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

Варианты в индивидуумах и популяциях

План лекции

1. Большие геномные проекты
2. Теория коалесценции
3. Оценки разнообразия нуклеотидов в геноме человека
4. ExAC, gnomAD, dbSNP
5. Исследования российской популяции

Большие геномные проекты

2001 Геном человека
2003 Encyclopedia of DNA Elements (ENCODE)
2004 Исследования по ресеквенированию (Геном человека – опять)
2005 HapMap: 11 популяций
2006 UK Biobank: 500,000 волонтеров
2007 Индивидуальные геномы: Craig Venter, James Watson
2009 Genome Reference Consortium Human Build 37
2012 100 геномов: 2,504 из 26 популяций. NHLBI Exome Sequencing Project: 6,500
2013 Genome Reference Consortium Human Build 38. NCBI ClinVar, ClinGen
2016 ExAC, gnomAD: 60,706 экзомов из 6 больших популяций и 14 когорт по заболеваниям; >125,000 экзомов, >71,000 полных геномов
2021 Telomere-to-Telomere (T2T) Consortium – снова Геном человека!
2022 UK Bioibank: >150,000 полных геномов

Случайный дрейф и мутации // revisited

Модель бесконечного числа аллелей: каждая мутация создает новый аллель в популяции

Гетерозиготность $H = \frac{\theta}{1+\theta}$, где $\theta = 4N_e\mu$

N_e: эффективный размер популяции, $\sim$10,000

μ: скорость мутации на 1 сайт на 1 поколение, $\sim 1.2 \cdot 10^{-8}$

$$\theta = 4 \cdot 10^4 \cdot 1.2 \cdot 10^{-8} \approx 5 \cdot 10^{-4}$$

$$\theta \ll 1 \Rightarrow H \approx \theta = 1/2000$$

Теория коалесценции

Цель: оценить число сегрегирующих (=неконсервативных) сайтов в выборке из N последовательностей

```
A A A A T T T T A G G G C C C C
A A A A T T T T G G G G C T C C
G A A A C T T T A G G G C C C C
G A A A T T T T A G G G C C C C
```

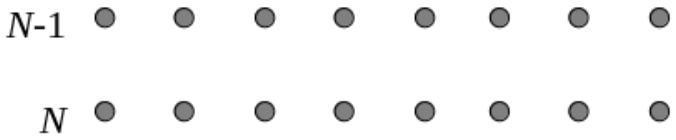
Предположения:

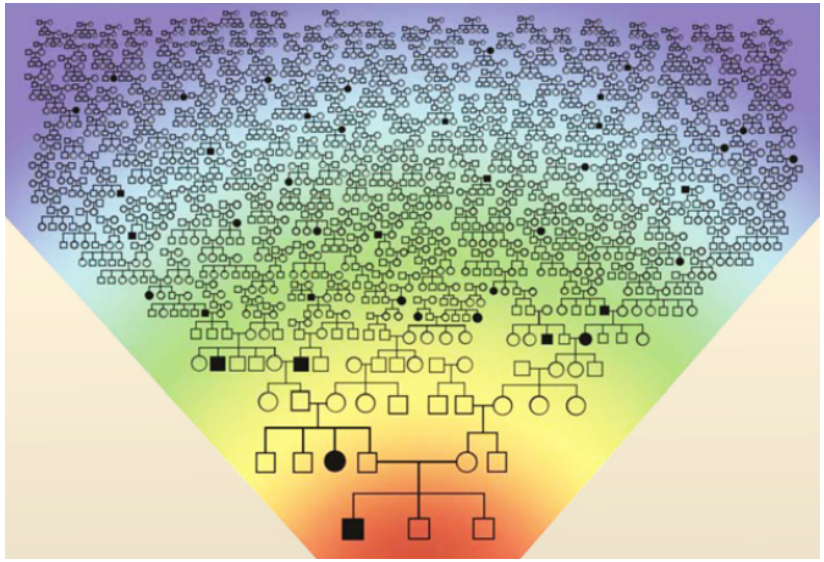
- Случайное размножение (= случайный дрейф) в популяции постоянного размера
- Случайные нейтральные мутации

Метод: генерировать случайную генеалогию индивидуумов обратно во времени, затем добавлять мутации

Теория коалесценции

У каждого человека:
$2^1 = 2$ родителя
$2^2 = 4$ бабушек и дедушек
$2^3 = 8$ прабабушек и прадедушек
...
Макро: у некоторых индивидуумов общие предки, некоторые не имеют потомков
Микро:

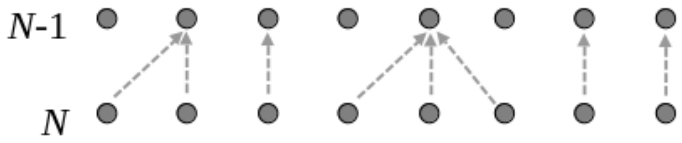


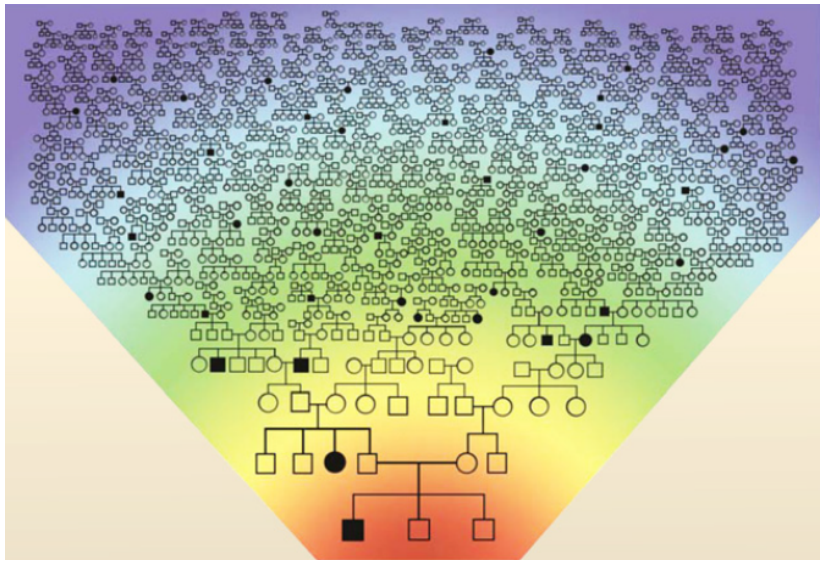


Lupski (2011) *Cell*

Теория коалесценции

У каждого человека:
$2^1 = 2$ родителя
$2^2 = 4$ бабушек и дедушек
$2^3 = 8$ прабабушек и прадедушек
...
Макро: у некоторых индивидуумов общие предки, некоторые не имеют потомков
Микро:





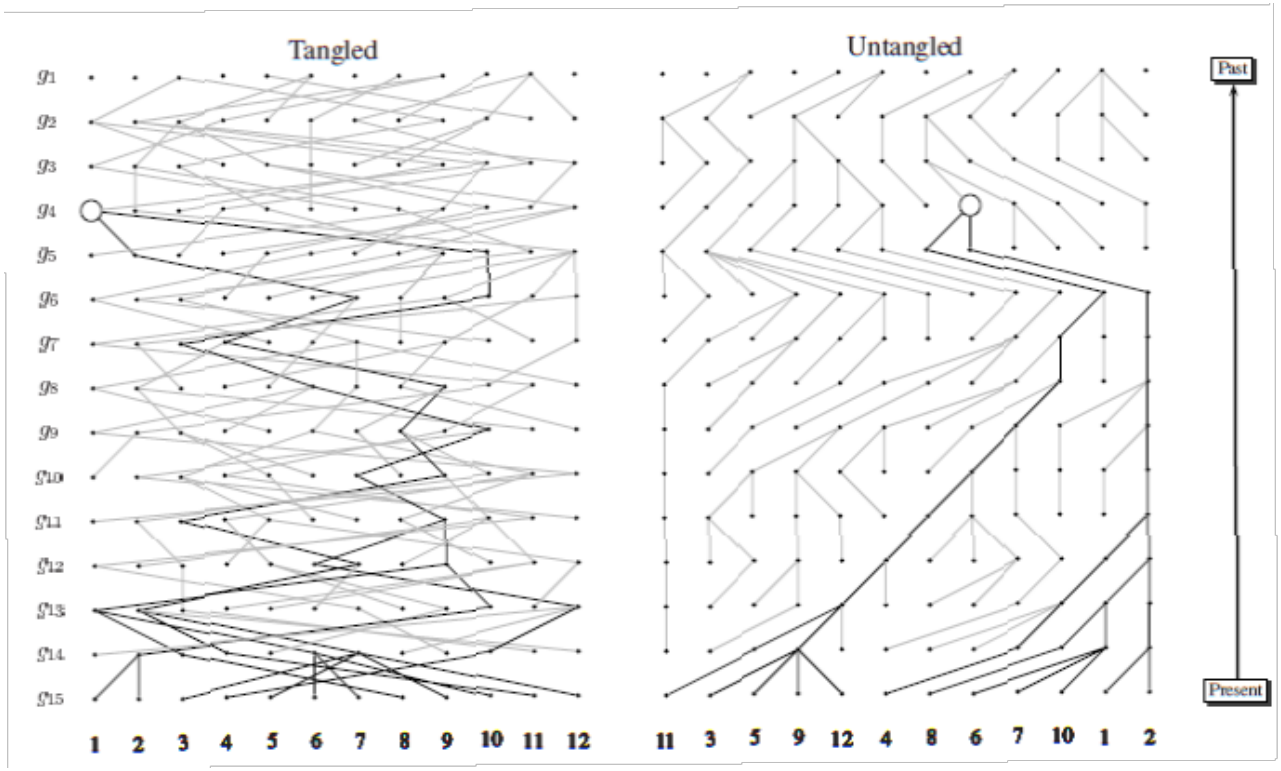
Lupski (2011) *Cell*

Появление самого близкого общего предка популяции, размножающейся половым путем, константного размера N, ожидается через $\sim log_2 N$ поколений
// Rhode (2004) *Nature*

Упражнение

Оцените время до самого близкого общего предка человеческой популяции

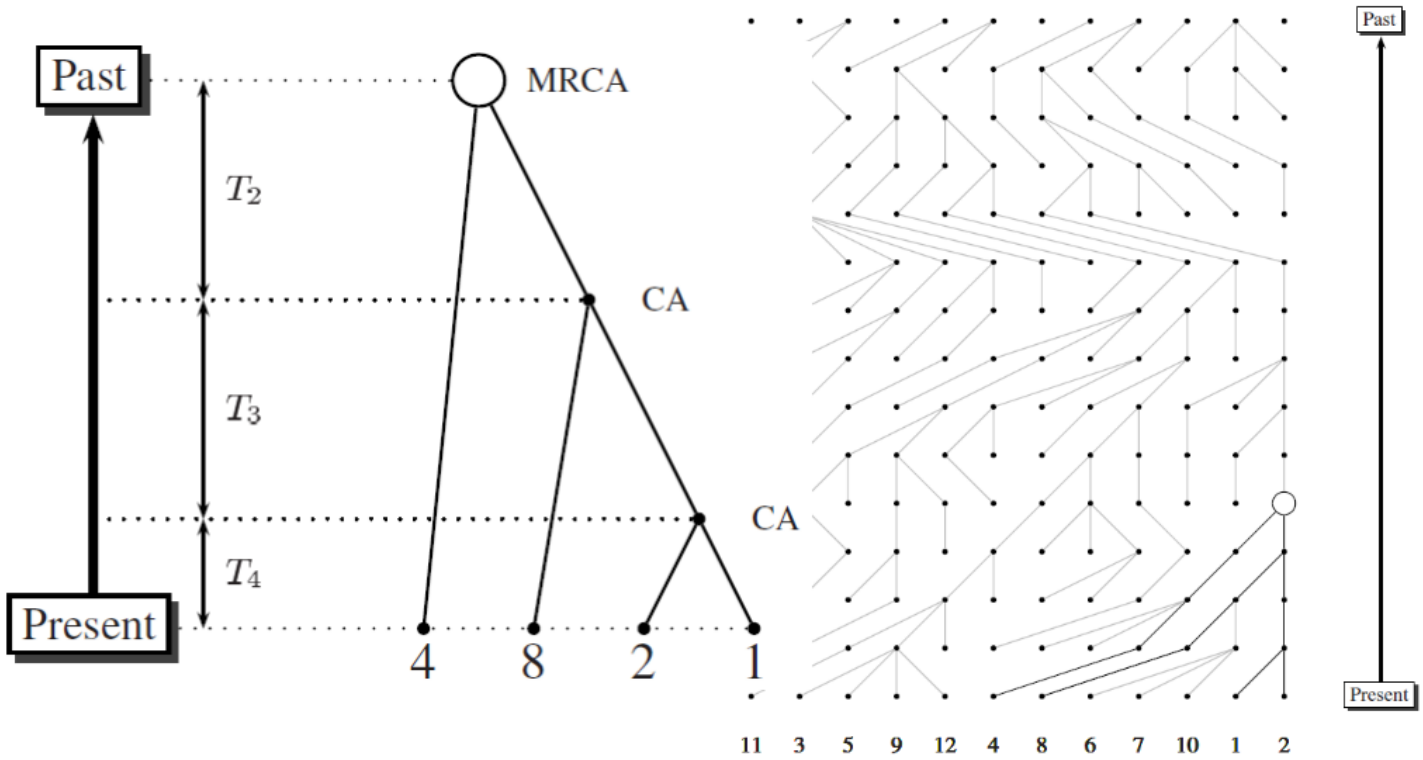
Теория коалесценции



Haubold & Wiehe (2006) – *Introduction to computational biology*

Родословная (генеалогия) для 12 генов в 15 поколениях при предположениях модели эволюции Райта-Фишера, где поколения появляются за счет случайного выбора с возвращением. ∘ означает самого близкого общего предка; черные линии — родословная сохранившихся генов, серые линии означают исчезнувшие родословные.

Теория коалесценции



Haubold & Wiehe (2006) – *Introduction to computational biology*

Линии потомства для 4 генов из подграфа генеалогии популяции на прошлом слайде. ○ означает самого близкого общего предка; T_i: интервал времени, в котором коалесценция состоит из i родословных.

Теория коалесценции

Соединение двух линий (родословных называется **событием коалесценции**. Полная топология событий коалесценции называется **коалесценцией**. Другими словами, **коалесценция** – это родословная последовательностей (аллелей, генов, локусов) в выборке, отслеженная обратно во времени к их (последней, самой близкой) предковой последовательности (last common ancestor - LCA, most recent common ancecter - MRCA). **Теория коалесценции** смотрит обратно во времени и объединяет последовательности, произошедшие от LCA.

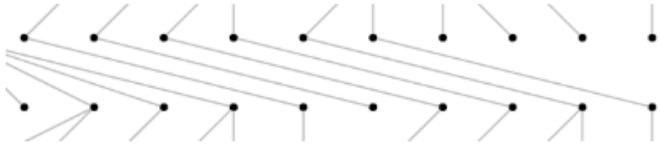
Можно описать свойства ансамбля деревьев коалесценции, совместимых с наблюдаемыми данными; при этом ни одно отдельно взятое дерево не может быть известно.

Коалесцентные деревья являются удобным и вычислительно эффективным способом для определения важных свойств изменчивости последовательностей генома.

Генетические события, такие как мутации, на основе которых можно различить последовательности, должны произойти после их происхождения от LCA. Напротив, любое событие произошедшее до LCA, одинаково повлияло на всех в популяции и поэтому остается незамеченным.

