Genomes and Genomics

part of "Gene Regulation" – Sequence Analysis module

Sonja J. Prohaska/Ella Cassidy

Leipzig, WS 2026/27

Computational EvoDevo
University of Leipzig

Content

1. What is a "genome"? Varying definitions
2. How big are genomes, and how do we measure that? (size, k-mers, repeats, ploidy, WGD)
3. What is inside a genome? (human genome, repeats, ERVs)
4. Reading genomes: reads, assemblies and references
5. One organism, many genomes: variation among cells

What Is a Genome?

Definitions, the tree of life, examples

What is a "genome"?

A universal definition has not been formulated so far.

The term "genome" refers to a **pluralistic**, not necessarily self-consistent concept.

Different disciplines have their own working definition.

The genome is, among others, equated with...

- the haploid set of chromosomes (eukaryote genetics)
- the total (nuclear) DNA in a cell (molecular biology)
- the heritable genetic material (theoretical biology)
- the genotype (population genetics)
- the sum of all genes (science journalist)
- the genetic information to build an organism (developmental biology)

Pluralism in concepts and definitions is a recurring problem in the science of biology.

The pragmatic bioinformatician's view of a genome

The genome is

the DNA sequence(s), the result of genome sequencing and assembly, that can be downloaded from a genome server using the **species name** as a key.

The pragmatic bioinformatician's view of a genome

The genome is

the DNA sequence(s), the result of genome sequencing and assembly, that can be downloaded from a genome server using the **species name** as a key.

In practice: a **text file** (FASTA format) with one sequence per chromosome or contig, written in the four letters A, C, G and T.

```
>chr1 Example species, assembly v1
CATGACGTCGCGGACAACCCAGAATTGTCTTGAGCGATGGTAAGATCTAACCTCACTGCCGGGGGAGGCTCATAC
CTGGGGCTTTACTGATGTCATACCGTCTTGCACGGGGATAGAATGACGGTGCCCGTGTCTGCTTGCCTCGAAGCA
ATTTTCTGAAAGTTACAGACTTCGATTAAAAAGATCGGACTGCGCGTGGGCCCGGAGAGACATGCGTGGTAGTCA
TTTTTCGACGTGTCAAGGACTCAAGGGAATAGTTTGGCGGGAGCGTTACAGCTTCAATTCCCAAAGGTCGCAAGA
CGATAAAATTCAACTACTGGTTTCGGCCTAATAGGTCACGTTTTATGTGAAATAGAGGGGAACCGGCTCCCAAAT
CCCTGGGTGTTCTATGATAAGTCCTGCTTTATAACACGGGGCGGTTAGGTTAAATGACTCTTCTATCTTATGGTG
ATCCAAGCGCCCGCTAATTCTGTTCTGTTAATGTTCATACCAATACTCACATCACATTAGATCAAAGGATCCCCG
AGCCCAGTCGCAAGGGTCTGCTGCTGTTGTCGACGCCTCATGTTACTCCTGGAATCTACCTGCCCTCCCCTCACC
GGTTAAGGCGTGTGATCGACGATGCAGGTATACATCGGCTCGGACCTACAGTGGTCGATCGACTGGCTACTGGCT
TCGCGGTTCGGCGCGTAGTTGAGTGCGATAACCCAACCGGTGGCAAGTAGCAAGAAGACCTACCTGGGTCACCTT
AGACAACCTAACTAATAGTCTCTAACGGGGAATTACCTTTACCAGTCTCATGCCTCCAATATATCTGCACCGCTT
CAATGATATCGCCCACAGAAAGTAGGGTCTCAGGTATCGCATACGCCGCGCCCGGGTCCCAGCTACGCTCAGGAC
GACAGTAGAGAGCTATTGTGTAATTCAGGCTCAGCATTCATCGACCTTTCCTGTTGTGAATATTGTGCTAATGCA
TCTCGTCCGTAACGATCTGGGGGGCAAAACCGAATATCCGTATTCTCGTCCTACGGGTCCACAATGAGAAAGTCC
TGCGCGTGATCGTCAGTTAAGTTAAATTAATTCAGGCTACGGTAAACTTGTAGTGAGCTAAGAATCACGGGAATC
ACGGGTTCGCTACAGATGAACTGAATTTATACACGGACAACTCATCGCCCATTTGGGCGTGGGCACCGCAGATCA
AAAGTGGCAGATTAGGAGTGCTTGATCAGGTTAGCAGGTGGACTGTATCCAACAGCGCATCAAACTTCAATAAAT
CCAAAGCGTTGTAGTGGTCTAAGCACCCCTGAACAGTGGCGCCCATCGTTAGCGTAGTACAACCCTTCCCCCTTG
AGGTGCGACATGGGGCCAGTTAGCCTGCCCTATATCCCTTGCACACGTTCAATAAGAGGGGCTCTACAGCGCCGC
TTTTTAAATTAGGATGCCGACCCCATCATTGGTAACTGTATGTTCATAGATATTTCTTCAGGAGTAATAGCGACA
AGCTGACACGCAAGGGTCAACAATAATTTCTACTATCACCCCGCTGAACGACTGTCTTTGCAAGAACCAACTGGG
CTTAGATTCGCGTCCTAACGTAGTGAGGGCCGAGTCATATCATAGATCAGGCATGAGAAACCGACGTCGAGTCTA
CACACGAGTTGTAAACAACTTGATTGCTATACTGTAGCTACCGCAAGGATCTCCTACATCAAAGACTACTGGGCG
ATCTGGATCCGAGTCAGAAATACGAGTTAATGCAAATTTACGTAGACCGGTGAAAACACGTGCCATGGGTTGCGT
AGACCGTAGTCAGAAGTGTGGCGCGCTATTCGTACCGAACCGGTGGAGTATACAGAATTGCTCTTCTACGACGTA
AGGAGCTCGGTCCCCAATGCACGCCAAAAAAGGAATAAAGTATTCAAACTGCGCATGGTCCCTCCGCCGGTGGCA
CTATTATCCATCCGAACGTTGAACCTACTTCCTCGGCTTATGCTGTCCTCAACAGTATCGCTTATGAATCGCATG
CGGCTGTGGATCTTAACGGCCACATTCTTAATTCCGACCGATCACCGATCGCCTTTCCTCGCTGGTACAATGAGT
ACTAAGTTATCCAGATCAAGGTTTGAACGGACTCGTATGACATGTGTGACTGAACCCGGGAGGAAATGCAGAGAA
CTGTTTCAAGGCCTCTGCTTTGGTATCACTCAATATATTCAGACCAGACAAGTGGCAAAATTTCGTGCGCCTCTC
CTAGGTATTCACGCAACCGTCGTAACATGCACTAAGGATAACTAGCGCCAGGGGGGCATACTAGGTCCCGGAGCT
AAAGACTACCCTATGGATTCCTTGGAGCGGGGACAATGCAGACCGGTTACGACACAATTATCGGGATCGTCTAGA
GGTATTATTAGCAAGACAATAAAGGACATTGCACAGAGACTTATTAGAATTCAACAAACAGGATCATATCATGCG
GTGTTGGGTCGGGCAAGTCCCCGAAGCTCGGCCAAAAGATTCGCCATGGAACCGTCTGGTCCTGTTAGCGTGTAC
GCCTGCTCCTGTTCCGGGTACCATAGATAGACTGAGATTGCGTCAAAAAATTGCGGCGAAAATAGAGGGGCTCCT
TGTAGAAATACCAGACTGGGGAATTTAAGCGCTTTCCACTATCTGAGCGACTAAACATCAACAAATGCGTCTACT
CGAATCCGCAGTAGGCAATTACAACCTGGTTCAGATCACTGGTTAATCAGGGATGTCTTCATAAGATTATACTTG
CCCCGACGCGACAGCTCTTCAAGGGGCCGATTTTTGGACTTCAGATACGCTAGAATTTAAAGGGTCTCTTACACC
TGCTGCGGCCTGCAGGGACCCCTAGAACTTGCCGCCTACTTGTCTCAGTCTAATAACGCGCGAAGCCGTGGGGCA
CGTGACCTTAAGTCGCAGAGCGAGTGATGAATTTGGGACGCTAATATGGGTGAATAGAGACTTATATCATCAGGG
```

Genomes and Genomics

- a geneticists' view in the year 2000: "the sum of all genes."
- **Dictionary "genome" definition**:
  "... the haploid set of chromosomes in a gamete or microorganism, or in each cell of a multicellular organism."
  "... the complete set of genes or genetic material present in a cell or organism."
- **Wikipedia "genome" definition**:
  "... the genetic material of an organism. It consists of DNA (or RNA in RNA viruses). The genome includes both the genes [the coding DNA] and the noncoding DNA, as well as mitochondrial DNA and chloroplast DNA."
- **Genotype**: the collection of genes/alleles responsible for the various traits of a given organism.
- **Genomics**: a field of study that uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to manipulate, sequence, assemble, and analyse the structure and function of genomes.

Why does mankind study genomes?

The genome...

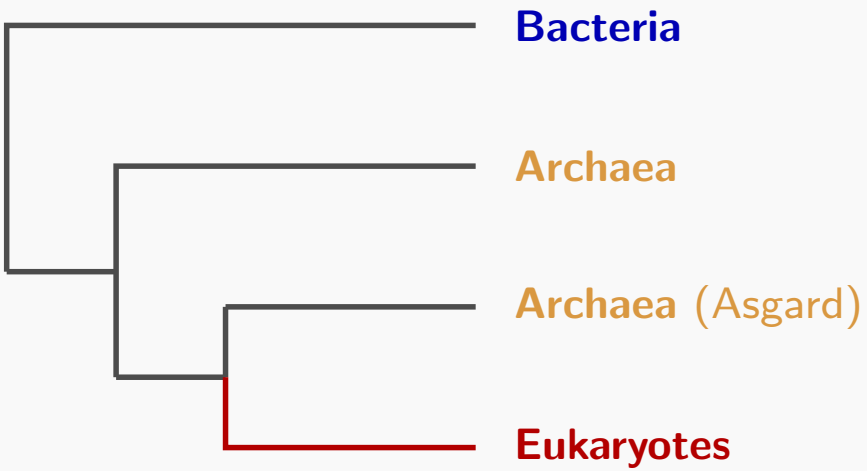
- is key to the **unity of life**: every[1] living cell or organism has a genome
- is the **heritable genetic material** that is passed from generation to generation and **changes (slowly)** over evolutionary time
- contains the genes and all[2] the **information** needed to build and maintain that organism
- is the same in all[3] cells of a multicellular organism
- is key to the heritable **(genetic) variation** among individuals of the same species
- reveals the **genetic predisposition**/risk for certain diseases
- allowing some simplifications, is a set of **strings of characters** from a 4 letter alphabet and, as such, convenient to work with

To learn why many of these statements are problematic you may want to read:
Goldman and Landweber (2016) *What is a Genome?* PLOS Genetics [1] exception: mature mammalian red blood cells do not contain DNA

[2] this is questionable

[3] exception: some organisms actively manipulate the DNA of certain cells (see later)

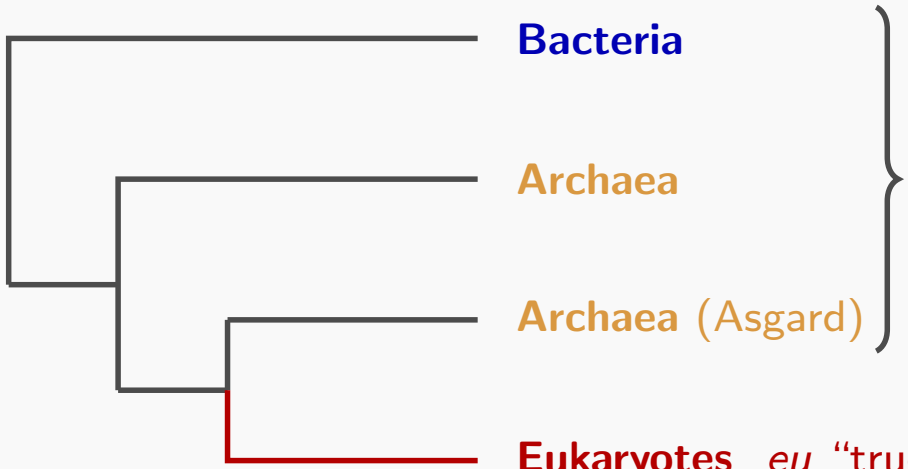
The tree of life: prokaryotes, eukaryotes and viruses



Two domains: Bacteria and Archaea.

Eukaryotes are not a third domain: they branch off from within the Archaea (Asgard lineage).

The tree of life: prokaryotes, eukaryotes and viruses



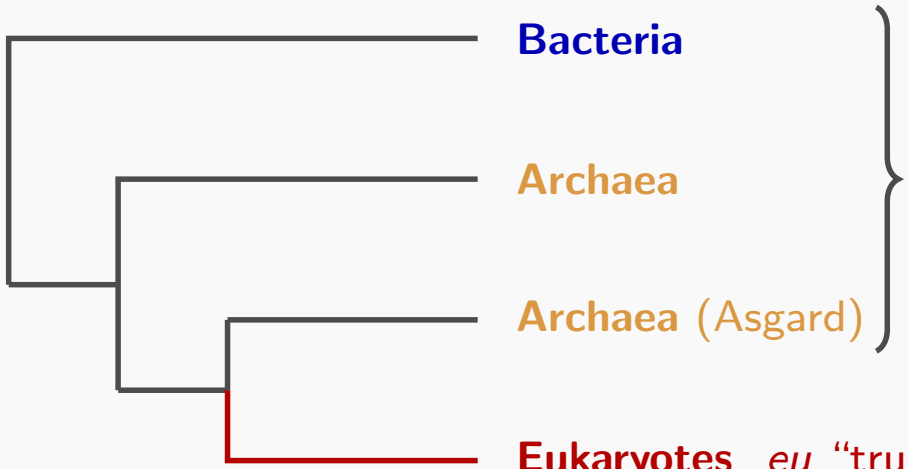
Prokaryotes

Greek *pro* "before" + *karyon* "kernel": **"before the nucleus"**

No nucleus. A grade of organisation, not a clade.

Eukaryotes *eu* "true" + *karyon*: "true nucleus"

The tree of life: prokaryotes, eukaryotes and viruses

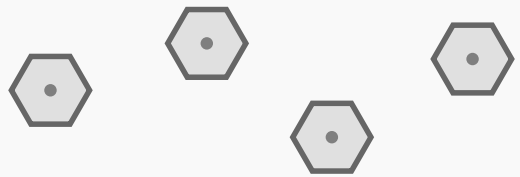


Prokaryotes

Greek *pro* "before" + *karyon* "kernel": **"before the nucleus"**

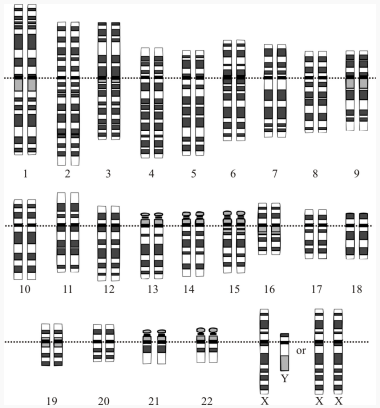
No nucleus. A grade of organisation, not a clade.

