NGS Data Types and Formats

Dr Luria Founou

Head of Research, CEDBCAM-RI

Advisory Member, WHO FERG

Honorary Senior Lecturer, UKZN

Learning outcomes

1. Describe what files are generated during the various steps of the analysis pipeline
2. Recognize the structure and information contained in the different file formats
3. Extract specific information from the different files
4. Summarize NGS file formats generated from QC to variant calling

Agenda

Overview of NGS data analysis

Single-end Reads vs Paired-end Reads

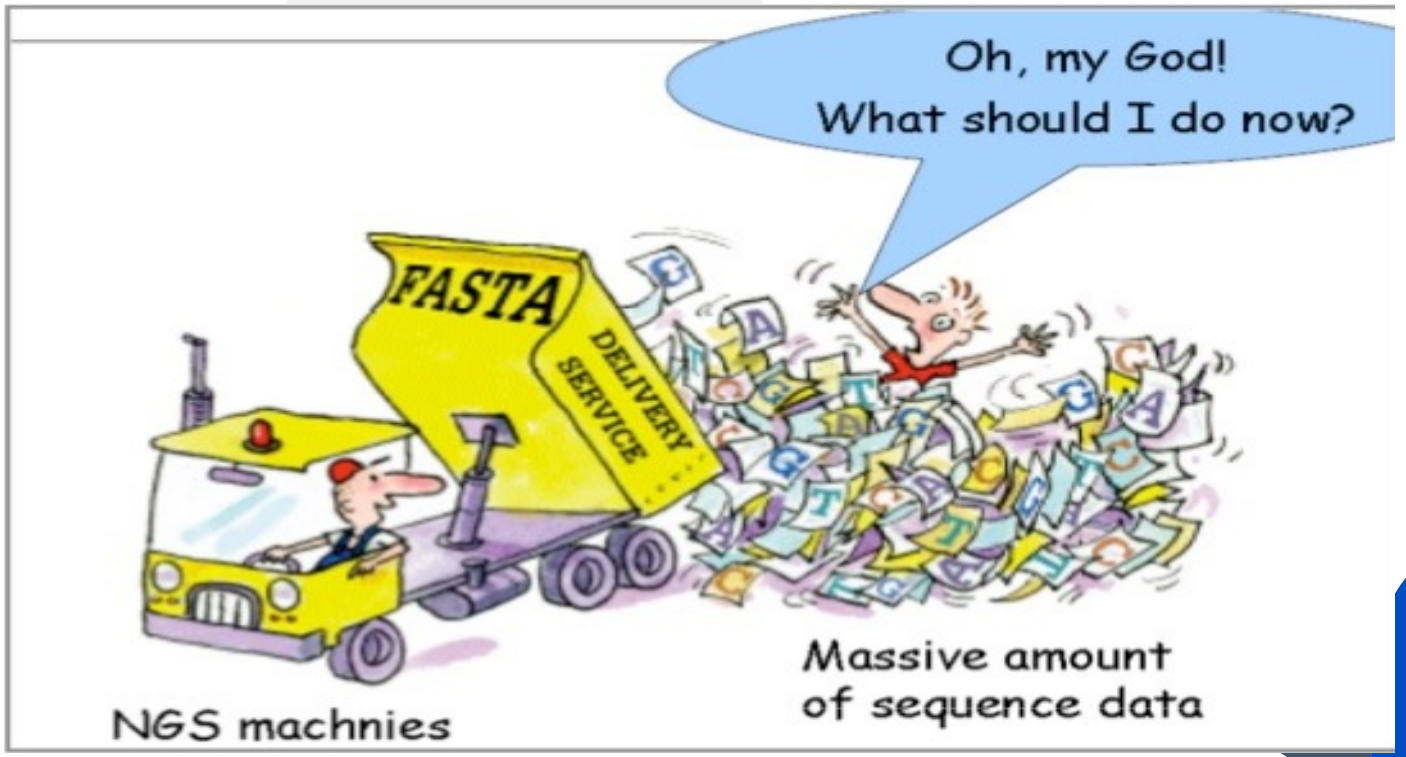
Raw Sequence data

Non-human readable formats

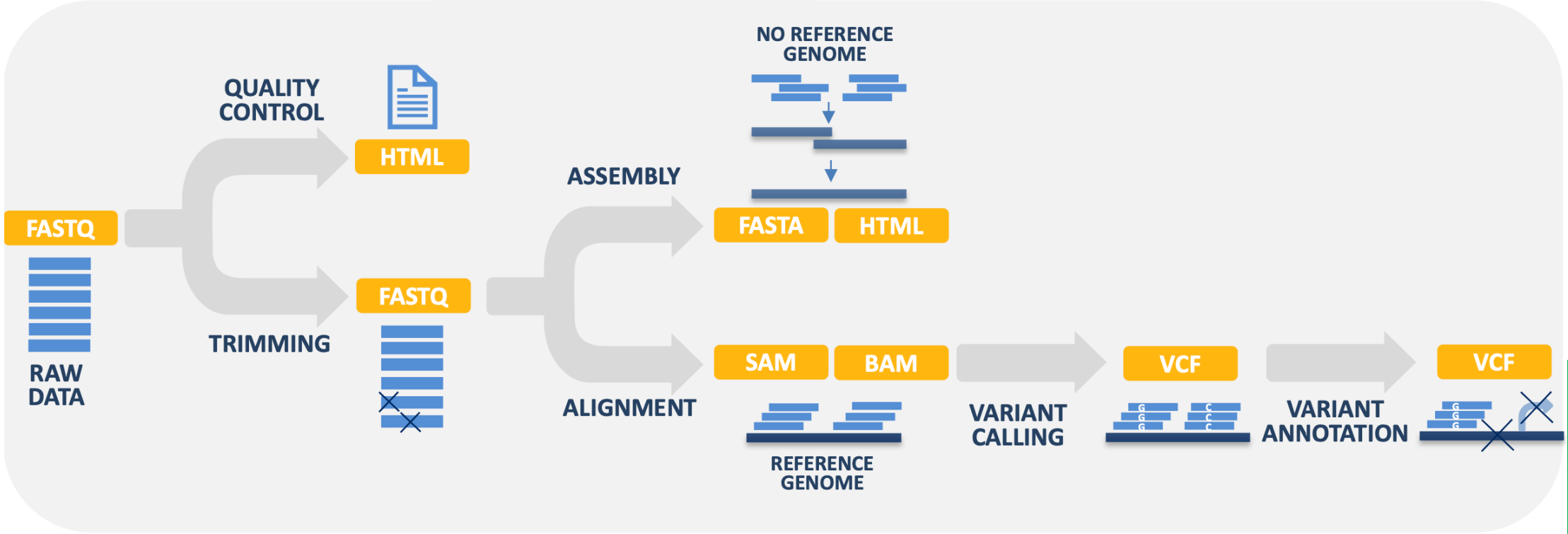
Human readable formats

Take home messages

Overview of NGS Data analysis



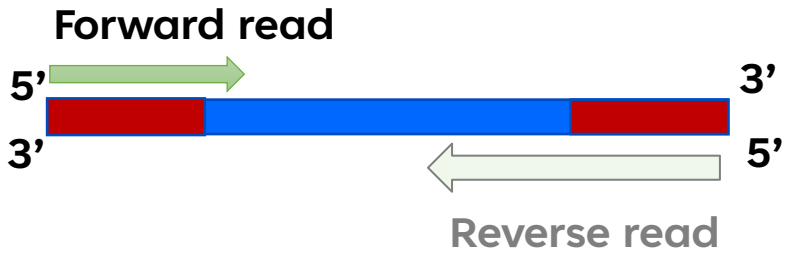
Overview of NGS Data analysis



Single (SR) vs Paired-end (PE) reads

Single-End Reads (SR)

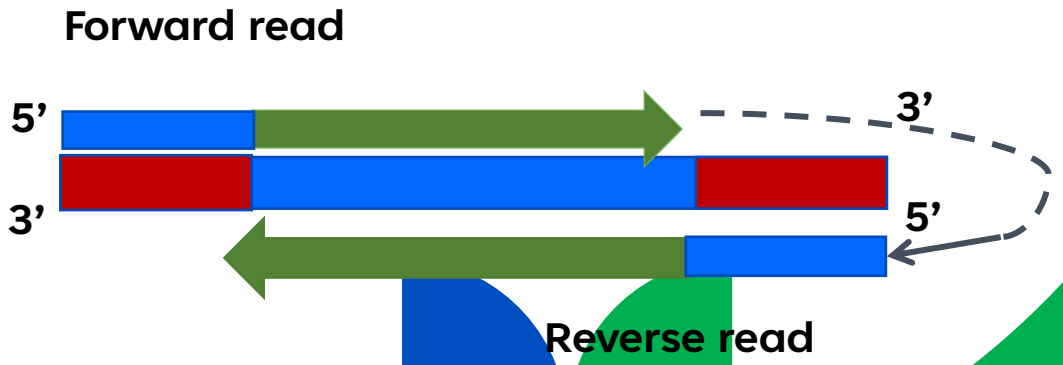