Теория коалесценции

Any n distinct alleles in generation G_i have ancestors in G_{i-1}. The probabilities that the ancestor of the allele 2 is distinct from the ancestor of 1; the 3 is distinct from 1 and 2, and so on:



$$\frac{2N-1}{2N} \rightarrow \frac{2N-1}{2N} \times \frac{2N-2}{2N} \rightarrow \ldots$$

The probability that n alleles all have distinct ancestors in G_{i-1};

$$\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\ldots\left(1-\frac{n-1}{2N}\right) \approx 1-\frac{1}{2N}-\frac{2}{2N}-\ldots-\frac{n-1}{2N}$$

The probability Pc that a coalescence occurs is one minus the probability that it does not:

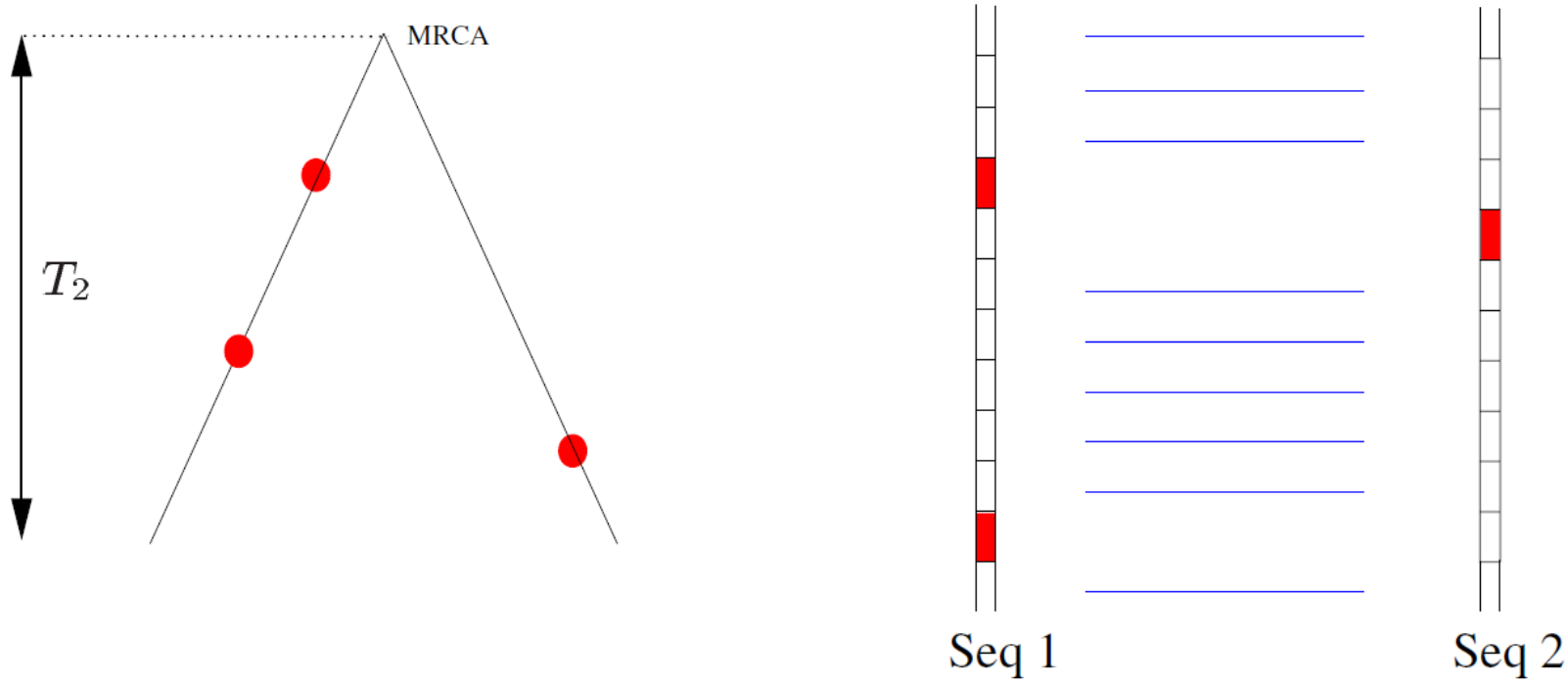
$$P_c = \frac{1+2+\ldots+(n-1)}{2N} = \frac{n(n-1)}{4N}$$

The probability that the first coalescence occurs after exactly t+1 generations is therefore $(1-Pc)^t Pc$. Coalescence times are geometrically distributed with parameter Pc. The mean of the geometric distribution is the reciprocal of the probability of success, giving **the mean time leading from a coalescent with *n* alleles to coalescent with *n*-1 alleles**

$$E\{T_n\} = \frac{4N}{n(n-1)}$$

Caption

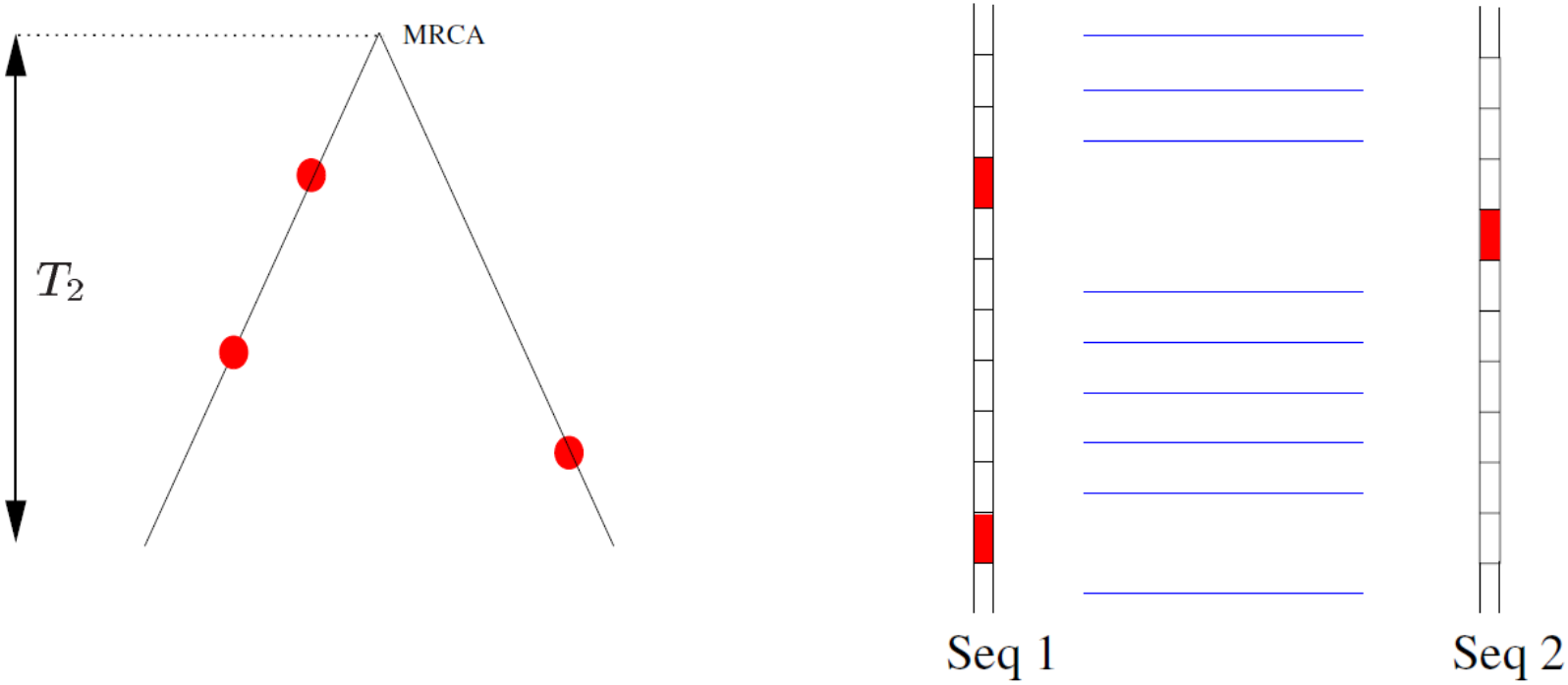
Теория коалесценции



Under the infinite sites model the number of (unobservable) mutations is equal to the number of observable segregating sites (variants) in the sample. For a given coalescence time T_2 the number of segregating sites S_2 per nucleotide is $2T_2\mu$, where μ is the mutation rate per site per generation. What is T_2 then?

Haubold & Wiehe (2006) – *Introduction to computational biology*

Теория коалесценции



The number of segregating sites per nucleotide S_2:

$$T_2 = 4N/2, S_2 = 2\mu T_2 = 4N\mu$$

Haubold & Wiehe (2006) – *Introduction to computational biology*

Теория коалесценции

The total time in all of the branches of a coalescent is

$$T_c = \sum_{i=2}^{n} iT_i,$$

which, using the fact that the expectation of the sum of random quantities is the sum of the expectations of those quantities (see Equation B.11 on page 162), is

$$E\{T_c\} = \sum_{i=2}^{n} iE\{T_i\} = 4N \sum_{i=2}^{n} \frac{1}{i-1}.$$

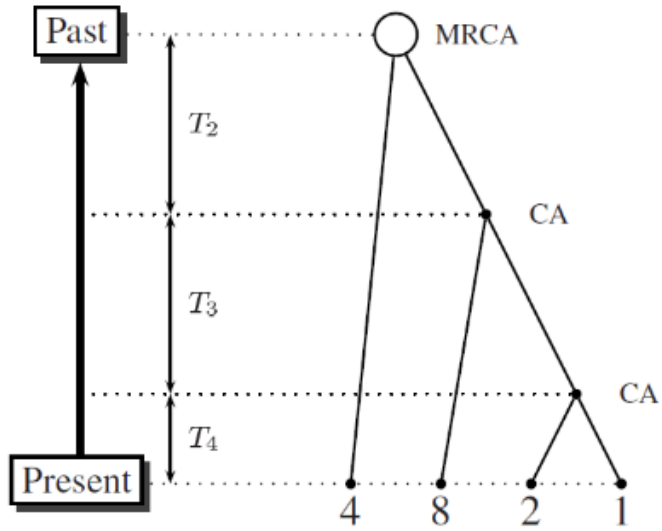
Recalling that the expected number of segregating sites is the neutral mutation rate, u, times the expected time in the coalescent, we have

$$E\{S_n\} = uE\{T_c\} = \theta \sum_{i=2}^{n} \frac{1}{(i-1)},$$

which suggests that

$$\hat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} \cdots + \frac{1}{n-1}}$$

should be a good estimator for $\theta = 4Nu$.



Gillespie – *Population genetics. A concise guide*

Теория коалесценции

Модель бесконечного числа сайтов: каждая мутации изменяет новый сайт в [бесконечно длинной] последовательности нуклеотидов

A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
G	A	A	A	**C**	T	T	T	**A**	G	G	G	**T**	C	C	C
A	**G**	A	A	T	**C**	T	T	G	**A**	G	G	C	**T**	C	C
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6

$$E(S) = \theta_S L \sum_{k=1}^{n-1} \frac{1}{k}, \text{ где } \theta_S = 4N_e\mu_S$$
$$E(\Pi) = \theta_S L, E(\pi) = \theta_S$$

Последовательностей $n = 4$, сегрегирующих сайтов $n = 8$, длина последовательности: $L = 16$, среднее число несовпадений $\Pi = 24/6 = 4$, Разнообразие нуклеотидов: $\pi = H = \Pi/L$, мутации на сайт на поколение μ_S

Упражнение

Как размер выборки влияет на поиск новых вариантов

Оценка разнообразия нуклеотидов в геноме человека

Разнообразие нуклеотидов π = Среднее количество несовпадений П / Длину L

$$E(\pi) = \theta_S, \theta_S = 4N_e\mu_S$$

N_e: эффективный размер популяции
μ_S: Частота мутации на сайт на поколение

$$E(S) = \theta_S L \sum_{k=1}^{n-1} \frac{1}{k}$$

S: общее количество сегрегирующих сайтов в наборе из n последовательностей

Оценка разнообразия нуклеотидов в геноме человека

Разнообразие нуклеотидов π = Среднее количество несовпадений П / Длину L

$$E(\pi) = \theta_S, \theta_S = 4N_e\mu_S$$

N_e: эффективный размер популяции $\sim 10,000$
μ_S: Частота мутации на сайт на поколение$\sim 1.2 \cdot 10^{-8}$

$$\theta_S = 4 \cdot 10^4 \cdot 1.2 \cdot 10^{-8} \approx 5 \cdot 10^{-4}$$

$$E(S) = \theta_S L \sum_{k=1}^{n-1} \frac{1}{k}$$

S: общее количество сегрегирующих сайтов в наборе из n последовательностей

Оценка разнообразия нуклеотидов в геноме человека

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

68 | NATURE | VOL 526 | 1 OCTOBER 2015

Всего 2,504 образцов, длина генома 2.84 Гб.
Ожидаемое количество аутосомных SNV:
$E(S) = \theta_S L(1 + 1/2 + ... + 1/(2 * 2504)) =$
$4.8 \cdot 10^{-4} \cdot 2.84 \cdot 10^{-9} \cdot 9.09 =$12.4 млн

Оценка разнообразия нуклеотидов в геноме человека

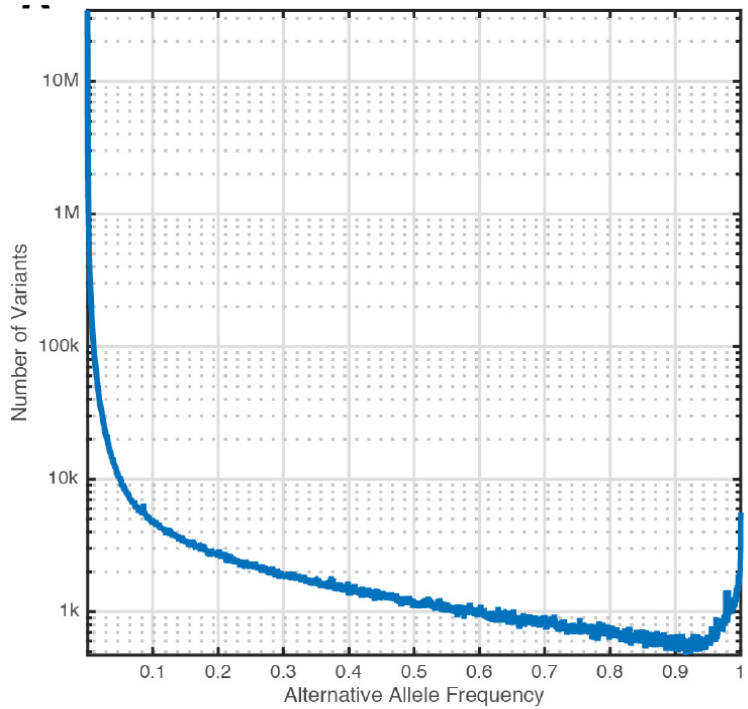
A global reference for human genetic variation

The 1000 Genomes Project Consortium*

68 | NATURE | VOL 526 | 1 OCTOBER 2015

Всего 2,504 образцов, длина генома 2.84 Гб.
Ожидаемое количество аутосомных SNV:
$E(S) = \theta_S L(1 + 1/2 + ... + 1/(2 * 2,504)) =$
$4.8 \cdot 10^{-4} \cdot 2.84 \cdot 10^{-9} \cdot 9.09 =$12.4 млн
Наблюдается:

- 64 млн с MAF $< 0.5\%$
- 12 млн (MAF 0.5-5%)
- 8 млн с MAF $> 5\%$



Caption

...Почему (а) так много (б) редких вариантов?

