VARIATION IN SPECIES

Dilvan Moreira (based on Prof. André Carvalho presentation)

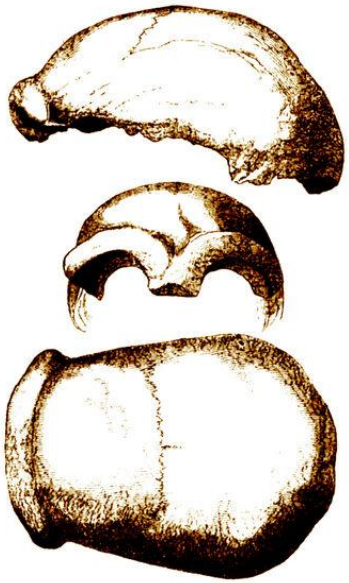
Leitura

- Introduction to Computational Genomics: A Case Studies Approach
  - Chapter 5

Introduction

- Introduction
- Origin of Human Being
- Variation in DNA sequences
- Mitochondrial DNA
- Variation among species
- Estimation of genetic distance
- Case Study: We descended from Neanderthals?

Introduction

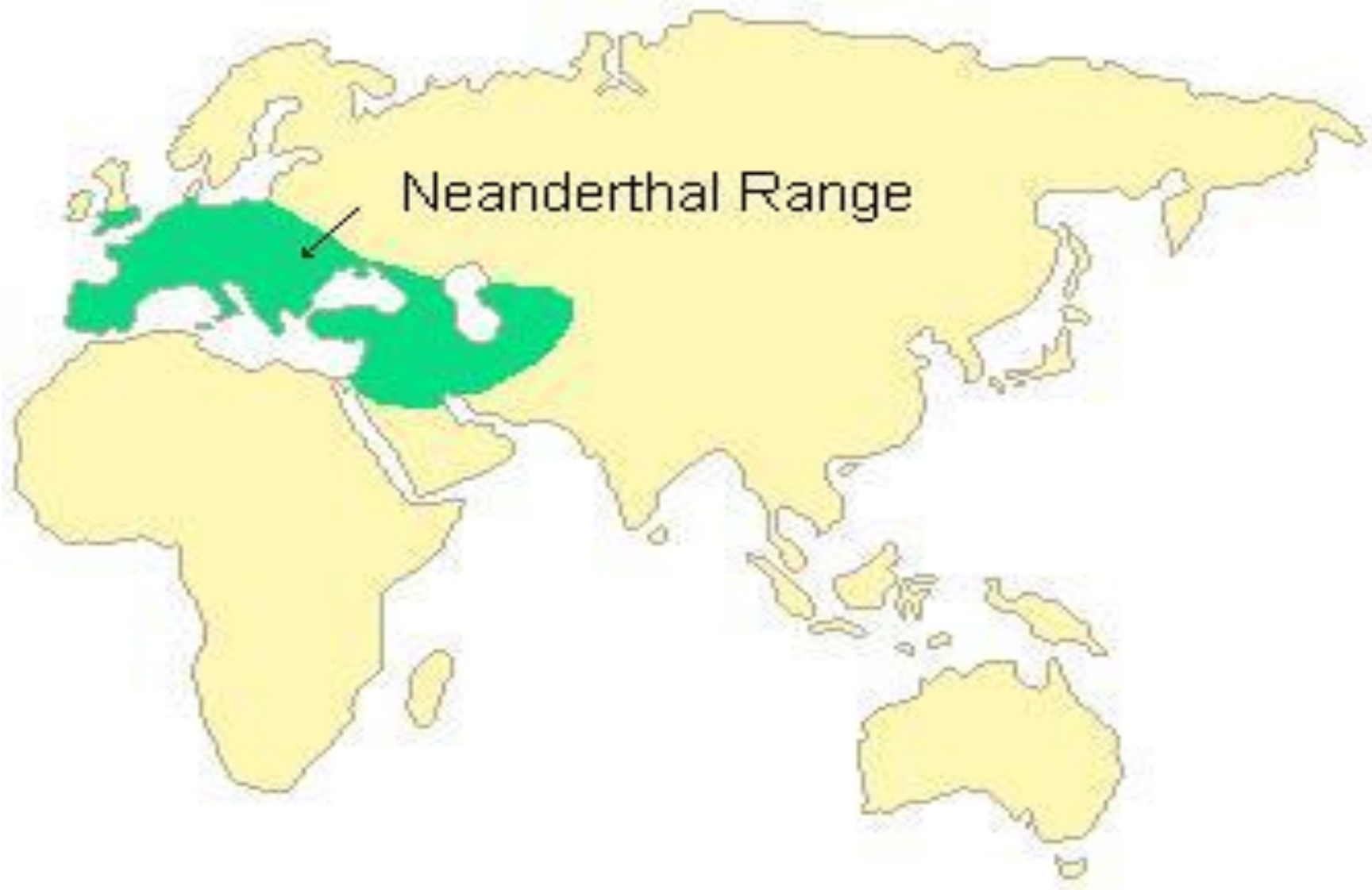


Neander Thal, Germany, 1856

Bear?

Deficient?

