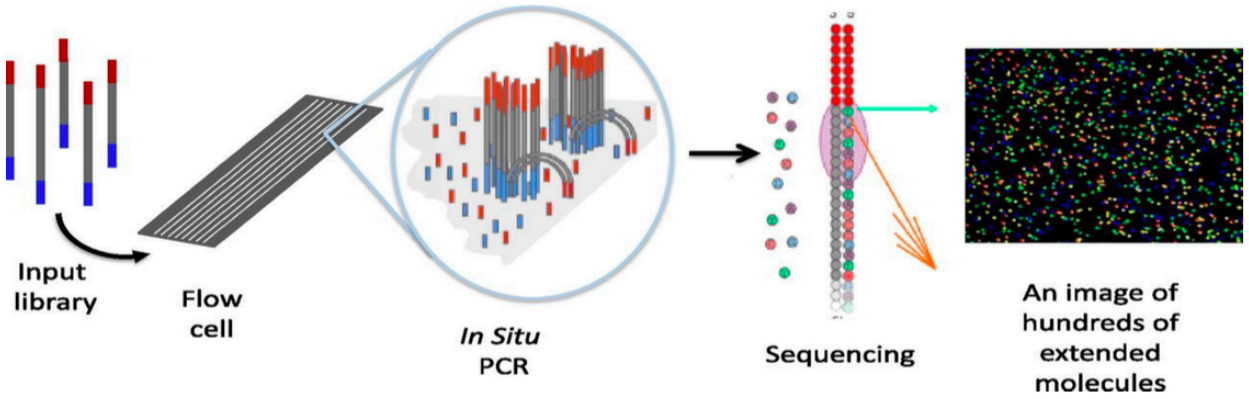
Notions de bases sur l'analyse des génomes I

(assemblage, mapping, comparaison de souches)

Théorie #3

Dr A. Jamet
Médecin Microbiologiste @Necker



FastQ
Reads sequences + quality

READ

Identifier **Sequence** **Quality scores *(as ASCII chars)***

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
```


Chadi Saad

2 files : Forward (1), Reverse (2)

Reads1.fq

```
@ERR229776.100000840
CTAGGAAGCGTAGTCCTGGGGTCATCTCTCCTATTAATACTGTTGGGGAATGTTTAGTA
+
BAEEAGEED96EHFE@BF><>EAAC;EBH<K<6:HJGFFHBC>DDIKG4AIHFFD@0/=
@ERR229776.100020365
CATTATTTCATAGTAGCCAAAAAGTGGAAACAGTCAAAATATCCGTCAGTGAATTGACC
+
1.*/.,/&((&3=;B@F860C>@51(3:).6GG-68C*:CG)#B4/=HDJ6;79)<@C/
@ERR229776.100104918
TATTTCTGGAATTTTCCATTTAATATTTTCAGACTGCAGTTGACTGCGGGTAACTGAAA
+
CEEEEEFEDAEGGGFDHGFFHGIHHHIIIIGKHBKJJIGHFHKILJKLEJLJJIFJMJK
```

Reads2.fq

```
@ERR229776.100000840
TTCTGGTCAGTAAGACCTCAAAAGGTTAAATACTAGCGATTTACACACCTTAAATGATT
+
CFIEEG@FFFGKFJHJ>HHKLLJIIJILLJIILJHKAKJKKJJJJJJLMKJMKJJJJKJ
@ERR229776.100020365
CCTAAAATGGTGTGTTTTCGTATATTCACAATGCTGTGGAACCATCACCACTATCTGAT
+
4B@EDFF=(/CHBHEHCE6@ED8E@@I6HJB6E:6%@C46FFIBGCIGKD,DN=CBBE@
@ERR229776.100104918
TCTTTCTTTTGTTTTTTTTTCTGAGATGTCTTTTGTTTTTGTTCTGAGGTCTTGTTATG
+
CFIGGGKHHHFHHFIJIIIJKLIIHJIIIKLJKKIJKLLKJFJJMHJJLFJMJIKKJJJ
```

Chadi Saad

Goal: read cleaning

RAW

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPPRJQRQQQQQRPQPRQQPFQSQQPRLIMHKSNRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJ#CFDGCCCB
@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBE#HCHG8EHFDCD6<329@<:69A<6,;<967>;=C:>AA8BBED########################
@SRR062635.15516129
AAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGGGCCCCCCTTTCCCCCCCGGGGGGGGGACAGGGGGGGTGTTCGGGCCCCGCGCCGCCCTTGACCACGG
+
EKLMPPPPPQQQQQQQQQQQQQQQH#################################################################################
```

