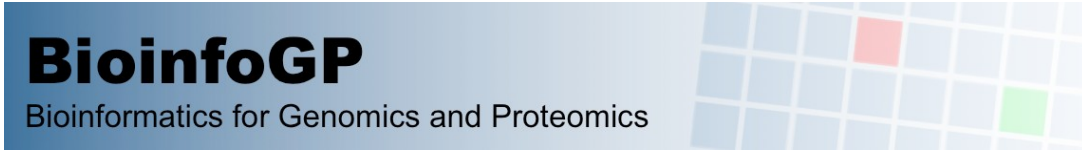
UNIX command line tools for manipulation and analysis of genomic data

#AprendeBioinformáticaEnCasa

YOUR HOME 13/04/2020

Juan Antonio García-Martín



File systems and file paths

File systems




A file system is the way an operative system stores and retrieves data, including:

- Logical drives
- Folders/directories
- Files

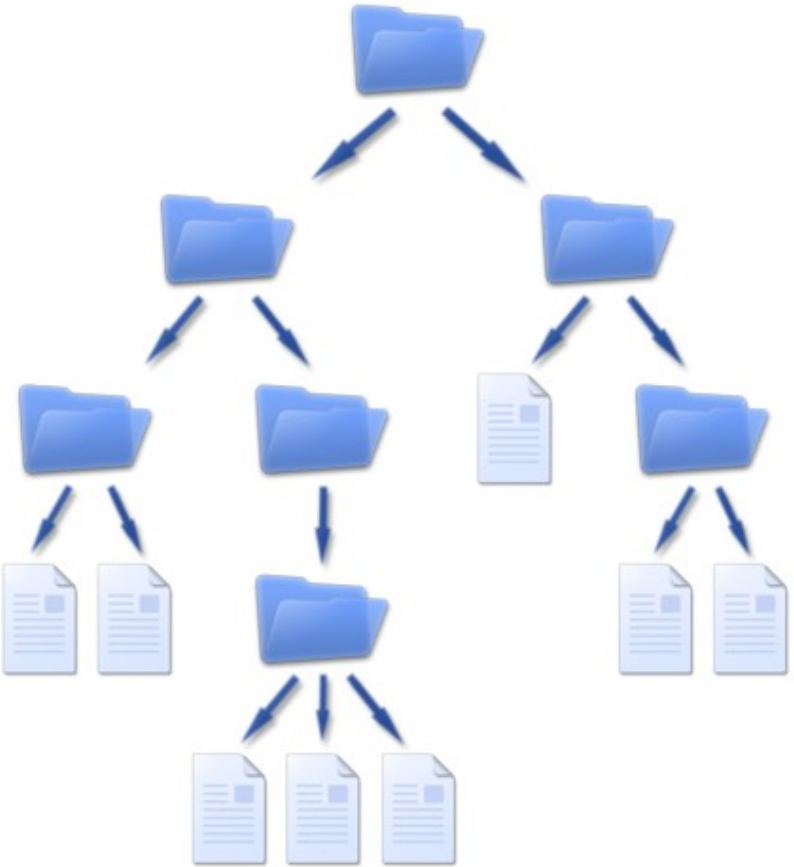
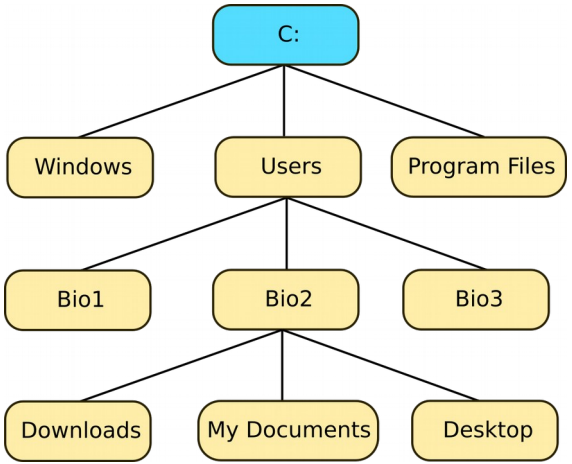


