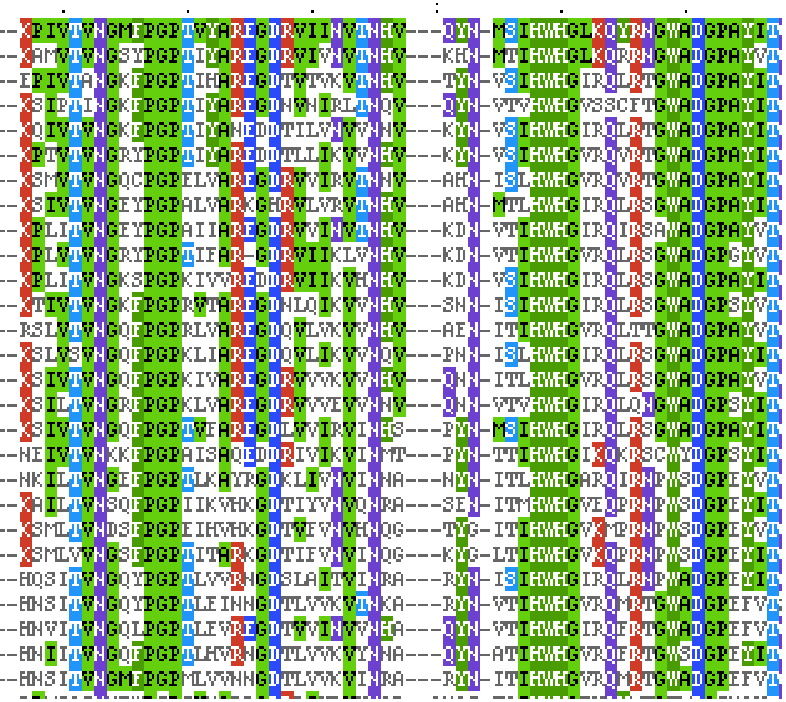
Aalto University
School of Science

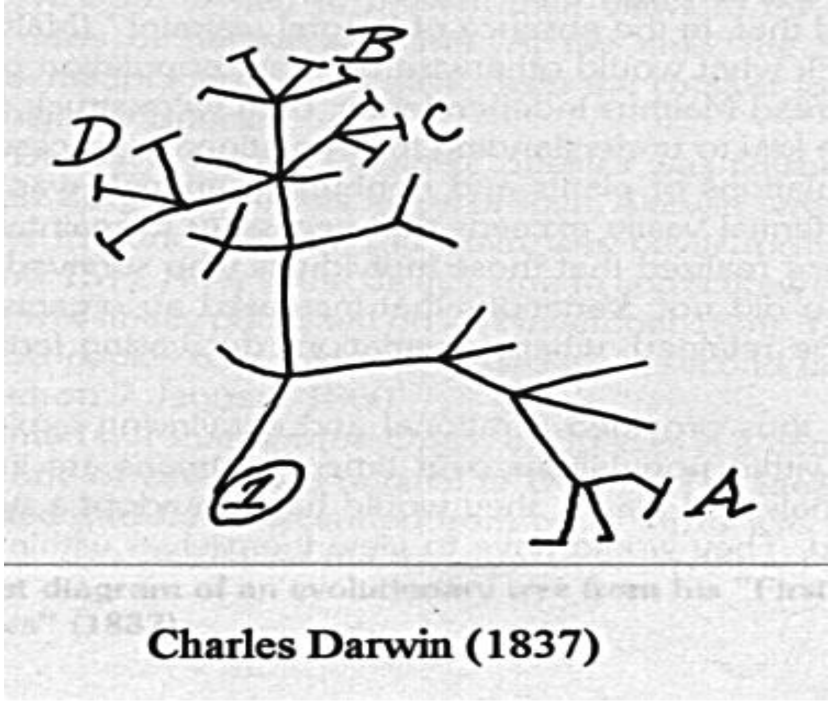
CS-E5865 Computational genomics

Autumn 2020, Lecture 6: Genome variation
Lecturer: Pekka Marttinen

Assistants: Alejandro Ponce de León, Zeinab Yousefi, Onur Poyraz

From sequences to genetic relationships

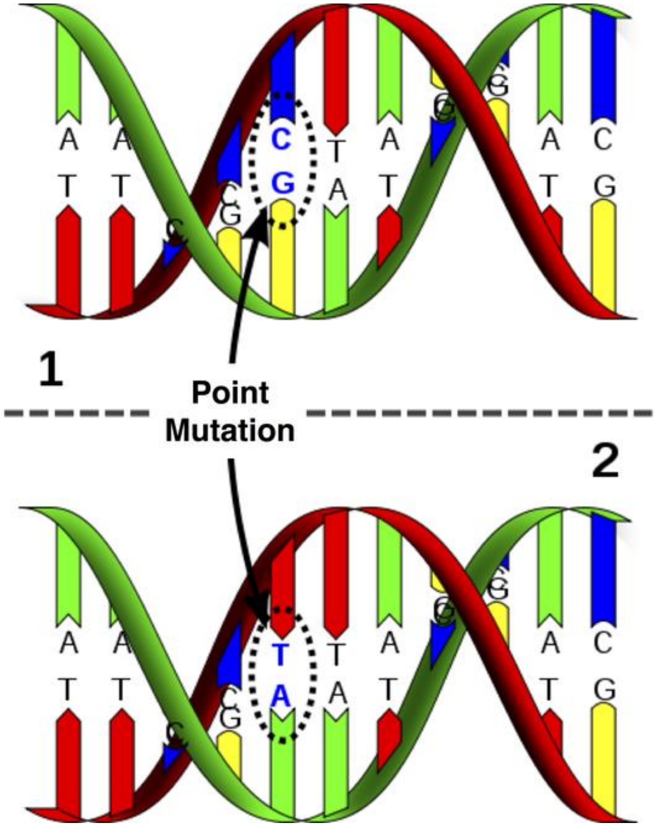




Charles Darwin (1837)

Variation in DNA sequences

- Mutations
  - mistakes in DNA replication
- Mutations are rare
  - on average, one mistake per 200 million to 1 billion nucleotides
- Consequently
  - most mutations in DNA are *inherited* from previous generations
- Shared mutations ≈ shared history

