Анализ данных NGS в медицинской генетике

Определение, аннотация и интерпретация генетических вариантов

Юрий Барбитов

Ресурсный центр «Биобанк» Научного Парка СПбГУ
Исследовательское подразделение Института биоинформатики

1 августа 2017 г.

Летняя школа по биоинформатике 2017

Наследственные болезни человека

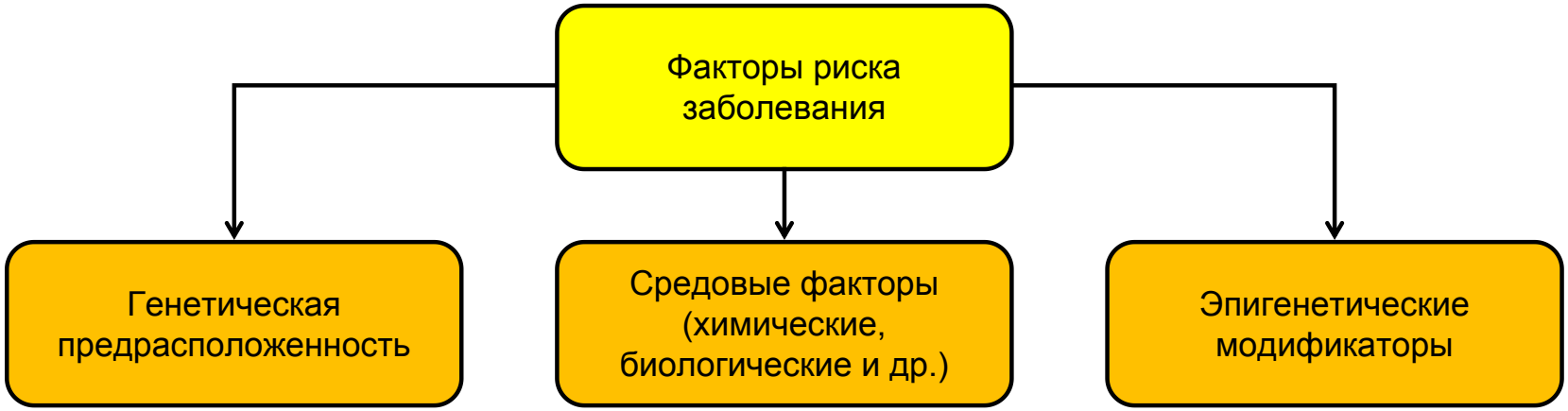


Таблица: Классификация наследственных заболеваний человека

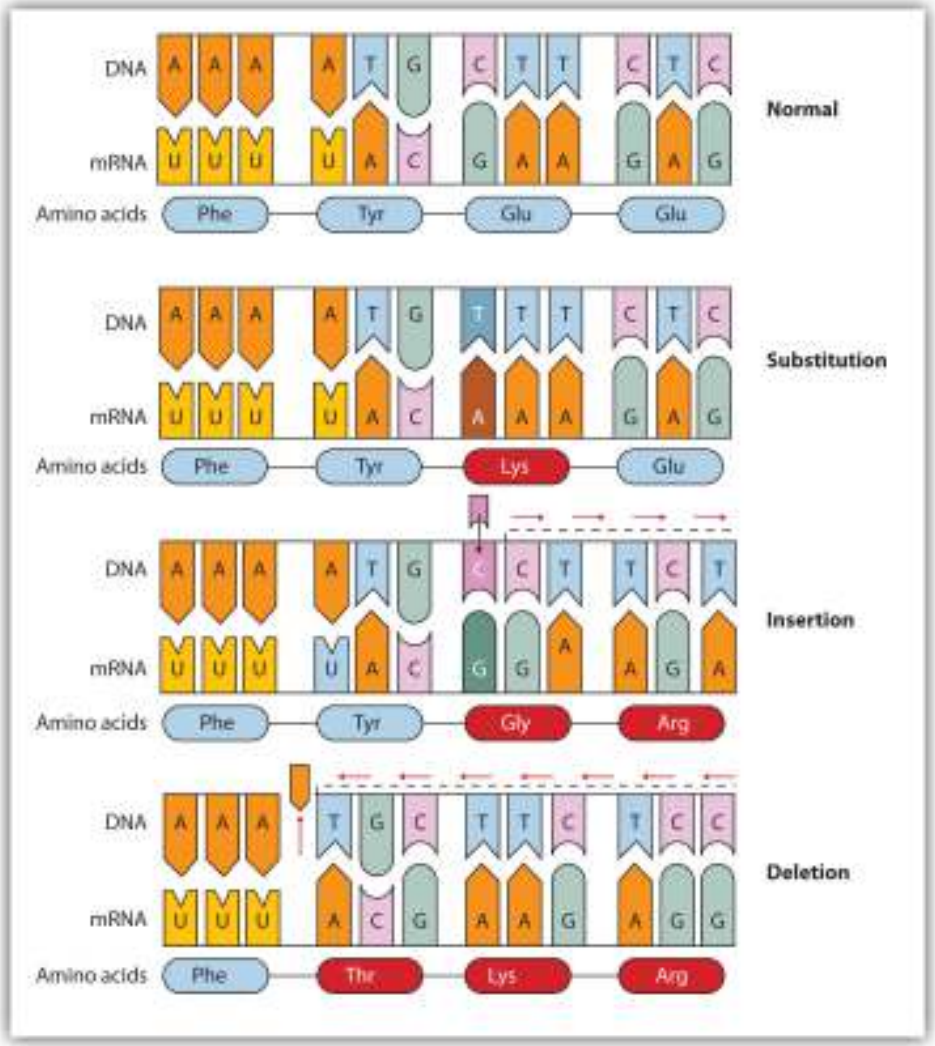
Тип	Гены	Наследуемость	Роль среды
Моногенные	1	Очень высокая	Низкая
Олигогенные	2-5	Высокая	Средняя
Мультифакторные	>5	Средняя	Высокая

Генетические вариации

В геноме человека ≈ 100,000,000 вариативных позиций (≈ 3% от длины, данные из dbSNP).

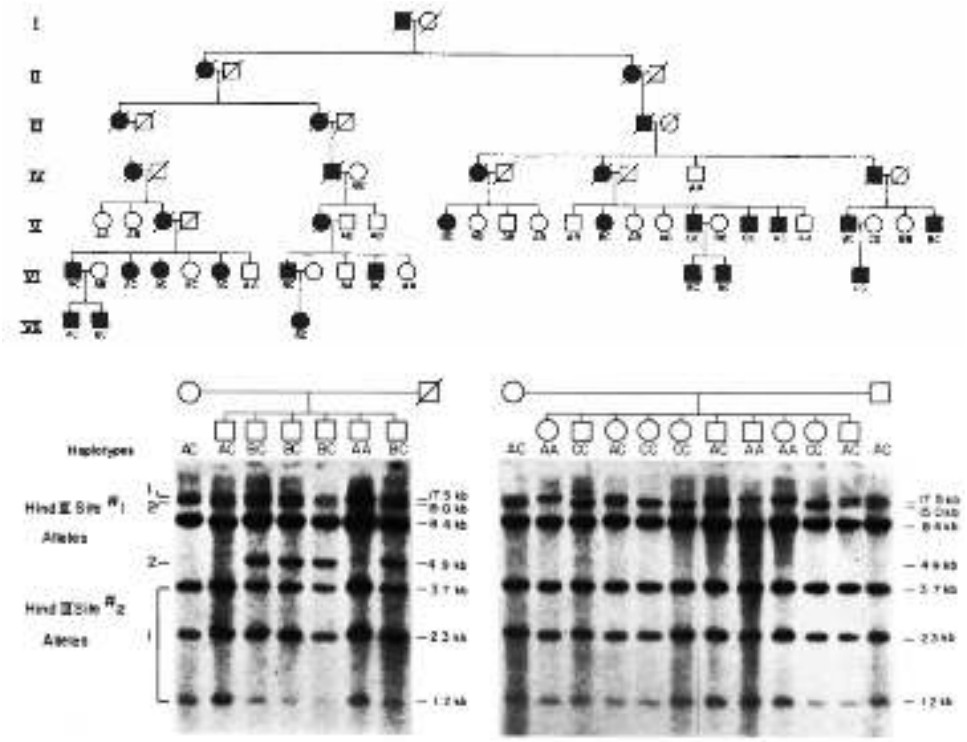
Основные типы вариаций:

