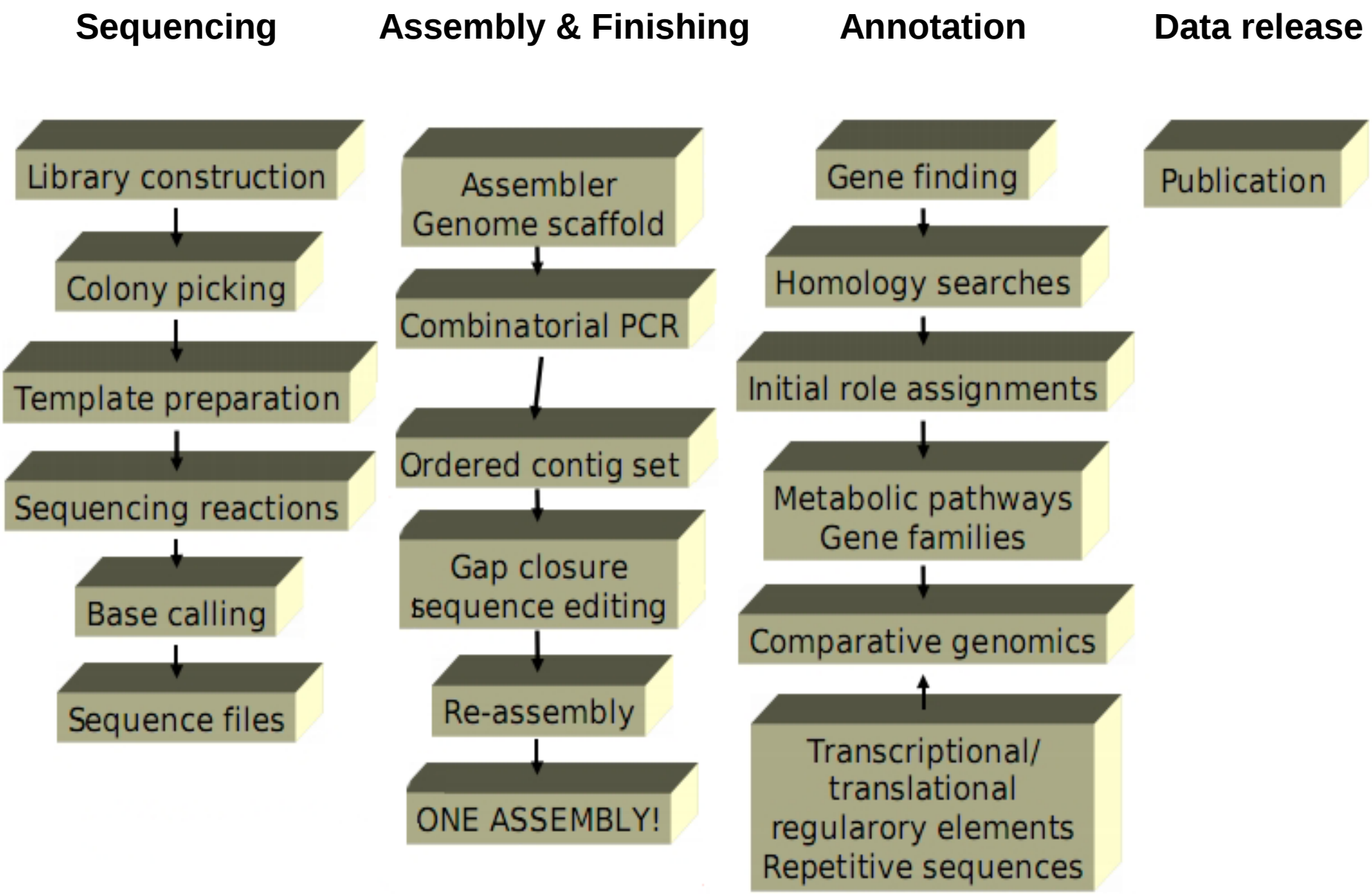
Genomes assemblies and finishing

Bioinformatics teachings

http://bioinfomed.fr - Olivier Croce -

Summary


Sequencing
Assembly & Finishing
Annotation
Data release
Library construction
Colony picking
Template preparation
Sequencing reactions
Base calling
Sequence files
Assembler
Genome scaffold
Combinatorial PCR
Ordered contig set
Gap closure
sequence editing
Re-assembly
ONE ASSEMBLY!
Gene finding
Homology searches
Initial role assignments
Metabolic pathways
Gene families
Comparative genomics
Transcriptional/
translational
regularory elements
Repetitive sequences
Publication

Data release

- Submission of the sequence on public databases
- Not always => publication

<u>3 main public databases</u>:

- EMBL-EBI - ENA (European Nucleotide Archive) ** http://www.ebi.ac.uk/embl/

- GenBank (USA) – NCBI ** http://www.ncbi.nlm.nih.gov/Genbank/

- DDBJ (DNA DataBank of Japon) – CIB ** http://www.ddbj.nig.ac.jp/

They are associated (International Nucleotide Sequence Database Collaboration) and exchange the same data which is periodically duplicated together

