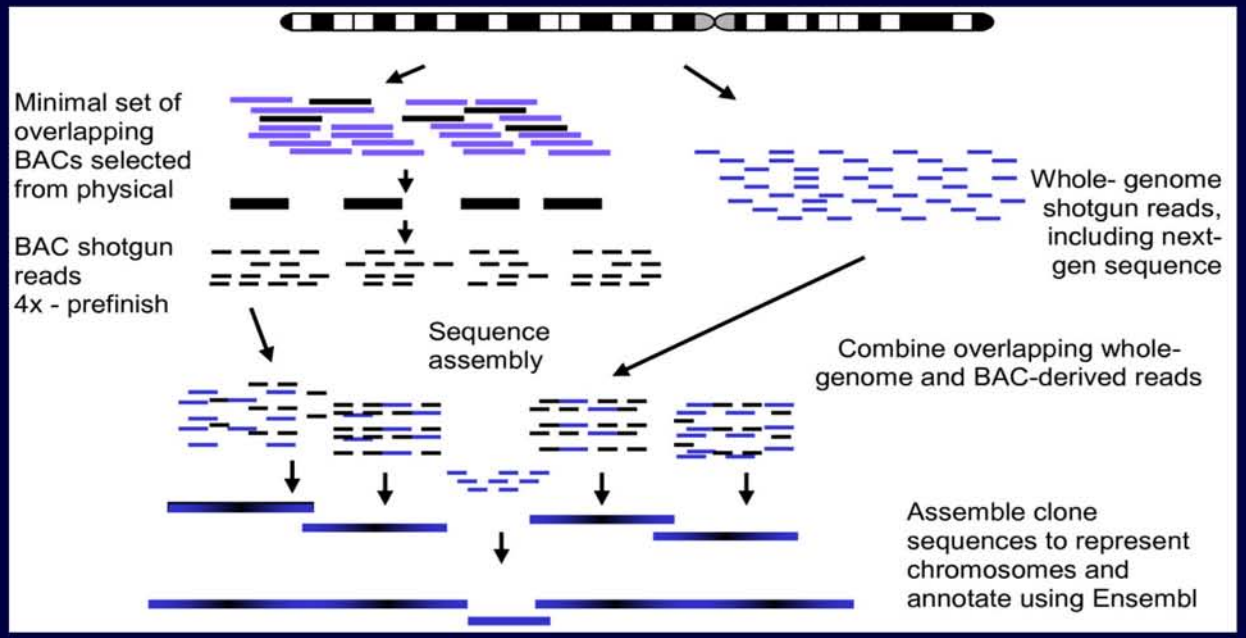
Next-Generation Sequencing: techniques and data processing

Domenico Simone

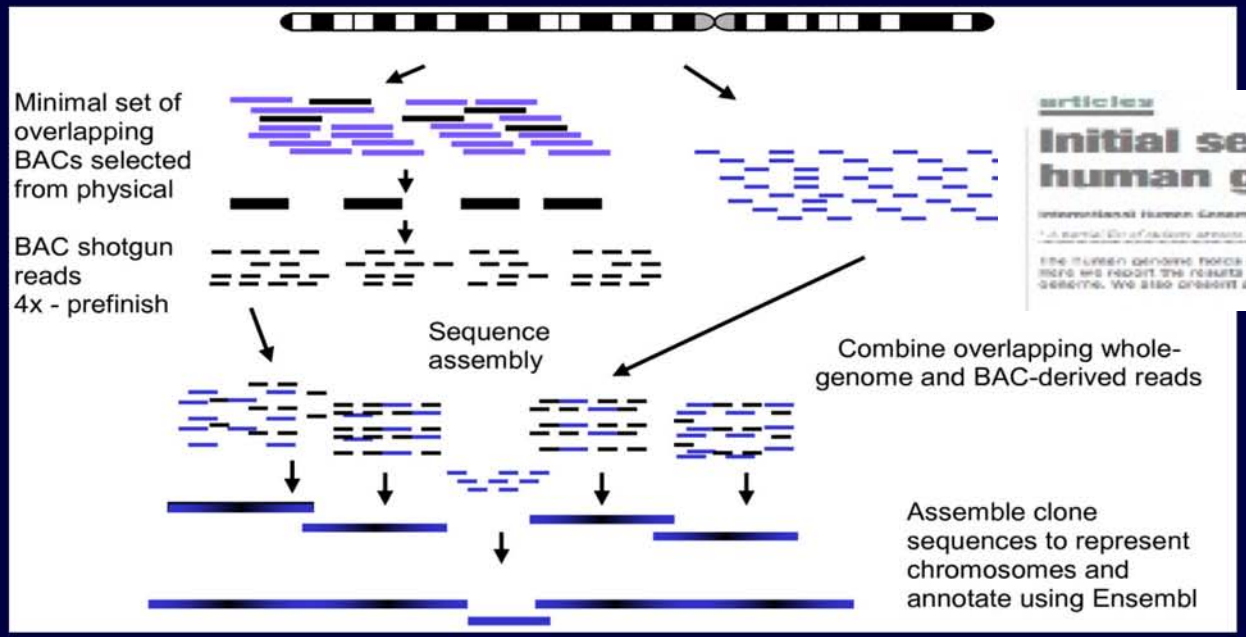
Genome sequencing and assembly

- The *traditional* method: Sanger
  DNA extraction (high amount)
  DNA cloning (hierarchical / shotgun)
  Sequencing (chain-termination)
  Assembly (*contigs*)
  Finishing, gap closing (*scaffolds* / *supercontigs*)



Genome sequencing and assembly

- The *traditional* method: Sanger
  DNA extraction (high amount)
  DNA cloning (hierarchical / shotgun)
  Sequencing (chain-termination)
  Assembly (*contigs*)
  Finishing, gap closing (*scaffolds* / *supercontigs*)




articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

Haemophilus influenzae (1995)
Drosophila melanogaster (2000)
Human Genome Project (1990-2001)

Next-Generation Sequencing

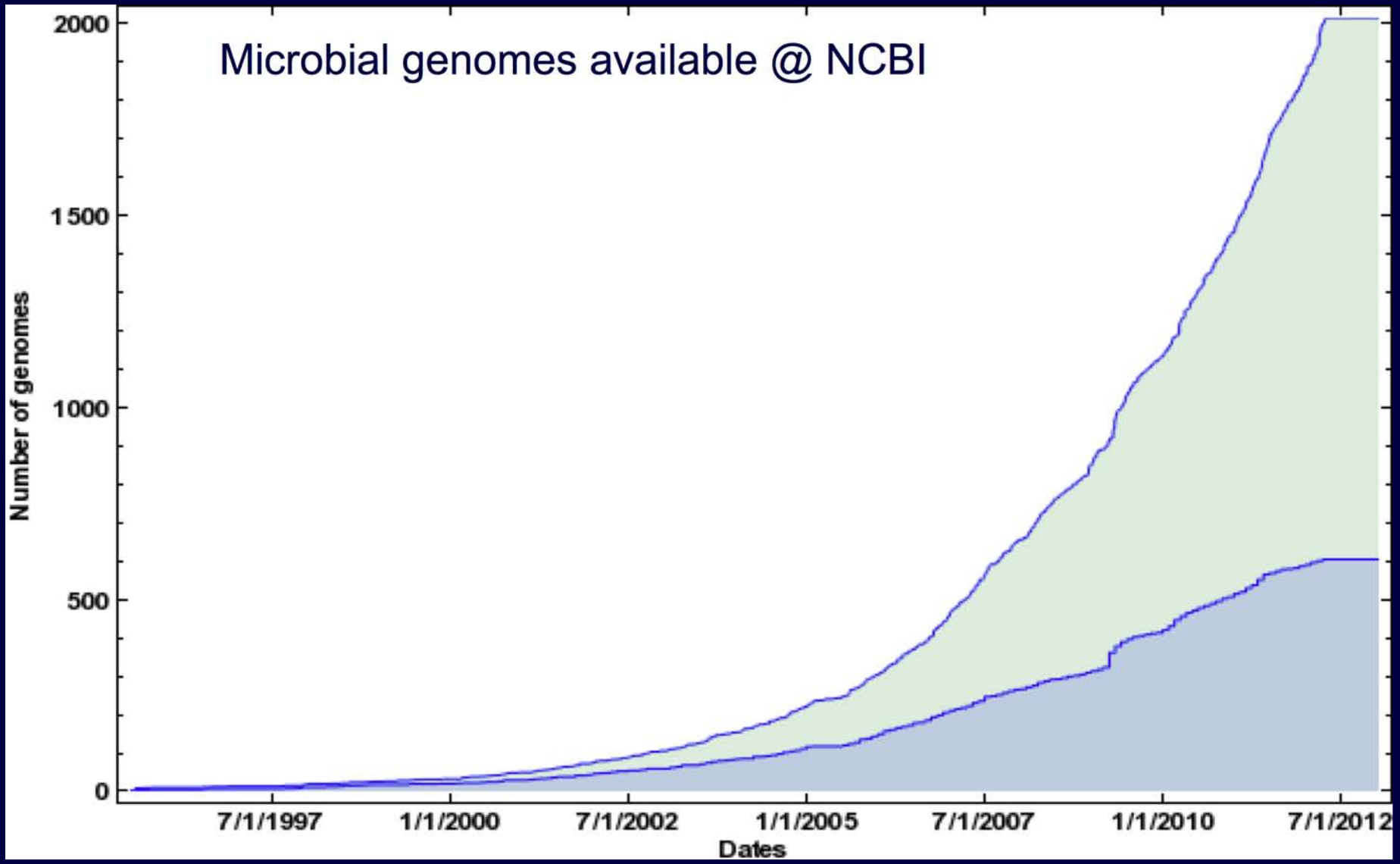
Based on several chemistries, but a common feature:
massive parallelism

Sanger capillary sequencing: 96 sequences / run

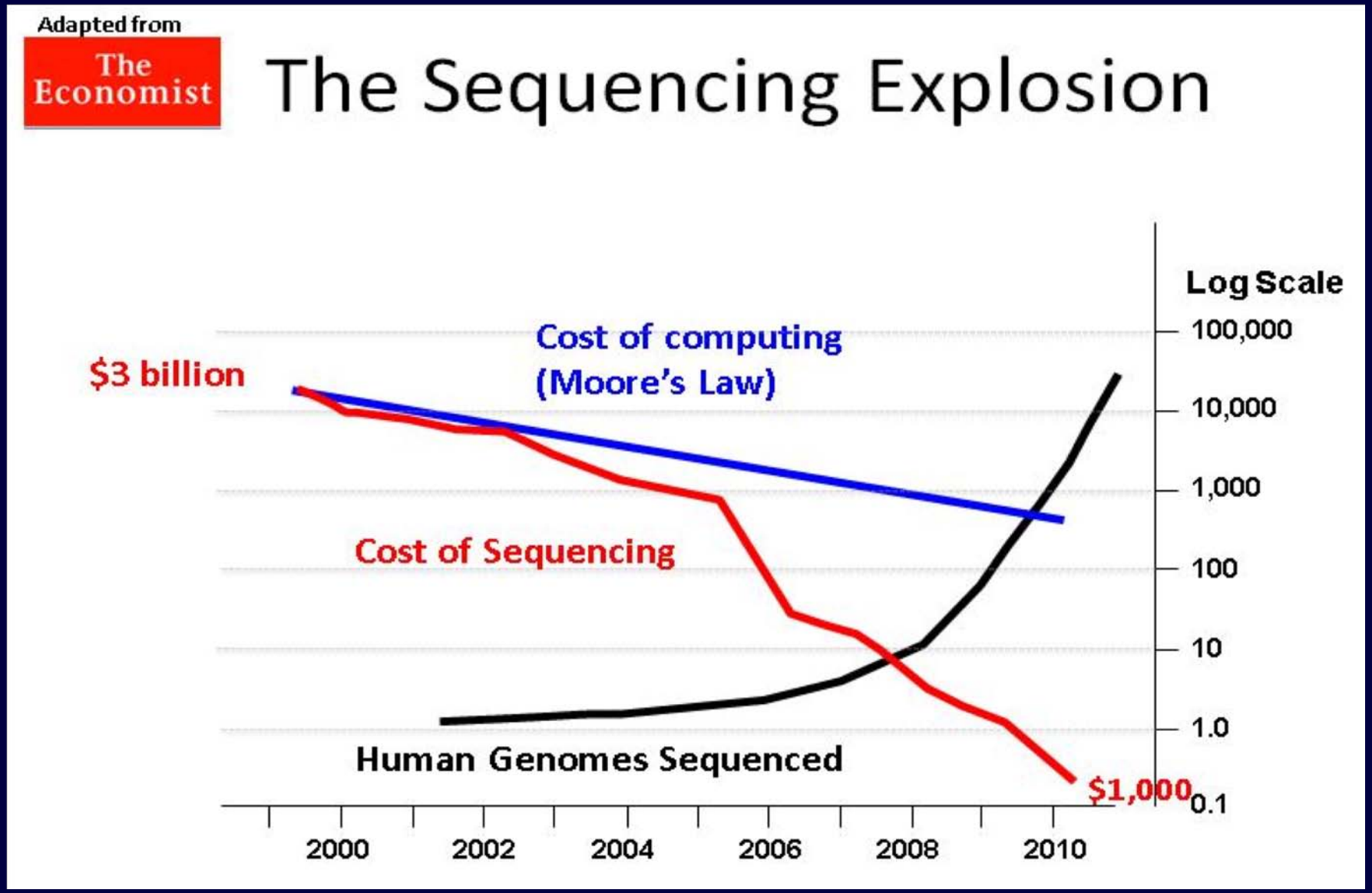
NGS: millions / run!!!

- Most used NGS methods:
  454 pyrosequencing
  Illumina sequencing-by-synthesis
  SOLiD
- Exploited by 2000s
  … not so "next"!!!

NGS: lower costs, higher throughput



NGS: lower costs, higher throughput



NGS: other features

- Lower sampling bias
  Sanger: strictly connected to cloning
  NGS doesn't require cloning in vectors!!!

