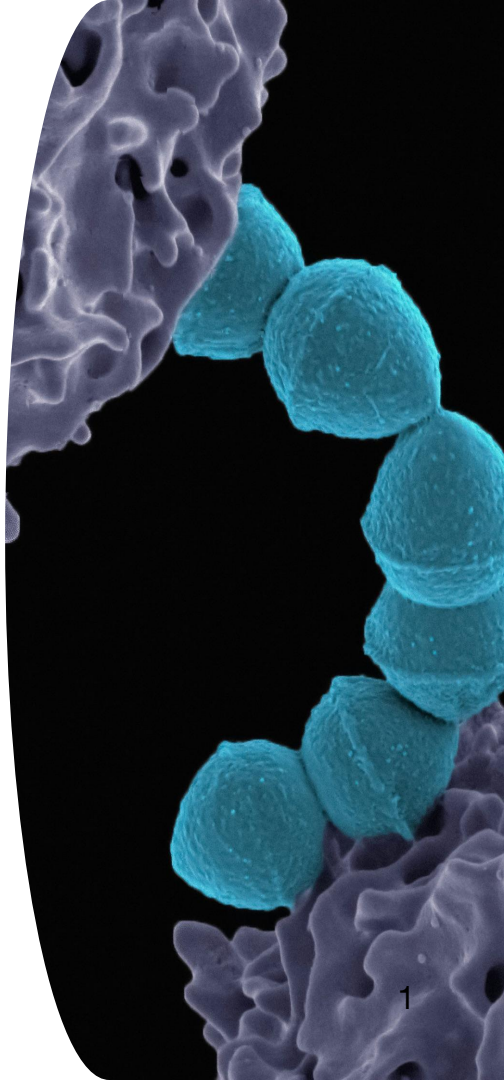
wellcome connecting science

ACORN

Bash, File formats and Quality Control

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



Instructor:
Rito Mikhari (PhD candidate)

AMR Virtual course (Africa and Asia)

21 May 2024

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

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

SAPPHGenE

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

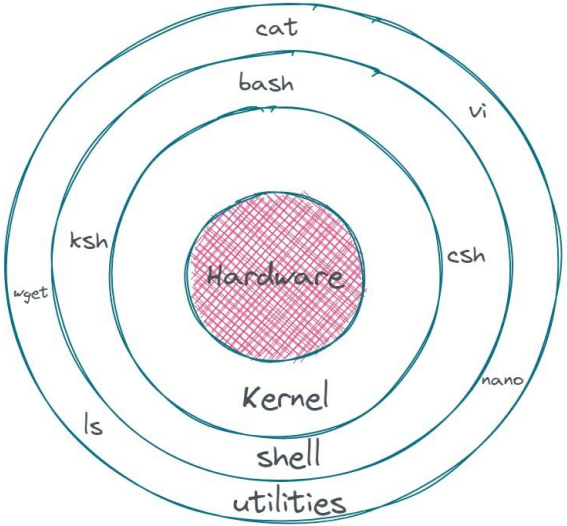
- An operating system (OS) is the software that functions as an interface between the computer user and its hardware
- The software enables and controls the set of programs that we install in our machines
- Some functions include: allocation of storage, processor management, security, error detection,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

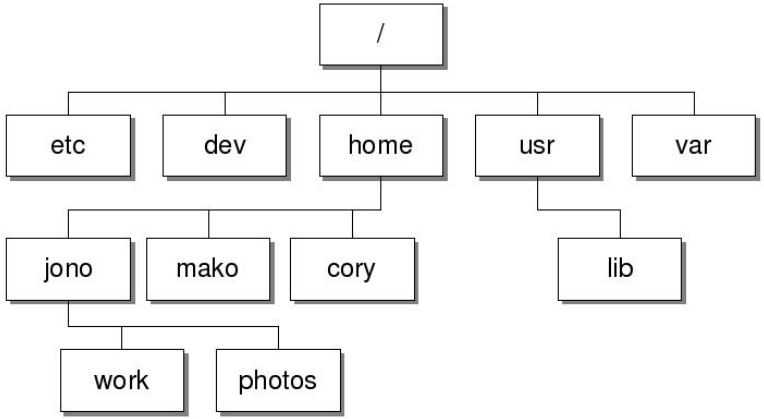
Background: Unix

- 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
  - Commands are case sensitive





