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

Мутации во времени:

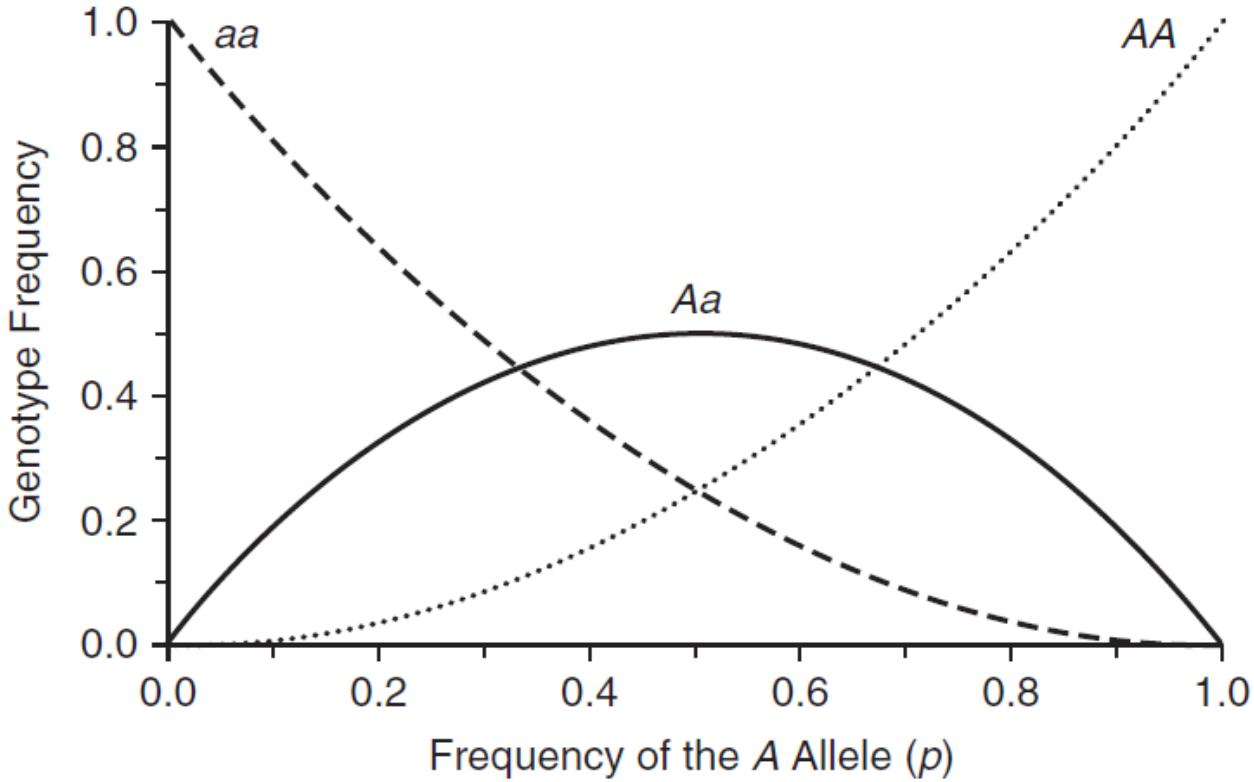
Основы популяционной генетики, ч. 1

План лекции

1. Равновесие Харди-Вайнберга
2. Случайный дрейф в отсутствие мутаций
3. Эффективный размер популяции
4. Случайный дрейф и мутации

Равновесие Харди-Вайнберга (1908)

Поколение N: $f_A = p, f_a = q, p + q = 1$
Поколение $N + 1$: $F_{AA} = p^2, F_{Aa} = 2pq, F_{aa} = q^2$



Равновесие Харди-Вайнберга

Поколение N: $f_A = p, f_a = q, p + q = 1$
Поколение $N + 1$: $F_{AA} = p^2, F_{Aa} = 2pq, F_{aa} = q^2$

Следствия:

1. Частоты аллелей не меняются:
   $p' = f'_A = F'_{AA} + \frac{F'_{Aa}}{2} = p^2 + pq = p$
2. Частоты равновесия Харди-Вайнберга достигаются не более, чем за два поколения

Задача

Выведите частоты аллелей в поколении $N + 1$

Равновесие Харди-Вайнберга

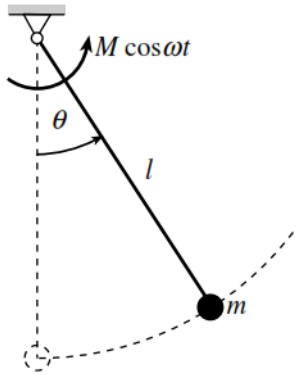
Предположения:

- Диплоидные организмы с половым размножением и случайным (не-выборочным) скрещиванием
- Одинаковые частоты аллелей у мужских и женских особей
- Неперекрывающиеся поколения
- Биаллельные (аутосомные) локусы
- Бесконечный размер популяции
- Отсутствует изменение частот аллелей за счет миграции, естественного отбора или мутаций
- Нет ошибок генотипирования

Равновесие Харди-Вайнберга

Есть ли смысл в равновесии Харди-Вайнберга с таким количеством предположений? Да:

- Основа для более реалистичных моделей
- Модель Харди-Вайнберга разделяет жизненный цикл на два интервала: гаметы → зиготы → взрослые особи



Равновесие Харди-Вайнберга

Проверка на равновесие Харди-Вайнберга:
$df = n - k - 1$, где $n = 3$ – количество классов, а $k = 1$ – количество независимых параметров. $\chi^2 = \sum \frac{(O-E)^2}{E}$

Genotype	Observed Number (O)	Expected Number (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
AA	90	83.2	6.8	46.24	0.5558
Aa	28	41.6	−13.6	184.96	4.4462
aa	12	5.2	6.8	46.24	8.8923

