Session 3.1 – Mapping against reference genome and Variant Calling

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

16-20 Octubre 2023, 5ª Edición
Programa Formación Continua, ISCIII

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:

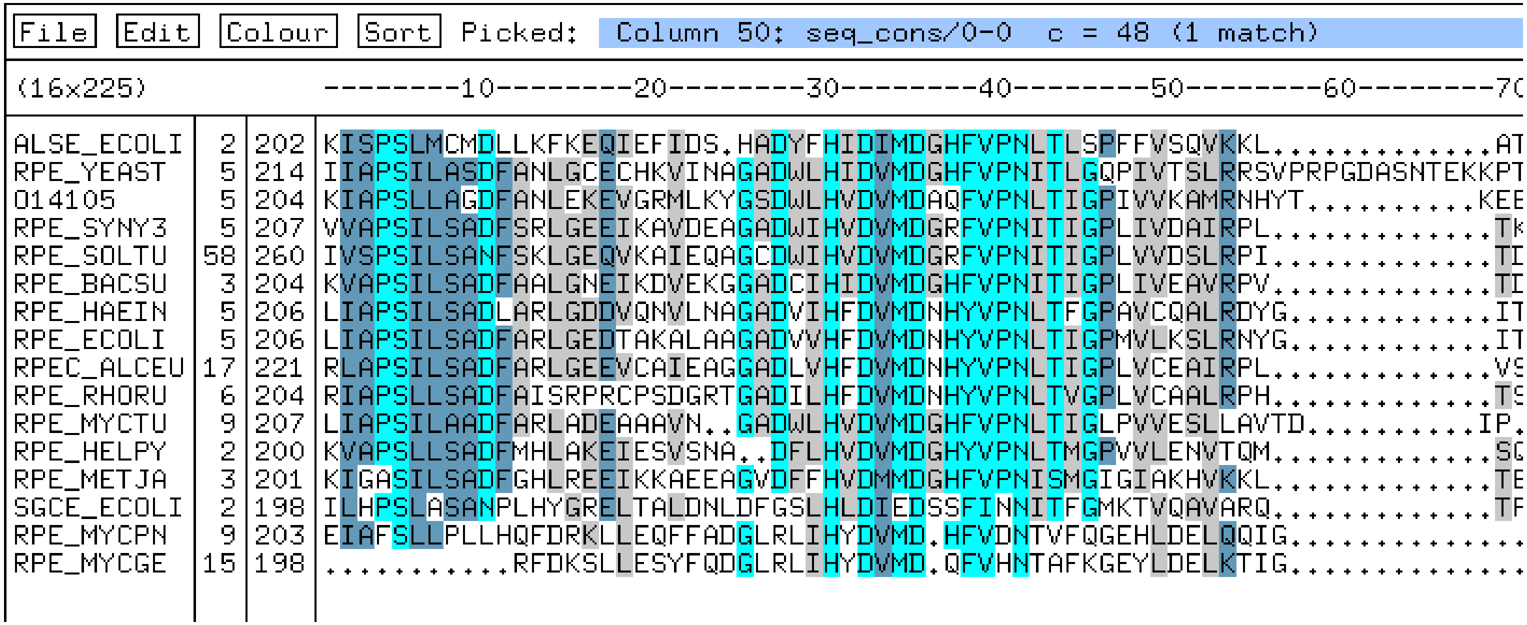
Arrange two or more nucleotide or aminoacid sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships.

```
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.

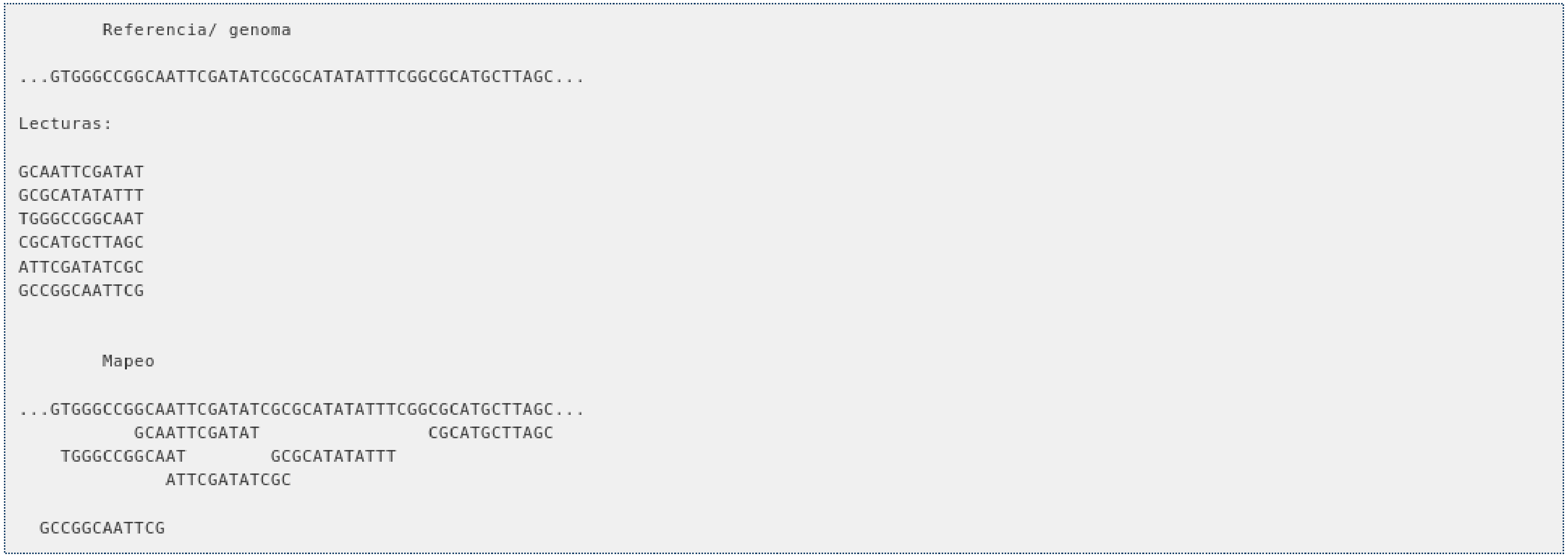


```
File  Edit  Colour  Sort  Picked:  Column 50: seq_cons/0-0  c = 48 (1 match)

(16x225)                  --------10--------20--------30--------40--------50--------60--------7

ALSE_ECOLI   2 202 KISPSLMCMDLLKFKEQIEFIDS.HADYFHIDIMDGHFVPNLTLSPFFVSQVKKL.............AT
RPE_YEAST    5 214 IIAPSILASDFANLGCECHKVINAGADWLHIDVMDGHFVPNITLGQPIVTSLRRSVPRPGDASNTEKKPT
O14105       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:

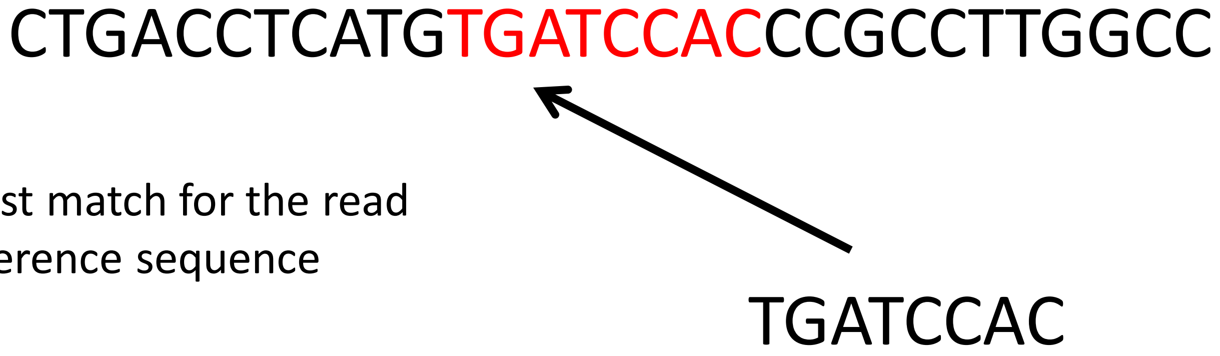
- An alignment is regarded to be correct only if each base is placed correctly.
- Minimizes differences among sequences
- From tens to hundred of sequences.

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…



CTGACCTCATGTGATCCACCCGCCTTGGCC

Find best match for the read
in a reference sequence

TGATCCAC

Challenges

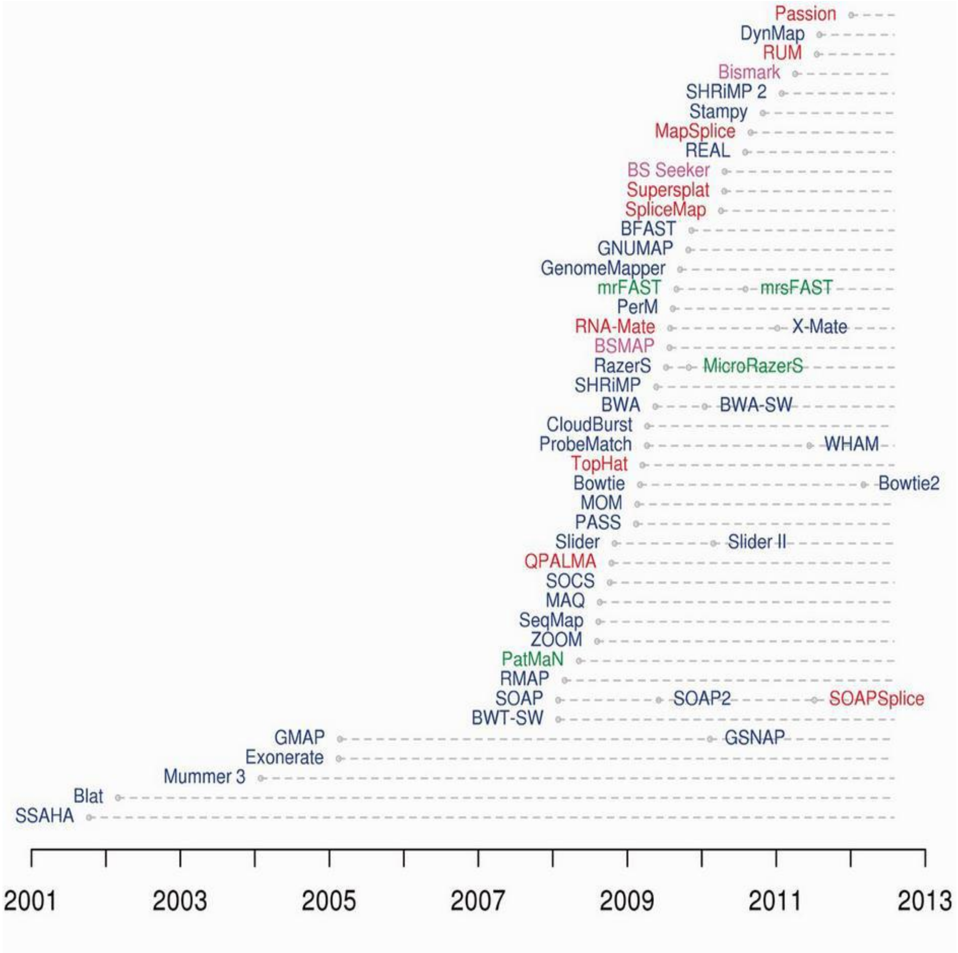
- Errors in reads
- Errors in libraries
- Repetitive regions (repeats, homologous regions)
- Homopolymers
- Individual polymorphisms

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

What mapper should I use?

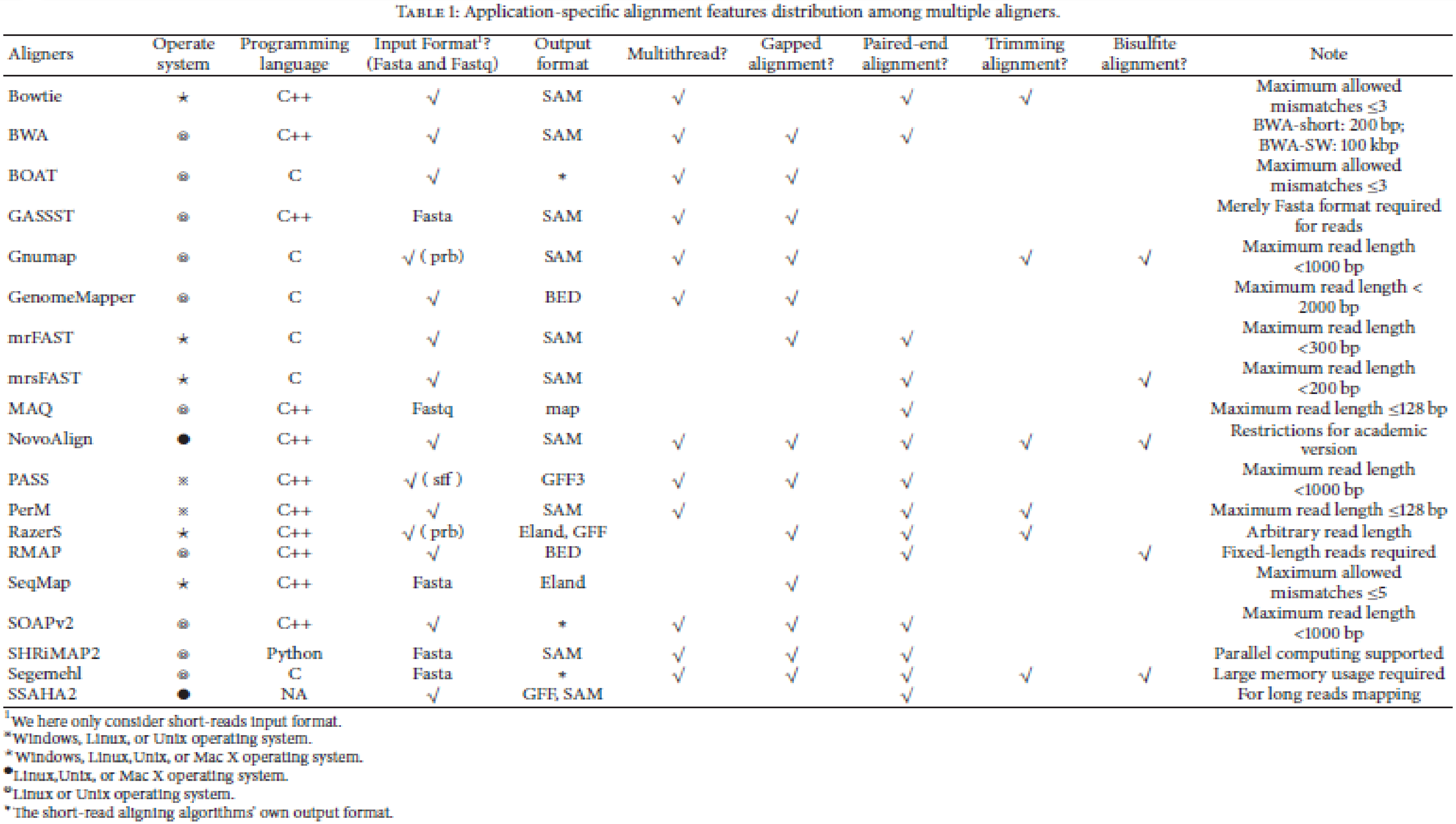


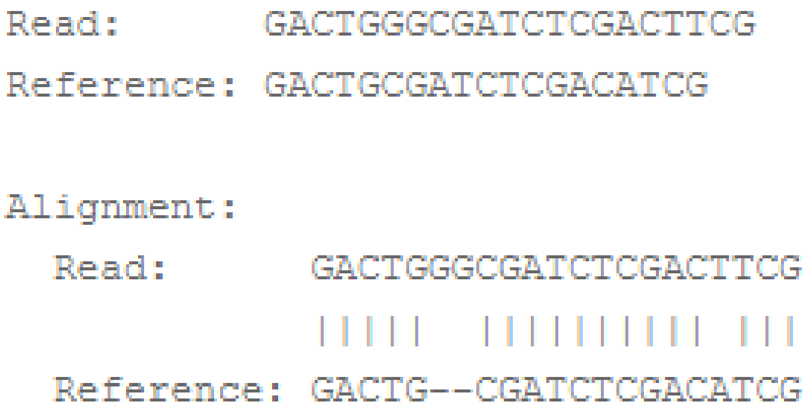
Table 1: Application-specific alignment features distribution among multiple aligners.

