Genomics & Bioinformatics

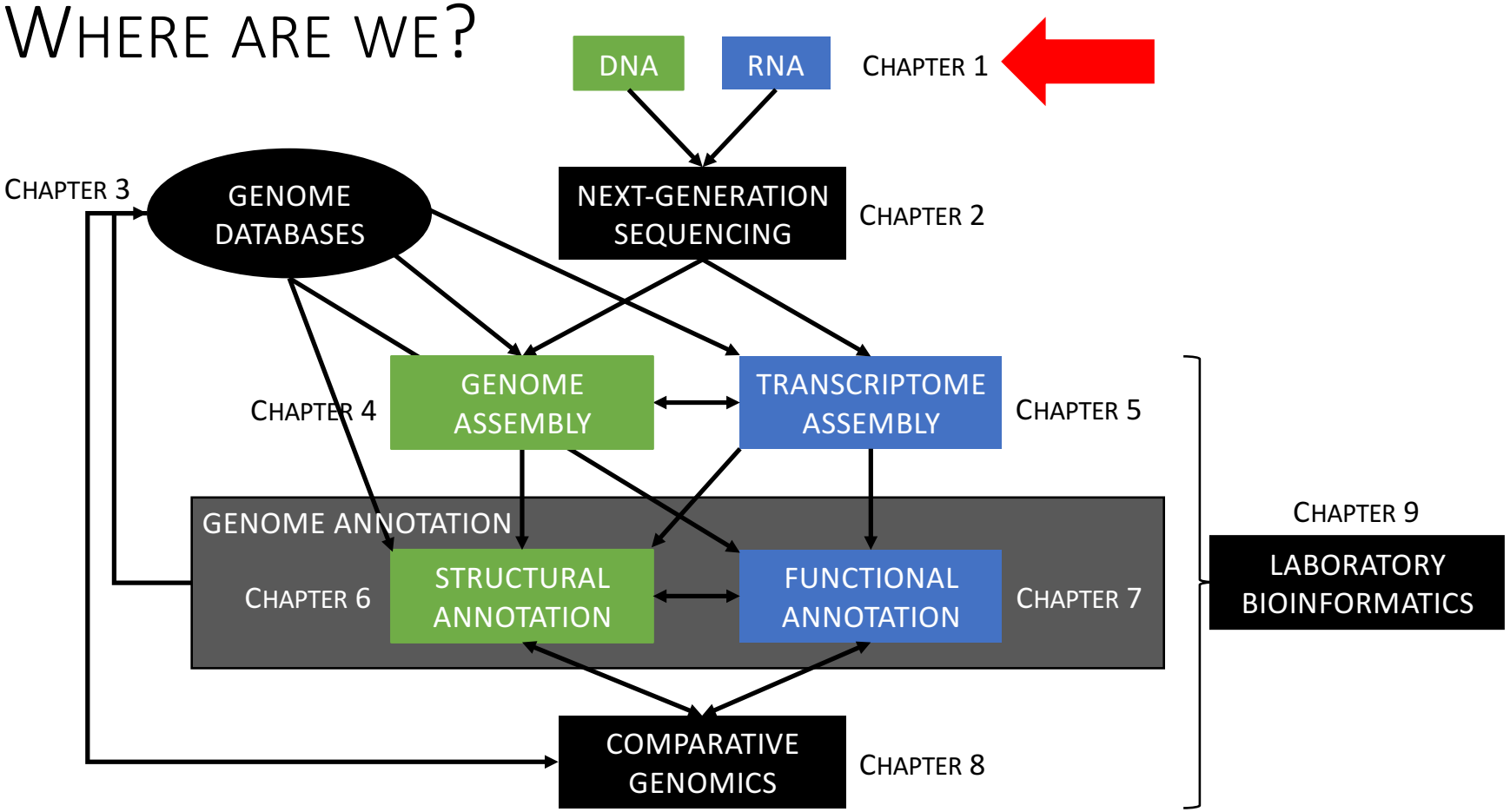
Chapter 1 – The "What is?" lecture

BIOL 497, 597

Boise State University

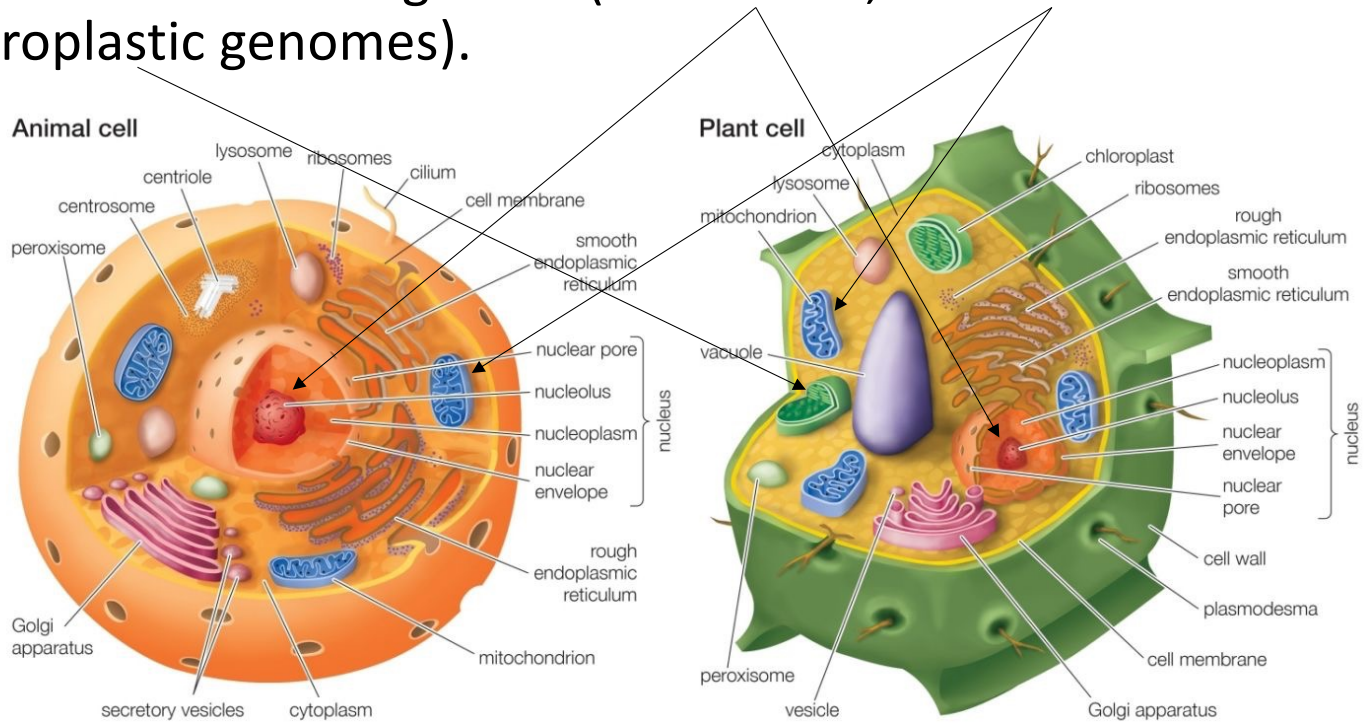
Spring

WHERE ARE WE?


DNA
RNA
CHAPTER 1
CHAPTER 3
GENOME DATABASES
NEXT-GENERATION SEQUENCING
CHAPTER 2
CHAPTER 4
GENOME ASSEMBLY
TRANSCRIPTOME ASSEMBLY
CHAPTER 5
GENOME ANNOTATION
CHAPTER 6
STRUCTURAL ANNOTATION
FUNCTIONAL ANNOTATION
CHAPTER 7
CHAPTER 9
LABORATORY BIOINFORMATICS
COMPARATIVE GENOMICS
CHAPTER 8

What is Genomics?

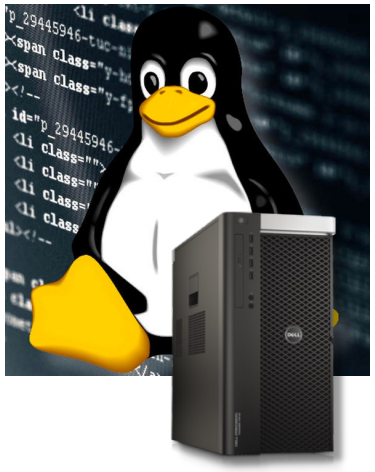
- **Genomics:** The study of genomes, the complete set of genetic material within an organism (i.e. nuclear, mitochondrial and chloroplastic genomes).

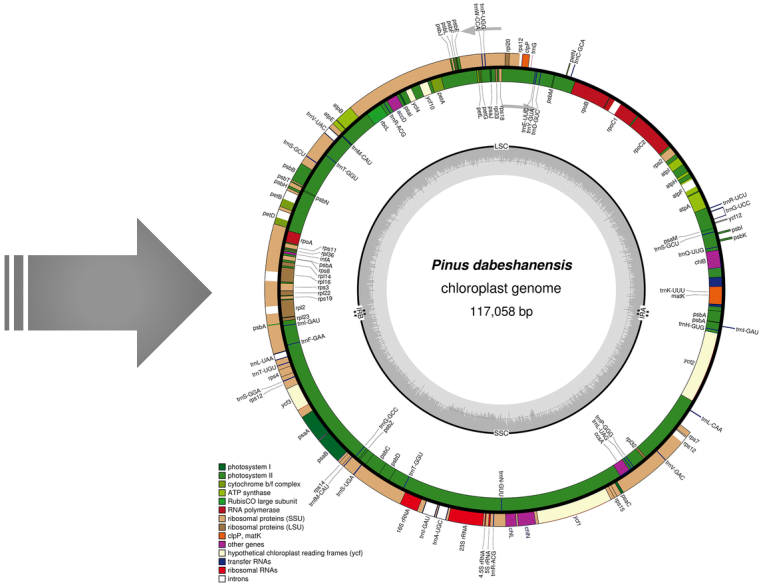


WHAT IS GENOMICS?

- **Genomics** uses an approach combining **next-generation sequencing** (hereafter NGS; Chapter 2 and Mini report 1) and **bioinformatics** to **assemble and annotate entire genomes**.

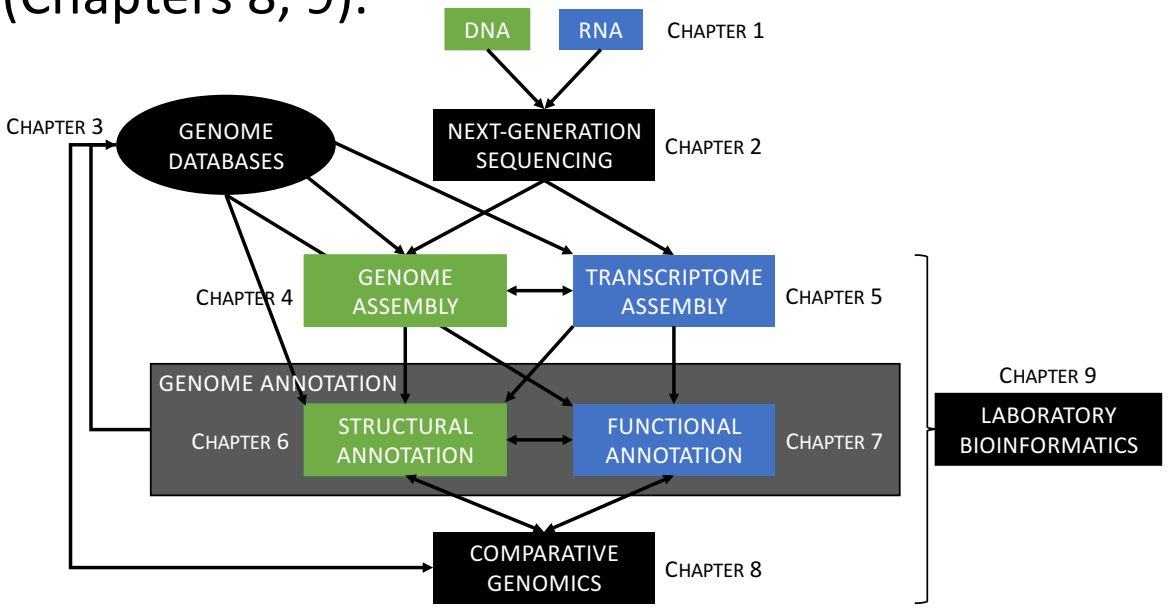
+





WHAT IS GENOMICS?

- Ultimately genomics aims at **inferring genomes' structures and functions** (Lab. on mining the sagebrush genome; Chapter 9).
- Such objectives are usually achieved by applying a comparative approach (Chapters 8, 9).



STRUCTURE AND FUNCTION OF CHLOROPLAST GENOME

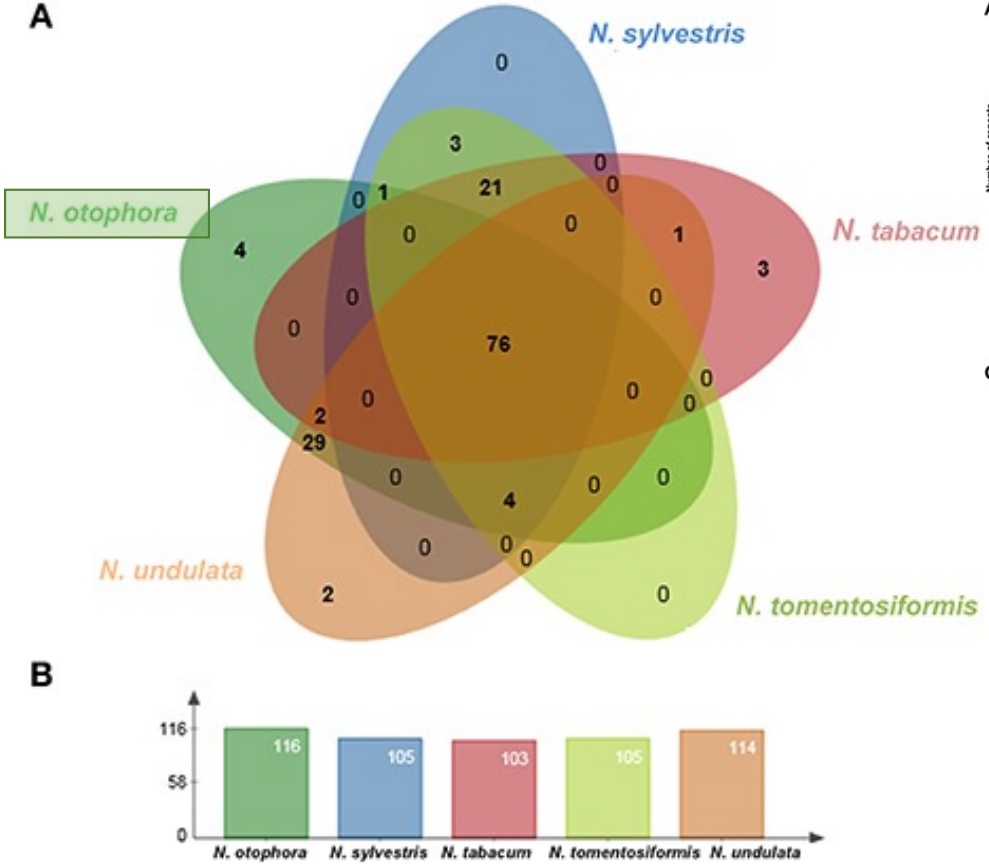
Nicotiana otophora, a wild parent species of tabaco

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl2, 14, 16, 20, 22, 23, 32, 33, 36*
	Small subunit of ribosomal proteins	*rps2, 3, 7, 8, 11, 12, 14, 15, 16, 18, 19*
	DNA dependent RNA polymerase	*rpoA, B, C1, C2*
	rRNA genes	*RNA*
	tRNA genes	*trnA-UGC, C-GCA, D-GUC, E-UUC, F-GAA, fM-CAU, G-UCC, H-GUG, I-CAU, L-CAA, M-CAU, N-GUU, P-GGG, P-UGG, Q-UUG, R-ACG, R-UCU, S-GCU, S-GGA, S-UGA, T-GGU, T-UGU, V-GAC, V-UAC, W-CCA, Y-GUA*
Photosynthesis	Photosystem I	*psaA, B, C, I, J*
	Photosystem II	*psbA, B, C, D, E, F, H, I, J, K*
	NadH oxidoreductase	*ndhA, B, C, D, E, F*
	Cytochrome b6/f complex	*petA, B, D, G, L, N*
	ATP synthase	*atpA, B, E, F, H, I*
	Rubisco	*rbcL, rbcLr*
Other genes	Translational initiation factor	*infA*
	Maturase	*matK*
	Protease	*clpP*
	Envelop membrane protein	*cemA*
	Subunit Acetyl-CoA-Carboxylate	*accD*
	c-type cytochrome synthesis gene	*ccsA*
Unknown	Conserved Open reading frames	*ycf1, 2, 3, 4, 15, 68*

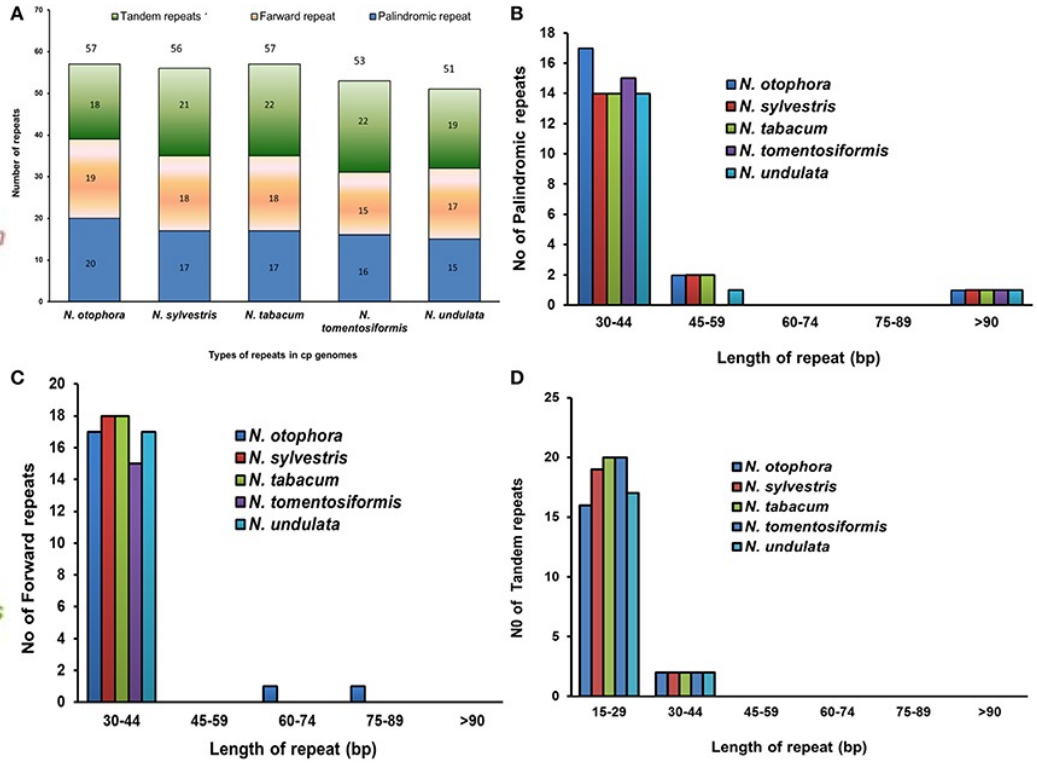
Asaf *et al.* (2016) *Front. Plant Sci.*