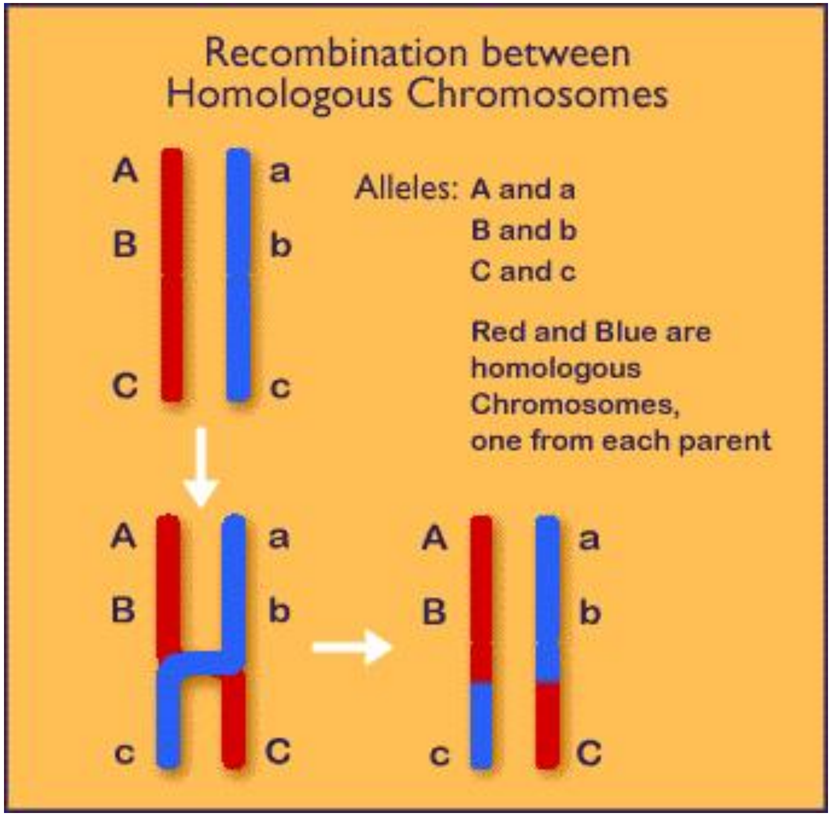


http://rosalind.info/media/point_mutation.png

Aalto University
School of Science

Variation in DNA sequeces

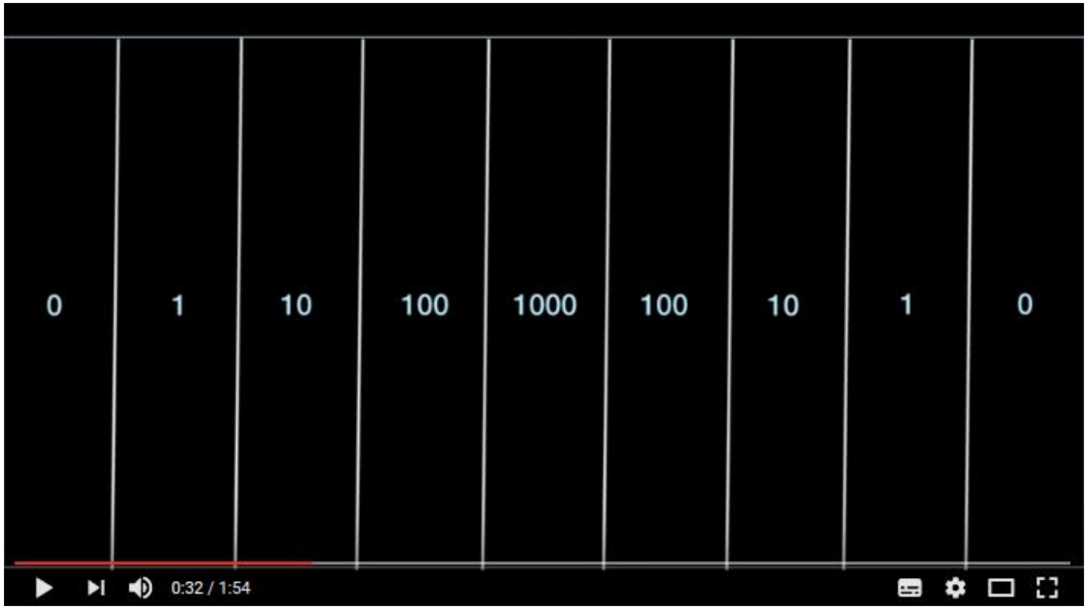
- Recombination:
  – mixing of DNA from multiple parents (in species with sexual reproduction, but also in clonally evolving bacteria)


Recombination between Homologous Chromosomes
A
a
B
b
C
c
Alleles: A and a
B and b
C and c
Red and Blue are homologous Chromosomes, one from each parent
A
a
B
b
c
C
A
a
B
b
c
C

http://members.cox.net/amgough/Fanconi-genetics-genetics-primer.htm

Natural selection acts on heritable variation

Growing *E.coli* bacteria on a Petri Dish with varying concentrations of antibiotic.

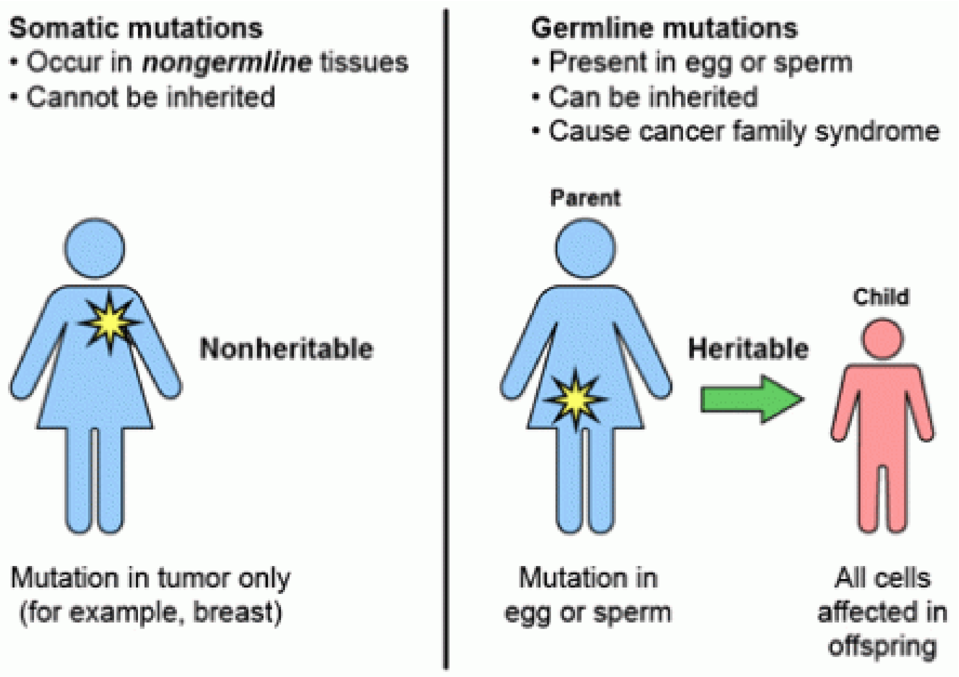


https://www.youtube.com/watch?v=plVk4NVIUh8

Types of mutations

- Mutations originate in single individuals
- Mutations can become *fixed* in a population
  - every individual has that mutation
- Neutral mutations: do not affect the organisms functions or ability to generate offspring
- Deleterious mutations: disrupt some functions
  - Under negative selection
- Advantageous mutations: enhance some function
  - Under positive selection

Germline mutations



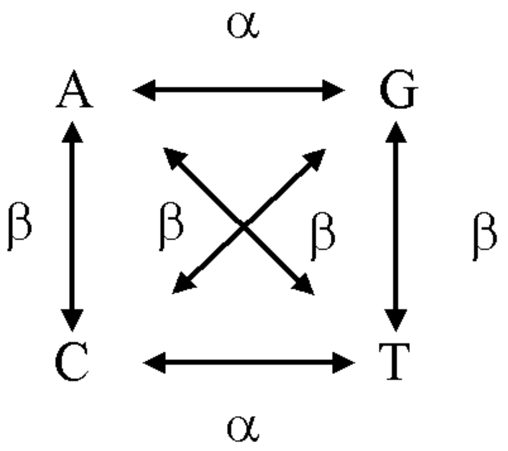
https://autismsciencefoundation.wordpress.com/2015/12/02/brain-tissue-reveals-more-genetic-clues-to-autism/

Substitution rate and mutation rate

- Two rates to consider:
  - Mutation rate: rate at which new mutations arise
  - Substitution rate: rate at which new mutations become fixed in a species
    - depends on mutation rate and selection.
- Mutation rates and substitution rates are related
  - substitutions can happen only after mutations occur
- But they refer to different processes.