Clean

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPPRJQRQQQQQRPQPRQQPFQSQQPRLIMHKSNRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJ#CFDGCCCB
@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAG
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBE#HCHG8EHFDCD
```

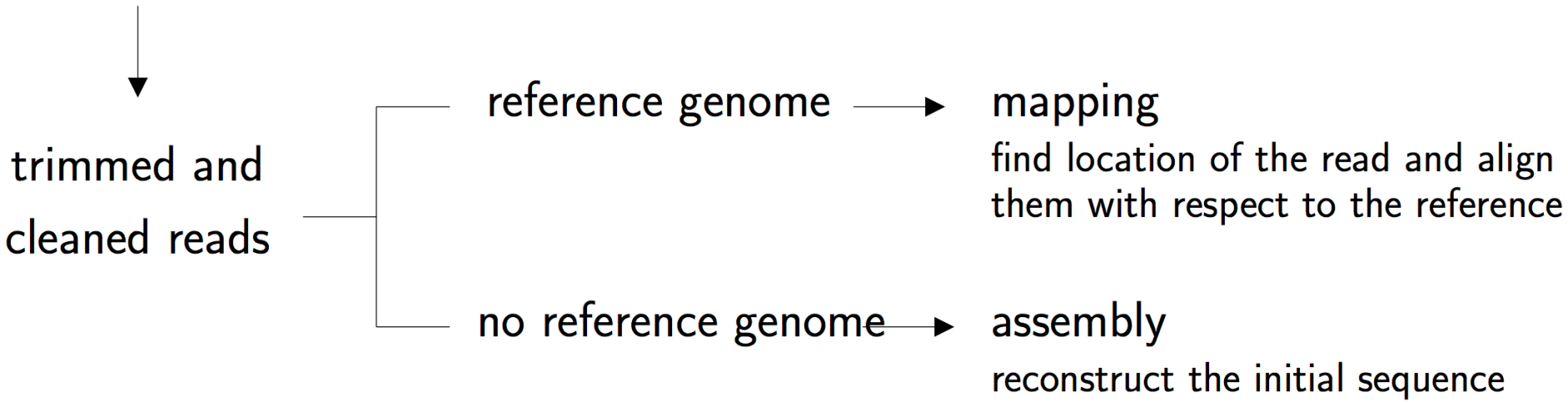
A. Jamet

Chadi Saad

WGS = 2 possibilités

Fragmented DNA is sequenced into “reads”

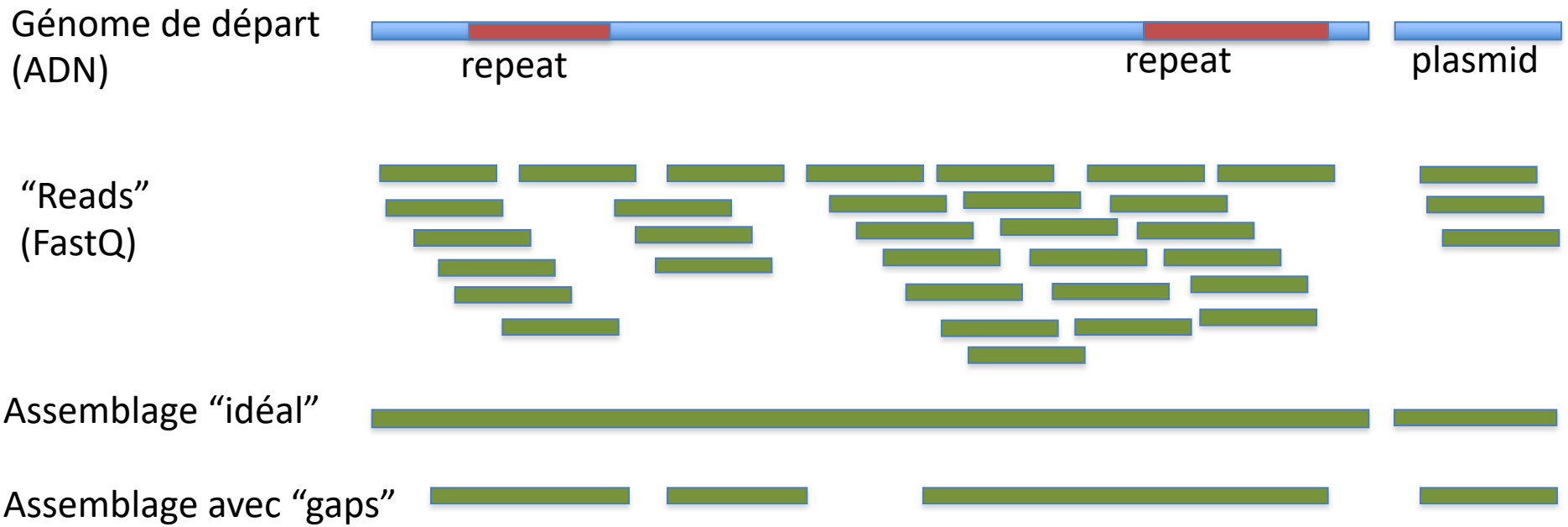
raw reads



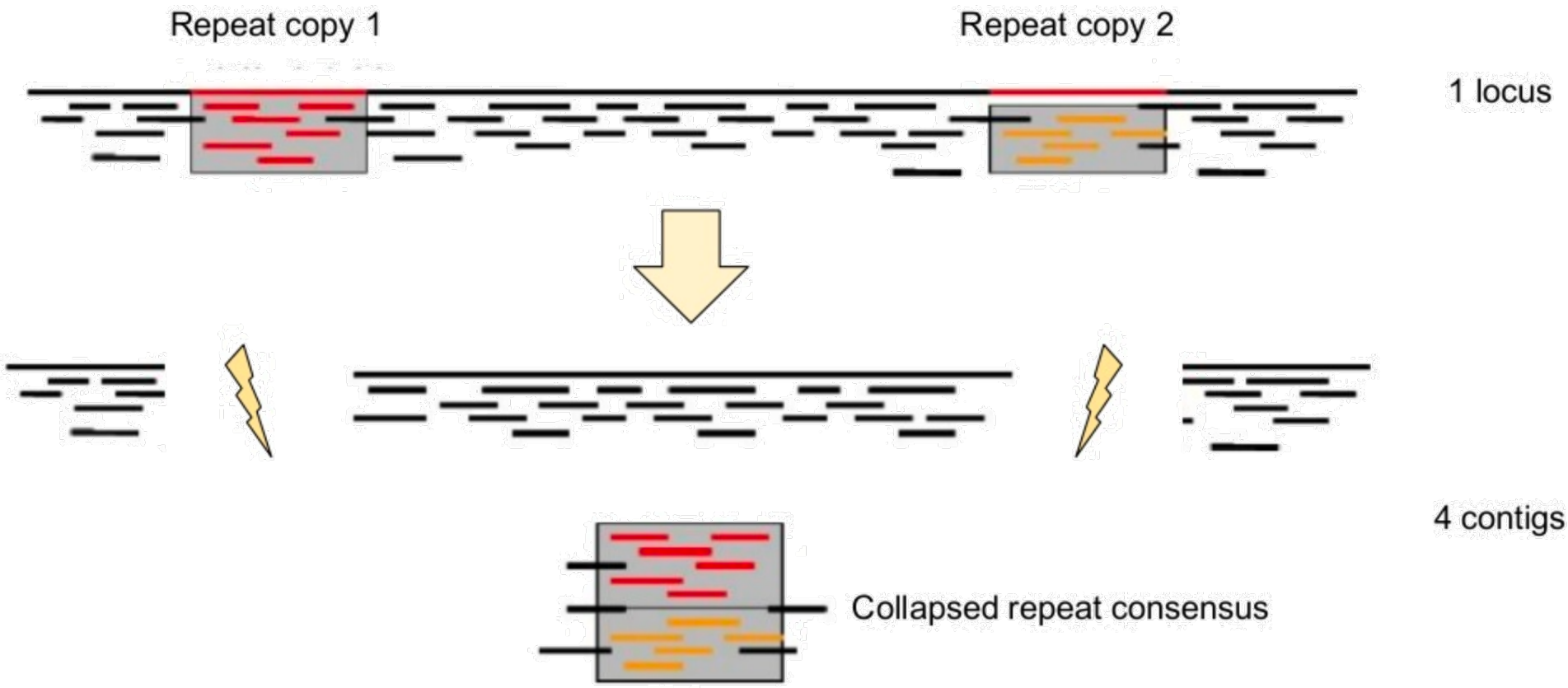
De novo assembly

De novo assembly of the reads++ in microbiology

Without any references



The problem with repeats

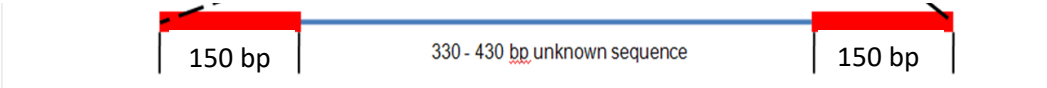


Handle unresolvable repeats by leaving them out
This breaks the assembly into fragments
Fragments called contigs (short for “contiguous”)

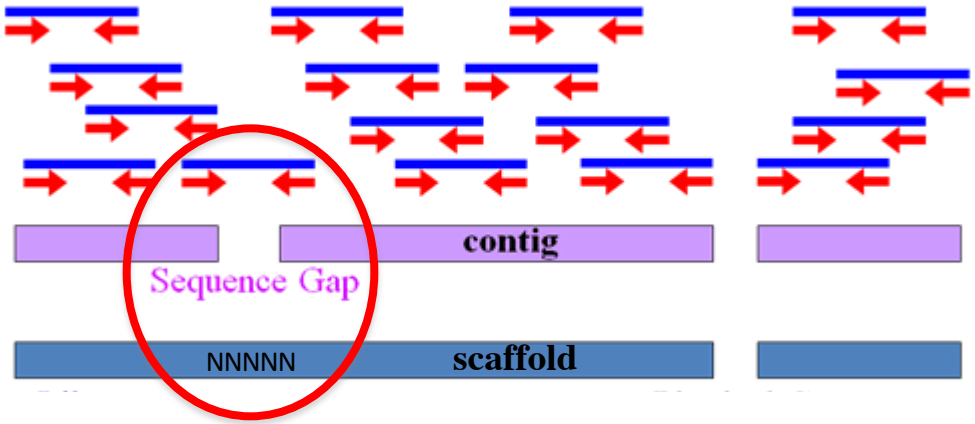
Slide credit: T. Seemann
langmead

Paired-reads

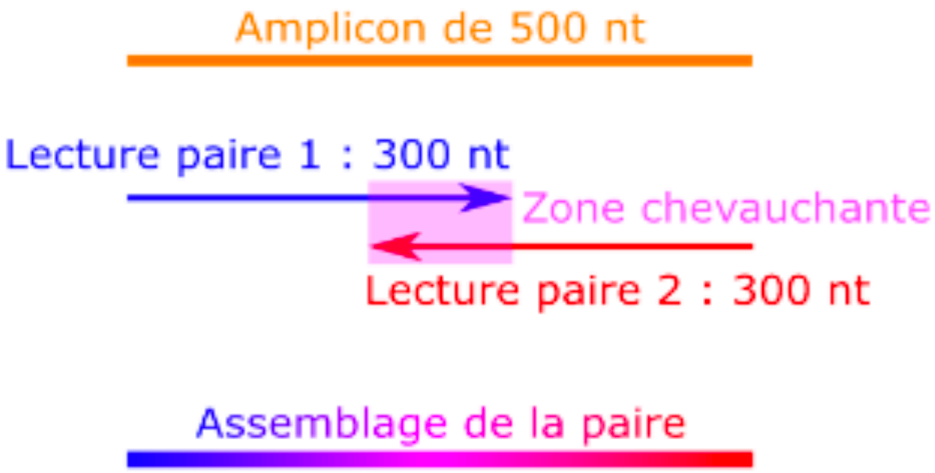
- Illumina reads are organized into read pairs



- Read pairs can be puzzled together into “contigs”
- Contigs can be ordered according to their expected position into “scaffolds”

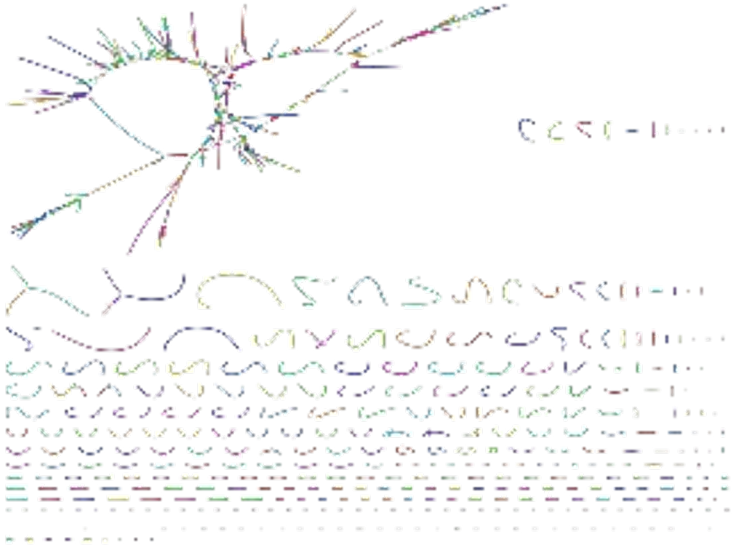


Illumina MiSeq

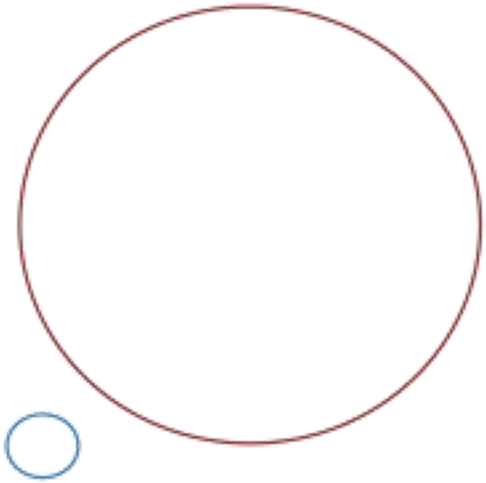


Séquençage Illumina paired-end :
Le séquençage de la cible est effectué dans les deux sens, générant deux reads par amplicon qui se chevauchent, permettant de les assembler en un reads plus long sur la base de cette zone chevauchante.

