Comparative Genomics

Ana C. Marques

ana.marques@dpag.ox.ac.uk

The Module - Overview

Lecture 1 - Comparative genomics- Ana Marques

Comparison of DNA sequences.

Lecture 2 - Comparative transcriptomics- Chris Ponting

Comparison of RNA/protein

Lecture 3 - Disease genomics- Caleb Webber

Using genomics to understand phenotype and disease

Practical - Ana Marques and Steve Meader

Using web-based data-mining tools to compare disease associated loci between human and mouse.

The Lecture - Overview

Overview:

1-Genome(s);

2-Genomics: comparative, functional and evolutionary;

3-Protein-coding genes and evolution;

The genome

The genome contains all the biological information required to build and maintain any given living organism.

The genome contains the organisms molecular history.

Decoding the biological information encoded in these molecules will have enormous impact in our understanding of biology.

Some history

1866- Gregor Mendel suggested that the traits were inherited.

1869-Friedrich Miescher isolated DNA.

1919-Phoebus Levene identified the nucleotides and proposed they were linked through phosphate groups.

1943- Avery, MacLeod and McCarty showed that DNA and not protein is the carrier of genetic information.

1953- Based on a X-ray diffraction taken by Rosalind Franklin and Raymond Gosling and the Erwin Chargaff discovery that DNA bases are paired James D. Watson and Francis Crick suggested the double helix structure for the DNA.

1957- Crick laid out the central dogma of molecular biology (DNA->RNA->protein).

1961 - Nirenberg and colleagues “cracked” the genetic code

Some history (cont.)

1975- Sanger sequencing

1976/79- First viral genome – MS2/fX174 (chromosomal walking- size ~5 kb)

1982 -First shotgun sequenced genome – Bacteriophage lambda (~50 kb)

1995 - First prokaryotic genome – *H. influenzae*

1996 - First unicellular eukaryotic genome – Yeast

1998 - The first multicellular eukaryotic genome – *C.elegans*

2000 - *Drosophila melanogaster* - fruitfly

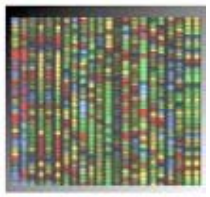
2000 - *Arabidopsis thaliana*

2001- Human Genome

~50 years

1865	Mendel discovers laws of genetics
1900	Rediscovery of Mendel's genetics
1944	DNA identified as hereditary material
1953	DNA structure
1960's	Genetic code
1977	Advent of DNA sequencing
1975-79	First human genes isolated
1986	DNA sequencing automated
1990	Human genome project officially begins
1995	First whole genome
1999	First human chromosome
2003	'Finished' human genome sequence






The Human genome project

The Human genome project promised to revolutionise medicine and explain every base of our DNA.

Large MEDICAL GENETICS focus

Identify variation in the genome that is disease causing

Determine how individual genes play a role in health and disease

The Human genome project

This was a huge technical undertaking so further aims of the project were…

This was a huge technical undertaking so further aims of the project were…

- Develop and improve technologies for: DNA sequencing, physical and genetic mapping, database design, informatics, public access
- Genome projects of 5 model organisms e.g. *E. coli, S. cerevisiae, C. elegans, D. melanogaster, M. musculus.*

Provide information about these organisms

As test cases for refinement and implementation of various tools required for the HGP

This was a huge technical undertaking so further aims of the project were…

- Develop and improve technologies for: DNA sequencing, physical and genetic mapping, database design, informatics, public access
- Genome projects of 5 model organisms e.g. *E. coli, S. cerevisiae, C. elegans, D. melanogaster, M. musculus.*

Provide information about these organisms

As test cases for refinement and implementation of various tools required for the HGP

- Train scientists for genomic research and analysis
- Examine and propose solutions regarding ethical, legal and social implications of genomic research (ELSI)

The 2 Human genome project

PUBLIC - Watson/Collins

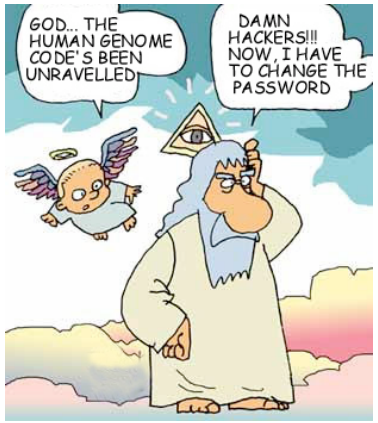
- Human Genome Project
- Officially launched in 1990
- Worldwide effort - both academic and government institutions
- Assemble the genome using maps
- 1996 Bermuda accord

PRIVATE - Craig Venter

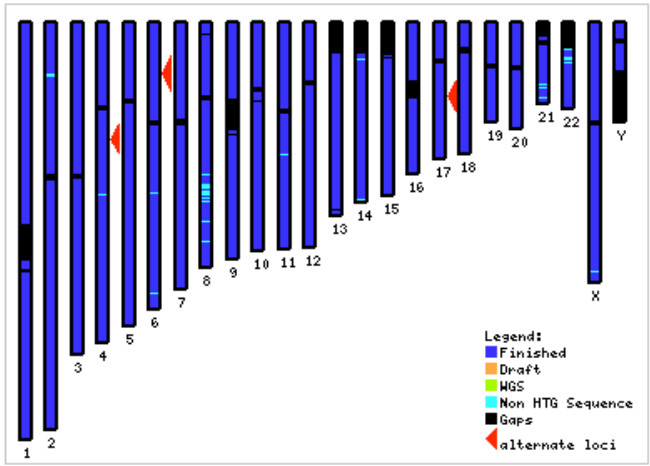
- 1998 Celera Genomics
- Aim to sequence the human genome in 3 years
- ‘Shotgun’ approach - no use of maps for assembly
- Data release NOT to follow Bermuda principles

The Human genome project

It cost 3 billion dollars and took 10 years to complete (5 less than initially predicted).

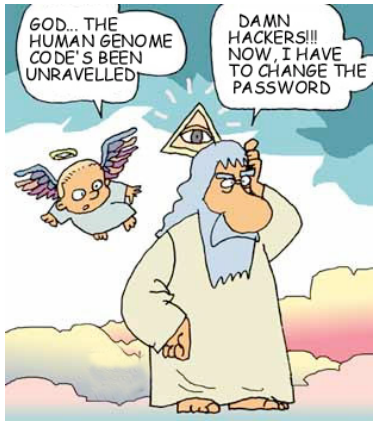


- Currently 3.2 Gb
- Approx 200 Mb still in progress
  - Heterochromatin
  - Repetitive
- Most recent human genome uploaded February 2009

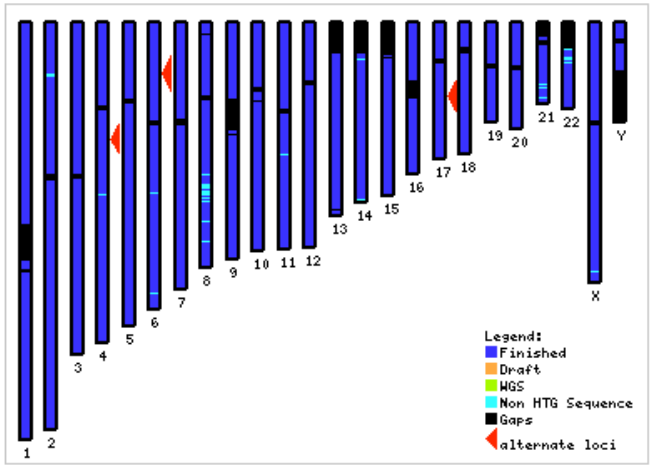


The Human genome project

It cost 3 billion dollars and took 10 years to complete (5 less than initially predicted).



- Currently 3.2 Gb
- Approx 200 Mb still in progress
  - Heterochromatin
  - Repetitive
- Most recent human genome uploaded February 2009



Finally, it is has not escaped our notice that the more we learn about the human genome, the more there is to explore.

"We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time."—T. S. Eliot[450] □

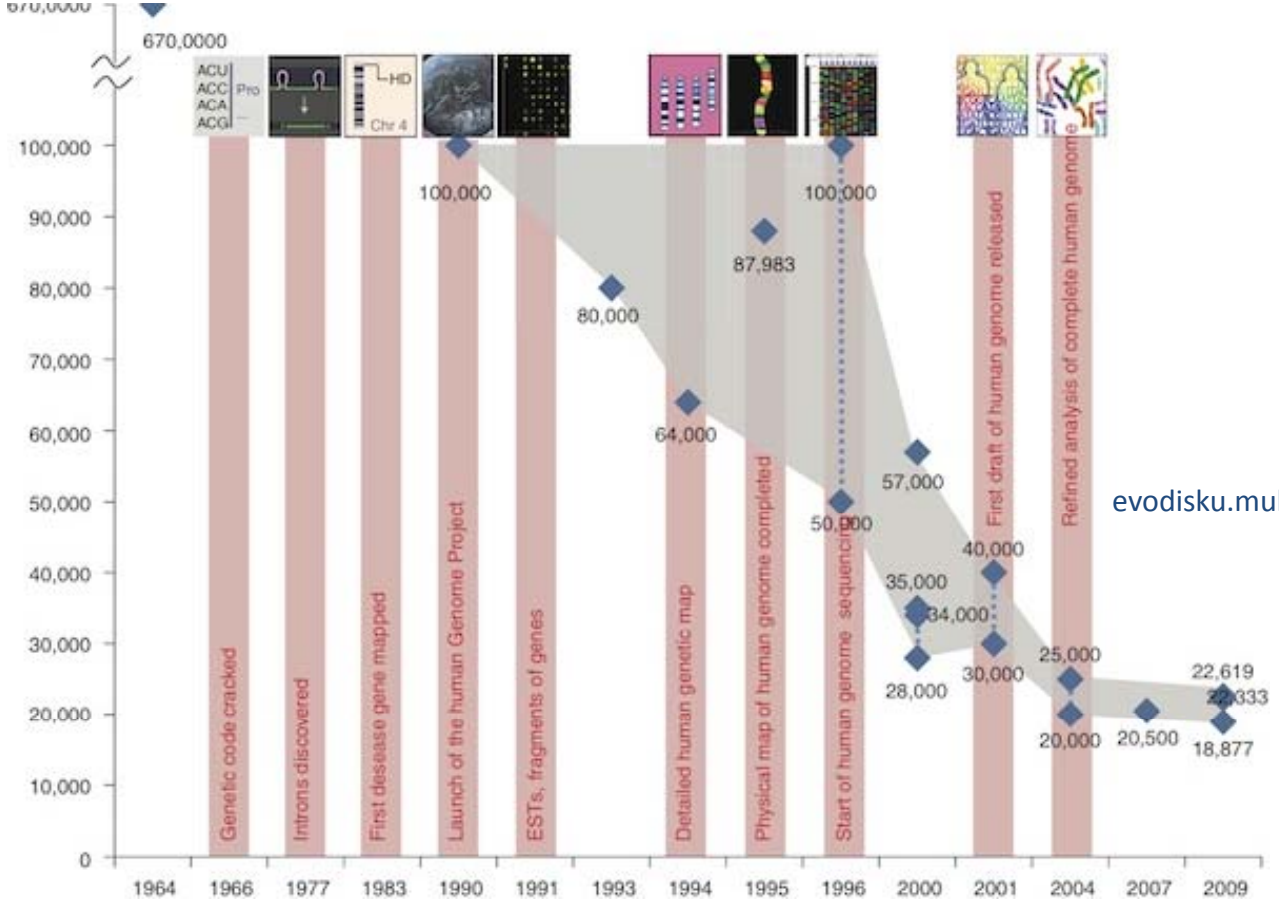
The Human genome.

From sequence to function

The scientific program outlined above focuses on how the genome sequence can be mined for biological information. In addition, the sequence will serve as a foundation for a broad range of functional genomic tools to help biologists to probe function in a more systematic manner. These will need to include improved techniques and databases for the global analysis of: RNA and protein expression, protein localization, protein–protein interactions and chemical inhibition of pathways. New computational techniques will be needed to use such information to model cellular circuitry. A full discussion of these important directions is beyond the scope of this paper.

The functional genome

- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly.



evodisku.multiply.com/notes/item/109

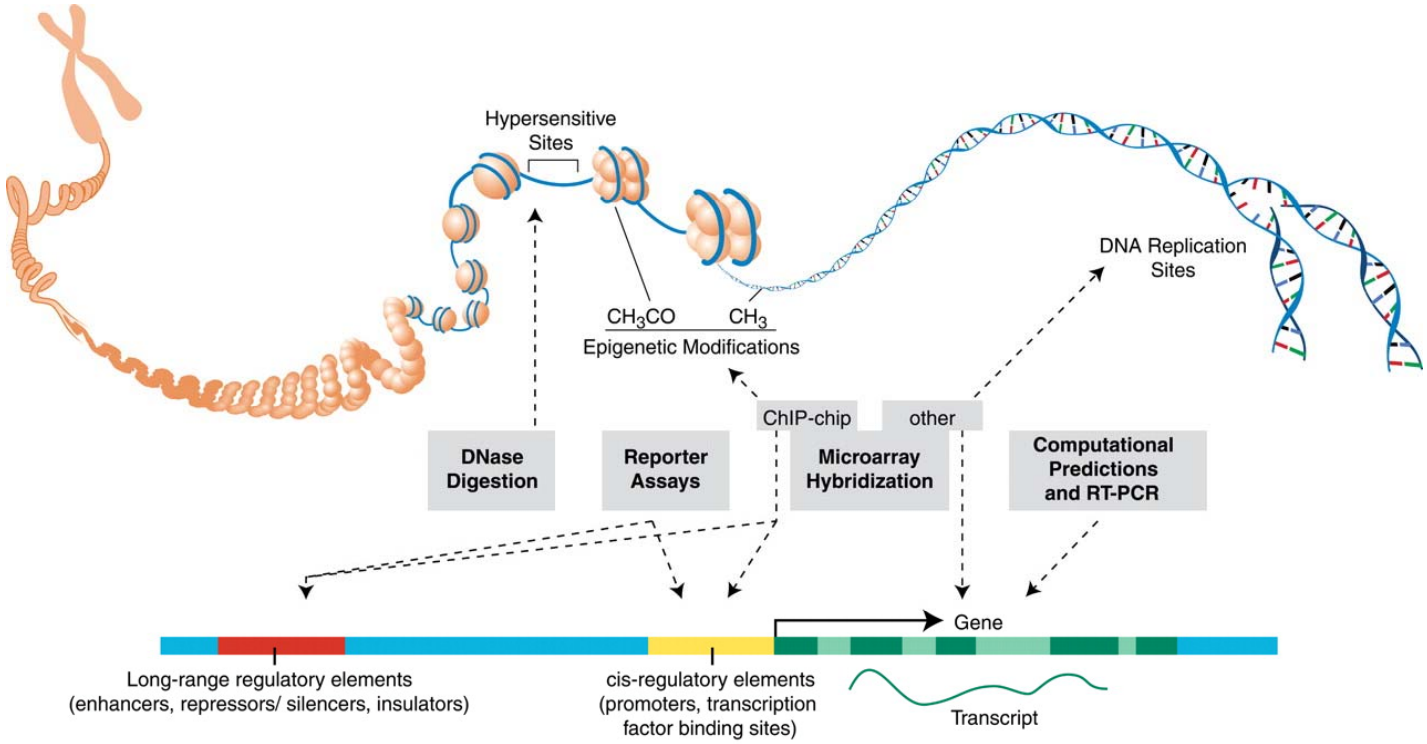
Protein-coding do not explain complexity/diversity.

