# Genomics

***Allele*** - each of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome. Variants of the same gene. Distinguishable and heritable (SNP, indel, microsat)

***Locus*** — any position (or unit) in the genome with one or more alleles

***Genotype***: combination of alleles carried by an individual in a particular locus *e.g. an individual has A and G alleles, and therefore has AG genotype, at locus in position 8,789,654 of chromosome 1.*

***Haplotype(单体型)***: Contains information about *where* the alleles sit on which particular copy of the chromosome. E.g. the AG genotype will have two hapolotypes

***Single Nucleotide Polymorphism (SNPs)*** — Variation at a single base *e.g. The C/T variation at position 478 in MC1R; individuals with two T’s tend to be gingers!*

***Indel*** — The insertion/deletion of a few nucleotides

***Microsatillites*** — DNA replication machinery tends to miscopy repeated sequences in the genome. Microsatellites are variants on the number of repeats transmitted during meiosis(减数分裂) (a type of cell division), with a small possibility of error. *e.g. sequence AGCTGCACACACACACACATGCTG has CA motif repeated seven times, while other individuals may have a different number of copies, thus (CA)n.*

***diploid*** species have two copies of its chromosomes, so for a collection of N diploid individuals, there are 2N gene copies at each locus, with one or more alleles.

* As mutations are rare in most organisms, ***di-allelic*** models are often used, with at most two alleles at each locus

***Diallelic*** - a gene, genetic locus, or genetic marker having two alleles.

# Alleles and Genotypes

## Allele frequencies:

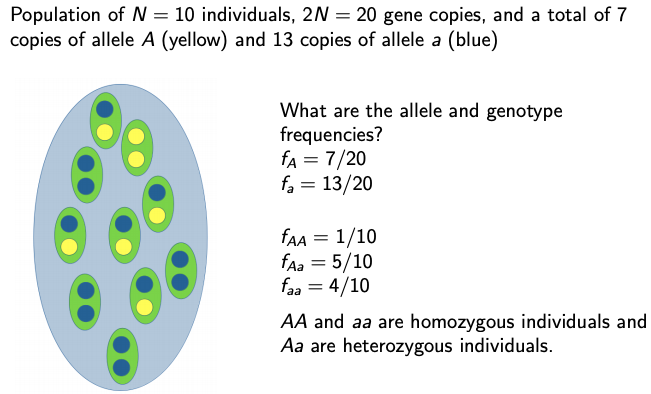
Where is the number of X alleles in the population and is the total number of alleles in the population

Hence with N diploid individuals with X and x alleles:

with:

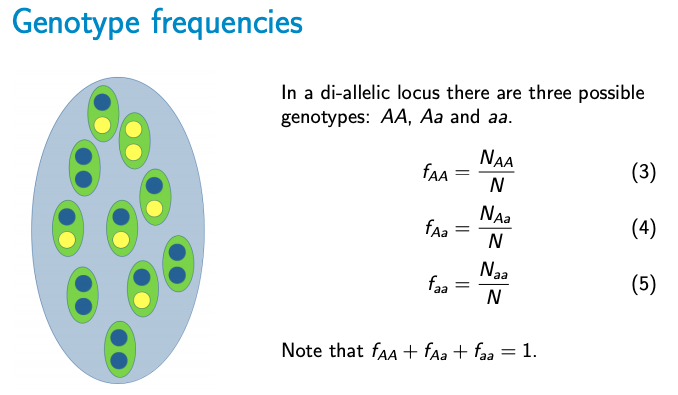
(Same logic for genotype frequencies)

**E.g.**



* AA and aa are ***homozygous*** individuals and Aa are ***heterozygous*** individuals.
* Much population genetics focuses on describing the changes of fA and fa with time.

## Genotype frequencies:



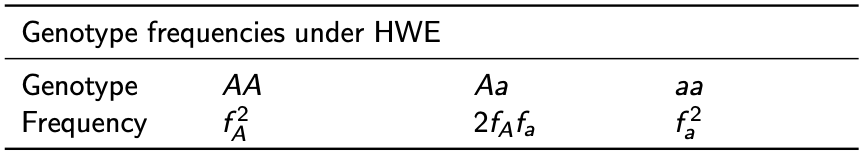
## Genotype frequencies → Allele frequencies:

* The proportion of heterozygous individuals in the population (fAa) is called the ***heterozygosity***.
* The proportion of homozygotes (1 − fAa = fAA + faa) is the ***homozygosity*** of the population.

