wellcome connecting science

ACORN

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

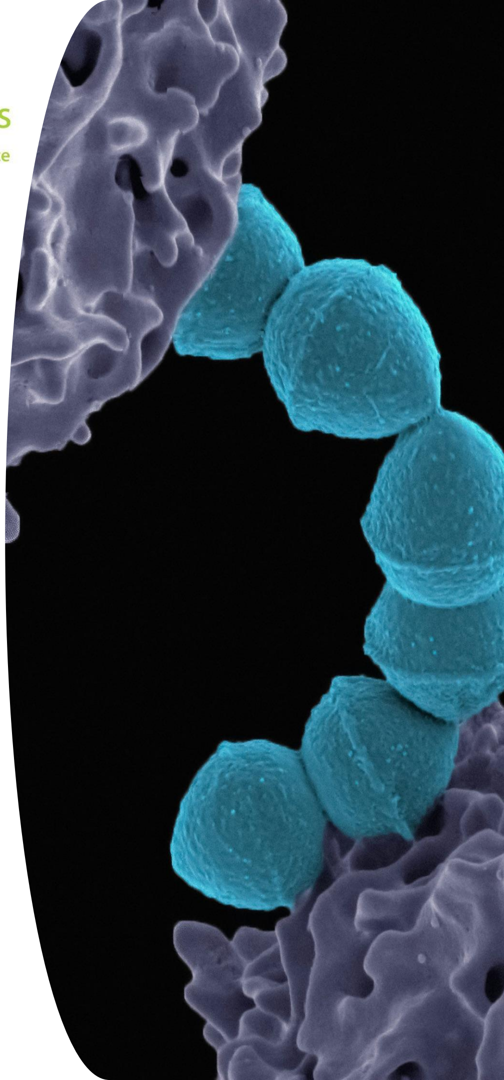
Bash, File formats and Quality Control

Lecture: Introduction to Linux systems, command line and file formats

Instructors:
Rito Mikhari
Collins Kigen

AMR Virtual course (Africa and Asia)

21 May 2024



Learning outcome

- Learn how to use the Unix command
- Understand the different file formats

Overview

- Background: Operating systems, Unix
- Basic Unix commands
- Introduction to colab notebooks (demo)
- Introduction to the basic file formats and their functions in genomics

Background: Operating systems

- An operating system (OS) is the software that functions as an interface between the computer user and its hardware
- The software enables and controls the set of programs that we install in our machines
- Some functions include: allocation of storage, processor management, security, error detection, file management
- Examples of most used OS include: Windows, Linux, Unix, MAC OS, Android, etc.

Background: Unix

- Unix was originally called AT&T Unix, and was first developed in 1969 at the Bell Lab research center
- A family of OS's that allow for multiple users and allow for multitasking
- Beneficial for working with large text files (or genomic data files) and accommodates several powerful yet flexible commands
- Ability to combine these multiple commands for different objectives

