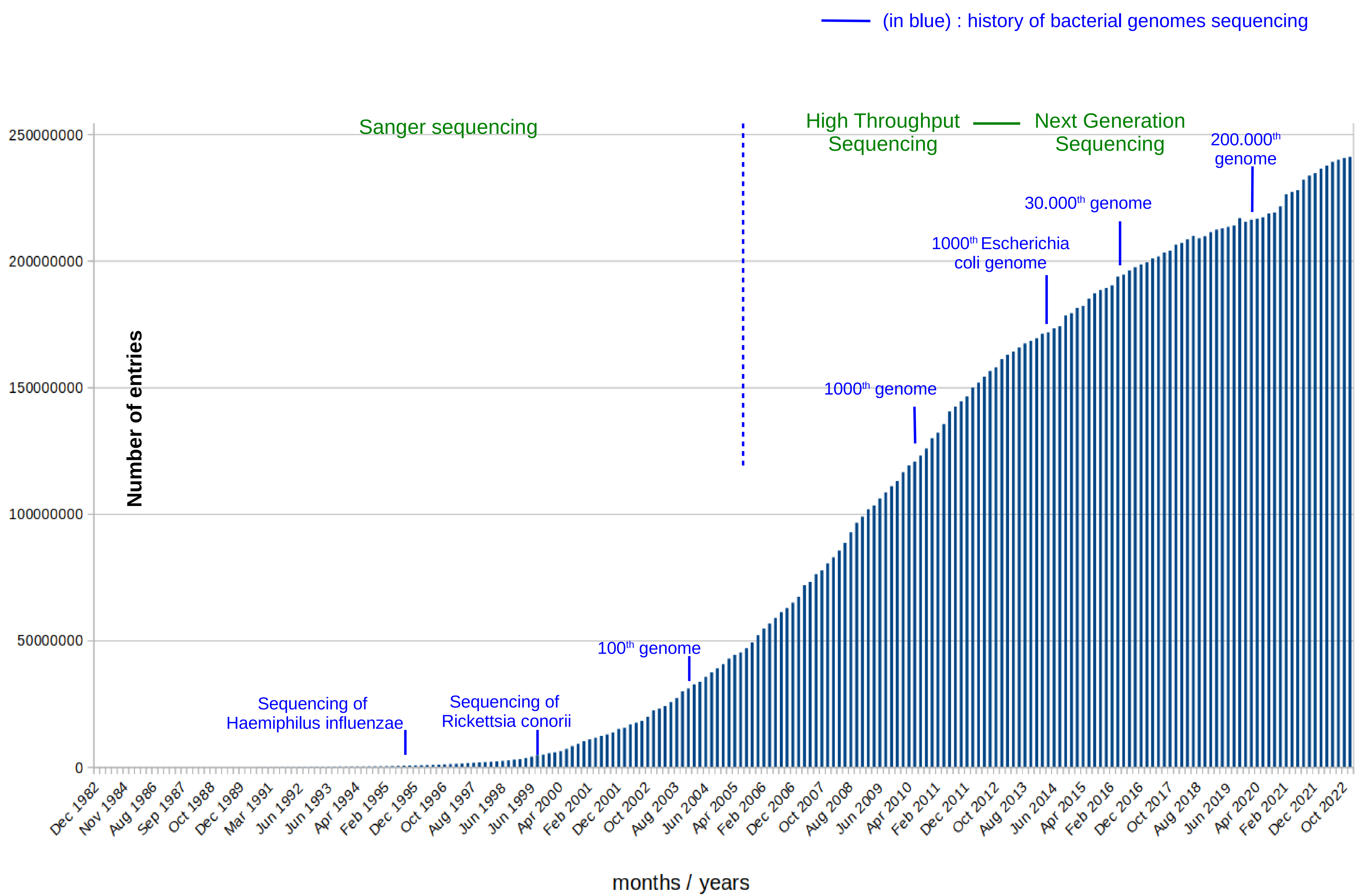
NGS and Genomes assemblies

Bioinformatics teachings

http://bioinfomed.fr - Olivier Croce (croce@unice.fr)

