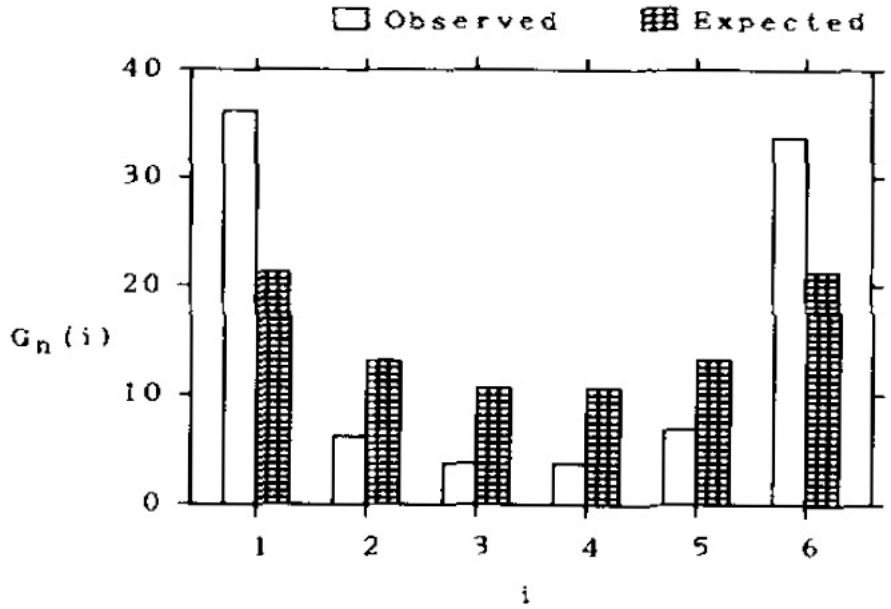
Intro to Population Genetics

Mutation, drift, and the nearly neutral theory of molecular evolution



(Effective) population size is the single largest determinant of evolutionary fate

Census population size (N_c) – Number of individuals in a population

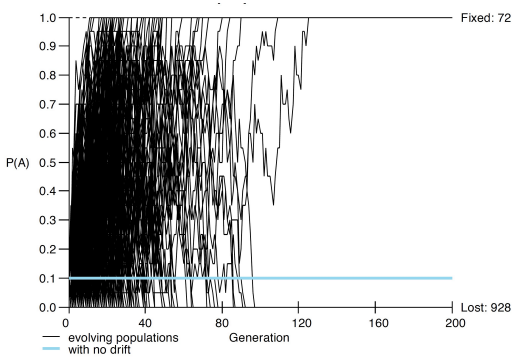
Effective population size (N_e) – Number of individuals in an idealized population that would explain the amount of variation in the population

N_e = # of genetically different individuals contributing to next generation

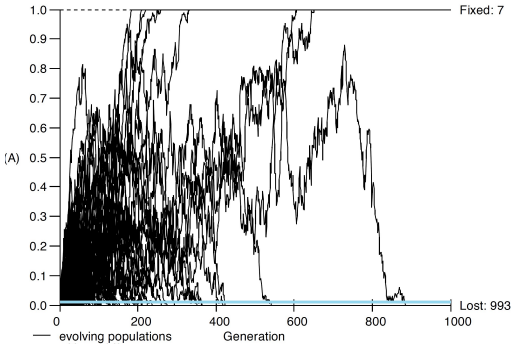
(Effective) population size is the single largest determinant of evolutionary fate

(Effective) population size is the single largest determinant of evolutionary fate

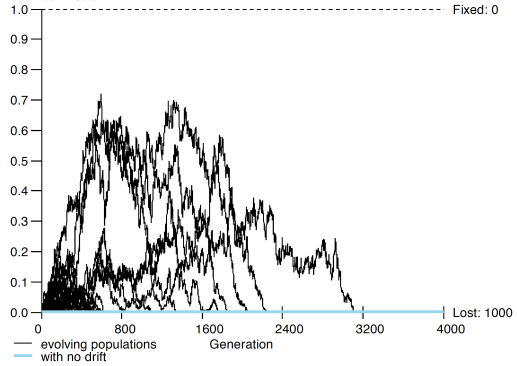
N=10



N=100



N=1000



What happened?

- All mutations start out at freq = $\frac{1}{Ne}$

  ($\frac{1}{2Ne}$ in diploid populations)

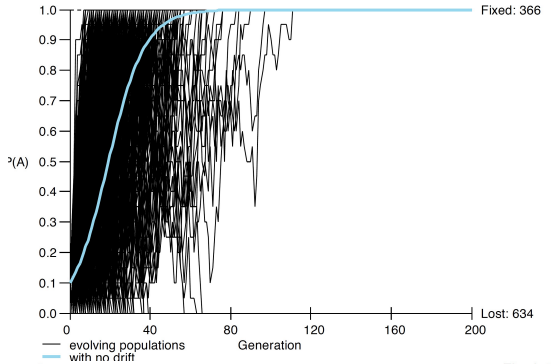
- In the absence of selection, P_{fix} = Freq

- Fixation/loss takes longer in larger populations (drift has less of an impact per generation)

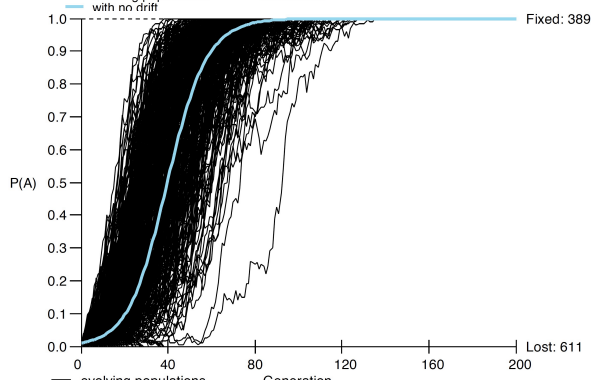
Beneficial mutations more likely to fix in larger populations

