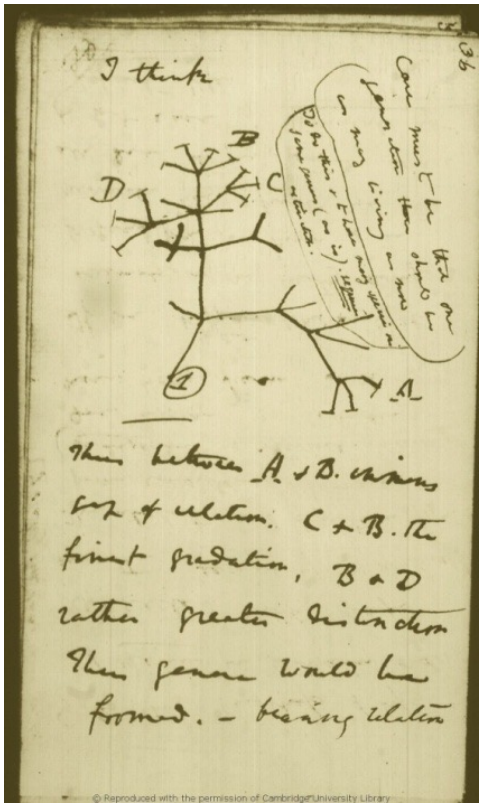
Gene expression and the evolutionary rate of genes

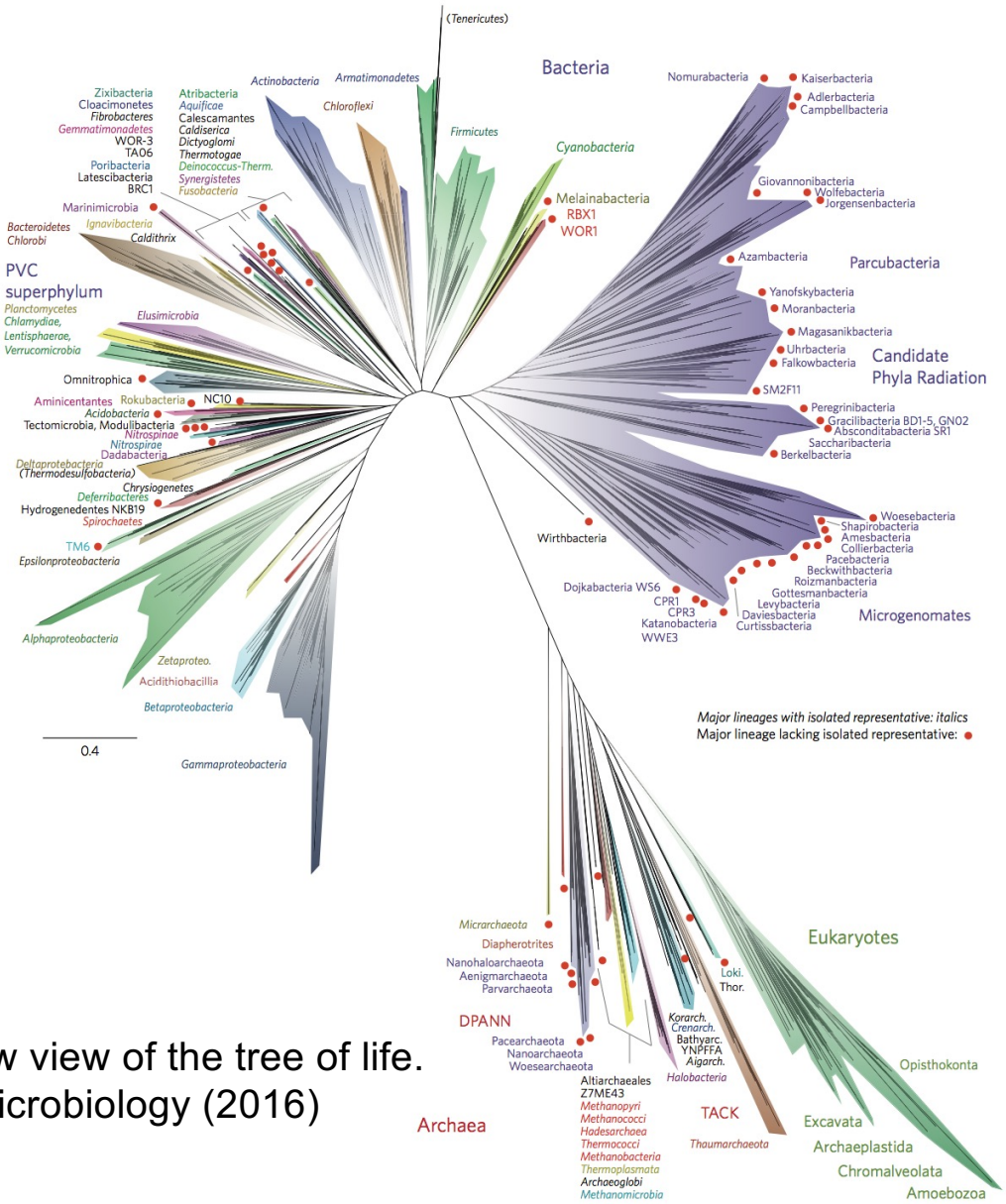
Searching for biological “laws”