Basic Unix: Commands

File system navigation commands
- pwd – print working directory
- ls – list directories and file
- cd – change directory
- mkdir – make new directory
- touch – create and empty file
- rm – remove
- mv – move
- cp – copy
- wget – downloads a file with a provided URL

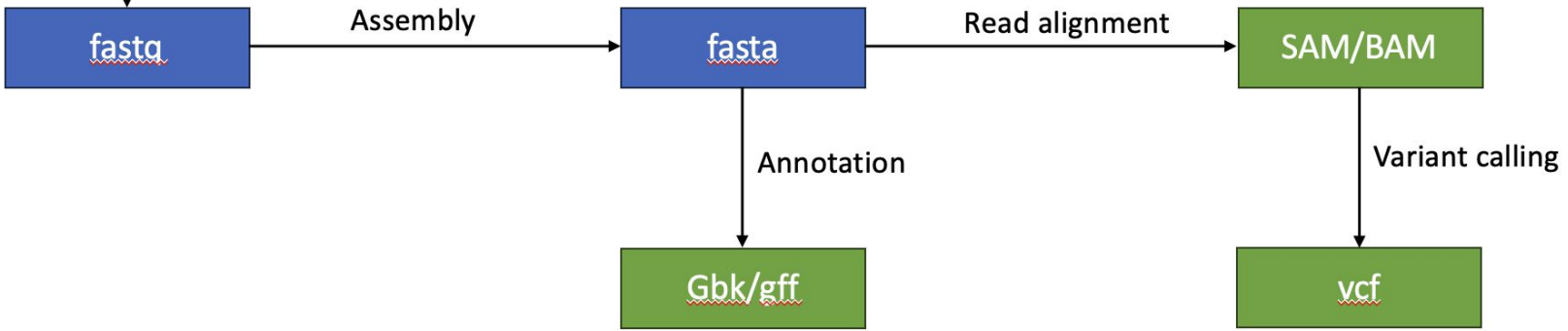
Tricks and other commands
- man – prints command manual
- Tab complete
- Wildcard
- Arrows
- history – prints your command history trail
- Pipes (|)
- kill command – terminates job

Text management
- nano/less – view file
- cat – concatenate and pint file
- head – print first lines from a file
- tail – print last file from a file
- cut – retrieves data from selected columns in a tab-delimited file
- grep – prints line matching a pattern
- wc – counts lines and words

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

SEQUENCING FILE FORMATS

Introduction to the basic file formats - Data transformation



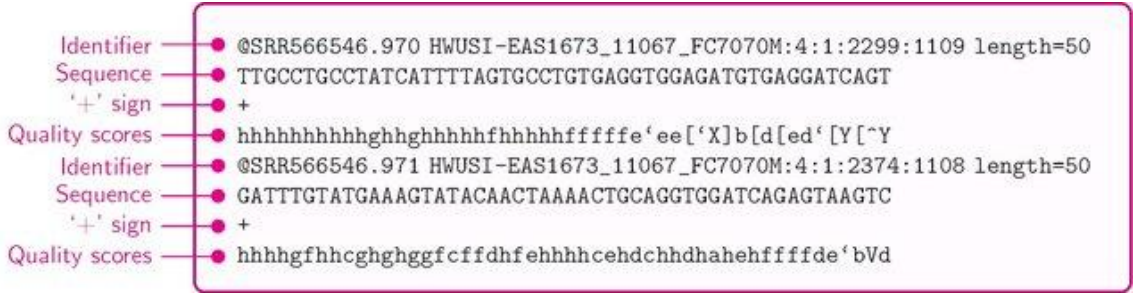
Introduction to the basic file formats

- 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



Fasta

- Following genome assembly
- fasta format is much simpler
- Contains only an identifier and the sequence

```
>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

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
SD SN:ref LN:45
```