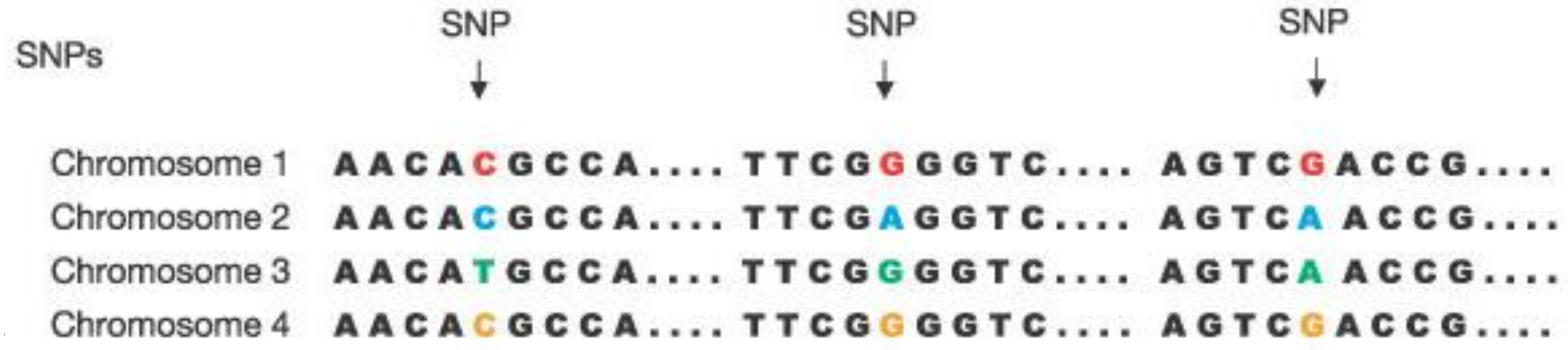
Transitions and transversions

- Not all point mutations are equally likely
  - There are 2 types of nucleotides: purines (A,G) and pyrimidines (T,C)
- Transitions (α) = mutations within the groups
- Transversions (β) = mutations between groups
- Transitions are more common
  - In humans, transitions are at least 2 times more likely than transversions
  - More SNP's of the type A/G and C/T



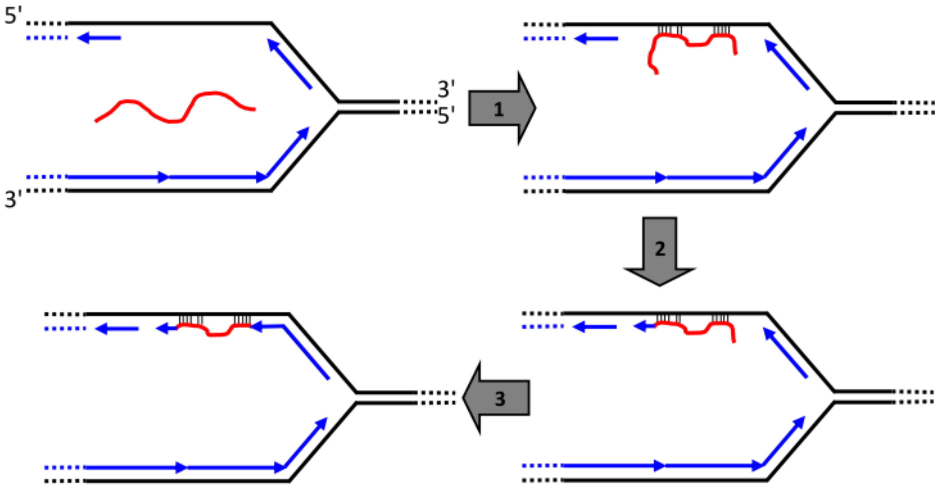
Types of genetic variations

- Polymorphism = occurrence of two or more distinct genetic variants at one genomic position (locus)
- Different types of genetic variations:
  - Single nucleotide polymorphism (SNP)
    - The most common mutations
  - Microsatellites: short regions of repeat sequences e.g. ACACACAC
    - Different individuals can have different number of repetitions
  - Indels: insertions or deletions of DNA sequence
  - Rearrangements: inversion, duplication or transpositions of DNA
- Different variants are called alleles


SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....

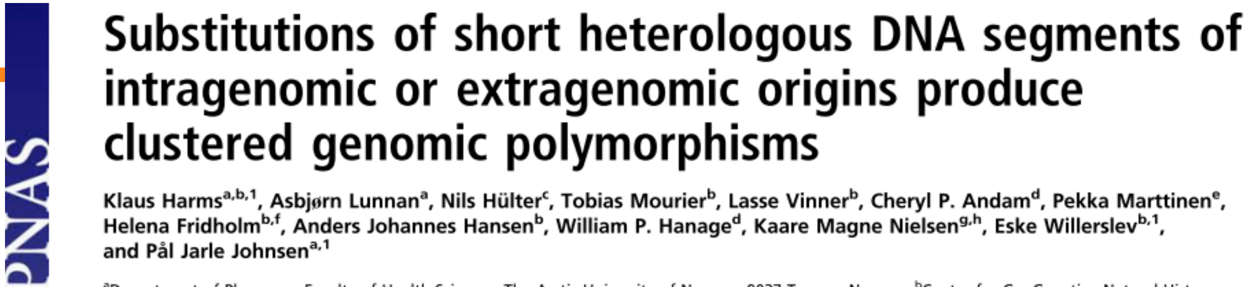
http://www.nature.com/nature/journal/v426/n6968/fig_tab/nature02168_F1.html

Research at aalto



```
P  GACTTCATCCGTGACTTCCATCAGCTAGTGAAGGCCCTACCCCAGTATCAGCACTCC
   |||||||||||||||||||| | ||     |||||||||||||||||||||||||||
R  GACTTCATCCGTGACTTCCACCTGCAGCATAAGGCCCTACCCCAGTATCAGCACTCC
    ||   |        ||||||||||||||||||||||||| | |
D  CACCAAAAAAGCCCGTTCCACCTGCAGCATAAGGCCCTACACGATAACTTTGTAATG
```

Aalto University
School of Science



PNAS

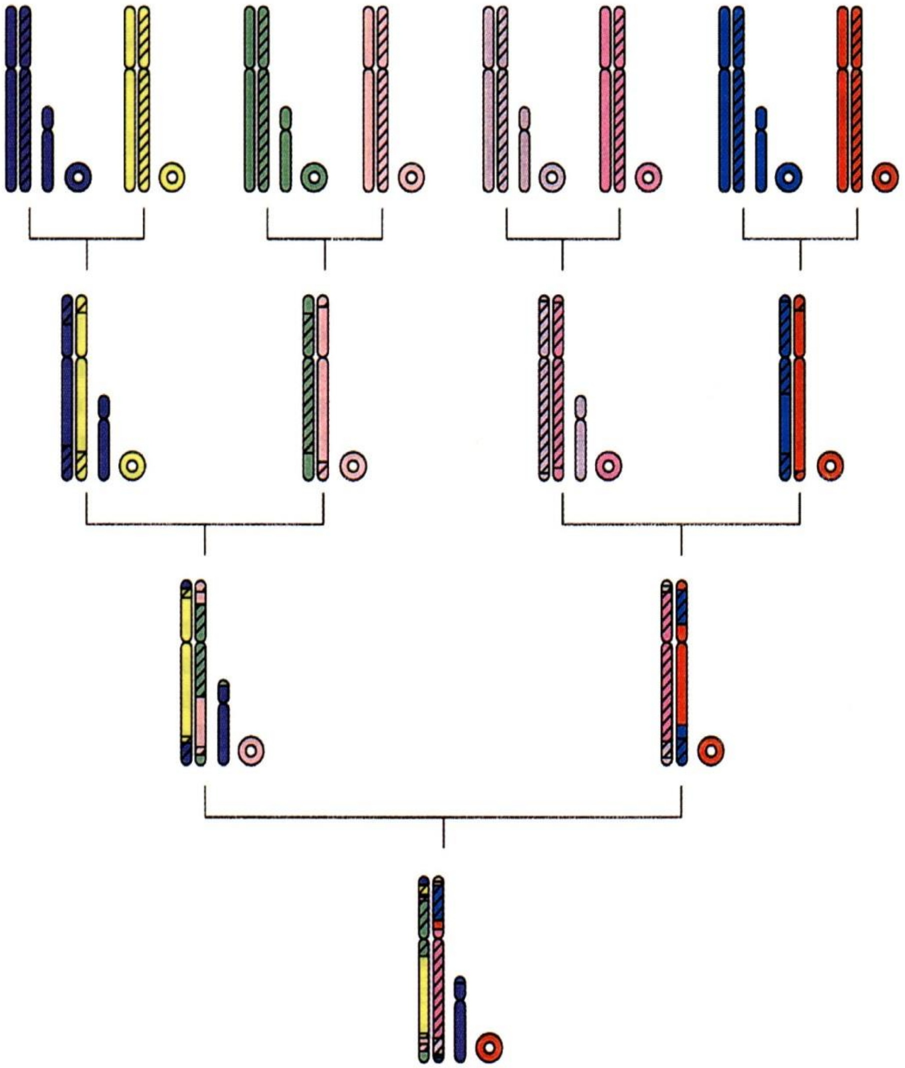
Substitutions of short heterologous DNA segments of intragenomic or extragenomic origins produce clustered genomic polymorphisms

