Heritability, complex diseases, genome-wide association studies, polygenic risk scores

Lecture plan

- Non-Mendelian inheritance: Mendelians vs. biometricians.
- Basics of quantitative genetics. Heritability.
- Liability threshold model
- Monogenic vs. complex disorders. Allelic architecture of genetic disorders
- Genome-wide association studies
- Polygenic risk scores

Non-Mendelian inheritance

Dichotomous (binary) phenotypes; complex (common, multifactorial) diseases

- Diabetes
- Schizophrenia
- Coronary artery disease

Quantitative phenotypes

- Height
- Body-mass index
- Blood lipid levels
- Blood pressure

Non-Mendelian inheritance

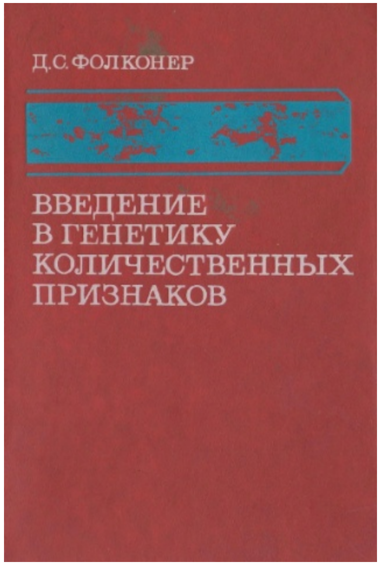
A controversy between Mendelians and biometricians (1900-1918)

- 1865 Gregor Mendel: “Versuche über Pflanzen-Hybriden” (“Experiments in plant hybridization”),
- 1865 Francis Galton: “Hereditary Talent and Character”

Biometricians: most of the characters likely to be important in evolution (fertility, body size, strength, and skill in catching prey or finding food) were continuous or quantitative characters and *not* amenable to Mendelian analysis

Controversy resolved:

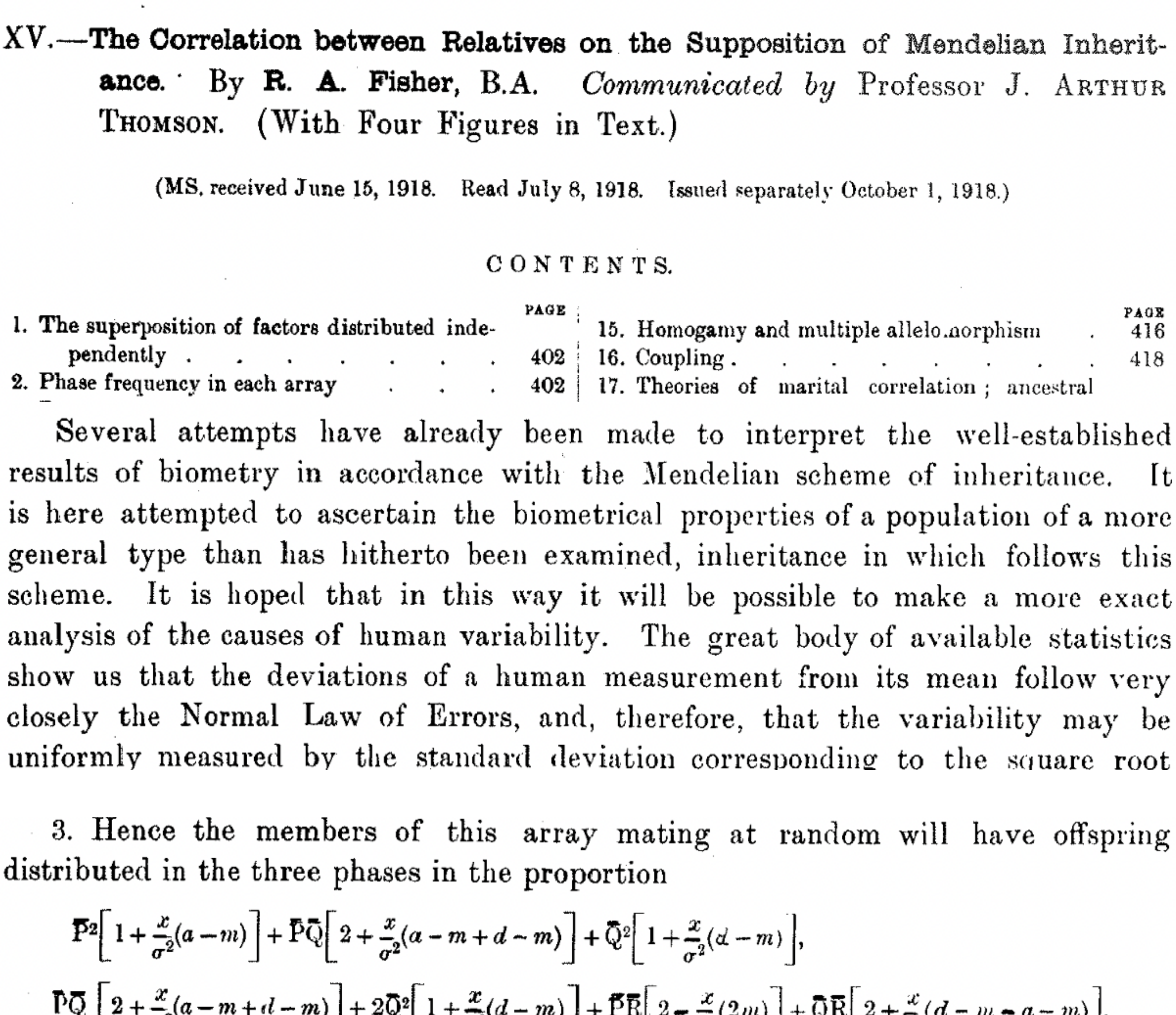
- 1918 R.A.Fisher: “The Correlation between Relatives on the Supposition of Mendelian Inheritance”
- 1965 D.S.Falconer: “The inheritance of liability to certain diseases, estimated from the incidence among relatives”



Д.С.ФОЛКОНЕР

ВВЕДЕНИЕ
В ГЕНЕТИКУ
КОЛИЧЕСТВЕННЫХ
ПРИЗНАКОВ

Non-Mendelian inheritance



XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher,** B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

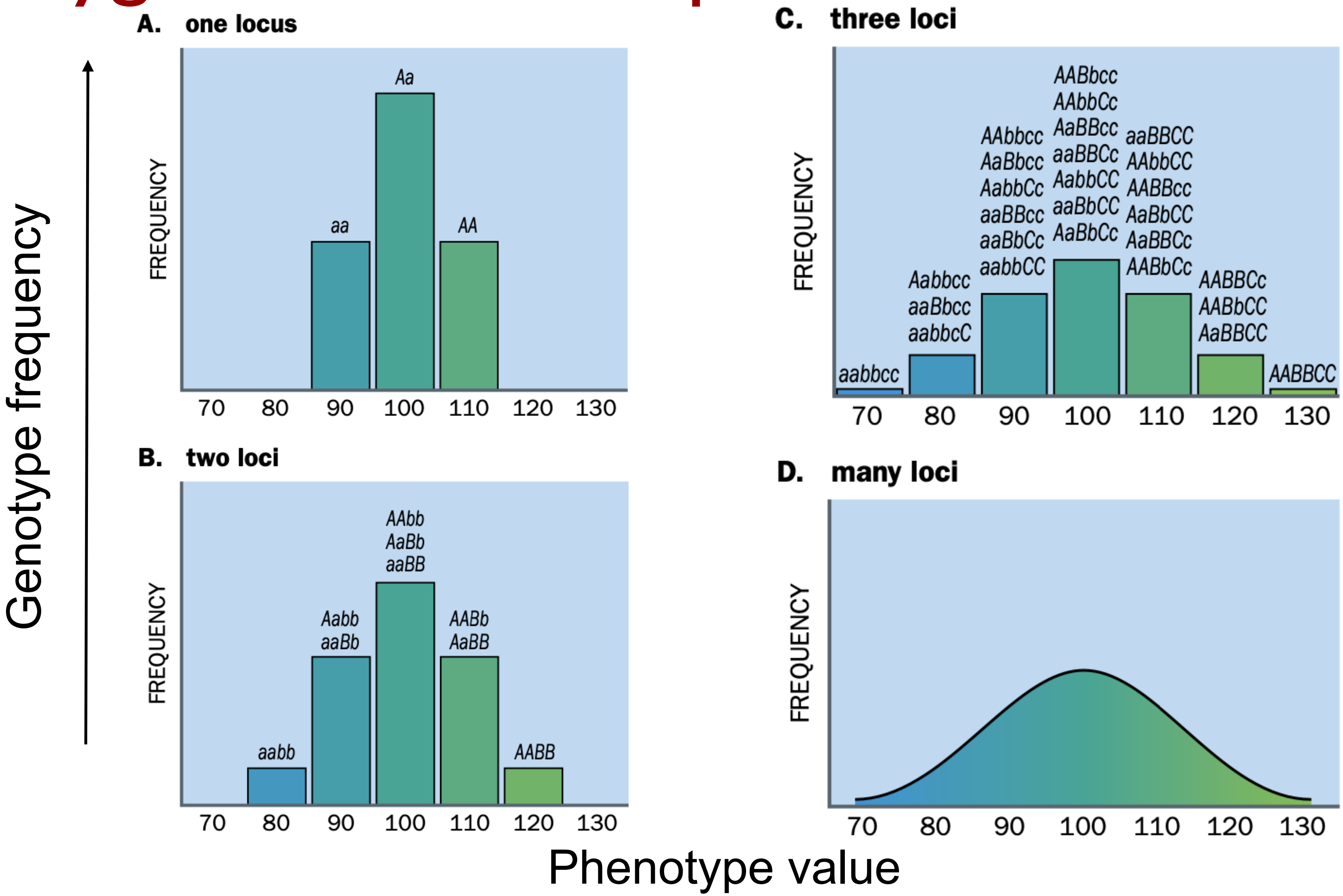
Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this scheme. It is hoped that in this way it will be possible to make a more exact analysis of the causes of human variability. The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root

3. Hence the members of this array mating at random will have offspring distributed in the three phases in the proportion

$$\bar{P}^2\left[1+\frac{x}{\sigma^2}(a-m)\right]+\bar{P}\bar{Q}\left[2+\frac{x}{\sigma^2}(a-m+d-m)\right]+\bar{Q}^2\left[1+\frac{x}{\sigma^2}(d-m)\right],$$

$$\bar{P}\bar{Q}\left[2+\frac{x}{\sigma^2}(a-m+d-m)\right]+2\bar{Q}^2\left[1+\frac{x}{\sigma^2}(d-m)\right]+\bar{P}\bar{R}\left[2-\frac{x}{\sigma^2}(2m)\right]+\bar{Q}\bar{R}\left[2+\frac{x}{\sigma^2}(d-m-a-m)\right],$$

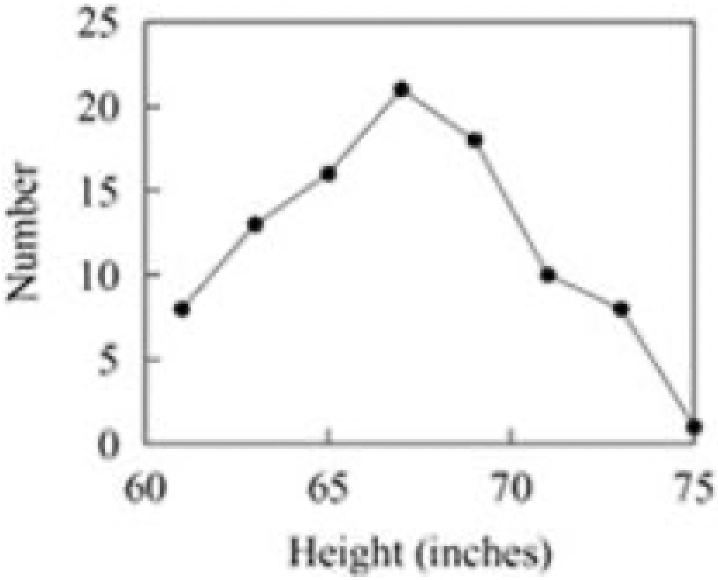
Polygenic nature of quantitative phenotypes



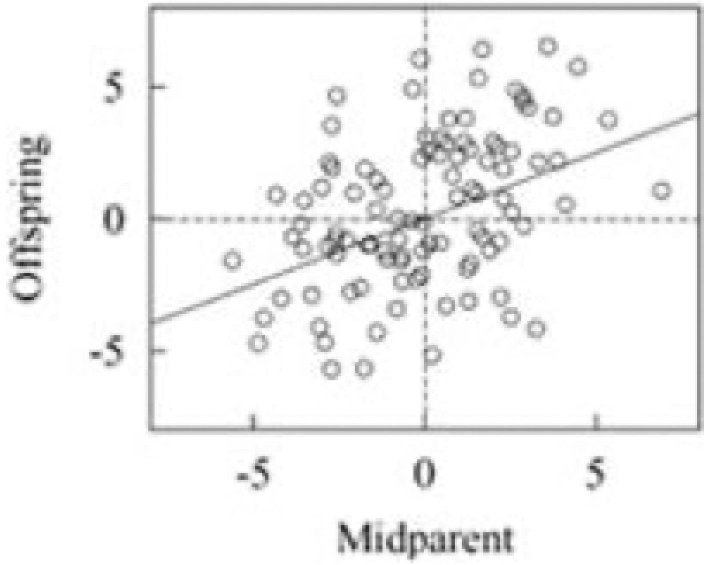
A hypothetical continuous character, mean = 100 units; additive (codominant) effects: each copy of A = +5 units, a = -5 units. All allele frequencies are 0.5.

Strachan, Read – *Human Molecular Genetics*

Example of a trait: human height



Histogram



Deviation from the population mean, in

Observations:

- The correlation coefficient $r^2 = 0.476 > 0$ → Relatives resemble each other → the trait has genetic component
- Scatter due to Mendelian segregation and environmental effects
- The coefficient $r^2 < 1$ → "regression to the mean"

Strachan, Read – *Human Molecular Genetics*

Phenotypic variance and heritability

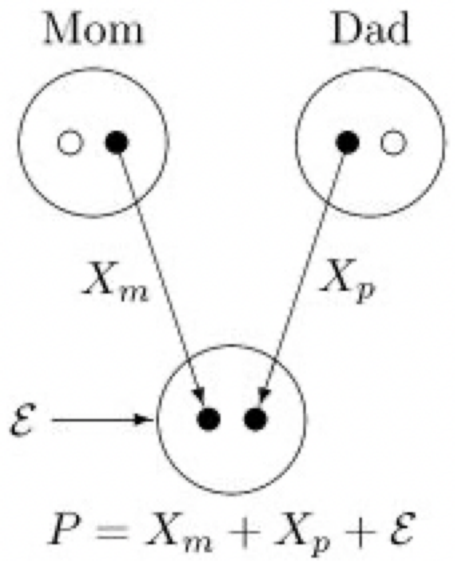
Phenotype value P

$$P = X_m + X_p + \epsilon$$

where $P = P_{individual} - P_{pop.mean}$ — deviation of the phenotype of an individual from the population mean, $X_{m,p}$ and ϵ are [normally distributed] random variables. Denote:

$$V(X_m) = V(X_p) = V_a/2,\ V(\epsilon) = V_e$$

J.Gillespie: "Quantitative genetics is all about variances, covariances, and correlations"



Phenotypic variance:

$$V_p = V(X_m) + V(X_p) + V(\epsilon) + 2C(X_m, X_p) + 2C(X_m, \epsilon) + 2C(X_p, \epsilon)$$

where $V(.)$ denotes variance, $C(\cdot,\cdot)$ covariance, $C(X_m, X_p) = 0$ if parents are not related, and $C(X, \epsilon)$ is the genotype-environment interactions, assumed to be zero.

An example of a genotype-environment interaction:

One allele adds +1 cm to the phenotype in a warm environment and -1 cm in a cold environment; another allele does exactly the opposite

Note: $C(X, X) = V(X) = V_a/2$

Gillespie – *Population genetics. A concise guide*

Phenotypic variance and heritability

$$V_p = V(X_m)+V(X_p)+V(\epsilon)+2C(X_m,X_p)+2C(X_m,\epsilon)+2C(X_p,\epsilon) = V_a/2+V_a/2+V(\epsilon)$$