- Замены нуклеотидов (SNV)
- Вставки и выпадения (indel)
- Вариации числа копий (CNV)
- Хромосомные перестройки



Картирование наследственных болезней

- Первое картирование - 1983 год
- Болезнь Хантингтона - АД, с поздней манифестацией
- Требуются большие родословные
- Сцепление с маркером

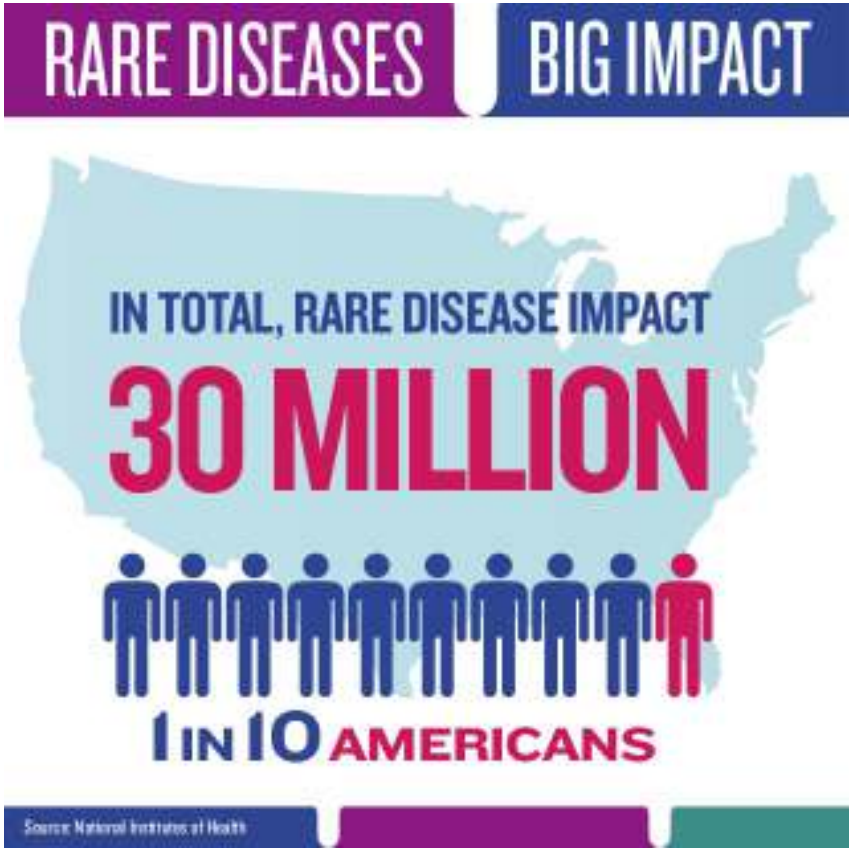


Gusella *et al.*, 1983

Наследственные болезни человека

Количество фенотипов, описанных в базе OMIM:

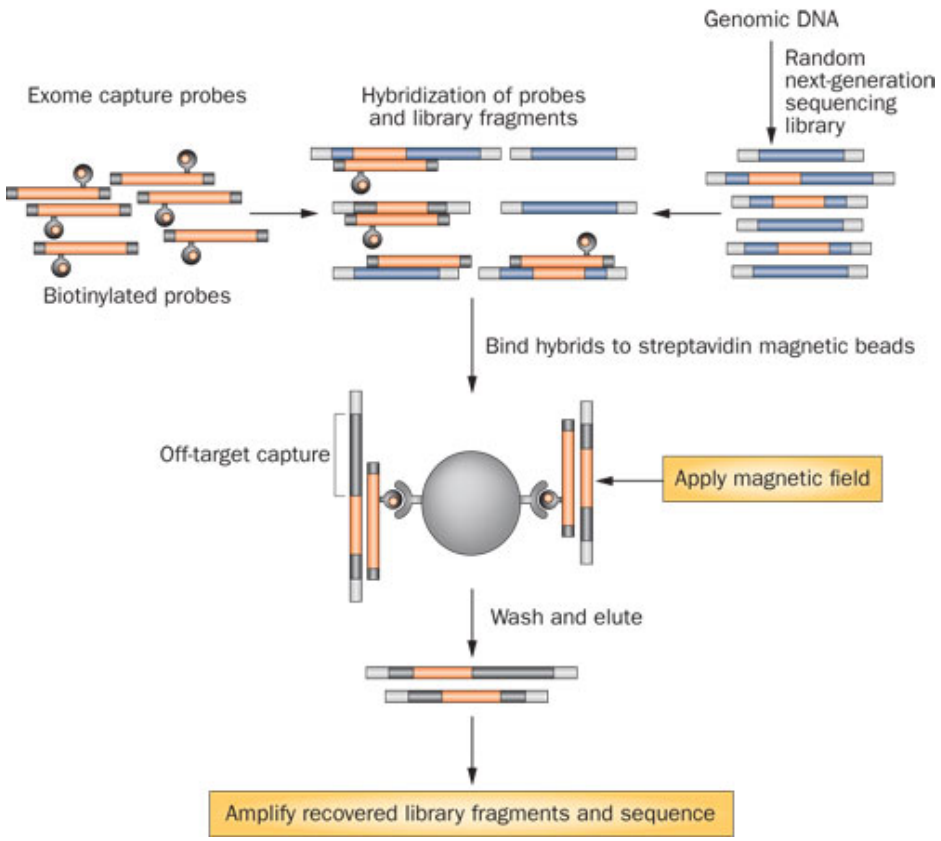
- Фенотипы с известной молекулярной природой: 5,035
- Фенотипы без известной природы: 3,385
- Предрасположенности: 702
- Моногенные болезни: 4,991



Подходы к анализу генома

Методы NGS в клинике:

- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)
- Таргетные панели, в т.ч. «клинический экзом» (*e.g.*, TruSight One Sequencing Panel)

