FRANCE GÉNOMIQUE

Institut Pasteur Genopole

“Finding variants in bacterial genomes: application to the analysis of epidemic or endemic clones”

Institut Pasteur

Objective

- Compare automatically a large number of genomes
- Minimize the number of false positive and false negative
- Minimize human curration of the data
- Visualization of the polymorphisms

Institut Pasteur

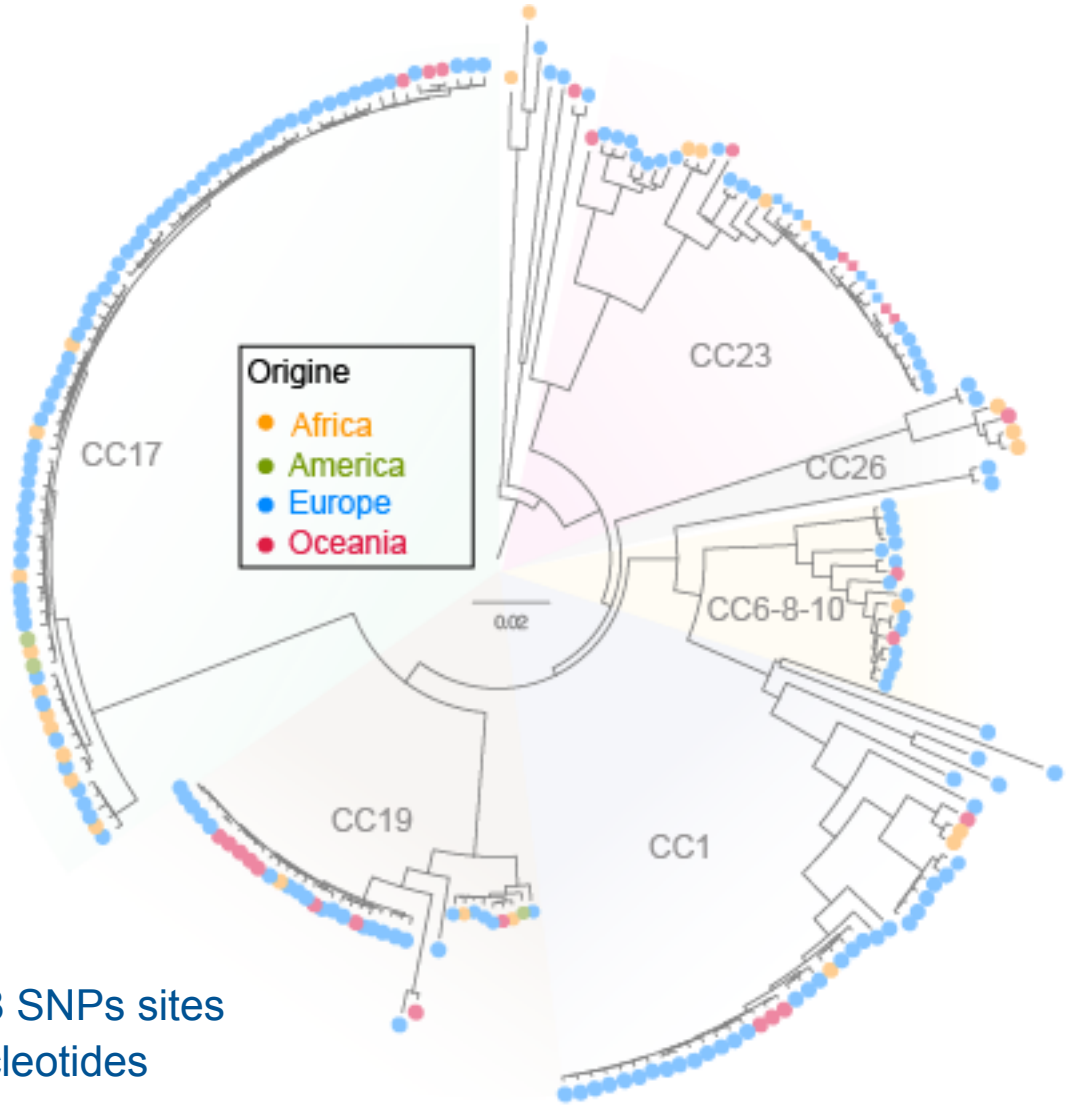
First generation pipeline – whole genome phylogeny

Objective: discovering the reason for the emergence of *Streptococcus agalactiae* neonatal infections in the 1960s

- **Sequencing of 230 isolates**
- **Use of different reference genome**
- **Filtering low quality reads**
- **Aligner: BWA**
- **SNA calling: SAMtools MPILEUP and varFilter**
- **Velvet assembly of the 230 genomes**

Institut Pasteur

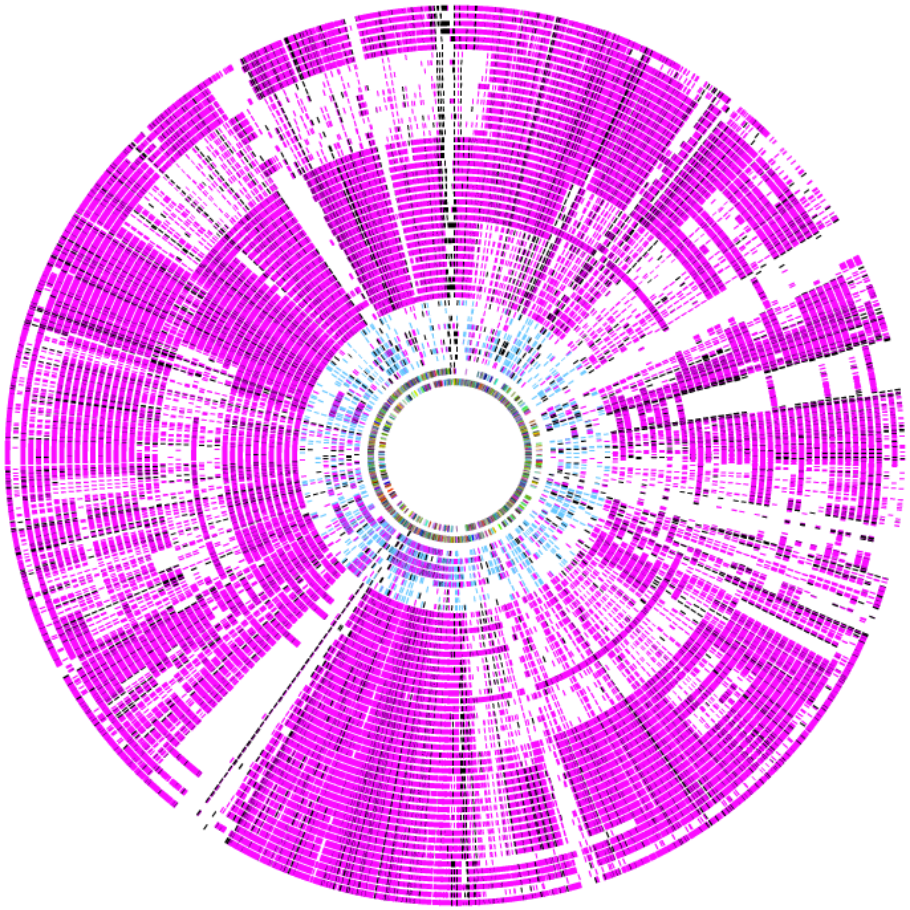
Phylogeny of *Streptococcus agalactiae* isolates