Basic Unix: Basic terms

- Shell

  Command-line interpreter

- Terminal

  Tool used for shell commands

- Command prompt

  (base) ubuntu@student-1:~$

- Directories

  - Equivalent to folders with files

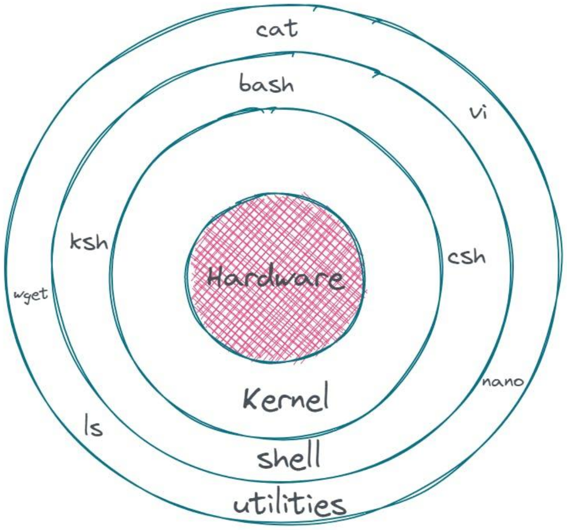
- Root

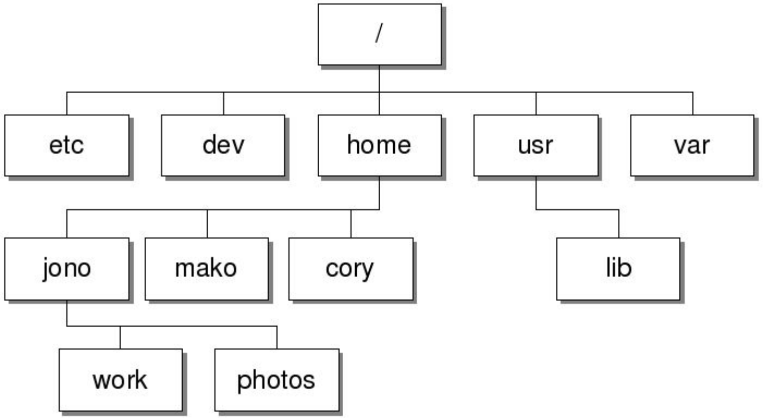
  - The mother directory where all directories stem from

- Syntax

  - Command [options] [arguments]

  - Different programs will have their own syntax with a number of options





/

etc dev home usr var

jono mako cory lib

work photos

Basic Unix: Commands

- Navigation through files and directories
  (*cd, ls, pwd*)

- File and directory management
  (mkdir, touch, nano, cat, less, rm, mv, cp, head, tail)

- Pipes
  (we, sort)

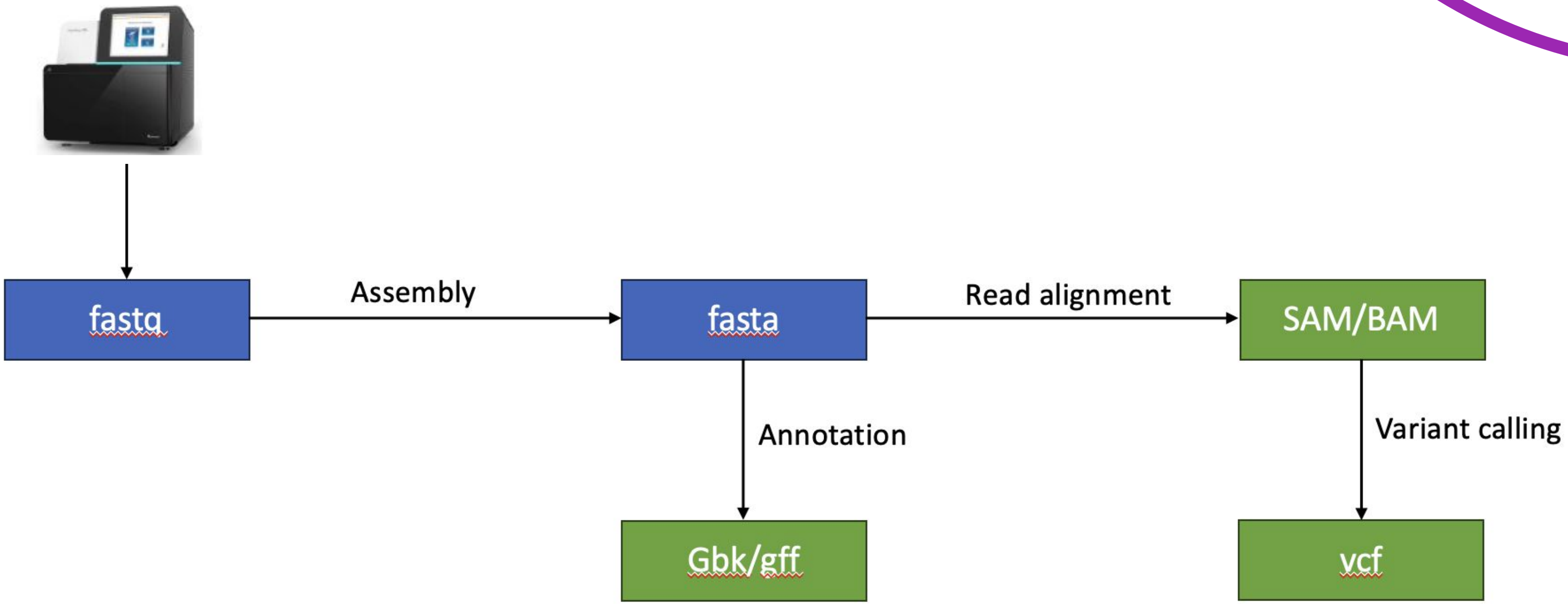
- Searching and filtering
  (*grep, find, locate*)