Избыток редких вариантов в геноме человека

Предположения в основе оценки $E(S)$ по теории коалесценции:

- Константный размер популяции
- Нейтральность вариантов

Ранние оценки: мало образцов $\Rightarrow$ частые (нейтральные) варианты

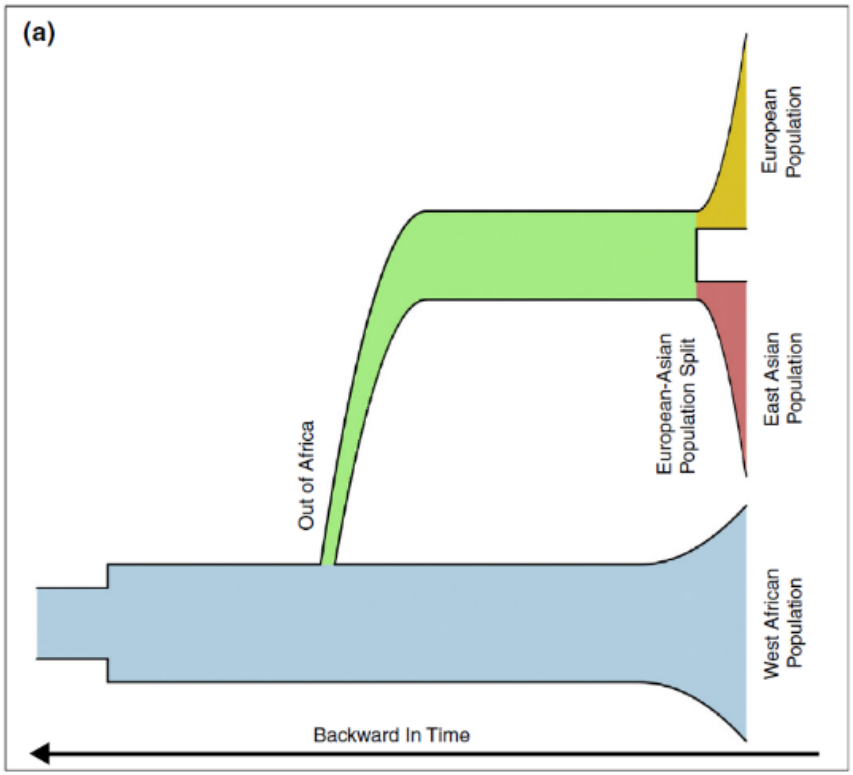
Более реалистично:

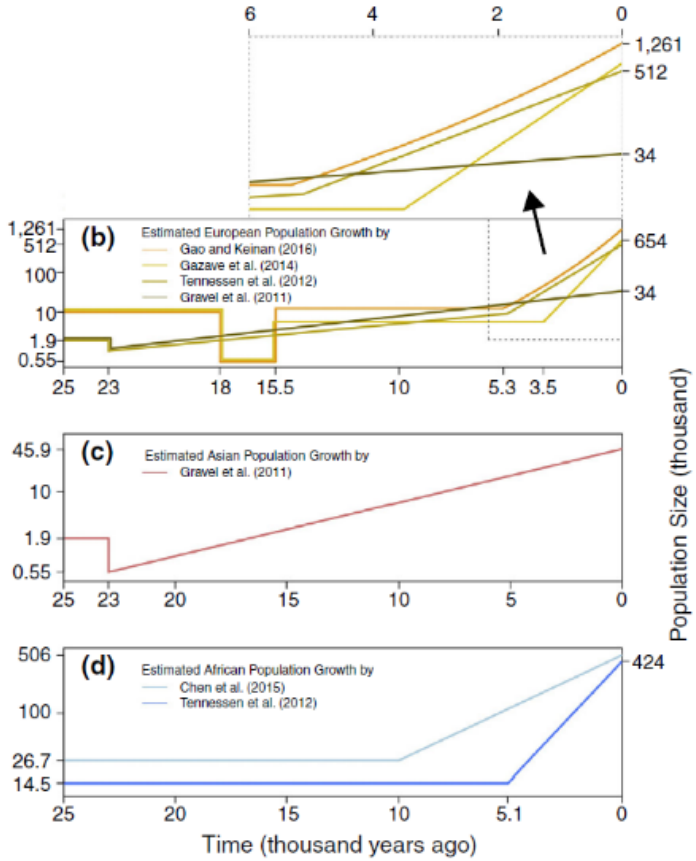
- Демографические модели с недавним **увеличением численности людей**
- **Отрицательный отбор**: уменьшение изменчивости + избыток редких аллелей в остающихся вариантах

Explosive genetic evidence for explosive human population growth

Current Opinion in Genetics & Development 2016, 41:130–139

Feng Gao and Alon Keinan





Explosive genetic evidence for explosive human population growth

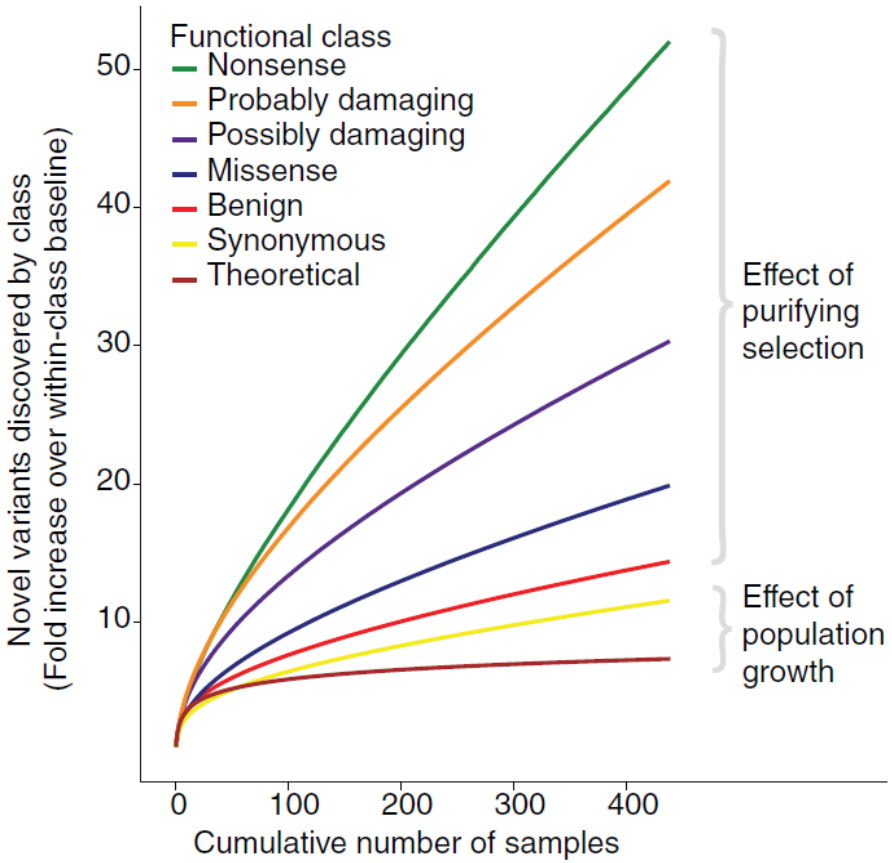
Current Opinion in Genetics & Development 2016, 41:130–139

Feng Gao and Alon Keinan

Implications

One consequence of recent explosive growth is the extreme excess of very rare variants, including those observed only in a single genome out of a large sample (singletons). In fact, explosive population growth predicts not only more rare variants, for example singletons, as the sample size increases, but also a larger proportion of such variants (e.g. [13,14]). A recent study characterized how population growth and purifying selection has shaped the fraction of variants private to an individual, hence the number of new variants that will be discovered with each newly sequenced individual [14]. Assuming 10,000 genomes from the exact same population have already been perfectly sequenced, with growth of the magnitude estimated for Europeans [12••] it predicts >6,000 novel variants to be discovered as heterozygous in the 10,001st sequenced genomes, which is 18-times more than that in the absence of growth. This entails that personalized medicine or personalized genomics will have to be much more personal in recently expanded populations than expected in the absence of growth.

Появление новых вариантов с увеличением выборки



"The number of nonsense variants discovered in 300 samples is 40 times greater than the average number discovered in a single sample, whereas the number of synonymous variants is only 10 times greater (although the absolute number of nonsense variants is a relatively minor proportion of the total variation discovered); this effect is due to purifying selection. All classes of variants are discovered at rates exceeding what would be predicted under a neutral model of evolution in a population of constant size, an effect of population growth."

Kiezun (2012) *Nature Genetics*

Медианное количество аутосомных вариантов на геном

	AFR		EAS		EUR	
Samples	661		504		503	
Mean coverage	8.2		7.7		7.4	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.55M	14.8k	3.53M	11.4k
Indels	625k	-	546k	-	546k	-
Large deletions	1.1k	5	940	7	939	5
CNVs	170	1	158	1	157	1
MEI (Alu)	1.03k	0	899	1	919	0
MEI (L1)	138	0	130	0	123	0
MEI (SVA)	52	0	56	0	53	0
MEI (MT)	5	0	4	0	4	0
Inversions	12	0	10	0	9	0
Nonsynon	12.2k	139	10.2k	144	10.2k	116
Synon	13.8k	78	11.2k	79	11.2k	59
Intron	2.06M	7.33k	1.68M	7.39k	1.68M	5.68k
UTR	37.2k	168	30.0k	169	30.0k	129
Promoter	102k	430	81.6k	425	82.2k	336
Insulator	70.9k	248	57.7k	252	57.7k	189
Enhancer	354k	1.32k	289k	1.34k	288k	1.02k
TFBSs	927	4	748	4	749	3
Filtered LoF	182	4	153	4	149	3
HGMD-DM	20	0	16	1	18	2
GWAS	2.00k	0	1.99k	0	2.08k	0
ClinVar	28	0	24	0	29	1

The 1000 Genomes Project Consortium (2015) *Nature*

ExAC

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285

60,706 экзомов неродственных индивидуумов без врожденных заболеваний

- 7,404,909 высококачественных вариантов (1 на каждые 8 п.н.)
- 99% вариантов имеют MAF<1%, 54% синглетоны
- 7.9% мультиаллели
- 317,381 инделы

- Приближение к **насыщению**: 62.8% всех возможных синонимичных C>T в CpG (gnomAD: $\sim$ 85%)
- **Повторение мутаций**: нарушение принципа бесконечного числа сайтов

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285

SNVs	Average	Deviation
PTV *HIGH*	97	6
Missense *MODERATE*	6291	139
Synonymous *LOW*	7192	88
Other *MODIFIER*	561	13
Indels		
Frameshift	69	3
Other	41	3

SNVs	Average	Deviation
Singleton	18	13
<0.01%	177	30
0.01-1%	273	23
1-10%	1308	72
>10%	12365	109
Indels		
<=5%	15	5
>5%	151	6

Упражнение

Почему большинство вариантов здесь частые, а не редкие?

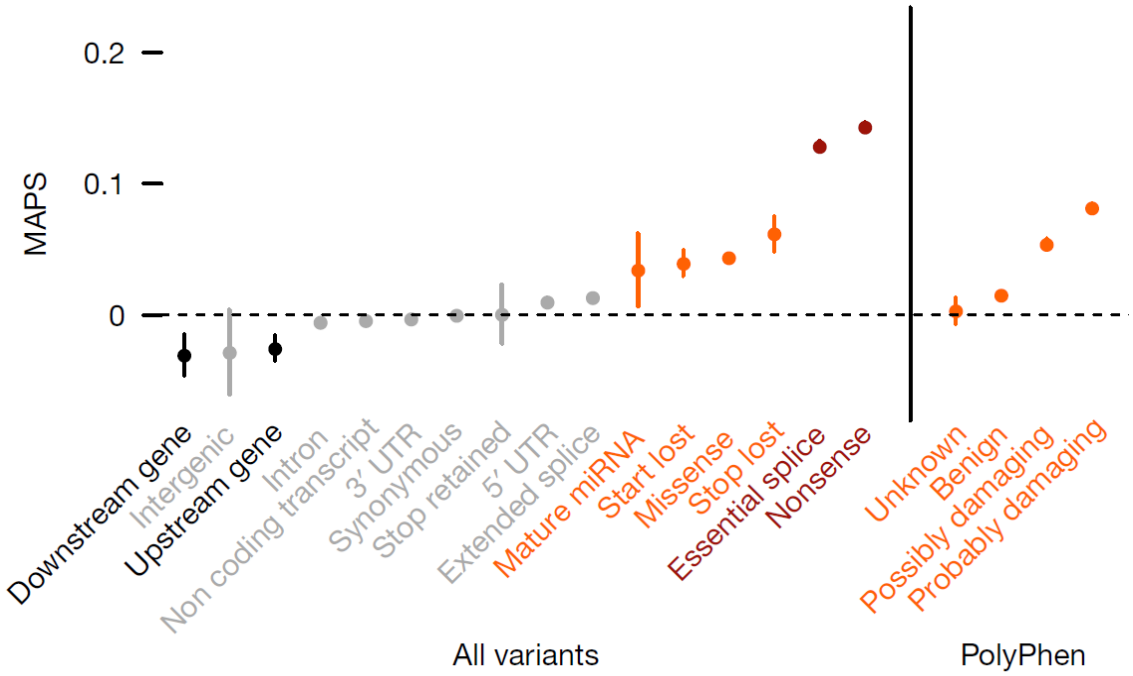
ExAC

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285



Mutability – adjusted proportion of singletons (MAPS)

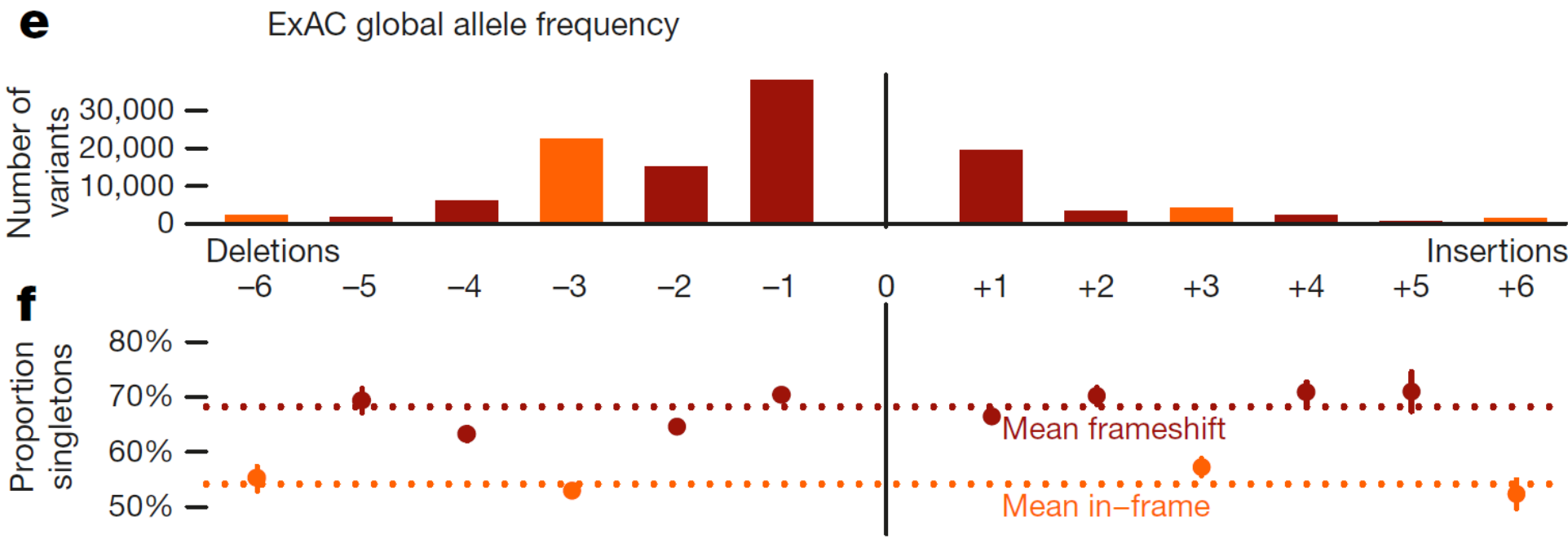
ExAC

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285



Frameshift and in-frame indels

Mutability – adjusted proportion of singletons (MAPS)

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285

Индивидуальные экзомы:

1) Известные болезнетворные варианты

53.7 болезнетворных аллелей из HGMD и CLinVar на экзом, у 47.2 из них AF_POPMAX>1%

Это несовместимо даже с рецессивным наследованием $\Rightarrow$ неверная классификация, неполная пенетрантность

2) Достоверные PTV

179,774 PTV высокой достоверности, 121,309 (67%) синглетоны

- 85 гетерозиготных и 35 гомозиготных PTV на экзом, из которых
- 18 (гетерозиготных) и 0.19 (гомозиготных) редкие (AF$<$ 1%), 2 синглетона

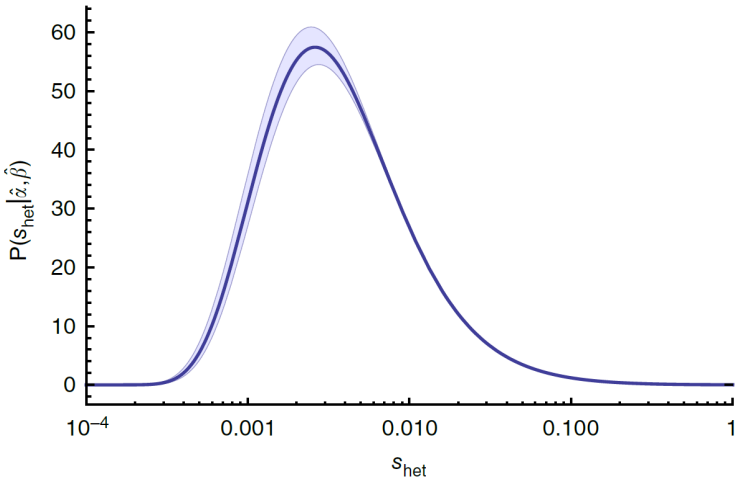
ExAC

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285



S_{het} **applications**:

- Discrimination between AR and AD modes of inheritance
- In dominant diseases, restricting to genes with Shet>0.04 provides a 3x reduction of candidate variants
- S_{het} helps predict phenotypic severity, age of onset, penetrance

"The cumulative frequency of rare deleterious PTVs [in a gene] is primarily determined by the **balance** between incoming mutations and purifying selection rather than genetic drift. This enables the estimation of the genome-wide distribution of selection coefficients for heterozygous PTVs and corresponding Bayesian estimates for individual genes."