After performing the calculations in this table, we get a chi-square (χ^2) statistic of

$$\chi^2 = 0.5558 + 4.4462 + 8.8923 = 13.8943$$

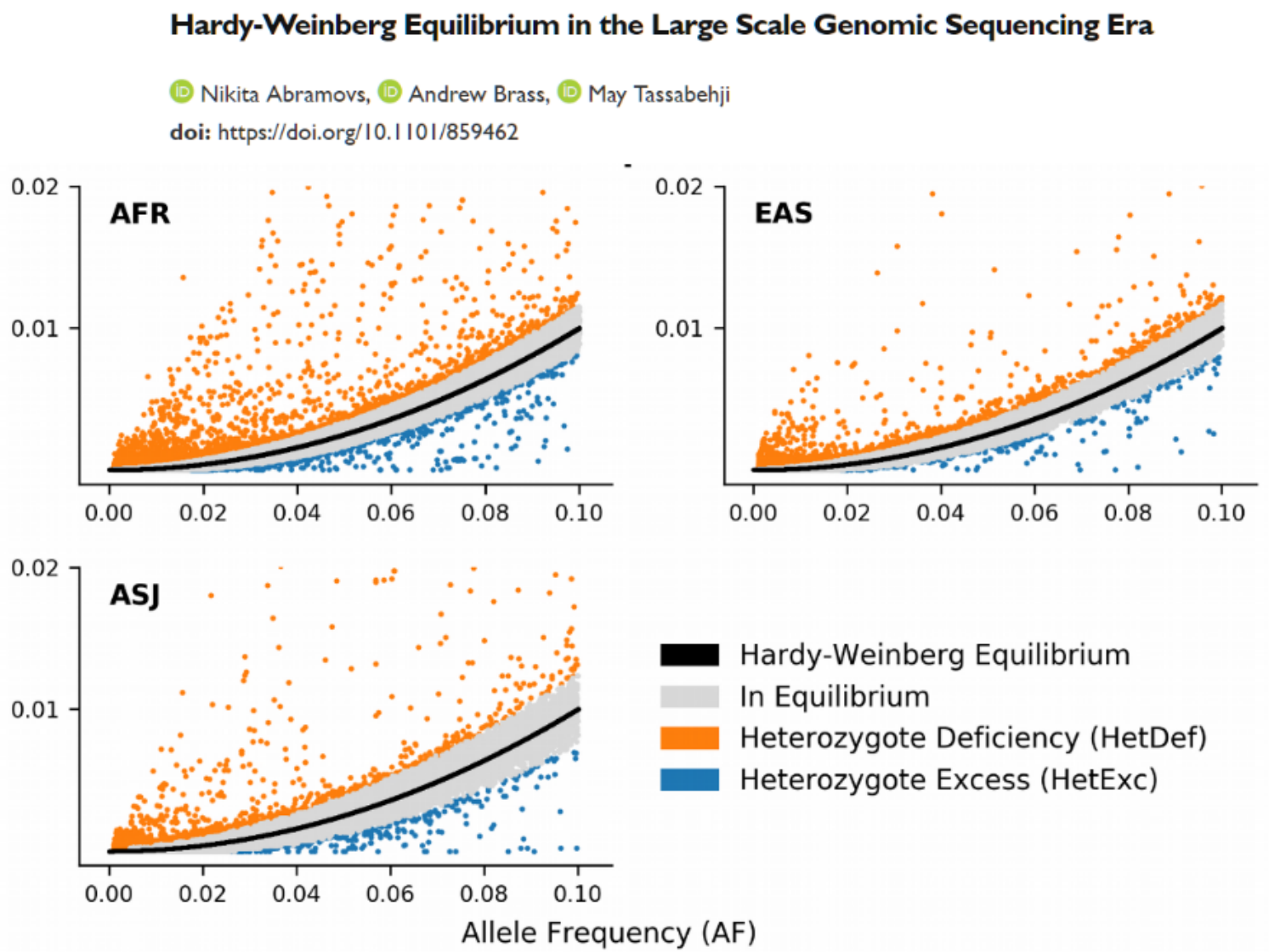
This value is *much* larger than the critical value of 3.841, so we reject the hypothesis of Hardy–Weinberg equilibrium.

Relethford – *Human Population Genetics*

Задача

Проделайте эти расчеты самостоятельно

Равновесие Харди-Вайнберга



gnomAD: 137,842 преимущественно здоровых индивидуумов из 7 основных популяций

Селекционисты и нейтралисты

Селекционисты (prerequisites):

- Понижающие приспособленность аллели удаляются отбором, повышающие — [быстро] фиксируются отбором.
- Источником полиморфизма генома является балансирующий отбор.

Нейтралисты:

- Большинство замен (фиксаций) происходят из-за случайного дрейфа, а не из-за мутаций, повышающих приспособленность.
- Источником полиморфизма генома является *случайный дрейф* [почти] нейтральных аллелей.

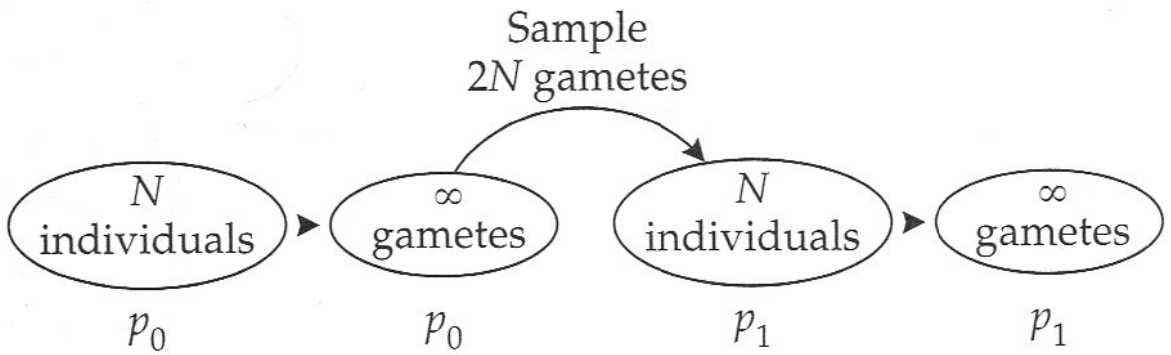
Случайный дрейф

Предположения (Райт-Фишер, 1930):