The functional genome

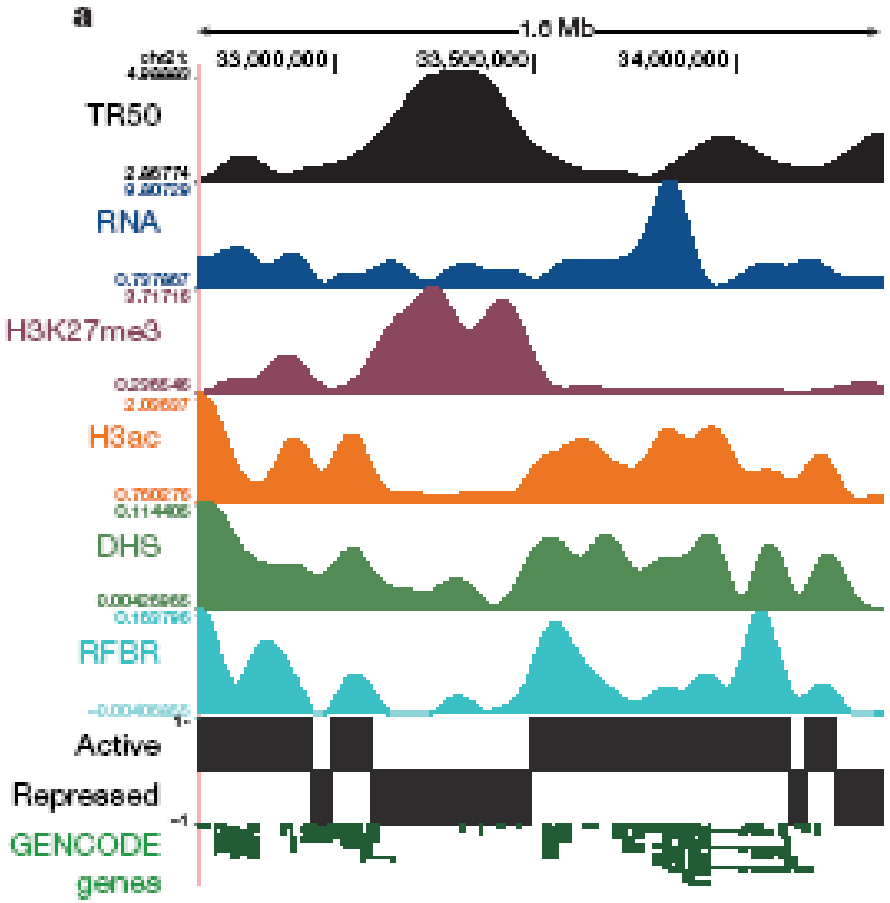
Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

35 Research groups threw everything at 30Mb (1%) of human DNA sequence. >200 experimental datasets (transcription, histone-modifications, chromatin structure, regulatory binding sites, replication timing, population variation and more.)



The functional genome map



Estimating the fraction of the genome that is functional

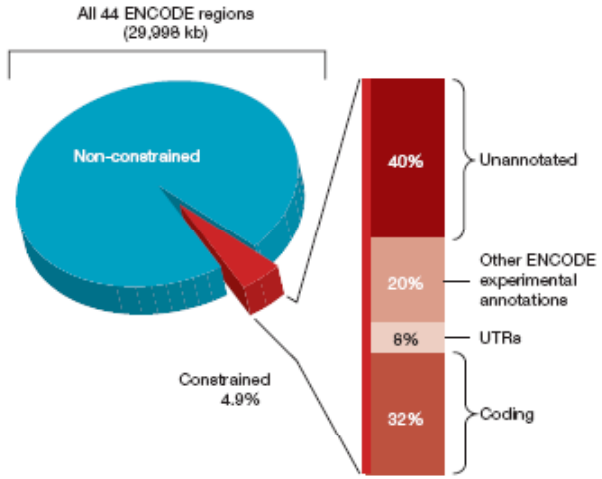


Figure 10 | Relative proportion of different annotations among constrained sequences. The 4.9% of bases in the ENCODE regions identified as constrained is subdivided into the portions that reflect known coding regions, UTRs, other experimentally annotated regions, and unannotated sequence.

- Only about 1.2% of the genome encodes protein sequence
- Most of it is composed of decaying transposons
- 5% appears “constrained” = likely functional
- >70% appears transcribed but unconstrained (lots fast evolving?)

2nd generation sequencing

Genome wide annotation of functional elements made easy!

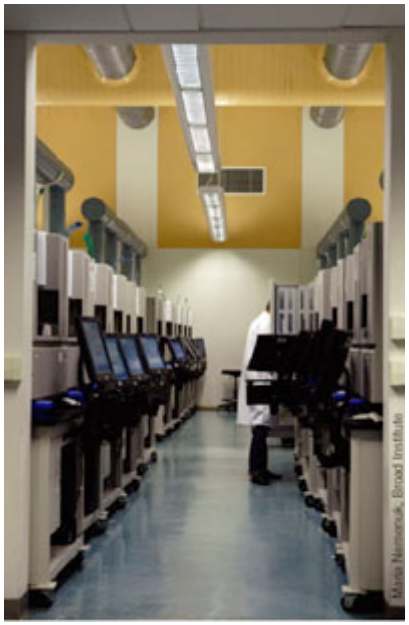
Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics

Applications

1-Genome sequencing and genome assembly (Panda genome, 2009)
2-Genome re-sequencing (Craig Venter, James Watson…1000 genomes project)
3- Transcriptome sequencing (unbiased)
4- Metagenomics
5-ChIP-seq
7-RIP-seq

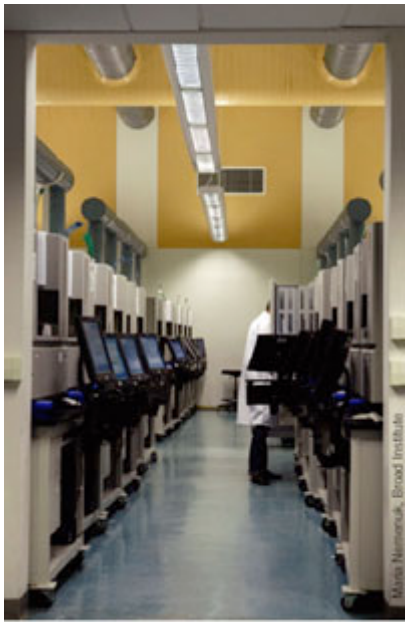
…seq.

Single molecule sequencing.
Potential to answer questions that remain open (somatic variation/ single cell transcription…)



3nd and counting generation sequencing

Single molecule sequencing.
Potential to answer questions that remain open (somatic variation/ single cell transcription…)



Next generation sequencing has (and will continue to) changed the way we do and understand biology!
More data but what should we do with it?

How we use this data to understand physiology, behaviour, disease and variation between species/individuals we need to:

- The evolutionary history of every genetic element (every base)
- Evolutionary forces shaping the genome
- Structural and sequence variation in the population and between species.

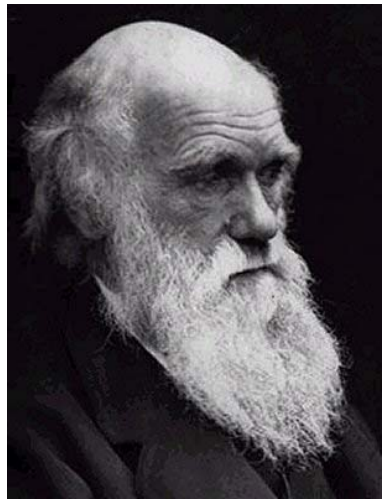
From genome to biology

How we use this data to understand physiology, behaviour, disease and variation between species/individuals we need to:

- The evolutionary history of every genetic element (every base)
- Evolutionary forces shaping the genome
- Structural and sequence variation in the population and between species.

Comparative genomics studies differences between genome sequences pin-pointing changes over time. Comparison of the number/type changes against the background “neutral” expected changes provides a better understanding of the forces that shaped genomes and traits.

"Nothing in Biology Makes Sense Except in the Light of Evolution."
Theodosius Dobzhansky

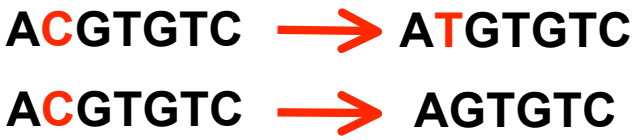


MUTATION

1. Small scale mutations

Nucleotide substitutions

Small Insertions / Deletions (Indels)



MUTATION

1. Small scale mutations

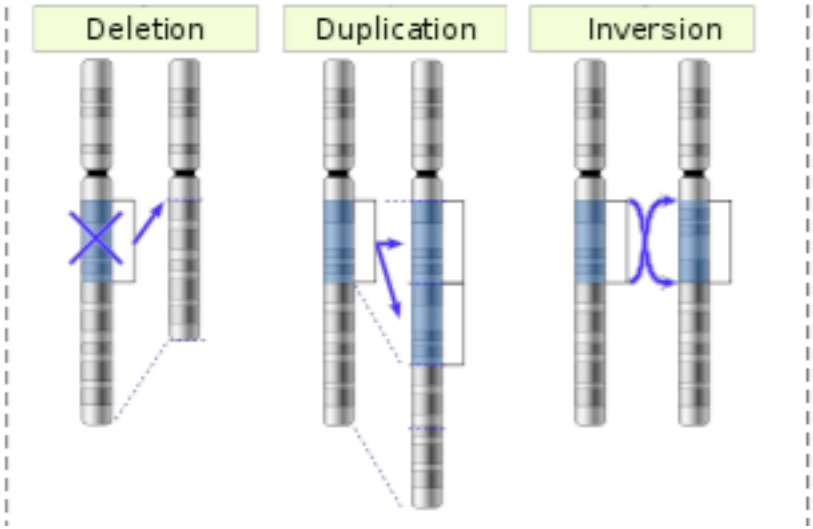
Nucleotide substitutions

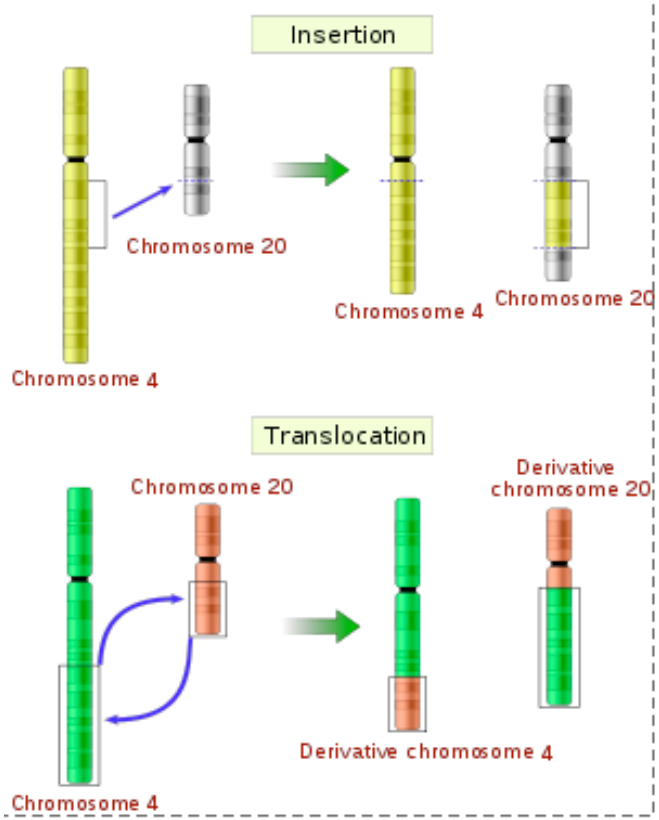
Small Insertions / Deletions (Indels)

ACGTGTC → ATGTGTC

ACGTGTC → AGTGTC

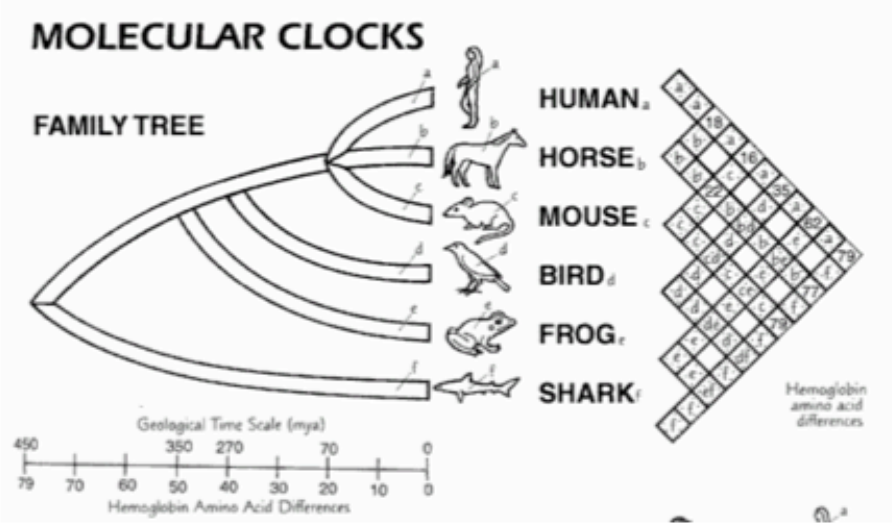
2. Large scale mutations (> 1kb)





How do changes accumulate in the genome?

In 1965 Pauling and colleagues showed that for any given protein the rate of molecular evolution is approximately constant in all lineages.



Motoo Kimura

1968, proposed that most mutations accumulated in genomes are neutral.

The Neutral Theory.

Neutral model

Aim: Identify regions of the genome that are not evolving neutrally!

LOCI X-Neutral

Species 1 CGACATTAAATAGGCGCAGGACCAGATACCAGATCAAAGCAGGCGCA

Species 2 CGACGTTAAATTGGCGCAGTATCAGATACCCGATCAAAGCAGACGCA

Neutral model

Aim: Identify regions of the genome that are not evolving neutrally!