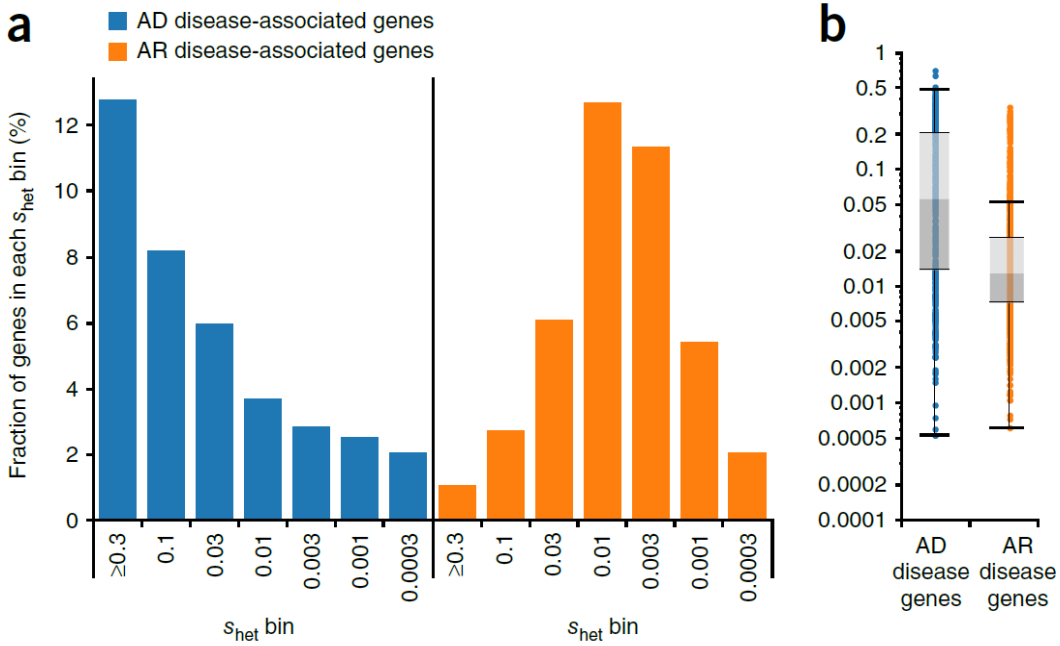
ExAC

ExAC

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2]

18 AUGUST 2016 | VOL 536 | NATURE | 285



Вопрос

Наблюдаются ли все значения S?

Действительно ли PTV являются LoF

Lek (2016) Nature, ExAC paper, ∼60,000 individuals:

- 13.2 ожидаемых LoF вариантов на ген, 62.8% генов имеют более >10 pLoF вариантов в каноническом транскрипте
- Каждый индивидуум имеет порядка ∼ 85 гетерозиготных и ∼34 гомозиготных PTV

Sulem (2015) Nat Genet, ∼101,000 Icelanders: // Популяция основателей

- 7.7% индивидуумов имеют один ген в состоянии полного нокаута LoF вариантом с MAF менее 2%
- Было обнаружено 552 человека, у которых нокаутировано >1 гена
- 1,171 из 19,135 генов (6.1%) полностью нокаутированы

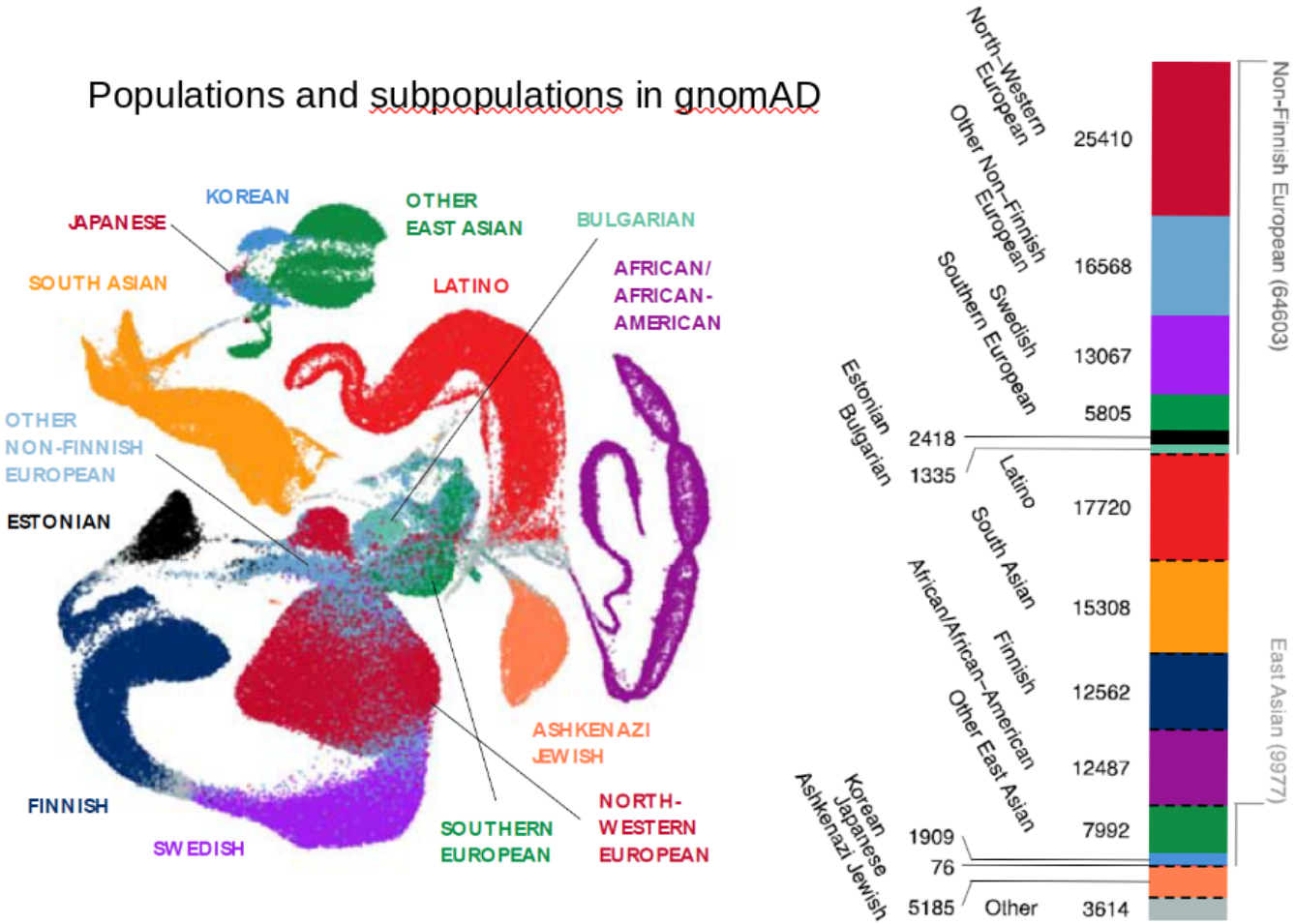
Saleheen (2017) Nature, 10,000 Pakistanis // родственники

- 1,317 различных генов были предсказаны как инактивированые из-за гомозиготных вариантов pLoF
- 17.5% участников имели минимум один нокаутированый ген за счет гомозиготной pLoF мутации, ∼18% имели более одного гена в состоянии нокаута

Backman (2021) Nature 454,787 UK Biobank participants

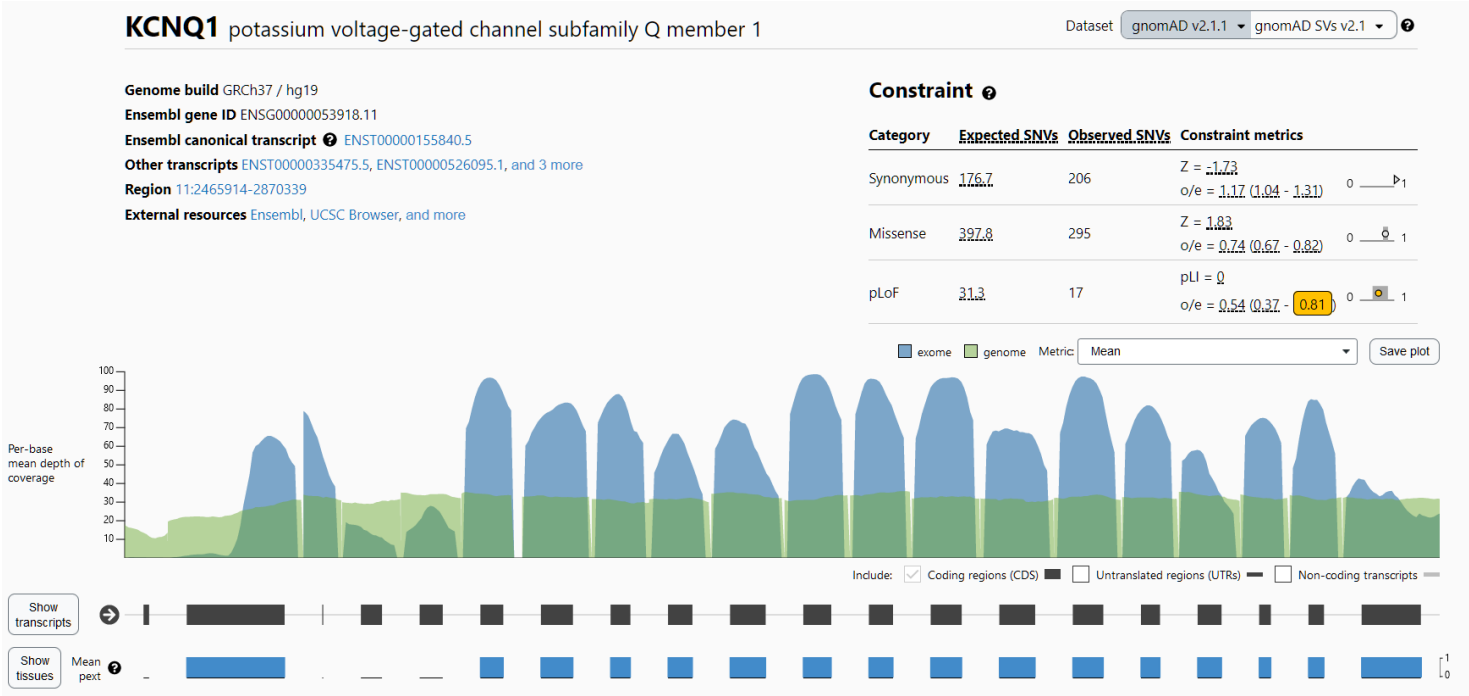
- В >80% генов не менее 50 участников имели предсказанный pLoF вариант