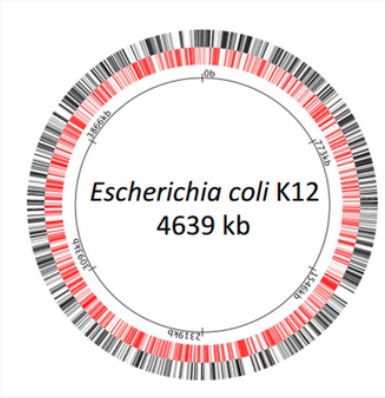
eu "true" + *karyon*: "true nucleus"

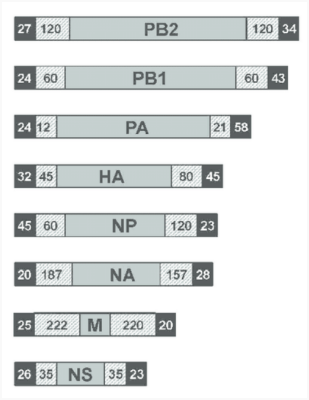


Viruses: outside the domains of life. No cells, no own metabolism or ribosomes; they replicate only inside host cells. **Not considered alive.**

Animal, bacterial and viral genome examples





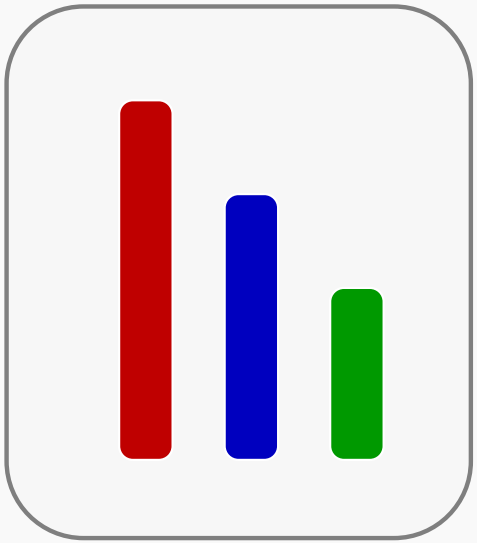


human genome *H. sapiens*	bacterial genome *E. coli*	viral genome Influenza A virus
multiple linear pieces of dsDNA	one[1] circular piece of dsDNA	multiple linear pieces of ssRNA

[1] several if plasmids are included

- **archaea also have circularized genomes**

Haploid vs. diploid: sets of chromosomes

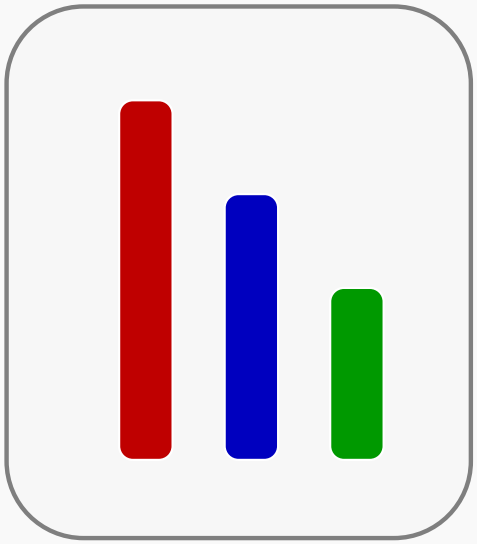


haploid (n)

one set of chromosomes

human gamete: n = 23

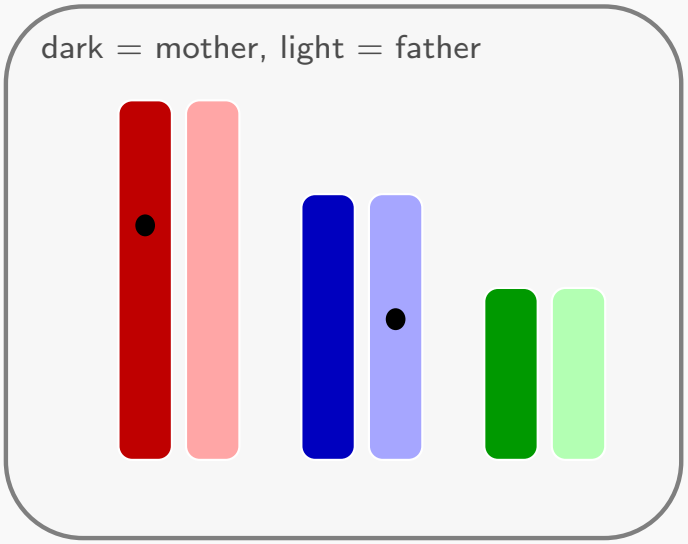
Haploid vs. diploid: sets of chromosomes



haploid (n)

one set of chromosomes

human gamete: n = 23

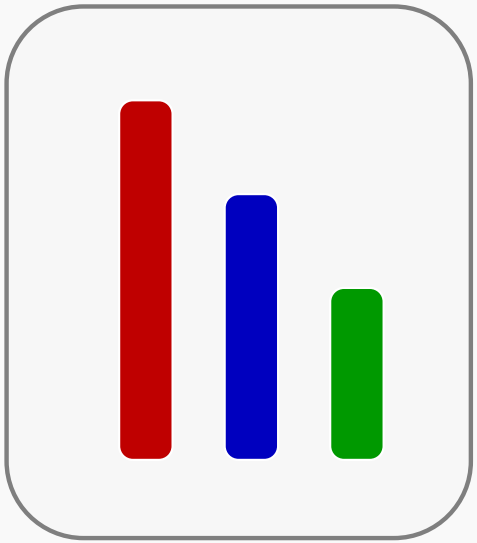


diploid (2n)

two sets: one from each parent

human body cell: 2n = 46

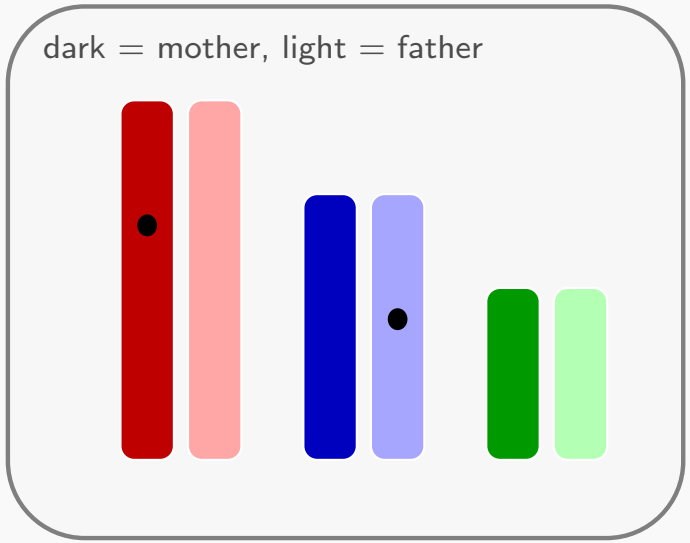
Haploid vs. diploid: sets of chromosomes



haploid (n)

one set of chromosomes

human gamete: n = 23



diploid (2n)

two sets: one from each parent

human body cell: 2n = 46

- The two sets are two **versions** of the genome, not identical copies (black dots: small differences).
- **Genome size is given per haploid set** (1C). A human has ∼3.1 Gb per set, so ∼6.2 Gb in a diploid cell.

An example: The human genome

The largest part of our body consists of **somatic cells**

- Their DNA does not participate in reproduction
- contain **46 chromosomes** (linear pieces of DNA)
- made up of 44 **autosomes** and 2 **sex chromosomes** (X,X or X,Y)
- they are *diploid*

Some cells belong to the **germline**

- Those that undergo meiosis become **germ cells**
- meiosis: a cell division process that halves the chromosome number
- contain **23 chromosomes**
- made up of **22 autosomes** and 1 **sex chromosome** (X or Y)
- they are *haploid*

The 44 autosomes form 22 homologous pairs (99.8% pairwise similarity).
Chromosome X (156 Mb) and chromosome Y (57 Mb) are less than 30% similar.

Human genome = one set of autosomes (22) + chrX / chrY (+ mtDNA*)

See `https://en.wikipedia.org/wiki/Human_genome` for details

* mitochondrial DNA (see later)

How Big Are Genomes?

Measuring size, k-mers, ploidy, giant genomes

Genome size: how is it measured?

Genome size = total amount of DNA in **one copy** of the complete genome (the "C-value": 1C = haploid DNA content). Since the early 1970s.

Conversion: 1 pg DNA $\approx$ 978 Mb

Genome size: how is it measured?

Genome size = total amount of DNA in **one copy** of the complete genome (the "C-value": 1C = haploid DNA content). Since the early 1970s.

Wet lab: flow cytometry

- stain the DNA in isolated nuclei with a fluorescent dye
- brightness $\propto$ amount of DNA
- compare with a standard of known size (e.g. onion)

$\rightarrow$ **mass** in picograms (pg)

Needs fresh cells. The reference method.

Conversion: 1 pg DNA $\approx$ 978 Mb

Genome size: how is it measured?

Genome size = total amount of DNA in **one copy** of the complete genome (the "C-value": 1C = haploid DNA content). Since the early 1970s.

Wet lab: flow cytometry

- stain the DNA in isolated nuclei with a fluorescent dye
- brightness $\propto$ amount of DNA
- compare with a standard of known size (e.g. onion)

$\rightarrow$ **mass** in picograms (pg)

Needs fresh cells. The reference method.

Computer: k-mer counting

- sequence the DNA, count k-mers in the reads
- genome size $\approx$ total k-mers / typical k-mer count
- no live cells and no reference genome needed

$\rightarrow$ number of **base pairs** (Mb, Gb)

An estimate: sensitive to coverage, sequencing errors and contamination.

Conversion: 1 pg DNA $\approx$ 978 Mb

k-mers: cut reads into short words

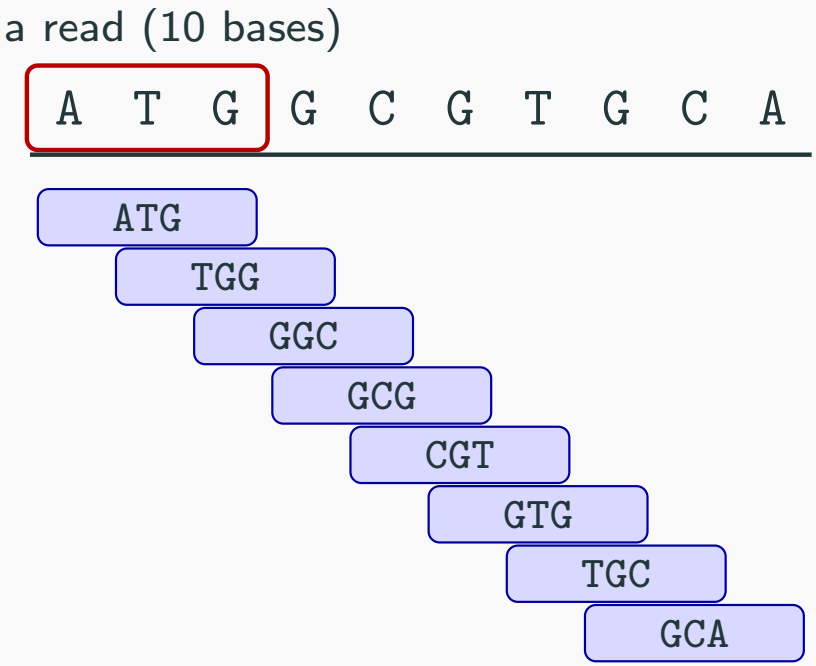
a read (10 bases)

A T G G C G T G C A

- A **k-mer** is a short word of k letters.
- Slide a window along the read, one base at a time (here $k = 3$).
- Then **count** how often each k-mer occurs in all reads.

In real data $k \approx 21$.

k-mers: cut reads into short words



- A **k-mer** is a short word of k letters.
- Slide a window along the read, one base at a time (here $k = 3$).
- Then **count** how often each k-mer occurs in all reads.

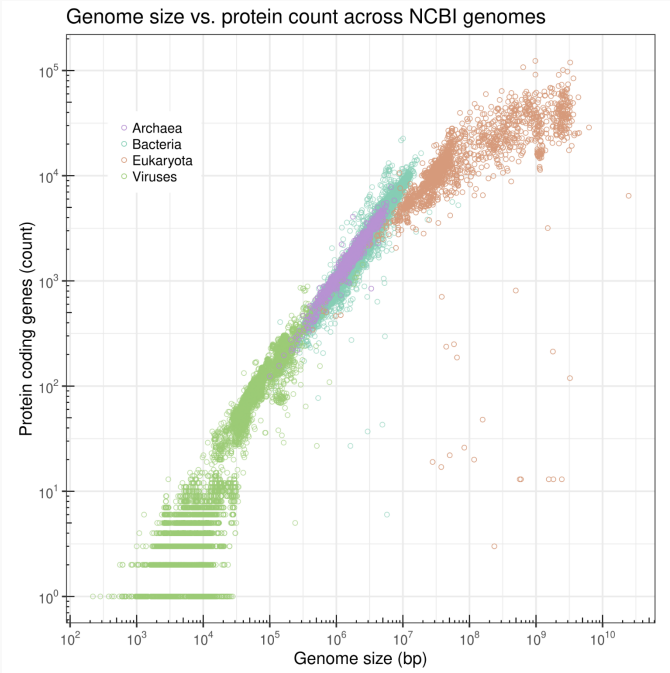
In real data $k \approx 21$.

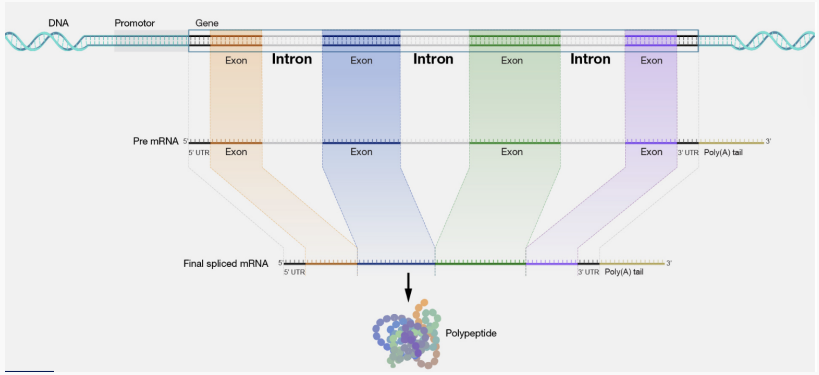
What are k-mers used for?

Counting short words is fast and needs no reference genome:

- **Genome size, repeats:** read from the shape of the k-mer histogram (e.g. GenomeScope)
- **Genome assembly:** k-mers that overlap by $k-1$ letters are chained into contigs (de Bruijn graphs)
- **Error correction:** k-mers seen only once are probably sequencing errors
- **Species identification:** match reads to k-mers of known genomes (e.g. Kraken)
- **Comparing genomes without alignment:** the more k-mers two genomes share, the closer they are (e.g. Mash)
- **Contamination screening:** k-mers from the wrong species stand out

Variation in genome size: overview





- Genome size reflects how much DNA an organism has, while gene number shows how many functional "recipes" it encodes.
- Prokaryotes and viruses: genome size $\approx$ protein-coding genes (most of the DNA is coding)
- Eukaryotes: genome size $\neq$ protein-coding genes (non-coding DNA, introns, repetitive elements)

What are repetitive elements?

Sequences that occur many times in a genome.

Tandem repeats
copies side by side



- satellites (centromeres)
- microsatellites
- telomeres

Interspersed repeats
copies scattered over the genome



- transposable elements: DNA that copies itself to new places ("jumping genes"), e.g. Alu and L1 (see appendix)
- segmental duplications

How much? Human: about half. Maize, wheat: ∼85%.

What does a repeat look like in the sequence?

Tandem repeat: the same short unit, copied side by side.

...GTCCTTC CAGCAGCAGCAGCAGCAGCAGCAGCAGCAG CAACAG...

(schematic) A CAG unit repeated 10×. In the *HTT* gene, about 40 or more copies cause Huntington's disease.

What does a repeat look like in the sequence?

Tandem repeat: the same short unit, copied side by side.

...GTCCTTC CAGCAGCAGCAGCAGCAGCAGCAGCAGCAG CAACAG...

(schematic) A CAG unit repeated 10×. In the *HTT* gene, about 40 or more copies cause Huntington's disease.

Interspersed repeat: similar copies of one element at different places in the genome (e.g. a transposon like Alu).

TTGACCGA GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCA GGATTCAA
CCATAGTC GGCCGGGCGTGGTGGCTCATGCCTGTAATCCTAGCA AAGCTTGG

(schematic) The element is almost identical at both places (red = differences), but the DNA around it is unique.

What does a repeat look like in the sequence?

Tandem repeat: the same short unit, copied side by side.

...GTCCTTC CAGCAGCAGCAGCAGCAGCAGCAGCAGCAG CAACAG...

(schematic) A CAG unit repeated 10×. In the *HTT* gene, about 40 or more copies cause Huntington's disease.

Interspersed repeat: similar copies of one element at different places in the genome (e.g. a transposon like Alu).

TTGACCGA GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCA GGATTCAA
CCATAGTC GGCCGGGCGTGGTGGCTCATGCCTGTAATCCTAGCA AAGCTTGG

(schematic) The element is almost identical at both places (red = differences), but the DNA around it is unique.

So: unique DNA occurs once. A repeat occurs many times, similar but not identical.

A 150-letter read taken from inside an element could come from any of its copies.

Repeats and disease: when a repeat expands

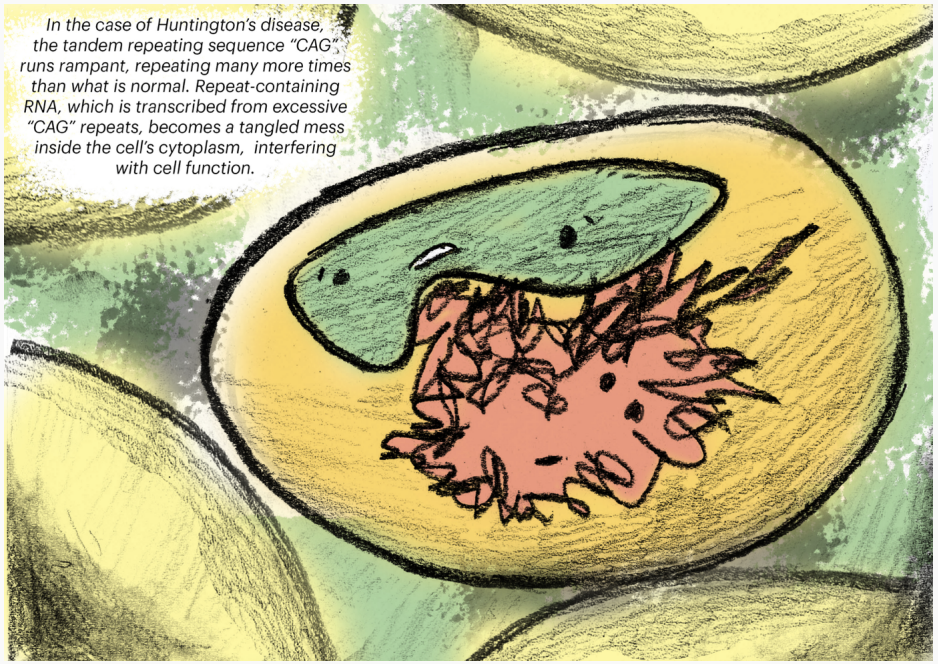


Image: Whitehead Institute, wi.mit.edu

Repeats are normal, but too many copies can make a gene toxic.