HEADER section

```
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
```

ALIGNMENT section

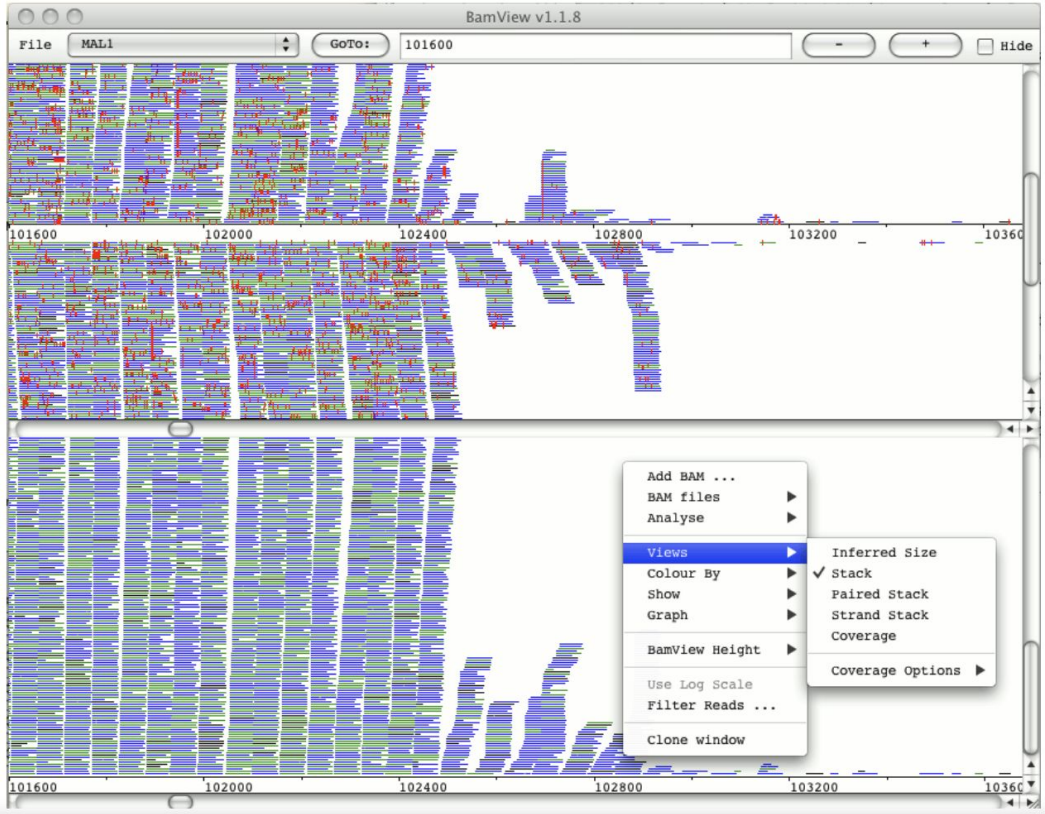
QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL

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

Introduction to the basic file formats

Binary alignment map (BAM)

