Genomics and the use of the UK Biobank in health informatics

Jeonghan Hong

Goal and Objectives

Goal of the Lecture

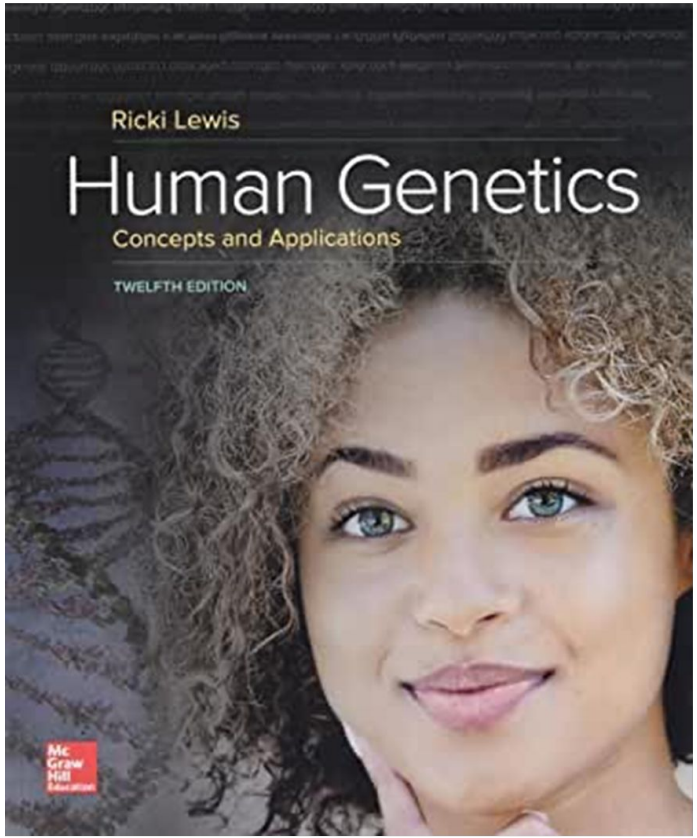
- To provide an overview of genomics and the utilization of the UK Biobank in health informatics, highlighting their importance and potential impact on healthcare and biomedical research.

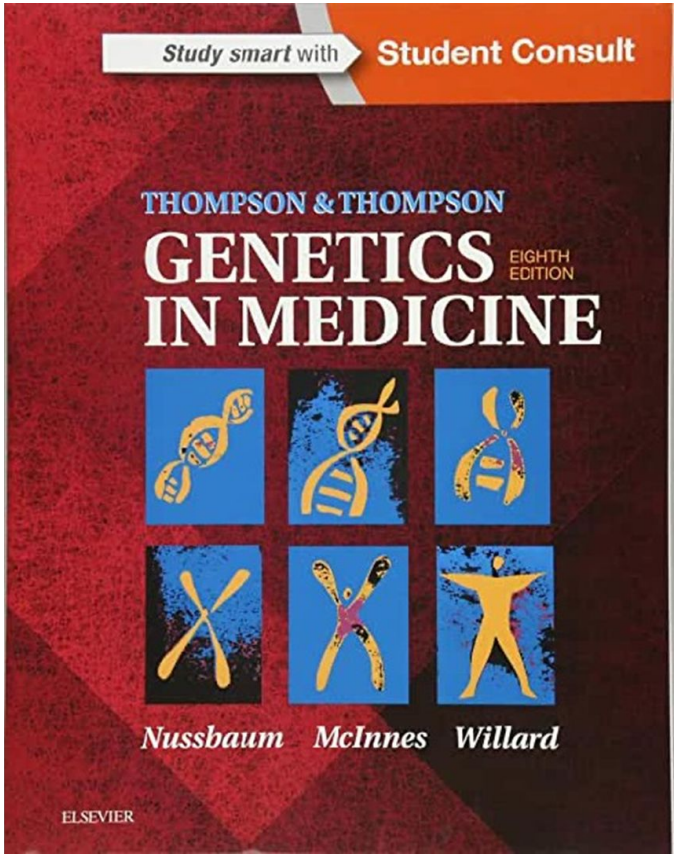
Objectives

- Introduce the fundamental concepts of genomics, including DNA, genes, and genome sequencing.
- Explore how genomics serves as a cornerstone for various research disciplines, including healthcare, biotechnology, and personalized medicine.
- Discuss approaches to accessing genomic data and the importance of data accessibility in research integration.
- Highlight methodologies for identifying and interpreting disease-causing genes in genomics research.
- Provide insights into how genomics can serve as a basis for integrating into one's research interests and contribute to advancements in personalized medicine and disease genetics.

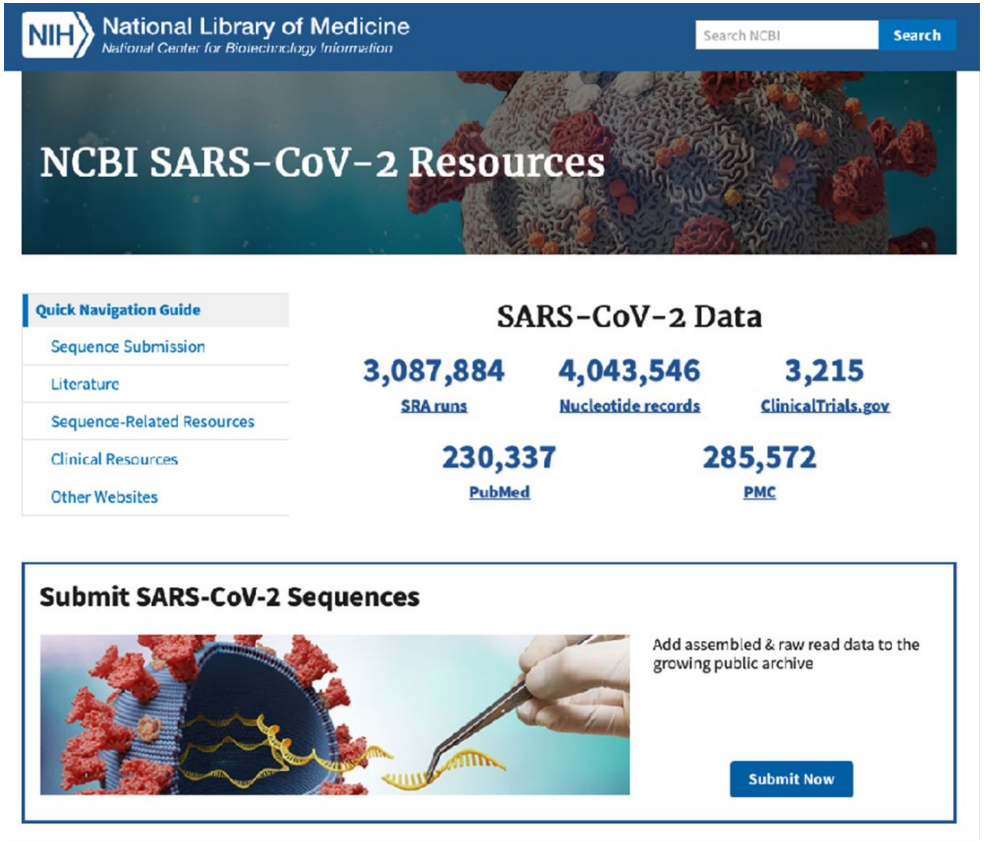
Table of Contents

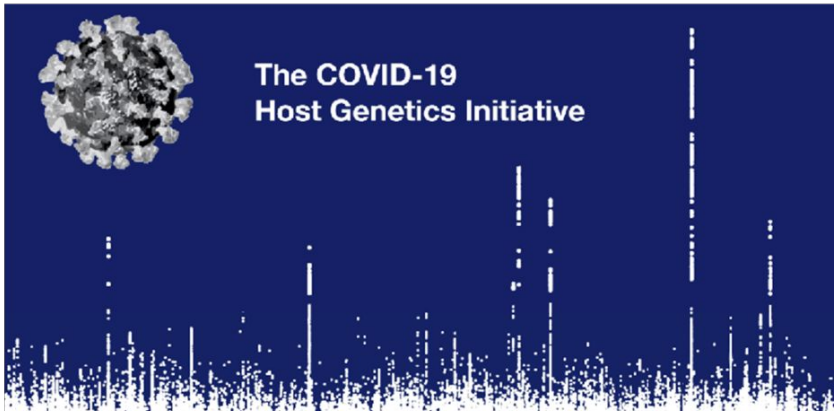
Reference books



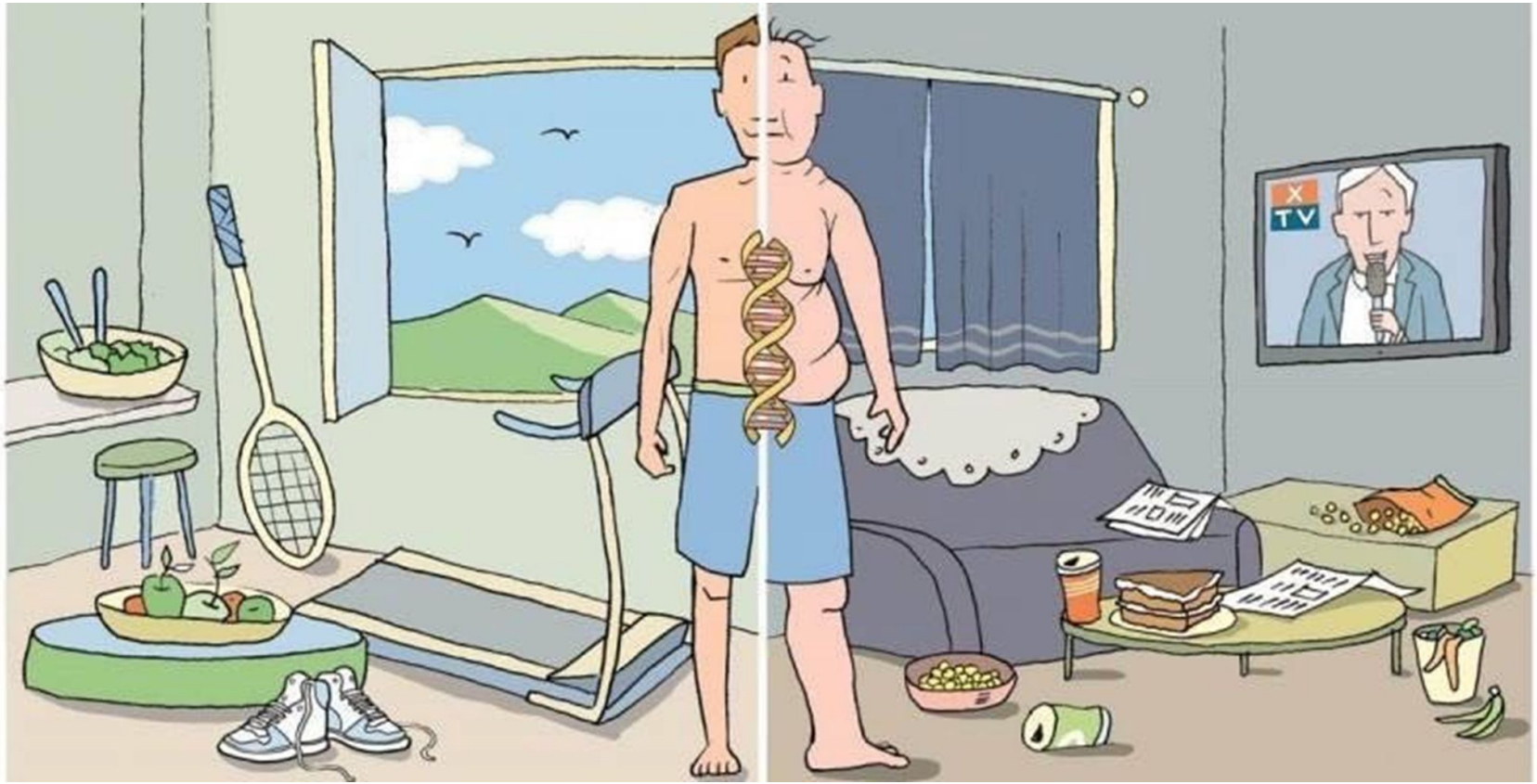


COVID-19 pandemic

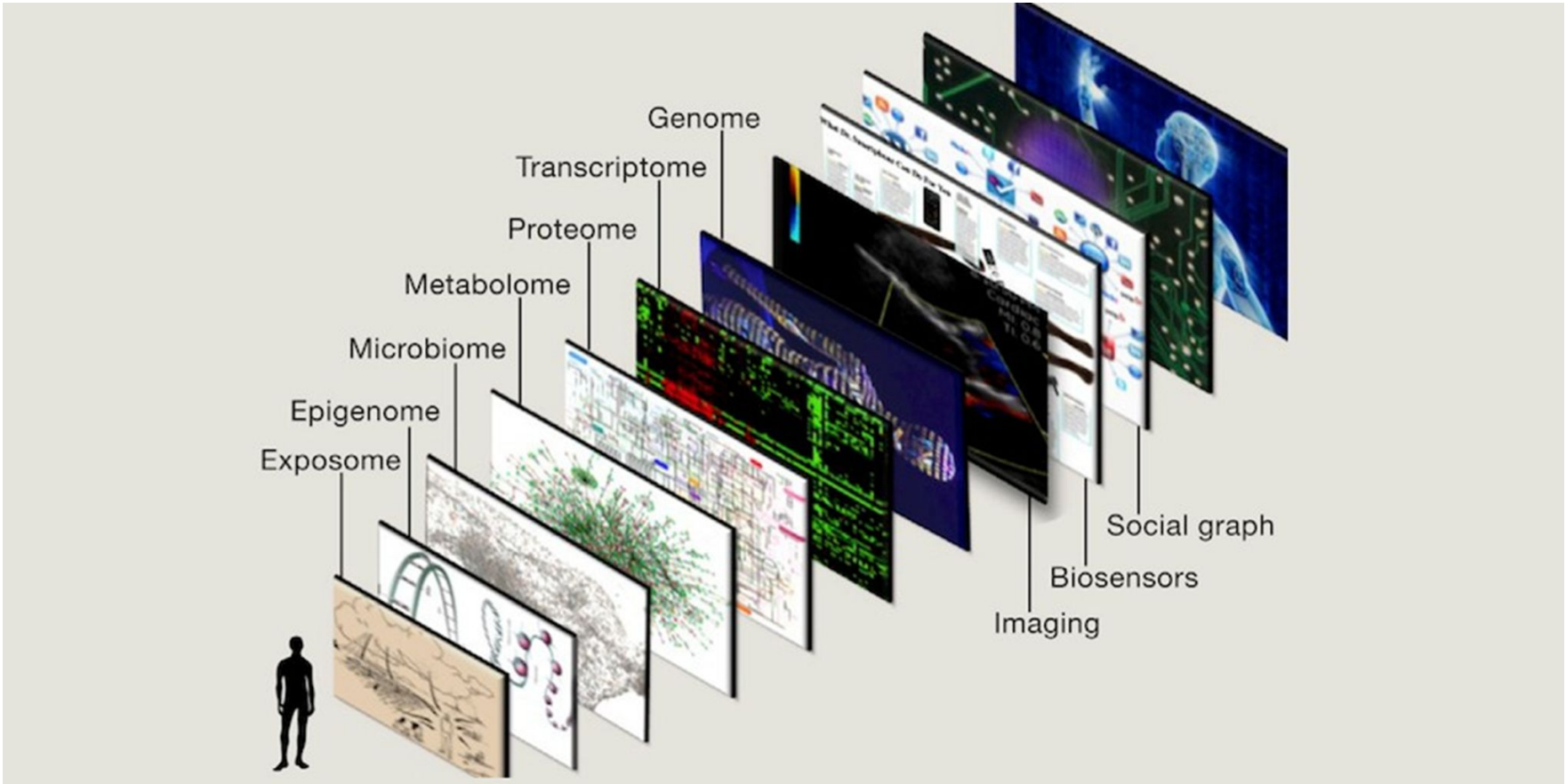




Phenotype (or Disease)
= Gene function + Environmental action



Integration of Data for Precision Medicine



Topol E. Cell 2014

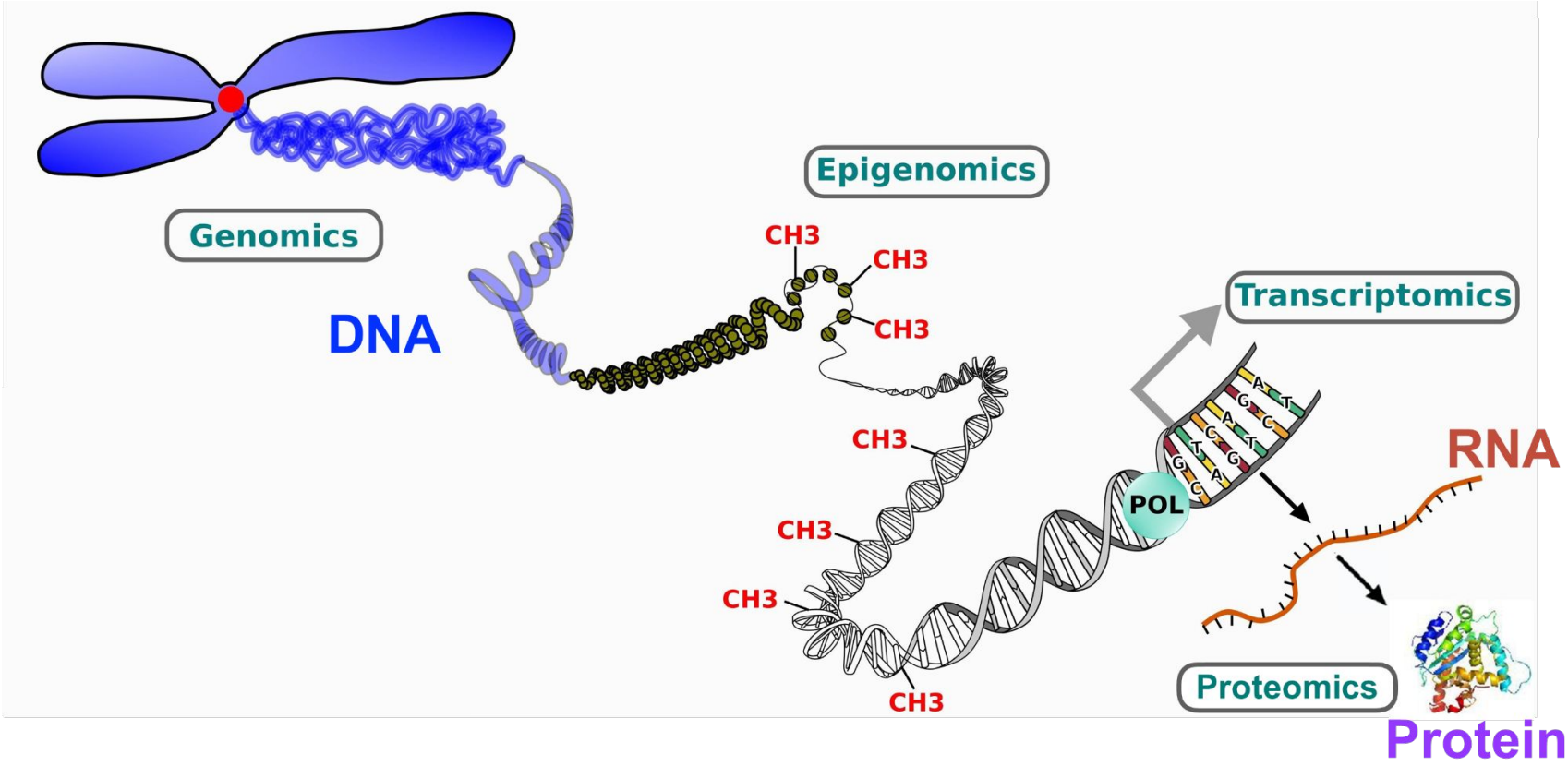
Genetics is the science of the variation of inherited traits

inherited traits and **their variations**.

