INTRODUCTION TO UNIX-LINUX – ALIGNMENT/MAPPING

Klebsiella Workshop

Sep 2024

ASLM AFRICAN SOCIETY FOR LABORATORY MEDICINE

UNIVERSITY *of the* WESTERN CAPE

SANBI South African National Bioinformatics Institute

BIG DATA INSTITUTE

UNIVERSITY OF OXFORD

African Union

AFRICA CDC Centres for Disease Control and Prevention Safeguarding Africa's Health

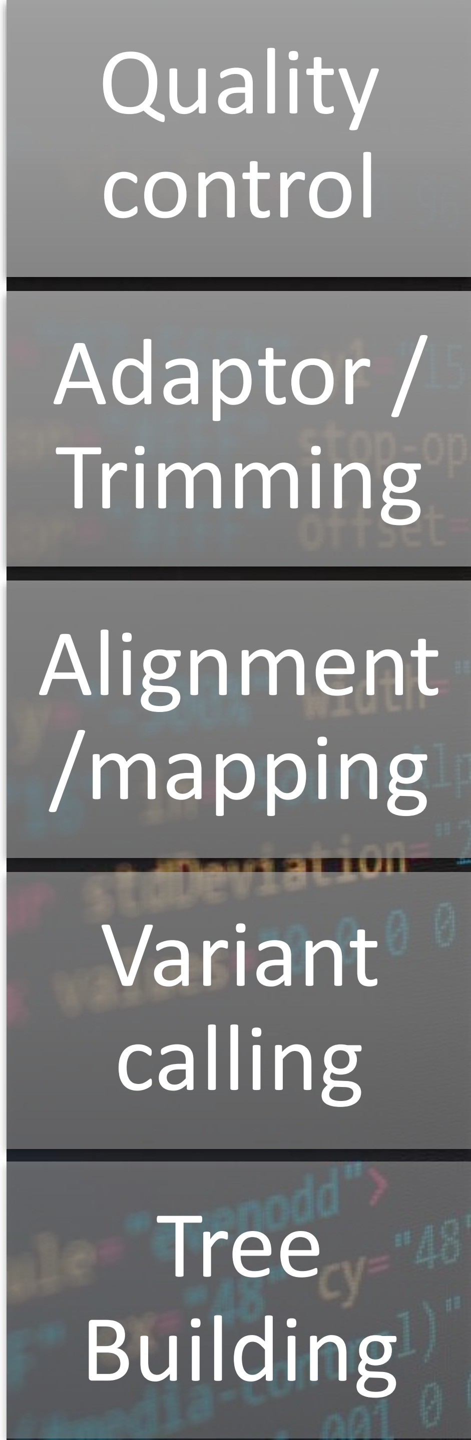
PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Introduction to Alignment / Mapping

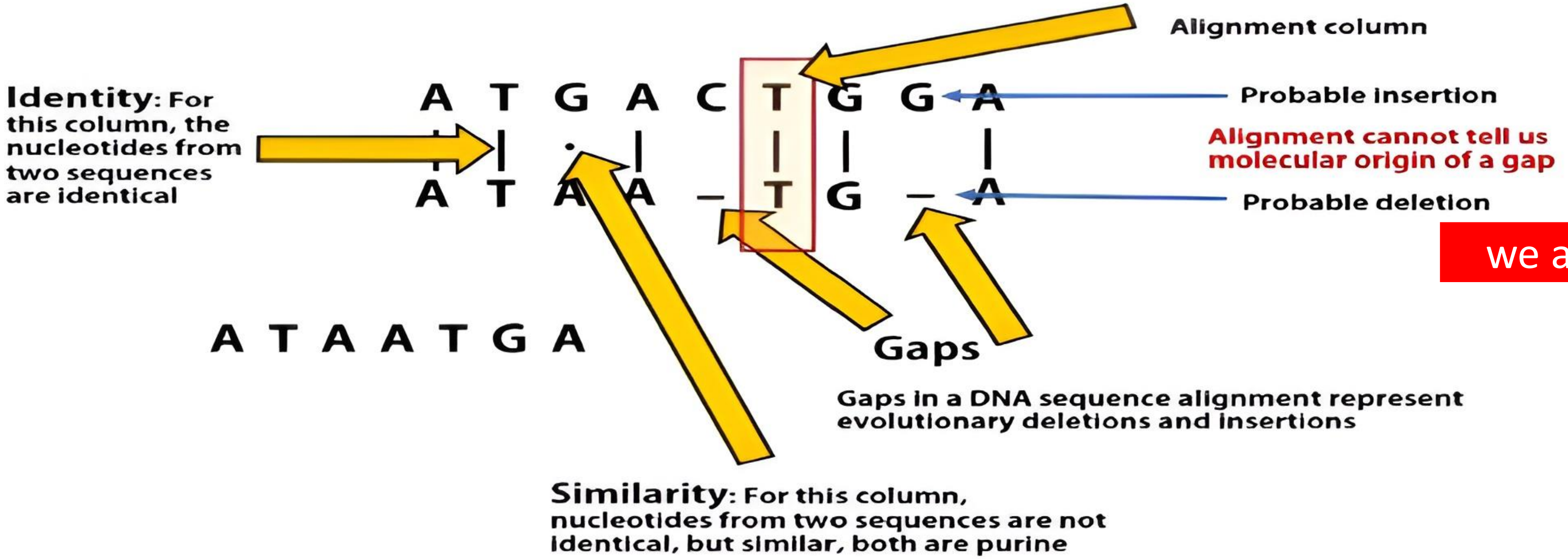
- A sequence alignment is a way of arranging the primary sequence/reads of DNA, RNA or Proteins to identify regions of similarity that may be a consequences of FUNCTIONAL, STRUCTURAL and EVOLUTIONARY relationship between the sequences

we are here

- Sequence alignment is the procedure of comparing two (pair-wise alignment) or more (multiple sequences) by searching for a series of individual characters/patterns that are in the same order in the sequences.



Introduction to Alignment / Mapping cont.



Alignment column

Probable insertion

Alignment cannot tell us molecular origin of a gap

Probable deletion

Identity: For this column, the nucleotides from two sequences are identical

Gaps

Gaps in a DNA sequence alignment represent evolutionary deletions and insertions

Similarity: For this column, nucleotides from two sequences are not identical, but similar, both are purine

- Known sequence is the REFERENCE SEQUENCE, and unknown is the QUERY SEQUENCE.
- Align QUERY SEQUENCE to CONTIGS

we are here

Quality control

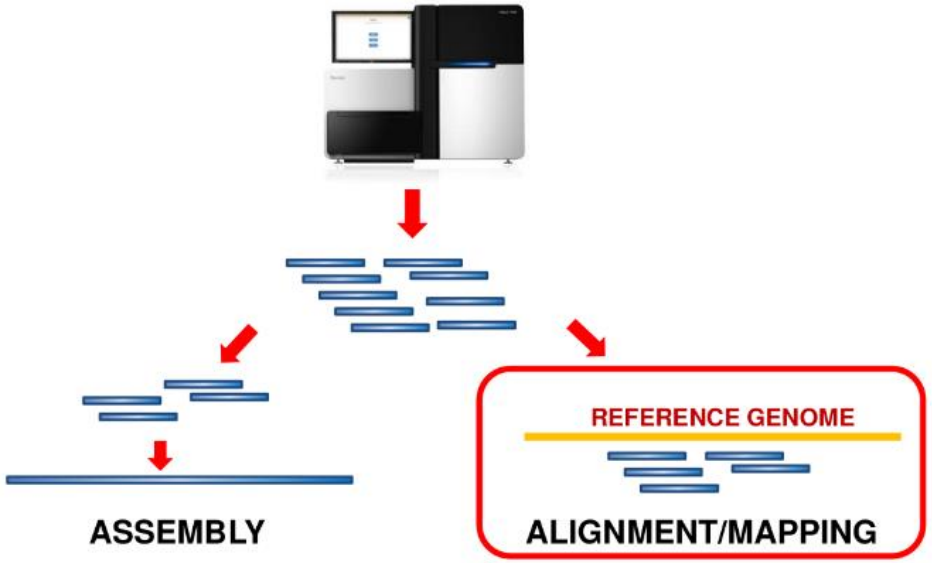
Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