The effect of read length



250 bp - Illumina - $100



8000 bp - Pacbio - $1000

Or Nanopore

Each of our sequenced genomes is a “draft”

Slide credit: T. Seemann

Quality check assembly

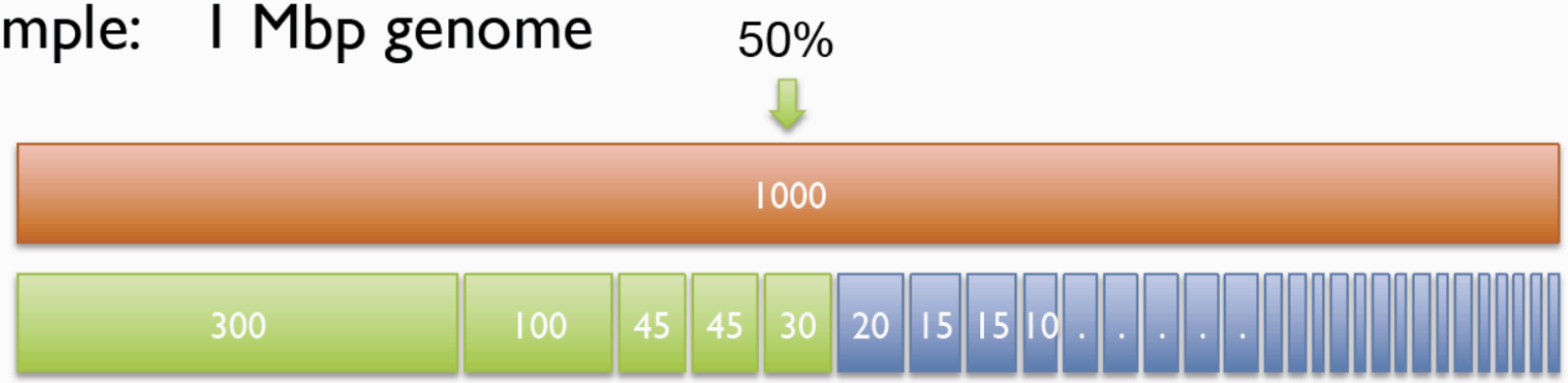
QUAST http://quast.bioinf.spbau.ru/

- **Number of contigs (< 250)**
- **Size of assembled genome +++++**
  - Chromosome + plasmides
- **N50** (> 15 kb is usually ok)
  - the *N50* length is defined as the shortest sequence length at 50% of the genome
- **Perform several assemblies and compare them ++**

N50 size

Def: 50% of the genome is in contigs larger than N50

Example: 1 Mbp genome



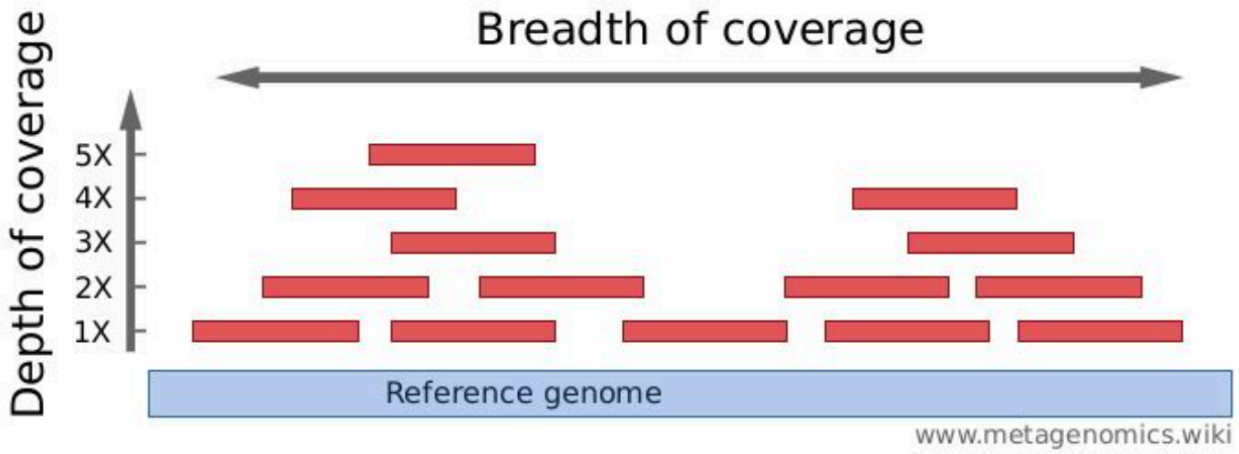
N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Definitions

- **Depth of coverage =** average number of reads covering a base (X) **(Profondeur)**
  - Example: 30X for normal sample, 100X for tumor sample

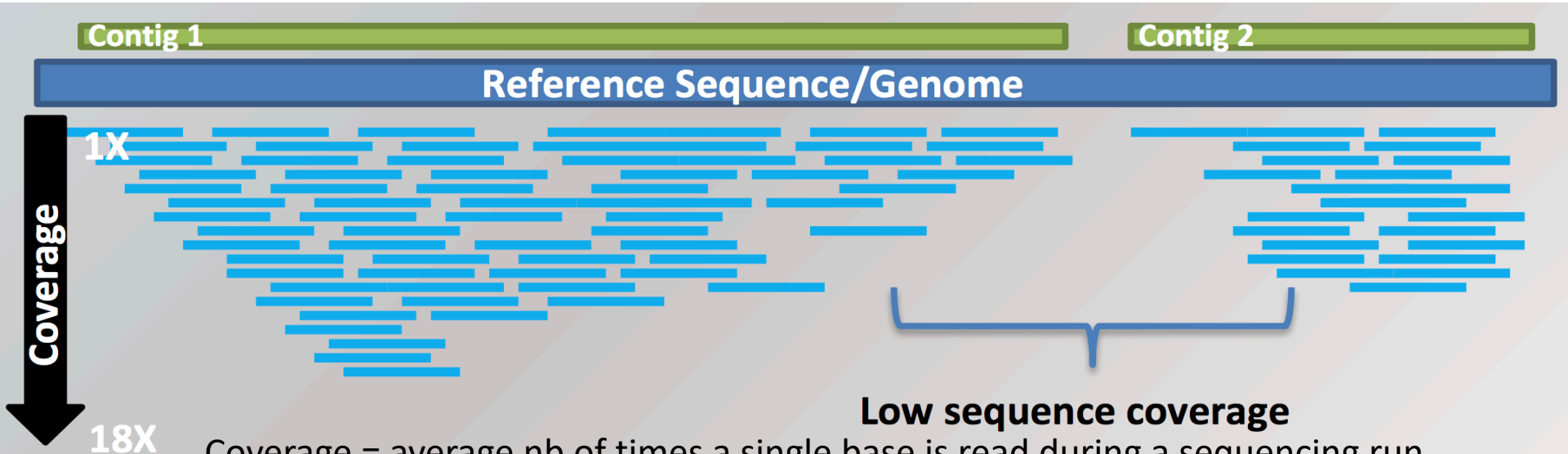
- **(Breadth of) Coverage =** percentage of the targeted regions covered by at least X read
  - For example: 90% of a genome is covered at 1X depth; and still 40% is covered at 4X depth.



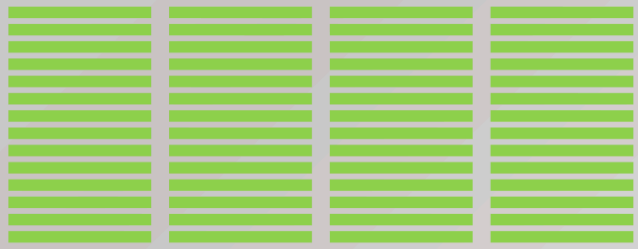
Mapping

Mapping of the reads

- **Comparison** to a reference genome (or a database of alleles)
- **Variant calling**
