Mapping against reference genome and Variant Calling

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

2-10 Noviembre 2022
Programa Formación, AESAN

Index

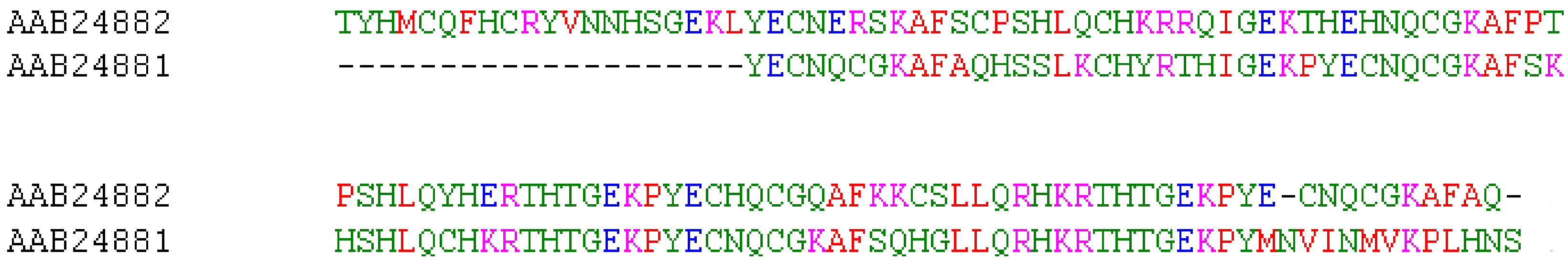
Mapping against reference genome and Variant Calling :

- Mapping vs Alignment
- What is mapping?
- How to choose a NGS mapper.
- SAM/BAM format
- Duplicate filter
- Variant Calling
- Source of error and mitigation strategies
- VCF and bed format
- GATK vs VARSCAN2
- High quality SNP selection

Alignment

Definition:

Arrange two or more nucleotide or aminoacid sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships.

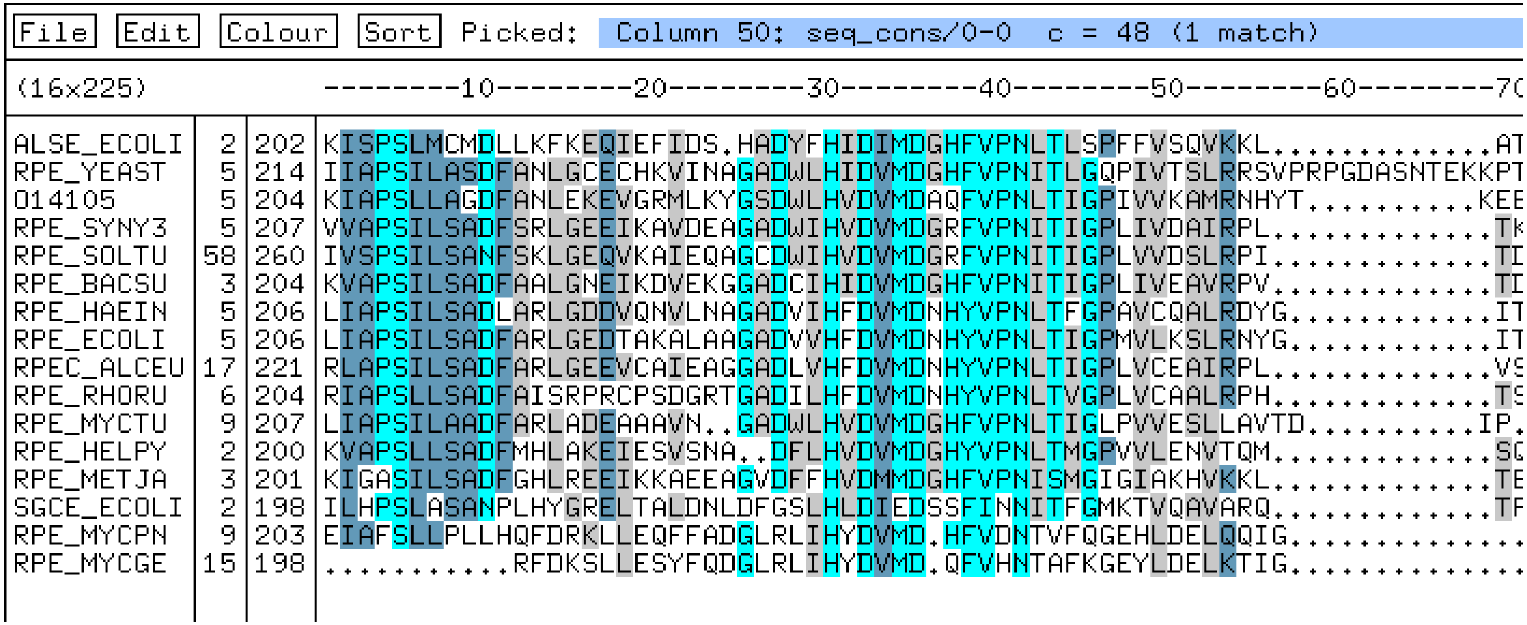


```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK


AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ-
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS
```

Multiple alignment (MSA)

Definition:

A multiple alignment is a colection of three or more sequences partial or completely aligned.

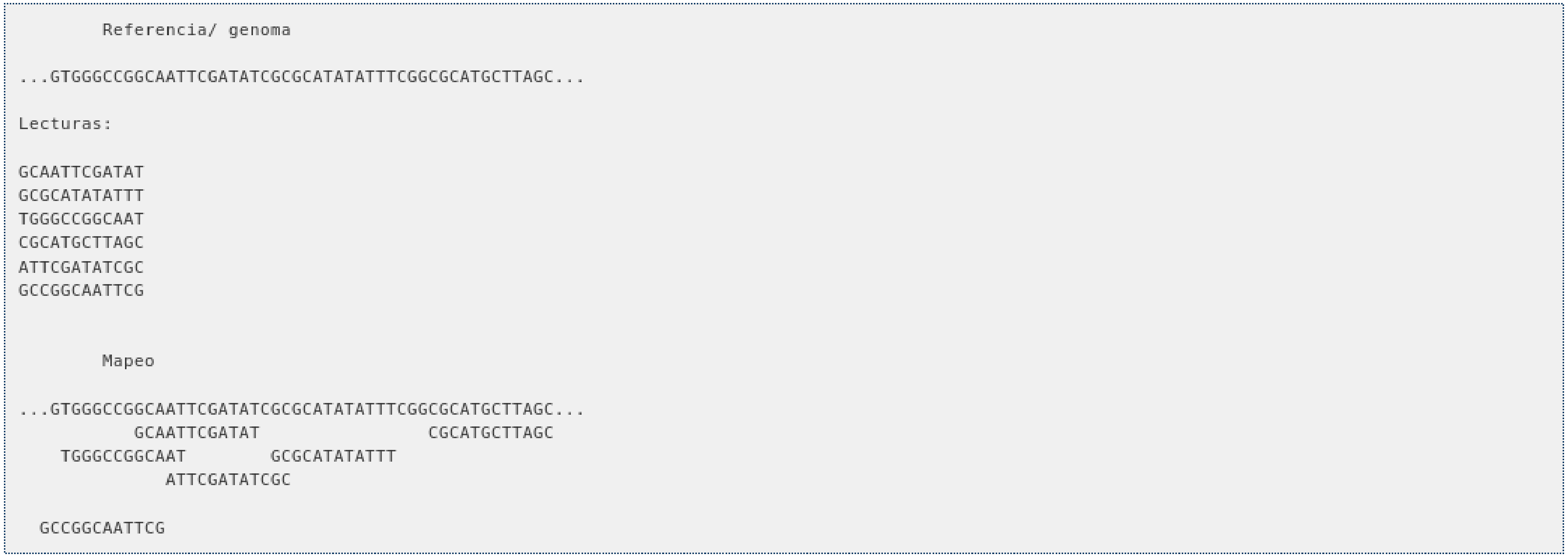


```
Picked:  Column 50: seq_cons/0-0   c = 48 (1 match)

(16x225)                 --------10--------20--------30--------40--------50--------60--------70

ALSE_ECOLI   2 202  KISPSLMCMDLLKFKEQIEFIDS.HADYFHIDIMDGHFVPNLTLSPFFVSQVKKL.............AT
RPE_YEAST    5 214  IIAPSILASDFANLGCECHKVINAGADWLHIDVMDGHFVPNITLGQPIVTSLRRSVPRPGDASNTEKKPT
014105       5 204  KIAPSLLAGDFANLEKEVGRMLKYGSDWLHVDVMDAQFVPNLTIGPIVVKAMRNHYT..........KEE
RPE_SYNY3    5 207  VVAPSILSADFSRLGEEIKAVDEAGADWIHVDVMDGRFVPNITIGPLIVDAIRPL.............TK
RPE_SOLTU   58 260  IVSPSILSANFSKLGEQVKAIEQAGCDWIHVDVMDGRFVPNITIGPLVVDSLRPI.............TI
RPE_BACSU    3 204  KVAPSILSADFAALGNEIKDVEKGGADCIHIDVMDGHFVPNITIGPLIVEAVRPV.............TI
RPE_HAEIN    5 206  LIAPSILSADLARLGDDVQNVLNAGADVIHFDVMDNHYVPNLTFGPAVCQALRDYG............IT
RPE_ECOLI    5 206  LIAPSILSADFARLGEDTAKALAAGADVVHFDVMDNHYVPNLTIGPMVLKSLRNYG............IT
RPEC_ALCEU  17 221  RLAPSILSADFARLGEEVCAIEAGGADLVHFDVMDNHYVPNLTIGPLVCEAIRPL.............VS
RPE_RHORU    6 204  RIAPSLLSADFAISRPRCPSDGRTGADILHFDVMDNHYVPNLTVGPLVCAALRPH.............TS
RPE_MYCTU    9 207  LIAPSILAADFARLADEAAAVN..GADWLHVDVMDGHFVPNLTIGLPVVESLLAVTD..........IP.
RPE_HELPY    2 200  KVAPSLLSADFMHLAKEIESVSNA..DFLHVDVMDGHYVPNLTMGPVVLENVTQM.............SG
RPE_METJA    3 201  KIGASILSADFGHLREEIKKAEEAGVDFFHVDMMDGHFVPNISMGIGIAKHVKKL.............TE
SGCE_ECOLI   2 198  ILHPSLASANPLHYGRELTALDNLDFGSLHLDIEDSSFINNITFGMKTVQAVARQ.............TF
RPE_MYCPN    9 203  EIAFSLLPLLHQFDRKLLEQFFADGLRLIHYDVMD.HFVDNTVFQGEHLDELQQIG..............
RPE_MYCGE   15 198  ...........RFDKSLLESYFQDGLRLIHYDVMD.QFVHNTAFKGEYLDELKTIG..............
```

Mapping definition

Definición:

Place a sequence inside a larger sequence. For example, determine the position of a read inside a reference genome.



```
        Referencia/ genoma

...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...

Lecturas:

GCAATTCGATAT
GCGCATATATTT
TGGGCCGGCAAT
CGCATGCTTAGC
ATTCGATATCGC
GCCGGCAATTCG


        Mapeo

...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...
           GCAATTCGATAT                CGCATGCTTAGC
    TGGGCCGGCAAT        GCGCATATATTT
              ATTCGATATCGC

  GCCGGCAATTCG
```

Alignment vs mapping

Mapping:

- A mapping is regarded to be correct if it overlaps the true region.
- Each read maps independently
- From thousand to millions of sequences.

Multiple alignment:

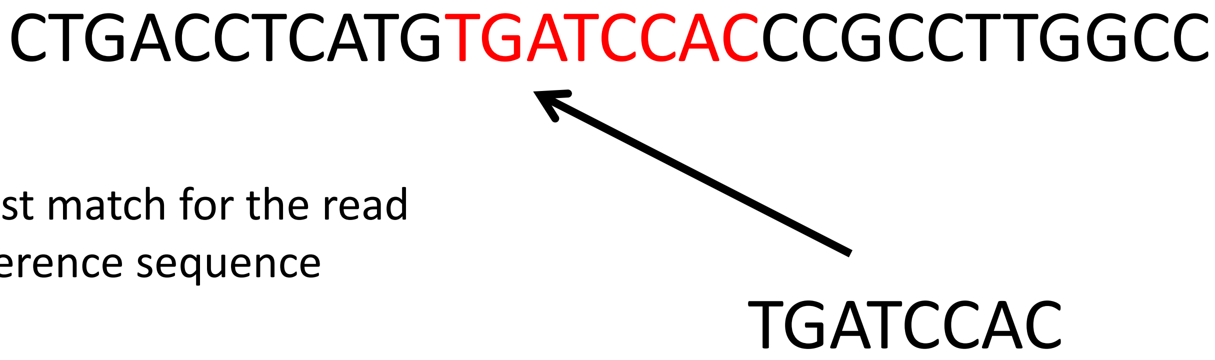
- An alignment is regarded to be correct only if each base is placed correctly.
- Minimizes differences among sequences
- From tens to hundred of sequences.

Consideratiosn:

- An algorithm can be good at mapping but may not be good aligning.
- This is because the true alignment minimizes differences between reads, but the read mapper only sees the reference.

Hen Li. Mapping, Alignment and SNP Calling. MPG Next Gen Workshop 2011

So in summary…



Find best match for the read
in a reference sequence

Challenges

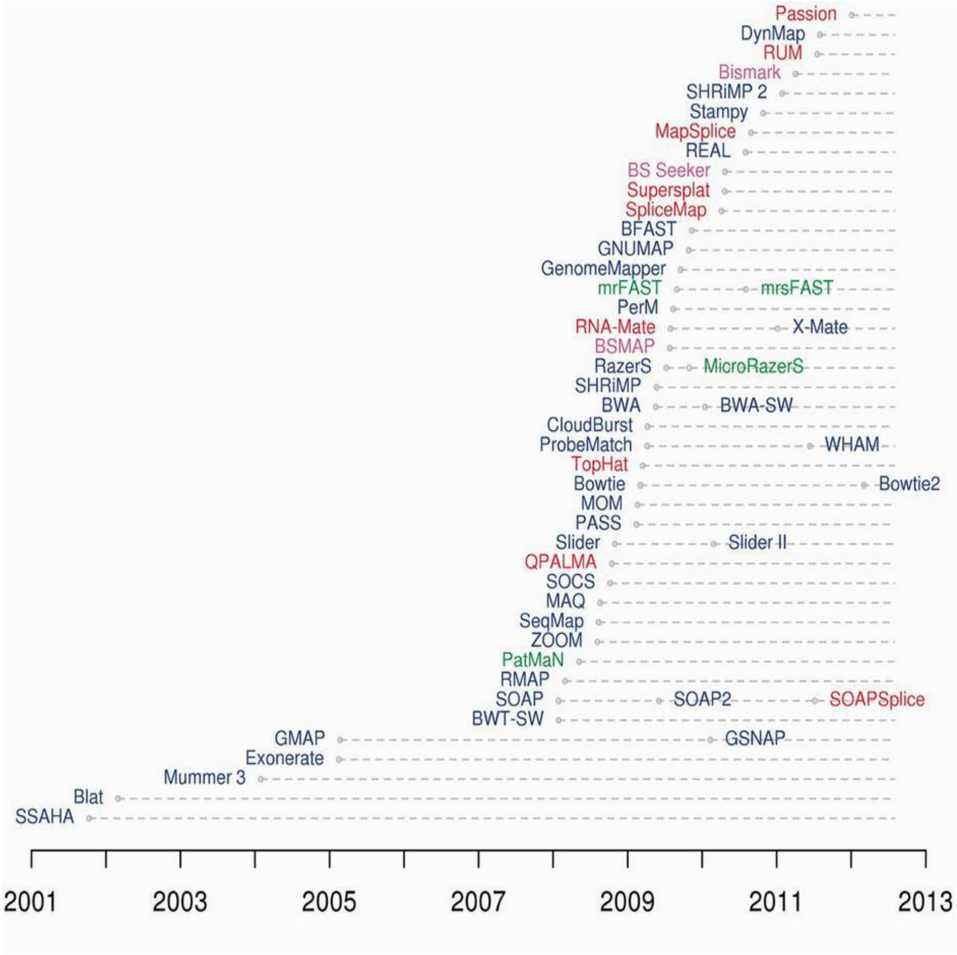
- Errors in reads
- Errors in libraries
- Repetitive regions (repeats, homologous regions)
- Homopolymers
- Individual polymorphisms

Pierre Lechat. Variants Calling lecture. Pasteur.fr

What mapper should I use?

Mappers:

- Más de 60 mappers available.
- Lots of papers reviewing its performamnce.



What mapper should I use?

Cosas a tener en cuenta:

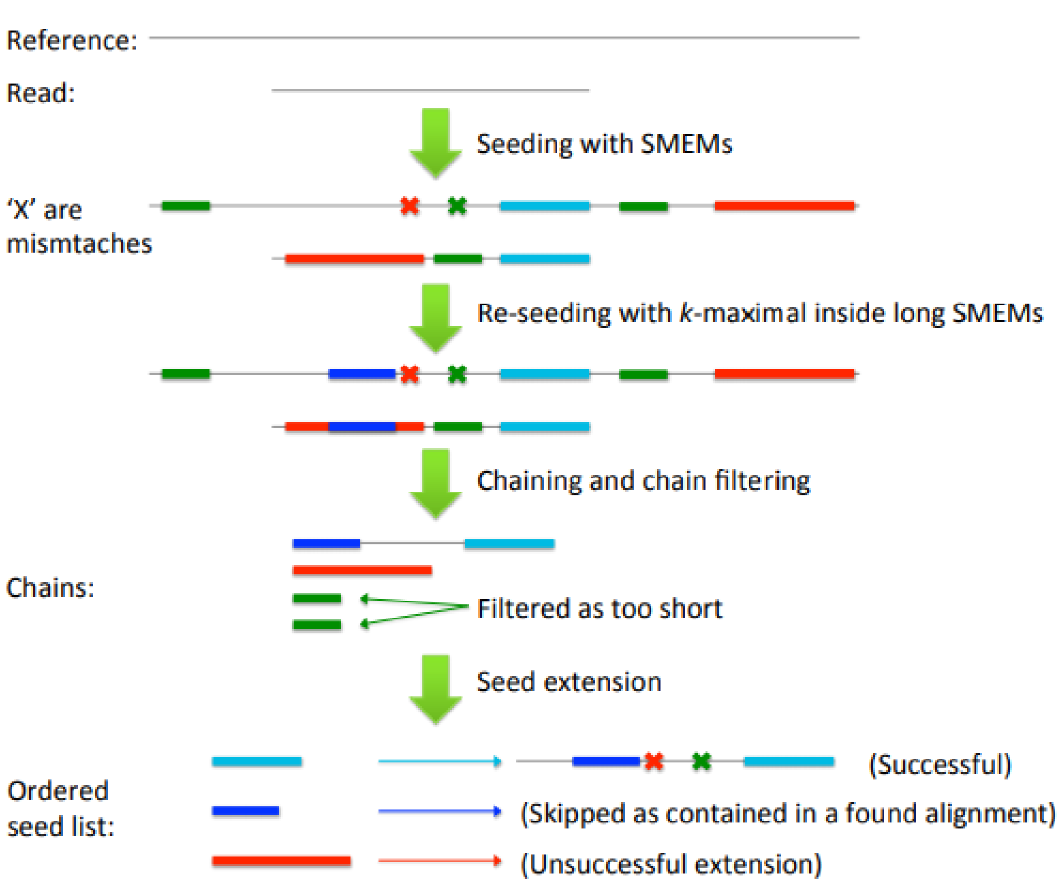
- Computational resources vs sensibility
- Platform and type of experiment (Illumina/454/etc,paired-end,DNA/RNA/etc)
- Variation (indels allowance, mistmatch number,etc.)
- Repetitions (all regions, best match, random, user defined number…)

Importante:

- Default options don’t have to be the best:

“… there is no tool that outperforms all of the others in all the tests. Therefore, the end user should clearly specify his needs in order to choose the tool that provides the best results.” - Hatem et al *BMC Bioinformatics* 2013, **14**:184

BWA MEM



SMEM strategy

- Maximal exact match (MEM): an exact match that cannot be extended further in either direction
- Super-maximal exact match (SMEM): a MEM that is not contained in any other MEMs on the query coordinate (Li, 2012). At any query position, the longest exact match covering the position must be a SMEM.