- Process control
  (*kill*)

- Tricks and shortcuts
  (tab complete, wildcard, arrows)

https://cdn.hostinger.com/tutorials/pdf/Linux-Commands-Cheat-Sheet.pdf?_gl=1*1qpygsr*_gcl_au*MTc3ODMzODAwMi4xNzE0OTM3MDUx&_ga=2.258160322.999754537.1714937051-998924758.1714937051

Introduction to the basic file formats
- Data transformation



fastq

Assembly

fasta

Read alignment

SAM/BAM

Annotation

Gbk/gff

Variant calling

vcf

Introduction to the basic file formats

- Data is information contained in a file

- Different file formats carry data in a specific design and optimization for some programs to be able to read and display the information in an understandable way

- The basic progression of sequencing data and formats involved:

    - □ **Fastq** – raw sequence reads

    - □ **Fasta** – genome assembly

    - □ **.gff/.gbk** – annotation files

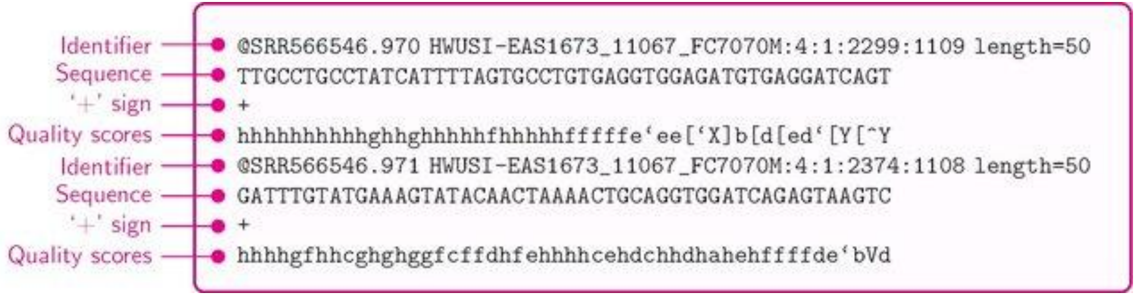
    - □ **SAM/BAM** – read alignment

    - □ **VCF** – variant calling

Introduction to the basic file formats

Fastq

- Standard format directly from the sequencing instrument



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

Fasta

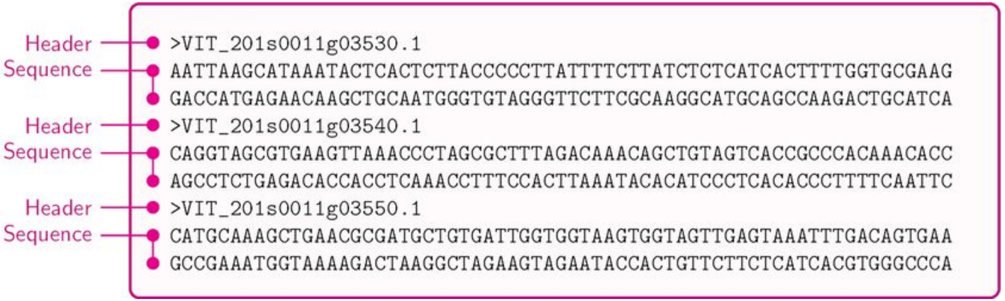
- Following genome assembly
- fasta format is much simpler
- Contains only an identifier and the sequence
- Can also be amino acids (20 letters)

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Introduction to the basic file formats