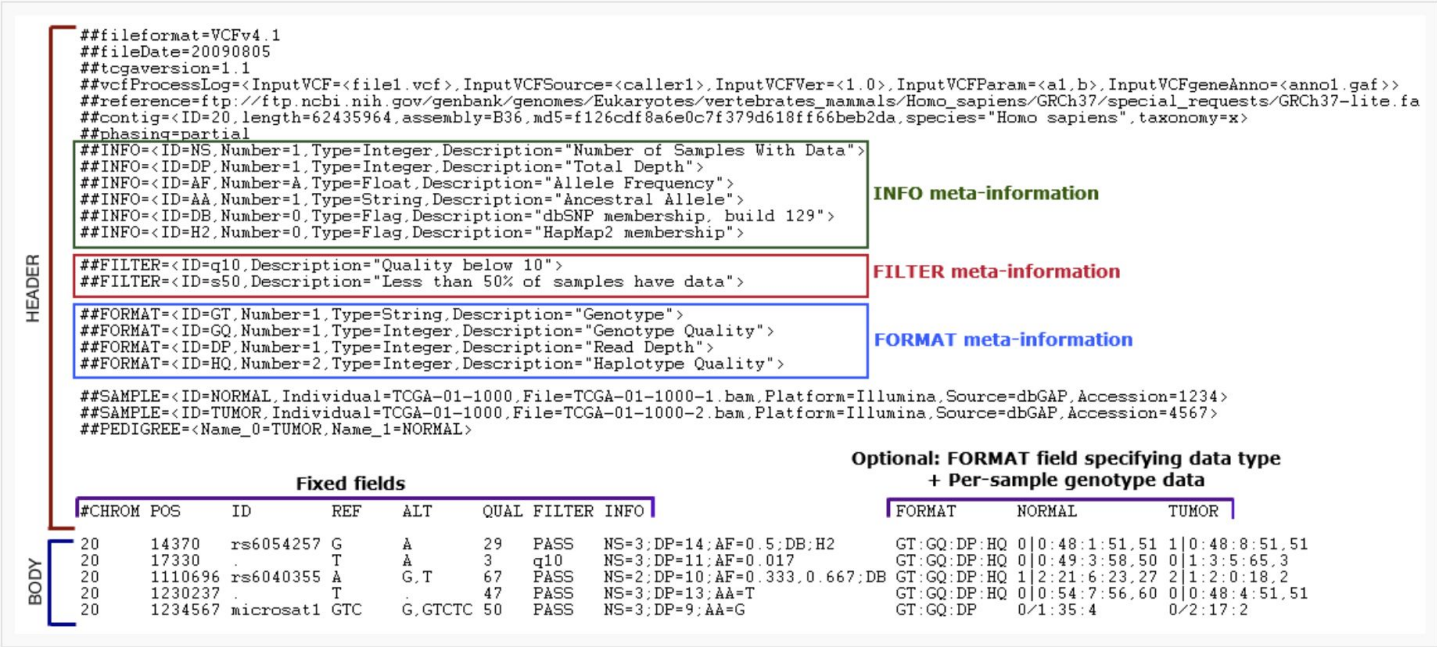
- Compressed version of SAM
- Not human-readable
- Can be recognized by certain programs that allow you to visualize the alignment (e.g IGV, Artemis)
- Ready for variant calling



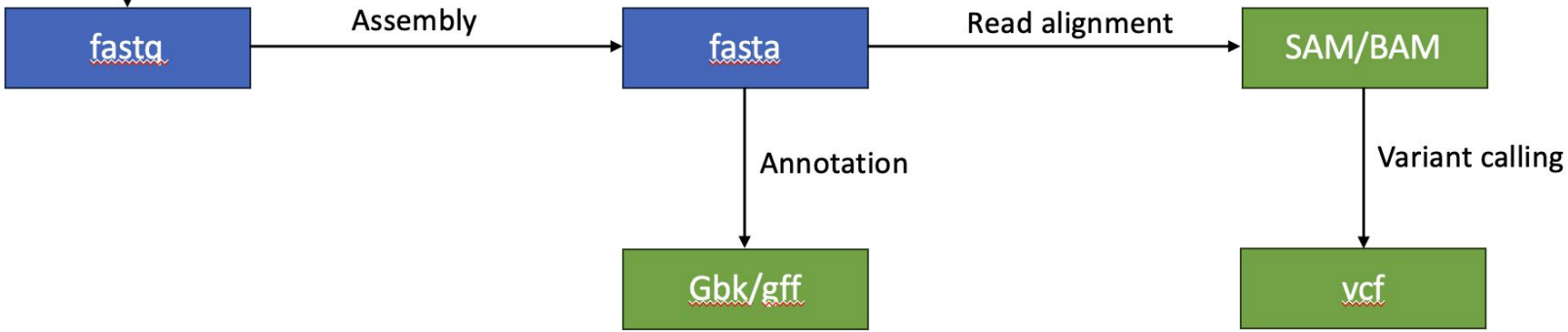
Introduction to the basic file formats

Variant call format (VCF)

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



Introduction to the basic file formats - Data transformation



Acknowledgements:

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