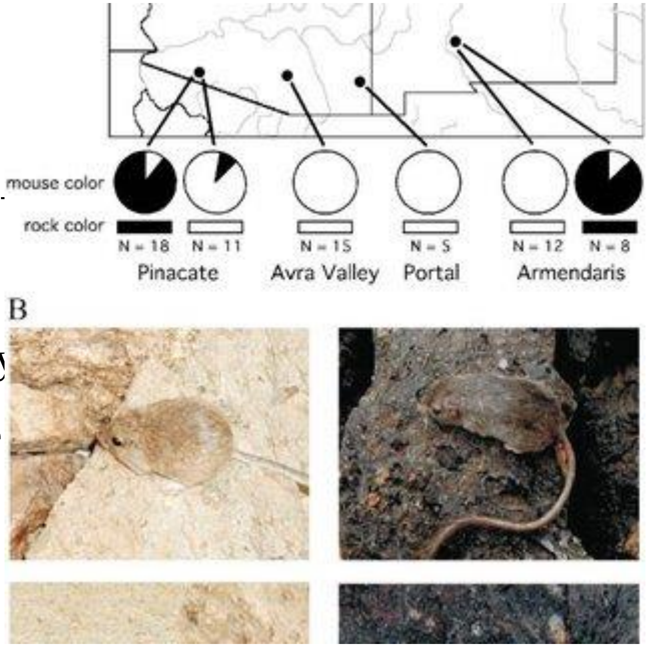
MOLECULAR EVOLUTION (22ZOONME33)

POPULATION STUDIES AND GENE FLOW

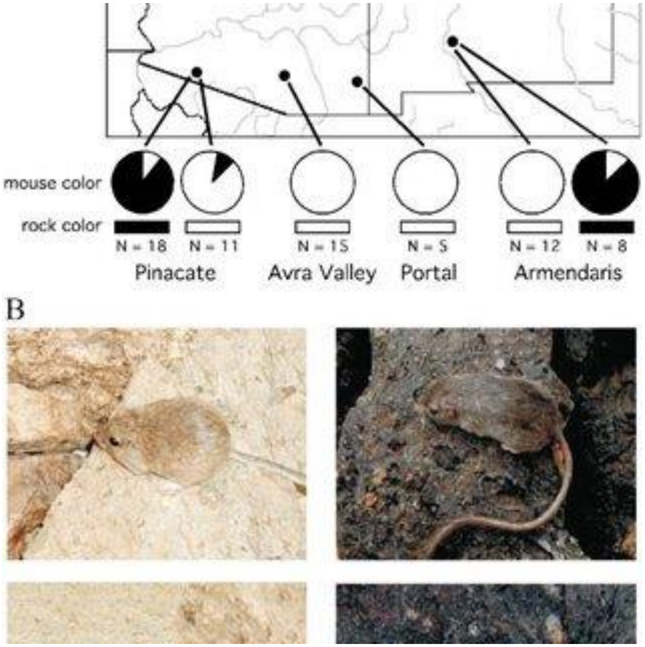
Mutation

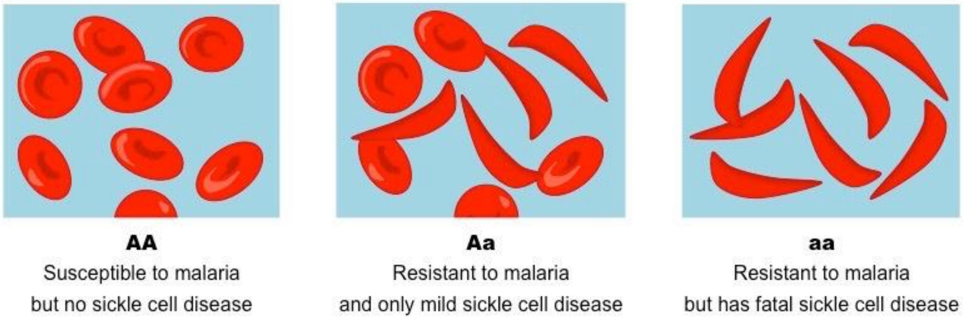
- Supply an important source of variation in population
- A population – established – particular environment – many genes – adopt for prevailing conditions
- Multicellular organisms – constrainted by their evolutionary history advantageous mutation generally confined to few of many intricate development and processed
- Plants and animals are seperated – more than a billion years ago
- Differences in 100 aminoacid of histone 4 protein binds and folds the DNA
- Alleles – low frequency – high relative fitness
  - Rapid genetic change in many insects – exposed to pesticides DDT – Resistance alleles appear in all chromosomes
  - Increased in the frequency of black alleles in populations of the peppered moth
    - Biston Betularia in industrial region
    - Receptor involved in pigmnent cells in light and dark strains of the rock pocket mouse
    - Chaetodipus intermedius in Southern west USA
  - Increased frequencies of resistant genes in plant population – exposed to herbicide and metallic toxins
  - Genes – modify RBCs physiology offer – protection against malaria – Sickle cell anemica



Mutation

- Alleles – low frequency – high relative fitness
  - Rapid genetic change in many insects – exposed to pesticides DDT – Resistance alleles appear in all chromosomes
  - Increased in the frequency of black alleles in populations of the peppered moth
    - Biston Betularia in industrial region
    - Receptor involved in pigmnent cells in light and dark strains of the rock pocket mouse Chaetodipus intermedius in Southern west USA
  - Increased frequencies of resistant genes in plant population – exposed to herbicide and metallic toxins
  - Genes – modify RBCs physiology offer – protection against malaria – Sickle cell anemica

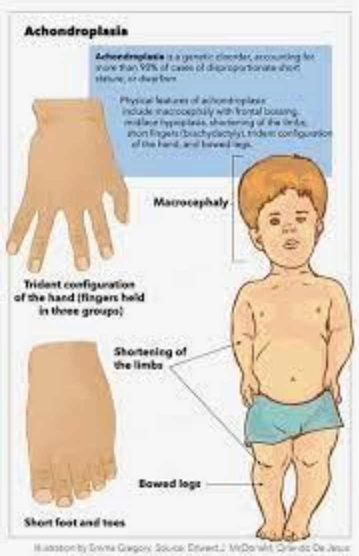



AA
Susceptible to malaria
but no sickle cell disease
Aa
Resistant to malaria
and only mild sickle cell disease
aa
Resistant to malaria
but has fatal sickle cell disease

Founder of the next generation

- Many individuals die
- Leaving these individuals with alleles
- Resistance
- Melanism
- Protection to pass their alleles to next generation
- Mutation rate and mutational equilibrium calculated

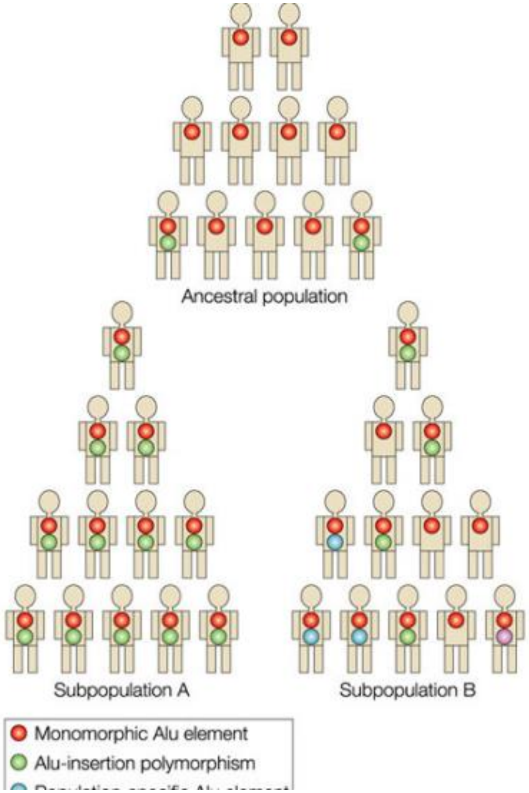
Mutation rate


Achondroplasia
Achondroplasia is a genetic disorder, accounting for more than 90% of cases of disproportionate short stature, or dwarfism
Physical features of achondroplasia include macrocephaly with frontal bossing, midface hypoplasia, shortening of the limbs, short fingers (brachydactyly), trident configuration of the hand, and bowed legs.
Macrocephaly
Trident configuration of the hand (fingers held in three groups)
Shortening of the limbs
Bowed legs
Short foot and toes
Illustration by Emma Gregory. Source: Edward J. McDonald, Orlando De Jesus

- Low – 1/100,000 copies of a gametes much variation
- In modern humans
- Achrondroplasia – form – dwarfism
- 0.6-1.3 mutation / 100,000 gamets
- Neurofibromatosis - 5-10 mutations/100,000 gametes
- Carrying an estimated 25,000 genes per haploid genome
- Ovum and sperm may carry less than one mutation average of <0.4 new mutation in a diploid fertilised zygote
- Mutation rate – not only low – not constant

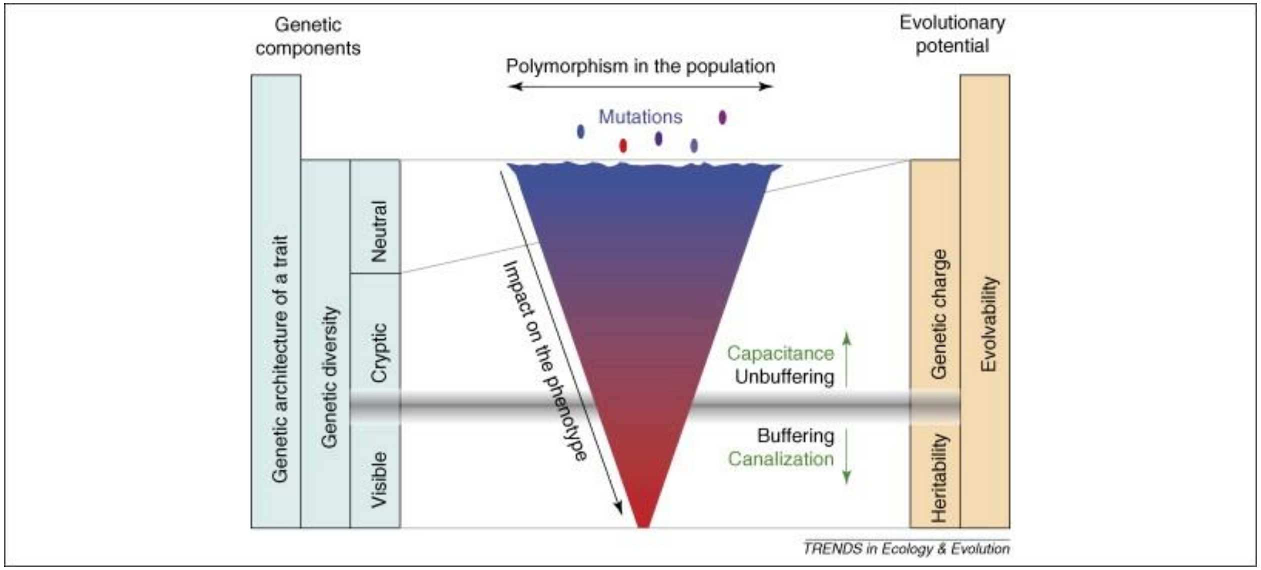
Hot spot mutation

- 100 times the “normal’ rate
- Specific nucleotide sequences may be the site of higher than average mutation rates
- Increased mutation rates among adjacent nucleotide
- Site for less ready for repair/compensated for them sequence change at another site
- DNA coiling specific pattern – results in DNA polymerase producing replication errors
- Alu insertion

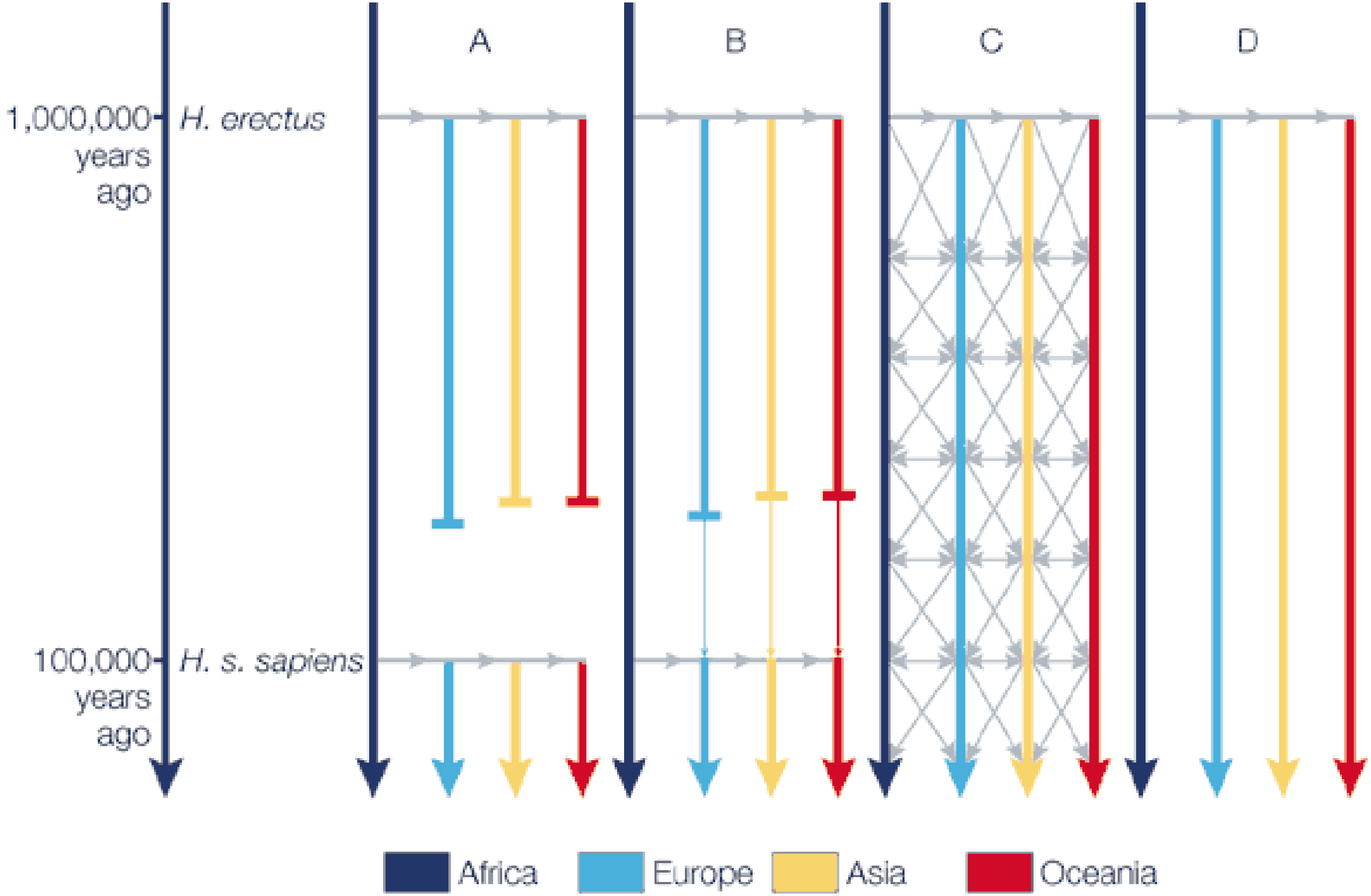


Genetic polymorphisms

- Coin
- New mutation – immediate beneficial effect – rare
  - Evolutionary potential
  - Genetic variation



