Полногеномный анализ ассоциаций: модели и методы

Роман Сергеев

н.с. лаб. математической кибернетики
ОИПИ НАН Беларуси

Минск, 2019

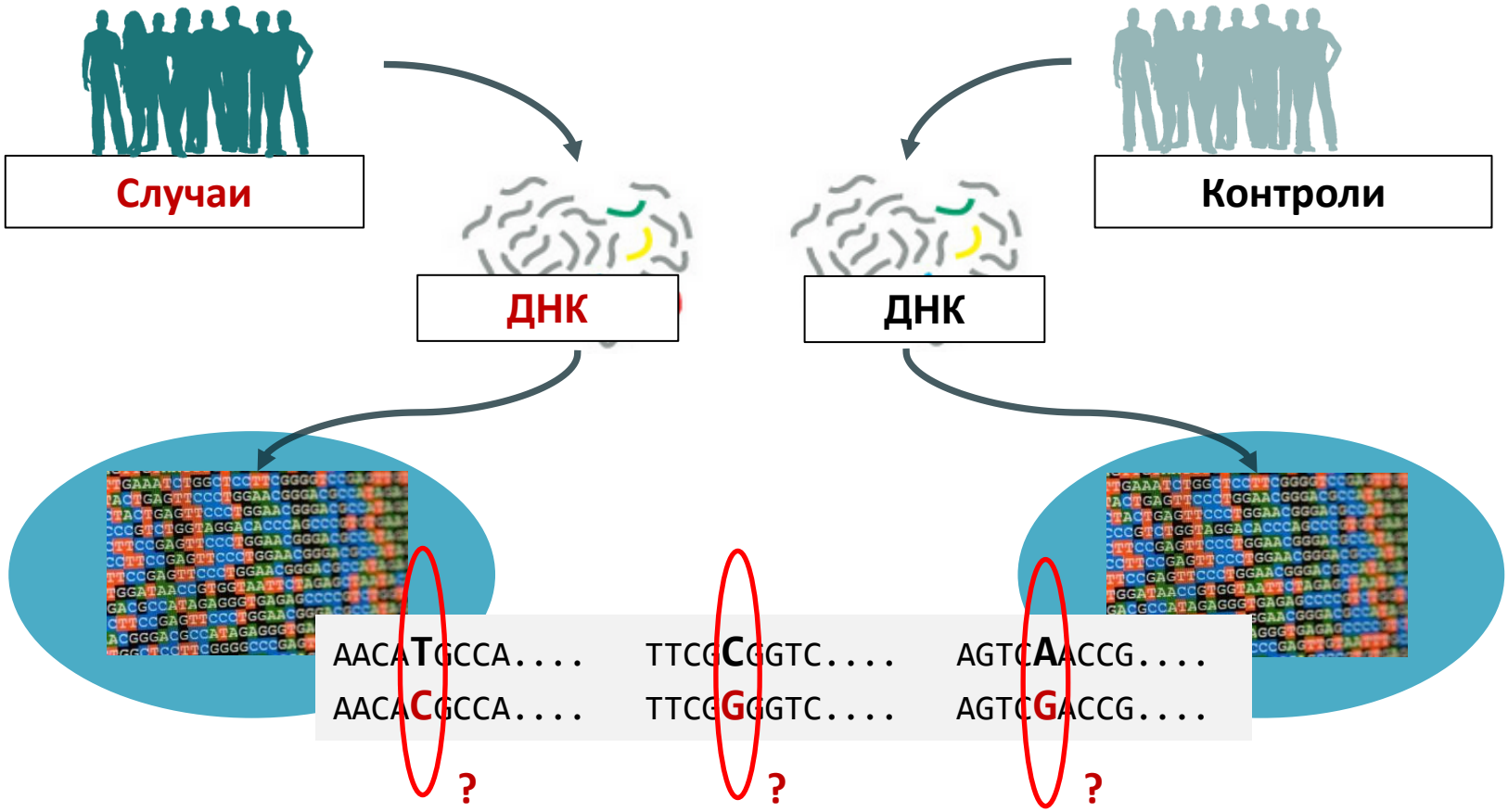
Содержание

1. Постановка задачи
2. Представление данных
3. Модели и методы
4. Анализ лекарственной устойчивости M.tuberculosis

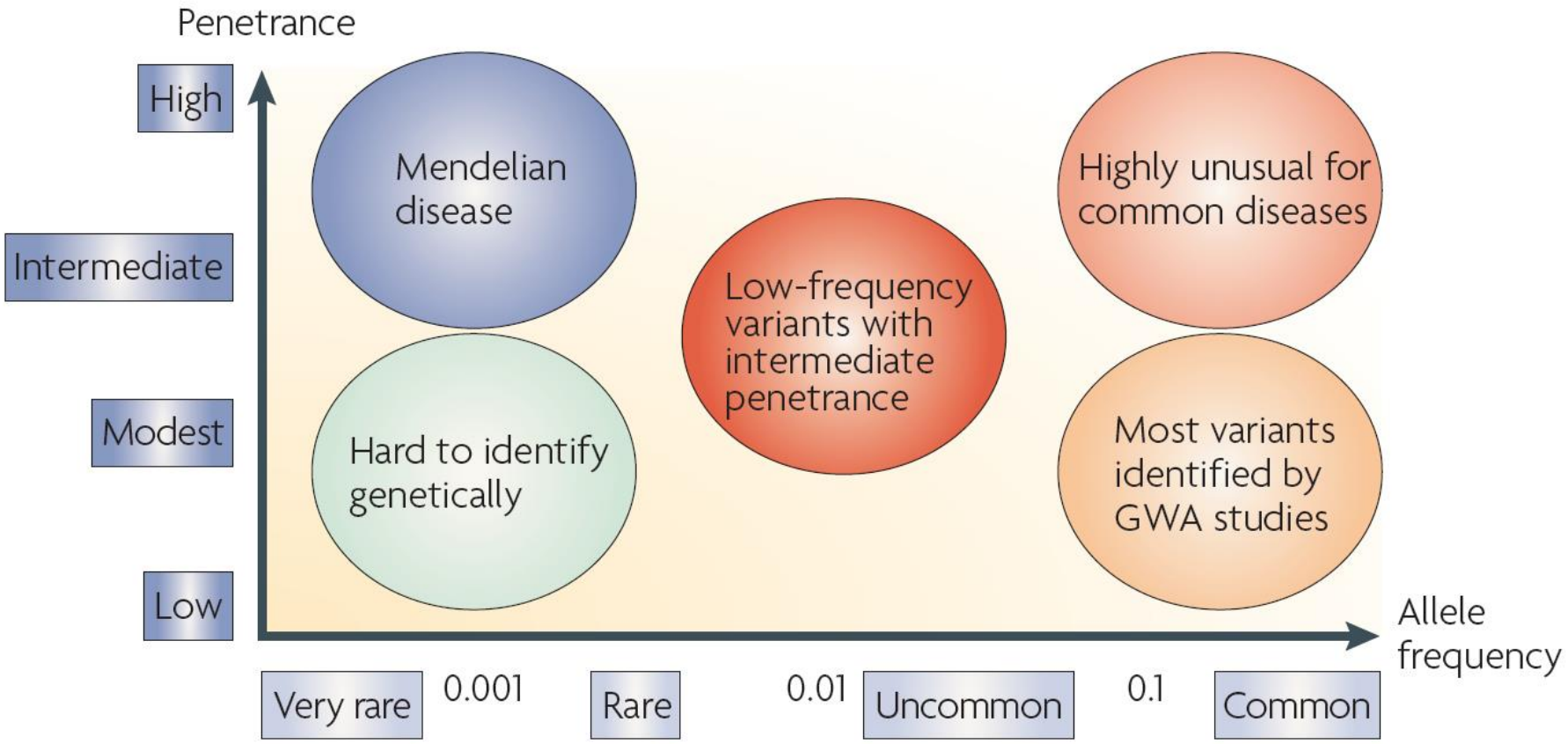
Постановка задачи

Полногеномный анализ ассоциаций (ПГАА)

- Цель ПГАА – поиск статистически значимых взаимосвязей между фенотипом и мутациями в геноме



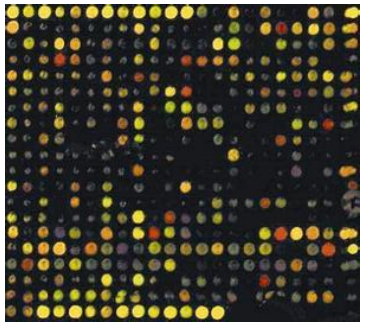
ПГАА для сложных признаков



McCarthy, M. I. *et al* Nature Reviews Genetics (2008)

Виды экспериментов ПГАА

- Эксперименты на *ДНК-микрочипах*
  - маркеры – распространенные SNPs
  - маркеры распределены по всему геному
- Цель – поиск маркеров геноме, которые ассоциированы с фенотипом
  - возможные функциональные мутации неравновесно сцеплены (LD – linkage disequilibrium) с маркерами
- Крупные исследования ПГАА стартовали с проекта HapMap в середине 2000-х