Multifasta

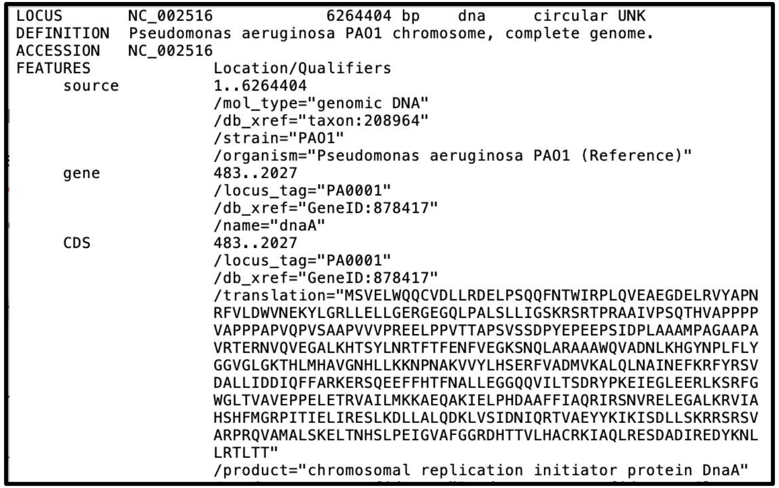
- Concatenated fasta files usually following multisequence alignment



```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

Genbank (gbk)

- Genome annotation output
- Combination of computer and human-readable information
- Still need another application to visualize annotations

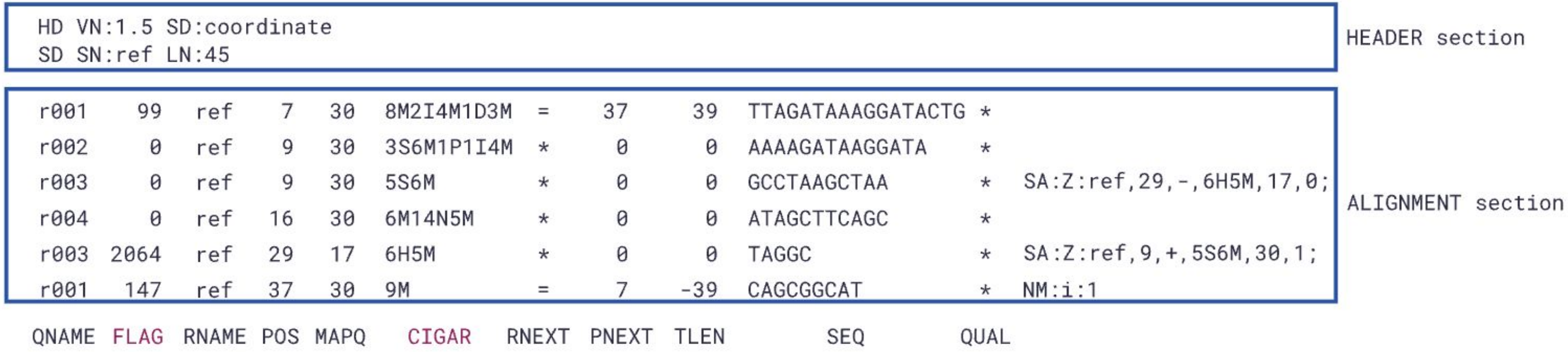