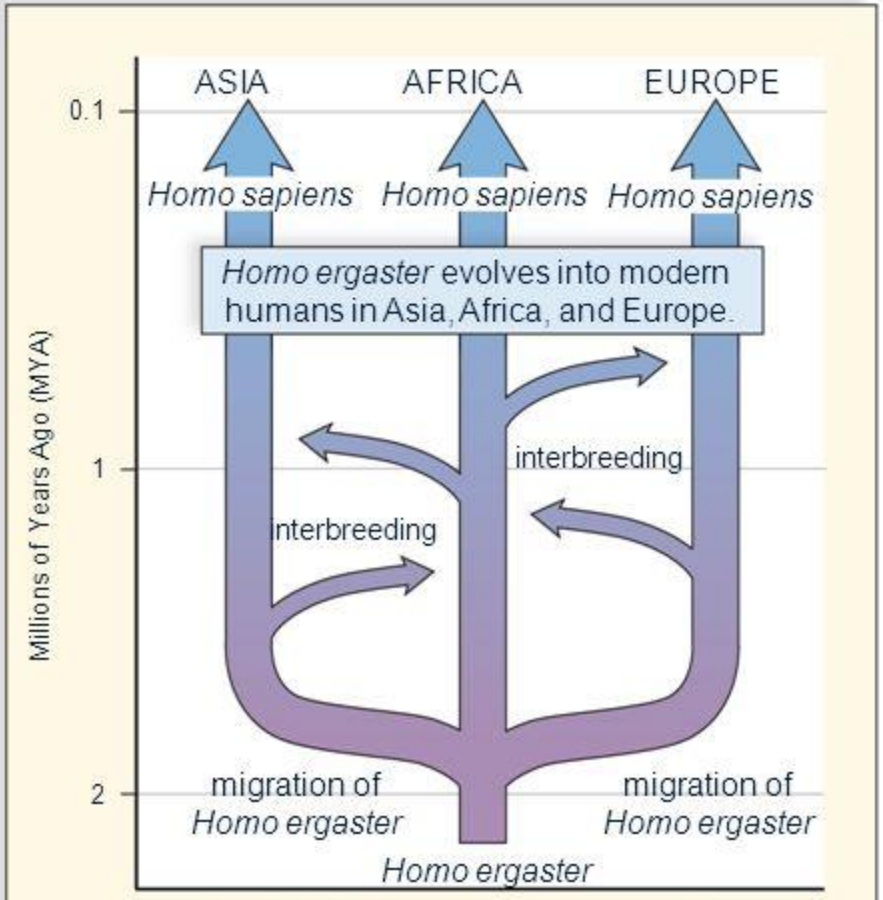
TRENDS in Ecology & Evolution



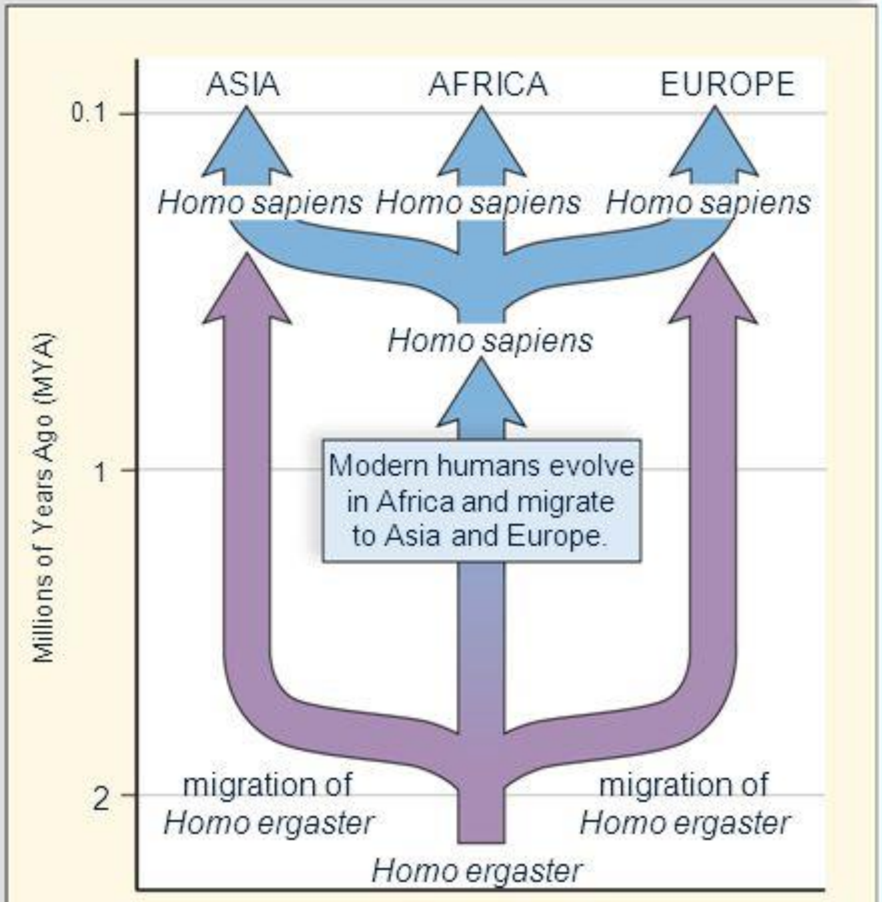
Nature Reviews | Genetics

- There are two competing hypotheses on the origin of modern humans:
- **Out-of-Africa hypothesis and the multiregional hypothesis.**
- Both agree that **Homo erectus originated in Africa and expanded to Eurasia** about one million years ago, but they differ in explaining the origin of modern humans (Homo sapiens sapiens).
- The first hypothesis proposes that a second migration out of Africa happened about 100,000 years ago, in which anatomically modern humans of African origin conquered the world by completely replacing archaic human populations (Homo sapiens; Model A).
- The multiregional hypothesis states that independent multiple origins (Model D) or shared multiregional evolution with continuous gene flow between continental populations (Model C) occurred in the million years since Homo erectus came out of Africa (the trellis theory).+
- A compromised version of the Out-of-Africa hypothesis emphasizes the African origin of most human populations but allows for the possibility of minor local contributions (Model B)

Two Hypotheses for the Origin of Modern Humans


ASIA
AFRICA
EUROPE
0.1
1
2
Millions of Years Ago (MYA)
Homo sapiens
Homo sapiens
Homo sapiens
Homo ergaster evolves into modern humans in Asia, Africa, and Europe.
interbreeding
interbreeding
migration of *Homo ergaster*
migration of *Homo ergaster*
Homo ergaster

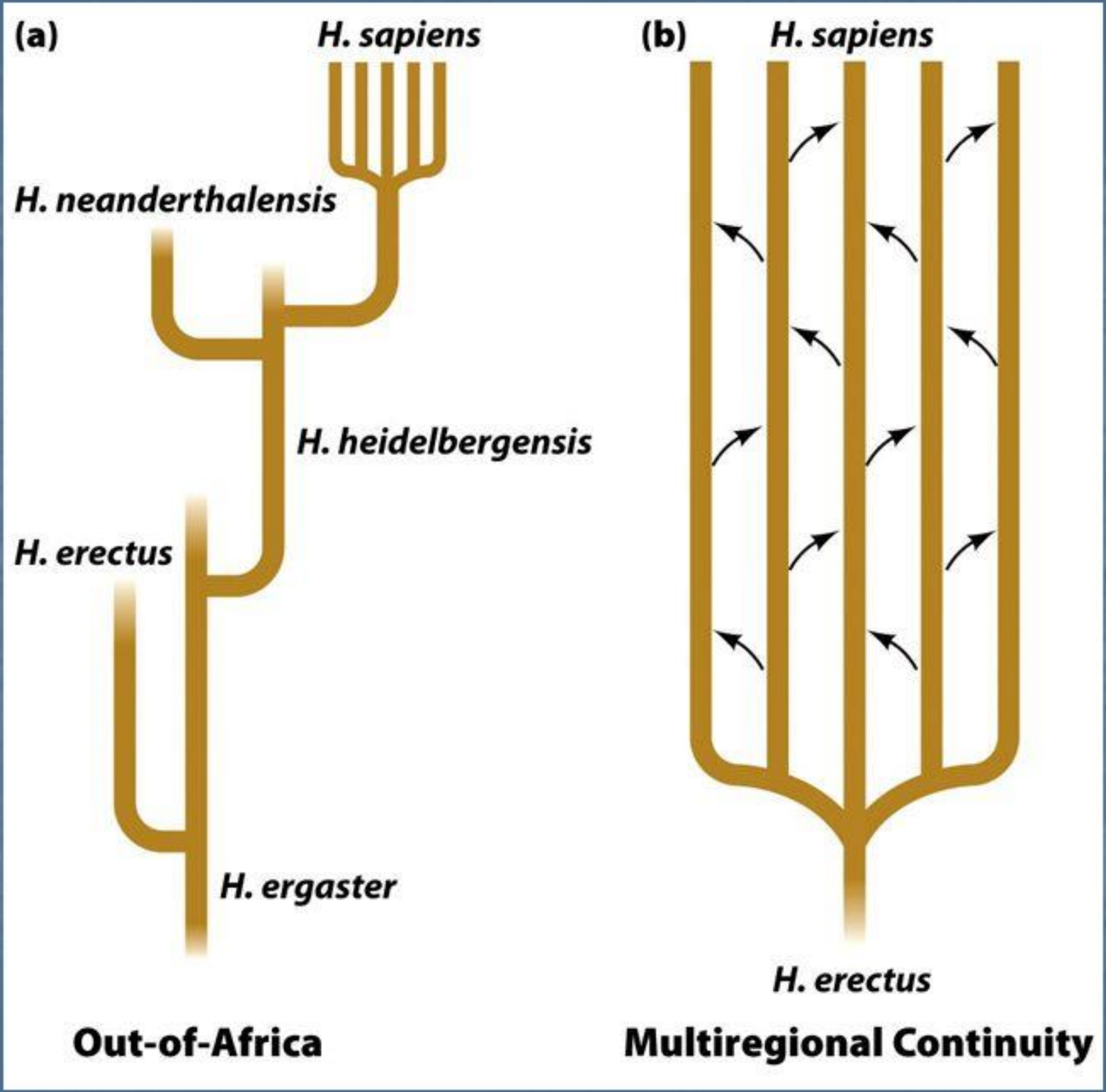
Multiregional continuity


ASIA
AFRICA
EUROPE
0.1
1
2
Millions of Years Ago (MYA)
Homo sapiens
Homo sapiens
Homo sapiens
Homo sapiens
Modern humans evolve in Africa and migrate to Asia and Europe.
migration of *Homo ergaster*
migration of *Homo ergaster*
Homo ergaster

Out of Africa

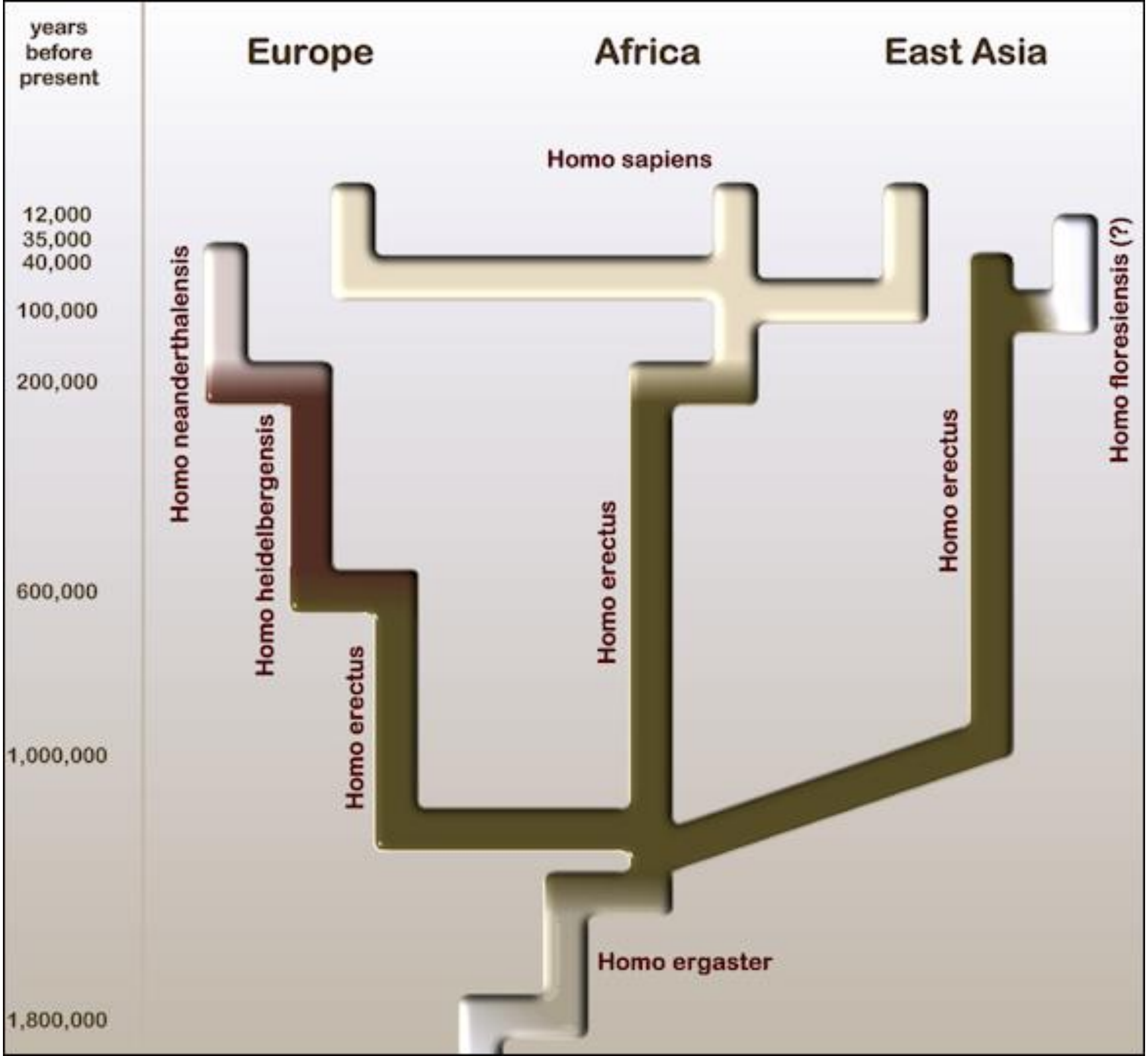
Out-of Africa hypothesis

- The ‘Out-of-Africa’ hypothesis of modern human origins suggests that local populations outside Africa were completely replaced by modern humans who originated in Africa.
- This hypothesis has been widely accepted since in the late 1980s
- It is supported by extensive genetic evidence and by archaeological
- Abundant hominid fossils found in China and in other regions of East Asia show evolutionary continuity, not only in morphological characters, but also in spatial and temporal distributions
- This observation implies that the evolution from Homo erectus to Homo sapiens and then to Homo sapiens sapiens (modern man), took place in East Asia as well as in Africa — a theory that has been used repeatedly to challenge the validity of the **Out-of Africa hypothesis**



Out-of-Africa

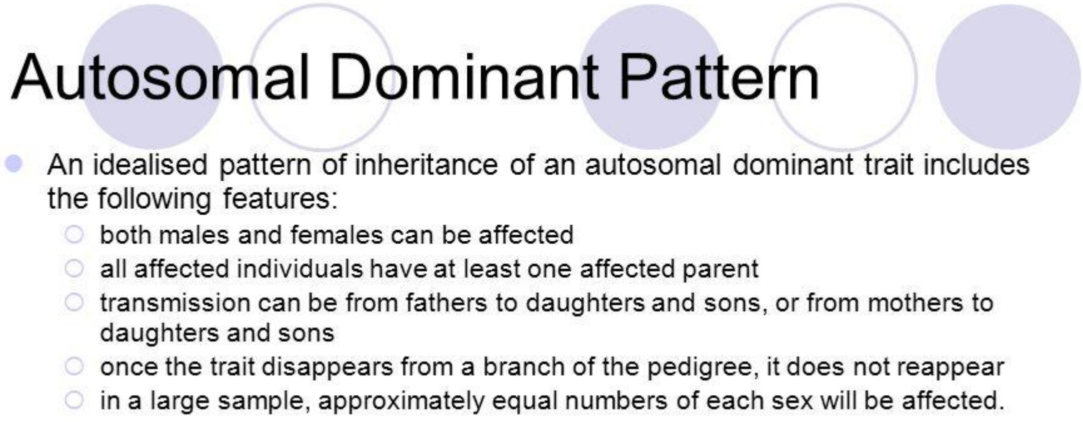
Multiregional Continuity


years before present
Europe
Africa
East Asia
Homo sapiens
12,000
35,000
40,000
100,000
200,000
600,000
1,000,000
1,800,000
Homo neanderthalensis
Homo heidelbergensis
Homo erectus
Homo erectus
Homo erectus
Homo floresiensis (?)
Homo ergaster

Mode of inheritance

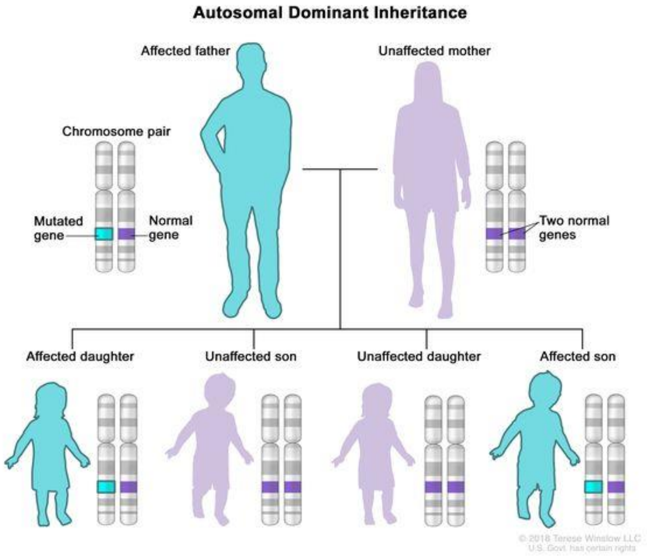
- Dominant mode of inheritance
- Recessive mode of inheritance
- Co-dominant mode of inheritance
- Addictive mode of inheritance

Dominant mode of inheritance



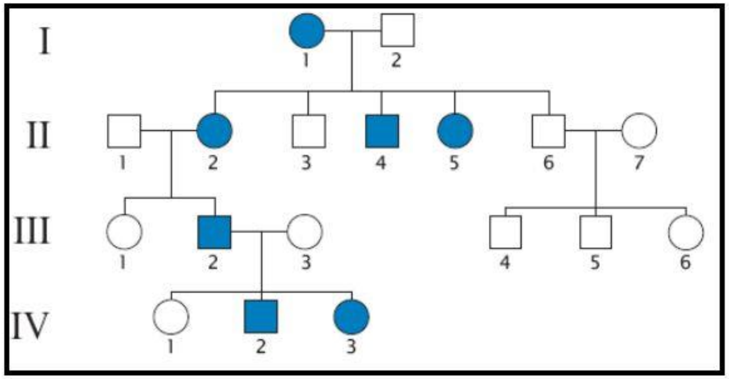
Autosomal Dominant Pattern

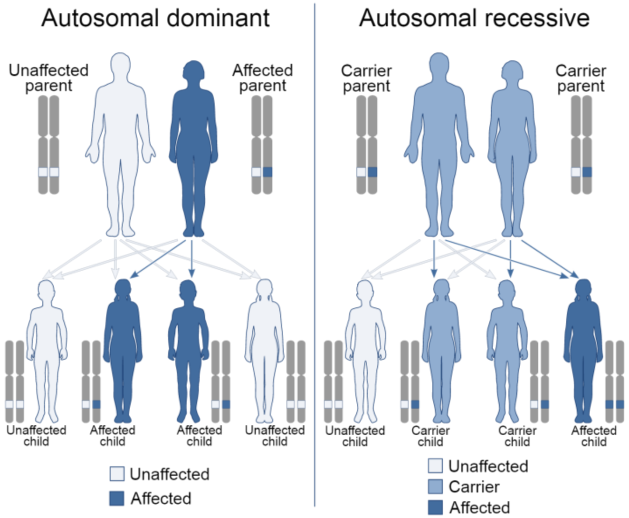
- An idealised pattern of inheritance of an autosomal dominant trait includes the following features:
  - both males and females can be affected
  - all affected individuals have at least one affected parent
  - transmission can be from fathers to daughters and sons, or from mothers to daughters and sons
  - once the trait disappears from a branch of the pedigree, it does not reappear
  - in a large sample, approximately equal numbers of each sex will be affected.