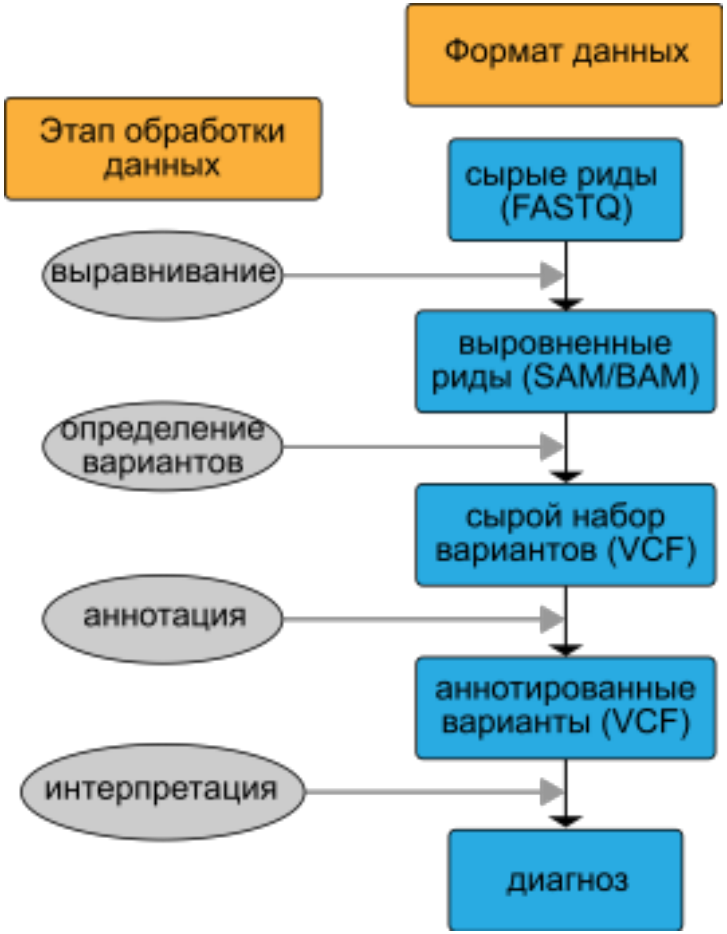


Анализ данных

Вариант - отличие генома образца от референсного генома.

Таблица: Среднее количество вариантов у представителя европейской популяции

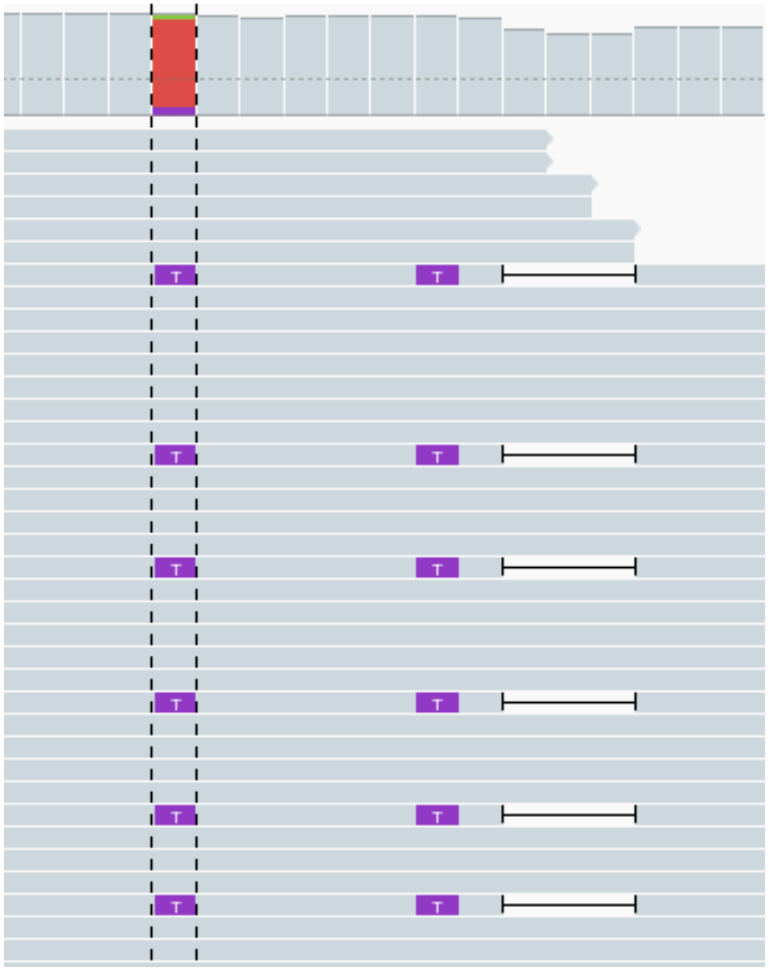
Эксперимент	N
WGS	3,000,000
WES	25,000
CES	7,500



А был ли вариант?

Проблемы данных NGS:

- Ошибки секвенирования
- Неоднородность покрытия генома
- Ошибки ПЦР и ПЦР-дубликаты
- Неточное выравнивание вокруг инделов
- Малая эффективность для выявления крупных перестроек