Cloning bias
Tendency of certain regions of the genome to be cloned less often than others during library preparation, and thus less likely to be sequenced

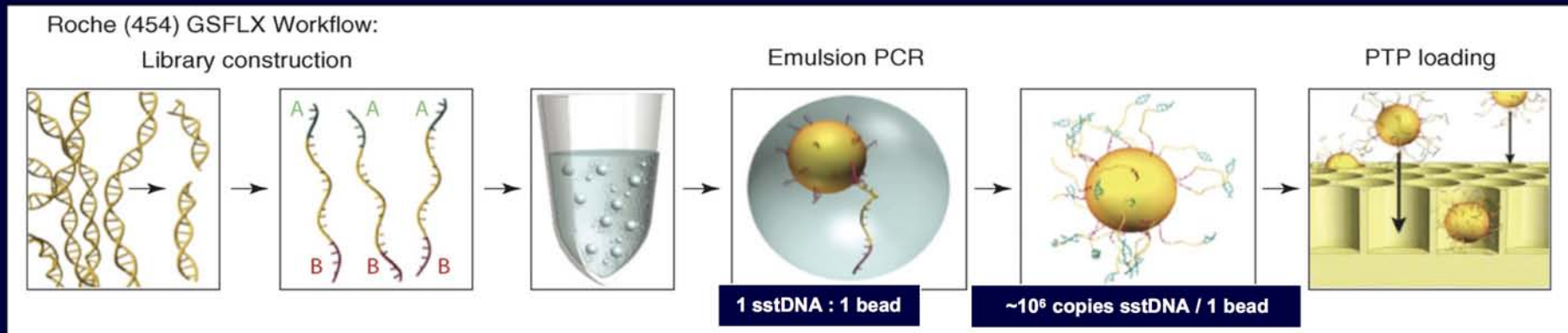
- Smaller reads but higher coverage depth

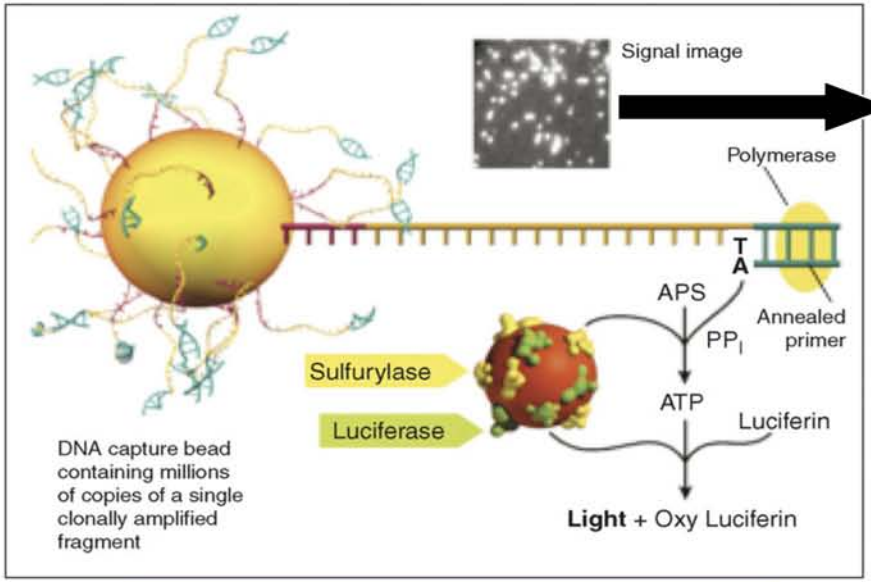
NGS: common applications

- Whole genome sequencing
  de novo genome assembly
  variant detection (mutations, SNPs, indels, copy number)
- Targeted resequencing
  exomes
- ChIP-seq (Chromatine Immunoprecipitation sequencing)
  protein-DNA binding
- Expression profiling
- sRNA sequencing

454 (Life Sciences, Roche)

Pyrosequencing
nucleotide incorporation leads to reactions resulting in releasing light (luciferase)





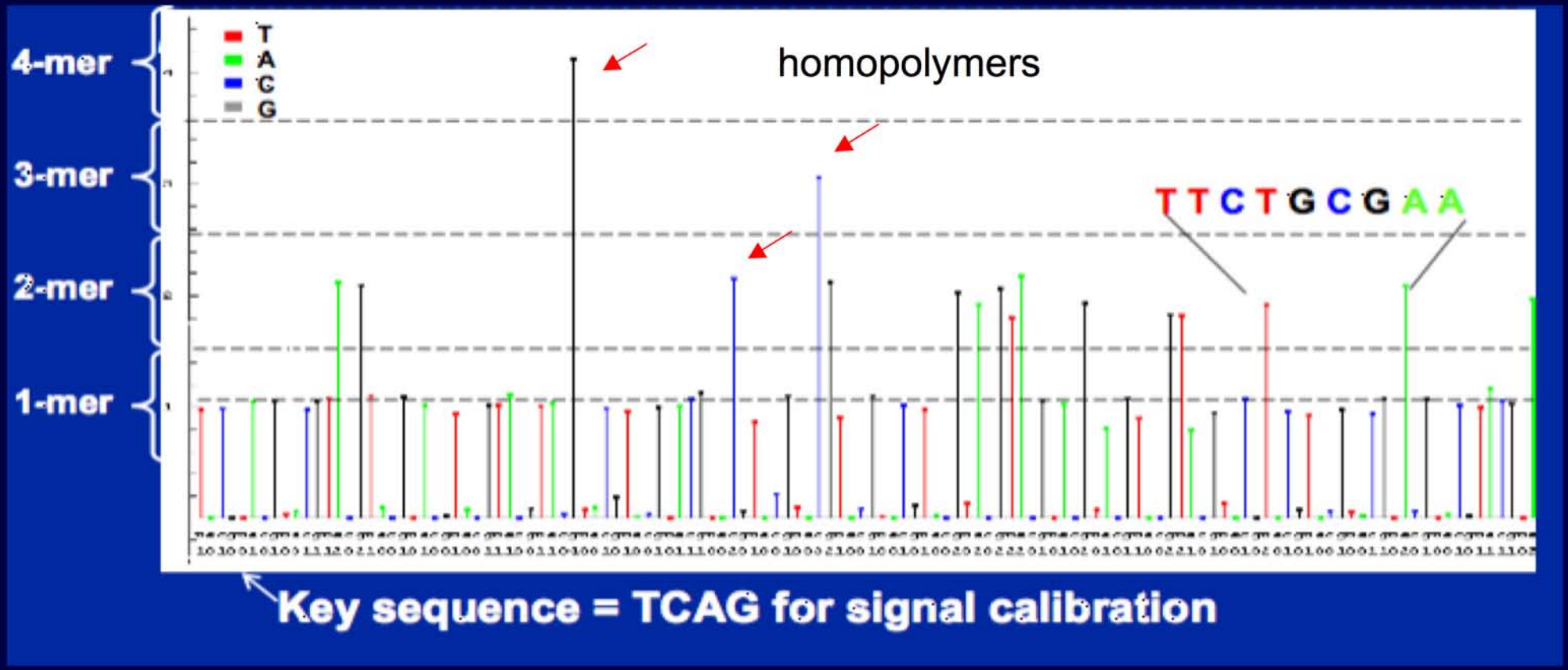
Light signal is proportional to the number of dNTPs incorporated
Loss of linearity for long homopolymeric stretches

Sequencing run data are provided as **flowgrams** (raw data of light intensity across cycles) in *.sff files (standard flowgram format)

TRENDS in Genetics

Pyrosequencing reaction
Cycles of dNTP addition

454 flowgram



Solexa/Illumina sequencing

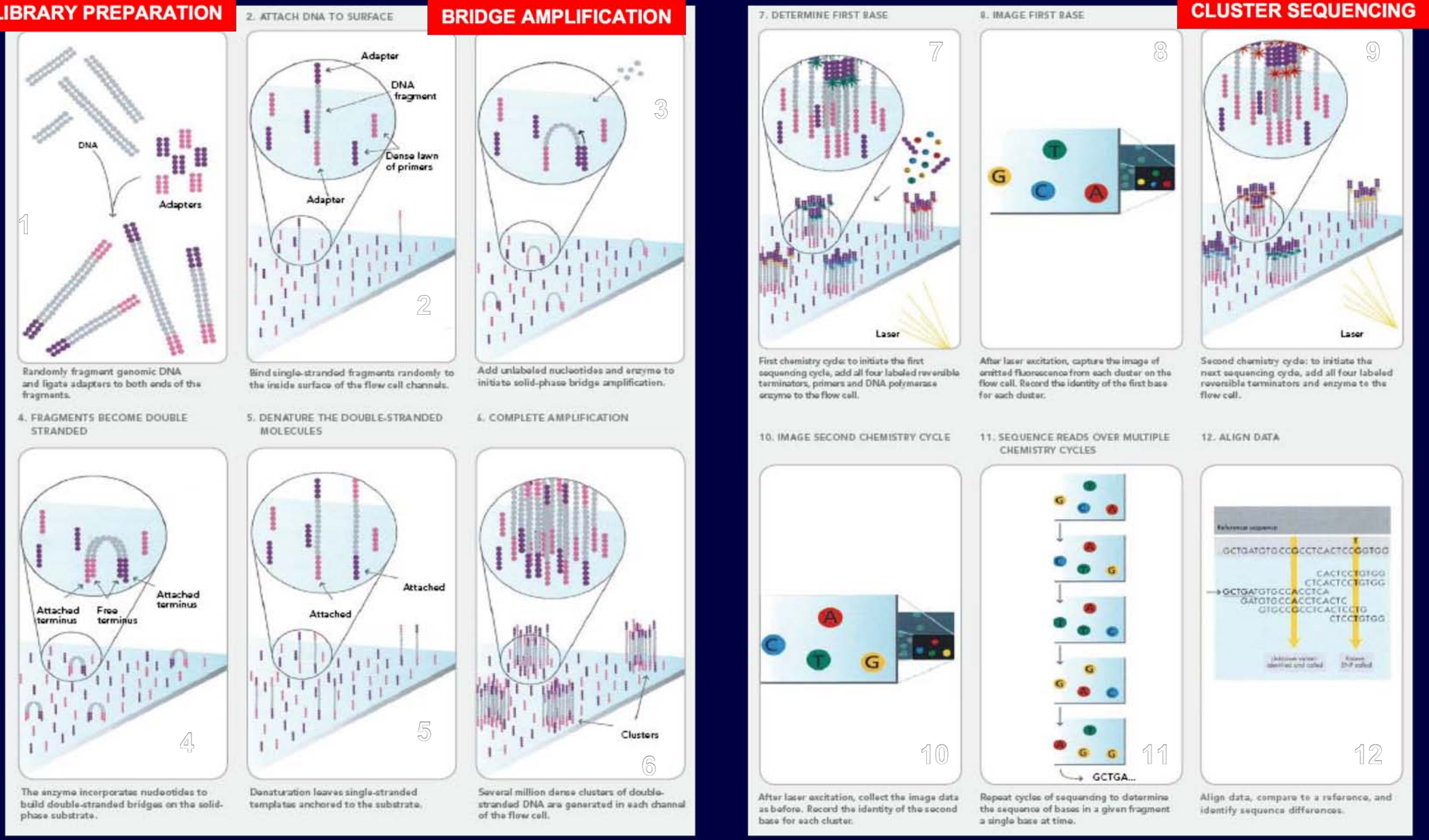
1. Library preparation
2. Clusters generation (**bridge amplification**)
3. Clusters sequencing (**sequencing by synthesis**)
4. Data analysis

Solexa/Illumina sequencing

LIBRARY PREPARATION

BRIDGE AMPLIFICATION

CLUSTER SEQUENCING



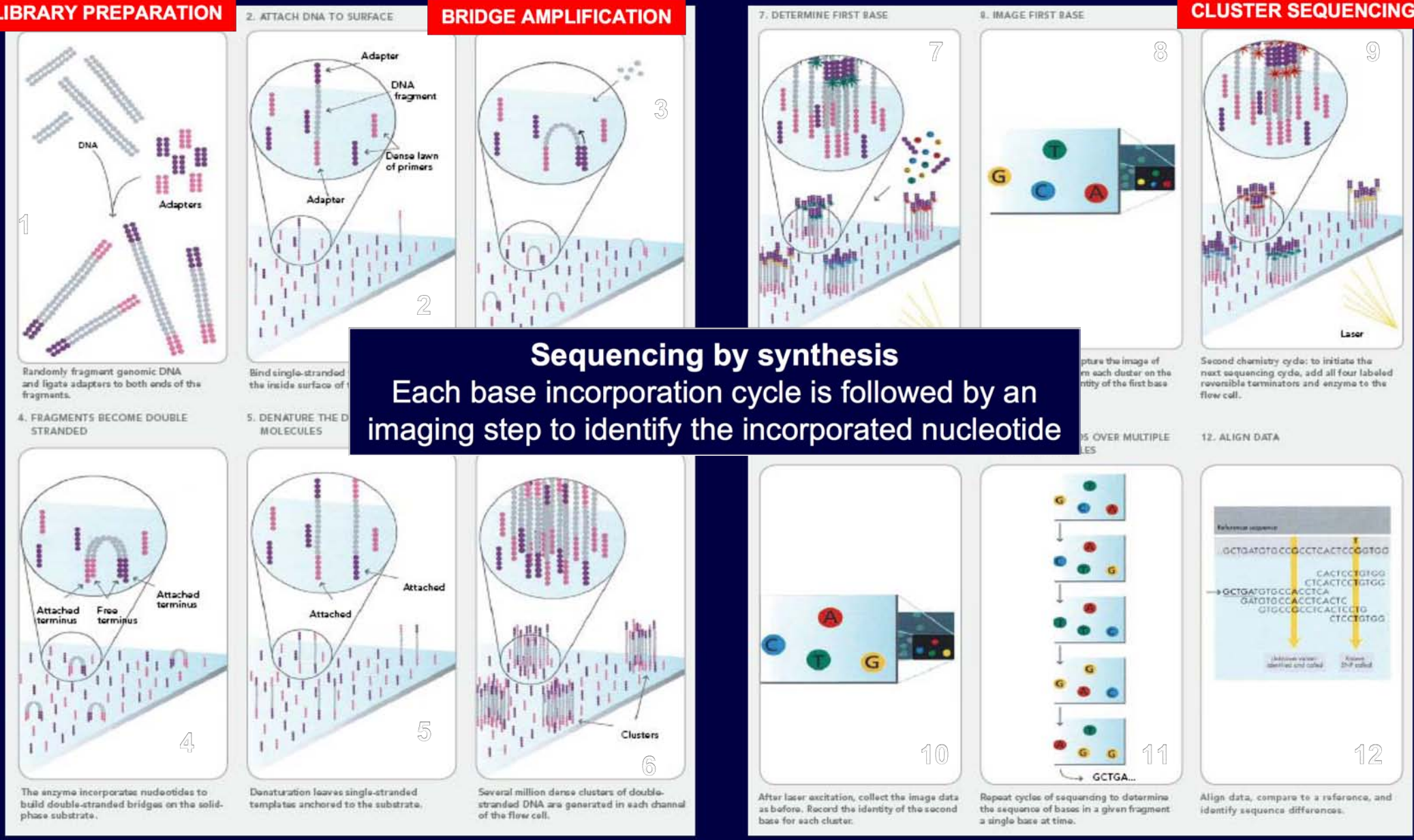
http://www.youtube.com/watch?v=45vNetkGspo (3'10'')

Solexa/Illumina sequencing

LIBRARY PREPARATION

BRIDGE AMPLIFICATION

CLUSTER SEQUENCING



Sequencing by synthesis

Each base incorporation cycle is followed by an imaging step to identify the incorporated nucleotide

http://www.youtube.com/watch?v=45vNetkGspo (3'10'')

Sequencing techniques:
a brief comparison

Method	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Chain termination (Sanger sequencing)
Read length	500bp (mean)	50 to 150bp	400 to 900bp
Reads per run	1 million	up to 3 billion	96
Time per run	24 hours	1 to 10 days	20 minutes to 3 hours
Cost per 1 million bases (in US$)	$10	$0.05 to $0.15	$2400

NGS data: quality check / pre-processing

NGS technologies: higher throughput but more prone to sequencing errors

…quality check!!!