COMPARATIVE GENOMICS

Coding genes



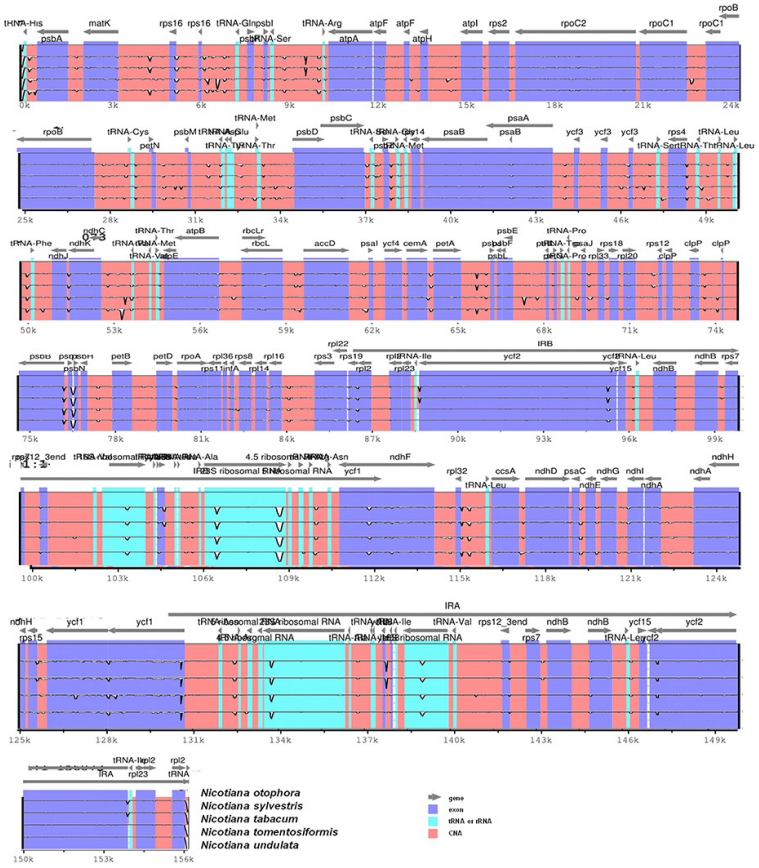
Repeat sequences



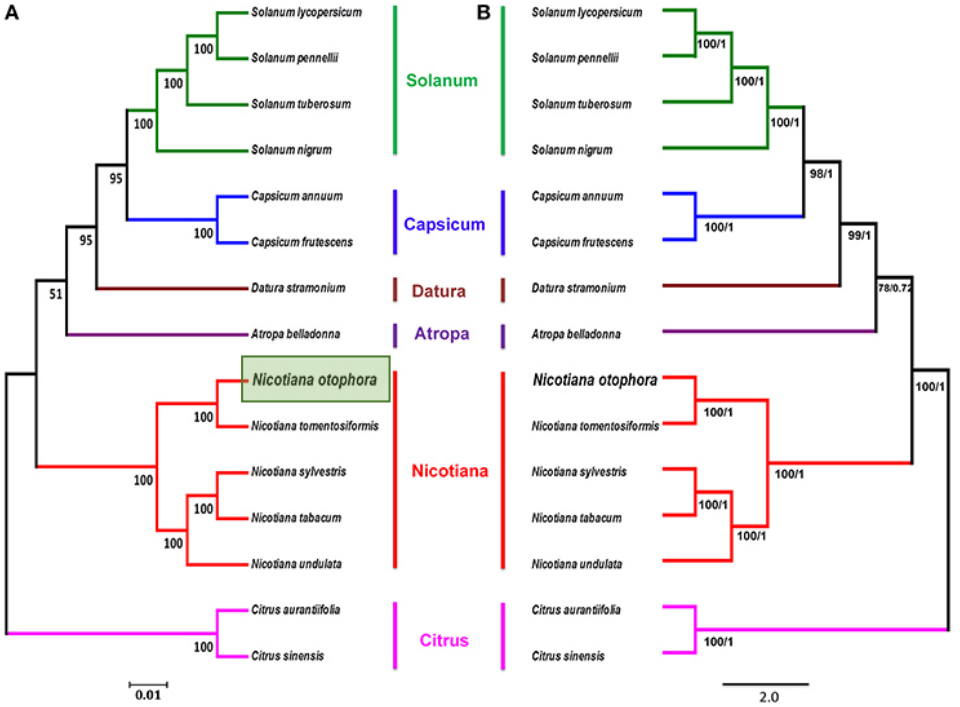
Asaf *et al.* (2016) *Front. Plant Sci.*

COMPARATIVE GENOMICS

Genomes alignment



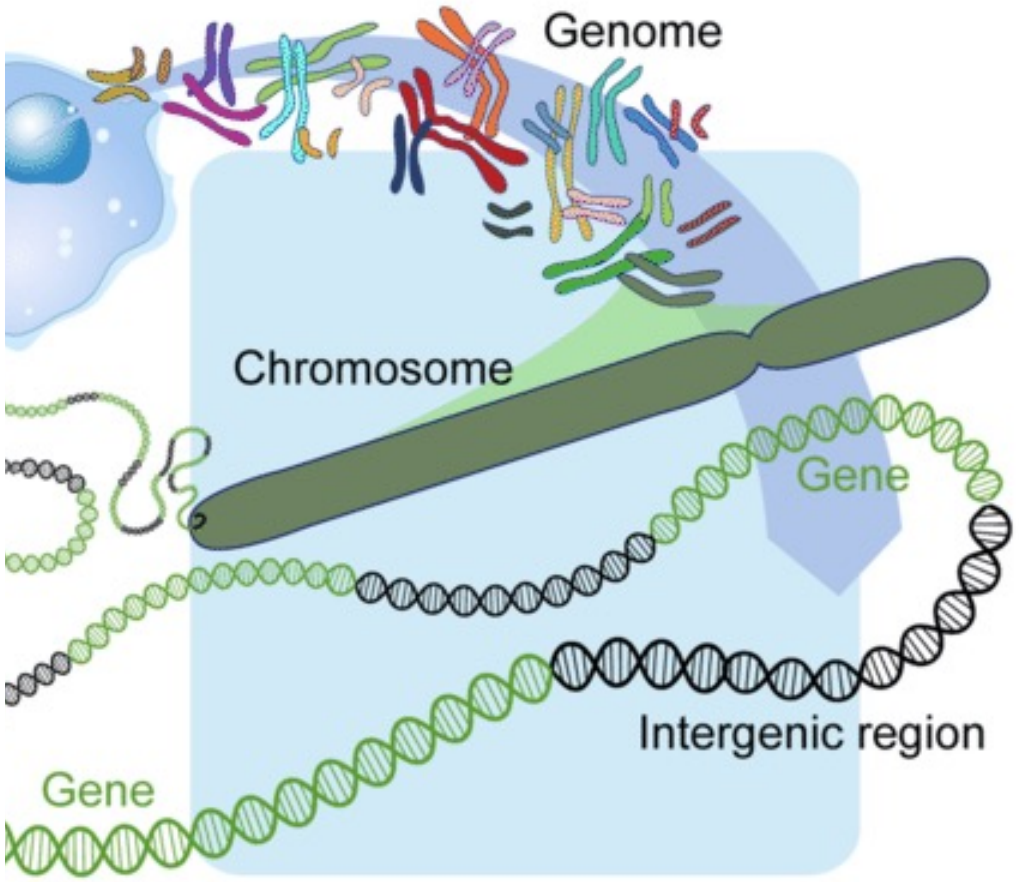
Phylogenetic relationships



Asaf *et al.* (2016) *Front. Plant Sci.*

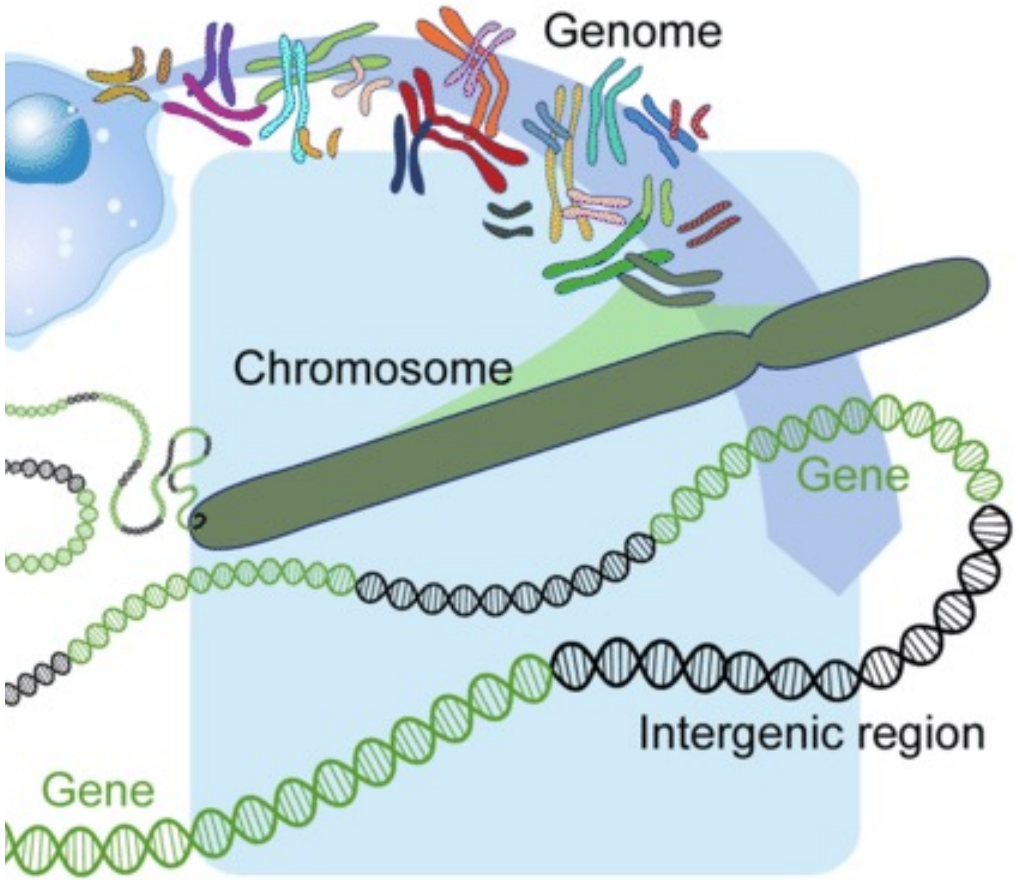
What is a Genome?

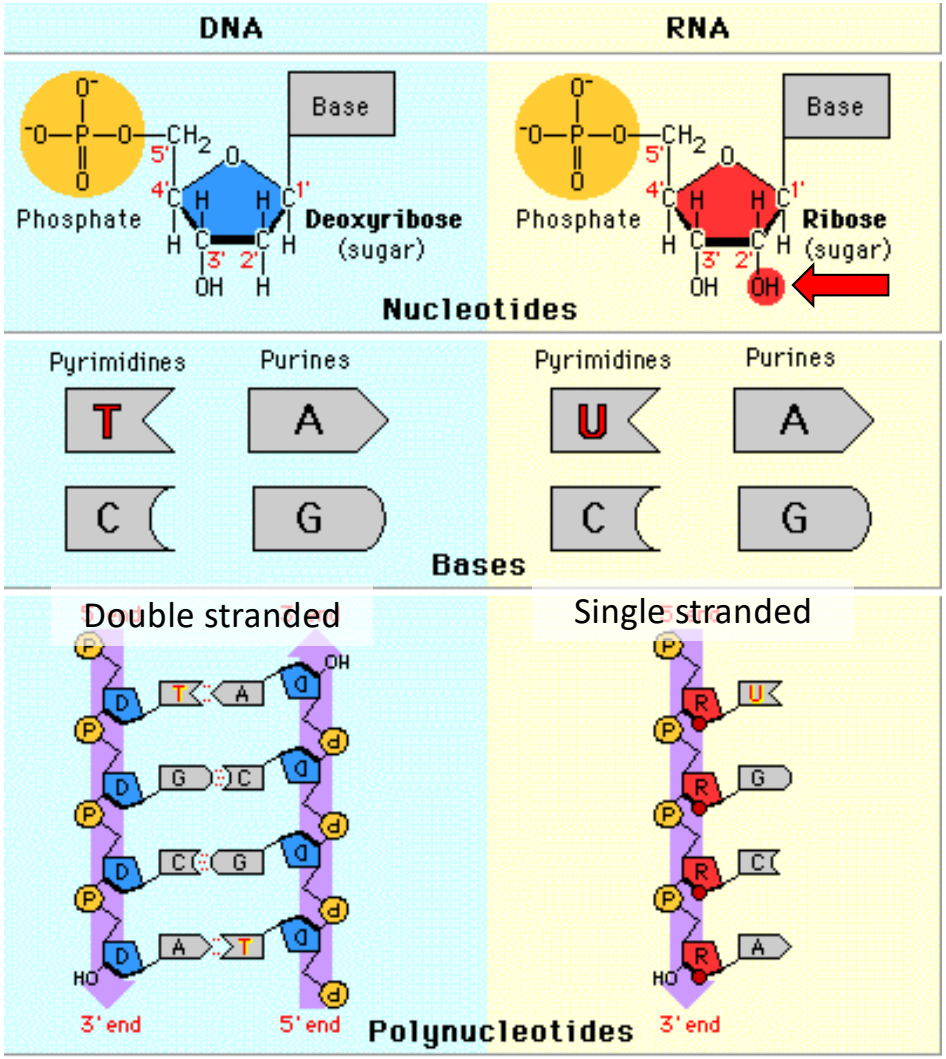
- A **genome** is the entire genetic complement of a living organism.
- Each organism possesses a genome containing the biological information needed to construct and maintain a living example of that organism.



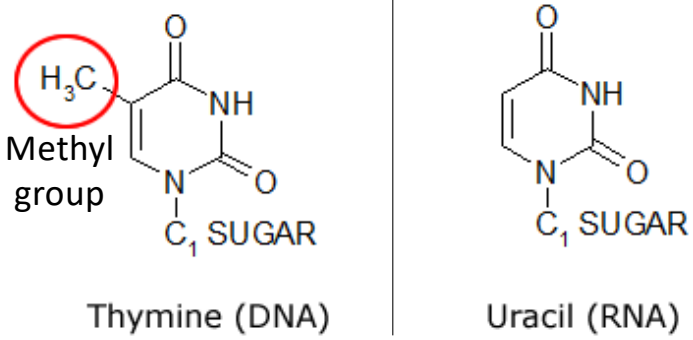
What is a Genome?

- Most genomes are made of **DNA** (deoxyribonucleic acid), but few viruses have **RNA** (ribonucleic acid) genomes.
- DNA and RNA are polymeric molecules made up of chains of monomeric subunits called nucleotides.



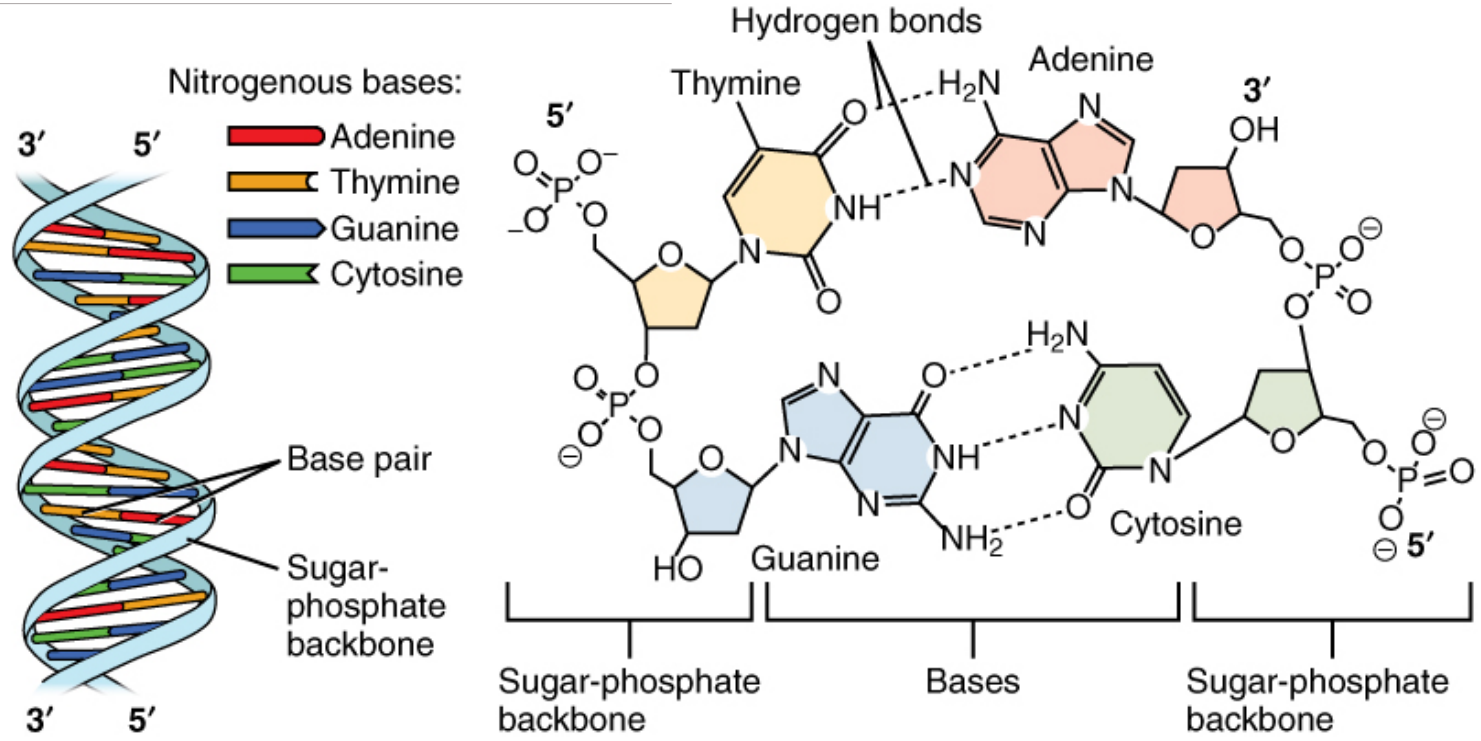


- The **OH bond** in RNA ribose makes molecule more reactive, especially in alkaline conditions.



- DNA is more resistant to enzymatic attacks (due to its helix structure) compared to RNA (impact on storage/preservation).