Seed-and-extend algorithm

Local alignment

Hen LI. Aligning sequence reads, clone sequences and assembly con*gs with BWA-MEM. Poster. Broad Institute.

BOWTIE2

End-to-end alignment by default.

Three reporting modes:

- Best alignment
- K alignments
- All alignments

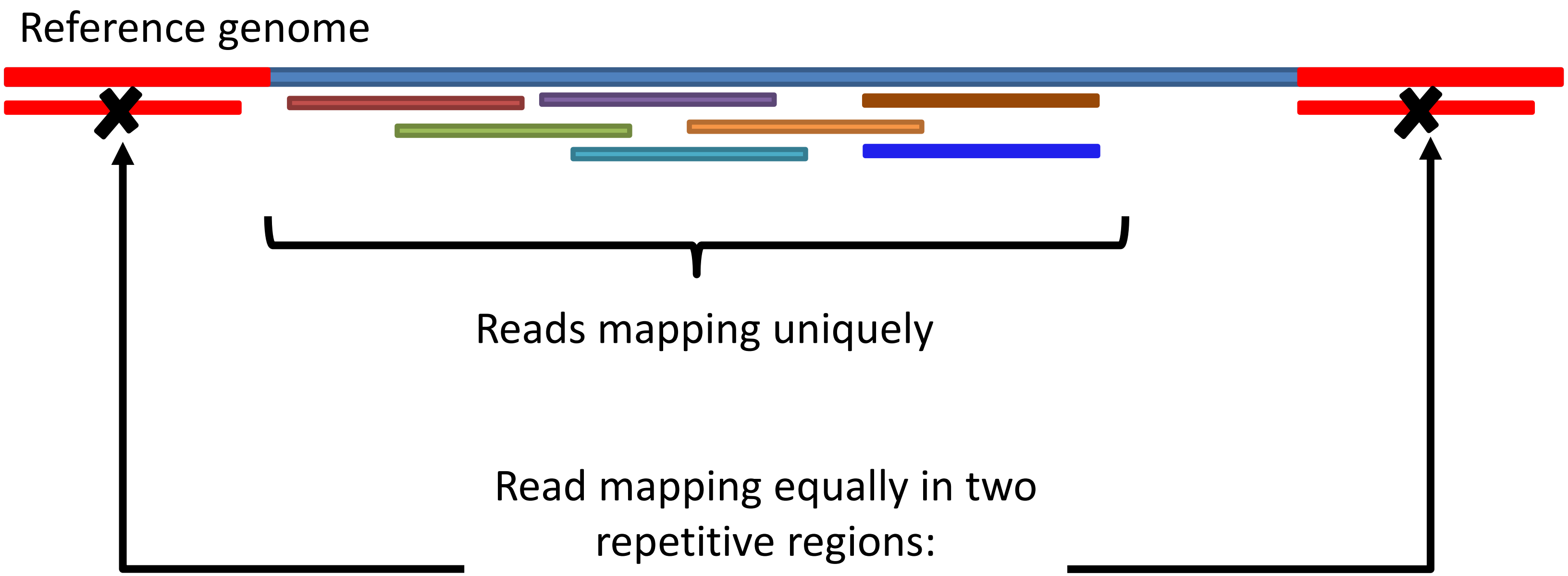
Lots of customizable parameters that change its performance.

Example whole genome aligner: MUMMER

- **Maximal Unique Matcher (MUM)**
  - match <- exact match of a minimum length
  - maximal <- cannot be extended in either direction without a mismatch
  - unique
    - occurs only once in both sequences (MUM)
    - occurs only once in a single sequence (MAM)
    - occurs one or more times in either sequence (MEM)

Adam M. Phillippy. Whole Genome Alignment with MUMmer. Presentation.

Which aligner should I use for aligning reads agains a complete genome for variant calling?



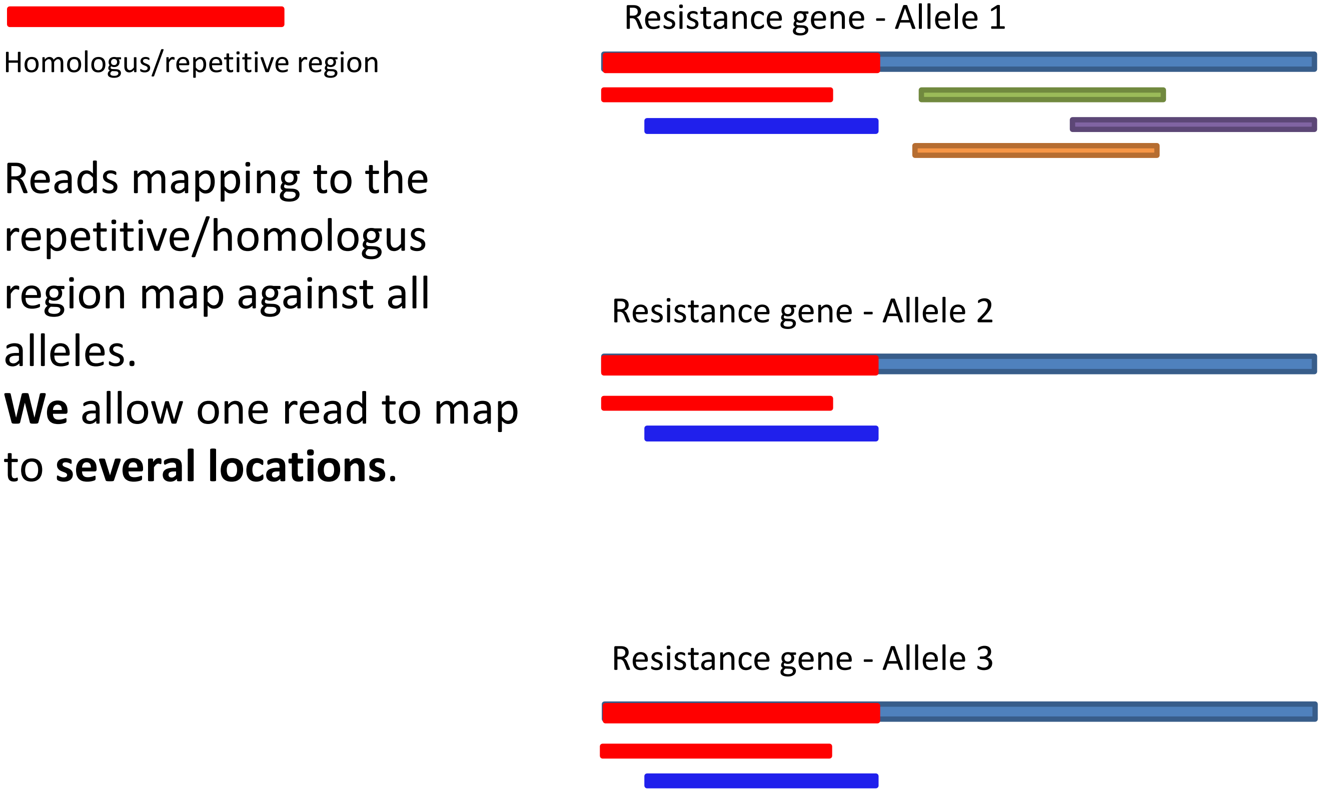
Reference genome

Reads mapping uniquely

Read mapping equally in two repetitive regions:

- MAPQ = 0
- Generate FP variant calls

Which aligner should I use for aligning reads against a resistance gene database for determining with resistance genes I have in my sample?


Homologus/repetitive region
Resistance gene - Allele 1
Resistance gene - Allele 2
Resistance gene - Allele 3

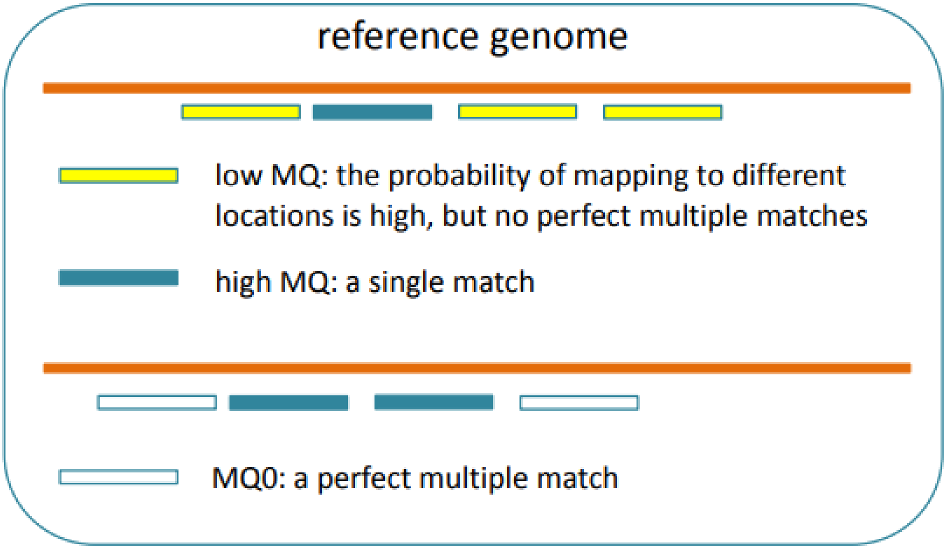
Reads mapping to the repetitive/homologus region map against all alleles.
We allow one read to map to **several locations**.

Reads mapping uniquely only map in Allele 1. Which is the one more **covered**

MAPQ

- What if there are several possible places to align your sequencing read? This may be due to:
  - Repeated elements in the genome
  - Low complexity sequences
  - Reference errors and gaps

MQ is a phredScore of the quality of the alignment



reference genome

low MQ: the probability of mapping to different locations is high, but no perfect multiple matches

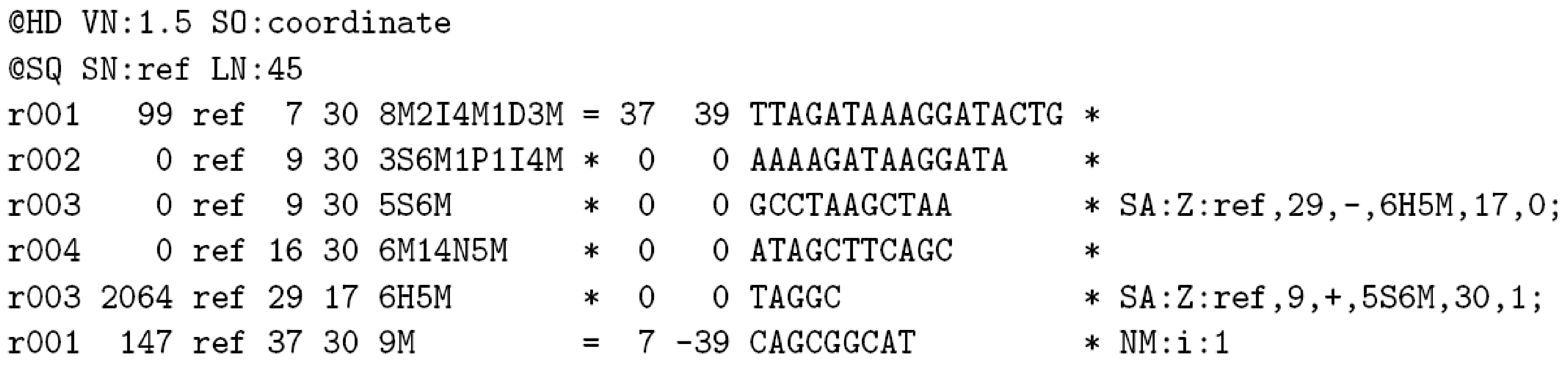
high MQ: a single match

MQ0: a perfect multiple match

SAM format

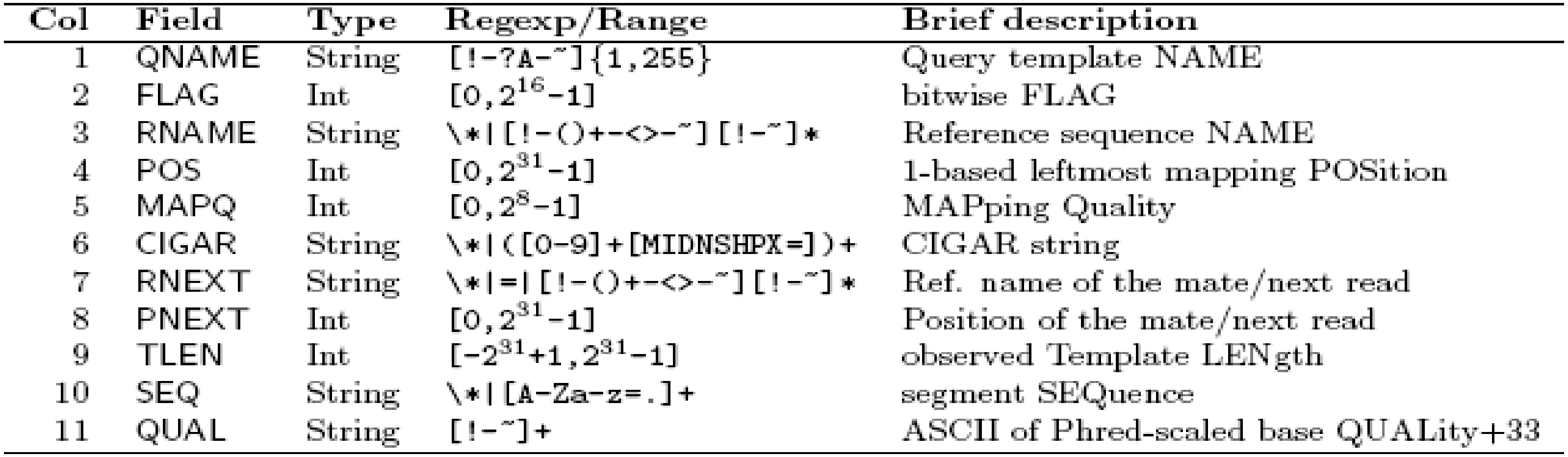
Definición:

It’s a specification that defines a generic format for storing nucleotide alignments. It describes a query alignment against a reference genome.

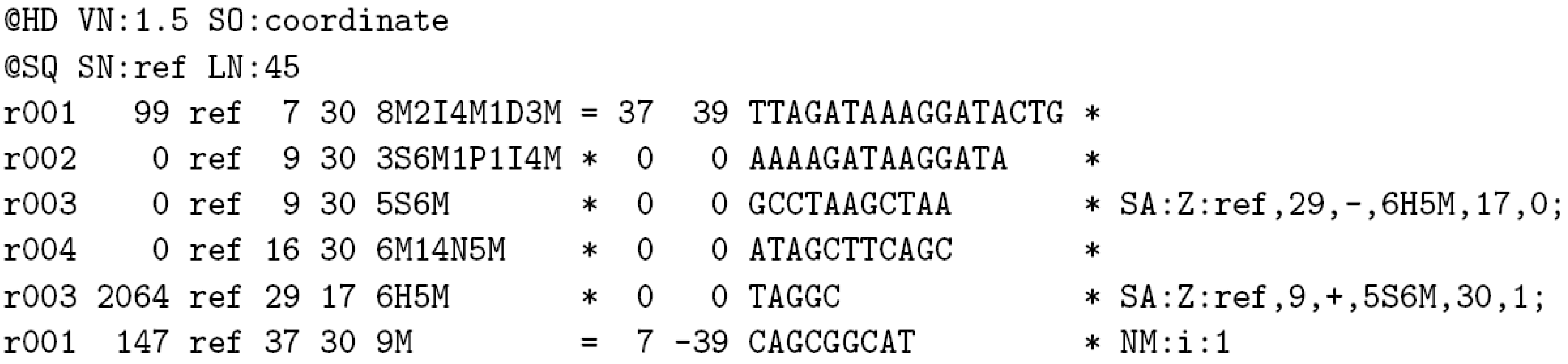


```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
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
```