- **Huntington's disease:** CAG repeat in *HTT*, normally <36 copies. Repeat-containing RNA forms aggregates in the nucleus and in the cytoplasm. The cytoplasmic ones grow over time and deform the nucleus (Jain lab).
- **Fragile X syndrome:** CGG repeat in *FMR1*, expansion to >200 copies silences the gene.

Repeat expansions are also a bioinformatics problem: short reads cannot span them, so they need long-read sequencing or dedicated tools to size them.

Chromosome number and ploidy in a nutshell

species	2n chrom.*	genome (1C)
Fruit fly	8	∼0.14 Gb
Human	46	∼3.1 Gb
Onion	16	∼16 Gb
Carp	100	∼1.7 Gb
*Oxytricha***	∼16,000	∼0.05 Gb

* body cell, two sets. Genome size per haploid set (1C).
** macronucleus (working nucleus)

Chromosome number and ploidy in a nutshell

species	2n chrom.*	genome (1C)
Fruit fly	8	∼0.14 Gb
Human	46	∼3.1 Gb
Onion	16	∼16 Gb
Carp	100	∼1.7 Gb
*Oxytricha***	∼16,000	∼0.05 Gb

* body cell, two sets. Genome size per haploid set (1C).
** macronucleus (working nucleus)

Chromosome number is not genome size: it only tells you into how many pieces the DNA is divided.

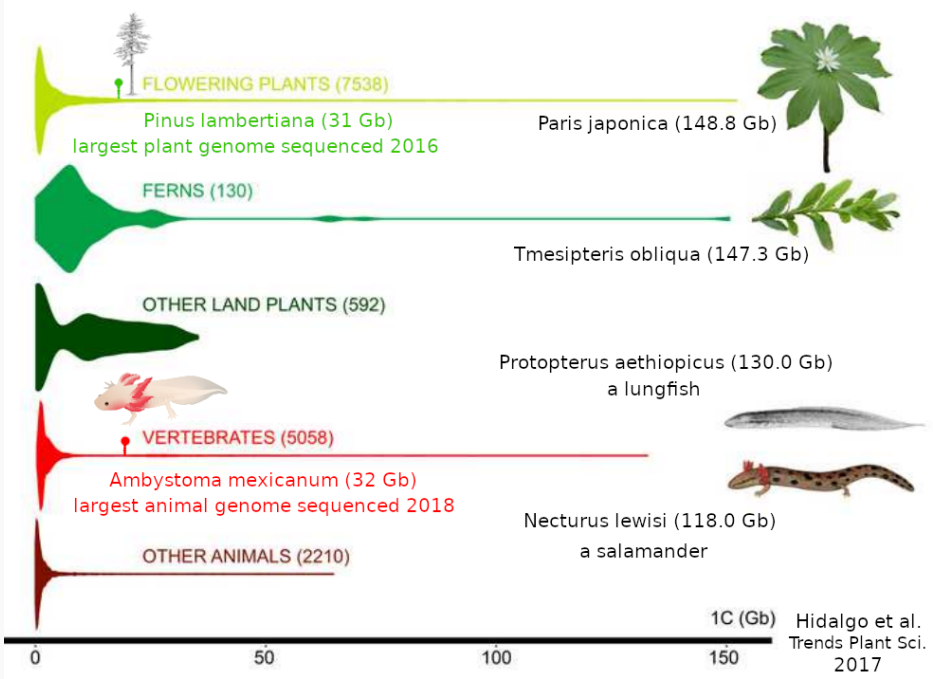
- The onion has fewer than half as many chromosomes as a human, but about five times the DNA.
- *Oxytricha*'s macronucleus has ∼16,000 tiny chromosomes, but only ∼50 Mb of DNA.

Extremes: where does “a genome” end and “life” begin?

Genome	Size	Note
Nasuia deltocephalinicola	∼112 kb	bacterial insect endosymbiont, smaller than many viral genomes
Carsonella ruddii	∼160 kb	bacterial endosymbiont of sap-feeding insects (psyllids)
Mycoplasma genitalium	∼580 kb	human urogenital pathogen
Mimivirus	∼1.2 Mb	giant virus of amoebae, bigger than the genomes above
Pandoravirus *salinus*	∼2.5 Mb	a virus with $>20\times$ the genome of *Nasuia*
Homo sapiens	∼3.1 Gb	
Tmesipteris oblanceolata	∼160 Gb	fork fern from New Caledonia, largest genome measured so far

Genome size alone cannot tell you whether something is cellular, viral, or “alive”.

Giant genomes: the largest genomes known to date



Violin plots showing the frequency and range of genome sizes in different eukaryotic groups (Hidalgo et al. 2017).

The 2024 record (*Tmesipteris oblanceolata*, 160.45 Gb, Fernández et al.) is not yet in this figure.

- Record 1C values are **measured** (flow cytometry): fork fern *Tmesipteris oblanceolata* 160.45 Gb (2024), *Paris japonica* 148.9 Gb, marbled lungfish 130 Gb.
- The largest genomes that have been **sequenced** are far smaller: Australian lungfish ∼43 Gb (2021), axolotl 32 Gb (2018), sugar pine 31 Gb (2016).

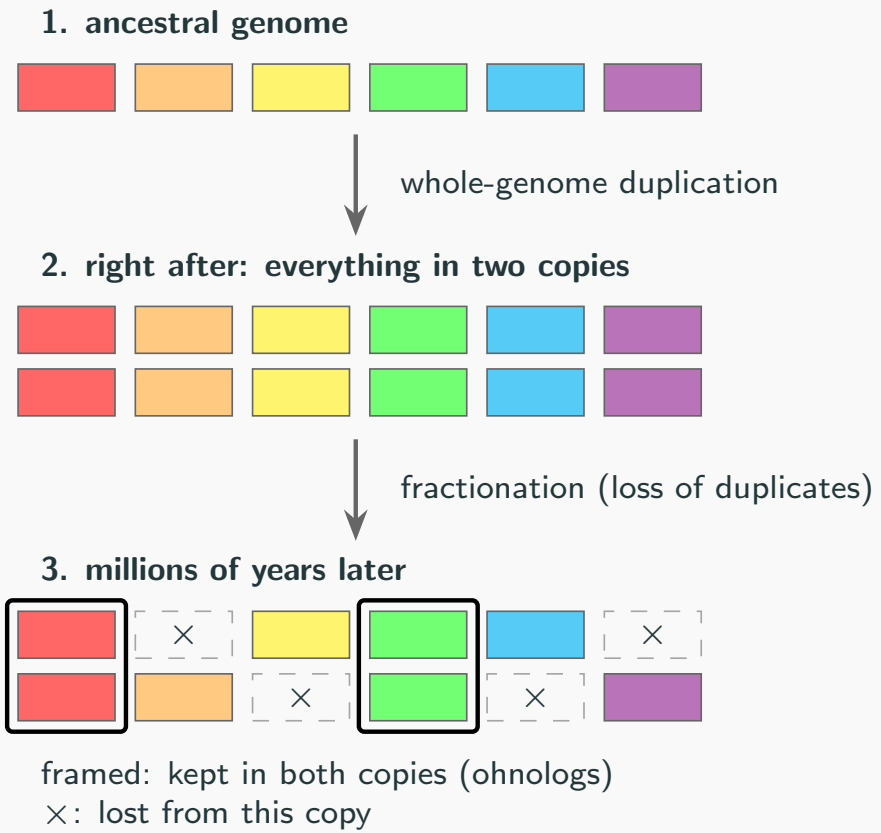
Giant genomes: why have they not been sequenced?

- A 100 Gb genome is roughly 30× the size of the human genome.
- Such genomes are dominated by repetitive DNA, which makes assembly extremely difficult even with long-read sequencing.
- 30× coverage of 100 Gb means **3 Tb of raw sequence**: storage and compute to assemble them.

Why does genome size vary so much?

- whole-genome duplication (polyploidy)
- expansion of repetitive elements and transposons
- the balance between DNA insertion and DNA loss (plants often have low rates of DNA removal)

Whole-genome duplication and fractionation



- Paleopolyploidy: two rounds in early vertebrates, at least one in most flowering plants.
- After a WGD every gene has a duplicate. Most are lost again, usually by pseudogenization.

What Is Inside a Genome?

Genes are the minority (at least in humans)

The human genome: a rough inventory

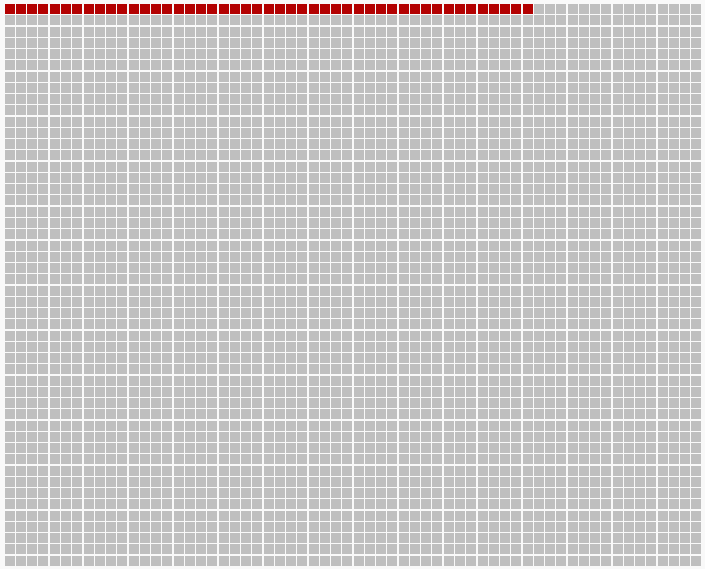
Component	Share of the genome
protein-coding exons	$\sim$1.5%
introns	$\sim$25%
repeats and transposable-element-derived DNA	about half
of which endogenous retroviruses	$\sim$8%
other non-coding: regulatory DNA, ncRNA genes, UTRs, pseudogenes, unique intergenic DNA	the rest

The categories overlap: many repeats lie inside introns and intergenic regions.

Only $\sim$1.5% codes for protein. What does the rest do?

Answer: it depends on what you mean by "function" (next slide).

If you printed one human genome as books...



1 square = 1 book of 500 pages

- 3.1 billion letters (haploid), 1 letter per base
- $\sim$**3,100 books** of 500 pages ($\sim$6,200 for the diploid genome)
- Reading 1 letter per second, non-stop: $\sim$**100 years**

Red: only $\sim$47 of the 3,100 books would hold the protein-coding exons ($\sim$1.5%).

Your genome is partly viral

- **Endogenous retroviruses (ERVs)**: remnants of ancient retroviral infections. Only an insertion into a germ cell (or early embryo) is passed on to offspring, so an ERV sits in every cell of every descendant. Insertions in somatic cells die with the individual. ERVs make up about 8% of human DNA, five times more than the protein-coding exons.
- **Syncytin**: a gene derived from a retroviral envelope gene that mediates the cell fusion needed to build the placenta. It acts in the placenta, but the gene is in all cells. Different mammal lineages captured different syncytins independently.

parasite, junk, or raw material for evolution?

How much of the genome is "functional"?

It depends on what you mean by "function":

1. It does something
Transcribed, bound by proteins, chemically marked (ENCODE 2012)

$\sim$80% of the genome

How much of the genome is "functional"?

It depends on what you mean by "function":

1. It does something
Transcribed, bound by proteins, chemically marked (ENCODE 2012)

$\sim$80% of the genome

2. It matters
Changes would be harmful, so the sequence is conserved between species

$\sim$5–15% of the genome

How much of the genome is "functional"?

It depends on what you mean by "function":

1. It does something
Transcribed, bound by proteins, chemically marked (ENCODE 2012)

$\sim$80% of the genome

2. It matters
Changes would be harmful, so the sequence is conserved between species

$\sim$5–15% of the genome

Active does not mean important: DNA can be transcribed or bound by chance. How much of the genome is "functional" is still debated (Graur et al. 2013; Doolittle 2013).

Important DNA outside the genes

Not protein-coding does not mean junk. Conserved non-coding DNA includes:

- **Regulatory DNA** (promoters, enhancers): switches that decide where and when a gene is on.
  Example: the ZRS enhancer lies $\sim$1 Mb away from the *Shh* gene. Single-base changes in it cause extra fingers and toes (Lettice et al. 2003).

Important DNA outside the genes

Not protein-coding does not mean junk. Conserved non-coding DNA includes:

- **Regulatory DNA** (promoters, enhancers): switches that decide where and when a gene is on.
  Example: the ZRS enhancer lies $\sim$1 Mb away from the *Shh* gene. Single-base changes in it cause extra fingers and toes (Lettice et al. 2003).
- **Non-coding RNA genes:** the product is an RNA, not a protein.
  Examples: rRNA, tRNA (necessary for protein biosynthesis)

Important DNA outside the genes

Not protein-coding does not mean junk. Conserved non-coding DNA includes:

- **Regulatory DNA** (promoters, enhancers): switches that decide where and when a gene is on.
  Example: the ZRS enhancer lies $\sim$1 Mb away from the *Shh* gene. Single-base changes in it cause extra fingers and toes (Lettice et al. 2003).
- **Non-coding RNA genes:** the product is an RNA, not a protein.
  Examples: rRNA, tRNA (necessary for protein biosynthesis)
- **Signals inside genes:** UTRs and splice sites control how a transcript is processed, but are not translated.

Important DNA outside the genes

Not protein-coding does not mean junk. Conserved non-coding DNA includes:

- **Regulatory DNA** (promoters, enhancers): switches that decide where and when a gene is on.
  Example: the ZRS enhancer lies $\sim$1 Mb away from the *Shh* gene. Single-base changes in it cause extra fingers and toes (Lettice et al. 2003).
- **Non-coding RNA genes:** the product is an RNA, not a protein.
  Examples: rRNA, tRNA (necessary for protein biosynthesis)
- **Signals inside genes:** UTRs and splice sites control how a transcript is processed, but are not translated.

Important DNA outside the genes

Not protein-coding does not mean junk. Conserved non-coding DNA includes:

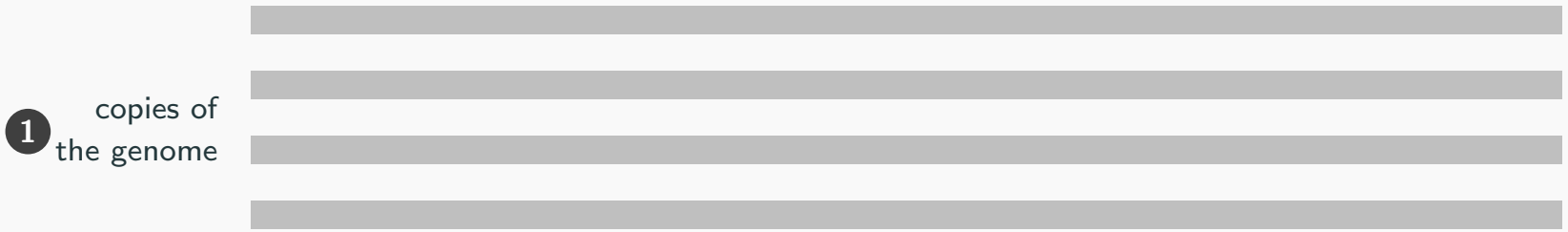
- **Regulatory DNA** (promoters, enhancers): switches that decide where and when a gene is on.
  Example: the ZRS enhancer lies ∼1 Mb away from the *Shh* gene. Single-base changes in it cause extra fingers and toes (Lettice et al. 2003).
- **Non-coding RNA genes:** the product is an RNA, not a protein.
  Examples: rRNA, tRNA (necessary for protein biosynthesis)
- **Signals inside genes:** UTRs and splice sites control how a transcript is processed, but are not translated.

Gene regulation is the topic of the rest of this course.