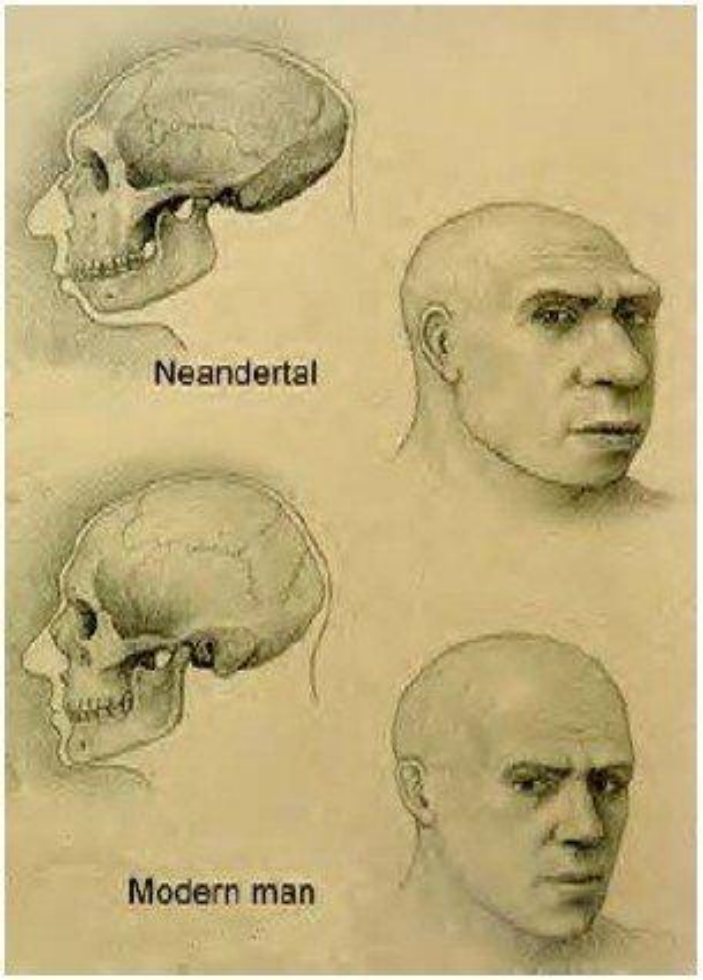
Introduction



Introduction

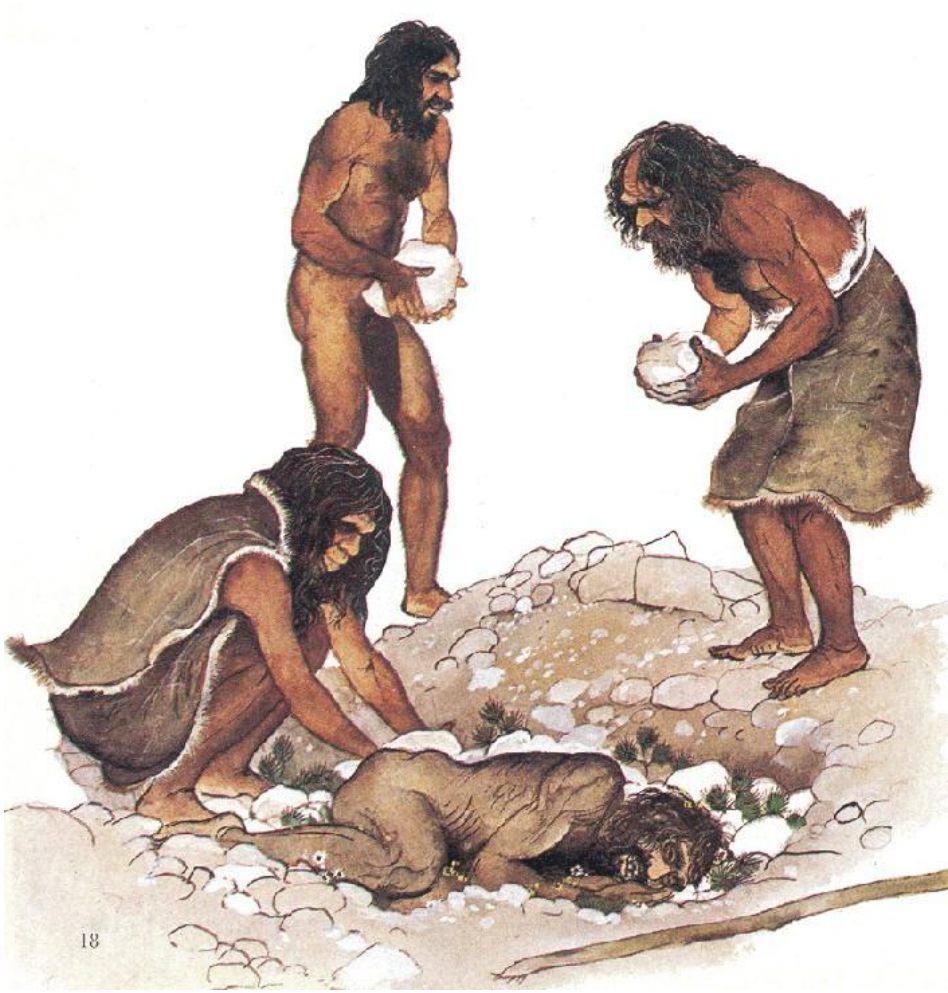
Homo Neandertalensis

Homo sapiens sapeins



Introduction

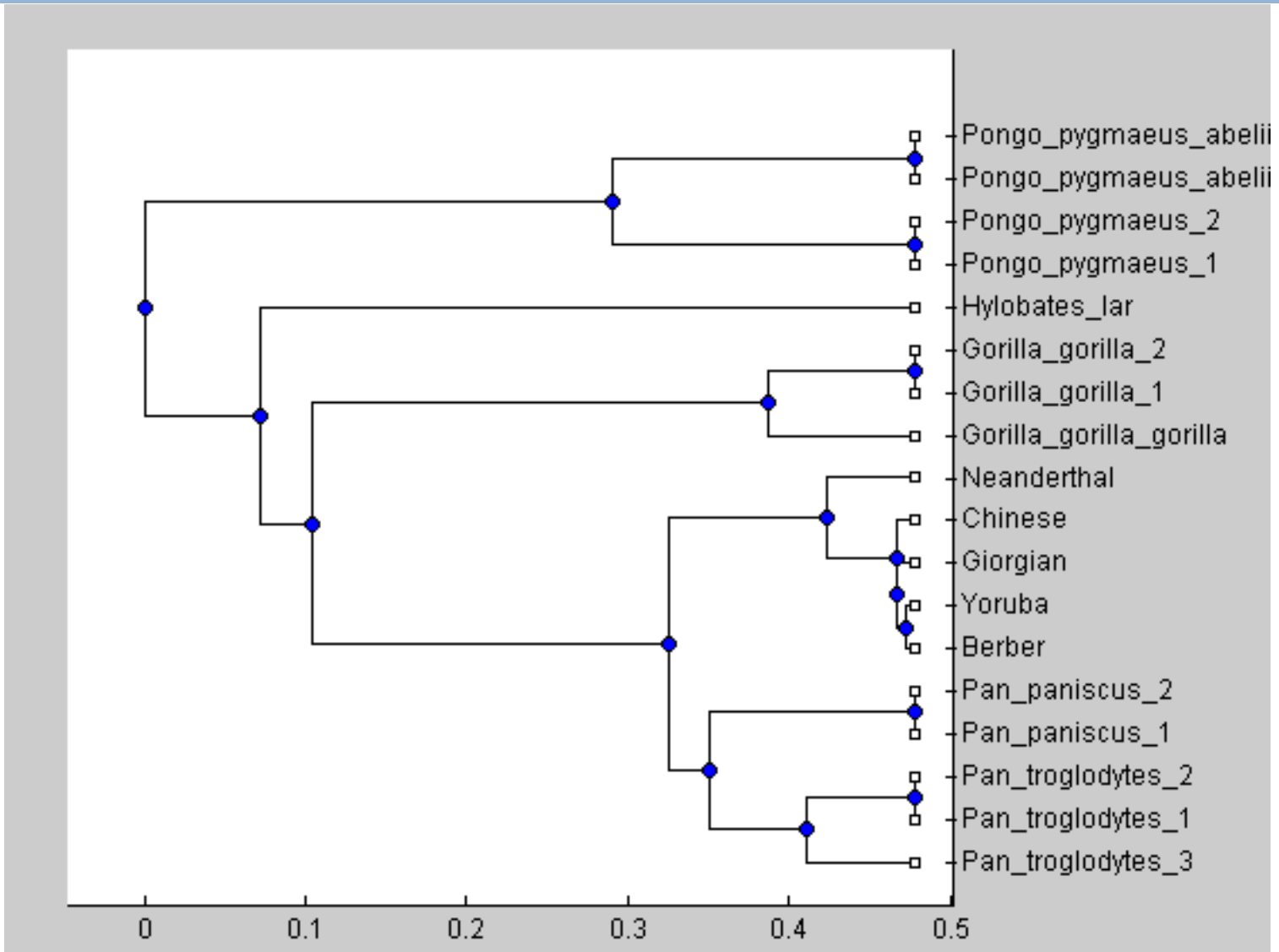
- Skeletons and tools were found
- Time estimated by radio-carbon technical



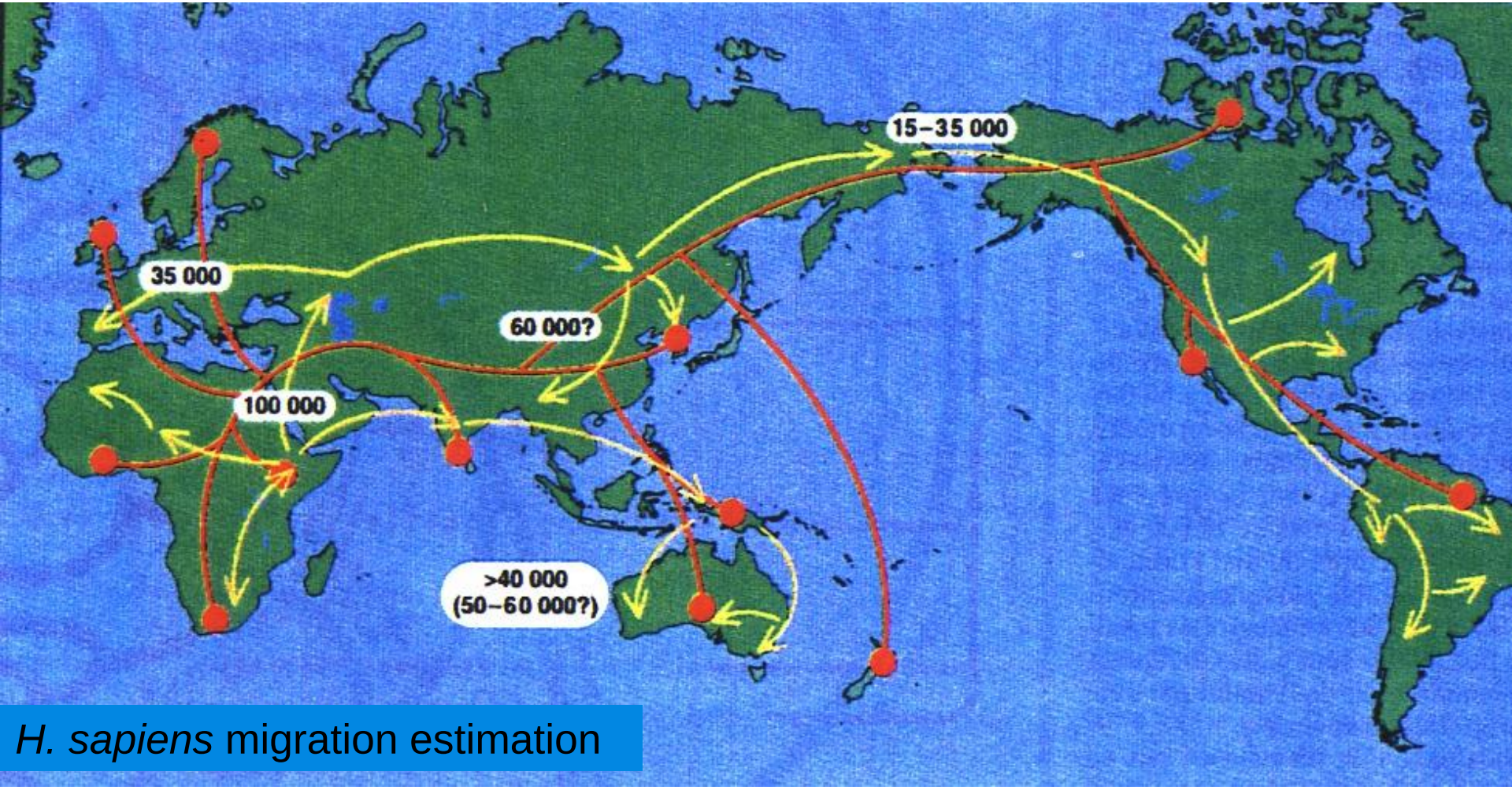
Historical Evolution

- 5 MYA, chimpanzee and bonobo (monkeys), Africa
- 2 MYA, *H. erectus* and *H. habilis*, Africa
- 250 KYA , H. neanderthalensis, Europe and Western Asia
- 130 KYA, H. sapiens, África
  - 60 KYA , Asia and Australia
  - 40 KYA , Europe
  - 30 KYA , Americas
  - They were not the only inhabitants of our genre at that time:
    - Up to 27 KYA , *H. erectus*, Indonesia
    - Up to 28 KYA , *H. neanderthalensis*, Europe and western Asia