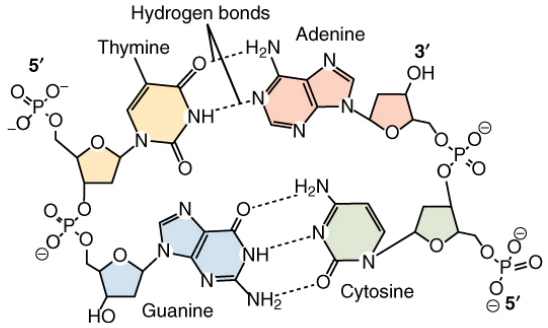
DNA – HYDROGEN BONDS

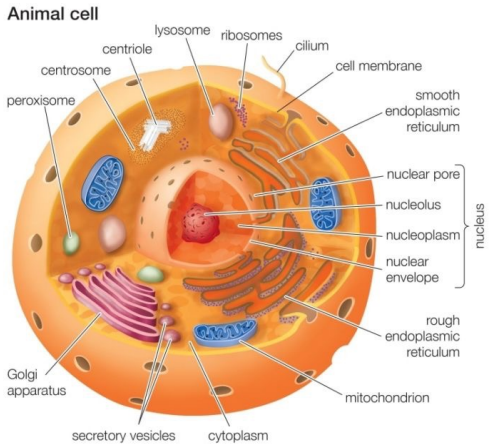


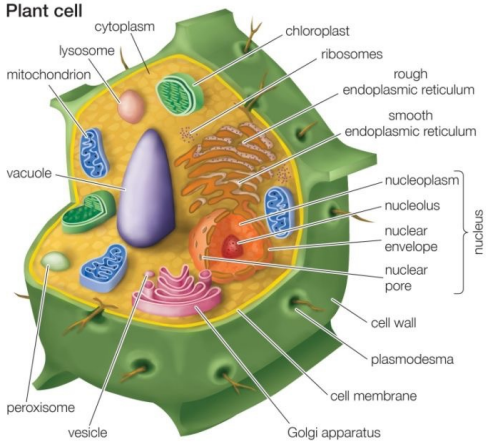
T-A: 2 hydrogen bonds
G-C: 3 hydrogen bonds

DNA – Hydrogen bonds

- Mitochondrial and chloroplastic genomes are enriched in AT.
- Nuclear genome (nrDNA) is enriched in GC.
- On average, plastid DNA GC-content is ~37%, whereas nrDNA GC-content is ~41%.
- These genome structural properties can be used to filter reads in bioinformatics pipelines and study gene trafficking between genomes.





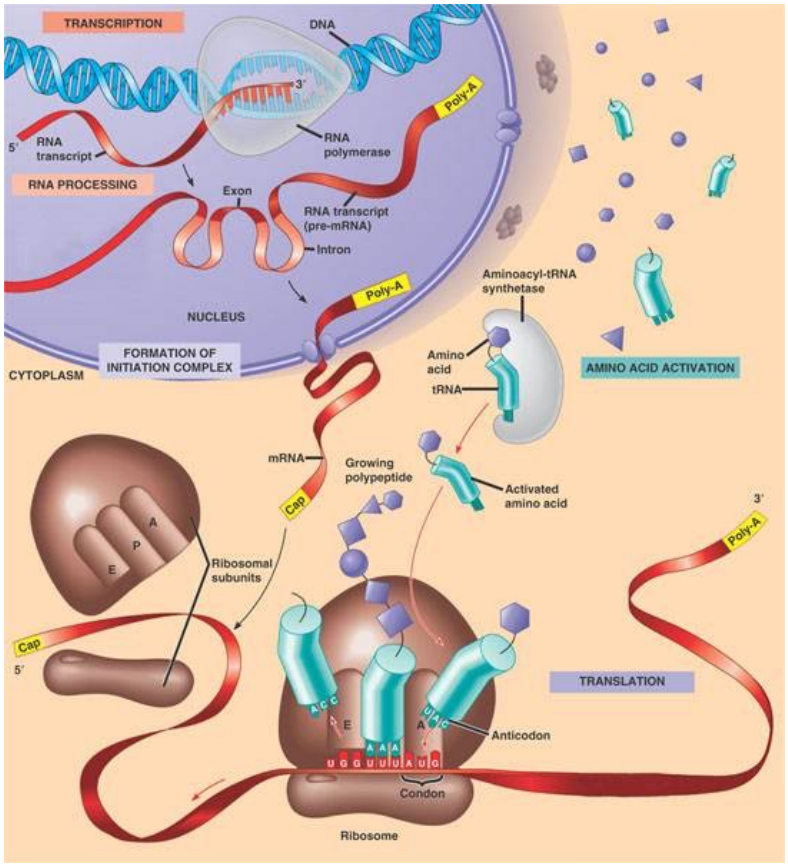


DNA vs. RNA – MODES OF TISSUE PRESERVATION

- The structural and chemical differences between DNA and RNA molecules must be considered to select a storage strategy.
- **What is the best strategy to preserve tissue for genomic/transcriptomic analyses?**
- Which tissue is best suited for the study? When should it be sampled?

Important for Lab. report

GENOME EXPRESSION



- **The genome is a store of biological information, but on its own it is unable to release the information to the cell.**
- This process is done through a complex series of biochemical reactions referred to as genome expression.

GENOME

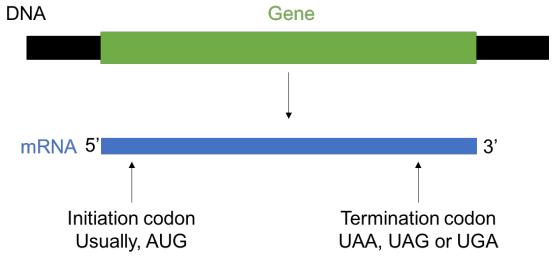
↓ Transcription

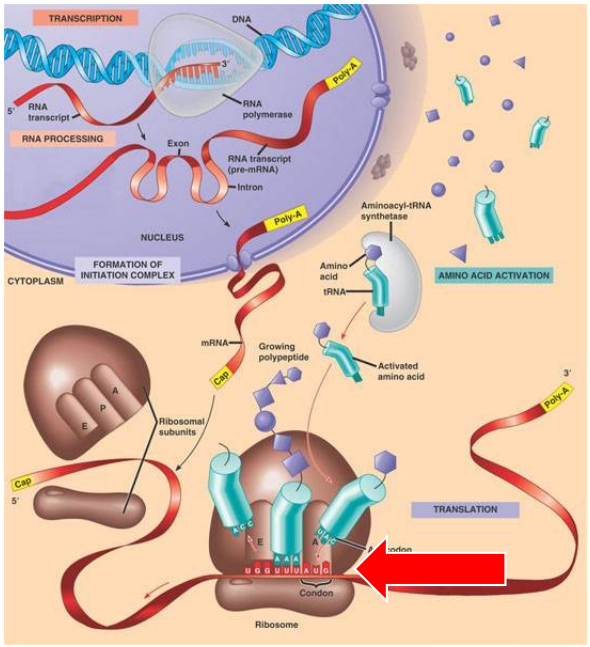
TRANSCRIPTOME
RNA copies of the active protein-coding genes

↓ Translation

PROTEOME
The cell's repertoire of proteins

GENETIC CODE – LINKING TRANSCRIPTOME & PROTEOME

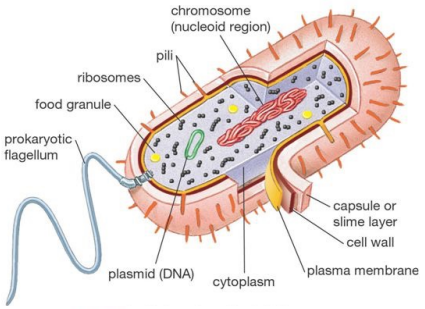


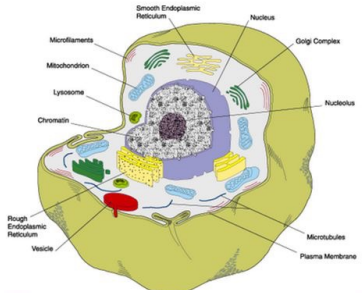


- The genetic code specifies **how the nucleotide sequence of an mRNA is translated into the amino acid (AA) sequence of a protein**.
- Different genetic codes depending on lineages.
- **20 amino acids:** high redundancy (except Met & Trp)
- **Start codon:** Met (sometimes Trp in cpDNA).

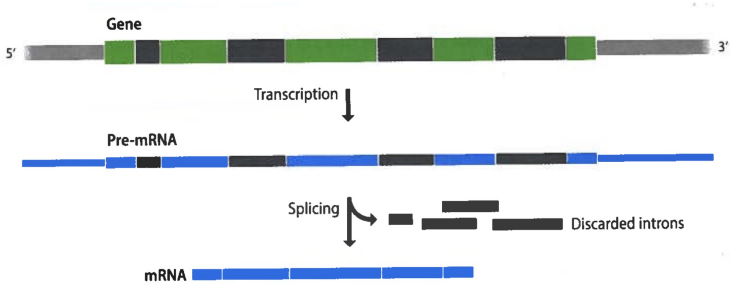
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC		UCC		UAC		UGC	
UUA	Leu	UCA		UAA	Stop	UGA	Stop
UUG		UCG		UAG		UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Gln	CGA	
CUG		CCG		CAG		CGG	
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lys	AGA	Arg
AUG	Met	ACG		AAG		AGG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

VARIETIES OF GENOME ORGANIZATION





Gene annotation is "easier" in prokaryotes due to the lack of introns



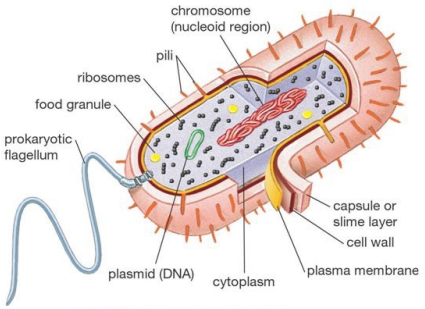
	Prokaryotes	Eukaryotes
DNA	DNA is naked	DNA bound to protein
	DNA is circular	DNA is linear
	Usually no introns	Usually has introns
Organelles	No nucleus	Has a nucleus
	No membrane-bound	Membrane-bound
	70S ribosomes	80S ribosomes
Reproduction	Binary fission	Mitosis and meiosis
	Single chromosome (haploid)	Chromosomes paired (diploid or more)
Average Size	Smaller (~1–5 μm)	Larger (~10–100 μm)

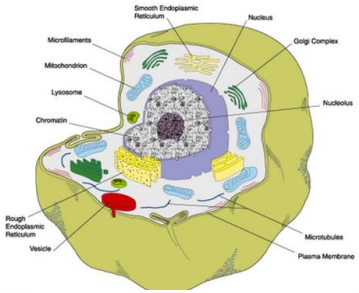
STRUCTURE AND FUNCTION OF CHLOROPLAST GENOME

Nicotiana otophora, a wild parent species of tabaco

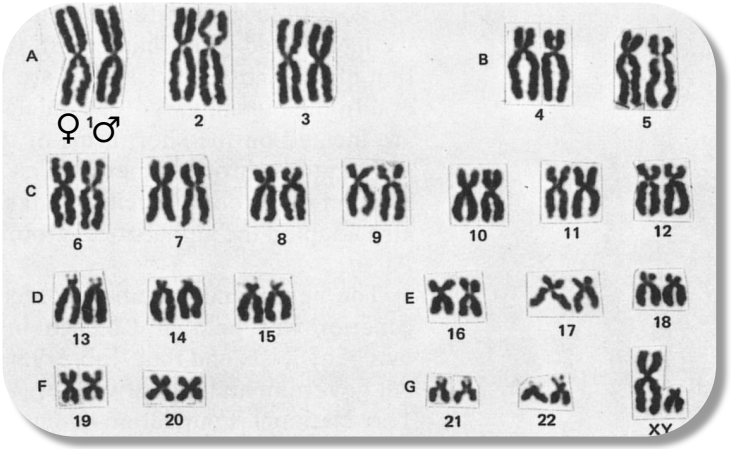
Although chloroplast genomes are of prokaryote origin several genes have introns!

VARIETIES OF GENOME ORGANIZATION





Human: 2n=2x=46

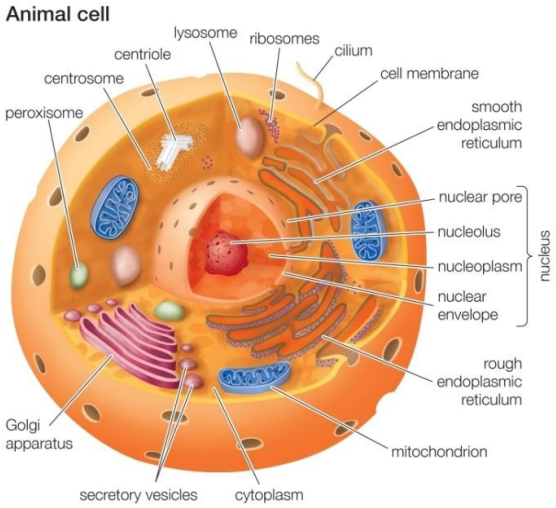


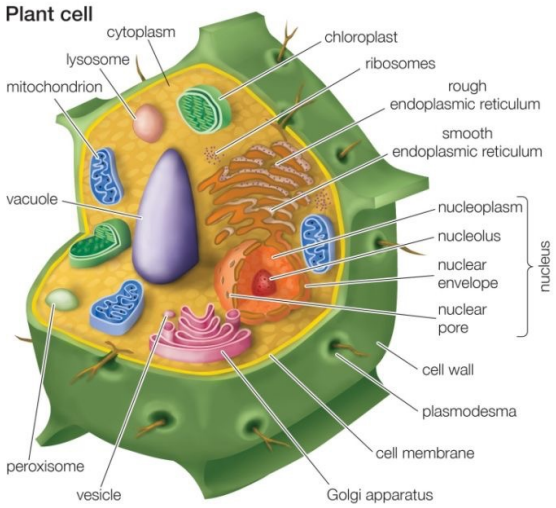
This difference impacts on genome complexity (heterozygosity) and influence bioinformatics (chromosome reconstruction and phasing)

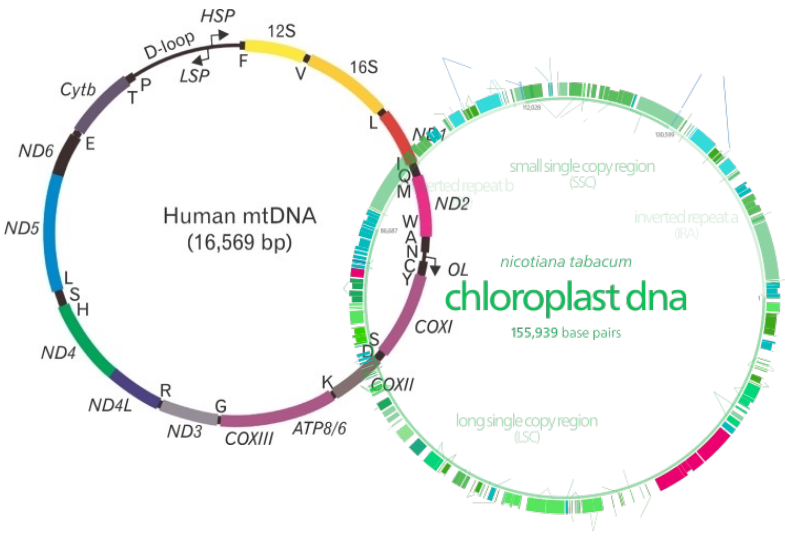
	Prokaryotes	Eukaryotes
DNA	DNA is naked	DNA bound to protein
	DNA is circular	DNA is linear
	Usually no introns	Usually has introns
Organelles	No nucleus	Has a nucleus
	No membrane-bound	Membrane-bound
	70S ribosomes	80S ribosomes
Reproduction	Binary fission	Mitosis and meiosis
	Single chromosome (haploid)	Chromosomes paired (diploid or more)
Average Size	Smaller (~1–5 μm)	Larger (~10–100 μm)

VARIETIES OF GENOME ORGANIZATION

- Eukaryotic cells also contain organelles with their own genomes.
- These organelles contain additional DNA usually in the form of single closed or circular molecules; un-complexed with histones, like the DNA of prokaryotes, but some genes have introns.

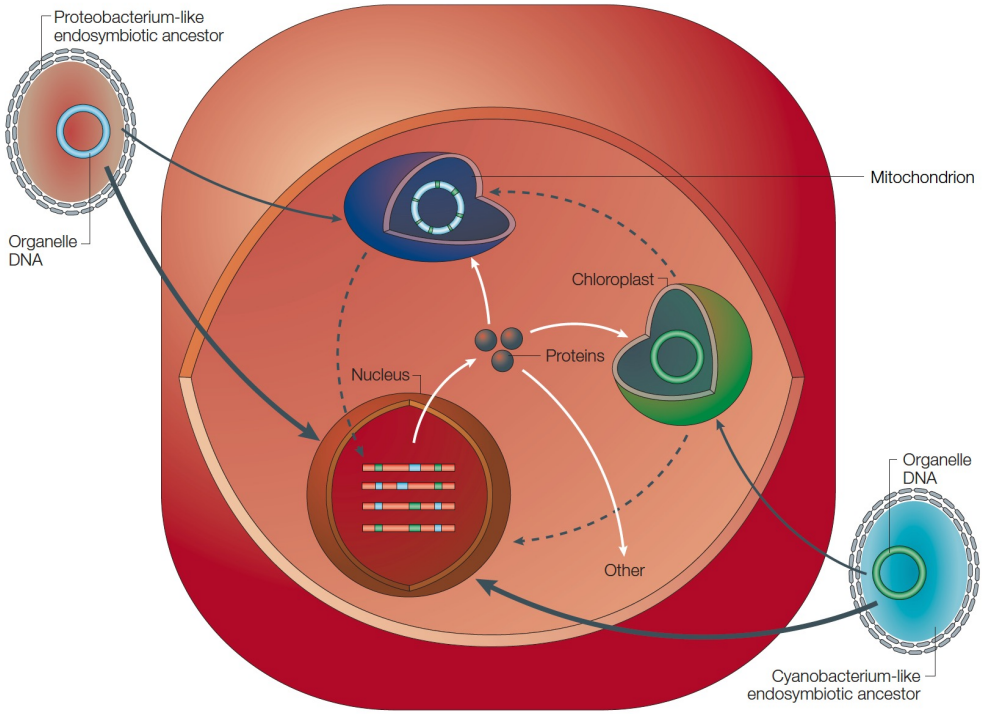






ORGANELLE GENOMES FORGE EUKARYOTIC CHROMOSOMES

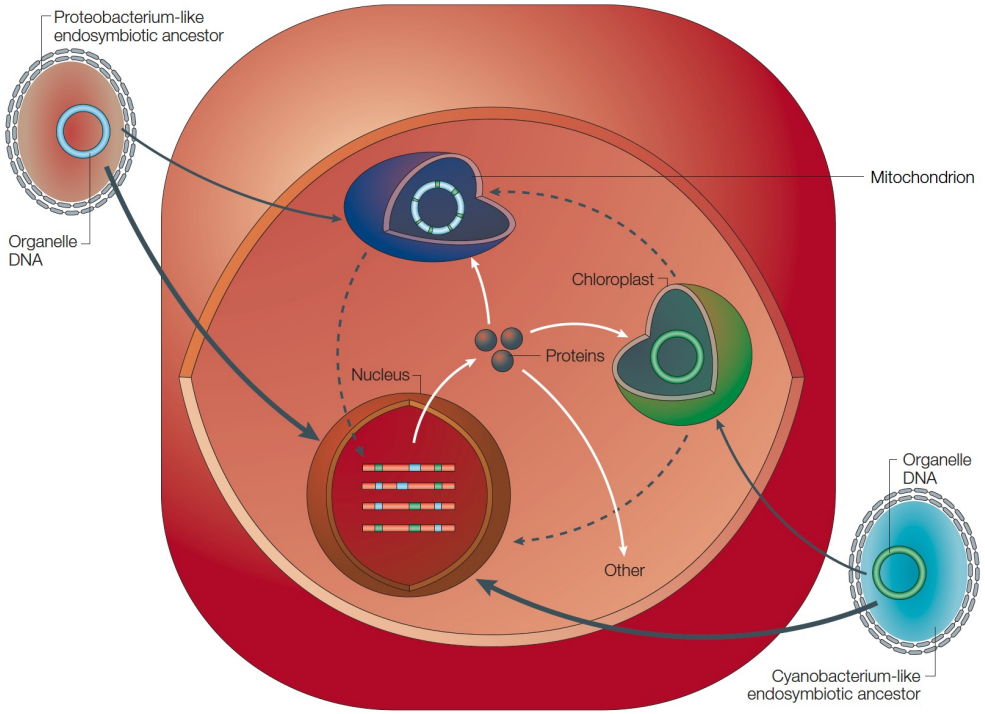
- Mitochondria and plastids were once free-living prokaryotes.
- **Most of the genes** originally present in the ancestors' genomes **have been transferred to the nucleus**, with few being retained in the organelles.
- These organelles heavily dependent on nuclear genes and import >90% of their proteins from the cytoplasm (e.g. RuBisCO).
- **Endosymbiotic gene transfer is not homogenous across lineages therefore complicating genome assembly.**



