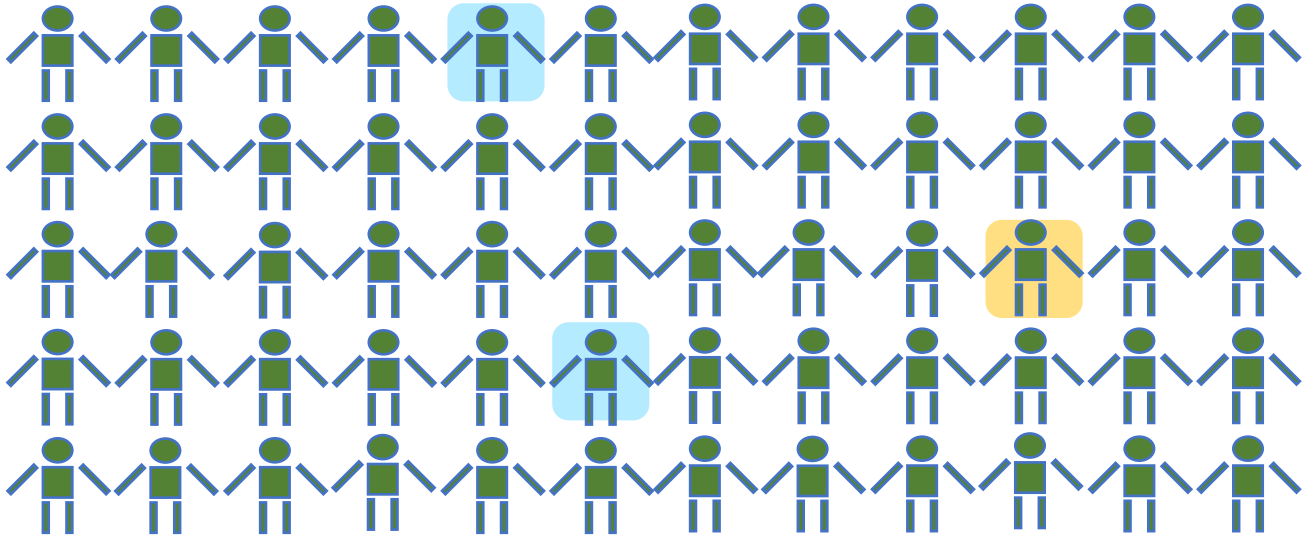
Genomic Epidemiology:
Linking Precision Health with Populations

Marta Gwinn, MD, MPH

Department of Epidemiology

Rollins School of Public Health

Emory University



ORISE Enrichment Event
September 7-8, 2023

Outline

- How genetics and epidemiology evolved in parallel
- Why group-level data are required to assess individual risk
- An early vision for genomic medicine
- How genetic association studies got so large
- When polygenic risk scores are biased
- How genomics can help evaluate environmental risk factors

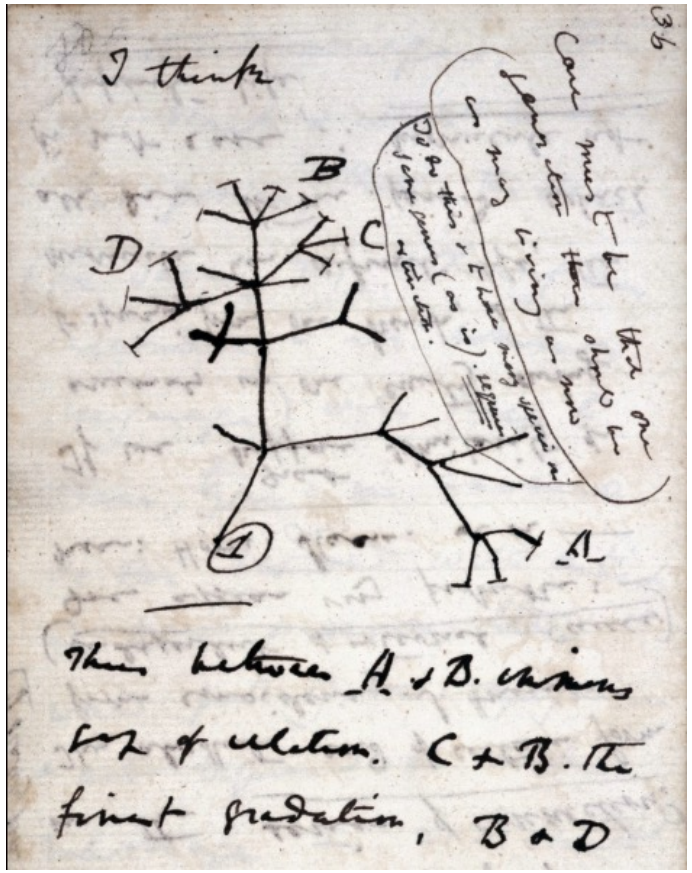
M Gwinn 2023

Outline

- How genetics and epidemiology evolved in parallel

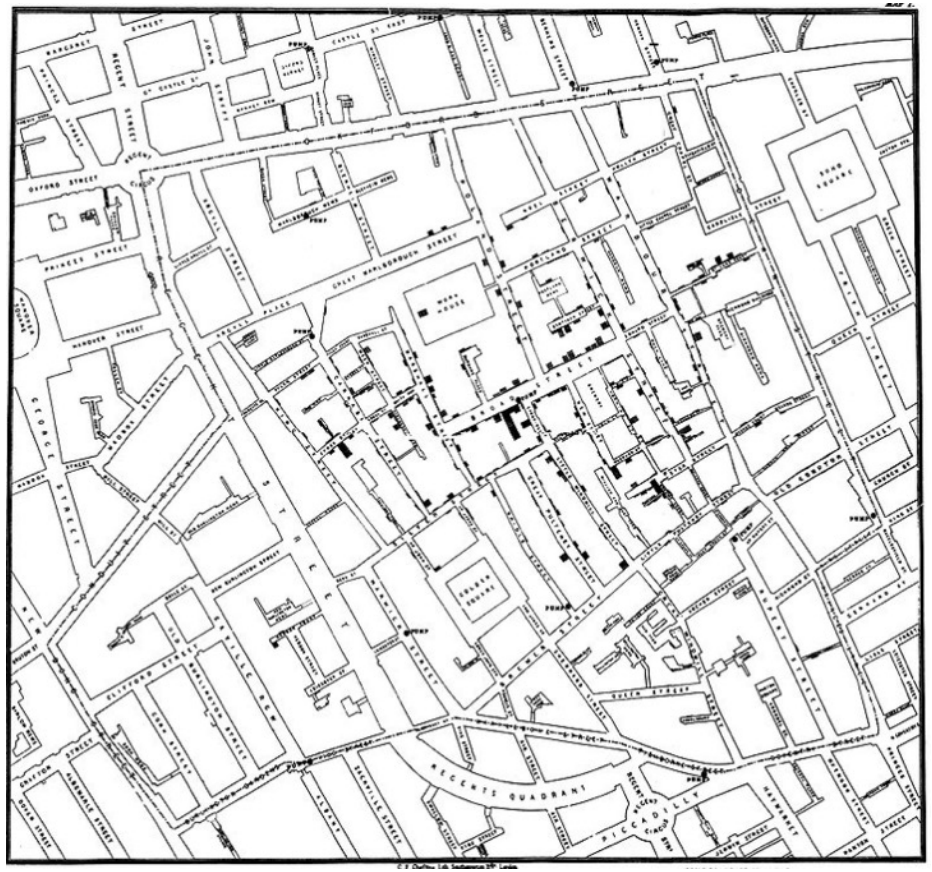
19th century

Genetics



Evolutionary tree
Charles Darwin, 1837
Origin of Species, 1859

Epidemiology

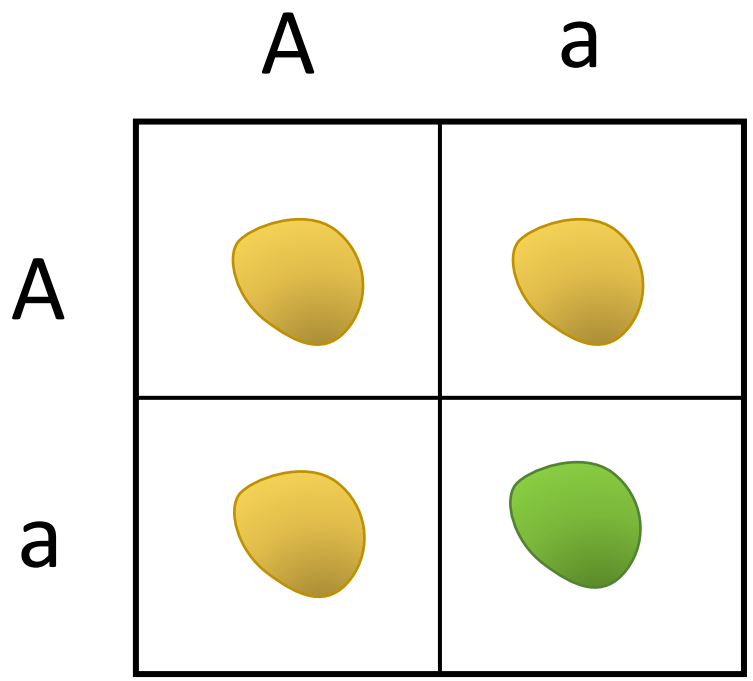


Cholera in London
John Snow, 1854

M Gwinn 2023

Early 20th century

Genetics



Punnett square, 1905

Epidemiology

	+	—
+	a	b
—	c	d

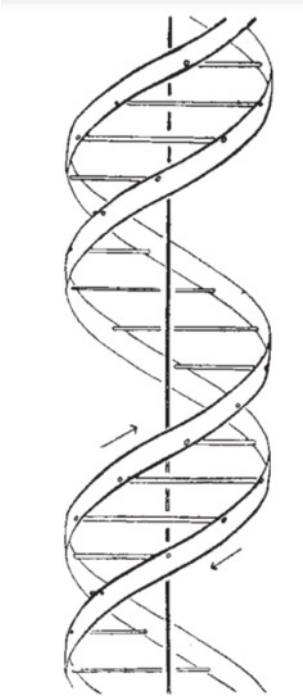
Fisher's exact test, 1922

Rediscovery of Mendel's experiments → Foundations of genetics and statistics

M Gwinn 2023

Mid 20th century

Genetics



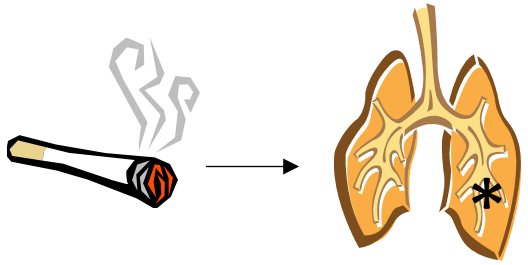
DNA structure
Watson and Crick, 1953

Molecular biology

Epidemiology

TABLE IV.—*Proportion of Smokers and Non-smokers in Lung-carcinoma Patients and in Control Patients with Diseases Other Than Cancer*

Disease Group	No. of Non-smokers	No. of Smokers	Probability Test
Males: Lung-carcinoma patients (649)	2 (0·3%)	647	P (exact method) = 0·00000064
Control patients with diseases other than cancer (649) ..	27 (4·2%)	622	
Females: Lung-carcinoma patients (60)	19 (31·7%)	41	$\chi^2 = 5{\cdot}76$; n = 1 0·01 < P < 0·02
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28	



Smoking and carcinoma of the lung
Doll and Hill, 1950

Risk factor epidemiology

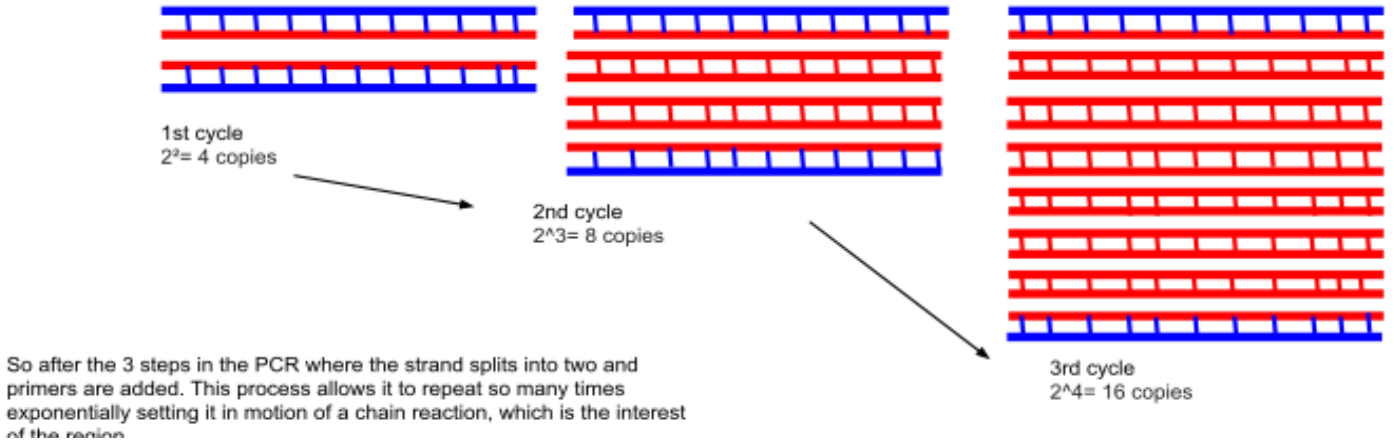
M Gwinn 2023

Late 20th century

High-throughput genotyping

Exponential amplification

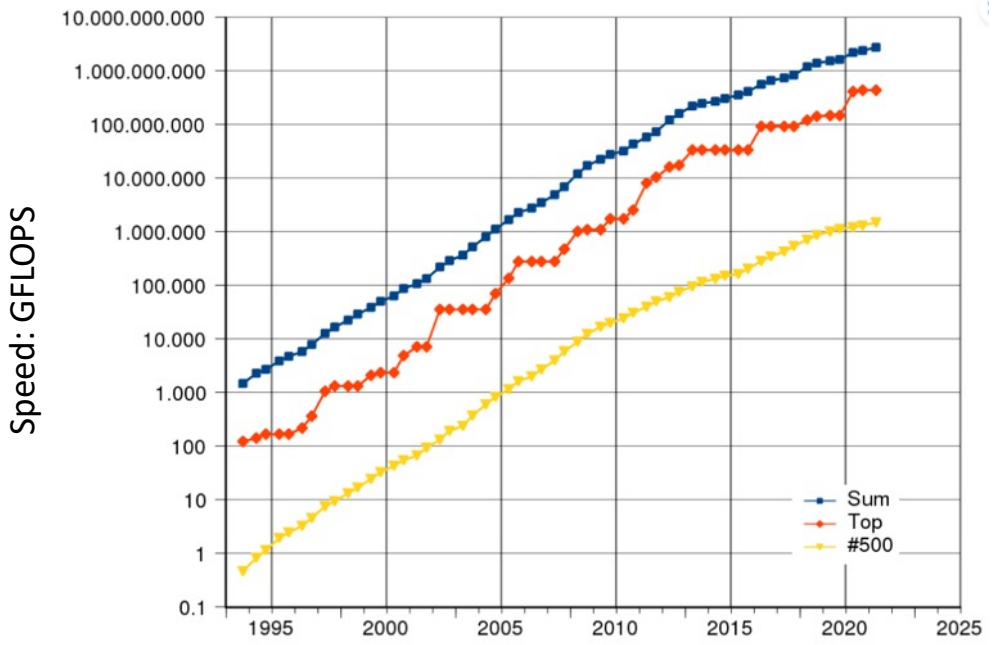
Polymerase chain reaction (PCR)
Mullis and Smith, 1983