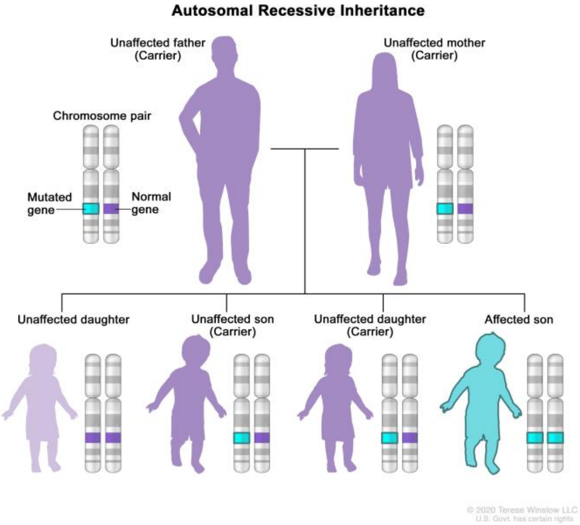
Examples include:

- Huntington disease
- Achondroplasia (a form of dwarfism)
- Familial form of Alzheimer disease
- Defective enamel of the teeth
- Neurofibromatosis (the 'Elephant man' disease)

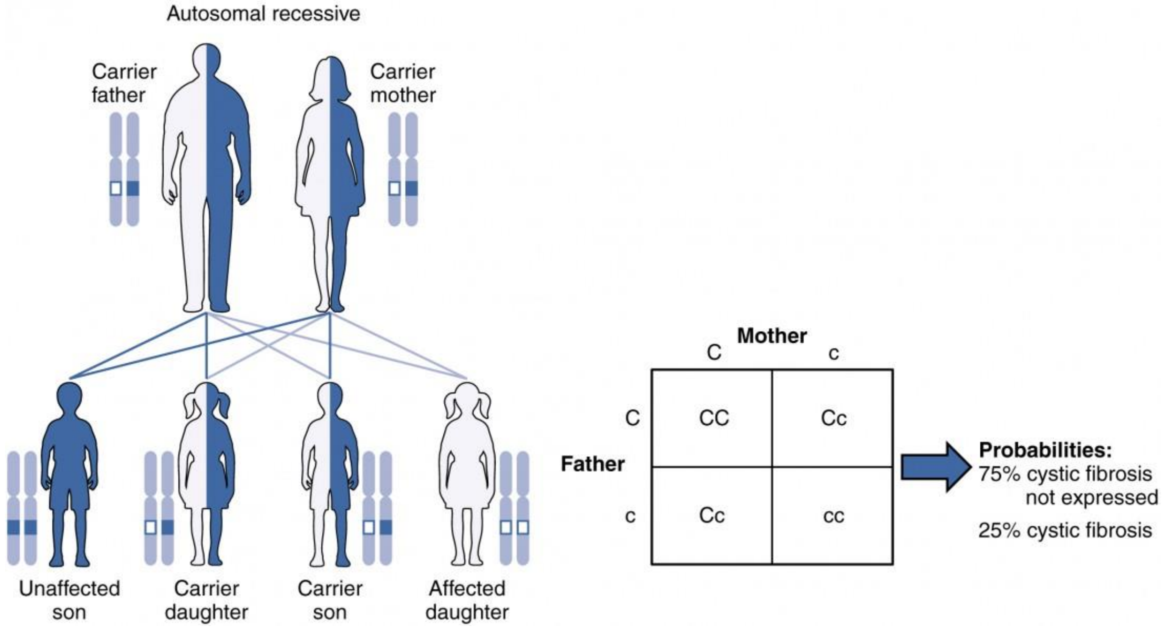




Recessive mode of inheritance

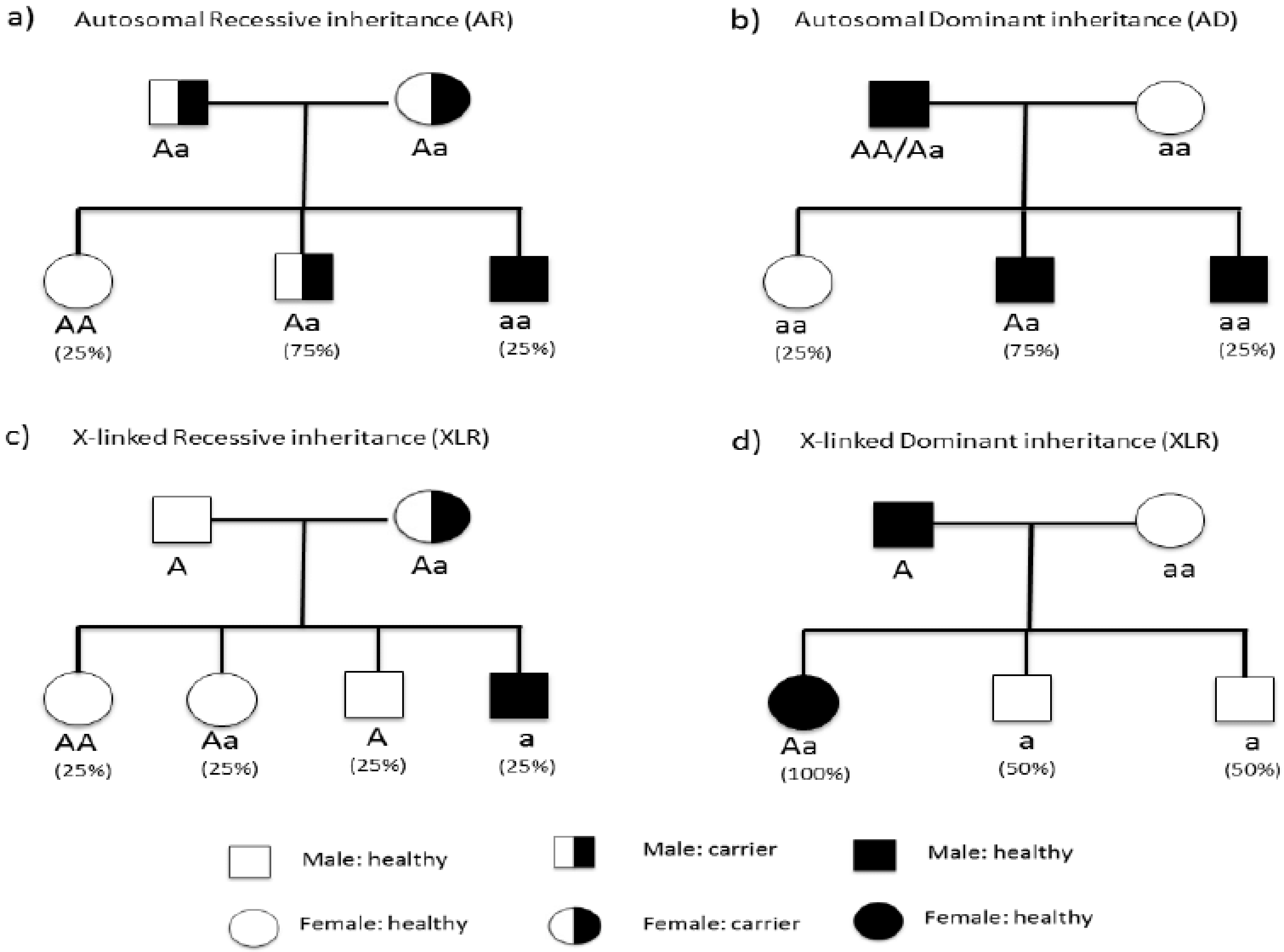


© 2020 Terese Winslow LLC
U.S. Govt. has certain rights



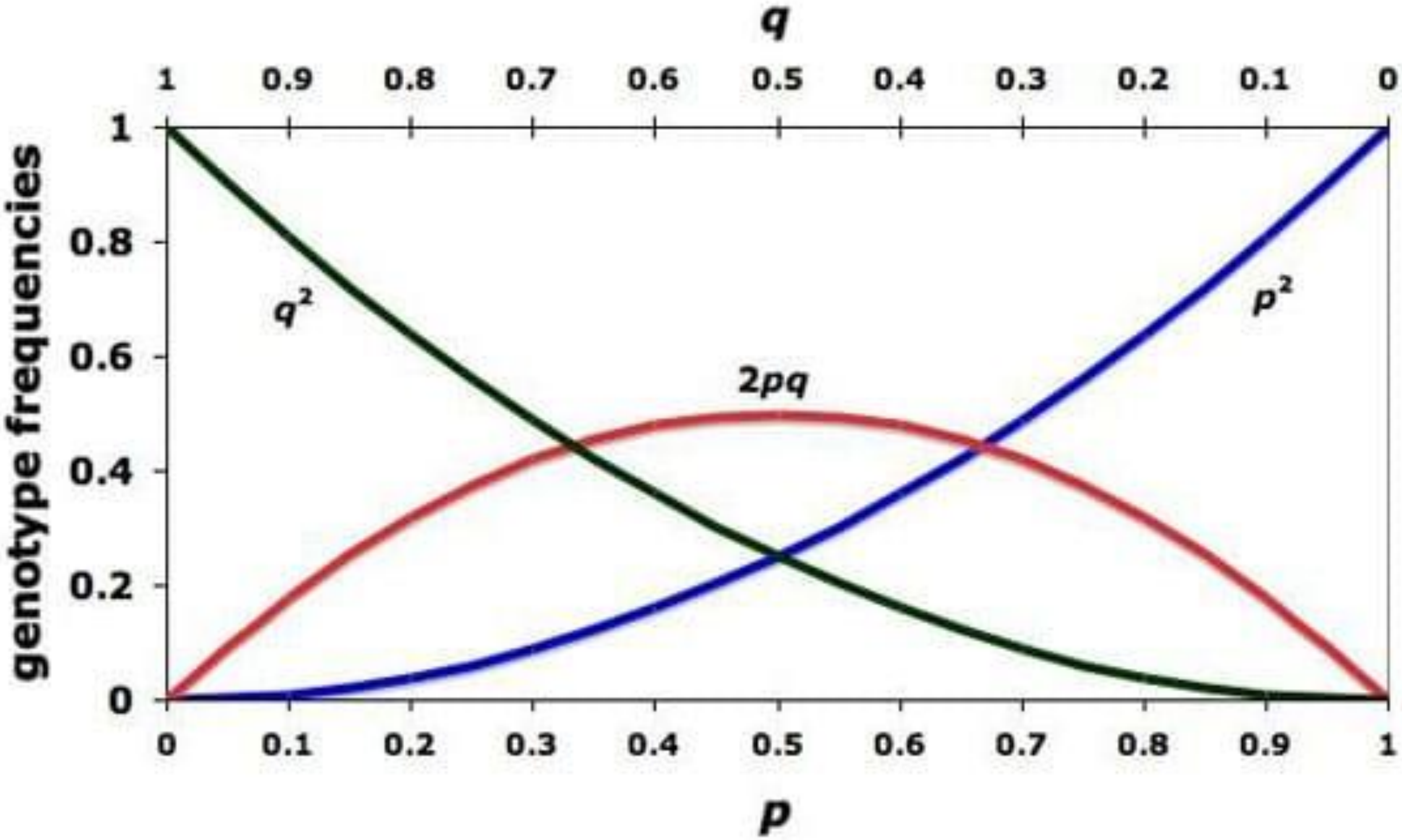
Father \ Mother	C	c
C	CC	Cc
c	Cc	cc

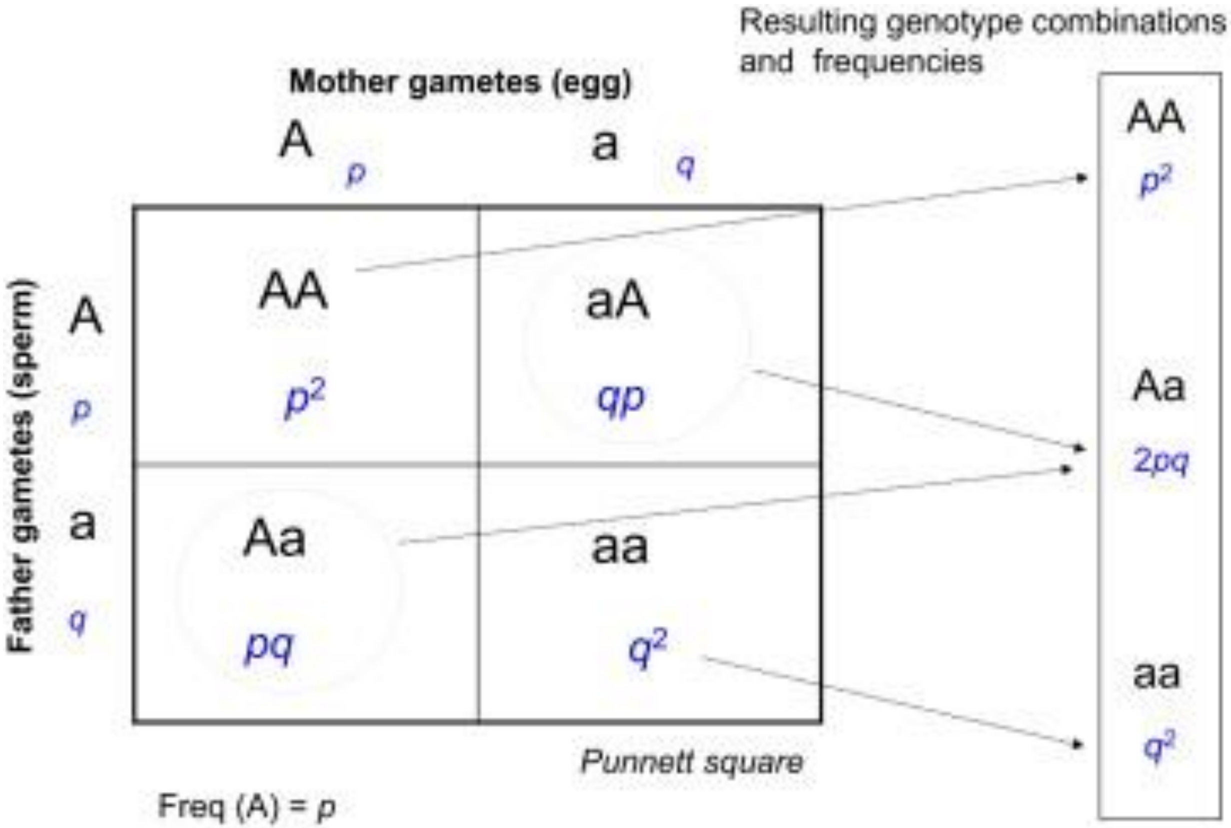
Probabilities:
75% cystic fibrosis
not expressed
25% cystic fibrosis



a) Autosomal Recessive inheritance (AR)
Aa
Aa
AA
(25%)
Aa
(75%)
aa
(25%)
b) Autosomal Dominant inheritance (AD)
AA/Aa
aa
aa
(25%)
Aa
(75%)
aa
(25%)
c) X-linked Recessive inheritance (XLR)
A
Aa
AA
(25%)
Aa
(25%)
A
(25%)
a
(25%)
d) X-linked Dominant inheritance (XLR)
A
aa
Aa
(100%)
a
(50%)
a
(50%)
Male: healthy
Male: carrier
Male: healthy
Female: healthy
Female: carrier
Female: healthy