Мир альтернатив

Выравнивание

bwa
bowtie2
Novoalign
MAQ

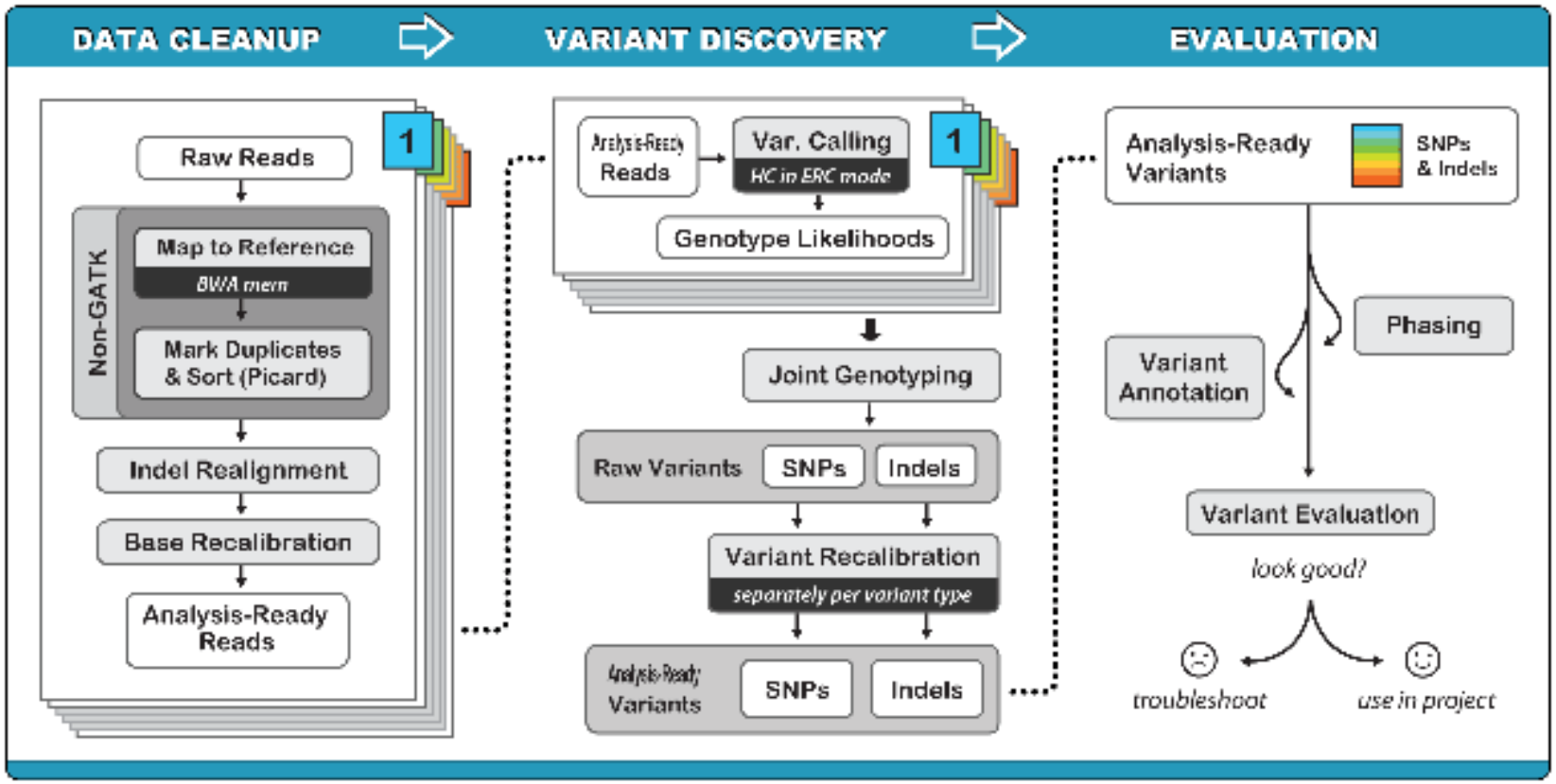
Variant calling

GATK
freebayes
SAMtools
TSVC
Varscan
Platypus
ScanIndel
fermi.kit

Аннотация

Ensembl VEP
ANNOVAR
SnpEff/SnpSift

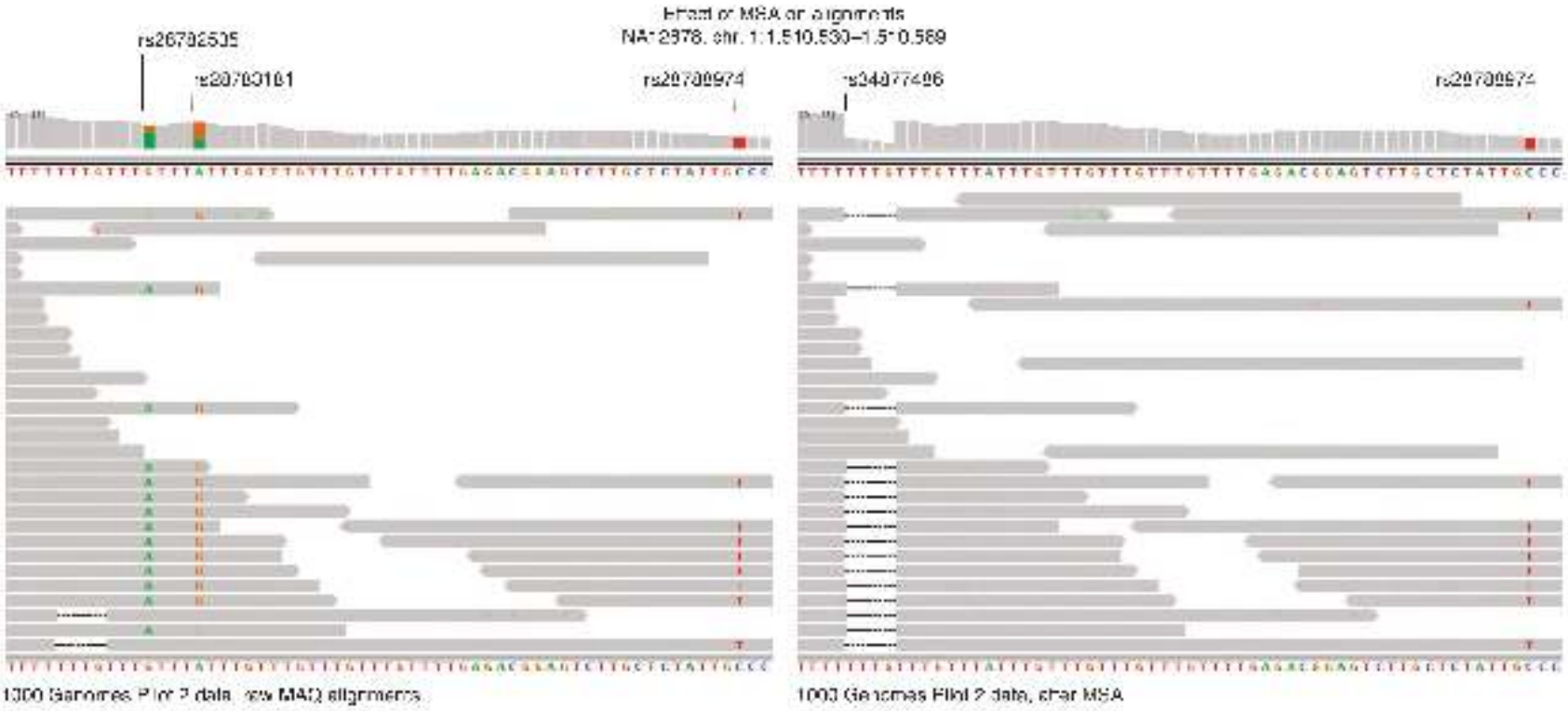
GATK Best Practices



Удаление дубликатов

Showing duplicate reads Hiding duplicate reads

Коррекция выравнивания в районах инделов



- Ложные SNP по краям от индела
- Неверные пропорции аллелей в месте индела

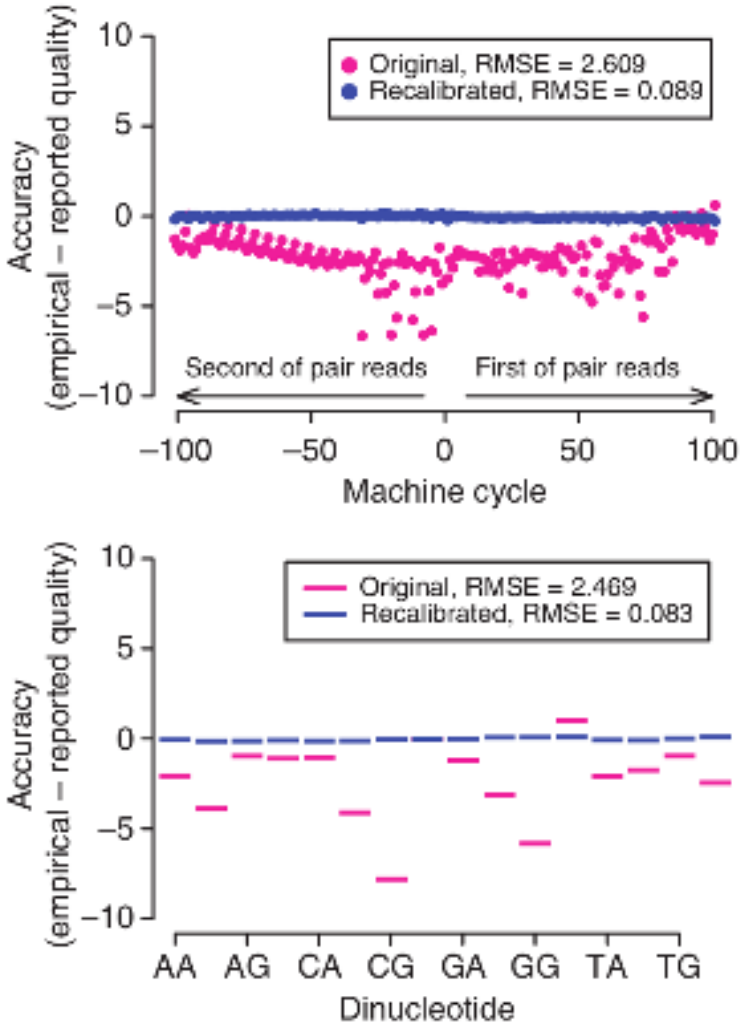
Перерасчет качества прочтений

Приборная оценка качества оснований не соответствует эмпирической!