gnomAD 125,748 exomes + 15,708 genomes

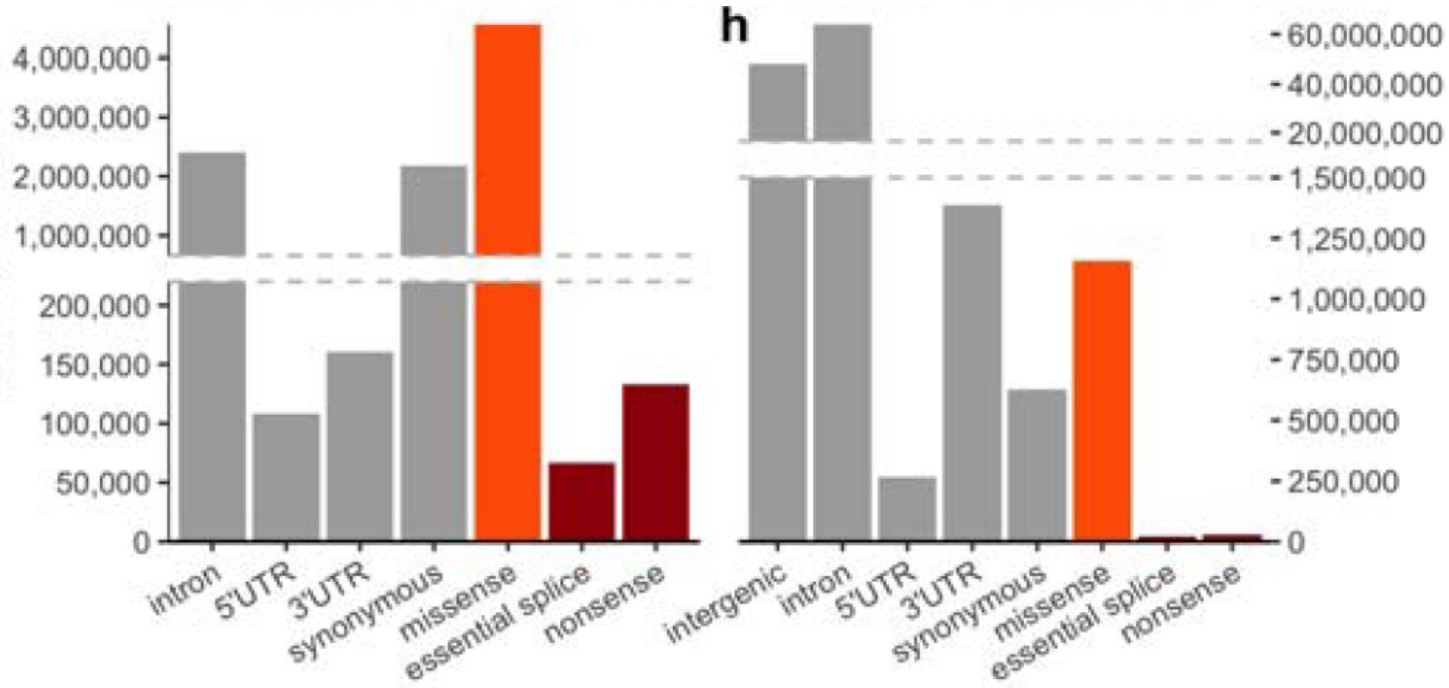


Karczewski *biorXiv* http://dx.doi.org/10.1101/531210

gnomAD 125,748 exomes + 15,708 genomes

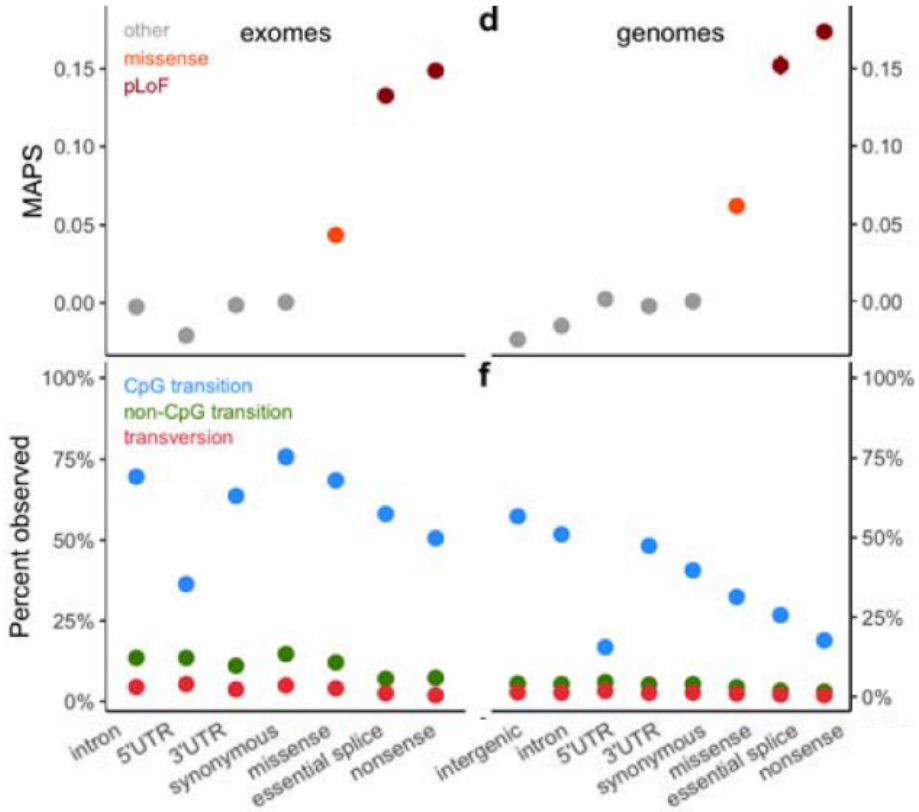


gnomAD 125,748 exomes + 15,708 genomes



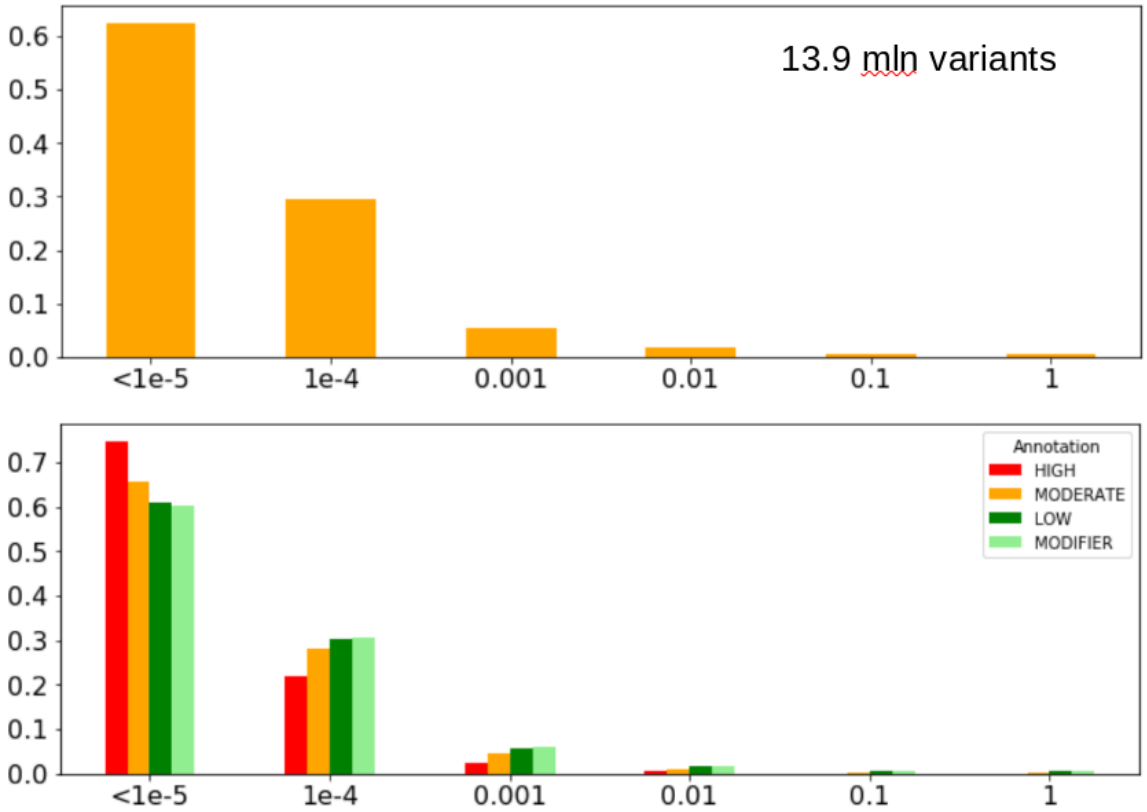
The total number of variants observed in each functional class for exomes (g) and genomes (h).
Karczewski *biorXiv* `http://dx.doi.org/10.1101/531210`

gnomAD 125,748 exomes + 15,708 genomes



(d) The mutability-adjusted proportion of singletons (MAPS) (f) The proportion of all possible variants. Karczewski *biorXiv* `http://dx.doi.org/10.1101/531210`

gnomAD. Частота вариантов в 125,748 экзомах

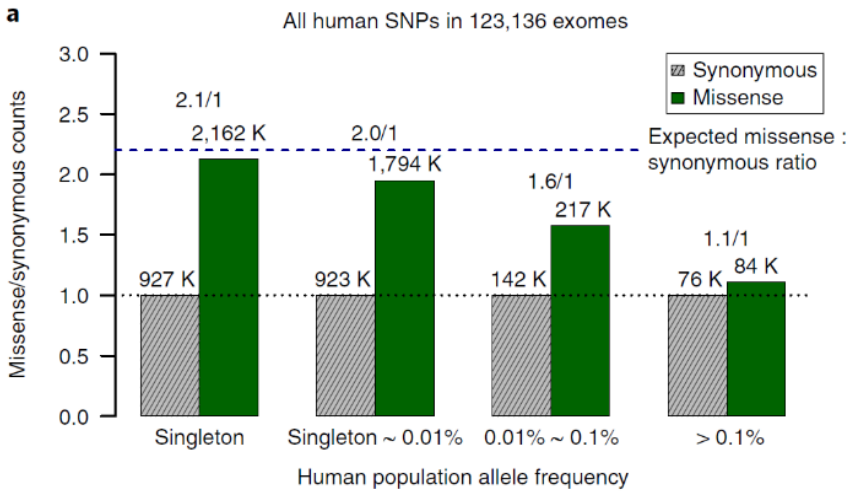


gnomad.broadinstitute.org

nature genetics
ARTICLES
https://doi.org/10.1038/s41588-018-0167-z

Predicting the clinical impact of human mutation with deep neural networks

Laksshman Sundaram[1,2,3,6], Hong Gao[1,6], Samskruthi Reddy Padigepati[1,3], Jeremy F. McRae[1], Yanjun Li[3], Jack A. Kosmicki[1,4], Nondas Fritzilas[1], Jörg Hakenberg[1], Anindita Dutta[1], John Shon[1], Jinbo Xu[5], Serafim Batzloglou[1], Xiaolin Li[3] and Kyle Kai-How Farh[1*]



Вопрос

Объясните утверждение: *~50% из всех новых миссенсов фильтруются очищающим отбором при достижении высоких частот*

LOEUF: гены, нетолерантные к pLoF-вариантам

"Мы классифицирем белок-кодирующие гены человека по спектру, представляющему степени нетолерантности к инактивации pLoF-вариантами"

- **pLoF, putative loss-of-function** $\approx$ PTV (protein-truncating variants)
- LOFTEE: достоверный набор из 443,769 pLoF вариантов (413,097 в канонических транскриптах 16,694 генов)
- Медианное ожидаемое количество pLoF 17.3 на ген, минимум один pLoF в 95.8% всех генов.
- LOEUF: наблюдаемых/ожидаемых pLoF вариантов, разделенных на децили по примерно 1,920 генов каждый
- 1,752 генов которые, скорее всего, толерантны к биаллельной инактивации
- 1,266 генов, в которых не наблюдается pLoF (obs_lof=0, у некоторых высокое exp_lof)

Упражнение

Найдите гены с obs_lof=0

LOEUF: гены, нетолерантные к pLoF-вариантам

ARPC4 actin related protein 2/3 complex subunit 4

Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	37.7	31	Z = 0.86 o/e = 0.82 (0.62 - 1.11)
Missense	106	42	Z = 2.21 o/e = 0.4 (0.31 - 0.51)
pLoF	11.3	0	pLI = 0.97 o/e = 0 (0 - 0.27)

ARPC3 actin related protein 2/3 complex subunit 3

Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	31.3	21	Z = 1.45 o/e = 0.67 (0.47 - 0.97)
Missense	91.6	81	Z = 0.39 o/e = 0.88 (0.74 - 1.06)
pLoF	11.4	3	pLI = 0.22 o/e = 0.26 (0.12 - 0.68)

PCSK9 proprotein convertase subtilisin/kexin type 9

Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	187.5	170	Z = 1.01 o/e = 0.91 (0.8 - 1.03)
Missense	435	419	Z = 0.27 o/e = 0.96 (0.89 - 1.04)
pLoF	26.9	26	pLI = 0 o/e = 0.97 (0.71 - 1.34)

APOBEC1 apolipoprotein B mRNA editing enzyme

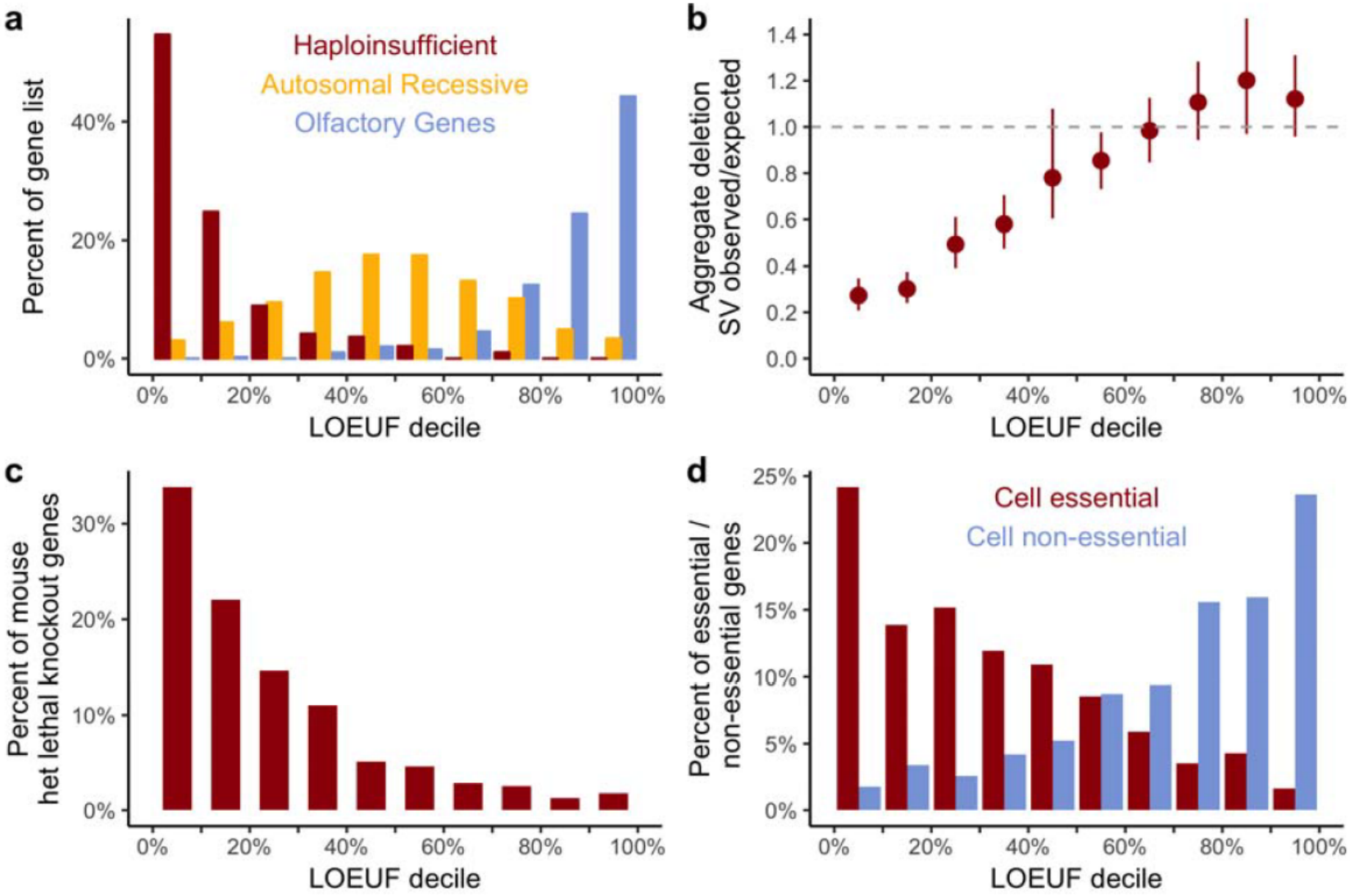
Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	46.7	42	Z = 0.54 o/e = 0.9 (0.7 - 1.16)
Missense	134.2	109	Z = 0.77 o/e = 0.81 (0.69 - 0.95)
pLoF	12.1	12	pLI = 0 o/e = 0.99 (0.63 - 1.59)

Although oe is a continuous value, we understand that it can be useful to use a threshold for certain applications. In particular, for the interpretation of Mendelian diseases cases, we suggest using the upper bound of the oe CI < 0.35 as a threshold if needed. Again, ideally oe should be used as a continuous value rather than a cutoff and evaluating the oe 90% CI is a must.

gnomad.broadinstitute.org

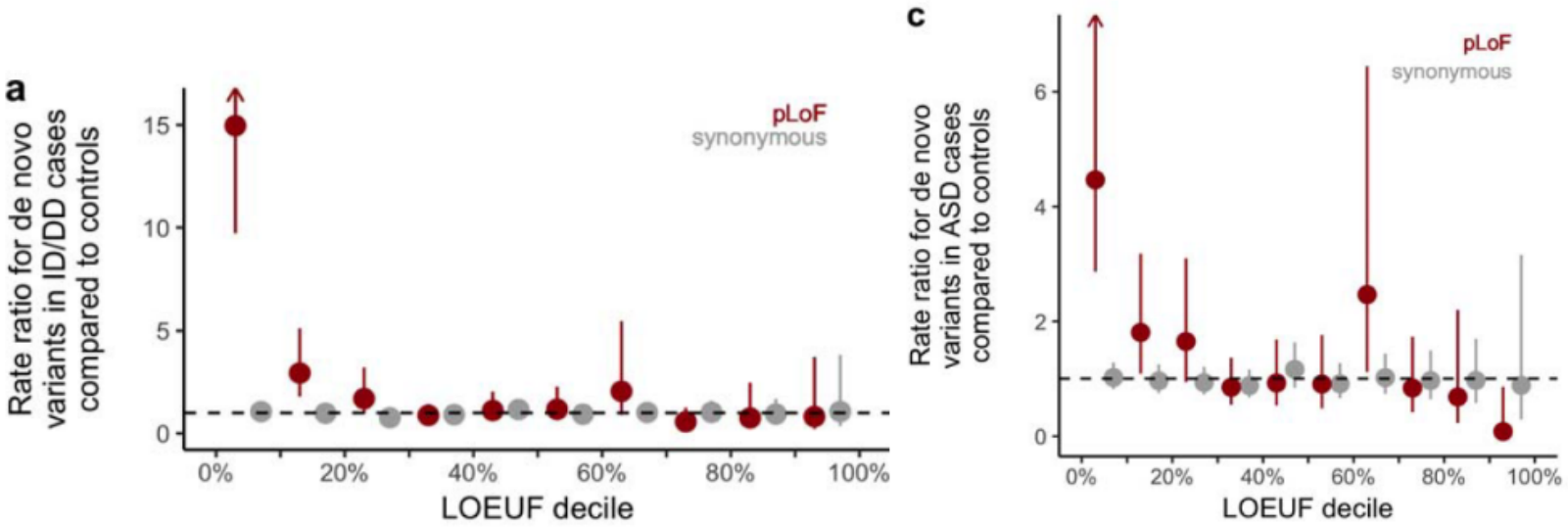
gnomad.broadinstitute.org

LOEUF: гены, нетолерантные к pLoF-вариантам



Karczewski *bioRxiv* http://dx.doi.org/10.1101/531210

LOEUF: гены, нетолерантные к pLoF-вариантам



Karczewski *biorXiv* `http://dx.doi.org/10.1101/531210`

Disease applications of constraint. (a) The rate ratio is defined by the number per patient of de novo variants in **intellectual disability / developmental delay (ID/DD)** cases divided by the rate in controls. pLoF variants in the most constrained decile of the genome are approximately 11-fold more likely to be found in cases compared to controls. (c) **Autism cases**. pLoF variants in the most constrained decile of the genome are approximately 4-fold more likely to be found in cases compared to controls.