- Диплоидные организмы с половым размножением и случайным (не-выборочным) скрещиванием
- Одинаковые частоты аллелей у мужских и женских особей
- Не перекрывающиеся поколения
- Биаллельные (аутосомные) локусы
- ~~Размер популяции бесконечный~~
- Отсутствует изменение частот аллелей за счет миграции, естественного отбора или мутаций
- Нет ошибок генотипирования

Случайный дрейф

Конечный размер популяции ⇒ Вариабельность выборки ⇒ Флуктуации частот аллелей ⇒ Случайный дрейф



p_0, p_1: частоты аллеля, k – число копий аллеля

$$P(k) = \binom{2N}{k} p_0^k (1-p_0)^{2N-k}$$

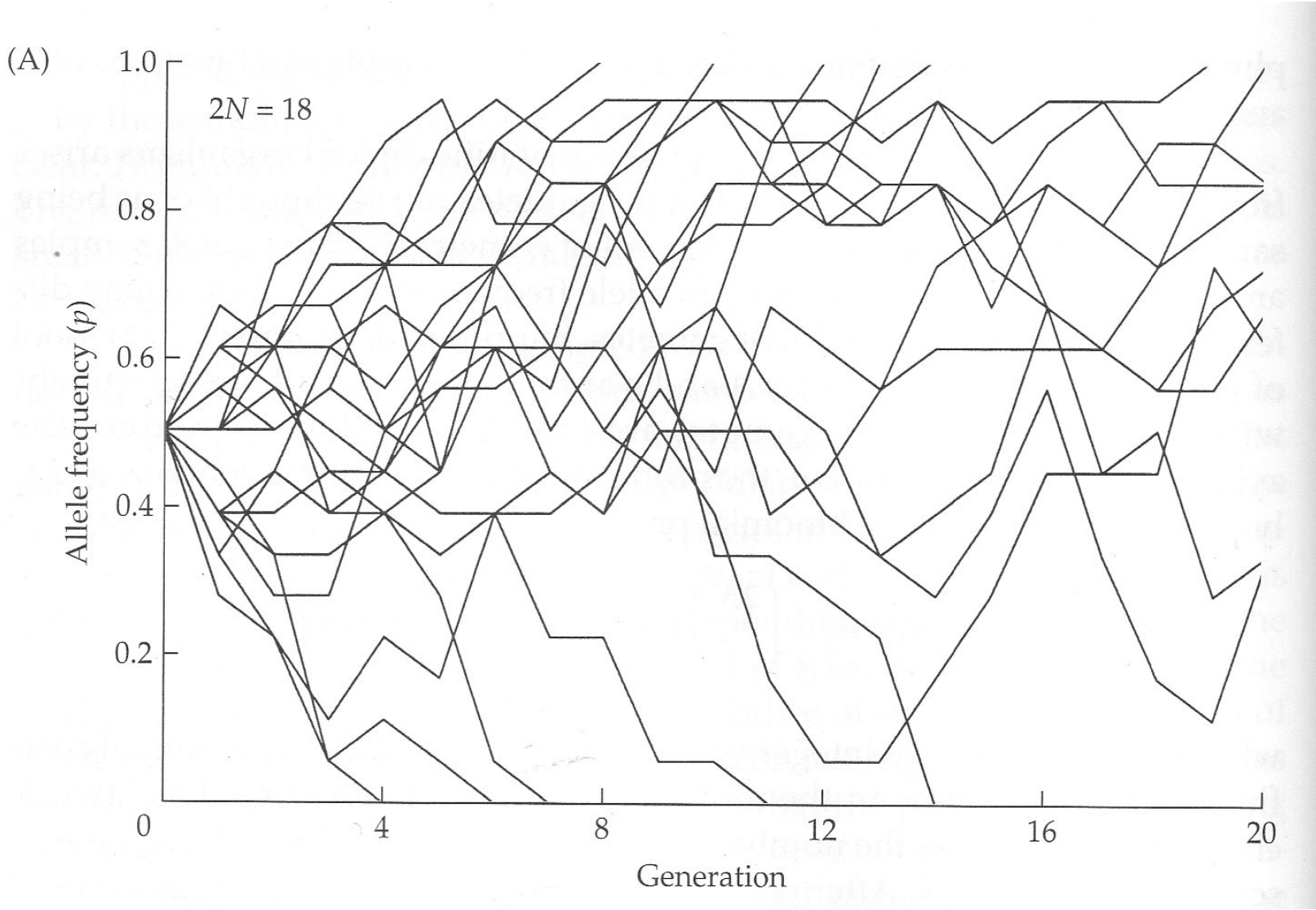
$$E(\Delta p \mid p) = E(k/2N - p \mid p) = 0$$

$$Var(\Delta p \mid p) = Var(k/2N - p \mid p) = p(1-p)/2N \quad \text{Сила дрейфа: } \approx 1/2N$$

