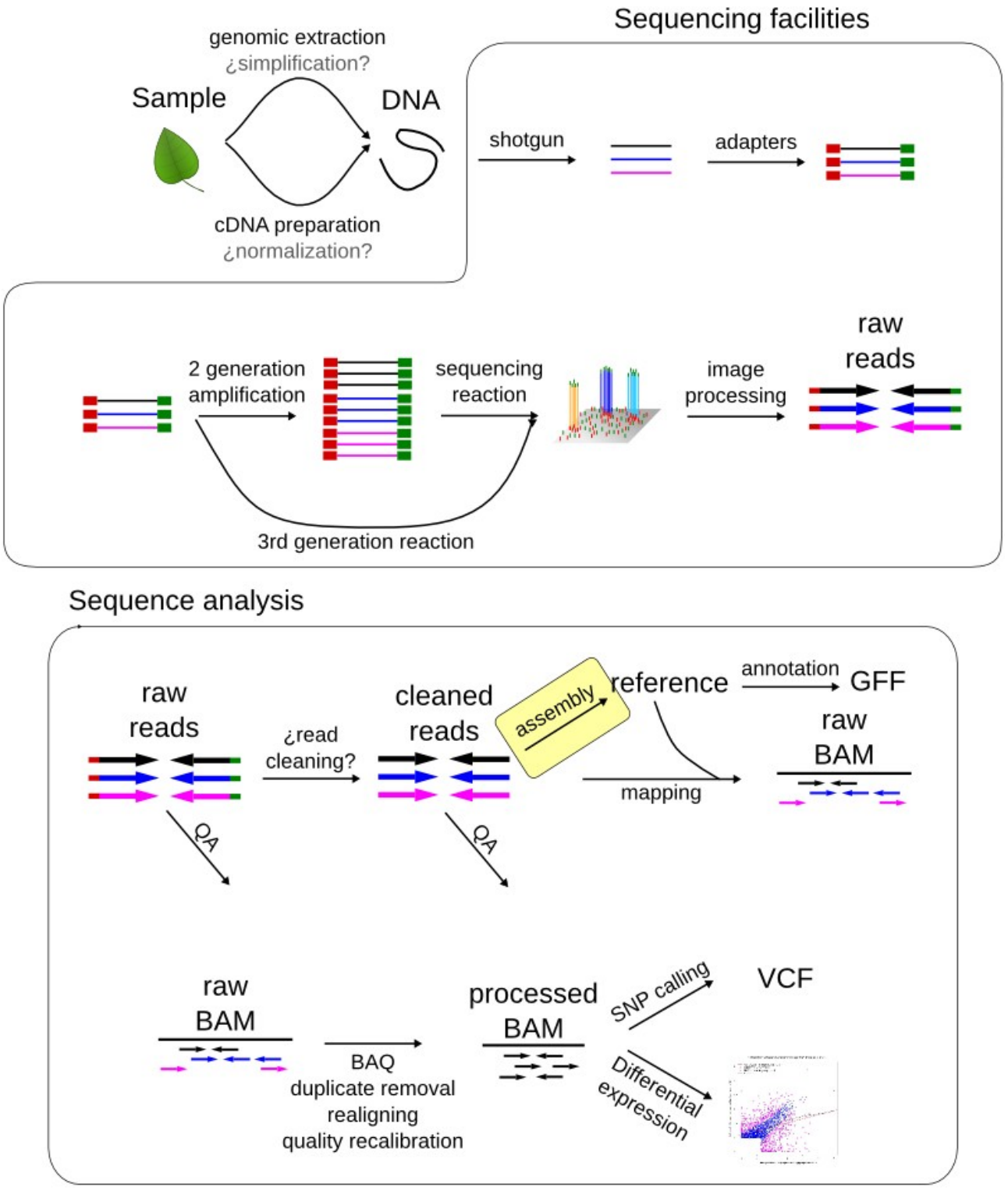
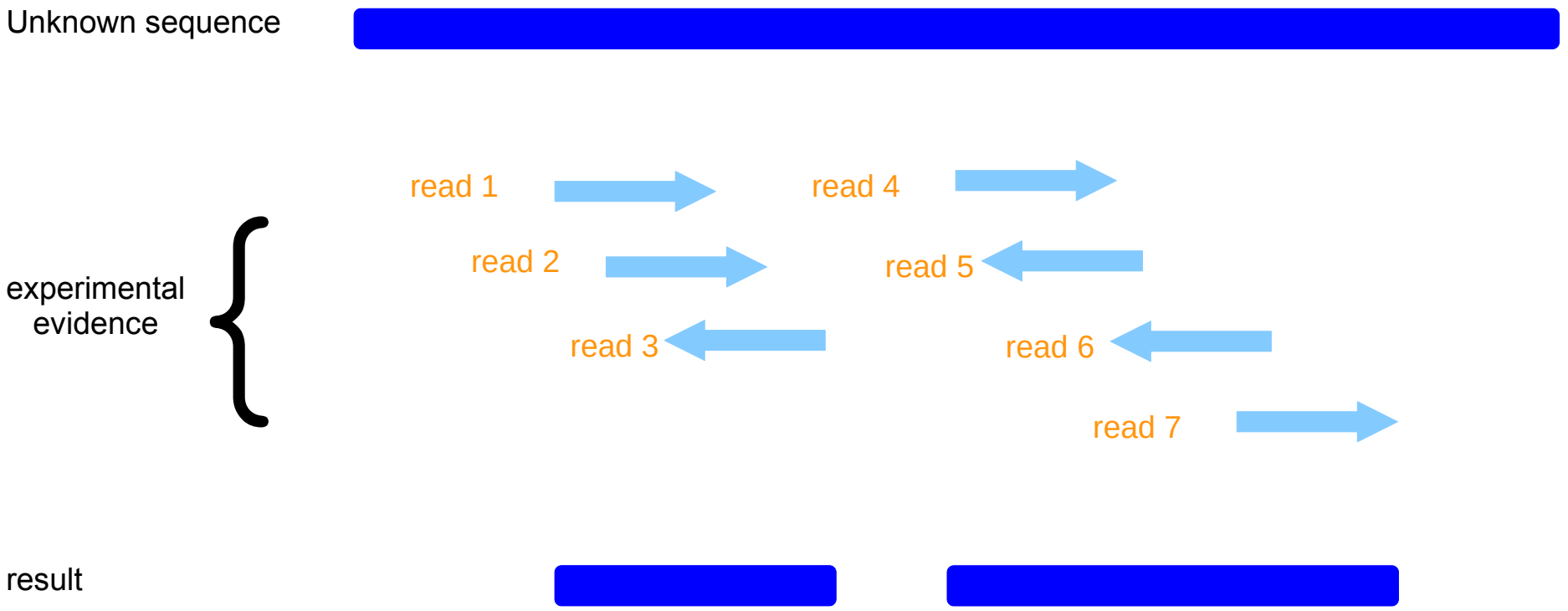
Sequence assembly

Jose Blanca
COMAV institute
bioinf.comav.upv.es

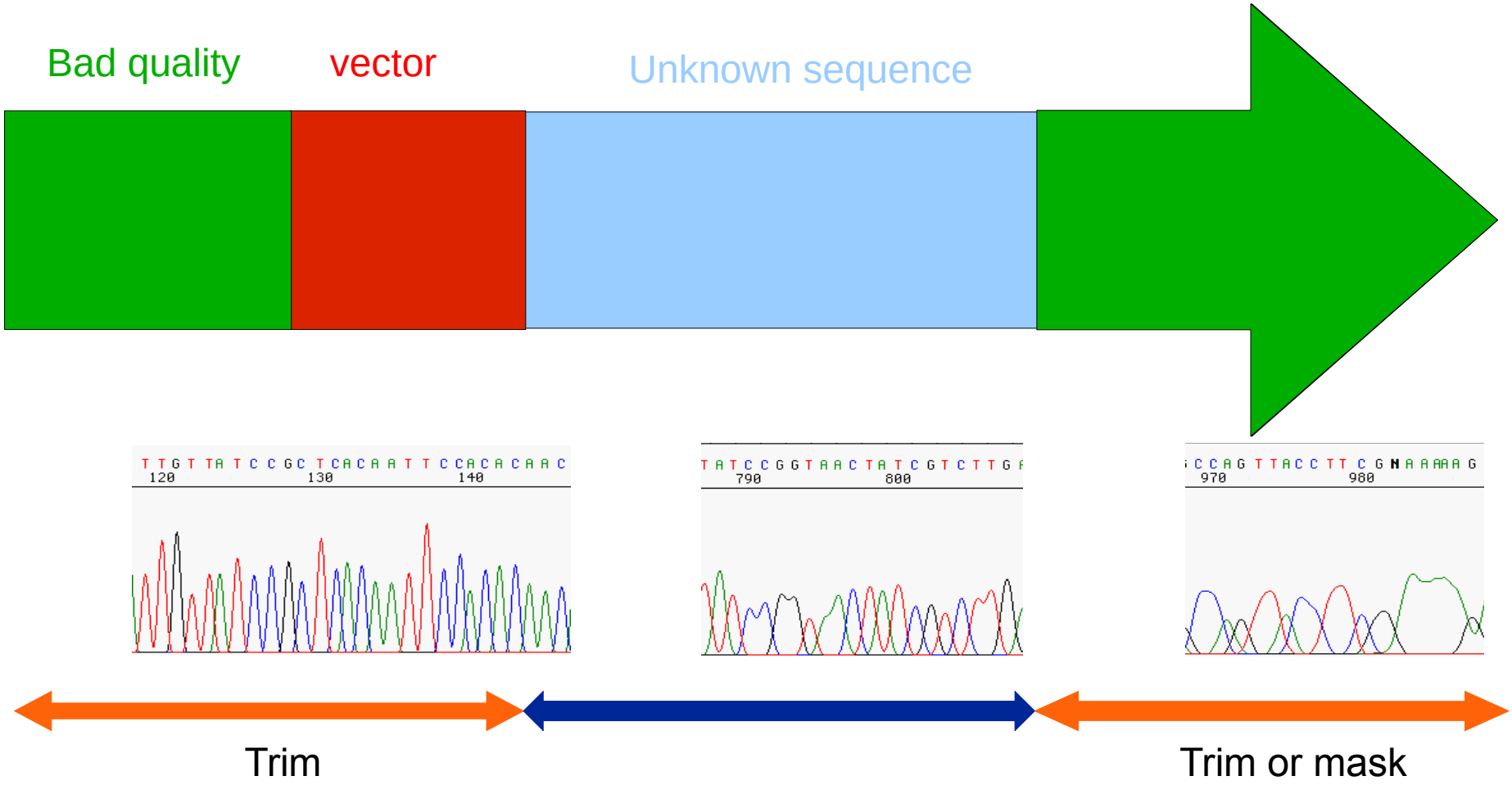

Sequencing facilities
genomic extraction
¿simplification?
Sample
DNA
cDNA preparation
¿normalization?
shotgun
adapters
2 generation
amplification
sequencing
reaction
image
processing
raw
reads
3rd generation reaction
Sequence analysis
raw
reads
¿read
cleaning?
cleaned
reads
assembly
reference
annotation
GFF
raw
BAM
mapping
QA
QA
raw
BAM
BAQ
duplicate removal
realigning
quality recalibration
processed
BAM
SNP calling
VCF
Differential
expression

Assembly project



Read cleaning

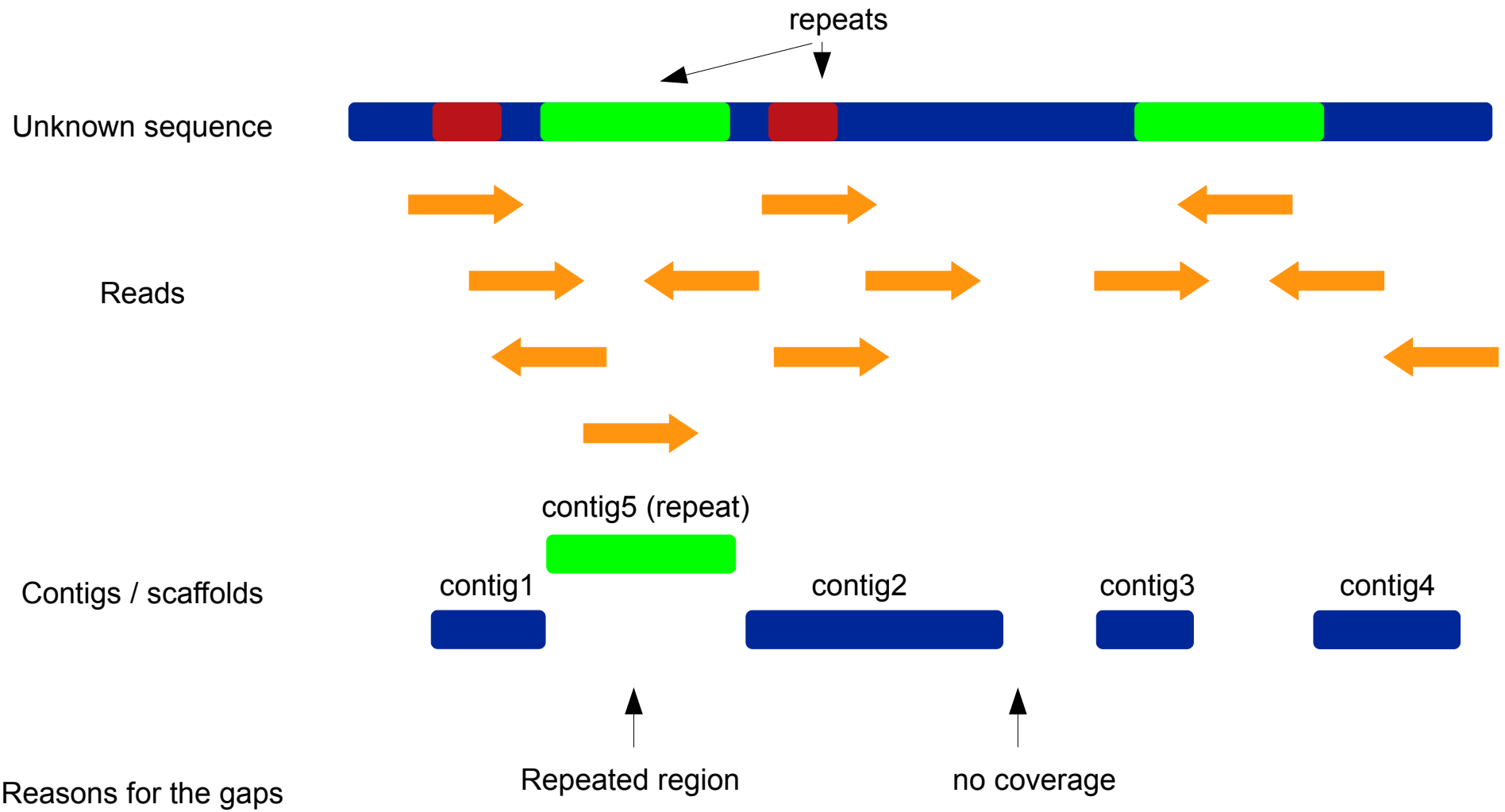
Some assemblers choke with: bad quality stretches, adapters, low complexity regions, contamination.



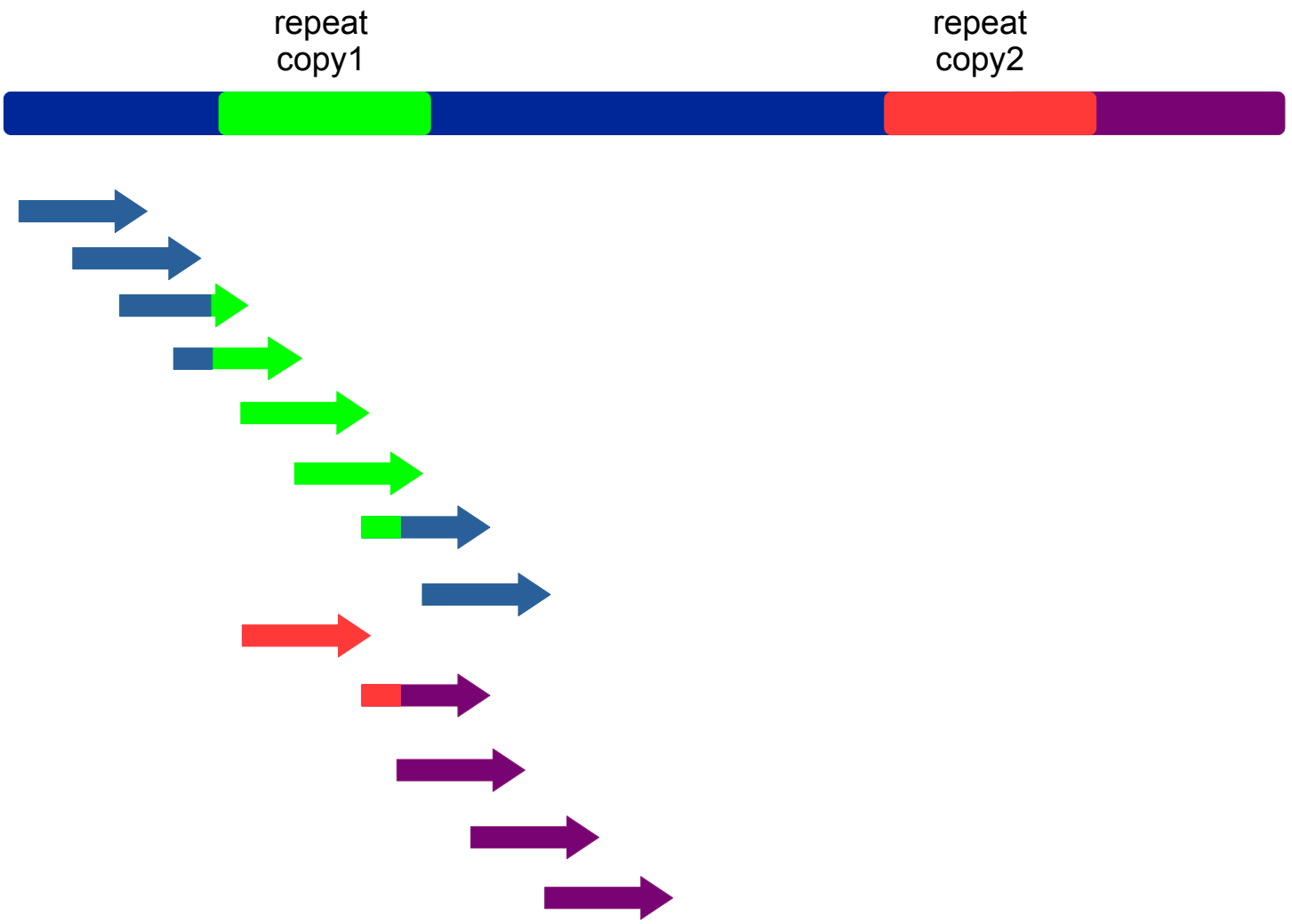
The repeat assembly problem

Only one read is usually not capable of producing the complete sequence.