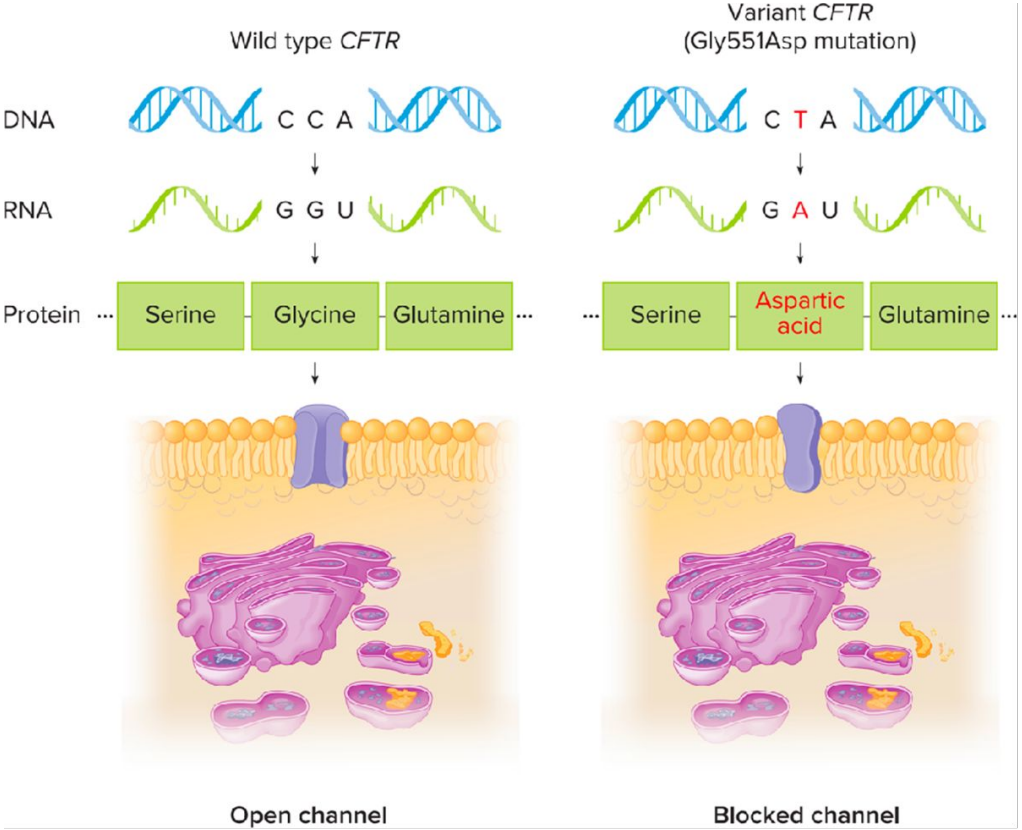
Phenotype

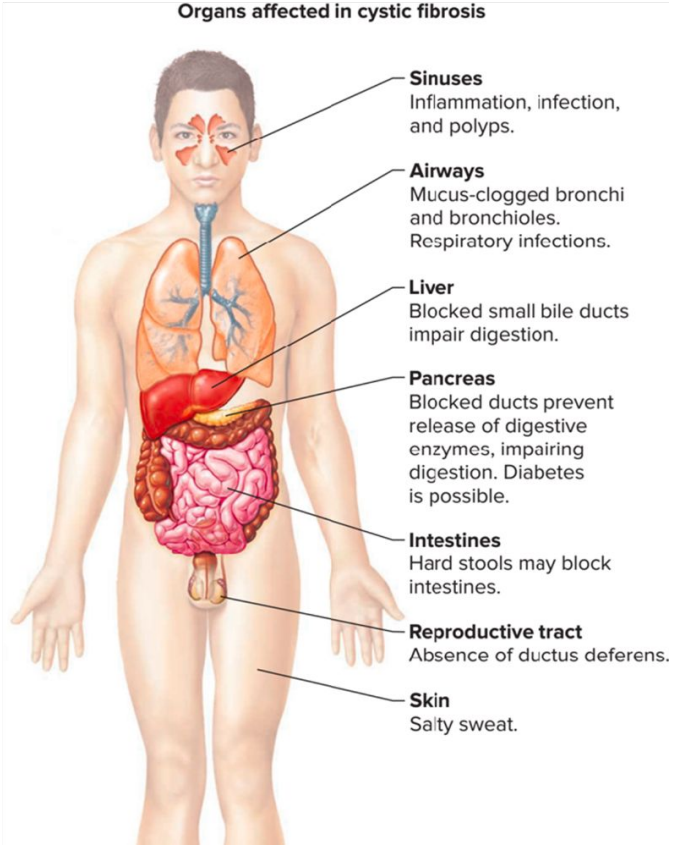
Genotype

Genomics


Genomics
Epigenomics
Transcriptomics
Proteomics
DNA
RNA
Protein
POL
CH3
CH3
CH3
CH3
CH3
CH3
CH3

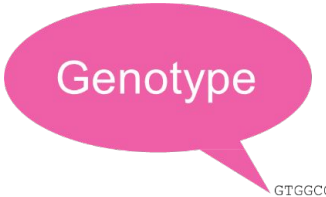
From gene to protein to person





Human Genetics. 12th edit. Ricki Lewis.

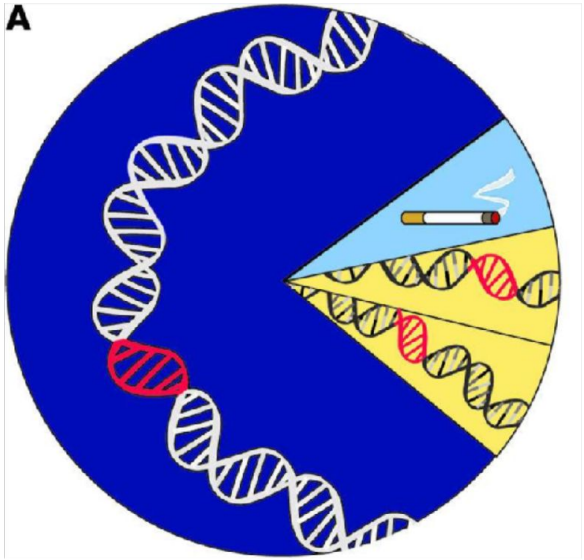
Human Genomic Variation

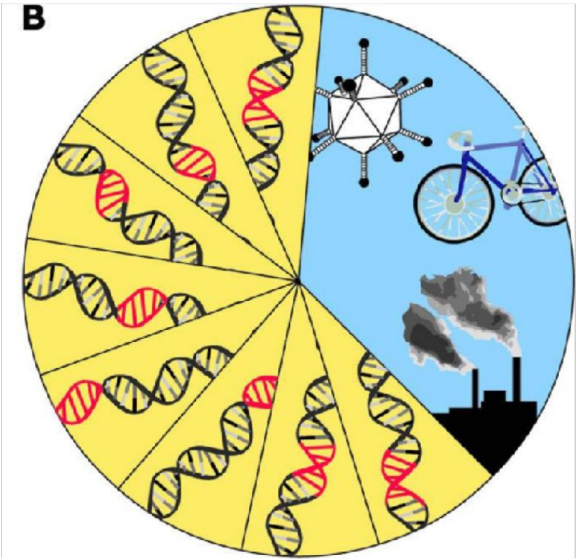


GTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCTGCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGG
AGAAGCGTGAGGGGACAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTGCACCTCTGGAGCGGGTTAGTGGTGGTGGTAGTGGGT
TGGGACGAGCGCGTCTTCCGCAGTCCCAGTCCAGCGTGGCGGGGGAGCGCCTCACGCCCCGGGTCGCTGCCGCGGCTTCTTGCCCTTTTGTCTCTGCCAACCCCC
ACCCATGCCTGAGAGAAAGGTCCTTGCCCGAAGGCAGATTTTCGCCAAGCAAATTCGAGCCCCGCCCCTTCCCTGGGTCTCCATTTCCCGCCTCCGGCCCGGCCT
TTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATCTGAAATTTCTTGGAAACACGATCACTTTAACGGAATATTGCTGT
TTTGGGGAAGTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTTACTTAAGCCCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGGAATTACCAGGCGGC
GTTGGTCTCTAACTGGAGCCCTCTGTCCCCACTAGCCACGCGTCACTGGTTAGCGTGATTGAAACTAAATCGTATGAAAATCCTCTTCTCTAGTCGCACTAGCCA
CGTTTCGAGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAAAATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGACTGA
ATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCTTCTGTTTTGCAGACT
TATTTACCAAGCATTGGAGGAATATCGTAGGTAAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTTTTAAGACACGCTGCAACAAAGCAGGT
ATTGACAAATTTTATATAACTTTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAAACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACT
GTTCCTTATGTGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACCACATGATAAATATAGAACGTCTAGTGGATAAAGAGGAAACTGGCCCCT
TGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGTTACTTTATGGCAGAAGTTGTCCAACTTTTTGGTTTCAGTACTCCTTATACTCTTAAAAA
TGATCTAGGACCCCCGGAGTGCTTTTGTTTATGTAGCTTACCATATTAGAAATTTAAAACTAAGAATTTAAGGCTGGGCGTGGTGGCTCACGCCTGTAATCCCAG
CACTTTGGGAGGCCGAGGTGGGCGGATCACTTGAGGCCAGAAGTTTGAGACCAGCCTGGCCAACATGGTGAAACCCTATCTCTACTAAAAATACAAAAAATGTGC
TGCGTGTGGTGGTGCGTGCCTGTAATCCCAGCTACACGGGAGGTGGAGGCAGGAGAATCGCTTGAACCCTGGAGGCAGAGGTTGCAGTGAGCCAAGATCATGCCA
CTGCACTCTAGCCTGGGCCACATAGCATGACTCTGTCTCAAAACAAACAAACAAACAAAAAACTAAGAATTTAAAGTTAATTTACTTAAAAATAATGAAAGCTAA
CCCATTGCATATTATCACAACATTCTTAGGAAAAATAACTTTTTGAAAACAAGTGAGTGGAATAGTTTTTACATTTTTGCAGTTCTCTTTAATGTCTGGCTAAAT
AGAGATAGCTGGATTCACTTATCTGTGTCTAATCTGTTATTTTGGTAGAAGTATGTGAAAAAAAATTAACCTCACGTTGAAAAAAGGAATATTTTAATAGTTTTC
AGTTACTTTTTGGTATTTTTCCTTGTACTTTGCATAGATTTTTCAAAGATCTAATAGATATACCATAGGTCTTTCCCATGTCGCAACATCATGCAGTGATTATTT
GGAAGATAGTGGTGTTCTGAATTATACAAAGTTTCCAAATATTGATAAATTGCATTAAACTATTTTAAAAATCTCATTCATTAATACCACCATGGATGTCAGAAA
AGTCTTTTAAGATTGGGTAGAAATGAGCCACTGGAAATTCTAATTTTCATTTGAAAGTTCACATTTTGTCATTGACAACAAACTGTTTTCCTTGCAGCAACAAGA
GTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCTGCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGG
AGAAGCGTGAGGGGACAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTGCACCTCTGGAGCGGGTTAGTGGTGGTGGTAGTGGGT
TGGGACGAGCGCGTCTTCCGCAGTCCCAGTCCAGCGTGGCGGGGGAGCGCCTCACGCCCCGGGTCGCTGCCGCGGCTTCTTGCCCTTTTGTCTCTGCCAACCCCC
ACCCATGCCTGAGAGAAAGGTCCTTGCCCGAAGGCAGATTTTCGCCAAGCAAATTCGAGCCCCGCCCCTTCCCTGGGTCTCCATTTCCCGCCTCCGGCCCGGCCT
TTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATCTGAAATTTCTTGGAAACACGATCACTTTAACGGAATATTGCTGT
TTTGGGGAAGTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTTACTTAAGCCCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGGAATTACCAGGCGGC
GTTGGTCTCTAACTGGAGCCCTCTGTCCCCACTAGCCACGCGTCACTGGTTAGCGTGATTGAAACTAAATCGTATGAAAATCCTCTTCTCTAGTCGCACTAGCCA
CGTTTCGAGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAAAATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGACTGA
ATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCTTCTGTTTTGCAGACT
TATTTACCAAGCATTGGAGGAATATCGTAGGTAAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTTTTAAGACACGCTGCAACAAAGCAGGT
ATTGACAAATTTTATATAACTTTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAAACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACT
GTTCCTTATGTGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACCACATGATAAATATAGAACGTCTAGTGGATAAAGAGGAAACTGGCCCCT
TGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGTTACTTTATGGCAGAAGTTGTCCAACTTTTTGGTTTCAGTACTCCTTATACTCTTAAAAA
TGATCTAGGACCCCCGGAGTGCTTTTGTTTATGTAGCTTACCATATTAGAAATTTAAAACTAAGAATTTAAGGCTGGGCGTGGTGGCTCACGCCTGTAATCCCAG

Human Genomic Variation

Genetic influences on disease





Dissected OMIM Morbid Map Scorecard (Updated February 22nd, 2022) :

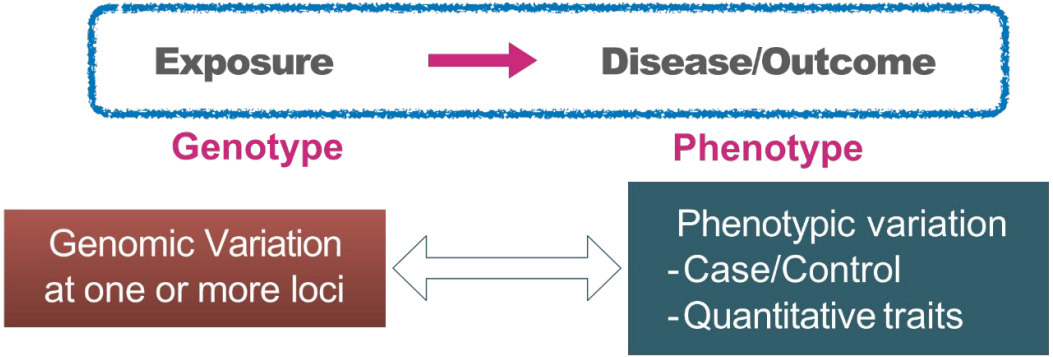
Class of phenotype	Phenotype	Gene *
Single gene disorders and traits	6,032	4,218

www.omim.org

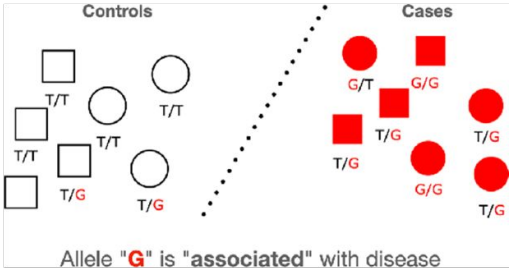
Manolio et al., J Clin Invest. (2008)

Genetic Association Study
between genetic variations and phenotype variations

- Objective: Is there a statistical association?

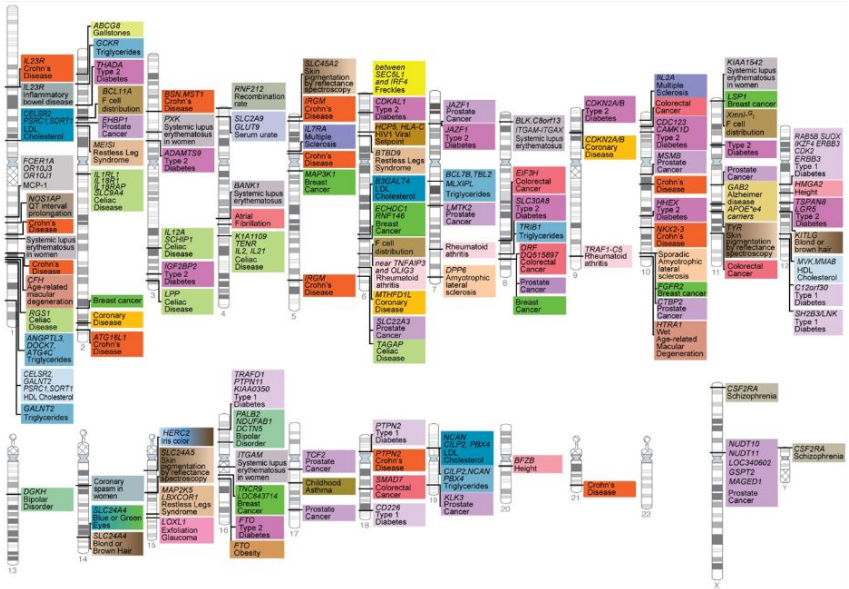


- Candidate genes
- Genome-wide association study (GWAS)
  - Whole Genome-Wide SNPs array (GWAS, genotyping)
  - Whole Genome Sequencing (WGS)
  - Whole Exome Sequencing (WES)



Genome-Wide Association Study (GWAS)

2008



Manolio, Brooks, Collins, J. Clin. Invest., May 2008

2019

As of 2021.02.10, the GWAS Catalog contains 4,865 publications and 247,051 associations.