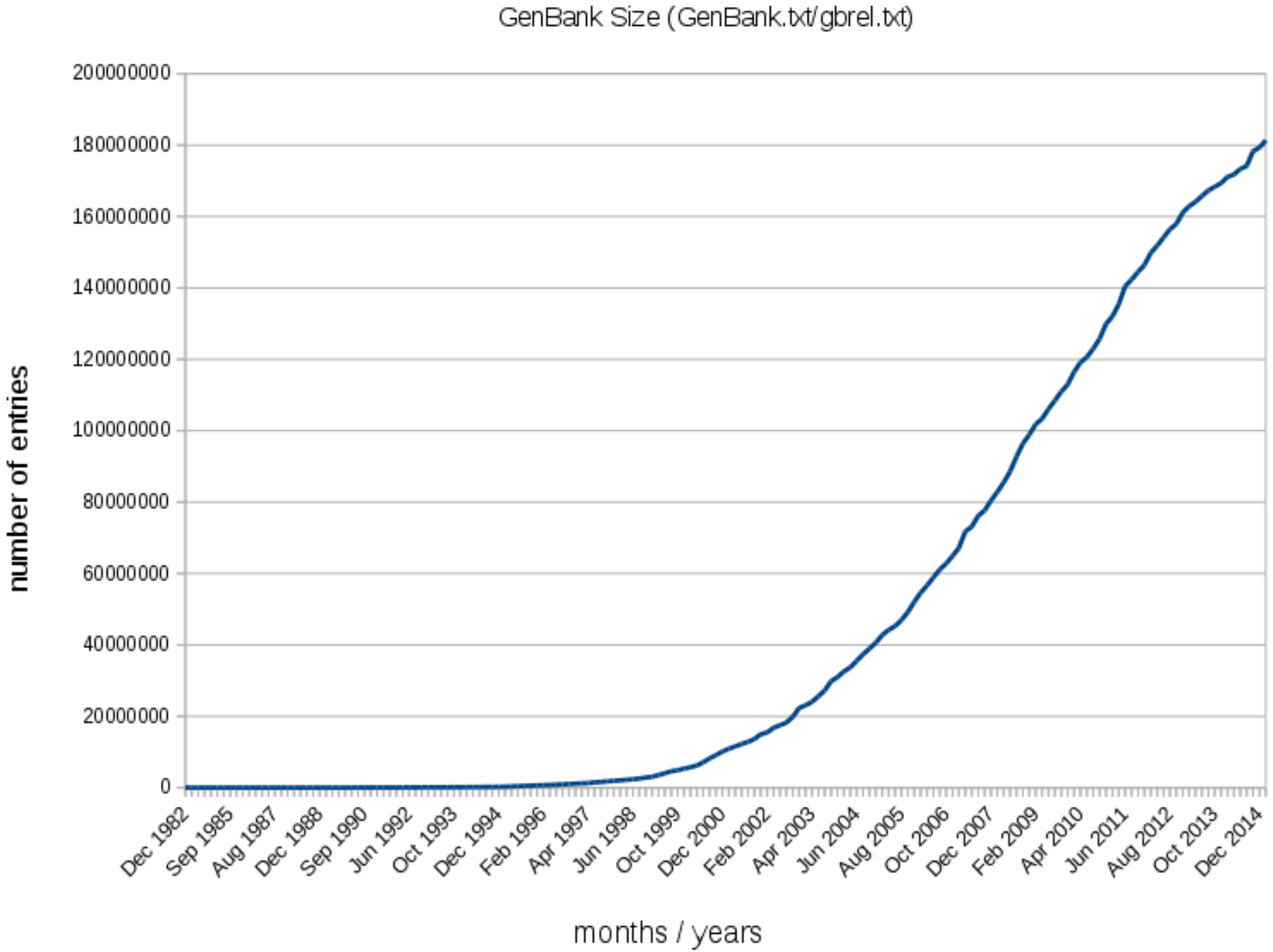
<u>Embl, Ebi, Ddbj contain</u>:

- Sequences of DNA or RNA from various sequencers technologies and from many labs

