Медицинская геномика

Василий Евгеньевич Раменский
Анастасия Александровна Жарикова и Мария Ильинична Зайченока

ramensky@gmail.com, azharikova89@gmail.com

НМИЦ Терапии и профилактической медицины
Факультет биоинженерии и биоинформатики МГУ
Институт искусственного интеллекта МГУ

2024

Комплексные (мультифакторные) заболевания

План лекции

1. Линейные модели
2. Моногенные и комплексные заболевания
3. Аллельная архитектура генетических заболеваний
4. Основы полногеномного поиска ассоциаций (GWAS)
5. Применение GWAS, шкалы генетического риска

Линейная модель с аддитивными генетическими эффектами

$Y =$	G	$+E$	$+\epsilon$
Фенотип	Генотип	Окружающая среда	случайная величина

Линейная модель с аддитивными генетическими эффектами

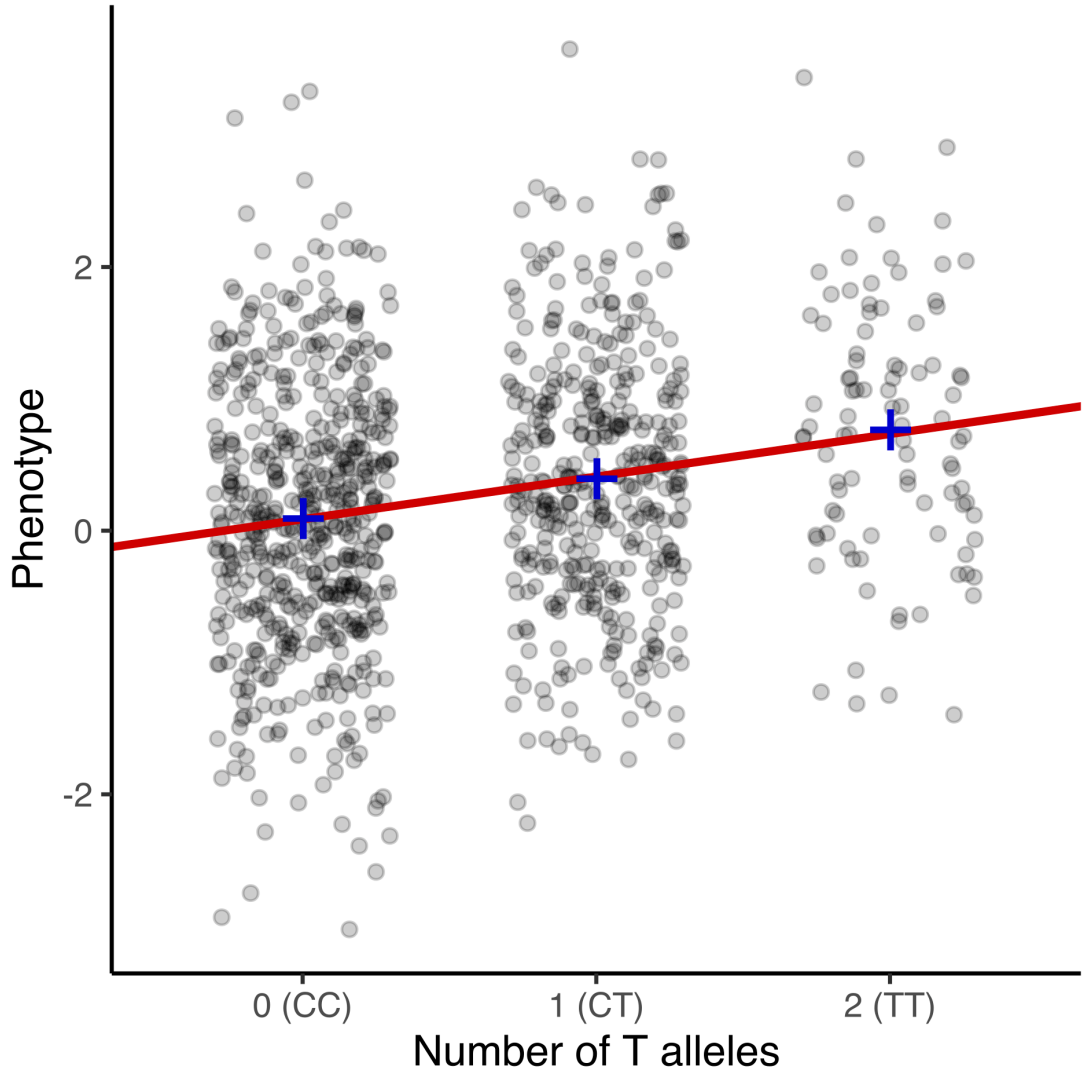
$$\underset{\text{Фенотип}}{Y =} \quad \underset{\text{Генотип}}{G} \quad \underset{\text{Окружающая среда}}{+E} \quad \underset{\text{случайная величина}}{+\epsilon}$$

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

- Y_i – фенотип i-го индивидуума
- μ – начальный уровень фенотипа
- G_i – генотип, число эффекторных аллелей a: $G(AA) = 0, G(Aa) = 1, G(aa) = 2$
- β – эффект: изменение значения на каждую копию эффекторного аллеля (количественный признак) или логарифм отношения шансов (бинарный признак)
- $\alpha_j X_{ij}$ – ковариаты: пол, возраст, статус курения, прием лекарств, этнос, другие варианты

