Биоинформатические подходы к изучению микроэволюции

Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species?
EM Leffler et al., 2012
PLoS Biol. 2012;10(9):e1001388.

Щелелистник *Schizophyllum commune* – самый изменчивый из изученных видов, H = 0.14.

План

1. Что такое микроэволюция?

2. Что такое популяция?

3. Характеристики популяции – мутации, отбор, дрейф, размножение, структура.

4. Прямая и обратная задачи в изучении микроэволюции.

5. Биоинформатика и обратная задача - сравнение геномов и генотипов.

6. Краткая история изучения изменчивости популяций.

7. Изучение мутационного процесса – мама, папа и потомок.

8. Изучение дрейфа: эффективная численность популяции, $H = 4N_e\mu$.

9. Отрицательный, положительный и балансирующий отбор.

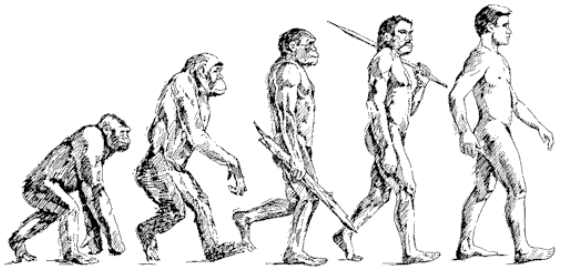
10. Изучение отбора.

11. Изучение размножения.

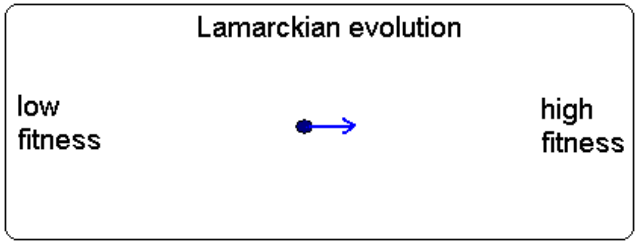
12. Изучение структуры.

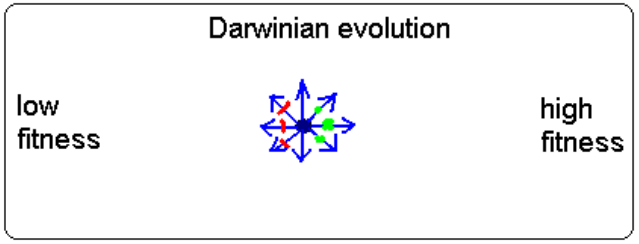
1. Что такое микроэволюция?

Макроэволюция – самое важное и очевидное.



Микроэволюция важна только потому, что механизмом дарвиновской эволюции является естественный отбор наследуемых изменений.





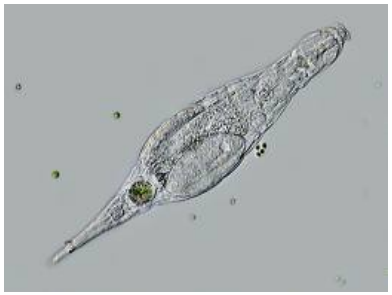
Ламарковская эволюция линии, состоящей, в каждый момент, из одной особи, была бы возможна, а дарвиновская – нет.

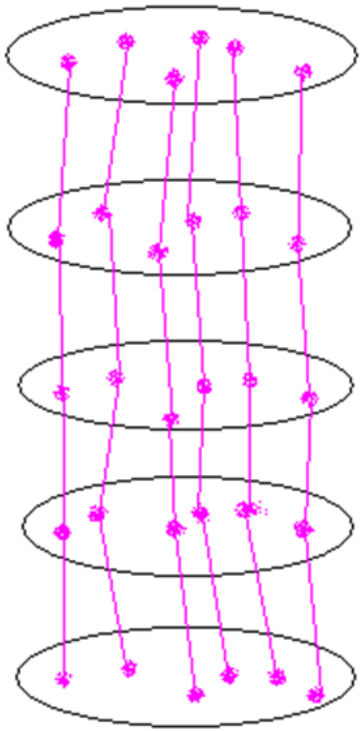
Микроэволюция = эволюция внутрипопуляционной изменчивости.

2. Что такое популяция?

Почему нельзя рассматривать линию каждой (бесполой) особи отдельно?

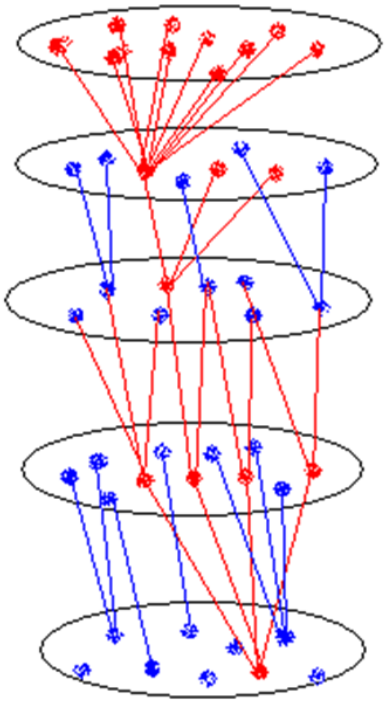
Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* (2013) doi:10.1038/nature12326



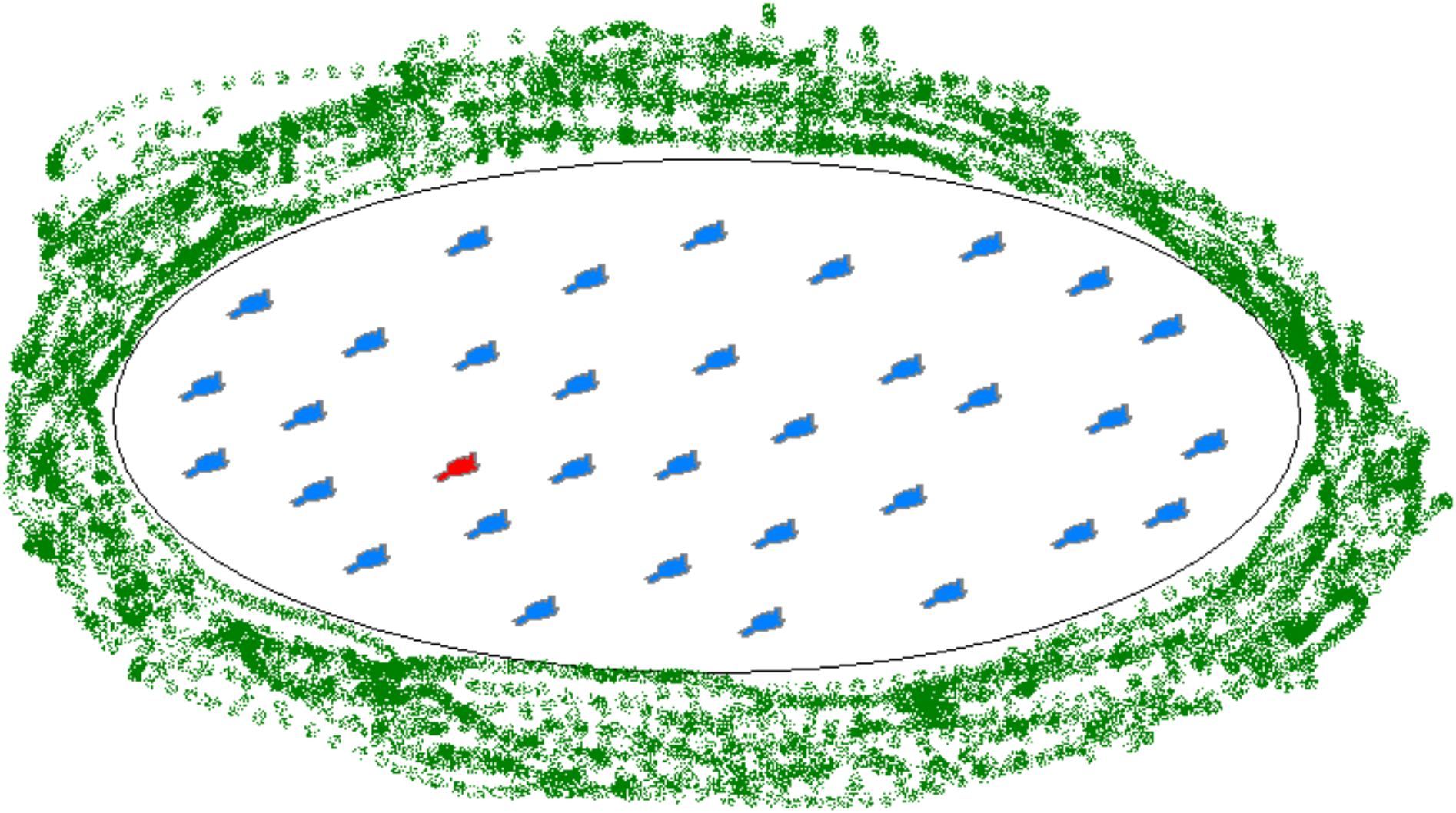


Не бывает.

Жизнь несправедлива!

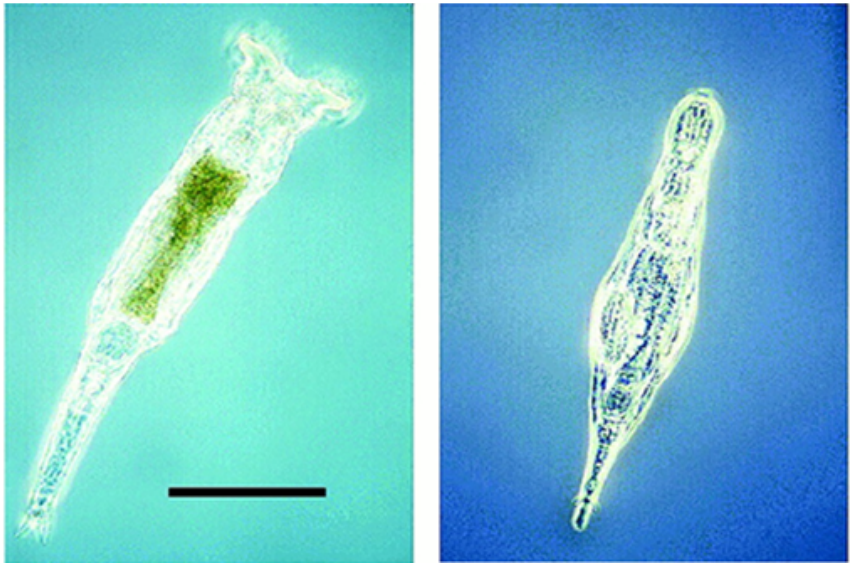


Только так.



Потомки красной останутся, всех остальных вымрут. Почему так? По двум причинам – из-за систематического преимущества (отбор) и игры случая (дрейф). Красная – хорошая, но не обязательно самая лучшая.

Бесполая популяция – множество особей, представляющих, в данный момент, минимальное долговечное множество линий, такое, что хоть одна из них будет жить долго.

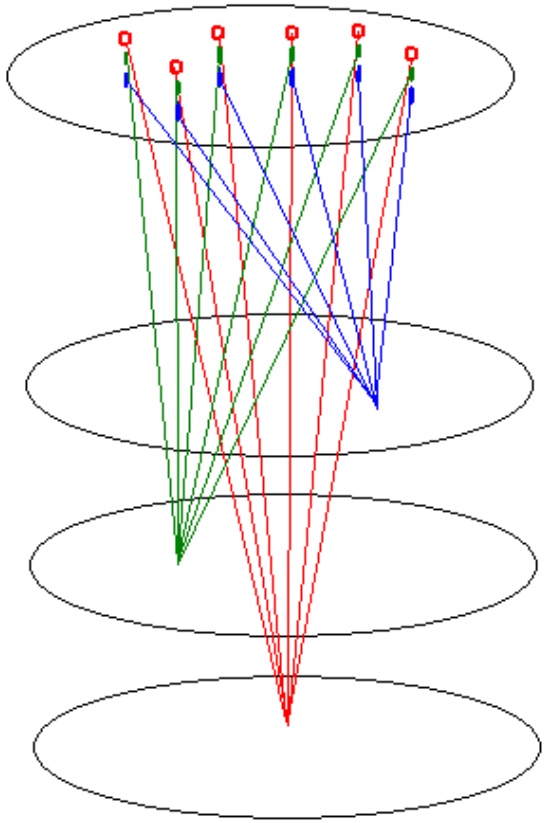


Philodina roseola ***Adineta vaga***

Филодина адинете глаз не выклюет (все наоборот!). Члены одной популяции экологически эквивалентны (принцип Гаузе) и играют в игру с нулевой суммой.

Популяция – множество особей, представляющих, в данный момент, все линии, которые не защищены друг от друга, так что размножение одной из них неизбежно повлечет вымирание остальных.

Размножение и вымирание линий происходит, конечно, и при половом размножении – но в этом случае, из-за рекомбинации, эти процессы происходят независимо в разных сегментах генома.



В современной популяции человека коалесцентные истории митохондриального генома (красный), (red), Y хромосомы (синий), и сегмента аутосомы (зеленый) различны.

Y-хромосомный "Адам", ~150 тысяч лет.

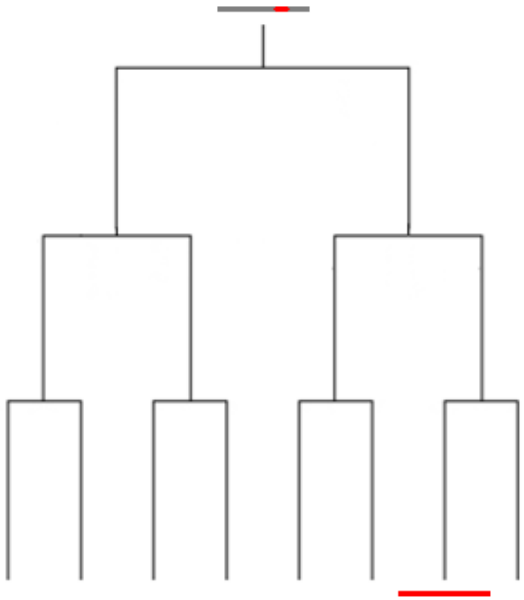
Общий предок аутосомного локуса – 100-400 тысяч лет.