Data release



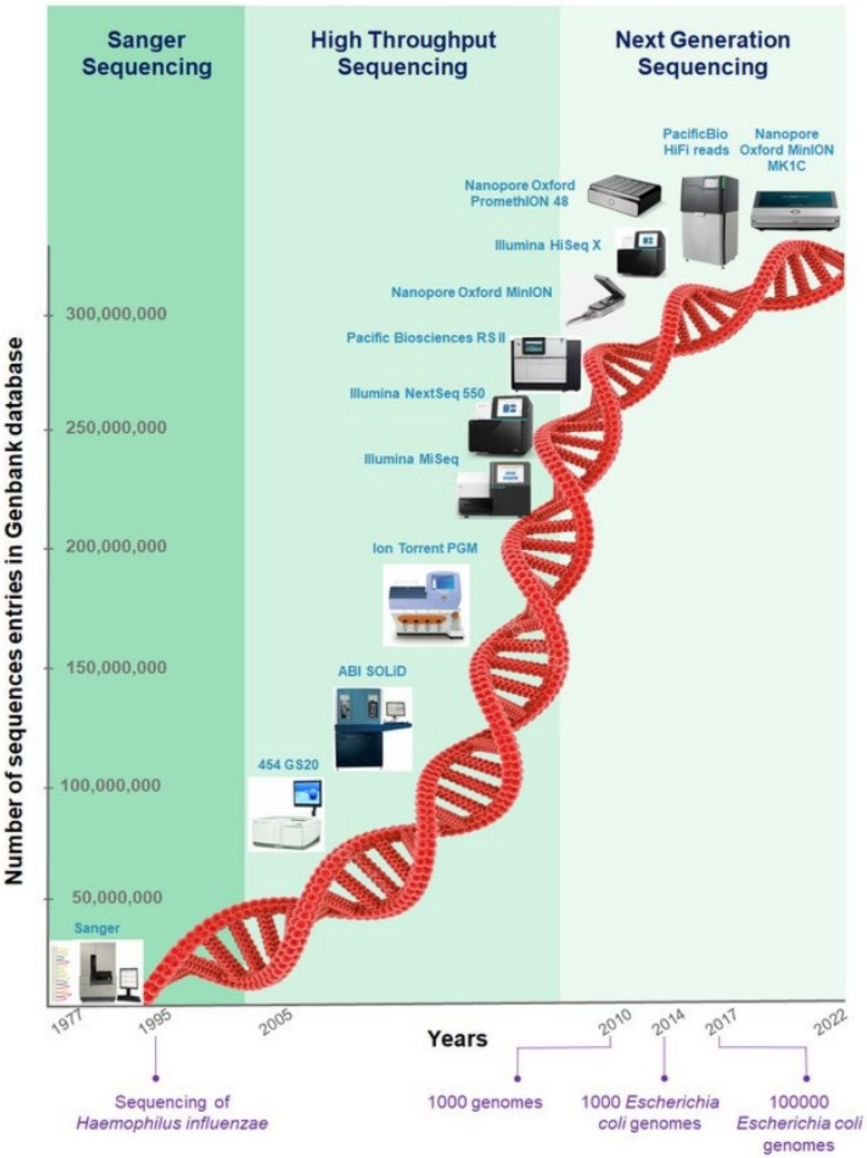
Data release

Application and Challenge of 3rd Generation Sequencing for Clinical Bacterial Studies

Mariem Ben Khedher, Kais Ghedira, Jean-Marc Rolain, Raymond Ruimy, Olivier Croce

International Journal of Molecular Sciences

(Int. J. Mol. Sci. 2022, 23(3), 1395; https://doi.org/10.3390/ijms23031395)



Abstract

Over the past 25 years, the powerful combination of genome sequencing and bioinformatics analysis has played a crucial role in interpreting information encoded in bacterial genomes. High-throughput sequencing technologies have paved the way towards understanding an increasingly wide range of biological questions. This revolution has enabled advances in areas ranging from genome composition to how proteins interact with nucleic acids. This has created unprecedented opportunities through the integration of genomic data into clinics for the diagnosis of genetic traits associated with disease. Since then, these technologies have continued to evolve, and recently, long-read sequencing has overcome previous limitations in terms of accuracy, thus expanding its applications in genomics, transcriptomics and metagenomics. In this review, we describe a brief history of the bacterial genome sequencing revolution and its application in public health and molecular epidemiology. We present a chronology that encompasses the various technological developments: whole-genome shotgun sequencing, high-throughput sequencing, long-read sequencing. We mainly discuss the application of next-generation sequencing to decipher bacterial genomes. Secondly, we highlight how long-read sequencing technologies go beyond the limitations of traditional short-read sequencing. We intend to provide a description of the guiding principles of the 3rd generation sequencing applications and ongoing improvements in the field of microbial medical research

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>Contain</u>:

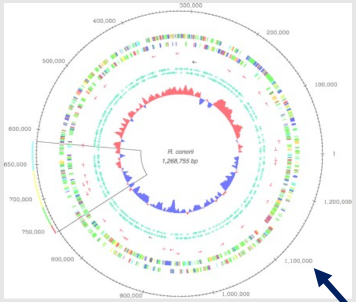
- Sequences of DNA or RNA from various sequencers technologies and from many labs

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

- Annotations

Aims

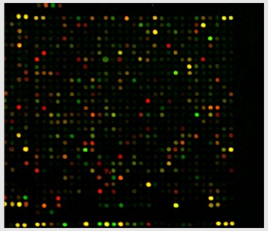
Molecular detection and identification



Comparative genomics

```
TAAAGTTTTTAAGCCATTTTTTAAAGCTTTACA
TAAAGTTTTTAAGCCCTCTTTTAAAGCATTGCA
TAAAGTTTTTAAGCCATCTTTTAGAGGTTTACA
TAAAGTTTTTAAGCCATCTTTTAAAGCTTCACA
```