Klaus Harms[a,b,1], Asbjørn Lunnan[a], Nils Hülter[c], Tobias Mourier[b], Lasse Vinner[b], Cheryl P. Andam[d], Pekka Marttinen[e], Helena Fridholm[b,f], Anders Johannes Hansen[b], William P. Hanage[d], Kaare Magne Nielsen[g,h], Eske Willerslev[b,1], and Pål Jarle Johnsen[a,1]

Mitochondrial DNA: a model for variation analysis

- Mitochondrial DNA (mtDNA) is inherited only from the mother

- mtDNA is useful for studying human evolution
  - The mutation rate is 10 times higher than for nuclear DNA



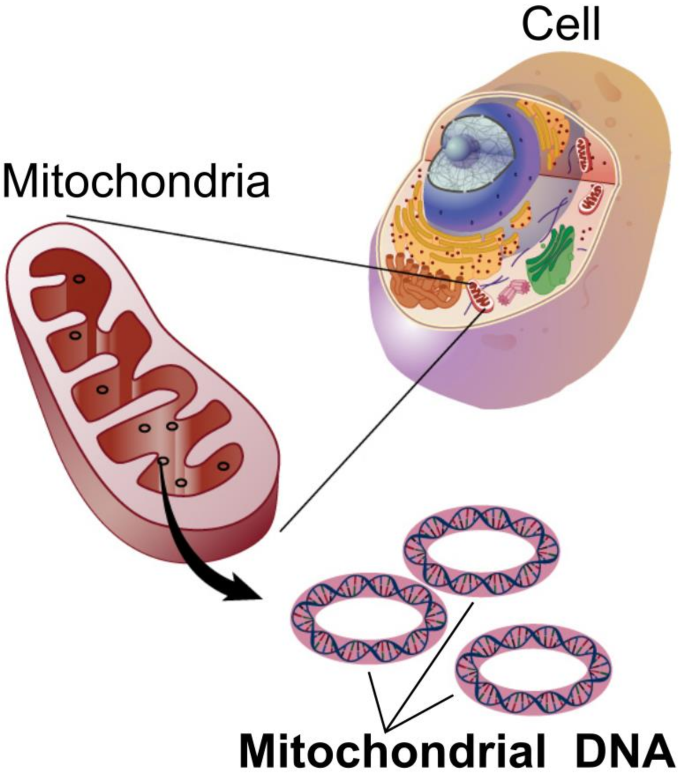
Aalto University
School of Science

Mitochondrial DNA: technical advantages

- mtDNA is inherited only from the mother → only a single haplotype
  - Inferring haplotype for nuclear DNA is a computational problem known as phasing
  - Suppose we have 2 polymorphisms in a nuclear gene of an individual, i.e., there are 2 differences between the maternal and paternal versions of the gene, e.g., one A/G and one C/T
  - There are 2 possible configurations: $\begin{matrix} -A-C- \\ -G-T- \end{matrix}$ and $\begin{matrix} -A-T- \\ -G-C- \end{matrix}$

Mitochondrial DNA: technical advantages

- Each cell has multiple copies of mtDNA → easy to isolate and sequence
- It is feasible to extract mtDNA from old tissue, e.g., mummies or Neanderthal skeletons



https://en.wikipedia.org/wiki/Mitochondrial_DNA

Estimating genetic distance

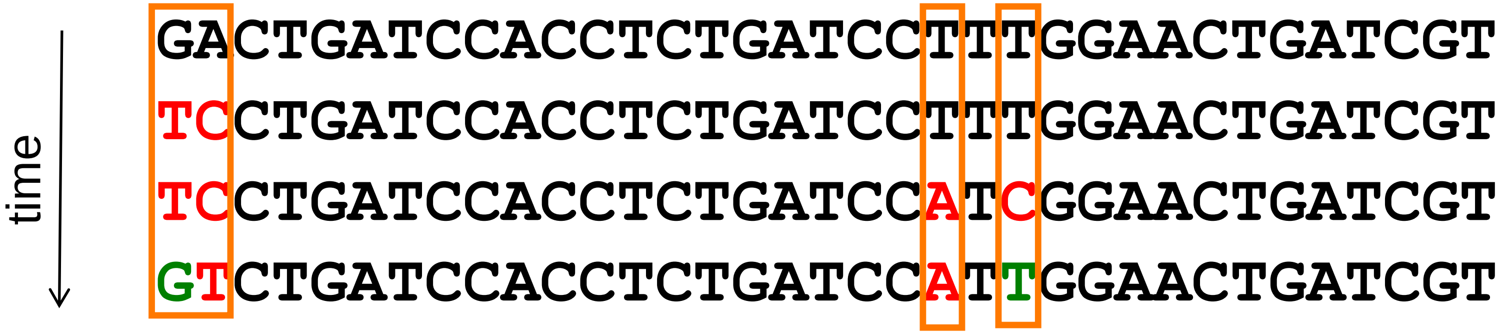
- *Genetic distance* = the number of substitutions that have accumulated between two homologous sequences after they diverged from a common ancestor
- First approximation: proportion of sites that are different between the two sequences
  - sometimes it is called the *p*-distance.

```
GACTGATCCACCTCTGATCCTTTGGAACTGATCGT
GTCTGATCCACCTCTGATCCATTGGAACTGATCGT
```

If 10 sites are different between two sequences, each 100 bp long, then *p*= 10% = 0.1

Estimating genetic distance

- Genetic distance will be underestimated by counting differences:
  - Mutations can also convert a nucleotide back to the original: G➔T➔G
  - Multiple mutations may occur at the same position: A➔T➔C



GACTGATCCACCTCTGATCCTTTGGAACTGATCGT

TCCTGATCCACCTCTGATCCTTTGGAACTGATCGT

TCCTGATCCACCTCTGATCCATCGGAACTGATCGT

GTCTGATCCACCTCTGATCCATTGGAACTGATCGT

Jukes-Cantor Model

- Simple probabilistic model for correcting for multiple substitutions per position
- Assumptions:
  - all positions in a sequence evolve independently
  - all 3 possible substitutions from one base to any of the other 3 are equally likely
- Denote by α the probability of a substitution occurring at a given position in a time unit
- The rate of substitution for each nucleotide is 3α per unit time

Jukes-Cantor Model

- Consider the Markov chain over nucleotide substitutions, with the transition matrix below

$$M_{JC}= \begin{array}{c|cccc} & \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \\ \hline \mathtt{A} & 1-3\alpha & \alpha & \alpha & \alpha \\ \mathtt{C} & \alpha & 1-3\alpha & \alpha & \alpha \\ \mathtt{G} & \alpha & \alpha & 1-3\alpha & \alpha \\ \mathtt{T} & \alpha & \alpha & \alpha & 1-3\alpha \end{array}$$

where **α** is the substitution probability *for a given pair of nucleotides in a position in a time unit* (the same for all nucleotide pairs)

Jukes-Cantor Model

- Consider we start with nucleotide A in a particular position (time 0)
- At time 1, the probability of still having A at this site is given by $P_{A(1)} = 1 - 3\boldsymbol{\alpha}$
- The probability of having A at time 2

$$P_{A(2)} = (1 - 3\boldsymbol{\alpha})\ P_{A(1)} + \boldsymbol{\alpha}\,(1 - P_{A(1)})$$

- We consider two possible scenarios:
  - the nucleotide has remained unchanged from time 0 to time 2, and
  - the nucleotide has changed to T, C, or G at time 1, but has subsequently reverted to A at time 2.