Trials and Tribulations of Genome Assemblies

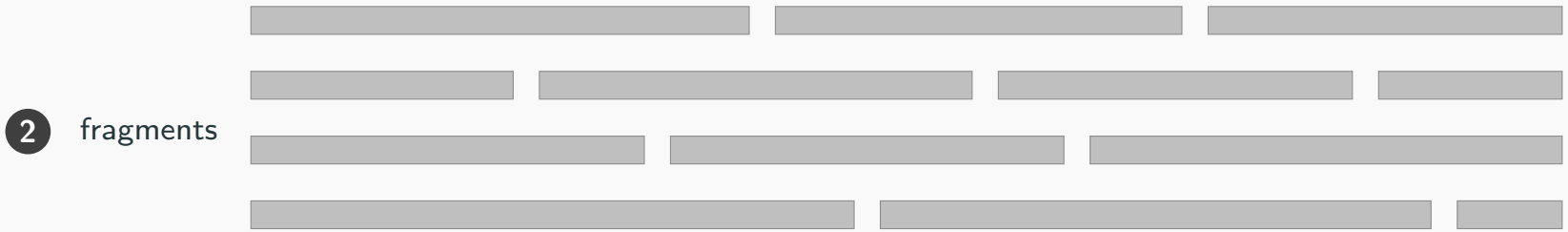
From reads to reference genomes

From DNA to reads



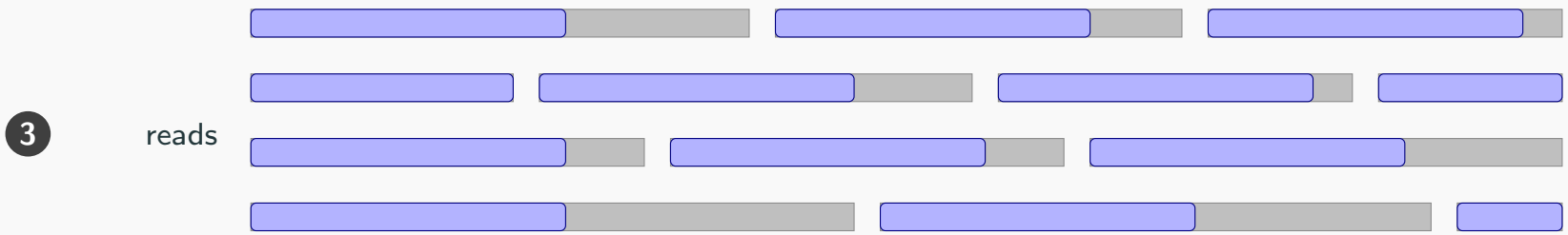
1 **DNA:** many copies of the genome (one per cell) are extracted

From DNA to reads



1. **DNA:** many copies of the genome (one per cell) are extracted
2. **fragmentation:** the DNA is broken at random places, differently in every copy

From DNA to reads



1. **DNA:** many copies of the genome (one per cell) are extracted
2. **fragmentation:** the DNA is broken at random places, differently in every copy
3. **sequencing:** the machine reads the start of each fragment (blue): short reads

The reads overlap because every position is covered by many fragments. But the machine only outputs the letters: **where each read came from is no longer known.**

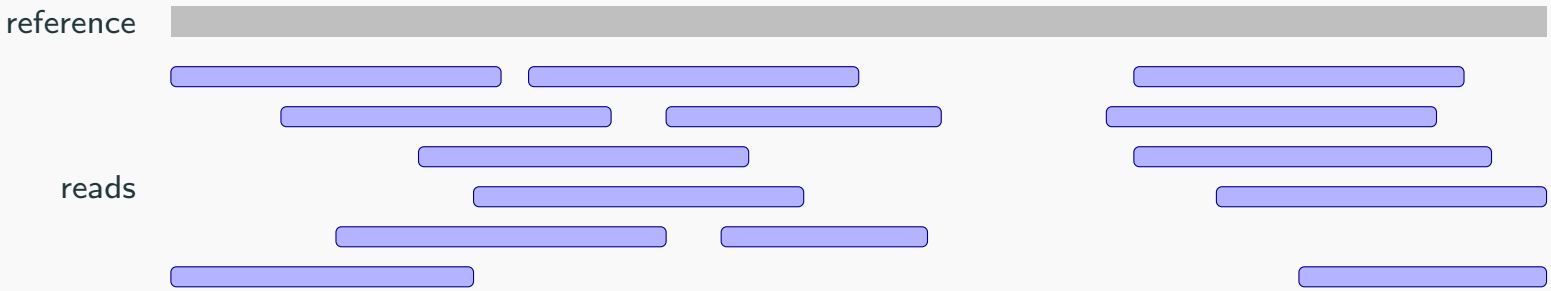
Genome assembly: mapping reads to a reference

With a reference genome, each short read is placed where its letters match:



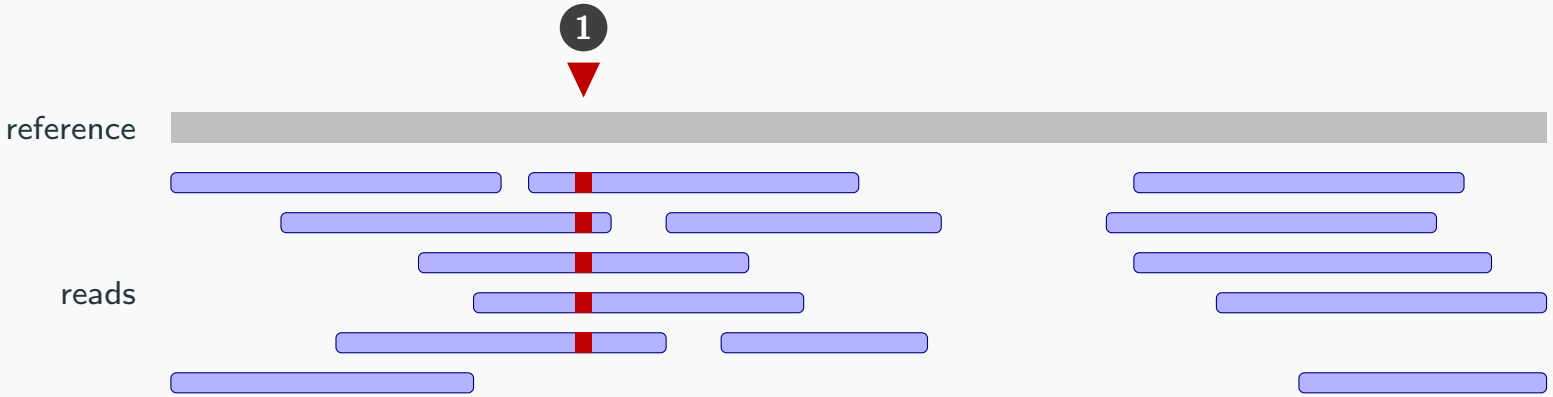
Genome assembly: mapping reads to a reference

With a reference genome, each short read is placed where its letters match:



Genome assembly: mapping reads to a reference

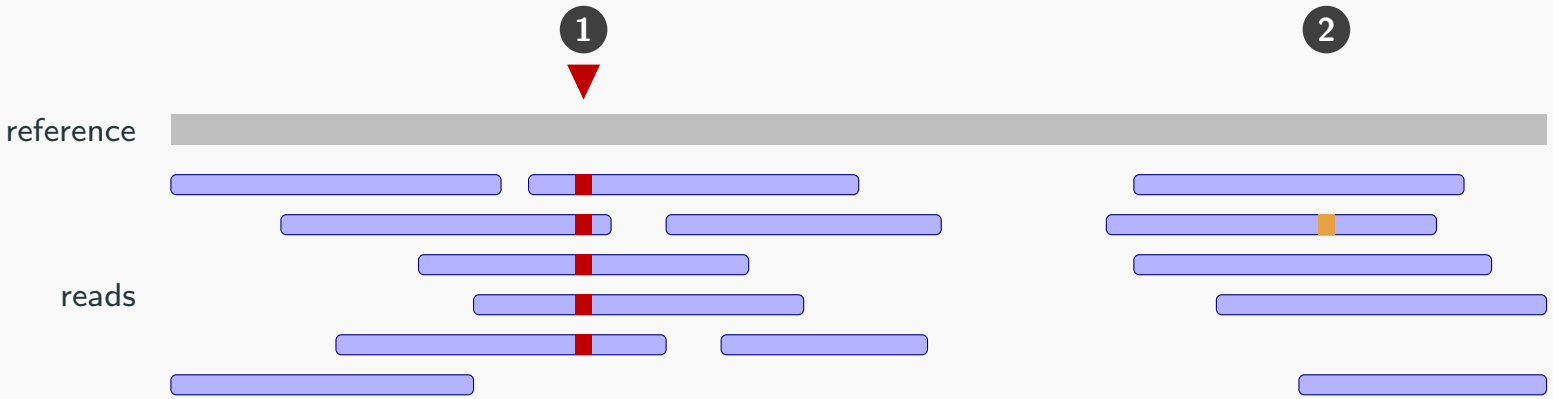
With a reference genome, each short read is placed where its letters match:



1 variant: all reads differ from the reference in the same way

Genome assembly: mapping reads to a reference

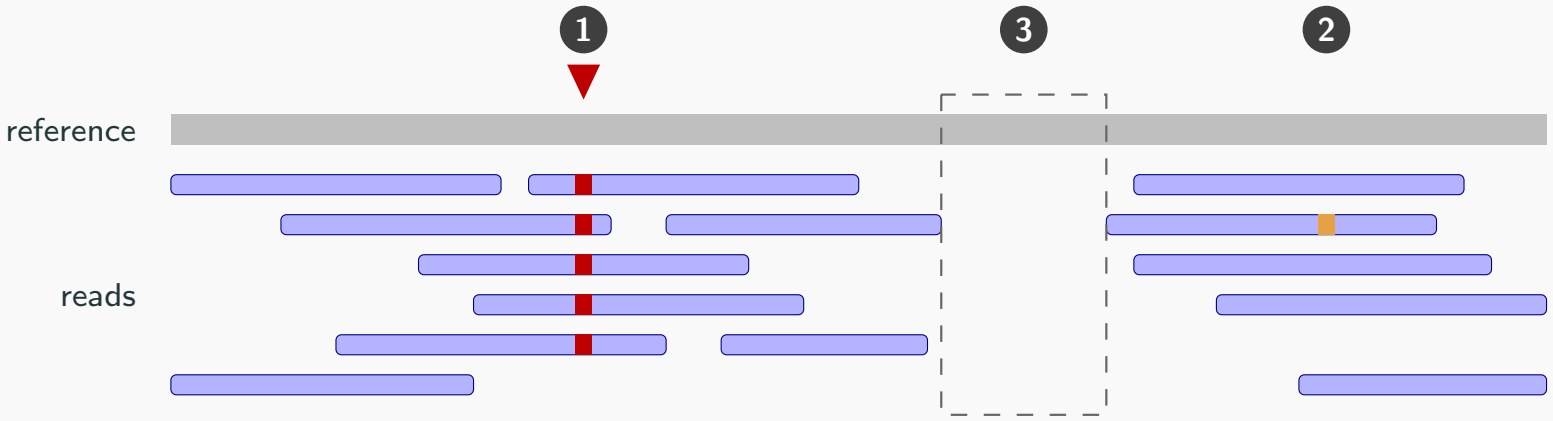
With a reference genome, each short read is placed where its letters match:



1. **variant:** all reads differ from the reference in the same way
2. **sequencing error:** only one read differs, so it is outvoted

Genome assembly: mapping reads to a reference

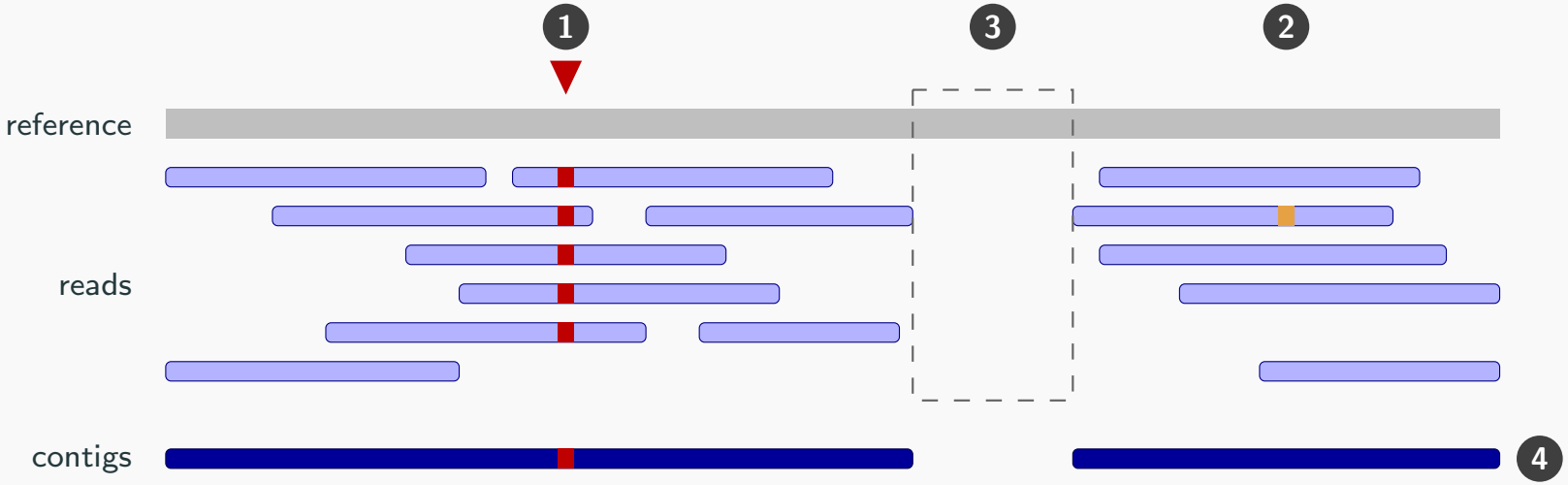
With a reference genome, each short read is placed where its letters match:



1 **variant:** all reads differ from the reference in the same way
2 **sequencing error:** only one read differs, so it is outvoted
3 **no reads:** sequence deleted in this person, or too repetitive to map

Genome assembly: mapping reads to a reference

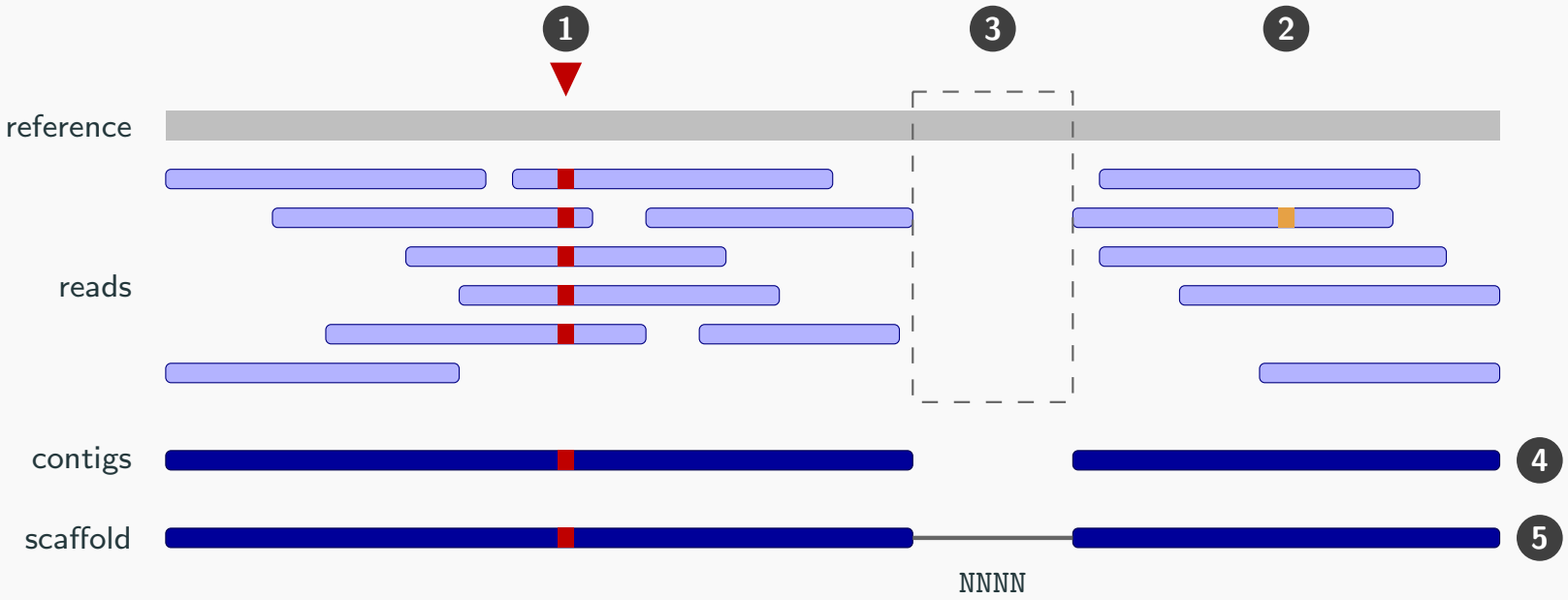
With a reference genome, each short read is placed where its letters match:



1 **variant:** all reads differ from the reference in the same way
2 **sequencing error:** only one read differs, so it is outvoted
3 **no reads:** sequence deleted in this person, or too repetitive to map
4 **contigs:** the consensus of the reads, continuous where reads overlap

Genome assembly: mapping reads to a reference

With a reference genome, each short read is placed where its letters match:



1 **variant:** all reads differ from the reference in the same way
2 **sequencing error:** only one read differs, so it is outvoted
3 **no reads:** sequence deleted in this person, or too repetitive to map
4 **contigs:** the consensus of the reads, continuous where reads overlap
5 **scaffold:** contigs in the order of the reference, gaps filled with `N`

What makes a good genome assembly?

- **Completeness:** Are all genomic regions represented?
- **Contiguity:** Are sequences assembled into long, continuous contigs/scaffolds?
- **Accuracy:** Are bases and structures correctly assembled (no misjoins or gaps)?
- **Purity:** Is the assembly free from contamination or mixed species?