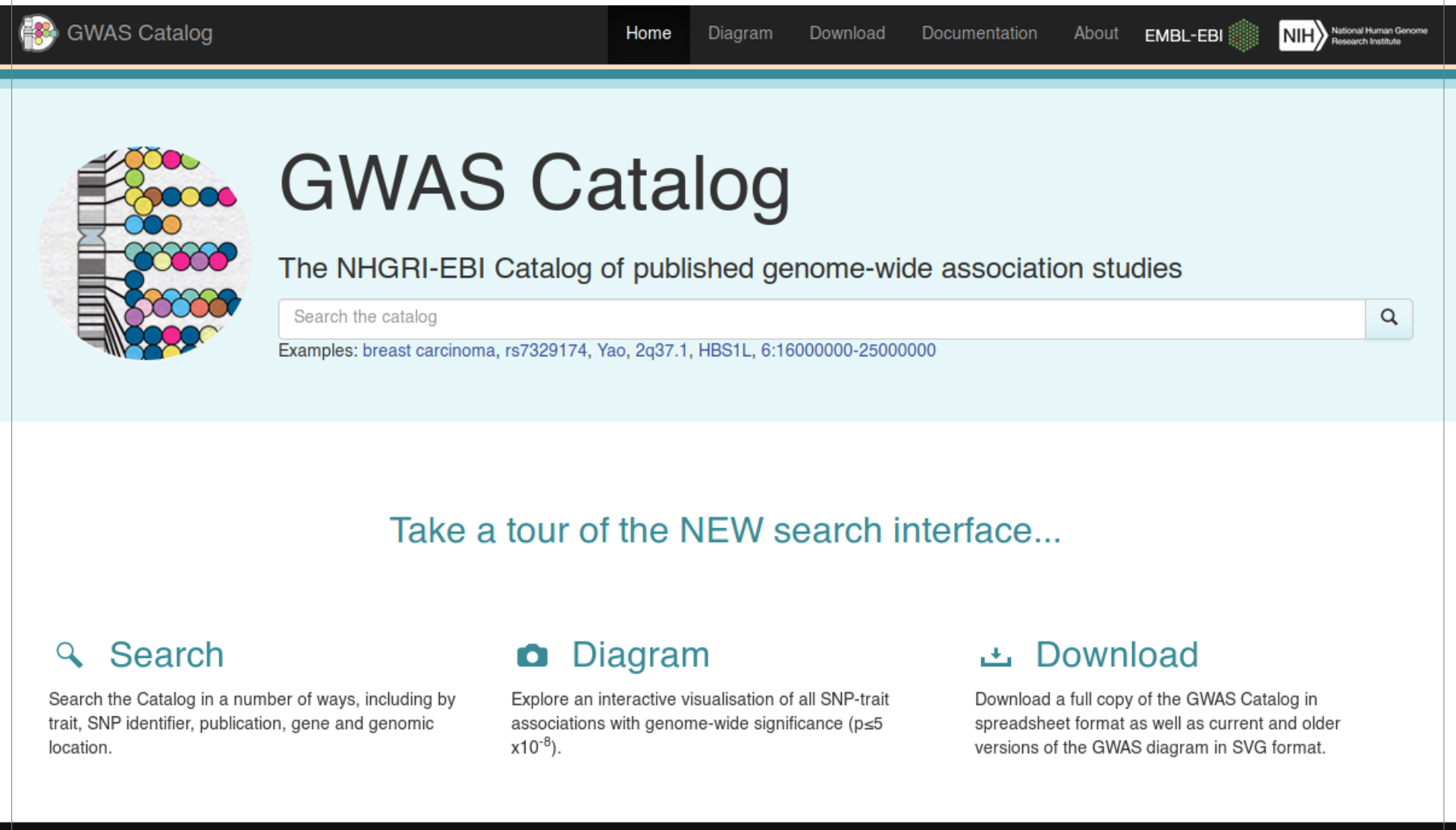


Виды экспериментов ПГАА

- Эксперименты на основе *полногеномного секвенирования*
  - получить и использовать сразу все мутации в геноме
  - позволяет обнаруживать редкие аллели, встречающиеся у <1% представителей популяции

- Часто используется при анализе бактериальных геномов

Каталог опубликованных исследований ПГАА



https://www.ebi.ac.uk/gwas/

Дизайн исследования ПГАА

- Прежде всего – что исследуем?
  - тип результирующей переменной (фенотипа)
- *Бинарная* => исследование “случай-контроль”
  - заболевшие / здоровые
  - чувствительные / устойчивые
- *Непрерывная* => количественный признак
  - уровень генной экспрессии
  - рост
  - уровень холестерина
- Наследуется ли исследуемый признак?

Представление данных

Тесты ПГАА в зависимости от данных

- Генотипические

	Генотипы			**Итого**
	aa	**aA**	**AA**	
Случай	r_0	r_1	r_2	r
Контроль	s_0	s_1	s_2	s
Итого	n_0	n_1	n_2	n

- Аллельные

	Аллели		**Итого**
	a	**A**	
Случаи	$u_0 = 2r_0 + r_1$	$u_1 = 2r_2 + r_1$	$u = 2r$
Контроли	$v_0 = 2s_0 + s_1$	$v_1 = 2s_2 + s_1$	$v = 2s$
Итого	$m_0 = 2n_0 + n_1$	$m_1 = 2n_2 + n_1$	$m = 2n$

Представление данных для теста ПГАА

Генотипы (по SNPs)

Матрица генотипов
(*признаки*)

$$X = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & \dots & 1 \\ 0 & 1 & 0 & 2 & 0 & 0 & \dots & 1 \\ 1 & 1 & 1 & 0 & 1 & 2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 0 & 2 & 0 & 1 & \dots & 1 \end{pmatrix}$$

Образцы

Вектор фенотипов
(*отклики*)
для эксперимента
“случай-контроль”

$$Y = \underbrace{(1, 0, 1, \dots, 1)^T}_{\text{Образцы}}$$

Контроль качества данных

- На уровне образца
  - пропуски
  - отклонения в гетерозиготности
  - дубликаты, родственные образцы
  - выбросы из популяции
- На уровне SNPs
  - пропуски / распределение пропусков по случаям и контролям
  - равновесие Харди-Вайнберга
  - MAF (minor allele frequency)

---

(1) Anderson C.A. et al. Data quality control in genetic case-control association studies. *Nature Protocols* 5, 1564-1573 (2010)
(2) Turner S. et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet.* Chapter 1: Unit1.19 (2011)

Модели и методы

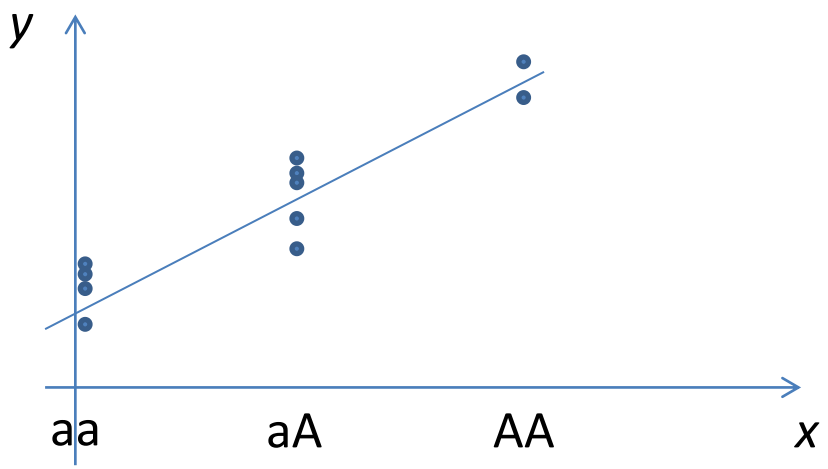
Модели регрессии

- Простейшая линейная регрессия для количественного признака:

$$y_i = \mu + \beta x_i + \epsilon_i$$

y_i - значение фенотипа для i-го образца

x_i - сколько раз (0,1 или 2) аллель ‘A’ встречается в рассматриваемом маркере i-го образца



Модели регрессии

- Логистическая регрессия для “случай-контроль” исследования:

$$ln\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)} = \mu + \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2}$$

y_i - значение фенотипа для i-го образца

$x_{i0} = I[n_{iA} = 0]$, $x_{i1} = I[n_{iA} = 1]$, $x_{i2} = I[n_{iA} = 2]$ – индикатор, что аллель ‘A’ встречается заданное число раз

- Также можно анализировать модели наследования:

$\beta_0 = 0, \beta_2 = 2\beta_1$ - аддитивная,
$\beta_0 = 0, \beta_2 = \beta_1$ - доминантная

Проверка гипотез и интерпретация

- Проверка статистических гипотез:

$$H_0: \beta_i = 0 \text{ vs } H_1: \beta_i \neq 0$$

- Как интерпретировать β_i?
  - *Размер эффекта* (effect size) в линейной регрессии: насколько сильно генотип влияет на фенотип
  - *Отношение шансов* (odds ration) в логистической регрессии

- Стандартная ошибка – в каких пределах может изменяться оценка для β?

- P-значение – какова значимость ассоциации?

Анализ таблиц сопряженности

- Критерий Хи-квадрат для анализа таблиц сопряженности

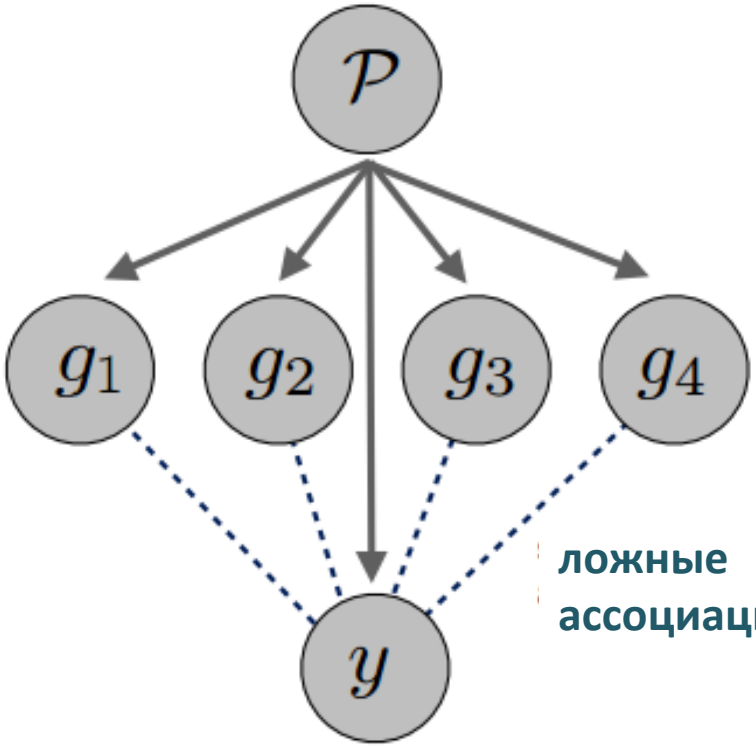
	Аллели		**Итого**
	a	**A**	
Случаи	u_0	u_1	u
Контроли	v_0	v_1	v
Итого	m_0	m_1	m

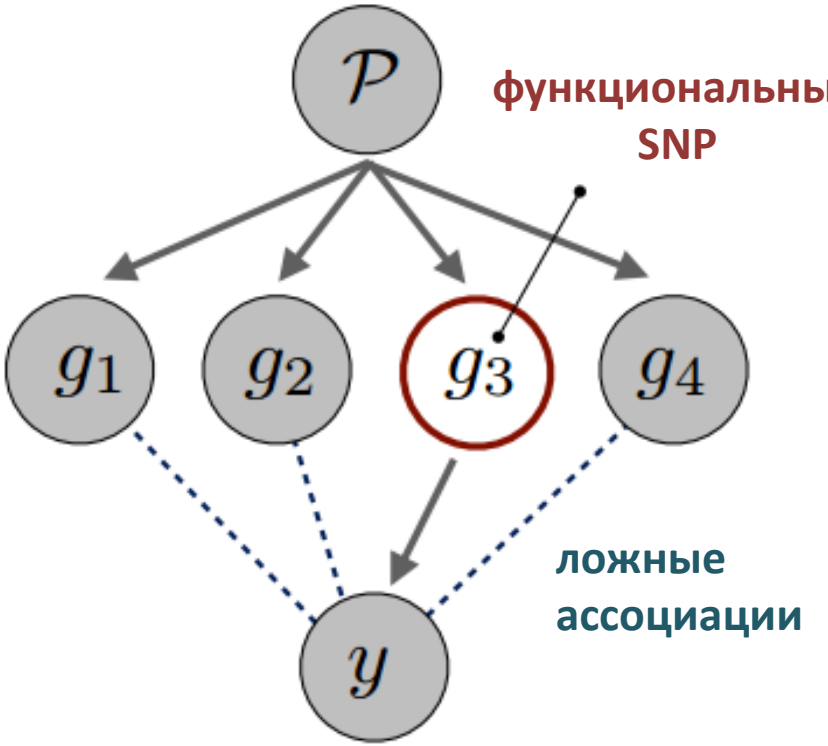
$$\chi^2_{\mathrm{df}} = \sum_i \frac{(Obs_i - Exp_i)^2}{Exp_i}, \quad \text{где} \quad Exp_1 = \frac{u\, m_0}{m}$$

df=(число столбцов-1) X (число строк-1)