Beneficial mutations more likely to fix in larger populations

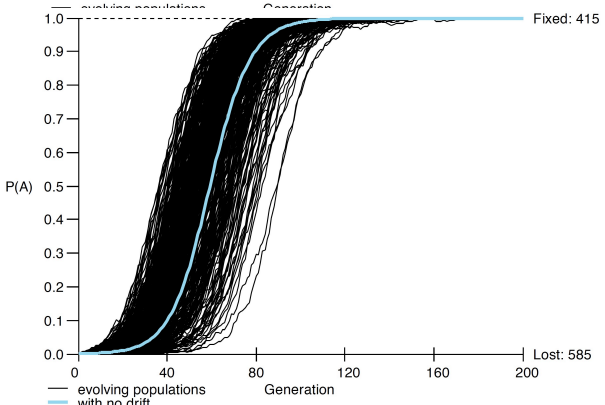
N=10



N=100



N=1000

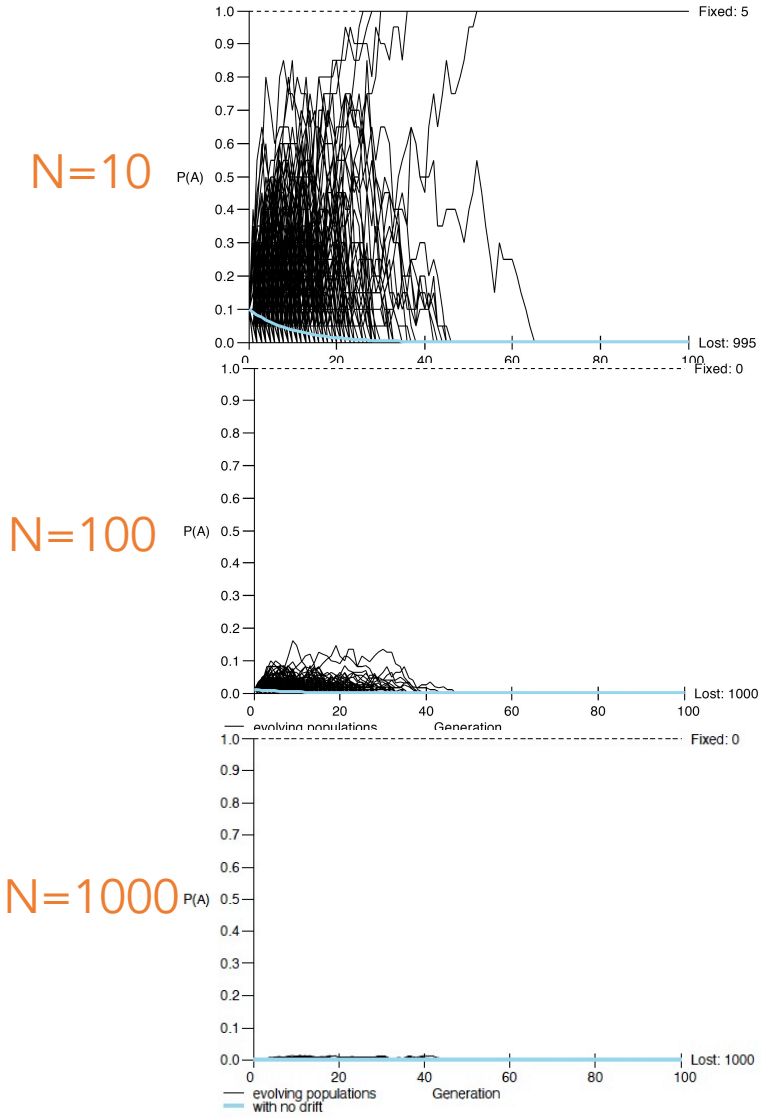


What happened?

- Most mutations go extinct immediately, even when they are beneficial (improve fitness)
- Selection is more effective in larger populations
- P_{fix} ~ selection coefficient (s = 0.1)

Harmful mutations less likely to fix in larger populations

Harmful mutations less likely to fix in larger populations



What happened?

- Harmful mutations can still fix in small populations
- Selection is more effective in larger populations
- P_{fix} inversely proportional to population size

Probability of fixation depends upon

- allele frequency
- selective co-efficient (s)
- degree of dominance (h)
- effective population size (Ne)

Mutations can therefore be fixed by drift (neutral, slightly deleterious)

Or by selection (beneficial, linkage)

Some Hardy Weinberg…

$1 = p + q$ TRUISM

$1 = p^2 + 2pq + q^2$ HWE Null Hypothesis

Hardy-Weinberg Equilibrium is not a good null model

HWE Assumptions:

- Infinite population size
- No mutation
- No selection
- No migration/gene flow
- Random mating

Hardy-Weinberg Equilibrium is not a good null model

HWE Assumptions:

- Infinite population size
- [illegible]
- Real populations cannot
- EVER meet these criteria
- Random mating

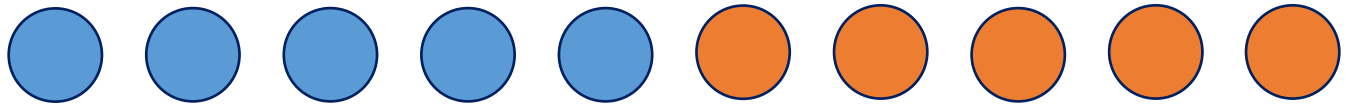
Genetic drift (random subsampling) is a feature of all finite populations

Genetic drift is the change in allele frequencies across generations that is due to random subsampling of the previous generation

Eventually results in fixation of one allele absent some other evolutionary force

Magnitude of allele frequency change due to drift is greatest in small populations, but probability of drift occurring in a single generation is higher in large populations

Genetic drift thought experiment



5 blue circles

5 orange circles

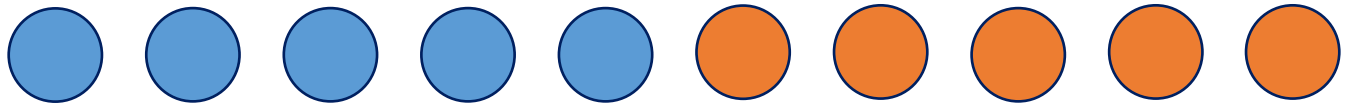
Constant population size

P(blue) = 0.5

P(orange) = 0.5

What's the probability of observing exactly 5 blue circles and 5 orange circles in the next generation?