Aligners	Operate system	Programming language	Input Format[1]? (Fasta and Fastq)	Output format	Multithread?	Gapped alignment?	Paired-end alignment?	Trimming alignment?	Bisulfite alignment?	Note
Bowtie	★	C++	√	SAM	√		√	√		Maximum allowed mismatches ≤3
BWA	⊛	C++	√	SAM	√	√	√			BWA-short: 200 bp; BWA-SW: 100 kbp
BOAT	⊛	C	√	*	√	√				Maximum allowed mismatches ≤3
GASSST	⊛	C++	Fasta	SAM	√	√				Merely Fasta format required for reads
Gnumap	⊛	C	√ (prb)	SAM	√	√		√	√	Maximum read length <1000 bp
GenomeMapper	⊛	C	√	BED	√	√				Maximum read length < 2000 bp
mrFAST	★	C	√	SAM		√	√			Maximum read length <300 bp
mrsFAST	★	C	√	SAM			√		√	Maximum read length <200 bp
MAQ	⊛	C++	Fastq	map			√			Maximum read length ≤128 bp
NovoAlign	●	C++	√	SAM	√	√	√	√	√	Restrictions for academic version
PASS	※	C++	√ (sff)	GFF3	√	√	√			Maximum read length <1000 bp
PerM	※	C++	√	SAM	√		√	√		Maximum read length ≤128 bp
RazerS	★	C++	√ (prb)	Eland, GFF		√	√	√		Arbitrary read length
RMAP	⊛	C++	√	BED			√		√	Fixed-length reads required
SeqMap	★	C++	Fasta	Eland		√				Maximum allowed mismatches ≤5
SOAPv2	⊛	C++	√	*	√	√	√			Maximum read length <1000 bp
SHRiMAP2	⊛	Python	Fasta	SAM	√	√	√			Parallel computing supported
Segemehl	⊛	C	Fasta	*	√	√	√	√	√	Large memory usage required
SSAHA2	●	NA	√	GFF, SAM			√			For long reads mapping

[1] We here only consider short-reads input format.
※ Windows, Linux, or Unix operating system.
★ Windows, Linux, Unix, or Mac X operating system.
● Linux, Unix, or Mac X operating system.
⊛ Linux or Unix operating system.
* The short-read aligning algorithms' own output format.

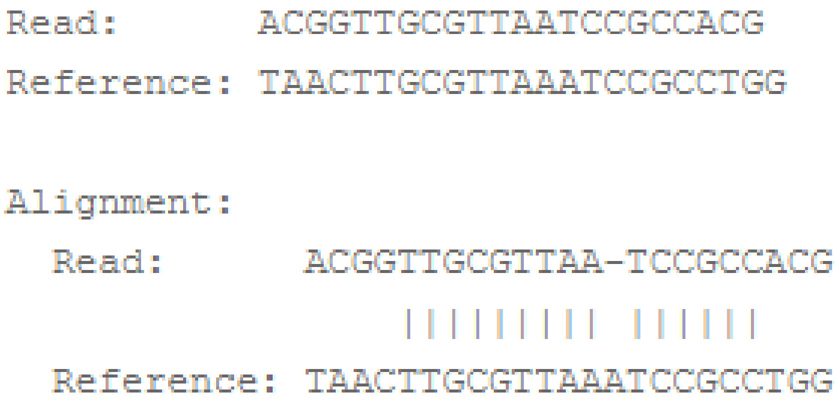
End-to-end vs local alignment

End-to-end



```
Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG

Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

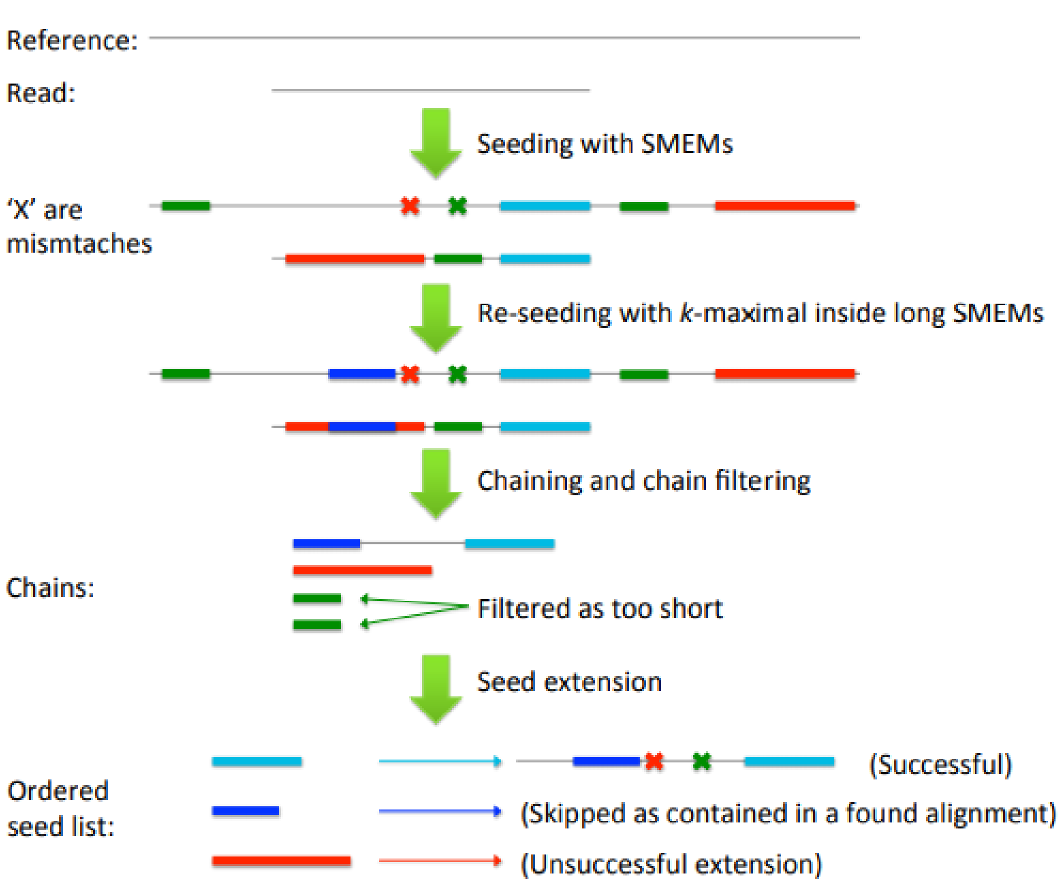
Local



```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG

Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

Bowtie2 manual.

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)

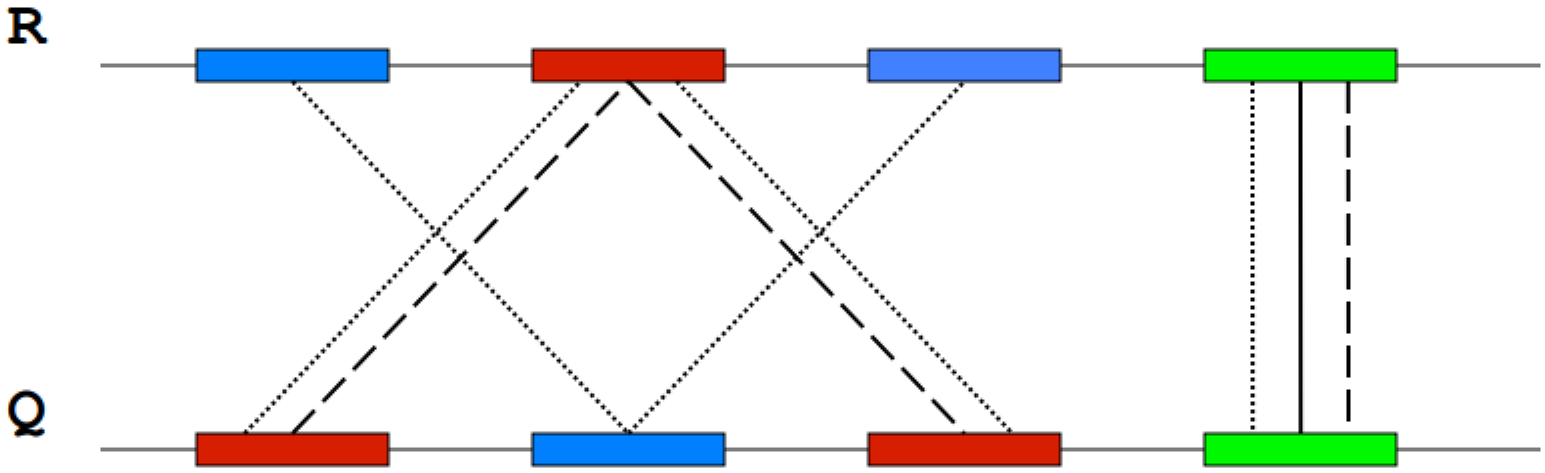
Adam M. Phillippy. Whole Genome Alignment with MUMmer. Presentation.

Example whole genome aligner: MUMMER

MUM : maximal unique match

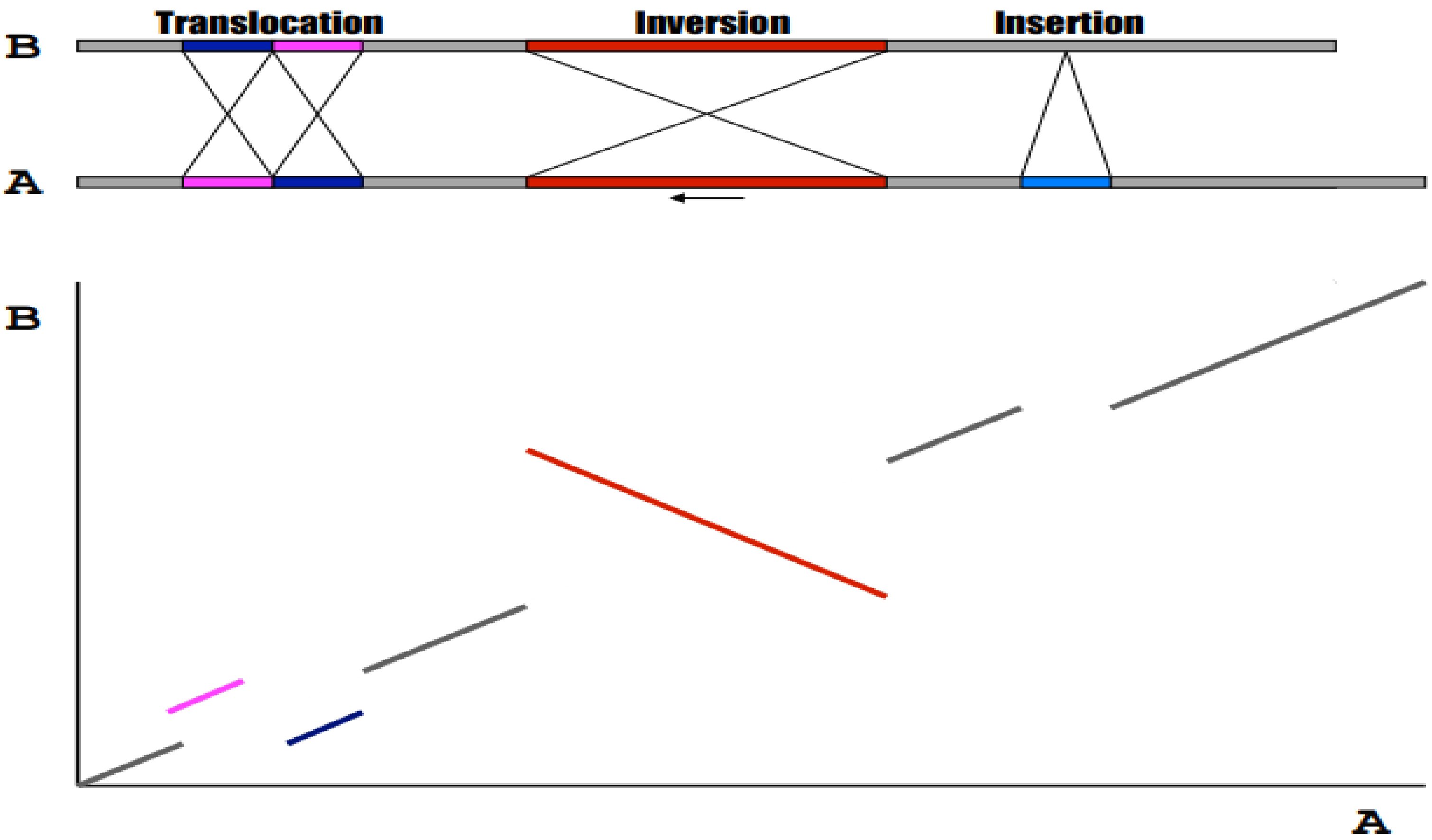
MAM : maximal almost-unique match

MEM : maximal exact match



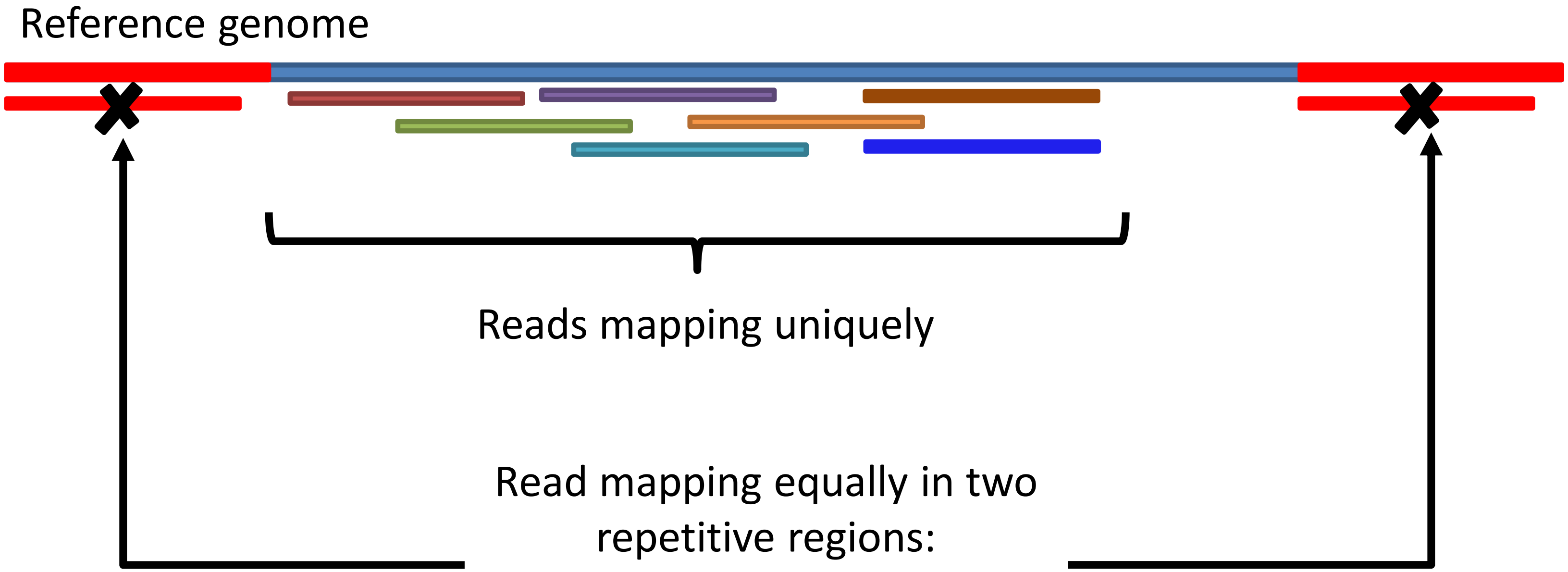
Adam M. Phillippy. Whole Genome Alignment with MUMmer. Presentation.