  - Identification of the differences (mutations, indels…)



Coverage = average nb of times a single base is read during a sequencing run



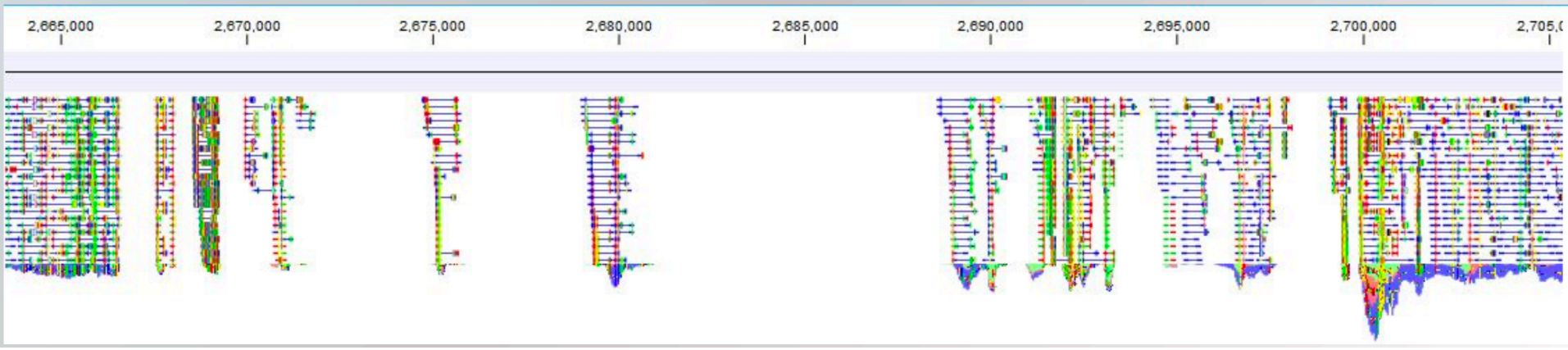
UNMAPPED READS

1. **Sequences not present in the reference.**
2. **Plasmids or other extrachromosomal.**
3. **DNA Structural Variation/Rearrangement**

Warning

- Reads that correspond to regions not found in the reference won't be mapped /!\
  - Incomplete genome reference = make a good choice !
  - Insertions in the sample relative to the reference

Selection of an Appropriate Reference



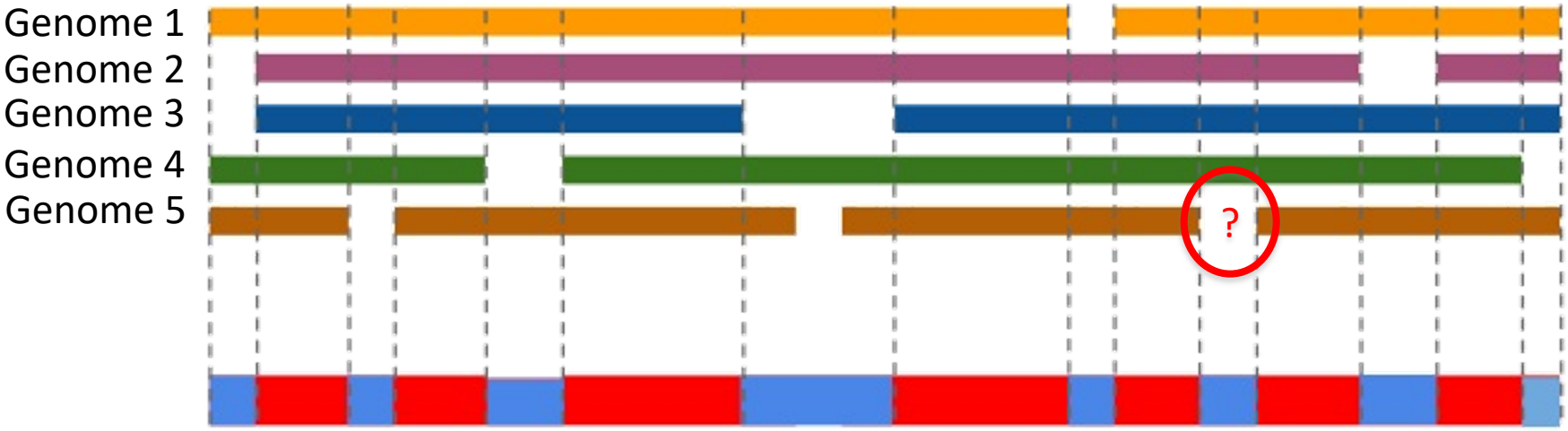
Definitions

- **Core** genome: genes shared by **all** members of a predefined group of bacteria
  - orthologues
  - diverge by accumulation of nt polymorphisms = sequence differences in **alleles**

- **Accessory** genome: genes present in a **subset** of a predefined group of bacteria
  - Gain or less of whole genes = **gene content**

- **Pan**-genome: the sum of **core and accessory** genes

The pan genome



Pan = Core + Accessory

Accessory genes are problematic because it is difficult to know if a gene is absent from the strain or just "missing" from the assembly

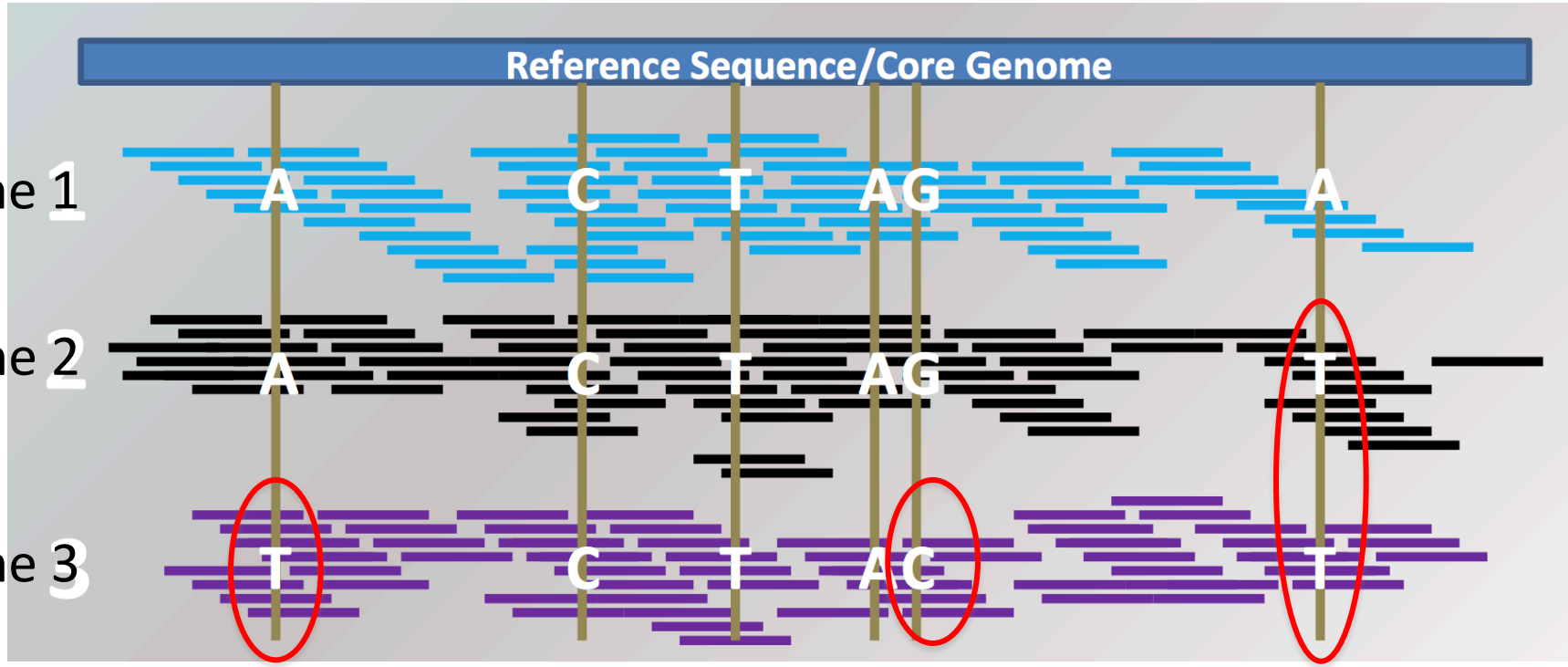
Slide credit: T. Seemann

Comparaison de souches

1- Comparer après assemblage des génomes (ex: cgMLST, variants...)