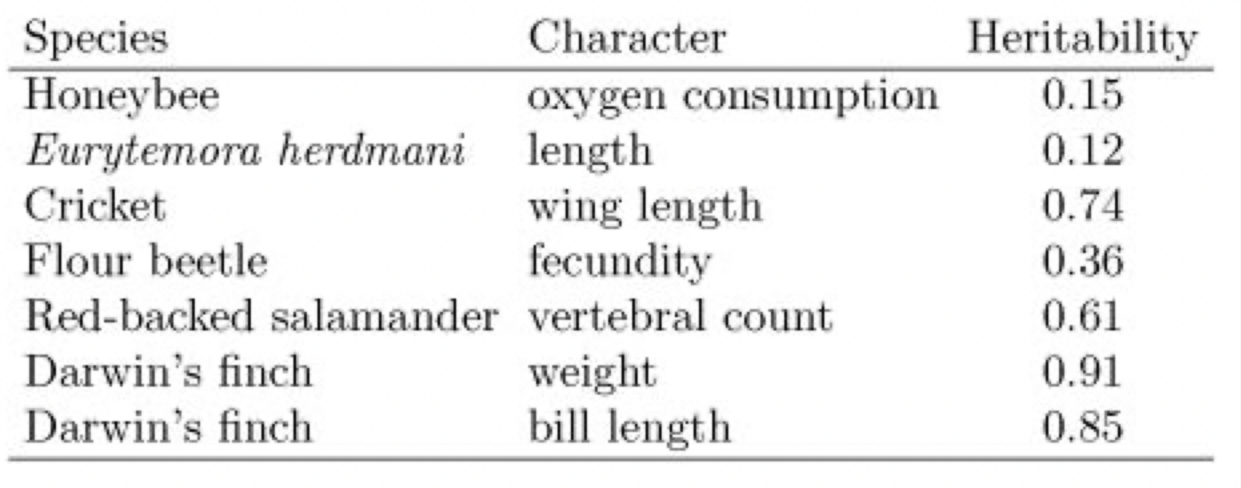
Phenotypic variance is the sum of the **additive** and **environmental** variances:

$$V_p = V_a + V_e$$

The [narrow-sense] heritability of the trait h^2:

$$h^2 = \frac{V_a}{V_p} = \frac{V_a}{V_a + V_e}$$

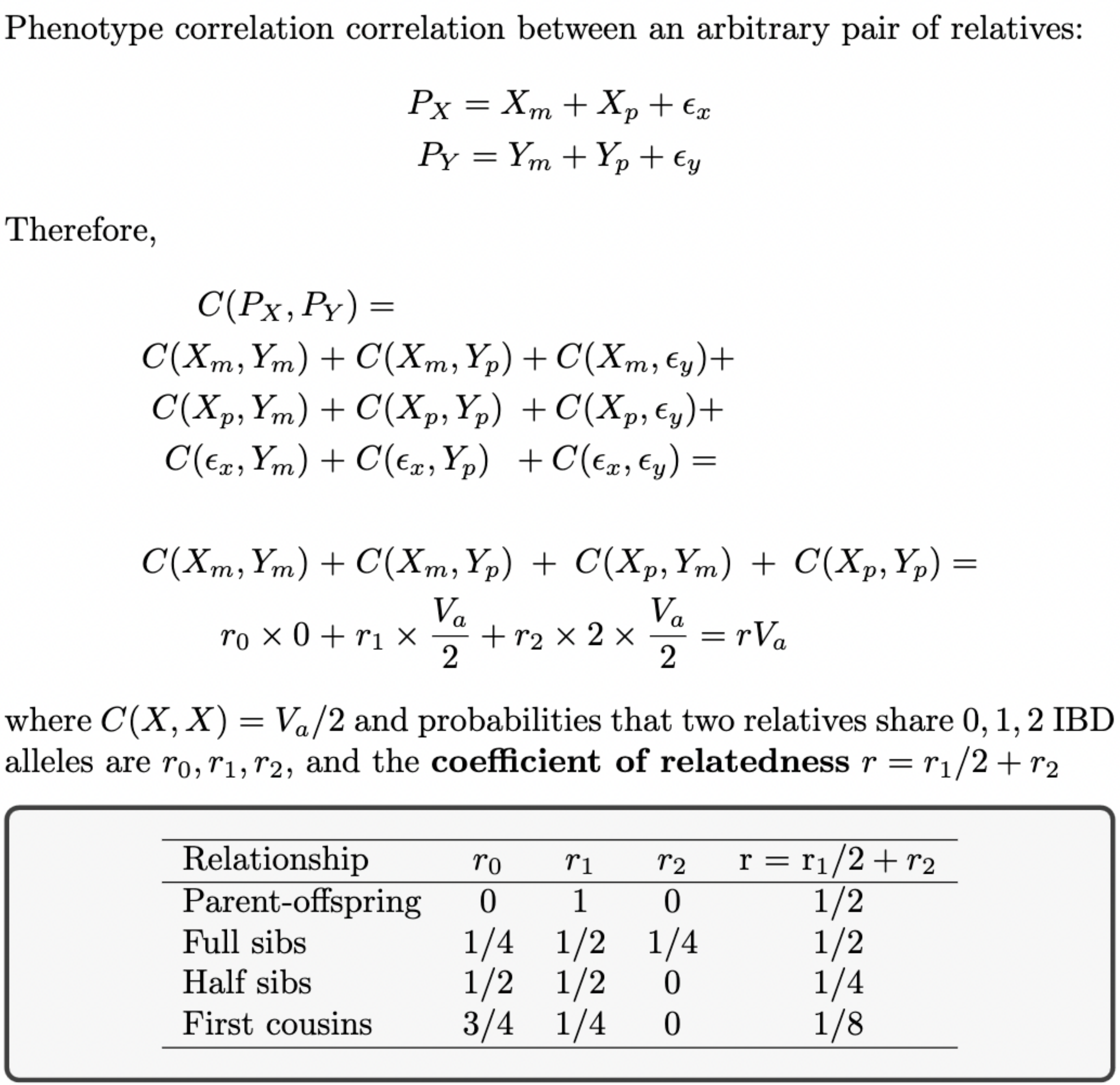
Additive variance = the genetic contribution is a simple sum of the contribution from each allele which do not interact



Species	Character	Heritability
Honeybee	oxygen consumption	0.15
Eurytemora herdmani	length	0.12
Cricket	wing length	0.74
Flour beetle	fecundity	0.36
Red-backed salamander	vertebral count	0.61
Darwin's finch	weight	0.91
Darwin's finch	bill length	0.85

Heritability estimates determined by parent-offspring correlations for a variety of traits and species taken from a paper by Mousseau and Roff (1987)

Phenotypic variance and heritability



Phenotype correlation correlation between an arbitrary pair of relatives:

$$P_X = X_m + X_p + \epsilon_x$$
$$P_Y = Y_m + Y_p + \epsilon_y$$

Therefore,

$$\begin{gathered} C(P_X, P_Y) = \\ C(X_m, Y_m) + C(X_m, Y_p) + C(X_m, \epsilon_y) + \\ C(X_p, Y_m) + C(X_p, Y_p) \;+ C(X_p, \epsilon_y) + \\ C(\epsilon_x, Y_m) + C(\epsilon_x, Y_p) \;\; + C(\epsilon_x, \epsilon_y) = \end{gathered}$$

$$\begin{gathered} C(X_m, Y_m) + C(X_m, Y_p) \; + \; C(X_p, Y_m) \; + \; C(X_p, Y_p) = \\ r_0 \times 0 + r_1 \times \frac{V_a}{2} + r_2 \times 2 \times \frac{V_a}{2} = rV_a \end{gathered}$$

where $C(X, X) = V_a/2$ and probabilities that two relatives share $0, 1, 2$ IBD alleles are r_0, r_1, r_2, and the **coefficient of relatedness** $r = r_1/2 + r_2$

Relationship	r_0	r_1	r_2	r = r$_1/2 + r_2$
Parent-offspring	0	1	0	1/2
Full sibs	1/4	1/2	1/4	1/2
Half sibs	1/2	1/2	0	1/4
First cousins	3/4	1/4	0	1/8

Phenotypic variance and heritability

$$C(P_X, P_Y) = rV_a$$

Recall that $Corr(x,y) = \frac{C(x,y)}{\sqrt{V(x)V(y)}}$ and $V(X) = V_a/2$, therefore

$$Corr(P_X, P_Y) = \frac{C(P_X, P_Y)}{V(P)} = \frac{rV_a}{V_p} = rh^2$$

$$Corr(P_X, P_Y) = rh^2$$

The correlation between [phenotypic values of] a pair of relatives is equal to the coefficient of relatedness times the heritability.

Therefore, given the deviation of the phenotype of an individual X from the population mean $P_X = x$, the expected phenotype of relative Y is

$$E(P_Y|P_X = x) = Corr(P_X, P_Y)x = rh^2x$$

Exercise: What is the correlation between each of the pairs of relatives in the table above if V_A = 2 and V_P = 3

Phenotypic variance and heritability

Earlier: simplified model with additive genetic effects only:

$$P = X_m + X_p + \epsilon$$
$$V_p = V_a + V_e$$

More realistic:

$$P = X_m + X_p + X_{mp} + \epsilon$$

where the new term, X_{mp}, captures the dominance relationships between the maternally and paternally derived alleles. Therefore, assuming that the additive and dominance contributions are uncorrelated,

$$V_p = V_a + V_d + V_e$$

where V_d is the so-called **dominance variance**. If we consider multiple and interacting loci, then

$$V_p = V_a + V_d + V_i + V_e$$

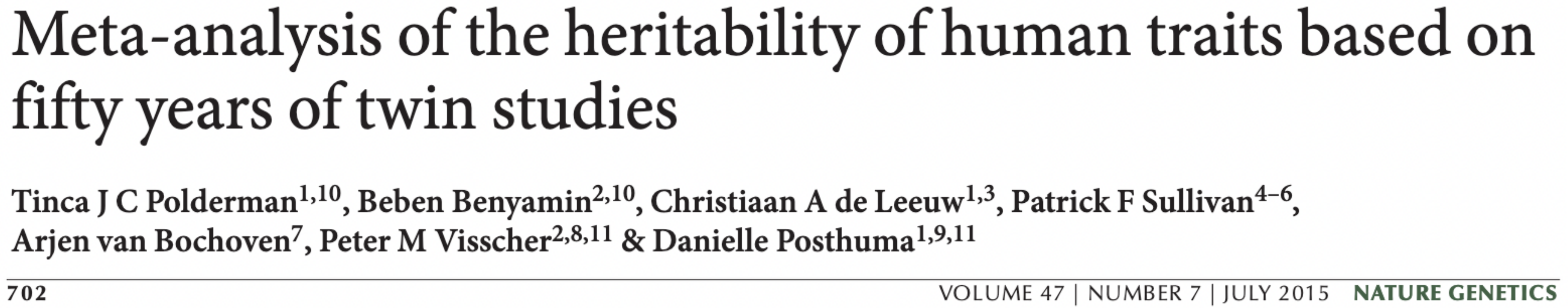
where V_i reflects the **epistatic variance** due to the interactions between loci. The [**narrow sense**] **heritability** h^2 is still

$$h^2 = \frac{V_a}{V_p}$$

but the **broad sense heritability** H^2 is the ratio of all of the genetic variances to the phenotypic variance,

$$H^2 = \frac{V_a + V_d + V_i}{V_p}$$

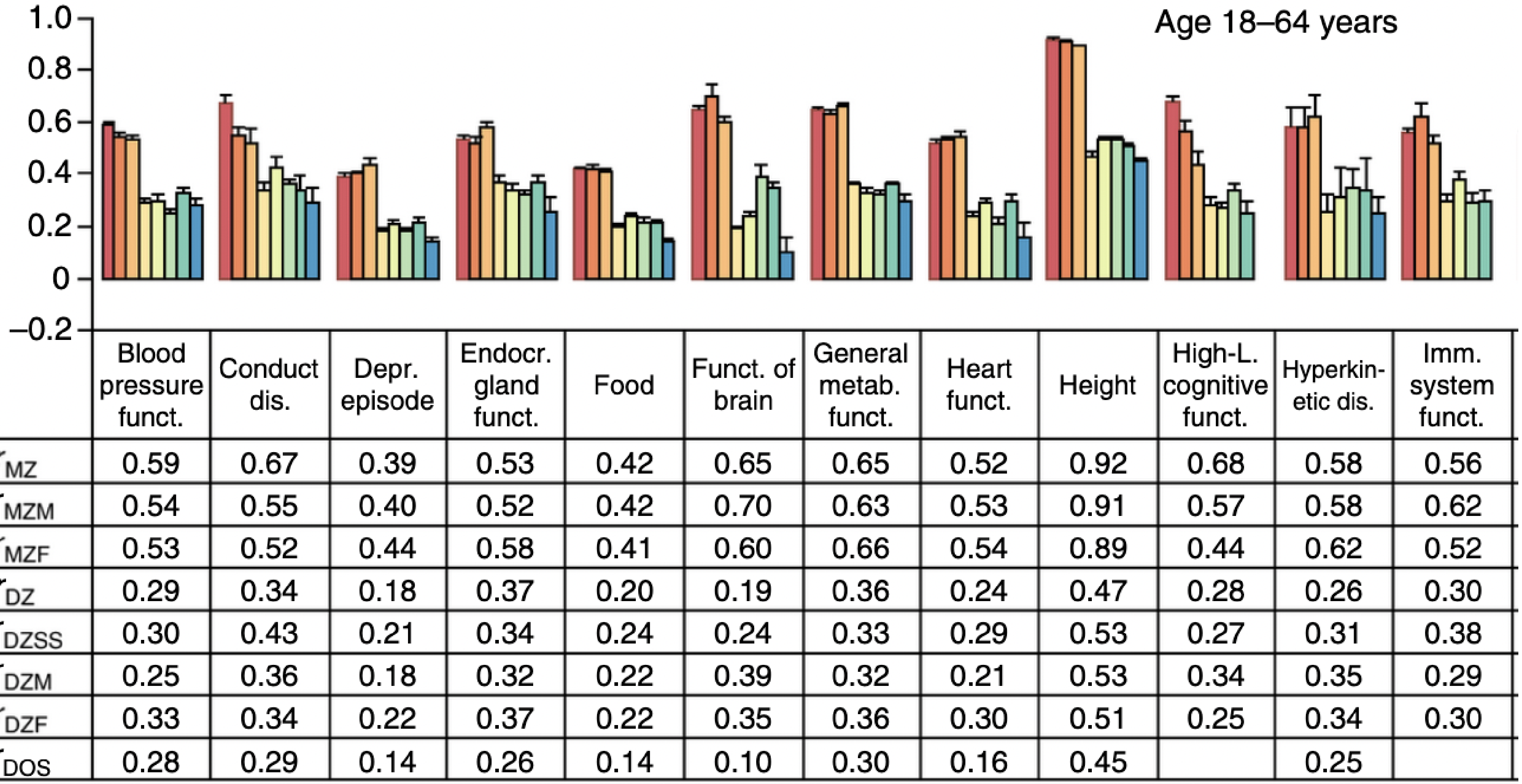
Exercise: In a population of beetles, the total variance of body weight is V_p = 130; the environmental variance is V_e = 35 and dominance genetic variance is V_d = 45. Assuming no epistatic effects, calculate heritability in the narrow sense



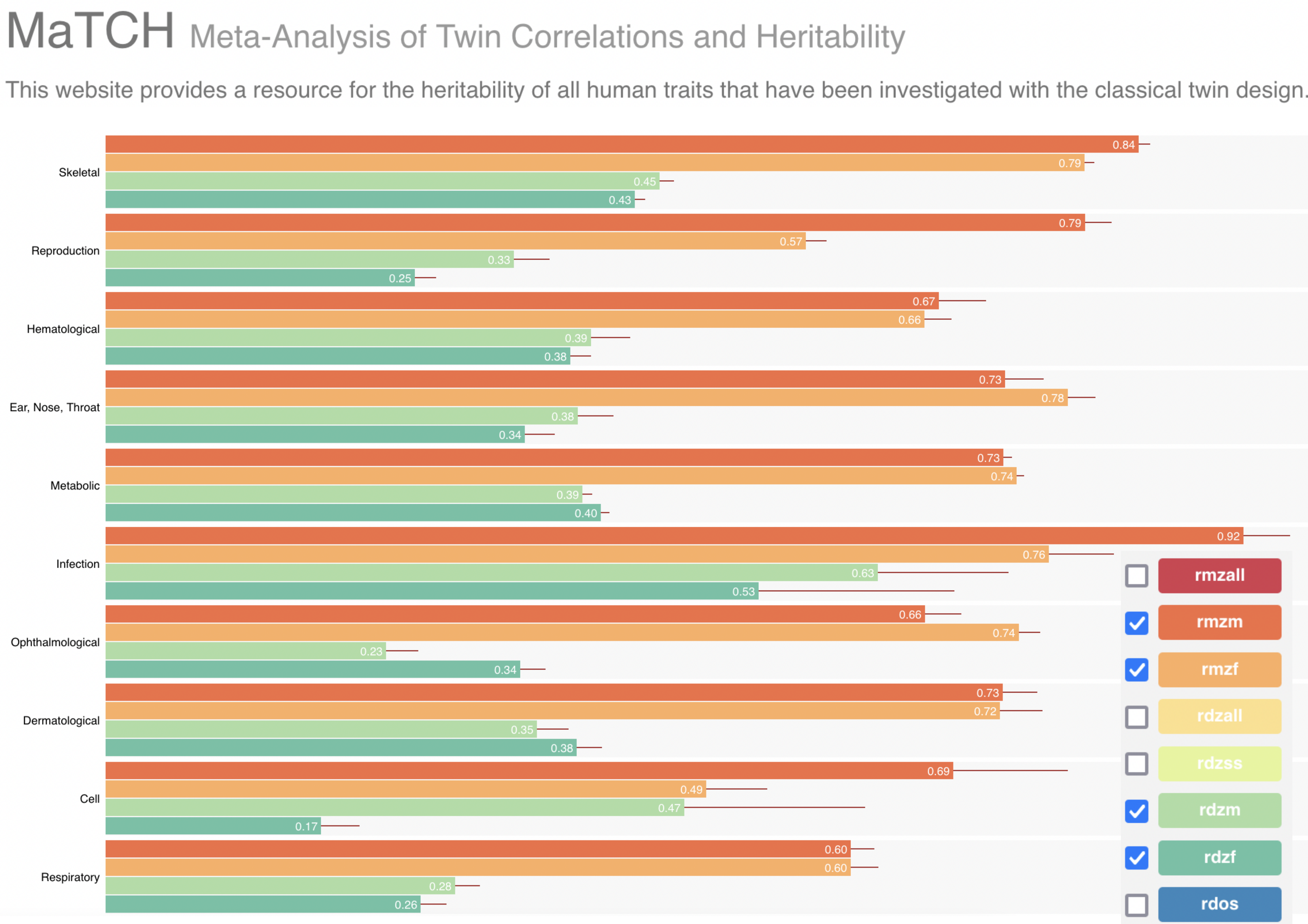
Meta-analysis of the heritability of human traits based on fifty years of twin studies

Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

702 VOLUME 47 | NUMBER 7 | JULY 2015 NATURE GENETICS



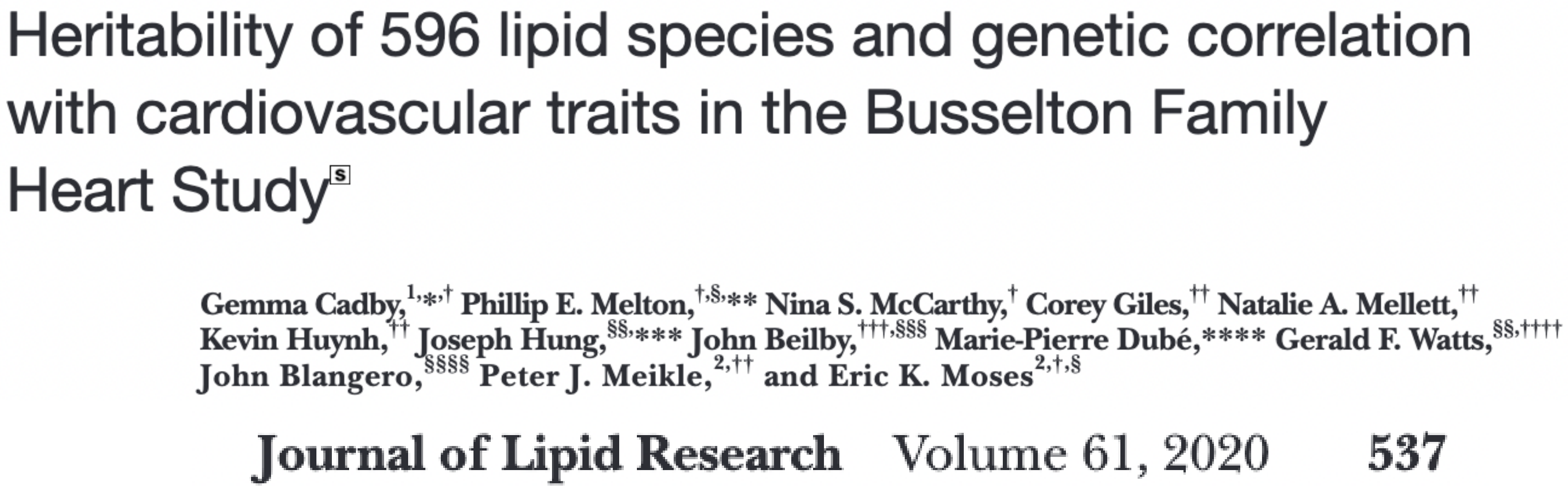
	Blood pressure funct.	Conduct dis.	Depr. episode	Endocr. gland funct.	Food	Funct. of brain	General metab. funct.	Heart funct.	Height	High-L. cognitive funct.	Hyperkin-etic dis.	Imm. system funct.
🟥 r_{MZ}	0.59	0.67	0.39	0.53	0.42	0.65	0.65	0.52	0.92	0.68	0.58	0.56
🟧 r_{MZM}	0.54	0.55	0.40	0.52	0.42	0.70	0.63	0.53	0.91	0.57	0.58	0.62
🟧 r_{MZF}	0.53	0.52	0.44	0.58	0.41	0.60	0.66	0.54	0.89	0.44	0.62	0.52
🟨 r_{DZ}	0.29	0.34	0.18	0.37	0.20	0.19	0.36	0.24	0.47	0.28	0.26	0.30
🟨 r_{DZSS}	0.30	0.43	0.21	0.34	0.24	0.24	0.33	0.29	0.53	0.27	0.31	0.38
🟩 r_{DZM}	0.25	0.36	0.18	0.32	0.22	0.39	0.32	0.21	0.53	0.34	0.35	0.29
🟩 r_{DZF}	0.33	0.34	0.22	0.37	0.22	0.35	0.36	0.30	0.51	0.25	0.34	0.30
🟦 r_{DOS}	0.28	0.29	0.14	0.26	0.14	0.10	0.30	0.16	0.45		0.25	