SAM format

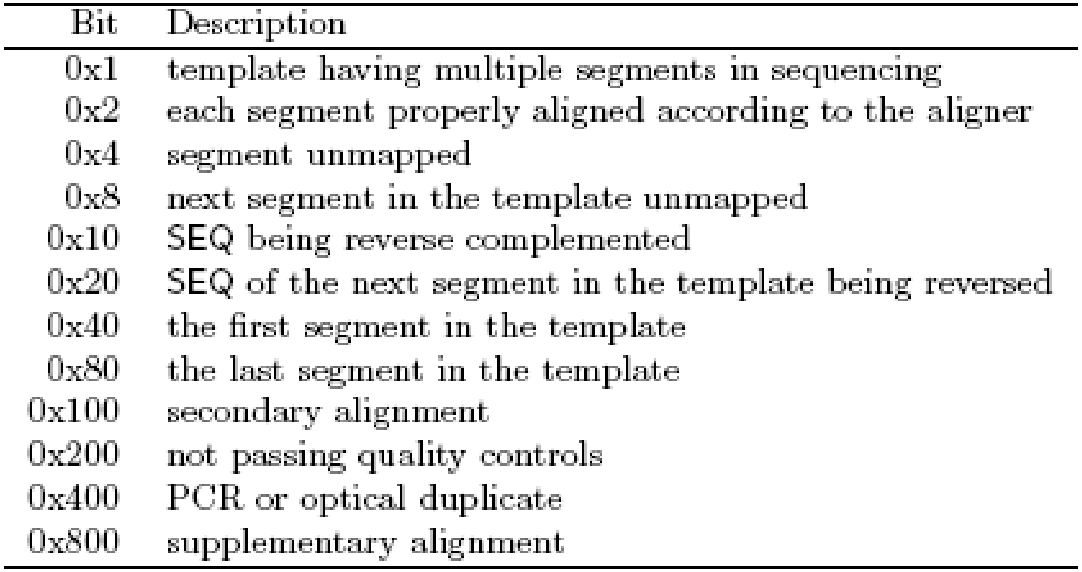


<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,255}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>[0,2^{16}-1]</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-<>-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>[0,2^{31}-1]</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>[0,2^{8}-1]</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-<>-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>[0,2^{31}-1]</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>[-2^{31}+1,2^{31}-1]</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>



```
@HD VN:1.5 SO:coordinate
@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
```

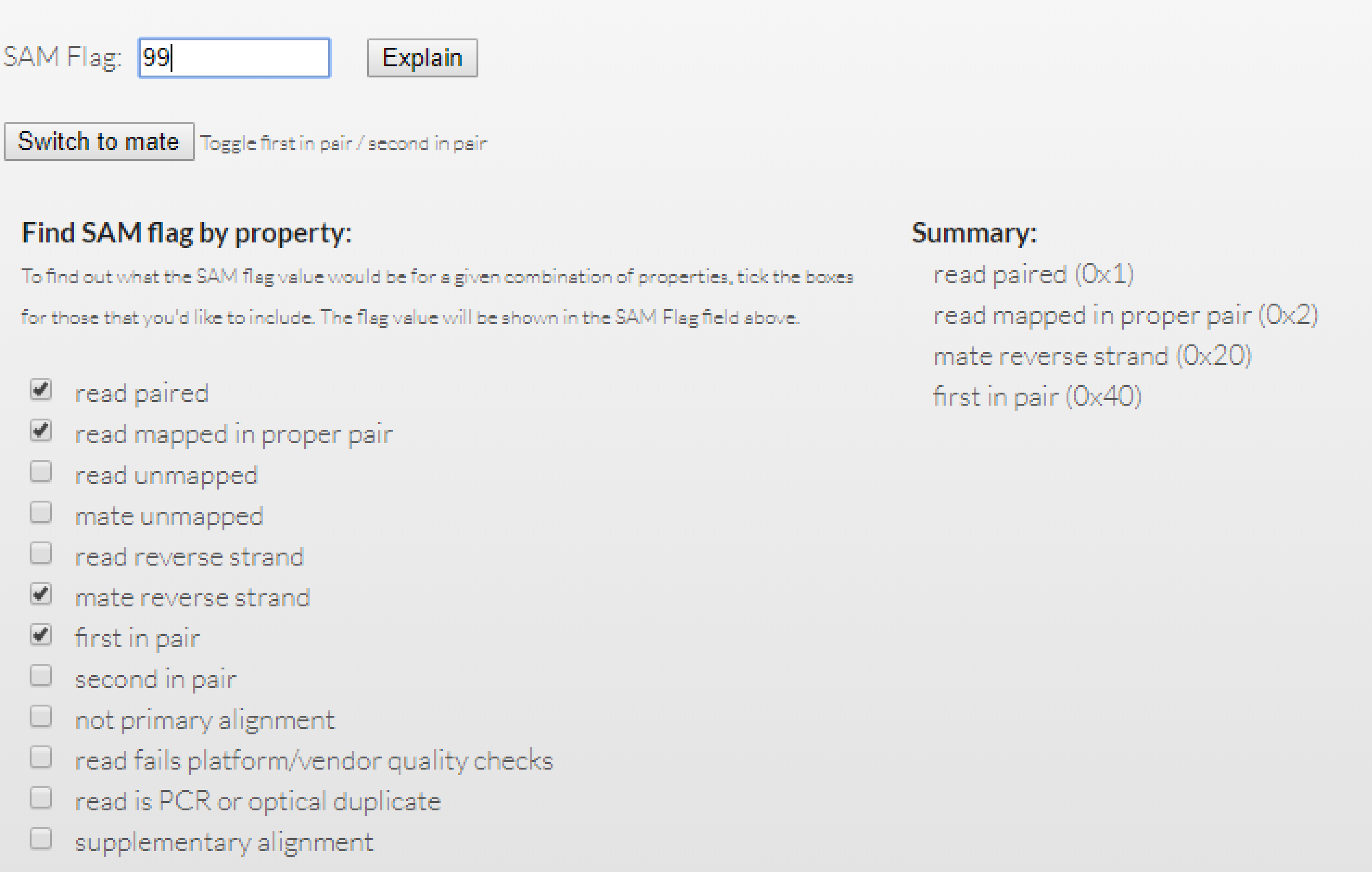
SAM format: flags



Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

https://broadinstitute.github.io/picard/explain-flags.html

Flag explanation example 1



SAM Flag: 99

Toggle first in pair / second in pair

Find SAM flag by property:

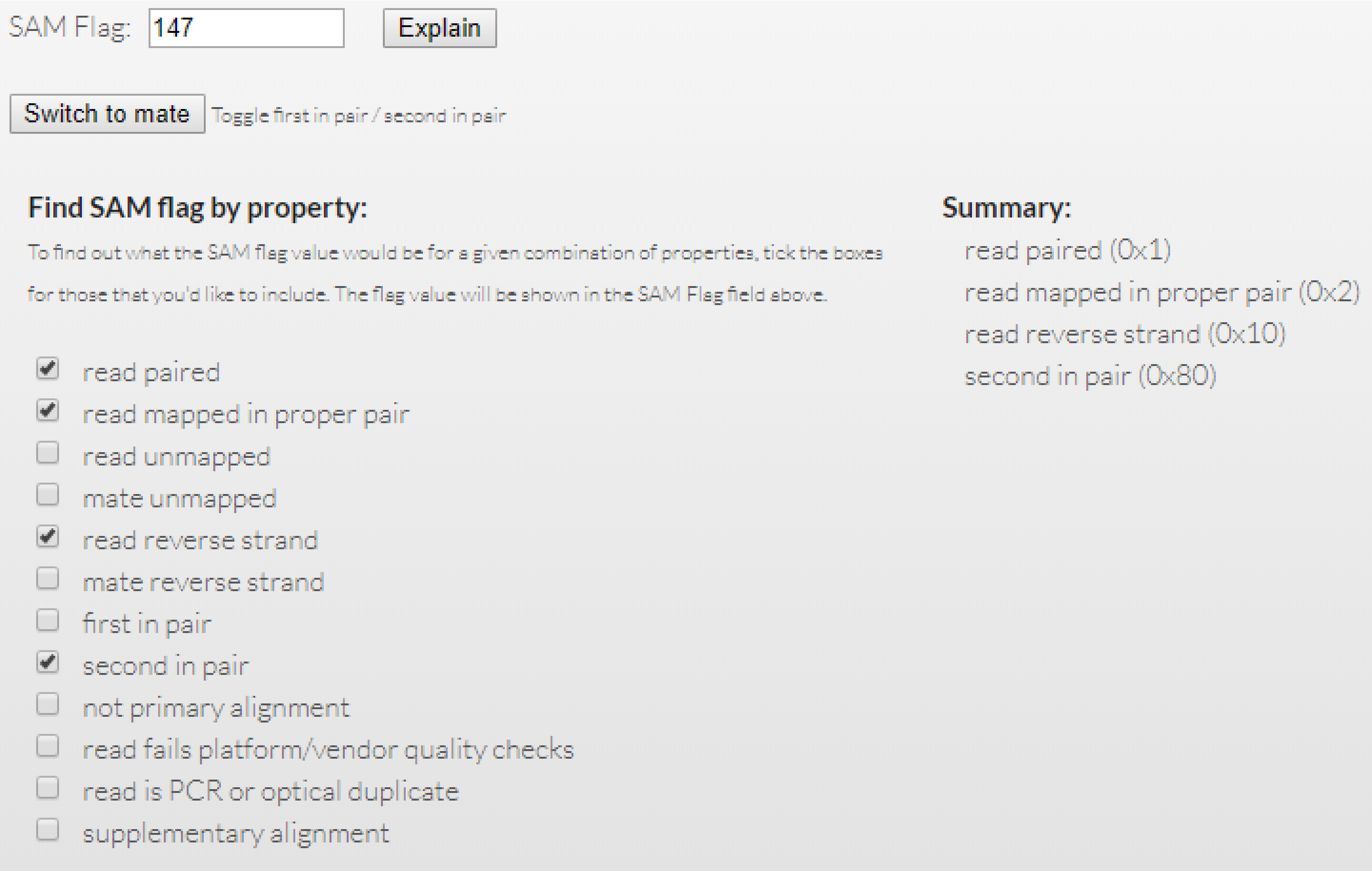
To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ☑ read paired
- ☑ read mapped in proper pair
- □ read unmapped
- □ mate unmapped
- □ read reverse strand
- ☑ mate reverse strand
- ☑ first in pair
- □ second in pair
- □ not primary alignment
- □ read fails platform/vendor quality checks
- □ read is PCR or optical duplicate
- □ supplementary alignment

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
first in pair (0x40)

Flag explanation example 2



SAM Flag: 147 Explain

Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

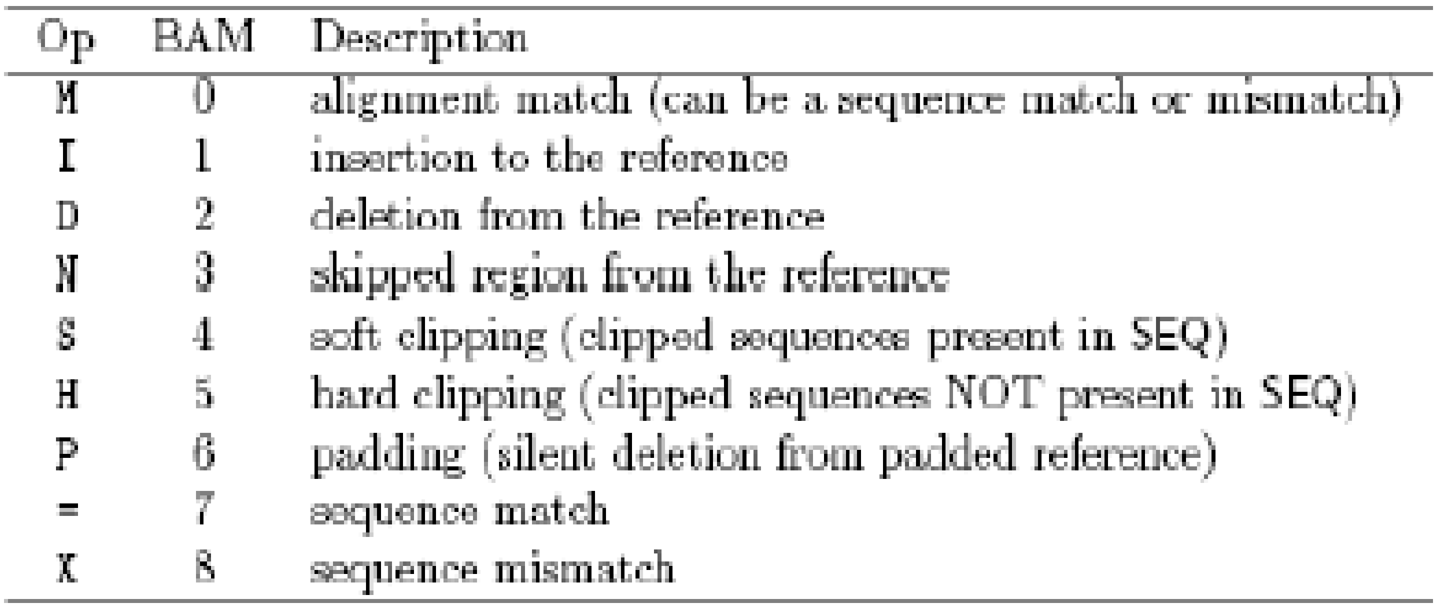
To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ☑ read paired
- ☑ read mapped in proper pair
- □ read unmapped
- □ mate unmapped
- ☑ read reverse strand
- □ mate reverse strand
- □ first in pair
- ☑ second in pair
- □ not primary alignment
- □ read fails platform/vendor quality checks
- □ read is PCR or optical duplicate
- □ supplementary alignment

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
read reverse strand (0x10)
second in pair (0x80)

SAM format: CIGAR string



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

SAM vs BAM format

- SAM and BAM format are exactly the same.
  - SAM is a tabular plain text file.
  - BAM is its binary format. Binary meaning is in a compress format not human readable.
  - We **MUST** always use BAM format because it is optimized for computer-reading

AND

BECAUSE IT SAVES A LOT OF DISK SPACE!!

Typical bam and sam format files weights from a S. grumpensis

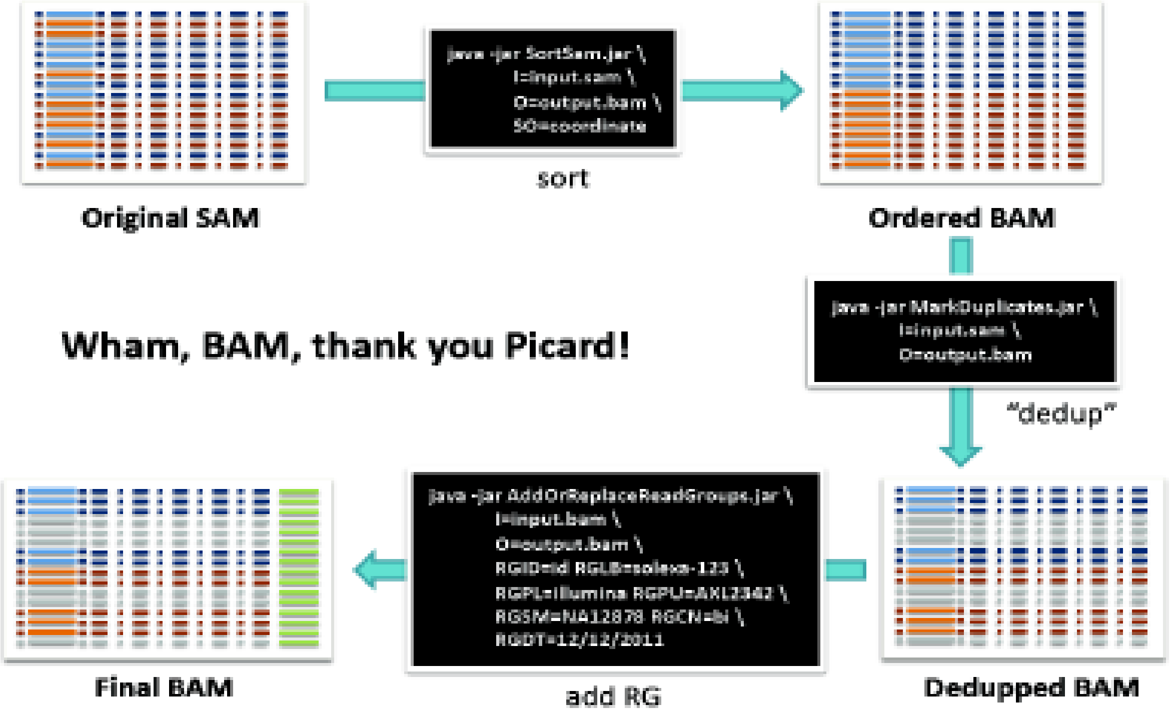
SAM format file: 3.6 GB

BAM format file: 689 M

Duplicate filter

- Duplicates are non-independent measurements of a sequence
  - Sampled from the exact same template of DNA
  - Violates assumptions of variant calling
- Errors in sample/library prep will get propagated to all the duplicates
- Just pick the “best” copy – mitigates the effects of errors
- **Definition**: sequences starting and finishing in the exact same coordinates. Both pairs if paired-end.

Duplicate filter



```
java -jar SortSam.jar \
        I=input.sam \
        O=output.bam \
        SO=coordinate
```

```
java -jar MarkDuplicates.jar \
        I=input.sam \
        O=output.bam
```

```
java -jar AddOrReplaceReadGroups.jar \
        I=input.bam \
        O=output.bam \
        RGID=id RGLB=solexa-123 \
        RGPL=illumina RGPU=AXL2342 \
        RGSM=NA12878 RGCN=bi \
        RGDT=12/12/2011
```

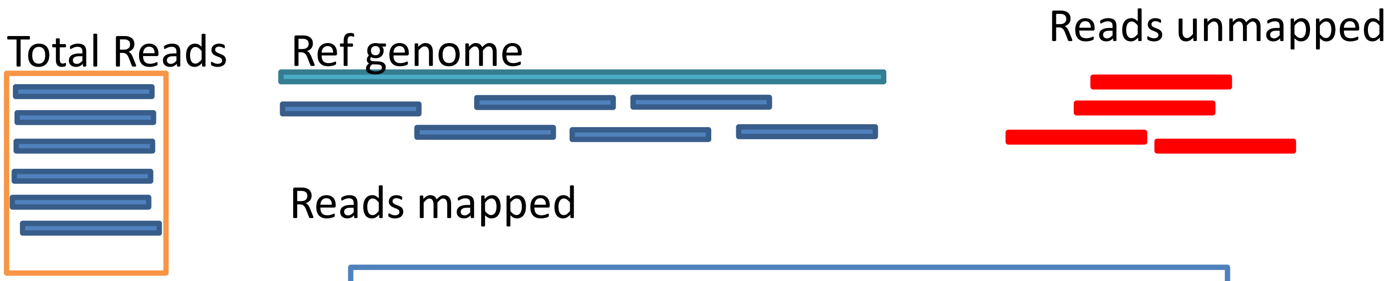
Mapping statistics

- % mapped: reads mapped/total reads
- % unmapped: reads unmapped/total reads
- % duplicates: reads belonging to same template/total reads
- Mean depth of coverage
- Coverage: % genome with at least one read mapped.

Mapping quality control

- **% mapping:** number of reads mapping againts reference genome.