Identification of specific signatures, genotyping

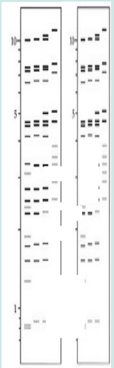


DNA microarrays or primers & probes design

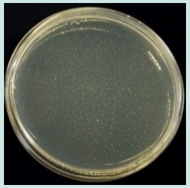
Identification of antigenic epitopes (surfaceome)

```
ATGTTCACAAGGACTATCATGAACATCGGCTATGTTGATGATGTACAACCTTTAAAACAGGGAGTACGTT
TAAATTTTTCTACGCGCTATGACATACAGAGTTTGGAAATTGGTGCATCGATTGCGTGTTCATGGATTTG
TCTGACAATTGTCGAGCGGGGGGTAAAACAAGCGACTGCTGGTTGGTTTGCTGTAGAAGCATGGGAAGAA
GCATTGCGTTTGACTAATCTTGCACAATGGACAAAAGGAACTTTTGTTAATTTGGAACGTTCGCTTCGAT
TAGGTGATGAAATAGGAGGACATTTGGTTTCCGGTCATATTGATGGTTTGGCTGAAATCATTGATCAAAA
AAATGAAGGGGATGCAATTCGTTTTTATTTAAAAGTTGTAAGACAATTTATGCCTTTCATTGTCAATAAG
GGATCTATTGCACTTAATGGGACATCTTTGACTGTTAATGGTGTTGAGGATTGTGTTTTTGATGTTCTTA
TTATTCGCCATACACTCGAAATGACAACGTGGGGACAAGCTAGAATTGGAGATTGGGTCAATTTGGAAAT
TGATCAACTTGCTCGTTATGCTGCGAAACTTTTTGCTTTAAAAAGAGAAGATGAATAA
```

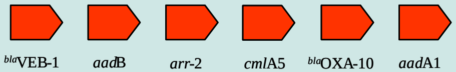
Phenotype prediction



Prediction of restriction profiles

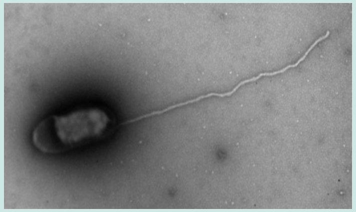


Culture medium design (metabolome)



Prediction of antibiotic resistance and antibiotic design (resistome)