Relationships between measured quantities that allow us to predict behavior outside of the measured range

Evolutionary theory



Darwin's sketch of the tree of life
On the origin of species (1869)

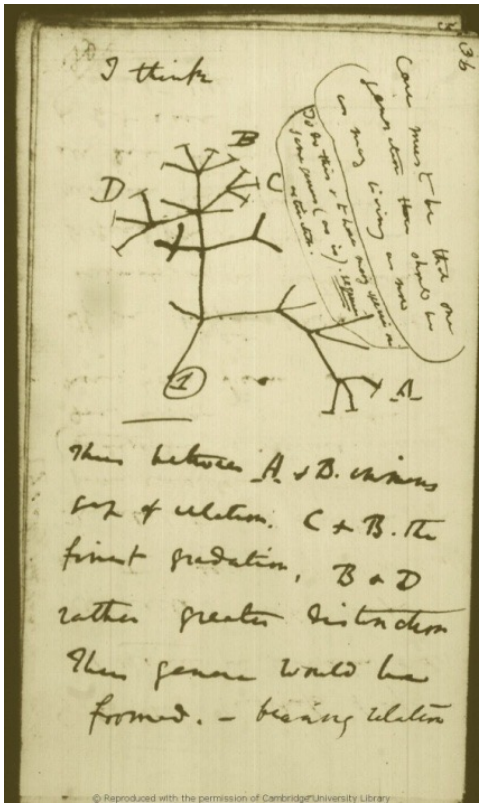


Hug et al. A new view of the tree of life.
Nature Microbiology (2016)

How do organisms evolve?

How do genes and proteins evolve?

Evolutionary theory



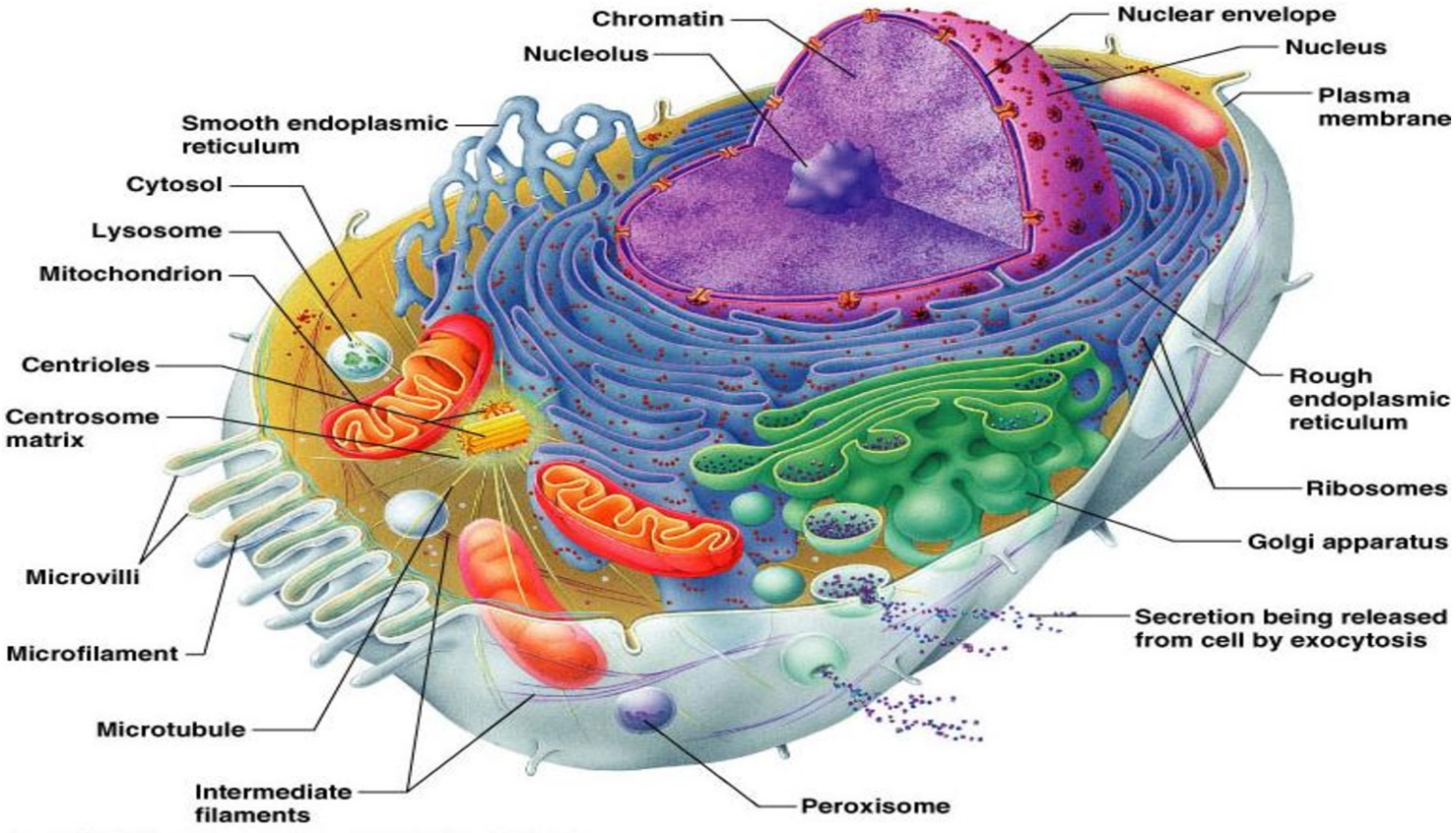
Darwin's sketch of the tree of life
On the origin of species (1869)