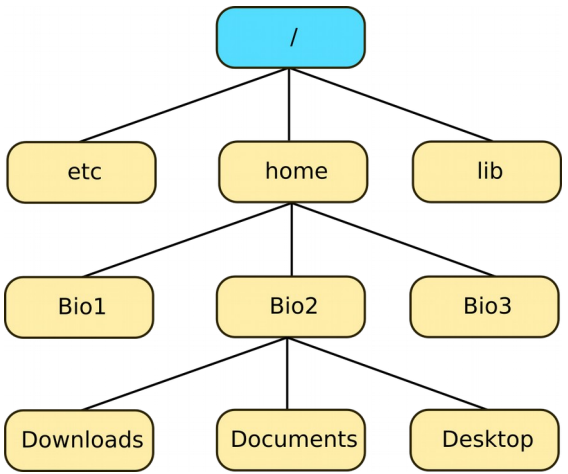
Image: De FrancoGG – Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=766938

Directory structure by OS

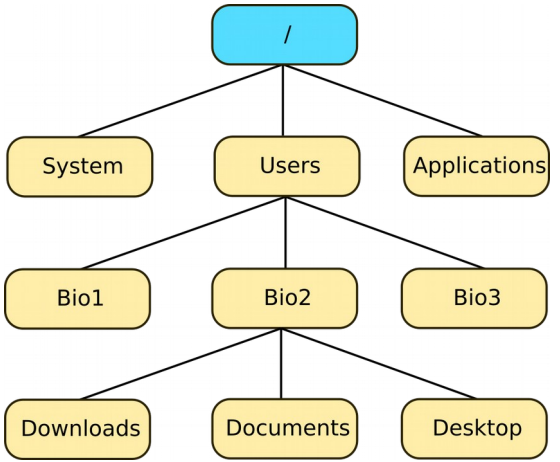
- Directory structure depends on the OS
- The root directory is the lowest level



