

Genome

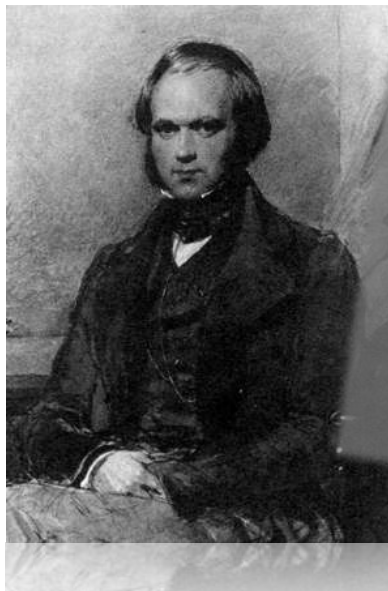
Walter Salzburger

Zoological Institute
University of Basel, Switzerland

A brief history

1859

Charles R. Darwin publishes “*On the origin of species by means of natural selection*” and establishes the theory of evolution



Charles R. Darwin (1809-1882)

1866

Gregor Mendel publishes “*Experiments in plant hybridization*”. This paper established what eventually became formalized as the Mendelian laws of inheritance.



Gregor Mendel
(1822-1884)



1866

Gregor Mendel publishes “*Experiments in plant hybridization*”. This paper established what eventually became formalized as the Mendelian laws of inheritance.



1869

Johann Friedrich Miescher extracts what comes to be known as DNA from the nuclei of white blood cells.



Johann F. Miescher (1844-1895)

1900

Independently of one another, **Hugo de Vries** (1848-1935), **Erich von Tschermak-Seysenegg** (1871-1962) and **Carl Correns** (1864-1933) rediscover Mendel's published, but long neglected, paper outlining the basic laws of inheritance.



Hugo de Vries



Erich v. Tschermack



Carl Correns

1902

Theodor Boveri and **Walter Sutton** propose that chromosomes bear hereditary factors in accordance with Mendelian laws.



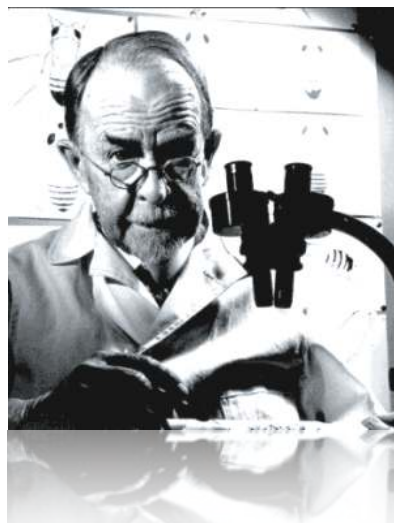
Walter Sutton
(1877-1916)



Theodor Boveri
(1862-1915)

1910

Thomas H. Morgan establishes the chromosomal theory of inheritance. He also discovered the recombination of homologous chromosomes during meiosis.



Thomas Hunt Morgan (1866-1945)

1944

Oswald T. Avery (1877-1955), **Maclyn McCarty** (1911-2005) and **Colin MacLeod** (1909-1972) identify deoxyribonucleic acid (DNA) as the “transforming principle”.



Oswald T. Avery
(1877-1955)

1950

Erwin Chargaff discovers regularity in proportions of DNA bases. In all organisms he studied, the amount of adenine (A) equaled that of thymine (T), and guanine (G) equaled cytosine (C).



Erwin Chargaff (1905-2002)

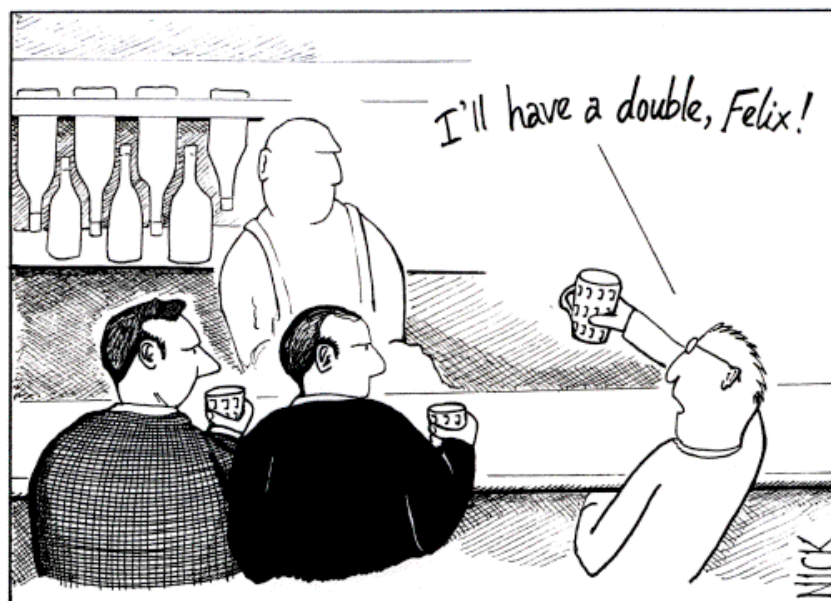
1953

James Watson and **Francis Crick** discover the double helical structure of the DNA and that this structure meets the unique requirements for a substance that encodes genetic information.



James D. Watson (1928-)
Francis H. C. Crick (1916-2004)

1953



Cambridge, 1953. Shortly before discovering the structure of DNA, Watson and Crick, depressed by their lack of progress, visit the local pub.

Watson and Crick, depressed by their lack of progress, visit the local pub in Cambridge, 1953, shortly before discovering the structure of DNA.

1960

Discovery of messenger RNA (mRNA) by **Sydney Brenner** (1927-), **Francis Crick** (1916-2004), **Francois Jacob** (1920-) and **Jacques Monod** (1910-1976).



1977

Frederick Sanger (1918-) and **Walter Gilbert** (1932) develop techniques for DNA sequencing

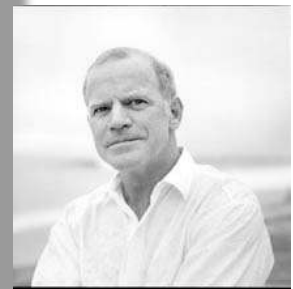


Walter Gilbert

Frederick Sanger

1983

Kary B. Mullis (1944-) invents and helps to develop the polymerase chain reaction (PCR)



Kary B. Mullis

1995

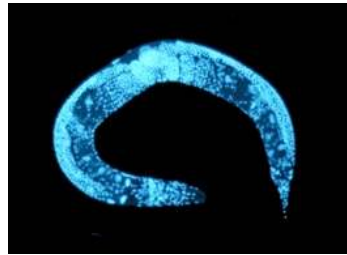
1,830,137 bp of *Hamophilus influenzae* sequenced: the first genome of a free living organisms determined



A brief history

1998

Caenorhabditis elegans sequenced



2000

Drosophila melanogaster sequenced

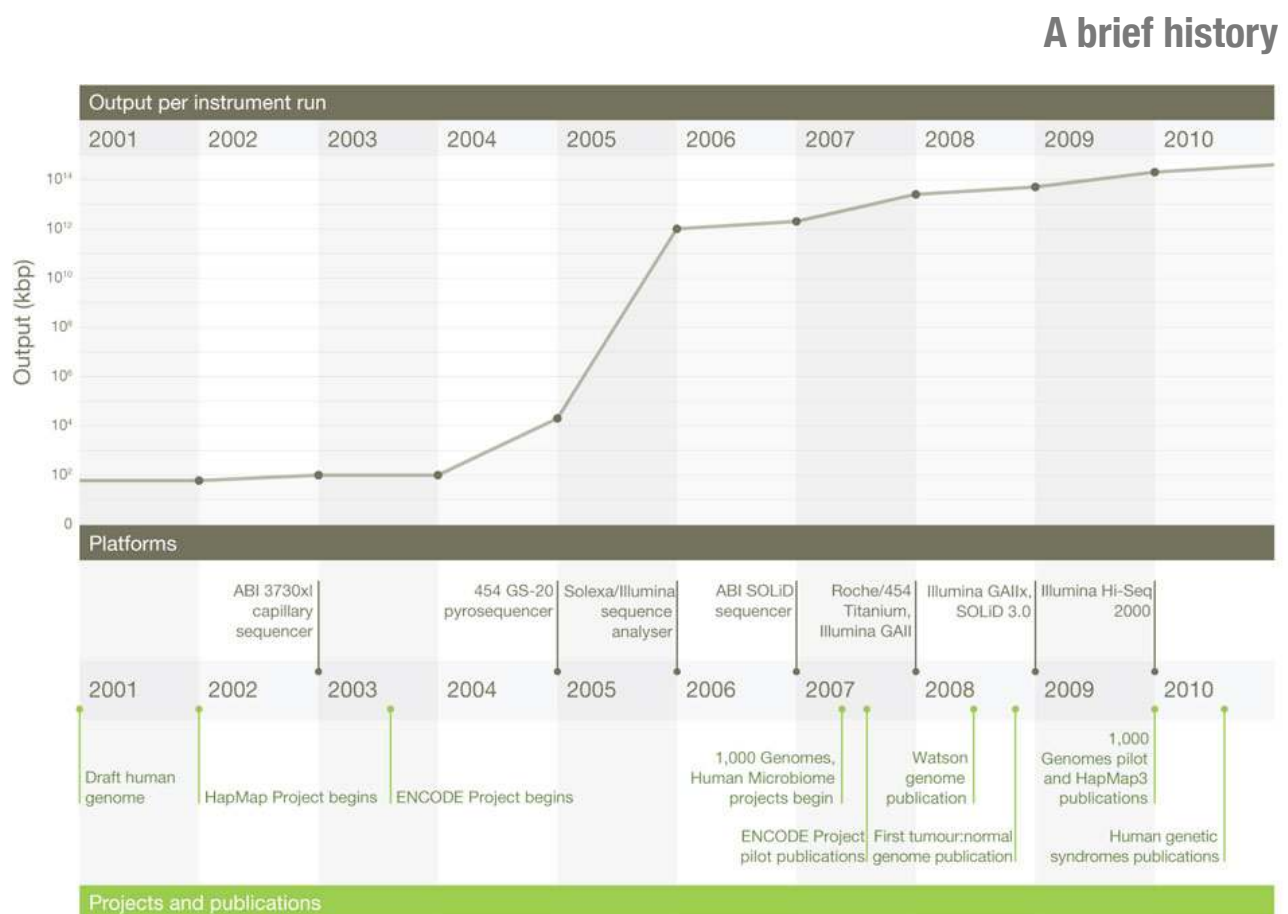


A brief history

2001

Homo sapiens sequenced

2001

Homo sapiens sequenced

Genome Size Evolution

Walter Salzburger

Genome Size Evolution

species	genome size (MB)	gene number
<i>Saccharomyces cerevisiae</i>	12.05	6213
<i>Plasmodium falciparum</i>	22.85	5268
<i>Trypanosoma spp.</i>	39.20	10000
<i>Aspergillus nidulans</i>	30.07	9541
<i>Dictyostelium doscoideum</i>	34	9000
<i>Arabidopsis thaliana</i>	125	25498
<i>Oryza sativa</i>	466	60256
<i>Lotus japonicus</i>	472	26000

Lynch (2006)

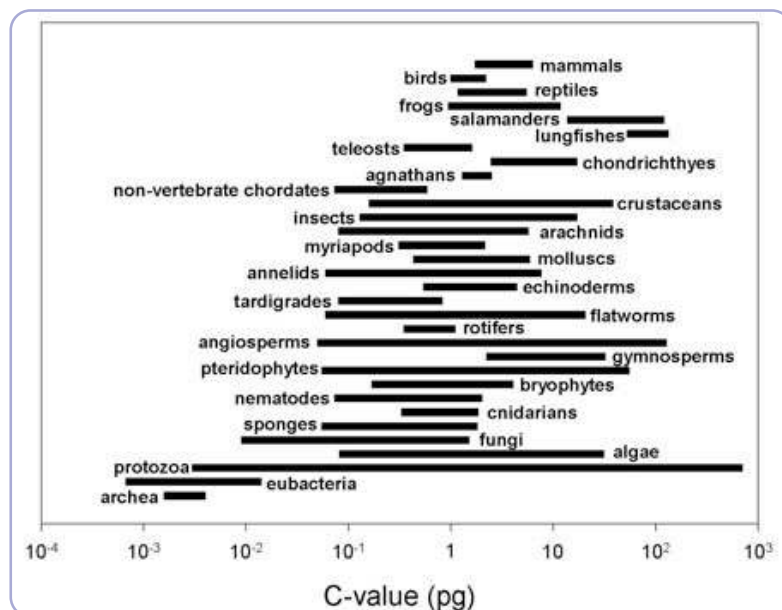
Genome Size Evolution

species	genome size (MB)	gene number
<i>Caenorhabditis elegans</i>	100.26	21200
<i>Drosophila melanogaster</i>	137	16000
<i>Ciona intestinalis</i>	156	16000
<i>Anopheles gambiae</i>	278	13683
<i>Fugu rubripes</i>	365	38000
<i>Gallus gallus</i>	1050	21500
<i>Mus musculus</i>	2500	24000
<i>Homo sapiens</i>	2900	24000

Lynch (2006)