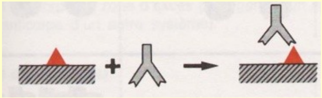
Virulence prediction



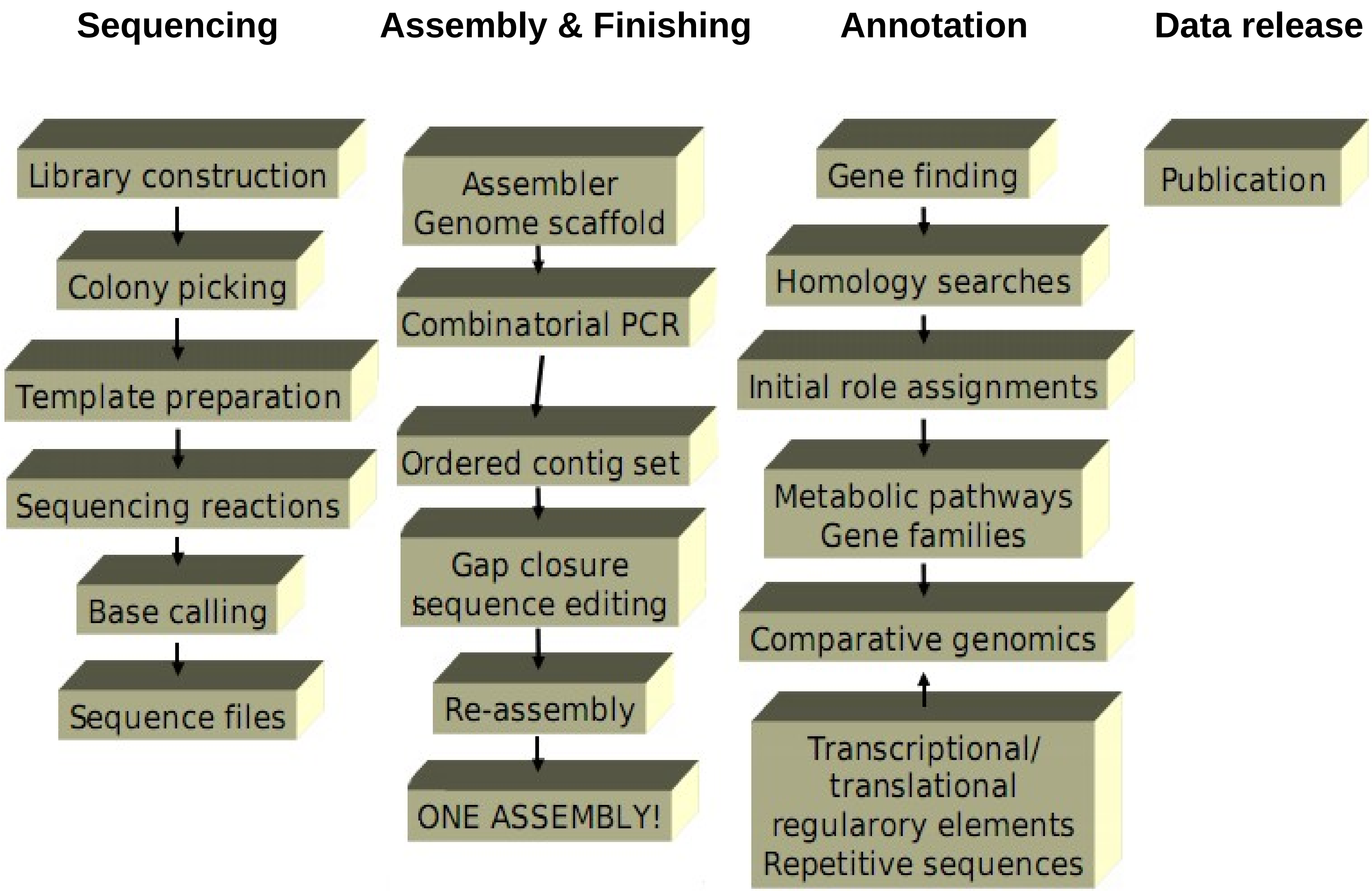
Vaccine & serological assay development

Vaccine design

Development of monoclonal antibodies

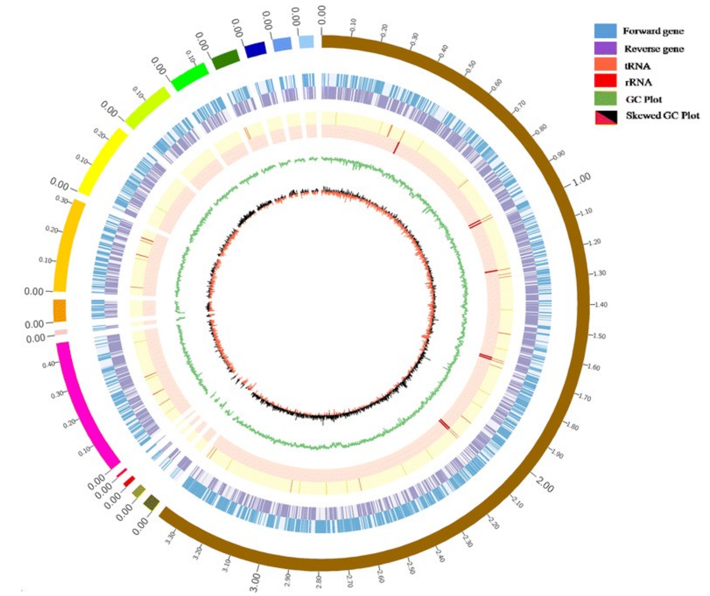


From the bench to the publication



Quality level of genomes

- The genome sequence must be completed with a high quality and annotated 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
=> now easier using long reads sequencing