If:

1. entities that reproduce
2. with *heritable* variations
3. affecting reproductive success
4. competing for resources

Then: Evolution!

Structure of a Generalized Cell

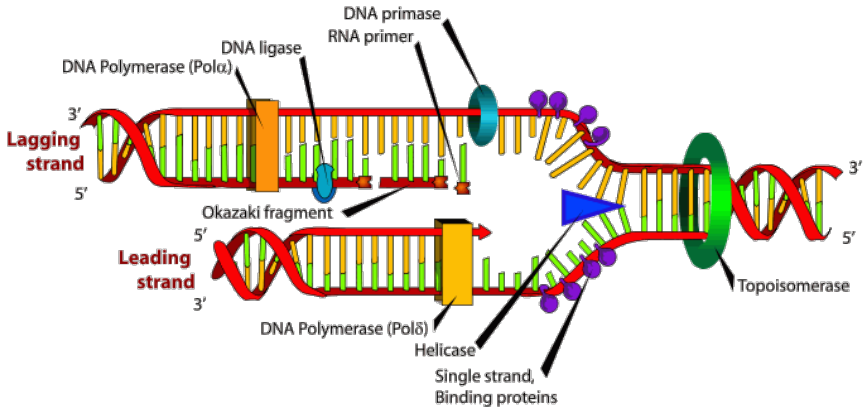


Pearson Education, Inc.

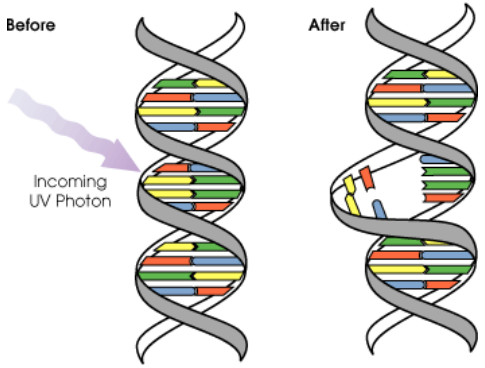
Genetic variation

- Living systems interact with their environment
- Molecular processes are subject to thermal ‘noise’
- Errors will occur
- As it is almost impossible to predict when, where and what error will occur, we describe them as *random*