Genome Size Evolution

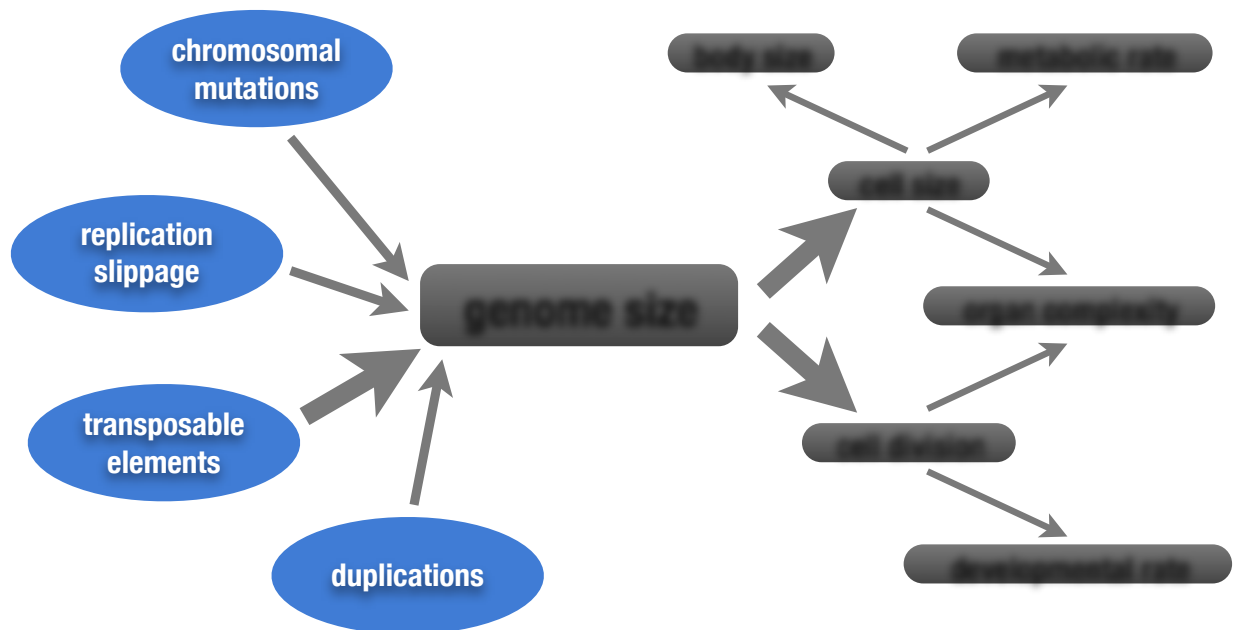
“C-value enigma”



Gregory (2004)

1 pg $\approx$ 1000 Mb = 1Gb

Genome Size Evolution



Lynch (2007)

Chromosomal Rearrangements

Chromosomal mutations

- ▶ Mutations that affect entire chromosomes or large parts thereof are called **chromosomal mutations**
- ▶ The phenotypic effects of chromosome mutations are **difficult to generalize**
- ▶ Phenotypic consequences also arise, as gene expression is at least partly regulated by the relations between neighboring genes

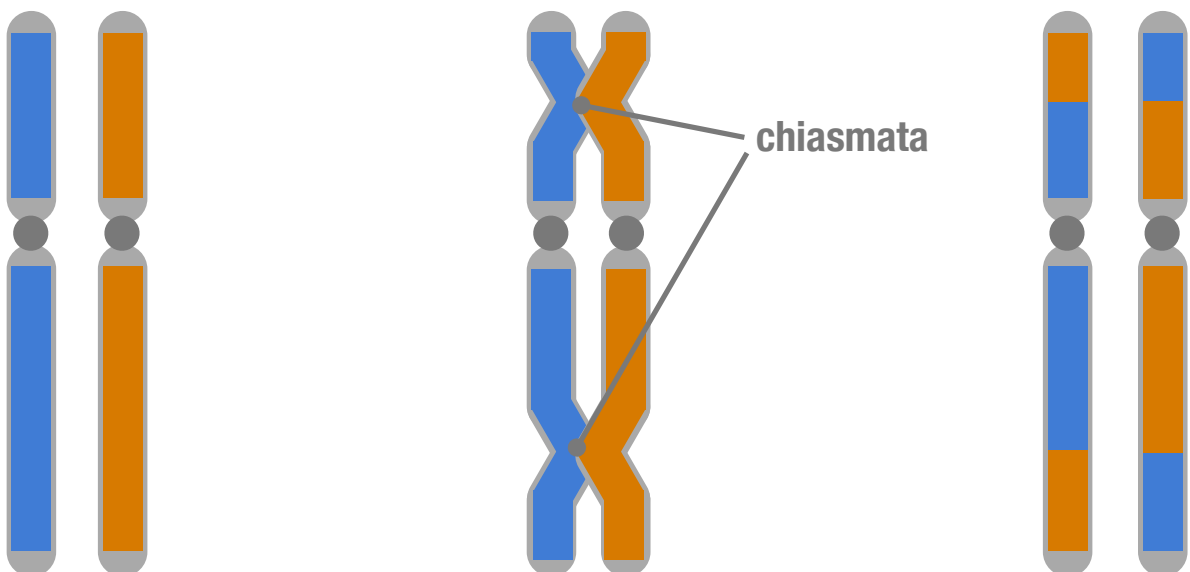
Chromosomal Rearrangements

Recombination...

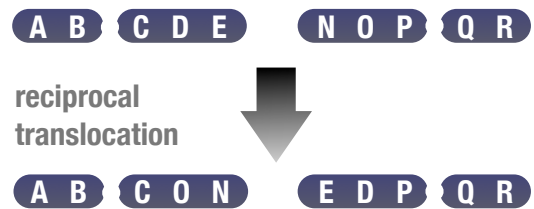
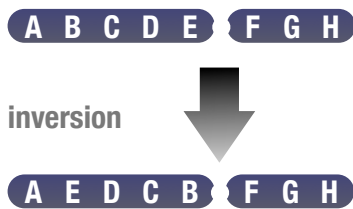
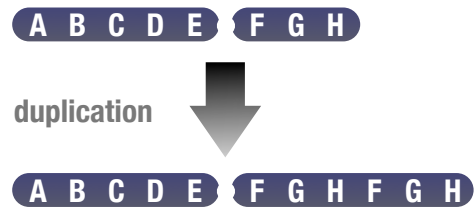
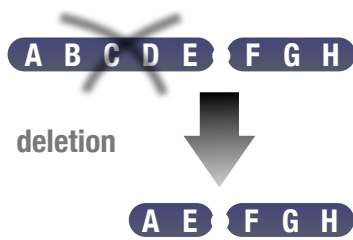
- ▶ ...occurs by the **crossing-over** of homologous chromosomes during **meiosis**
- ▶ ...leads to the **exchange of DNA** between a pair of chromosomes
- ▶ As a consequence, two previously unlinked genes may become linked or *vice versa*
- ▶ ...may lead to rearrangements

Chromosomal Rearrangements

Recombination

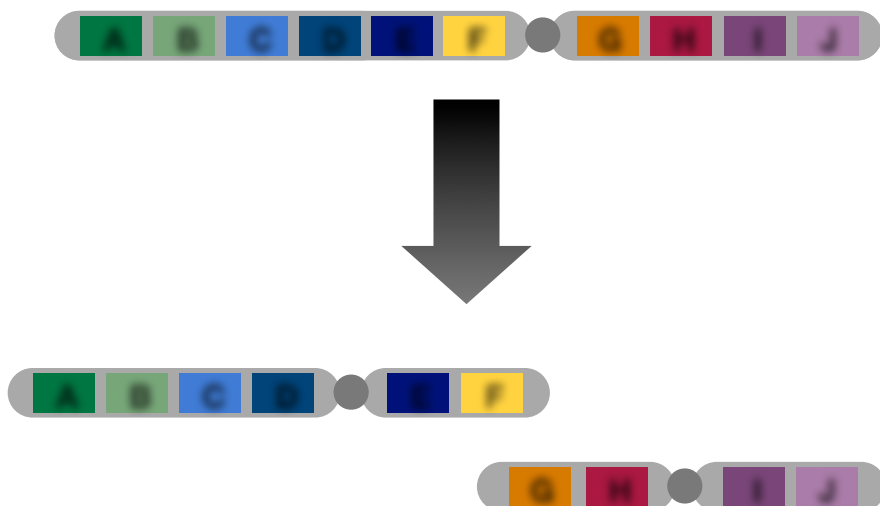


Chromosomal Rearrangements



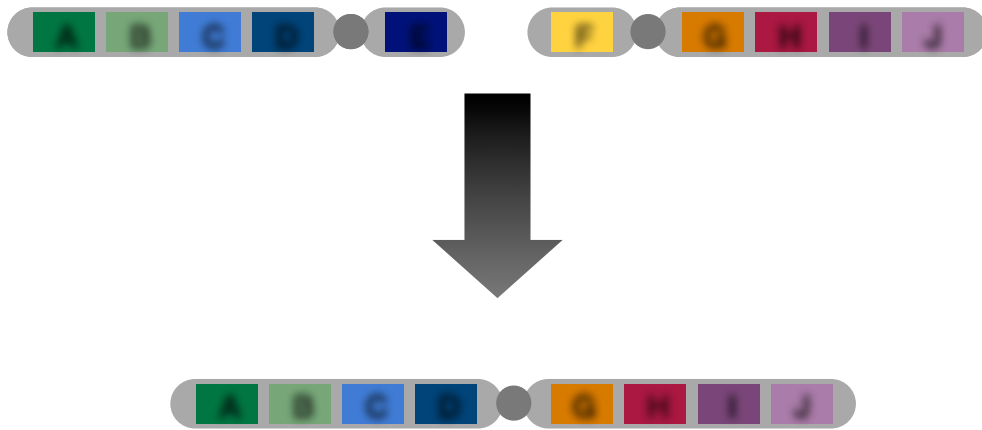
Chromosomal Rearrangements

Chromosomal Fission



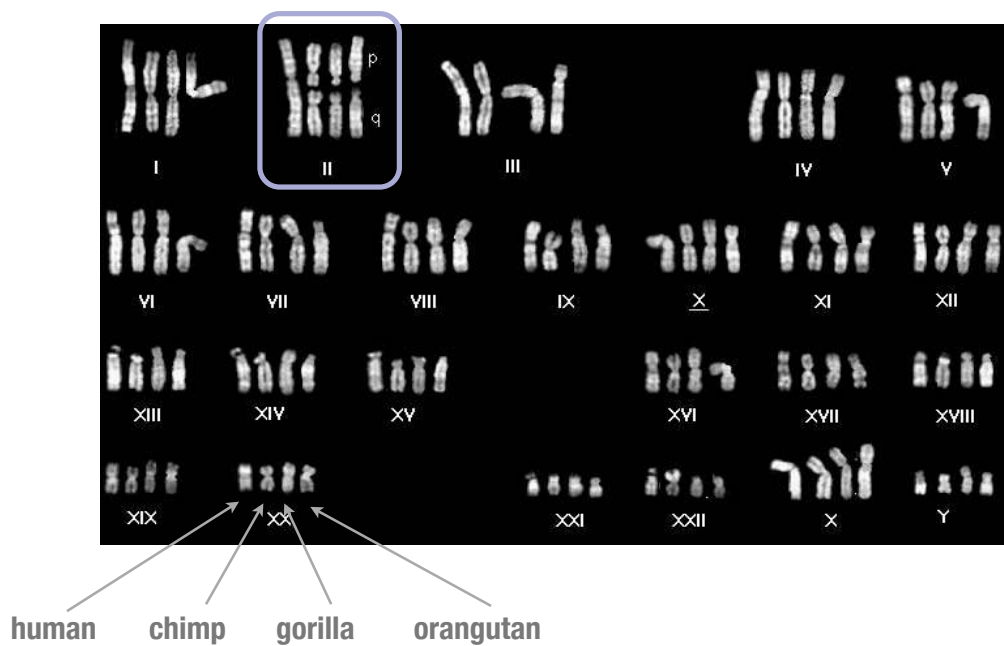
Chromosomal Rearrangements

Chromosomal Fusion



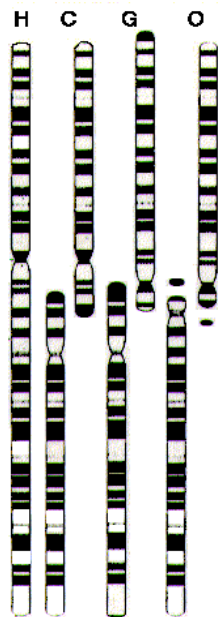
Chromosomal Rearrangements

Chromosomal fusion: human chromosome 2



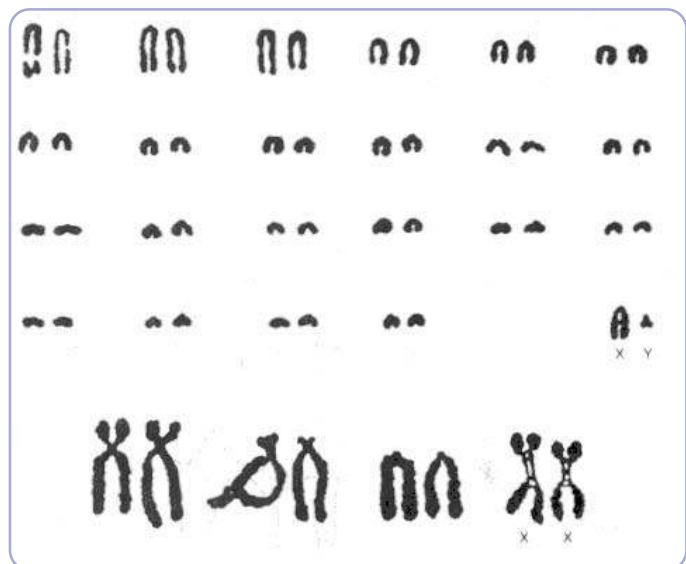
Chromosomal Rearrangements

Chromosomal fusion: human chromosome 2



Chromosomal Rearrangements

An extreme case: muntjak deers



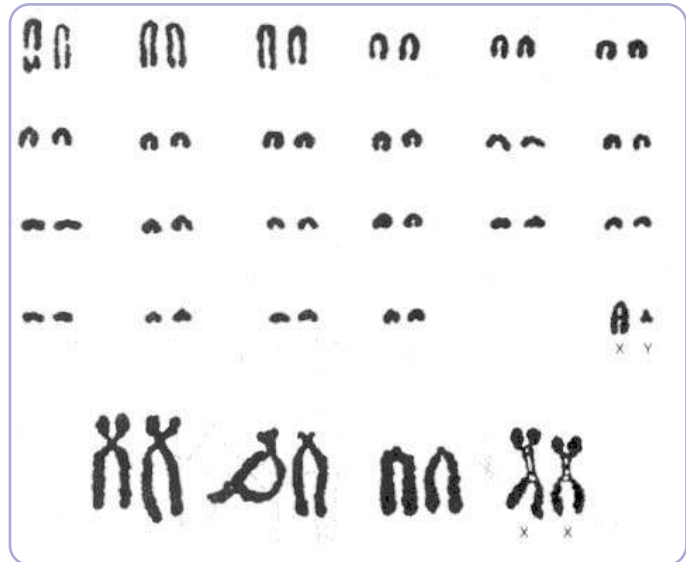
Chromosomal Rearrangements

An extreme case: muntjak deers

chinese muntjac deer

Translocation,
chromosome fusion,
and/or fission
explain why these
two very similar
species of hoofed
mammal have such
different karyotypes.