Tree based on 40,898 SNPs sites
among 1,384,073 nucleotides

Institut Pasteur

SNP distribution in 63 representative isolates



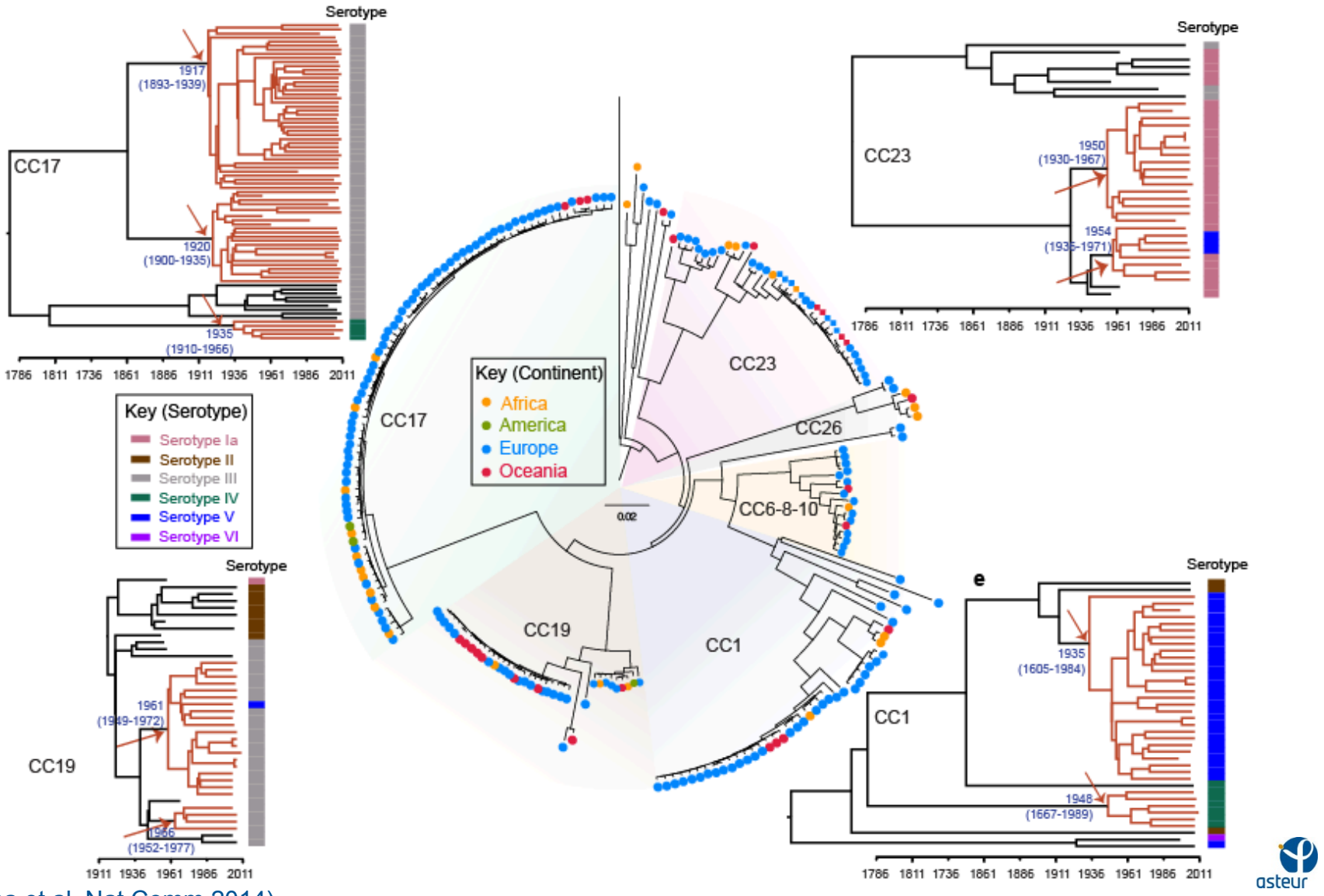
⇒ All parts of the genome have recombined

⇒ A true phylogeny of the species is not feasible

⇒ Analysis of the 6 CC with specific reference genomes

Institut Pasteur

Understanding the emergence of *S. agalactiae* neonatal infections



Da Cunha et al. Nat Comm 2014)

Variant calling workflow

Comparing every strain to a reference sequence :


```
@HISEQ:376:C3JL4ACXX:2:1101:3736:2085 1:N:0:CGTACG
AAGANCTCCAGAGCCTTACAAAGGTAAAGGTATTCGTTACCAAGGTGAATACGTTCGCCGTAAAGAAGGGAAAACTGGTAAATAATAGATAAACTCTAAAG
+
@@@B#2=BFBH<FDGIGEH@E>GECFGEIG*?CEFHFHEIDCHGGDDFGGHHGDFEGGIIGB82=7>@B'.;?CCACCC@A:@>AA>CCD@3,>CCC@###

31 31 31 33 2 17 28 33 37 […] 29 32 32 29 34 34 35 31 18 11 29 34 34 34 31 2 2 2
```

*Criscuolo A, Brisse S : **AlienTrimmer : A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads.** *Genomics* 102 (2013) 500-506

Institut Pasteur

Variant calling workflow

Comparing every strain to a reference sequence :


```
@HISEQ:376:C3JL4ACXX:2:1101:3736:2085 1:N:0:CGTACG
AAGANCTCCAGAGCCTTACAAAGGTAAAGGTATTCGTTACCAAGGTGAATACGTTCGCCGTAAAGAAGGGAAAACTGGTAAATAATAGATAAACTCTAAAG
+
@@@B#2=BFBH<FDGIGEH@E>GECFGEIG*?CEFHFHEIDCHGGDDFGGHHGDFEGGIIGB82=7>@B'.;?CCACCC@A:@>AA>CCD@3,>CCC@###

31 31 31 33 2 17 28 33 37 […] 29 32 32 29 34 34 35 31 18 11 29 34 34 34 31 2 2 2
```

```
@HISEQ:376:C3JL4ACXX:2:1101:3736:2085 1:N:0:CGTACG
AAGANCTCCAGAGCCTTACAAAGGTAAAGGTATTCGTTACCAAGGTGAATACGTTCGCCGTAAAGAAGGGAAAACTGGTAAATAATAGATA
+
@@@B#2=BFBH<FDGIGEH@E>GECFGEIG*?CEFHFHEIDCHGGDDFGGHHGDFEGGIIGB82=7>@B'.;?CCACCC@A:@>AA>CCD@
```

*Criscuolo A, Brisse S : **AlienTrimmer : A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads.** *Genomics* 102 (2013) 500-506

Institut Pasteur

Variant calling workflow

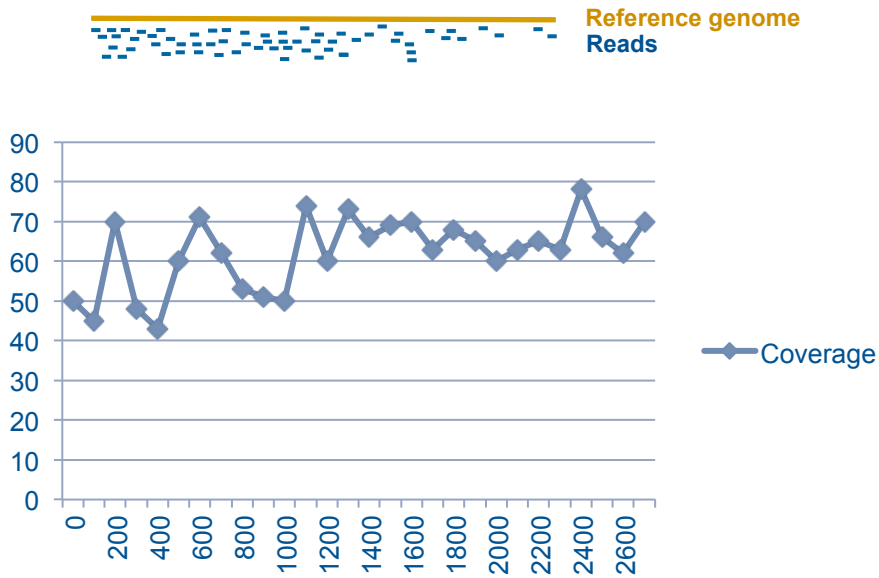
Comparing every strain to a reference sequence :

*Li H : **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** *arXiv:1303.3997v2*

Institut Pasteur

Variant calling workflow

- **Comparing every strain to a reference sequence :**

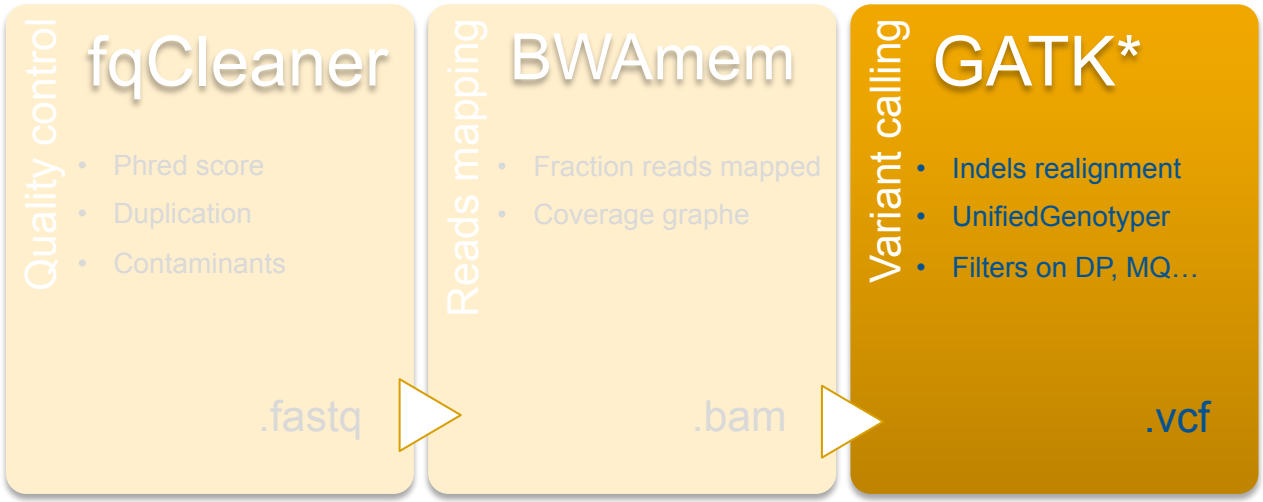


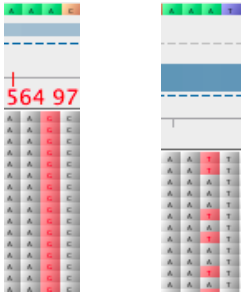
*Li H : **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** *arXiv:1303.3997v2*

Institut Pasteur

Variant calling workflow

Comparing every strain to a reference sequence :





Are these positions SNPs ?