Jukes-Cantor Model

- We actually have for any time t:

$$P_{A(t+1)} = (1 - 3\boldsymbol{\alpha})\ P_{A(t)} + \boldsymbol{\alpha}(1 - P_{A(t)})$$

- By solving the ’non-homogeneous linear recurrence relation’ above, we get the probability that after time t we have nucleotide A in the position

$$P_{A(t)} = \frac{1}{4} + (\frac{3}{4})e^{-4\alpha t}$$

- The general form of this equation is

$$P_{ii(t)} = \frac{1}{4} + (\frac{3}{4})e^{-4\alpha t}$$

Jukes-Cantor Model

- If the initial nucleotide is G instead of A, then

$$P_{GA(t)} = \frac{1}{4} - (\frac{1}{4})e^{-4\alpha t}$$

- More generally, the probability $\mathrm{P}_{\mathrm{ij(t)}}$ that a nucleotide will become j at time t, given that it was i ($\mathrm{i} \neq \mathrm{j}$) at time 0 is

$$P_{ij(t)} = \frac{1}{4} - (\frac{1}{4})e^{-4\alpha t}$$

Jukes-Cantor Model

- The probabilities of observing substitutions i→j after t time units are given by the matrix $M(t)$

$$M(t) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & r(t) & s(t) & s(t) & s(t) \\ C & s(t) & r(t) & s(t) & s(t) \\ G & s(t) & s(t) & r(t) & s(t) \\ T & s(t) & s(t) & s(t) & r(t) \end{array}$$

where $r(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$ and $s(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$

- $s(t)$ = the probability of observing a substitution after t time steps

Jukes-Cantor Model

- Probability that two homologous sequences differ at a given position after time t:

P=1 – prob they are identical

=1- (prob. of both staying the same

+

prob. of both changing to the same thing)

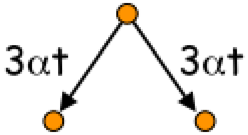
$=1- \{(P_{AA(t)})^2 + (P_{AT(t)})^2 + (P_{AC(t)})^2 + (P_{AG(t)})^2\}$

$= \frac{3}{4} (1-e^{-8\alpha t})$

Aalto University
School of Science

Jukes-Cantor Model

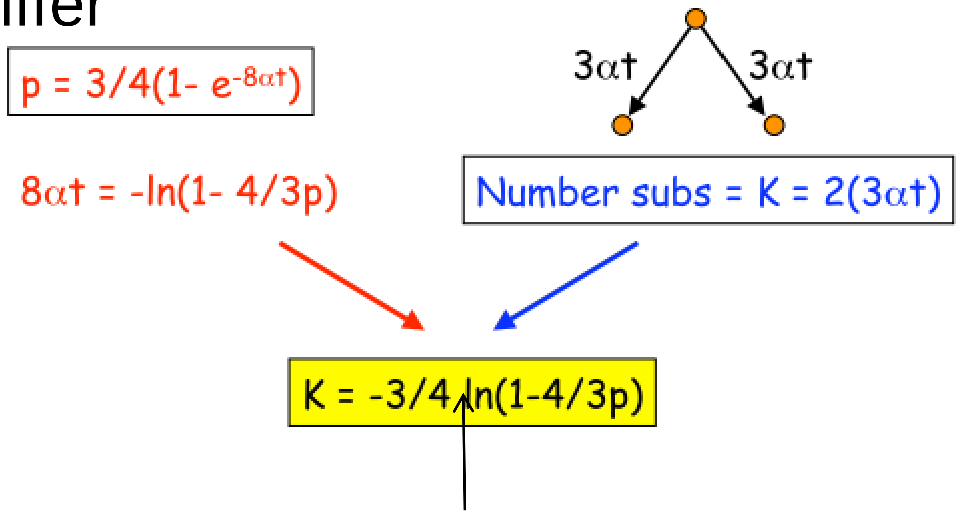
- Let now K be the number of substitutions per site since the time of divergence between two sequences.



Number subs = K = 2(3αt)

Jukes-Cantor Model

- p is the same as the proportion of differences observed between 2 sequences
- Calculate number of substitutions in terms of proportion of sites that differ



$p = 3/4(1- e^{-8\alpha t})$

$8\alpha t = -\ln(1- 4/3p)$

Number subs = $K = 2(3\alpha t)$

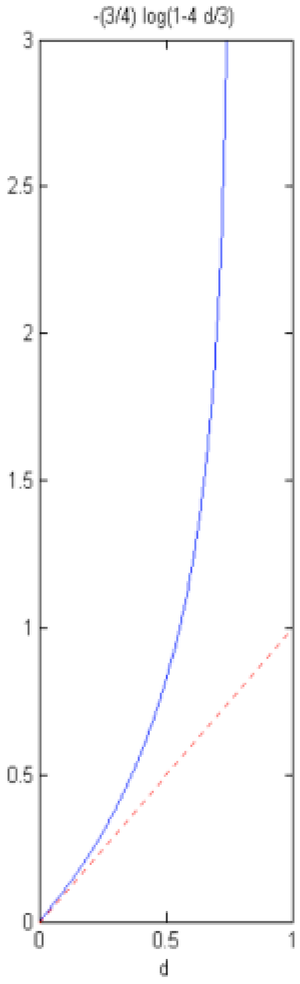
$K = -3/4 \ln(1-4/3p)$

Jukes-Cantor estimate of genetic distance (p is the observed proportion of different nucleotides between two sequences, K is the average number of substitutions per site)

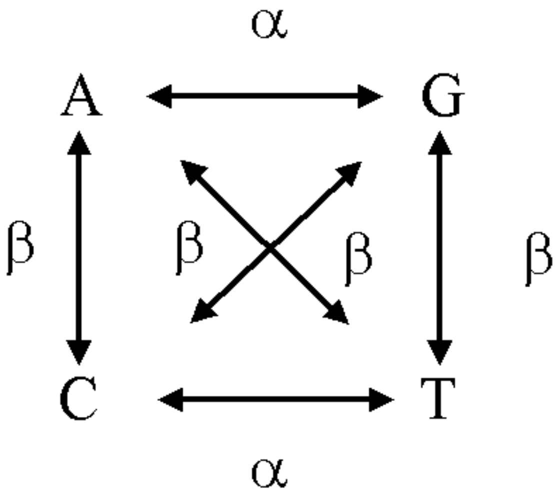
Jukes-Cantor Model

- For small d using the approximation $\ln(1+x) \approx x$, we obtain that $K \approx d$
  - So: actual distance ≈ observed distance

- When aligning random sequences ($d \rightarrow$ ¾), we have that $K \rightarrow \infty$
  - So: if the sequences are random, Jukes-Cantor estimates infinite genetic distance



Kimura 2-parameter model



- α = transition probability
- β = transversion probability

- The Jukes-Cantor model assumes that all substitutions are equaly likely
  - This is not always the case
  - Substitutions A ⬄ G and T ⬄ C (called transitions) happen more frequently than A ⬄ T, A ⬄C, T ⬄G, C ⬄G (transversions)

- Kimura model takes this into account by introducing a separate substitution rate for the two groups

Kimura 2-parameter model

- The Kimura 2-parameter estimate of the genetic distance between 2 sequences is

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

- *P is the proportion of sites with observed transitions and Q is the proportion of sites with observed transversions*

Studying variation – why?