High-performance computing

Exponential performance

TOP500, founded 1993



M Gwinn 2023

Early 21st century

The selfie

The “quantified self”

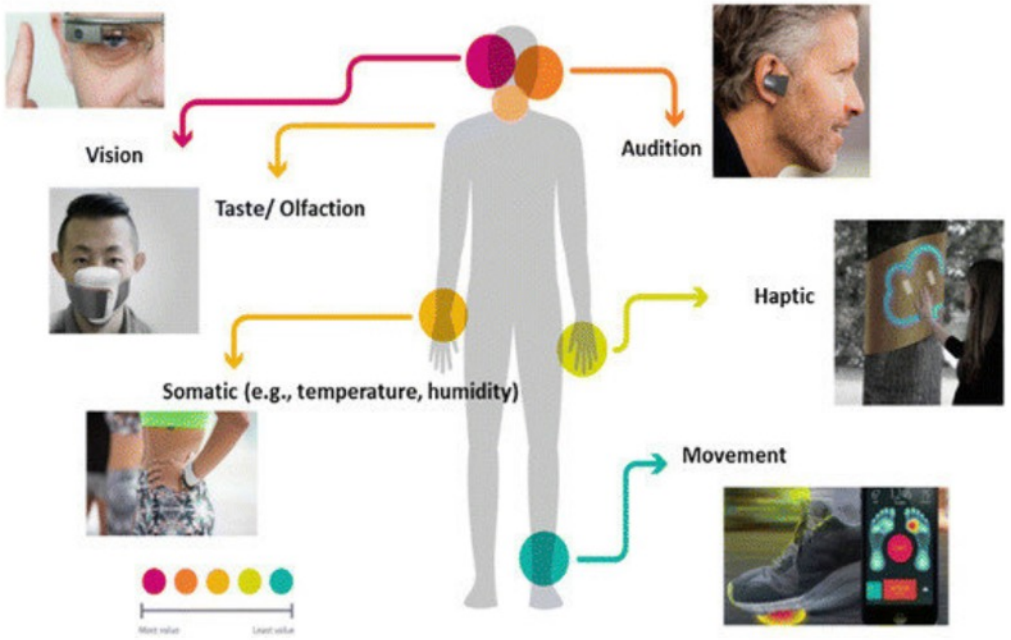


Figure 1: Applications and wearable devices used for ‘Quantified-Self’ adapted from Kim and Fesenmaier (2015)

M Gwinn 2023

Outline

- How genetics and epidemiology evolved in parallel
- Why group-level data are required to assess individual risk

Why do we need population-level data to assess individual risk?

Epidemiology:

The *determinants* and *distribution* of health and disease in *defined populations*.

Medicine:

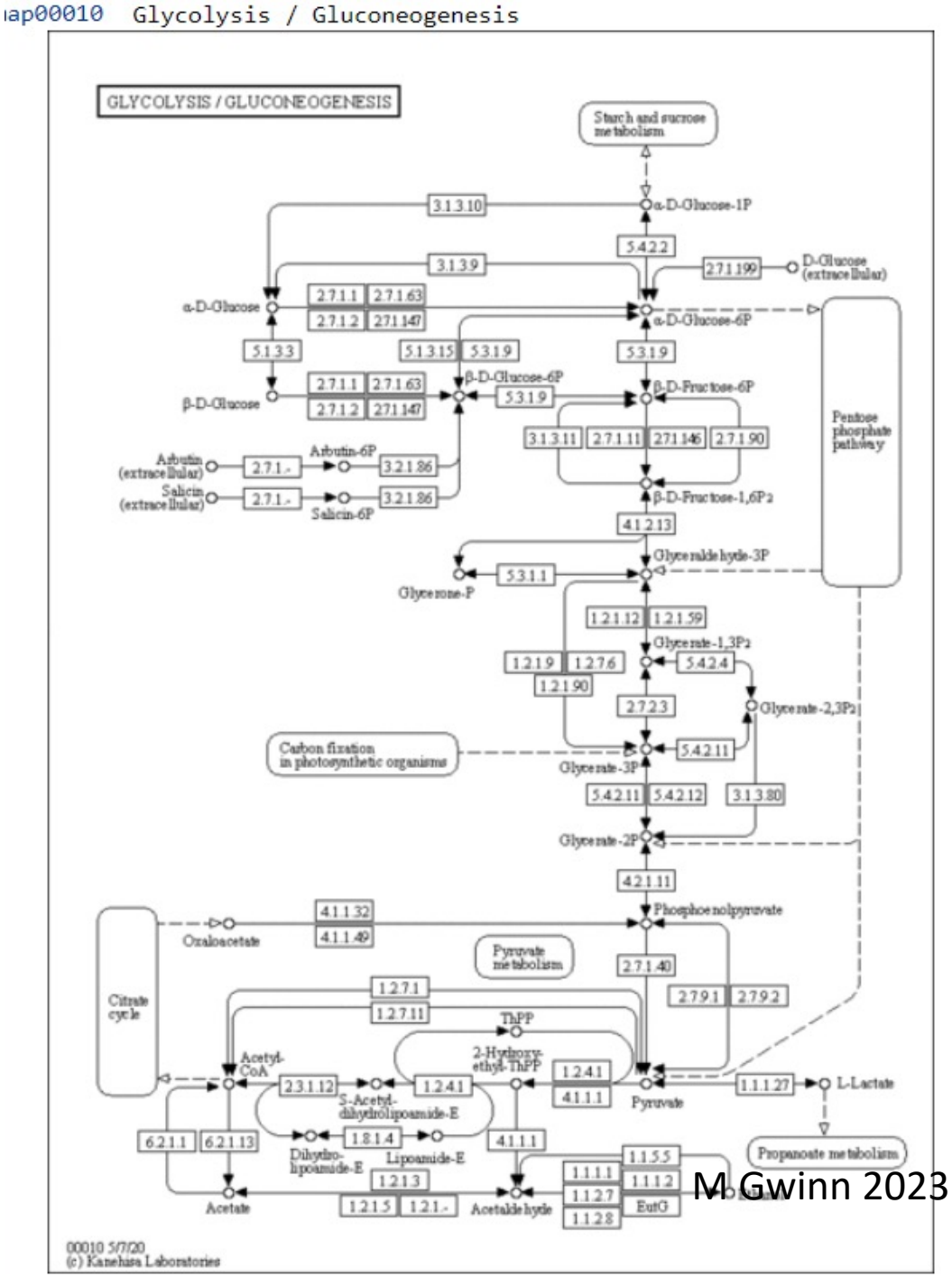
The *cause* and *occurrence* of disease in an individual.

Why can't we just use more—and more precise—individual measurements?

M Gwinn 2023

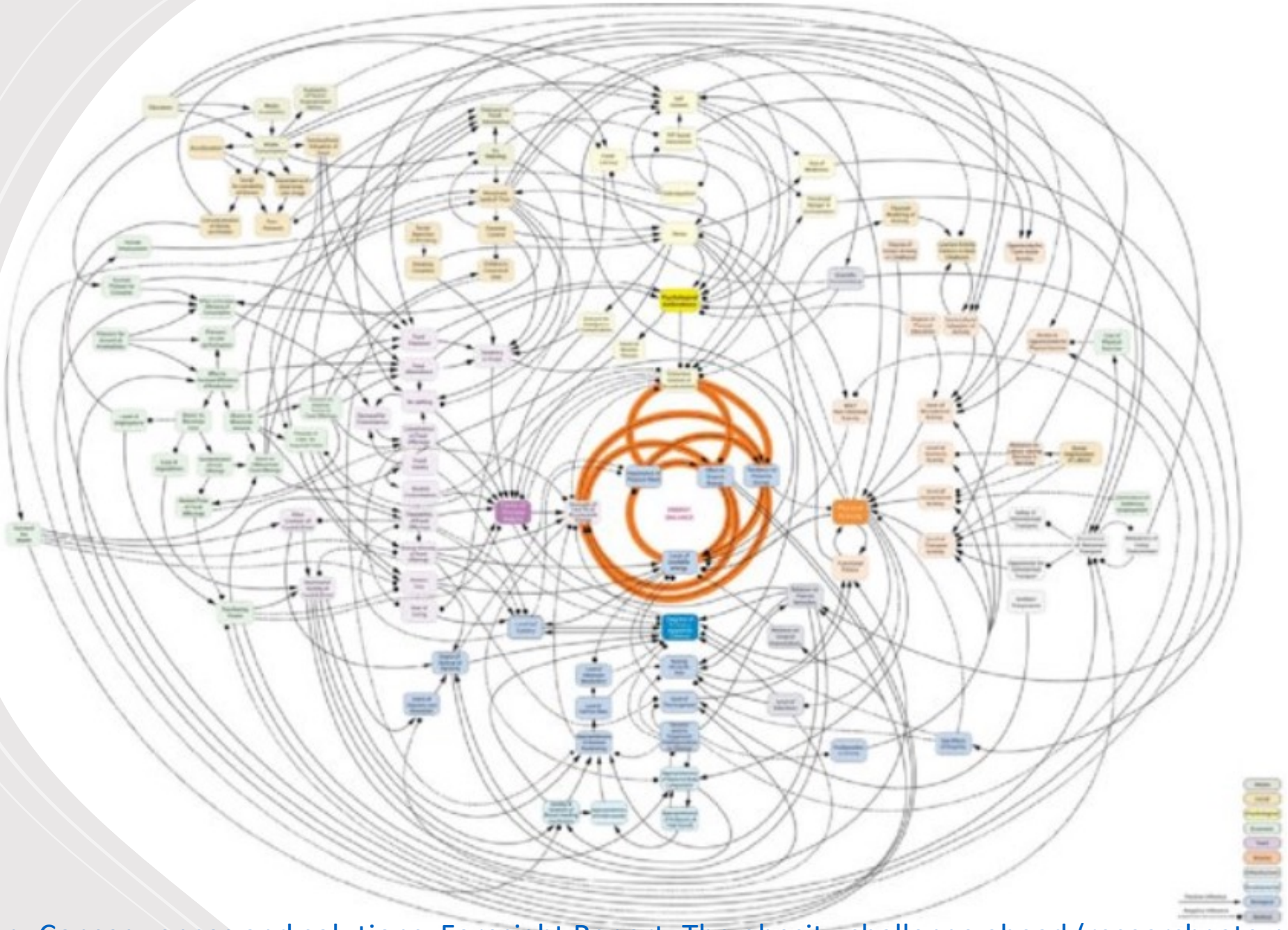
The causes of common diseases are too complex and dynamic to predict by using data from a single individual.

[KEGG PATHWAY: map00010 (genome.jp)]



M Gwinn 2023

Causal processes are also subject to random variation.



[Symposium 1: Overnutrition: Consequences and solutions. Foresight Report: The obesity challenge ahead (researchgate.net)](#)
Kopelman P. 2009

Epidemiology:

The *determinants* and *distribution* of health and disease in *defined populations*.

The *cause* and *occurrence* of disease in an individual.