Hardy weignberg equilibrium





Resulting genotype combinations and frequencies

Mother gametes (egg)

Father gametes (sperm)

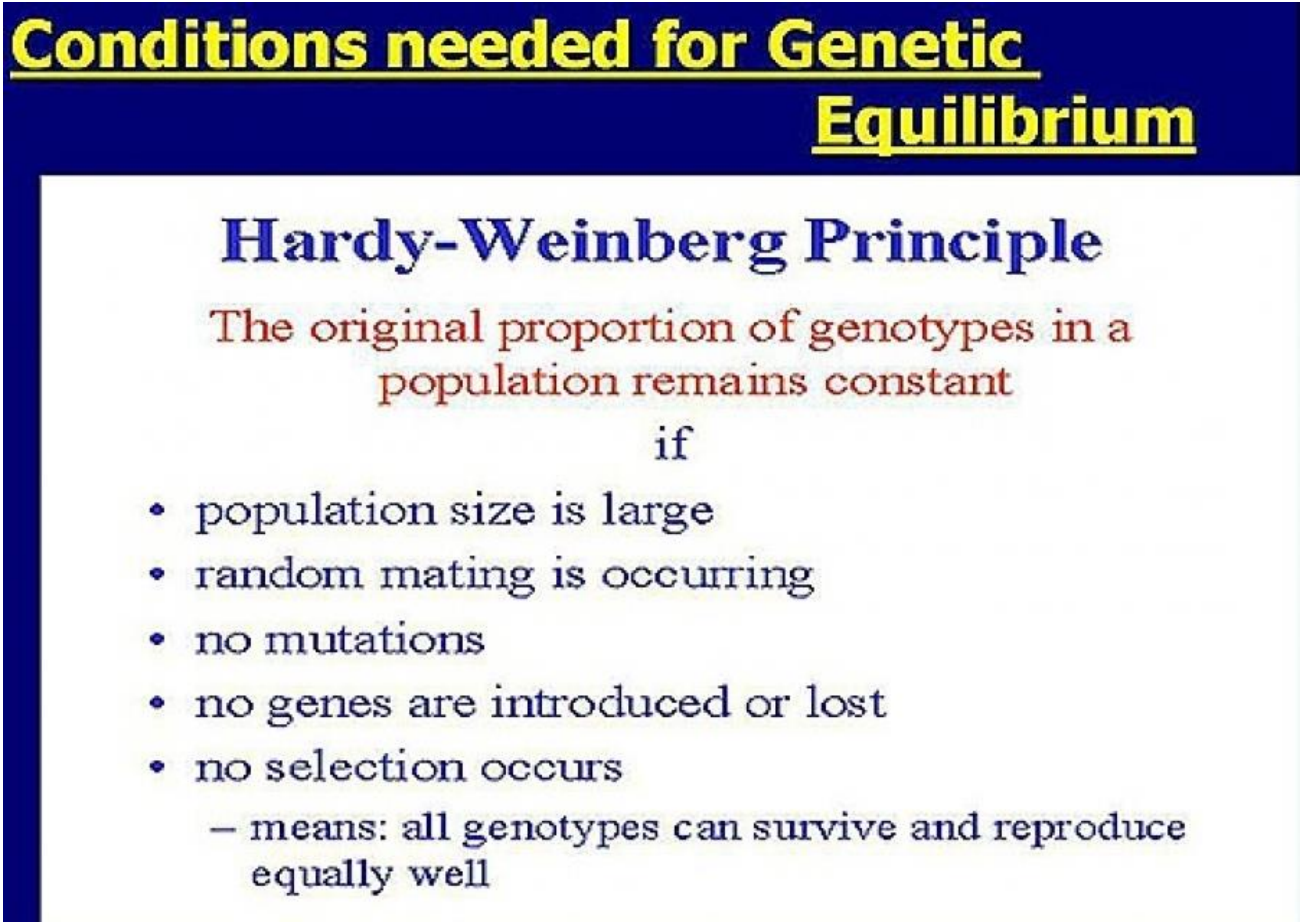
	A p	a q
A p	AA p^2	aA qp
a q	Aa pq	aa q^2

Punnett square

AA	p^2
Aa	$2pq$
aa	q^2

Freq (A) = p

Freq (a) = q $\quad p + q = 1 \quad (p + q)^2 = p^2 + 2pq + q^2$



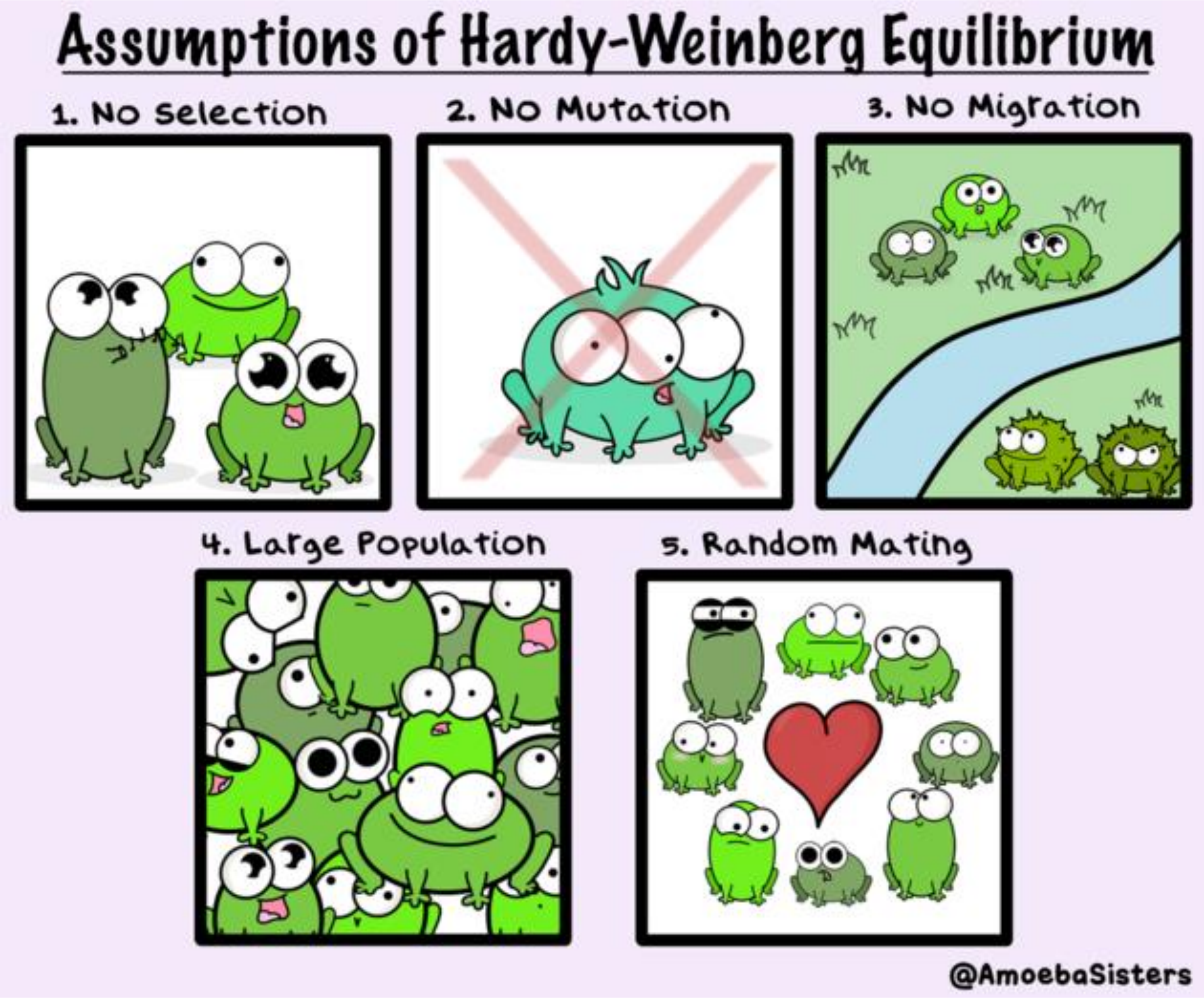
Conditions needed for Genetic Equilibrium

Hardy-Weinberg Principle

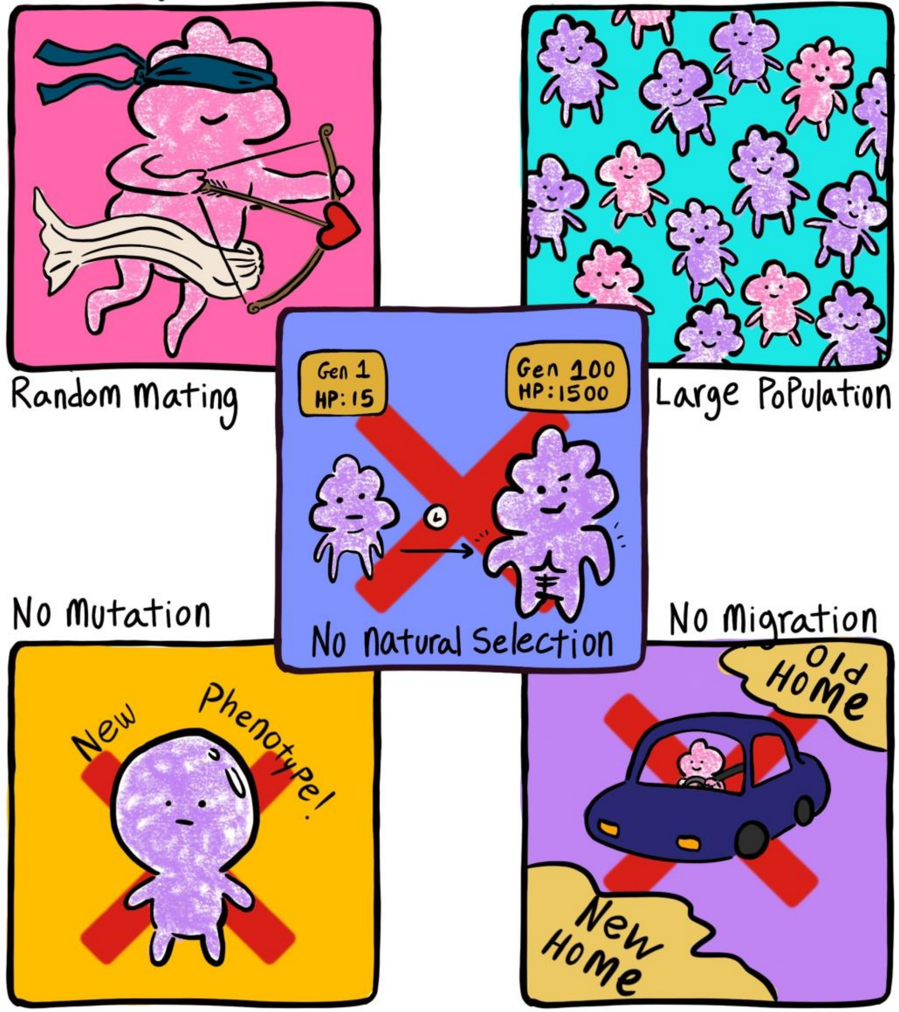
The original proportion of genotypes in a population remains constant

if

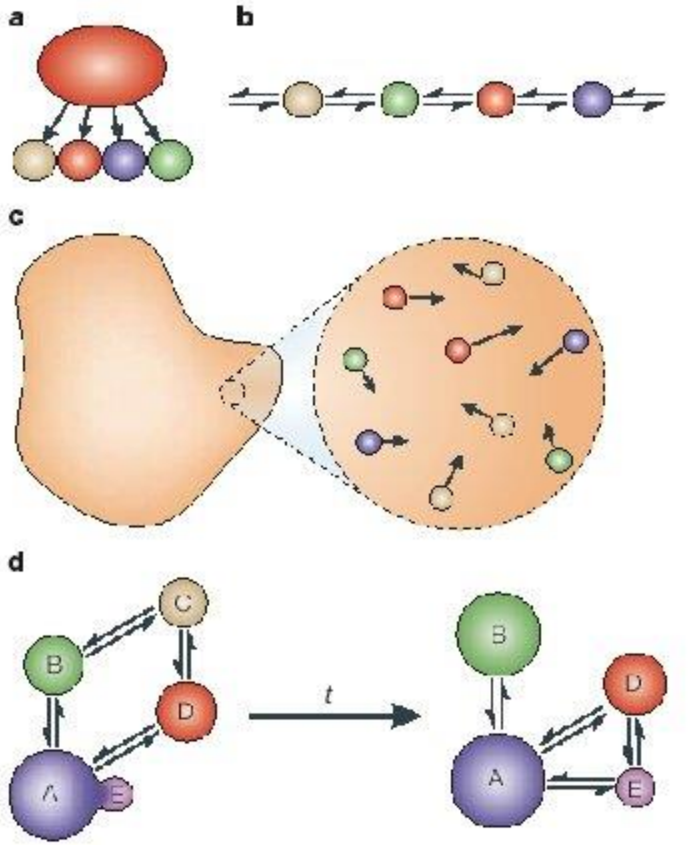
- population size is large
- random mating is occurring
- no mutations
- no genes are introduced or lost
- no selection occurs
  - means: all genotypes can survive and reproduce equally well


Assumptions of Hardy-Weinberg Equilibrium
1. No Selection
2. No Mutation
3. No Migration
4. Large Population
5. Random Mating
@AmoebaSisters

Hardy-Weinberg Assumptions



Models of population structure that allow for migration (Gene Flow)



Idealized Population Models

a. Island model

b. Stepping stone

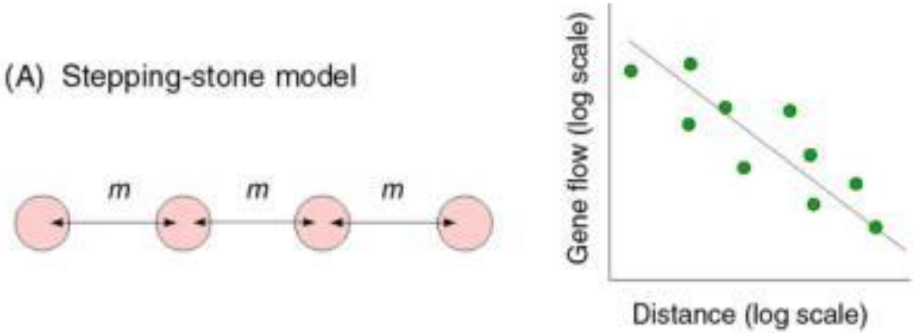
c. Isolation by distance

d. Metapopulation

Dispersal models

Continuous populations

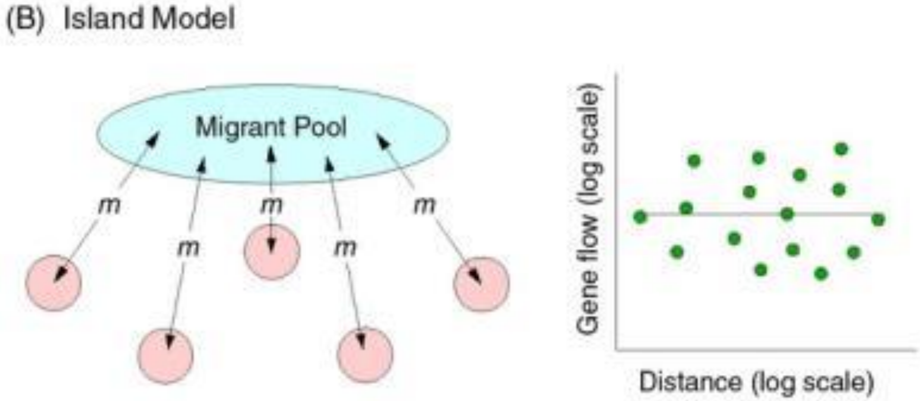
- Isolation-by-distance



Kimura, M., and W. H. Weiss. 1964. The stepping stone model of genetic structure and the decrease of genetic correlation with distance. Genetics 49:561-576.

