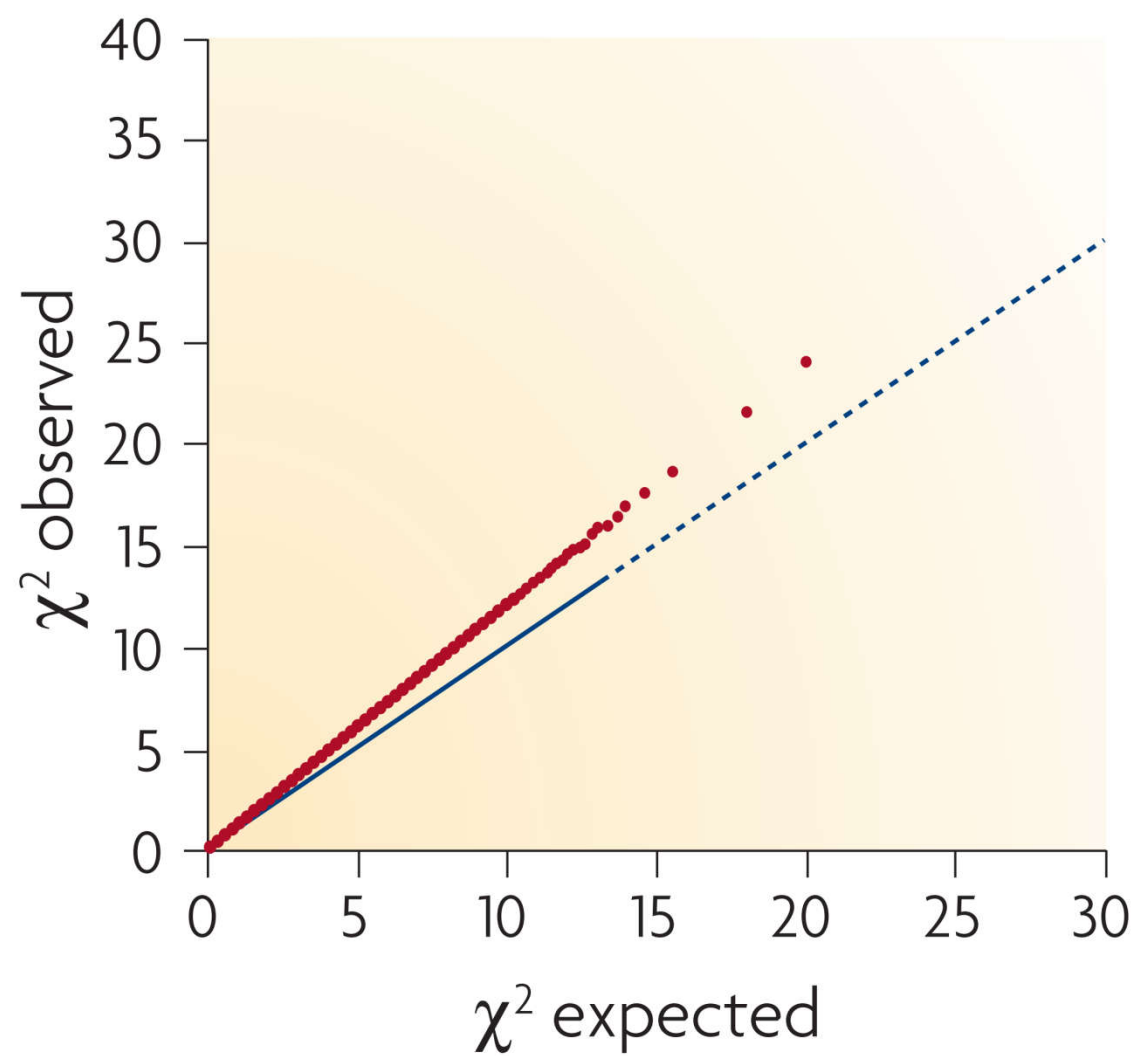




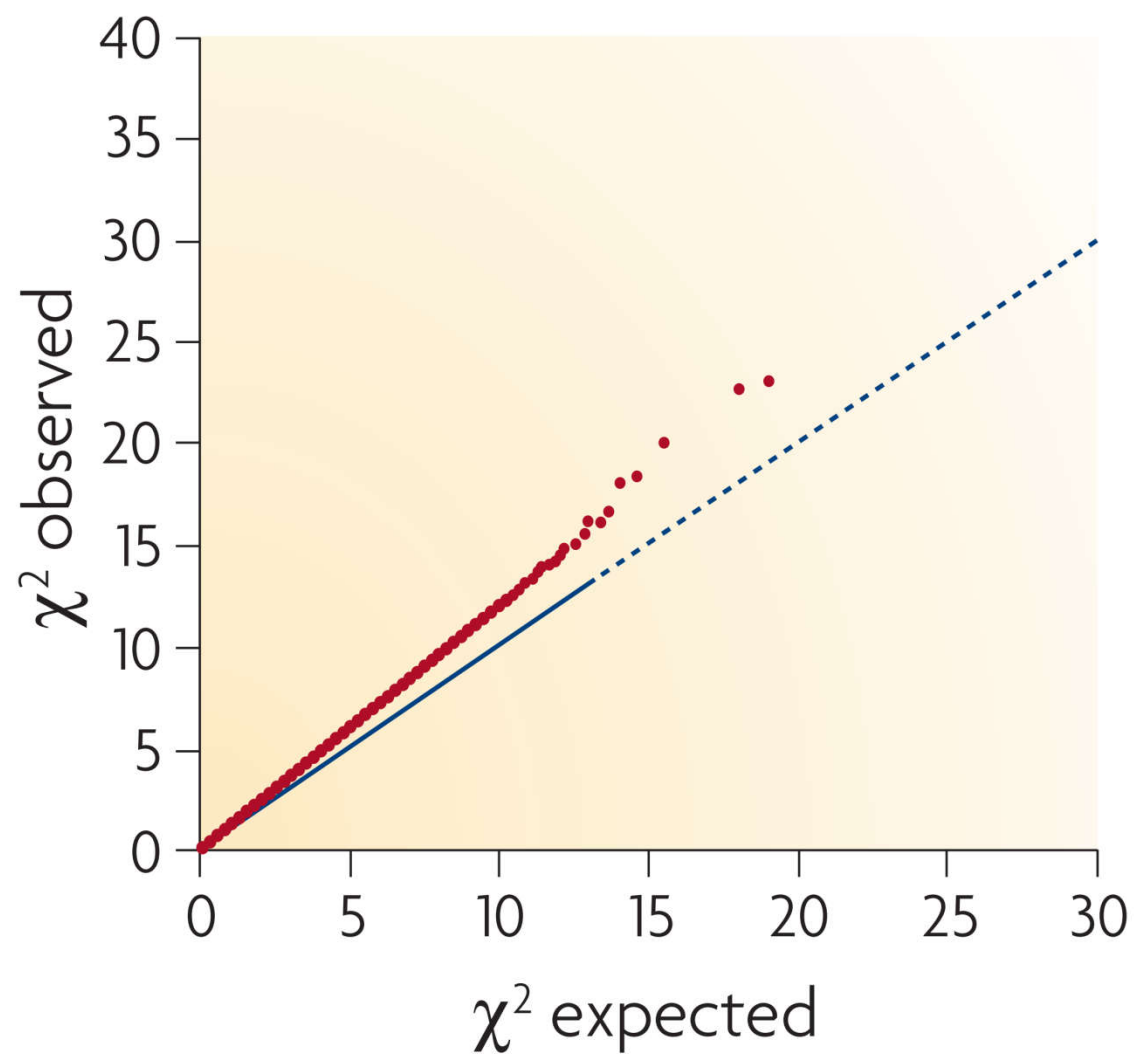
**a**



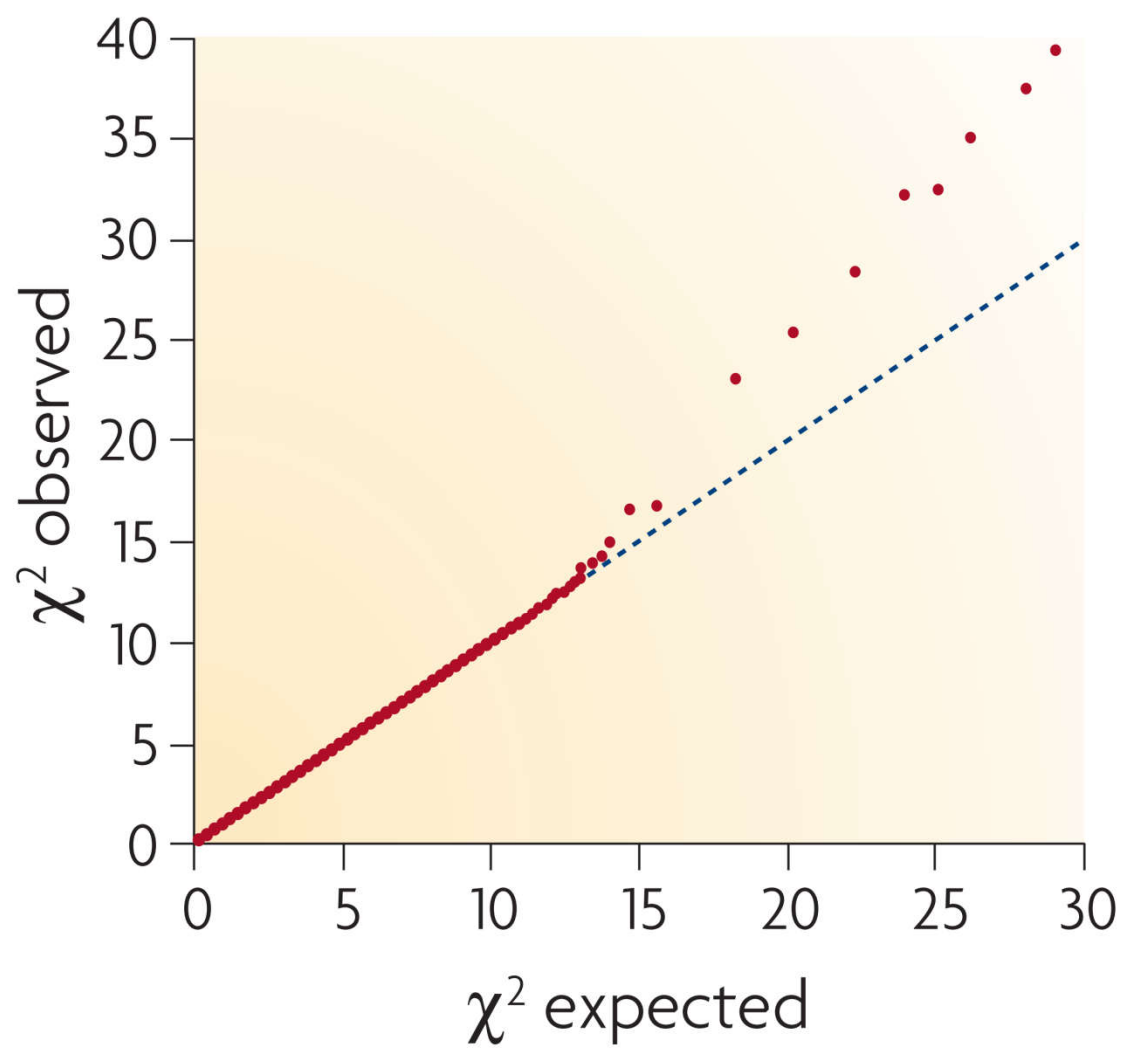
**b**

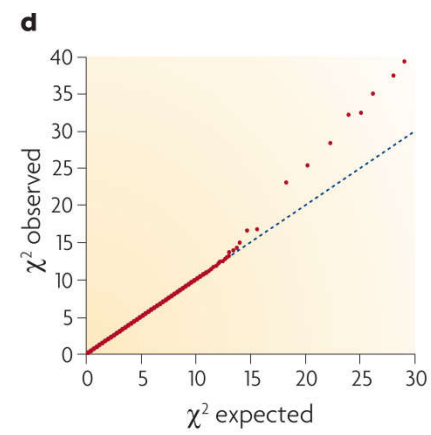
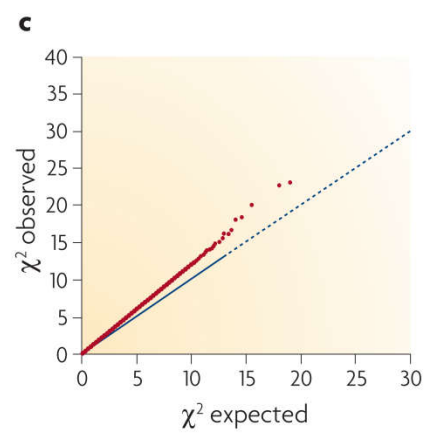
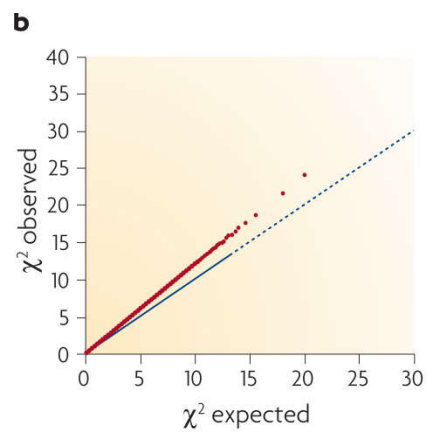
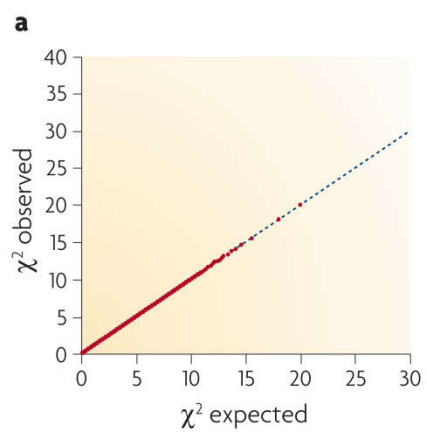