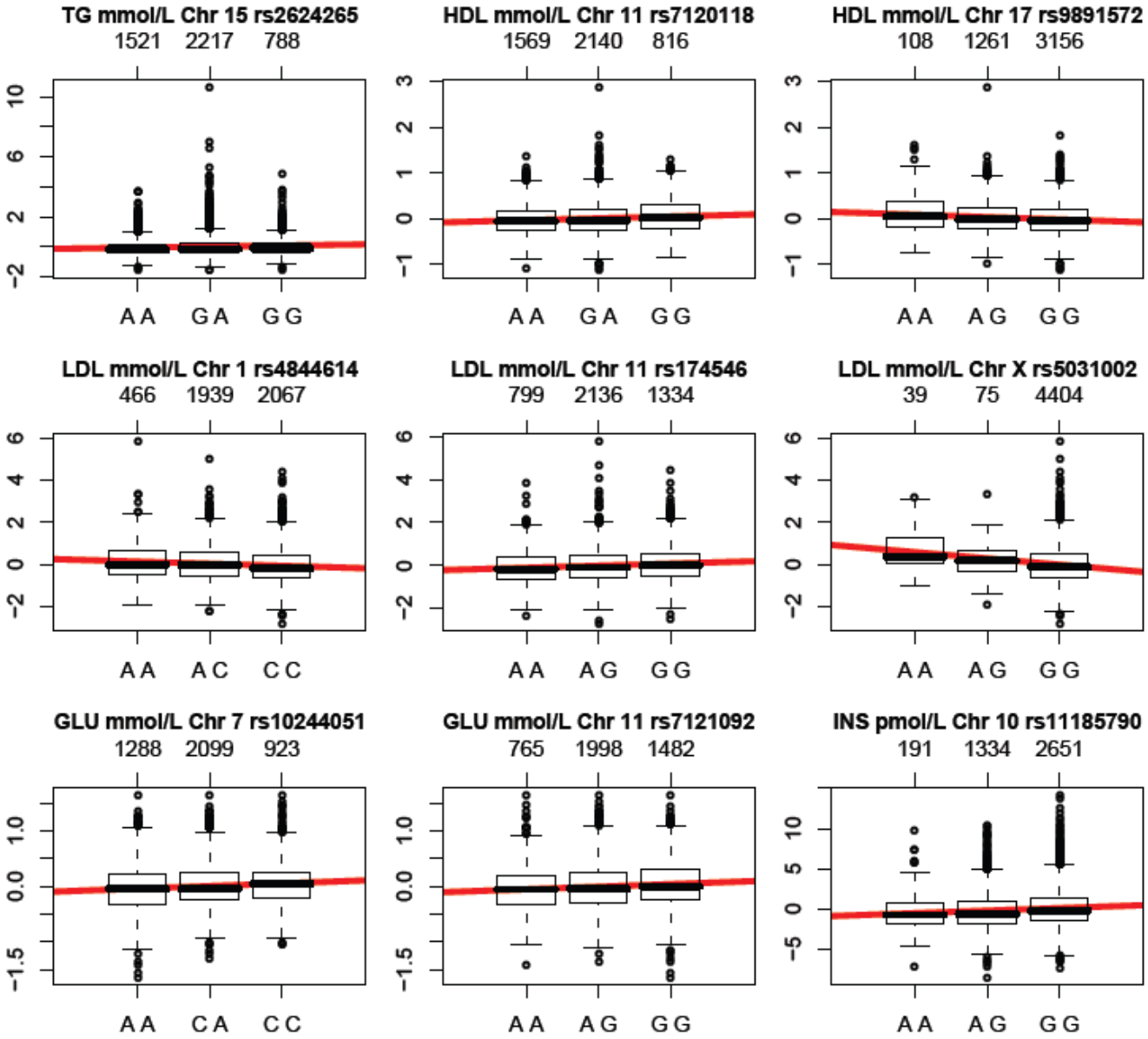
Morris and Cardon (2019) *Handbook of Stat Genomics*

Линейная модель с аддитивными генетическими эффектами

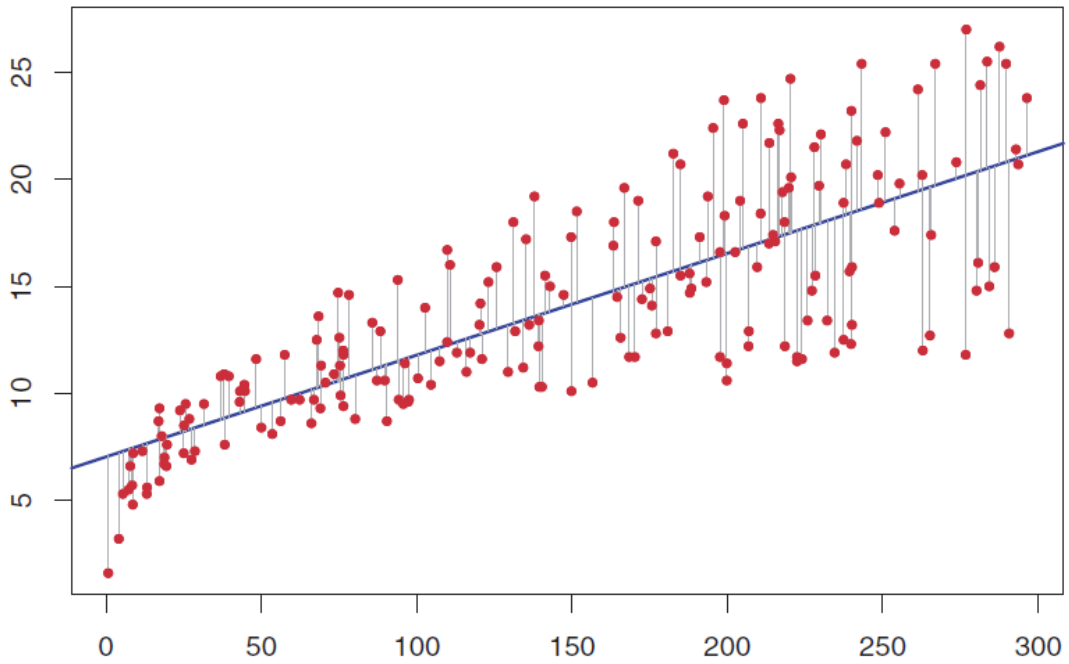


https://en.wikipedia.org/wiki/Genome-wide_association_study

Более реалистичный пример // Sabatti (2009) *Nat Genet*



Простая модель линейной регрессии



$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

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

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

Оценка связи независимой и зависимой переменной

- P – вероятность нулевой гипотезы H_0 (нет взаимосвязи)
- β – коэффициент регрессии (оценка эффекта)
- R^2 – коэффициент детерминации: пропорция вариабельности в Y, которая может быть объяснена X