Windows



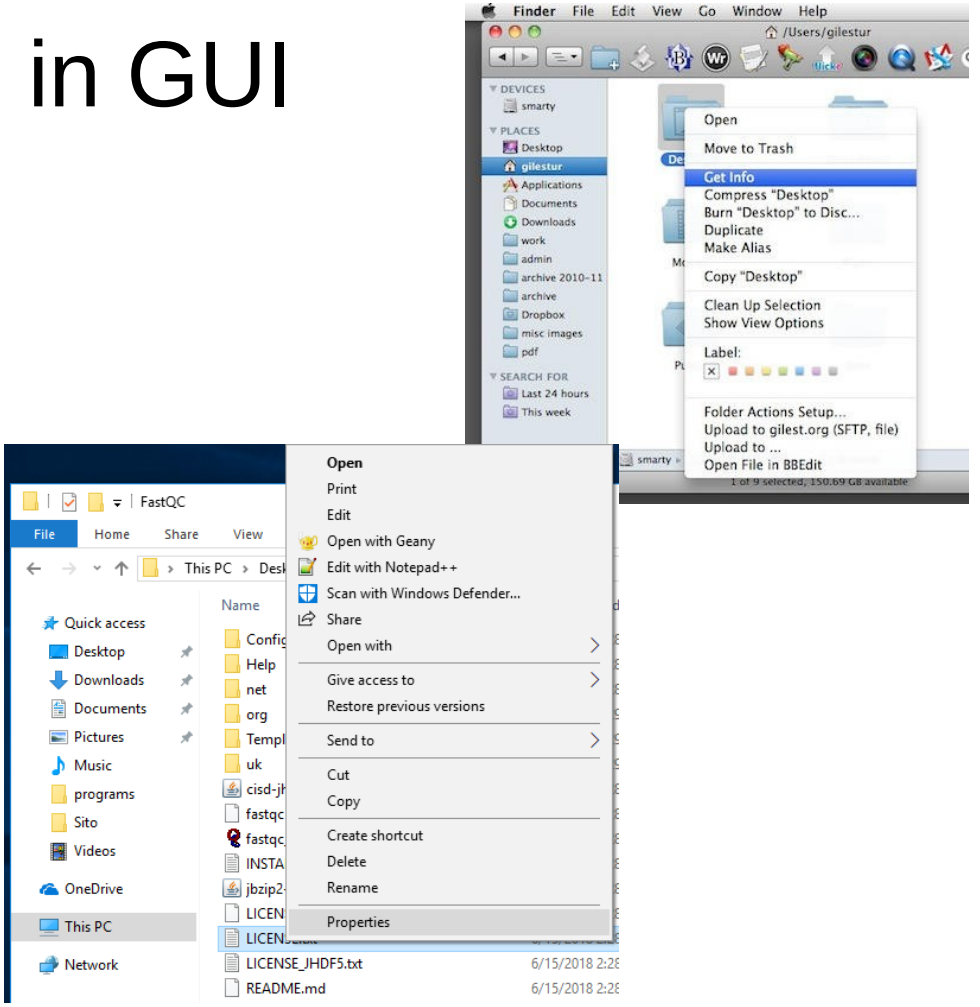
Linux



OsX (Mac)

File path in GUI

- Graphical user interfaces (GUI) facilitate navigation through the file system.
- The vast majority of bioinformatic tools do not use the GUI.
- We must learn to obtain the **FILE PATH**, the full name assigned file or directory in our file system.



Writing file paths

File path: Text with the location of a file or directory.

	Separator	Case sensitive
• Windows – DriveLetter:\Directory\Filename.extension – e.g. • C:\Users\Bio2\Dowloads\MyFile.txt	\	NO
• Os X (Mac) – /Directory/Filename.extension – e.g. • /Users/Bio2/Dowloads/MyFile.txt	/	YES
• Linux – /Directory/Filename.extension – e.g. • /home/Bio2/Dowloads/MyFile.txt	/	YES

Spaces and quotation in paths

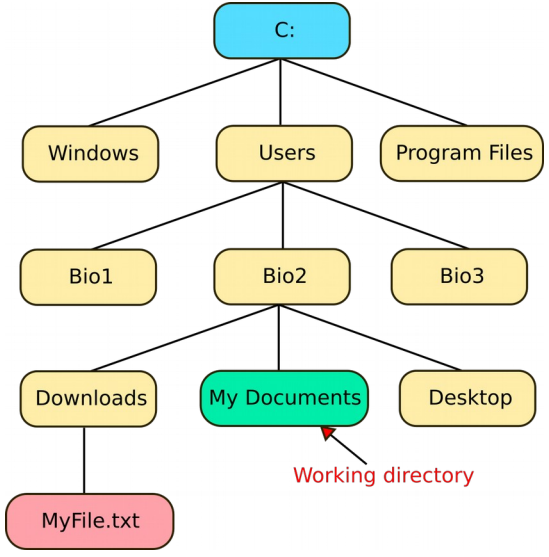
- Spaces are allowed in file paths, however they require an special notation.

- The best way to manage these file paths is enclosing names in quotes, which can be either single ' or double "

- Quotes must be straight, (" , ') be careful to avoid curved quotation symbols (“ ”, ‘ ’, `)

Working directory and relative paths

- The working directory is that from where we execute a command or program.
- In a path, a dot . indicates the current directory, and a double dot .. the parent directory.
- Paths can be:
  - Absolute: The full route of the file or directory.
  - Relative: The route of the file or directory relative to the working directory.



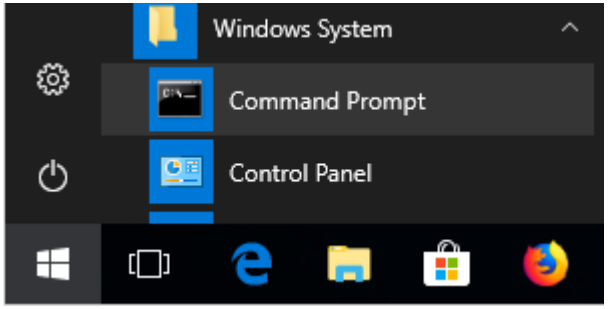
- Absolute:
  `C:\Users\Bio2\Downloads\MyFile.txt`
- Relative to working directory:
  `..\Downloads\MyFile.txt`
- Mixed
  `C:\Users\Bio1\..\Bio2\.\Downloads\MyFile.txt`