Picard
Samtools



% mapping = Reads mapped / Total Reads

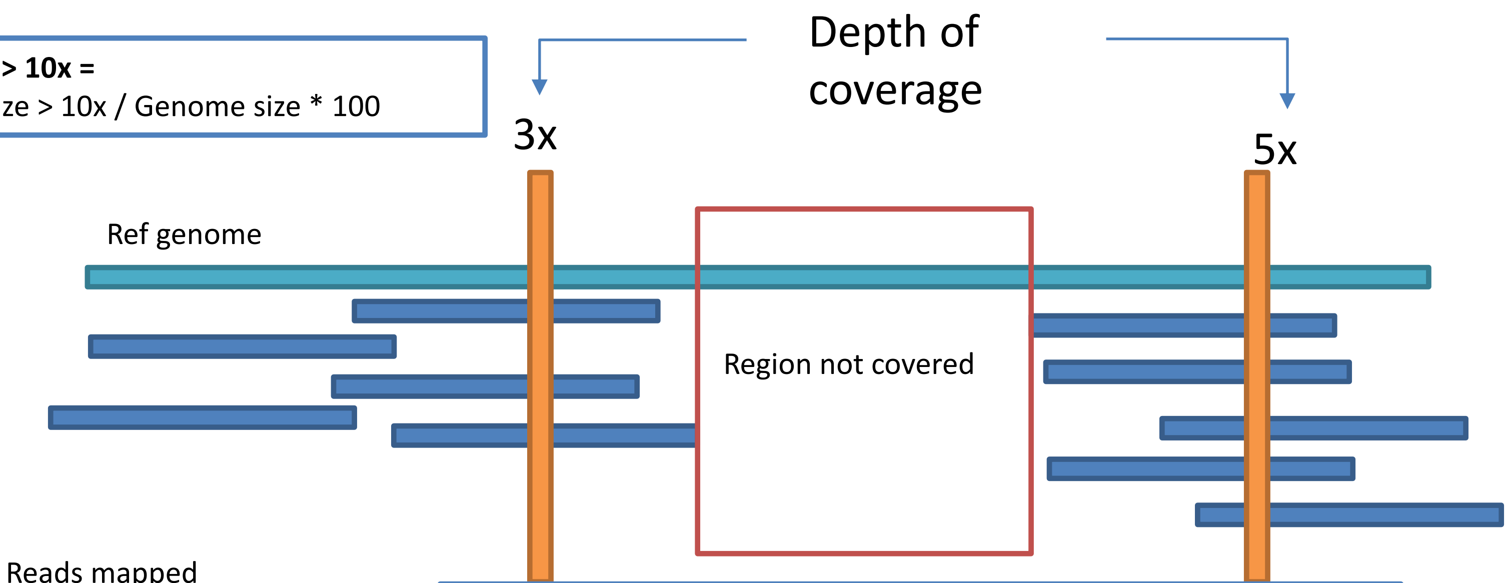
Mandatory parameter for microbial genomics!! It indicates us how many reads we have from our organism of interest. In human genomics this is almost always 99.99% unless something terrible happens. Not here!!!

Mapping quality control

- **% genome > 10x:** percentage of genome covered with more than 10 reads.
- **Mean Depth of coverage:** mean of reads covering a genome position.

Picard
Samtools

% genome > 10x =
Genome size > 10x / Genome size * 100



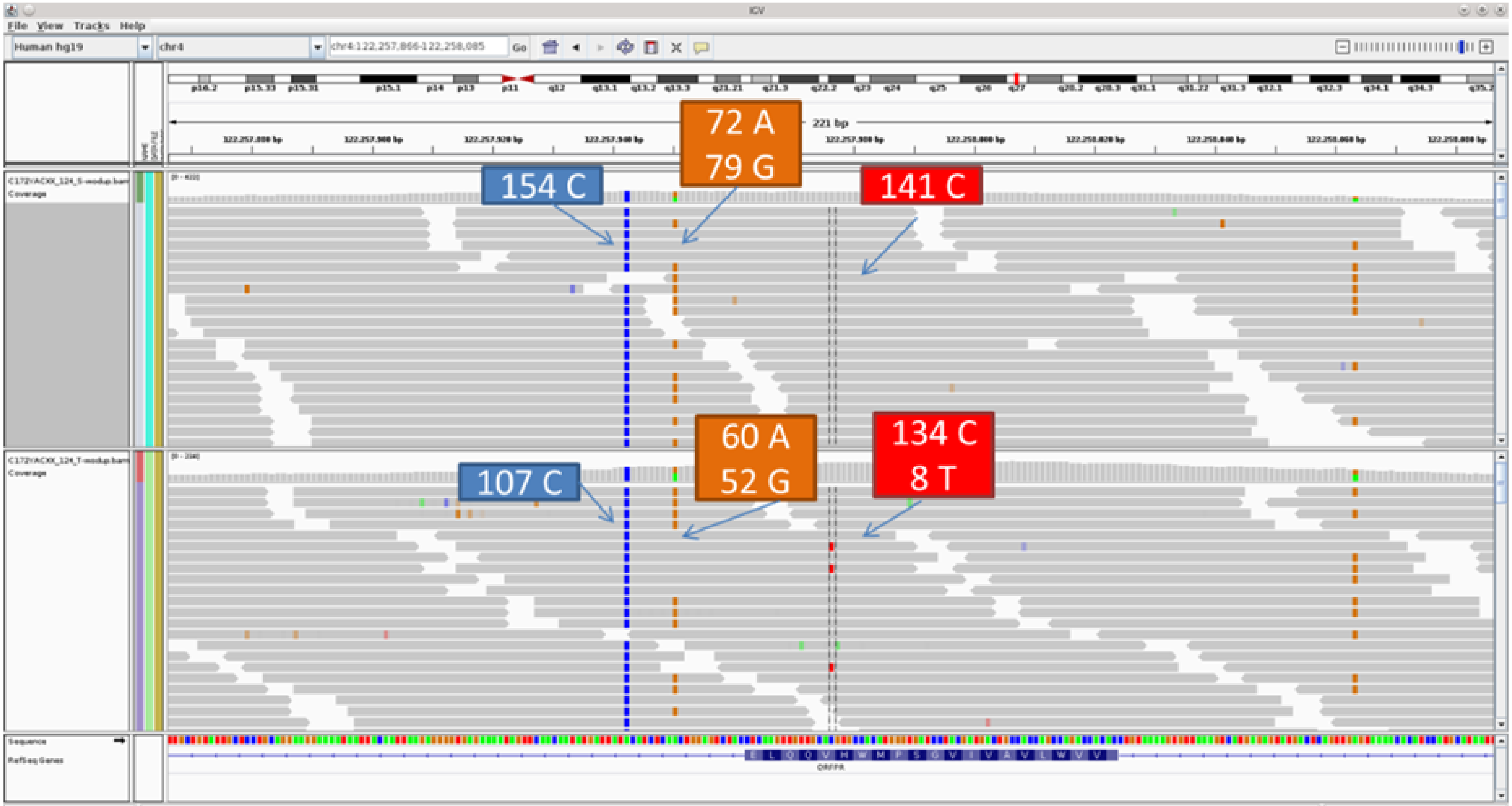
Mean Depth of coverage =
Depth of coverage in position 1, 2, 3 / Genome size

Variant Calling

- **Variant calling concept is simple:**

Find positions in our reads different from the reference.

- We start with our secuences mapped against our reference genome, and we walk trough every column of the alignment counting the number of alleles found and comparing them against the reference.



154 C

72 A
79 G

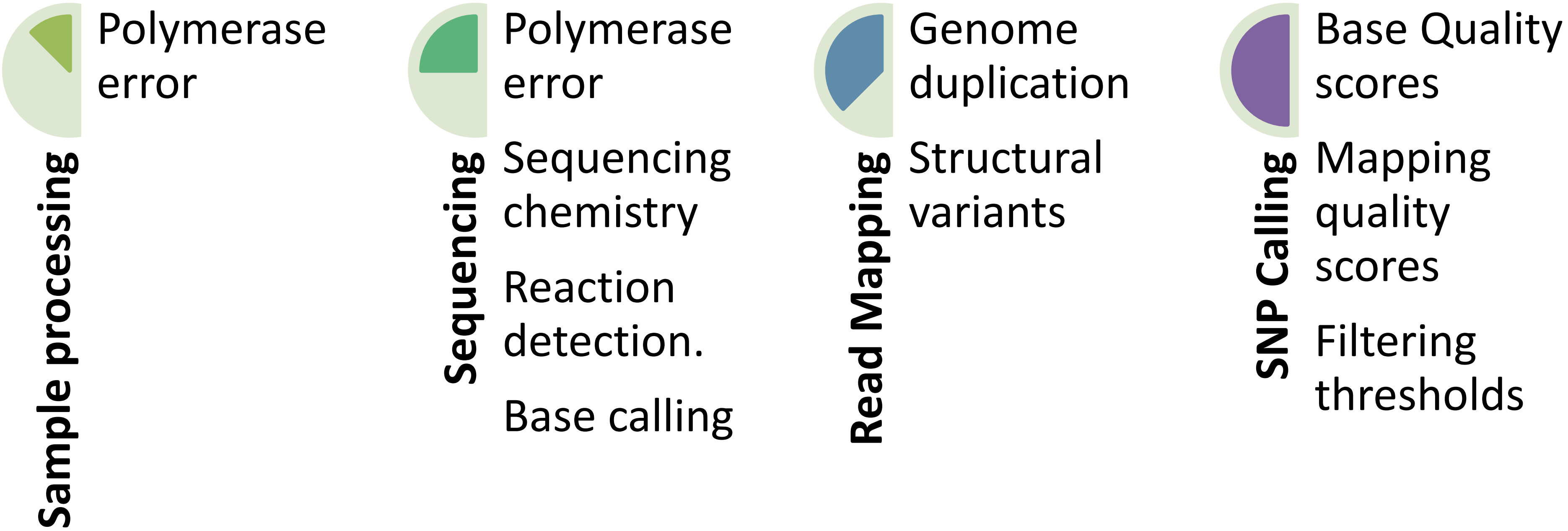
141 C

107 C

60 A
52 G

134 C
8 T

Sources of error and mitigation strategies



Sample processing

Polymerase error

Sequencing

Polymerase error

Sequencing chemistry

Reaction detection.

Base calling

Read Mapping

Genome duplication

Structural variants

SNP Calling

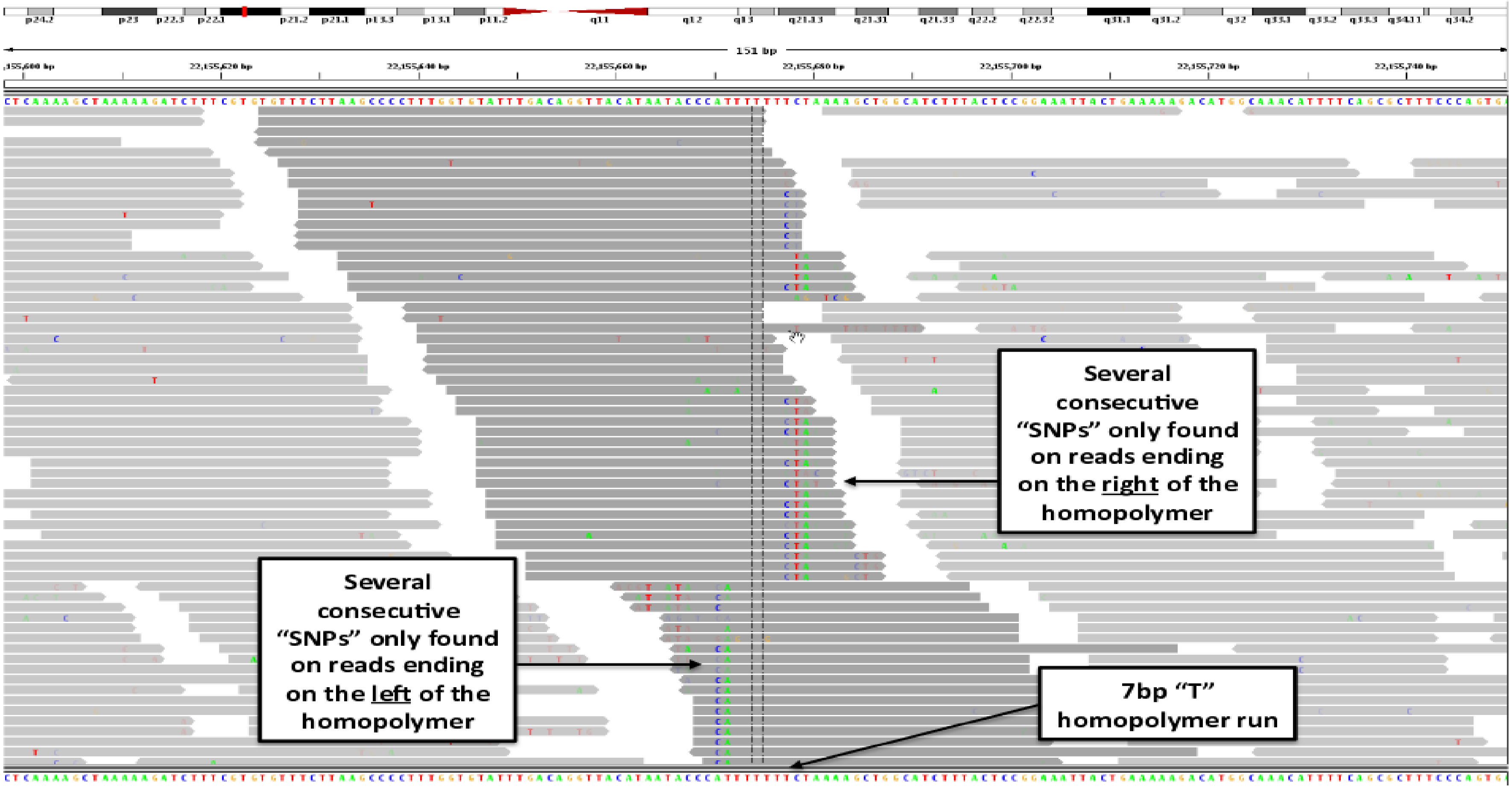
Base Quality scores

Mapping quality scores

Filtering thresholds

Adapted from Olson et al. Frontiers in Genetics. 2015

Sources of error and mitigation strategies

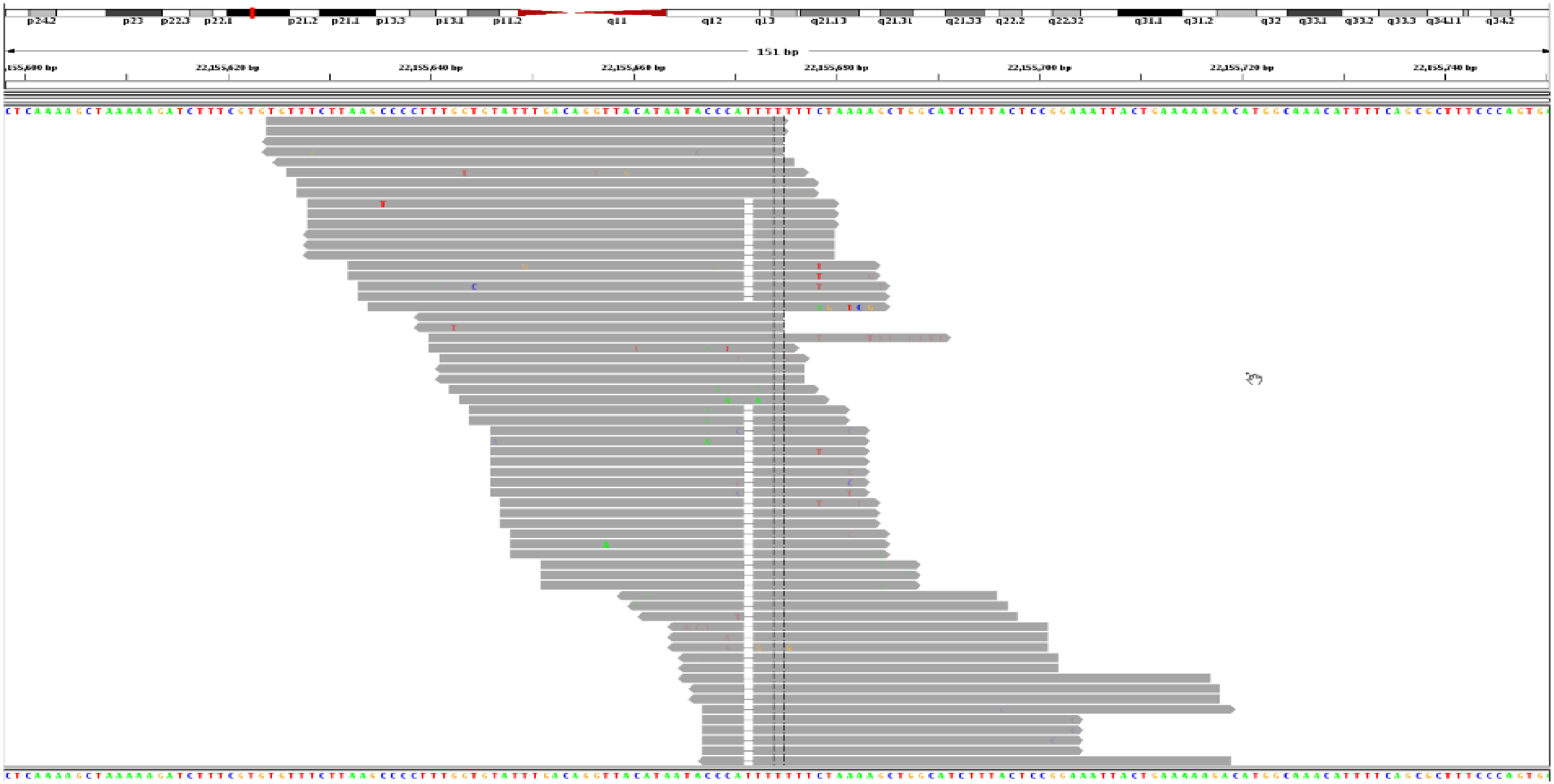


Several consecutive "SNPs" only found on reads ending on the right of the homopolymer

Several consecutive "SNPs" only found on reads ending on the left of the homopolymer

7bp "T" homopolymer run

Sources of error and mitigation strategies



Reference selection

- Critial step <- Bias which SNPs are called.
- SNPs in genes not present in the reference **WON’T** be called.
- Less effect in clonal bacteria.
- Number of SNPs called vary **A LOT!**