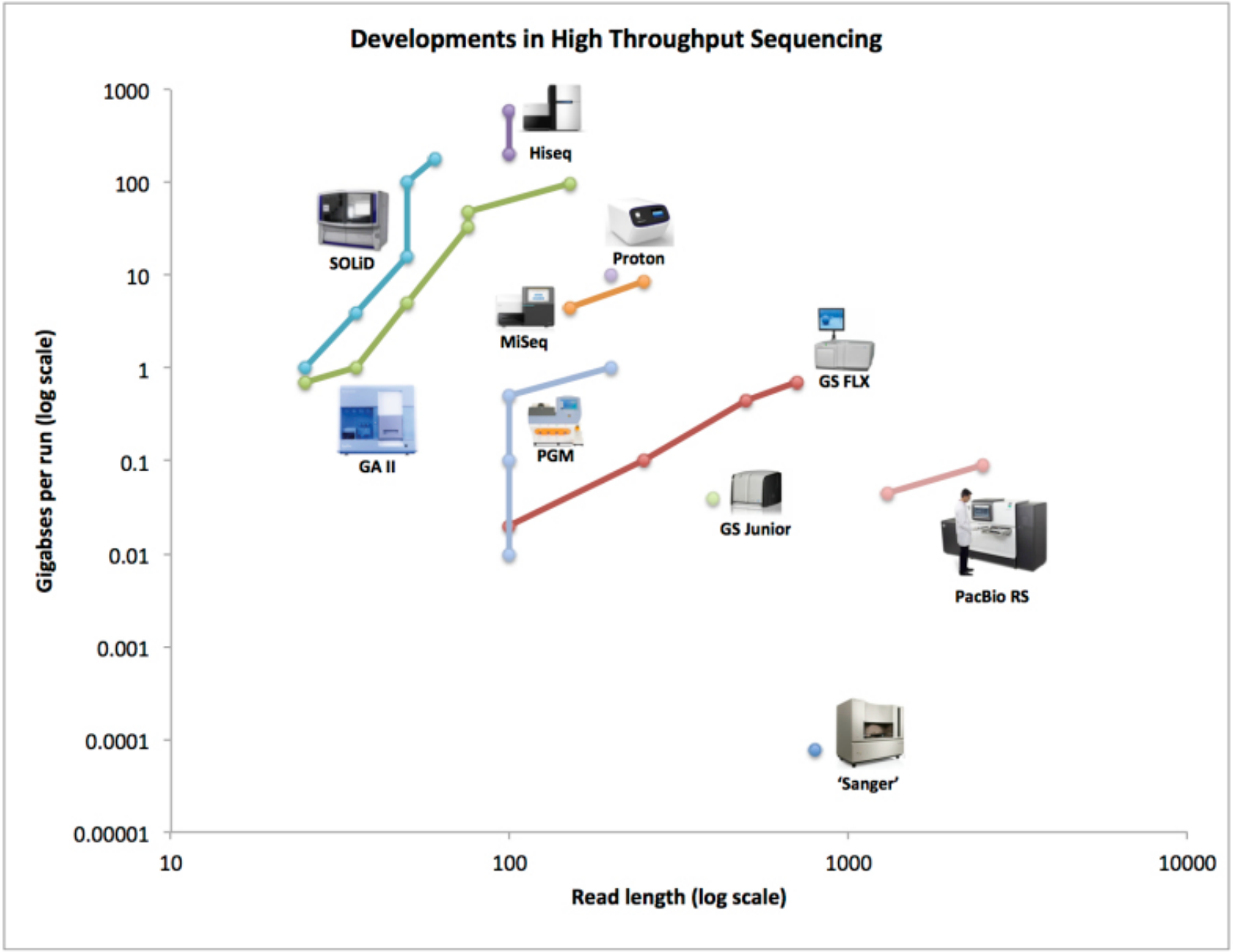
    * Some genome fragments : one or more genes, intergenic sequences, parts of a genome
    * Completed genomes
    * mRNA, tRNA, rRNA (ie. 16s)

- Annotations

Data release



Sequencing


Developments in High Throughput Sequencing
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Hiseq
SOLiD
Proton
MiSeq
GS FLX
GA II
PGM
GS Junior
PacBio RS
'Sanger'

Finishing goals

<u>2 conceptions of the sequences finishing</u>:

- The genome sequence must be completed and with a high quality before the release.