Порог восприимчивости/уязвимости

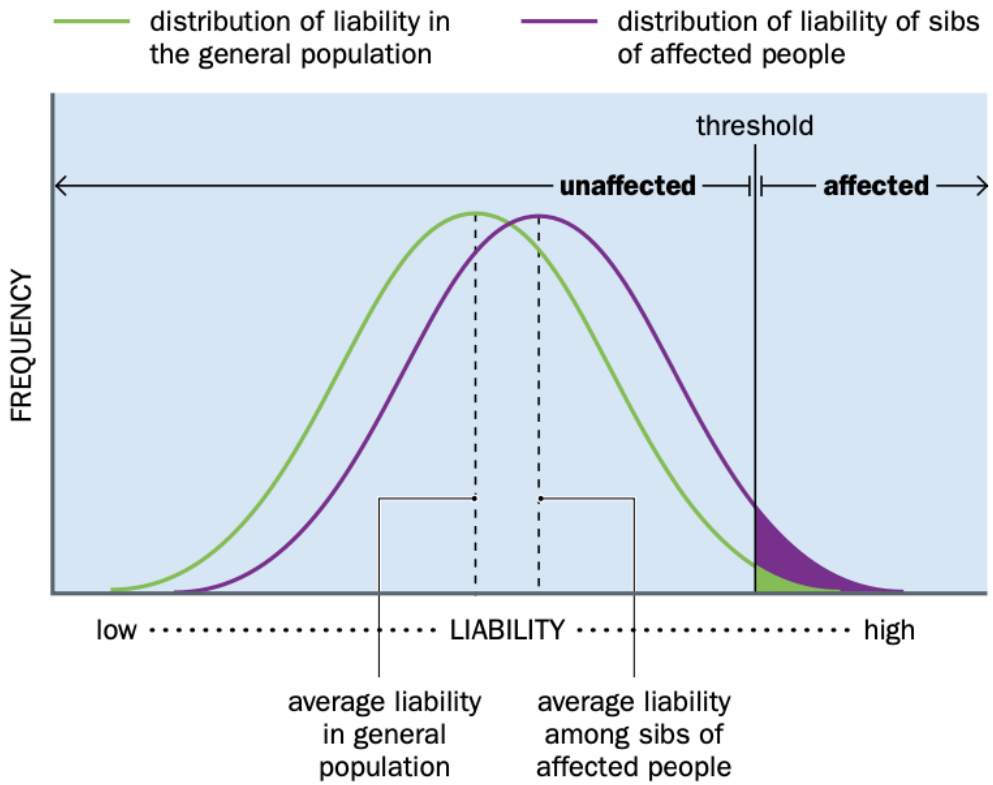


Figure 5.23 A polygenic threshold model for dichotomous non-Mendelian characters. Liability to the condition is polygenic and Normally distributed (green curve). People whose liability is above a certain threshold value (the balance point in **Figure 5.22**) are affected. The distribution of liability 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.

Объясняет неменделевское накопление (обогащение) бинарных признаков (например, комплексных заболеваний) в потомстве.
В отличие от менделевских заболеваний, риск повторения увеличивается с количеством уже больных детей.

Strachan, Read – *Human Molecular Genetics*

Менделевские	**Комплексные**
Индивидуально редки в популяции	Часты в популяции
Паттерны наследования в семье: AD, AR, и т.д.	Неменделевское накопление в семьях
Один или несколько генов с сильным эффектом	Несколько локусов, нет одного необходимого и достаточного локуса
Вызывается кодирующим аллелем с большой или полной пенетрантностью	Сложная аллельная архитектура, некодирующие варианты
В основном генетические факторы	Комбинация генетических факторов, факторов окружающей среды и образа жизни
Примеры: муковисцидоз, семейная гиперхолистеренемия, наследственные кардиомиопатии, нарушения ритма сердца	Примеры: ИБС, артериальная фибрилляция, гипертензия, шизофрения, сердечная недостаточность

Аллельная архитектура генетических заболеваний

Редкие (Менделевские) заболевания

- Очень редкие (AF<1%) и сильно вредные варианты
- Подвержены балансу мутирования и отбора

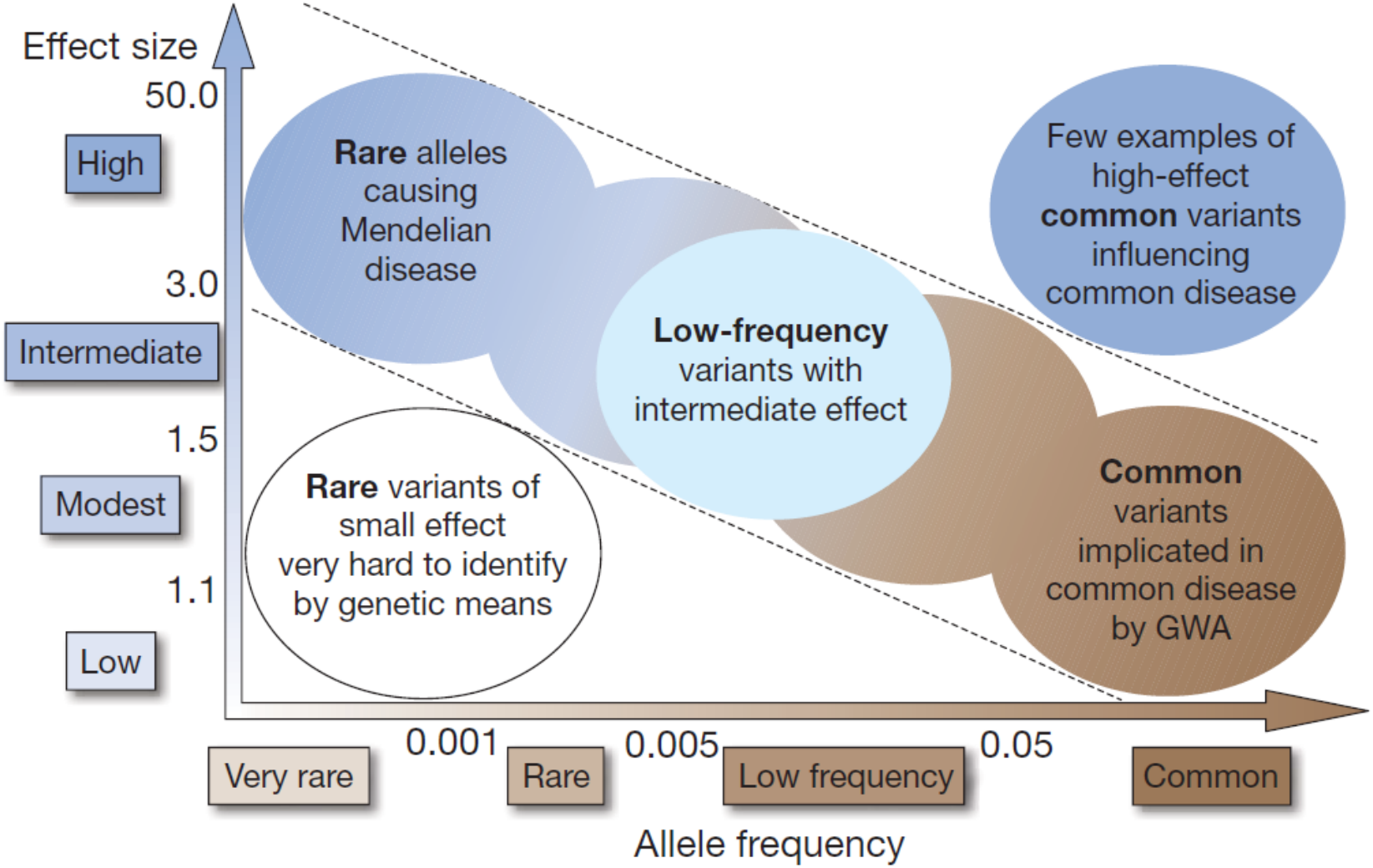
Комплексные заболевания, частые варианты //Reich, Lander (2001)

- Относительно небольшое количество старых, частых (AF>1%) вариантов
- Не подвержены отбору?
  - Начинается после периода размножения, не подвержены очищающему отбору (Диабет 2 типа)
  - Балансирующие отбор (Заболевание почек/стойкость к паразитам в Африке)
  - Гипотеза «бережливости» (ожирение, диабет)

