Mutations in time:

Some basics of population genetics

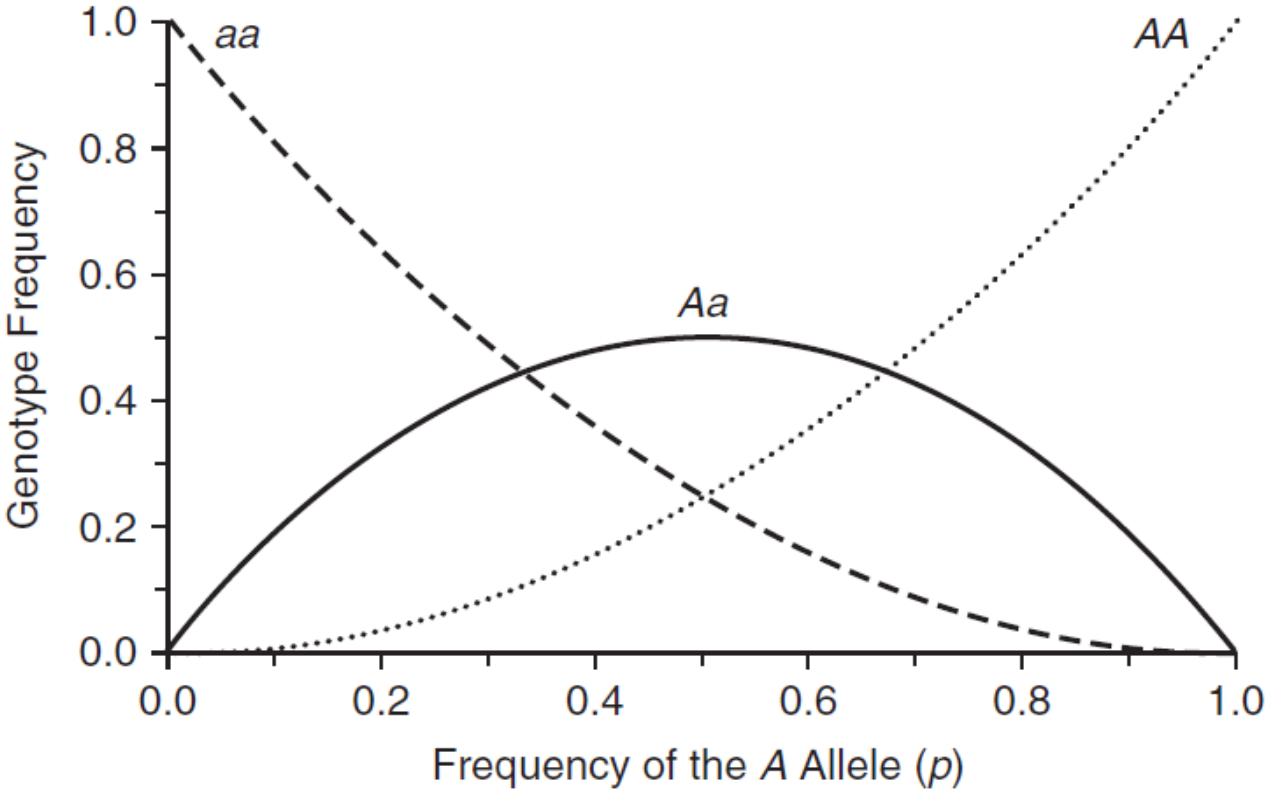
Lecture plan

- Hardy-Weinberg equilibrium
- Random genetic drift without mutations
- Effective population size
- Random genetic drift and mutations
- The coalescent theory
- Natural selection. Mutation-selection balance
- Random genetic drift, positive selection
- Selection coefficients, deleterious alleles
- Non-random mating, population subdivision, gene flow, admixture, adaptation

Hardy-Weinberg equilibrium (1908)

$$\text{Generation } N: \quad f_A = p, \; f_a = q, \; p + q = 1$$

$$\text{Generation } N+1: \quad F_{AA} = p^2, \; F_{Aa} = 2pq, \; F_{aa} = q^2$$



Hardy-Weinberg equilibrium

$$\text{Generation } N: \quad f_A = p, \; f_a = q, \; p + q = 1$$

$$\text{Generation } N+1: \quad F_{AA} = p^2, \; F_{Aa} = 2pq, \; F_{aa} = q^2$$

Implications:

1. The allele frequencies does not change:

$$p' = f'_A = F'_{AA} + F'_{Aa}/2 = p^2 + pq = p$$

Exercise: derive this

2. HWE frequencies are attained in one generation

Hardy-Weinberg equilibrium

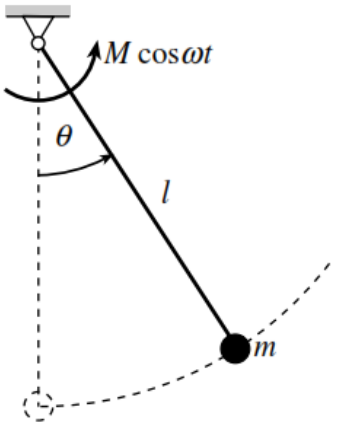
Assumptions:

- Diploid species with sexual reproduction and random (not assortative) mating
- Same allele frequencies in males and females
- Non-overlapping generations
- Biallelic (autosomal) locus
- Population size is infinite
- No change in allele frequencies by migration, natural selection or mutation
- No genotyping errors

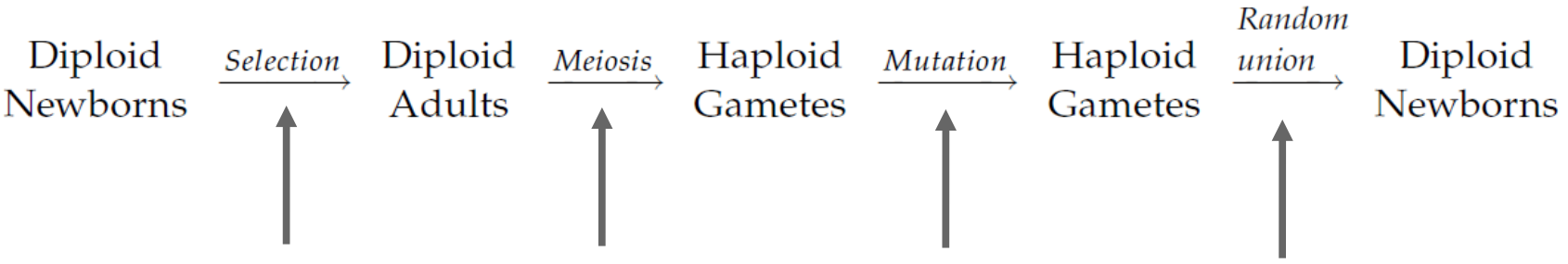
Hardy-Weinberg equilibrium

Does it still make sense with so many assumptions? Yes:

1. A baseline for more realistic models



2. The H-W model splits life history into two intervals: gametes → zygotes and zygotes → adults



Hardy-Weinberg equilibrium

Testing for HWE:

df = $n - k - 1$, where $n = 3$ is the number of classes

and $k = 1$ is the number of independent parameters

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.