Even if the problem sequence is short one read might have sequencing errors.



The repetitive problem is unsurmontable



Long reads

Read size influence

In the ideal case each piece would be a chromosome or a transcript

Mate pairs and paired-ends

Read length is critical, but constrained by sequencing technology, we can sequence molecule ends.

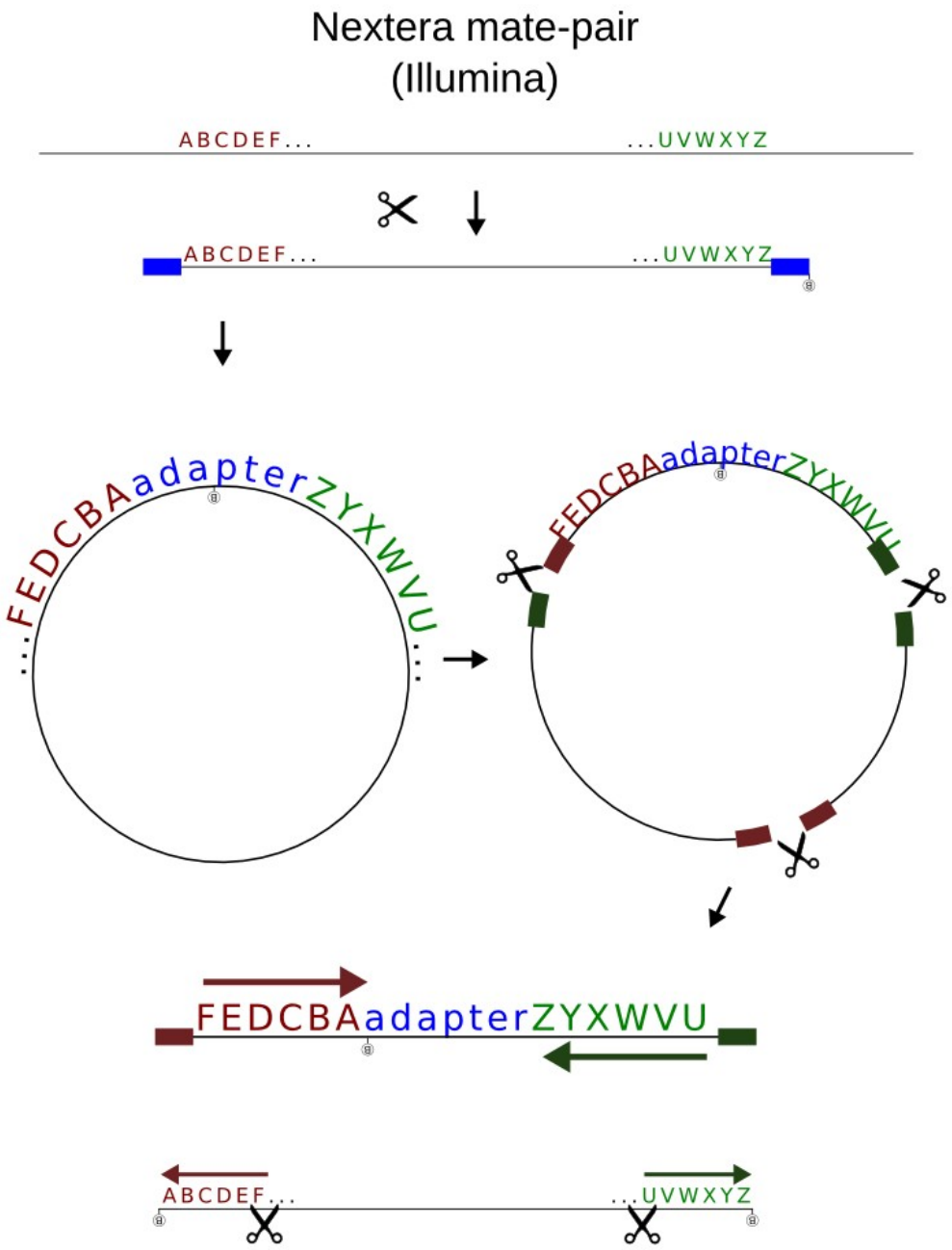
Useful for dealing with repetitive genomic DNA and with complex transcriptome structure.

Paired-ends: Illumina can sequence from both ends of the molecules. (150-500 bp)

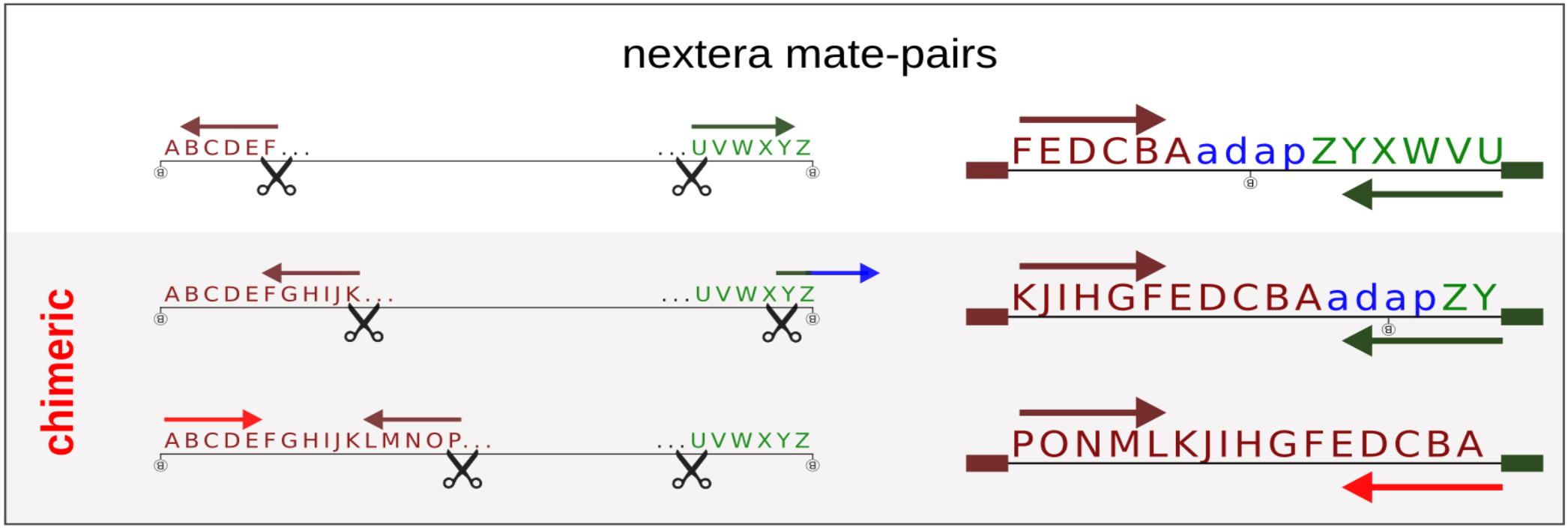
Mate-pairs: Can be generated from libraries with different lengths (2-20 Kb).

BAC-ends: Usually sequenced with Sanger.

Mate-pairs


Nextera mate-pair
(Illumina)
ABCDEF...
...UVWXYZ
ABCDEF...
...UVWXYZ
FEDCBAadapterZYXWVU
FEDCBAadapterZYXWVU
FEDCBAadapterZYXWVU
ABCDEF...
...UVWXYZ

Illumina mate-pair chimeras



Long reads

Pacbio:

- Standard
- HiFi

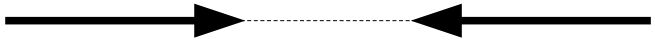
Nanopore:

- Long
- Ultralong

Library types

Single reads

Illumina Pair Ends (150-500 pb)



Mate Pairs (2-10 kb)

Illumina

Long reads

	Illumina TruSeq Synthetic Long Reads	*Pacific Biosciences*	*Oxford Nanopore Technologies*
Technology	Barcoded & Amplified Synthetic long reads	Single Molecule Real Time Sequencing	Nanopore Sequencing
Mean Length	3-5kbp	10-15kbp	5-10kbp
Raw Error Rate	0.1%	10-15%	10-30%
Costs / GB	~$2500*	~$500†	~$1000†

Third-generation sequencing and the future of genomics. Lee et al.

Short reads are harder to assemble

1. Overlap Effect: For same number of sequenced bases, shorter reads require more coverage to achieve comparable N50.

Assembly 1.
9 reads of length = 30bp.
Total sequenced bases = 270.
Assemble with min overlap = 20bp.
Result = 7 contigs.

Same inputs

Assembly 2.
3 read length = 90.
Total sequenced bases = 270.
Assemble with min overlap = 20bp.
Result = 1 contig.

2. Repeat Effect: Shorter reads resolve fewer repeats.

Repeat length = 600bp.
Read length = 800bp (Sanger).
Reads span the repeats.

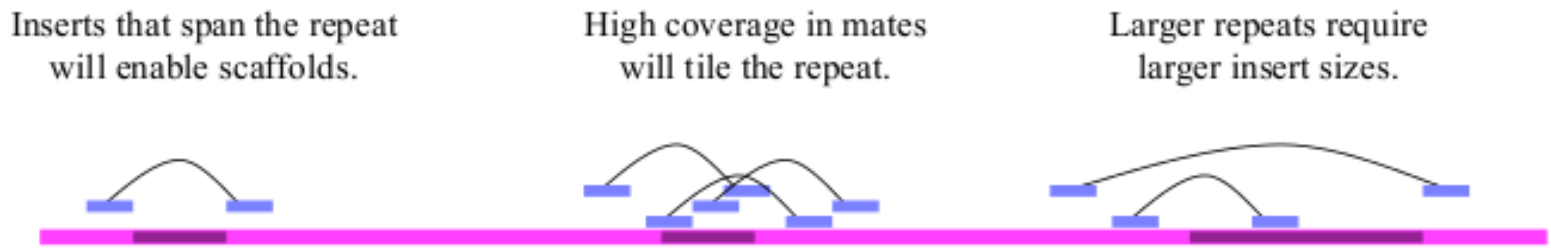
Repeat length = 600bp.
Read length = 400bp (454).
Reads bridge the repeats.

Repeat length = 600bp.
Read length = 75bp (Solexa).
Repeats not resolved.

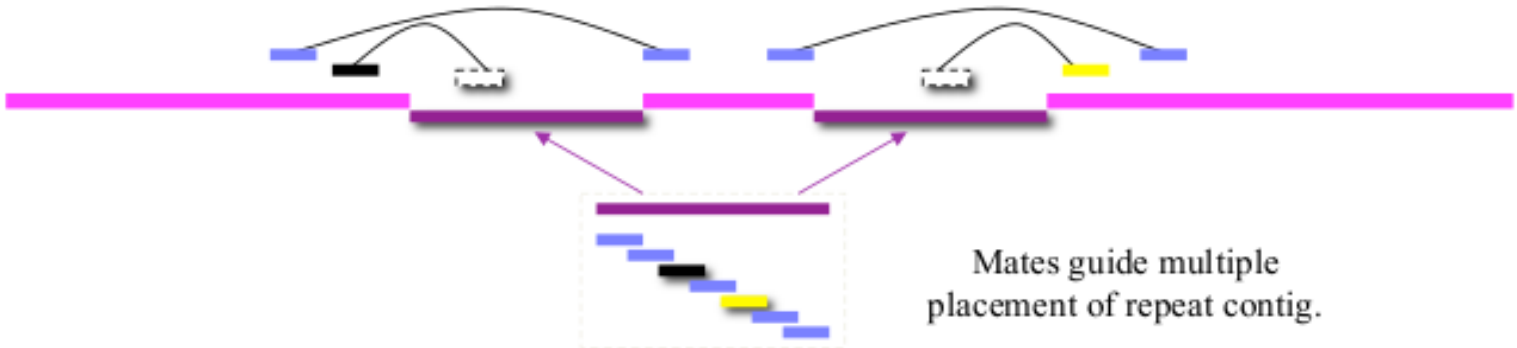
Taken from Jason Miller

The need for paired reads

1. Variety of insert sizes will span variety of repeats.



2. Mates can resolve repeats even if not possible to tile with reads.



Taken from Jason Miller

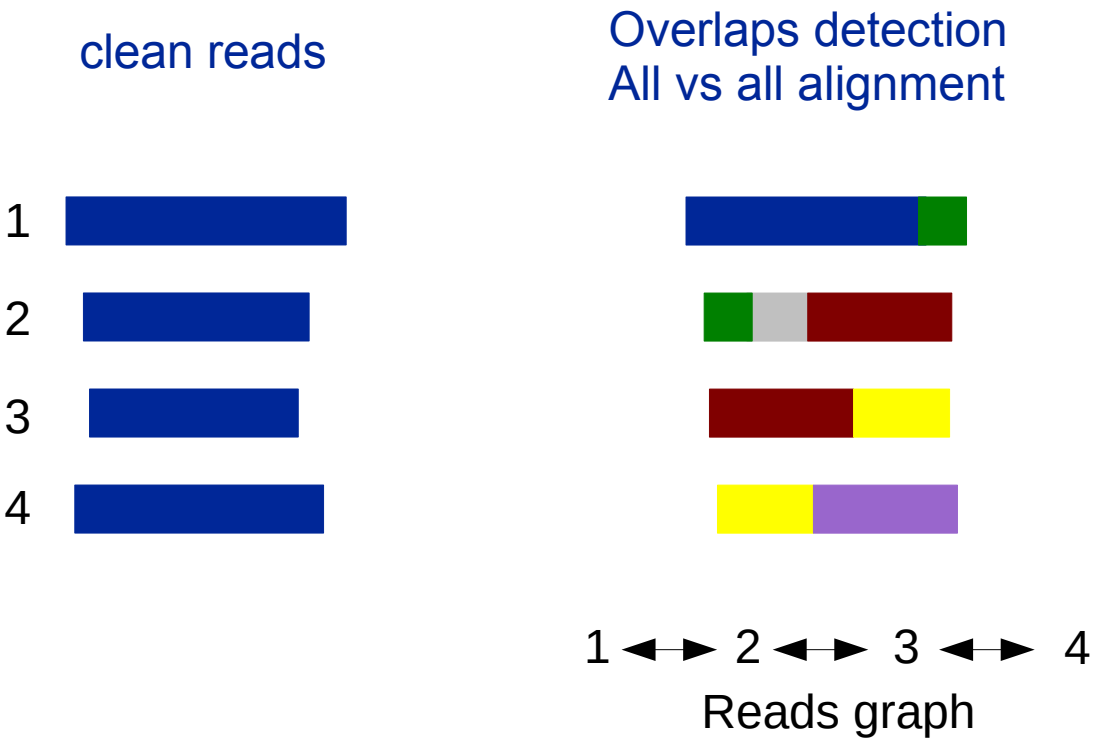
Algorithm I

Overlap – Layout - Consensus

Overlap – Layout - Consensus

Algorithm:

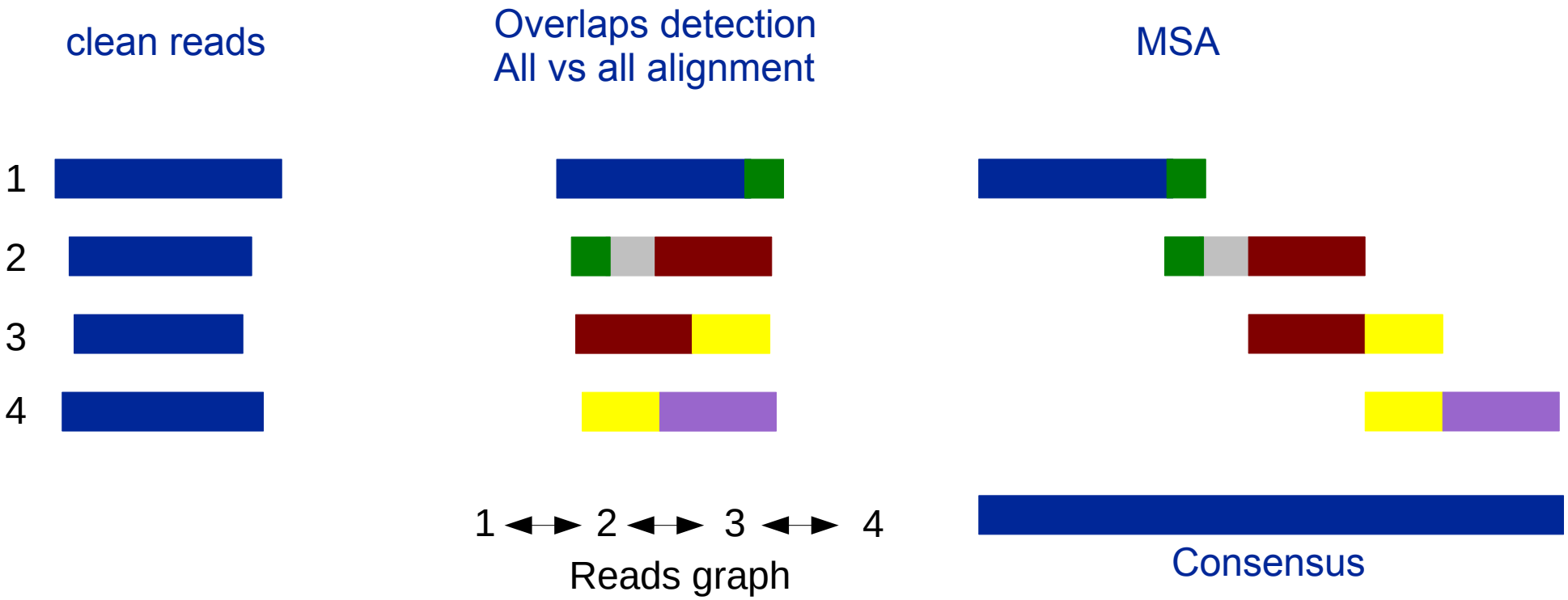
- Overlap: All-against-all, pair-wise read comparison.
- Layout: construction of an overlap graph with approximate read layout.



