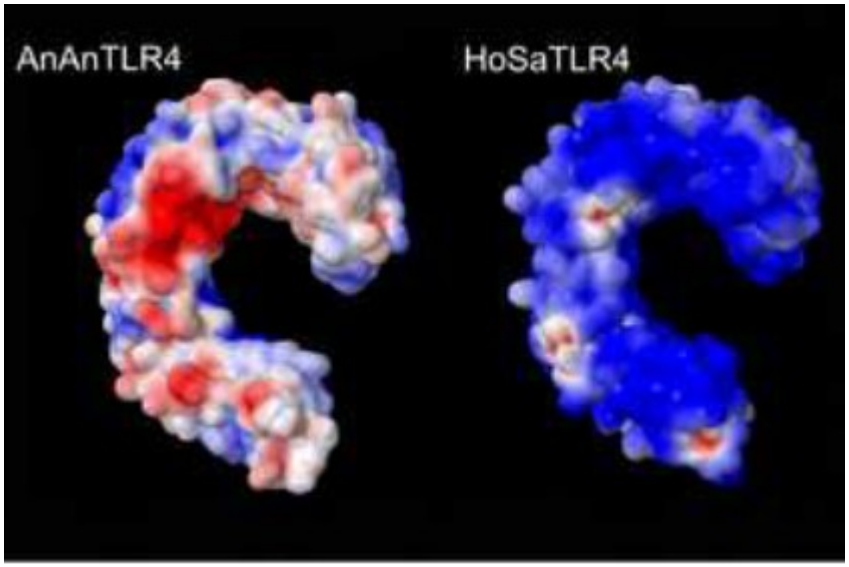
Molecular Applications in Zoology



X.

Functional genetic variability:

From SNP to selection

Michal Vinkler

Charles University

Department of Zoology

e-mail: ***michal.vinkler@natur.cuni.cz***

EMBL-EBI

European Molecular Biology Laboratory (EMBL)
European Bioinformatics Institute (EBI)

http://www.ebi.ac.uk/training/

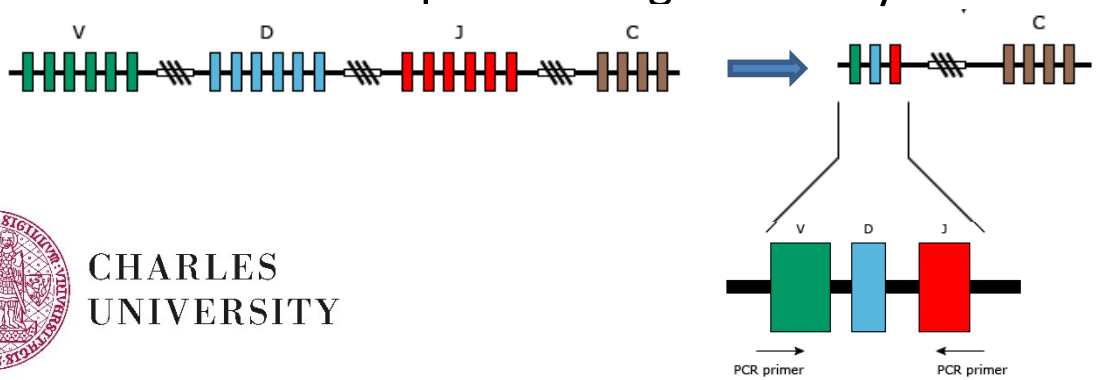
CHARLES UNIVERSITY

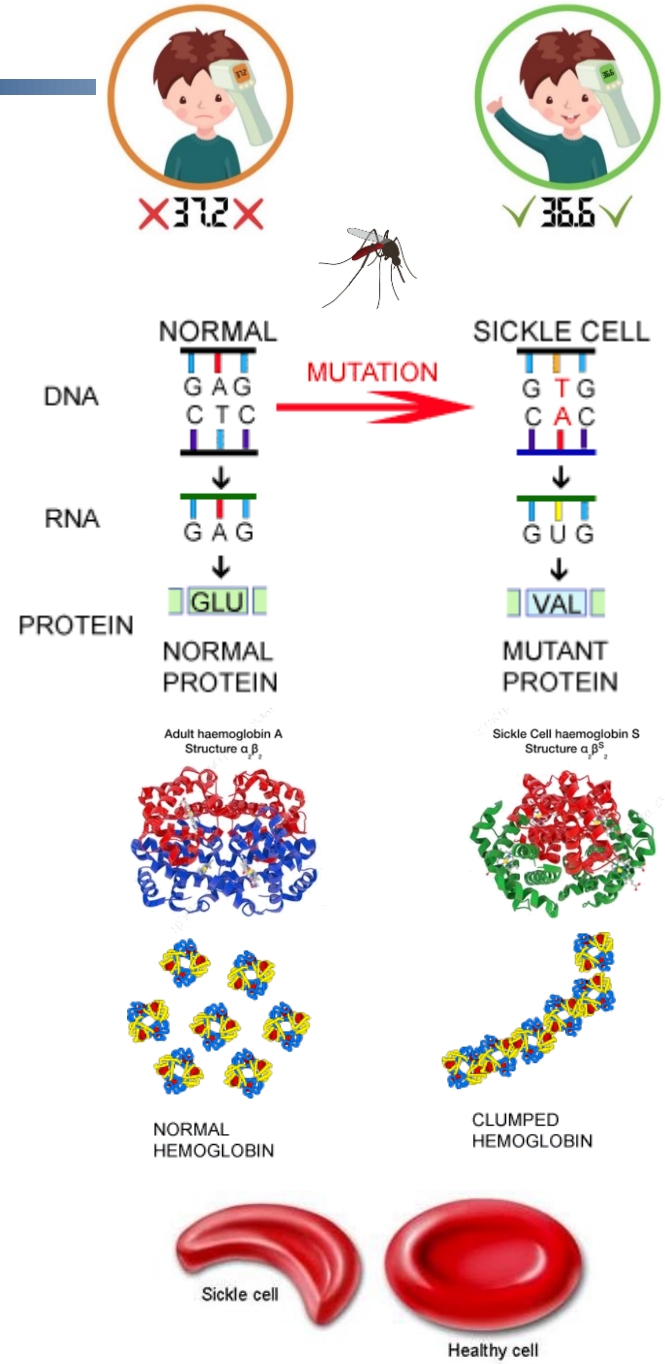
Outline of the lecture

- Brief introduction to genetic variability
- Basic methods of genetic variability detection
- Examples of polymorphism at different types of sites with different effects
- Recombination
- Natural selection – principles and methods of detection in molecular data

Molecular variation

- Phenotypic variability
  - observed traits
  - interaction of genotype & environment
- Genotypic variability
  - germline encoded differences in NA sequences
- Somatic variability
  - somatic mutations
  - germline encoded sequences rearranged in somatic cells
  - increase of pre-existing variability





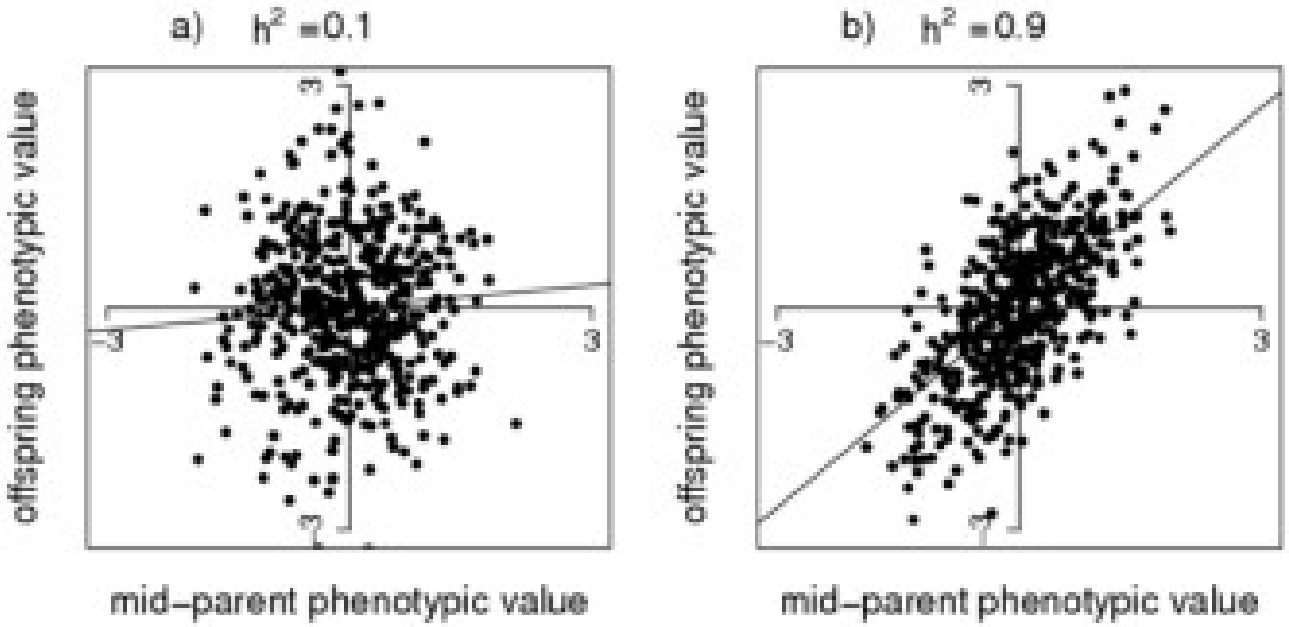
CHARLES UNIVERSITY

Heritability

Heritability

- how much of the variation in a trait is due to variation in genetic factors
- $H^2 = V_G/V_P$ V_P = proportion of phenotypic variation

  V_G = proportion of variation due to genetic factors



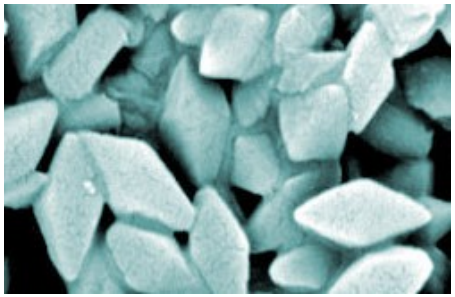
CHARLES UNIVERSITY

Heritability in host-parasite interactions

Slash pine (*Pinus elliotii*) - Fungal rust (*Cronartium quercuum*)

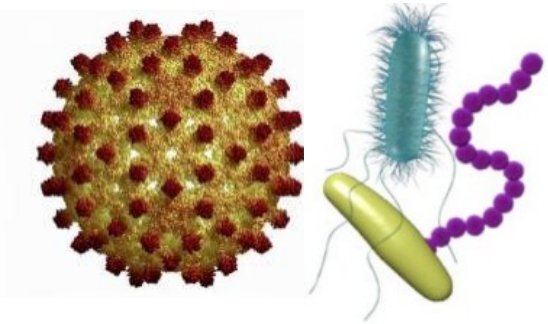
h^2=0.21

Corn borer (*Ostrinia nubilalis*) - bacteria *Bacillus thuringiensis*



h^2=0.31

Human (*Homo sapiens*) - various viral and bacterial diseases

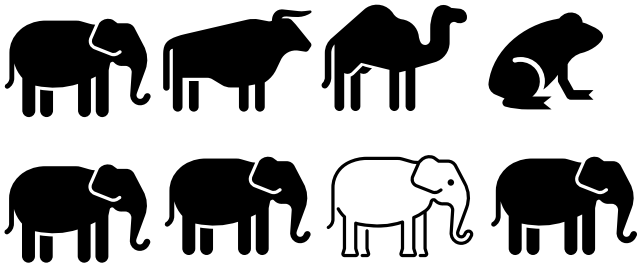


32-65%

CHARLES UNIVERSITY

Molecular variation

- Interspecific vs. Intraspecific
  - Principally the same, difference in gene flow
- Interpopulation vs. Intrapopulation



- Polymorphism

Definition:

- „The condition in which the DNA sequence shows variation between individuals in a population.“ (Patthy 2008)
- Convention: genetic variability with minor allele frequency > 0.01

How many common SNPs diversify living humans?

Polymorphism

Genotypic variability ~ Genetic polymorphism

- Human vs. Chimpanzee – 1.2% divergence
- Human vs. Neanderthal – 0.50% divergence
- Humans vs. Human – variability in 0.10% positions (frequency > 1%)

→All humans are from 99.9% genetically identical




BUT human genome >3‘000‘000‘000 bp

→0.10% > 3‘000‘000 bps are commonly variable in humans

→many more are variable with frequency < 1%

- mostly no phenotypic effect – 3%-5% SNPs functional

→ ca. 100‘000 common functional SNPs

CHARLES UNIVERSITY

Polymorphism

Markers of genetic variability:

- **Single nucleotide polymorphism (SNP) and short indels**
- > 99.9% of variants
- Human genome – ca. 5 million common SNPs (8M over 5%; >80M known)
- Every 6kb on average, linkage between neighbouring loci

- Short tandem repeats (STRs) = microsatellites

- short (usually 2-5 bp) sequences repeated in genome
- highly variable in length
- usually neutral

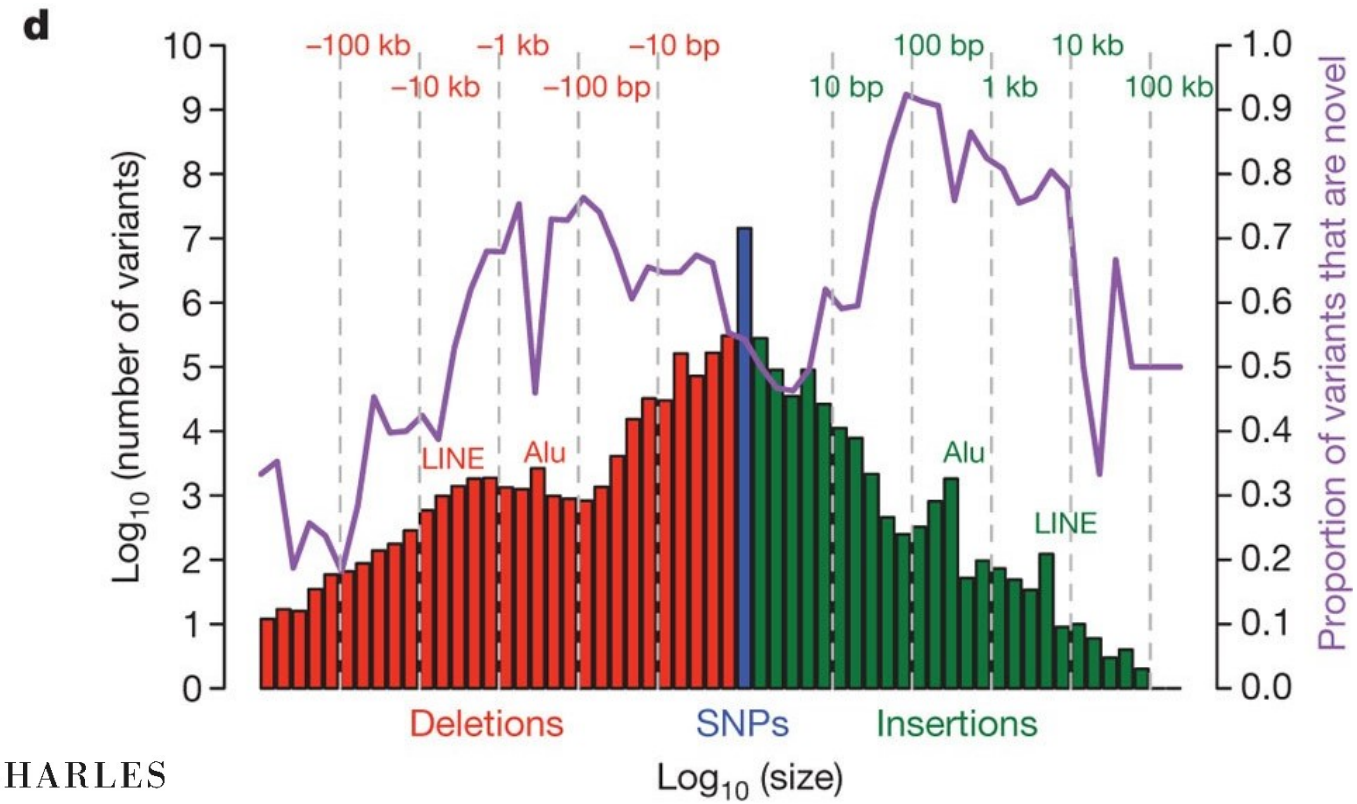
In a typical genome contains ~ 2500 structural variants:

- Insertions (Alu, L1, SVA): ~ 1000
- **Copy number variants (CNVs):** 160; 4.8–9.7% of the human genome
- larger than 50 bp, often longer sequences (1kb-1Mb) and genes
- **Large deletions**: 1 000
- **Inversions**: 10

Polymorphism

Markers of genetic variability:

- **Single nucleotide polymorphism (SNP) and short indels**
- > 99.9% of variants
- Human genome – ca. 5 million common SNPs (8M over 5%; >80M known)
- Every 6kb on average, linkage between neighbouring loci

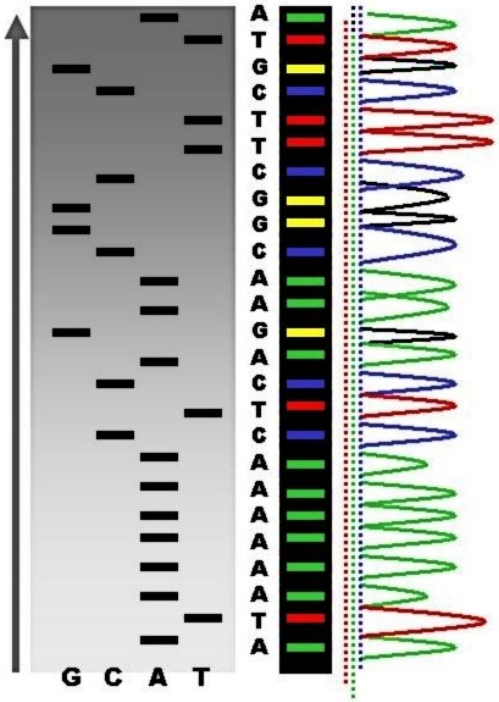


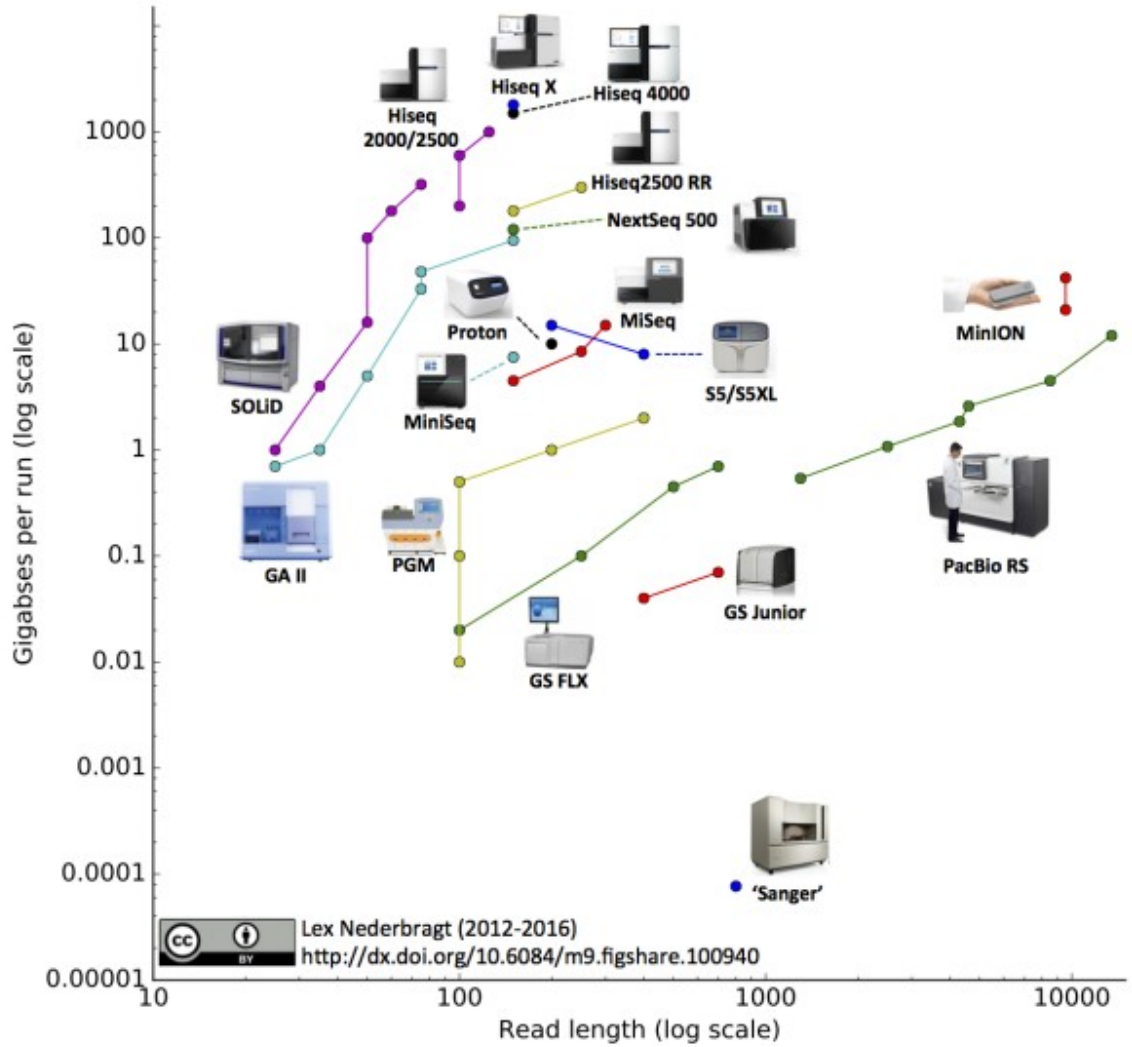
CHARLES UNIVERSITY

Techniques of SNP detection

Basic methods of SNP detection:

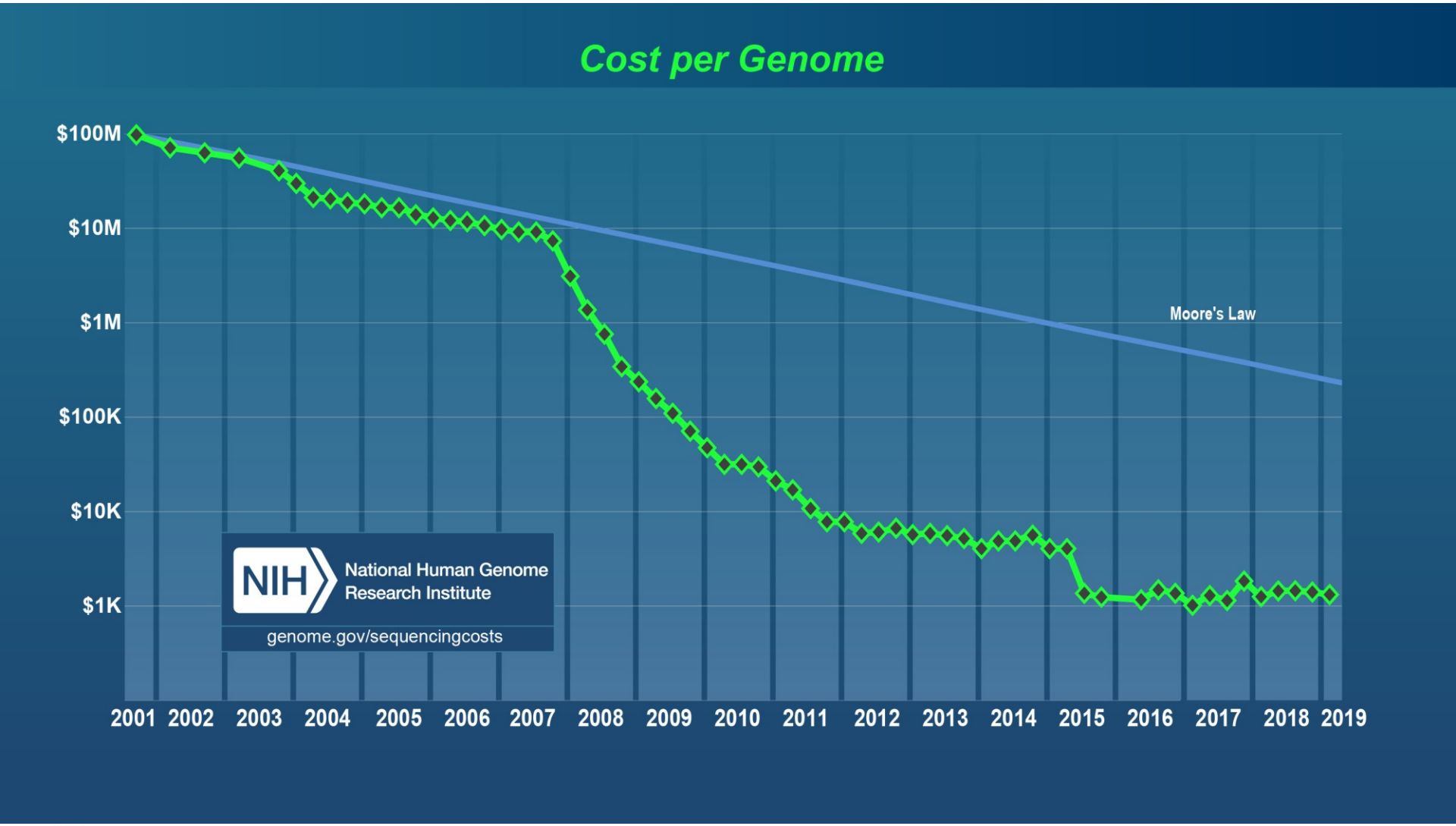
- **Sequencing** – Sanger / 2nd generation / 3rd generation