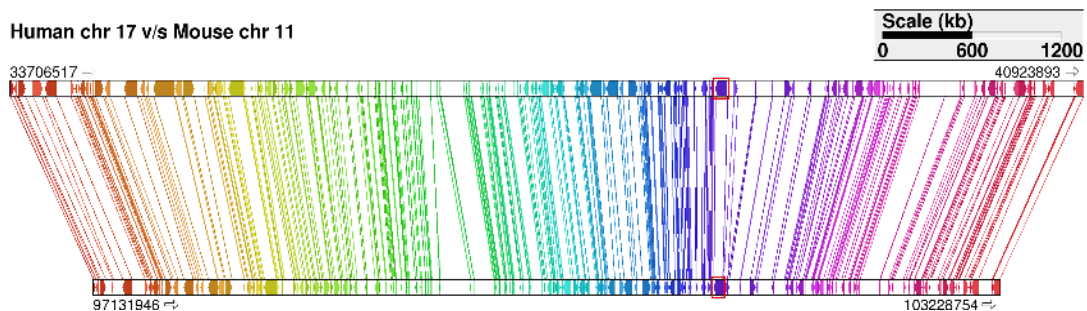
indian muntjac deer



Chromosomal Rearrangements

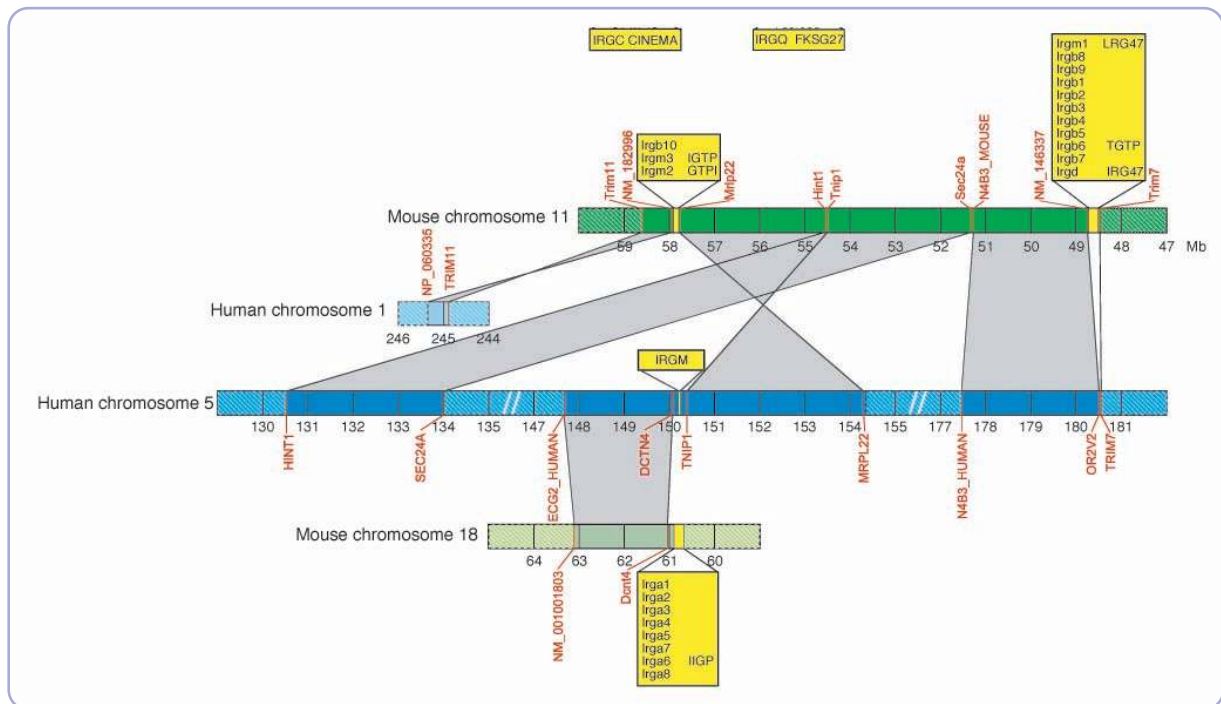
Synteny

- ▶ In comparative genomic terms, synteny describes the **preserved order of genes** on chromosomes as a result of common ancestry



Chromosomal Rearrangements

Synteny



Bekpen et al. (2005)

Gene and Genome Duplications

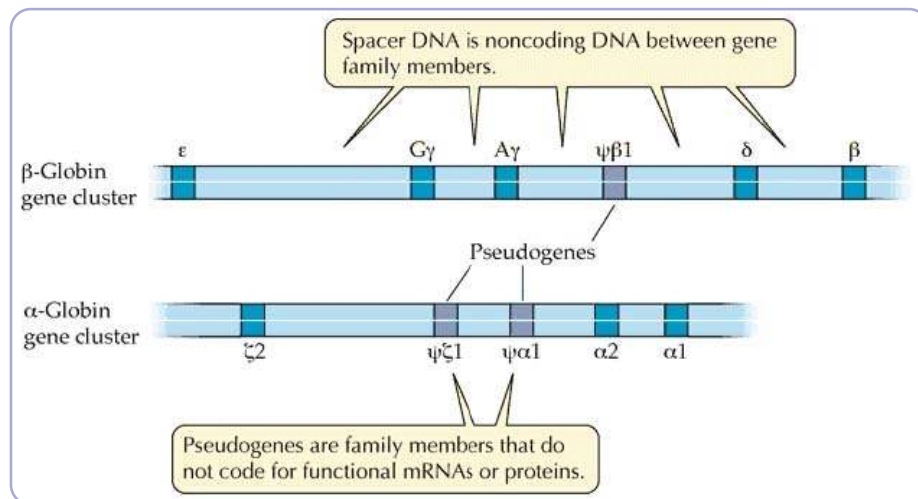
Gene duplication

- ▶ A gene can be duplicated by **various kinds of mechanisms**, e.g., genome duplications, chromosome mutations, unequal crossing over, etc.
- ▶ Any such duplication will be rare initially but may increase its frequency by **natural selection** or **random drift**.
- ▶ Duplicated genes may undergo different **evolutionary fates** such as non-functionalization, neo-functionalization or sub-functionalization.

Gene and Genome Duplications

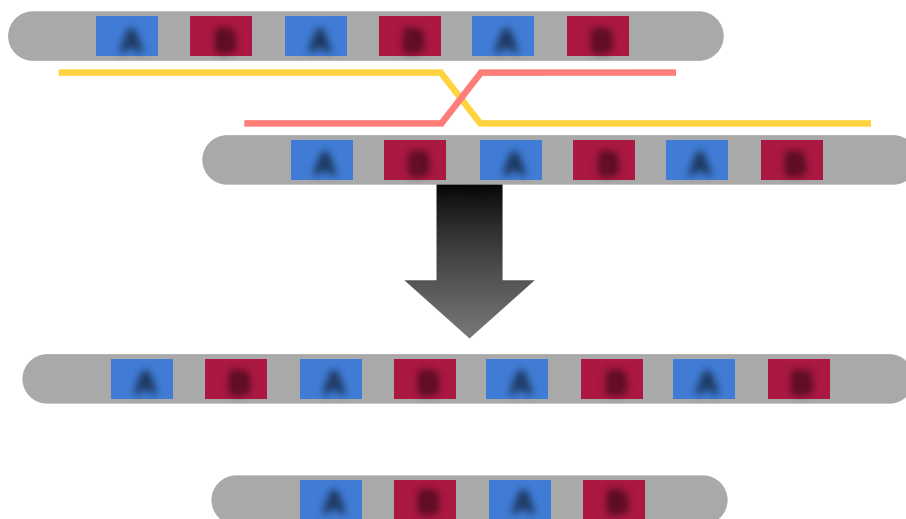
Gene clusters

- ...is a set of two or more genes of common ancestry that encode similar products



Gene and Genome Duplications

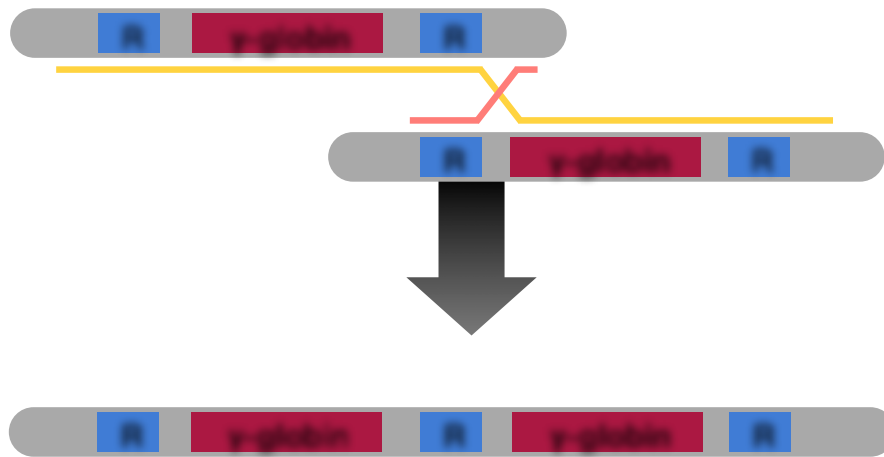
Unequal crossing over: misalignment



Misalignment is more likely when several copies of similar sequences are already present

Gene and Genome Duplications

Unequal crossing over: misalignment



Unequal crossing over is likely at repetitive sequence motives

Gene and Genome Duplications

Genome duplication

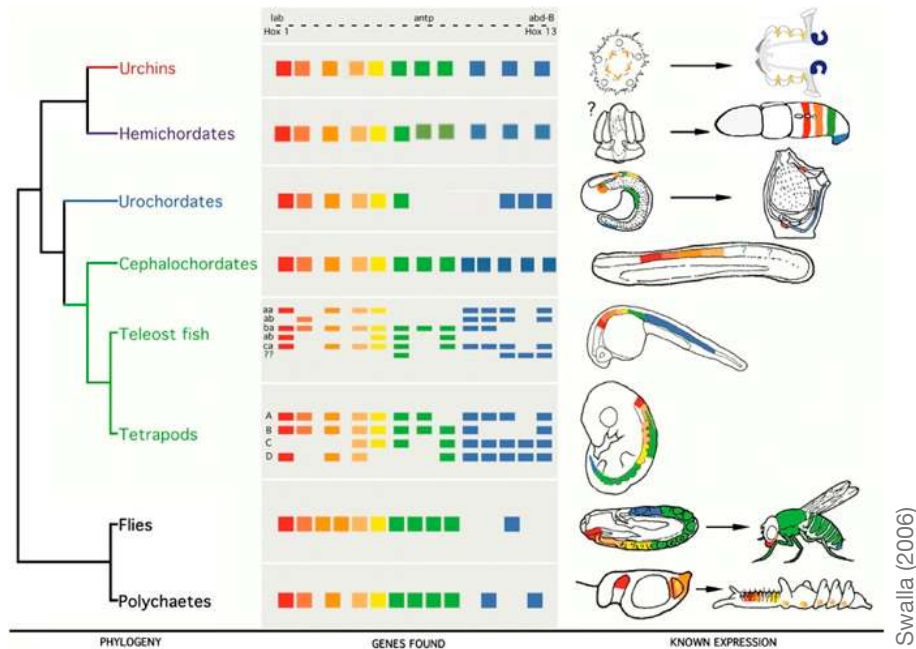
- ▶ Sosumo Ohno (1970):
Evolution by Gene Duplication
- ▶ Ohno also postulated that whole genome duplications exist and argued that the whole genome had duplicated twice near the origin of the vertebrates (**2R hypothesis**)
- ▶ Many more genome duplication events have been suggested since, e.g., the fish-specific genome duplication (**3R hypothesis**)



Sosumo Ohno
(1928-2000)