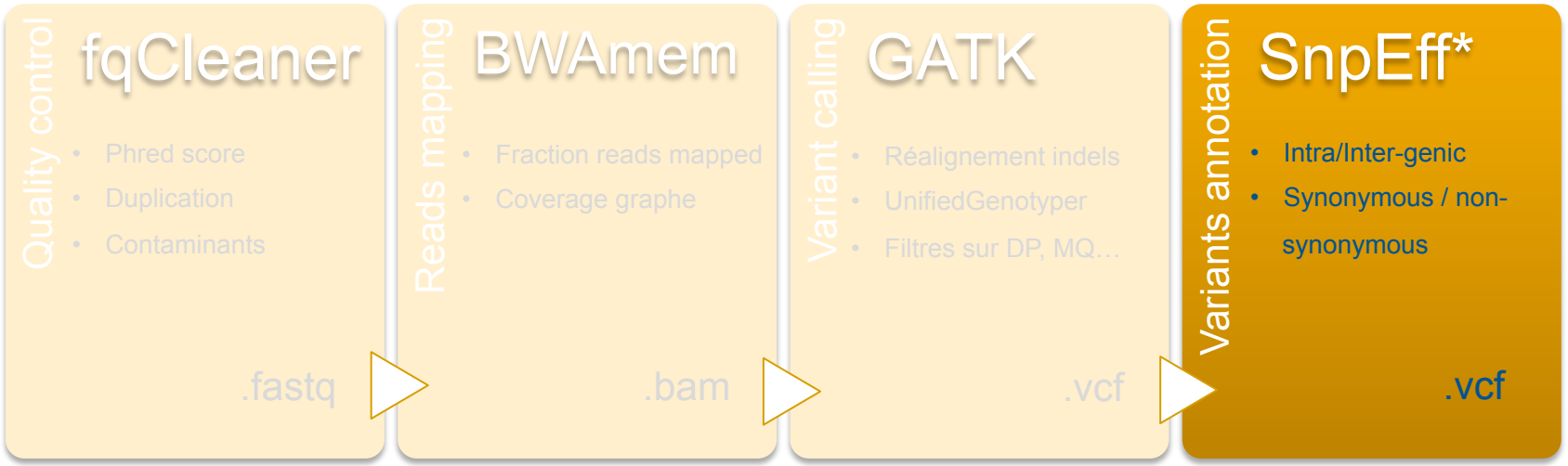
- Reads coverage
- Strand bias
- Mapping quality
- Found in another strain
- …

*DePristo M, *et. al* : **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nature Genetics* (2011) **43**:491-498

Institut Pasteur

Variant calling workflow

- **Comparing every strain to a reference sequence :**



*Cingolani P, *et. al* : **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.** *Fly* 2012 Apr-Jun;**6**(2):80-92

Institut Pasteur

Validation of the pipeline on 512 *bacterial isolates*

Sequenced at the Genopole (PF1)

- 67 strains from the French NRC (1x100bp, Hiseq 2000)
- 6 old (~10-15yo) american strains (1x150bp, Miseq)

HiSeq 2000

From the Short Reads Archive database

- 403 american strains*, (2x100bp, Hiseq 2000)
- 36 american strains**, (2x100bp Genome Analyzer IIx)

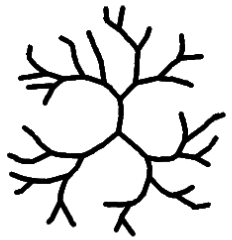
Institut Pasteur

Building a phylogeny

- **>14000 polymorphic positions :**
  - **~ 50-100 SNPs for most strains**
  - **~ 200-700 SNPs for some american strains**
  - **2200 SNPs for the outgroup**

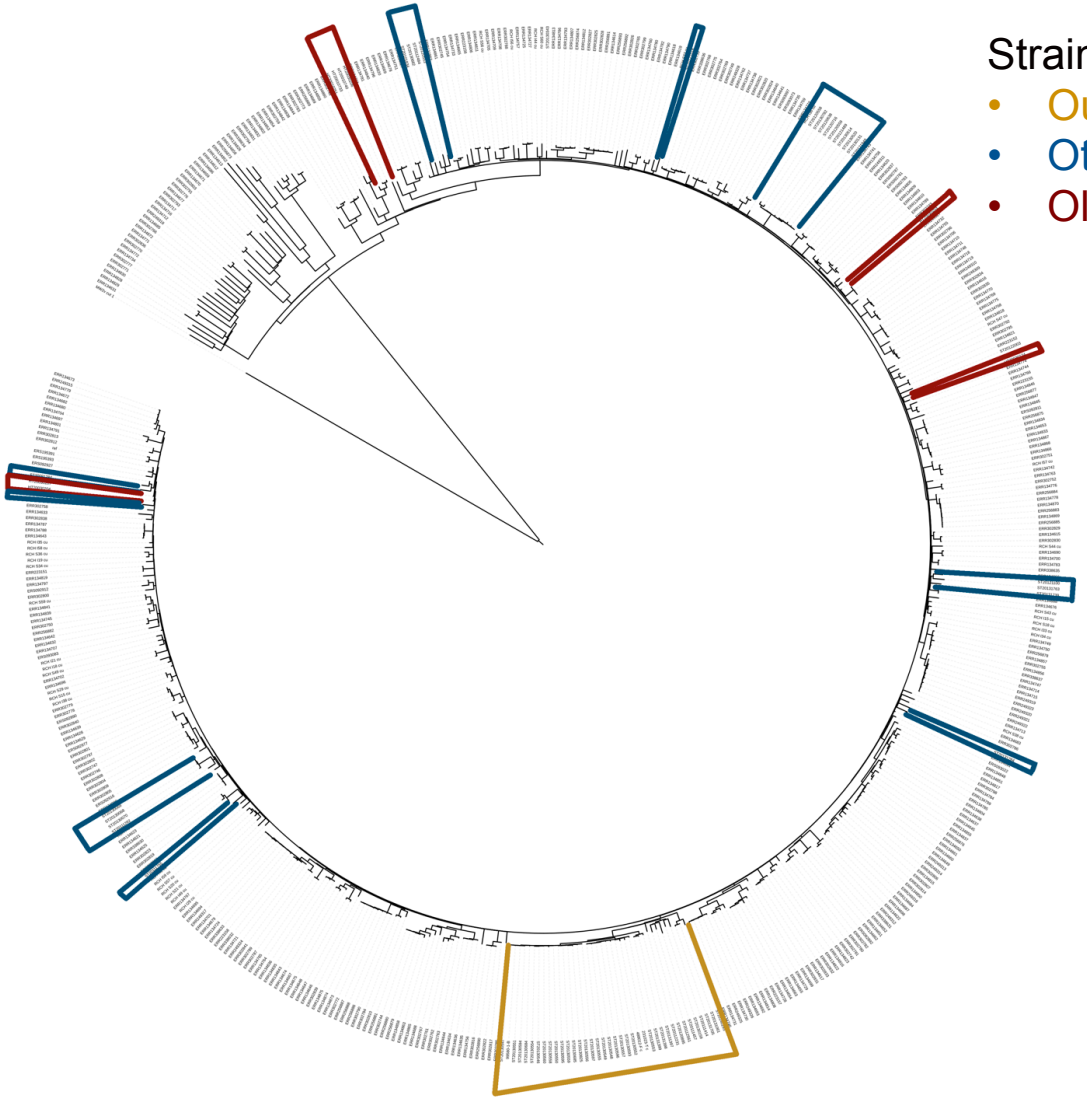
- **Additionnal filters :**
  - Known plasmids and repeated regions were removed
  - Cluster of isolates are remove - In one isolate, SNPs must be distant from each other (e. g. 100 bp)

- **Phylogeny was built using PhyML***



*Guindon S, *et. al* : **New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.** *Systematic Biology*, **59**(3):307-21, 2010.