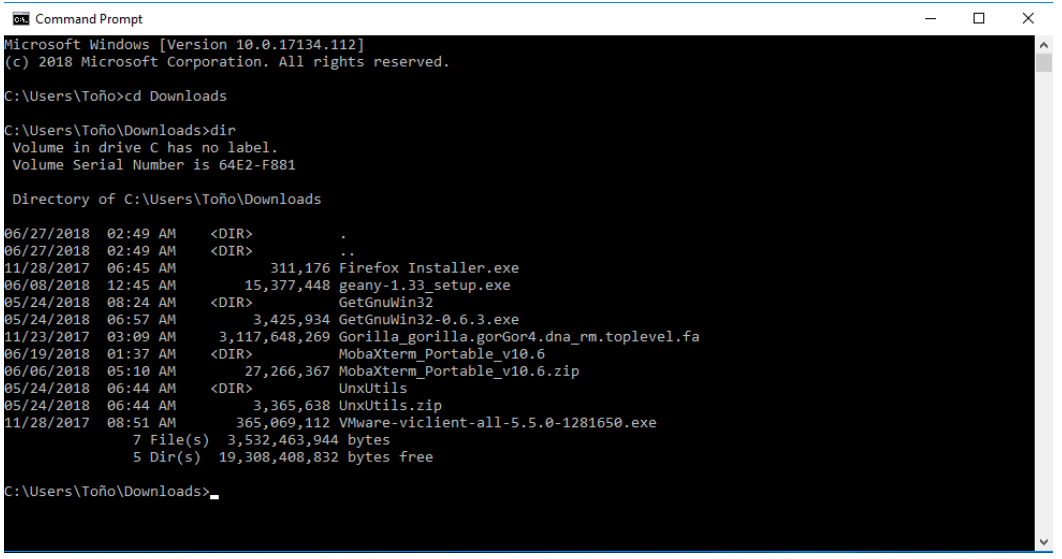
Using the terminal to access files

- Launch a command line terminal
  - Windows → Command prompt / cmd
  - Os X/Linux → Terminal
- Change directory
  - Change working dir to Directory

    `cd directory`
  - Go to the parent directory

    `cd ..`
- List directory contents
  - Windows → `dir`
  - Os X/Linux → `ls`





Biological data file formats and quality control

File types

Plain text

- Plain text tables separated by:
  - Commas (.csv)
  - Tabs (.txt,.tab)
  - Spaces
- Text files (.txt)
- Biological data
  - Fasta (.fas, .fasta,.faa)
  - Sequence reads (.fastq)
  - Genomic features file (.gff, .gtf)
  - Feature track files (.bed)
  - Raw alignments (.sam)
  - Variant calling files (.vcf)
- Source code (example.R)

Formatted text

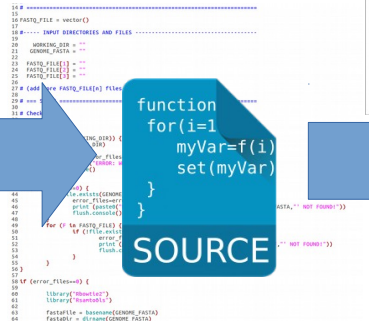
- Excel (.xls,.xlsx)
- Adobe PDF (.pdf)
- Word (doc, docx)
- Enriched format(.rtf)

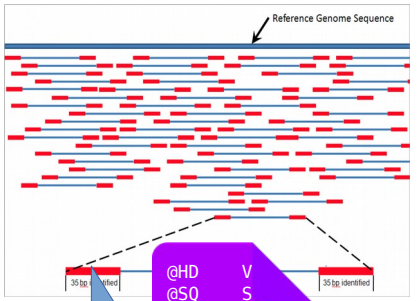
Binary

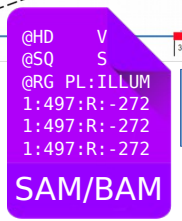
- Images
- Compressed alignments (.bam)
- Executables (.exe)