What is Alignment / Mapping



- Short reads must be combined into larger fragments
- **Mapping:** use a reference genome as a guide
- **Alignment:** referred to as alignment of query sequences to contigs. Also have gene alignment
- **De Novo assembly:** without a reference genome

we are here

Quality control

Adaptor / Trimming

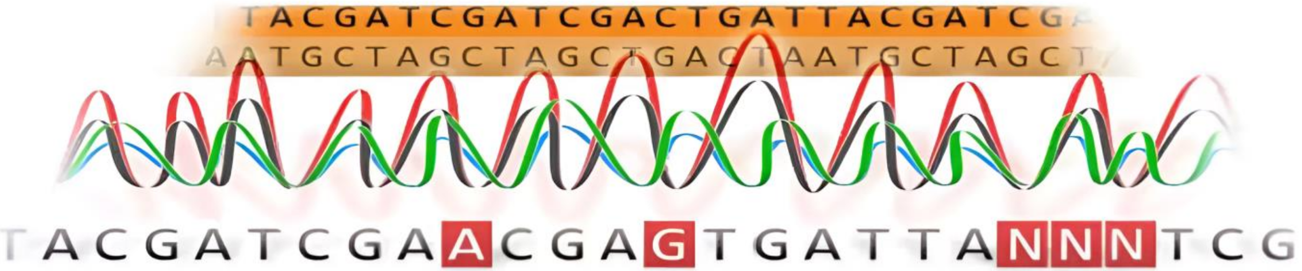
Alignment /mapping

Variant calling

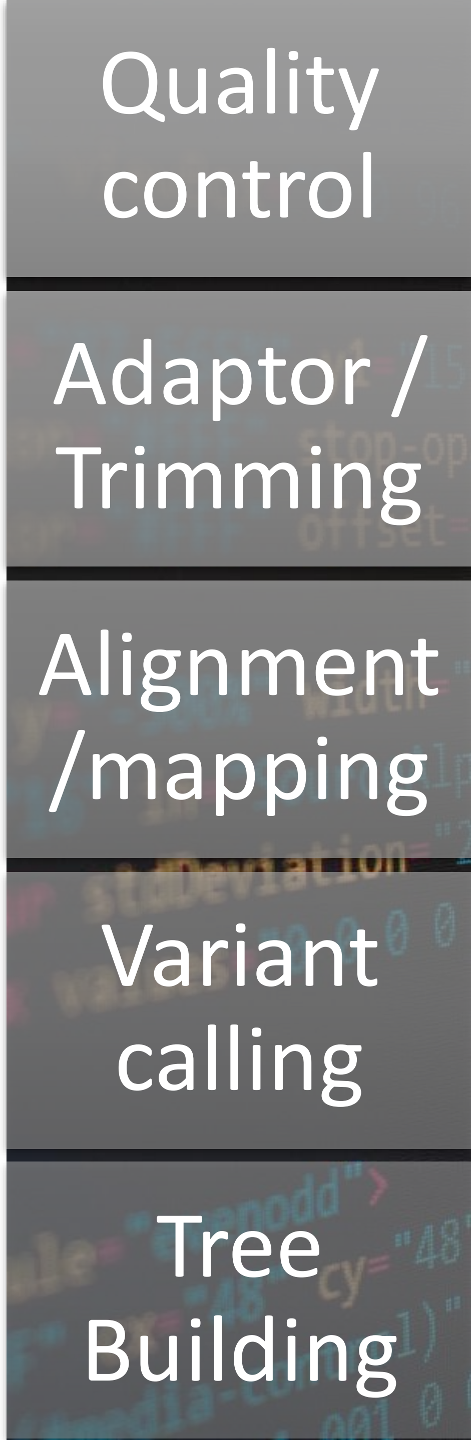
Tree Building

Why is Alignment needed?

- To compare sequences for similarities and differences
  - Mutations
- Often, we are looking for similarities
- Homology: similarity due to descent from a common ancestor
- For traceability
- To infer differences and similarities in structures, function and sequences


TACGATCGATCGACTGATTACGATCGA
AATGCTAGCTAGCTGACTAATGCTAGCT
TACGATCGAACGAGTGATTANNNTCG

we are here



Quality control

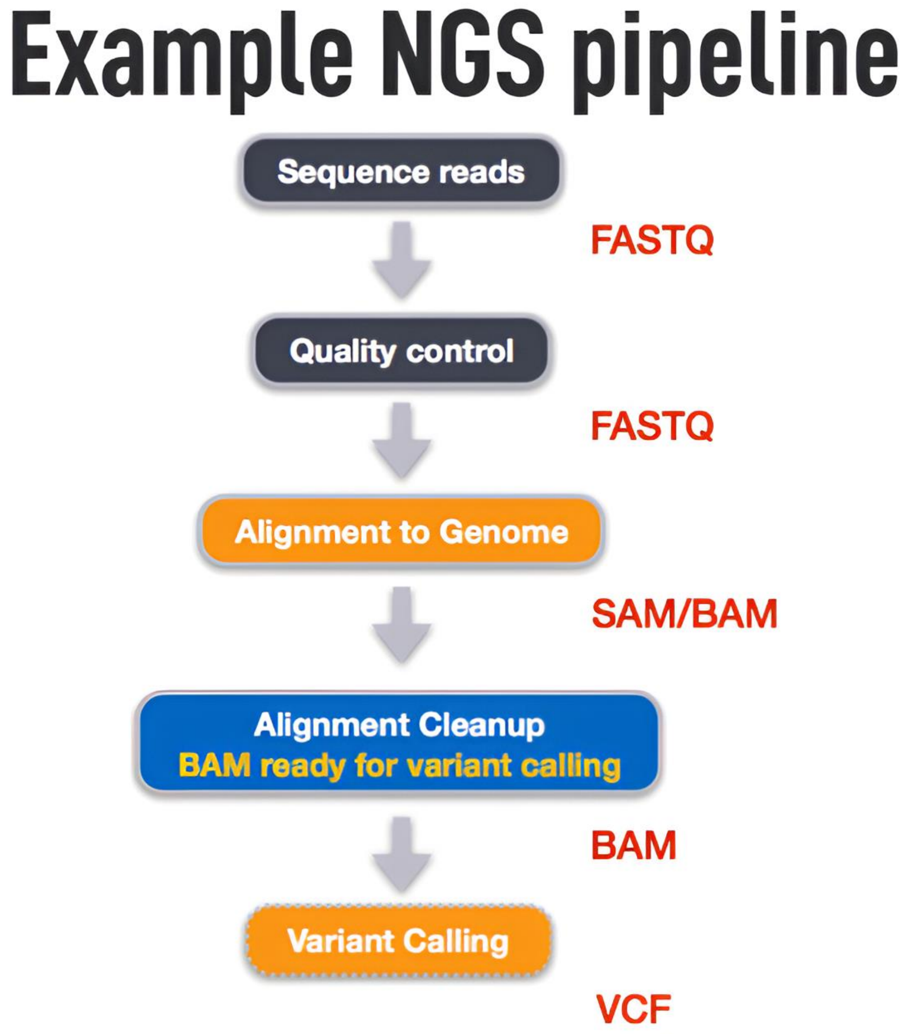
Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