Timmis *et al.* (2004) *Nature Reviews*

ORGANELLE GENOMES FORGE EUKARYOTIC CHROMOSOMES

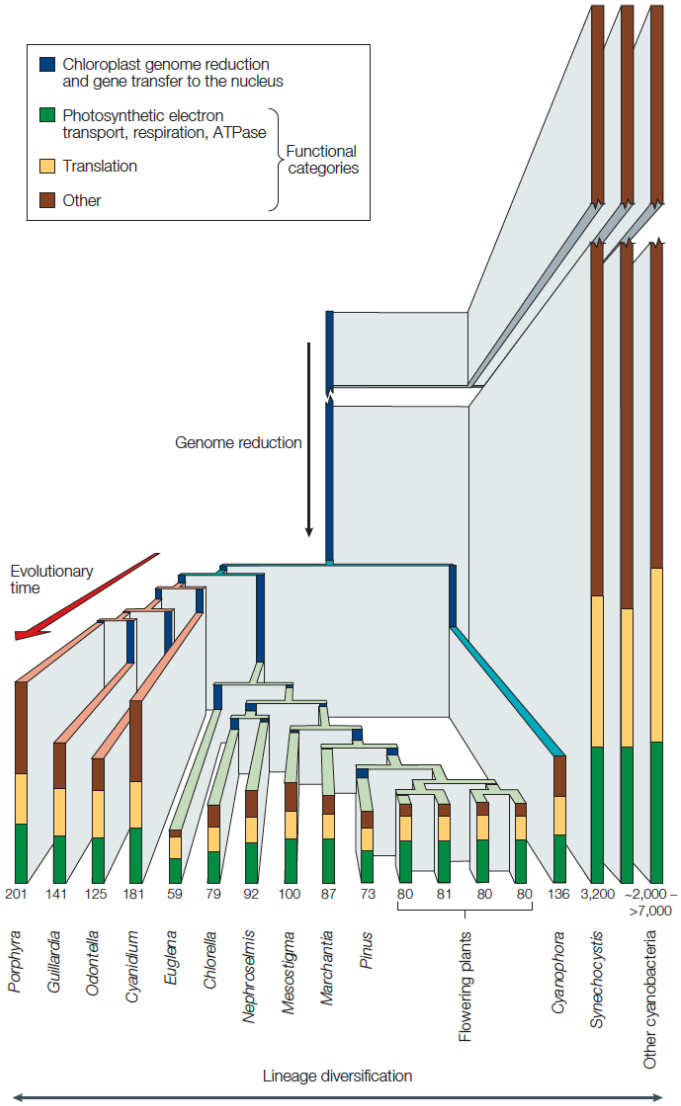
- Gene transfers into the nucleus is tricky to assess when conducting genome assemblies.
- **Tendency for genes to increase their GC-contents when they are transferred from organelles to the nucleus.**



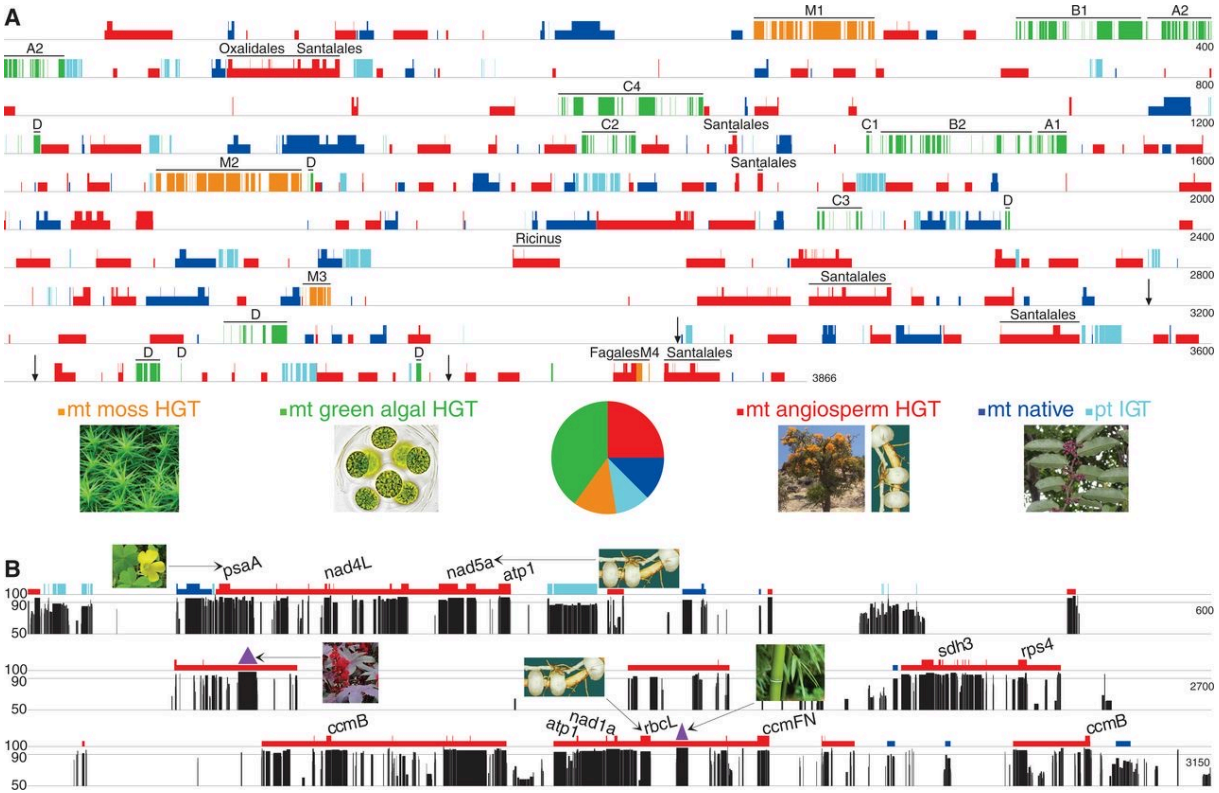
Timmis *et al.* (2004) *Nature Reviews*

REDUCTION OF SIZE OF CHLOROPLAST GENOMES OVER TIME

- Ancestor of plastids was a free-living cyanobacterium and therefore must have possessed several thousand genes as did its contemporaries.
- **Plastids have relinquished most of their genes to the genome of their host cell.**
- This gene relocation process occurred massively at the onset of endosymbiosis and continued in parallel during algal diversification.
- We are still at the infancy of unraveling this process, but NGS provides a boost.



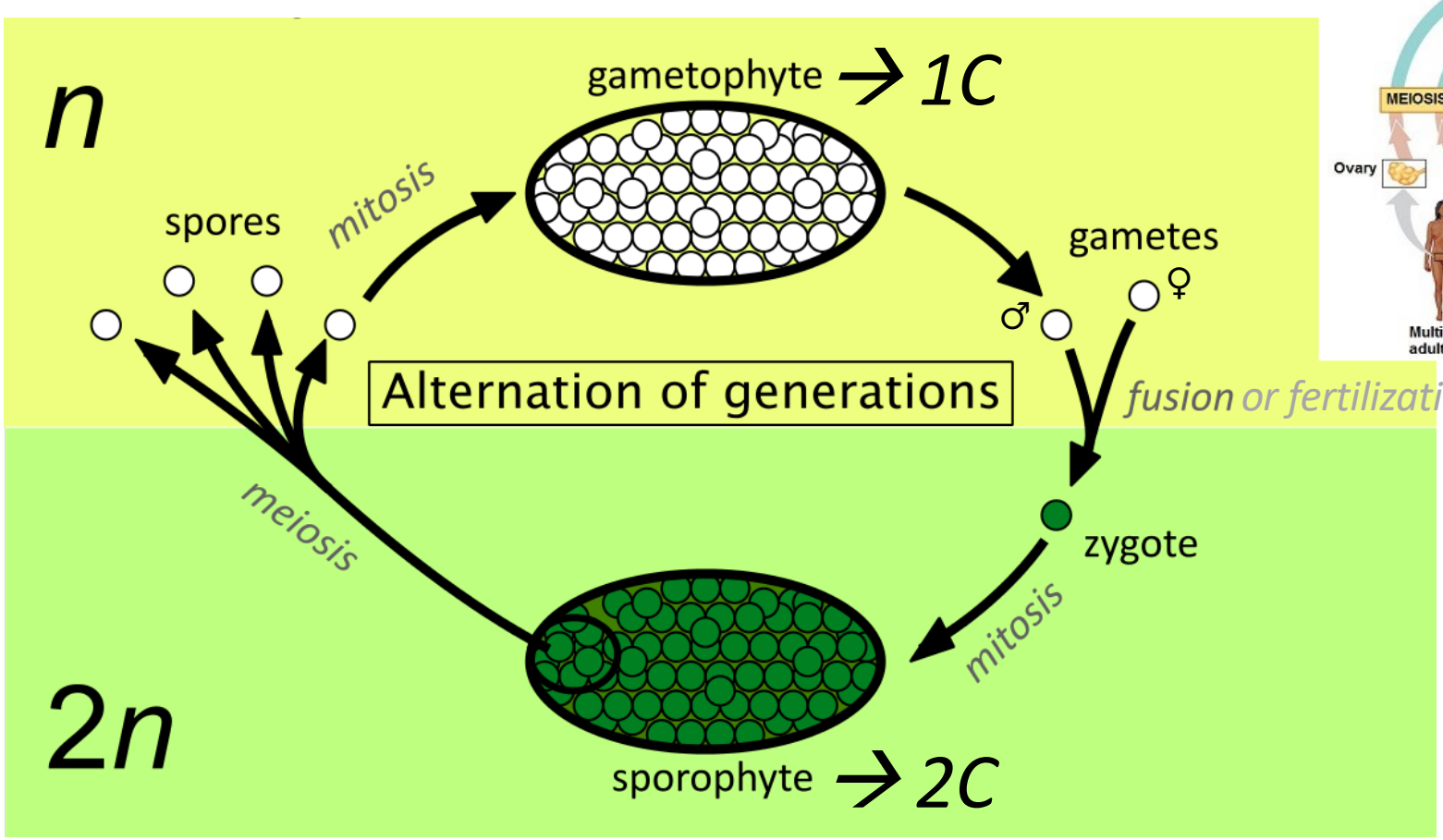
HORIZONTAL GENE TRANSFERS IN PLANTS' MITOCHONDRIAL GENOMES

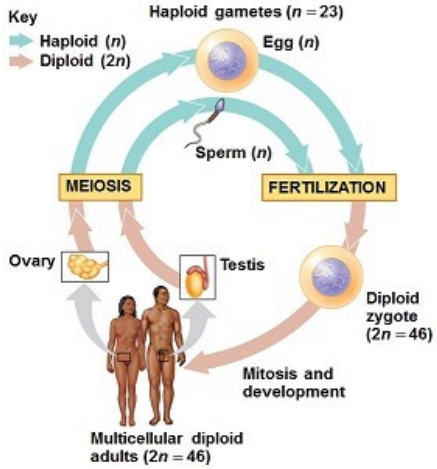


- Plant mitochondrial genomes (mtDNA) are not as conserved as those of animals.
- In plants, mtDNA genomes greatly vary in sizes and number of genes due to horizontal gene transfers (e.g. entire mtDNA genomes in *Amborella*)

Rice *et al.* (2013) *Science*

QUANTIFYING GENOME SIZE – LIFE CYCLE

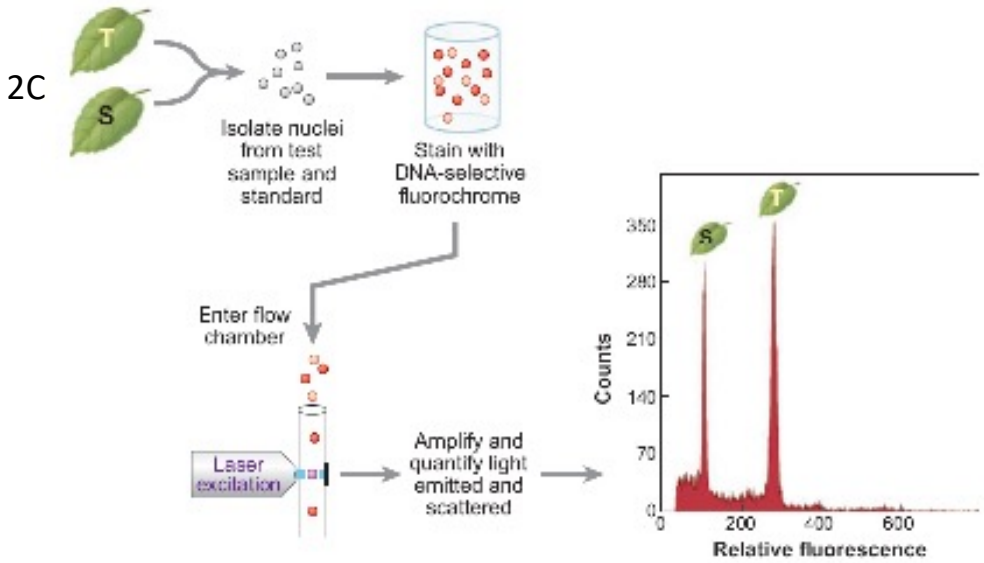




1C = DNA content of the gametophytic (n) set of chromosomes

QUANTIFYING GENOME SIZE – FLOW CYTOMETRY

- Flow cytometry is the best strategy to estimate genome size (and ploidy).
- It is key to design sequencing strategy (e.g. how many sequence data should be produced to assemble genome).



S 2C value (pg) = T 2C value x S 2C mean peak position / T 2C mean peak position

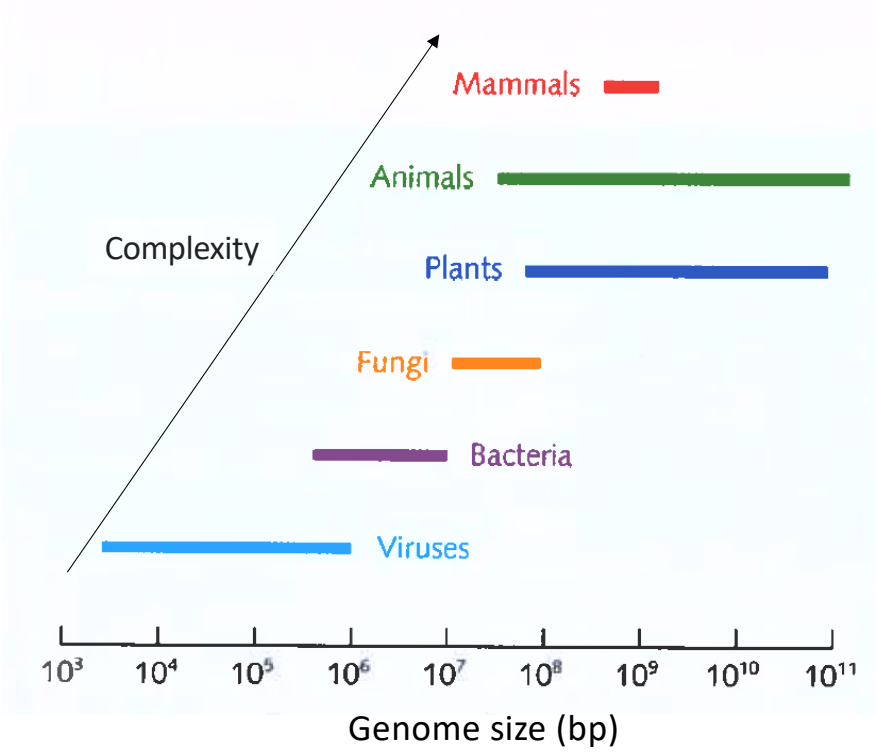
S Nb bp = mass in pg x 0.978 x 10^9
or
1 pg = 978 Mbp

C= DNA content of the haploid set of chromosomes
S= Sample
T= Reference (= sample of known genome size used as standard to estimate size of sample)

Doležel *et al.* (2007) *Nature Protocols*

GENOME SIZES AND ORGANISMS' COMPLEXITY

- Large variation in genome sizes across the Tree of Life!
- **Broad scale: correlation between genome size (= amount of DNA per cell) and the complexity of an organism**.
- Fine scale: this correlation does not necessarily hold true within closely related species (especially in plants with processes of whole genome doubling).
- The number of genes is usually used as a proxy to reflect an organism's complexity.



GENOME SIZES AND ORGANISMS' COMPLEXITY

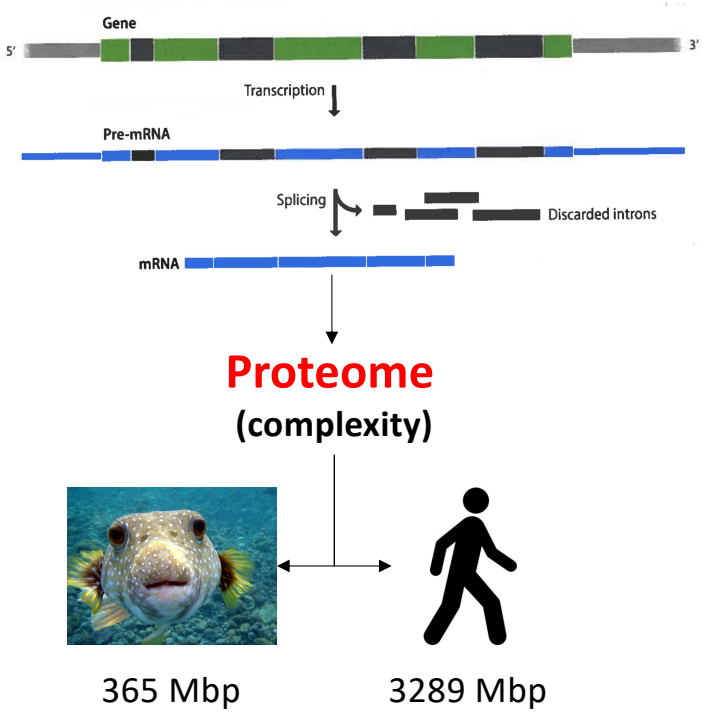
Species	Genome size (Mb)	Coding (%)	Approx. number of genes	Gene density (kb/gene)
Escherichia coli	4.64	88	4485	1.03
Yeast	12.5	70	6000	2.1
Pufferfish	365	15	23000	10
Arabidopsis thaliana	115	29	23000	6
Human	3289	1.3	23000	143

Pufferfish and human exhibit different levels of complexity, but they share the same number of genes

→ Where does complexity take place?

