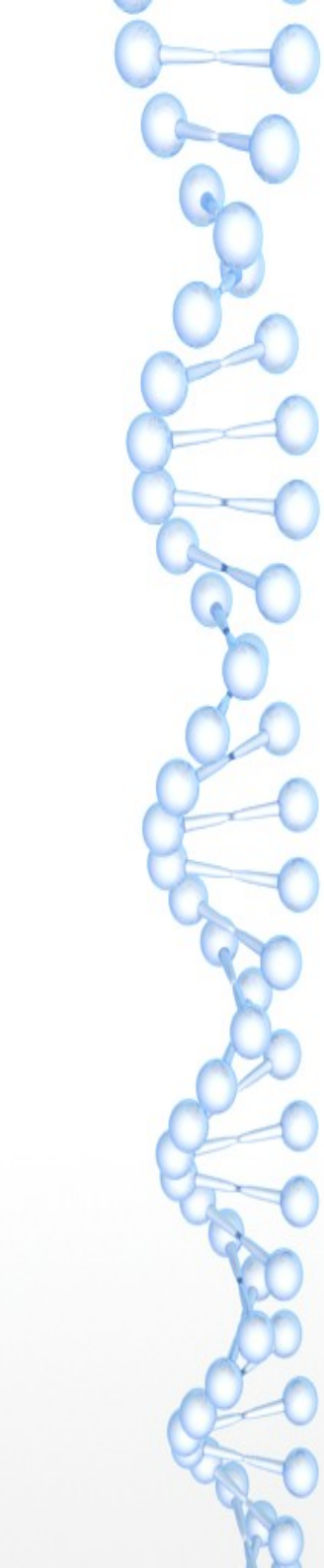
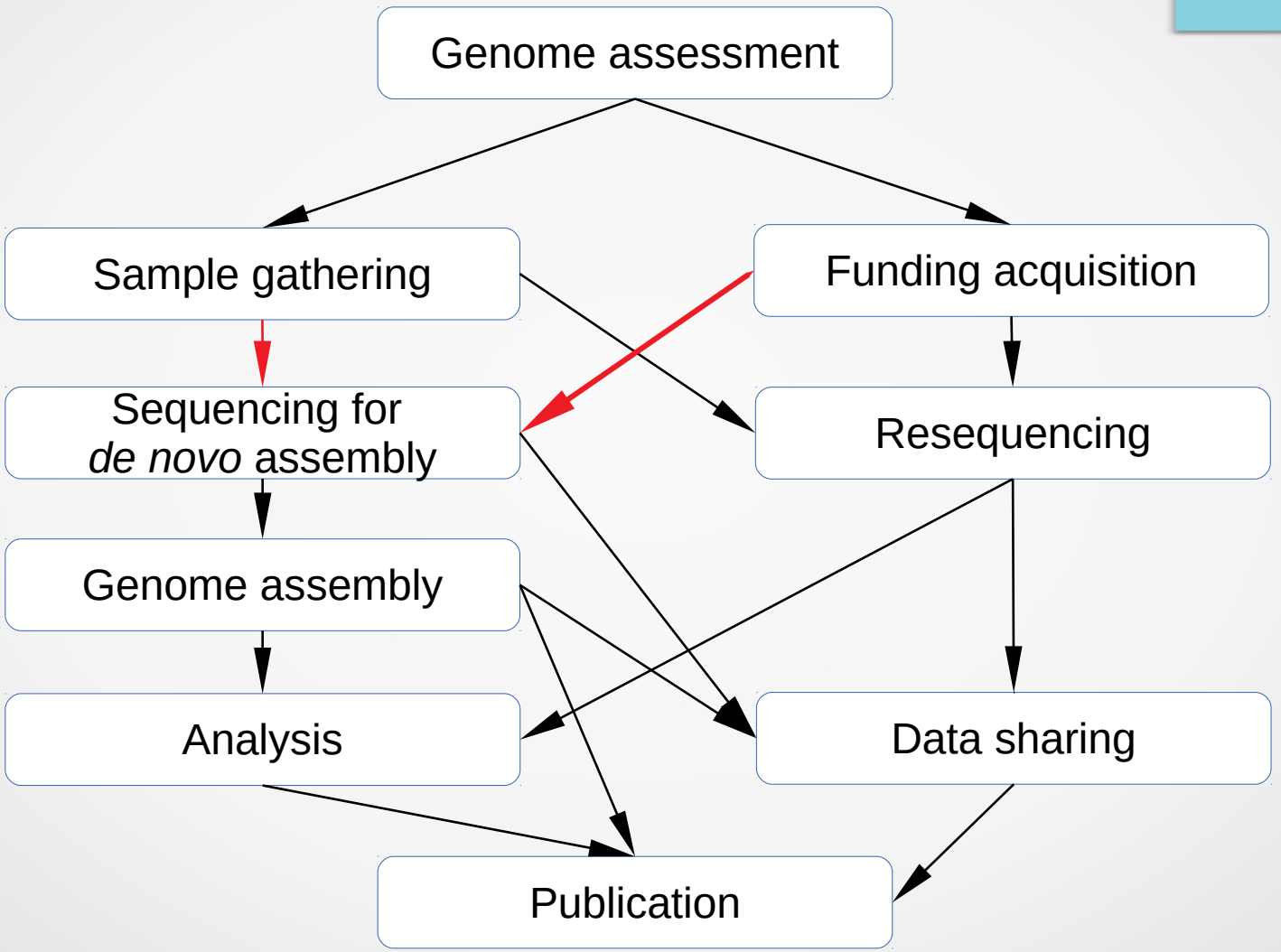
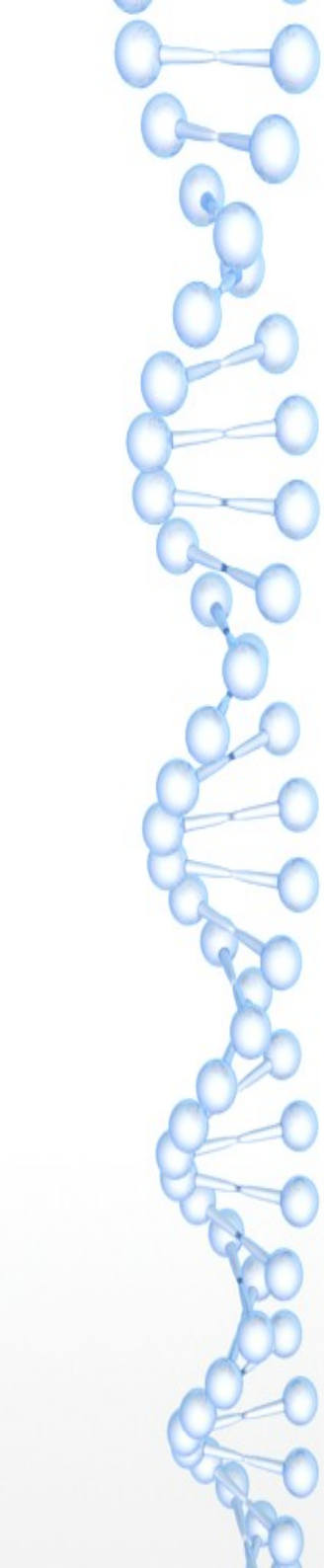


III. Genome Projects

Stages of the project





III. Genome Projects

Genome assessment

Questions you need to answer before starting project

I. How big is the genome?

II. Does it have heterochromatin?

III. Is it diploid or polyploid?

IV. Is it highly heterozygous or not?

V. Is your species hybrid or not?

VI. Are the genome rearrangements widespread in your species?

Genome size of your species

Ways to assess

I. Check literature and databases

for animals	https://www.genomesize.com
for plants	https://cvalues.science.kew.org
for fungi	http://www.zbi.ee/fungal-genomesize/

II. Estimate using flow cytometry

- require cell line of your species
- require reference (at least two) species with known genome size
- require cytometer and cytogenetisist

III. Estimate from reads using k-mer distribution

- require preliminary sequencing

C-value (haploid genome size) might be in picograms and Mbp.

1 pg = 978 Mbp

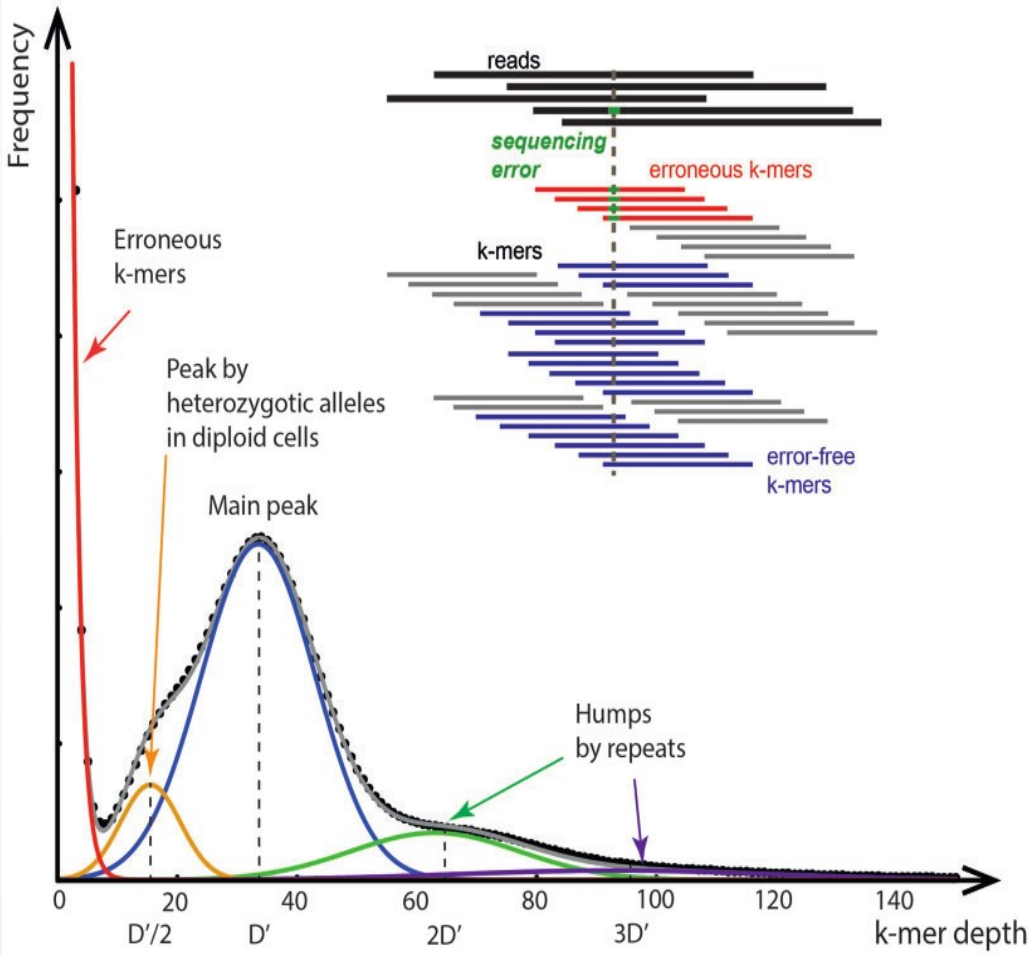
Genome size estimation from reads (1)

K-mer based approach:

I. Count k-mers and create database
II. Count histogram
III. Assess genome size from histogram

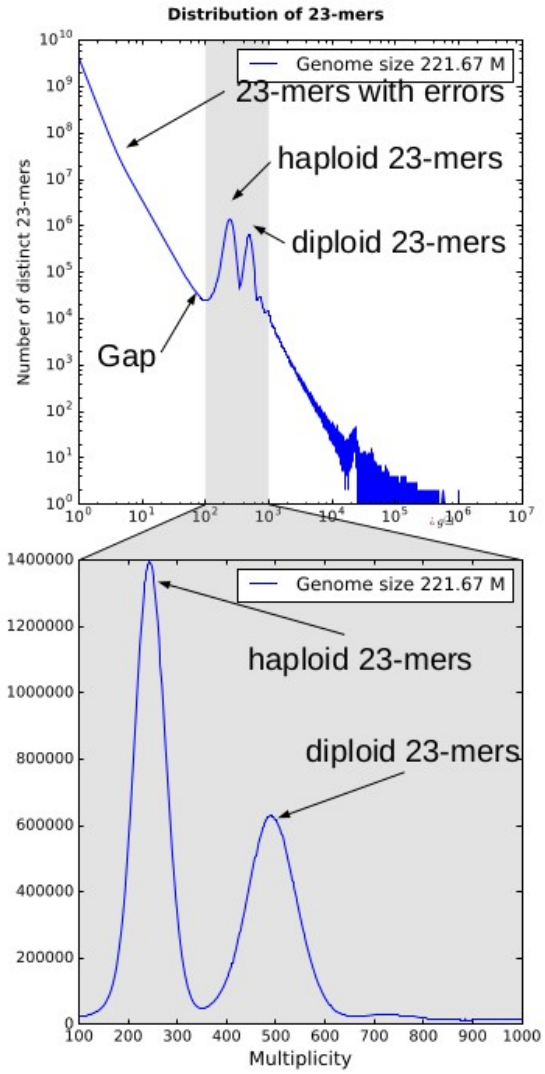
Tools:

I. Jellyfish 2
II. Jellyfish 2
III. Genomescope 2, KrATER, etc



Sohn and Nam, 2016

Genome size estimation from reads (2)



23-mer distribution for PE reads of hybrid plant *B. divaricarpa*

Genome size estimation

$$\text{Genome size} = \frac{\sum_{i=g} N_i * m_i}{C}$$

$g-k-mer\ multiplicity\ at\ gap\ between\ peak\ of\ k-mers\ with\ errors\ and\ corresponding\ to\ unique\ part\ of\ genome$

$m_i-multiplicity\ of\ distinct\ k-mers$

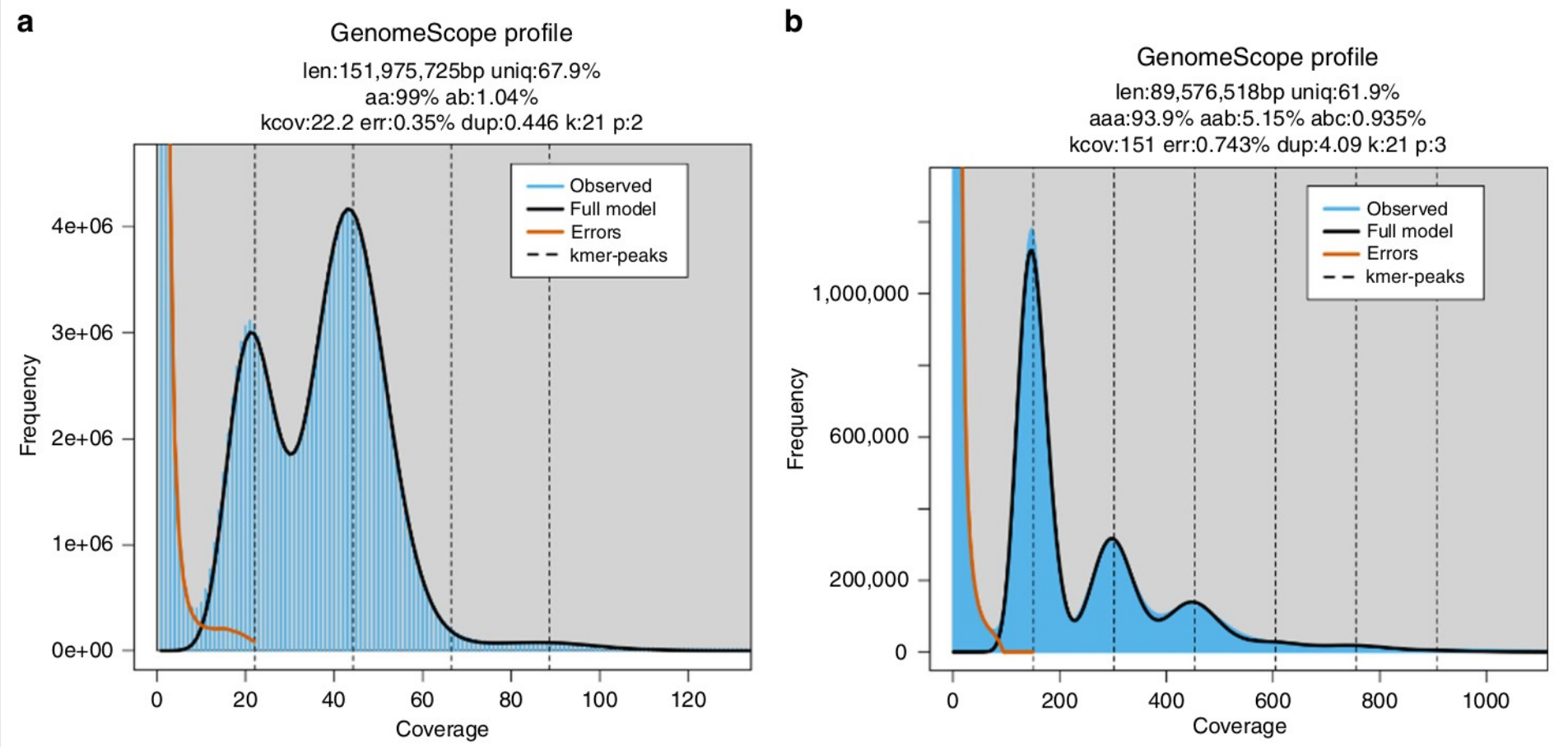
$N_i-number\ of\ distinct\ k-mers\ with\ m_i\ multiplicity$