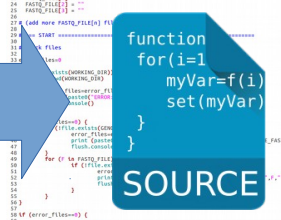
Biological file types

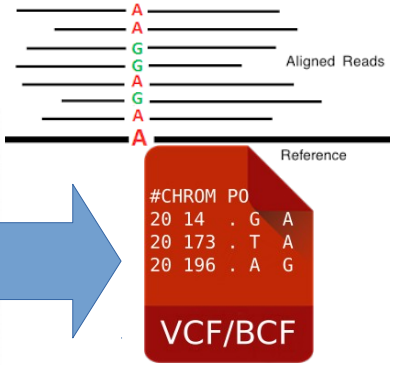
@ SEQNAM
AGCTAGCG
+ SEQNAME
FFBFFIFIFF
@ SEQNAME2
FASTQ











> SEQ
AGCTAGCG
CGTACAGCGT
GCTACTGCGA
GCTATCGCAT
FASTA

1 ncbi m
1 ncbi f
1 ncbi gene
1 ncbi mRNA
1 ncbi exon
1 ncbi cds
GFF/GTF

(Optional)

NCBI

Ensembl

chr7 0
chr7 12
chr7 50 110
chr7 150 532
chr7 607 805
BED

TABLE

Sequence files (FASTA)

- The simplest format to store biological sequences (nucleic acids and proteins)
- Common file extensions (.fa .fas, .fasta)
- Genomes are usually stored in this format, where each sequence is a chromosome or a contig in not fully sequenced genomes.

```
>ENA|AL022645|AL022645.1 Mus musculus cDNA clone 528-1F10 5&apos;;
CGACACGCGTCCGCTTTGTGACTTCACCATGGCGTACCGCGGCCAGGGCCAAAAGGTGCA
GAAGGTGATGGTGCAGCCCATCAACCTTATCTTCAGATACTTGCAAAATAGATCTCGAAT
TCAGGTGTGGCTGTATGAACAAGTGAATATGCGGATAGAAGGTTGTATTATTGGCTTTGA
TGAGTACATGAACCTCGTATTAGATGATGCAGAAGAGATTCATTCTAAAACAAAGTCAAG
AAAACAACTGGGTCGGATCATGCTAAAAGGAGATAATATTACTCTGCTCCAAAGTGTTTC
CAACTAGCAGTGAATGGTGAAGTCTGTAGGATGTTGAGAAGACCCCTTGAGCGTGTTTAA
AGATGTCTGCTGCAACCTGCATTTACTCAACTTGTTTTACTTGCACATTATTATTAGGTG
ACAATAAATGCTGTAGGAAGTTTTTGT

>ENA|AL022645|AL022645.1 Mus musculus cDNA clone 528-1F10 5&apos;;
CGACACGCGTCCGCTTTGTGACTTCACCATGGCGTACCGCGGCCAGGGCCAAAAGGTGCA
GAAGGTGATGGTGCAGCCCATCAACCTTATCTTCAGATACTTGCAAAATAGATCTCGAAT
TCAGGTGTGGCTGTATGAACAAGTGAATATGCGGATAGAAGGTTGTATTATTGGCTTTGA
TGAGTACATGAACCTCGTATTAGATGATGCAGAAGAGATTCATTCTAAAACAAAGTCAAG
AAAACAACTGGGTCGGATCATGCTAAAAGGAGATAATATTACTCTGCTCCAAAGTGTTTC
CAACTAGCAGTGAATGGTGAAGTCTGTAGGATGTTGAGAAGACCCCTTGAGCGTGTTTAA
AGATGTCTGCTGCAACCTGCATTTACTCAACTTGTTTTACTTGCACATTATTATTAGGTG
ACAATAAATGCTGTAGGAAGTTTTTGT
```

Header starts with a right angle **>**

No empty lines between header and sequence.

Sequences can be in one or multiple lines

Fasta files can contain one or many sequences

Genomic features (GTF/GFF)

