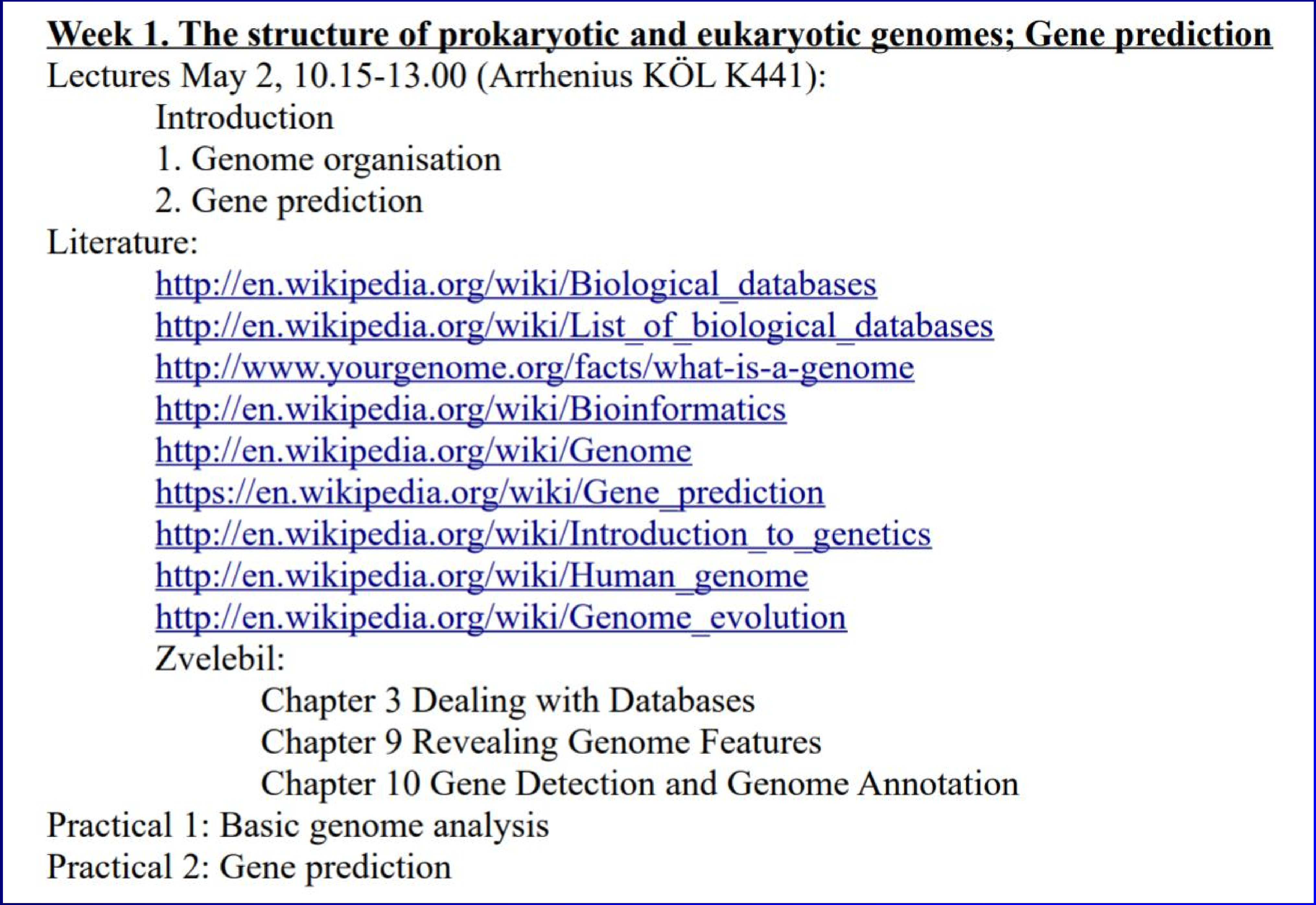
Comparative Genomics

1: Genome organization

Zvelebil Chapters 9, 10

Schedule

http://www.nada.kth.se/~erison/KB8019_Comparative_Genomics_Schedule_2018.pdf



Week 1. The structure of prokaryotic and eukaryotic genomes; Gene prediction

Lectures May 2, 10.15-13.00 (Arrhenius KÖL K441):

Introduction

1. Genome organisation
2. Gene prediction

Literature:

http://en.wikipedia.org/wiki/Biological_databases
http://en.wikipedia.org/wiki/List_of_biological_databases
http://www.yourgenome.org/facts/what-is-a-genome
http://en.wikipedia.org/wiki/Bioinformatics
http://en.wikipedia.org/wiki/Genome
https://en.wikipedia.org/wiki/Gene_prediction
http://en.wikipedia.org/wiki/Introduction_to_genetics
http://en.wikipedia.org/wiki/Human_genome
http://en.wikipedia.org/wiki/Genome_evolution

Zvelebil:

Chapter 3 Dealing with Databases
Chapter 9 Revealing Genome Features
Chapter 10 Gene Detection and Genome Annotation

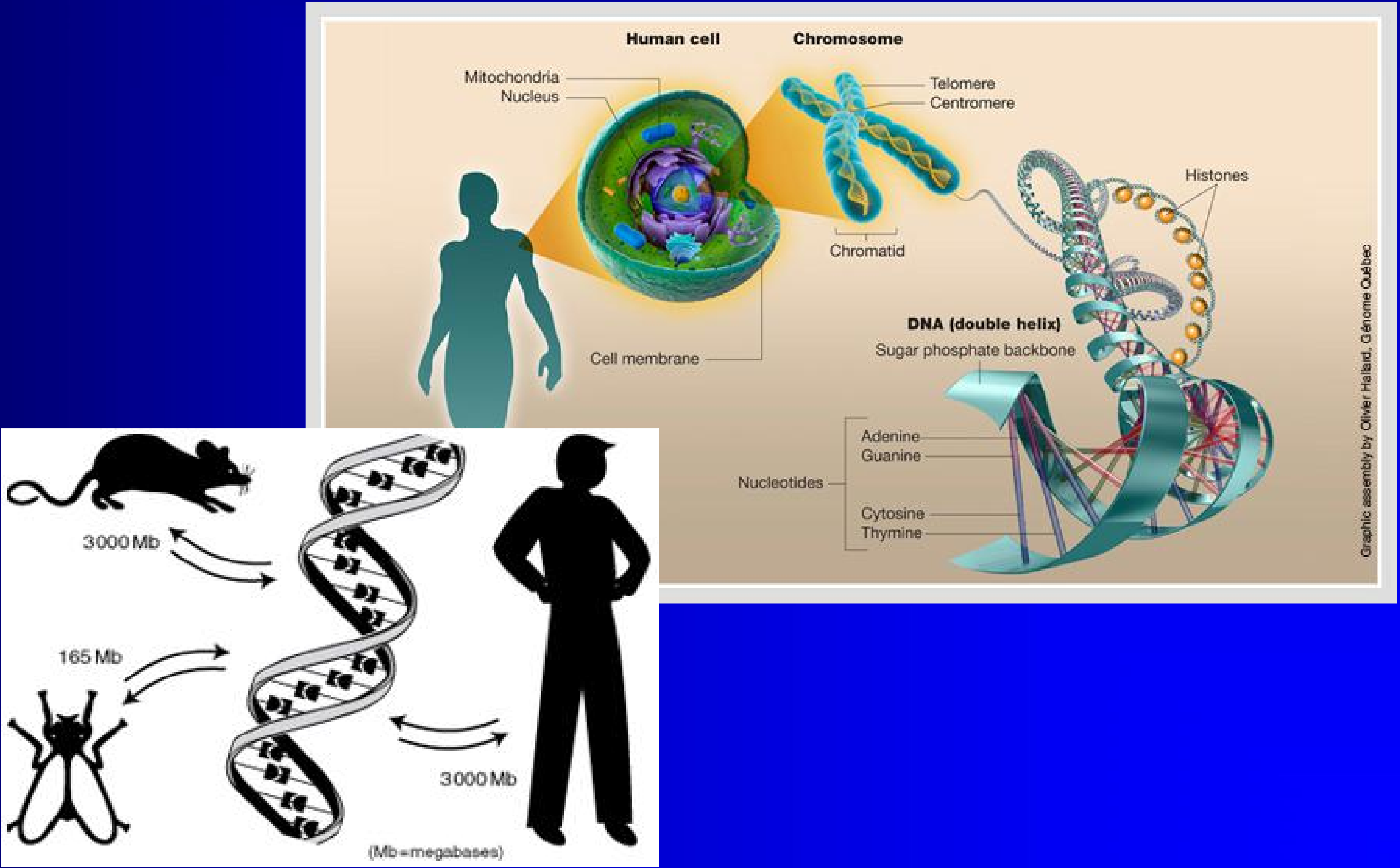
Practical 1: Basic genome analysis

Practical 2: Gene prediction

Outline

- Introduction to genomes
- Genomes in different species
- Gene content
- Regulatory sequences
- Non-coding sequences
- Metagenomics

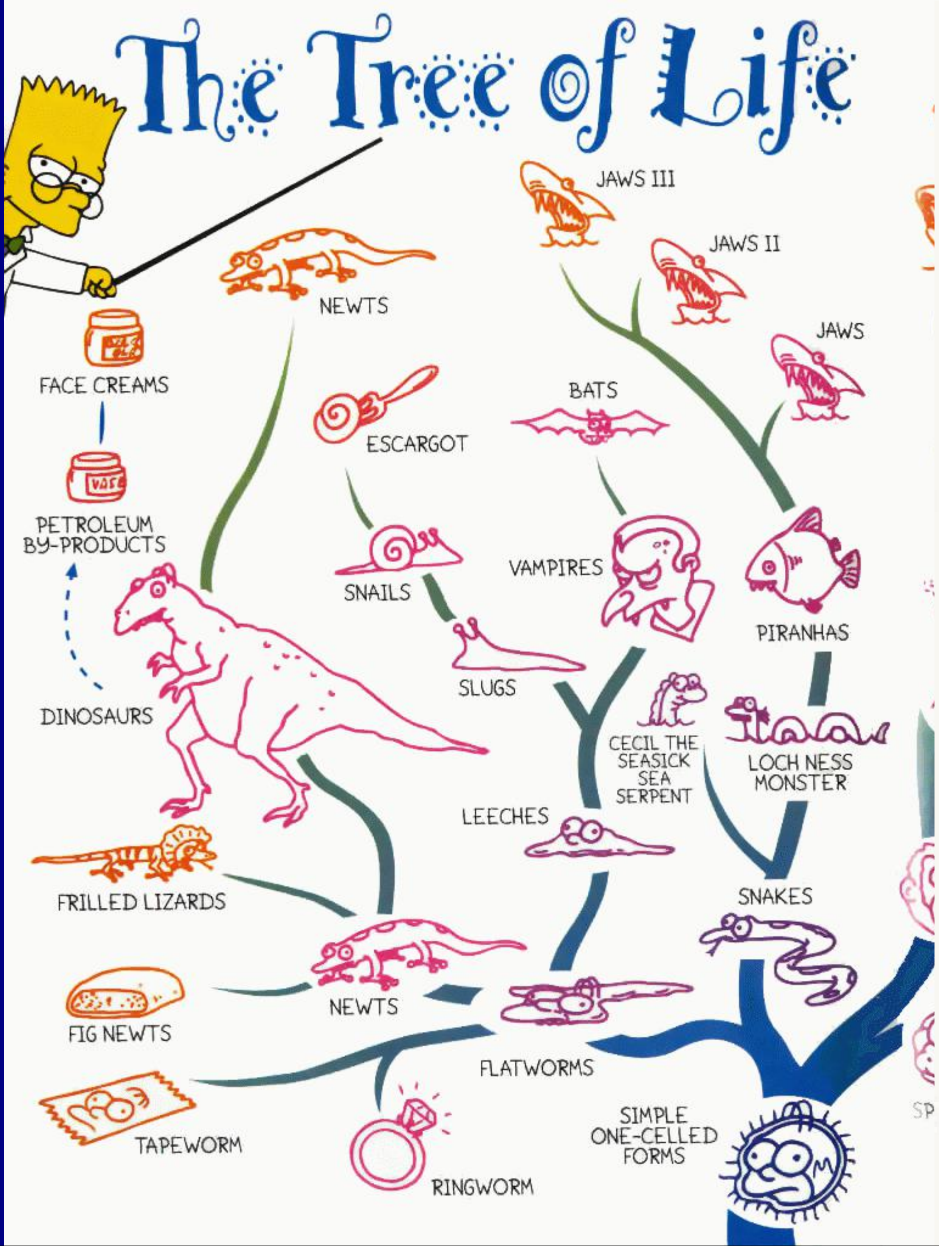
The genome


Human cell
Chromosome
Mitochondria
Nucleus
Telomere
Centromere
Histones
Chromatid
DNA (double helix)
Sugar phosphate backbone
Cell membrane
Adenine
Guanine
Nucleotides
Cytosine
Thymine
Graphic assembly by Oliver Hallard, Génome Québec
3000 Mb
165 Mb
3000 Mb
(Mb=megabases)

What’s a genome?