Mapping in Bioinformatics



A high level view of a typical NGS bioinformatics workflow

we are here

Quality control

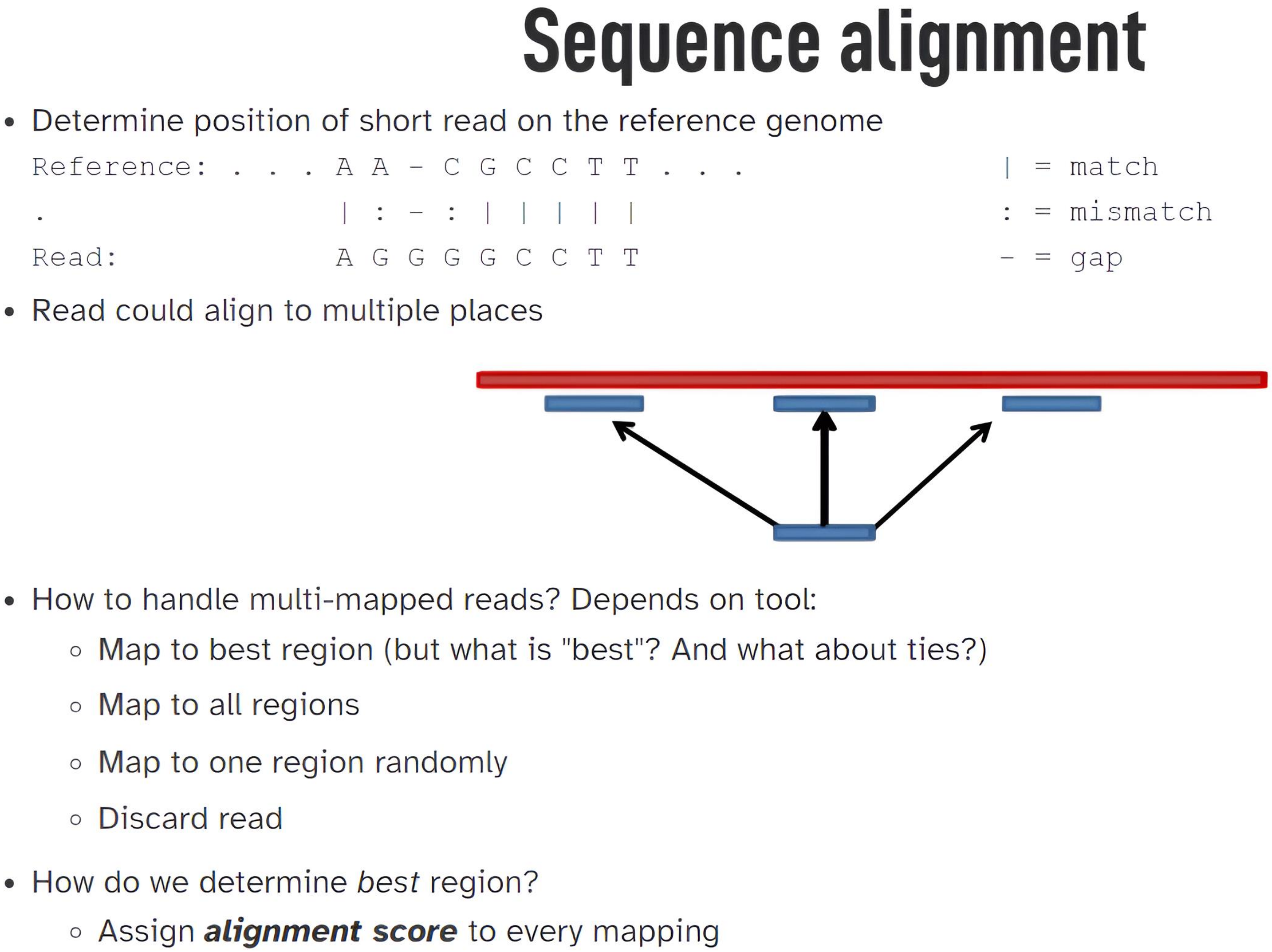
Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

Mapping in Bioinformatics



Sequence alignment

- Determine position of short read on the reference genome

```
Reference: . . . A A - C G C C T T . . .               | = match
.                | : - : | | | | |                     : = mismatch
Read:            A G G G G C C T T                     - = gap
```

- Read could align to multiple places
- How to handle multi-mapped reads? Depends on tool:
  - Map to best region (but what is "best"? And what about ties?)
  - Map to all regions
  - Map to one region randomly
  - Discard read
- How do we determine *best* region?
  - Assign ***alignment score*** to every mapping

we are here

Quality control

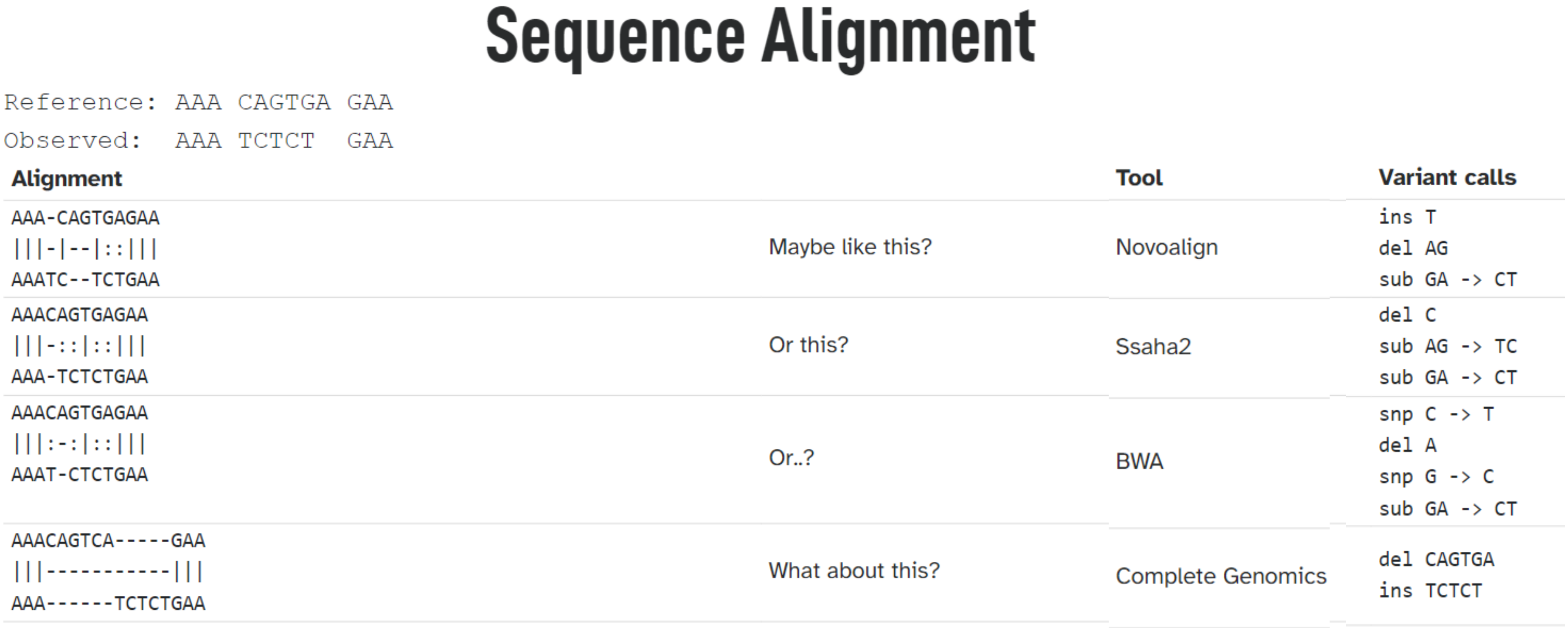
Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