- General Feature Format (GFF) is a tab separated file used to encode features in genomic sequences. Is the most common format for annotating genomes.
- The features defined include genes, transcripts, coding and non-coding regions, exons, ncRNAs, etc.
- Provide the coordinates, strand, hierarchy, coding frame and other attributes of each feature.
- There are three versions with slight modifications (GFF, GTF and GFF3)

```
1   ensembl_havana   gene   923928   944581   .   +   .   ID=gene:ENSG00000187634;Name=SAMD11;biotype=protein_coding;description=sterile
alpha motif domain containing 11;
1   havana   mRNA  923928   939291   .   +   .   ID=transcript:ENST00000420190;Parent=gene:ENSG00000187634;Name=SAMD11-203;
1   havana   five_prime_UTR 923928   924431   .   +   .   Parent=transcript:ENST00000420190
1   havana   exon  923928   924948   .   +   .   Parent=transcript:ENST00000420190;Name=ENSE00001637883;exon_id=ENSE00001637883;
1   havana   CDS   924432   924948   .   +   0   ID=CDS:ENSP00000411579;Parent=transcript:ENST00000420190;protein_id=ENSP00000411579
1   havana   exon  925922   926013   .   +   .   Parent=transcript:ENST00000420190;Name=ENSE00003794726;exon_id=ENSE00003794726;
1   havana   CDS   925922   926013   .   +   2   ID=CDS:ENSP00000411579;Parent=transcript:ENST00000420190;protein_id=ENSP00000411579
1   havana   exon  930155   930336   .   +   .   Parent=transcript:ENST00000420190;Name=ENSE00002727207;exon_id=ENSE00002727207;
1   havana   CDS   930155   930336   .   +   0   ID=CDS:ENSP00000411579;Parent=transcript:ENST00000420190;protein_id=ENSP00000411579
1   havana   exon  931039   931089   .   +   .   Parent=transcript:ENST00000420190;Name=ENSE00002696520;exon_id=ENSE00002696520;
```

Sequence reads (FASTQ)

- The most common output format for sequenced reads.
- Each read information is encoded in four lines

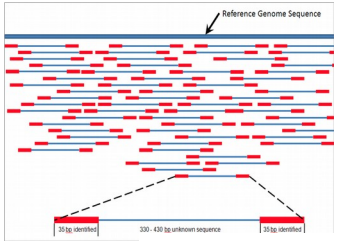
```
@SRR3214175.10 HISEQ01:422:HABGFADXX:1:1101:1547:2189/1
CGTGTCACAAATTTGCTTGTACACTCTGCCAAGGCATCATTCCTAGCCTTT
+
BBBFFFFFFFFFFIIIIIIFIIIIIFIIIIIIIFIFFFBFFIFIFFFFFIF
@SRR3214175.20 HISEQ01:422:HABGFADXX:1:1101:2196:2034/1
GGGGCTGGCACGGAGCAGAGTGATTCGTAGAGCAGAGGGTCAGCTCCCTGG
+
BBBFFFFFFFFFFIIIIIIIFFIIIIIFFFFIIIIIIIIBFFFFFFIIFII
@SRR3214175.30 HISEQ01:422:HABGFADXX:1:1101:2768:2070/1
CTTATTTGCTGAGAACAGTAAGATCTGCAGGCAGTTTCAGATGAAACAGGC
+
BBBFFFFFFFFFFIIIIIFIIIIIIIIIIFIIIIIIIIIFIFIIIFFIFF
@SRR3214175.40 HISEQ01:422:HABGFADXX:1:1101:3177:2234/1
AATCATTTGAGAGGTTTTGCCATAATTATTACATGAGGCATAGAAAGTTTG
+
BBBFFFFFFFFFFI<FFIIIIIIIIIIIIIIIIIIIIIIFIIIIIIIIIII
@SRR3214175.50 HISEQ01:422:HABGFADXX:1:1101:3661:2200/1
CAAAAGTATGTCCTTCTAGTTCCGTGTGGCCTTTGTCAGCATGGACAGCAT
+
BBBFFFFFFFFFFIIIIIIIIIIIFIFFIIIIIIIFFIFIFIIIIIIFIIF
```

Line 1: Begins with '@' followed by a **sequence identifier** and an **optional description.**

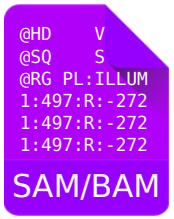
Line 2: **Raw sequence** (IUPAC single letter codes).

Line 3: begins with a '+' , optionally followed by the same sequence identifier (and any description) again.

Line 4: **Quality values** for the sequence in Line 2 . Quality ranges from 1 to 40 in most platforms, however the letters used for encoding them can vary.



