H3ABioNet
Pan African Bioinformatics Network for H3Africa

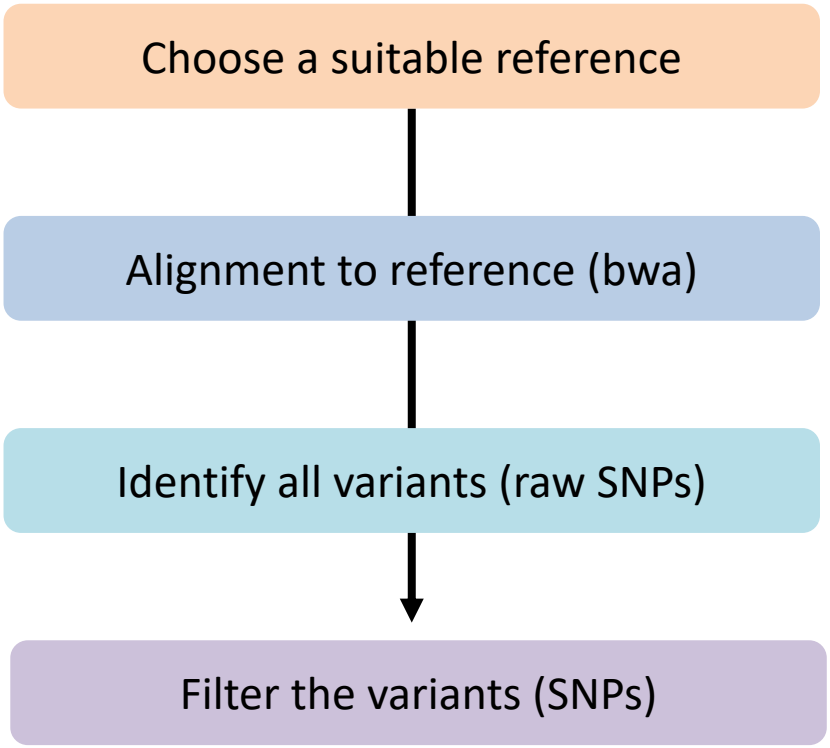
wellcome connecting science

Next Generation Sequencing Bioinformatics Course 2021

Module 6 – Pathogen variant calling

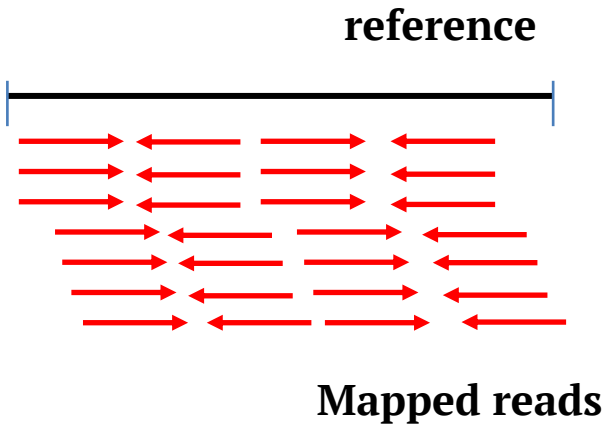
Session 1: Alignment to reference

Detection of variants (Variant calling)