$C-sample\ coverage\ estimated\ by\ mode\ of\ multiplicy\ of\ diploid\ k-mers$

Naive approach:

- direct assessment from histogram
- works always if there distinguishable peaks
- huge error

Genome size estimation from reads (3)

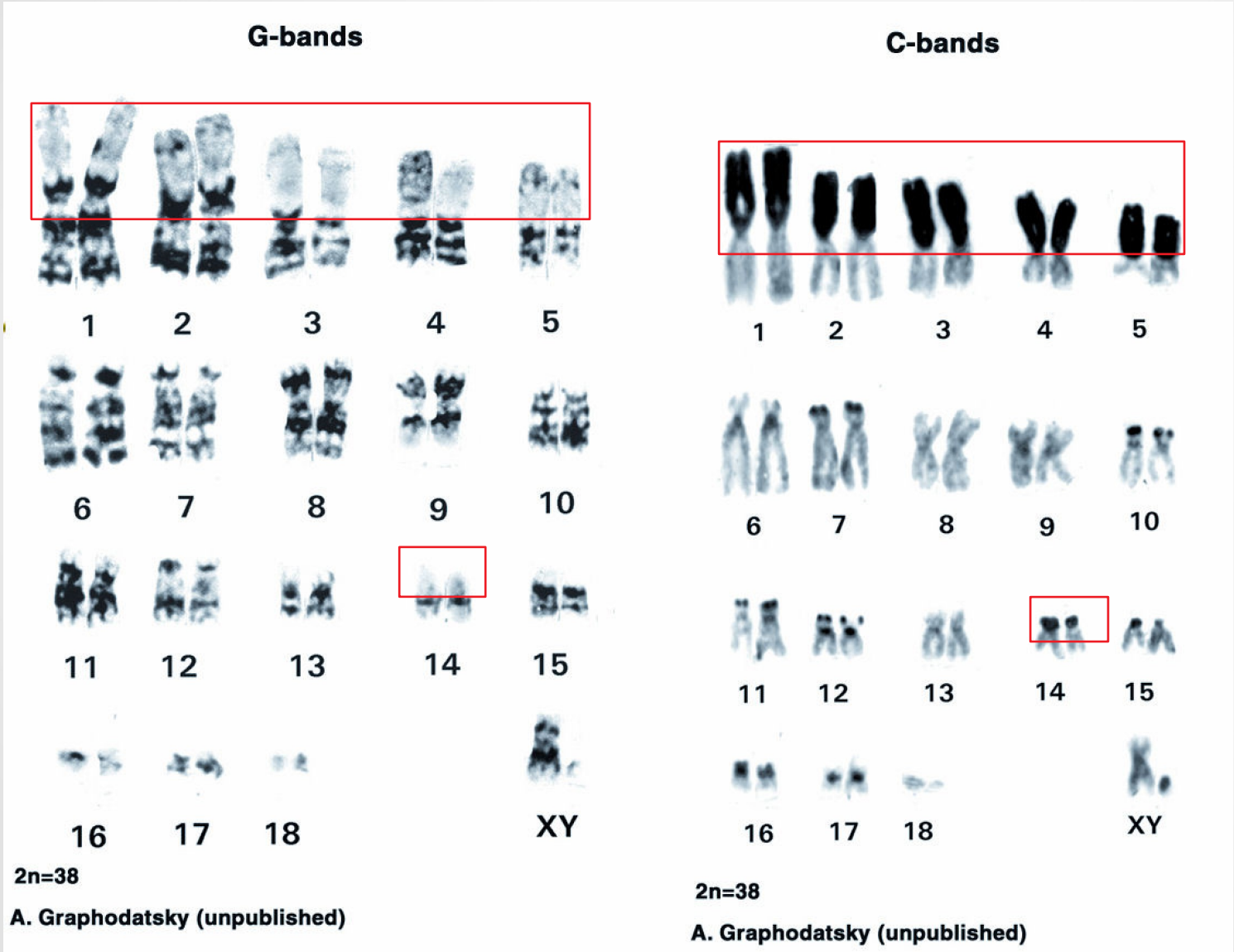


Modeling approach:

- assessment of genome size from histogram by fitting it with model (sum of several negative binominal distributions)
- works not always
- more precise

Ranallo-Benavidez et al, 2020

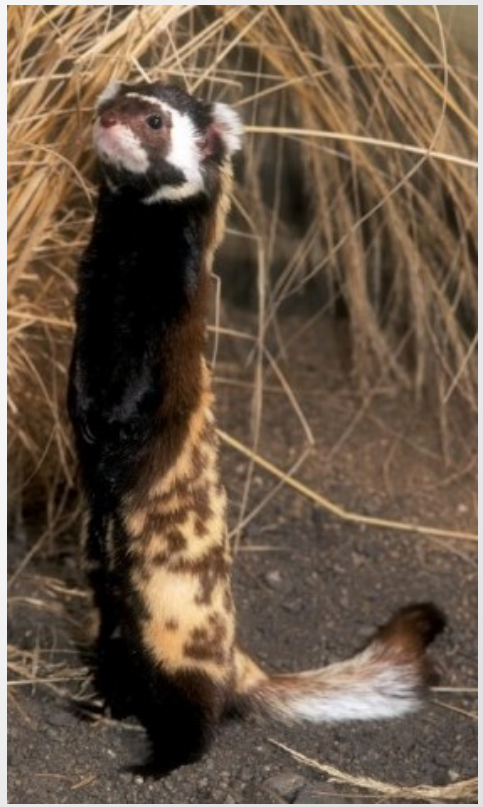
Heterochromatine and repeat content



Graphodatsky et al, 2020

Vormela peregusna

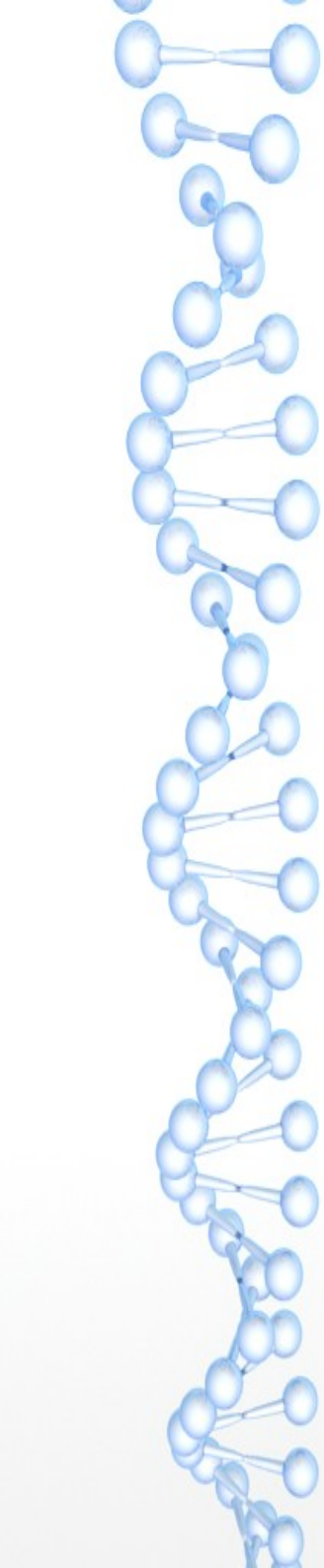
Marbled polecat



~ 6.4 Gbp (!)

Widespread chromosomal rearrangements

Graphodatsky et al, 2020



III. Genome Projects

Samples

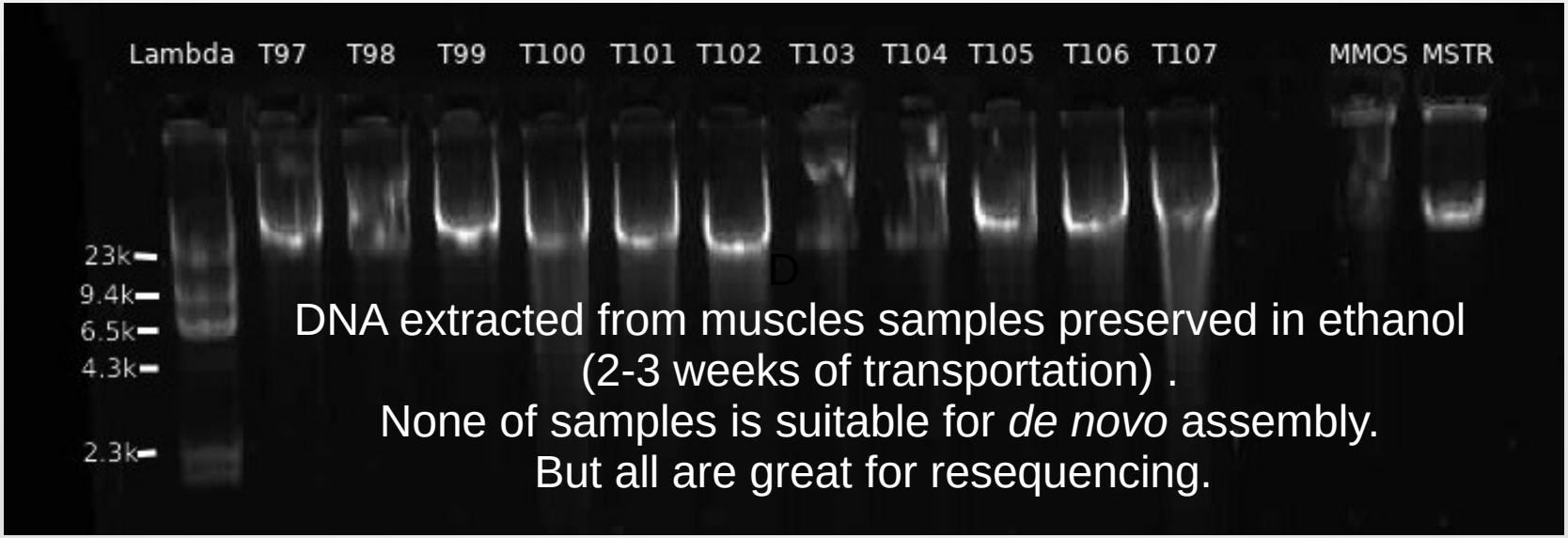
Fragment size of extracted DNA

HMW DNA
high molecular weight DNA ~ 50 - 300 kbp fragments

UHMW DNA
ultrahigh molecular weight DNA ~300+ kbp fragments

Modern approaches for *de novo* assembly require at least HMW DNA.

Not all sample types and DNA extraction methods could produce HMW DNA!



DNA extracted from muscles samples preserved in ethanol (2-3 weeks of transportation) .
None of samples is suitable for *de novo* assembly.
But all are great for resequencing.

DNA sources

		Fragments, bp
• Cell lines	Suitable for *de novo* assembly	up to 500000

• Fresh tissue samples
  • blood
  • biopsy
  • necropsy
		up to 300000

• Preserved tissue samples		up to 50000 - 100000
• Secretions • saliva		~100 - 10000

• Museum samples
  • skins
  • bones
		~50-200

• Ancient samples • bones		~25-200

Primary cell lines

Advantages:

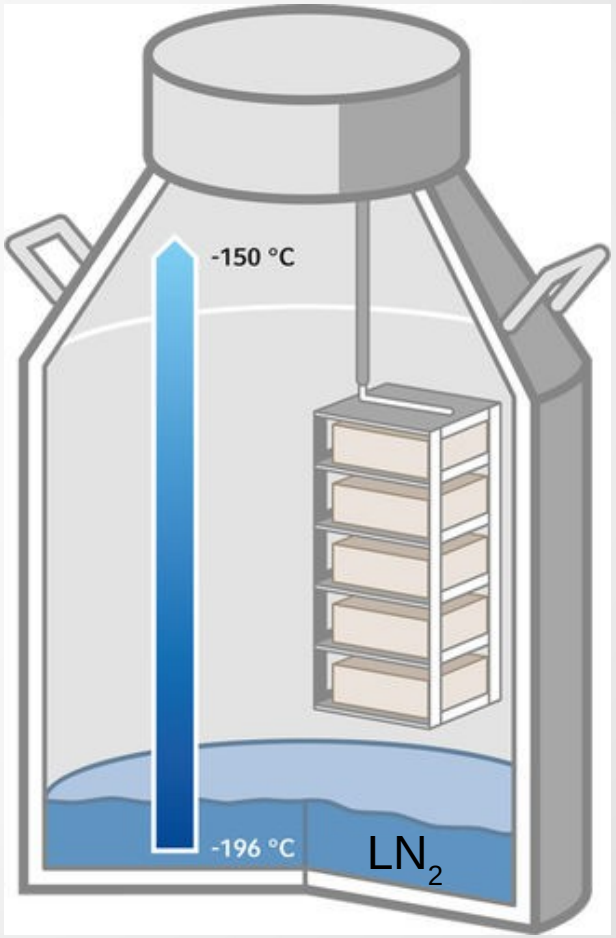
- Available living cells
- Allows cytogenetic experiments
- Source for DNA of excellent quality
- Could be stored for decades

Disdvantages:

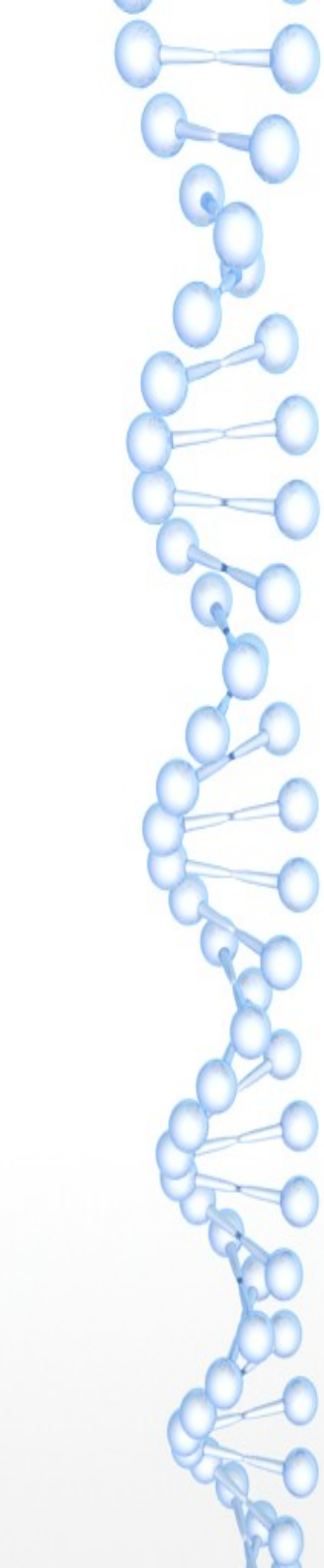
- Require cytogenetic lab and cytogenetisists
- Strict requirements for samples to establish cell line (cell must be alive)
- Storage in liquid nitrogen (LN_2)
- Will stop growing after specific number of divisions

Immortalized cell lines are not suitable for *de novo* genome assembly of new species!

cryogenic storage dewar



eppendorf.com



III. Genome Projects

Genome assembly

Major definitions related to genome assemblies

Read

small fragment generated by sequencing

Contig

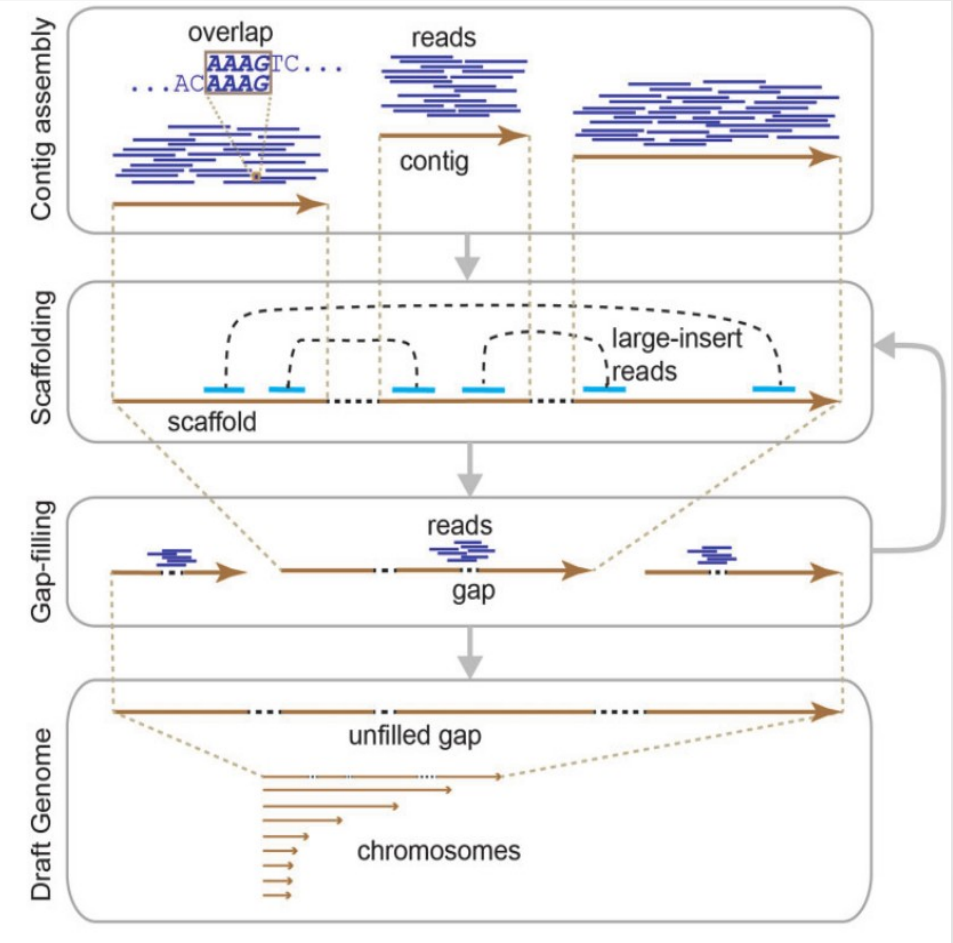
ungapped sequence assembled from reads

Scaffold

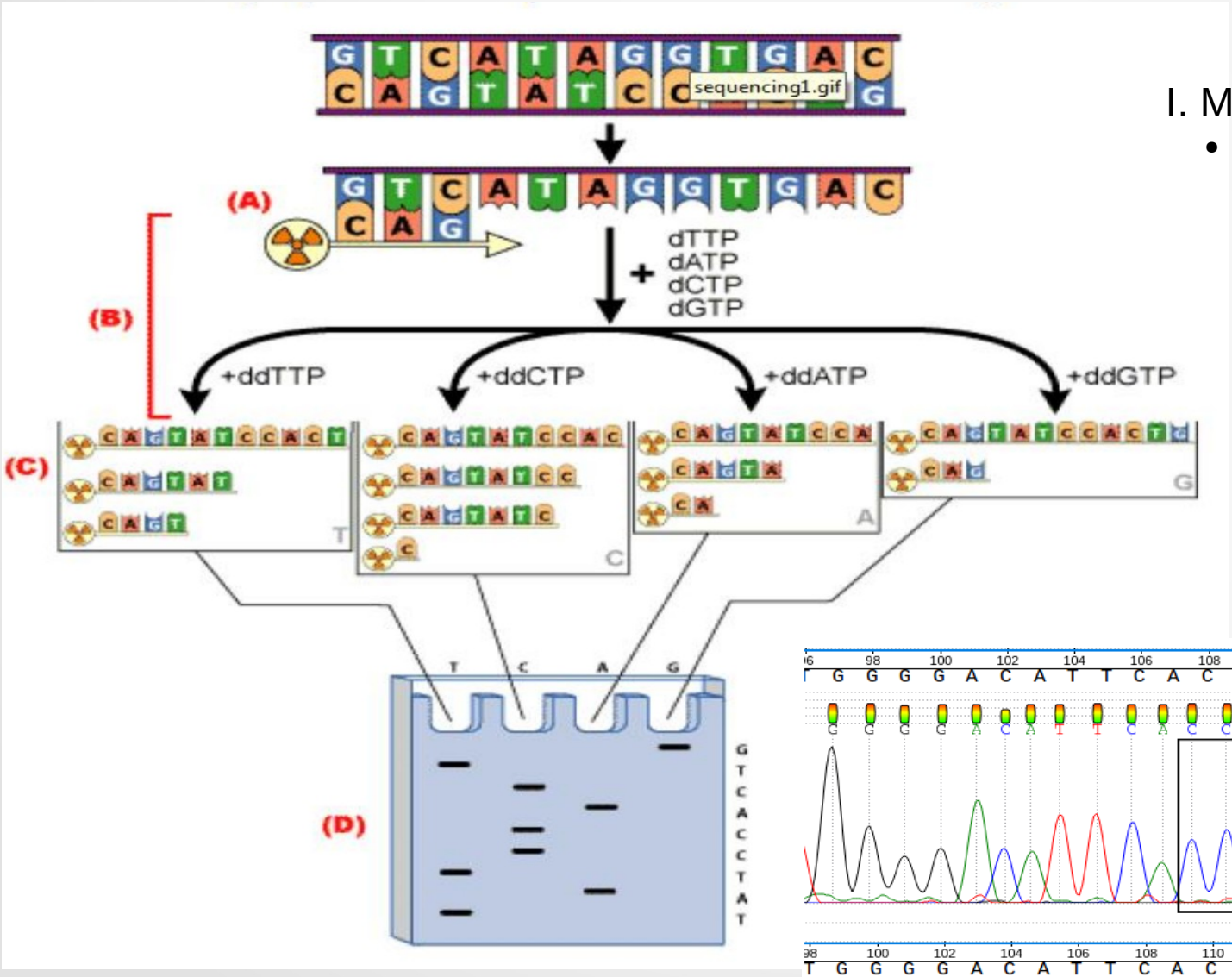
sequence with gaps, generated from contigs (set of oriented contigs)

C-scaffold

chromosome (or close to) scaffold, representing whole chromosome or its significant part

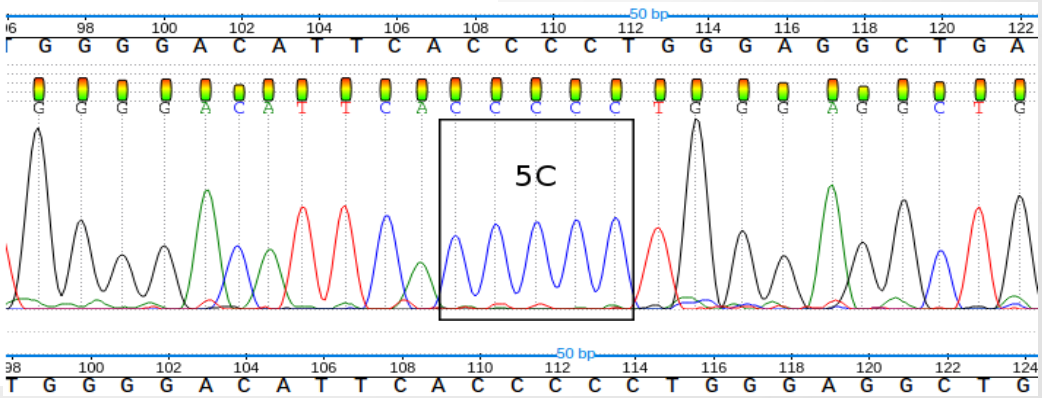


Sequencing technologies: first generation



I. Maxam-Gilbert sequencing
- R.I.P

II. Sanger sequencing
- still alive
- used in validation and preliminary analysis
- reads up to 800 bp



Sequencing technologies: second (next) generation

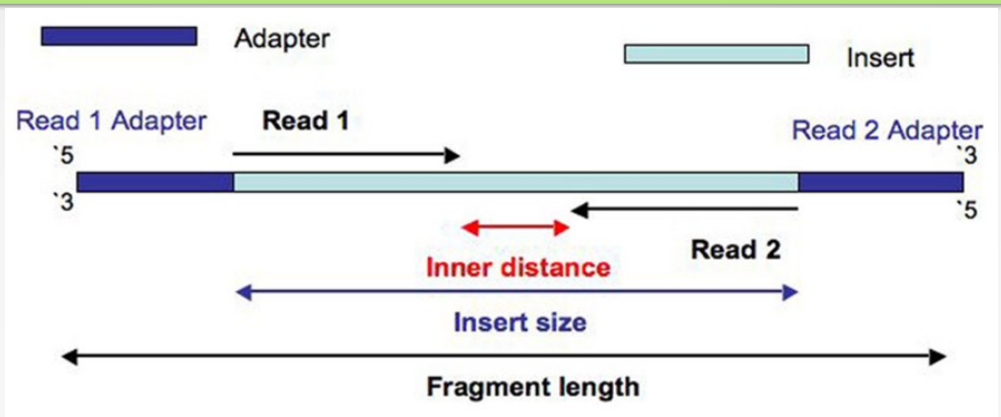
Platform	Max read length, bp	
• SOLID • R.I.P.	35	
• Roche 454 (pyrosequensing) • R.I.P.	400	Issues with sequencing of homopolymers
• IonTorrent • R.I.P. for *de novo* assembly	400	
• Illumina	250	
• MGIseq	300	

All NGS platforms are based on sequencing-by-synthesis (SBS) and can't sequence a single molecule!

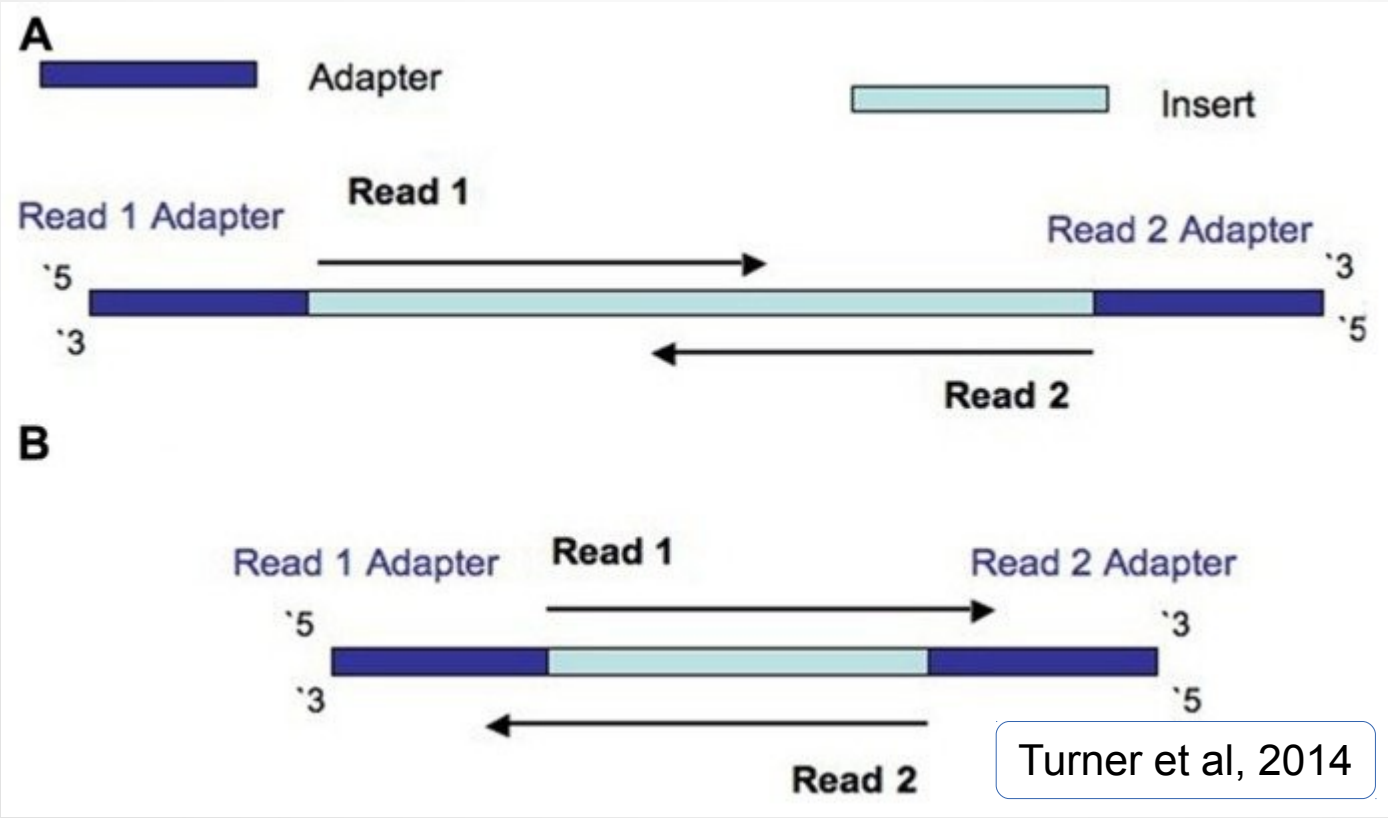
Major differences between platforms are related to what is detected during synthesis and how amplification is performed.

