Genome alignment and assembly

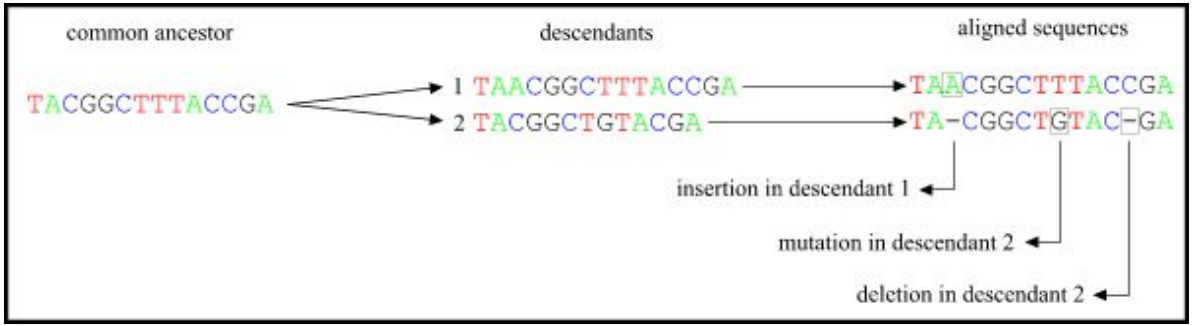
Genome Biology (BIOL7263)
12Sept24

* this lecture includes materials adapted from Ryan Chikhi's, Antoine Limasset's and Camille Marchet's presentations at the 2024 Workshop on Genomics

1. Alignment
- Why?
- How?
- Limitation and Biases

2. Assembly
- What and Why?
- Kmers and the de Bruijn graph
- Coverage and quality
- Limitation and technical solutions

Part 1 - Alignment - Why align sequences?



- Reveal evolutionary relationship
- Variant calling
- taxonomic classification
- Quantify gene expression - RNA-seq quantification
- Identify epigenetic modification - ATAC-seq

Why align sequences?

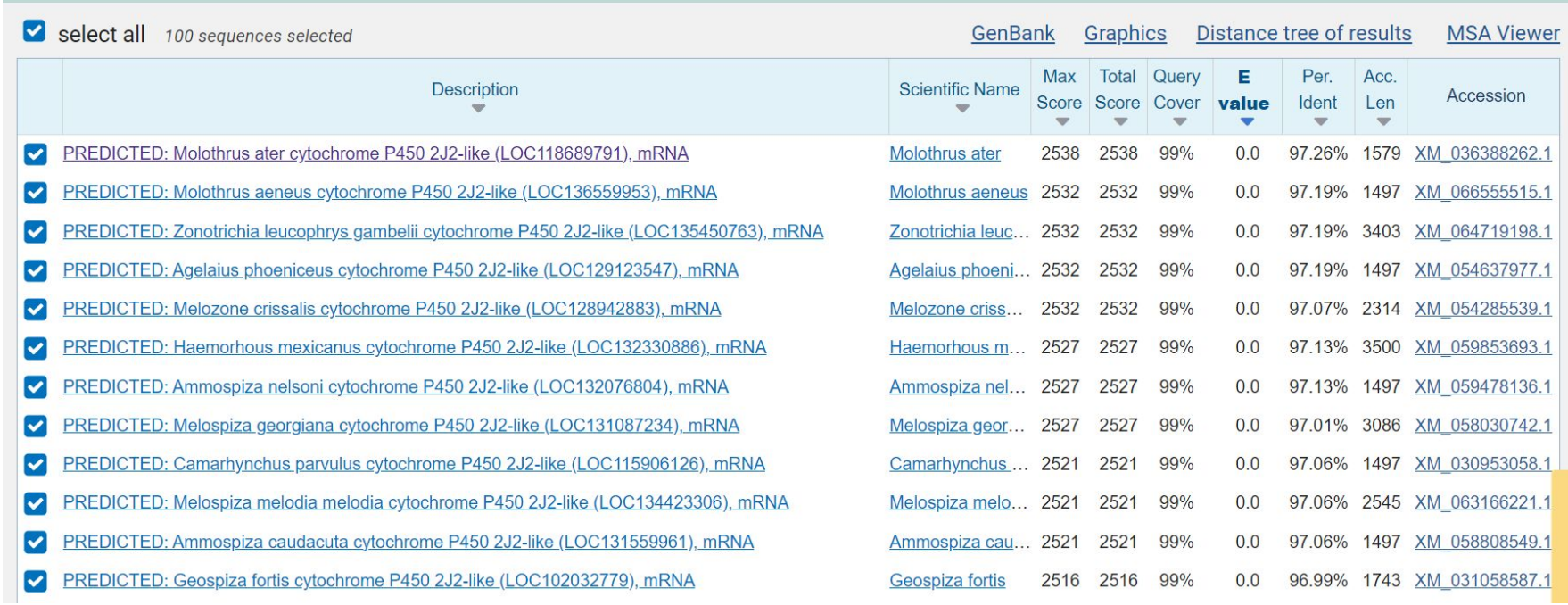
Types of alignment

Pairwise - 2 sequences

https://blast.ncbi.nlm.nih.gov

Types of alignment

1 sequence vs. database

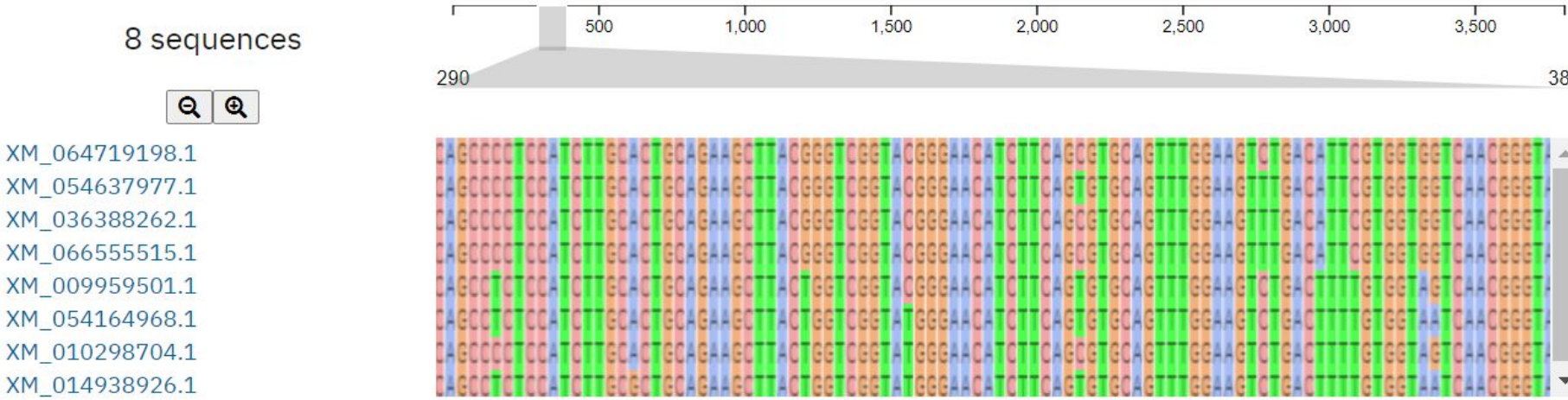


select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Molothrus ater cytochrome P450 2J2-like (LOC118689791), mRNA	Molothrus ater	2538	2538	99%	0.0	97.26%	1579	XM_036388262.1
PREDICTED: Molothrus aeneus cytochrome P450 2J2-like (LOC136559953), mRNA	Molothrus aeneus	2532	2532	99%	0.0	97.19%	1497	XM_066555515.1
PREDICTED: Zonotrichia leucophrys gambelii cytochrome P450 2J2-like (LOC135450763), mRNA	Zonotrichia leuc…	2532	2532	99%	0.0	97.19%	3403	XM_064719198.1
PREDICTED: Agelaius phoeniceus cytochrome P450 2J2-like (LOC129123547), mRNA	Agelaius phoeni…	2532	2532	99%	0.0	97.19%	1497	XM_054637977.1
PREDICTED: Melozone crissalis cytochrome P450 2J2-like (LOC128942883), mRNA	Melozone criss…	2532	2532	99%	0.0	97.07%	2314	XM_054285539.1
PREDICTED: Haemorhous mexicanus cytochrome P450 2J2-like (LOC132330886), mRNA	Haemorhous m…	2527	2527	99%	0.0	97.13%	3500	XM_059853693.1
PREDICTED: Ammospiza nelsoni cytochrome P450 2J2-like (LOC132076804), mRNA	Ammospiza nel…	2527	2527	99%	0.0	97.13%	1497	XM_059478136.1
PREDICTED: Melospiza georgiana cytochrome P450 2J2-like (LOC131087234), mRNA	Melospiza geor…	2527	2527	99%	0.0	97.01%	3086	XM_058030742.1
PREDICTED: Camarhynchus parvulus cytochrome P450 2J2-like (LOC115906126), mRNA	Camarhynchus …	2521	2521	99%	0.0	97.06%	1497	XM_030953058.1
PREDICTED: Melospiza melodia melodia cytochrome P450 2J2-like (LOC134423306), mRNA	Melospiza melo…	2521	2521	99%	0.0	97.06%	2545	XM_063166221.1
PREDICTED: Ammospiza caudacuta cytochrome P450 2J2-like (LOC131559961), mRNA	Ammospiza cau…	2521	2521	99%	0.0	97.06%	1497	XM_058808549.1
PREDICTED: Geospiza fortis cytochrome P450 2J2-like (LOC102032779), mRNA	Geospiza fortis	2516	2516	99%	0.0	96.99%	1743	XM_031058587.1

Types of alignment

Multiple sequence alignment



https://www.ebi.ac.uk/jdispatcher/msa/clustalo

Types of alignment

Vs. a profile (motif)

Motif logo constructed from the ~3700 predicted Hox sites

What can we align?

DNA vs. DNA

RNA vs. RNA

RNA vs. DNA

Protein vs. protein

DNA vs. protein, RNA vs. protein

Alignment terminology

Query: sequence to align

Reference (or target): sequence to align to

Hit (or match or alignment): part of query aligned to part of reference

Homology: shared ancestry

Similarity, identity: mathematical ways to detect homology

String: sequence

Global vs. Local Alignment:

Global: must align all nucleotides, using insertions/deletions if necessary

Local: you're allowed to skip beginning and/or end of either sequence

Figure: Ganesh, Prakhar, et al. "Nucl2Vec: Local alignment of DNA sequences using Distributed Vector Representation." bioRxiv (2018): 401851.

There are many possible alignments, how do we chose among them?