**c**

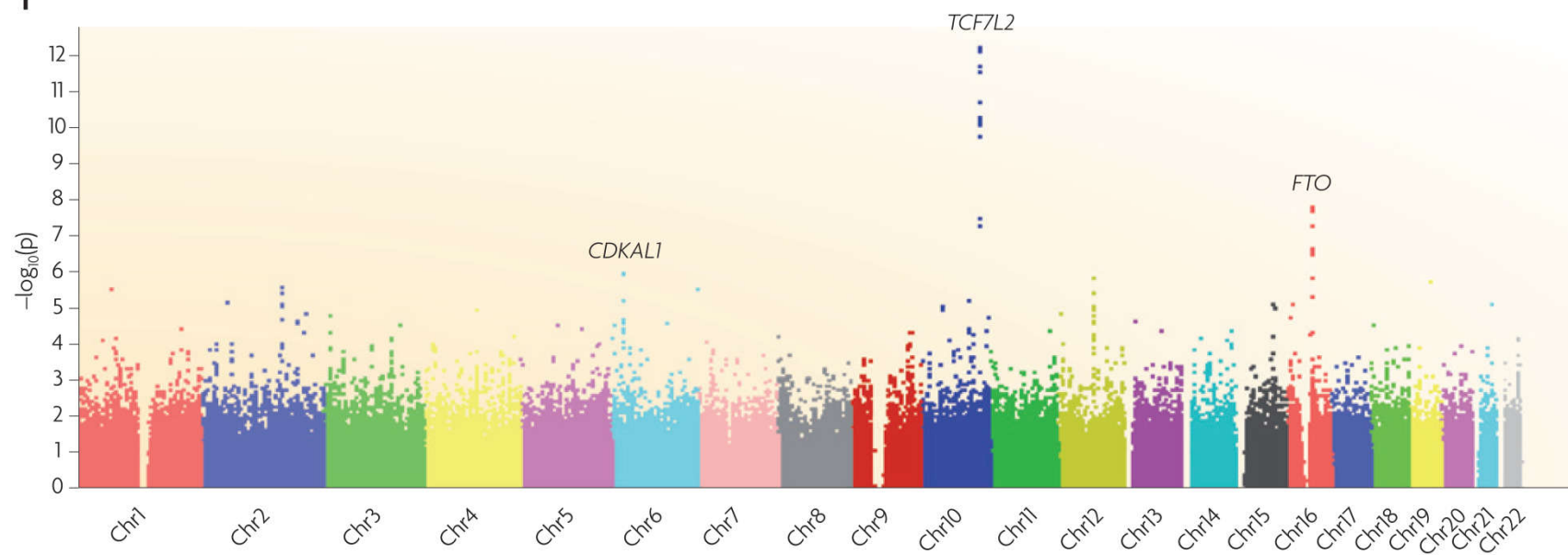


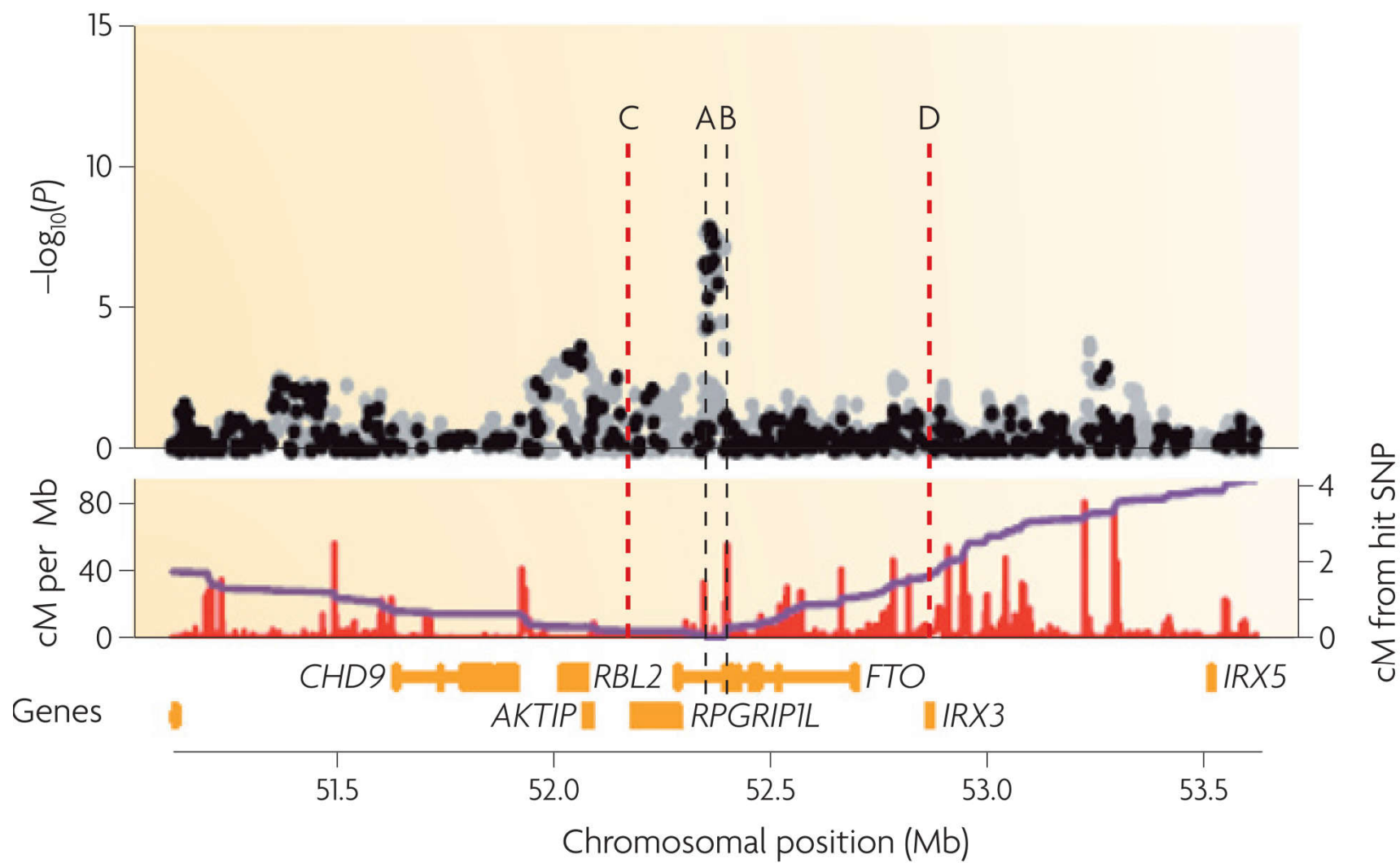
**d**





**i**





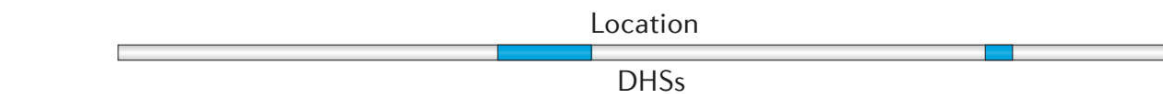
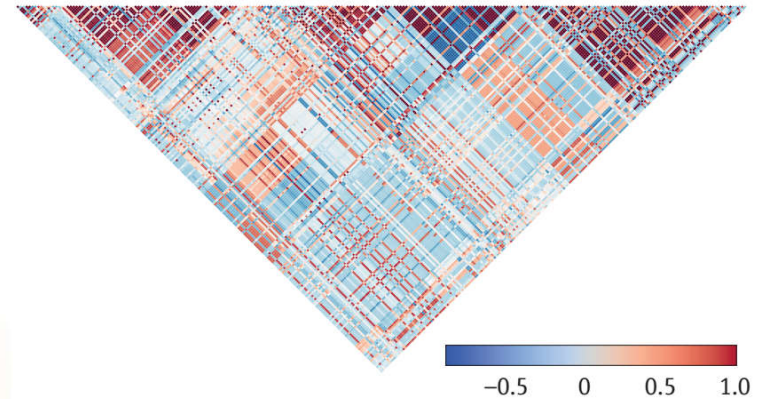
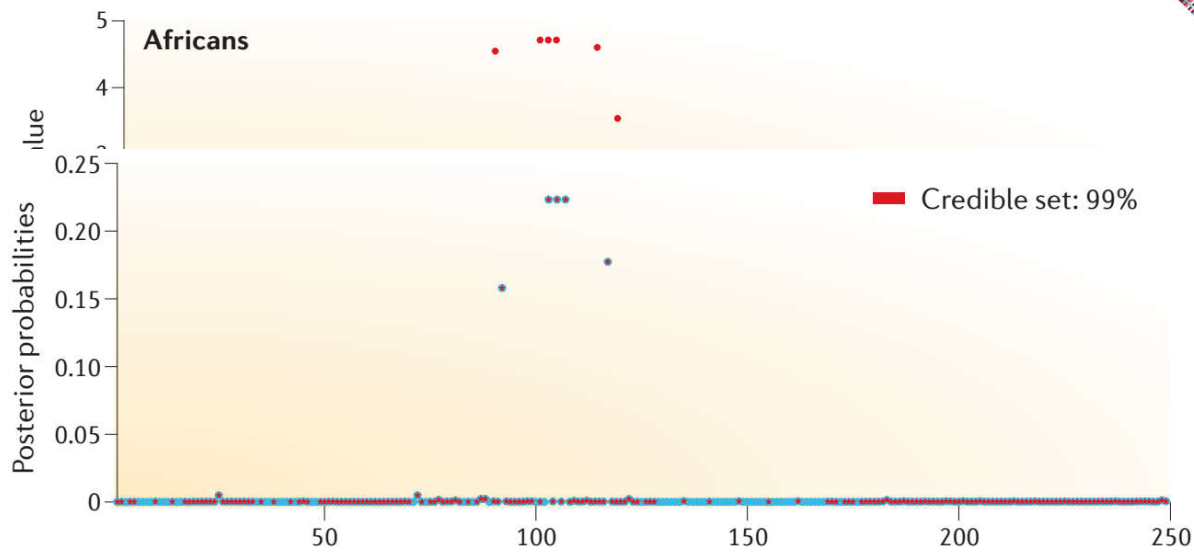
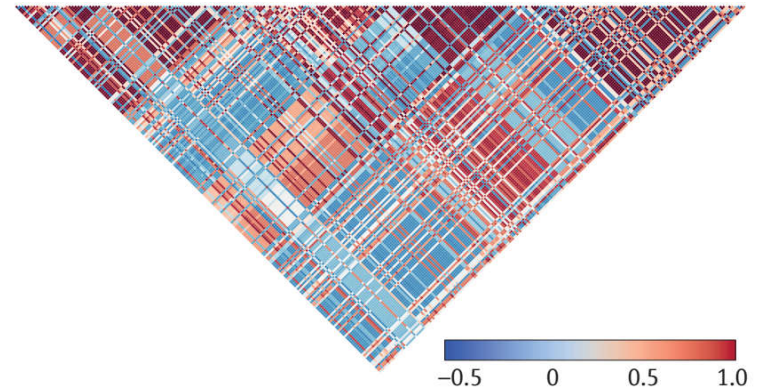
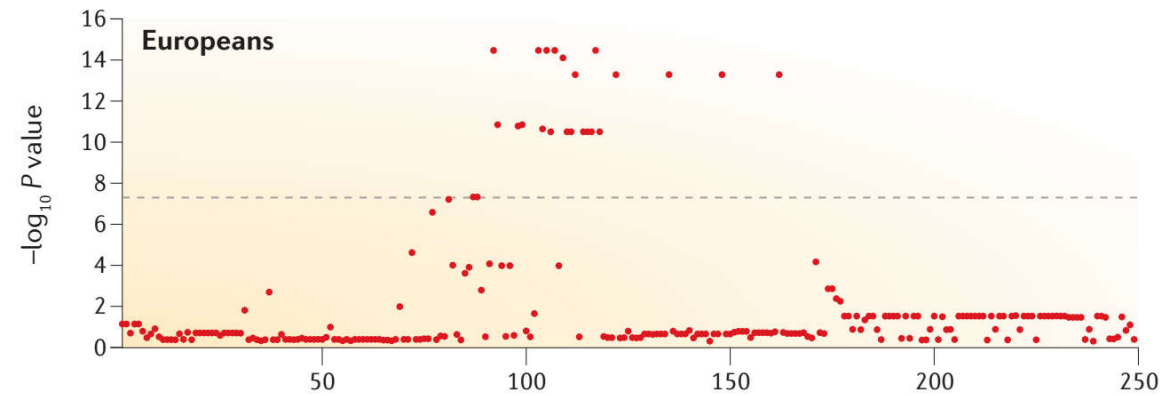
$\gamma$ (allelic odds ratio)	Frequency of susceptibility allele in controls					
	1%	5%	10%	20%	30%	40%
1.1	221 927	46 434	24 626	13 987	10 759	9505
1.2	58 177	12 217	6509	3730	2896	2581
1.3	27 055	5702	3051	1763	1380	1240
1.5	10 604	2249	1213	712	566	516
2.0	3193	687	377	229	188	177
4.0	598	134	78	52	46	47

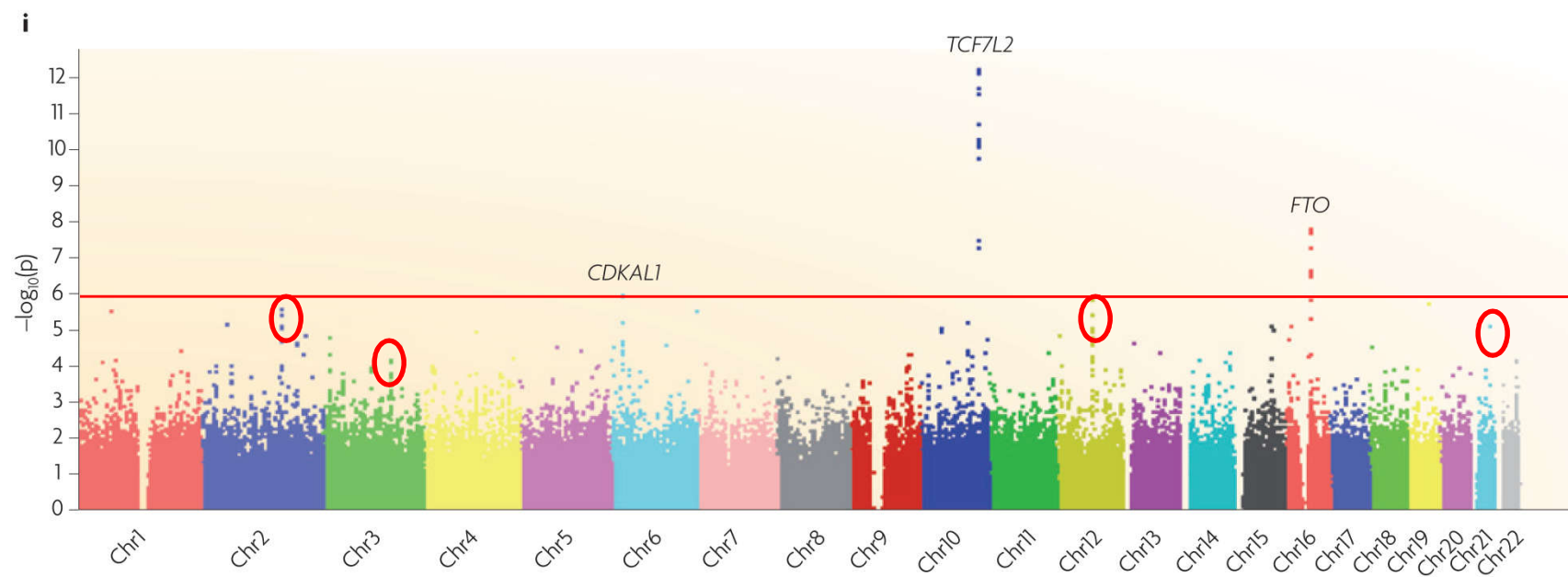
Calculations assume multiplicative effect on disease risk (ie, homozygous susceptibility genotype has penetrance that exceeds that of heterozygote by factor  $\gamma$ , the genotype relative risk, and that of wild-type homozygote by  $\gamma^2$ ). Under such model, each allele has independent effects on disease risk, and allelic odds ratio is also equal to  $\gamma$ . Sample sizes presented are total number of cases needed in case control study where controls are present in equal numbers. These sample size derivations assume best-case scenario in which susceptibility variant itself (or a perfect proxy) has been typed.

**Table 3: Approximate sample sizes necessary to detect significant association (power=90%, two-sided  $\alpha=0.001$ ) by effect size and allele frequency for predisposing allele**