```
LOCUS       NC_002516             6264404 bp    dna     circular UNK
DEFINITION  Pseudomonas aeruginosa PAO1 chromosome, complete genome.
ACCESSION   NC_002516
FEATURES             Location/Qualifiers
     source          1..6264404
                     /mol_type="genomic DNA"
                     /db_xref="taxon:208964"
                     /strain="PAO1"
                     /organism="Pseudomonas aeruginosa PAO1 (Reference)"
     gene            483..2027
                     /locus_tag="PA0001"
                     /db_xref="GeneID:878417"
                     /name="dnaA"
     CDS             483..2027
                     /locus_tag="PA0001"
                     /db_xref="GeneID:878417"
                     /translation="MSVELWQQCVDLLRDELPSQQFNTWIRPLQVEAEGDELRVYAPN
                     RFVLDWVNEKYLGRLLELLGERGEGQLPALSLLIGSKRSRTPRAAIVPSQTHVAPPPP
                     VAPPPAPVQPVSAAPVVVPREELPPVTTAPSVSSDPYEPEEPSIDPLAAAMPAGAAPA
                     VRTERNVQVEGALKHTSYLNRTFTFENFVEGKSNQLARAAAWQVADNLKHGYNPLFLY
                     GGVGLGKTHLMHAVGNHLLKKNPNAKVVYLHSERFVADMVKALQLNAINEFKRFYRSV
                     DALLIDDIQFFARKERSQEEFFHTFNALLEGGQQVILTSDRYPKEIEGLEERLKSRFG
                     WGLTVAVEPPELETRVAILMKKAEQAKIELPHDAAFFIAQRIRSNVRELEGALKRVIA
                     HSHFMGRPITIELIRESLKDLLALQDKLVSIDNIQRTVAEYYKIKISDLLSKRRSRSV
                     ARPRQVAMALSKELTNHSLPEIGVAFGGRDHTTVLHACRKIAQLRESDADIREDYKNL
                     LRTLTT"
                     /product="chromosomal replication initiator protein DnaA"
```

Introduction to the basic file formats

Sequence alignment map (SAM)

- A tab-delimited, line oriented text format of information on alignments
- Header section with metadata in each column
- Alignment section with corresponding information on the alignment



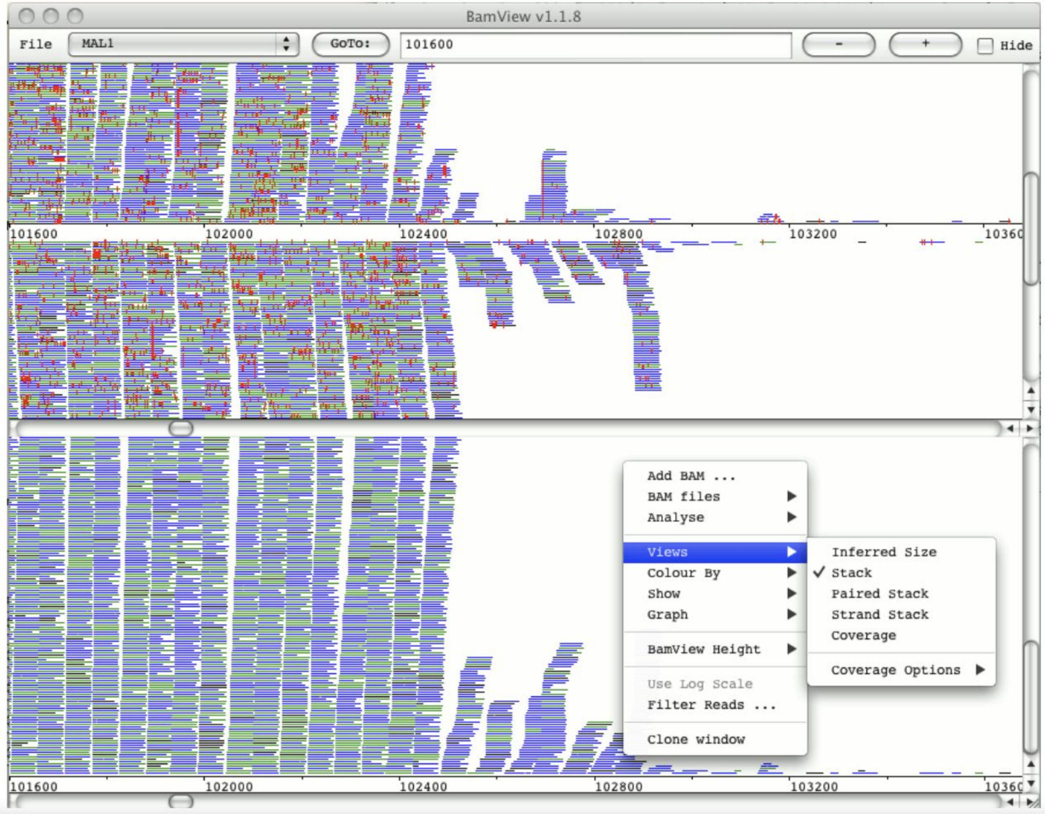
```
HD VN:1.5 SD:coordinate                                                          HEADER section
SD SN:ref LN:45

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;
                                                                                          ALIGNMENT section
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

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

https://samtools.github.io/hts-specs/SAMv1.pdf

Introduction to the basic file formats

Binary alignment map (BAM)