Read length and insert size

IS - insert size
RL - read length



IS > 2 * RL

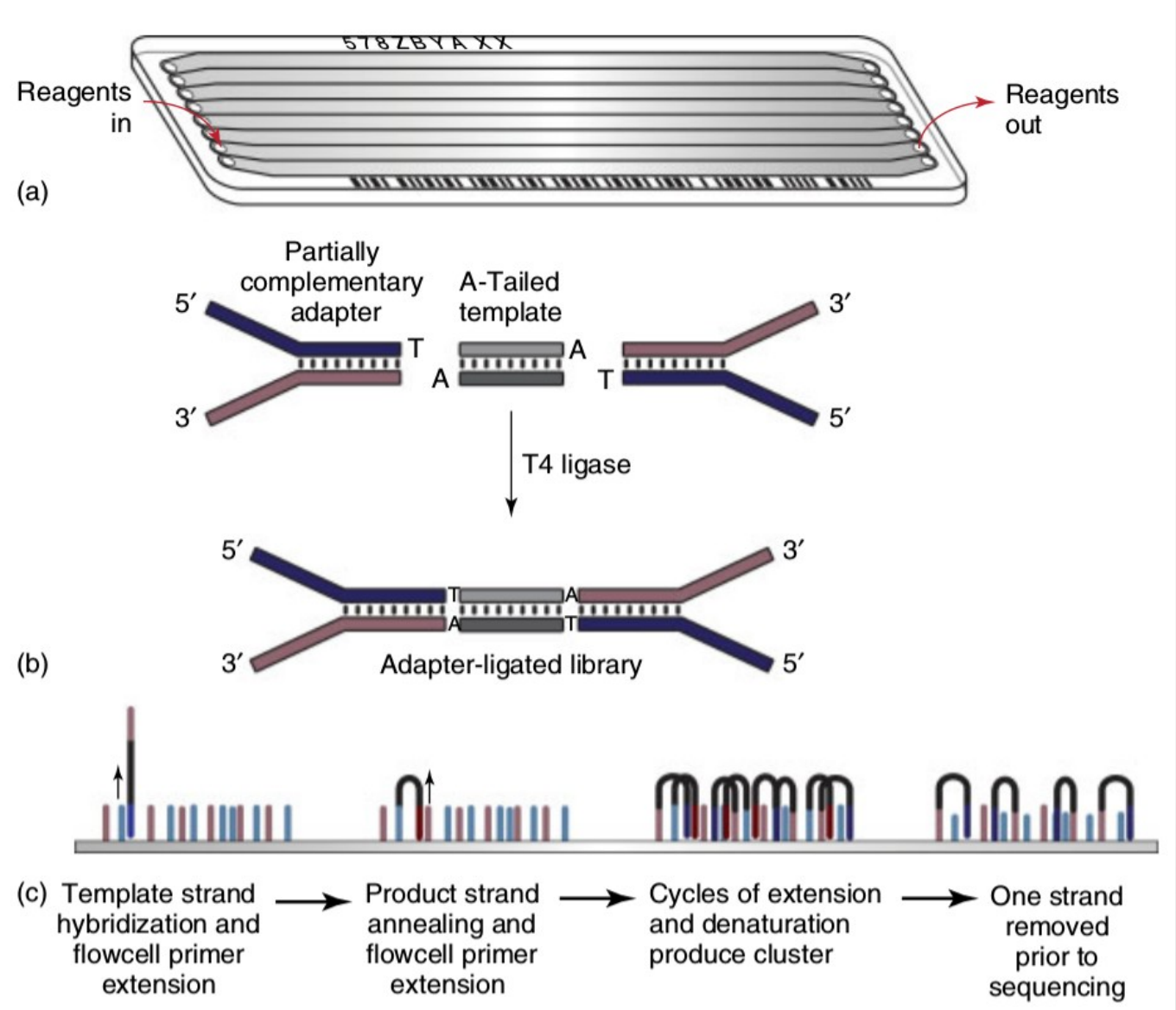


RL < IS < 2 *RL

IS < RL

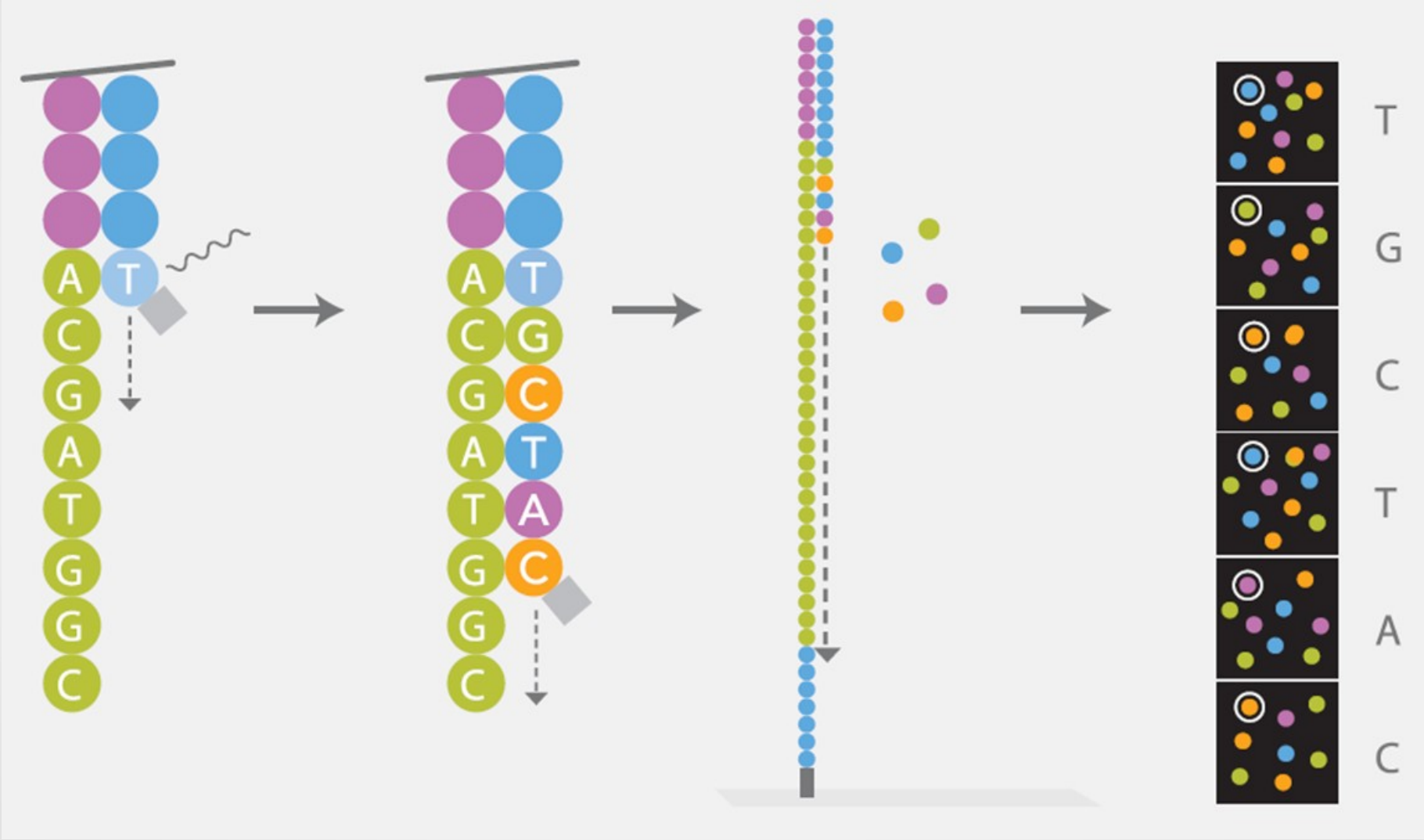
Turner et al, 2014

Illumina platform



Turner et al, 2014

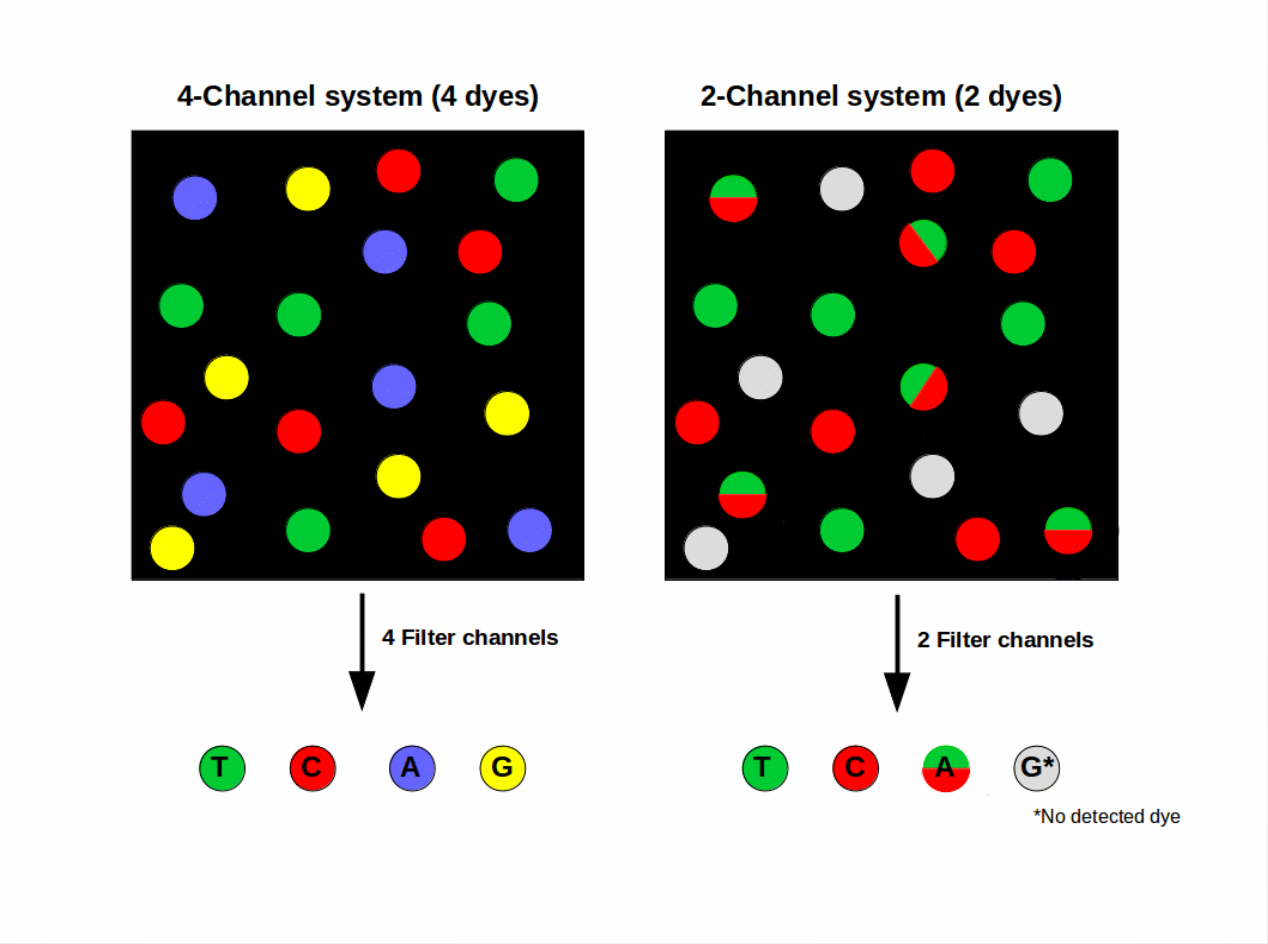
Sequencing-by-synthesis on Illumina platform



Ansorge et al, 2009

Illumina platform is heterogenious

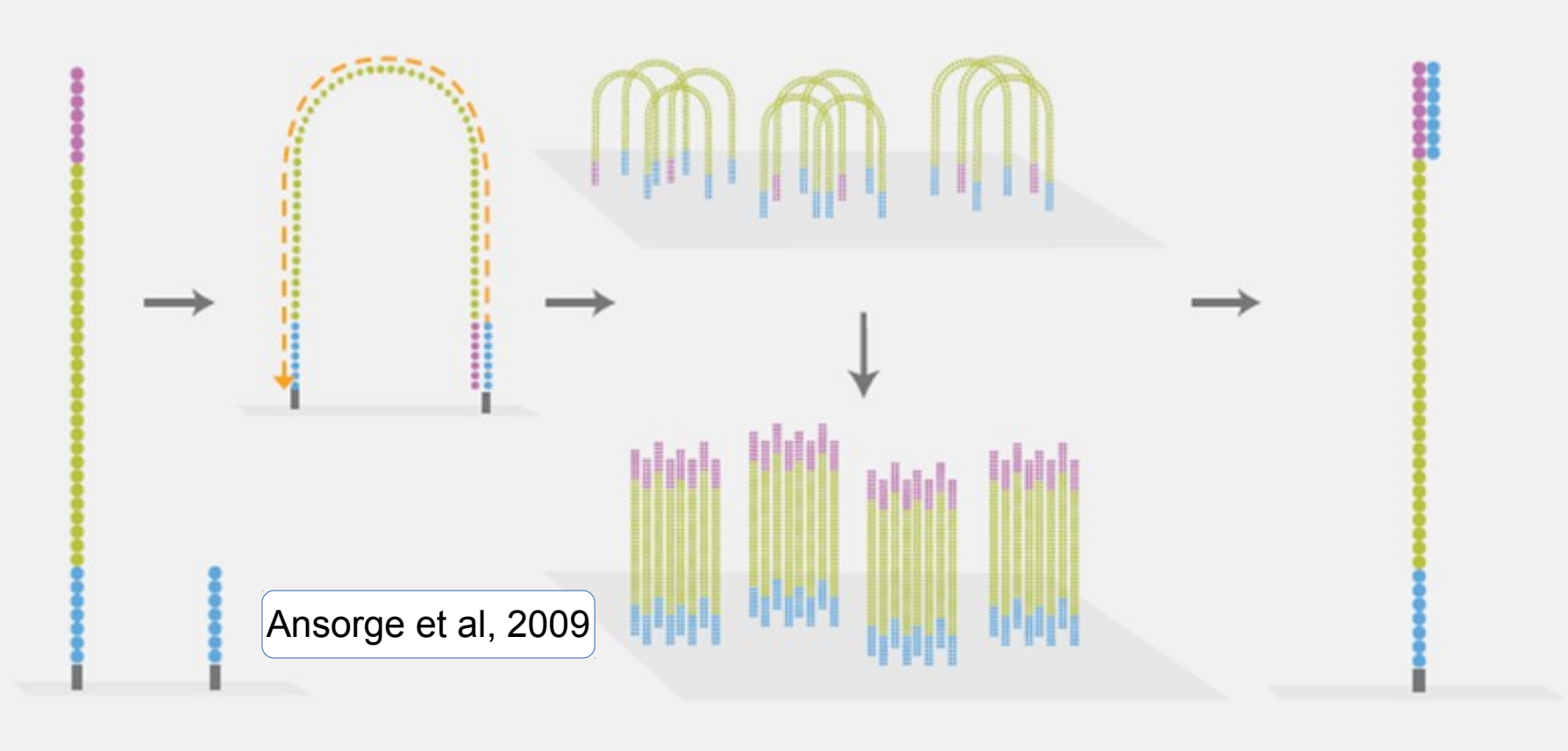
GAII
HiSeq2000
HiSeq2500
HiSeq3000
HiSeq4000