Remove low quality regions and adaptor sequences (**clipping**)

…pre-processing

Many assemblers have built-in utilities for these aims

Quality score

- Basically, is the probability that a base is called incorrectly
- Most used equation for expressing this: **Phred quality score** (Sanger variant)

$$Q_{sanger} = -10 \log_{10} p$$

FYI: Solexa/Illumina had a similar codification but it has been dropped

Example…

p = 1/1000 "*There is 1 probability in 1000 that the base is called incorrectly*"

↓

Q = 30

The range of scores is usually upper-delimited to **40**

Quality score

- The QS info is integrated with the sequence in the **FASTQ** (fasta quality) format

```
@SRR001666.1
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Sequence id

Sequence

separator

QS

QS line: same length as sequence, each character encodes the quality of the base (ASCII coded)

ASCII coding for base qualities
(FASTQ format)

Quality score	Error probability	ASCII characters
0..9	1	!"#$%&'()*
10..19	0.1	+,-./01234
20..29	0.01	56789:;<=>
30..39	0.001	?@ABCDEFGH
40	0.0001	I

FastQC

a quality control tool for NGS data

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

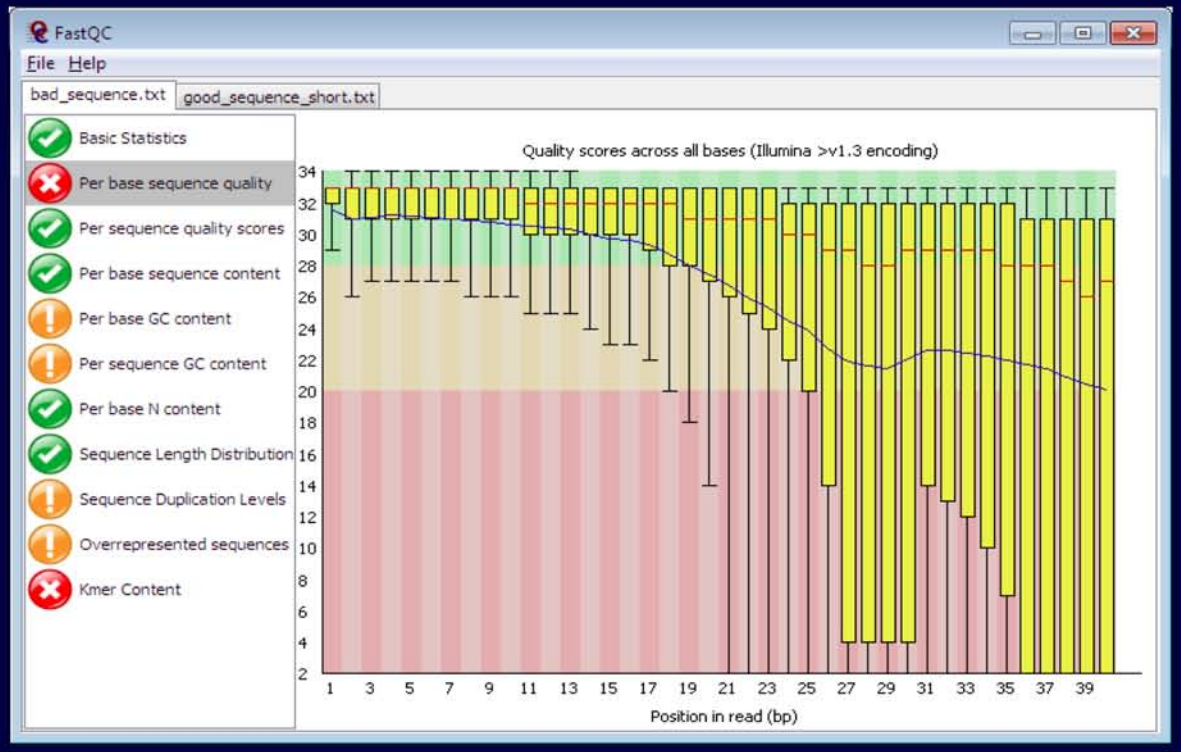
- Modular set of analyses to get a quick impression about data… potential issues (quality score problems, bias etc)
- “Are you satisfied with your raw data and are they worth to be processed?”
- Each module gives evaluations about dataset features in a pass/warning/fail fashion

PASS

WARNING

FAIL

NOTE:
a “normal” sample for FastQC is random and diverse (eg a WGS)



FastQC
module examples

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequence length	40
%GC	45

No test effectively, but …

FastQC
module examples

Basic Statistics

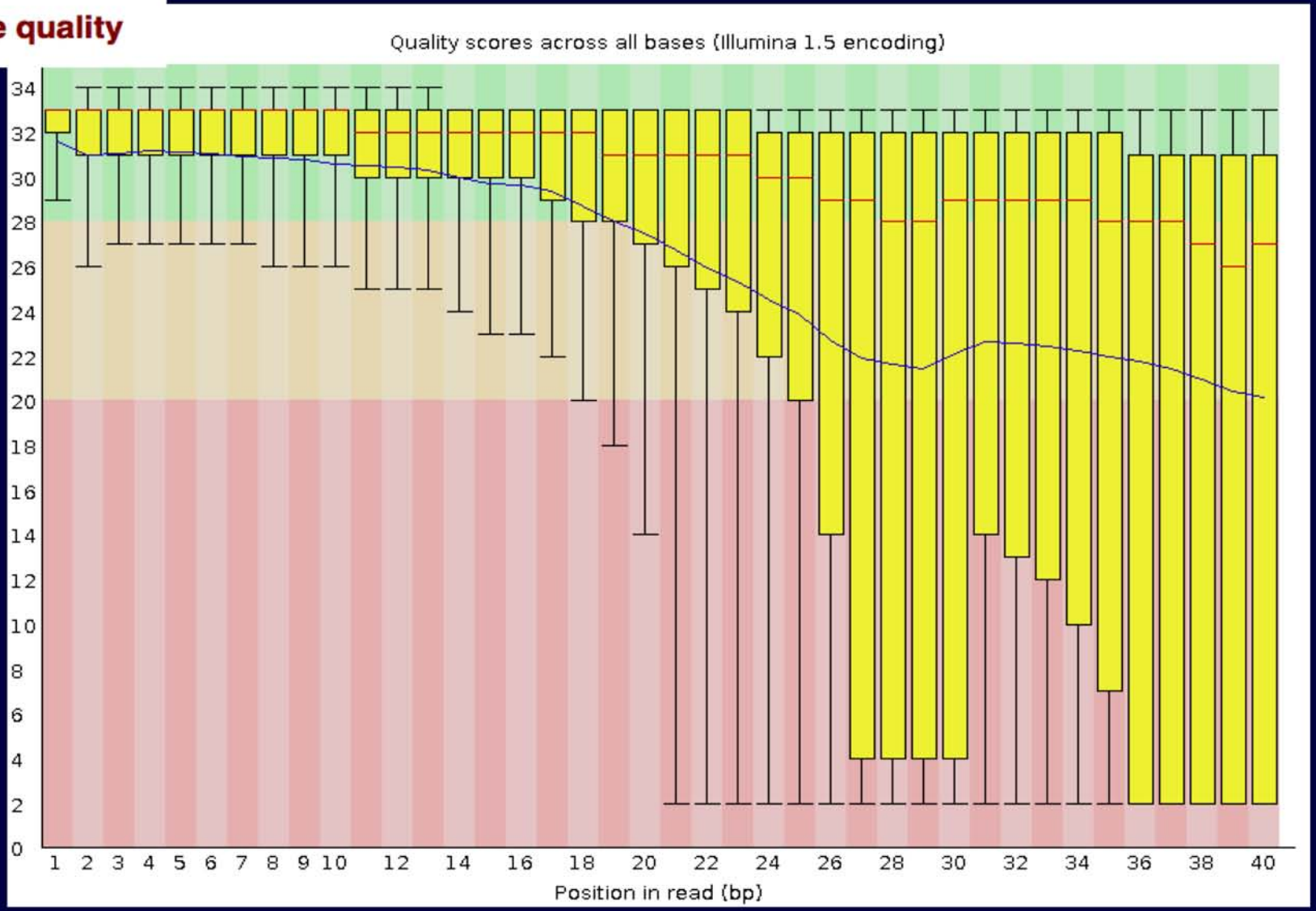
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequence length	40
%GC	45

Is this what you expected?

No test effectively, but …

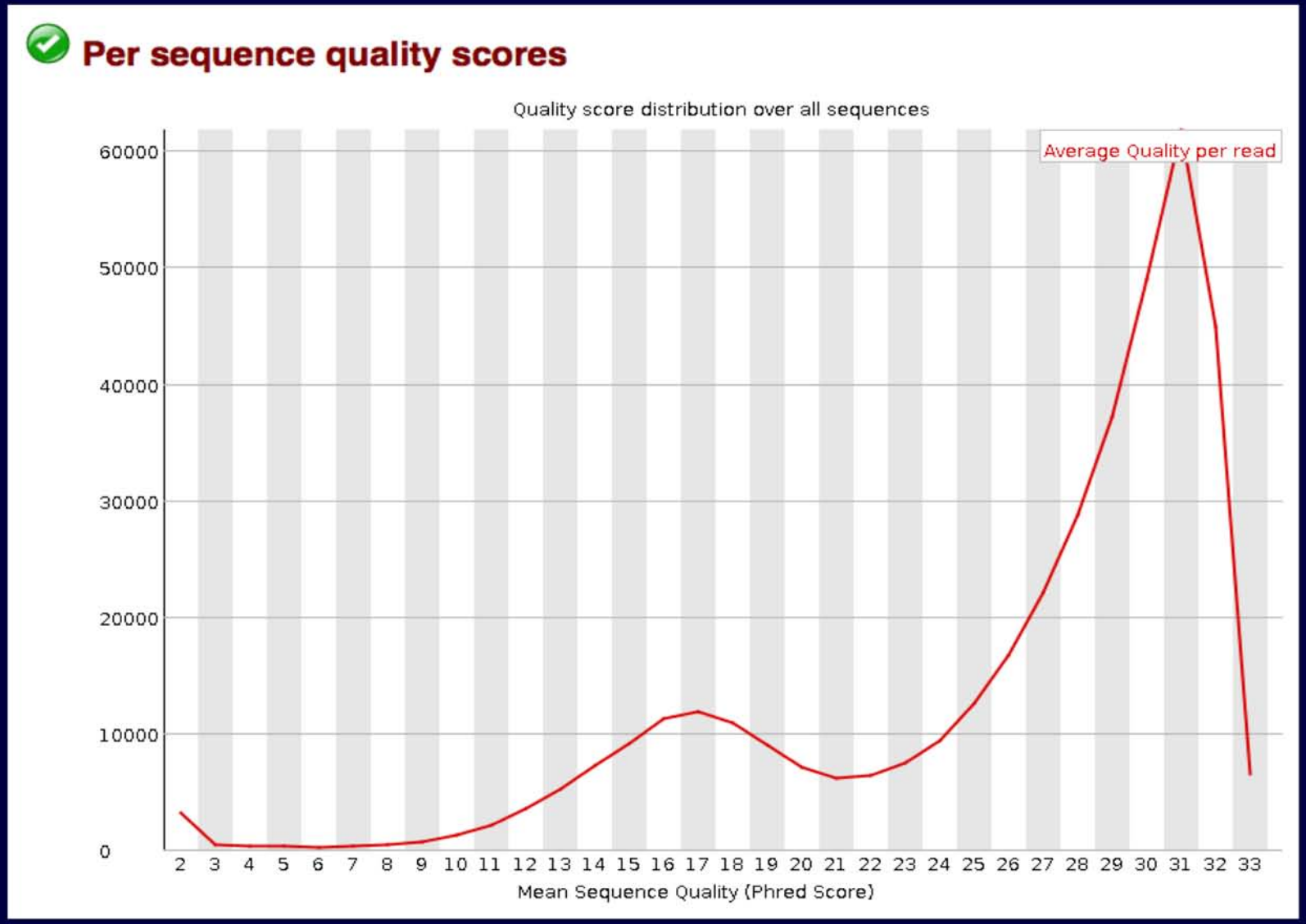
FastQC
module examples

Per base sequence quality

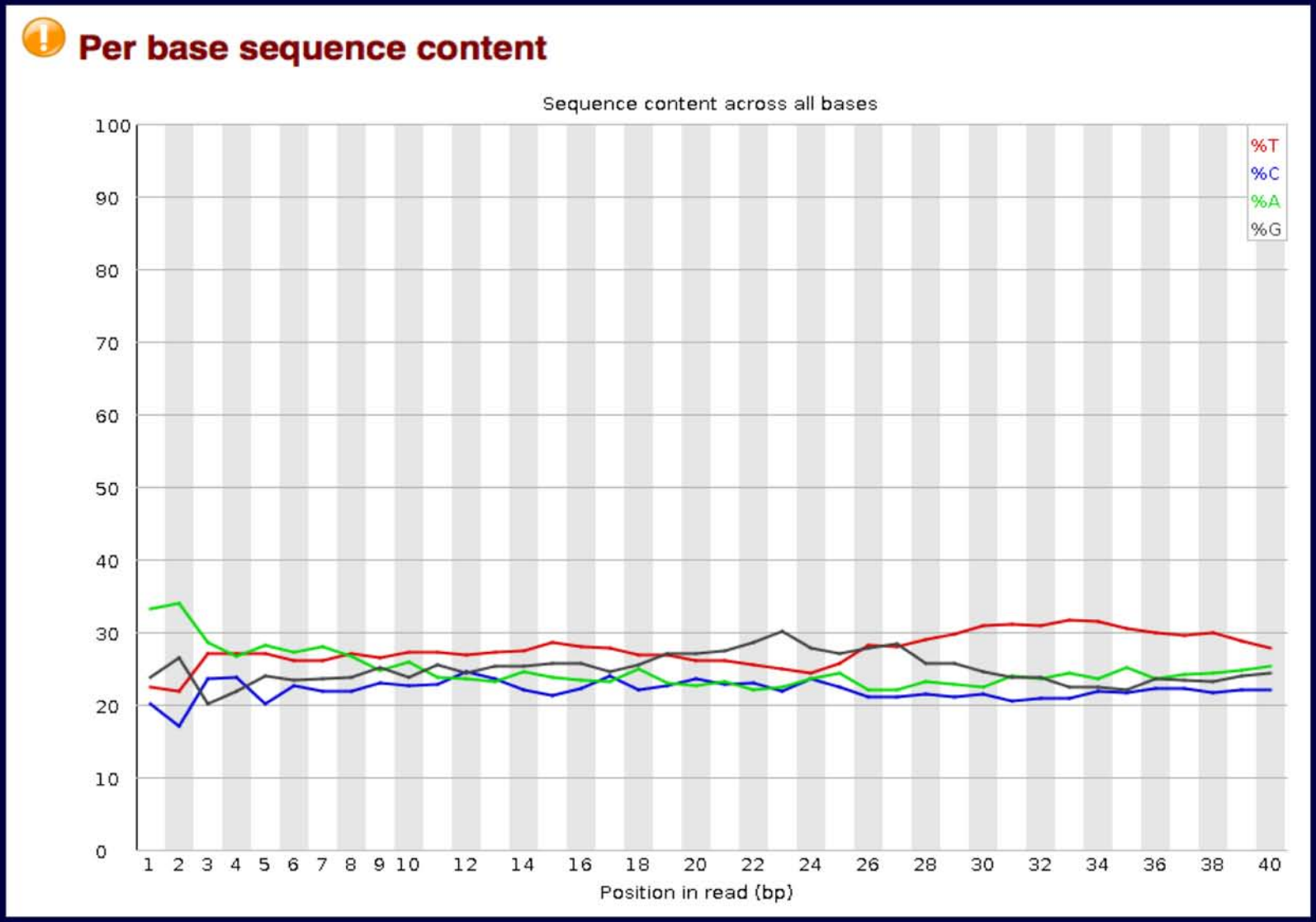


End clipping is needed

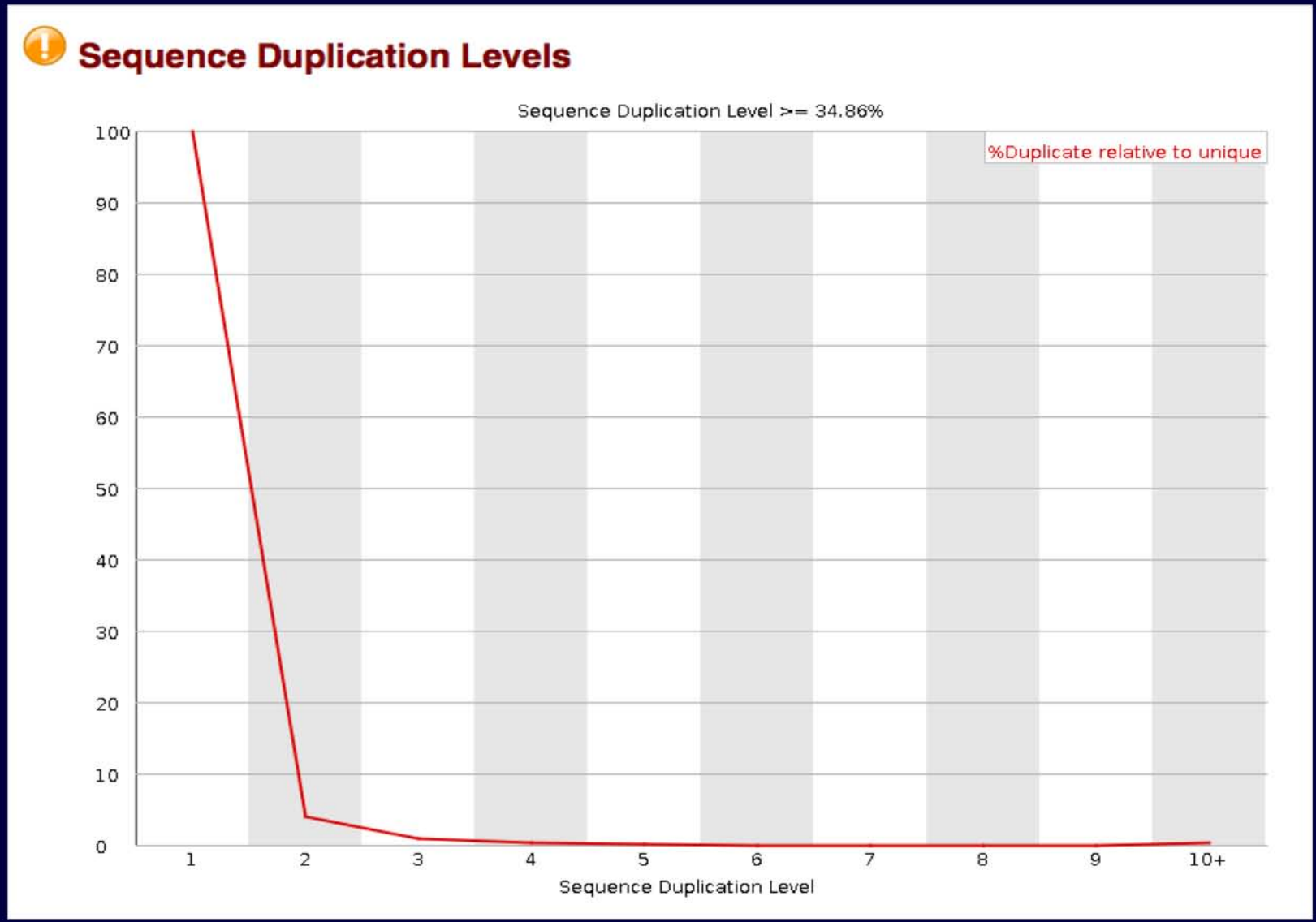
FastQC
module examples



FastQC
module examples



FastQC
module examples



Reads: alignment / assembly

- If the sequenced sample has a **reference genome**, reads are **aligned** on it (*e.g.* for variant finding)
- Otherwise, reads are **aligned with each other**, *i.e.* they are **assembled** (*e.g. de novo* sequencing projects)

Short read alignment

- Given a reference and a set of reads, report at least one "good" local alignment for each read if one exists
  (Approximate answer to question: where in genome did read originate?)

- "Good" .. ?
  "Fewer mismatches are better"
  "Failing to align a low-quality base is better than failing to align a high-quality base"

Short read alignment

- Given a reference and a set of reads, report at least one "good" local alignment for each read if one exists
  (Approximate answer to question: where in genome did read originate?)

- "Good" .. ?
  "Fewer mismatches are better"
  "Failing to align a low-quality base is better than failing to align a high-quality base"