MaTCH Meta-Analysis of Twin Correlations and Heritability

This website provides a resource for the heritability of all human traits that have been investigated with the classical twin design.

https://match.ctglab.nl/#/multiple/twin_cor



Heritability of 596 lipid species and genetic correlation with cardiovascular traits in the Busselton Family Heart Study[S]

Gemma Cadby,[1,*,†] Phillip E. Melton,[†,§,**] Nina S. McCarthy,[†] Corey Giles,[††] Natalie A. Mellett,[††] Kevin Huynh,[††] Joseph Hung,[§§,***] John Beilby,[†††,§§§] Marie-Pierre Dubé,[****] Gerald F. Watts,[§§,††††] John Blangero,[§§§§] Peter J. Meikle,[2,††] and Eric K. Moses[2,†,§]

Journal of Lipid Research Volume 61, 2020 **537**

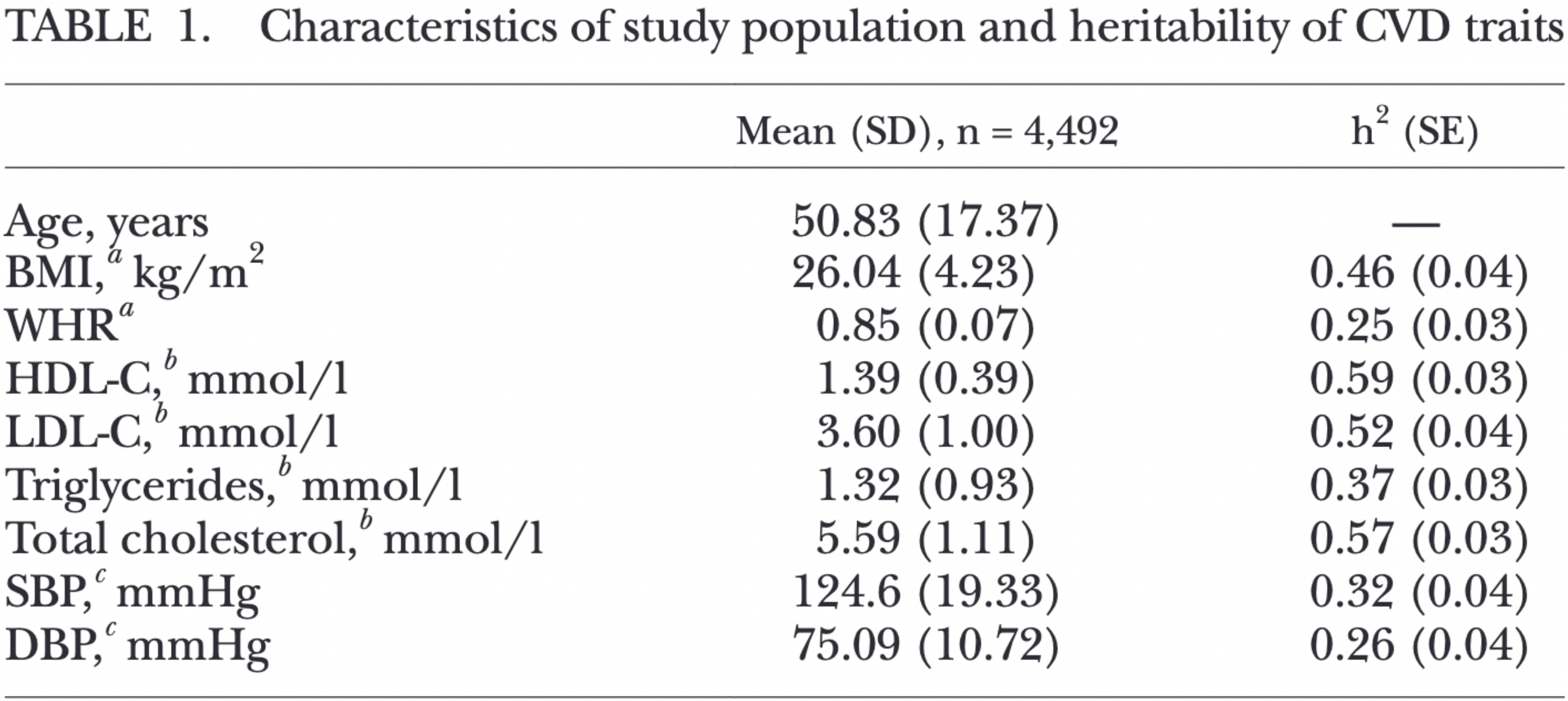


TABLE 1. Characteristics of study population and heritability of CVD traits

	Mean (SD), n = 4,492	h^2 (SE)
Age, years	50.83 (17.37)	—
BMI,[a] kg/m^2	26.04 (4.23)	0.46 (0.04)
WHR[a]	0.85 (0.07)	0.25 (0.03)
HDL-C,[b] mmol/l	1.39 (0.39)	0.59 (0.03)
LDL-C,[b] mmol/l	3.60 (1.00)	0.52 (0.04)
Triglycerides,[b] mmol/l	1.32 (0.93)	0.37 (0.03)
Total cholesterol,[b] mmol/l	5.59 (1.11)	0.57 (0.03)
SBP,[c] mmHg	124.6 (19.33)	0.32 (0.04)
DBP,[c] mmHg	75.09 (10.72)	0.26 (0.04)

Linear modeling framework

Linear model with additive genetic effects:

$$\underset{\text{Phenotype}}{\underline{\;Y\;}} = \underset{\text{Genotype}}{\underline{\;G\;}} + \underset{\text{Environment}}{\underline{\;E\;}} + \underset{\text{Random}}{\underline{\;\epsilon\;}}$$

Linear modeling framework

Linear model with additive genetic effects:

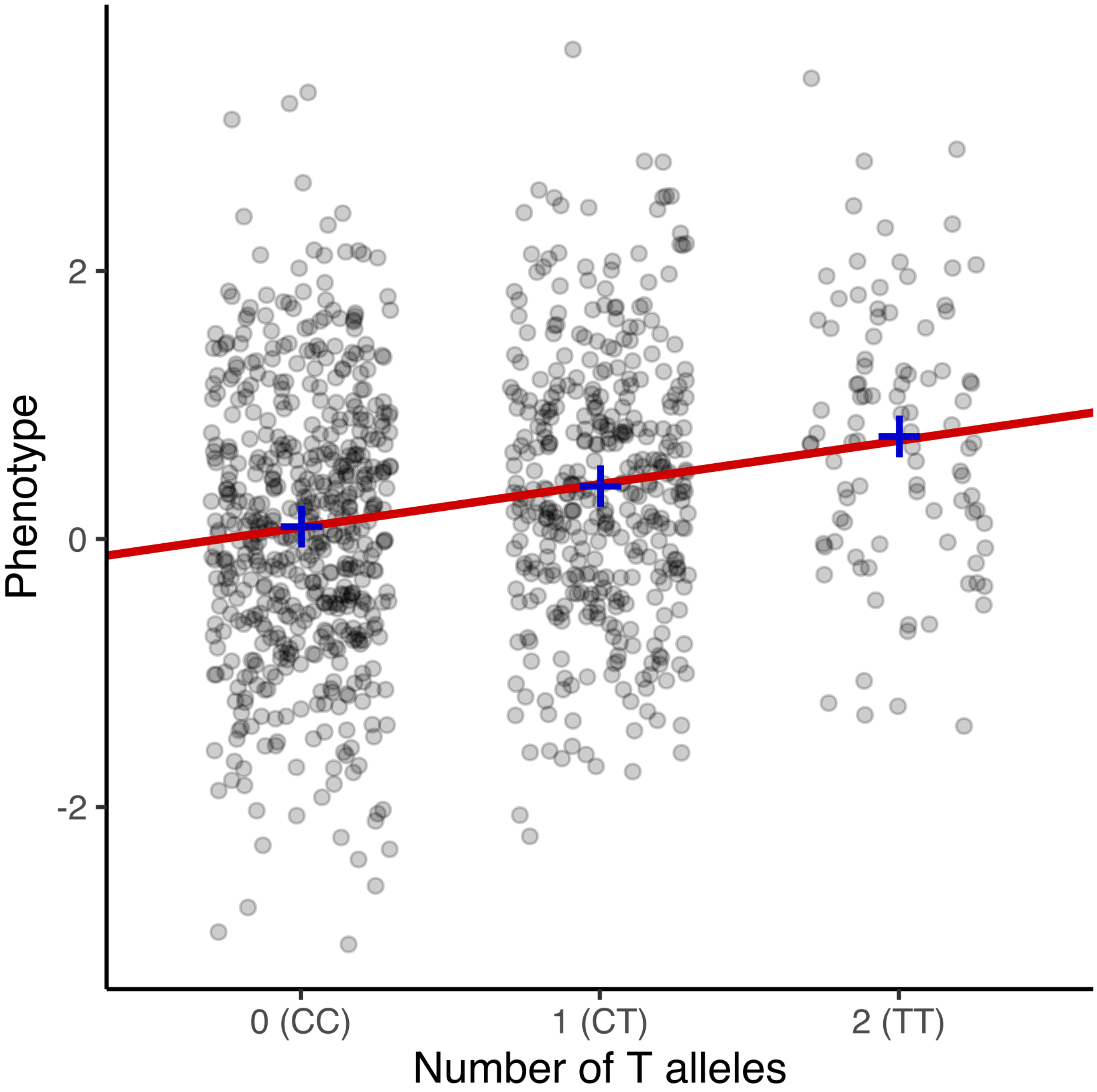
$$\underset{\text{Phenotype}}{\underline{\;Y\;}} \;=\; \underset{\text{Genotype}}{\underline{\;G\;}} + \underset{\text{Environment}}{\underline{\;E\;}} + \underset{\text{Random}}{\underline{\;\epsilon\;}}$$

$$Y_i = \mu + \beta G_i + \sum_j \alpha_j X_{ij} + \epsilon_i$$

- Y_i – phenotype of the *i*-th individual
- μ – baseline phenotype value
- G_i – genotype, effector allele a count: *G*(AA)=0, *G*(Aa)=1, *G*(aa)=2
- β – effect of the effector allele: change in value for each copy of effector allele (quantitative trait) or log odds ratio (binary trait)
- $\alpha_k X_{ij}$ – covariates: age, sex, smoking, medication use, ethnicity, other variants

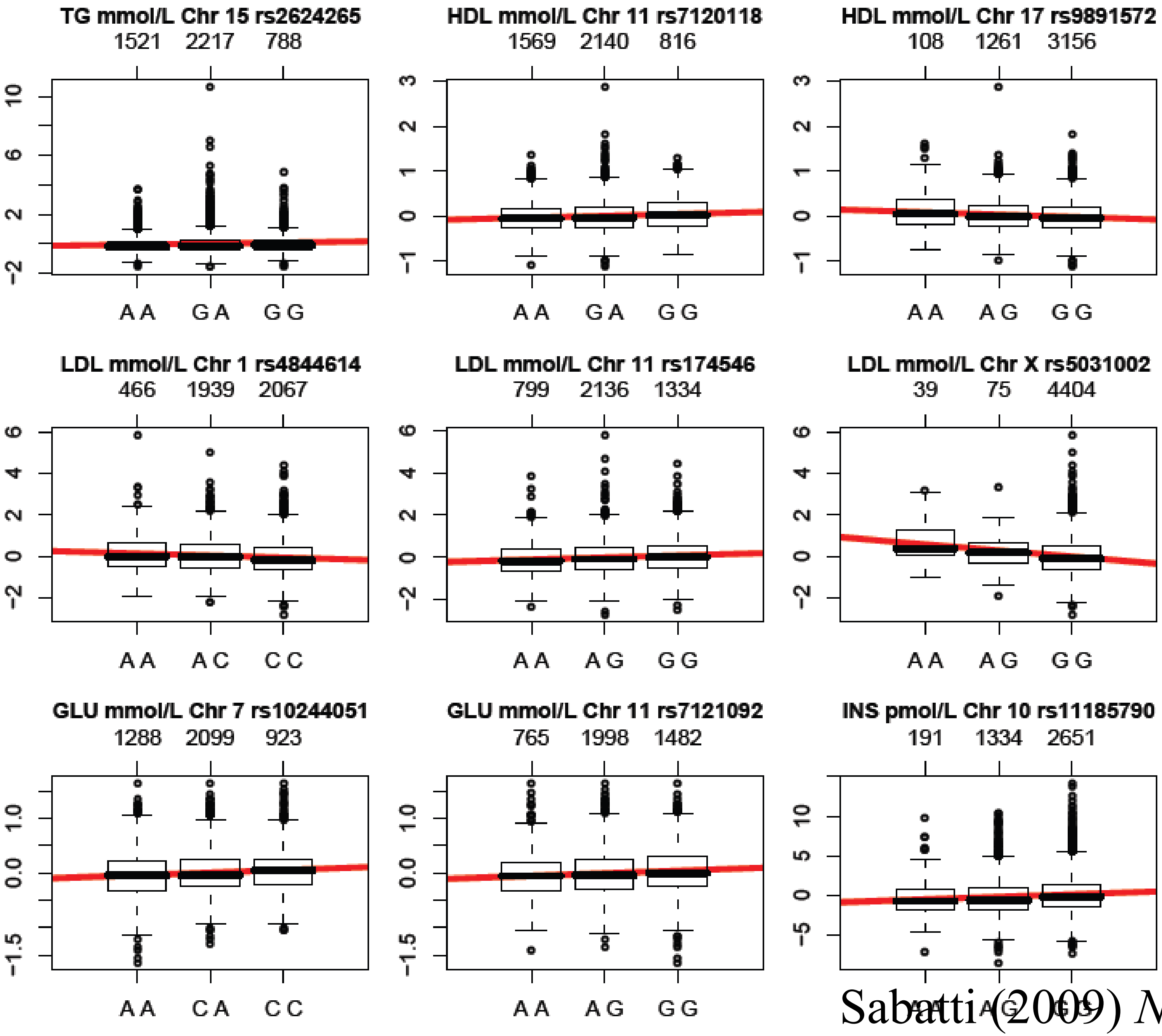
Morris and Cardon (2019) *Handbook of Stat Genomics*

Linear modeling framework



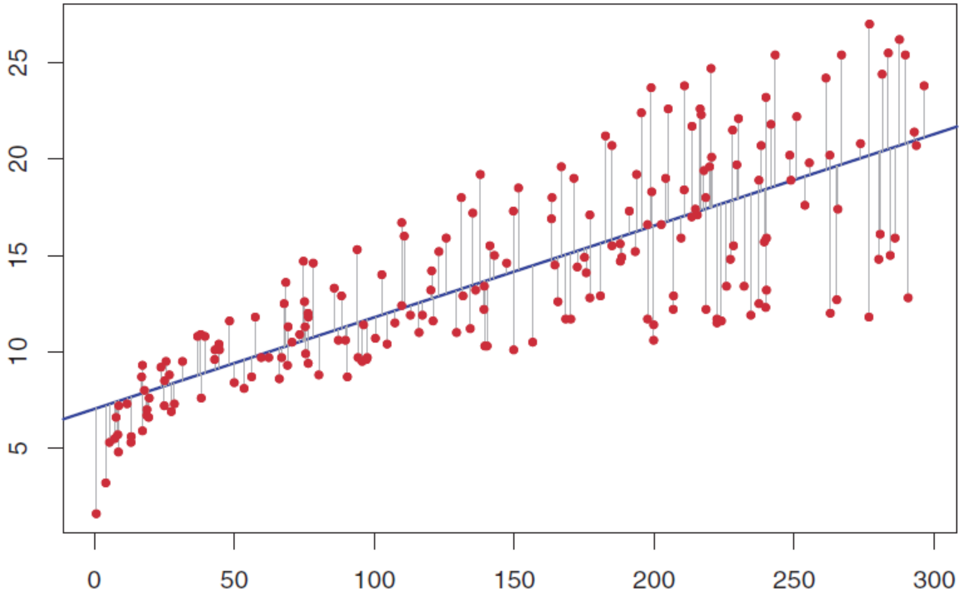
https://en.wikipedia.org/wiki/Genome-wide_association_study

Linear modeling framework



Sabatti (2009) *Nat Genet*

Linear modeling framework



$$(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$$

$$Y \approx \beta_0 + \beta_1 X$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Assessing the relationship in the simple linear regression model

- P – probability of the null hypothesis H_0: no relationship
- β – regression coeeficient (effect size)
- R^2 – coefficient of determination, the proportion of variability in Y that can be explained using X