Mapping in Bioinformatics



Sequence Alignment

```
Reference: AAA CAGTGA GAA
Observed:  AAA TCTCT  GAA
```

<table>
<tr><th>Alignment</th><th></th><th>Tool</th><th>Variant calls</th></tr>
<tr><td>AAA-CAGTGAGAA
|||-|--|::|||
AAATC--TCTGAA</td><td>Maybe like this?</td><td>Novoalign</td><td>ins T
del AG
sub GA -> CT</td></tr>
<tr><td>AAACAGTGAGAA
|||-::|::|||
AAA-TCTCTGAA</td><td>Or this?</td><td>Ssaha2</td><td>del C
sub AG -> TC
sub GA -> CT</td></tr>
<tr><td>AAACAGTGAGAA
|||:-:|::|||
AAAT-CTCTGAA</td><td>Or..?</td><td>BWA</td><td>snp C -> T
del A
snp G -> C
sub GA -> CT</td></tr>
<tr><td>AAACAGTCA-----GAA
|||-----------|||
AAA------TCTCTGAA</td><td>What about this?</td><td>Complete Genomics</td><td>del CAGTGA
ins TCTCT</td></tr>
</table>

Quality control

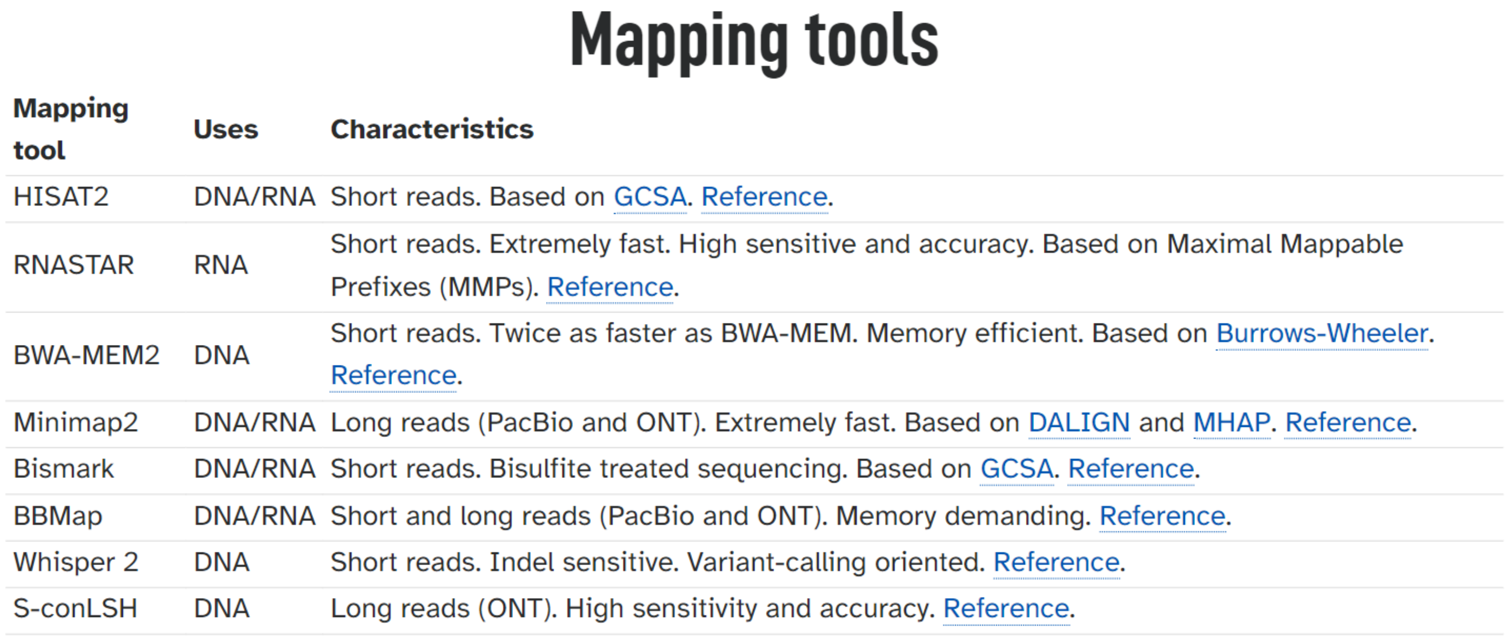
Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