DNA replication is error-prone

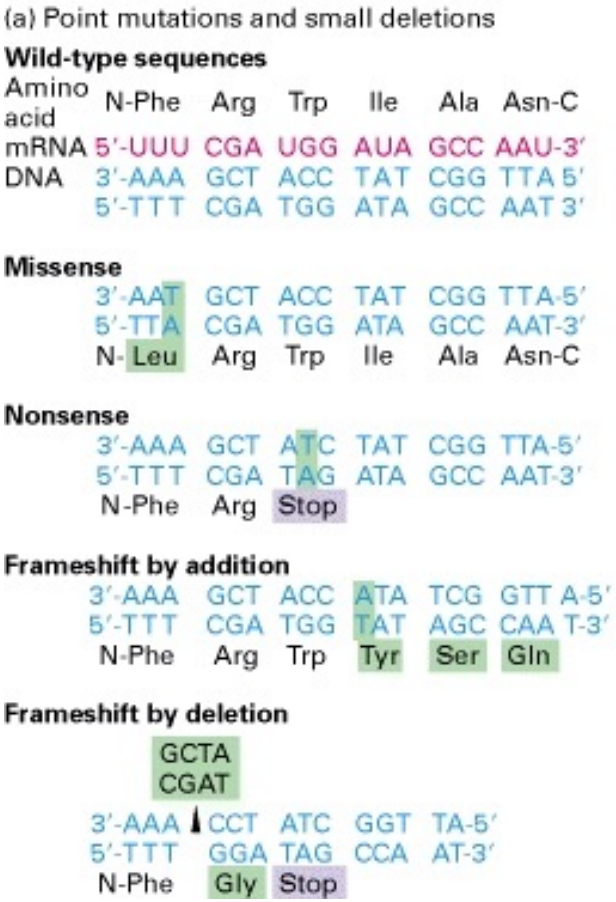


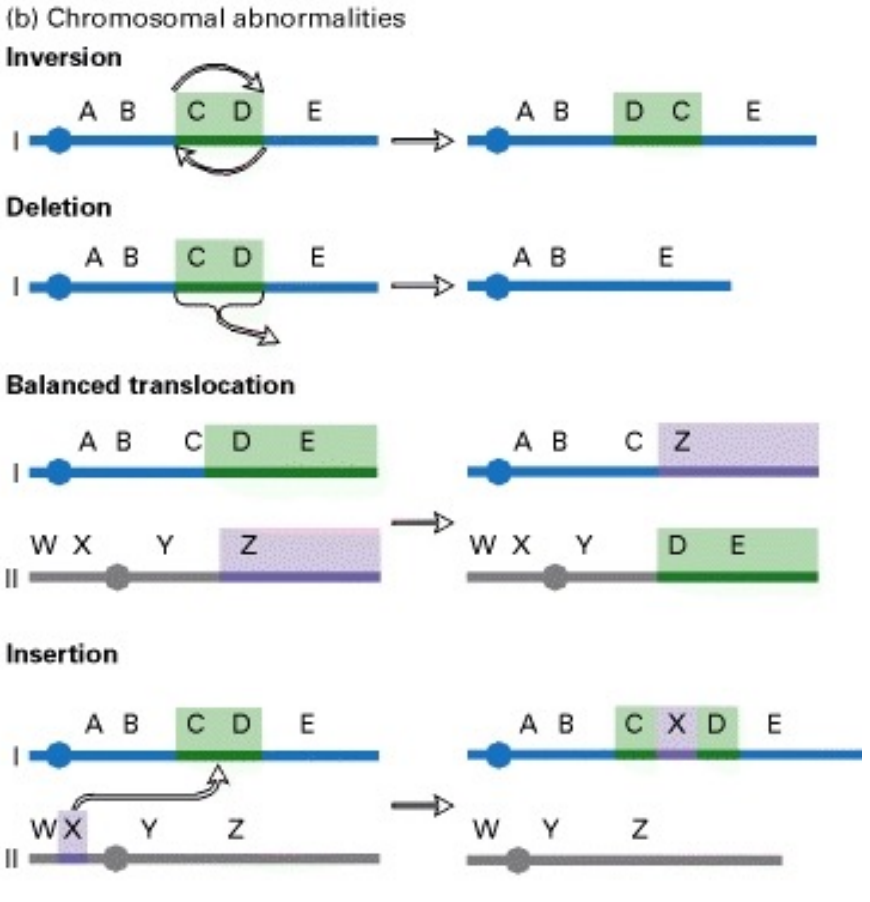
The DNA can get damaged



Selection operates on the background of random genetic variation

Some nomenclature





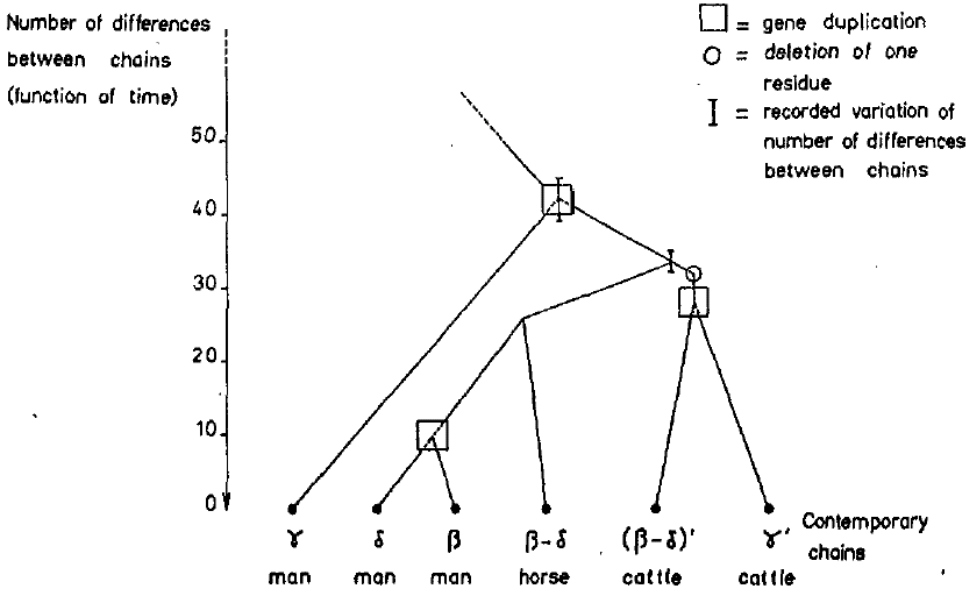
From: Lodish et al. *Molecular Cell Biology*

What patterns of genetic variation do we observe among living organisms?