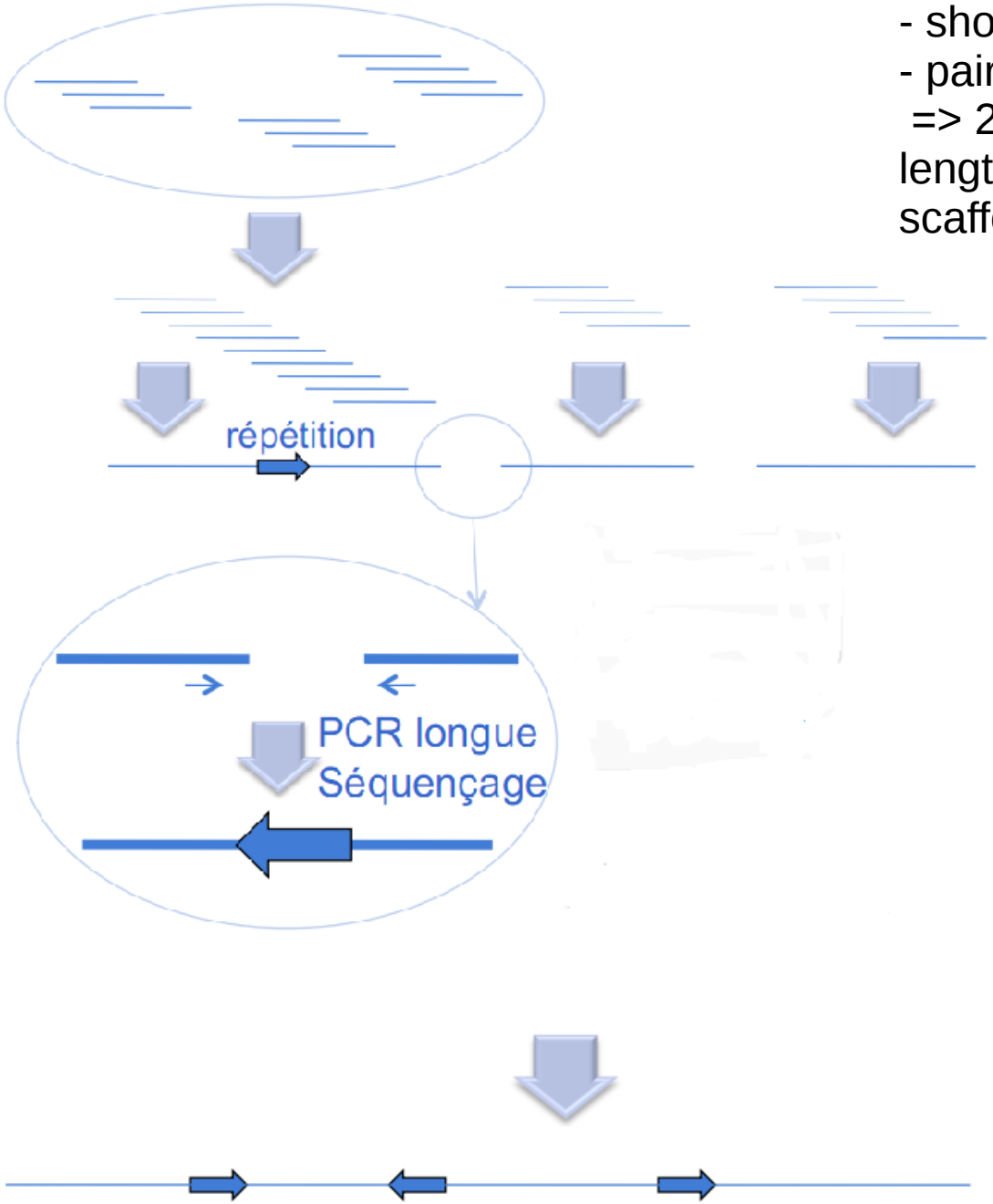
Of course the best, but very time consuming.
Actually, 90-95 % of a microorganism genome could be easy covered without finishing, but the 5-10 % remained can take many weeks or months to be ended.

- The sequence should be uncompleted with a draft quality, whether most of the genes are sequenced and identified.

Many eukaryote genomes are only draft genomes, because of the complexity of finishing

=> In general, fundamental research usually performs high quality genomes
and applicative research (industry, our lab) usually performs draft genomes

Construction of a library of genomes fragments:
- shotgun = single fragments
- paired-end (or mate-pair)
 => 2 fragments linked by an insert of a known length (~5 kb for 454 or Illumina), needed for scaffolding