- Penalize mismatches and gaps
- Search possible alignments to find ones that **minimize** the penalty

E.g. here a mismatch gives 1 penalty, a deletion gives 2 penalties:

```
ref:   TAC    GAT

query: TTC    G-T

penalty=1     penalty=2
```

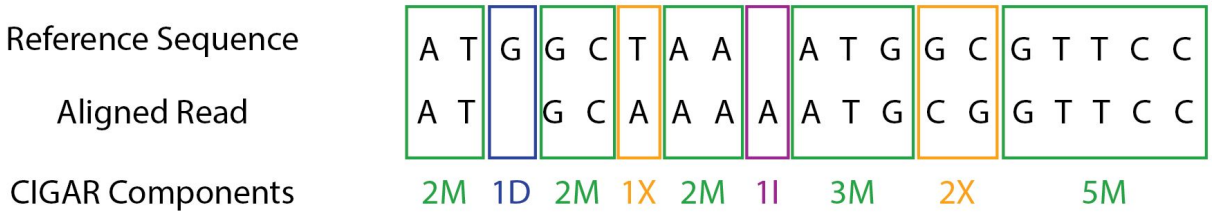
CIGAR strings ("Concise Idiosyncratic Gapped Alignment Report")

Commonly used format to encode alignments

M = match I = insertion (gap in the target sequence)

X = mismatch D = deletion (gap in the query sequence)

*note this may vary among programs



CIGAR String: 2M1D2M1X2M1I3M2X5M

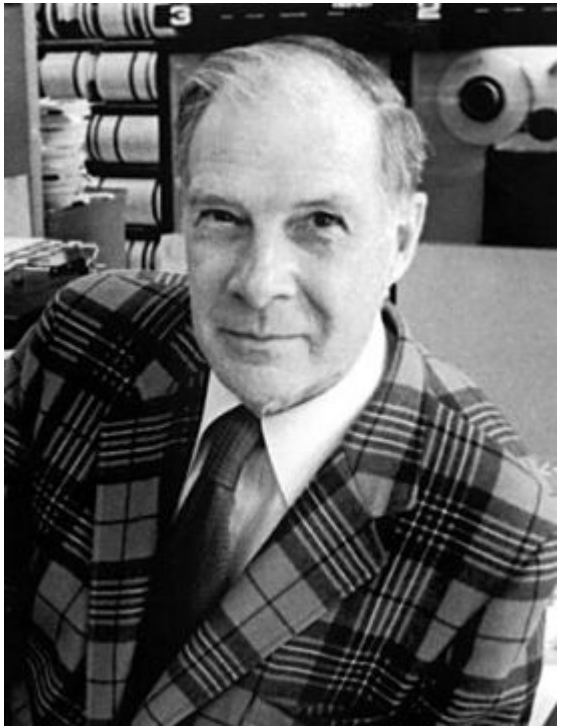
Hamming distance

Minimum number of substitutions needed to turn sequence A into sequence B. No insertions or deletions.

A - ACTAGATG

B - CGTACATG

- Important metric for designing synthetic sequences for barcoding applications e.g.



Bystrykh, Leonid V. "Generalized DNA barcode design based on Hamming codes." *PloS one* 7.5 (2012): e36852.

How do you find the best alignment? Smith-Waterman Algorithm (1981)

1. Allow gaps at beginning
2. Find the highest scoring cell
3. Trace it back to a zero

http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Smith-Waterman

AATCGATAGC
AACGAAAGC

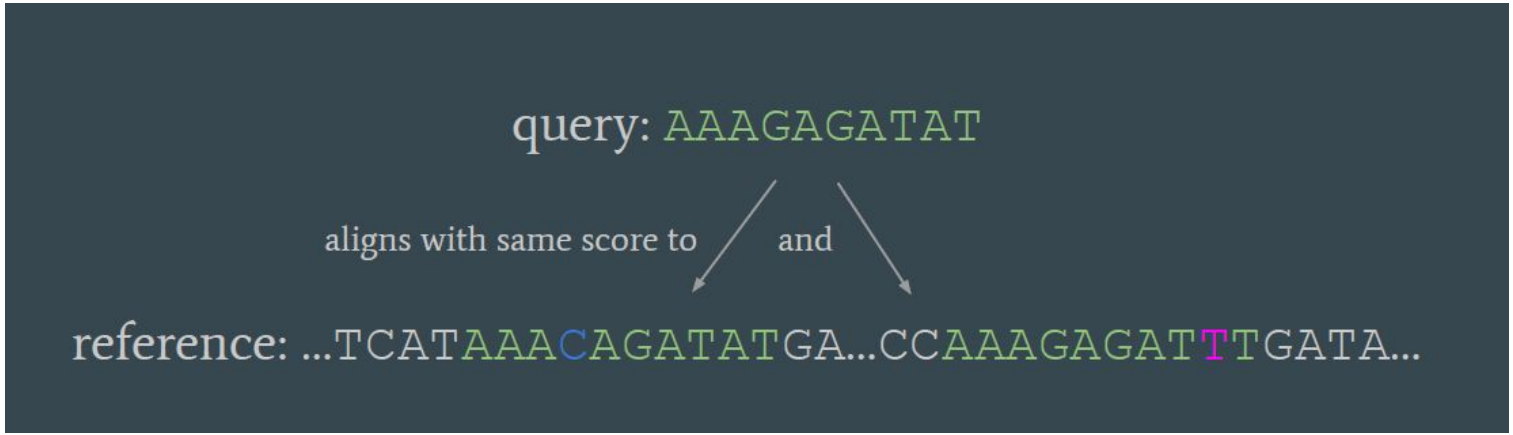
Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$
$W_1 = 2$

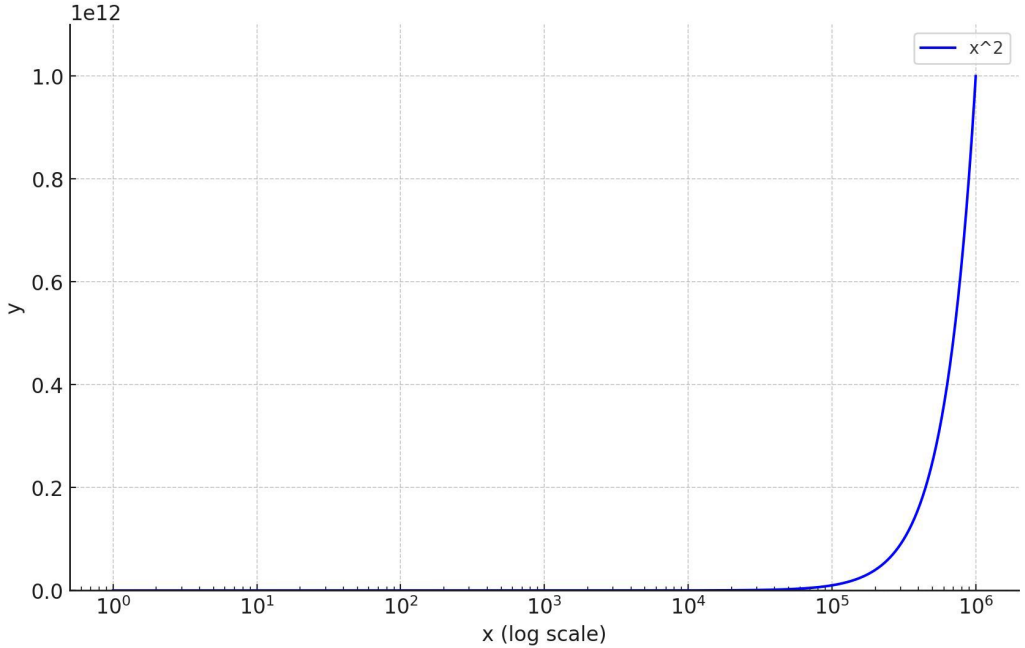
Limits of Smith-Waterman: Equally good alignments



- Most tools will either report a fixed number of equally good alignments, or just one arbitrarily with a warning (‘low mapping quality’). Either way, beware.

Limits of Smith-Waterman: Computationally intensive

- It requires (n*m) operations, where n and m are the sequence lengths.

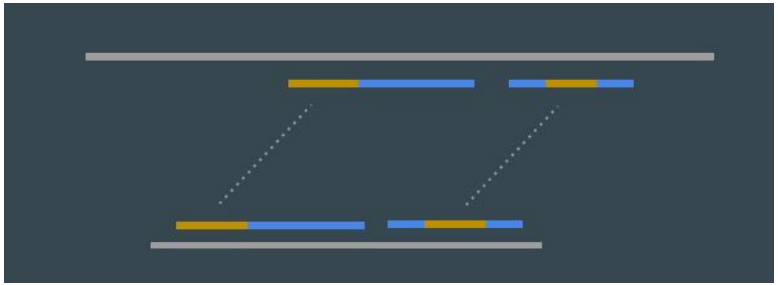


How BLAST works

Seeds: short sequences found in both the query and the reference.

1) Finds seeds using a table

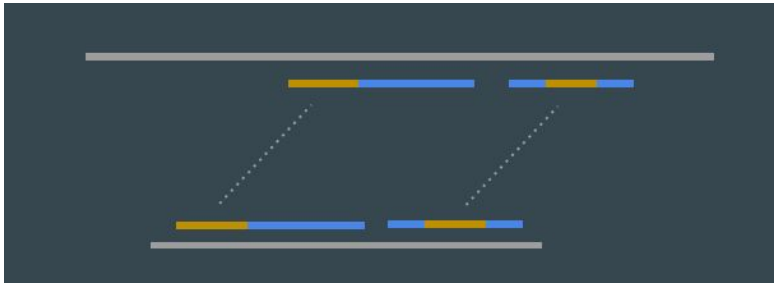
2) Aligns with SW-like method around seeds



Sequence	Found in ref at position(s)
AAAAA	10, 65, 147, …
AAAAC	80
….	
CTTAA	none
….	
CCCCC	49, 101