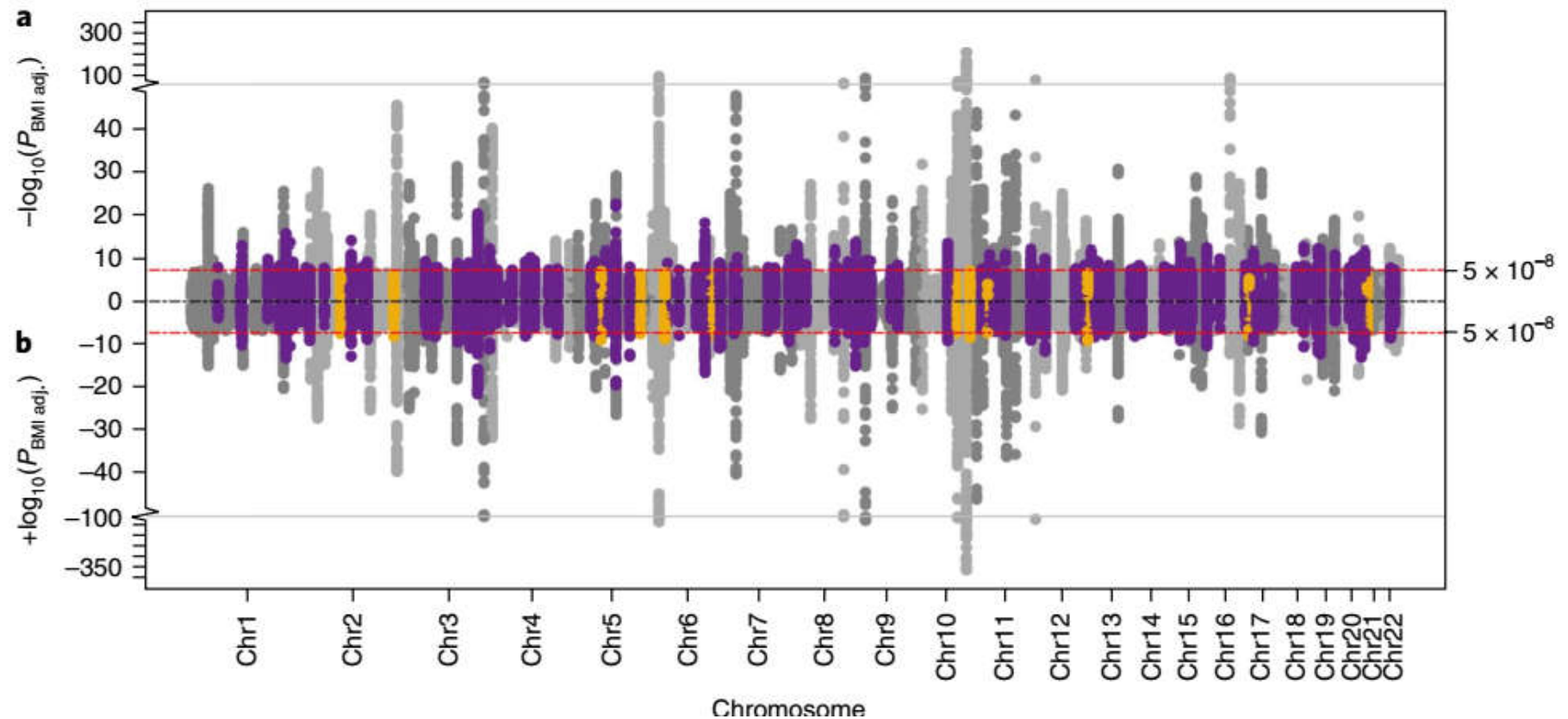


# Entre populações



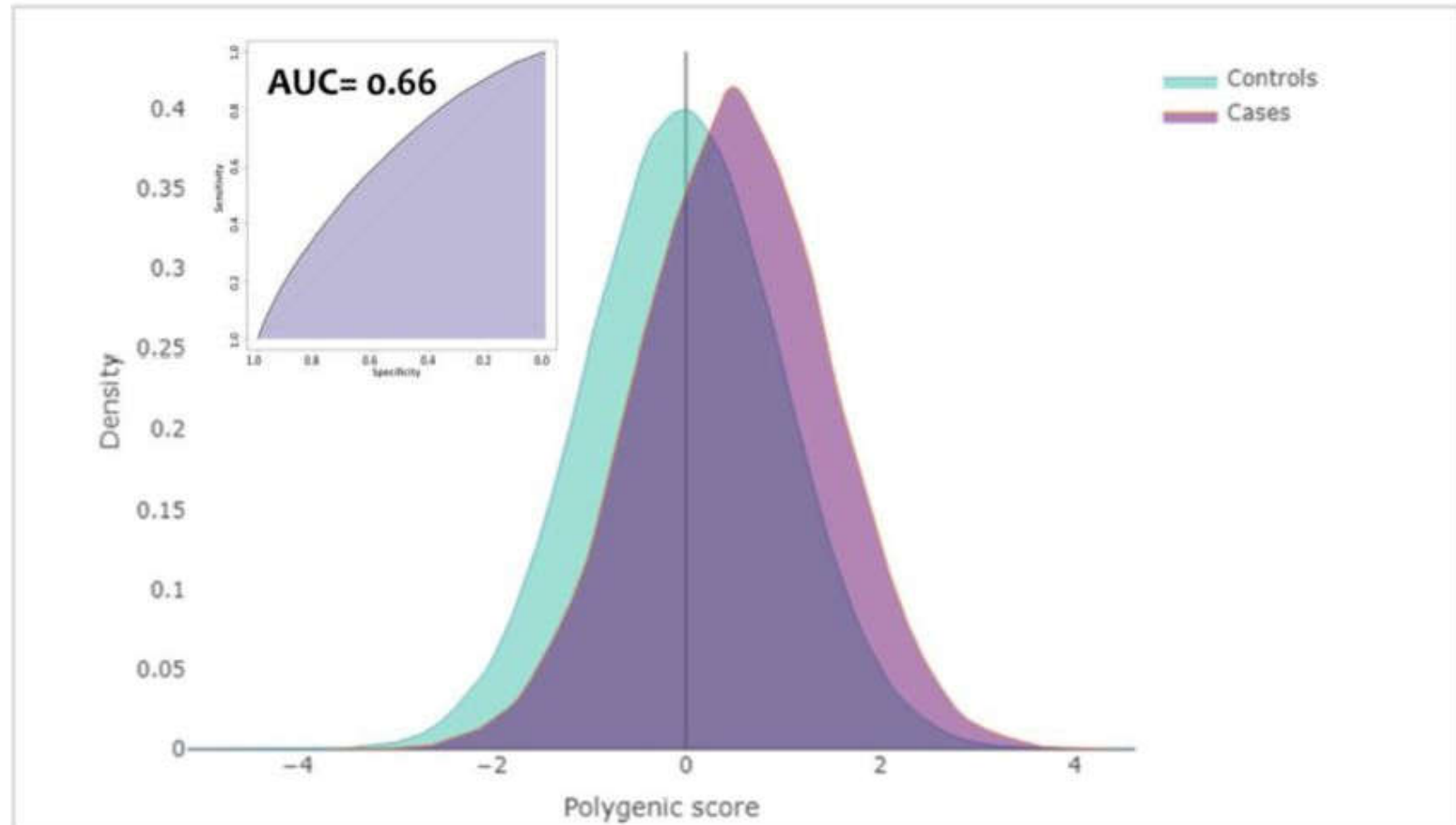


# Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps



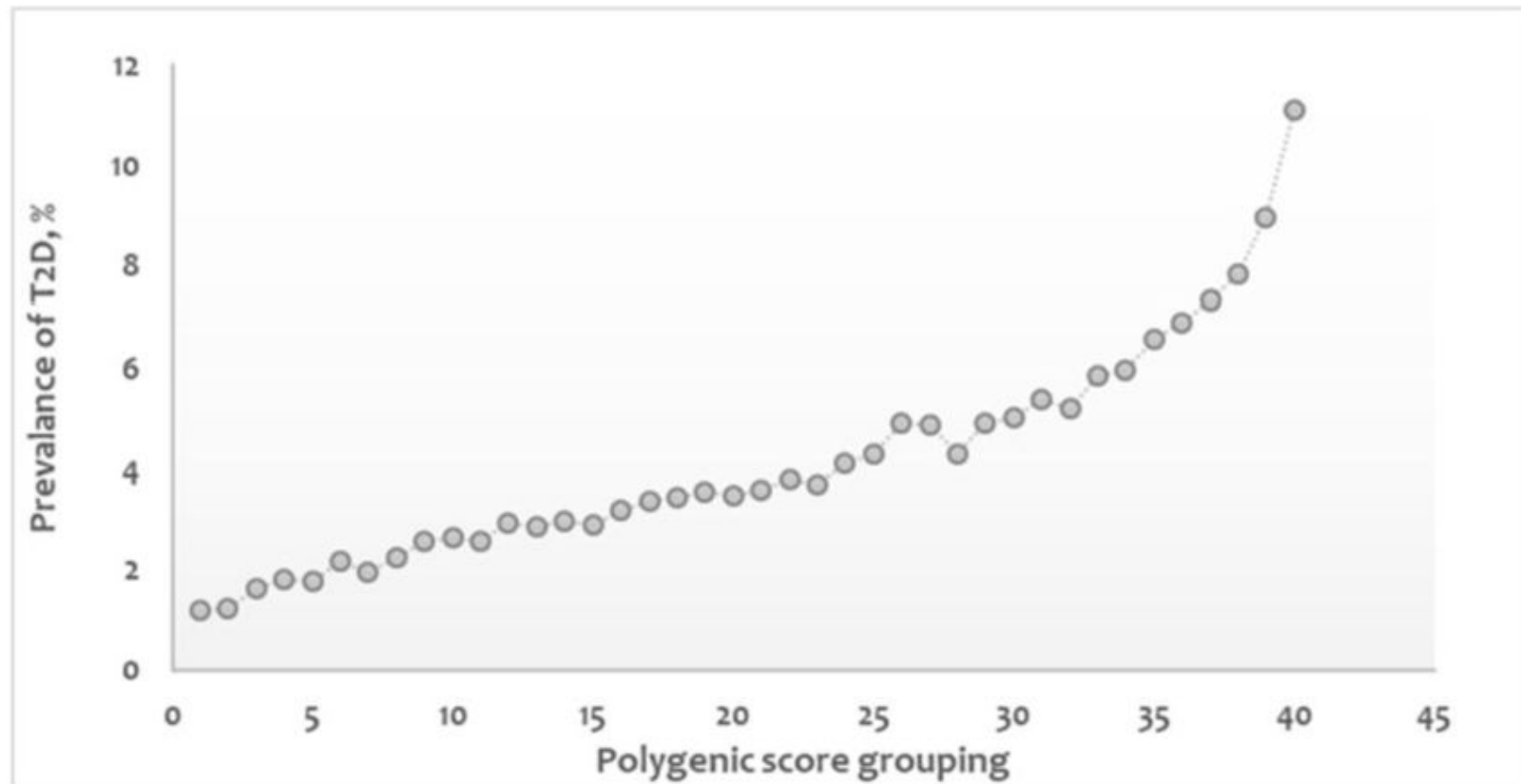
# Polygenic risk scores

a)



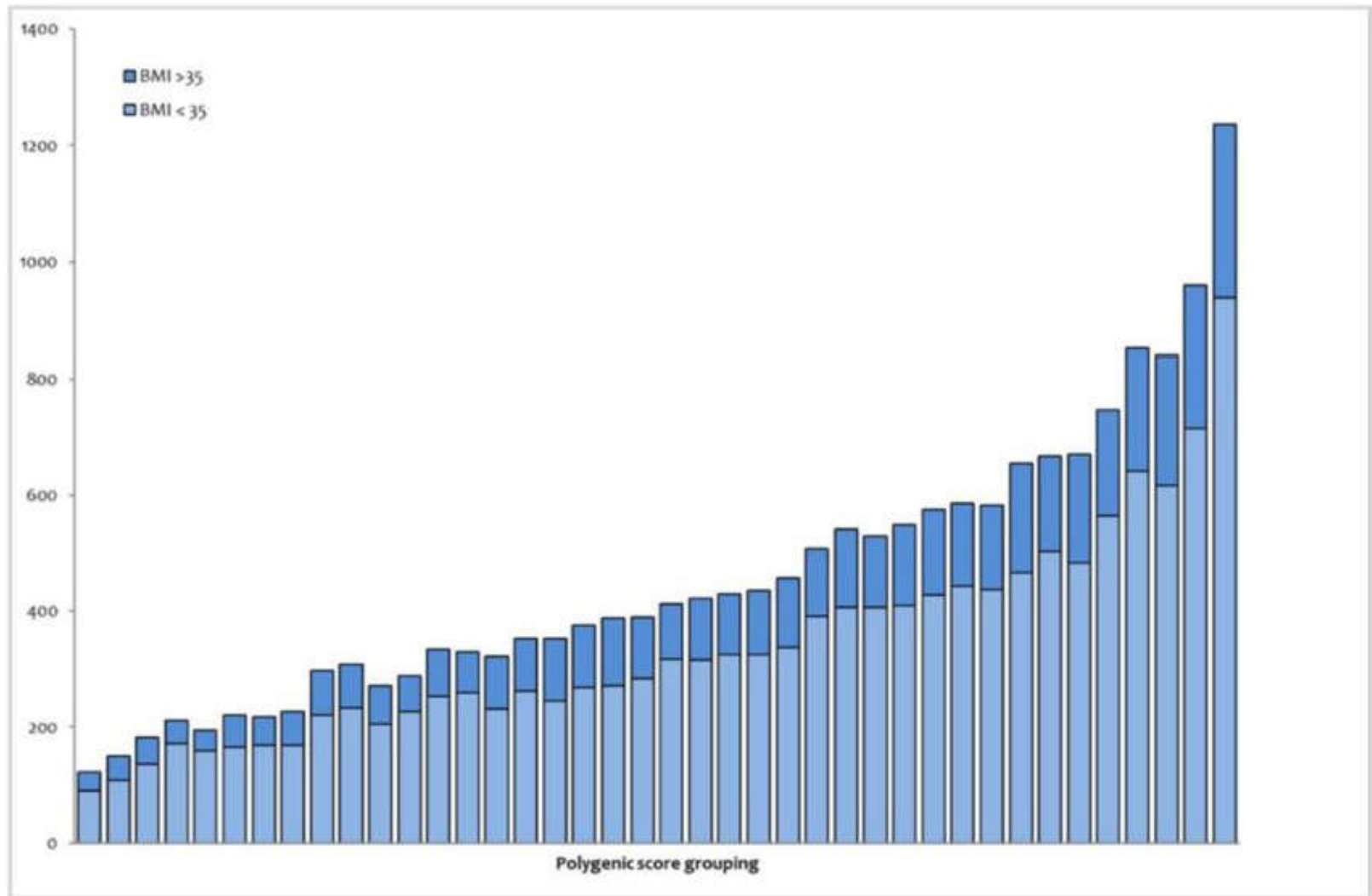
# Polygenic risk scores

b)



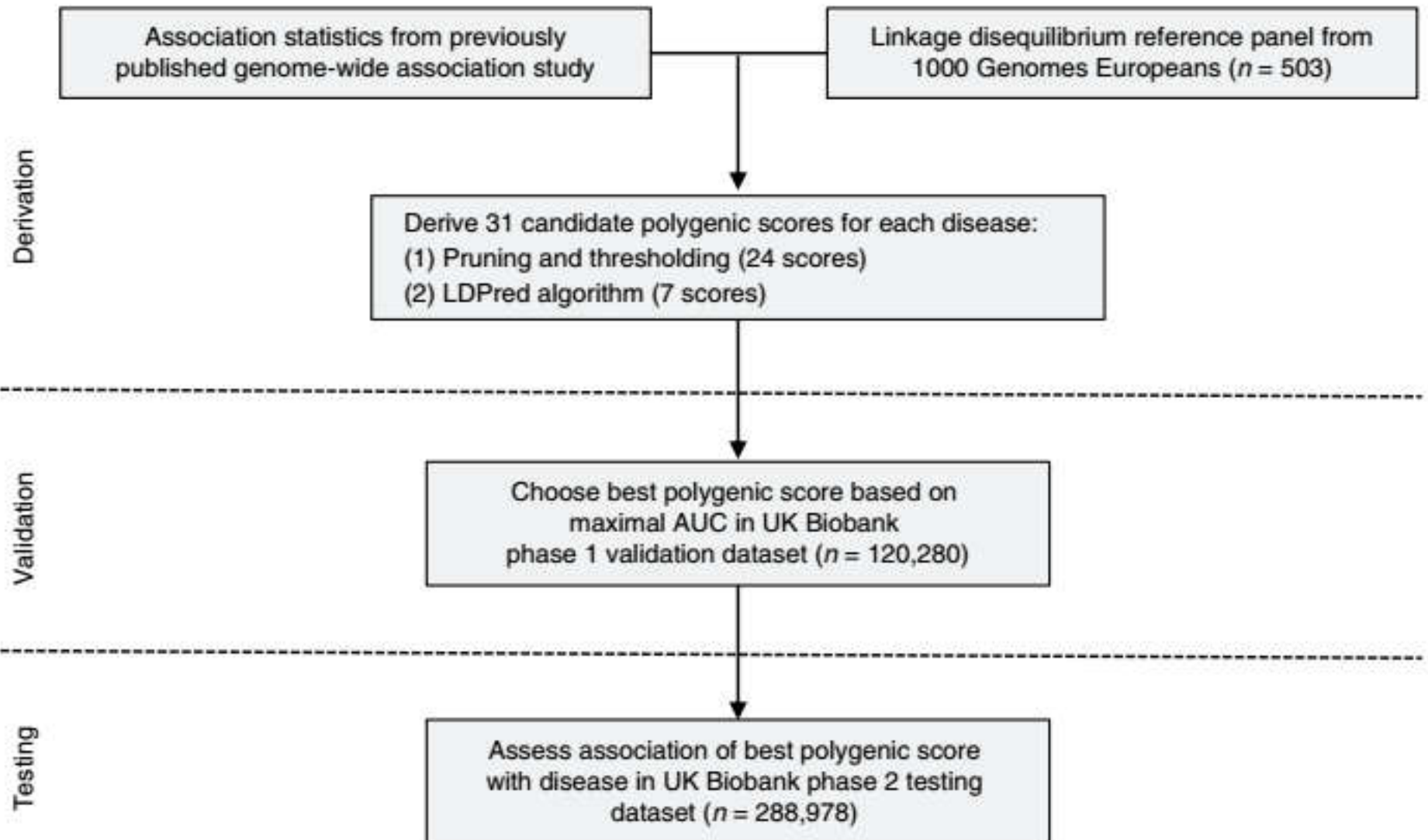
# Polygenic risk scores

c)