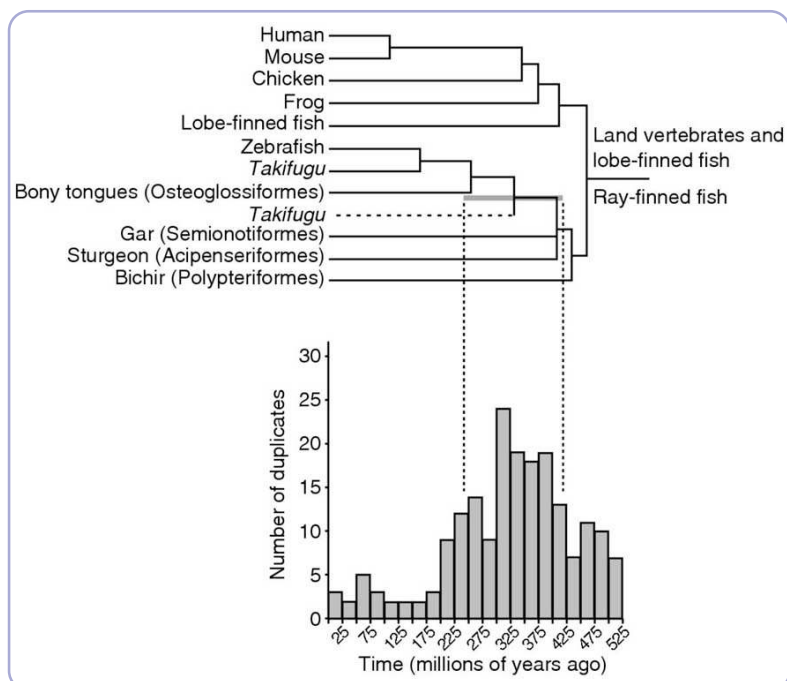
Gene and Genome Duplications

Hox gene clusters



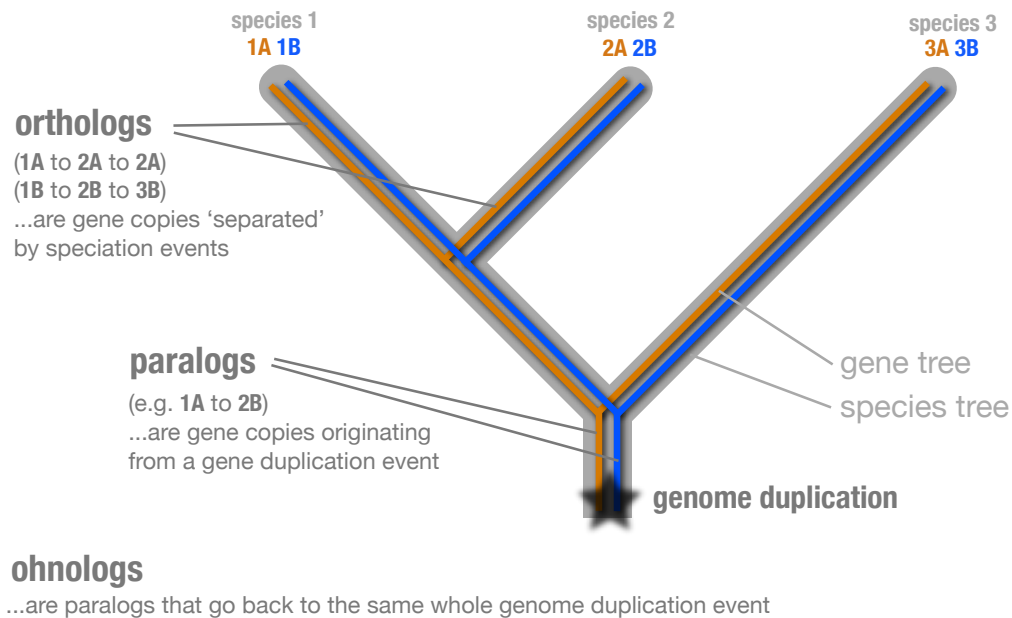
Gene and Genome Duplications

fish specific genome duplication



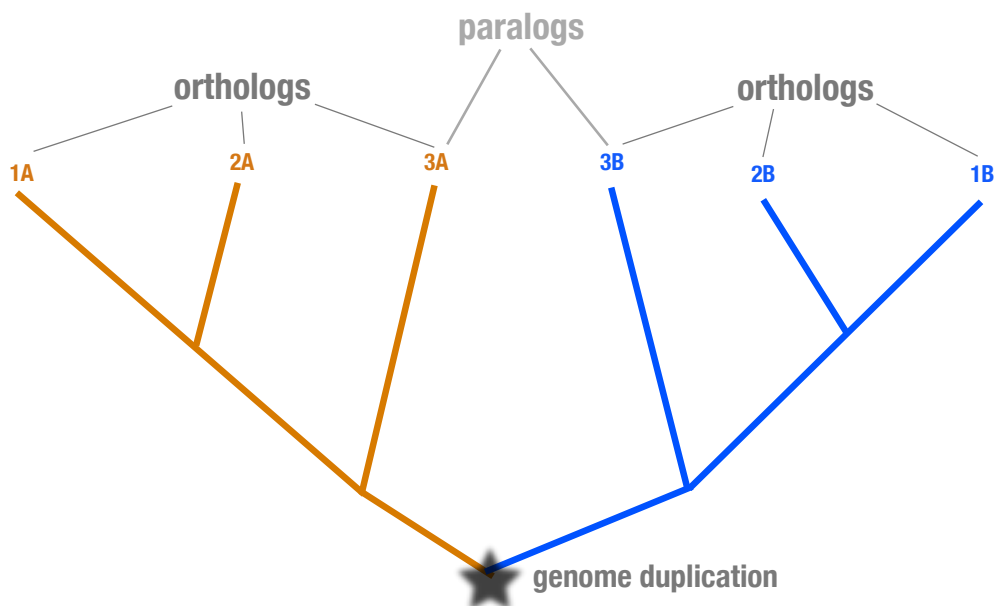
Gene and Genome Duplications

paralogs, orthologs and ohnologs



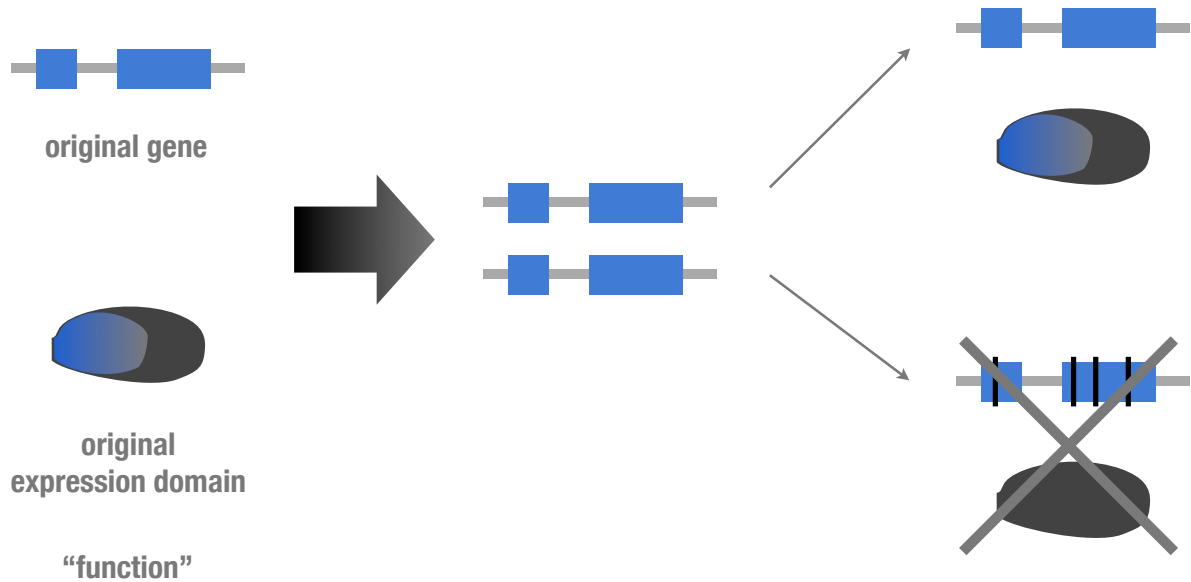
Gene and Genome Duplications

paralogs, orthologs and ohnologs



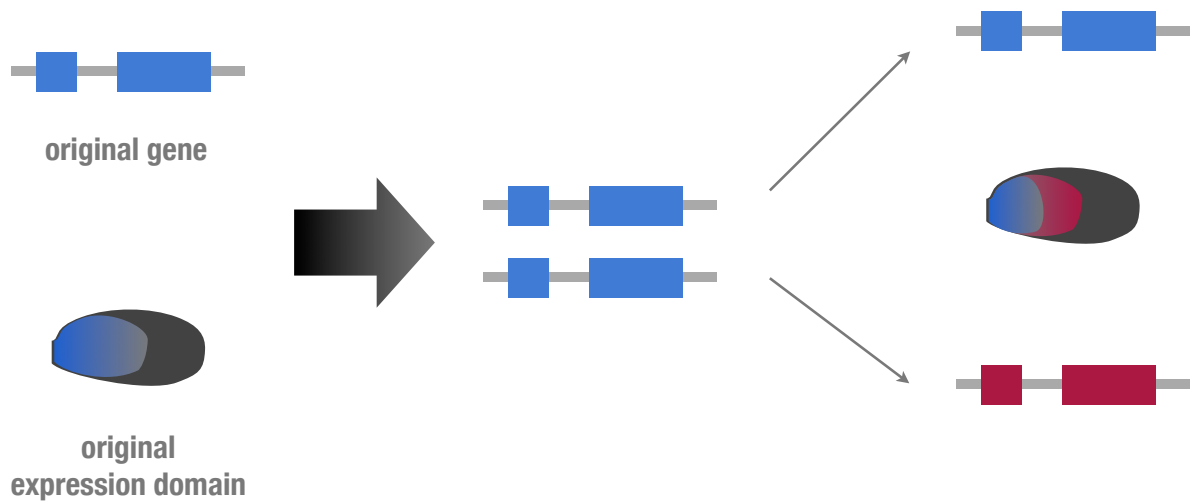
Gene and Genome Duplications

non-functionalization



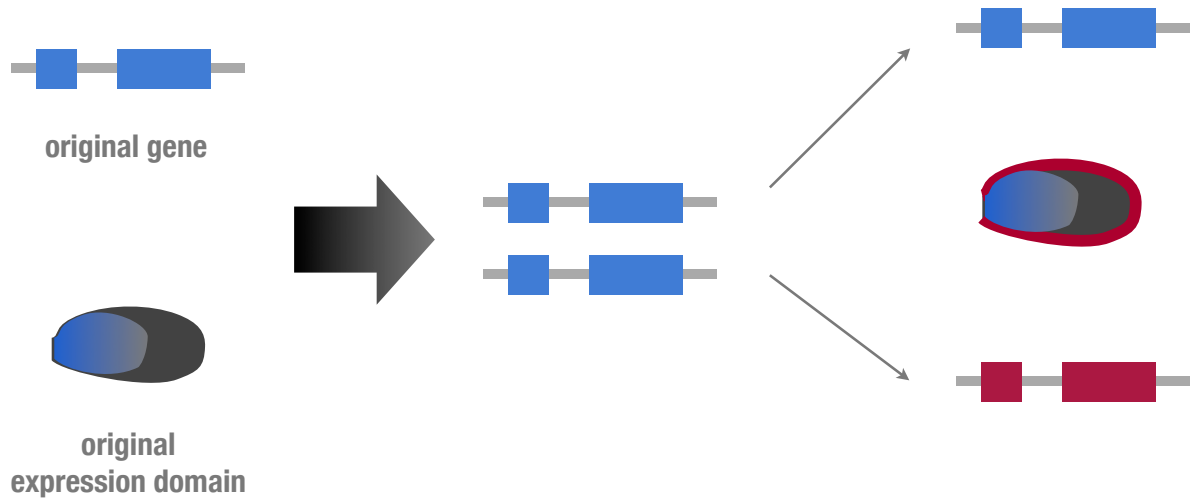
Gene and Genome Duplications

sub-functionalization



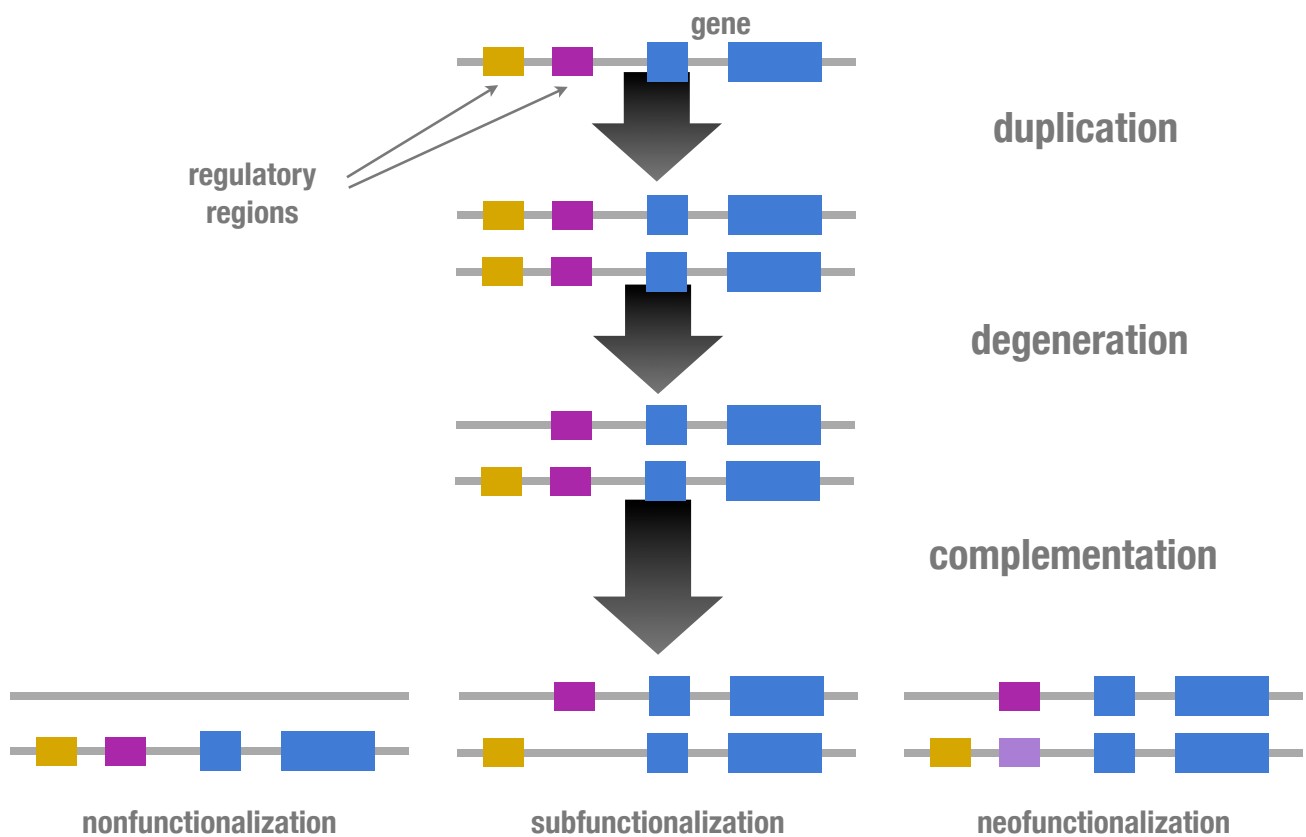
Gene and Genome Duplications

neo-functionalization



Gene and Genome Duplications

DDC model

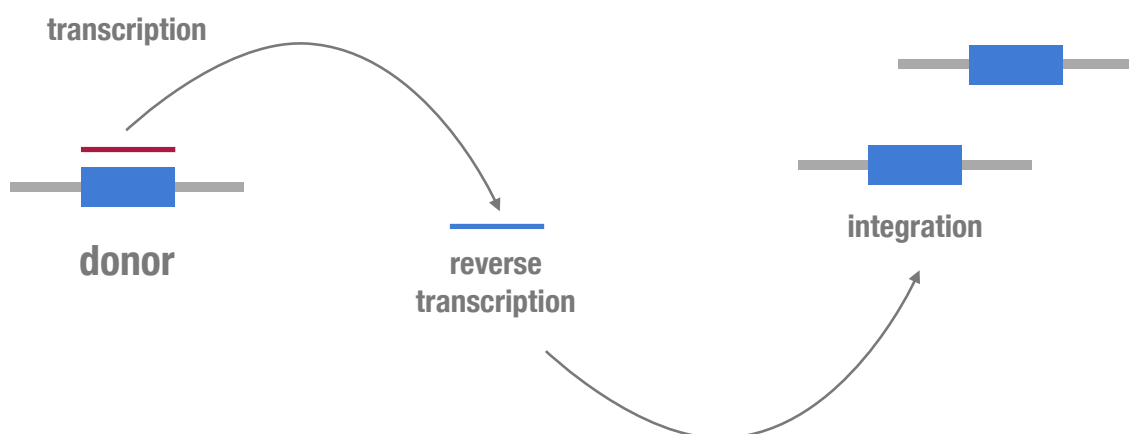


Transposable Elements

TEs

- ▶ Transposable elements (TEs) are discrete DNA sequences that move from one location to another within the genome
- ▶ TEs were discovered by Barbara McClintock (1902-1992) in the 1940s and 1950s. She received the Nobel Prize in 1983.
- ▶ TEs are found in nearly all species and constitute a large fraction of some genomes, including the human genome
- ▶ TEs can generate variation in the host genome

Transposable Elements



Transposable Elements

“retrotransposons”

LTRs

These transposable genetic elements are characterized by flanking **long terminal repeats**. LTRs are similar to retroviruses and contain a group specific antigen (*gag*).

LINEs