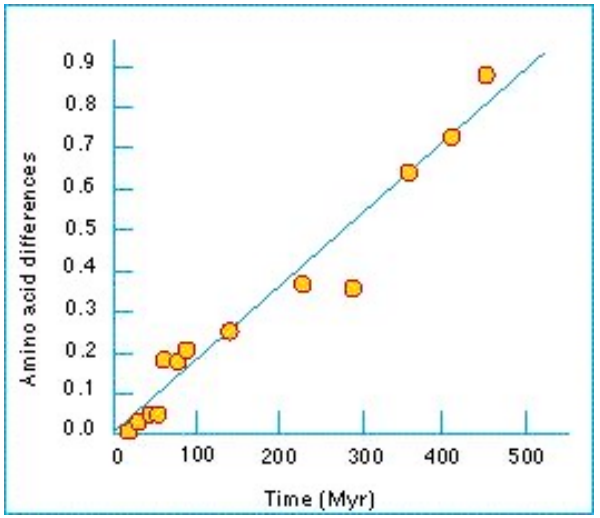
Can we explain them?

Describing patterns of genetic variation

Zuckerkandl & Pauling – Molecular Disease, Evolution and Genic Heterogeneity (1962)



Reconstructed phylogenetic tree of hemoglobin protein chains from man, horse, and cattle.



Comparison of the number of amino acid differences with the evolutionary distance estimated from the fossil record.