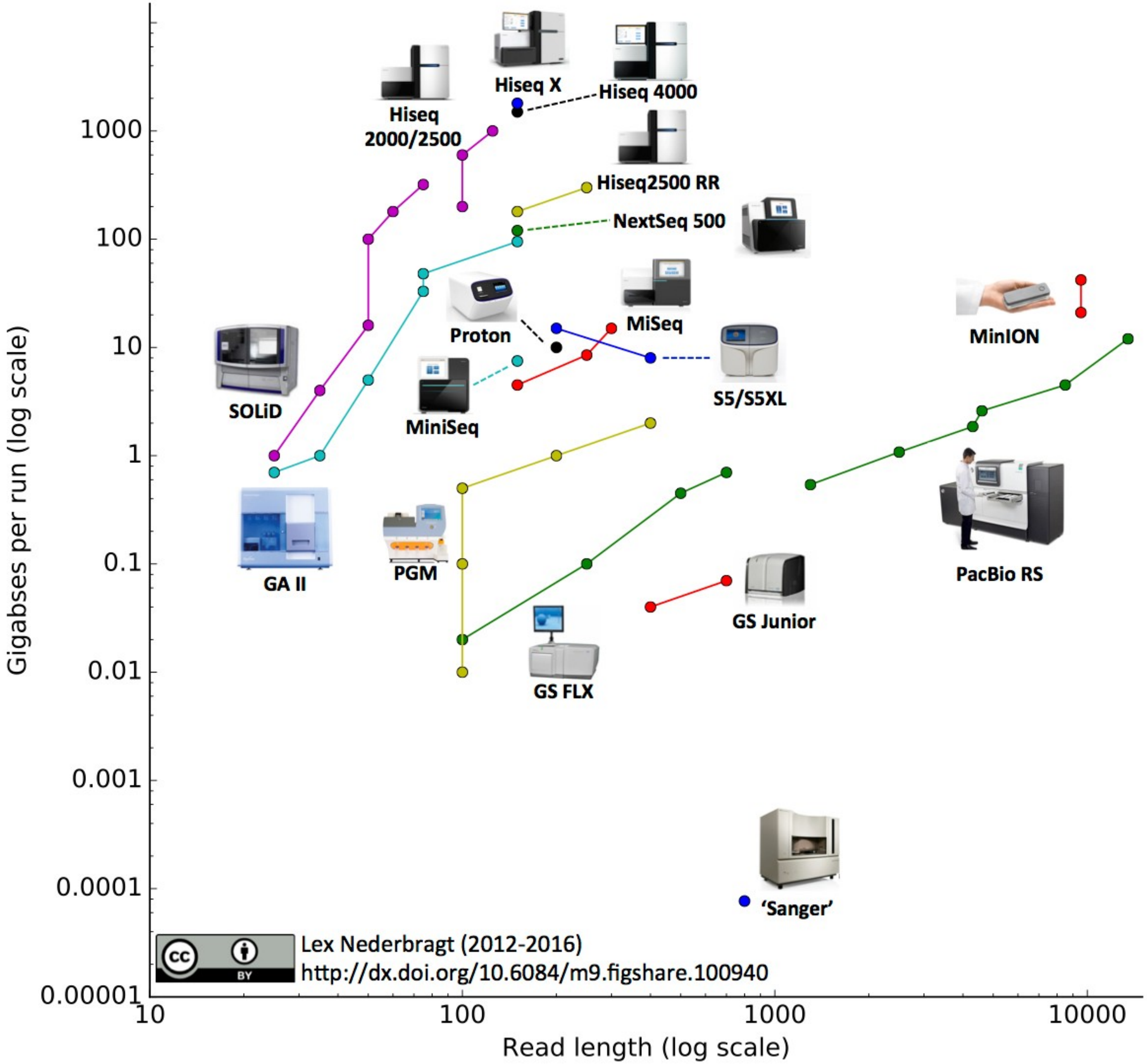
- The sequence should be uncompleted with a draft quality, whether we suppose most of the genes are sequenced (and identified)

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, part of clinical) usually performs draft genomes

=> depending of the project : time and experience (bioinformatician), money (coverage of NGS), organisms, the question to answer