```
TGATATTA
||||   ||
TGATcaTA
```

reference

read

```
TGGCCATA
||   ||||
TGATcaTA
```

[lowercase: low quality]

Short read alignment

- Given a reference and a set of reads, report at least one "good" local alignment for each read if one exists
  (Approximate answer to question: where in genome did read originate?)

- "Good" .. ?
  "Fewer mismatches are better"
  "Failing to align a low-quality base is better than failing to align a high-quality base"

```
TGATATTA
||||   ||
TGATcaTA
```

reference

read

```
TGGCCATA
||   ||||
TGATcaTA
```

[lowercase: low quality]

Softwares: BWA, Bowtie, SSAHA, …

Assembly

one of the biggest computational problems

Traditional approach…

Greedy

1 Calculate **pairwise alignments** of all fragments;
2 Choose two fragments with the **largest overlap**;
3 Merge chosen fragments;
4 repeat
5 step 2;
6 step 3;
7 until only one fragment is left ;
8 return **Set of contigs**

Clusters of reads consistently aligning with each other, forming contiguous sequences

Adopted in the earliest Genome Projects
(softwares: TIGR, Phrap, CAP3)

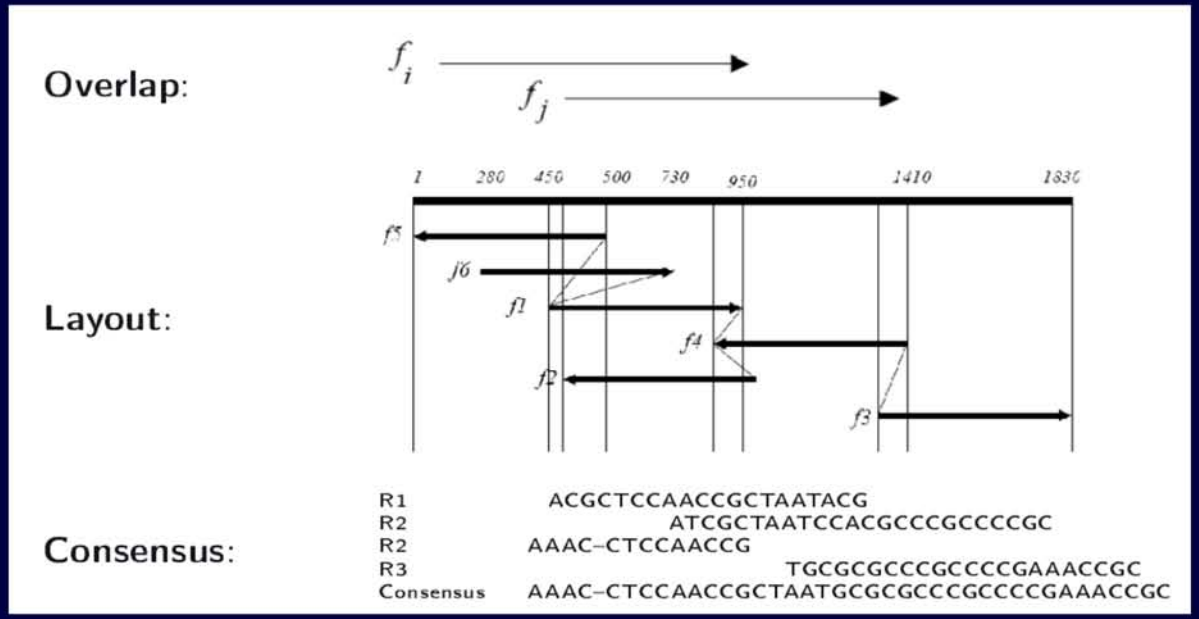
Not suitable for NGS datasets

Algorithms for NGS dataset assembly

Overlap-Layout-Consensus (OLC)

Newbler (Roche)

1. Overlap discovery (Overlap)
2. Build and use the overlap graph (Layout)
3. Multiple sequence alignment (Consensus)



Algorithms for NGS dataset assembly

De Brujin graph

ABySS

- Reads are not aligned directly; they are indexed in *k*-mers; then assembled in paths and contigs

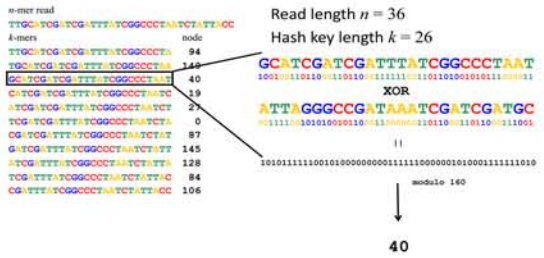
↑ k = ↑ specificity — More fragmented assemblies

↓ k = ↓ specificity — Less fragmented assemblies

k-mer choice is empirical

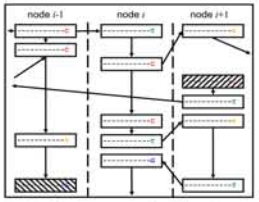
Partitioning Read Space

Distribute sub-reads and reverse-complements over nodes



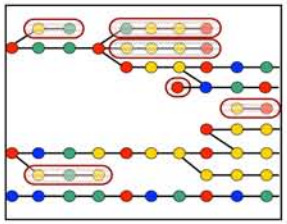
Graph Generation

- A given *k*-mer can have up to 8 extensions
- Each node announces the list of *k*-mers that it has to the nodes that hold their possible extensions
- Each node records if there are any extensions of the *k*-mers that it stores
- This forms adjacency information for *k*-mers over a distributed de Bruijn graph



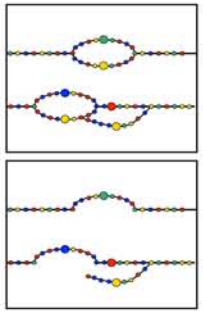