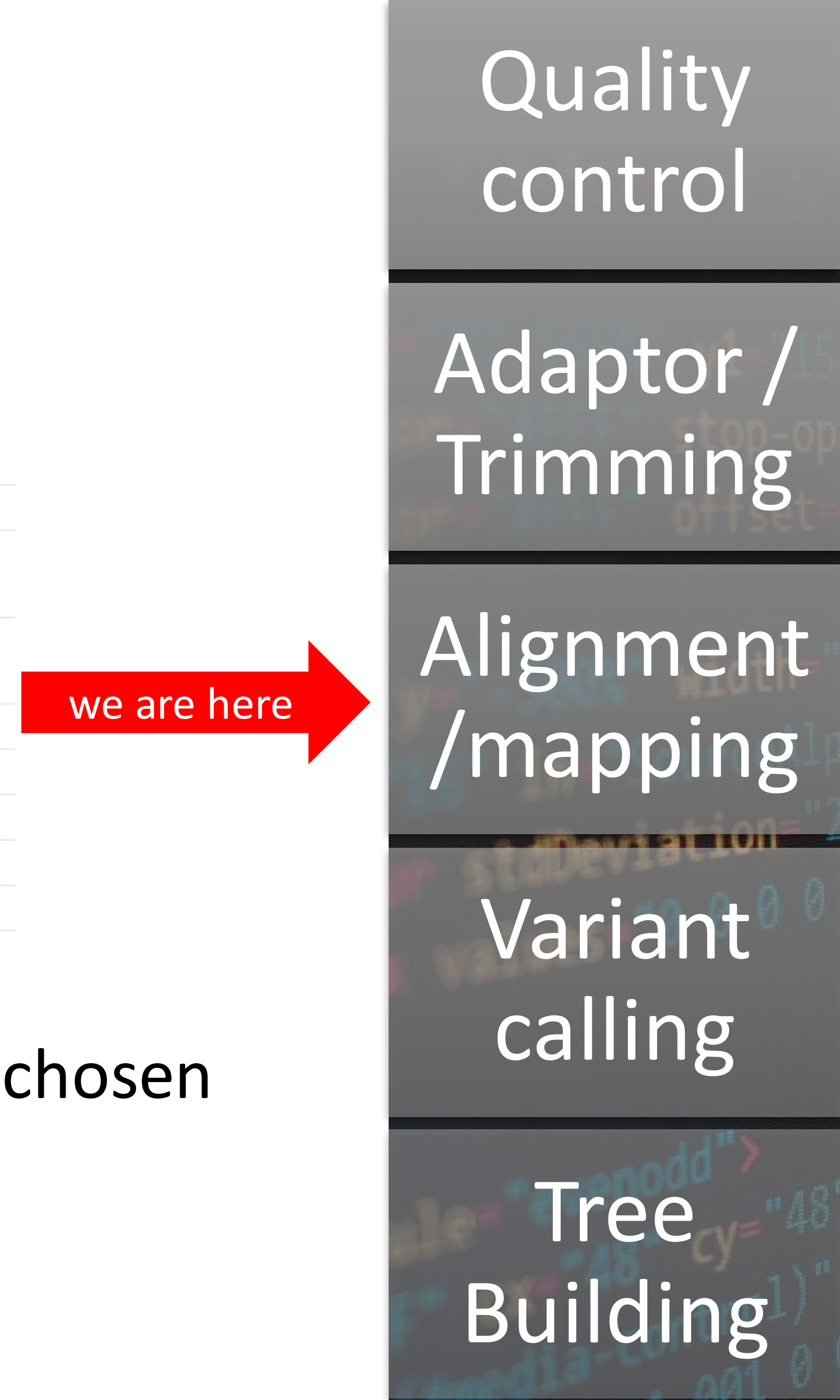
Mapping in Bioinformatics



Mapping tools

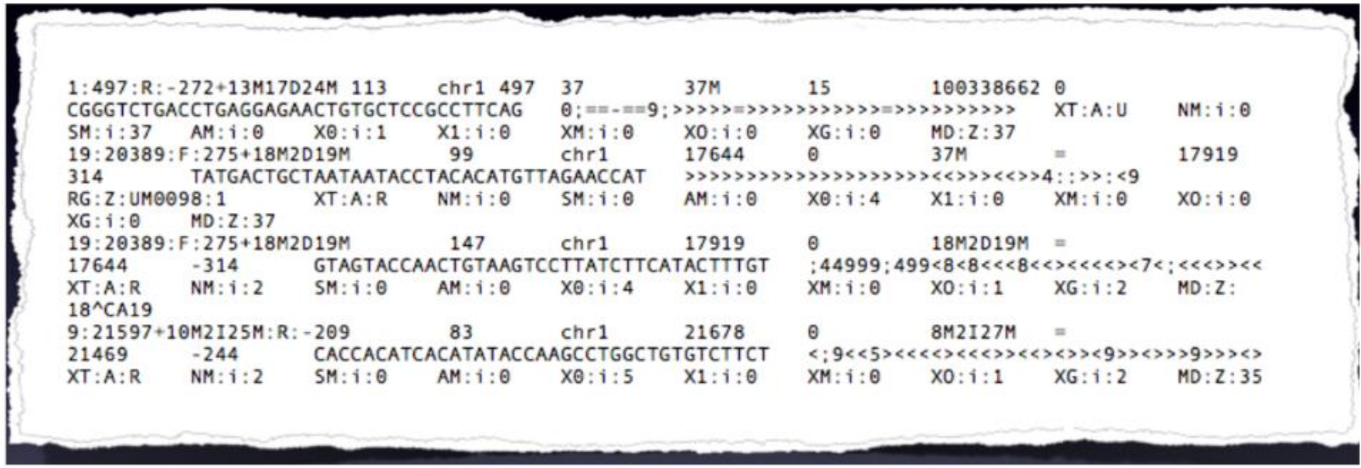
Mapping tool	Uses	Characteristics
HISAT2	DNA/RNA	Short reads. Based on GCSA. Reference.
RNASTAR	RNA	Short reads. Extremely fast. High sensitive and accuracy. Based on Maximal Mappable Prefixes (MMPs). Reference.
BWA-MEM2	DNA	Short reads. Twice as faster as BWA-MEM. Memory efficient. Based on Burrows-Wheeler. Reference.
Minimap2	DNA/RNA	Long reads (PacBio and ONT). Extremely fast. Based on DALIGN and MHAP. Reference.
Bismark	DNA/RNA	Short reads. Bisulfite treated sequencing. Based on GCSA. Reference.
BBMap	DNA/RNA	Short and long reads (PacBio and ONT). Memory demanding. Reference.
Whisper 2	DNA	Short reads. Indel sensitive. Variant-calling oriented. Reference.
S-conLSH	DNA	Long reads (ONT). High sensitivity and accuracy. Reference.

- There are many more tools, the mapping tool is chosen based on which best fits your data



Mapping in Bioinformatics

SAM/BAM file format



```
1:497:R:-272+13M17D24M 113    chr1 497  37        37M       15        100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>   XT:A:U    NM:i:0
SM:i:37   AM:i:0    X0:i:1    X1:i:0    XM:i:0    XO:i:0    XG:i:0    MD:Z:37
19:20389:F:275+18M2D19M        99       chr1      17644     0         37M       =         17919
314       TATGACTGCTAATAATACCTACACATGTTAGAACCAT   >>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9
RG:Z:UM0098:1       XT:A:R    NM:i:0    SM:i:0    AM:i:0    X0:i:4    X1:i:0    XM:i:0    XO:i:0
XG:i:0    MD:Z:37
19:20389:F:275+18M2D19M        147      chr1      17919     0         18M2D19M  =
17644     -314      GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT   ;44999;499<8<8<<<8<<><<<<><7<;<<<>><<
XT:A:R    NM:i:2    SM:i:0    AM:i:0    X0:i:4    X1:i:0    XM:i:0    XO:i:1    XG:i:2    MD:Z:
18^CA19
9:21597+10M2I25M:R:-209        83       chr1      21678     0         8M2I27M   =
21469     -244      CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<><<<>><<><>><9>><>>>9>>><>
XT:A:R    NM:i:2    SM:i:0    AM:i:0    X0:i:5    X1:i:0    XM:i:0    XO:i:1    XG:i:2    MD:Z:35
```