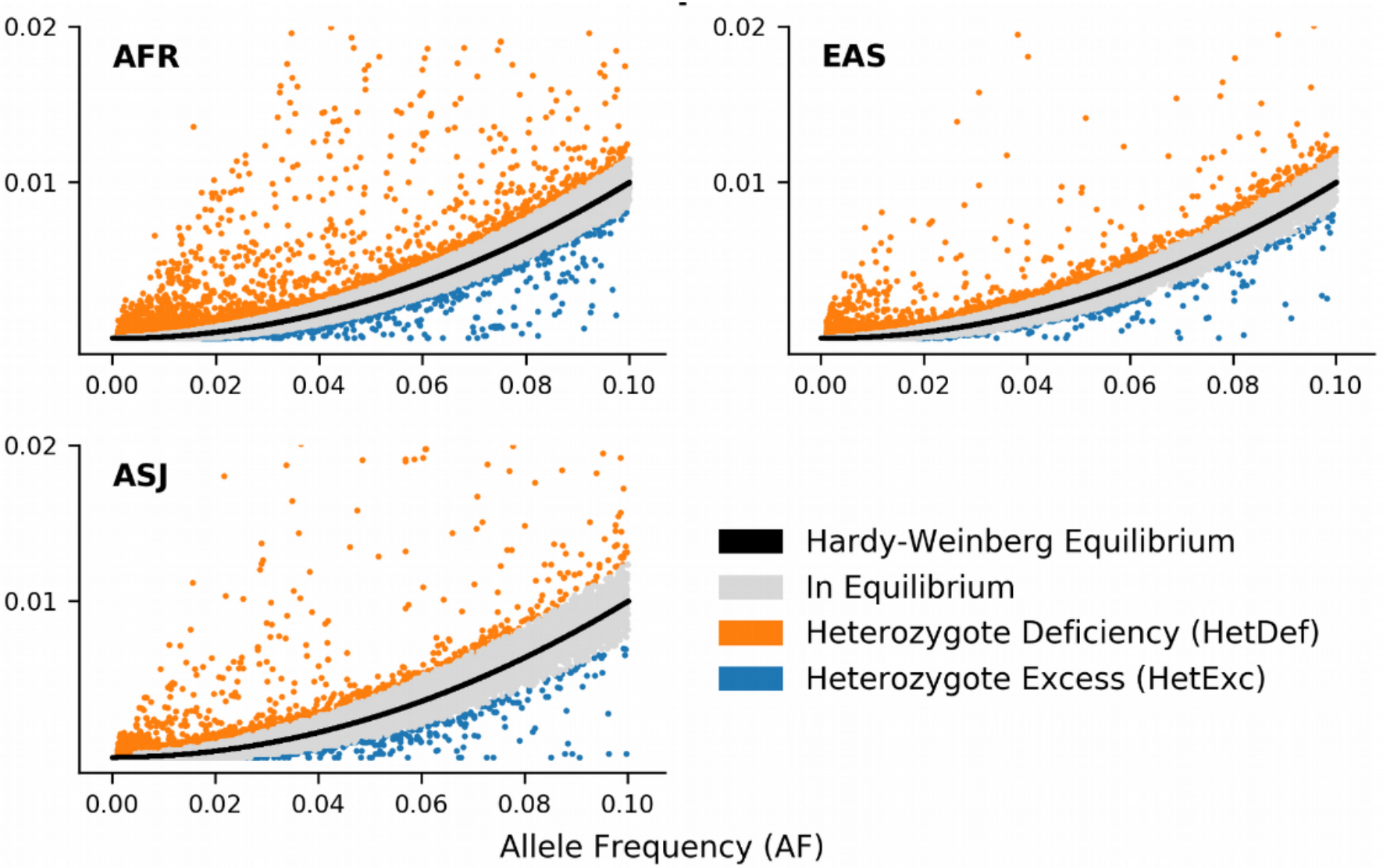
Exercise: do it yourself

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era

Nikita Abramovs, Andrew Brass, May Tassabehji

doi: https://doi.org/10.1101/859462



gnomAD: 137,842 predominantly healthy individuals from 7 major ethnic populations

Random genetic drift (Wright-Fisher, 1930)

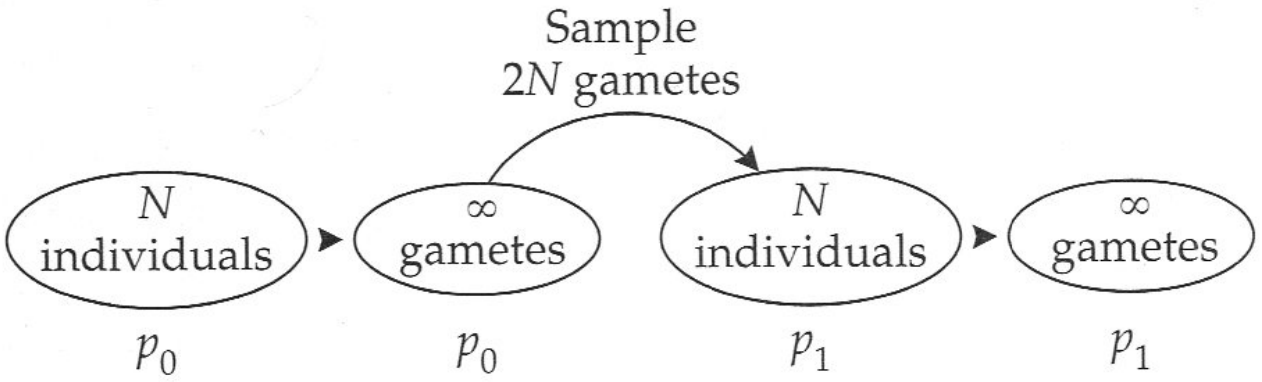
Assumptions:

- Diploid species with sexual reproduction and random (not assortative) mating
- Same allele frequencies in males and females
- Non-overlapping generations
- Biallelic (autosomal) locus
- ~~Population size is infinite~~
- No change in allele frequencies by migration, natural selection or mutation
- No genotyping errors

Random genetic drift

Finite population $\Longrightarrow$ Sampling variation $\Longrightarrow$

Allele frequency fluctuations $\Longrightarrow$ Random genetic drift



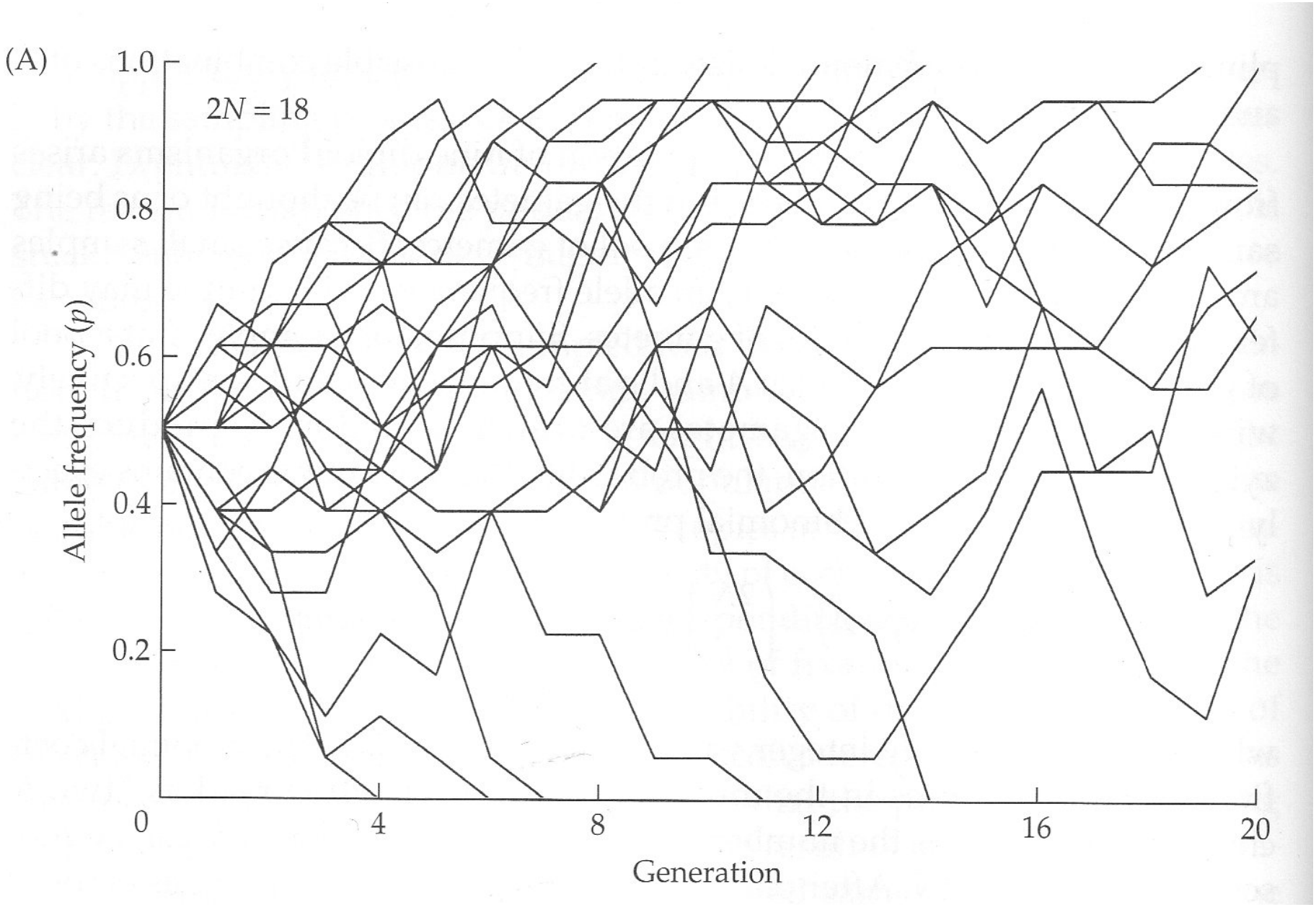
$$P(k) = \binom{2N}{k} p_0^k (1-p_0)^{2N-k}$$