Проблема ложных зависимостей

- *Confounders* – сторонние неконтролируемые факторы, являющиеся причиной ложных ассоциаций





Контроль дополнительных внешних факторов

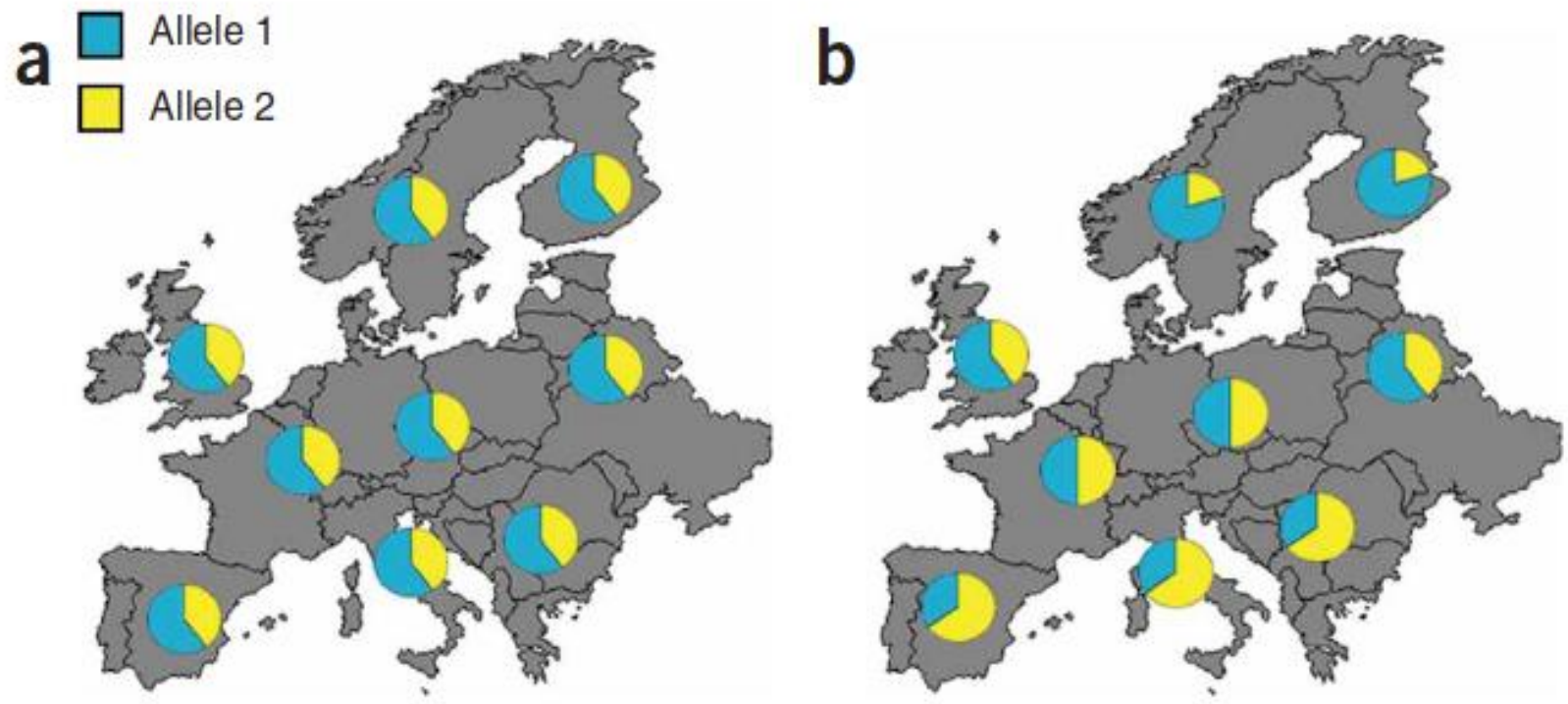
- В моделях регрессии – можно включать в модель:

$$y_i = \mu + \beta x_i + \underbrace{\alpha_1 C_{i1} + \alpha_2 C_{i2} + \cdots}_{\text{доп.факторы (ковариаты)}} + \epsilon_i$$

- В моделях на основе таблиц сопряженности – использовать стратификационный анализ:
  - тест Кохрана-Мантеля-Хаензеля (Cochran-Mantel-Haenszel) - проверяет ассоциации между генотипом и фенотипом в группах
  - тест Бреслоу-Дэйя (Breslow-Day) - проверяет однородность величины отношения шансов при переходе от группы к группе

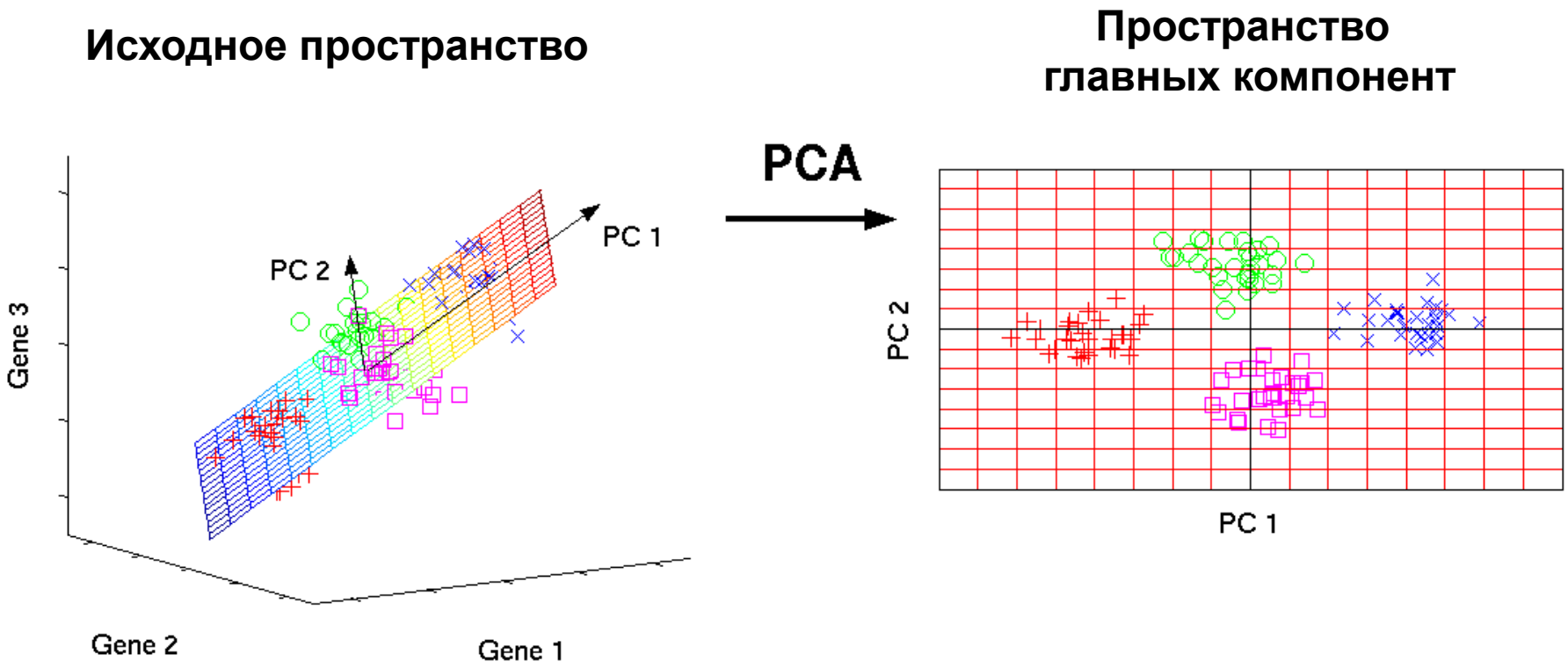
Популяционная подразделенность

- Распределение фенотипа соответствует распределению частот геномного маркера и зависит от субпопуляции => ложные зависимости (случай b)