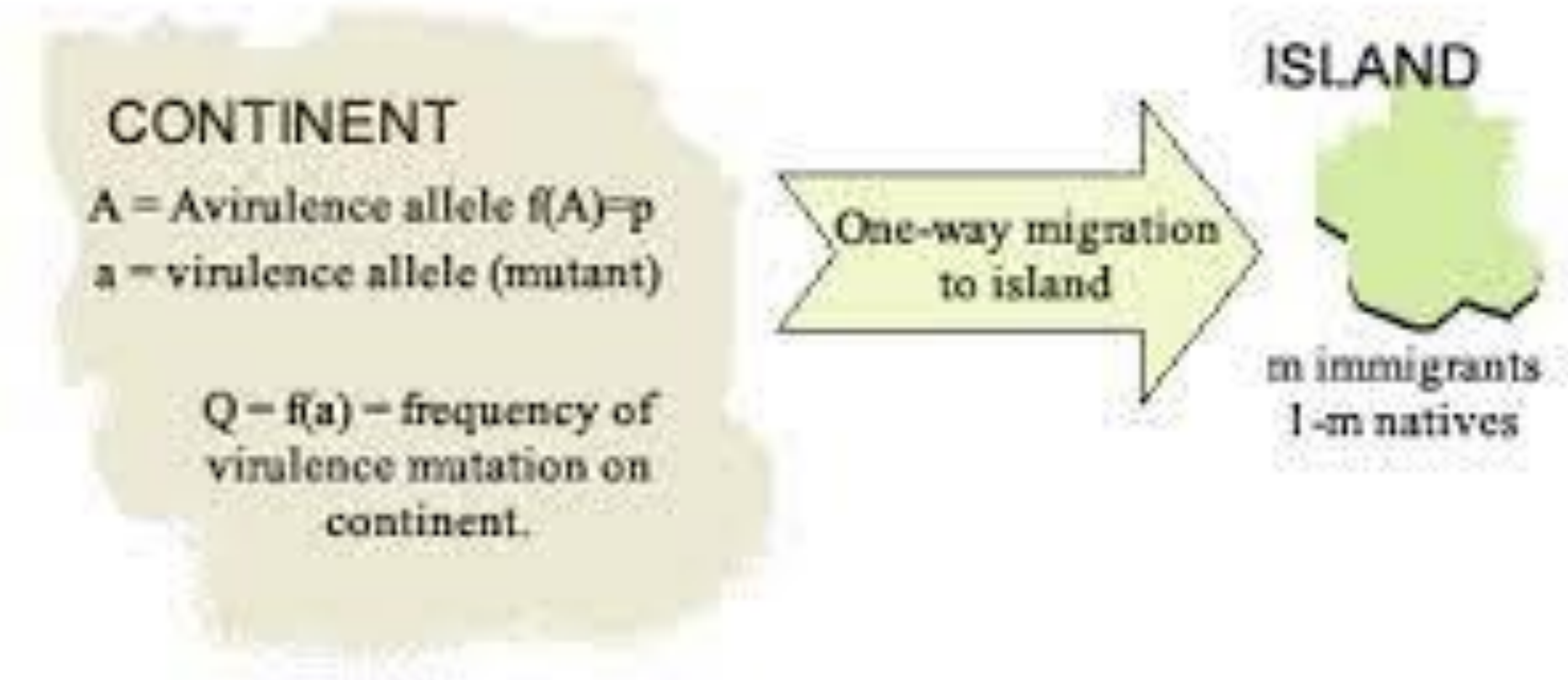
Discrete populations

- Stepping-stone
- Island model



Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97-159..

Continent-island model of Sewall Wright



CONTINENT

A = Avirulence allele f(A)=p

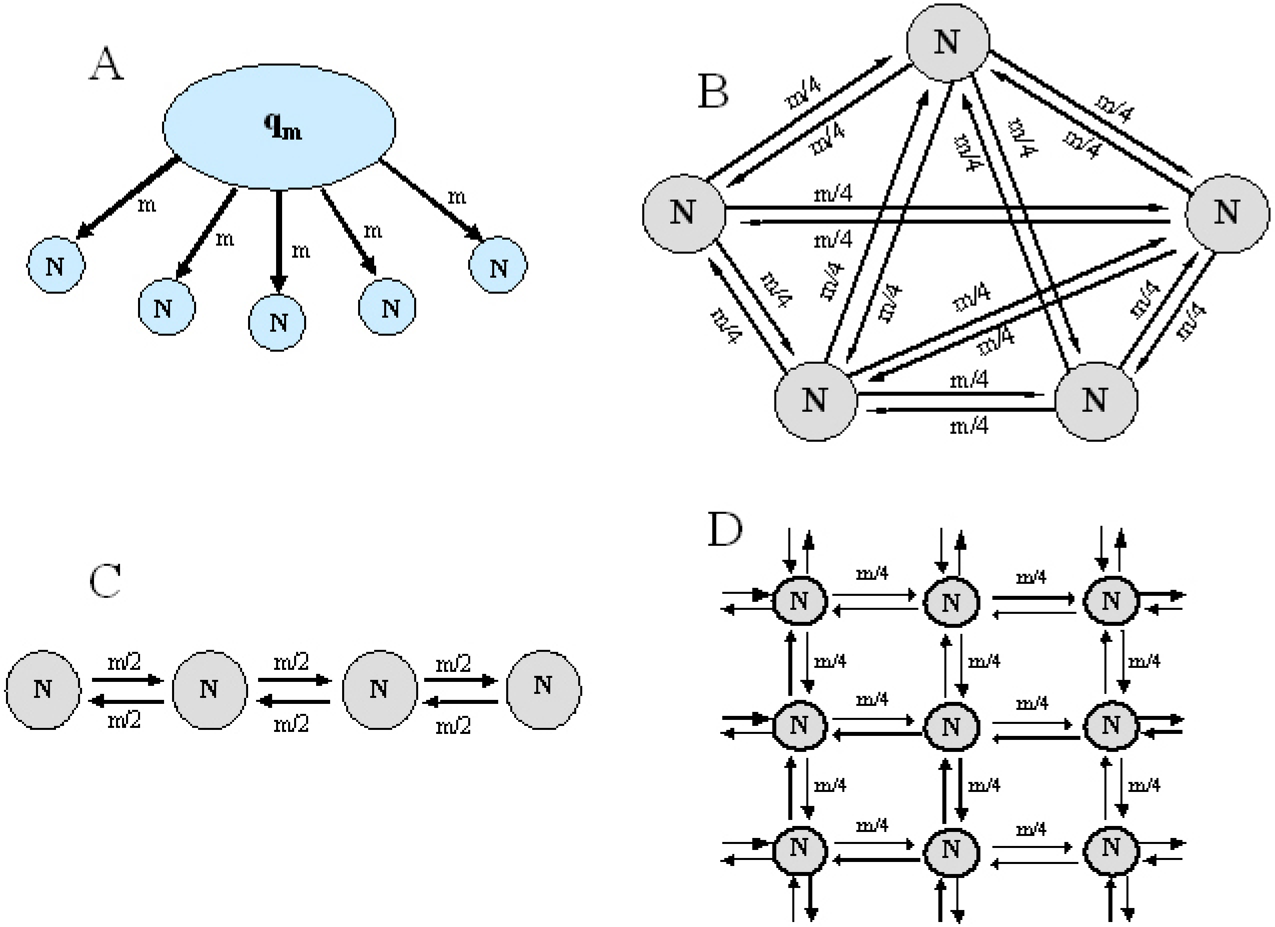
a = virulence allele (mutant)

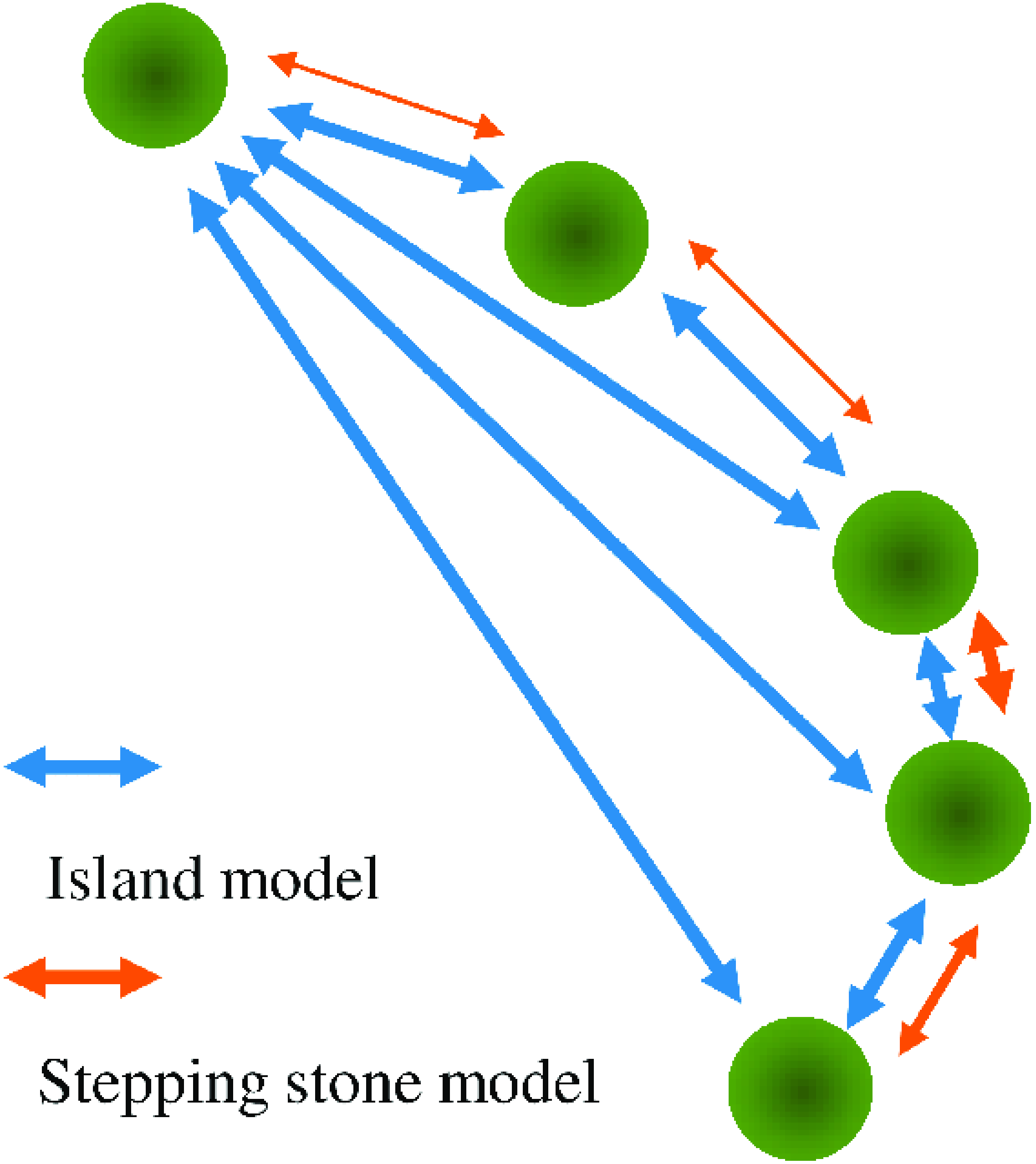
Q = f(a) = frequency of virulence mutation on continent.

One-way migration to island

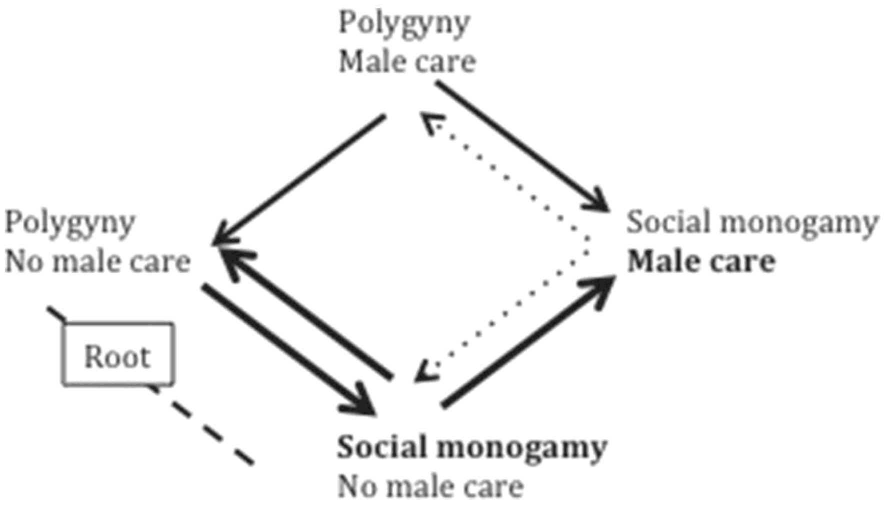
ISLAND

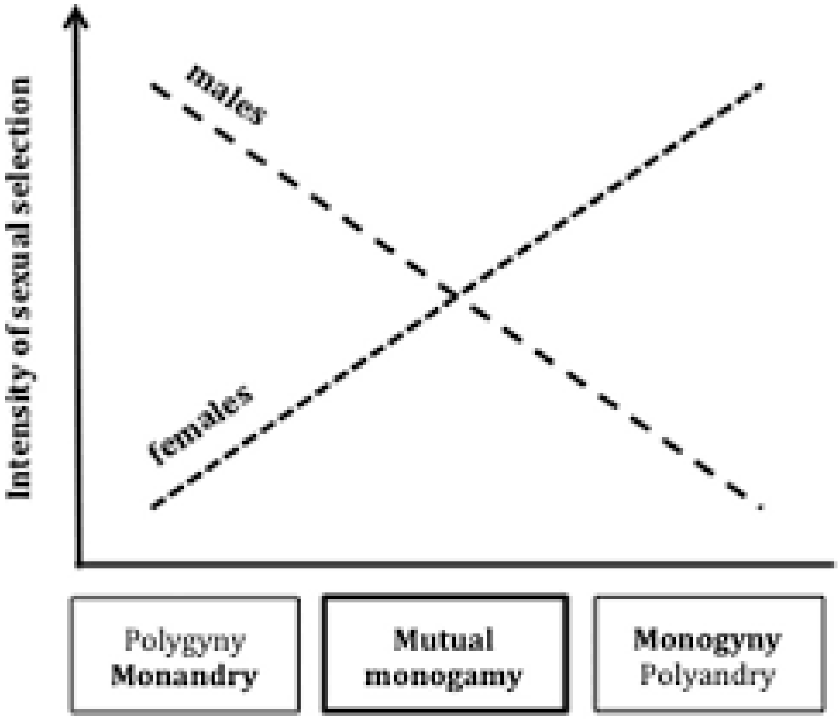
m immigrants
1-m natives


A
q_m
m
m
m
m
m
N
N
N
N
N
B
m/4
N
N
N
N
N
C
m/2
m/2
m/2
m/2
m/2
m/2
N
N
N
N
D
m/4
N

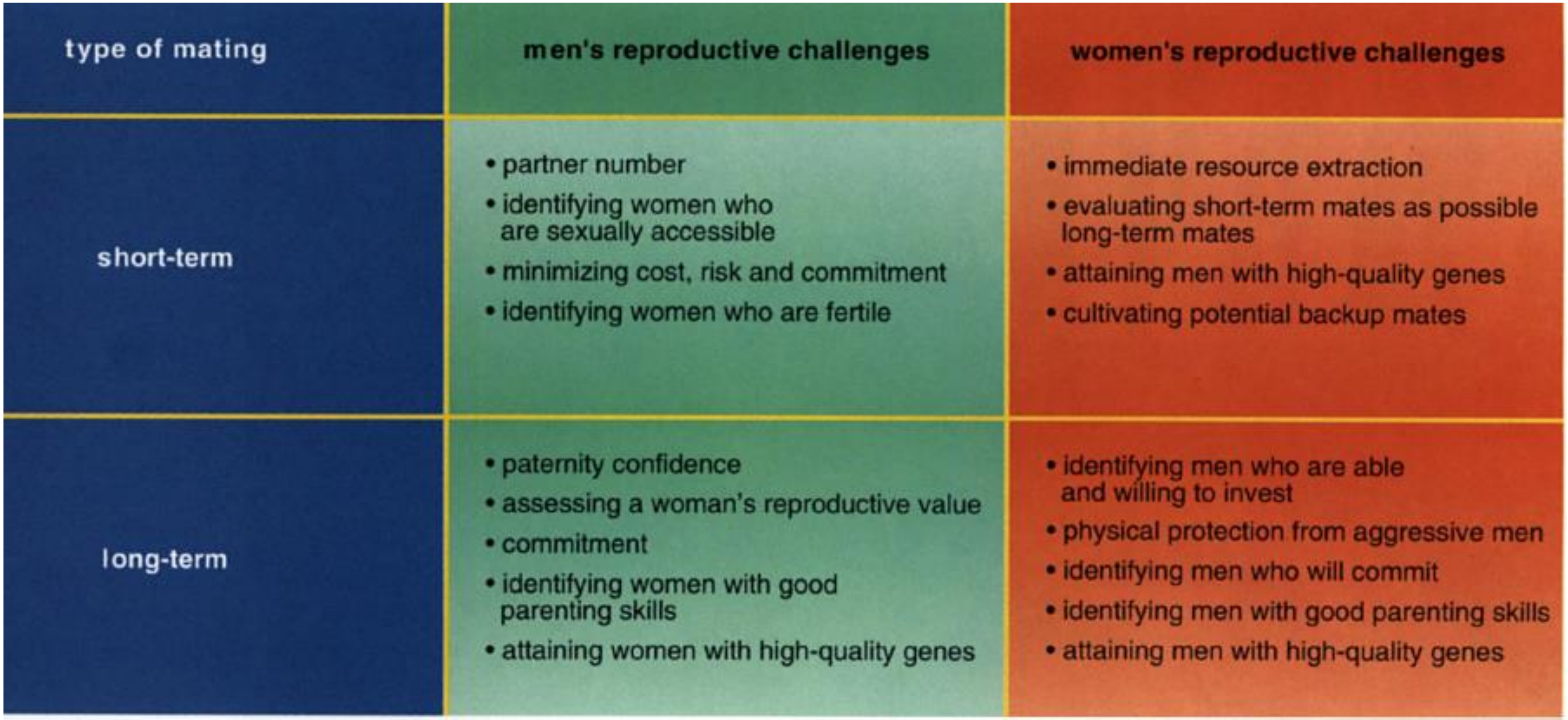

Island model
Stepping stone model

- **Mating pattern:** shown by an individual, but can be shared by many individuals within a population or species; a description of how many partners a male and/or female has, often loosely classified as monogamy or polygamy, or, more strictly, as monandry, monogyny, mutual monogamy, polyandry, polygyny or polygynandry.