NovaSeq
NextSeq

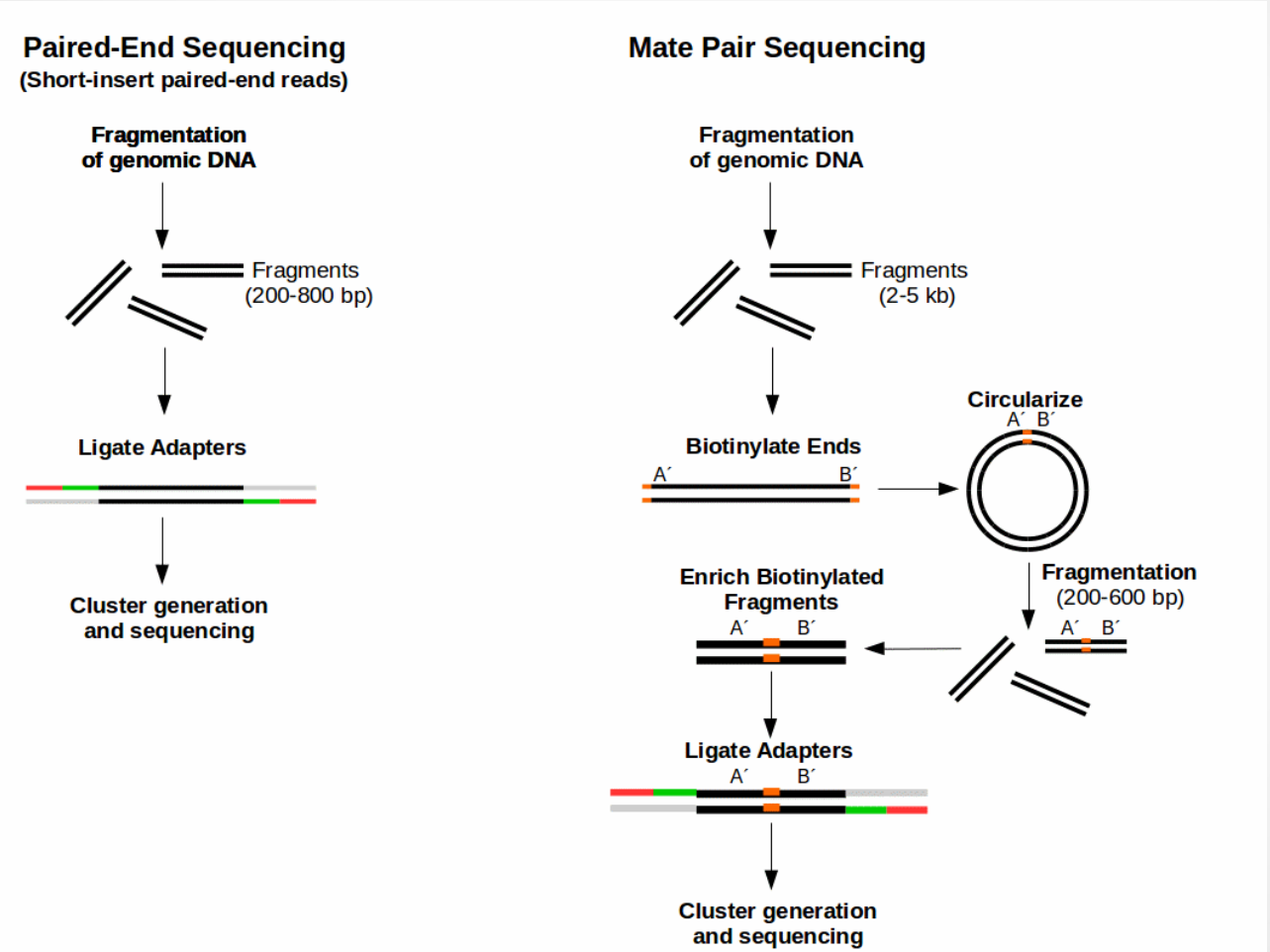
www.ecseq.com

Bridge amplification on illumina platform



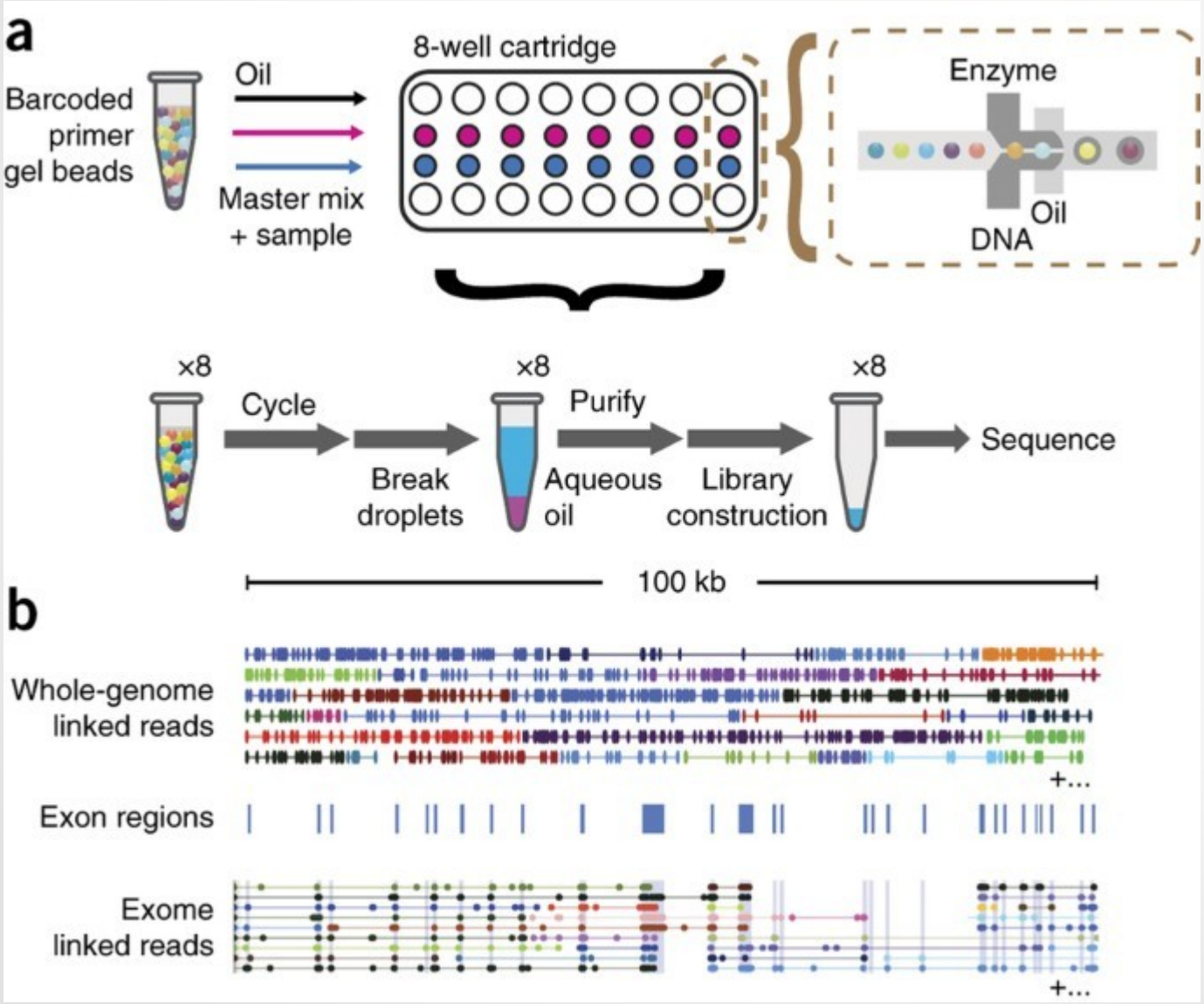
Bridge amplification doesn't work for fragments longer ~ 1500 bp.
It is a maximal threshold for insert size (IS), and it is difficult to achieve.
Commonly use IS are 250, **350** and 550

Special types of libraries: Mate pairs



www.ecseq.com

Special types of libraries: linked reads



Microfluidics based:

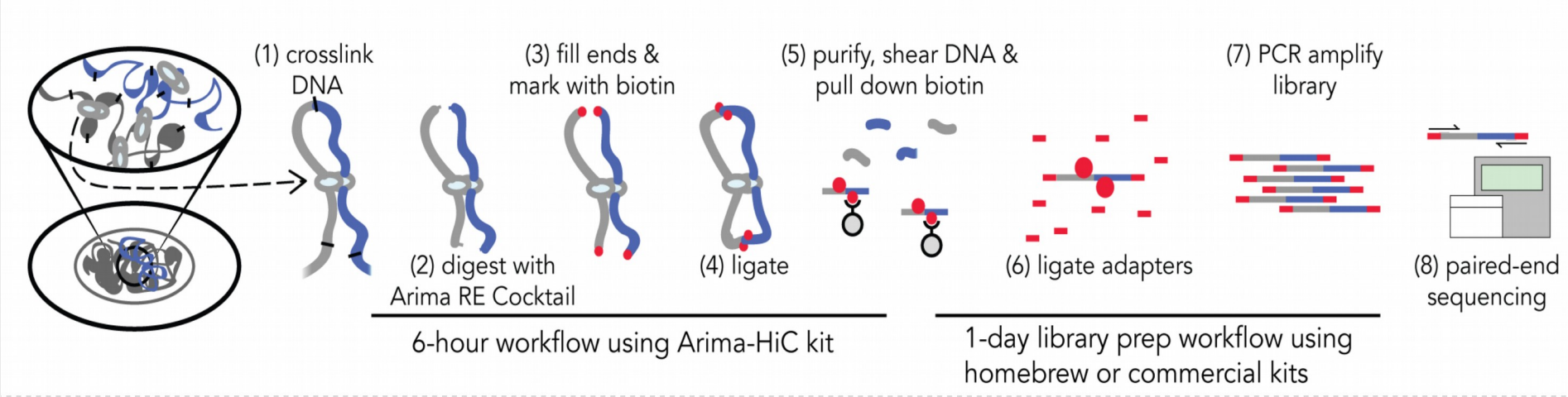
- 10X Genomics

Transposase-based

- Tell-Seq
- stLFR

Zheng et al, 2016

Special types of libraries: HiC



HiC-sequencing allows to scaffold even very fragmented draft genomes to the chromosome level.

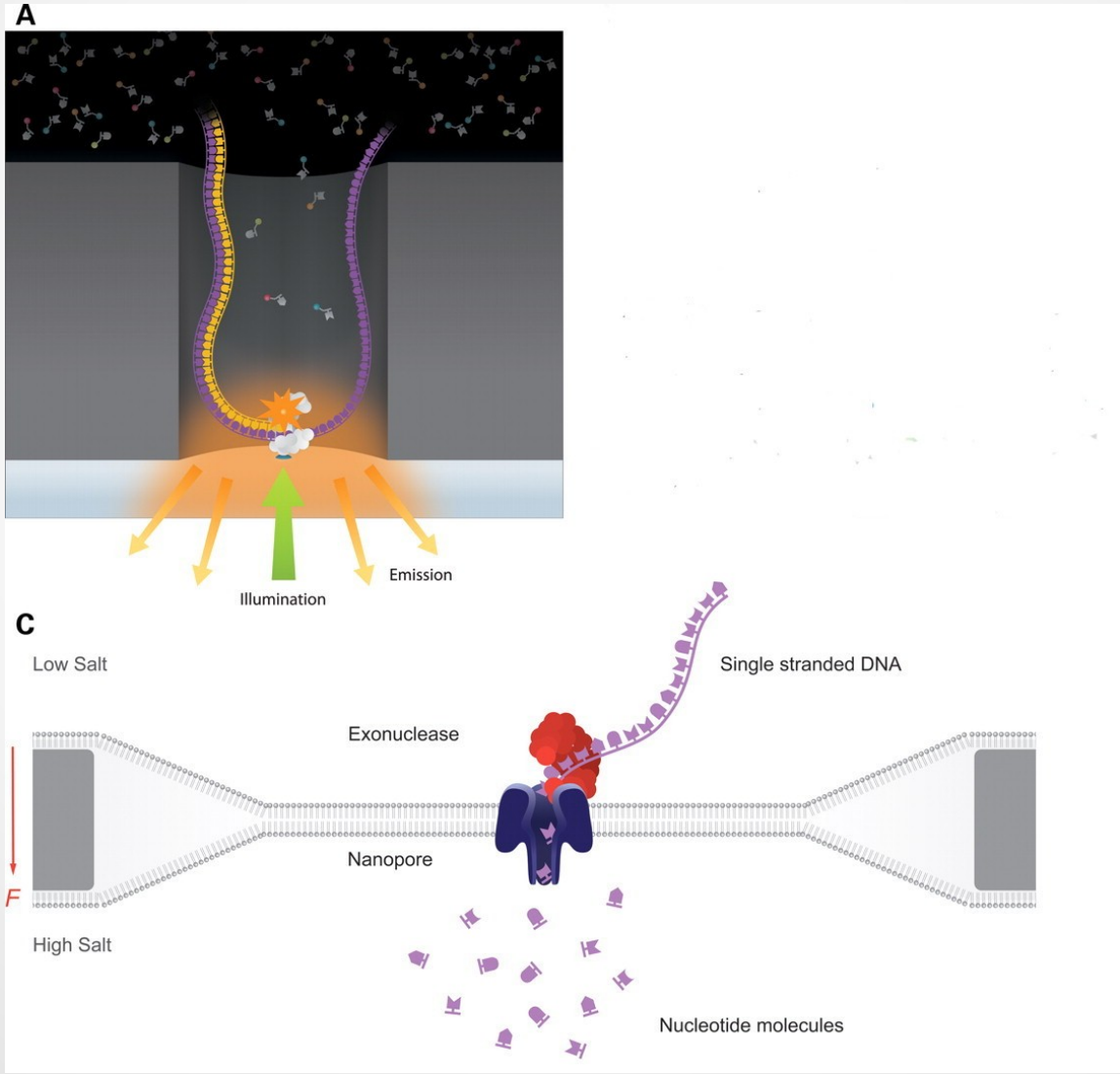
https://arimagenomics.com/kit/

Sequencing technologies: third generation

PacBio vs Nanopore

PacBio

- Shorter reads
- Less errors
- More expensive



Nanopore

- Longer reads
- More(?) errors
- Cheaper

Waiting forthcoming release of new chemistry soon!

Assembly strategies and assemblers

Short read based Sequencing	**Assemblers**
• Linked reads + HiC	Supernova + 3D-DNA/Salsa
• Overlapping PE reads + HiC	w2rap + 3D-DNA/Salsa

Long read based Sequencing	**Assemblers**
• PacBio HiFi + HiC	HiFiasm
• Nanopore + Illumina + HiC	MaSuRCA/Flye/Canu + Medaka + 3D-DNA/Salsa
• Pacbio + Nanopore + HiC	Falcon/Flye/Canu + Arrow + 3D-DNA/Salsa

For highly repetitive genomes an optical mapping (Bionano) could be added as intermediate step before HiC-scaffolding

Assembly of mtDNA from WGS

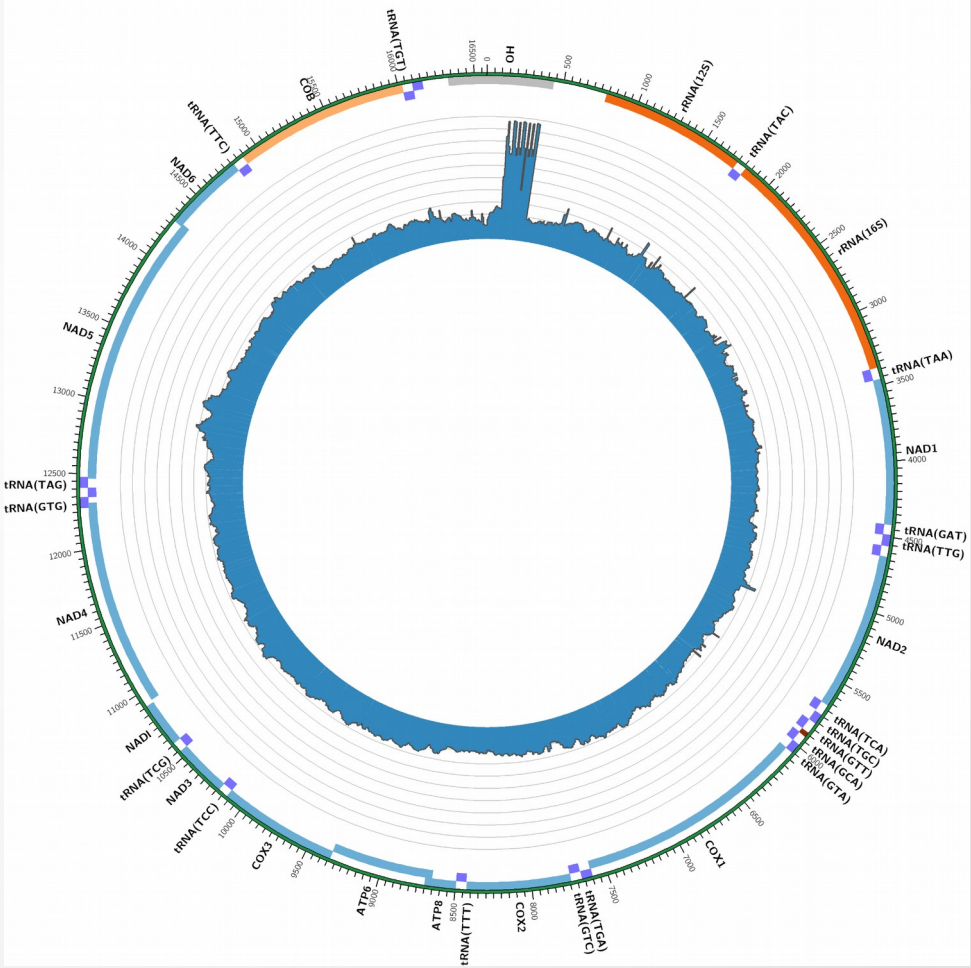
Main issue

dramatic (~100-1000x) difference in coverage between nuclear and mtDNA genome.