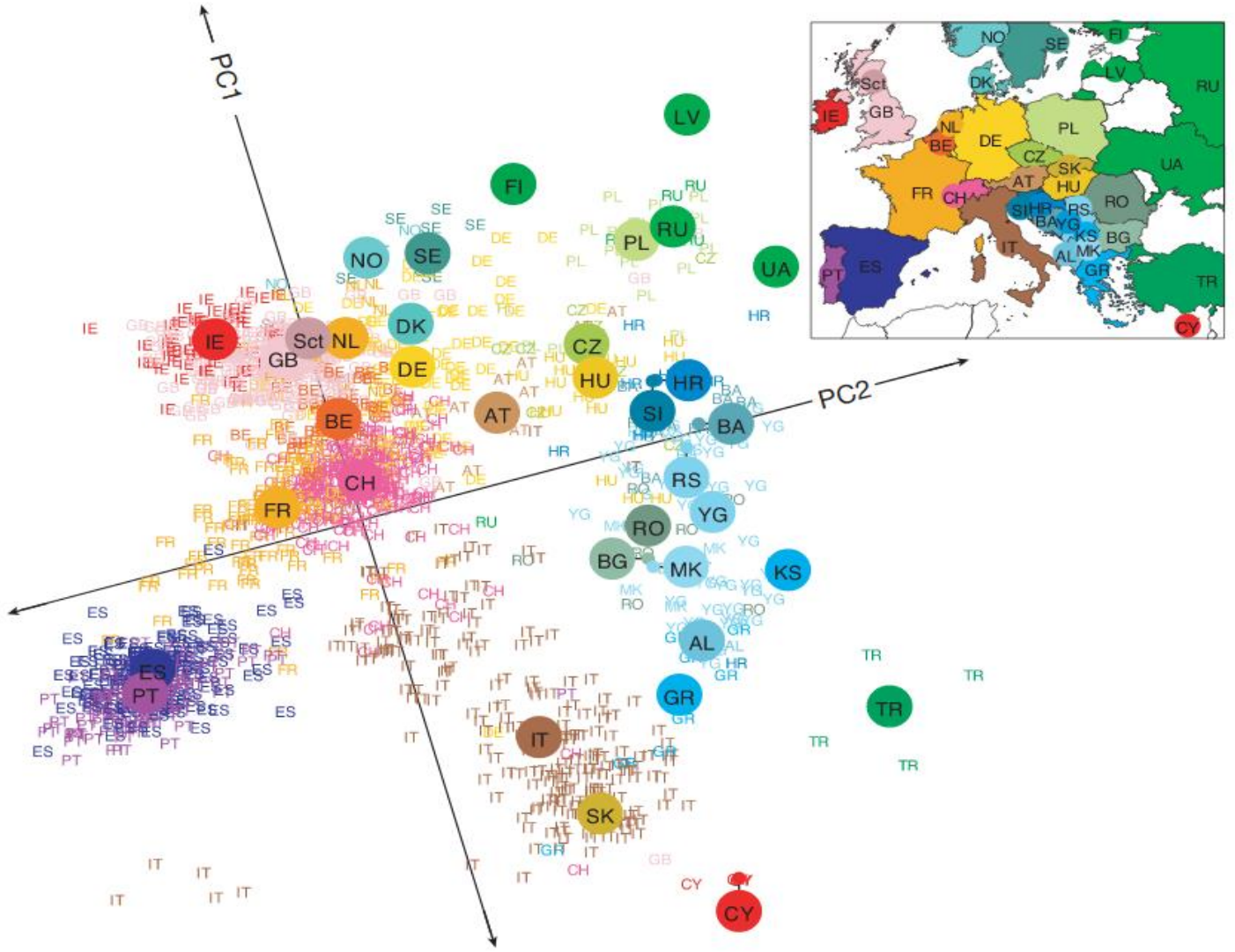


Cambel, C.D. *et al* Nature Genetics (2005)

Контроль популяционной подразделенности с помощью PCA



Контроль популяционной подразделенности с помощью PCA



Novembre, J. et al., Nature (2008)

Контроль популяционной подразделенности с помощью PCA

- Для моделей регрессии – включить главные компоненты как ковариаты:

$$y_i = \mu + \beta x_i + \underbrace{\alpha_1 PC_{i1} + \alpha_2 PC_{i2} + \cdots}_{\text{гл.компоненты}} + \epsilon_i$$

- Для других моделей :
  - Выполнить регрессию значений генотипического маркера $X \sim PC_1, \ldots, PC_L$, вычислить остатки регрессии $X_{adj,L}$
  - Выполнить регрессию значений фенотипов $Y \sim PC_1, \ldots, PC_L$, вычислить остатки регрессии $Y_{adj,L}$
  - Выполнить тест ПГАА на скорректированных значениях $X_{adj,L}$ и $Y_{adj,L}$

EIGENSOFT/EIGENSTRAT – анализ с помощью PCA (Price, A.L. et al. Nat Genet)

Линейная смешанная модель

- Учитывает популяционную структуру и родство:

$$\boldsymbol{y} = \boldsymbol{x}\beta + \mathbf{C}\boldsymbol{\alpha} + \boldsymbol{u} + \boldsymbol{\epsilon}$$

$\boldsymbol{y}$ – вектор фенотипов

$\boldsymbol{x}$ – вектор значений маркера, β – эффект маркера

$\mathbf{C}$ – матрица дополнительных факторов (ковариат), $\boldsymbol{\alpha}$ – эффекты ковариат

$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{R})$ – вектор случайных эффектов, $\mathbf{R}$ – матрица родства (kinship matrix)

- Эквивалентная запись:

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{x}\beta + \mathbf{C}\boldsymbol{\alpha}, \sigma_g^2 \mathbf{R} + \sigma_\epsilon^2 \mathbf{I}\right)$$

отражает ковариационную структуру значений фенотипа

Линейная смешанная модель

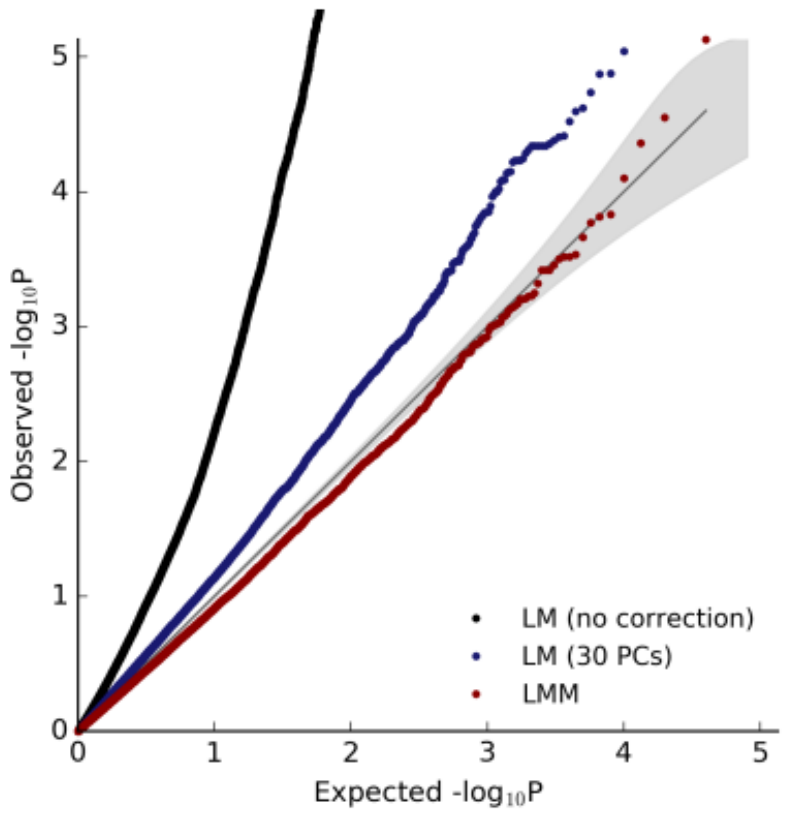
- Позволяет оценить пропорцию дисперсии фенотипа, объясняемую генетическими причинами при $\beta = 0$ (PVE – proportion of variance explained): $PVE = \sigma_g^2 / (\sigma_g^2 + \sigma_\epsilon^2)$

Препарат	PVE,%	Ошибка,%	Препарат	PVE,%	Ошибка,%
Препараты первого ряда			Препараты второго ряда		
Изониазид	99.997	0.021	Циклосерин	75.716	12.386
Рифампицин	99.997	0.021	Капреомицин	73.903	11.048
Пиразинамид	99.997	0.049	Амикацин	69.831	11.925
Стрептомицин	99.997	0.036	Офлоксацин	58.682	12.922
Этамбутол	97.119	1.695	Этионамид	45.680	24.906
			Кислота пара-аминосалициловая	29.998	17.010

GEMMA – анализ с помощью линейной смешанной модели (Zhou , X. and Stephens, M. Nat Genet, 2012)

Анализ результатов: квантиль-квартиль диаграммы

- Визуализировать распределение всех p-значений из исследования
- При нулевой гипотезе p-значения распределены равномерно
- Большинство значений будет в нижней части графика
- Иногда также приводят $\lambda = \frac{med(\chi^2)}{0{,}456}$ - коэф. геномного контроля



PLINK – универсальный инструмент для выполнения ПГАА (Purcell, S. et al. Am J Hum Gen, 2007; Chang, C.C. et al. GigaScience, 4, 2015)

Проблема множественных сравнений

- Если $\alpha = 0{,}05$ и необходимо проверить 100 мутаций – получим 5 ложных срабатываний
- В большинстве исследований используют:
  - контроль по групповой вероятности ошибки (FWER – family-wise rate)
  - контроль по средней доле ложных отклонений среди отклоненных нулевых гипотез (FDR – false discovery rate)

$$\frac{E(FP)}{m} \leq FDR \leq FWER \leq E(FP)$$

- Поправки p-значений:
  - Бонферрони: $\alpha^* = \alpha/m$
  - Сидака: $\alpha^* = 1 - (1-\alpha)^{1/m}$
  - тест перестановок

Одномаркерные тесты ПГАА

- Количественные признаки – проверяем, соответствуют ли тренд в генотипах и тренд в фенотипах
  - Линейная регрессия
- Бинарные признаки – сопоставляем частоты аллеля в случаях и контролях
  - Хи-квадрат, точный тест Фишера, трендовый тест Корхана-Армитажа, логистическая регрессия
- Необходимо контролировать:
  - Дополнительные внешние факторы (ковариаты)
  - Популяционную структруру
  - В линейной смешанной модели генетические confounders контролируются автоматически с помощью матрицы родства
  - Много независимых тестов – применять поправки p-значений на множественные сравнения

Многомаркерные тесты ПГАА

- Могут строиться на разных моделях:
  - Регуляризованная линейная или логистическая регрессия (регуляризация по l_1 и l_2 нормам)
  - Линейная смешанная модель с множественными геномными эффектами
  - Байесовская многомерная регрессия, где эффекты геномных мутаций $\boldsymbol{\beta} \sim N\left(0, \mathbf{I}\sigma_\beta^2\right)$ – случайные векторы и т.п.
  - …
- Включать или не включать взаимодействующие переменные
- Часто $m >> n$ – трудности при оценивании

Регрессия с регуляризацией

- Улучшает свойства оценок

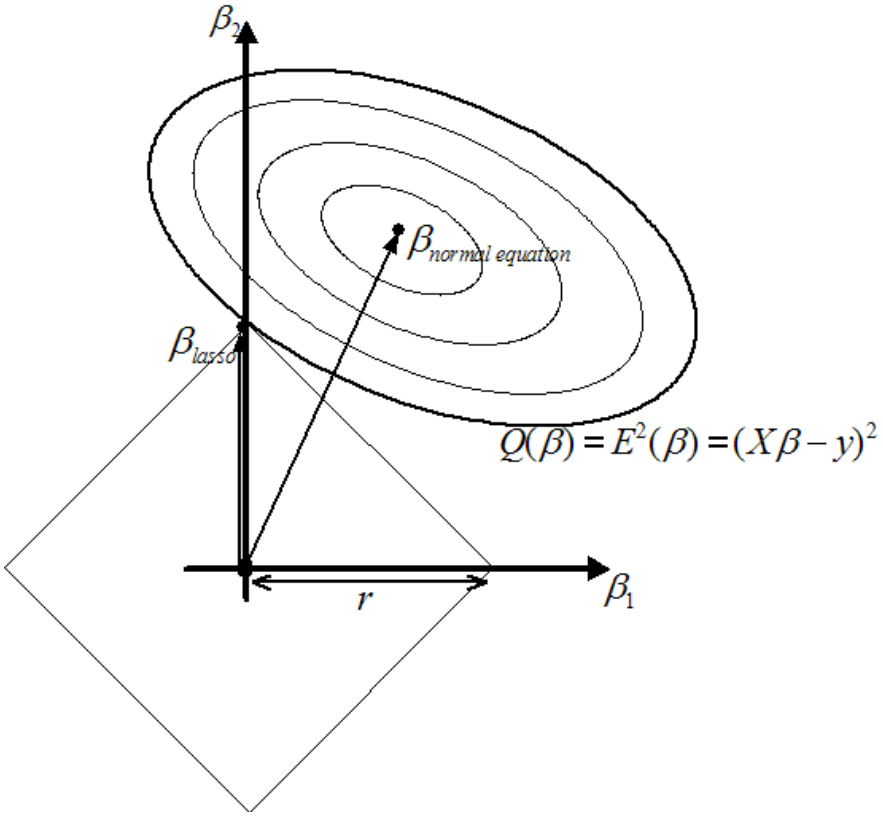
- Оптимизируется функционал (*elastic net* - смешанная регуляризация по $l1$ и $l2$ норме):