Задача

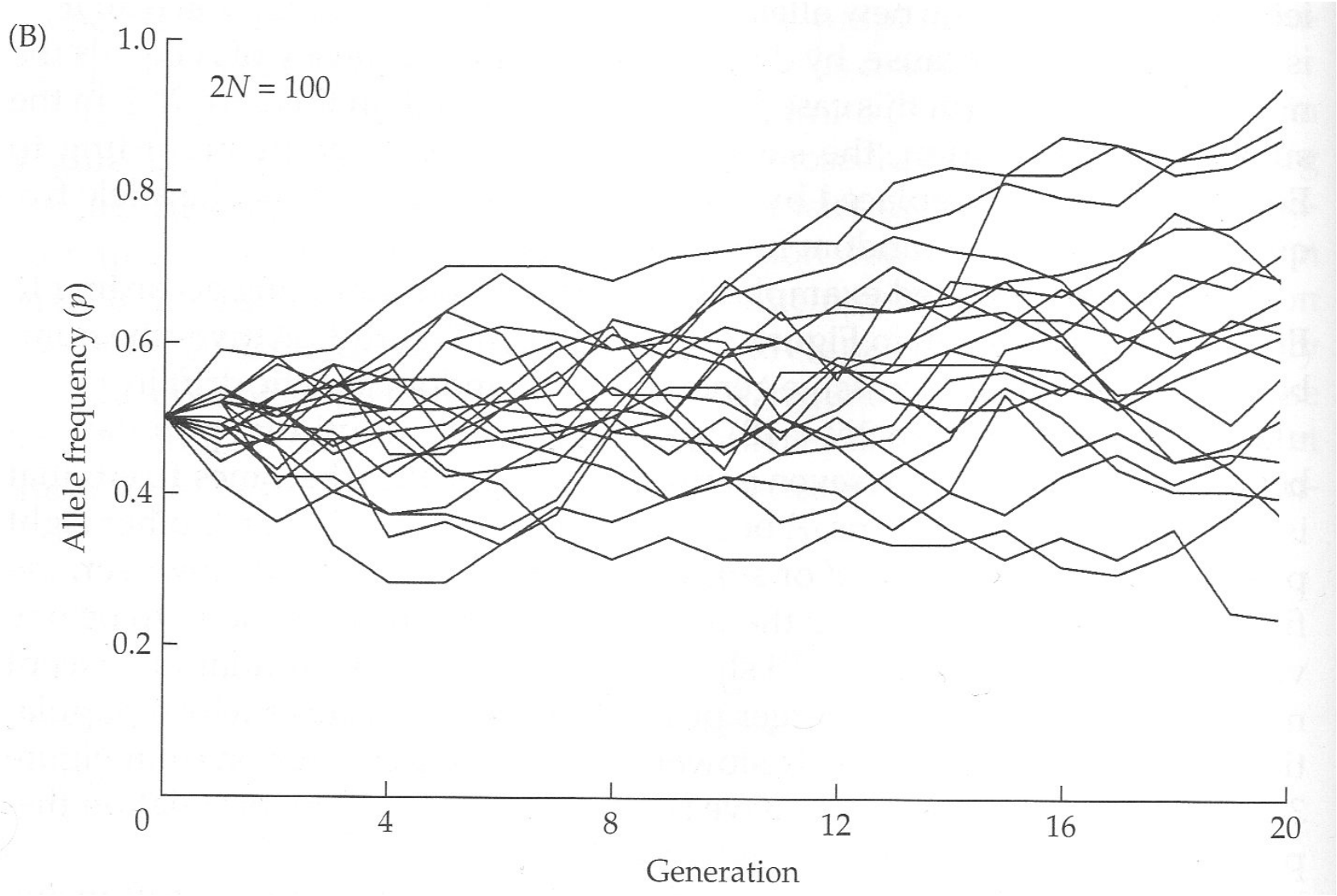
Выведите эти формулы

Случайный дрейф



Hartl & Clark – Principles of population genetics

Случайный дрейф



Hartl & Clark – Principles of population genetics

Случайный дрейф

Конечная точка дрейфа: потеря или фиксация аллеля: $P(F \mid p) = p$

Среднее время до фиксации: $\overline{t_F}(p) = -4N(\frac{1-p}{p})ln(1-p)$

Среднее время до потери: $\overline{t_L}(p) = -4N(\frac{p}{1-p})ln(p)$

Среднее время существование аллеля:
$\overline{t}(p) = p \times \overline{t_F}(p) + (1-p) \times \overline{t_L}(p) = -4N[(1-p)ln(1-p) + pln(p)]$

Задача

1. При каком p время существования аллеля максимально и чему оно равно?
2. Оцените $\overline{t_F}(p)$ при $p \to 0$.

«Зачем это нужно?»

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*]

Millions of human genomes and exomes have been sequenced, but their clinical applications remain limited due to the difficulty of distinguishing disease-causing mutations from benign genetic variation. Here we demonstrate that common missense variants in other primate species are largely clinically benign in human, enabling pathogenic mutations to be systematically identified by the process of elimination. Using hundreds of thousands of common variants from population sequencing of six non-human primate species, we train a deep neural network that identifies pathogenic mutations in rare disease patients with 88% accuracy and enables the discovery of 14 new candidate genes in intellectual disability at genome-wide significance. Cataloging common variation from additional primate species would improve interpretation for millions of variants of uncertain significance, further advancing the clinical utility of human genome sequencing.

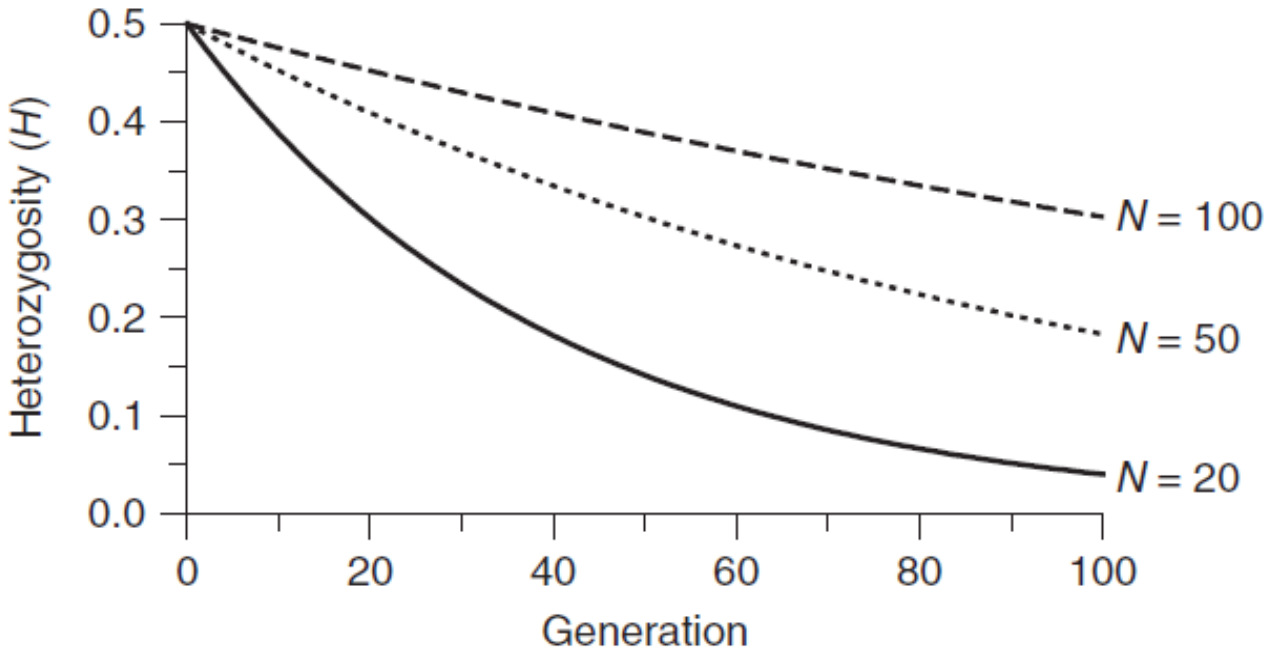
Outside of modern human populations, chimpanzees comprise the next closest extant species, and share 99.4% amino acid sequence identity[10]. The near-identity of protein-coding sequence in humans and chimpanzees suggests that purifying selection operating on chimpanzee protein-coding variants might also model the consequences on fitness of human mutations that are identical-by-state. Because the mean time for neutral polymorphisms to persist in the ancestral human lineage (~$4N_e$ generations) is a fraction of the species' divergence time (~6 mya)[11], naturally occurring chimpanzee variation explores mutational space that is largely non-overlapping except by chance, aside from rare instances of haplotypes maintained by balancing selection[12,13]. If polymorphisms that are identical-by-state similarly affect fitness in the two species, the presence of a variant at high allele frequencies in chimpanzee populations should indicate benign consequence in human, expanding the catalog of known variants whose benign consequence has been established by purifying selection.