Trimming

- Data would have experimental noise
- de Bruijn graph would have false branches
- Some read errors are filtered by removing such branches
- Trimming prevents the later assembly step to come to a premature end because of read errors



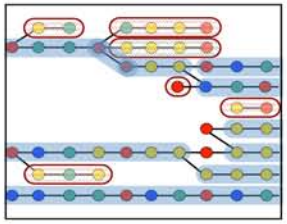
Bubble Popping

- Repeat read errors and single nucleotide allelic differences would cause "bubbles" of length $2k$-1
- Bubbles are popped by removing either of those branches
- Complex bubbles can form when multiple bubbles intersect
  - Bubble popping step either reduces the bubble orders by one
  - Or creates dead branches
- Popped bubbles are recorded in a log file to study potential allelic differences



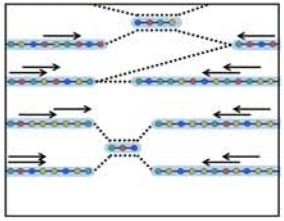
Assembly - SET

- Remaining de Bruijn graph is analyzed for contig extension ambiguities
- If there is a multiplicity in the inbound or outbound contig extensions, then contig growth is terminated
- SET assembly step then concatenates the remaining connected nodes in the di-graph, creating independent contigs that overlap by no more than *k*-1 bases



Assembly - PET

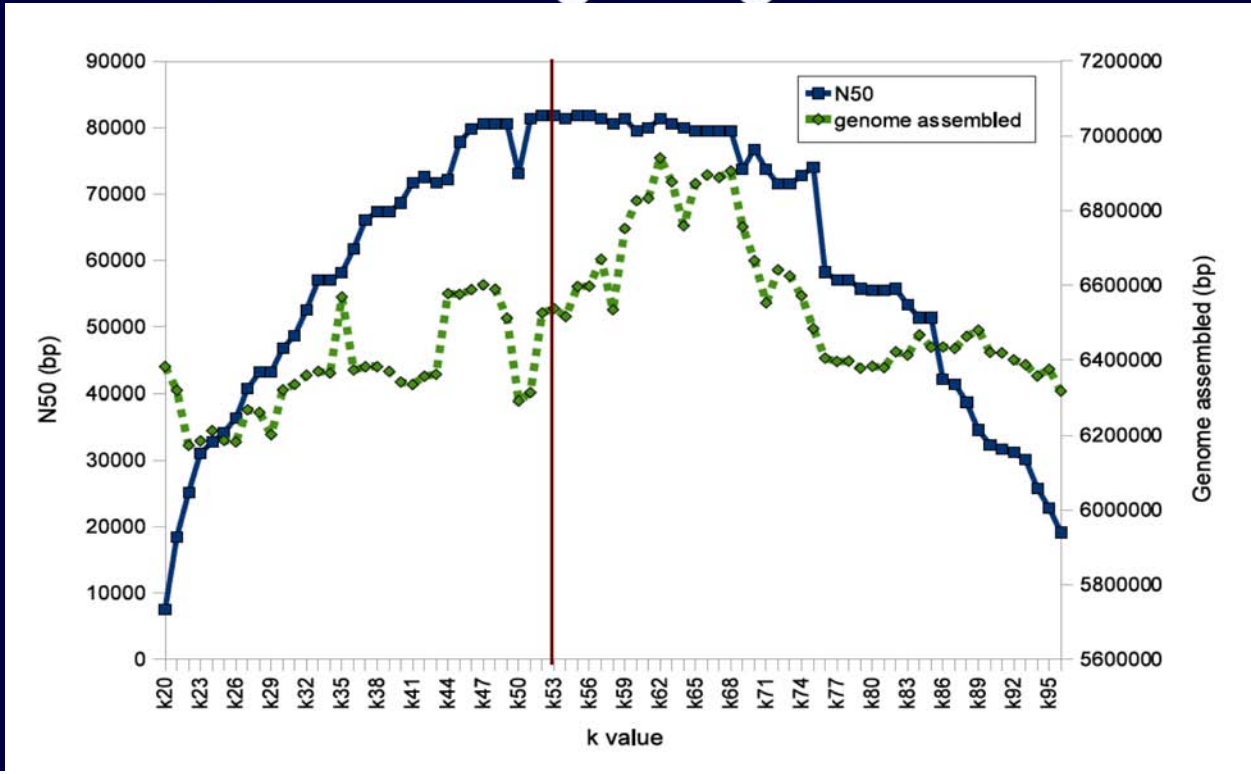
- After SET assembly, reads are aligned to contigs
- Using reads that hit the same contig, empirical fragment size distribution(s) is (are) calculated
- Using reads that hit multiple contigs, inter-contig distances are inferred with a maximum likelihood estimator
- Contigs with coherent and unambiguous distances are joined



Assembly evaluation

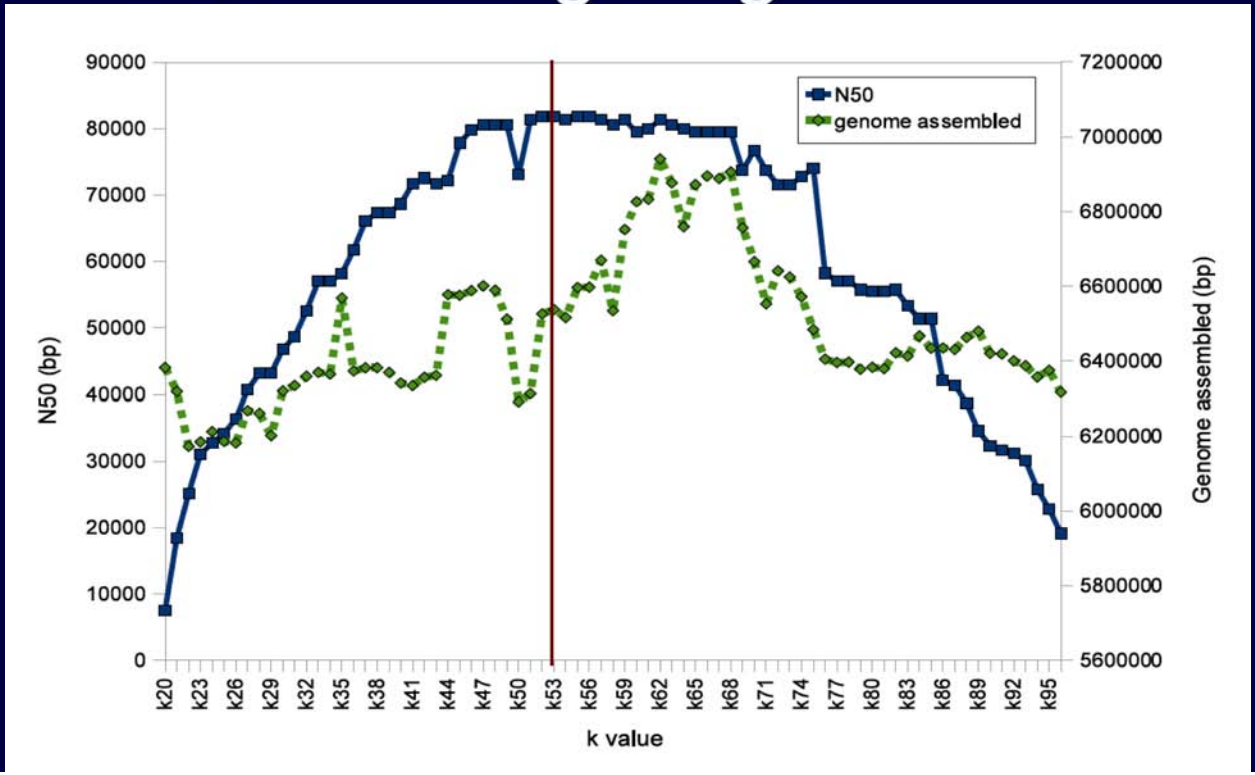
- Total assembly length
- N50
- Max contig length
- Mean contig length

N50 is the contig length such that 50% of assembled bases are in equal or longer contigs.

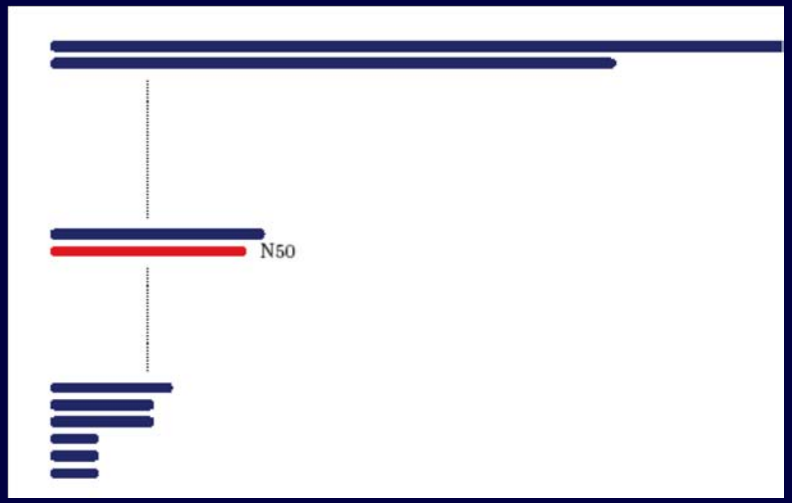


Assembly evaluation

- Total assembly length
- N50
- Max contig length
- Mean contig length



N50 is the contig length such that 50% of assembled bases are in equal or longer contigs.



higher N50
=
lower assembly fragmentation

Genome finishing

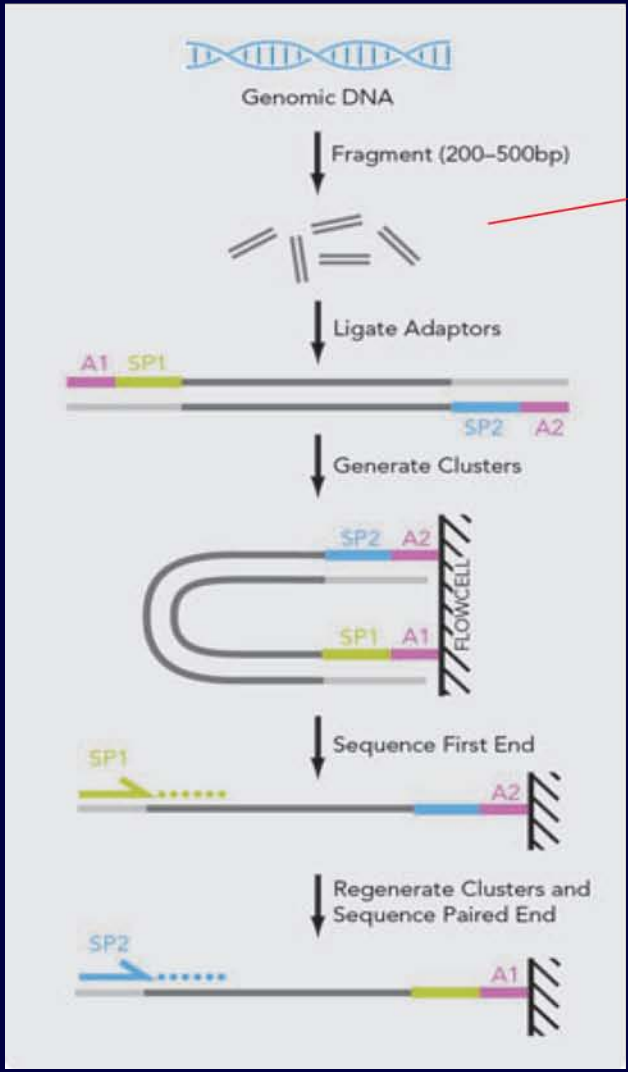
(*de novo* assembly)

- NGS, assembly: a kind of magic, but usually you never get a complete genome at the first shot!

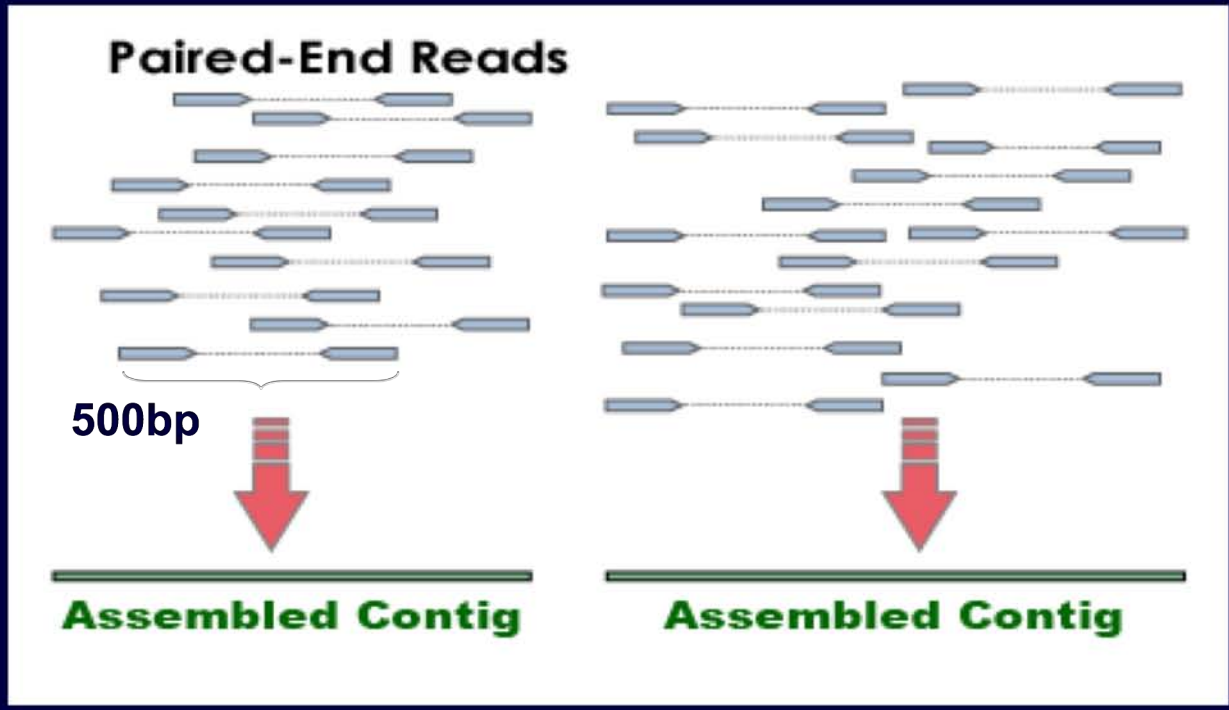
- Main reasons: limits in…
  assembly algorithms