Example whole genome aligner: MUMMER



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

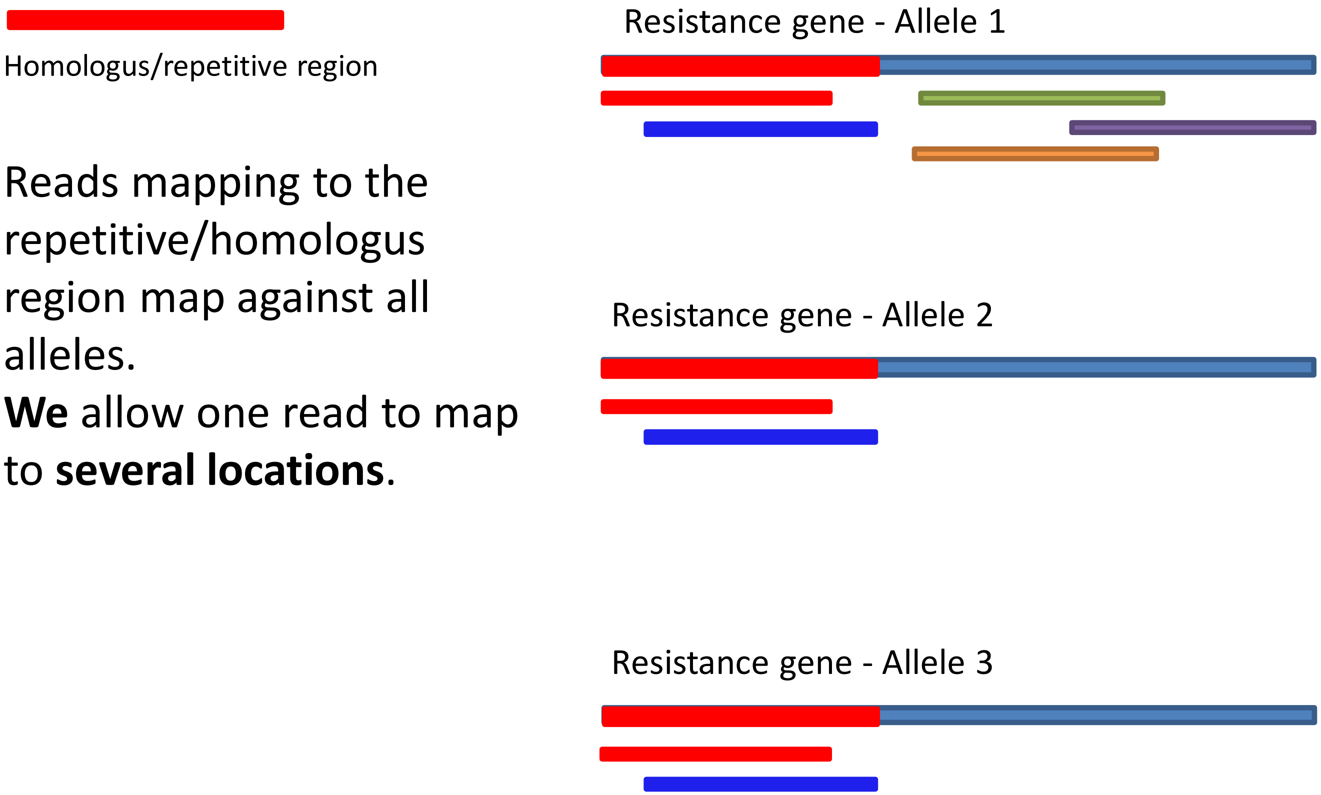
Which aligner should I use for aligning reads agains a complete genome for variant calling?



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

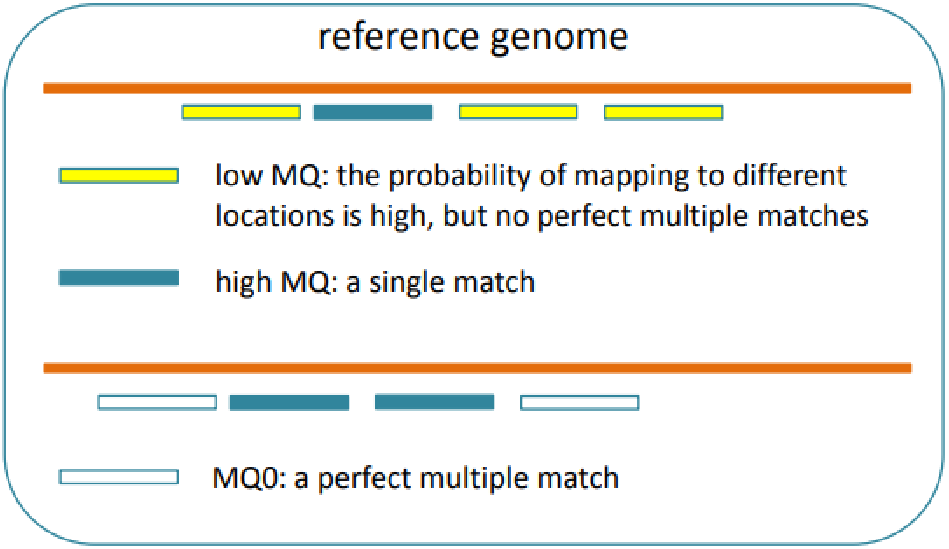
Reads mapping to the repetitive/homologus region map against all alleles.
We allow one read to map to **several locations**.

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

MAPQ

MAPQ is NOT comparable among mappers.

BWA:

- MAPQ represents the probability of the read to be mapped correctly.
- MAPQ = 0 identifies unmapped reads and…

Reads mapping to different locations!

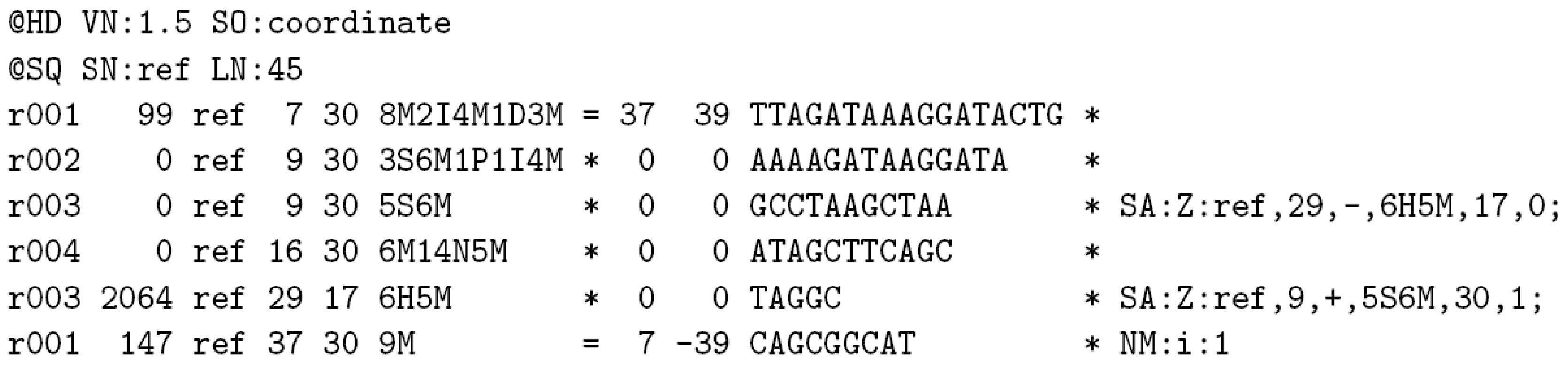
BOWTIE2:

- MAPQ represents the “uniqueness” of the read. A MAPQ < 10 indicates that there is at least a 1 in 10 chance that the read truly originated elsewhere
- MAPQ = 0 identifies unmapped reads

SAM format

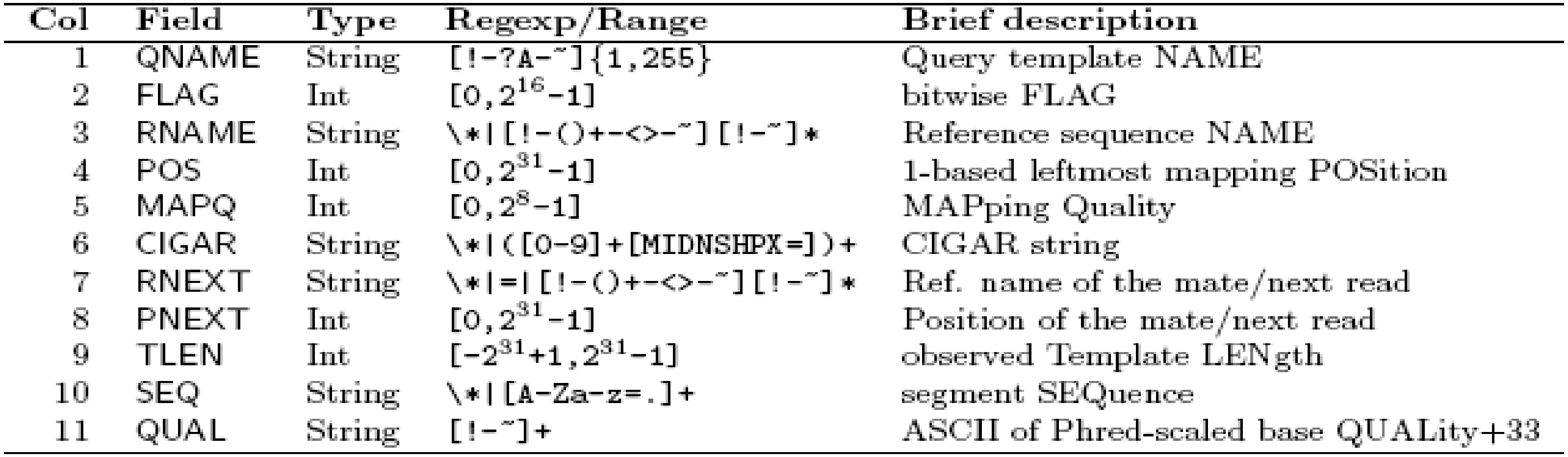
Definición:

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

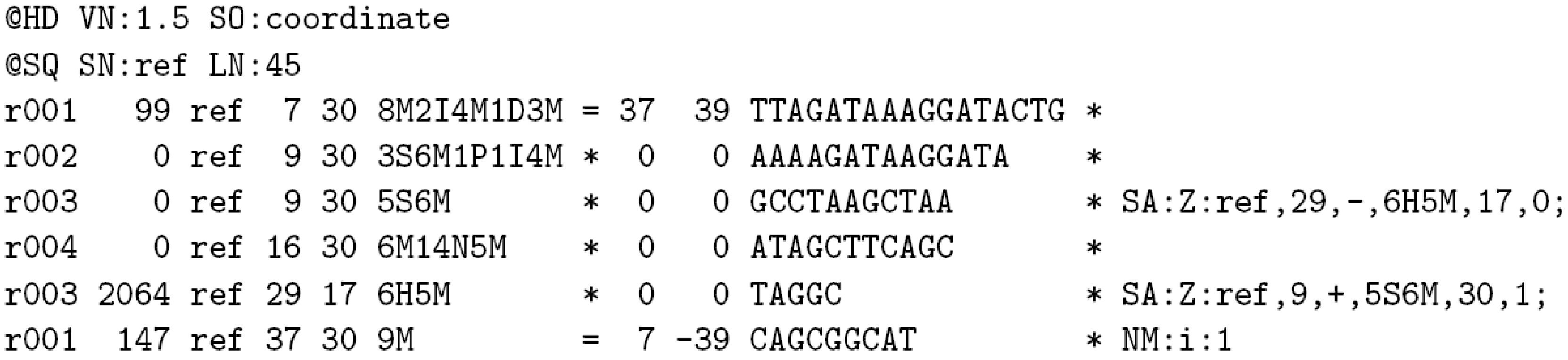


```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC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

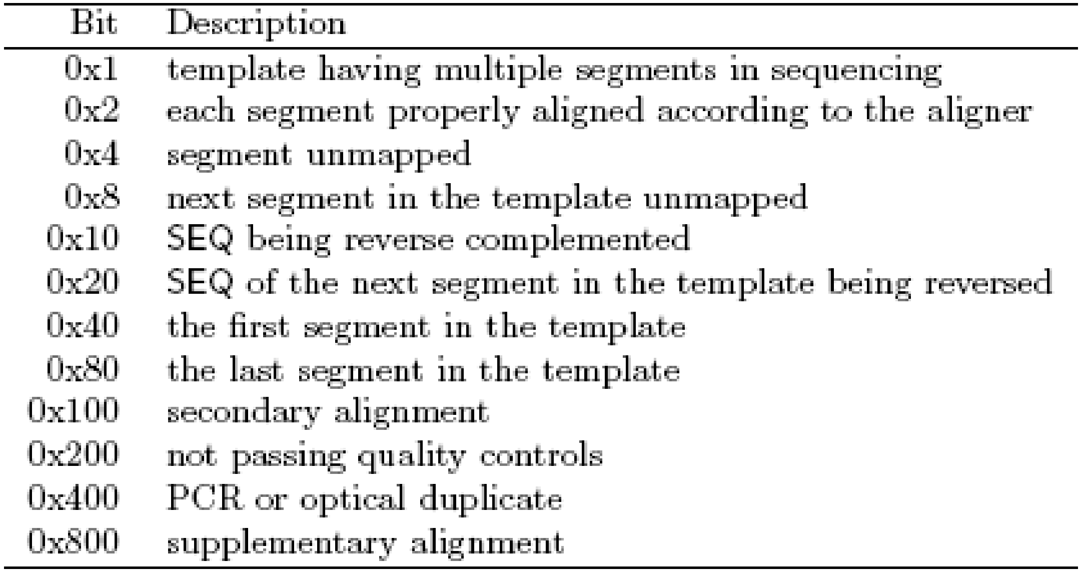


Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String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