All strains phylogeny



Strains sequenced at IP

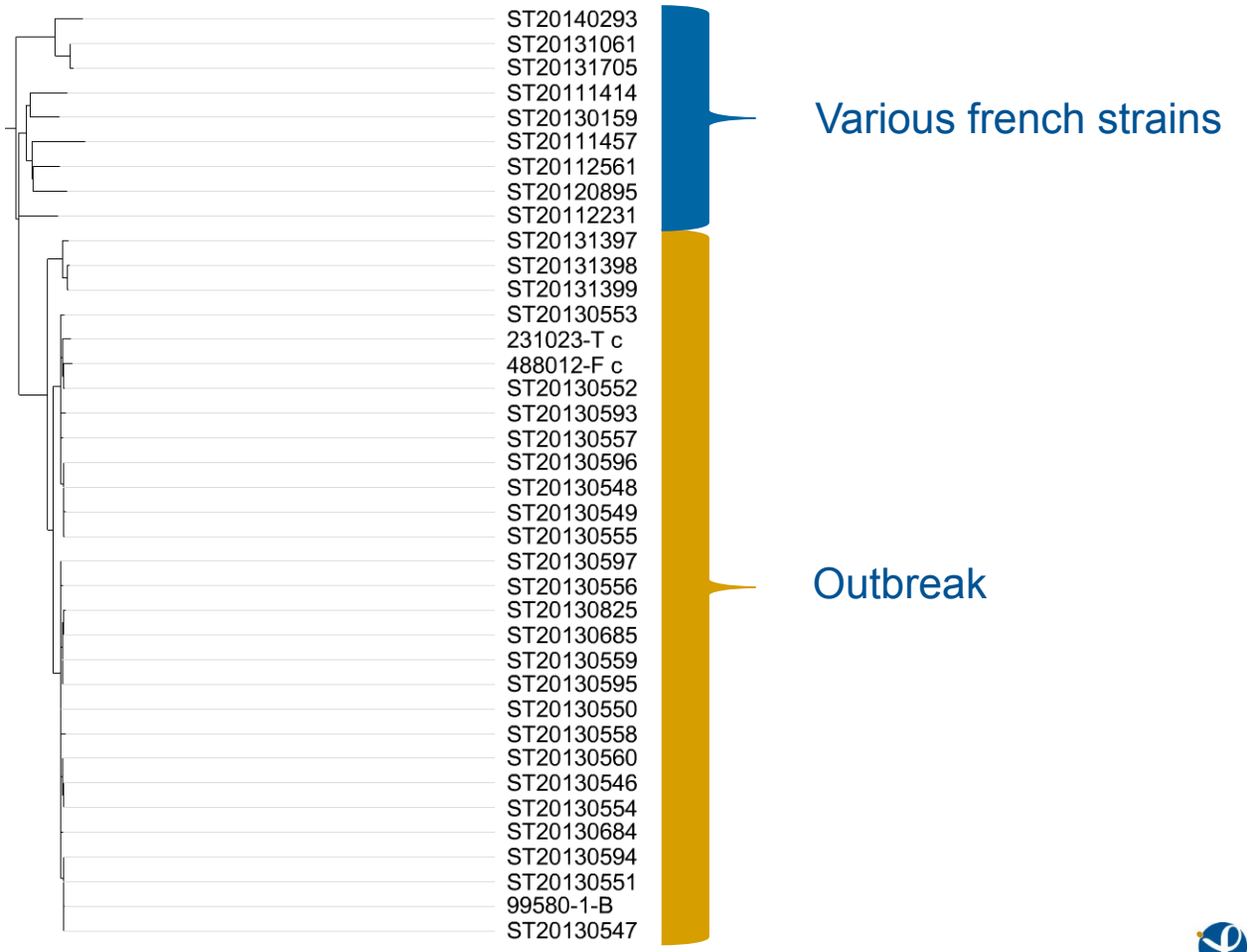
- Outbreak
- Other recent french cases
- Old american cases