LOCI X-Neutral

Species 1 CGACATTAAATAGGCGCAGGACCAGATACCAGATCAAAGCAGGCGCA

Species 2 CGACGTTAAATTGGCGCAGTATCAGATACCCGATCAAAGCAGACGCA

LOCI Y

Species 1 CATGGGTCATCACTCTAGCTGTACGTCTACTTCATCATCGCGCTACG

Species 2 CATGAGTCATCACTCTAGCTGTACGTCTACTTCATCATCGCGTTACG

Neutral model

Aim: Identify regions of the genome that are not evolving neutrally!

LOCI X-Neutral

Species 1 CGACATTAAATAGGCGCAGGACCAGATACCAGATCAAAGCAGGCGCA

Species 2 CGACGTTAAATTGGCGCAGTATCAGATACCCGATCAAAGCAGACGCA

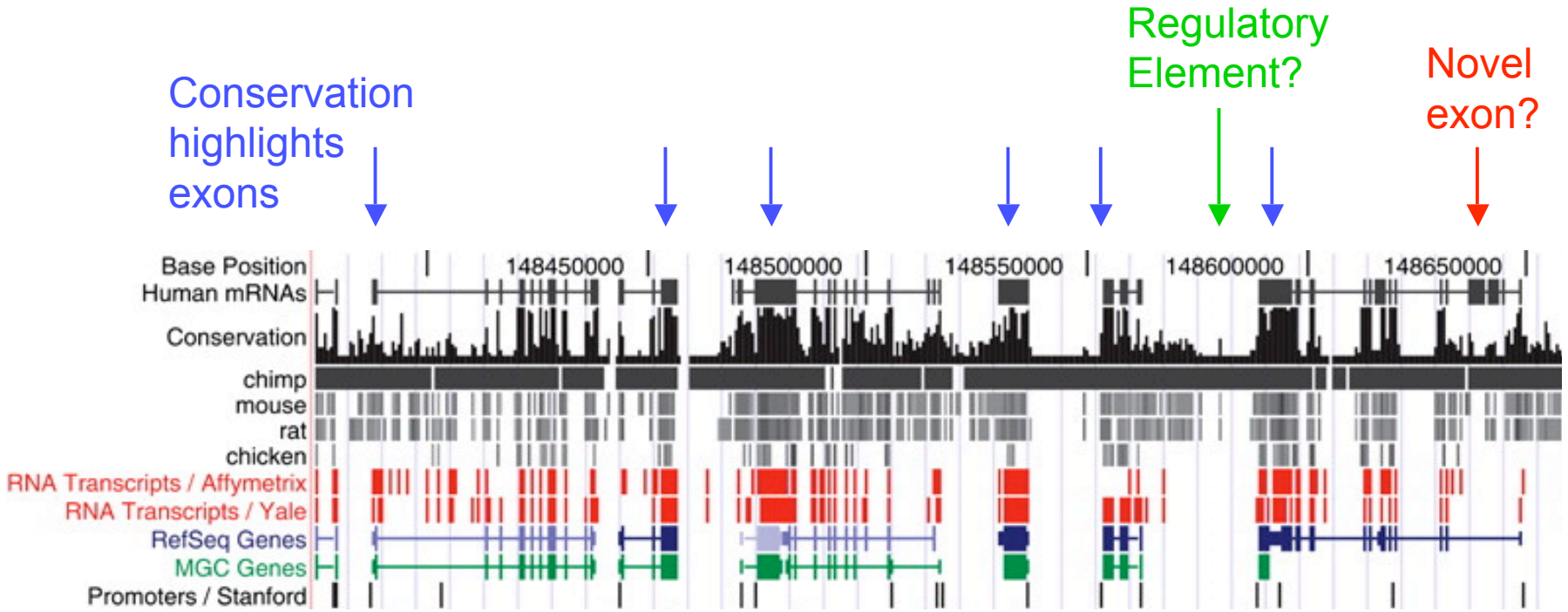
LOCI Y

Species 1 CATGGGTCATCACTCTAGCTGTACGTCTACTTCATCATCGCGCTACG

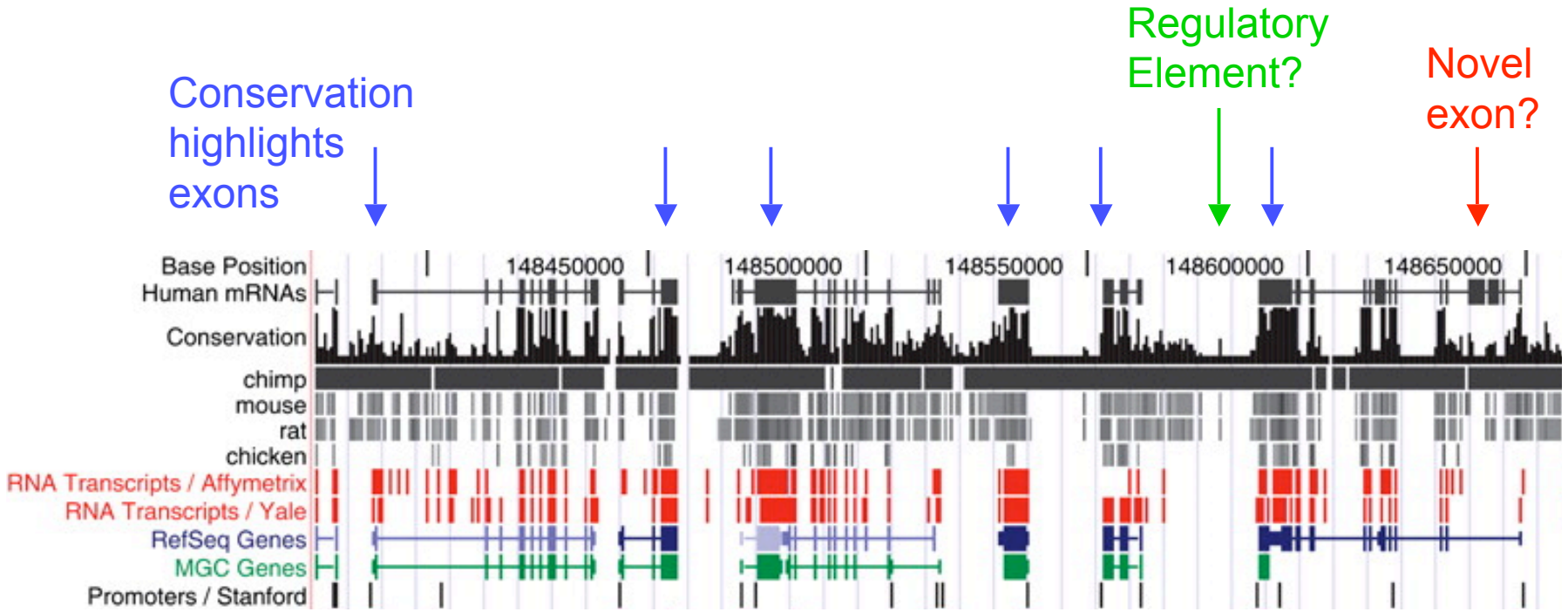
Species 2 CATGAGTCATCACTCTAGCTGTACGTCTACTTCATCATCGCGTTACG

Sequence that is conserved over long evolutionary distances is likely to be under selective constraint

Conservation is often a good predictor of functionality



Conservation is often a good predictor of functionality

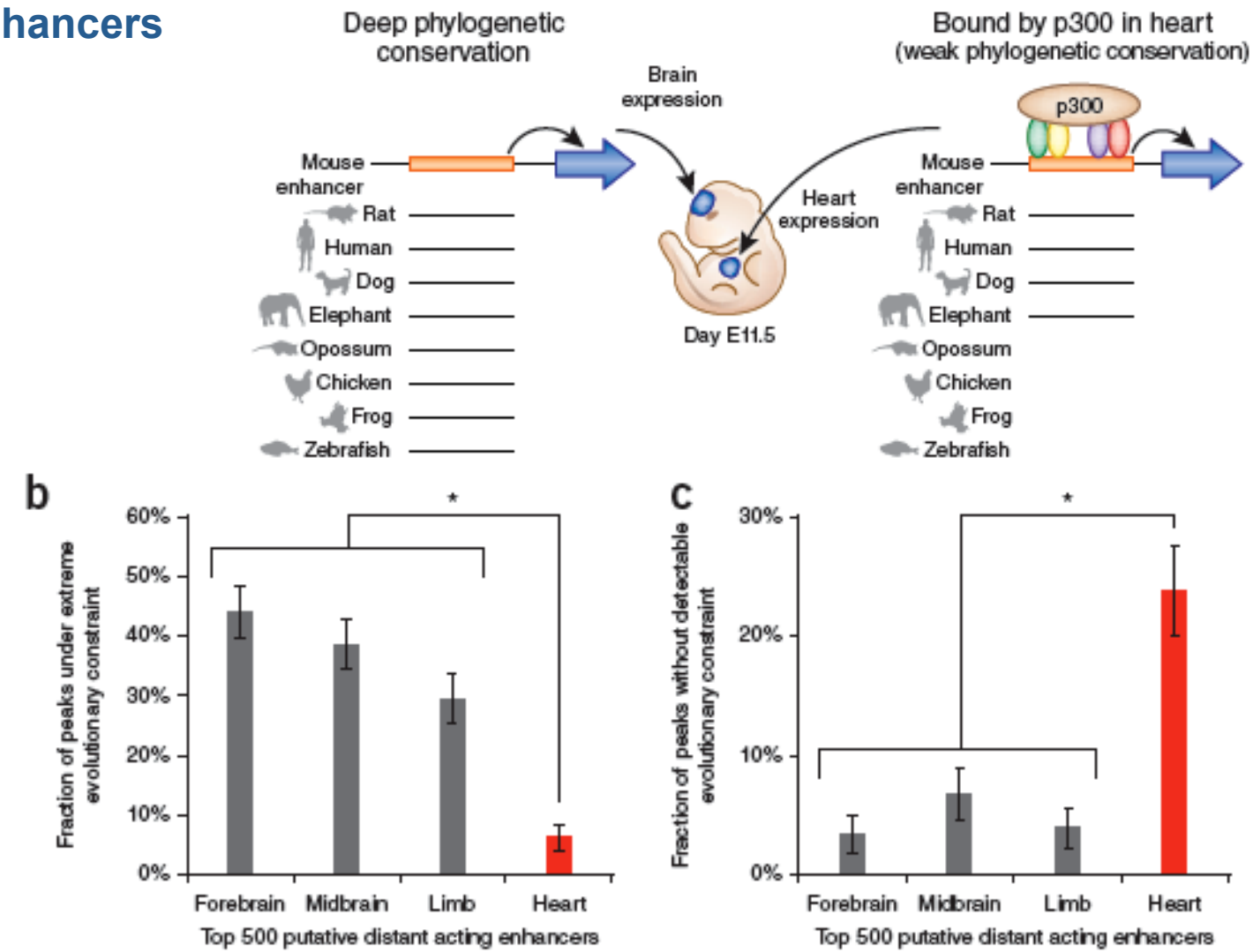


BUT…

Conservation is not synonymous of function

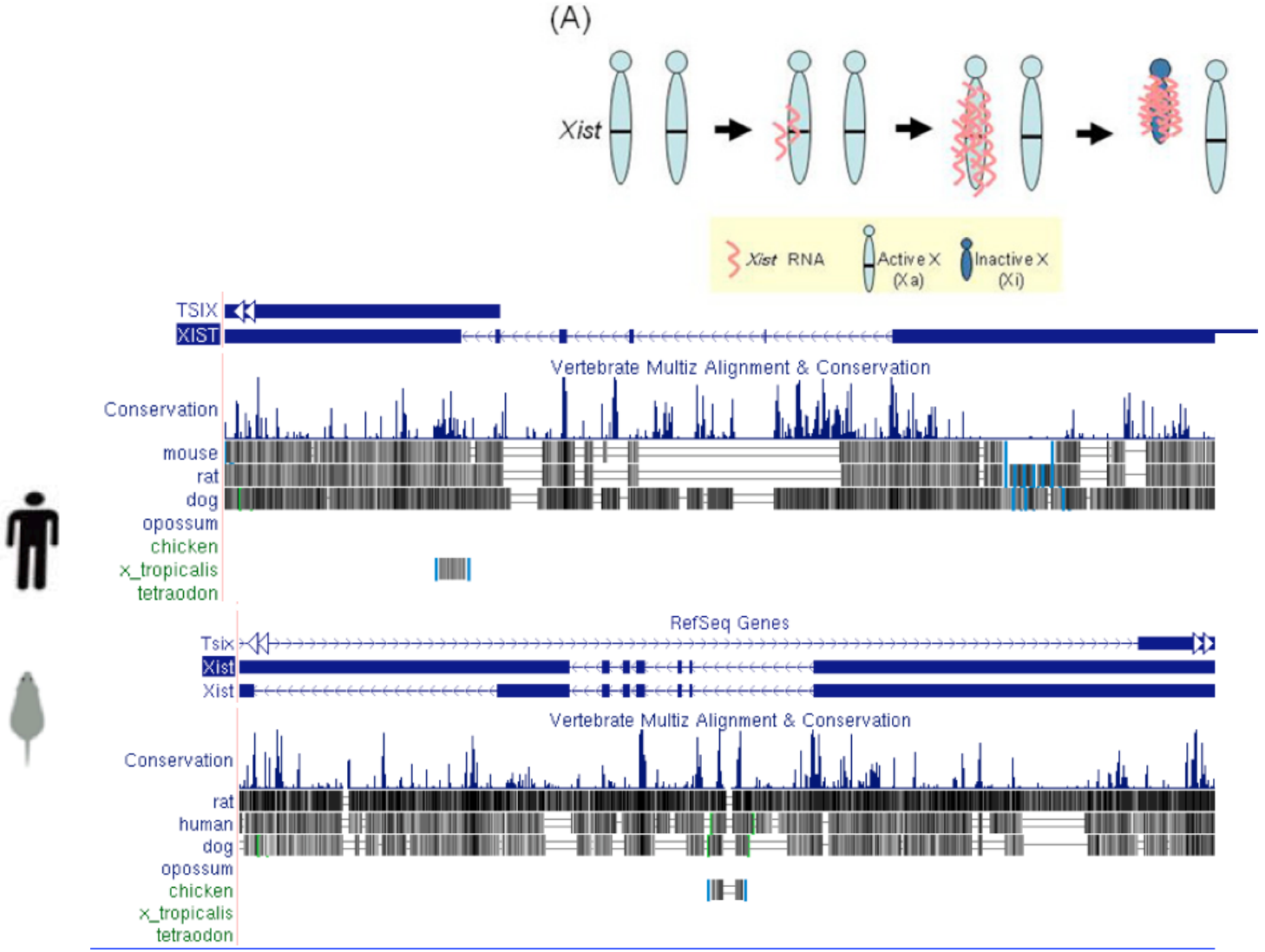
Not all functional sequence is conserved across long evolutionary distance.