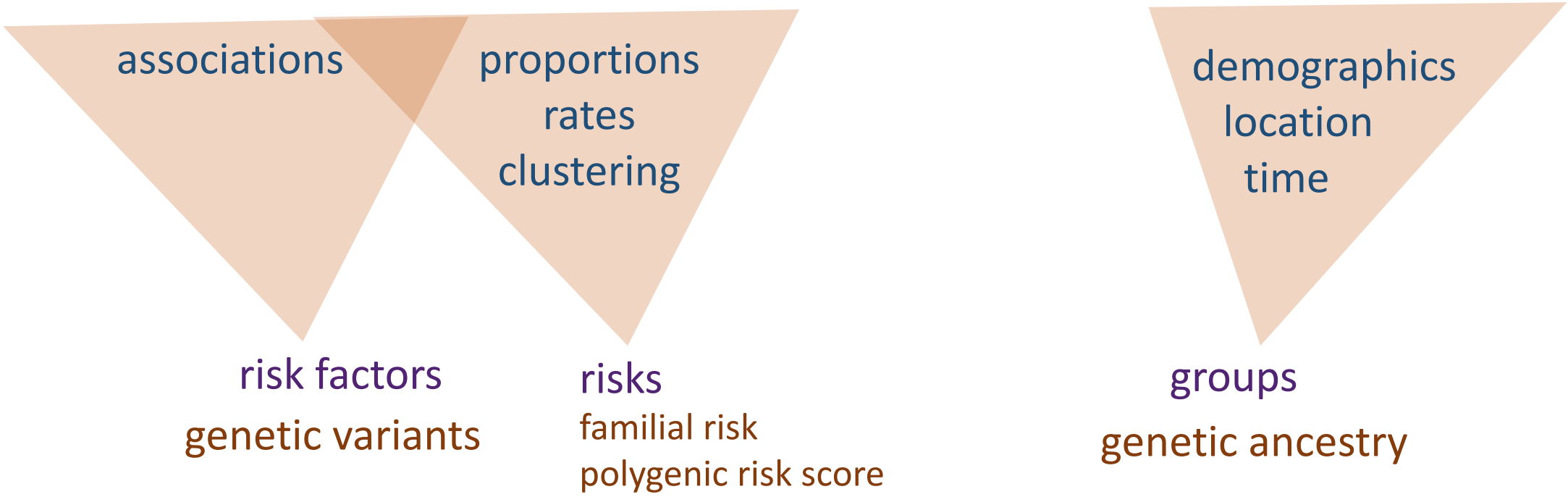
Epidemiologic analysis of population-based data can identify "risk factors" and estimate their effects, incorporating the uncertainty due to random variation.

M Gwinn 2023

Genomic

Epidemiology:

The *determinants* and *distribution* of health and disease in *defined populations*.



M Gwinn 2023

Outline

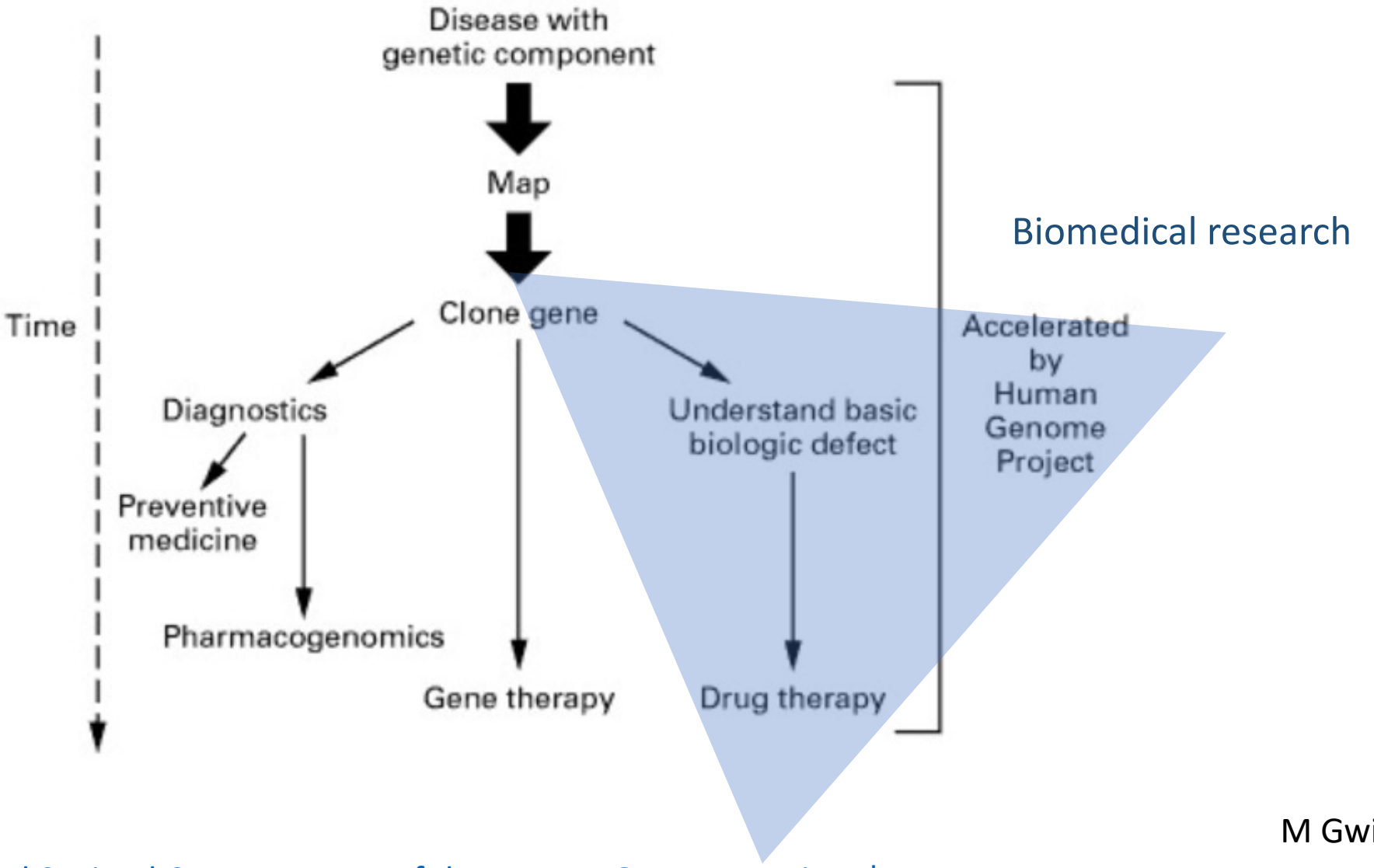
- How genetics and epidemiology evolved in parallel
- Why group-level data are required to assess individual risk
- An early vision for genomic medicine

Early vision for genomic medicine (through the retrospect-o-scope)

M Gwinn 2023

Medical and Societal Consequences of the Human Genome Project

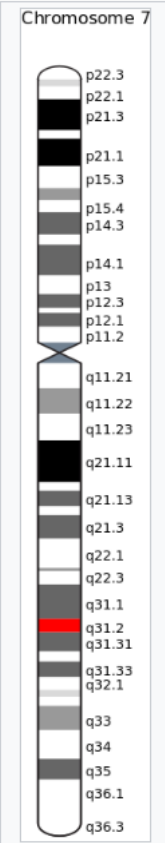
Francis S. Collins, MD, PhD. N Engl J Med **1999**; 341:28-37



M Gwinn 2023

Medical and Societal Consequences of the Human Genome Project | NEJM

“Single-gene disorder”: Cystic fibrosis



The location of the *CFTR* gene on chromosome 7

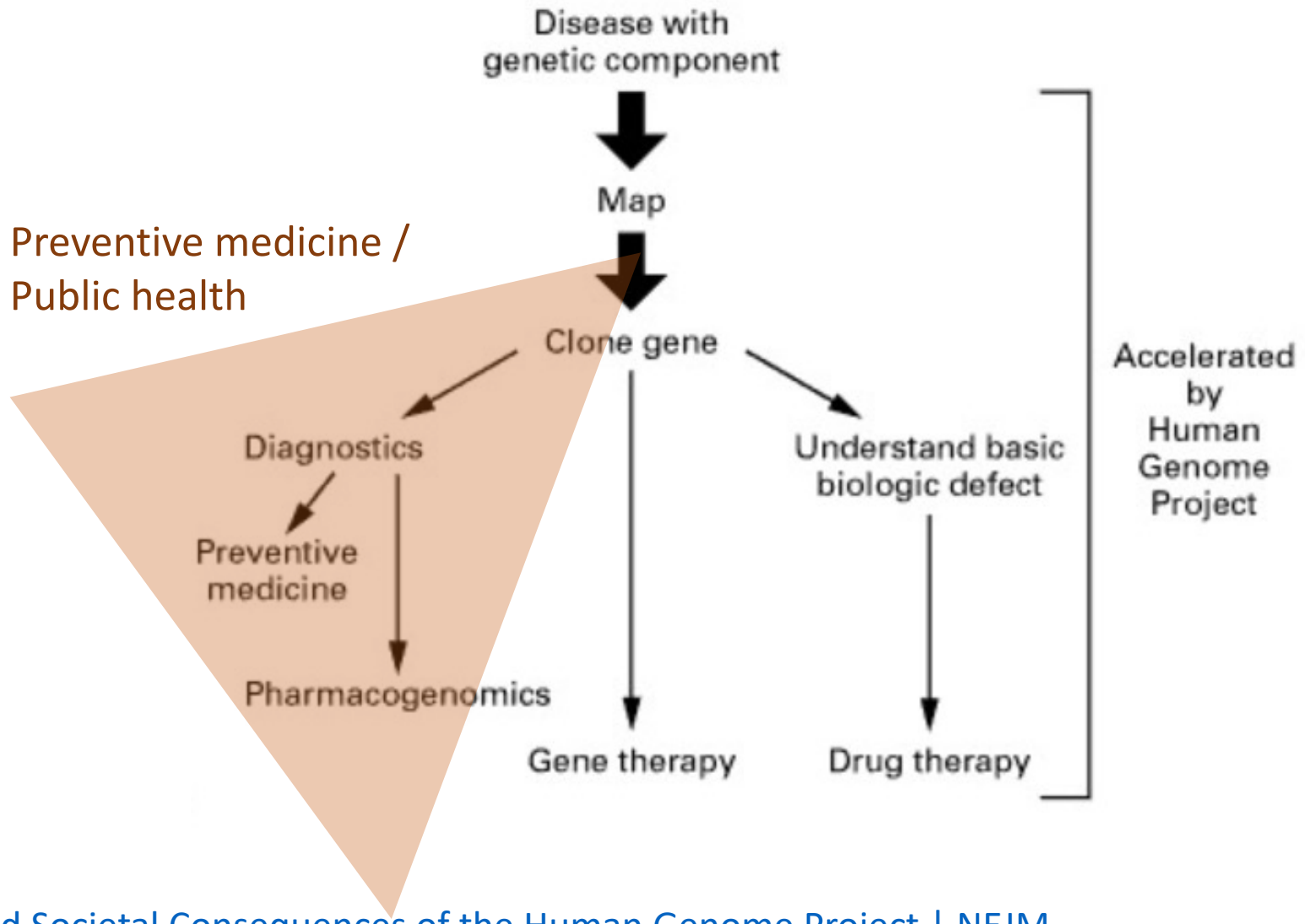
- Cystic fibrosis is the most common single-gene disorder in populations of European descent
- The most common (70-90%) causative mutation was discovered in 1989: *CFTR* $\Delta F508$ at 7q31.2
- Drug therapy targeted to this mutation was approved by FDA in 2019 (30 years in translation!)

Dare to Dream: The Long Road to Targeted Therapies for Cystic Fibrosis.
Collins F, NIH Director’s Blog, Oct 31, 2019

M Gwinn 2023

Medical and Societal Consequences of the Human Genome Project

Francis S. Collins, MD, PhD. N Engl J Med **1999**; 341:28-37



M Gwinn 2023

[Medical and Societal Consequences of the Human Genome Project | NEJM]

Medical and Societal Consequences of the Human Genome Project

Francis S. Collins, MD, PhD. N Engl J Med **1999**; 341:28-

Epidemiology

TABLE 1. RESULTS OF GENETIC TESTING IN A HYPOTHETICAL PATIENT IN 2010.

CONDITION	Candidate genes	RELATIVE RISK	LIFETIME RISK (%)
Reduced risk			
Prostate cancer	*HPC1, HPC2, HPC3*	0.4	7
Alzheimer's disease	*APOE, FAD3, XAD*	0.3	10
Elevated risk			
Coronary artery disease	*APOB, CETP*	2.5	70
Colon cancer	*FCC4, APC*	4	23
Lung cancer	*NAT2*	6	40

Cohort studies
Case-control studies

Cohort studies

M Gwinn 2023

Outline

- How genetics and epidemiology evolved in parallel
- Why group-level data are required to assess individual risk
- An early vision for genomic medicine
- How genetic association studies got so large

Why are association studies so large?