Assembly process => construction of contigs (and scaffolds) from reads

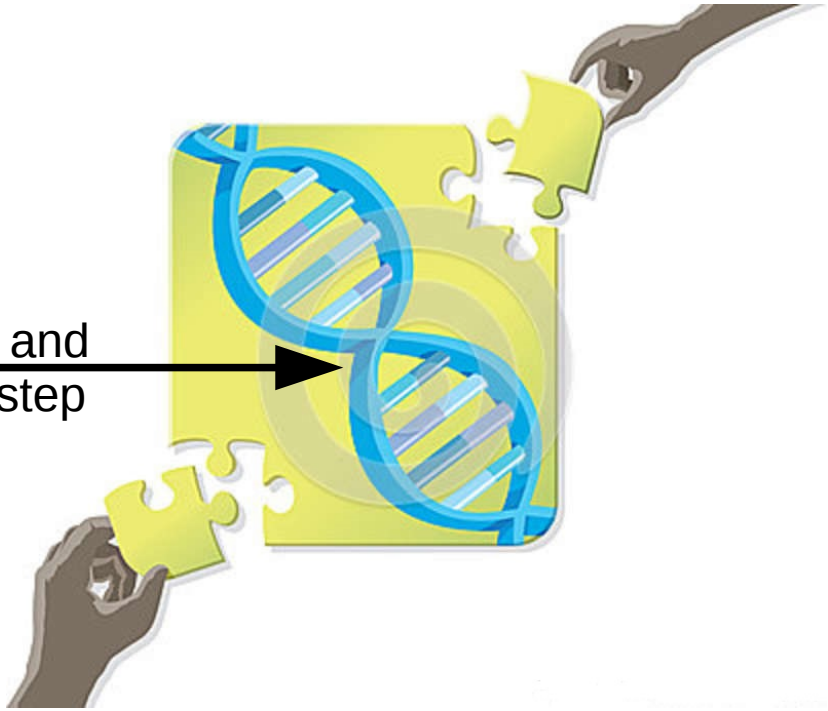
In silico finishing + PCR to fill gaps

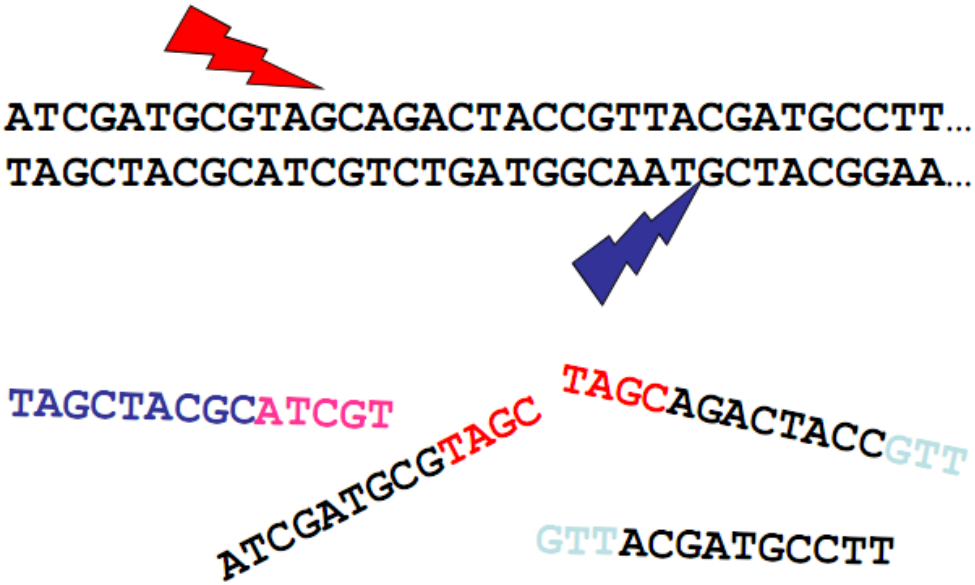
Checking with a closed reference or using annotations

Sequencing

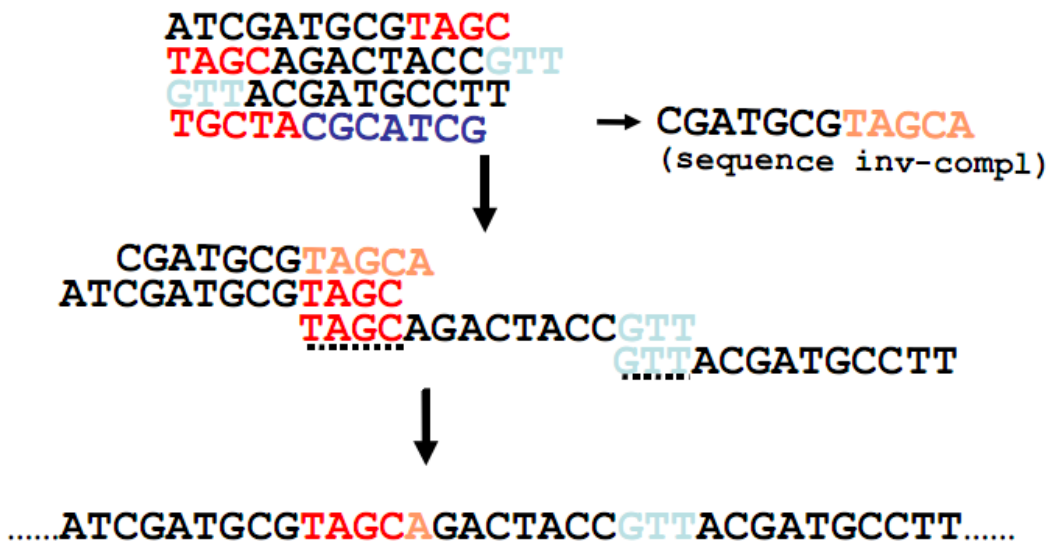
Sequencing step: reads have heterogeneous distribution