- **Short-term mating pattern:** when a mating pattern only lasts over one breeding event (also referred to as serial or sequential; thus, for example, serial monogyny becomes polygyny over time).
- **Long-term mating pattern:** when a mating pattern lasts over multiple breeding events.

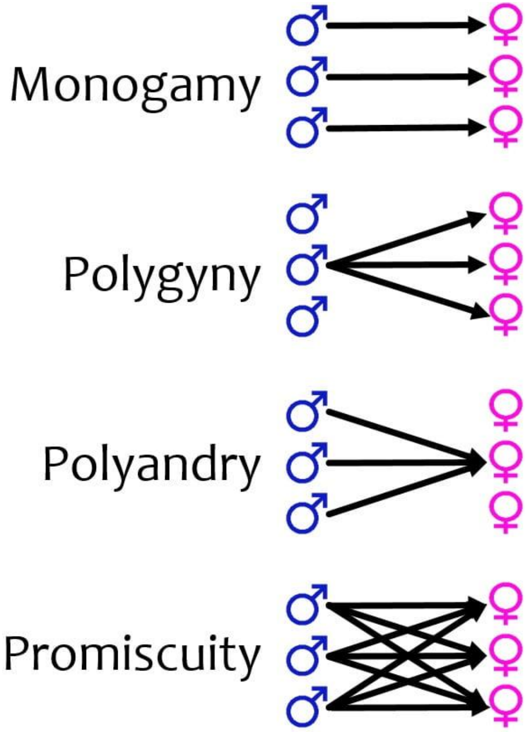


type of mating	men's reproductive challenges	women's reproductive challenges
short-term	• partner number • identifying women who are sexually accessible • minimizing cost, risk and commitment • identifying women who are fertile	• immediate resource extraction • evaluating short-term mates as possible long-term mates • attaining men with high-quality genes • cultivating potential backup mates
long-term	• paternity confidence • assessing a woman's reproductive value • commitment • identifying women with good parenting skills • attaining women with high-quality genes	• identifying men who are able and willing to invest • physical protection from aggressive men • identifying men who will commit • identifying men with good parenting skills • attaining men with high-quality genes

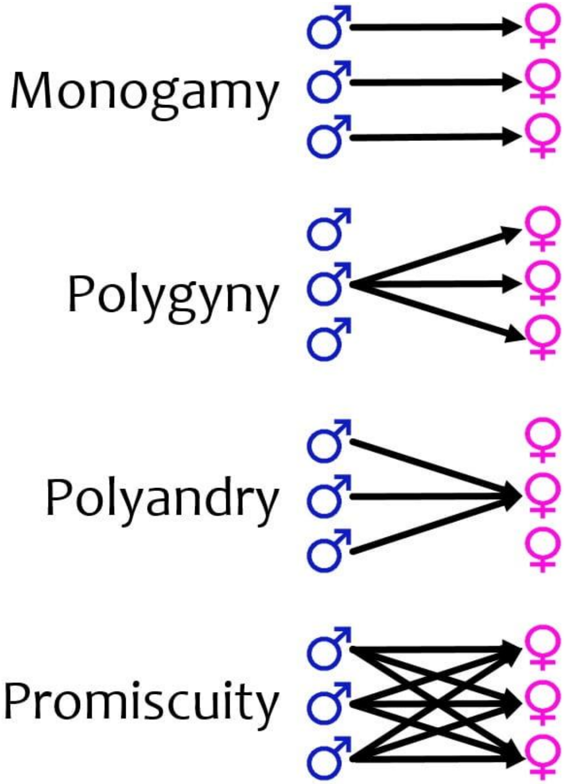
Social mating pattern

- A mating pattern that is based on behavioural observations of males and females.
- Often inferred indirectly by **formation or lack of pair-bonds**.
- **Social mating patterns** are interesting in their own right but at least in some taxa are poor indicators of actual reproduction.
- **Reproductive mating pattern:**
  - used to describe a mating pattern that is based on the sexual or genetic mating pattern:
- **Genetic mating pattern:**
  - mating pattern determined using DNA samples from offspring, showing which males and females are reproducing together.
- **Sexual mating pattern:**
  - mating pattern based on observed matings (copulations or spawnings), indicating which males and females are reproducing together.
- The sexual mating pattern may differ from the genetic mating pattern because not all matings are easily observed, and not all matings result in fertilised offspring, because many factors (timing of mating, first/last male sperm precenence, sperm competition, cryptic mate choice, etc.) can influence whether, for example, extra-pair copulations result in extra-pair young.

- **Monogamy:** males and females typically mate and reproduce with only one partner:
- **Monandry:** each female mates and reproduces with only one male
- **Monogyny:** each male mates and reproduces with only one female
- **Mutual monogamy (monogynandry):** one male and one female mate and reproduce only with each other



- **Polygamy:** males and females typically mate and reproduce with more than one partner
- **Polyandry:** females mate and reproduce with more than one male
- **Polygyny:** males mate and reproduce with more than one female
- **Polygynandry:** both sexes mate and reproduce with multiple mates

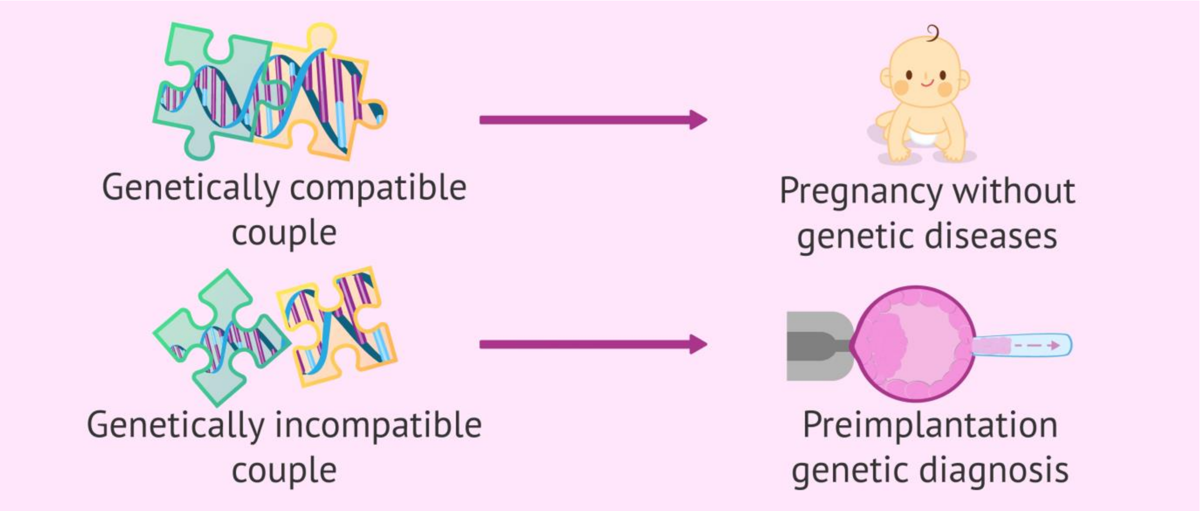


- **Facultative monogamy:** monogamy that varies, within species or even within individuals, for example in relation to density of mates or some other resource, such as habitat suitable for breeding or feeding.
- **Type I monogamy.**
- occurs when the male is not fully committed to one female.
- But, he chooses to stay with her because there are no other mating opportunities available to him.
- occurs because of low density.
- The species rarely spend time with their families.
- There is a lack of paternal care for the offspring.
- Elephant shrews, Agoutis, Grey duikers and Pacaranas are the most common examples of facultative monogamous animals.

- **Obligate monogamy:** monogamy that does not vary in relation to density of mates or other resources (cf . facultative monogamy).
- **Type Il monogamy**
- practiced by species that live in overlapping territories
- occurs females cannot rear their young without the help of their partners.
- Factors associated with Type Il monogamy:
  - High paternal investment
  - Delayed sexual maturation in juveniles
  - Juveniles contributing greatly to the rearing of their sibling
- **Breeding event:** usually a brood of offspring. Depending on the organism studied, there can be few or many breeding events (or brood cycles) within a breeding season, over a lifetime, etc.
- **Degree of breeding synchrony:** proportion of males or females that are fertile at the same time.

- **Pair-bonding behaviour:** behavioural association between a male and female, often indicating that they are breeding together, based on, for example, joint care of young, joint defence of a breeding or feeding territory, mutual grooming, courtship or greetings.
- Social mate: a partner that shows pair-bonding behaviour, regardless of whether joint reproduction occurs.
- **Extra-pair copulations:** for socially monogamous animals, the term refers to copulations that occur outside the social pair.
- **Extra-pair paternity:** number or proportion of young in a brood fathered by other male(s) than the social mate.
- Intra-pair (or within pair) paternity: number or proportion of young in a brood that is fathered by the social mate.
- Genetic incompatibility: inviability of offspring caused by negative interactions between maternally and paternally inherited genetic elements.
- Genetic compatibility: increased fitness of offspring generated through positive interactions between maternally and paternally inherited genes or genetic elements.

- **Genetic incompatibility:** inviability of offspring caused by negative interactions between maternally and paternally inherited genetic elements.
- **Genetic compatibility:** increased fitness of offspring generated through positive interactions between maternally and paternally inherited genes or genetic elements.


Genetically compatible couple
Pregnancy without genetic diseases
Genetically incompatible couple
Preimplantation genetic diagnosis

Adult sex ratio

- Adult sex ratio: the number of sexually mature males to females in a population, expressed as a ratio (m/f), or as a proportion (m/m + f)

Monogamy

- Monogamy refers to a mating system in which males and females typically mate with only one partner.
- Causes of Monogamy
  - (i)spatial constraints (habitat limitation, mate availability),
  - (ii) time constraints (breeding synchrony, length of breeding season),
  - (iii) need for parental care, and
  - (iv) genetic compatibility,
- Important consequences of reproductive monogamy:
  - (i) parentage,
  - (ii) parental care,
  - (iii) eusociality and altruism,
  - (iv) infanticide,
  - (v)effective population size,
  - (vi) mate choice before mating,
  - (vii) sexual selection, and
  - (viii) sexual conflict.

- **Social relationships:**
- **Social Monogamy.** by the socio-sexual relationship between two animals.
- The pair bond exists outside the times of courtship and copulation, either before or after mating.
- Paternal care may or may not exist. The maintenance of a socio-sexual relationship does not preclude the possibility for extra-pair mating by either sex.
- This system appears to be the most common in birds passerine

Polygyny

- The socio-sexual relationship where one male has a specific relationship or bond with more than one female that lasts outside the period of courtship and mating. This male typically mates with all sexually mature females to which he is bonded. Paternal care may or may not occur. The maintenance of a relationship does not preclude the possibility for extra-group mating by either sex.
- Classic examples of this type of mating system includes red-winged blackbirds and black-tailed prairie dogs

Polyandry

- The socio-sexual relationship where one female has a specific relationship or bond with more than one male that lasts outside the period of courtship and mating. This female typically mates with all sexually mature males to which she is bonded.
- Paternal care may or may not occur. Sex-role reversal is typical, with females often having more exaggerated phenotypes.
- The maintenance of a social relationship does not preclude the possibility for extra group mating by either sex.
- Examples of this social mating system include Jacana

Polygynandry

- The socio-sexual relationship between >1 male and >1 female that lasts outside the period of courtship and mating.
- All animals typically mate with all other opposite-sexed animals.
- Paternal care may or may not occur.
- The maintenance of a social relationship does not preclude the possibility for extra-group mating by either sex.
- Classic examples include acorn woodpeckers

Promiscuity

- The socio-sexual system that is characterized by the **lack of pair bonding**. Typically, both males and females mate with multiple individuals of the opposite sex.
- Classical lekking species

Genetic relationships

- Genetic monogamy. The genetic outcome of mating, and is characterized by
  - (1) the absence of extra pair males in the resultant clutch, brood, or litter and
  - (2) the absence of paternal representation in any other female clutch, brood, or litter.
- This genetic outcome may be associated with social monogamy, but this link is not necessary.

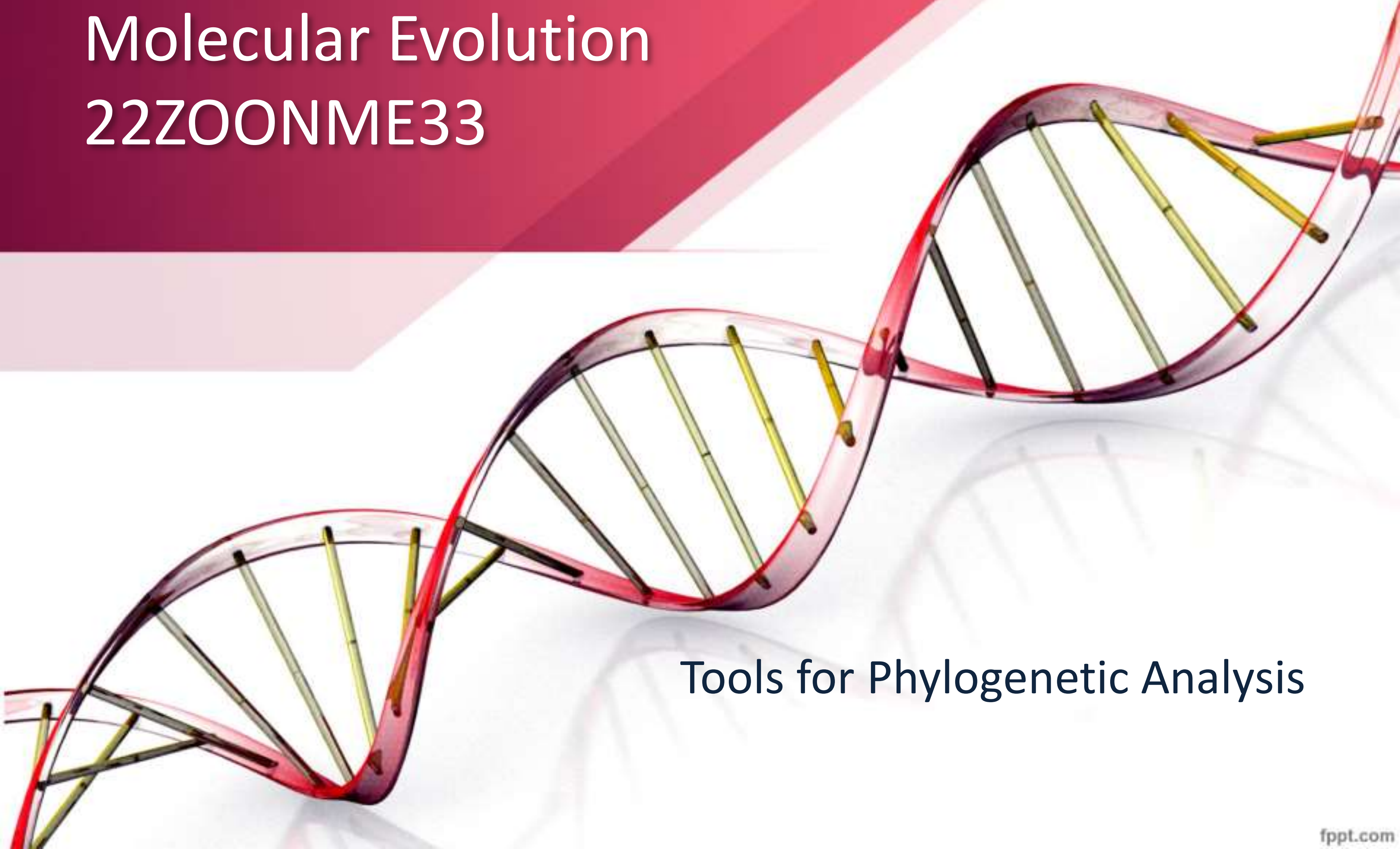
Multiple male mating (MMM).

- The system where only males may mate multiply, and these matings result in fertilizations.
- Systems where socially monogamous males engage in extra pair fertilizations with other females would fall into this category.
- **Multiple female mating (MFM).** Multiple female mating describes the system where only females may mate multiply, and these matings result in fertilizations.
- A typical example of this is a system where females solicit matings from unpaired males.
- Genetic Promiscuity. Genetic promiscuity describes the situation where males sire offspring with multiple females, and females bear multiply sired clutches, broods, or litters. Polygynandry cannot be distinguished from promiscuity when looking only at genetic data

Multiple female mating (MFM).