Long interspersed nuclear elements are autonomous retrotransposons. LINEs have two open reading frames (ORFs); ORF2 encodes for a reverse transcriptase (rvt)

SINEs

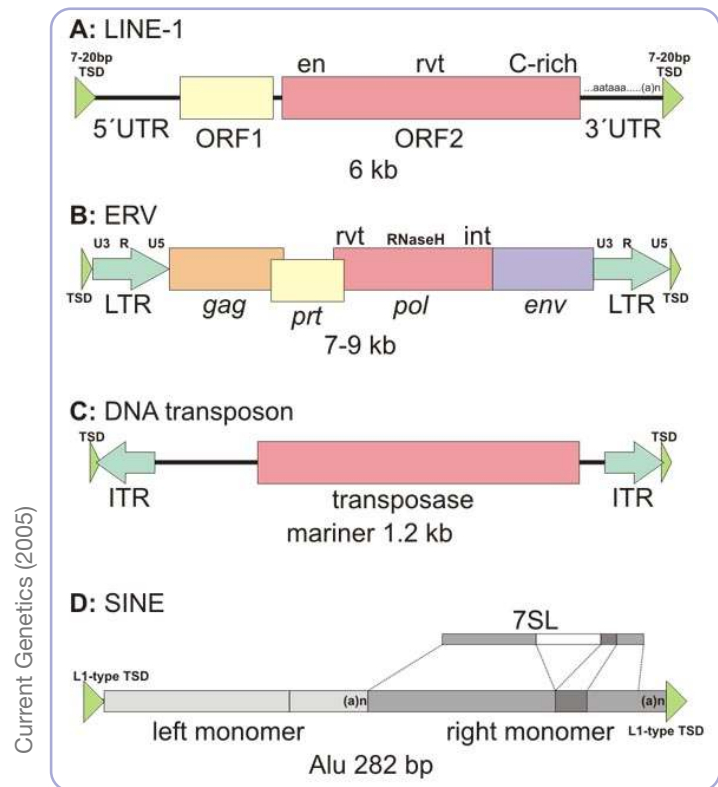
Short interspersed nuclear elements are non-autonomous retrotransposons that exploit the enzymatic retrotransposition machinery of LINEs

Transposable Elements

DNA transposons

- ▶ DNA transposons follow a different way of transposition and do not use an RNA stage and reverse transcription
- ▶ There are two classes of DNA transposons:
 - ▶ The majority of DNA transposons use a **cut-and-paste** mechanism
 - ▶ The other group uses a **rolling circle** (RC) mechanism

Transposable Elements



Transposable Elements

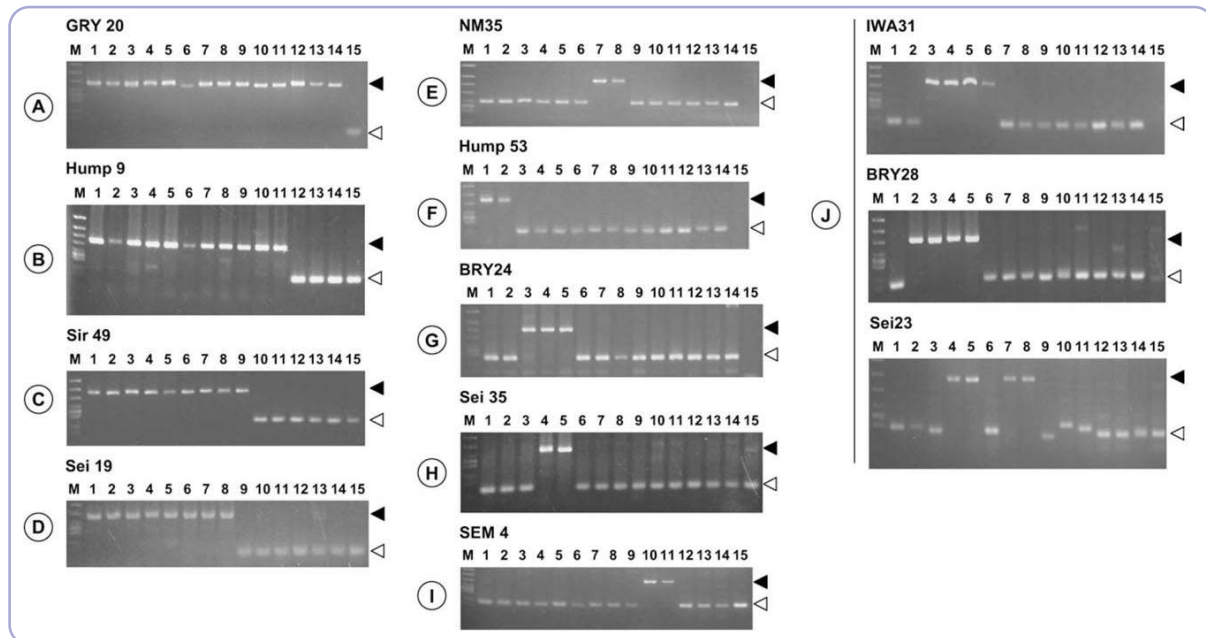
	yeast	slime mold	<i>C. elegans</i>	<i>Arabidopsis</i>	human
LTRs	3.1	4.4	0.1	6.4	7.9
SINEs, LINEs	0	3.7	0.4	0.7	31.2
DNA transposons	0	1.5	5.3	6.8	2.8
total	3.1	9.6	6.5	14	44.8

in % of genomes

Kidwell (2005)

Transposable Elements

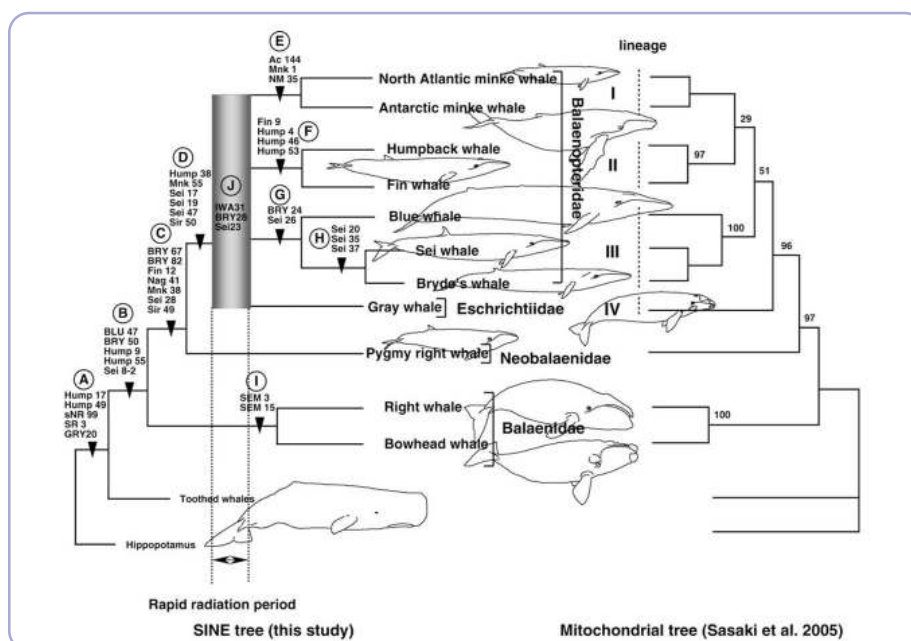
...as phylogenetic markers



Nikaido et al. (2005)

Transposable Elements

...as phylogenetic markers



Nikaido et al. (2005)

Other 'features'

gene deserts

- ▶ ...are long regions of more than 500 kb in length that are devoid of protein coding sequences (Venter et al. 2001)
- ▶ 25% of the human genome are organized in gene deserts
- ▶ Gene deserts are primarily found in mammalian genomes
- ▶ Some gene deserts are conserved over long evolutionary times. Some gene deserts have been deleted in mouse without any noticeable effect.

Other 'features'

"junk" DNA...