Understanding biological pathways of disease

>70,000 loci at genome-wide significance, for 100s of diseases and traits

Inflammatory Bowel Disease
Autophagy, TGFβ signaling, other pathways

Age-related Macular Degeneration
Complement system

Heart Disease
HDL not protective, non-lipid pathways

Atrial Fibrillation
Sarcomere and contractile proteins

Schizophrenia
synaptic pruning

Sickel-cell complication
Control of fetal hemoglobin

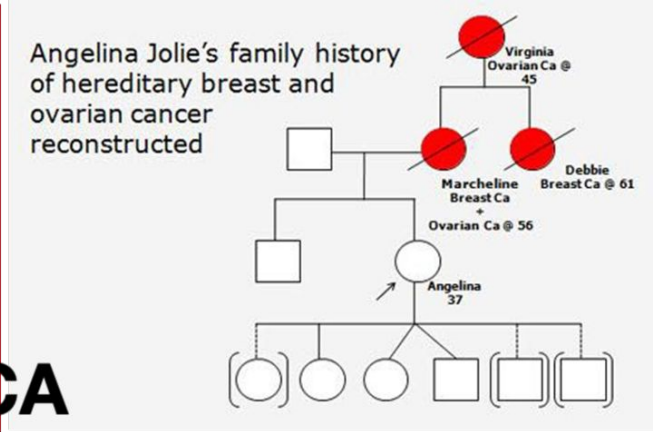
Alzheimer's
microglia

Obesity
regulation of thermogenesis

INTERNATIONAL COMMON DISEASE ALLIANCE | Maps to Mechanisms to Medicine

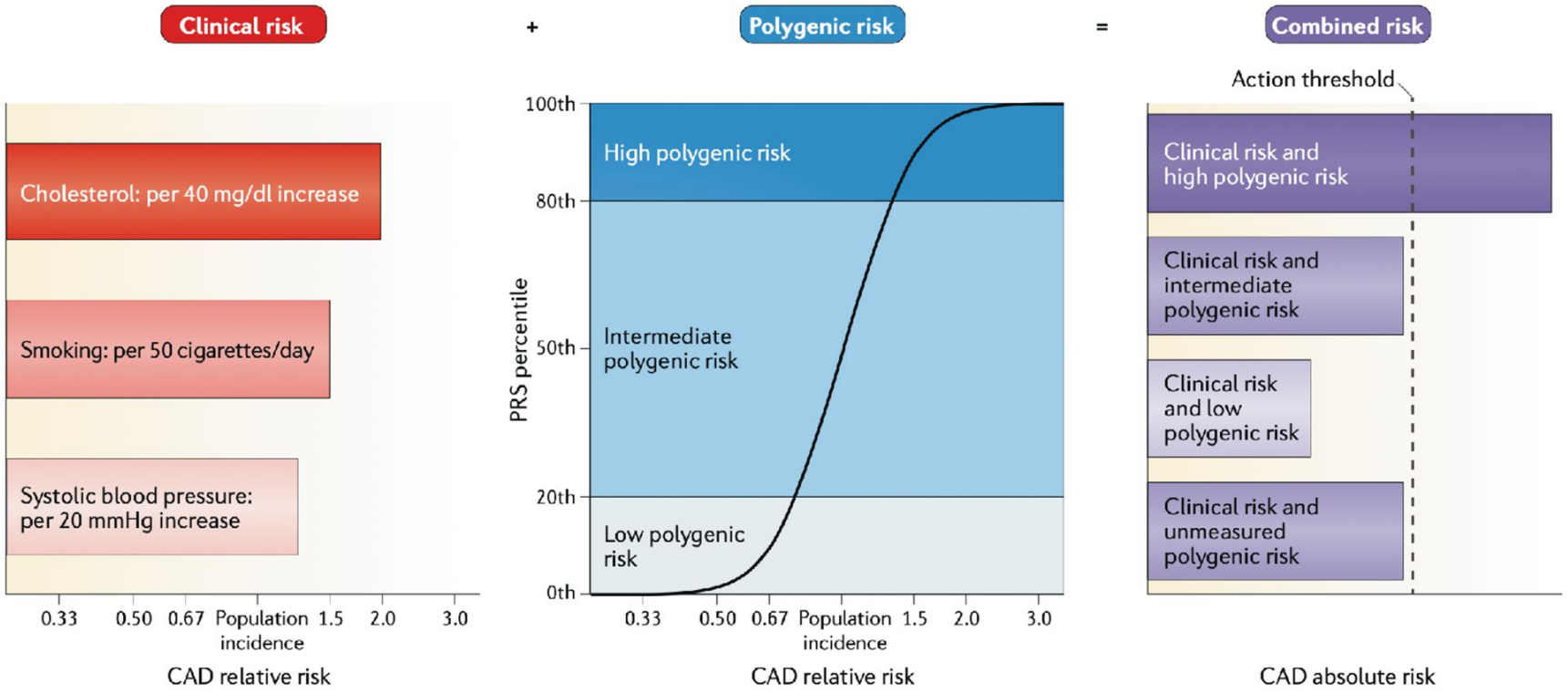
Co-Hosted by NIH

Post Human Genome Project



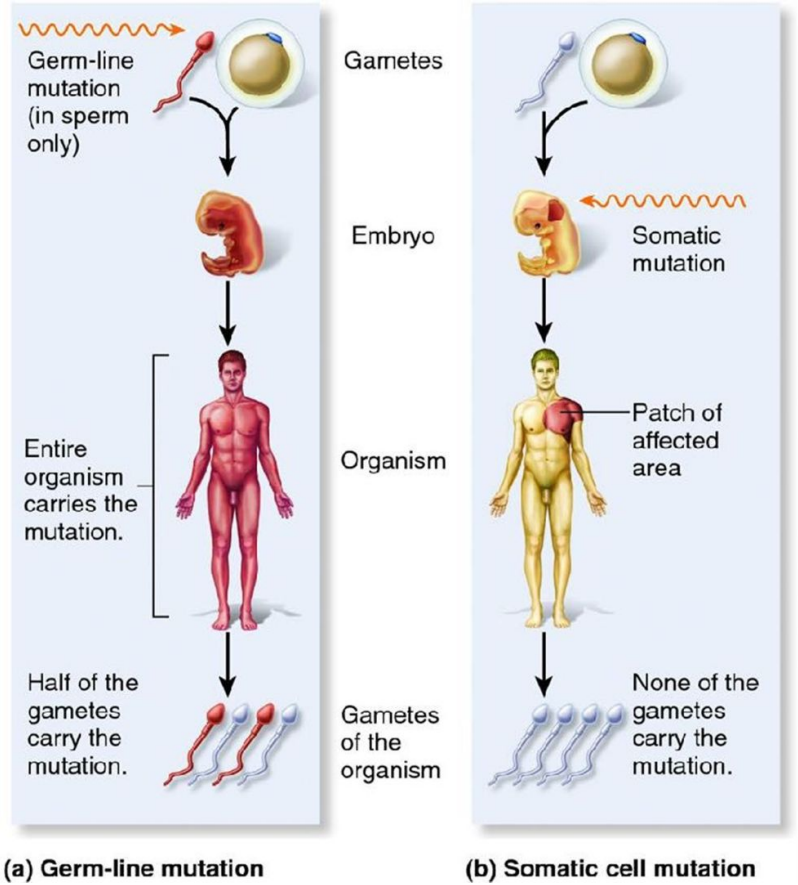
2019.
$1,000~2,000 (HiSeq X)
<$1,000 (NovaSeq)

Prediction & Prevention

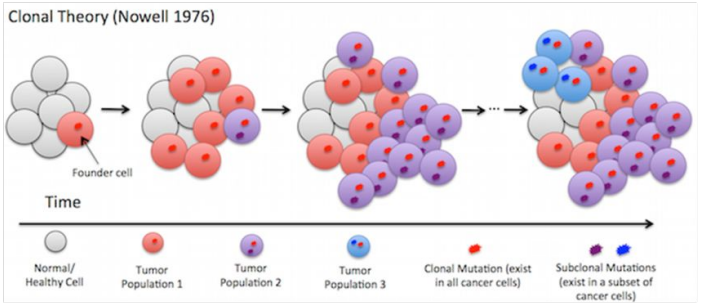


Nat Rev Genet. 2018 Sep;19(9):581-590.

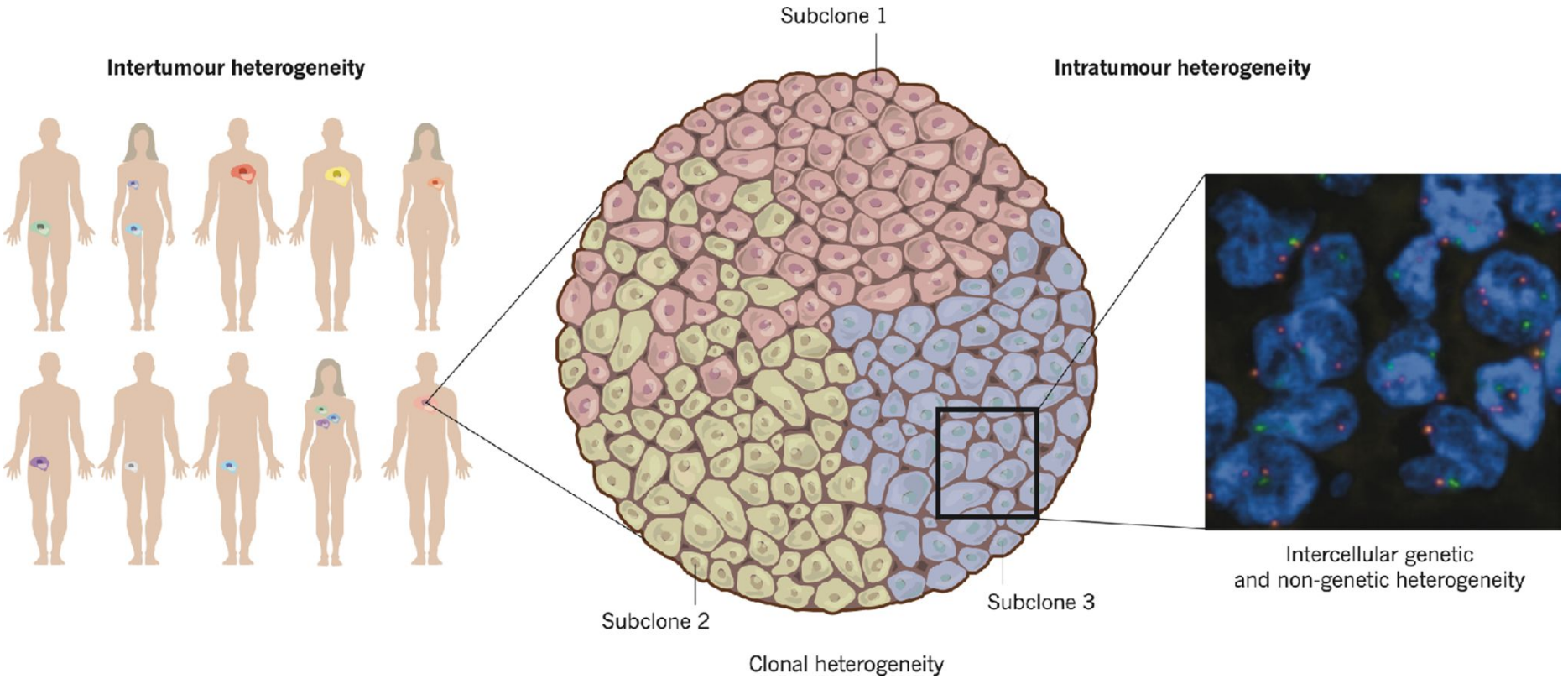
Germ-line or Somatic mutations



- Inherited disease : germ-line mutation in every cell in our body
- Cancer : somatic mutation in a specific cell or tissue (except inherited cancer)



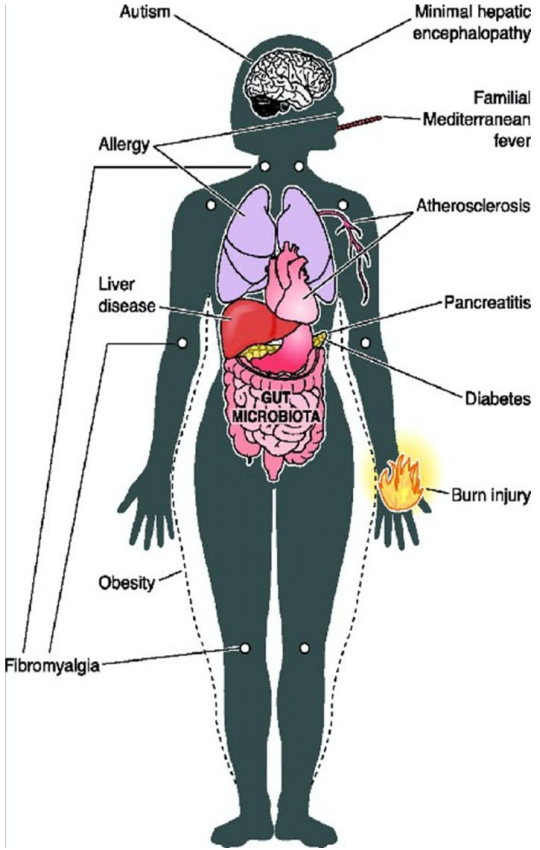
Cancer Genomics



Nature 501(7467):338-45 (2013)

Microbiome

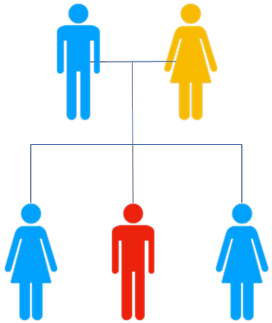
- Kill you by acute infection
- Prevent same infection
- Make you fat(ter)
- Give you a heart attack
- Give you cancer
- Rescue you from cancer



Physiological Reviews Published, 2010 Vol. 90 no. 3, 859-904

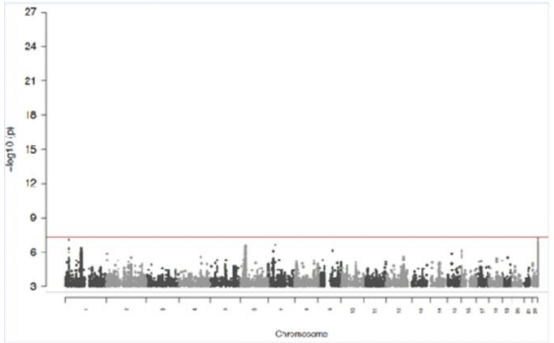
Personalized Medicine, P4 medicine, Precision Medicine

- **Preventive:** Shifting from a treatment-centric approach to a focus on prevention and health promotion.
- **Prediction:** Predicting the likelihood of disease occurrence and preparing accordingly.
- **Personalized:** Tailored medicine, individualized treatment, and customized healthcare.
- **Participatory:** Empowering patients and doctors to interact on equal footing, actively utilizing personal health information, and shifting from hospital-centric to patient-centric care.

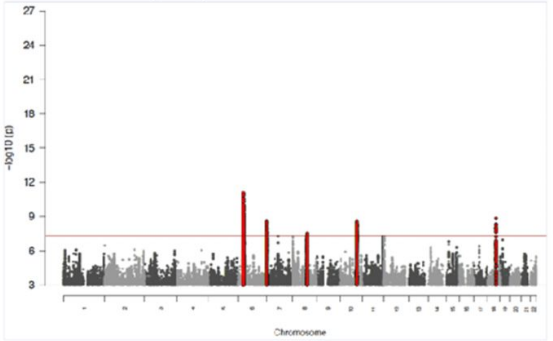


We need larger sample size

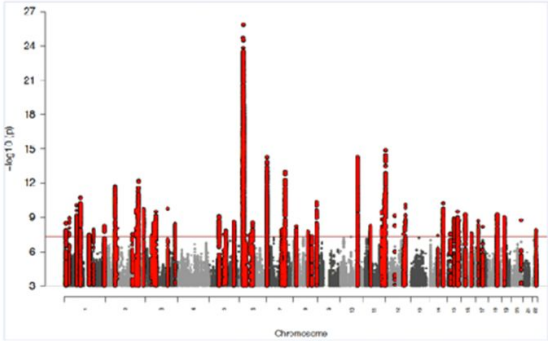
6,000 people 0 genes!



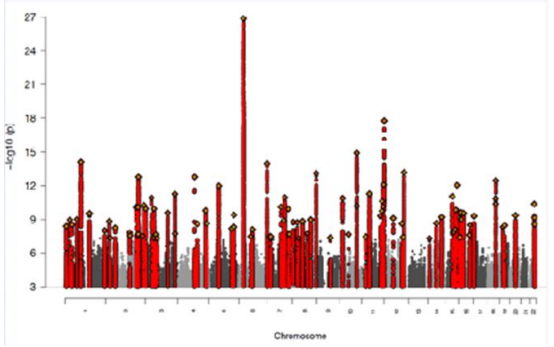
20,000 people 5 genes



50,000 people 62 genes

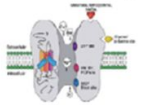


110,000 people 108 genes !