Overlap – Layout - Consensus

Algorithm:

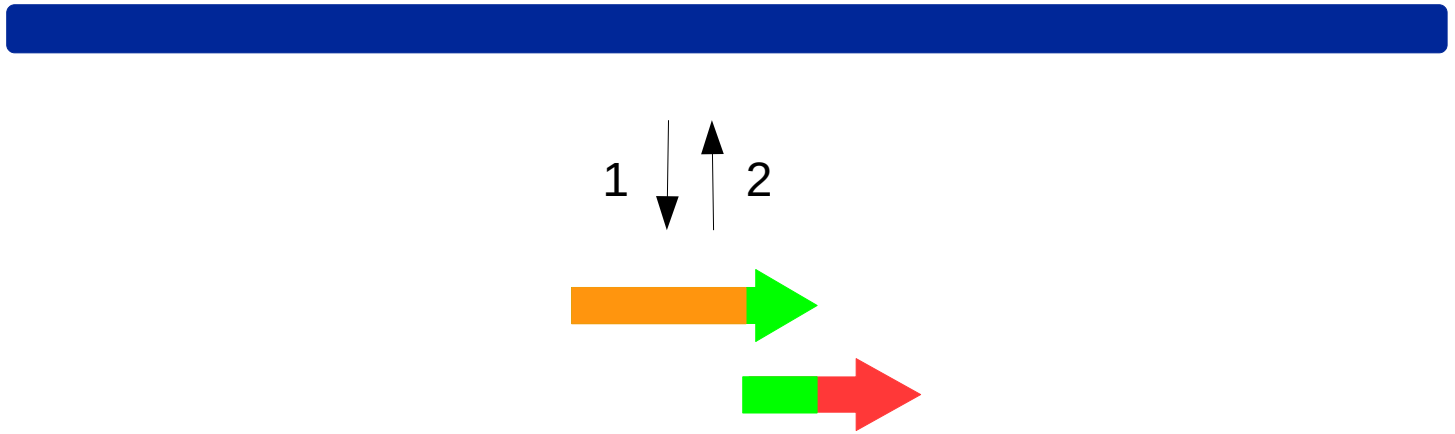
- All-against-all, pair-wise read comparison.
- Construction of an overlap graph with approximate read layout.
- Consensus: Multiple sequence alignment (MSA) determines the consensus sequence.



Overlap assumption

We are assuming:

- 1) Two reads originated from the same region will overlap
- 2) Two reads that overlap come from the same region



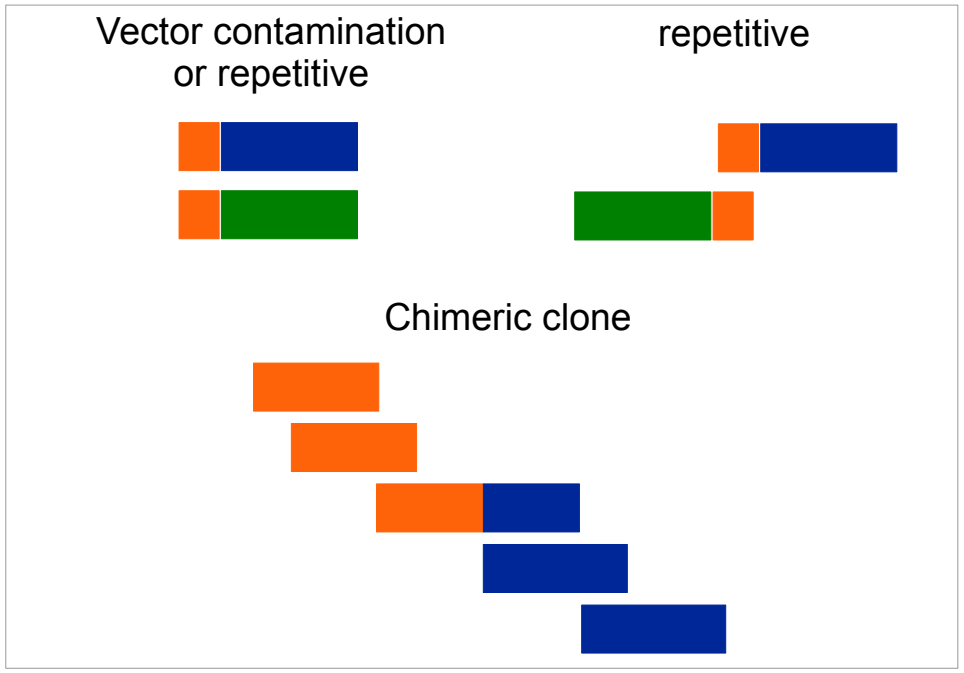
Overlap problems

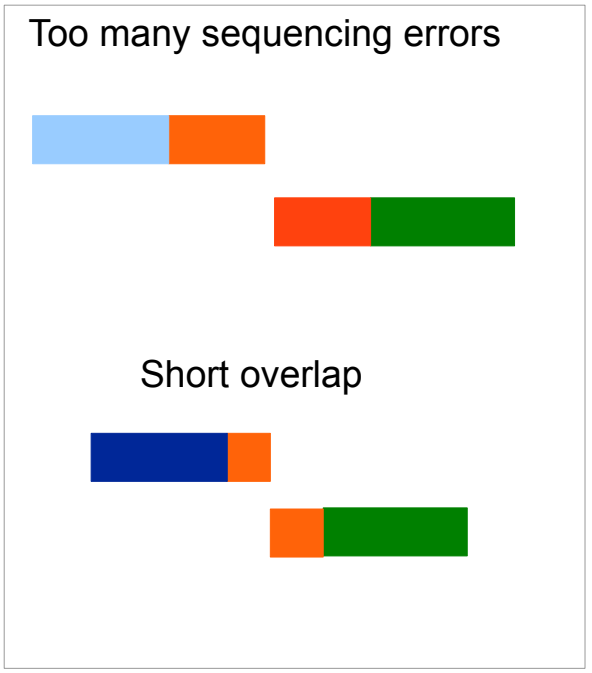
Two reads are similar if they have a good overlap.

We assume that overlap implies common origin in the genome.

This goodness depends on the overlap quality and similarity.

But several things might go wrong



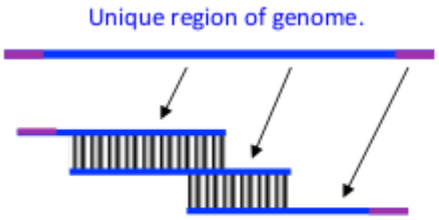


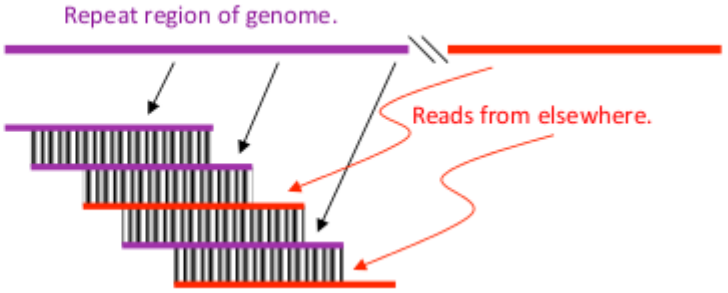
Overlap problems

False positives will be induced by chance and repeats

Avoid false positives with stringent criteria:

- Overlaps must be long enough
- Sequence similarity must be high (identity threshold)
- Overlaps must reach the ends of both reads
- Ignore high-frequency overlaps (repetitive regions in genomic?)





But stringency will induce false negatives.

Assembly terminology

Consensus

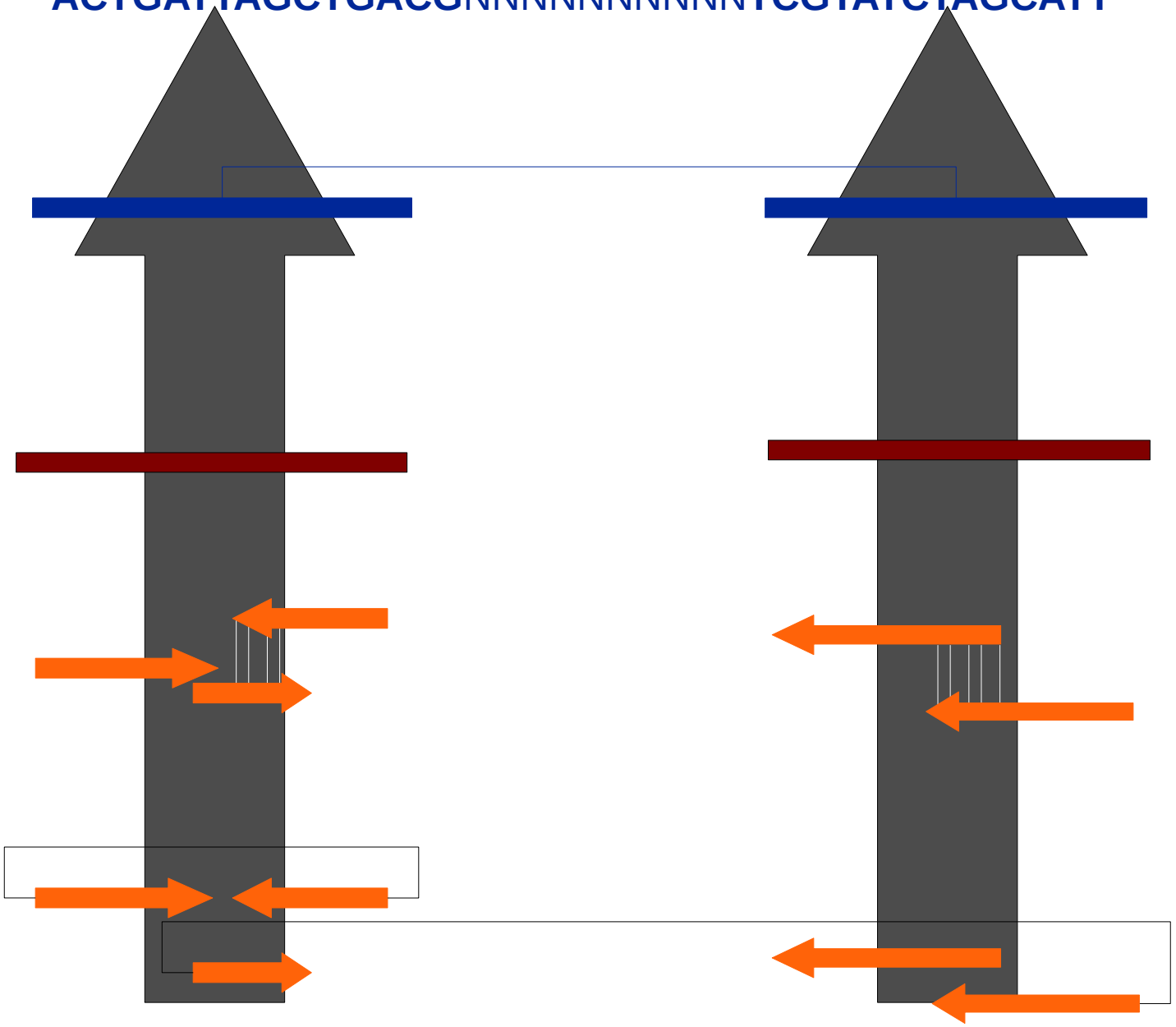
ACTGATTAGCTGACGNNNNNNNNNN**TCGTATCTAGCATT**

Scaffolds = Contigs + Parings
(gap lengths derived from pairings)

Contigs = Reads + Overlaps

Overlaps

Reads & Pairing

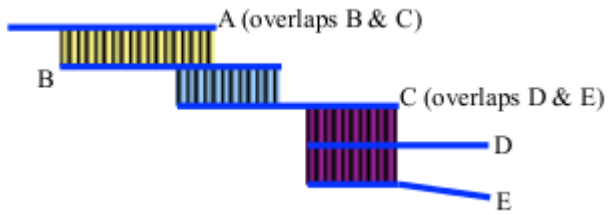


Taken from Jason Miller

Contigs

Contigs:

- High-confidence overlaps
- Maximal contigs with no contradiction (or almost no contradiction) in the data
- Ungapped
- Contigs capture the unique stretches of the genome.
- Usually where unitigs end, repeats begin



Contig example. The ideal Contig is A, B, C.
C, D and E have a repetitive sequence.

Scaffolds

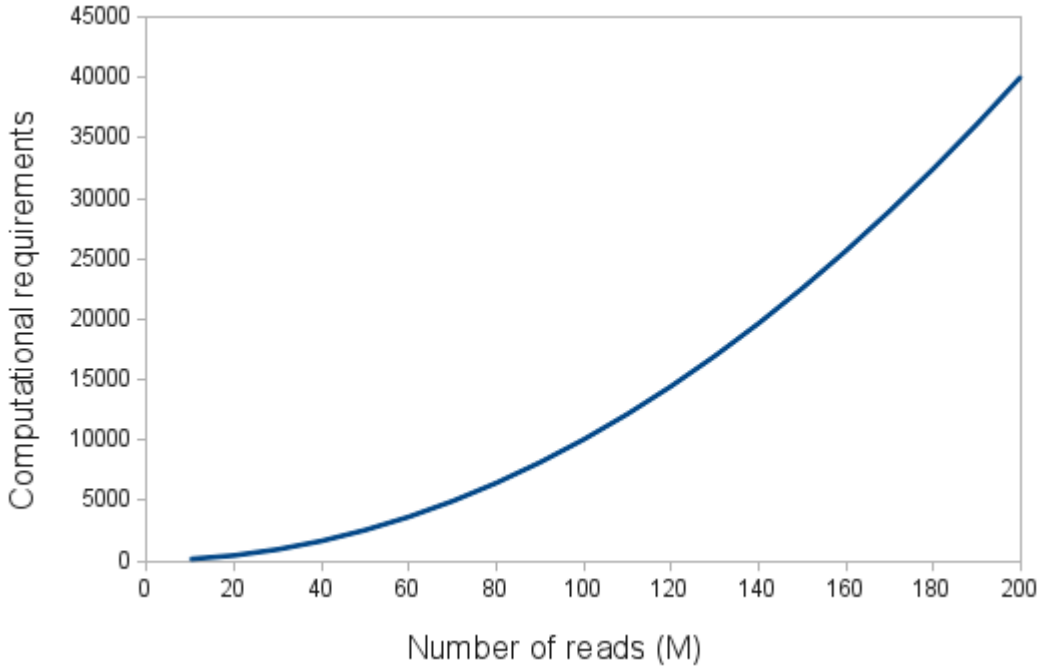
Build with mate pairs

- Mate pairs help resolve ambiguity in overlap patterns

Scaffolds

- Every contig is a scaffold
- Every scaffold contains one or more contigs
- Scaffolds can contain sequence gaps of any size (including negative)

Overlap-Consensus-Layout limitation



30 Million 5Kb (long) reads

Number of alignments: 30 M reads x 30 M reads = 900 trillion alignments (It does not scale!!)

Main requirements:

Memory.

Time.

The kmer-based assemblers requirements depend on the genome complexity and not so much on the reads.