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

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

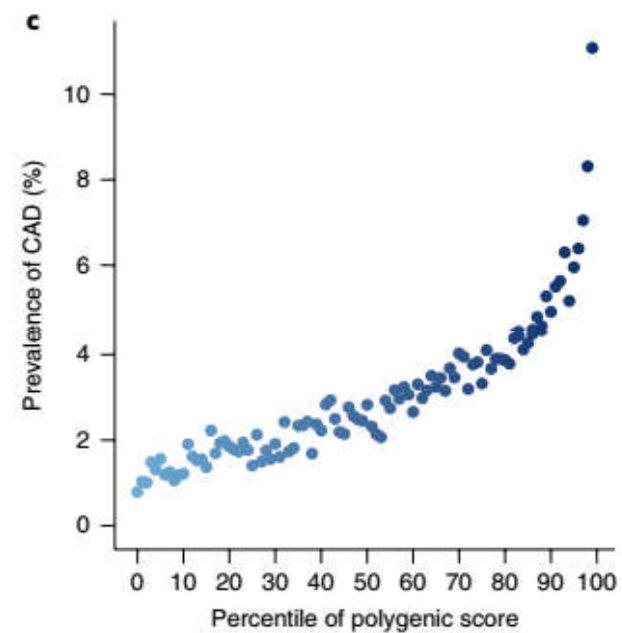
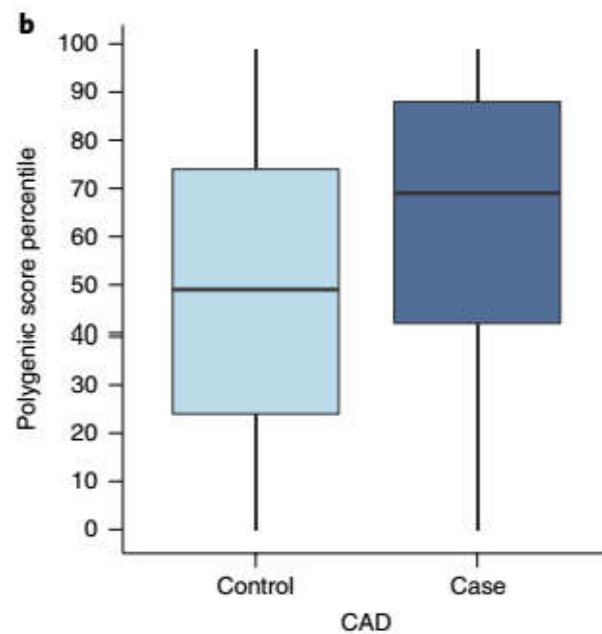
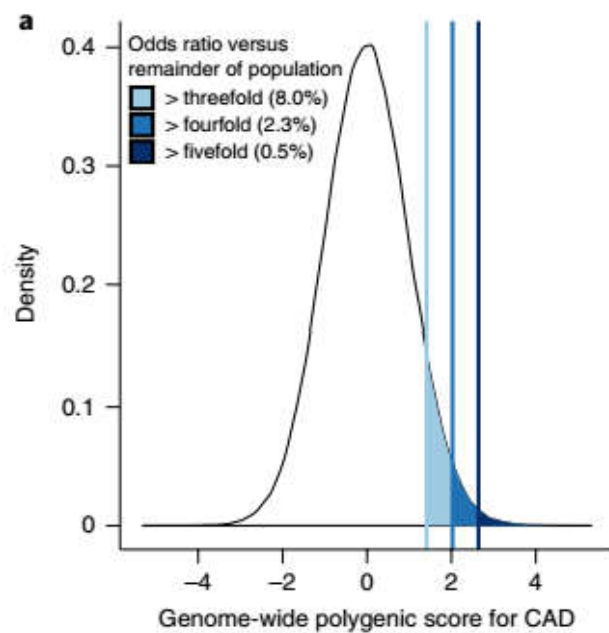


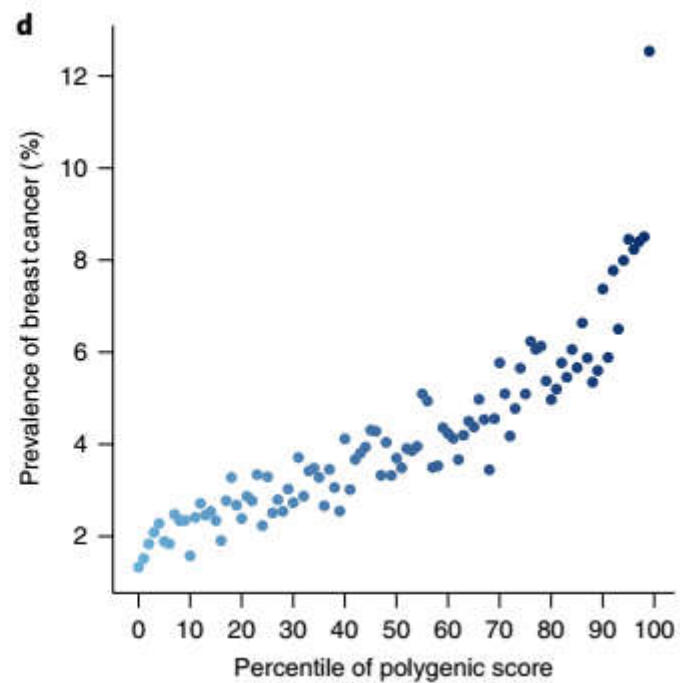
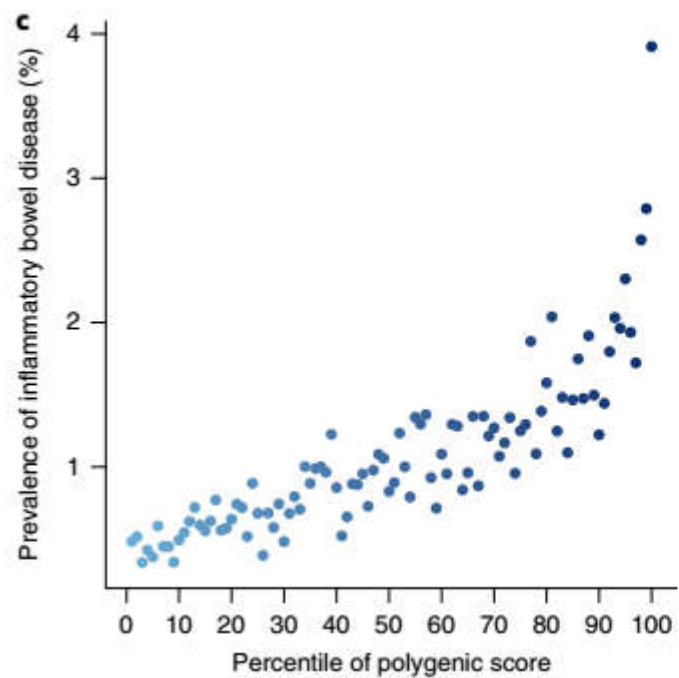
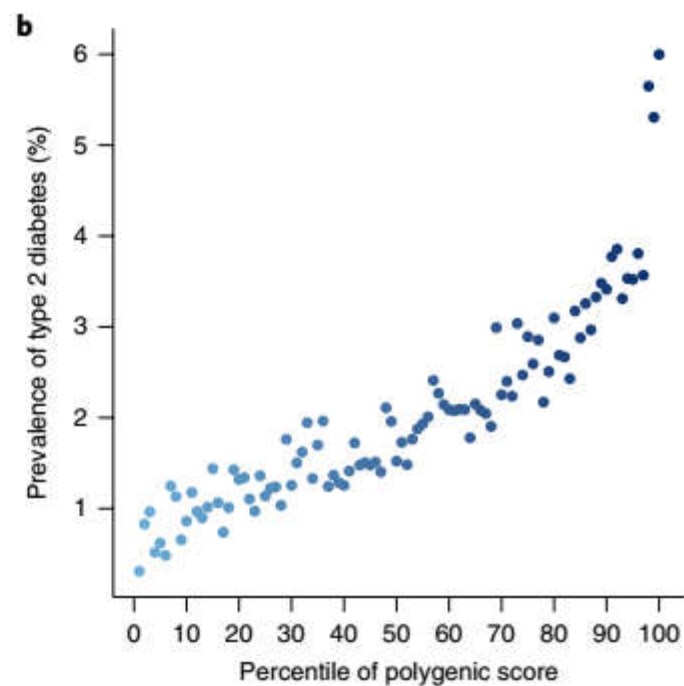
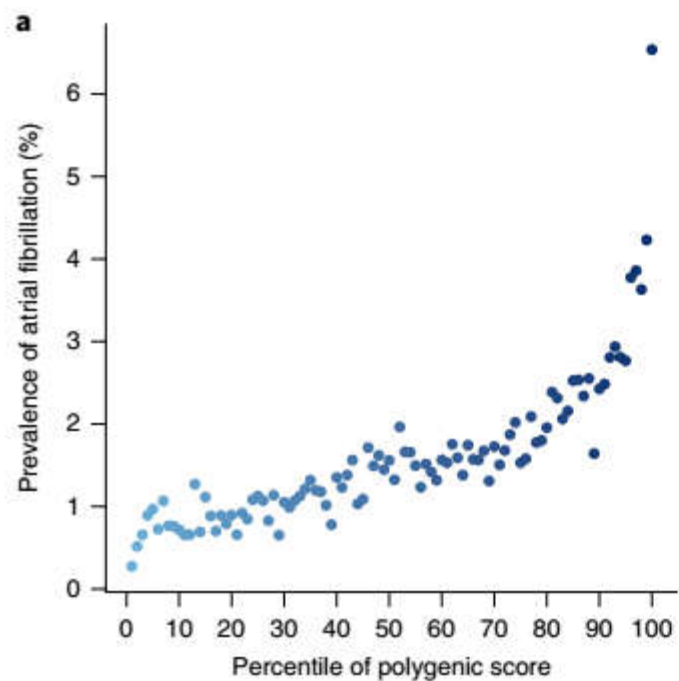


**Table 1 | GPS derivation and testing for five common, complex diseases**

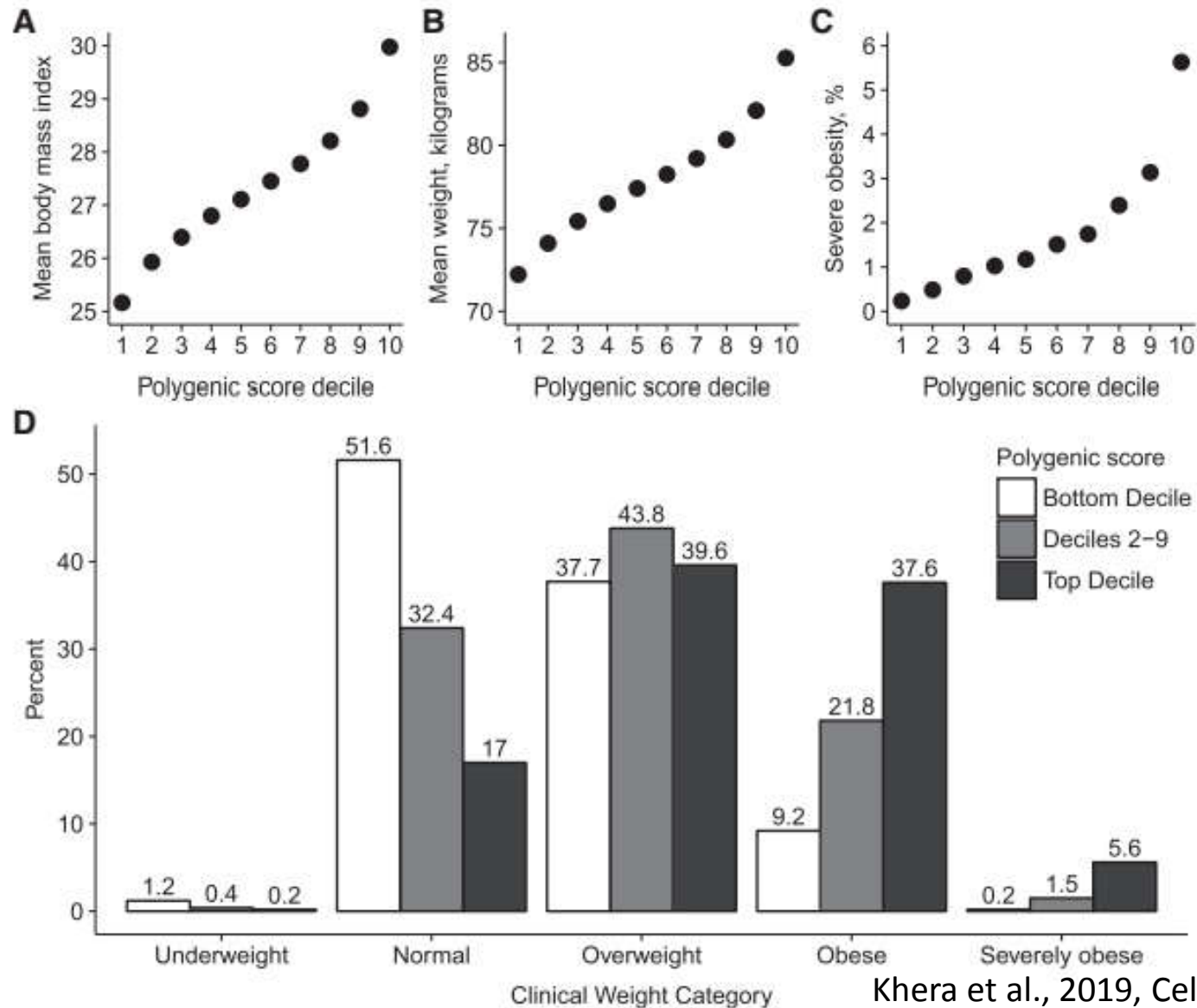
Disease	Discovery GWAS (n)	Prevalence in validation dataset	Prevalence in testing dataset	Polymorphisms in GPS	Tuning parameter	AUC (95% CI) in validation dataset	AUC (95% CI) in testing dataset
CAD	60,801 cases; 123,504 controls <sup>16</sup>	3,963/120,280 (3.4%)	8,676/288,978 (3.0%)	6,630,150	LDPred ( $\rho = 0.001$ )	0.81 (0.80–0.81)	0.81 (0.81–0.81)
Atrial fibrillation	17,931 cases; 115,142 controls <sup>30</sup>	2,024/120,280 (1.7%)	4,576/288,978 (1.6%)	6,730,541	LDPred ( $\rho = 0.003$ )	0.77 (0.76–0.78)	0.77 (0.76–0.77)
Type 2 diabetes	26,676 cases; 132,532 controls <sup>31</sup>	2,785/120,280 (2.4%)	5,853/288,978 (2.0%)	6,917,436	LDPred ( $\rho = 0.01$ )	0.72 (0.72–0.73)	0.73 (0.72–0.73)
Inflammatory bowel disease	12,882 cases; 21,770 controls <sup>32</sup>	1,360/120,280 (1.1%)	3,102/288,978 (1.1%)	6,907,112	LDPred ( $\rho = 0.1$ )	0.63 (0.62–0.65)	0.63 (0.62–0.64)
Breast cancer	122,977 cases; 105,974 controls <sup>33</sup>	2,576/63,347 (4.1%)	6,586/157,895 (4.2%)	5,218	Pruning and thresholding ( $r^2 < 0.2$ ; $P < 5 \times 10^{-4}$ )	0.68 (0.67–0.69)	0.69 (0.68–0.69)

AUC was determined using a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. The breast cancer analysis was restricted to female participants. For the LDPred algorithm, the tuning parameter  $\rho$  reflects the proportion of polymorphisms assumed to be causal for the disease. For the pruning and thresholding strategy,  $r^2$  reflects the degree of independence from other variants in the linkage disequilibrium reference panel, and  $P$  reflects the  $P$  value noted for a given variant in the discovery GWAS. CI, confidence interval.