Assembly and finishing step





Fragmentation + sequencing
=> sets of reads



Build of contigs with overlapping regions

Assembly :
=> alignements of reads + consensus

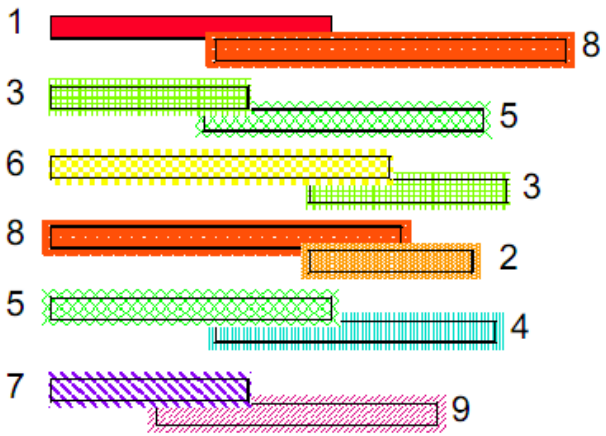
Search for best pairings

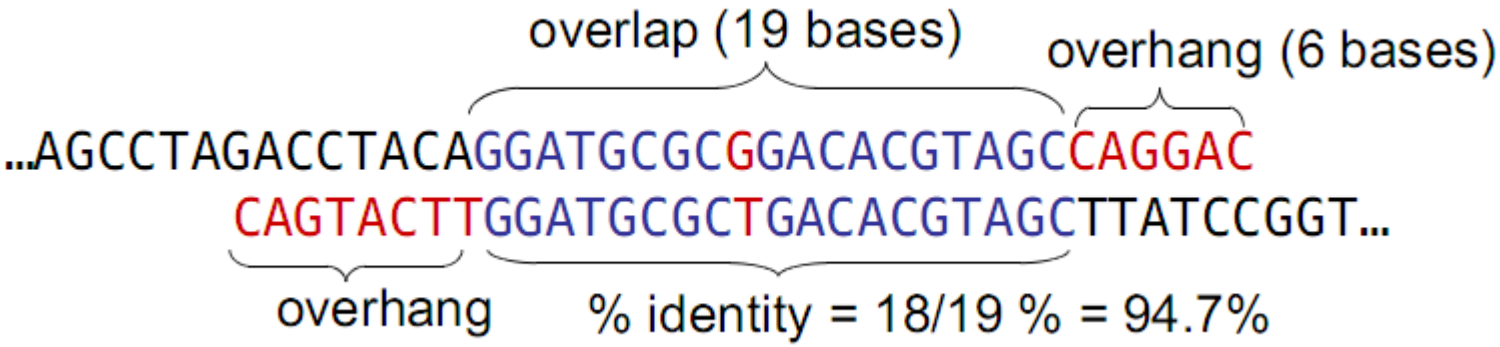
- Compare each sequence (and its reverse complement) against every others sequence to find the best overlapping

=> list of best candidates with similarities criteria

Best candidate is a compromise between :

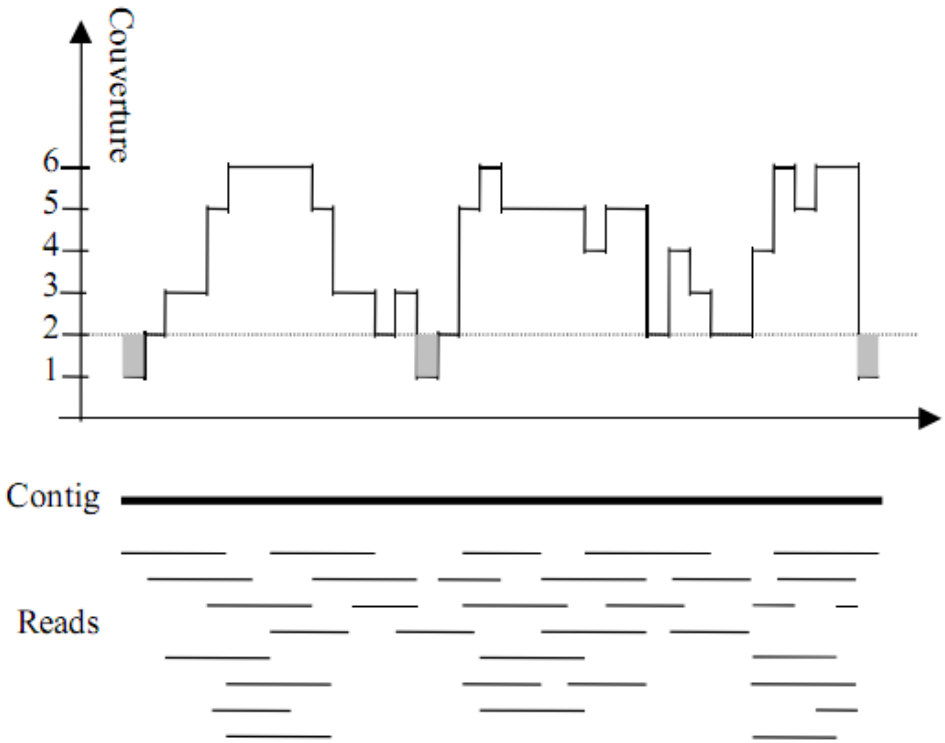
- maximum overlap length - region of similarity between regions
- minimum overhang length - unaligned ends of the sequences
- maximum % identity in overlap region
- minimum repeat length