2- Comparer sans assembler par **mapping contre une reference commune**

- Identifier les SNPs dans les cores gènes par mapping



Mutation rate

Mutation rates are modulated by varying levels of efficacy of the DNA mismatch repair systems

Pathogen	Mutations per site per year	Mutations per genome per year
Staphylococcus aureus	3.0×10^{-6}	8.4
Clostridium difficile	5.3×10^{-7}	2.3
Mycobacterium tuberculosis	1.1×10^{-7}	0.5
Streptococcus pneumoniae	1.6×10^{-6}	3.5
Helicobacter pylori	1.9×10^{-5}	30.4
Vibrio cholerae	8.3×10^{-7}	3.3
Escherichia coli	2.26×10^{-7}	1.1

Didelot 2012 Nat Rev G

Outbreak thresholds

Table 1
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

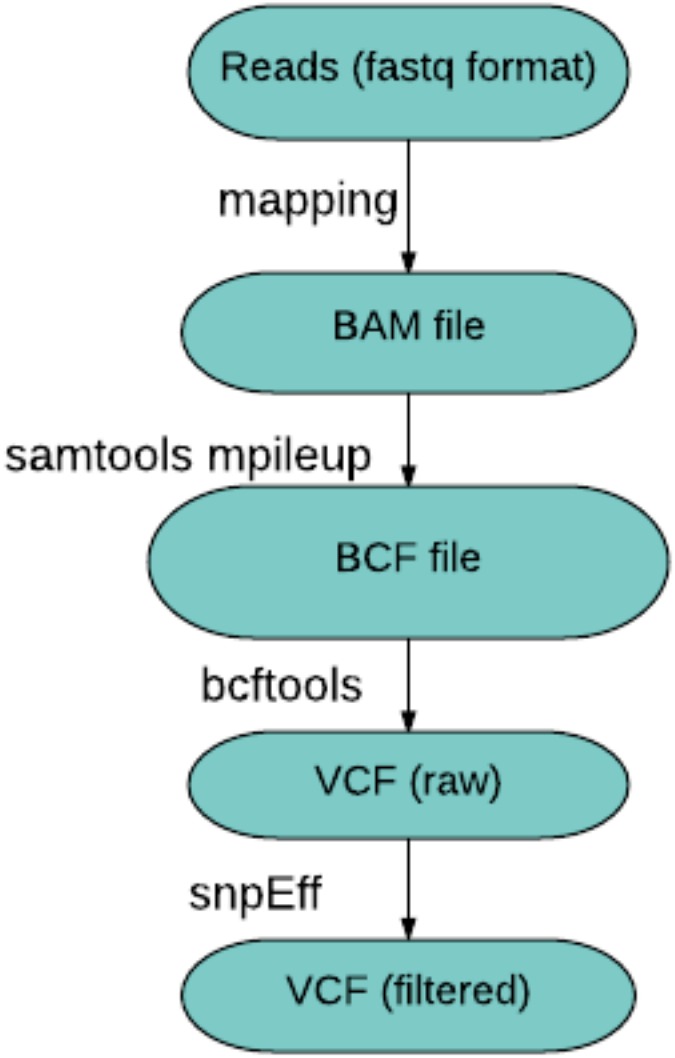
Organism	Relatedness threshold[a]		References
	wg/cgMLST (allele) SNPs		
Acinetobacter baumannii	≤8	≤3	[25,26]
Brucella spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Campylobacter coli, C. jejuni	≤14	≤15	[27,28]
Cronobacter spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Clostridium difficile	Epidemiologic validation in progress[b]	≤4	[29], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst
Enterococcus faecium	≤20	≤16	[30]
Enterococcus raffinosus	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Escherichia coli	≤10	≤10	[31,32], https://enterobase.warwick.ac.uk/
Francisella tularensis	≤1	≤2	[33,34]