Algorithm II

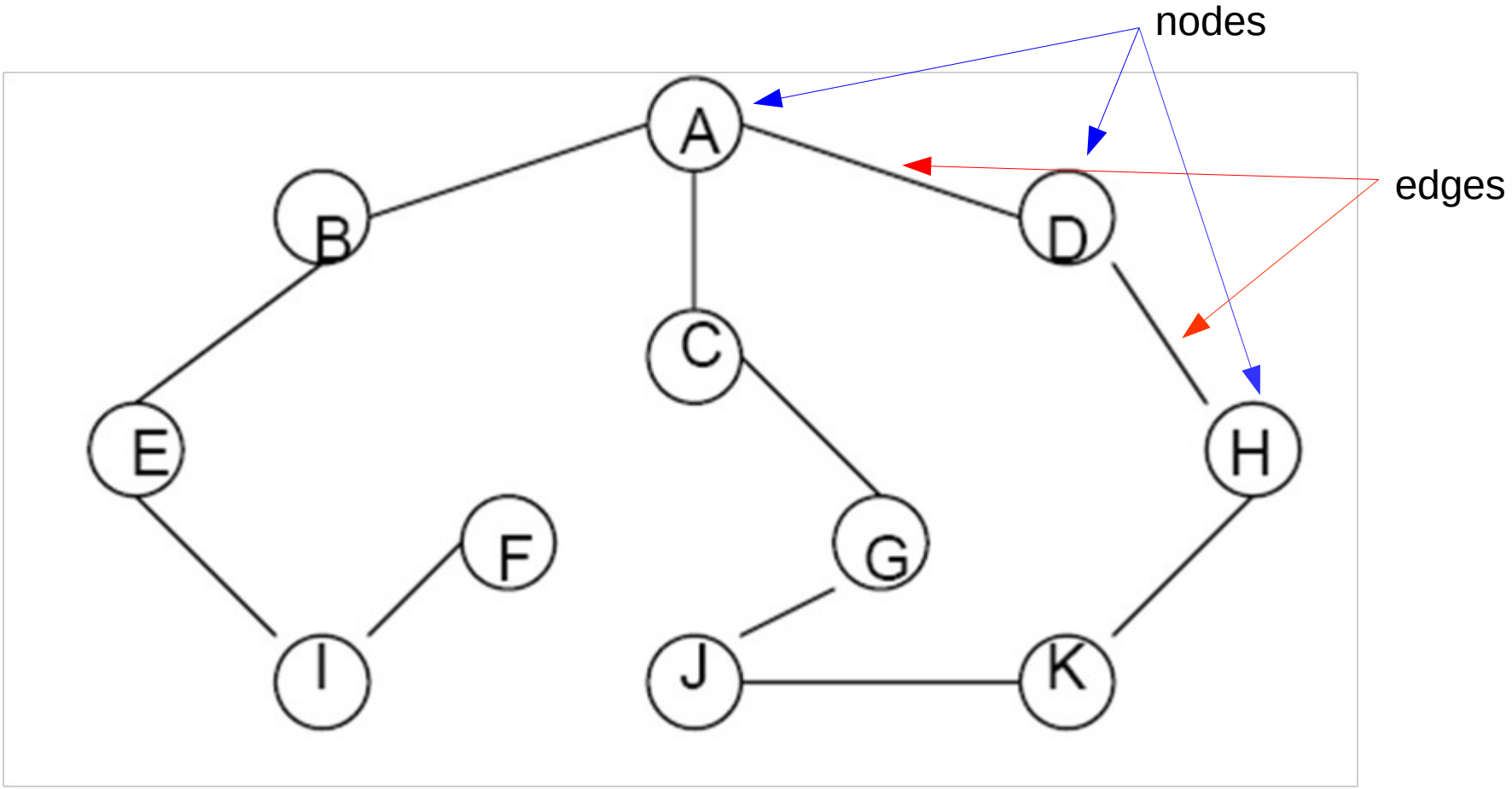
kmer-based

K-mer based assembly

```
Seq              ACTGGTCAT
K-mers 5pb       ACTGG
                  CTGGT
                   TGGTC
                    GGTCA
                     GTCAT
```

K-mer based assembly

K-mer graphs are de Bruijn **graphs**.



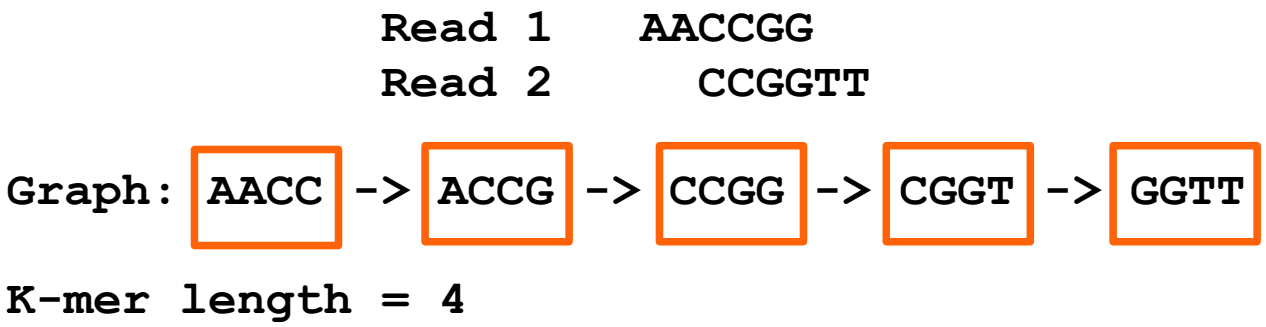
K-mer based assembly

K-mer graphs are de Bruijn graph.

Nodes represent all K-mers.

Edges join consecutive K-mers.

Assembly is reduced to a graph reduction problem.



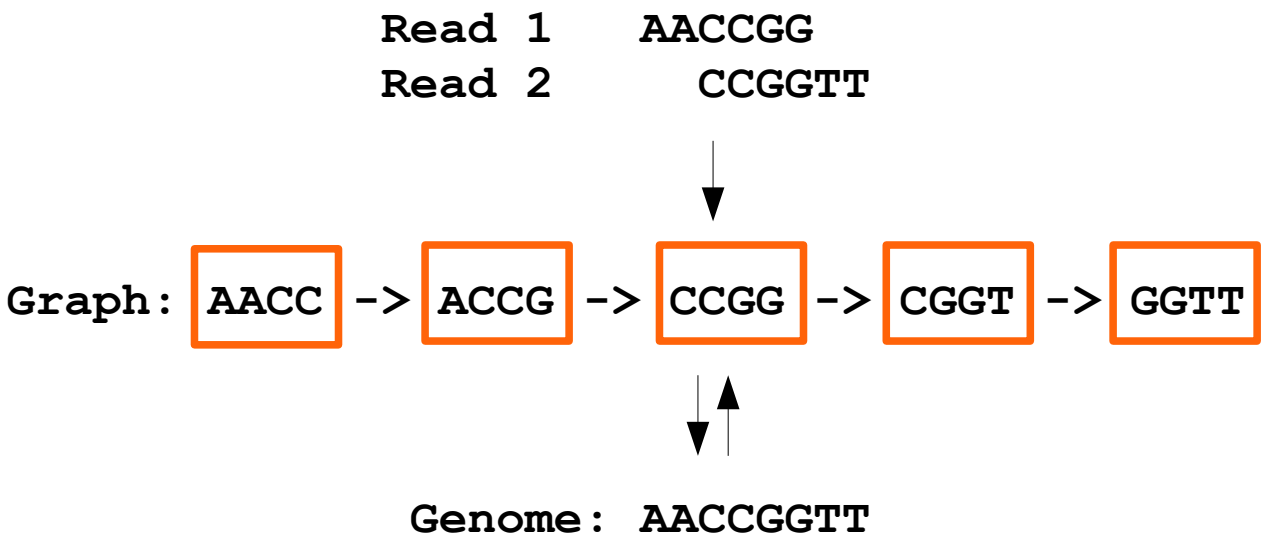
K-mer based assembly

K-mer graphs derived from reads and genome is very similar

From the graph we can recreate the genome sequence

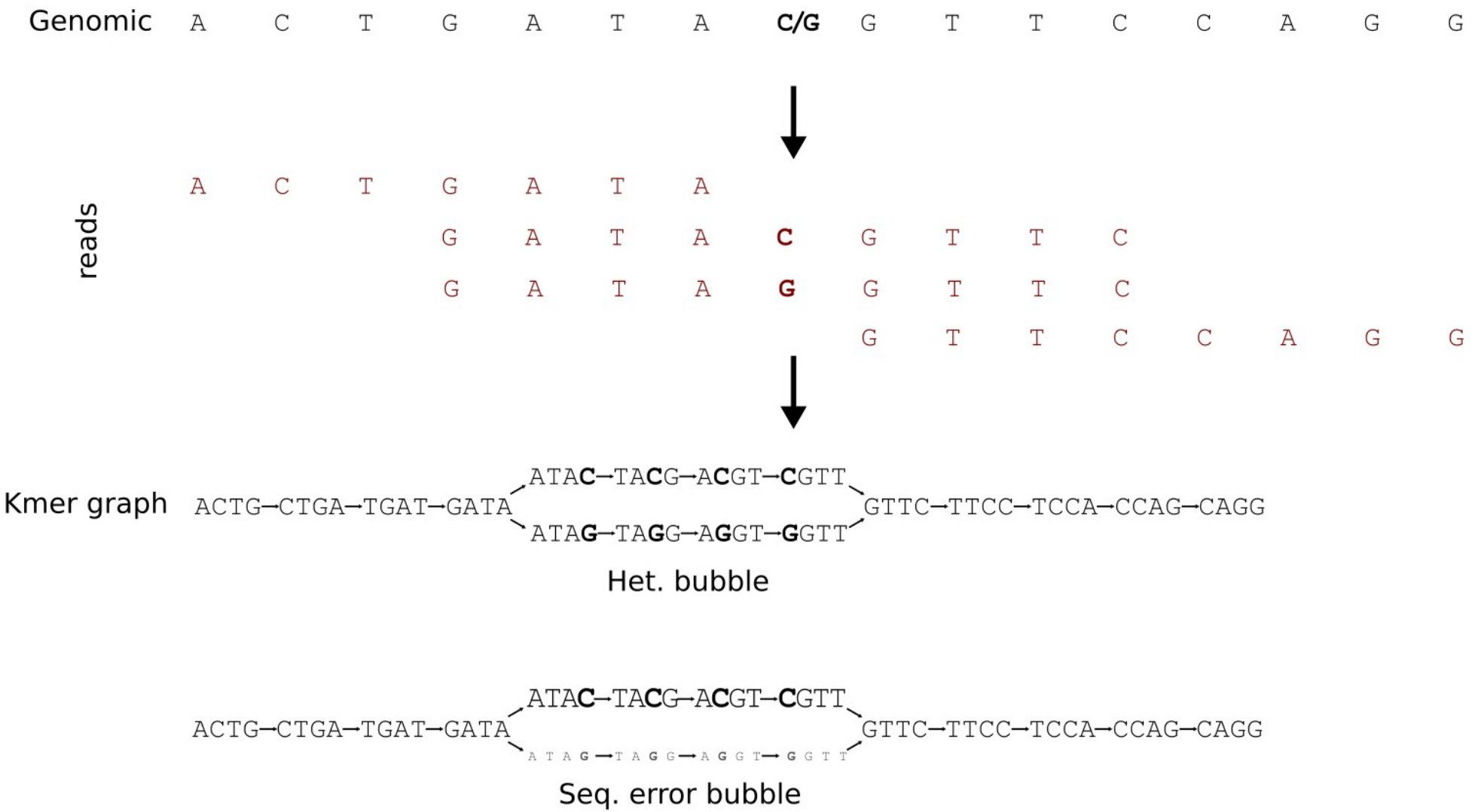
K-mer graph size depends, mostly on genome size, not on number of reads

Reads are only read once to prepare the kmer graph

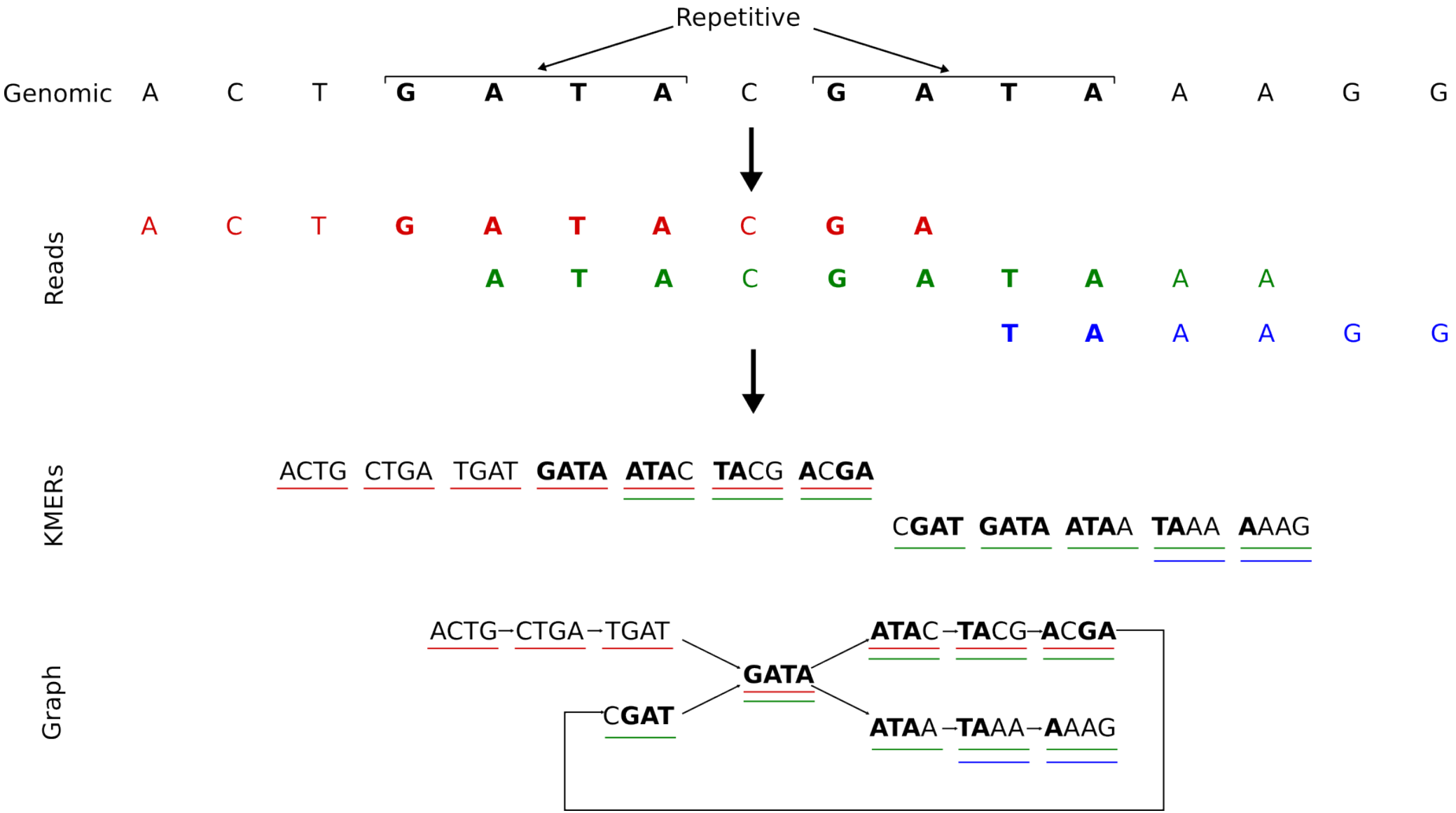


K-mer based assembly

K-mer bubbles



K-mer repetitive



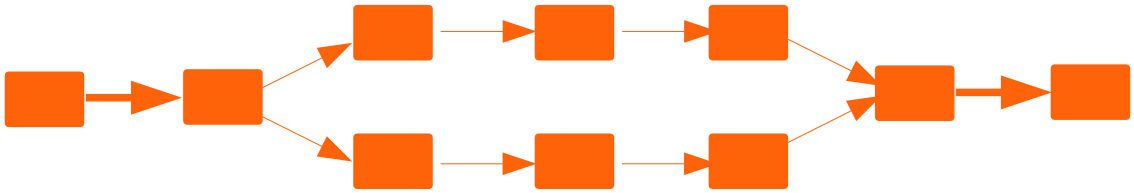
Graph problems

Repeats, sequencing errors and polymorphisms increase graph complexity, leading to tangles difficult to resolve

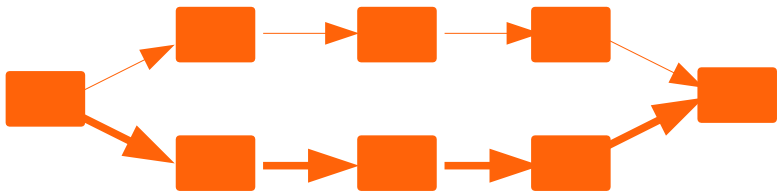
K-mer graphs are more sensitive to repeats and sequencing errors than overlap based methods.