## Allele frequencies → Genotype frequencies (predictions):

* Some assumptions must be made here:
  + Random mating: individuals mate with each other without regard to genotype
  + Males and females have equal allele frequency
  + Di-allelic locus

### We use the ***Hardy-Weinberg Equilibrium (HWE)*** to predict genotype frequency from allele frequency:



* Expected homozygosity:
* Expected heterozygosity:
* Homozygosity + heterozygosity = 1 =

**What about when we are not in Hardy-Weinberg Equilibrium?**

I.e. there is

* ***Assortative mating***: non-random with respect to genotype
* ***Inbreeding***: mating of related individuals
* Population structure: sample of individuals from two or more subpopulations
  + I.e. separate gene pools with diff alleles; no heterozygous individuals in the whole population
* Deleterious natural selection
  + E.g. Insertion in HEXA gene associated with Tay-Sachs disease

### Testing deviation from Hardy-Weinberg Equilibrium

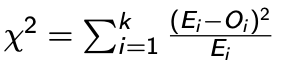
#### ***Inbreeding coefficient***

* Measures the degree to which heterozygosity is reduced both within individuals and populations
* If F = 0 the population is in HWE, if F = 1 there are no heterozygotes (completely inbred).

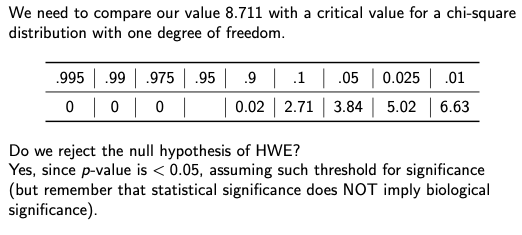
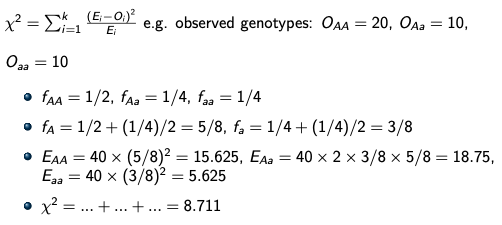
This can be rearranged to find the heterozygosity:

* The proportion of heterozygotes in the population is reduced by a factor of F from that expected under HWE.
* **If we know F and the allele frequencies, we can predict genotype frequencies without assuming HWE.**
* Species likely to deviate from HWE: self-breeding (asexual) species
* A random sample for a population in HWE may deviate from HWE.
* We need a formal statistical test:

#### ***Chi-sqd Test***

* + - 
    - observed values (Oi , genotype counts)
    - expected values (Ei , expected genotype counts under HWE)
  + Null hypothesis: genotype frequencies follow those predicted by HWE
  + Alternative hypothesis: genotype frequencies do not follow those predicted by HWE
  + If chi-sqd is large enough (i.e. compare to critical values to find p value) , we reject the null hypothesis

E.g.



# Genetic Drift

Population genetics often focuses on describing the changes of allele frequencies through time.

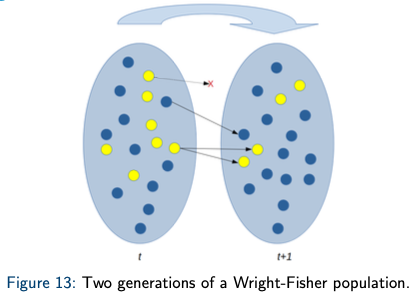
The three most important factors that cause allele frequencies to change are:

* natural selection
* mutations
* ***Genetic drift*** — the random change of allele frequencies in populations of finite size
  + some individuals leave many offspring, others fewer, other none
  + heterozygous individuals will randomly transmit allele A or a
  + Due to these, it is likely that allele frequencies will change from one generation to another, and over many generations, this process can produce large changes in allele frequencies

## Wright-Fisher Model

Assumptions:

* haploid population
* asexual (no mating)
* discrete generations



* The distribution of offspring in generation t + 1 is given by a binomial distribution
* Under the Wright-Fisher model, we can easily characterise the change in allele frequency mathematically.

## Calculating expected allele freqs

What is the probability that any gene copy in generation t + 1 is A?