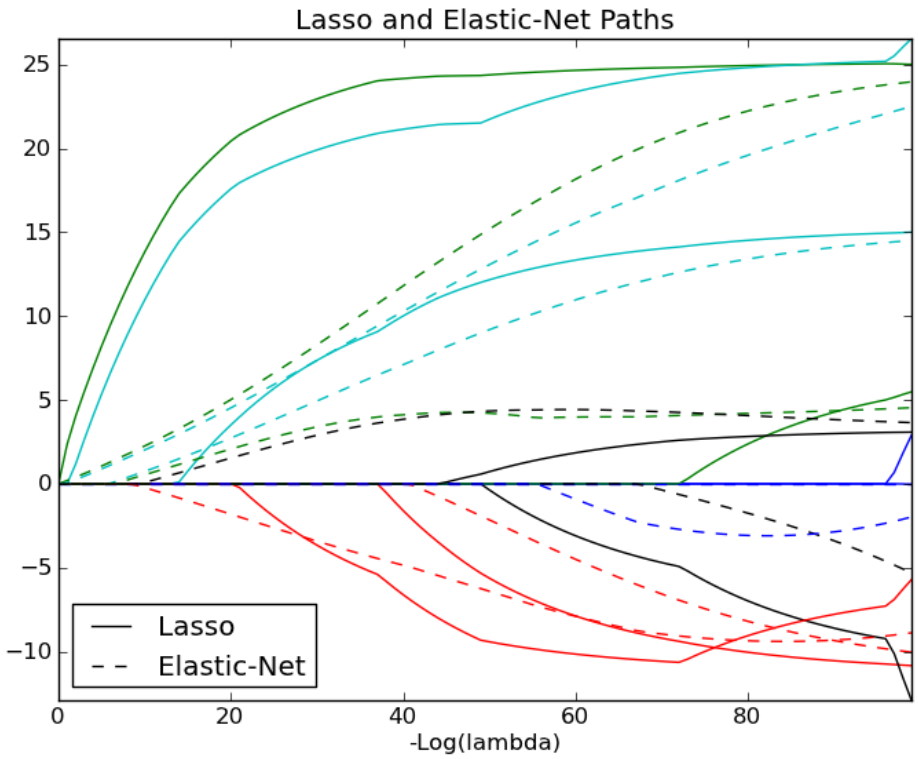
$$\sum_{i=1}^{n}\left[(1-y_i)\beta^T x_i + \ln(1+\exp(-\beta^T x_i))\right] + \lambda_1\|\beta\|_{l_1} + \lambda_2\|\beta\|_{l_2} \to \min_{\beta}$$

- Уменьшает количество ненулевых признаков, уменьшает дисперсии их оценок и “сужает” размер $|\beta_j|$

Регрессия с регуляризацией

- Пример регуляризации методом lasso и elastic-net





Найденная ассоциация – это не причинно-следственная связь!

- Найденный маркер имеет функциональный эффект на фенотип:
  - несинонимичная мутация (меняет белок)
  - меняет кодон на стоп-кодон
  - влияет на регуляцию синтеза белка

ИЛИ

- Найденный маркер неравновесно сцеплен с функциональным
  - обычно в экспериментах на основе ДНК-микрочипов

SnpEff– анализ эффектов SNPs (Cingolani, P. et al. PMID: 22728672, 2012)

Резюме: три простых шага для ПГАА

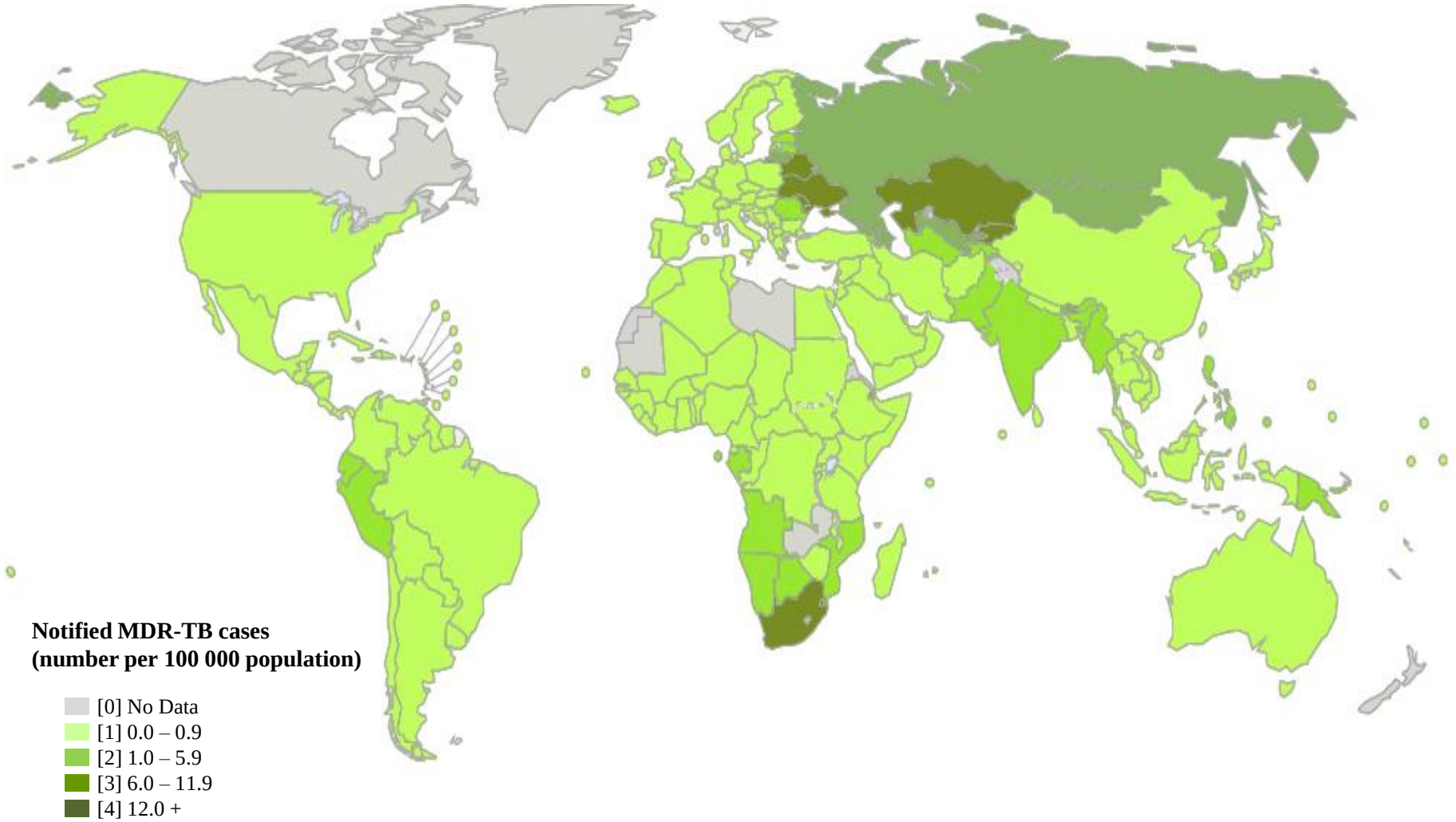
1. Определиться с дизайном исследования

2. Получить данные и выполнить контроль качества

3. Выбрать модель, оценить параметры и критически отнестись к найденным ассоциациям:
   - возможны ложные зависимости
   - воспроизводимость результатов, вторичный анализ

Анализ лекарственной устойчивости M. tuberculosis

Диагносцированные случаи МЛУ ТБ

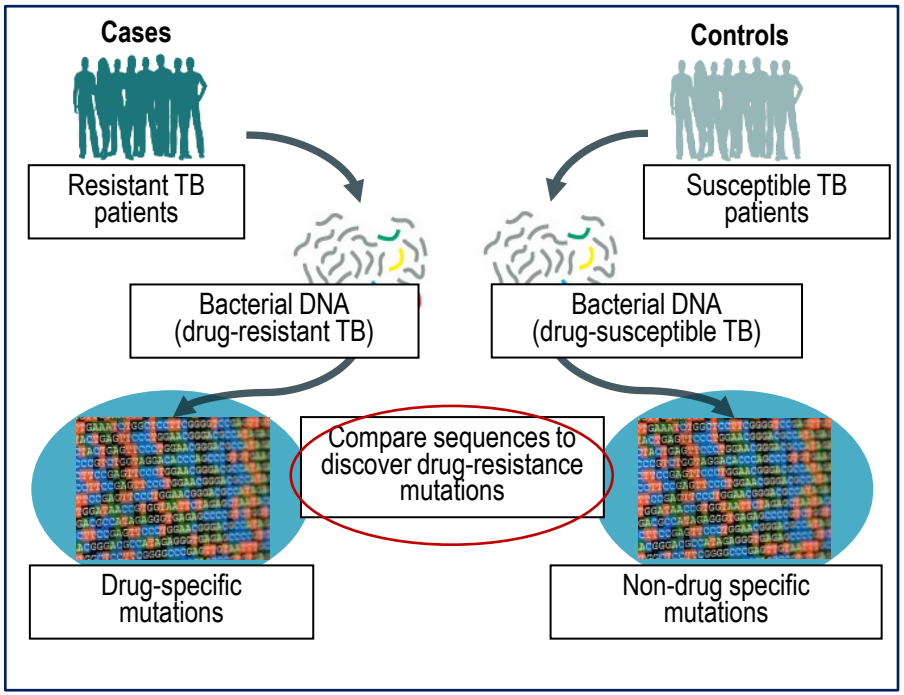


TBDreamDB “не объясняла” всю резистентность

	RR*	SR	RS	SS	% resistance ‘explained’?	% susceptibility ‘explained’
rifampicin 400ugml	106	3	0	29	97.25%	100%
isoniazid 02ugml	105	3	5	22	97.22%	81.48%
streptomycin 40ugml	104	10	0	24	91.23%	100%
ethambutol 20ugml	94	11	3	30	89.52%	90.91%
ofloxacin 20ugml	60	10	6	61	85.71%	91.04%
pyrazinamide 100ugml	14	13	0	7	51.85%	100%
ethionamide 400ugml	0	36	0	104	0%	100%
kanamycin 300ugml	0	83	0	57	0%	100%
amikacin 400ugml	0	66	0	74	0%	100%
cycloserine 400ugml	0	49	0	83	0%	100%
capreomycin 400ugml	0	70	0	64	0%	100%