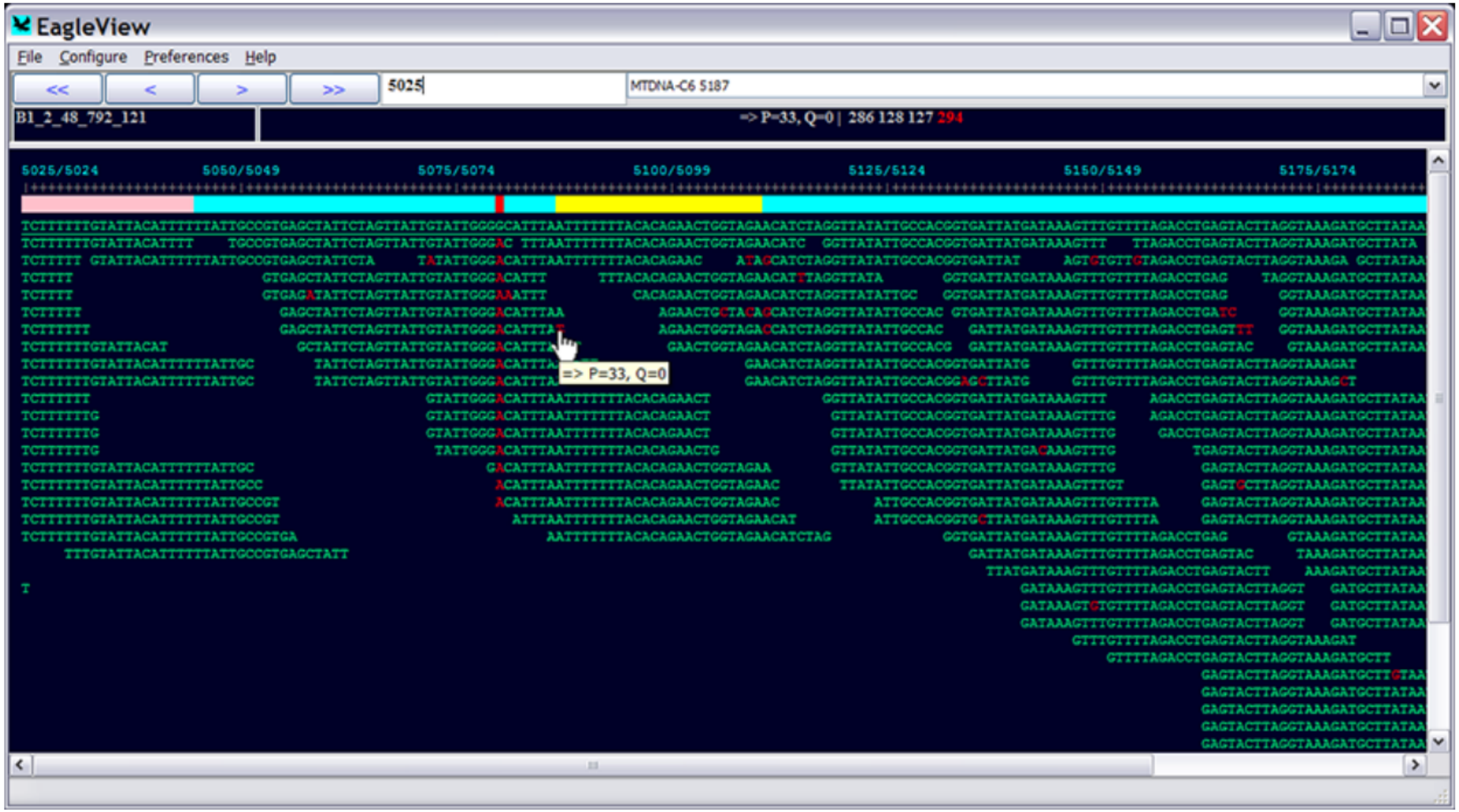
Main remaining problems:

- Bad assembly of reads
- Low coverage of reads



- Bad insert size estimation
- Different orientation of contigs
- Error of sequencing
- Repeat sequence ambiguities

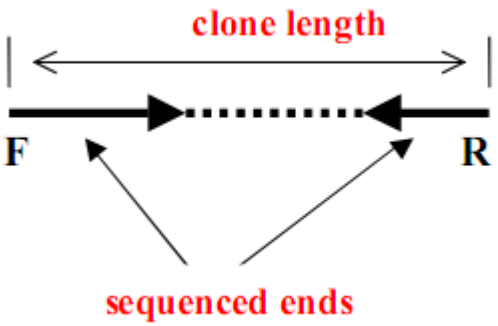
Remap reads on a reference genome or assembled genome itself
Highlighting errors, ie. sequencing error or SNP, show coverage



Reads paired-end (similar to **mate-pair**)

- The distance between the reads is known (length of the insert), with some experimental uncertainty