Klebsiella oxytoca	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Klebsiella pneumonia	≤10	≤18	[35,36]
Legionella pneumophila	≤4	≤15	[37]
Listeria monocytogenes	≤10	≤3	[38,39]
Mycobacterium abscessus		≤30	[40]
Mycobacterium tuberculosis	≤12	≤12	[41]
Neisseria gonorrhoeae	Epidemiologic validation in progress[b]	≤14	[42], http://www.applied-maths.com/applications/wgmlst
Neisseria meningitidis	Epidemiologic validation in progress[b]		http://www.cgmlst.org/ncs
Pseudomonas aeruginosa	≤14	≤37	[31,43]
Salmonella dublin	Epidemiologic validation in progress[b]	≤13	[44], https://enterobase.warwick.ac.uk/
Salmonella enterica	Epidemiologic validation in progress[b]	≤4	[45], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/
Salmonella typhimurium	Epidemiologic validation in progress[b]	≤2	[46], https://enterobase.warwick.ac.uk/
Staphylococcus aureus	≤24	≤15	[47,48]
Streptococcus suis		≤21	[49]
Vibrio parahaemolyticus	≤10		[50]
Yersinia spp.	0		[51]

cg, core genome; MLST, multilocus sequence typing; SNP, single nucleotide polymorphism; wg, whole genome.

[a] Data often represent single studies that can be used to begin formulation of species-specific interpretation criteria. Thus, these data should be coupled with newly published similar studies to ensure that resulting values are not atypical and can be generally applied.