```
@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
```

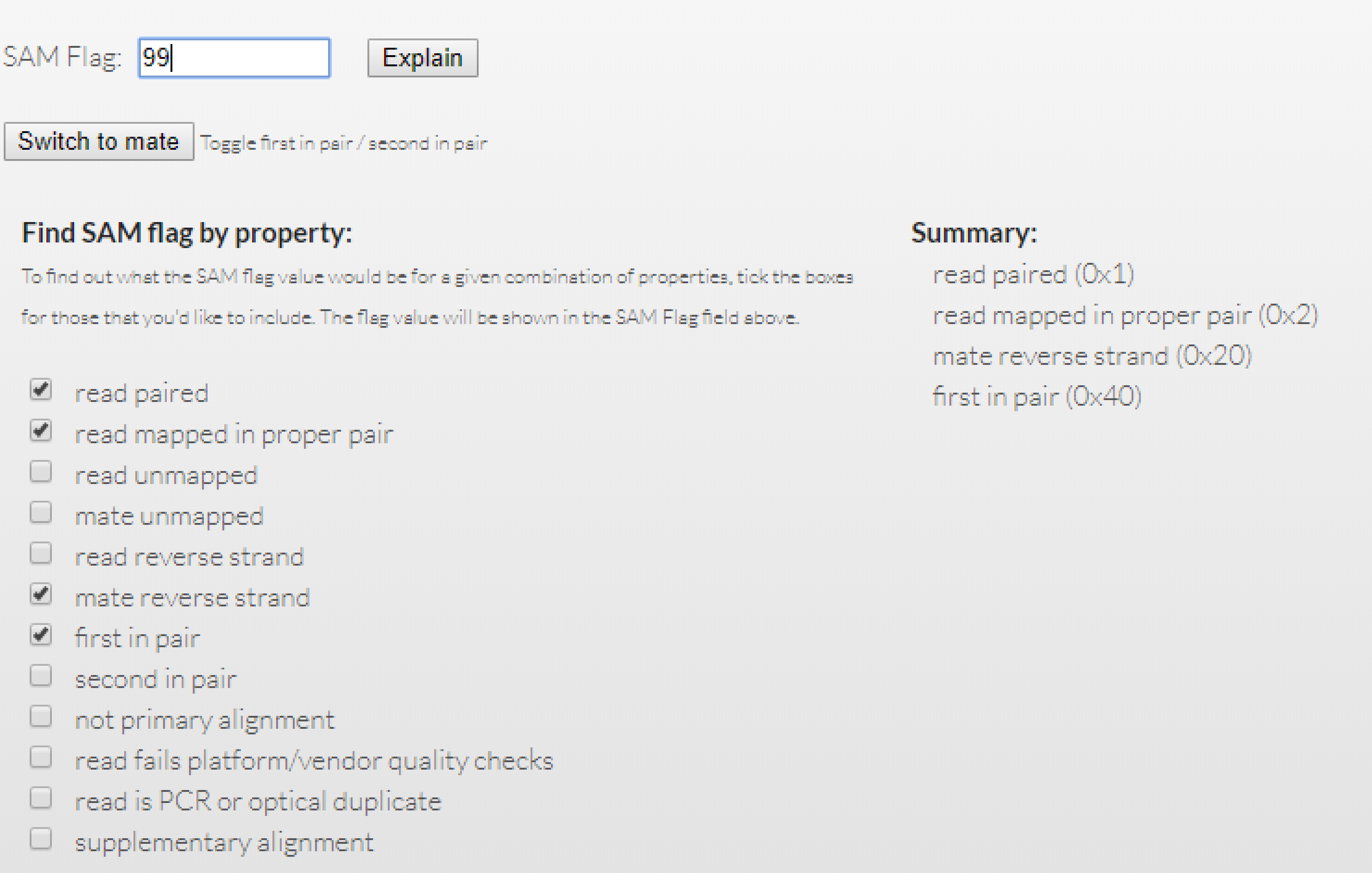
SAM format: flags



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 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:

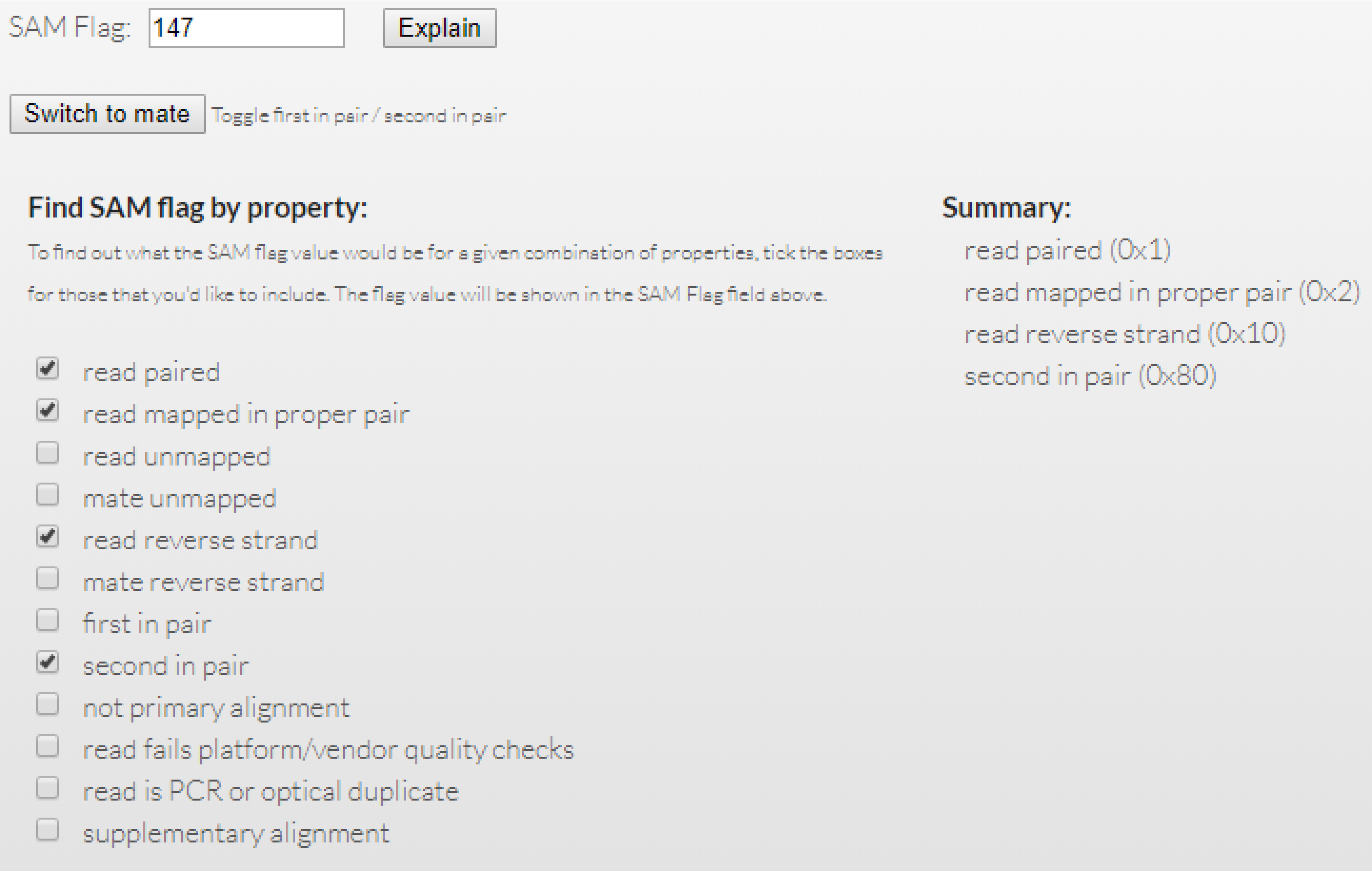
read paired (0x1)

read mapped in proper pair (0x2)

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:

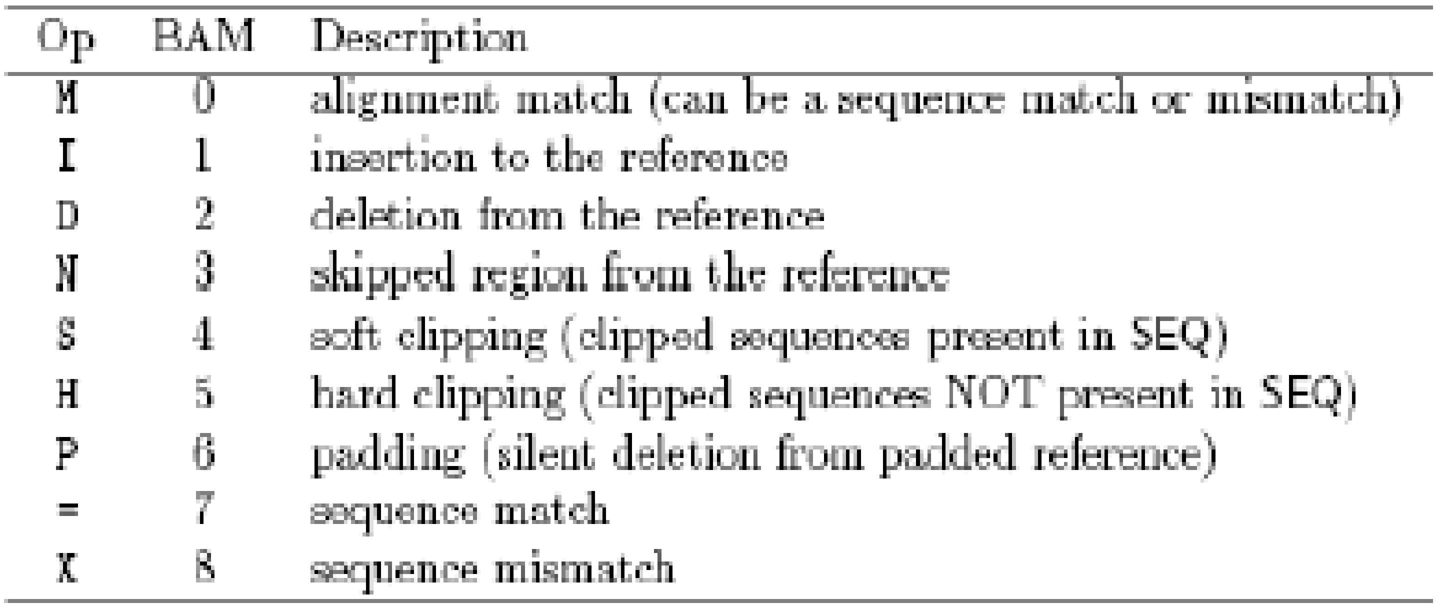
To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

☑ 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

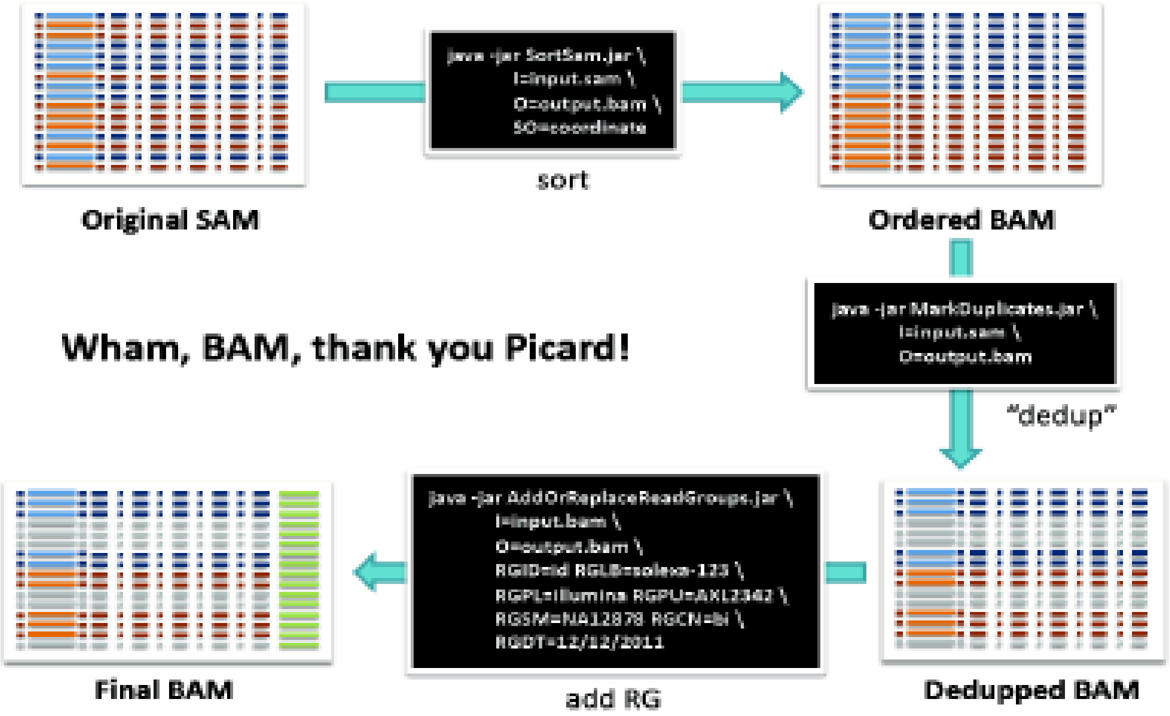
BECAUSE IT SAVES A LOT OF DISK SPACE!!

<u>Typical bam and sam format files weights from a S. grumpensis</u>
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



Original SAM

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

sort

Ordered BAM

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

"dedup"

Dedupped 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
```

add RG

Final BAM

Wham, BAM, thank you Picard!

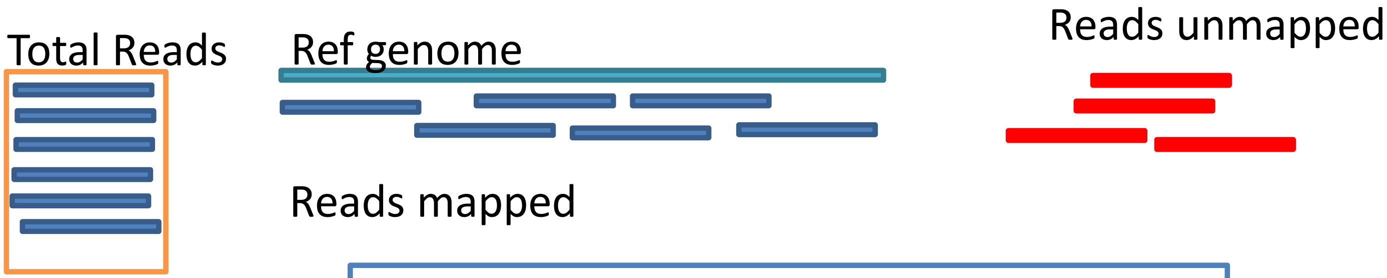
Mapping statistics

- % 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

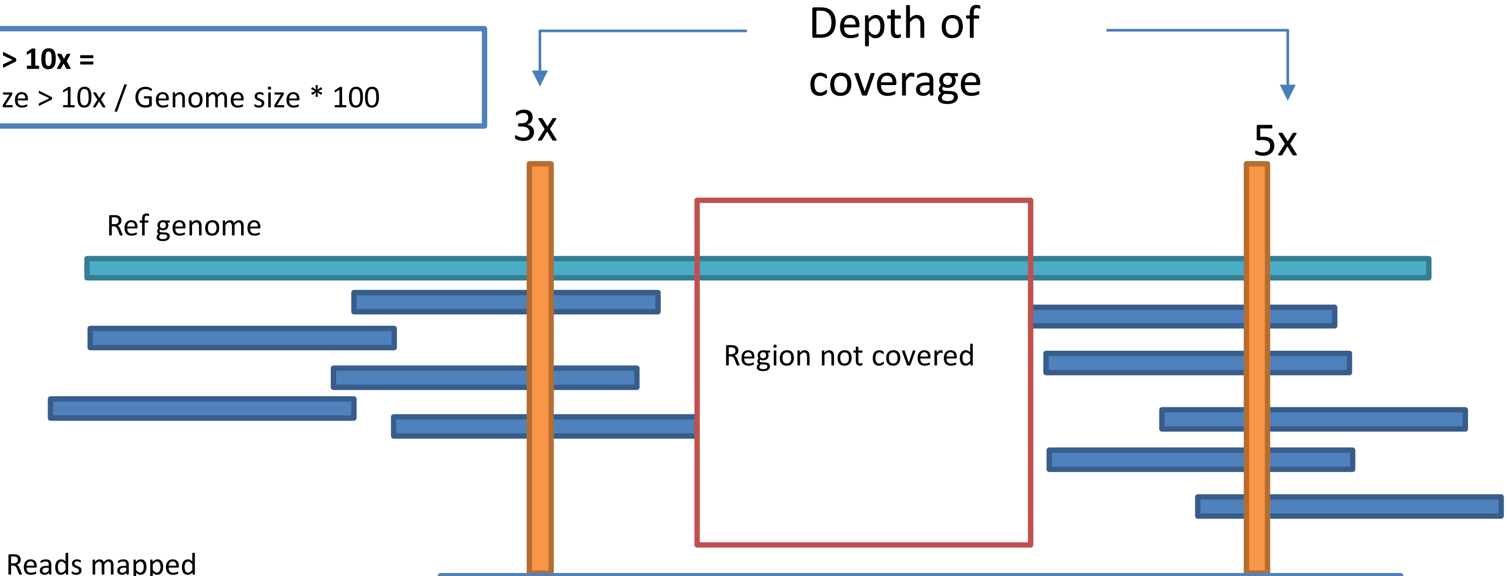
Mandatory parameter for microbial genomics!! It indicates us how many reads we have from our organism of interest. In human genomics this is almost always 99.99% unless something terrible happens. Not here!!!