- The system where only females may mate multiply, and these matings result in fertilizations.
- A typical example of this is a system where females solicit matings from unpaired males.
- Genetic Promiscuity. situation where males sire offspring with multiple females, and females bear multiply sired clutches, broods, or litters.
- Polygynandry cannot be distinguished from promiscuity when looking only at genetic data

Molecular Evolution
22ZOONME33



Tools for Phylogenetic Analysis

fppt.com

Resources

- NCBI-National Center for Biotechnology Information (NIH)
- EBI- European Bioinformatics Institute-
- ExPasy- Expert Protein Analysis System
- Types of Databases
- Nucleotide Sequence Database
  - Genbank -allows for the storage of information in addition to a DNA/protein sequence
  - DDBJ (DNA Data Bank of Japan)- Provides freely available nucleotide sequence data and supercomputer system
  - EMBL(European Molecular Biology Laboratory)- computational modelling, statistical data analysis, or image analysis.

fppt.com

Resource

- Protein Sequence Database
  - Swissport
    - SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, structure of its domains, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
  - TrEMBL
    - TrEMBL consists of entries in a SWISS-PROT format that are derived from the translation of all coding sequences in the EMBL nucleotide sequence database, that are not in SWISS-PROT
  - PIR
    - The Protein Information Resource (PIR) is an integrated public resource of protein informatics that supports genomic and proteomic research and scientific discovery.

fppt.com

Resources

- Protein Structure Database - PBD
- Literature Databases - Pubmed, OMIM
- Chemical Databases - pubchem
- Metabolic pathway databases- KEGG (Kyoto Encyclopedia of Genes and Genomes)
- MEGA
- Phylip

fppt.com

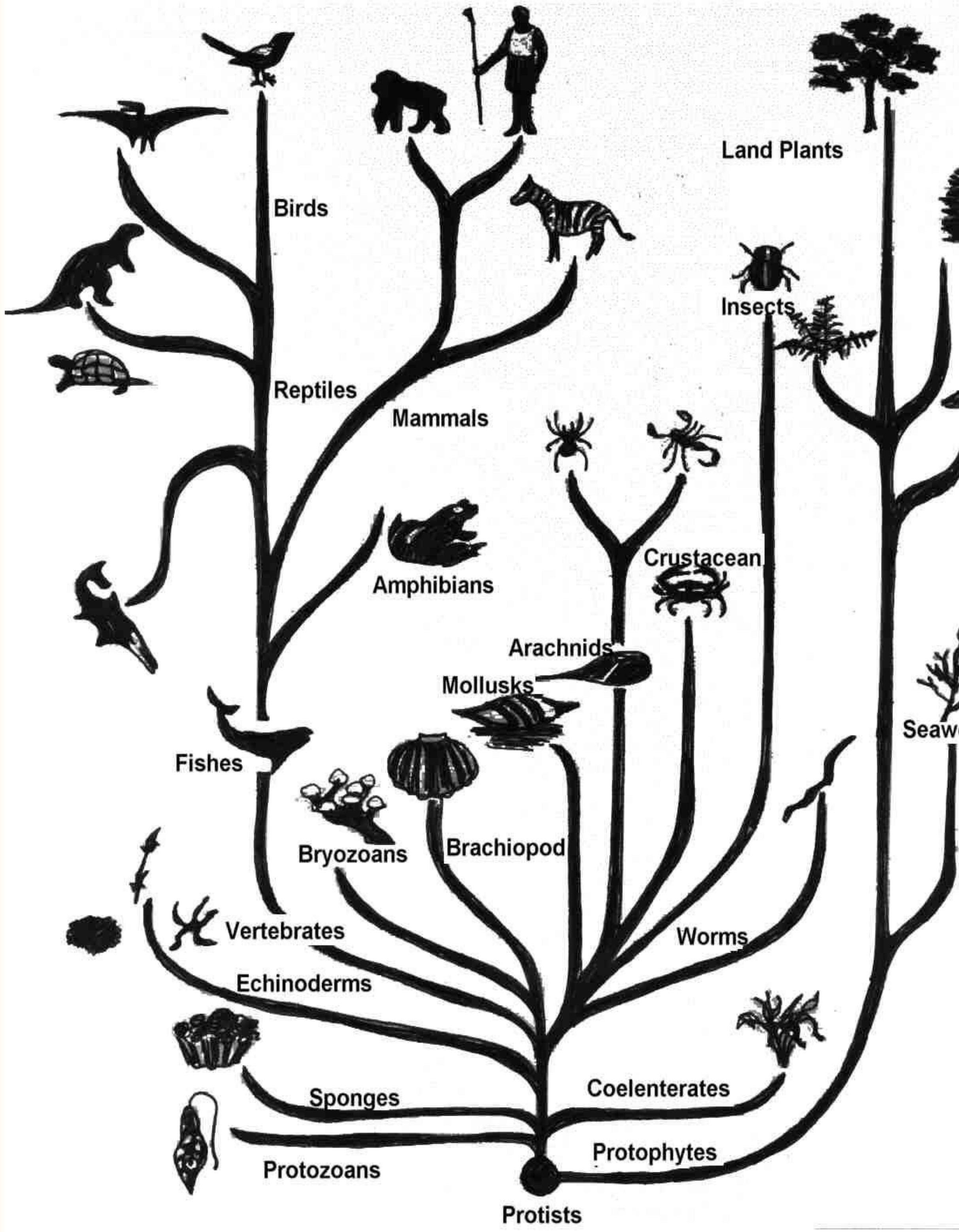
Molecular Evolution

22ZOONME33

Phylogenetic Tree Reconstruction

Phylogenetic Tree Reconstruction

- Mathematically model for the process of molecular evolution, or genome sequences changing over time.
- To explore new mutations occur in genetic sequences and are passed down through generations.
- Mathematical tools to understand branch length in phylogenetic trees and compare different methods for reconstructing evolutionary relationships between species based on their genetic sequences.



Introduction to Molecular Evolution

1

Genetic Mutations

In a genetic sequence, new mutations may occur in the germline that are passed down to the next generation. Humans have a very low rate of mutations per generation (50-100). Increased maternal/paternal age may lead to more mutations being passed down.

2

Modeling Evolution

In this lecture, we will think of generations as very short, and model mutations as a continuous process over many generations. We will discuss mathematical tools to understand branch length in trees.

3

Shift from Morphology

In the past, people used morphology to infer phylogenetic trees, grouping organisms according to observable characteristics. A major weakness of this approach is convergent evolution, where organisms that are not closely related have independently evolved similar physical traits.

Sequence Comparison for Phylogenetic Analysis

1

Sequence Alignment

Take some sequences, align all sequence pairs, and define a distance metric between the pairs.

2

Clustering

Try to cluster the sequences based on distance.

3

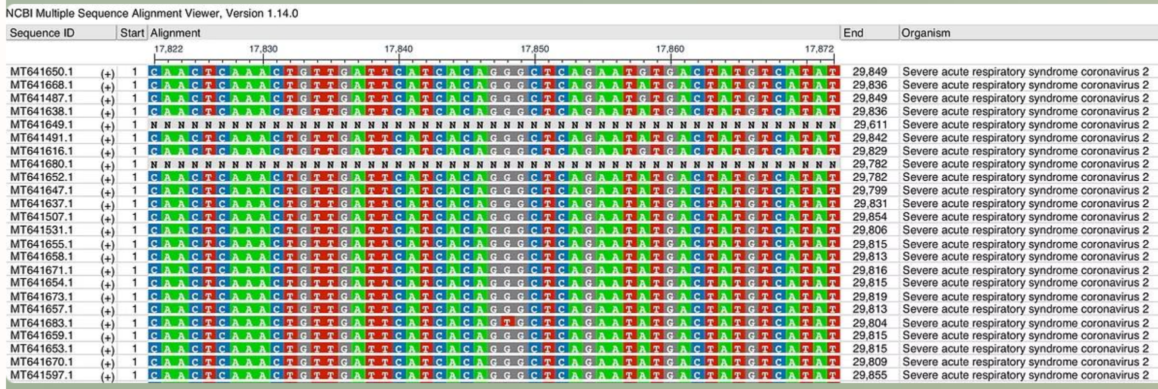
Continuous Process Model

Start with a base at a given position. Consider its change over time to be a continuous process with mutation rate μ.

4

Multiple Substitutions

Note that if the same genomic site undergoes multiple substitutions, some of the changes may mask others, and we won't be able to detect all of them.



NCBI Multiple Sequence Alignment Viewer, Version 1.14.0

Sequence ID		Start	Alignment	End	Organism
MT641650.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATGTGACTATGTCATAT	29,849	Severe acute respiratory syndrome coronavirus 2
MT641668.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,836	Severe acute respiratory syndrome coronavirus 2
MT641487.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATGTGACTATGTCATAT	29,849	Severe acute respiratory syndrome coronavirus 2
MT641638.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,836	Severe acute respiratory syndrome coronavirus 2
MT641649.1	(+)	1	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	29,611	Severe acute respiratory syndrome coronavirus 2
MT641491.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,842	Severe acute respiratory syndrome coronavirus 2
MT641616.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATGTGACTATGTCATAT	29,829	Severe acute respiratory syndrome coronavirus 2
MT641680.1	(+)	1	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	29,782	Severe acute respiratory syndrome coronavirus 2
MT641652.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,782	Severe acute respiratory syndrome coronavirus 2
MT641647.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,799	Severe acute respiratory syndrome coronavirus 2
MT641637.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,831	Severe acute respiratory syndrome coronavirus 2
MT641507.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,854	Severe acute respiratory syndrome coronavirus 2
MT641531.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,806	Severe acute respiratory syndrome coronavirus 2
MT641655.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,815	Severe acute respiratory syndrome coronavirus 2
MT641658.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,813	Severe acute respiratory syndrome coronavirus 2
MT641671.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,816	Severe acute respiratory syndrome coronavirus 2
MT641654.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,815	Severe acute respiratory syndrome coronavirus 2
MT641673.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,819	Severe acute respiratory syndrome coronavirus 2
MT641657.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,813	Severe acute respiratory syndrome coronavirus 2
MT641683.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGTGCTCAGAATATGACTATGTCATAT	29,804	Severe acute respiratory syndrome coronavirus 2
MT641659.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,815	Severe acute respiratory syndrome coronavirus 2
MT641653.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,815	Severe acute respiratory syndrome coronavirus 2
MT641670.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,809	Severe acute respiratory syndrome coronavirus 2
MT641597.1	(+)	1	CAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATAT	29,855	Severe acute respiratory syndrome coronavirus 2

Mathematical Modeling of Molecular Evolution

Mutation Rate

If the base in question is 'A', we can express its mutation rate μA as the sum of the rates at which the 'A' mutates to each of the other bases: $\mu A = \mu AC + \mu AT + \mu AG$

Probability Over Time

The probability of the given base being 'A' at $t + \Delta t$ is $pA(t + \Delta t) = pA(t) - pA(t)\mu A\Delta t + pC(t)\mu CA\Delta t + pG(t)\mu GA\Delta t + pT\,(t)\mu T\,A\Delta t$

Matrix Notation

The expression can also be written in matrix/vector notation: $P(t+\Delta t)=P(t)+Q\,P(t)\Delta t$

Differential Equation

As differential equation: $P'(T) = QP(T)$

Made with Gamma

Evolutionary Models

Jukes-Cantor Model

The simplest model where every base substitution has the same rate. This is not really true in reality.

Kimura Model

Most frequently used across the literature. Separate rate defined for transitions ($K > 1$) and transversions ($= 1$).

Felsenstein Model

Incorporates the fact that in a given genome, the frequencies of each base may be different. This can be modeled by having separate rates of change between the different bases.

The Jukes-Cantor Model in Detail

1

Probability Calculation

Starting with the differential equation $P'(t) = QP(t)$, we can solve for the probability that a base is different between two sequences, or solve for the expected evolutionary time at which they will be different.

2

Closed Form Formulas

$r(t)=1/4(1+3e^{(-\mu t)})$ and $s(t)=1/4(1-e^{(-\mu t)})$, where $r(t)$ is the probability of the base remaining the same and $s(t)$ is the probability of a base substitution.

3

Solving for Evolutionary Time

Let p be the probability that a given base is different between two sequences. We can calculate this by aligning the sequences and counting the number of mismatches relative to matches. Once we have p, we can solve for t: $\mu t = -\ln(1 - 4p/3)$

Building Phylogenetic Trees: UPGMA Method

1

Initialize

Each node is in its own cluster. Define one leaf per sequence, each with height 0.

2

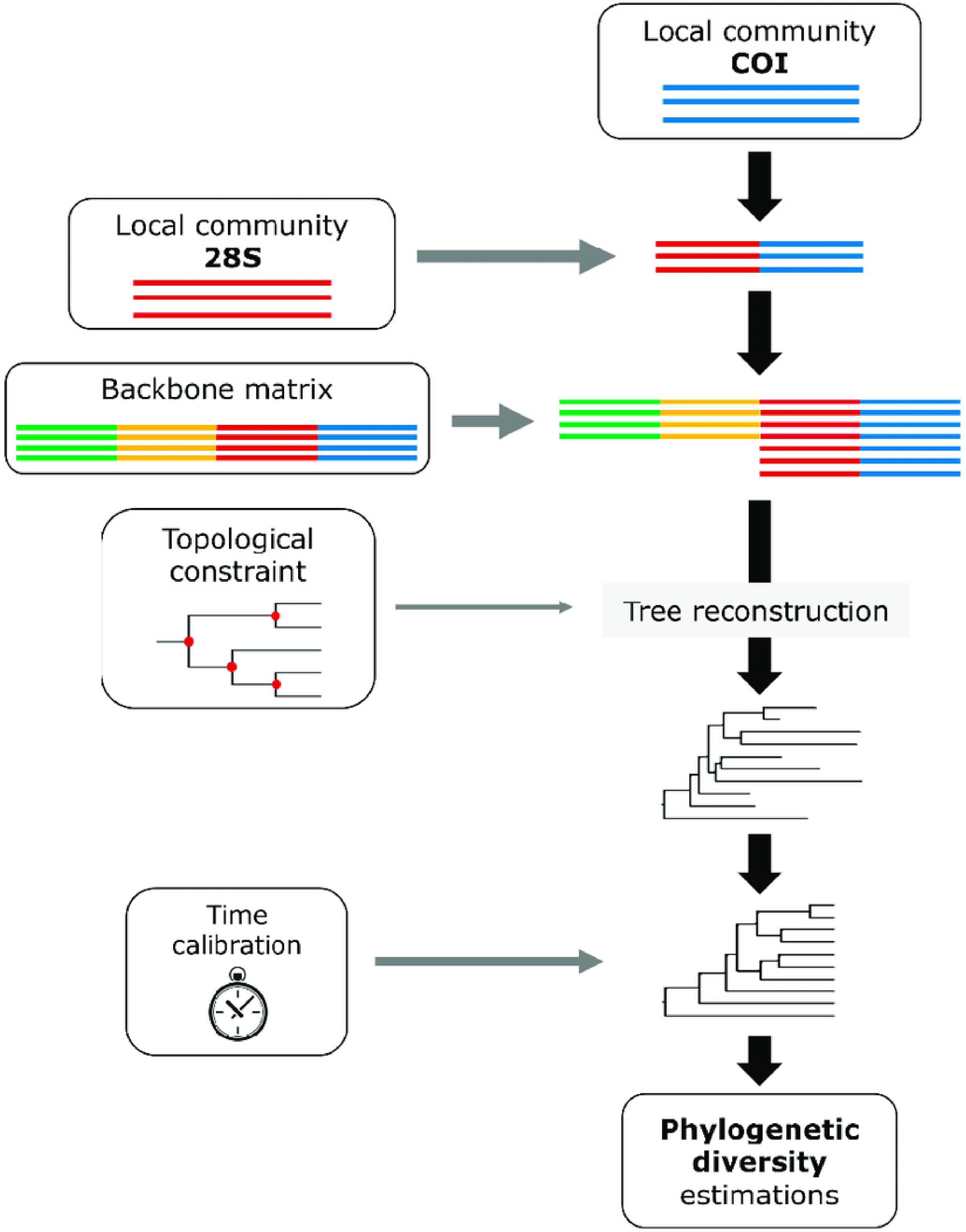
Iterate

Select the two clusters Ci and Cj with the smallest pairwise distance. Let Ck = Ci ∪ Cj. Define a node connecting Ci and Cj, and place it at height dij/2. Delete Ci and Cj.