- Small set of huge DNA molecules, identical in (almost) all cells of an organism.
- Carrier of hereditary information and the ultimate substrate on which evolution works by mutation and selection.
- Defines virtually all activities of a cell by encoding its proteins, catalytic RNAs and regulatory mechanisms for these.
- Organised in chromosomes.
- Circular (prokaryotes) or linear (eukaryotes).
- Also viruses and some organelles (cell parts) have genomes

Comparative genomics is based on species relations



How many species exist?

General definition: different species don't interbreed (because of genome differences)

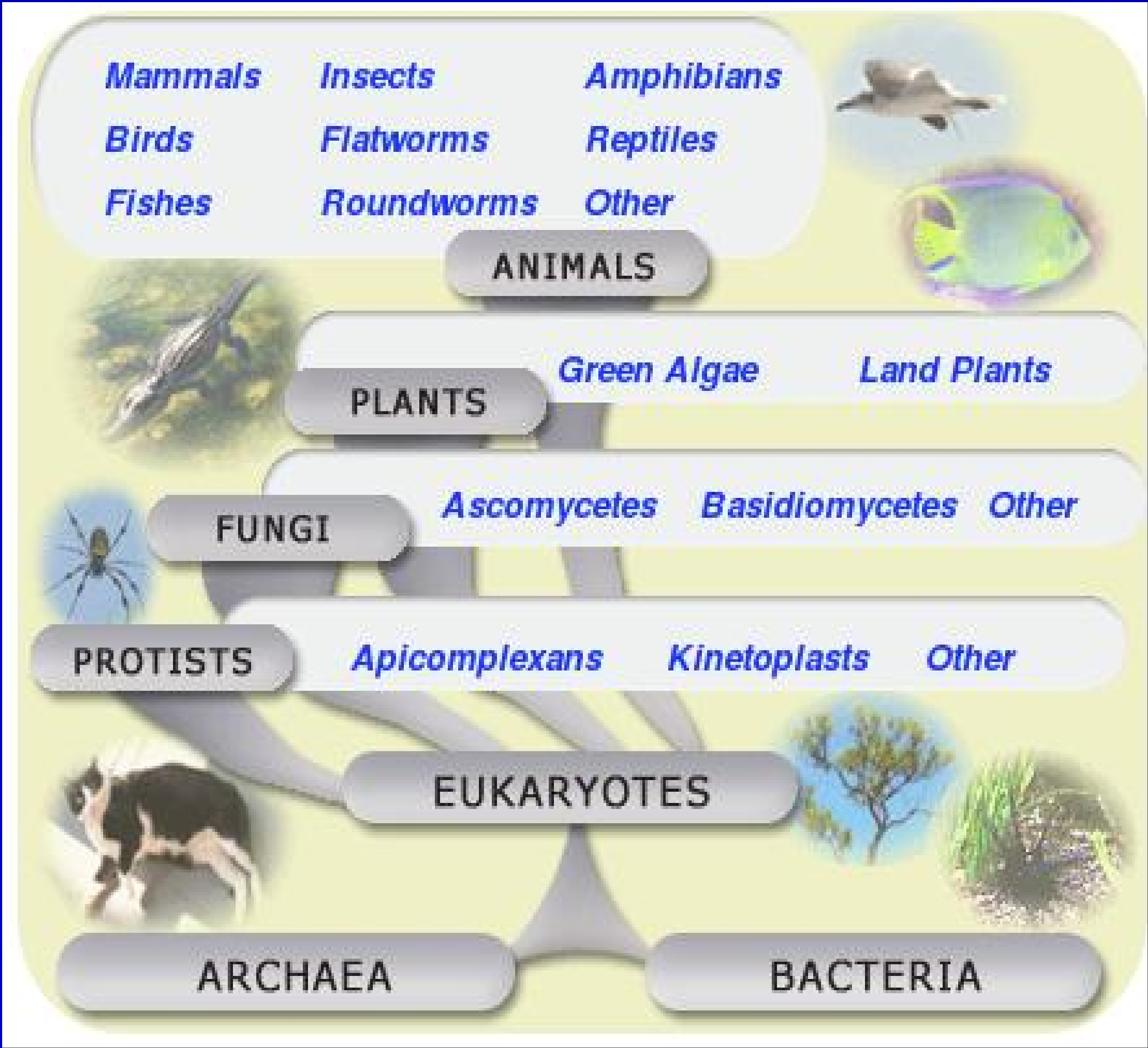
~1.7 million species are named so far.

Each year, about 13,000 more species are added

Estimates of up to 100 million species.

How many genomes are completely sequenced?

www.ncbi.nlm.nih.gov/genome



How many genomes are completely sequenced?

www.ncbi.nlm.nih.gov/genome 2016-04-26:

	‘Complete genomes’ (some gaps)
Virus:	5496
Bacteria:	5915
Archaea:	257
Eukaryota:	363

(Only genomes up to ~42 Mbases are 100.0% complete)

How many genomes are completely sequenced?

www.ncbi.nlm.nih.gov/genome 2017-04-27:

	'Complete genomes' (some gaps)
Virus:	7141
Bacteria:	8379
Archaea:	277
Eukaryota:	692

(Largest 100.0% complete genome: *C. elegans*, 100 Mbp)

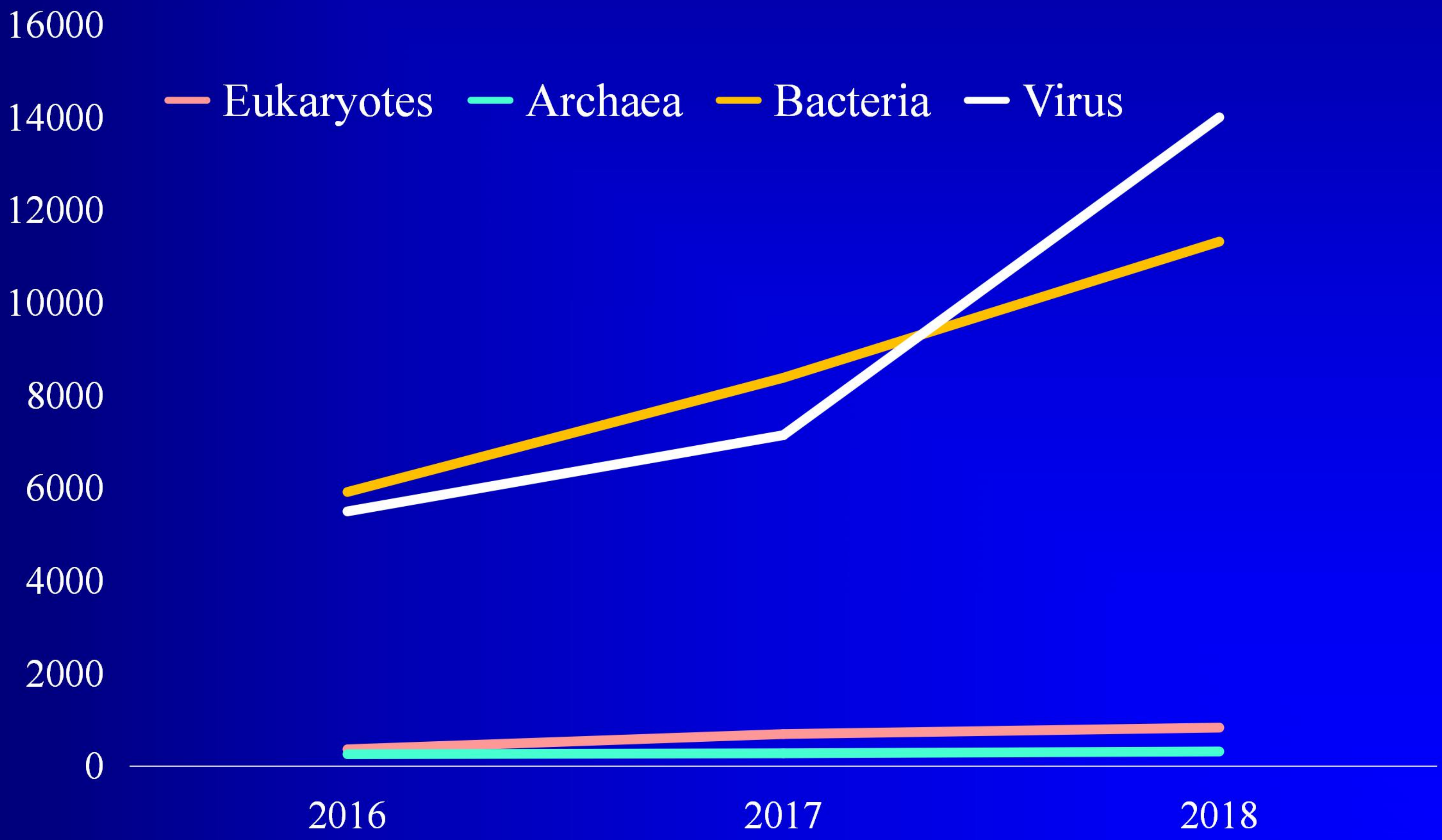
How many genomes are completely sequenced?

www.ncbi.nlm.nih.gov/genome 2018-04-19:

	‘Complete genomes’ (some gaps)
Virus:	13999
Bacteria:	11317
Archaea:	316
Eukaryota:	831

(The largest 100.0% complete is 100 Mbp)

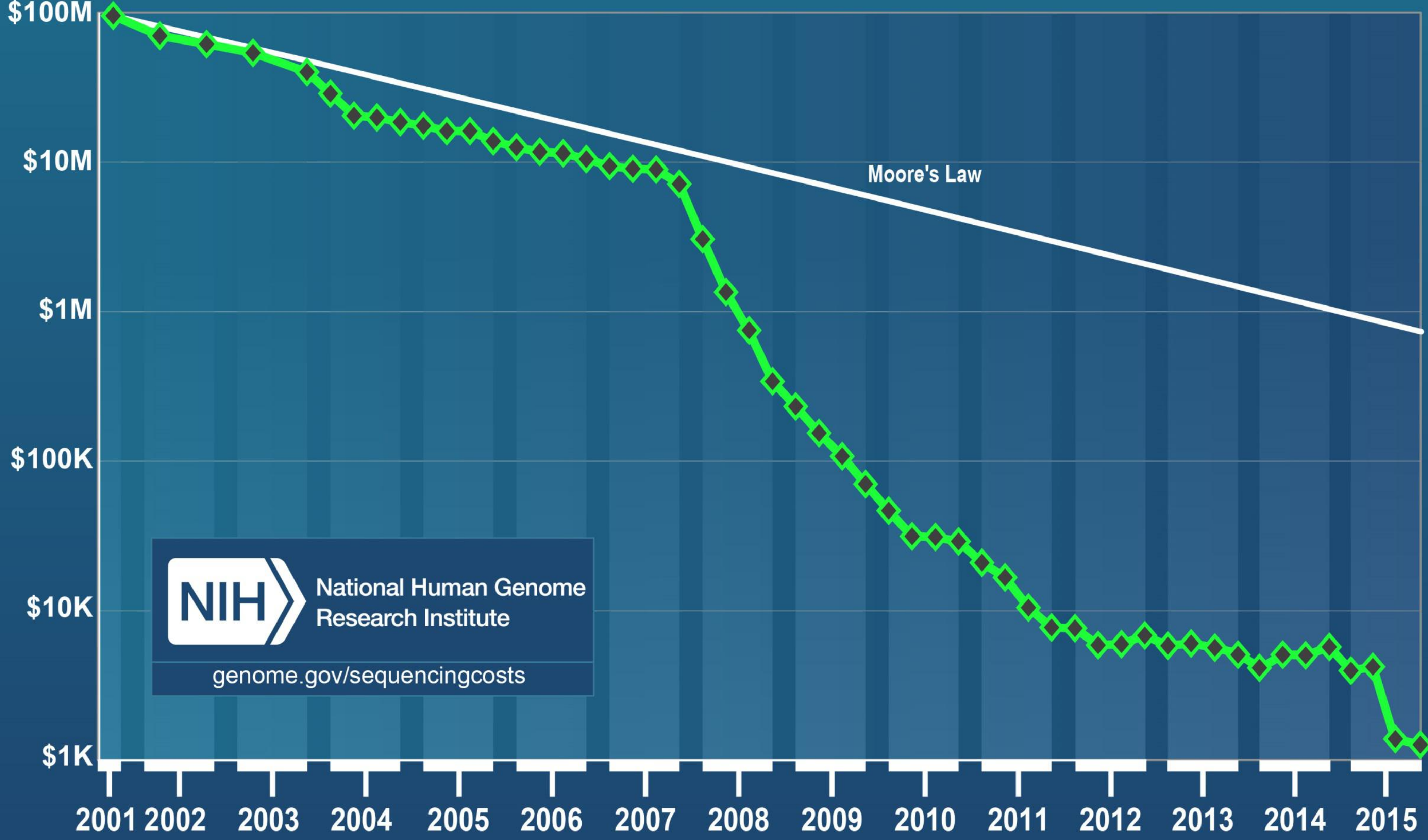
Number of completely sequenced genomes



Milestones in whole genome sequencing

Year	Clade	Species	Genome size, bp
1977	Bacteriophage	Φ X174	5386
1995	Bacterium	*H. influenzae*	1.8 x 10^{6}
1996	Eukaryote	Brewer’s yeast	1.2 x 10^{7}
1998	Animal	*C. elegans*	1.0 x 10^{8}
2000	Plant	*A. thaliana*	1.3 x 10^{8}
2001	Mammal	Human	3.2 x 10^{9}
2013	Tree	Spruce	2.5 x 10^{10}