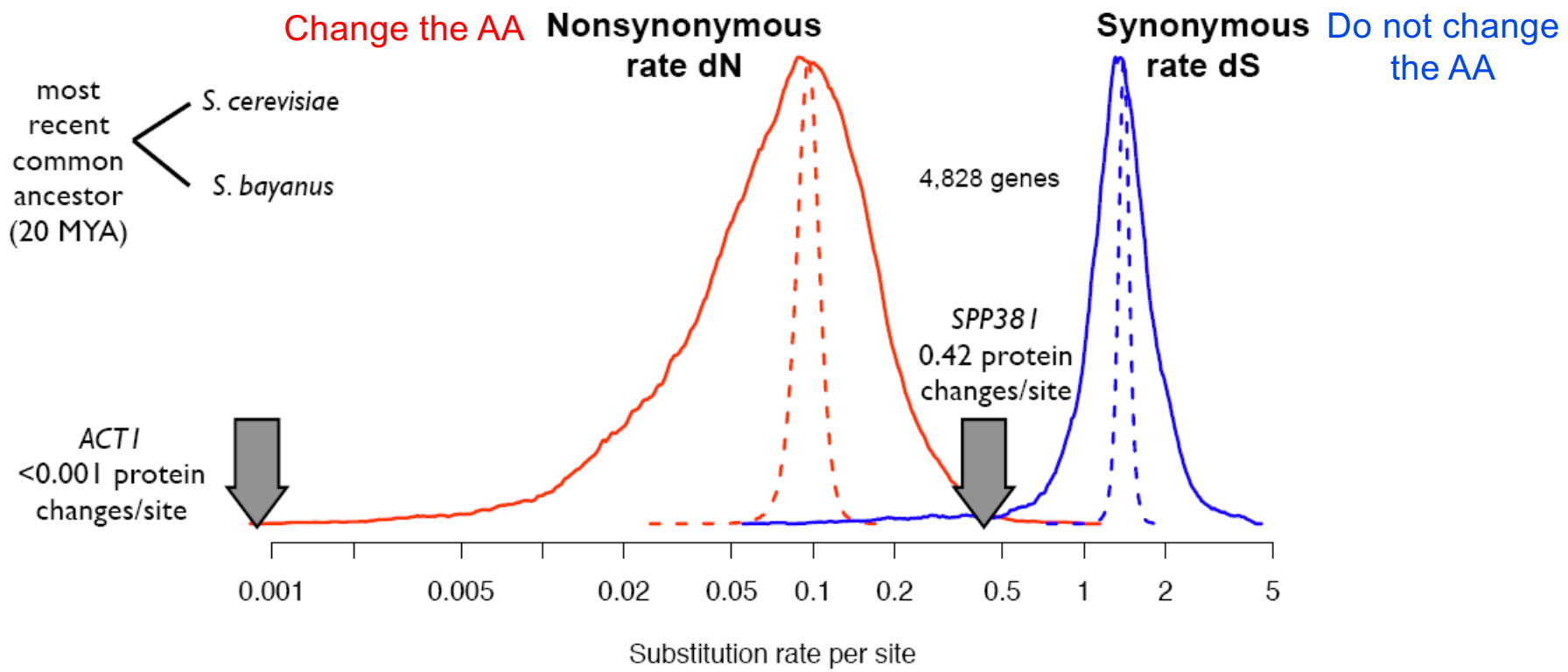
A quantitative theory that can give rise to the observed 'molecular clock' was proposed by Kimura in 1968: if the vast majority of single nucleotide changes are selectively *neutral,* then the probability of a mutation arising in an individual and spreading to the entire population will be roughly constant, and equal to the mutation rate.

Are we done?