Per-allele relative risks are much smaller

Many more genes are involved

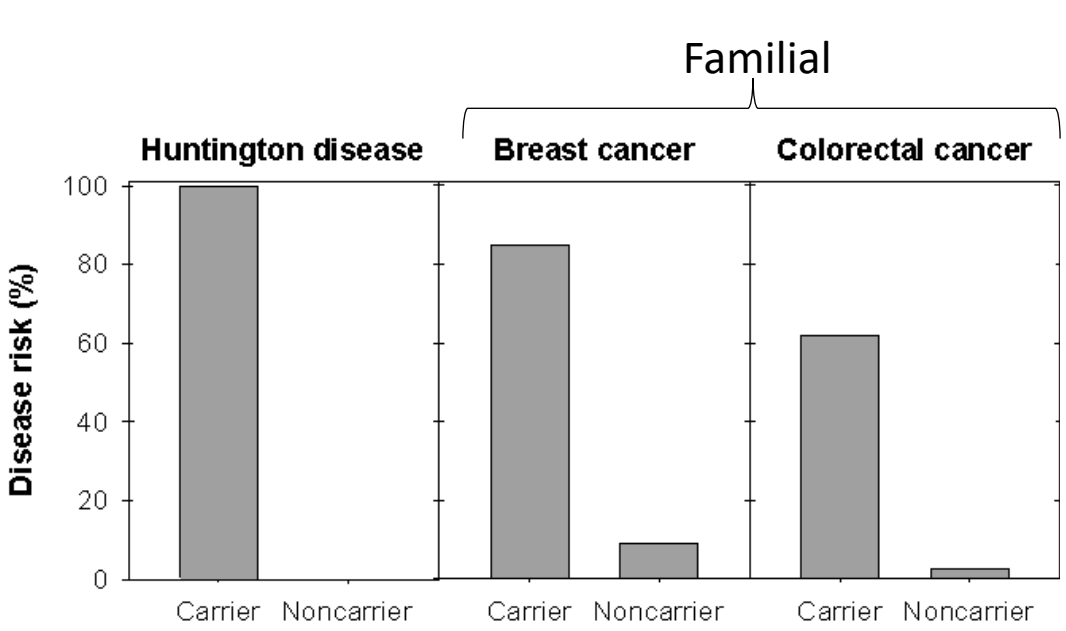
Condition	Genes Involved*	Relative Risk	Lifetime Risk (%)
Reduced risk			
Prostate cancer	*HPC1, HPC2, HPC3*	0.4	7
Alzheimer's disease	*APOE, FAD3, XAD*	0.3	10
Elevated risk			
Coronary artery disease	*APOB, CETP*	2.5	70
Colon cancer	*FCC4, APC*	4	23
Lung cancer	*NAT2*	6	40

Medical and Societal Consequences of the Human Genome Project

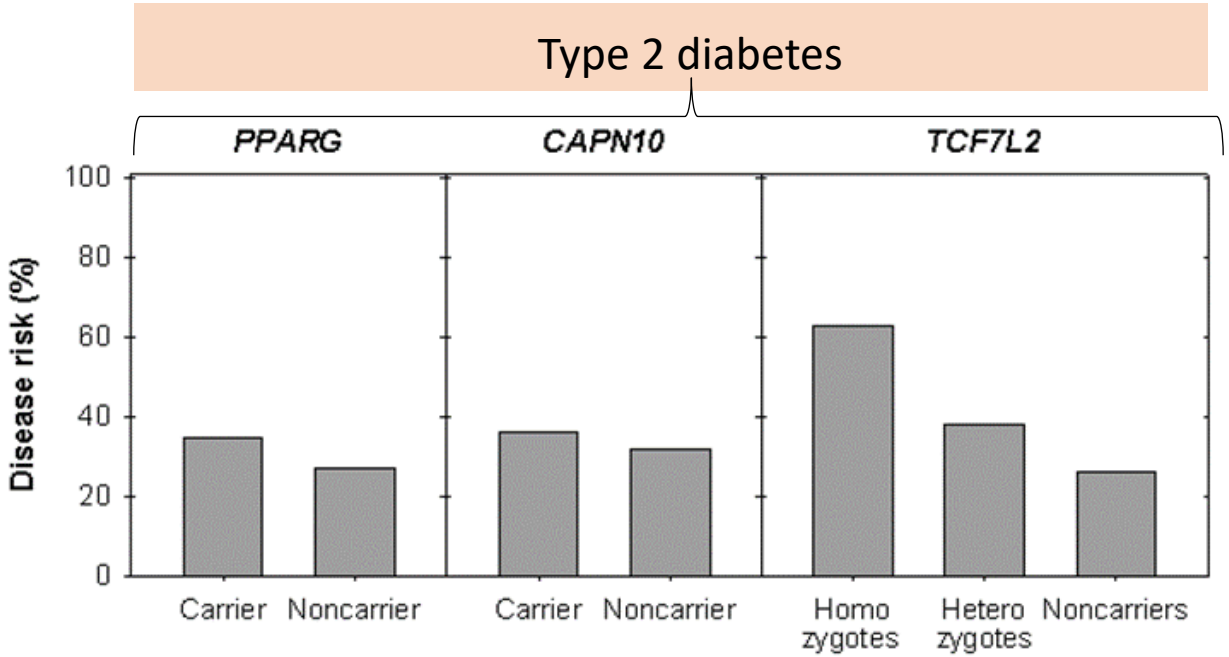
Collins. N Engl J Med 1999; 341:28-37

M Gwinn 2023

Rare vs. common genetic causes of complex diseases



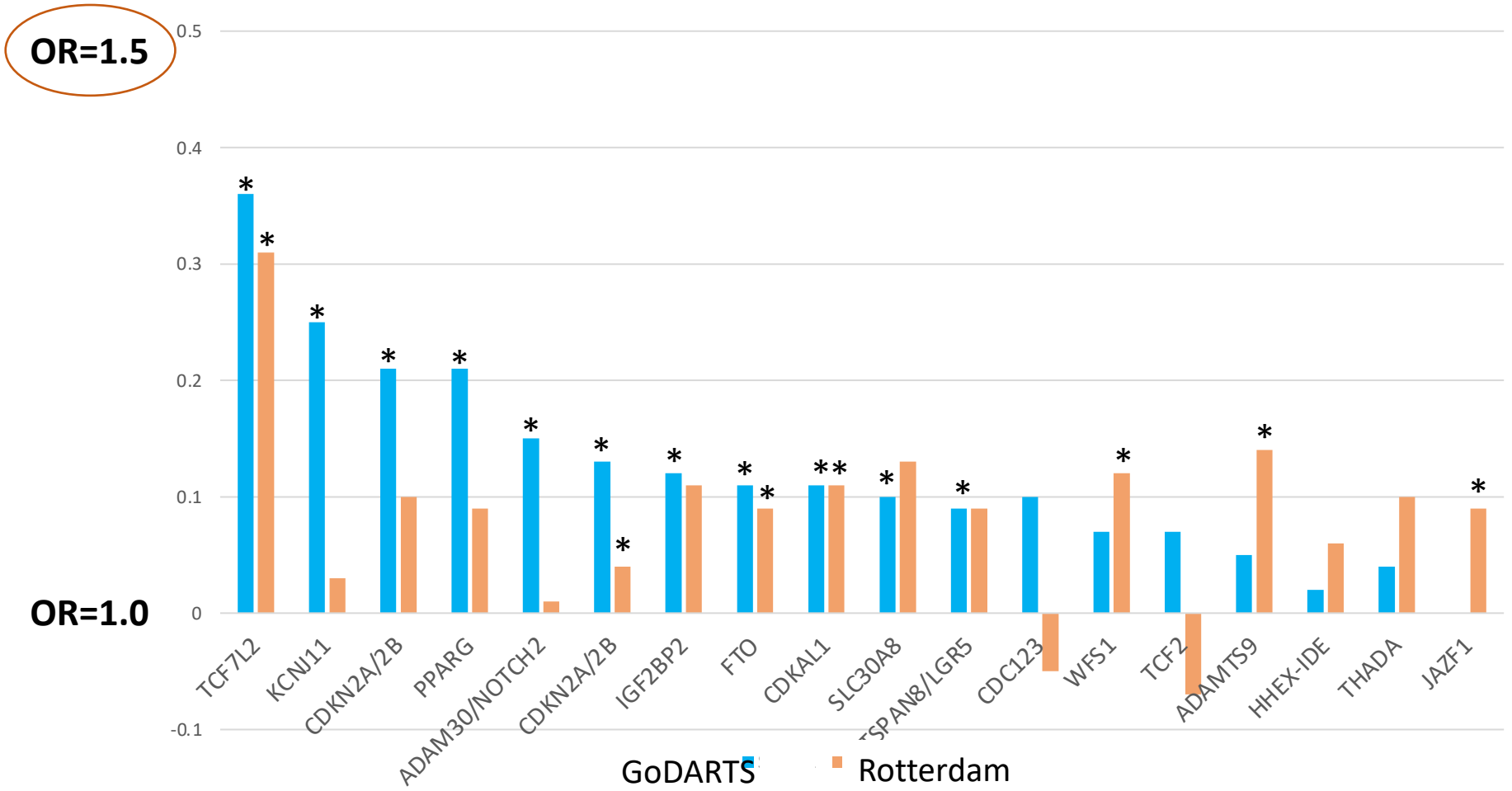
- Rare mutations
- Autosomal dominant inheritance
- *Low population risk*
- Very high relative risk



- Common variants
- Family history?
- *High population risk*
- Low relative risk

After Janssens 2008

Type 2 diabetes: Common variants with small effect sizes



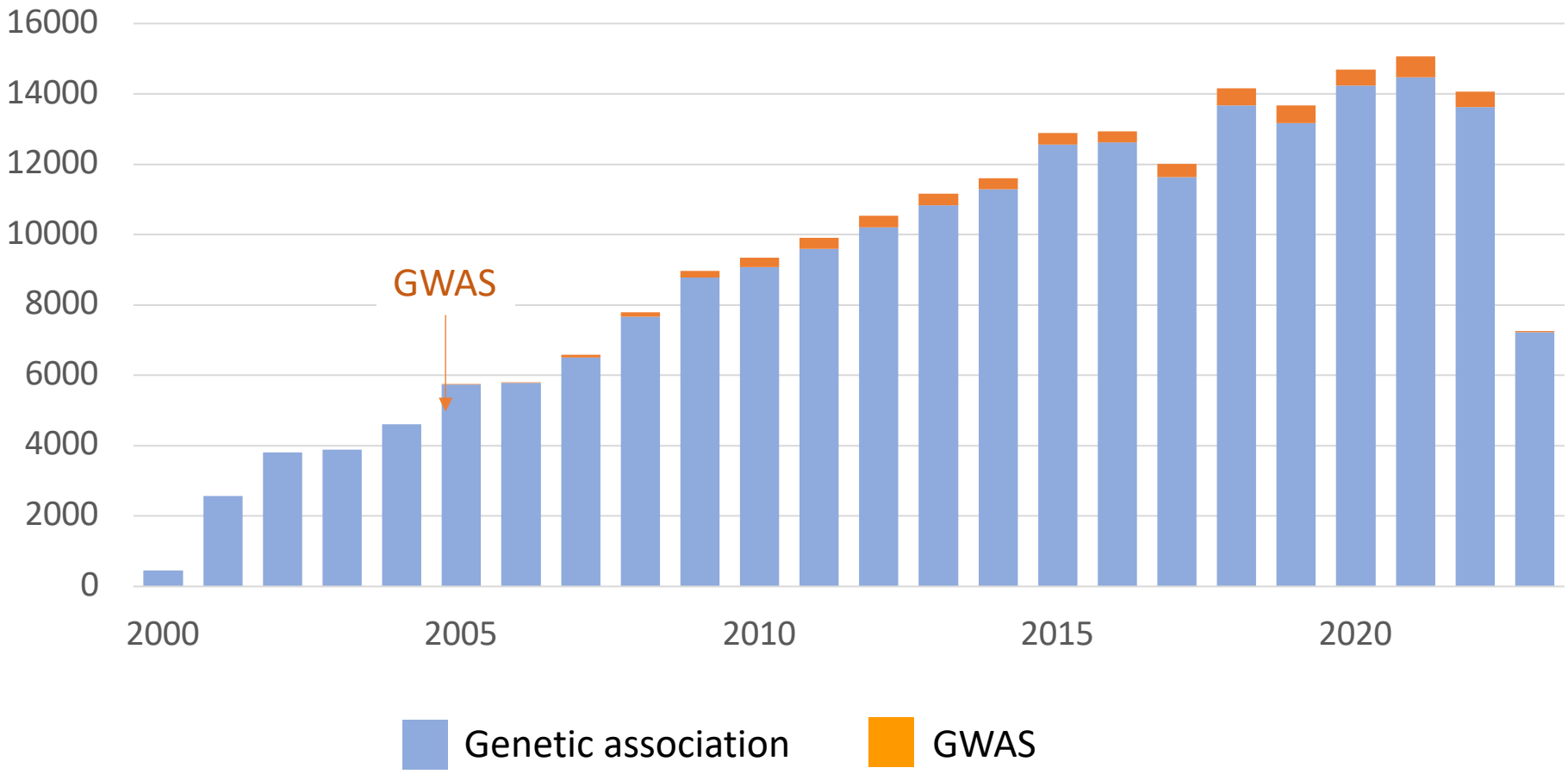
Janssens, van Duijn. Genome Medicine, 2008.

M Gwinn 2023

Forward to the present: Genetic association studies

Human Genome Epidemiology Literature Finder*

Publications in PubMed



*As of August 2023

M Gwinn 2023

“Population”-based genomic epidemiology

UK Biobank

In-depth genetic and health information from half a million volunteer UK participants
Established 2006

All of Us

In-depth genetic and health information from one million volunteer US participants
Established 2015 as the Precision Medicine Initiative Cohort Program