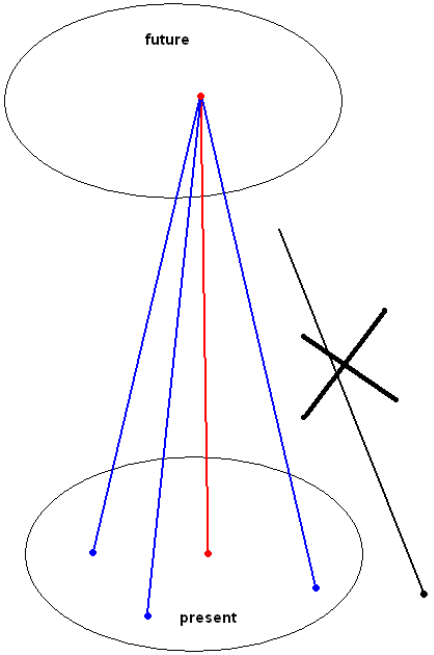
Митохондриальная "Ева" ~ 200 тысяч лет.

Митохондриальная Ева не была женой Y-хромосомного Адама!

Половая популяция – минимальное множество особей, такое, что в любом сегменте генома аллель, присутствующий хотя бы в одной из них, сохранится надолго.

Половое размножение добавляет еще одну причину, заставляющую нас рассматривать популяции вместо особей – перекрестное оплодотворение.





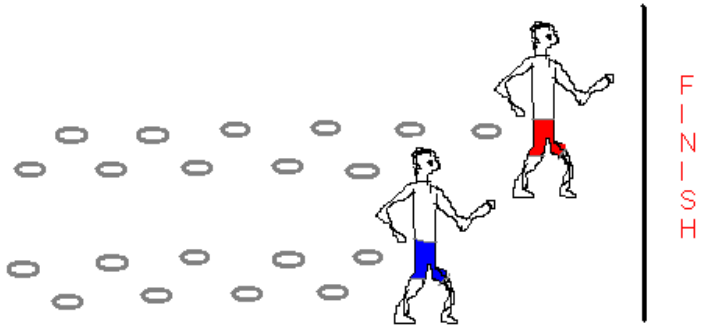
Половая популяция – минимальное репродуктивно замкнутое множество особей, такое, что в потомстве каждой из них будут присутствовать только аллели других членов популяции, но не посторонних.

Обычно эти два подхода не противоречат друг другу, хотя есть и исключения.

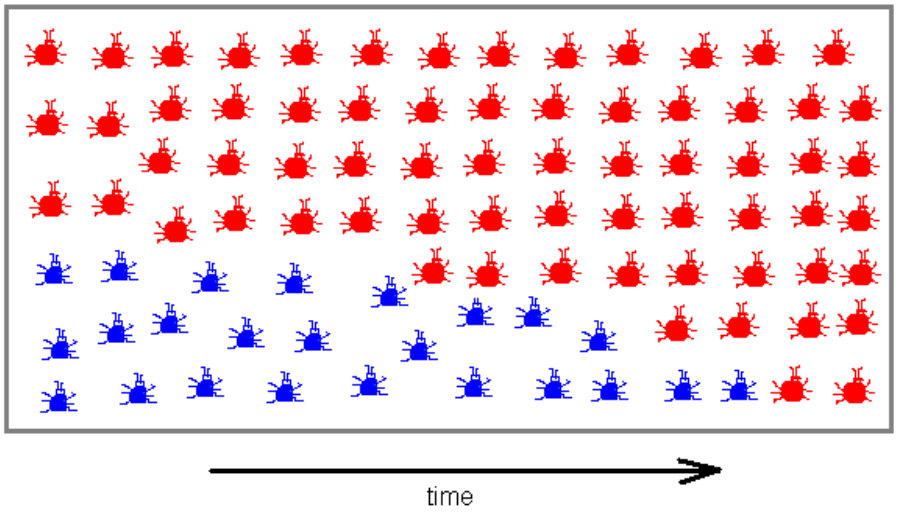
3. Характеристики популяции – мутации, отбор, дрейф, размножение, структура.

Мутационный процесс – случайные изменения последовательности ДНК. Замены, выпадения, вставки, сложные события. Скорость мутаций – вероятность изменения на нуклеотид за поколение.

Отбор – дифференциальное размножение генотипов. Кто бежит быстрее, придет к финишу первым – на уровне популяций все просто.

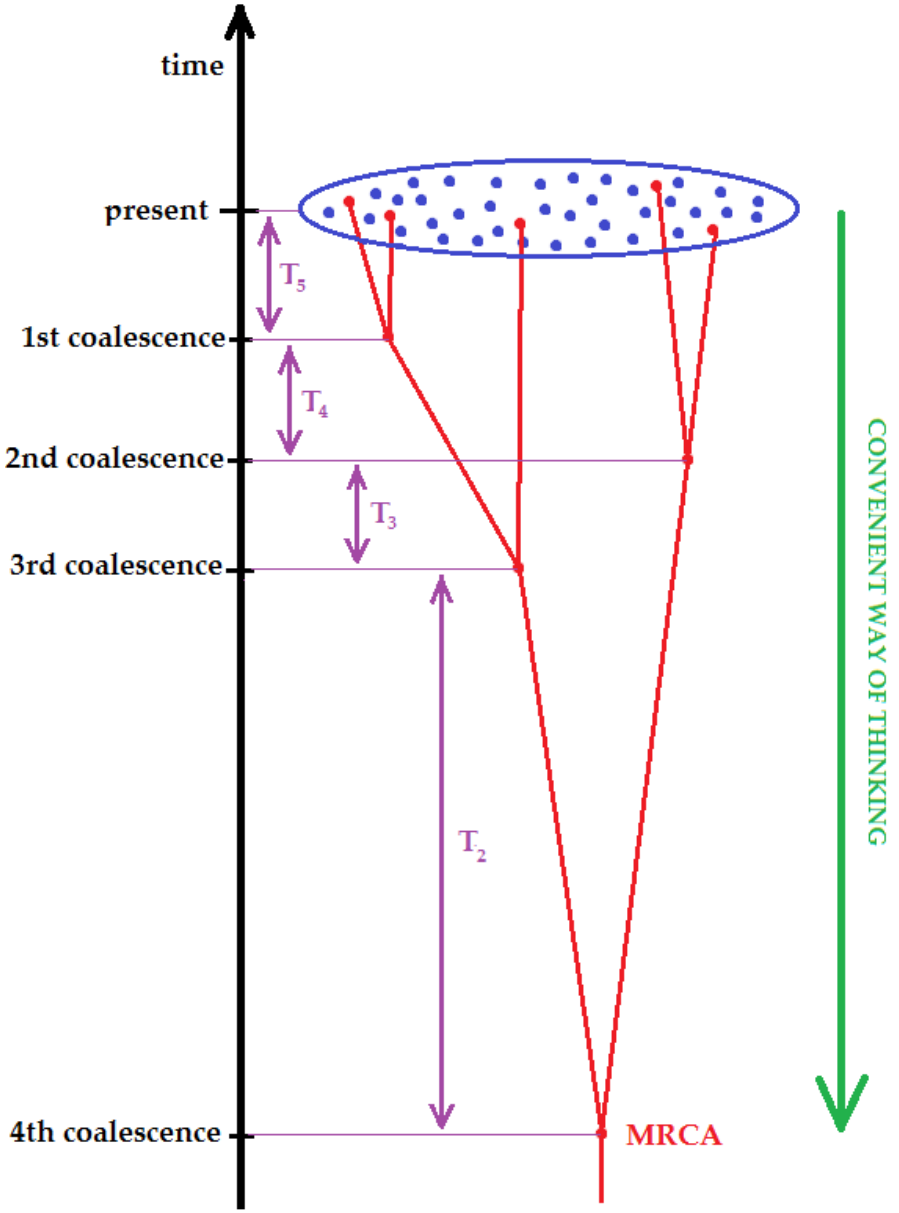


Приспособленность генотипа – ожидаемая эффективность размножения особи с таким генотипом.



Дрейф – влияние на генетический состав популяцию случайного процесса вымирания и размножения линий. Наиболее важный аспект дрейфа – случайные изменения частот аллелей в локусе.

Эффективная численность N_e популяции – численность эквивалентной Райт-Фишеровской популяции (в которой число потомков особи распределено по Пуассону).



Размножение может быть бесполым (апомиксис) и половым (амфимиксис).

Структура может быть возрастной и пространственной.

4. Прямая и обратная задачи в изучении микроэволюции.

Прямая задача – мы знаем все параметры популяции, что с ней будет?

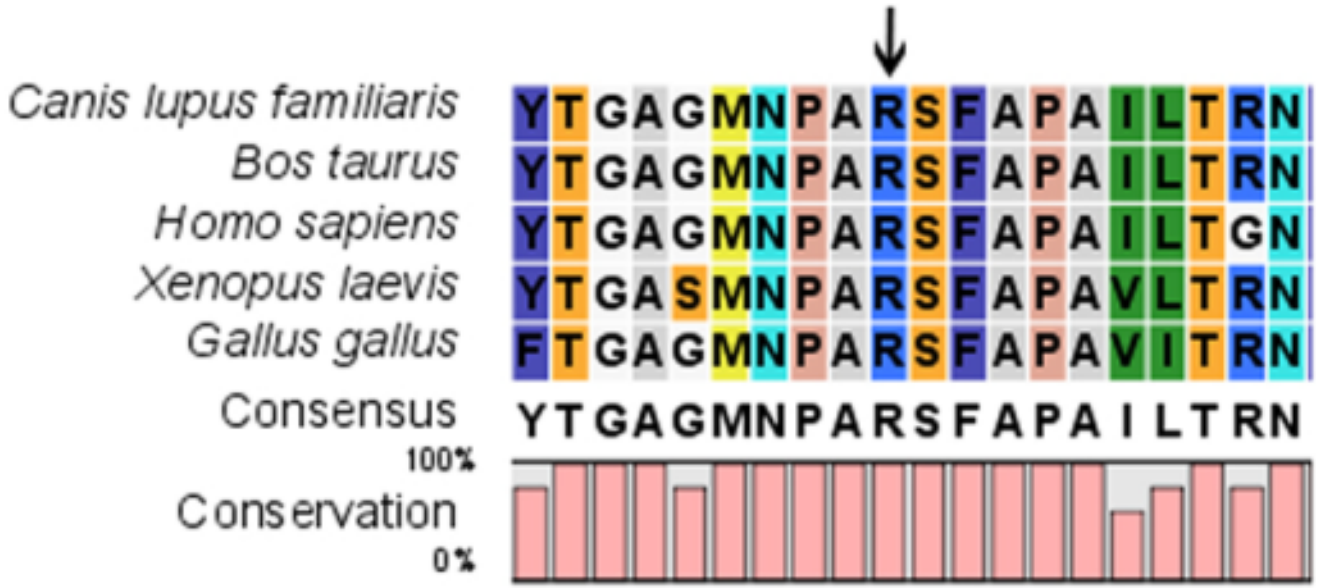
$$d[A]/dt = s[A](1-[A])$$

Обратная задача – мы знаем состояние популяции, каковы ее параметры?

$H = 0.01$, $\mu = 10^{-8}$. Поскольку $H = 4N_e\mu$, $N_e = 2.5 \times 10^5$.

5. Биоинформатика и обратная задача - сравнение геномов и генотипов.

Геном – у вида, а генотип (или два у диплоида) – у особи.



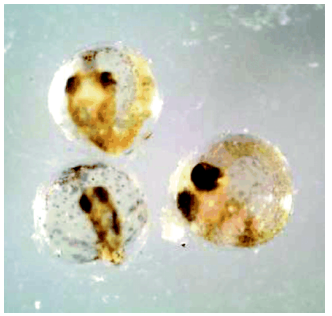
Сравнение с близким видом позволяет отличить производные аллели от анцестральных. Производные аллели обычно редки и вредны – хотя, конечно, бывают исключения (иначе не было бы эволюции).

6. Краткая история изучения изменчивости популяций.

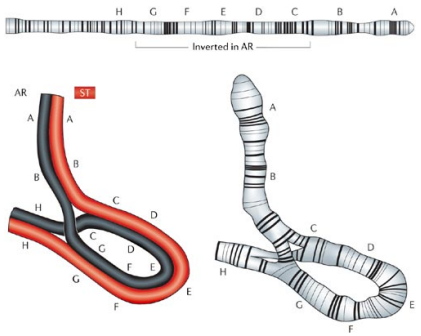
1909 Арчибальд Гаррод - «Врожденные ошибки метаболизма». Открыл редкие вредные рецессивные аллели, заметив, что алкаптонурия чаще встречается у потомков от браков между родственниками.

1926 Николай Владимирович и Елена Александровна Тимофеевы-Ресовские - Оценка числа видимых рецессивных мутаций на генотип – около 0.1.

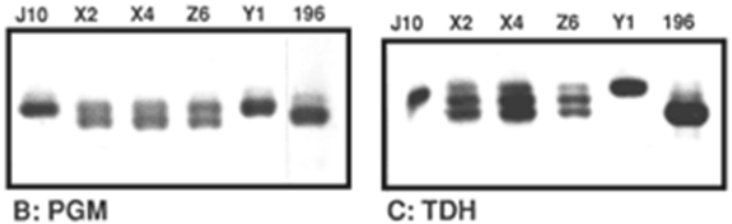
1928 Герман Меллер – Оценка числа рецессивных леталей на генотип – 1-2.