Случайный дрейф и генетическая изменчивость

Гетерозиготность: вероятность того, что индивидуум является гетерозиготным по локусу: $H = 2pq$

Убывание гетерозиготности за счет случайного дрейфа:
$H_{t+1} \simeq H_t - H_t/2N \Rightarrow H_t = H_0(1 - 1/2N)^t$ Сила дрейфа: $\approx 1/2N$

Убывание гетерозиготности происходит медленно (но необратимо!):
$H_t = H_0/2: \ \ t \approx 2N ln(2)$ для $N \gg 1$



People living on Earth

7,849,058,679

All on this page, one by one

watch as we increase

Эффективный размер популяции

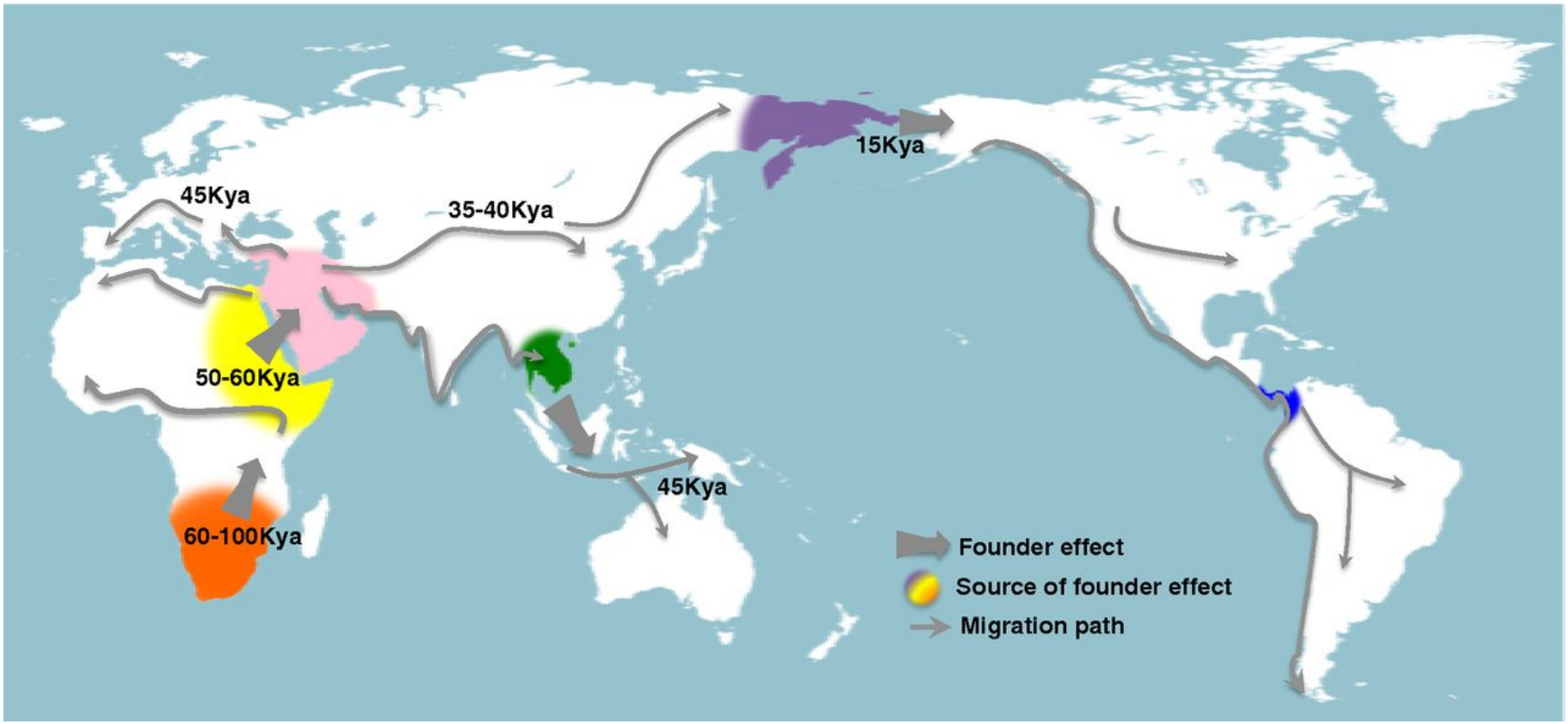
Эффективный размер популяции в случае реальной популяции – количество индивидуумов в теоретически идеальной популяции, для которых сила случайного дрейфа будет такой же, как и в реальной популяции (Hartl & Clark, *Principles of population genetics*)