Many, many research consortia

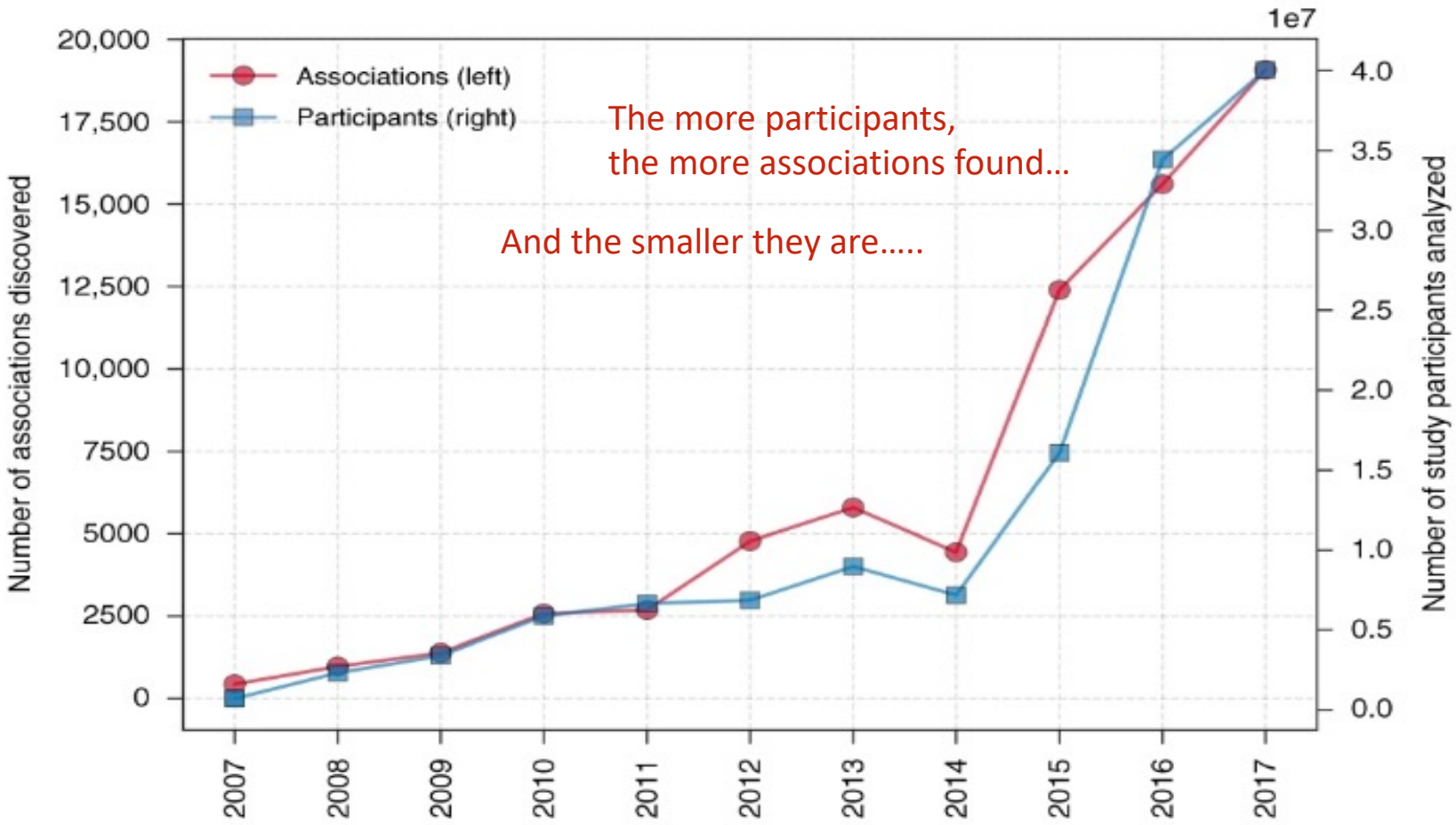
Regional, national, international; population, disease, or exposure-based; hundreds to millions of participants

23andme

Personal genomics and biotechnology company with genotype data for 5 million people
Established 2007

M Gwinn 2023

Increasingly large genome-wide association studies



Number of participants,
in millions

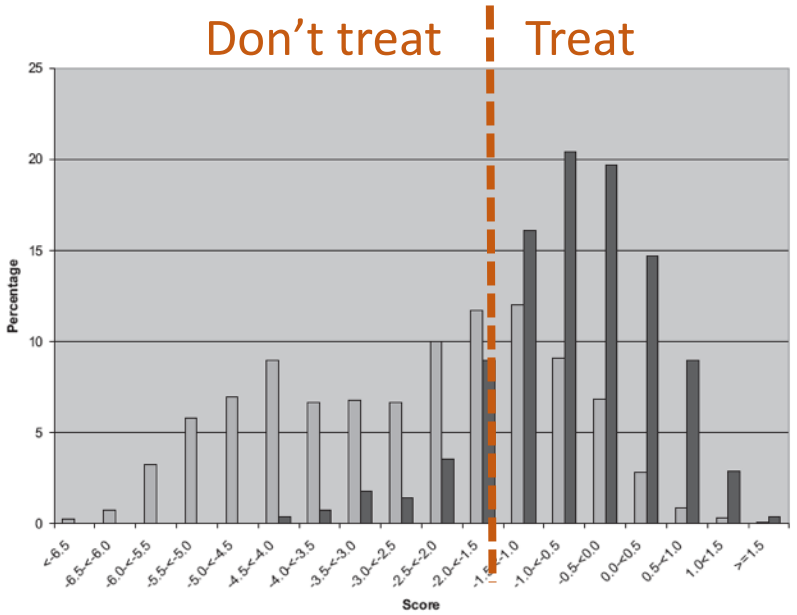
[A scientometric review of genome-wide association studies | Communications Biology (nature.com](#) Mills, Rahal, 2019.

M Gwinn 2023

Polygenic risk score (PRS) based on validated SNP associations

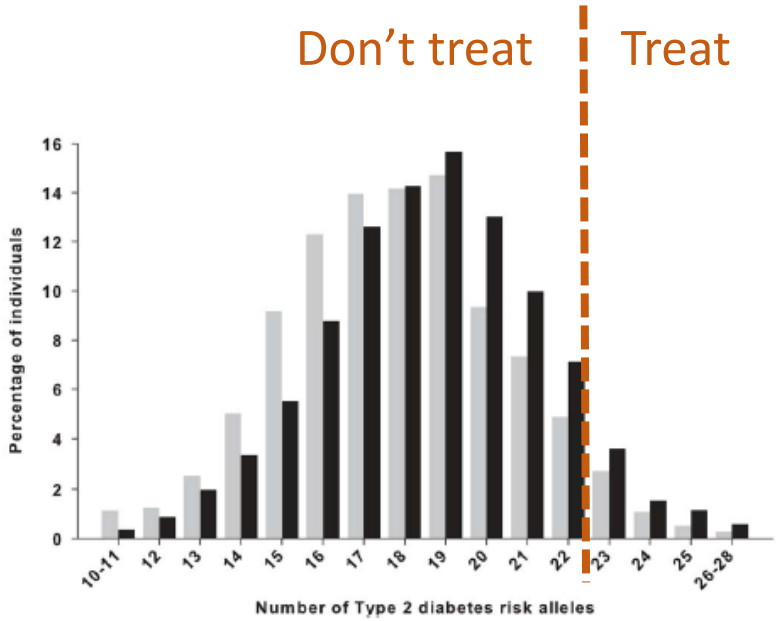
Higher genetic score → higher disease risk

Age-related macular degeneration



Seddon et al. *IOVS* 2009

Type 2 diabetes



Lango et al *Diabetes* 2008

Disease

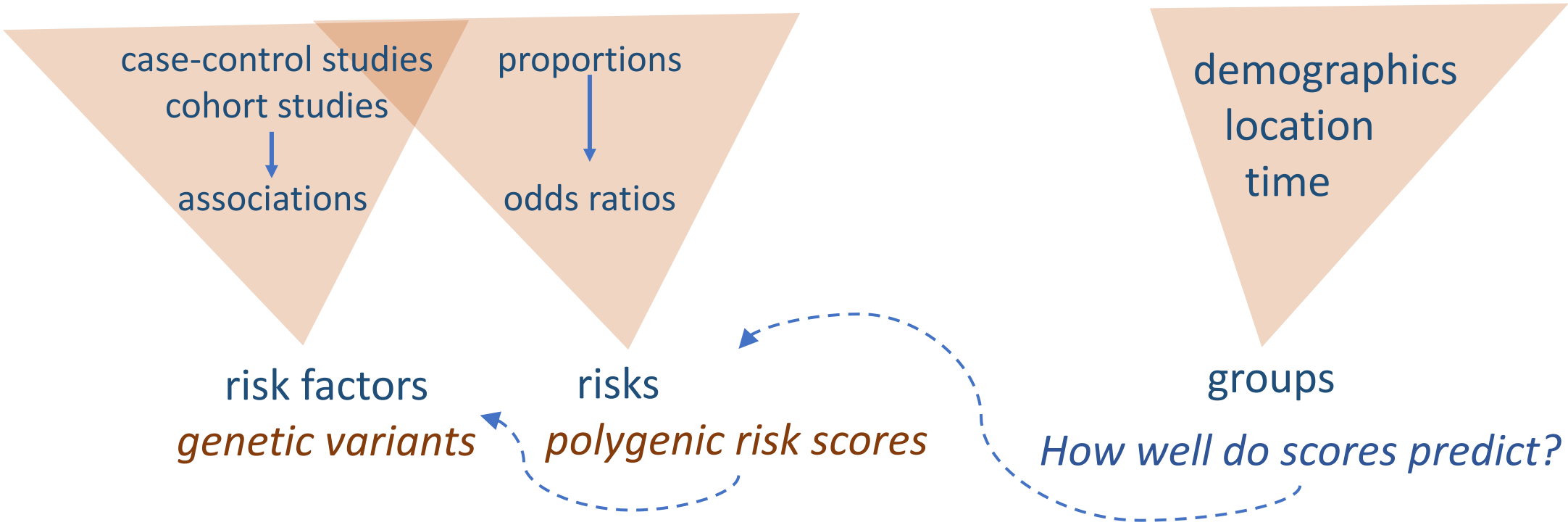
No disease

Janssens 2022

M Gwinn 2023

Epidemiology:

The *determinants* and *distribution* of health and disease in *defined populations*.

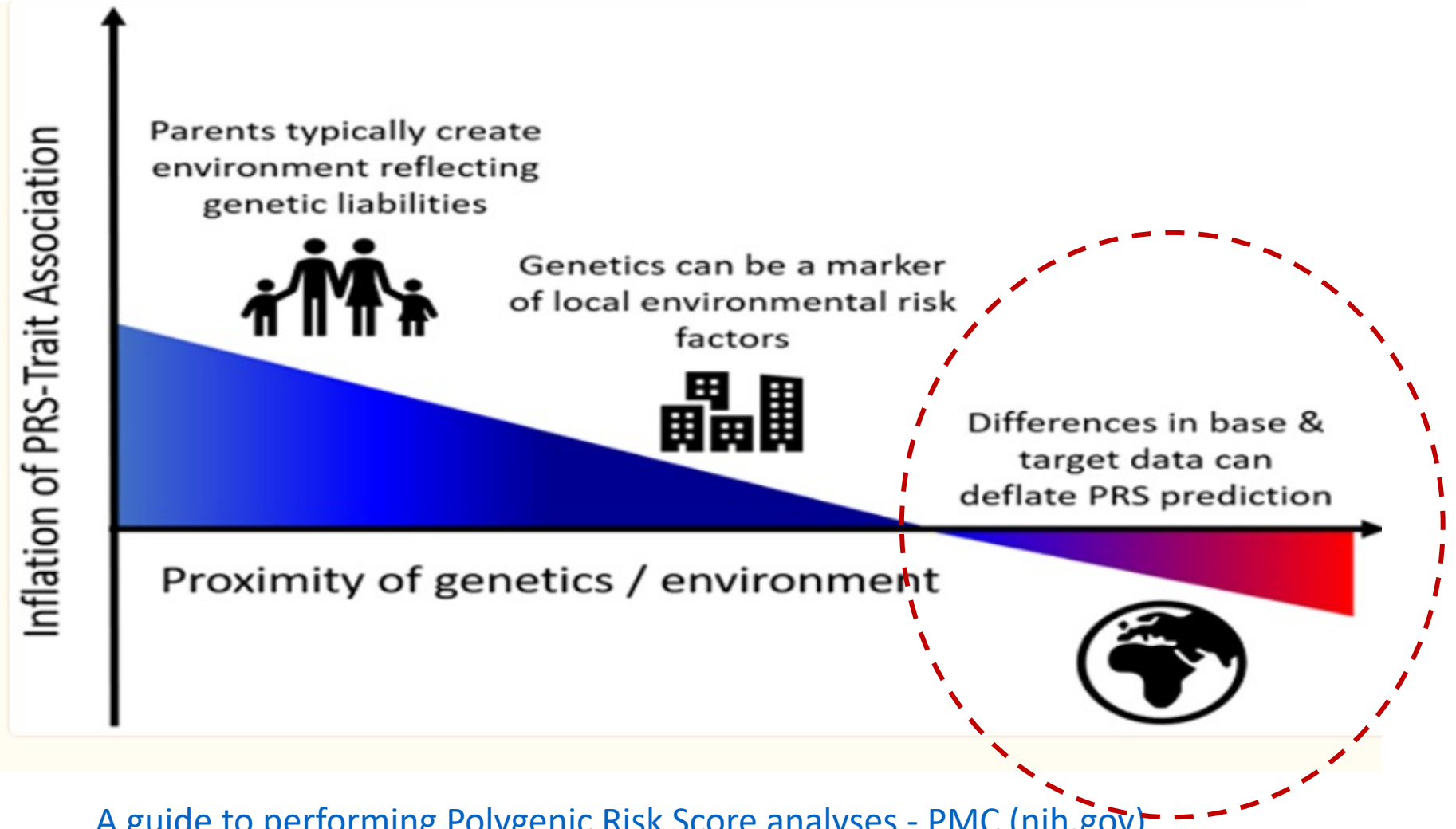


M Gwinn 2023

Outline

- How genetics and epidemiology evolved in parallel
- Why group-level data are required to assess individual risk
- An early vision for genomic medicine
- How genetic association studies got so large
- When polygenic risk scores are biased