Solution

Independent assembly from downsampled reads (~1-5x coverage of nuclear genome) using different tools.

Tool	Function
MitoZ	assembly
Mitos	annotation



Tomarovsky, 2021

Quality control of the assembly

Assembly quality metrics

- Number of C-scaffolds should be equal to number of chromosome pairs
- Number of breaks introduced by HiC-scaffolding
- N50 and L50 (not informative for chromosome length assemblies)
- Number of Ns in the assembly
- Number of unplaced scaffolds
- BUSCO metrics

and so on ...

N50 - maximal length of contig/scaffold in the assembly for which all sequences of such length and longer encompass no less than 50% of the assembly.

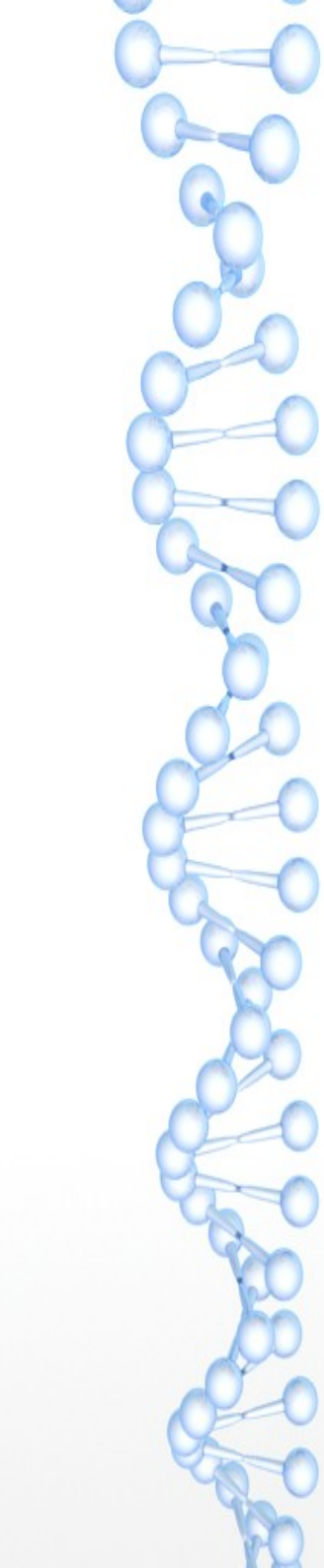
BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

- Assesses number of conservative genes present in the assembly
- specific databases for different taxa

Example for 4 mustelid species (Mammalian database, 9226 BUSCOs)

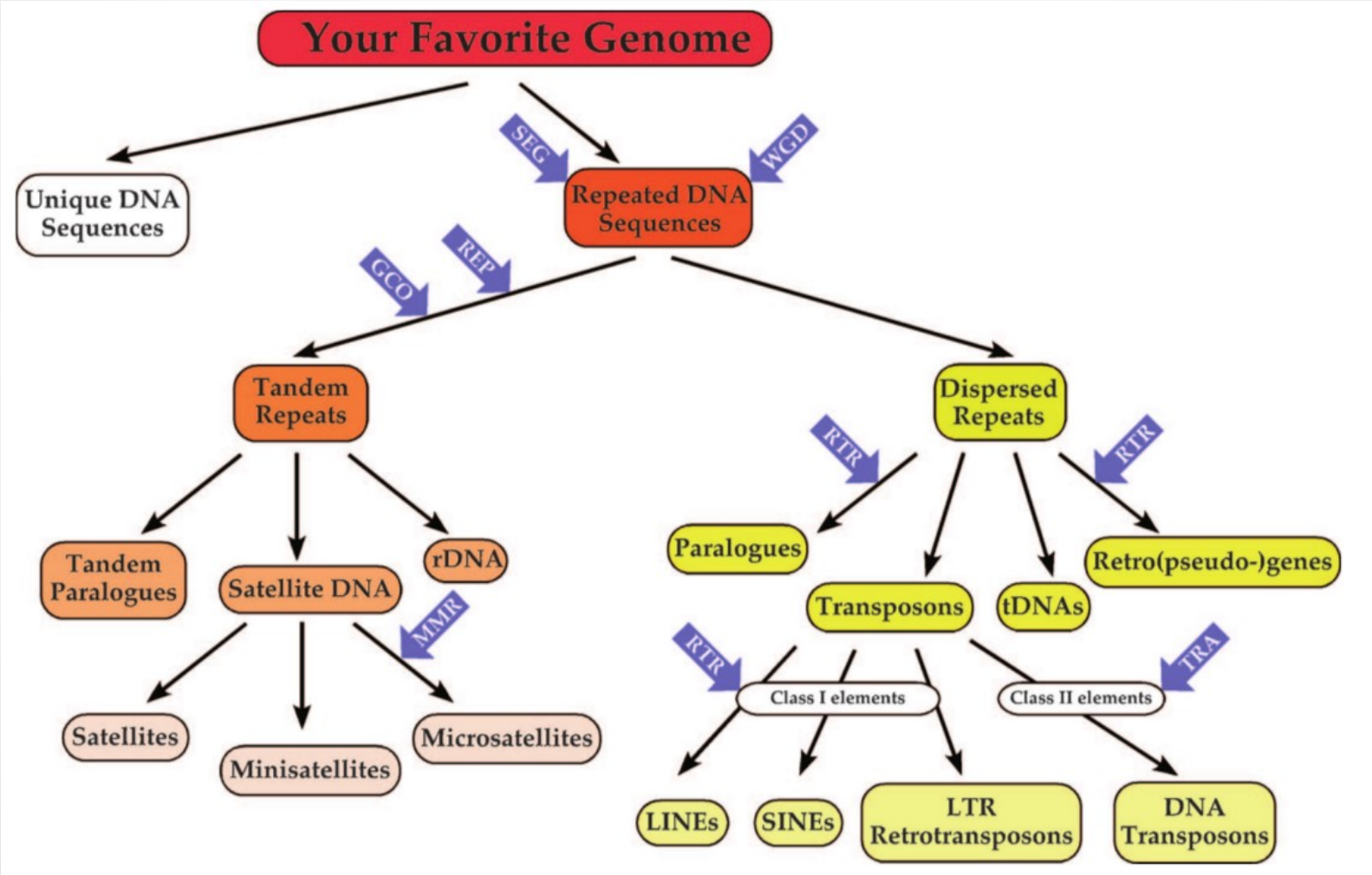
Species	Complete, %	Complete and single copy, %	Complete and duplicated, %	Fragmented, %	Missing, %
Mustela nigripes	96.2	94.0	2.2	1.0	2.8
Mustela putorius furo	94.0	92.8	1.2	1.2	4.8
Enhydra lutris	96.4	95.5	0.9	0.8	2.8
Pteronura brasiliensis	95.2	94.1	1.1	1.4	3.4