Quality depends on:

- sequencing technology (read length, error rate)
- genome features (repeats, GC content, ploidy)
- assembly algorithms and polishing steps

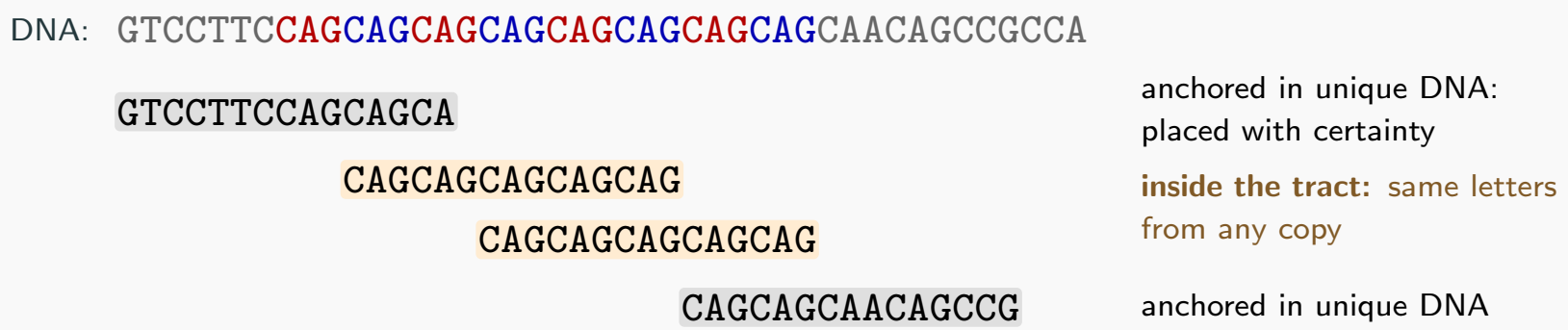
Why repeats are hard: counting copies (1/3)

Example: the CAG repeat in the huntingtin (*HTT*) gene. The tract is CAG copied again and again, with unique DNA on both sides (red/blue = alternating copies). Sequencing cuts the DNA into short reads, here 15 letters each:

DNA: GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA

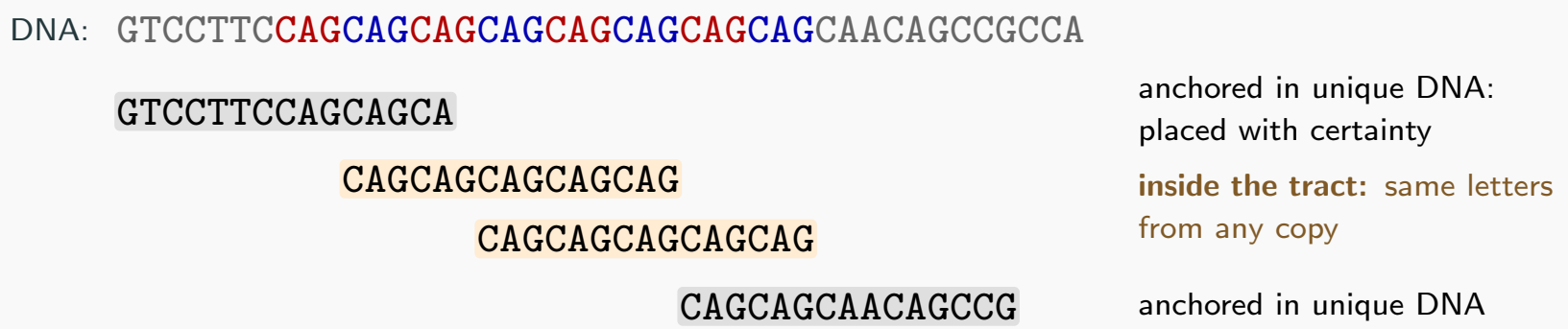
Why repeats are hard: counting copies (1/3)

Example: the CAG repeat in the huntingtin (*HTT*) gene. The tract is CAG copied again and again, with unique DNA on both sides (red/blue = alternating copies). Sequencing cuts the DNA into short reads, here 15 letters each:



Why repeats are hard: counting copies (1/3)

Example: the CAG repeat in the huntingtin (*HTT*) gene. The tract is CAG copied again and again, with unique DNA on both sides (red/blue = alternating copies). Sequencing cuts the DNA into short reads, here 15 letters each:



Problem: the two anchored reads only say that *some* CAG tract lies between them. The orange reads all say `CAGCAGCAGCAGCAG`, which fits anywhere in a tract of 8, 12 or 40 copies. The reads contain no count (next slide: what the assembler makes of this).

Why repeats are hard: counting copies (2/3)

Person 1 has 8 CAG copies, person 2 has 12. They give the *same* short reads (previous slide):

person 1 GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA

person 2 GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA

Why repeats are hard: counting copies (2/3)

Person 1 has 8 CAG copies, person 2 has 12. They give the *same* short reads (previous slide):

person 1 GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA

person 2 GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA



assembly from short reads: **same result for both**

assembled GTCCTTCCAGCAGCAGCAGCAACAGCCGCCA

only 4 copies: the repeat is **collapsed**

The assembler builds **one** sequence from the reads. Because the reads cannot say how long the tract is, it keeps **too few copies** (here 4; the exact number is arbitrary) or leaves a gap. Both people look alike, and the real length difference, which can separate healthy from Huntington's disease, is lost.

Why repeats are hard: counting copies (3/3)

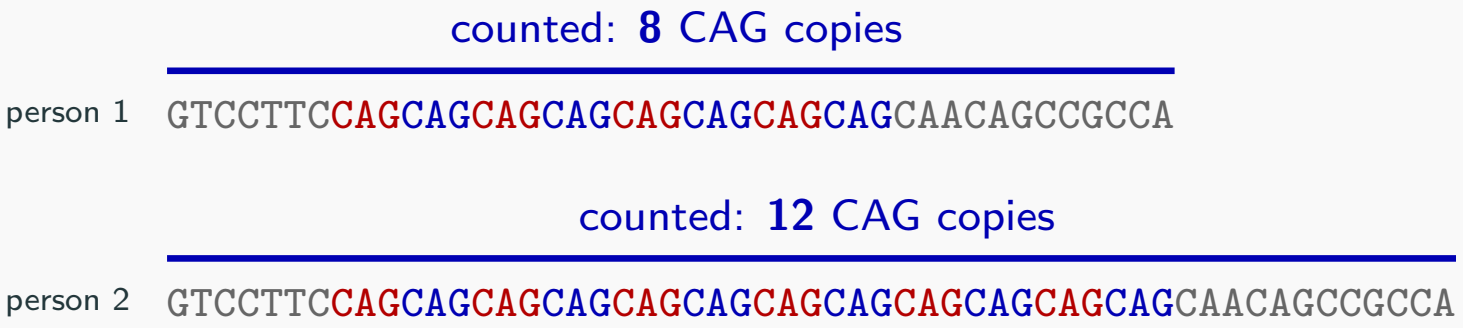
The same region in two people: the unique flanks are identical, only the number of CAG copies differs. Healthy people have about 10–35 copies, Huntington's disease is linked to about 40 or more (shortened here).

person 1 GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA

person 2 GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCA

Why repeats are hard: counting copies (3/3)

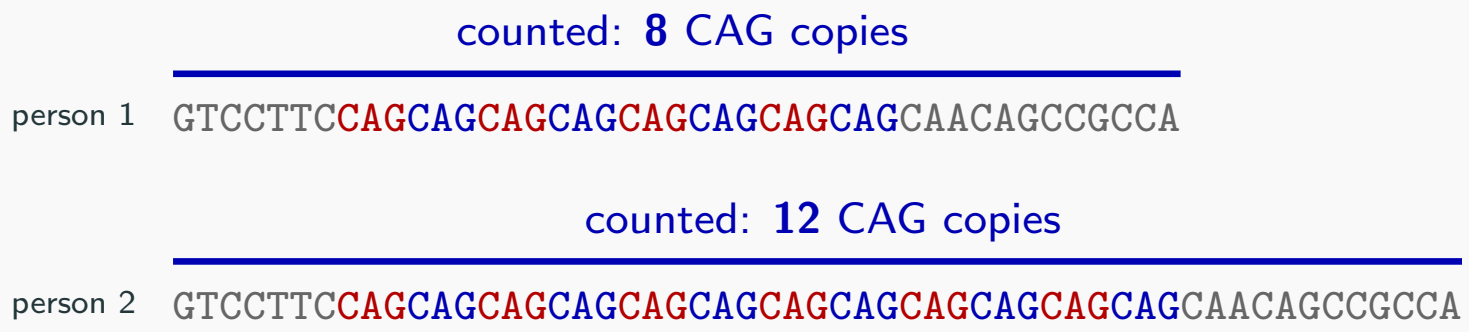
The same region in two people: the unique flanks are identical, only the number of CAG copies differs. Healthy people have about 10–35 copies, Huntington's disease is linked to about 40 or more (shortened here).



A **long read** (blue) that starts in the left flank and ends in the right flank contains the whole tract, so the copies can simply be counted. 40 copies are 120 letters, more than a 100–150 letter short read can span; long reads (thousands of letters) can.

Why repeats are hard: counting copies (3/3)

The same region in two people: the unique flanks are identical, only the number of CAG copies differs. Healthy people have about 10–35 copies, Huntington's disease is linked to about 40 or more (shortened here).



A **long read** (blue) that starts in the left flank and ends in the right flank contains the whole tract, so the copies can simply be counted. 40 copies are 120 letters, more than a 100–150 letter short read can span; long reads (thousands of letters) can.

Take-away: a repeat longer than the read cannot be resolved, because the read has no unique DNA to anchor the count.

The cost of reading a genome

Year	Approx. cost of one human genome
2001	\$100,000,000
2007	\$10,000,000
2015	\$4,000
today	below \$1,000

Rounded values based on the NHGRI genome sequencing cost data. After 2007 the cost fell faster than Moore's law.

Sequencing is cheap now; **turning reads into a trustworthy genome is the hard part.**

The Human Genome Project

Ran from 1990 to 2003 and cost nearly **$3 billion**. Produced the first human reference genome, about **92% complete**, using *Sanger sequencing*.

What Was Still Missing

- **Gaps and repeats:** Large regions of heterochromatin, centromeres, and telomeres were unresolved. Short read sequencing could not handle long, repetitive regions.
- **Representation:** The reference came from a few individuals, not a complete picture of human diversity.
- **Continuity:** Some chromosomes contained gaps and uncertain joins; the assembly was fragmented.

The Legacy

- Provided the foundation for all modern genome assembly and annotation methods.
- Highlighted that even a $3B, 13-year effort could not yield a truly complete genome.

The reference genome is not a person

- About 70% of the original reference came from a single anonymous donor; the rest from a handful of others.
- **2022, Telomere-to-Telomere (T2T-CHM13):** the first gapless human genome. It had **no Y chromosome**; a complete Y followed in 2023.
- **2023, Human Pangenome Reference:** a draft built from 47 individuals, because one linear reference cannot represent the variation among people.

So which sequence is "the human genome"? (cf. our definitions slide)

One Organism, Many Genomes

Cell variation, programmed rearrangements, other genomes, microbiome

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Counter-examples in humans:

- **Red blood cells:** no DNA at all (nucleus expelled, mitochondria lost); platelets too. Birds and fish keep the nucleus.

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Counter-examples in humans:

- **Red blood cells:** no DNA at all (nucleus expelled, mitochondria lost); platelets too. Birds and fish keep the nucleus.
- **Polyploid cells:** cardiomyocytes (4n), hepatocytes ($\sim$50% are 4n or 8n), megakaryocytes (16n), trophoblast giant cells (up to $\sim$1000n in rodents)

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Counter-examples in humans:

- **Red blood cells:** no DNA at all (nucleus expelled, mitochondria lost); platelets too. Birds and fish keep the nucleus.
- **Polyploid cells:** cardiomyocytes (4n), hepatocytes ($\sim$50% are 4n or 8n), megakaryocytes (16n), trophoblast giant cells (up to $\sim$1000n in rodents)

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Counter-examples in humans:

- **Red blood cells:** no DNA at all (nucleus expelled, mitochondria lost); platelets too. Birds and fish keep the nucleus.
- **Polyploid cells:** cardiomyocytes (4n), hepatocytes ($\sim$50% are 4n or 8n), megakaryocytes (16n), trophoblast giant cells (up to $\sim$1000n in rodents)

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Counter-examples in humans:

- **Red blood cells:** no DNA at all (nucleus expelled, mitochondria lost); platelets too. Birds and fish keep the nucleus.
- **Polyploid cells:** cardiomyocytes (4n), hepatocytes ($\sim$50% are 4n or 8n), megakaryocytes (16n), trophoblast giant cells (up to $\sim$1000n in rodents)
- **Neurons:** can carry somatic CNVs[1], causing subtle DNA differences

[1] **Copy number variation (CNV):** gain or loss of **chromosomal segments** ($\geq$ 1 kb).

Changes to the Amount of DNA in the Cell

Every cell descends from one zygote and starts with the same genome, **but mature cells can differ in the DNA they contain.**

Counter-examples in humans:

- **Red blood cells:** no DNA at all (nucleus expelled, mitochondria lost); platelets too. Birds and fish keep the nucleus.
- **Polyploid cells:** cardiomyocytes (4n), hepatocytes ($\sim$50% are 4n or 8n), megakaryocytes (16n), trophoblast giant cells (up to $\sim$1000n in rodents)
- **Neurons:** can carry somatic CNVs[1], causing subtle DNA differences
- **Cancer cells:** often aneuploid[2] or polyploid

[1] **Copy number variation (CNV):** gain or loss of **chromosomal segments** ($\geq$ 1 kb).

[2] **Aneuploidy:** gain or loss of **whole chromosomes**

Programmed genome rearrangement

Germline = cells that give rise to eggs and sperm (inherited). **Soma** = all other body cells (not inherited). Some organisms **deliberately** give them different DNA:

Programmed genome rearrangement

Germline = cells that give rise to eggs and sperm (inherited). **Soma** = all other body cells (not inherited). Some organisms **deliberately** give them different DNA:

- **Sea lamprey:** about 20% of the germline DNA ($\sim$0.5 Gb), including hundreds of germline genes, is lost from somatic cells in the early embryo.
  Why? Possibly genes needed for germ cells but harmful in body cells (e.g. cancer). Deletion is irreversible; silencing is not.

Programmed genome rearrangement

Germline = cells that give rise to eggs and sperm (inherited). **Soma** = all other body cells (not inherited). Some organisms **deliberately** give them different DNA:

- **Sea lamprey:** about 20% of the germline DNA ($\sim$0.5 Gb), including hundreds of germline genes, is lost from somatic cells in the early embryo.
  Why? Possibly genes needed for germ cells but harmful in body cells (e.g. cancer). Deletion is irreversible; silencing is not.

Programmed genome rearrangement

Germline = cells that give rise to eggs and sperm (inherited). **Soma** = all other body cells (not inherited). Some organisms **deliberately** give them different DNA:

- **Sea lamprey:** about 20% of the germline DNA ($\sim$0.5 Gb), including hundreds of germline genes, is lost from somatic cells in the early embryo.
  Why? Possibly genes needed for germ cells but harmful in body cells (e.g. cancer). Deletion is irreversible; silencing is not.
- **Ciliates** (*Oxytricha*): the most extreme case. One cell, two nuclei:
  - **micronucleus** (germline): silent and inherited; **macronucleus** (soma): transcribed, runs the cell

Programmed genome rearrangement

Germline = cells that give rise to eggs and sperm (inherited). **Soma** = all other body cells (not inherited). Some organisms **deliberately** give them different DNA:

- **Sea lamprey:** about 20% of the germline DNA (∼0.5 Gb), including hundreds of germline genes, is lost from somatic cells in the early embryo.
  Why? Possibly genes needed for germ cells but harmful in body cells (e.g. cancer). Deletion is irreversible; silencing is not.
- **Ciliates** (*Oxytricha*): the most extreme case. One cell, two nuclei:
  - **micronucleus** (germline): silent and inherited; **macronucleus** (soma): transcribed, runs the cell
  - after mating, the macronucleus is rebuilt from the micronucleus: ∼96% of the DNA is discarded, scrambled gene pieces are put in order, and the rest is cut into ∼16,000 gene-sized "nanochromosomes", each copied ∼2,000×

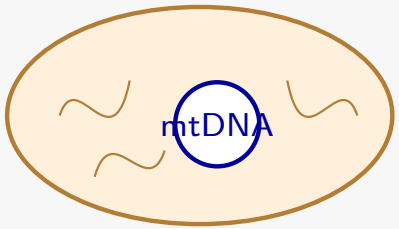
Programmed genome rearrangement

Germline = cells that give rise to eggs and sperm (inherited). **Soma** = all other body cells (not inherited). Some organisms **deliberately** give them different DNA:

- **Sea lamprey:** about 20% of the germline DNA (∼0.5 Gb), including hundreds of germline genes, is lost from somatic cells in the early embryo.
  Why? Possibly genes needed for germ cells but harmful in body cells (e.g. cancer). Deletion is irreversible; silencing is not.
- **Ciliates** (*Oxytricha*): the most extreme case. One cell, two nuclei:
  - **micronucleus** (germline): silent and inherited; **macronucleus** (soma): transcribed, runs the cell
  - after mating, the macronucleus is rebuilt from the micronucleus: ∼96% of the DNA is discarded, scrambled gene pieces are put in order, and the rest is cut into ∼16,000 gene-sized "nanochromosomes", each copied ∼2,000×
  - Why? Not fully known. One idea: it removes transposon-derived DNA from the working genome.