All strains phylogeny

Institut Pasteur

French subtrees

The hospital outbreak strains form a monophyletic group

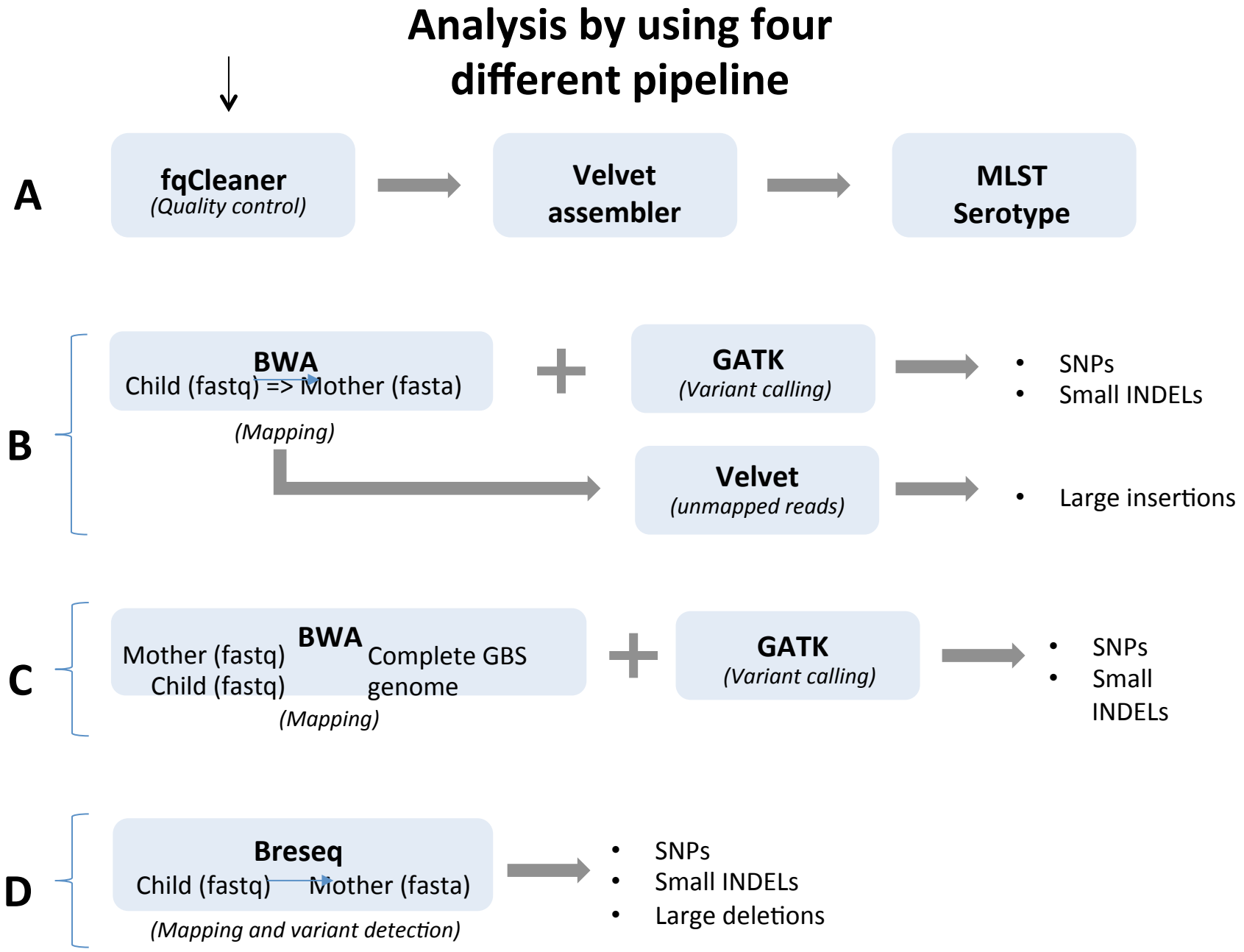


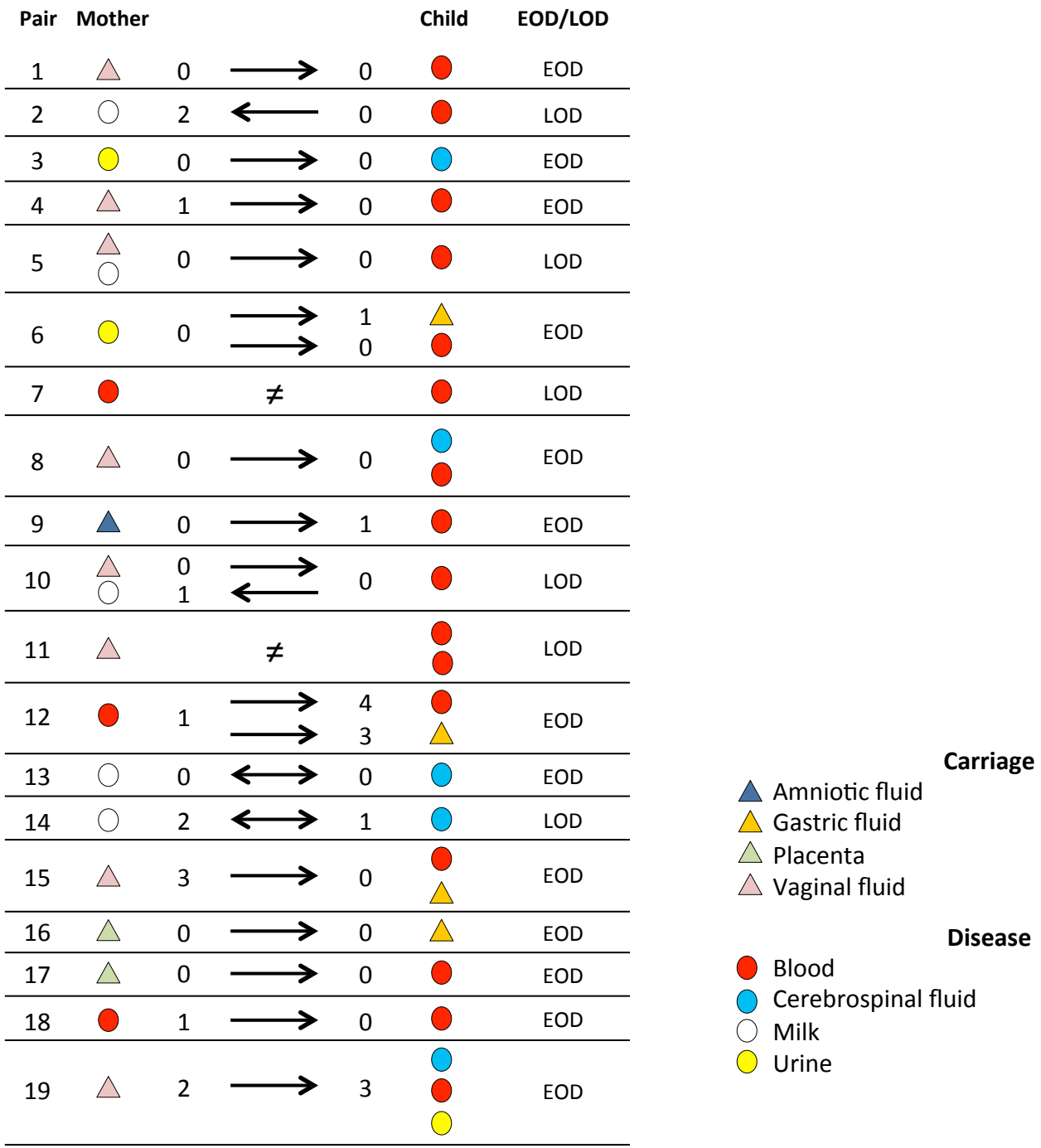
Institut Pasteur

In vivo evolution

- **Objective comparing S. agalactiae strains isolated from the infantand from their mother.**
- **Sequencing of 47 isolates from 19 pairs (mother / infant)**
- **Disease or carriage associated isolates**
- **Identify polymorphisms linked to the transition from comensal to pathogen**

Institut Pasteur


Analysis by using four different pipeline
A
fqCleaner
(Quality control)
Velvet assembler
MLST Serotype
B
BWA
Child (fastq) => Mother (fasta)
(Mapping)
GATK
(Variant calling)
SNPs
Small INDELs
Velvet
(unmapped reads)
Large insertions
C
Mother (fastq)
Child (fastq)
BWA
Complete GBS genome
(Mapping)
GATK
(Variant calling)
SNPs
Small INDELs
D
Breseq
Child (fastq) Mother (fasta)
(Mapping and variant detection)
SNPs
Small INDELs
Large deletions


Pair
Mother
Child
EOD/LOD
1
0
0
EOD
2
2
0
LOD
3
0
0
EOD
4
1
0
EOD
5
0
0
LOD
6
0
1
0
EOD
7
≠
LOD
8
0
0
EOD
9
0
1
EOD
10
0
1
0
LOD
11
≠
LOD
12
1
4
3
EOD
13
0
0
EOD
14
2
1
LOD
15
3
0
EOD
16
0
0
EOD
17
0
0
EOD
18
1
0
EOD
19
2
3
EOD
Carriage
Amniotic fluid
Gastric fluid
Placenta
Vaginal fluid
Disease
Blood
Cerebrospinal fluid
Milk
Urine

Visualisation with SynTView

SynTView*

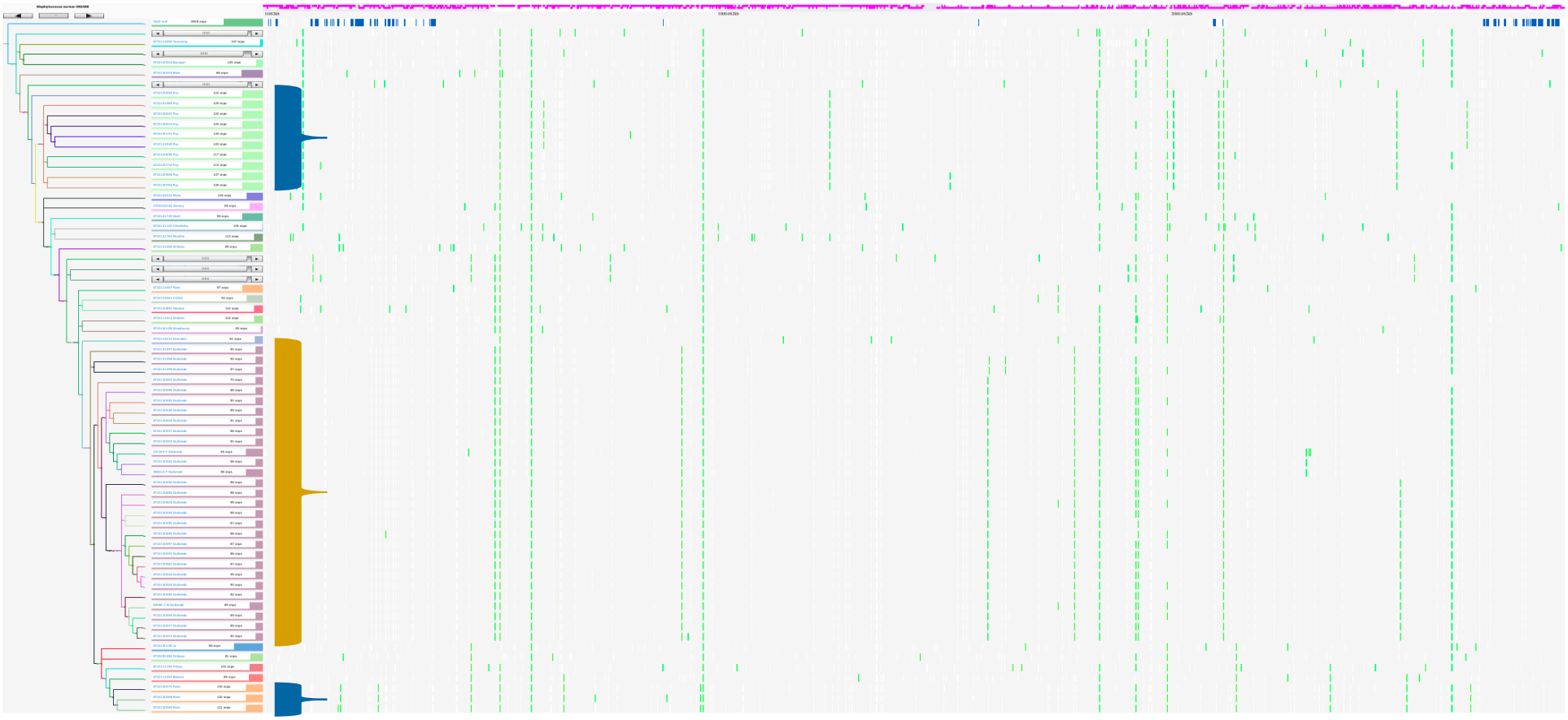
- Genome browser with multiple view (synteny, circular view, SNPs…)
- Polymorphisms and short indels visualisation for n strains versus a reference chromosome
- Integration of metadata : virulence, date of isolation…
- Ordering of strains with a phylogenetic tree
- genopole.pasteur.fr/SynTView/

SynTView 2

- Developped in collaboration with Genostar (Wallgene)
- Reference as contigs is now supported
- Better interface

*Lechat P, Souche E, Moszer I **: SynTView — an interactive multi-view genome browser for next-generation comparative microorganism genomics.** *BMC Bioinformatics* 2013, **14**:277