How BLAST works

BLAST (megablast) scoring
- Match = +1
- Mismatch = -2
- Indel = -2.5



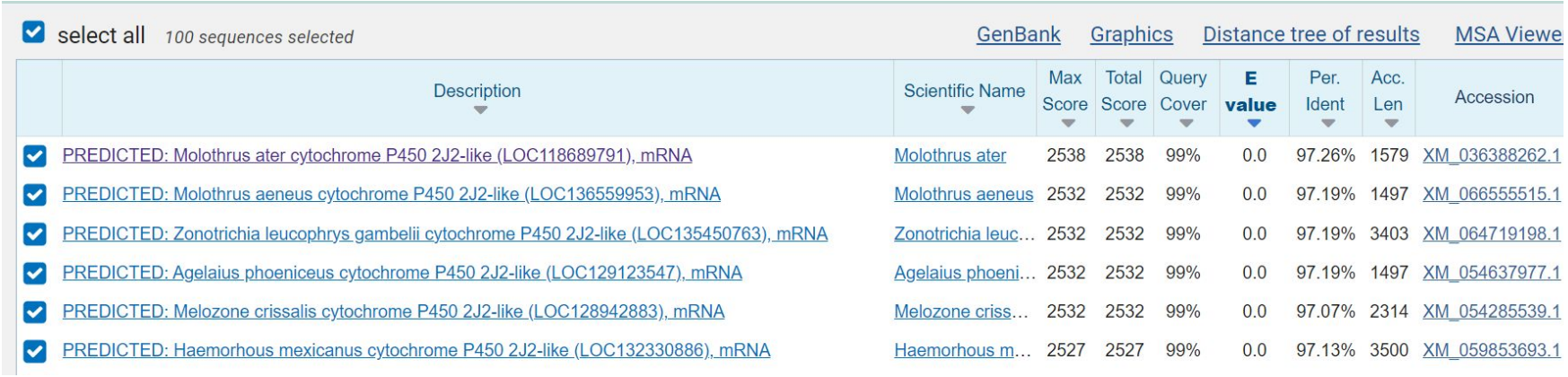
Sequence	Found in ref at position(s)
AAAAA	10, 65, 147, …
AAAAC	80
….	
CTTAA	none
….	
CCCCC	49, 101

How BLAST works

E-value = number of hits one can “expect” to see by chance on a database this size.

Common thresholds: < 0.01, or < 1e-5.

If your E-value is >= 0.01 you should question this match.



select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewe

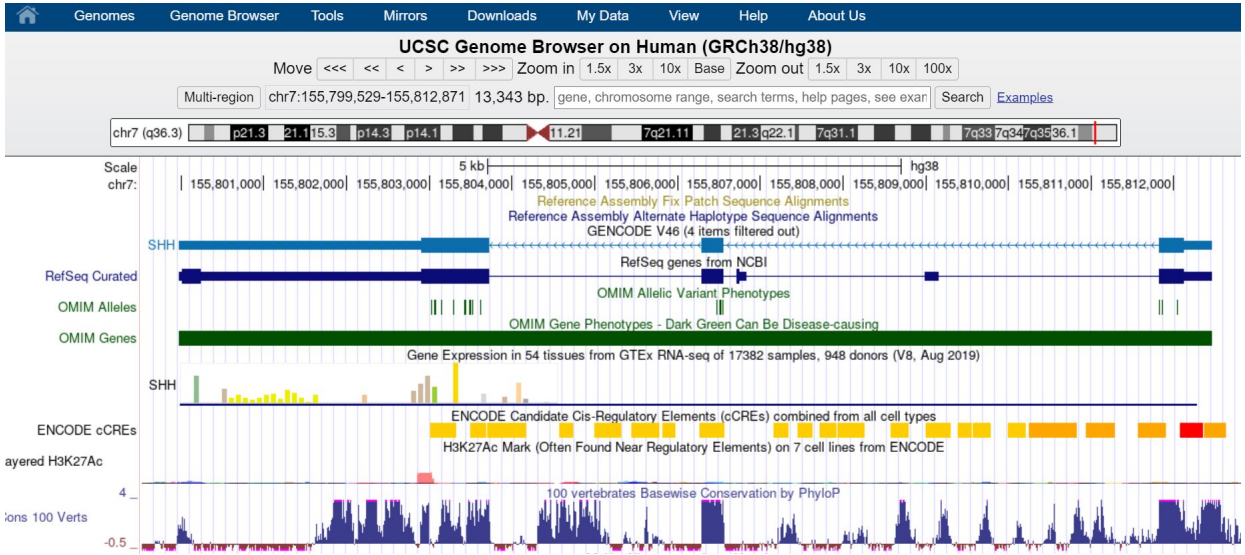
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Molothrus ater cytochrome P450 2J2-like (LOC118689791), mRNA	Molothrus ater	2538	2538	99%	0.0	97.26%	1579	XM_036388262.1
PREDICTED: Molothrus aeneus cytochrome P450 2J2-like (LOC136559953), mRNA	Molothrus aeneus	2532	2532	99%	0.0	97.19%	1497	XM_066555515.1
PREDICTED: Zonotrichia leucophrys gambelii cytochrome P450 2J2-like (LOC135450763), mRNA	Zonotrichia leuc…	2532	2532	99%	0.0	97.19%	3403	XM_064719198.1
PREDICTED: Agelaius phoeniceus cytochrome P450 2J2-like (LOC129123547), mRNA	Agelaius phoeni…	2532	2532	99%	0.0	97.19%	1497	XM_054637977.1
PREDICTED: Melozone crissalis cytochrome P450 2J2-like (LOC128942883), mRNA	Melozone criss…	2532	2532	99%	0.0	97.07%	2314	XM_054285539.1
PREDICTED: Haemorhous mexicanus cytochrome P450 2J2-like (LOC132330886), mRNA	Haemorhous m…	2527	2527	99%	0.0	97.13%	3500	XM_059853693.1

BLAT is not BLAST

1) Sequence-vs-genome (BLAT), instead of sequence-vs-database (BLAST)

2) Only find hits with >= 95% identity, over >= 40 bases

3) Faster than BLAST, integrated into UCSC Genome Browser



Efficient alignment of millions of reads requires massive amounts of computation but the Burrows-Wheeler Transform (BWT) can help

- Algorithm invented in 1994
- Rearranges and sorts sequence elements
- Reduces computational cost of search
- Core component of most commonly used aligners
  - BWT → FM-Index → SW alignment

Transformation				
1. Input	2. All rotations	3. Sort into lexical order	4. Take the last column	5. Output
^BANANA$	^BANANA$ $^BANANA A$^BANAN NA$^BANA ANA$^BAN NANA$^BA ANANA$^B BANANA$^	ANANA$^B ANA$^BAN A$^BANAN BANANA$^ NANA$^BA NA$^BANA ^BANANA$ $^BANANA	ANANA$^B ANA$^BAN A$^BANAN BANANA$^ NANA$^BA NA$^BANA ^BANANA$ $^BANANA	BNN^AA$A

Tools for short-read alignment

Bowtie2

BWA-MEM (BWA-MEM2)

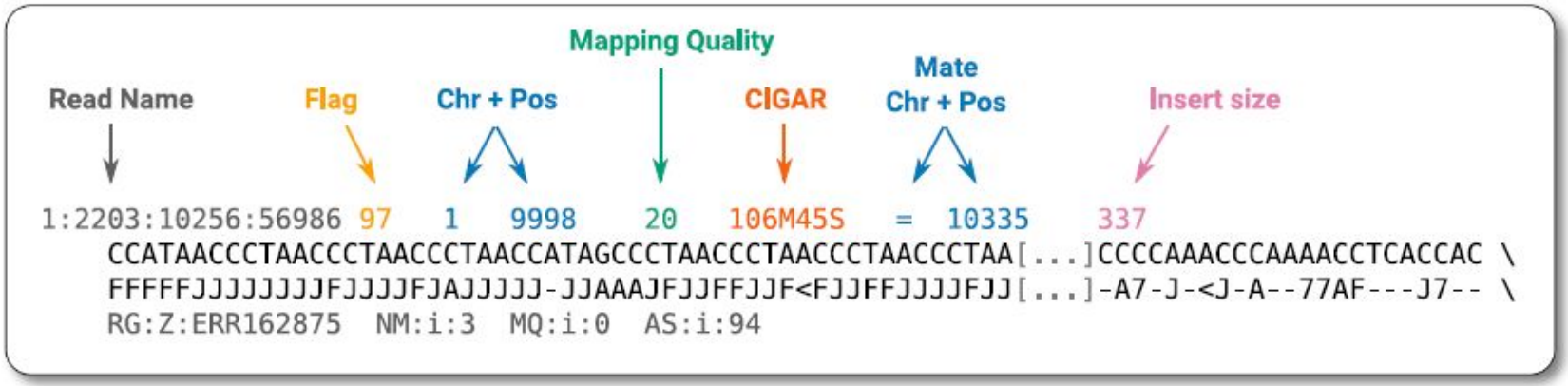
Minimap2 - faster, but cannot map <= 100 bp reads

How do I choose?

- Actively maintained
- Well documented
- Commonly used in your field
- Easy to install

Alignment output

- SAM - Sequence Alignment Map / BAM - Binary Alignment MAP
  - For each read
    - Coordinates in the reference
    - Sequences - carried over from FastQ format
    - Alignment details - CIGAR string
    - Quality information - carried over from FastQ format



BAM/SAM flags

Flag

Hex	Dec	Flag	Description
0x1	1	PAIRED	paired-end (or multiple-segment) sequencing technology
0x2	2	PROPER_PAIR	each segment properly aligned according to the aligner
0x4	4	UNMAP	segment unmapped
0x8	8	MUNMAP	next segment in the template unmapped
0x10	16	REVERSE	SEQ is reverse complemented
0x20	32	MREVERSE	SEQ of the next segment in the template is reversed
0x40	64	READ1	the first segment in the template
0x80	128	READ2	the last segment in the template
0x100	256	SECONDARY	secondary alignment
0x200	512	QCFAIL	not passing quality controls
0x400	1024	DUP	PCR or optical duplicate
0x800	2048	SUPPLEMENTARY	supplementary alignment

Flag lookup tool: https://broadinstitute.github.io/picard/explain-flags.html

BAM/SAM flags

- Flags are especially useful in selecting reads with samtools

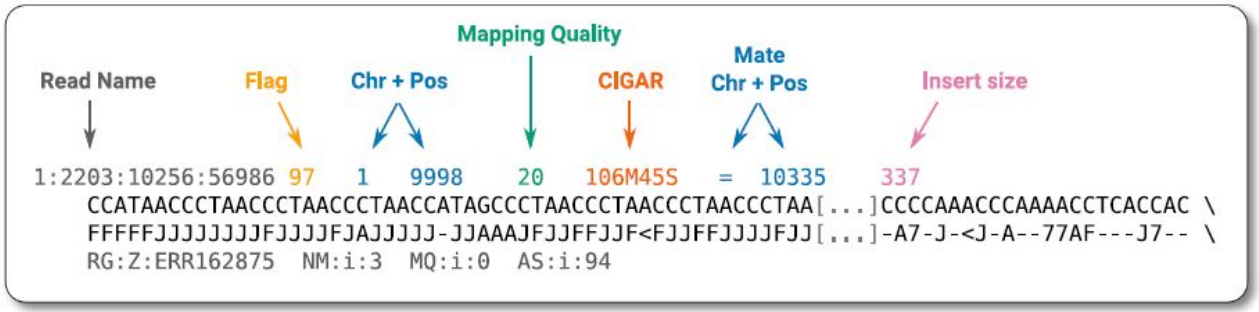
```
samtools view -f 4 file.sam > unmapped.sam

samtools view -F 4 file.sam > mapped.sam
```