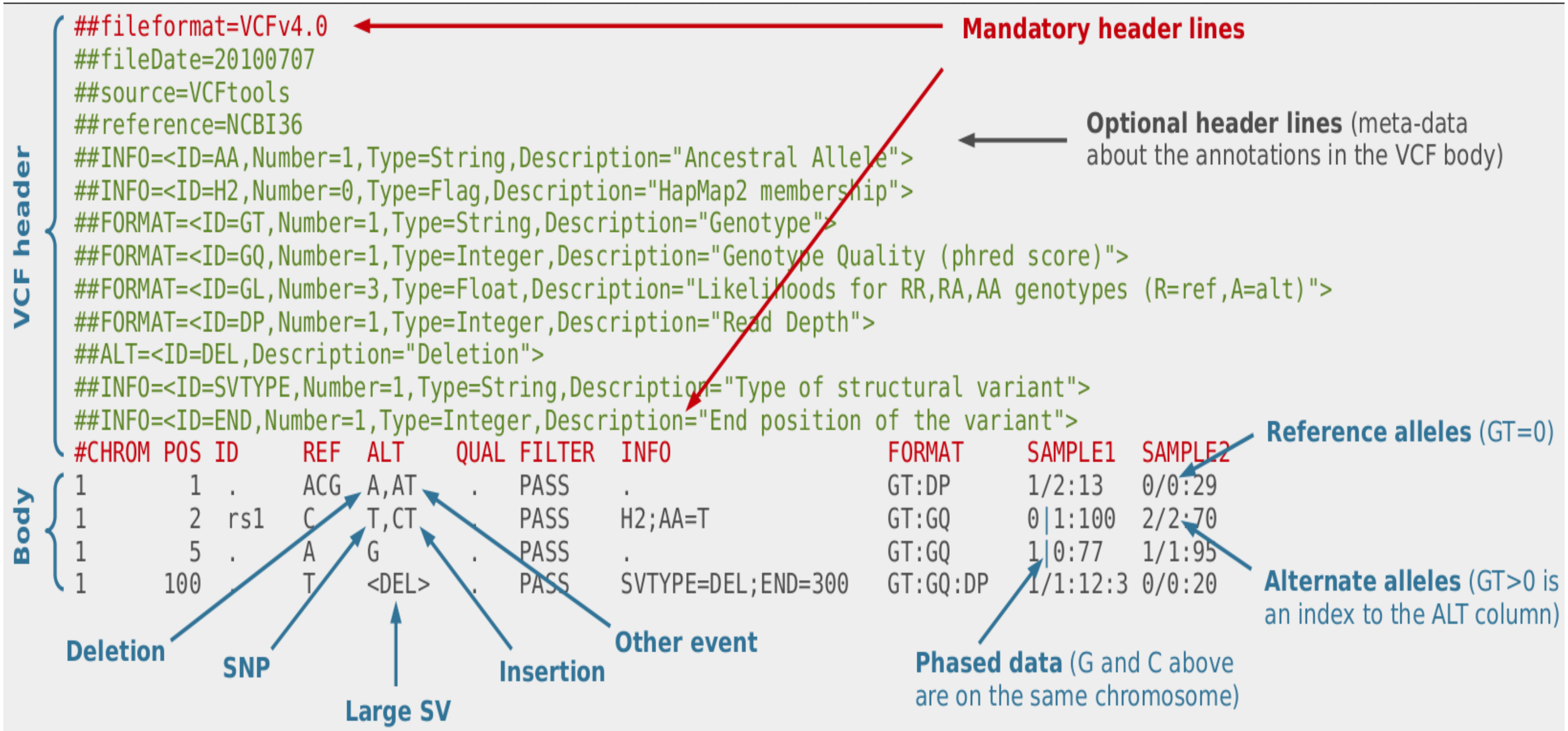
- **Solutions:**
  - Kmerfinder

Repetitive/Phage regions filtering

- **PHASTER**

- We can remove/mask phague/repetitive regions where reads won’t map.
- This way those areas will be out of analysis.
- Problem: those areas could be important!

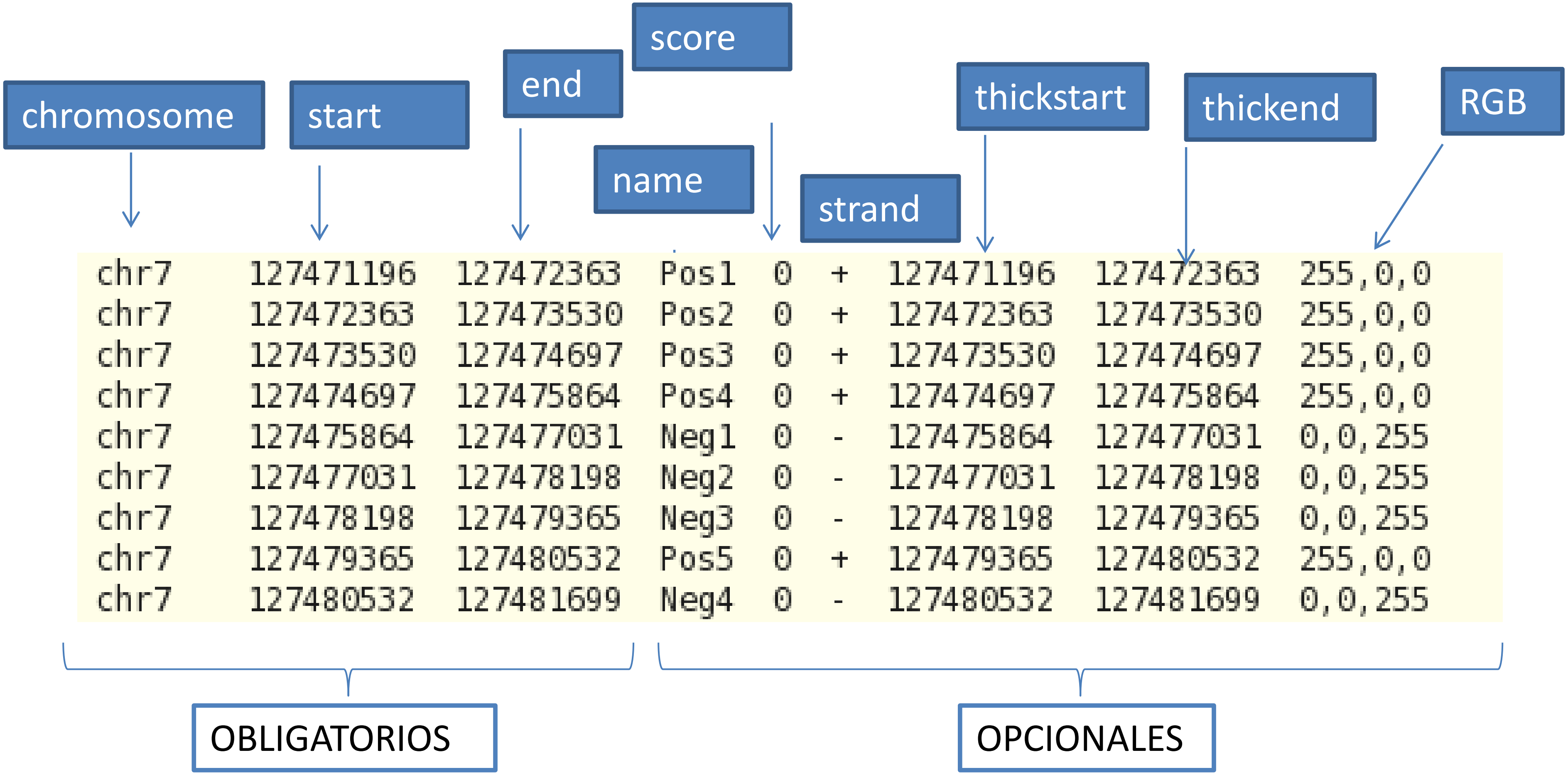
VCF format



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID      REF  ALT     QUAL FILTER  INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1 .       ACG  A,AT    .    PASS    .                    GT:DP      1/2:13   0/0:29
1        2 rs1     C    T,CT    .    PASS    H2;AA=T              GT:GQ      0|1:100  2/2:70
1        5 .       A    G       .    PASS    .                    GT:GQ      1|0:77   1/1:95
1      100 .       T    <DEL>   .    PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3 0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Bed format