- Sequencing determines the DNA sequence of just one end of each DNA fragment.
- **Random**



Single-End Reads

Paired-end Reads

- Sequencing yields both ends of each DNA fragment.
  - More expensive
  - Increase mappability for repetitive regions
  - Easier identification of SNPs and Indels
  - Increase precision

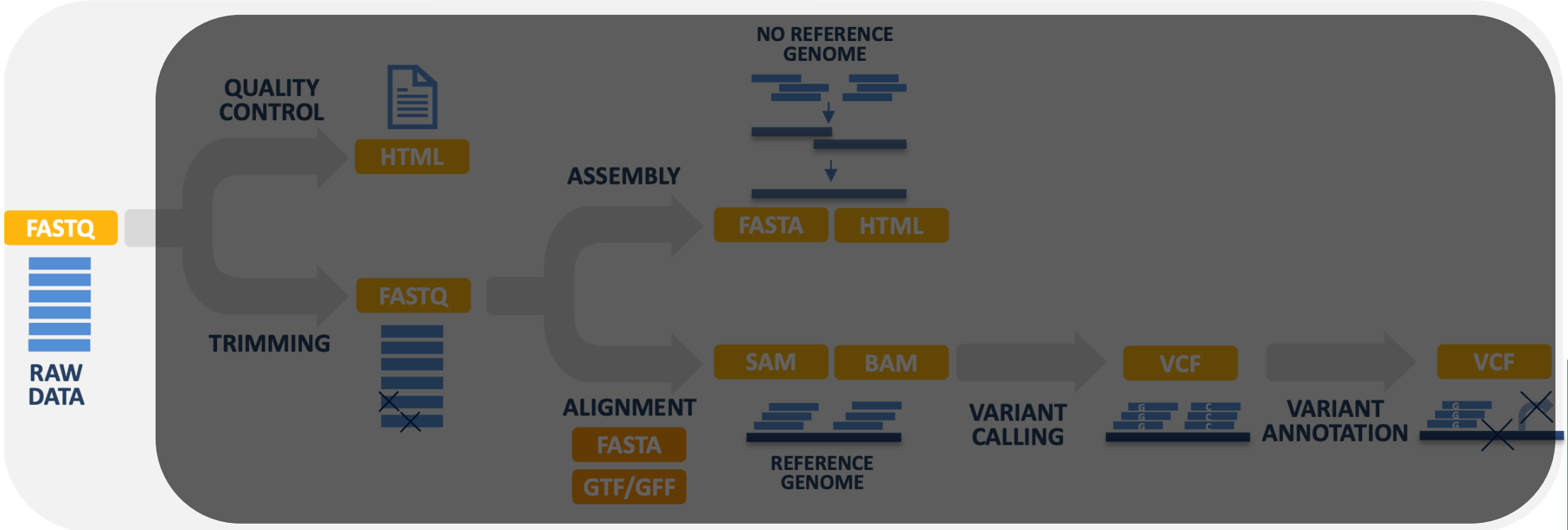


Paired-end Reads

File format for defining genomic regions

- Various file formats exist to store information about the location of transcription start sites, exons, introns etc.
- All formats agree on having
  - Header
  - Features
  - Sequence
- Nature of the information contained in each row can vary strongly between the formats

Overview of NGS Data analysis



Raw Sequence Data

Raw Sequence Data

FASTQ format

- Text-based file derived from FASTA format
- Original Sanger standard from capillary sequence data
- Sequence description, sequence and associated per base quality score