Historical Evolution



Padrões de Migração



H. sapiens migration estimation

Variação

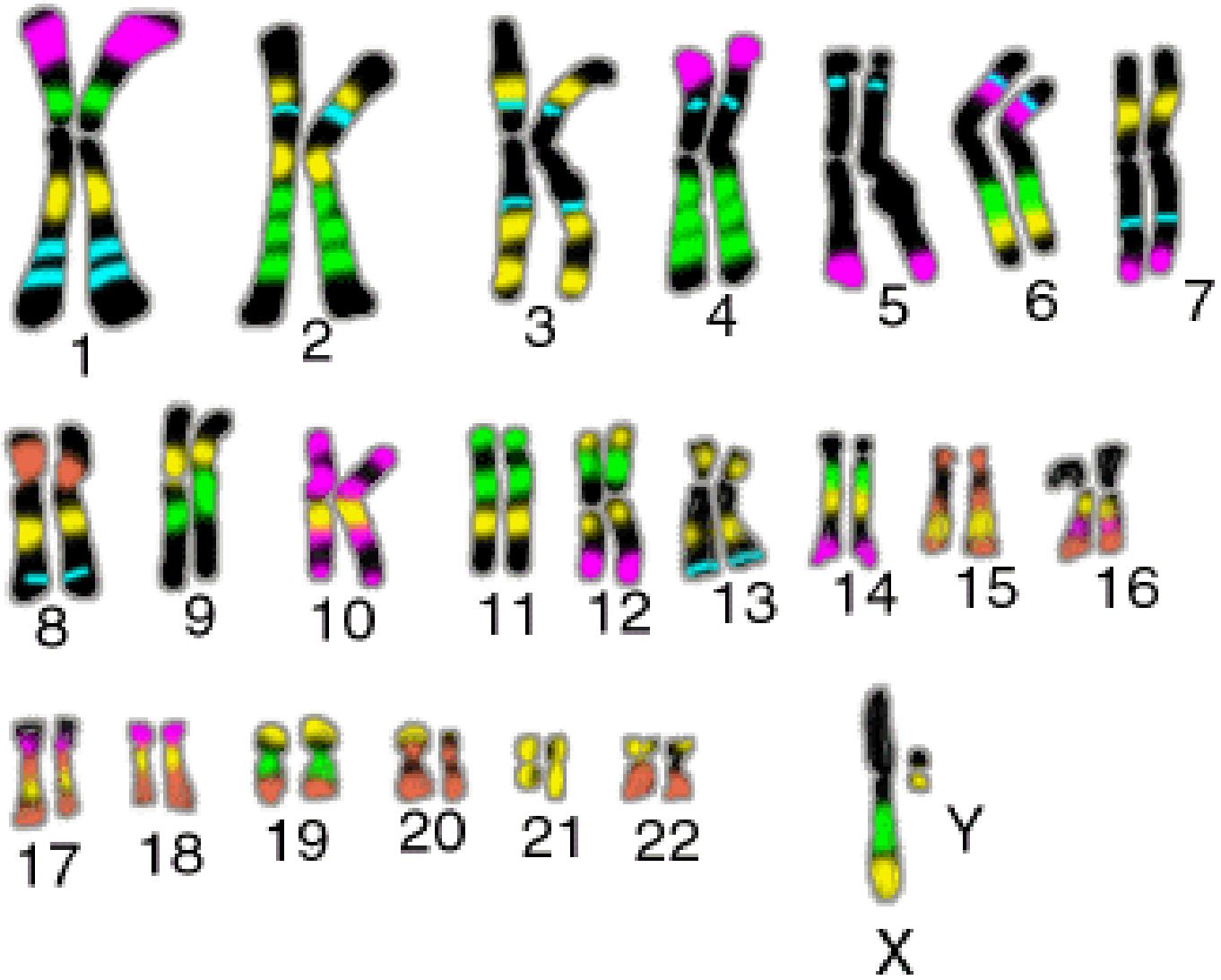
- DNA can be used to explain Origin of Humans
  - Each individual has a different genome sequence
    - Different species and the same species
    - Including children of the same parents

Variação

- Variation is due to:
  - Mutations
    - Errors made by the cellular machinery that are incorporated into the genome
  - Recombination
    - Exchange of genome segments of diploid organisms (two copies of each chromosome)
    - Consequence of the process of reproduction

Variação

Diploid organism (humans)



Mutações

- Mutations occur for several reasons
  - Usually because of an error during genome replication
    - How many mistakes can be made during the process of re-entering a text with 2500 letters?
    - Human cells do this several times to sequences of 3.5 billion letters
      - Our correction mechanism is very good
    - Estimate of the rate of human mutation: an error every 200M to 1B replicated bases

Mutações

- Mutation rates
  - They can be augmented by external factors
    - UV ray, viruses, drinks, cigarettes, drugs
  - They differ for different organisms and types of genome
    - In animals, the mitochondria mutation rate for DNA is an order of magnitude greater than the nuclear
    - In plants the opposite occurs

Mutações

- Some mutations are inherited
  - It allows to study history of individuals and species
  - Shared mutations may indicate shared ancestry
- *Germline* mutations
  - Ocorrem em células dos testículos / ovários
  - They occur in testicles/ovaries cells
  - They can be transmitted to future generations

Mutações

- Every mutation occurs initially in a single individual
- Mutation can be:
  - Neutral: No effect
  - Harmful: makes a biological function worse than it was
  - Advantageous: improve some biological function
- If mutation is not passed to a child, it is lost

Polimorfism

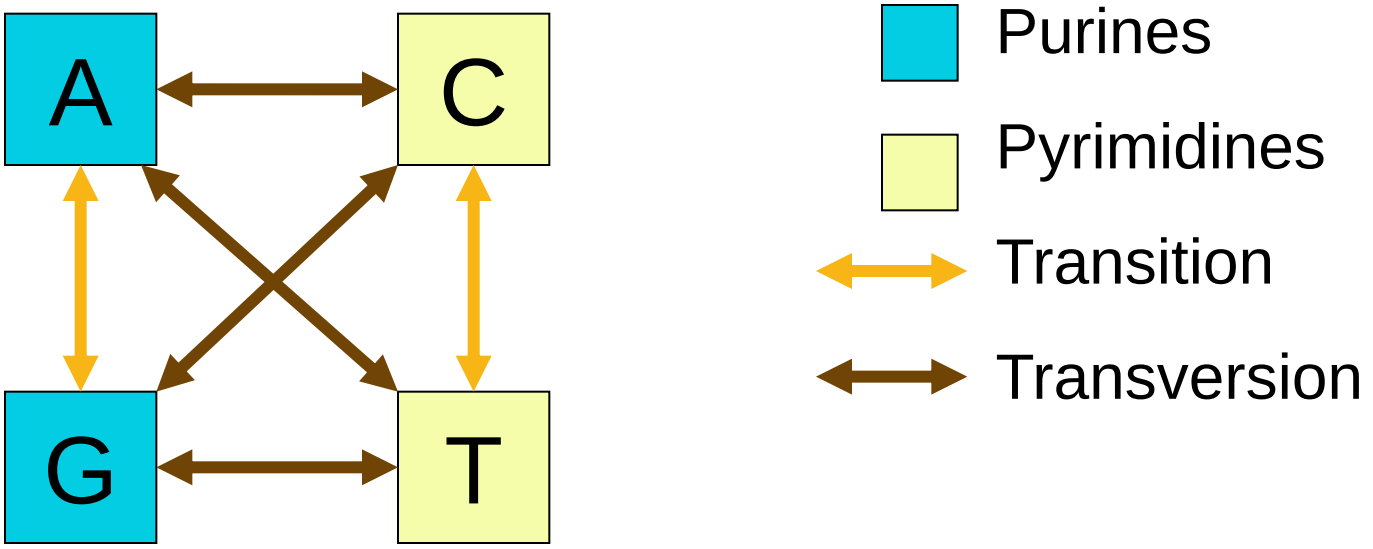
- Difference between individuals to a given position in the genome
  - Several possibilities for a nucleotide = allele
  - Polymorphism of a single nucleotide = punctual mutation
    - Single Nucleotide Polymorphism – SNP ("snip")
    - Ex.: AAATAAA
      AAACAAA
    - Humans: SNP = 1/1500 bases = 0.067%

Polimorfism

- SNP map
  - List all positions where a polymorphism was documented with a certain frequency
  - Different types of nucleotide polymorphism
    - Transitions
    - Transversions

```
G T C C T T C A T A A T C A T C A C G G G
G A C C T T C A T A A C C A T C A C G G G
A A C C T T C A T A A C C A T C T C C G G
x o - - - - - - - - - x - - - - o - o - -
```

Polimorfism



Transition is more frequent

Polimorfism

- Small repeated sequences
  - Second main source of variation among human genomes
  - STR = *Short Tandem Repeats* (microsatellites)
  - Ex.: CACACACACACACACACA …
- Mutation rate for STRs is much higher than for SNPs
  - What makes them very useful for identification

Matrizes de substituição

- Not all mutations of nucleotides occur at the same frequency
  - Because:
    - Chemical nature of the bases
    - Harmful consequences of change
  - The same applies to amino acids
    - Codons for two amino groups differ by more than one nucleotide
    - Some amino acids are easier to be exchanged than others
      - They share biochemical features

Matrizes de substituição

- Matrices may represent differences in substitutions
- Differences between nucleotide substitutions
  - Separate transitions from transversions
  - Default parameters of BLAST do not make that difference