Комплексные заболевания, редкие варианты // Pritchard (2001) AJHG

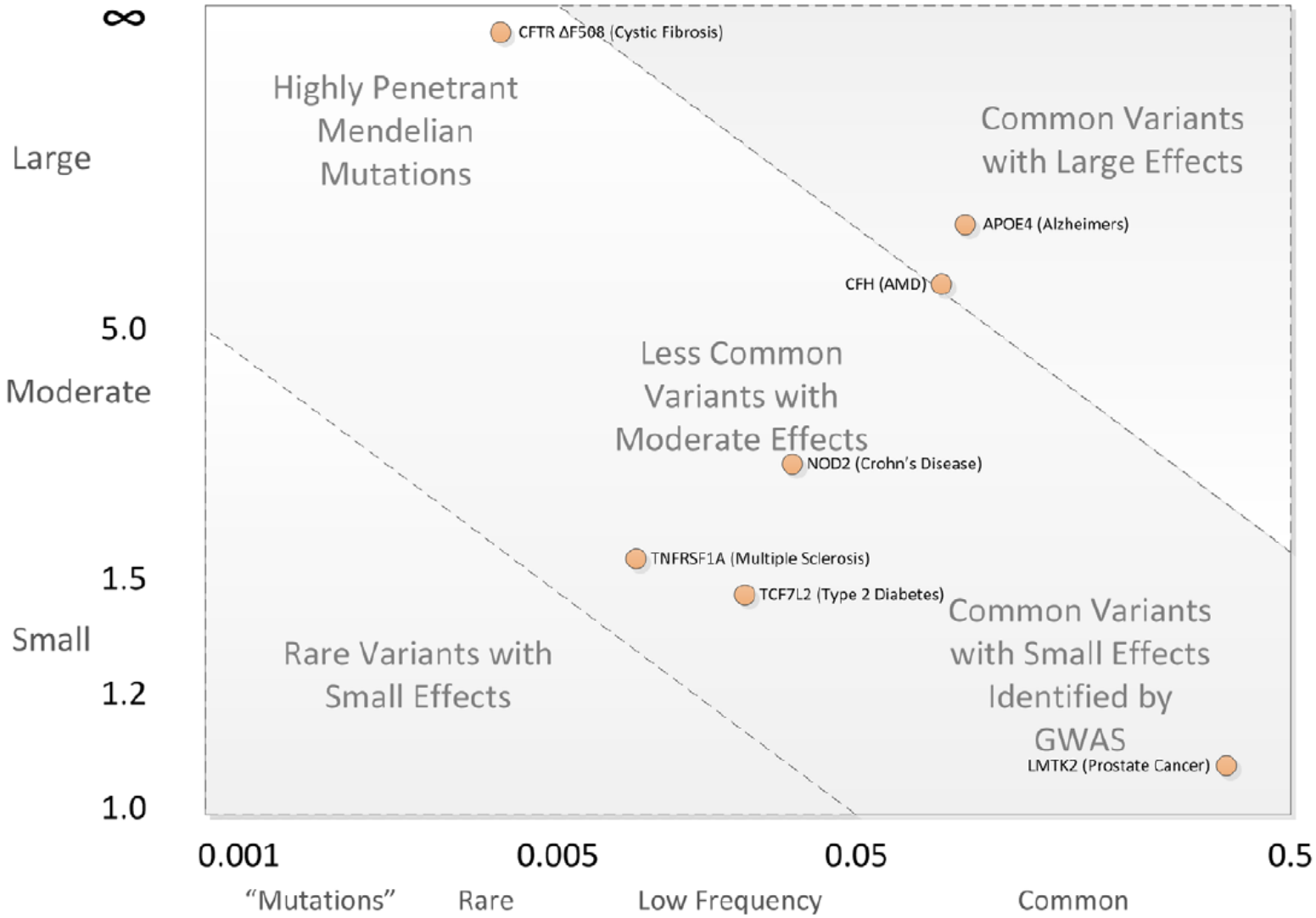
- Множество редких (AF<1%) вариантов со средним эффектом
- Недавнее увеличение размера человеческой популяции $\Rightarrow$ множество слабо повреждающих вариантов