Heart Enhancers



Conservation is not synonymous of function

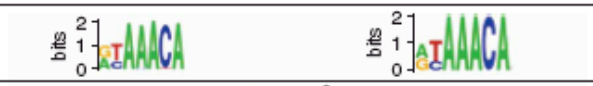
Long Intergenic ncRNA



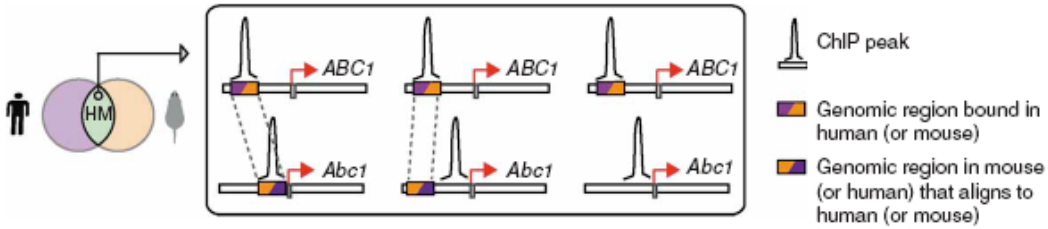
Sequence conservation doesn't imply function conservation

Despite conservation of binding preferences and binding sites only a small proportion of TF binding events is conserved across species

a

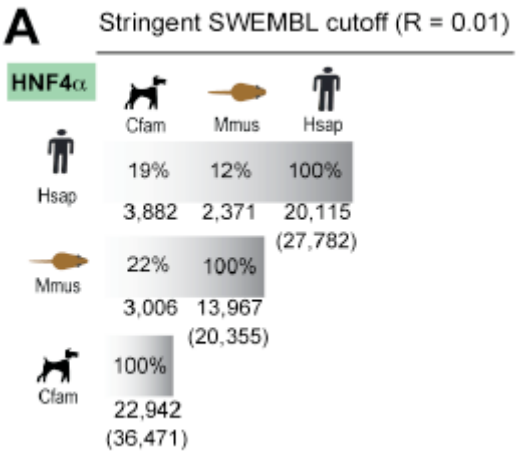
Regulator	PFAM category	HS bound	MM bound	Intersection	*P* value	HS binding sequence	MM binding sequence
FOXA2	Forkhead	151	574	68	1.0E–45		
HNF1A	POU-homeodomain	251	224	45	1.0E–29		
HNF4A	Nuclear receptor	1,251	654	387	1.0E–136		
HNF6	CUT-homeodomain	157	324	41	1.0E–27		

c



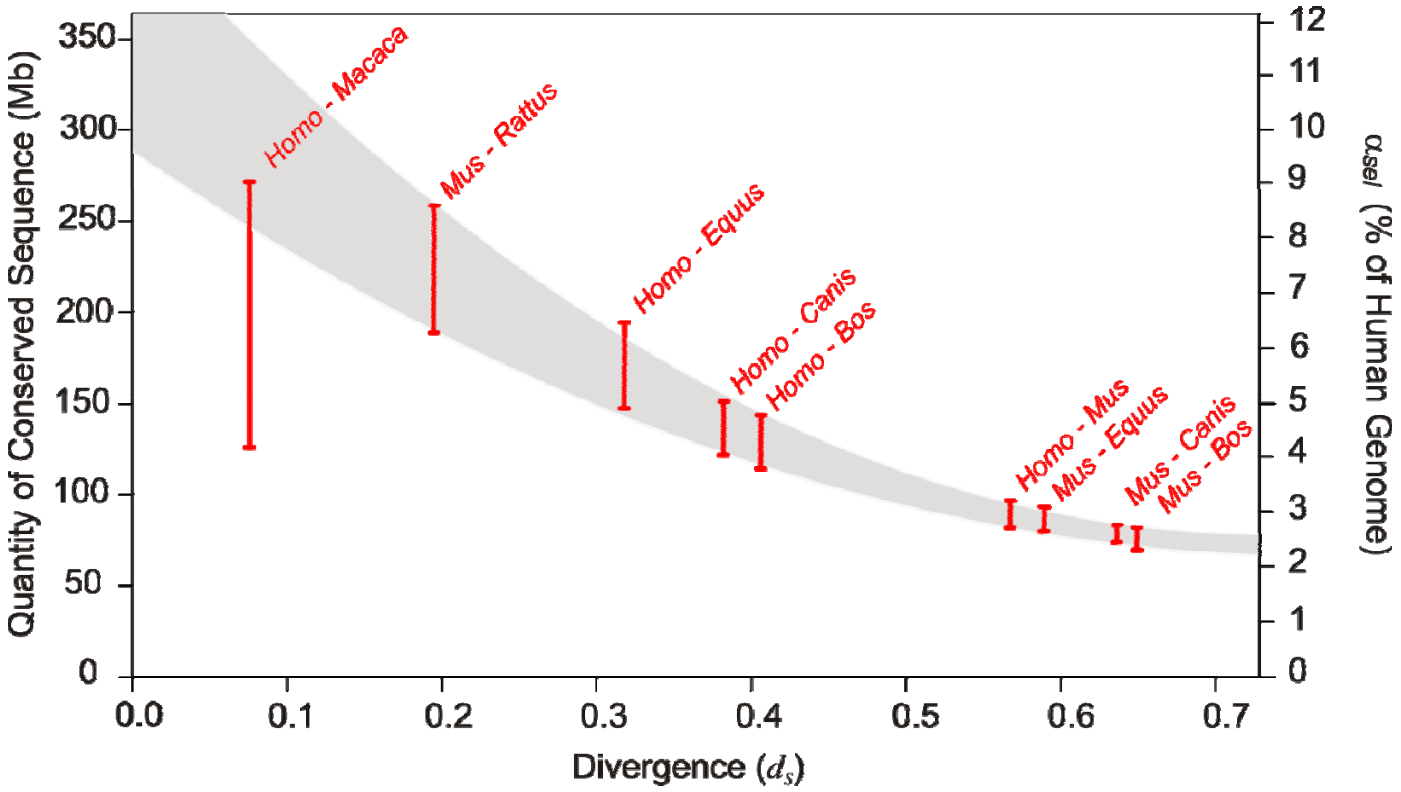
FOXA2	34%	39%	28%
HNF1A	41%	31%	28%
HNF4A	23%	38%	38%
HNF6	30%	41%	30%

A



Odom D. et al (2007)
Schmidt D. et al (2010)

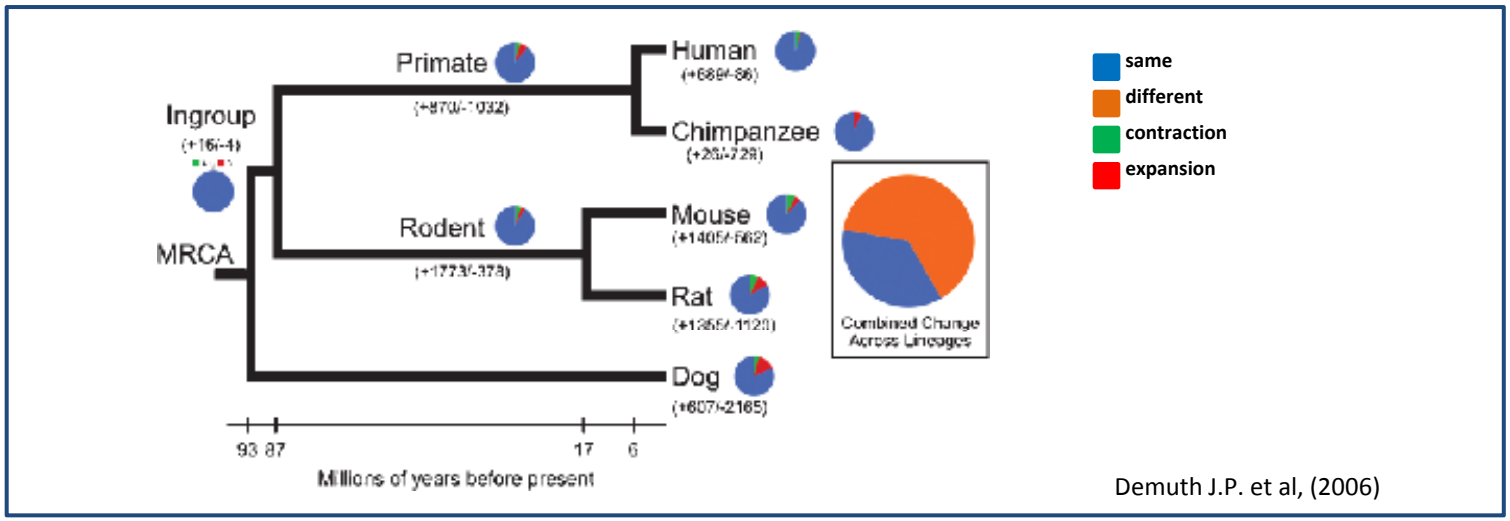
Sequence conservation doesn't imply function conservation



Massive turnover of functional sequence in mammalian genomes

Meader S et al. (2011)

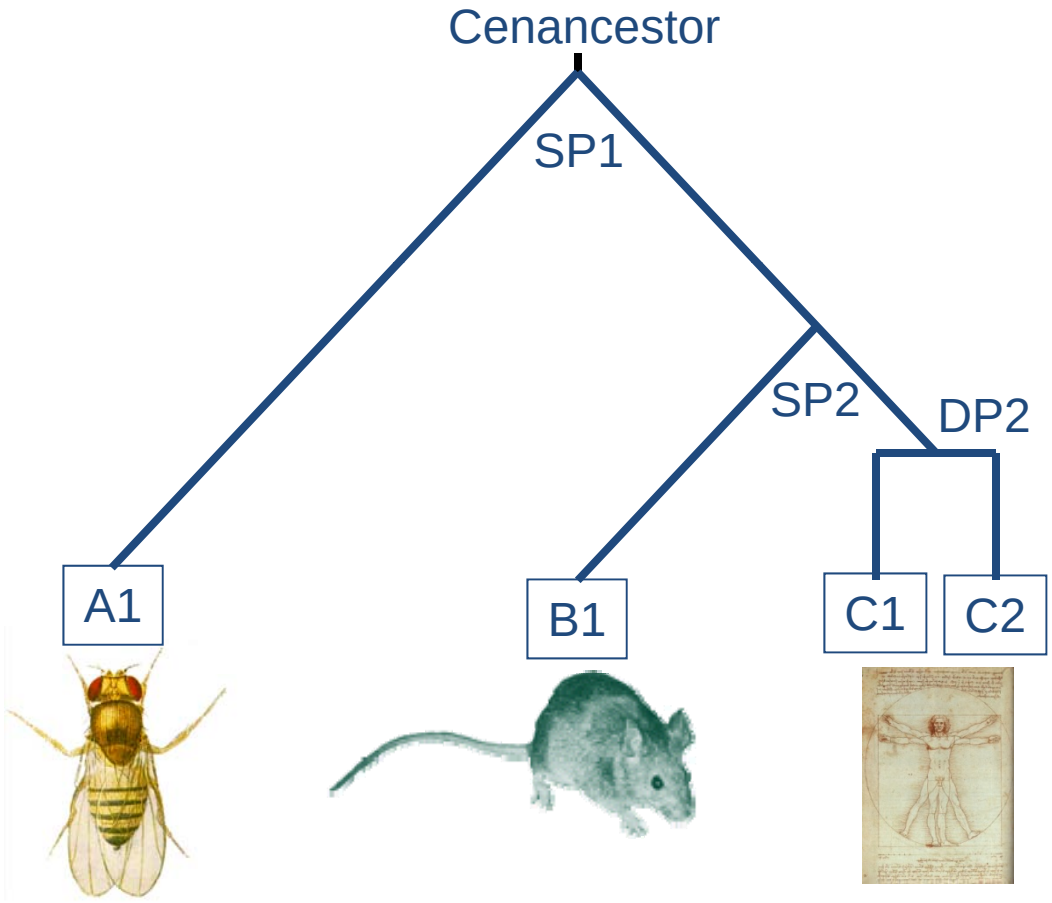
Lessons from comparative genomics: Changes of protein coding repertoires and contributions to phenotypic differences



Demuth J.P. et al, (2006)

Identifying changes in the genome requires resolving evolutionary relationships for all bases.

Homologues: Common descent from an ancestral sequence
Paralogues: Homologues in the same genome which are the result of gene duplication; Often short hand for:
In-paralogues: Genes which have arisen from duplications in one lineage (E.g. mouse- or human- specific gene duplications)
Orthologues: Corresponding genes in two species which were derived from a single gene in the last common ancestor



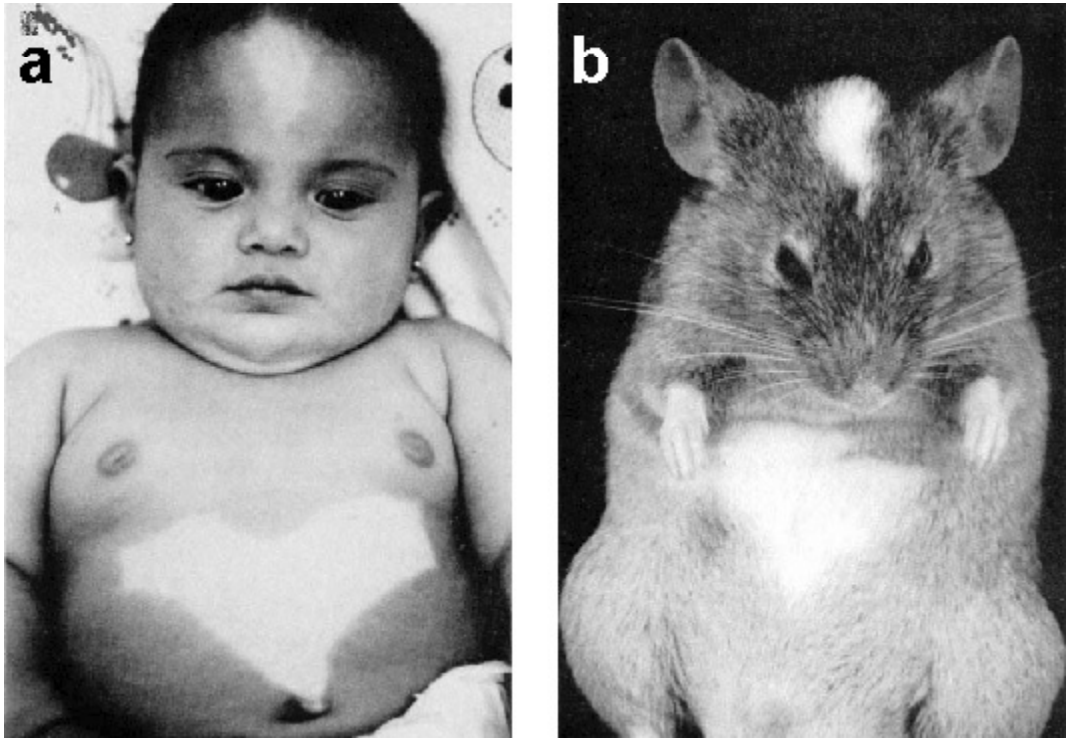
C1 and C2 are paralogues
A1 and B1 and (C1 and C2) are orthologues

Orthologs

We share ~80% (~16,000) of our genes with the mouse!