1938 Феодосий Григорьевич Добржанский и Альфред Стертевант - Полиморфизм популяций по инверсиям.



1966. Ричард Левонтин и др. Полиморфизм по изоферментам.



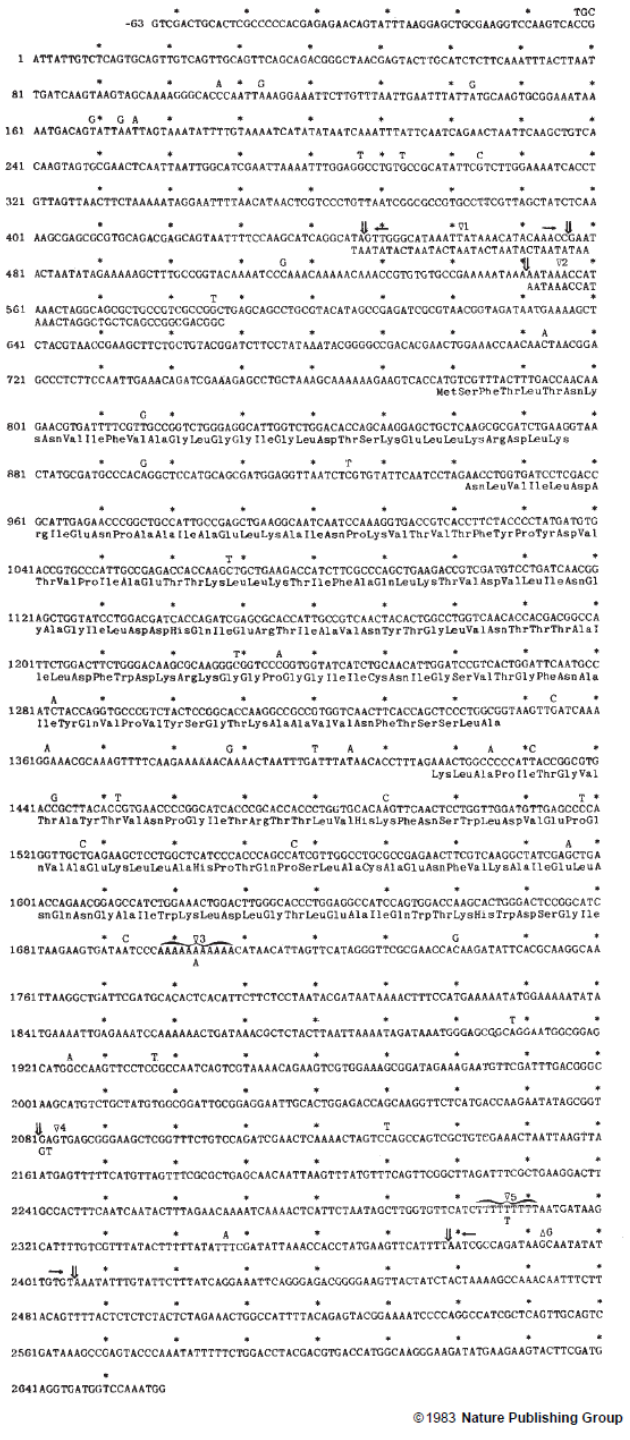
1983 – Начало новой зры – Мартин Крейтман описал полиморфизм на уровне ДНК. 11 аллелей алкогольдегидрогеназы дрозофилы, всего 43 отличия.

Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*

Martin Kreitman

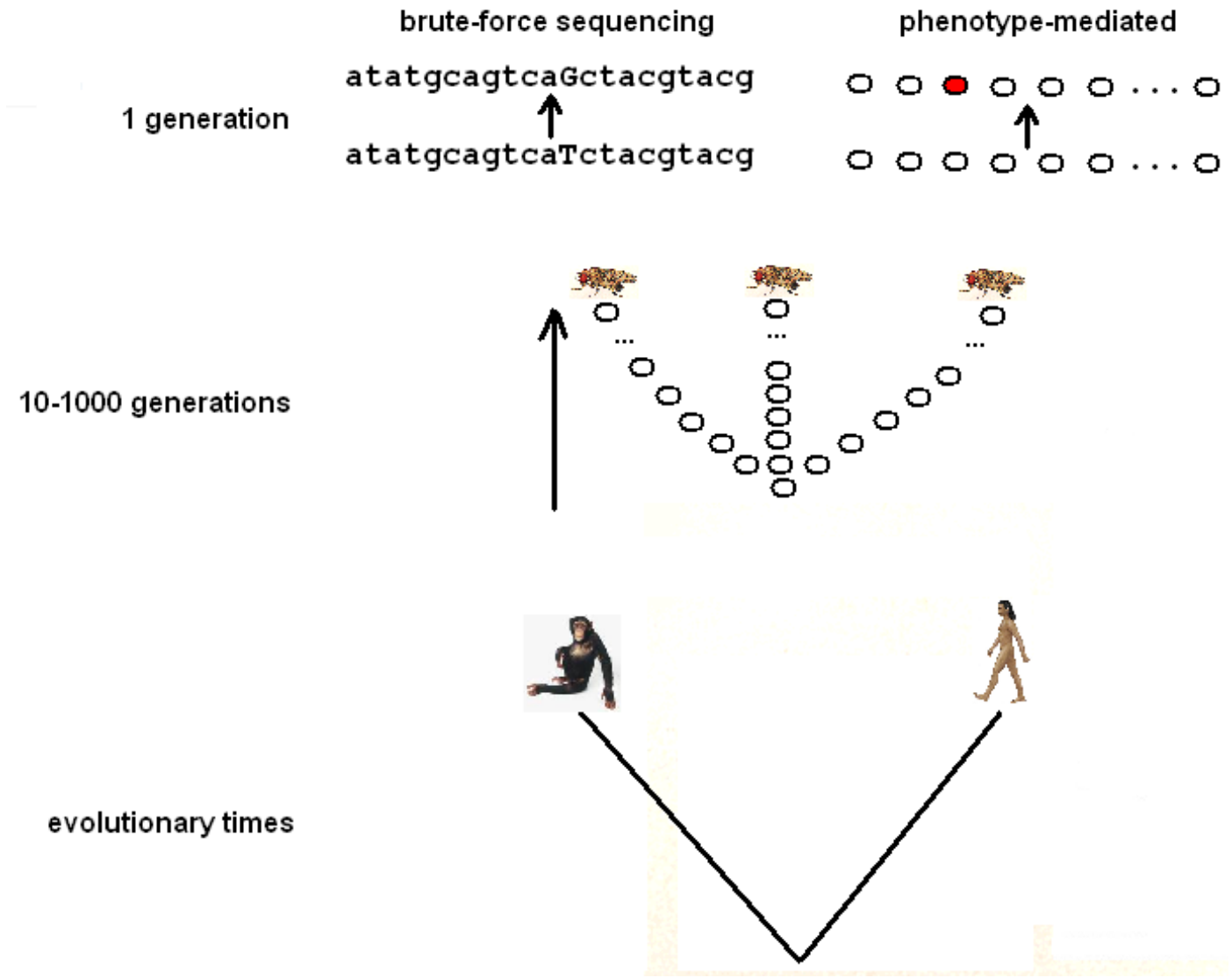
Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA

The sequencing of eleven cloned Drosophila melanogaster *alcohol dehydrogenase* (Adh) *genes from five natural populations has revealed a large number of previously hidden polymorphisms. Only one of the 43 polymorphisms results in an amino acid change, the one responsible for the two electrophoretic variants (fast, Adh-f, and slow, Adh-s) found in nearly all natural populations. The implication is that most amino acid changes in Adh would be selectively deleterious.*



7. Изучение мутационного процесса – мама, папа и потомок.

Появление методов секвенирования нового поколения сделало ненужными много красивых непрямых методов изучения мутационного процесса.



Rate of *de novo* mutations and the importance of father's age to disease risk

Augustine Kong[1], Michael L. Frigge[1], Gisli Masson[1], Soren Besenbacher[1,2], Patrick Sulem[1], Gisli Magnusson[1], Sigurjon A. Gudjonsson[1], Asgeir Sigurdsson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Wendy S. W. Wong[3], Gunnar Sigurdsson[1], G. Bragi Walters[1], Stacy Steinberg[1], Hannes Helgason[1], Gudmar Thorleifsson[1], Daniel F. Gudbjartsson[1], Agnar Helgason[1,4], Olafur Th. Magnusson[1], Unnur Thorsteinsdottir[1,5] & Kari Stefansson[1,5]

Mutations generate sequence diversity and provide a substrate for selection. The rate of *de novo* mutations is therefore of major importance to evolution. Here we conduct a study of genome-wide mutation rates by sequencing the entire genomes of 78 Icelandic parent–offspring trios at high coverage. We show that in our samples, with an average father's age of 29.7, the average *de novo* mutation rate is 1.20×10^{-8} per nucleotide per generation. Most notably, the diversity in mutation rate of single nucleotide polymorphisms is dominated by the age of the father at conception of the child. The effect is an increase of about two mutations per year. An exponential model estimates paternal mutations doubling every 16.5 years. After accounting for random Poisson variation, father's age is estimated to explain nearly all of the remaining variation in the *de novo* mutation counts. These observations shed light on the importance of the father's age on the risk of diseases such as schizophrenia and autism.

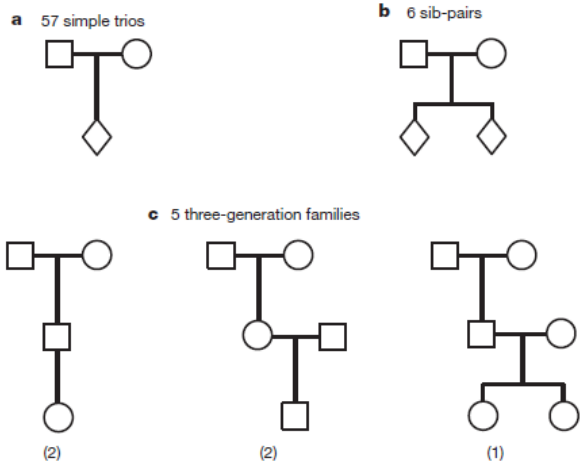
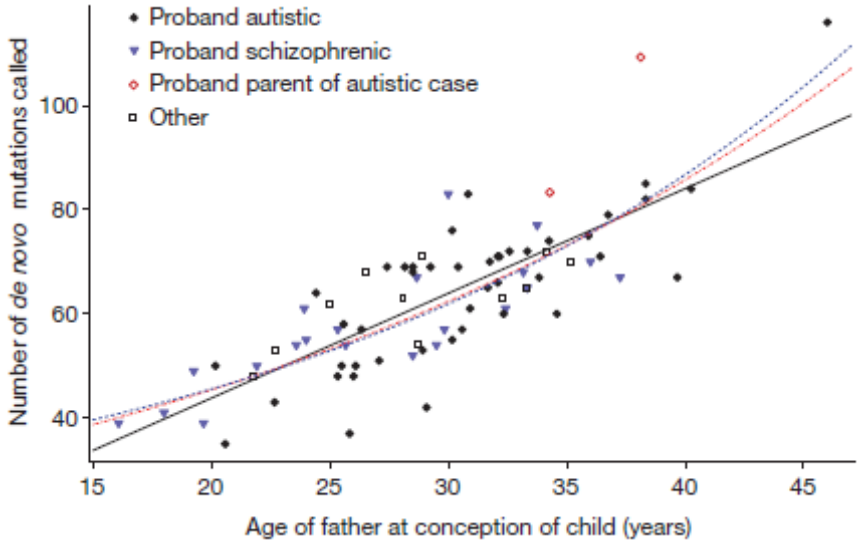


Figure 1 | A summary of the family types. a, Fifty-seven simple trios. b, Six sib-pairs accounting for 12 trios. c, Five three-generation families accounting for nine trios.



8. Изучение дрейфа: эффективная численность популяции, $H = 4N_e\mu$.

Эта замечательная формула дает нам непрямую оценку скорости дрейфа, если мы знаем скорость мутирования, поскольку H померить относительно несложно (хотя есть проблемы с отбором). Такие оценки незаменимы, поскольку измерить дрейф напрямую нереально – он слишком медленный.