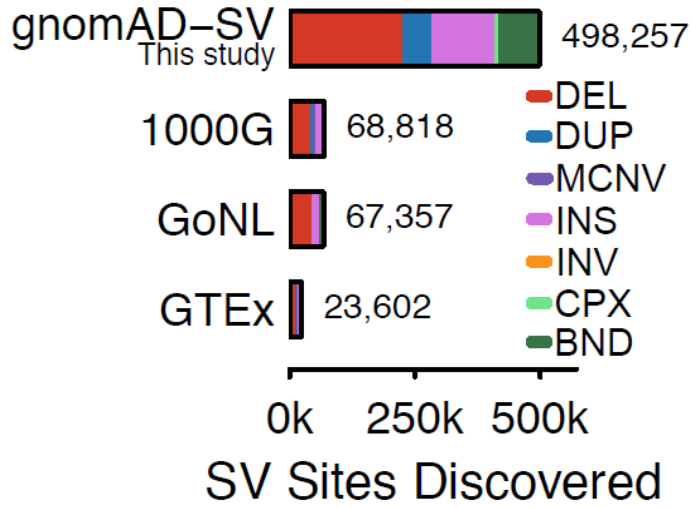
gnomAD: структурные варианты в 14,891 геномах

Структурные варианты (SVs): геномные перестройки, которые изменяют сегменты ДНК длиной более 50 п.н.

- Несбалансированые (CNV) и балансированые (инверсии, транслокации) + более экзотические SV
- Метод: 4 ортогональных сигнатур, 498,257 различных SV
- После фильтрации: 382,460 уникальных, полностью определенных SV из 12,549 неродственных геномов

SV на геном:

- 1000 геномов: 3,441
- GTEx project: 3,658
- **gnomAD-SV: 8,202**
- WGS на длинных ридах: 24,825



Collins *biorXiv*
http://dx.doi.org/10.1101/578674

gnomAD: структурные варианты в 14,891 геномах

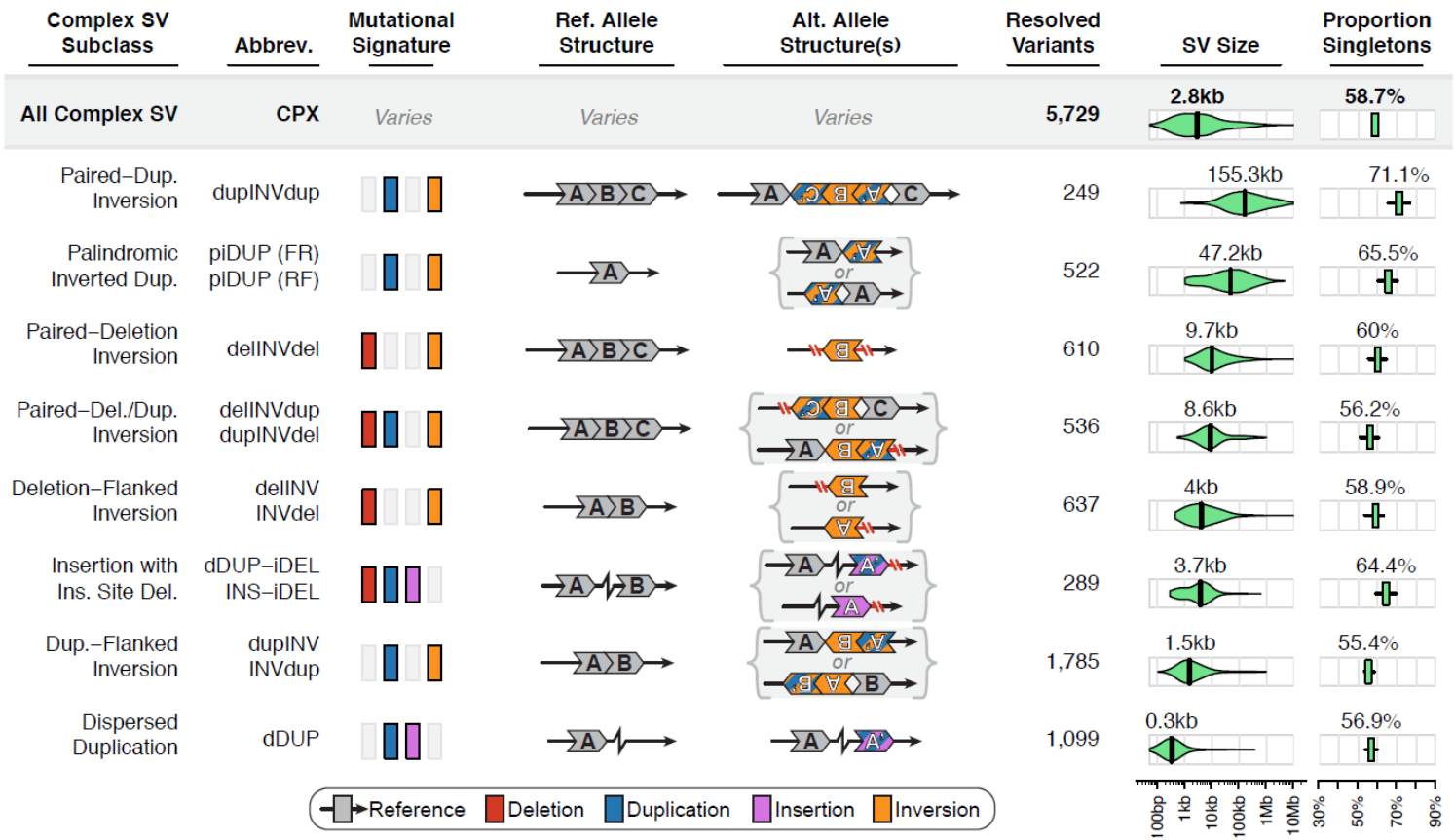
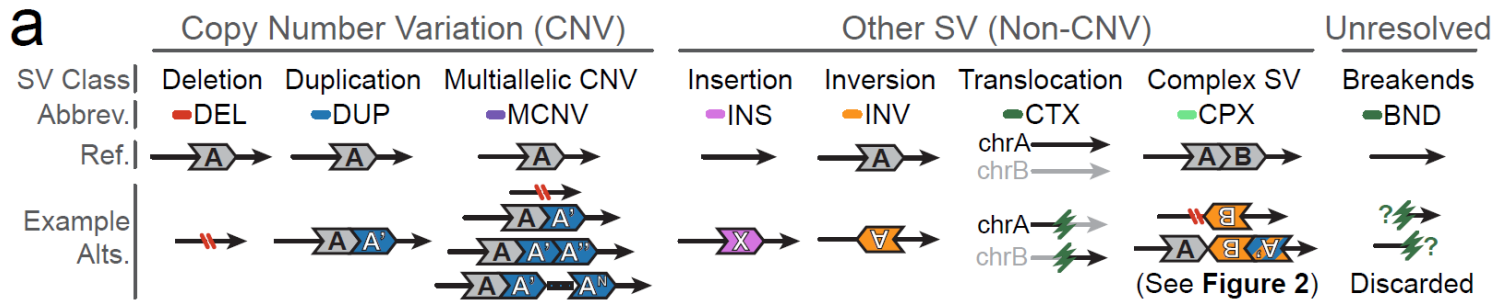


Figure 2 | Complex SVs are abundant in the human genome

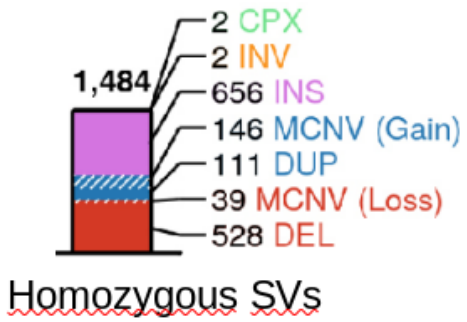
Collins *biorXiv* `http://dx.doi.org/10.1101/578674`

gnomAD: структурные варианты в 14,891 геномах

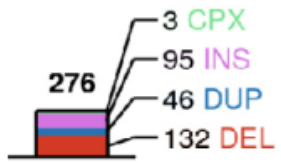


В среднем геноме: **8,202 SV**:

- малые (медианный размер SV=374 п.н.)
- ... и редкие (92% с AF<1%)
- 46.4% синглетоны
- Восемь генов изменяются редкими SV
- Большие (более 1Мб), редкие аутосомные SV в 3.1% геномов



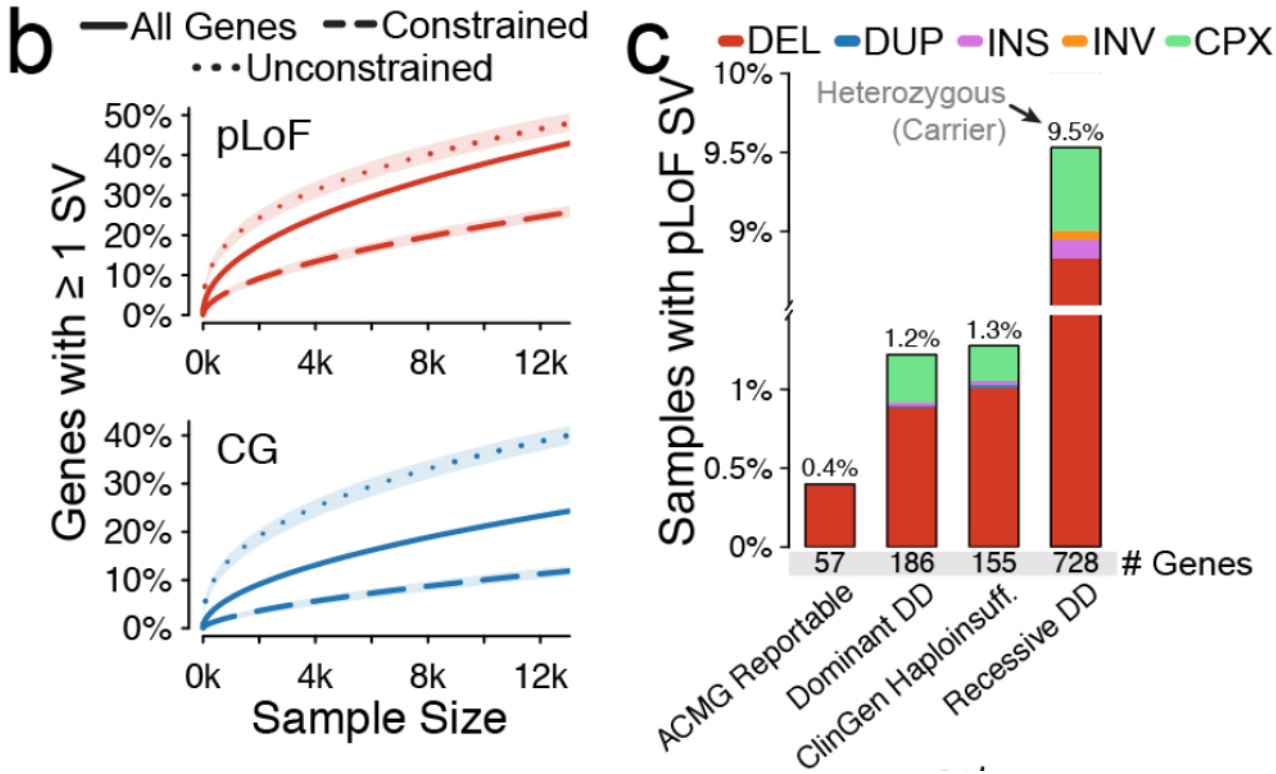
Homozygous SVs



Rare SVs

Collins *biorXiv*
http://dx.doi.org/10.1101/578674

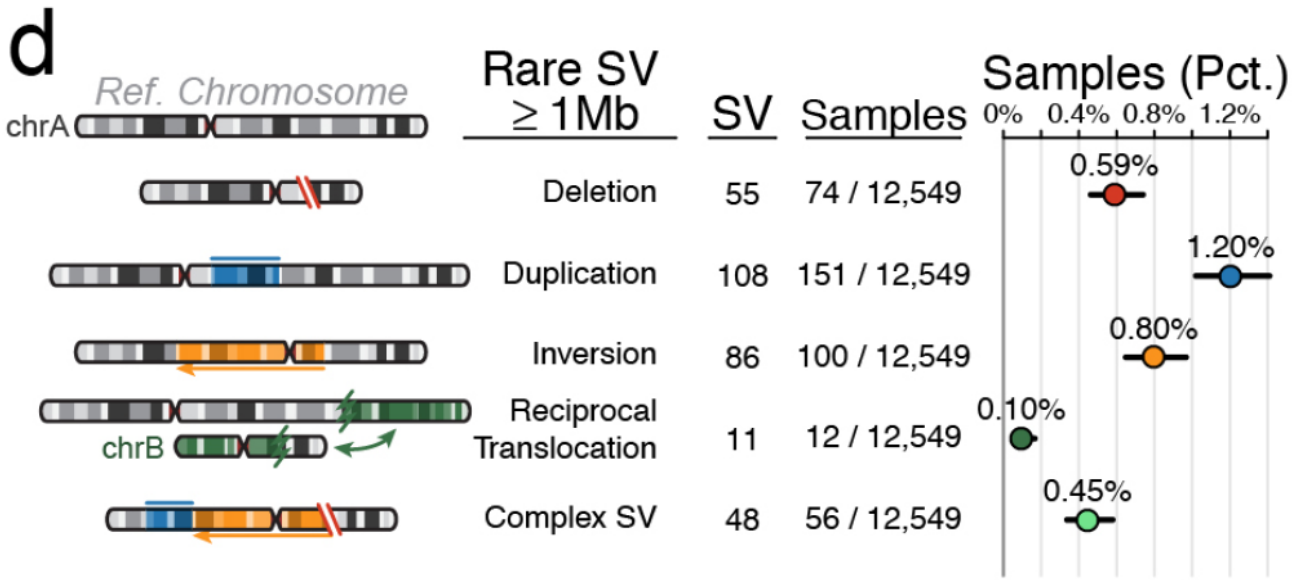
gnomAD: структурные варианты в 14,891 геномах



Collins *biorXiv* `http://dx.doi.org/10.1101/578674`

(b) At least one pLoF or CG SV was detected in 40.4% and 23.5% of all autosomal genes, respectively. (c) Up to 1.3% of genomes in gnomAD-SV harbored a very rare (AF<0.1%) pLoF SV in a medically relevant gene across several gene lists.