They are more than 90 million years old

1:1 orthologues are most likely to retain the common ancestral function



Human and mouse c-kit mutations show similar phenotypes.
The utility of mouse as a biomedical model for human disease is enhanced when mutations in orthologous genes give similar phenotypes in both organisms.
In a visually striking example of this, the same pattern of hypopigmentation is seen in (a) a patient with the piebald trait and (b) a mouse with dominant spotting, both resulting from heterozygous mutations of the c-kit proto-oncogene.

Measuring evolutionary rates on protein coding genes

There are 2 type of mutations:

synonymous - don't change the encode aa.

non-synonymous-change aa.

Steps:

1. Synonymous and nonsynonymous changes treated separately

2. Compute the number of potential synonymous and nonsynonymous sites in the two sequences and get the average

Example:

Ser	Thr	Glu	Met	Cys	Leu
TCA	ACT	GAG	ATG	TGT	TTA

Leu	Thr	Glu	Ile	Cys	Leu
TTA	ACA	GAG	ATA	TGT	CTA

Genetic code

TABLE 1.2 The universal genetic code

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Example:

Ser	Thr	Glu	Met	Cys	Leu
TCA	ACT	GAG	ATG	TGT	TTA
N					

Leu	Thr	Glu	Ile	Cys	Leu
TTA	ACA	GAG	ATA	TGT	CTA

Genetic code

TABLE 1.2 The universal genetic code

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Example:

Ser	Thr	Glu	Met	Cys	Leu
TCA	ACT	GAG	ATG	TGT	TTA
NN					

Leu	Thr	Glu	Ile	Cys	Leu
TTA	ACA	GAG	ATA	TGT	CTA

Genetic code

TABLE 1.2 The universal genetic code

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Example:

Ser	Thr	Glu	Met	Cys	Leu
TCA	ACT	GAG	ATG	TGT	TTA
NNS					

Leu	Thr	Glu	Ile	Cys	Leu
TTA	ACA	GAG	ATA	TGT	CTA

Measuring evolutionary rates on protein coding genes

Example:

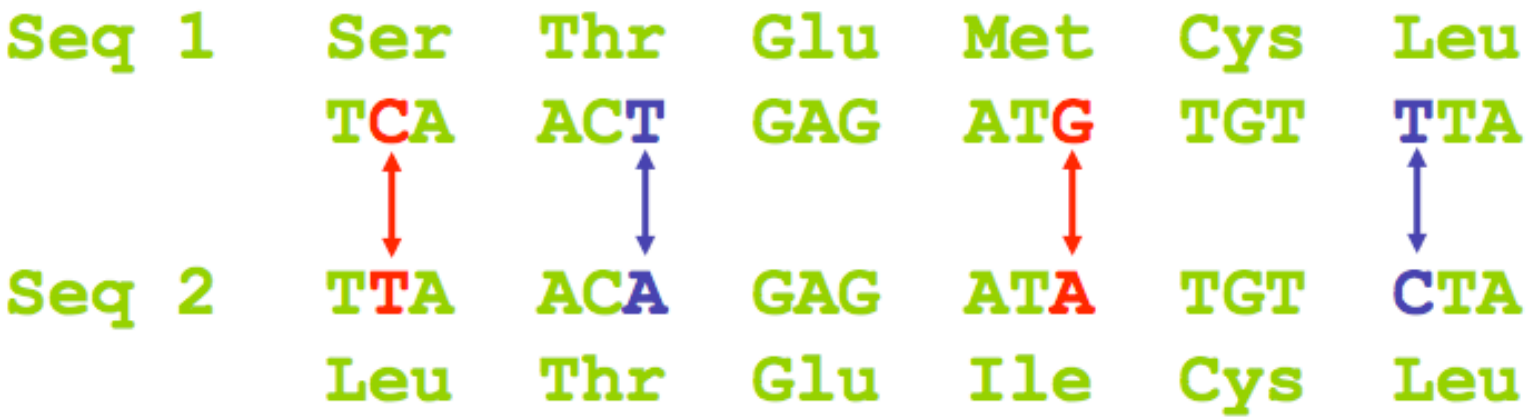
Ser	Thr	Glu	Met	Cys	Leu
TCA	ACT	GAG	ATG	TGT	TTA
NNS	NNS	$NN_{1/3S}$	NNN	$NN_{1/2S1/3S}N_{1/3S}$	
		$_{2/3N}$		$_{1/2N2/3N}$ $_{2/3N}$	

Leu	Thr	Glu	Ile	Cys	Leu
TTA	ACA	GAG	ATA	TGT	CTA
NNS	NNS	$NN_{1/3S}$	$NN_{2/3S}$	$NN_{1/2S1/3S}NS$	
		$_{2/3N}$	$_{1/3N}$	$_{1/2N2/3N}$	

$$S = \frac{2+1/3+1/2+1/3+1/3+2+1/3+2/3+1/2+1+1/3}{2} = 4.1667$$

$$N = \frac{6+2/3+5+1/2+2/3+1+2/3+6+2/3+2+1/3+2+1/2+1+2/3}{2} = 13.8333$$

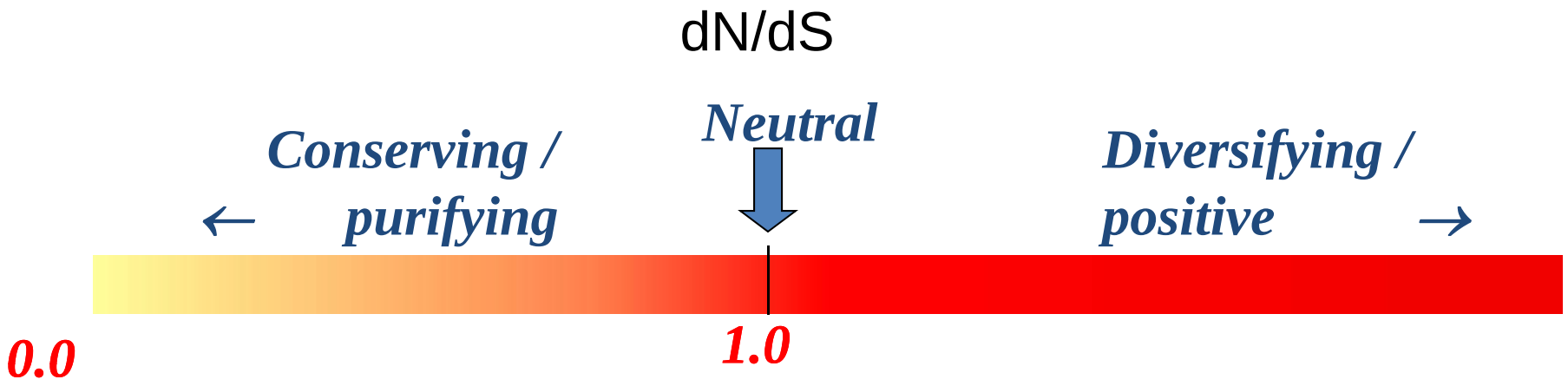
Estimation of substitution rates in protein coding regions



d_N = 2/13.83 = 0.14

d_S = 2/4.1667 = 0.48

Measuring evolutionary rates on protein-coding genes

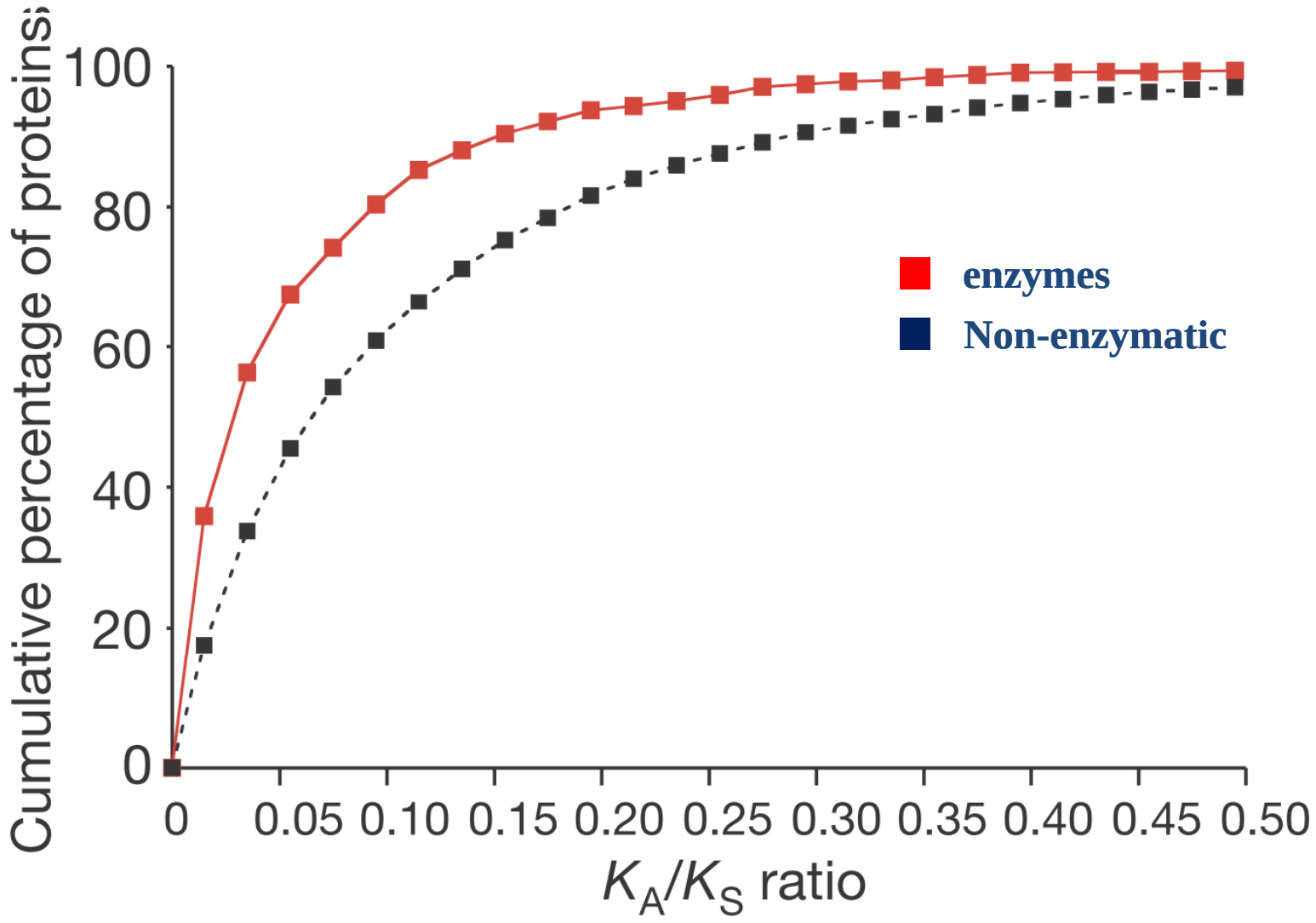


<<	1	purifying selection
=	1	neutral
>>	1	positive diversifying selection

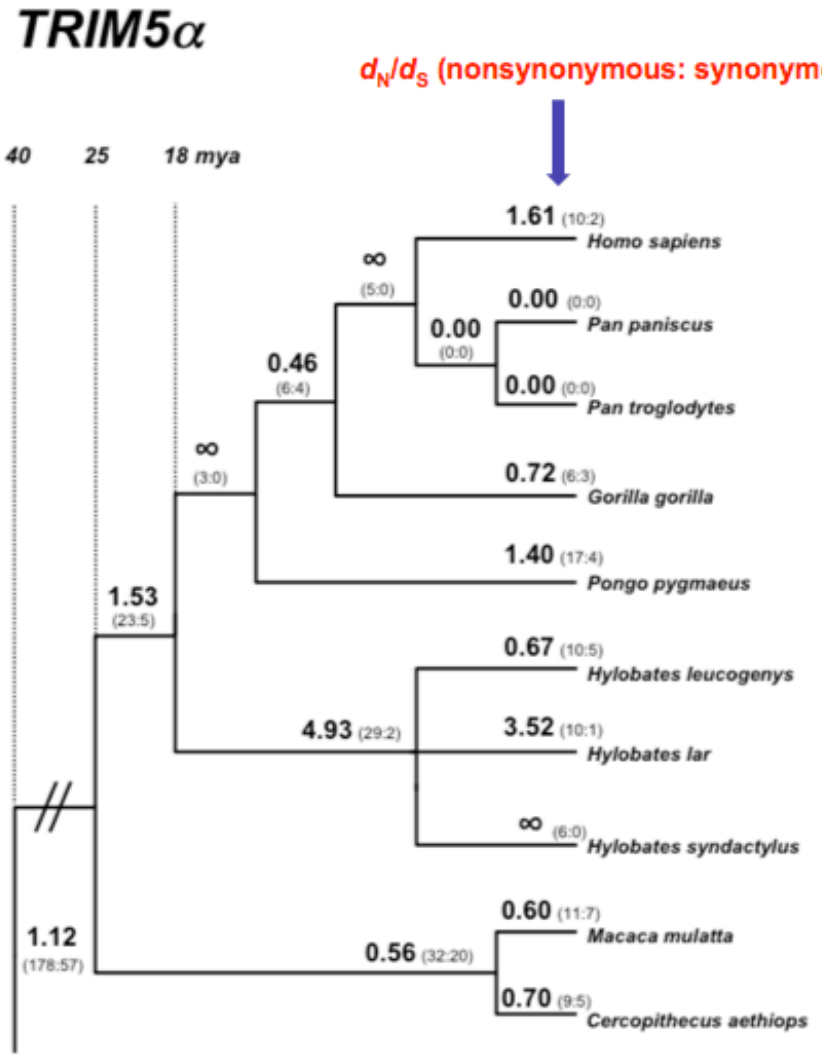
N.B: d_N non-synonymous substitution rate