- Флуктуация размеров популяции: $\frac{1}{N_e} = \frac{1}{t}(\frac{1}{N_0} + \frac{1}{N_1} + ... + \frac{1}{N_{t-1}})$
- Разность числа особей обоих полов: $N_e = \frac{4N_mN_f}{N_M+N_f}$
- Вариация числа потомков: σ, ξ – среднее и дисперсия количества потомков $N_e = \frac{N-1}{(\sigma^2/\xi)+(\xi-1)}$
- Разделение на субпопуляции: d субпопуляций размера N; m – сила миграции $N_e = Nd(1 + \frac{1}{4Nm})$

Задача

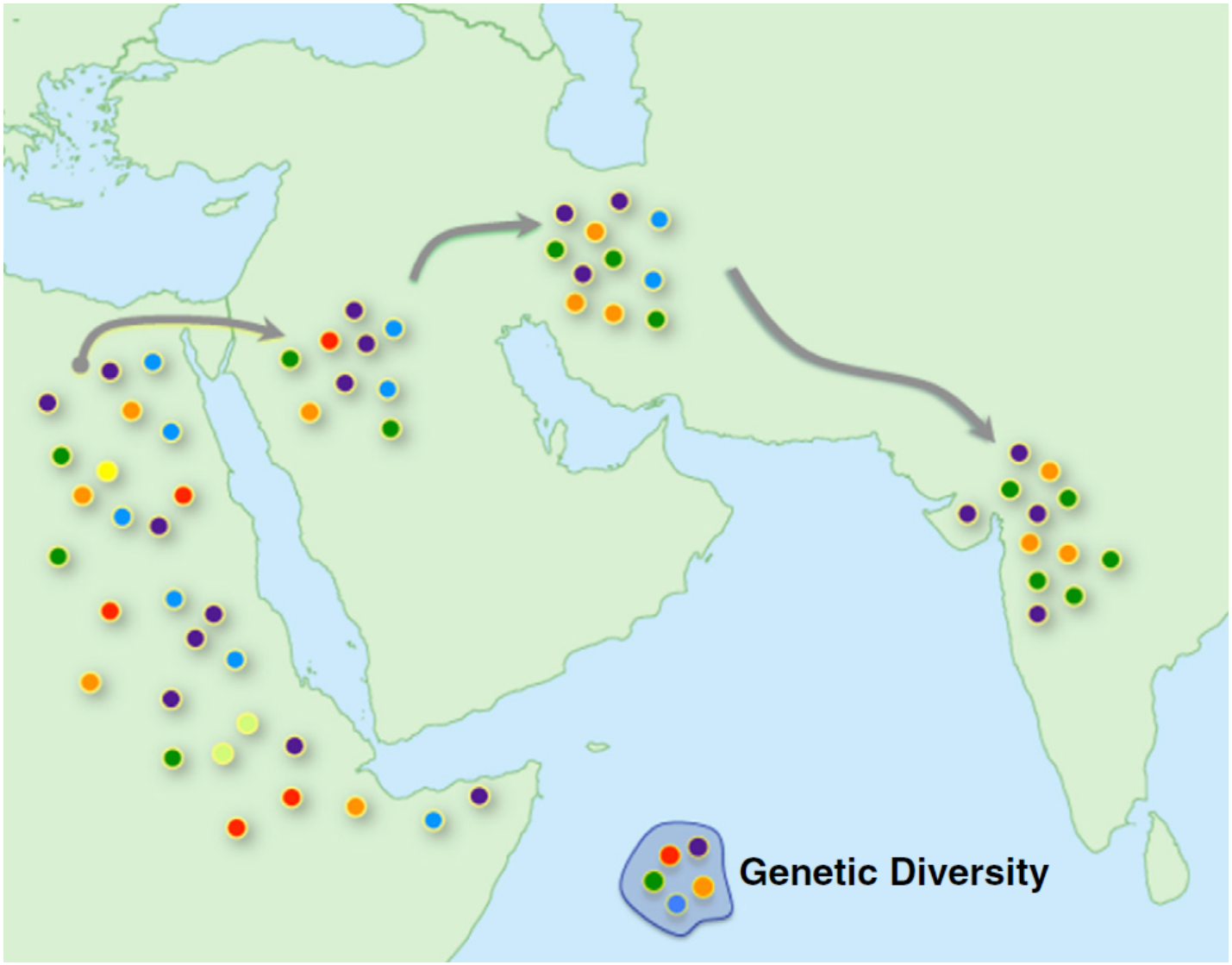
Оцените влияние «бутылочного горлышка» (флуктуации размера) на N_e

Расселение человечества и «бутылочные горлышки»



В исследовании оценили эффективный размер предковой популяции как 12,800-14,400, с 5-10 кратным эффектом бутылочного горлышка примерно 50,000-65,000 лет назад (Henn *et al* (2012) PNAS)