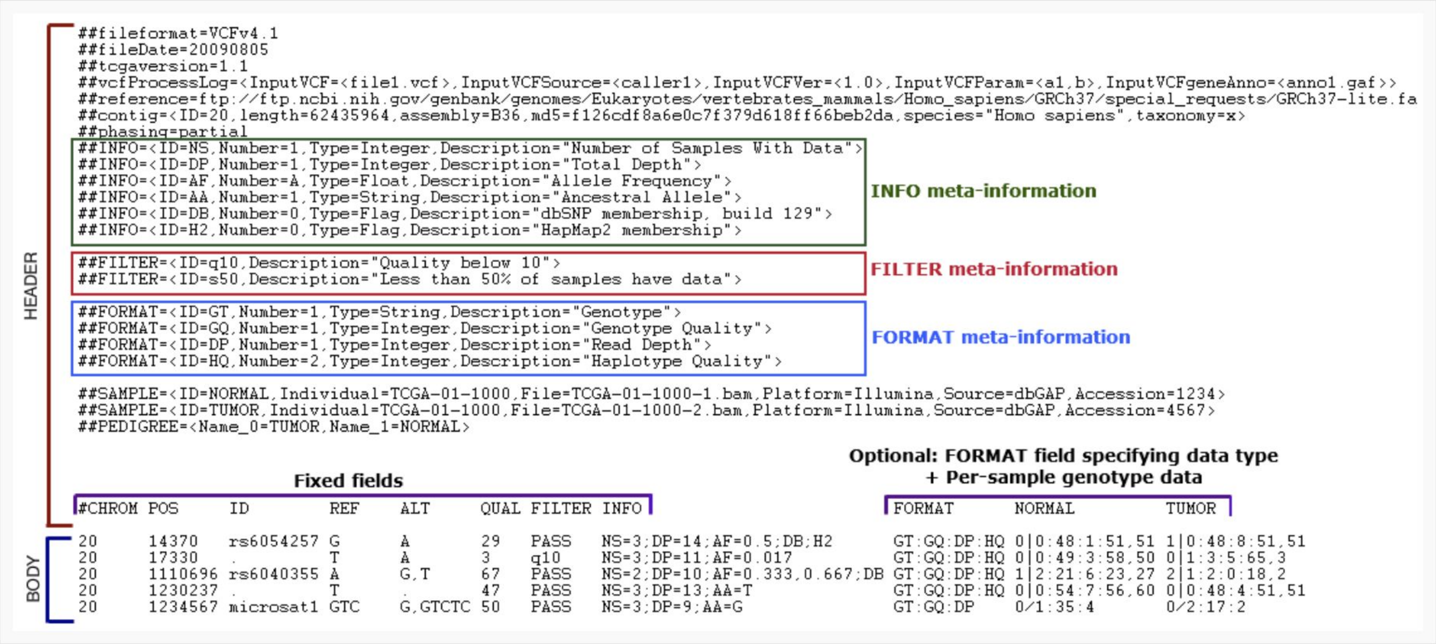
- Compressed version of SAM
- Not human-readable
- Can be recognized by certain programs that allow you to visualize the alignment (e.g IGV, Artemis)
- Ready for variant calling



Introduction to the basic file formats

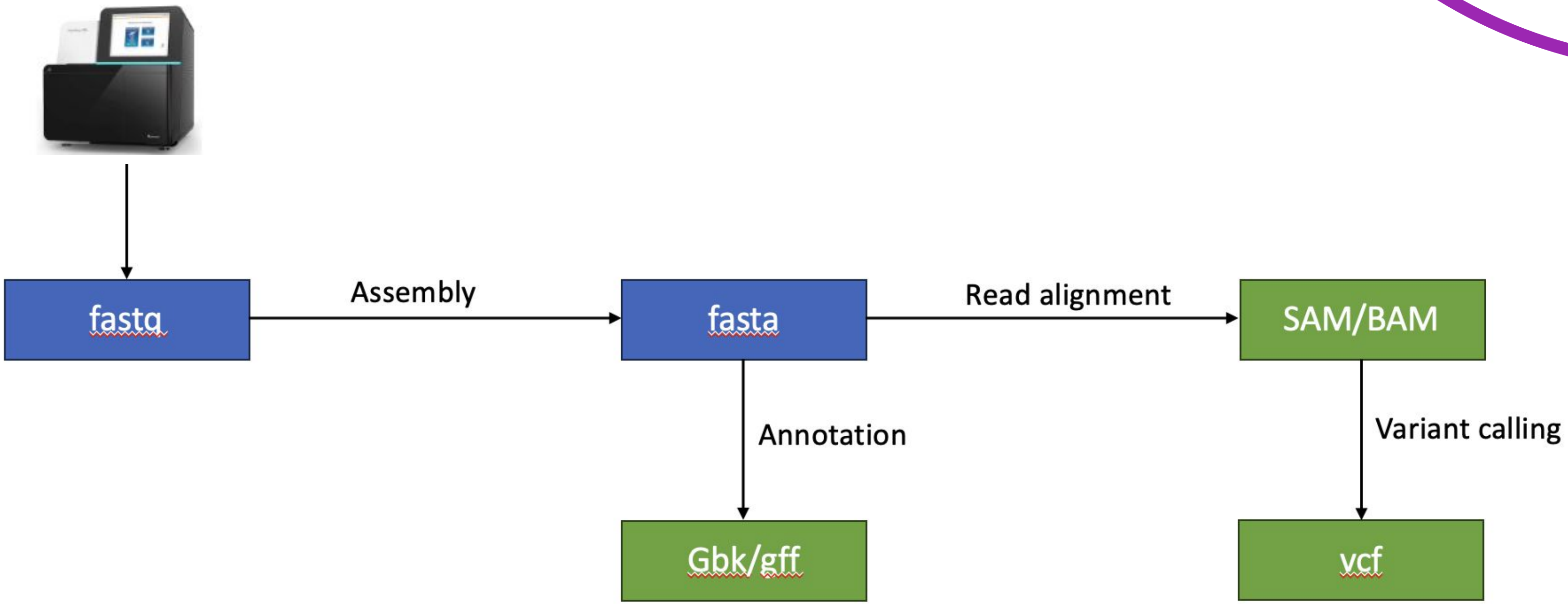
Variant call format (VCF)