Аллельная архитектура генетических заболеваний



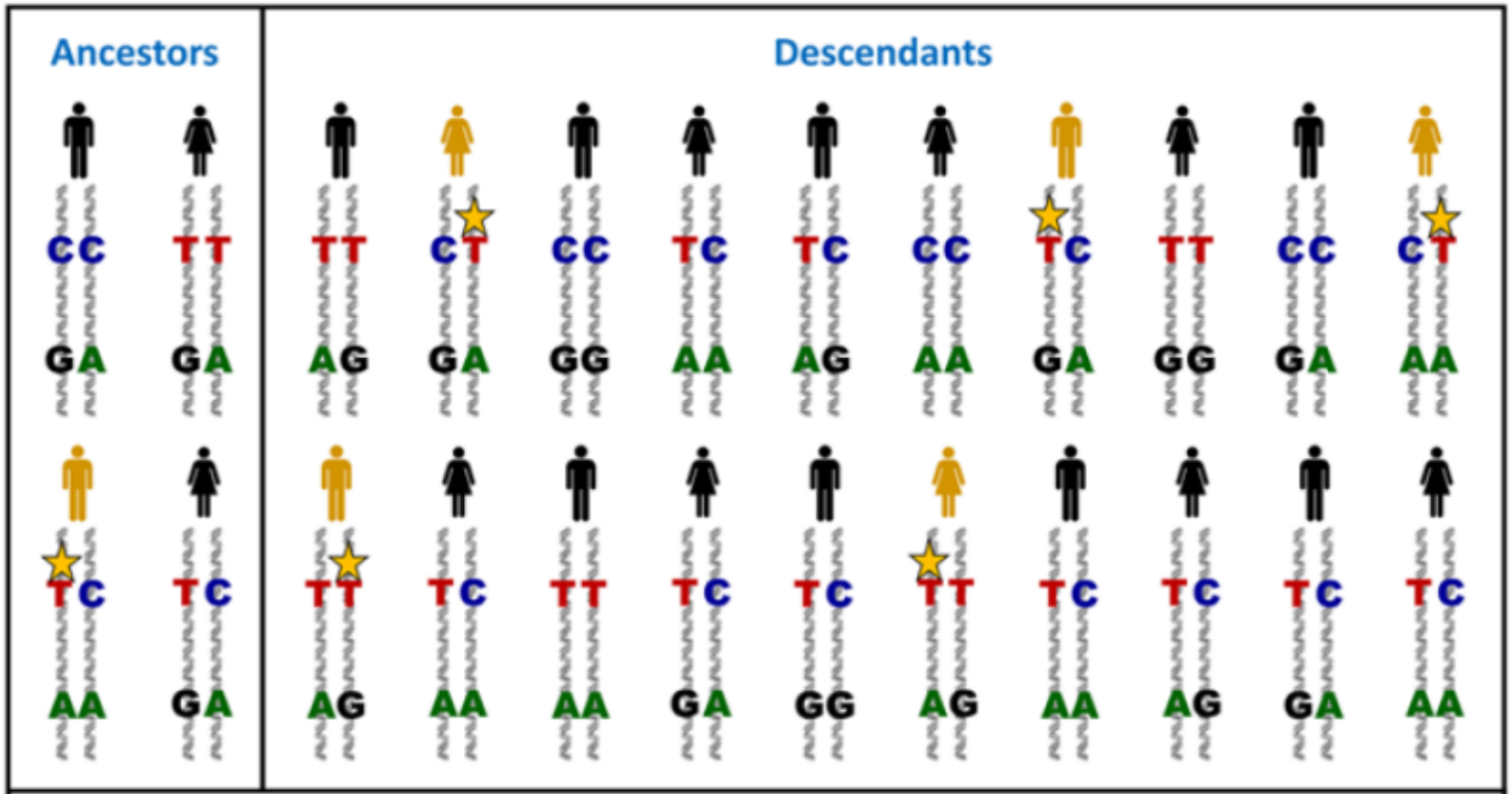
Manolio (2009) *Nature*

Аллельная архитектура генетических заболеваний



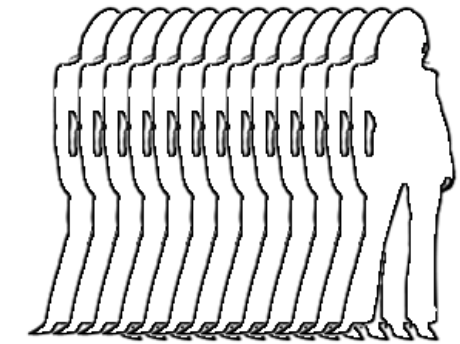
Bush and Moore (2012) *PloS Comp Bio*

Jackson (2018) *Essays Biochem*

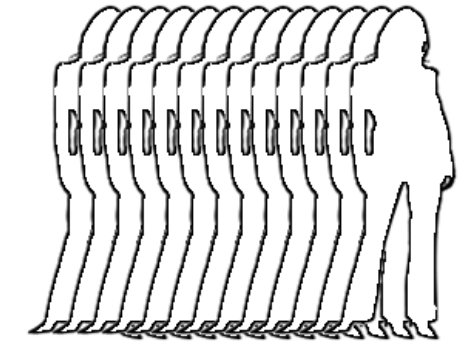


Allele counts (descendants)

	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%)



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

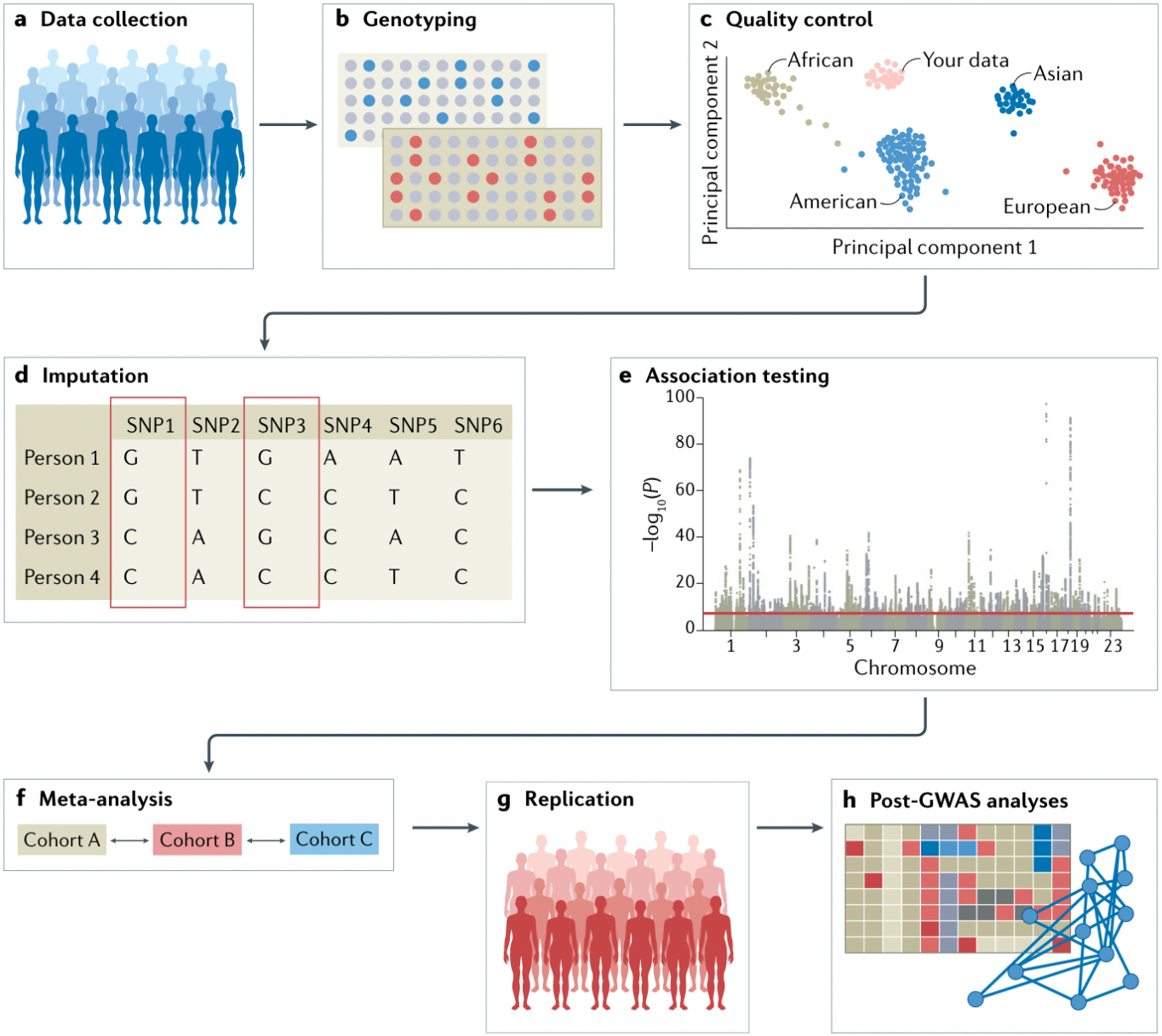
SNP1	*SNP2*	*SNP ...*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$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