Первоочередное влияние оказывают:

- Цикл секвенирования (вверху).
- Нуклеотидный контекст (внизу).

Решение: перекалибровка качества, исходя из известных паттернов ковариации.



Преимущества когортного анализа

Почему полезно использовать дополнительные образцы?

- Информация о частоте аллелей.
- Детекция систематических отклонений покрытия.
- Различение референсной гомозиготы и отсутствия покрытия.
- Легкое фазирование и импутация.

Основная проблема биоинформатического анализа

100%-ная точность, чувствительность и воспроизводимость анализа на практике недостижимы!

precisionFDA Consistency Challenge

- Данные WGS образца **NA12878** (консорциум Genome In A Bottle, GIAB).
- Два прибора в двух точках мира - Garvan, HLI.
- Воспроизводимость и аккуратность.
- 22 команды-участницы.

Интерпретация вариантов

«OK, Google»

Таблица: Среднее количество вариантов у представителя европейской популяции

Эксперимент	N
WGS	3,000,000
WES	25,000
CES	7,500

Функциональные классы вариантов

AUG UUG CAU GUA UUG AUA GGG UAU UAG

DNA	HAS	ALL	YOU	CAN	ASK	FOR	normal
DNA	HAS	ALL	YOO	CAN	ASK	FOR	silent
DNA	HAS	ALL	LOU	CAN	ASK	FOR	missense
DNA	HAS	ALL	YOU	STOP			nonsense
DNA	HAS	ALY	OUC	ANA	SKF	OR	frameshift

Интерпретация вариантов

Какая информация полезна?

- Эффект на белок/транскрипт
- Частота аллели (1000 Genomes, ExAC, ESP6500, gnomAD)
- Предсказания функционального эффекта (SIFT, Polyphen2, PROVEAN, мета-скоры - собрано в dbNSFP)
- Клинические данные (ClinVar, OMIM, HGMD, NHGRI GWAS Catalog)
- Семейный анализ, аннотация *de novo* mutations (DNM)

ACMG Standards and Guidelines

Классификация генетических вариантов:

- Pathogenic
- Likely pathogenic
- Likely benign
- Benign
- Variant of uncertain significance

ACMG
American College of Medical Genetics and Genomics
Translating Genes Into Health®

Некоторые критерии ACMG

Very strong

PVS1
Loss-of-function (LoF) варианты в генах с известной ролью в патогенезе.

Strong

PS1
Аминокислотная замена, аналогичная ранее описанной.
PS2
DNM у пациента без семейной истории болезни.
PS3
In vitro или *in vivo* исследования.

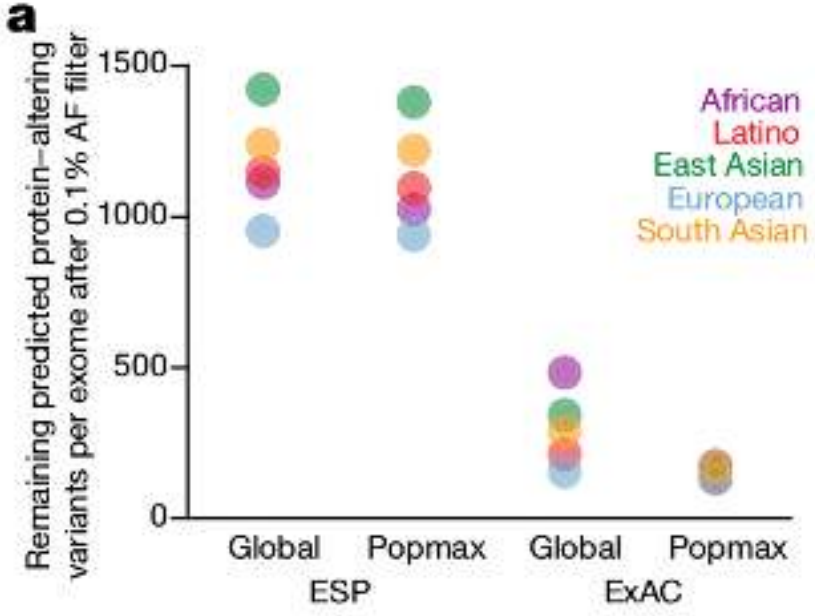
Moderate

PM2
Отсутствие в референсных выборках.
PM6
Предполагаемая *de novo* природа.

Патогенный ли?

Находки Exome Aggregation Constortium (ExAC, BROAD Institute of Harvard and MIT):

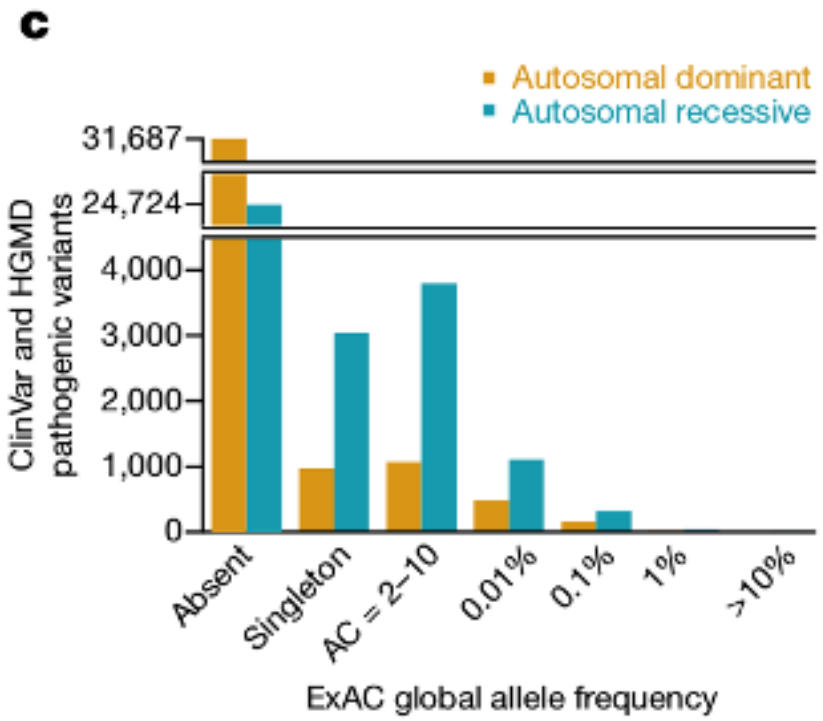
- Каждый индивид несет ≈ 100 ***редких*** значащих вариантов
- Большинство патогенных по литературе вариантов - слишком частые, чтобы вызывать заболевание



Патогенный ли?