CHARLES UNIVERSITY

Techniques of SNP detection



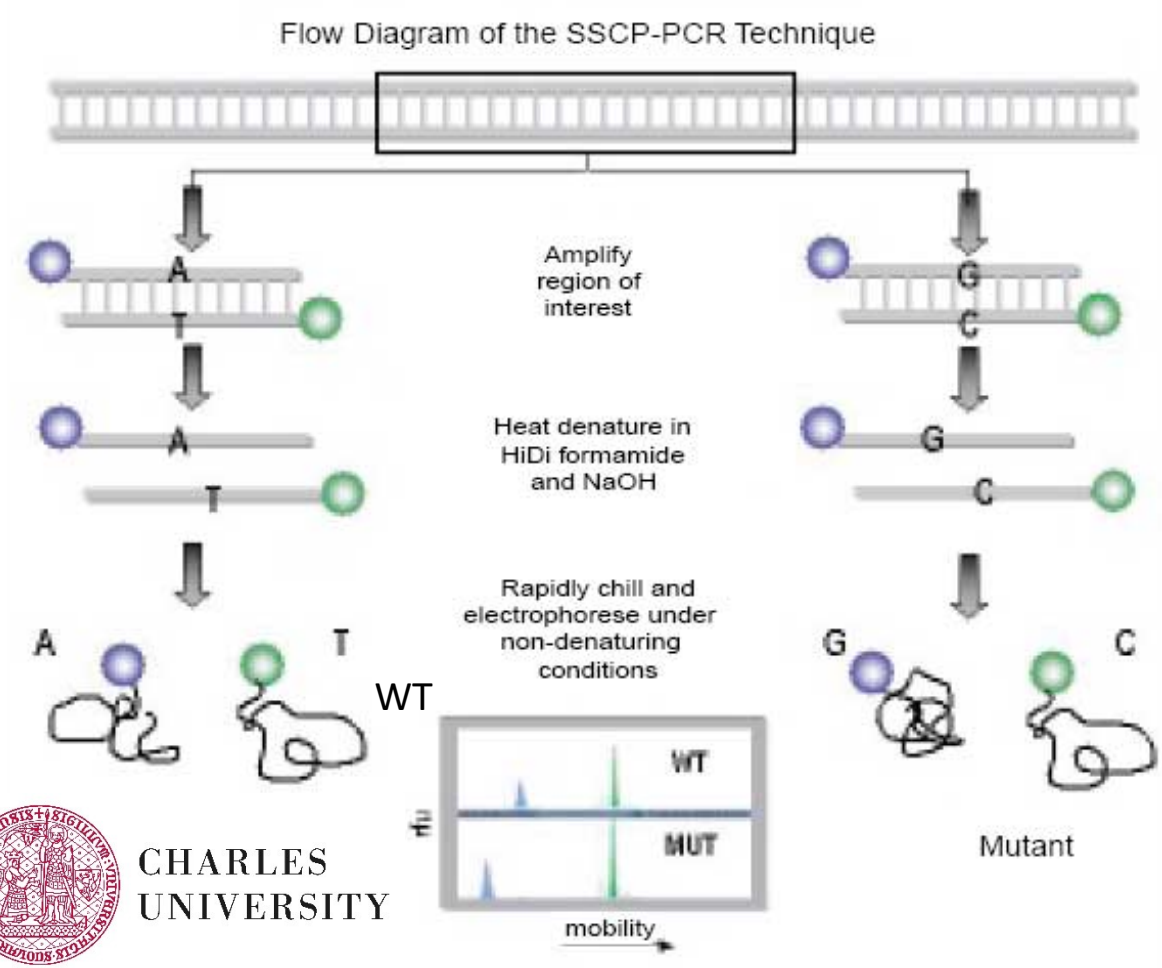
CHARLES UNIVERSITY

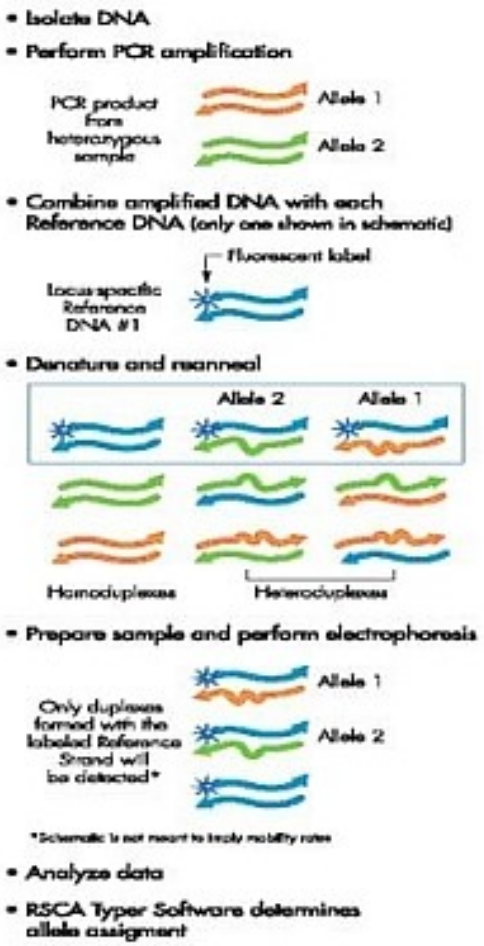
Presetly 1 genome for ~ $1000

Techniques of SNP detection

Basic methods of SNP detection:

- **Single strand conformation polymorphism** (SSCP) and related approaches (e.g. Reference strand conformation analysis, RSCA)





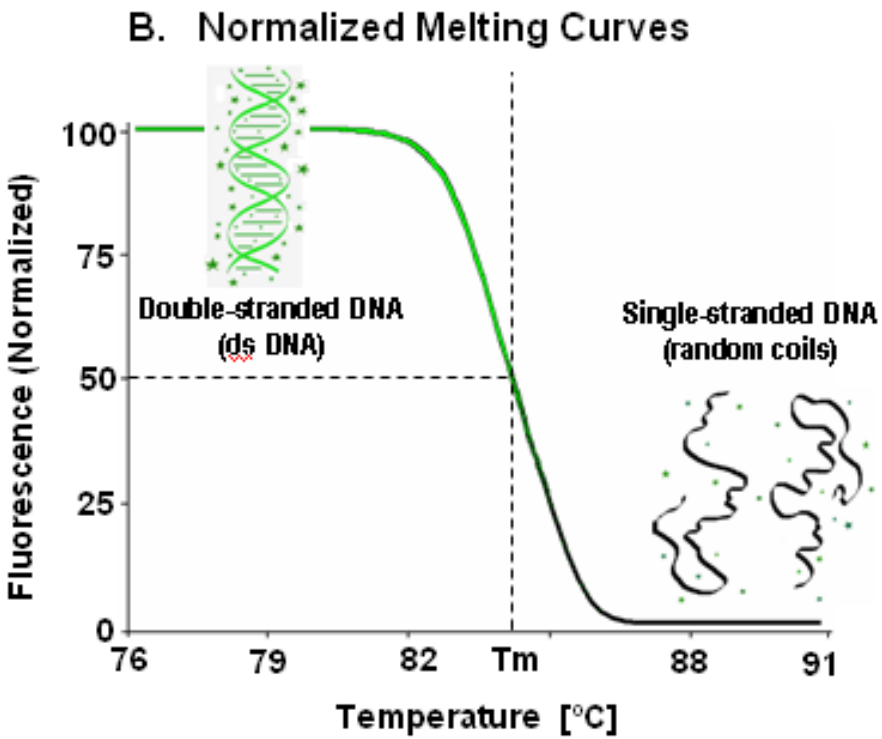
CHARLES UNIVERSITY

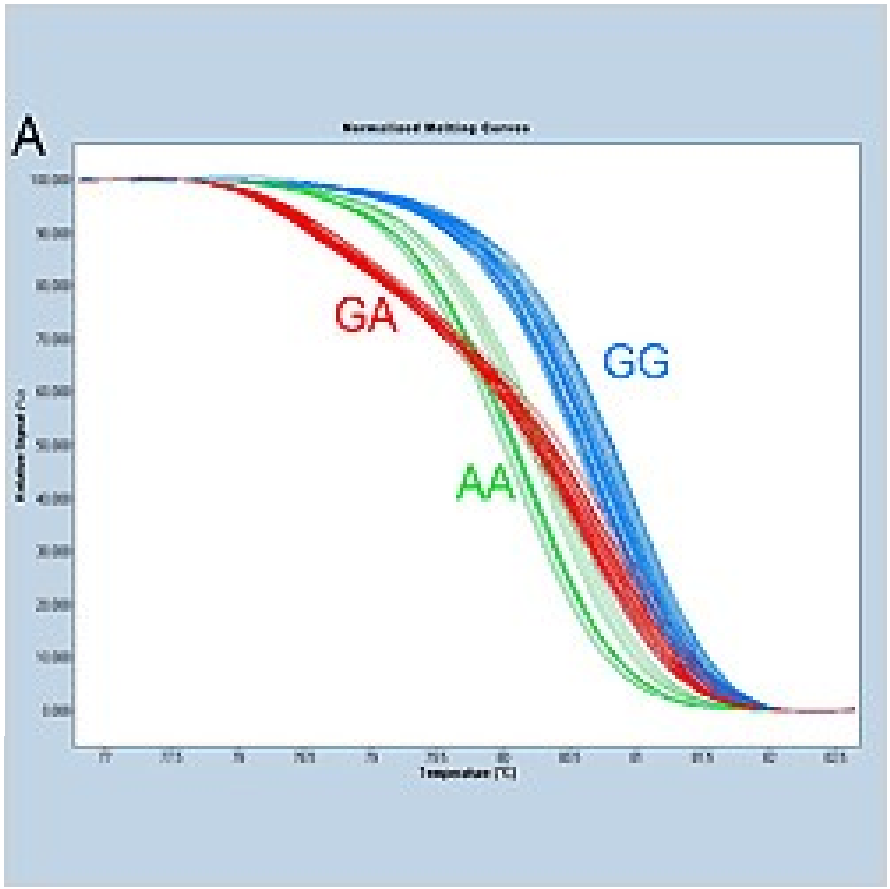
Techniques of SNP detection

Basic methods of SNP detection:

- **High resolution melting (temperature) analysis (HRMA)**

  – PCR →warming from 50˚C up to 95˚C → real-time fluorescent detection of double-stranded DNA



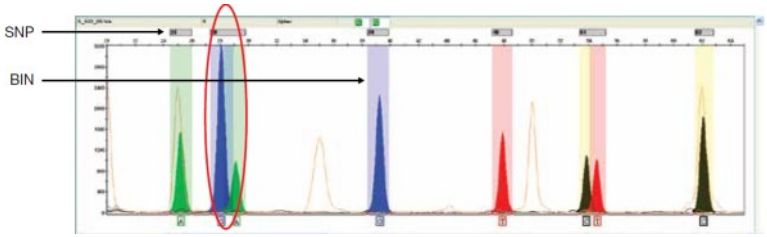


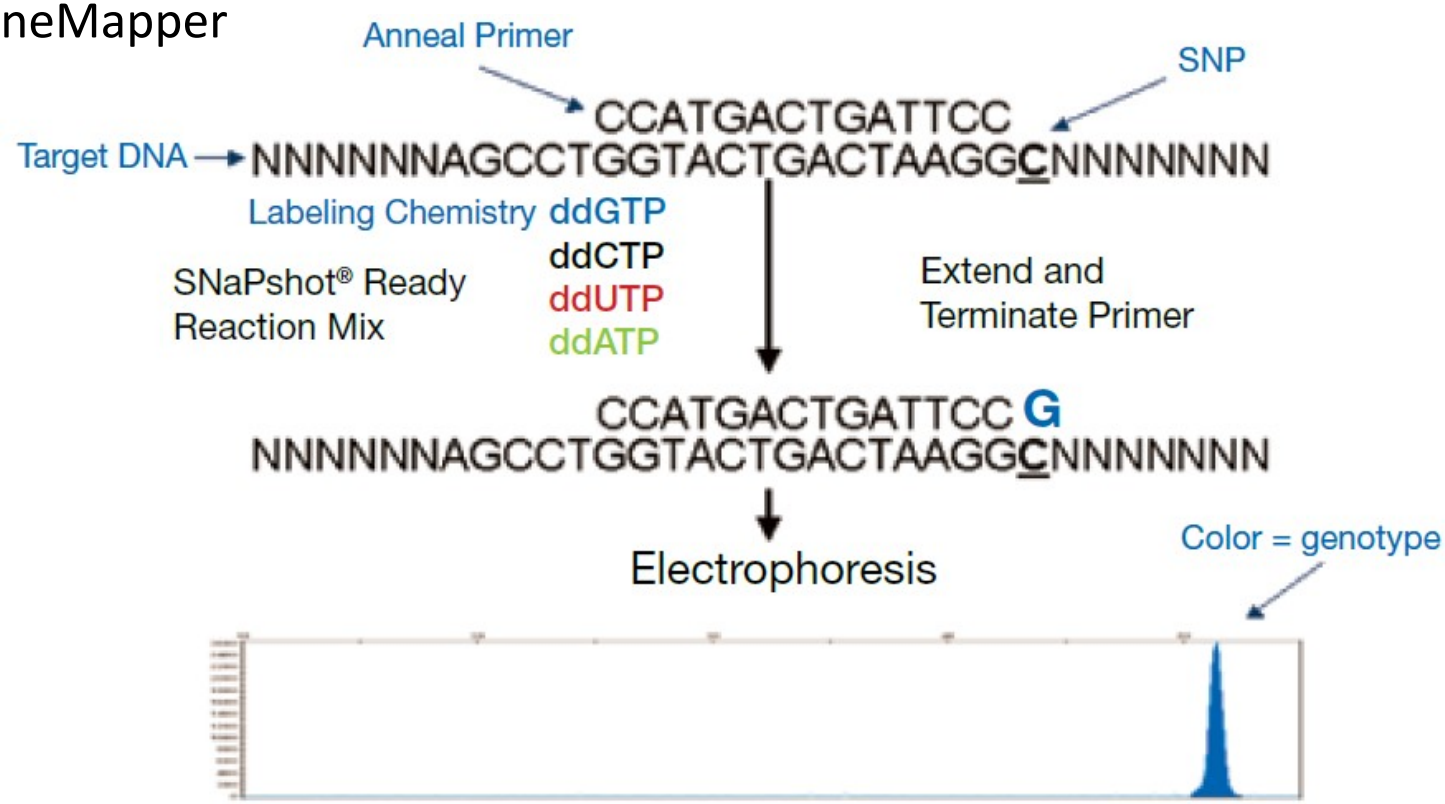
CHARLES UNIVERSITY

Techniques of SNP detection

Basic methods of SNP detection:

- **SNaPshot**
- primer extension–based method
- multiplexing capability (up to 10-plex)
- sensitive allele-frequency detection (typically 5%)
- analysis - GeneMapper

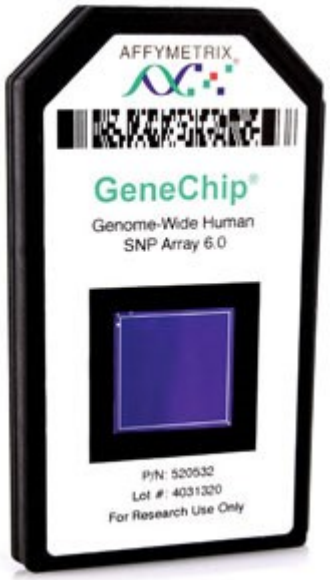


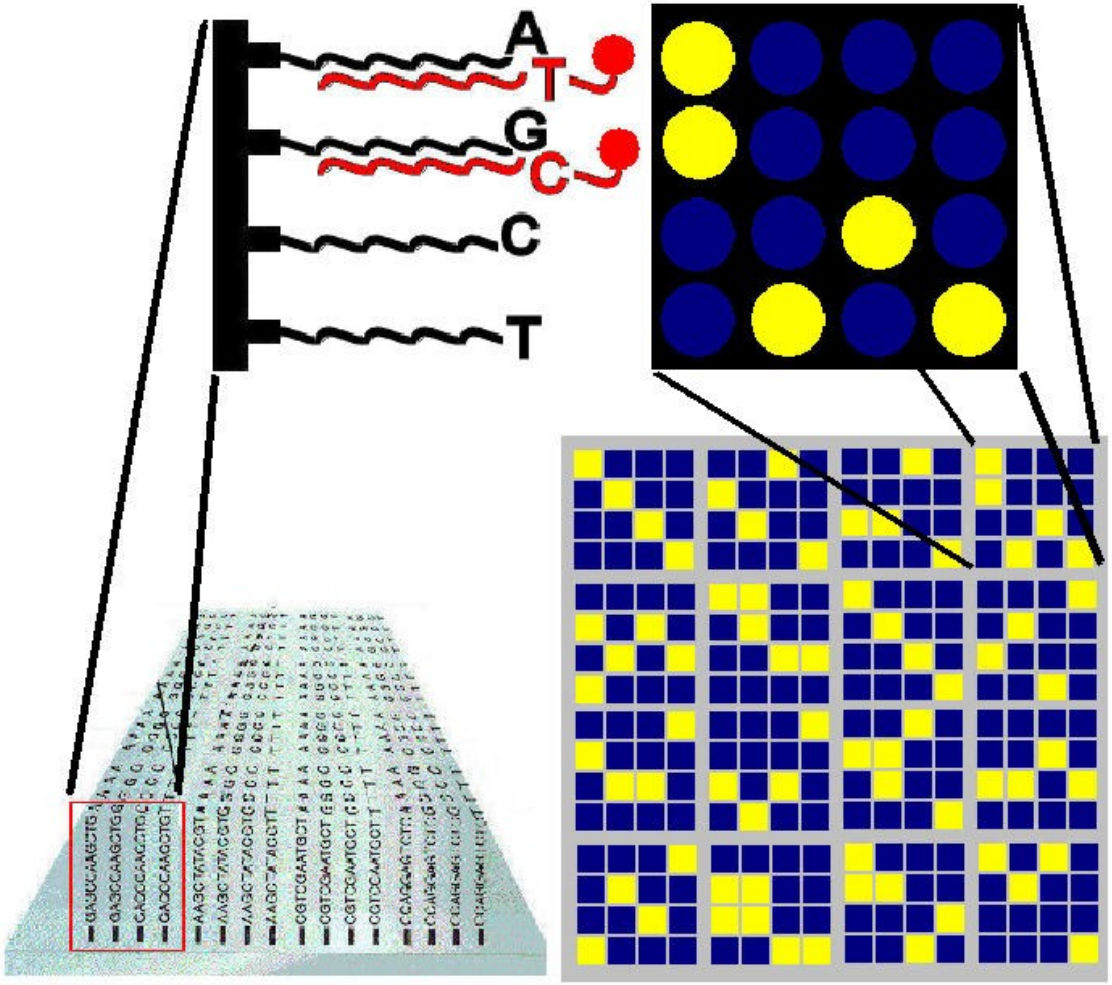


Techniques of SNP detection

Basic methods of SNP detection:

- **SNP microarrays (SNP chip)**





CHARLES UNIVERSITY

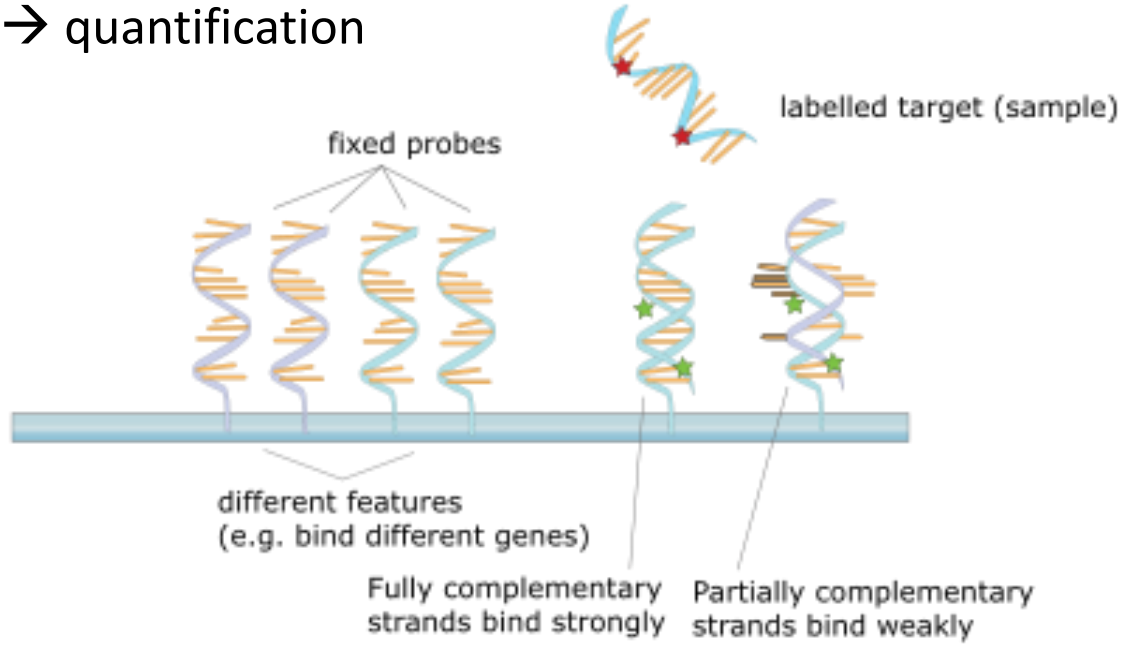
Techniques of SNP detection

Microarrays

- **Probes** (or reporters or oligos) = picomoles (10^{-12} moles) of a specific DNA sequence – synthesised and attached covalently to the chip surface
- tens of thousands of probes per chip

→ hybridize cDNA or cRNA (~ anti-sense RNA) sample (**Target**)

→ detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target → quantification



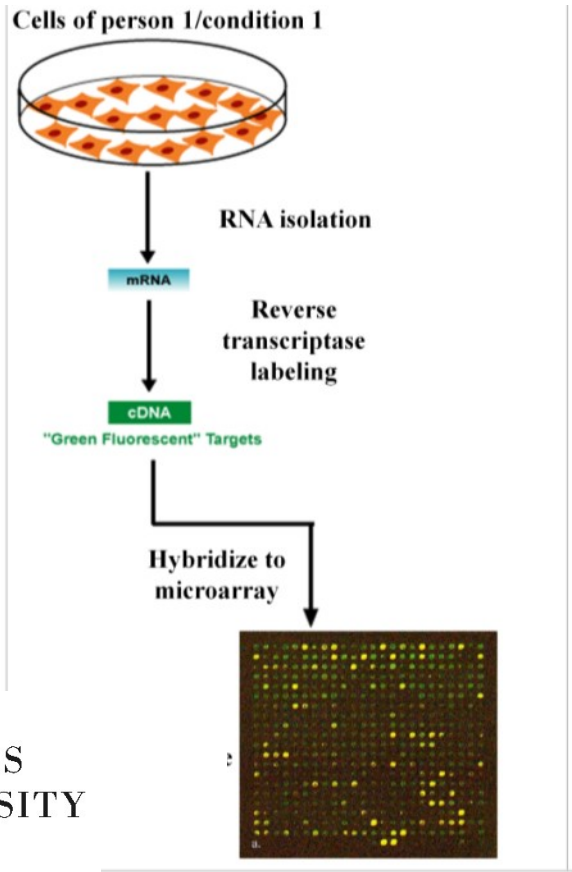
Techniques of SNP detection

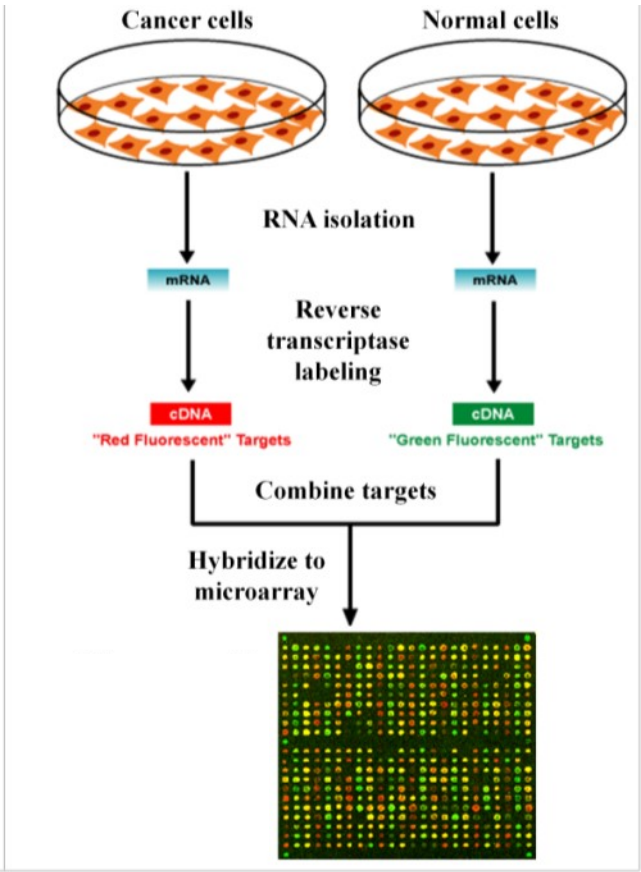
One-channel (colour) detection

- relative abundance when compared to other samples (on the same slide)
- aberrant samples cannot affect raw data