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

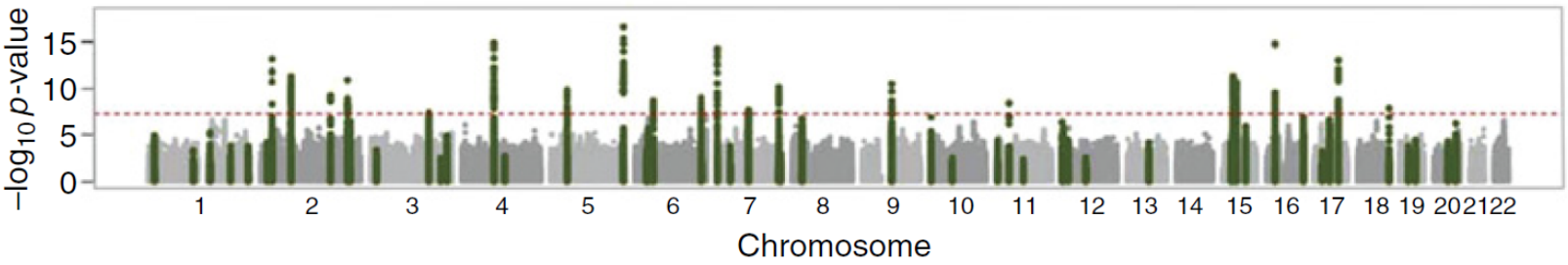
Пример сводной статистики GWAS // Selvaraj (2022) *Nat Commun*

	A	B	C	D	E	F	G	H	I	J	K
1	**TOPMed Freeze 8 variants summary.** Variants which passed the significance criteria were clumped (window 250 kb, r2 0.5) and compared agains summary statistic and GWAS catalog. Variants were binned to three categories, Known-Position (variant previously associated), Known-Loci (varia previously significantly associated with the corresponding lipid phenotype but within 500 kb of a known locus) and Novel. The list of variants is tabu each lipid phenotype and each category of is ordered based on chromosome position. Summary statistics reported were obtained from two-sided g 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										
2											
3	**HDL**										
4	**CHR**	**POS**	**A1**	**A2**	**rs_dbSNP151**	**BETA**	**SE**	**p.value**	**MAF**	**VEP ensembl precedent consequence**	**VEP ensembl precedent gene**	**Ca**
5	2	21008652	G	A	rs676210	0.978	0.111	1.08E-18	0.246	missense_variant	*APOB*	Known
6	7	17872129	G	T	rs1917368	-0.595	0.093	1.91E-10	0.468	intron_variant	*SNX13*	Known
7	7	80671133	T	G	rs3211938	2.823	0.290	1.93E-22	0.026	stop_gained	*CD36*	Known
8	8	9326086	A	G	rs4841132	1.782	0.149	9.17E-33	0.101	non_coding_transcript_e	*AC022784.1*	Known
9	9	104827463	C	T	rs4149307	1.240	0.104	1.61E-32	0.388	intron_variant	*ABCA1*	Known
10	11	116830638	G	A	rs138326449	12.784	1.142	4.21E-29	0.002	splice_donor_variant	*APOC3*	Known
11	11	61785208	G	T	rs174537	-0.859	0.105	2.98E-16	0.301	intron_variant	*TMEM258*	Known
12	12	124853983	C	T	rs10773112	0.923	0.094	1.49E-22	0.394	intron_variant	*SCARB1*	Known
13	17	43848758	C	T	rs72836561	-3.641	0.345	4.87E-26	0.017	missense_variant	*CD300LG*	Known
14	18	49583585	A	G	rs77960347	4.739	0.494	8.39E-22	0.008	missense_variant	*LIPG*	Known
15	19	54295230	G	A	rs380267	-0.987	0.119	8.87E-17	0.202	downstream_gene_varia	*AC245884.12*	Known
16	19	8364439	G	A	rs116843064	4.256	0.375	7.11E-30	0.014	missense_variant	*ANGPTL4*	Known
17	19	44908684	T	C	rs429358	-1.649	0.125	8.06E-40	0.153	downstream_gene_varia	*TOMM40*	Known
18	20	44413724	C	T	rs1800961	-2.561	0.290	1.06E-18	0.024	missense_variant	*HNF4A*	Known
19	1	109274623	C	T	rs11102967	-0.642	0.100	1.49E-10	0.435	3_prime_UTR_variant	*CELSR2*	Known
20	1	109274968	G	T	rs12740374	0.900	0.109	1.56E-16	0.214	3_prime_UTR_variant	*CELSR2*	Known
21	1	230144512	C	G	rs11122400	0.585	0.092	1.91E-10	0.417	intron_variant	*GALNT2*	Known
22	1	230148510	C	A	rs4846906	0.767	0.129	2.75E-09	0.143	intron_variant	*GALNT2*	Known
23	1	230158438	A	T	rs910502	1.115	0.155	7.44E-13	0.093	intron_variant	*GALNT2*	Known

Основные этапы GWAS // Uffelmann (2021) *Nat Rev Methods*



Визуализация результатов GWAS



Manhattan plot Каждая точка соответствует SNP, помещенному на график в точке своей геномной позиции по оси x, ось y: сила ассоциации ($-log_{10}p - value$) по оси y. SNP, отмеченные зеленым соответствуют локусам, для которых ранее была показана ассоциация с признаком.

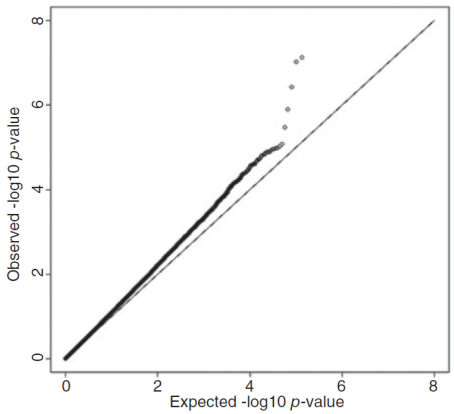


График квантиль-квантиль. Каждая точка соответствует SNP, ось y: ранжированные значения $-log_{10}p$, ось x: ожидаемые значения $-log_{10}p$ в случае нулевой гипотезы отсутствия ассоциации. Нахождение $-log_{10}p$ выше линии y = x указывает на неучтенную в анализе популяционную структуру. // Morris and Cardon (2019) *Handbook of Stat Genomics*