Genetic drift thought experiment



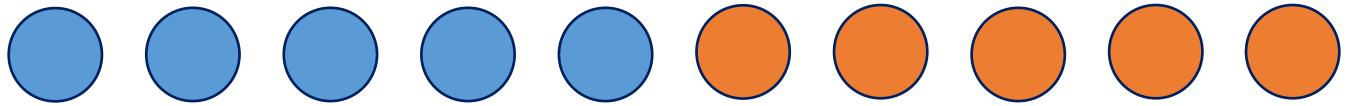
5 blue circles

5 orange circles

Constant population size

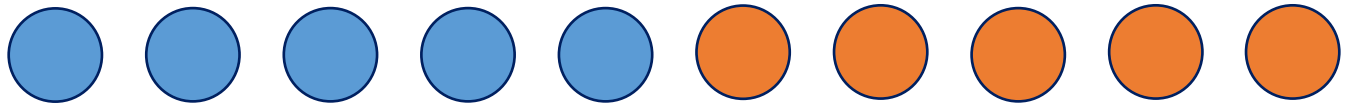
P(blue) = 0.5

P(orange) = 0.5



24.61% chance 5 blue/5 orange, 75.39% chance something else

Genetic drift thought experiment



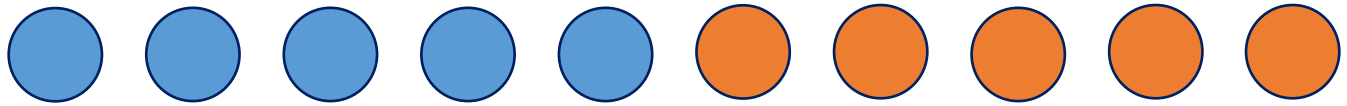
500 blue circles

500 orange circles

Constant population size

P(blue) = 0.5

P(orange) = 0.5



2.52% chance 500 blue/500 orange, 97.48% chance something else

Genetic drift (random subsampling) is a feature of all finite populations

Genetic drift is the change in allele frequencies across generations that is due to random subsampling of the previous generation

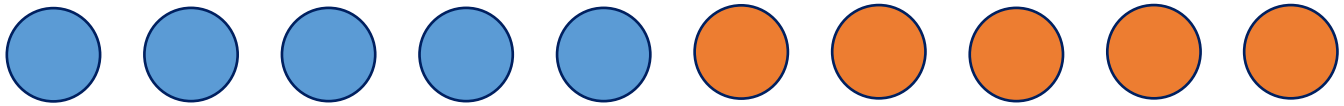
Eventually results in fixation of one allele absent some other evolutionary force

Magnitude of allele frequency change due to drift is greatest in small populations, but probability of drift occurring in a single generation is higher in large populations

Probability of fixation depends upon allele frequency, selective co-efficient, and effective population size (*Ne*)

$$P_{fix} = \frac{1 - e^{-s}}{1 - e^{-s \bullet Ne}}$$

$$s = 0, P_{fix} = freq = \frac{1}{2Ne}$$

P_{fix} (blue) $= 0.5$  P_{fix} (orange) $= 0.5$

What is a good null model for molecular evolution?

Neutral Theory (Motoo Kimura 1968)

“This neutral theory claims that the overwhelming majority of evolutionary changes at the molecular level are not caused by selection acting on advantageous mutants, but by random fixation of selectively neutral or very nearly neutral mutants through the cumulative effect of sampling drift (due to finite population number) under continued input of new mutations"

Nearly Neutral Theory (Tomoko Ohta 1973)

-Added in that slightly deleterious changes can accumulate depending on population size (drift-barrier hypothesis)

Selectionist vs. Neutralist debate

The Neutral Theory in Light of Natural Selection

Andrew D. Kern*,[1] and Matthew W. Hahn[2]

[1]Department of Genetics, Rutgers University, Piscataway, NJ

[2]Department of Biology and Department of Computer Science, Indiana University Bloomington, IN

***Corresponding author:** E-mail: kern@biology.rutgers.edu.

Associate editor: Sudhir Kumar

Abstract

In this perspective, we evaluate the explanatory power of the neutral theory of molecular evolution, 50 years after its introduction by Kimura. We argue that the neutral theory was supported by unreliable theoretical and empirical evidence from the beginning, and that in light of modern, genome-scale data, we can firmly reject its universality. The ubiquity of adaptive variation both within and between species means that a more comprehensive theory of molecular evolution must be sought.

Key words: **natural selection, neutral theory, population genetics.**

Calculating allele frequencies

$$1 = p + q$$

$$f(p) = 1 - f(q)$$

$$f(q) = 1 - f(p)$$

Calculating allele frequencies

$1 = p + q$

$p = A,\ q = B$

$f(p) = 1 - f(q)$

$$f(p) = \frac{(2 \times \boldsymbol{AA}) + \boldsymbol{AB}}{2 \times (\boldsymbol{AA} + \boldsymbol{AB} + \boldsymbol{BB})}$$

$f(q) = 1 - f(p)$

$$f(q) = \frac{(2 \times \boldsymbol{BB}) + \boldsymbol{AB}}{2 \times (\boldsymbol{AA} + \boldsymbol{AB} + \boldsymbol{BB})}$$

Calculating allele frequencies

$1 = p + q$

$p = A,\ q = B$

30 – AA, 25 – AB, 50 – BB

$f(p) = 1 - f(q)$

$$f(p) = \frac{(2 \times AA) + AB}{2 \times (AA + AB + BB)}$$

$f(p) = ?$

$f(q) = 1 - f(p)$

$$f(q) = \frac{(2 \times BB) + AB}{2 \times (AA + AB + BB)}$$

$f(q) = ?$

Calculating allele frequencies

$1 = p + q$

$p = A,\ q = B$

30 – AA, 25 – AB, 50 – BB

$f(p) = 1 - f(q)$

$$f(p) = \frac{(2 \times AA) + AB}{2 \times (AA + AB + BB)}$$

$f(p) = 0.405$

$f(q) = 1 - f(p)$

$$f(q) = \frac{(2 \times BB) + AB}{2 \times (AA + AB + BB)}$$

$f(q) = 0.595$

Expected Heterozygosity (H_{exp})

The fraction of individuals in a population predicted to be heterozygous

$$H_{exp} = \mathbf{2pq}$$

30 – AA, 25 – AB, 50 – BB

f(p) = 0.405

f(q) = 0.595

H_{exp} = ?