Susceptibility/liability threshold

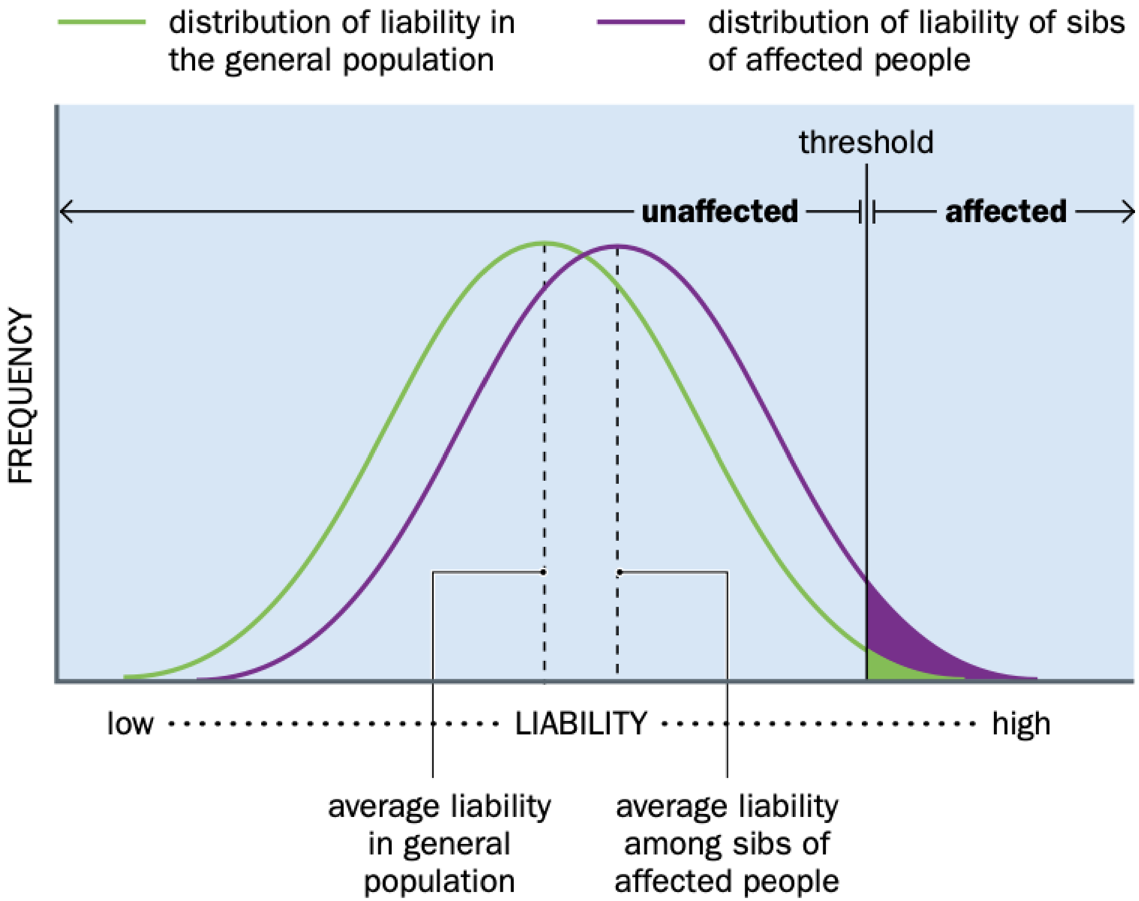


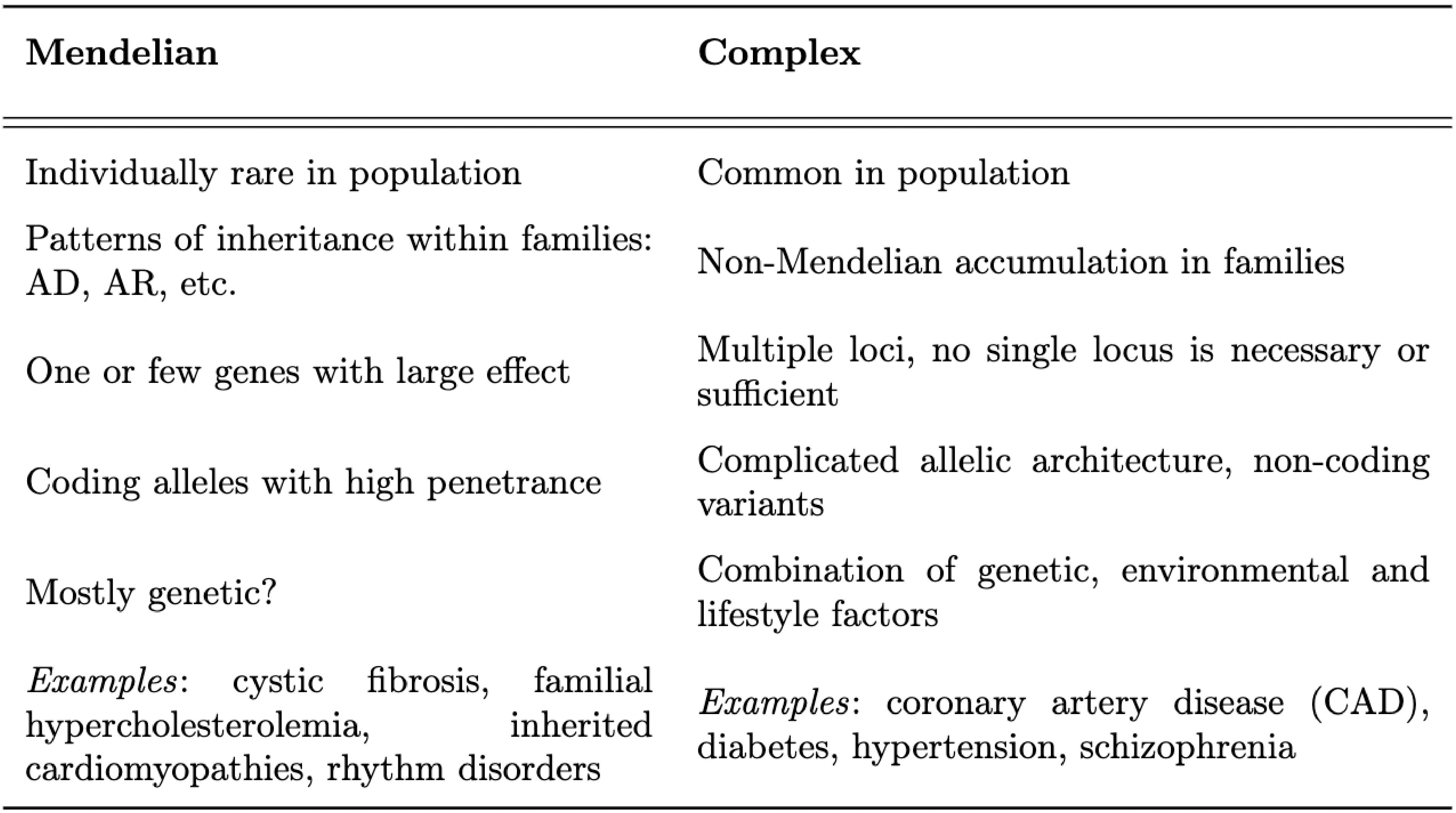
Figure 5.23 A polygenic threshold model for dichotomous non-Mendelian characters. Liability to the condition is polygenic and Normally distributed (green curve). People whose liability is above a certain threshold value (the balance point in **Figure 5.22**) are affected. The distribution of liability among sibs of an affected person (purple curve) is shifted toward higher liability because they share genes with their affected sib. A greater proportion of them have liability exceeding the fixed threshold. As a result, the condition tends to run in families.

Explains non-Mendelian accumulation (enrichment) of binary traits (e.g., complex diseases) in pedigrees

Unlike in Mendelian conditions, the recurrence risk increases with the number of previous affected children.

Strachan, Read – *Human Molecular Genetics*

Mendelian vs. complex disorders



Mendelian	**Complex**
Individually rare in population	Common in population
Patterns of inheritance within families: AD, AR, etc.	Non-Mendelian accumulation in families
One or few genes with large effect	Multiple loci, no single locus is necessary or sufficient
Coding alleles with high penetrance	Complicated allelic architecture, non-coding variants
Mostly genetic?	Combination of genetic, environmental and lifestyle factors
Examples: cystic fibrosis, familial hypercholesterolemia, inherited cardiomyopathies, rhythm disorders	*Examples*: coronary artery disease (CAD), diabetes, hypertension, schizophrenia

Allelic architecture of genetic disorders

Rare (Mendelian) disorders

- Very rare (AF<<1%) and highly deleterious variants
- Subject to mutation-selection balance

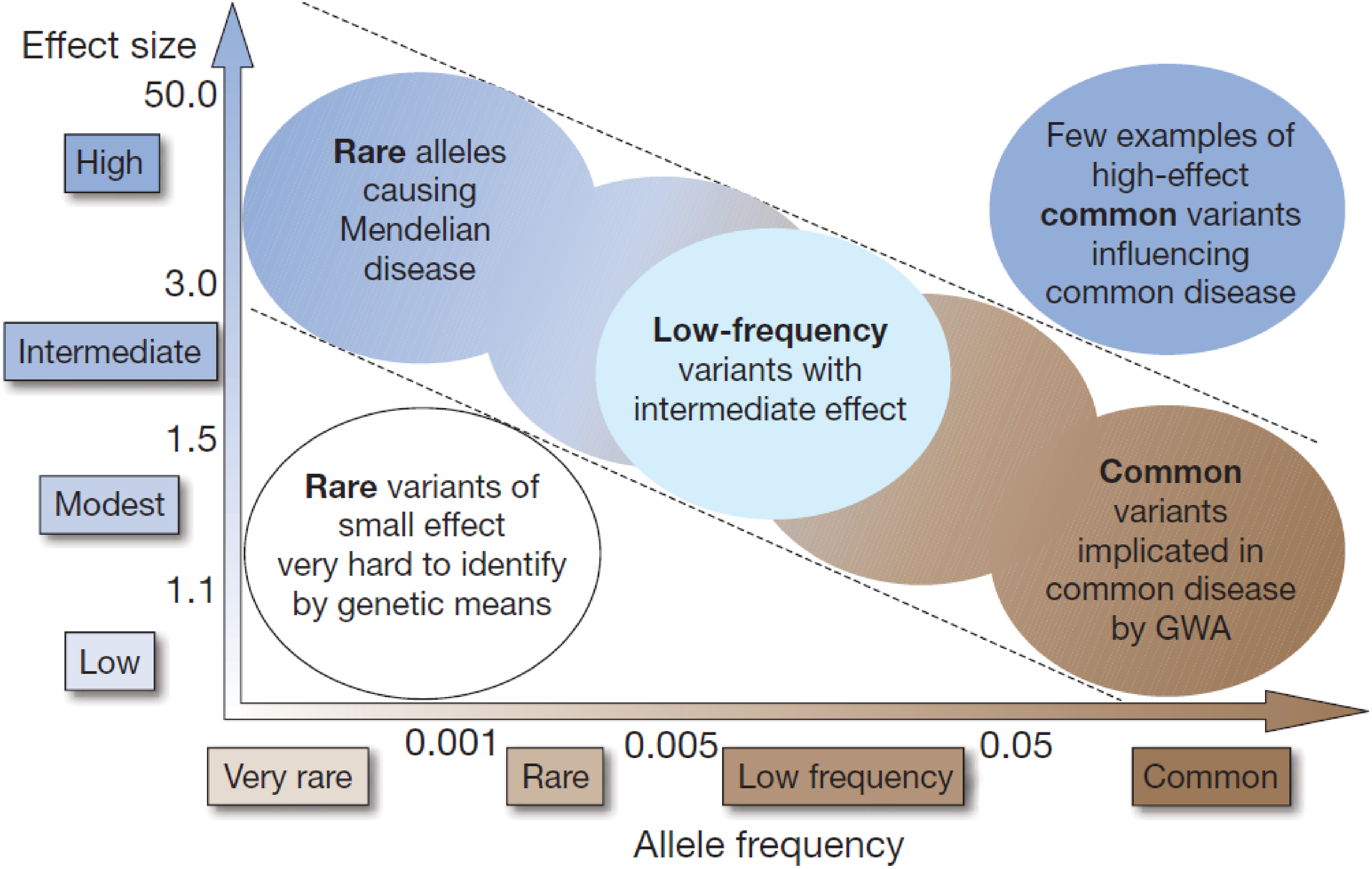
Complex disorders, common variants // Reich, Lander (2001)

- Relatively few old, common (AF>1%) variants
- Experience no selection?
  - Post-reproductive onset, no purifying selection (T2D)
  - Balancing selection (Kidney disease/parasite resistance in Africa)
  - «Thrifty» hypothesis (Obesity, diabetes)

Complex disorders, rare variants // Pritchard (2001) *AJHG*

- Numerous low frequency (AF<1%) variants with intermediate effect
- Recent human expansion ⇒ multiple mildly deleterious variants

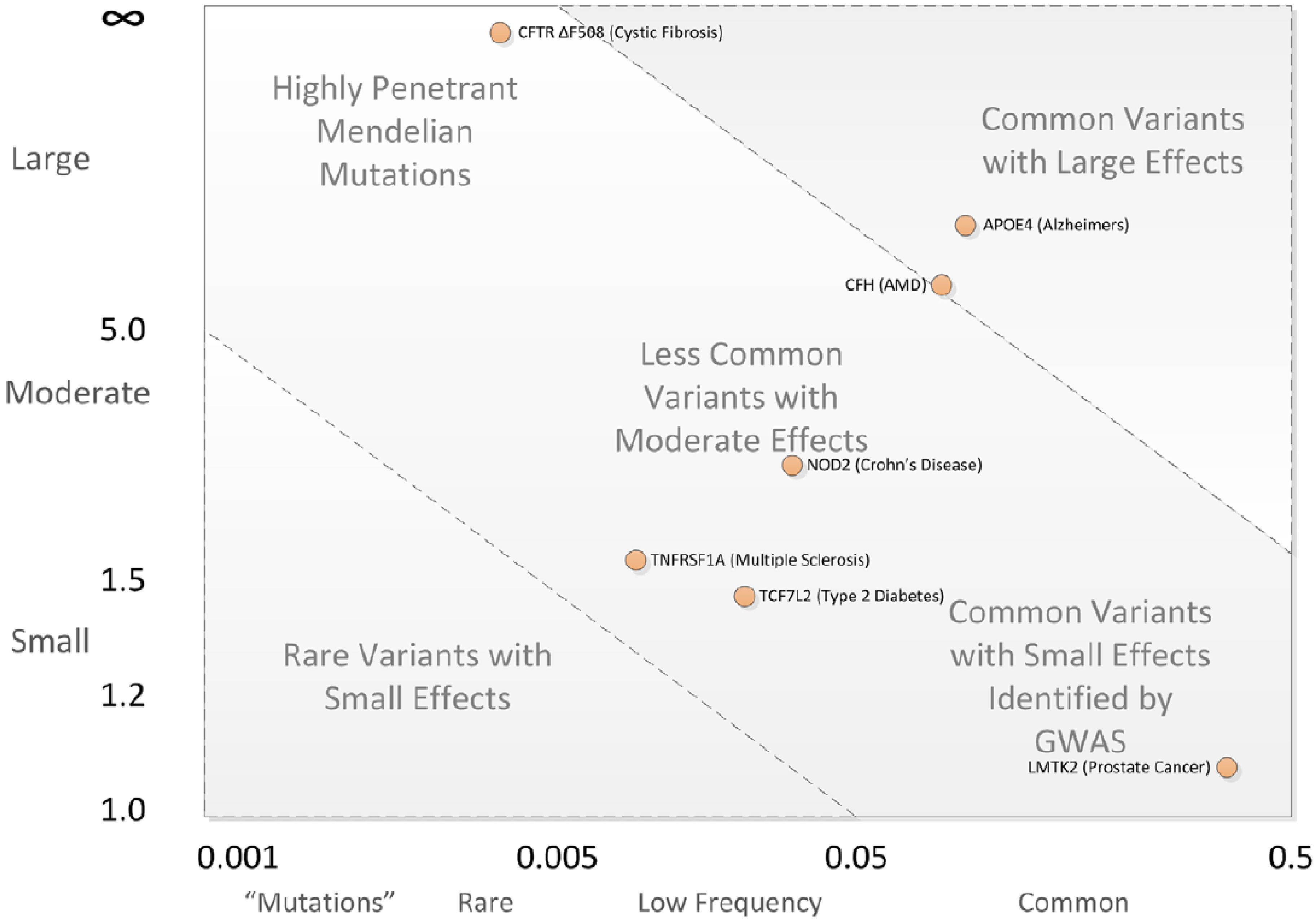
Allelic architecture of genetic disorders


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect **common** variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Effect size (odds ratio) vs. frequency of risk alleles

Manolio (2009) *Nature*

Allelic architecture of genetic disorders



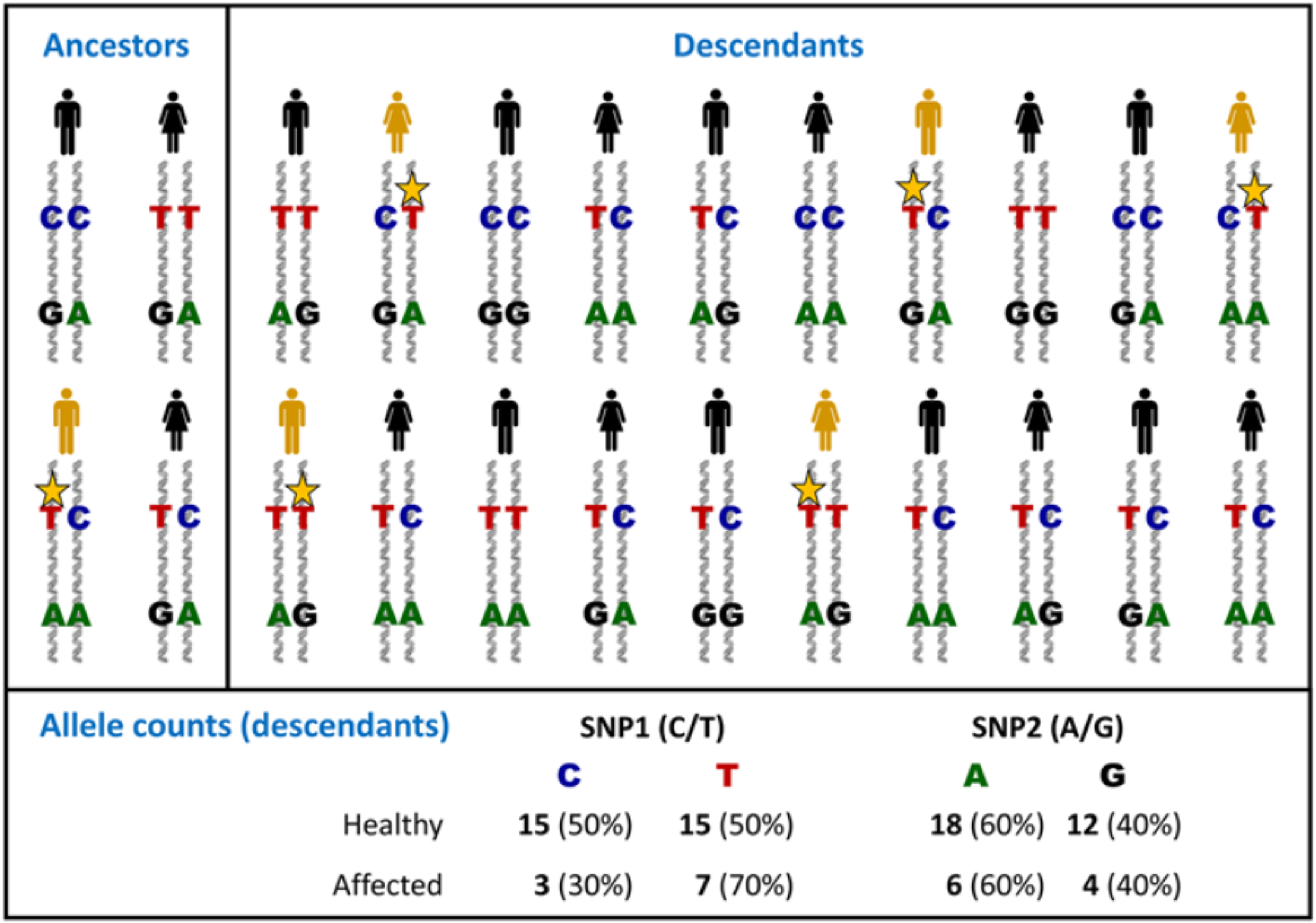
Effect size (odds ratio) vs. frequency of risk alleles