*RR = resistant genotype/resistant phenotype, RS = resistant genotype/sensitive phenotype, SR = sensitive genotype/resistant phenotype, SS = sensitive genotype/sensitive phenotype

Дизайн исследования ПГАА



Raw data upload from NGS sequencer or NCBI database

Converting short paired reads to FASTQ-format

Data preprocessing to improve reads quality, elongation of paired-end reads

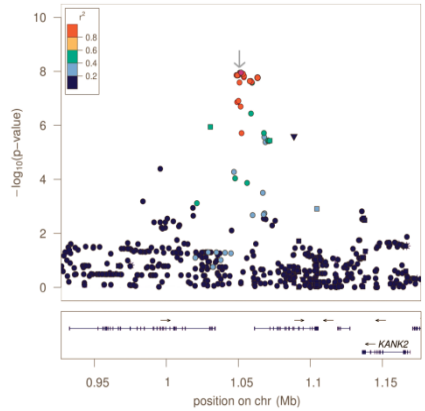
Genome assembly by mapping to the reference strain

Variants calling, detecting genome variations

Analysis of mutations, selecting drug-resistance genes

Mutation profiles of the strains and lists of drug-resistance mutations

	Mutation (SNP)		Mutation (SNP)		Mutation (SNP)	
	Gene 1		*Gene 2*		*Gene 3*	
ReferenceH37Rv	AACA**C**GCCA....		TTCG**G**GGTC....		AGTC**G**ACCG....	CGG
Sequence1	AACACGCCA....		TTCG**A**GGTC....		AGTC**A**ACCG....	.AA
Sequence2	AACA**T**GCCA....		TTCGGGGTC....		AGTC**A**ACCG....	T.A
Sequence3	AACACGCCA....		TTCGGGGTC....		AGTCGACCG....	...



Данные полногеномного секвенирования

NCBI BioProject	Год	# образцов	Страна
PRJNA200335	2013	144	Belarus
PRJNA421446	2017	96	Azerbaijan
PRJNA384815	2017	41	Romania
PRJNA436997	2018	*NA*	Moldova
PRJNA429460	2018	82	Belarus
PRJNA318002	2016-2018	568	Azerbaijan, Georgia, Moldova, Romania
Total		***931***	

Данные полногеномного секвенирования

Секвенирование + Метаданные на портале

Batch 1 (2013):
134 MTB strains
from Belarus

Paired-end frags and
jumping libraries,
4 SRRs per genome

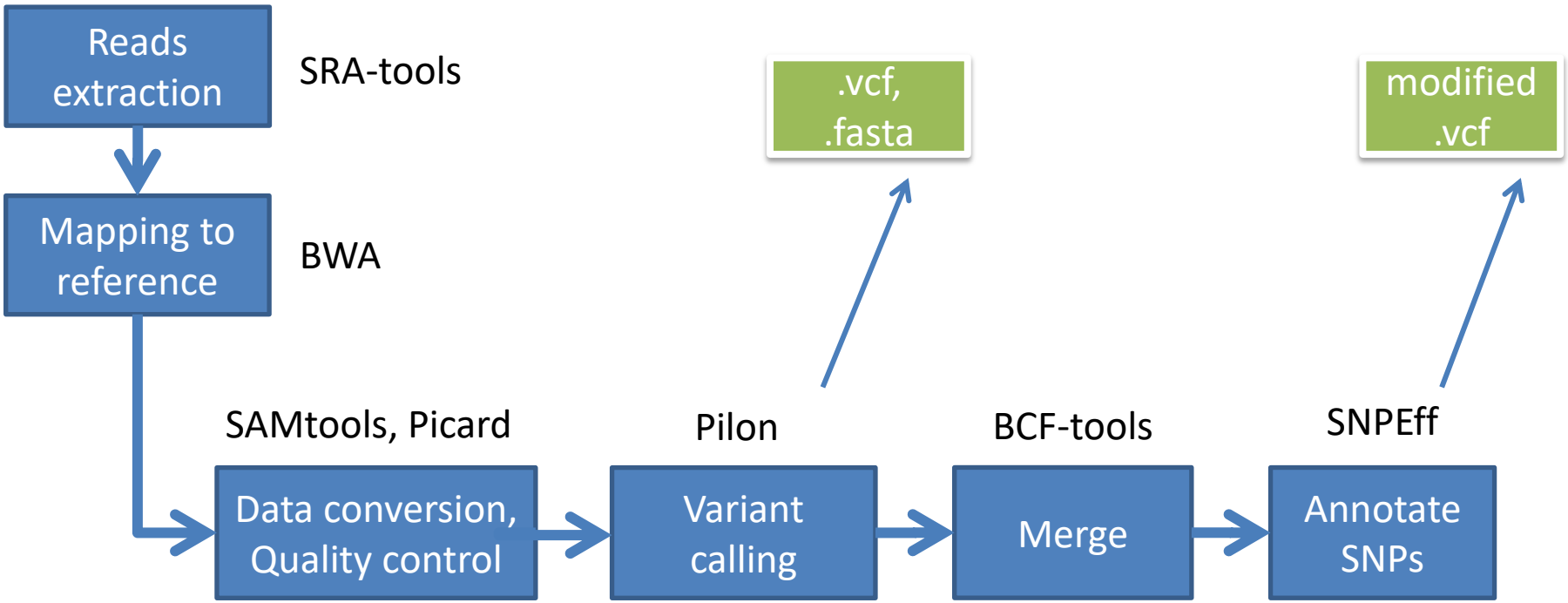
Batch 2 (2016-2018):
605 MTB strains from
AZ, BY, GE, MO, RO

Paired-end frags libraries,
1 SRR per genome

739 геномов с клиническими метаданными (сентябрь, 2018)

Пайплайн для аннотации SNPs

- Процедура картирования прочтений на референс и аннотации мутаций (SNPs)



Пример отчета программы SNPEff

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	3,982	0.158%
LOW	164,924	6.554%
MODERATE	68,609	2.726%
MODIFIER	2,278,965	90.562%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	66,933	28.796%
NONSENSE	649	0.279%
SILENT	164,854	70.924%

Missense / Silent ratio: 0.406

Number of effects by type and region

Type

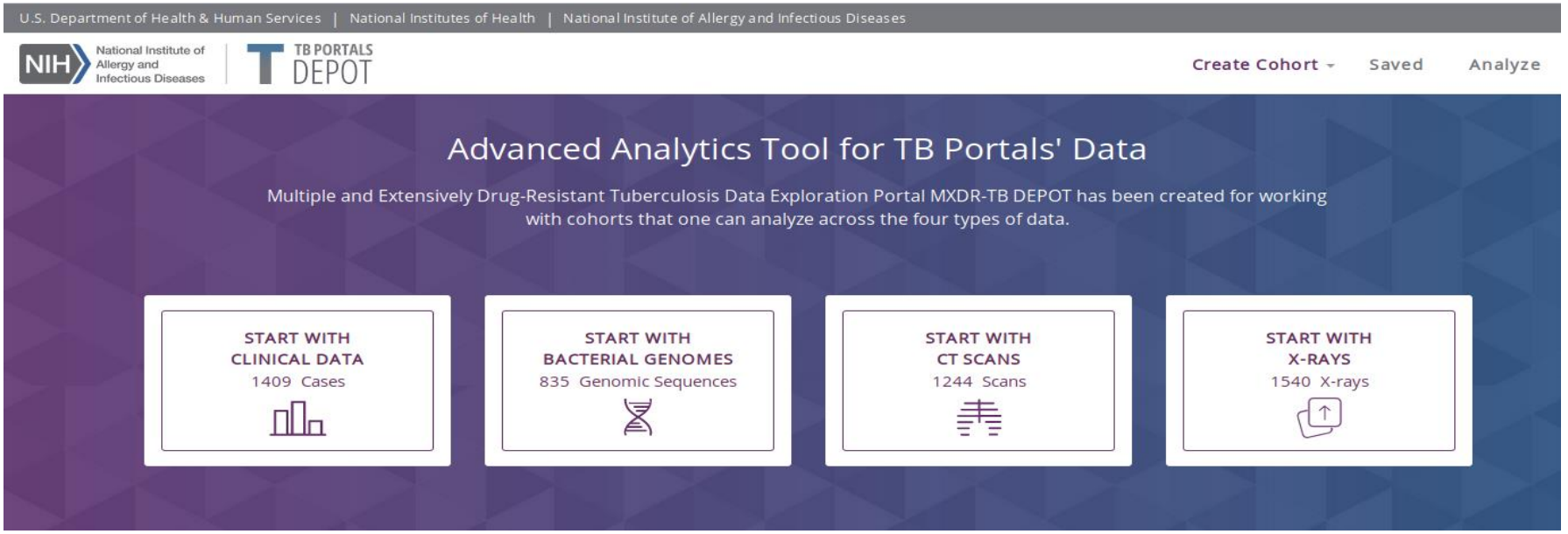
Type (alphabetical order)	Count	Percent
bidirectional_gene_fusion	22	0.001%
conservative_inframe_deletion	459	0.018%
conservative_inframe_insertion	214	0.008%
disruptive_inframe_deletion	736	0.029%
disruptive_inframe_insertion	224	0.009%
downstream_gene_variant	1,121,328	44.525%
feature_ablation	89	0.004%
frameshift_variant	2,614	0.104%
gene_fusion	79	0.003%
initiator_codon_variant	14	0.001%
intergenic_region	6,672	0.265%
intron_variant	78	0.003%
missense_variant	67,342	2.674%
non_canonical_start_codon	3	0%
non_coding_transcript_exon_variant	661	0.026%
splice_region_variant	606	0.024%
start_lost	388	0.015%
stop_gained	806	0.032%
stop_lost	290	0.012%
stop_retained_variant	121	0.005%
synonymous_variant	165,112	6.556%
transcript_ablation	246	0.01%
upstream_gene_variant	1,150,322	45.676%