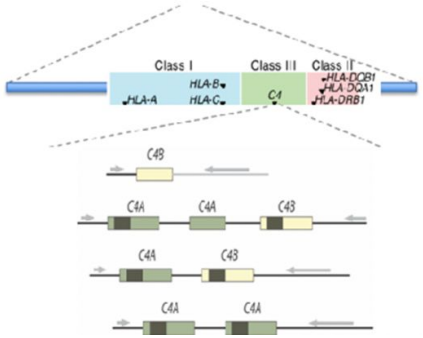


Calcium channels

Glutamate signaling

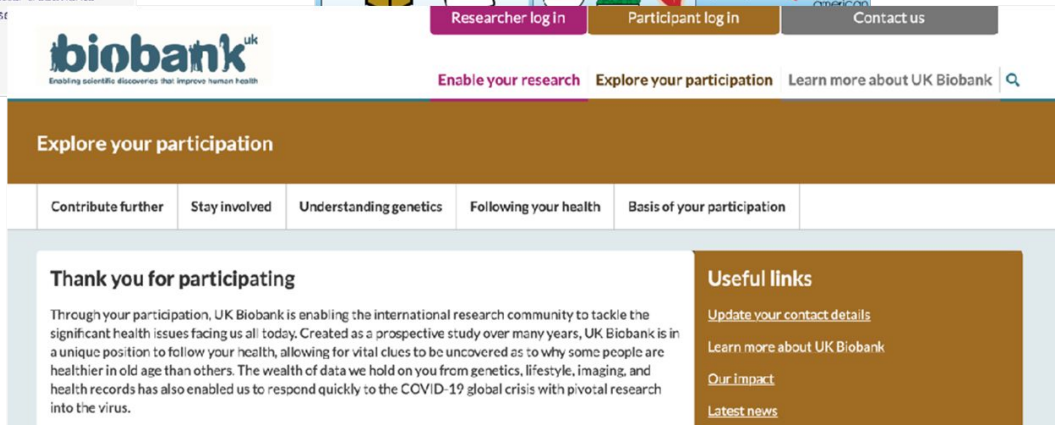


-- Chr 6 --



Eric Lander, ASHG (2018)

Participation



Genotype to Phenotype and Need for Large Cohorts

History of UK Biobank

2003 UK Biobank established

2006 Recruitment begins

2010 500,000 recruited

2012 Open for research

2013 Physical activity data collection starts
Death and cancer registries made available for study

2014 Hospital outpatient data made available

2015 First study using genetic data is published

2016 Biochemistry studies begin
Mental health questionnaire conducted
Imaging of 100,000 begins

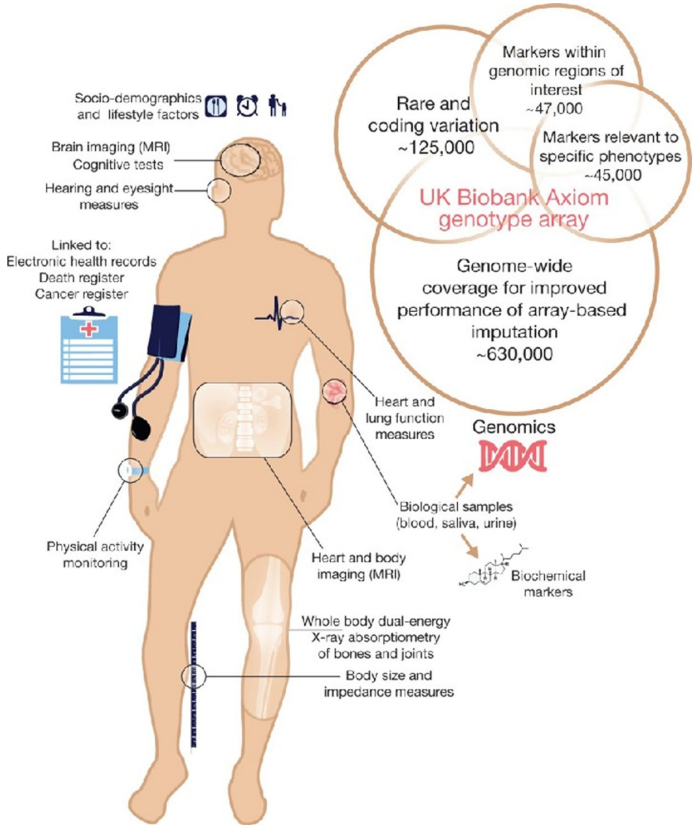
2017 Exome sequencing begins
Genetic data of full cohort released for research

2018 Whole genome sequencing begins (on 50,000 participants)

2019 10,000th registered researcher
Clinical, prescription data for ~45% of participants made available

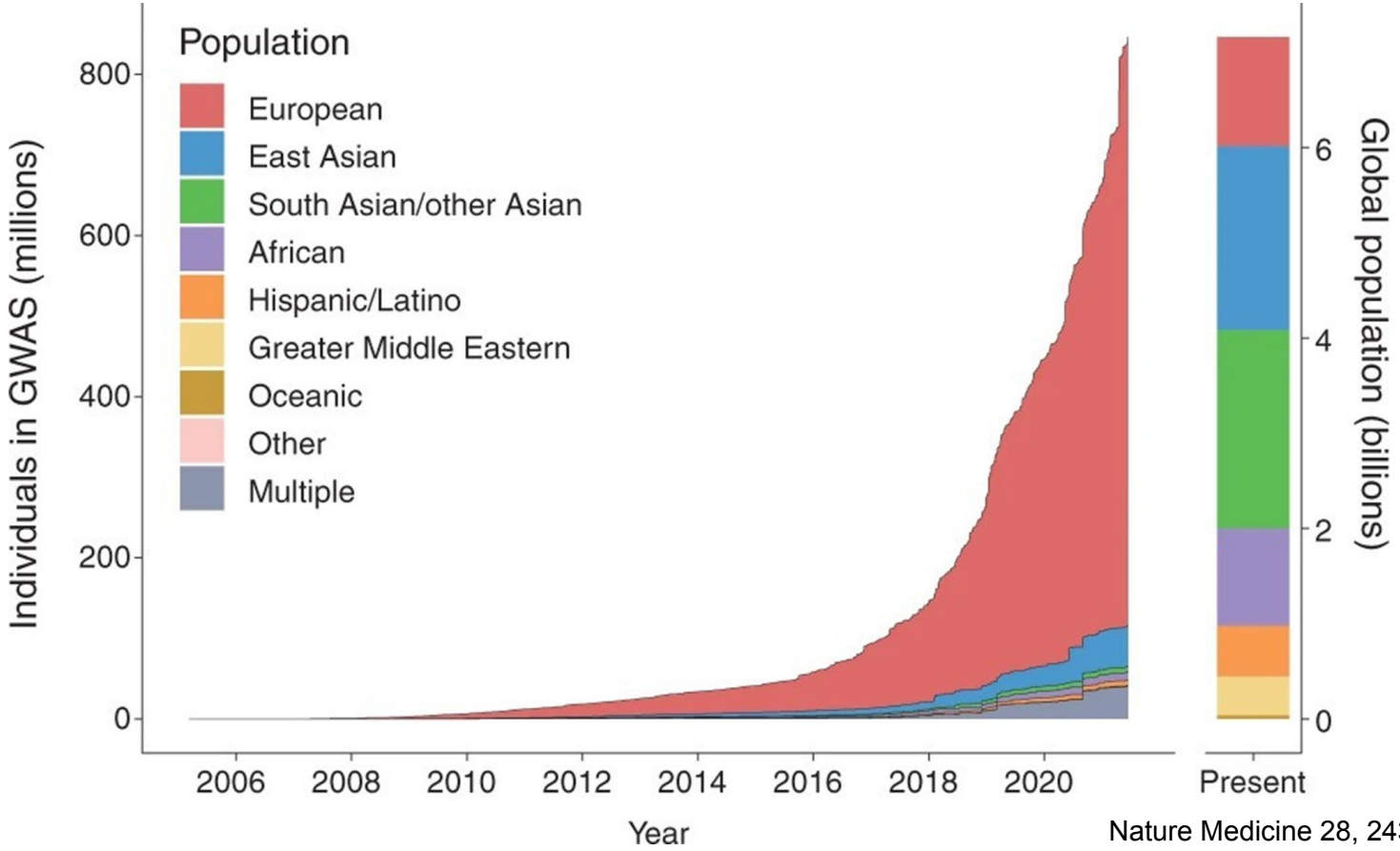
SOURCE: UK BIOBANK

KNOWABLE MAGAZINE



Bycroft et al., Nature (2018)

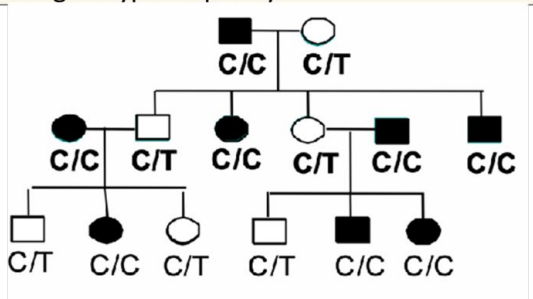
Major challenge & opportunity

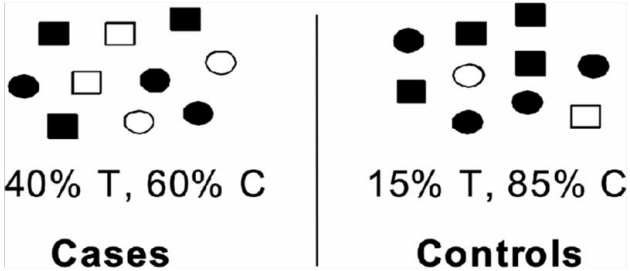


Nature Medicine 28, 243-250 (2022)

Linkage vs. Association Study

Linkage Analysis	Association Analysis
Requires families (within)	Families or **unrelated population** (between)
Recombination fraction 사용	Allele or genotype frequency 사용
Matching/ethnicity generally unimportant	Matching/ethnicity generally important
Simple inheritance, rare traits	Complex inheritance, common traits
Powerful for rare variants	Powerful for common variants
The genotype frequency of offspring is determined by parental information and thus is not influenced by population genotype frequency.	It is affected by population stratification.





Candidate gene approaches

- Within the frame of conventional epidemiologic study designs
- Rely on a *priori* knowledge about disease etiology
  - Known region?
  - Biological support
- Based on previous studies, such as GWAS and functional studies..
  - e.g. type I diabetes : the human leukocyte antigen (*HLA*) DR3/DR4 alleles
  - Alzheimer's disease: Apolipoprotein E (*APOE*) ε2/ ε3/ ε4 alleles
- Genotyping only specific variants within the gene of interest or conducting targeted sequencing of the gene.

Candidate Gene: Where do I start?

- Location
  - What chromosome? What position on the chromosome?
- Exons/UTR
  - How many exons? UTR regions?
- Size of gene?
- Effect of the variants
  - a potential biological impact?
  - missense variant?
- Use the genome browsers
  - UCSC genome browser
  - Ensemble genome browser

Genotyping using genetic marker

TaqMan and Fluidigm
Efficient for analyzing multiple SNP markers, ranging from one to several dozen, across numerous samples simultaneously.

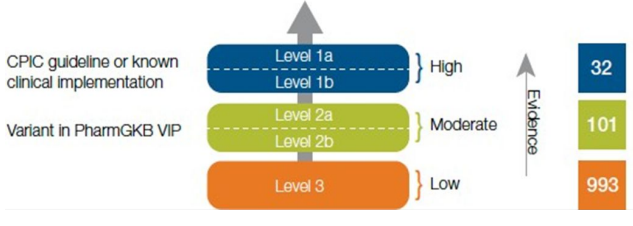
Microarray (DNA chip)
Illumina: Infinium Global Screening Array (GSA)
ThermoFisher: Axiom Precision Medicine Research Array (PMRA)
Efficient for analyzing over one million SNP markers simultaneously.
Commonly used in GWAS (Genome-Wide Association Studies) research.

Table 1. Axiom Asia PMRA key marker groups.

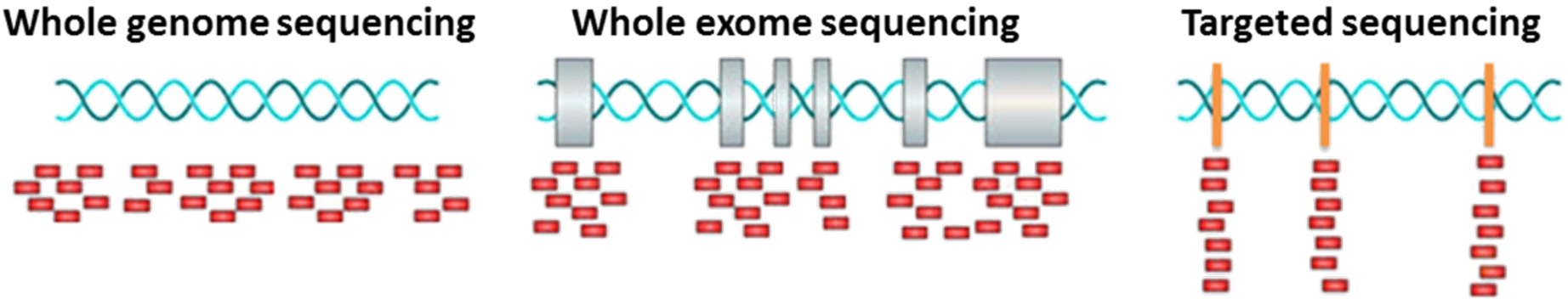
Variant category	Number of markers*
Genome-wide imputation grid**—focus on EAS and SAS populations	>540,000
NHGRI-EBI GWAS catalog	>23,400
Markers of clinical relevance	
ClinVar	>43,000
ACMG	>9,200
Pharmacogenomics and ADME	>2,600
Additional high-value markers (subset of ClinVar: *APOE, BRCA1/2, DMD, CFTR*)	> 2,000
Immune-related markers	
Human leukocyte antigen (HLA)	>9,000
Killer immunoglobulin-like receptor (KIR)	>1,400
Autoimmune and inflammatory	>250
Functional markers	
LOF	>43,000
Very rare nonsynonymous variants (minor allele frequency (MAF) > 0.01%)	>35,000
Expression quantitative trait loci (eQTL)	>15,000
Lung function phenotypes	>7,600

Variant category	Number of markers*
Disease-related markers	
Alzheimer's disease	>900
Cardio-metabolic	>360
Neurological disorders	~16,000
Diabetes	>500
Common variants in cancer	>300
Rare missense variants in cancer predisposition genes	>2,600
Rare variants in cardiac predisposition genes	>830
Rare polymorphic variants from Exome Aggregation Consortium (ExAC) data	>4,700
Miscellaneous	
Fingerprinting and sample tracking	>300
Y chromosome	~400
Mitochondrial	~500
Gender determination	~1,000
Chromosome X SNPs and indels	~25,000
Custom variants**	
Add 50,000 custom markers, or fully customize as required	
Total markers	**>750,000**

* Content in categories may overlap.
** 50,000 markers in the GWAS grid can be replaced with custom content without impacting coverage or accuracy.



Sequencing using NGS



Designing studies that utilize genomic research findings

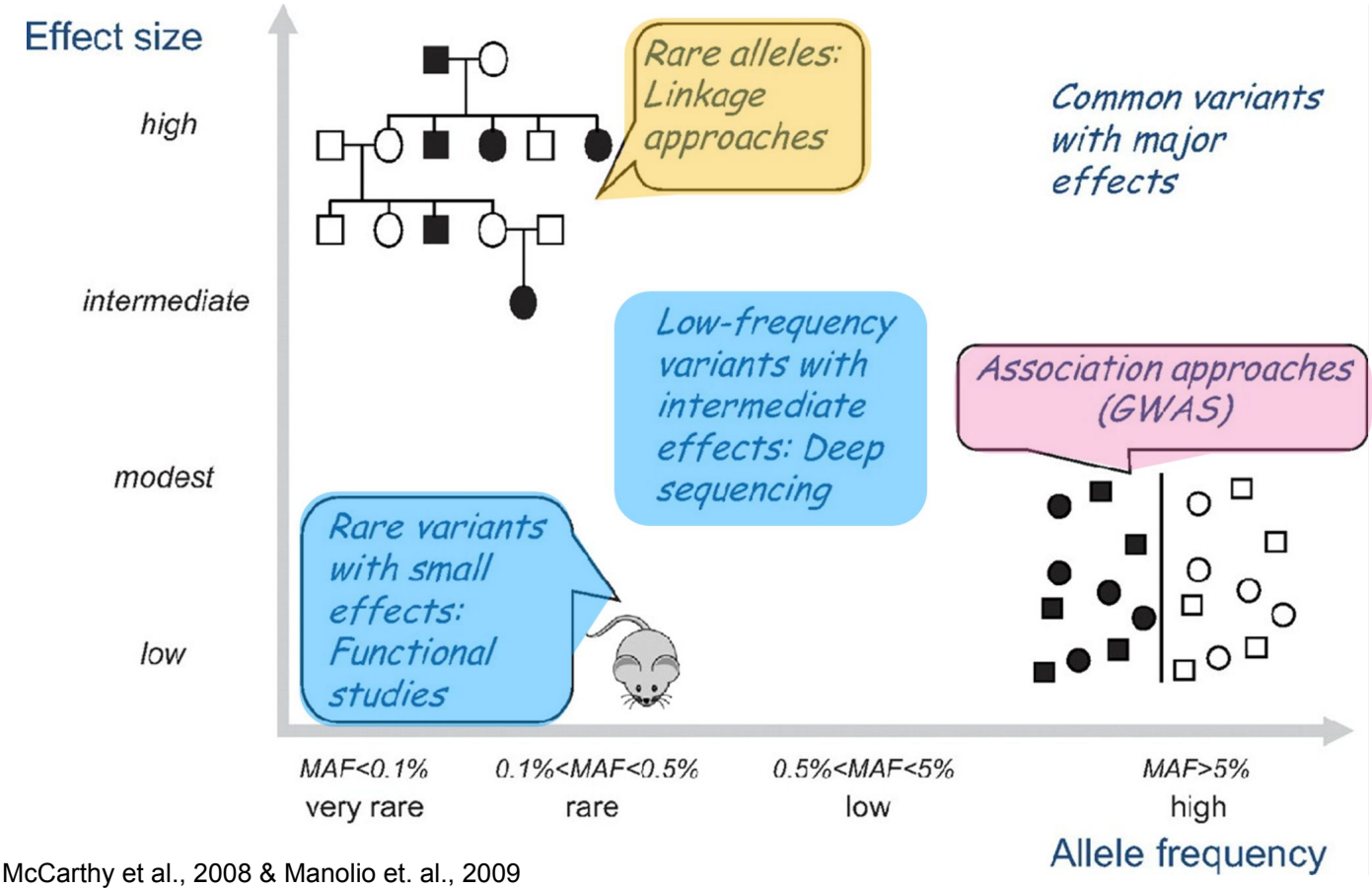
- Mostly utilizing GWAS research findings (GWAS catalog: https://www.ebi.ac.uk/gwas/)
- Predicting disease risk using Polygenic Risk Score
- Utilizing genetic information to increase the efficiency of Randomized Controlled Trials (RCTs)
- Mendelian Randomization: demonstrating causality between exposure and outcome

Genomic research design

- The study design must align with the study goal.
  - ➢ Identifying new candidate genes
  - ➢ predicting risk for known genes in genomic research.
  - ➢ Establishing causal relationships in epidemiological studies.