  sampling biases in library preparations

Paired end assembly

infer distance between contigs

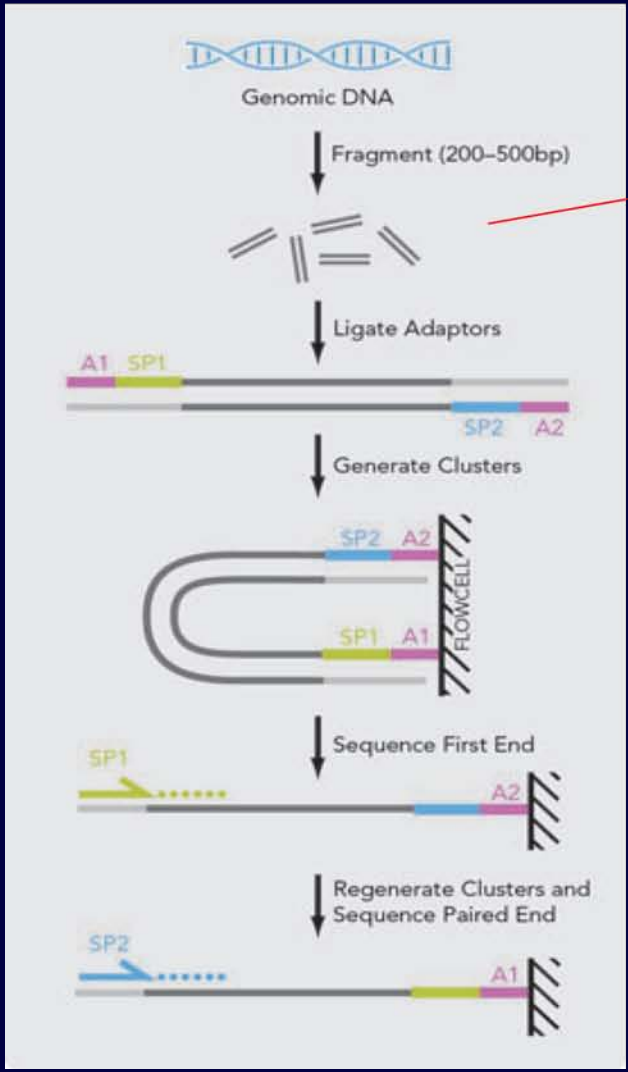


Library fragments are selected based on their size and sequenced at their ends

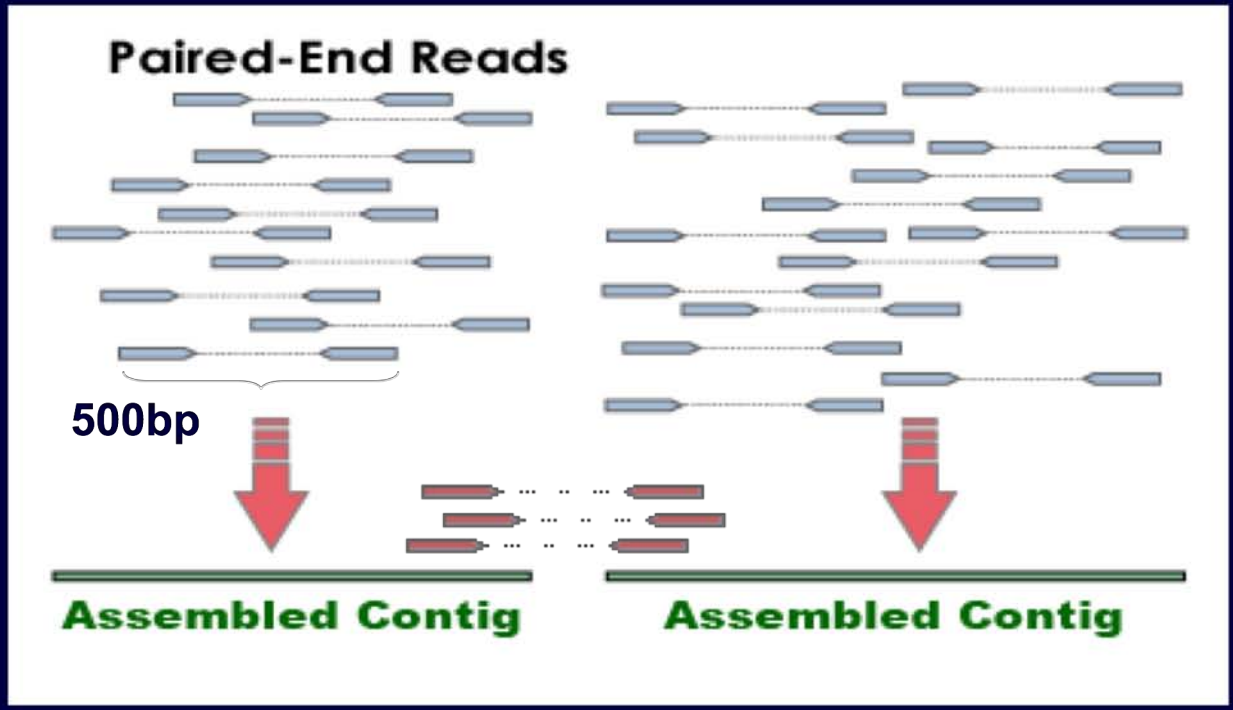


Paired end assembly

infer distance between contigs

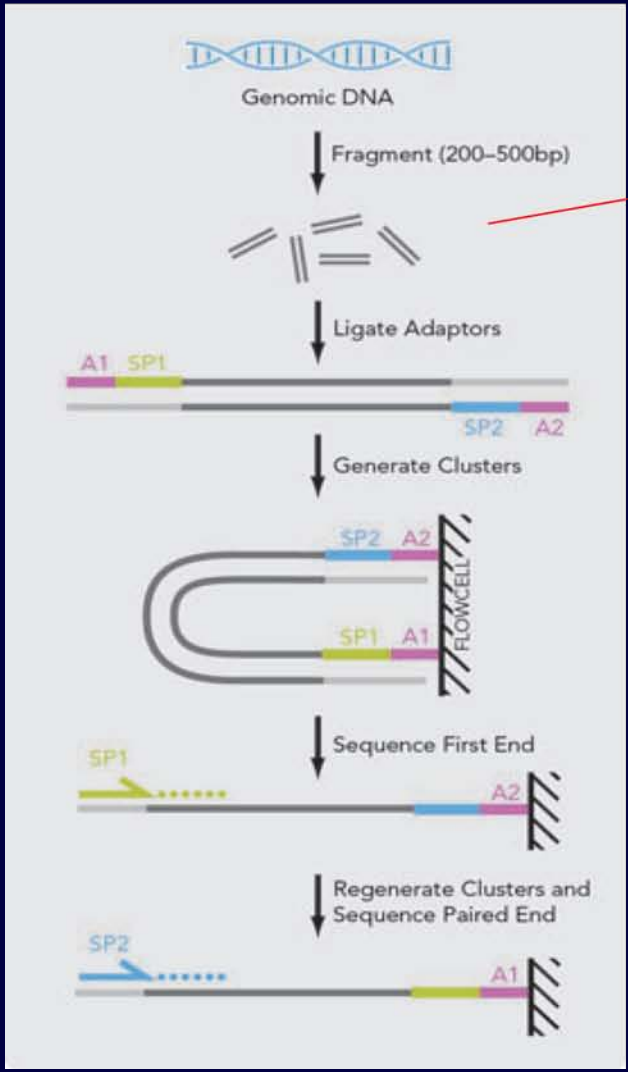


Library fragments are selected based on their size and sequenced at their ends

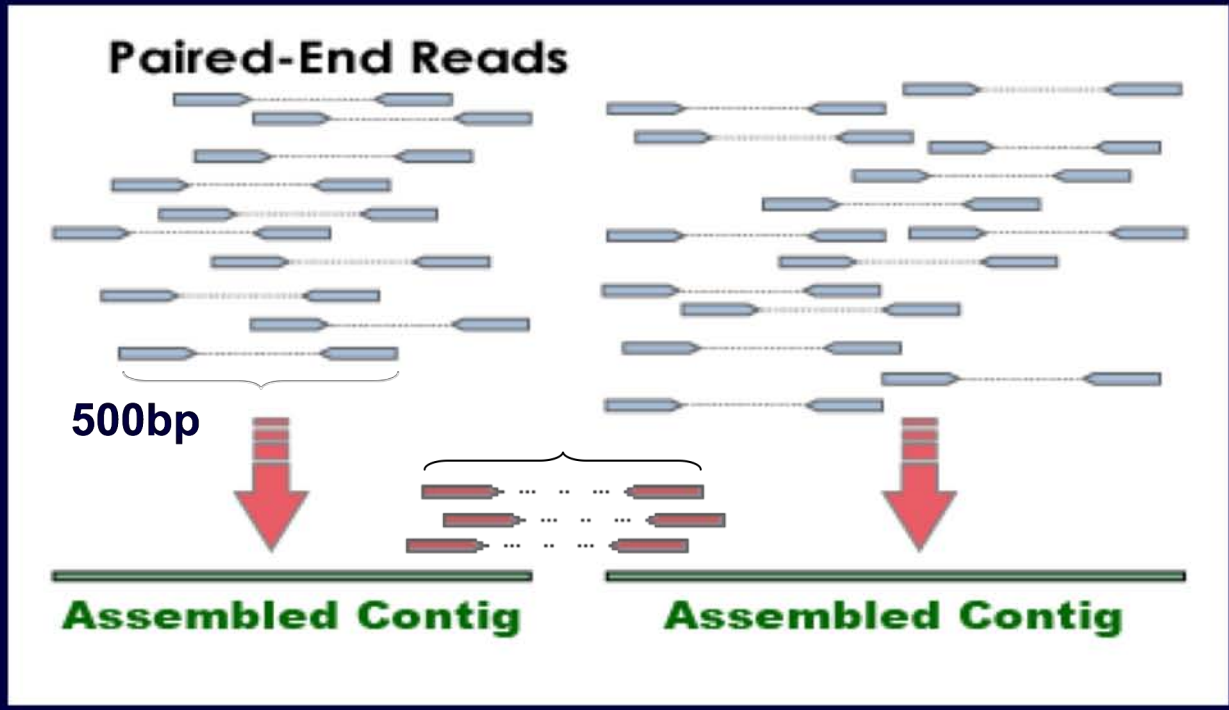


Paired end assembly

infer distance between contigs

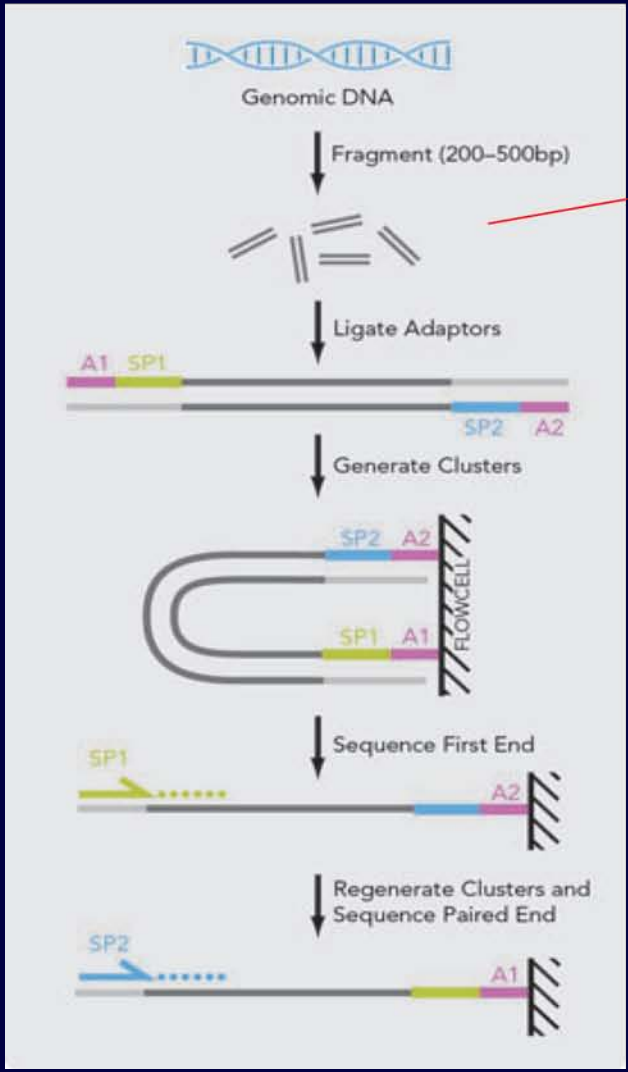


Library fragments are selected based on their size and sequenced at their ends

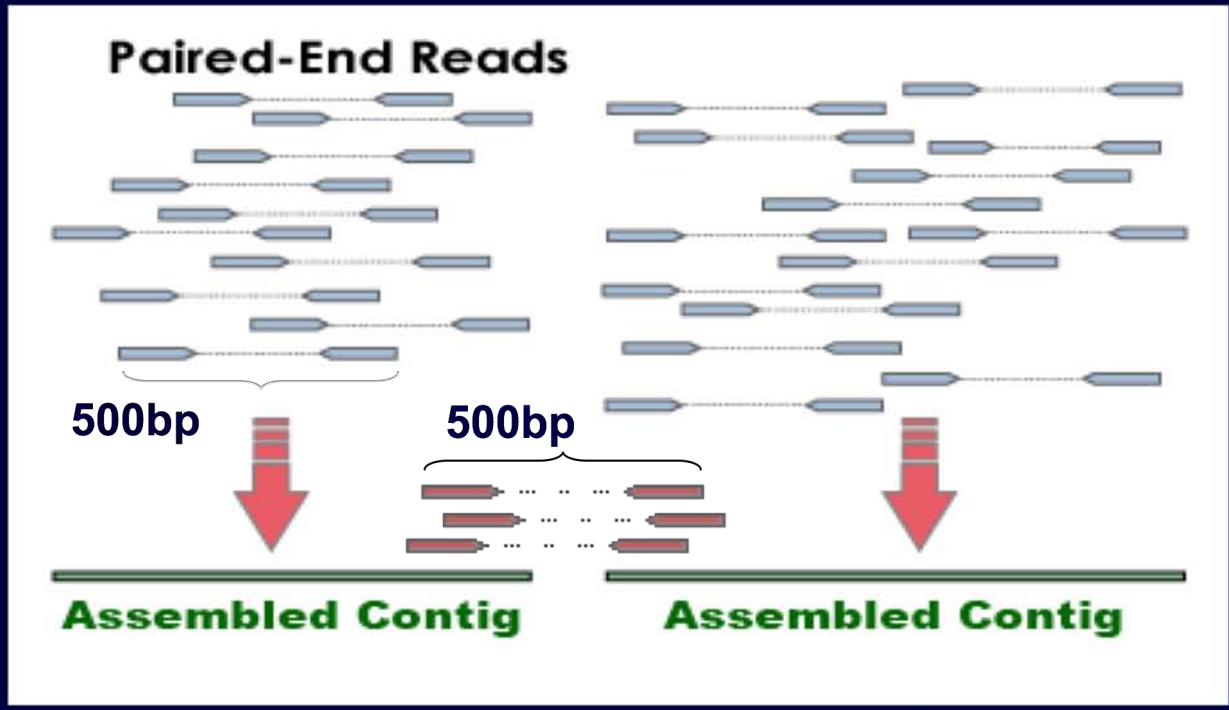


Paired end assembly

infer distance between contigs

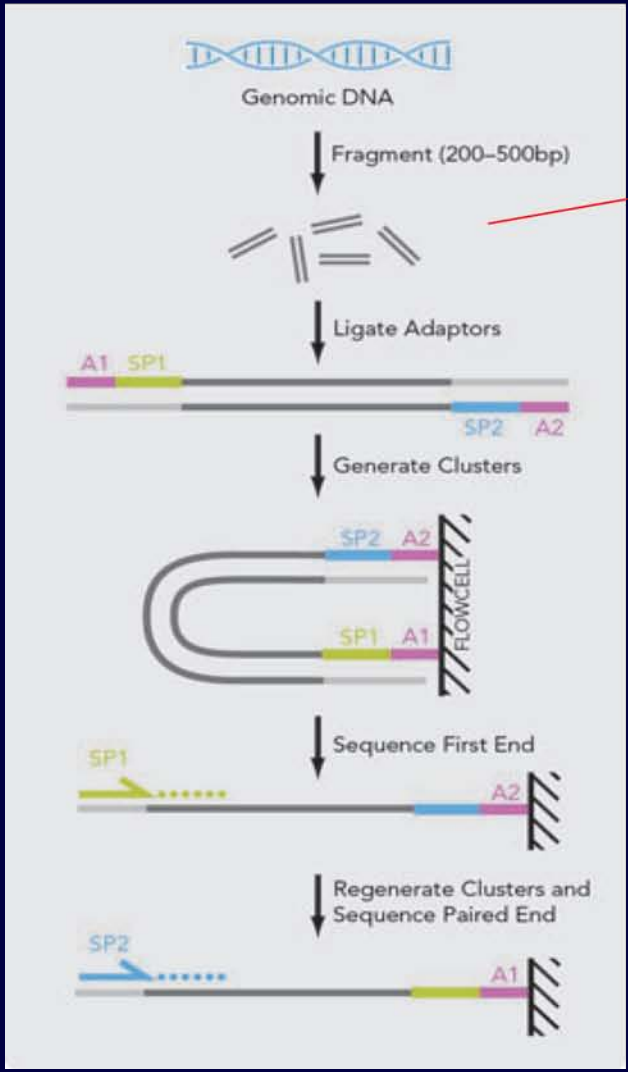


Library fragments are selected based on their size and sequenced at their ends

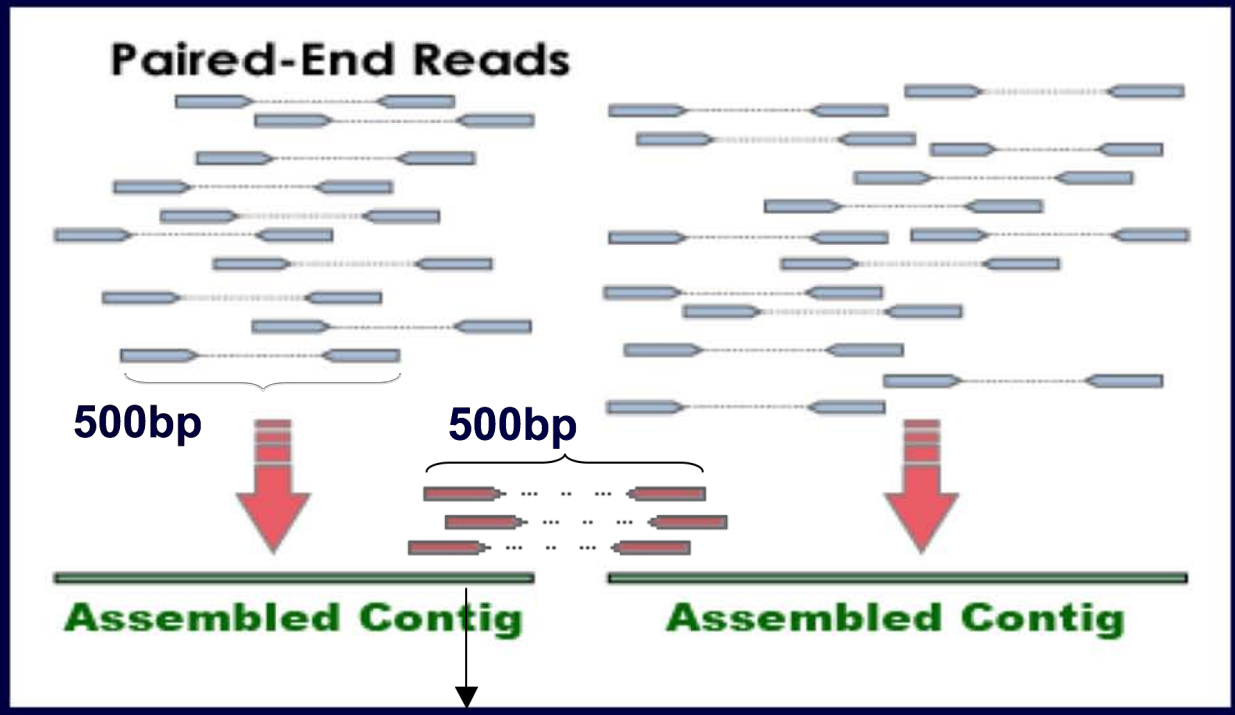


Paired end assembly

infer distance between contigs

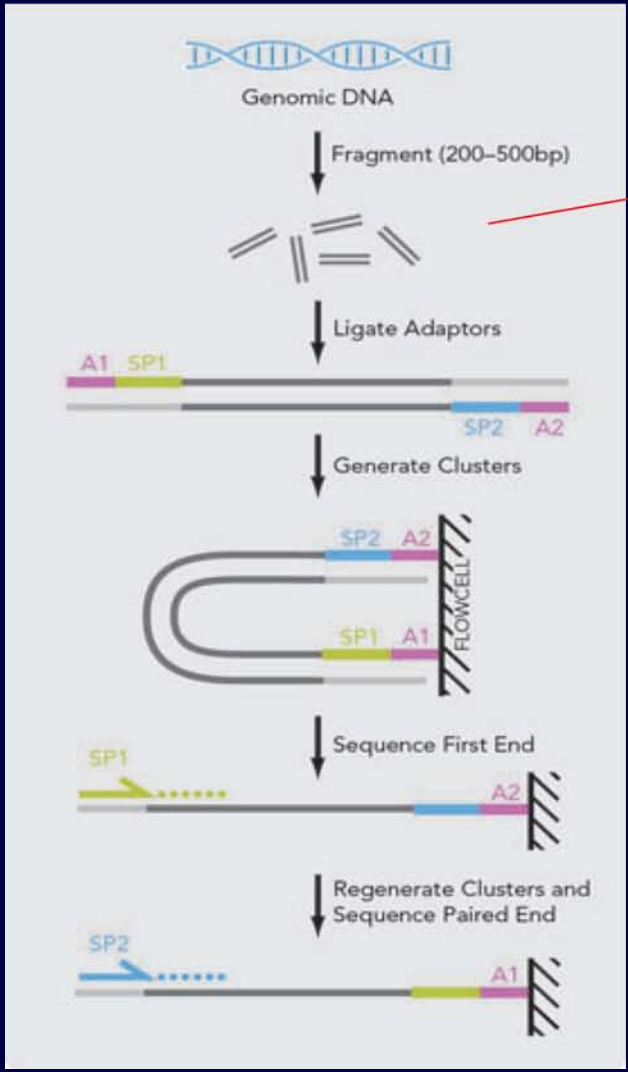


Library fragments are selected based on their size and sequenced at their ends

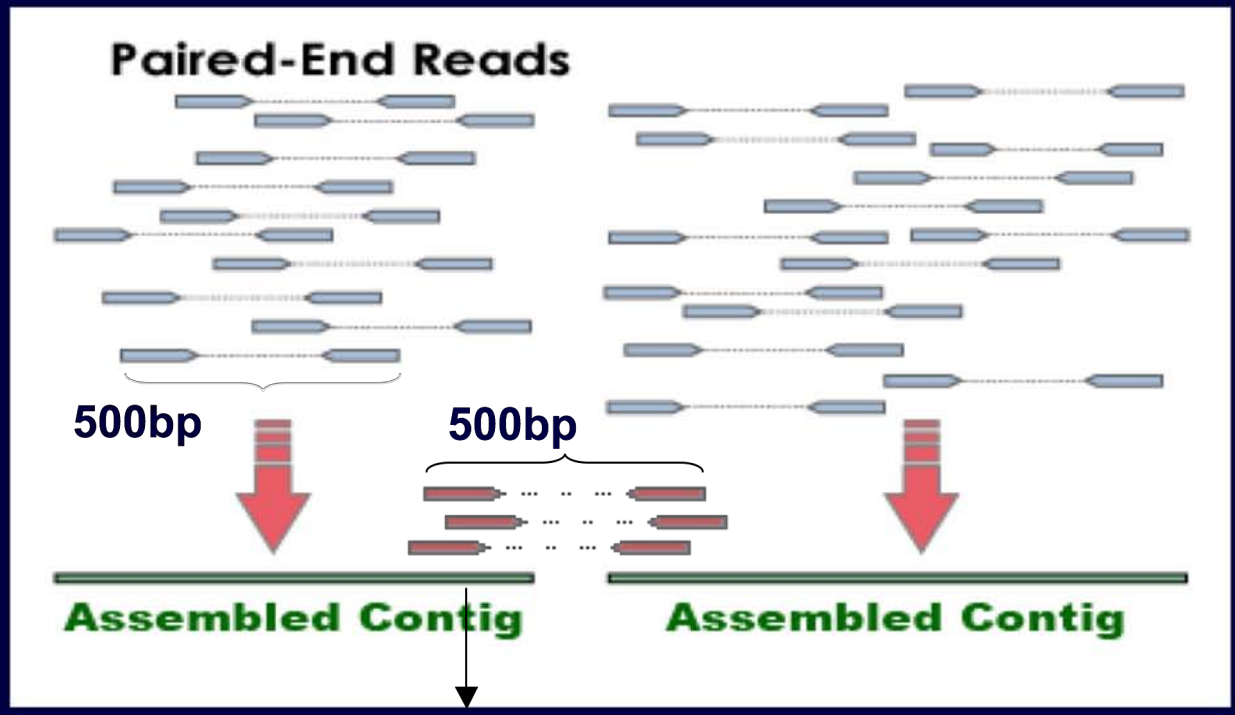


Paired end assembly

infer distance between contigs

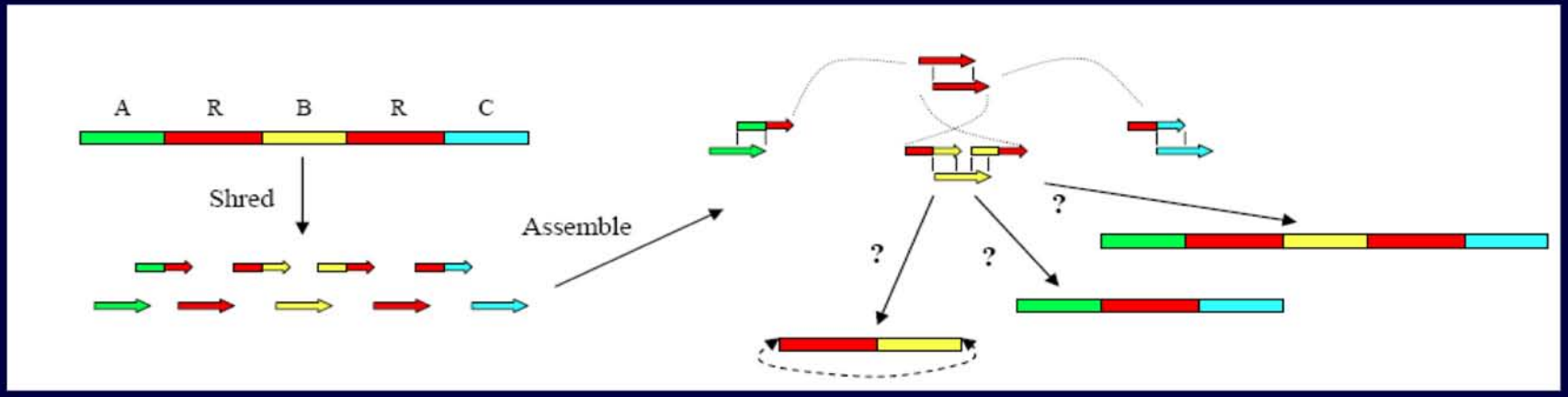


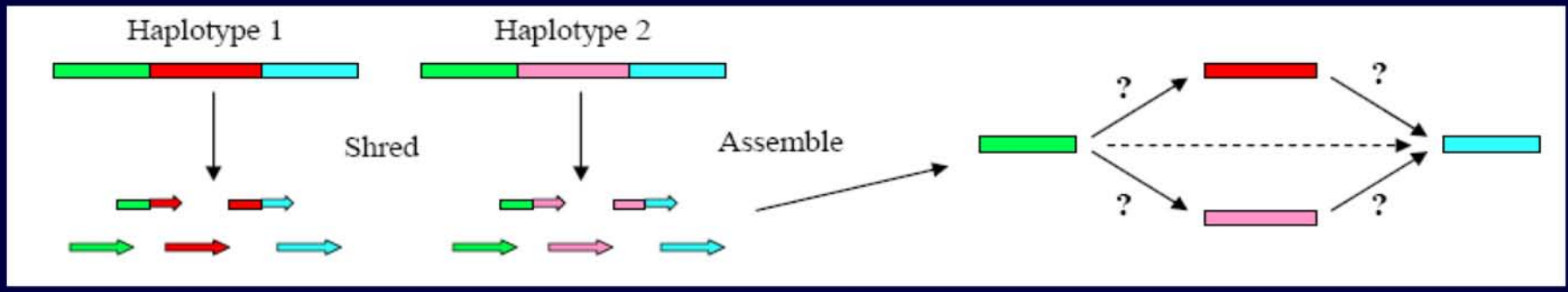
Library fragments are selected based on their size and sequenced at their ends



Paired end reads mapping on different contigs help in building **scaffolds** (ordered series of contigs)

The dark side of short reads…

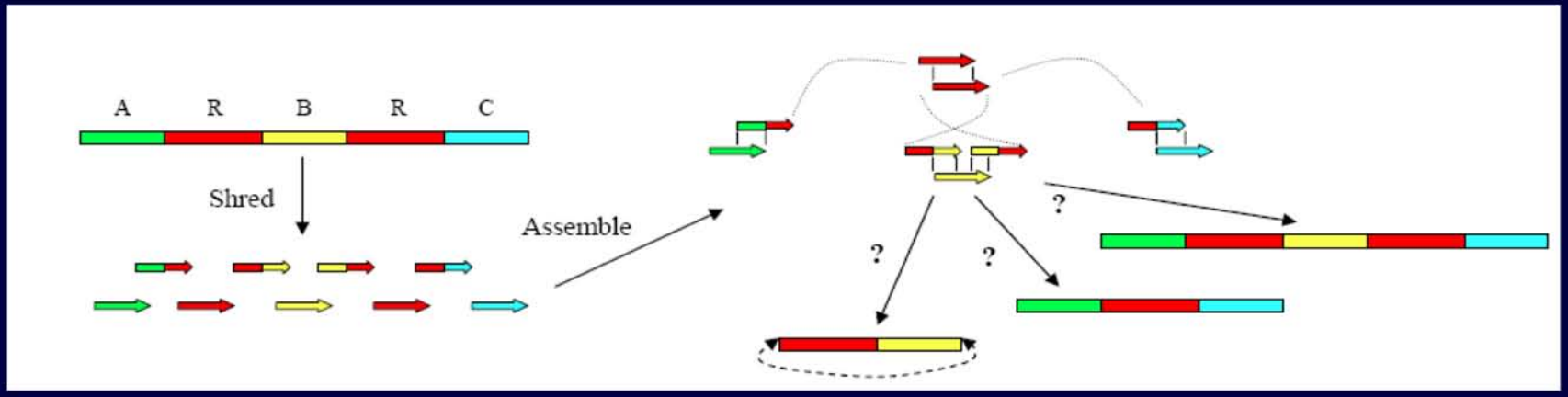


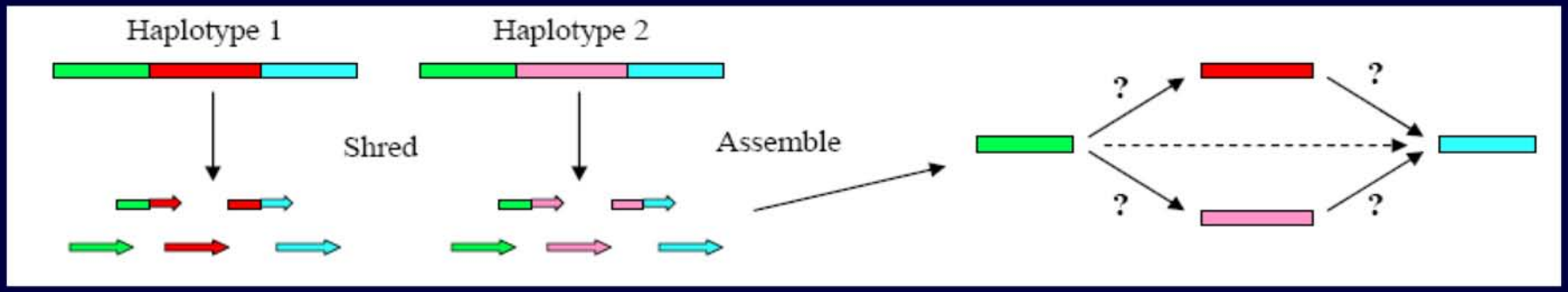


If repeated elements are <u>longer</u> than reads, getting a consistent *de novo* assembly is not trivial

Courtesy of Alberto Policriti (IGA)

The dark side of short reads…





If repeated elements are <u>longer</u> than reads, getting a consistent *de novo* assembly is not trivial

A common strategy involves several libraries with different insert sizes (eventually, validation with Sanger sequencing of troubling regions)

Courtesy of Alberto Policriti (IGA)

Genome finishing

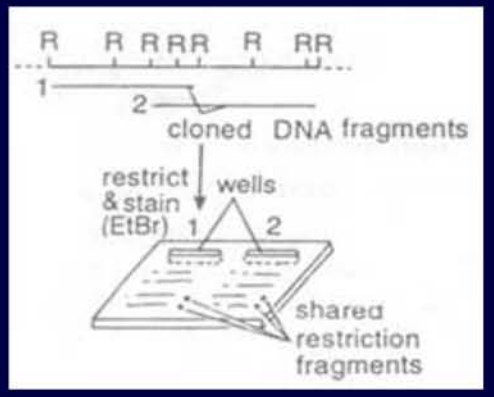
Scaffold

Genome projects need scaffolding:

Determination of relative placement and orientation of contigs returned by assembly process