Expected Heterozygosity (H_{exp})

The fraction of individuals in a population predicted to be heterozygous

$$H_{exp} = \mathbf{2pq}$$

30 – AA, 25 – AB, 50 – BB

f(p) = 0.405

f(q) = 0.595

H_{exp} *= 0. 482*

Observed Heterozygosity (H_{obs})

The fraction of individuals in a population that are heterozygous

30 – AA, 25 – AB, 50 – BB

H_{obs} = ?

$$H_{obs} = \frac{\boldsymbol{AB}}{(\boldsymbol{AA} + \boldsymbol{AB} + \boldsymbol{BB})}$$

Observed Heterozygosity (H_{obs})

The fraction of individuals in a population that are heterozygous

30 – AA, 25 – AB, 50 – BB

$$H_{obs} = \frac{\boldsymbol{AB}}{(\boldsymbol{AA} + \boldsymbol{AB} + \boldsymbol{BB})}$$

$$H_{obs} = \frac{\mathbf{25}}{\mathbf{105}} = 0.238$$

$$H_{exp} - H_{obs} = 0.482 - 0.238 = 0.244$$

Fewer heterozygotes present than expected

Observed Heterozygosity (H_{obs})

The fraction of individuals in a population that are heterozygous

30 – AA, 25 – AB, 50 – BB

Use χ^2 test* to evaluate if

$H_{obs} \neq H_{exp}$

$$H_{obs} = \frac{AB}{(AA + AB + BB)}$$

$$H_{obs} = \frac{25}{105} = 0.238$$

$H_{exp} - H_{obs} = 0.482 - 0.238 = 0.244$

* χ^2 test requires #s, not frequencies!!!

Fewer heterozygotes present than expected

χ^2 test of Independence

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

$$df = (r - 1) * (c - 1) = 1$$

χ^2 test of Independence

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

If $\chi^2 > 3.841$, we can reject H_o that $H_{obs} = H_{exp}$

Allele frequency practice:

44 red flowers (RR); 33 pink flowers (RW); 23 white flowers (WW)

Calculate:

- p and q
- H_{obs}
- H_{exp}
- Perform χ^2 test

Allele frequency practice:

44 red flowers (RR); 33 pink flowers (RW); 23 white flowers (WW)

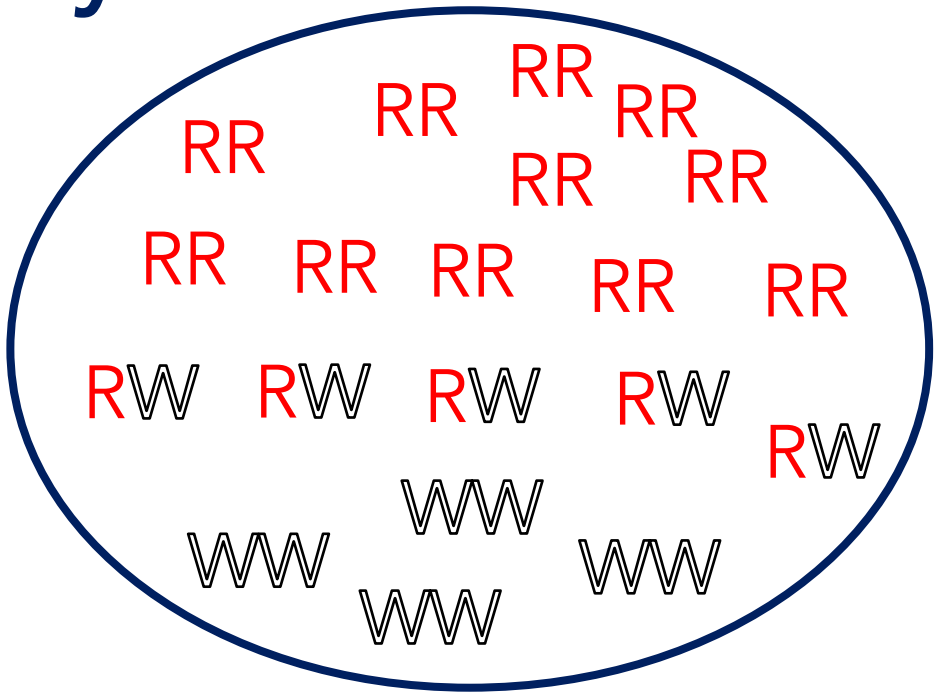
p = freq(R) = ((2*44) + 33)/200 = 0.605

q = 1 – p = 0.395

- H_{obs} = 33/100 = 0.33
- H_{exp} = $2pq$ = 2*0.605*0.395 = 0.478

Causes of low heterozygosity

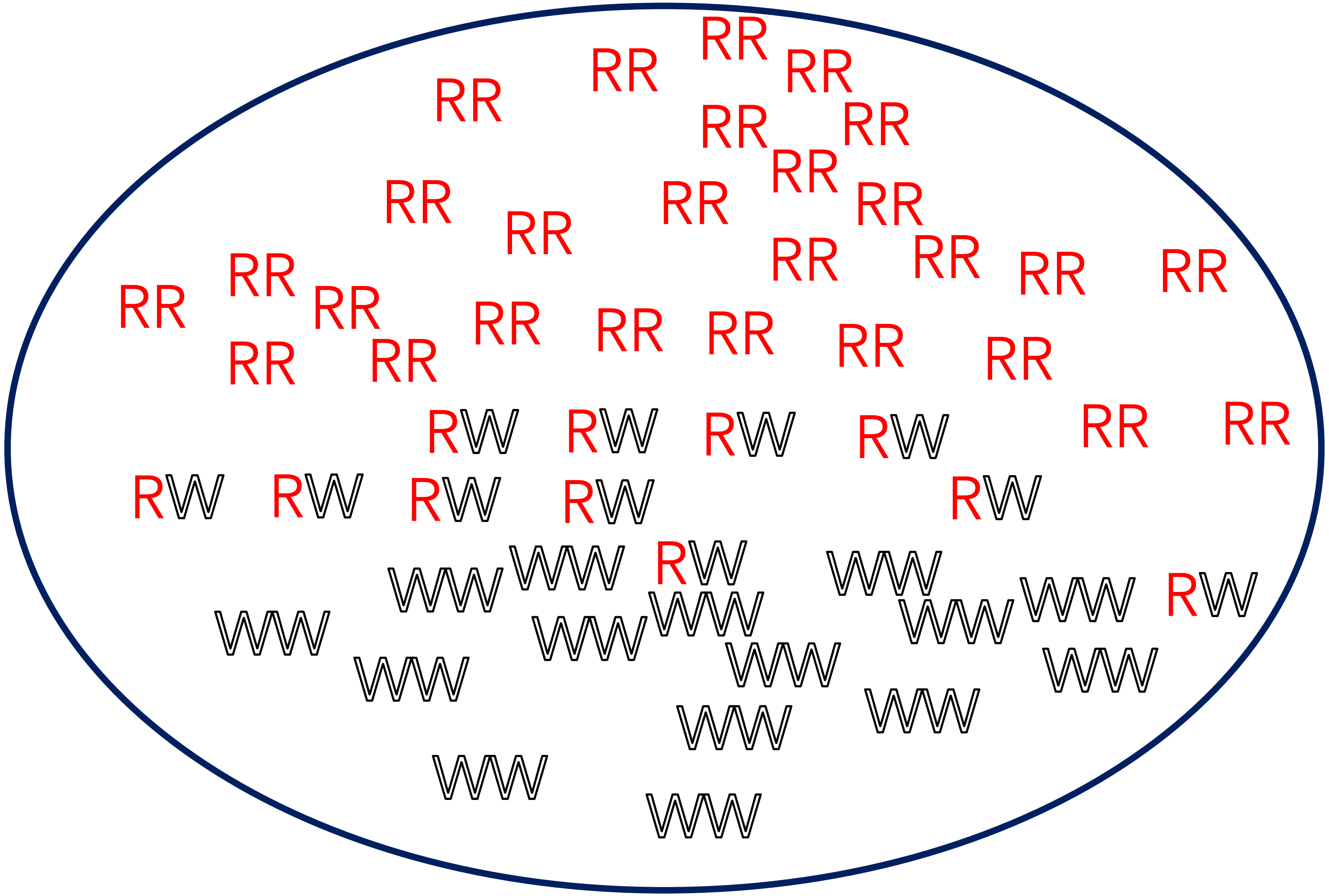
- Population structure
- Inbreeding/ selfing
- Recent population admixture
- Assortative mating
  (mating with similar individuals)
  - Heterozygote disadvantage
  - Sexual selection