- -f 4 - option selects all unmapped reads
- -F 4 - option excludes all unmapped reads

https://broadinstitute.github.io/picard/explain-flags.html

BAM/SAM CIGAR strings



CIGAR string

compact representation of sequence alignment:

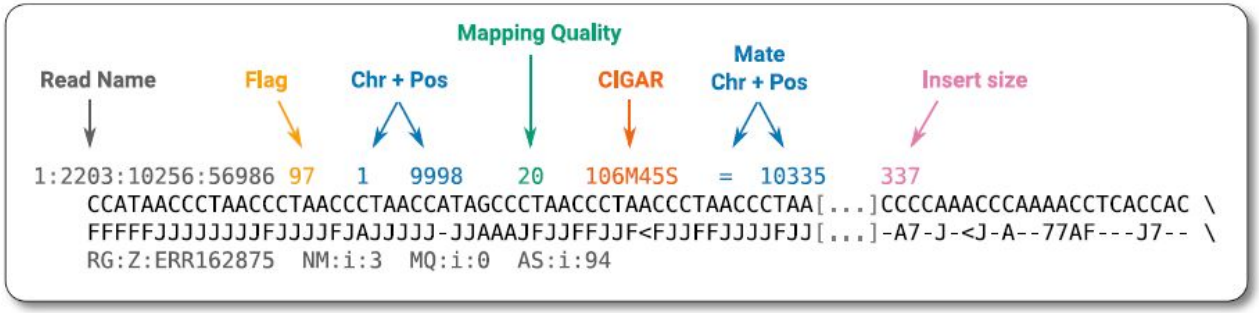
M	alignment match or mismatch
=	sequence match
X	sequence mismatch
I	insertion to the reference
D	deletion from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
N	skipped region from the reference
P	padding (silent deletion from padded reference)

```
Ref:   ACGTACGTACTGT
Read:  ACGT----ACTGA
Cigar: 4M 4D 5M
```

```
Ref:   ACGT----ACGTA
Read:  ACGTACGTACGTA
Cigar: 4M 4I 5M
```

```
Ref:   CTCAGTG-GTCATCGTT
Read:  CGCA-TGAGTCTAGACG
Cigar: 4M 1D 2M 1I 3M 6S
```

BAM/SAM insert size



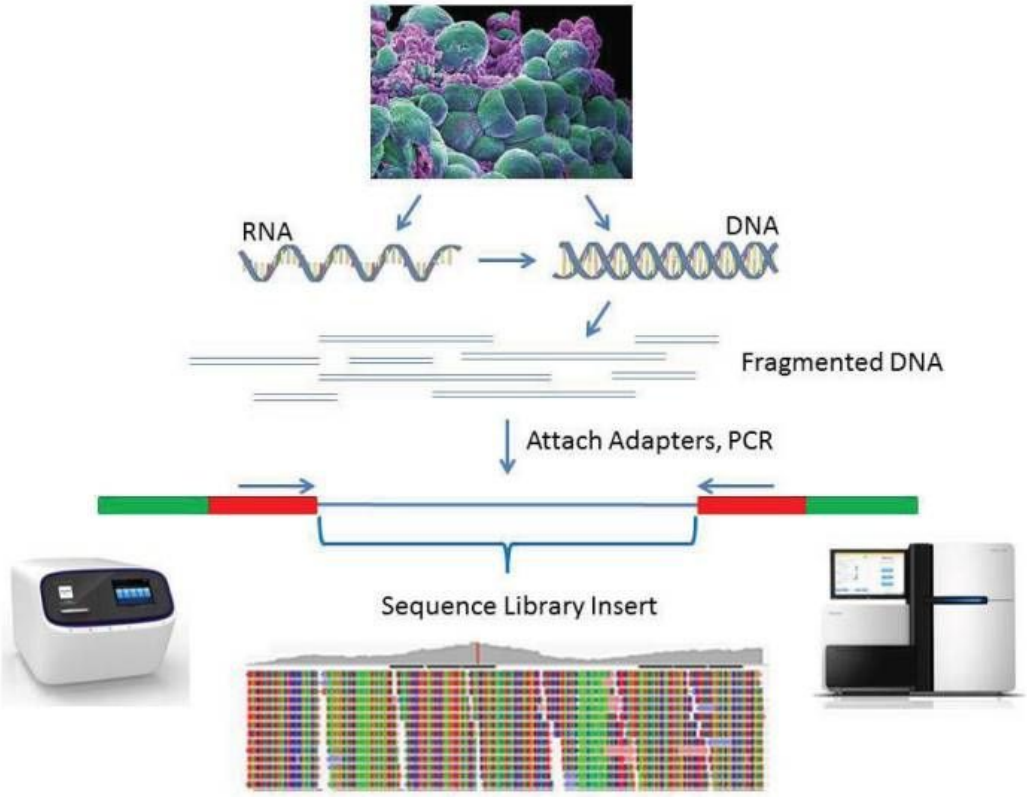
Insert size

length of the DNA fragment sequenced from both ends by paired-end sequencing:

BAM/SAM Visualization - https://igv.org

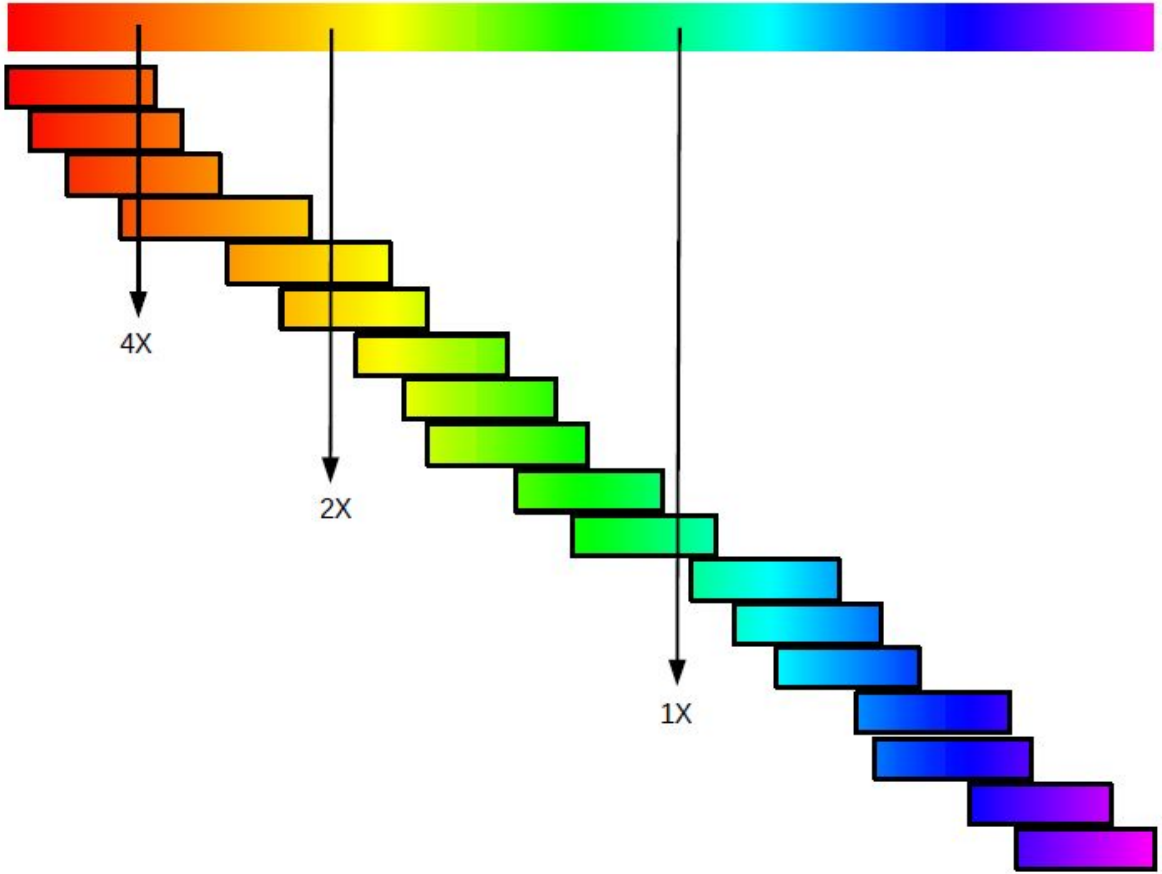
http://tinyurl.com/yuq9ab9z

Part 2 - Genome Assembly



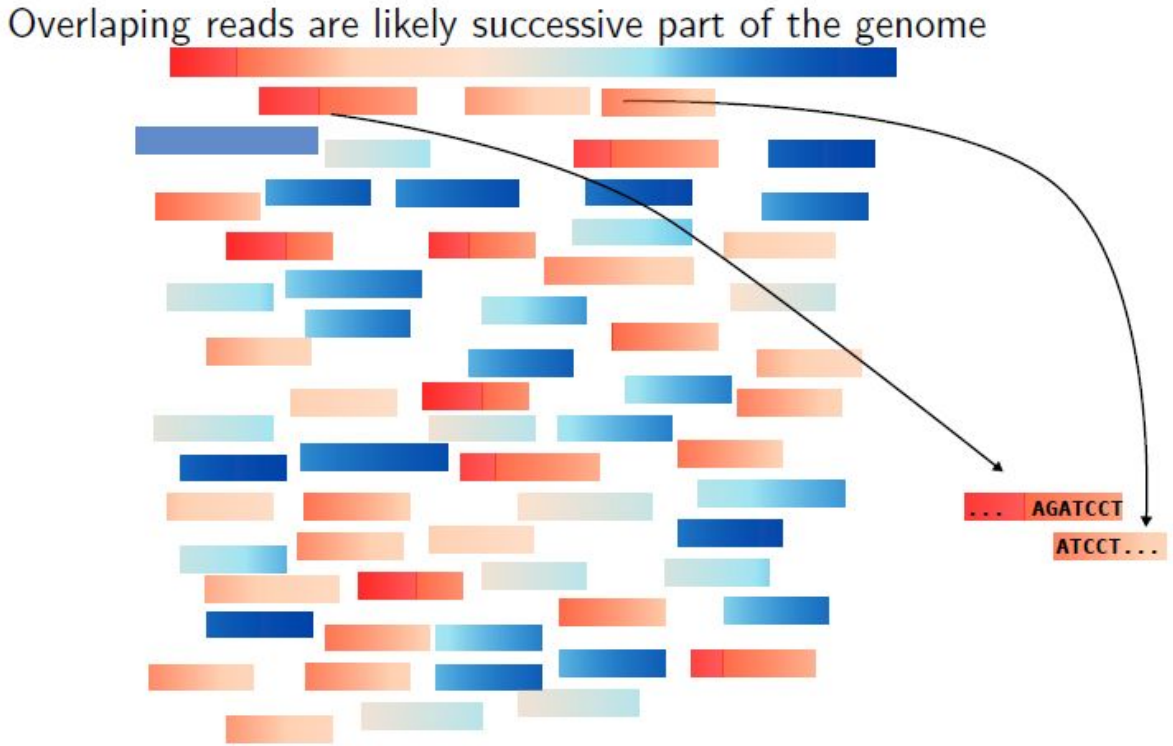
- Sequencing libraries consist of a millions of small fragments of the genome
- How do we put these pieces back together?