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

Depth of coverage vs coverage

Breadth of coverage

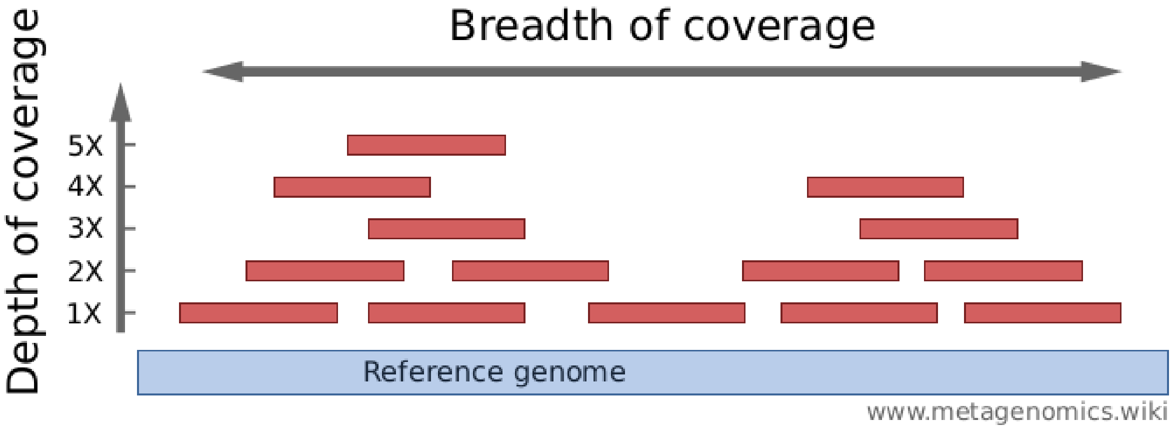
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: 90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth.

Depth of coverage

How strong is a genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).



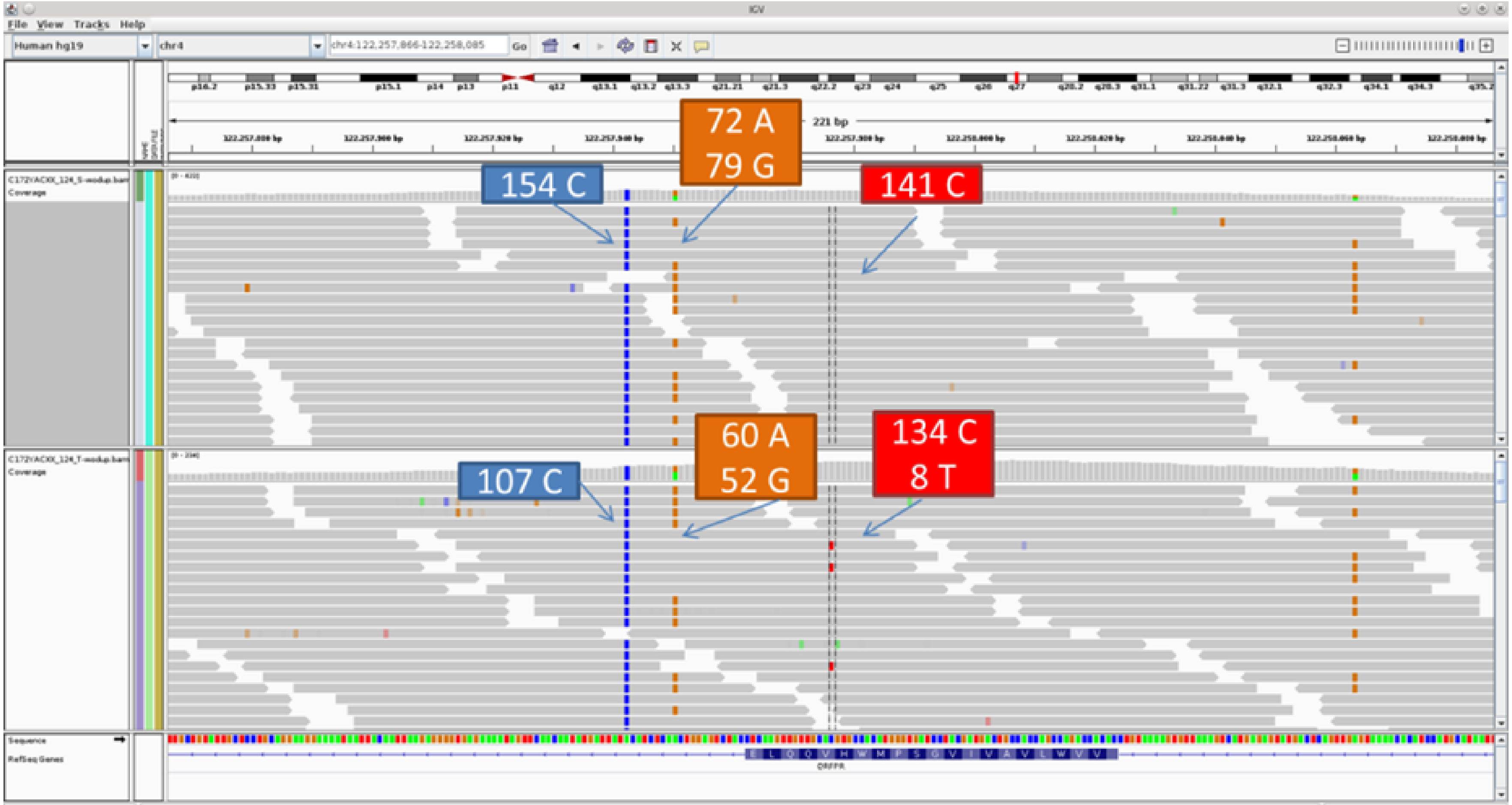
www.metagenomics.wiki

Variant Calling

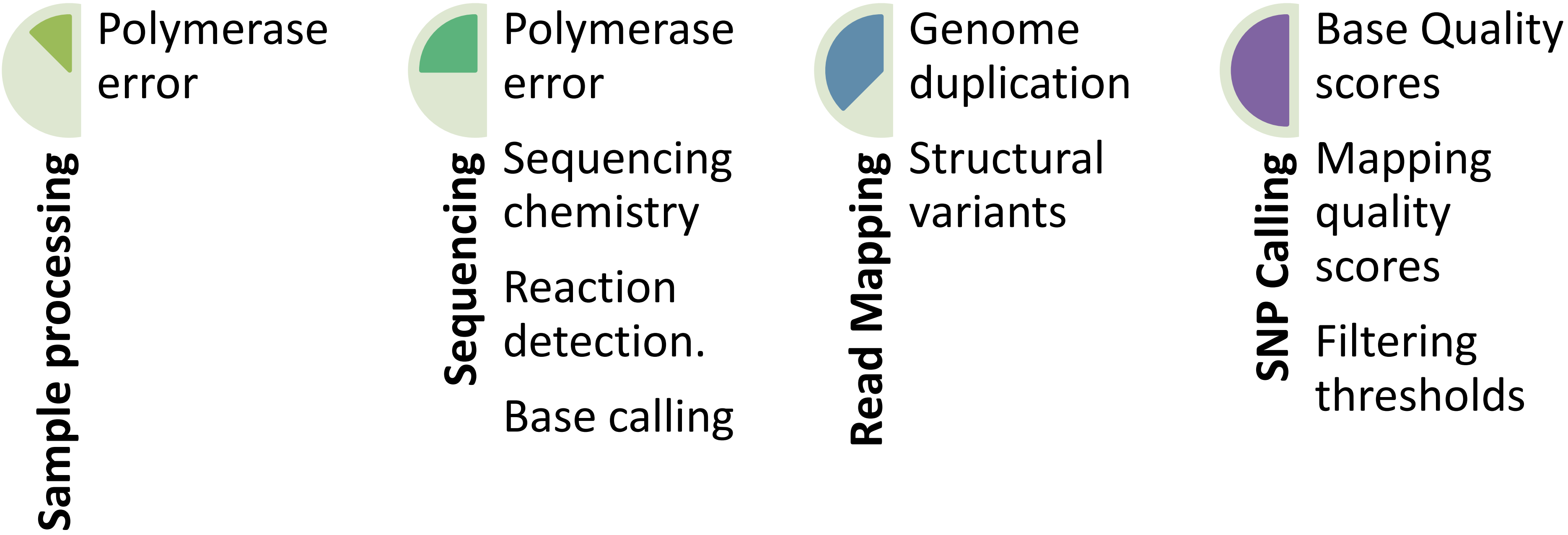
- **Variant calling concept is simple:**

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.


IGV
File View Tracks Help
Human hg19
chr4
chr4:122,257,866-122,258,085
Go
221 bp
C172YACXX_124_S-wodup.bam Coverage
C172YACXX_124_T-wodup.bam Coverage
Sequence
RefSeq Genes
QRFPR
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

- **Sample processing errors.**
  - Random errors.
  - Associated with polymerase errors . (1 in 10^{2-3} bases)
  - Homopolymers and tandem repeats experience higher indel error rates.
- **Solutions:**
  - Paired-end libraries.
  - Minimization of PCR cycles.

Adapted from Olson et al. Frontiers in Genetics. 2015

Sources of error and mitigation strategies

- **Sequencing:**
  - Dependent on the platform.
  - Can be random and systematic.
  - 6% Illumina, 50% Roche (Ross et al.2013)
  - P.e Illumina commits error in the G/T channels.

- **Solutions:**
  - Strand bias.

Adapted from Olson et al. Frontiers in Genetics. 2015

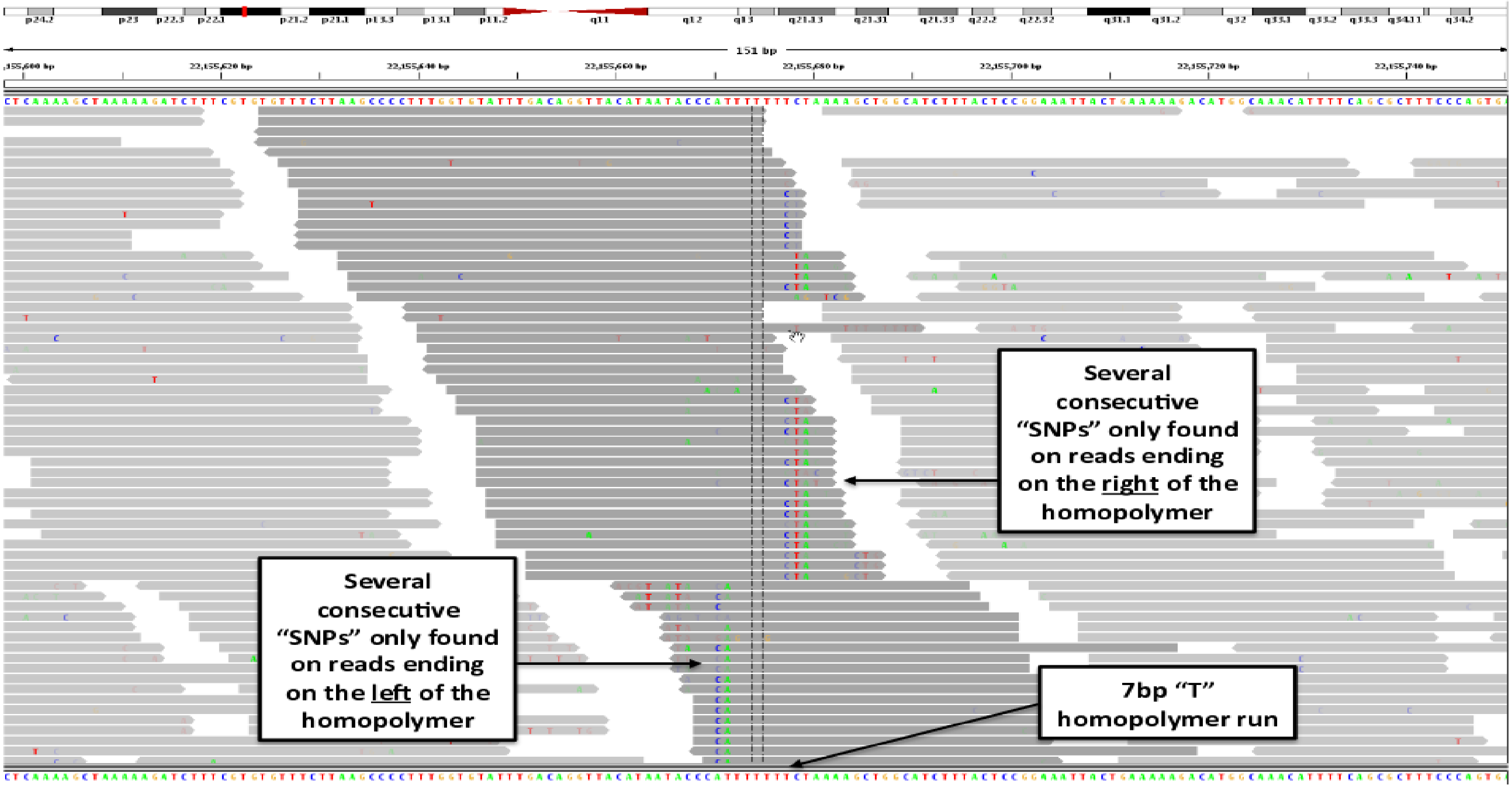
Sources of error and mitigation strategies

- **Mapping errors:**
  - Genomic duplication and structural variation.
  - High diverse areas.

- **Solutions**
  - Paired-end libraries.
  - Long reads / fragments.
  - MAPQ
  - Realignment around indels.