Optimal graph reductions algorithms are NP-hard, so assemblers use heuristic algorithms

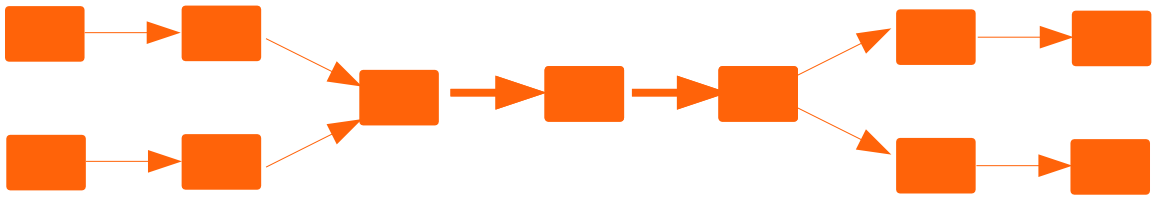
Polymorphism



Sequencing error



Repeat



Quality Assessment

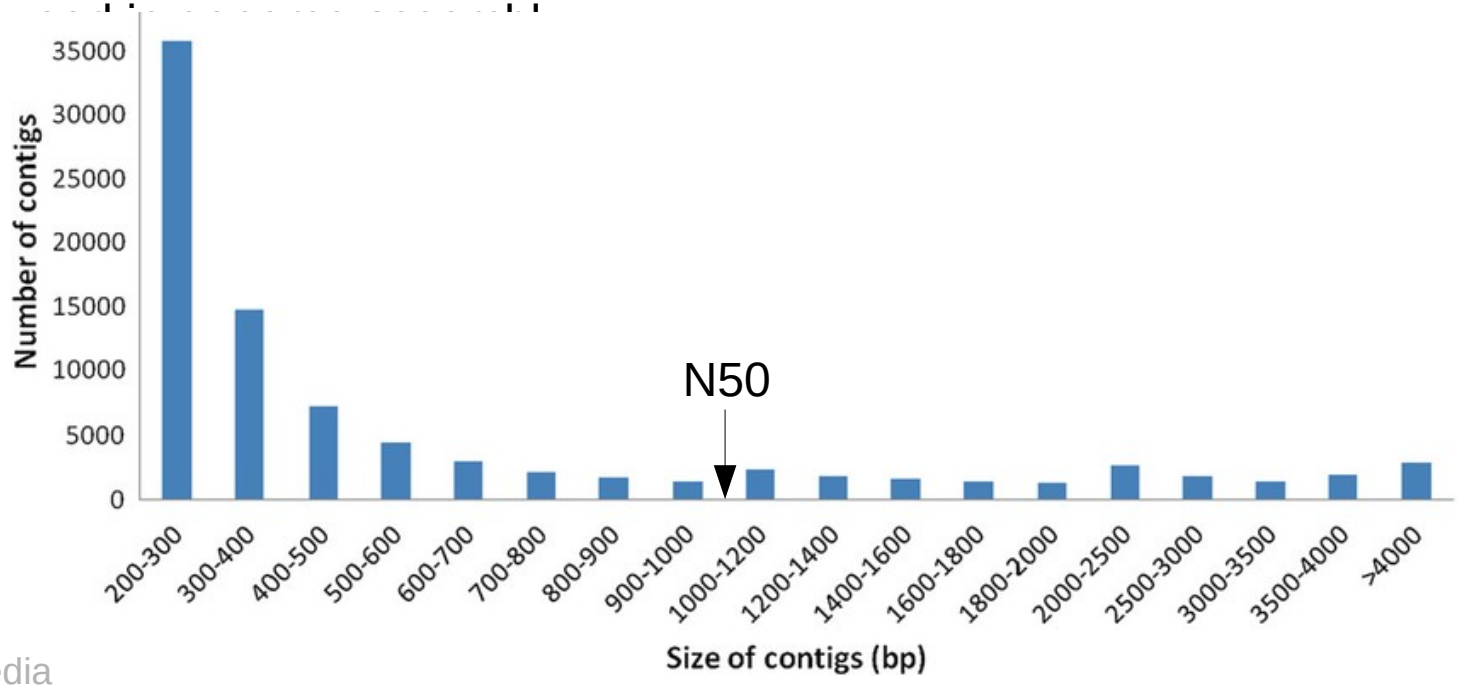
N50

N50 is defined as the contig length such that using equal or longer contigs produces half the bases of the assembly.

- NG50: 50% of the genome (if available)

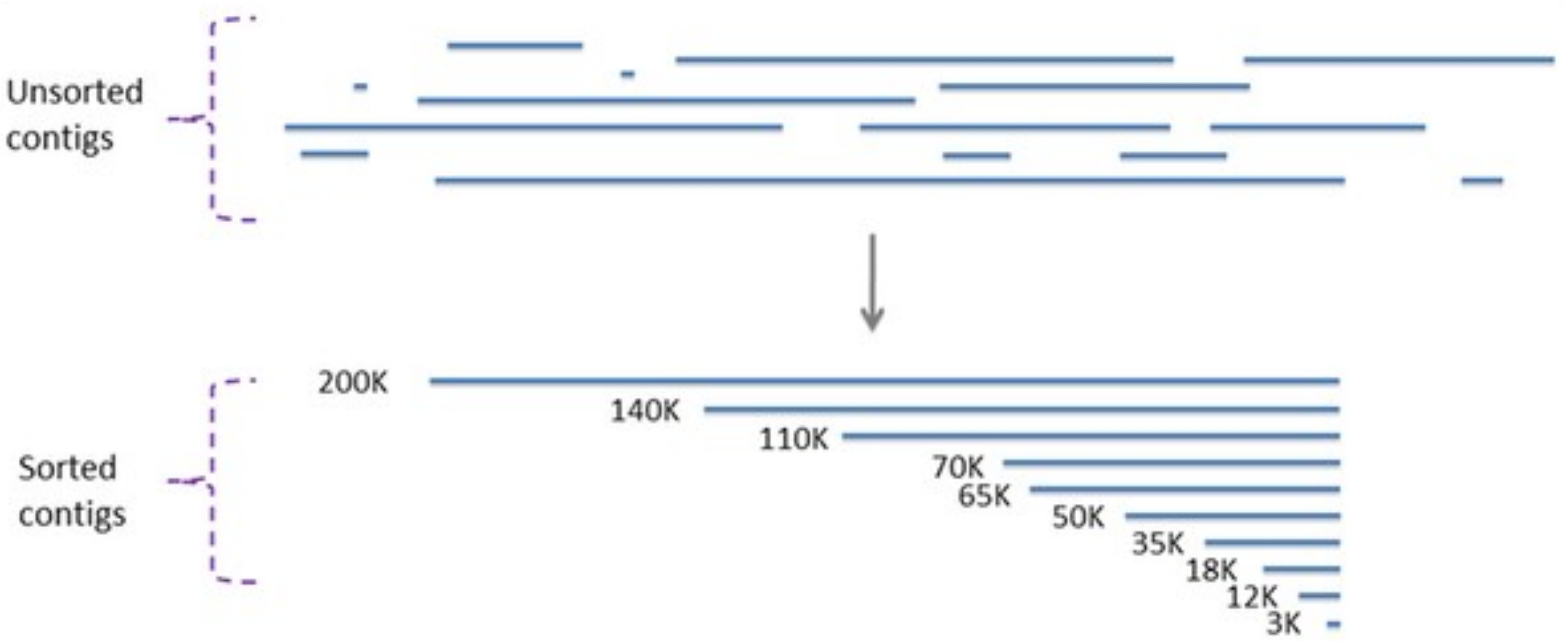
The N50 statistic is a measure of the average length of a set of sequences, with greater weight given to longer sequences.

It is widely



Taken from Wikipedia

N50

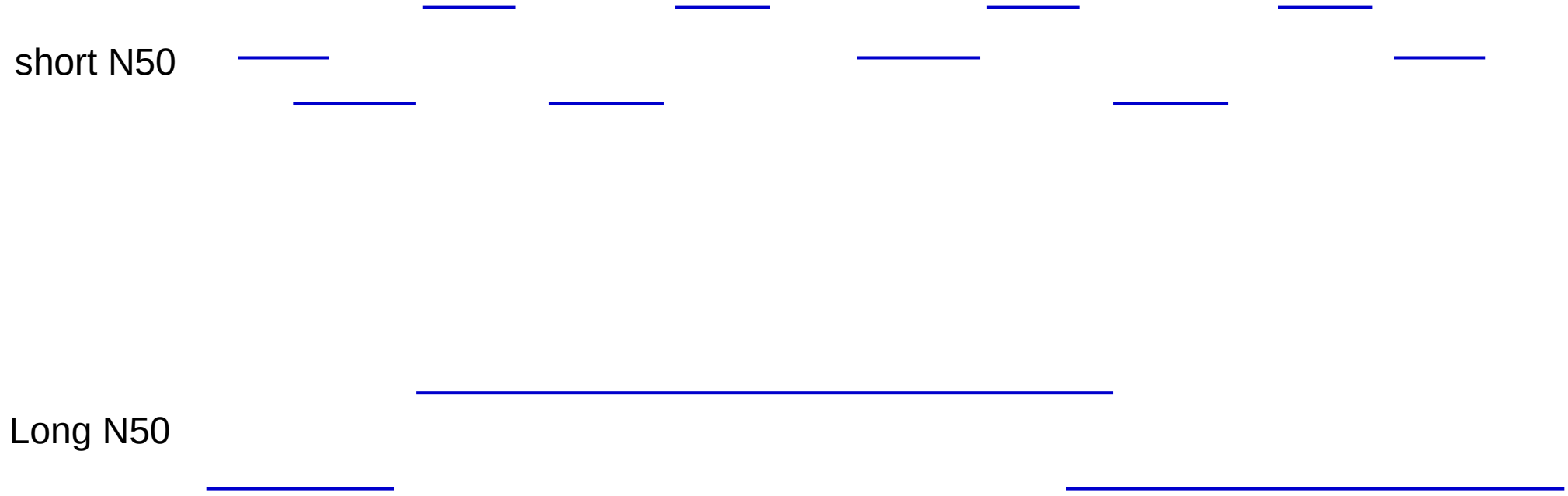


Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

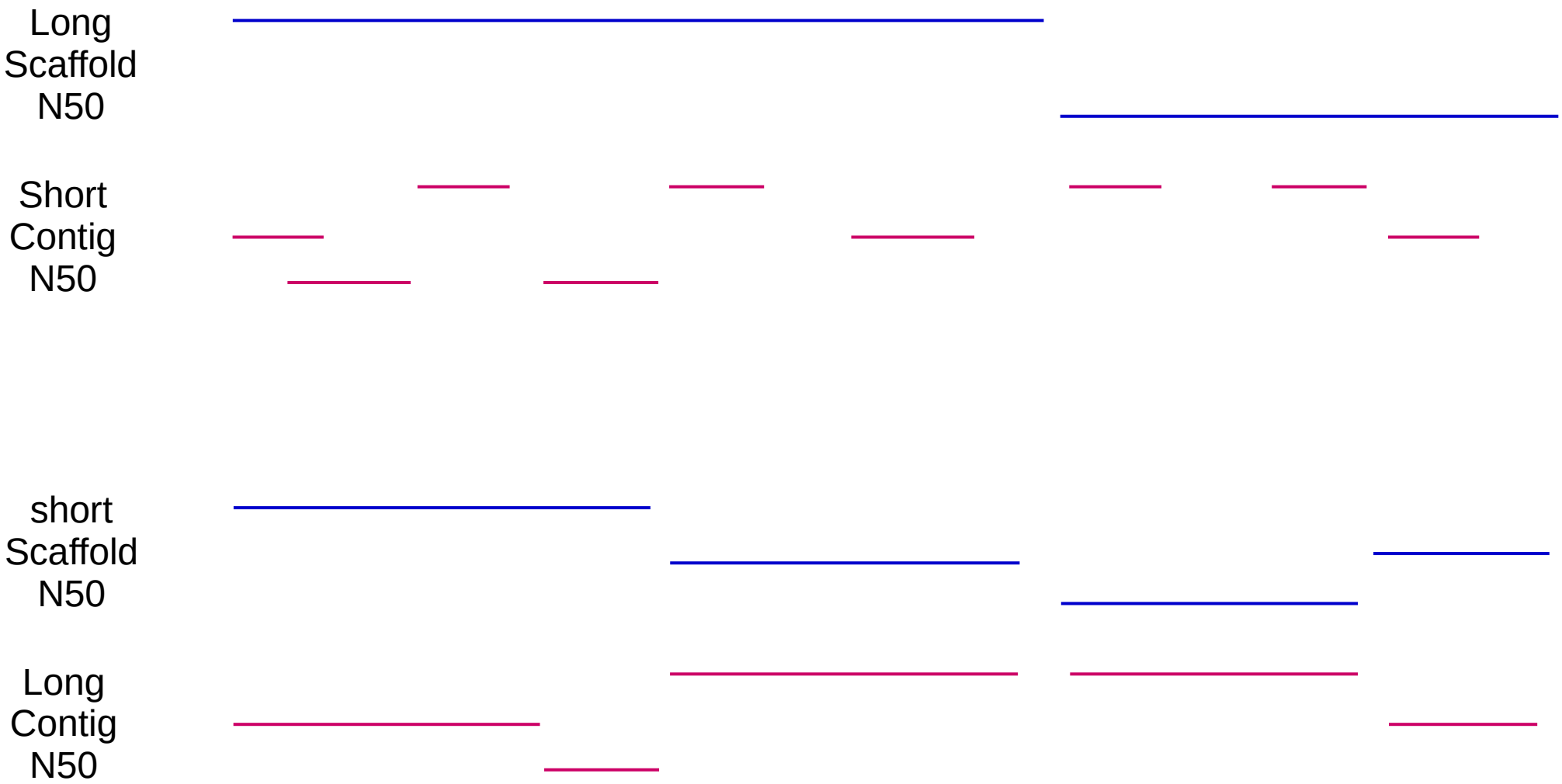
50% total contig length= 703K x 50%= 351.5K

∵200K+140K +110K> 351.5K，∴N50= 110K

N50

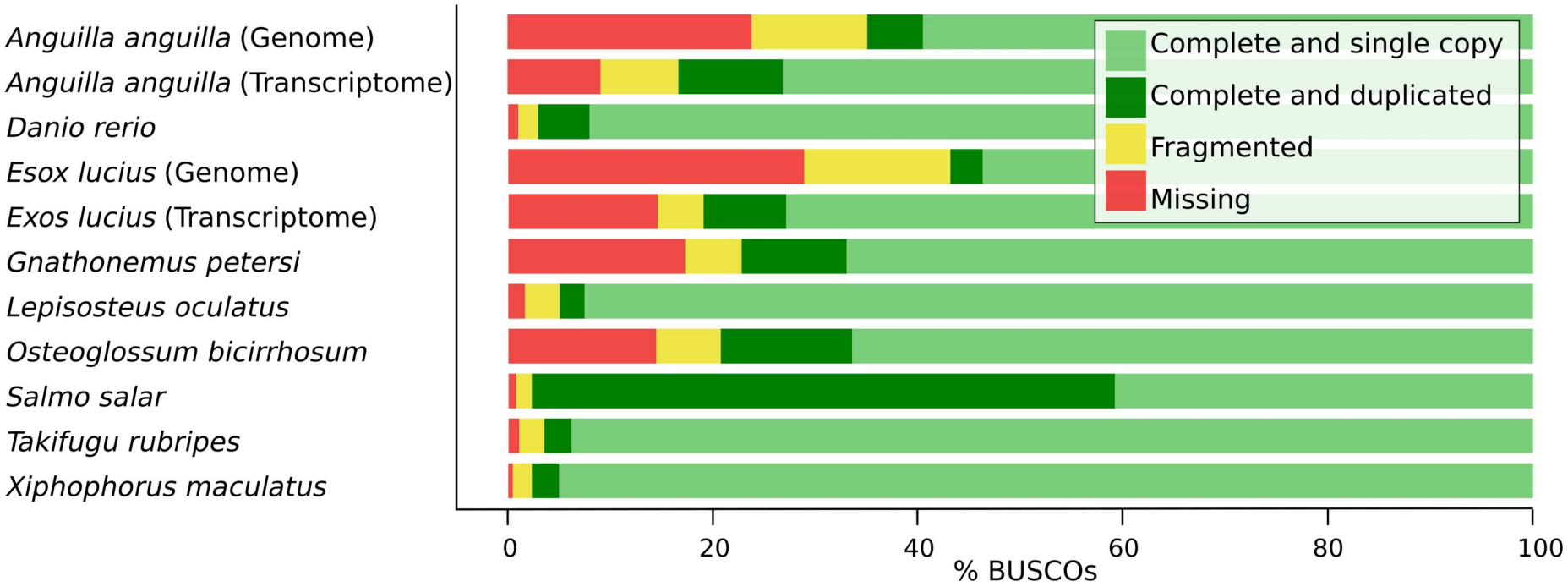


N50 for scaffolds and contigs



BUSCO orthologs

http://busco.ezlab.org/



Mapping reads against the assembly

Reads used to create the assembly

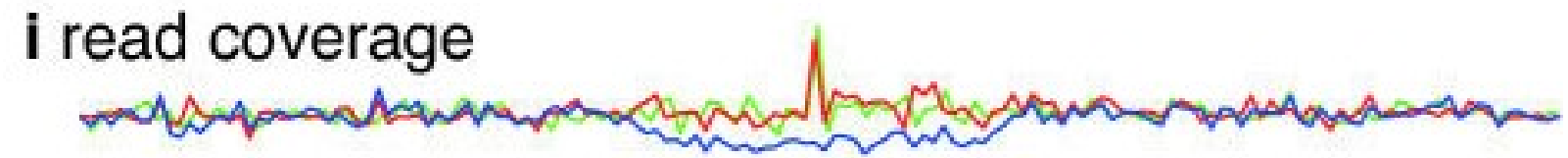
Other reads from the same species

Percentage of reads mapped

- Most reads should properly map

Taken from REAPR paper: https://doi.org/10.1186/gb-2013-14-5-r47

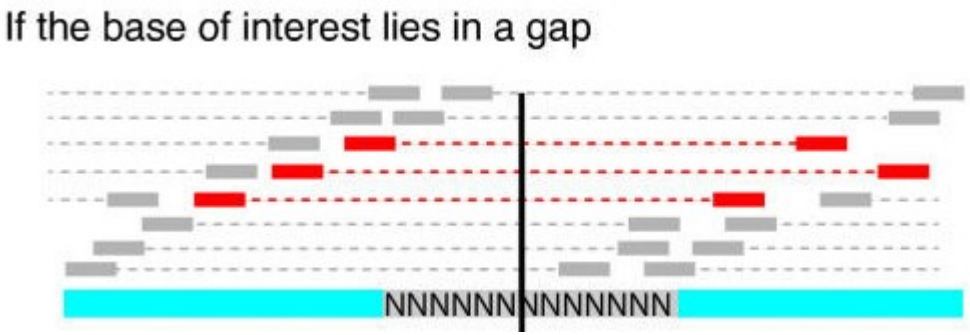
Mapping reads against the assembly



Properly paired

Orphaned

Reads only partially aligned



REAPR corrected N50

Taken from REAPR paper: https://doi.org/10.1186/gb-2013-14-5-r47

Main assembly hurdles

Main Problems:

- Short, inaccurate sequence reads
- genomic repeats
  - Difficulty of the assembly depends a lot on the genome: easy for bacteria, very difficult for long highly repetitive polyploids

Main solutions:

- Long reads
- Low sequencing error rates:
  - Simplify graph problem
  - Complex graphs typically yield worse assemblies
  - Allow to differentiate between repeats

Common assembly errors

Collapsed repeats

- Align reads from distinct (polymorphic) repeat copies
- Fewer repeats in assembly than in genome
- Fewer tandem repeat copies in assembly than genome

Missed joins

- Missed overlaps due to sequencing error
- Contradictory evidence from overlaps
- Contradictory or insufficient evidence from pairs

Chimera

- Enter repeat at copy 1 but exit repeat at copy 2
- Assembly joins unrelated sequences

Ingredients for Good Assembly

Coverage and read length

- 20X for (corrected) PacBio, 150X for Illumina

Paired reads

- Read lengths long enough to place uniquely
- Inserts longer than long repeats
- Pair density sufficient to traverse repeat clusters
- Tight insert size variance
- Diversity of insert sizes

Read Quality:

- Sequencing errors
- No vector contamination
- Contamination: mitochondrial or chloroplastic

The requirements depend greatly on the quality: it is not the same a draft that high quality genome.