Coverage

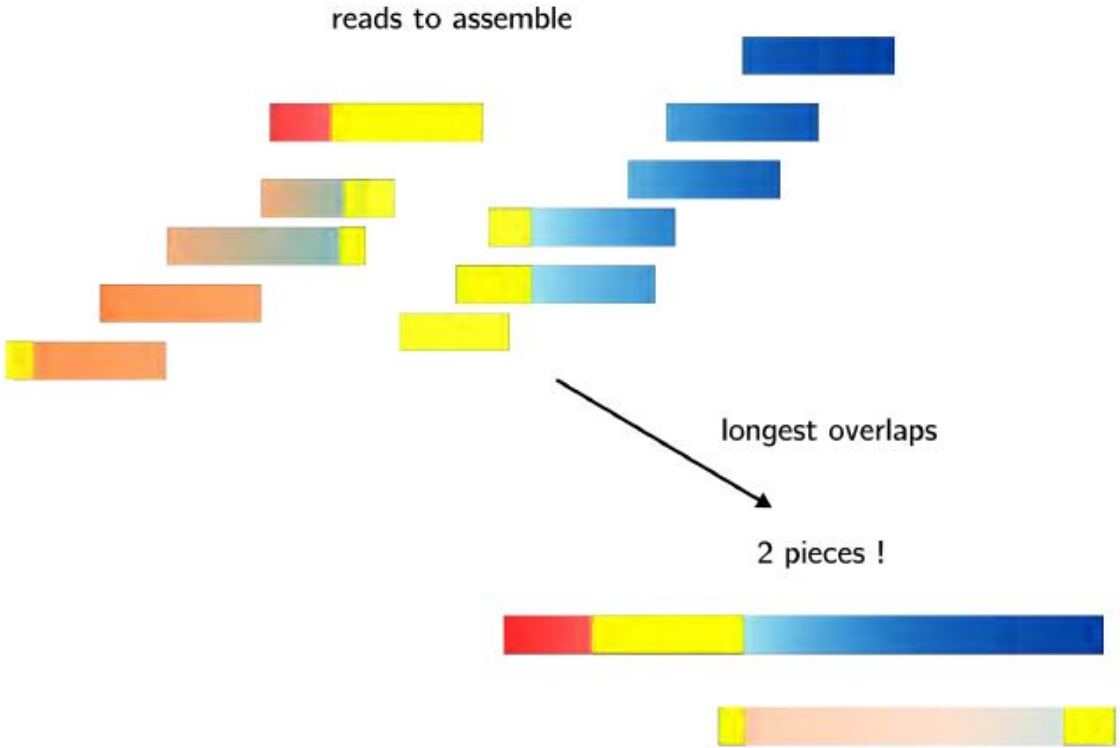


- Typically 30-60x coverage for short-read assemblies

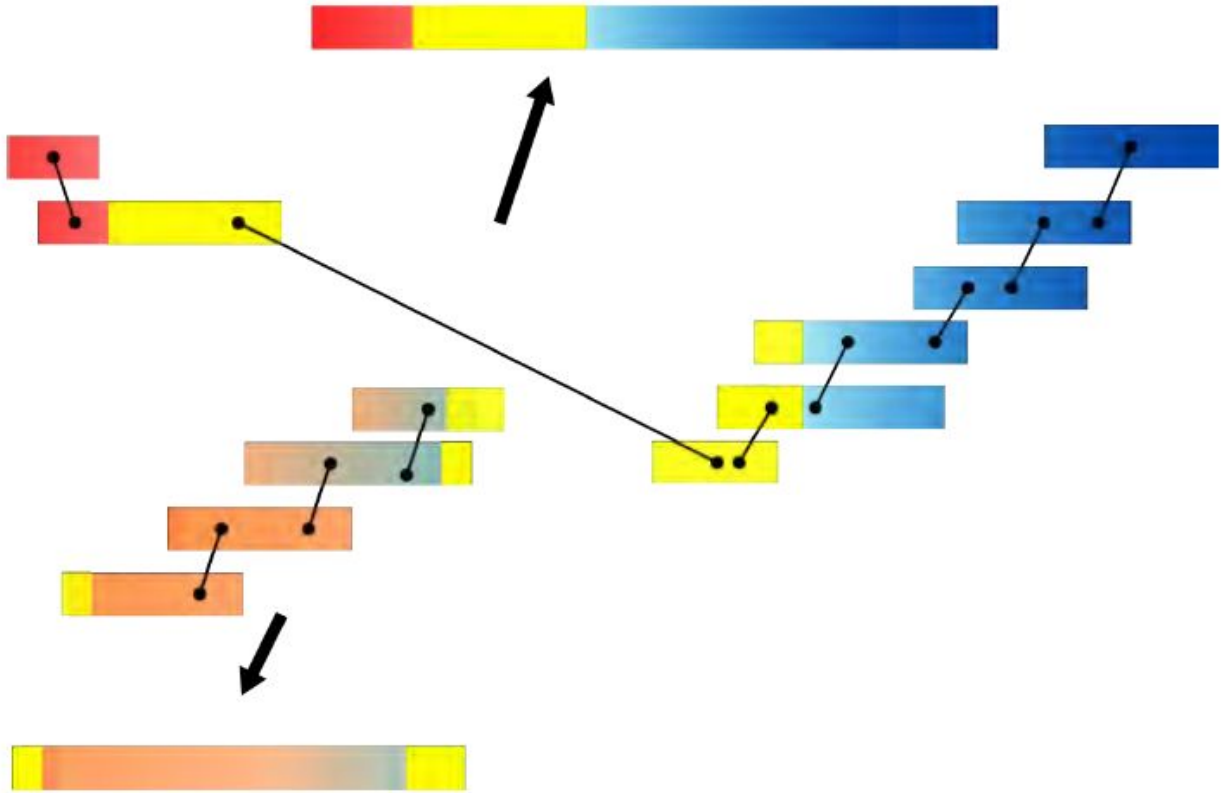
Approach 1 - Order reads according to overlaps



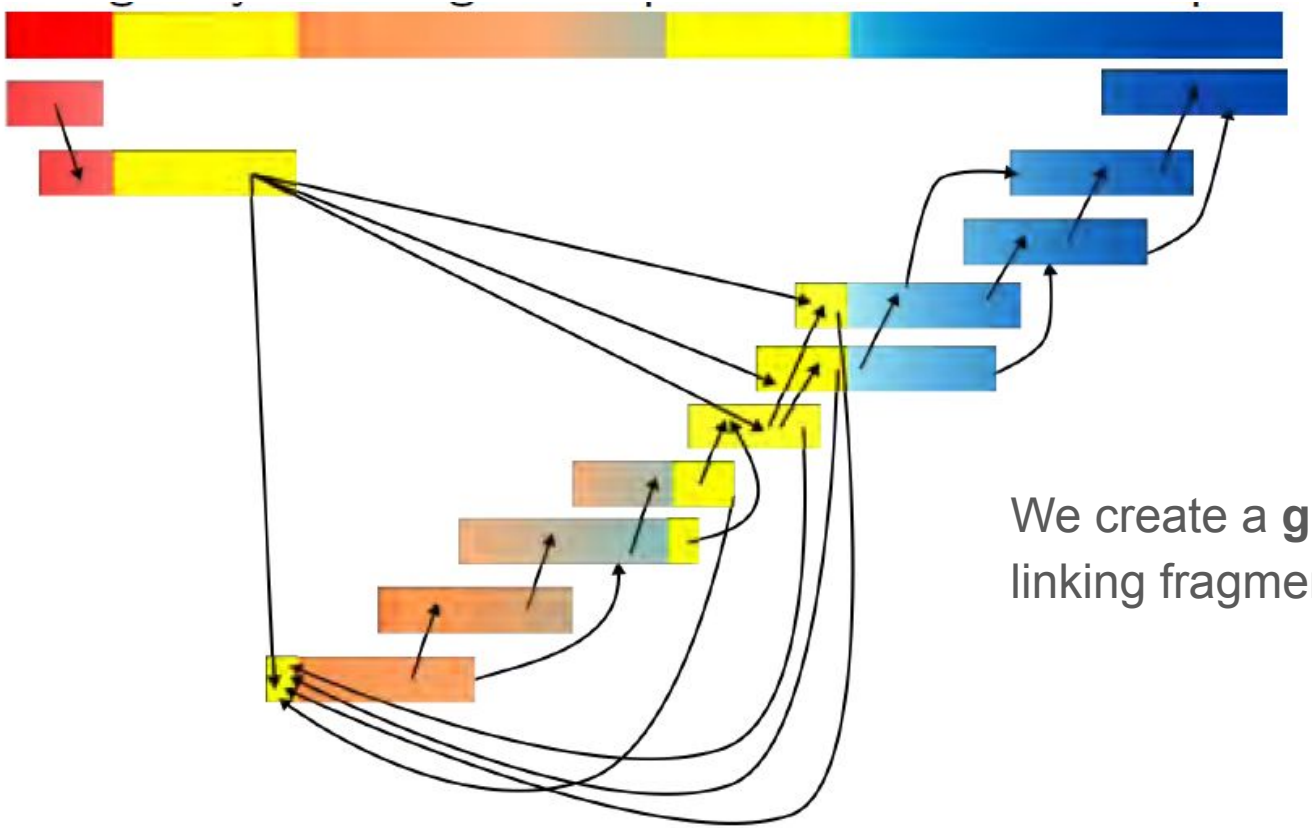
Assemble by longest overlaps



Assemble by longest overlaps



Accounting for other overlaps



We create a **graph** with paths linking fragments.