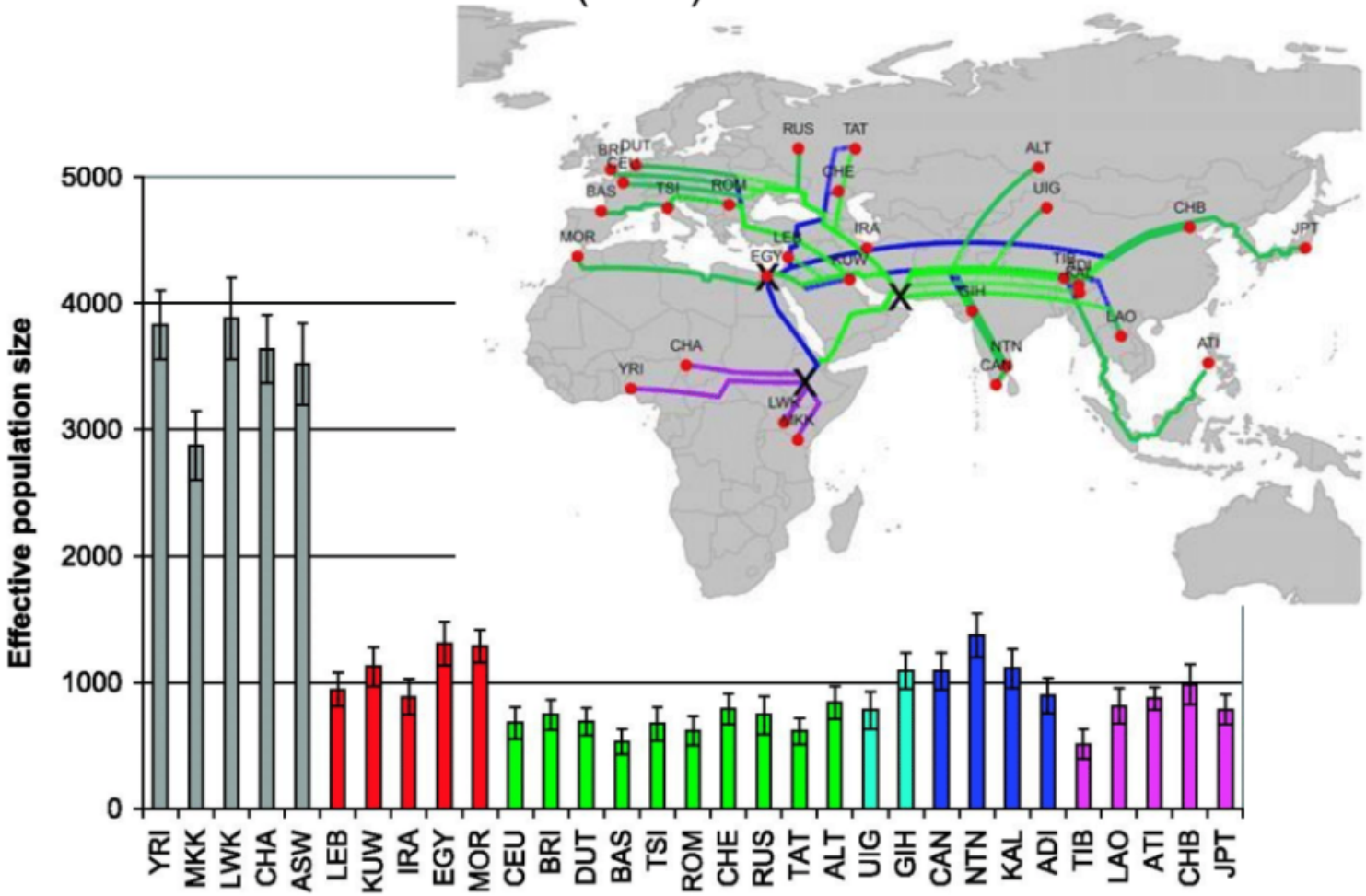
Расселение человечества и «бутылочные горлышки»



Henn *et al* (2012) PNAS

Расселение человечества и «бутылочные горлышки»

Recombination Gives a New Insight in the Effective Population Size and the History of the Old World Human Populations Mele *et al* (2011) *Mol Biol Evol*



Случайный дрейф и мутации

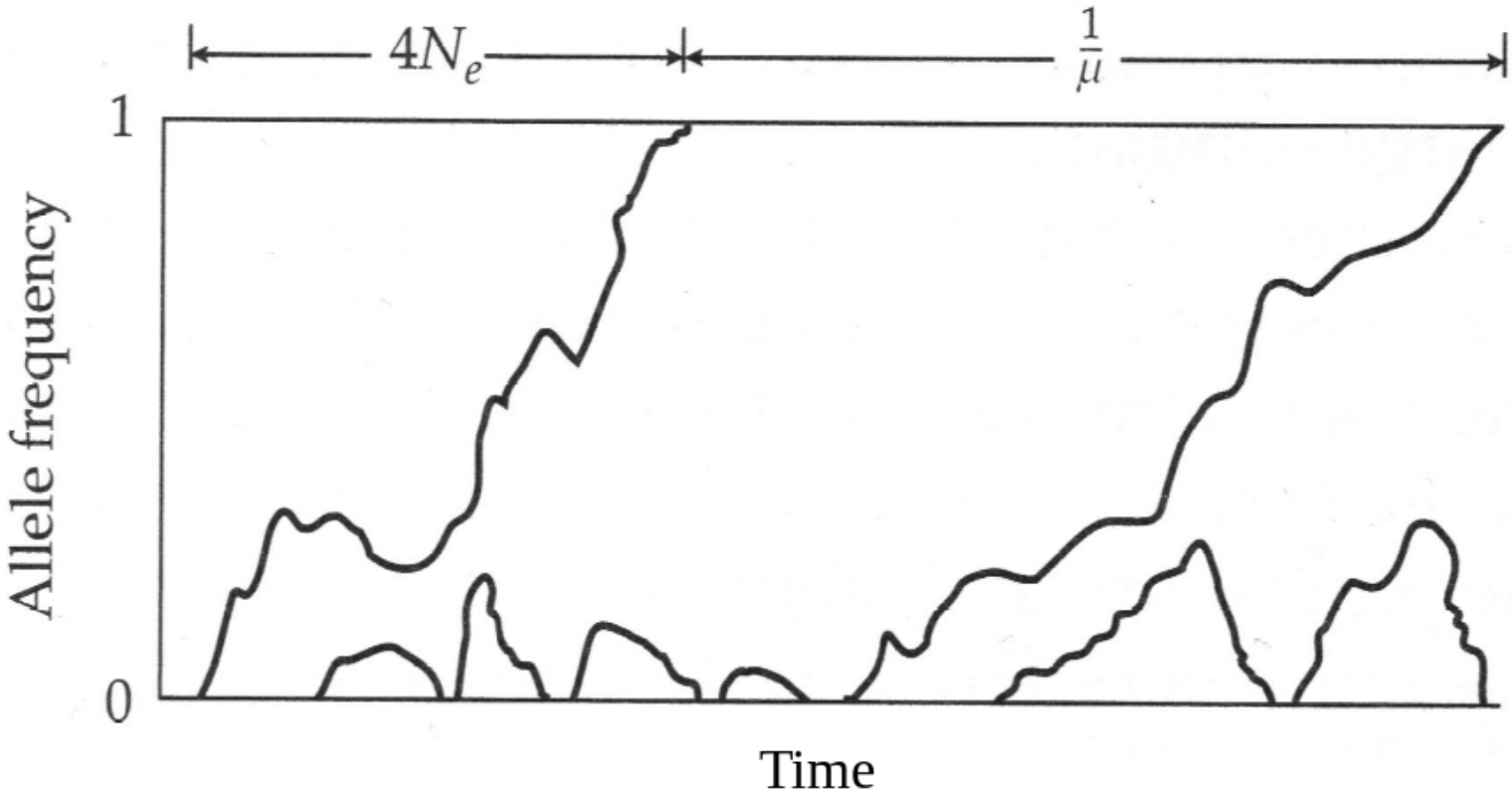
Нейтральная теория (Кимура, 1968): большинство мутаций селективно нейтрально; частота их аллелей определяется случайным дрейфом.