Находки Exome Aggregation Constortium (ExAC, BROAD Institute of Harvard and MIT):

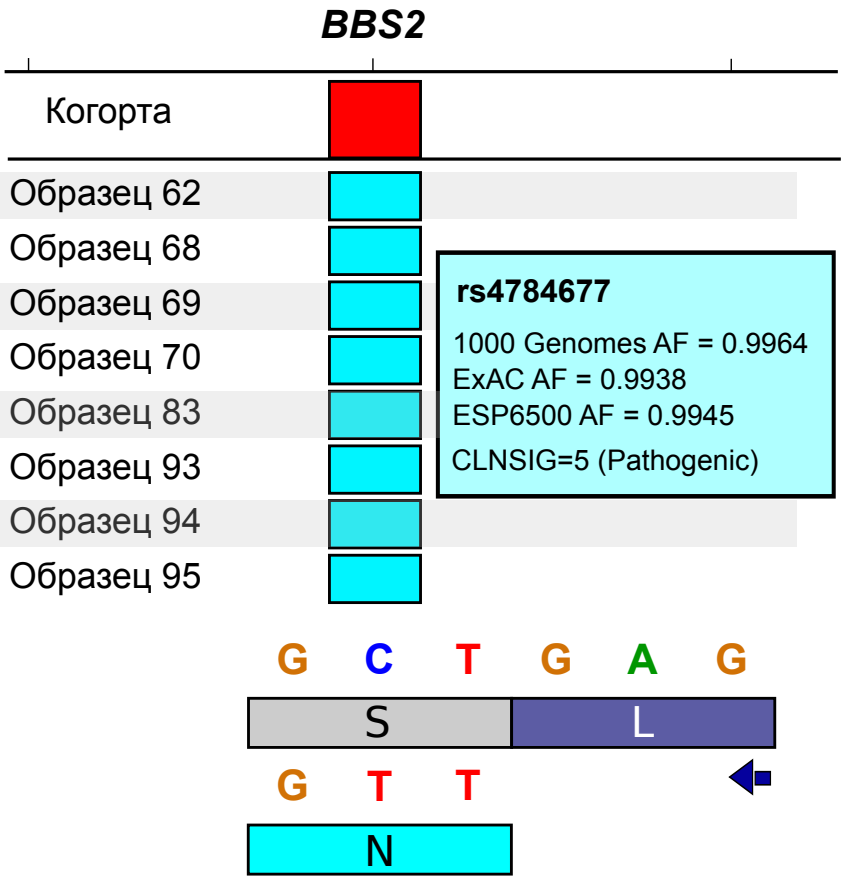
- Каждый индивид несет ≈ 100 редких значащих вариантов
- Большинство патогенных по литературе вариантов - слишком частые, чтобы вызывать заболевание



Референсные минорные аллели

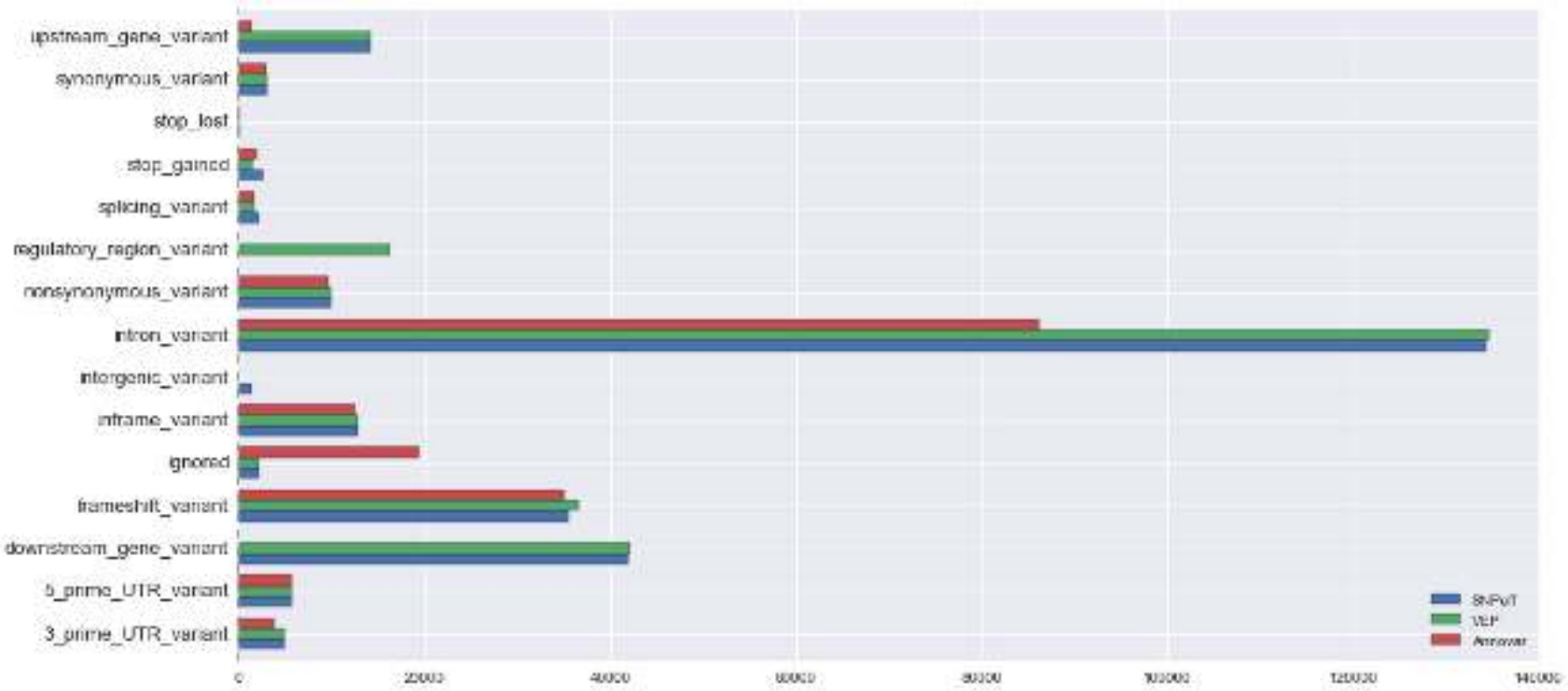
Референсный геном содержит множество редких и патогенных аллелей.

- Нормальная аллель определяется как вариант
- Патогенная аллель не определяется как вариант
- Изменяется эффект близлежащих вариантов