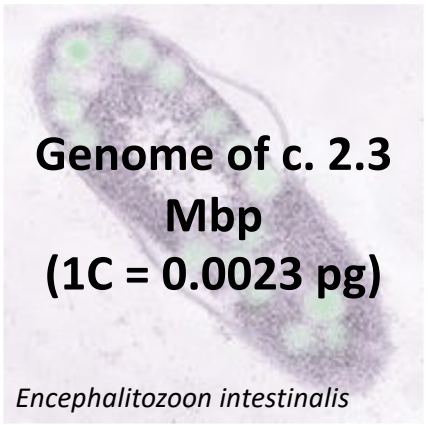
GENOME SIZES AND ORGANISMS' COMPLEXITY

- Number of genes is not a good proxy of an organism's "complexity".
- **Complexity takes place at the proteome level, but it is orchestrated at the transcriptome level** via two main processes:
  - ✓ Alternative splicing.
  - ✓ RNA editing.
- **Alternative splicing is a process during gene expression that results in a single gene coding for multiple proteins.** In this process, particular exons of a gene may be included within or excluded from the final mRNA. For instance, in mammals, billions of antibodies arise from <100 exons.

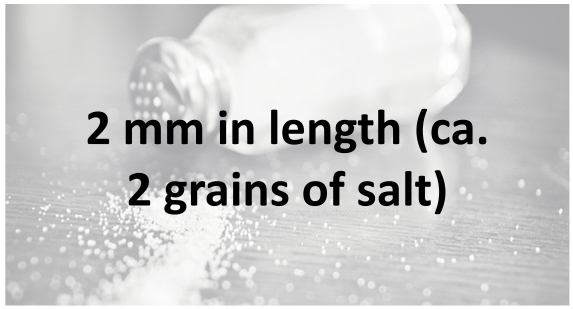


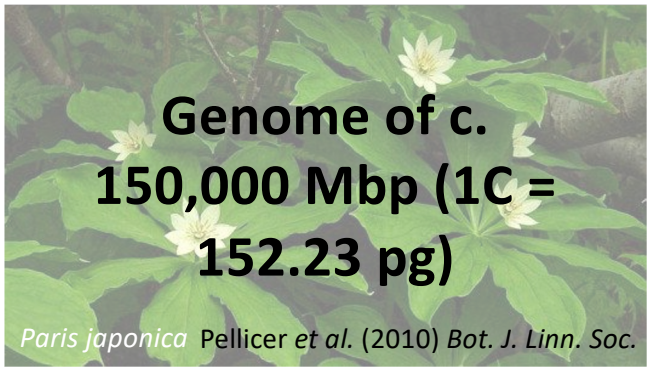
VARIATION OF GENOME SIZES – EUKAROYOTIC GENOMES

66,000-fold variation



Encephalitozoon intestinalis





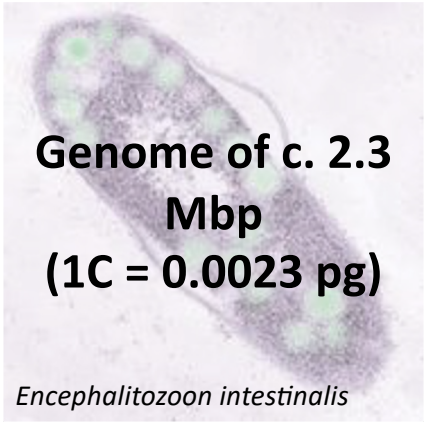
Paris japonica Pellicer *et al.* (2010) *Bot. J. Linn. Soc.*

96 m in length (Big Ben)

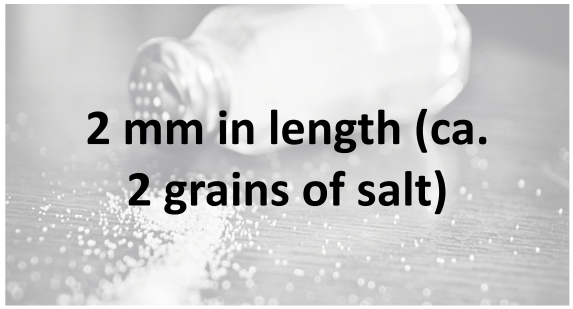
Mbp = mega base pairs = 1,000,000 bp
1 nucleotide = c. 0.34 nm

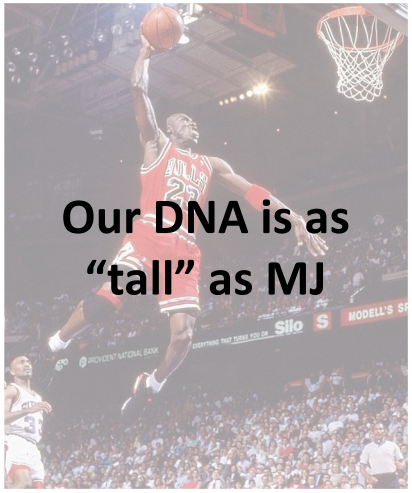
VARIATION OF GENOME SIZES – EUKAROYOTIC GENOMES

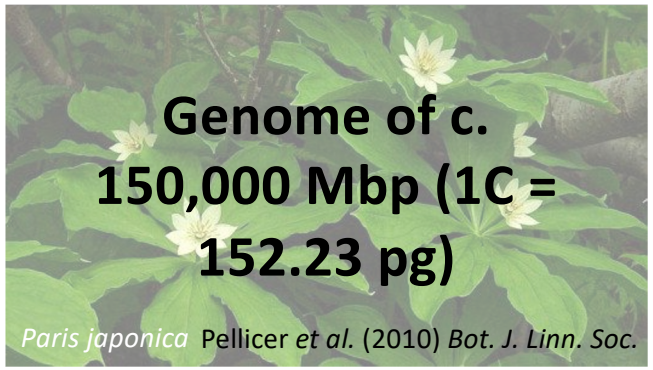
66,000-fold variation



Encephalitozoon intestinalis





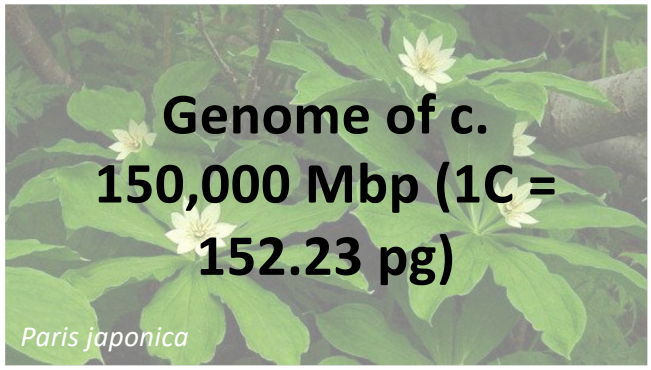


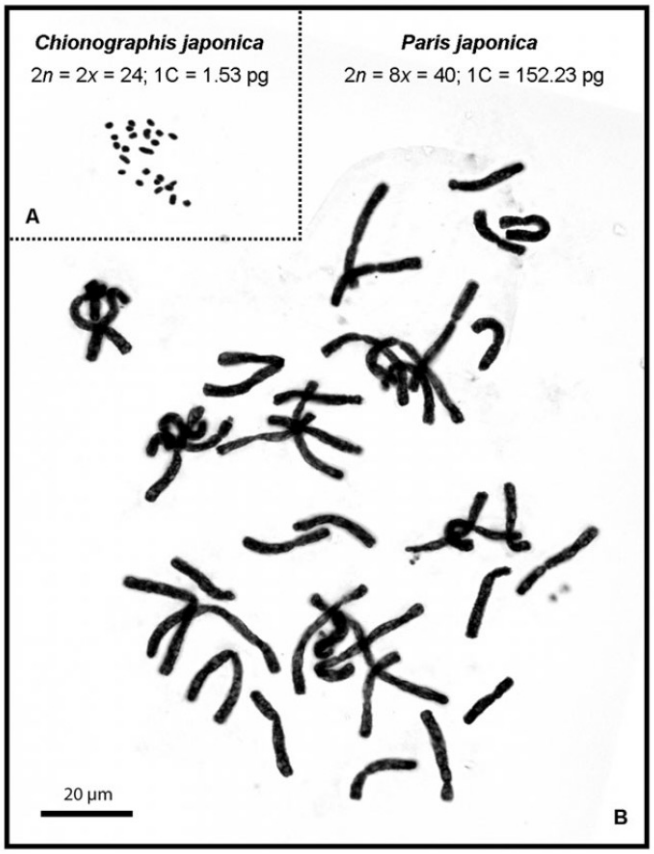
Paris japonica Pellicer *et al.* (2010) *Bot. J. Linn. Soc.*

96 m in length (Big Ben)

Mbp = mega base pairs = 1,000,000 bp
1 nucleotide = c. 0.34 nm

VARIATION OF GENOME SIZES – EUKAROYOTIC GENOMES

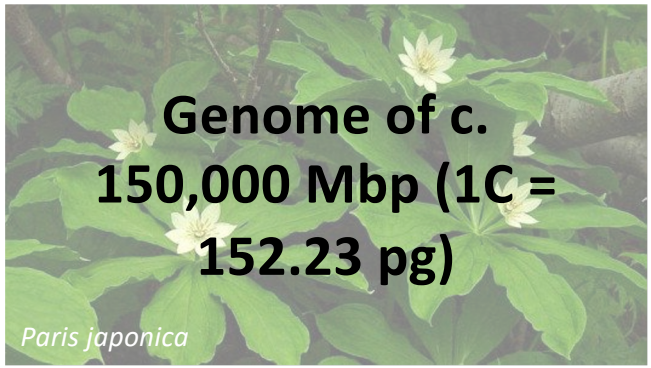




The "obese" genome of *P. japonica* is associated with:

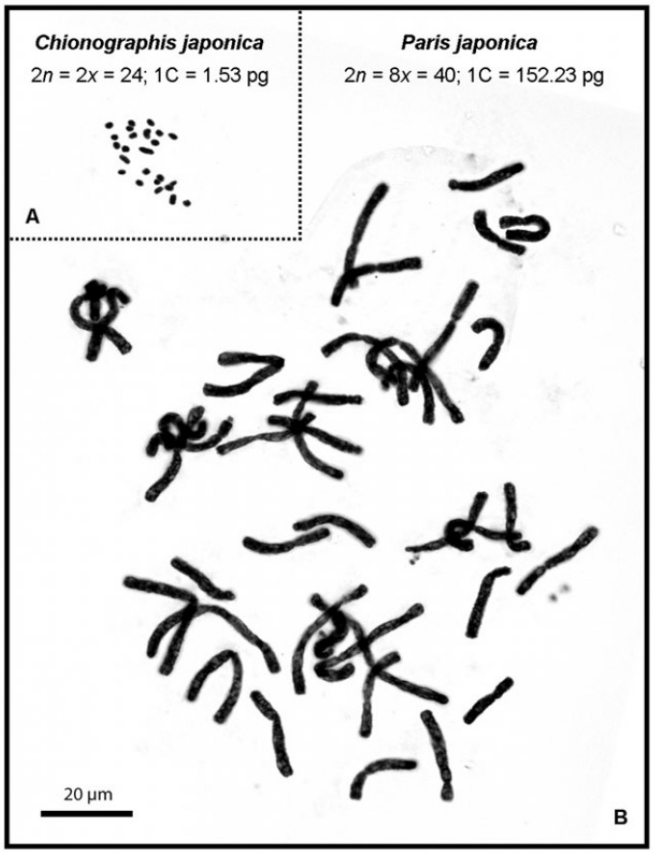
1. Polyploidization event(s) or whole genome doubling (8x).
2. Extreme karyological rearrangements (> 10-fold difference in chromosome length with other species in family).

VARIATION OF GENOME SIZES – EUKAROYOTIC GENOMES



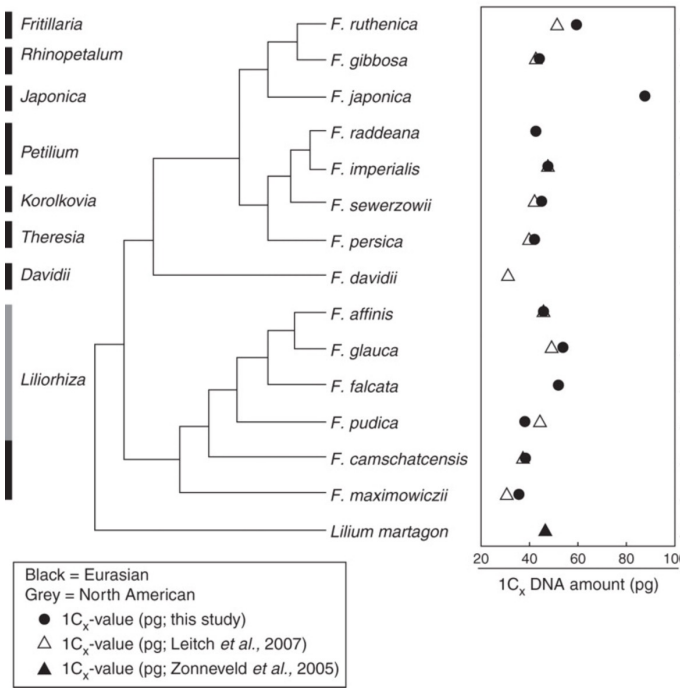
Basic knowledge on chromosome numbers and genome sizes are essential to design any genomic project, especially to develop a sequencing strategy.

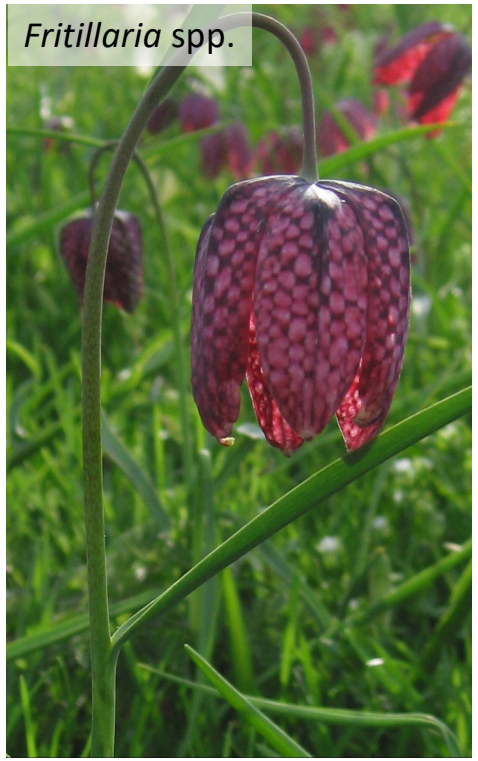
Rule of thumb: Aim for sequencing depth between 100-150x to assemble an eukaryote nuclear genome (but this depends on the level of heterozygosity)



VARIATION OF GENOME SIZES – EUKAROYOTIC GENOMES

- Large variation of genome sizes in *Fritillaria*, most likely connected to transposable elements.
- **Phylogenetic evidence can't predict species genome sizes.**
- First step in designing a genomic project is therefore to estimate the organism's genome size and ploidy level.

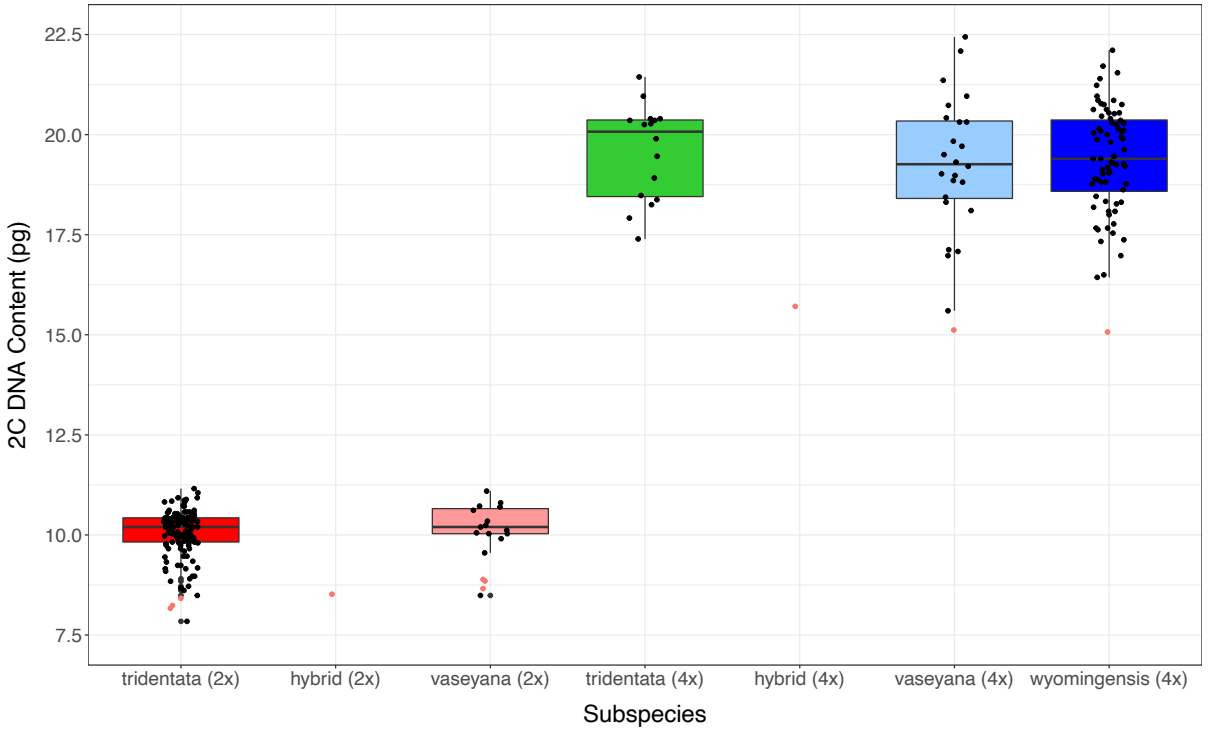




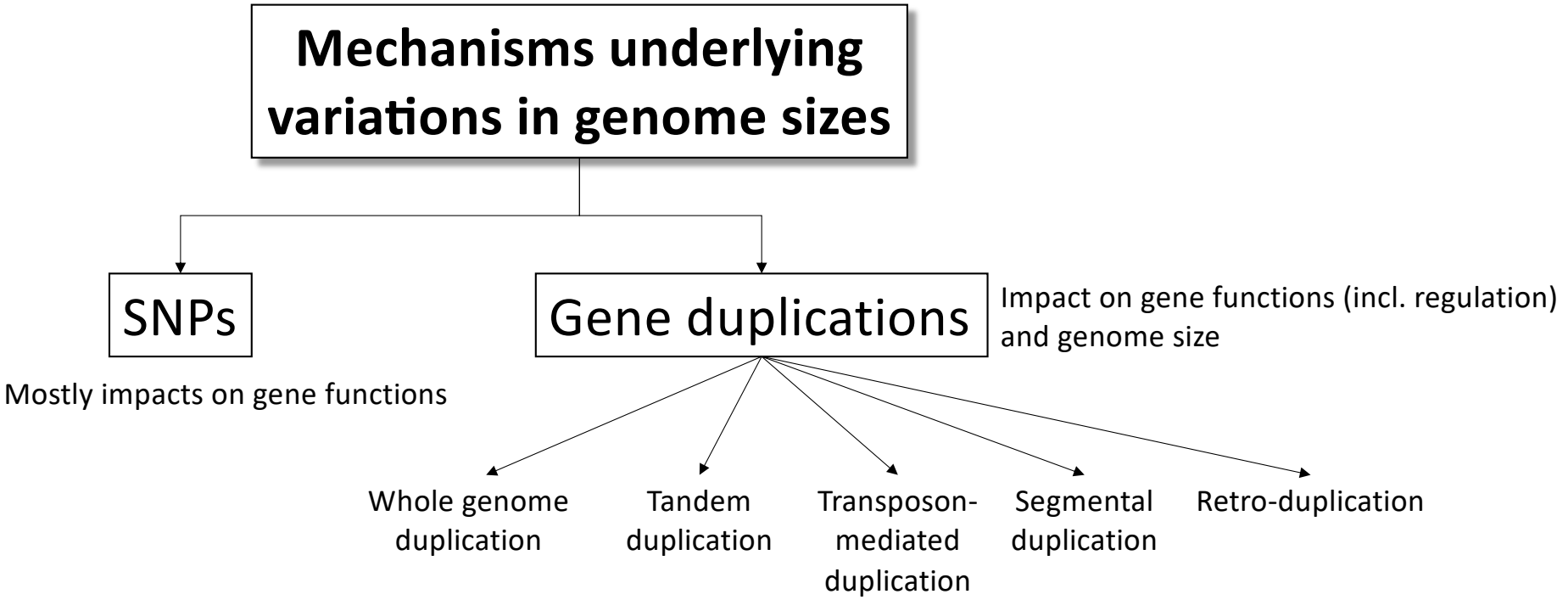
Ambrozova *et al.* (2011) *Ann. Bot.*

VARIATION OF GENOME SIZES – EUKAROYOTIC GENOMES

Sagebrush genome is 3x bigger than human genome (ca. 9Gbp)



How do genomes differ?