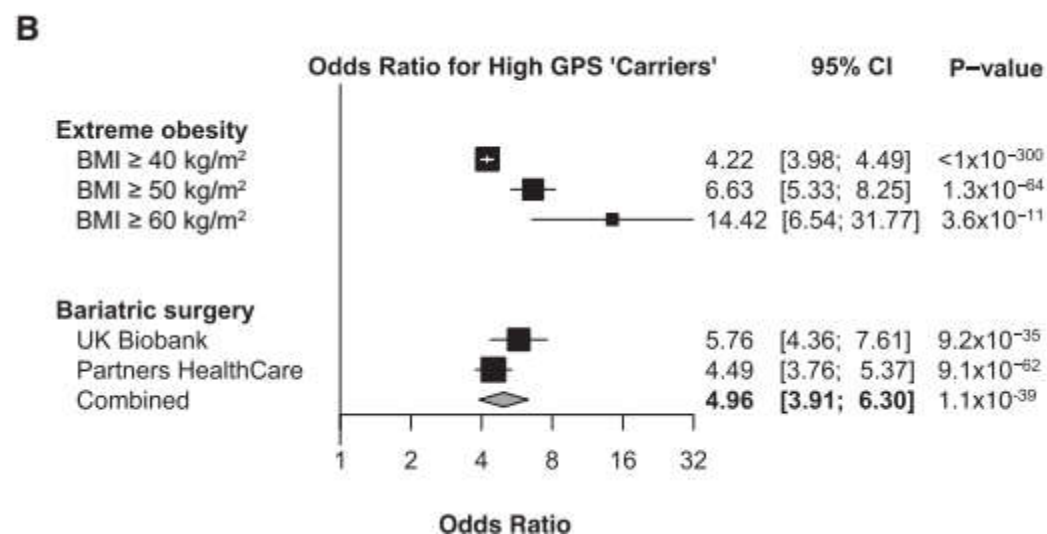
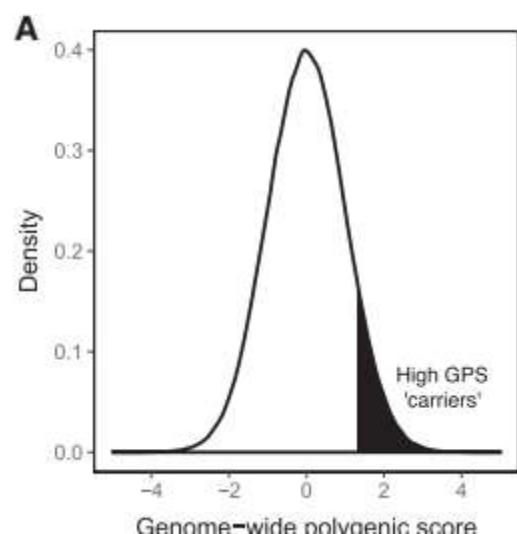
# BMI – “age-penetrance”

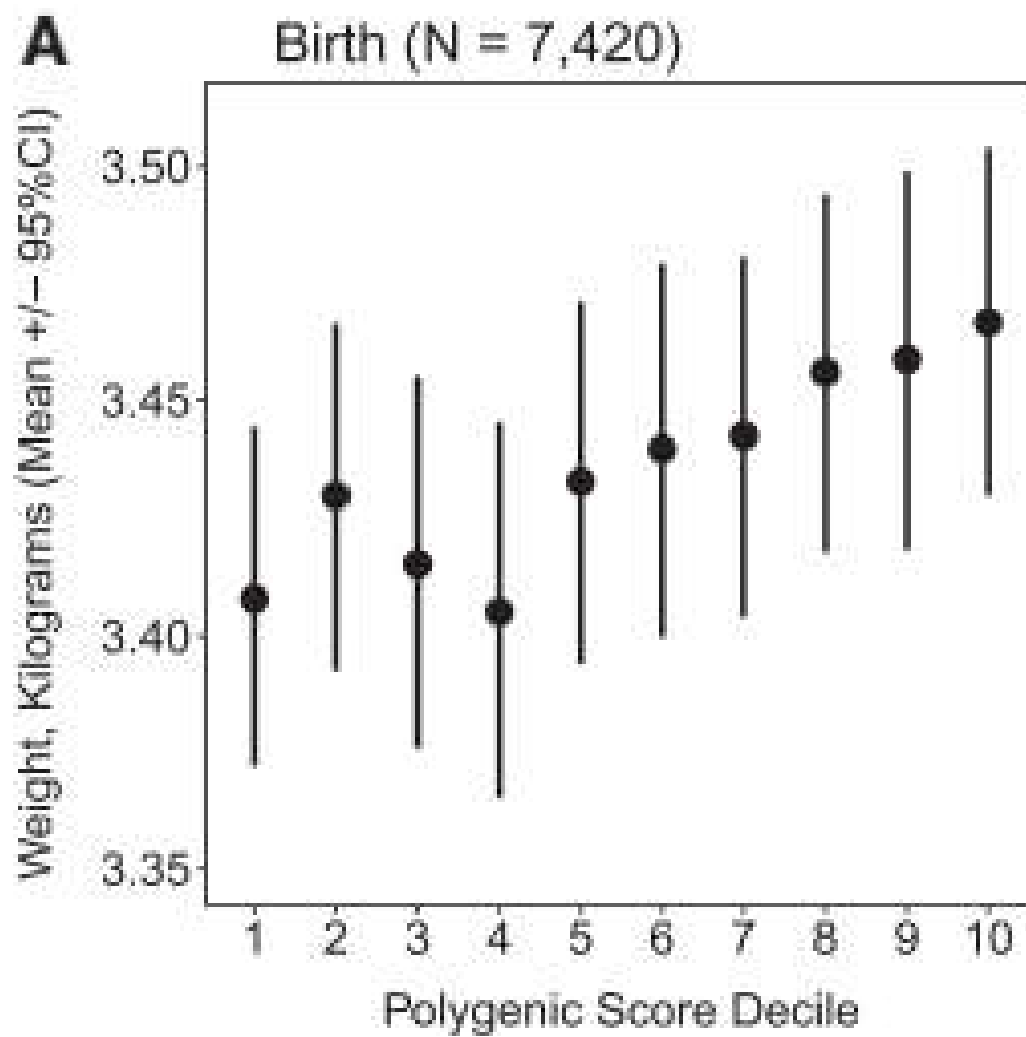


**Table 1. Genome-wide Polygenic Score for Obesity, Assessed in Four Independent Testing Datasets**

	UK Biobank	Partners HealthCare	Framingham Offspring/CARDIA	Avon Longitudinal Study of Parents and Children
n participants	288,016	6,536	3,722	7,861
Study design	cross-sectional	case-control	longitudinal	longitudinal
Age range	40–69 years	≥ 18 years	18–40 years	birth
Female sex	55%	61%	48%	49%
Outcomes	weight, severe obesity, bariatric surgery, cardiometabolic diseases, mortality	bariatric surgery	incident severe obesity (27 years median follow-up)	weight at birth and subsequent visits (0–18 years)

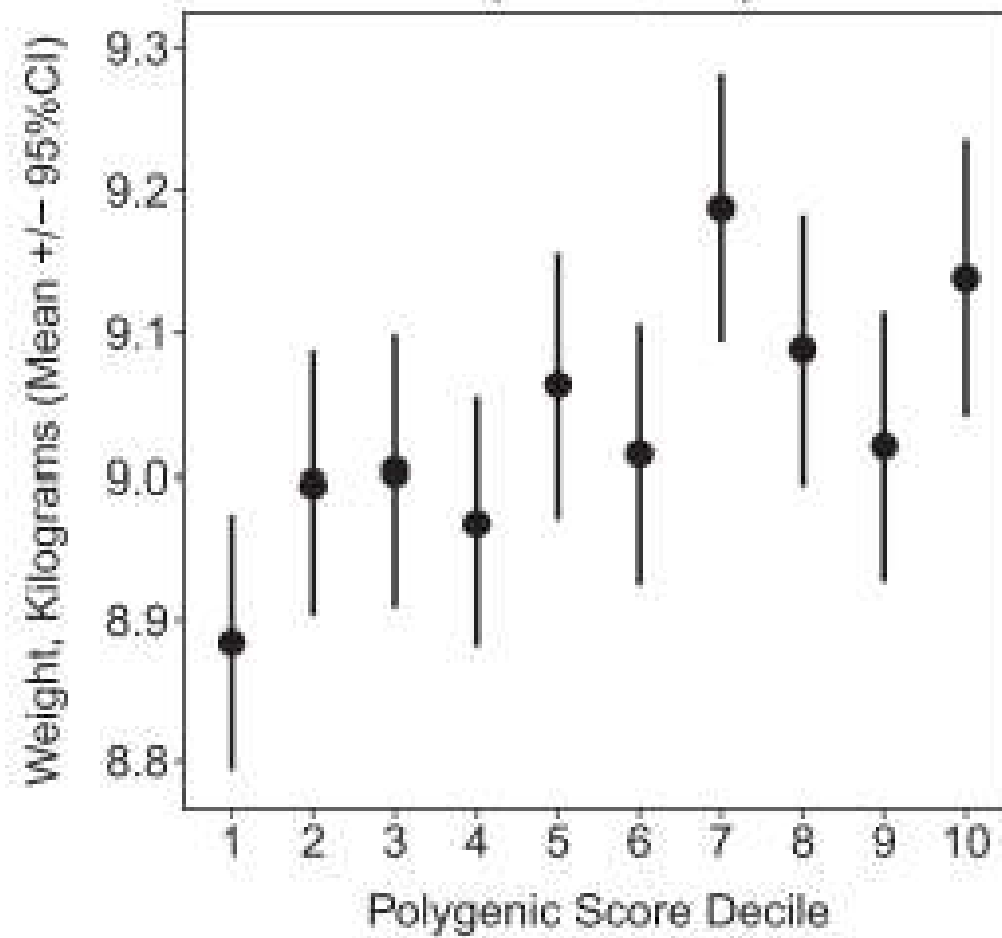
CARDIA, Coronary Artery Risk Development in Young Adults.





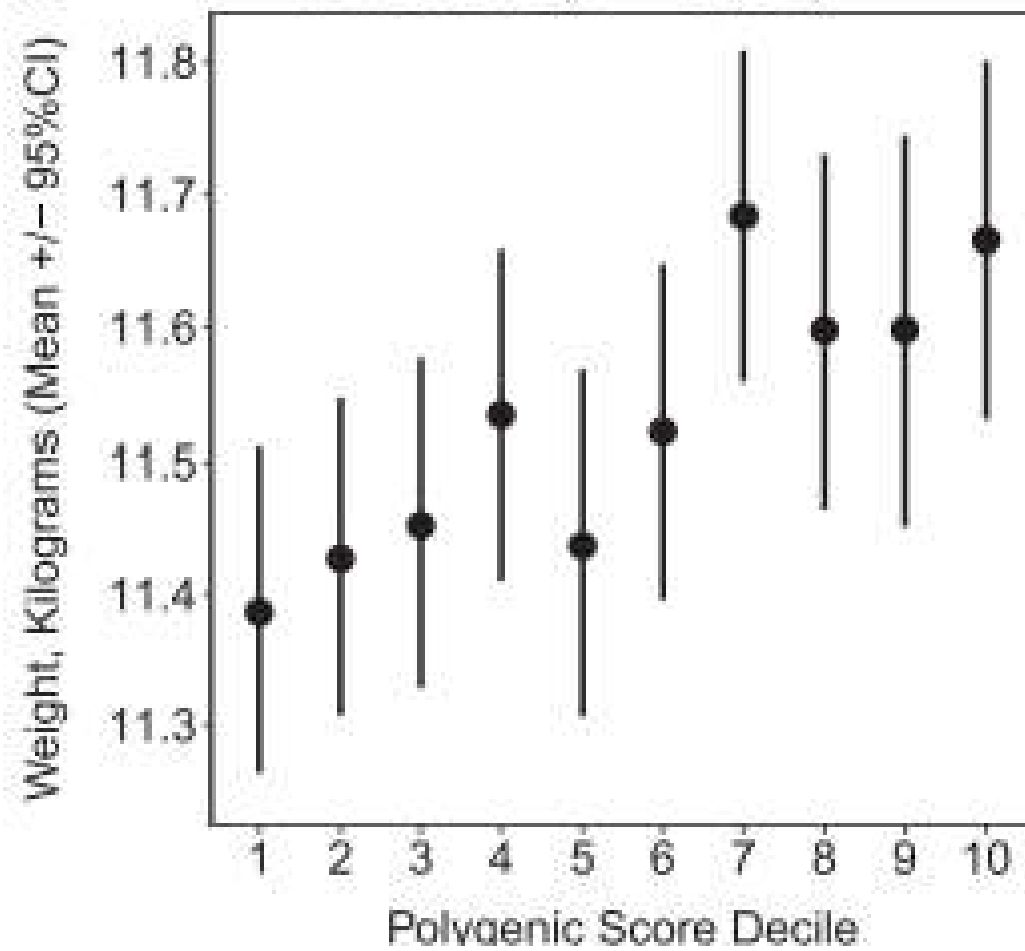
**B**

8 Months (N = 5,011)



**C**

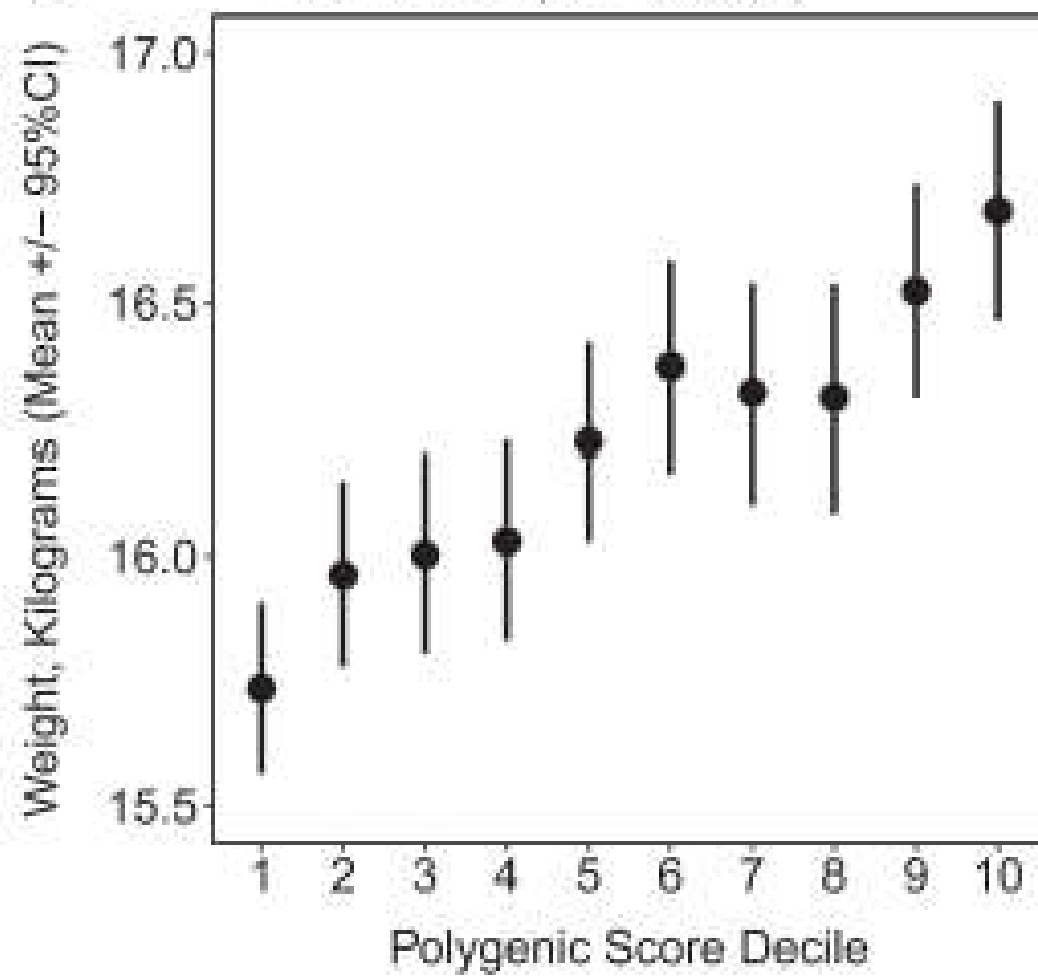
18 Months (N = 3,954)