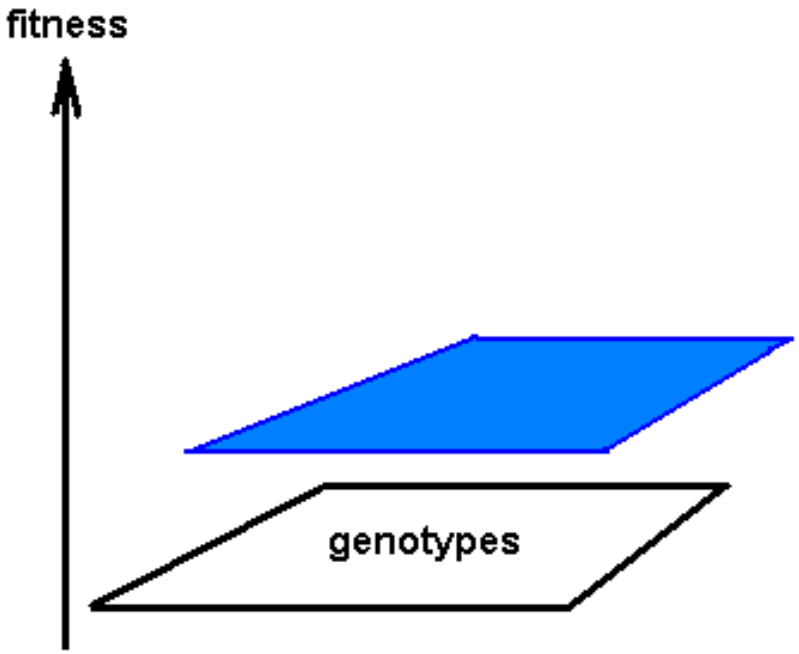
Table 1 | **Effective population size (N_e) estimates from DNA sequence diversities**

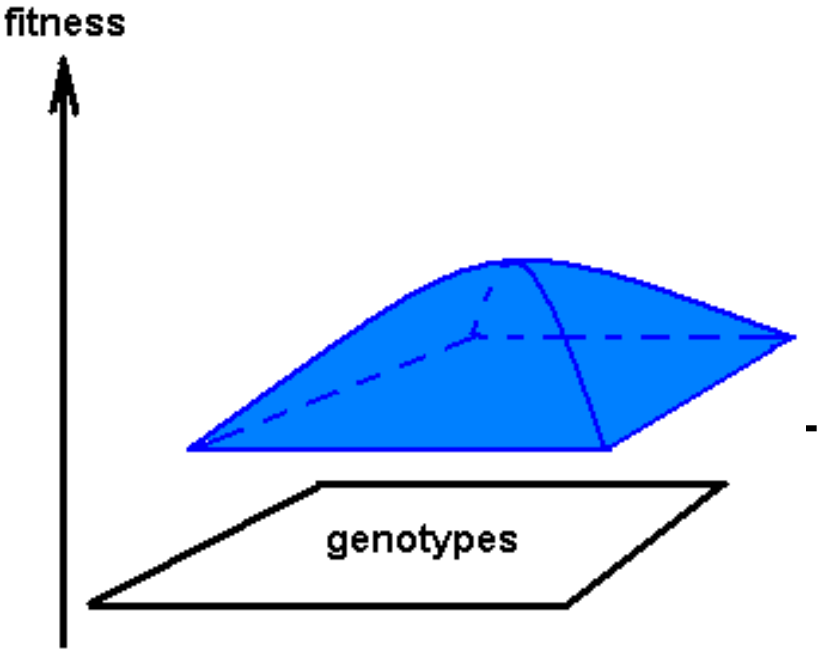
Species	N_e	Genes used	Refs
Species with direct mutation rate estimates			
Humans	10,400	50 nuclear sequences	145
Drosophila melanogaster (African populations)	1,150,000	252 nuclear genes	108
Caenorhabditis elegans (self-fertilizing hermaphrodite)	80,000	6 nuclear genes	41
Escherichia coli	25,000,000	410 genes	146
Species with indirect mutation rate estimates			
Bonobo	12,300	50 nuclear sequences	145
Chimpanzee	21,300	50 nuclear sequences	145
Gorilla	25,200	50 nuclear sequences	145
Gray whale	34,410	9 nuclear gene introns	147
Caenorhabditis remanei (separate sexes)	1,600,000	6 nuclear genes	43
Plasmodium falciparum	210,000 –300,000	204 nuclear genes	148

For data from genes, synonymous site diversity for nuclear genes was used as the basis for the calculation, unless otherwise stated.

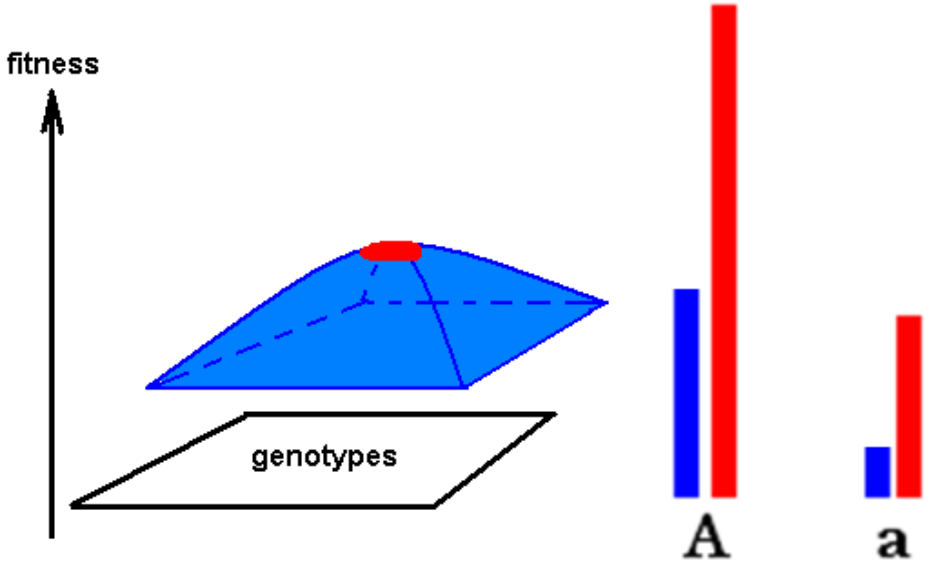
9. Отрицательный, положительный и балансирующий отбор.

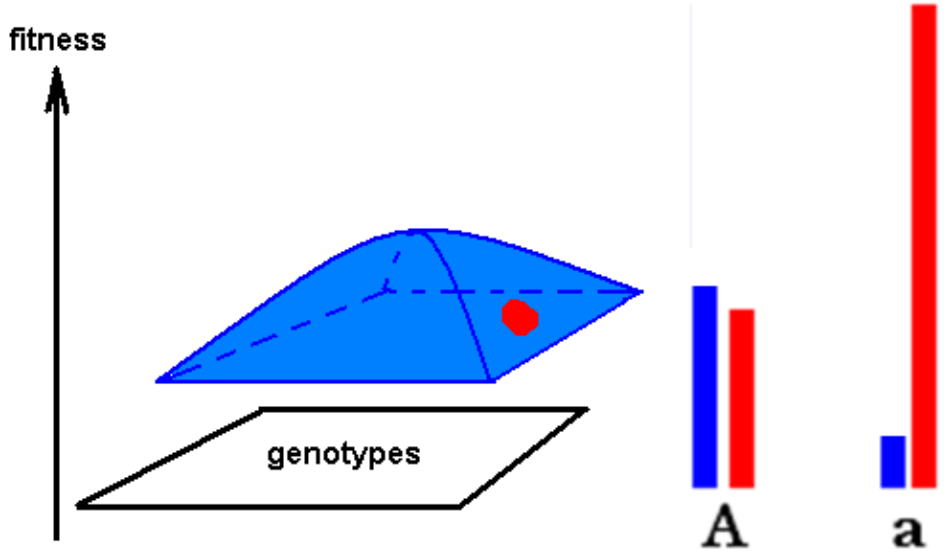
Адаптивный ландшафт это график функции генотип -> приспособленность. Конечно, пространство генотипов очень многомерно.





Форма отбора зависит от адаптивного ландшафта и от того, где находится популяция в пространстве генотипов.





Отрицательный (очищающий) отбор – самый приспособленный генотип – частый.

Положительный (дарвинов) отбор – самый приспособленный генотип – редкий.

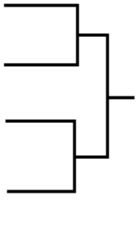
Отрицательный отбор поддерживает статус кво и препятствует изменениям. Положительный отбор способствует изменениям. После того, как положительный отбор сделал свое дело, и наиболее приспособленный генотип стал частым, отбор делается отрицательным – при неизменности адаптивного ландшафта.

Балансирующий отбор благоприятствует любому редкому генотипу, что возможно только если адаптивный ландшафт меняется.

What kinds of sequence data can be used to study selection?

First, we may be dealing with interspecies differences - with an alignment of homologous (orthologous or paralogous) sequences from different species, with known phylogeny.

```
Homo     fkvmnasdfrtshnmcvadnmd
Macacca  fklmnasdfrtshnmcvqdnmd
Rattus   fklmnatdfrtshnmcvadnmd
Mus      fkvmnasdfrtshnicvadnmd
```



Variable sites are shown in red.

Second, we can be dealing with within-population variation. This variation can also be represented by alignment of different genotypes (preferably polarized by an outgroup, a genotype from related species, in order to distinguish ancestral and derived alleles)

```
Homo genotype 1: catgccagca-cgtctagcatatacgcagactcgctatttacgtcacgatgagcat
Homo genotype 2: catgccagcatcgtctagcatatacacagactcgctatttacgtcacgatcagcat
Homo genotype 3: catgccagcatcgtctagcatatacgcagactcgctatttacgtcacgatcagcat
Pan (outgroup) : catgccagcatcgtgtagcatataggcagactcgctatttacgtcacgatcagtat
```

Derived human alleles are shown in red. Sites where the outgroup diverged are in blue.

Detecting negative selection

This is a relatively easy task - because negative selection is very common. Negative selection affects evolving sequences in two ways:

1) it reduces the probability of fixation of a mutation with s < 0

2) it reduces the time until elimination of a mutation with s < 0

As a result, negative selection leaves two kinds of footprints:

1) reduced rate of evolution and the level of within-population variation

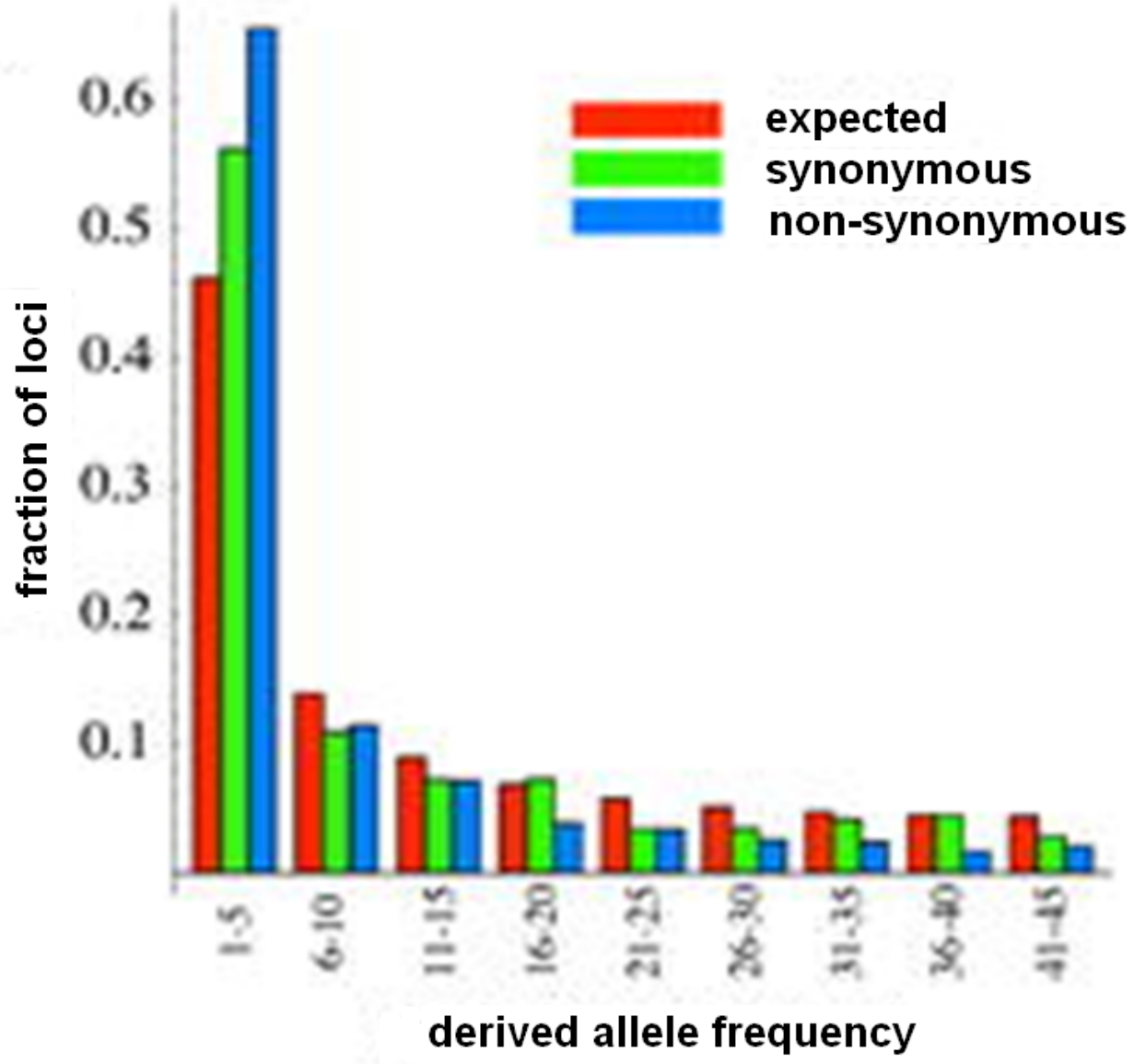
Reduced relative to what? - to the rate of evolution at selectively neutral sites. According to the fundamental theorem of neutral evolution, neutral sites evolve at the mutation rate (this is intuitively obvious). Practically, negative selection is detected by comparing the amount of interspecies divergence or within-population polymorphism to that at plausibly neutral sequence sites.

Alignment of orthologous regulatory regions of 4 mammals. A transcription factor-binding site with low divergence is marked by blue. If the alignment includes only a few sequences, we can only detect substantial segments with reduced divergence rates (never call them mutation rates!) - for example, using Hidden Markov Model technique.

A typical segment of an alignment of orthologous proteins from different species. Here the number of sequences makes it possible to detect negative selection even at individual sites.

Data on within-population variation usually allow us only to detect negative selection in wide classes of sites, for example to show that non-synonymous coding sites are under stronger selection than synonymous sites. However, with high H making inferences about individual sites may become possible. We badly need 100 genotypes of *Ciona savignyi*.

2) An excess of rare alleles



Distribution of allele (nucleotide) frequencies in *Arabidopsis thaliana*. *PLoS Biology* 3, 1289-1299, 2005.

At non-synonymous sites an excess of rare alleles, relative to the neutral expectation, is higher. Of course, here we cannot make inferences about individual sites.