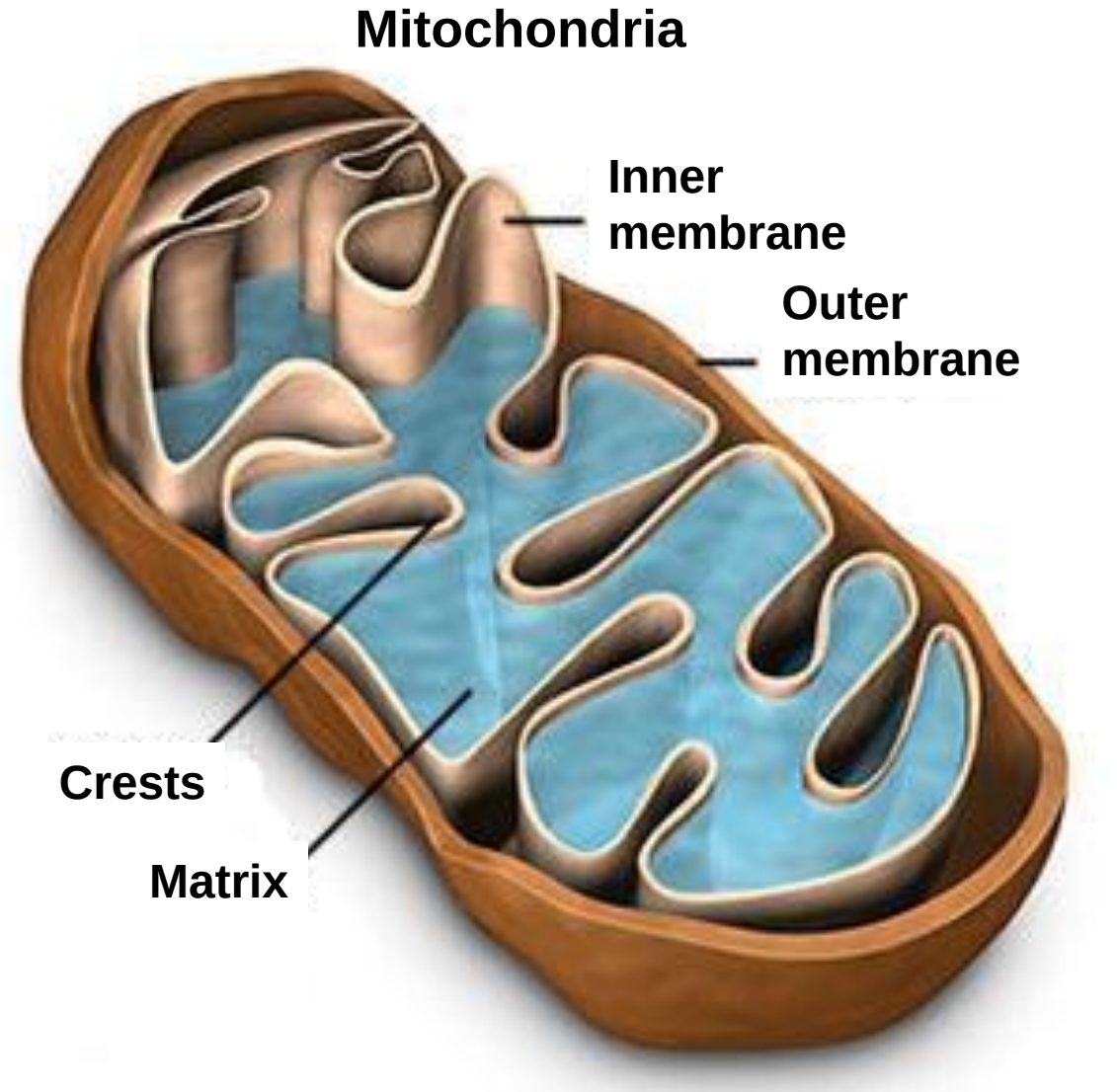
Matrizes de substituição

- Differences in amino acid substitutions
  - Consider all the mutation probabilities
    - Amino acid substitution matrix
  - Matrices families commonly used:
    - PAM (*Percent or Point Accepted Mutation*)
    - BLOSUM (*BLOcks Substitution Matrix*)

Mitochondrial DNA

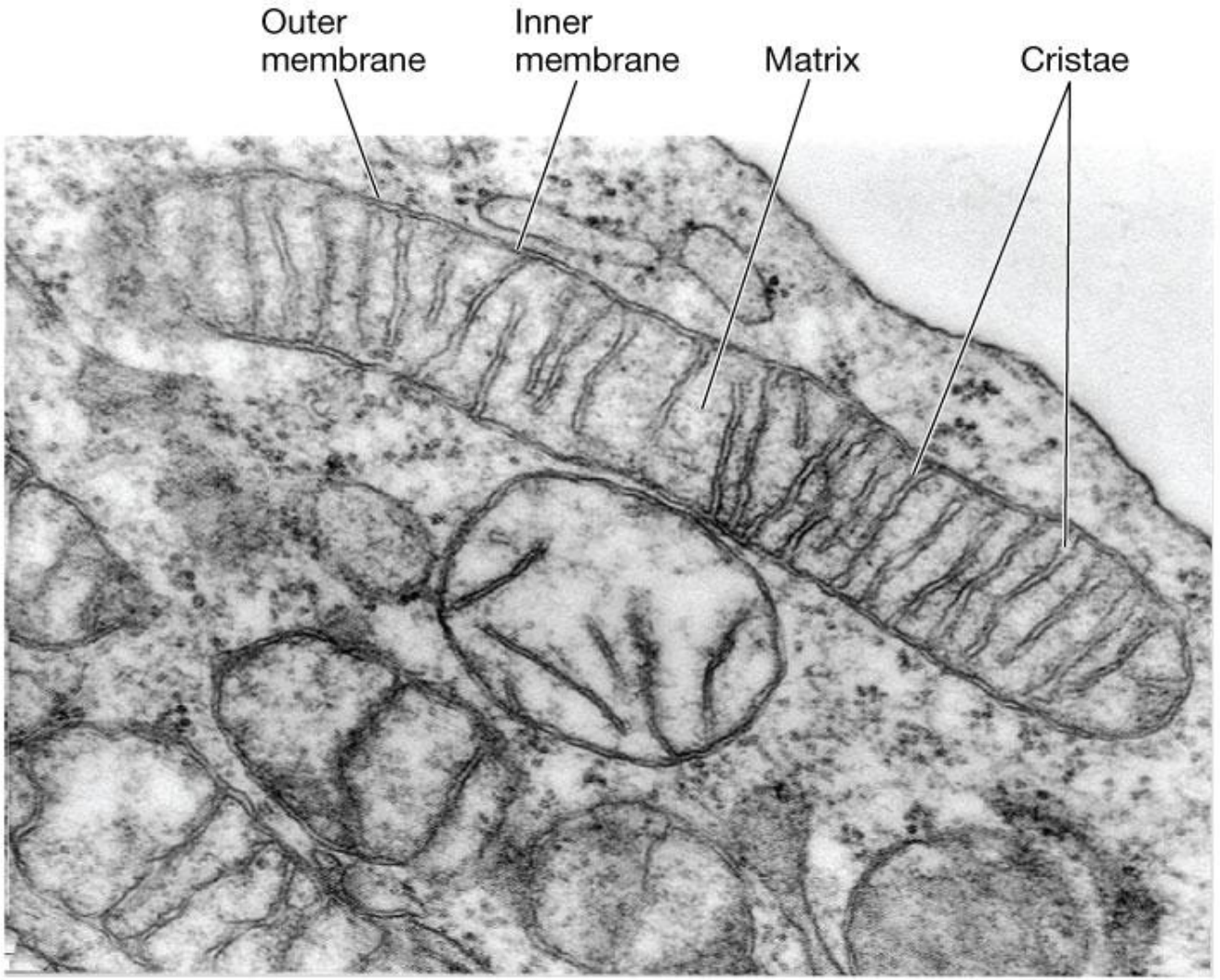
- Ideal for studying human evolution
  - Compact
    - Circular chromosome with 16,569 bases and 37 genes in humans
  - It has high mutation rate
  - Inherited by mother only
    - Each individual will have only one version
  - Each cell contains multiple copies of mtDNA
    - Easier to isolate and sequence