Example of a fastq file

@title	@SRR010930.8436795/1
sequence	ACCCCAGGATCAACACTTCACATGCATTAGCAGAGAGAGATAAATCAA
+optional_text	+
quality	=>=??A?<@B@A:?B?D;AC@@CAAAD<AAA:99?:@=?@B@77C><4

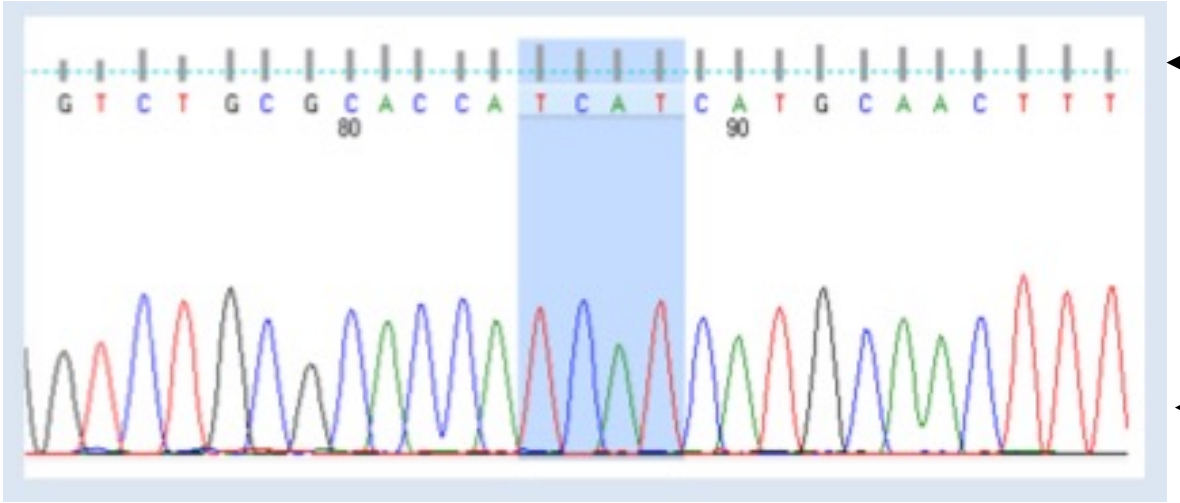
Line1: begin with @, Read sequence identifier (encoded descriptions of instrument, lane...)
Line2: raw sequence letters
Line3: « + » sign (optional: « + » followed by seq identifier)
Line4: quality values for the sequence in line 2

Raw Sequence Data

- PHRED quality scores encoded as ASCII printable characters (ASCII 33-126)
- Encodes the probability of a call being an error

- **Other raw sequence data**
  - fastq.gz files
  - fastq.ora files
  - bcl2fastq Conversion: NextSeq, HiSeq, NovaSeq 6000
  - Basecall file format: PacBio & ONT

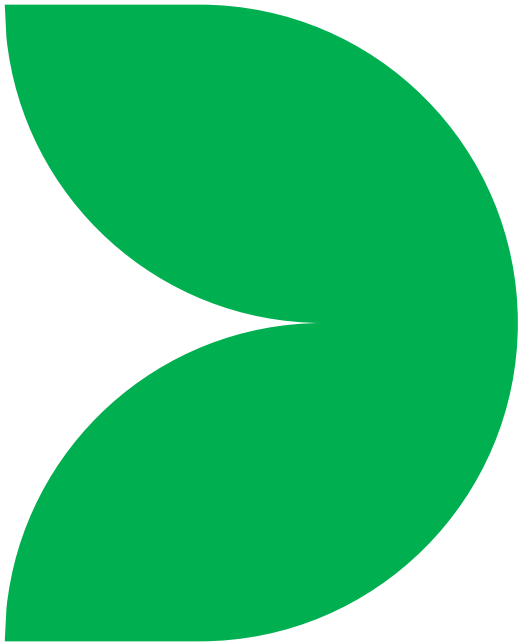
Reminder: SANGER SEQUENCING



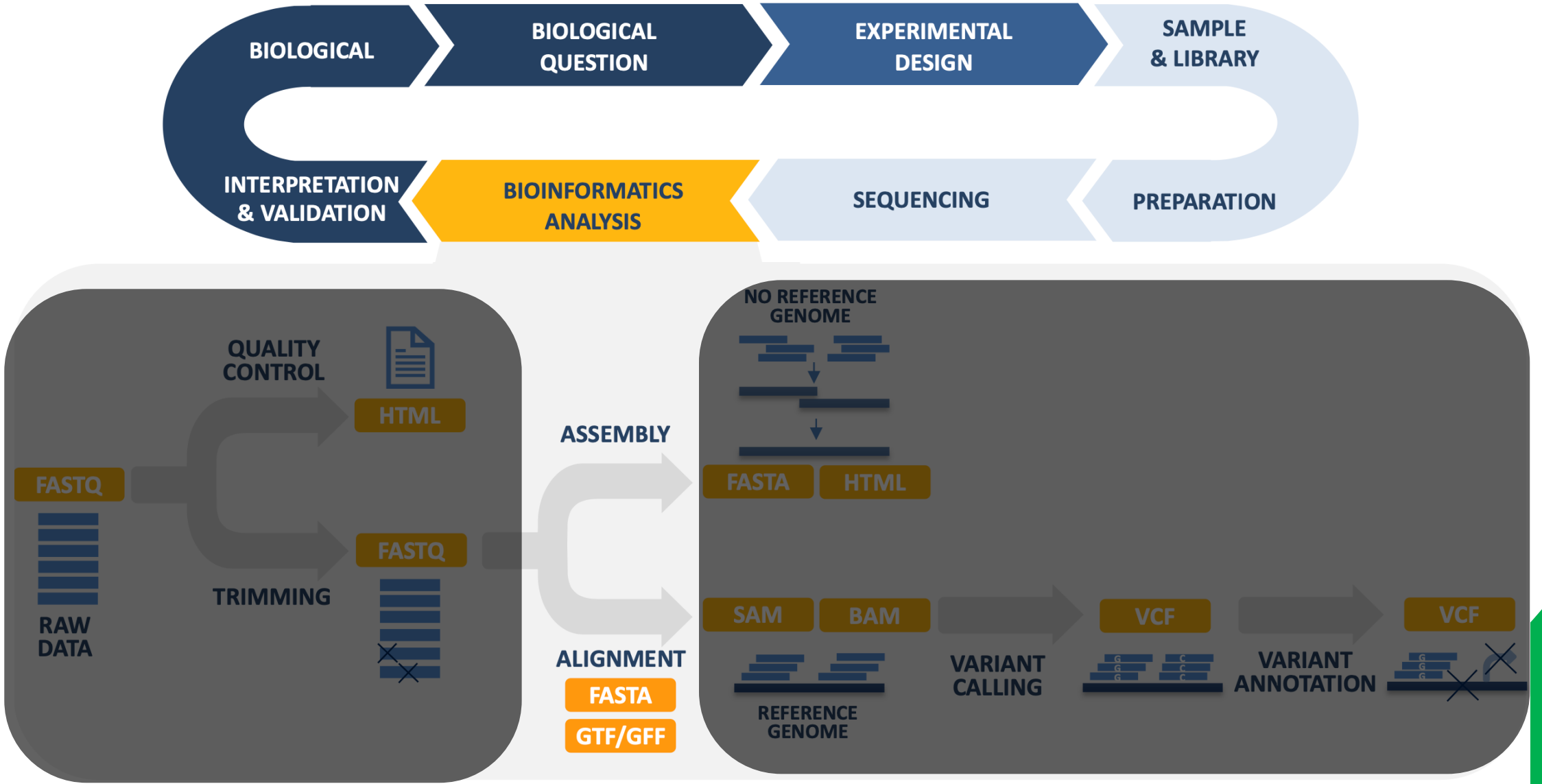
FATSQ file is a file containing :