Adapted from Olson et al. Frontiers in Genetics. 2015

Sources of error and mitigation strategies

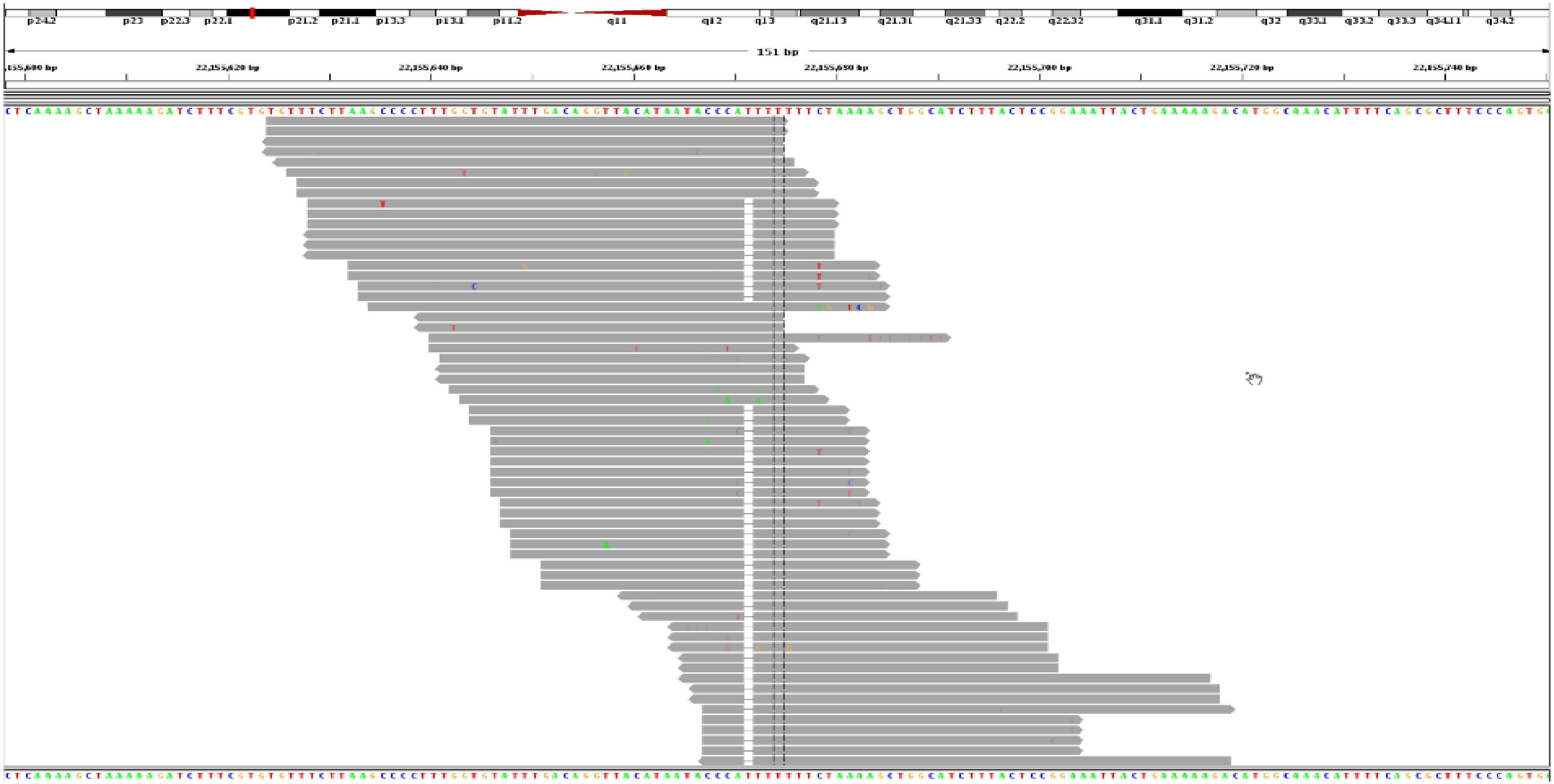


Several consecutive "SNPs" only found on reads ending on the right of the homopolymer

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

7bp "T" homopolymer run

Sources of error and mitigation strategies



Sources of error and mitigation strategies

- **SNP calling step**
  - Errors may result in base calling errors.
  - FP and FN calls.
- **Solutions**
  - Strand bias
  - Base quality rank sum
  - MAPQ
  - Hard filters:
    - Depth of coverage
    - Minimun base call frequency.

Adapted from Olson et al. Frontiers in Genetics. 2015

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!

GMI Proficiency test

1. **Proficiency Testing for bacterial WGS, 2012 an end-user survey of current capabilities, requirements and priorities**
2. **Proficiency Test Pilot, 2014 Wet lab and Dry lab**

Escherichia coli, Staphilococus aureus and Salmonella typhimurium

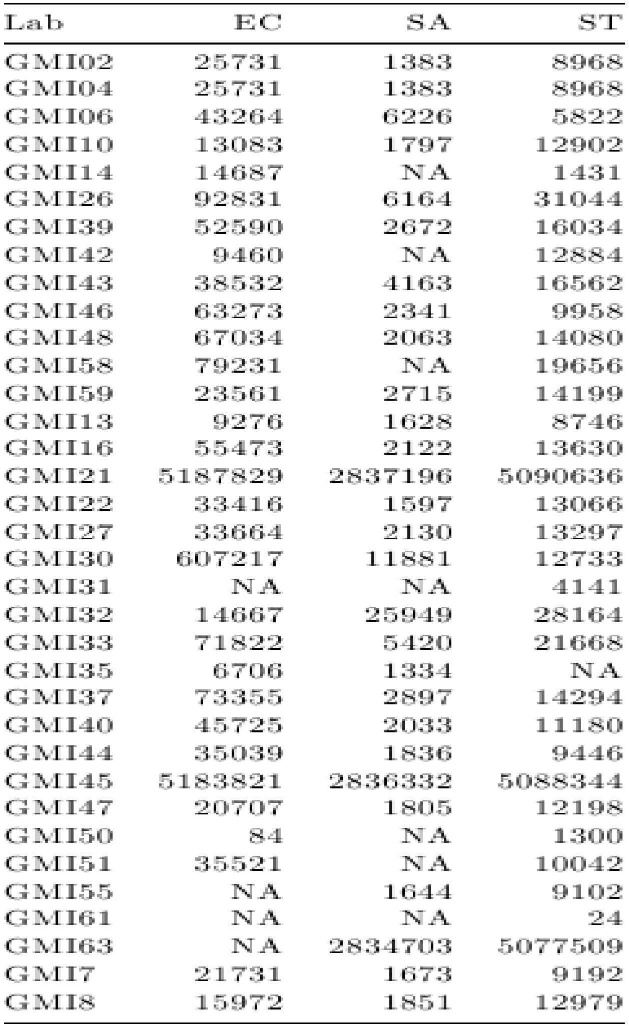
3. **Full Proficiency Test, 2015**

Escherichia coli, Staphilococus aureus and Salmonella tiphimurium

4. **Full Proficiency Test, 2016 Wet lab and Dry lab**

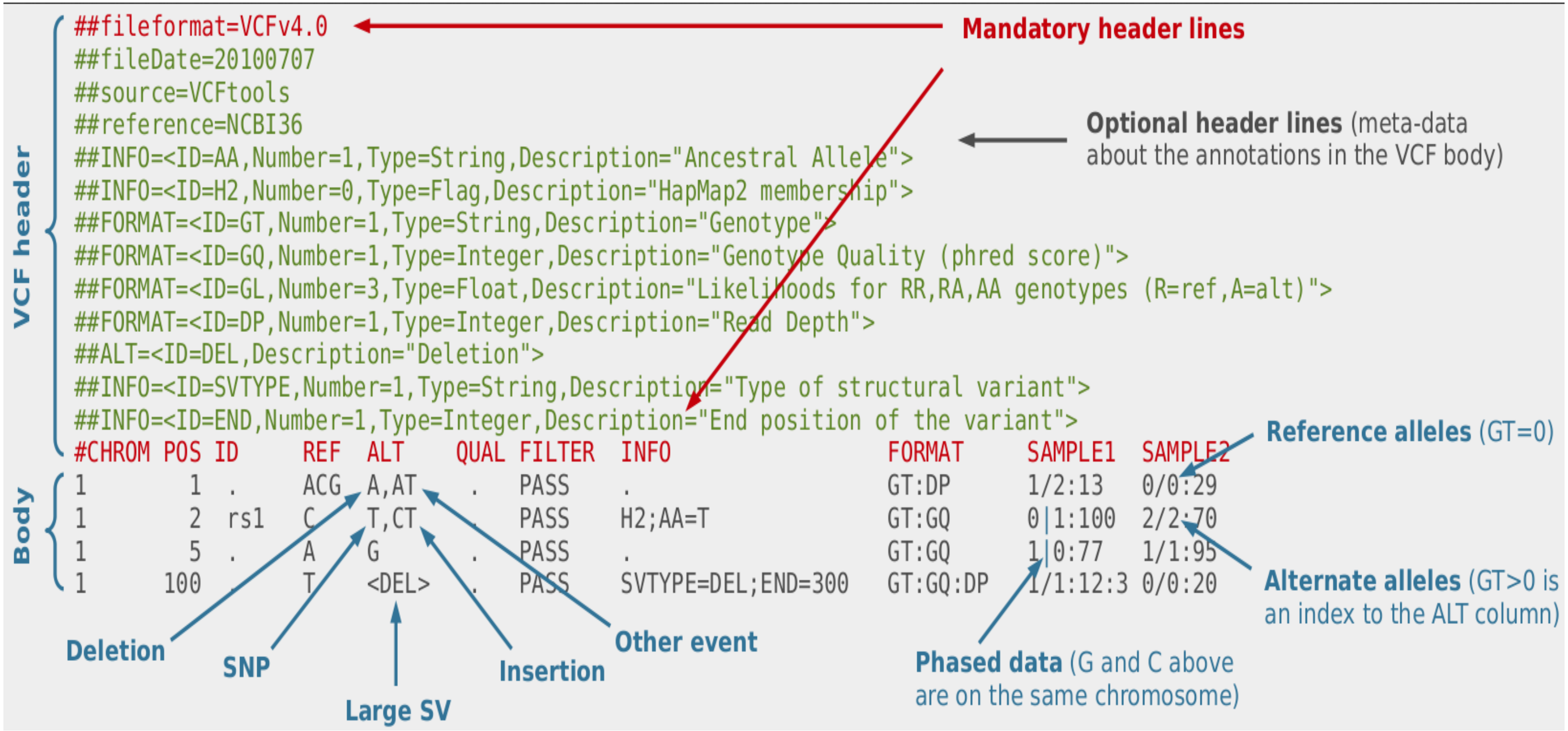
Campylobacter coli and C. jejuni, Listeria monocytogenes and klebsiella pneumoniae

- **Número de SNPs reportado por cada laboratorio parcipante**



Lab	EC	SA	ST
GMI02	25731	1383	8968
GMI04	25731	1383	8968
GMI06	43264	6226	5822
GMI10	13083	1797	12902
GMI14	14687	NA	1431
GMI26	92831	6164	31044
GMI39	52590	2672	16034
GMI42	9460	NA	12884
GMI43	38532	4163	16562
GMI46	63273	2341	9958
GMI48	67034	2063	14080
GMI58	79231	NA	19656
GMI59	23561	2715	14199
GMI13	9276	1628	8746
GMI16	55473	2122	13630
GMI21	5187829	2837196	5090636
GMI22	33416	1597	13066
GMI27	33664	2130	13297
GMI30	607217	11881	12733
GMI31	NA	NA	4141
GMI32	14667	25949	28164
GMI33	71822	5420	21668
GMI35	6706	1334	NA
GMI37	73355	2897	14294
GMI40	45725	2033	11180
GMI44	35039	1836	9446
GMI45	5183821	2836332	5088344
GMI47	20707	1805	12198
GMI50	84	NA	1300
GMI51	35521	NA	10042
GMI55	NA	1644	9102
GMI61	NA	NA	24
GMI63	NA	2834703	5077509
GMI7	21731	1673	9192
GMI8	15972	1851	12979

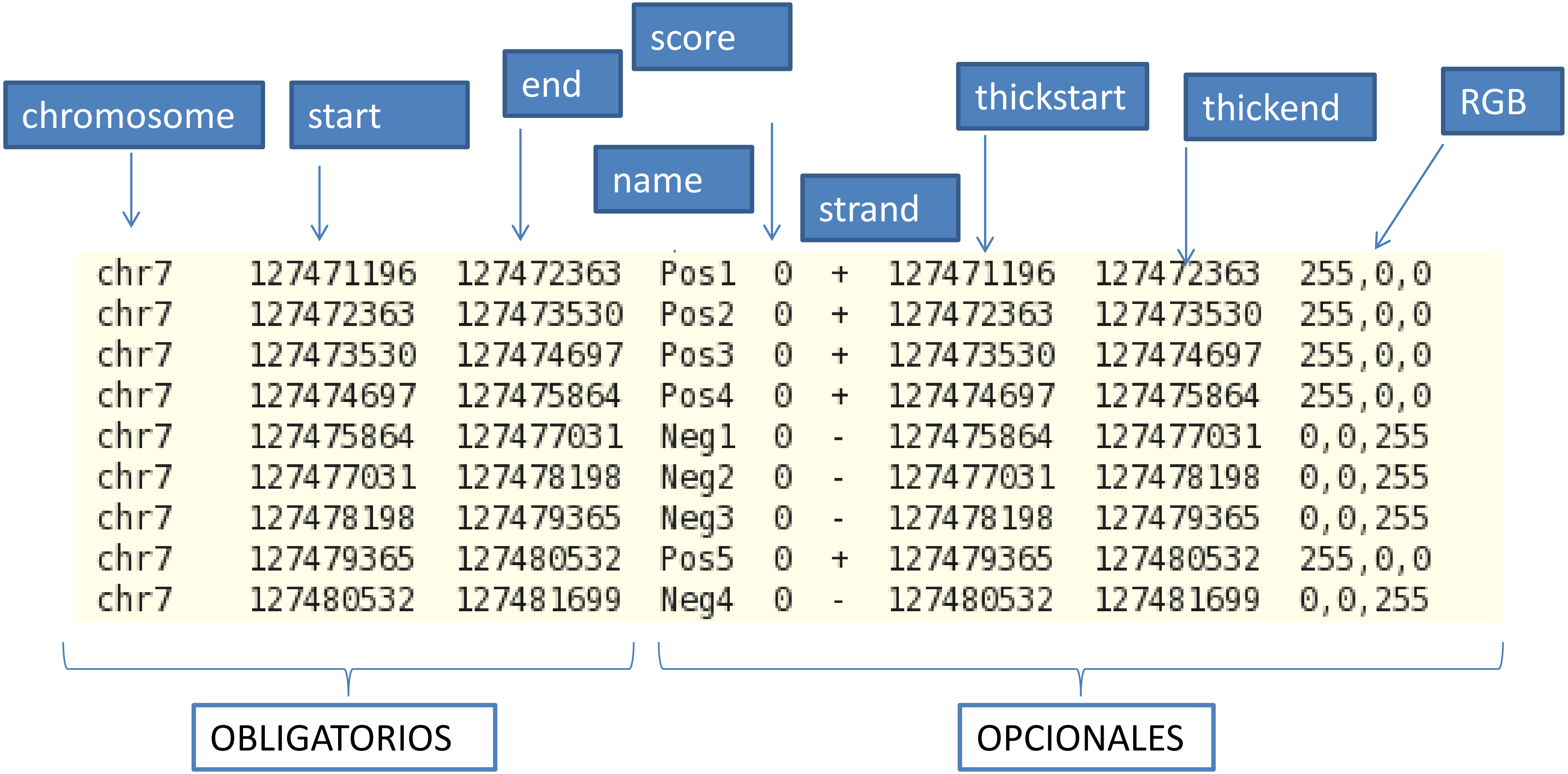
VCF format



```
##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)
Phased data (G and C above are on the same chromosome)
Deletion
SNP
Large SV
Insertion
Other event

Bed format



chromosome

start

end

score

name

strand

thickstart

thickend

RGB