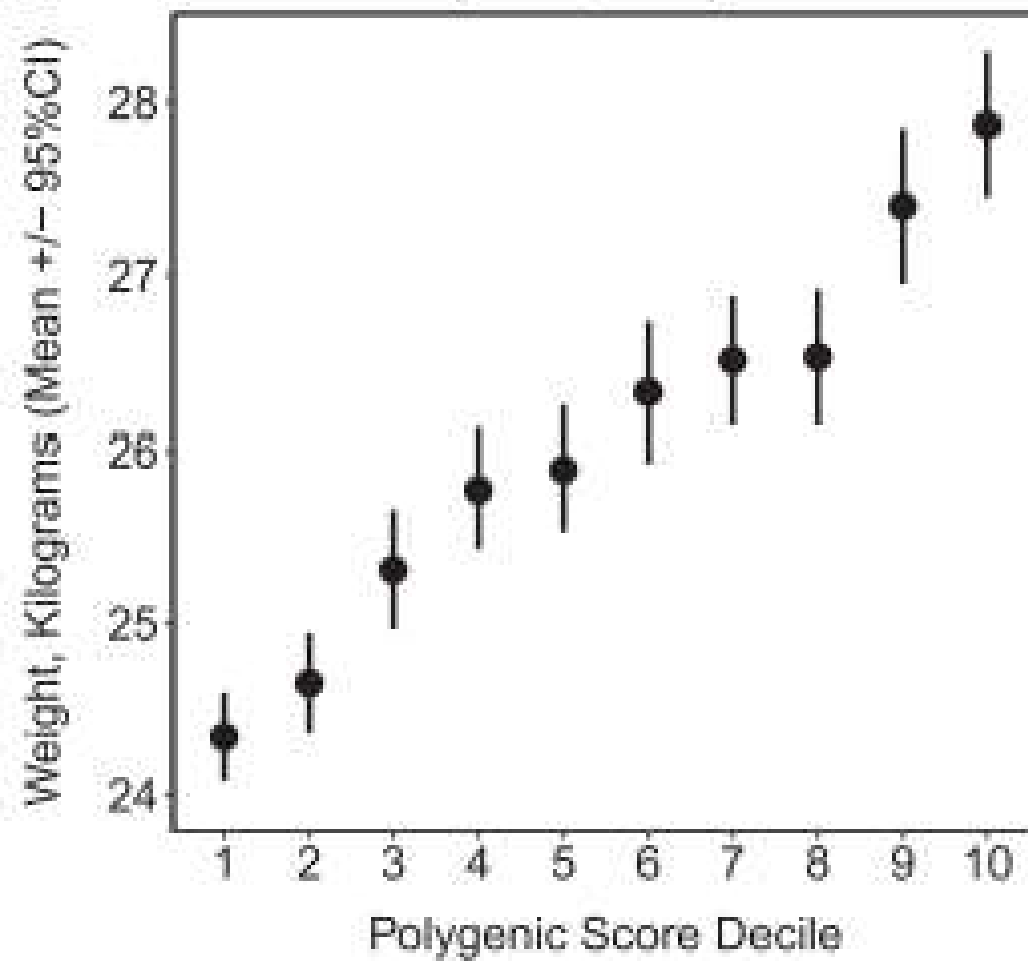
**D**

3.5 Years (N = 3,322)



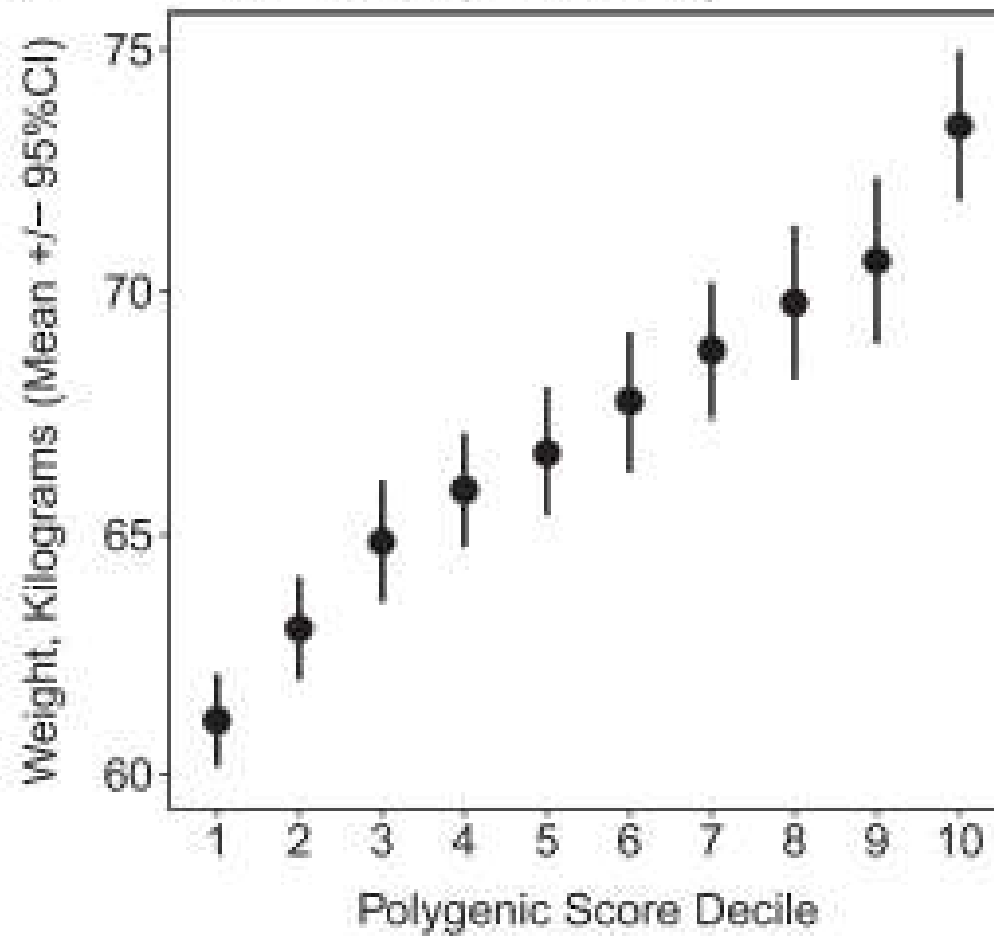
**F**

8 Years (N = 6,049)

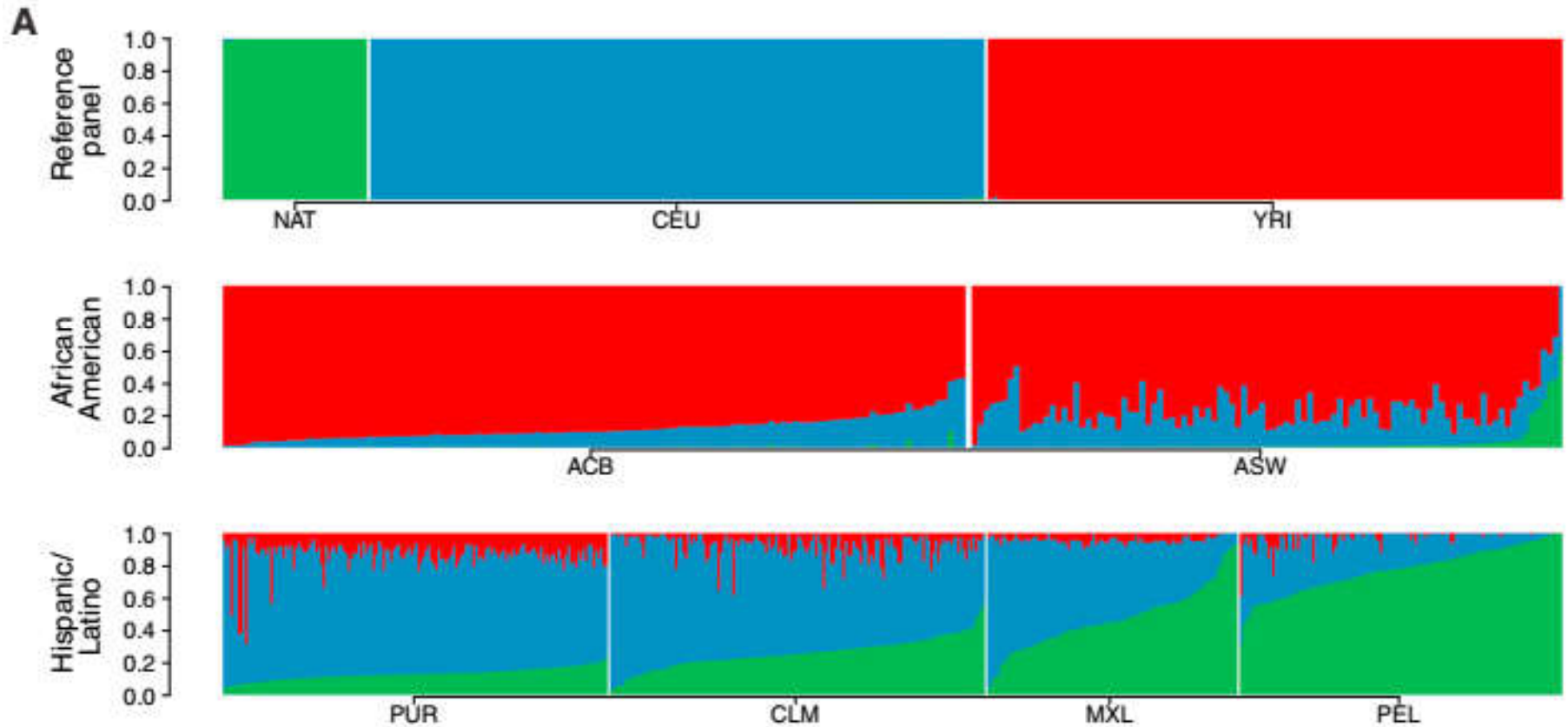


**F**

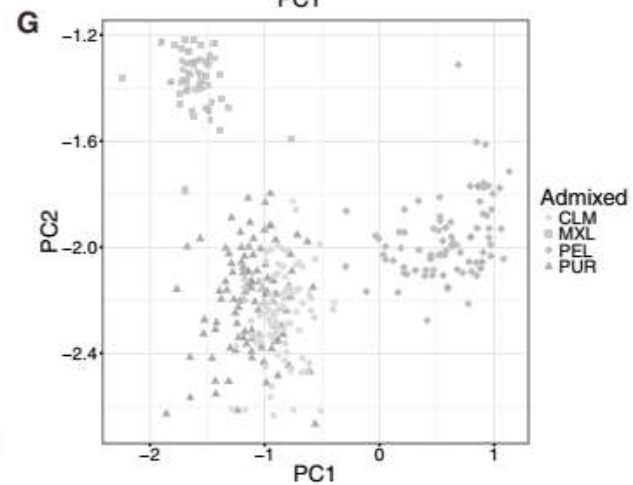
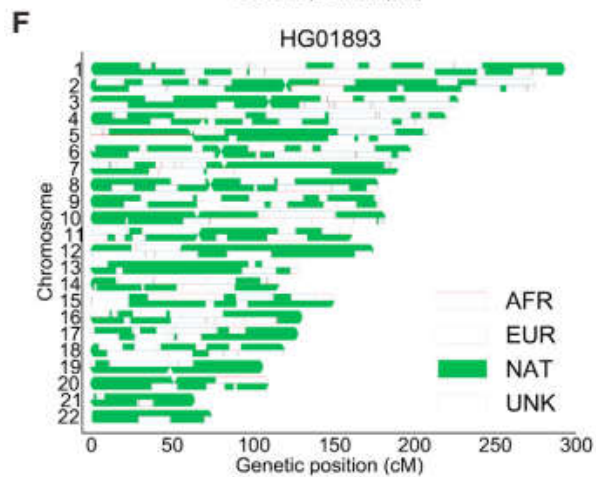
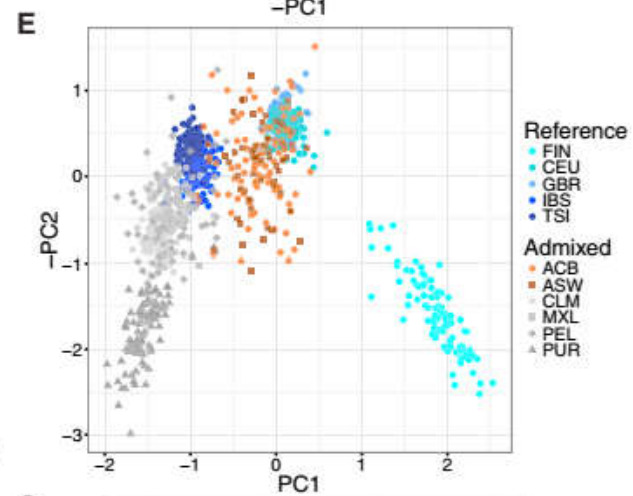
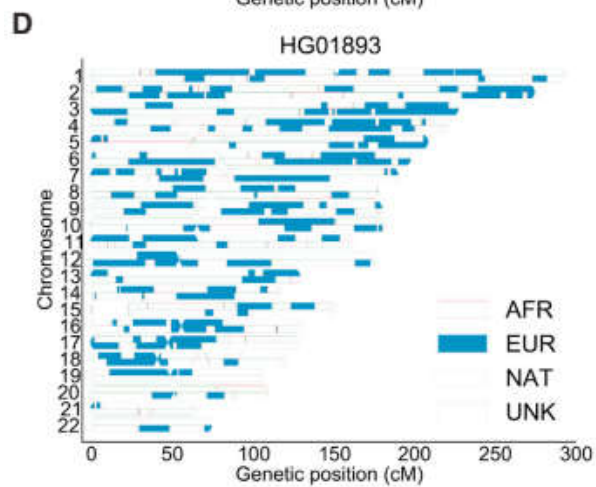
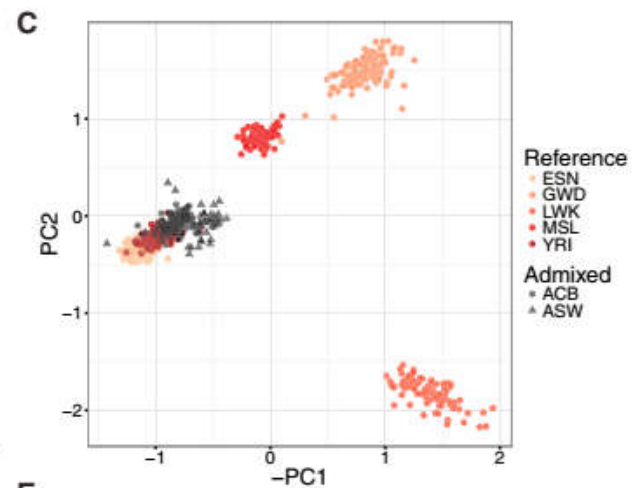
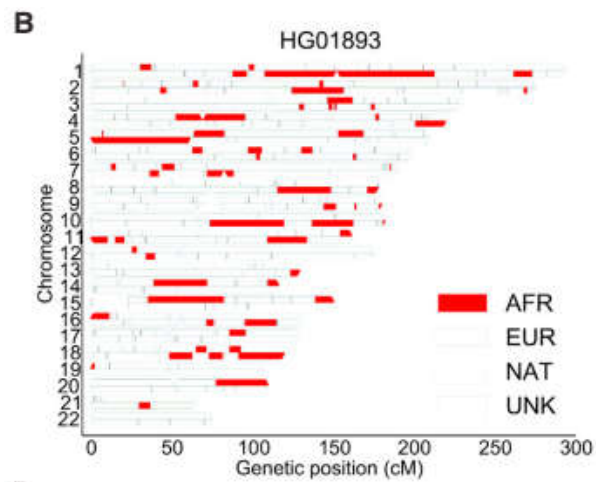
18 Years (N = 3,612)

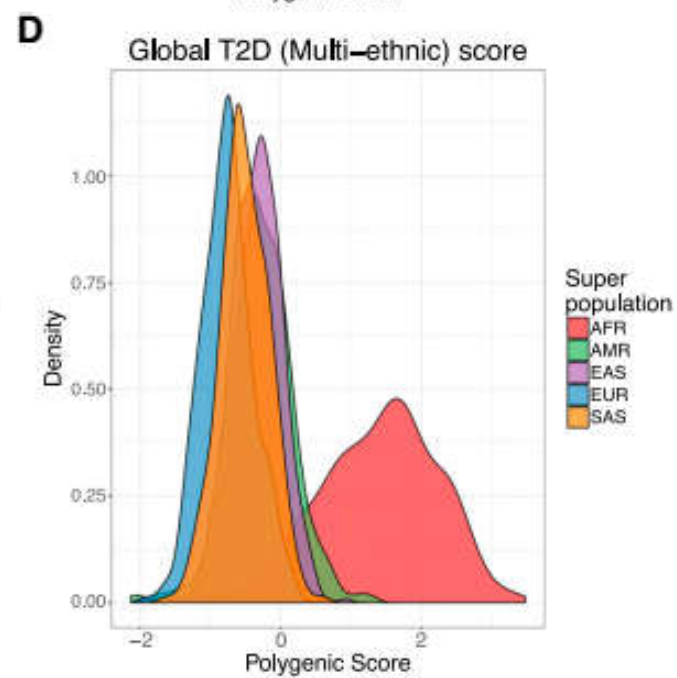
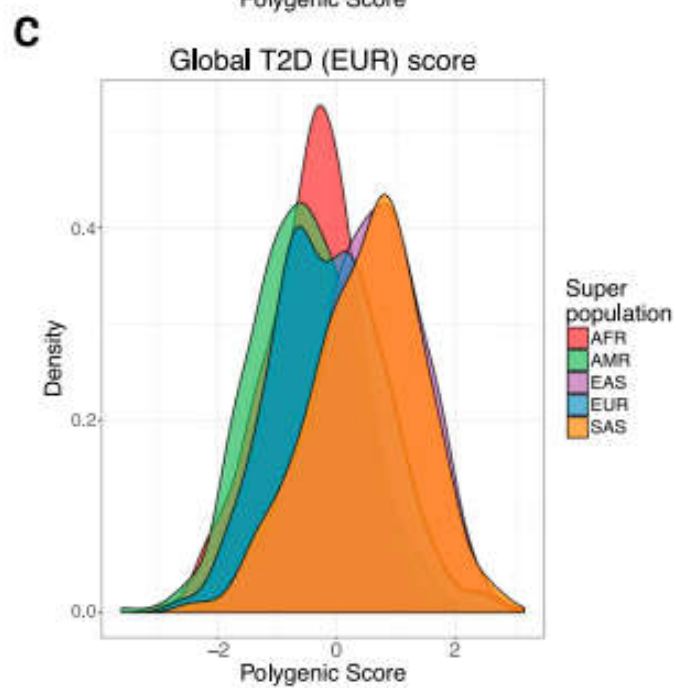
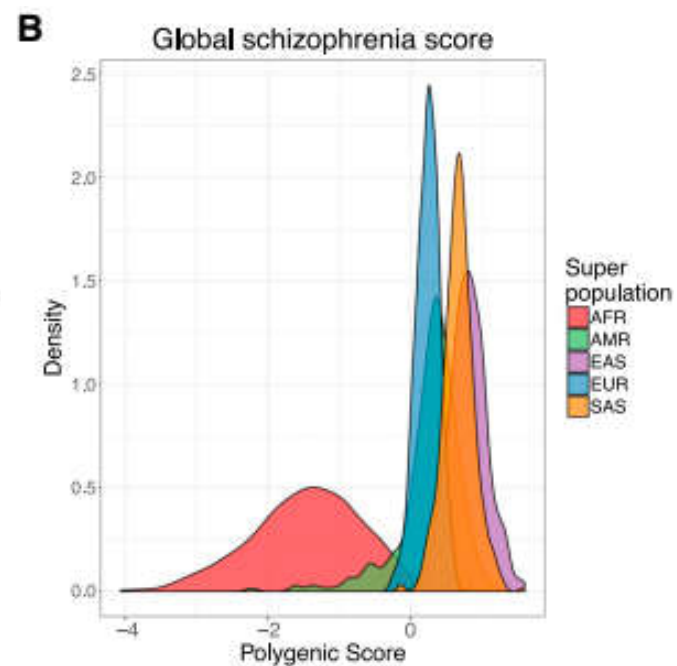
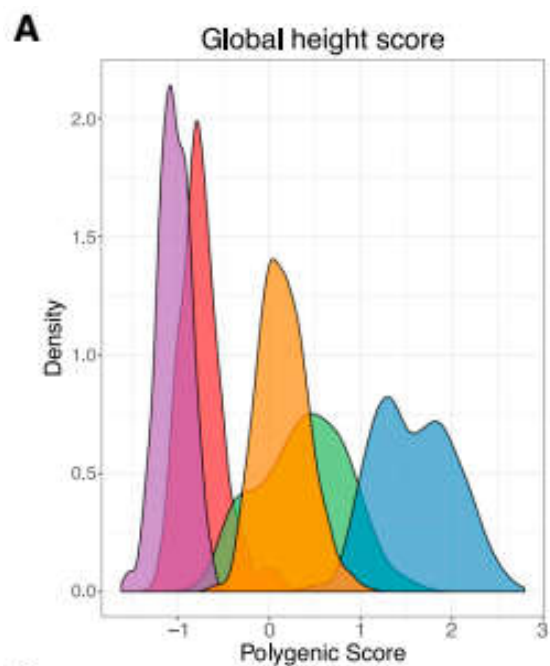