Bush and Moore (2012) *PloS Comp Bio*

The basis of genome-wide association studies

Affected –

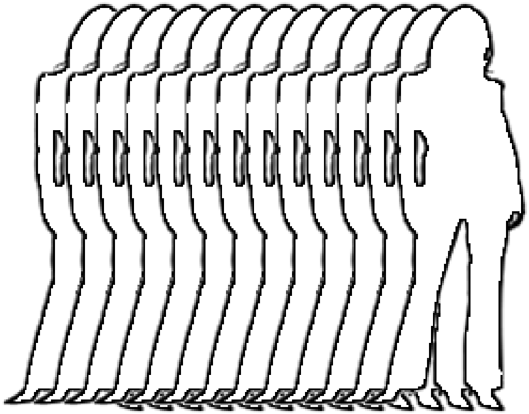
Risk allele –



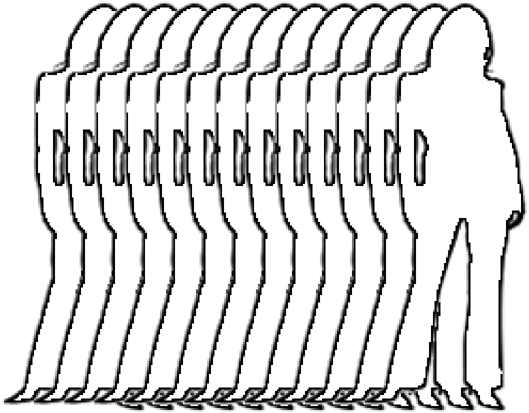
Allele counts (descendants)	SNP1 (C/T)		SNP2 (A/G)	
	C	T	A	G
Healthy	**15** (50%)	**15** (50%)	**18** (60%)	**12** (40%)
Affected	**3** (30%)	**7** (70%)	**6** (60%)	**4** (40%)

Jackson (2018) *Essays Biochem*

The basis of genome-wide association studies



GC CC GG GC CC GC GC
GG CC GC GG GC GG

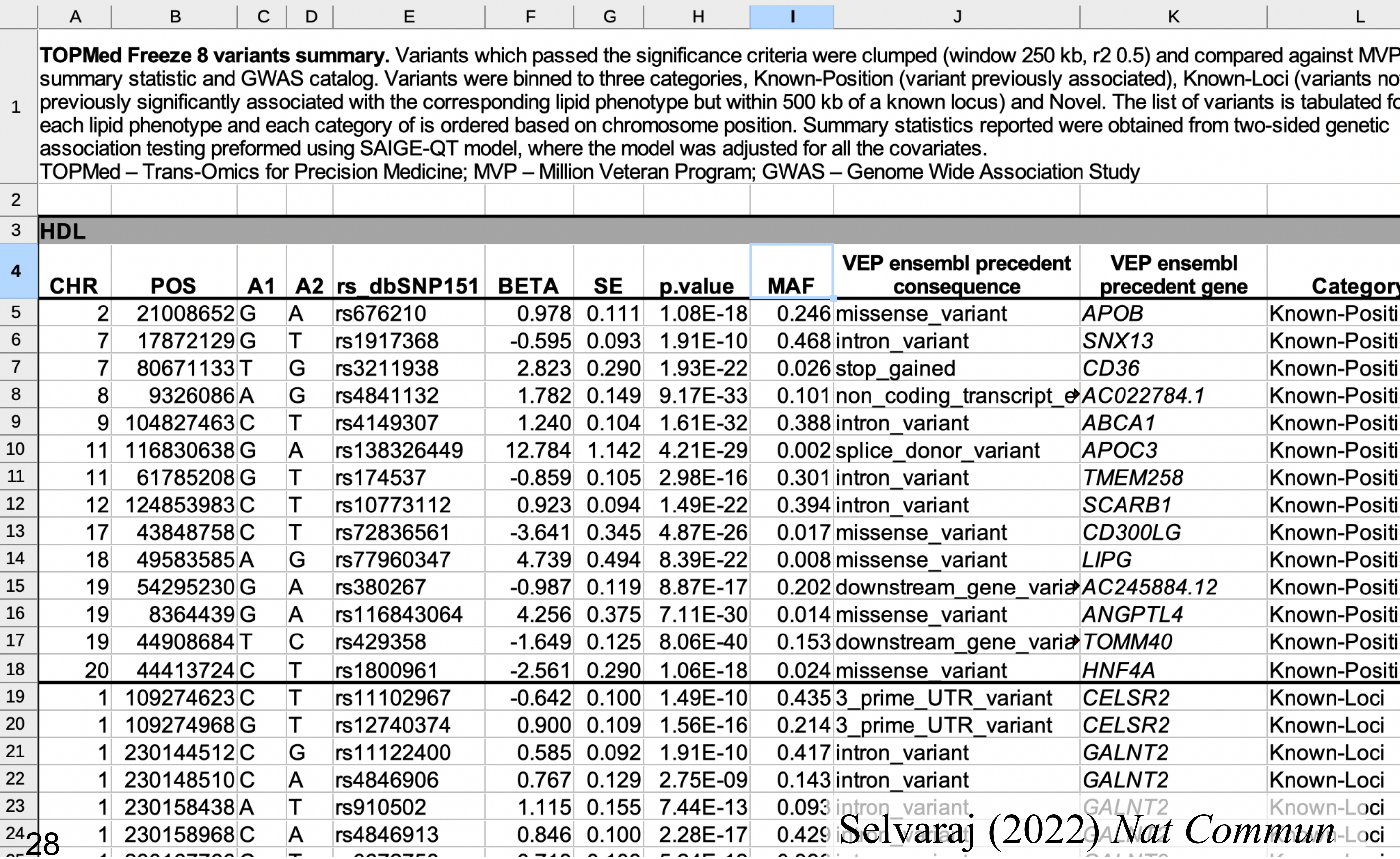


GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1	*SNP2*	*SNP …*
Cases	**Cases**	*Repeat for all SNPs*
Count of G: 2104 of 4000	Count of G: 1648 of 4000	
Frequency of G: 52.6%	Frequency of G: 41.2%	
Controls	**Controls**	
Count of G: 2676 of 6000	Count of G: 2532 of 6000	
Frequency of G: 44.6%	Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

The numbers: 2007 study of coronary artery disease (CAD) that showed that the individuals with the G-allele of SNP1 (rs1333049) were overrepresented amongst CAD-patients. doi:10.1038/nature05911

Example of GWAS summary statistics



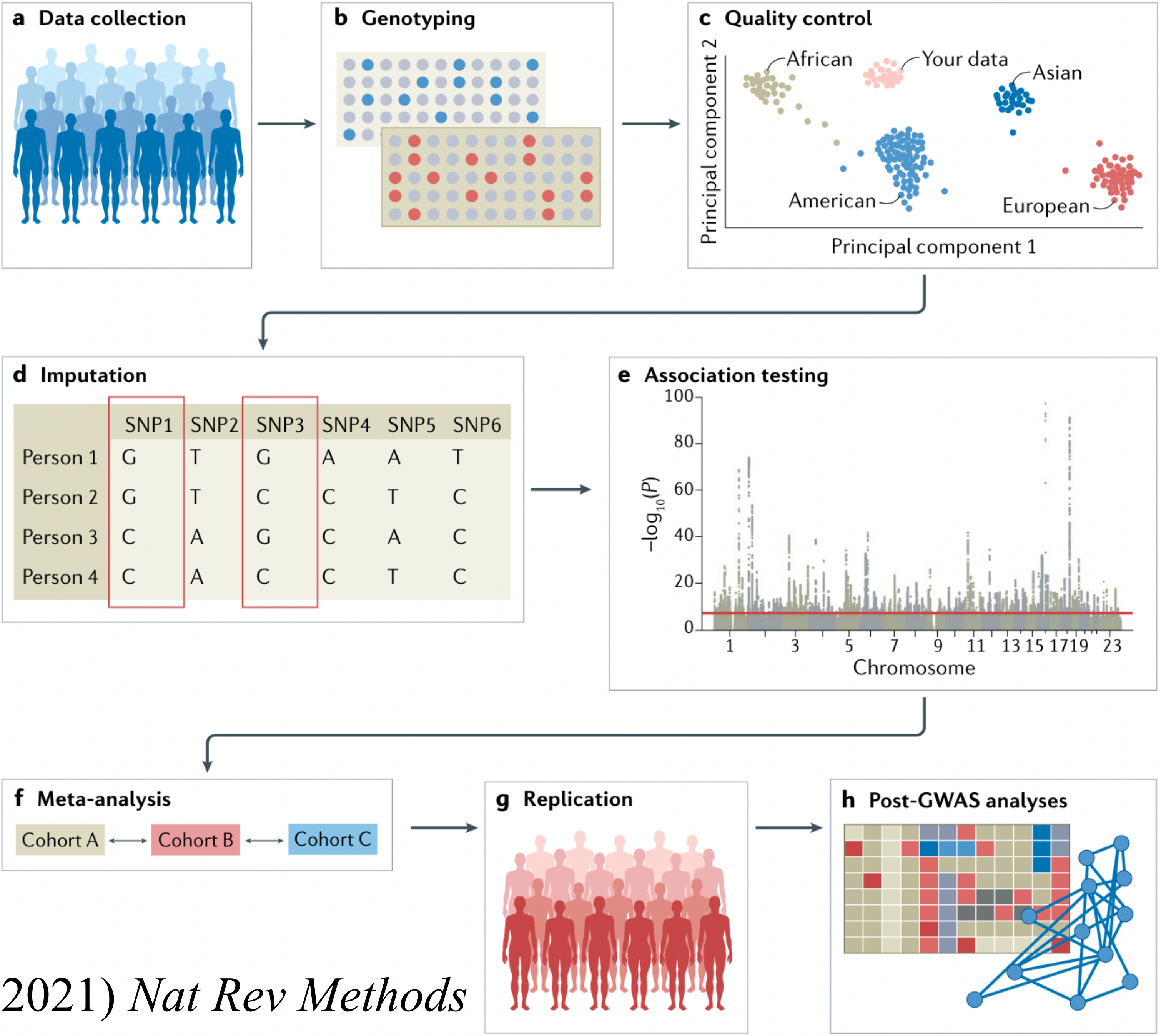
TOPMed Freeze 8 variants summary. Variants which passed the significance criteria were clumped (window 250 kb, r2 0.5) and compared against MVP summary statistic and GWAS catalog. Variants were binned to three categories, Known-Position (variant previously associated), Known-Loci (variants no previously significantly associated with the corresponding lipid phenotype but within 500 kb of a known locus) and Novel. The list of variants is tabulated fo each lipid phenotype and each category of is ordered based on chromosome position. Summary statistics reported were obtained from two-sided genetic association testing preformed using SAIGE-QT model, where the model was adjusted for all the covariates.
TOPMed – Trans-Omics for Precision Medicine; MVP – Million Veteran Program; GWAS – Genome Wide Association Study

HDL

CHR	POS	A1	A2	rs_dbSNP151	BETA	SE	p.value	MAF	VEP ensembl precedent consequence	VEP ensembl precedent gene	Category
2	21008652	G	A	rs676210	0.978	0.111	1.08E-18	0.246	missense_variant	*APOB*	Known-Positi
7	17872129	G	T	rs1917368	-0.595	0.093	1.91E-10	0.468	intron_variant	*SNX13*	Known-Positi
7	80671133	T	G	rs3211938	2.823	0.290	1.93E-22	0.026	stop_gained	*CD36*	Known-Positi
8	9326086	A	G	rs4841132	1.782	0.149	9.17E-33	0.101	non_coding_transcript_e	*AC022784.1*	Known-Positi
9	104827463	C	T	rs4149307	1.240	0.104	1.61E-32	0.388	intron_variant	*ABCA1*	Known-Positi
11	116830638	G	A	rs138326449	12.784	1.142	4.21E-29	0.002	splice_donor_variant	*APOC3*	Known-Positi
11	61785208	G	T	rs174537	-0.859	0.105	2.98E-16	0.301	intron_variant	*TMEM258*	Known-Positi
12	124853983	C	T	rs10773112	0.923	0.094	1.49E-22	0.394	intron_variant	*SCARB1*	Known-Positi
17	43848758	C	T	rs72836561	-3.641	0.345	4.87E-26	0.017	missense_variant	*CD300LG*	Known-Positi
18	49583585	A	G	rs77960347	4.739	0.494	8.39E-22	0.008	missense_variant	*LIPG*	Known-Positi
19	54295230	G	A	rs380267	-0.987	0.119	8.87E-17	0.202	downstream_gene_varia	*AC245884.12*	Known-Positi
19	8364439	G	A	rs116843064	4.256	0.375	7.11E-30	0.014	missense_variant	*ANGPTL4*	Known-Positi
19	44908684	T	C	rs429358	-1.649	0.125	8.06E-40	0.153	downstream_gene_varia	*TOMM40*	Known-Positi
20	44413724	C	T	rs1800961	-2.561	0.290	1.06E-18	0.024	missense_variant	*HNF4A*	Known-Positi
1	109274623	C	T	rs11102967	-0.642	0.100	1.49E-10	0.435	3_prime_UTR_variant	*CELSR2*	Known-Loci
1	109274968	G	T	rs12740374	0.900	0.109	1.56E-16	0.214	3_prime_UTR_variant	*CELSR2*	Known-Loci
1	230144512	C	G	rs11122400	0.585	0.092	1.91E-10	0.417	intron_variant	*GALNT2*	Known-Loci
1	230148510	C	A	rs4846906	0.767	0.129	2.75E-09	0.143	intron_variant	*GALNT2*	Known-Loci
1	230158438	A	T	rs910502	1.115	0.155	7.44E-13	0.093	intron_variant	*GALNT2*	Known-Loci
1	230158968	C	A	rs4846913	0.846	0.100	2.28E-17	0.429	intron_variant	*GALNT2*	Known-Loci

Selvaraj (2022) *Nat Commun*

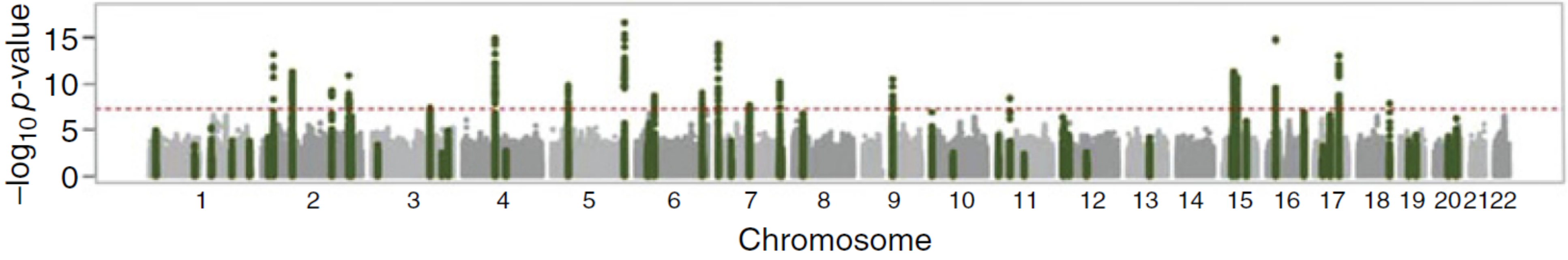
Overview of GWAS steps



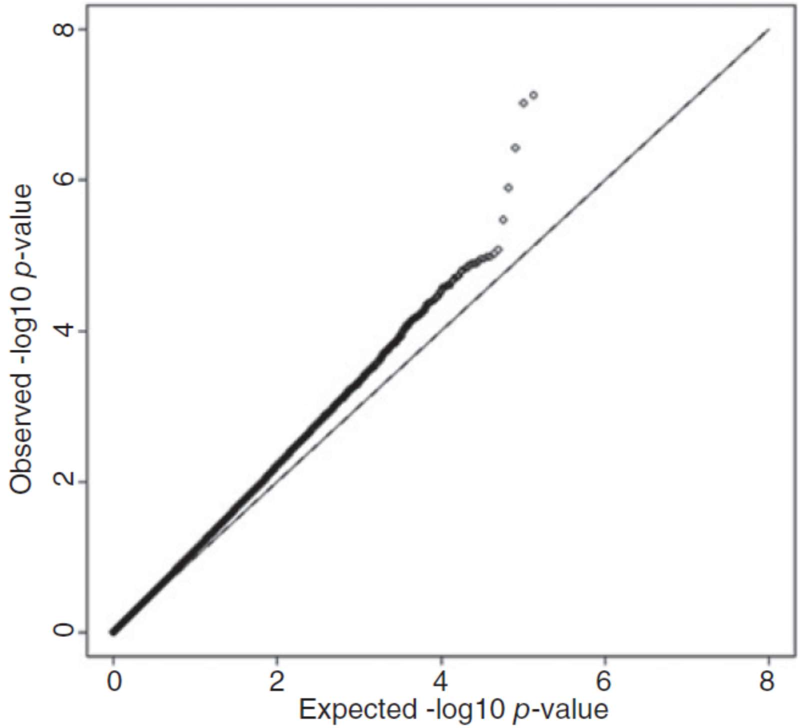
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