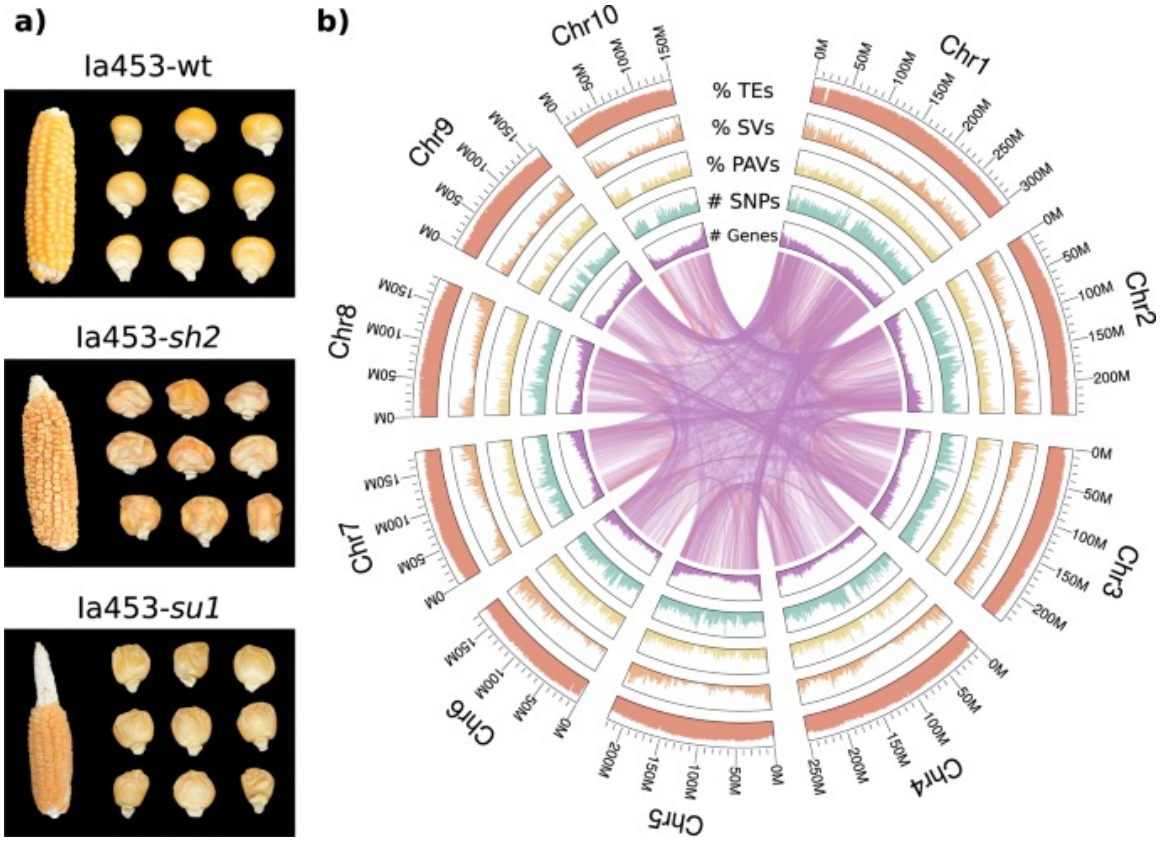


How do genomes differ?

Single nucleotide polymorphisms (SNPs)

- The most common type of genetic variation.
- Each SNP represents a difference in a single nucleotide.
  - ✓For instance, a C is replaced by a T.
- In the human genome, SNPs occur on average once in every 300 nucleotides.
  - ✓There are roughly 10 million SNPs in our genome.

EXAMPLE OF SNPS AND GENOME FEATURES

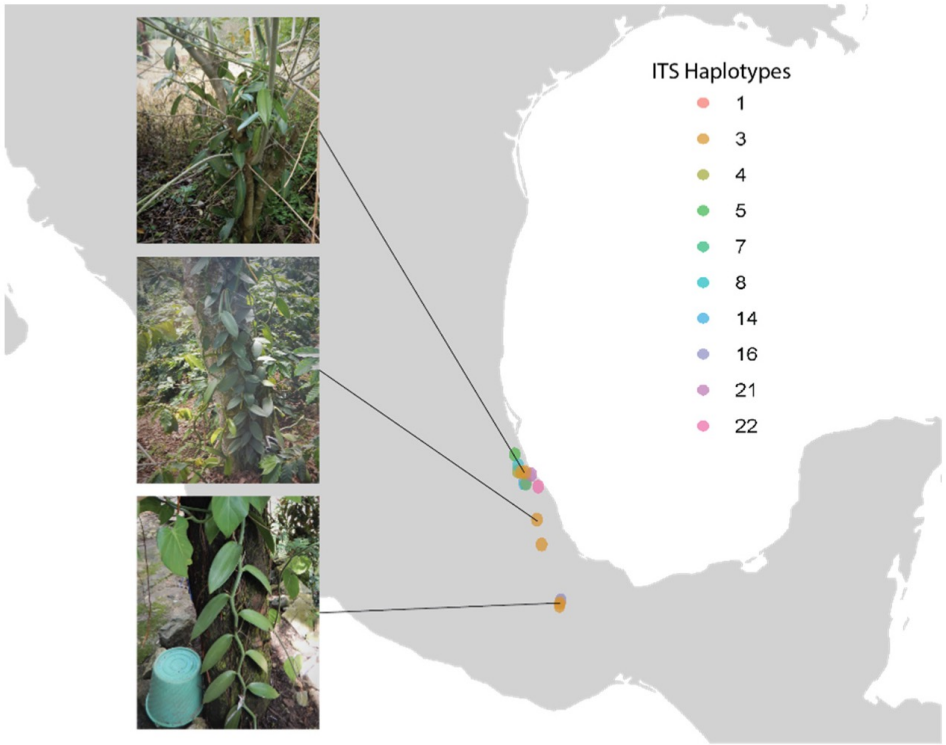


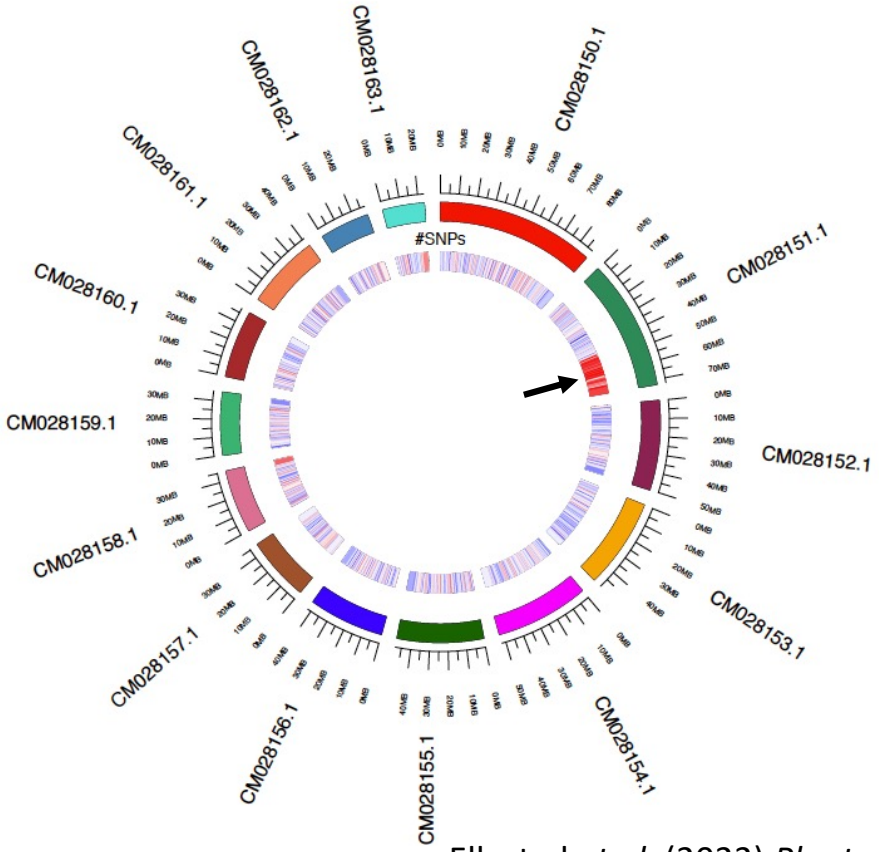
Hu *et al.* (2021)

Mini Report 3

Example of SNPs and Genome Features

Vanilla in Mexico



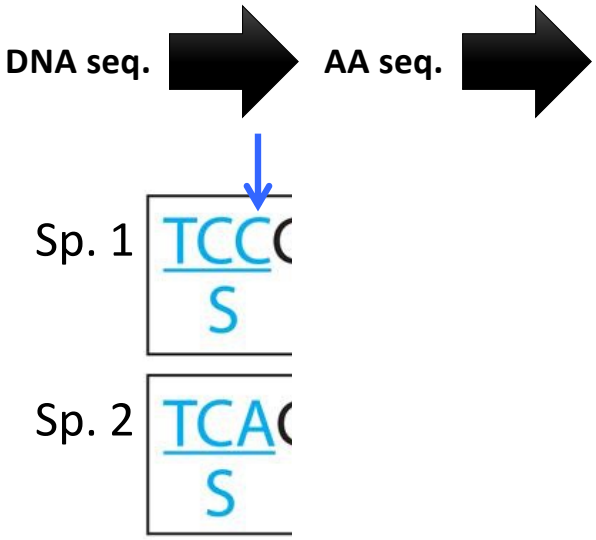


Ellestad *et al.* (2022) *Plants*

How do genomes differ?

Impact of SNPs on coding genes

- Synonymous vs. non-synonymous substitutions



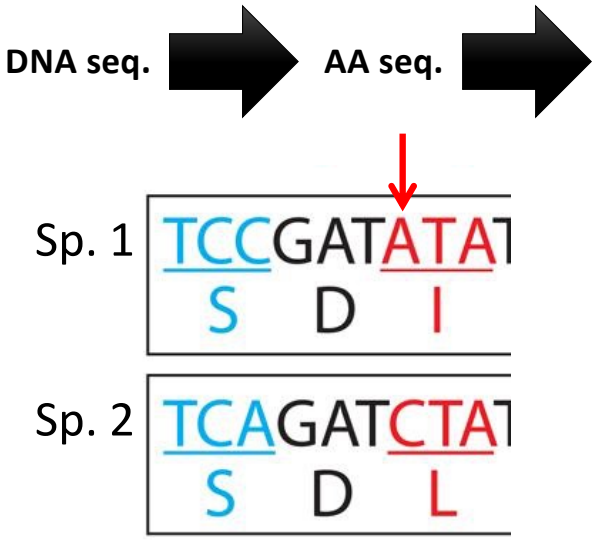
Synonymous substitution – mutation that does not alter amino acid seq.

1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

How do genomes differ?

Impact of SNPs on coding genes

- Synonymous vs. non-synonymous substitutions

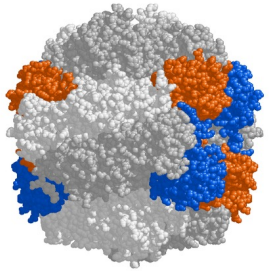


Non-synonymous substitution – mutation altering the amino acid seq.

1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G

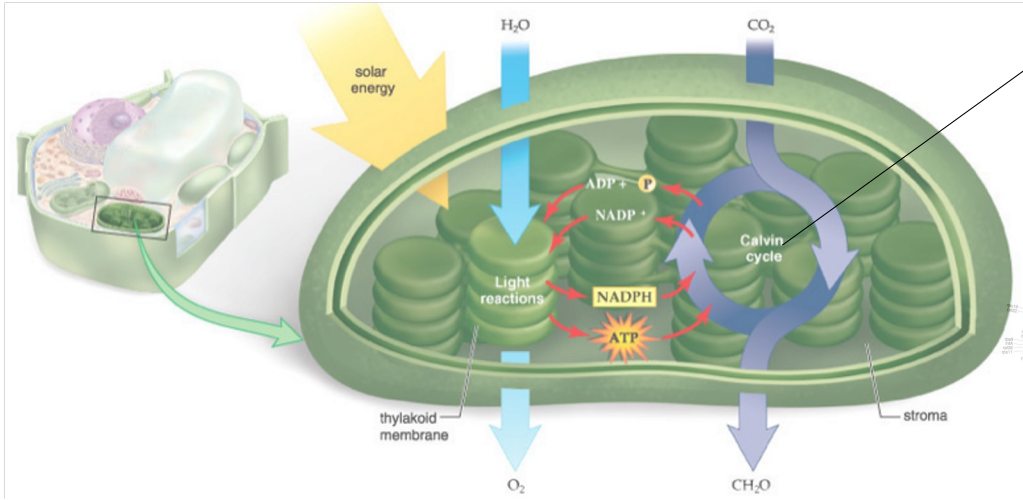
EVOLUTION OF CARBON FIXATION IN FACE OF CLIMATE CHANGE

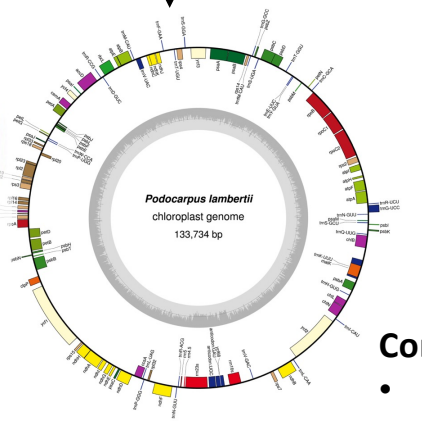
RuBisCO relies on genes located in nuclear and chloroplast genomes



Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO)

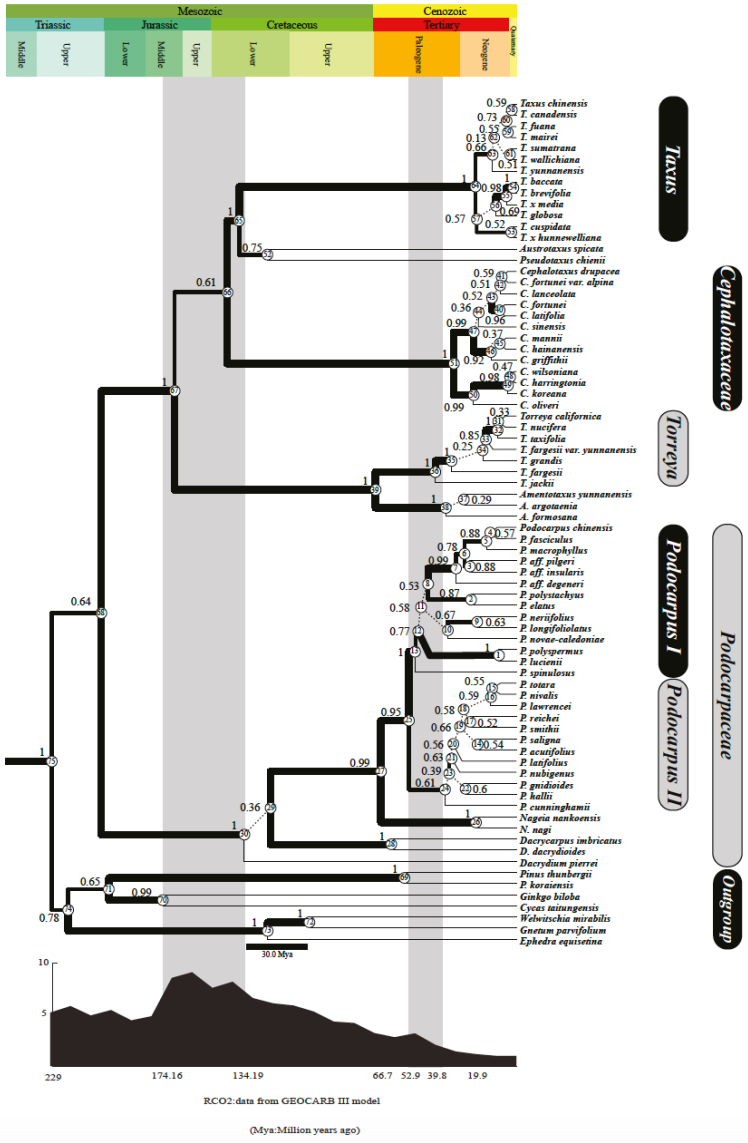
Carbon fixation



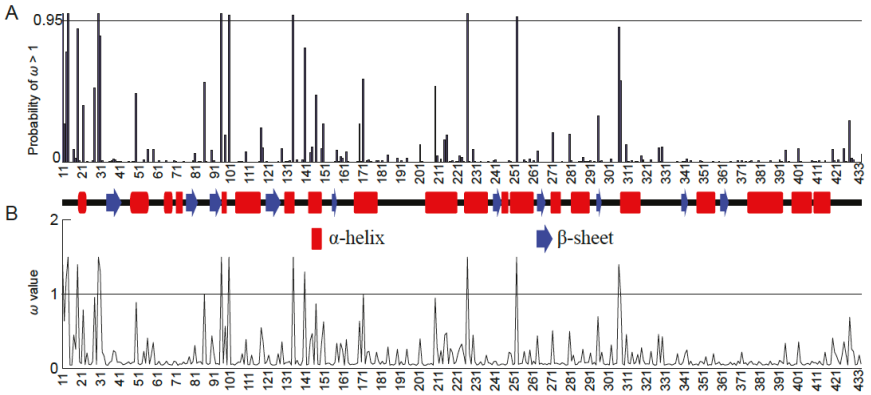


Consists of two subunits:

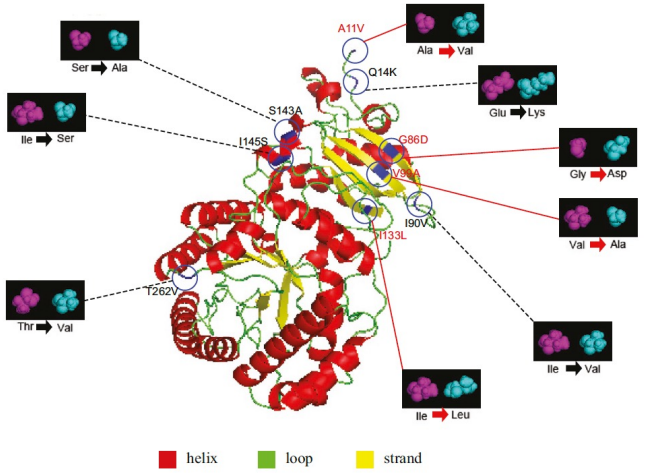
- Large (about 55,000 Da) subunit encoded by chloroplastic ***rbcL***.
- Small (about 13,000 Da) subunit encoded by several nuclear genes.