- **Reads sequences**
- **a Quality score associated to each Read**

Human Readable formats



Overview of NGS Data analysis



FASTA

- Text-based format for DNA, RNA or Amino Acid data
- Structure
  - Description line or header always start with “>”
  - Lines of sequence data typically 50 to 8
  - May be upper or lower case.
- FASTA Extensions
  - .fa or .fasta: generic Fasta
  - .fna:FASTA nucleic acid
  - .ffn: FASTA nucleotide of gene regions
  - .faa: FASTA amino acid
  - .frn: FASTA non-coding RNA

```
>contig00001 len=304256 cov=17.6 corr=0 origname=Contig_206_17.5601 sw=shovill-skesa/1.0.4
date=20221108
TTTTCAACTTCTGATGGATGCGAGTGATTGAACTCATACATTAATGTTTTCCCACGAAGT
CTTTTTTCAGGTAAGCCTTCGCACATATCGGTAAATAGATTGCCTGCTTTTATTTTTTCT
GTCATCGACATGTTCATTTTTAACATTCCGTCCTGATAAGTTGGTCGGATAAGGCGCTCG
CGCCGTATCCGACATTAATTTCTTAAGCGACTTCATTCACCTGGCGACGCAGCAGGGAAA
GTGGGCCGGGGCCGCTAAGCGTGAACACGGAAATTAAGGTGAAGCCCAGCGCCACCAGAC
CCAGCACCAGGTAAGCGCCCTGGAAACCGATGCTTTCATACATATTGCCCGCCAGAATAG
ACATAAAAACCATCGCCAGTTGCTTAAAGAAGCAGAAACAGACCAGATAAATCGTCGCTG
AAAAACGCACTTCAAACTGGCTGGTAATATATTTAAAGCAGCCCACCAGCAGGAACGGTA
CTTCAAACATATGCAGCGTTTTCAGAATAACCACTTCCAGTGCTGAGGTAGCGAACGATG
AGCCAATAATACGTACTGACATAATAGTGCCAGCCAGCAGCAGGGCGTTTTTCCCACCGA
TGCGATTAATGATCAGTGGCGCAAAAAACATAATCGAGGCGTTAAGTAATTCGCCCATTG
TCGTTACGTAGCCAAATACCCGCGTACCCTGTTCACCGGTAGCAAAGAACGAAGTAAAGA
AATTAGCAAACTGTTGGTCAAAAACATCGTAGGTGCAGGAAACGCCAATAACATACAGTG
ACAAAAACCACAGTTTTGGCTGTCTGAACAGTTCCAGCGCCAGCTTAAGGCTAAATGCCG
AATGGTTGGCACCTACCGCATTGGCAACCGTGGCAGAAGAGGGCGCATCCGTTTTGGCGA
AAAAGAGTAAAACGGCGAGGATGAGTGCACAGCCAGAACCCAGCCAGAAAACAAACTGAT
TATTGATGGTGAACATGATGCCGACAATCGAGGCACACAGCGCCCAGCCAACACAGCCAA
ACATCCGCGCGCGACCAAATTCGAAATTACTGCGACGGCTGACTTTCTCGATAAATGCCT
CTACTGCTGGCGCACCGGCGTTAAAACAAAAGCCTAGATAAATACCACCAACAATCGATC
CTACTAAAATGTTGTATTGTAACAGTGGCCCGAAGATAAAAATAAAGAACGGCGCAAACA
TCACTAACATGCCGGTAATAATCCACAGCAGGTATTTGCGCAGCCCGAGTTTGTCAGAAA
GCAGACCAAACAGCGGTTGGAATAATAGCGAGAACAGAGAAATAGCGGCAAAAATAATAC
CCGTATCACTTTTGCTGATATGGTTGATGTCATGTAGCCAAATCGGGAAAAACGGGAAGT
AGGCTCCCATGATAAAAAAGTAAAAGAAAAAGAATAAACCGAACATCCAAAAGTTTGTGT
TTTTTAAATAGTACATAATGGATTTCCTTACGCAAAATACGGGCAGACATGGCCTGCCCG
GTTATTATTATTTTTGACACCAGAGCAACTGGTAATGGTAGCTACCGGCGCTAAGCTGGA
ATTCCGCCGACACTGACGGGCTCCAGGAGTCGTCGCCACCAATCCCCATATGGAAACCGT
CGATATTCAGCCATGTGCCTTCTTCCGCGTGCAGCAGATGGCGATGGCTGGTTTCCATCA
```