Other genomes in the body

Sources of DNA other than the nuclear genome



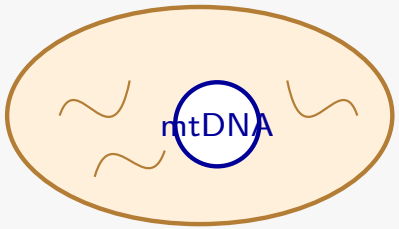
Organelles

Mitochondria and plastids descend from endosymbionts.

Human mtDNA: ∼16.6 kb, 37 genes, many copies per cell.

Other genomes in the body

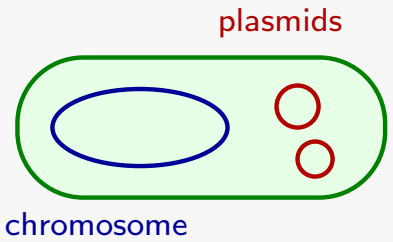
Sources of DNA other than the nuclear genome



Organelles

Mitochondria and plastids descend from endosymbionts.

Human mtDNA: ∼16.6 kb, 37 genes, many copies per cell.



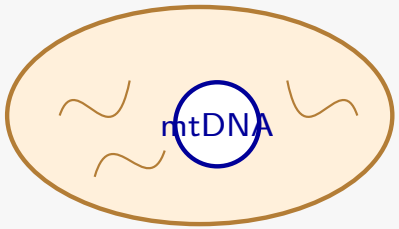
Plasmids

In bacteria: small, usually circular DNA, copied independently of the chromosome.

Often carries extra genes, e.g. antibiotic resistance.

Other genomes in the body

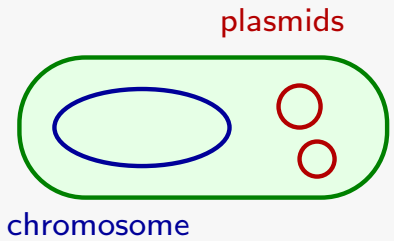
Sources of DNA other than the nuclear genome



Organelles

Mitochondria and plastids descend from endosymbionts.

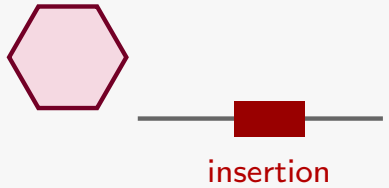
Human mtDNA: ∼16.6 kb, 37 genes, many copies per cell.



Plasmids

In bacteria: small, usually circular DNA, copied independently of the chromosome.

Often carries extra genes, e.g. antibiotic resistance.



Viruses, mobile DNA

Symbionts, parasites, viruses and mobile elements add DNA to a host.

E.g. endogenous retroviruses: ∼8% of the human genome.

The microbiome: a second genome you carry

The **microbiome** is the community of commensal, symbiotic and pathogenic microorganisms living on surfaces and in body cavities.

How many cells?

Reference adult (70 kg man). Sender, Fuchs and Milo (2016), *PLOS Biology*.

The microbiome: a second genome you carry

The **microbiome** is the community of commensal, symbiotic and pathogenic microorganisms living on surfaces and in body cavities.

How many cells?

Myth: 10 : 1

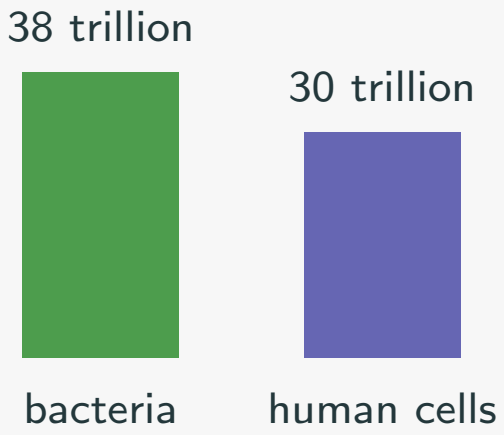
Reference adult (70 kg man). Sender, Fuchs and Milo (2016), *PLOS Biology*.

The microbiome: a second genome you carry

The **microbiome** is the community of commensal, symbiotic and pathogenic microorganisms living on surfaces and in body cavities.

How many cells?

~~**Myth:** 10 : 1~~



Revised: about 1 : 1

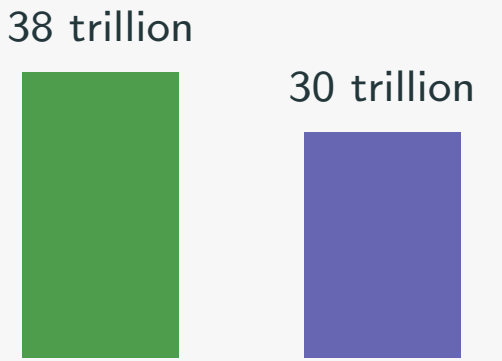
Reference adult (70 kg man). Sender, Fuchs and Milo (2016), *PLOS Biology*.

The microbiome: a second genome you carry

The **microbiome** is the community of commensal, symbiotic and pathogenic microorganisms living on surfaces and in body cavities.

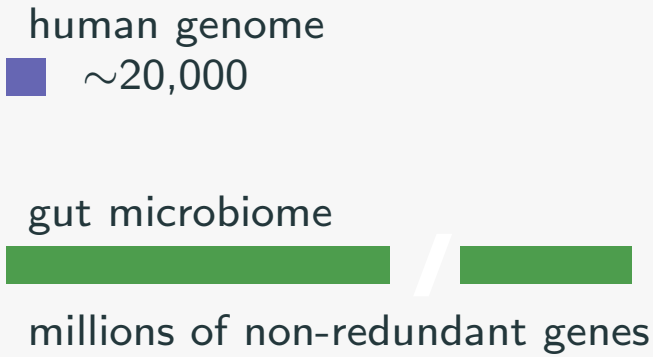
How many cells?

~~**Myth:** 10 : 1~~



Revised: about 1 : 1

How many genes?



By gene count the microbiome still wins.

Reference adult (70 kg man). Sender, Fuchs and Milo (2016), *PLOS Biology*.

Take-home messages

1. There is **no single definition** of "genome".

Take-home messages

1. There is **no single definition** of "genome".
2. **Size is not complexity**: repeats and duplications drive the variation.

Take-home messages

1. There is **no single definition** of "genome".
2. **Size is not complexity**: repeats and duplications drive the variation.
3. Most of the human genome is **not protein-coding**, and "functional" depends on how you define it.

Take-home messages

1. There is **no single definition** of "genome".
2. **Size is not complexity**: repeats and duplications drive the variation.
3. Most of the human genome is **not protein-coding**, and "functional" depends on how you define it.
4. A reference genome is **a model**, not a person.

Take-home messages

1. There is **no single definition** of "genome".
2. **Size is not complexity**: repeats and duplications drive the variation.
3. Most of the human genome is **not protein-coding**, and "functional" depends on how you define it.
4. A reference genome is **a model**, not a person.
5. **One organism can have many genomes**: across cells, between germline and soma, and beyond the nucleus.

Take-home messages

1. There is **no single definition** of "genome".
2. **Size is not complexity**: repeats and duplications drive the variation.
3. Most of the human genome is **not protein-coding**, and "functional" depends on how you define it.
4. A reference genome is **a model**, not a person.
5. **One organism can have many genomes**: across cells, between germline and soma, and beyond the nucleus.

Next: how genes are switched on and off in these genomes (gene regulation).

Literature (1/3): genomes, size and the tree of life

- Goldman and Landweber (2016), *What is a genome?*, PLoS Genetics 12(7): e1006181
- Williams et al. (2013), *An archaeal origin of eukaryotes supports only two primary domains of life*, Nature 504: 231–236
- Eme et al. (2023), *Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes*, Nature 618: 992–999
- Hidalgo et al. (2017), *Is there an upper limit to genome size?*, Trends in Plant Science 22: 567–573
- Fernández et al. (2024), *A 160 Gbp fork fern genome shatters size record for eukaryotes*, iScience 27(6): 109889
- Meyer et al. (2021), *Giant lungfish genome elucidates the conquest of land by vertebrates*, Nature 590: 284–289
- Otto (2007), *The evolutionary consequences of polyploidy*, Cell 131: 452–462
- Comai (2005), *The advantages and disadvantages of being polyploid*, Nature Reviews Genetics 6: 836–846

Literature (2/3): inside the genome, reading it

- International Human Genome Sequencing Consortium (2001), *Initial sequencing and analysis of the human genome*, Nature 409: 860–921
- ENCODE Project Consortium (2012), *An integrated encyclopedia of DNA elements in the human genome*, Nature 489: 57–74
- Graur et al. (2013), *On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE*, Genome Biology and Evolution 5: 578–590
- Doolittle (2013), *Is junk DNA bunk? A critique of ENCODE*, PNAS 110: 5294–5300
- Rands et al. (2014), *8.2% of the human genome is constrained*, PLoS Genetics 10(7): e1004525
- Lettice et al. (2003), *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly*, Human Molecular Genetics 12: 1725–1735
- Mi et al. (2000), *Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis*, Nature 403: 785–789
- Blond et al. (2000), *An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor*, J. Virology 74: 3321–3329
- Nurk et al. (2022), *The complete sequence of a human genome*, Science 376: 44–53
- Liao et al. (2023), *A draft human pangenome reference*, Nature 617: 312–324

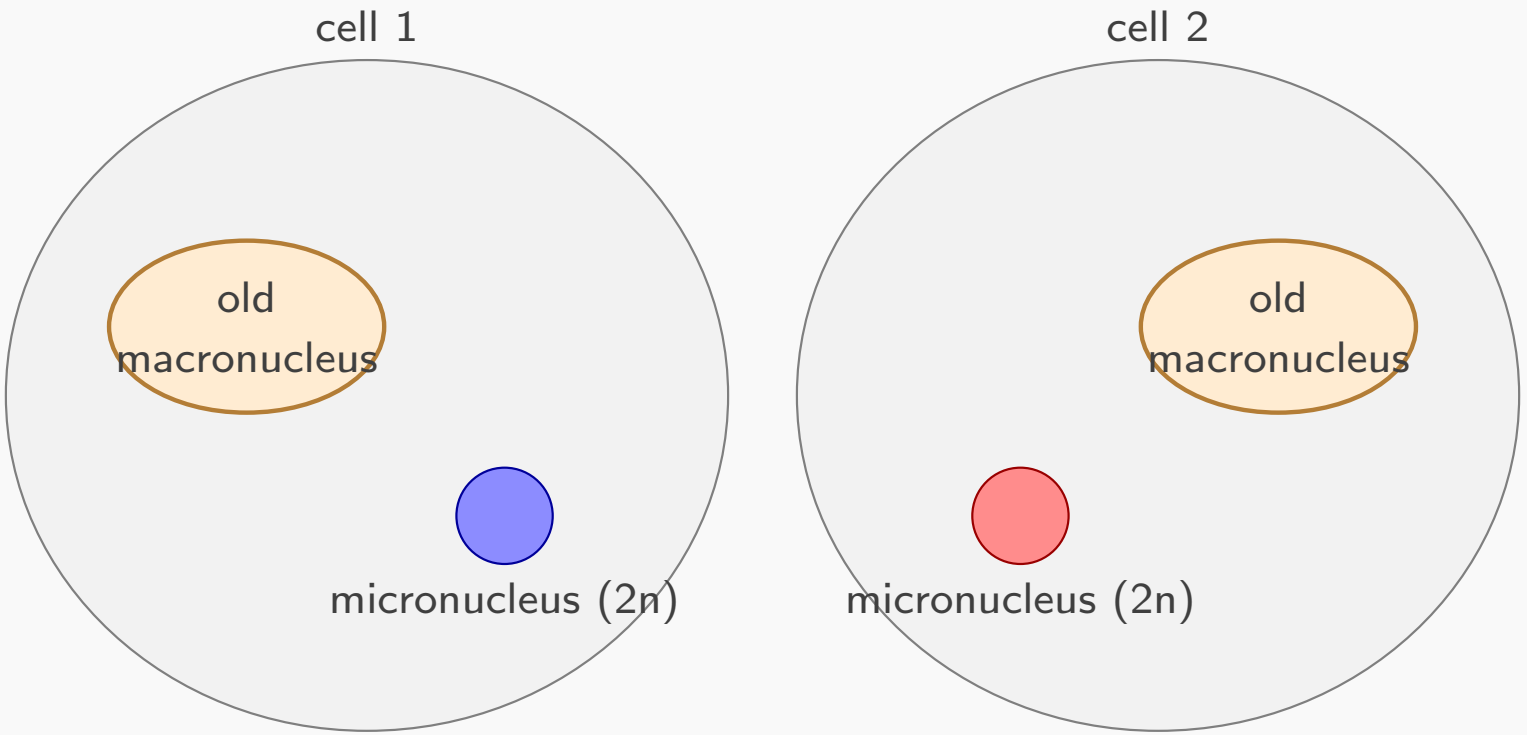
Literature (3/3): k-mers, one organism with many genomes

- Marçais and Kingsford (2011), *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers*, Bioinformatics 27: 764–770 (Jellyfish)
- Ranallo-Benavidez et al. (2020), *GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes*, Nature Communications 11: 1432
- Wood, Lu and Langmead (2019), *Improved metagenomic analysis with Kraken 2*, Genome Biology 20: 257
- Ondov et al. (2016), *Mash: fast genome and metagenome distance estimation using MinHash*, Genome Biology 17: 132
- Compeau, Pevzner and Tesler (2011), *How to apply de Bruijn graphs to genome assembly*, Nature Biotechnology 29: 987–991
- Sender, Fuchs and Milo (2016), *Revised estimates for the number of human and bacteria cells in the body*, PLoS Biology 14(8): e1002533
- Smith et al. (2018), *The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution*, Nature Genetics 50: 270–277
- Kinsella et al. (2019), *Programmed DNA elimination of germline development genes in songbirds*, Nature Communications 10 (doi:10.1038/s41467-019-13427-4)
- Swart et al. (2013), *The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes*, PLoS Biology 11(1): e1001473