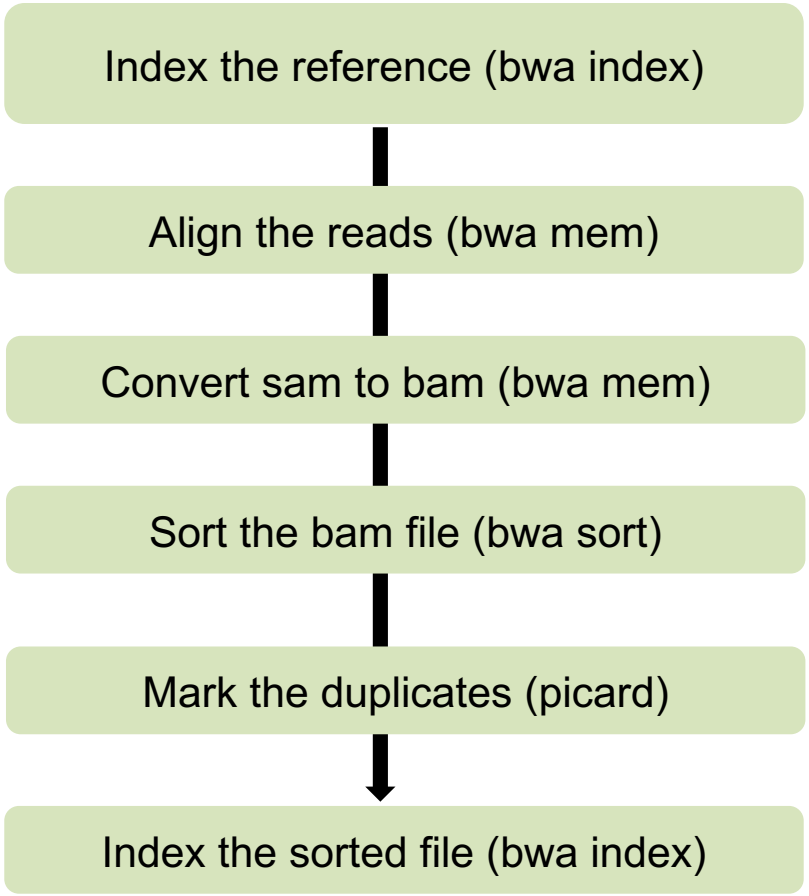


Selecting suitable reference

- Best source: NCBI reference database or prior published studies
- Complete genome sequence available
  - NCBI has versions : select the latest
- Genome annotation
- From the same species or close relative
- Present case : *M. tuberculosis* H37Rv (NC_000962.3)



Step1: Alignment to reference



Duplicates

1) picard MarkDuplicates I=sorted.bam O=markdup.bam M=metrics.txt

2) grep –A 2 "^## METRICS" metrics.txt

```
1 ## METRICS CLASS        picard.sam.DuplicationMetrics
2 LIBRARY         UNPAIRED_READS_EXAMINED
  READ_PAIRS_EXAMINED     SECONDARY_OR_SUPPLEMENTARY_RDS
  UNMAPPED_READS  UNPAIRED_READ_DUPLICATES
  READ_PAIR_DUPLICATES    READ_PAIR_OPTICAL_DUPLICATES
  PERCENT_DUPLICATION     ESTIMATED_LIBRARY_SIZE
3 Unknown Library         231556  2524511         30570   289914
  120543  300973  49885   0.13682         11354842
```

Mapping statistics

1) samtools stats in.bam>bamstat.txt

2) grep "^SN" bamstats.txt >stats.txt

```
SN	raw total sequences:	5570492
SN	filtered sequences:	0
SN	sequences:	5570492
SN	is sorted:	1
SN	1st fragments:	2785246
SN	last fragments:	2785246
SN	reads mapped:	5280578
SN	reads mapped and paired:	5049022	# paired-end technology bit set + both mates mapped
SN	reads unmapped:	289914
SN	reads properly paired:	4995486	# proper-pair bit set
SN	reads paired:	5570492	# paired-end technology bit set
SN	reads duplicated:	722489	# PCR or optical duplicate bit set
SN	reads MQ0:	69525	# mapped and MQ=0
SN	reads QC failed:	0
SN	non-primary alignments:	0
SN	total length:	835573800	# ignores clipping
SN	total first fragment length:	417786900	# ignores clipping
SN	total last fragment length:	417786900	# ignores clipping
SN	bases mapped:	792086700	# ignores clipping
SN	bases mapped (cigar):	776568548	# more accurate
SN	bases trimmed:	0
SN	bases duplicated:	108373350
SN	mismatches:	5722273	# from NM fields
SN	error rate:	7.368664e-03	# mismatches / bases mapped (cigar)
SN	average length:	150
SN	average first fragment length:	150
SN	average last fragment length:	150
SN	maximum length:	150
SN	maximum first fragment length:	150
SN	maximum last fragment length:	150
SN	average quality:	35.9
SN	insert size average:	350.1
SN	insert size standard deviation:	109.3
SN	inward oriented pairs:	2483610
SN	outward oriented pairs:	34508
SN	pairs with other orientation:	19681
SN	pairs on different chromosomes:	0
SN	percentage of properly paired reads (%):	89.7
```

Thank you