- Output from extraction of variance/variations between the query sequence and the reference sequence directly from the BAM file.



```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##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">
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>

#CHROM POS     ID         REF ALT     QUAL FILTER INFO                       FORMAT      NORMAL         TUMOR
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
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
20     1110696 rs6040355  A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
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
20     1234567 microsat1  GTC G,GTCTC 50   PASS   NS=3;DP=9;AA=G             GT:GQ:DP    0/1:35:4       0/2:17:2
```

Introduction to the basic file formats
- Data transformation



fastq

Assembly

fasta

Read alignment

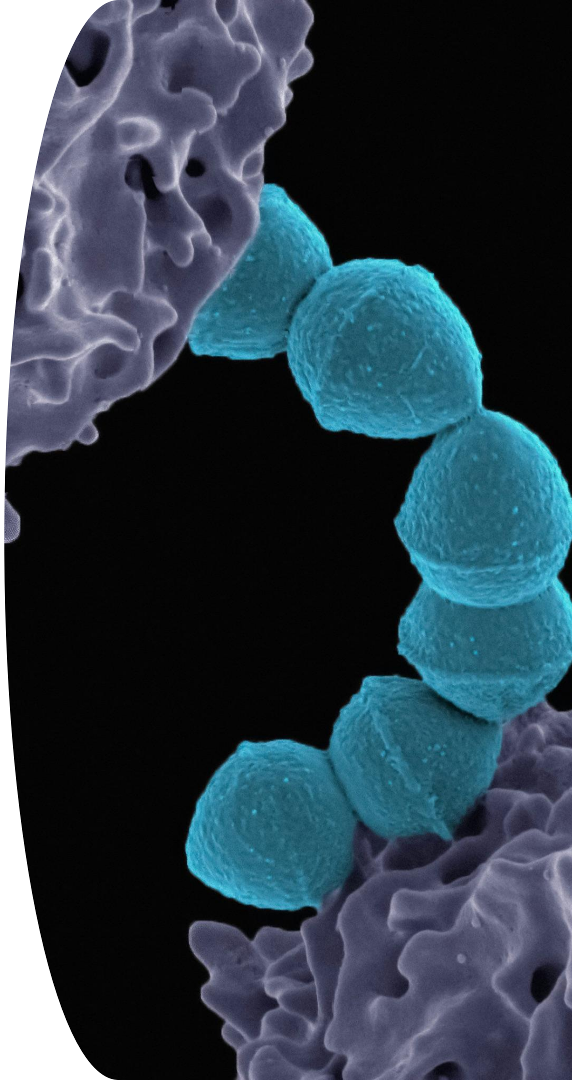
SAM/BAM

Annotation

Gbk/gff

Variant calling

vcf



Acknowledgements:

- Collins Kigen (Kenya)
- Jorge da Rocha (United kingdom)
- Progress Dube (Zimbabwe)
- Marcela Suarez Esquivel (Cosat Rica)