Mitochondrial DNA

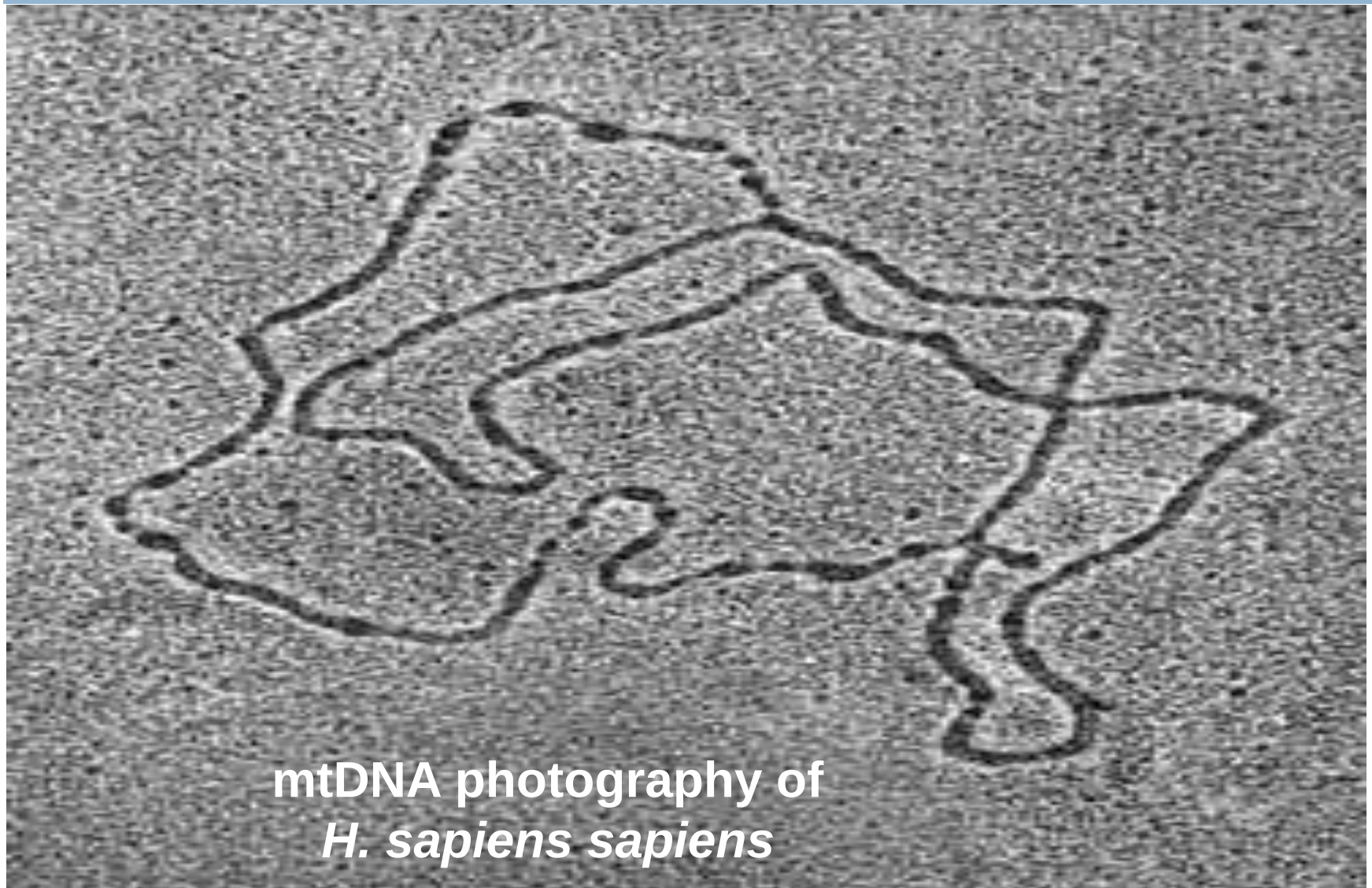


Mitochondrial DNA

H. sapiens sapiens
Mitochondria

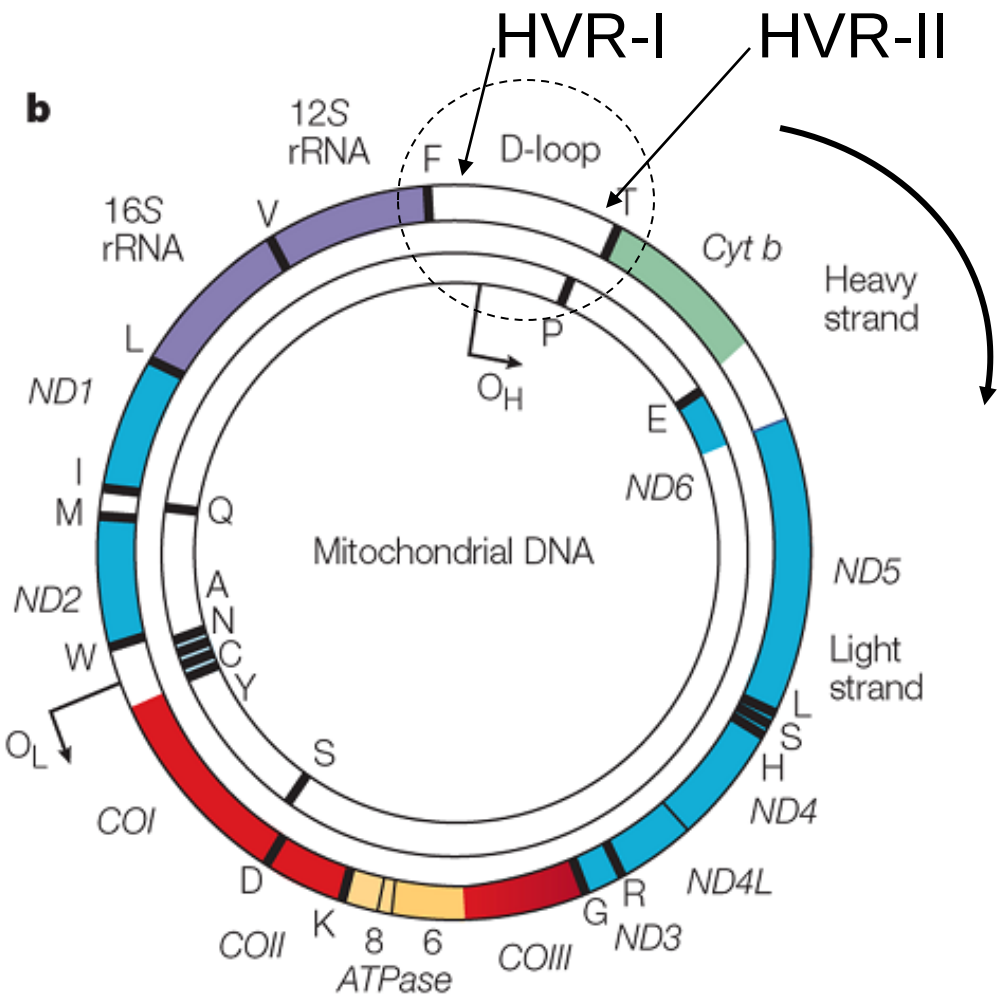


Mitochondrial DNA



Mitochondrial DNA

mitochondrial human DNA



Mitochondrial DNA

- Control region or D-loop region
  - Single noncoding portion in the mitochondrial genome (does not contain genes)
  - contains:
    - Origin (beginning) of replication
    - Mitochondrial promoter
    - Two hyper variable regions
      - HVR-I
      - HVR-II

Mitochondrial DNA

- Hyper variable regions
  - Most of the D-loop (400 to 500 bp each)
  - Situated at the ends of the replication start site
  - Sequences show high variability in humans
    - Ideal for studying the relationship among individuals

Variation among specie

- Can explain:
  - Relationship among different species
    - Closest species have on average more similar DNA
  - How did the evolution occurred in millions of years
    - Relationship between genetic variation and phylogenetic tree
- It is necessary to measure the distance between sequences
  - How many nucleotide substitutions separate any two DNA sequences
    - Replacement rate

Variation among specie

- Replacement rate
  - Mutations arise from a single individual
    - They may be lost
      - If a mutation carrier individual have no children
    - They can increase in frequency and become permanent or fixed in the species
      - Every individual of this specie will have this new allele at a specific position
  - Rate at which species accumulate such fixed mutations

Variation among specie

- Replacement rate
  - Based on homologous sequences from different species
  - allows estimate
  - When there was divergence between the two species
    - Biological function of genomic sequences
    - Relationships between species

Variation among specie

- Neutral mutations
  - Do not affect reproductive capacity of the individual
  - Lead to a relationship between mutation rate and substitution rate:
  - The replacement rate (ρ) = the number of new mutations X probability that one of them is perpetuated

Variation among specie

- The replacement rate (ρ) = the number of new mutations X probability that one of them is perpetuated
  - Prob. of new mutation perpetuate = 1 / (2N)
  - Number of new mutations = 2Nμ
  - ρ = 2Nμ*1/(2N) = μ
  - ρ = K/(2T)

μ - mutation rate
N - number of diploid individuals
T - divergence time
K - number of substitutions (**genetic distance**)

Variation among specie

- Replacement rate is independent of population size
  - Equal to neutral mutation rate
    - larger populations
      - Have more mutations, but with small chance of becoming fixed
    - smaller populations
      - They have fewer mutations, but most likely to become fixed
  - When the changes are not neutral, rates can be quite different

Estimation of genetic distance

- Genetic distance between two homologous sequences
  - Number of substitutions that have accumulated between them since they diverged from a common ancestor
    - Counting of different positions generates a pessimistic estimate when there are multiple substitutions