The expected allele frequency in generation t + 1 is equal to the allele frequency in generation t (the probability that any gene copy in generation t+1 is A):

### Drift

* At each generation, allele frequency might change a bit
* Small changes add up and, after many generations, allele frequency may have changed significantly
* Many small changes may result in large evolutionary changes over sufficiently long periods of time.
* Allele frequency may increase or decrease with equal probabilities
* In some cases, allele has become **fixed** or **lost**
  + When an allele first has become fixed or lost, its frequency cannot change anymore (e.g. if fA(t) = 0 then fA(t + 1) = 0)
  + Is it always true? Yes if we assume no recurrent mutation: in the absence of recurrent mutation, it can be shown mathematically that an allele must eventually become fixed or lost.

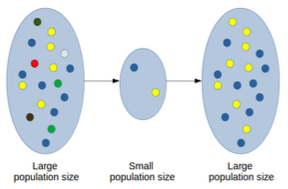
## Effect of population size on drift

How fast can genetic drift change allele frequencies? **It depends on the population size, N.**

* Large changes in allele frequency are unlikely in large populations, but happen more easily by chance in small populations.
* Genetic drift works much faster in small populations than in large populations.
* This has impotant implications for our understanding of natural pops

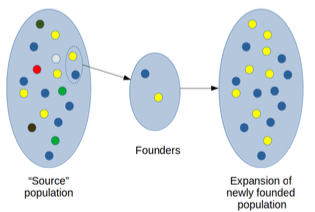
### Bottlenecks:

* You can lose genetic variation through population size bottlenecks, as in the short period of time when the population size is very small many alleles become either fixed or lost in the population.
  + Sea elephants were hunted nearly to extintion to a populazione of just 2-20 individuals. Today the population rebounded to 175,000 individuals. From historical (before hunting) and modern samples, their genetic diversity\* was reduced from 0.90 to 0.41.



### Founder Effect:

* Reduction in variability caused by a bottleneck in the population size during the founding of a new population.
* Genetic divergence after speciation may be helped along by the strong effects of genetic drift in the founders of a population.



# Mutation

New mutations arise to produce new genetic variation that genetic drift can act on:

* deletions
* insertions
* translocations
* point mutations

## Effect on allele frequency

If the *a* allele in each individual randomly mutates to *A* with probability *μ* (mutation rate) in each generation, then the expected frequency of the A allele in the next generation is:

If mutations occur in both directions, e.g. mutations occur at rate *μa→A*  from *a* to *A* and *μA→a* from *A* to *a*, then:

In the absence of other forces (e.g. genetic drift and selection), an equilibrium will eventually be established:

## Mutation rate

* Mutation is a weak force in higher organisms
* With no drift, it takes a long time for the elle frequency to equilibrate
* We can often ignore recurrent mutations

## Probability of fixation

* *In the absence of selection and mutation, the probability of fixation of an allele is simply its allele frequency.*

## Rate of substitution

* Rate at which mutations accumulate **between species**. Substitution refers to mutations that have gone to fixation.
* Assume:
  + mutation rate *μ*: in each generation μ new mutations occur in each gene copy (e.g. per site, per gene, ...)
  + 2N gene copies
* Then the expected no of mutations each generation that will eventually lead to fixation is simply the mutation rate:

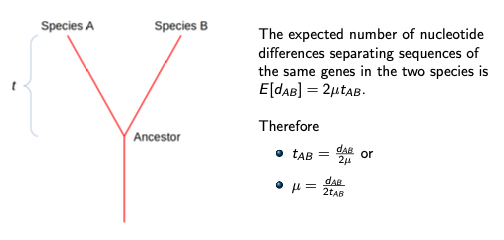
## Molecular Clock

If:

* no selection
* ”low” mutation rate (not affecting allele frequencies much)
* constant mutation rate

then the rate of substitution should be constant in time.

Mutations can be used to date divergence between species. How?