SAM: Sequence **A**lignment **M**ap

BAM: Binary (compressed) SAM; not human-readable

we are here

Quality control

Adaptor / Trimming

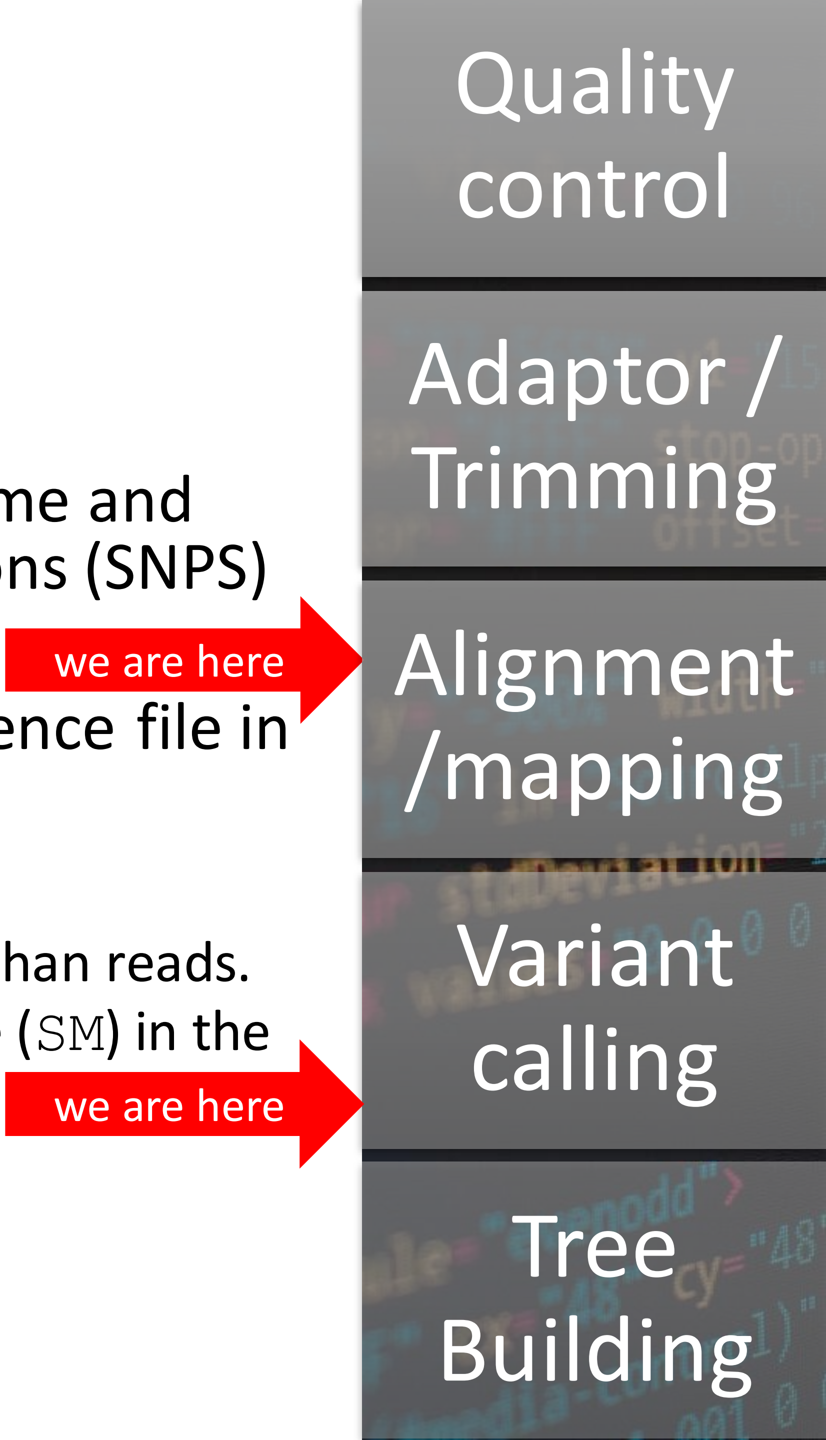
Alignment /mapping

Variant calling

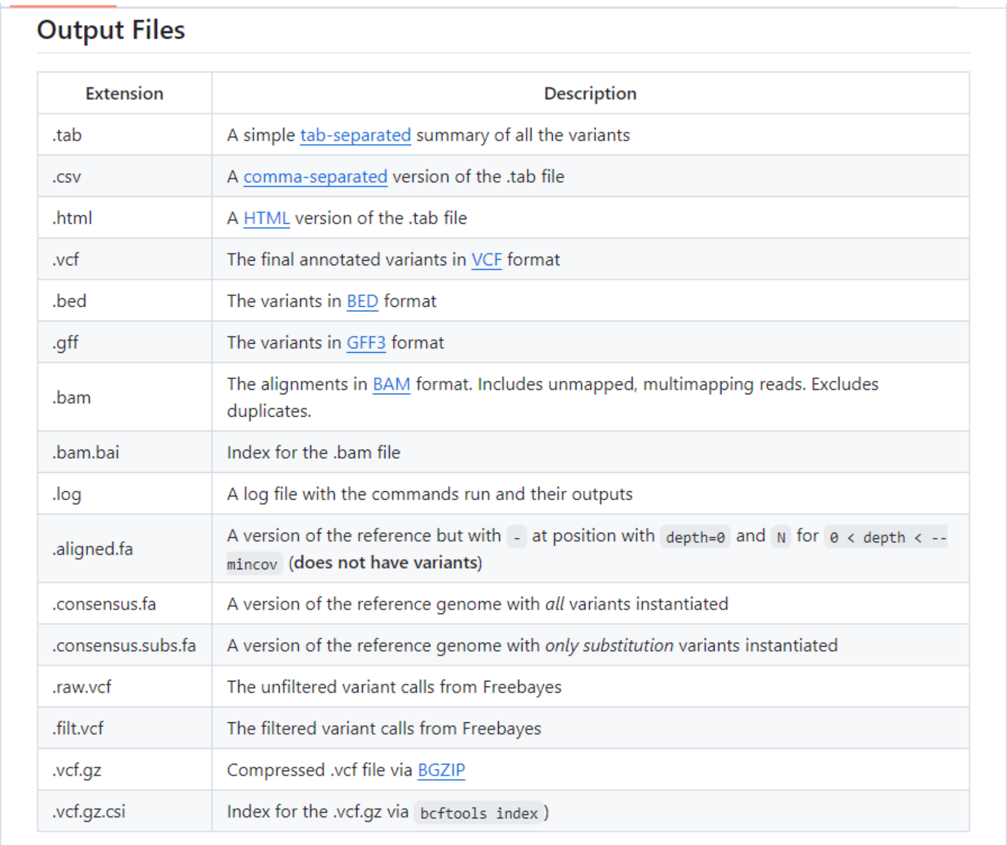
Tree Building

Tool: SNIPPY

- An alignment and variant calling tool – all-in-one
- Snippy finds SNPs between a haploid reference genome and your NGS sequence reads. It will find both substitutions (SNPS) and insertions/deletions (indels).
- Input: NGS Reads in fastq format (SE or PE) & a Reference file in either fasta or genbank format
- Various parameters can be defined "options"
  - `--contigs` allows you to call SNPs from contigs rather than reads.
  - `--rgid` will set the Read Group (`RG`) ID (`ID`) and Sample (`SM`) in the BAM and VCF file.
- Simplicity, input → output



How SNIPPY works



Output Files

Extension	Description
.tab	A simple tab-separated summary of all the variants
.csv	A comma-separated version of the .tab file
.html	A HTML version of the .tab file
.vcf	The final annotated variants in VCF format
.bed	The variants in BED format
.gff	The variants in GFF3 format
.bam	The alignments in BAM format. Includes unmapped, multimapping reads. Excludes duplicates.
.bam.bai	Index for the .bam file
.log	A log file with the commands run and their outputs
.aligned.fa	A version of the reference but with `-` at position with `depth=0` and `N` for `0 < depth < --mincov` **(does not have variants)**
.consensus.fa	A version of the reference genome with *all* variants instantiated
.consensus.subs.fa	A version of the reference genome with *only substitution* variants instantiated
.raw.vcf	The unfiltered variant calls from Freebayes
.filt.vcf	The filtered variant calls from Freebayes
.vcf.gz	Compressed .vcf file via BGZIP
.vcf.gz.csi	Index for the .vcf.gz via `bcftools index`)

- SNIPPY has the following incorporated:
  - BWA-MEM – maps individual sequence
  - SAM tools – converts SAM→ BAM
  - GNU parallel – executing job in parallel (isolates)
  - Freebayes – variant calling program
  - VCFlib – parsing and manipulating VCF files
  - VCFtools – output .vcf file
  - SNPEFF – Prediction tool, annotates and predicts the effects of genetic variants