Genome Assembly - Overlap Graph

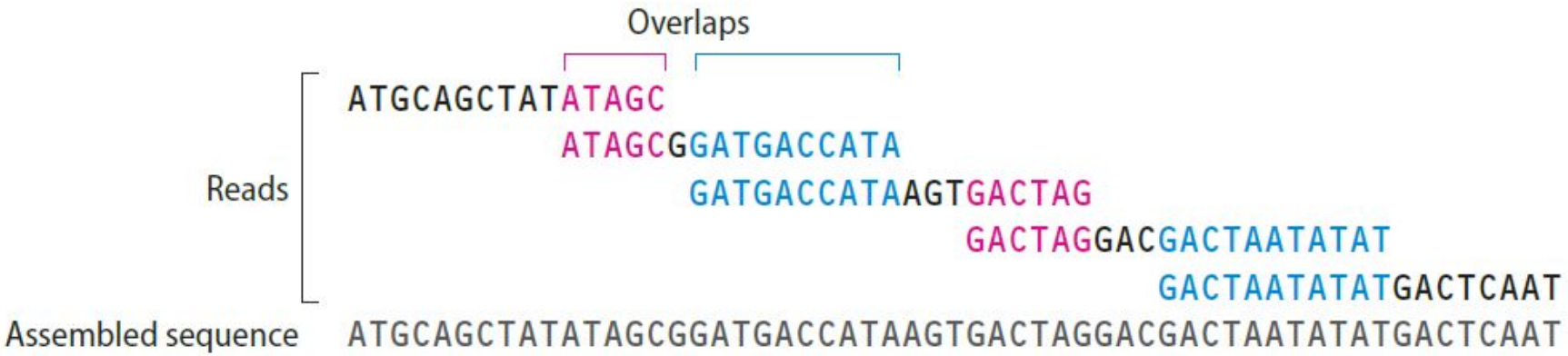
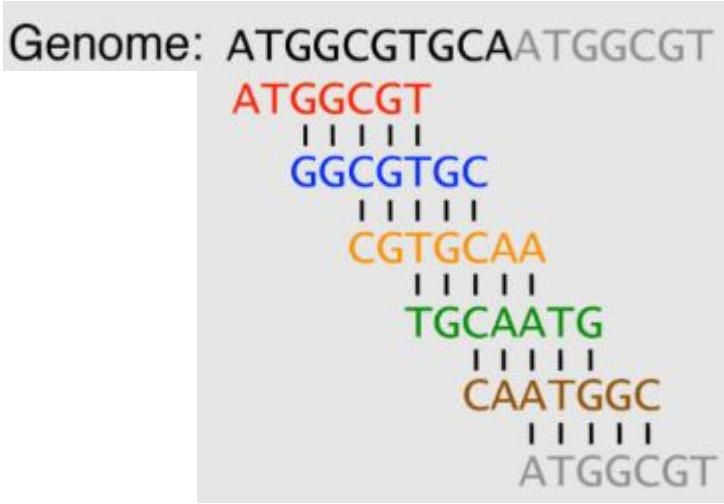


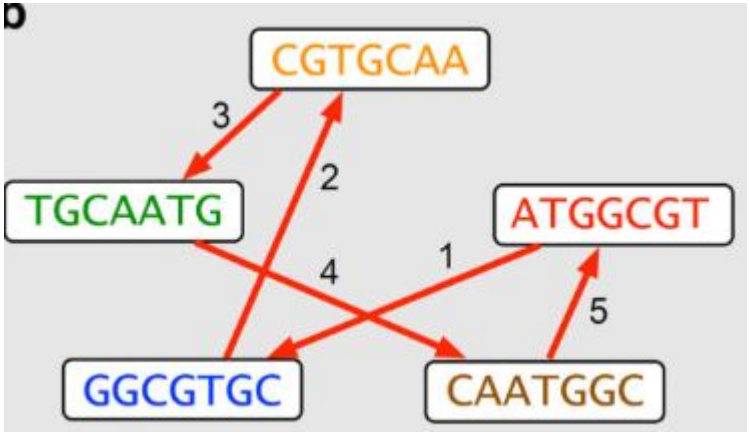
Figure 4.20 Sequence assembly using an overlap graph. Overlaps between pairs of reads are identified in order to build up the master sequence.

- Original approach to assembly w/ sanger sequencing - Example - Celera Assembler (CA)
- Computationally expensive - requires pairwise alignment of all reads to find overlaps. Not suitable for short-read sequencing
- Today it is useful for long-read sequencing

Kmers

- Sub-strings of fixed length
- Smaller than read size (k = 21-55)
- If small enough, can find all unique kmers in a sequencing dataset

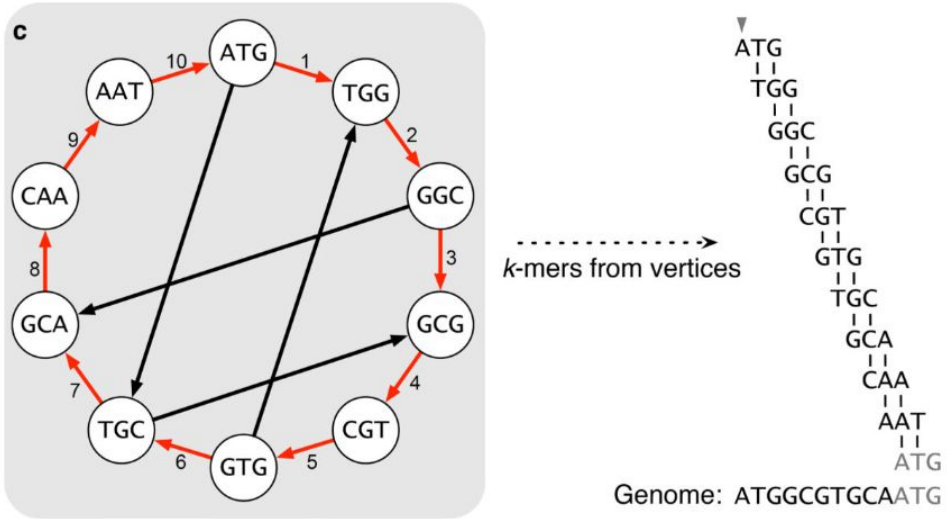




How do we find the best graph?

Hamiltonian path

- Overlap graph - Kmers are nodes and edges are the overlaps
- Visit every **node** once
- NP-complete problem
  - No known deterministic algorithm

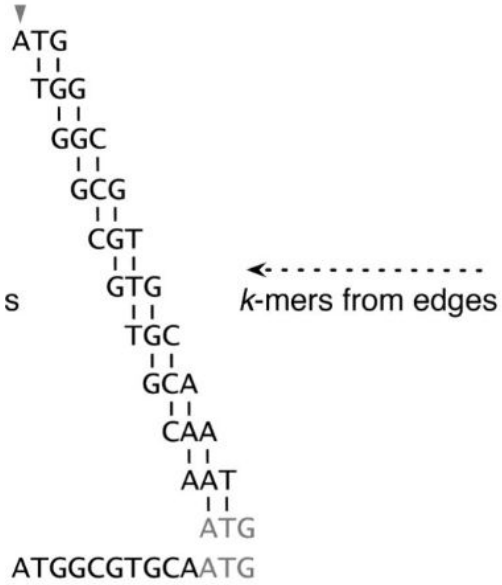


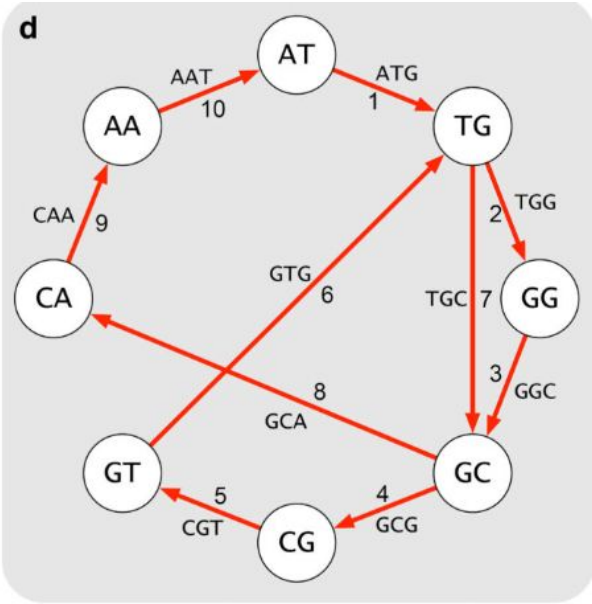
k = 3

How do we find the best graph?

Eulerian path

- de Bruijn graph - Kmers are edges and overlaps are the nodes
- Visit every **edge** once
- Computationally tractable