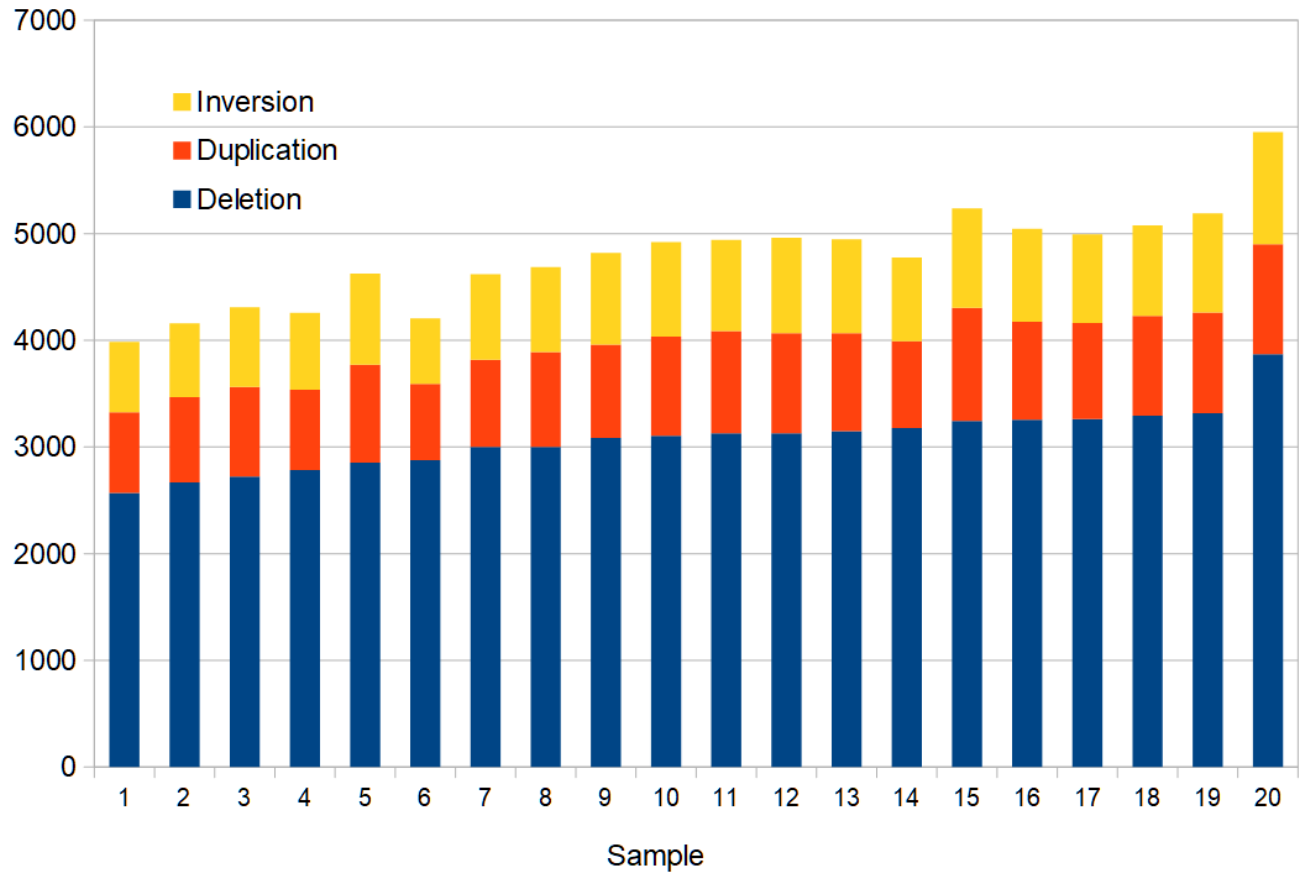
gnomAD: структурные варианты в 14,891 геномах



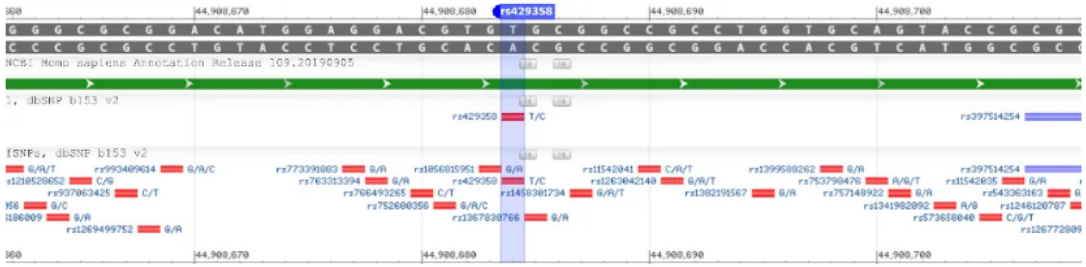
Collins *biorXiv* `http://dx.doi.org/10.1101/578674`

(d) We found **308 rare autosomal SVs** $\geq$ **1Mb**, revealing that $\sim$3.1% of genomes carry a large, rare chromosomal abnormality.

Структурные варианты в 20 геномах, согласно Delly



dbSNP: главный и открытый архив вариантов генома



NCBI Variation Summary

Description:

Summary of human variation data available from dbSNP and dbVar.

Report date: Tuesday, April 21, 2020

Total Variants:

- SubSNP count: 1,803,563,957
- RefSNP count: 660,773,127
- Variant Call count: 36,118,602
- Variant Region count: 6,023,949

dbVar is NCBI's database of human genomic Structural Variation – large variants >50 bp including insertions, deletions, duplications, inversions, mobile elements, translocations, and complex variants

Organism	Common Name	Taxon ID	dbSNP	dbVar
Homo sapiens	human	9606	**Last Updated**: Build 151 (Mar 22, 2018) **RefSNP Count**: 660.8 Million **SubSNP Count**: 1803.6 Million **Assembly**: GRCh37.p13, GRCh38.p7 **Data**: Search, FTP **Genome Data Viewer**: GRCh37.p13, GRCh38.p7	**Last Updated**: Apr 19, 2020 **Variant Regions**: 6 Million **Variant Calls**: 35.9 Million **Assembly**: GRCh37, GRCh37.p13, GRCh38, GRCh38.p12, GRCh38.p13, GRCh38.p9, NCBI36 **Data**: Search, FTP **dbVar Browser**: GRCh37, GRCh38, NCBI34, NCBI35, NCBI36 **Genome Data Viewer**: GRCh37, GRCh38

Проект «Геномы России»

Original Article

Genome-wide sequence analyses of ethnic populations across Russia

Daria V. Zhernakova[a,b,*], Vladimir Brukhin[a], Sergey Malov[a,c], Taras K. Oleksyk[a,d,r], Klaus Peter Koepfli[a,e], Anna Zhuk[a,f], Pavel Dobrynin[a,e], Sergei Kliver[a], Nikolay Cherkasov[a], Gaik Tamazian[a], Mikhail Rotkevich[a], Ksenia Krasheninnikova[a], Igor Evsyukov[a], Sviatoslav Sidorov[a], Anna Gorbunova[a,g], Ekaterina Chernyaeva[a], Andrey Shevchenko[a], Sofia Kolchanova[a,d], Alexei Komissarov[a], Serguei Simonov[a], Alexey Antonik[a], Anton Logachev[a], Dmitrii E. Polev[h], Olga A. Pavlova[h], Andrey S. Glotov[u], Vladimir Ulantsev[i], Ekaterina Noskova[i,j], Tatyana K. Davydova[s], Tatyana M. Sivtseva[k], Svetlana Limborska[l], Oleg Balanovsky[m,n,o], Vladimir Osakovsky[k], Alexey Novozhilov[p], Valery Puzyrev[q], Stephen J. O'Brien[a,t,*]

Zhernakova (2019) *Genomics*

Российская Федерация является **самой большой и одной из самых этнически разнообразных** стран в мире, однако до сих пор не существует централизованной базы данных генетической вариации. Такие данные крайне важны для медицинской генетики и незаменимы для изучения истории популяции.
Проект геном России нацелен на решение этой проблемы путем проведения полногеномного секвенирования и анализа жителей Российской Федерации. В данной статье мы представляем характеристику полногеномной вариации у **264 здоровых взрослых участников**, включая 60 новых образцов. Жители России являются носителями известных и новых вариантов с адаптивными, клиническими и функциональными последствиями, которые, во многих случаях, имеют отличия в частотах аллелей от соседних популяций.

Проект «Геномы России»

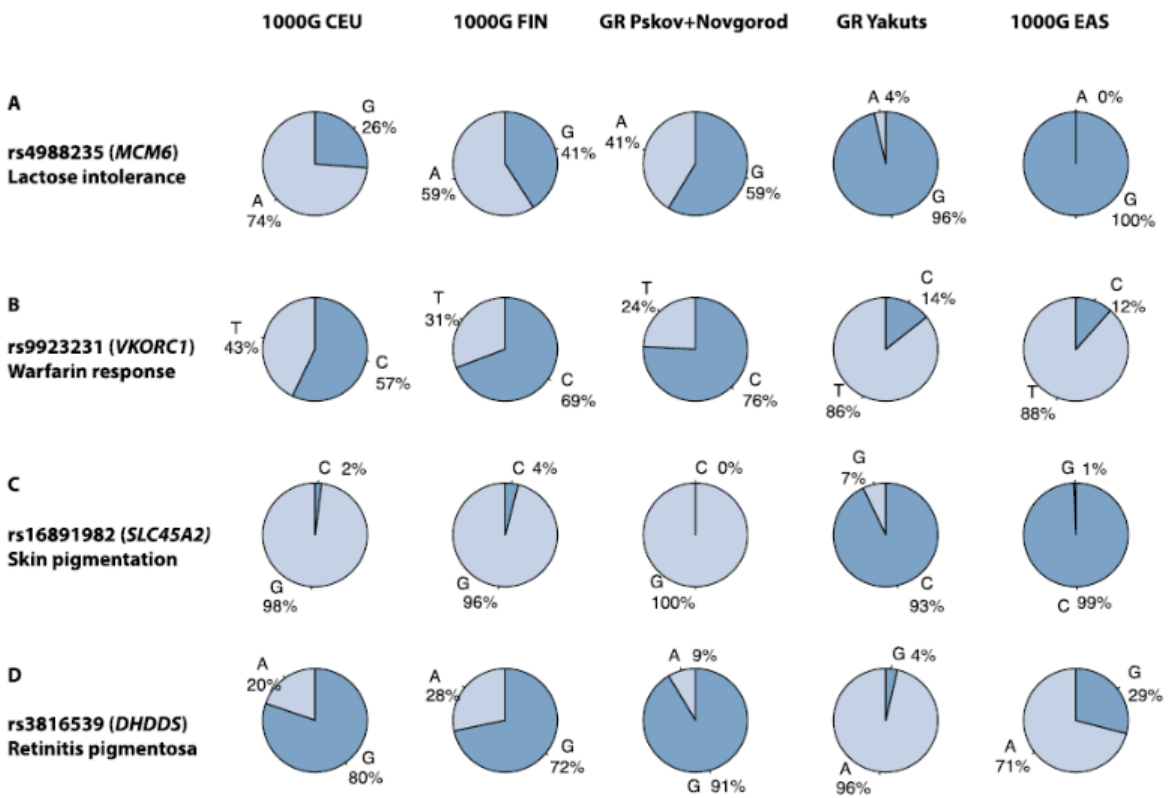


Fig. 3. Differences in Genome Russia allele frequencies of SNPs in notable genes with important phenotypes differentiate among Eurasian ethic groups. Allele frequencies for populations of Pskov and Novgorod (combined) and Yakut are shown together with allele frequencies of 1000G populations: Europeans (CEU), Finnish (FIN), East Asians (EAS) and South Asians (SAS) for four SNPs: (a) rs4988235, located in *MCM6* gene. This SNP is associated with adult type lactose intolerance. G allele tags the lactose intolerant haplotype [58,59]; (b) rs9923231, located in *VKORC1* gene. This SNP is associated with Warfarin response. T allele carriers need reduced dose of warfarin; (c) rs16891982 located in *SLC45A2* gene. G allele related to lighter skin pigmentation; (d) rs3816539 located in DHDDS gene. A allele is associated with retinitis pigmentosa.

Zhernakova (2019) *Genomics*

Популяция Ивановской популяции: 242 гена, 1,685 образцов

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 October 2021
doi: 10.3389/fgene.2021.709419

Targeted Sequencing of 242 Clinically Important Genes in the Russian Population From the Ivanovo Region

Vasily E. Ramensky[1,2], Alexandra I. Ershova[1], Marija Zaicenoka[3], Anna V. Kiseleva[1], Anastasia A. Zharikova[1,2], Yuri V. Vyatkin[1,4], Evgeniia A. Sotnikova[1], Irina A. Efimova[1], Mikhail G. Divashuk[1,5], Olga V. Kurilova[1], Olga P. Skirko[1], Galina A. Muromtseva[1], Olga A. Belova[6], Svetlana A. Rachkova[6], Maria S. Pokrovskaya[1], Svetlana A. Shalnova[1], Alexey N. Meshkov[1†] and Oxana M. Drapkina[1†]*

[1] National Medical Research Center for Therapy and Preventive Medicine, Moscow, Russia, [2] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [3] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow, Russia, [4] Novosibirsk State University, Novosibirsk, Russia, [5] All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia, [6] Cardiology Dispensary, Ivanovo, Russia

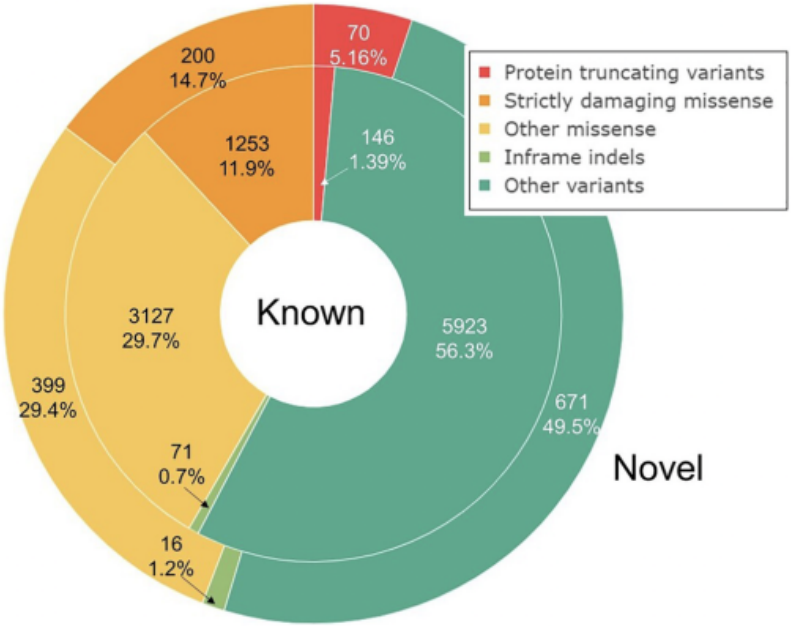
OPEN ACCESS

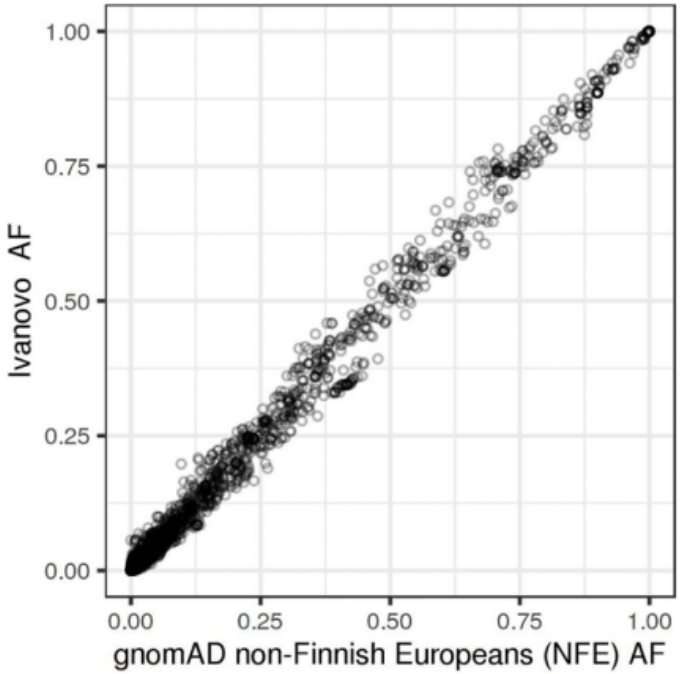
Edited by:
Tatiana V. Tatarinova,
University of La Verne, United States

Популяция Ивановской популяции: 242 гена, 1,685 образцов

	Rare, AF<0.1%		Common, AF≥0.1%	
	Known	Novel (Not in NWR)	Known	Novel (Not in NWR)
Protein truncating variants	112	70 (69)	34	2 (2)
Strictly damaging missense variants	907	193 (190)	346	7 (5)
Other missense	1957	395 (379)	1170	4 (4)
Inframe indels	49	15 (15)	22	1 (1)
Other variants	3227	657 (635)	2696	14 (3)
Total	6252	1330	4268	28

Популяция Ивановской популяции: 242 гена, 1,685 образцов





Популяция Ивановской популяции: 242 гена, 1,685 образцов

Gene	Disease	Variant	HGVS	gnomAD	Ivanovo AC	Ivanovo AF	Ivanovo/ gnomAD
KCNQ1	Long QT syndrome (AD, OMIM:192500)	rs1337409061	ENSP00000155840.2:p.Thr96Arg	3.459E-05	3	0.00089	25.7
MYBPC3	Hypertrophic cardiomyopathy (AD, OMIM:115197)	rs376395543	ENST00000545968.1:c.26-2A>G	5.1837E-05	3	0.00089	17.2
GAA	Glycogen storage disease (Pompe disease) (AR, OMIM:232300)	rs375470378	ENST00000302262.3:c.1552-3C>G	0.0002713	8	0.00237	8.8
GLB1	GM1-gangliosidosis (AR, OMIM:253010, 230600)	rs376663785	ENSP00000306920.4:p.Tyr270Asp	4.6641E-05	4	0.00119	25.4
LAMA2	Merosin-deficient congenital muscular dystrophy type 1A (AR, OMIM:607855)	rs398123387	ENST00000421865.2:c.7536del	1.7651E-05	4	0.00119	67.2
MTO1	Combined oxidative phosphorylation deficiency (AR, OMIM:614702)	rs201544686	ENSP00000402038.2:p.Arg517His	0.0002322	6	0.00178	7.7
SCO2	Mitochondrial complex IV deficiency (AR, OMIM:604377)	rs74315511	ENSP00000444433.1:p.Glu140Lys	0.0001784	4	0.00119	6.7
SURF1	Mitochondrial complex IV deficiency, Leigh syndrome (AR, OMIM:220110)	rs782316919	ENST00000371974.3:c.845_846del	0.0001476	4	0.00119	8.0
ALMS1	Alstrom syndrome (AR, OMIM:203800)	rs797045228	ENST00000264448.6:c.4150dup	4.675E-05	3	0.00089	19.0
ALMS1	Alstrom syndrome (AR, OMIM:203800)	rs747272625	ENST00000264448.6:c.11310_11313	5.34E-05	3	0.00089	16.7

Известные болезнетворные варианты, которые значимо более частые в Ивановской популяции

APOB и гипобеталипопротеинемия

HGNC Approved Gene Symbol: APOB

Cytogenetic location: 2p24.1 *Genomic coordinates (GRCh38): 2:21,001,428-21,044,072* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype Clinical Synopses	Phenotype MIM number	Inheritance	Phenotype mapping key
2p24.1	Hypercholesterolemia, familial, 2	144010	AD	3
	Hypobetalipoproteinemia	615558	AR	3

Hypobetalipoproteinemia (FHBL) and abetalipoproteinemia (ABL; 200100) are rare diseases characterized by hypocholesterolemia and malabsorption of lipid-soluble vitamins leading to retinal degeneration, neuropathy, and coagulopathy. Hepatic steatosis is also common. The root cause of both disorders is improper packaging and secretion of apolipoprotein B-containing particles.