However, we can make inferences about the strength of negative selection - because only alleles with small s are observed as rare polymorphisms.

In contrast, reduced rate of evolution tells us very little about the strength of selection: s = -0.001 is enough to stop evolution.

Detecting positive selection

This is a difficult and important problem - because positive selection is rare, relatively to negative selection (this was proposed in 1935 by Ivan Schmalhausen) and because positive selection is the only driving force of adaptive evolution.



Positive selection affects evolving sequences in two ways:

1) it increases the probability of fixation of a mutation with s > 0

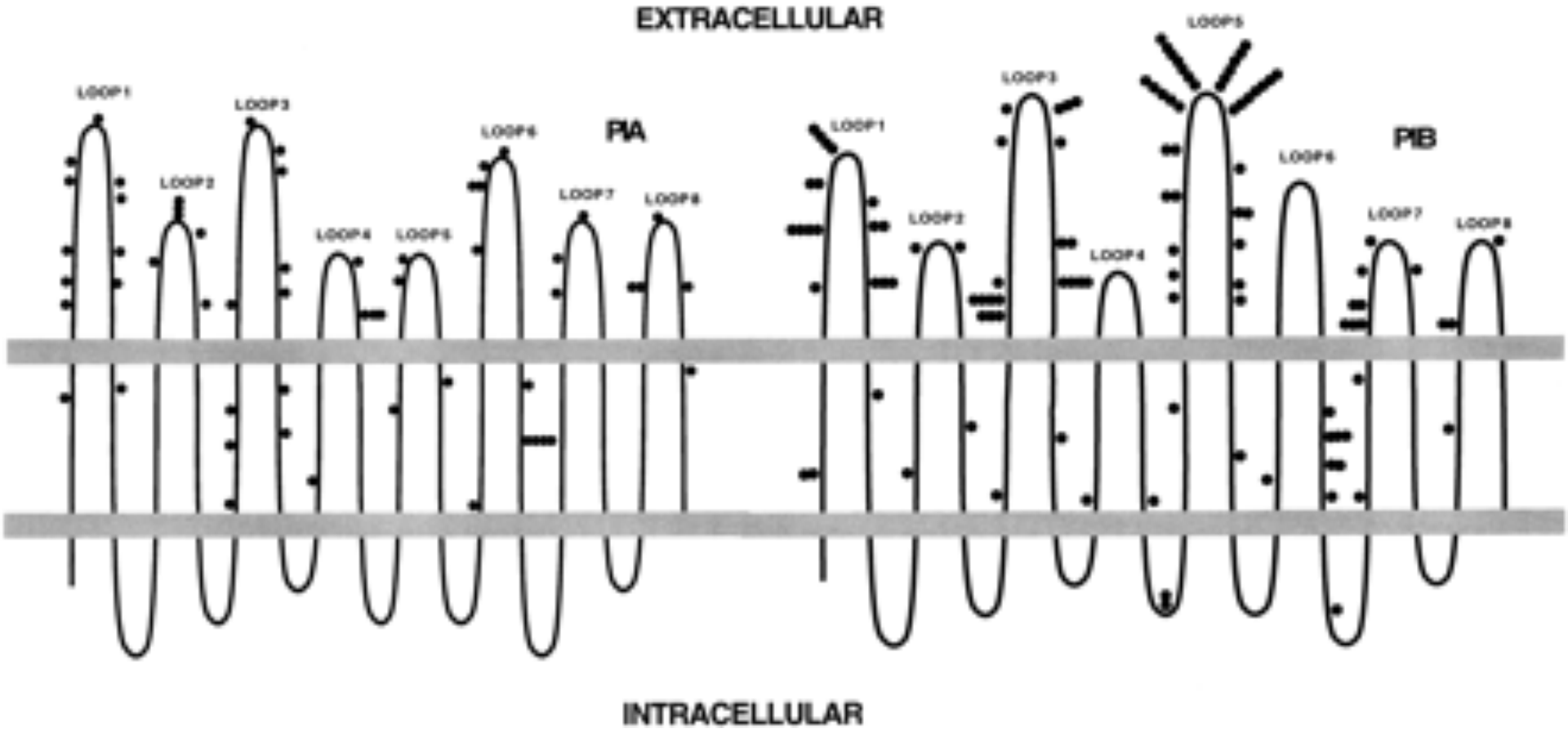
2) it reduces the time until fixation of a mutation with s > 0

Footprint of positive selection looks rather differently depending on its age.

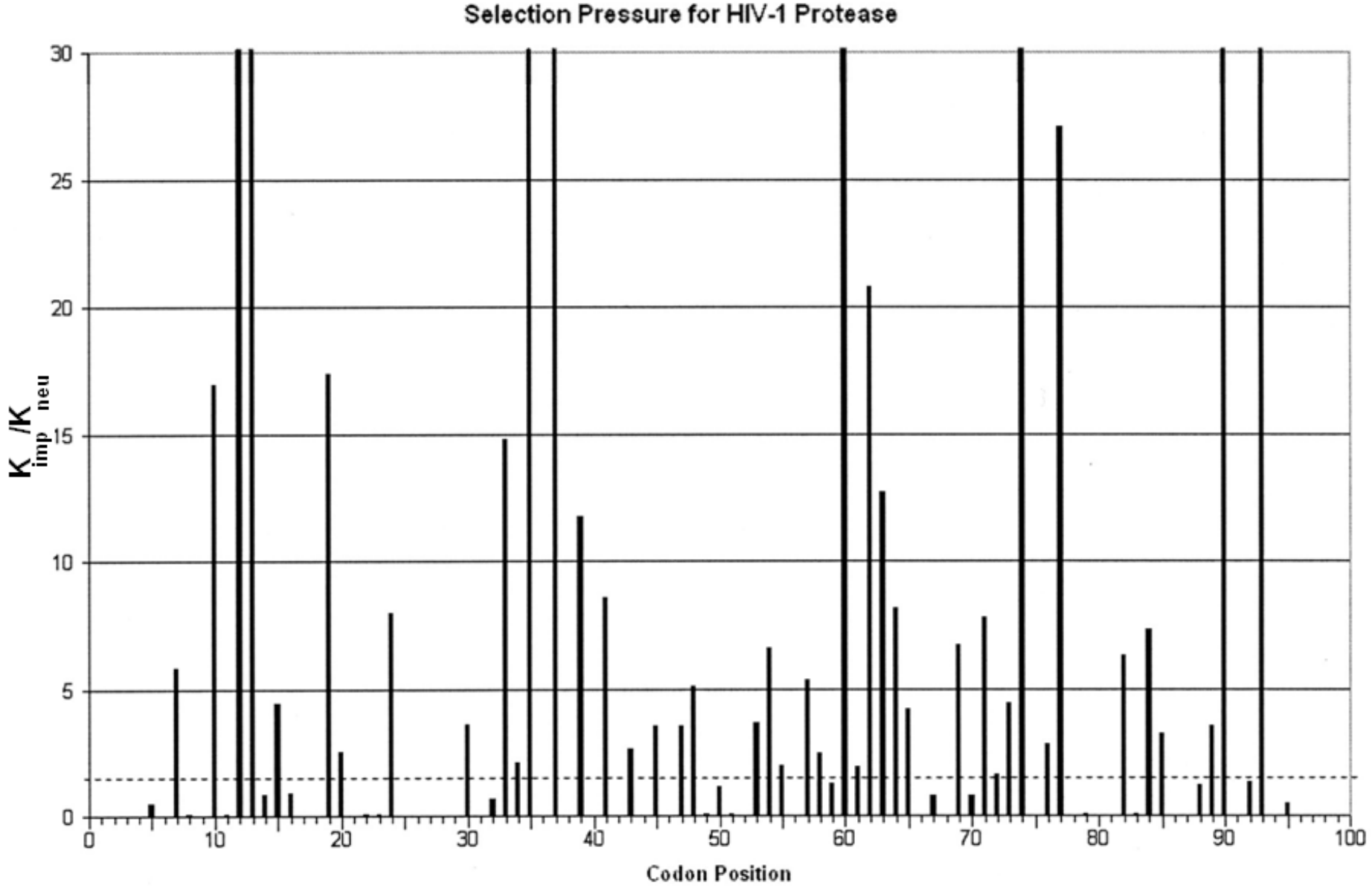
1) Positive selection accomplished a long time ago - interspecies comparisons

In contrast to negative selection, positive selection accelerates evolution (not the rate of evolution!). Thus, it makes sites or segments to evolve faster than neutrally. As a result, we can detect positive selection only from comparing relatively close species, such that the number of accepted substitutions between them per neutral site, K_{neu}, is ~1-3. Ancient actions of positive selection, that occurred more than 1/m generations ago (m is the per nucleotide mutation rate) could never be detected.

So, if we have a large number of close enough sequences, even individual sites where K > K_{neu} (K_{neu} is measured for sites that are probably under no selection) can be detected. This approach works well for pathogens, with multiple moderately different strains.



Distribution of amino acid replacements along the *Neisseria gonorrhoeae* transmembrane porin sequence. Each dot represents one replacement. Obviously, sequence segments exposed outside the cell evolve much faster, probably due to positive selection. *Molecular Biology and Evolution* 17, 423-436, 2000.



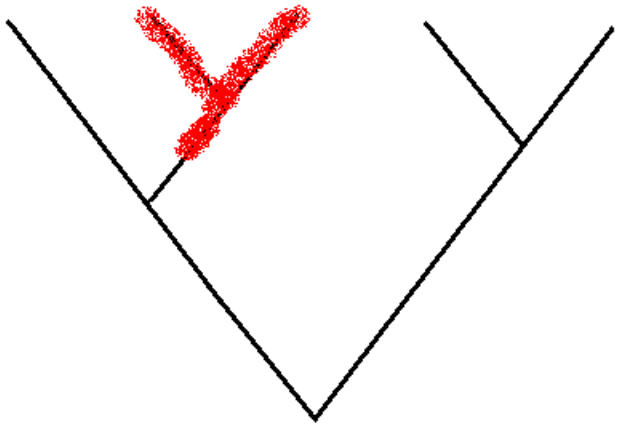
Positive selection in HIV-1 protease, detected on samples from 40,000 patients. For each codon site, the ratio of the rate of the most common allele replacement over the neutral rate is shown (*Journal of Virology* 78, 3722-3732, 2004).

However, there are two problems with this approach:

1) Positive selection can act only within one clade, with negative selection acting at the same site in the rest of the phylogeny. Then, overall K will be low at the site.

2) There may be not enough species to measure K for individual sites. If so, all probably important sites are treated together, and their average per site number of changes, K_{imp}, is calculated. Trouble is, sites under positive are generally scattered between numerous sites under negative selection, leading to $K_{imp} < K_{neu}$. Only very rarely, there are long enough segments with a majority of sites under positive selection.

Positive selection acting in one clade, on a sparse phylogenetic tree.



Sophisticated statistical methods can be used to analyze such data - but, in my opinion, they reliably detect positive selection only if a substantial fraction of sites to $K_{imp} > K_{neu}$. at least within a large clade - and this is generally very rare. Most of "important" sites are, most of the time, under negative, and not positive selection.

A clever idea of MacDonald and Kreitman can offer some help. They realized that the condition $K_{imp} > K_{neu}$ (or $K_{imp}/K_{neu} > 1$) can be relaxed. If negative selection is strong, "important" sites under it will not be polymorphic in the population. Sites under positive selection also make only minimal contribution to polymorphism (because polymorphism in the course of an allele replacement is very short-lived). Thus, instead of asking for

$$K_{imp}/K_{neu} > 1$$

as a signature of positive selection it is enough to ask for

$$K_{imp}/K_{neu} > H_{imp}/H_{neu}$$

H_{imp}/H_{neu} can be as low as 0.2-0.3 (due to a large fraction of sites under negative selection among "important" sites), so this is a much less stringent condition.

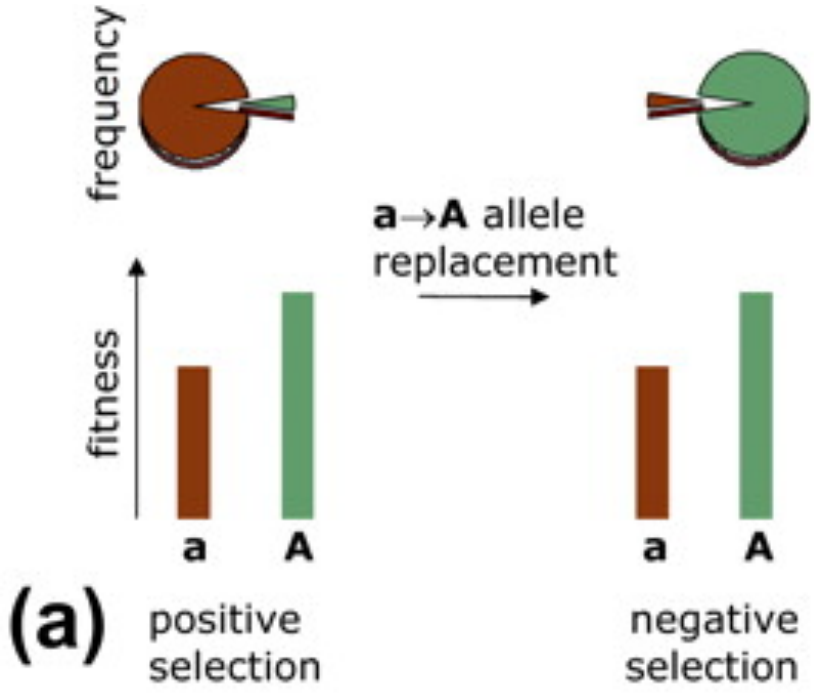
One problem with this approach is that slightly deleterious variants with $-s \sim 1/N_e$ can segregate within the population, but are only rarely fixed, and thus inflate H_{imp}/H_{neu}. A possible way of dealing with this problem is to ignore rare variants.