Annotation file output - GFF

- **General Feature Format: 9 required fields**
- The first three fields form the basic name, start, end tuple that allows for the identification of the location in respect to the reference genome
- Fields must be separated by a single **TAB (/)**, but no white space.
- All but the final field in each feature line must contain a value; missing values should be denoted with a '.'
- GFF2 and GFF3 are similar but not compatible

```
# GFF-version 2
IV      curated exon    5506900 5506996 . + .   Transcript B0273.1
IV      curated exon    5506026 5506382 . + .   Transcript B0273.1
IV      curated exon    5506558 5506660 . + .   Transcript B0273.1

# GFF-version 3
ctg123  .  exon  1300  1500  .  +  .  ID=exon00001
ctg123  .  exon  1050  1500  .  +  .  ID=exon00002
ctg123  .  exon  3000  3902  .  +  .  ID=exon00003
```

Annotation file output - GTF

- **Gene Transfer Format (GTF)** based on the GFF and sometimes referred to as GFF2.5
- Contrary to GFF files, the **TYPE VALUE** pairs of GTF files are **separated by one space** and must end with a **semi-colon**
- **GFF and GTF files can contain various types of annotations**

```
# GTF example
chr1  HAVANA  gene  11869 14412 . + . gene_id "ENSG00000223972.4";
    transcript_id "ENSG00000223972.4"; gene_type "pseudogene"; gene_status "
    KNOWN"; gene_name "DDX11L1"; transcript_type "pseudogene"; transcript_status
     "KNOWN"; transcript_name "DDX11L1"; level 2; havana_gene "
    OTTHUMG00000000961.2";
chr1  HAVANA  transcript  11869 14409 . + . gene_id "ENSG00000223972.4";
    transcript_id "ENST00000456328.2"; gene_type "pseudogene"; gene_status "
    KNOWN"; gene_name "DDX11L1"; transcript_type "processed_transcript";
    transcript_status "KNOWN"; transcript_name "DDX11L1-002"; level 2; tag "
    basic"; havana_gene "OTTHUMG00000000961.2"; havana_transcript "
    OTTHUMT00000362751.1";
```

Annotation file output – Genbank

- GenBank or .gbk allows the storage of annotation information for sequence in addition to a DNA/protein sequence itself
- Consists of an annotation section and a sequence section.
- The **start of the annotation** section is marked by a line beginning with the word **"LOCUS"**
- The **start of the sequence** section is marked by a line beginning with the word **"ORIGIN"** and the end of the section is marked by a line with only "//".
- Check example https://www.ncbi.nlm.nih.gov/nuccore/MH814718.1

Genome Browsers

- **BED(Browser Extensible Data) file:**
  - A tab-delimited text file describes genomic intervals (reads, peaks, genes, etc)
  - 0-based
  - Essential columns : chromosome – start – stop
  - Specific descriptions of file formats: https://genome.ucsc.edu/FAQ/FAQformat.html

bed format file of aligned sequencing reads :

Read identifier

Alignment score

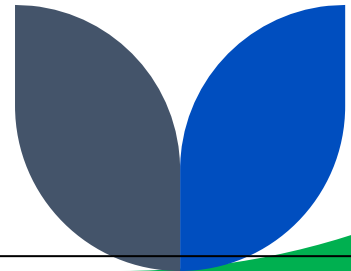
Chromosome	start	End	Name	Score	Strand
chr1	92	153	NS500322:23:H0UM0AGXX:1:22305:20603:1636		
	0	+			
chr1	205	264	NS500322:23:H0UM0AGXX:3:22506:4037:7916		
	23	+			

bed format file of TF binding sites (peaks) :

Peak identifier

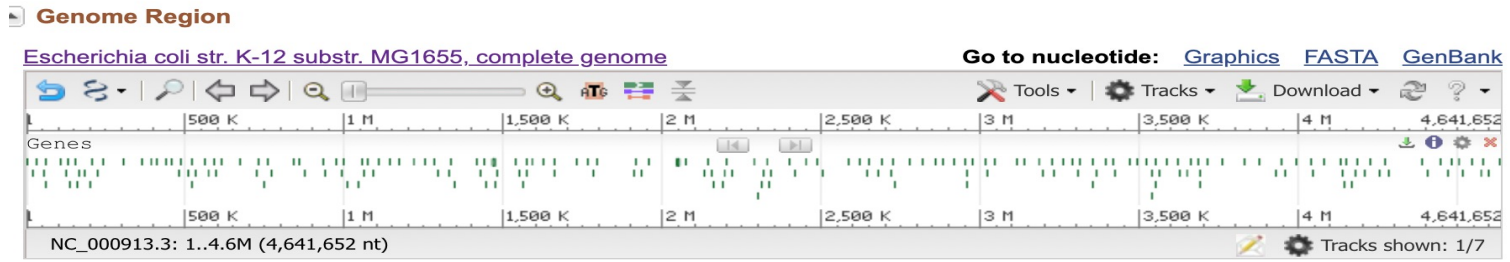
Peak significance score

Chromosome	start	End	Name	Score
Chr1	3006970	3007100	chip_TF1_1	120
chr1	3015000	3015120	chip_TF1_2	100
Chr1	4007004	4007145	chip_TF1_3	5



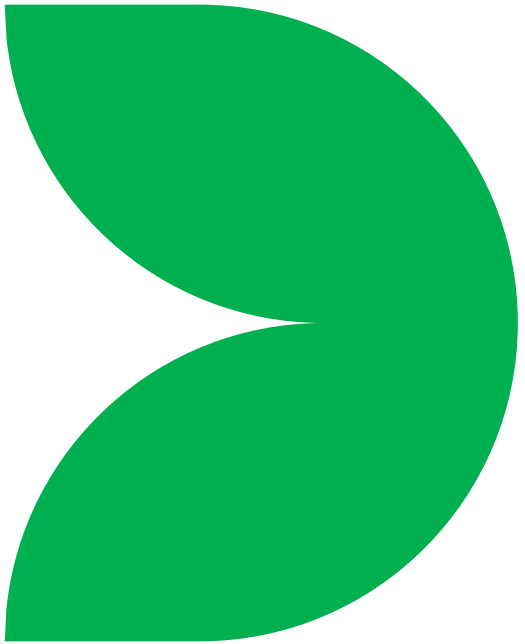
Genome Browsers

- **BedGraph/Wiggle** file format:
  - Used to represent quantitative information across genomic regions.
  - e.g. read depth from ChIP-seq experiments.
    - chr1 3000095 3000131 1
    - chr1 3003970 3004006 2
    - chr1 3006970 3007004 2
  - **Visualization using genome browser (https://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521) other browser https://genome.ucsc.edu/**