Common assemblers

Genomic

- SOAPdenovo, Illumina
- Canu: Pacbio and Illumina
-

Staden for Sanger reads.

Transcriptomic assemblers are specialized software

Assembly comparison

Empirical evaluation of methods for de novo genome assembly

- https://peerj.com/articles/cs-636/

The best assemblers and parameters depend a lot on the genome and on the information available

Arabidopsis thaliana:

- SOAPdenovo2 produces much larger contigs than any other assembler
- it has many assembly errors
- SPAdes appeared to be preferable

Bacillus cereus

- SOAPdenovo2 has the fewest contigs, as well as a low N50 value
- its error rate was among the best

Human genome

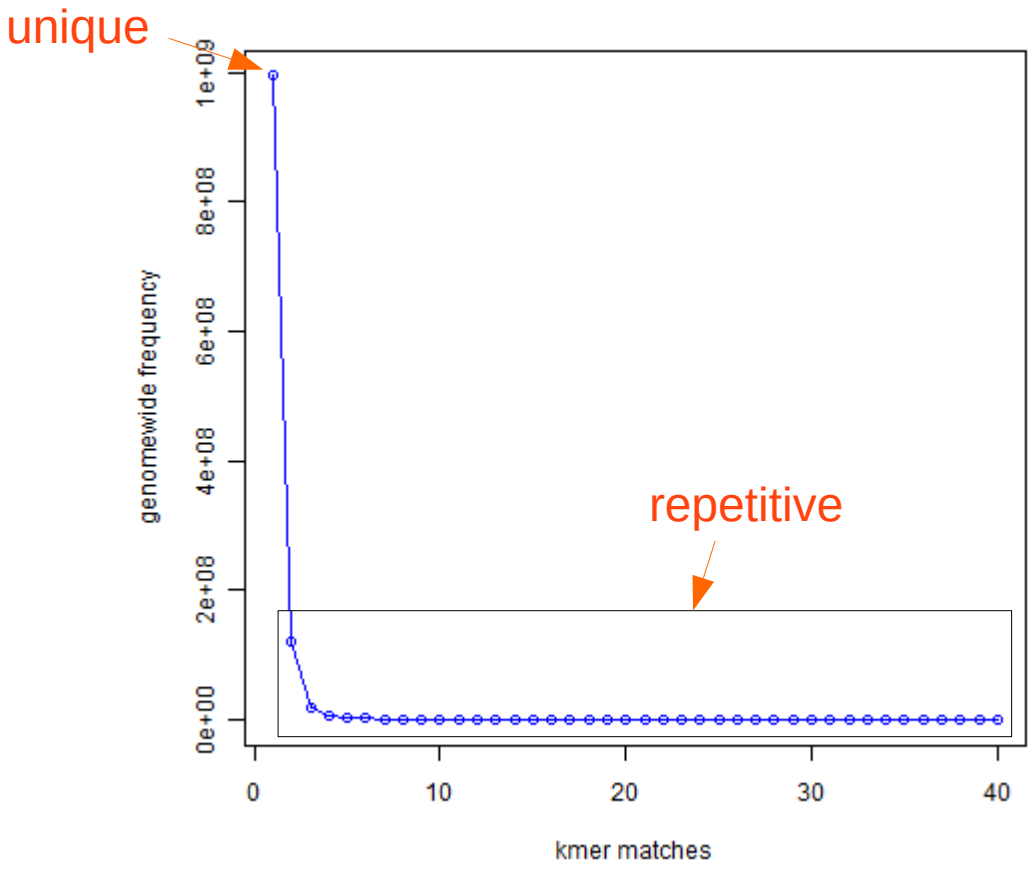
- HiFiasm outperformed other assemblers
- problems with mismatches and misassemblies.

Flye has consistently demonstrated superior output based on contig size, with the trade-off between scale and error rate. Hinge and Canu, though they demonstrate more errors than Flye, both performed reasonably well too.

K-mer analysis

K-mer analysis

21-mer distribution of sea urchin (Strongylocentrotus purpuratus) genome (900 Mbase long)



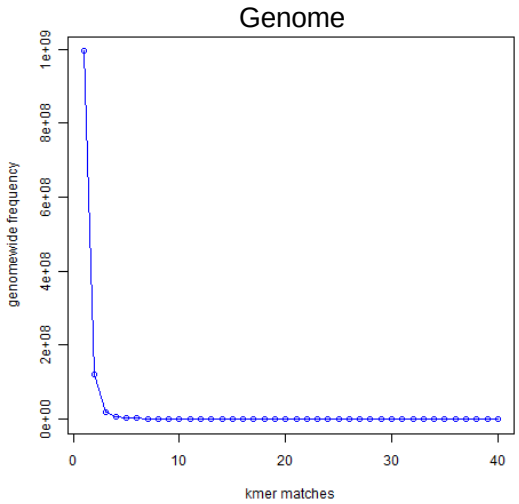
K-mer analysis

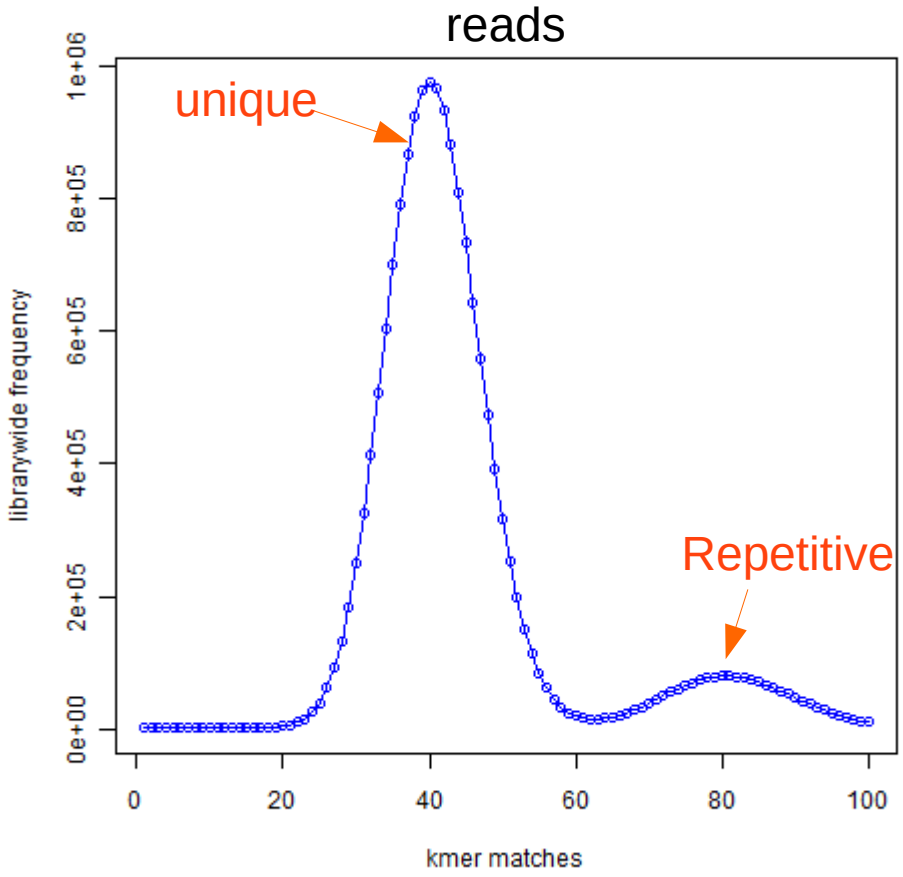
K-mer distribution

Simulated reads.

40X coverage

No errors





(duplication)

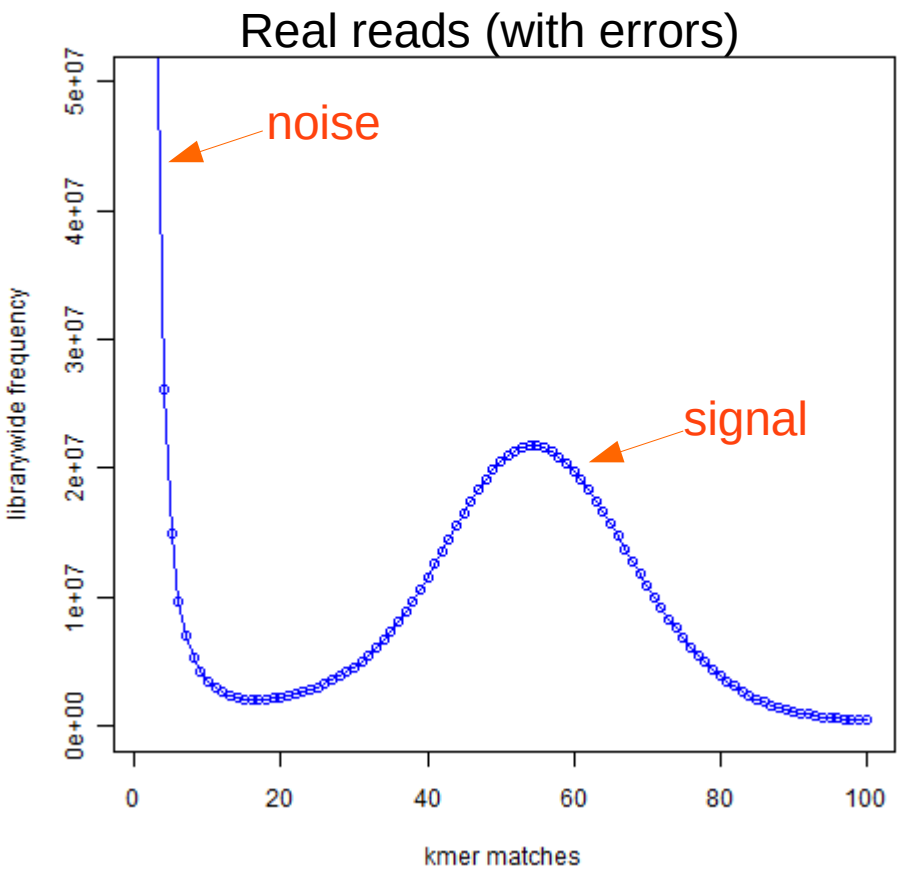
K-mer analysis

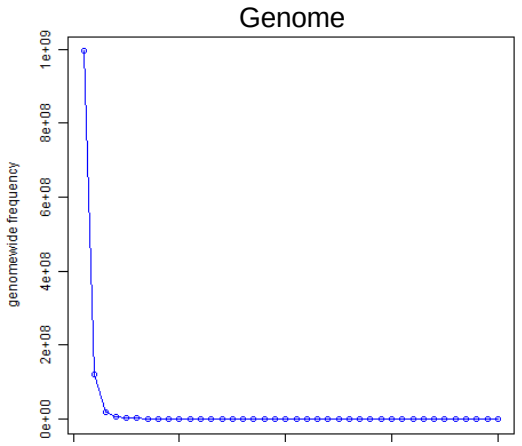
K-mer distribution

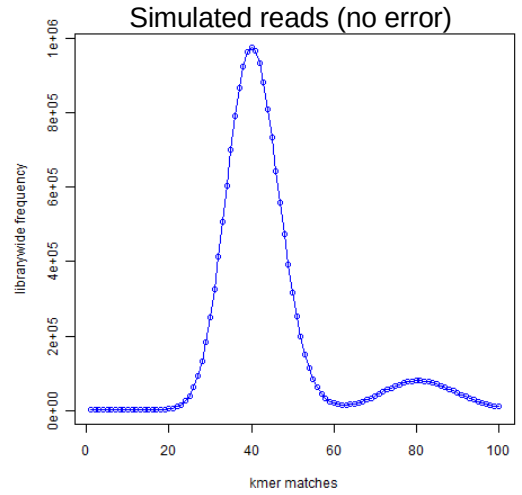
real reads.

50X coverage

With errors

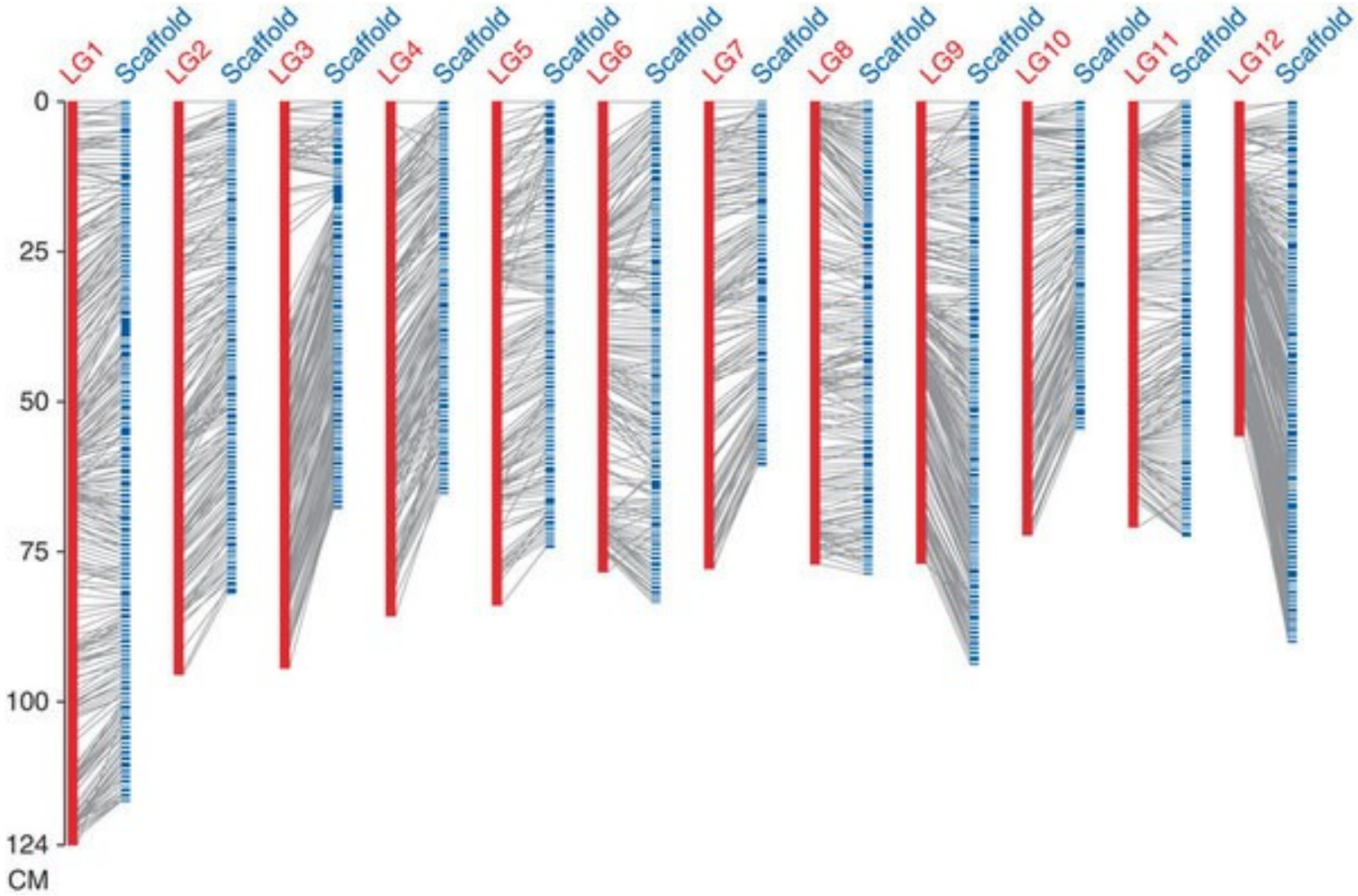






From scafflods to chromosomes

Genetic maps



Mapping platforms

	BioNanoGenomics	*10X Genomics*	*Dovetail cHiCago*
Technology	Optical mapping of fluorescent probes	Barcoded "Read Clouds"	Chromatin mate-pairs
Mean Span	100-250kbp	30-100kbp	25-100kbp
Error Modes	Fragile sites, incomplete labeling	Barcode reuse, Short read mapping	Variable span, short read mapping
Costs / Mammalian	~$3,000	~$2,000*	~$10,000*

Third-generation sequencing and the future of genomics. Lee et al.