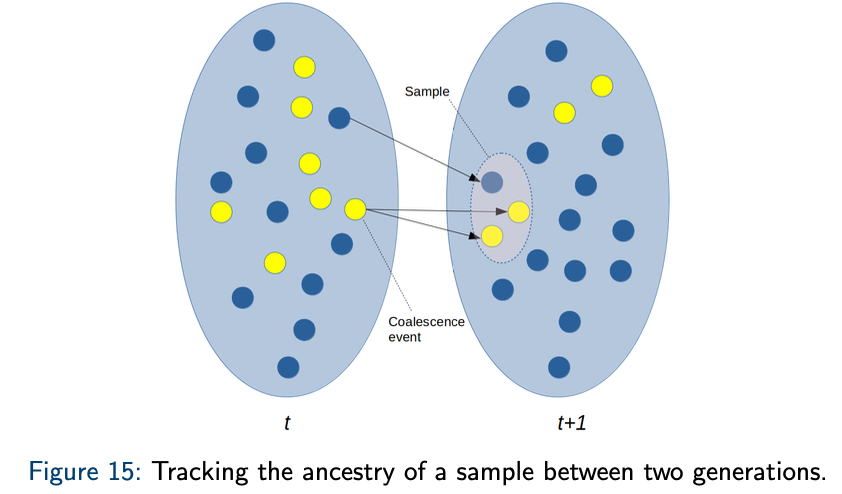
Caveats:

* it depends on an estimate of μ
* it assumes no natural selection acting upon
* it assumes that mutation rate is constant and equal among different species

Not very realistic but a good approximation for closely related species.

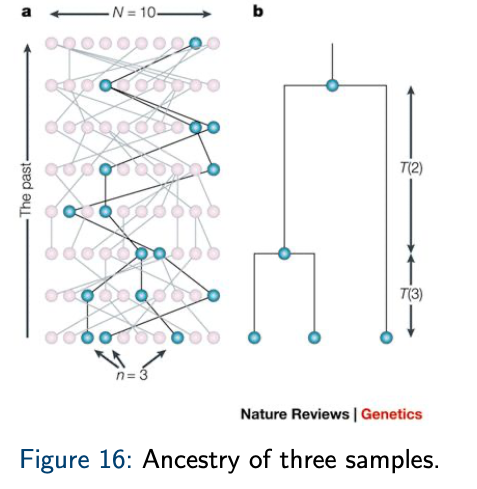
# Coalescence Theory

We use **coalescence theory,** which is based on Wright-Fisher model, to consider the genealogy history of a sample and make inferences about populations instead of modelling changes of allele frequencies forward in time.



* If two individual gene copies have the same parent in the previous generation, we say that the **ancestral lineage** representing these two individuals have **coalesced**.
* They have a **common ancestor** and a **coalescent event** has occurred.

## Coalescence Tree



* The ancestry of an individual gene copy is represented by a line (or **edge**).
* The time until two lineages find a **most recent common ancestor (MRCA)** is called **coalescence time**.

How can we find the coalescence time?

## Coalescence in sample n=2

* The probability of two individuals having the same parent in the previous generation is (as there are 2N potential parents ”chosen” with equal probability)
* Hence, the probability that they did NOT share the same parent is
* The probability that two gene copies did not share the same parent in the past r generations is
* The probability of not finding any common ancestor in generation **r − 1** but then finding the first common ancestor in generation **r** is
  + This gives us the probability distribution of the coalescence time in a sample of size **n=2**
  + This is a geometric random variable: the probability distribution of the number of Bernoulli trials needed to get one success.

## Coalescence in large populations

* If we consider the limit of an infinitely large population, calculations simplify but we can still consider the effect of genetic drift.
* It is convenient to measure time in 2N generations, by setting r = 2Nt with t measuring time in 2N generations.
* The probability that two gene copies do not find a common ancestor in 2Nt generations becomes

as

* As N becomes large, the distribution of the coalescence times follows an exponential distribution with mean 1.
* As time is measured in 2N generations, the mean (expected) time to coalescence is actually 2N generations. In other words, there is a constant rate of coalescence of 1 per 2N generations.
* The random process of following the lineages backward in time until a most recent common ancestor has been found is called a **coalescence process**.
* The coalescence process in a large randomly mating diploid population with two sexes is the same as that in the simple haploid model.

## Genetic Variability and Population Size

* A simple relationship between the amount of genetic variability and population sizes:

Where θ = the expected number of mutations separating two gene copies in a neutral population where only drift and mutation is happening.

### ?

How is this derived?

* With a mutation rate of μ we expect μr mutations in r generations. If we measure time t in 2N generations, then this expectation becomes 2Nμt. We also know that the coalescence rate is 1 per 2N generations and there are two lineages separating the two gene copies. Therefore E[t] = 1 and 2Nμ×2 leading to θ = 4Nμ.

* The expected number of mutations occuring in a lineage during any time interval of length τ is 2Nμτ = τθ/2.
* As such, we can think of the data generated by a coalescence process producing a coalescence tree and a subsequent process in which mutations are distributed across the lineage of the tree at rate θ/2.