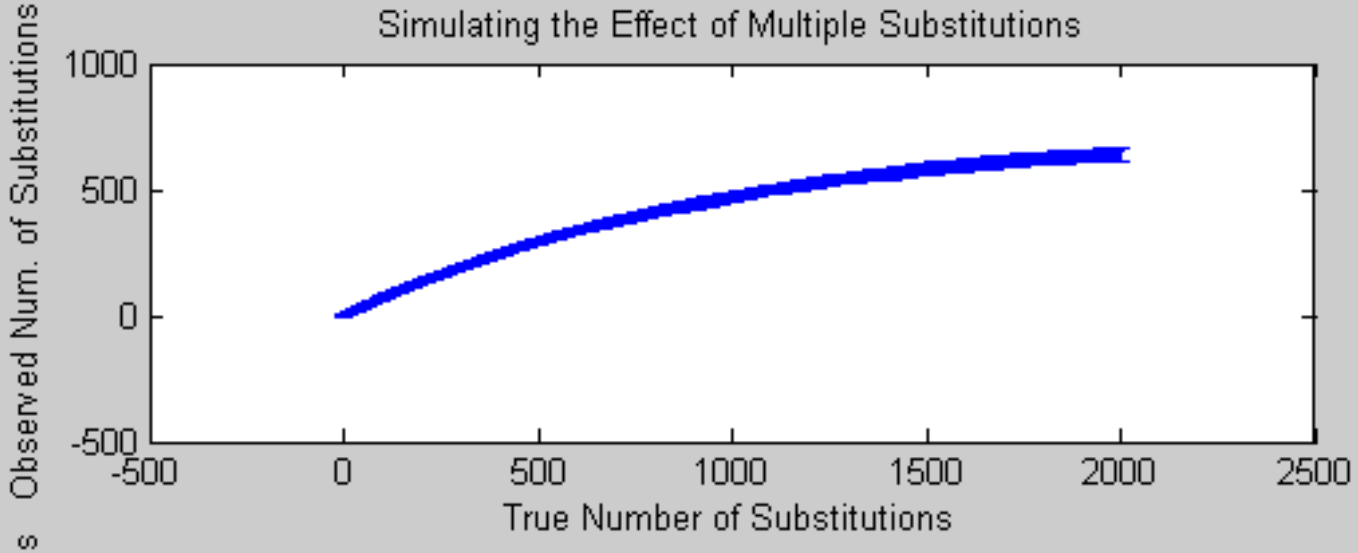
Estimation of genetic distance

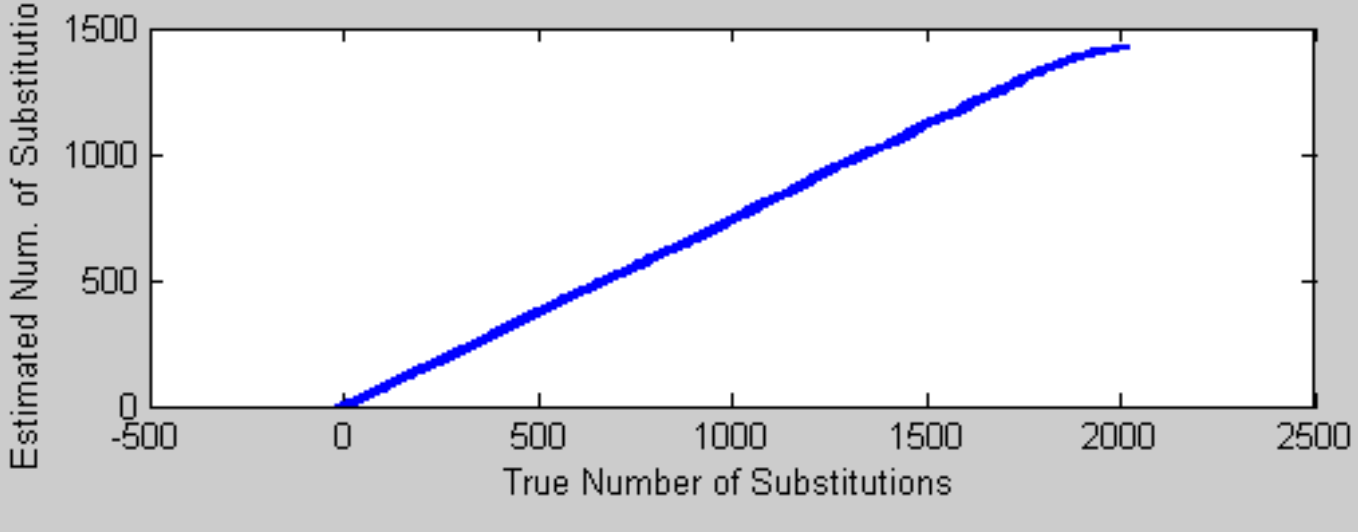

```
GACTGATCCACCTCTGATCCTTTGGAACTGATCGT
TTCTGATCCACCTCTGATCCTTTGGAACTGATCGT
TTCTGATCCACCTCTGATCCATCGGAACTGATCGT
GTCTGATCCACCTCTGATCCATTGGAACTGATCGT
```

Observed subistitutions: 2 (= *d*)

Real subistitutions: 4 (= *K*)

Estimation of genetic distance





Estimation of genetic distance

- Substitutions extreme: saturation
  - When there is on average one substitution for each position of the sequence
  - Limit of divergence between two homologous sequences
    - Only a quarter of their positions matches
      - Easy to get for two random sequences

Estimation of genetic distance

- Problem:
  - Multiple (several) replacements and return mutations may hide real genetic distance
- Solution:
  - Estimate distance between two sequences using probabilistic model

Estimation of genetic distance

- There are several models to estimate replacement rate (distance)
  - Simpler: each replacement has the same probability
    - Jukes-Cantor Model
    - Kimura Model
  - Evolution can be viewed as a Markov process
    - Sequence in time (generation) t depends on the sequence in time t-1

Jukes-Cantor Model

- Probabilistic model based on Markov processes
  - It assumes that all replacements are equally likely

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}d\right)$$

d - number of observed differences

Kimura Model

- Probabilistic model based on Markov processes
  - Uses different probabilities for transitions and transversions

$$K = -\frac{1}{2}\ln(1-2P-Q) - \frac{1}{4}\ln(1-2Q)$$

P – Number of transitions
Q – Number of transversions
d = P + Q

Other Models

- There are more complex models
  - Distinguishing different types of transitions and transversions
  - With different parameters for each pairwise probability of substitution
  - Limitation of existing models:
    - They assume that the substitution probabilities are symmetric
      - But they are not

Case Study: Neanderthals

- Discovery of skeletons of Neanderthals
  - Generated several questions about our relationship with this species
    - Origin of human and primate
  - Many answers can be given by Mitochondrial DNA study
    - Using:
      - Complete mitochondrial genome or
      - Regiões hipervariáveis

Case Study: Neanderthals

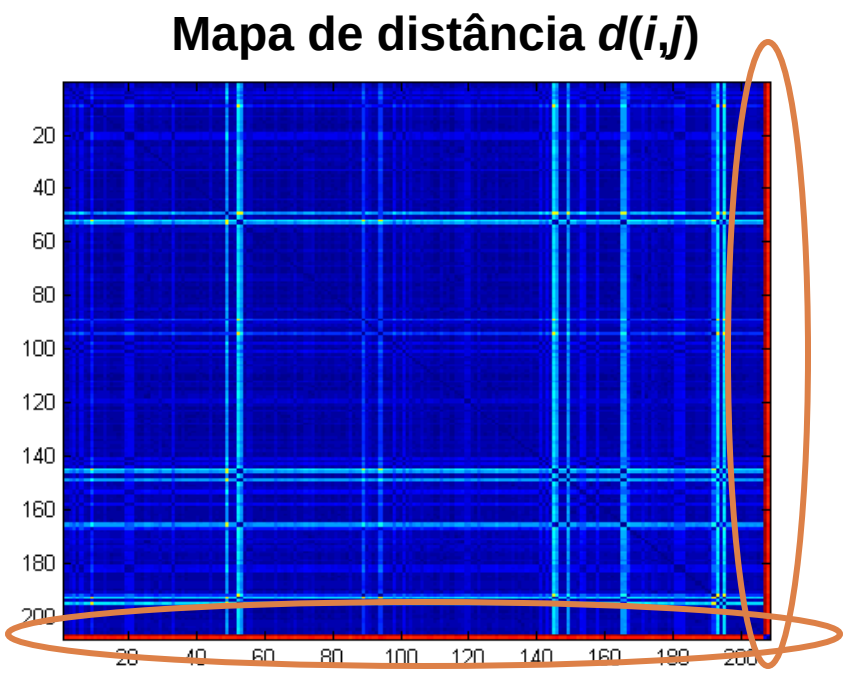
- To answer the questions, they were analyzed:
  - 206 mtDNA of modern humans
  - Parts of 2 mtDNA of Neanderthals
    - Including unique individual found in German cave
    - All available in GenBank
  - Corresponding regions were used
  - FSet with 208 homologous with 800 bp each

Case Study: Neanderthals

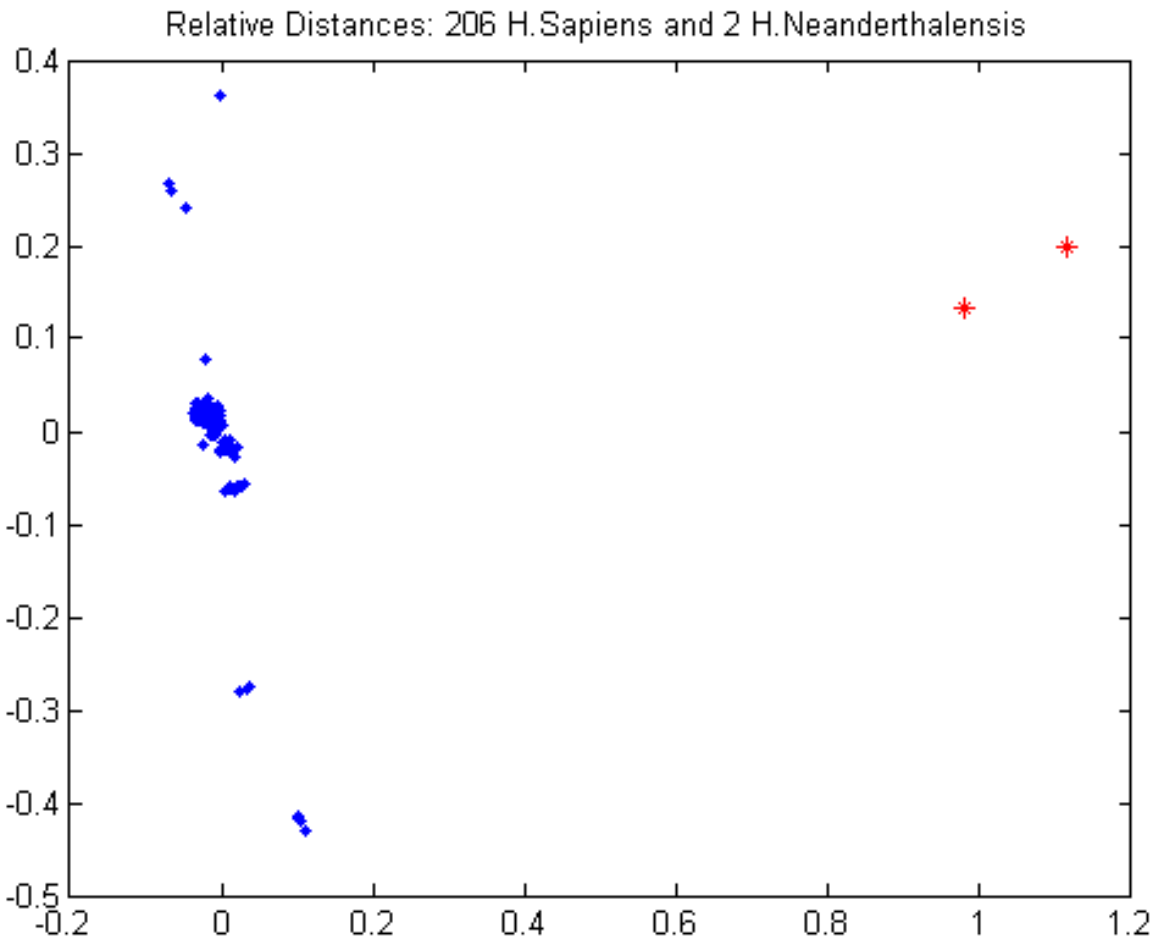
- Genetic distances between pairs of sequences were calculated
  - For all the 208 sequences
  - Corrected by Jukes-Cantor formula
- Medium distances:
  - Between any 2 H. sapiens = 0.025 (25 of each 1000 bases are different)
  - Between Neanderthal and modern human = 0.140