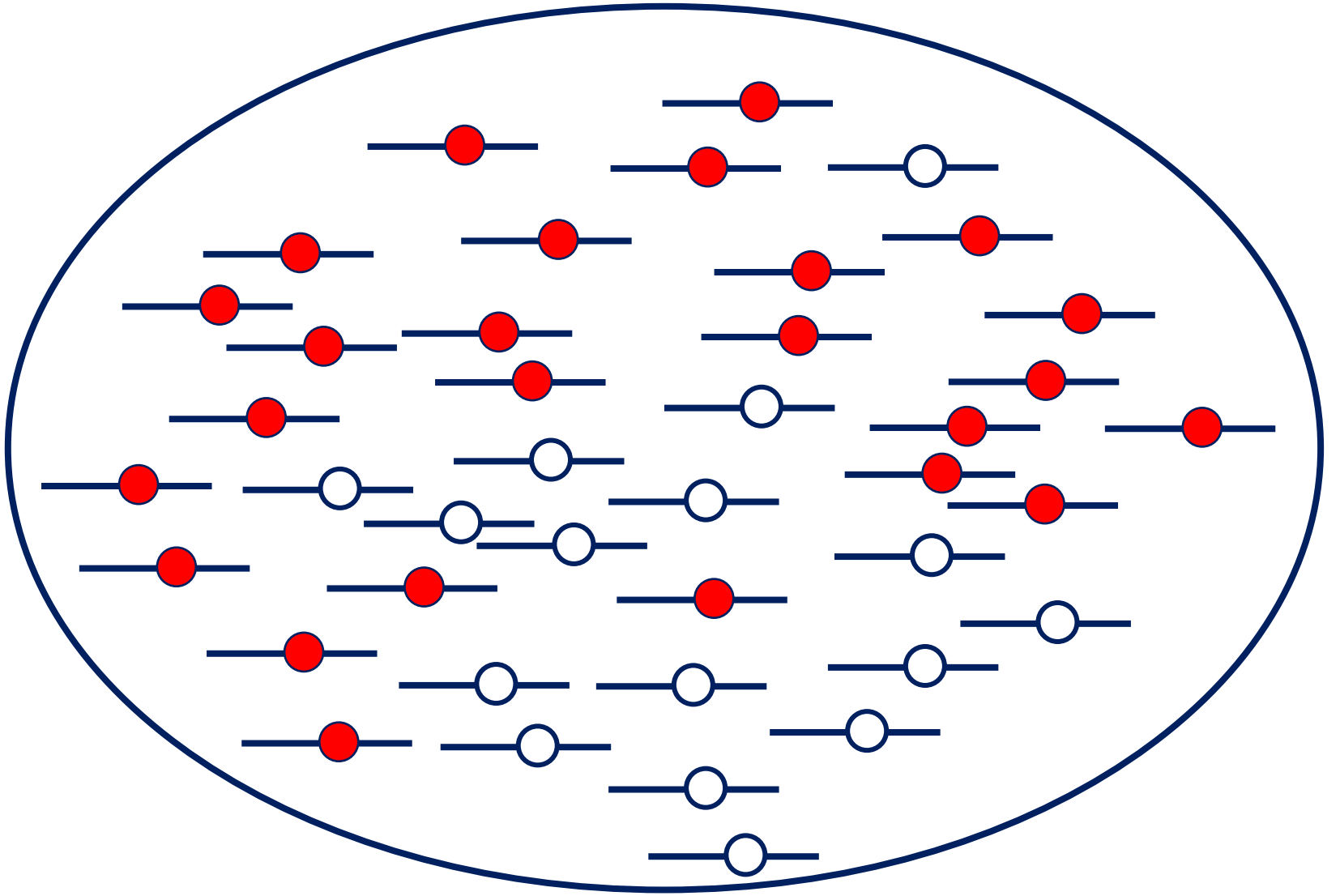
Causes of high heterozygosity

- Outcrossing
- Balancing selection
- Older population admixture
- Disassortative mating
  (mating with different individuals)
  - Heterozygote advantage
  - Sexual selection

Thinking about populations…



Thinking about populations…pop gen style



Population geneticists think about chromosomes, not about individuals

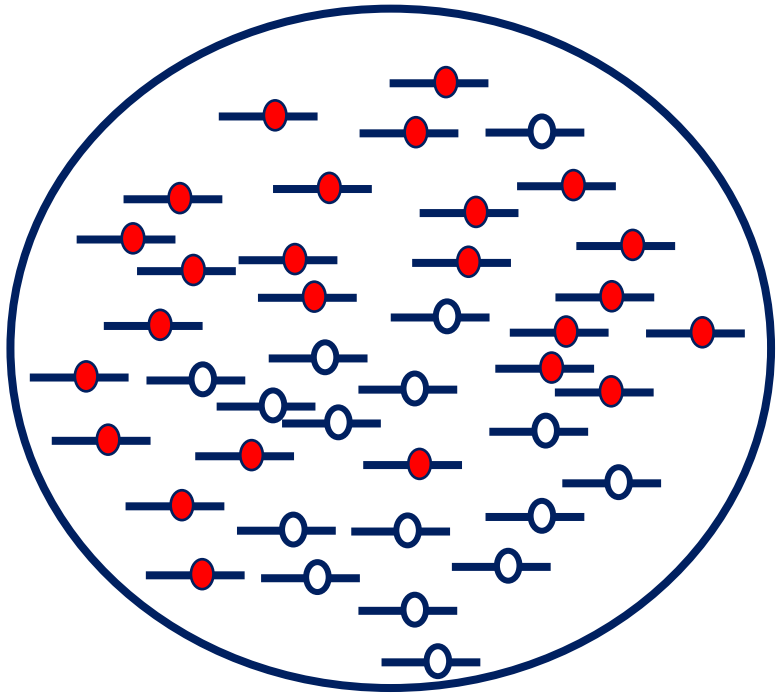
How can we describe populations of chromosomes?