Uffelmann (2021) *Nat Rev Methods*

Visualization of GWAS results



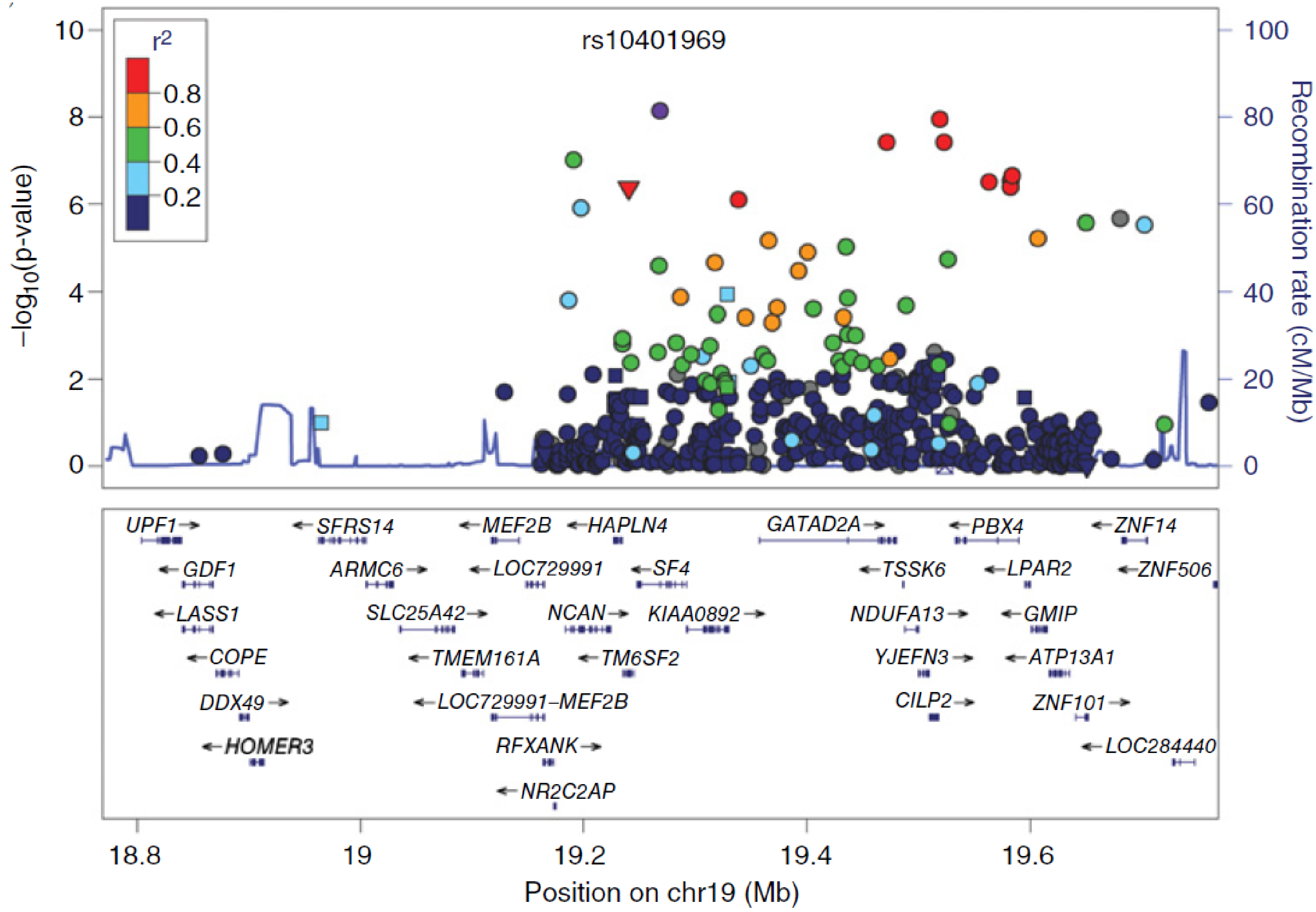
Manhattan plot. Each point corresponds to a SNP, plotted according to genomic position on the *x*-axis and the evidence in favour of association ($-\log_{10}$ *p*-value) on the *y*-axis. SNPs highlighted in green map to loci previously reported for the trait.



Quantile–quantile plot. Each point corresponds to a SNP, plotted according to the ranked $-\log_{10}$ *p*-value for association on the y-axis against the expected ranked $-\log_{10}$ *p*-value under the null hypothesis of no association on the x-axis. Inflation of $-\log_{10}$ *p*-values above the $y = x$ line is indicative of population structure that has not been accounted for in the association analysis.

Morris and Cardon (2019) *Handbook of Stat Genomics*

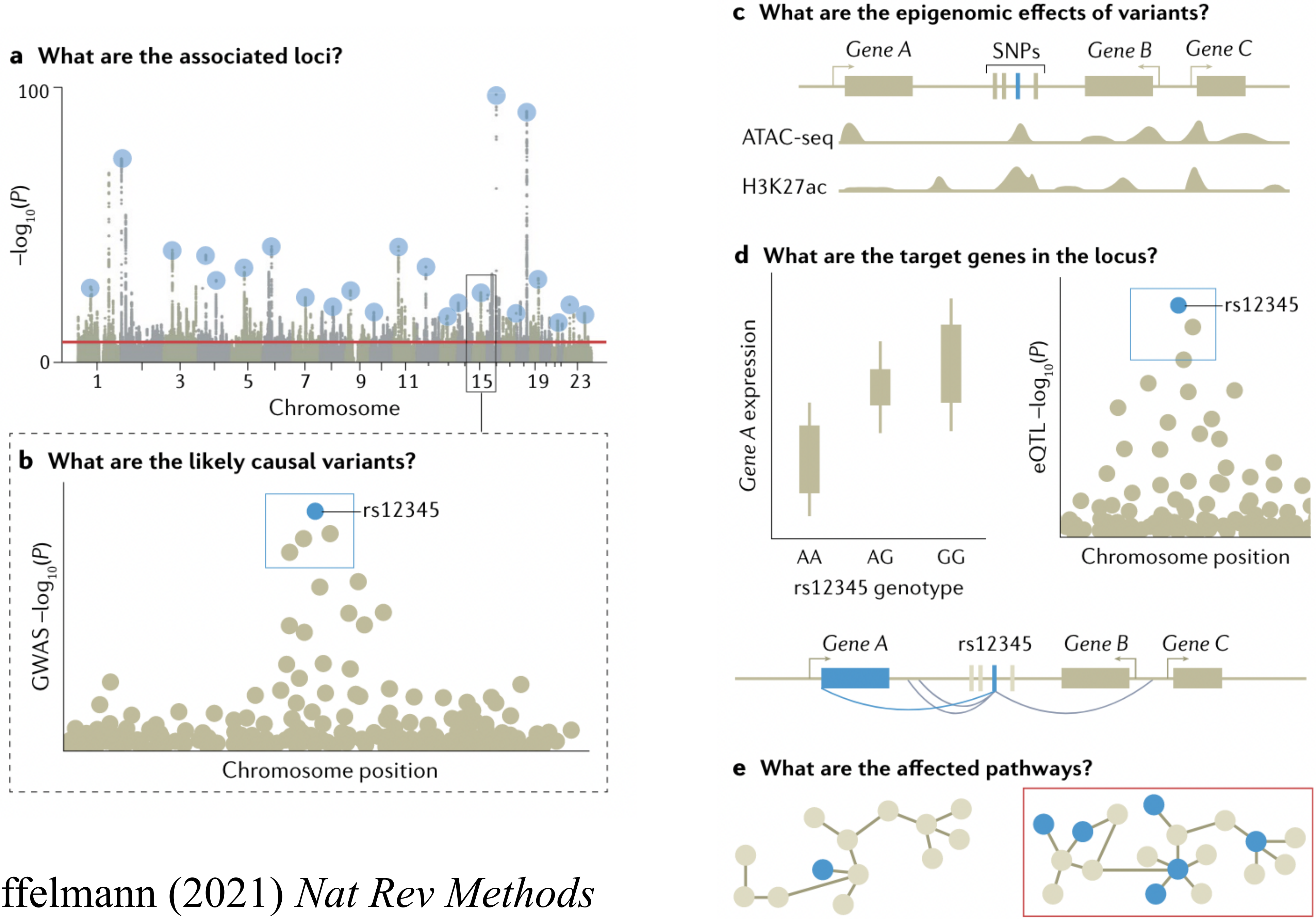
Visualization of GWAS results



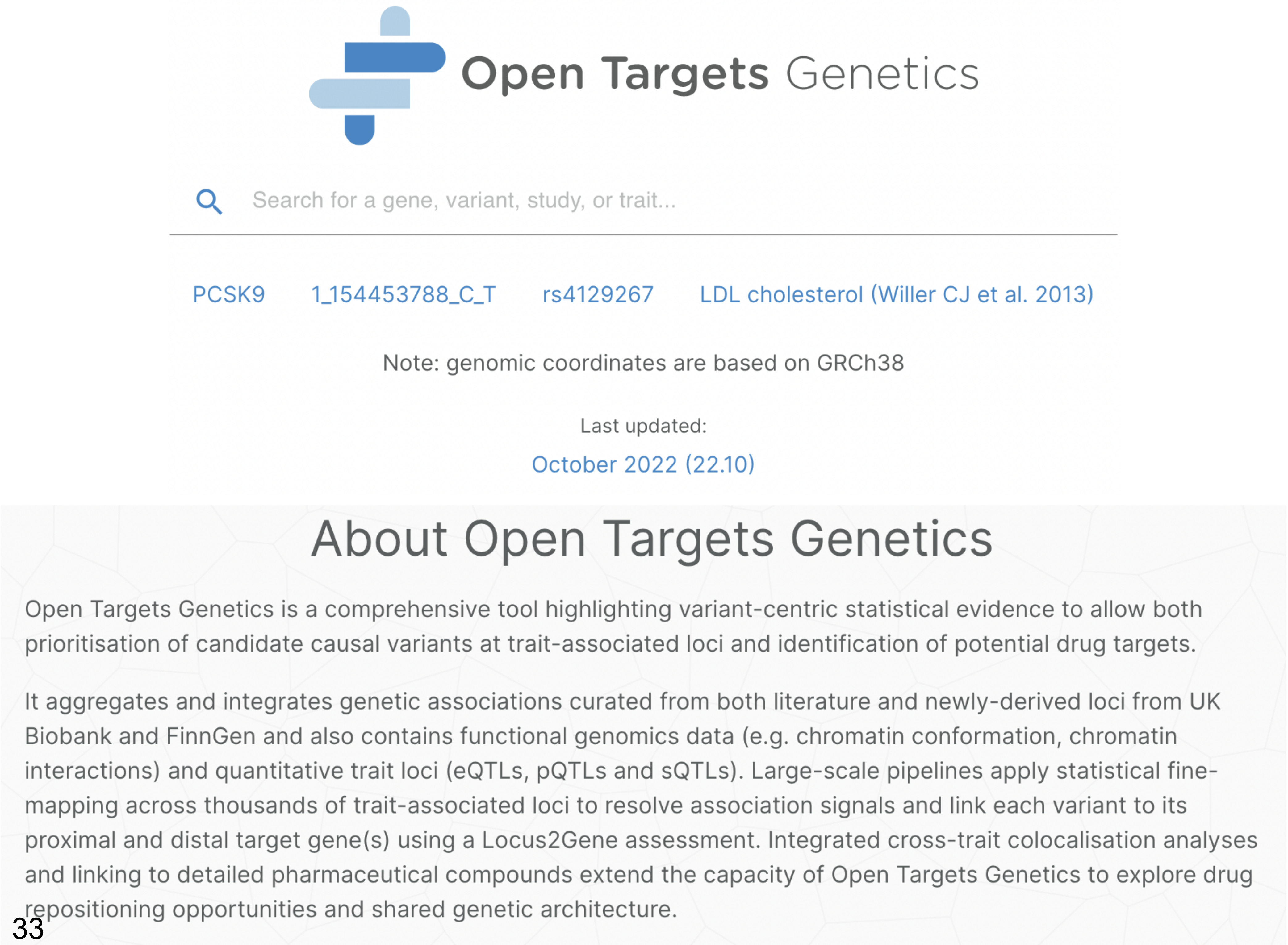
Signal plot. The index SNP: purple. Other colours: LD with the index variant in European ancestry haplotypes from the 1000 Genomes. The shape: upward triangle for frameshift, stop or splice; downward triangle for non-synonymous; square for synonymous or UTR; and circle for intronic or non-coding. Recombination rates are estimated from Phase II HapMap

Morris and Cardon (2019) *Handbook of Stat Genomics*

Functional follow-up of GWAS



Uffelmann (2021) *Nat Rev Methods*



Open Targets Genetics

Search for a gene, variant, study, or trait...

PCSK9 1_154453788_C_T rs4129267 LDL cholesterol (Willer CJ et al. 2013)

Note: genomic coordinates are based on GRCh38

Last updated:

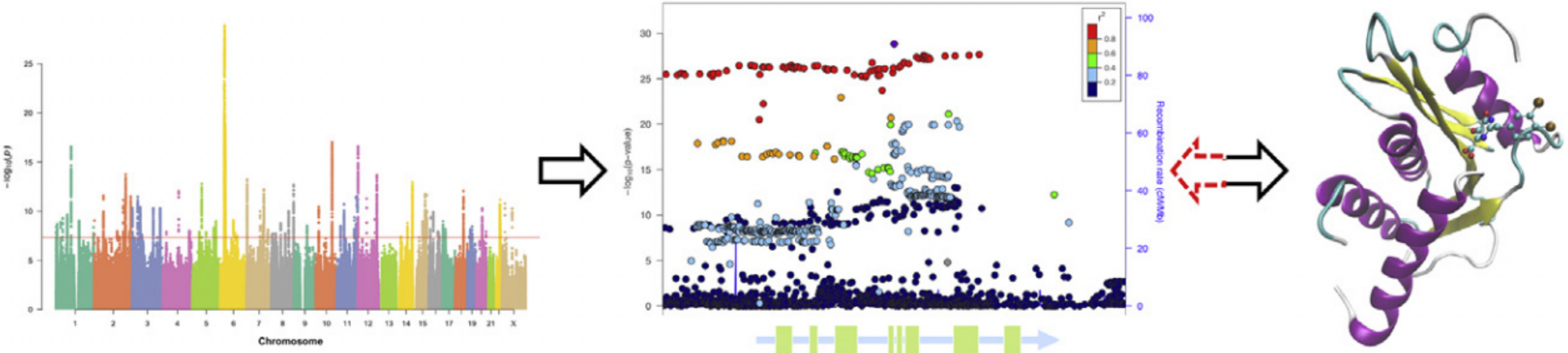
October 2022 (22.10)

About Open Targets Genetics

Open Targets Genetics is a comprehensive tool highlighting variant-centric statistical evidence to allow both prioritisation of candidate causal variants at trait-associated loci and identification of potential drug targets.

It aggregates and integrates genetic associations curated from both literature and newly-derived loci from UK Biobank and FinnGen and also contains functional genomics data (e.g. chromatin conformation, chromatin interactions) and quantitative trait loci (eQTLs, pQTLs and sQTLs). Large-scale pipelines apply statistical fine-mapping across thousands of trait-associated loci to resolve association signals and link each variant to its proximal and distal target gene(s) using a Locus2Gene assessment. Integrated cross-trait colocalisation analyses and linking to detailed pharmaceutical compounds extend the capacity of Open Targets Genetics to explore drug repositioning opportunities and shared genetic architecture.

GWAS applications: drug targeting



GWAS hits

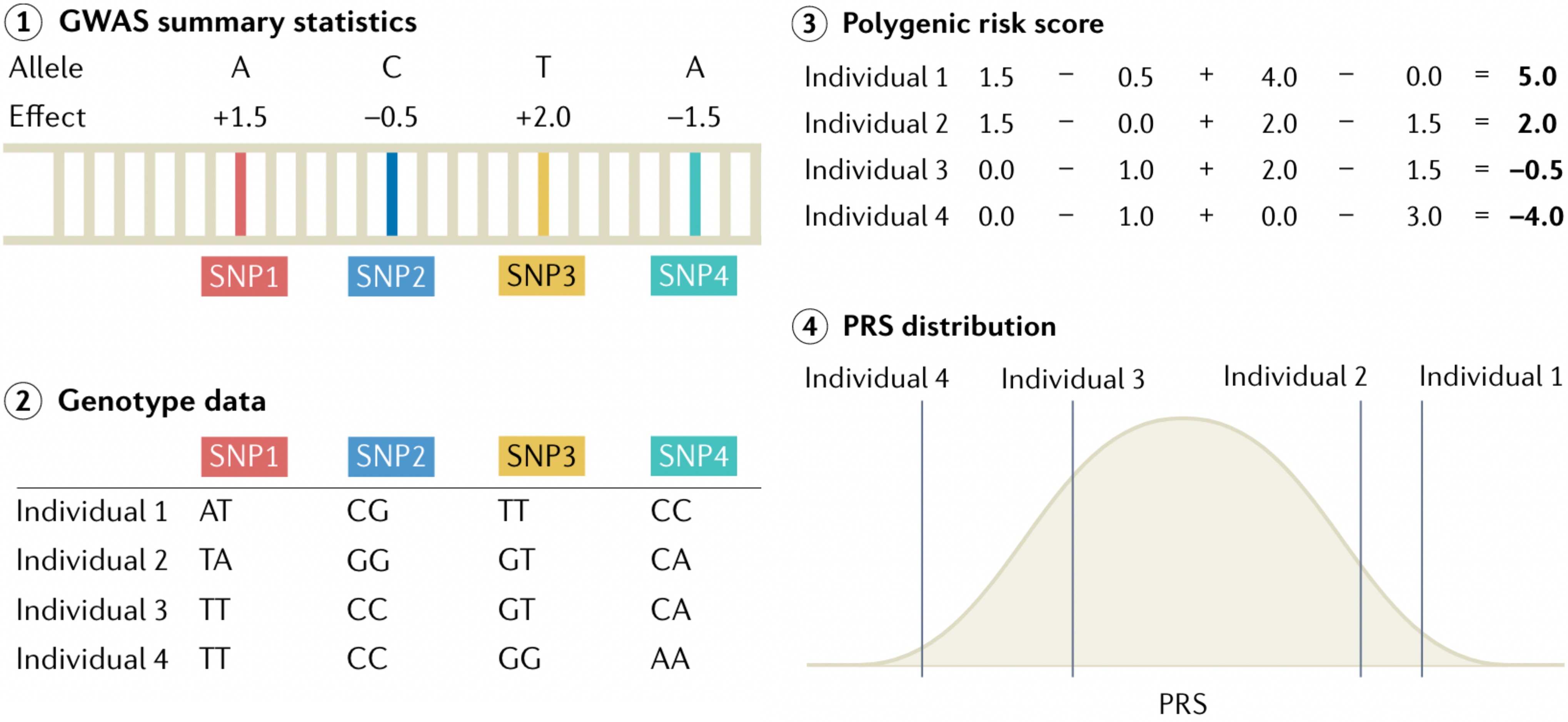
Gene

Drug

Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

Visscher (2017) *Am J Hum Genet*

GWAS applications: polygenic risk scores



1 GWAS summary statistics

Allele	A	C	T	A
Effect	+1.5	−0.5	+2.0	−1.5

2 Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

3 Polygenic risk score

Individual 1	1.5	−	0.5	+	4.0	−	0.0	=	**5.0**
Individual 2	1.5	−	0.0	+	2.0	−	1.5	=	**2.0**
Individual 3	0.0	−	1.0	+	2.0	−	1.5	=	**−0.5**
Individual 4	0.0	−	1.0	+	0.0	−	3.0	=	**−4.0**