Bionano

1 **Sequence-specific labeling**

Nickase (Nt.BspQI)

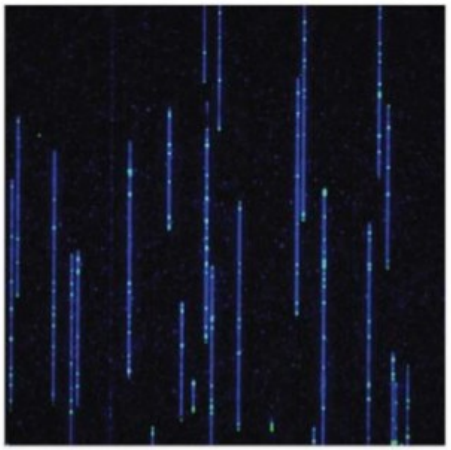
5′–ATGC***GCTCTTC***CATGAATGCGAGC–3′
3′–TACG***CGAGAAG***GTACTTACGCTCG–5′

Nick labeling

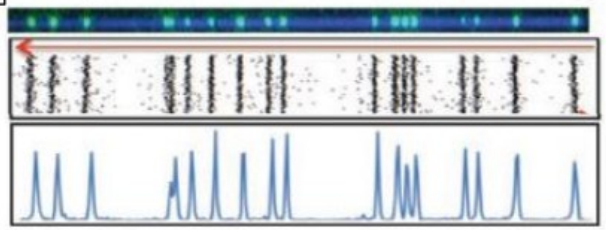
ATGAATGC

5′–ATGC***GCTCTTC***CATGAATGCGAGC–3′
3′–TACG***CGAGAAG***GTGCTTACGCTCG–5′

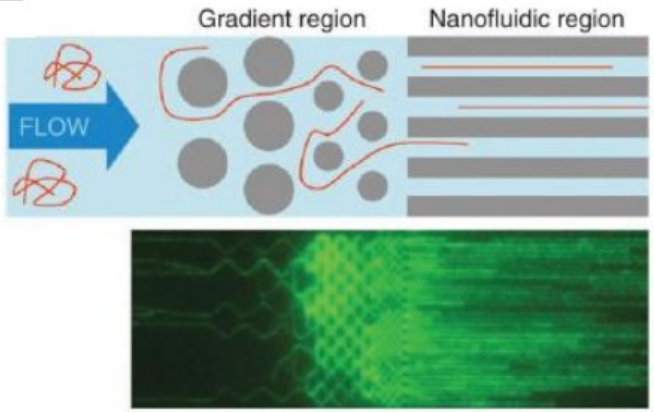
3 **Fluorescence imaging**



4 **Map construction**



2 **DNA linearization**

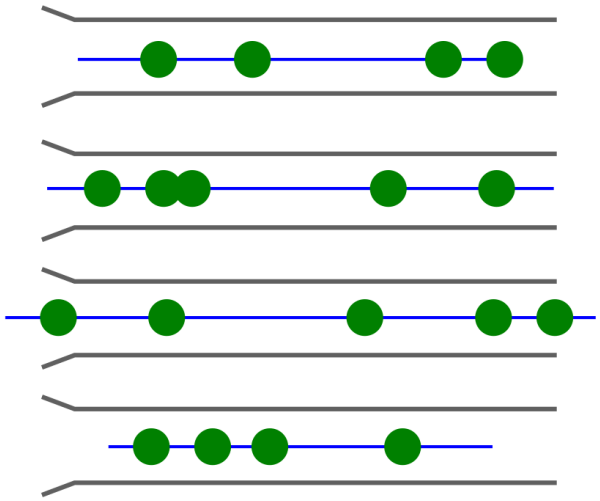


Lam et al., Nat. Biotechnol. 30(8) 2012

Output:

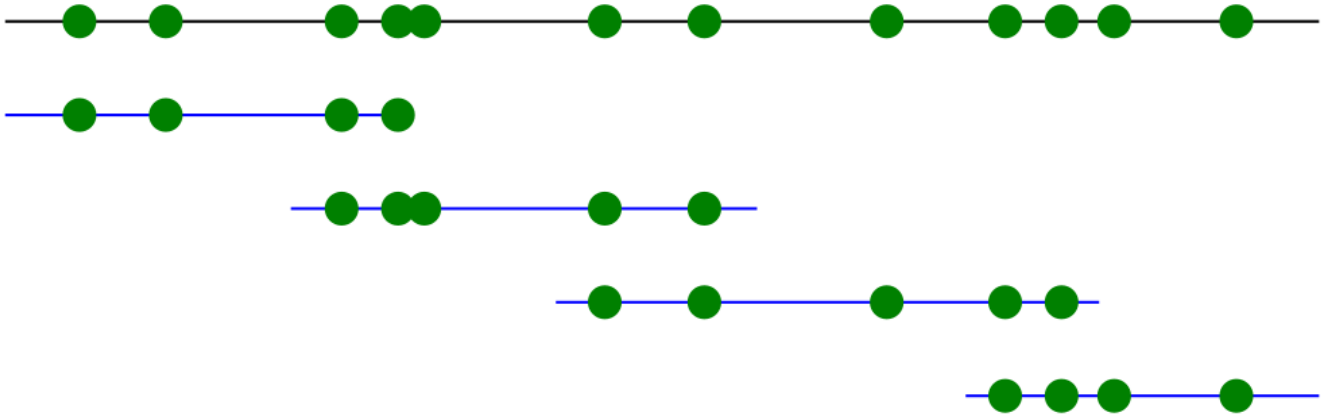
sequence motif (GCTCTTC) map along hundreds-kb to megabase DNA stretches

Bionano



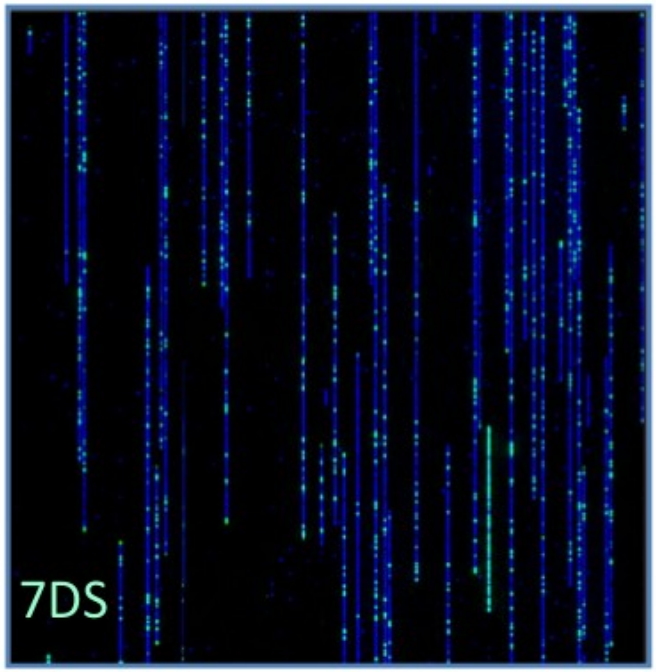
1

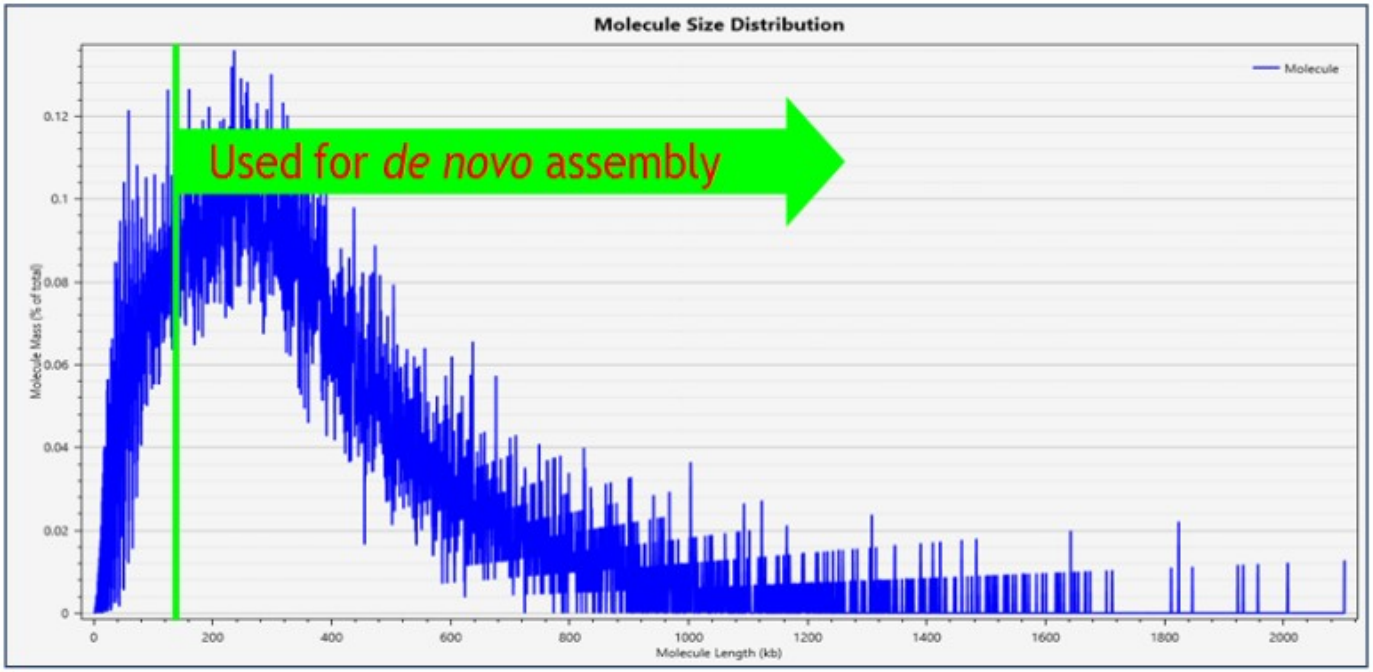
Bionano helps you solve long range structures
(although not chromosome wide assemblies)



2

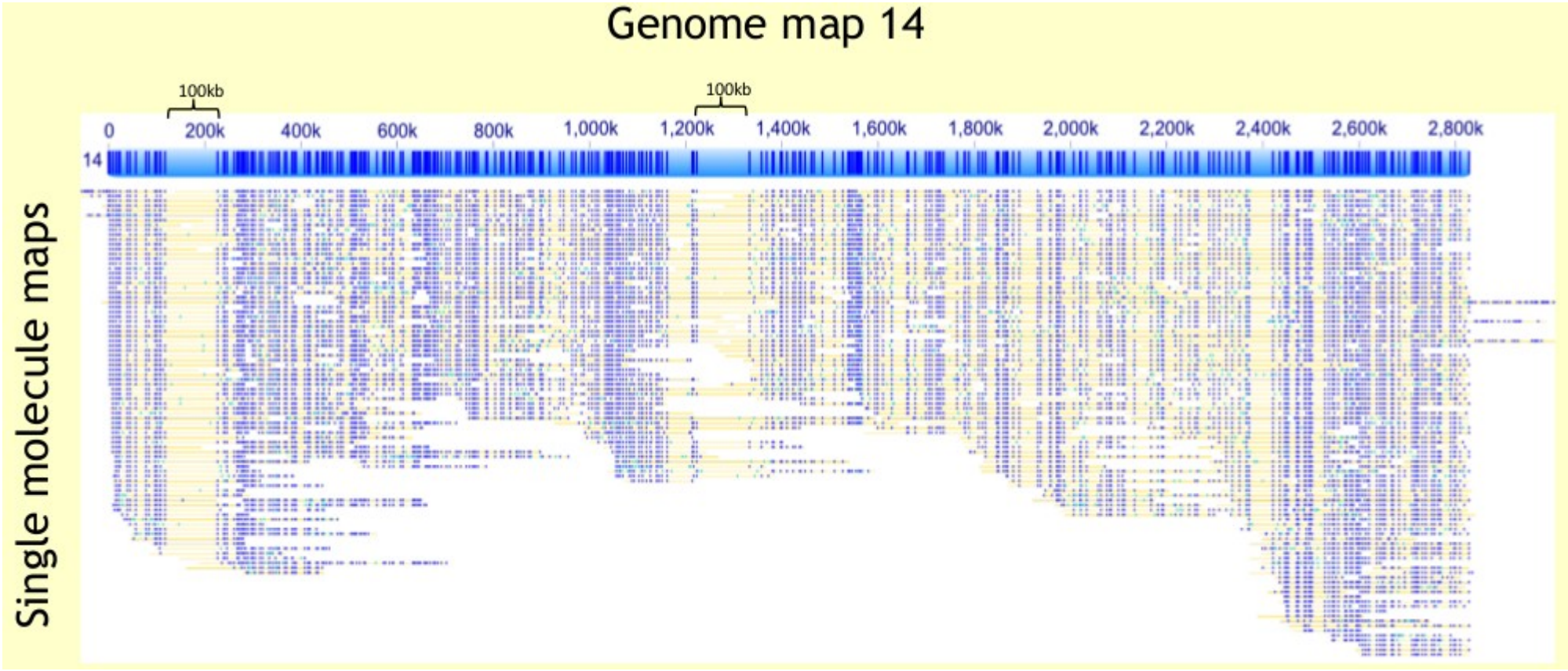
Bionano



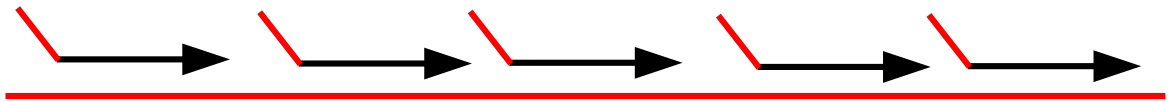


http://olomouc.ueb.cas.cz/

Bionano


Genome map 14
100kb
100kb
0
200k
400k
600k
800k
1,000k
1,200k
1,400k
1,600k
1,800k
2,000k
2,200k
2,400k
2,600k
2,800k
14
Single molecule maps