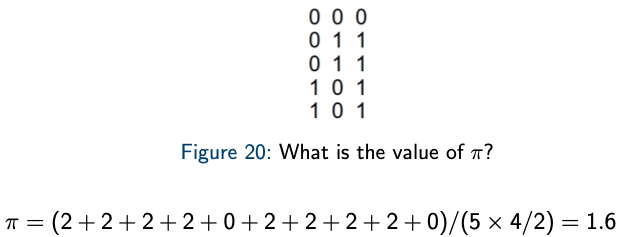
## Infinite Sites Model

* Each new mutation creates a new variable site, i.e. that each new mutation hits a new site in the sequence, such that no site experiences more than one mutation.
* The sites in which some of the individuals differ are called **segregating sites** or **single nucleotide polymorphisms (SNPs)**.
* Under the infinite sites model, we can deduce which mutations occurred in the ancestry of a sample of sequences.
* The model does not distinguish between different nucleotides and does not care about invariable sites.
* Infinite sites model
  + Labelling with zeros and ones is arbitrary.
  + Good approximation if the rate of mutation is low.
  + DNA sequences with different mutations are different haplotypes.
  + **MORE HERE - WATCH VIDEO**

## Tajima’s estimator

* We want an estimate of θ = 4Nμ (population genetic parameter) under the infinite sites model from the expected number of mutations separating two individuals based on the DNA sequences obtained from data.
* Data can be summarised as the **average number of pairwise differences**, or π.

with n sequences, di,j number of differences between sequence i and j.



* The **expected number of nucleotide differences between two sequences** is the expected number of mutations, **θ = 4Nμ**.
  + E[di,j] = θ
  + E[π] = θ (17) θT = π is called Tajima’s estimator of θ.

## Effective Population Size

The number of individuals in a Wright-Fisher model that would produce the same amount of genetic drift as in the real population.

The amount of genetic drift can be measured as

* the expected heterozygosity
* expected number of pairwise differences [π]
* rate of coalescence
* …
* If a population fluctuates between sizes N1, N2, ..., Nk at a proportion p1, p2, ..., pk of the time, the coalescent effective population size is the harmonic mean

which is smaller than the arithmetic mean and gives more weight to smaller sizes.

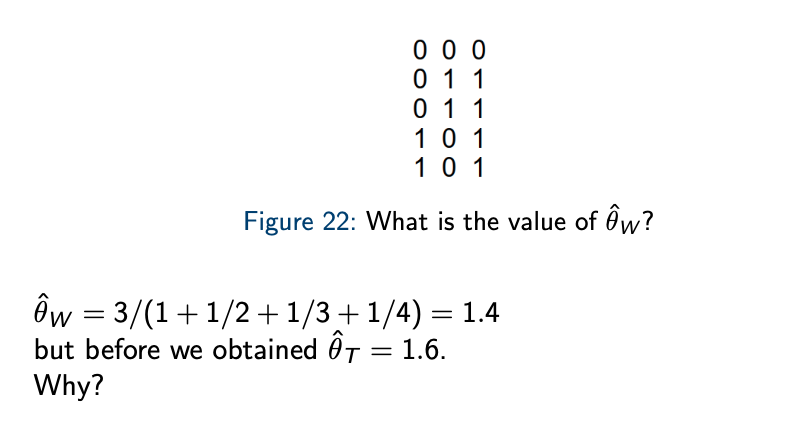
* The effective population size with unequal sex ratio is:

which is smaller than Nm + Nf .

## Watterson’s estimator

Another estimator for θ = 4Nμ

Where S is no. of segregating sites (SNPs) and n is no. of samples



## Summary Stats

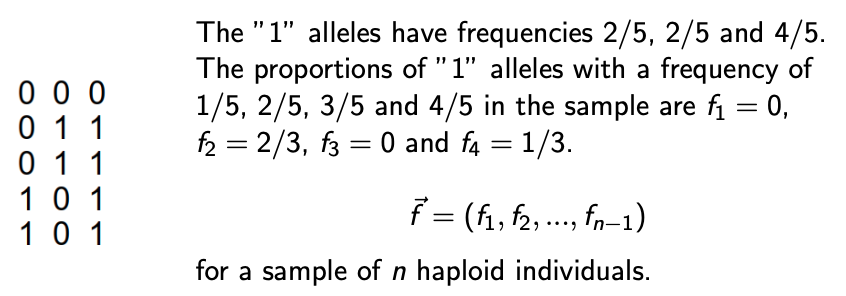
Possible summaries of DNA sequence data are:

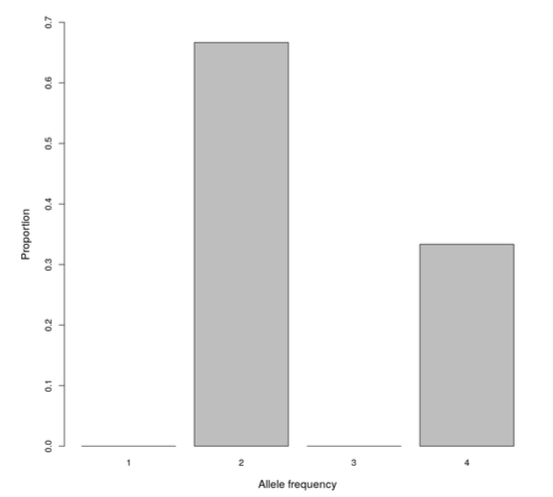
* the number of segregating sites (S)
* the average number of pairwise differences (π)

but they don’t provide much information regarding allele frequencies.

### Site Frequency Spectrum (SFS)

* + Interpret as the distribution of allele frequencies in a given population
  + obtained by tabulating the sample allele frequencies of all mutations.

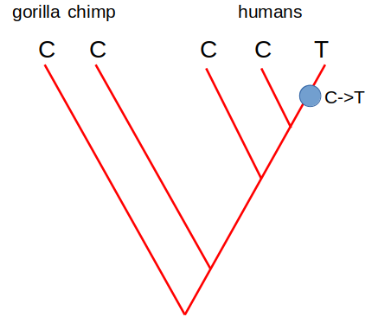




**Ancestral allele** — the allele found in the MRCA of the sample

**Derived allele** — the allele that’s not ancestral

The ancestral allele is often inferred using outgroups. e.g. if C/T polymorphism in humans and primate have C, then C is likely to be the ancestral allele.



* We infer tell this if we have knowledge of the outgroup and no ambiguity in allelic configuration of outgroup

### Folded Site Frequency Spectrum (SFS)

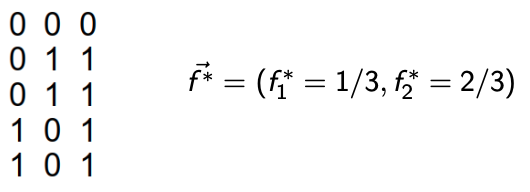
* If no information on the ancestral allele is available, we can *fold* the frequency spectrum.
* The folded frequency spectrum f ∗ is obtained by adding together the frequencies of the derived and ancestral alleles.

for

for

only defined for values of f\* ≤ n/2.

**E.g.**



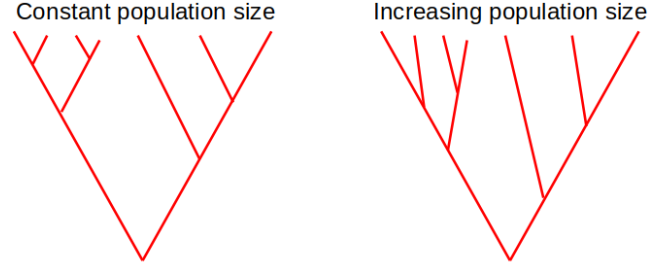
* S and π can be calculated directly from f but the opposite is not true.
* Alleles segregating at frequency of 1/n are called **singletons**.
* The expected SFS under the standard coalescence model with infinite sites mutations is

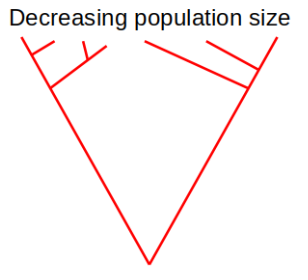
with j = 1, 2, ..., n − 1

## Tree shape and population size

Measured in number of generations, the expected coalescence time for k lineages is