Cost per Genome

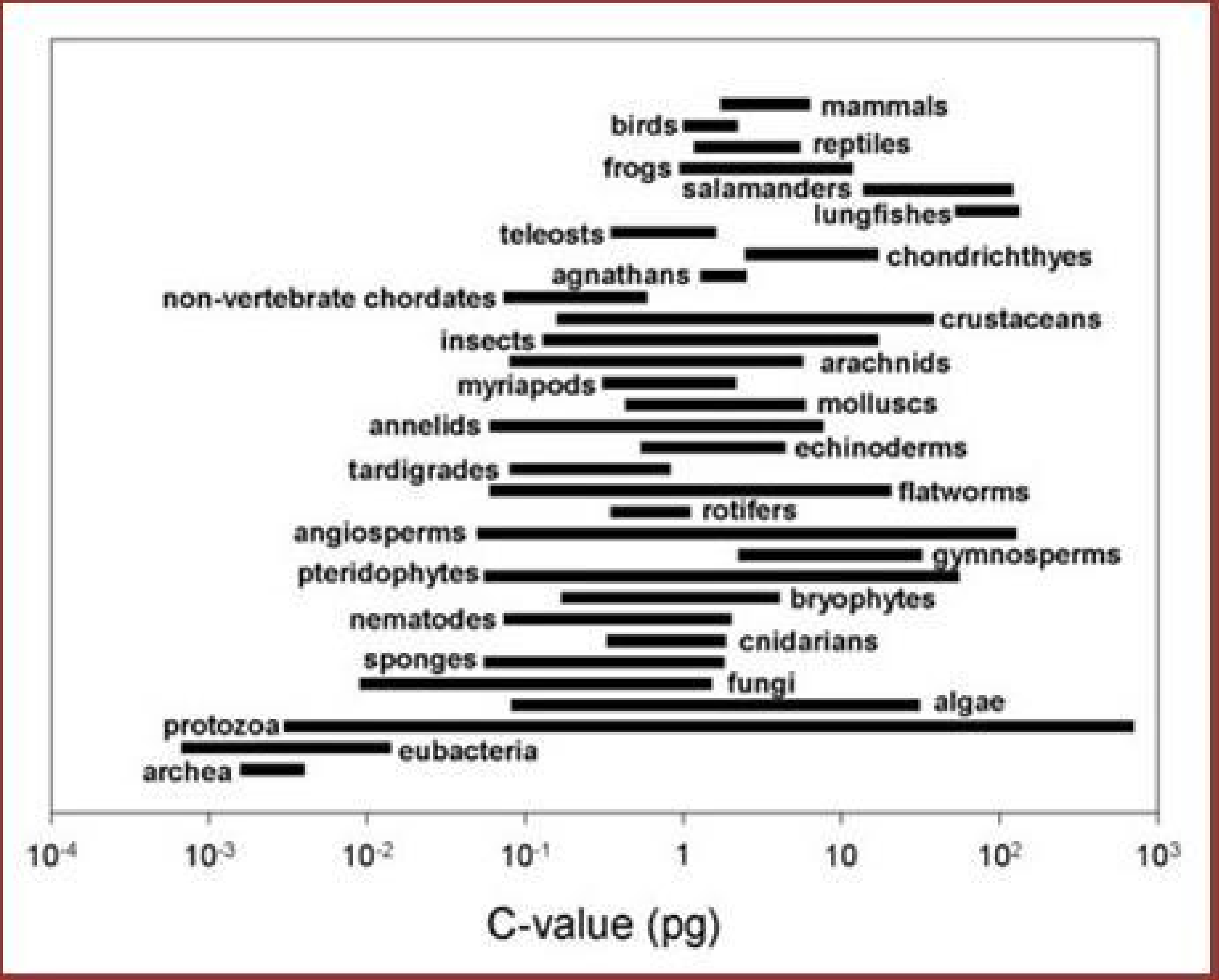


Genome size

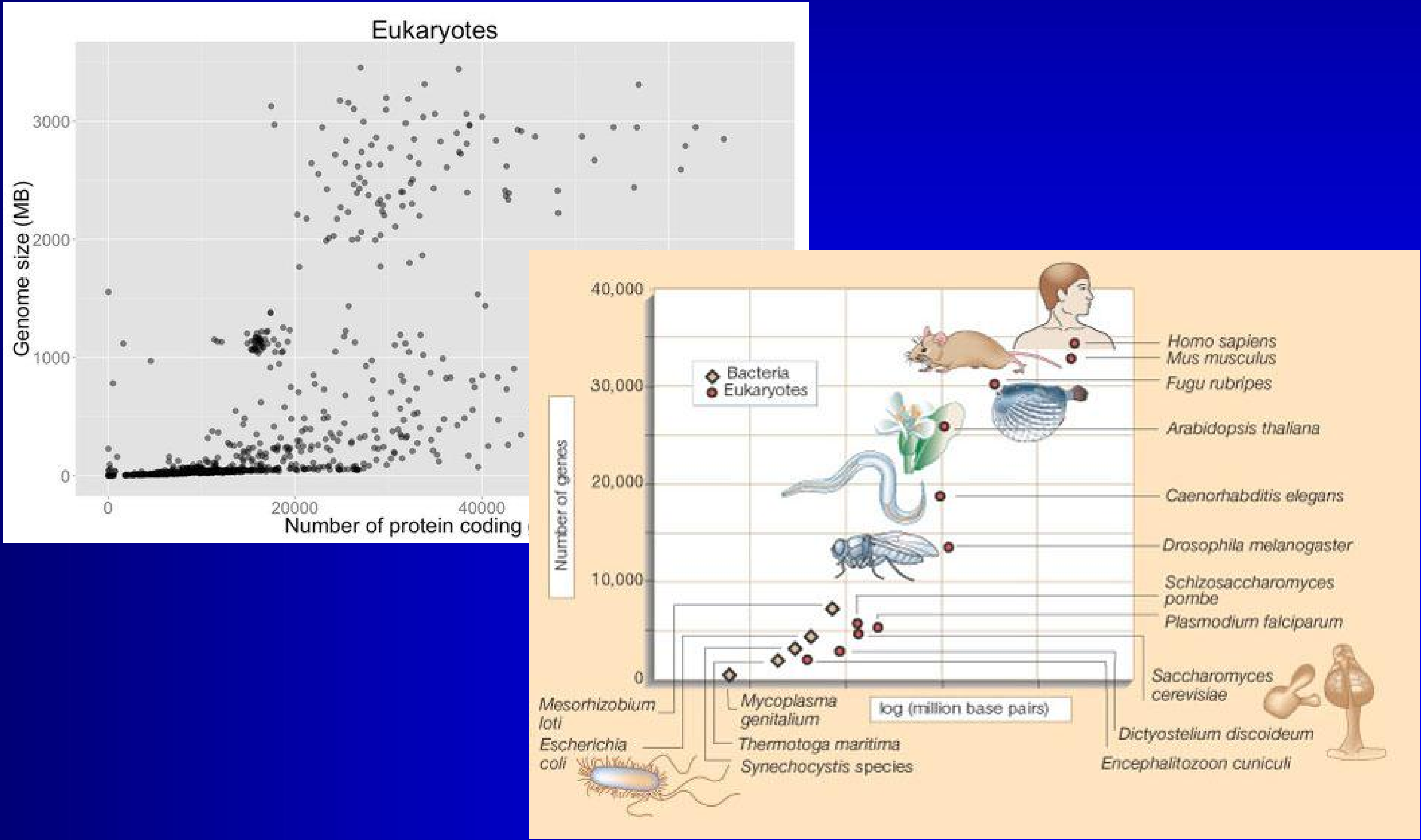
- Absolute minimum unknown – a few hundred thousand kilobases for a minimal bacterium?
- Viruses are smaller but do not count as actual organisms
- Most bacteria 1-5 megabases
- Humans ~3 gigabases, similar for other "higher" eukaryotes – some plants have extremely large genomes
- Gene count ranges from 500-1000 for simplest bacteria, up to ~5000 for the most complex
- Eukaryotes range from ~5000 genes for yeast to ~60000 for rice and Trichomonas.
- Humans have somewhere around 20-25000 genes

Genome size – C-value paradox

- Paradox: the ra constant, but is the latter!
- How is this po
- Likewise, the c vs yeast) is mu
- Small genome
- In large genom of non-coding