Look for signatures of natural selection along sequence



Positively selected sites evolving in ancestor of Podocarpaceae

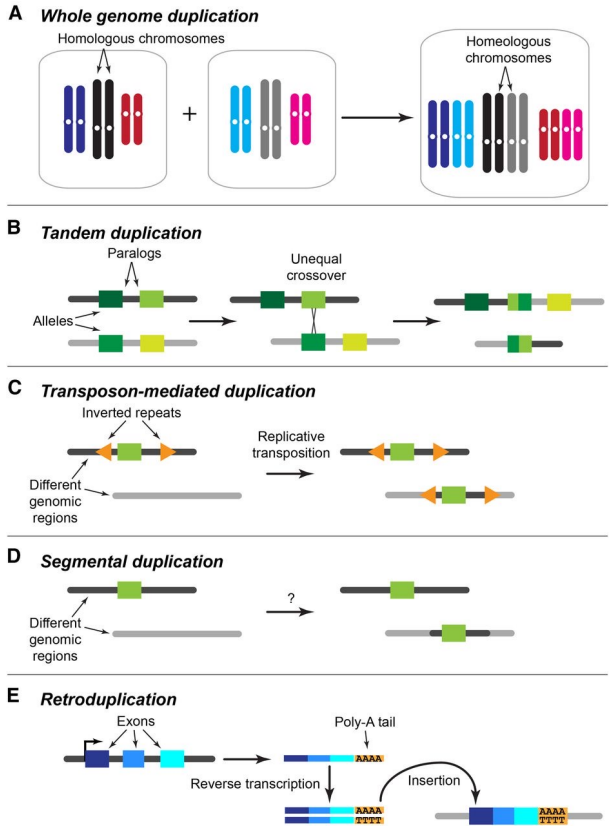


Sen *et al.* (2011) *Biology Direct*

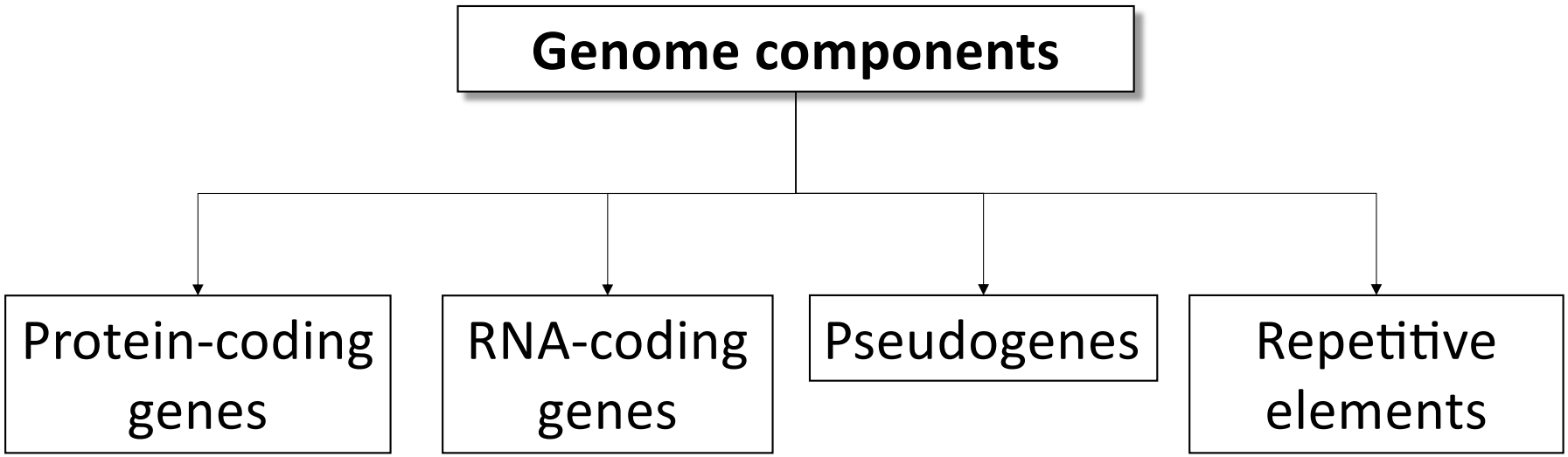
How do genomes differ?

Gene duplications

- This process drives the evolution of organisms.
- **Plant diversity has arisen largely following the duplication and adaptive specialization** of pre-existing genes (e.g. flower).
- On average, **65% of annotated genes in plant genomes have a duplicate copy**. Most copies were derived from whole genome doubling (WGD).
- In animals, WGD is less frequent, but there are several cases of large-scale segmental duplications. For instance, this process explains the difference in genome sizes between the human and chimpanzee genomes.



WHAT IS "IN" A GENOME?



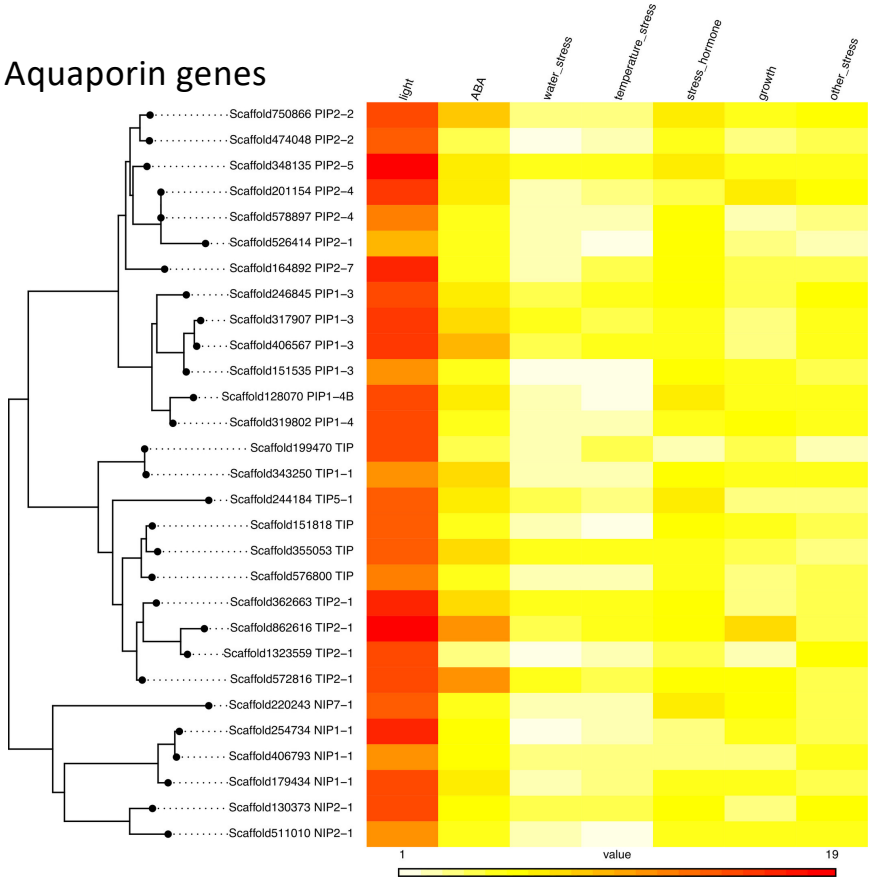
What is "in" a genome?

Protein-coding genes

- Small fraction of genomes. For instance, >2-3% in the human genome representing ca. 23,000 genes.
- Many protein-coding genes appear in multiple copies (= paralogues), either identical or diverging into families obtained through gene duplication.

Important for Lab. report

Example of paralogues in Sagebrush



Aquaporin genes in sagebrush genome

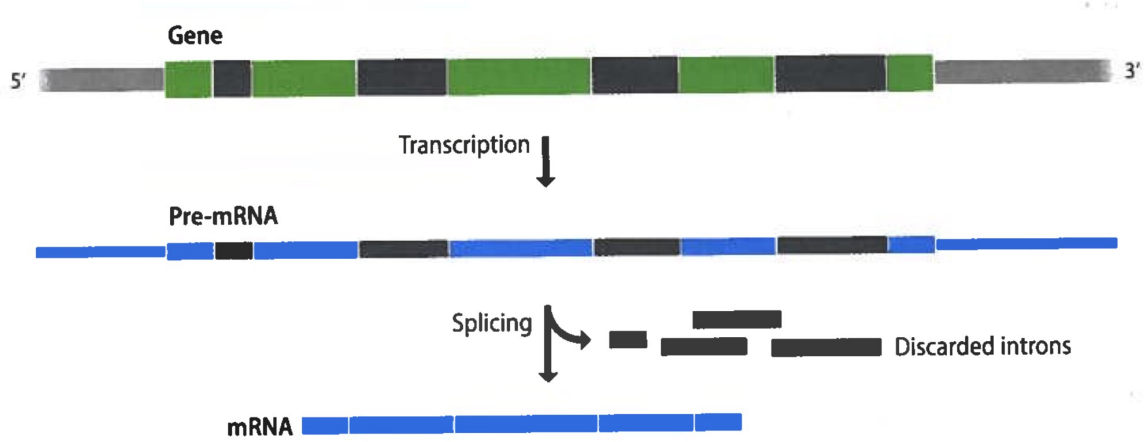
- Mining genome unveiled ca. 50 genes.
- Phylogenetic analyses and protein reconstructions allowed identifying that most of these genes are paralogues, most likely evolving through multiple rounds of whole genome doubling.
- Promoter sequence analysis suggest that they serve different purposes.

Melton *et al.* (2021)

WHAT IS "IN" A GENOME?

Protein-coding genes

- RNA-Seq helps identifying coding genes in genome sequence, but this process is made difficult due to splicing in eukaryotes.



What is "in" a genome?

Genes coding for RNAs

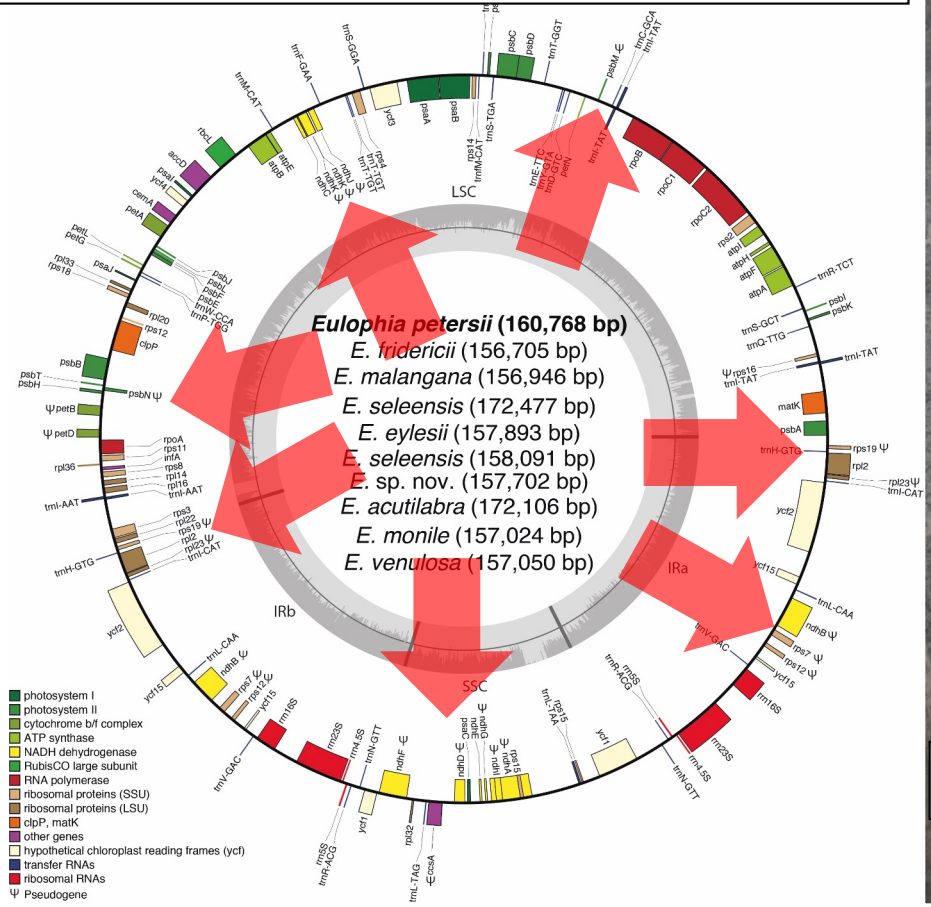
- About 3000 genes coding for RNAs (exclusive of the mRNAs translated to proteins).
- It is becoming clear that the RNA-ome is much richer than previously suspected.
- Except for RNAs involved in the machinery of protein synthesis (e.g. tRNAs and the ribosome itself), most non-coding RNAs are involved in gene expression (e.g., miRNA).

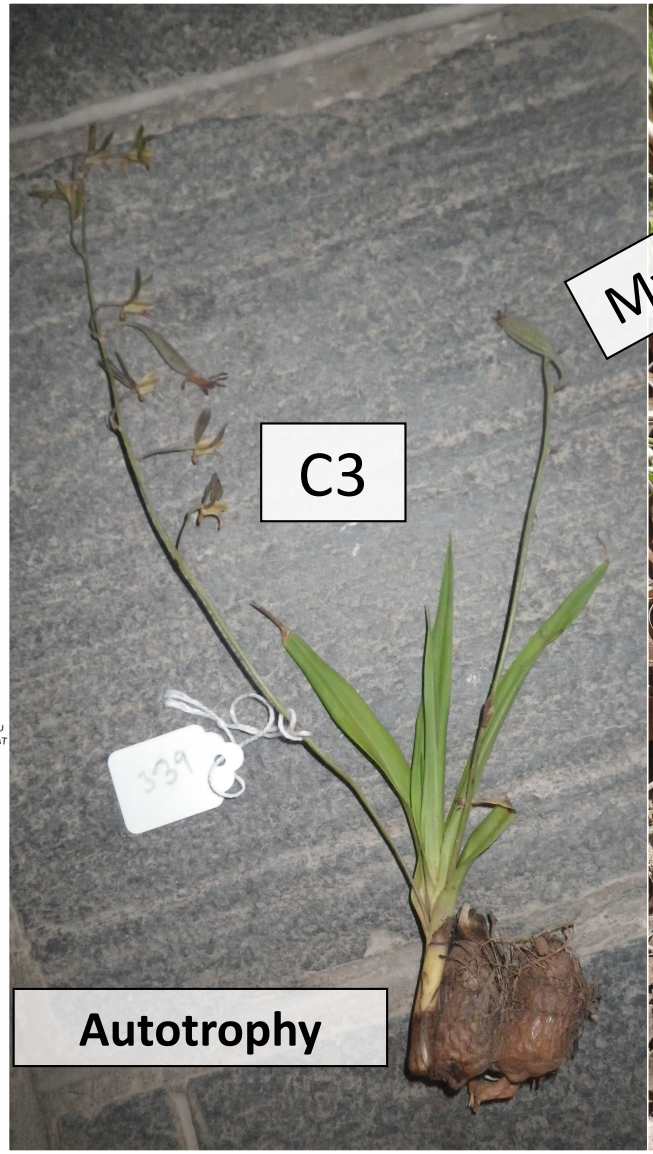
WHAT IS "IN" A GENOME?

Pseudogenes

- Degenerated genes that have mutated so far from their original sequences that the polypeptide sequence they encode will not be functional.
- First step before “losing” genes that are not anymore under selection.

Pseudogenes associated with shifts from autotrophy to heterotrophy in orchids





What is "in" a genome?

Repetitive elements (mostly of unknown function)

- About 50% of human genome consists of repeats.
- In humans, **long and short interspersed elements account for 21% and 13%** of the genome.
- Even more highly repeated sequences – **minisatellites** (about 10-100 base pairs) and **microsatellites** (mostly 2-4 base pairs) – may appear in hundreds of thousands of copies **totaling up to 15%** of the genome. These regions are widely used for population genetic analyzes.
- **Polyploidy and the accumulation of repetitive DNA sequences** (often derived from retrotransposons) **are the main factors driving the diversification of genome sizes.**

Liao *et al.* (2023)

What is "in" a genome?

Repetitive elements (mostly of unknown function)

- The **sagebrush** genome (*Artemisia tridentata*; Asteraceae) consists of 77.99% of repeats.
- This genome is the most highly repetitive plant genomes to be sequenced and assembled.

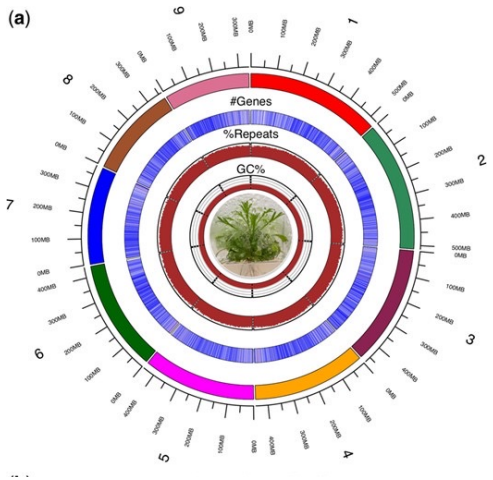


Table 1. Summary statistics for the 9 pseudo-chromosomal scaffolds within the IDT3 "G1_b2" genome assembly.