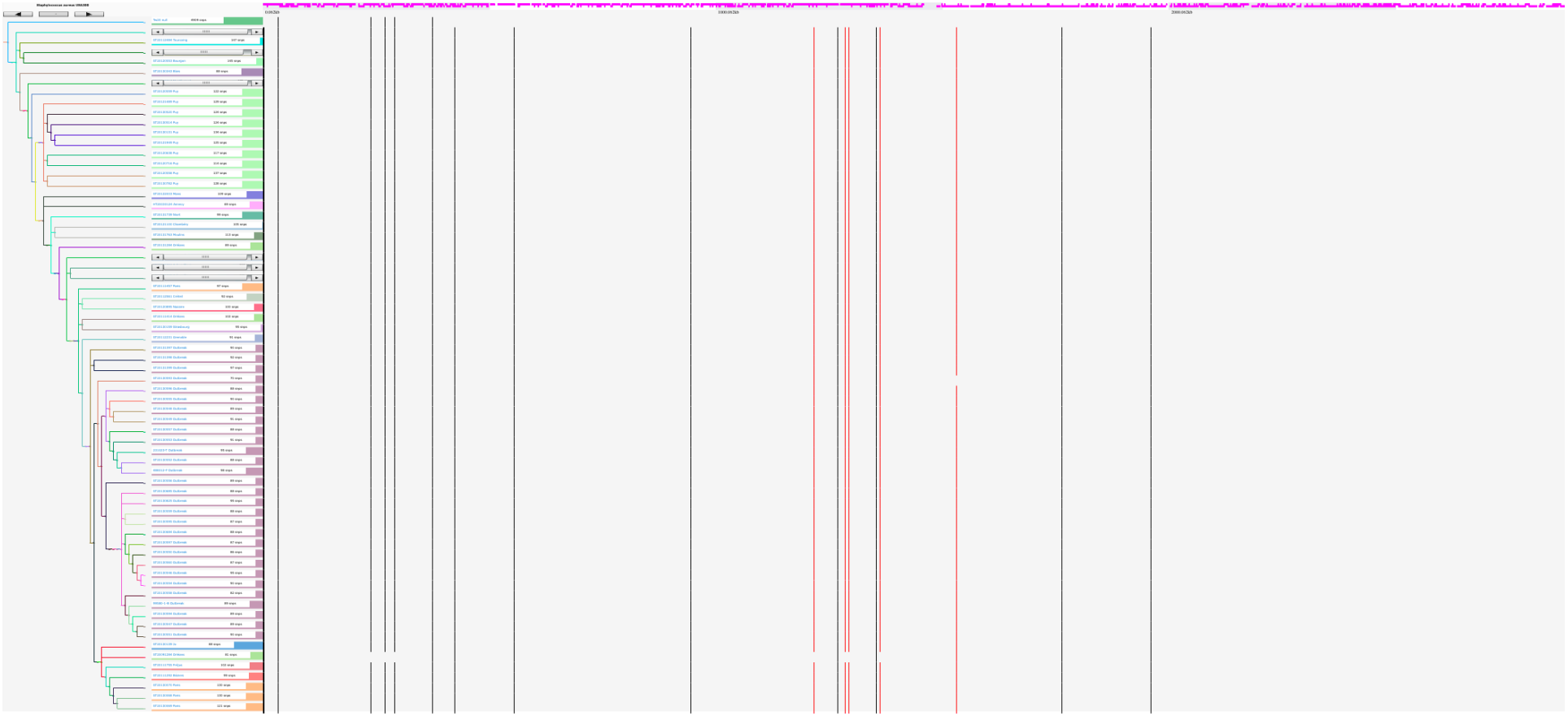
Analysing French strains in SynTView



Insertions and deletions

Institut Pasteur

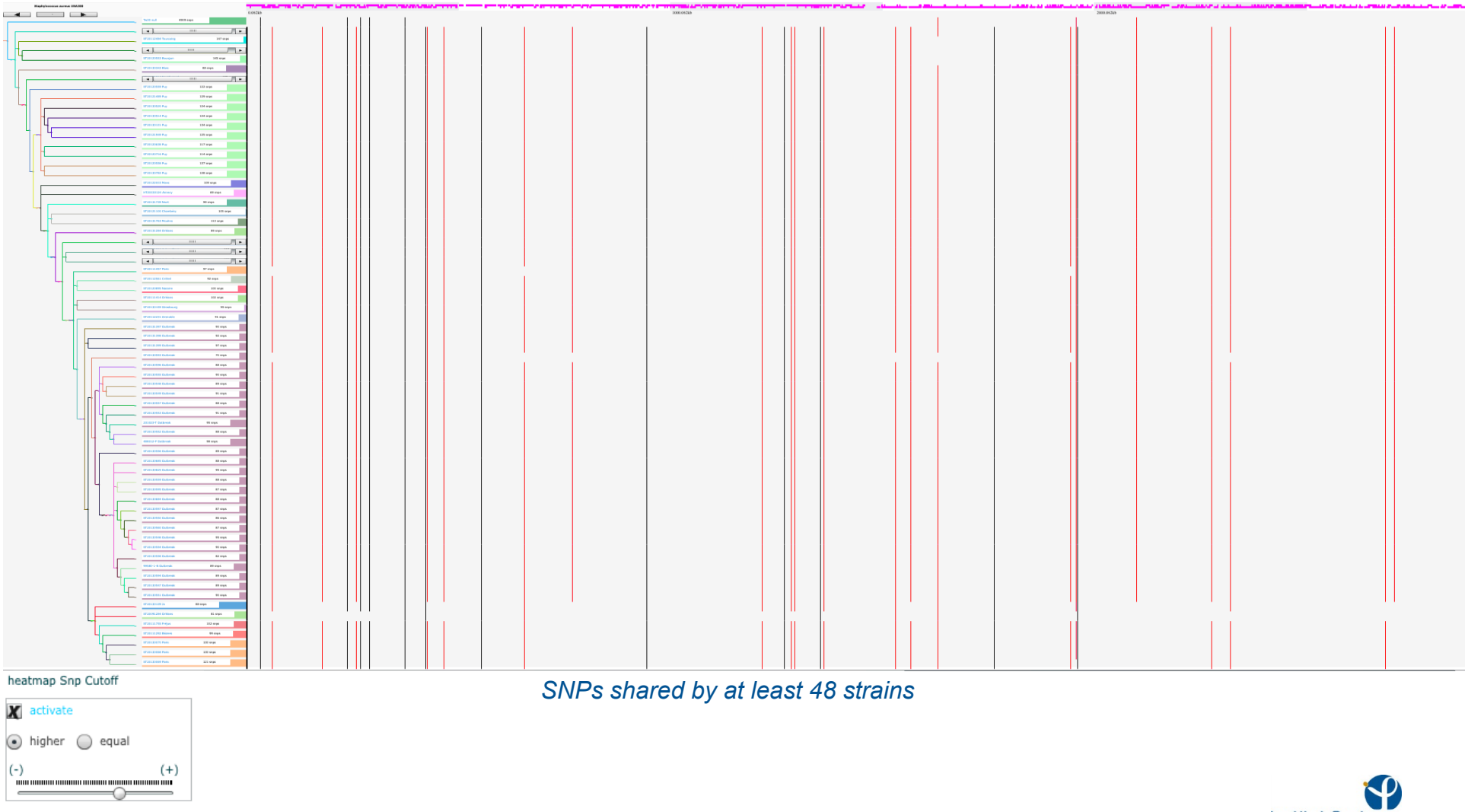
Analysing French strains in SynTView



SNPs shared by at least 67 strains

Institut Pasteur

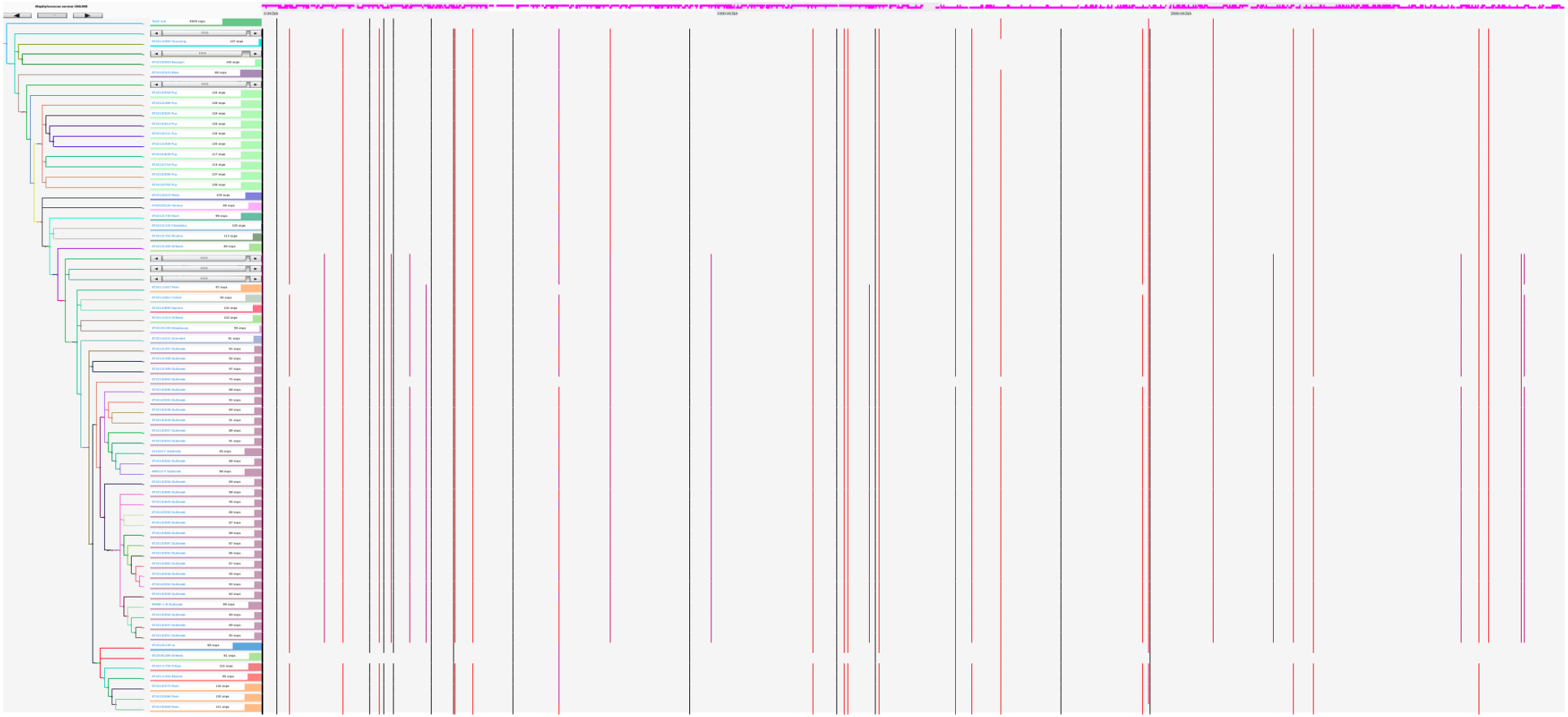
Analysing French strains in SynTView



SNPs shared by at least 48 strains

Institut Pasteur

Analysing French strains in SynTView



heatmap Snp Cutoff