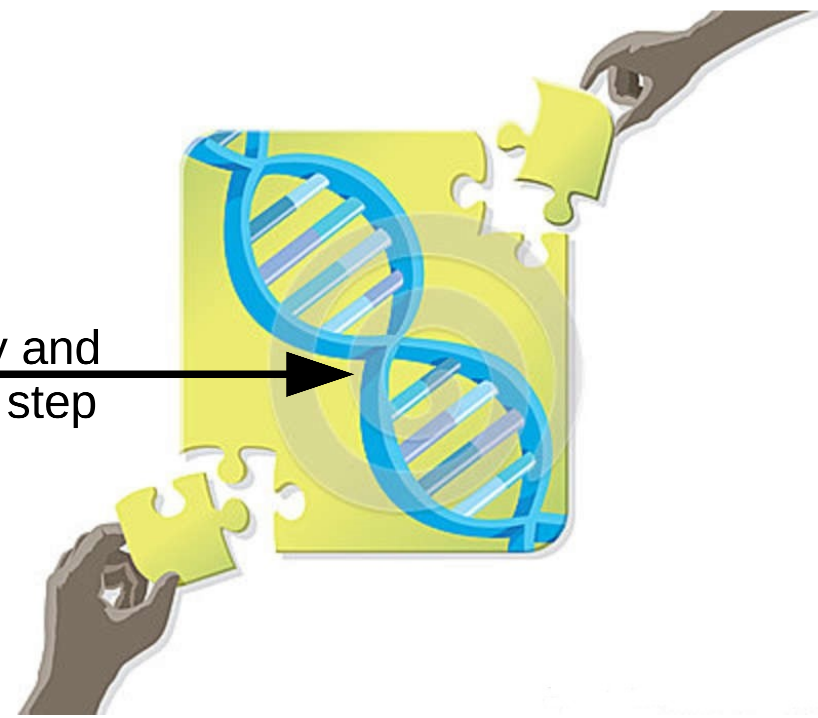
Sequencing technologies



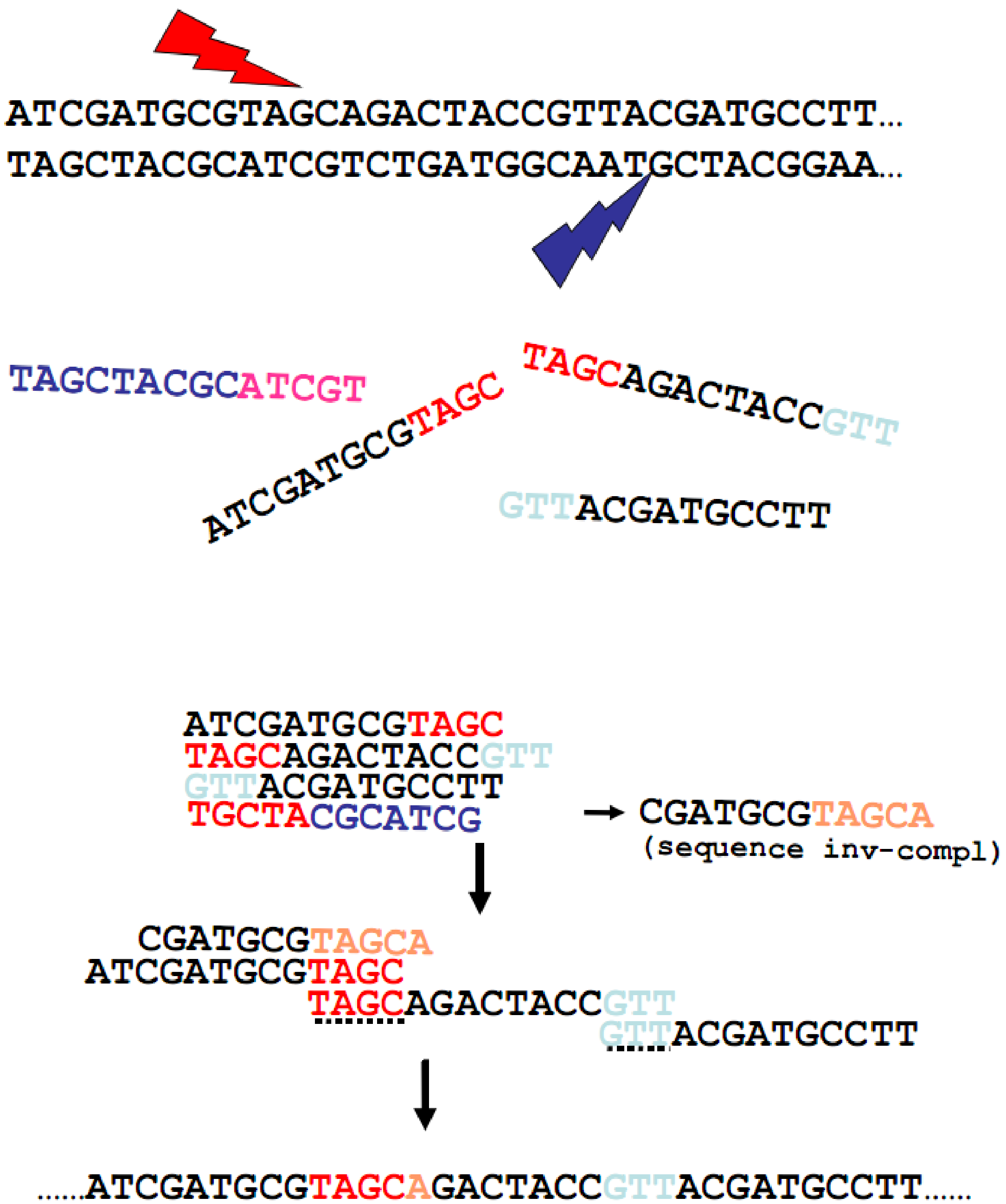
Principle of sequencing and assembly

Sequencing step: reads have heterogeneous distribution

Assembly and finishing step



Principle of sequencing and assembly



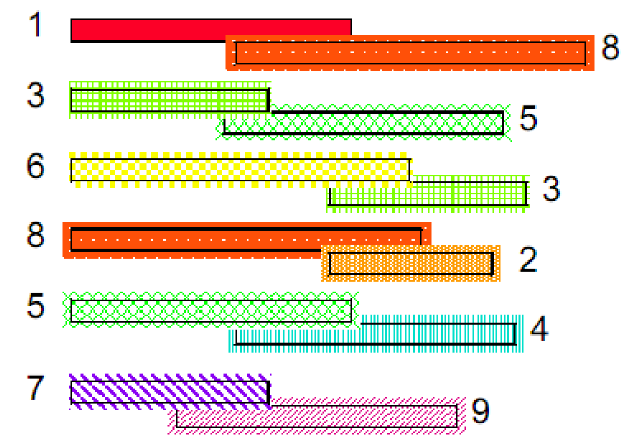
Fragmentation + sequencing
=> sets of reads