[b] Proposed wg/cgMLST schemes are available online (http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/) but as yet have not been epidemiologically validated.

The Snippy tool



Snippy is a pipeline encompassing several tools which :

- align reads,
- filter the good quality variants
- annotate them using an annotated reference genome

Types of variants

- Substitutions
  - single nucleotide polymorphism *(snp)* A→C
  - multiple nucleotide polymorphism *(mnp)* AG→TC
- Indels
  - insertion *(ins)* A→AC
  - deletion *(del)* ACCG→AG
- Complex
  - compound events AC→T

A. Jamet

T. Seemann

Visualiser un mapping

Aligned reads - sam

- **S**equence **A**lignment **M**atrix (sam) http://samtools.github.io/hts-specs/SAMv1.pdf (http://samtools.github.io/hts-specs/SAMv1.pdf)
- *Header* lines followed by tab-delimited lines
  - Header gives information about the alignment and references sequences used

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr10        LN:135534747
@SQ     SN:chr11        LN:135006516
```

```
HWI-ST1001:137:C12FPACXX:7:1115:14131:66670     0       chr1    12805   1
 42M4I5M *
0       0       TTGGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCACCAATATG
CCCFFFFFHHGHHJJJJJHJJJJJJJJJJJJJJJIJJJJJJJJJJJJIJJ
AS:i:-28        XN:i:0  XM:i:2  XO:i:1XG:i:4   NM:i:6  MD:Z:2C41C2     YT:Z:UU
NH:i:3
CC:Z:chr15      CP:i:102518319  XS:A:+  HI:i:0
```