As indicated in the listing of allelic variants, a number of mutations resulting in a truncated apolipoprotein B have been found as the basis of hypobetalipoproteinemia. Other patients with this disorder have been found to have reduced concentrations of a full-length apoB100 (Young et al., 1987; Berger et al., 1983; Gavish et al., 1989).

APOB и гипобеталипопротеинемия

Gene	ACMG	Variant	HGVS	Phenotype (Source)
II. Novel protein truncating: 27 variants, 27 carriers				
APOB	Yes	chr2:21232683_G/A	ENSP00000233242.1: p.Gln2353Ter	Hypobetalipoproteinemia, LDL-C=1.47 mmol/l (Biochemical assay)
APOB	Yes	chr2:21234967_GA/G	ENSP00000233242.1: p.Phe1591SerfsTer19	Hypobetalipoproteinemia, LDL-C=0.95 mmol/l (Biochemical assay)
APOB	Yes	chr2:21260870_AC/A	ENSP00000233242.1: p.Val166PhefsTer66	Hypobetalipoproteinemia, LDL-C=0.72 mmol/l (Biochemical assay)
MYH7	Yes	chr14:23889261_CT/C	ENSP00000347507.3: p.Lys1173ArgfsTer41	Hypertrophic cardiomyopathy (Medical record)

Table 6 Variants with confirmed phenotypes. **Variant**: dbSNP rsID for known variants or chr:pos_ref/alt identifier for novel PTVs. **HGVS**: variant description. **Phenotype**: disease phenotype confirmed by evaluation of clinical data; source of clinical data is specified in the parentheses.

RUSeq

Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples

Yury A. Barbitoff[1,3,4,✉], Darya N. Khmelkova[2], Ekaterina A. Pomerantseva[2], Aleksandr V. Slepchenkov[3], Nikita A. Zubashenko[2], Irina V. Mironova[2], Vladimir S. Kaimonov[2], Dmitrii E. Polev[1], Victoria V. Tsay[1,5], Andrey S. Glotov[1,4], Mikhail V. Aseev[1,4], Oleg S. Glotov[1,4,5], Arthur A. Isaev[2], and Alexander V. Predeus[3,✉]

1. We construct an expanded reference set of genetic variants by analyzing **6,096 exome samples** collected in two major Russian cities of Moscow and St. Petersburg.

2. An approximately tenfold increase in sample size compared to previous studies allowed us to identify genetically **distinct clusters of individuals within an admixed population** of Russia.

3. We show that **up to 18 known pathogenic variants are overrepresented in Russia** compared to other European countries.

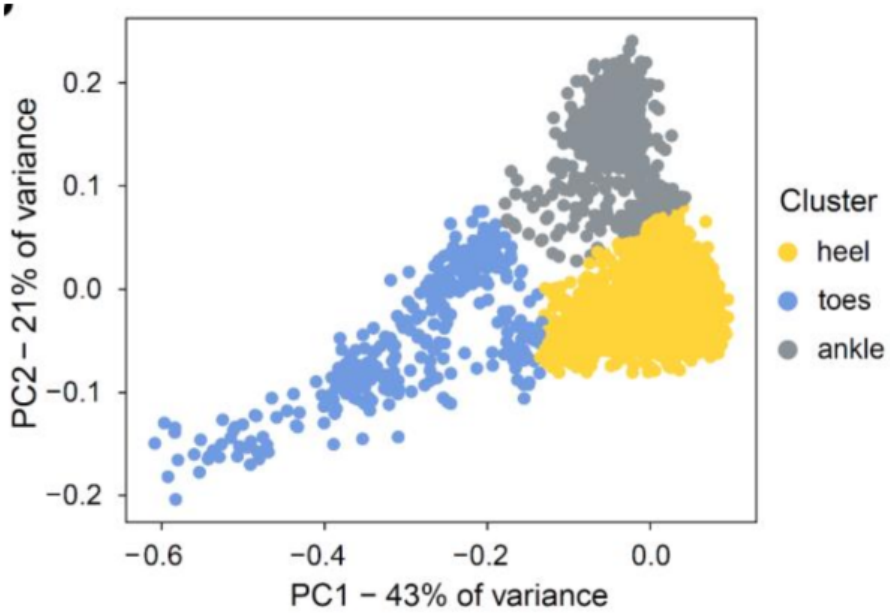
4. We also identify several dozen high-impact **variants that are present in healthy donors** despite either being annotated as pathogenic in ClinVar or falling within genes associated with autosomal dominant disorders.

5. **The constructed database of genetic variant frequencies in Russia** has been made available to the medical genetics community through a variant browser available at http://ruseq.ru

RUSeq

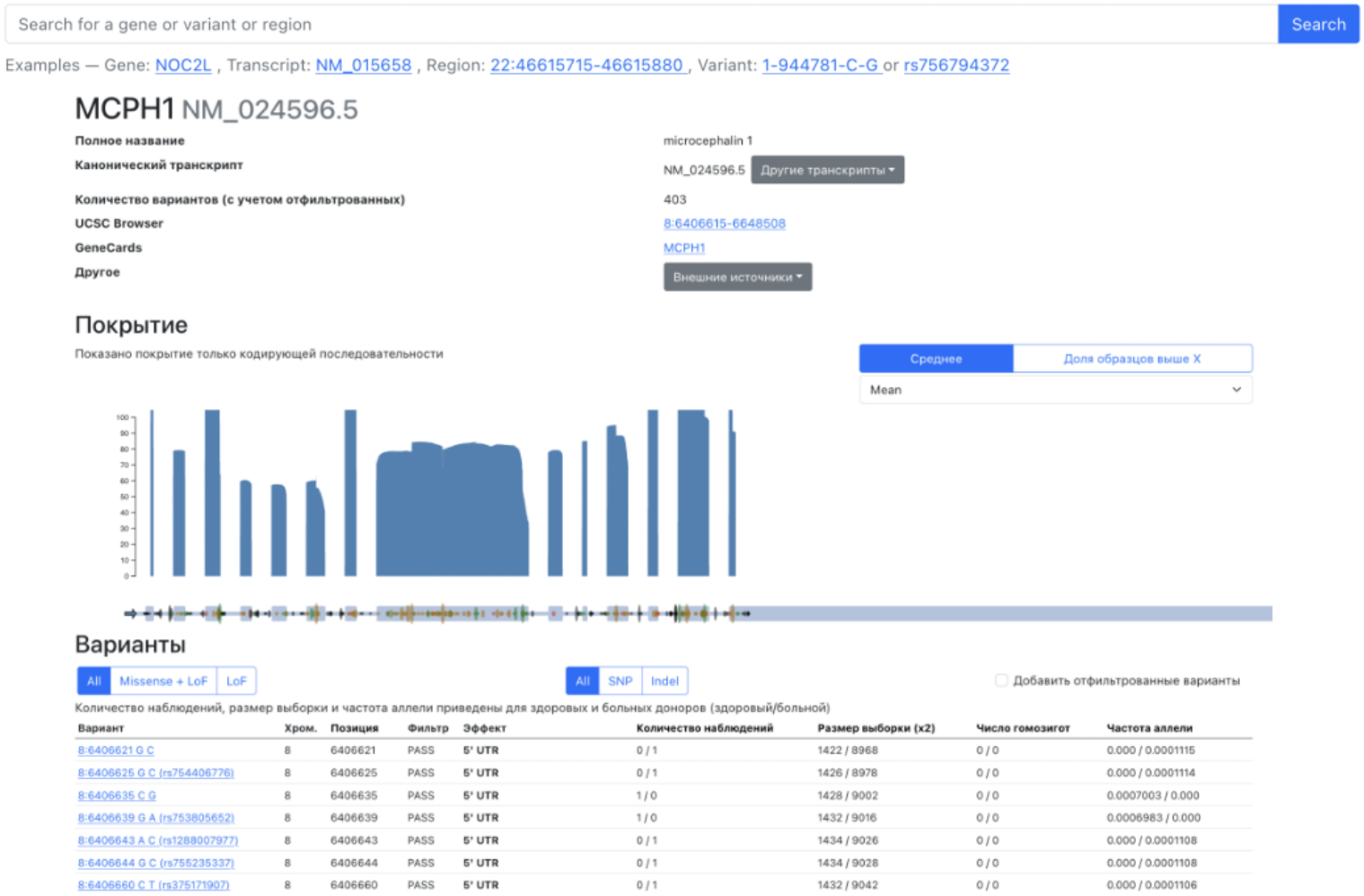
Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples

Yury A. Barbitoff[1,3,4,✉], Darya N. Khmelkova[2], Ekaterina A. Pomerantseva[2], Aleksandr V. Slepchenkov[3], Nikita A. Zubashenko[2], Irina V. Mironova[2], Vladimir S. Kaimonov[2], Dmitrii E. Polev[1], Victoria V. Tsay[1,5], Andrey S. Glotov[1,4], Mikhail V. Aseev[1,4], Oleg S. Glotov[1,4,5], Arthur A. Isaev[2], and Alexander V. Predeus[3,✉]



We identified several genetically distinct clusters of the study participants. **Yellow:** most likely represents European part of Russia; **gray:** represents Caucasus; **blue:** unites diverse samples from East part of Russia (e.g., originating from Syberia, the “Far East”, etc.). Variant frequencies at this website are provided for all three clusters.

RUSeq


Search for a gene or variant or region
Search
Examples — Gene: NOC2L , Transcript: NM_015658 , Region: 22:46615715-46615880 , Variant: 1-944781-C-G or rs756794372
MCPH1 NM_024596.5
Полное название
microcephalin 1
Канонический транскрипт
NM_024596.5
Другие транскрипты
Количество вариантов (с учетом отфильтрованных)
403
UCSC Browser
8:6406615-6648508
GeneCards
MCPH1
Другое
Внешние источники
Покрытие
Показано покрытие только кодирующей последовательности
Среднее
Доля образцов выше X
Mean
Варианты
All
Missense + LoF
LoF
All
SNP
Indel
Добавить отфильтрованные варианты
Количество наблюдений, размер выборки и частота аллели приведены для здоровых и больных доноров (здоровый/больной)
Вариант
Хром.
Позиция
Фильтр
Эффект
Количество наблюдений
Размер выборки (x2)
Число гомозигот
Частота аллели
8:6406621 G C
8 6406621 PASS 5' UTR 0 / 1 1422 / 8968 0 / 0 0.000 / 0.0001115
8:6406625 G C (rs754406776)
8 6406625 PASS 5' UTR 0 / 1 1426 / 8978 0 / 0 0.000 / 0.0001114
8:6406635 C G
8 6406635 PASS 5' UTR 1 / 0 1428 / 9002 0 / 0 0.0007003 / 0.000
8:6406639 G A (rs753805652)
8 6406639 PASS 5' UTR 1 / 0 1432 / 9016 0 / 0 0.0006983 / 0.000
8:6406643 A C (rs1288007977)
8 6406643 PASS 5' UTR 0 / 1 1434 / 9026 0 / 0 0.000 / 0.0001108
8:6406644 G C (rs755235337)
8 6406644 PASS 5' UTR 0 / 1 1434 / 9028 0 / 0 0.000 / 0.0001108
8:6406660 C T (rs375171907)
8 6406660 PASS 5' UTR 0 / 1 1432 / 9042 0 / 0 0.000 / 0.0001106

Уроки, полученные из секвенирования

- PCA может показать локальные субпопуляции, частоты вариантов могут различаться
- RUSeq: объединяет генетическую информацию между клиническими лабораториями и геномными центрами в России
- Примерно 10% вариантов являются новыми, они обогащены вариантами с большим влиянием (PTV, миссенсы)
- Наблюдается перепредставленность некоторых известных болезнетворных вариантов
- Известные и возможные болезнетворные варианты обнаруживаются у здоровых доноров
- Открываются новые и подтверждаются известные варианты, связанные с фенотипами
- Нужно различать здоровых доноров и пациентов при оценке частоты вариантов

Выводы

- Ранние оценки разнообразия нуклеотидов не учитывали очень быстрый рост человеческой популяции и естественный отбор. Поэтому фактическая оценка разнообразия гораздо выше; также наблюдается избыток редких аллелей.
- Недавние крупномасштабные проекты по секвенированию (1000 геномов, ExAC, gnomAD, UK Bioibank) проливают свет на ранее неизвестные паттерны вариабельности в геноме человека и дают новый взгляд на человеческую популяцию и генетику заболеваний.
- В частности, варианты с популяционной частотой, несовместимой с рецессивной наследуемостью и раннее считавшиеся болезнетворными, переклассифицируются.
- Накопление образцов делает возможным делать оценки на уровне генов, например, меры нетерпимости генов к pLoF-вариантам или коэффициенты отбора для них.
- Появляются исследования по спектру вариантов в российской популяции.

Список литературы

- Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291
- Cassa, C.A., Weghorn, D., Balick, D.J., Jordan, D.M., et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat. Genet. 49, 806–810
- Saleheen, D., Natarajan, P., Armean, I.M., Zhao, W., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 544, 235–239
- Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 531210
- Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., et al. (2019). An open resource of structural variation for medical and population genetics. BioRxiv 578674
- Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nature Genetics 44, 623–630

Список литературы

- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526, 68–74
- Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nature Reviews Genetics 18, 599
- Rehm, H.L., Berg, J.S., and Plon, S.E. (2018). ClinGen and ClinVar – Enabling Genomics in Precision Medicine. Human Mutation 39, 1473–1475
- Gao, F., and Keinan, A. (2016). Explosive genetic evidence for explosive human population growth. Current Opinion in Genetics & Development 41, 130–139
- Shah, N., Hou, Y.-C.C., Yu, H.-C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. The American Journal of Human Genetics 102, 609–619.
- Barbitoff, Y.A., et al. (2022). Expanding the Russian allele frequency reference via cross-laboratory data integration: insights from 6,096 exome samples. https://doi.org/10.1101/2021.11.02.21265801
- Zhernakova, D.V., Brukhin, V., Malov, S., Oleksyk, T.K., Koepfli, K.P., et al. (2019). Genome-wide sequence analyses of ethnic populations across Russia. Genomics.