$2N$ гамет $\Rightarrow 2N\mu$ мутаций в каждом поколении, где μ – число мутаций на 1 гамету на 1 поколение.

Частота каждой мутации $p_0 = 1/2N \Rightarrow P_{Fix} = 1/2N$ (см. выше)

Частота фиксации нейтральных мутаций: $k = 2N\mu \times P_{Fix} = \mu$
Среднее время фиксации, при фиксации: $\overline{t_F}(p) = 4N_e$ для $p \approx 0$

Случайный дрейф и мутации



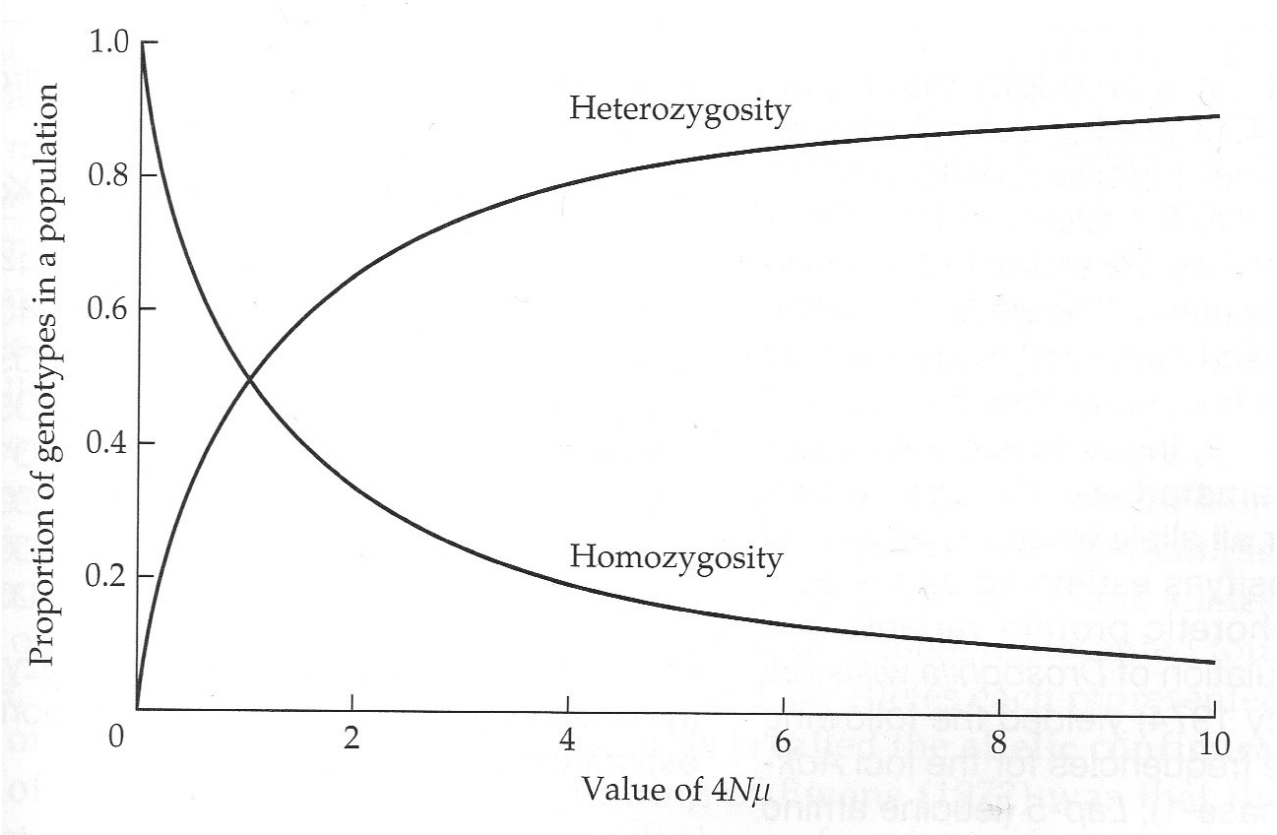
Hartl & Clark – *Principles of population genetics*

Частота фиксации нейтральных мутаций: $k = 2N\mu \times P_{Fix} = \mu$
Среднее время фиксации, при фиксации: $\overline{t_F}(p) = 4N_e$ для $p \approx 0$

Случайный дрейф и мутации

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

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



Hartl & Clark – *Principles of population genetics*

Случайный дрейф и мутации

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: частота мутаций на 1 сайт на 1 поколение, $\sim 1.2 \cdot 10^{-8}$

$$\theta = 4 \cdot 10^4 \cdot 1.2 \cdot 10^{-8} \approx 5 \cdot 10^{-4}$$
$$\theta \ll 1 \Rightarrow H \approx \theta = 1/2000$$

Случайный дрейф и мутации

Нейтральная (Мотоо Кимура) и почти нейтральная (Томоко Ота) теория молекулярной эволюции (1960-1970):

- Источником полиморфизма генома является случайный дрейф [почти] нейтральных аллелей, а не балансирующий отбор.
- Большинство замен (фиксаций) происходят из-за случайного дрейфа, а не из-за мутаций, повышающих приспособленность.
- Отсутствующие замены эволюционно запрещены.