4 PRS distribution

Uffelmann (2021) *Nat Rev Methods*

GWAS applications: polygenic risk scores

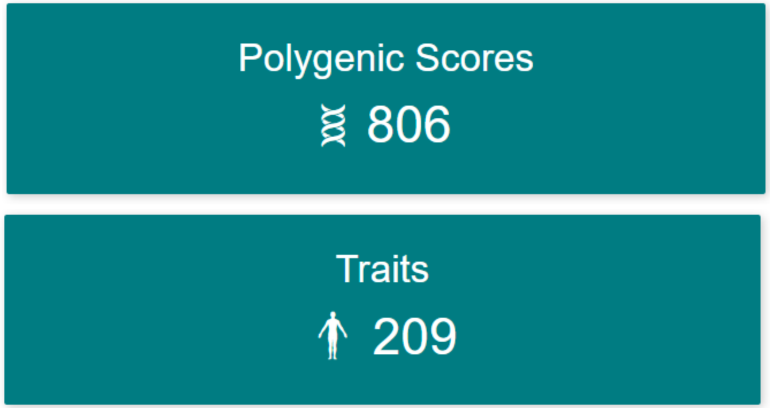
$$S = \sum_{i}^{N} \beta_i G_i, \quad G_i = \{0, 1, 2\}$$

- PRS: sum of allele dosages
- weighted by their effect sizes,
- allele effects derived from GWAS, $N = 10^2$–10^6

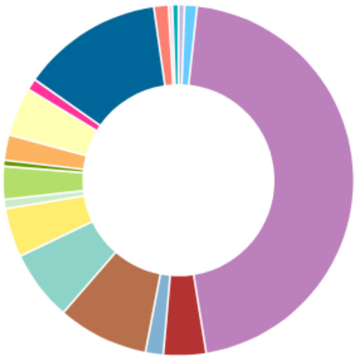
- PRS can be calculated at birth
- Carriers of high PRS cannot be identified with conventional risk factors or biomarkers
- Top 5% high coronary artery disease PRS carriers are at 3.7-fold increased odds for myocardial infarction
- Polygenic background may modify penetrance of monogenic mutations

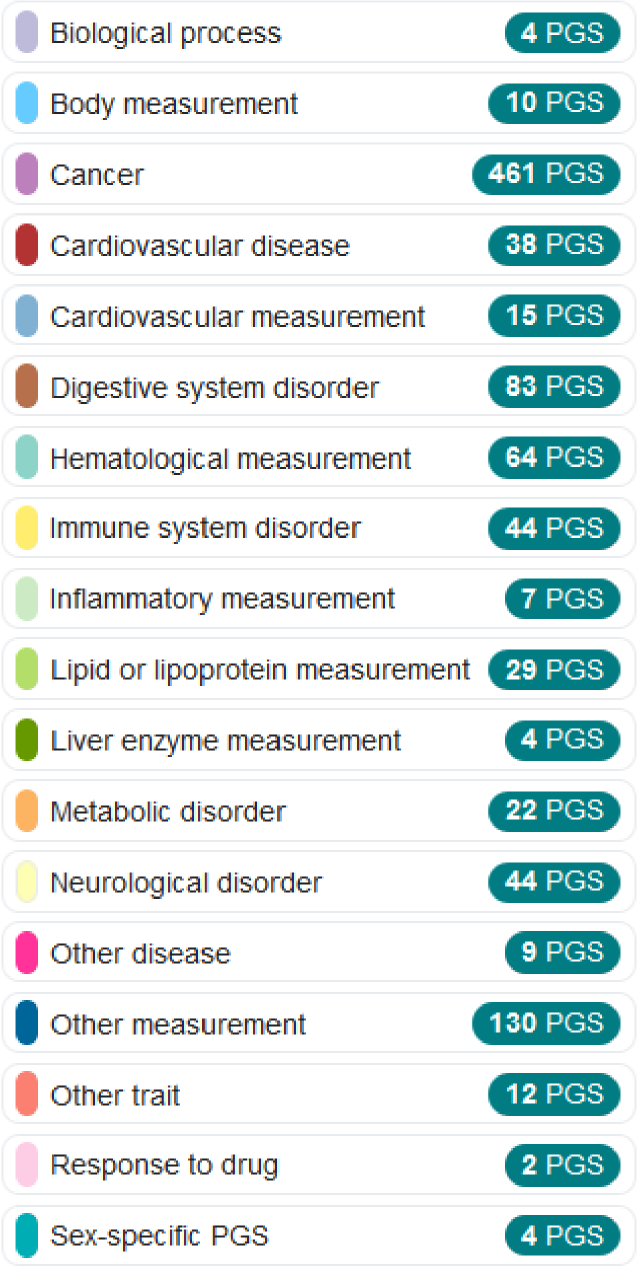
Polygenic risk scores

PGS Catalog | Home | Browse ▾


Polygenic Scores
806

Traits
209





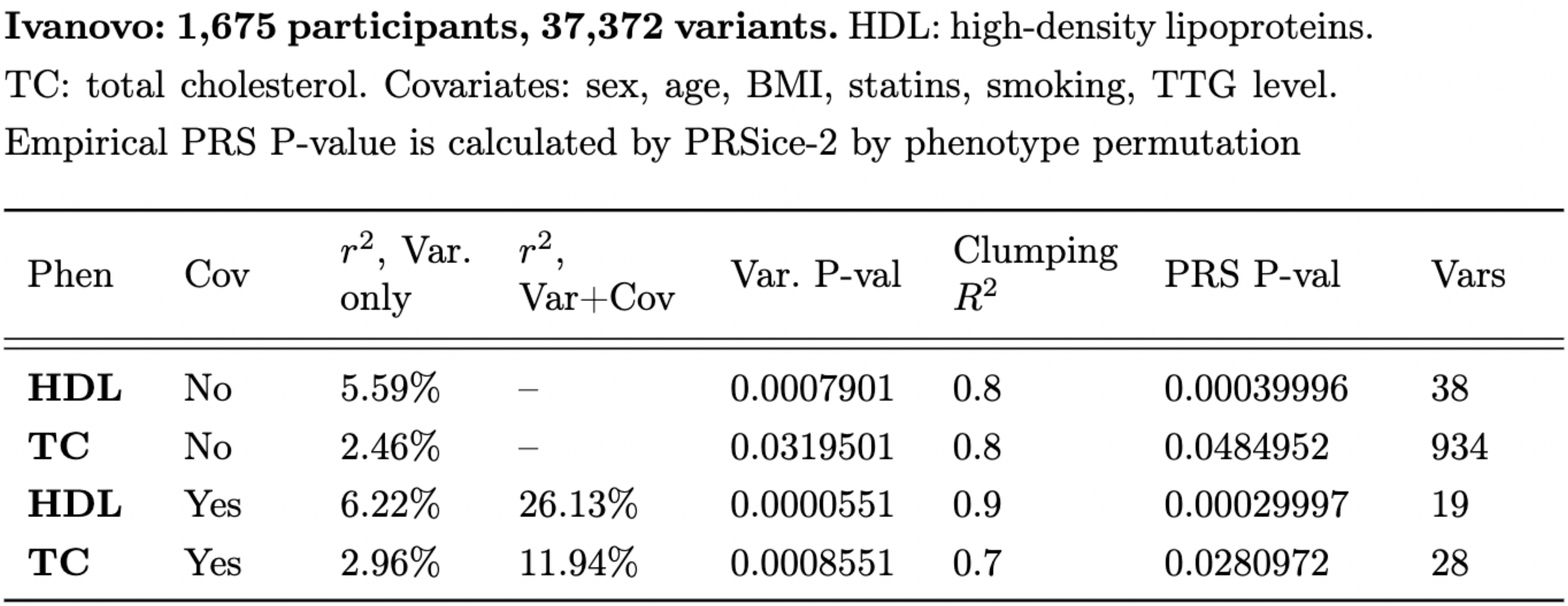
●	Biological process	4 PGS
●	Body measurement	10 PGS
●	Cancer	461 PGS
●	Cardiovascular disease	38 PGS
●	Cardiovascular measurement	15 PGS
●	Digestive system disorder	83 PGS
●	Hematological measurement	64 PGS
●	Immune system disorder	44 PGS
●	Inflammatory measurement	7 PGS
●	Lipid or lipoprotein measurement	29 PGS
●	Liver enzyme measurement	4 PGS
●	Metabolic disorder	22 PGS
●	Neurological disorder	44 PGS
●	Other disease	9 PGS
●	Other measurement	130 PGS
●	Other trait	12 PGS
●	Response to drug	2 PGS
●	Sex-specific PGS	4 PGS

What is a Polygenic Score?

A **polygenic score** (PGS) aggregates the effects of many genetic variants into a single number which predicts genetic predisposition for a phenotype. PGS are typically composed of hundreds-to-millions of genetic variants (usually SNPs) which are combined using a weighted sum of allele dosages multiplied by their corresponding effect sizes, as estimated from a relevant genome-wide association study (GWAS).

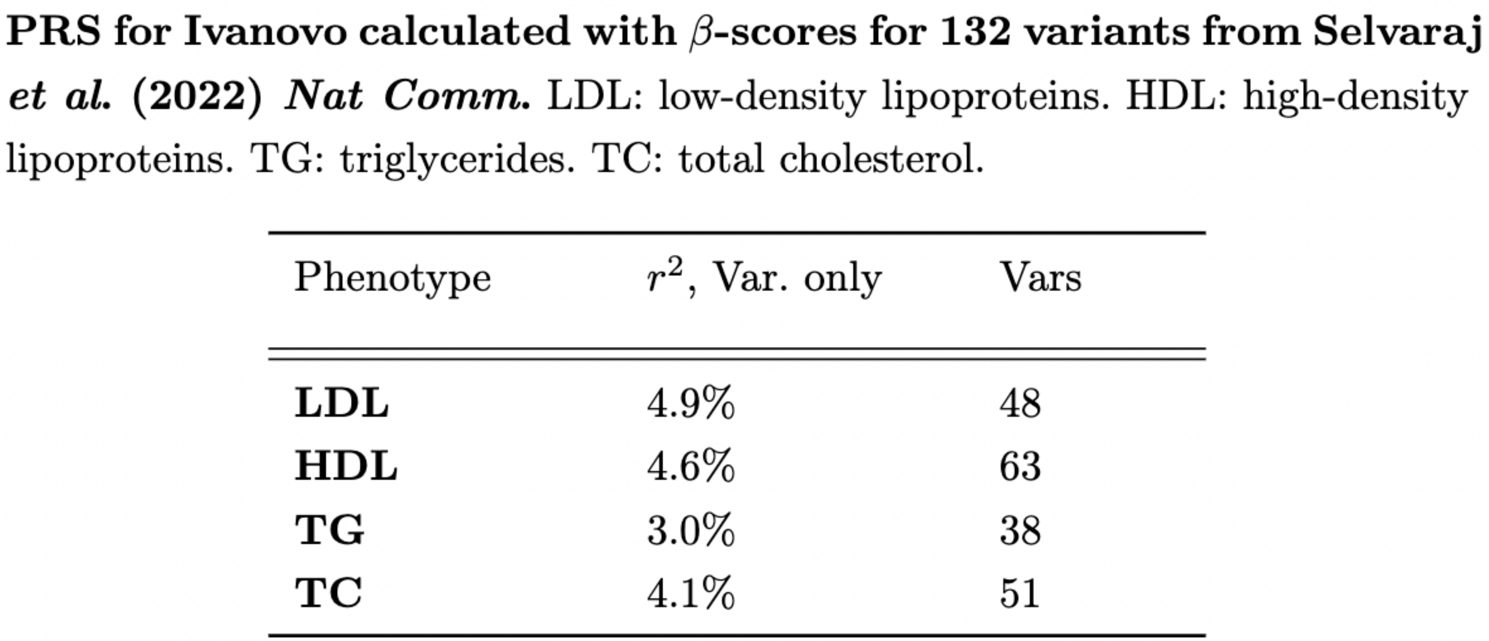
PGS nomenclature is heterogeneous: they can also be referred to as **genetic scores** or **genomic scores**, and as **polygenic risk scores (PRS)** or **genomic risk scores (GRS)** if they predict a discrete phenotype, such as a disease.

Lipid PRS for Ivanovo



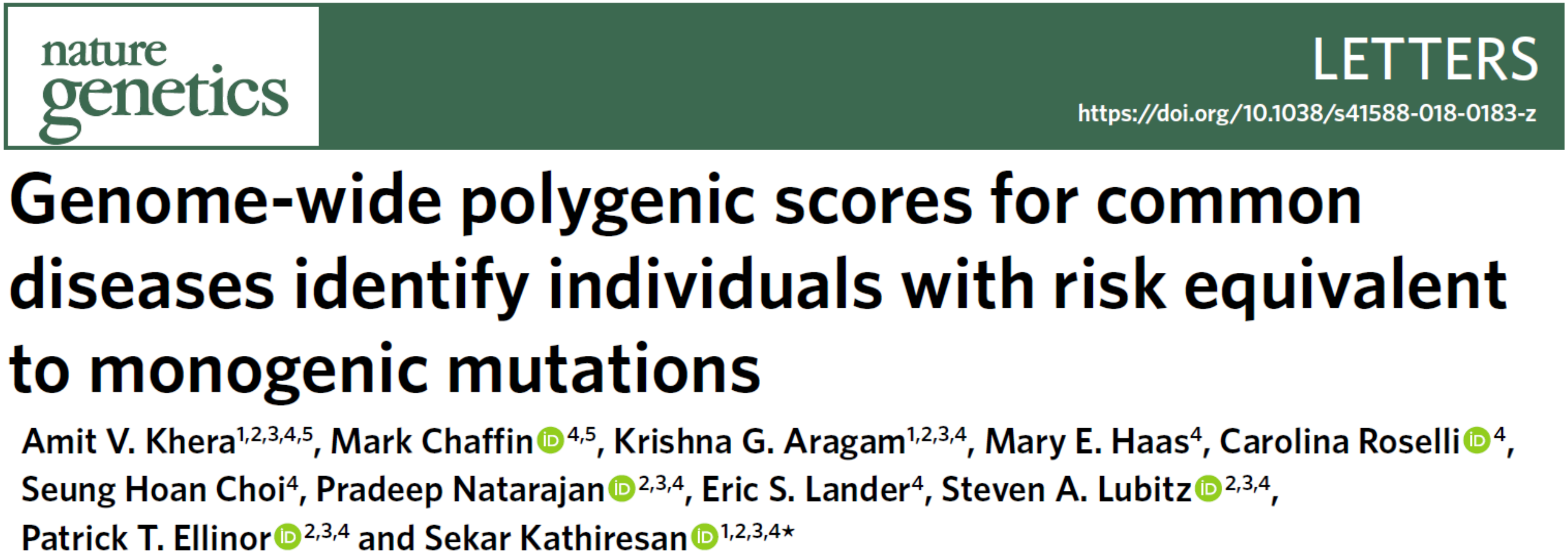
Ivanovo: 1,675 participants, 37,372 variants. HDL: high-density lipoproteins. TC: total cholesterol. Covariates: sex, age, BMI, statins, smoking, TTG level. Empirical PRS P-value is calculated by PRSice-2 by phenotype permutation

Phen	Cov	r^2, Var. only	r^2, Var+Cov	Var. P-val	Clumping R^2	PRS P-val	Vars
HDL	No	5.59%	–	0.0007901	0.8	0.00039996	38
TC	No	2.46%	–	0.0319501	0.8	0.0484952	934
HDL	Yes	6.22%	26.13%	0.0000551	0.9	0.00029997	19
TC	Yes	2.96%	11.94%	0.0008551	0.7	0.0280972	28



PRS for Ivanovo calculated with β-scores for 132 variants from Selvaraj *et al.* (2022) *Nat Comm.* LDL: low-density lipoproteins. HDL: high-density lipoproteins. TG: triglycerides. TC: total cholesterol.

Phenotype	r^2, Var. only	Vars
LDL	4.9%	48
HDL	4.6%	63
TG	3.0%	38
TC	4.1%	51



nature genetics

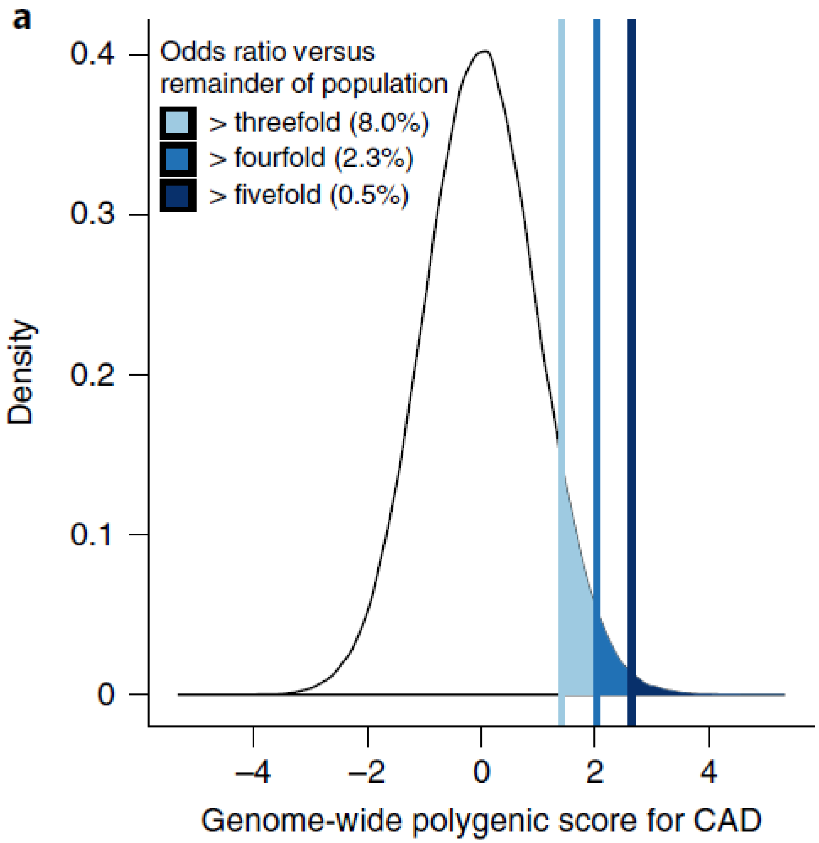
LETTERS

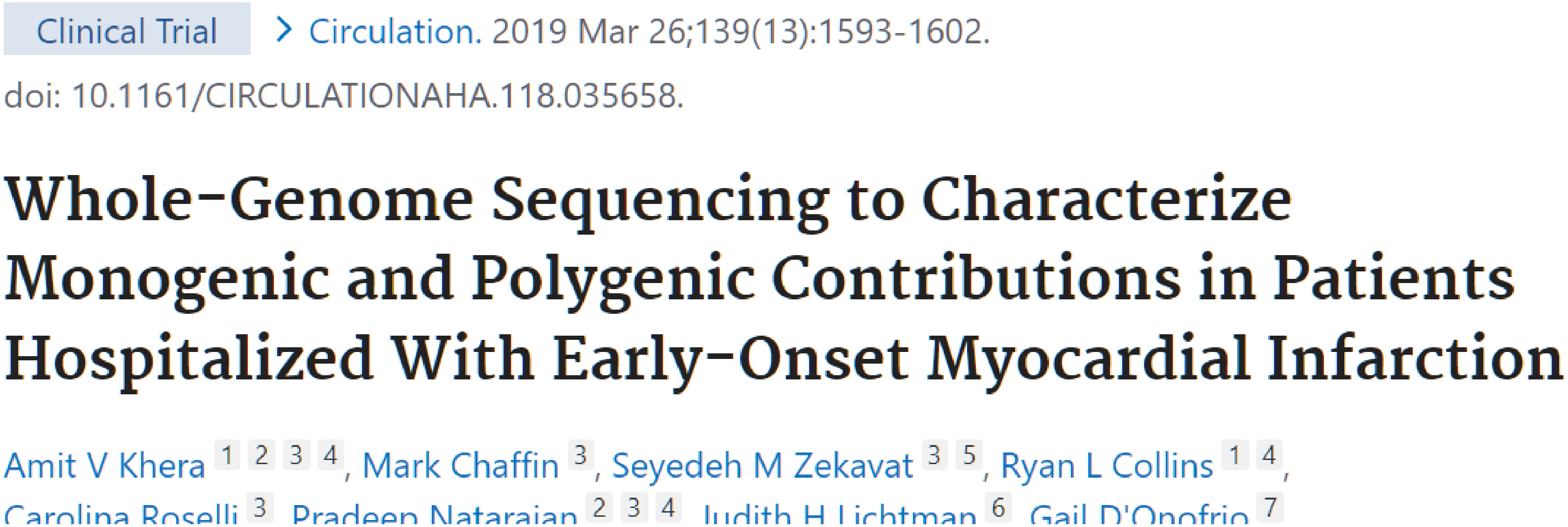
https://doi.org/10.1038/s41588-018-0183-z

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

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

Here, we develop and validate genome-wide polygenic scores for five common diseases. The approach identifies 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively. For coronary artery disease, this prevalence is 20-fold higher than the carrier frequency of rare monogenic mutations conferring comparable risk. We propose that it is time to contemplate the inclusion of polygenic risk prediction in clinical care, and discuss relevant issues.