Two-channel (colour) detection

- two different fluorophores:
- Laser → e.g. Cy3 (570 nm = orange) and Cy5 (emission of 670 nm = red)
- control probes → normalization





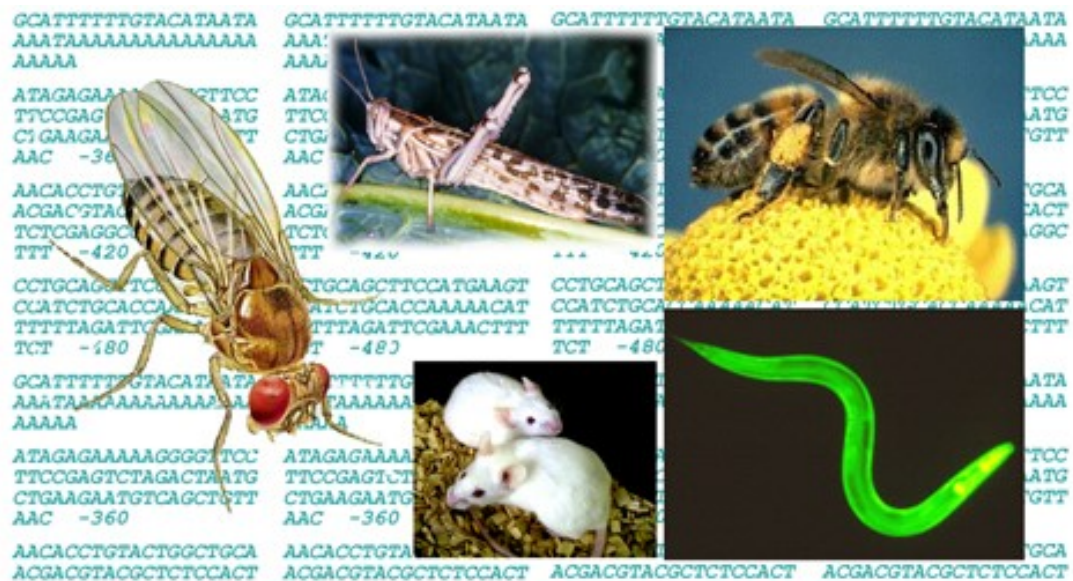
CHARLES UNIVERSITY

Techniques of SNP detection

Usage

- **Gene expression microarrays** – control vs. treatment
- **SNP microarray (SNP chip)** – allele A vs. allele B
- **Comparative genomic hybridization**
- **Alternative splicing (Exon arrays)**

Applicable only to species with complete genome (model species)



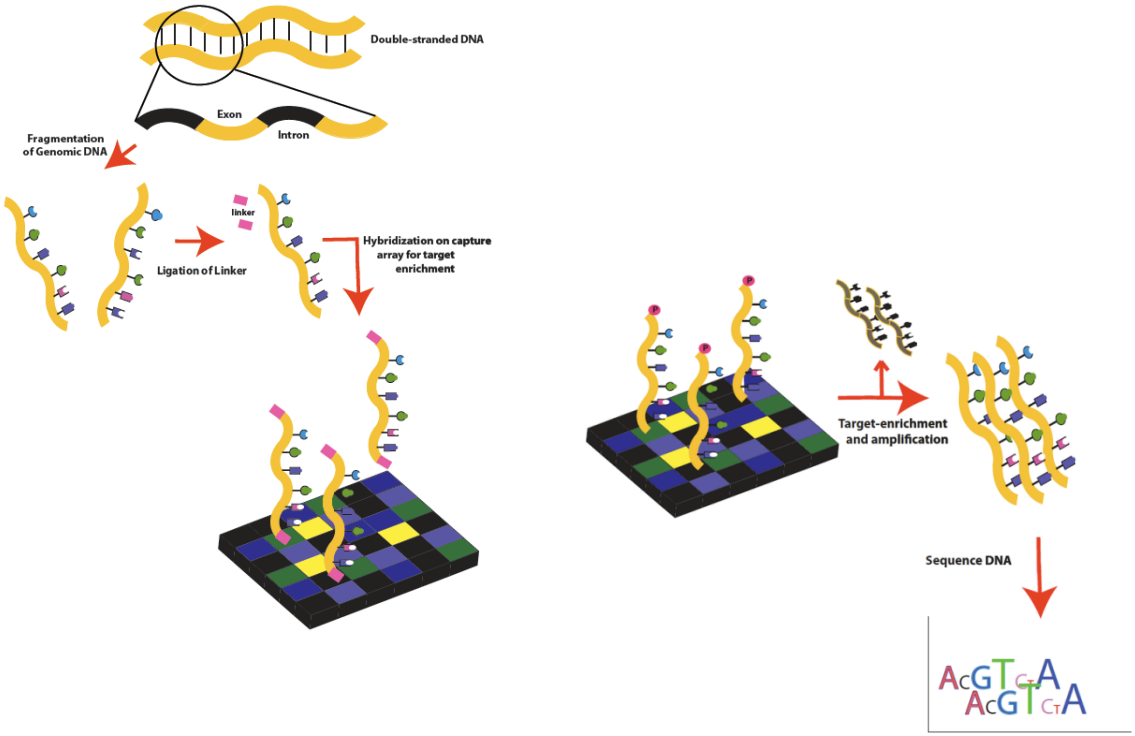
CHARLES UNIVERSITY

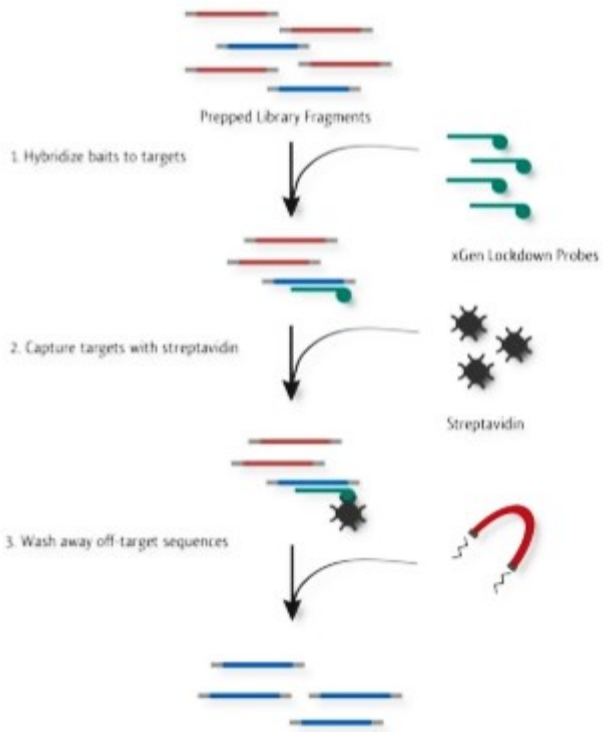
Techniques of SNP detection

Exome sequencing

Target-enrichment strategies:

- **Array-based capture**





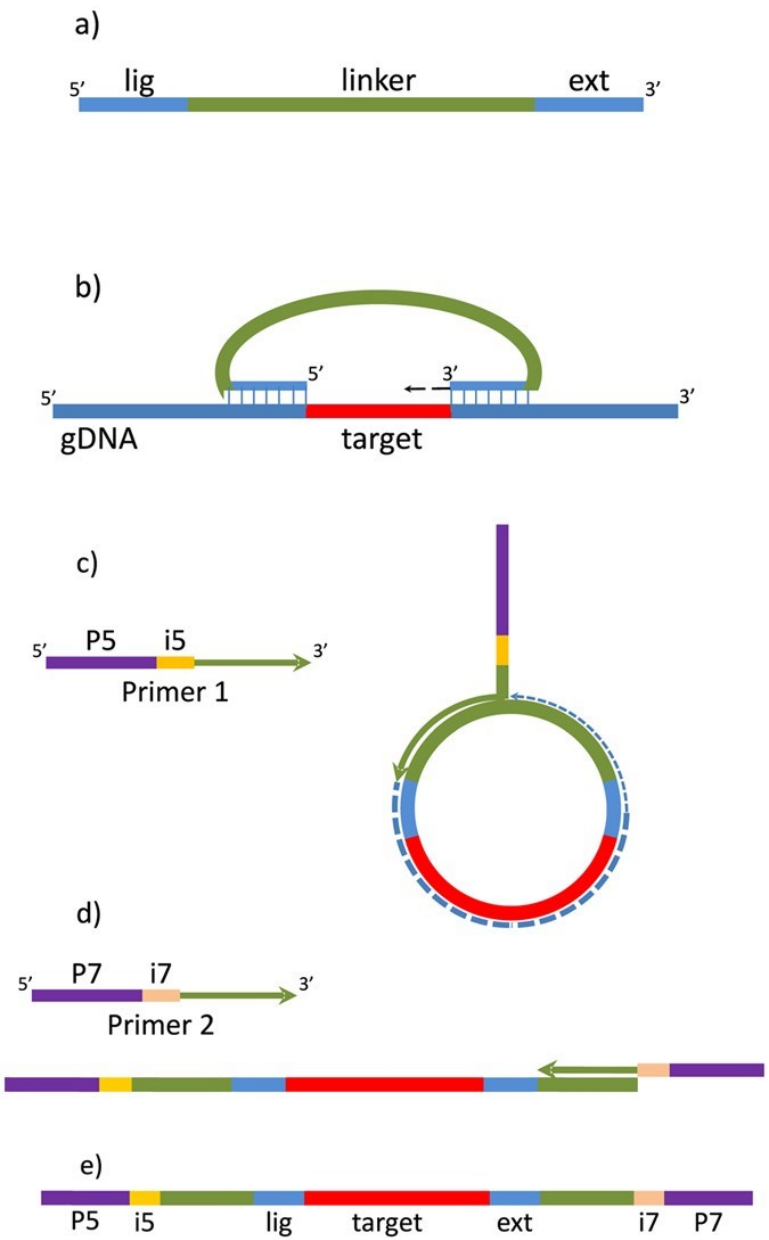
- **In-solution capture**

Techniques of SNP detection

Molecular Inversion Probes (MIP)

Target-enrichment

→ **Resequencing (NGS)**

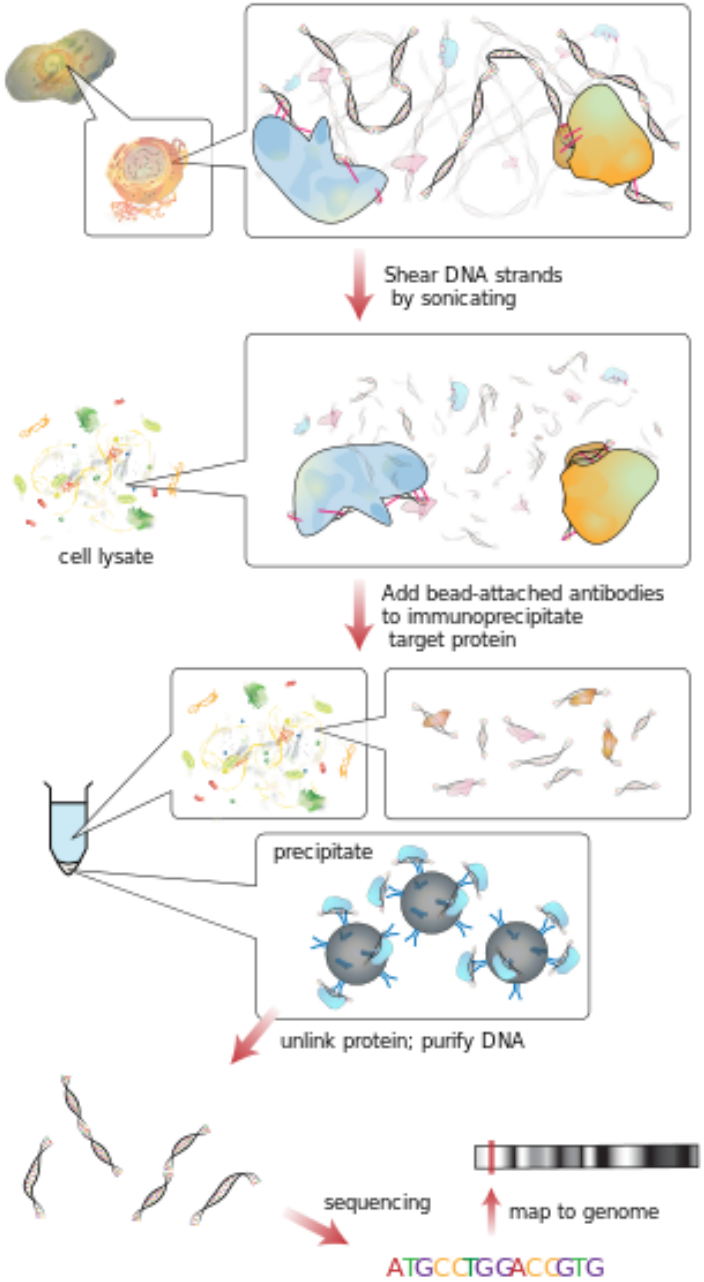


Niedzicka et al. 2016

Techniques of SNP detection

ChIP-seq

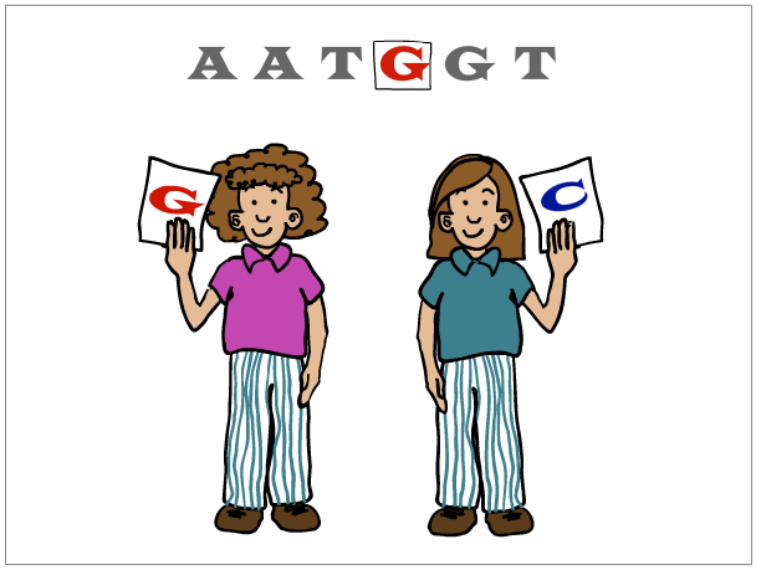
- determines how transcription factors and other chromatin-associated proteins influence phenotype-affecting mechanisms
- combines chromatin immunoprecipitation (ChIP – antibodies attached to beads) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins



Well, and what now?

```
                    10        20        30        40        50        60        70        80        90
GaGaTLR4-AY0 ATGCCCAGCAGAGCGGCTCCCACCGCGCTGACTCTTGGGGTGCTGCTGCAGCTGCTGCTCGTGCTGTCCCTGCTGGCAGGATGCATCCCCAGTC
GaLaTLR4-FJ9 ........................................................A.......................................
GaSoTLR4-FJ9 ........................................................................................................
GaVaTLR4-FJ9 ....T...................................................................T...........................
PePeTLR4-JQ7 ......................TT..A.A...G..C............G...........C...............................
MeGaTLR4-XM  ......................TT..A.A...G..CA...........G...........................................
AnPlTLR4-JN0 ........G.....A....T.CT.C..TG...GT.C.T..G.............G.....A.C....TG.A........C...C...T.....
AnAnTLR4-HQ4 ........G.....A....T.CT.C..TG...GT.C....A........G....GC....A.C.....G.A........TC...C...T.....
```

???



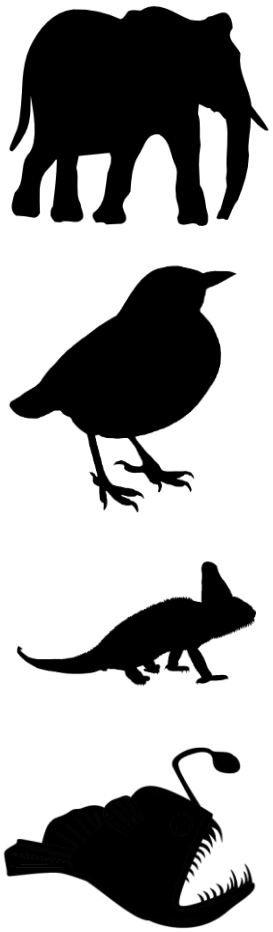
Which questions may I now ask?

CHARLES UNIVERSITY

Well, and what now?

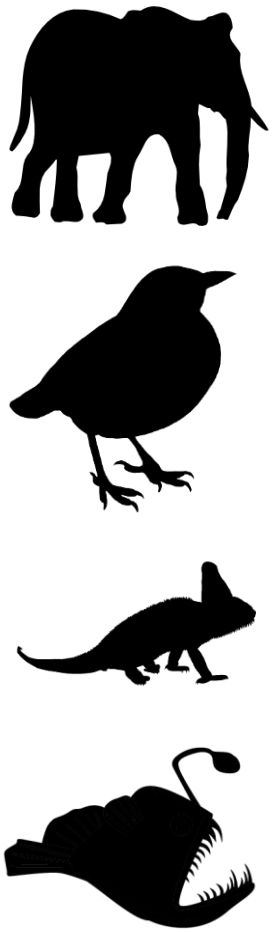
CHARLES UNIVERSITY

Well, and what now?



CHARLES UNIVERSITY

Well, and what now?

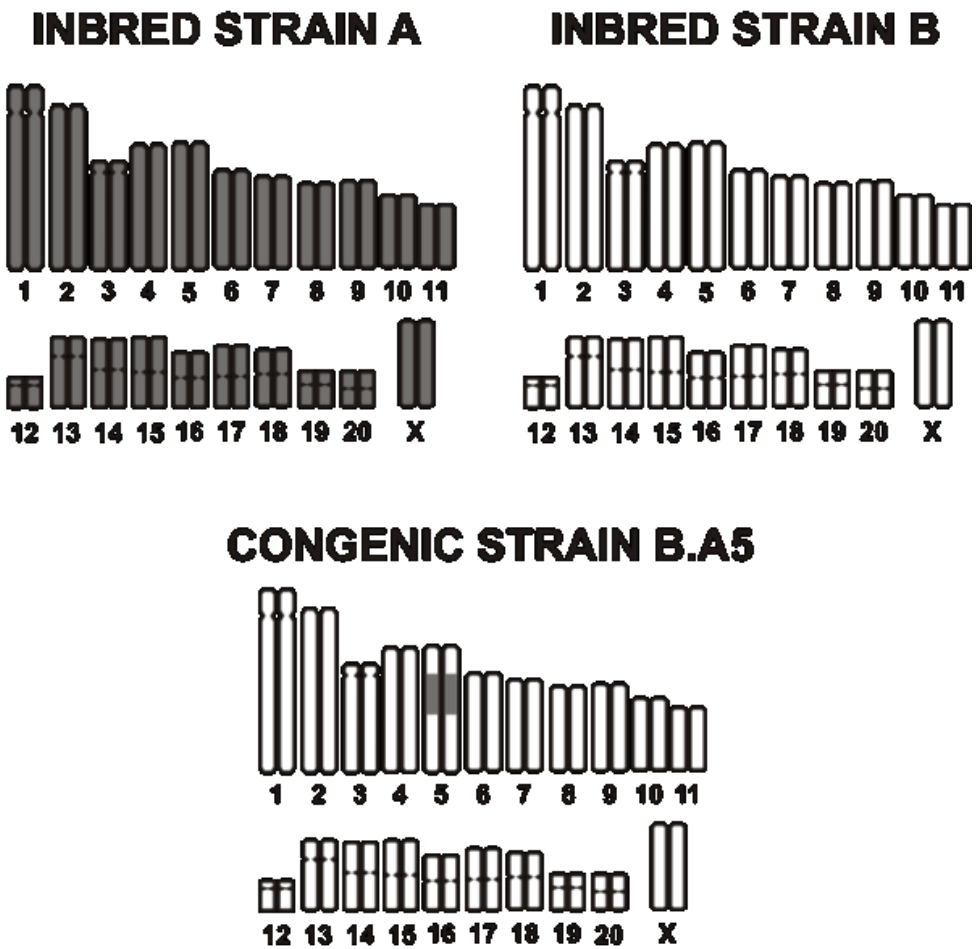


CHARLES UNIVERSITY

Quantitative trait loci (QTLs)

Congenic animals and strains

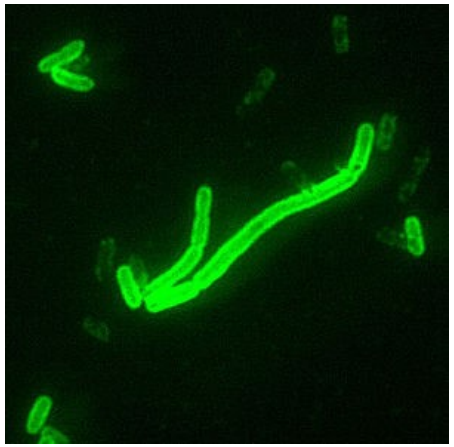
- animals / strains in which a specific and defined part of the genome from one inbred strain (strain A) is introgressed on the genetic background of second inbred strain (strain B)

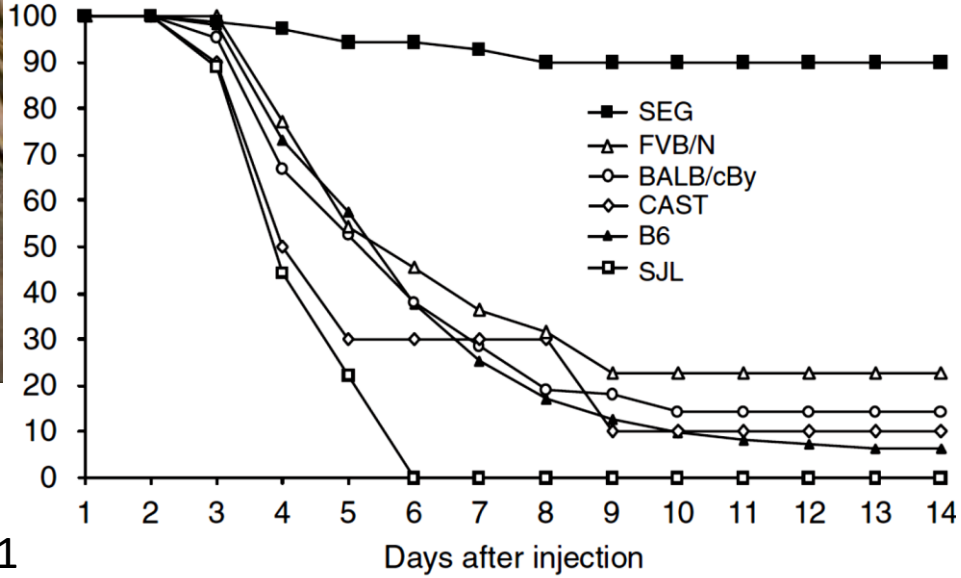


Quantitative trait loci (QTLs)

Quantitative traits - polygenic effects on phenotype

- QTL mapping – effect of SNPs
- *Mus musculus* (e.g. C57BL/6)
  - susceptible to *Y. pestis* (10^2 CFU →<8% survival)
- *Mus spretus* (SEG/Pas)
  - resistant to *Y. pestis* (10^2 CFU → >70% survival)