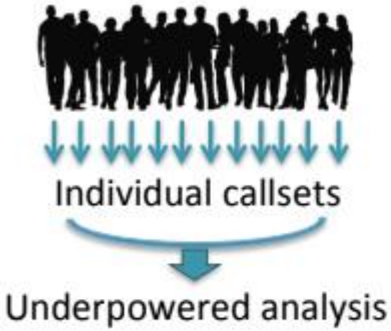
```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

Pipelines for bacterial SNP-based analysis

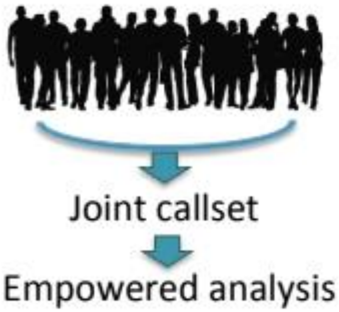
Software	Description	Other	References
CFSAN	VARSCAN variant calling	Terminal	Davis et al., 2015
NASP	Variant calling with VarScan, solSNP,samtools and GATK	Terminal	Sahl et al., 2016
Lyve-Set	VARSCAN variant calling	Terminal	Katz et al., 2017
KSNP	Reference free variant calling.	Terminal	Gardner et al., 2015
SNVPhyl	Variant calling with freebayes and samtools	Galaxy	Petkau et al., 2017
Snippy	Variant calling with freebayes and snp matrix generation	Terminal	Tseeman et al. (github)
CSI phylogeny	Variant calling with samtools.	Web	Kaas et al., 2014

Cohorts need to be analyzed together at variant calling step

- If we simply call variants on individual samples then merge lists of their variants, we miss a lot of important information

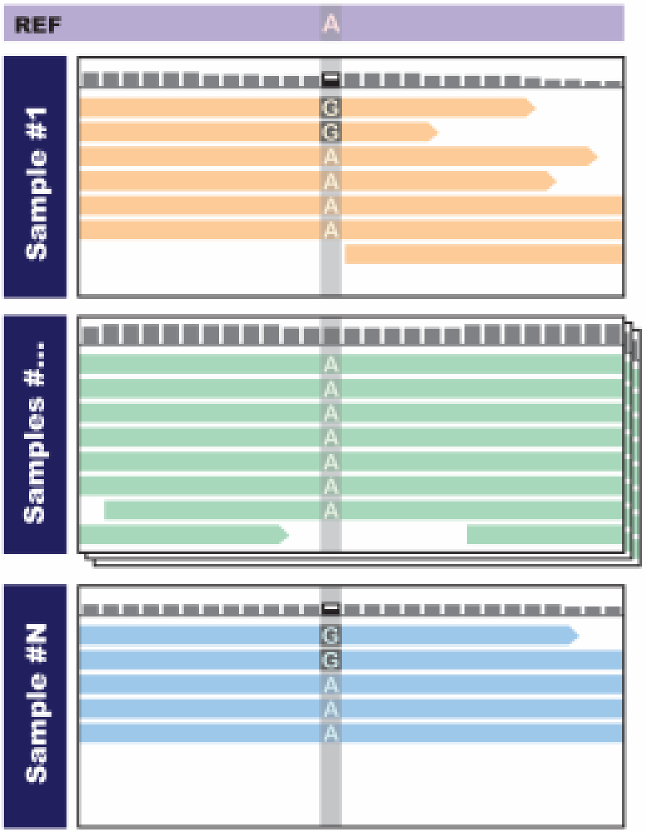


- Joint variant discovery rescues a lot of valuable information



Sequencing and variant calling pipelines MPG Primer @ Broad Institute Cambridge, 15 October, 2015

Joint analysis empowers calls in difficult sites

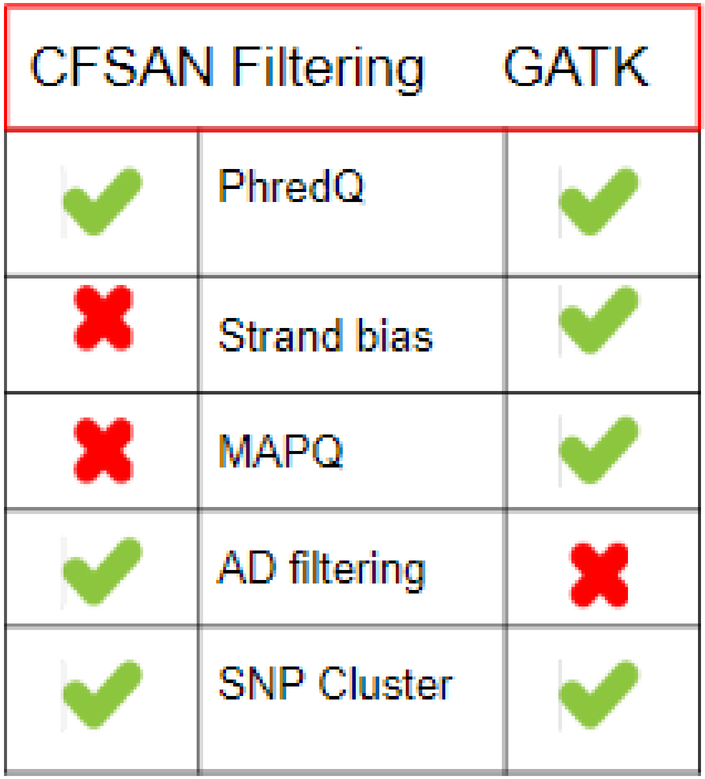


- If we analyze Sample #1 or Sample #N alone we are not confident that the variant is real

- If we see both samples then we are more confident that there is real variation at this site in the cohort

Sequencing and variant calling pipelines MPG Primer @ Broad Institute Cambridge, 15 October, 2015

High Quality SNP selection



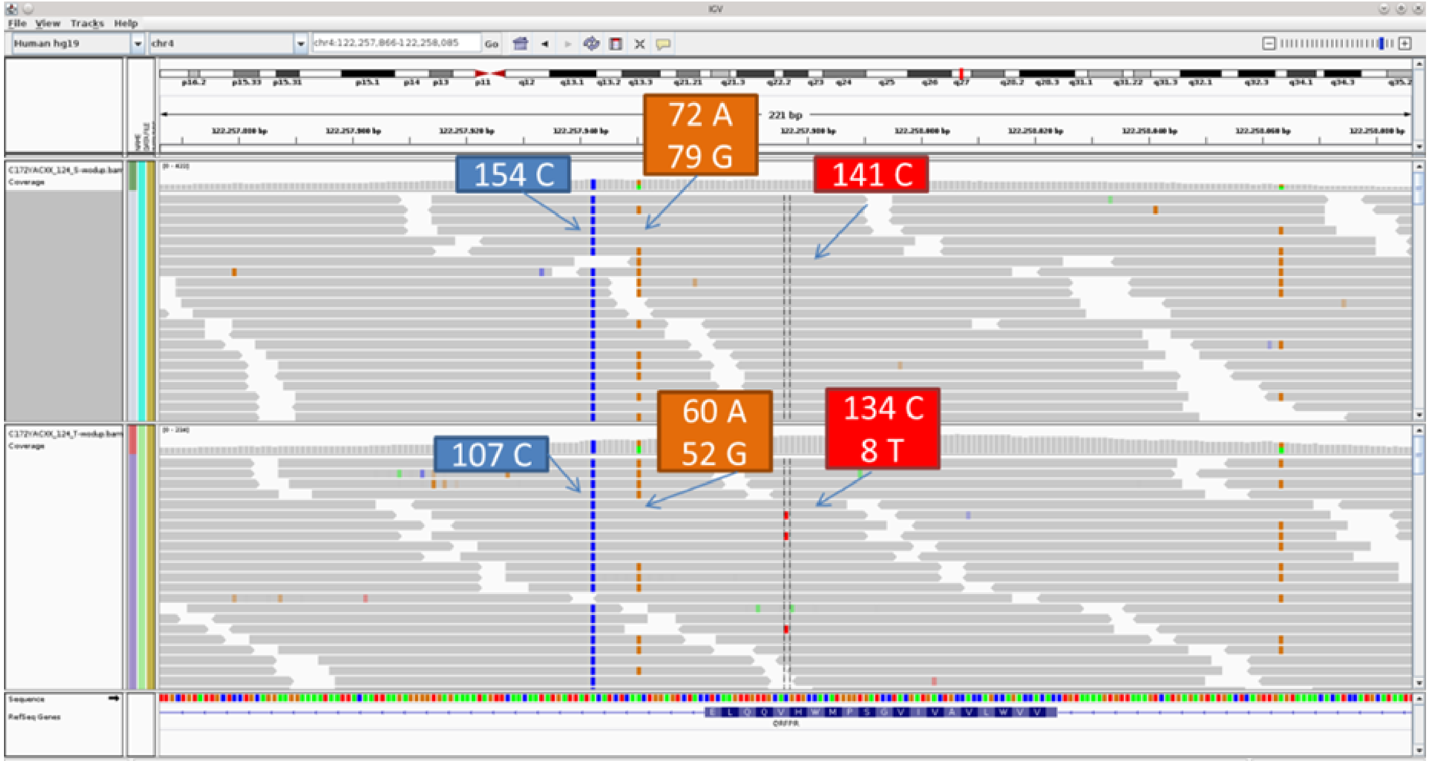
CFSAN Filtering		GATK
✓	PhredQ	✓
✗	Strand bias	✓
✗	MAPQ	✓
✓	AD filtering	✗
✓	SNP Cluster	✓

Population Allele frequency vs Sample Allele frequency

- **Population allele frequency:** probability of finding an allele in the population. Number of individuals carrying an allele vs total of individuals in the population.

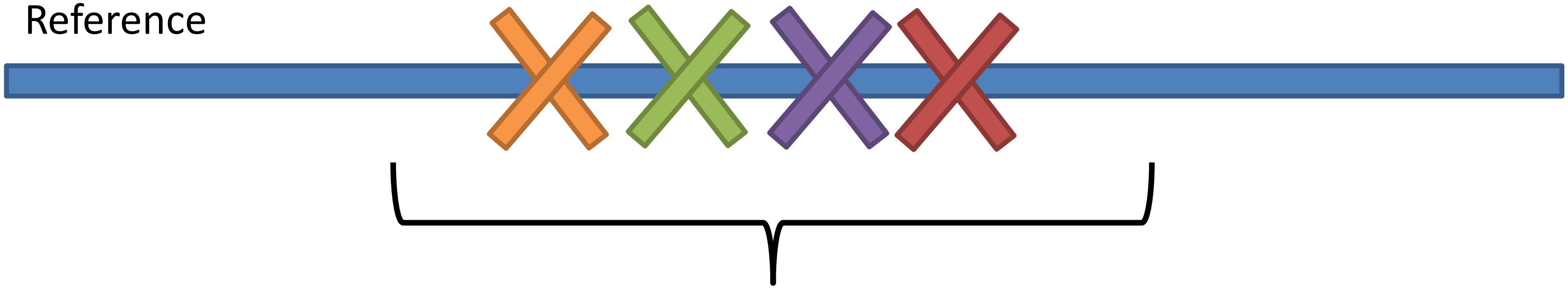
Population Allele frequency vs Sample Allele frequency

- **Alternate/Base allele frequency**: number of reads supporting the alternate allele vs total of reads.



SNP cluster filtering

Acummulation of more than 3 SNPs in 1000 pb



1000 pb window

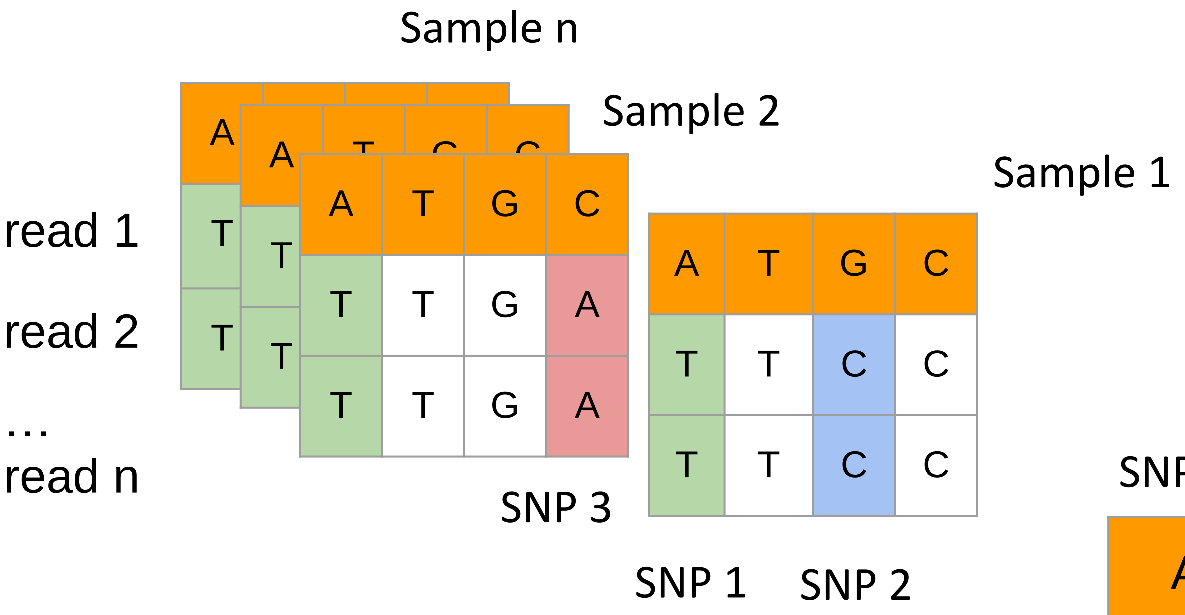
What’s next?

SNP matrix creation

And

Phylogeny!

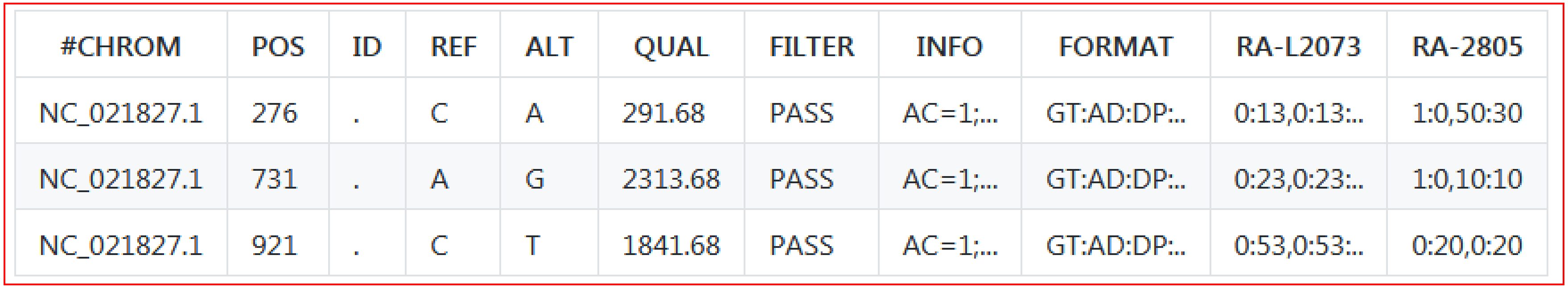
Building a SNP matrix



	SNP 1	SNP 2	SNP 3
Reference	A	G	C
Sample 1	T	C	C
Sample 2	T	G	A
Sample n			

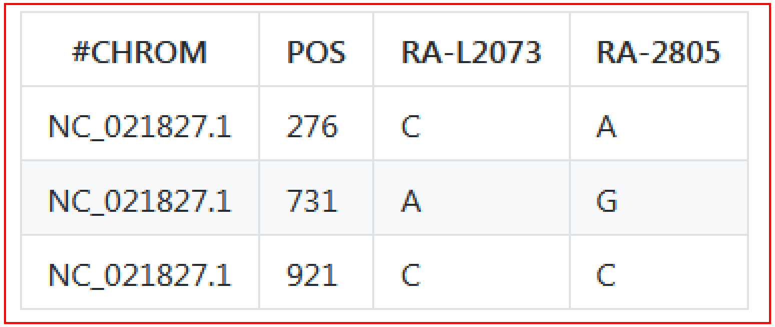
Building a SNP matrix

- Once we have our multisample vcf:



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	RA-L2073	RA-2805
NC_021827.1	276	.	C	A	291.68	PASS	AC=1;...	GT:AD:DP:..	0:13,0:13:..	1:0,50:30
NC_021827.1	731	.	A	G	2313.68	PASS	AC=1;...	GT:AD:DP:..	0:23,0:23:..	1:0,10:10
NC_021827.1	921	.	C	T	1841.68	PASS	AC=1;...	GT:AD:DP:..	0:53,0:53:..	0:20,0:20

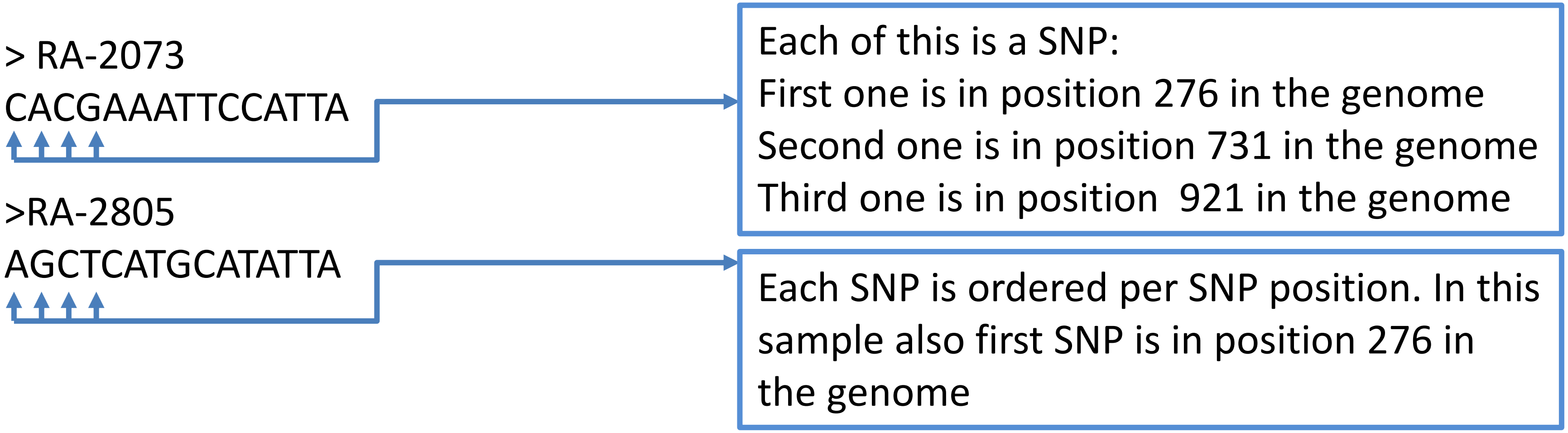
- We can generate the genotype for each sample



#CHROM	POS	RA-L2073	RA-2805
NC_021827.1	276	C	A
NC_021827.1	731	A	G
NC_021827.1	921	C	C

Building a SNP matrix

- So… now we have a simple multifasta, where each nucleotide represents a SNP.
- This means that even the nucleotide positions are sequentially in the fasta, they don’t have to be near each other in the genome!
- The SNP matrix file will look like this:



> RA-2073
CACGAAATTCCATTA

>RA-2805
AGCTCATGCATATTA

Each of this is a SNP:
First one is in position 276 in the genome
Second one is in position 731 in the genome
Third one is in position 921 in the genome

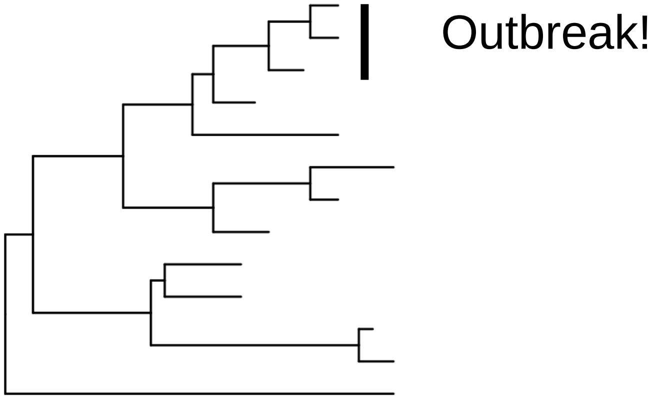
Each SNP is ordered per SNP position. In this sample also first SNP is in position 276 in the genome

Phylogeny

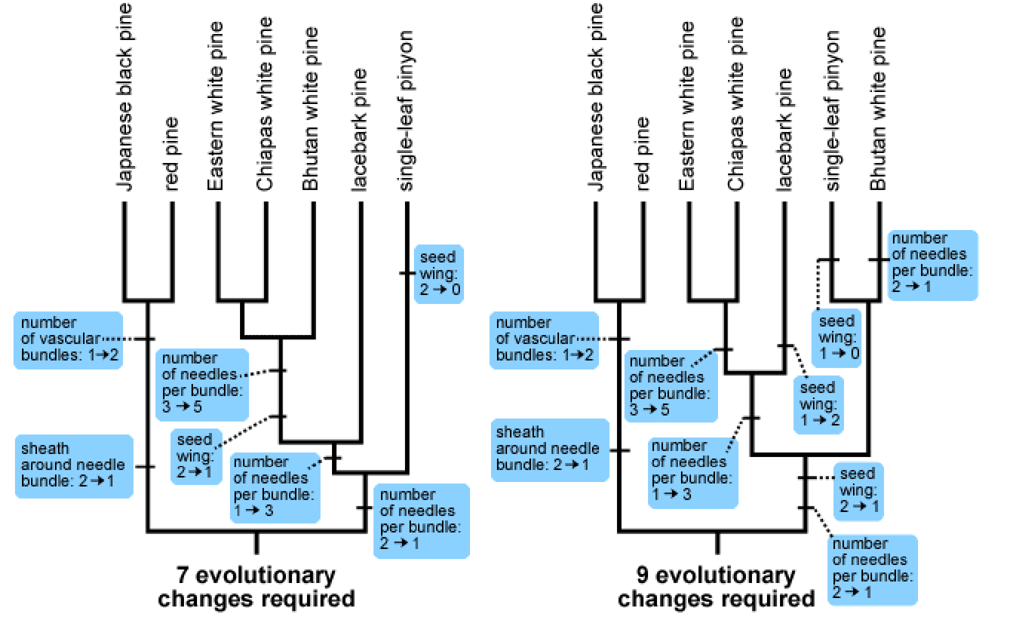
SNP matrix

SNP 1	SNP 2	SNP 3	
A	G	C	Reference
T	C	C	Sample 1
T	G	A	Sample 2
			Sample n

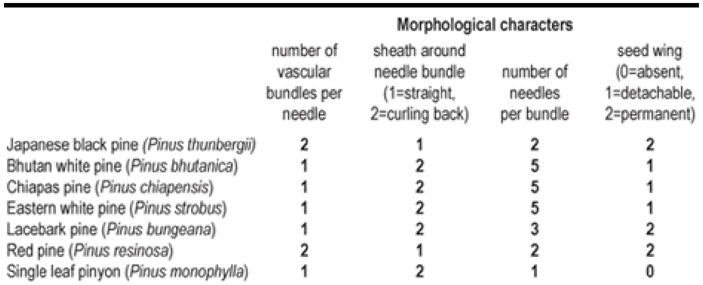
Phylogeny



Maximum parsimony



Japanese black pine
red pine
Eastern white pine
Chiapas white pine
Bhutan white pine
lacebark pine
single-leaf pinyon
seed wing: 2 → 0
number of vascular bundles: 1 → 2