III. Genome Projects

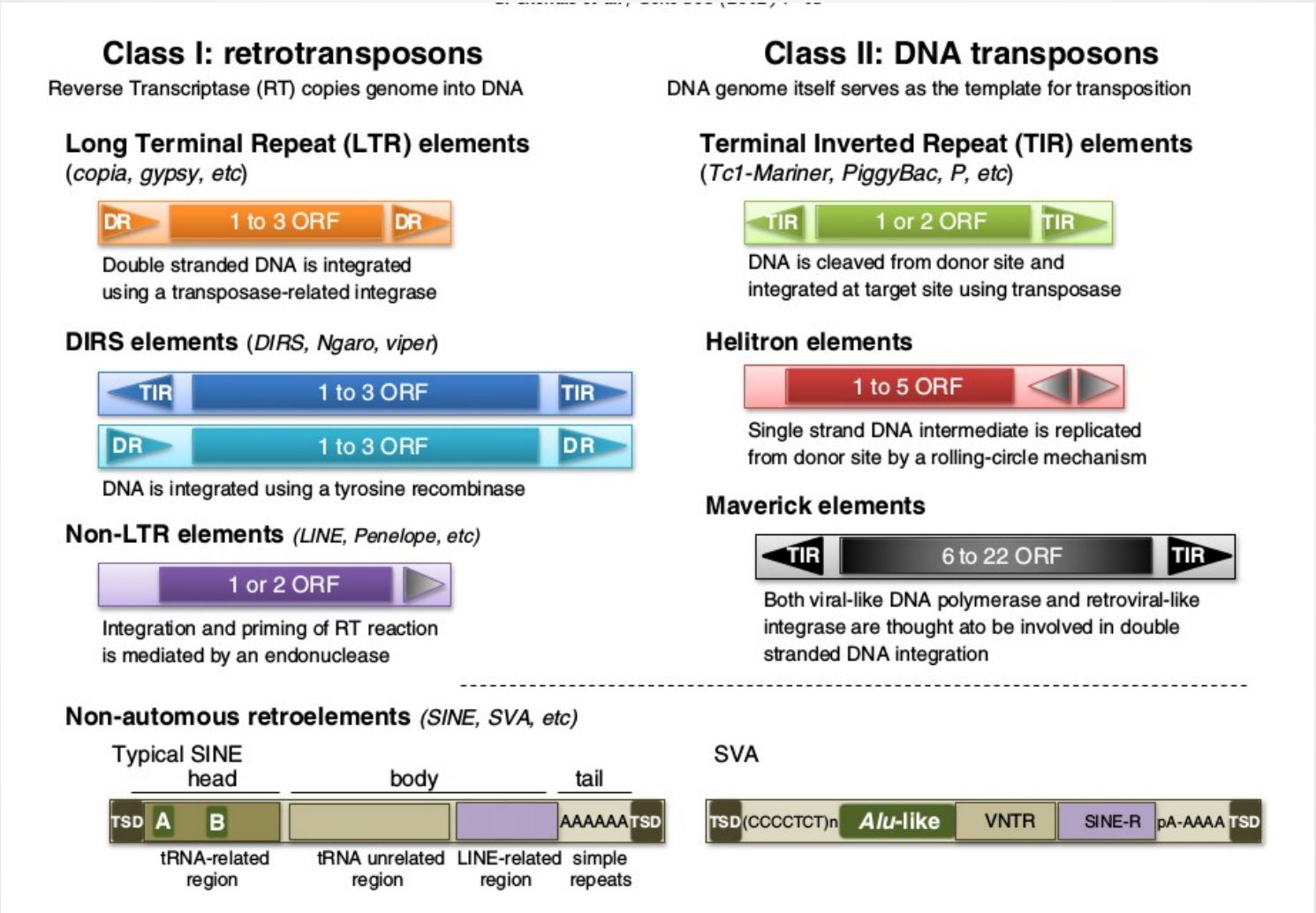
Analysis

Repeat types



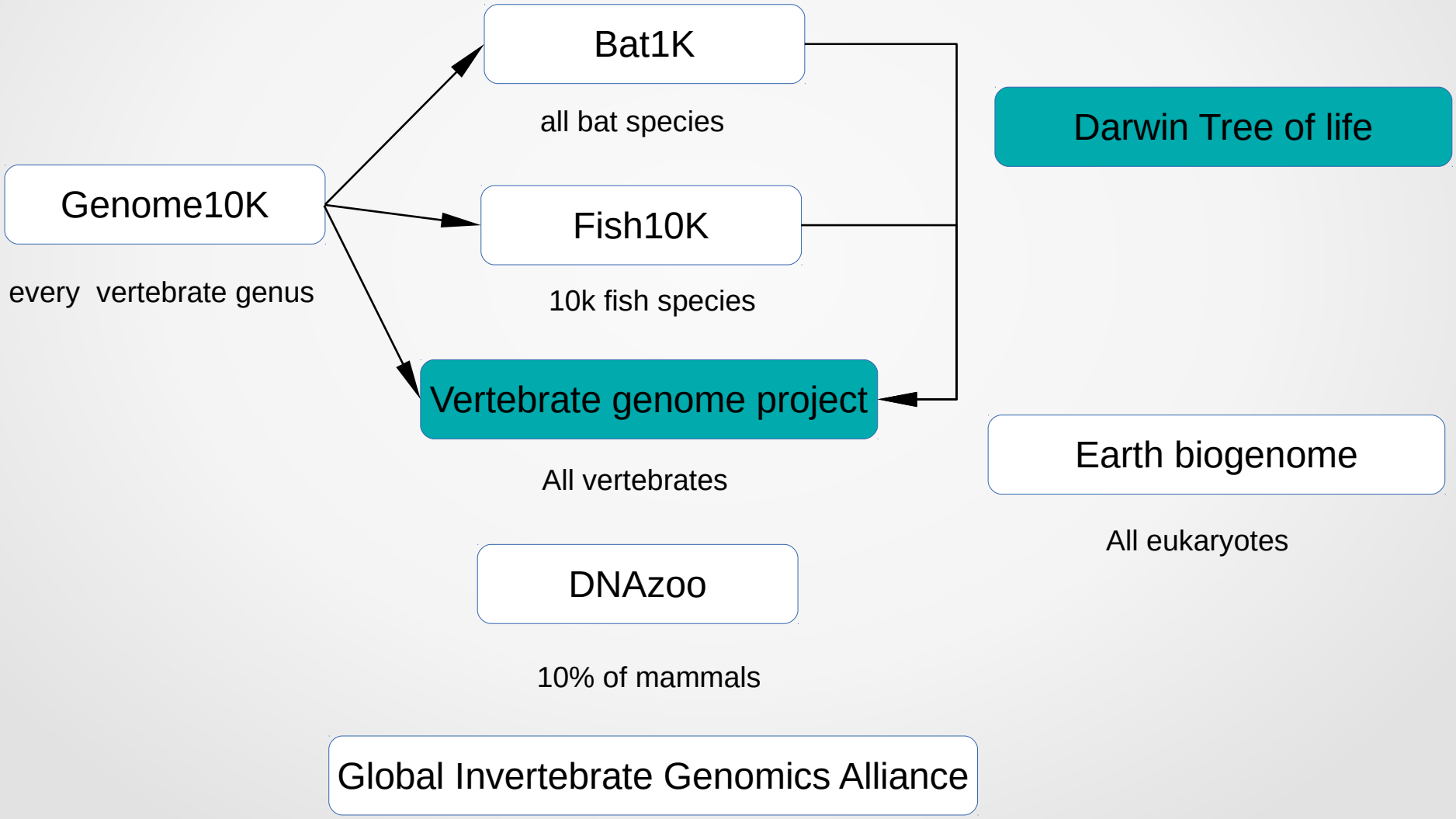
Richard et al, 2008

Major types of mobile elements

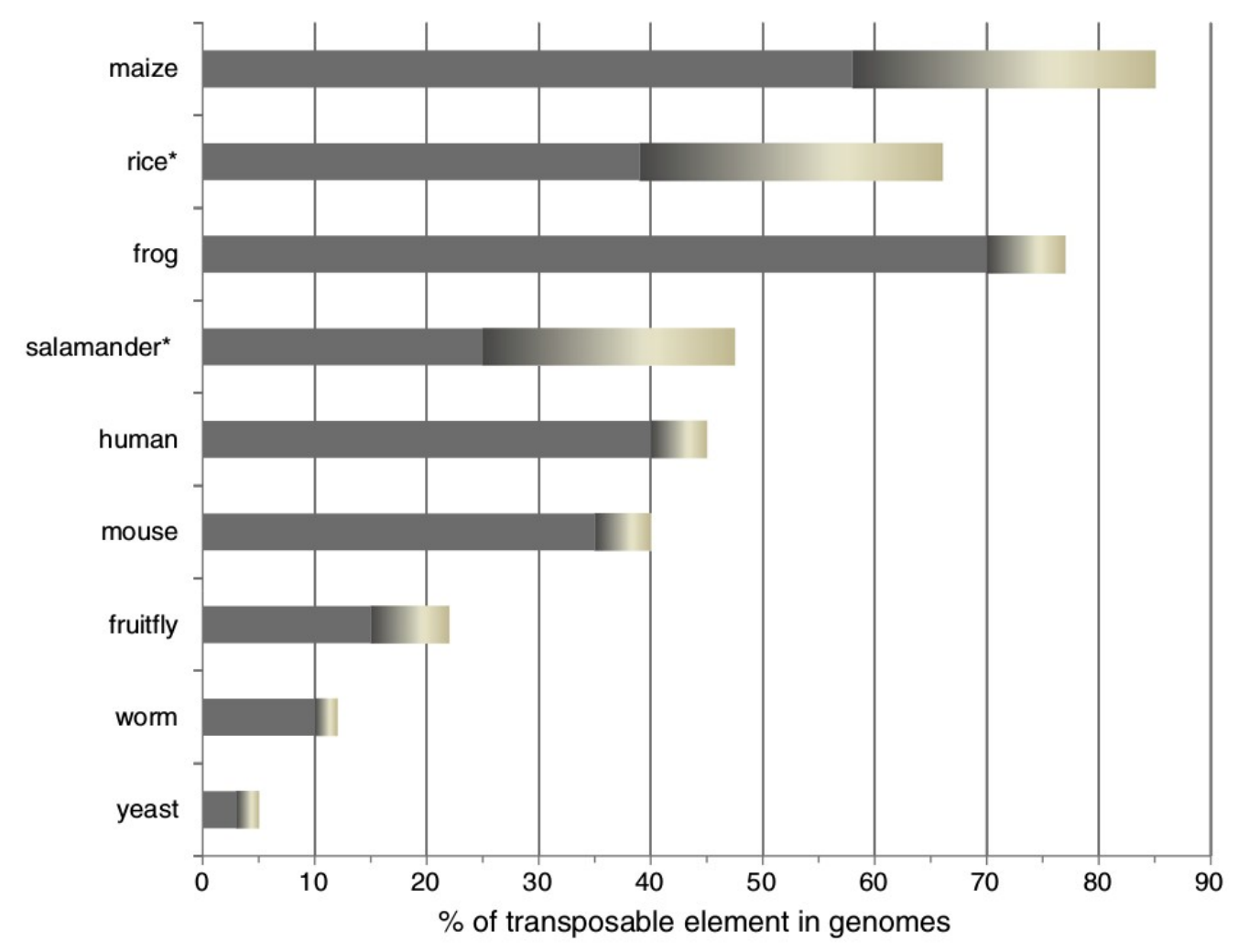


Chénais et al, 2012

Big multigenome projects and their aims

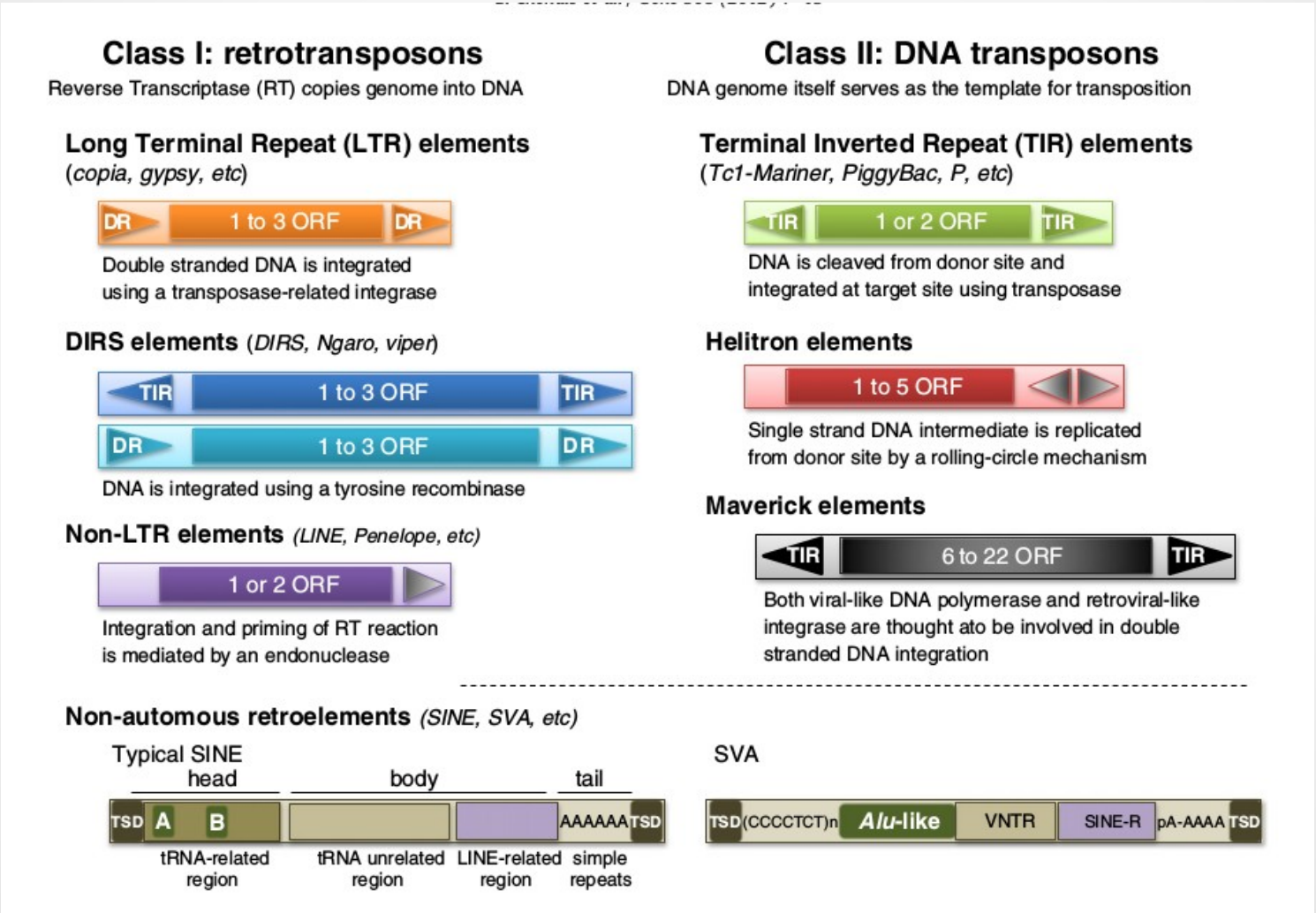


Presence of mobile elements in eukaryotic genomes



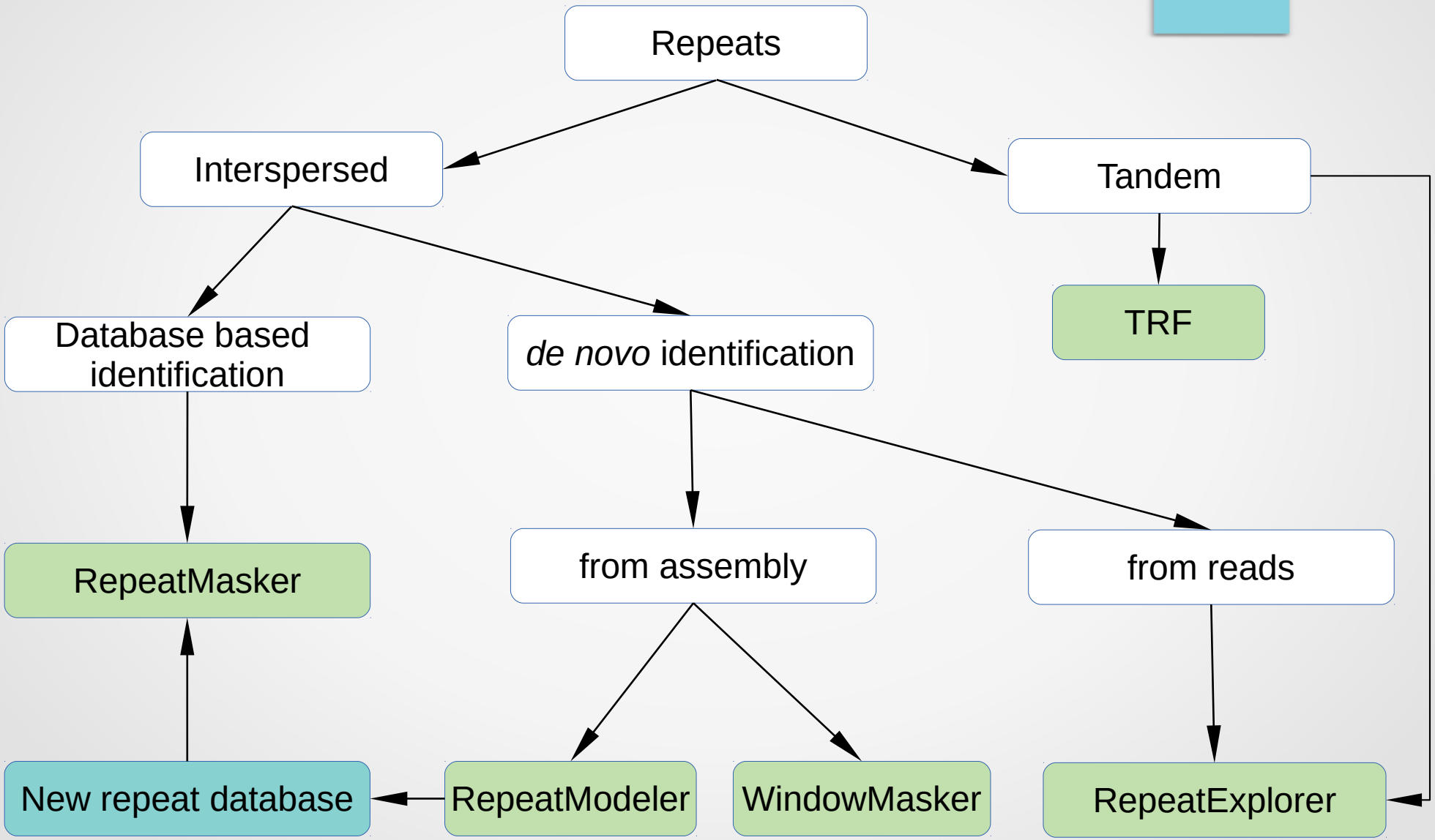
Chénais et al, 2012

Major types of mobile elements

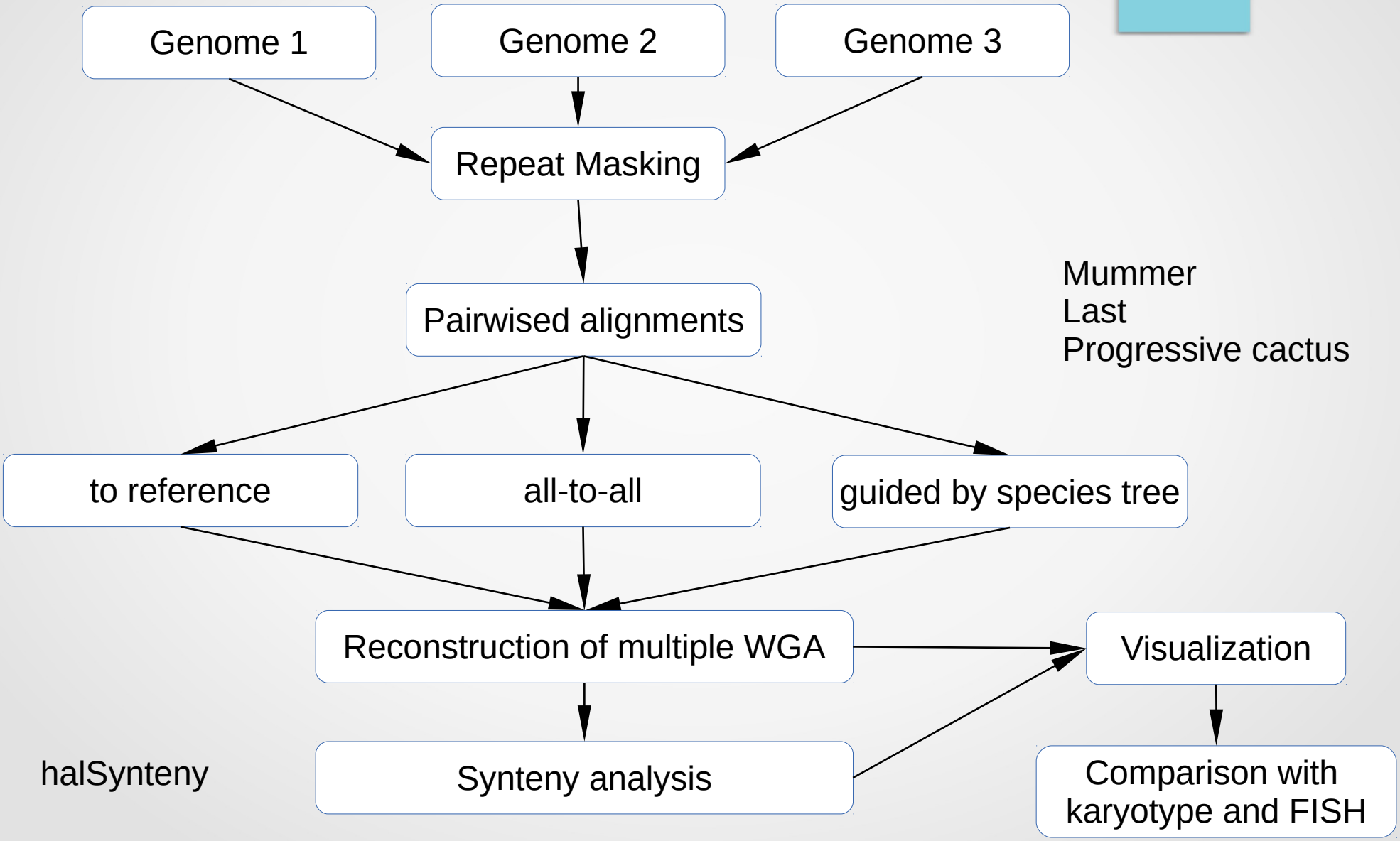


Richard et al, 2008

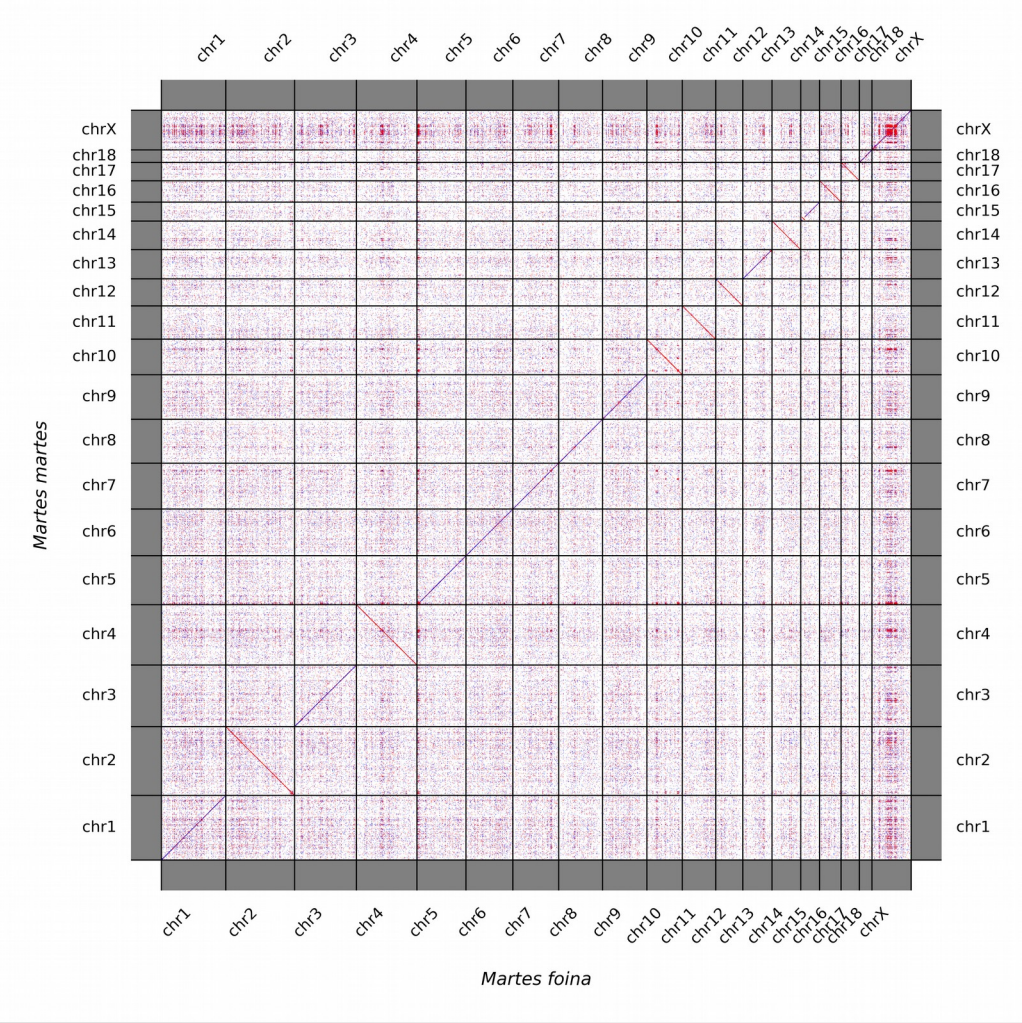
Tools for repeat identification