Alignment files (SAM/BAM)



- Sequence alignment map files store biological sequences (reads) aligned to a reference sequence.
- The SAM format consists of a header and an alignment section.
  - Headings begin with a @, and precede the alignment data.
  - Alignment lines have 11 mandatory fields, containing information such as the sequence identifier, position, quality and length of the alignment, if the read align in more than one locus, etc.

```
@HD    VN:1.0 SO:coordinate
@SQ    SN:1   LN:249250621  AS:NCBI37    UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ    SN:2   LN:243199373  AS:NCBI37    UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:a0d9851da00400dec1098a9255ac712e
@SQ    SN:3   LN:198022430  AS:NCBI37    UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:fdfd811849cc2fadebc929bb925902e5
@RG    ID:UM0098:1   PL:ILLUMINA   PU:HWUSI-EAS1707-615LHAAXX-L001   LB:80  DT:2010-05-05T20:00:00-0400      SM:SD37743    CN:UMCORE
@RG    ID:UM0098:2   PL:ILLUMINA   PU:HWUSI-EAS1707-615LHAAXX-L002   LB:80  DT:2010-05-05T20:00:00-0400      SM:SD37743    CN:UMCORE
@PG    ID:bwa VN:0.5.4

1:497:R:-272+13M17D24M     113    1      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     1      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    1      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     1      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
```

Basic formats: Tables

- Tables in plain text use delimiters
  - Common delimiters: comma (.csv), tabs (.tsv .txt .tab)
  - Fields are sometimes enclosed by double or single quotes
  - All spreadsheet editors can export to plain text.

```
!strand         slot            start           stop    type                    label           mouseover                       hyperlink
forward         1               34              894     open_reading_frame      ORF             ORF rf1 34-894                   http://wishart.biology.ualberta.ca/cgview/
forward         2               34              894     gene                    AMPr            amp resistance 34-894            http://wishart.biology.ualberta.ca/cgview/
forward         1               1049            1668    origin_of_replication   pBR322          pBR322 orgin 1049-1668           http://wishart.biology.ualberta.ca/cgview/
forward         1               2657            2963    origin_of_replication   f1              f1 orgin 2657-2963               http://wishart.biology.ualberta.ca/cgview/
forward         3               1301            1301    unique_restriction_site AlwNI           AlwNI 1301                       -
forward         3               2099            2099    unique_restriction_site StuI            StuI 2099                        -
forward         3               2101            2101    unique_restriction_site XhoI            XhoI 2101                        -
forward         3               2105            2105    unique_restriction_site SfiI            SfiI 2105                        -
forward         3               2111            2111    unique_restriction_site EcoRI           EcoRI 2111                       -
forward         3               2124            2124    unique_restriction_site BamHI           BamHI 2124                       -
forward         3               2137            2137    unique_restriction_site NotI            NotI 2137                        -
forward         3               2150            2150    unique_restriction_site HindIII HindIII 2150                     -
forward         3               3193            3193    unique_restriction_site ClaI            ClaI 3193                        -
forward         3               3244            3244    unique_restriction_site AccI            AccI 3244                        -
forward         3               3798            3798    unique_restriction_site NcoI            -                               -
forward         1               1977            2006    promoter          lacZ          lacZ promoter 1977-2006          -
forward         1               2180            2339    gene                    lacZ            lacZ reporter 2180-2339          -
forward         1               3853            4084    regulatory_sequence     attP            attP recombination site 3853-4084  -
reverse         1               2156            2174    promoter          T7            T7 promoter 2156-2174            -
```

Other genomic features (BED)

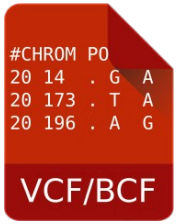
- The Browser Extensible Data (BED) format consists of one line per feature, each containing 3-12 columns of data, plus optional track definition lines. This format is used to include any feature spanning a genomic region.
- Only the the first three fields, which indicate the position of each feature, are mandatory.

```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

- Custom BED track files can contain headers in the first lines providing information for the browser, but these are specific of each program.

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
```

Variant calling (VCF/BCF)



- Standard format for storing SNPs and INDELS
- It starts with meta-information lines, a header line, and then data lines, each containing information about one SNP or INDEL detected.


```
##fileformat=VCFv4.0
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##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
```