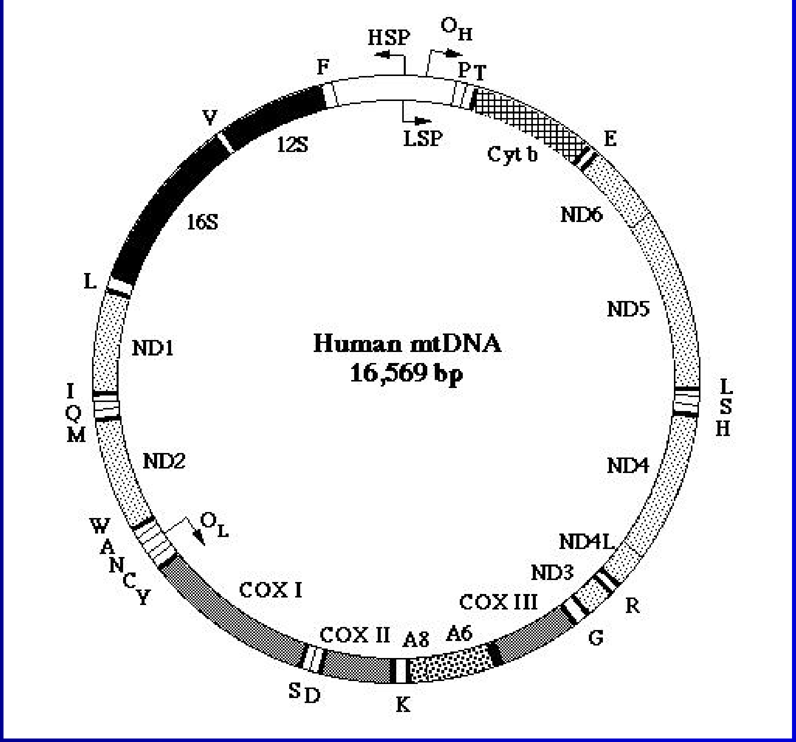


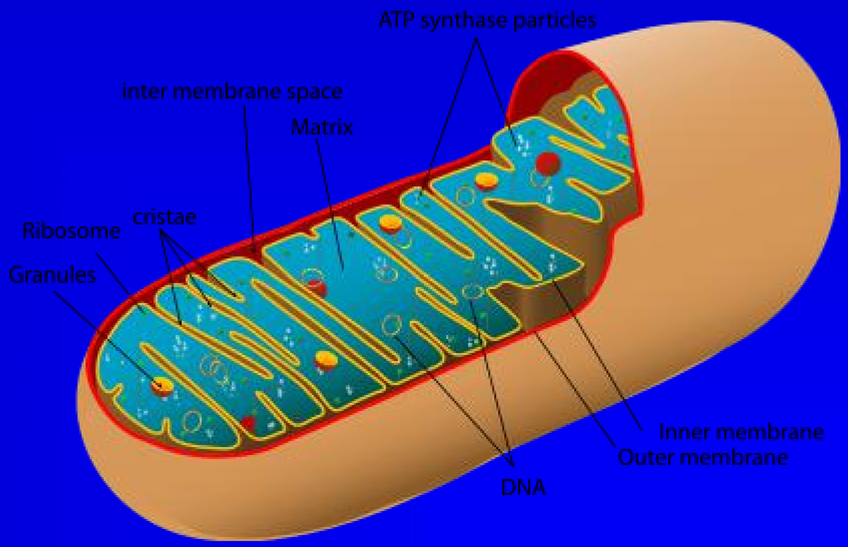
Genome size – C-value paradox



Smaller sub-genomes

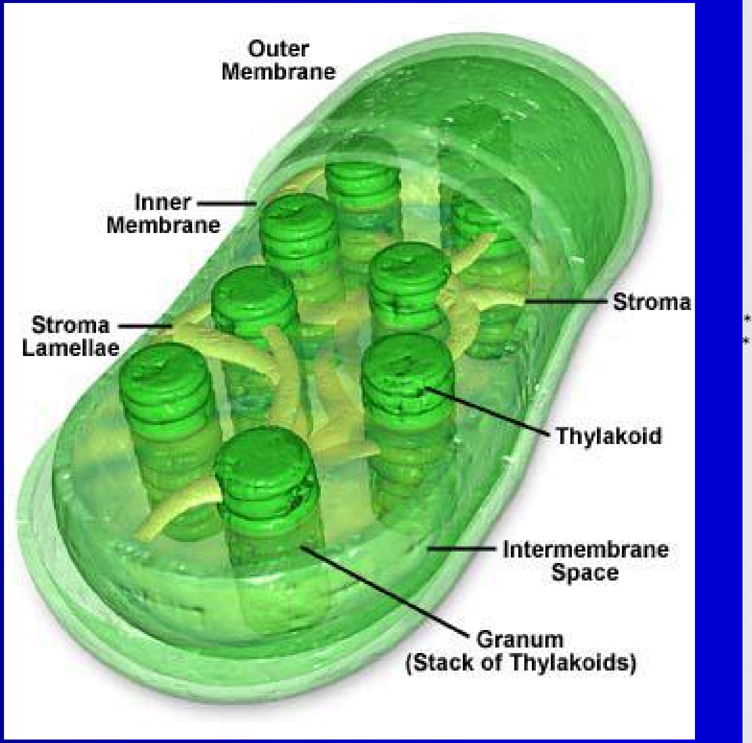
- Aside from chromosomal genome, there is mitochondrial genome in eukaryotes
- Mitochondrion appears to be degenerate symbiotic bacterium – circular genome

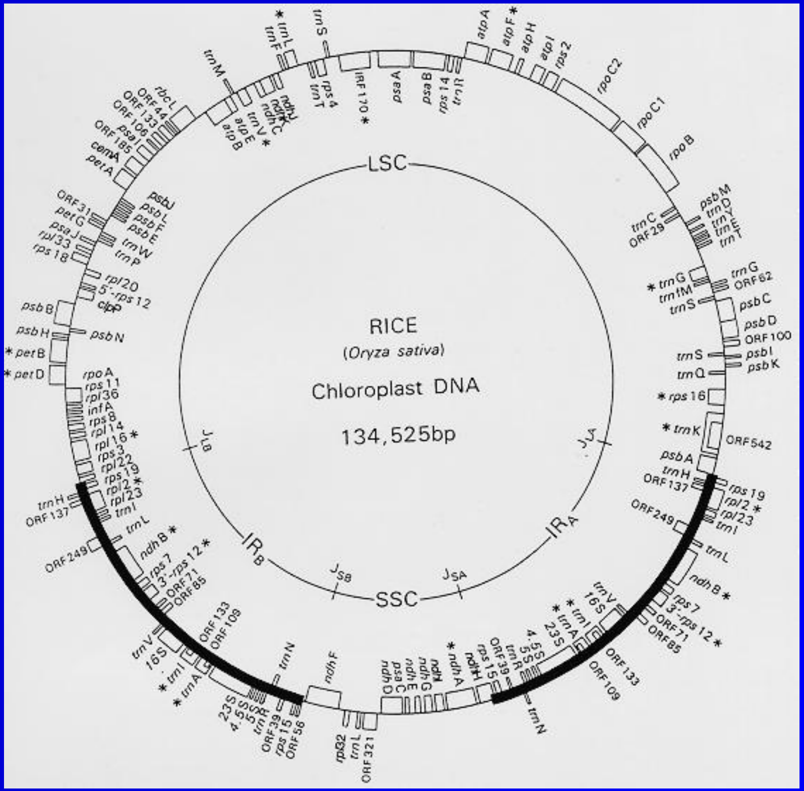




Smaller sub-genomes

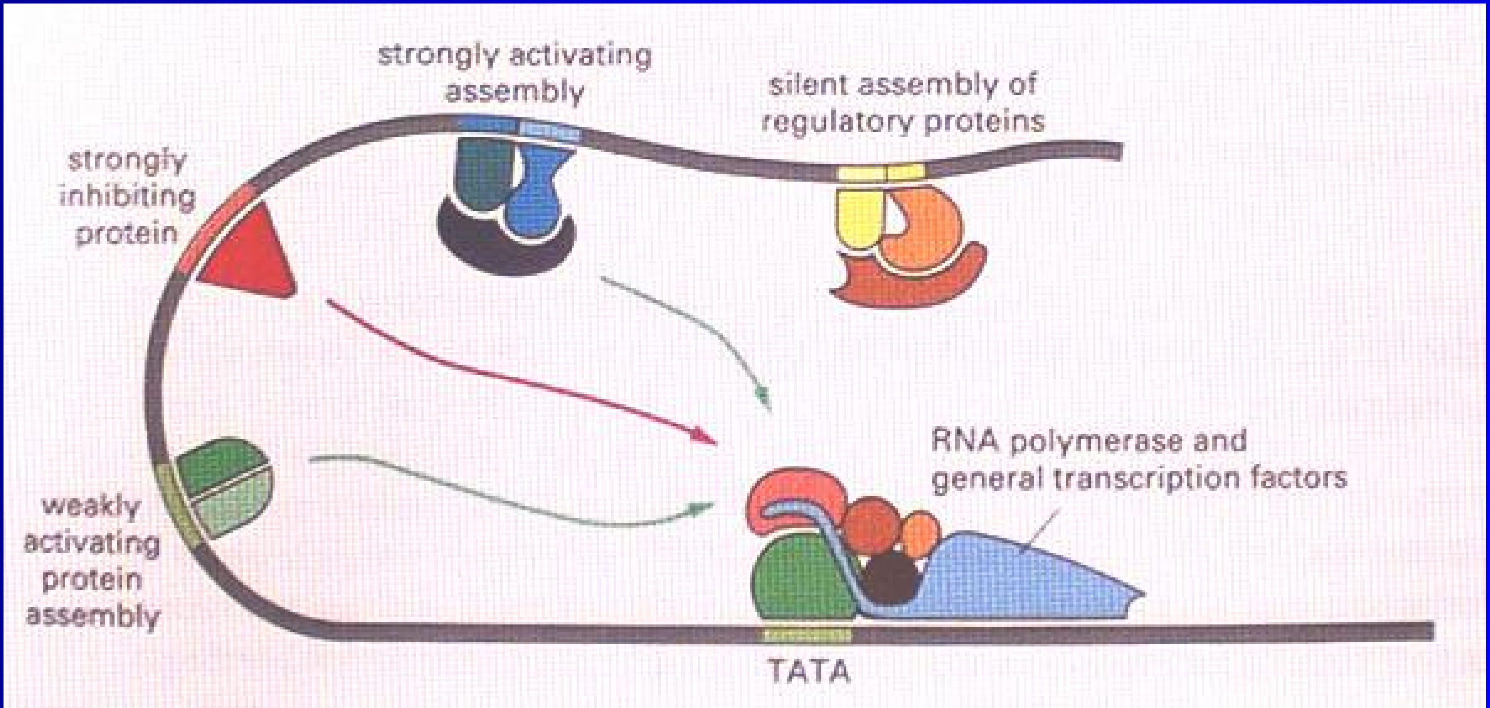
- In plants, there is (circular) chloroplast genome
- Similarly, a symbiont
- These (sub-)genomes resemble their bacterial origins
- Many genes seem to have migrated to the nuclear genome





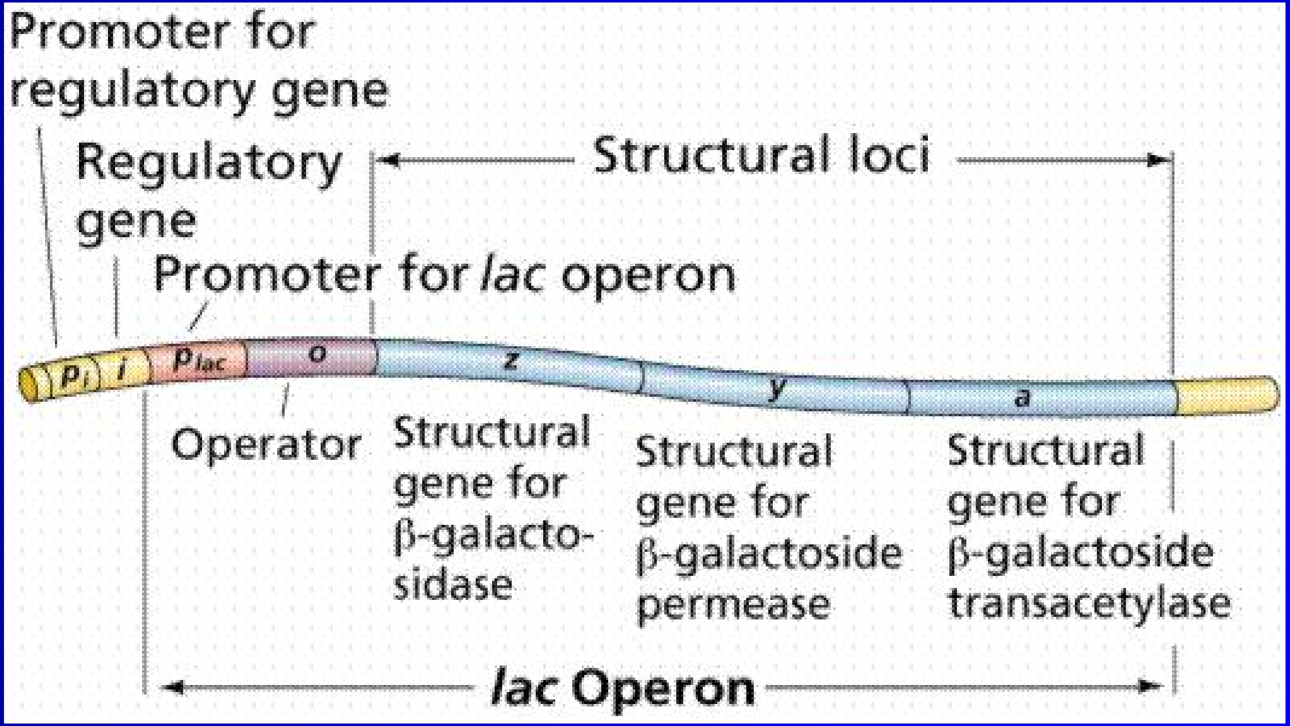
Genes and regulatory sequences

- Around gene sequences we find regulatory elements
- Promoters
- Enhancers
- Repressors


strongly activating assembly
silent assembly of regulatory proteins
strongly inhibiting protein
weakly activating protein assembly
RNA polymerase and general transcription factors
TATA

Genes and regulatory sequences

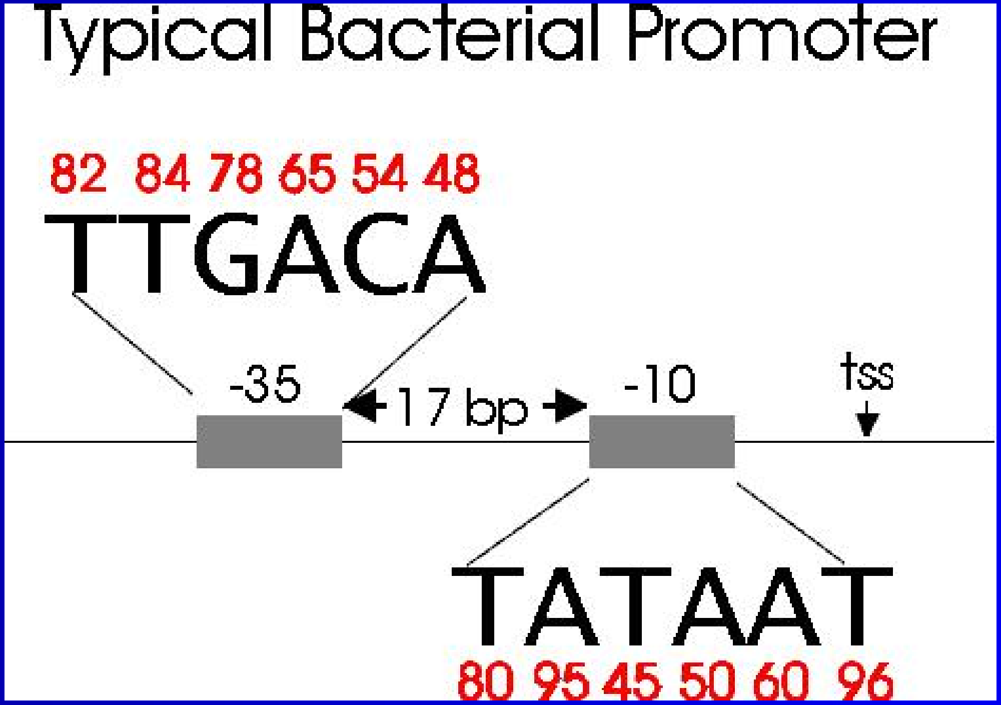
- Prokaryotic genes often form operons, several genes lying in a row and expressed (transcribed) as a result of the same signals
- These genes are co-regulated