Case Study: Neanderthals

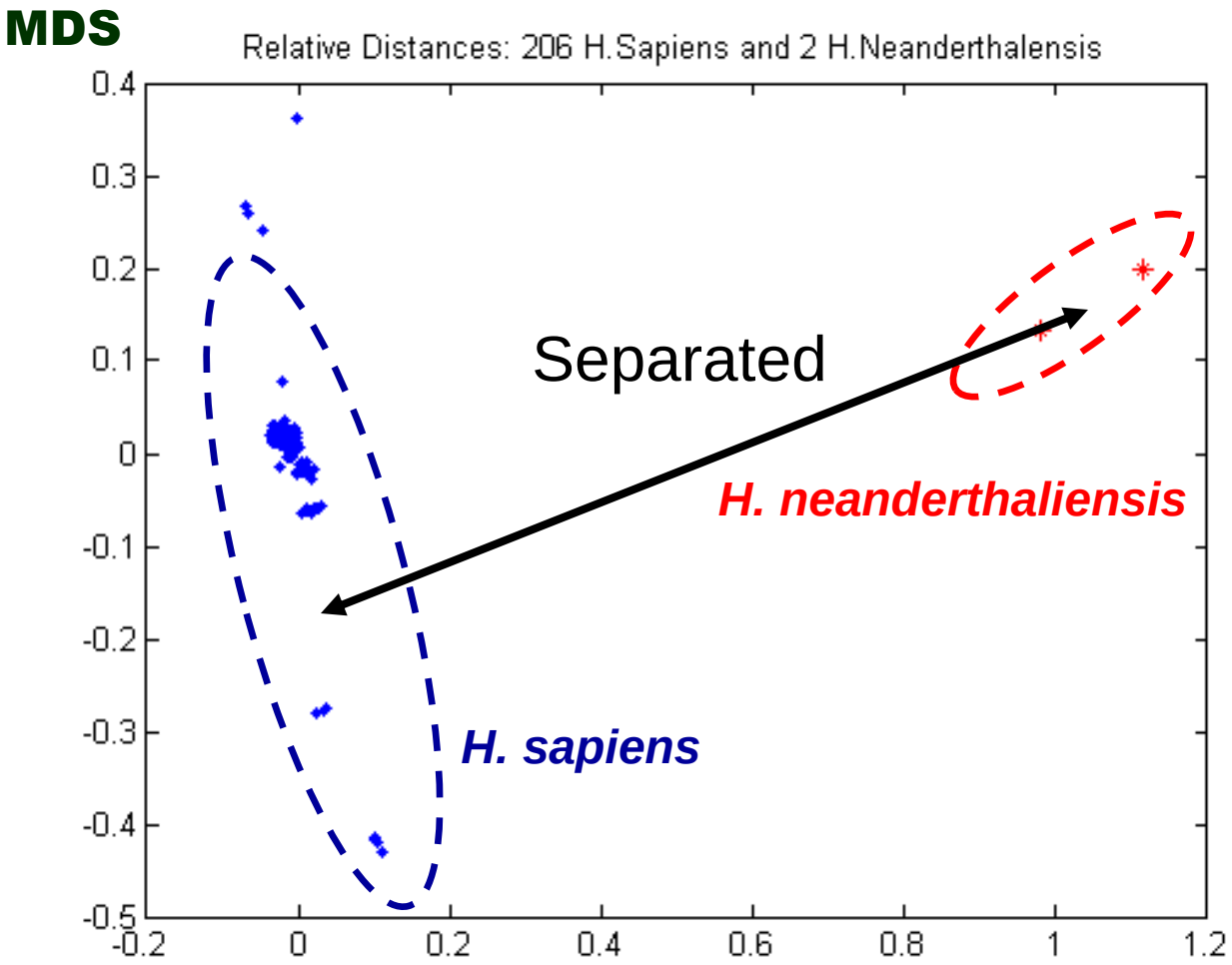
- Multi -dimensional scaling (MDS)
- Translates distance table in an n-dimensional map