3

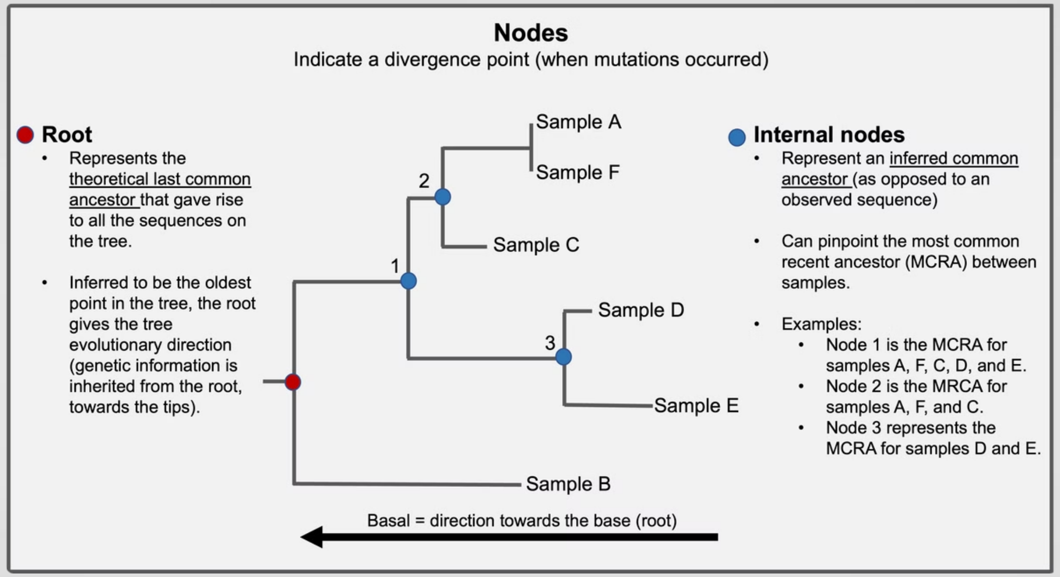
Terminate

When only two clusters i and j remain, merge them by placing the root node at height dij/2.



Additive Distances and Neighbor Joining

Method	Description	Advantage
Additive Distances	Given a tree, distance measurements are additive given that the distance between any two nodes is equal to the sum of the edges connecting them.	Can reconstruct edge lengths accurately
Neighbor Joining	Transforms the original distance matrix to find pairs of leaves that are neighbors in the correct tree.	Guaranteed to produce correct tree if distances are additive, robust to small errors



Nodes

Indicate a divergence point (when mutations occurred)

● **Root**

- Represents the theoretical last common ancestor that gave rise to all the sequences on the tree.
- Inferred to be the oldest point in the tree, the root gives the tree evolutionary direction (genetic information is inherited from the root, towards the tips).

● **Internal nodes**

- Represent an inferred common ancestor (as opposed to an observed sequence)
- Can pinpoint the most common recent ancestor (MCRA) between samples.
- Examples:
  - Node 1 is the MCRA for samples A, F, C, D, and E.
  - Node 2 is the MRCA for samples A, F, and C.
  - Node 3 represents the MCRA for samples D and E.

Basal = direction towards the base (root)

Phylogenetic Analysis: Methods and Applications

- Phylogenies are crucial tools in biology for understanding relationships among species, genes, viral infections, and more.
- With advances in DNA sequencing, phylogenetic analysis has become increasingly powerful and widely used across biological disciplines.
- This review covers the major methods of phylogenetic inference, including parsimony, distance, likelihood and Bayesian approaches, discussing their strengths, weaknesses, and appropriate applications.

Applications of Phylogenetic Analysis

Systematics and Taxonomy

Describing relationships among species on the tree of life

Gene Families

Analyzing relationships between paralogues

Population Histories

Tracing histories and dynamics of populations

Pathogen Evolution

Studying evolutionary and epidemiological dynamics of pathogens

Made with Gamma

Phylogenetic Methods: Distance-Based Approaches

1 **Calculate Pairwise Distances**

Compute distances between all pairs of sequences using a substitution model

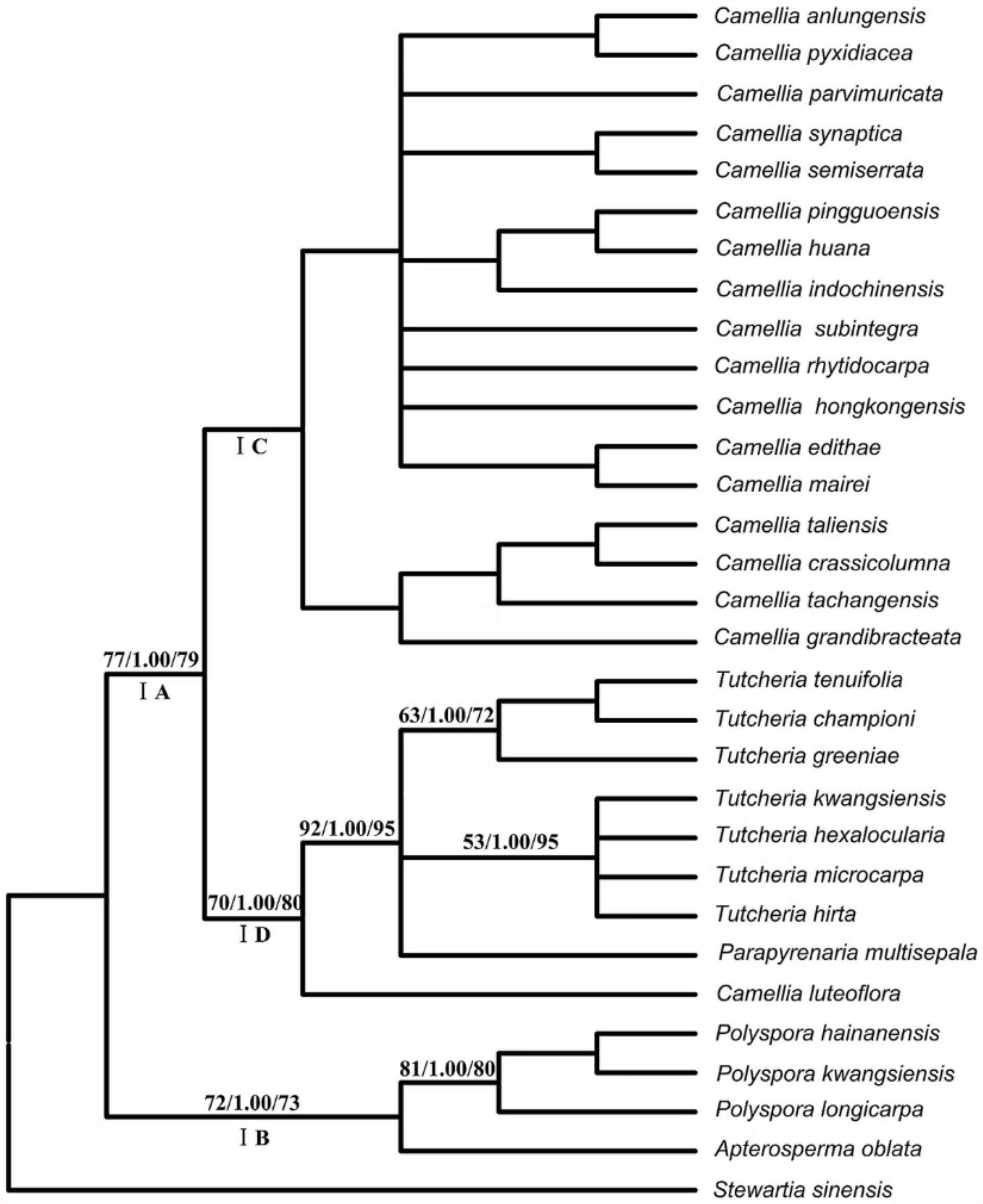
2 **Construct Distance Matrix Matrix**

Organize pairwise distances into a matrix

3 **Apply Clustering Algorithm**

Use algorithms like Neighbor-Joining to construct a tree from the distance matrix

Distance methods like Neighbor-Joining are computationally efficient and useful for large datasets with low sequence divergence. However, they can perform poorly for highly divergent sequences and are sensitive to alignment gaps.



Maximum Parsimony Method

1

Principle

Minimize the number of evolutionary changes required to explain the observed data

2

Tree Score

Sum of minimum number of changes required for each site

3

Informative Sites

Focus on parsimony-informative sites where at least two distinct characters are observed at least twice

Parsimony is simple and computationally efficient, but lacks explicit evolutionary assumptions and can suffer from long-branch attraction, potentially leading to incorrect tree inference in some cases.

Maximum Likelihood Method

1 **Define Likelihood Function**

Probability of observing the data given the tree and model parameters

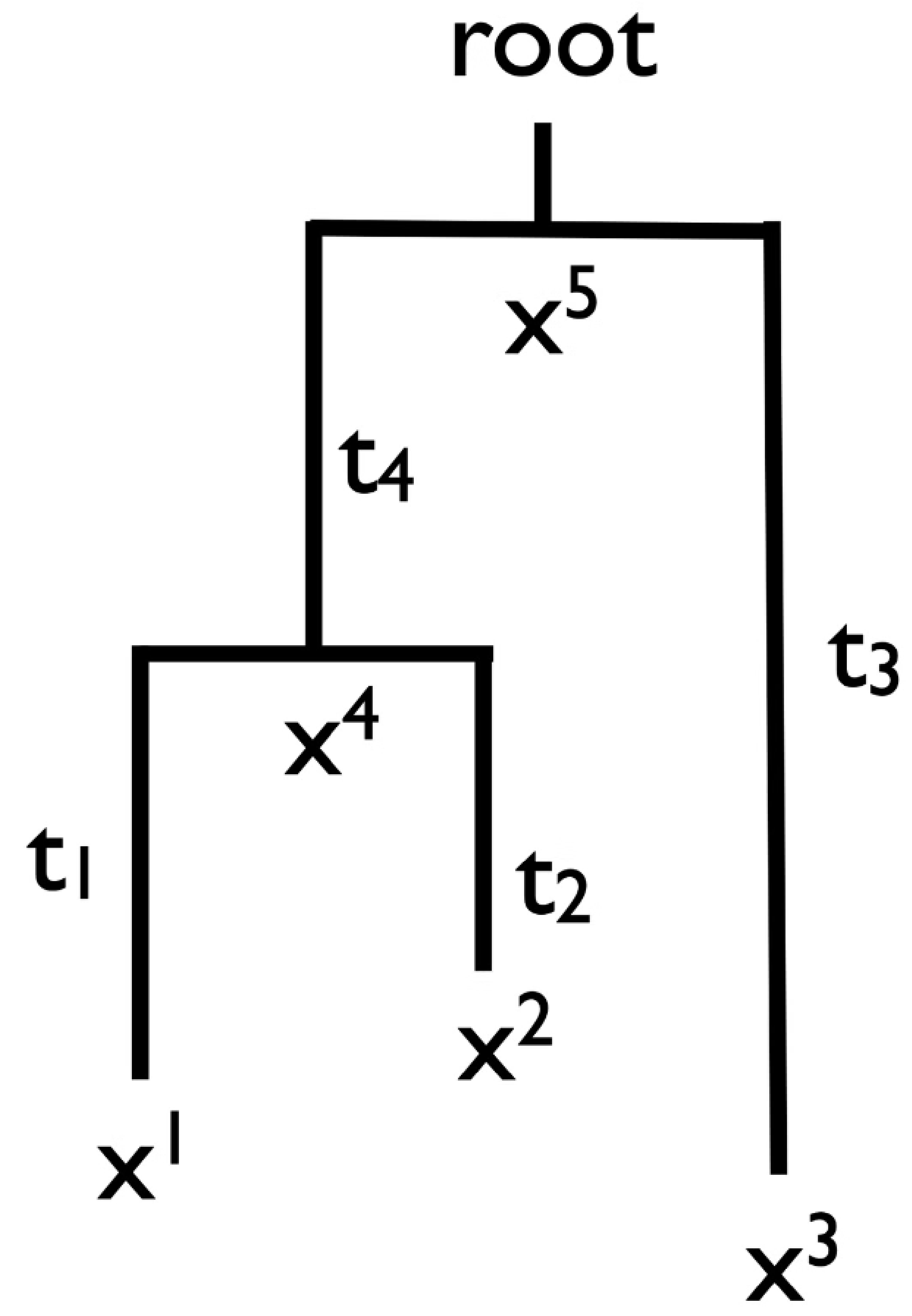
2 **Optimize Branch Lengths**

For each candidate tree, find optimal branch lengths

3 **Search Tree Space**

Find the tree that maximizes the likelihood function

Maximum likelihood allows for explicit modeling of sequence evolution and can accommodate sophisticated evolutionary models. It is widely used for inferring deep phylogenies but is computationally intensive.



Bayesian Inference in Phylogenetics

Posterior Probability

$P(T,\theta|D) \propto P(T,\theta)P(D|T,\theta)$

The posterior probability of the tree and parameters given the data is proportional to the prior probability times the likelihood

MCMC Sampling

Use Markov chain Monte Carlo to sample trees and parameters from the posterior distribution

Advantages

Provides measures of uncertainty, accommodates complex models, allows incorporation of prior information

Phylogenomic Analysis of Large Datasets

1

Supertree Approach

Analyze genes separately, then combine into a supertree

2

Supermatrix Approach

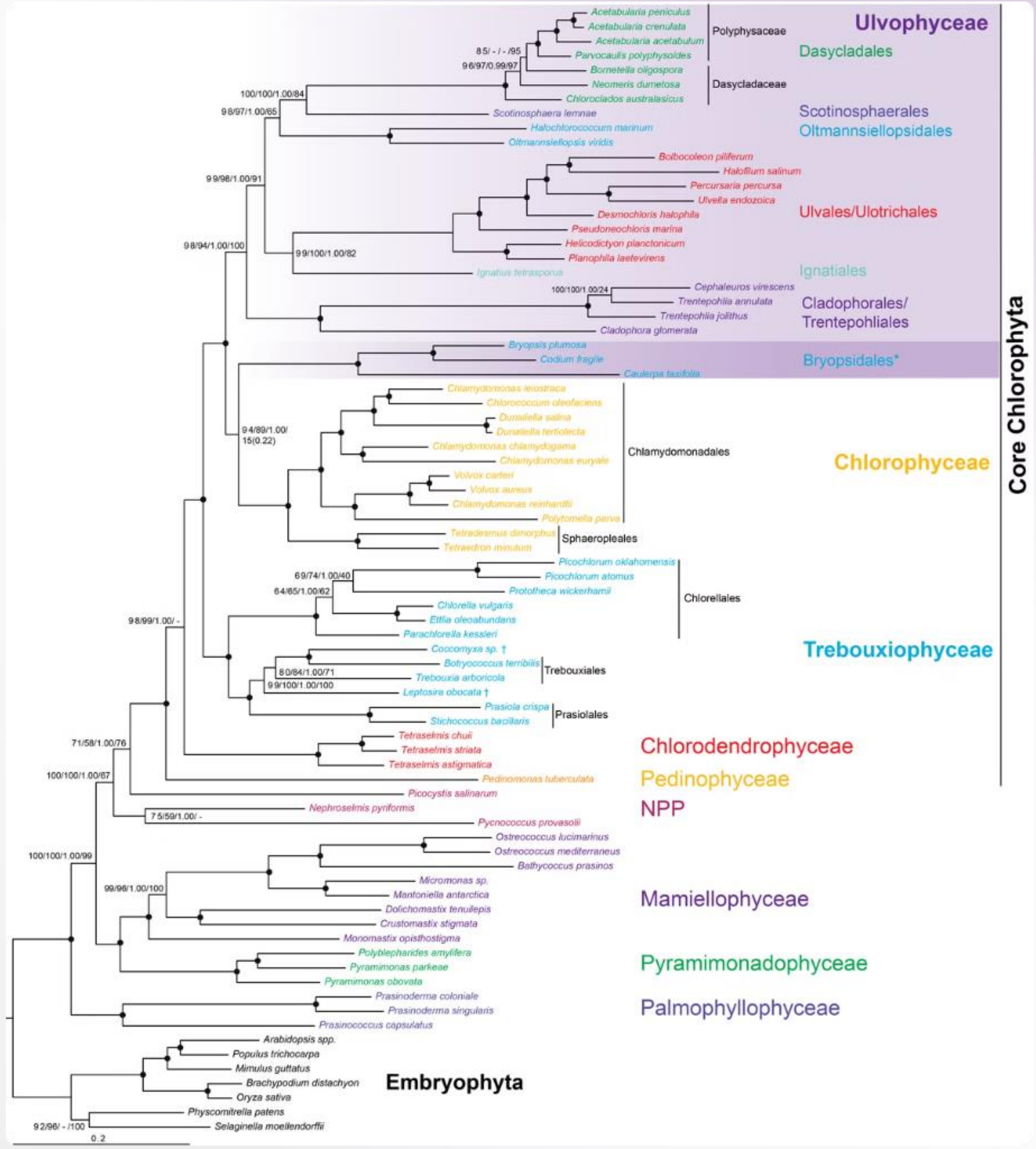
Concatenate sequences into a large matrix for combined analysis

3

Data Partitioning

Group genes or sites with similar evolutionary characteristics

Analysis of genome-scale datasets poses challenges in handling missing data, addressing systematic errors, and developing appropriate partitioning strategies. Combined analysis using likelihood methods to accommodate among-gene heterogeneity is often ideal.



Future Directions in Phylogenetics

Multiple Sequence Alignment

Developing statistical methods for joint inference of alignment and phylogeny

Molecular Clock Dating

Improving models of rate evolution and incorporation of fossil calibrations

Statistical Phylogeography

Integrating population genetics and phylogenetics to address evolutionary questions

Population Genomics

Analyzing whole genome data to infer species histories and demography