Rates of substitution differ widely among genes

Comparison of substitution rates of orthologous genes

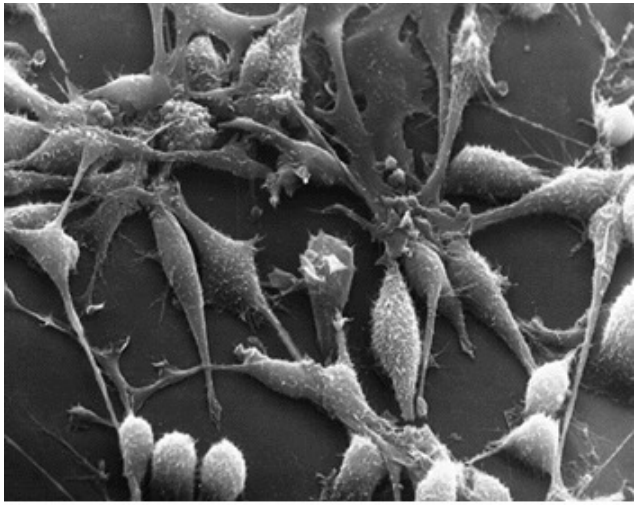
Courtesy Alain Drummond

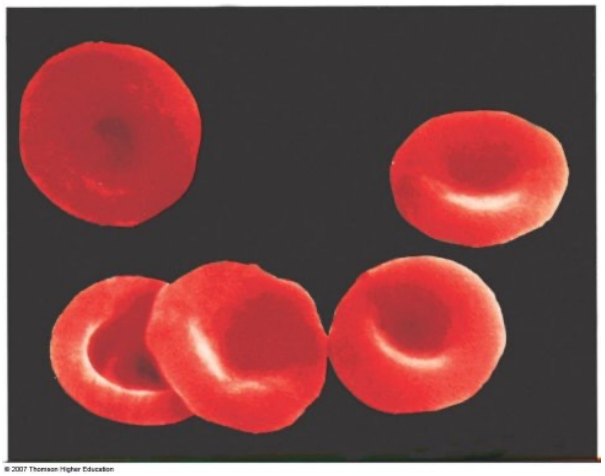


Distribution of rates per gene is much wider than expected for a Poisson process

Gene expression

Same DNA – different phenotypes





Cell type-specific gene expression

Genome annotation

Homo sapiens (human) **Build 37.1 (Current)** BLAST The Human Genome

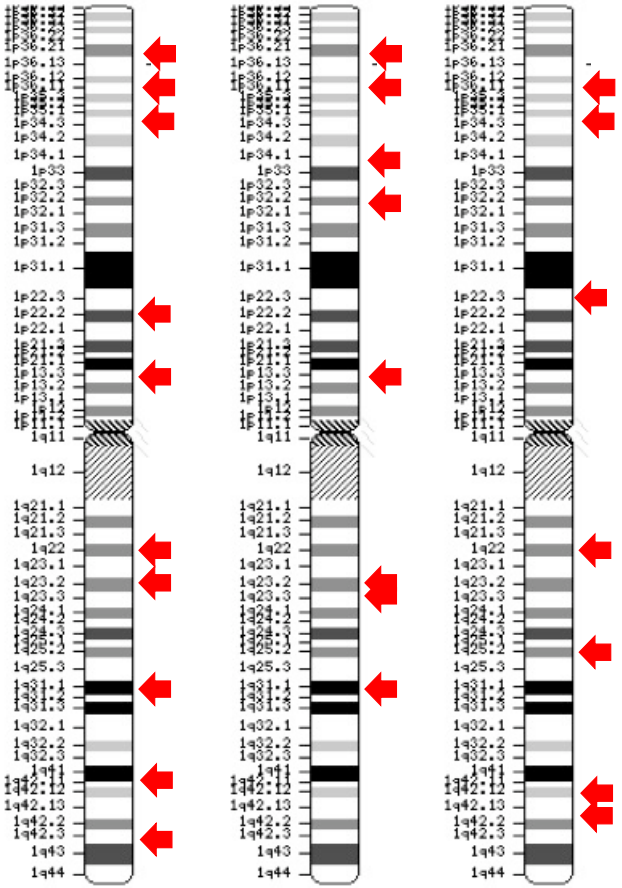
Chromosome: [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Master Map: Genes On Sequence Summary of Maps Maps & Options

Region Displayed: **0-250M bp** Download/View Sequence/Evidence

Ideogram Genes_seq

Symbol	O	Links	E	Cyto	Description
CALML6	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	calmodulin-like 6
LOC100287506	↑	sv dl ev mm	mRNA	1	hypothetical protein LOC100287506
RNU5E	↓	HGNC sv dl ev mm	best RefSeq	1	RNA, U5E small nuclear
HTR6	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1p36-p35	5-hydroxytryptamine (serotonin) receptor 6
SNIP1	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	Smad nuclear interacting protein 1
HEYL	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1p34.3	hairy/enhancer-of-split related with YRPW motif-like
DMBX1	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	diencephalon/mesencephalon homeobox 1
CYP4A11	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	cytochrome P450, family 4, subfamily A, polypeptide 11
L1TD1	↓	HGNC sv pr dl ev mm hm sts SNP	best RefSeq	1	LINE-1 type transposase domain containing 1
TGFBR3	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1p33-p32	transforming growth factor, beta receptor III
IGSF2	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1p13	immunoglobulin superfamily, member 2
LOC100132999	↓	sv pr dl ev mm hm sts SNP	mRNA	1	hypothetical protein LOC100132999
TDRKH	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1q21	tudor and KH domain containing
NBPF18P	↓	HGNC sv dl ev mm sts	best RefSeq	1	neuroblastoma breakpoint family, member 18 (pseudogene)
ATP8B2	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	ATPase, class I, type 8B, member 2
RXFP4	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	relaxin/insulin-like family peptide receptor 4
MAEL	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1	maelstrom homolog (Drosophila)
PRELP	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	1q32	proline/arginine-rich end leucine-rich repeat protein
RPL13AP11	↓	HGNC sv dl ev mm	best RefSeq	1q32.1	ribosomal protein L13a pseudogene 11
OR14A2	↑	HGNC sv pr dl ev mm	best RefSeq	1	olfactory receptor, family 14, subfamily A, member 2