number of needles per bundle: 3 → 5
sheath around needle bundle: 2 → 1
seed wing: 2 → 1
number of needles per bundle: 1 → 3
number of needles per bundle: 2 → 1
7 evolutionary changes required
Japanese black pine
red pine
Eastern white pine
Chiapas white pine
lacebark pine
single-leaf pinyon
Bhutan white pine
number of needles per bundle: 2 → 1
seed wing: 1 → 0
number of vascular bundles: 1 → 2
number of needles per bundle: 3 → 5
seed wing: 1 → 2
sheath around needle bundle: 2 → 1
number of needles per bundle: 1 → 3
seed wing: 2 → 1
number of needles per bundle: 2 → 1
9 evolutionary changes required

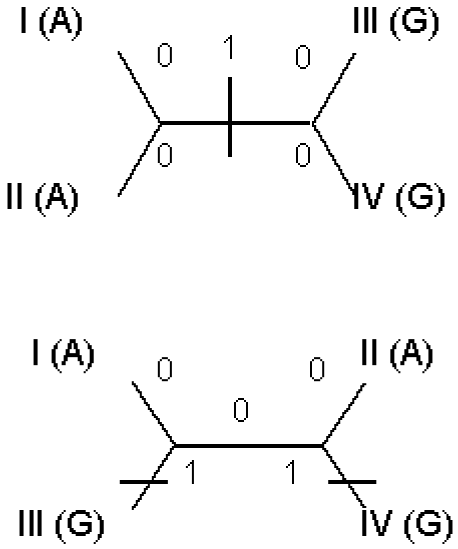


	Morphological characters			
	number of vascular bundles per needle	sheath around needle bundle (1=straight, 2=curling back)	number of needles per bundle	seed wing (0=absent, 1=detachable, 2=permanent)
Japanese black pine (*Pinus thunbergii*)	2	1	2	2
Bhutan white pine (*Pinus bhutanica*)	1	2	5	1
Chiapas pine (*Pinus chiapensis*)	1	2	5	1
Eastern white pine (*Pinus strobus*)	1	2	5	1
Lacebark pine (*Pinus bungeana*)	1	2	3	2
Red pine (*Pinus resinosa*)	2	1	2	2
Single leaf pinyon (*Pinus monophylla*)	1	2	1	0

- Search the most parsimonious tree
- The most simple hipothesis must be the correct.
- Search the tree that explains the relationships with the less changes possible.

Maximum Likelihood

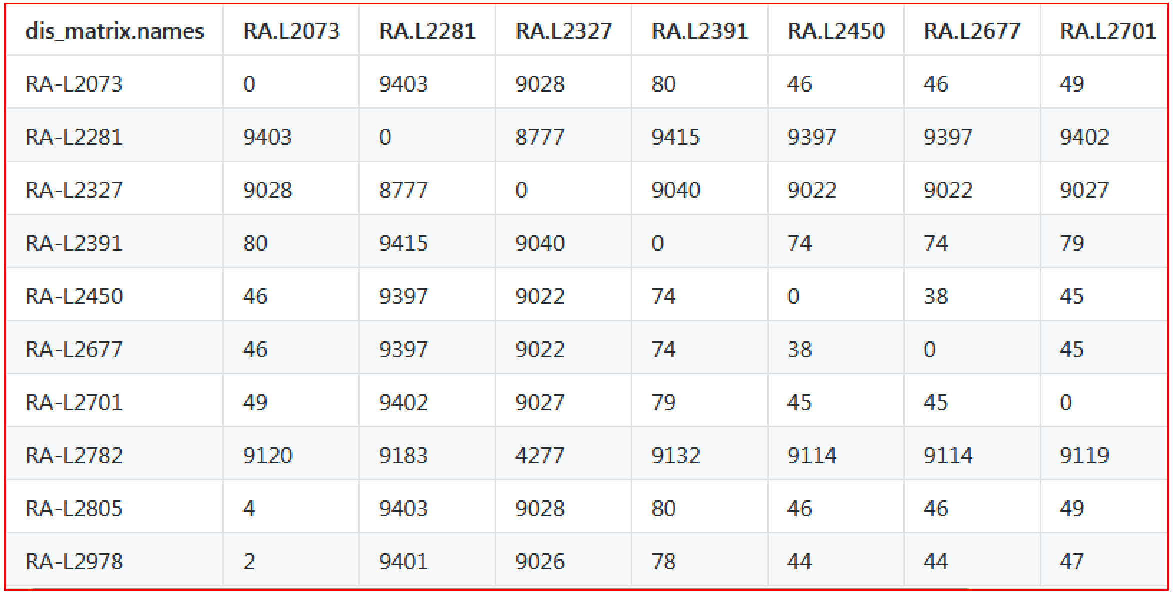
- Searchs the most likely tree given the data and based in a evolutionary model.
- More sofisticated.
- Not prepared a priori for snp matrix.
- RAxML
  - Heterogenity rate disabled.
  - Branchs indicate the expected number of substitution per site.



- 0,1 = differences along that branch
- Which hypothesis is more likely, given that the change is rare?

SNP distance

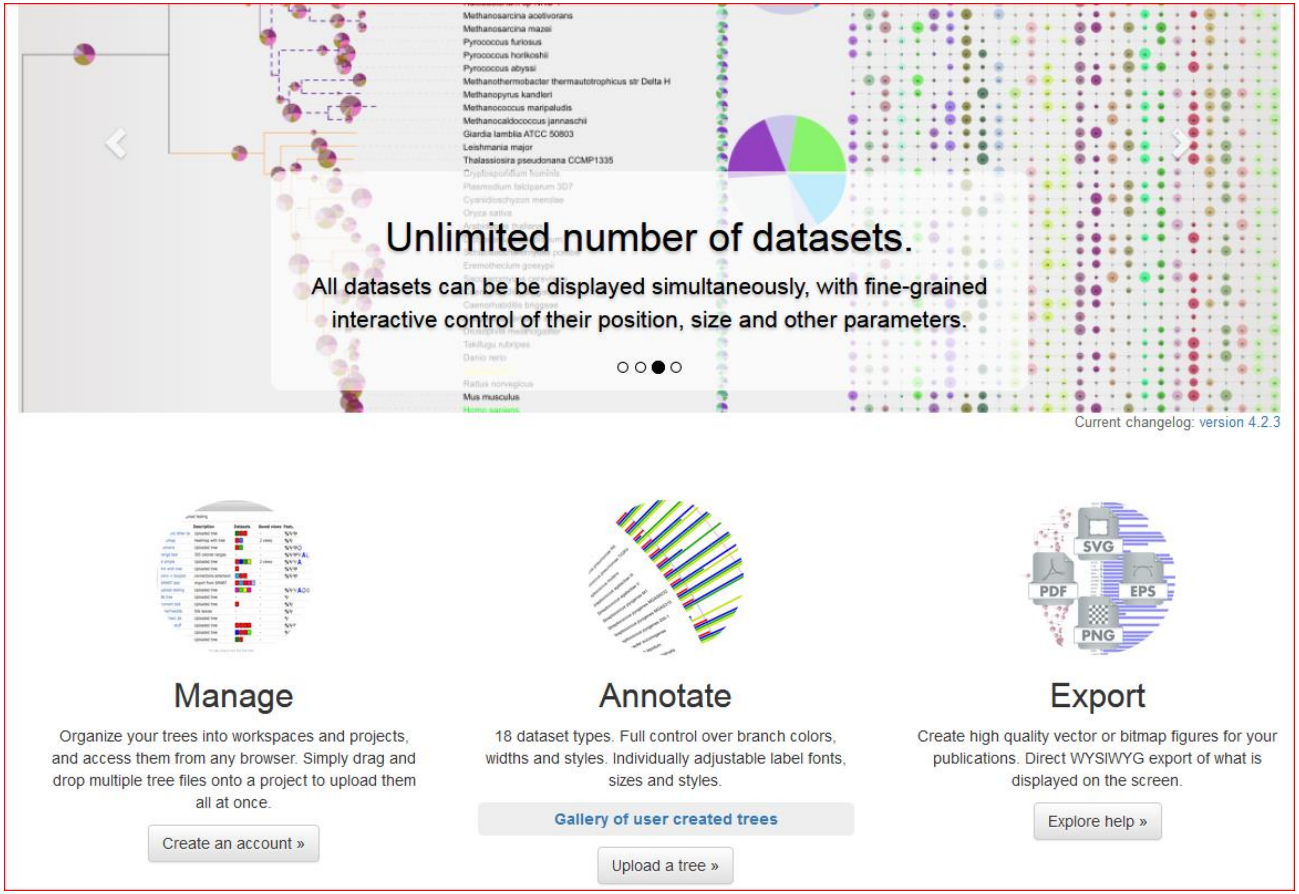
- SNP distance is calculated with N model. Simply the number of sites that differ between each pair of sequences.
- By default sites with at least one missing data is deleted for all sequences in R (complete deletion option in MEGA).



dis_matrix.names	RA.L2073	RA.L2281	RA.L2327	RA.L2391	RA.L2450	RA.L2677	RA.L2701
RA-L2073	0	9403	9028	80	46	46	49
RA-L2281	9403	0	8777	9415	9397	9397	9402
RA-L2327	9028	8777	0	9040	9022	9022	9027
RA-L2391	80	9415	9040	0	74	74	79
RA-L2450	46	9397	9022	74	0	38	45
RA-L2677	46	9397	9022	74	38	0	45
RA-L2701	49	9402	9027	79	45	45	0
RA-L2782	9120	9183	4277	9132	9114	9114	9119
RA-L2805	4	9403	9028	80	46	46	49
RA-L2978	2	9401	9026	78	44	44	47

iTOL

https://itol.embl.de/



Unlimited number of datasets.

All datasets can be be displayed simultaneously, with fine-grained interactive control of their position, size and other parameters.

Current changelog: version 4.2.3

Manage

Organize your trees into workspaces and projects, and access them from any browser. Simply drag and drop multiple tree files onto a project to upload them all at once.

Create an account »

Annotate

18 dataset types. Full control over branch colors, widths and styles. Individually adjustable label fonts, sizes and styles.

Gallery of user created trees

Upload a tree »

Export

Create high quality vector or bitmap figures for your publications. Direct WYSIWYG export of what is displayed on the screen.

Explore help »

How to interpret our phylogeny

- **Combination of:**
  - SNP counts
  - Tree topologies
  - Bootstrap support

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny

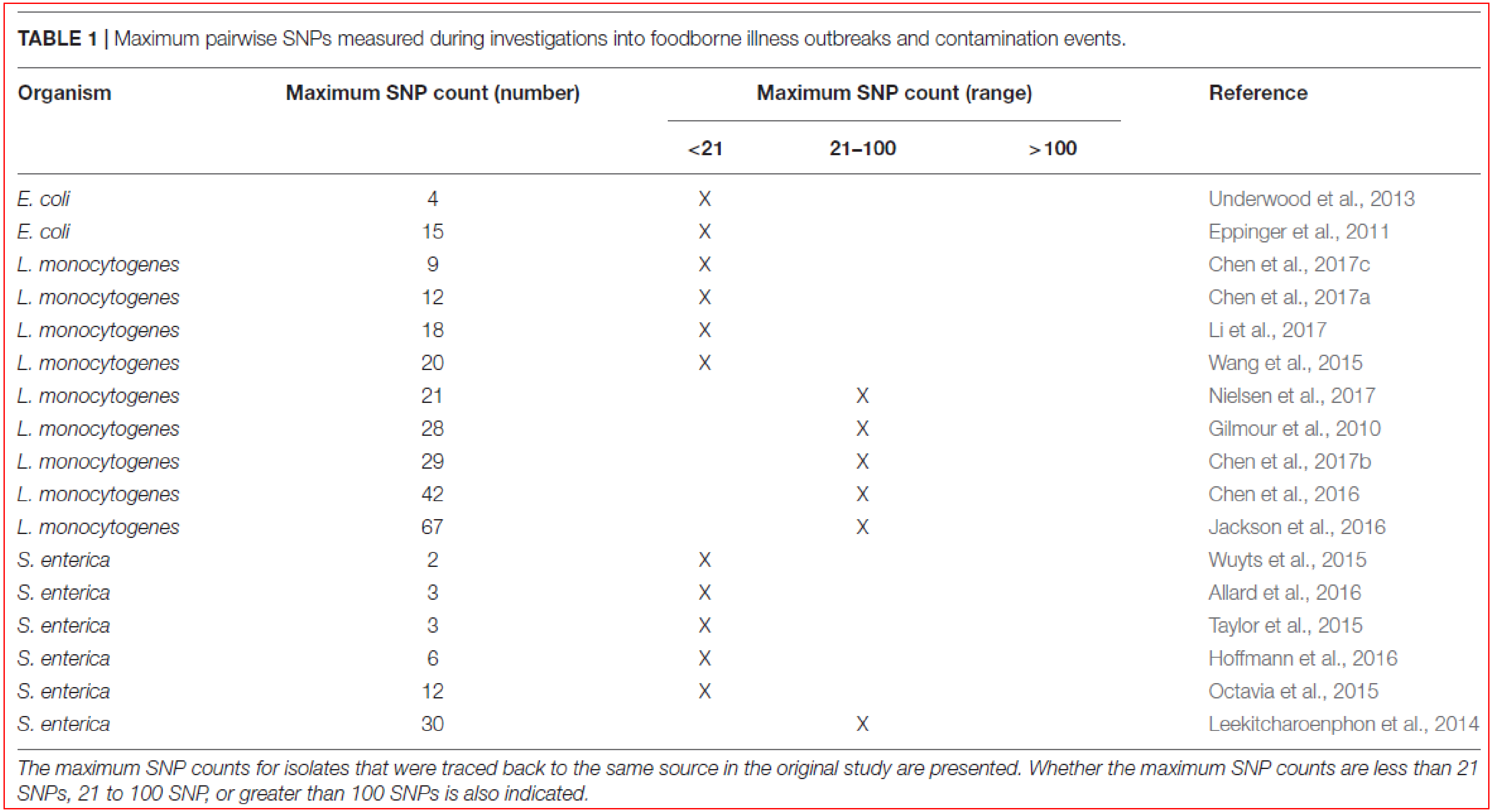


TABLE 1 | Maximum pairwise SNPs measured during investigations into foodborne illness outbreaks and contamination events.

Organism	Maximum SNP count (number)	Maximum SNP count (range)			Reference
		<21	21–100	>100	
E. coli	4	X			Underwood et al., 2013
E. coli	15	X			Eppinger et al., 2011
L. monocytogenes	9	X			Chen et al., 2017c
L. monocytogenes	12	X			Chen et al., 2017a
L. monocytogenes	18	X			Li et al., 2017
L. monocytogenes	20	X			Wang et al., 2015
L. monocytogenes	21		X		Nielsen et al., 2017
L. monocytogenes	28		X		Gilmour et al., 2010
L. monocytogenes	29		X		Chen et al., 2017b
L. monocytogenes	42		X		Chen et al., 2016