Appendix slides

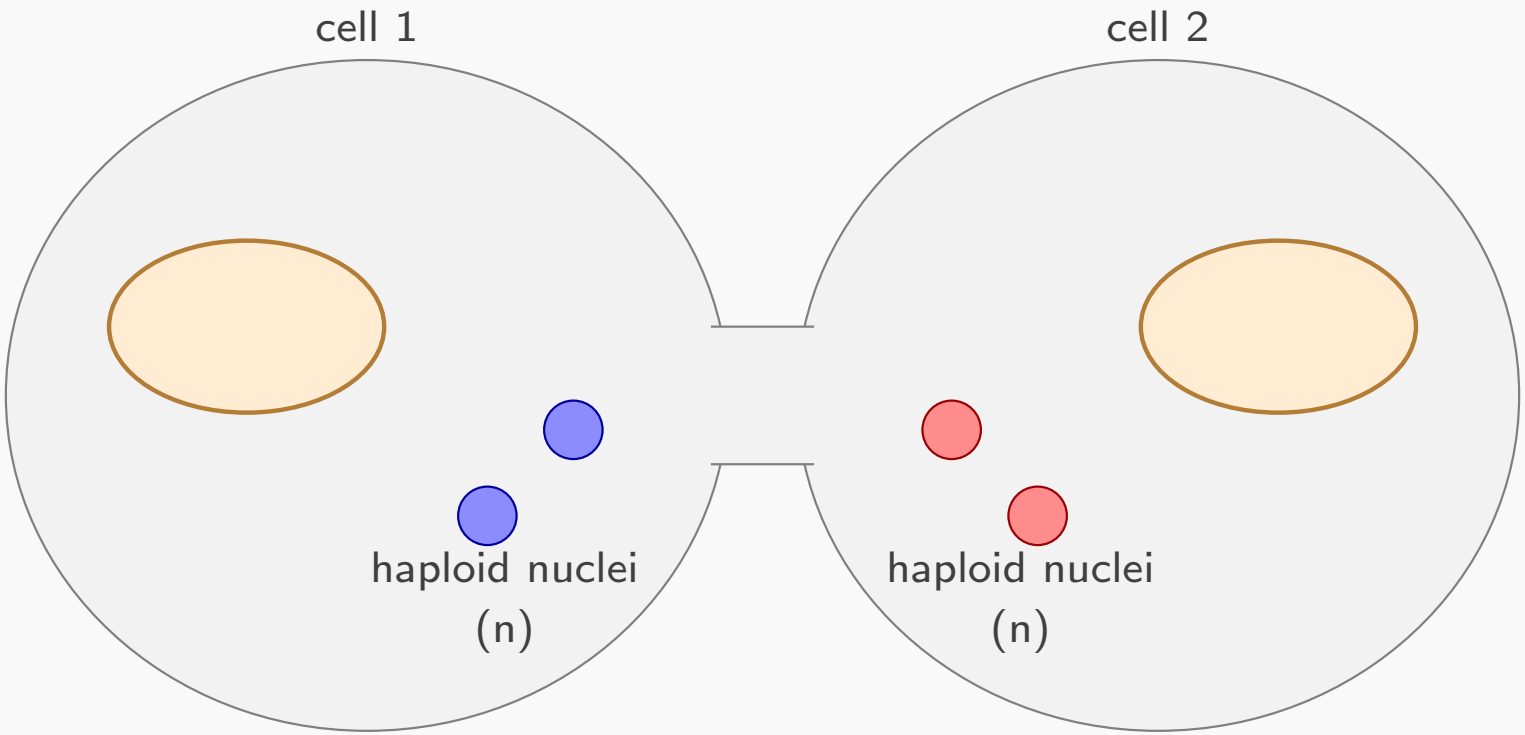
Backup material

Ciliates: mating and a new working nucleus



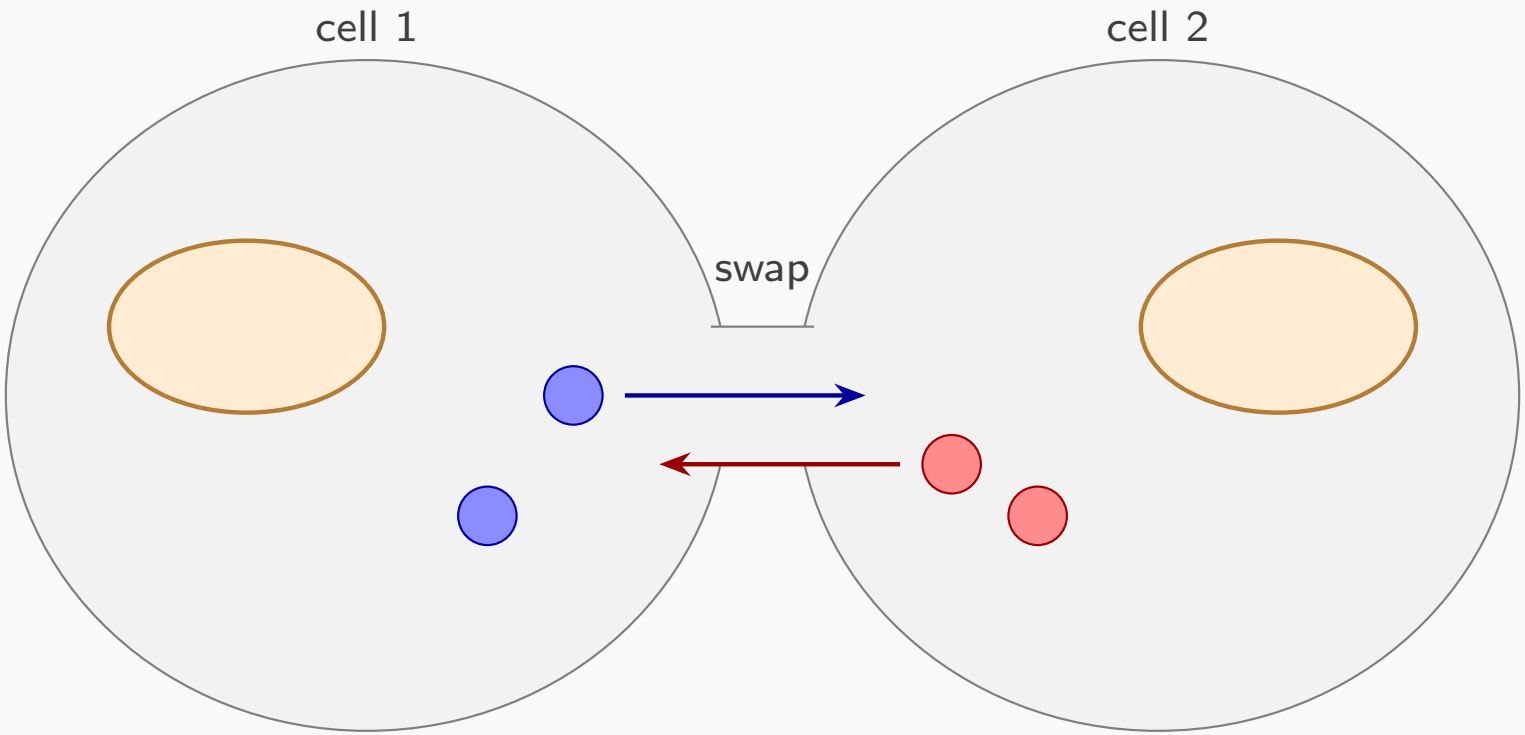
1. Two cells, each with an old macronucleus (working genome) and a micronucleus (germline).

Ciliates: mating and a new working nucleus



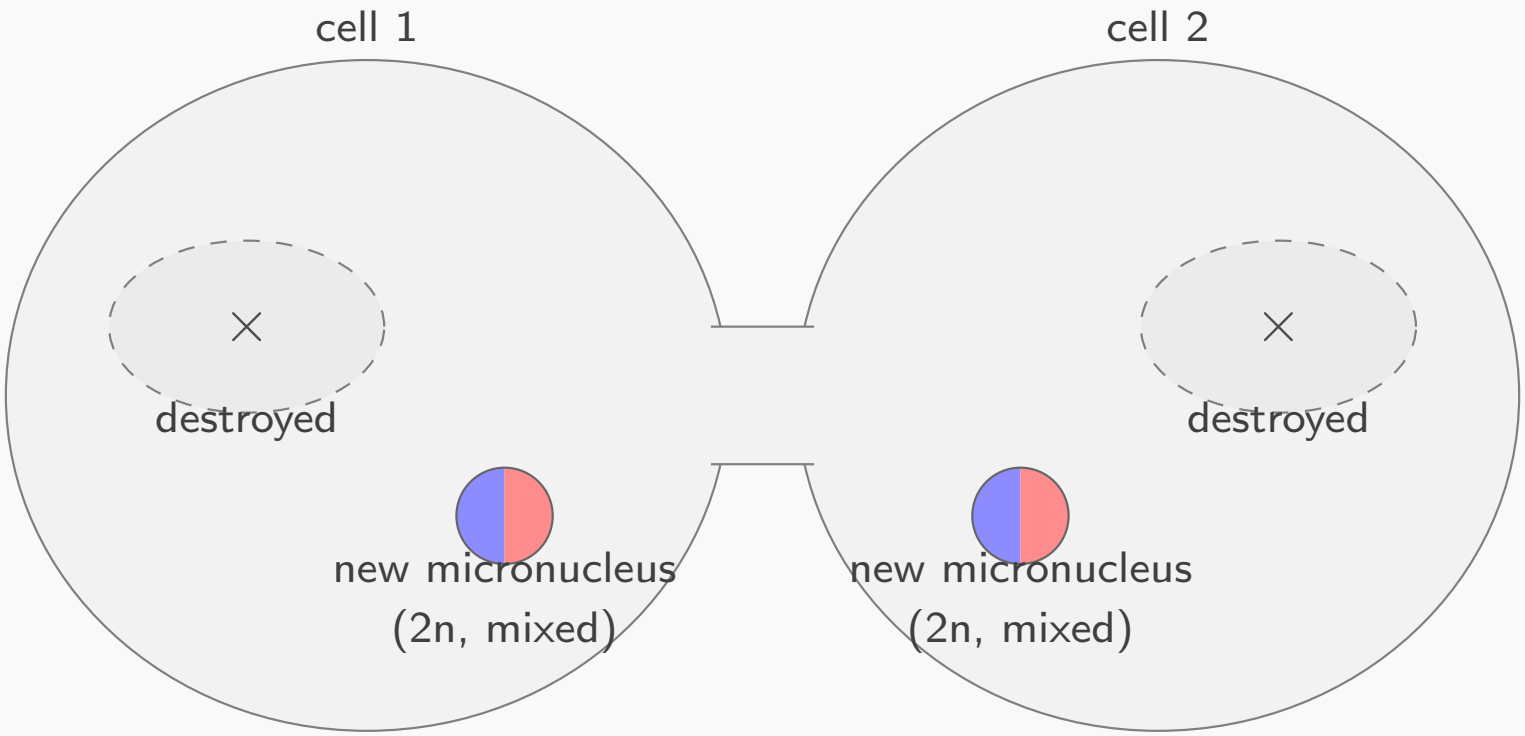
2. The cells join. Each micronucleus undergoes meiosis and gives haploid nuclei.

Ciliates: mating and a new working nucleus



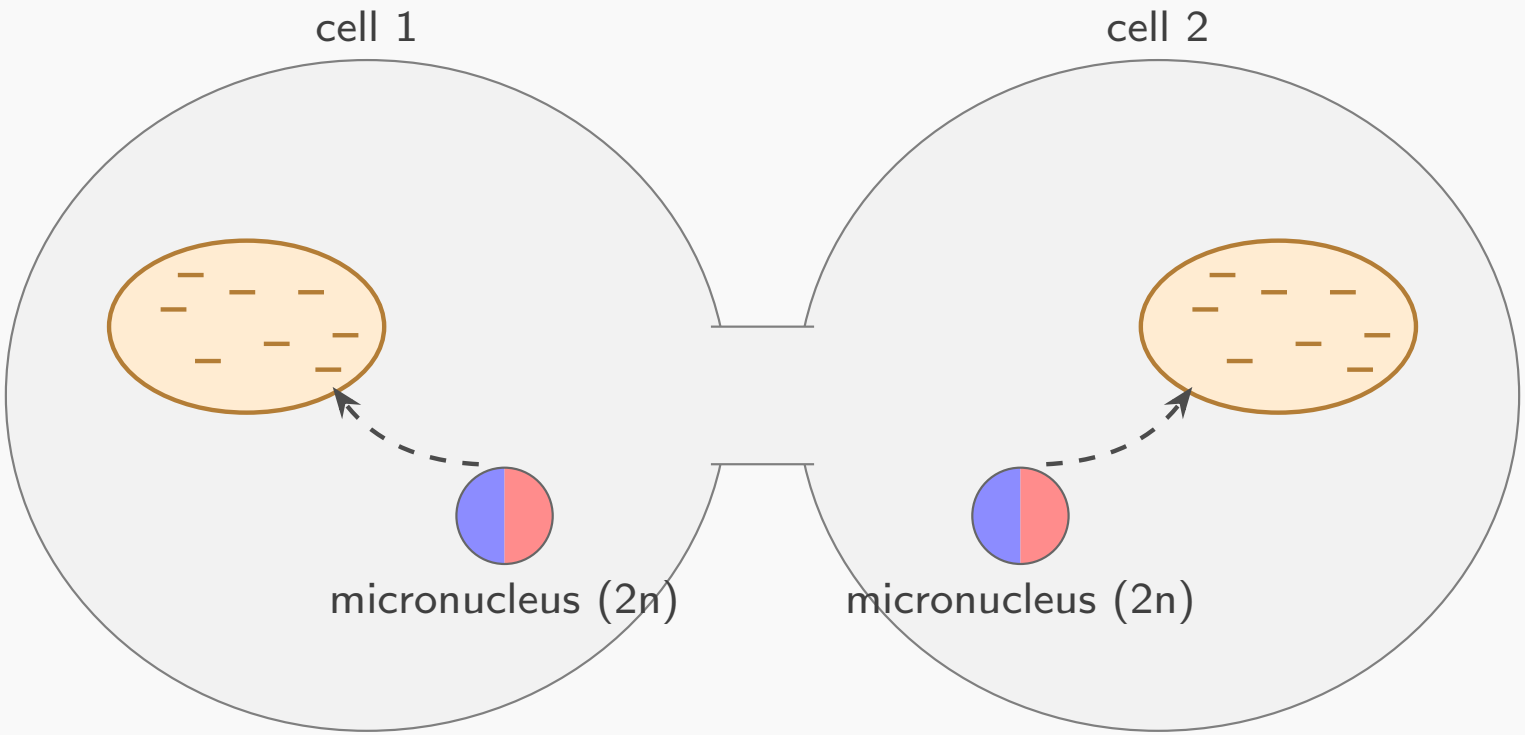
3. Each cell passes one haploid nucleus to the other.

Ciliates: mating and a new working nucleus



4. Swapped and resident nuclei fuse into a new, mixed micronucleus. The old macronucleus is destroyed.

Ciliates: mating and a new working nucleus



5. A new macronucleus (orange, many small chromosomes) is built from a copy of the new micronucleus. Then the cells separate.

k-mers: cut a read into overlapping words

a read (length $L = 10$)

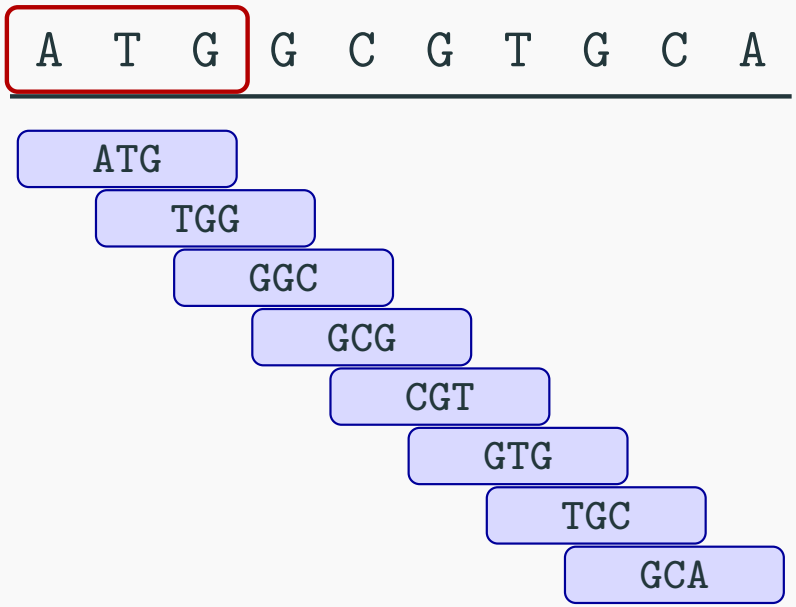
A T G G C G T G C A

- A **k-mer** is a word of length k.
- Slide a window of k letters along the read, one base at a time (here $k = 3$).
- A read of length L gives $L - k + 1$ k-mers: $10 - 3 + 1 = 8$.

Why $k \approx 21$? Too short: every k-mer occurs everywhere by chance (only 4^k possible words). Too long: one sequencing error spoils k k-mers.

k-mers: cut a read into overlapping words

a read (length $L = 10$)

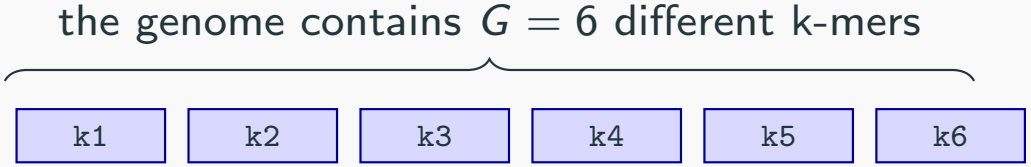


- A **k-mer** is a word of length k.
- Slide a window of k letters along the read, one base at a time (here $k = 3$).
- A read of length L gives $L - k + 1$ k-mers: $10 - 3 + 1 = 8$.

Why $k \approx 21$? Too short: every k-mer occurs everywhere by chance (only 4^k possible words). Too long: one sequencing error spoils k k-mers.