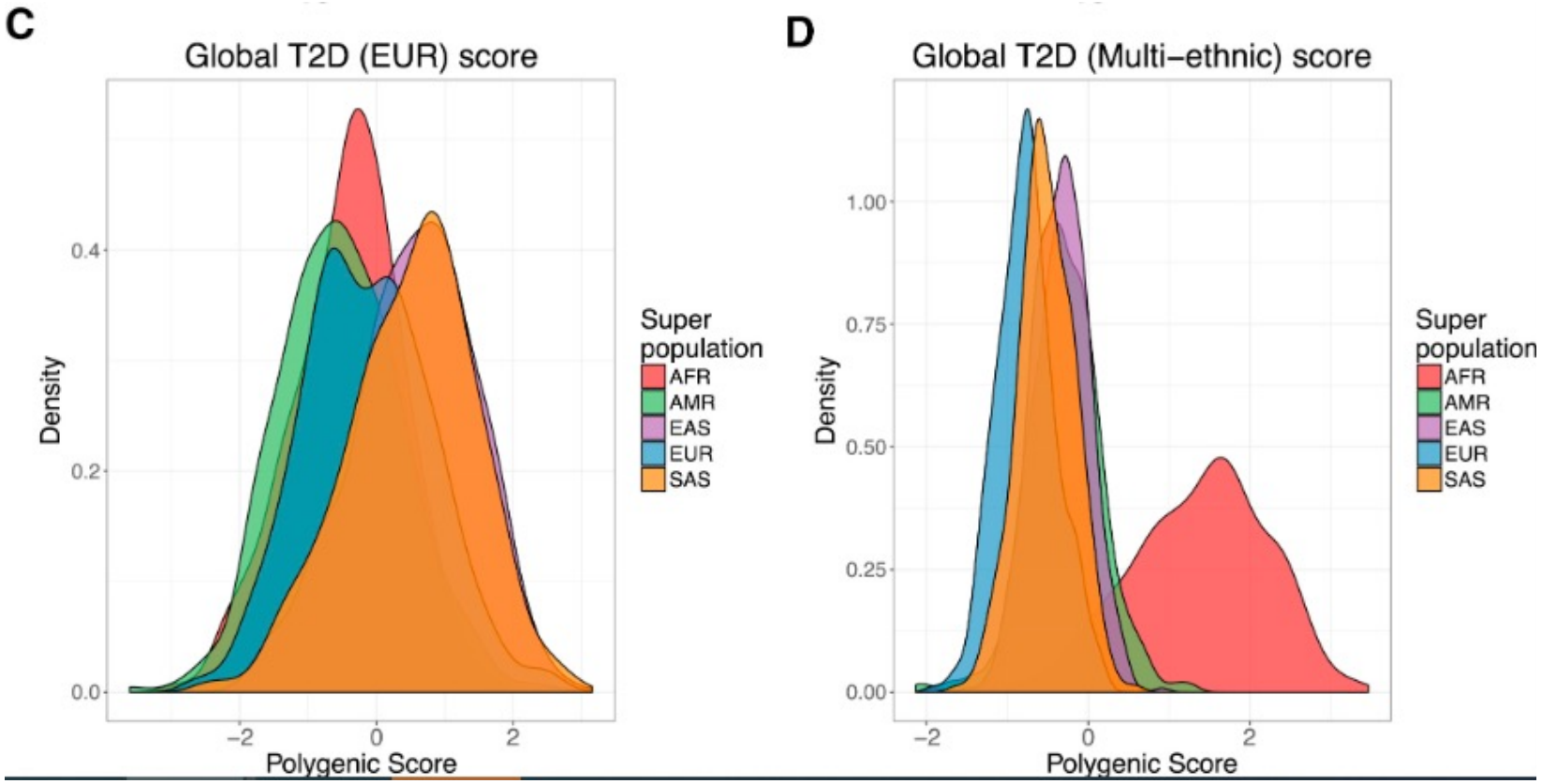
Genetic and environmental correlations can bias PRS associations



A guide to performing Polygenic Risk Score analyses - PMC (nih.gov)
Choi SW, et al. Nat Protoc. 2020 Sep 1; 15(9): 2759–2772.

M Gwinn 2023

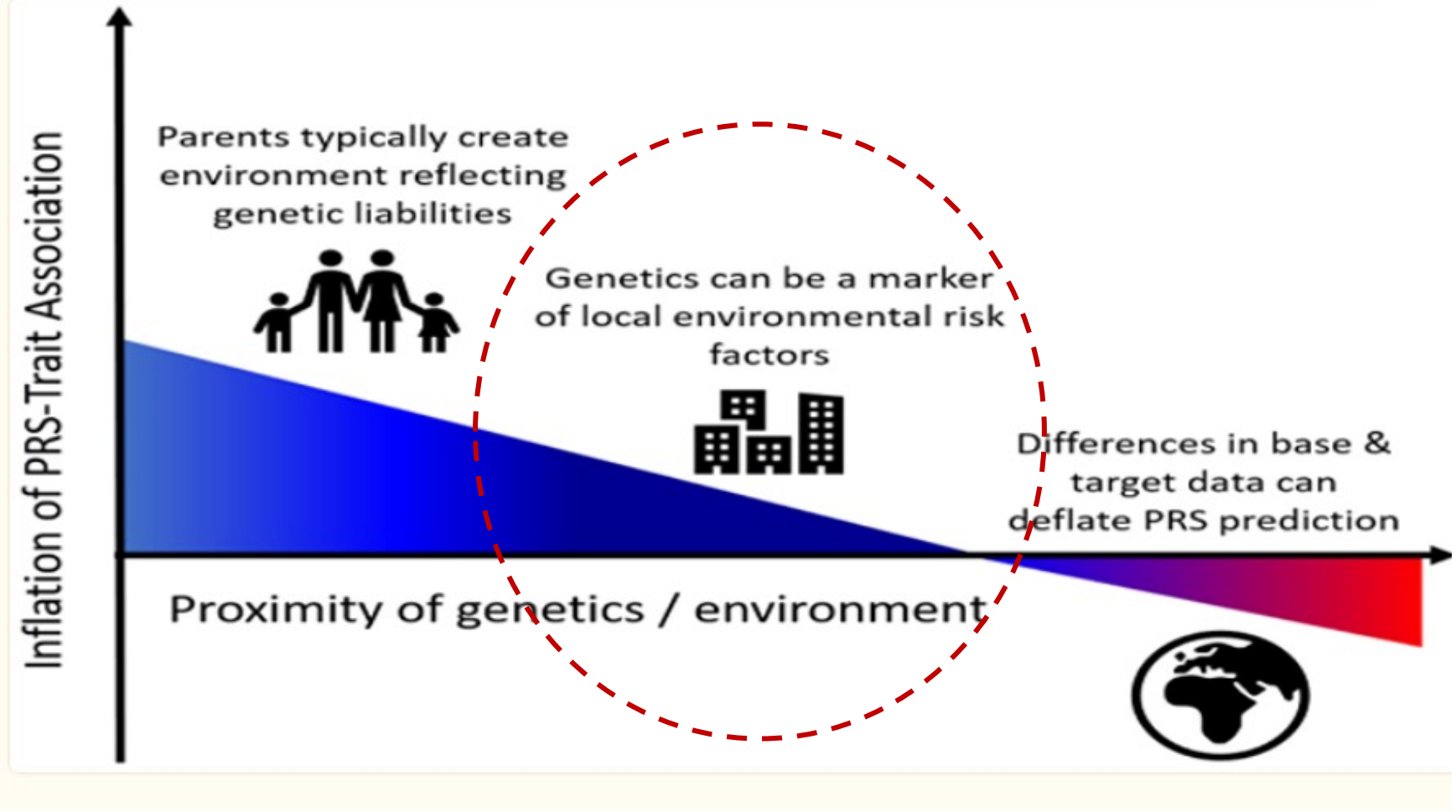
Biased genetic discoveries influence disease risk inferences



Inferred and standardized polygenic risk scores for type 2 diabetes by population, based on summary statistics from European and multi-ethnic studies.

Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations Martin AR, et al. 2017;100:635-649.

Genetic and environmental correlations can bias PRS associations



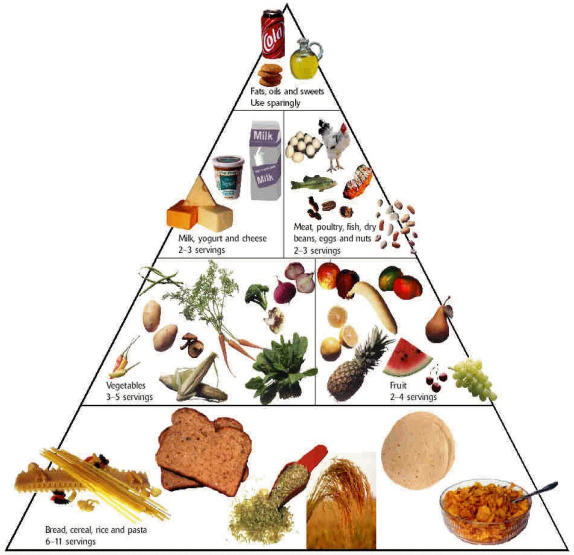
[A guide to performing Polygenic Risk Score analyses - PMC (nih.gov)](#)
Choi SW, et al. Nat Protoc. 2020 Sep 1; 15(9): 2759–2772.

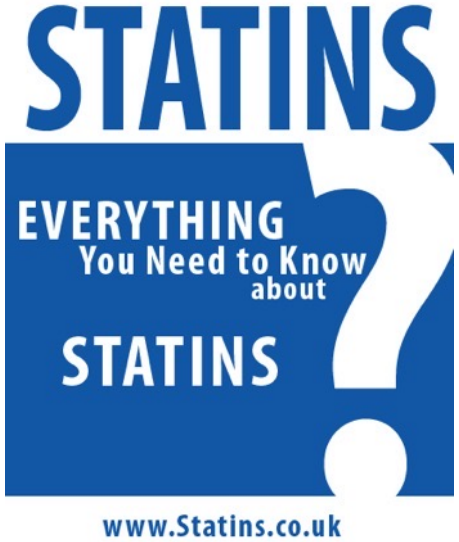
Outline

- How genetics and epidemiology evolved in parallel
- Why group-level data are required to assess individual risk
- An early vision for genomic medicine
- How genetic association studies got so large
- When polygenic risk scores are biased
- How genomics can help evaluate environmental risk factors

environmental

Disease prevention: eliminate or modify risk factors





Eliminate trans fats

Environment

Eat better

Behavior

Take statins

Preventive medicine

M Gwinn 2023

Why study genomics of common diseases with environmental causes?

- Stratify disease risks → Prediction
- Find environmental causes → Prevention
- Understand patterns of disease occurrence → Diagnosis and prognosis

Do we need genomic research for the prevention of common diseases with environmental causes? - PubMed (nih.gov)
Khoury, et al. 2005;161:799–805

M Gwinn 2023

Epidemiology: observational studies

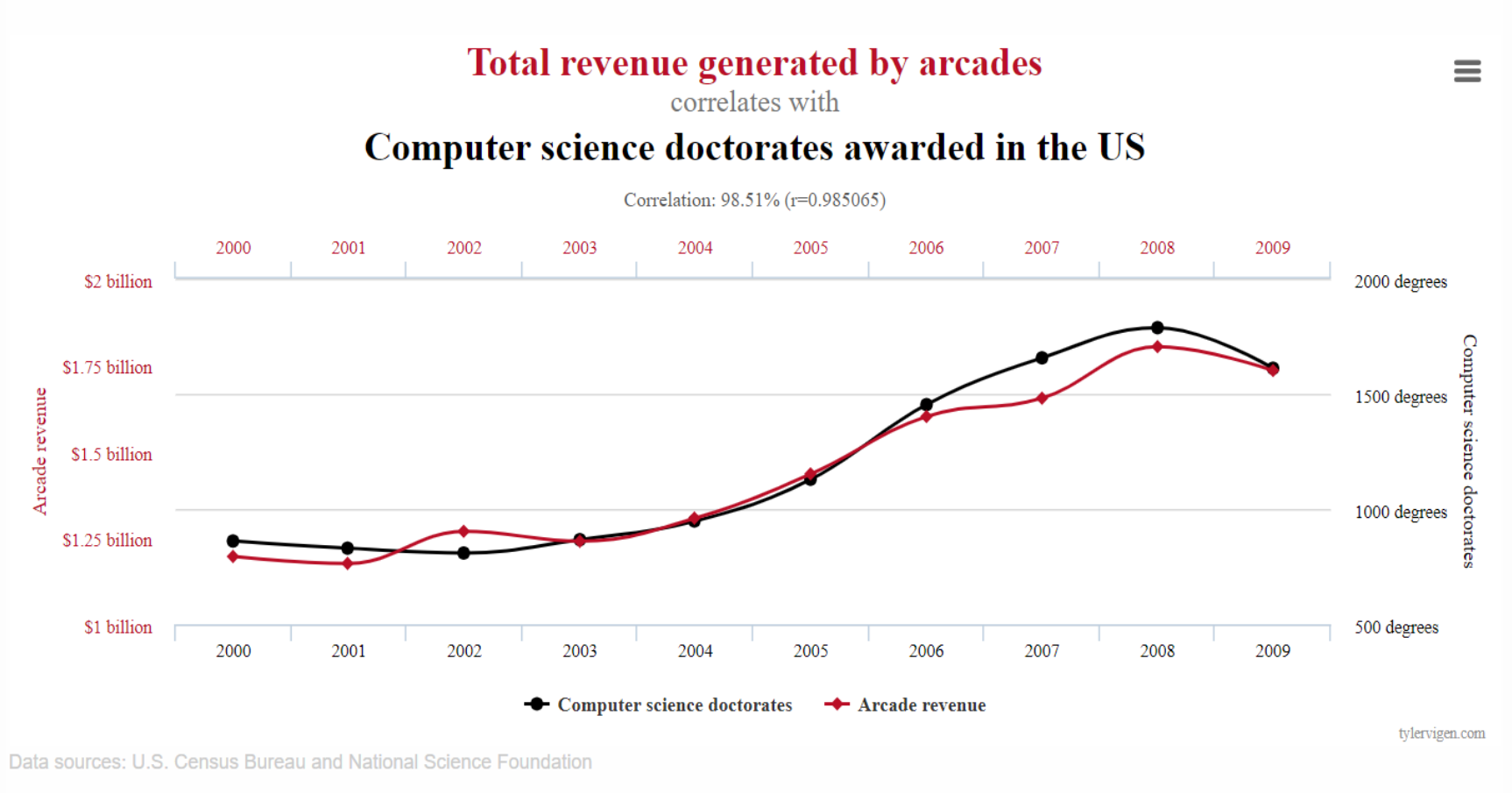
Observational study designs

- Cohort
- Case-control

associations
"correlations"

causes?