Region

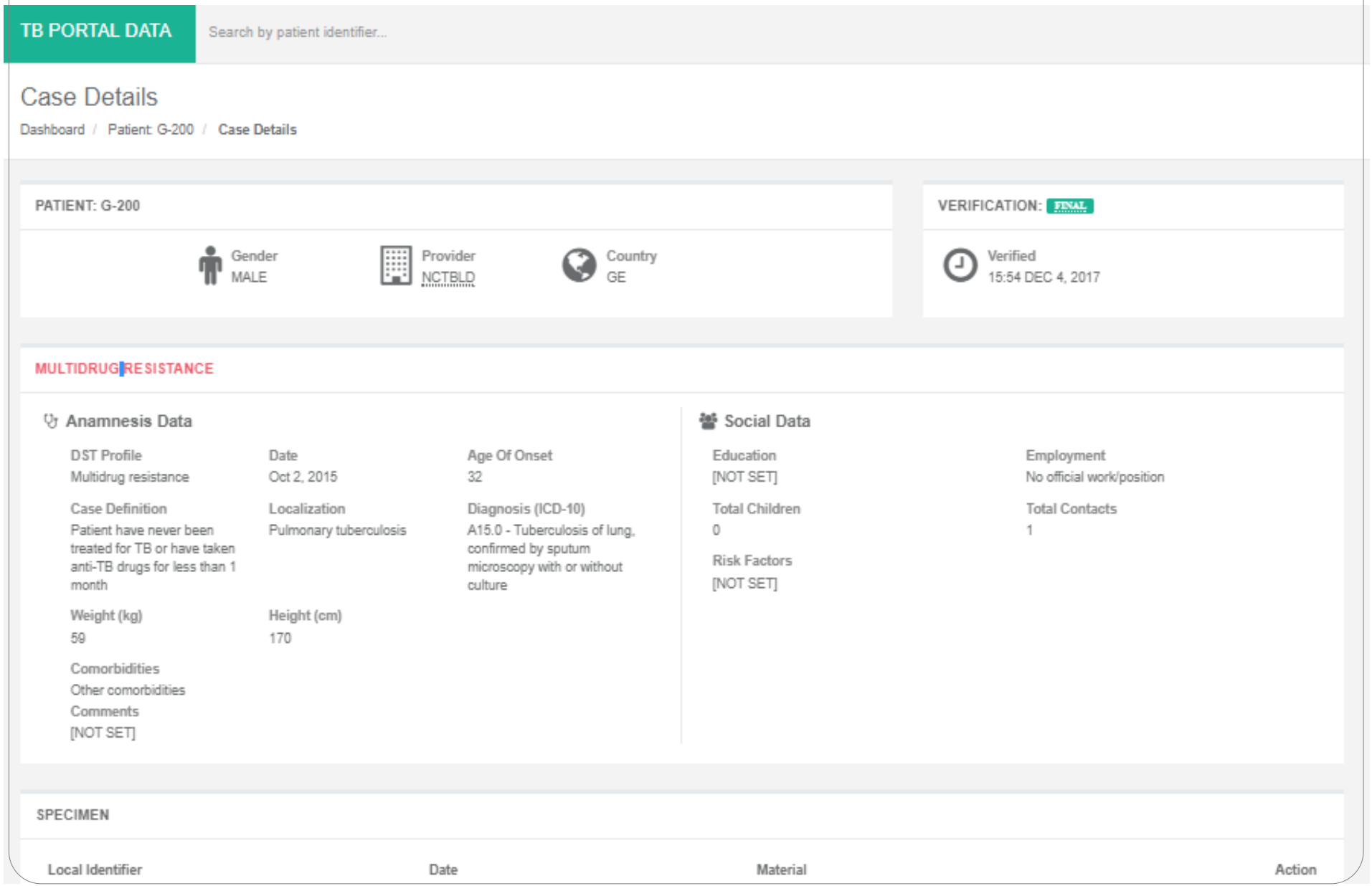
Type (alphabetical order)	Count	Percent
DOWNSTREAM	1,121,328	44.559%
EXON	237,657	9.444%
GENE	190	0.008%
INTERGENIC	6,672	0.265%
SPLICE_SITE_REGION	65	0.003%
TRANSCRIPT	246	0.01%
UPSTREAM	1,150,322	45.712%

Доступные метаданные

- Портал: https://depot.tbportals.niaid.nih.gov/
  - 998 случаем из 5 стран
    - **Анамнез, социальные данные**
    - **Схема лечения**
    - **Результаты тестов на чувствительность к препаратам**
    - **Изображения (КТ, рентген)**
    - **Геномы**
  - Внимание к МЛУ/ШЛУ образцам



Метаданные: пример записи на портале



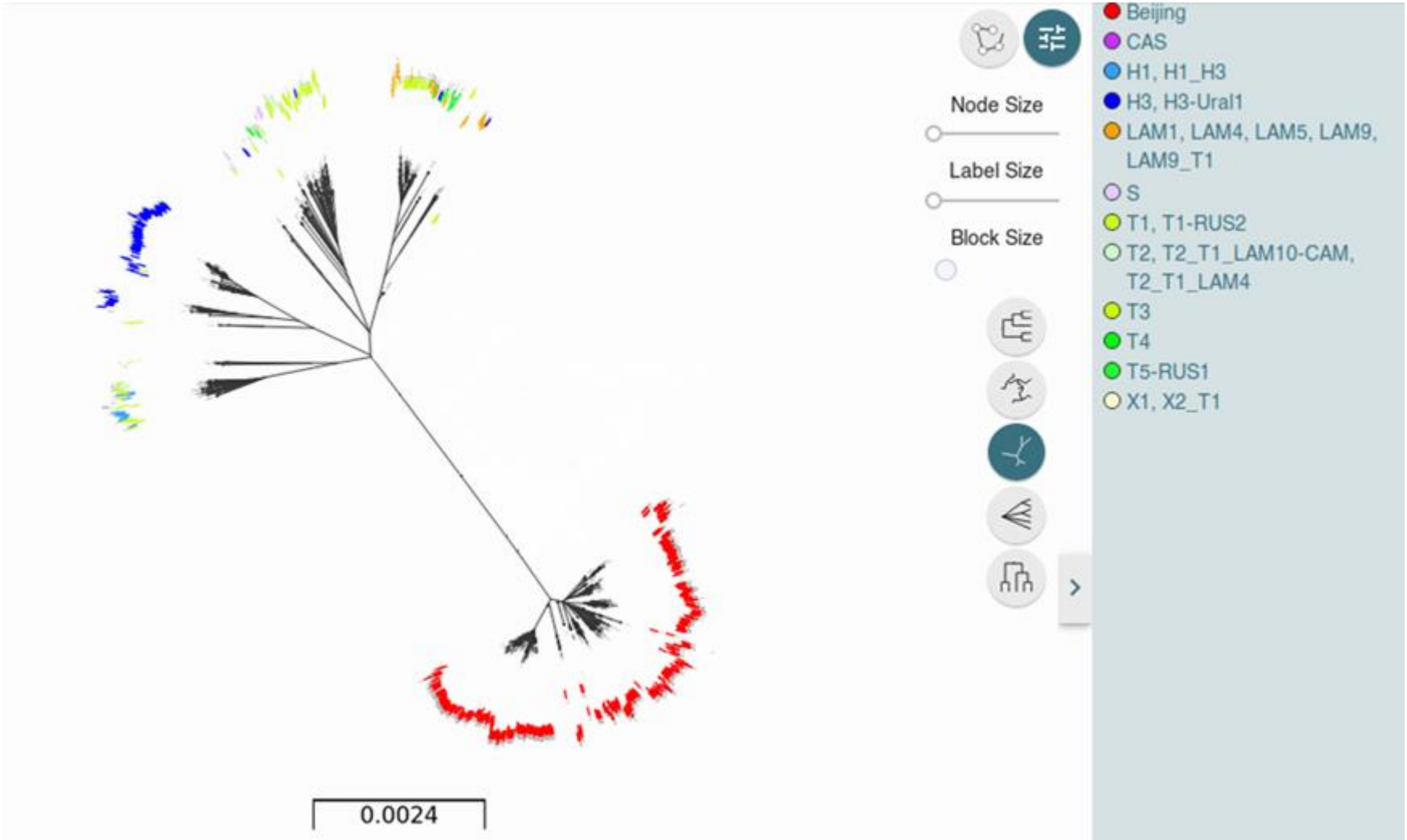
Цели экспериментов ПГАА

- Подтвердить эффект известных мутаций резистентности на нашем наборе данных (или выявить новые специфичные для страны мутации?)
- Поиск компенсаторных мутаций (‘fitness-cost’ , …)
- Исследовать связь с другими факторами внешней среды:
  - treatment outcomes (cures/failed patients)
  - infected and relapsed cases
  - TB lineage/sub-lineage
  - cavity size (in connection with virulence factors)
  - outbreaks (transmission analysis)
  - geography (country, region)

Набор данных для исследования (2018)

Drug	Susceptible	Resistant	Total
ISON (Isoniazid)	114	496	610
RIF (Rifampicin)	116	494	610
STRE (Streptomycin)	113	476	589
ETHA, EMB (Ethambutol)	205	383	588
OFLO (Ofloxacin)	234	236	470
CAPR (Capreomycin)	333	172	505
AMIK (Amikacin)	252	102	354
KANA (Kanamycin)	147	178	325
PYRA (Pyrazinamide)	34	31	65
LEVO (Levofloxacin)	79	31	110
MOXI (Moxifloxacin)	30	-	30
PARA (Para aminosalicylic acid)	387	40	427
PTH (Prothionamide)	218	114	332
CYCL (Cycloserine)	164	48	212

Филогенетический анализ (2018)