```
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
```

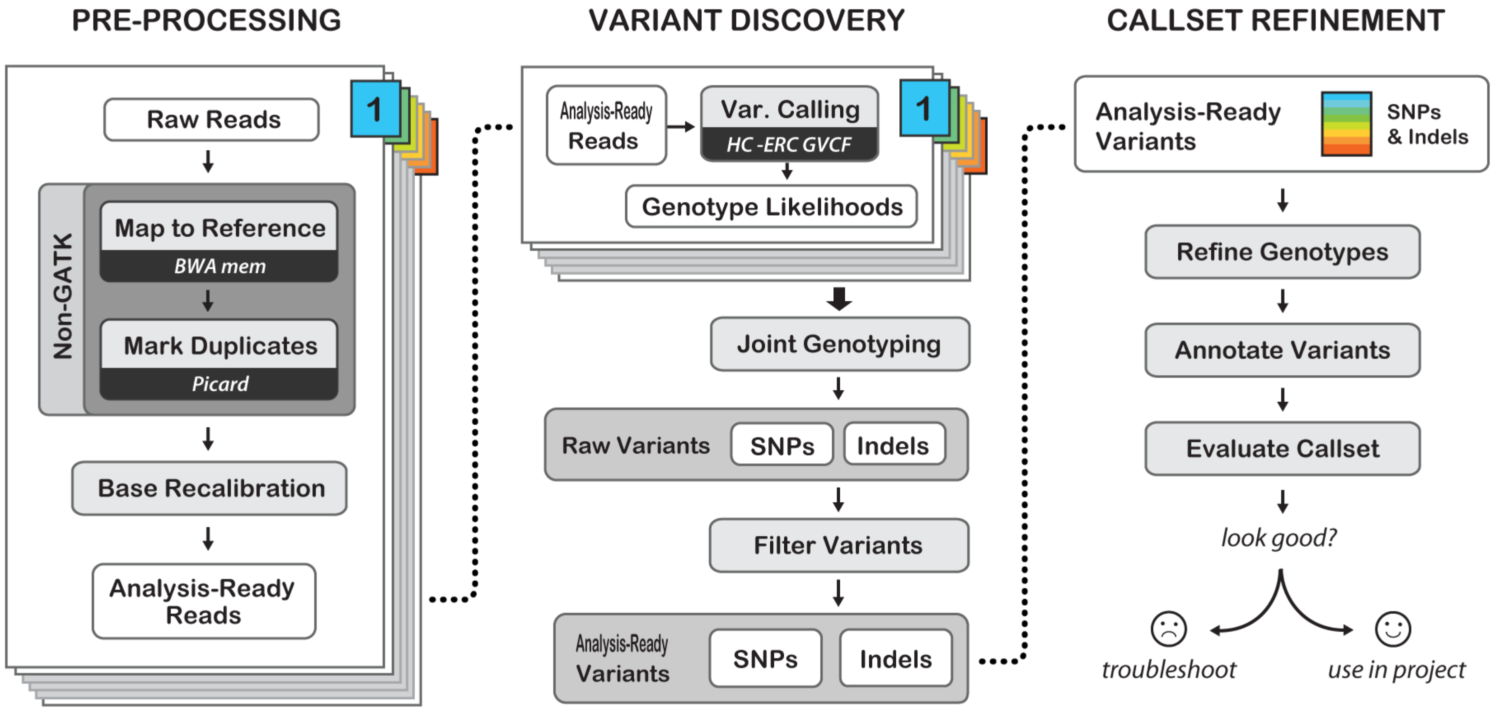
OBLIGATORIOS

OPCIONALES

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
CSI phylogeny	Variant calling with samtools.	Web	Kaas et al., 2014

GATK



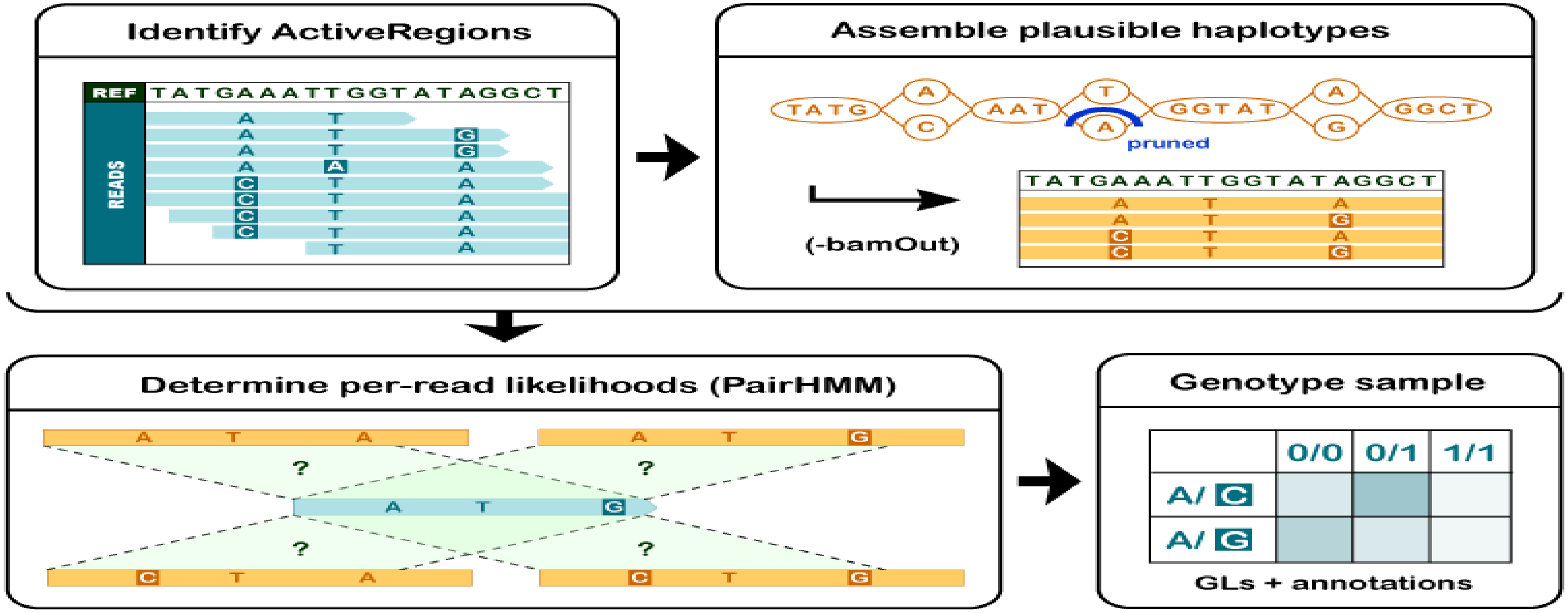
Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

GATK

- **1. Define active regions**
  - The program determines which regions of the genome it needs to operate on, based on the presence of significant evidence for variation.
- **2. Determine haplotypes by assembly of the active region**
  - For each ActiveRegion, the program builds a De Bruijn-like graph to reassemble the ActiveRegion, and identifies what are the possible haplotypes present in the data. The program then realigns each haplotype against the reference haplotype using the Smith-Waterman algorithm in order to identify potentially variant sites.
- **3. Determine likelihoods of the haplotypes given the read data**
  - For each ActiveRegion, the program performs a pairwise alignment of each read against each haplotype using the PairHMM algorithm. This produces a matrix of likelihoods of haplotypes given the read data. These likelihoods are then marginalized to obtain the likelihoods of alleles for each potentially variant site given the read data.
- **4. Assign sample genotypes**
  - For each potentially variant site, the program applies Bayes' rule, using the likelihoods of alleles given the read data to calculate the likelihoods of each genotype per sample given the read data observed for that sample. The most likely genotype is then assigned to the sample.

Best GATK practice guide.

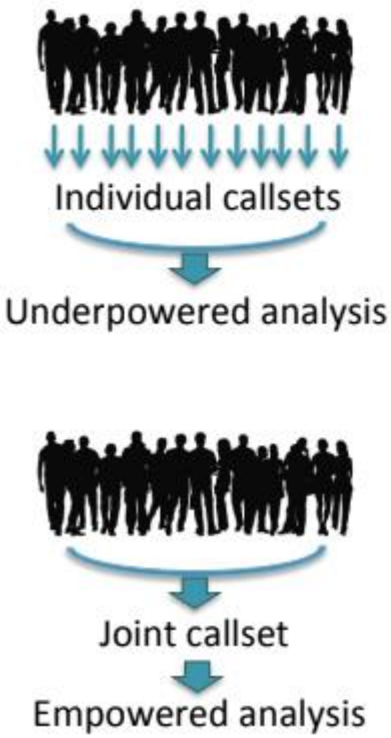
GATK


Identify ActiveRegions
REF
TATGAAATTGGTATAGGCT
READS
Assemble plausible haplotypes
TATG
A
C
AAT
T
A
pruned
GGTAT
A
G
GGCT
(-bamOut)
TATGAAATTGGTATAGGCT
Determine per-read likelihoods (PairHMM)
?