Promoter for regulatory gene
Regulatory gene
Structural loci
Promoter for *lac* operon
P_i
i
P_{lac}
o
z
y
a
Operator
Structural gene for β-galactosidase
Structural gene for β-galactoside permease
Structural gene for β-galactoside transacetylase
lac Operon

Genes and regulatory sequences

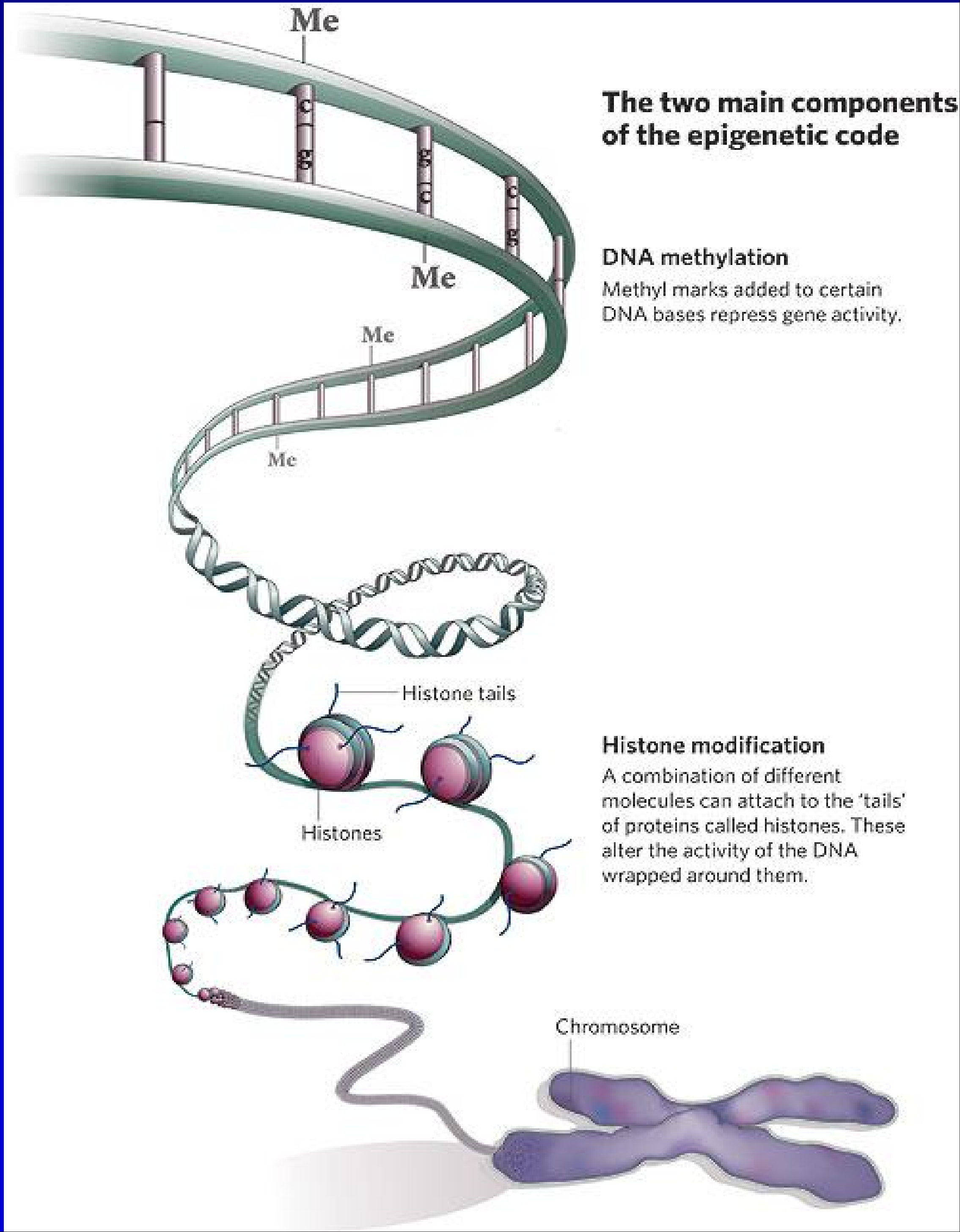
- In bacteria, often a small number of transcription factors each turn on or off large classes of genes (such as heat shock response genes)
- Eukaryotes have more regulatory elements, and they stretch over longer regions around the genes



- Epigenetics - not covered in this course, but is a growing field.

- Some epigenetic changes are genetic mutations that are passed along in families; some stem from environmental factors.

- Many diseases and conditions are linked to epigenetic changes.



The two main components of the epigenetic code

DNA methylation

Methyl marks added to certain DNA bases repress gene activity.

Histone modification

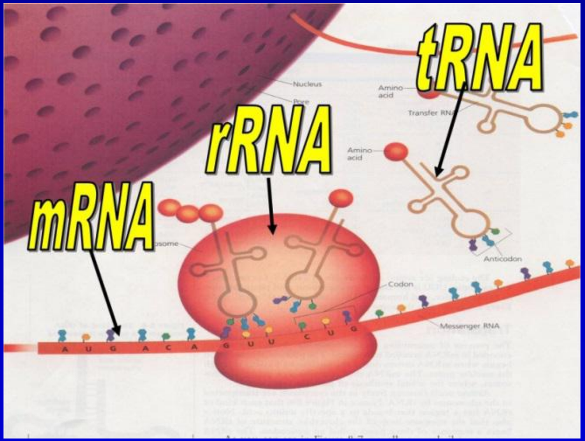
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

Pseudogenes

- Compact organisms remove unnecessary genomic components
- Less compact organisms might not, which leads to defunct copies of genes – pseudogenes – remaining in the genome although no longer expressed
- Processed pseudogenes – inserts of other genes, but lacking introns and promoters and hence being nonfunctional
- Likely the result of some reverse transcriptase and integrase acting on an arbitrary mRNA molecule

RNA genes

- RNAs involved in protein synthesis:



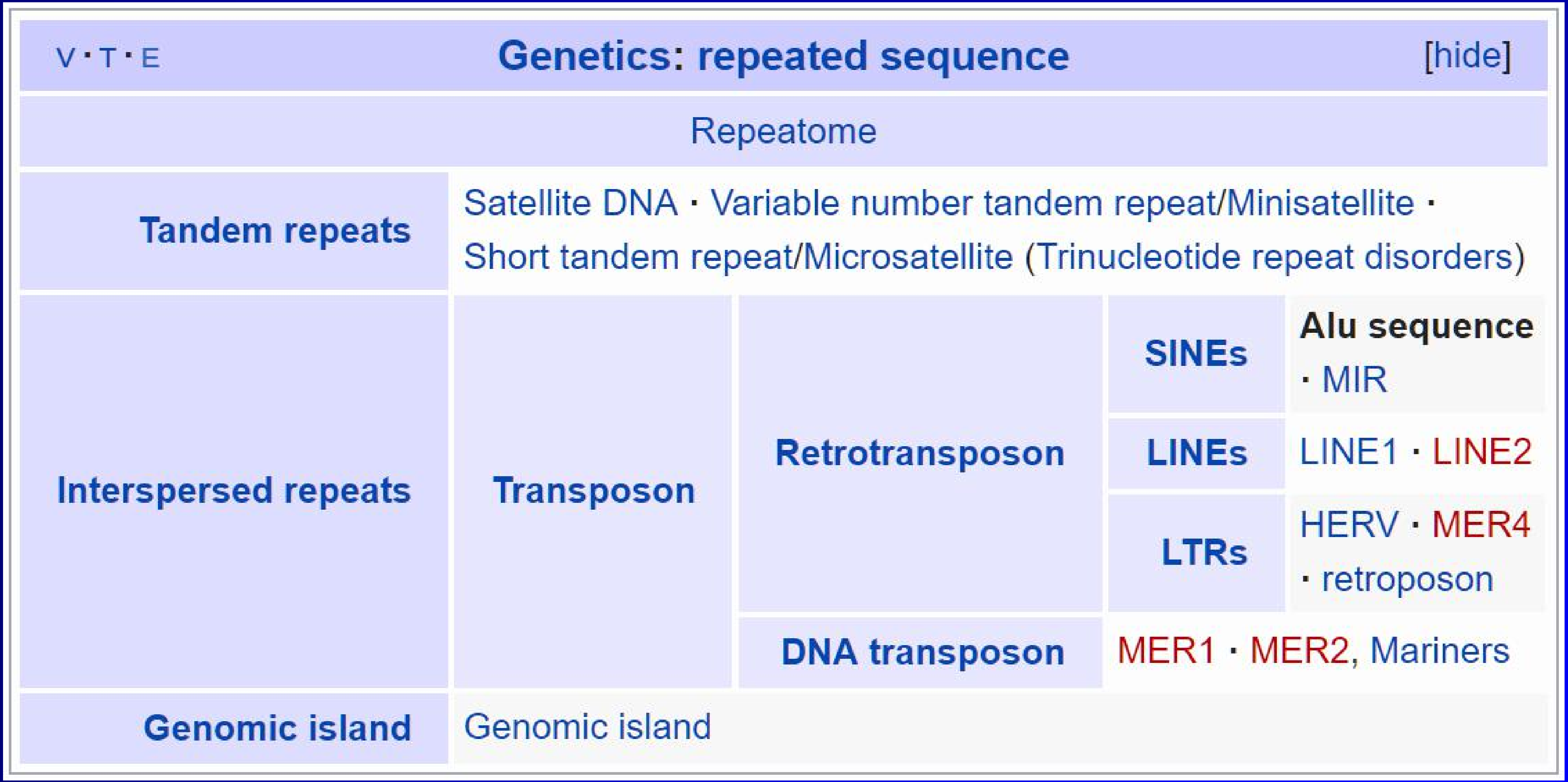
- Regulatory RNAs:
  - aRNA, asRNA, cis-NAT, crRNA, lncRNA, miRNA, piRNA, siRNA, tasiRNA, rasiRNA, 7SK

- RNAs involved in post-transcriptional modification or DNA replication:
  - gRNA, RNase MRP, RNase P, scaRNA, SL RNA, SmY, snoRNA, snRNA, TERC

Non-coding DNA

- The rest of the genome?
- Some seemingly random “spacer sequence” between elements – more in larger genomes.
- Prokaryotes are almost only coding sequences, promoters and little spacer.
- In higher eukaryotes, only a few percent are protein coding.
- >85% of the genome is transcribed.
- Repetitive elements.
- Transposable elements.