- Understanding Evolution
- Determine disease risk
- Individualised medicine (pharmacogenomics)
- Forensic studies
- Biological markers

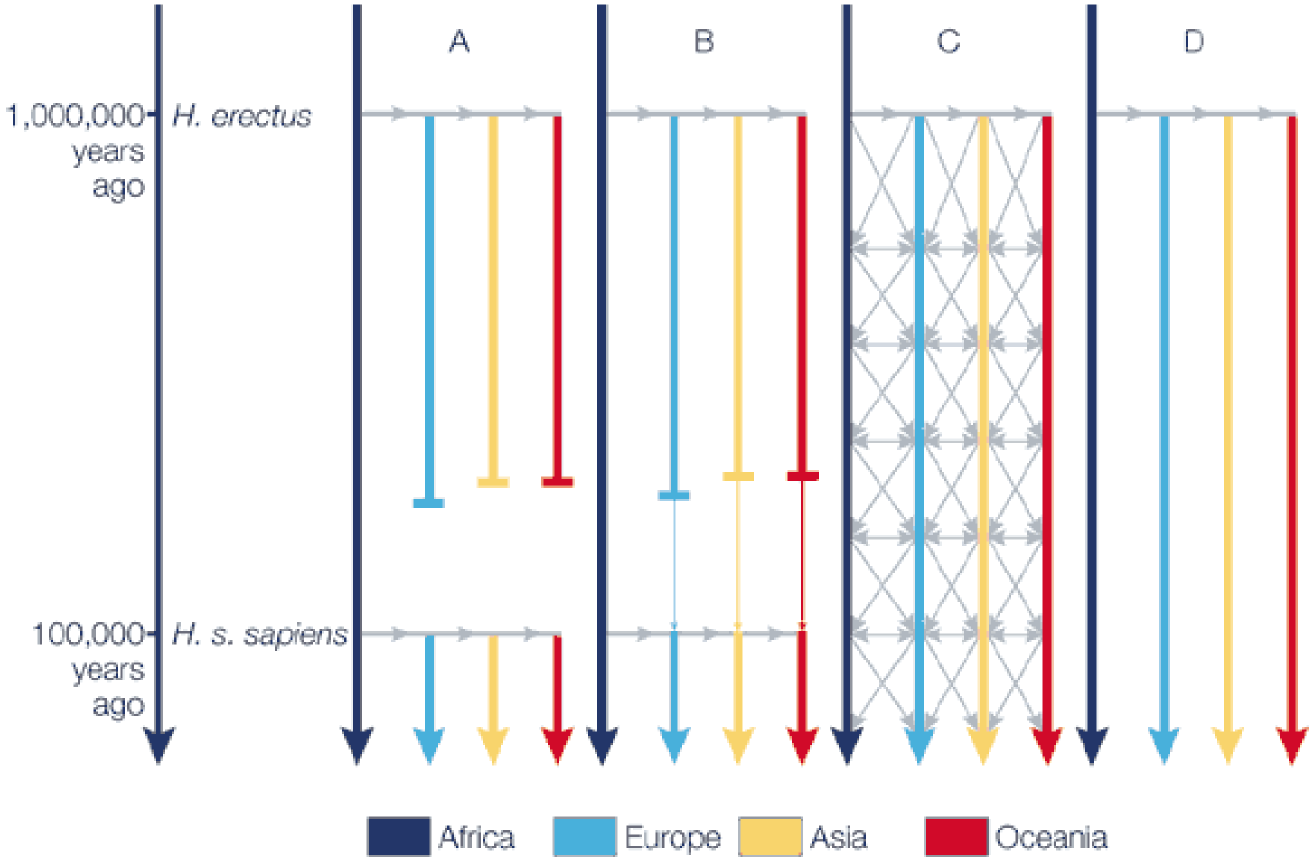
Origin of modern humans:

Hypothesis A:

“Out of Africa”

Hypothesis D:

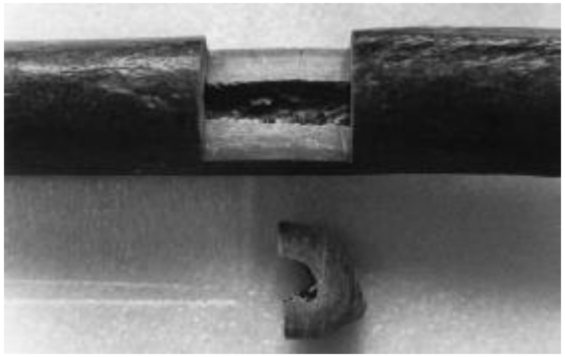
Multiregional

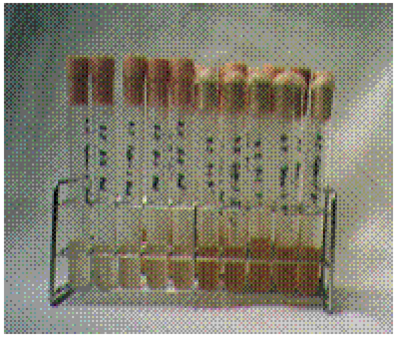


Nature Reviews | Genetics (2000)

mtDNA Analysis Supports “Out of Africa”

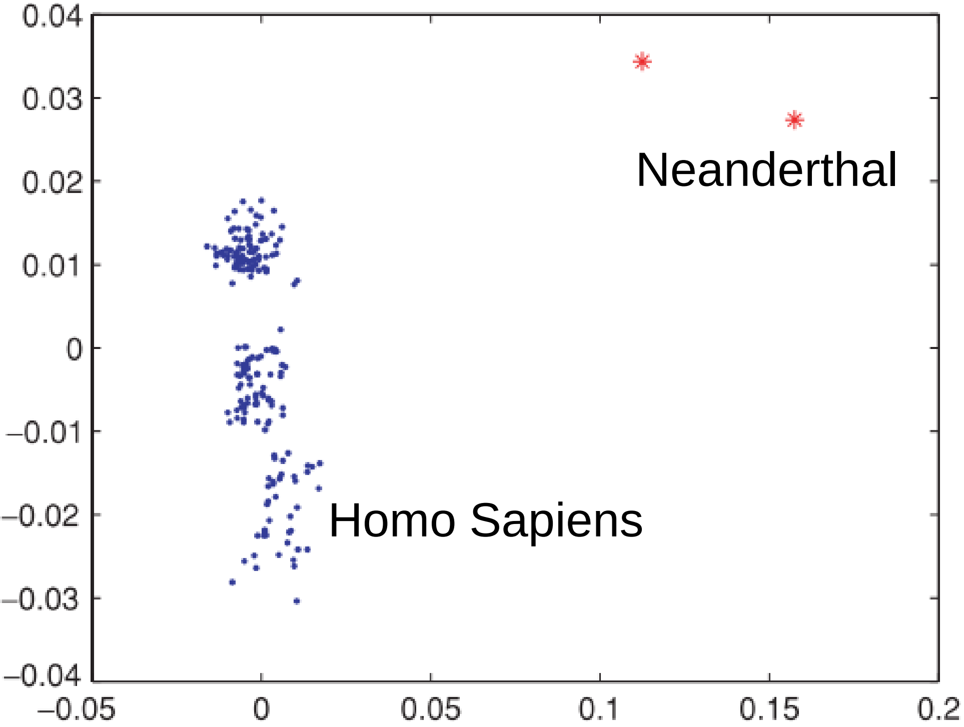
- mtDNA was sequenced from several Neanderthal skeletons and compared with modern human mtDNA (around year 2000)
  - 206 mtDNA samples from modern humans
  - 2 mtDNA from 2 Neanderthals