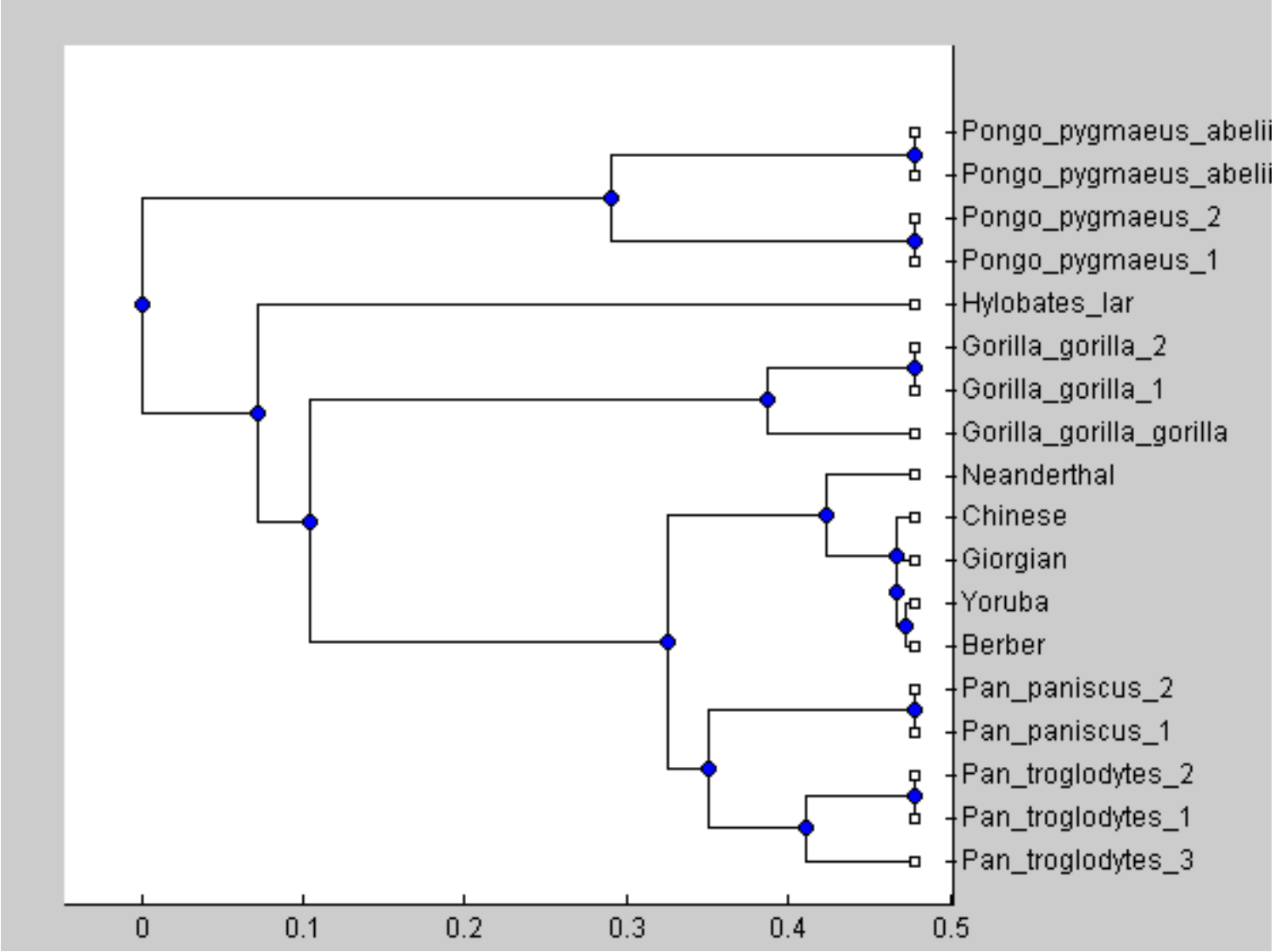
Case Study: Neanderthals



Case Study: Neanderthals



Case Study: Neanderthals



Case Study: Neanderthals

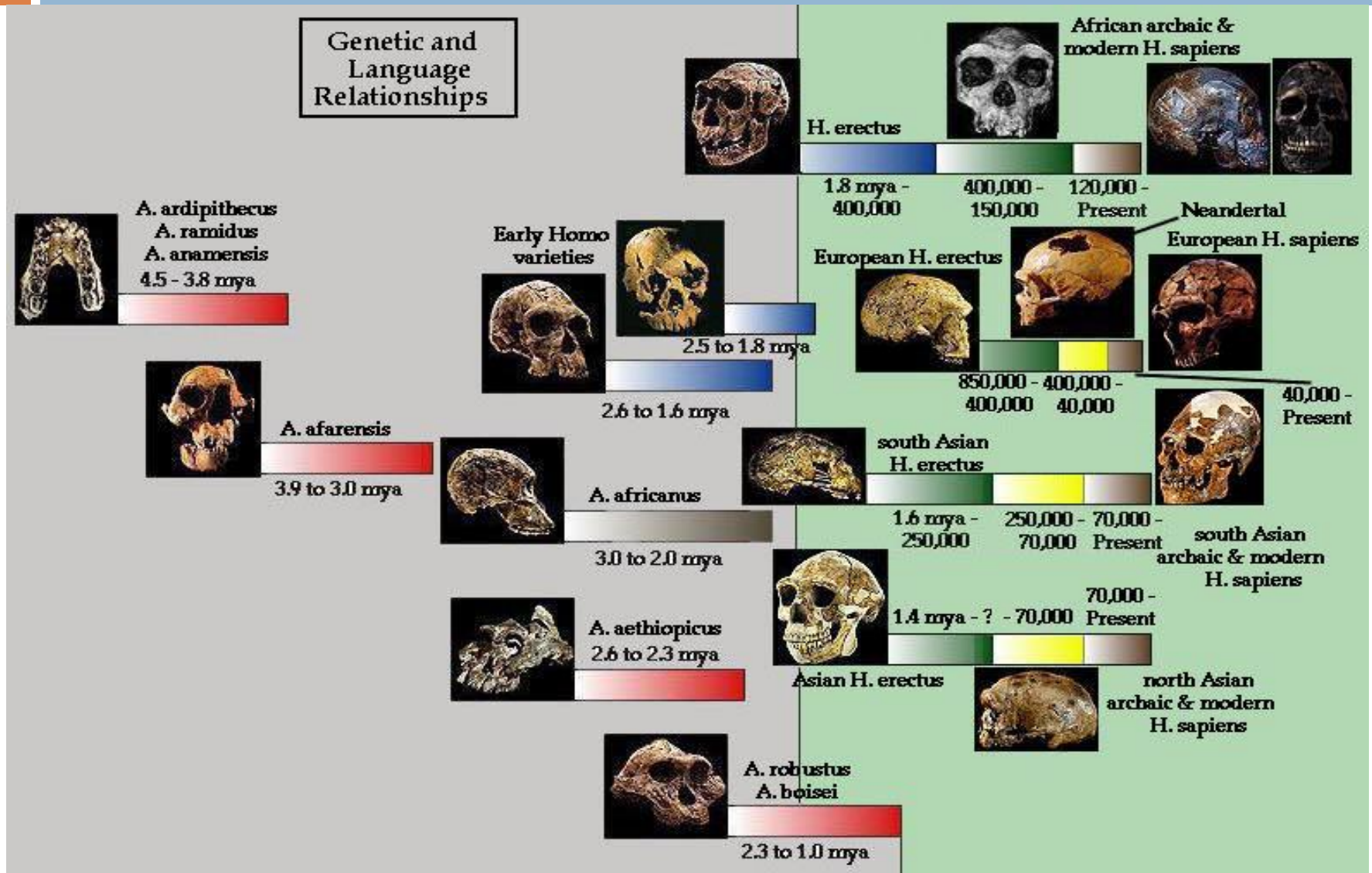
- Yoruba
  - Inhabitants of West Africa (21% of Nigeria)
- Berber
  - Northern population of Africa, west of the Nile River
    - Mainly Morocco and Egypt
- Georgian
  - South Caucasian
    - Georgia, Iran, Russia, Turkey

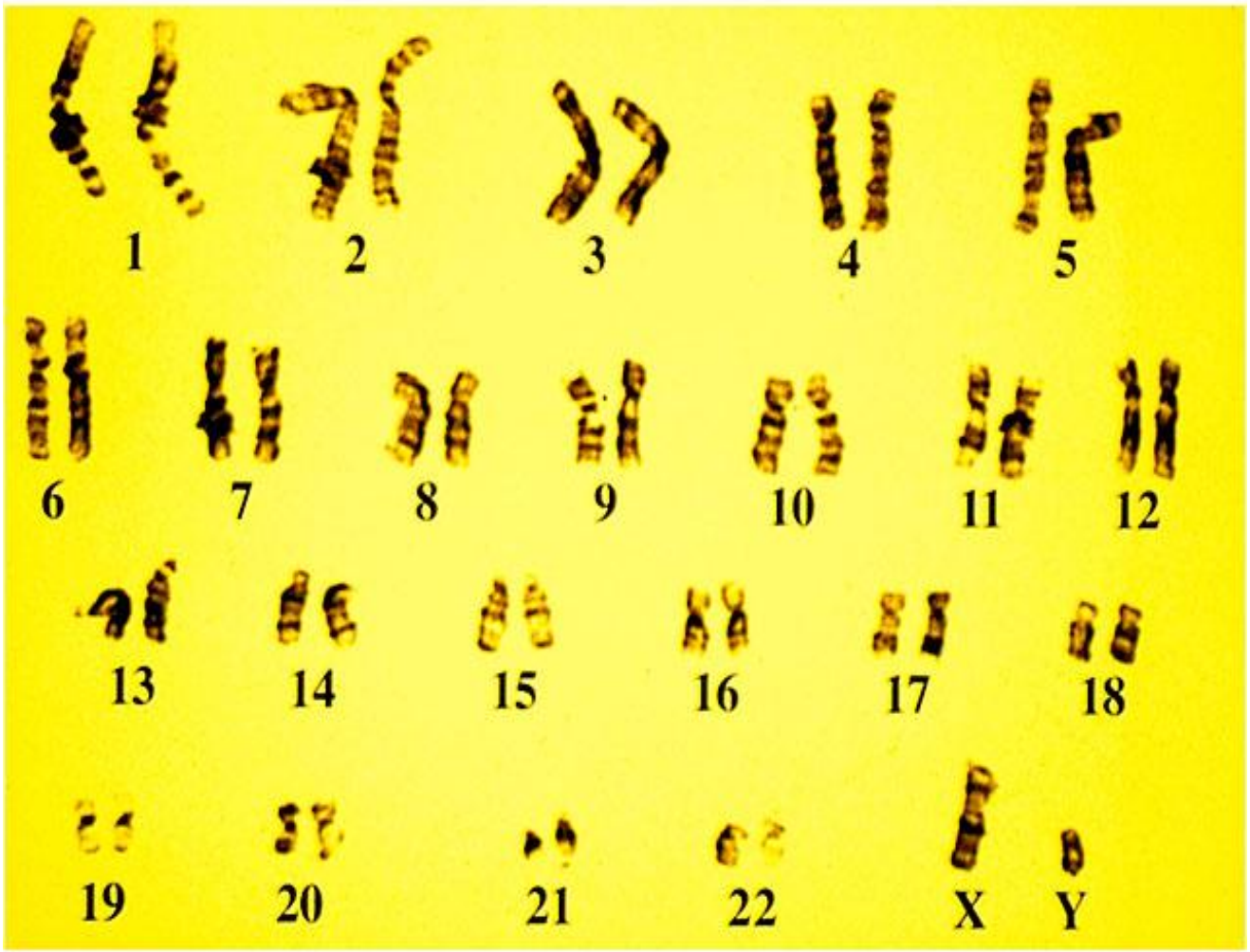
Conclusion

- Origin of Human Being
- Variation in DNA sequences
- Mitochondrial DNA
- Variation Among specie
- Estimation of genetic distance
- Neanderthal is not ancestor of modern man

Questions?

Historical Evolution





Chromosomes

- Human are diploid organisms
  - 23 pairs of chromosomes
- Diploid organisms have two types of cell:
  - somatic cells
    - Body cells
  - Germ cells (gametes)
    - Eggs and sperm

Mitosis

- DNA replication followed by cell division (body cell)
  - Generates two genetically identical cells
    - Each one receives a copy of each chromosome
  - Mutations may occur during DNA replication

The reason of Mitosis

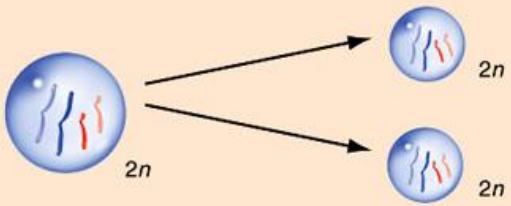
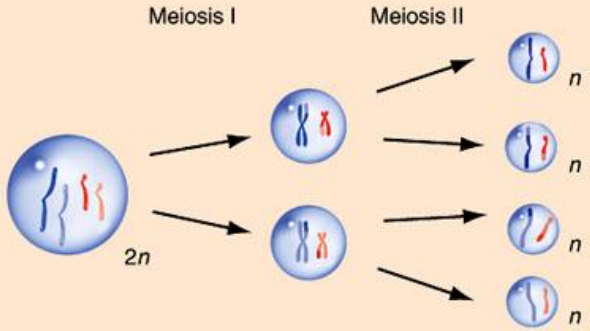
- Growth of multi-cellular organisms by increasing the number of cells
- Cells Replacement
  - Cells die and must be replaced by new cells
    - Blood cells live only 4 months
  - New cells must be exact copies of those replaced
- Regeneration of body parts
- Vegetative reproduction generating genetically similar offspring (clones)

Meiosis

- DNA replication followed by two rounds of cell division
- Generates four haploid cells
  - Gametes
  - Each half of the original cell DNA
    - Only one copy of each chromosome
  - Mutation and recombination may lead to new combinations DNA
    - None is identical to the original

Mitosis X Meiosis

TABLE 4.2 Mitosis and Meiosis: A Comparison

Mitosis	Meiosis
 The centromere splits at the beginning of anaphase.	 The centromere does not split during meiosis I.
	 Sister chromatids attach to spindle fibers from opposite poles during metaphase II.
	 The centromere splits at the beginning of anaphase II.
 Mitosis produces two new daughter cells, identical to each other and the original cell. Mitosis is thus genetically conservative.	 Meiosis produces four haploid cells, one (egg) or all (sperm) of which can become gametes. None of these are identical to each other or to the original cell, because meiosis results in combinatorial change.

Mutações Geminativas

- Can cause disease or susceptibility to them
  - Albinism
  - Dwarfism
  - Huntington's Disease
  - Alzheimer