Aligned reads bam

Exactly the same information as a sam file except that it is binary version of sam compressed around x4
Attempting to read will print garbage to the screen bam files can be indexed
Produces an index file with the same name as the bam file, but with .bai extension

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

Alignement

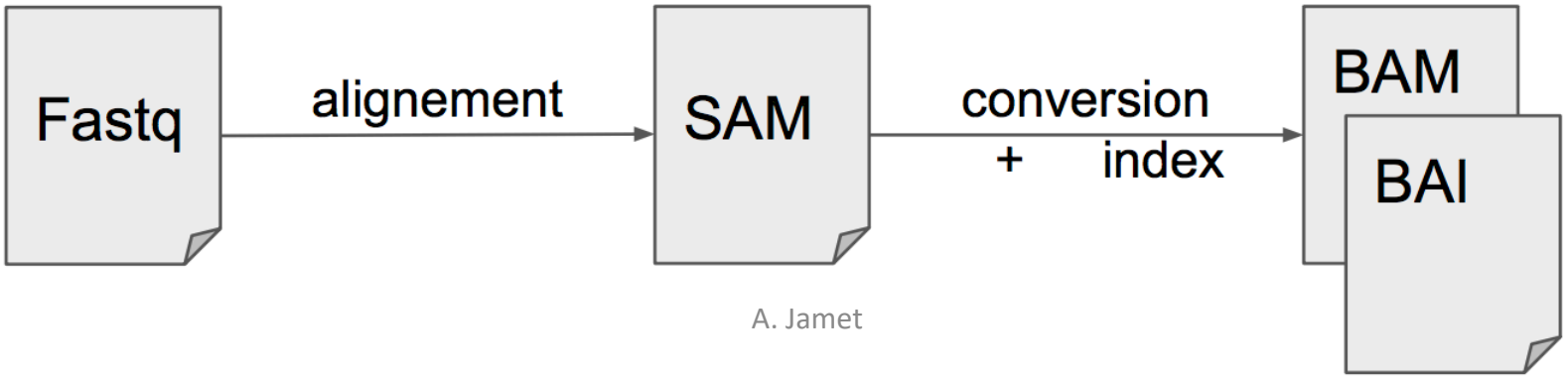
The corresponding SAM format is:

SAM file

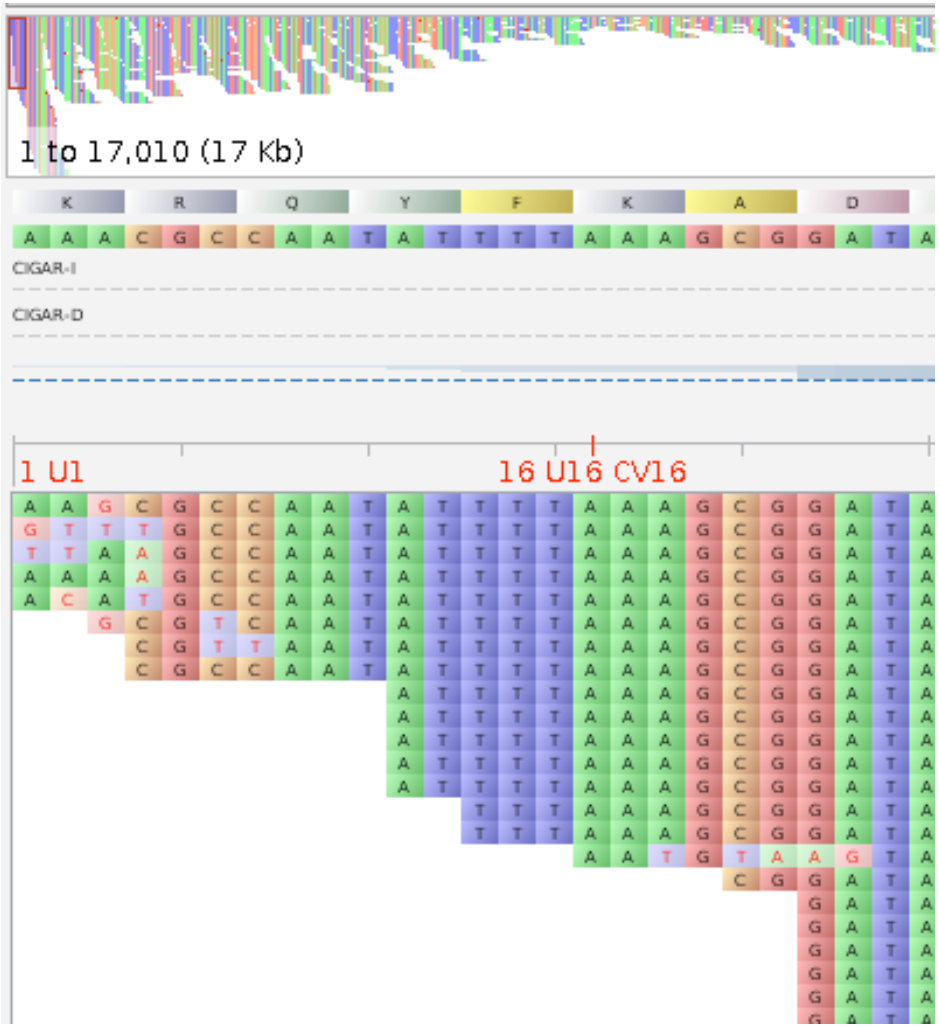
Header @SQ : Reference sequence
SN (sequence name) = ref, LN (sequence length) = 45

1 line per read, 11 columns

```
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```



Visualization with Tablet



Reference

Aligned reads