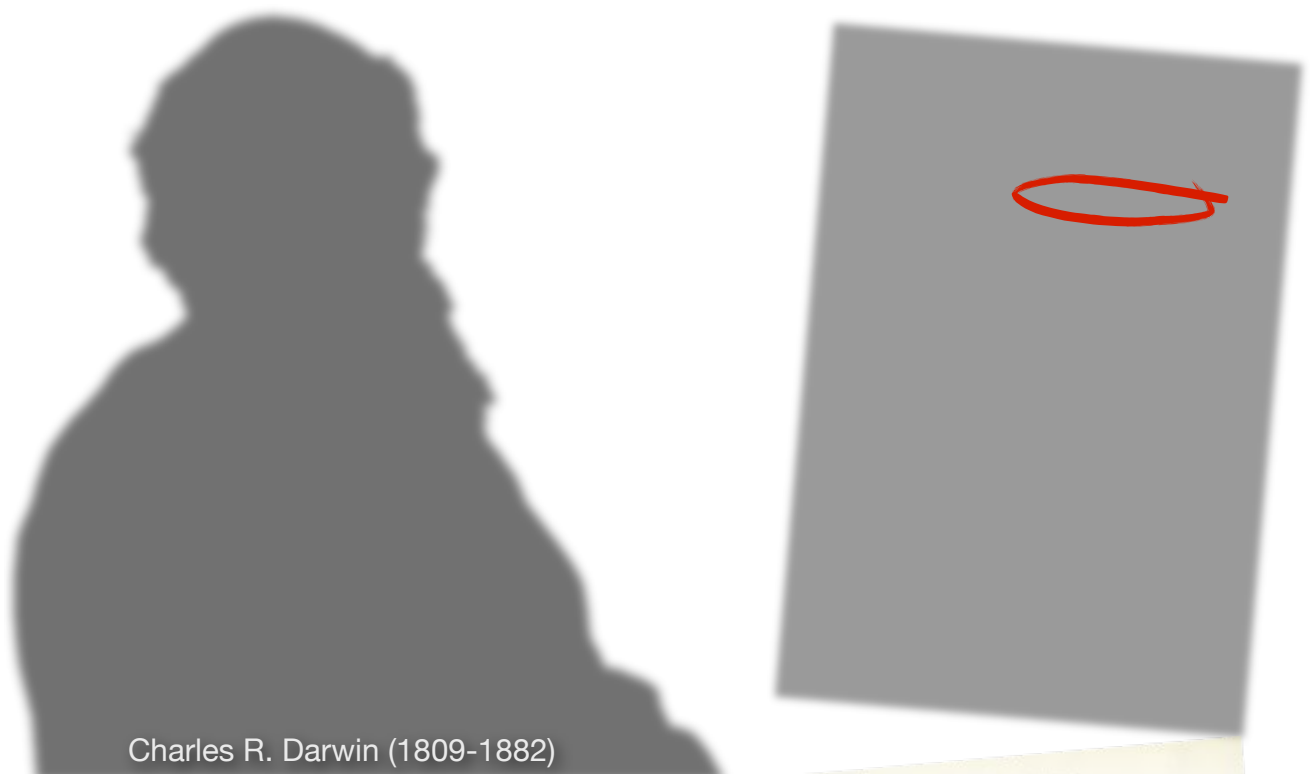
- ▶ ...is widely used as descriptor of noncoding DNA
- ▶ ...has originally been used to refer to pseudogenes, *i.e.* defunct copies of protein-coding genes (Ohno 1970)
- ▶ S. Ohno (1972): "The earth is strewn with fossil remains of extinct species; is it any wonder that our genome too is filled with the remains of extinct genes?"

Genome Evolution & Phenotypic Diversity

Walter Salzburger

Zoological Institute
University of Basel, Switzerland

Natural Selection...



Charles R. Darwin (1809-1882)

Natural Selection...

...Sexual Selection

	fitness	competitors
sexual selection	individual fitness	other members of the same sex
natural selection	fitness of the genotype	other individuals in the same population

Natural Selection...

- ▶ ... “is the process by which the forms of organisms in a population that are **best adapted** to the environment **increase in frequency** relative to less well-adapted forms over a **number of generations**” (Ridley 1996)



Sexual Selection...

- ▶ ...“is the selection on mating behavior, either through **competition among members of one sex** (usually males) for access to members of the other sex or through **choice by members of one sex** (usually females) for certain members of the other sex” (Ridley 1996)



Natural Selection...

...Sexual Selection

- ▶ ...**operates** if the following **conditions** are met:

reproduction

organisms must reproduce to form new generations

heredity

offspring resemble parents
("like must produce like")

trait variation

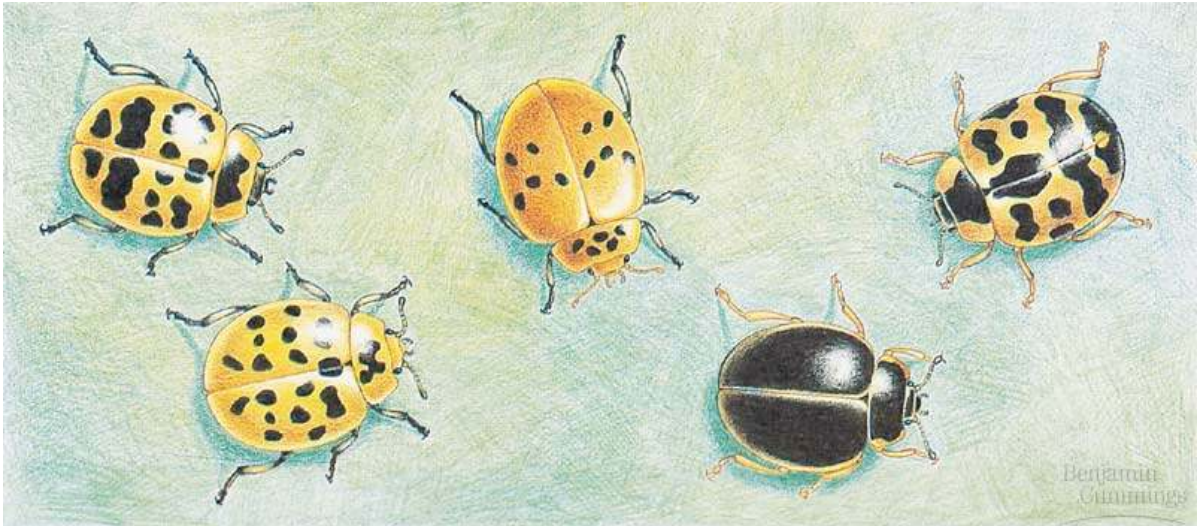
individuals in natural populations vary in
(adaptive) traits

variation in fitness

individuals in natural populations vary in the
number of their offspring that survive to
reproduce ('lifetime reproductive success')

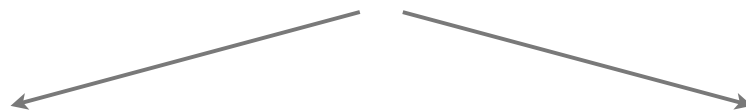
Natural Variation

- ▶ **Natural populations** show **variation** at all levels, from gross morphology to DNA sequences. Selection can only operate, if heritable variation exists.



Natural Variation

- ▶ **Natural populations** show **variation** at all levels, from gross morphology to DNA sequences. Selection can only operate, if heritable variation exists.
- ▶ Natural variation is generated by two processes:



recombination

“reshuffling” of genetic material by introducing or breaking up physical linkage

mutation

generation of new genetic variation by “mistakes” during the copying of a DNA strand

Natural Variation

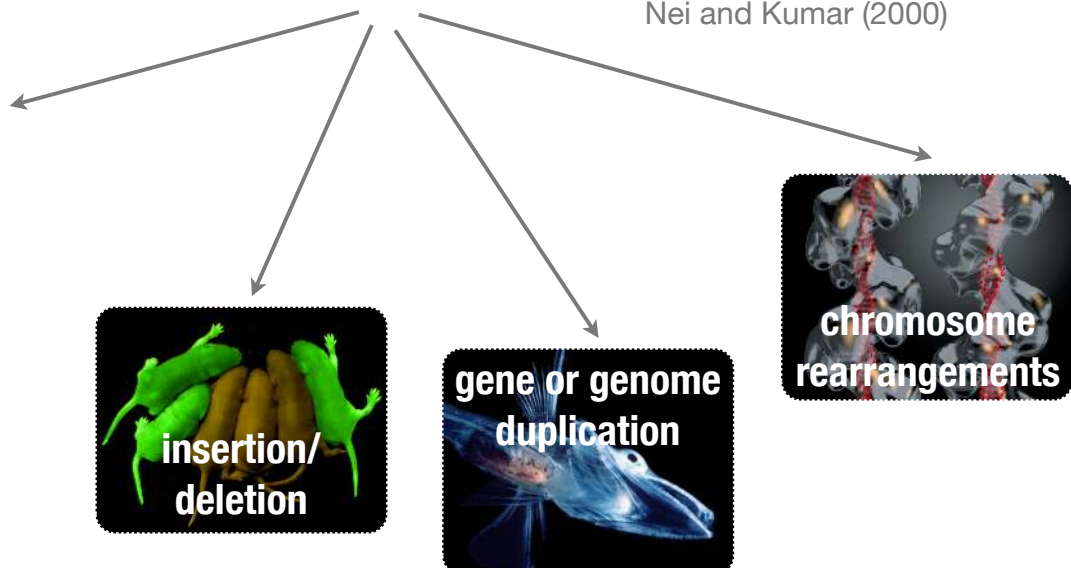
- ▶ New mutations are only transmitted to the next generation, if they occur in **germinal tissue**!



Phenotypic Evolution

“The primary cause of evolution is the mutational change of genes”

Nei and Kumar (2000)



Phenotypic Evolution

trait-based



sequence-based



Phenotypic Evolution

variation in:

phenome

genome

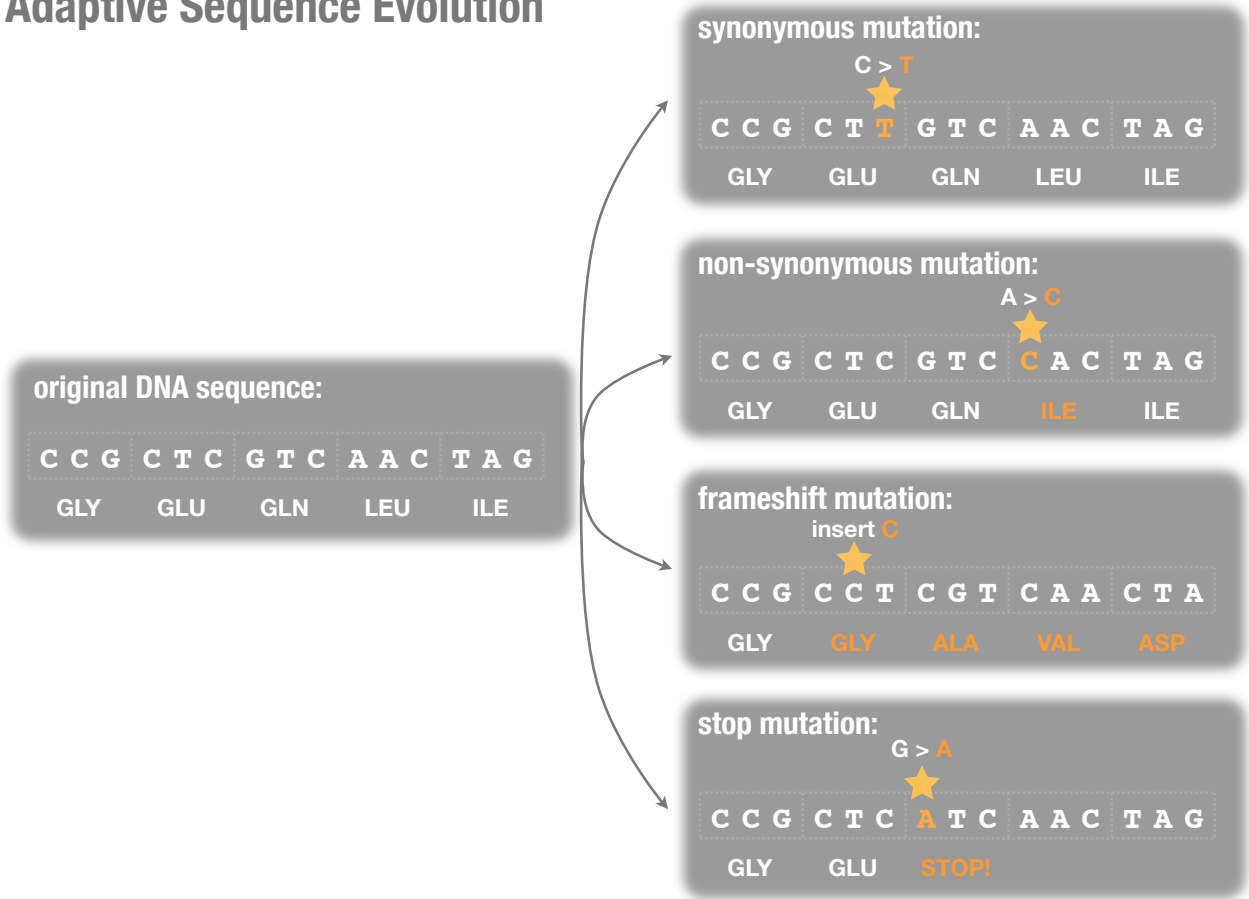


T
Val

C
Ala



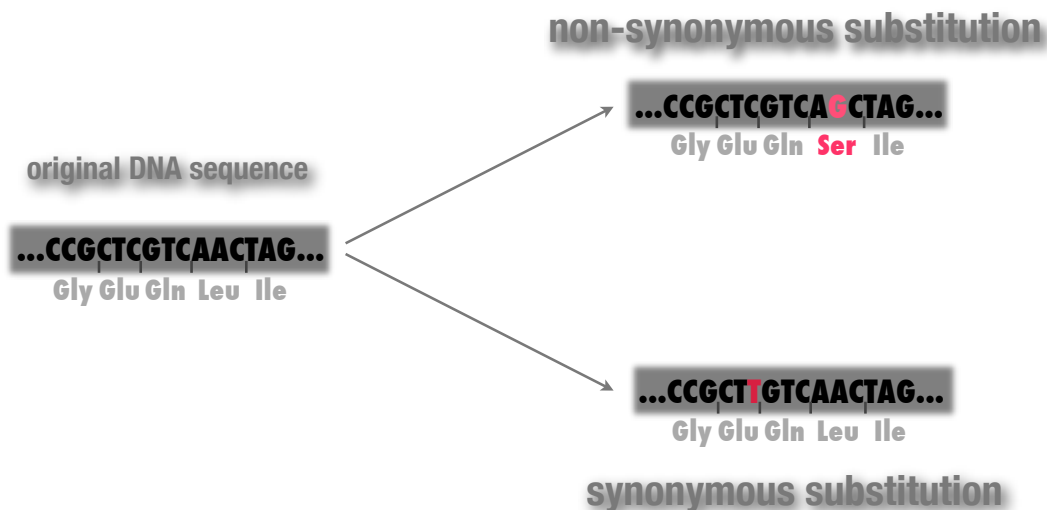
Adaptive Sequence Evolution



Adaptive Sequence Evolution



Adaptive Sequence Evolution



Adaptive Sequence Evolution

dN/dS ratio

- ▶ dN/dS (Ka/Ks) provides information about the form of sequence evolution in a gene*:
 - ▶ **$dN/dS < 1$: purifying selection** (non-synonymous sites are evolving slower than synonymous sites)
 - ▶ **$dN/dS = 1$: neutrality** (non-synonymous and synonymous sites are evolving at equal rates)
 - ▶ **$dN/dS > 1$: positive selection** (non-synonymous sites are evolving faster than synonymous sites; natural selection has acted to change the amino acid sequence)