Wait a minute! “Correlation does not imply causation”



[Spurious correlations. Tylervigen.com]

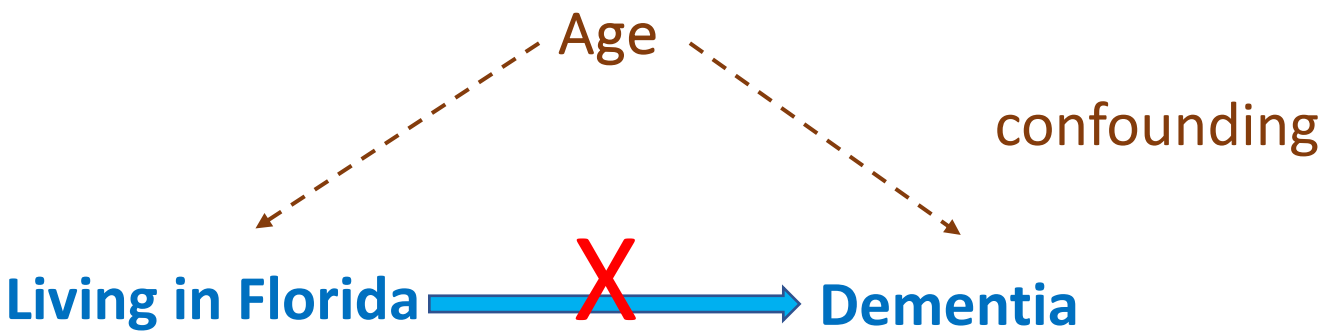
M Gwinn 2023

Epidemiology: observational studies

Define “imply”

- In logic, *“a implies b”* means *“if a, then b”*
- In common discourse, “*implies*” means “*suggests*”

Correlation is **necessary** to infer causation—but **not sufficient** to prove it.



ACSH Explains 'Confounding': Why Correlation Does Not Mean Causation | American Council on Science and Health

Epidemiology: control of confounding

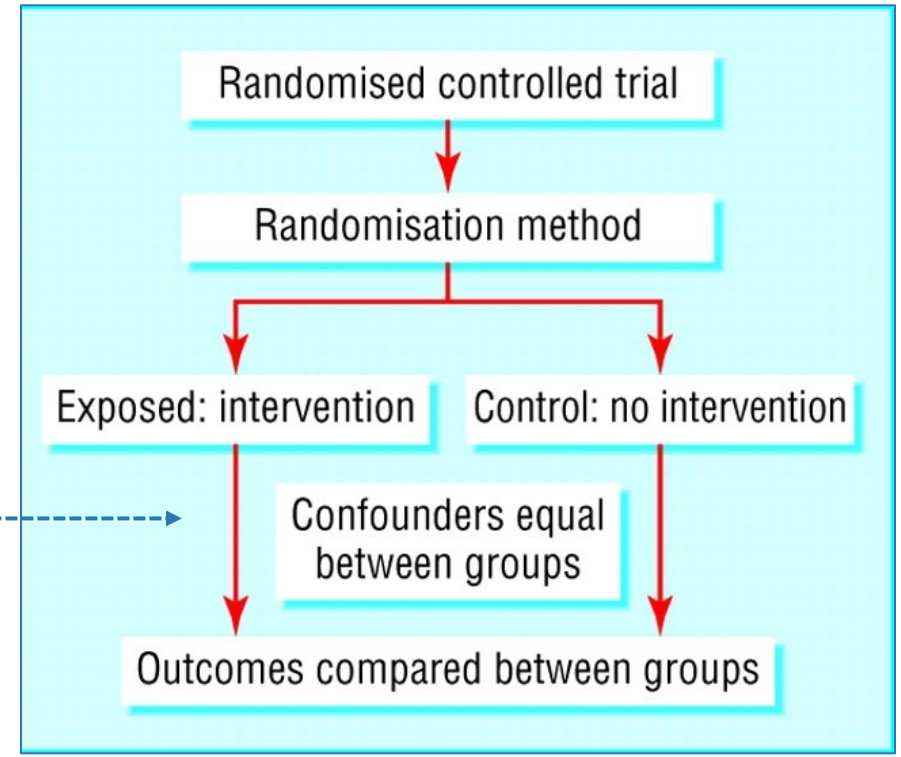
Observational study designs / statistical methods

- Cohort
- Case-control

Experimental study design

- Randomized clinical trial

Control for confounding



M Gwinn 2023

Epidemiology: the problem of confounding

Coronary heart disease (CHD) is less frequent in women taking menopausal hormone replacement therapy (HRT)

- *Observational studies (case-control, cohort)*

Taking HRT *has no effect* on risk of CHD

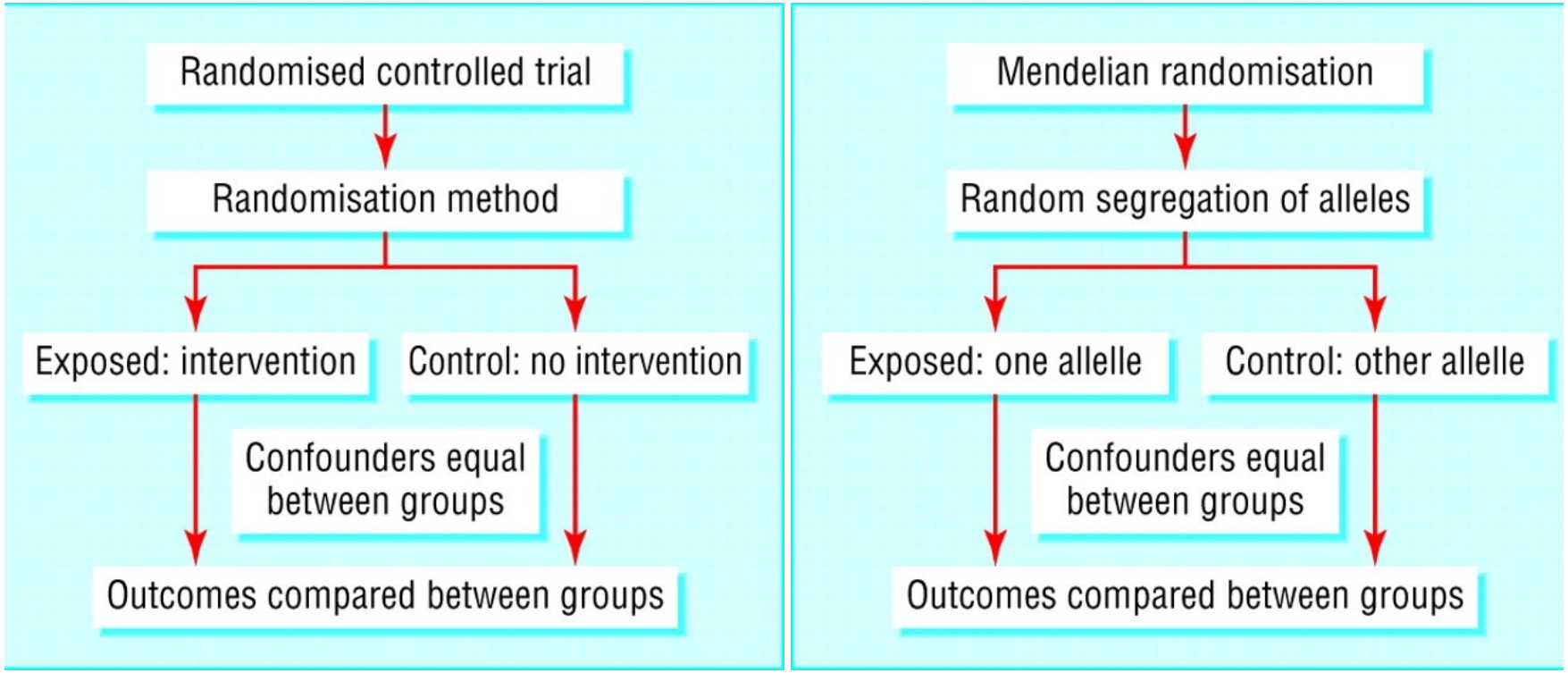
- *Randomized clinical trial*

Women taking HRT are *more likely to develop breast cancer*

- *Observational studies (case-control, cohort)*

Mendelian randomization

- **Strategy**: counter unmeasured confounding in observational studies.
- **Principle:** random allocation of alleles from parents to offspring.

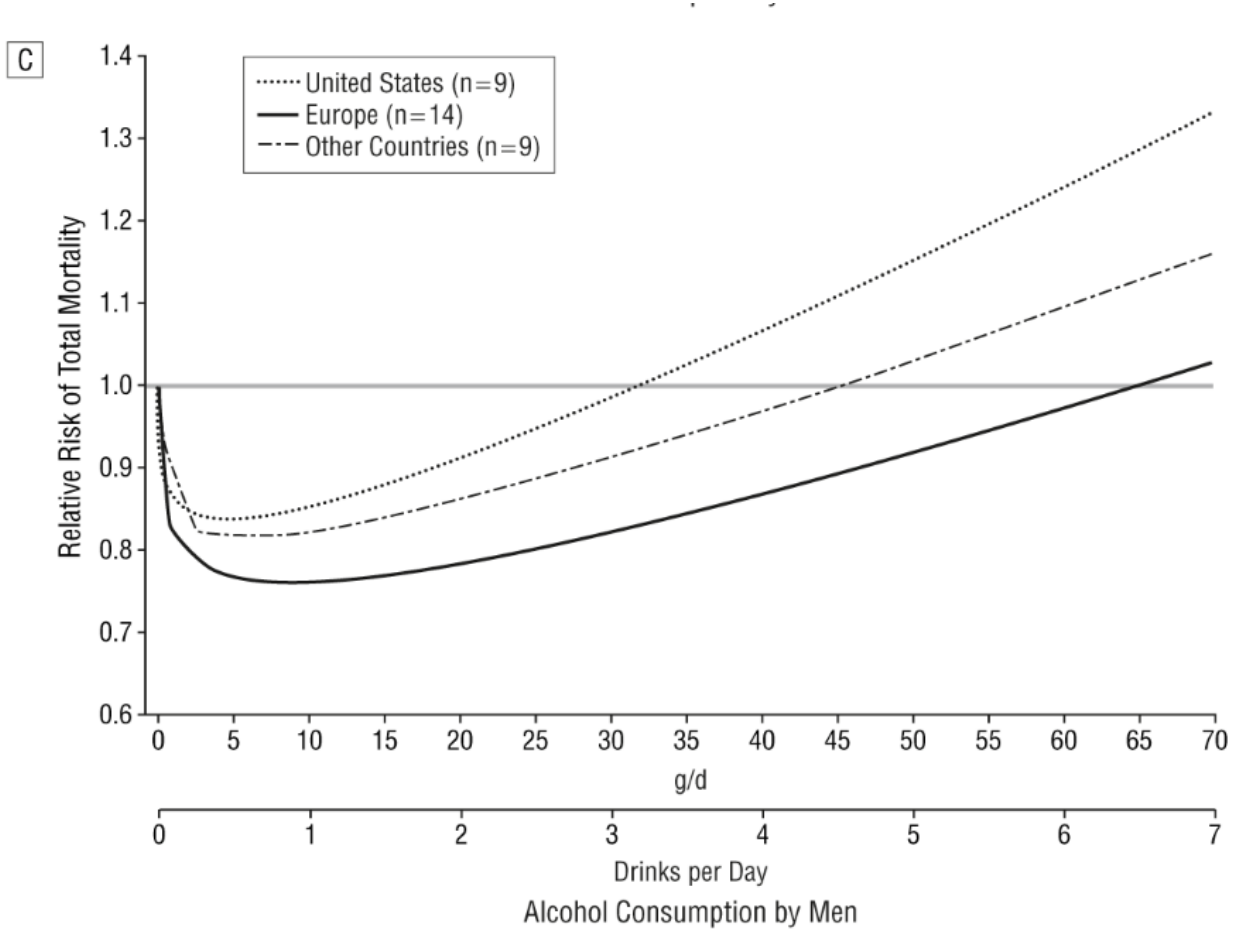


Random inheritance of risk alleles

[Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians | The BMJ](#)
Davies et al. 2018

M Gwinn 2023

The “J-shaped curve” for alcohol consumption and mortality



Alcohol dosing and total mortality in men and women: an updated meta-analysis of 34 prospective studies - PubMed (nih.gov)

DiCastelnuovo, et al. Ann Int Med, 2006.