Build of contigs with overlapping regions

Assembly :
=> alignments of reads + consensus

Principle of sequencing and assembly

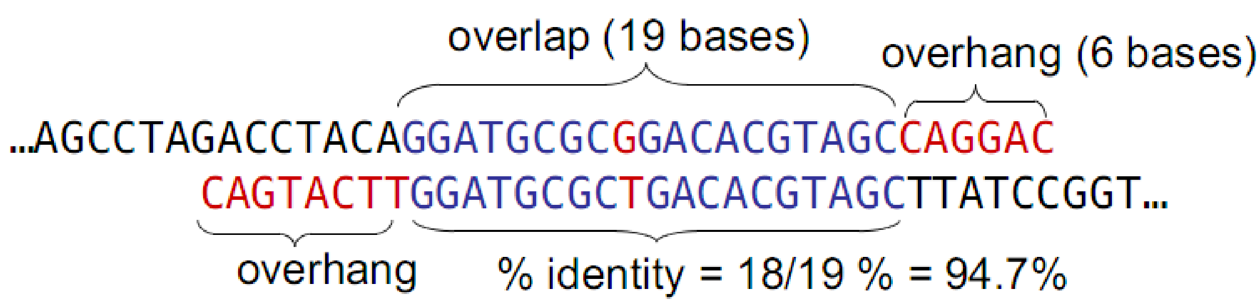
Search for best pairings by comparing each sequence (and its reverse complement) against every others sequences to find the best overlapping



=> 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



=> Many **assemblers** tools existed (depending of sequencing technologies, libraries, genomes size, etc..)

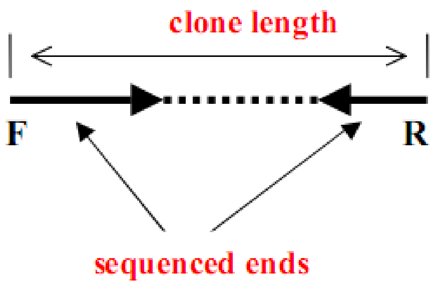
Constructions of library from genomes fragments

*** single - end (= shotgun)** : reads sequenced independently

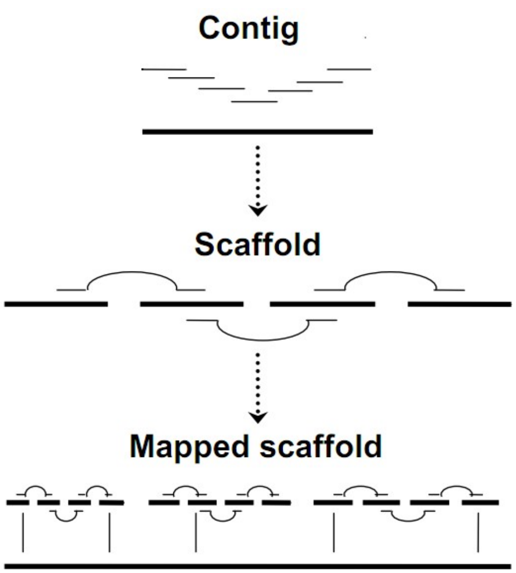
*** paired-end** (similar to **mate-pair**) : reads are sequenced by pairs

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

- Distance of insert depends of technology (ie. Illumina ~150 nt for paired-end, ~1-5 kb mate-paired)



Why using PE/MP ? length of reads is limited => assemble repetitive regions by using reads as “anchors”

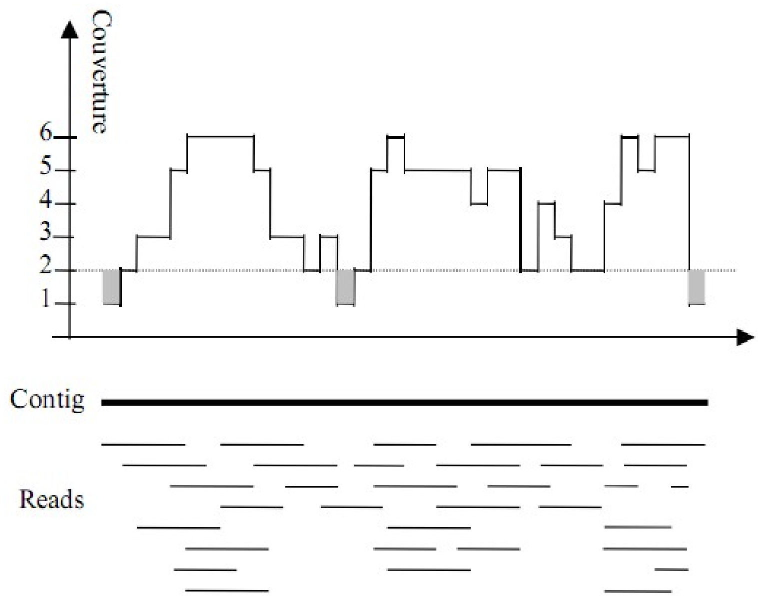


Contigs : group of overlapping reads, without gap

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 !