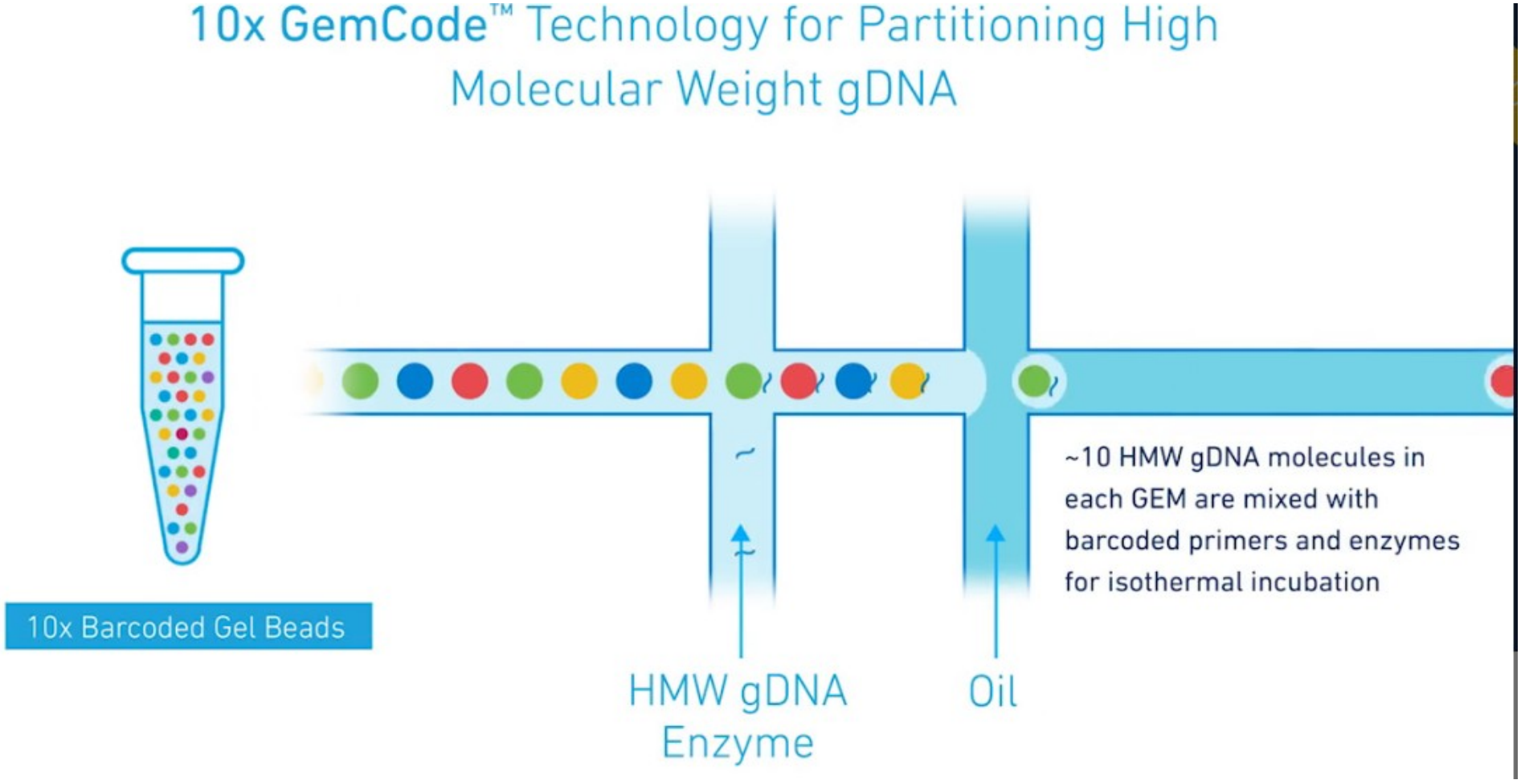
10X linked reads



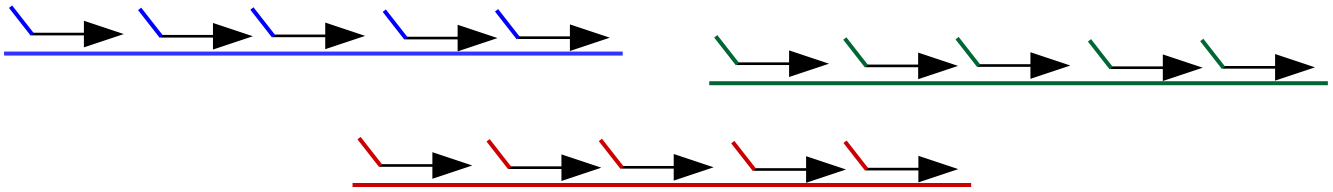
Reads generated from the same long DNA sequence are tagged with the same index and sequenced using Illumina

Reads with the same index convey medium range information (Up to 100Kb)

10X GemCode


10x GemCode™ Technology for Partitioning High Molecular Weight gDNA
10x Barcoded Gel Beads
HMW gDNA
Enzyme
Oil
~10 HMW gDNA molecules in each GEM are mixed with barcoded primers and enzymes for isothermal incubation

10X linked reads assembly



10x possible uses

De novo genome assembly

Haplotype phasing

Structural Variants:

- Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases (https://doi.org/10.1186/s13073-017-0447-8)
- Integrative analysis of genomic alterations in triple-negative breast cancer in association with homologous recombination deficiency. (DOI: 10.1371/journal.pgen.1006853)

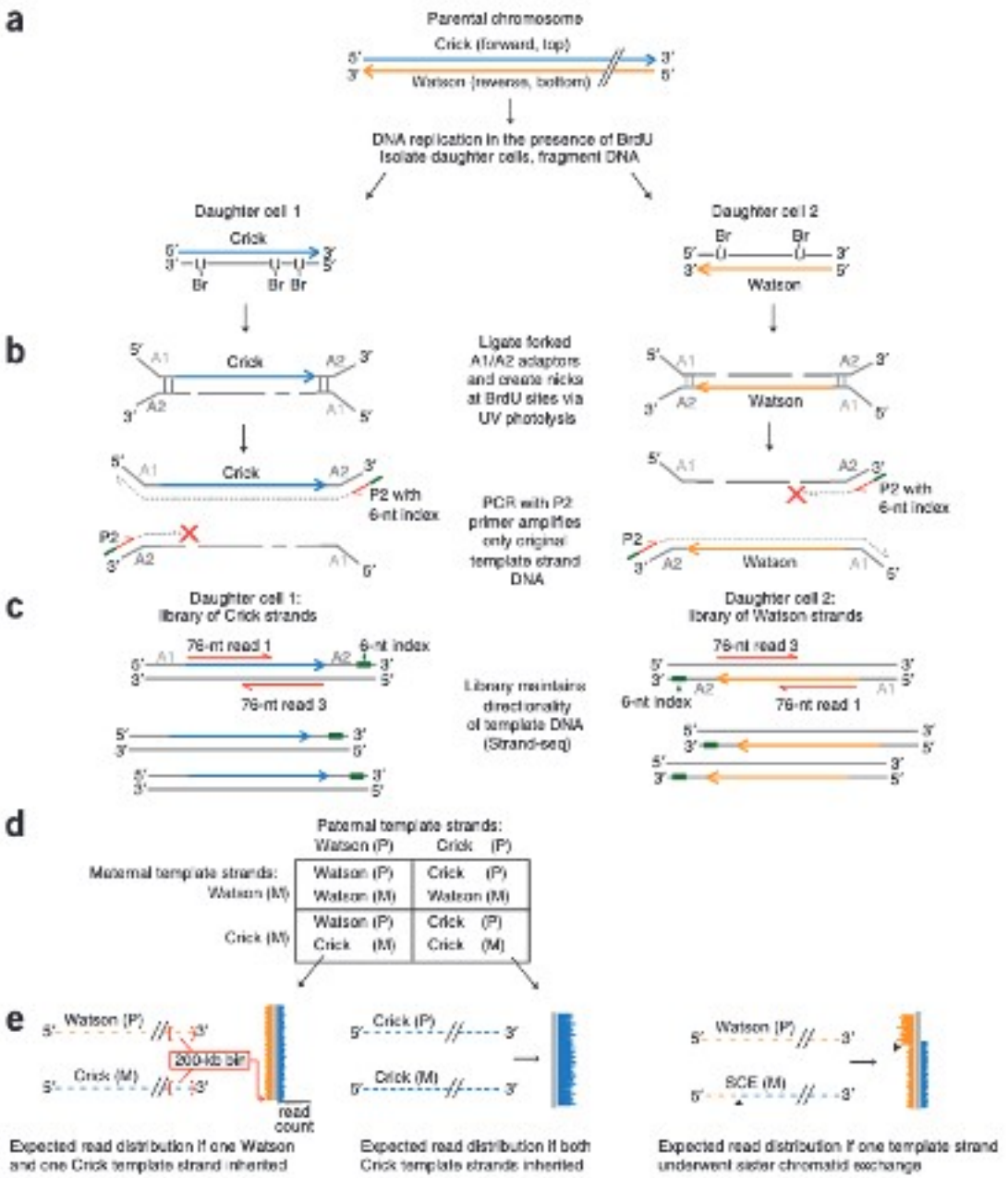
Single seq strand

short-read, single-cell sequencing method

tymidine analog (BrdU) to selectively label and remove one of the DNA strands

References:

- Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads
- Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs



Tomato assembly example

	PacBio	10X	PacBio-Bionano	10X-Bionano
N75	1.789.934	813.655	16.105.568	6.758.801
L75	141	288	18	40
N’s per 100 Kb	0	6071	2325	14885
Cost	$$$	$	$$$$	$$

Pacbio HiFi vs Nanopore ultralong

Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore

- https://doi.org/10.1093/gigascience/giaa123

Comparison with rice genome

ONT ultralong:

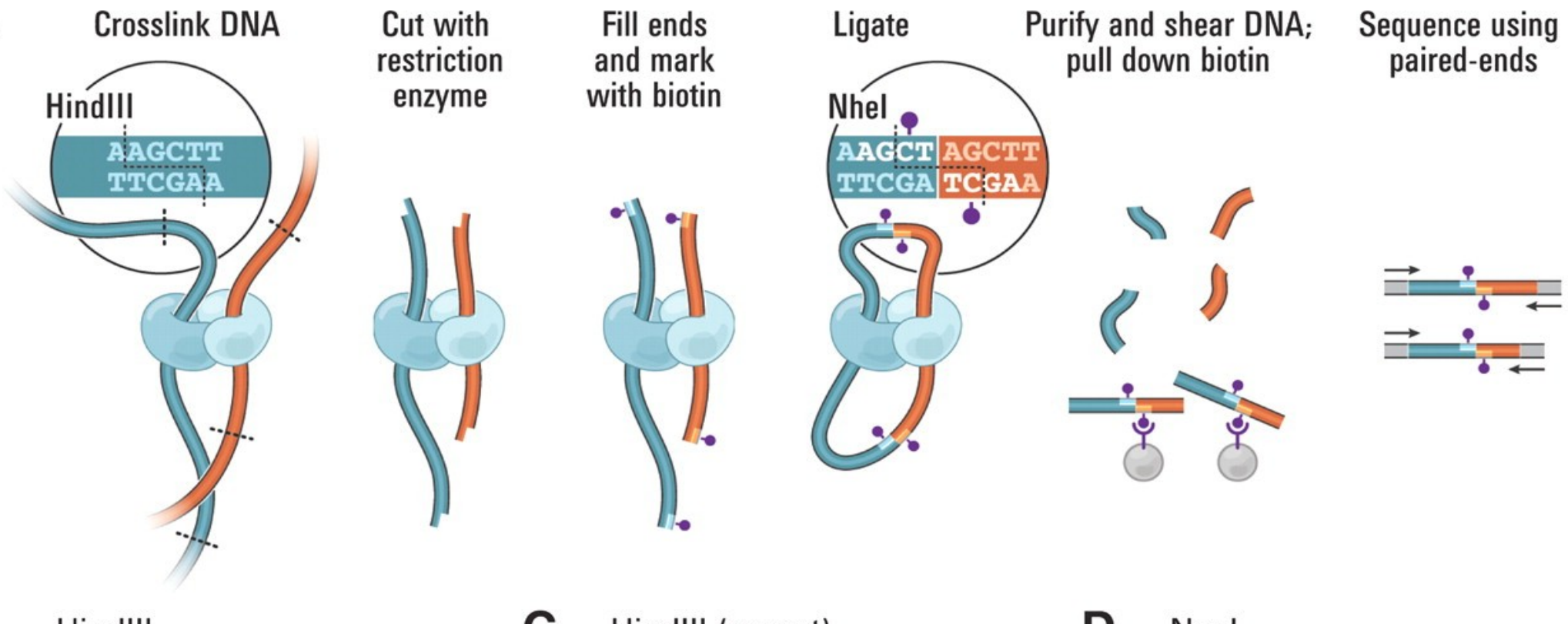
- 92 Gb data (230×) with an N50 of 41,473 bp
- higher contiguity
  - 18 contigs of which 10 were assembled into a single chromosome compared to 394 contigs and 3 chromosome-level contigs for the PacBio assembly
  - prevented assembly errors caused by long repetitive regions. PacBio assembly: over- or underestimation of the gene families

Pacbio HiFi:

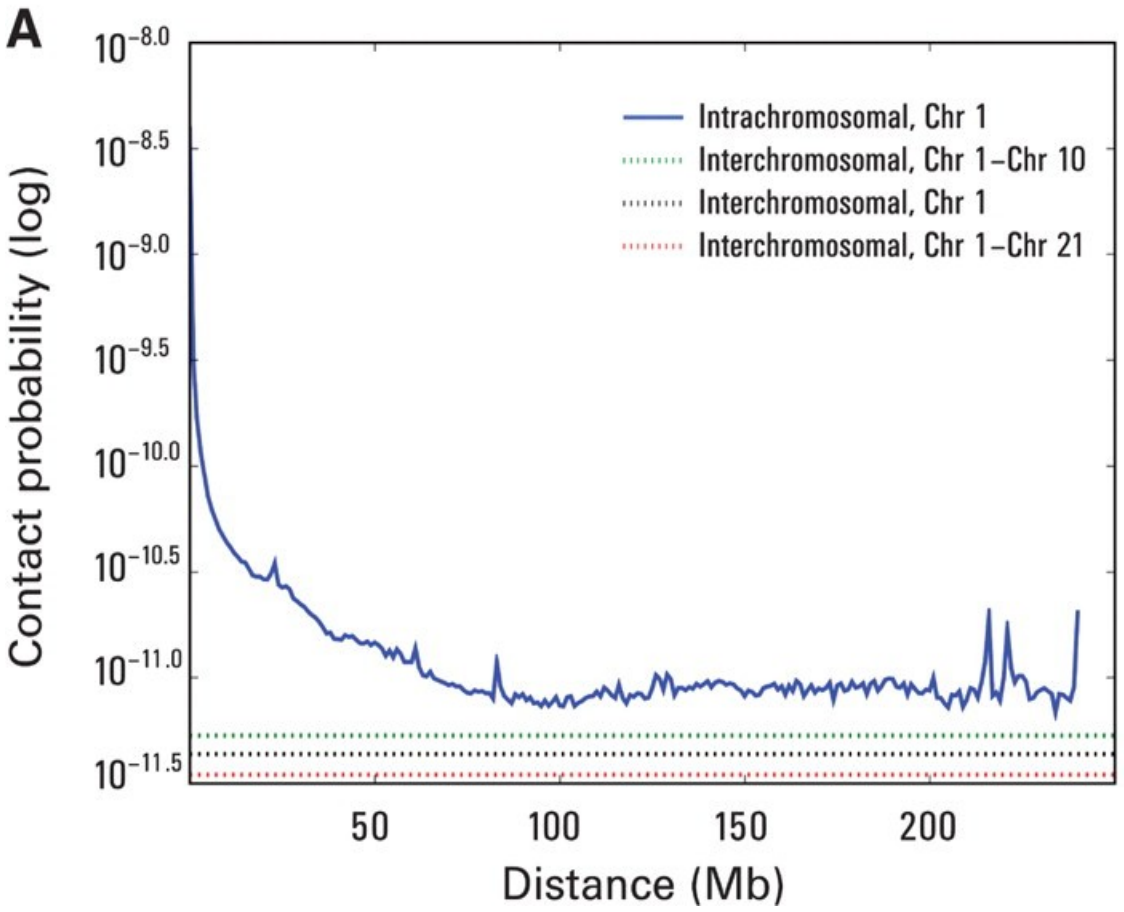
- 20 Gb HiFi reads (50×) average length 13,363 bp
- fewer errors at the level of single nucleotides and small insertions and deletions

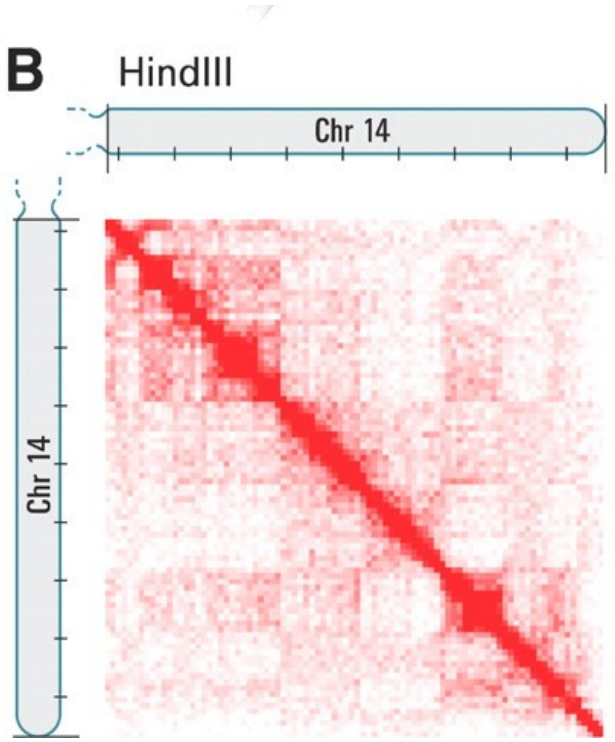
Hi-C

Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. (DOI: 10.1126/science.1181369)



Hi-C


A
Contact probability (log)
$10^{-8.0}$
$10^{-8.5}$
$10^{-9.0}$
$10^{-9.5}$
$10^{-10.0}$
$10^{-10.5}$
$10^{-11.0}$
$10^{-11.5}$
Intrachromosomal, Chr 1
Interchromosomal, Chr 1–Chr 10
Interchromosomal, Chr 1
Interchromosomal, Chr 1–Chr 21
50
100
150
200
Distance (Mb)


B
HindIII
Chr 14
Chr 14

Hi-C Assembly

De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. (DOI: 10.1126/science.aal3327)

Hi-C data provide links across a variety of length scales

Unlike mate-pair and pair-end reads Hi-C contact spans an unknown length and may connect loci on different chromosomes

Human genome assembly:

- Pair-end Illumina reads (67X coverage)
- Hi-C data (6.7 coverage)
- 23 scaffold that span the 99.5% of the 23 human chromosomes
- Errors remain in the ordering of short distances than could be fixed by mate-pairs, long reads or bionano

Haplotype assembly with hifiasm

Additional short read data:

- Hi-C pair ends or Strand-seq
  - improves contiguity and phasing accuracy
- Parental trios
  - Improves phasing accuracy
  - Advisable specially for high heterozygosity

Hifiasm does not perform scaffolding for now

Checked with human and California redwood with a ~30-Gb hexaploid genome

Similar to HiCanu

Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, 18:170-175. https://doi.org/10.1038/s41592-020-01056-5

Jose Blanca
COMAV institute
bioinf.comav.upv.es