M Gwinn 2023

Mendelian randomization

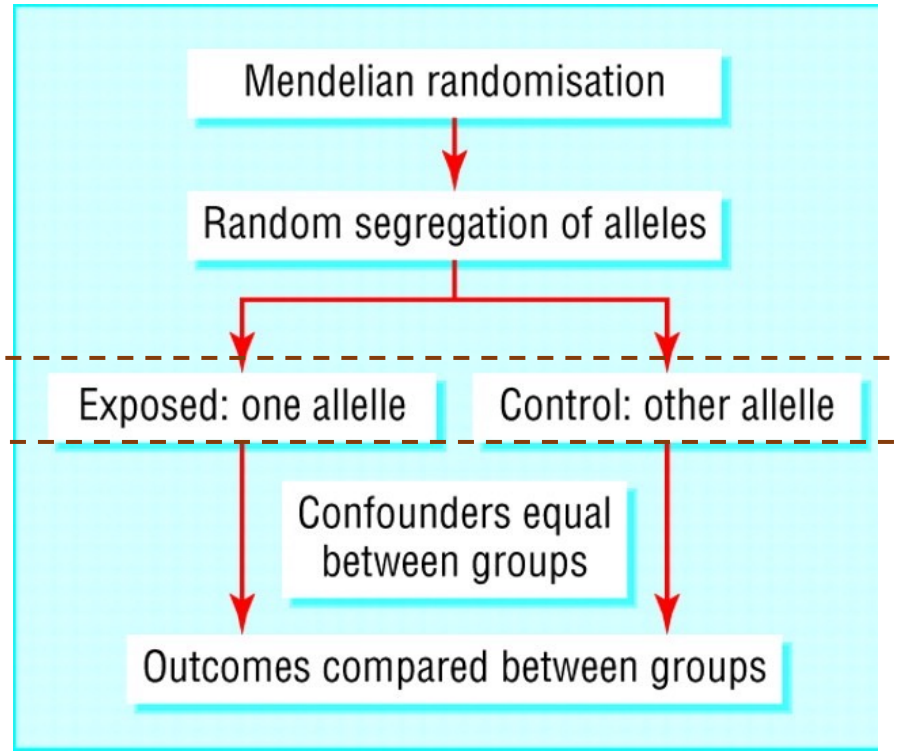
Environmental exposure:
Alcohol consumption

Genetic proxy for exposure

ALDH2 (alcohol dehydrogenase 2)

*1*1 *1*2 *2*2

low alcohol consumption

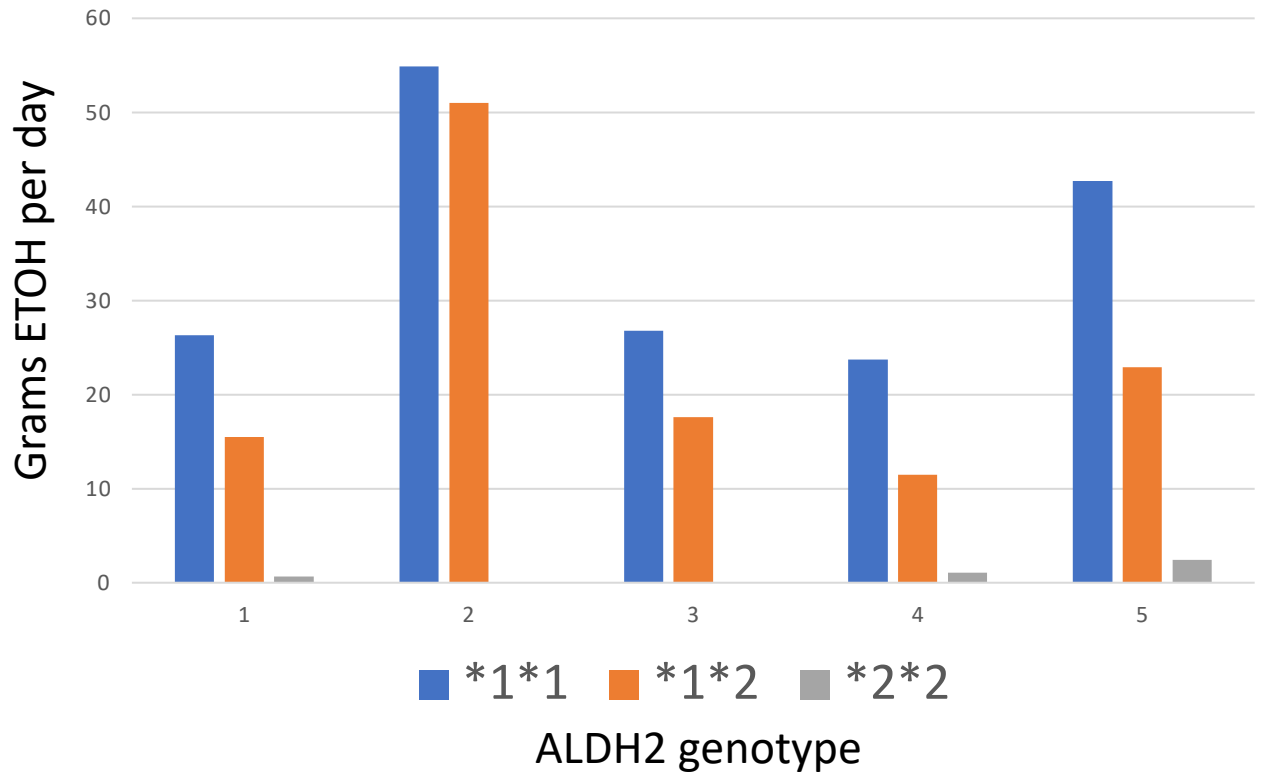


Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians | The BMJ
Davies et al. 2018

M Gwinn 2023

Mendelian randomization: example

Mean alcohol consumption by ALDH2 genotype



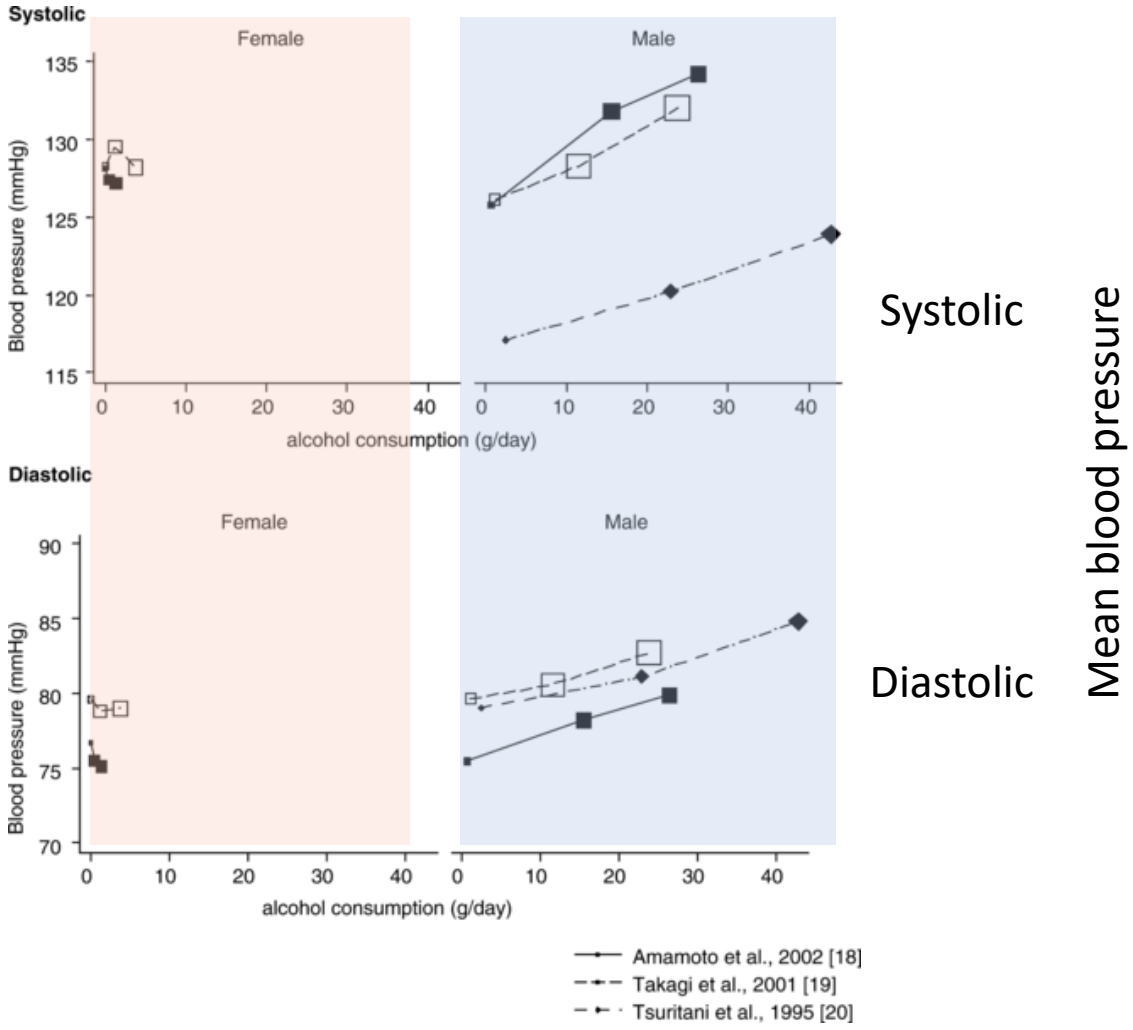
Five studies that reported alcohol consumption as a continuous variable, males only

Chen L, Davey Smith G, Harbord RM, Lewis SJ (2008) Alcohol Intake and Blood Pressure: A Systematic Review Implementing a Mendelian Randomization Approach. PLOS Medicine 5(3): e52. https://doi.org/10.1371/journal.pmed.0050052
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050052

M Gwinn 2023

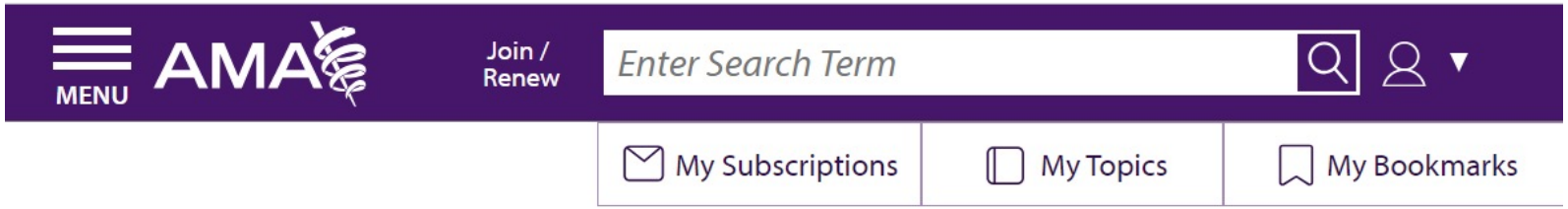
Is moderate alcohol consumption good for your blood pressure?

No



Chen L, Davey Smith G, Harbord RM, Lewis SJ (2008) Alcohol Intake and Blood Pressure: A Systematic Review Implementing a Mendelian Randomization Approach. PLOS Medicine https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050052

M Gwinn 2023



What doctors wish patients knew about precision medicine

JUL 21, 2023 • 11 MIN READ

By Sara Berg, MS, Senior News Writer

"While we used to call it **personalized medicine**, we're actually starting to move away from that terminology because it gives the implication that the therapy is really uniquely tailored for that person, which is probably a bit of an overstatement," he said. Instead, "we are identifying genetic differences to **help us tailor the therapy to the entire group** with the same genetic differences, and we would treat them differently as we would another group with a different set of genetics."

*Jordan Laser, MD, chair of the Personalized Medicine Committee for the College of American Pathologists

What doctors wish patients knew about precision medicine | American Medical Association (ama-assn.org)

M Gwinn 2023

What I hope everyone will remember about precision medicine

- The goal of medicine is to protect, preserve, and prolong health
- The determinants of health and disease are complex and dynamic
- No matter how many or precise our measurements, deterministic models are unrealistic
- Critical thinking—including formal systems of probability and statistics—remains crucial in the age of “big data”

M Gwinn 2023

ORISE Enrichment Event
September 7-8, 2023

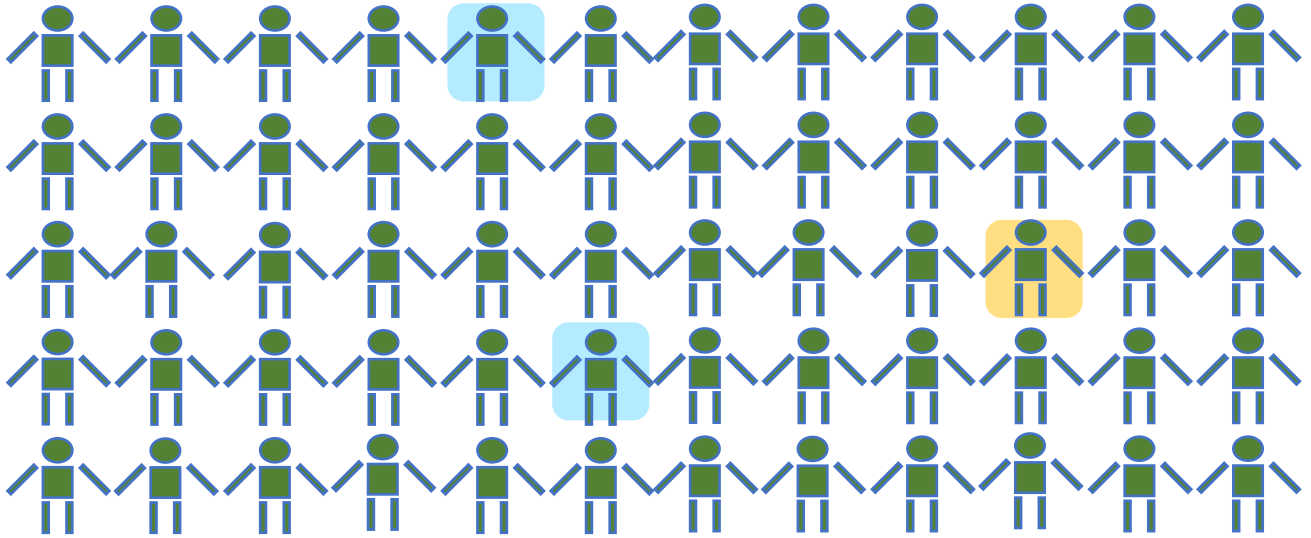
Marta Gwinn, MD, MPH

Department of Epidemiology

Rollins School of Public Health

Emory University

mgwinn2@emory.edu



Learning objectives

1. Name two technologies that have made large-scale, population-based genomic epidemiology studies possible.

   High-throughput genotyping, high-performance computing

2. Offer two reasons why population-based data are needed to assess individual risk of common diseases.

   The causes of common diseases are too dynamic and complex, causal processes are subject to random variation.

3. Identify the main reason why genome-wide association studies are conducted using very large study populations.

   They are searching for very small effects.

4. Describe the problem in epidemiologic studies that Mendelian randomization is designed to address.

   By using a genetic proxy for an environmental exposure, MR studies are designed to reduce the risk of confounding, that is, spurious association due to other factors related to both exposure and outcome.

M Gwinn 2023