activate

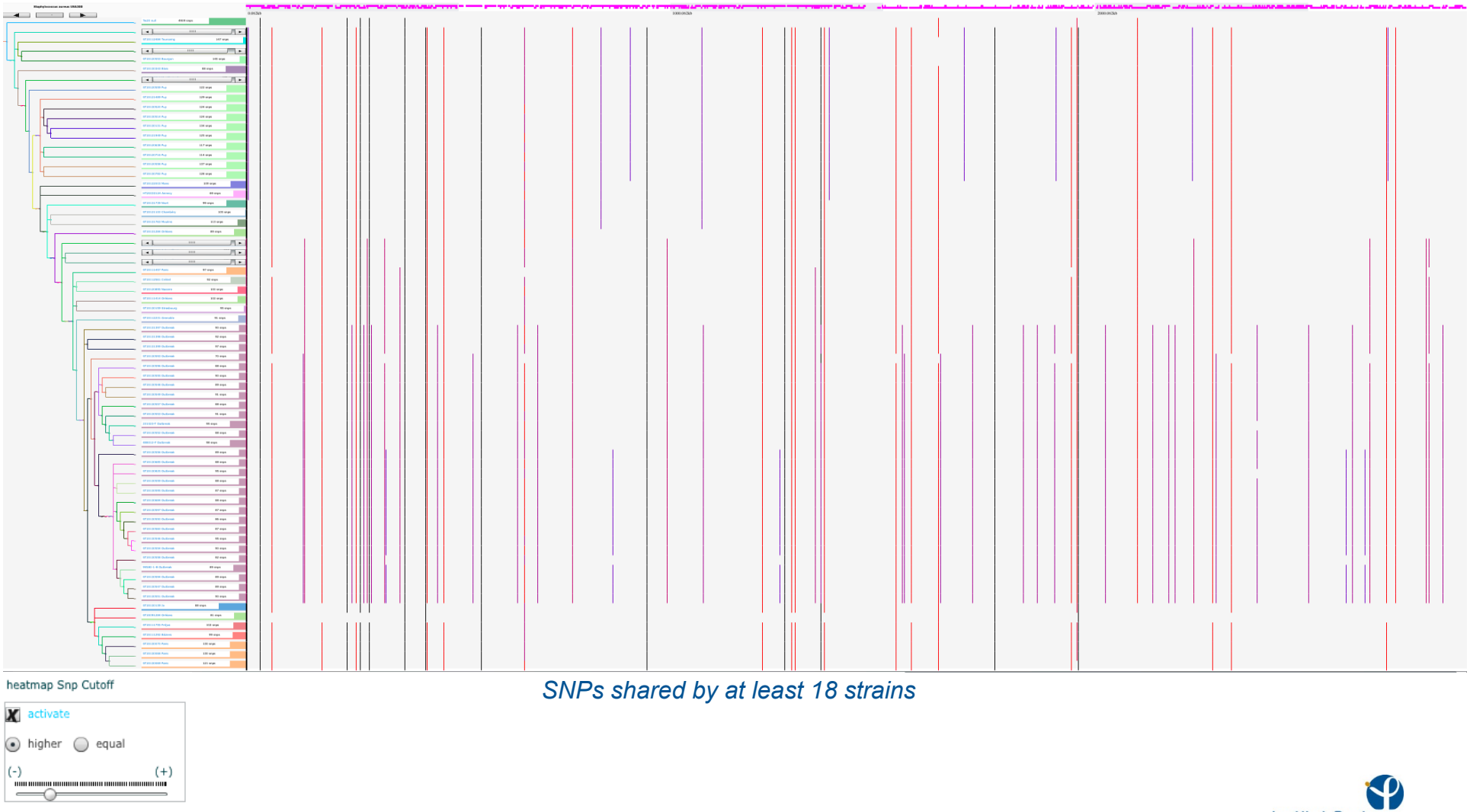
higher equal

(-) (+)

SNPs shared by at least 32 strains

Institut Pasteur

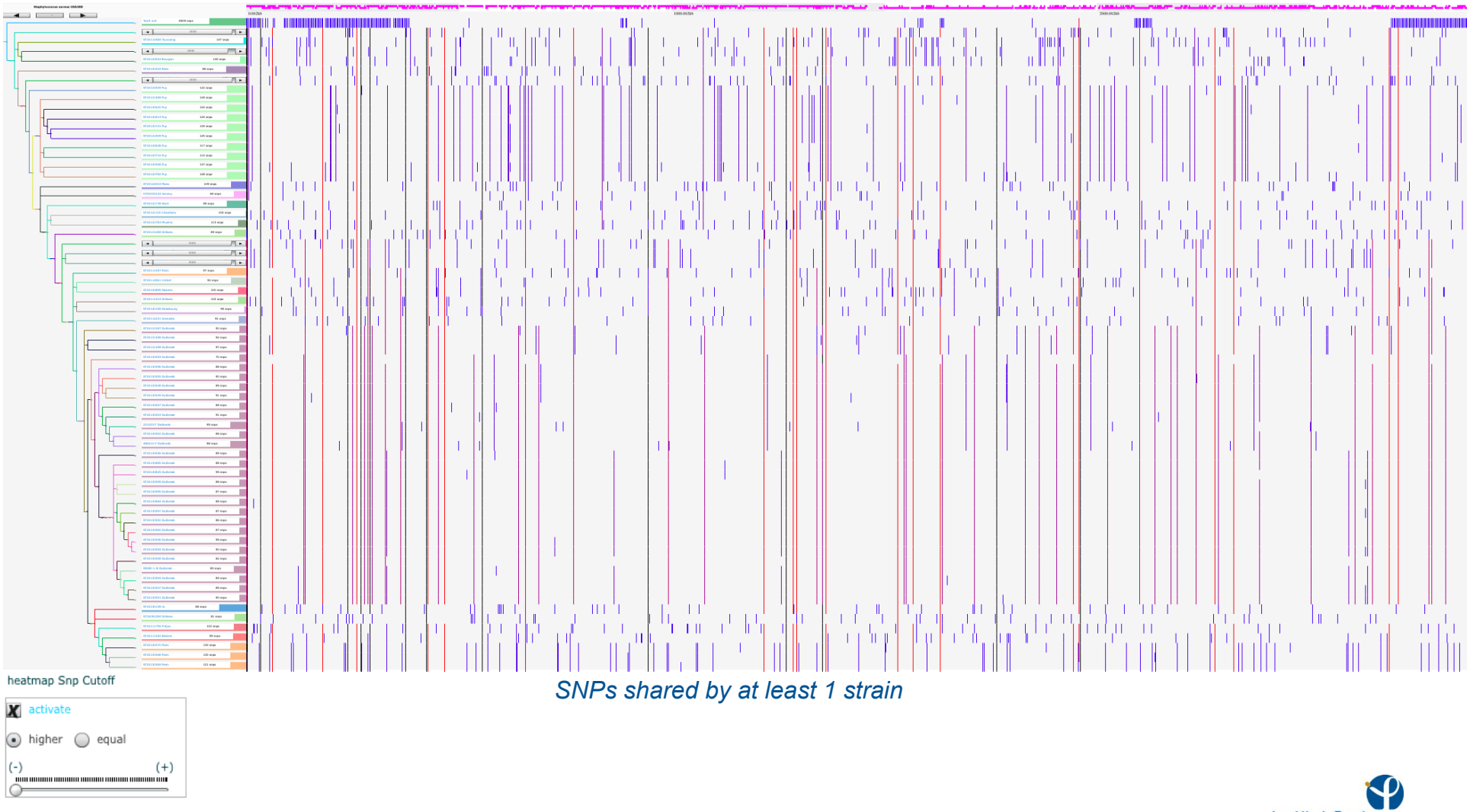
Analysing French strains in SynTView



SNPs shared by at least 18 strains

Institut Pasteur

Analysing French strains in SynTView



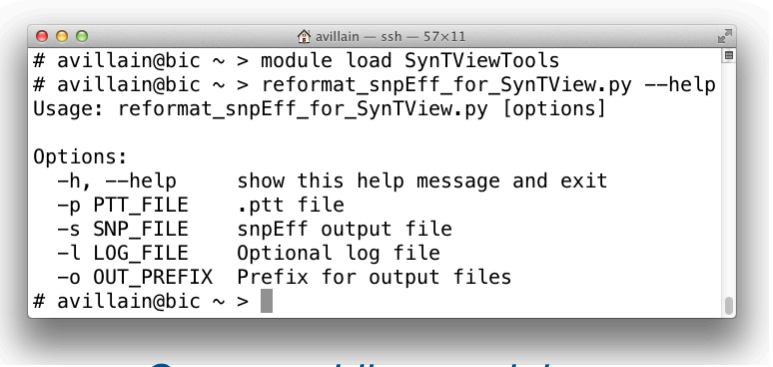
SNPs shared by at least 1 strain

Institut Pasteur

Set up a SynTView website

Conversion to SynTView formats :