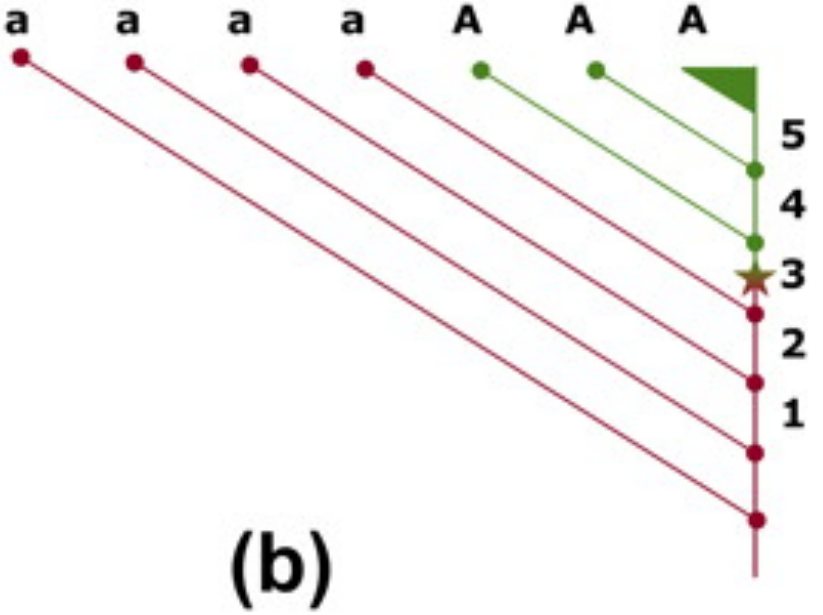
Some applications of MacDonald-Kreitman test to *Drosophila* species suggest that as many as 50% of allele replacements in fly evolution were driven by positive selection, because

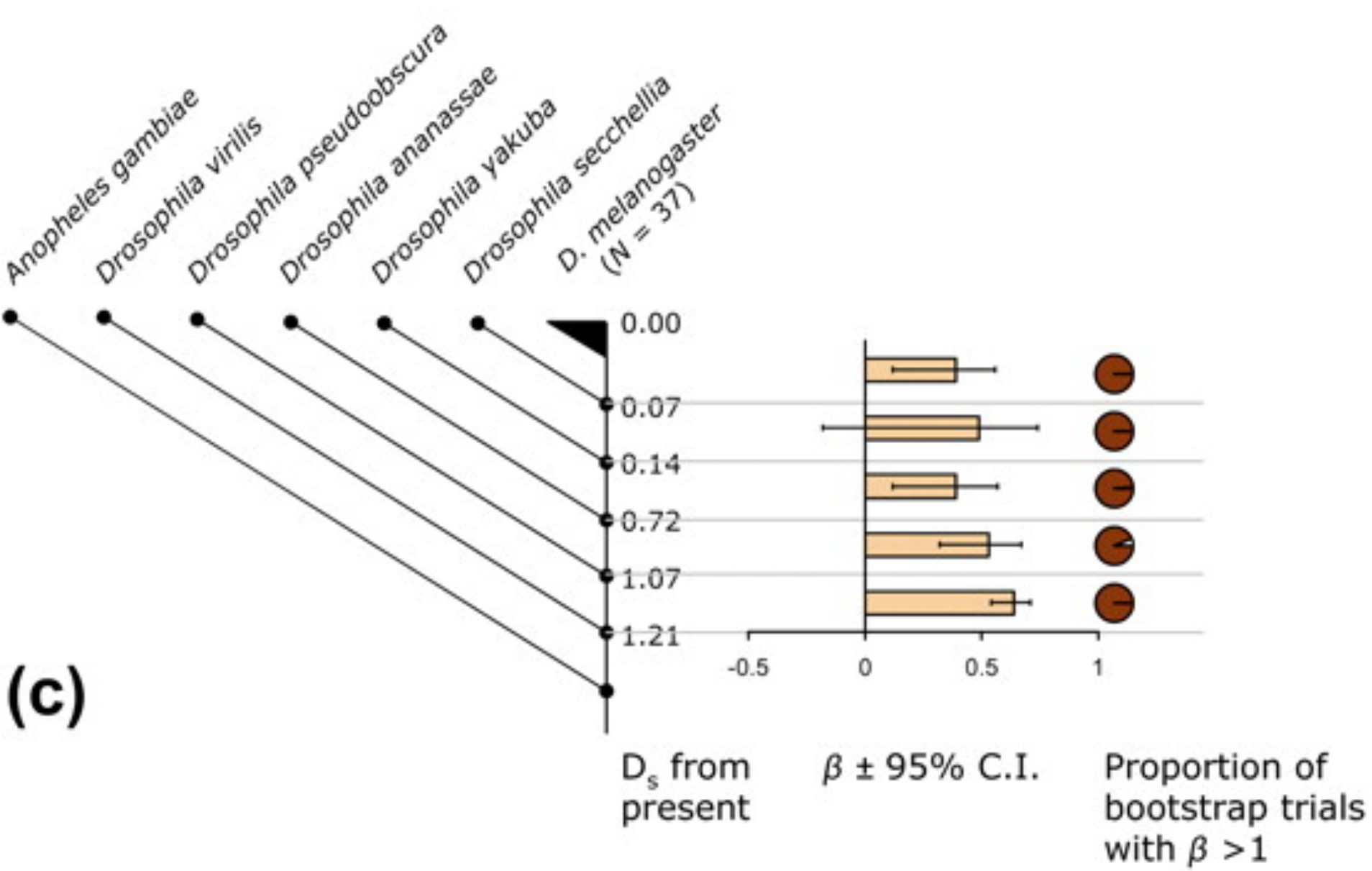
$$K_{imp}/K_{neu} = 2H_{imp}/H_{neu}$$

In contrast, in mammals $K_{imp}/K_{neu} < H_{imp}/H_{neu}$, suggesting no positive selection. The reasons for such contrast are unclear. Anyway, MK test could never establish identities of individual sites under positive selection.

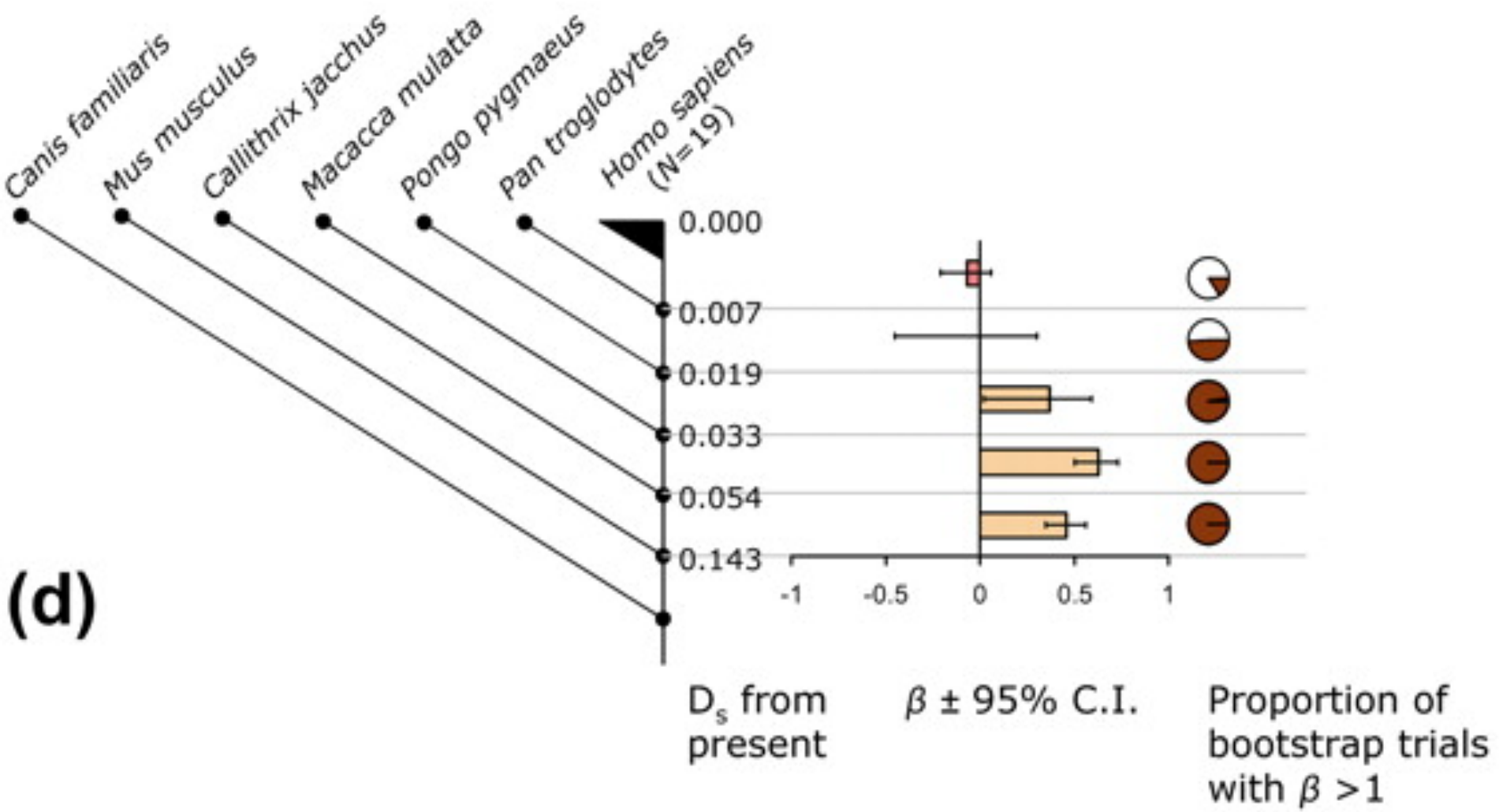
Обнаружение положительного отбора в прошлом по отрицательному отбору в настоящем.







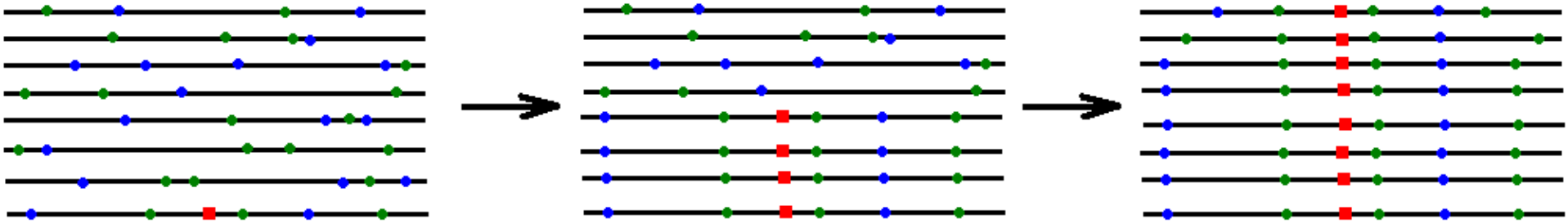
В линии мухи положительный отбор шел все время.



В линии человека положительный отбор ослаб около 20 миллионов лет назад – видимо, из-за усиления дрейфа.

2) Positive selection accomplished recently - within-population variation

A recent allele replacement driven by positive selection produces a region of very low variation, flanked by regions with some high-frequency derived alleles. Such a scar of an allele replacement is due to an effect called hitch-hiking, and it remains visible for $<< 1/N_e$ generations, where N_e is the effective population size per nucleotide mutation rate.

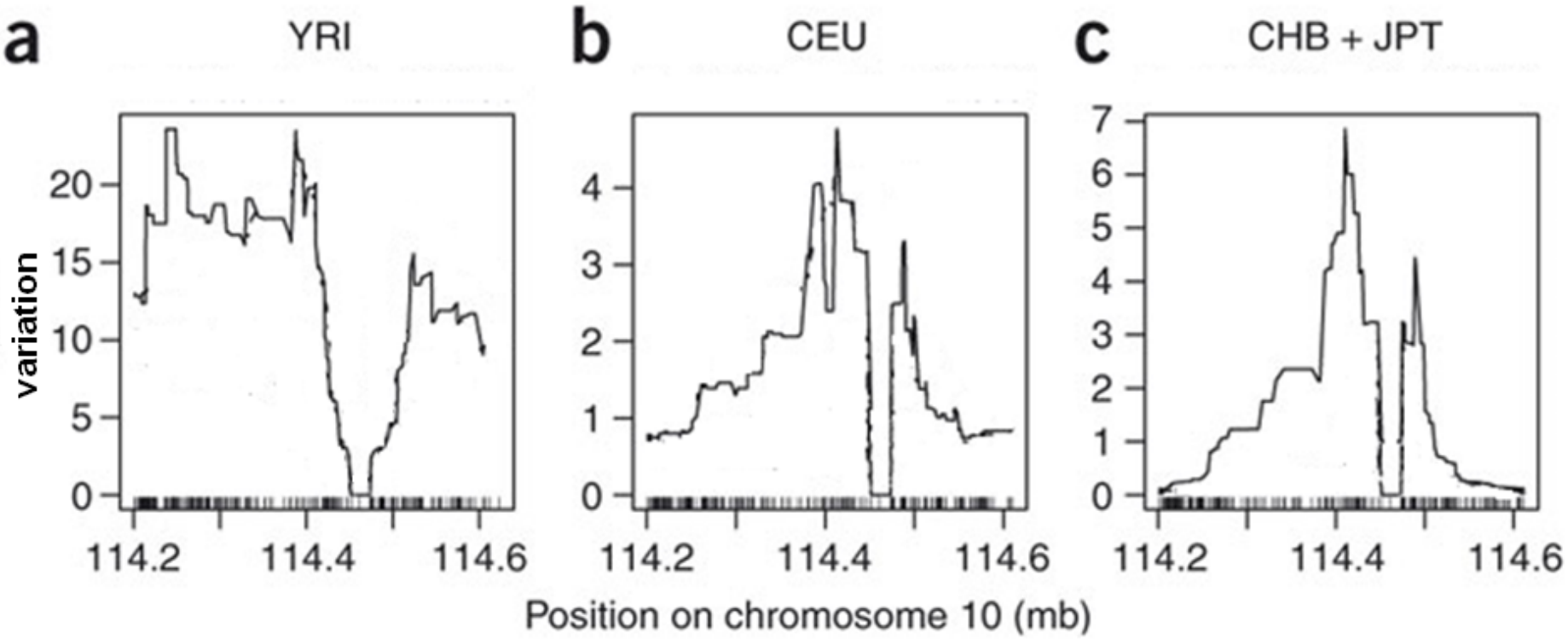


A beneficial mutation (red) in a population with many segregating neutral (green) and slightly deleterious (blue) variants.