L. monocytogenes	67		X		Jackson et al., 2016
S. enterica	2	X			Wuyts et al., 2015
S. enterica	3	X			Allard et al., 2016
S. enterica	3	X			Taylor et al., 2015
S. enterica	6	X			Hoffmann et al., 2016
S. enterica	12	X			Octavia et al., 2015
S. enterica	30		X		Leekitcharoenphon et al., 2014

The maximum SNP counts for isolates that were traced back to the same source in the original study are presented. Whether the maximum SNP counts are less than 21 SNPs, 21 to 100 SNP, or greater than 100 SNPs is also indicated.

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny

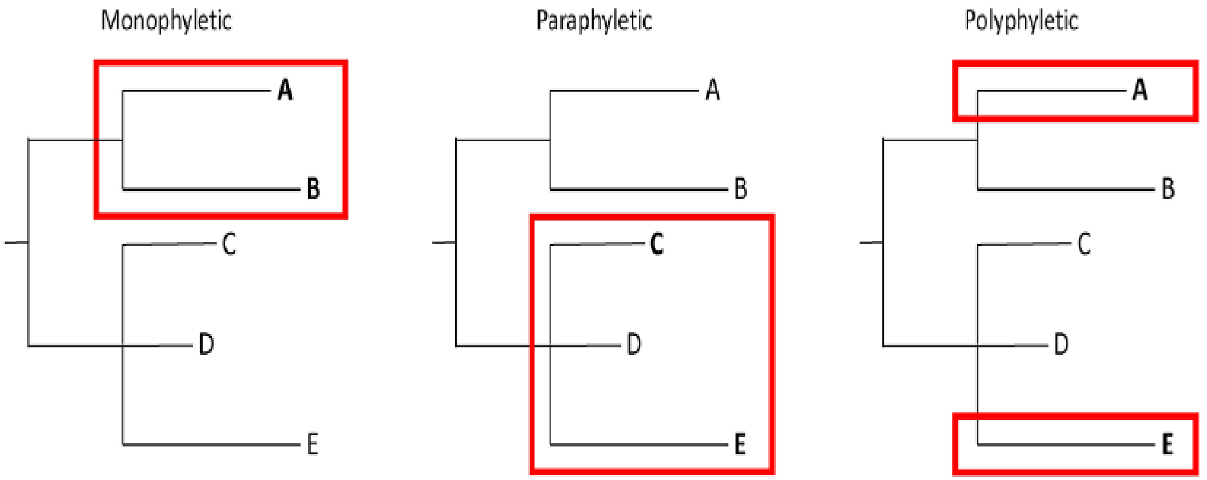


FIGURE 2 | Illustration of monophyletic, paraphyletic, and polyphyletic groupings. A monophyletic topology exists when isolates of interest (e.g., A and B) group together to the exclusion of all others. A paraphyletic topology is one in which isolates of interest (e.g., C and E) group together but not to the exclusion of all others (e.g., D). A polyphyletic topology exists when isolates of interest do not form a group (e.g., A and E).

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny

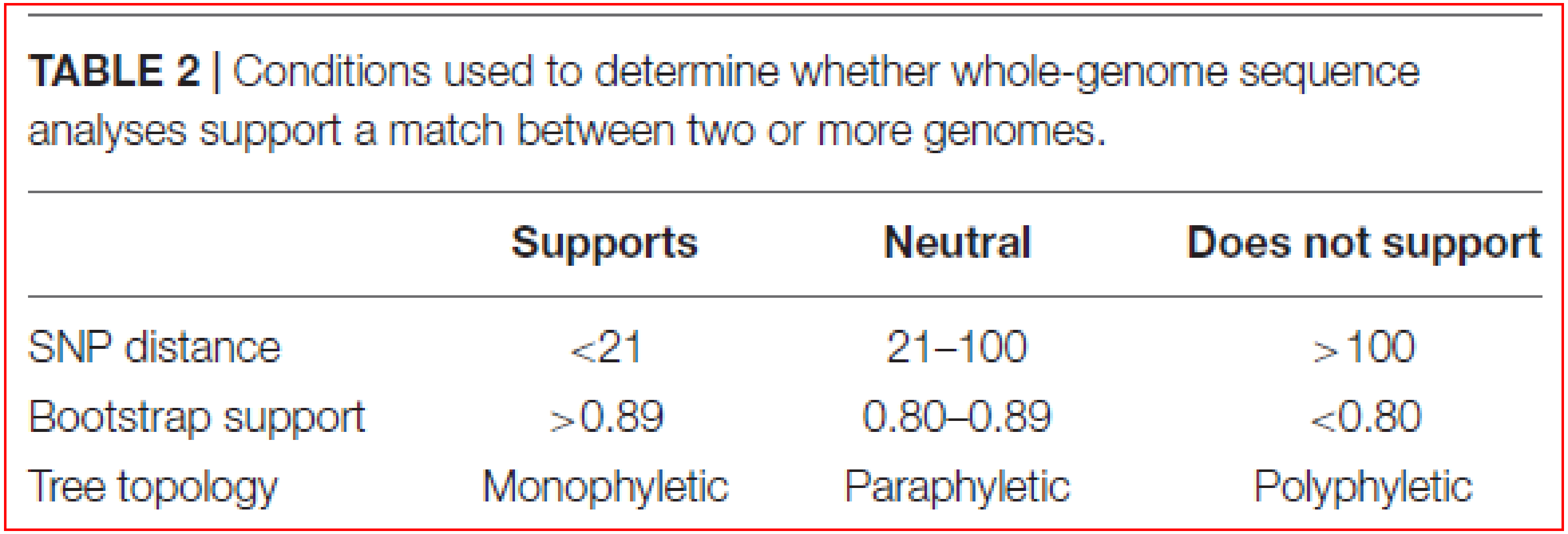
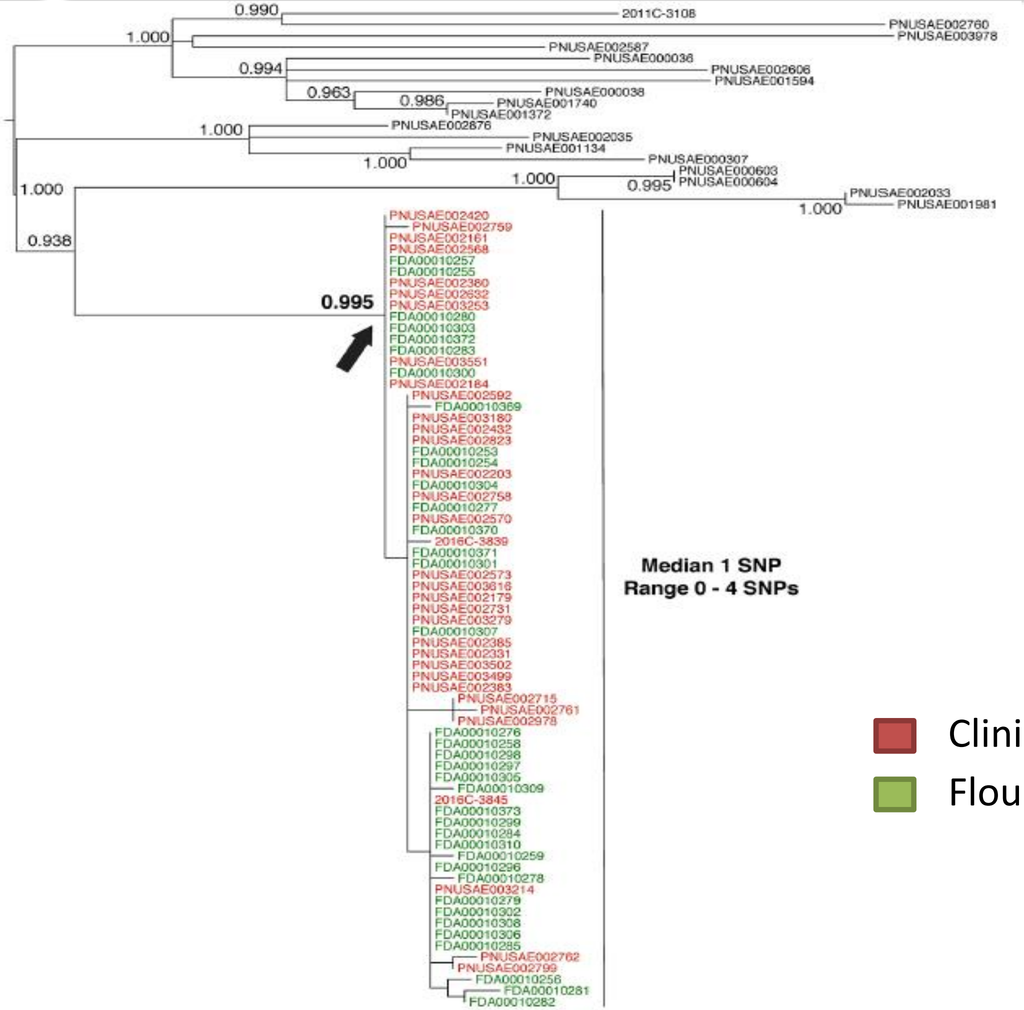


TABLE 2 | Conditions used to determine whether whole-genome sequence analyses support a match between two or more genomes.

	Supports	Neutral	Does not support
SNP distance	<21	21–100	>100
Bootstrap support	>0.89	0.80–0.89	<0.80
Tree topology	Monophyletic	Paraphyletic	Polyphyletic

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny



- Bootstrap support
- SNP count support
- Topology support
- Epidemiology support

E. coli Clinical isolates - source

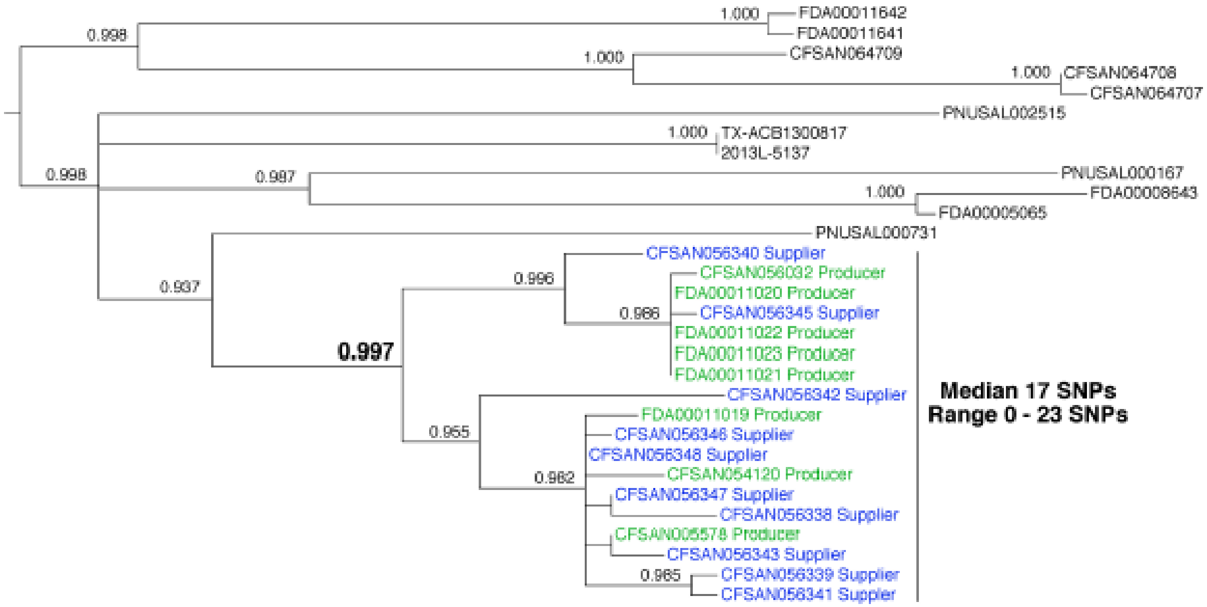
🟥 Clinical isolates

🟩 Flour isolates

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny

L. monocytogenes ingredient supplier – Ice cream producer

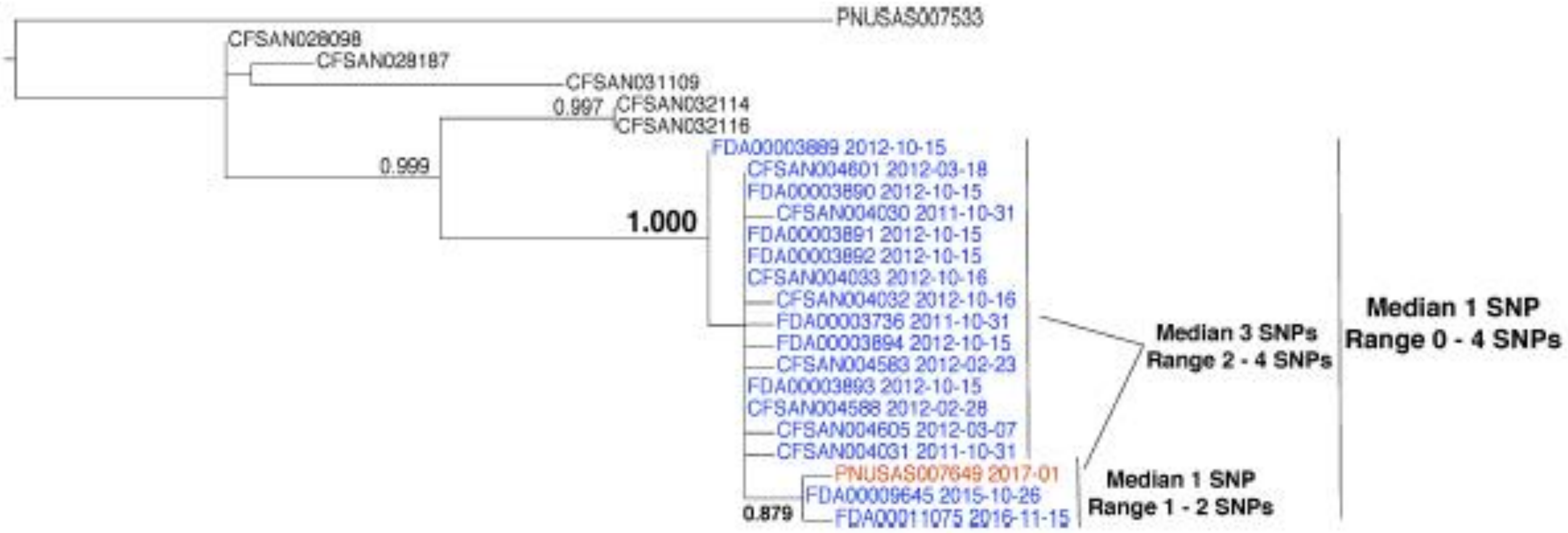


- Bootstrap support
- SNP count support
- Topology support
- Epidemiology support

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny

Resident pathogen



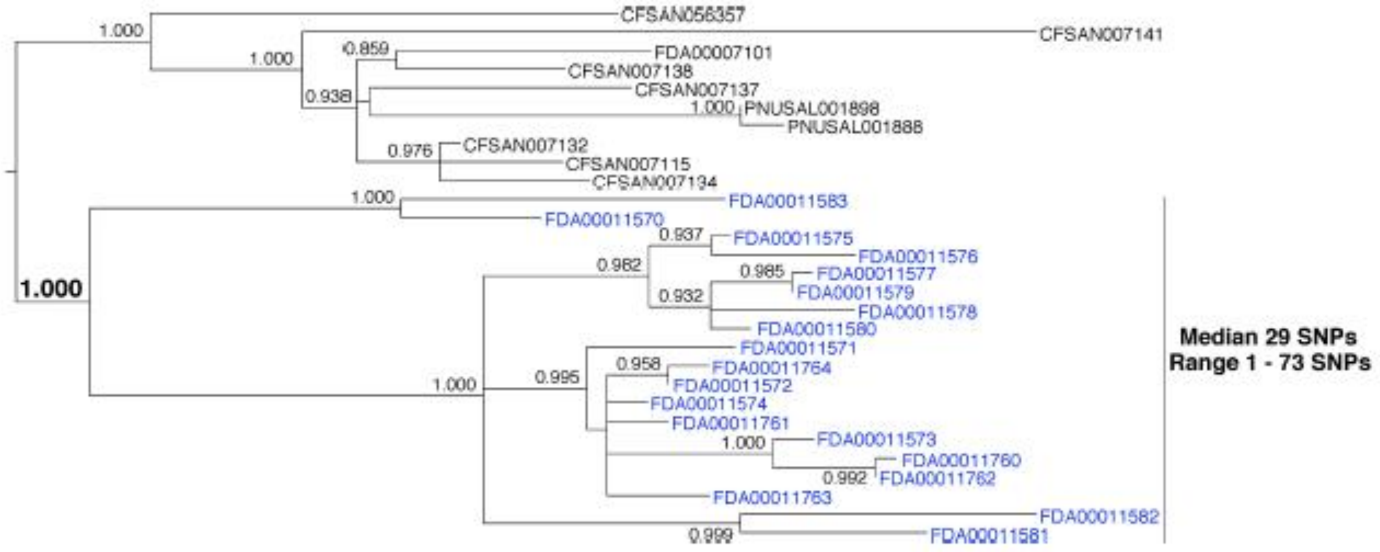
Pightling et al. Frontiers in Microbiology. 2018

- Bootstrap support
- SNP count support
- Topology support
- Epidemiology not support

🟥 Clinical isolates

🟦 Environmental isolates

How to interpret our phylogeny



Pightling et al. Frontiers in Microbiology. 2018

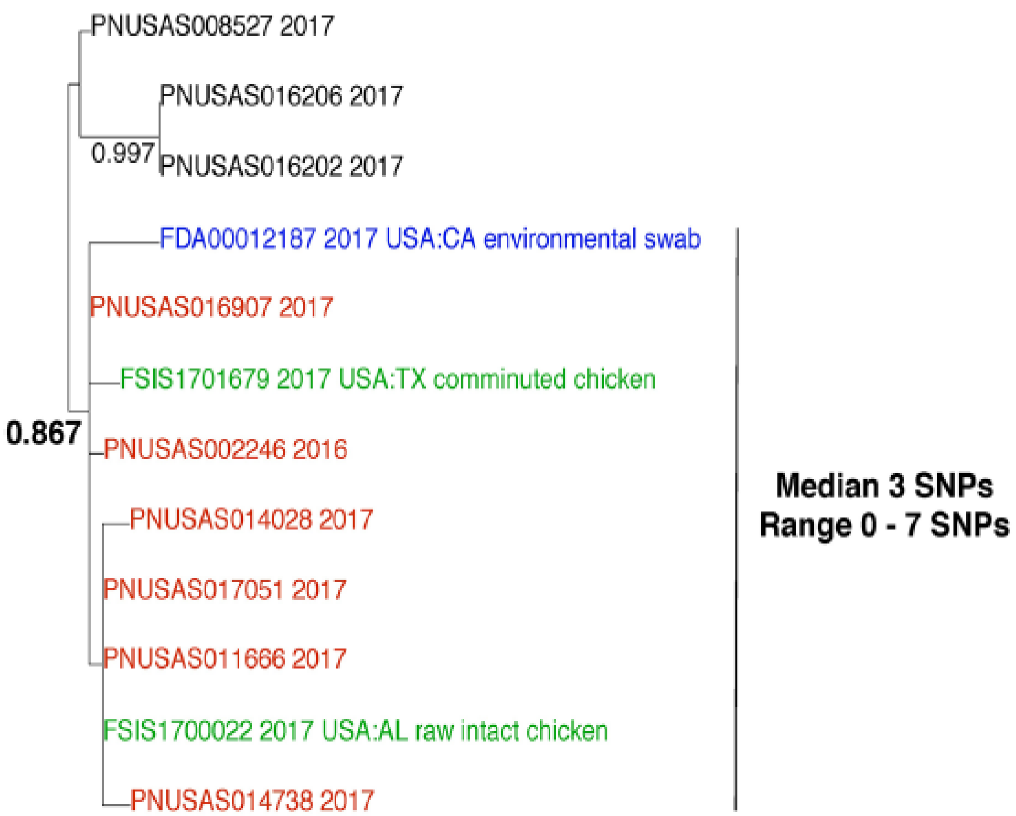
- Bootstrap support
- SNP count neutral
- Topology support

🟥 Clinical isolates

🟦 Environmental isolates

Environmental isolates from an inspection.

How to interpret our phylogeny

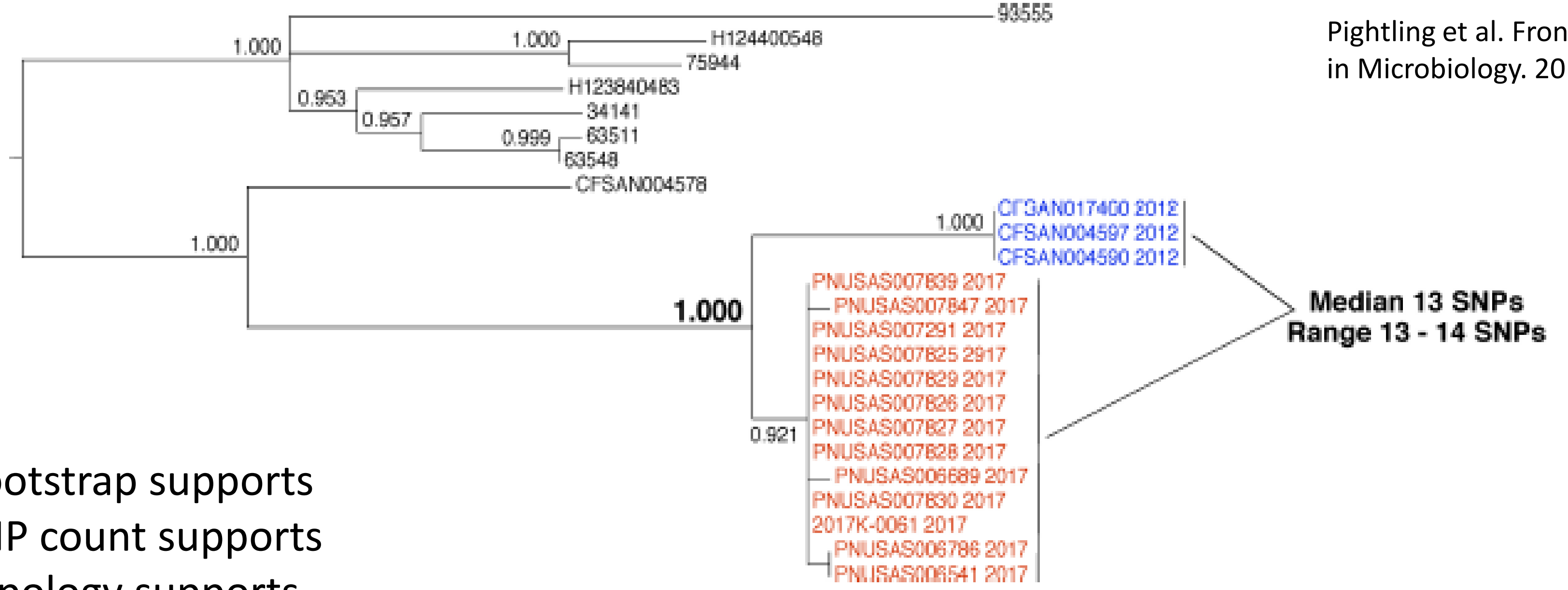


- Bootstrap neutral
- SNP count support
- Topology neutral
- Epidemiology Not support

🟥 Clinical isolates
🟩 Food isolates
🟦 Environmental isolates

Pightling et al. Frontiers in Microbiology. 2018

How to interpret our phylogeny



Pightling et al. Frontiers in Microbiology. 2018

- Bootstrap supports
- SNP count supports
- Topology supports
- Epidemiology does not support

🟥 Clinical isolates

🟦 Environmental isolates

How to interpret our phylogeny

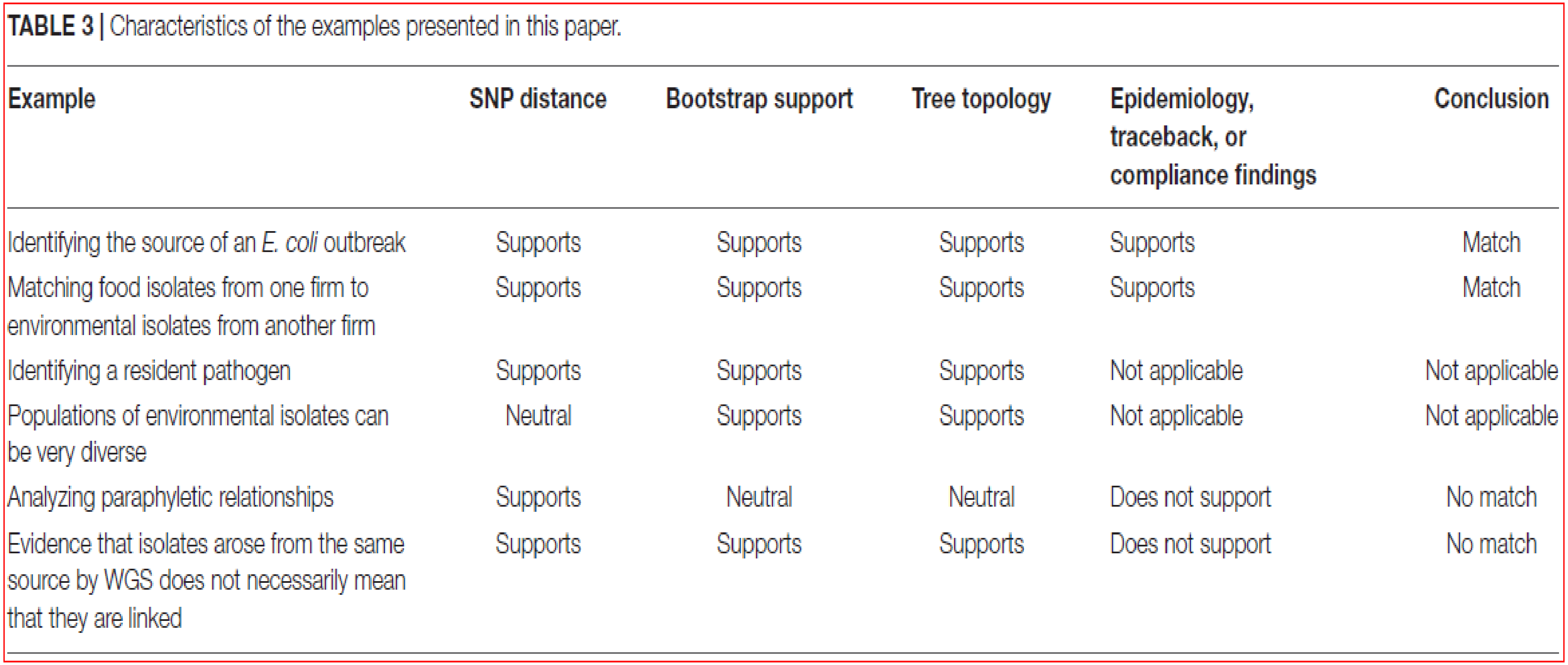


TABLE 3 | Characteristics of the examples presented in this paper.

Example	SNP distance	Bootstrap support	Tree topology	Epidemiology, traceback, or compliance findings	Conclusion
Identifying the source of an *E. coli* outbreak	Supports	Supports	Supports	Supports	Match
Matching food isolates from one firm to environmental isolates from another firm	Supports	Supports	Supports	Supports	Match
Identifying a resident pathogen	Supports	Supports	Supports	Not applicable	Not applicable
Populations of environmental isolates can be very diverse	Neutral	Supports	Supports	Not applicable	Not applicable
Analyzing paraphyletic relationships	Supports	Neutral	Neutral	Does not support	No match
Evidence that isolates arose from the same source by WGS does not necessarily mean that they are linked	Supports	Supports	Supports	Does not support	No match

Pightling et al. Frontiers in Microbiology. 2018

Thanks for your attention!