Несовпадение аннотаций

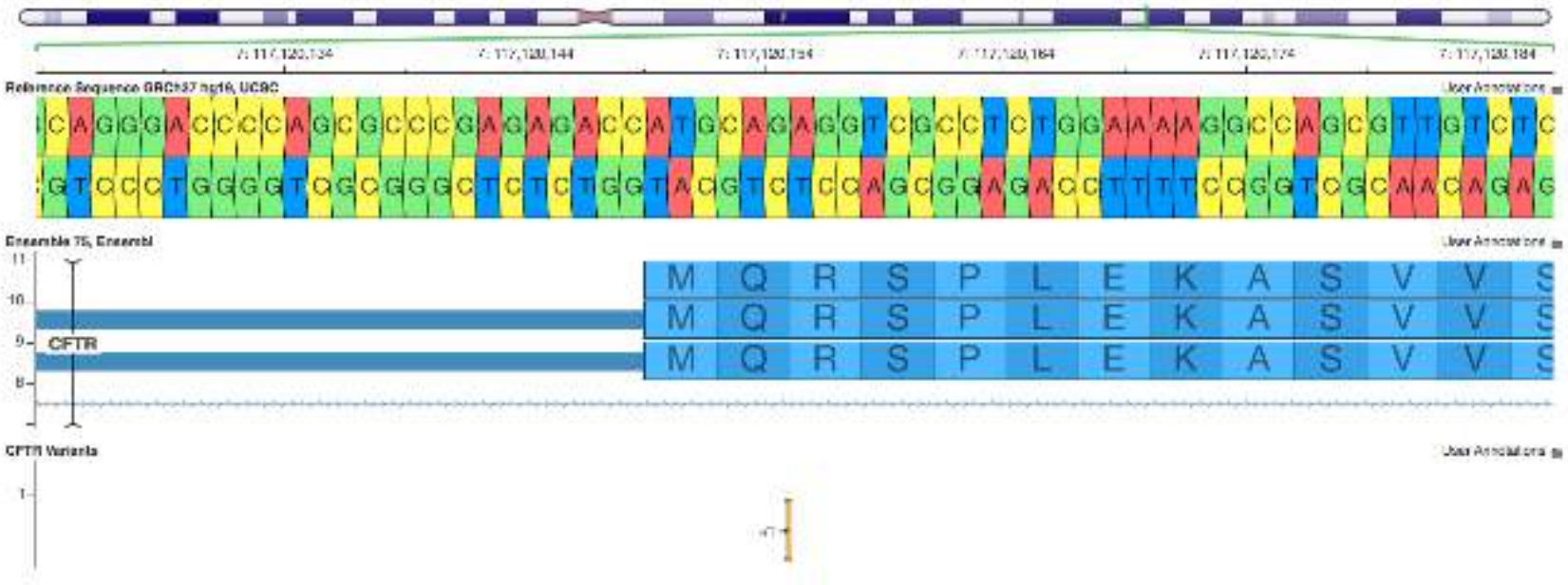
Сравнение классификации одного и то же набора симулированных вариантов разными программами:



http://blog.goldenhelix.com/goldenadmin/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/

Несовпадение аннотаций

Инсерция T в 6-ом положении приводит к стоп-кодону из-за сдвига рамки считывания:



Сдвиг рамки или нонсенс-мутация?
http://blog.goldenhelix.com/goldenadmin/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/

Интерпретация аллельных частот инделов

Частота короткой делеции $< 1\%$. Частота длинной в том же участке - $\approx 24\%$.

ACMG Incidental Findings

Случайные находки

Случайная находка - известный или ожидаемо патогенный вариант, не имеющий связи с основным диагнозом.

Таблица: Примеры рекомендованных к выдаче в заключении случайных находок

Фенотип	Ген	Тип варианта
Breast/ovarian cancer	*BRCA1*	Known, expected
Neurophybromatosis type II	*NF2*	Known, expected
Familial hypercholesterolemia	*APOB*	Known

Green *et al.* (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.

Случайные находки: случай #1

Число патогенных вариантов у 6503 индивидов из проекта ESP6500 (NHLBI):

- По HGMD: 615
- По критериям ACMG всего: 81
- **В списке генов ACMG Incidental Findings: 63**

Из Amendola *et al.* (2015) *Genome Res*

National Heart Lung and Blood Institute

Случайные находки: случай #2

OPEN ACCESS Freely available online PLOS ONE

Whole-Genome Sequencing of the World's Oldest People

Hinco J. Gierman[1], Kristen Fortney[1], Jared C. Roach[2], Natalie S. Coles[3,4], Hong Li[2], Gustavo Glusman[2], Glenn J. Markov[1], Justin D. Smith[1], Leroy Hood[2], L. Stephen Coles[3,4], Stuart K. Kim[1]*

1 Dept. of Developmental Biology and Genetics, Stanford University, Stanford, CA, United States of America, 2 Institute for Systems Biology, Seattle, WA, United States of America, 3 Gerontology Research Group, Los Angeles, CA, United States of America, 4 David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America

Molecular Genetics & Genomic Medicine

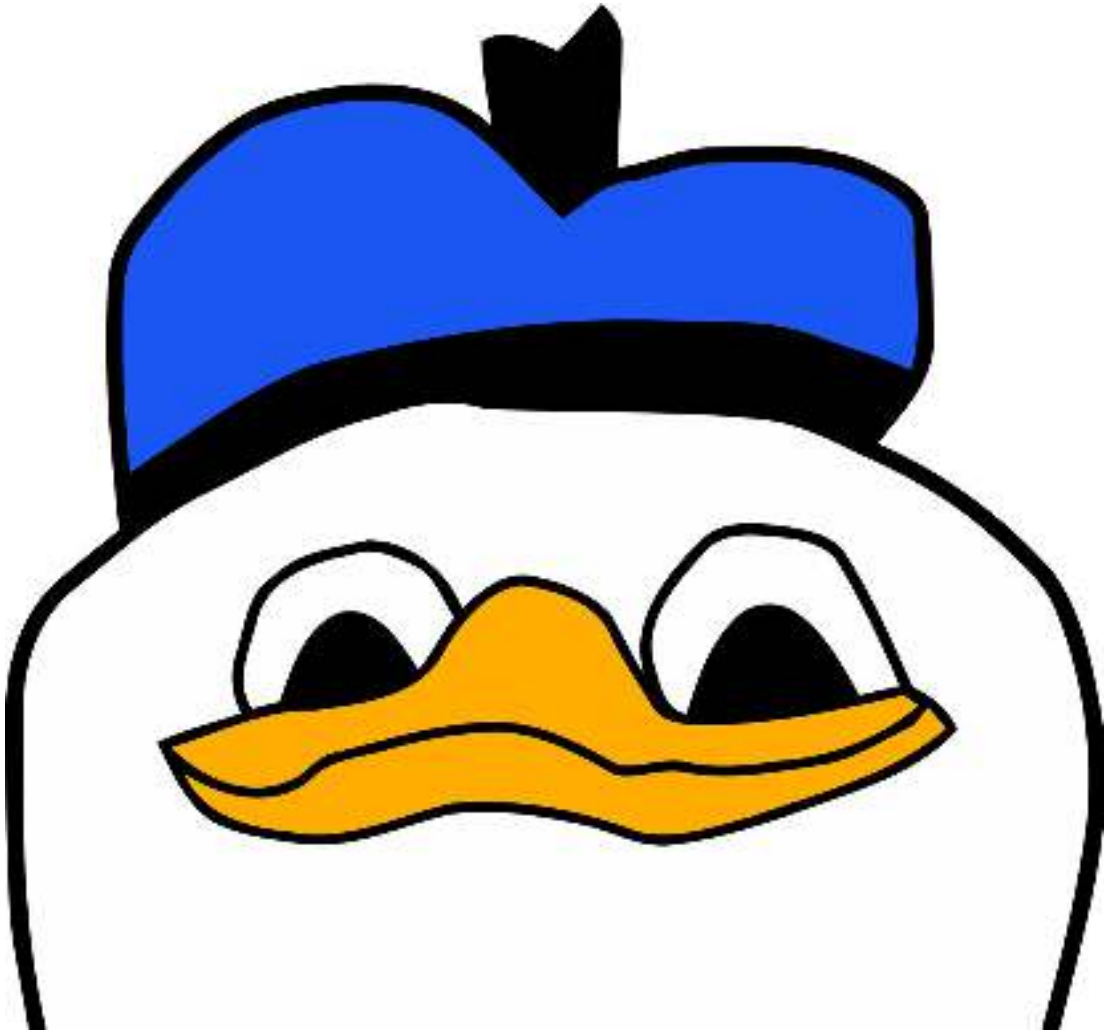
Open Access

ORIGINAL ARTICLE

Disease variants in genomes of 44 centenarians

Yun Freudenberg-Hua[1,2], Jan Freudenberg[3], Vladimir Vacic[4], Avinash Abhyankar[4], Anne-Katrin Emde[4], Danny Ben-Avraham[5], Nir Barzilai[5], Dayna Oschwald[4], Erika Christen[1], Jeremy Koppel[1,2], Blaine Greenwald[2], Robert B. Darnell[4,6], Soren Germer[4], Gil Atzmon[5] & Peter Davies[1]

Все очень плохо?