Repetitive DNA



V · T · E **Genetics: repeated sequence** [hide]

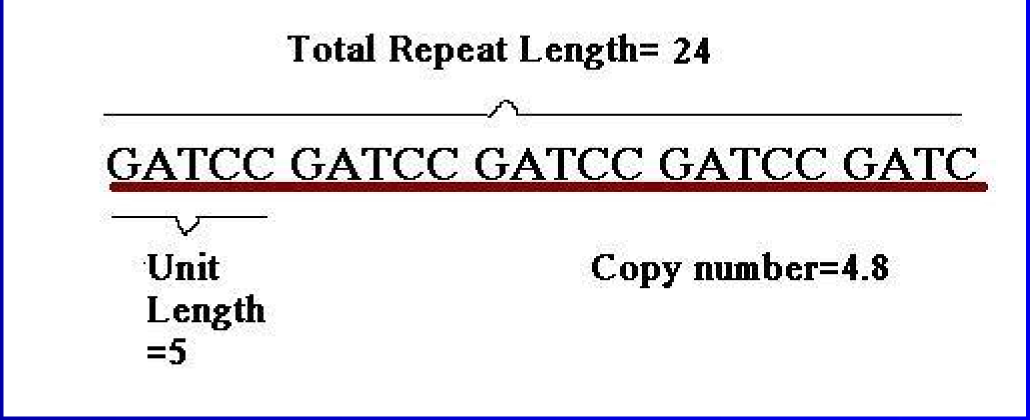
Repeatome

Tandem repeats	Satellite DNA · Variable number tandem repeat/Minisatellite · Short tandem repeat/Microsatellite (Trinucleotide repeat disorders)			
Interspersed repeats	**Transposon**	**Retrotransposon**	**SINEs**	**Alu sequence** · MIR
			LINEs	LINE1 · LINE2
			LTRs	HERV · MER4 · retroposon
		DNA transposon	MER1 · MER2, Mariners	
Genomic island	Genomic island			

from Wikipedia

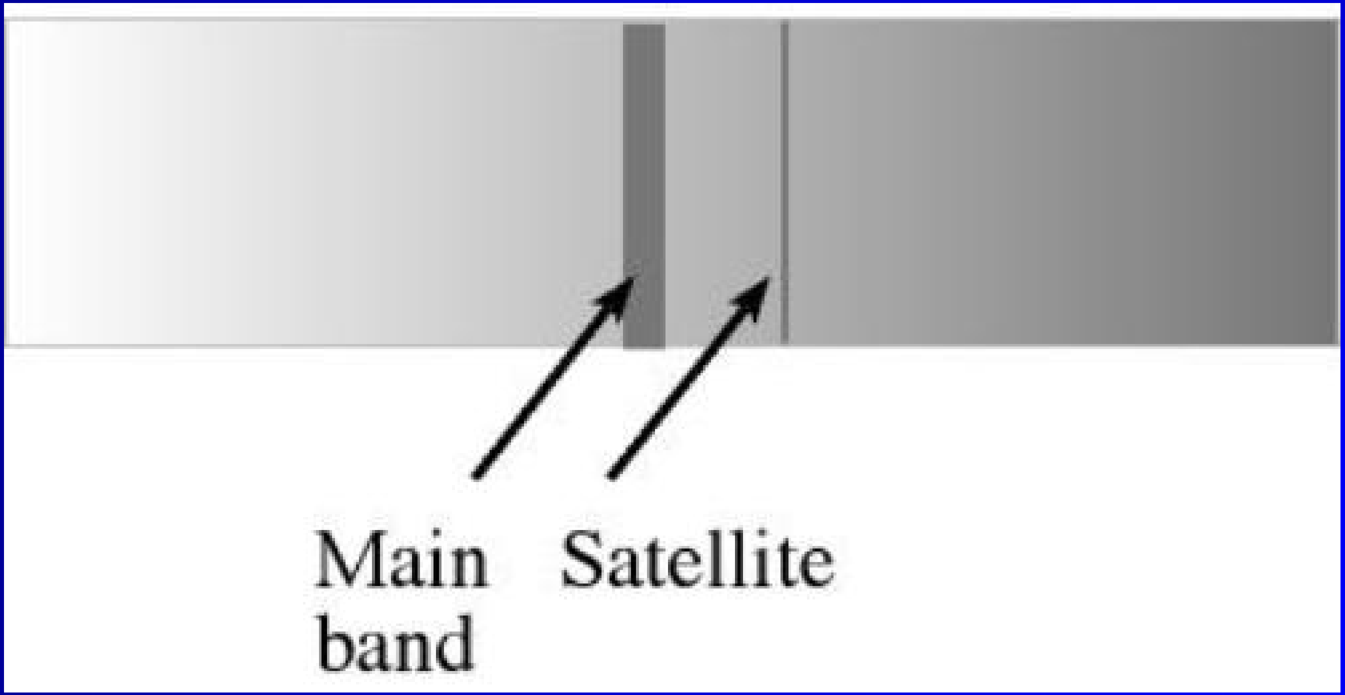
Repetitive DNA

- Repeats are short or long stretches of repeating patterns of nucleotides, in turn either short or long patterns.
- Many different kinds
- Tandem repeats – can be short (2-10) or long (10-100+), and occurs many times (a few to hundreds) in a row
- Can increase or decrease in size as the replicating polymerase "slips" when copying it
- Repeats make assembly of shotgun sequencing reads difficult!



Tandem repeats

- VNTR – variable number tandem repeat
- Forms a “satellite” band different from rest of genome when spread out on a gel, because of genomically atypical base composition

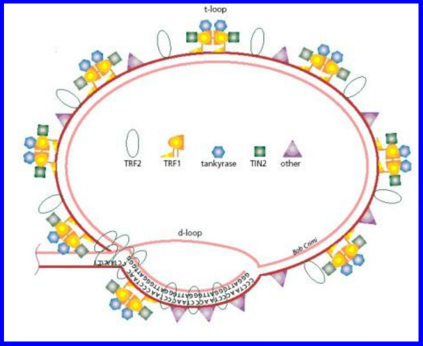


Tandem repeats

- Satellite – longest tandem sequences
- Minisatellite – longer tandem sequences
- Microsatellite – shorter tandem sequences
- Minisatellites increase chance of recombination between chromosomes
- Number of repeats is useful as a molecular marker, both for identification (paternity tests, forensics etc.) or for population studies etc.

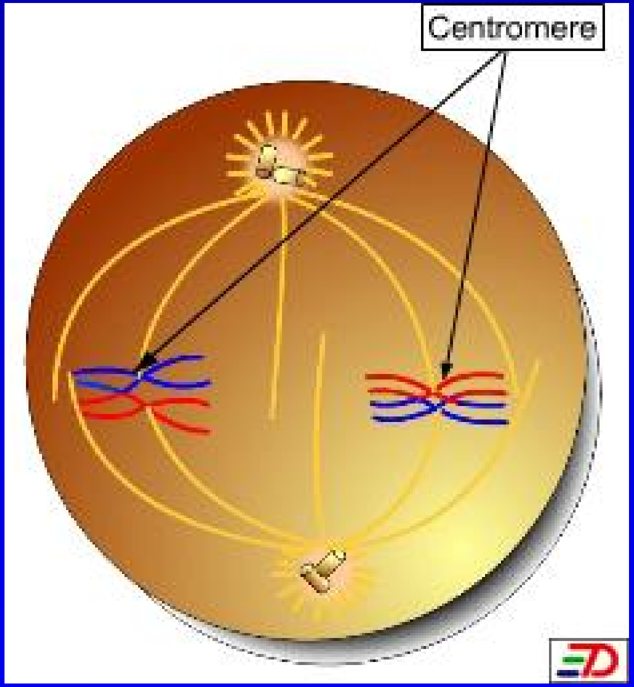
Tandem repeats – telomeric repeats

- Because of limitations of telomerases, linear chromosomes shorten with each replication
- Telomeric repeats – long sequences of repeats at the ends of the chromosomes – a type of microsatellite.
- Associated with the telomere are several types of protein, forming the T-loop at the end of the chromosome
- They are not used for anything, but when too short, the cell cannot divide and enters senescence (cellular old age) – may be aging mechanism!
- Telomeres can be lengthened by telomerase enzyme in stem cells or cancer cells.



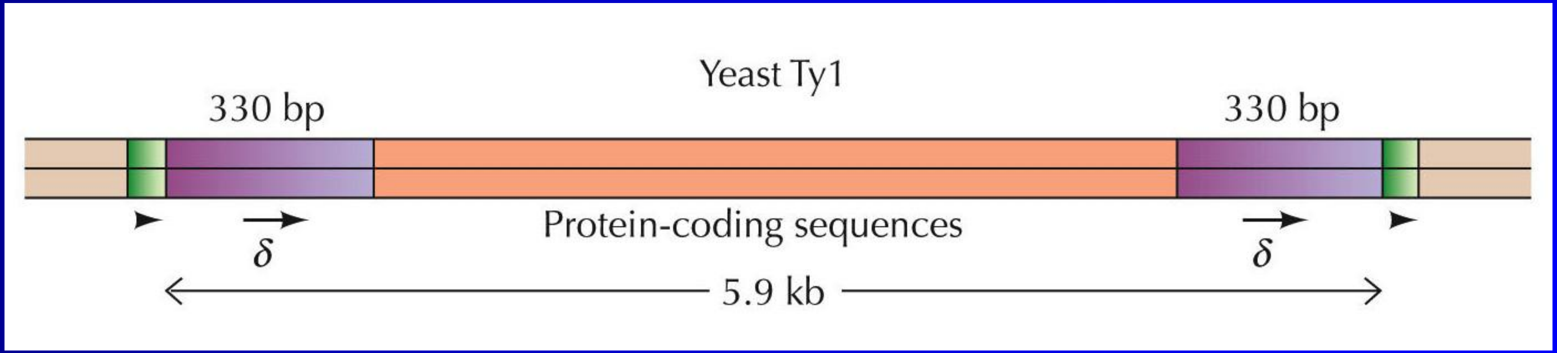
Tandem repeats – centromere & ori

- Centromeric satellite repeats near middle of (eukaryotic) chromosomes
- Used during cell division to separate the sister chromatids into different daughter cells
- Failure to do so yields chromosome number changes
- Another type are the replication origin sequences, which are starting points for replication forks
- Circular genomes have just one, linear chromosomes may have many origins



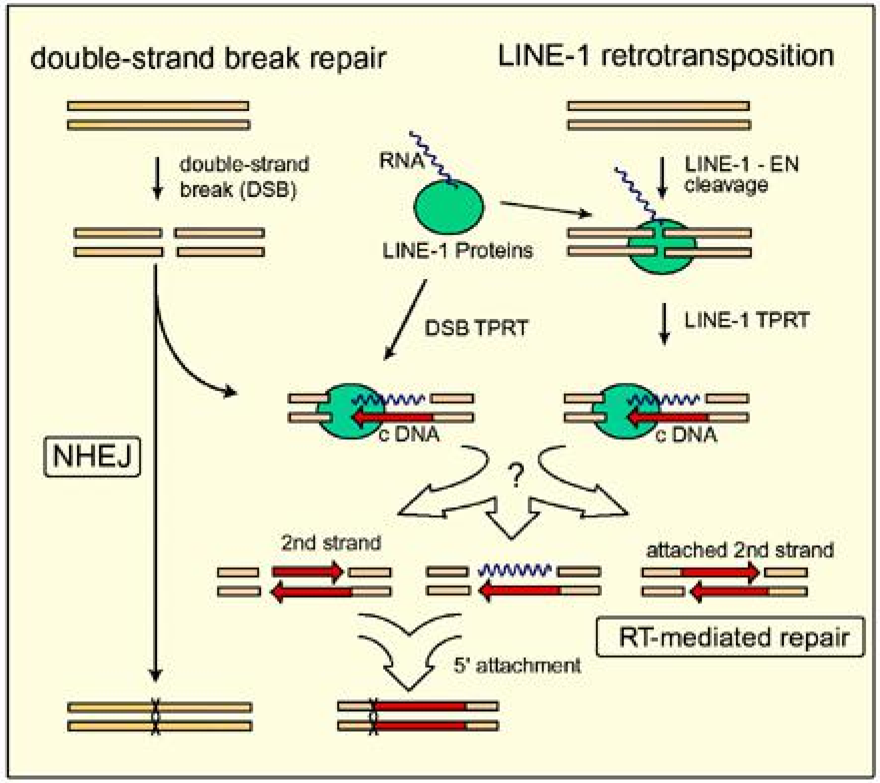
Interspersed repeats

- Interspersed repeats
- Repeated motifs found interspersed throughout the genome, rather than in tandem
- Mobile or transposable genetic elements
- LINE – Long Interspersed Nuclear Element
- SINE – Short Interspersed Nuclear Element
- LTR – Long Terminal Repeats



Interspersed repeats - LINEs

- Long Interspersed Nuclear Element (LINE)
- About 5000 bases
- About 20% of the human genome are LINEs!
- Encodes proteins for copying itself and inserting again elsewhere in the genome – similar to intracellular retroviruses.