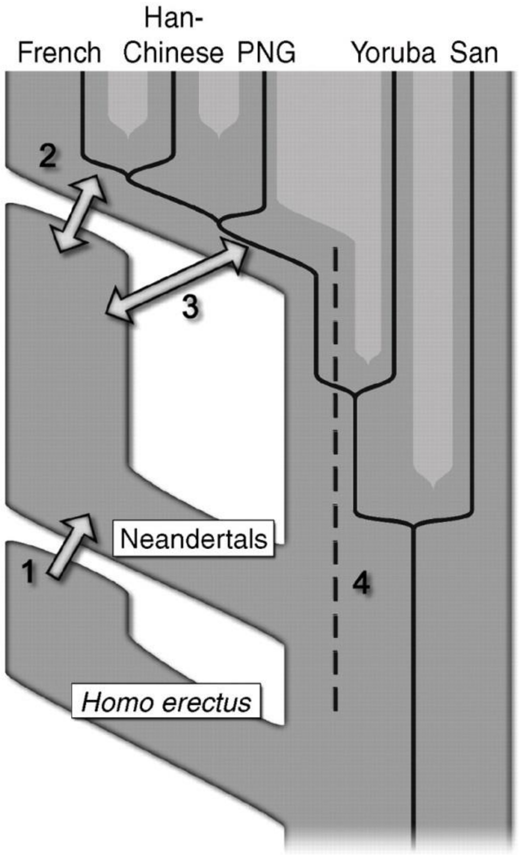
mtDNA Analysis Supports “Out of Africa”

- Pairwise genetic distances using the Jukes-Cantor formula
  - The average distance between H. Sapiens was 0.025
  - The average distance between H. Sapiens and Neanderthal was 0.14
  - Neanderthal DNA is different from H. Sapiens.



Analysis of Neanderthal whole-genomes in 2010

- Interbreeding between Neanderthals and modern humans contributed 1 to 4% of the genome of present-day non-Africans.
- Scenario 3 supported by the data



Richard E. Green et al. Science 2010;328:710-722

Forensic studies: Example

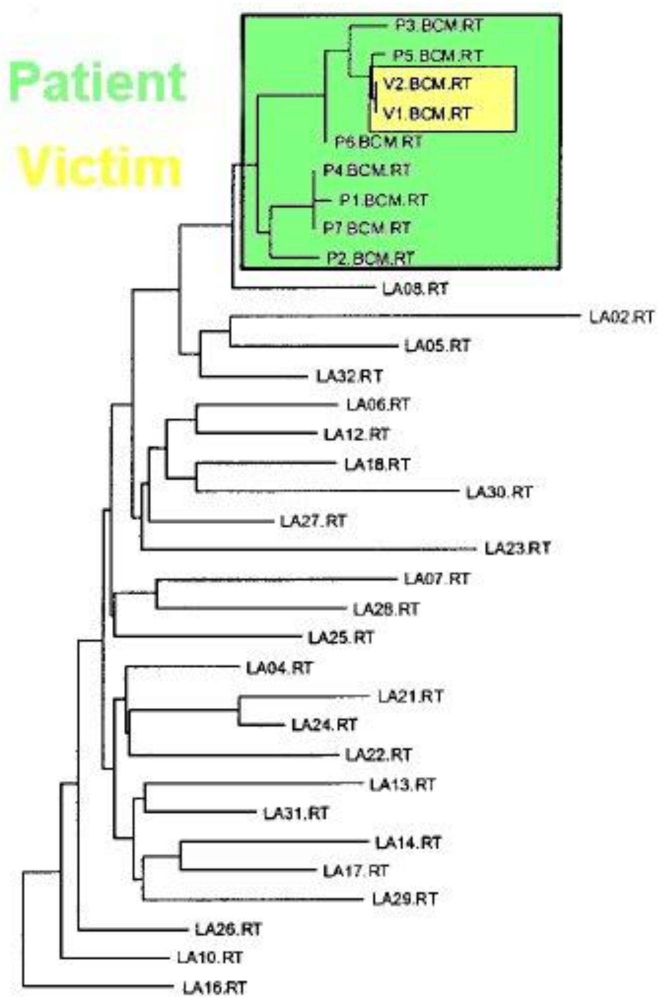
- Lafayette, Louisiana, 1994 – A woman claimed her ex-boyfriend (who was a physician) injected her with HIV+ blood.
- Records show the physician had drawn blood from an HIV+ patient that day.
- But how can we prove that the blood from that specific HIV+ patient ended up in the woman?

Forensic studies: Example

- HIV has a high mutation rate, which can be used to trace paths of transmission.
- Two people who got the virus from two different people will have very different HIV sequences.
- *Tree reconstruction* methods were used to track changes in HIV genes.

Forensic studies: Example

- Took samples from the patient, the woman, and control HIV+ patients.
- In tree reconstruction, the woman’s sequences were found to be evolved from the patient’s sequences, indicating a close relationship between the two.
- This was the first time phylogenetic analysis was used in court.