Выявляемость патогенных вариантов

Выявляемость патогенных вариантов на образец по данным DG MacArthur (BROAD Institute):

- $\approx$ 30% - по данным WES
- $\approx$ 50% - по данным WGS

Daniel G. MacArthur

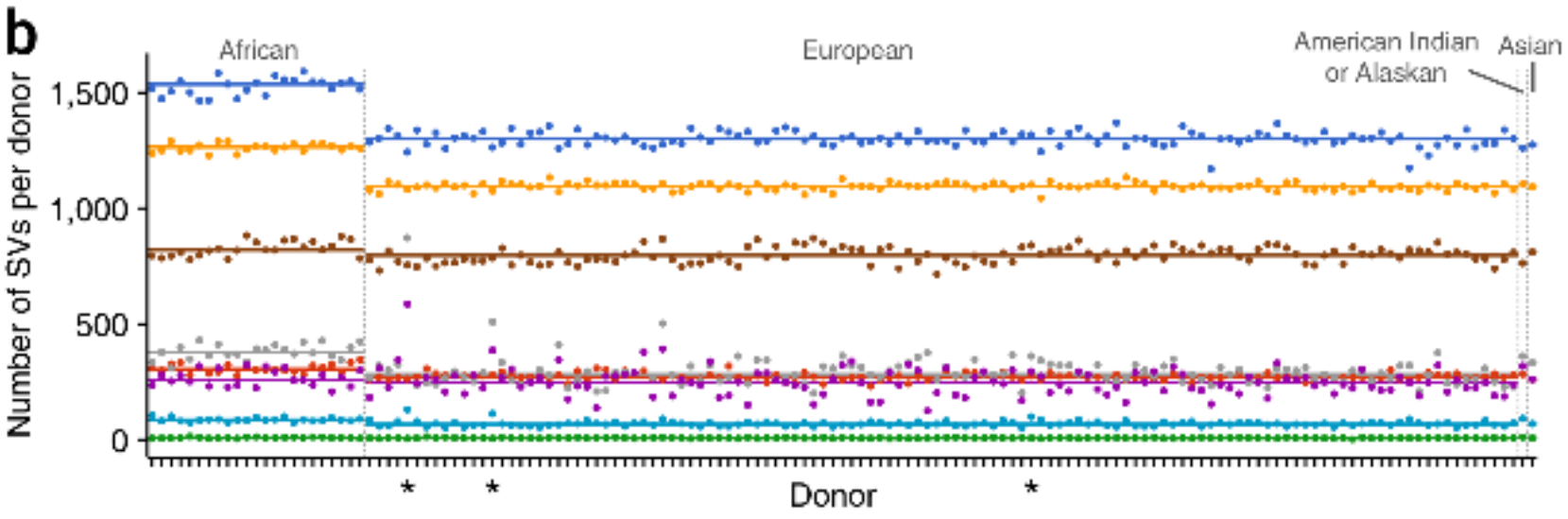
Куда уходят 50%?

Факторы, понижающие диагностируемость

- Структурные вариации (+ вариации числа копий, CNV)
- Эпистатические взаимодействия, влияние экспрессии генов и хроматинового окружения.
- Сложность интерпретации некодирующей (регуляторной) вариации.
- Регионы с систематическим отсутствием покрытия.
- Вырожденные (неуникальные) регионы.

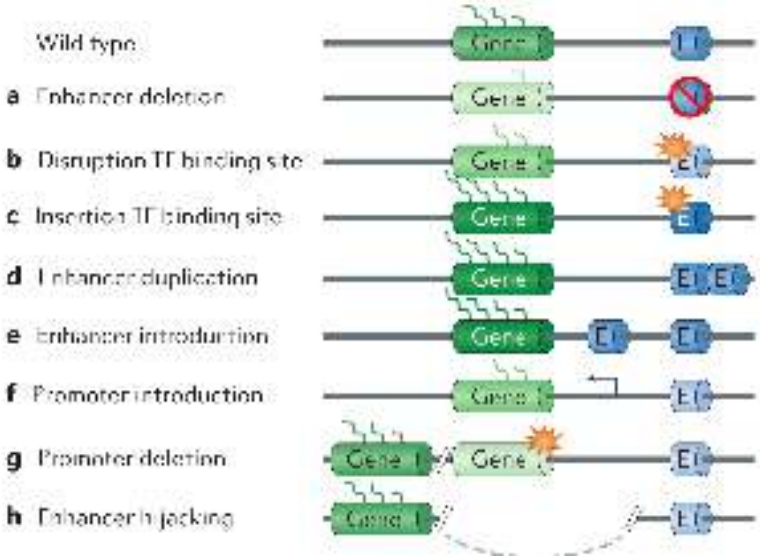
Структурные вариации (SV)

- Эффективность определения SV по данным NGS невелика
- Стандартный инструмент - LUMPY
- Методами NGS можно уловить $\approx$ 4,500 SV



Chiang *et al.* (2017) The impact of structural variation on human gene expression.

Роль экспрессии генов



Disease	Affected gene	Enhancer	Refs
β-Thalassaemia	β-globin genes	LCR	3,4
HPE	SHH	SBE2	36
PDD2	SHH	ZRS	35,122
Lung adenocarcinoma	MYC	3′~450 kb SE	124
T ALL	TAL1	NA	121
α-Thalassaemia	α-Globin genes	α-Globin enhancers	46
α-Thalassaemia*	α-Globin genes/NME4	α-Globin enhancers	44
Burkitt lymphoma	MYC	IgH enhancer	116,117

Krijger P. and de Laat W. (2016) Regulation of disease-associated gene expression in the 3D genome.

Third generation sequencing

Oxford Nanopore PromethION:

- 144 000 нанопор (выход ≈ 2000 Гб)
- Дешевое секвенирование
- Быстрая пробоподготовка
- Длина прочтений до 1,000,000 о.

Нужен ли референсный геном?

Проблемы выравнивания на референсную последовательность генома:

- Высоковариативные регионы
- Неуникальные регионы
- Регионы с инделами

Большинство программ ориентировано на локальную сборку гаплотипов (**fermi.kit** - на полную).

Резюме

Основные проблемы и пути развития

- Аккуратность определения вариантов далека от 100%.
- Интерпретация вариантов затруднена техническими неточностями:
  - Ошибками в литературе и базах данных.
  - Патогенными аллелями в референсе.
  - Неверным картированием и аннотацией транскриптов.
  - *to be continued...*
- Современных технологий недостаточно для полноценных исследований всех типов вариации.
- **Понимание молекулярных и клеточных механизмов развития заболеваний недостаточно**.

Further reading

1. Lek M., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans.
2. Richards S., *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.
3. GATK Best Practices workflow: https://software.broadinstitute.org/gatk/best-practices/
4. *Saleheen *et al.* (2017) Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity.

Спасибо за внимание!