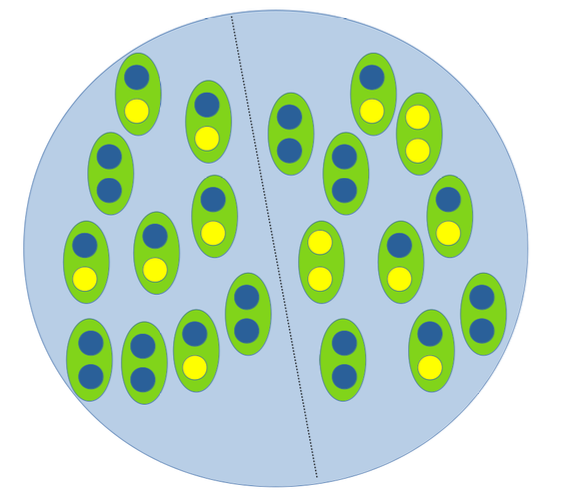
**2N/[k(k − 1)]**.





# Population subdivision

* There is population subdivision, or structure, when the population is not randomly mating because of geographic or social structure.
* Population subdivision is important to:
  + understand the effects of drift and natural selection
  + plan conservation strategies for rare or endangered species
* Hypothetical scenario: barrier between populations



## Allele Frequencies in Subdivided Population

Assume two subpopulations, each one in HW equilibrium with N1 and N2 individuals, respectively.

The average frequency of allele A when pooling the two subpopulations is

Hence if N1=N2:

## Heterozygosity in a Subdivided Population

* The **proportion of heterozygous individuals** is

which is the **expected heterozygosity when both populations are sampled**.

* (S in HS stands for ”in the subdivided population”)
* Rearranged:
* However, the **expected proportion of heterozygous individuals in a population with frequency fA** is

T in HT stands for ”in the total (pooled) population”

* Rearranged:

Where

* Hence, HT =HS +δ2/2
* **If δ = 0** then
  + HT = HS and the total (pooled) population is also in HWE.
* **If δ >> 0**
  + then HT > HS and the total (pooled) population contains fewer heterozygous individuals than expected given the pooled allele frequency.

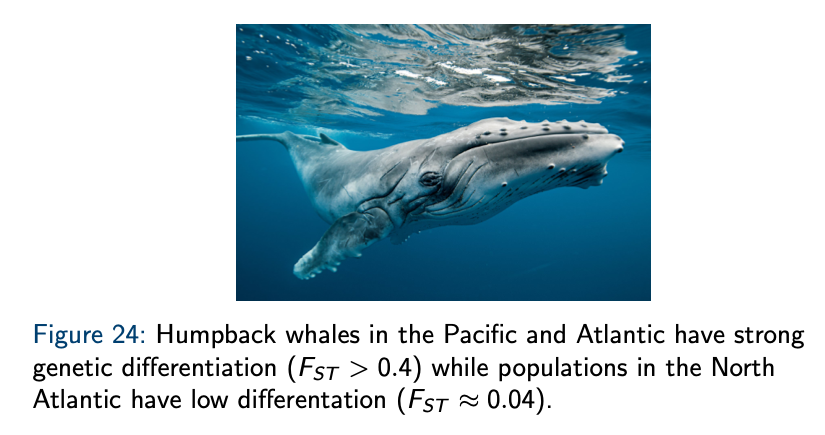
### Wahlund effect

* The decrease of heterozygosity in a subdivided population compared to a randomly mating one with the same (total) allele frequency.

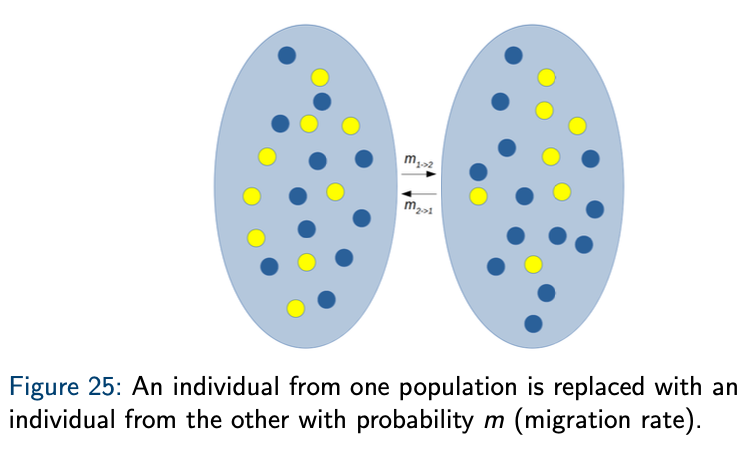
## Quantifying Population Subdivision

**Population Genetic differentiation:**

* FST has a range defined as
  + If δ=0 then HT = HS and FST =0
  + If δ>>0 then FST ≈1
* FST can be calculated for more than two subpopulations.