# Podemos testar em populações diferentes daquelas do GWAS original?

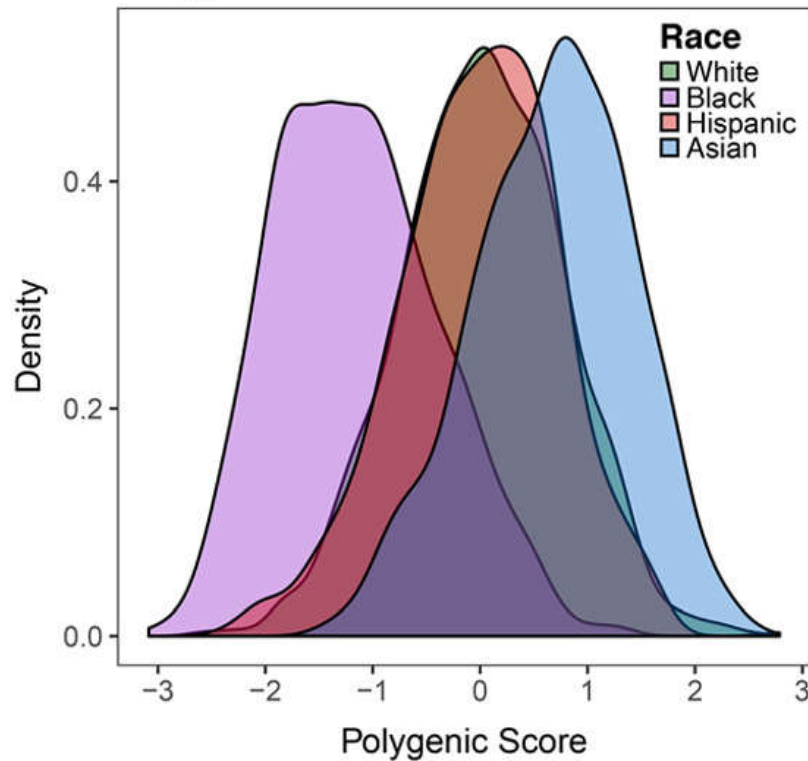


- Martin et al, 2017 AJHG

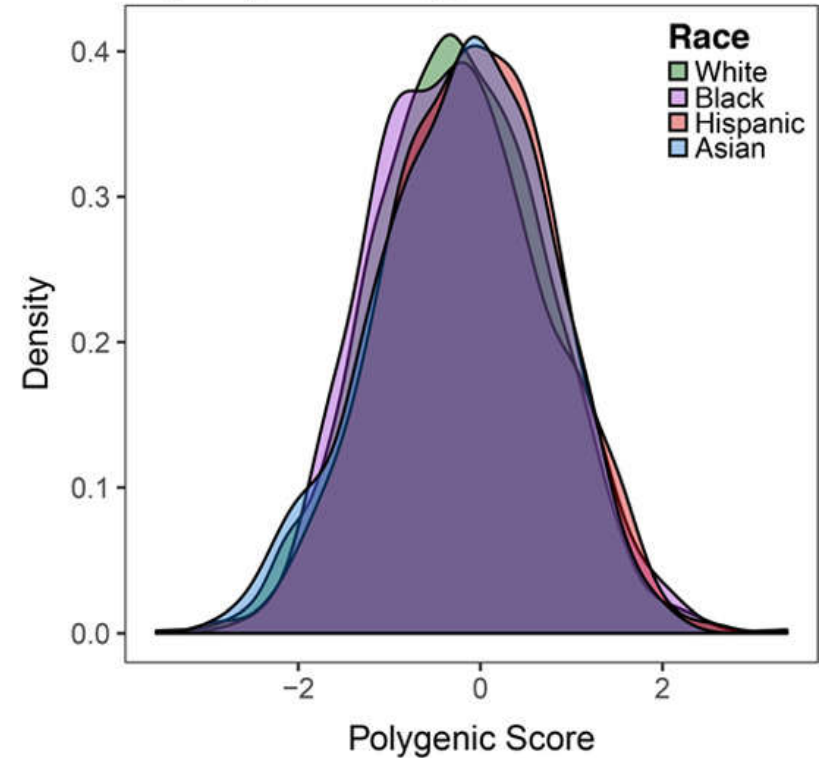




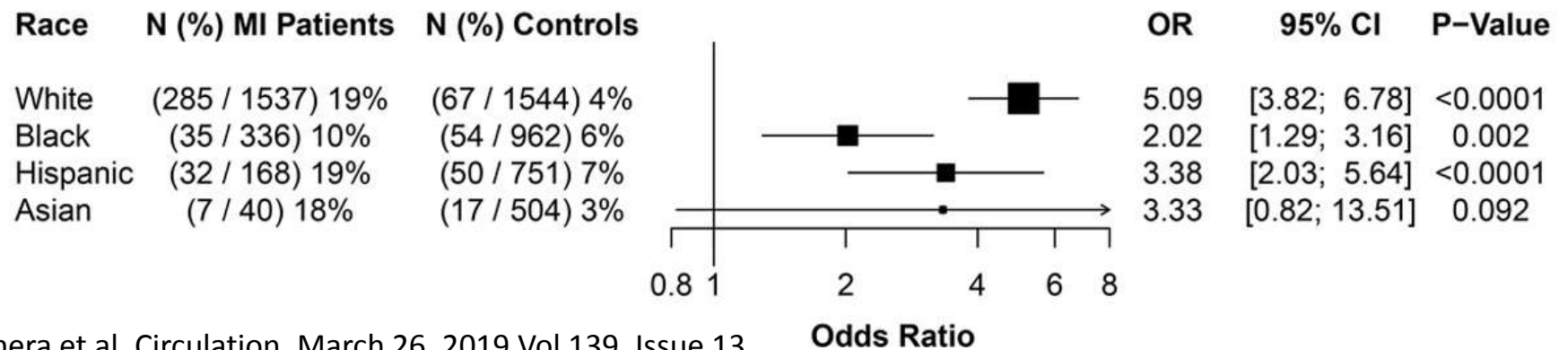
### A Raw Polygenic Score



### B Ancestry Adjusted Polygenic Score

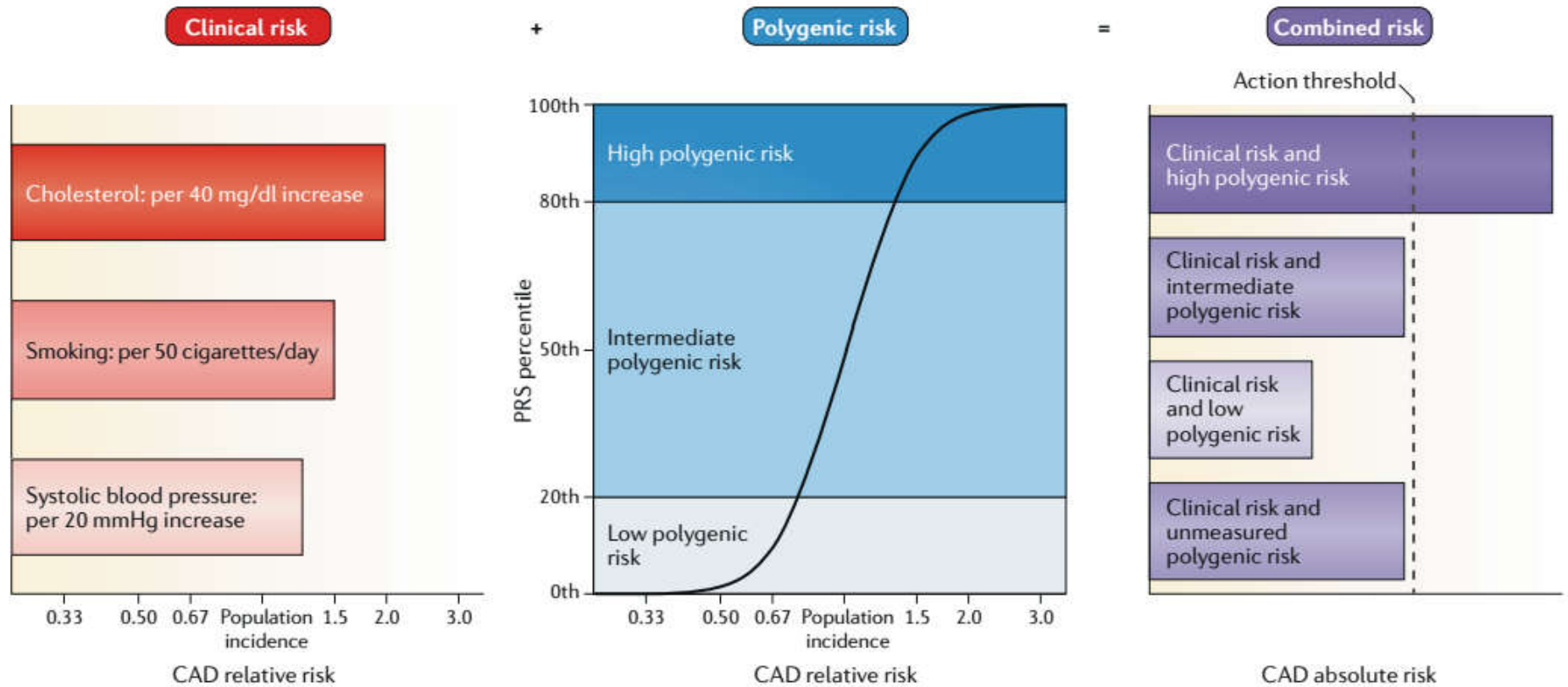


### C High Polygenic Score and Risk of Early-onset Myocardial Infarction



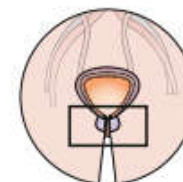
- Khera et al, Circulation. March 26, 2019 Vol 139, Issue 13

# Qual o impacto de PRS?



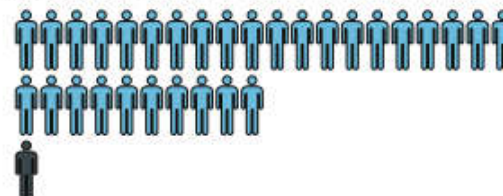
- Torkamani et al, Nature Reviews Genetics, 2018





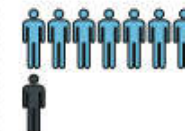
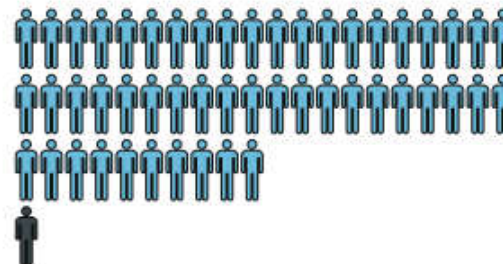
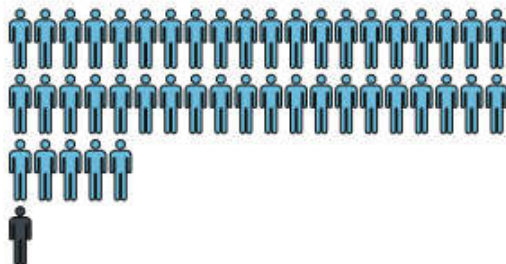
Top 20%

High risk



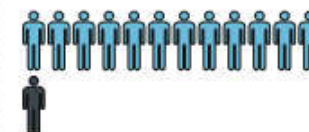
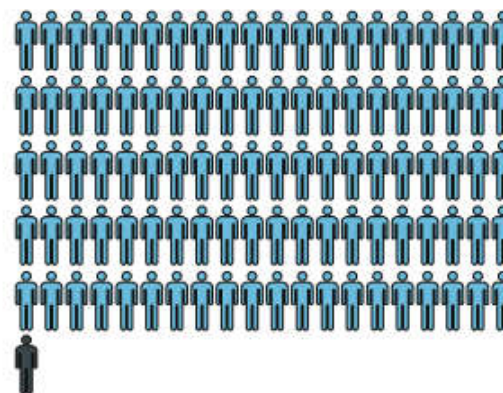
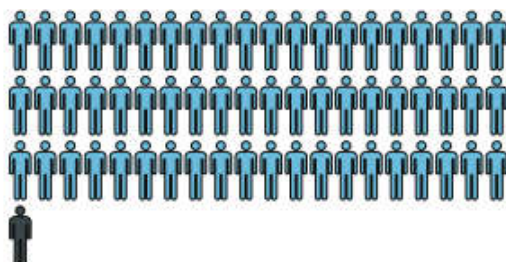
Middle 60%

Average risk



Bottom 20%

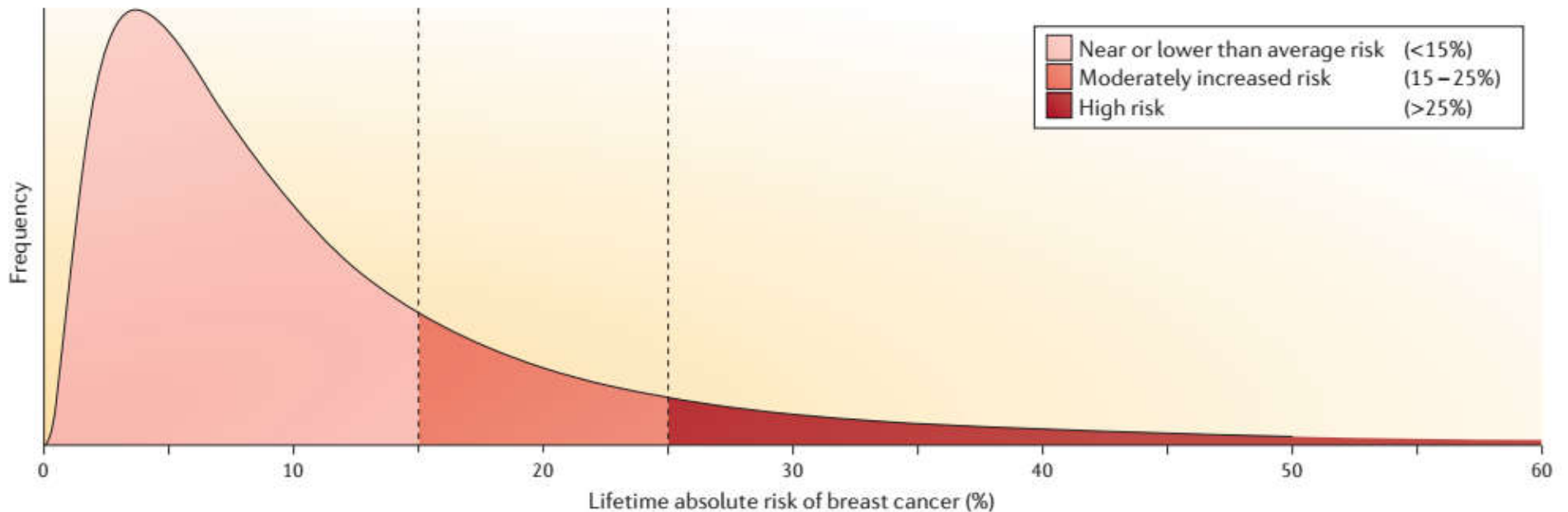
Low risk



### Possible clinical decisions

- |  |   |   |
|--|---|---|
| <ul style="list-style-type: none"> <li>• General advice on having a healthy lifestyle</li> <li>• Mammography screening frequency tailored to risk</li> </ul> | <ul style="list-style-type: none"> <li>• Lifestyle changes</li> <li>• Frequent mammography screening</li> <li>• Discuss preventive therapies</li> </ul> | <ul style="list-style-type: none"> <li>• Individual counselling in primary care and referral to secondary or tertiary care</li> <li>• Enhanced screening and surveillance</li> <li>• Chemoprevention and/or endocrine therapy</li> <li>• Risk-reducing surgery (mastectomy, salpingo-oophorectomy)</li> </ul> |
|--|---|---|

Absolute risk



### Possible risk factor profile

- |  |  |  |
|--|--|--|
| <ul style="list-style-type: none"> <li>• No family history of breast cancer, low to moderate polygenic risk, and none or few environmental risk factors</li> </ul> | <ul style="list-style-type: none"> <li>• No family history of breast cancer, moderate polygenic risk and several environmental risk factors</li> </ul> | <ul style="list-style-type: none"> <li>• Moderate to high polygenic risk with family history of breast cancer and many environmental risk factors, or known <i>BRCA1</i> and <i>BRCA2</i> or <i>TP53</i> mutation carriers for very high risk</li> </ul> |
|--|--|--|

# Genome-wide association studies for complex traits: consensus, uncertainty and challenges

Mark I. McCarthy<sup>\*†</sup>, Gonçalo R. Abecasis<sup>§</sup>, Lon R. Cardon<sup>\*||</sup>, David B. Goldstein<sup>¶</sup>, Julian Little<sup>#</sup>, John P. A. Ioannidis<sup>\*\*\*\*†</sup> and Joel N. Hirschhorn<sup>§§||||††</sup>

**Abstract** | The past year has witnessed substantial advances in understanding the genetic basis of many common phenotypes of biomedical importance. These advances have been the result of systematic, well-powered, genome-wide surveys exploring the relationships between common sequence variation and disease predisposition. This approach has revealed over 50 disease-susceptibility loci and has provided insights into the allelic architecture of multifactorial traits. At the same time, much has been learned about the successful prosecution of association studies on such a scale. This Review highlights the knowledge gained, defines areas of emerging consensus, and describes the challenges that remain as researchers seek to obtain more complete descriptions of the susceptibility architecture of biomedical traits of interest and to translate the information gathered into improvements in clinical management.