k = 3

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5531759/

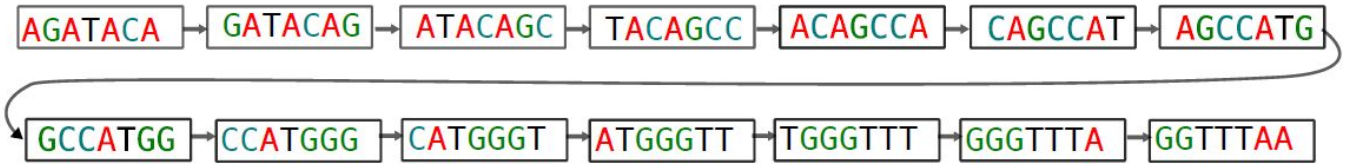
de Bruijn graph abstract redundancy

read overlaps

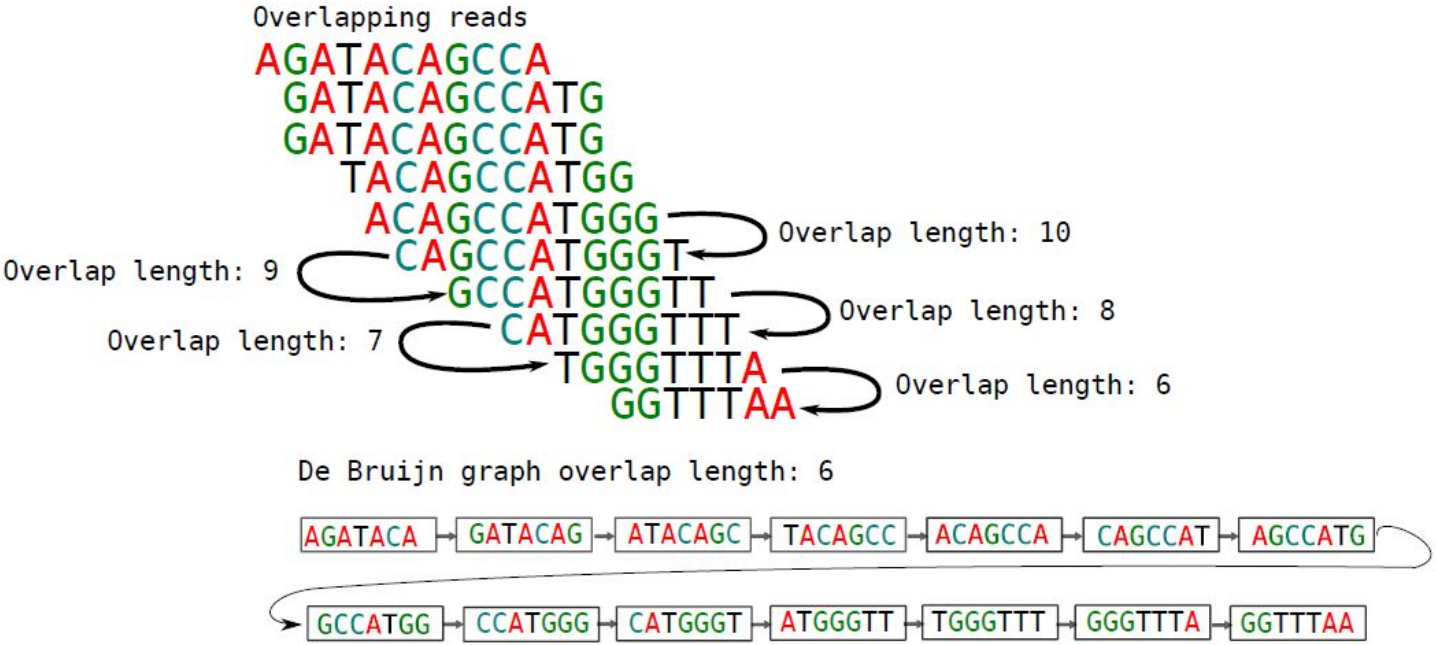
```
AGATACAGCCA
 GATACAGCCAT
 GATACAGCCAT
  ATACAGCCATG
   TACAGCCATGG
    ACAGCCATGGG
    ACAGCCATGGG
     CAGCCATGGGT
      AGCCATGGGTT
       GCCATGGGTTT
       GCCATGGGTTT
        CCATGGGTTTA
         CATGGGTTTAA
```

65 non distinct 7-mers in reads

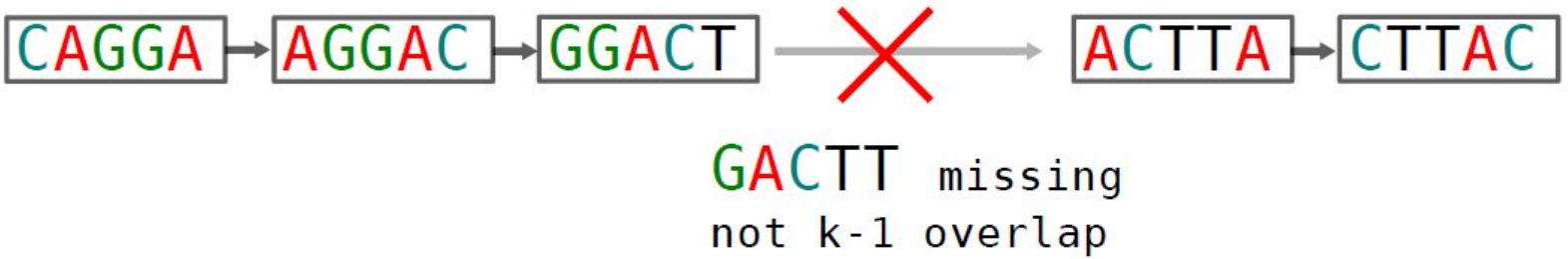
14 **distinct** 7-mers in the de Bruijn graph



de Bruijn graph only rely on k-1 overlap



de Bruijn limitations



GGACT and ACTTA overlap is only of size 3 !

- Low sequencing depth
- Sequencing errors

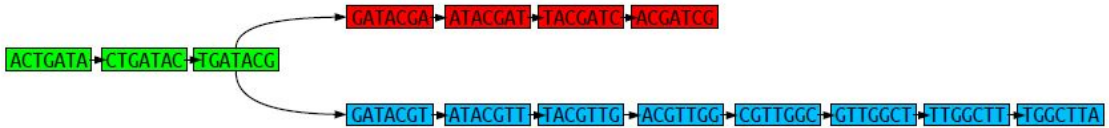
de Bruijn limitations

...TACAGGACTTA... ...TATAGGACTGA...

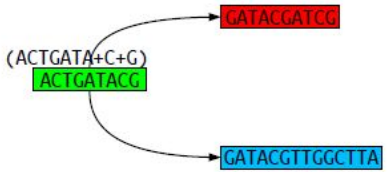
each k-mer appears only once in a de Bruijn graph

de Bruijn limitations - repeats lead to forking paths

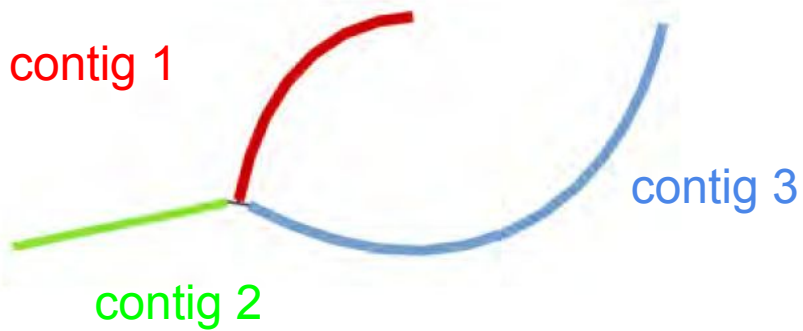
De Bruijn graph:



Compacted De Bruijn graph:

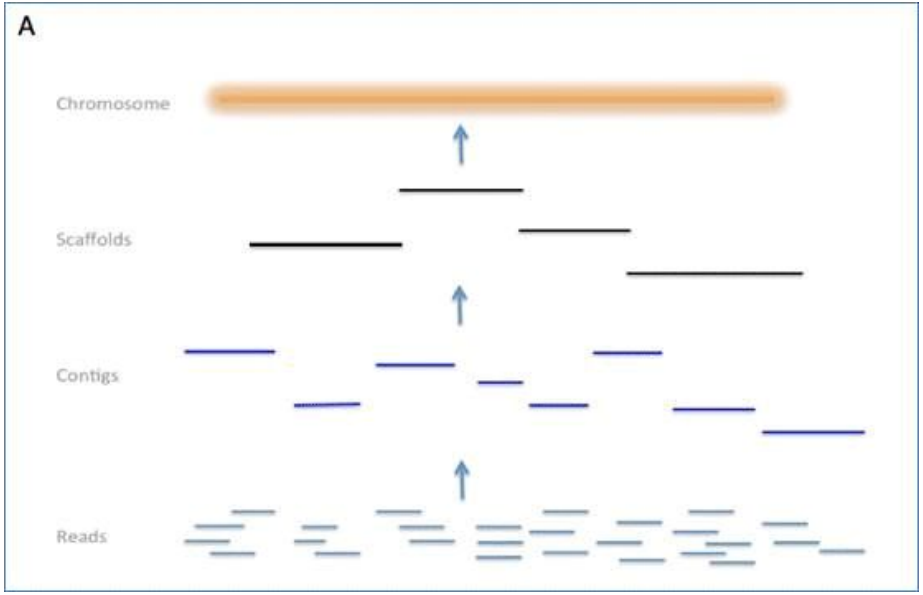


Graphical representation
(.gfa plot using Bandage):



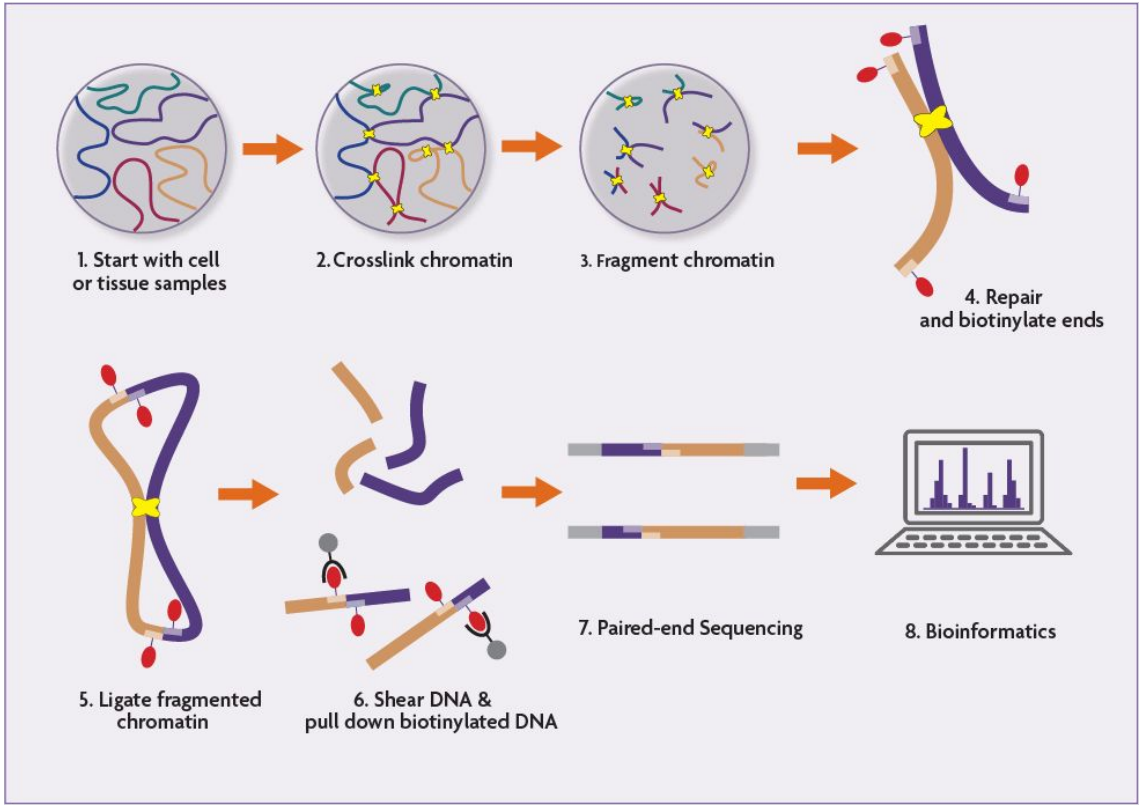
Chromosome level assembly requires scaffolding

- Link contigs with information from
  - External reference
  - Long reads
    - HiFi
    - Nanopore
  - Chromatin conformation information
    - HiC



Chromosome conformation - HiC

- **Intra**chromosomal contact probability is on average much higher than **inter**chromosomal.
- Interaction probability rapidly **decays** with increasing genomic distance.