Half-way towards fixation, the beneficial mutation carries with it the close-by variants.

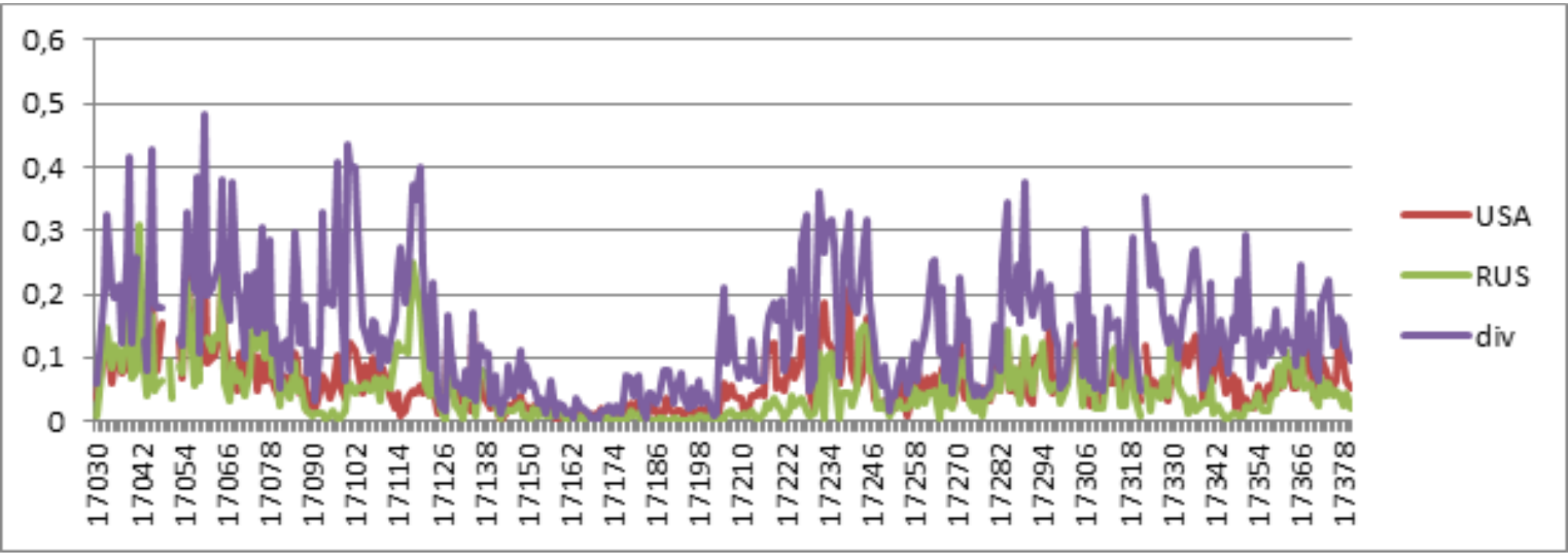
Some of these variants become detached, due to crossing-over, by the time of the fixation.

There are several definite known cases of recently accomplished selective sweeps.

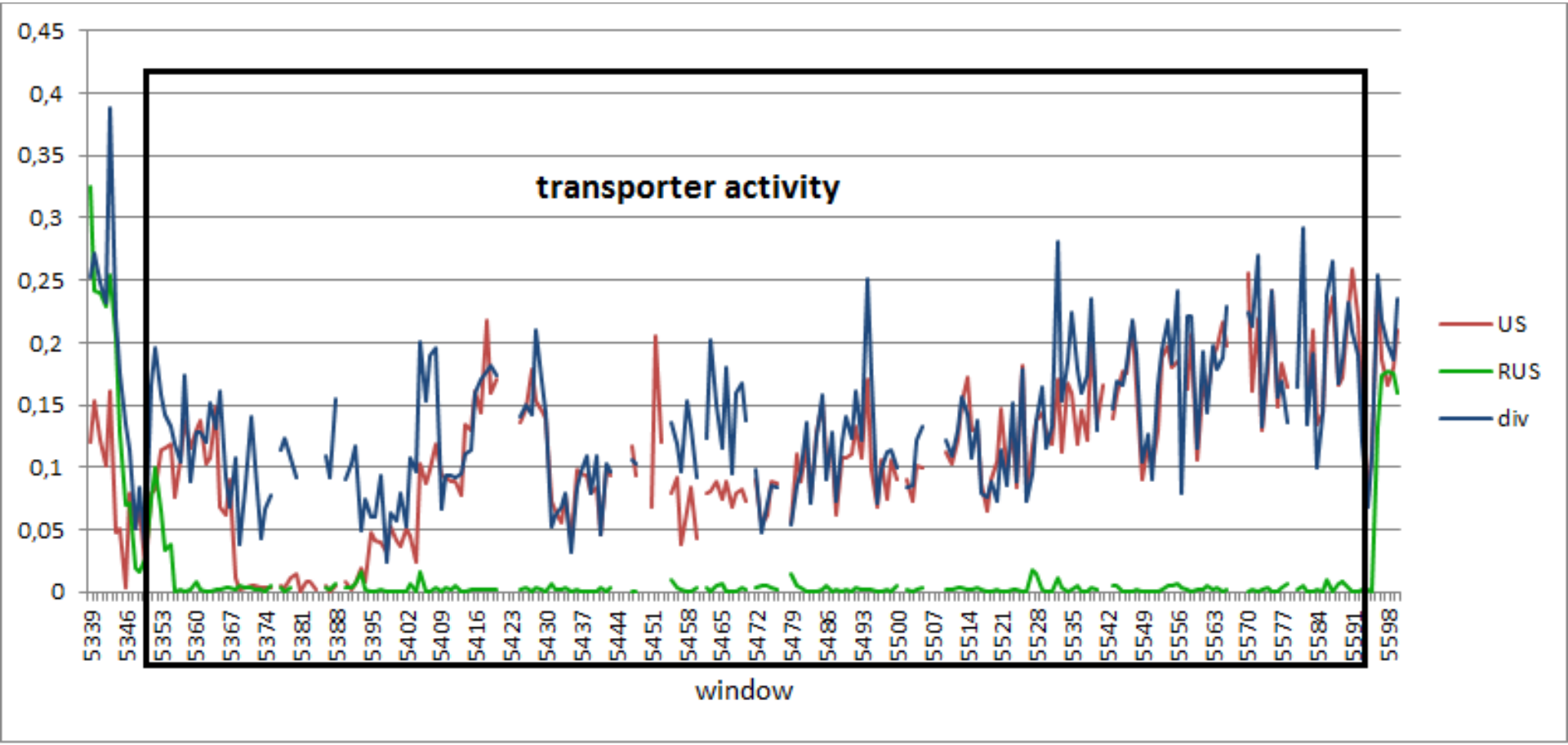


Reduced levels of genetic variation around the site of recent positive selection-driven allele replacement (selective sweep) in human populations from Africa (a), Europe (b), and East Asia (c) (*Nature Genetics* 39, 218 - 225, 2007).

«Свипы» у шизофиллума.



Глобальный – изменчивость снижена везде. Видимо, полезный аллель перелетел с одного континента на другой.

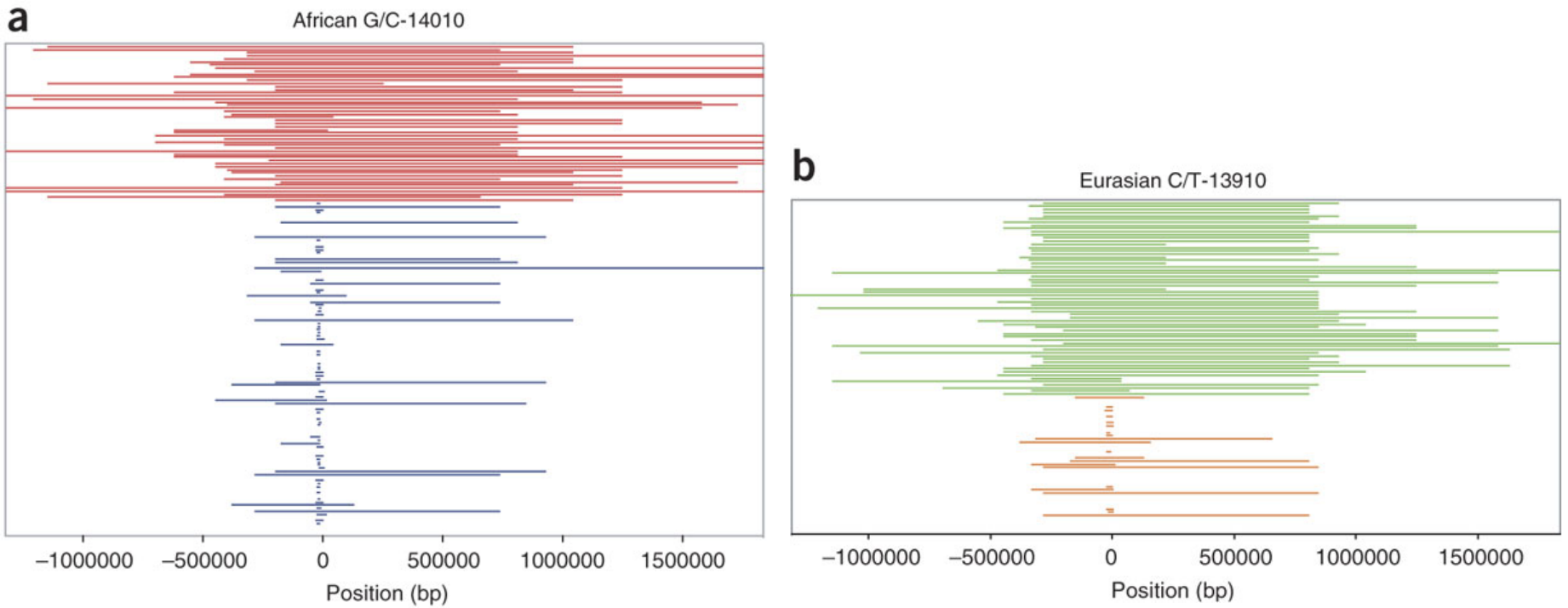


Scaffold 10

Локальный – приспособленность снижена только в России. Русских свипов больше, чем Американских – похоже, что Русская популяция моложе.

3) Ongoing positive selection - within-population variation

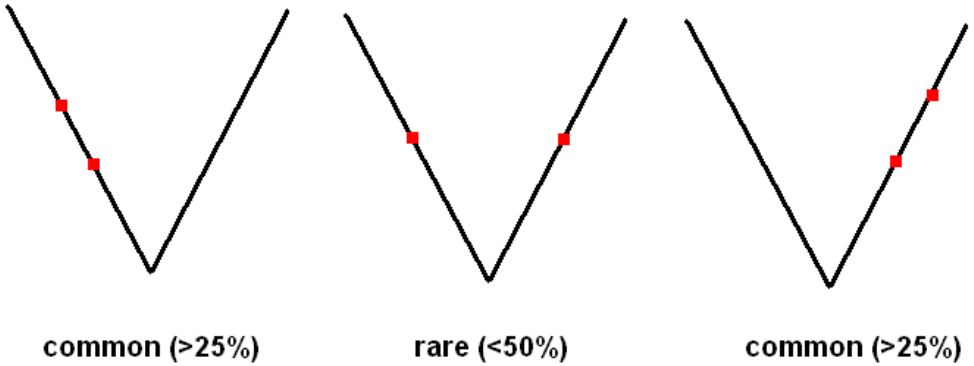
One must be lucky to study the right population at the right time. Still, there are some definite cases of ongoing allele replacements driven by strong positive selection. One of them is parallel acquisition the ability of adults to digest milk (due to persistent expression of lactase) in Africans and non-Africans. These ongoing sweeps left clear-cut signatures.



(a) Kenyan and Tanzanian C-14010 lactase-persistent (red) and non-persistent G-14010 (blue) homozygosity tracts. (b) European and Asian T-13910 lactase-persistent (green) and C-13910 non-persistent (orange) homozygosity tracts. Positions are relative to the start codon of lactase locus (*Nature Genetics* 39, 31 - 40, 2006).