## Wright-Fisher Model with Migration



### 

### 

### 

### 

# Next Generation Sequences

## DNA

* Four bases/nucleotides: Adenine, Thiamine, Guanine, Cytosine.
* Grouped by complementary base: AT, CG.
* DNA replication explanation with enzymes in bold:
  + **Helicase** separates DNA into two strands and the **DNA polymerase** adds complementary bases to them both.
* DNA replication is exploited in PCR (polymerase chain reaction), amplifying the amount of DNA in your sample.

## Sanger sequencing

* Dd nucleotides (ddATP) are attached to the portion of the sequence you're interested in and you can infer the nucleotides that form it by a classic electrophoresis analysis.
* Today the technology still uses the dd nucleotides but instead we obtain a chromatography from capillary gel electrophoresis.

## NGS

* Is highly parallel, sequencing millions of fragments simultaneously per run. It is faster, cheaper, requires less DNA and is higher throughput.
* Types of NGS include: (not very important to know exactly how they work)
  + Short reads: second generation sequencing. Eg:
    - Illumina - bridge amplification of small sequencing, sequencing by synthesis and imaging
    - IonTorrent - emulsion PCR, sequencing by synthesis and analysis of changes on pH
  + Long reads are third generation sequencing. Eg:
    - PacBio - single molecule real time sequencing. Sequencing by synthesis and imaging
    - Oxford Nanopore: tracks changes in current as DNA passes through a pore. There is a portable version that can be taken to space.

## Workflow

### Experimental design

* Careful design saves time and money.
* Think of these steps: question, sample and sequencing experiment

### Library preparation for illumina sequencing

Libraries are short fragments of DNA with attached:

* Indexes: differentiate one fragment from another (run at the same time)
* Adapters: 30-50 base pairs long and link the fragment to the slide

These fragments need preprocessing such as:

* Genome sequencing: DNA extraction, fragmentation through sonication or enzymatic reaction, selection of fragments through size.
* Amplicon sequencing: DNA extraction
* RNA sequencing: RNA extraction, fragmentation, reverse transcription (to DNA)

### Data analysis

(file types underlined)

* Go from raw to clean reads
  + Raw data are FASTQ files that start with @ and a sequence name, have the sequence itself, the character + and the phred quality scores represented as ASCII characters.
  + Phred quality scores (Q) represent the certainty of the assignment of bases. Q=-10log10(P) (Q=10 is 90% accuracy because there's a 1/10 chance of an incorrect base, Q=20 is 99% accuracy because there's a 1/100 of an incorrect base, etc.).
* Do a quality check
  + Remove the adapters, check for short or low quality reads, quality check with FastQC (graphs the distribution of quality scores per base) and finally trim the reads.
* Figure out alignment
  + Mapping to a reference with BWA, Bowtie2, SOAP
  + *De novo* assembly with Velvet, Trinity, SPAdes, ABySS
* Extract information
  + Important concepts:
    - Coverage: average number of reads that include the given nucleotide in the reconstructed sequence
    - N50: measure of the contiguits of assembly. Its the sequence length of the shortest contig at 50% of the total genome length
  + Output files
    - From mapping to a reference: SAM/BAM files that are tab delimited. They contain a header with sample information and a section on alignment with the location and quality of every read.
    - From de novo assembly: Fasta files are tab delimited and are composed of an identifier and the sequence itself.
  + Variant calling compares sequences with a reference and finds the differences . Before this we need to clean the data a bit more (realignment, base quality recalibration, removal of PCR duplicates). To identify differences we use:
    - GATK and SAMtools for SNPs and indels
    - Pindel and GRIDSS for structural variants

These tools output a VCF file (variant call format) that is tab delimited and has metadata, a header and data lines

* + Annotation is the extraction of biological meaning and outputs GFF (general feature format) files
    - Whole genome sequencing (WGS) to look at structural elements of the genome (coding vs non-coding regions, location of specific genes, etc)
    - Genome re-sequencing looks at differences between variants: are they associated with disease? Are some more frequent in given populations?
    - RNA sequencing describes what the genes do

Examples of online gene banks at the end of the lecture.