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

Exercise: derive

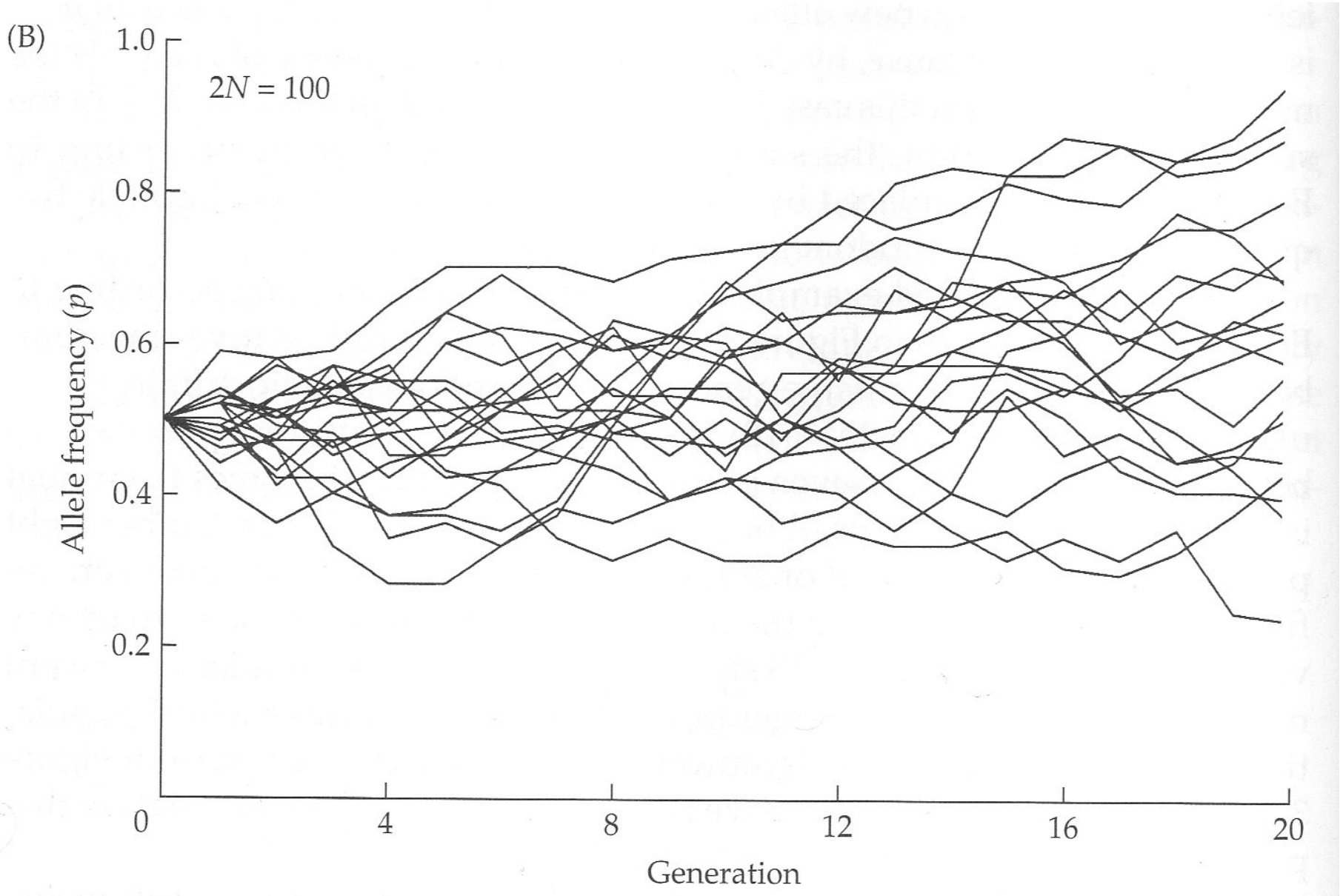
$$Var(\Delta p \,|\, p) = Var(k/2N - p \,|\, p) = \; p(1-p)/2N$$