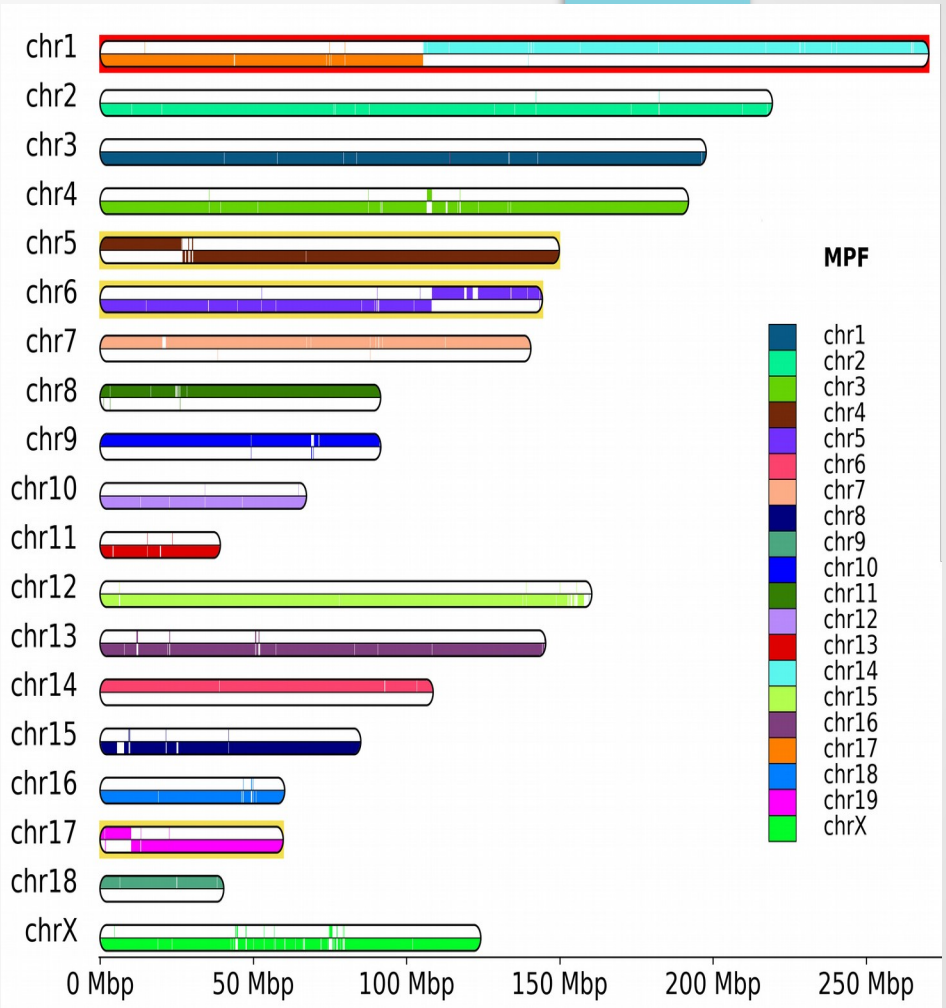


Whole genome alignment (WGA)



Whole genome alignment and synteny blocks





Protein-coding gene annotation

Combination

I. RNAseq-based

II. homology-based

CESAR2.0
Comparative annotation toolkit
Exonerate

III. de novo predictions

Augustus
Metaeuk

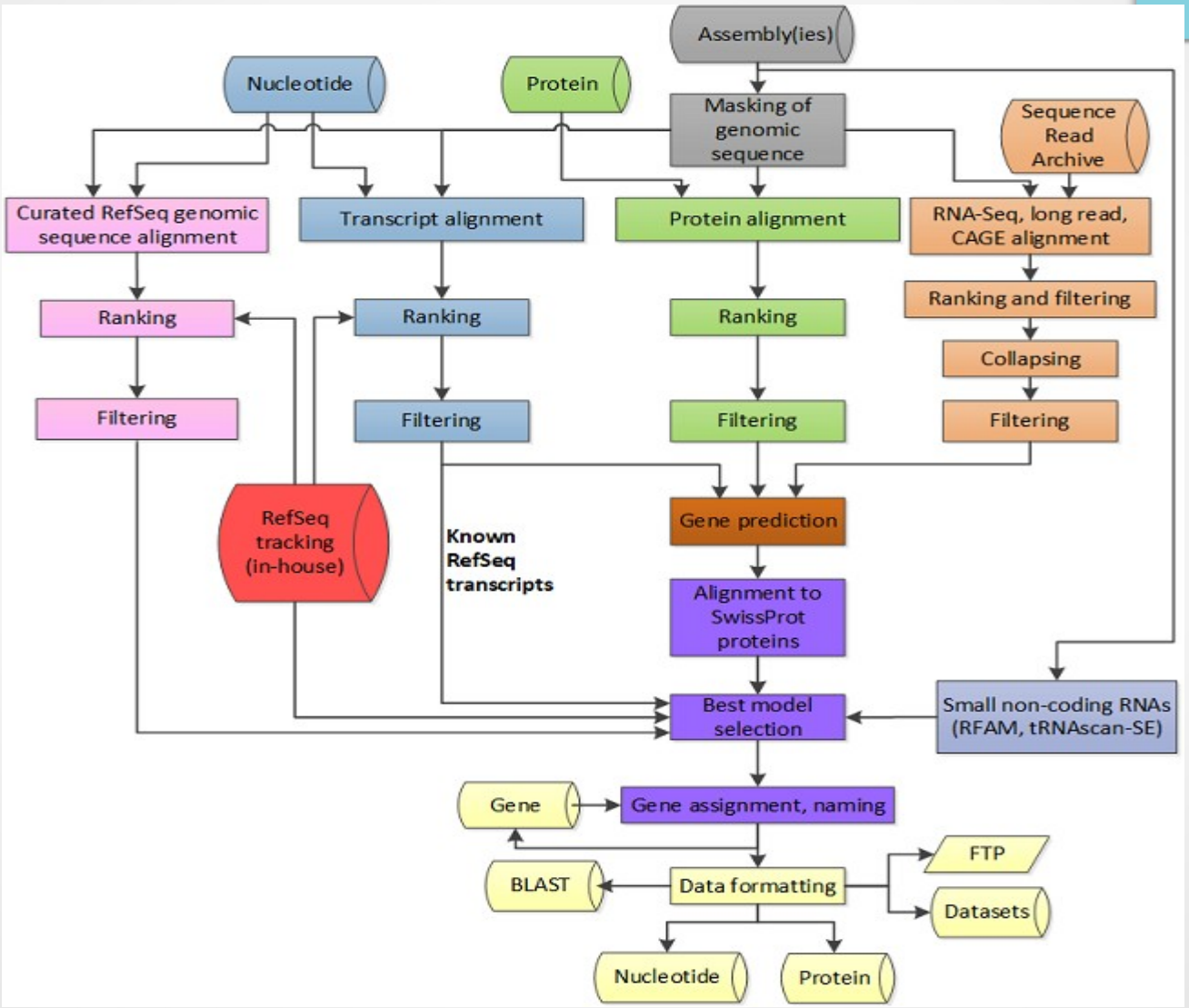
IV. Hybrid approach

MAKER2

Quality

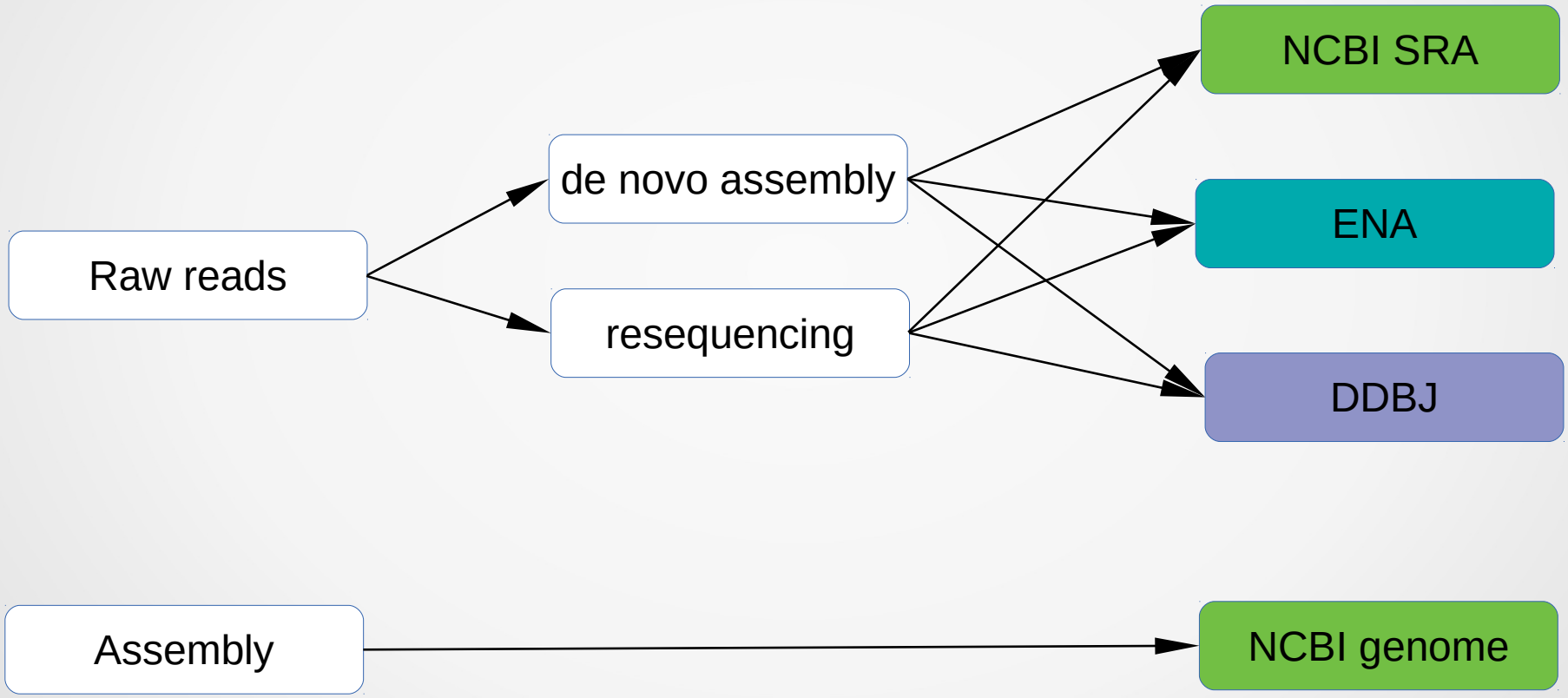
Price

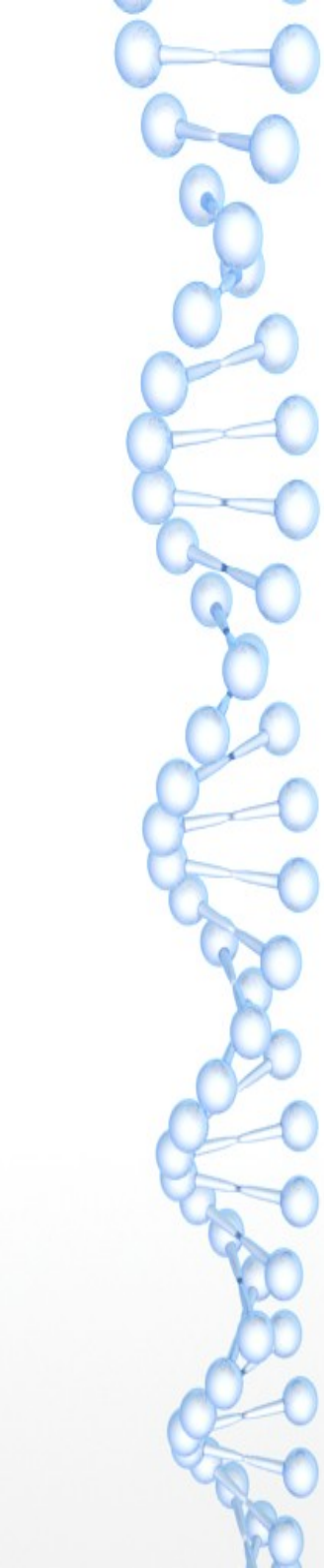
The NCBI Eukaryotic Genome Annotation Pipeline



https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

Data sharing





End of module III