Traditionally… physical or genetic maps (DNA fingerprinting)

- Cut clones with RE
- Gel separation
- Pattern of DNA migration
- Merging of patterns from different clones



Optical mapping

Extends restriction mapping by providing, in addition to the set of fragments sizes, information about their order on DNA

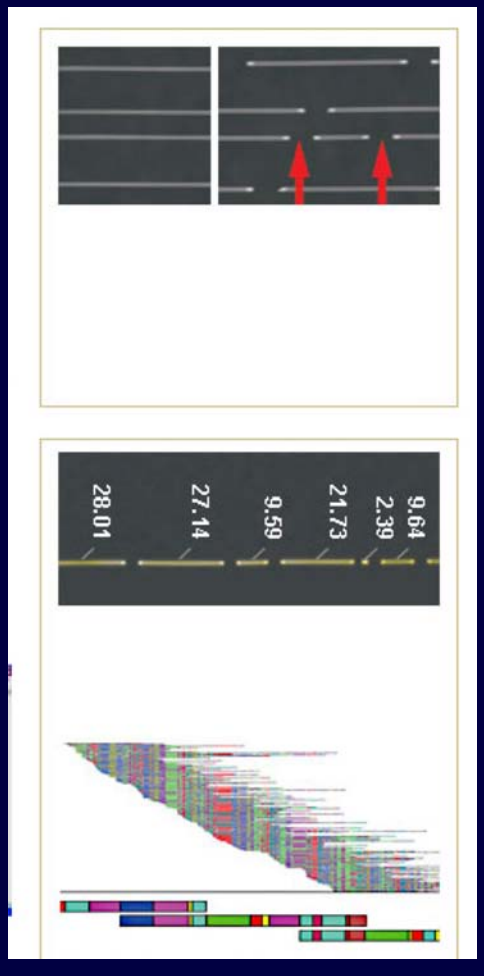
(ordered restriction mapping, *Samad et al. 1995*)

Contigs restricted *in silico* are placed on the restriction optical map

Optical mapping

Extends restriction mapping by providing, in addition to the set of fragments sizes, information about their order on DNA

(ordered restriction mapping, *Samad et al. 1995*)



Contigs restricted *in silico* are placed on the restriction optical map

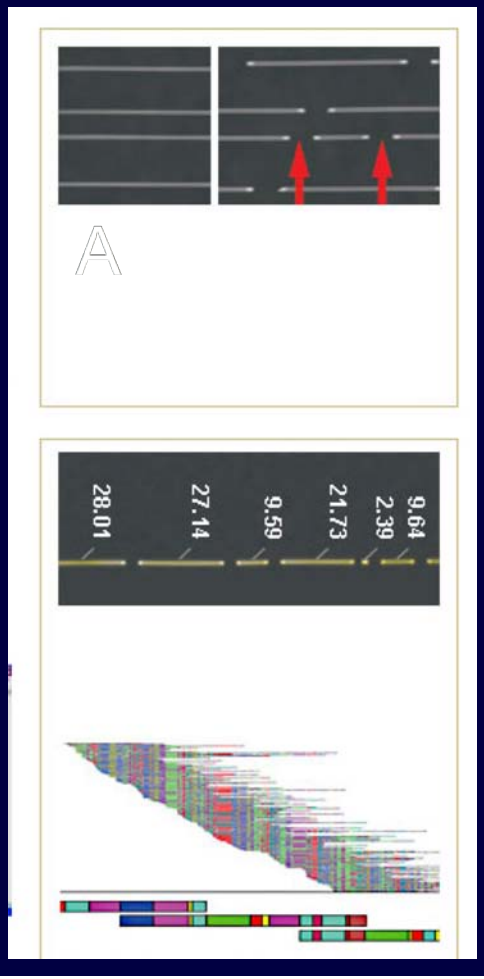
OpGen®

www.opgen.com

Optical mapping

Extends restriction mapping by providing, in addition to the set of fragments sizes, information about their order on DNA

(ordered restriction mapping, *Samad et al. 1995*)



Contigs restricted *in silico* are placed on the restriction optical map

A) DNA single molecules electrostatically **immobilized**

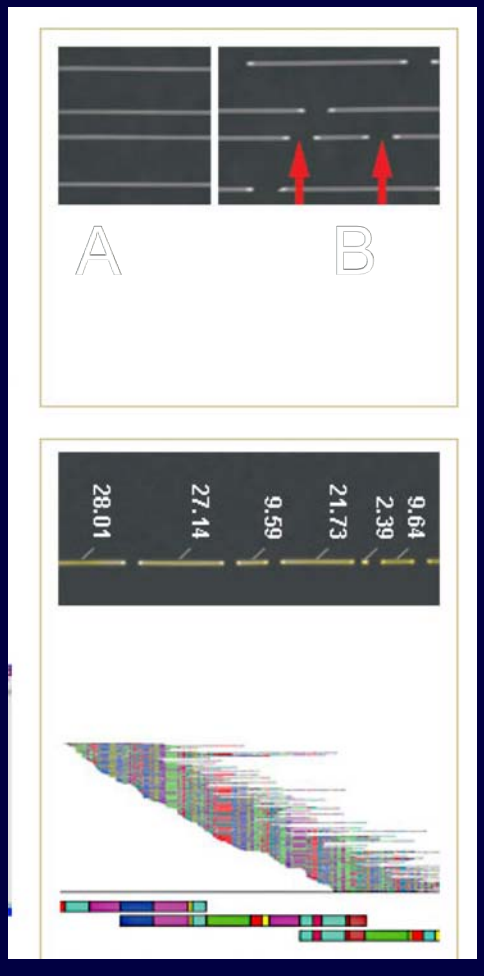
OpGen®

www.opgen.com

Optical mapping

Extends restriction mapping by providing, in addition to the set of fragments sizes, information about their order on DNA

(ordered restriction mapping, *Samad et al. 1995*)



Contigs restricted *in silico* are placed on the restriction optical map

A) DNA single molecules electrostatically **immobilized**

B) DNA digestion; restriction fragments mantain their **order**

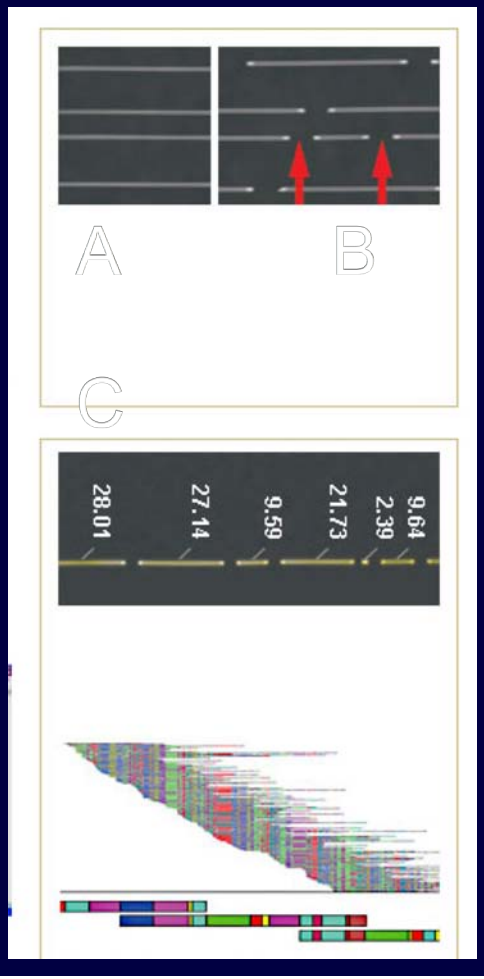
OpGen®

www.opgen.com

Optical mapping

Extends restriction mapping by providing, in addition to the set of fragments sizes, information about their order on DNA

(ordered restriction mapping, *Samad et al. 1995*)



Contigs restricted *in silico* are placed on the restriction optical map

A) DNA single molecules electrostatically **immobilized**

B) DNA digestion; restriction fragments mantain their **order**

C) DNA fragments are stained with fluorescent dye; **fluorescence intensity is proportional to fragment length**

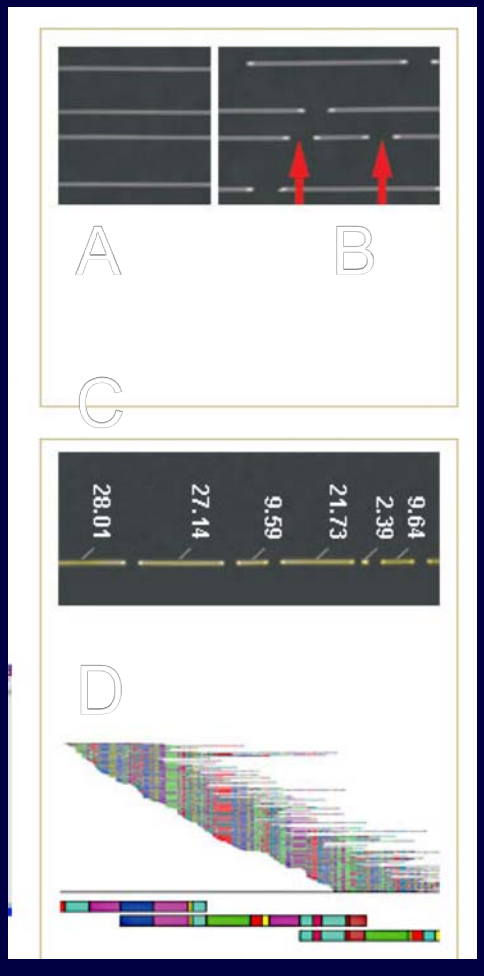
OpGen®