- π – Nucleotide heterozygosity
  (allele frequencies within populations)

- θ – Nucleotide diversity
  (number of polymorphisms within populations)

- **Tajima's *D*** – Statistic to compare polymorphism counts to frequencies

- F_{ST} – Population divergence



Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \bullet \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration

Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration

```
          1 2 3
Seq1    – A T G
Seq2    – A T A
Seq3    – A C G
Seq4    – A T G
Seq5    – A T G
Seq6    – A T A
Seq7    – A C G
Seq8    – A C G
Seq9    – A T A
Seq10   – A C A
```

Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \bullet \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration

```
          1  2  3
Seq1   —  A  T  G
Seq2   —  A  T  A
Seq3   —  A  C  G
Seq4   —  A  T  G
Seq5   —  A  T  G
Seq6   —  A  T  A
Seq7   —  A  C  G
Seq8   —  A  C  G
Seq9   —  A  T  A
Seq10  —  A  C  A
```

i = 1

10 As

0 Ts

0 Gs

0 Cs

p = 1.0

q = 0.0

2pq = 2•1.0•0.0 = 0.0

Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \bullet \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

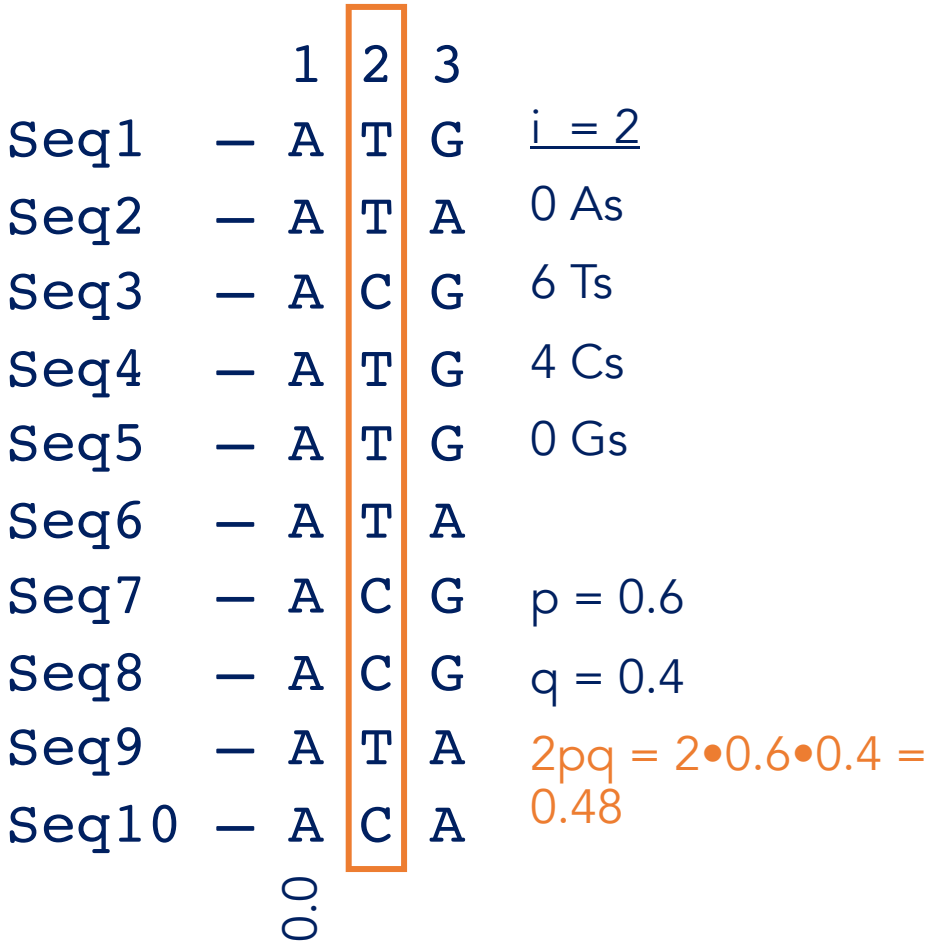
n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration



Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \bullet \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

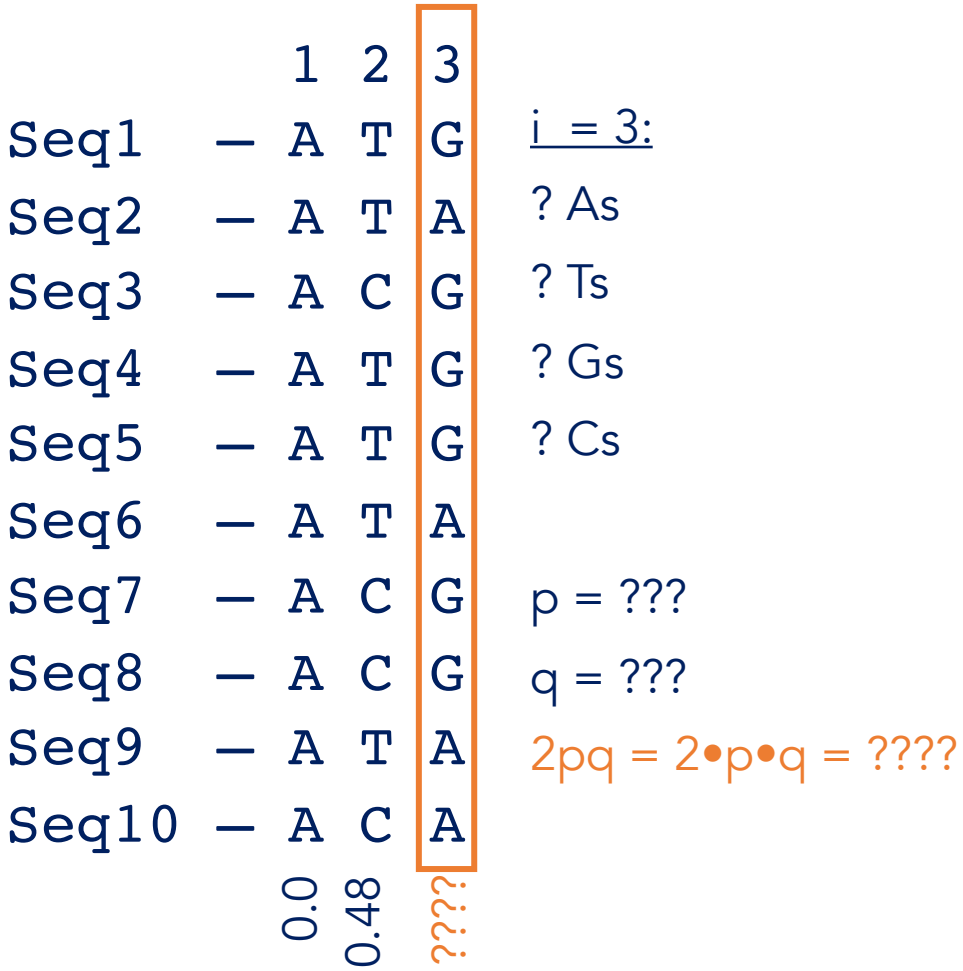
n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration



Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration

```
          1  2  3
Seq1   —  A  T  G
Seq2   —  A  T  A
Seq3   —  A  C  G
Seq4   —  A  T  G
Seq5   —  A  T  G
Seq6   —  A  T  A
Seq7   —  A  C  G
Seq8   —  A  C  G
Seq9   —  A  T  A
Seq10  —  A  C  A
         0.0 0.48 0.48
```

i = 3:

4 As

0 Ts

6 Gs

0 Cs

p = 0.6

q = 0.4

2pq = 2•0.6•0.4 = 0.48

Nucleotide heterozygosity (π)

$$\pi = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{L} 2p_i q_i}{L}$$

n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration

```
          1  2  3
Seq1      A  T  G
Seq2      A  T  A
Seq3      A  C  G
Seq4      A  T  G
Seq5      A  T  G
Seq6      A  T  A      n = 10
Seq7      A  C  G      L = 3
Seq8      A  C  G
Seq9      A  T  A
Seq10     A  C  A
          0.0 0.48 0.48   Sum = 0.96
```

Nucleotide heterozygosity (π)

$$\pi = \frac{10}{9} \bullet \frac{0.96}{3} = 0.356$$

n = # of sequences

L = Length of sequence

p = Frequency of allele A

q = Frequency of allele B

i = Site under consideration

```
          1 2 3
Seq1    – A T G
Seq2    – A T A
Seq3    – A C G
Seq4    – A T G
Seq5    – A T G
Seq6    – A T A     n = 10
Seq7    – A C G     L = 3
Seq8    – A C G
Seq9    – A T A
Seq10   – A C A
          0.0 0.48 0.48   Sum = 0.96
```

Nucleotide diversity (θ)

$$\theta = \frac{S}{L \bullet a_n}$$

a_n = Statistical correction for # of sequences

L = Length of sequence

S = Number of polymorphic (segregating) sites

n = # of sequences

Nucleotide diversity (θ)

$$\theta = \frac{S}{L \bullet a_n} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

a_n = Statistical correction for # of sequences

L = Length of sequence

S = Number of polymorphic (segregating) sites

n = # of sequences

Nucleotide diversity (θ)

$$\theta = \frac{S}{L \bullet a_n} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i} = ?$$

a_n = Statistical correction for # of sequences

L = Length of sequence

S = Number of polymorphic (segregating) sites

n = # of sequences

```
          1 2 3
Seq1    – A T G
Seq2    – A T A
Seq3    – A C G
Seq4    – A T G
Seq5    – A T G
Seq6    – A T A
Seq7    – A C G
Seq8    – A C G
Seq9    – A T A
Seq10   – A C A
```

Nucleotide diversity (θ)

$$a_n = \frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\frac{1}{9} = 2.829$$

$$\theta = \frac{S}{L \bullet a_n} \quad a_n = \sum_{i=1}^{n-1} \frac{1}{i} = 2.829$$

a_n = Statistical correction for # of sequences

L = Length of sequence

S = Number of polymorphic (segregating) sites

n = # of sequences

```
          1 2 3
Seq1    – A T G
Seq2    – A T A
Seq3    – A C G
Seq4    – A T G
Seq5    – A T G
Seq6    – A T A
Seq7    – A C G
Seq8    – A C G
Seq9    – A T A
Seq10   – A C A
```

Nucleotide diversity (θ)

$$a_n = 2.829$$

$$\theta = \frac{2}{3 \bullet 2.829} = 0.236$$

		1	2	3
Seq1	–	A	T	G
Seq2	–	A	T	A
Seq3	–	A	C	G
Seq4	–	A	T	G
Seq5	–	A	T	G
Seq6	–	A	T	A
Seq7	–	A	C	G
Seq8	–	A	C	G
Seq9	–	A	T	A
Seq10	–	A	C	A

a_n = Statistical correction for # of sequences

L = Length of sequence

S = Number of polymorphic (segregating) sites

n = # of sequences

Tajima's D

$$D = \pi - \theta \text{ (approximately)}$$

Compares average allele frequencies of polymorphisms
to the number of polymorphisms

D > 0.01 – allele frequencies are higher than expected

–0.01 < D < 0.01 – allele frequencies are consistent with NNMOME

D < –0.01 – allele frequencies are lower than expected

Tajima's D

$$D = \pi - \theta \text{ (approximately)}$$

Compares average allele frequencies of polymorphisms to the number of polymorphisms

D > 0.01 – allele frequencies are higher than expected

- Population structure (genome-wide)
- Population bottleneck (genome-wide)
- Ongoing/soft selective sweep (gene-specific)
- Balancing selection (gene-specific, unlikely)

Tajima's D

$$D = \pi - \theta \text{ (approximately)}$$

Compares average allele frequencies of polymorphisms
to the number of polymorphisms

D > 0.01 – allele frequencies are higher than expected

–0.01 < D < 0.01 – allele frequencies are consistent with NNMOME

D < – 0.01 – allele frequencies are lower than expected

Tajima's D

$$D = \pi - \theta \text{ (approximately)}$$

Compares average allele frequencies of polymorphisms
to the number of polymorphisms