Chromosome conformation - HiC

- **Intra**chromosomal contact probability is on average much higher than **inter**chromosomal.
- Interaction probability rapidly **decays** with increasing genomic distance.

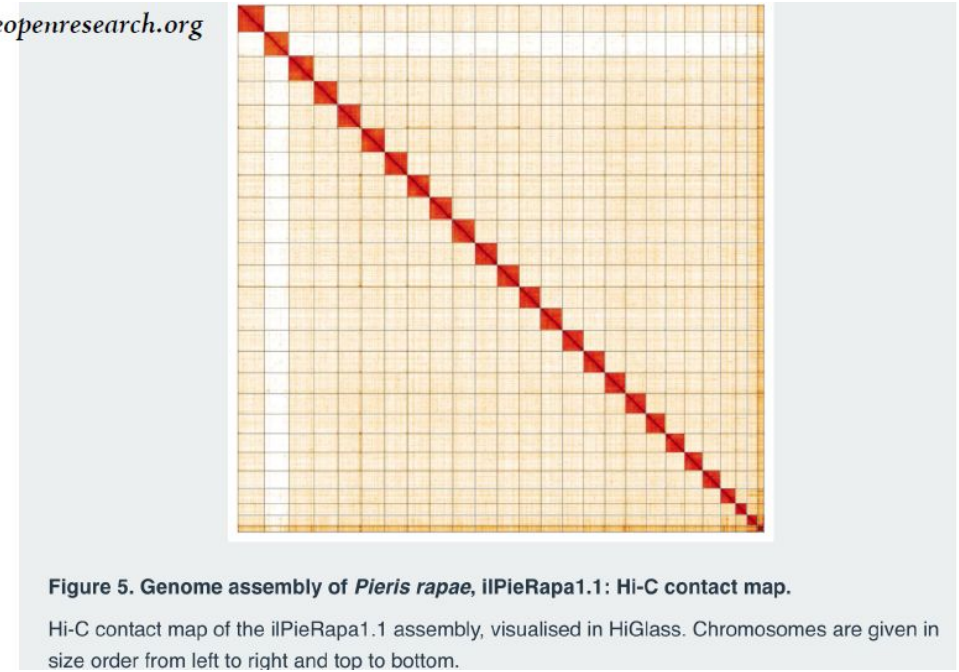
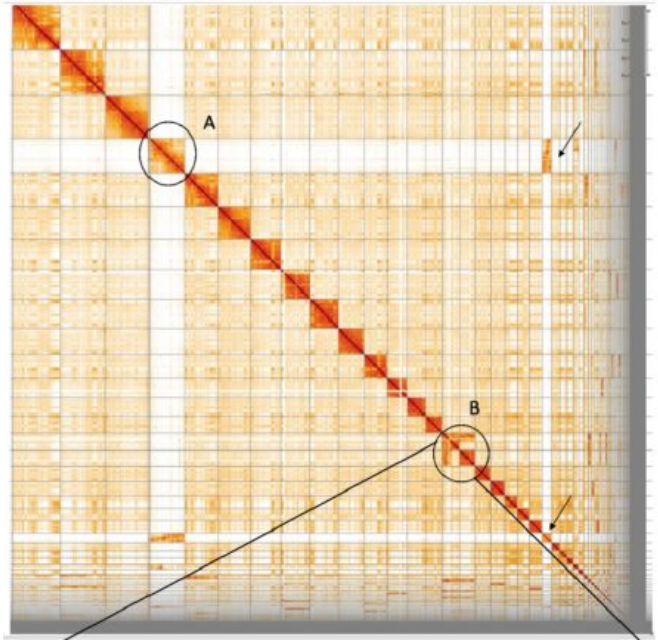


Figure 5. Genome assembly of *Pieris rapae*, ilPieRapa1.1: Hi-C contact map.

Hi-C contact map of the ilPieRapa1.1 assembly, visualised in HiGlass. Chromosomes are given in size order from left to right and top to bottom.

Chromosome conformation - HiC

- **Intra**chromosomal contact probability is on average much higher than **inter**chromosomal.
- Interaction probability rapidly **decays** with increasing genomic distance.

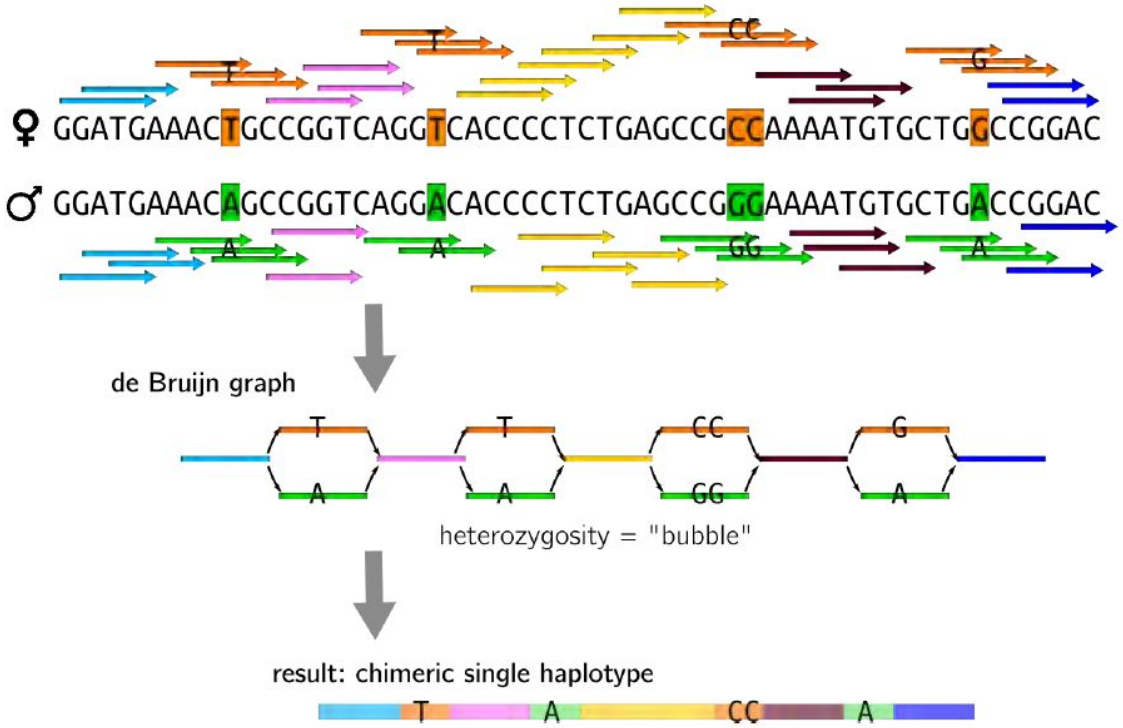


Choloepus didactylus VGP

Non-curated output

3.2 Gb, 281 scaffolds, N50 = 161 Mb

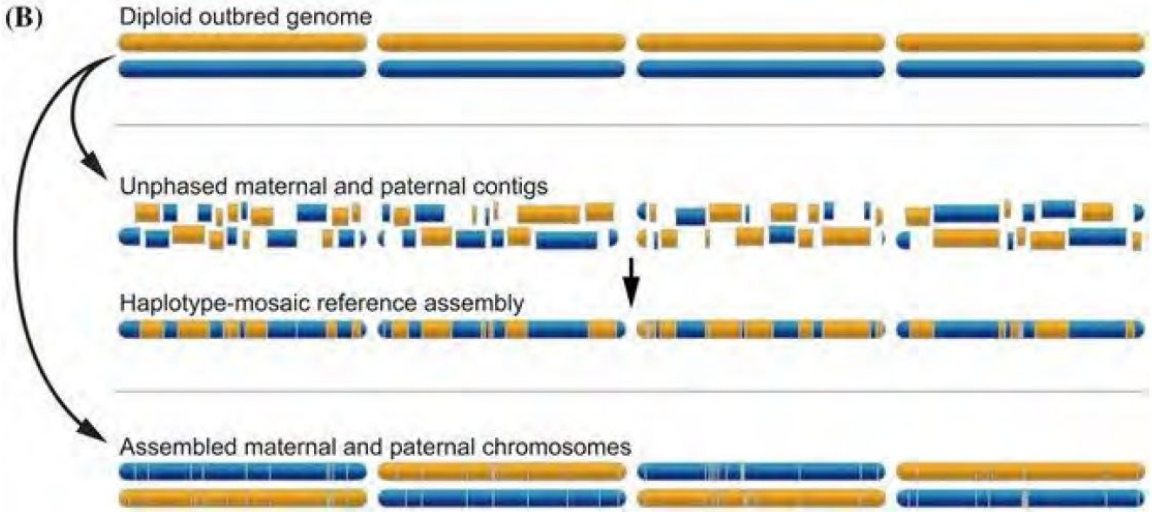
Haploid assembly



Assembly concession: haplotypes are collapsed when using short reads

Haplotype phasing

- Requires long-reads
- Built in function in some assemblers (HiFiasm)



Popular assemblers

SPAdes

- Designed to assemble megabase-sized genomes
- Multiple k de Bruijn graph assembly from short reads
- Can use long reads to solve repeats

Mandatory - Short reads

Optional - Long reads

Popular assemblers

SPAdes

- Designed to assemble megabase-sized genomes
- Multiple k de Bruijn graph assembly from short reads
- Can use long reads to solve repeats

Mandatory - Short reads

Optional - Long reads

Popular assemblers

Hifiasm assembler

- Build an overlap graph from HiFi reads
- Generate both haploid and diploid assemblies
- Can use (very) long reads to solve repeats

Mandatory - HiFi reads

Optional - Long reads

Popular assemblers

Flye assembler

- Build a repeat graph from long reads
- Can use any kind of long reads
- Can also assemble metagenomes

Mandatory - HiFi/Long reads

Optional - HiFi/Long reads

Popular assemblers

Unicycler (long read mode)

- Build a overlap graph from long reads
- Polish the assembly
- Also has a short-reads-first similar to SPAdes

Mandatory - Long reads

Optional - Short reads