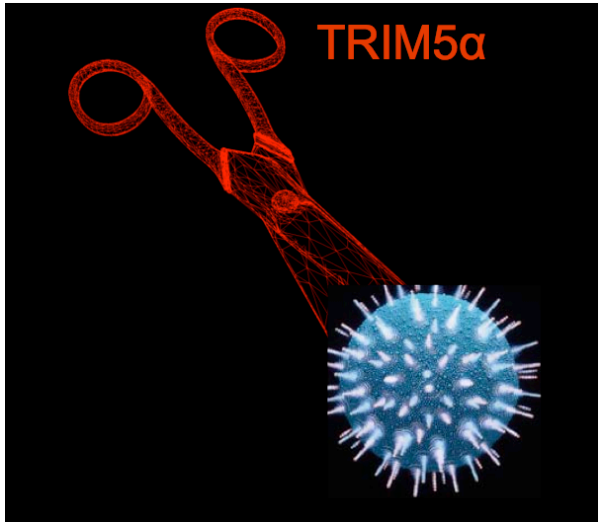
d_s synonymous rate

Slow evolvers



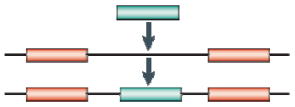
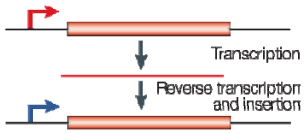
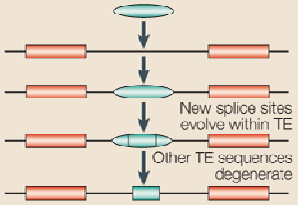
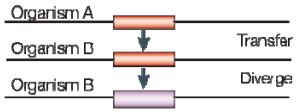
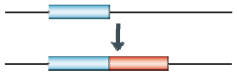
Fast evolvers





Origin of new elements in the genome

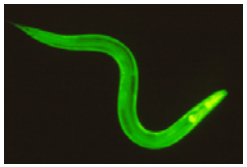
Table 1 | Molecular mechanisms for creating new gene structures

Mechanism	Process	Examples	Comments	References
Exon shuffling: ectopic recombination of exons and domains from distinct genes		*fucosyltransferase*, *jingwei*, *Tre2*	~19% of exons in eukaryotic genes have been formed by exon shuffling	8,32,40,62, 65–68,105
Gene duplication: classic model of duplication with divergence		*CGβ*, *Cid* , *RNASE1B*	Many duplicates have probably evolved new functions	9–11,29,35,39, 47,48,106
Retroposition: new gene duplicates are created in new genomic positions by reverse transcription or other processes		*PGAM3*, *Pgk2*, *PMCHL1*, *PMCHL2*, *Sphinx*	1% of human DNA is retroposed to new genomic locations	23,43,61,76, 80–82,107–110
Mobile element: a mobile element, also known as a transposable element (TE), sequence is directly recruited by host genes		*HLA-DR-1*, human *DAF*, *lungerkine* mRNA, *mNSC1* mRNA	Generates 4% of new exons in human protein-coding genes	16,78,111,112
Lateral gene transfer: a gene is laterally (horizontally) transmitted among organisms		*acytylneuraminate lysase*, *Escherichia coli* *mutU* and *mutS*	Most often reported in prokaryotes and recently reported in plants	18–20,113
Gene fusion/fission: two adjacent genes fuse into a single gene, or a single gene splits into two genes		Fatty-acid synthesis enzymes, *Kua-UEV*, *Sdic*	Involved in the formation of ~0.5% of prokaryotic genes	21,22,42, 114,115
De novo origination: a coding region originates from a previously non-coding genomic region		*AFGPs*, *BC1RNA*, *BC200RNA*	Rare for whole gene origination; might not be rare for partial gene origination	52–53,116,117

AFGP, antifreeze glycoprotein; *CGβ*, chorionic gonadotropin β polypeptide; *Cid*, centromere identifier; *DAF*, decay-accelerating factor; *HLA-DR-1*, major histocompatibility complex DR1; *PGAM3*, phosphoglycerate mutase 3; *Pgk2*, phosphoglycerate kinase 2; *PMCHL*, pro-melanin-concentrating hormone-like; *RNASE*, ribonuclease; *Sdic*, sperm-specific dynein intermediate chain; *UEV*, tumour susceptibility gene.

Proportion of (paralogous) genes in gene families

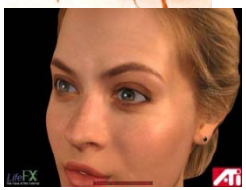
***Saccharomyces* (yeast): 30%**



***C. elegans*: 48%**

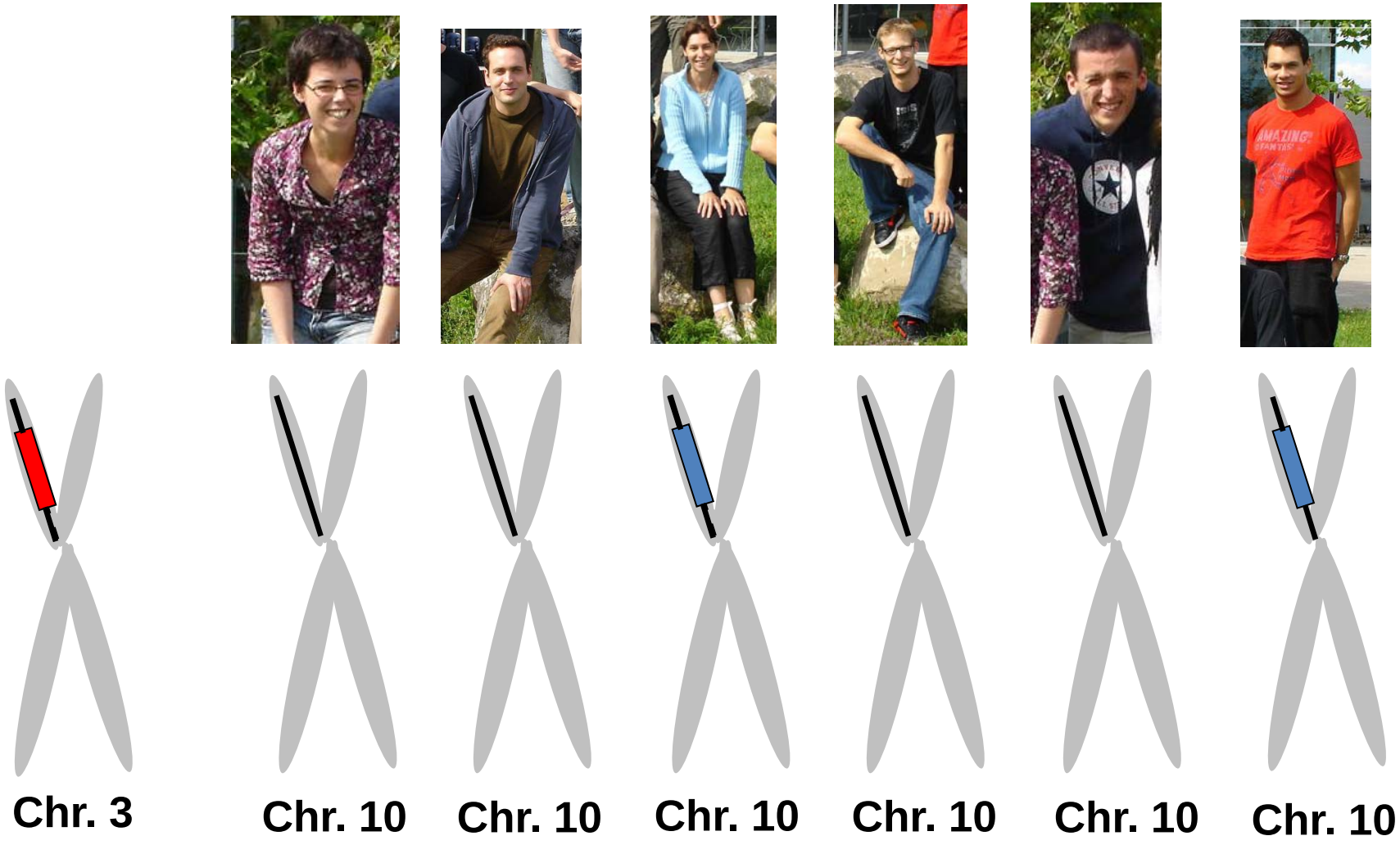
***Arabidopsis*: 60%**

***Drosophila:* 40%**

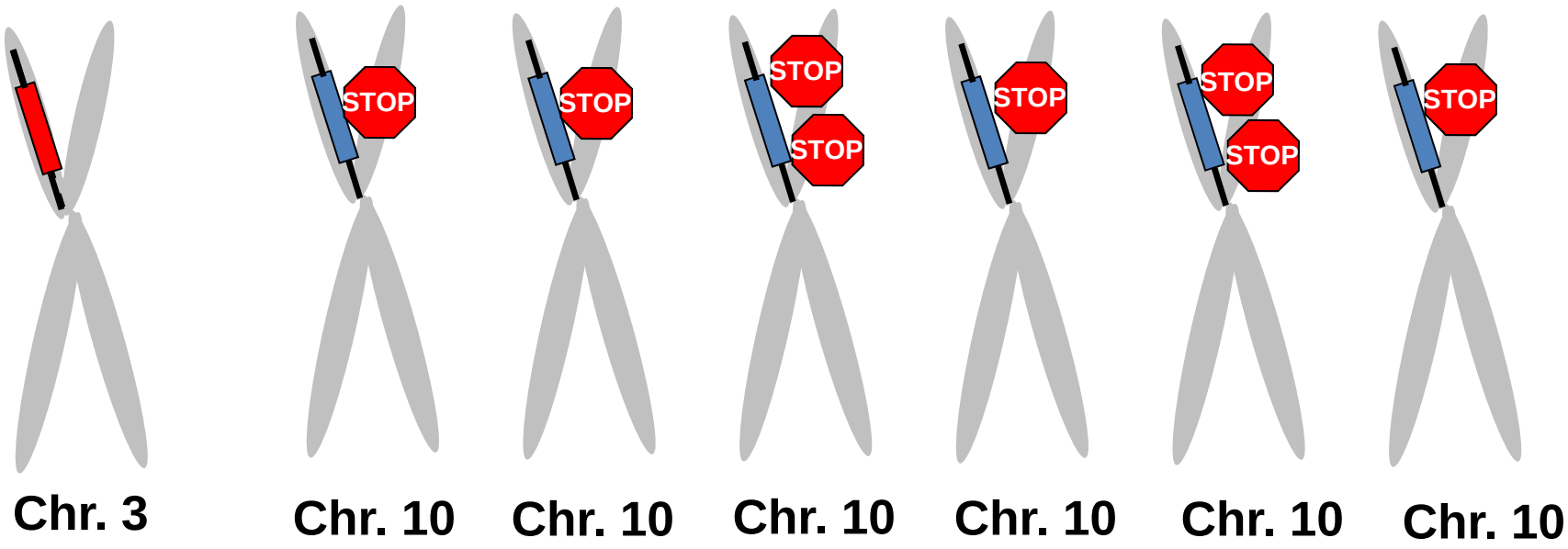


Humans: 40%

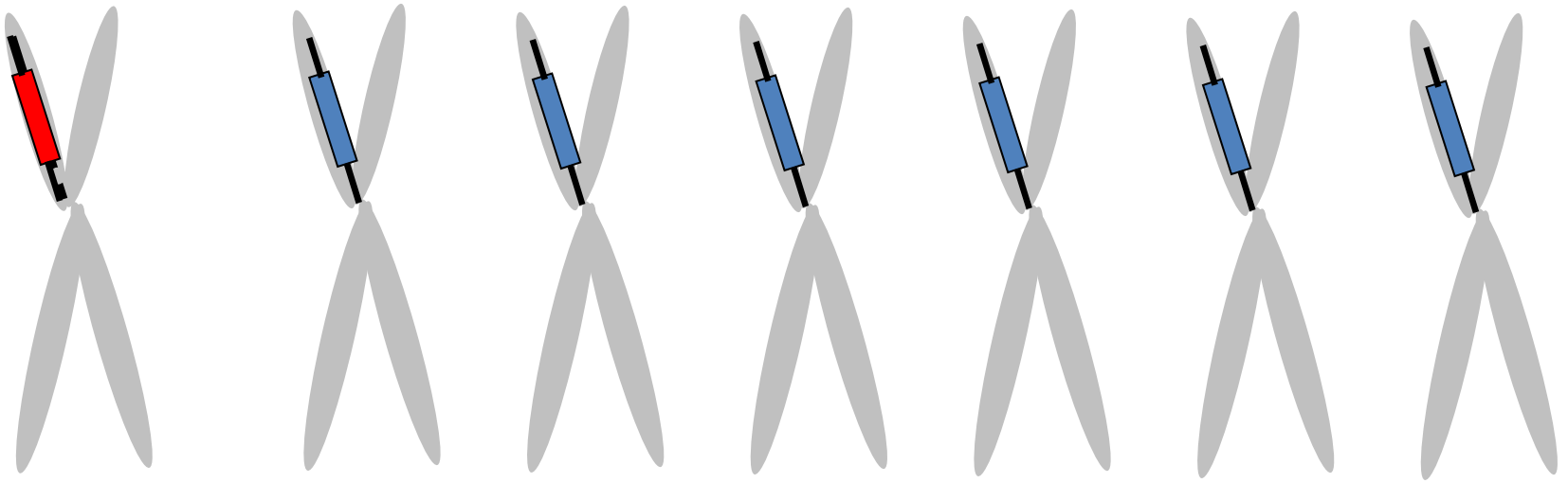
1. Duplication occurs but **does not reach fixation** in the population



2. Duplication occurs and fixes in the population but **degenerates** becoming a **pseudogene**: deletions, insertions and stop codons



3. Duplication occurs and fixes in the population – **new gene** is kept in the genome **with function** !



Evolutionary fate/role of new functional gene

- **Duplication for the sake of producing more of the same.**

Duplication of protein coding genes

TABLE 10.2 *Numbers of rRNA and tRNA genes per haploid genome in various organisms*

Genome source	Number of rRNA genes[a]	Number of tRNA genes	Approximate genome size (bp)
Human mitochondrion	1	22	1.7×10^4
Mycoplasma genitalium	2	33	5.8×10^5
Escherichia coli	7	~100	4×10^6
Saccharomyces cerevisiae	~140	320–400	1.3×10^7
Tetrahymena thermophila	1	ND[b]	2×10^8
Drosophila melanogaster	130–250	~750	2×10^8
Human	~300	~1,300	3×10^9
Xenopus laevis	400–600	~7,800	8×10^9

Updated from Li (1983).