Методы машинного обучения для ПГАА

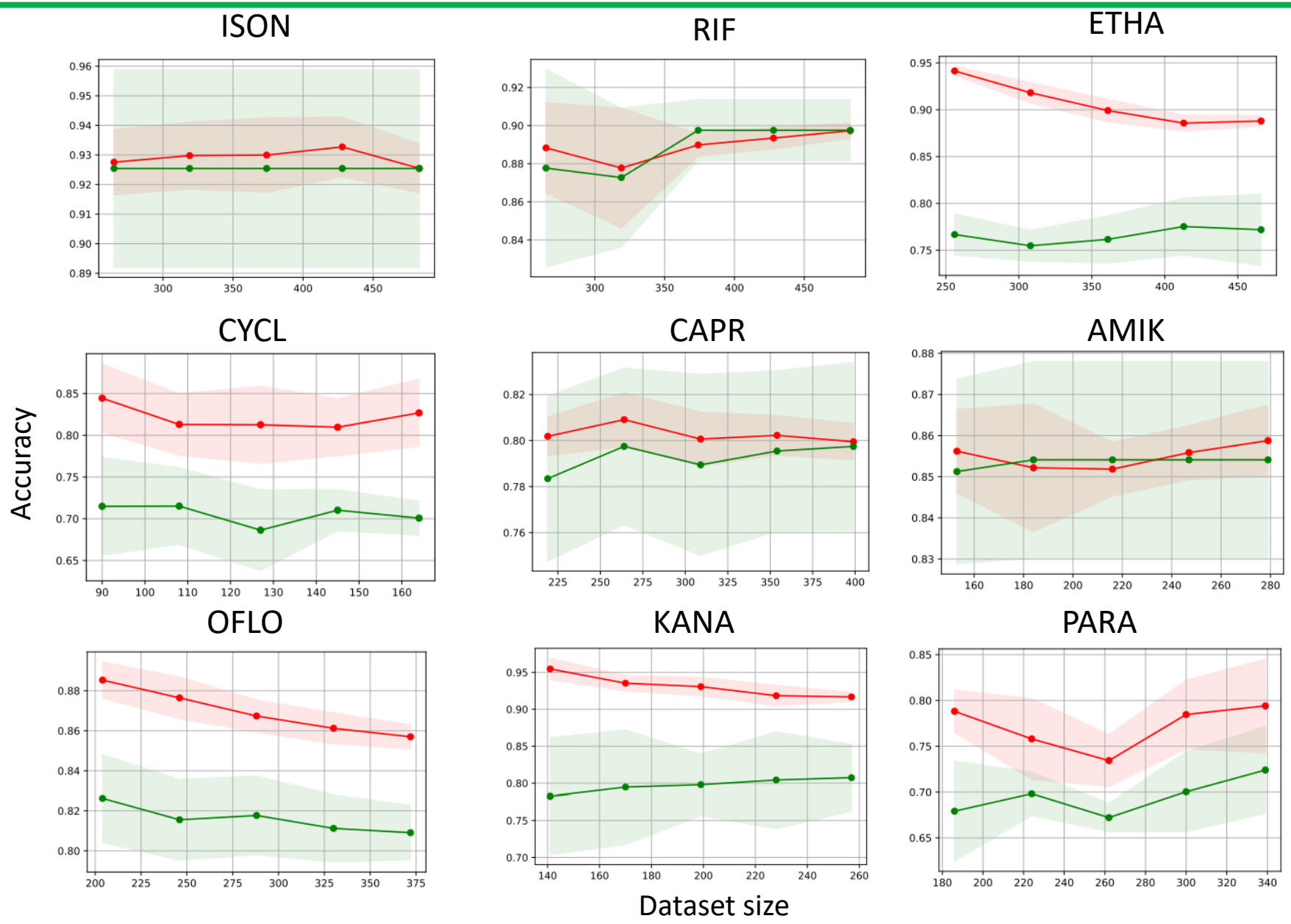
- Помимо классических методов ПГАА (ЛММ, модели регрессии) проводились эксперименты со следующими методами машинного обучения:
  - RF – случайный лес
  - GB – градиентный бустинг
  - SVM – машина опорных векторов

Результаты экспериментов ПГАА (2018)

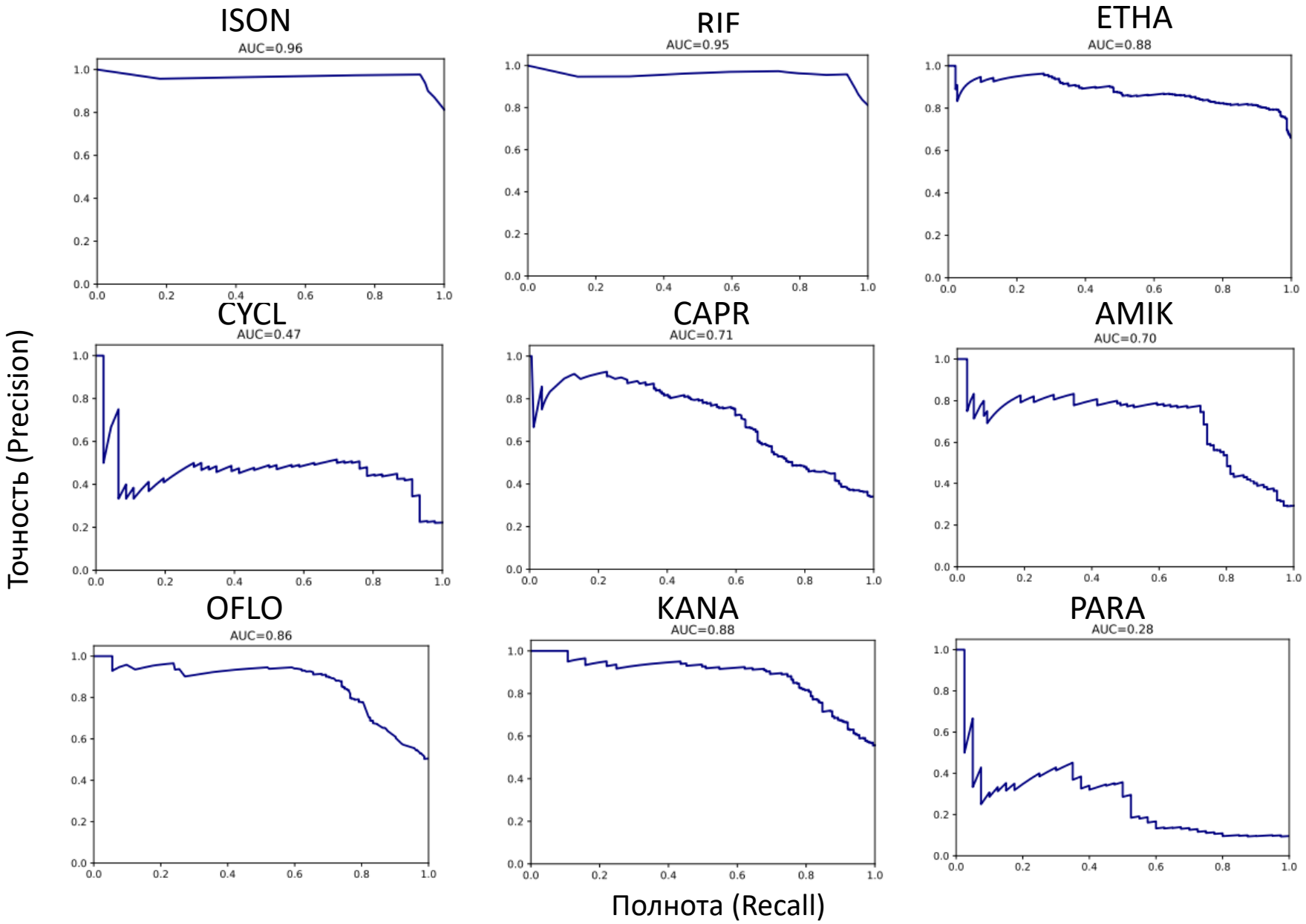
- Градиентный бустинг (GB) – лучший среди LR, RF, GB and SVM

Drug	R	S	Tree Depth*	Precision	Recall	F1	Accuracy	Kappa
ISON	492	113	1	0.977	0.931	0.953	0.926	0.773
RIF	491	114	1	0.957	0.914	0.935	0.898	0.688
STRE	473	111	1	0.962	0.922	0.942	0.908	0.719
ETHA	380	203	3	0.821	0.808	0.814	0.76	0.475
CYCL	46	161	4	0.442	0.826	0.576	0.729	0.403
CAPR	169	330	2	0.756	0.586	0.66	0.796	0.517
AMIK	101	249	1	0.76	0.723	0.741	0.854	0.64
OFLO	235	231	2	0.897	0.702	0.788	0.809	0.619
KANA	177	145	3	0.896	0.729	0.804	0.804	0.613
PTH	108	214	3	0.448	0.685	0.542	0.612	0.23
PARA	40	384	1	0.19	0.55	0.282	0.736	0.165

Кривые обучения градиентного бустинга



Точность/полнота и значения ROC AUC градиентного бустинга



Результаты ПГАА и аннотации

- Наиболее значимые SNPs, полученные логистической регрессией и градиантным бустингом по **721** геномам

Drug	SNP (LogReg)	Weight (LogReg)	Annotation	SNP (GradBoost)	Weight (GradBoost)	Annotation
OFLO	7582	1	GyrA	1633595	1	PE_PGRS27
	7570	0.865	GyrA	7582	0.829	GyrA
	7581	0.486	GyrA	7581	0.225	GyrA
	7582	0.426	GyrA			
AMIK	1473246	1	rrs	1473246	1	rrs
CAPR	1473246	1	rrs	1473246	1	rrs
KANA	1473246	1	rrs	1473246	1	rrs
	3841083	0.388	alr			
PARA	2338194	1	Rv2081c	927110	1	PE_PGRS13
	4247429	-0.946	embB	2986827	0.729	Rv2670c
	927110	-0.937	PE_PGRS13			
	7582	0.924	GyrA			

Результаты ПГАА и аннотации

- Наиболее значимые SNPs, полученные логистической регрессией и градиентным бустингом по **721** геномам

Drug	SNP (LogReg)	Weight (LogReg)	Annotation	SNP (GradBoost)	Weight (GradBoost)	Annotation
ISON	2155168	1	katG	2155168	1	katG
RIF	2155168	1	katG	2155168	1	katG
	761155	0.683	rpoB	761155	0.026	rpoB
ETHA	2155168	1	katG	4387729	1	Rv3902c
	1473246	0.81	rrs	2155168	0.948	katG
	761155	0.553	rpoB	3894784	0.881	PPE60
	4247429	0.422	embB	4247429	0.861	embB

Результаты ПГАА и аннотации

- Сопоставление результатов нескольких моделей, включая классические методы и машинное обучение

Drug	SNP position	Significant in models / Total models	Annotation
Ofloxacin	7570	All	DNA gyrase subunit A GyrA
	7582	3 / 5	DNA gyrase subunit A GyrA
	7579	3 / 5	DNA gyrase subunit A GyrB
Levofloxacin	7570	All	DNA gyrase subunit A GyrA
	7582	3 / 5	DNA gyrase subunit A GyrA
Amikacin	1473252	All	rrs
Capreomycin	1473252	All	rrs
Kanamycin	1473252	All	rrs
	2715379	4 / 5	intergenic
Para-aminosalicyclic acid	2747481	3 / 5	Dihydrofolate synthase

Заключение

Резюме

- Методы ПГАА успешно идентифицируют сложные признаки во многих исследованиях
  - При этом многие локусы являются некодирующими или с неизвестной функцией гена
  - Чем больше выборка – тем больше новых вариантов с малым эффектом демонстрируют значимость
- В результатах почти всегда будут статистические артефакты!
  - Чувствительность к популяционной стратификации
- Разные модели не всегда дают одинаковые результаты
- Значительная часть работы должна быть сделана еще на этапе подготовки и контроля качества данных

Q&A