Визуализация результатов GWAS

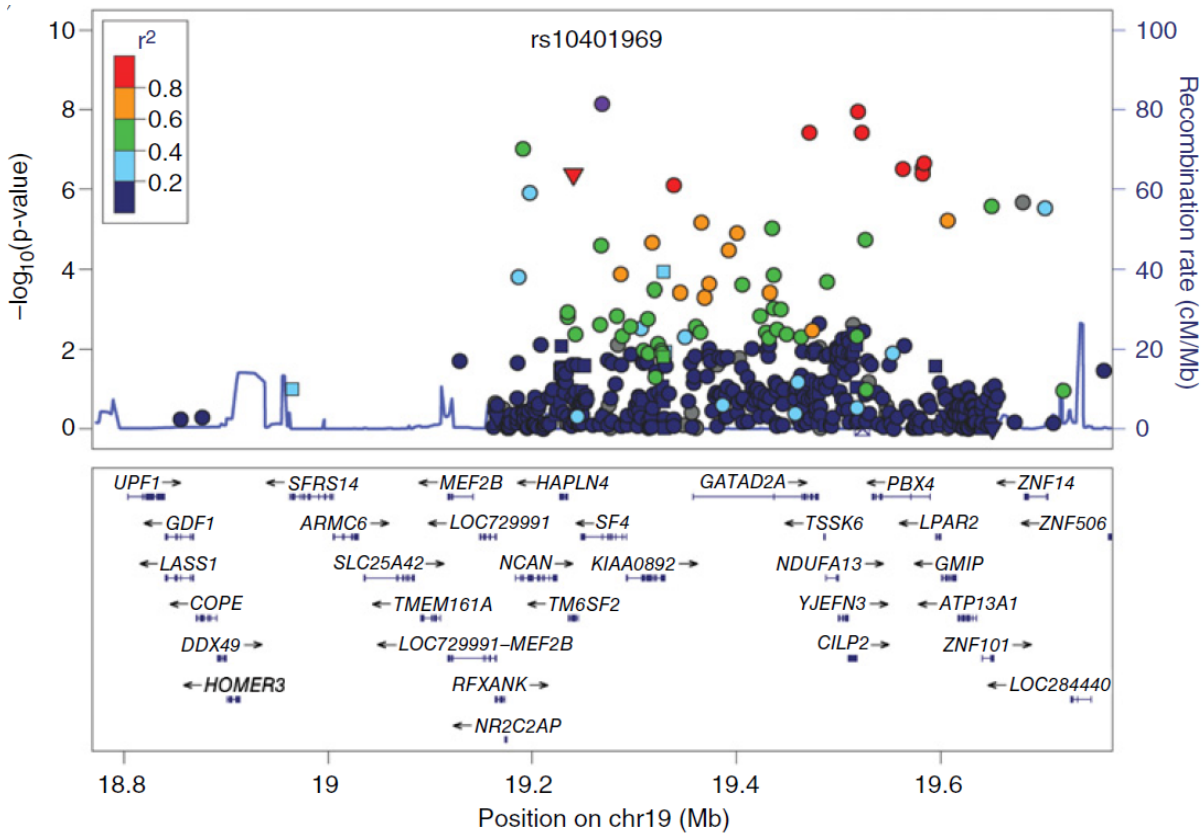
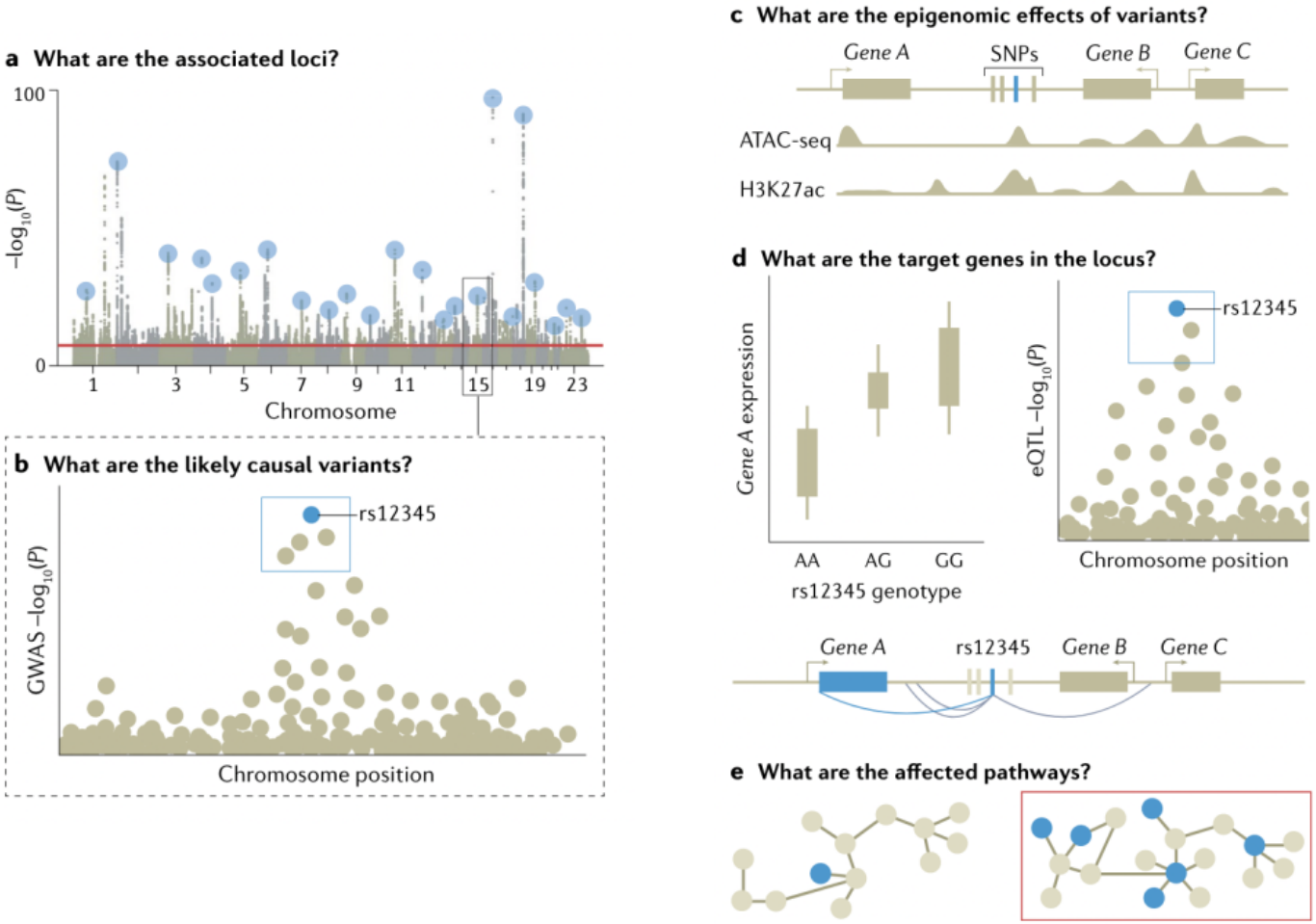


График сигнала. Лидирующие SNP: фиолетовый. Цвета: сцепление с лидирующим вариантом в гаплотипах европейского происхождения в 1000 Геномов. Форма: треугольник вверх для укорачивающего белок варианта, треугольник вниз для несинонимичных вариантов, квадрат для синонимичных или UTR, круг для интронных или некодирующих. Частота рекомбинация оценена по HapMap. Morris and Cardon (2019) *Handbook of Stat Genomics*

Исследования после GWAS // Uffelmann (2021) *Nat Rev Methods*



Ресурсы с результатами GWAS

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:

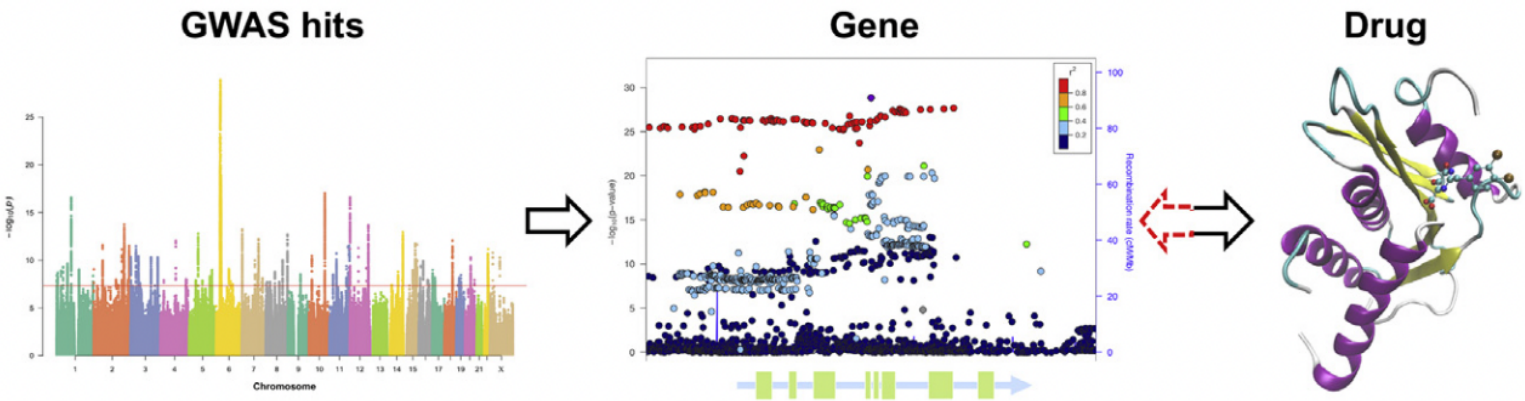
October 2022 (22.10)

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.

Мишени лекарств // Visscher (2017) *Am J Hum Genet*

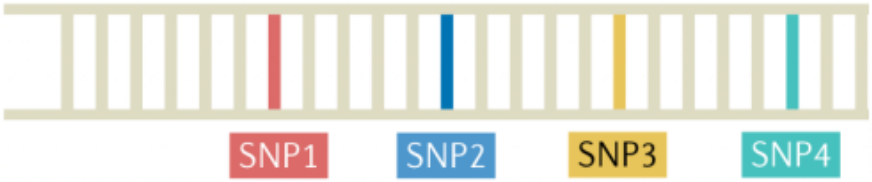


Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis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

Шкалы генетического риска // Uffelmann (2021) *Nat Rev Methods*

① GWAS summary statistics

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



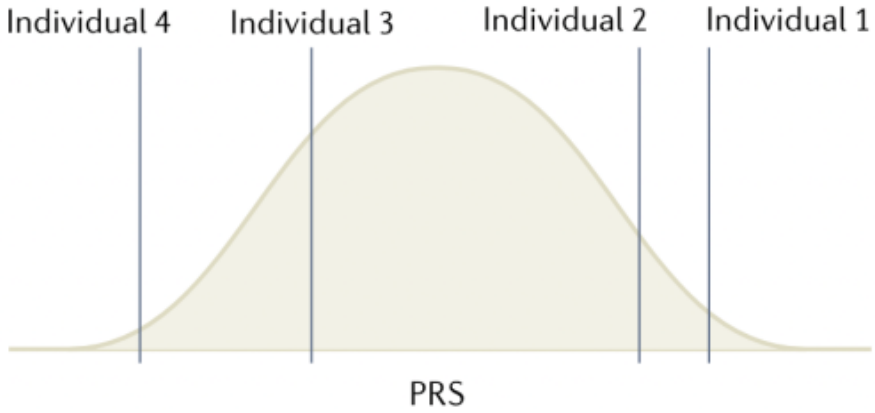
② Genotype data

	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



Шкалы генетического риска

- Шкалы генетического риска (ШГР): сумма количества аллелей, взвешенных по своей оценке эффекта
- Оценки эффекта аллелей получены из GWAS,

$$S = \sum_i^N \beta_i G_i, \text{ где } G_i = 0, 1, 2 \text{ и } N = 10^2 - 10^6$$

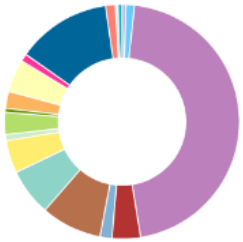
- Шкала генетического риска агрегирует эффекты многих генетических вариантов в одно-единственное число, которое предсказывает склонность к развитию фенотипа.
- ШГР можно рассчитать при рождении
- Носители высоких значений ШГР не могут быть идентифицированы по классическим факторам риска или биомаркерам
- Индивидуумы из топ-5% высоких значений ШГР для коронарной болезни сердца подвержены в 3.7 раз более высокому шансу инфаркта
- Полигенный фон может изменять пенетрантность моногенных мутаций

PGS catalog

Polygenic Scores
806

Traits
209

PGS Catalog | Home | Browse



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

ШГР для уровней липидов в Ивановской популяции

Иваново: 1,675 участников, 37,372 вариантов. HDL – липопротеины высокой плотности, TC – общий холестерин. Ковариаты: пол, возраст, ИМТ, прием статинов, курение, уровень ТТГ. Эмпирические значения p-value рассчитаны PRSice-2 через перестановки фенотипов.

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

ШГР для Иваново с оценками β для 132 вариантов // Selvaraj et al (2022) *Nat Comm*. LDL: липопротеины низкой плотности; HDL: липопротеины высокой плотности; TG:триглицериды; TC:общий холестерин.

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

nature genetics

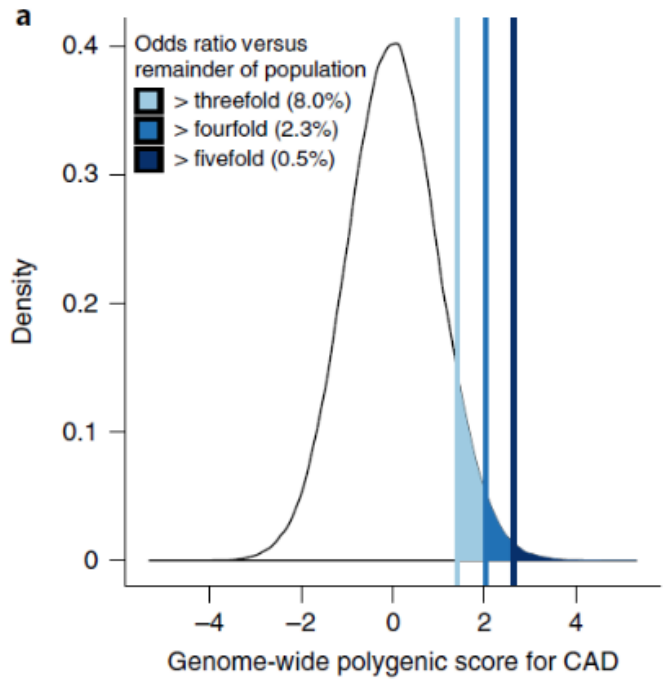
LETTERS

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. Cassa, 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.

Список литературы

- 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