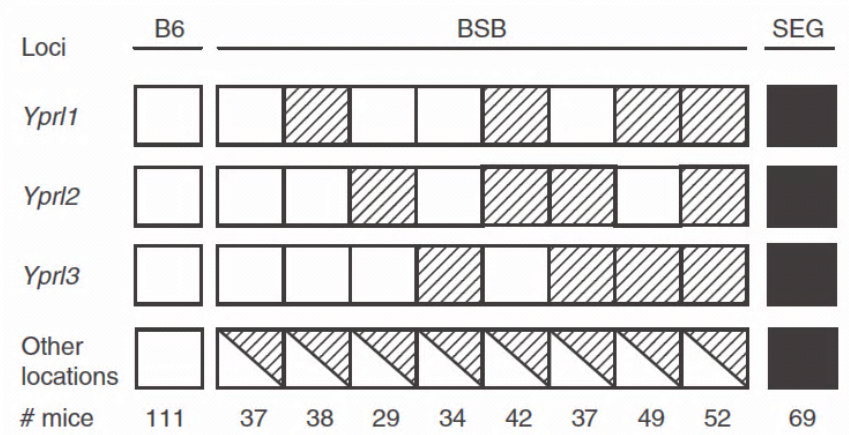
CHARLES UNIVERSITY

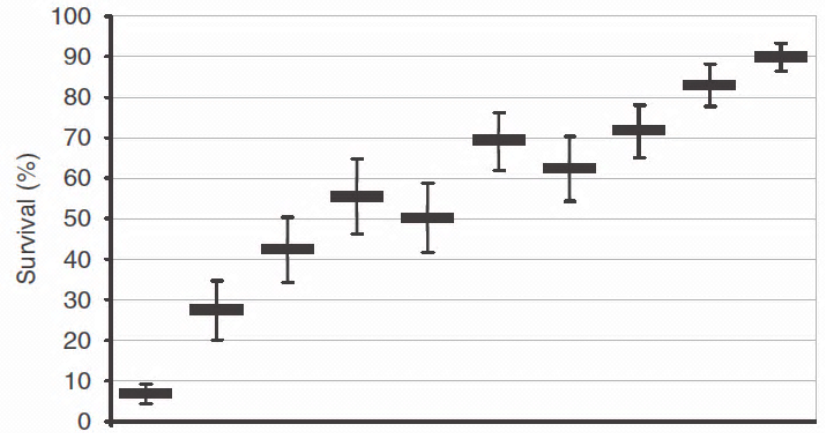
Blanchet et al. 2011

Quantitative trait loci (QTLs)

QTL mapping in mice using 322 backcrosses

- B6xSEG F1 females + B6 males → F2 progeny
  - 721 polymorphic markers covering the entire genome
- quantitative trait loci (QTLs) on chromosomes 3, 4 and 6, with dominant SEG protective alleles:
  - *Y. pestis* resistance loci (Yprl1-3)
  - each QTL contributes with ~20%
    → 67% in total
  - large chromosomal segments (between 50 and 84 Mb)
- Candidate genes:
  - Yprl1: *Pglyrp3, Pglyrp4* (AMPs), *IL-6*$^{\alpha}$
  - Yprl2: *Tlr4*, *IFNs*
  - Yprl3: *Nod1, IL-17R subunits, IL-23R*



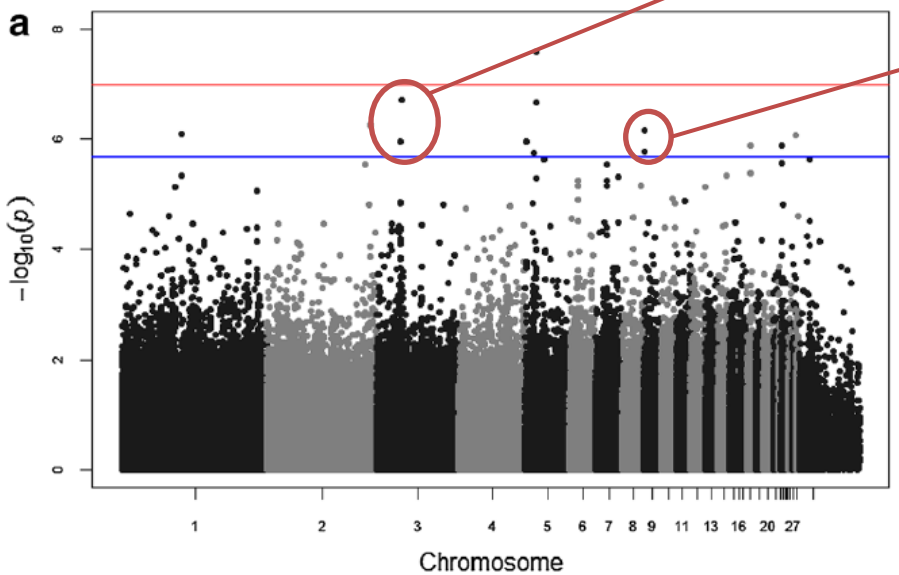


CHARLES UNIVERSITY

Blanchet et al. 2011

Genome-wide association studies

Indigenous Ethiopian chicken ecotype Horro



Infectious bursal disease (IBDV) antibody titres

Affx-51084536 – missense variant within the *XK-related protein 8* (*XKR8*) gene

Affx-51878048 - region for IBDV response on chromosome 9: two putative candidate genes -
- *protein tyrosine phosphatase nonreceptor type 1* (*PTPN1*)
- *nuclear factor of activated T-cells cytoplasmic calcineurin-dependent 2* (*NFATC2*)

Most significant SNPs were located in intergenic or intronic regions!

Psifidi et al. 2016

Table 3 Significant SNPs identified for traits in Horro chickens

Trait	SNP	Location Chr (bp)	GWAS P-value	Additive effect (P-value)	Dominance effect (P-value)	Phenotypic variance (%)	p	q
IBDV	*Affx-5Affx-51526157*[*,a]*	*5 (15315358)*	*2.55E−08*	*0.033 (0.05)*	*0.035 (0.09)*	*2*	*0.03**	*0.97*
	AfAffx-51242536[*,a]*	*3 (3148207)*	*1.96E−07*	*0.033 (0.01)*	*−0.014 (0.14)*	*10*	*0.12**	*0.88*
	Affx-50862142[a,b]	*2 (139341263)*	*5.47E−07*	*0.065 (8E−05)*	*−0.041 (0.02)*	*21*	*0.07**	*0.93*
	Affx-51878048[a,b]	*9 (866678)*	*1.68E−06*	*0.0270.04)*	*0.034 (0.05)*	*2*	*0.07**	*0.93*
	Affx-51183095[a,b]	*28 (581149)*	*8.47E−07*	*−0.025 (0.04)*	*0.117 (0.01)*	*2*	*0.03*	*0.97**
	Affx-50756295[b]	*18 (5404597)*	*1.25E−06*	*−0.003 (0.37)*	*0.032 (0.00)*	*7*	*0.13*	*0.87**
	Affx-51884018[a]	*Z (15058127)*	*2.31E−06*	*0.043 (6E−0.4)*	*−0.025 (0.07)*	*12*	*0.08**	*0.92*
	Affx-51084536[a,b]	23 (1467133)	2.72E−06	0.072 (0.002)	−0.048 (0.05)	18	0.04*	0.96
	Affx-50584797[a,b]	12 (19824359)	3.88E−06	0.025 (0.027)	0.000 (0.39)	4	0.09*	0.91

CHARLES UNIVERSITY

Polymorphism in coding regions

What applies to non-synonymous substitutions?

Where is the detected polymorphism localised?

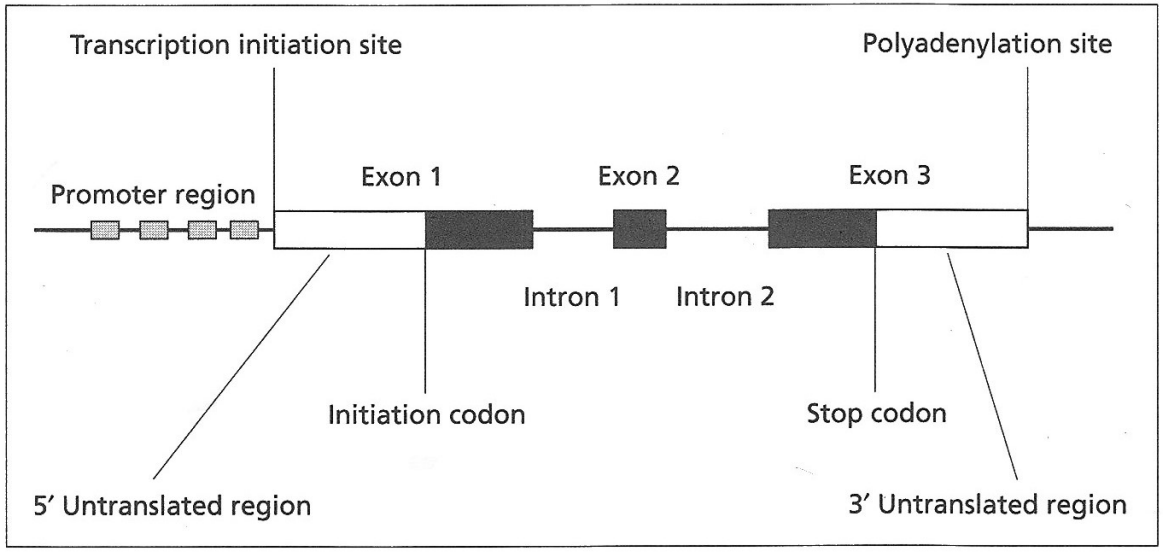
Polymorphism types:

- Single nucleotide polymorphism (SNPs)
- Indels – insertions & deletions
- Rearrangements

Look at position:

- Non-coding – more common
- Coding – synonymous (silent) vs. non-synonymous (missense & nonsense)

  ~ 1:1
- Regulatory regions



Patthy 2008

CHARLES UNIVERSITY

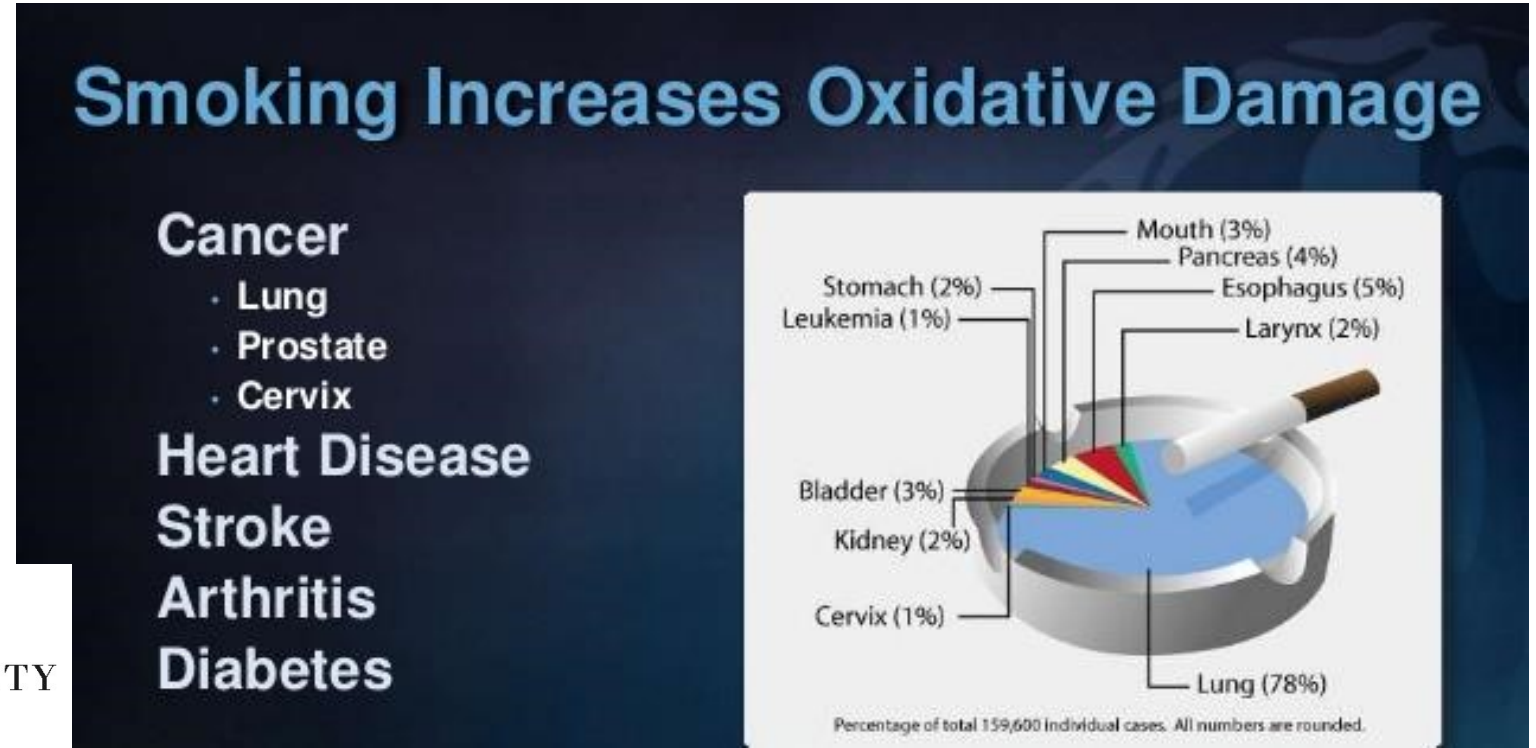
Polymorphism in regulatory regions

Promoter sequence

- combination of SNP database (NCBI) and expression microarrays

Human antioxidant response elements (AREs)

- cis-acting enhancer sequences found in the promoter regions of many genes that encode antioxidant and Phase II detoxification enzymes/proteins



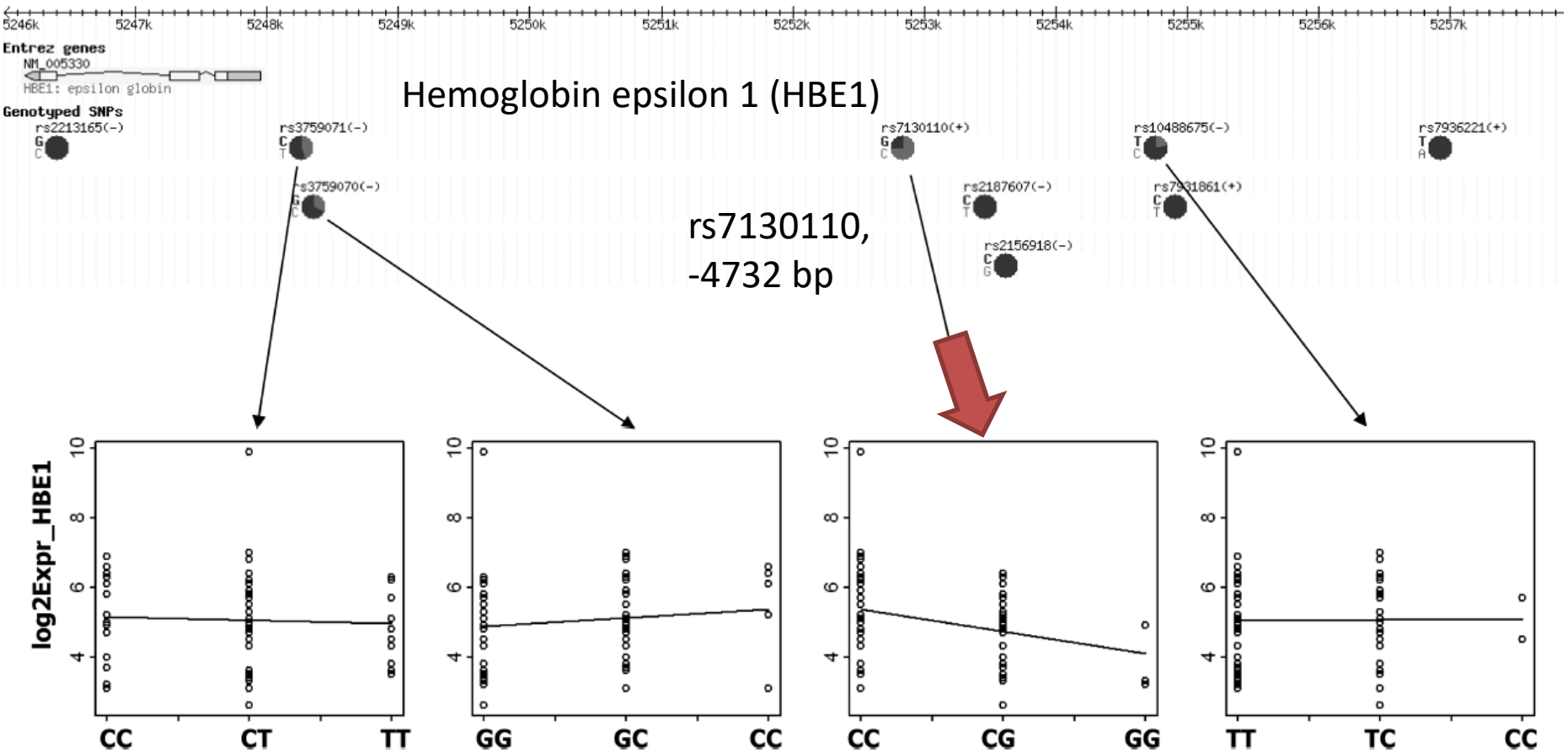
CHARLES UNIVERSITY

Polymorphism in regulatory regions

Promoter sequence

- combination of SNP database (NCBI) and expression microarrays

Human antioxidant response elements (AREs)



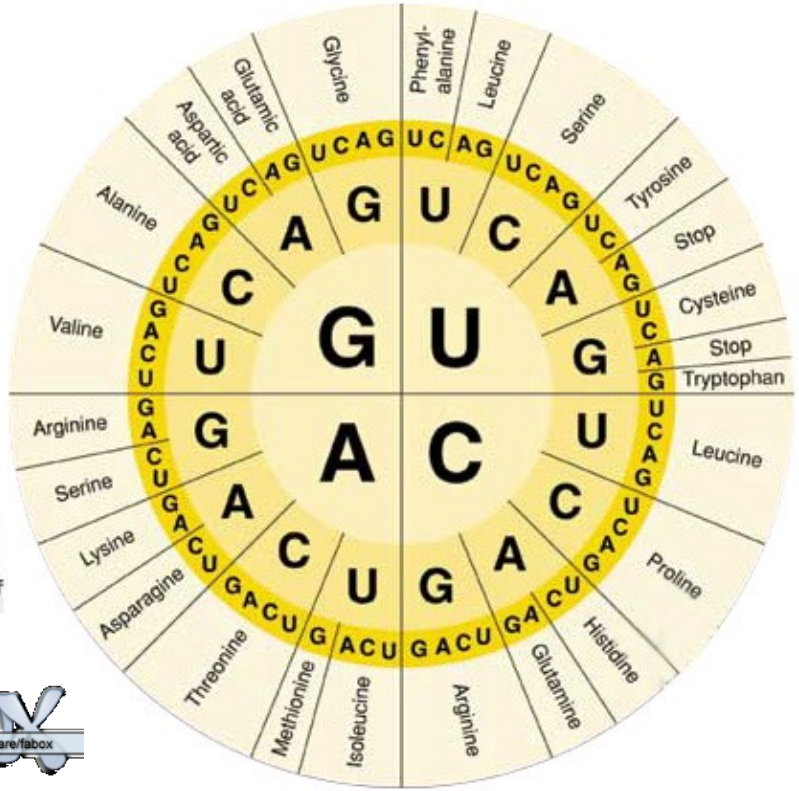
Polymorphism in coding regions

Synonymous vs. non-synonymous substitutions:

- **Translate to amino acid sequence**

```
                    10        20        30        40        50        60        70        80        90
GaGaTLR4-AY0 ATGCCCAGCAGAGCGGCTCCCACCGCGCTGACTCTTGGGGTGCTGCTGCAGCTGCTGCTCGTGCTGTCCCTGCTGGCAGGATGCATCCCCAGTC
GaLaTLR4-FJ9 ..................................................A.....................................
GaSoTLR4-FJ9 ......................................................................................
GaVaTLR4-FJ9 ....T.................................................................T.................
PePeTLR4-JQ7 ......................TT..A.A...G..C............G...........C.........................
MeGaTLR4-XM  ......................TT..A.A...G..CA...........G.........................................
AnPlTLR4-JN0 ........G.....A....T.CT.C..TG...GT.C.T..G............G.....A.C....TG.A.........C...C...T.....
AnAnTLR4-HQ4 ........G.....A....T.CT.C..TG...GT.C....A........G....GC....A.C.....G.A........TC...C...T.....

                    10        20        30
GaGaTLR4-AY0 MPSRAAPTALTLGVLLQLLLVLSLLAGCIPSP
GaLaTLR4-FJ9 ..............................
GaSoTLR4-FJ9 ..............................
GaVaTLR4-FJ9 .L............................
PePeTLR4-JQ7 .......I.Q......R...A..........
MeGaTLR4-XM  .......I.Q..R...R..............
AnPlTLR4-JN0 ..R...LLPW.FVG....V.I.LQ....LL..
AnAnTLR4-HQ4 ..R...LLPW.F.E..R.A.I..Q..V.LL..
```



Software:

- e.g. *Genious*, *BioEdit, web tools*: *ExPASy*
- show variable sites (*FaBox*) → count

geneious

ExPASy SIB Swiss Institute of Bioinformatics

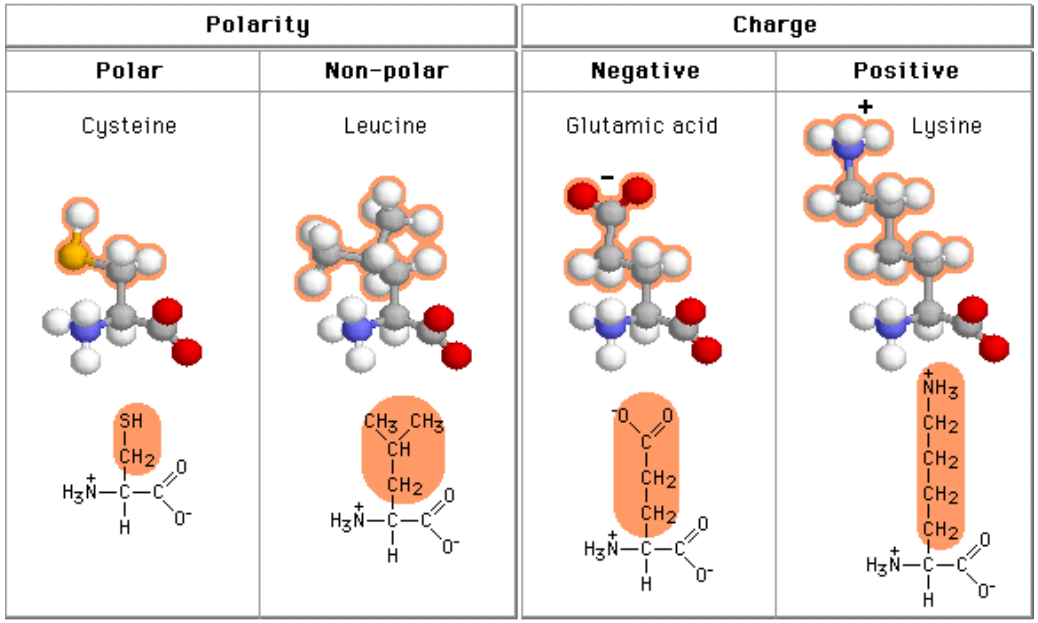
CHARLES UNIVERSITY

FaBox HTTP://www.birc.au.dk/software/fabox

Polymorphism in coding regions

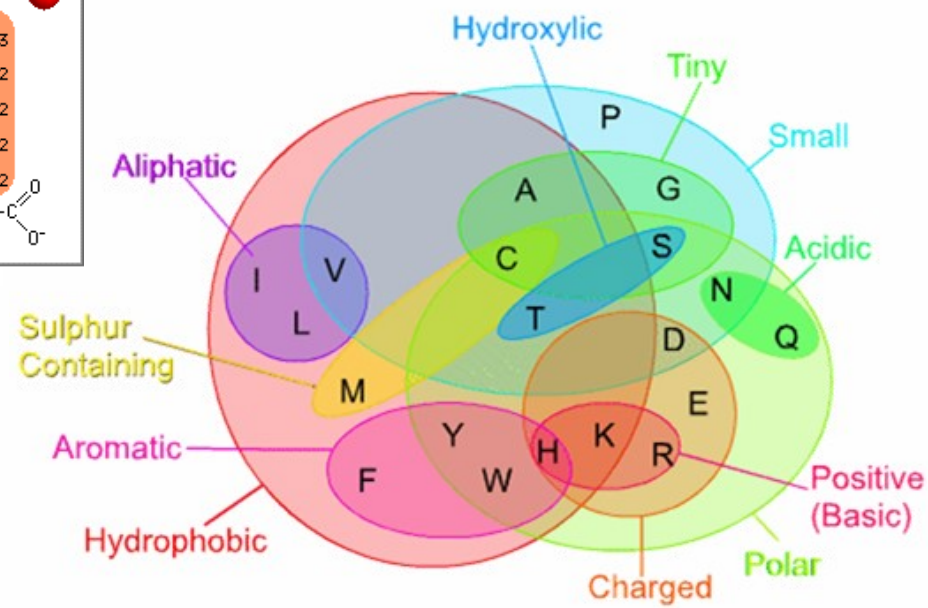
In coding regions may influence protein structure:

- **Electrostatic forces**
  - charges – within protein, surface electrostatic potential