- Considerations during sample recruitment:
  - ➢ Generalizability
  - ➢ Potential bias

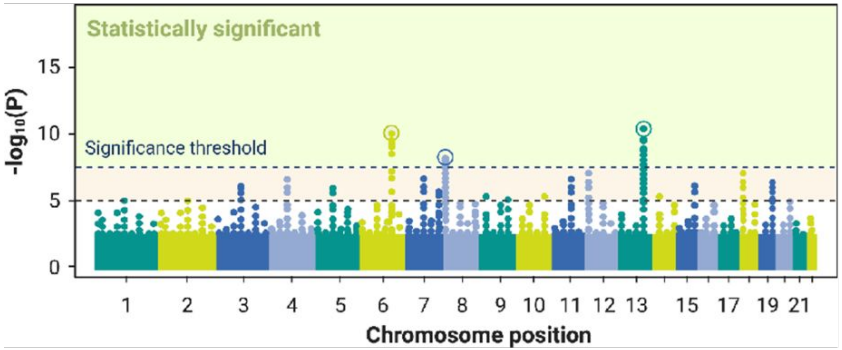
Risk allele frequencies, effect size, and study design



Adapted from McCarthy et al., 2008 & Manolio et. al., 2009

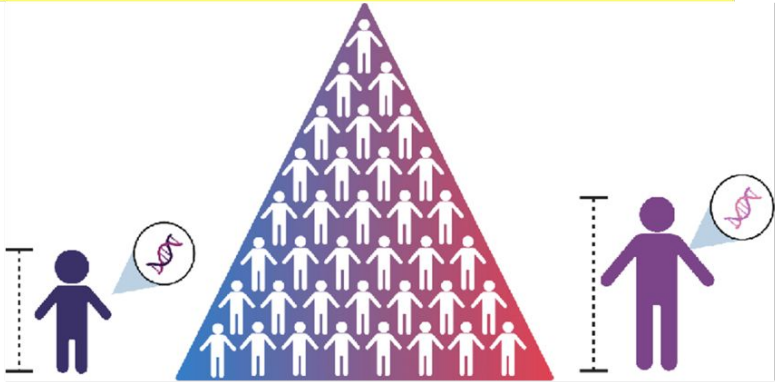
PRS Overview

Discovery data:
GWAS summary statistics



1. Select associated variants
2. Obtain risk allele and effect sizes

Test data:
Independent population

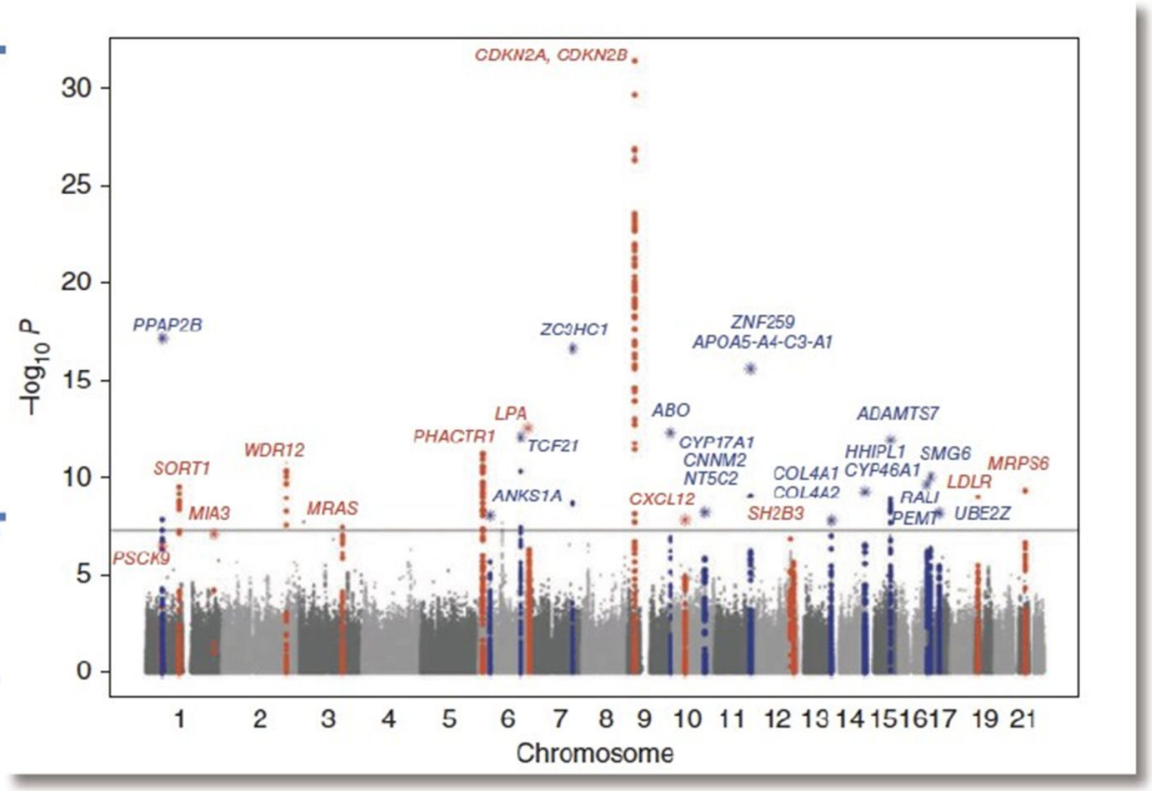


1. Calculate PRS: sum of weighted alleles
2. Evaluate associations with outcome

Move from top SNPs to a genome-wide set for prediction

Kathiresan, *N Engl J Med* (2008)
Ripatti, *Lancet* (2010)
Khera, *N Engl J Med* (2016)

Khera*, Chaffin*,
Nat Genet 2018



Nat Genet. 2018 Sep;50(9):1219-1224.

Predictive power of PRSs

	P-value threshold									
	1.00E-05	0.0005	0.001	0.01	0.03	0.05	0.1	0.3	0.5	1
CAD	0.587	0.607	**0.608**	0.584	0.569	0.562	0.555	0.548	0.546	0.546
DM	0.604	**0.607**	0.598	0.582	0.568	0.564	0.560	0.555	0.554	0.554
HDL	**0.161**	0.121	0.109	0.064	0.042	0.034	0.026	0.020	0.018	0.018
LDL	0.280	**0.298**	0.293	0.217	0.163	0.143	0.120	0.096	0.091	0.088
TG	0.182	**0.196**	0.192	0.136	0.099	0.085	0.070	0.055	0.051	0.050
TC	0.254	**0.275**	0.271	0.207	0.156	0.133	0.109	0.084	0.079	0.077
SCZ	0.694	0.764	0.777	0.817	**0.820**	0.817	0.812	0.805	0.802	0.801
BD	0.524	0.555	0.562	0.609	0.630	0.636	0.654	0.671	0.673	0.671
MDD_PGC	0.515	0.521	0.521	0.531	0.537	0.536	0.537	0.539	**0.540**	0.540
MDD_CONVERGE	0.532	0.539	0.544	0.575	0.582	**0.585**	0.582	0.578	0.577	0.577
Anxiety	0.515	0.519	0.523	0.539	**0.543**	0.541	0.539	0.538	0.538	0.538

DM, type 2 diabetes; CAD, coronary artery disease; TG, triglycerides; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TC, total cholesterol; SCZ, schizophrenia; BD, bipolar disorder; MDD_PGC, study of major depressive disorder by the Psychiatric Genomics Consortium; MDD_CONVERGE, study of major depressive disorder by the CONVERGE Consortium; Anxiety, anxiety disorders (case-control study). For HDL, LDL, TG and TC, predictive power is measured by R^2. Predictive power is measured by AUC for the rest of the traits. Full tables are available in Supplementary Tables S4 and S5.

Bioinformatics. 15;33(6):886-892 (2017)

Evaluating of predictive performance

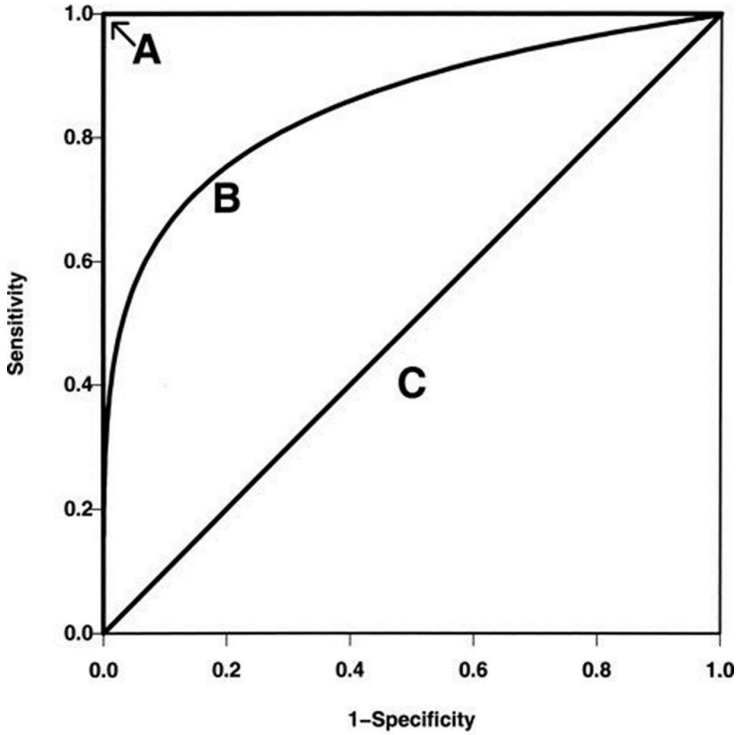
- Receiver operating characteristic curves (ROCs)

  The sensitivity and specificity of the predictions are ranked at various cut-off values.

- Area under a ROC curve (AUC)

  Probability of the examined model correctly identifying a case out of a randomly chosen pair of case and control samples

- AUC results range from 0.5 (i.e., random) to 1 (i.e., 100 % accuracy)

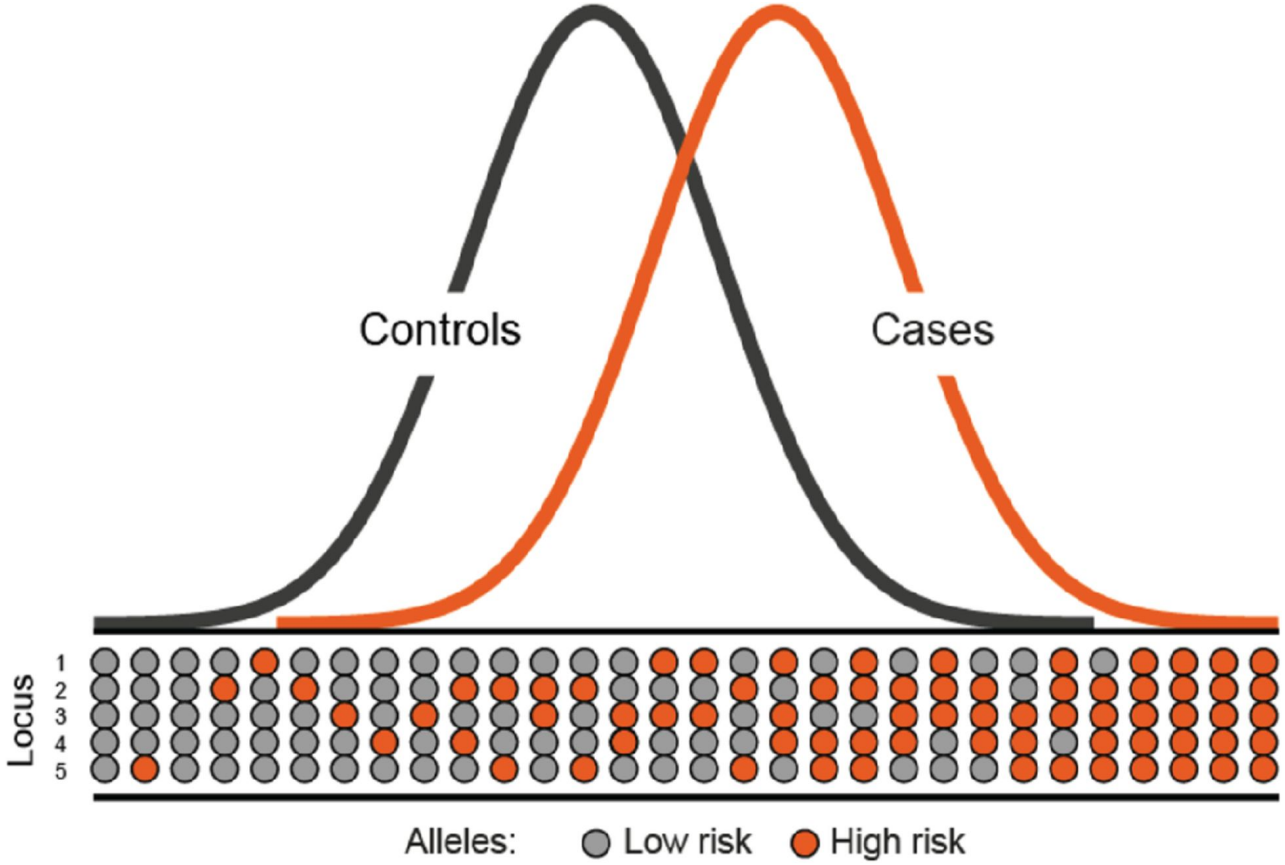


Performance evaluation

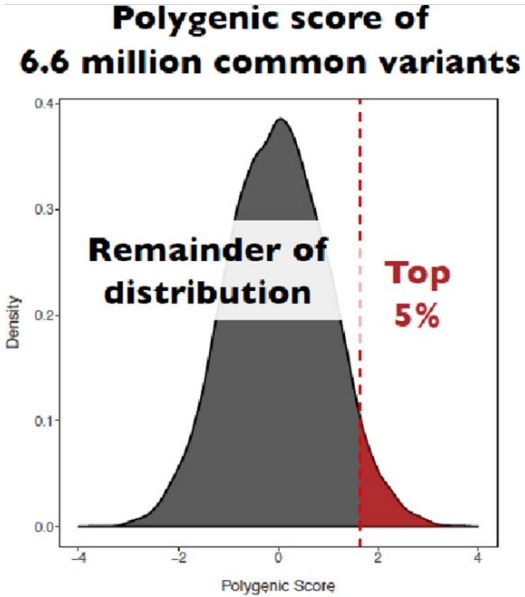
- Accuracy: (TP+TN)/(TP+FN+TN+FP)
- Sensitivity: TP/(TP+FN)
- Specificity: TN/(TN+FP)
- Positive predictive value (PPV): TP/(TP+FP)
- Negative predictive value (NPV): TN/(TN+FN)

		Predicted condition	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP)**, hit	**False negative (FN)**, type II error, miss, underestimation
	Negative (N)	**False positive (FP)**, type I error, false alarm, overestimation	**True negative (TN)**, correct rejection
	Prevalence $= \frac{P}{P+N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$
	Accuracy (ACC) $= \frac{TP+TN}{P+N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$
	Balanced accuracy (BA) $= \frac{TPR+TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$

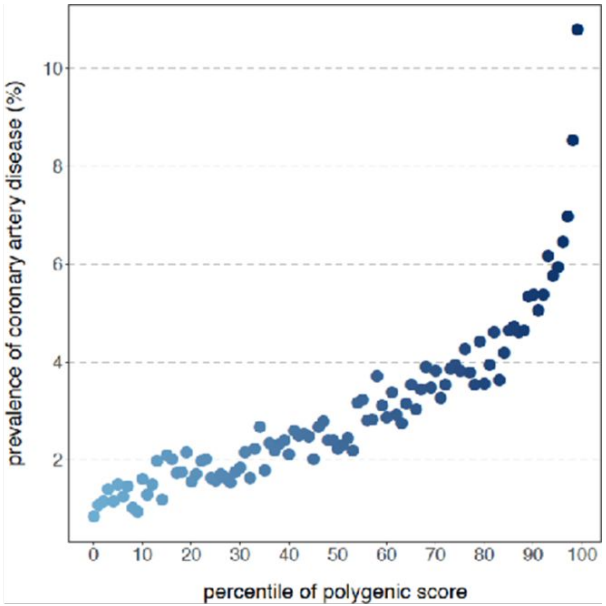
Distribution of PRS



Top 5% of polygenic MI score: risk equivalent to monogenic mutations



High polygenic score definition	Odds ratio
Top 5%	**3.3**
Top 1%	4.7



Khera's slide
Nat Genet. 2018 Sep;50(9):1219-1224

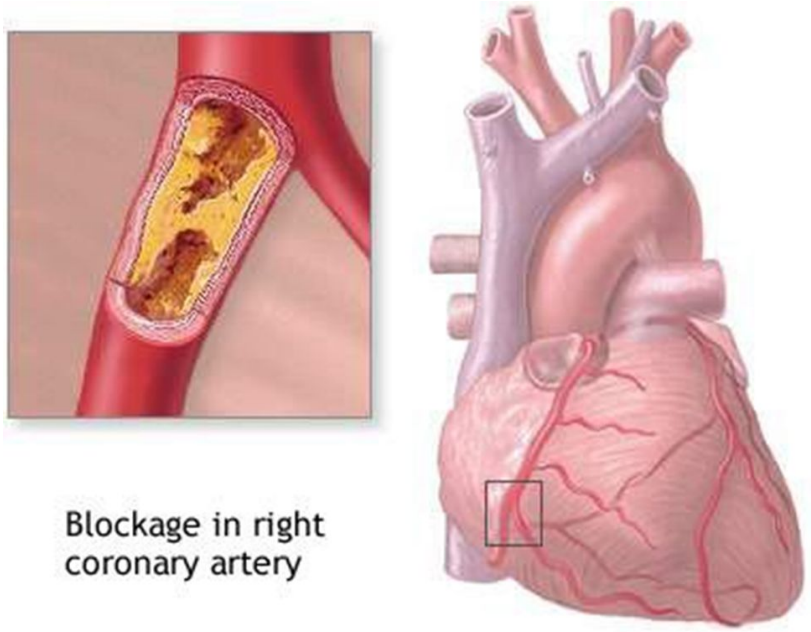
Traditional Approach for Genetic Prediction

- Traditional genetic prediction has mainly focused on rare and monogenic mutations.

  - Familial hypercholesterolemia (FH): Mutations in the LDLR gene, inherited in an autosomal dominant manner, leading to high LDL levels.