Random genetic drift



Hartl & Clark – *Principles of population genetics*

Random genetic drift



Hartl & Clark – *Principles of population genetics*

Random genetic drift

The endpoint is allele fixation or loss: $P(F|p) = p$

Mean time to fixation, if fixed: $\overline{t_F}(p) = -4N\left(\frac{1-p}{p}\right) ln(1-p)$

Mean time to loss, if lost: $\overline{t_L}(p) = -4N\left(\frac{p}{1-p}\right) ln(p)$

Mean perstistence time: $\bar{t}(p) = p\overline{t_F}(p) + (1-p)\overline{t_L}(p) =$

$= -4N[(1-p)ln(1-p) + p \cdot ln(p)]$

Exercise: at which *p* persistence time is maximal and what is it?

Exercise: estimate $t_F(p)$ when $p \to 0$

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.

Random genetic drift and genetic variation

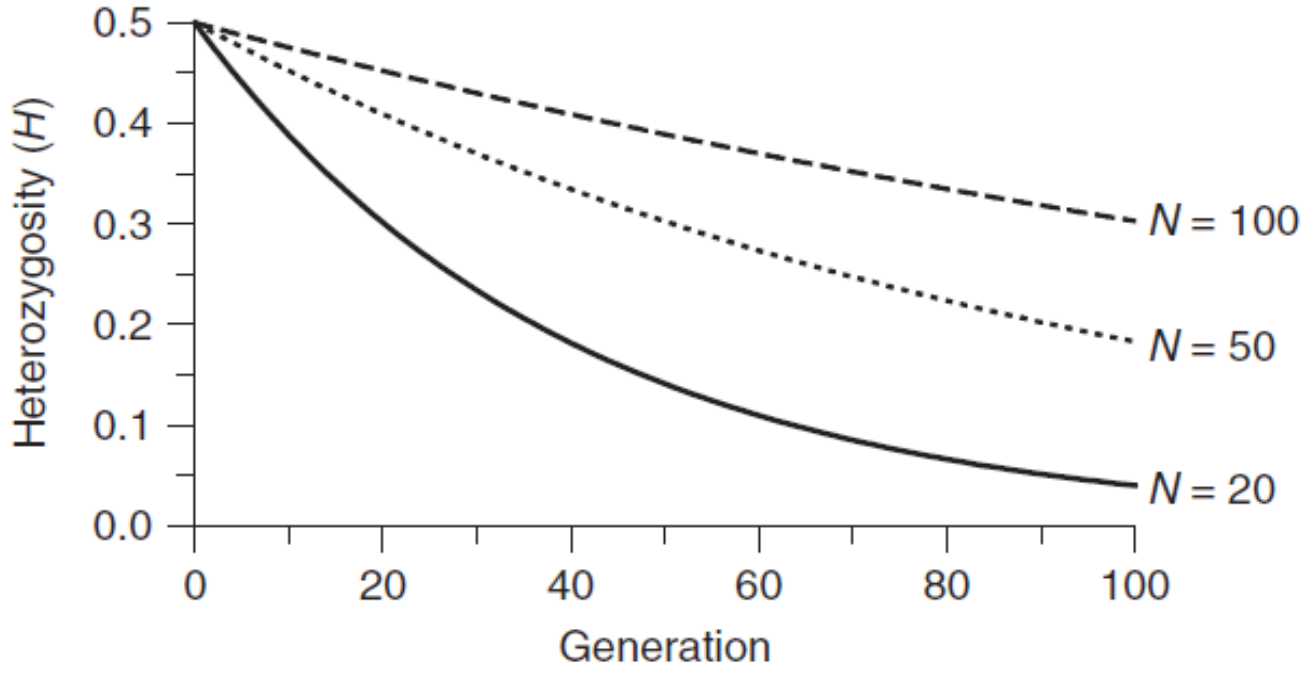
Heterozygosity: probability that an individuum is heterozygous at a locus: $H = 2pq$

$$H_{t+1} \simeq H_t - H_t/2N$$

Heterozygosity decay due to drift:

$$H_t = H_0(1 - 1/2N)^t$$

Decay is slow: $H_t = H_0/2 : \; t \approx 2N \ln(2) \; for \; N \gg 1$



Random genetic drift and genetic variation

Heterozygosity: probability that an individuum is heterozygous at a locus: $H = 2pq$