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

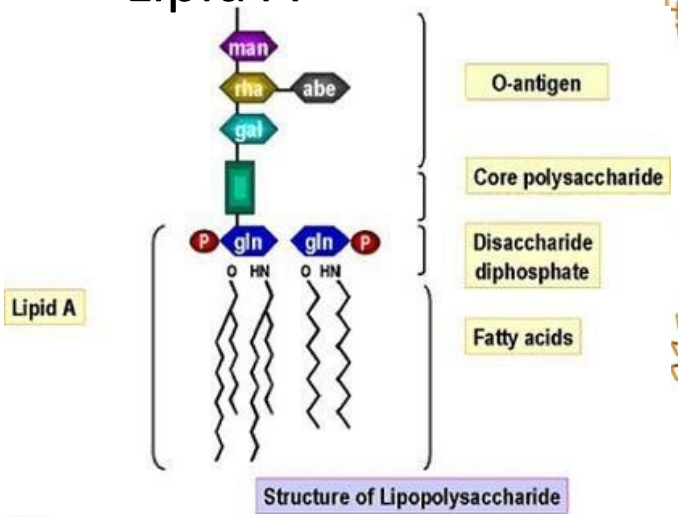


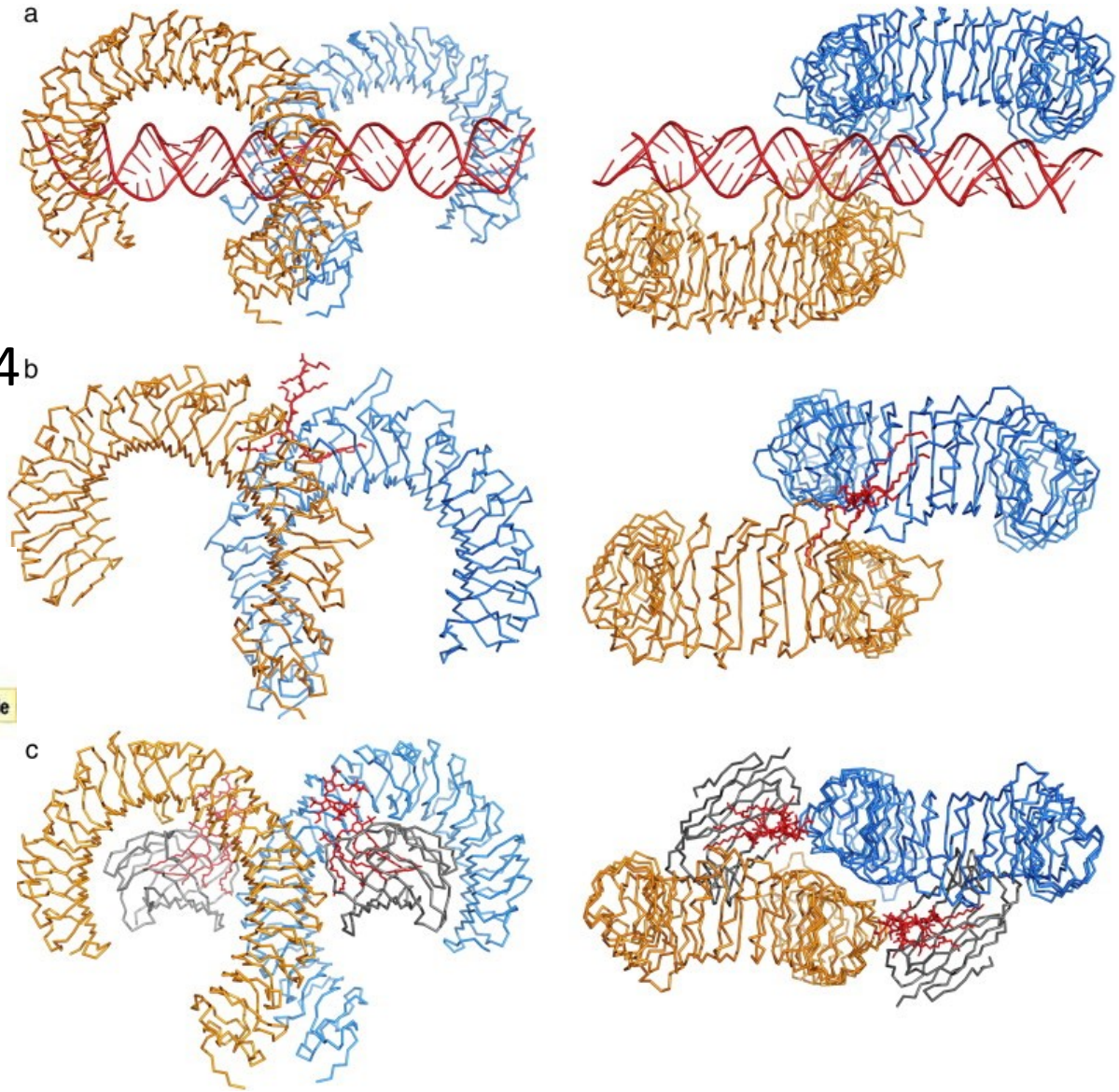
CHARLES UNIVERSITY

Polymorphism in coding regions

TLRs-ligand:

- Dimerisation
  - Homodimerisation
  - Heterodimerisation
- G- LPS → MD-2/TLR4
- Bound based on Lipid A

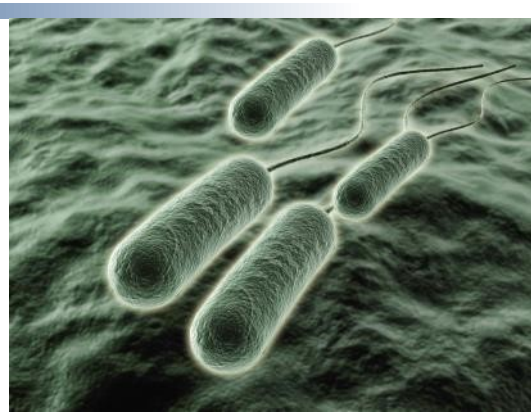




CHARLES UNIVERSITY

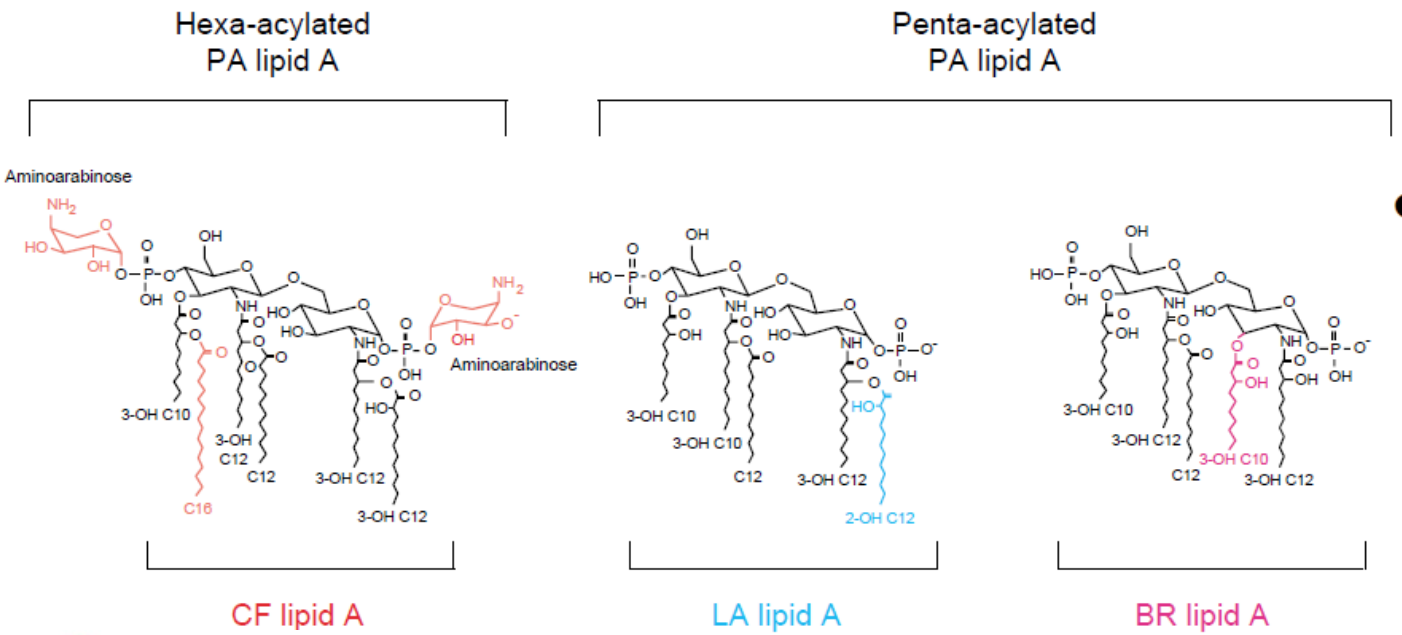
Botos et al. 2009

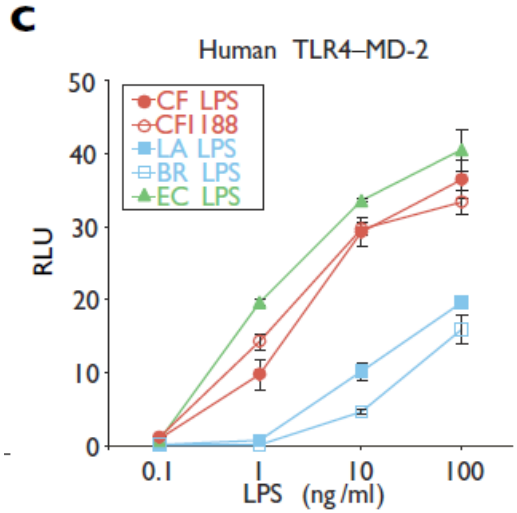
Polymorphism in coding regions



Pseudomonas aeruginosa

- opportunistic bacterium
- during infection
  - Down-regulates the flagellin expression
  - Increases → decreases the acylation state of LPS lipid A



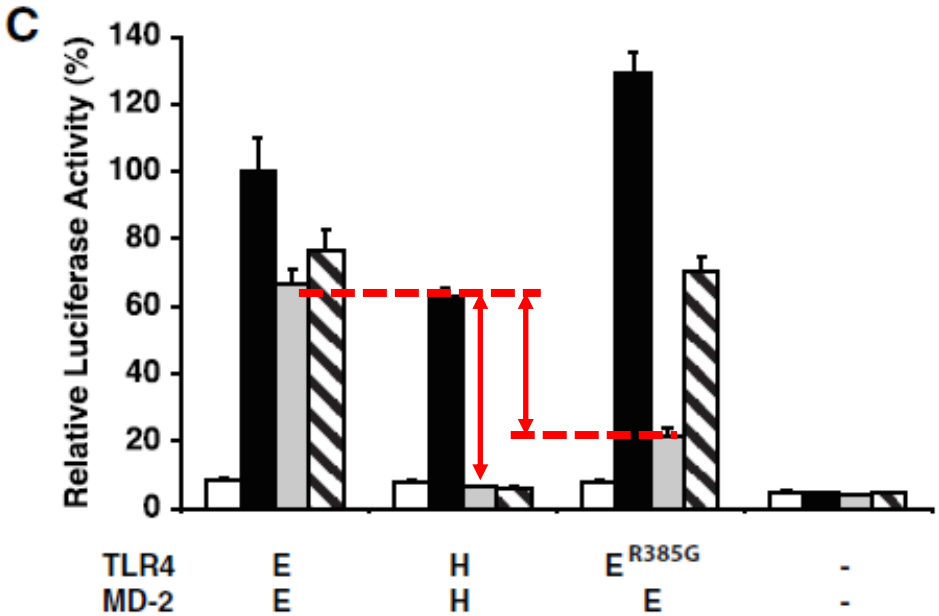


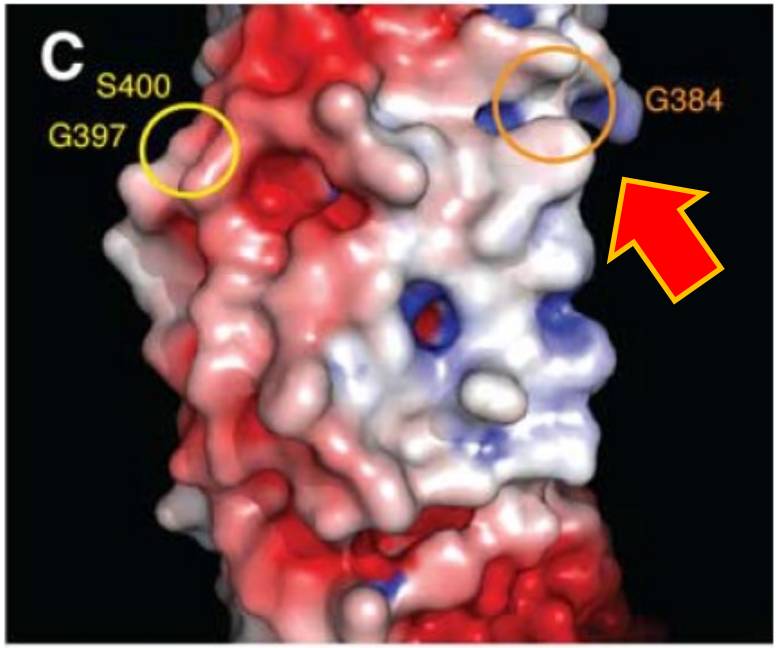
CHARLES UNIVERSITY

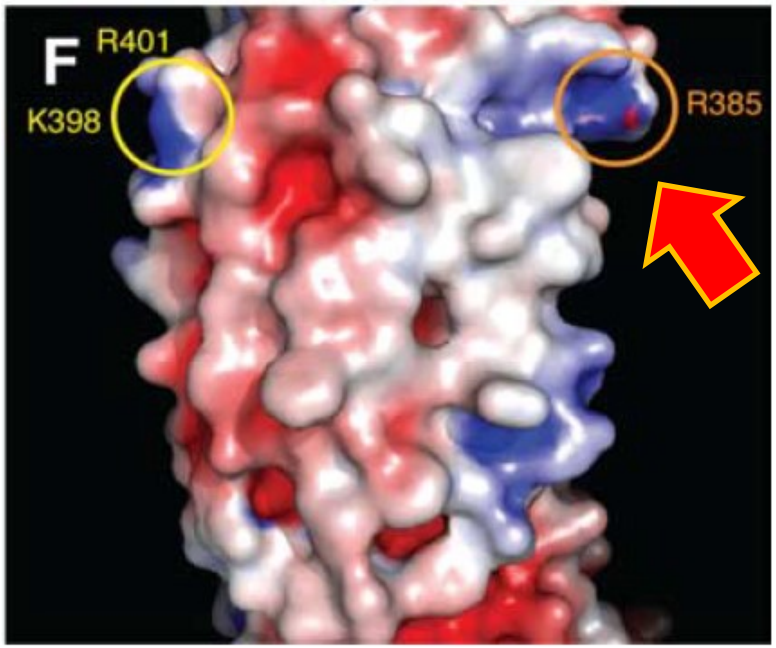
Hajjar et al. 2002

Polymorphism in coding regions

- Lipid IVa = precursor in Lipid A synthesis
- **agonist** in horse and mouse but an **antagonist** in humans and cat
- TLR4: R385G in the glycan-free flank of the horse TLR4 solenoid confers the ability to signal in response to lipid IVa







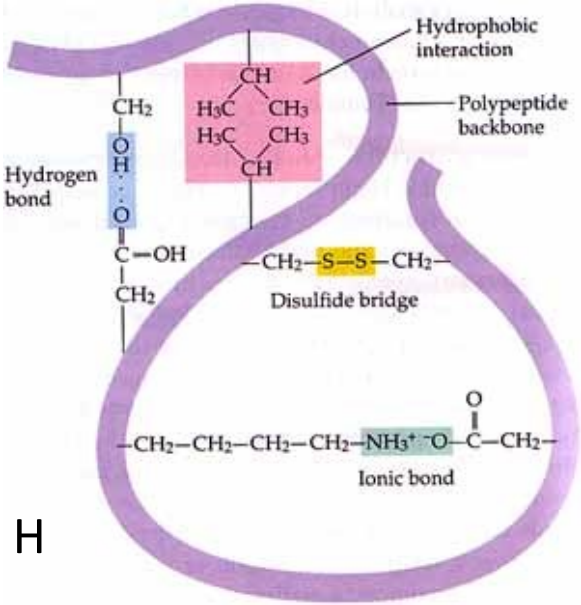
CHARLES UNIVERSITY

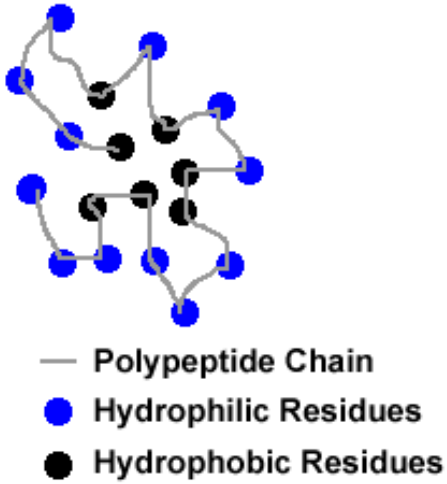
Walsh et al. 2008

Polymorphism in coding regions

In coding regions may influence protein structure:

- **Van der Waals interactions**
  - charged groups induce dipole → dipole-dipole interaction
- **Disulphide bonds**
  - oxidation of the sulfhydryl groups on cysteine
- **Hydrogen bonds**
  - two electronegative atoms compete for the same H atom
- **Hydrophobic interactions**
  - non-polar groups cannot interact with polar groups & water → keep together
- **role of posttranslational modifications**
  - functional groups – e.g. acetate, phosphate, various lipids and carbohydrates





CHARLES UNIVERSITY

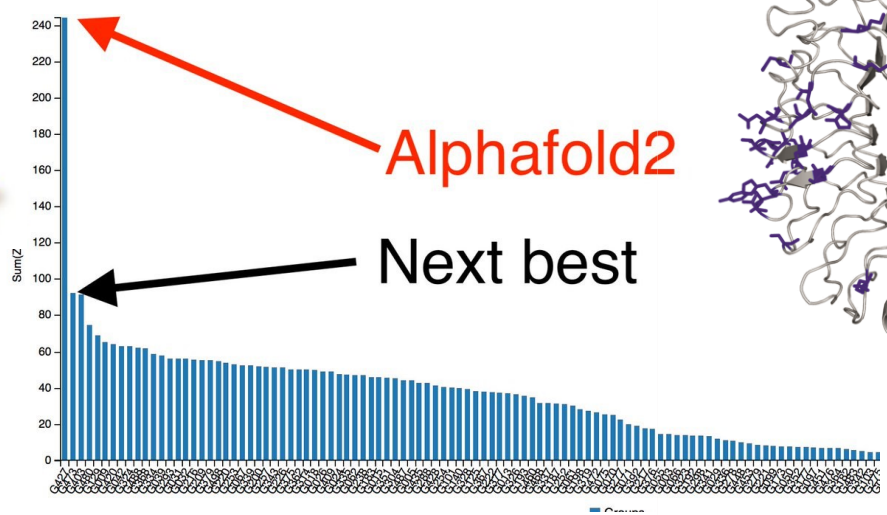
Polymorphism in coding regions

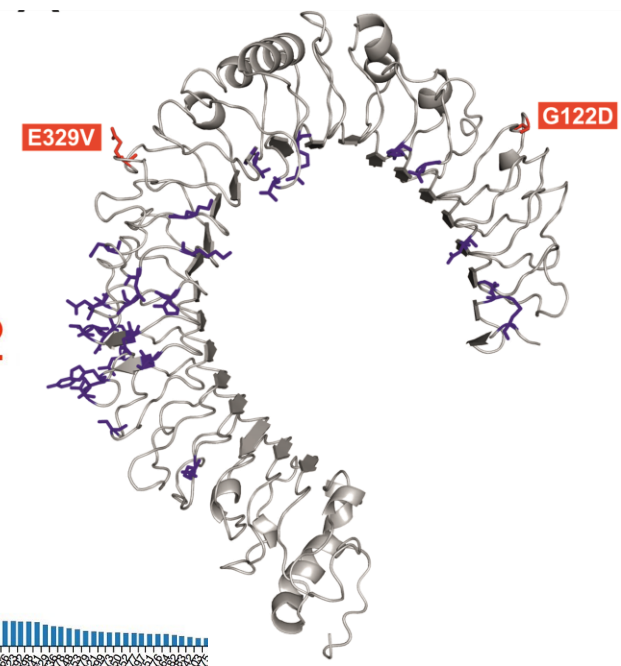
Protein structure prediction

Software:

- SMART – domain architecture

- Specialised domain prediction tools (SignalP, LRR-finder, DAS-Tmfilter, etc.)
- 3D modelling
  AlphaFold2
  Modeller
  I-TASSER