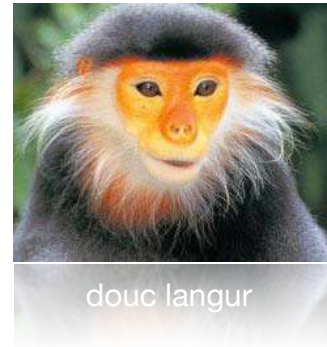
* dS is used as proxy for the neutral rate of molecular evolution

Adaptive Sequence Evolution

Pancreatic ribonuclease in colobine monkeys



Zhang et al. (2002)



Adaptive Sequence Evolution

Pancreatic ribonuclease in colobine monkeys

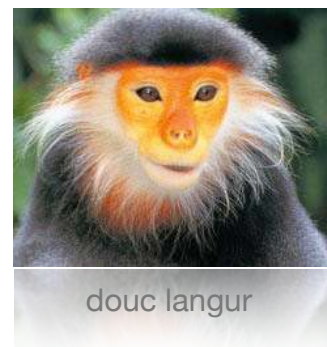
$$K_a = 0.0310$$

substitutions per non-synonymous sites

$$K_s = 0.0077$$

substitutions per synonymous sites
(and non-coding sites)

$$K_a/K_s = 4.026$$

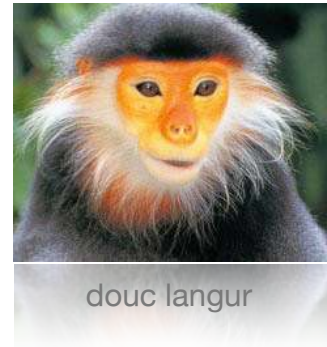


Adaptive Sequence Evolution

Pancreatic ribonuclease in colobine monkeys



Zhang et al. (2002)



douc langur

adaptation to different pH in the colobine small intestine

Adaptive Sequence Evolution

Pancreatic ribonuclease in leaf-eating monkeys



Prud'homme & Carroll (2006)



douc langur



guereza

Adaptive Sequence Evolution

Pancreatic ribonuclease in leaf-eating monkeys



Zhang (2006)



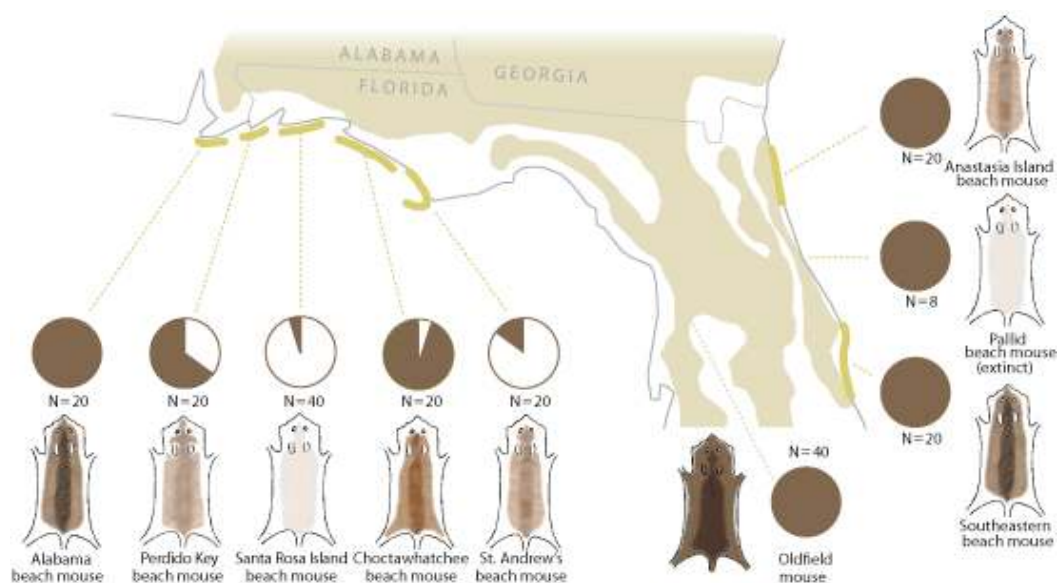
douc langur



guereza

Coding Sequence Evolution

beach mice



Hoekstra et al. (2006)

Coding Sequence Evolution

beach mice



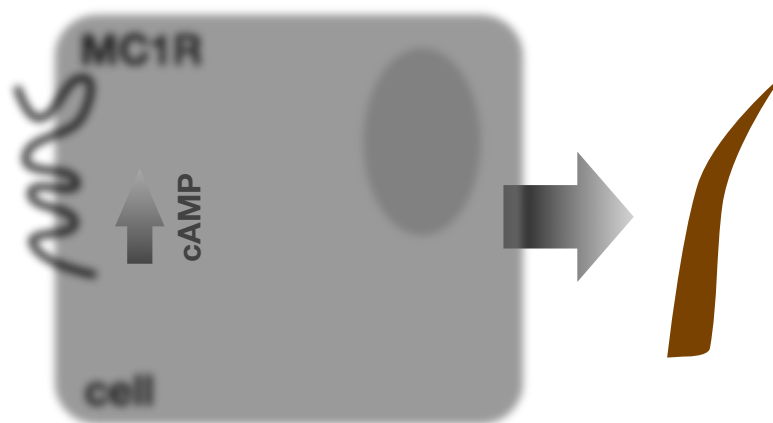
Nachman et al. (2003)

Coding Sequence Evolution

melanocortin 1 receptor

eumelanin

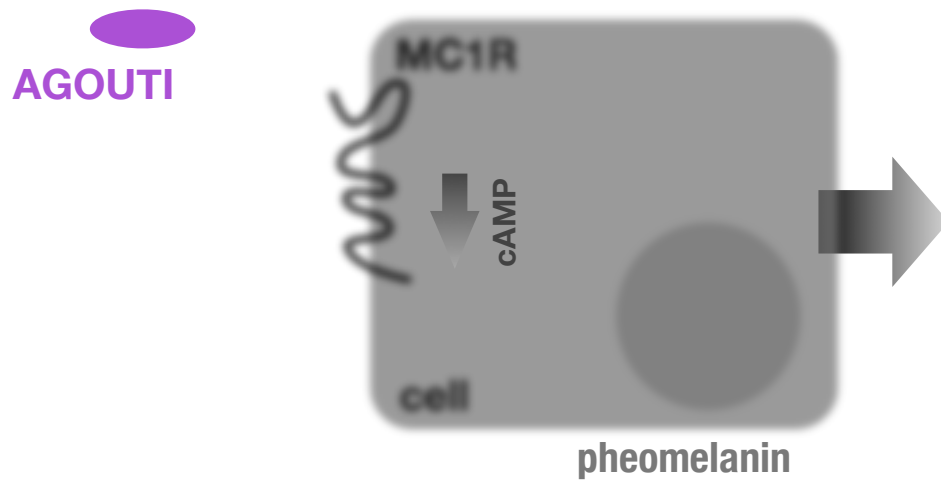
MSH



Hoekstra & Nachman (2005)

Coding Sequence Evolution

melanocortin 1 receptor



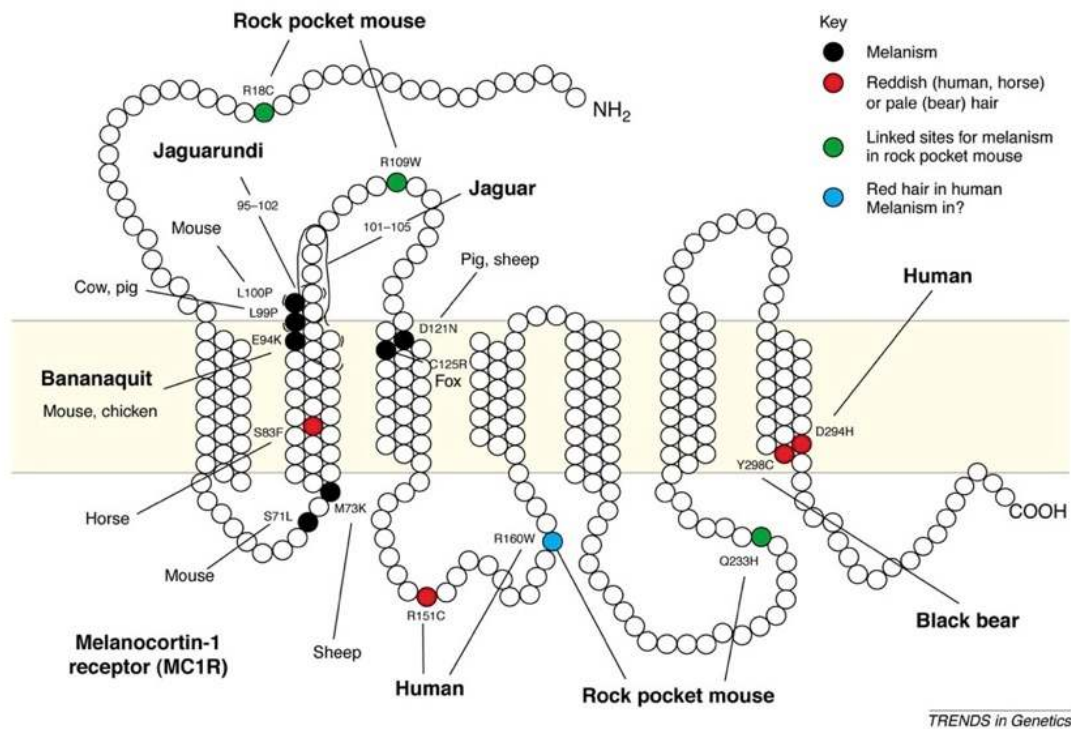
Hoekstra & Nachman (2005)

Coding Sequence Evolution

phenotype	<i>mc1r</i> alignment position								
	52	205	313	325	478	633	699	720	819
dark (1371)	T	C	G	T	T	C	C	T	C
	.	.	.	.	.	.	.	.	.
dark (1373)	.	.	T	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
dark (1384)	.	.	T	.	.	.	.	.	.
	C	.	.	C	C	T	A	.	.
dark (1412)	.	.	.	.	.	.	.	.	.
	C	T	.	C	C	T	A	.	.
light (1387)	C	.	.	C	C	T	A	G	T
	C	.	.	C	C	T	A	G	T
light (1403)	C	.	.	C	C	T	A	.	.
	C	.	.	C	C	T	A	.	.

Nachman et al. (2003)

Coding Sequence Evolution



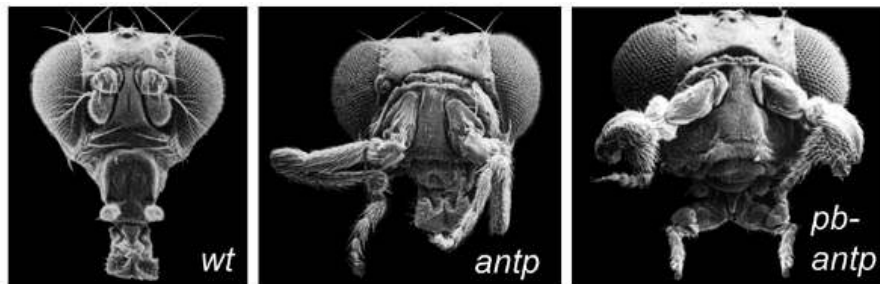
Regulatory Evolution

Hox genes

- ▶ **=homeotic genes:** genes that regulate the identity of body regions
- ▶ Hox genes contain a **homeobox**, a 180 bp region of DNA encoding a particular class of DNA-binding domains. Approximately 20 families of homeobox-containing genes exist.
- ▶ The homeobox encodes the **homeodomain**
- ▶ Mutations in homeotic genes cause the transformation of one body region or part into the likeness of another