Quality control

Adaptor / Trimming

Alignment /mapping

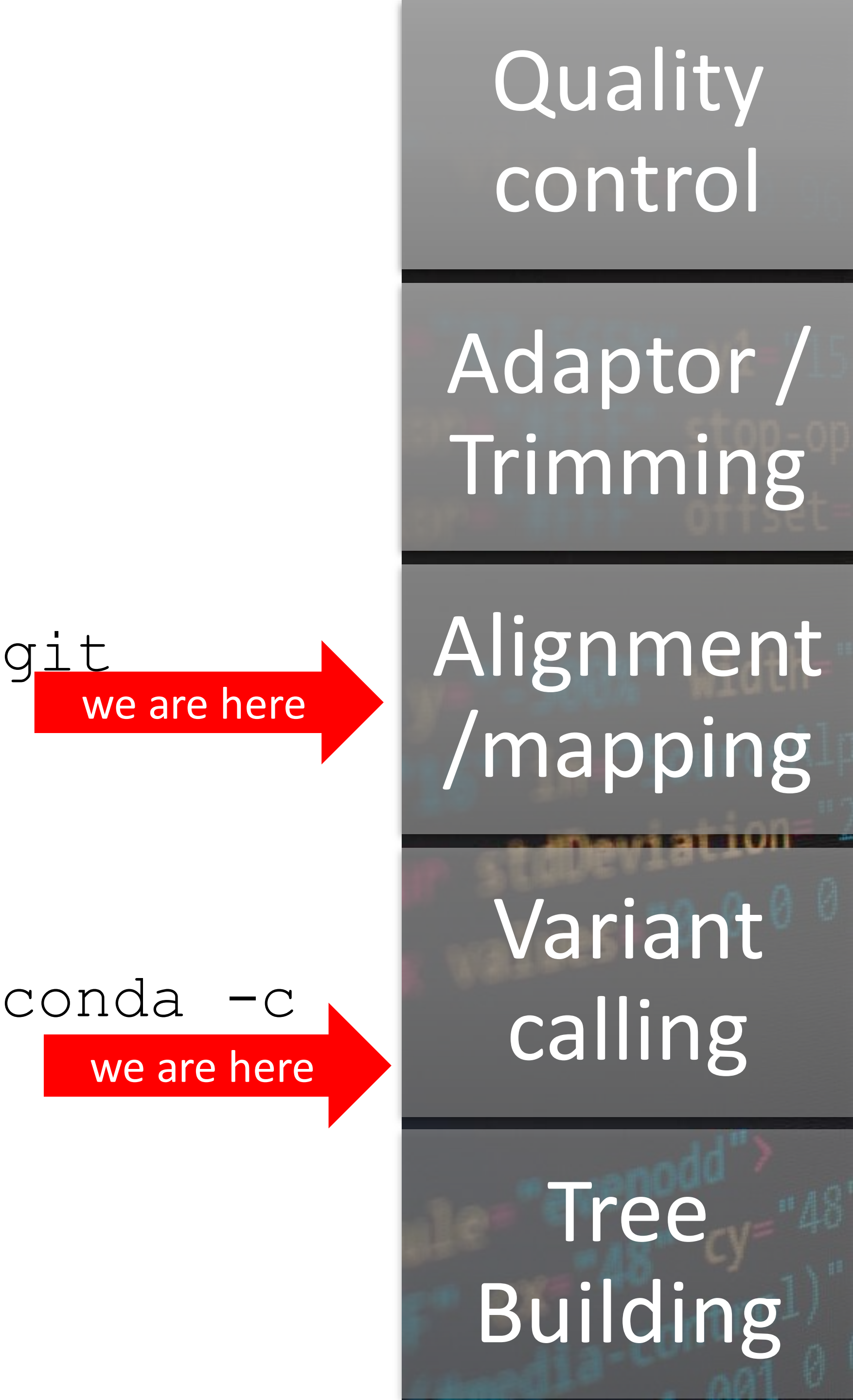
Variant calling

Tree Building

Installation - SNIPPY

- **Installation - CLI:**
  - `cd $HOME`
  - `git clone https://github.com/tseemann/snippy.git`
  - `$HOME/snippy/bin/snippy -help`

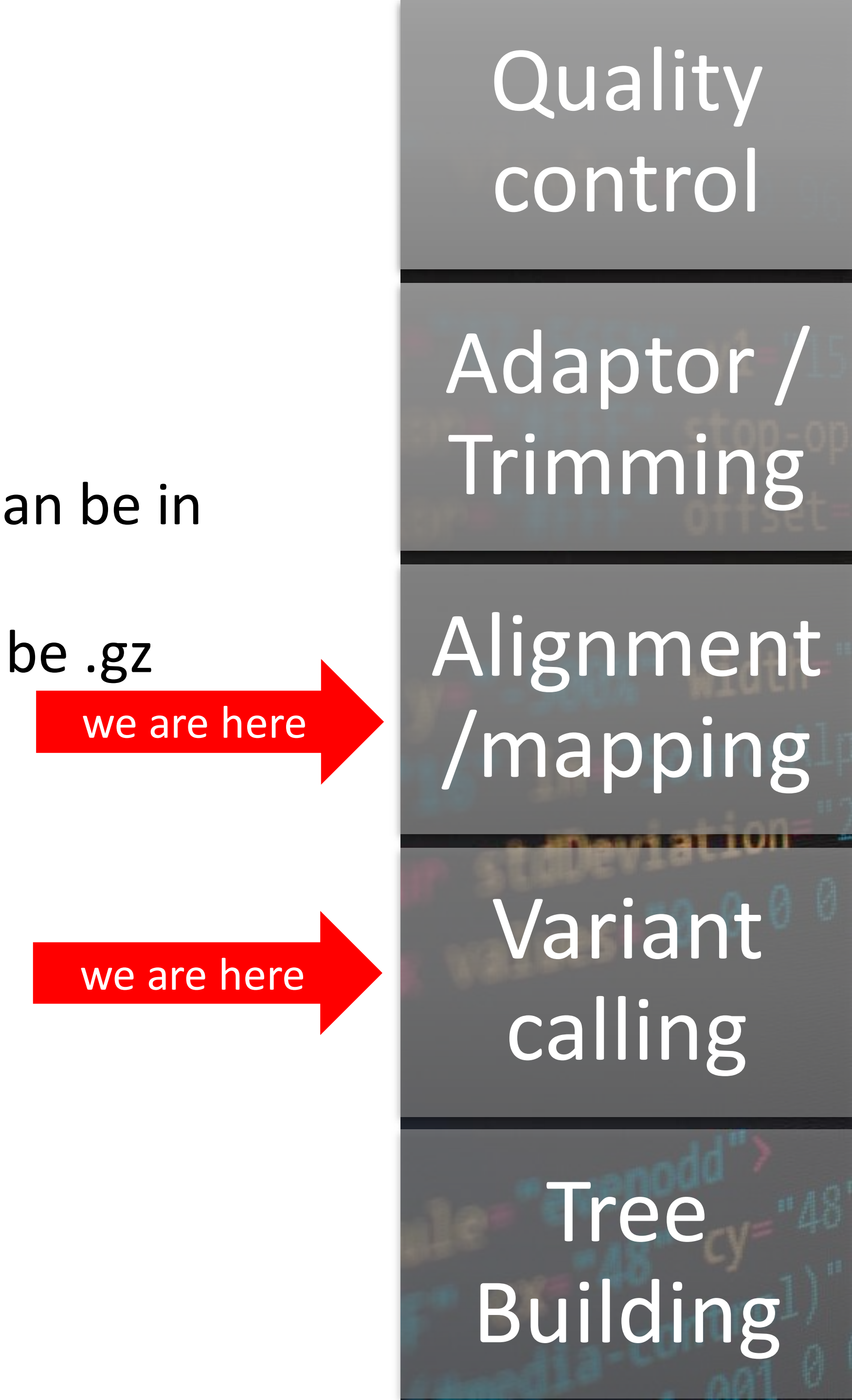
- **Installation – Bioconda:**
  - `conda install -c conda-forge -c bioconda -c defaults snippy`



Usage - SNIPPY

- Usage:
  - a reference genome in FASTA or GENBANK format (can be in multiple contigs)
  - sequence read file(s) in FASTQ or FASTA format (can be .gz compressed) format
  - a folder to put the results in