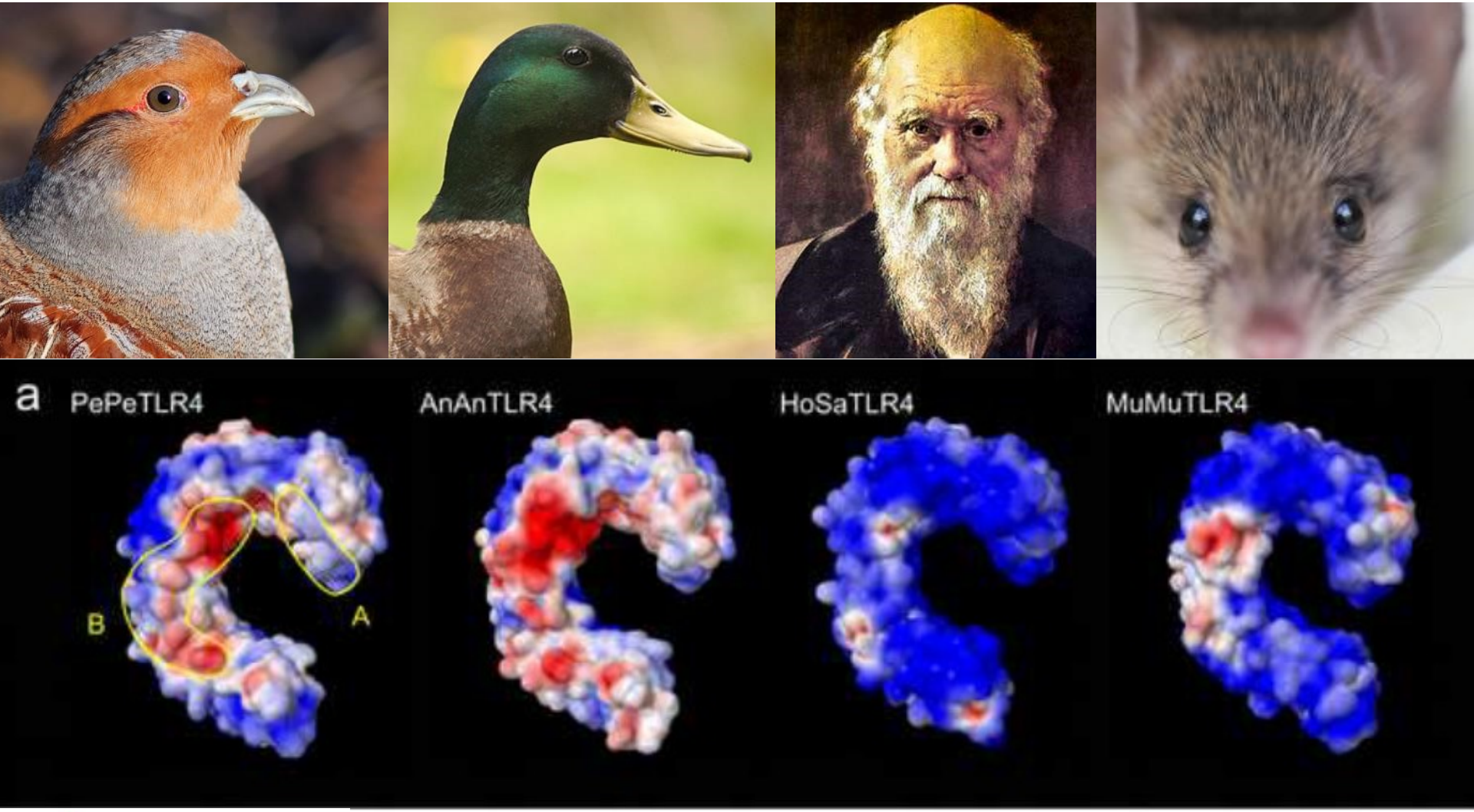




CHARLES UNIVERSITY

Polymorphism in coding regions

Vinkler et al. 2014

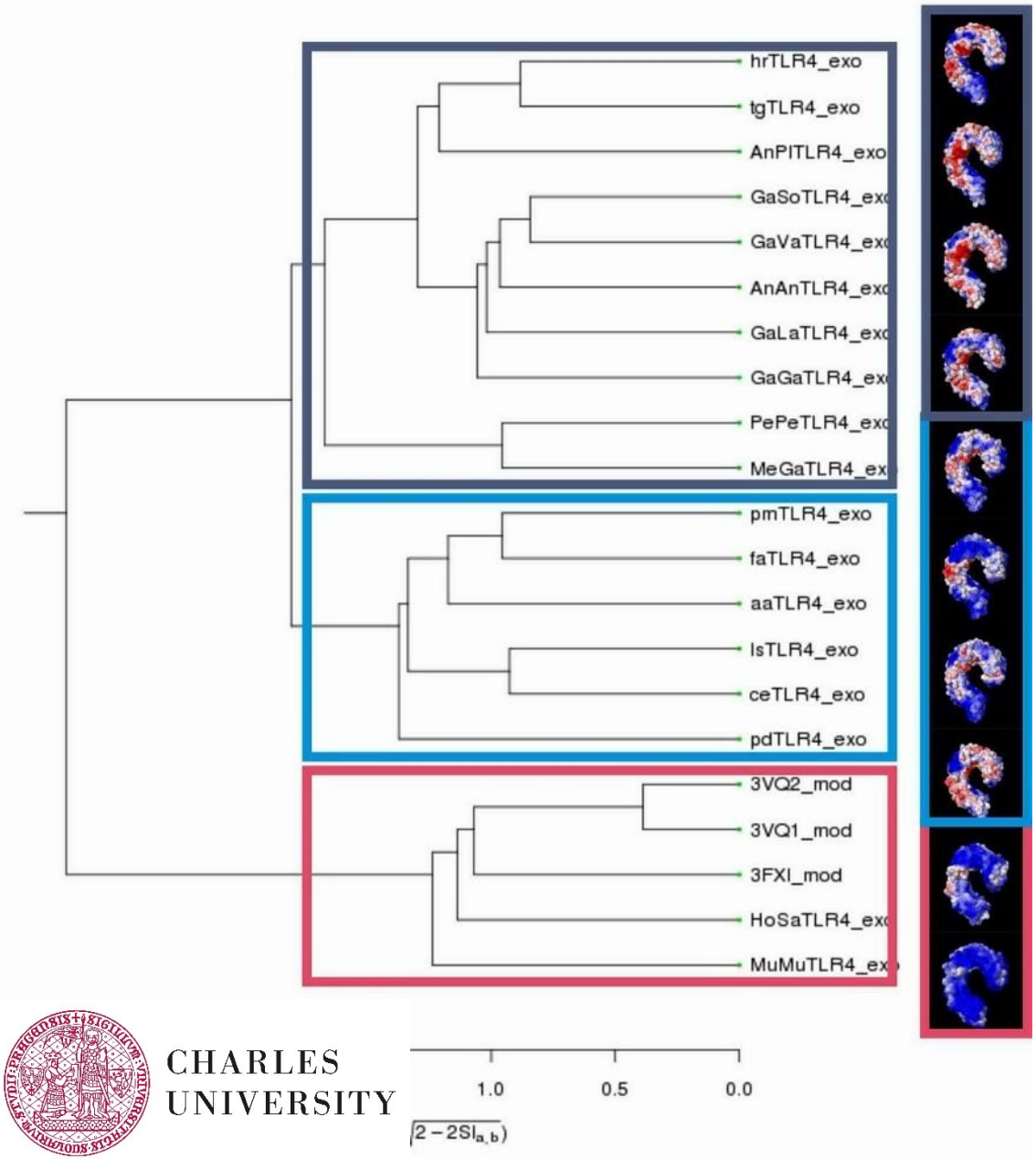


CHARLES UNIVERSITY

Surface electrostatic potential in avian-mammal TLR4s

Software: PDB2PQR Server → visualisation in Jmol

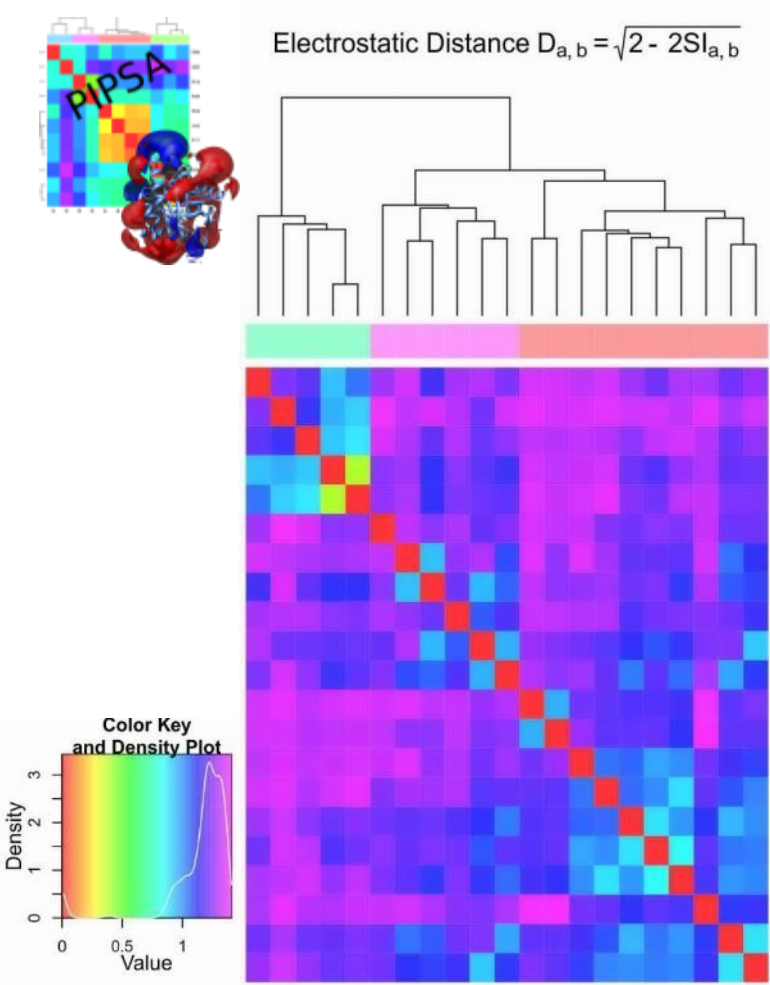
Polymorphism in coding regions



Surface electrostatic potential

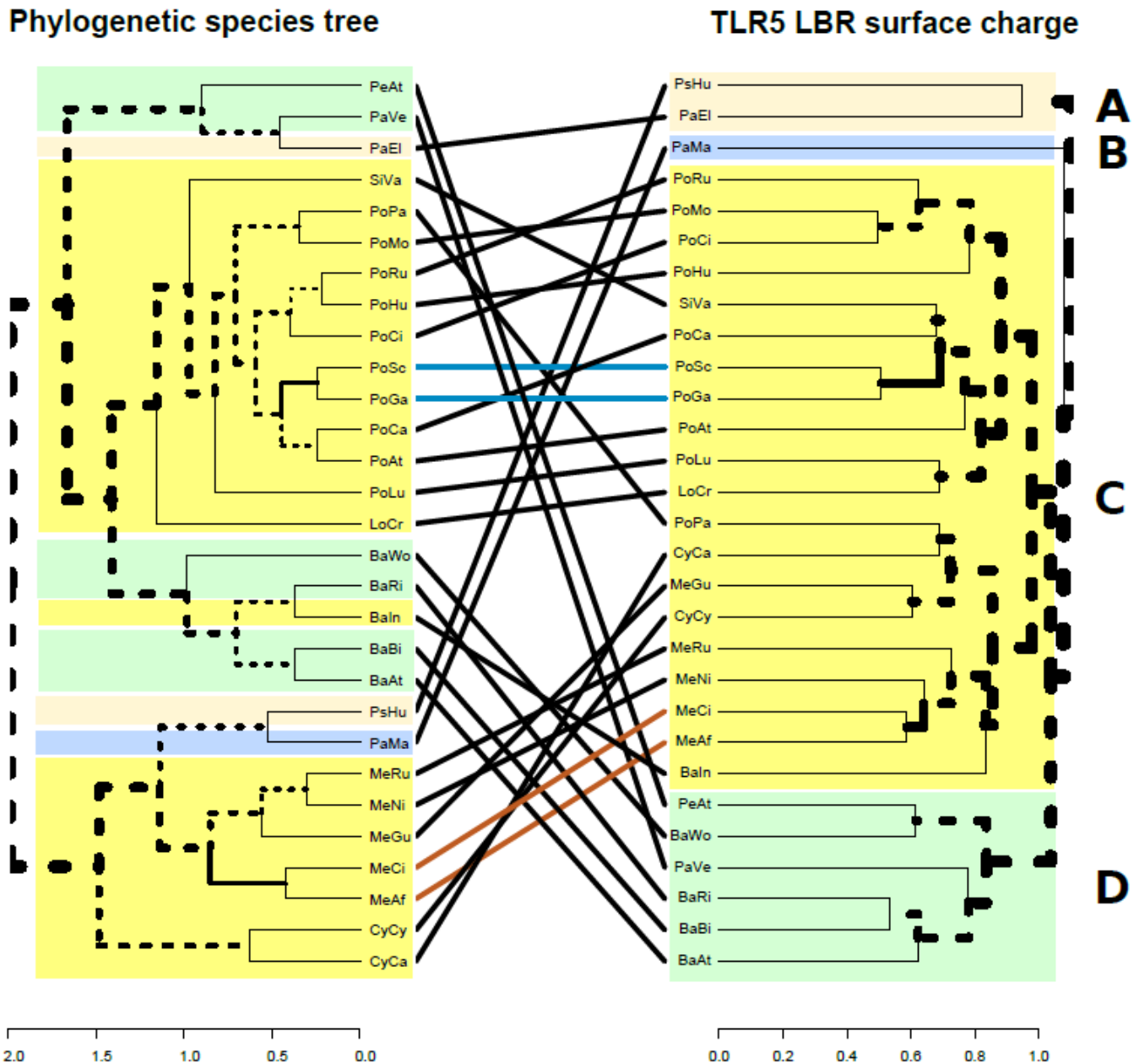
Software:

- PDB2PQR Server
- PIPSA



CHARLES UNIVERSITY

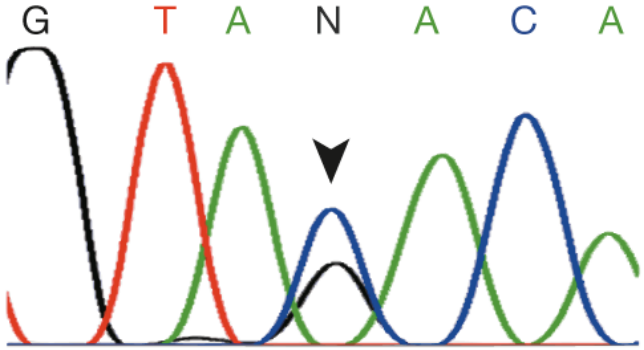
Polymorphism in coding regions



Těšický et al. 2020

CHARLES UNIVERSITY

Alleles



Software:

- DnaSP-PHASE

```
                          60        70        80        90       100       110       120       13
GgTLR4 DNA 1part.   GCTCGTGCTGTCCCTGCTGGCAGGATGCATCCCCAGTCCGTGCCTGGAGGTCATCCCCAGCACAGCTTTCAGAT
OR-2012GG0013       ..A.......................................................................
OR-2012GG0014       ..A.......................................................................
AR-2012GG0002       ....................y.........y...........................................
FX-2012GG0030       ...........................................................Y...........
FX-2012GG0031       ...........................................................Y...........
JK-2012GG0033       ....................Y......................................................
JK-2012GG0034       ....................Y......................................................
```

```
                          60        70        80        90       100       110       120       13
FJ915527.1-1     [b GCTCGTGCTGTCCCTGCTGGCAGGATGCATCCCCAGTCCGTGCCTGGAGGTCATCCCCAGCACAGCTTTCAGAT
OL-2012GG0013-1     ..A.......................................................................
OL-2012GG0013-2     ..A.......................................................................
OL-2012GG0014-1     ..A.......................................................................
OL-2012GG0014-2     ..A.......................................................................
AR-2012GG0002-1     ...........................................................................
AR-2012GG0002-2     ....................T.........T...........................................
FX-2012GG0030-1     ...........................................................................
FX-2012GG0030-2     ...........................................................T...........
FX-2012GG0031-1     ...........................................................................
FX-2012GG0031-2     ...........................................................T...........
JK-2012GG0033-1     ...........................................................................
JK-2012GG0033-2     ....................T......................................................
JK-2012GG0034-1     ...........................................................................
JK-2012GG0034-2     ....................T......................................................
```

- FaBox – haplotype collapser

CHARLES UNIVERSITY

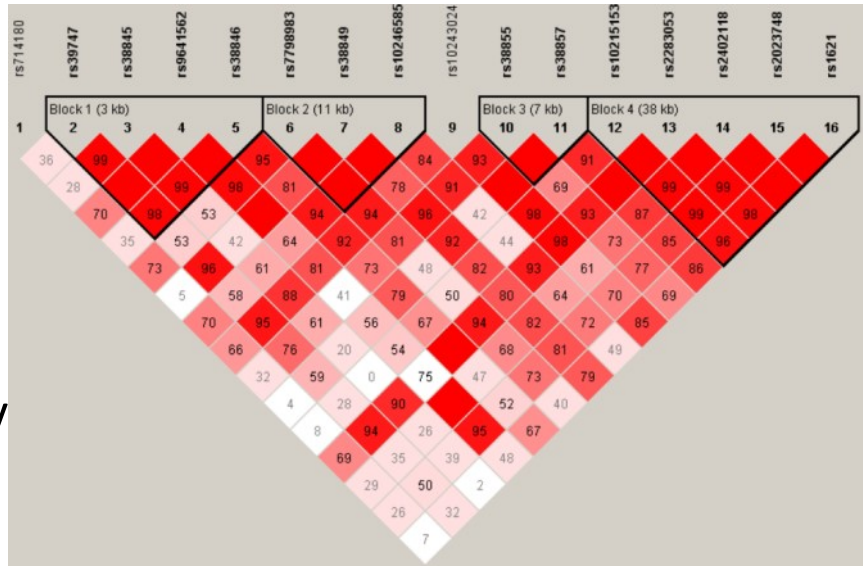
HTTP://www.birc.au.dk/software/fabox

Genetic recombination

Two DNA molecules exchange genetic information → new DNA variant

- Neighbouring SNPs in linkage = **haplotypes**
- **Linkage disequilibrium**
  - = degree of genotype (combination of alleles on several loci) frequency deviation from the expected independent assortment
  - – lowers with distance (probability of recombination increases)
  - – 1cM (=1% frequency of crossover) ≈ 1Mb, but not uniform
  - – considered in association studies
  - – **Recombination hot spots** = regions (1-2kb) with 10x higher recombination than surroundings – humans ca. 50‘000 hot spots

  → genomes in blocks (haplotype blocks), 5-100kb



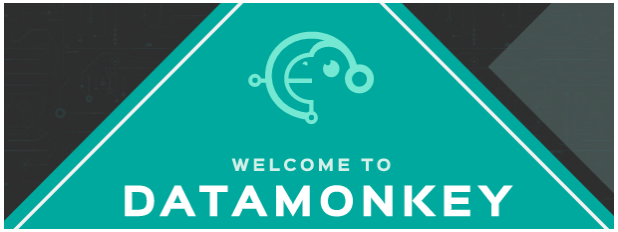
- Tag SNPs – SNPs selected to identify individual haplotypes

CHARLES UNIVERSITY

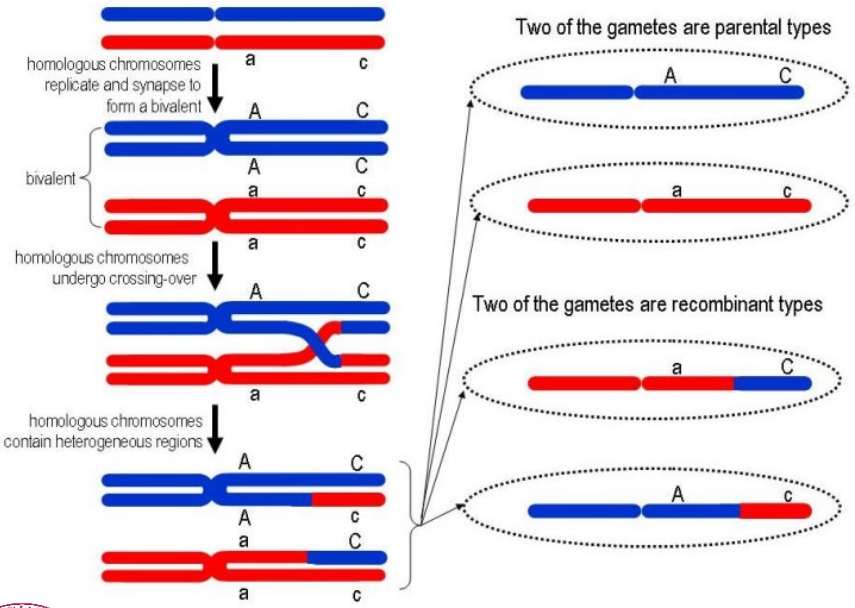
Genetic recombination

Software:

– http://www.datamonkey.org/ - GARD



BPs ?	AIC_c ?	Δ AIC_c ?	Segments ?
0	8694.58		1-2526
1	8537.02	157.556	1378
2	8453.61	83.4092	675, 1755
3	8424.32	29.2955	675, 1293, 1755

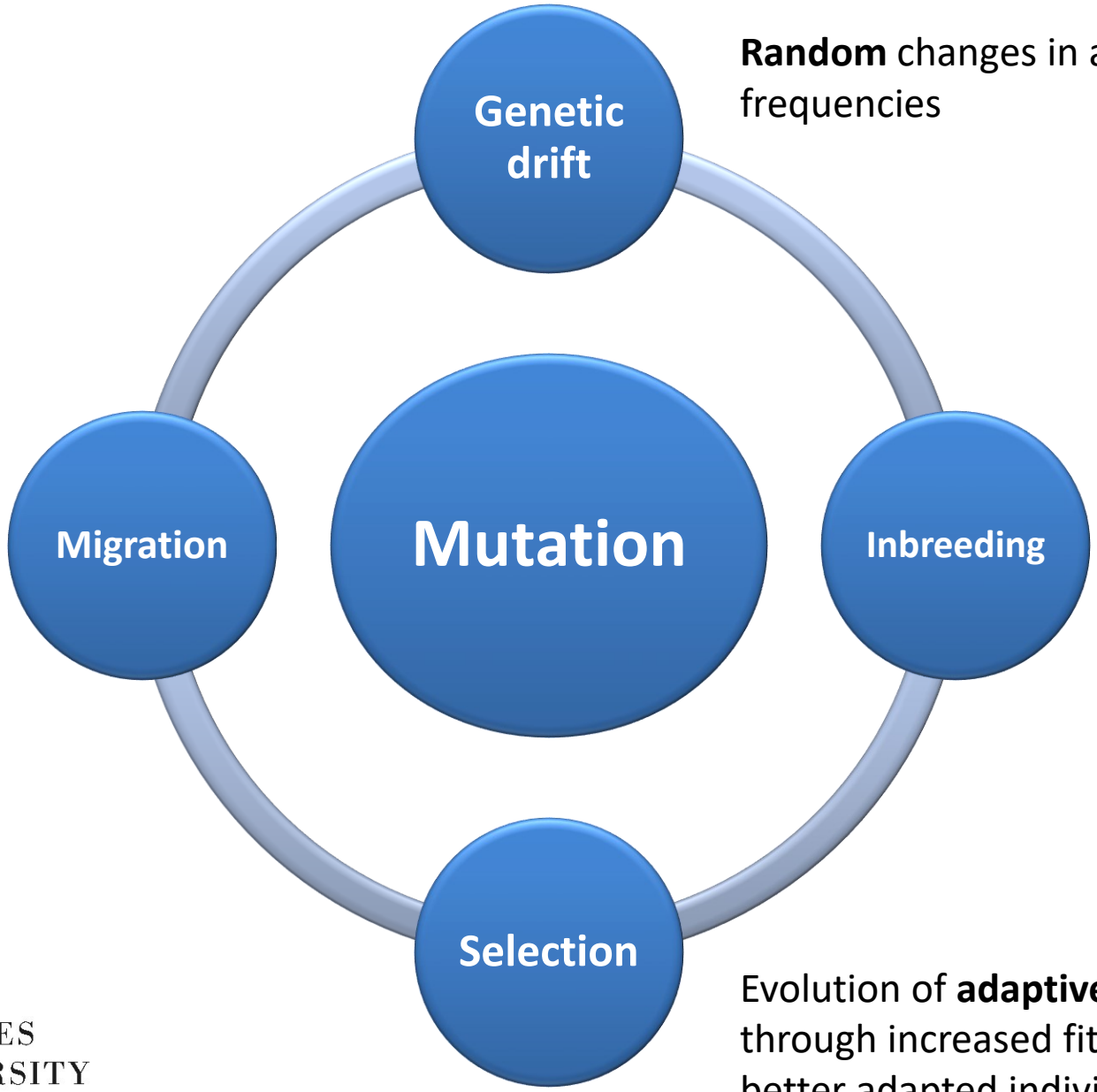


```
          111111111222
       123789111336677112
      501184822225484437 3
      147589738960802742 1
H_1   ACTCTGCGTATTATGACGT
H_4   ............T..C
H_5   G.C...TAC.C.GCA...C
H_2   G.......C.C........
H_19  G.C...TAC.C....T..C
```

RH

CHARLES UNIVERSITY

Genotype evolution: neutral or adaptive?



Random changes in allele frequencies

Gene flow

Evolution of **adaptive** traits through increased fitness of better adapted individuals

CHARLES UNIVERSITY

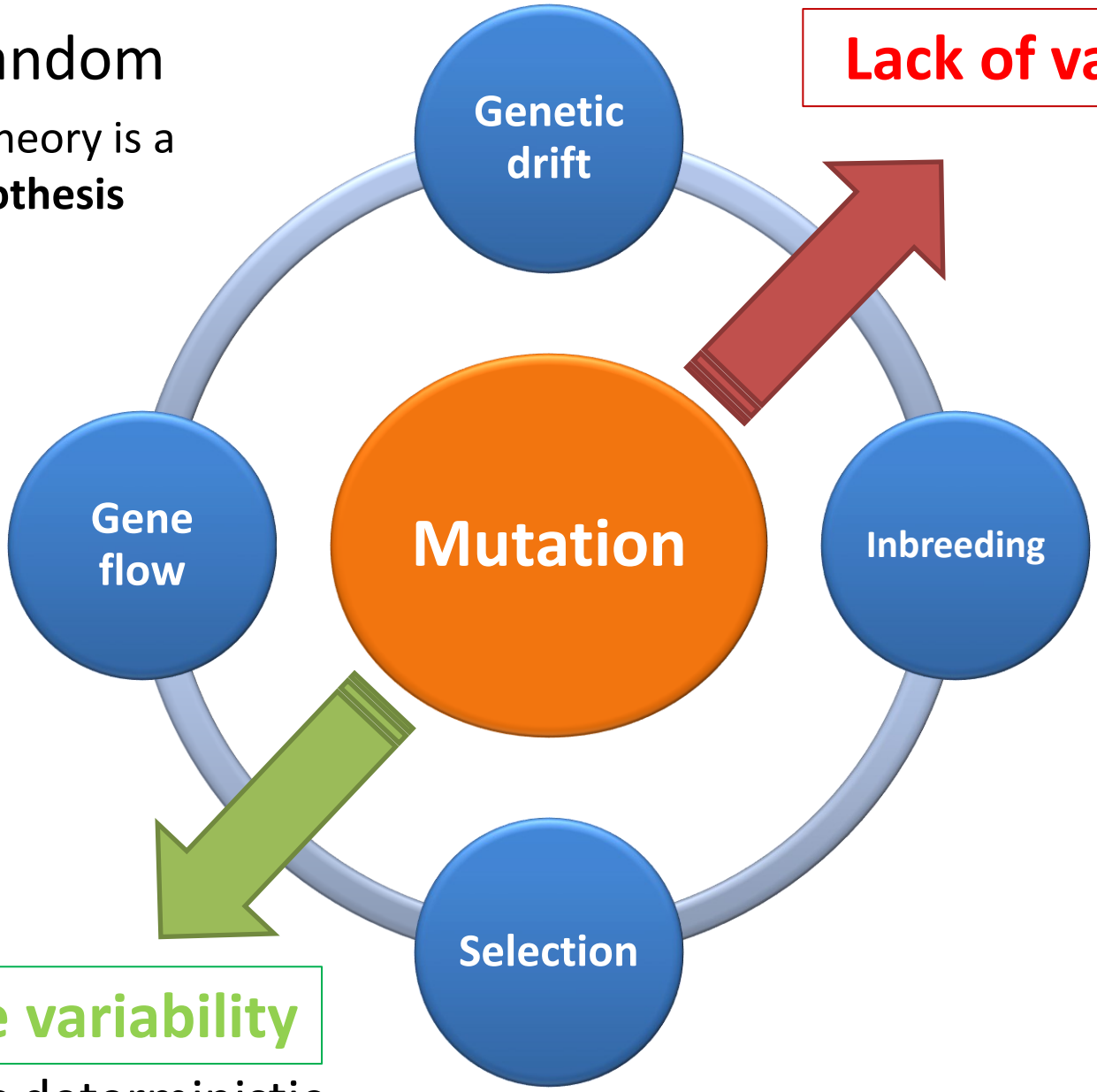
Polymorphism in coding regions

Which processes decrease genetic variation in a population?

Genotype evolution: neutral or adaptive?