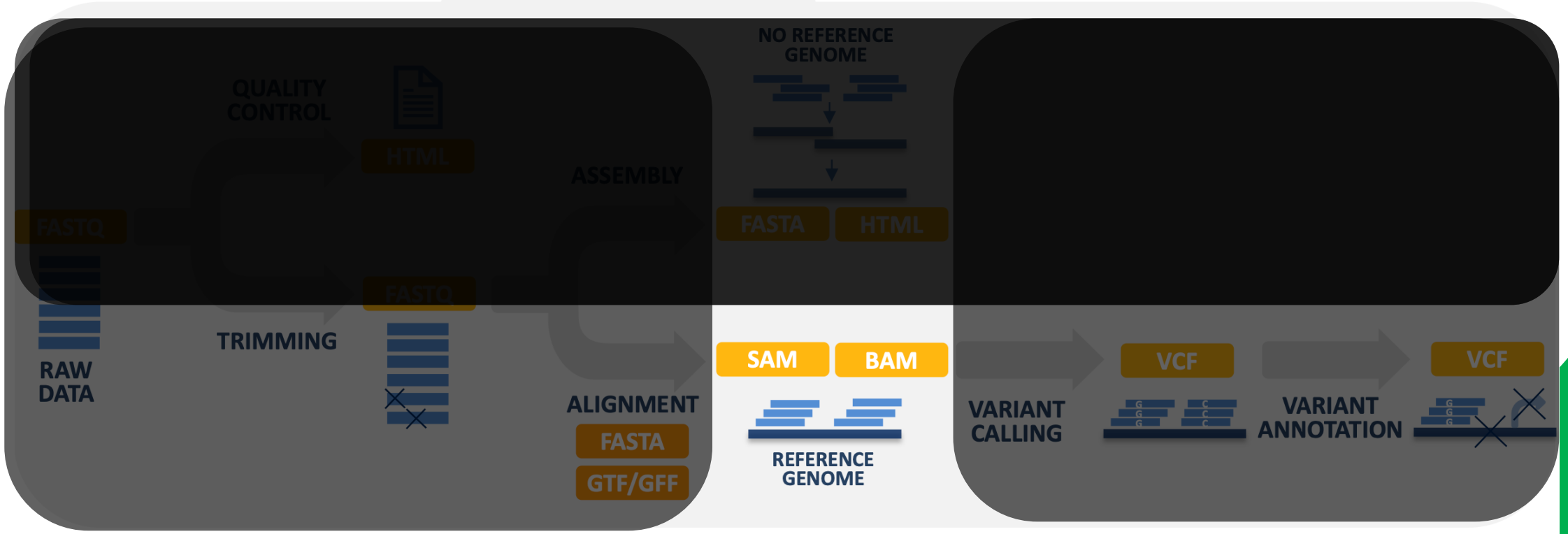


- **bigWig** file format (most widely used for data visualization)
  - indexed binary format
  - only the portions of the files needed to display a particular region are transferred

Not Human Readable formats



Overview of NGS Data analysis



Sequence alignment

- Basic aim of genomics studies :
  - What are the elements in the sample?
  - How abundant are the elements in the sample?
- Purpose: To relate sequencing reads to existing knowledge
  - Existing knowledge: reference genome; gene structures etc.
- A common technique for mapping reads to features is **alignment**
- Multiple alignmers

Alignment output - SAM

- Input QC'd FASTQ (Tool BWA)
- Output of read alignment and mapping → **S**equence **A**lignment **M**ap **(SAM)** file
- Standardised method for storing all information relevant to how reads aligns to a reference genome
- **SAM files are rather big when dealing with large NGS datasets**

Alignment output – SAM

- **SAM** (Sequence Alignment/Map format) file:
  - a tab-delimited text file that contains aligned sequence data information (human readable)
  - **11 compulsory fields:** mapping position, mapping quality, segment sequence...
  - The optional header section followed by the alignment section where each line corresponds to one sequenced read.
  - Detailed description of SAM file format:
  - http://samtools.sourceforge.net/SAM1.pdf
  - https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

@HD VN:

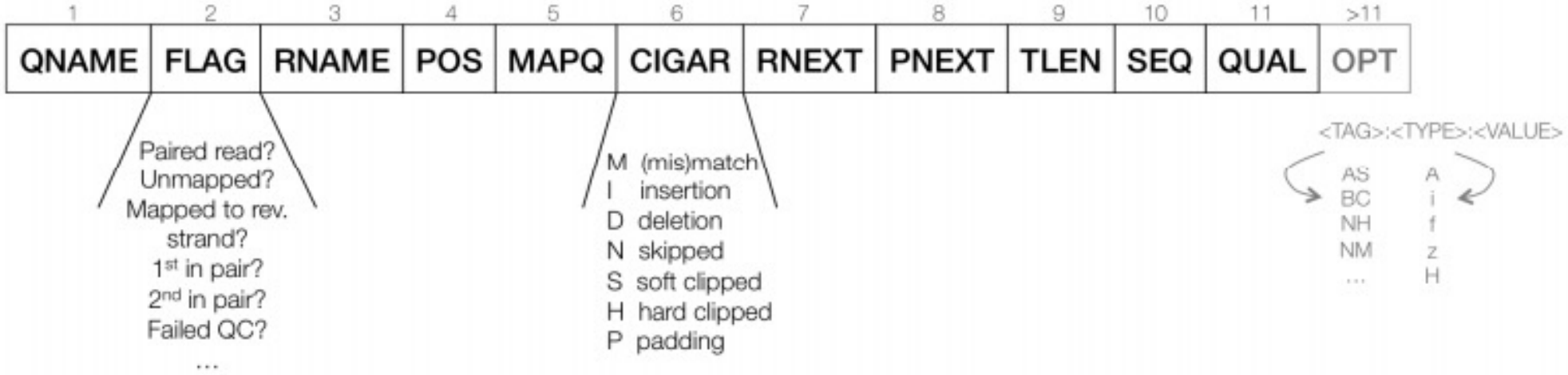
@SQ SN: LN:

@RG ID: SM:

@PG ID:

@CO

(theoretically) optional
HEADER SECTION
general information about the file



ALIGNMENT
SECTION
1 line per locus

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

SAM Structure

Alignment output - SAM

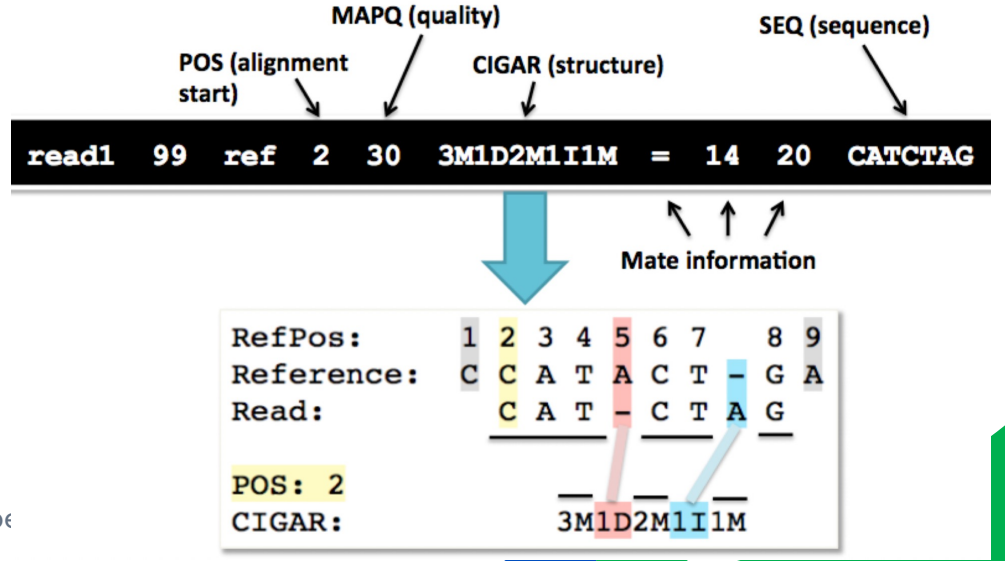
NS500322:23:H0UM0AGXX:1:22305:20603:1636	0	chr1	93	0	61M	*	0	0	
CCCTGTAGTTAAAATTGACTAAGTATTGGAAGGGGCCTATAGACCTTGAGTATTCTCAAGG									
<AAAAFAFFF7FFFFFFFFF.FFFAFFFFFFFFFFFFFFF.F.F)FFFFFFFF<FAFFFFF				XT:A:R	NM:i:0	X0:i:2	X1:i:0		
XM:i:0	XO:i:0	XG:i:0	MD:Z:61	XA:Z:chr7,-92852201,61M,0;					
NS500322:23:H0UM0AGXX:1:13301:15368:13300	0	chr1	265	37	58M	*	0	0	
AGTTATTTATTGGCCCTTCAATTTTCATTTTTATAACCTACTATTACCTTGCAAAAAA									
7AAAAFFFFFFFFFFFFFFFFFFFFFFFFFFFF<<FFFFFFFFFFFFFFFFFFFFF				XT:A:U	NM:i:0	X0:i:1	X1:i:0		
XM:i:0	XO:i:0	XG:i:0	MD:Z:58						
ERR458493	.552967	16	chrI	140	255	12		**M61232N37M2S**	*	0	0
CCACTCGTTCACCAGGGCCGGCGGGCTGATCACTTTATCGTGCATCTTGGC									
BB?HHJJIGHHJIGIIJJIJGIJIJJIIIGHBJJJJJJHHHHFFDDDA1+B				NH:i:1	HI:i:1	AS:i:41	nM:i:2		

CIGAR string

compact representation of sequence alignment:

M	alignment match or mismatch
=	sequence match
X	sequence mismatch
I	insertion to the reference
D	deletion from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
N	skipped region from the reference
P	padding (silent deletion from padded reference)

```
CGTACGTACTGT        Ref:   ACGT----ACGTA        Ref:   CTCAGTG
CGT----ACTGA        Read:  ACGTACGTACGTA        Read:  CGCA-TG
M 4D 5M             Cigar: 4M 4I 5M             Cigar: 4M 1D 2
```