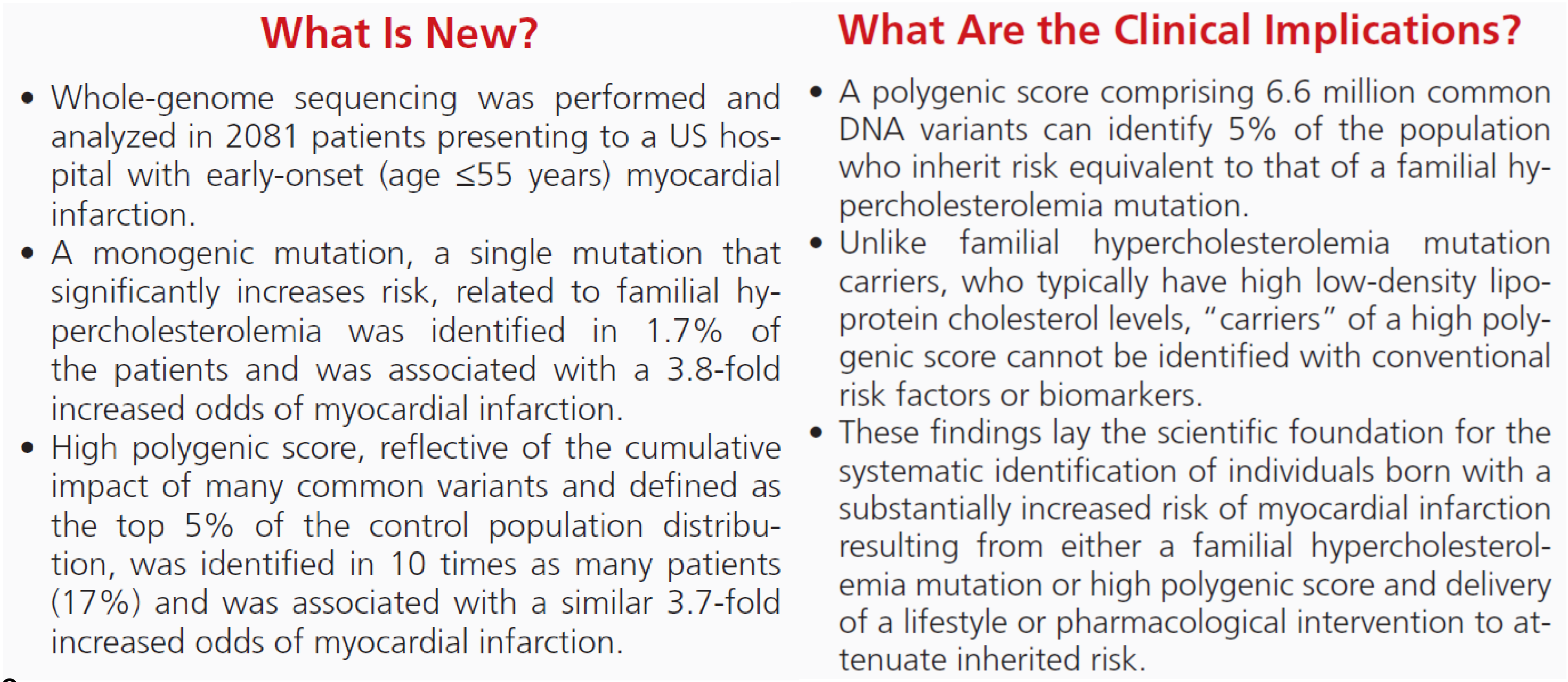




Clinical Trial > Circulation. 2019 Mar 26;139(13):1593-1602.
doi: 10.1161/CIRCULATIONAHA.118.035658.

Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction

Amit V Khera [1] [2] [3] [4], Mark Chaffin [3], Seyedeh M Zekavat [3] [5], Ryan L Collins [1] [4], Carolina Roselli [3], Pradeep Natarajan [2] [3] [4], Judith H Lichtman [6], Gail D'Onofrio [7]

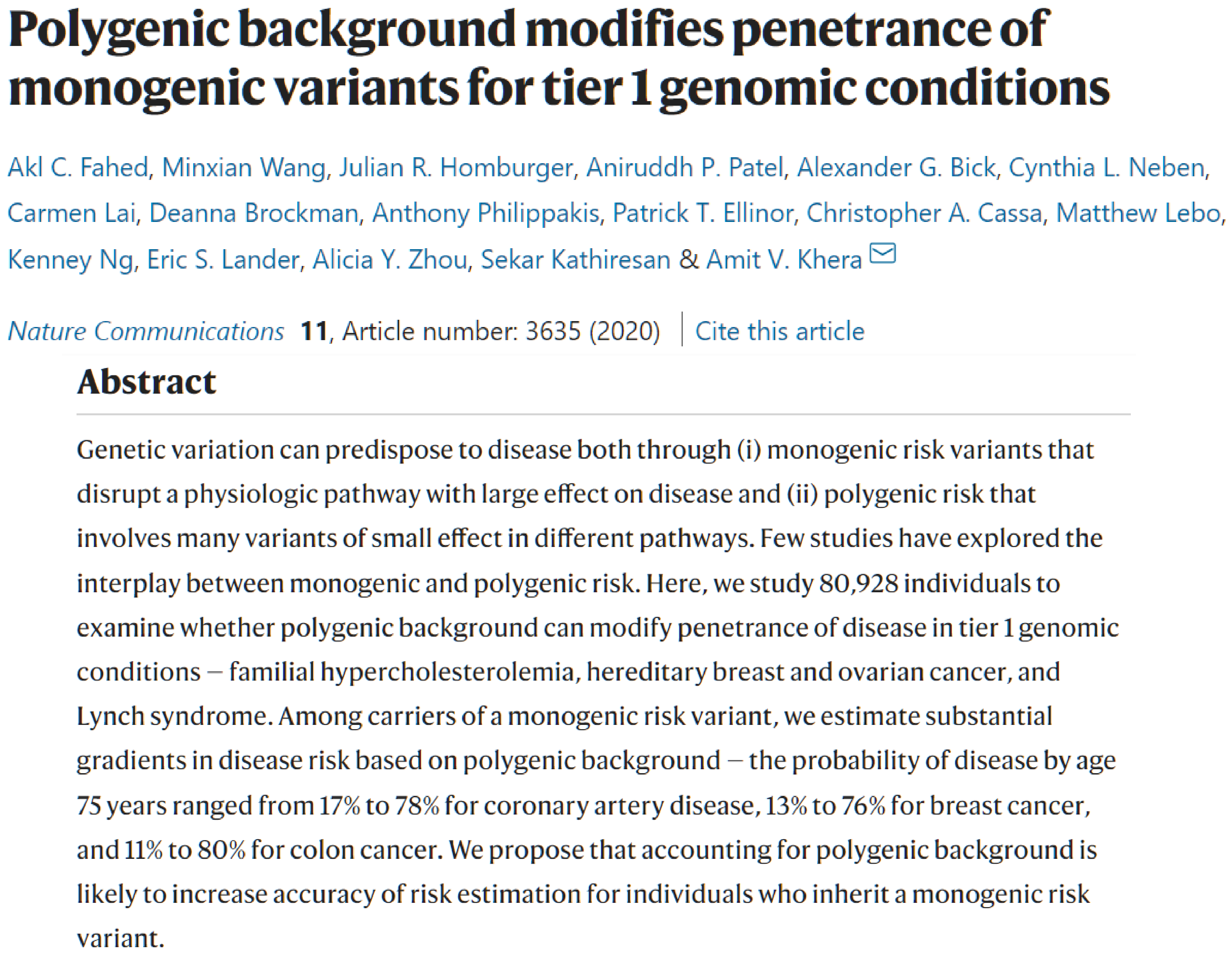


What Is New?

- Whole-genome sequencing was performed and analyzed in 2081 patients presenting to a US hospital with early-onset (age ≤55 years) myocardial infarction.
- A monogenic mutation, a single mutation that significantly increases risk, related to familial hypercholesterolemia was identified in 1.7% of the patients and was associated with a 3.8-fold increased odds of myocardial infarction.
- High polygenic score, reflective of the cumulative impact of many common variants and defined as the top 5% of the control population distribution, was identified in 10 times as many patients (17%) and was associated with a similar 3.7-fold increased odds of myocardial infarction.

What Are the Clinical Implications?

- A polygenic score comprising 6.6 million common DNA variants can identify 5% of the population who inherit risk equivalent to that of a familial hypercholesterolemia mutation.
- Unlike familial hypercholesterolemia mutation carriers, who typically have high low-density lipoprotein cholesterol levels, "carriers" of a high polygenic score cannot be identified with conventional risk factors or biomarkers.
- These findings lay the scientific foundation for the systematic identification of individuals born with a substantially increased risk of myocardial infarction resulting from either a familial hypercholesterolemia mutation or high polygenic score and delivery of a lifestyle or pharmacological intervention to attenuate inherited risk.



Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions

Akl C. Fahed, Minxian Wang, Julian R. Homburger, Aniruddh P. Patel, Alexander G. Bick, Cynthia L. Neben, Carmen Lai, Deanna Brockman, Anthony Philippakis, Patrick T. Ellinor, Christopher A. Casa, Matthew Lebo, Kenney Ng, Eric S. Lander, Alicia Y. Zhou, Sekar Kathiresan & Amit V. Khera

Nature Communications **11**, Article number: 3635 (2020) | Cite this article

Abstract

Genetic variation can predispose to disease both through (i) monogenic risk variants that disrupt a physiologic pathway with large effect on disease and (ii) polygenic risk that involves many variants of small effect in different pathways. Few studies have explored the interplay between monogenic and polygenic risk. Here, we study 80,928 individuals to examine whether polygenic background can modify penetrance of disease in tier 1 genomic conditions – familial hypercholesterolemia, hereditary breast and ovarian cancer, and Lynch syndrome. Among carriers of a monogenic risk variant, we estimate substantial gradients in disease risk based on polygenic background – the probability of disease by age 75 years ranged from 17% to 78% for coronary artery disease, 13% to 76% for breast cancer, and 11% to 80% for colon cancer. We propose that accounting for polygenic background is likely to increase accuracy of risk estimation for individuals who inherit a monogenic risk variant.

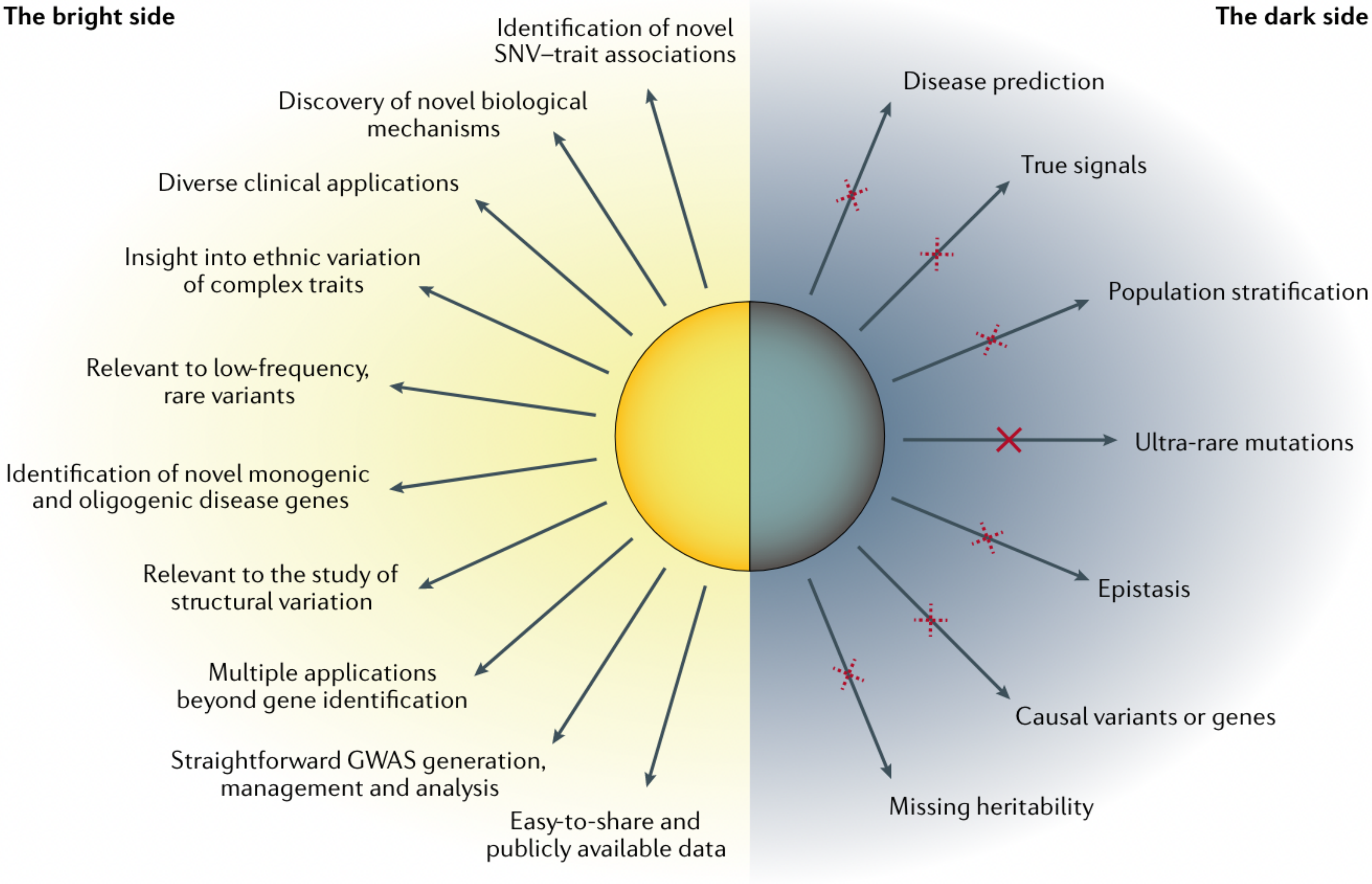

The bright side
Identification of novel SNV–trait associations
Discovery of novel biological mechanisms
Diverse clinical applications
Insight into ethnic variation of complex traits
Relevant to low-frequency, rare variants
Identification of novel monogenic and oligogenic disease genes
Relevant to the study of structural variation
Multiple applications beyond gene identification
Straightforward GWAS generation, management and analysis
Easy-to-share and publicly available data
The dark side
Disease prediction
True signals
Population stratification
Ultra-rare mutations
Epistasis
Causal variants or genes
Missing heritability

Fig. 4 | **Benefits and limitations of GWAS using SNP arrays.** A visual depiction of the current benefits (the bright side) and limitations (the dark side) of genome-wide association studies (GWAS). The solid X indicates a permanent limitation. The dotted Xs represent limitations that have the potential to be overcome, at least to some extent, in the future (for example, with larger sample sizes, technological and methodological advancements, and a shift from the use of single-nucleotide polymorphism (SNP) arrays to whole-genome sequencing). SNV, single-nucleotide variant.

Tam (2019) *Nat Rev Genet*

Further reading

- Polderman TJC, Benyamin B, de Leeuw CA, et al (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet 47:702–709. https://doi.org/10.1038/ng.3285
- Selvaraj MS, Li X, Li Z, et al (2022) Whole genome sequence analysis of blood lipid levels in >66,000 individuals. Nat Commun 13:5995. https://doi.org/10.1038/s41467-022-33510-7
- Uffelmann E, Huang QQ, Munung NS, et al (2021) Genome-wide association studies. Nat Rev Methods Primers 1:1–21. https://doi.org/10.1038/s43586-021-00056-9
- Xu Y, Ritchie SC, Liang Y, et al (2023) An atlas of genetic scores to predict multi-omic traits. Nature. https://doi.org/10.1038/s41586-023-05844-9
- Khera AV, Chaffin M, Aragam KG, et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50:1219–1224. https://doi.org/10.1038/s41588-018-0183-z
- Khera AV, Chaffin M, Zekavat SM, et al (2019) Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. Circulation 139:1593–1602. https://doi.org/10.1161/CIRCULATIONAHA.118.035658
- Fahed AC, Wang M, Homburger JR, et al (2020) Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nature Communications 11:3635. https://doi.org/10.1038/s41467-020-17374-3
- Tam V, Patel N, Turcotte M, et al (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467–484. https://doi.org/10.1038/s41576-019-0127-1