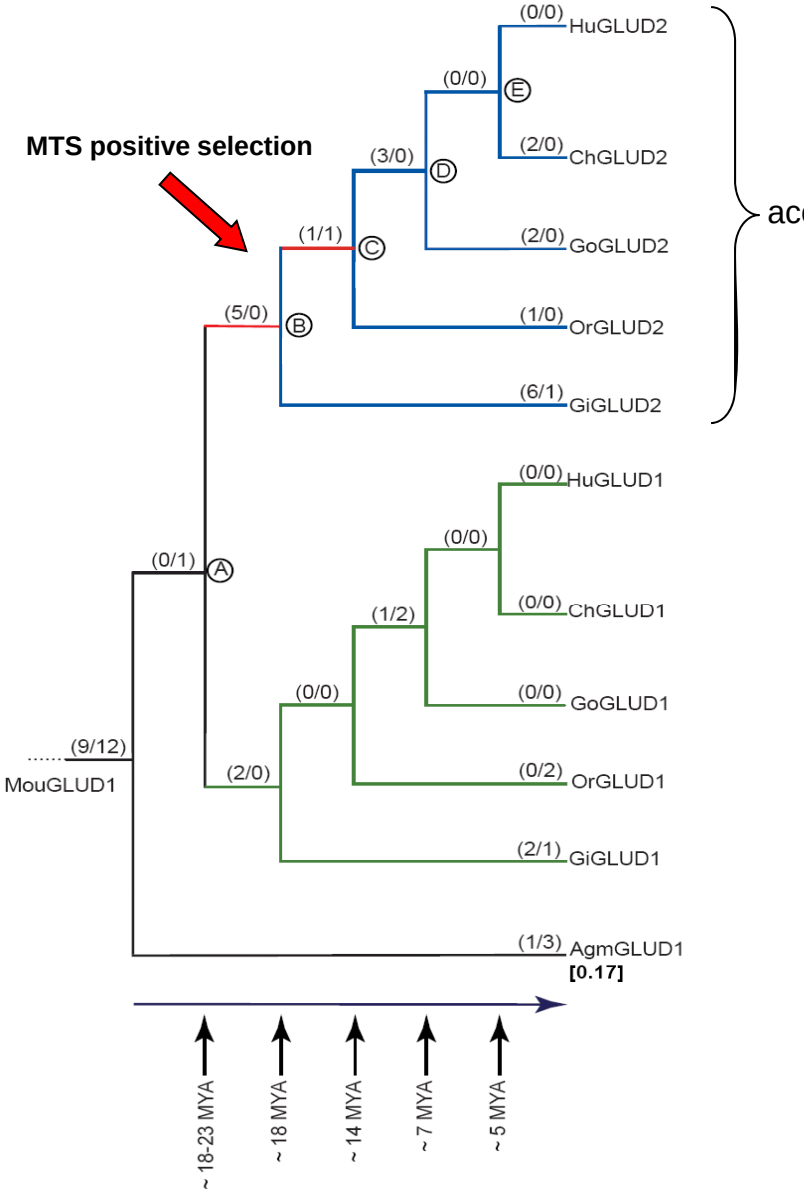
[a]For rRNA genes, the values refer to the number of complete sets of rRNA genes.

[b]ND = not determined.

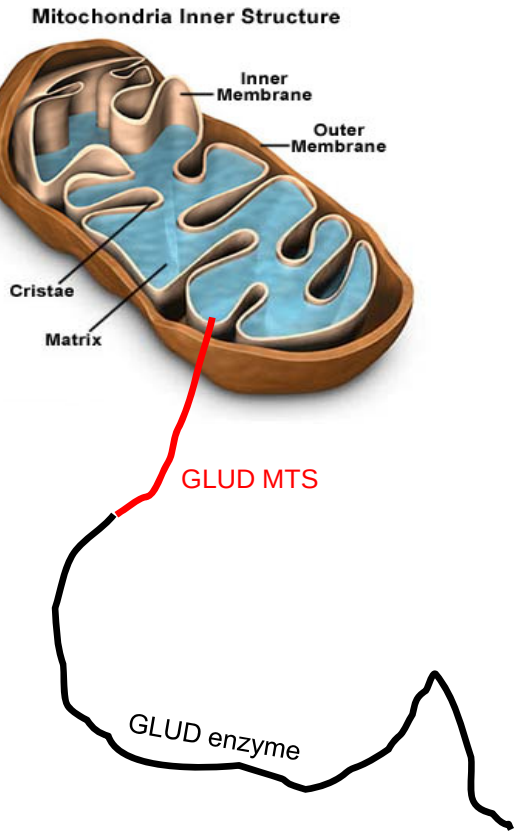
Evolutionary fate/role of new functional gene

- **Duplication for the sake of producing more of the same.**
- **Subfunctionalization**

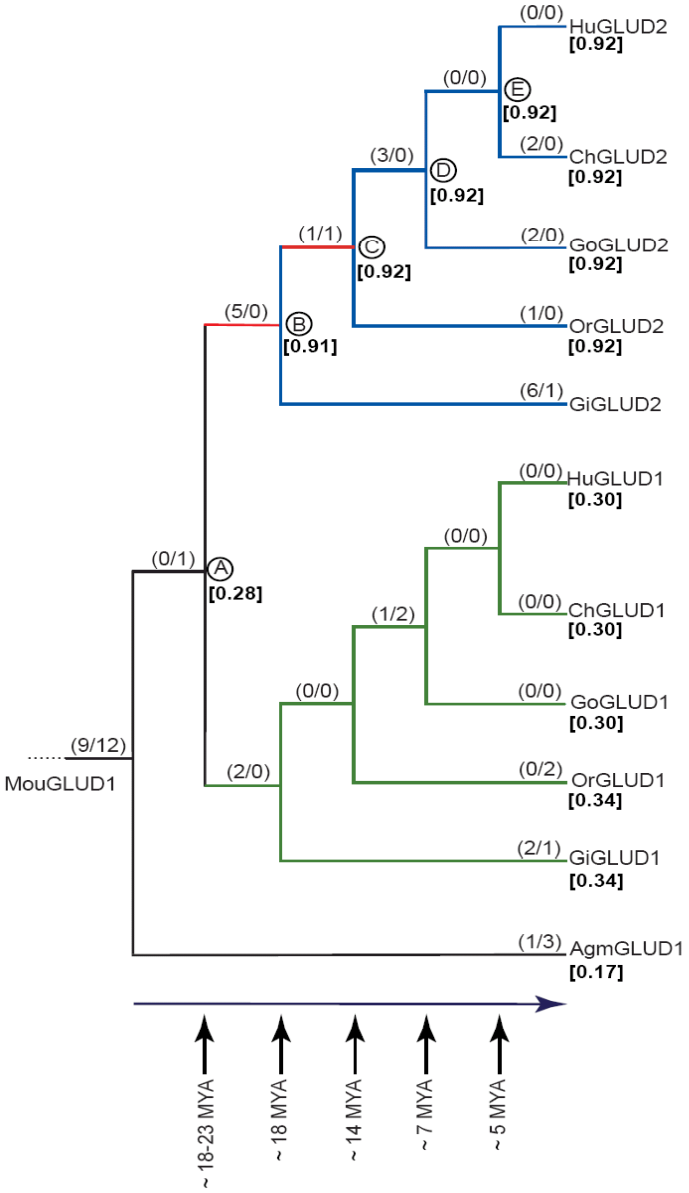
Subfunctionalization

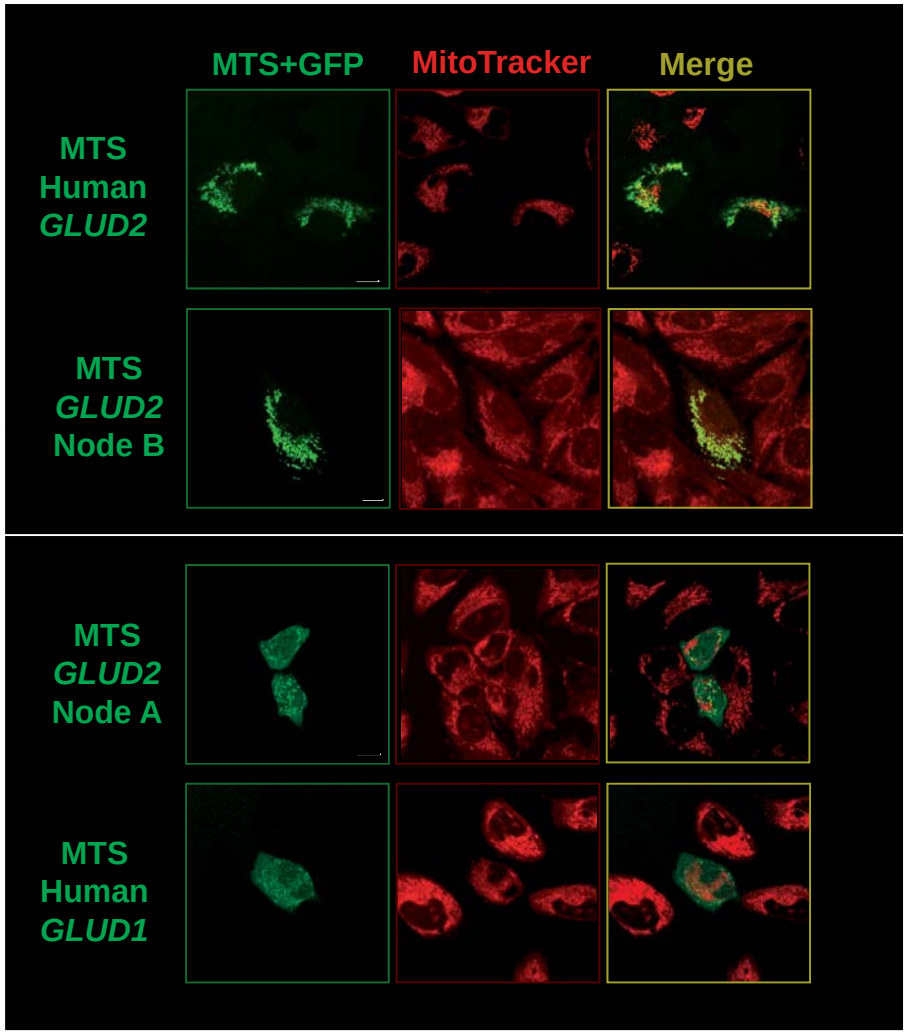


accelerated evolution of GLUD2 MTS ($P < 0.001$)



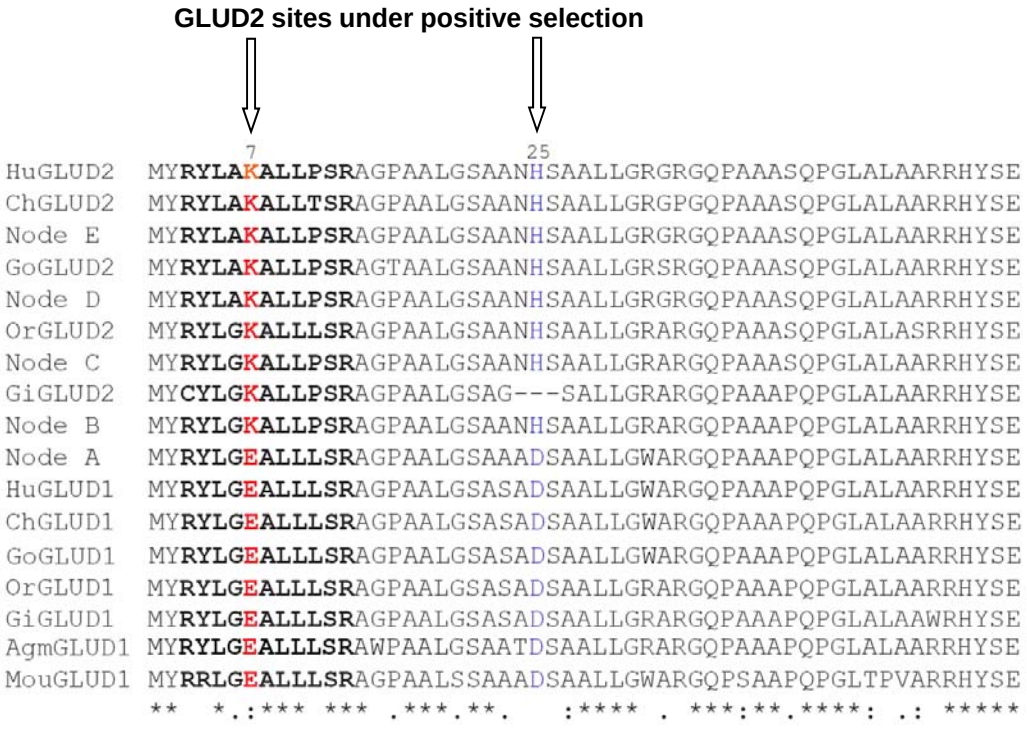
Subfunctionalization



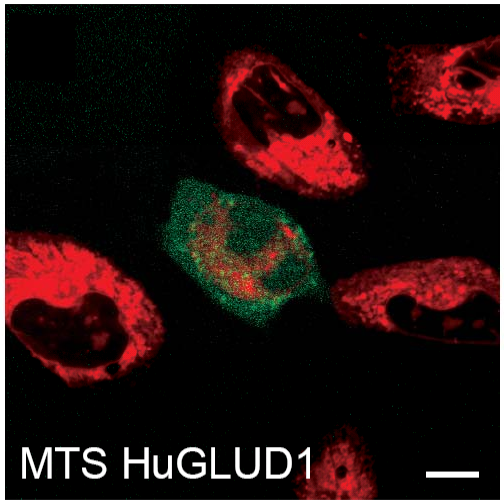


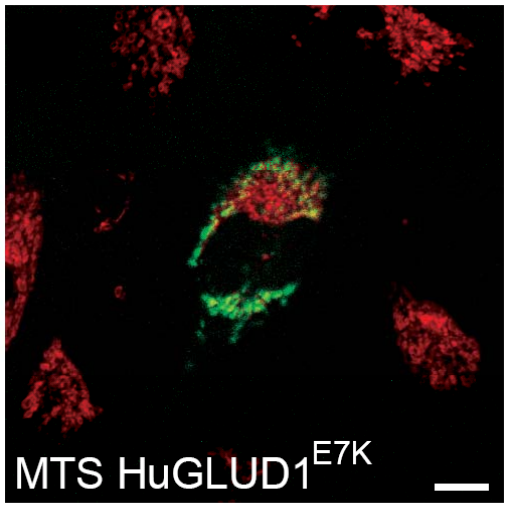
Subfunctionalization

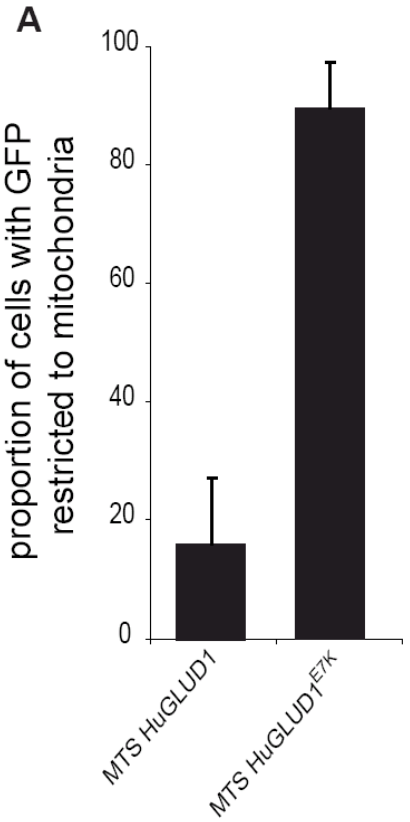
MTS alignment:



Subfunctionalization



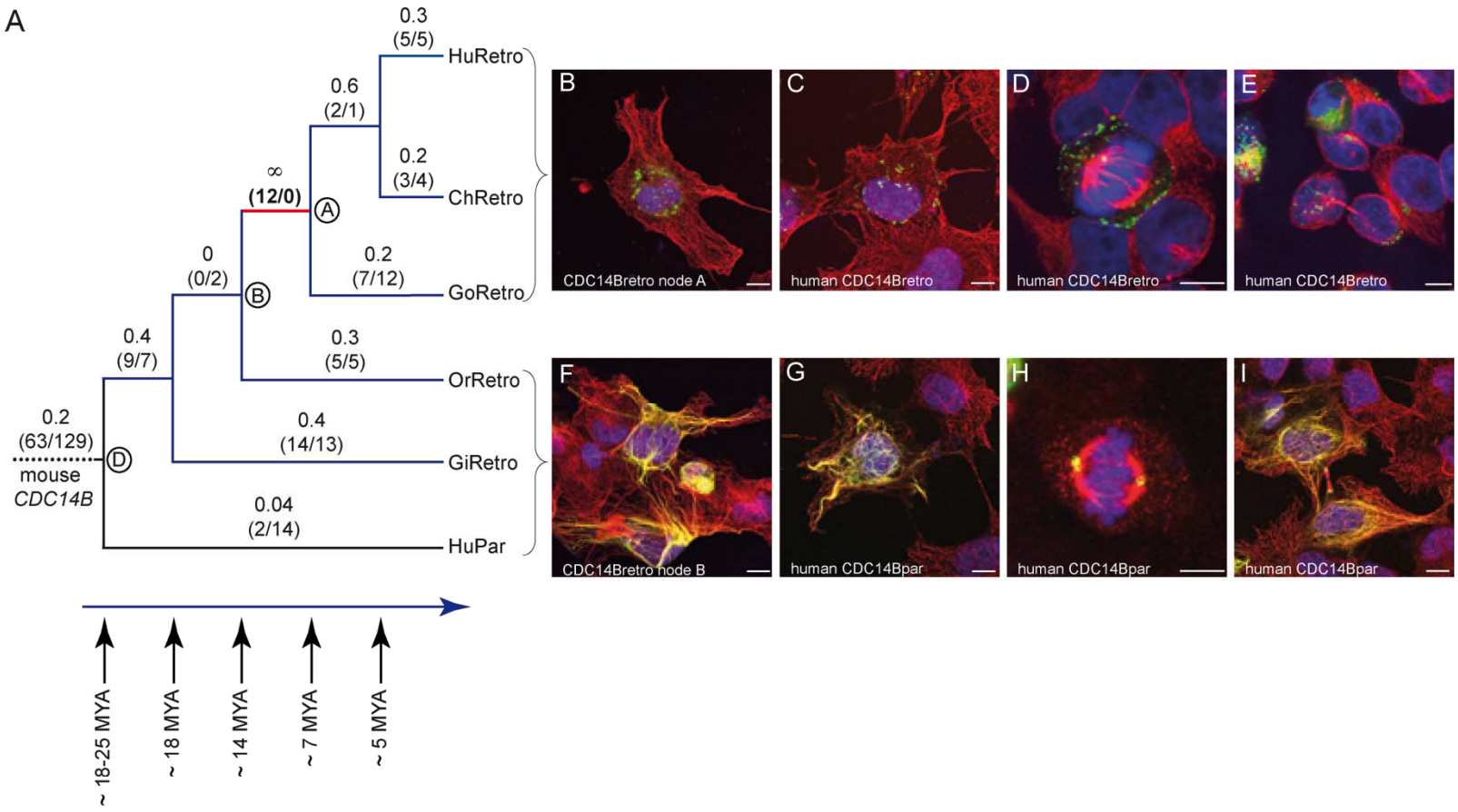




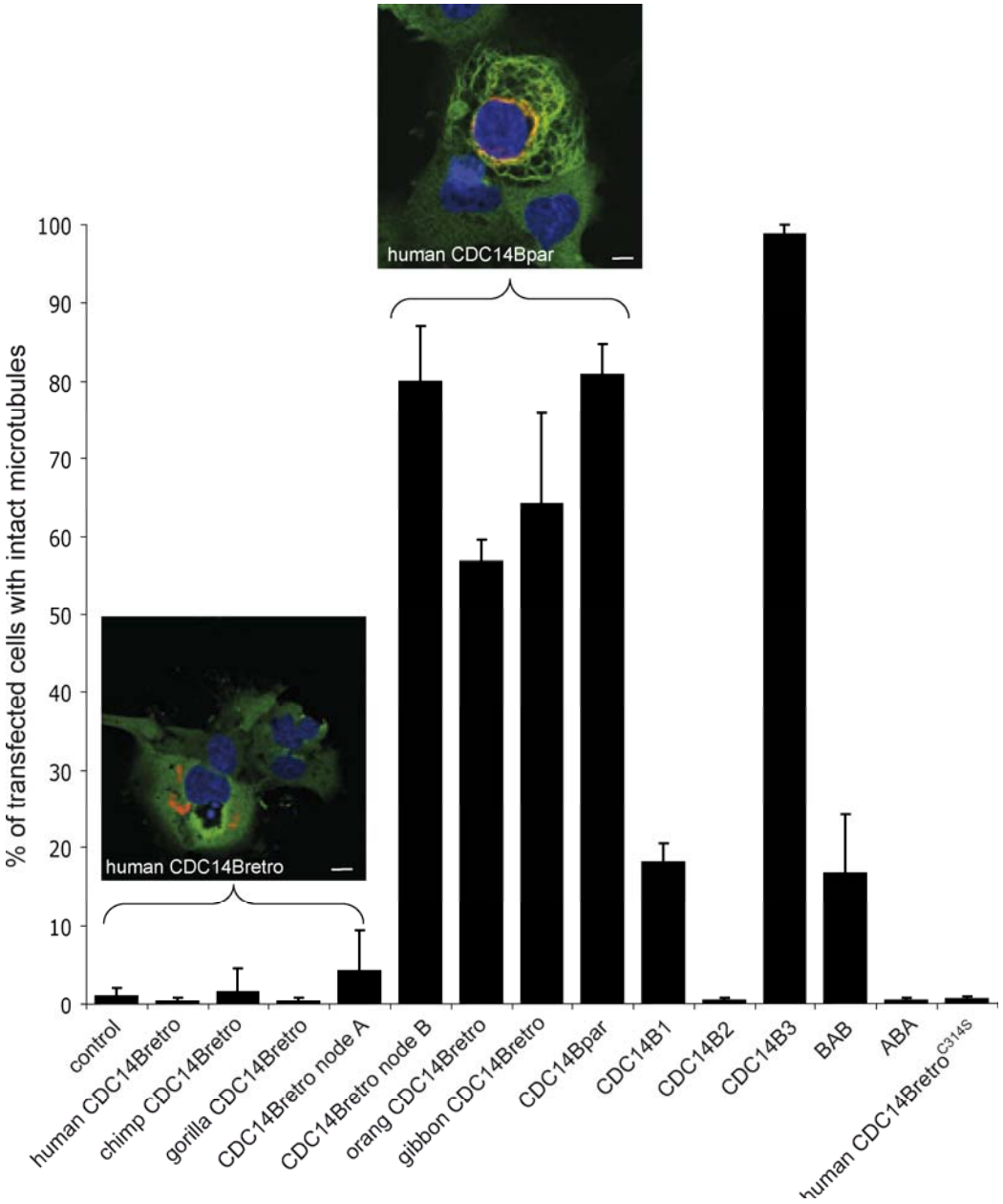
Evolutionary fate/role of new functional gene

- **Duplication for the sake of producing more of the same.**
- **Subfunctionalization**
- **Creation of a new gene function from a duplicate of an existing gene**

Duplication of protein coding genes



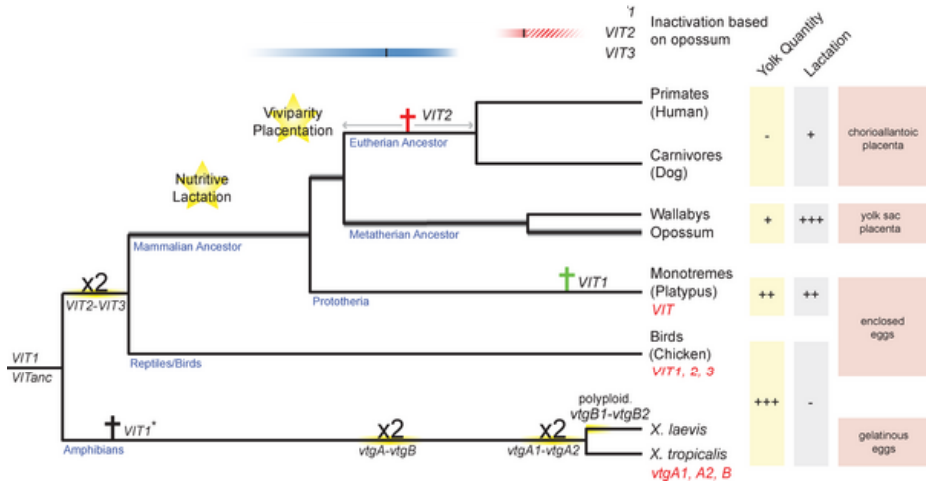
Duplication of protein coding genes



Gene Loss

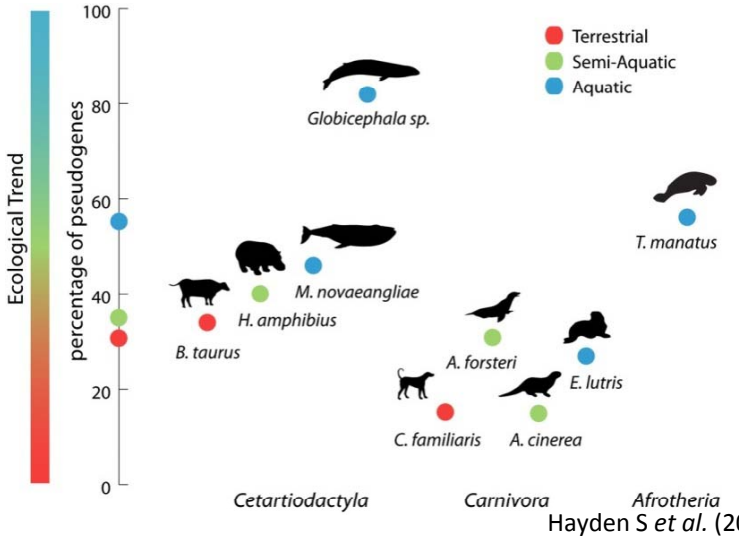
Gene loss is also associated with the origin of new traits.

Loss of egg yolk genes in mammals



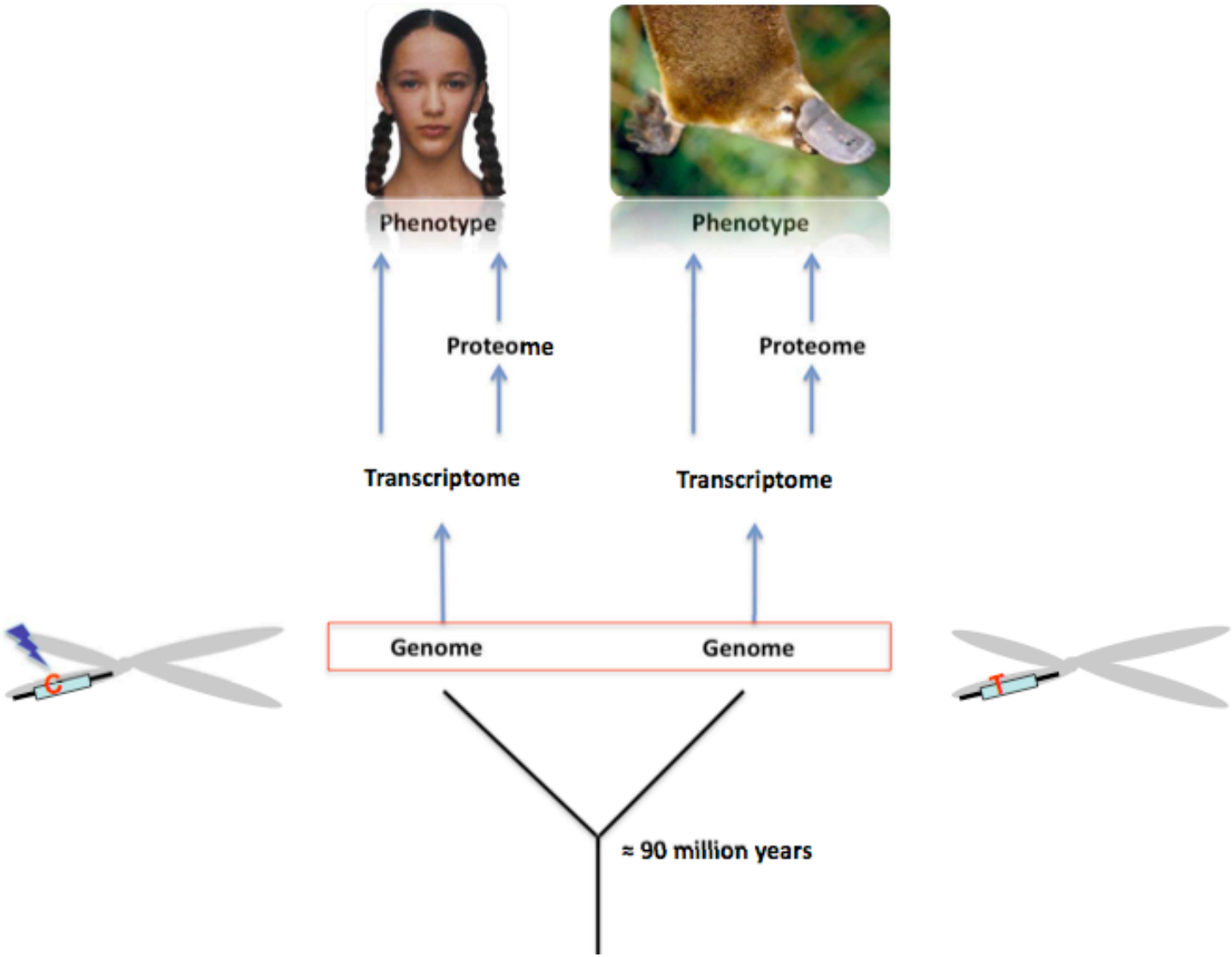
Brawand D. *et al.* (2006)

Loss of olfactory receptors



Hayden S *et al.* (2010)

How is this important ?



Challenges ahead

- How are the apparent differences in species “complexity” encoded?
- Are the ~19,000 genes in the genome the “important” bits?
  - How much genetic variation is determined epigenetically?
  - What is the function of the thousands of non-protein coding transcripts we find within the cell?
- Which genes are switched on in which tissues and at what developmental time-points?
- How much somatic variation is there?
- Currently, we can only explain <20% of the causes underlying many important diseases. How do we identify the cause of the rest?