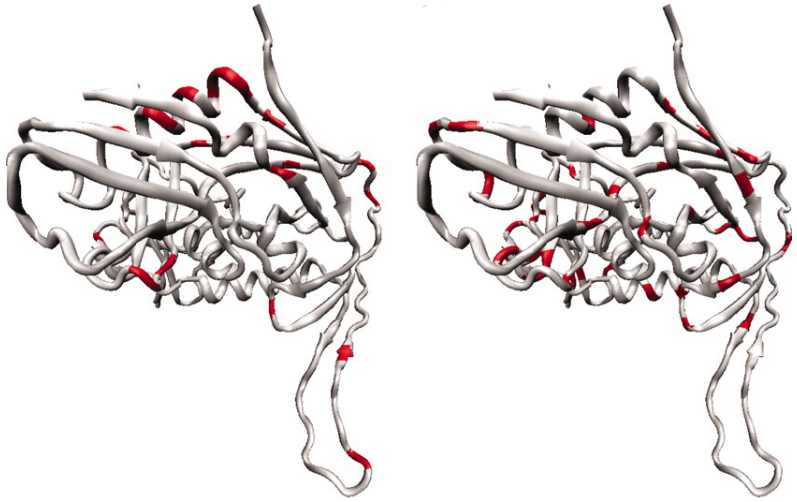
4) A different approach - detecting positive selection by bursts of substitutions

Suppose that at a codon site fitness landscape was suddenly changed. The new optimal amino acid may not be reachable from the old one by a single nucleotide substitution. Then, a clump of two or even three non-synonymous substitutions may follow. Such clumps were observed in evolution of mammals and HIV-1 (*PNAS* 103, 19396-19401, 2006).



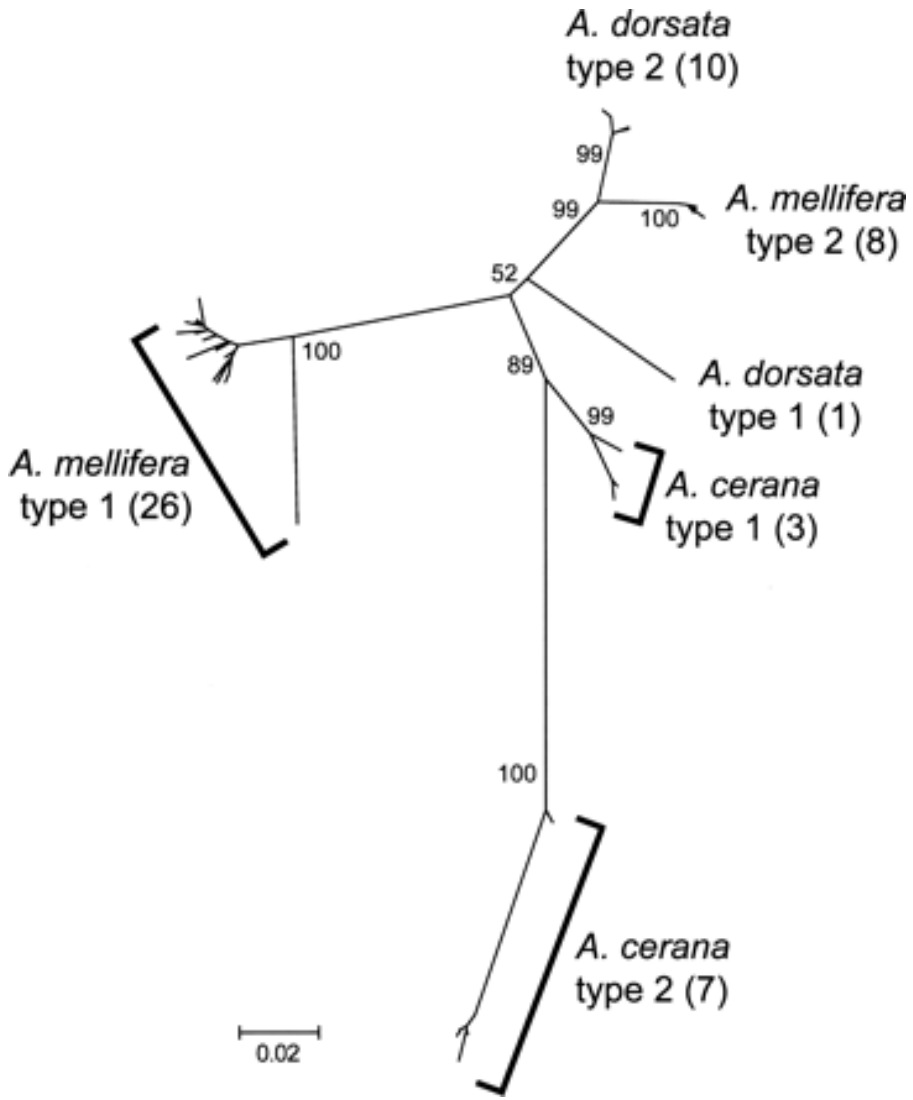
Clumping of nonsynonymous substitutions is the strongest in conservative regions of proteins, where the 1:1 situations occur only in ~20% of codons. Indeed - if an important amino acid is replaced, this must be beneficial. This approach reveals a number of slowly-evolving sites that occasionally undergo positive selection.

Amino acid sites inferred to be under positive selection in HIV-1 gp120. Left: rapidly evolving sites previously inferred to be under positive selection. Right: conservative sites with strongly clumped substitutions.



Detecting balanced selection

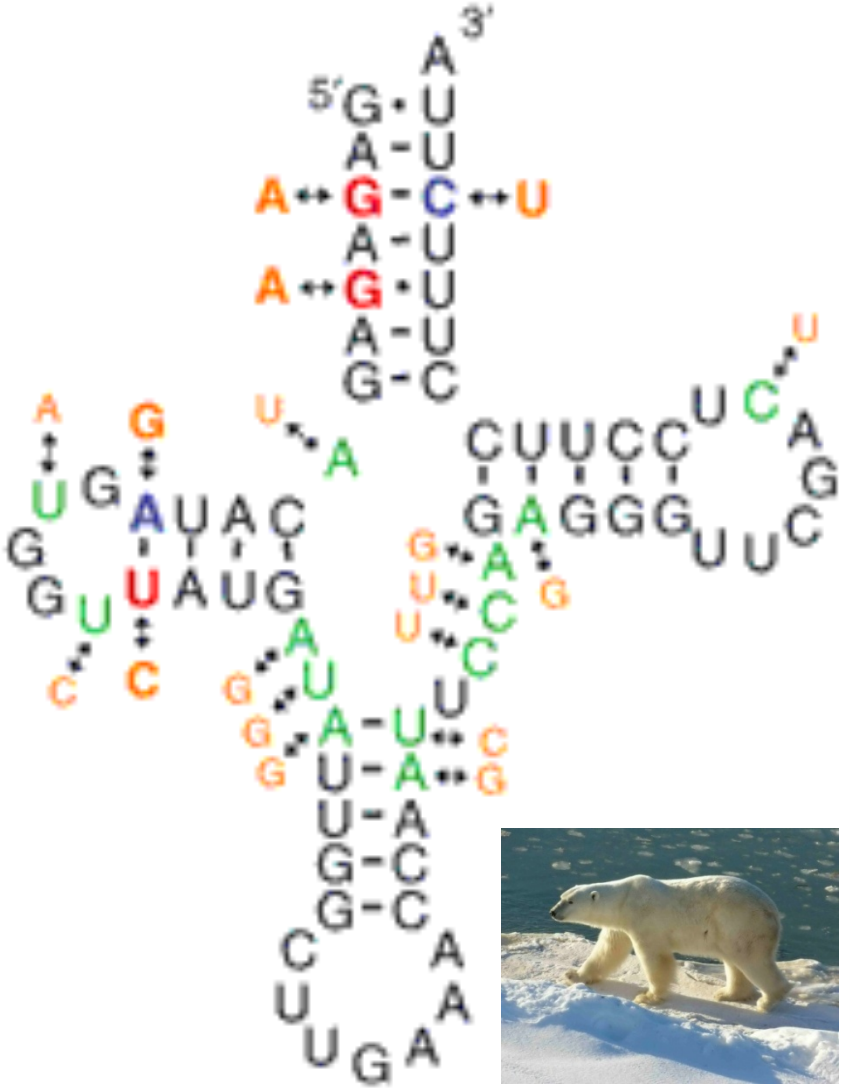
Balancing selection, which requires changing fitness landscapes, favors rare alleles. It prevents fixations and losses of the alleles involved, leading to durable polymorphisms.



In the extreme case this can lead to transspecies polymorphisms, persisting from the time of species divergence. This is the case for sad *csd* (complementary sex determination) locus in bees. Female must be heterozygous at this locus, and homozygotes develop into sterile males, causing strong selection against common alleles (*Genome Res.* 16, 1366-1375, 2006*).*

Detecting epistatic selection

Let us consider just one salient manifestation of epistasis: compensated pathogenic deviations, such that an nucleotide or amino acid normal for one species would be severely deleterious, at the same site of the orthologous molecules, for another species.



CDPs are very common in tRNAs. Three of them are present in mitochondrial tRNA^{Ser} of *Ursus maritimus*. Nucleotides corresponding to human pathogenic mutations are shown in red; predicted compensatory substitutions are shown in blue; and other deviations from the human ortholog, unrelated to the pathogenic mutations or their compensations, are shown in green. Nucleotides found in healthy humans are shown in orange alongside the bear sequence.

At least five mechanisms of compensation are known for pathogenic mutations that destroy a Watson-Crick pair in one of the four tRNA stems: restoration of the affected Watson-Crick interaction, strengthening of another pair, creation of a new pair, changes of multiple interactions in the affected stem and changes involving the interaction between the loop and stem structures. (Nature Genetics 36, 1207-1212, 2004).

11. Изучение размножения.

Independent joint distribution of alleles at several loci means that frequency of a genotype equals to the product of frequencies of its constituent alleles. In the case of two diallelic loci, alleles at A and B are distributed independently if and only if

[AB] = [A][B], [Ab] = [A][b], [aB] = [a][B], and [ab] = [a][b].

A convenient measure of non-independence of distributions of alleles is the coefficient of association (or of linkage disequilibrium - a horrible term!) of a pair of loci:

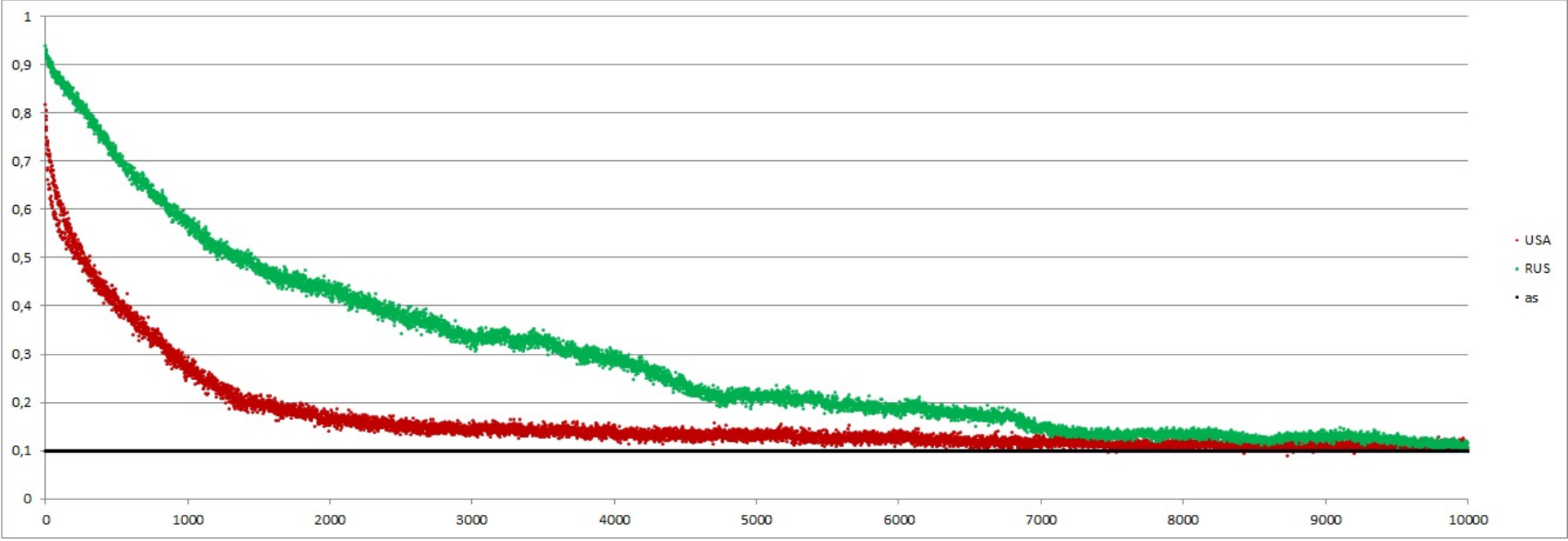
d = [AB] [ab] - [Ab] [aB]

independence **dependence**

In sexual populations only alleles at tightly linked loci are associated, and for pairs of more distant loci d ~ 0. Some distances between nucleotide sites at which d vanishes are:

~100,000 in *Homo sapiens*
~1000 in *Drosophila melanogaster*

In species with high H, d disappears at shorter distances - there is a reason for this!



Убывание LD с расстоянием между локусами в Американской и Русской популяциях шизофиллума. Сразу видно, что этот вид размножается половым путем.

12. Изучение структуры.

Есть много методов получения непрямых оценок степени географической структурированности популяций на основании их генетической дифференциации.

$$G_{ST} = (H_T - H_S)/H_T$$

Species	Phylum	Pelagic larva	Swim ability	G_{ST}
Coris julis	Vertebrata	yes	strong	0.133
Amphiprion clarkii	Vertebrata	yes	strong	0.008
Dascyllus trimaculatus	Vertebrata	yes	strong	0.720
Holocentrus ascensionis	Vertebrata	yes	strong	0.091
Botrylloides magnicoecum	Urochordata	no	weak	0.202
Stolonica australis	Urochordata	no	weak	0.201
Pyura gibbosa	Urochordata	yes	weak	0.002
Phallusia nigra	Urochordata	yes	weak	0.083
Strongylocentrotus droebachiens	Echinodermata	yes	weak	0.002
Strongylocentrotus purpuratus	Echinodermata	yes	weak	0.033
Cucumaria pseudocurata	Echinodermata	no	weak	0.966
Cucumaria miniata	Echinodermata	yes	weak	0.025
Tridacna maxima	Mollusca	yes	weak	0.156