Genotype sample
0/0
0/1
1/1
A/C
A/G
GLs + annotations

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



Individual callsets

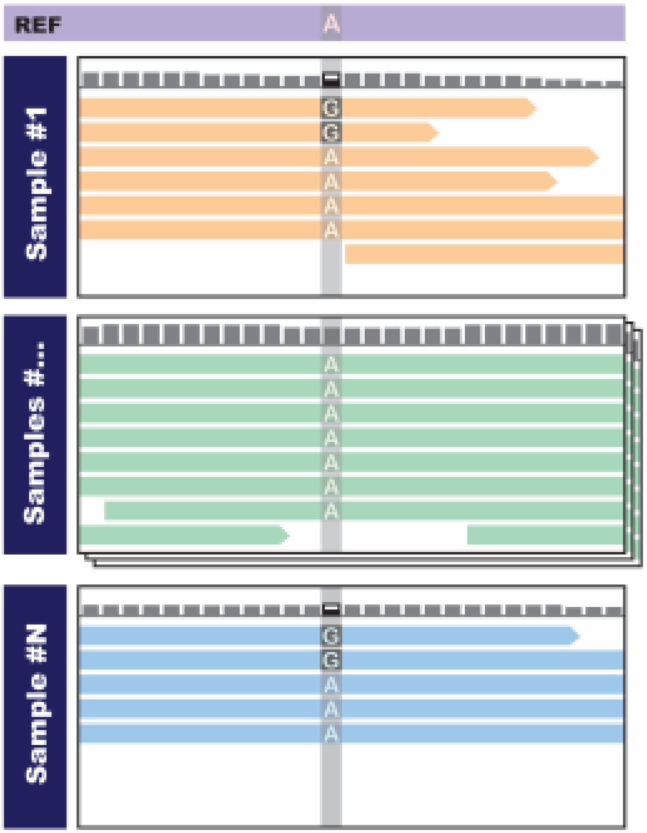
Underpowered analysis

Joint callset

Empowered analysis

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

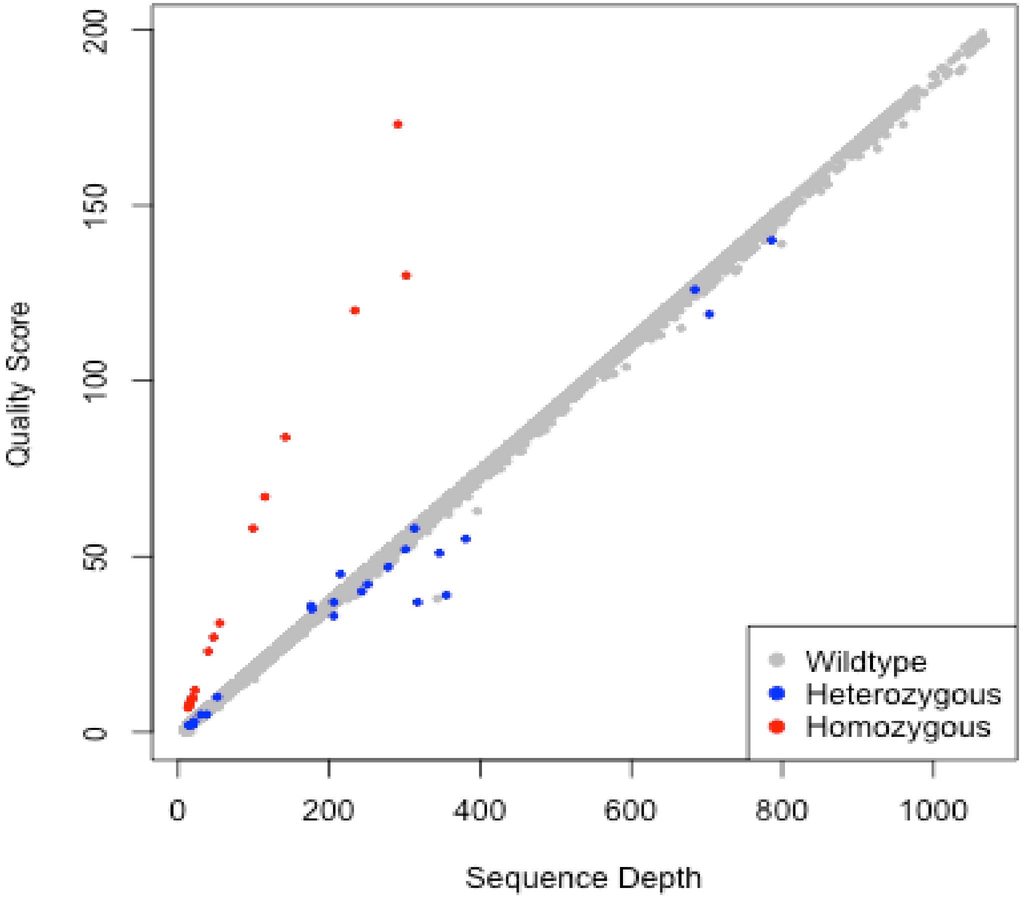
GATK problems

- Haploid variant calling is a side project. GATK is mainly for diploid organism, and the development and improvement of the haploid algorithm is slow.
- Manual filtering is needed for high quality SNPs selection.

VARSCAN2

- Uses a heuristic/statistic method instead of bayesian.
- Allows more flexibility and hard filters.
- Used in many bacterial variant calling pipelines. P.e CFSAN snp-pipeline.

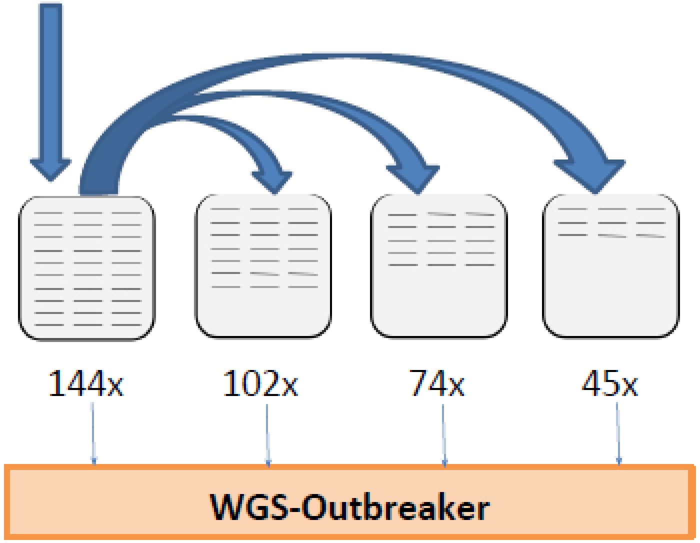
VARSCAN2


Quality Score
0
50
100
150
200
0
200
400
600
800
1000
Sequence Depth
● Wildtype
● Heterozygous
● Homozygous

Comparative VARSCAN - GATK

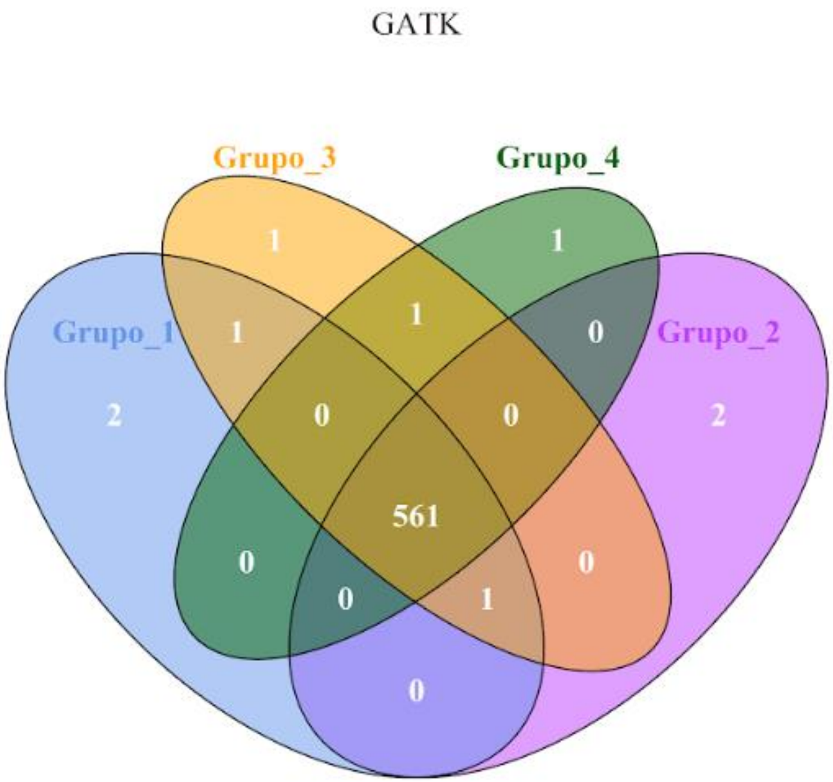
Año: 2016
41 aislados secuenciados

NextSeq 500
Paired end 150 bp

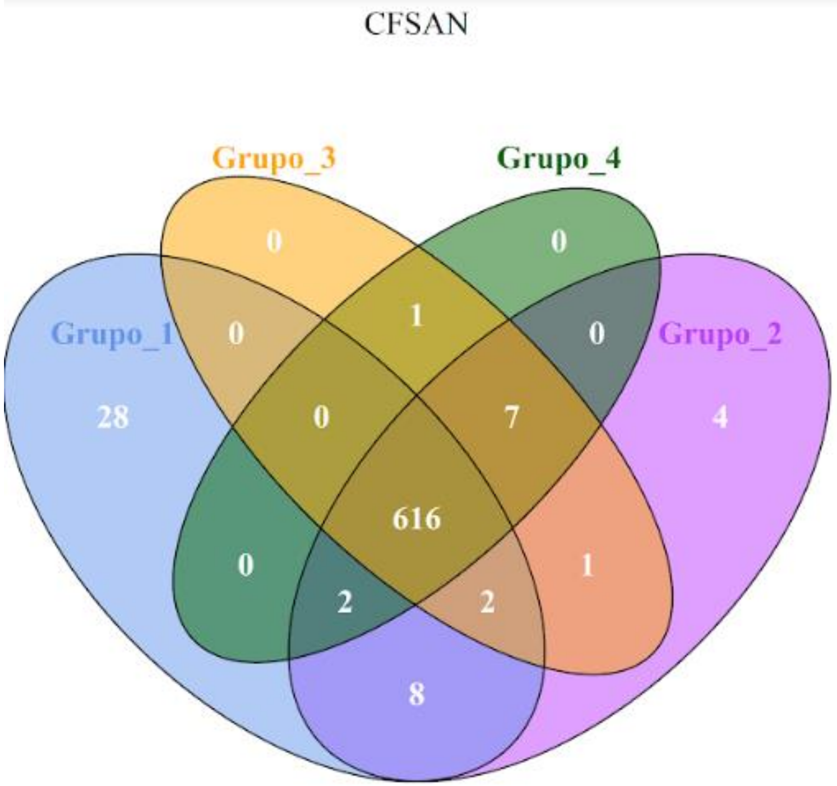


Comparative VARSCAN - GATK

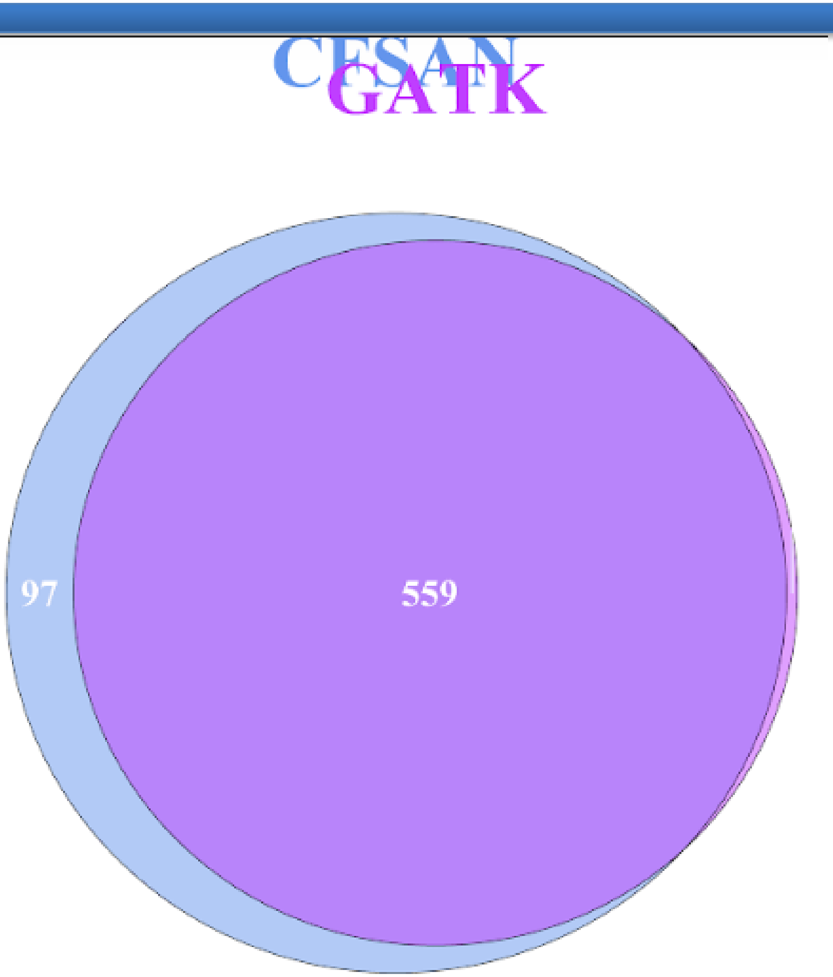
WGS-Outbreaker - GATK



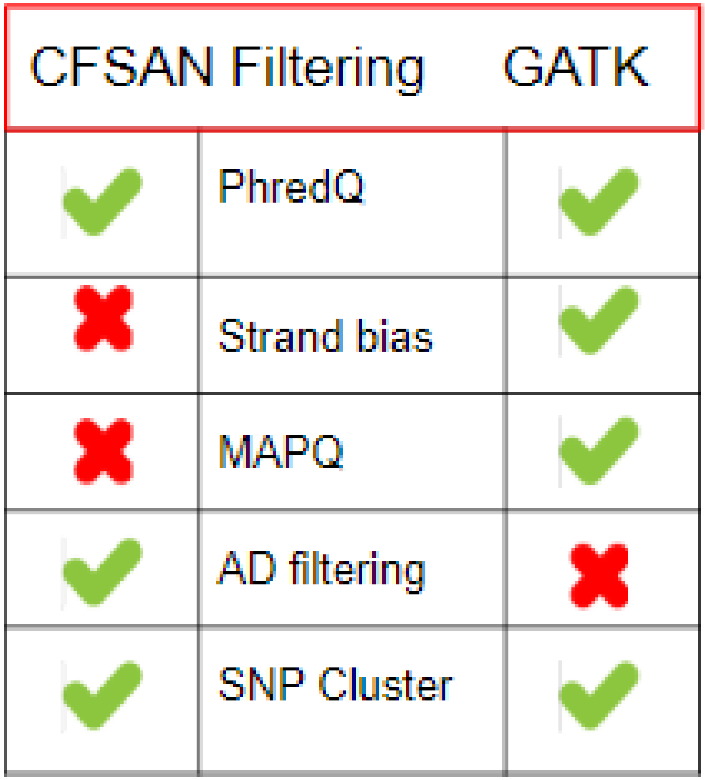
CFSAN - VARSCAN



Comparative VARSCAN - GATK



High Quality SNP selection



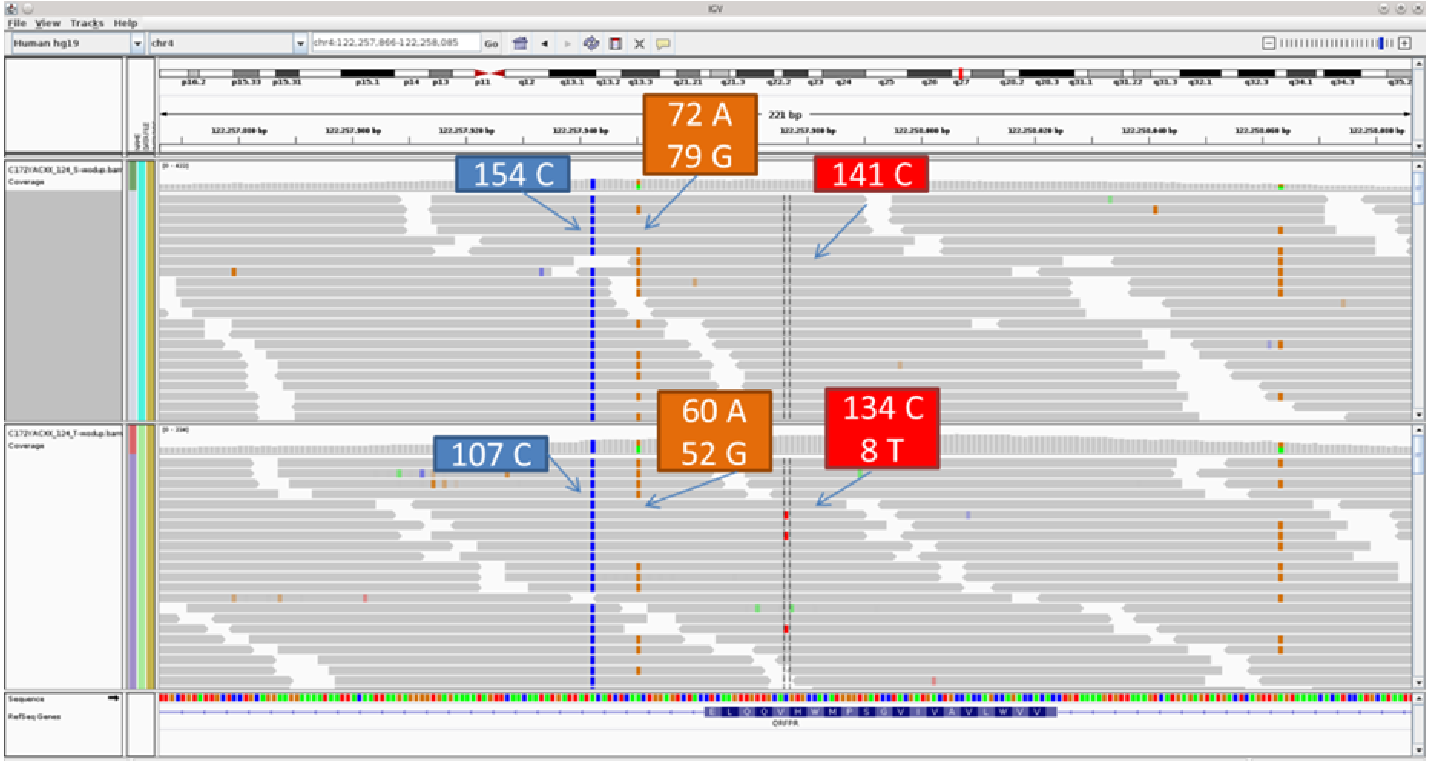
CFSAN Filtering		GATK
✓	PhredQ	✓
✗	Strand bias	✓
✗	MAPQ	✓
✓	AD filtering	✗
✓	SNP Cluster	✓

Population Allele frequency vs Sample Allele frequency

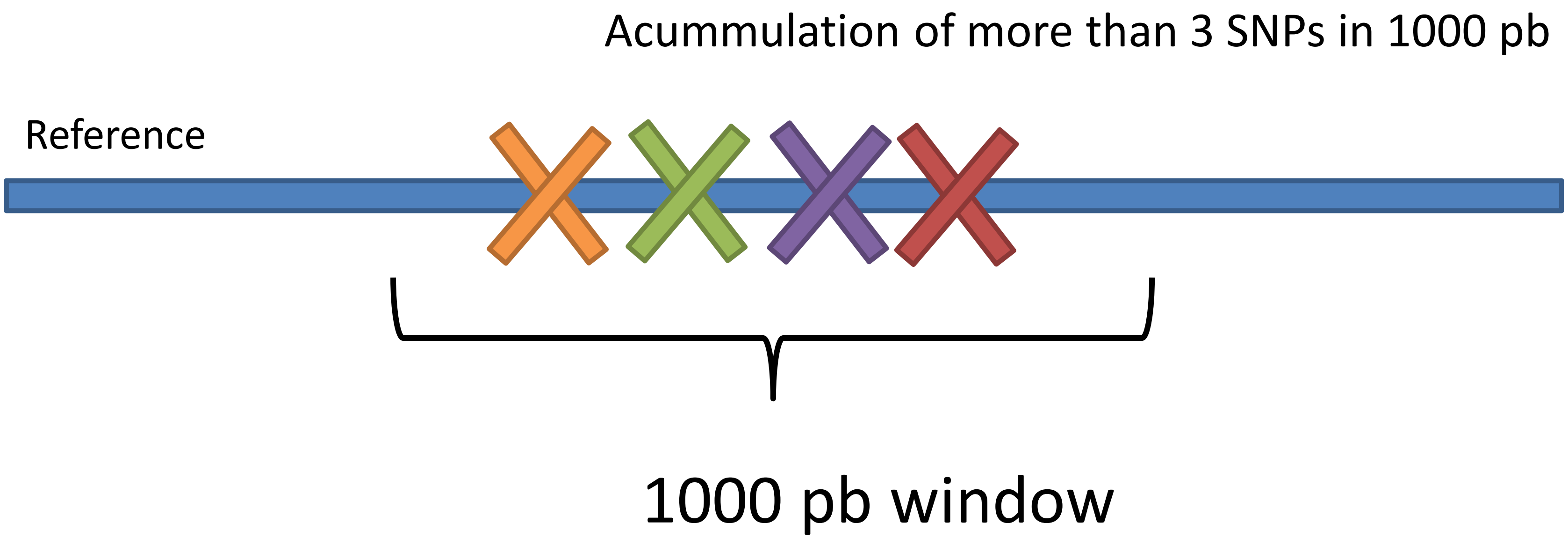
- **Population allele frequency:** probability of finding an allele in the population. Number of individuals carrying an allele vs total of individuals in the population.

Population Allele frequency vs Sample Allele frequency

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



SNP cluster filtering



What’s next?

SNP matrix creation

And

Phylogeny!

Thanks for your attention!