Interspersed repeats - LINEs

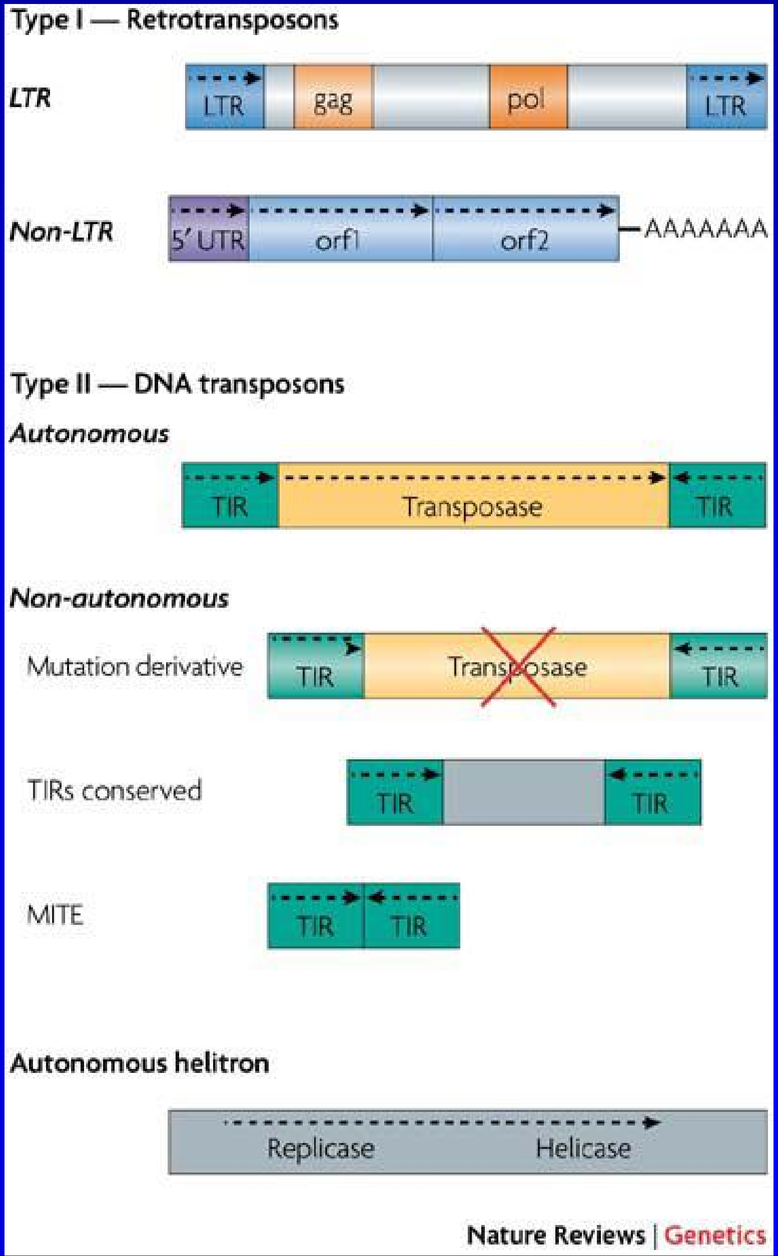
- Other transposable elements uses the LINE copying machinery for the same purpose
- These changes increase genome size (other factors decrease it) and may prevent recently duplicated genes from being removed by gene conversion
- Different types of LINEs present in different organisms
- L1 family by far dominant in mammals

Short interspersed repeats - SINEs

- Around 500 nucleotides
- About 13% of the human genome
- Common primate SINE: Alu repeat
- Rely on the LINE machinery for multiplying

Interspersed repeats - LTRs

- Long Terminal Repeats
- Several retroviruses (like HIV) have LTR regions flanking their genomes
- LTR-based retrotransposons share this mechanism
- Integrase enzymes exist that are specific for the LTRs, so anything between LTRs will be copied into different places across the genome
- Do retrotransposons come from retrovirusses or vice versa?



Nature Reviews | Genetics

Interspersed repeats – DNA transposons

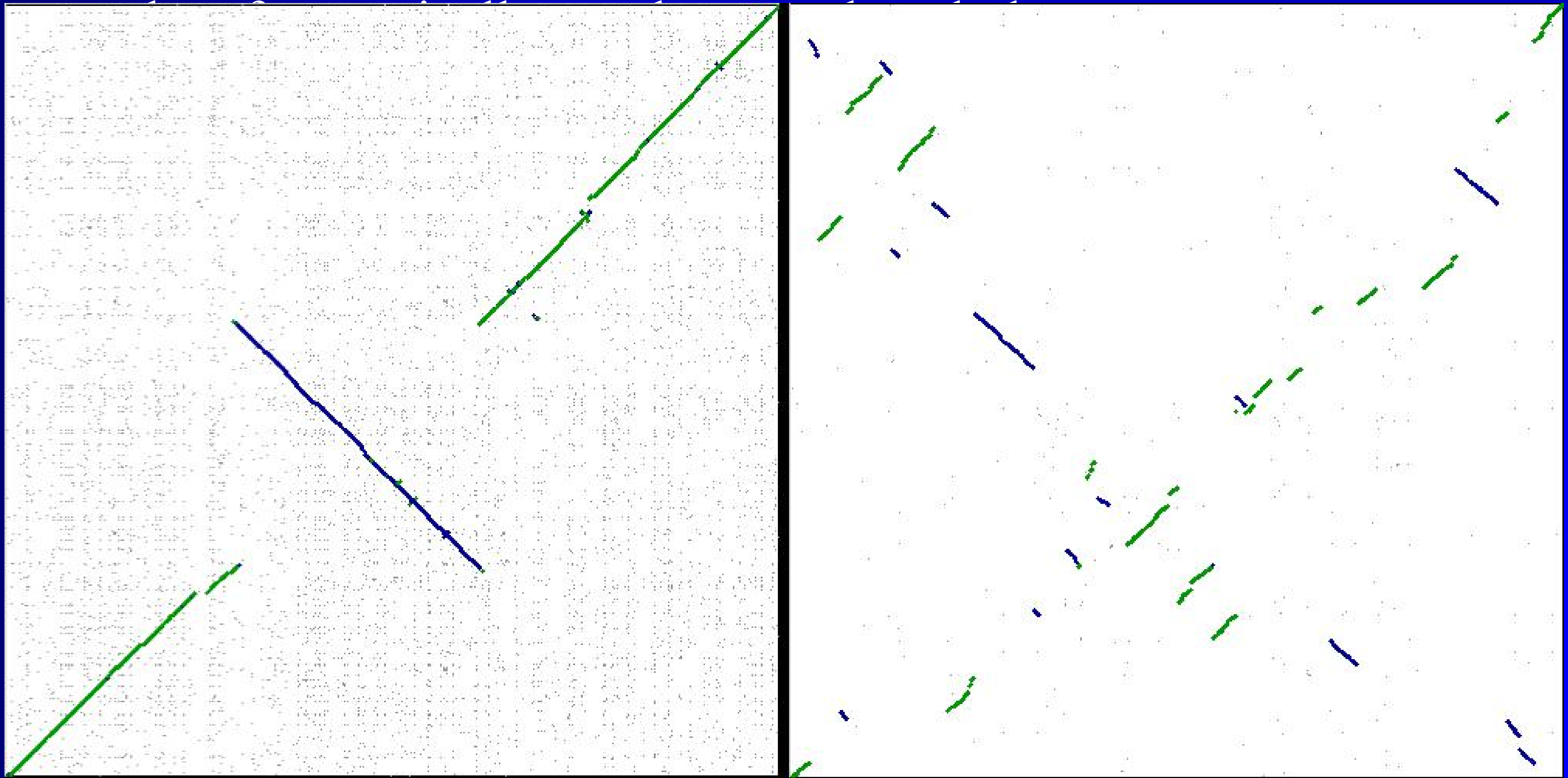
- These are not copied as such, but cut out and reinserted – i.e. they typically do not multiply
- All of the elements mentioned so far are mutation mechanisms – they can be inserted into a gene and inactivate it
- Many transposable elements eventually mutate and lose functionality
- All of these elements may also be used as genetic markers

Summary of different genomic features

- Prokaryote circular genomes – mainly genes and promoter sequences
- Eukaryote organelle genomes – similar to bacteria
- Eukaryote nuclear genomes – centromeric and telomeric repeat regions, tandem repeat regions, interspersed self-replicating elements, gene regulatory elements (promoter, enhancer, suppressor), RNAs and protein-coding genes, intron-exon-based, against a background of spacer sequence

Synteny

- How things are ordered in a genome (chromosomal organisation, gene order etc.) change much faster than the sequence of encoded components - comparative genomics is

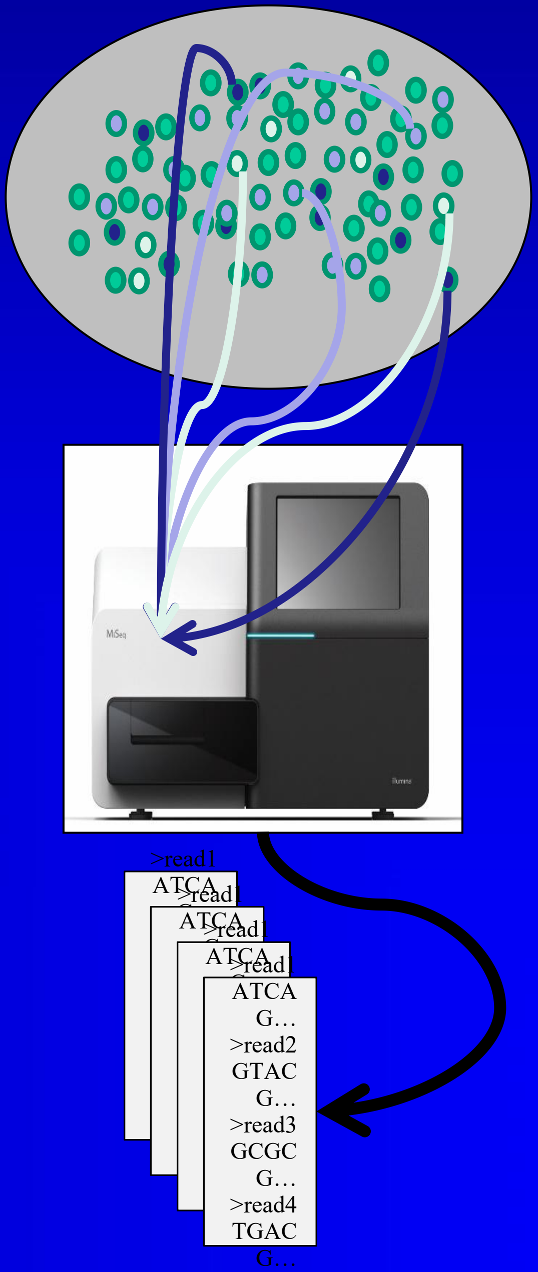


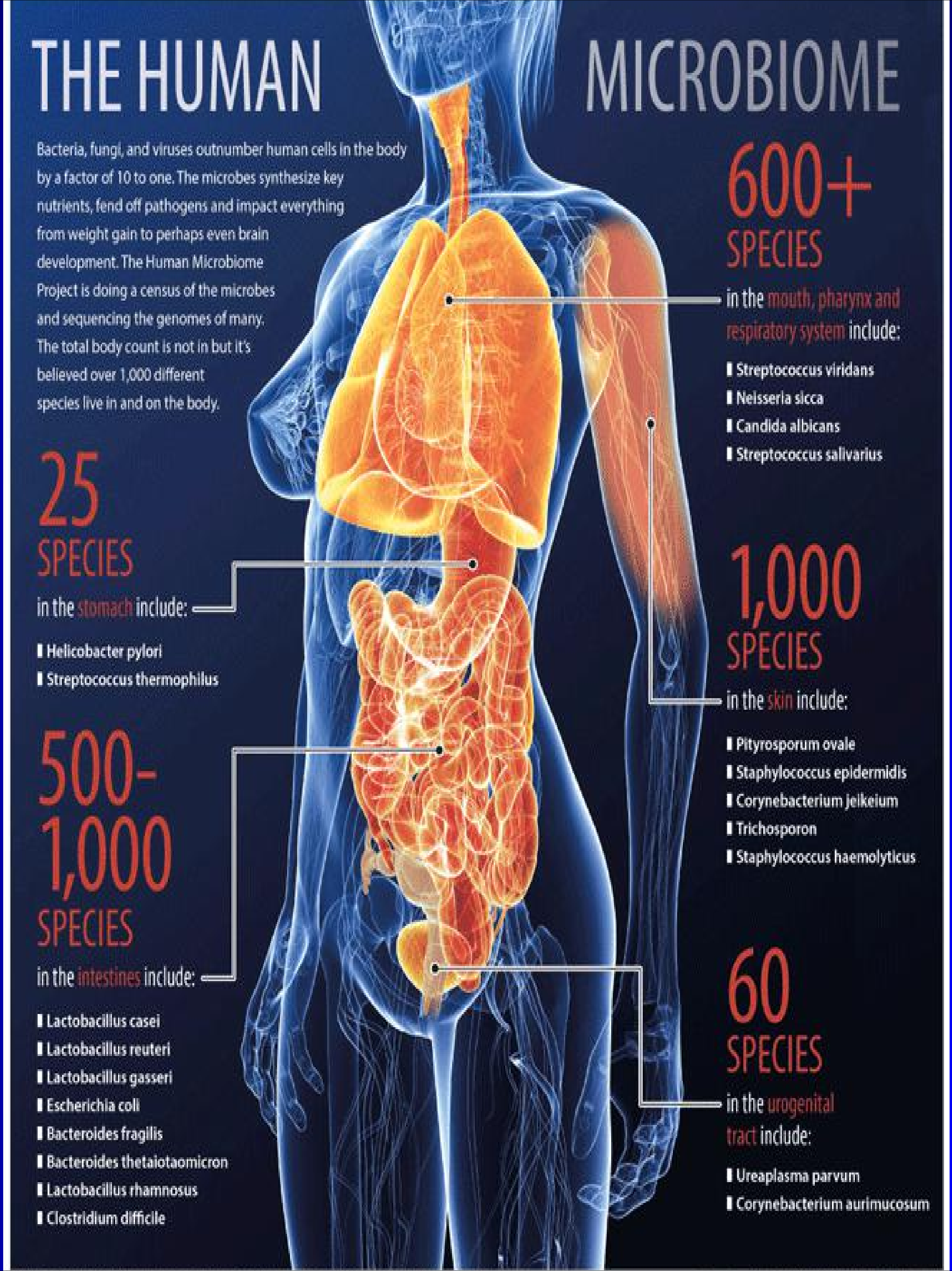
Metagenomics

- Samples, e.g. bacterial communities of interest

- Shotgun sequencing of random pieces of all bacterial genomes in community

- DNA sequence fragments





THE HUMAN MICROBIOME

Bacteria, fungi, and viruses outnumber human cells in the body by a factor of 10 to one. The microbes synthesize key nutrients, fend off pathogens and impact everything from weight gain to perhaps even brain development. The Human Microbiome Project is doing a census of the microbes and sequencing the genomes of many. The total body count is not in but it's believed over 1,000 different species live in and on the body.