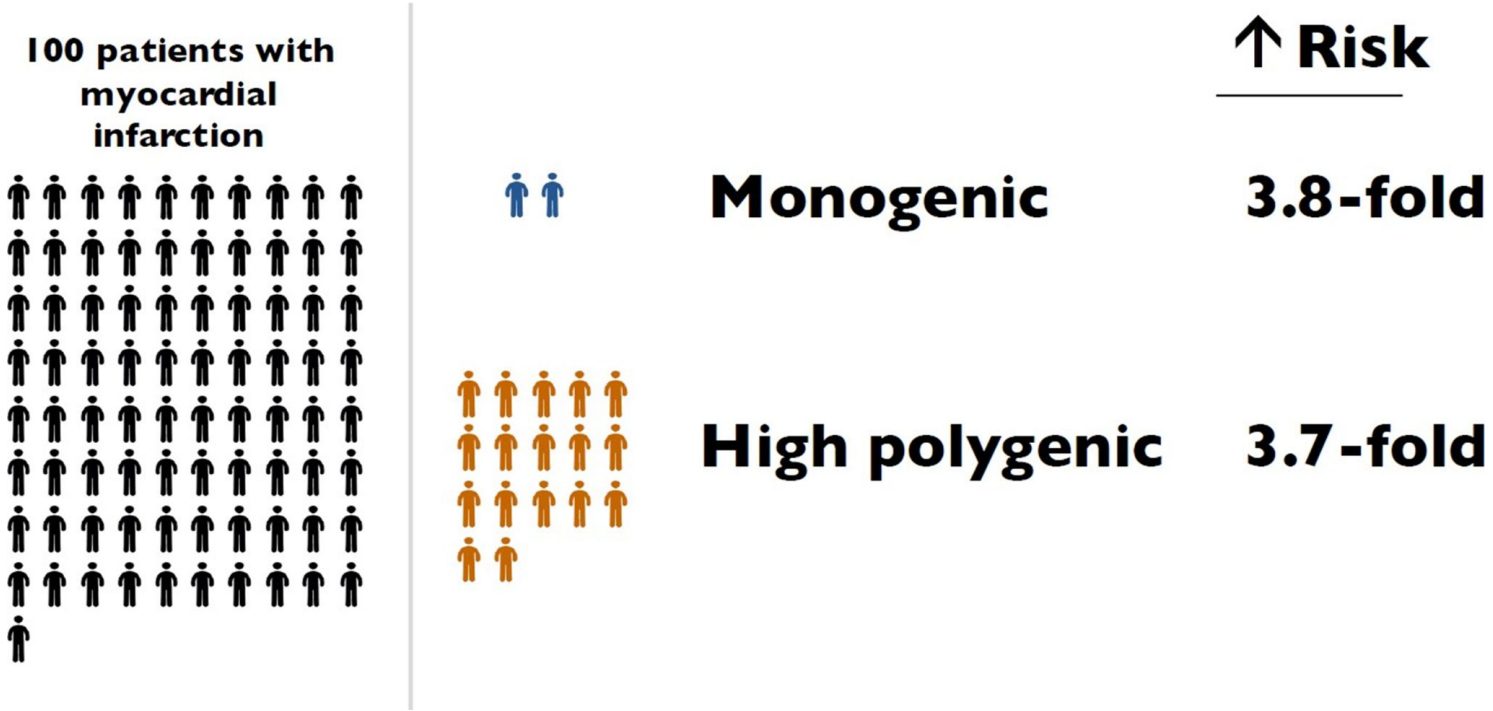
- However, FH affects only 0.4% of the general population, making it rare, and accounts for approximately 2% of early myocardial infarctions (MIs). So, how do we predict and prevent the remaining 98%?

  Approximately half of all MIs present as sudden death on first occurrence.



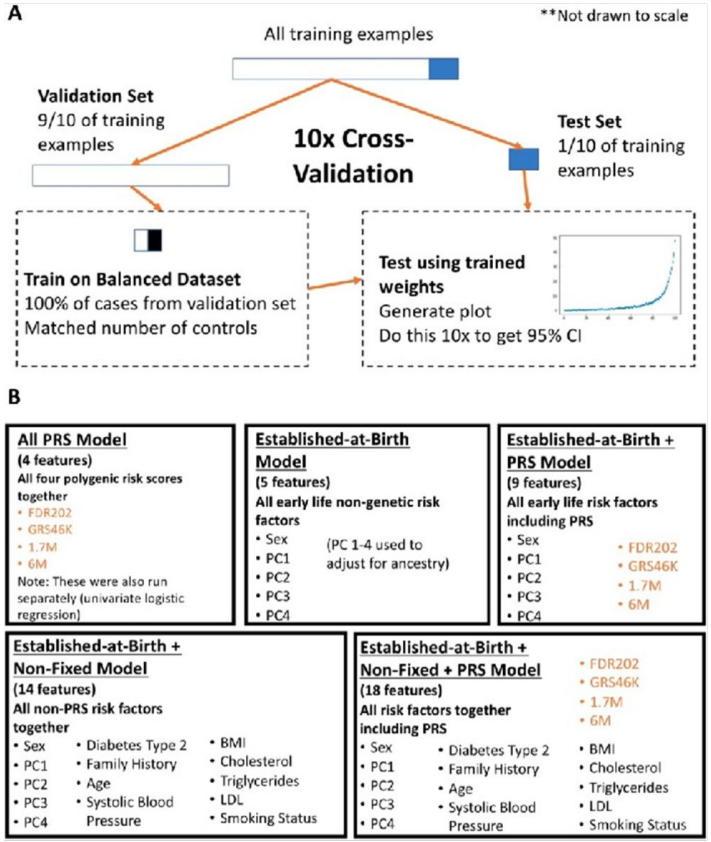
Khera's slide
Nat Genet. 2018 Sep;50(9):1219-1224

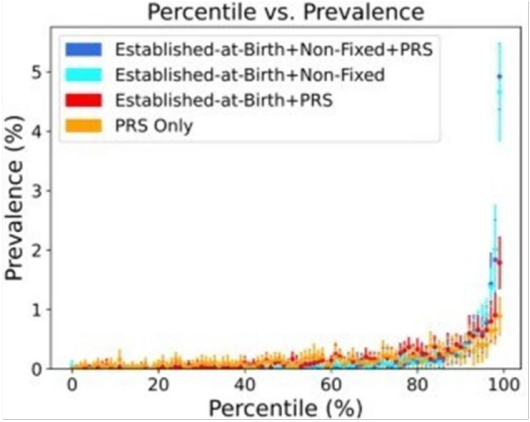
Can we identify additional at-risk individuals with a polygenic risk model?

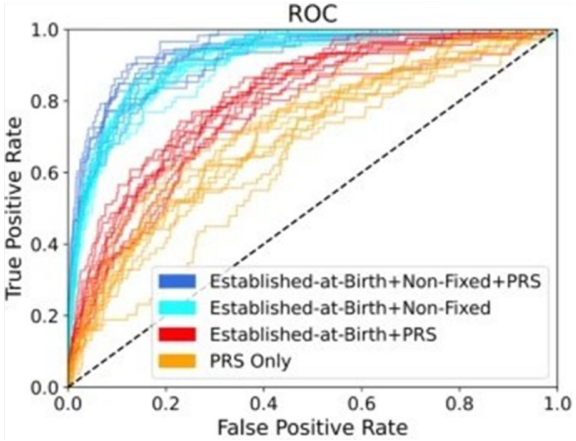


Circulation. 2019 Mar 26;139(13):1593-1602.

PRS predicts early onset MI





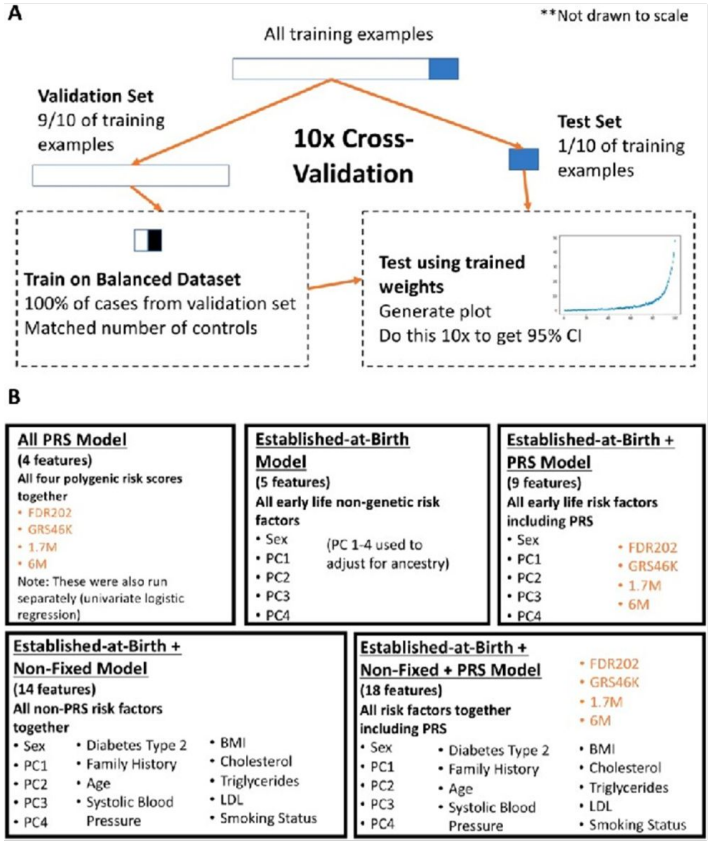


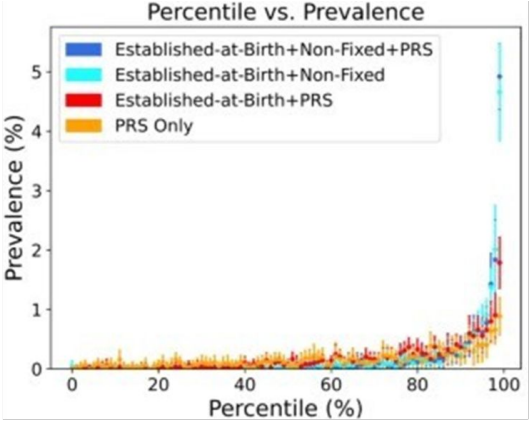
Genome Medicine. 13 (2021)

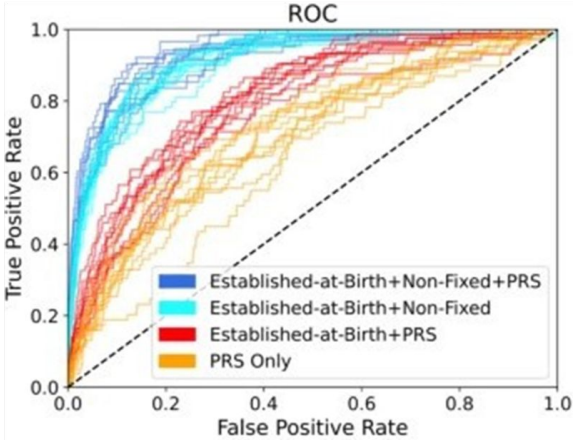
Heritability varies considerably between complex diseases

- Highly heritable (~70% or greater) diseases
  - autoimmune and immune-mediated diseases
  - e.g. celiac disease (CD), type-1 diabetes (T1D), and rheumatoid arthritis
  - the strongest associations typically localizing to the human leukocyte antigen (HLA) region.
  - both in HLA and outside of HLA, many of which are in linkage-disequilibrium (LD) and with different effect sizes
- Less heritable (~50%)
  - common diseases that incur substantial mortality and morbidity worldwide
  - e.g. cardiovascular disease (CVD)
  - weaker genetic associations spread over a large number of genomic loci
- The simplified assumptions underlying polygenic scoring have been shown to reduce the predictive power achieved in HLA-associated diseases including CD and T1D, but not in coronary artery disease and bipolar disorder

PRS predicts early onset MI

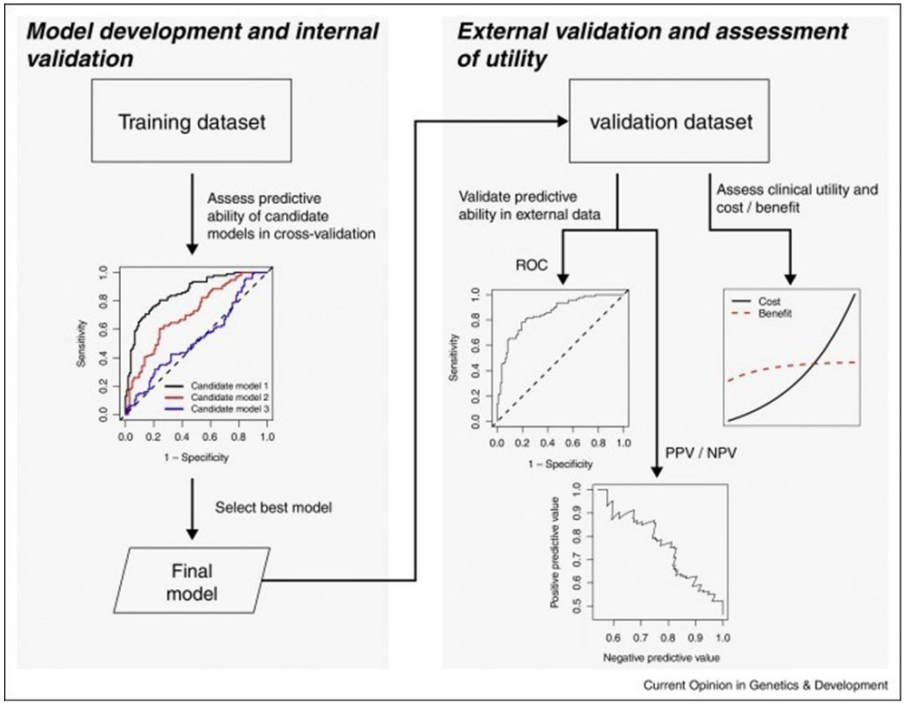


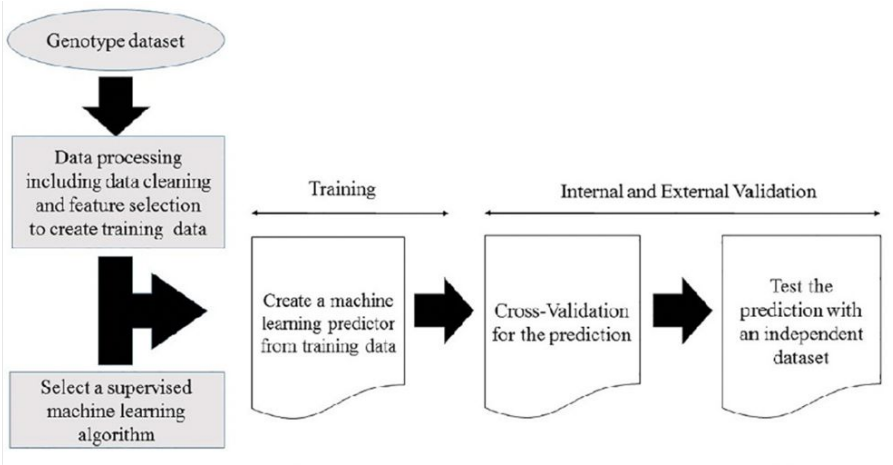


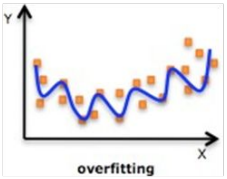


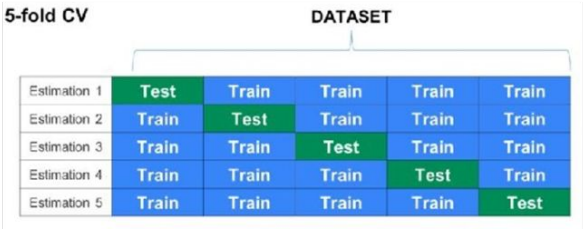
Genome Medicine. 13 (2021)

Machine Learning Disease Prediction Models

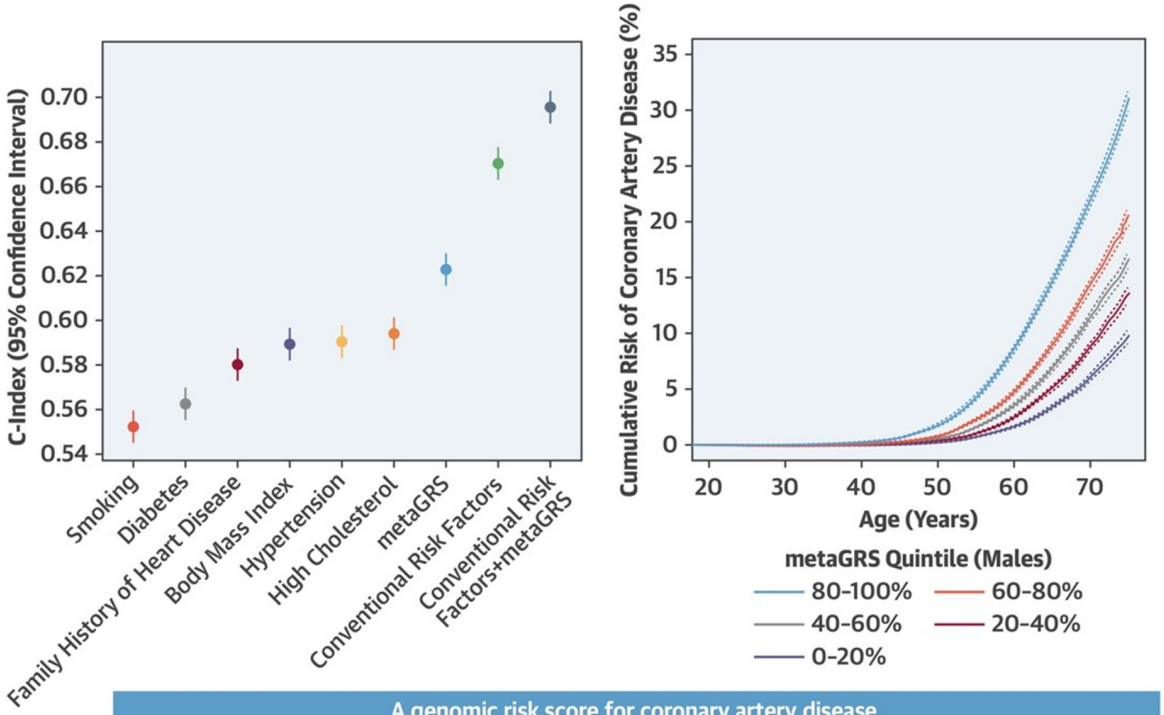








Genomic Risk Score for Coronary Artery Disease

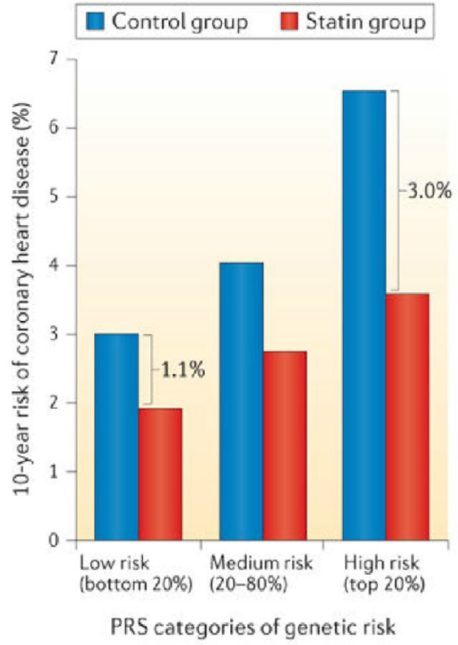


A genomic risk score for coronary artery disease

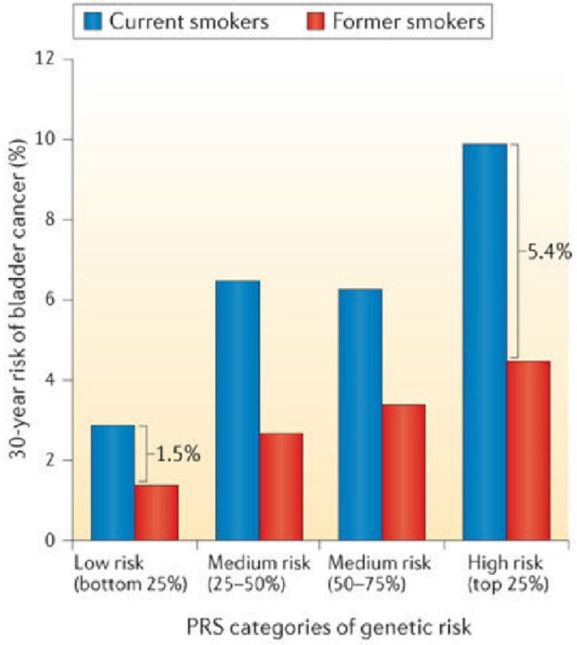
Greater association with future coronary artery disease than any single conventional risk factor
Independent of yet complements conventional risk factors
Provides meaningful lifetime risk estimates of coronary artery disease
Quantifiable at or before birth and shows potential for risk screening in early life

J Am Coll Cardiol. 2018;72(16):1883–93.