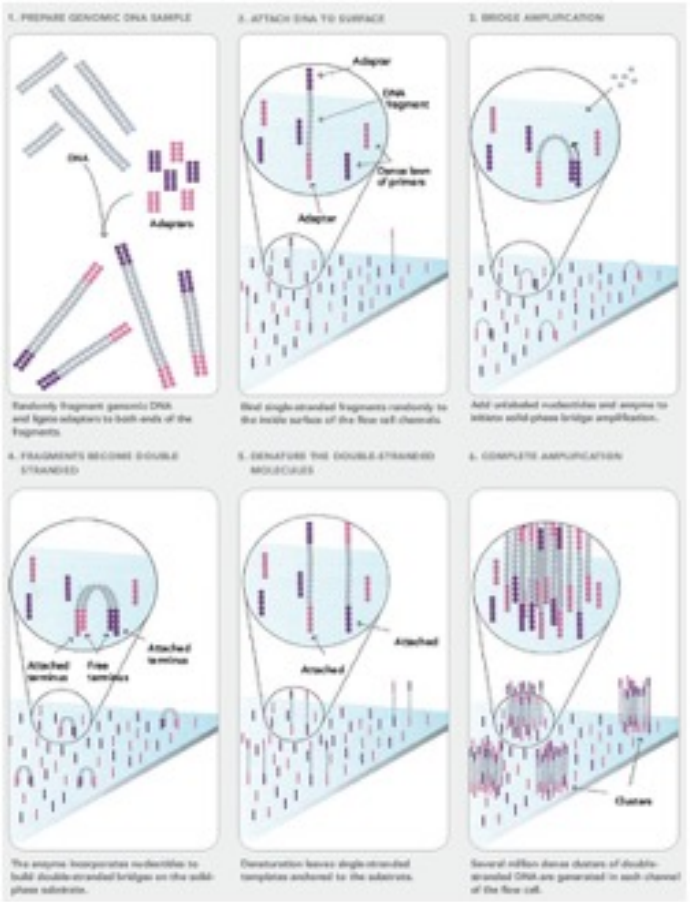


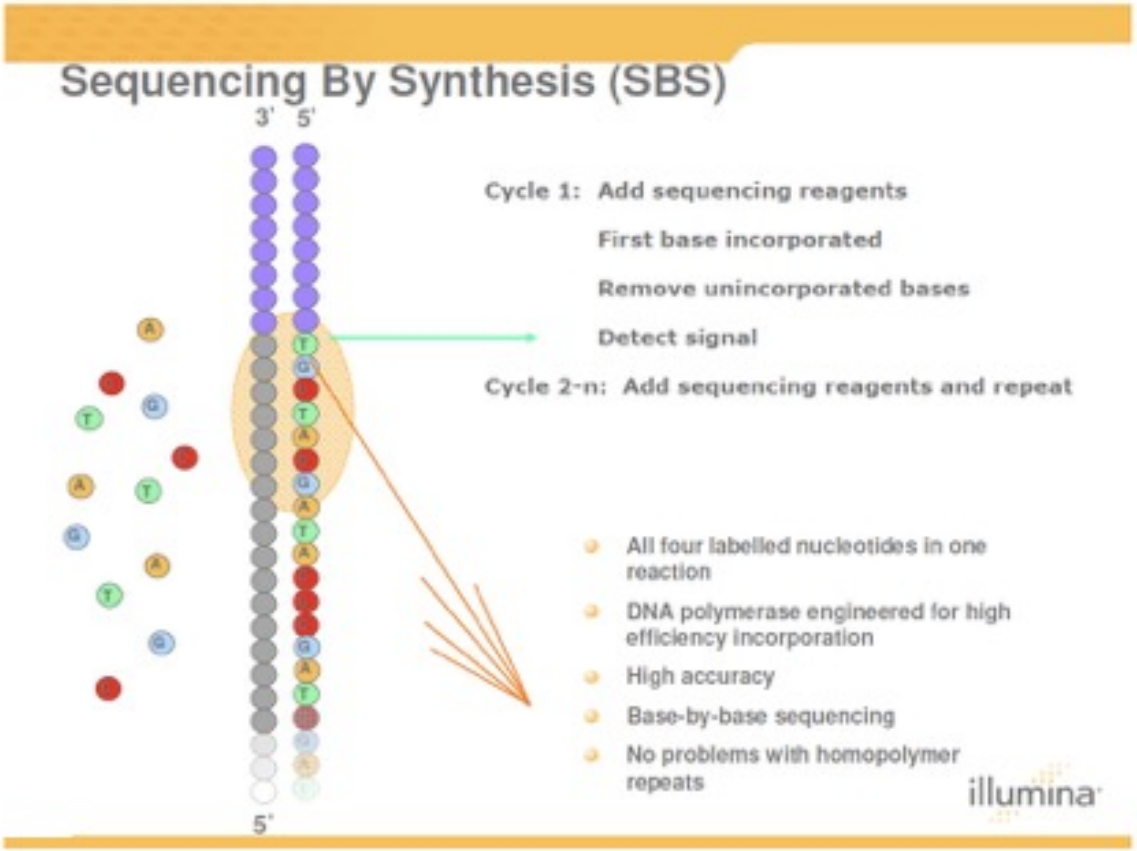
Cell type 1 Cell type 2 Cell type 3

How do we know that different cells express different sets of genes?

We can identify all RNAs that are present in a cell

Short read sequencing on the Illumina platform





What drives the variation in the substitution rate among genes?