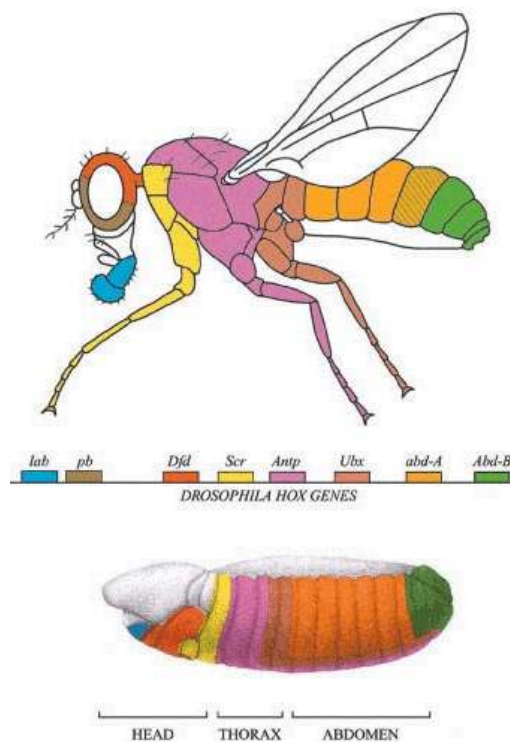
Regulatory Evolution

Hox genes, *Drosophila*



Regulatory Evolution

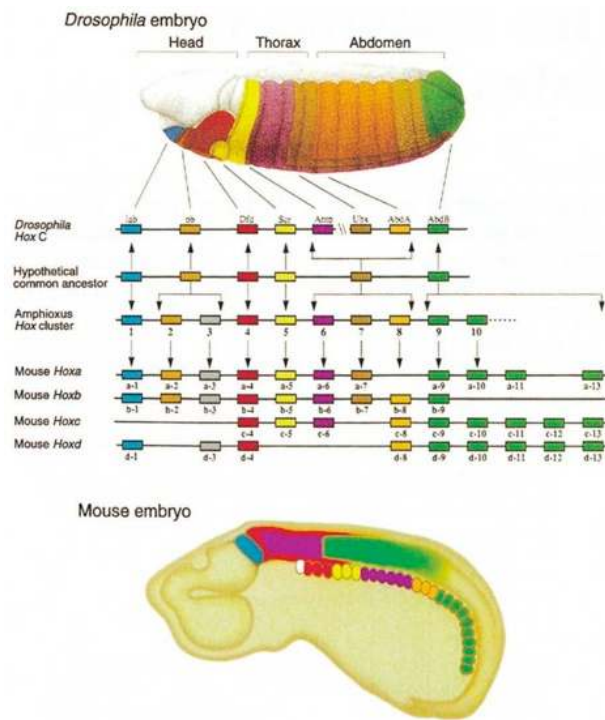
Hox genes and animal body plan



Carroll et al. (2001)

Regulatory Evolution

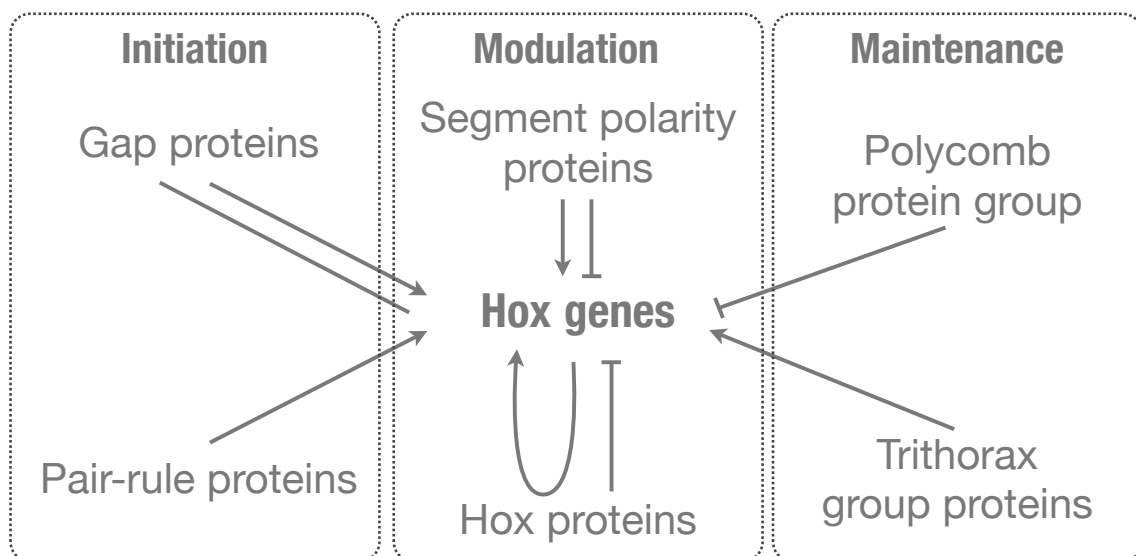
Hox genes and animal body plan



Carroll et al. (2001)

Regulatory Evolution

Hox gene regulation



Regulatory Evolution

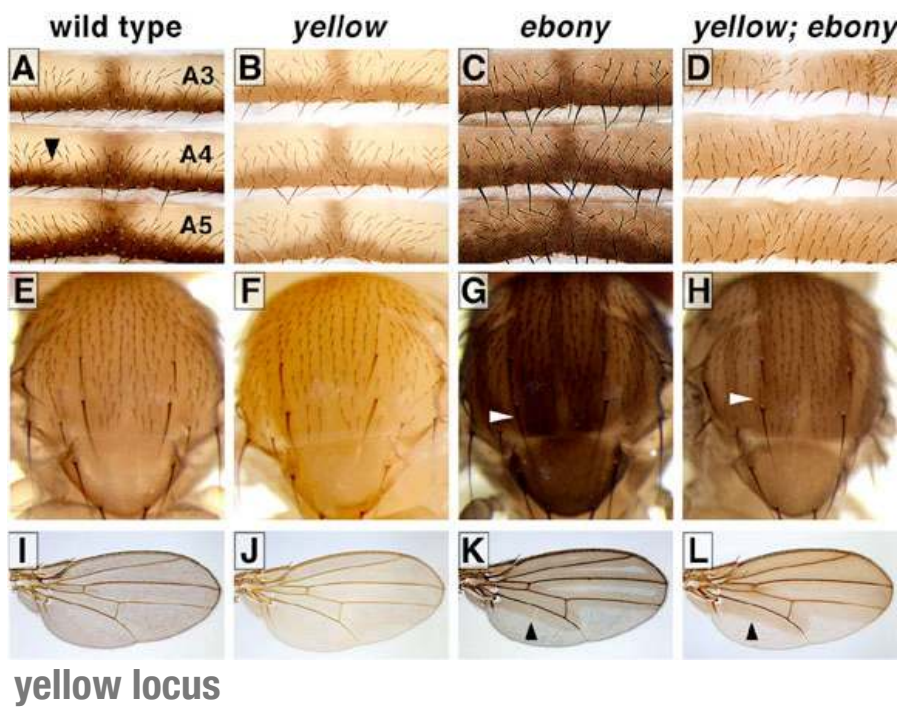
cis-regulatory evolution, *Drosophila*



Gompel et al. 2005

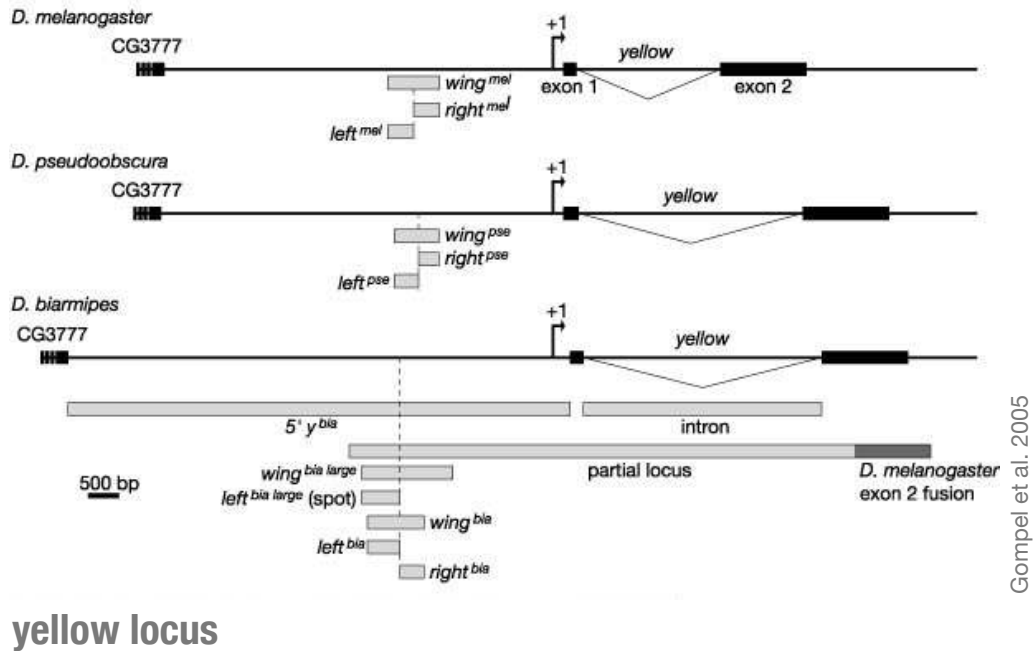
Regulatory Evolution

cis-regulatory evolution, *Drosophila*



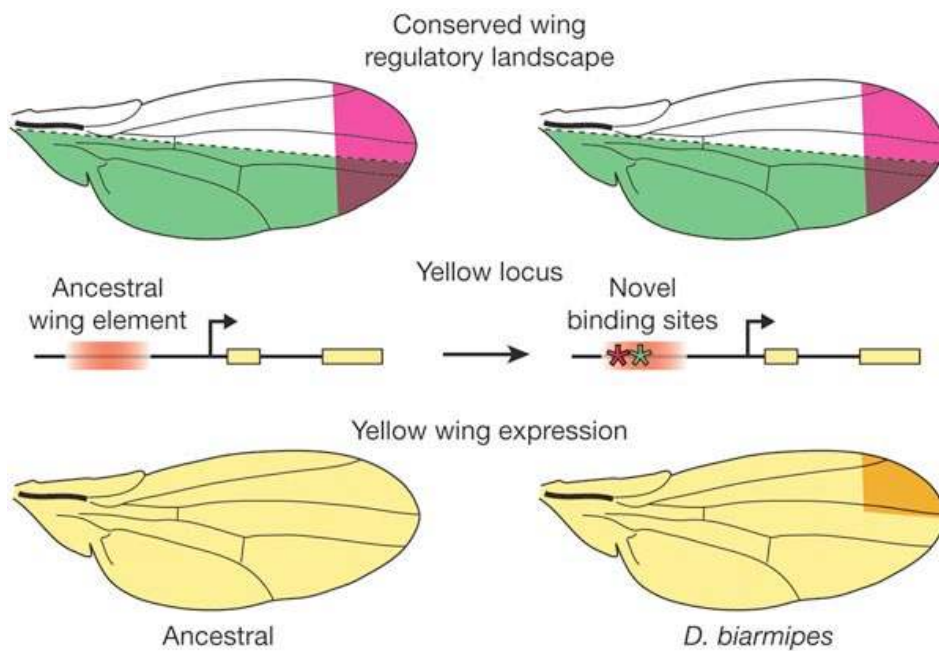
Regulatory Evolution

cis-regulatory evolution, *Drosophila*



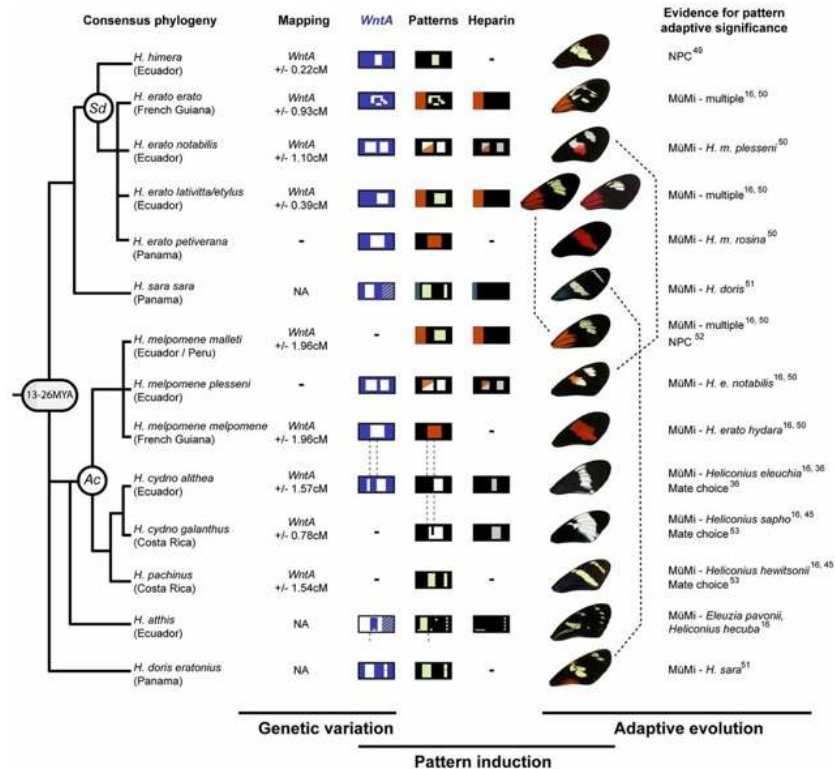
Regulatory Evolution

cis-regulatory evolution, *Drosophila*



Regulatory Evolution

cis-regulatory evolution, *Heliconius*



Martin et al. 2012

Regulatory Evolution

explosive speciation in cichlids



Regulatory Evolution

transcriptome comparison in cichlids

		
total number of reads	647'219	647'816
average read length	349	344
total number of bases	226 millions	223 millions
total number of contigs	49'311	46'298
contigs > 500 bp	19'408	17'164
average contig size [bp]	956	950
N50 contig size [bp]	1'016	1'003
largest contig size [bp]	8'335	7'430



L Baldo, E Santos & W Salzburger (2011) *Genome Biology and Evolution*

Regulatory Evolution

nucleotide divergence

overall*: 1.75%

annotated CDS & UTR: 1.33%



CDS: 0.95%



5' UTR coding sequence (CDS) 3' UTR polyA

*i.e. all sequences, open reading frames (ORFs) and untranslated regions (UTRs)

L Baldo, E Santos & W Salzburger (2011) *Genome Biology and Evolution*