Role of PRS in absolute risk reduction

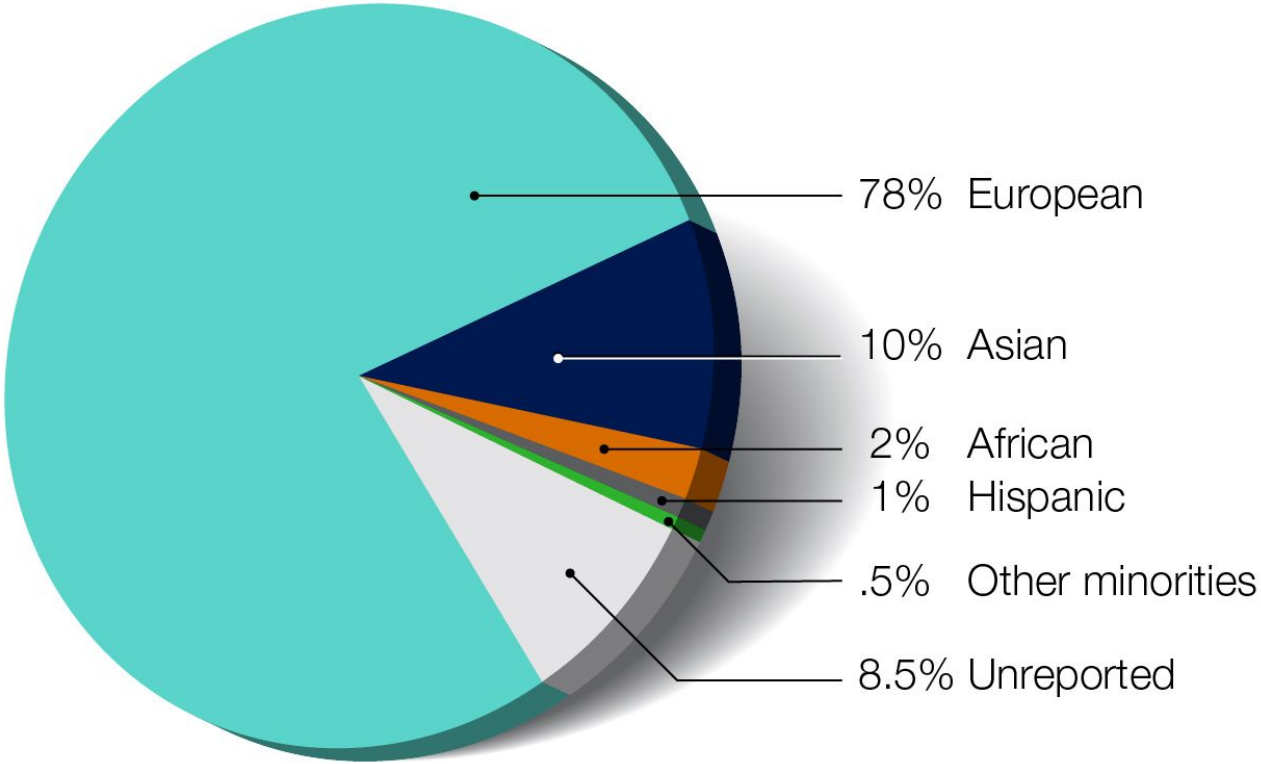


ARR (%)	1.1	1.3	3.0
RRR	0.36	0.32	0.46



ARR (%)	1.5	3.8	2.9	5.4
RRR	0.52	0.60	0.46	0.55

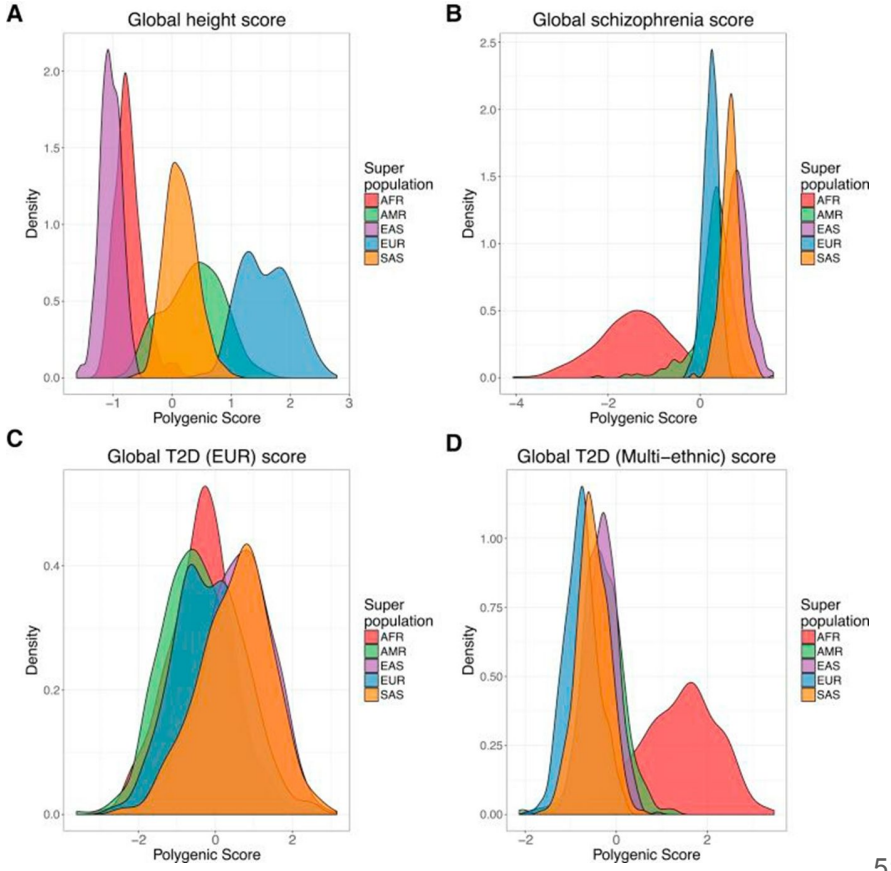
The percentage of ancestry populations in GWAS



What effect does ancestry have on prediction?

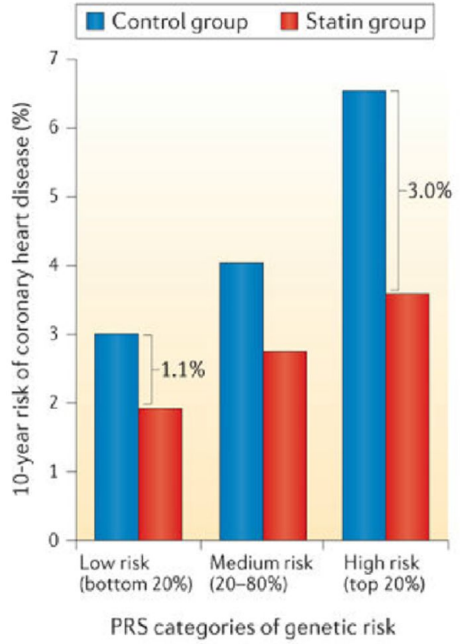
Genetic prediction accuracy decays with increasing genetic distance between discovery and target data

European ascertainment of GWAS signals yield unpredictably biased risk scores in other populations

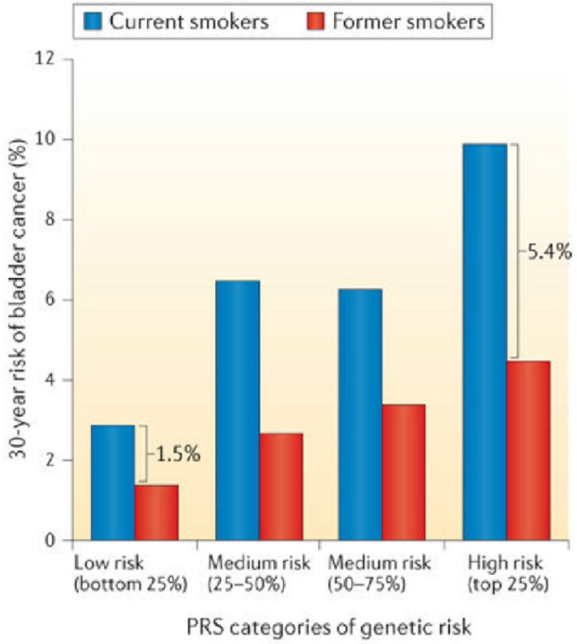


Am J Hum Genet. 2017 Apr 6; 100(4): 635–649.

Role of PRS in absolute risk reduction

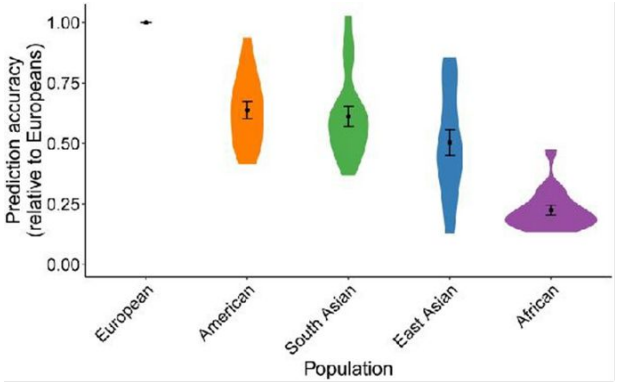


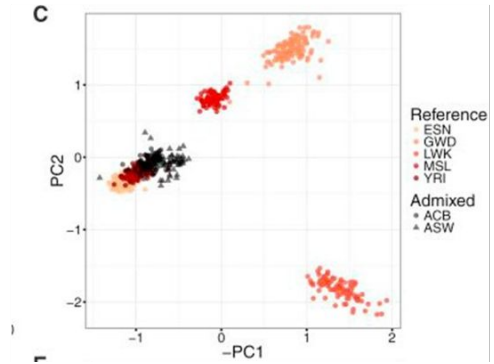
ARR (%)	1.1	1.3	3.0
RRR	0.36	0.32	0.46

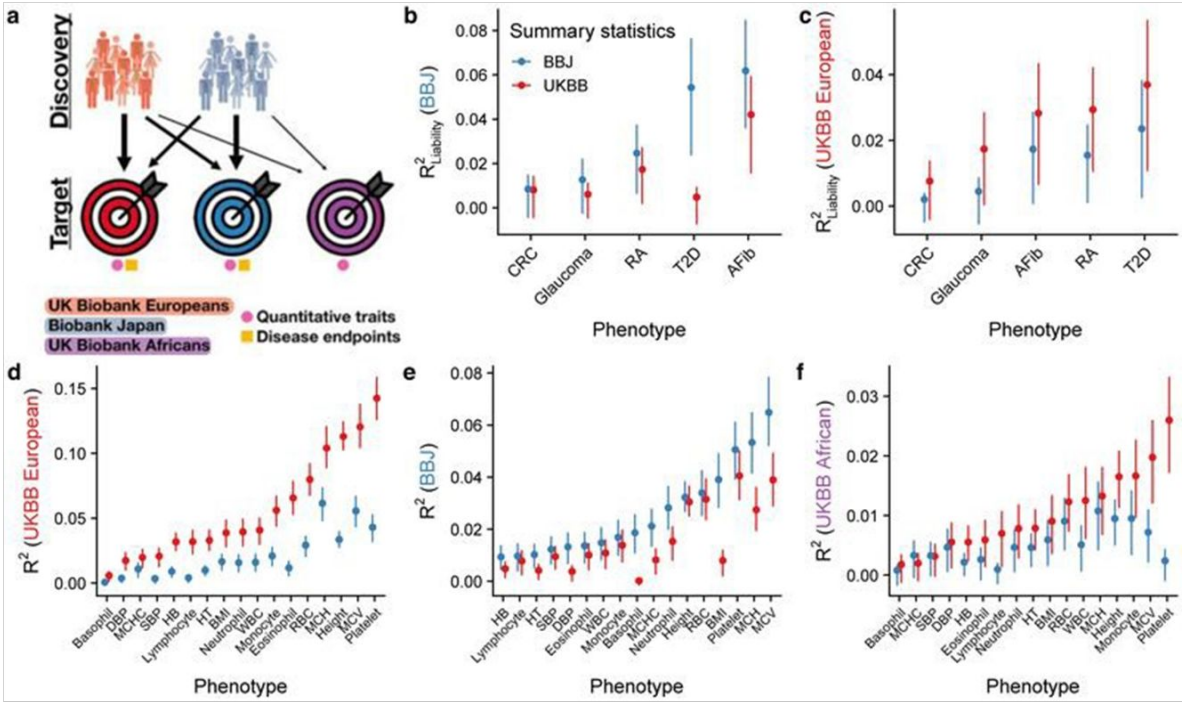


ARR (%)	1.5	3.8	2.9	5.4
RRR	0.52	0.60	0.46	0.55

Polygenic risk prediction accuracy : Ethnicity

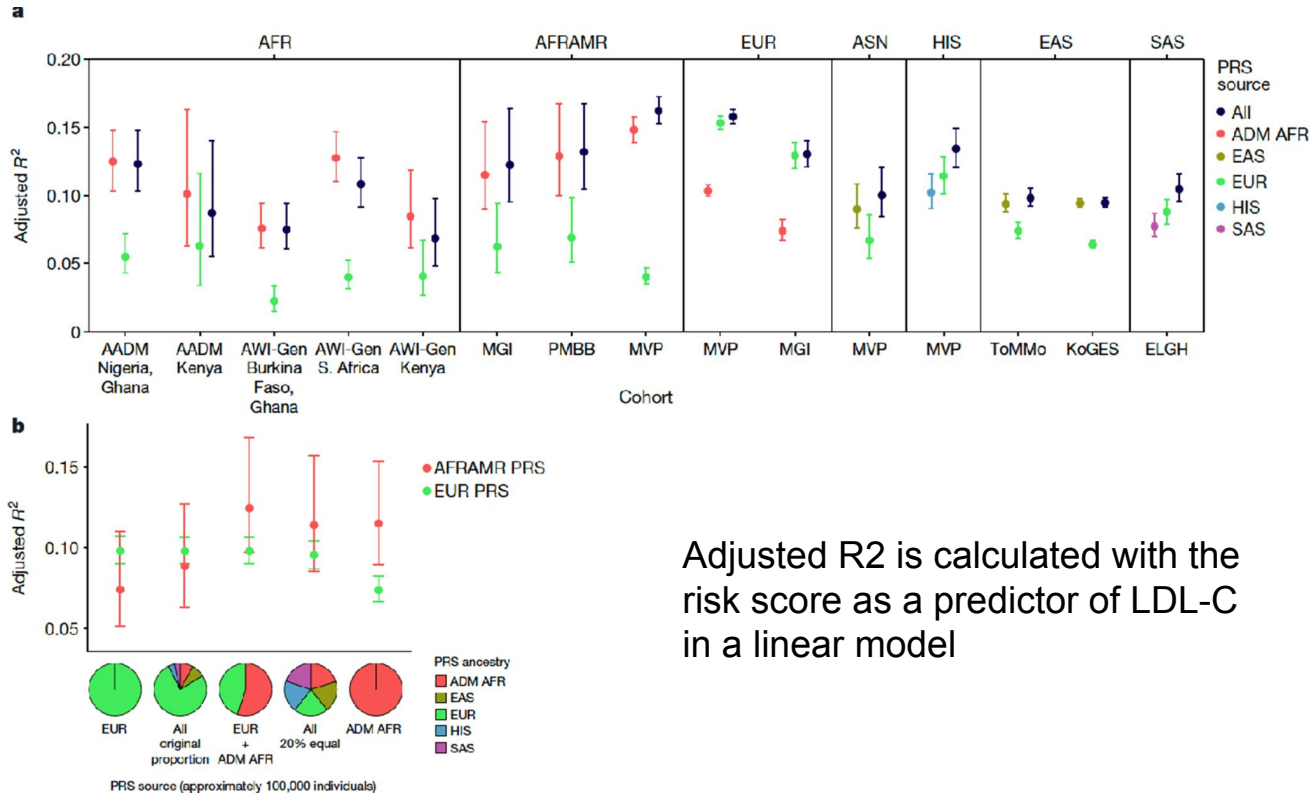






Nat Genet. 2019 Apr; 51(4): 584–591.