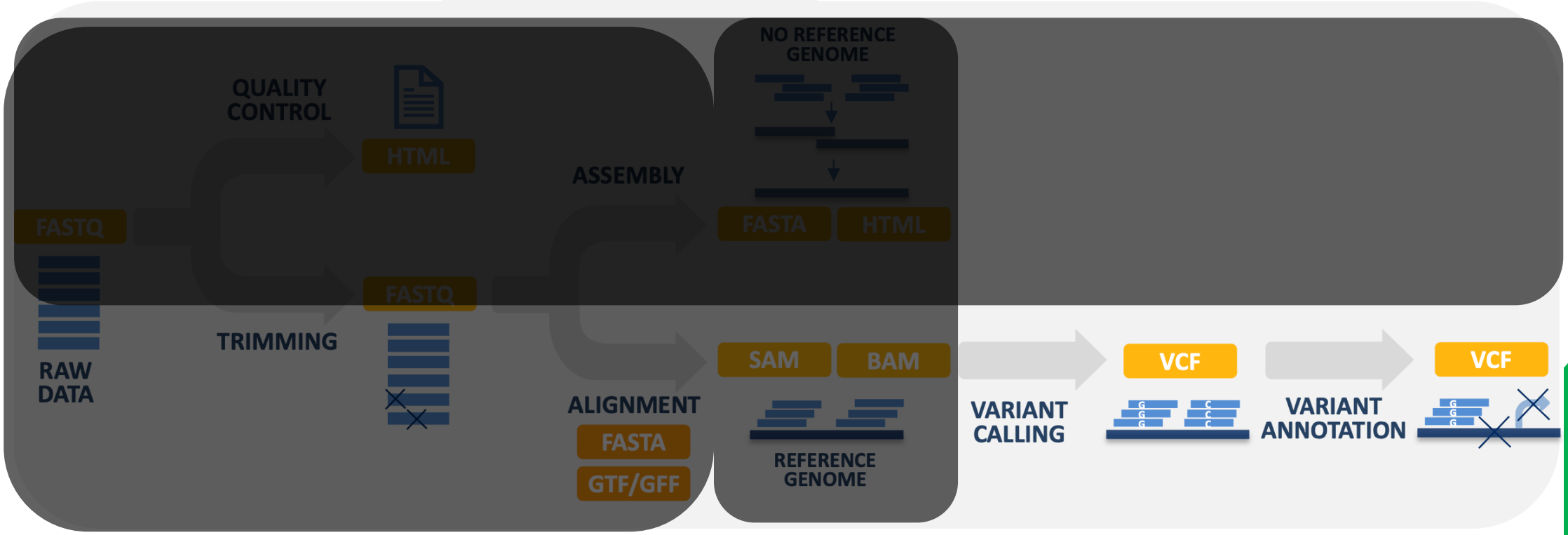
Alignment output file – BAM

- **BAM** file:
  - Compressed binary version of SAM file (**NOT human readable**)
  - → size reduction of around 5-10 fold, and BAM files can be processed much more quickly by NGS tools.
  - Can be visualized after converting to SAM file
  - Tools to convert SAM to BAM and process BAMs
    - Samtools
    - Picard
    - htslib
    - Galaxy
  - .bai file: index file for the corresponding bam file

Alignment output - CRAM

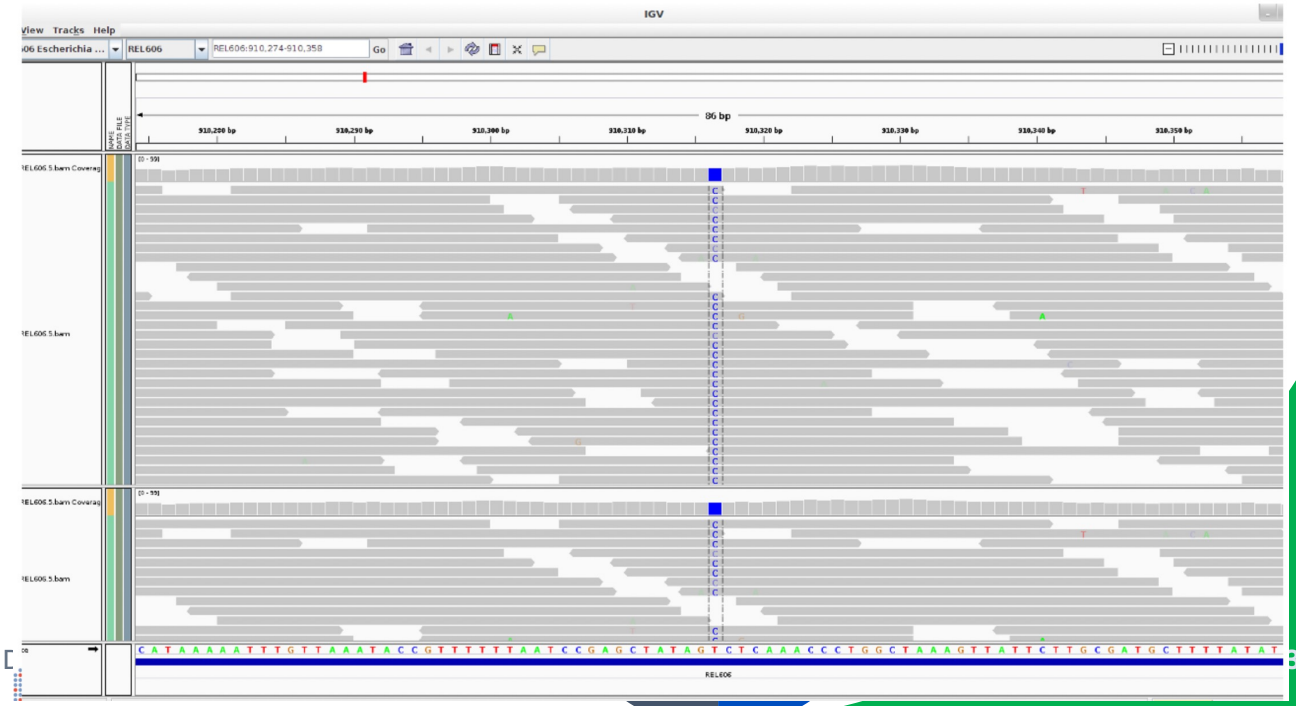
- BAM files are still relatively large ~1.5 – 2 bytes per base pair
- Computer disk capacity is falling behind storage requirements for sequencing data
- CRAM is a reference-based compression technique
- Some quality information lost but in a controlled manner
- Used in most production pipelines now and results in up to 40% reduction in disk space usage

Overview of NGS Data analysis



Variant calling

- Is there a variant (SNP, indel or structural variant) at a particular position?
- Based on per-base evidence provided for all the reads that have mapped to a particular position in the sequence
- Useful to aggregate the evidence from all reads that relate to a particular base in the sequence



Variant calling output (VCF)

- Input for variant callers is a QC'd BAM file
- Variant callers → **Variant Calling Format (VCF)**
- Standardised format for representing variant calls
- Format for SNPs, indels, structural variants and CNVs

Variant calling output (VCF)

- General structure
  - Header rows (start **##**), followed by tab-separated columns with the actual data.
  - The first 8 columns are mandatory
  - The **FORMAT** column defines the format of subsequent columns
  - *Sample level* information follows the **FORMAT** column
  - The header rows =***Meta-information* rows**, describe the coding used in each field via a series of tags.
  - More description https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF structure

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Take home messages

Human readable formats

- **.fasta:** text-based format containing DNA, RNA or protein sequence
- **.gbk**: file containing both sequence and annotation information
- **.bed** : tab delimitated files representing genomic regions

Not human readable

- **.fastq** : raw sequencing reads (sequences and sequencing qualities)
- **.sam** : aligned reads
- **.bam** : binary and compressed version of sam files
- **.cram**: binary and compressed version of bam files

Genome Browser formats

- **Bedgraph/wiggle** : read density across genomic positions
- **Bigwig**: binary version of wiggle files

All files are easily convertible

Thank you

Dr Luria Leslie Founou

luriafounou@gmail.com

www.cedbcam.com

Creative Commons