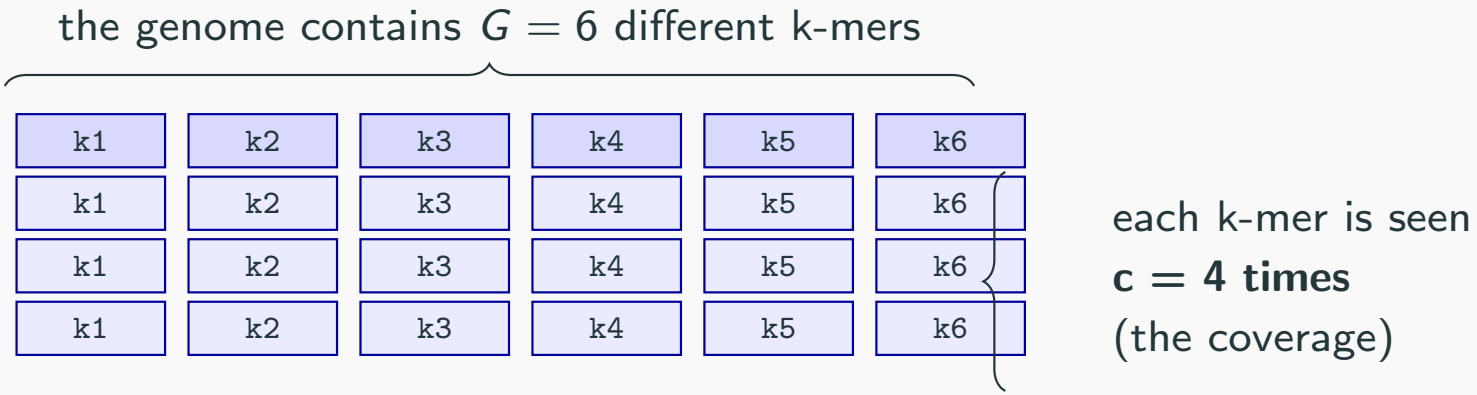
k-mers: why counting them tells you the genome size

Sequence the genome many times over: every part of it is read again and again.



k-mers: why counting them tells you the genome size

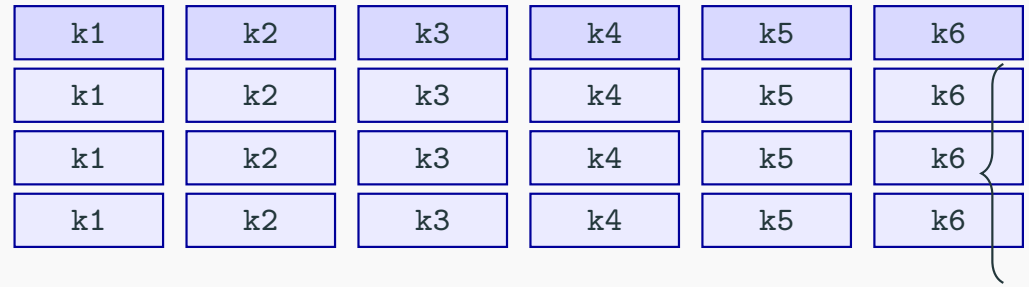
Sequence the genome many times over: every part of it is read again and again.



k-mers: why counting them tells you the genome size

Sequence the genome many times over: every part of it is read again and again.

the genome contains $G = 6$ different k-mers



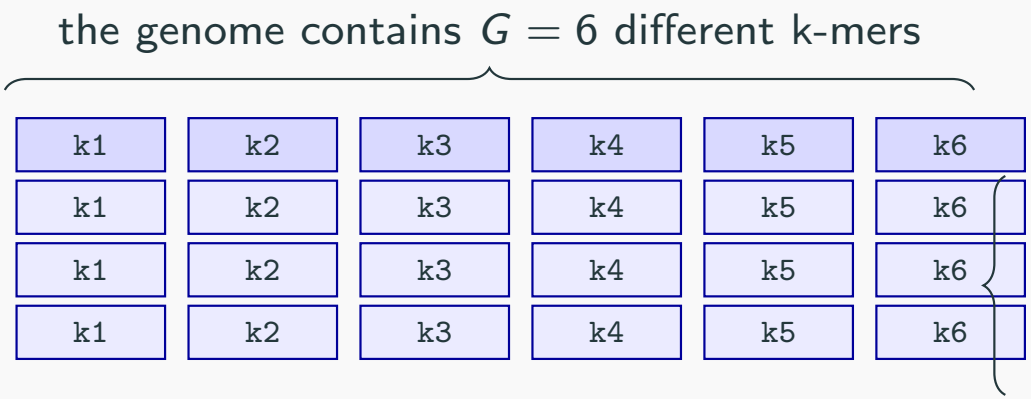
each k-mer is seen
c = 4 times
(the coverage)

Total k-mers counted: $N_k = G \times c = 6 \times 4 = 24$

$$\Rightarrow \quad G = \frac{N_k}{c} = \frac{24}{4} = 6$$

k-mers: why counting them tells you the genome size

Sequence the genome many times over: every part of it is read again and again.

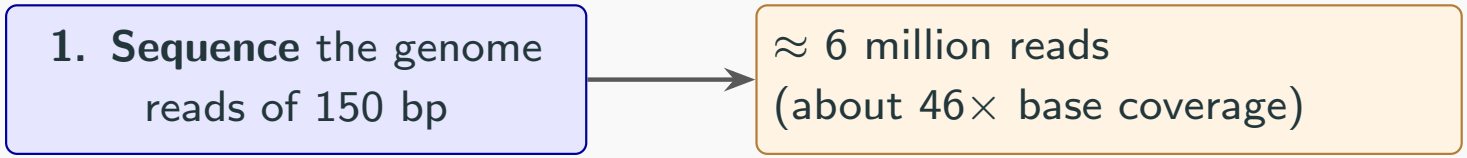


Total k-mers counted: $N_k = G \times c = 6 \times 4 = 24$

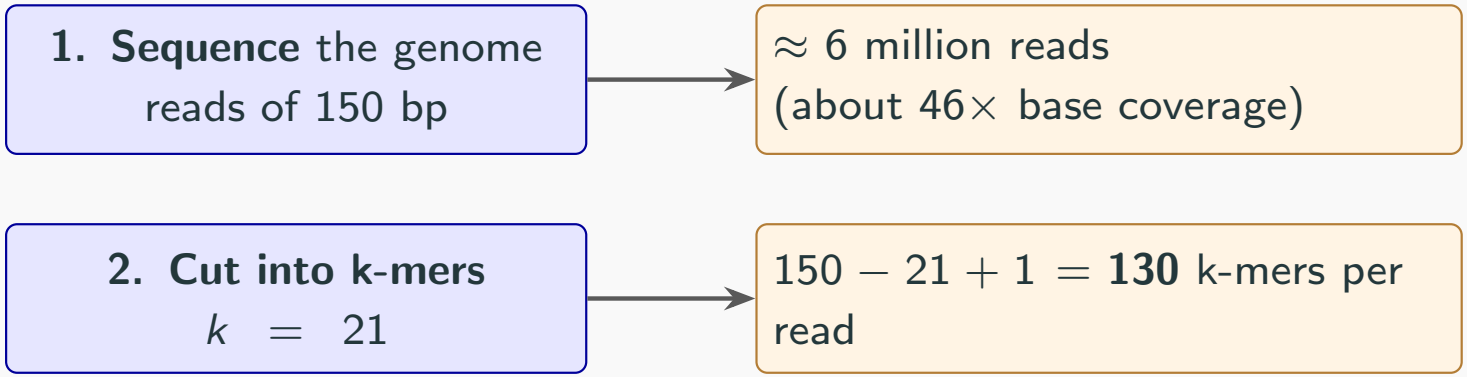
$$\Rightarrow \quad G = \frac{N_k}{c} = \frac{24}{4} = 6$$

In real data you don't see the stacks, but you can measure both numbers: N_k by counting, c from a histogram

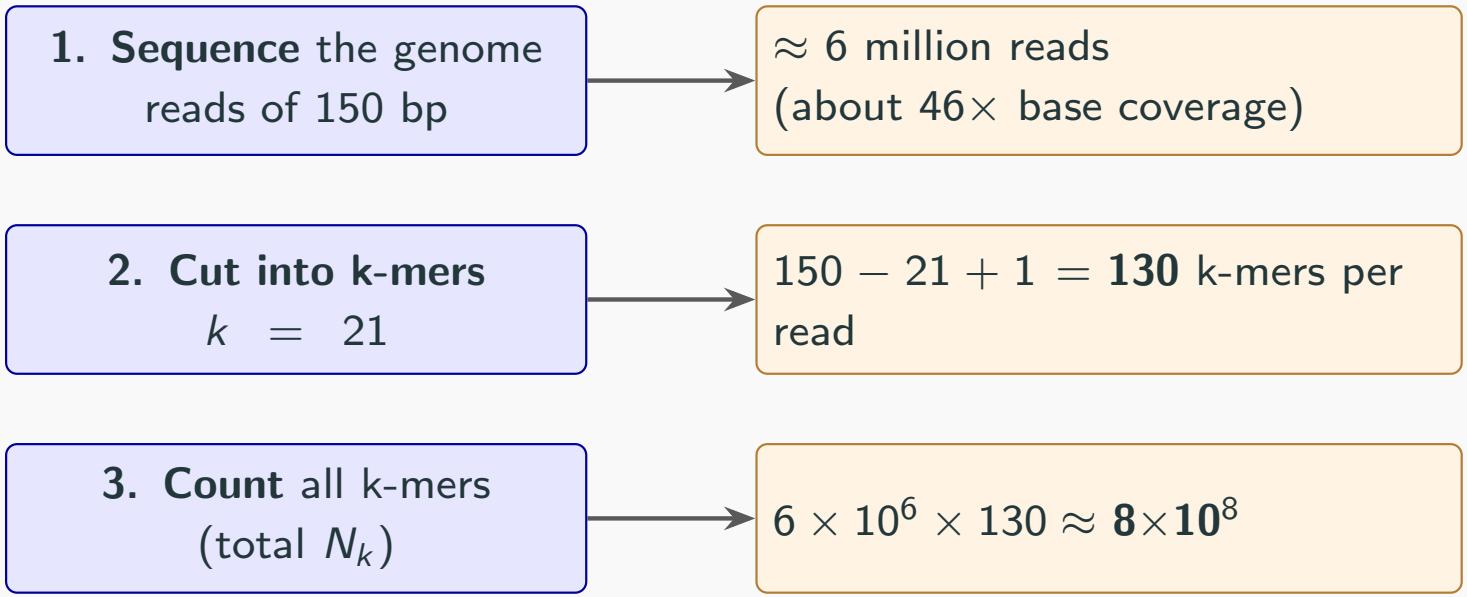
k-mers: a worked example



k-mers: a worked example



k-mers: a worked example



k-mers: a worked example

1. Sequence the genome
reads of 150 bp

→ $\approx$ 6 million reads
(about 46$\times$ base coverage)

2. Cut into k-mers
$k = 21$

→ $150 - 21 + 1 = \mathbf{130}$ k-mers per read

3. Count all k-mers
(total N_k)

→ $6 \times 10^6 \times 130 \approx \mathbf{8 \times 10^8}$

4. Histogram: how
often is a k-mer seen?

→ peak at $c = 40$: $G \approx \frac{8 \times 10^8}{40}$
$= 2 \times 10^7$ bp $= \mathbf{20\ Mb}$

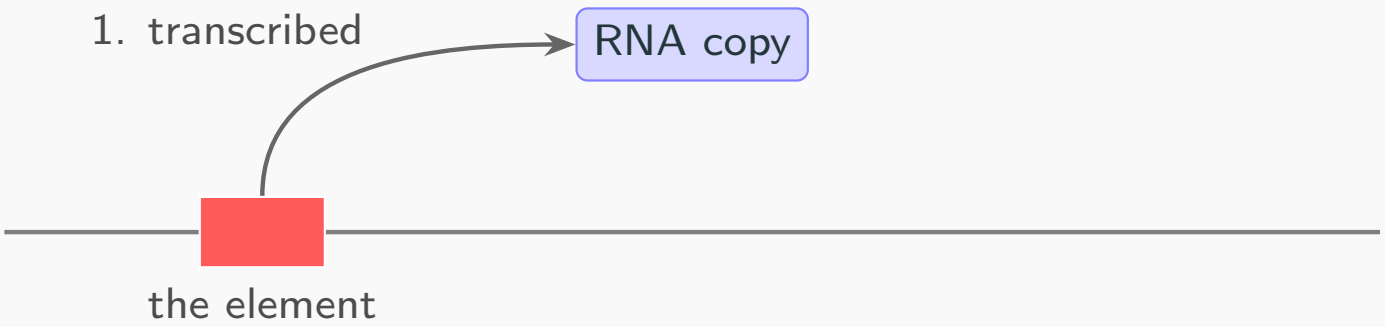
Transposable elements: the genome's copy machines

Transposable elements (TEs) are DNA sequences that can make new copies of themselves elsewhere in the genome.



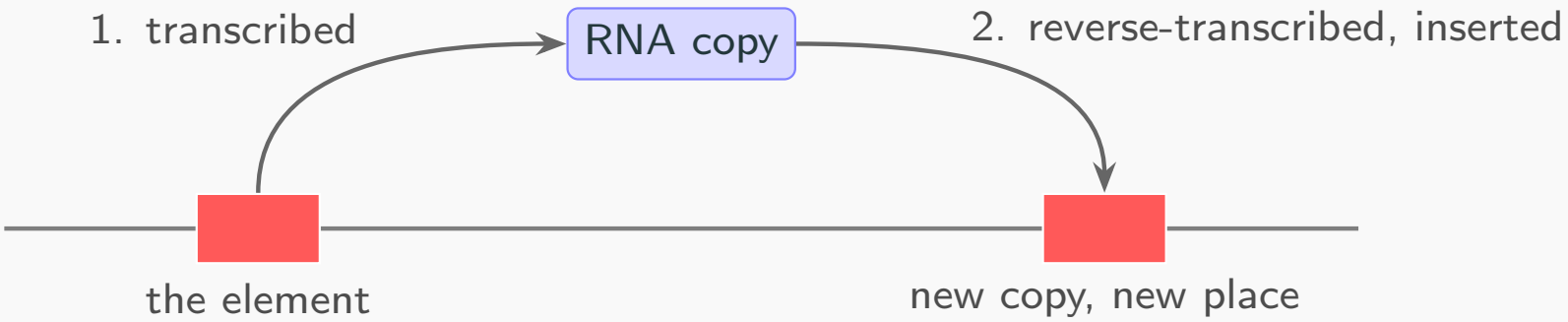
Transposable elements: the genome's copy machines

Transposable elements (TEs) are DNA sequences that can make new copies of themselves elsewhere in the genome.



Transposable elements: the genome's copy machines

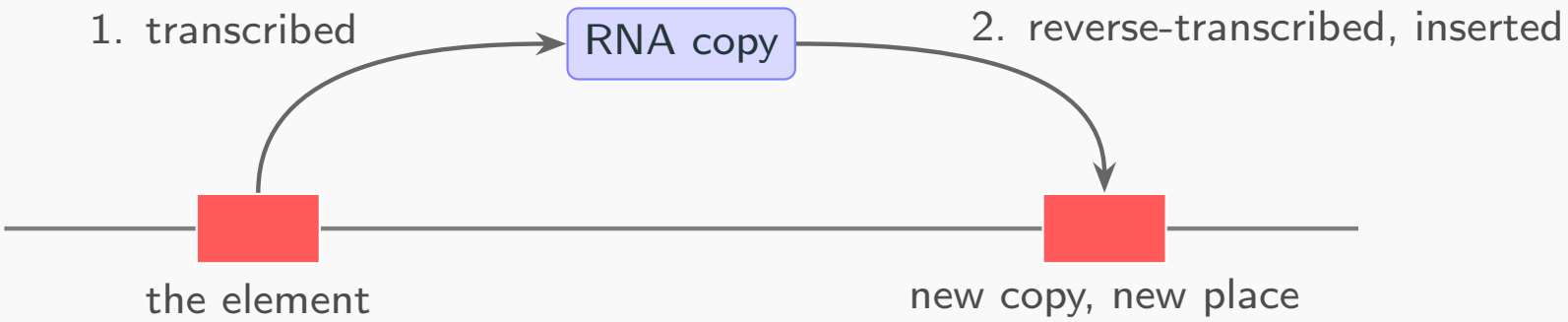
Transposable elements (TEs) are DNA sequences that can make new copies of themselves elsewhere in the genome.



"Copy and paste": the original stays, so the number of copies grows. (Some TEs, the DNA transposons, "cut and paste": the element itself moves.)

Transposable elements: the genome's copy machines

Transposable elements (TEs) are DNA sequences that can make new copies of themselves elsewhere in the genome.



"Copy and paste": the original stays, so the number of copies grows. (Some TEs, the DNA transposons, "cut and paste": the element itself moves.)

Most copies are old and dead (truncated or mutated). New insertions are rare, but can break a gene or add a regulatory sequence.

Transposable elements in the human genome

- **L1 (LINE-1):** $\sim$6 kb, $\sim$17% of the genome. Encodes its own copying enzyme (reverse transcriptase). Only a small fraction of copies is still active.

Transposable elements in the human genome

- **L1 (LINE-1):** $\sim$6 kb, $\sim$17% of the genome. Encodes its own copying enzyme (reverse transcriptase). Only a small fraction of copies is still active.
- **Alu (a SINE):** $\sim$300 bp, $\sim$10% of the genome, over a million copies. Has no enzyme of its own and borrows L1's.

Transposable elements in the human genome

- **L1 (LINE-1):** ∼6 kb, ∼17% of the genome. Encodes its own copying enzyme (reverse transcriptase). Only a small fraction of copies is still active.
- **Alu (a SINE):** ∼300 bp, ∼10% of the genome, over a million copies. Has no enzyme of its own and borrows L1's.
- **Endogenous retroviruses (ERVs):** ∼8%. Remnants of retroviruses.

Transposable elements in the human genome

- **L1 (LINE-1):** $\sim$6 kb, $\sim$17% of the genome. Encodes its own copying enzyme (reverse transcriptase). Only a small fraction of copies is still active.
- **Alu (a SINE):** $\sim$300 bp, $\sim$10% of the genome, over a million copies. Has no enzyme of its own and borrows L1's.
- **Endogenous retroviruses (ERVs):** $\sim$8%. Remnants of retroviruses.
- **DNA transposons:** $\sim$3%. Cut and paste. All inactive in humans.

Transposable elements in the human genome

- **L1 (LINE-1):** ∼6 kb, ∼17% of the genome. Encodes its own copying enzyme (reverse transcriptase). Only a small fraction of copies is still active.
- **Alu (a SINE):** ∼300 bp, ∼10% of the genome, over a million copies. Has no enzyme of its own and borrows L1's.
- **Endogenous retroviruses (ERVs):** ∼8%. Remnants of retroviruses.
- **DNA transposons:** ∼3%. Cut and paste. All inactive in humans.

Together, TEs make up roughly half of our DNA. **Because they are so similar to each other, they are hard to assemble** (see "Why repeats are hard").