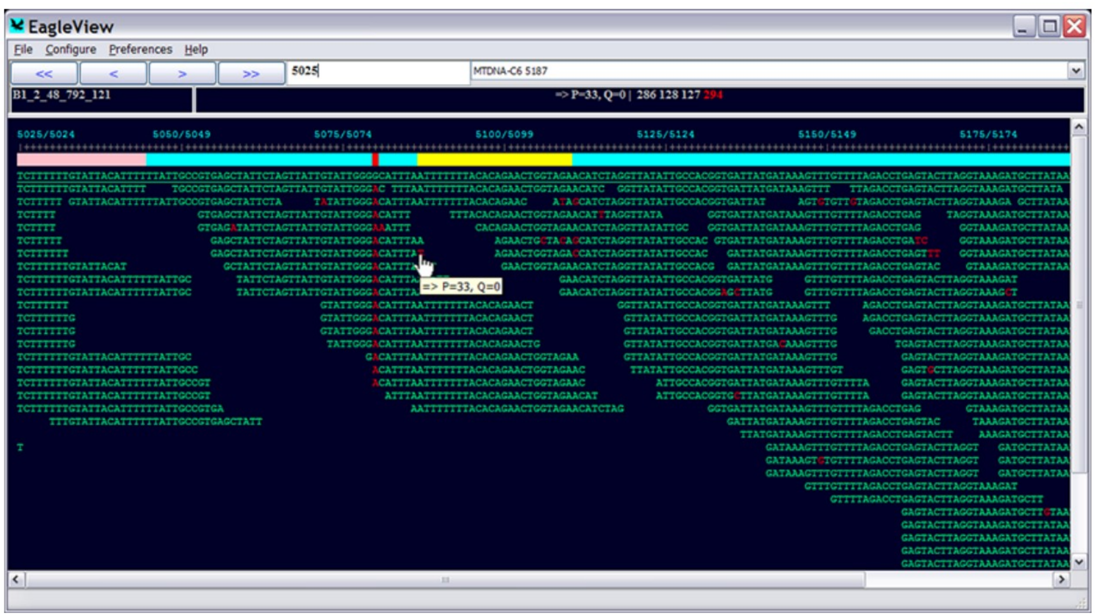
Mapped scaffolds : scaffolds mapped along a reference. Order, orientation and length of gaps are estimated, but not sure !

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



Re-Mapping, to see coverage, SNP and potential errors

Finishing

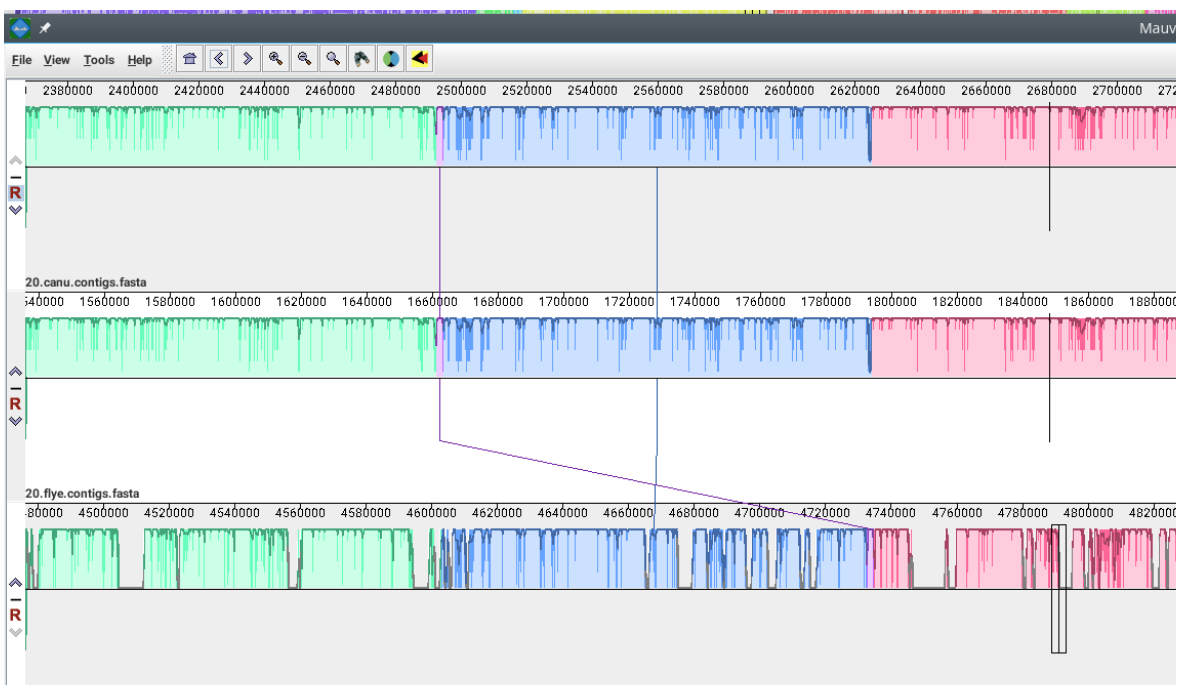
- (re)Mapping of reads along the assembled genome (or/and a reference)

- 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)

- Compare syntheny between multiple assemblers (global alignment)

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



Bacillus cereus assemblies using 3 assemblers tools.
2 first genomes are very similar, the third show many differences

=> High improvement with new long-reads technology (MinIon Nanopore, PacBio)