- Reduce the size of files loaded on the site
- Easy to use scripts available :
  - **On central-bio or bic@pasteur.fr**



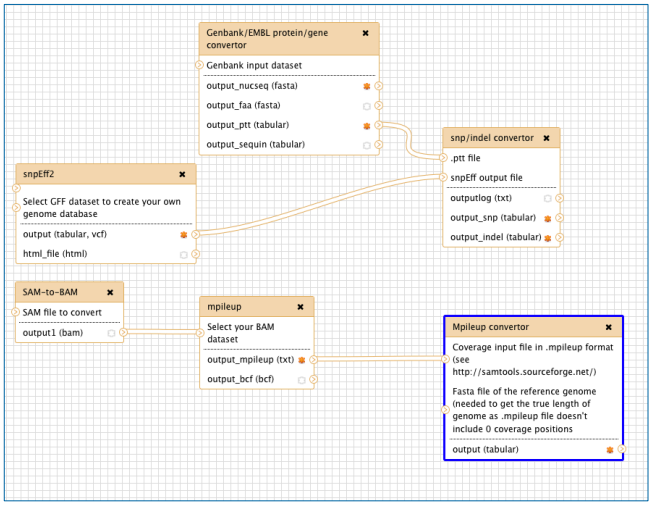
Command-line module

Institut Pasteur

Set up a SynTView website

Conversion to SynTView formats :

- Reduce the size of files loaded on the site
- Easy to use scripts available :
  - **On central-bio or bic@pasteur.fr**
  - **In galaxy@pasteur.fr toolshed**



Galaxy workflow

Institut Pasteur

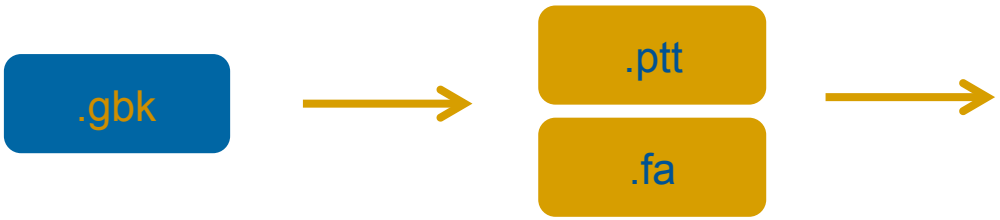
Set up a SynTView website

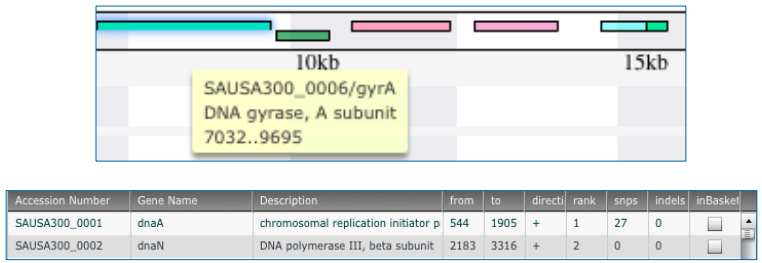
Conversion to SynTView formats :

- Reduce the size of files loaded on the site
- Easy to use scripts available :
  - **On central-bio or bic@pasteur.fr**
  - **In galaxy@pasteur.fr toolshed**

Steps :

- Get coding sequences information





ORFs and genes list

Institut Pasteur

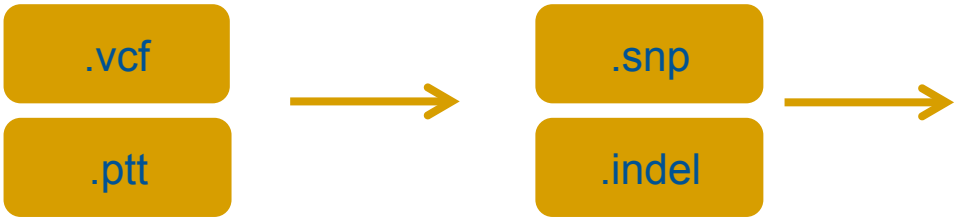
Set up a SynTView website

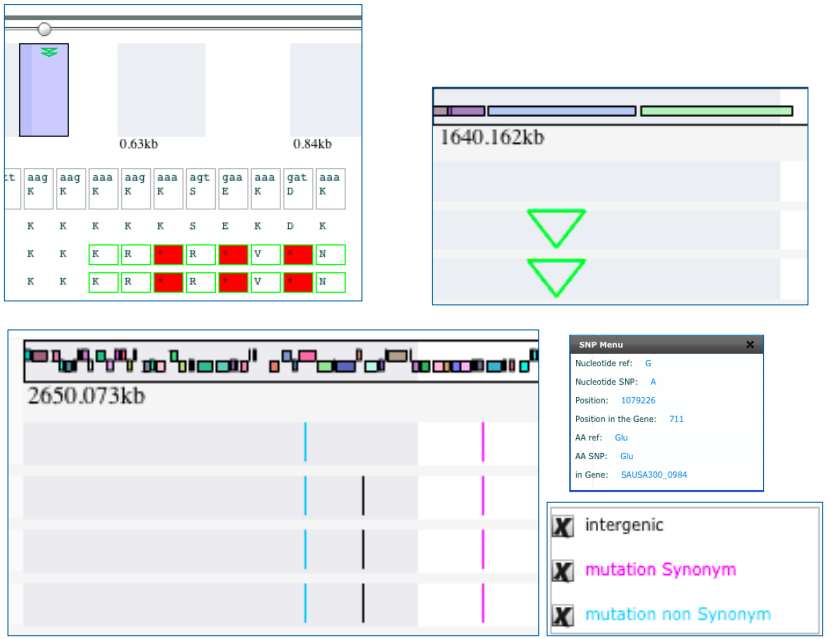
Conversion to SynTView formats :

- Reduce the size of files loaded on the site
- Easy to use scripts available :
  - **On central-bio or bic@pasteur.fr**
  - **In galaxy@pasteur.fr toolshed**

Steps :

- Get coding sequences information
- Convert variant file





SNPs and short indels visualisation

Institut Pasteur

Set up a SynTView website

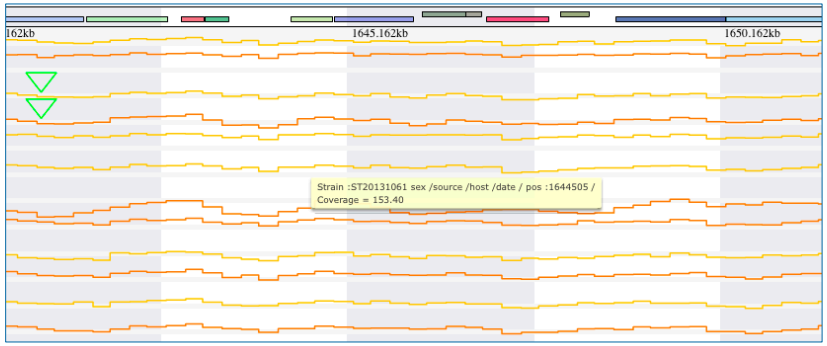
Conversion to SynTView formats :

- Reduce the size of files loaded on the site
- Easy to use scripts available :
  - **On central-bio or bic@pasteur.fr**
  - **In galaxy@pasteur.fr toolshed**

Steps :

- Get coding sequences information
- Convert variant file
- Add coverage track

.mpileup → .coverage →



Coverage along genome

Institut Pasteur

Conclusion and perspective

- The pipeline has been validated on the 500 bacterial isolates
- SynTView is particularly suited for the exploration of results
- Workflows for SNPs detection and results integration into SynTView are available on the servers and Galaxy
- Current work on detection of large indels and CNVs
- Implementation of tools for detecting recombination events and homoplasy.
- SynTView 2 is in active development and constantly improving

Thank you for your attention

Institut Pasteur