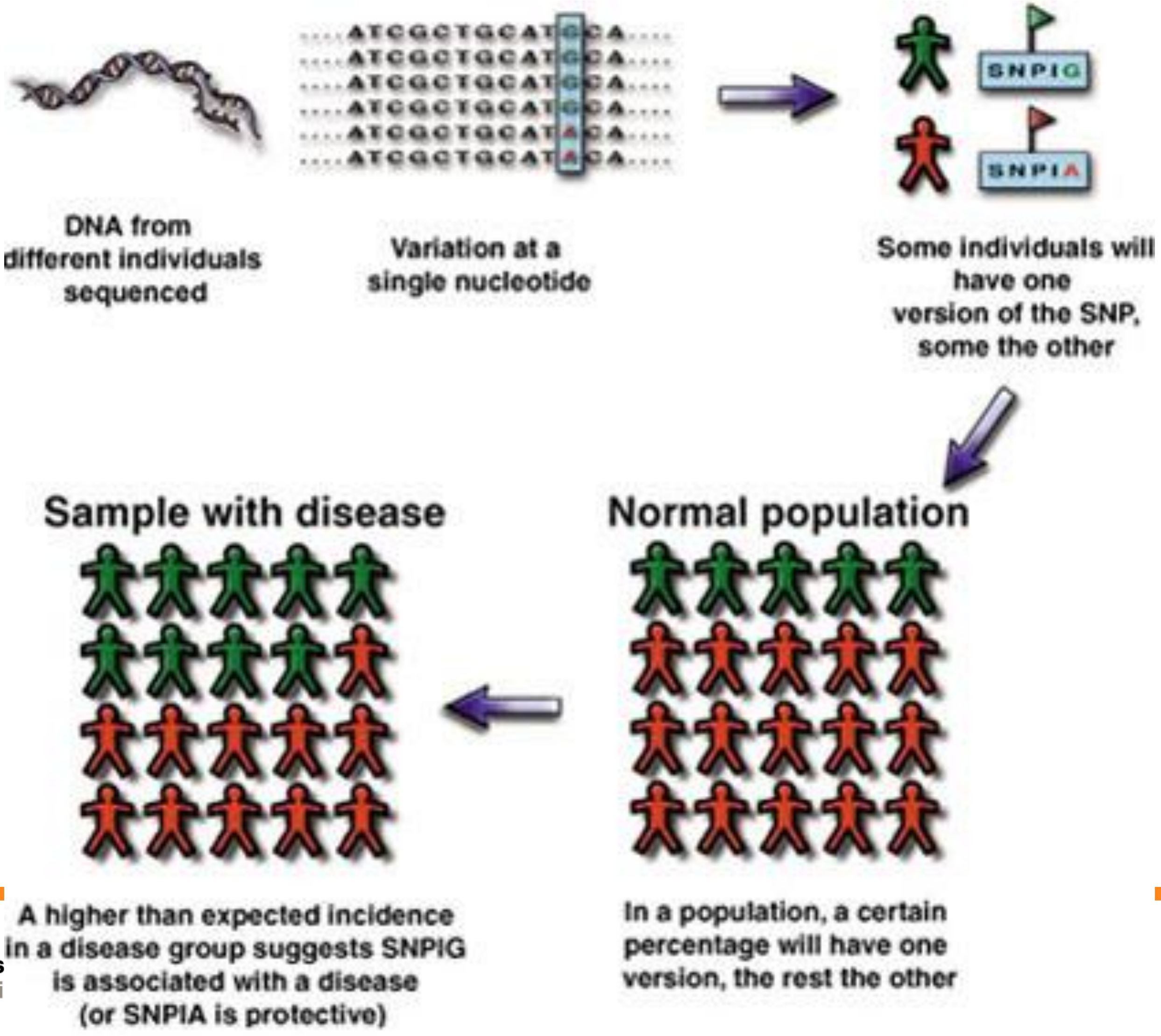
Genetic variance and diseases

- Disease gene discovery
  - Association studies, certain SNPs are susceptible for diabetes
  - Chromosome aberrations, duplication / deletion might cause cancer
- Personalized Medicine
  - Drug only effective if you have one allele

Types of Genetic Disorders

- Chromosome abnormalities
  - Addition or deletion of entire chromosomes or parts of chromosomes
  - Typically more than 1 gene involved
  - Classic example is trisomy 21 - Down syndrome
- Single gene disorders
  - Huntington's Disease caused by excess CAG repeats in huntington's protein gene
- Polygenic Disorders
  - In many cancers (solid tumors) somatic mutations that induce cell proliferation

Using SNPs to Track Predisposition to Disease and other Genetic Traits


....ATCGCTGCATGGCA....
....ATCGCTGCATGGCA....
....ATCGCTGCATGGCA....
....ATCGCTGCATGGCA....
....ATCGCTGCATACA....
....ATCGCTGCATACA....
SNPIG
SNPIA
DNA from different individuals sequenced
Variation at a single nucleotide
Some individuals will have one version of the SNP, some the other
Sample with disease
Normal population
A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)
In a population, a certain percentage will have one version, the rest the other

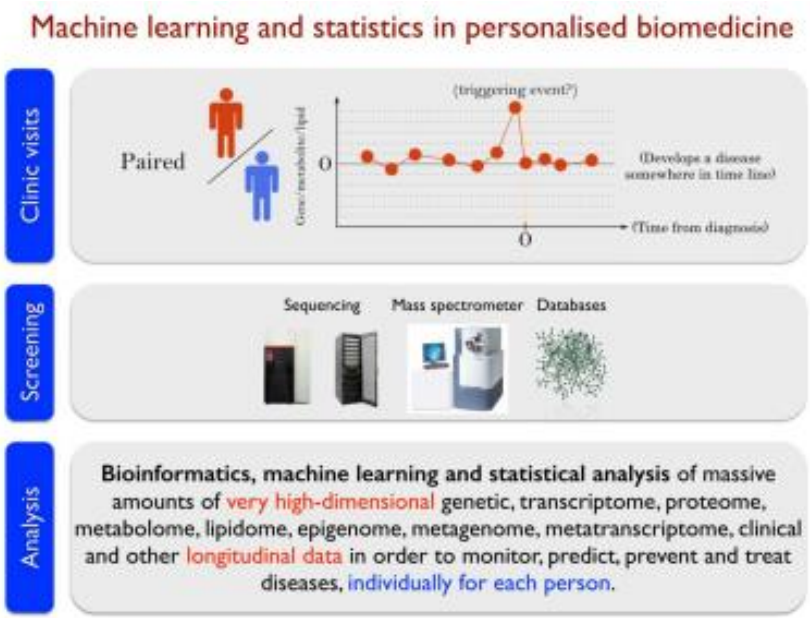
© Gibson & Muse, A Primer of Genome Science

Haplotype Map of the Human Genome

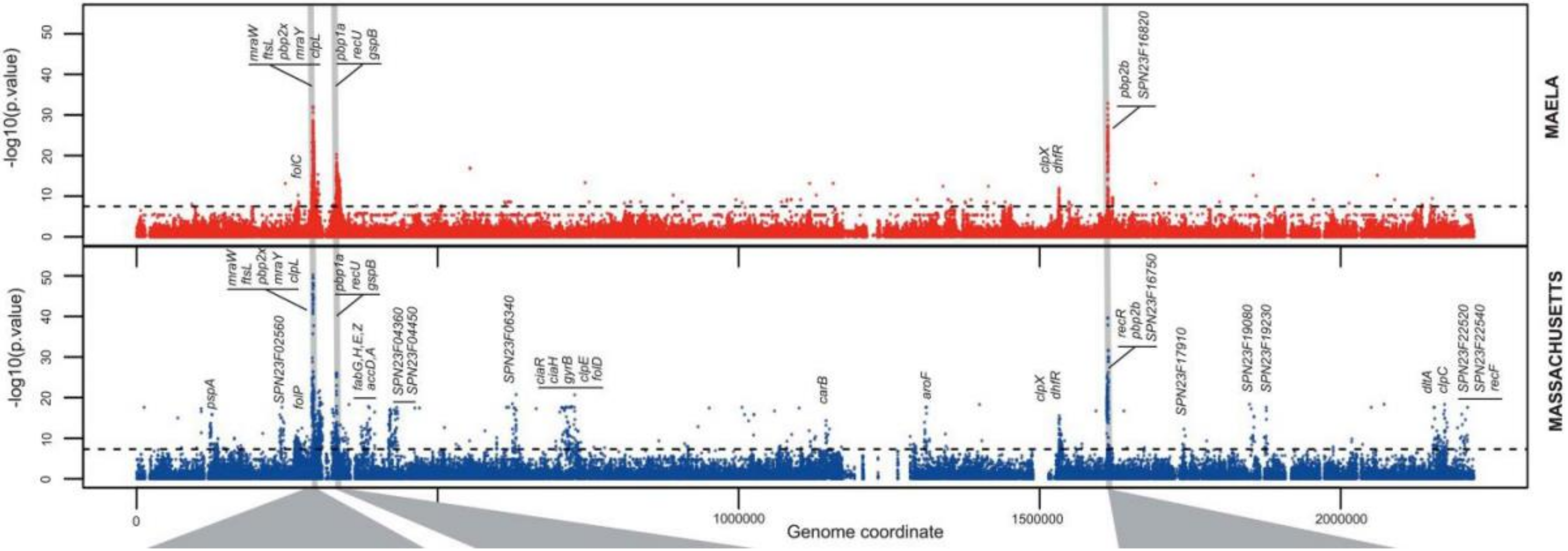
- HapMap is a catalog of common genetic variants that occur in humans
- The Project is designed to provide information that other researchers can use to link genetic variants to the risk for specific illnesses

Research at Aalto

- Personalized medicine
- Statistical genetics
- etc.



Bioinformatics, machine learning and statistical analysis of massive amounts of very high-dimensional genetic, transcriptome, proteome, metabolome, lipidome, epigenome, metagenome, metatranscriptome, clinical and other longitudinal data in order to monitor, predict, prevent and treat diseases, individually for each person.



Chewapreecha et al. 2014, Plos Genetics