Drift is random

→ Neutral theory is a **null hypothesis**



Selection is deterministic

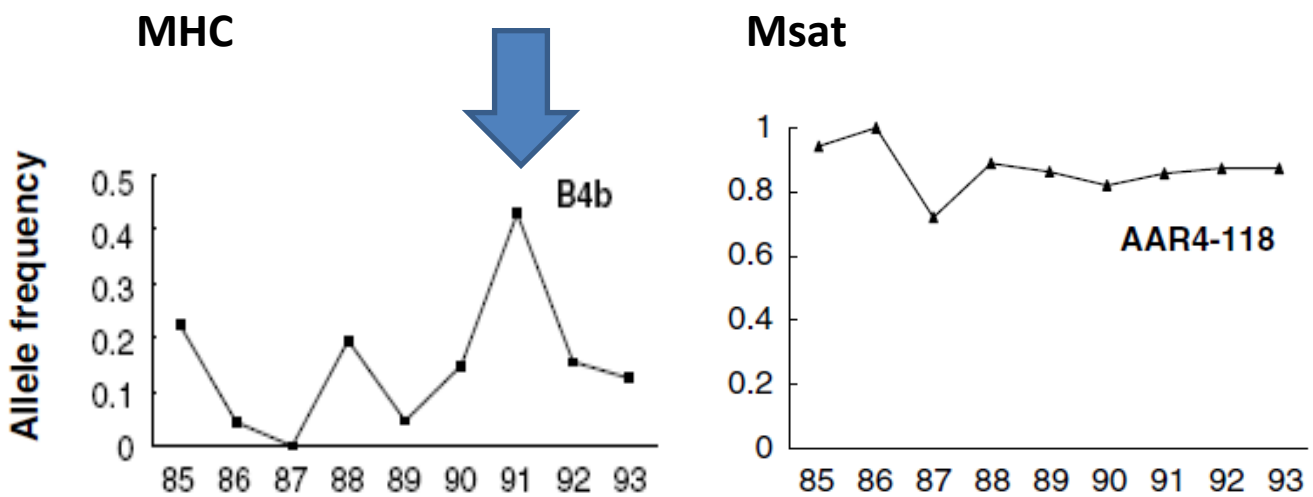
Genetic drift

Fluctuations of allele frequencies in time

Great reed warbler (*Acrocephalus arundinaceus*)

- MHC I
- Comparison of frequency changes in 23 MHC alleles and 23 microsatellite alleles in time
- Non-random fluctuation of frequencies of 2 alleles in time

→ evidence for variability in selection

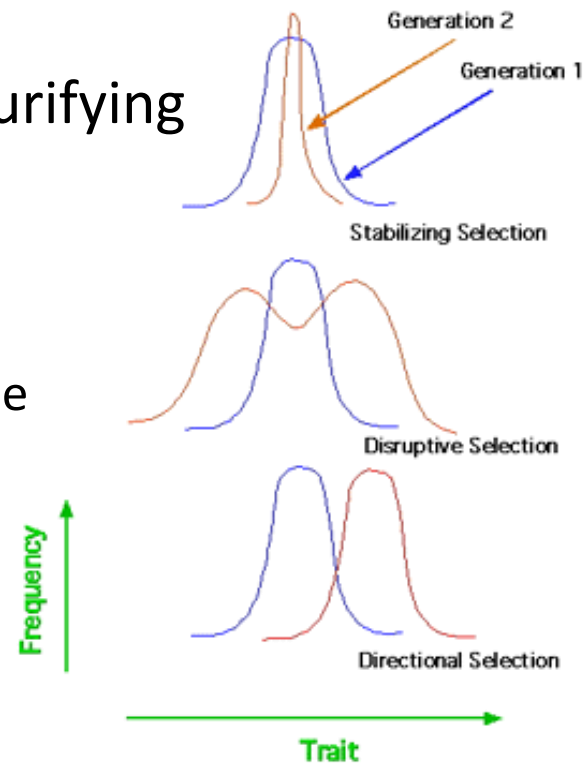


Westerdahl et al. 2004

Selection types

Selection on host genes:

- **Negative** = Stabilising = Purifying
  - Elimination of deleterious
  - Shallower genealogy
- **Disruptive** = Diversifying
  - Polymorphism maintenance
  - Deeper genealogy
- **Positive** = Directional
  - Fixation of advantageous
  - Shallower genealogy



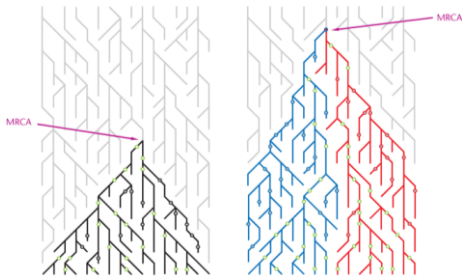


Table 10.5 Effects of various evolutionary processes on genetic variation (modified after Nielsen 2005).

Process	Degree of genetic variation[1]			Frequency spectrum[3]
	Within species	Between species	Ratio (B/W)[2]	
Mutation accumulation	+	+	No effect	No effect
Negative directional selection	–	–	(–)	More low-frequency alleles
Positive directional selection	+ or –	+	+	More high-frequency alleles
Balancing selection	+	+ or –	–	More medium-frequency alleles
Selective sweeps	–	No effect, or (+)	+	More low-frequency alleles

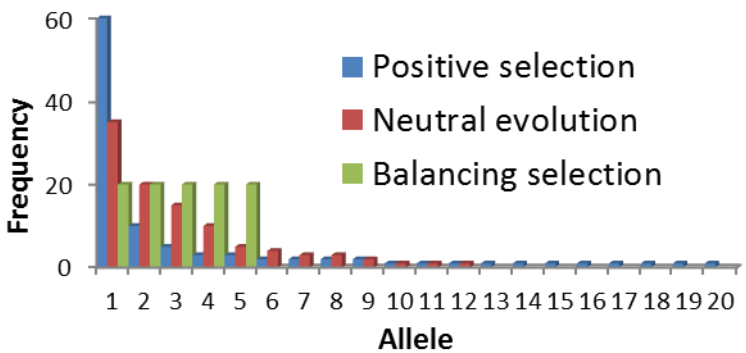
Distribution of allele frequencies

Tajima's D test

- Compares observed **nucleotide heterozygosity** (θπ) and observed **number of polymorphic sites** corrected for sample size (θw)

At equilibrium θπ = θw

- $D = \frac{\theta\pi - \theta w}{Var\ D}$
  - **D>0** if θπ>θw
  - **D<0** if θπ<θw



- <u>Excess of low-frequency alleles</u>

→ more SNP sites than heterozygosity (θπ<θw)

→ **D<0** = **positive** or **negative selection** (or population expansion)

- <u>Intermediate allele frequencies</u>

→ High heterozygosity per SNP

→ **D>0** = **balancing selection** (or population contraction)

- Roughly +2<D or D<-2 is likely to be significant

Divergence vs. polymorphism

Ratio Polymorphism : Divergence is the same for different loci under neutrality

HKA-test (Hudson-Kreitman-Aquadé)

- Expected levels of divergence and polymorphism vs. observed levels of divergence and polymorphism at several loci (at least 2)
- Distinguish locus-specific selection from population-level effects
- Due to linkage disequilibrium the selection may not be on the tested site but somewhere in the neighbourhood

MK-test (MacDonald-Kreitman)

- Comparison of dN and dS within species and among species

= compares two types of sites within the same locus! (same genealogy)

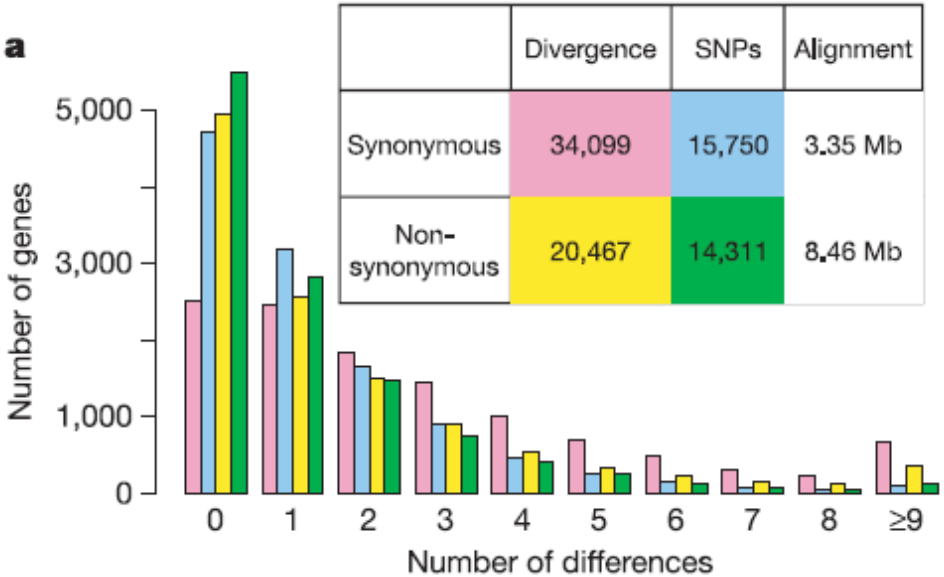
- X^2 test or Fisher's exact test

McDonald-Kreitman Test

Human variability vs. chimpanzee

- exon-specific PCR amplification of 11,624 genes in 39 humans and one chimp
- Variability in 10,767 genes (92.6%), 8,292 had >1 NS SNP or fixed difference
- 9.0% under rapid amino acid evolution
- 13.5% with low divergence between human and chimp
- Negative selection in cytoskeletal proteins, vesicle transport and related functions
- Positive selection in transcription factors, mRNA transcription, receptors and immunity

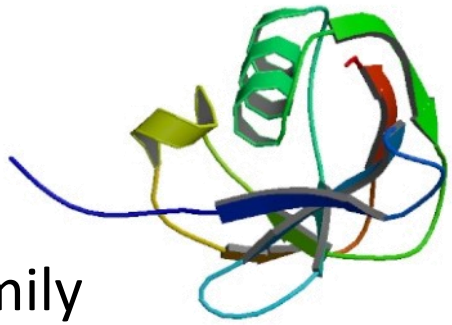
Bustamante et al. 2005



	Divergence	SNPs	Alignment
Synonymous	34,099	15,750	3.35 Mb
Non-synonymous	20,467	14,311	8.46 Mb

Population differentiation

CD5



- is a lymphocyte surface co-receptor
- scavenger receptor cysteine-rich (SRCR) superfamily
- Poorly known function:
  - cell-to-cell immune interactions
  - recognition of fungal β-glucans
- 27 polymorphic sites:
  - comprising 17 intronic, 3 synonymous, 7 nonsynonymous substitutions
  - selection for **A471V** substitution in cytoplasmatic region

  → differences in MAPK cascade activation

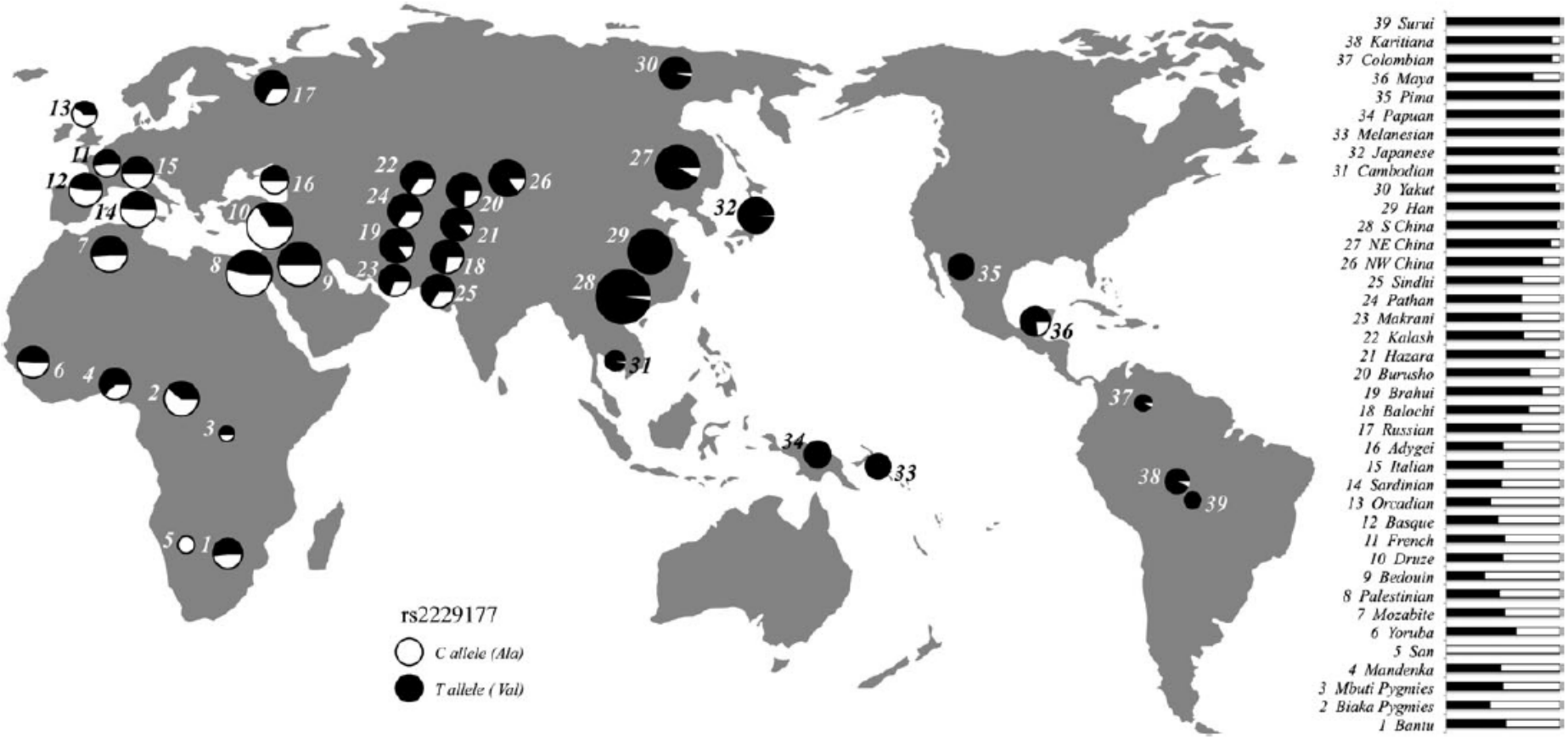
  → higher IL-8 in V471

CHARLES UNIVERSITY

Carnero-Montoro et al. 2012

Selective sweep

CD5

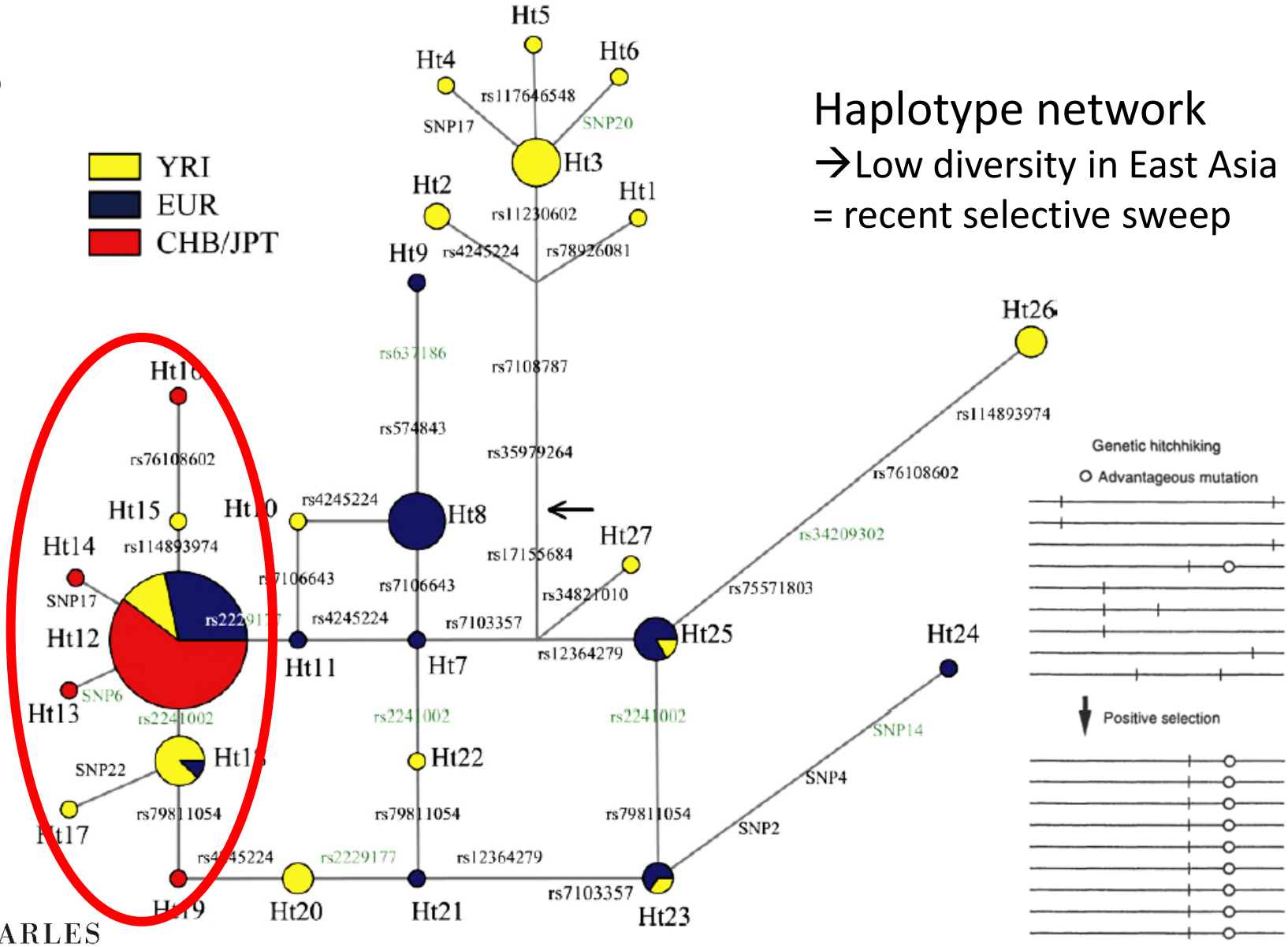


CHARLES UNIVERSITY

Carnero-Montoro et al. 2012

Selective sweep

CD5



Haplotype network

→Low diversity in East Asia

= recent selective sweep

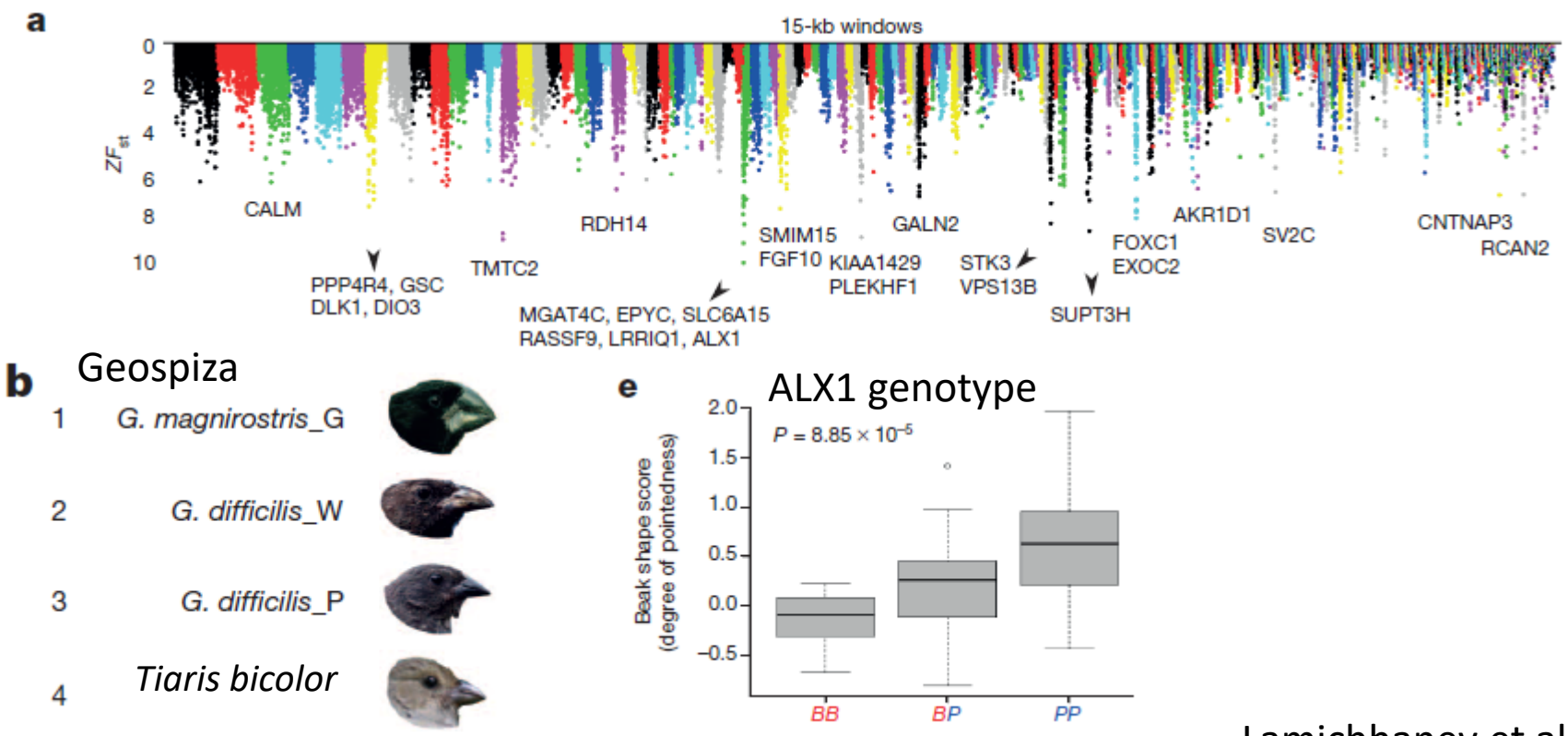
CHARLES UNIVERSITY

Carnero-Montoro et al. 2012

Fixation index (F_{ST})

- measure of population differentiation due to genetic structure
- based on the variance of allele frequencies between populations / probability of Identity by descent

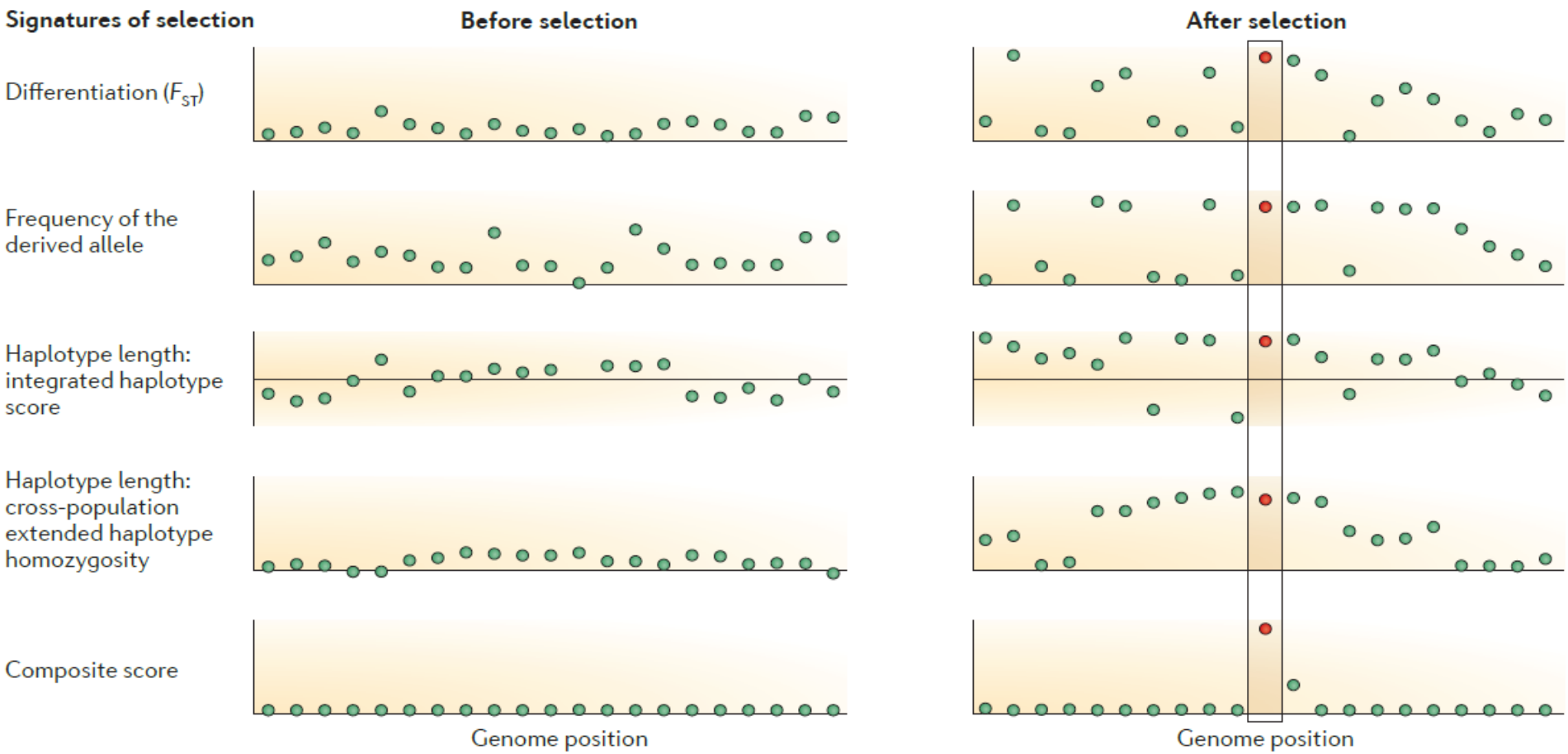
Evolution of Darwin's finches' beaks



Lamichhaney et al. 2015

Fixation index (F_{ST}) & GWAS

Genome-wide association studies (GWAS).