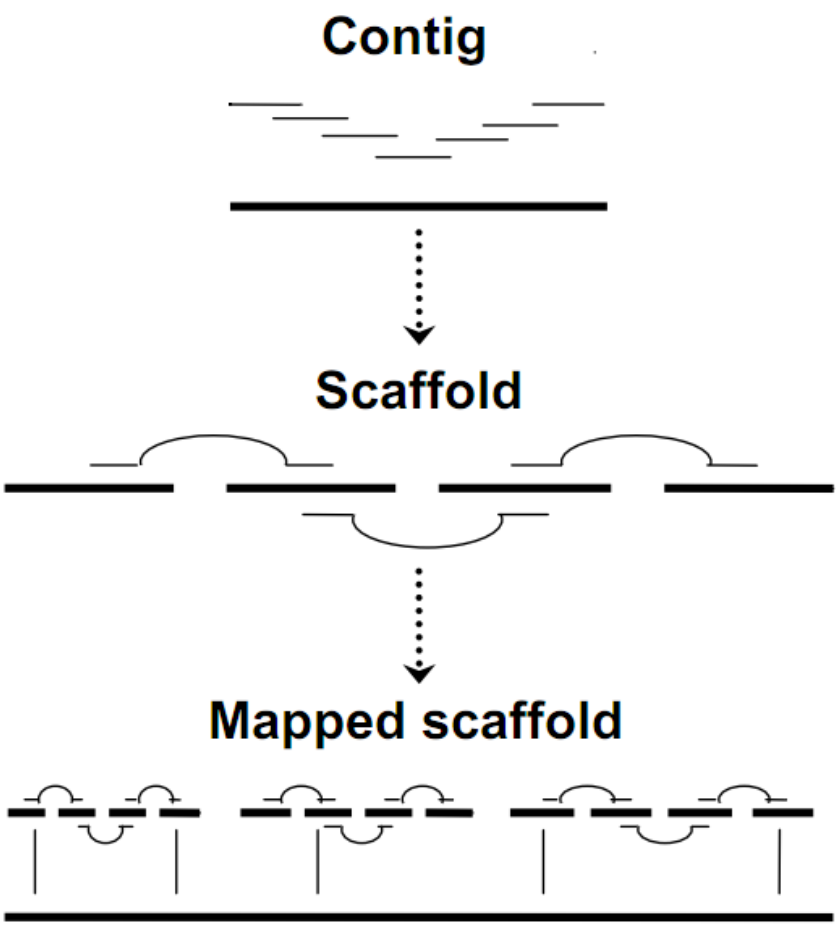
- Distance of insert depends of technology (454 or Illumina => 3-8 kb)

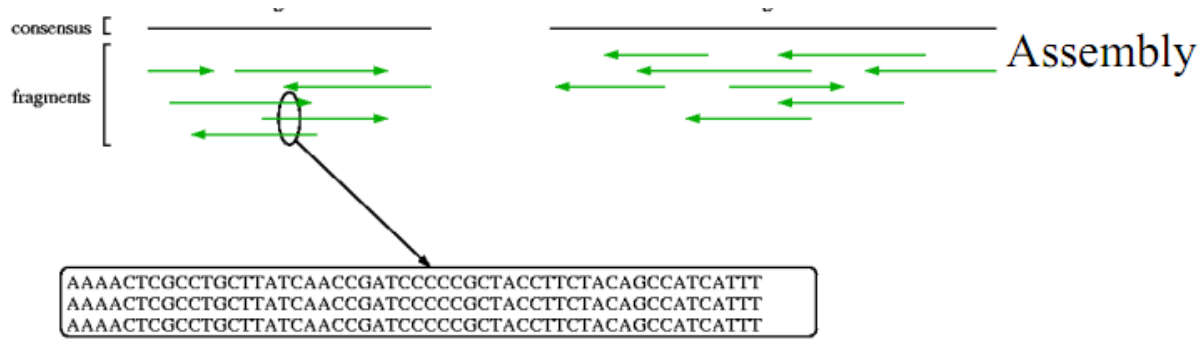


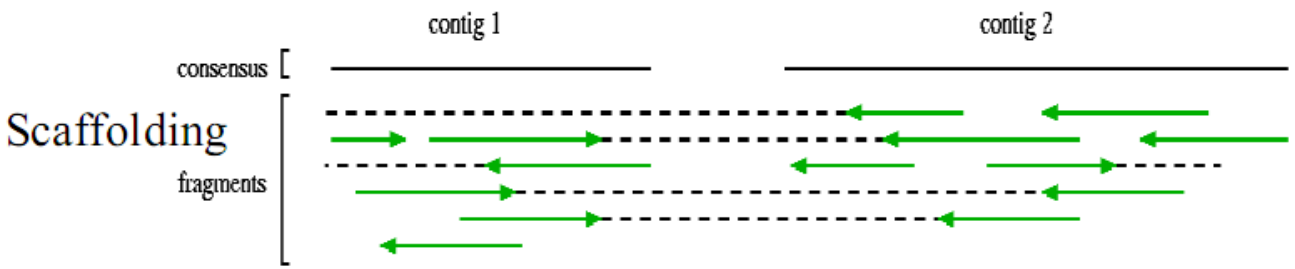
Contigs : group of overlapping reads, without gap

Scaffold : group of contigs order and in the same sens. Gap ("NNN") could existed and their length are known. Scaffolds exists only if a paired-ends (or mate pairs) sequencing was performed !

Mapped scaffolds : scaffolds mapped along a reference. Order, orientation and length of gaps are estimated, but not sure !







Finishing :

Mapping of reads along the assembled genome (or/and a reference) :

- help to correct the low quality/coverage areas

- Check the order of contigs
- Check the redundancy of contigs (false contigs or true repeat contigs like rRNA operons)

- Fill the gaps by extending the boundaries of each gap using ends of mapping reads
- Order (or reorder) contigs
- Desassemble some areas if they seem to be false