- https://github.com/tseemann/snippy



Quality control

Adaptor / Trimming

we are here

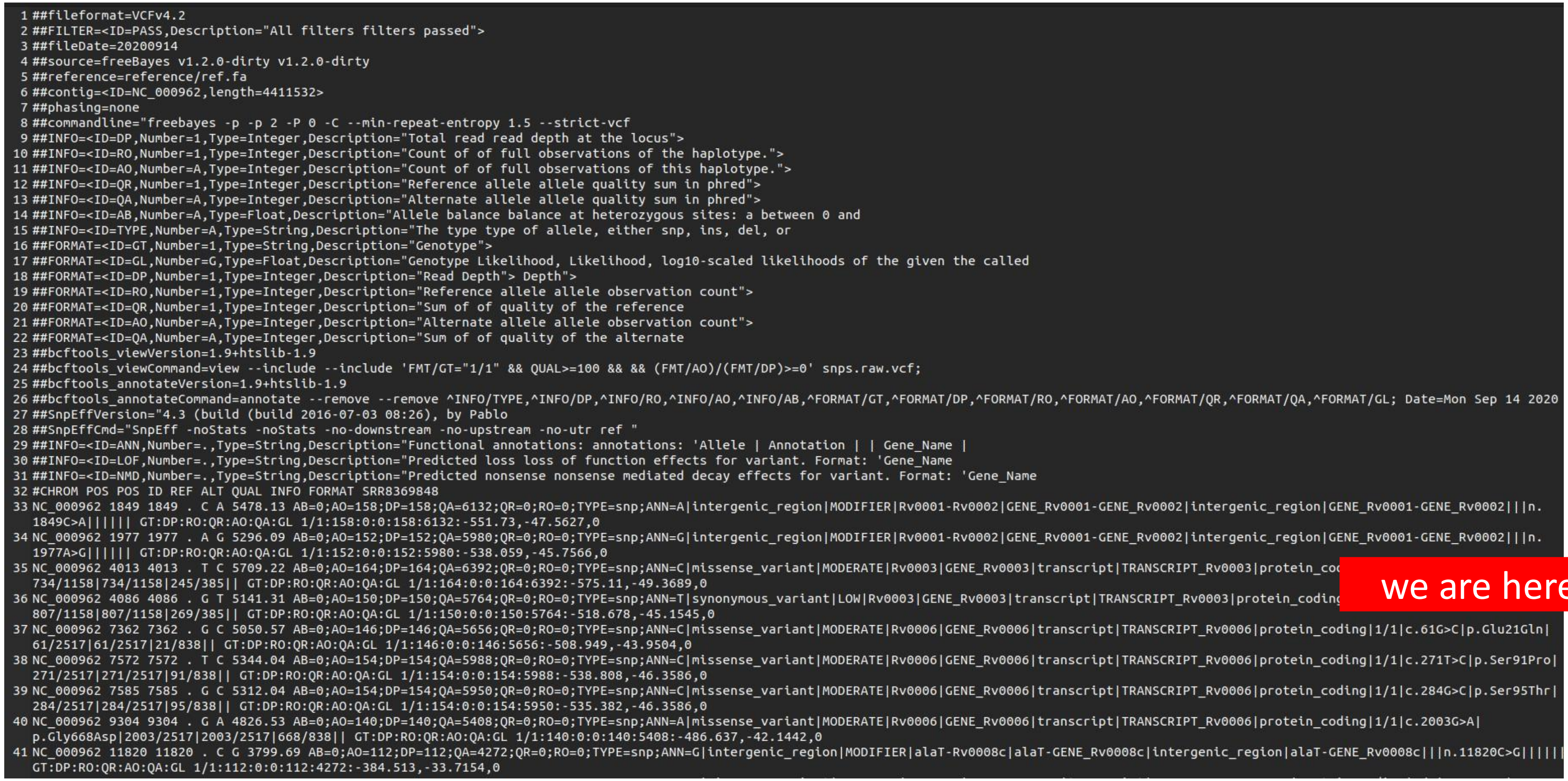
Alignment /mapping

we are here

Variant calling

Tree Building

VCF - Output



```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters filters passed">
##fileDate=20200914
##source=freeBayes v1.2.0-dirty v1.2.0-dirty
##reference=reference/ref.fa
##contig=<ID=NC_000962,length=4411532>
##phasing=none
##commandline="freebayes -p -p 2 -P 0 -C --min-repeat-entropy 1.5 --strict-vcf
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read read depth at the locus">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of of full observations of the haplotype.">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of of full observations of this haplotype.">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele allele quality sum in phred">
##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance balance at heterozygous sites: a between 0 and
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type type of allele, either snp, ins, del, or
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, Likelihood, log10-scaled likelihoods of the given the called
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth"> Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele allele observation count">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of of quality of the reference
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele allele observation count">
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of of quality of the alternate
##bcftools_viewVersion=1.9+htslib-1.9
##bcftools_viewCommand=view --include --include 'FMT/GT="1/1" && QUAL>=100 && && (FMT/AO)/(FMT/DP)>=0' snps.raw.vcf;
##bcftools_annotateVersion=1.9+htslib-1.9
##bcftools_annotateCommand=annotate --remove --remove ^INFO/TYPE,^INFO/DP,^INFO/RO,^INFO/AO,^INFO/AB,^FORMAT/GT,^FORMAT/DP,^FORMAT/RO,^FORMAT/AO,^FORMAT/QR,^FORMAT/QA,^FORMAT/GL; Date=Mon Sep 14 2020
##SnpEffVersion="4.3 (build (build 2016-07-03 08:26), by Pablo
##SnpEffCmd="SnpEff -noStats -noStats -no-downstream -no-upstream -no-utr ref "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: annotations: 'Allele | Annotation | | Gene_Name |
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss loss of function effects for variant. Format: 'Gene_Name
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense nonsense mediated decay effects for variant. Format: 'Gene_Name
#CHROM POS POS ID REF ALT QUAL INFO FORMAT SRR8369848
NC_000962 1849 1849 . C A 5478.13 AB=0;AO=158;DP=158;QA=6132;QR=0;RO=0;TYPE=snp;ANN=A|intergenic_region|MODIFIER|Rv0001-Rv0002|GENE_Rv0001-GENE_Rv0002|intergenic_region|GENE_Rv0001-GENE_Rv0002|||n.
1849C>A|||||| GT:DP:RO:QR:AO:QA:GL 1/1:158:0:0:158:6132:-551.73,-47.5627,0
NC_000962 1977 1977 . A G 5296.09 AB=0;AO=152;DP=152;QA=5980;QR=0;RO=0;TYPE=snp;ANN=G|intergenic_region|MODIFIER|Rv0001-Rv0002|GENE_Rv0001-GENE_Rv0002|intergenic_region|GENE_Rv0001-GENE_Rv0002|||n.
1977A>G|||||| GT:DP:RO:QR:AO:QA:GL 1/1:152:0:0:152:5980:-538.059,-45.7566,0
NC_000962 4013 4013 . T C 5709.22 AB=0;AO=164;DP=164;QA=6392;QR=0;RO=0;TYPE=snp;ANN=C|missense_variant|MODERATE|Rv0003|GENE_Rv0003|transcript|TRANSCRIPT_Rv0003|protein_coc
734/1158|734/1158|245/385|| GT:DP:RO:QR:AO:QA:GL 1/1:164:0:0:164:6392:-575.11,-49.3689,0
NC_000962 4086 4086 . G T 5141.31 AB=0;AO=150;DP=150;QA=5764;QR=0;RO=0;TYPE=snp;ANN=T|synonymous_variant|LOW|Rv0003|GENE_Rv0003|transcript|TRANSCRIPT_Rv0003|protein_coding
807/1158|807/1158|269/385|| GT:DP:RO:QR:AO:QA:GL 1/1:150:0:0:150:5764:-518.678,-45.1545,0
NC_000962 7362 7362 . G C 5050.57 AB=0;AO=146;DP=146;QA=5656;QR=0;RO=0;TYPE=snp;ANN=C|missense_variant|MODERATE|Rv0006|GENE_Rv0006|transcript|TRANSCRIPT_Rv0006|protein_coding|1/1|c.61G>C|p.Glu21Gln|
61/2517|61/2517|21/838|| GT:DP:RO:QR:AO:QA:GL 1/1:146:0:0:146:5656:-508.949,-43.9504,0
NC_000962 7572 7572 . T C 5344.04 AB=0;AO=154;DP=154;QA=5988;QR=0;RO=0;TYPE=snp;ANN=C|missense_variant|MODERATE|Rv0006|GENE_Rv0006|transcript|TRANSCRIPT_Rv0006|protein_coding|1/1|c.271T>C|p.Ser91Pro|
271/2517|271/2517|91/838|| GT:DP:RO:QR:AO:QA:GL 1/1:154:0:0:154:5988:-538.808,-46.3586,0
NC_000962 7585 7585 . G C 5312.04 AB=0;AO=154;DP=154;QA=5950;QR=0;RO=0;TYPE=snp;ANN=C|missense_variant|MODERATE|Rv0006|GENE_Rv0006|transcript|TRANSCRIPT_Rv0006|protein_coding|1/1|c.284G>C|p.Ser95Thr|
284/2517|284/2517|95/838|| GT:DP:RO:QR:AO:QA:GL 1/1:154:0:0:154:5950:-535.382,-46.3586,0
NC_000962 9304 9304 . G A 4826.53 AB=0;AO=140;DP=140;QA=5408;QR=0;RO=0;TYPE=snp;ANN=A|missense_variant|MODERATE|Rv0006|GENE_Rv0006|transcript|TRANSCRIPT_Rv0006|protein_coding|1/1|c.2003G>A|
p.Gly668Asp|2003/2517|2003/2517|668/838|| GT:DP:RO:QR:AO:QA:GL 1/1:140:0:0:140:5408:-486.637,-42.1442,0
NC_000962 11820 11820 . C G 3799.69 AB=0;AO=112;DP=112;QA=4272;QR=0;RO=0;TYPE=snp;ANN=G|intergenic_region|MODIFIER|alaT-Rv0008c|alaT-GENE_Rv0008c|intergenic_region|alaT-GENE_Rv0008c|||n.11820C>G|||||
GT:DP:RO:QR:AO:QA:GL 1/1:112:0:0:112:4272:-384.513,-33.7154,0
```

we are here

Quality control

Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building