www.opgen.com

Optical mapping

Extends restriction mapping by providing, in addition to the set of fragments sizes, information about their order on DNA

(ordered restriction mapping, *Samad et al. 1995*)



Contigs restricted *in silico* are placed on the restriction optical map

A) DNA single molecules electrostatically **immobilized**

B) DNA digestion; restriction fragments mantain their **order**

C) DNA fragments are stained with fluorescent dye; **fluorescence intensity is proportional to fragment length**

D) Fragment patterns are overlapped; a **Whole Genome Map** is assembled with a **minimum 30X coverage depth**.

OpGen®

www.opgen.com

Optical mapping methods

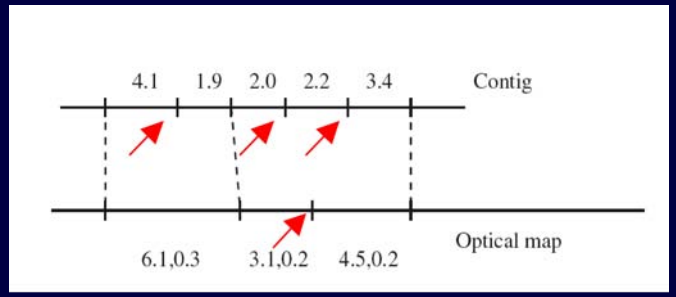
Contigs are restricted *in silico*
Restriction patterns aligned vs optical map

Optical mapping methods

Contigs are restricted *in silico*
Restriction patterns aligned vs optical map

A series of fragment sizes should match a unique region of the map, but it should be taken into account…

- Sequencing errors (false positive or false negative RSs)
- Assembly errors
- Small contigs and/or contigs originating in repeat regions (non-unique placement)
- Sequences from foreign DNA (no placement)

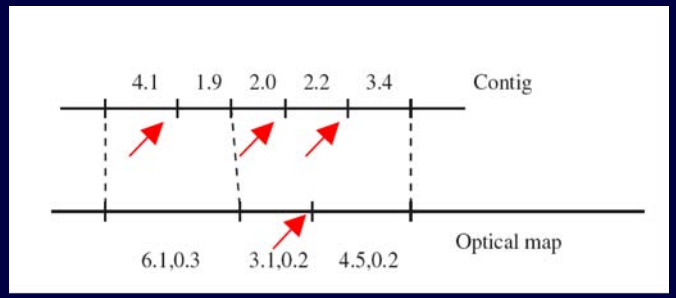


Optical mapping methods

Contigs are restricted *in silico*
Restriction patterns aligned vs optical map

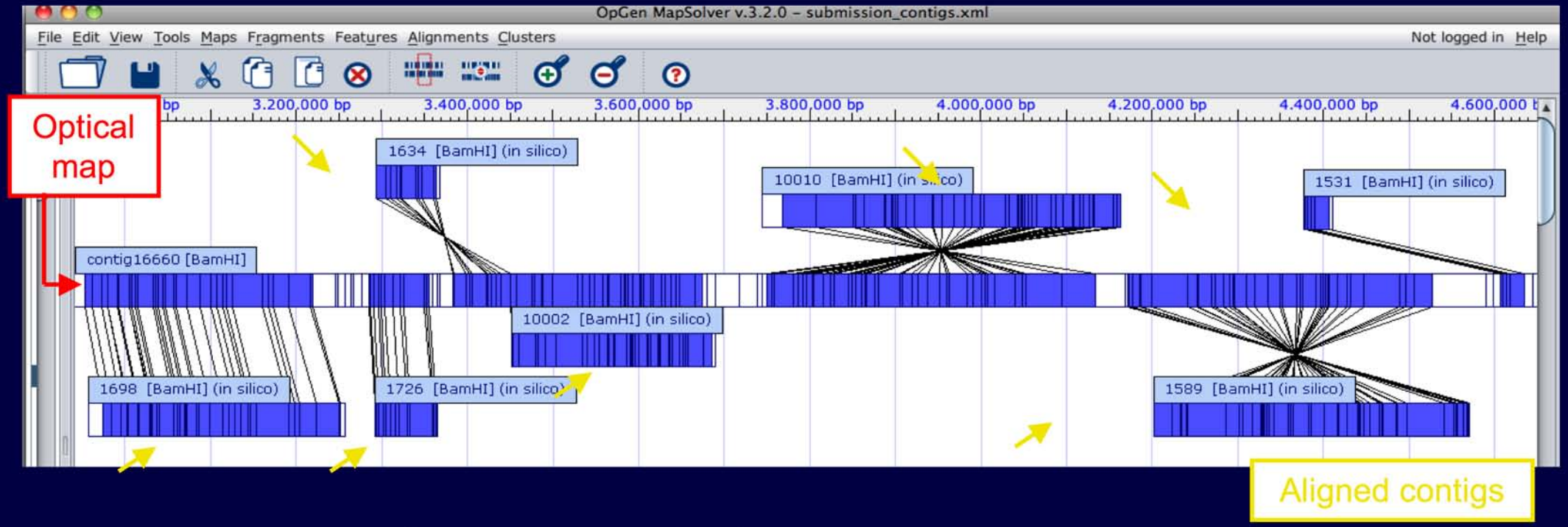
A series of fragment sizes should match a unique region of the map, but it should be taken into account…

- Sequencing errors (false positive or false negative RSs)
- Assembly errors
- Small contigs and/or contigs originating in repeat regions (non-unique placement)
- Sequences from foreign DNA (no placement)



MapSolver (OpGen)
Reliable mapping for contigs:
- Longer than 40Kb
- With at least 4 RSs

Optical mapping



KF707 OM excerpt (region 3Mb - 4.6Mb)

Further readings

Sequencing technologies

- Mardis ER - *Trends in Genetics* 2008 (attached)
- Liu L *et al* - 2012 (attached)
- Lee H, Tang H - 2012 (attached)

} Also an introduction to 3rd generation sequencing methods

Aligners / assemblers

- Flicek P, Birney E - *Nature Methods* 2009 (attached)
- Lee H, Tang H - 2012 (attached)

Contact:
dome.simone@gmail.com