600+ SPECIES

in the mouth, pharynx and respiratory system include:

- Streptococcus viridans
- Neisseria sicca
- Candida albicans
- Streptococcus salivarius

25 SPECIES

in the stomach include:

- Helicobacter pylori
- Streptococcus thermophilus

1,000 SPECIES

in the skin include:

- Pityrosporum ovale
- Staphylococcus epidermidis
- Corynebacterium jeikeium
- Trichosporon
- Staphylococcus haemolyticus

500-1,000 SPECIES

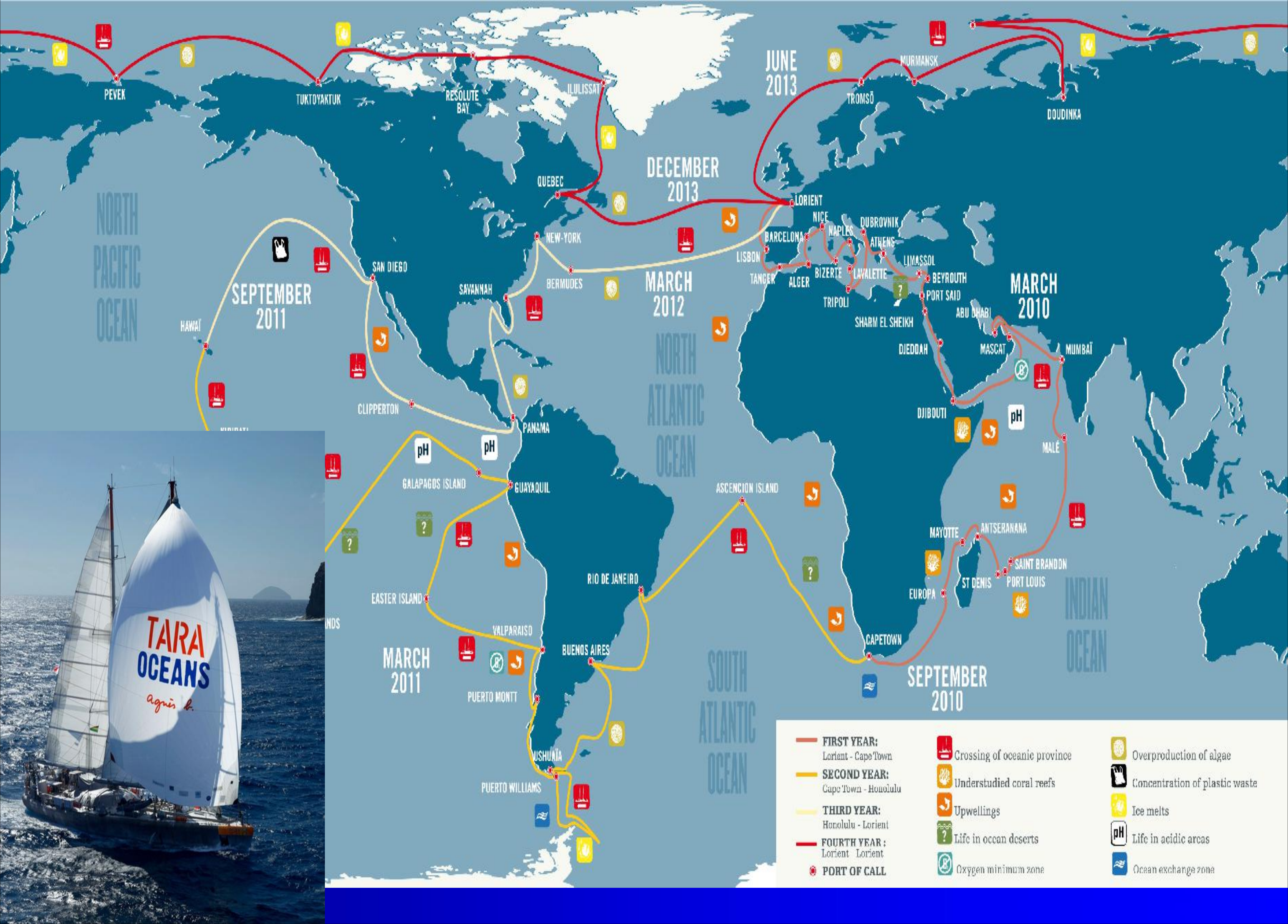
in the intestines include:

- Lactobacillus casei
- Lactobacillus reuteri
- Lactobacillus gasseri
- Escherichia coli
- Bacteroides fragilis
- Bacteroides thetaiotaomicron
- Lactobacillus rhamnosus
- Clostridium difficile

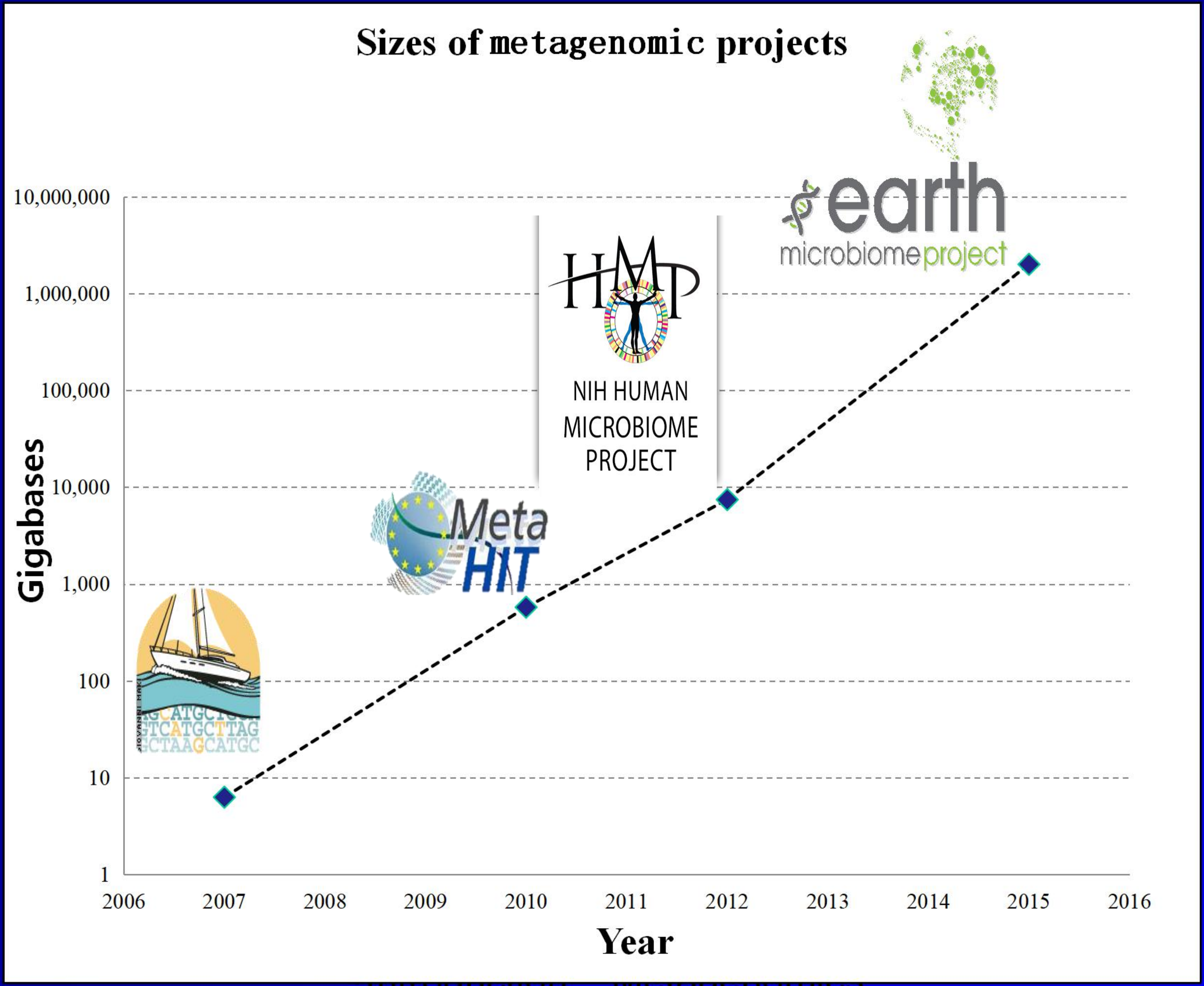
60 SPECIES

in the urogenital tract include:

- Ureaplasma parvum
- Corynebacterium aurimucosum


PEVEK
TUKTOYAKTUK
RESOLUTE BAY
ILULISSAT
JUNE 2013
MURMANSK
TROMSÖ
DOUDINKA
NORTH PACIFIC OCEAN
QUEBEC
DECEMBER 2013
LORIENT
NICE
DUBROVNIK
BARCELONA
NAPLES
ATHENS
LISBON
TANGER
ALGER
BIZERTE
LAVALETTE
LIMASSOL
BEYROUTH
PORT SAID
TRIPOLI
SHARM EL SHEIKH
ABU DHABI
MARCH 2010
SAN DIEGO
NEW-YORK
SAVANNAH
BERMUDES
MARCH 2012
SEPTEMBER 2011
HAWAÏ
DJEDDAH
MASCAT
MUMBAÏ
NORTH ATLANTIC OCEAN
CLIPPERTON
PANAMA
DJIBOUTI
pH
MALÉ
KIRIBATI
pH
pH
GALAPAGOS ISLAND
GUAYAQUIL
ASCENCION ISLAND
?
?
?
MAYOTTE
ANTSERANANA
RIO DE JANEIRO
SAINT BRANDON
ST DENIS
PORT LOUIS
EUROPA
INDIAN OCEAN
NDS
EASTER ISLAND
VALPARAISO
BUENOS AIRES
CAPETOWN
MARCH 2011
PUERTO MONTT
SOUTH ATLANTIC OCEAN
SEPTEMBER 2010
USHUAÏA
PUERTO WILLIAMS
FIRST YEAR:
Lorient - Cape Town
SECOND YEAR:
Cape Town - Honolulu
THIRD YEAR:
Honolulu - Lorient
FOURTH YEAR :
Lorient Lorient
PORT OF CALL
Crossing of oceanic province
Understudied coral reefs
Upwellings
Life in ocean deserts
Oxygen minimum zone
Overproduction of algae
Concentration of plastic waste
Ice melts
Life in acidic areas
Ocean exchange zone
TARA
OCEANS
agnès b.

Metagenomic data revolution



GOS: http://www.jcvi.org/cms/research/projects/gos/overview/, **Meta-HIT**: http://www.metahit.eu/, **HMP**: http://www.hmpdacc.org/, **EMP**: http://www.earthmicrobiome.org/. Accessed 2013-10-14.

- Wikipedia definition: ”*data* sets so *large* or complex that traditional *data* processing applications are inadequate.”

- Grand challenge: Big Data to Knowledge.

- A single (metagenomic) sequencing project can produce many terabytes, which is difficult to analyse.

- All vs all comparison: 10^{15} x 10^{15} = 10^{30} (Nonillion, e.g. estimated number of bacterial cells on Earth)