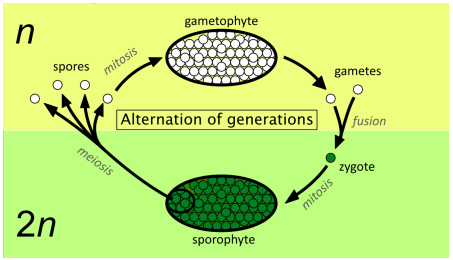
Scaffold	Length in Gb (% of assembly)	Protein coding genes	Total gene length in Gb (% of assembly)	Repeat occurrences	Repeat length total in Gb (% of assembly)
1	0.528 (12.58)	5,869	0.018 (3.49)	709,220	0.444 (84.00)
2	0.514 (12.23)	5,153	0.015 (2.99)	682,886	0.443 (86.21)
3	0.472 (11.24)	4,781	0.015 (3.15)	624,680	0.406 (86.04)
4	0.446 (10.62)	4,707	0.015 (3.33)	591,412	0.378 (84.73)
5	0.445 (10.59)	4,951	0.017 (3.73)	591,818	0.371 (83.43)
6	0.439 (10.46)	4,358	0.013 (3.04)	580,217	0.379 (86.38)
7	0.385 (9.18)	4,096	0.013 (3.30)	513,867	0.330 (85.52)
8	0.361 (8.61)	3,520	0.011 (3.03)	480,240	0.311 (86.11)
9	0.338 (8.06)	3,430	0.011 (3.11)	446,444	0.295 (87.12)
Total	3.929 (93.58)	40,865	0.128 (3.25)	5,220,784	3.356464852 (85.43)

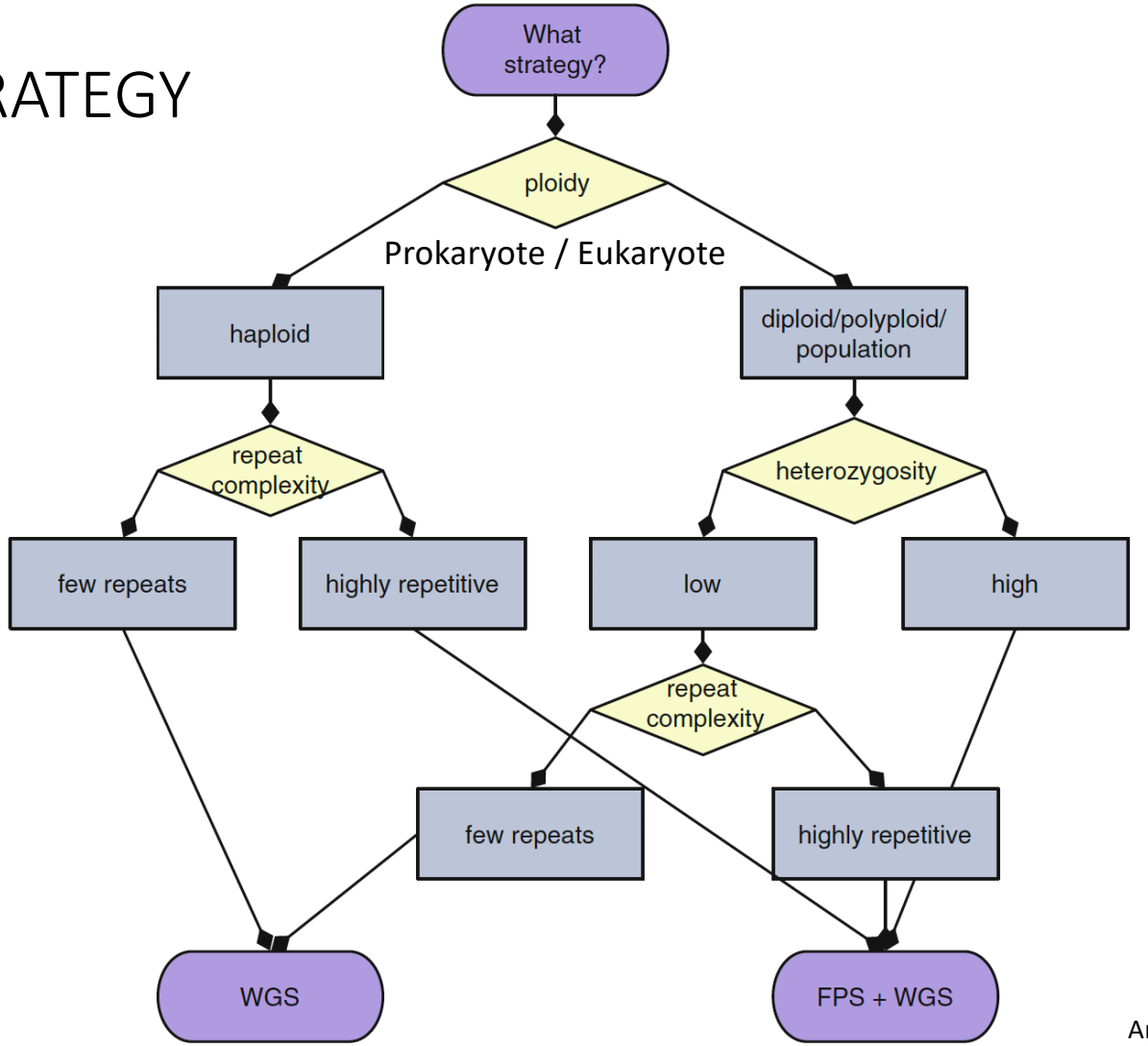
Melton *et al.* (2022)

SEQUENCING STRATEGY

Other aspects to consider:

- Test species monophyly.
- If species is a polyploid complex, sample diploids.
- If heterozygosity is still too high, consider sampling gametophyte (n).



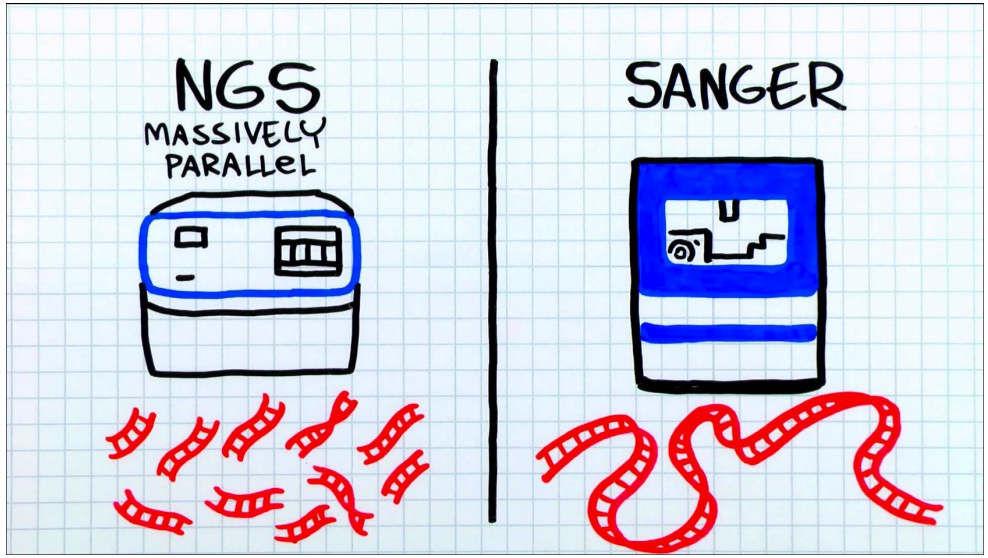


FPS: fosmid pool sequencing (cloning)
WGS: whole-genome shotgun sequencing

Aransay (2016)

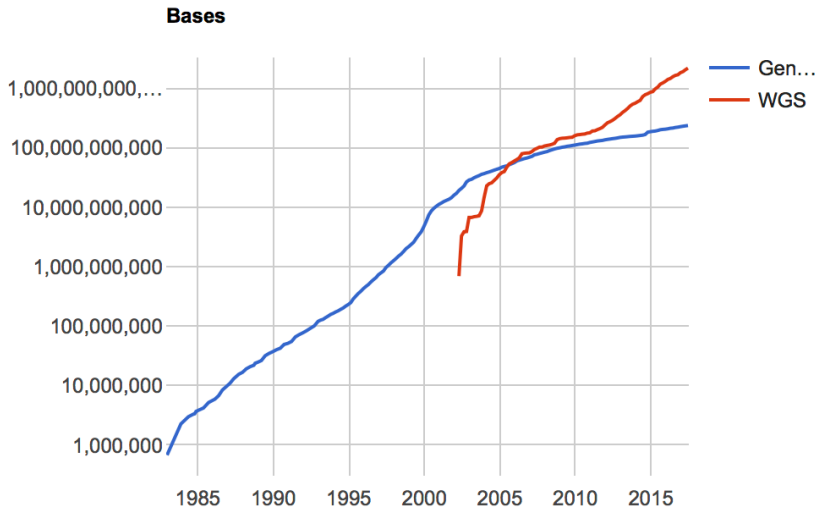
WHAT IS NGS?

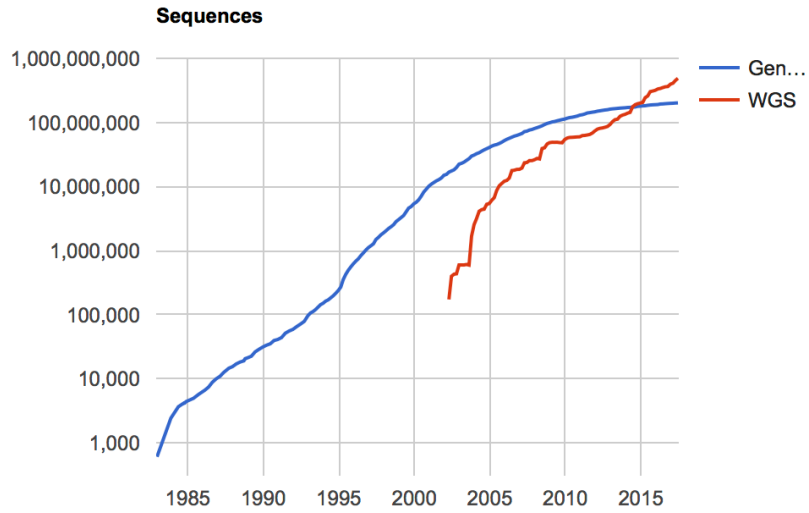
- **NGS** is the term applied to methods enabling **thousands of millions of DNA fragments to be sequenced in parallel in a single experiment**.
- NGS outperformed Sanger sequencing, which was able to sequence only individual DNA fragments, each fragment obtained by a different PCR.



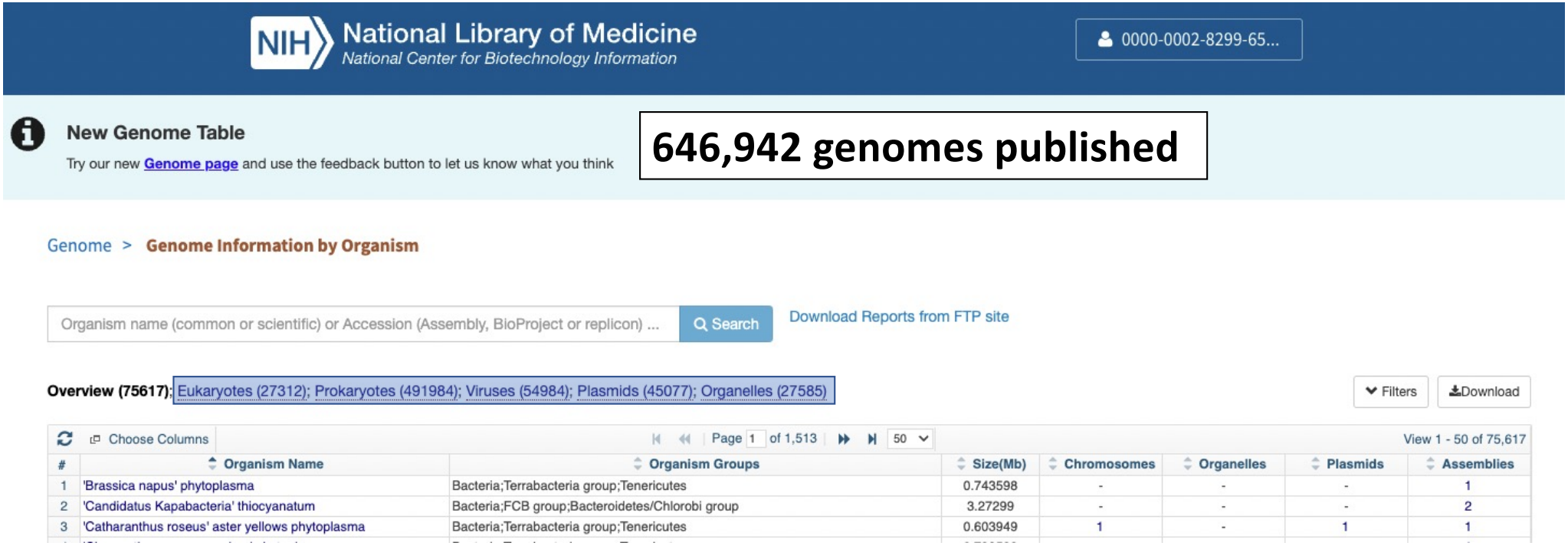
WHAT IS NGS?

- NGS methods enable the vast amounts of data needed to assemble an entire genome sequence to be obtained much more rapidly and cheaply than the Sanger sequencing method.
- The emergence of NGS technologies (between 2005-2011) had an impact on the availability of DNA sequences on GenBank.





How many genomes are published?



#	Organism Name	Organism Groups	Size(Mb)	Chromosomes	Organelles	Plasmids	Assemblies
1	'Brassica napus' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.743598	-	-	-	1
2	'Candidatus Kapabacteria' thiocyanatum	Bacteria;FCB group;Bacteroidetes/Chlorobi group	3.27299	-	-	-	2
3	'Catharanthus roseus' aster yellows phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.603949	1	-	1	1

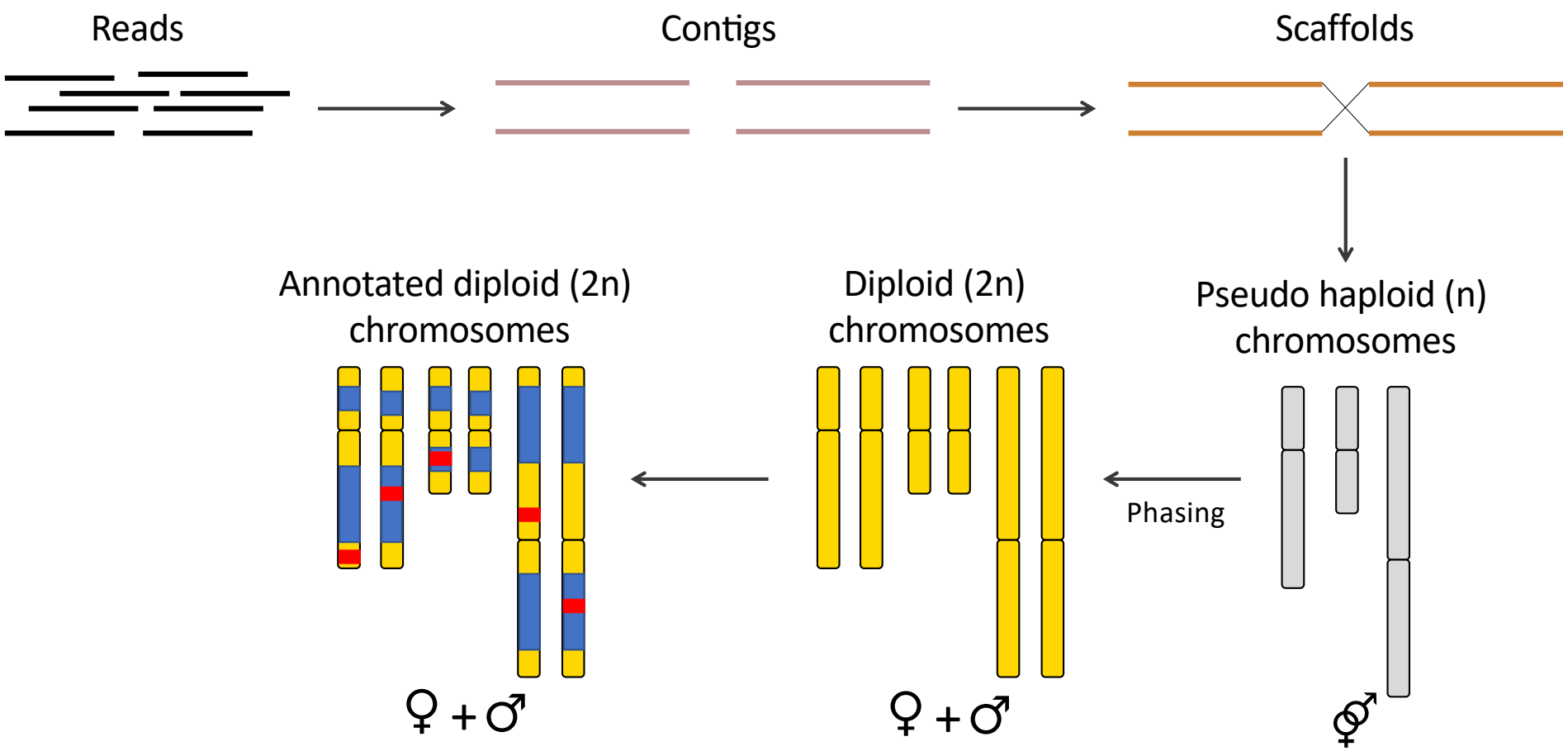
BUT NOT ALL GENOMES ARE OF THE SAME QUALITY…

- Not all genomes are assembled and annotated at the same depth.
- Genome assemblies could be submitted as:
  i. Complete,
  ii. Chromosome,
  iii. Scaffold,
  iv. Contig.

Quality gradient

- The quality of the submitted genomes depend on the complexity of the genome (i.e. amount of repeated DNA, level of ploidy), but also on the aims of the researchers submitting the genome (why producing a fully annotated genome when we do not need it?)

Genome assembly and annotation workflow



WHAT DOES IT MEAN TO SEQUENCE A GENOME?

- Ideally, **a draft genome would represent the complete nucleotide base sequence for all chromosomes** in the species of interest, a 'physical map' of its genetic content.
- There are four major complications with the concept of a "genome sequence":
  1. There is not one true sequence for a species due to individual SNPs.
  2. It is essentially impossible to sequence and assemble all nucleotides in the genome due to repetitive DNA.
  3. There will always be some degree of error in the characterized genome sequence: sequencing and assembly errors.
  4. Every genome assembly is the result of a series of assembly heuristics and should accordingly be treated as a working hypothesis.

https://www.science.org/toc/science/376/6588

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33], Sébastien Boisvert[10,11], Jarrod A Chapman[20], Guillaume Chapuis[7,9], Rayan Chikhi[7,9], Hamidreza Chitsaz[6], Wen-Chi Chou[14,16], Jacques Corbeil[10,13], Cristian Del Fabbro[17], T Roderick Docking[33], Richard Durbin[34], Dent Earl[40],

Conclusions: Many current genome assemblers produced useful assemblies, containing a significant representation of their genes and overall genome structure. However, the high degree of variability between the entries suggests that there is still much room for improvement in the field of genome assembly and that approaches which work well in assembling the genome of one species may not necessarily work well for another.

This field, especially bioinformatics, is exploding and methods for *de novo* genome assembly are improving every day!

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

RESEARCH Open Access

Assembl genom

Keith R Bradnar
Sébastien Boisv
Wen-Chi Chou[1]

Conclusions: M
their genes and
there is still mu
assembling the

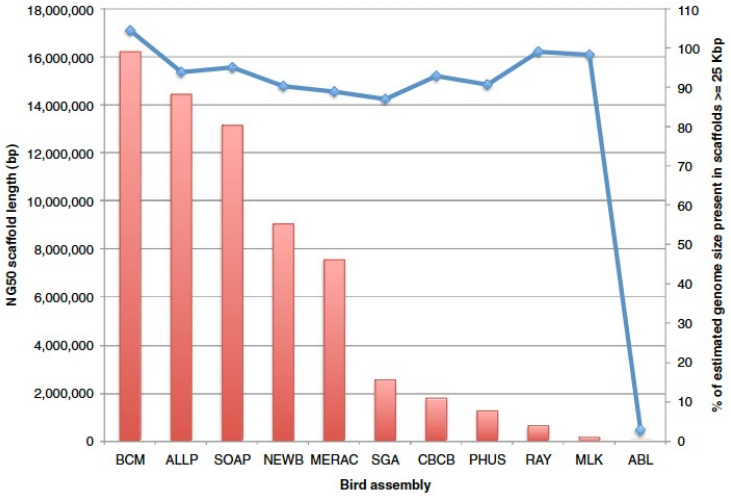


Figure 4 NG50 scaffold length distribution in bird assemblies and the fraction of the bird genome represented by gene-sized scaffolds. Primary Y-axis (red) shows NG50 scaffold length for bird assemblies: the scaffold length that captures 50% of the estimated genome size (~1.2 Gbp). Secondary Y-axis (blue) shows percentage of estimated genome size that is represented by scaffolds ≥25 Kbp (the average length of a vertebrate gene).

ods of
ies

Birol[33],
saz[6],
n[34], Dent Earl[40],

resentation of
uggests that
ork well in

This field, especially bioinformatics, is exploding and methods for *de novo* genome assembly are improving every day!