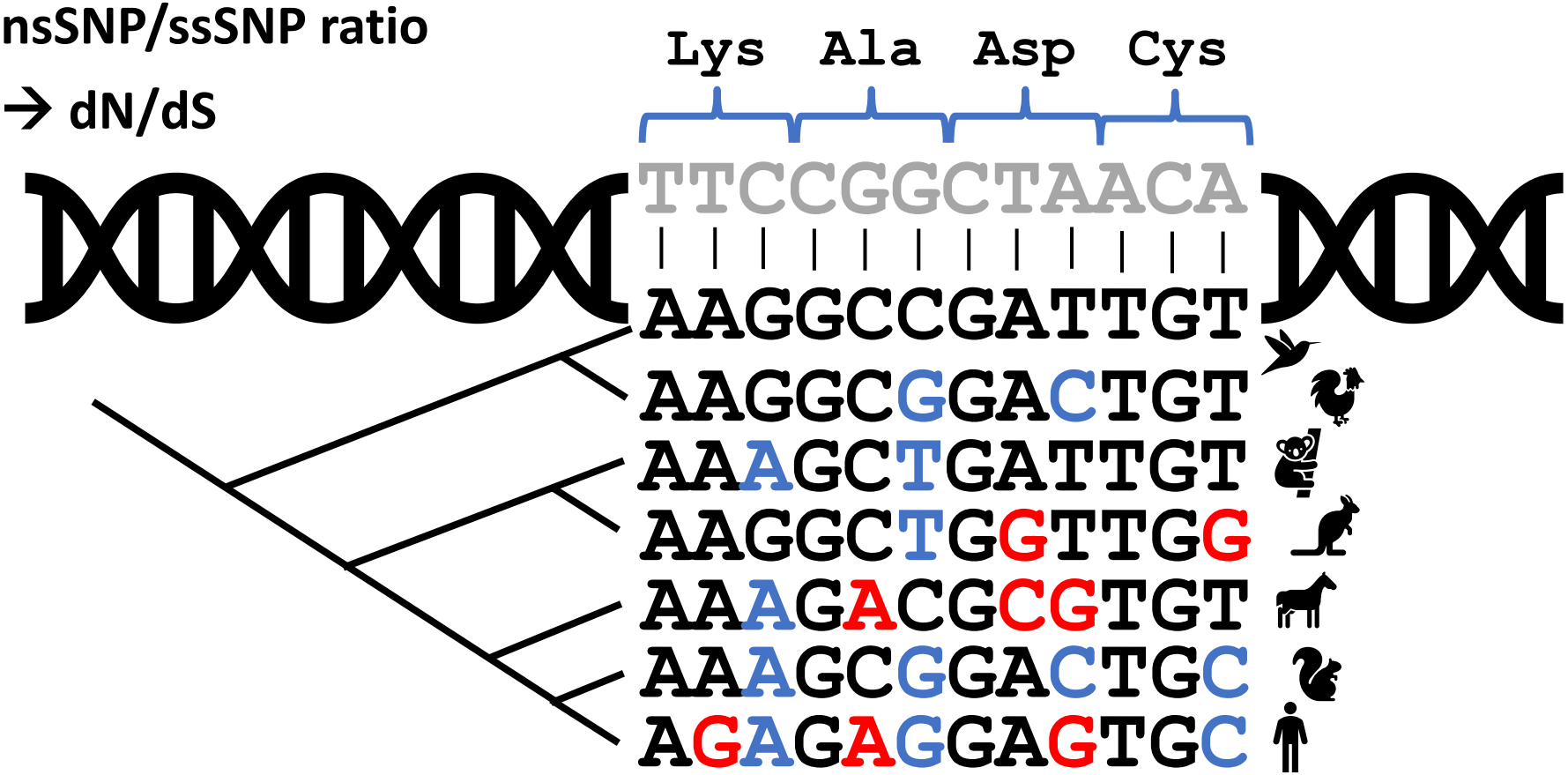


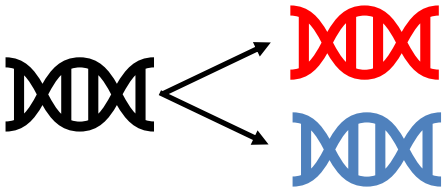
Karlsson et al. 2014

Pattern of variability in sequences

nsSNP/ssSNP ratio

→ dN/dS



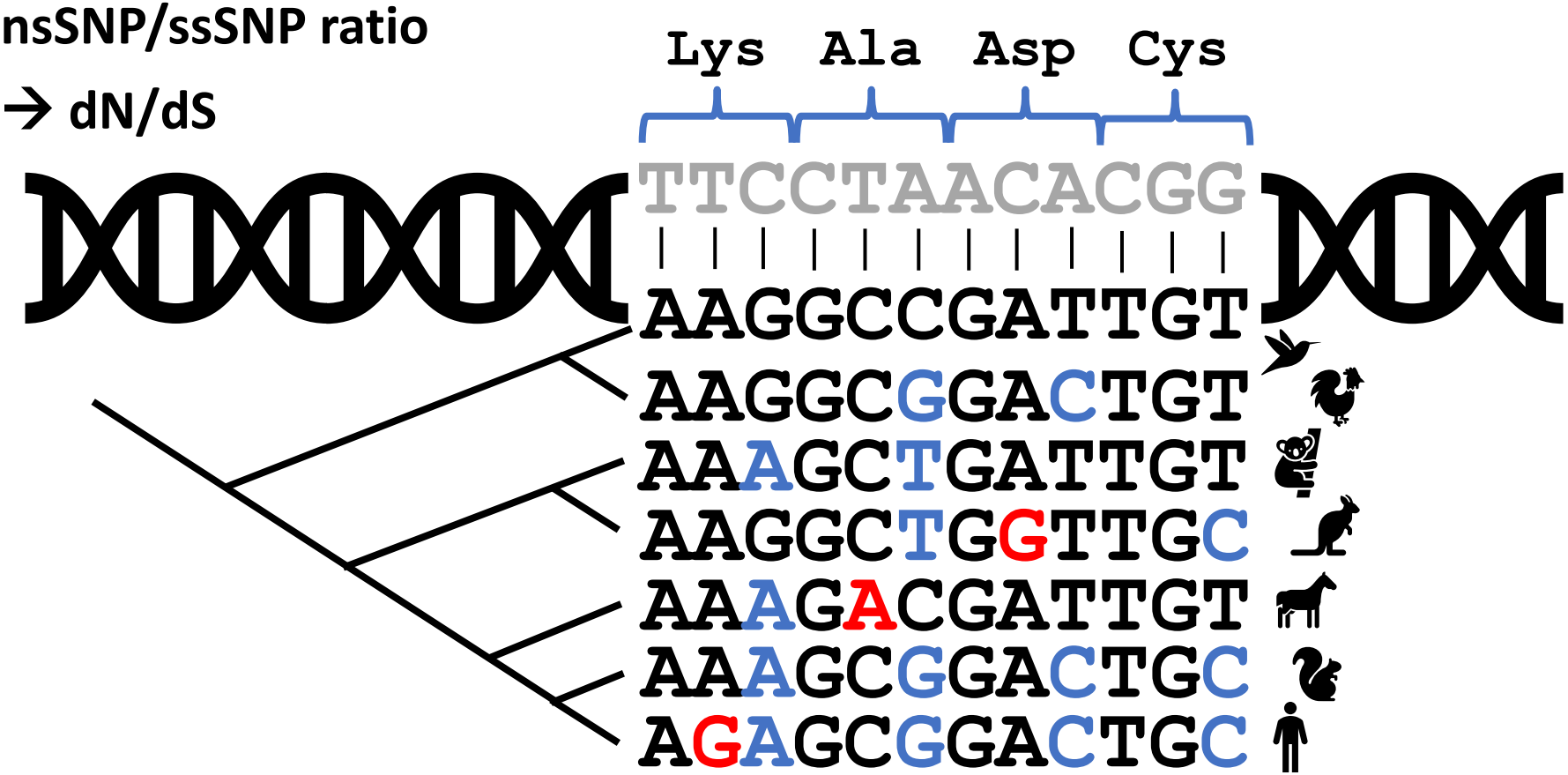


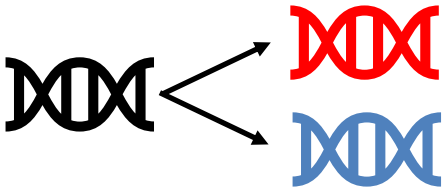
$$\frac{dN}{dS} \sim \frac{8}{8} = 1$$

Pattern of variability in sequences

nsSNP/ssSNP ratio

→ dN/dS



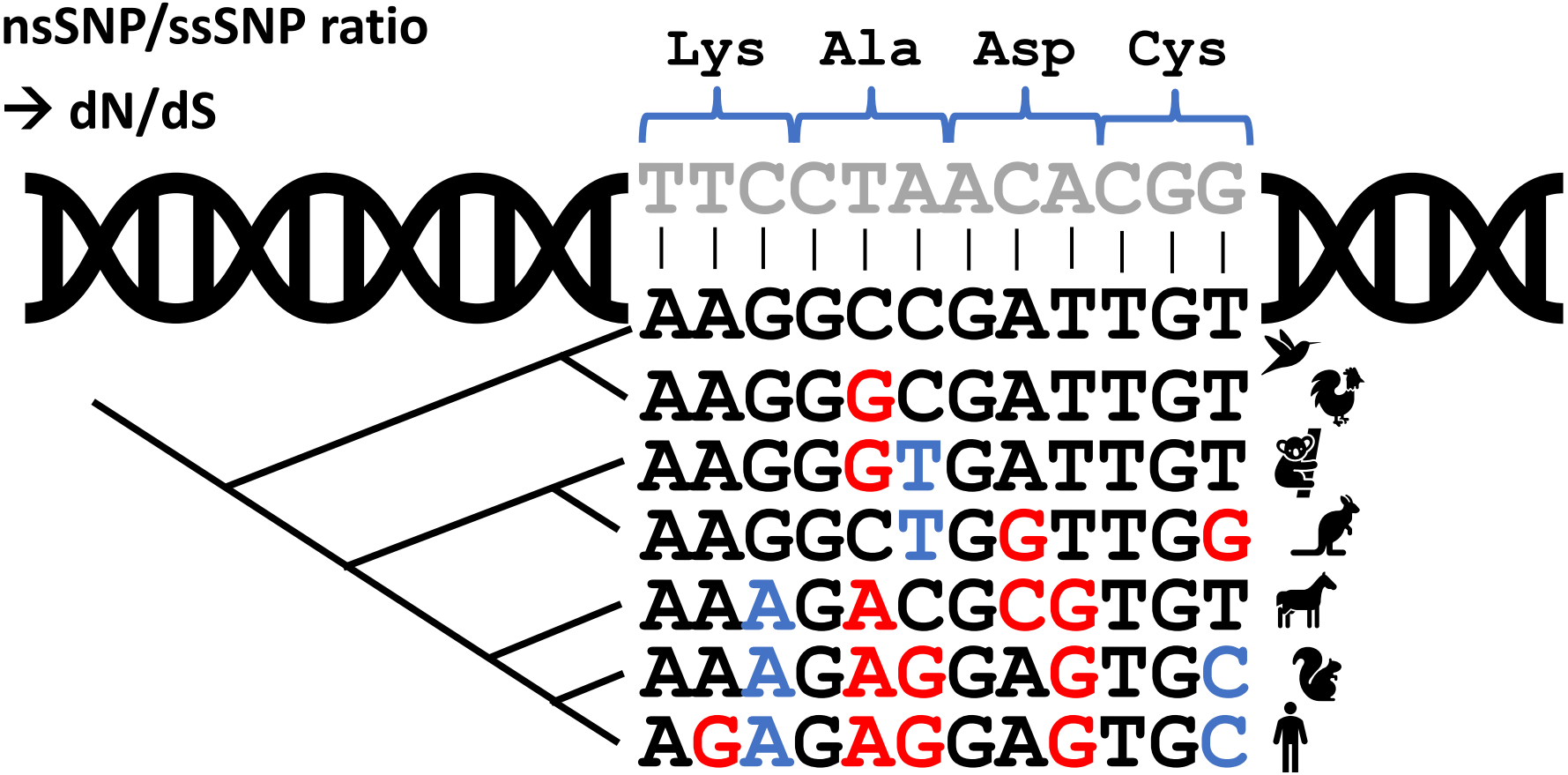


$$\frac{dN}{dS} \sim \frac{3}{9} < 1$$

Pattern of variability in sequences

nsSNP/ssSNP ratio

→ dN/dS



$$\frac{dN}{dS} = \frac{8}{3} > 1$$

Pattern of variability in sequences

Ratio of non-synonymous (d_N=Ka) substitutions to synonymous (d_S~Ks) substitutions

- Ka/Ks = ω
  - per locus
  - per site
  - Interspecific & intraspecific
  - neutral: **Ka/Ks=1**, positive: **Ka/Ks>1**, negative: **Ka/Ks<1**
  - little power to detect weak positive selection
  - power increases with sample size (species number)

Software:

- PAML
- http://www.datamonkey.org/ - interspecific: *SLAC* / *FEL* / *FUBAR* (maximum likelihood methods) and branch-specific models (*MEME*)

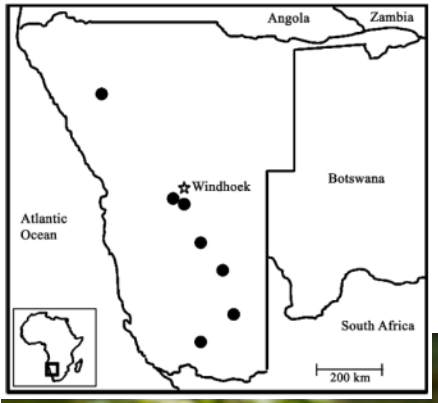
CHARLES UNIVERSITY

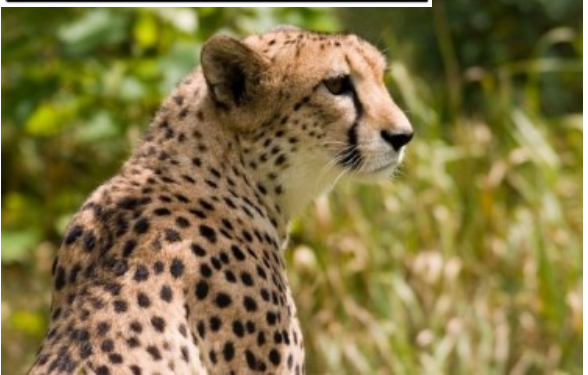
WELCOME TO DATAMONKEY

Distinct selection at different sites

Selection in different domains

- **MHC I and MHC II**
  - The d_N/d_S was higher in antigen-binding sites (ABS) than in non-ABS
  - Differentiating selection on antigen binding features

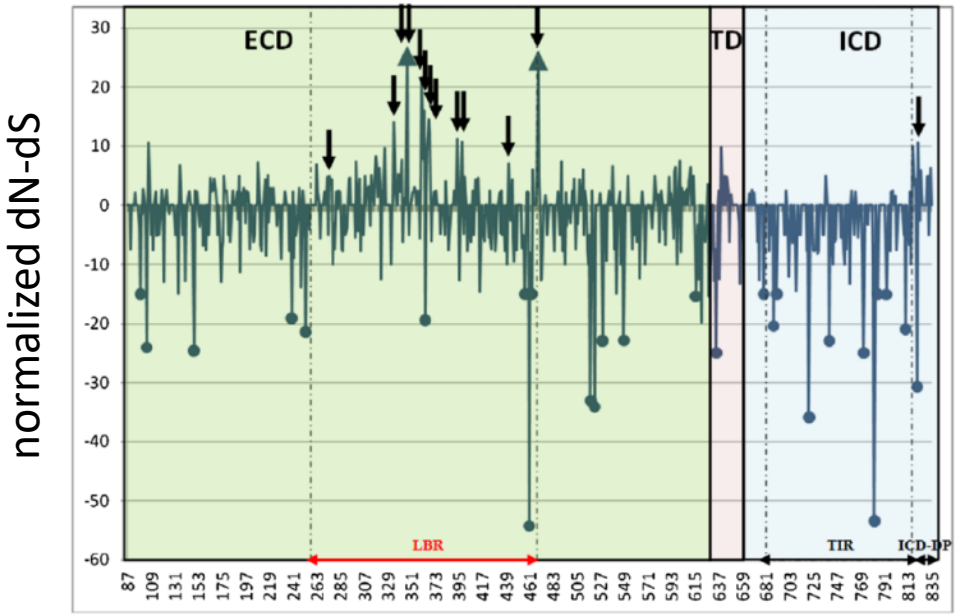




MHC	Region	Sites	d_N/d_S	P
Class I	Exon 2	ABS	2.87	<0.01
		Non-ABS	1.00	0.71
		All	1.60	0.04
	Exon 3	ABS	1.20	0.51
		Non-ABS	0.15	<0.01
		All	0.38	0.02
Class II-DRB	Exon 2	ABS	1.40	0.35
		Non-ABS	0.31	0.06
		All	0.69	0.25

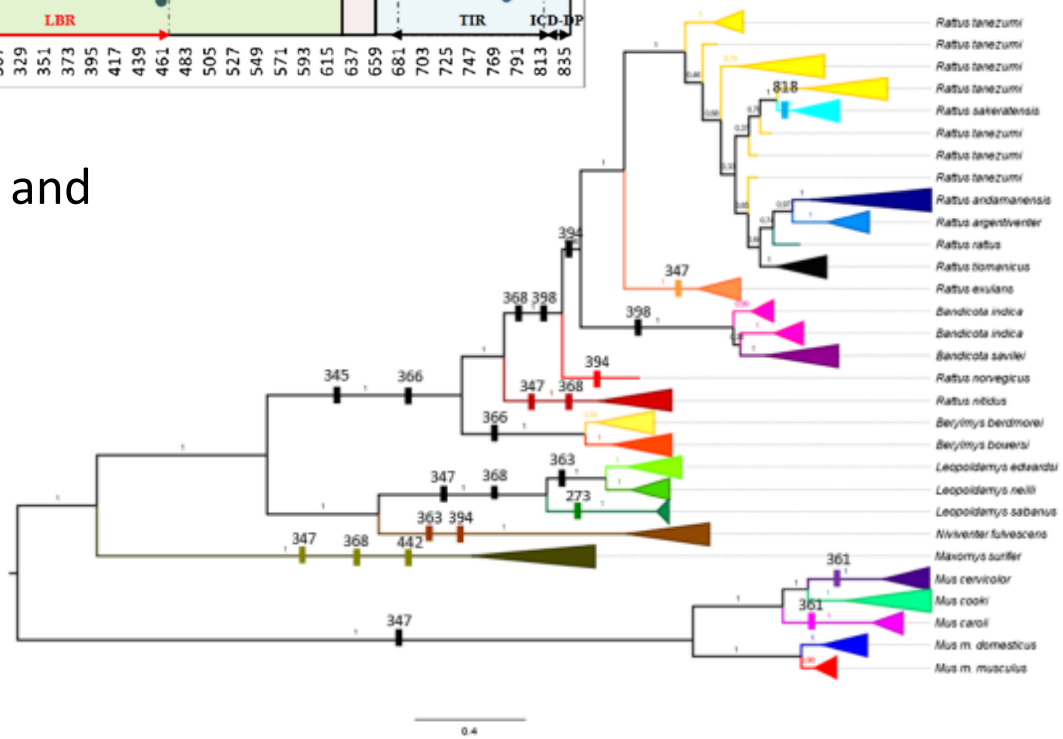
Castro-Prieto et al. 2011

Pattern of variability in sequences



Fornuskova et al. 2013

– lineage-specific effects and codon-specific effects (*MEME*)



CHARLES UNIVERSITY

Genome searches

Ka/Ks usage for pathogen elicitor detection

- G^- bacteria core genome
  - genes represented in all studied species (6)
  - 1‘322 orthologous
- Ka>Ks
  - 35% of the core genes had at least one positively selected
  - 56 proteins exhibited significant signatures of positive selection
- comparison with selection in soil-inhabiting bacteria
  - 48 proteins positively selected in pathogenic and not in non-pathogenic
- positively selected sites in clusters
  - 45 loci

→ Functional testing in *A. thaliana*

CHARLES UNIVERSITY

Footprints of selective events in genes

Advantages:

- Detection of events with small but biologically relevant selection coefficients
- Selection on evolutionary not ecological time scale (= selection in the past)
- Testing selection in genes without knowledge of the phenotype
- May allow detection of genes responsible for emergence of novel traits

Disadvantages:

- Demographic changes may give similar results as selection if only 1 gene observed
- When selection is found, linking with phenotype is difficult

→ We know about selection but we do not understand its functional consequences

→ Linking detected selection with phenotype is a major challenge

CHARLES UNIVERSITY

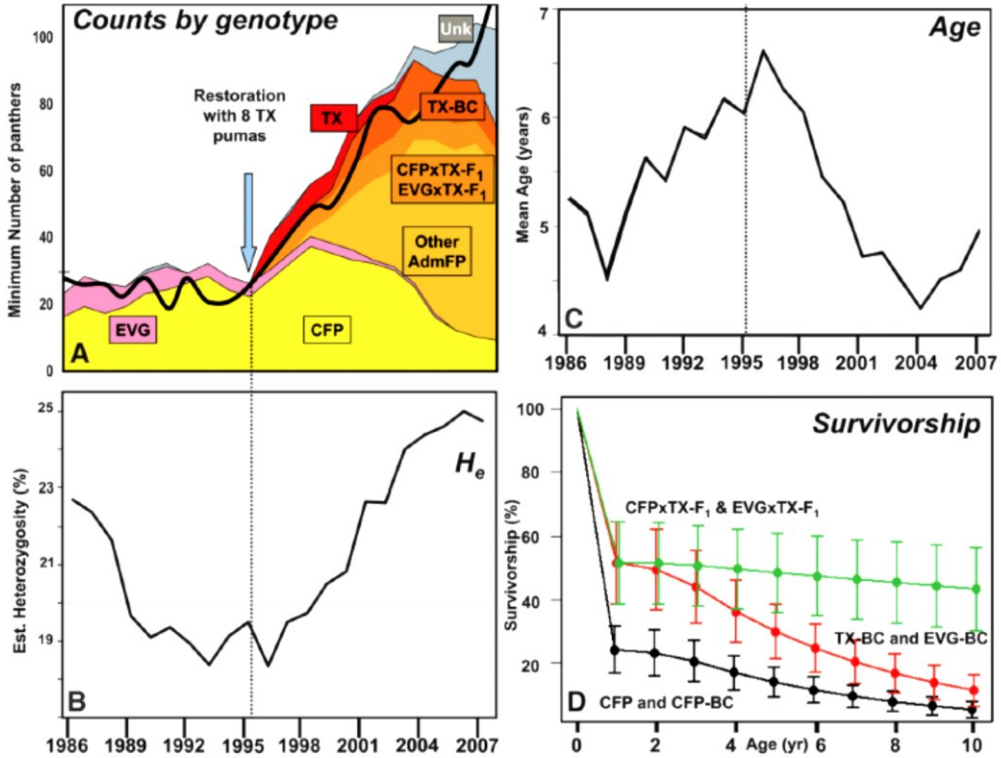
Nature conservation efforts

Florida panthers (*Puma concolor coryi*)

Inbred population in Florida

→ in 1995 translocation of 8 female pumas from Texas (*P. c. stanleyana*)

→ Fitness improvement



Two pre-1995 groups (CFP and EVG)
TX females → TX-backcross (TX-BC)
admixed Florida panthers (AdmFPs)

Johnson et al. 2010

Conclusion

- There are multiple techniques to detect genetic variation

What can I do with sequence data to reveal functional differences between individuals / populations / species?

1. Check the position of your SNPs
2. Model protein structures and predict functional effects of substitutions
3. Identify alleles and non-synonymous protein variants and assess their frequencies in distinct populations
4. Detect recombination
5. Detect selection and recognize adaptive evolution

Practical training

- The latest urging e-mail has been sent today by Zuznana Starostová

→ **ACTION REQUIRED !!!!!**

- Fill in information on the preference of the training time:
- **2 May:**
    - Morning?
    - Afternoon?