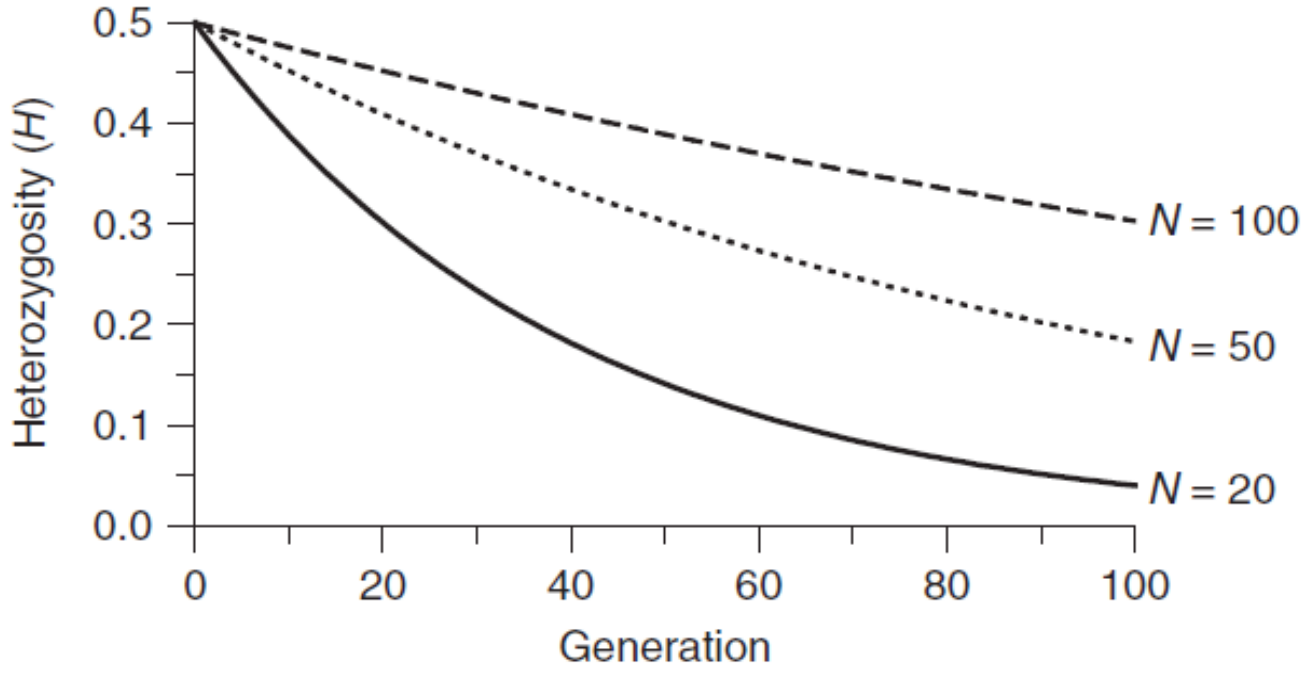
Drift strength is ≈1/2*N*

$$H_{t+1} \simeq H_t - H_t/2N$$

Heterozygosity decay due to drift:

$$H_t = H_0(1 - 1/2N)^t$$

Decay is slow: $H_t = H_0/2 : \ t \approx 2N \ln(2) \ for \ N \gg 1$



Effective population size

Effective population size of an actual population is the number of indivduals in a theoretically ideal population having the same magnitude of genetic drift as the actual population (Hartl & Clark, *Principles of population genetics*)

- Fluctuation in population size $\frac{1}{N_e} = \frac{1}{t}(\frac{1}{N_0} + \frac{1}{N_1} + \cdots + \frac{1}{N_{t-1}})$
- Unequal sex ratio: $N_e = \frac{4N_m N_f}{N_m + N_f}$

Exercise: bottleneck consequences for N_e

- Variance in offspring number:
σ, ξ – offspring mean and variance

$$N_e = \frac{N-1}{(\sigma^2/\xi) + (\xi - 1)}$$

- Subdivided population:
d sub-populations of size *N*; *m*, migration

$$N_e = Nd\left(1 + \frac{1}{4Nm}\right)$$