## Genome-wide association (GWA) studies

Studies in which a dense array of genetic markers, which captures a substantial proportion of common variation in genome sequence, is typed in a set of DNA samples that are informative for a trait of interest. The aim is to map susceptibility effects through the detection of associations between genotype frequency and trait status.

The first wave of large-scale, high-density genome-wide association (GWA) studies has improved our understanding of the genetic basis of many complex traits<sup>1</sup>. For several diseases, including type 1 (REFS 2, 3) and type 2 diabetes<sup>4–9</sup>, inflammatory bowel disease<sup>10–14</sup>, prostate cancer<sup>15–20</sup> and breast cancer<sup>21–23</sup>, there has been rapid expansion in the numbers of loci implicated in predisposition. For others, such as asthma<sup>24</sup>, coronary heart disease<sup>25–27</sup> and atrial fibrillation<sup>28</sup>, fewer novel loci have been found, although opportunities for mechanistic insights are equally promising. Several common variants influencing important continuous traits, such as lipids<sup>29–31</sup>, height<sup>32–35</sup> and fat mass<sup>36–38</sup>, have also been found. An updated list of published GWA studies can be found at the National Cancer Institute (NCI)-National Human Genome Research Institute (NHGRI)'s catalog of published genome-wide association studies.

These findings are providing valuable clues to the allelic architecture of complex traits in general. At the same time, many methodological and technical issues that are relevant to the successful prosecution of large-scale association studies have been addressed. However, despite understandable celebration of these achievements, sober reflection reveals many challenges ahead. Compelling signals have been found, often highlighting previously unsuspected biology, but, for most of the

traits studied, known variants explain only a fraction of observed familial aggregation<sup>39</sup>, limiting the potential for early application to determine individual disease risk. Because current technology surveys only a limited subset of potentially relevant sequence variation, this should come as no surprise. Much work remains to obtain a complete inventory of the variants at each locus that contribute to disease risk and to define the molecular mechanisms through which these variants operate. The ultimate objectives — full descriptions of the susceptibility architecture of major biomedical traits and translation of the findings into clinical practice — remain distant.

With completion of the initial wave of GWA scans, it is timely to consider the status of the field. This Review considers each major step in the implementation of a GWA scan, highlighting areas where there is an emerging consensus over the ingredients for success, and those aspects for which considerable challenges remain.

## Subject ascertainment and design

Although there is a growing focus on the application of GWA methodologies to population-based cohorts, most published GWA studies have featured case-control designs, which raise issues related to the optimal selection of both case and control samples.

<sup>\*</sup> Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.  
Correspondence to M.I.M.  
e-mail: mark.mccarthy@dr1.ox.ac.uk  
doi:10.1038/nrg2344  
Published online 9 April 2008



# Dissecting the genetics of complex traits using summary association statistics

Bogdan Pasaniuc<sup>1</sup> and Alkes L. Price<sup>2,3</sup>

**Abstract** | During the past decade, genome-wide association studies (GWAS) have been used to successfully identify tens of thousands of genetic variants associated with complex traits and diseases. These studies have produced extensive repositories of genetic variation and trait measurements across large numbers of individuals, providing tremendous opportunities for further analyses. However, privacy concerns and other logistical considerations often limit access to individual-level genetic data, motivating the development of methods that analyse summary association statistics. Here, we review recent progress on statistical methods that leverage summary association data to gain insights into the genetic basis of complex traits and diseases.

## Individual-level data

Genome-wide single nucleotide polymorphism genotypes and trait values for each individual included in a genome-wide association study.

## Summary association statistics

Estimated effect sizes and their standard errors for each single nucleotide polymorphism analysed in a genome-wide association study.

<sup>1</sup>Departments of Human Genetics, and Pathology and Laboratory Medicine, University of California, Los Angeles, California 90095, USA.

<sup>2</sup>Departments of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA.

<sup>3</sup>Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142, USA. pasaniuc@ucla.edu; aprice@hsph.harvard.edu

doi:10.1038/nrg.2016.142  
Published online 14 Nov 2016

Genome-wide association studies (GWAS) have been broadly successful in identifying genetic variants associated with complex traits and diseases, explaining a significant fraction of narrow-sense heritability and occasionally pinpointing biological mechanisms<sup>1</sup>. These studies have produced extensive databases of genetic variation (typically at the level of common single nucleotide polymorphisms (SNPs) included on genotyping arrays) in large numbers of individuals across hundreds of complex traits. Further analyses of these data can yield important insights into the genetics of complex traits, but privacy concerns and other logistical considerations often restrict access to individual-level data. Nevertheless, summary association statistics are often readily available and can be used to compute *z*-scores (FIG. 1). Here, we define summary association statistics as per-allele SNP effect sizes (log odds ratios for case-control traits) together with their standard errors, although we note that some applications may also require allele frequencies. A list of selected publicly available summary association statistics from large GWAS is provided in TABLE 1. Analyses of summary statistics also offer advantages in computational cost, which does not scale with the number of individuals in the study. These advantages have motivated the recent development of many new methods for analysing summary association data, often in conjunction with linkage disequilibrium (LD) information from a population reference panel such as 1000 Genomes<sup>2</sup>.

Here, we review these summary statistic-based methods. First, we review methods for performing single-variant association tests, including meta-analysis,

conditional association and imputation using summary statistics. Second, we review methods for performing gene-based association tests by incorporating transcriptome reference data or aggregating signals across multiple rare variants. Third, we review methods for fine-mapping causal variants, including the integration of functional annotation and/or trans-ethnic data. Fourth, we review methods for constructing polygenic predictions of disease risk and inferring polygenic architectures. Finally, we review methods for jointly analysing multiple traits. We conclude with a discussion of research areas for which further work on summary statistic-based methods is needed.

## Single-variant association tests

**Meta-analysis using fixed-effects or random-effects models.** Large consortia often combine multiple GWAS into a single aggregate analysis to boost power for discovering SNP associations with small effects. Studies are combined either by jointly analysing summary association results from each study (meta-analysis) or by re-analysing individual-level data across all studies (mega-analysis)<sup>3</sup>. It has been shown that a meta-analysis attains similar power for association as a mega-analysis, with fewer privacy constraints and logistical challenges (because only summary association data are shared across studies)<sup>4</sup>. A meta-analysis is usually performed using fixed-effects approaches, which assume that true effect sizes are the same across studies. Under the assumption that causal effect sizes may differ across studies, this heterogeneity can be explicitly modelled using random-effects methods. These methods include

## TRANSLATIONAL GENETICS

## The personal and clinical utility of polygenic risk scores

Ali Torkamani<sup>1,2\*</sup>, Nathan E. Wineinger<sup>1,2</sup> and Eric J. Topol<sup>1,3</sup>

**Abstract** | Initial expectations for genome-wide association studies were high, as such studies promised to rapidly transform personalized medicine with individualized disease risk predictions, prevention strategies and treatments. Early findings, however, revealed a more complex genetic architecture than was anticipated for most common diseases — complexity that seemed to limit the immediate utility of these findings. As a result, the practice of utilizing the DNA of an individual to predict disease has been judged to provide little to no useful information. Nevertheless, recent efforts have begun to demonstrate the utility of polygenic risk profiling to identify groups of individuals who could benefit from the knowledge of their probabilistic susceptibility to disease. In this context, we review the evidence supporting the personal and clinical utility of polygenic risk profiling.

## Polygenic risk scores

(PRSs). A weighted sum of the number of risk alleles carried by an individual, where the risk alleles and their weights are defined by the loci and their measured effects as detected by genome wide association studies.

## Genetic architecture

The underlying genetic basis of a trait or disease. The combination of the number, type, frequency, relationship between and magnitude of effect of genetic variants contributing to a trait.

Estimating the probabilistic susceptibility of an individual to disease — risk prediction — is central to clinical decision-making, especially in the context of early disease detection and prevention of common adult-onset conditions. Moreover, it can be a powerful tool for personal health management when communicated and understood effectively. Today, clinical risk prediction for common adult-onset diseases often relies on basic demographic characteristics, such as age, gender and ethnicity; basic health parameters and lifestyle factors, such as body mass index, smoking status, alcohol consumption and physical exercise habits; measurement of clinical risk factors proximal to overt disease onset, such as blood pressure levels, blood chemistries or biomarkers indicative of ongoing disease processes; ascertainment of environmental exposures, such as air pollution, heavy metals and other environmental toxins; and family history. Routine genetic profiling is conspicuously absent from this list, often relegated to use only when testing clarifies individual-level risks in the context of a known family history for some common adult-onset diseases.

Early disease detection, prevention and intervention are fundamental goals for advancing human health. Meanwhile, genetic risk estimation is, for all intents and purposes, the earliest measurable contributor to common heritable disease risk. Thus, in theory, genetic profiling could be considered a useful component of health management. Indeed, recent studies suggest that, for a subset of diseases, our knowledge of the genetic factors underlying these conditions has improved to a point where polygenic risk profiling on the basis of calculated polygenic risk scores (PRSs) provides personal and clinical utility.

Here, we review the utility of genetic risk profiling for common adult-onset polygenic conditions, focusing on the leading heritable causes of death in the developed world: Alzheimer disease, cancer (breast and prostate), coronary artery disease and type 2 diabetes mellitus. For these conditions, recent studies have linked polygenic risk prediction to actionable outcomes, including the prioritization of preventive interventions and screening<sup>1–3</sup>, prediction of age of disease onset<sup>4</sup>, benefit from lifestyle modifications<sup>5–7</sup> and modification of familial disease risk leading to changes in clinical decision-making<sup>6–8</sup>. We begin with an overview of the genetic architecture of common adult-onset diseases. We then describe how genetic risk factors can be combined to produce PRSs and review recent studies that have demonstrated the utility of PRSs for disease risk stratification as well as their implications for early disease detection, prevention, therapeutic intervention and/or life planning. We describe some of the limitations of PRSs and the remaining barriers to clinical and personal utility and lay out potential future directions for the enhancement of the predictive capacity, generalizability and utility of PRSs.

## Genetic inheritance of common diseases

The basic components of disease risk are usually broken down into genetic susceptibility, environmental exposures and lifestyle factors. The relative contribution of genetic susceptibility to the predisposition to disease in a population can be quantified by the heritability of the disease in that population. Heritability itself can be defined in several ways<sup>9</sup>; from a quantitative genetics perspective — especially as it relates to missing heritability in genome-wide association studies (GWAS)<sup>10</sup> — it is usually

<sup>1</sup>The Scripps Translational Science Institute, The Scripps Research Institute, La Jolla, CA, USA.

<sup>2</sup>Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA.

<sup>3</sup>Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA.

\*e-mail: atorkama@scripps.edu

<https://doi.org/10.1038/s41576-018-0018-x>

# Developing and evaluating polygenic risk prediction models for stratified disease prevention

Nilanjan Chatterjee<sup>1-3</sup>, Jianxin Shi<sup>3</sup> and Montserrat Garcia-Closas<sup>3</sup>

**Abstract** | Knowledge of genetics and its implications for human health is rapidly evolving in accordance with recent events, such as discoveries of large numbers of disease susceptibility loci from genome-wide association studies, the US Supreme Court ruling of the non-patentability of human genes, and the development of a regulatory framework for commercial genetic tests. In anticipation of the increasing relevance of genetic testing for the assessment of disease risks, this Review provides a summary of the methodologies used for building, evaluating and applying risk prediction models that include information from genetic testing and environmental risk factors. Potential applications of models for primary and secondary disease prevention are illustrated through several case studies, and future challenges and opportunities are discussed.

## Penetrance

The proportion of individuals in a population with a genetic variant who develop the disease associated with that variant. Common single-nucleotide polymorphisms (SNPs) are referred to as low-penetrant, as risk alleles typically confer modest risk.

<sup>1</sup>Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University.

<sup>2</sup>Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA.

<sup>3</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20892, USA.

Correspondence to N.C. [nilanjan@jhu.edu](mailto:nilanjan@jhu.edu)

doi:10.1038/nrg.2016.27  
Published online 5 May 2016

Common chronic diseases have complex, multifactorial aetiologies that involve the interplay of both genetic susceptibility and environmental risk factors, which are broadly defined as lifestyle, behavioural, occupational or environmental exposures, and other health conditions. Historically, family-based linkage studies have led to the identification of rare high-penetrant mutations underlying some of these diseases, such as those in the breast cancer 1 (*BRCA1*) and *BRCA2* genes for breast and ovarian cancers, and in multiple genes involved in Lynch syndrome, which predisposes individuals to colorectal and other cancers. With these discoveries, genetic testing became part of the clinical management of individuals in high-risk families in whom there is a high disease burden caused by the variants. The cost of genetic testing has declined following technological advances and the recent ruling by the US Supreme Court stating that genes cannot be patented; consequently, debate has now shifted towards the implications of performing genetic testing in the general population (for example, *BRCA1* and *BRCA2* mutation testing)<sup>1,2</sup>. Debate has also begun on standards for the regulation and clinical utility of increasingly available commercial gene-panel tests, which may screen for high- to moderate-penetrance susceptibility variants for various diseases<sup>3-6</sup>.

As the majority of cases of common diseases do not occur in highly affected families, the development of broad public health strategies for disease prevention requires the identification of risk factors that contribute to the substantial burden of disease in the general population. Recent genome-wide association studies (GWAS)

have clearly shown that common single-nucleotide polymorphisms (SNPs) have important roles in defining susceptibility to common diseases. For any given disease, there could be a large number of underlying susceptibility SNPs, each exhibiting only modest disease association, but in combination they could explain a significant portion of the variation in disease incidence in the general population. The success of GWAS indicates that gene-panel and whole-genome tests will continue to emerge in the future for the assessment of polygenic disease risks. This will require critical evaluation of both the statistical validity of the estimated risk and its potential clinical or public health utility.

The utility of genetic testing for disease prevention cannot be fully evaluated unless it is assessed along with environmental factors, which may not only be important determinants of risk but could also be potentially modifiable through changes in lifestyle or appropriate interventions. Thus, there is a need for continuous development and evaluation of risk models that incorporate our expanding knowledge of the risk factors for diseases. Critical to this research are epidemiological prospective cohort studies that can take advantage of the increasingly available electronic medical records, technological advances in the collection and analyses of biological specimens, and big data management platforms and analytics. Steps are being taken towards attaining these goals, as demonstrated by the establishment of new cohorts and biobanks, including [UK Biobank](#), [China Kadoorie Biobank](#), the German National Cohort<sup>7</sup>, the American Cancer Society's [Cancer](#)



*Article*

# Genotype Fingerprints Enable Fast and Private Comparison of Genetic Testing Results for Research and Direct-to-Consumer Applications

Max Robinson  and Gustavo Glusman \* 

Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA; Max.Robinson@SystemsBiology.org

\* Correspondence: Gustavo@SystemsBiology.org; Tel.: +1-206-732-1273

Received: 31 August 2018; Accepted: 2 October 2018; Published: 4 October 2018