- Population expansion (genome-wide)
- Exceptionally strong purifying selection (gene-specific)
- Historical selective sweep (gene-specific)

D < –0.01 – allele frequencies are lower than expected

Tajima's D

$$D = \pi - \theta$$

$$D = 0.356 - 0.236 = 0.120$$

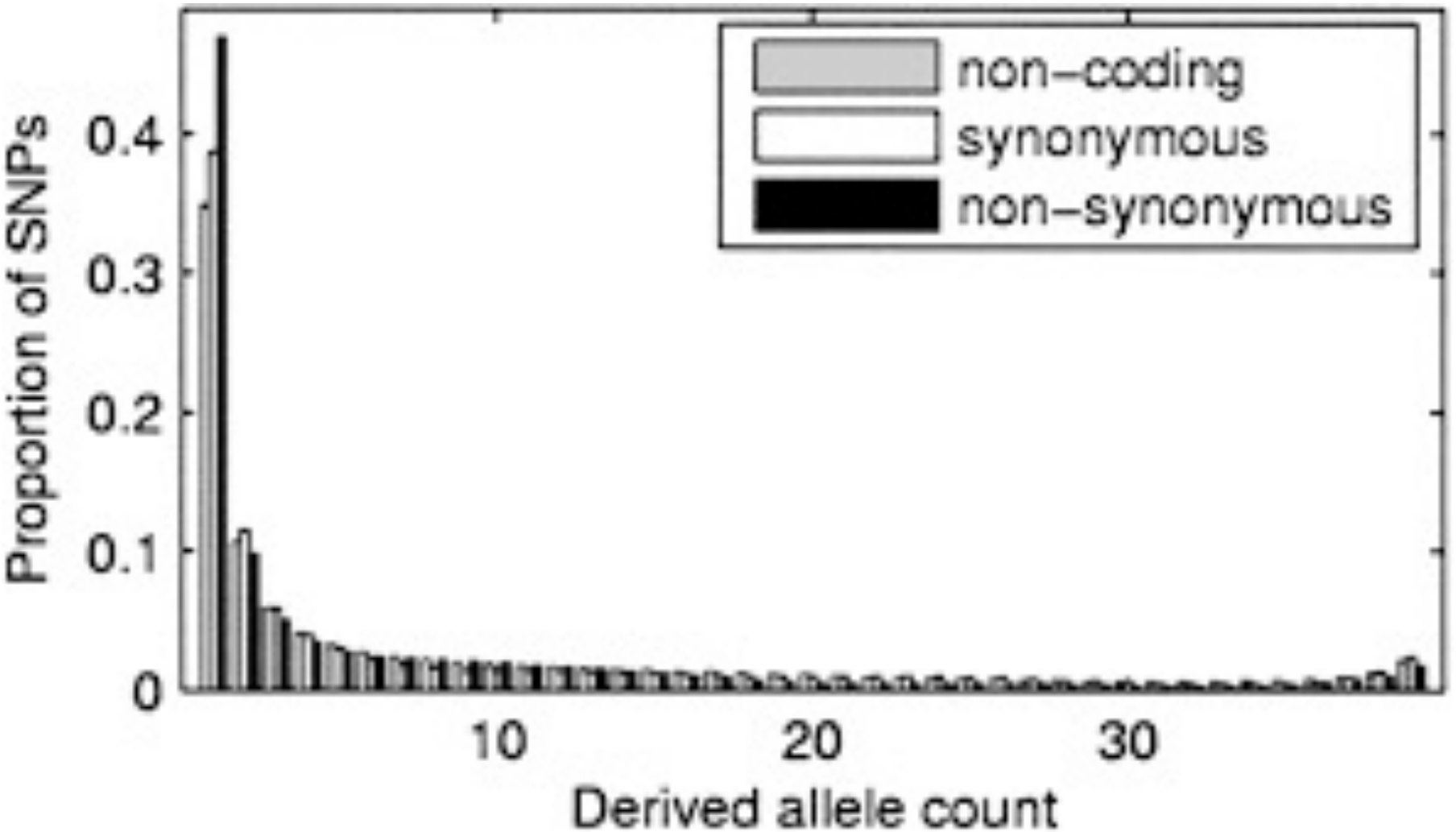
Compares average allele frequencies of polymorphisms to the number of polymorphisms

D > 0.01 – allele frequencies are higher than expected

–0.01 < D < 0.01 – allele frequencies are consistent with NNMOME

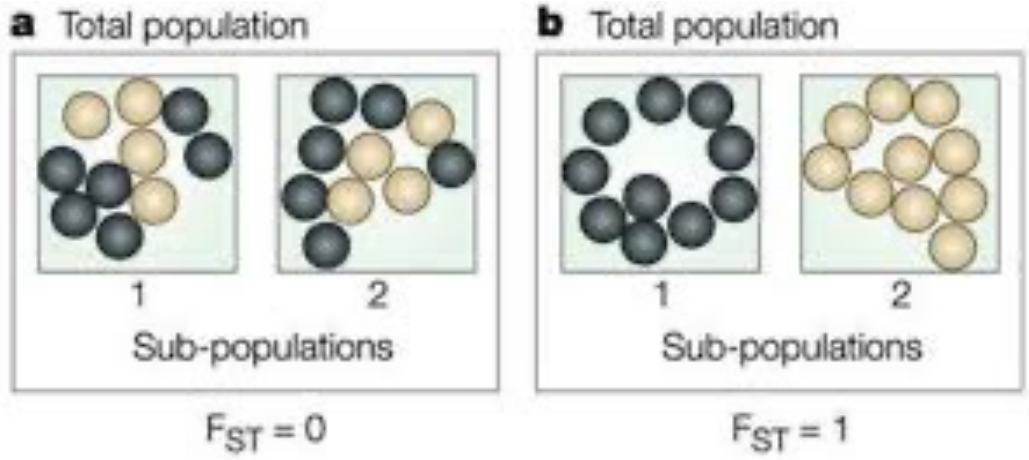
D < –0.01 – allele frequencies are lower than expected

Tajima's D – A measure of the Site Frequency Spectrum



F_{ST} – a measure of population divergence

$$F_{ST} = \frac{\pi_W - \pi_B}{\pi_B}$$



π_W = Nucleotide heterozygosity within a population

π_B = Nucleotide heterozygosity in combined populations

Compare across the genome, regions of exceptionally high divergence candidates for loci under selection

Next up: Genome-wide Association Studies

Please Watch: https://youtu.be/J97rj__zCxE

Please Read: Coop and Przeworski 2022