Multi-ancestry PRS show similar performance across ancestry



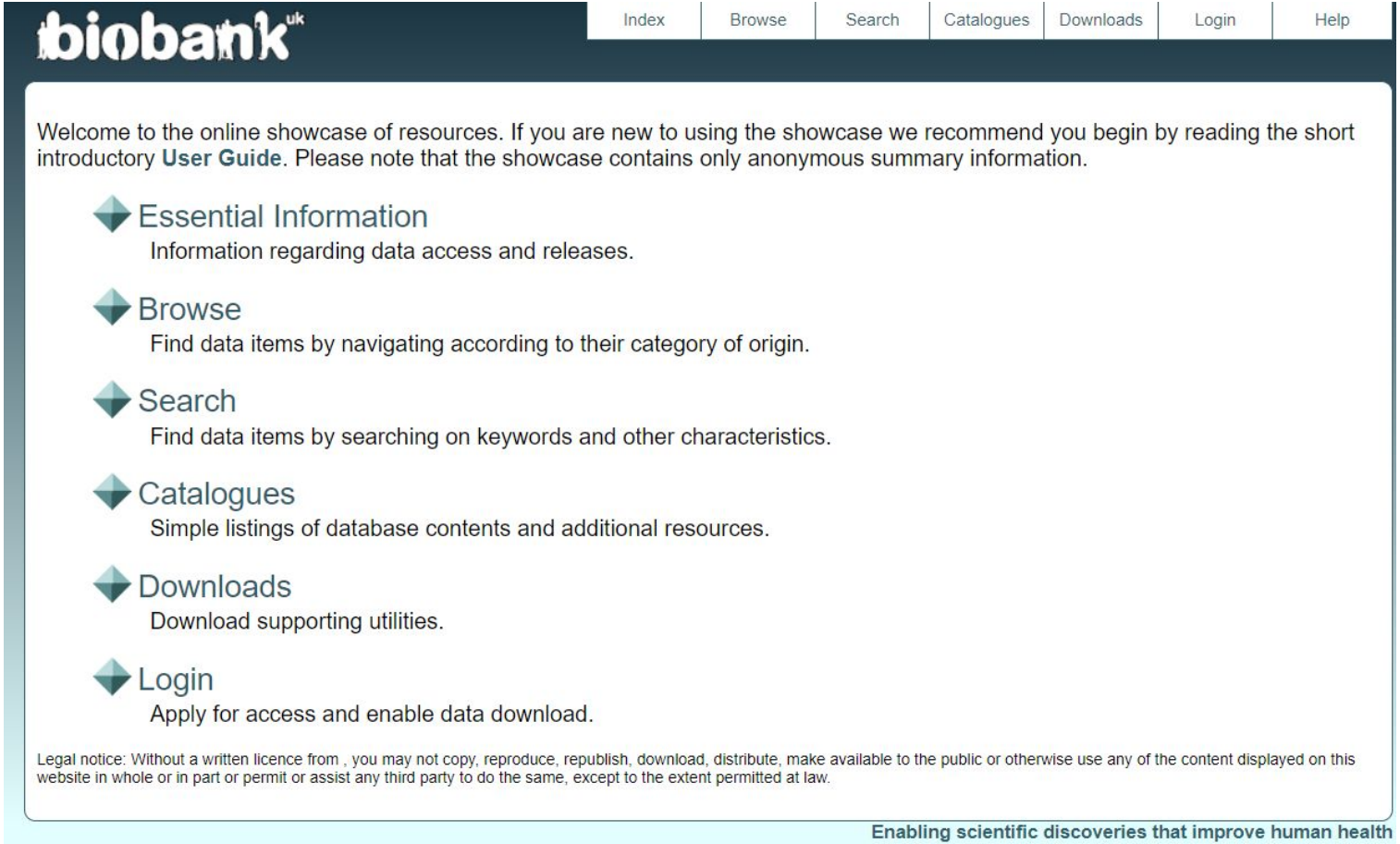
Adjusted R2 is calculated with the risk score as a predictor of LDL-C in a linear model

Nature. 600, 675–679 (2021)

GWAS Catalog: knowledgebase and deposition resource

- NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas)
- PheGenI (https://www.ncbi.nlm.nih.gov/gap/phegeni)
- Open Targets Genetics (https://genetics.opentargets.org/)
- HuGeAMP Knowledge Portals (https://hugeamp.org/)
- MRC IEU OpenGWAS (https://gwas.mrcieu.ac.uk/)
- PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk/)
- GWAS Central (https://www.gwascentral.org/)

Showcase of resources provided by the UK Biobank online



New data & enhancements to UK Biobank

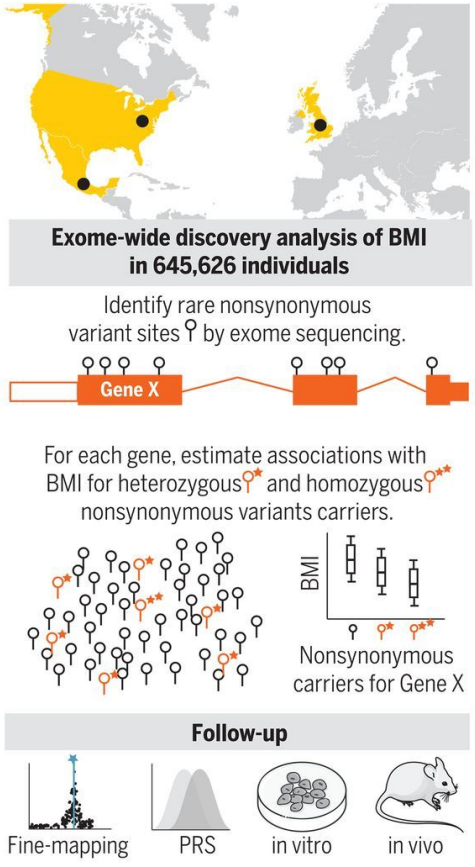
- **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- **Genetics:** Whole genome sequencing for all 500,000 participants, whole exome sequencing for 470,000 participants, genotyping (800,000 genome-wide variants and imputation to 90 million variants).
- **Health linkages:** Linkage to a wide range of electronic health-related records, including death, cancer, hospital admissions and primary care records.
- **Biomarkers**: Data on more than 30 key biochemistry markers from all participants, taken from samples collected at recruitment and the first repeat assessment.
- **Activity monitor:** Physical activity data over a 7-day period collected via a wrist-worn activity monitor for 100,000 participants plus a seasonal follow-up on a subset.
- **Online questionnaires:** Data on a range of exposures and health outcomes that are difficult to assess via routine health records, including diet, food preferences, work history, pain, cognitive function, digestive health and mental health.
- **Repeat baseline assessments:** A full baseline assessment is undertaken during the imaging assessment of 100,000 participants.
- **Samples:** Blood & urine was collected from all participants, and saliva for 100,000.

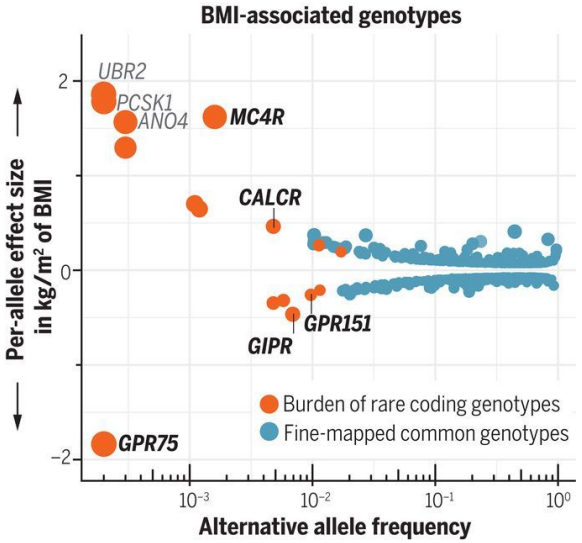
How genes affect human obesity

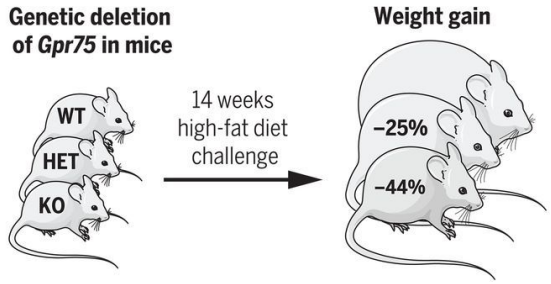
Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity

Exome sequencing–based discovery of BMI-associated genes.

(Left) Design for the discovery gene-burden analysis, with a depiction of follow-up analyses along the bottom. (Top right) Relationship between allele frequency and effect-size estimates for BMI-associated genotypes. (Bottom right) Weight gain for Gpr75+/+ (wild type, WT), Gpr75−/+ (heterozygous, HET), and Gpr75−/− (knockout, KO) mice during a high-fat diet challenge. PRS, polygenic risk score.



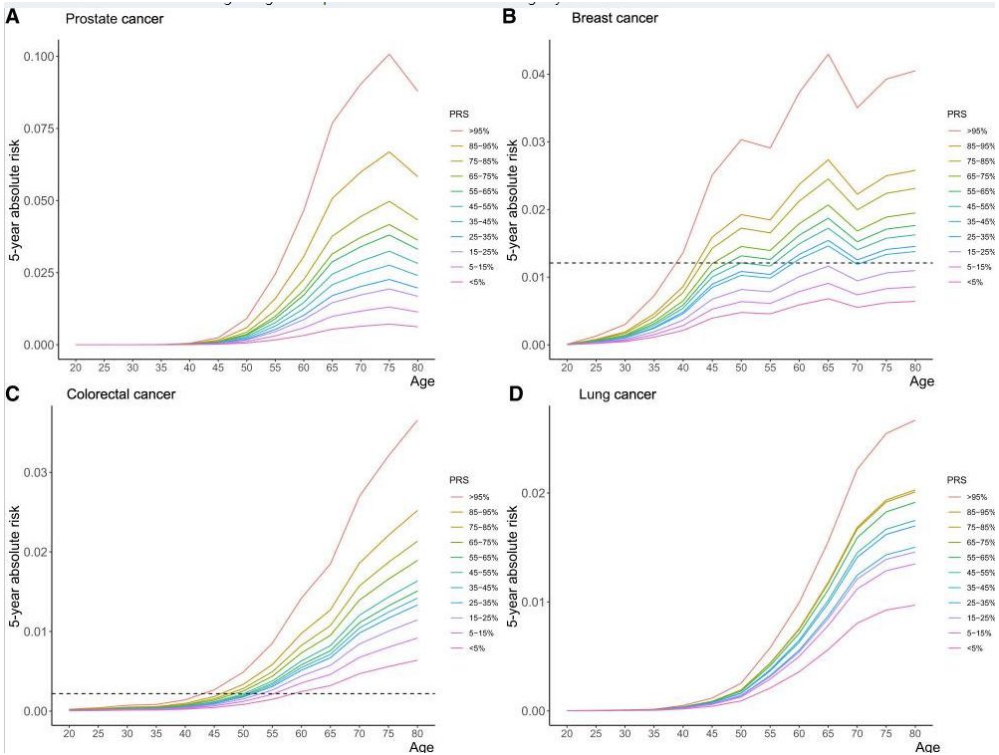




Parsa Akbari et al. ,Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity.Science373,eabf8683(2021).

Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers

Five-year absolute risks of site-specific cancers by PRS groups. Five-year absolute risk of developing cancer of (A) prostate, (B) breast, (C) colorectal, (D) and lung. The horizontal lines show the estimated 5-year risk for individuals with median PRS (45%-55%) at the age of 50 years for (B) breast cancer or (C) colorectal cancer. PRS = polygenic risk score.



Jia G, Lu Y, Wen W, Long J, Liu Y, Tao R, Li B, Denny JC, Shu XO, Zheng W. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. JNCI Cancer Spectr. 2020 Mar 12;4(3):pkaa021.

Free Access to UK Biobank GWAS Catalog

- **Access the Catalog:** Go to the UK Biobank GWAS Catalog website at https://www.ebi.ac.uk/gwas/ and navigate to the search page.
- **Search for Traits or Diseases:** Use the search bar or filters provided on the website to search for specific traits, diseases, or phenotypes of interest. You can also explore the available studies and associated data.
- **Browse Results:** Browse through the search results or study listings to find relevant GWAS studies related to your research interests. Each study entry typically includes information about the phenotype studied, associated genetic variants, and links to relevant publications.
- **View Study Details:** Click on the title or entry of a specific study to view more detailed information about the GWAS, including study design, sample size, statistical methods, significant genetic variants, and other relevant details.

Create a Polygenic Risk Score (PRS) for a specific disease of interest utilizing UK Biobank GWAS Catalog

- **Data Collection:** Collect genetic information and DNA sequence data related to the disease of interest. Utilize the UK Biobank GWAS Catalog or other publicly available databases to access relevant GWAS data associated with the disease.
- **Gene Selection:** Select genes associated with the disease based on the GWAS findings available in the UK Biobank GWAS Catalog or other relevant resources. These genes can be identified through significant associations with the disease in previous studies.
- **Assign Gene Weights:** Assign weights to the selected genes based on their effect sizes from the GWAS results. Utilize statistical models to calculate the PRS considering the contribution of each gene.
- **PRS Evaluation:** Use the calculated PRS to predict and evaluate the disease risk for specific individuals or populations. Assess the predictive performance of the PRS using relevant metrics such as sensitivity, specificity, and area under the curve (AUC).
- **Validation:** Validate the PRS using independent datasets or through cross-validation techniques to ensure its effectiveness and reliability in predicting disease risk.

Resources helpful for conducting PRS using GWAS catalog

- **ComPaSS-GWAS:** an alternative method for replication in GWAS studies, which can reduce type I errors when appropriate replication data are not available.
- **r2VIM:** This resource offers a recurrency-based variable selection method in random forests specifically designed for genome-wide genetic association studies.
- **Tiled Regression Analysis:** This software framework can assist in selecting a set of genetic predictors that explain trait variation using an additive regression model. This can be useful for identifying relevant genetic variants to include in PRS analysis.

Lecture Summary

1. Introduction to Genomics and the UK Biobank

- The lecture begins with an introduction to the basics of genomics, covering essential concepts such as DNA, genes, and genome sequencing. It also introduces the UK Biobank, highlighting its role as a significant resource in genomic research, particularly in how it collects and utilizes vast arrays of genetic data to advance health informatics.

2. Integration of Genomics with Health Informatics

- This section discusses the integration of genomic data with health informatics, demonstrating how such data can enhance healthcare outcomes through improved disease prediction and personalized medicine. It emphasizes the value of genomic data in understanding complex diseases and developing targeted treatments.

3. Methodologies and Applications

- The lecture details various methodologies used in genomic research, such as genome-wide association studies (GWAS) and polygenic risk scoring. It explores practical applications of these methodologies using data from the UK Biobank, showcasing real-world examples of how genomic research contributes to advancements in disease prediction and prevention.

4. Advances and Innovations in Genomic Research

- Advances in genomic technologies and research are covered, including the impact of the Human Genome Project and subsequent innovations in sequencing and data analysis. The section also highlights how these advances have enabled researchers to uncover complex genetic interactions and their implications for disease mechanisms.

5. Challenges and Future Directions

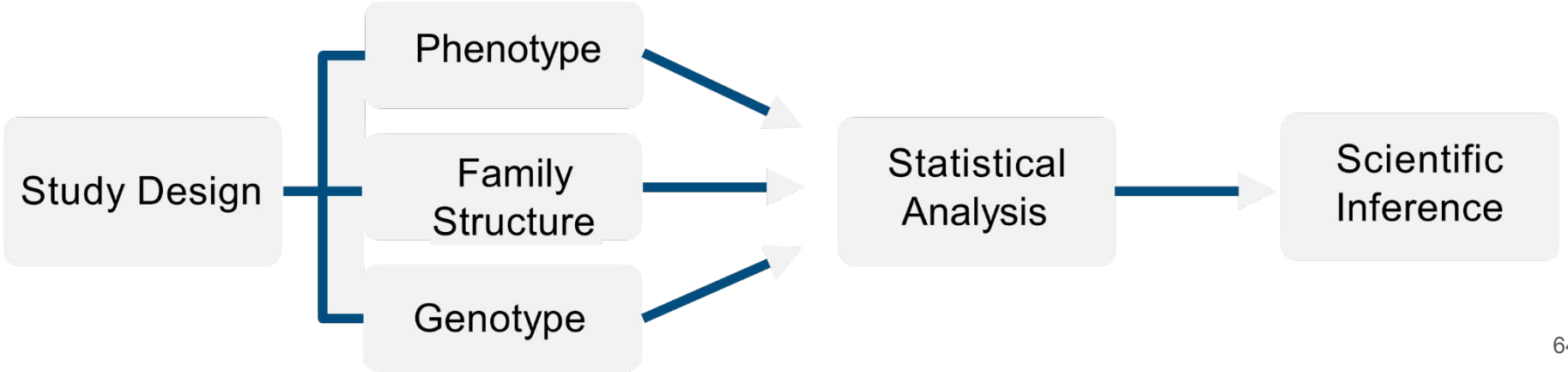
- The lecture addresses several challenges facing genomic research, such as ethical issues, the need for large and diverse datasets, and the technical challenges of data integration and analysis. It also discusses future directions, including the potential for genomics to further revolutionize personalized medicine, and the ongoing efforts to enhance genomic databases like the UK Biobank for broader research applications.

Conclusion

We outline the comprehensive approaches used in genomic research, starting with study design. The research design is pivotal as it sets the foundation for data collection and analysis methods. This applies directly to three main components: :Phenotype, family structure and genotype.

All elements will then be subjected to statistical analysis across the board to engage the data, controlling for various confounding factors and extracting meaningful patterns. This analysis is very important because it is very promising and ultimately leads to conclusions that can inform further research, and clinical applications.

This structured approach allows us to leverage the reliability and validity of our research findings to contribute to a broader understanding of genomic research.



References

Akbari, P., & others. (2021). Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. Science, 373(6550), Article eabf8683. https://doi.org/10.1126/science.abf8683

Bush, W. S., Oetjens, M. T., & Crawford, D. C. (2016). Unravelling the human genome–phenome relationship using phenome-wide association studies. Nature Reviews Genetics, 17(3), 129-145. https://doi.org/10.1038/nrg.2015.36

Jia, G., Lu, Y., Wen, W., Long, J., Liu, Y., Tao, R., Li, B., Denny, J. C., Shu, X. O., & Zheng, W. (2020). Evaluating the utility of polygenic risk scores in identifying high-risk individuals for eight common cancers. JNCI Cancer Spectrum, 4(3), pkaa021. https://doi.org/10.1093/jncics/pkaa021. PMID: 32596635; PMCID: PMC7306192.

Khera, A. V., Chaffin, M., Aragam, K. G., & others. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics, 50, 1219-1224. https://doi.org/10.1038/s41588-018-0183-z

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics, 16(5), 275-284. https://doi.org/10.1038/nrg3908

Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., & Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. Nature Reviews Genetics, 19(9), 566-580. https://doi.org/10.1038/s41576-018-0020-3

Suhre, K., & Gieger, C. (2012). Genetic variation in metabolic phenotypes: study designs and applications. Nature Reviews Genetics, 13(11), 759-769. https://doi.org/10.1038/nrg3314