Managing biological flat text files

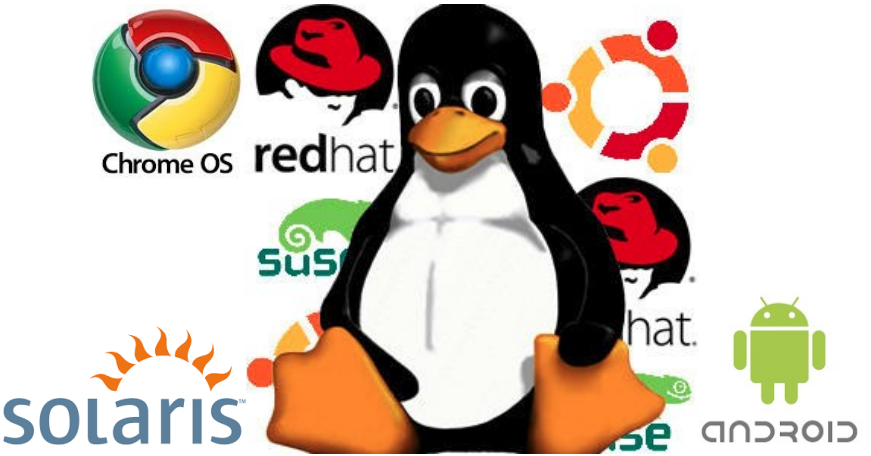
- Biological files are usually huge, so it is often impossible to open these files with text editors.
- It is recommended to check that our files are correct, however visual inspection can become an impossible task due to:
  - File sizes
  - Computational capabilities (Speed, memory)
- Routine check can be performed without specific computational software using basic commands.

Take home message

- Being able to locate files in your computer is necessary for using most bioinformatic analysis tools.
- Biological information is stored in different formats, intended for specific purposes.
- Avoid mixing or changing versions of biological data files during your analysis.

UNIX command line tools

Unix based operating systems



Unix commands in Windows

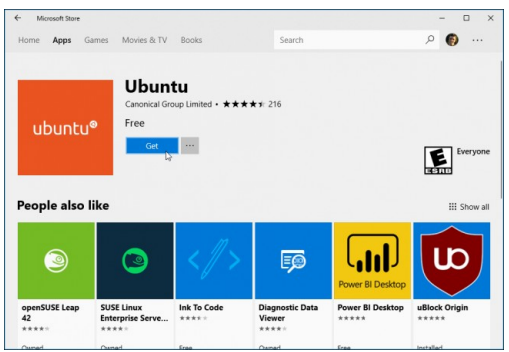
Virtual machines

Microsoft Virtual PC

vmware

VirtualBox

Linux subsystem for Windows



Unix environment emulation

CygWin

MobaXterm

MinGW
Minimalist GNU for Windows

MinGW + MSSYS

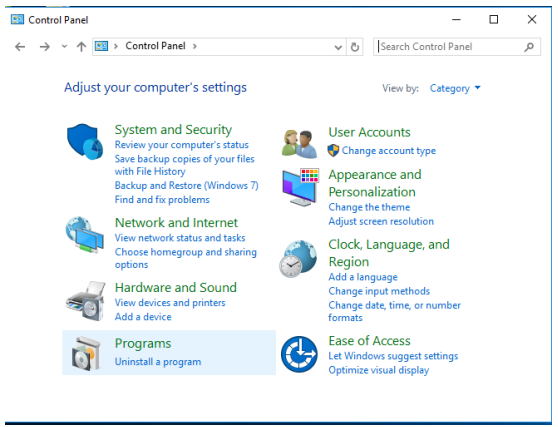
Unix ports for Windows

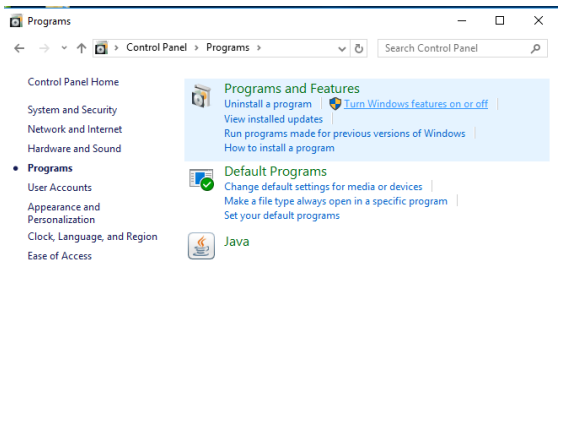
GnuWin

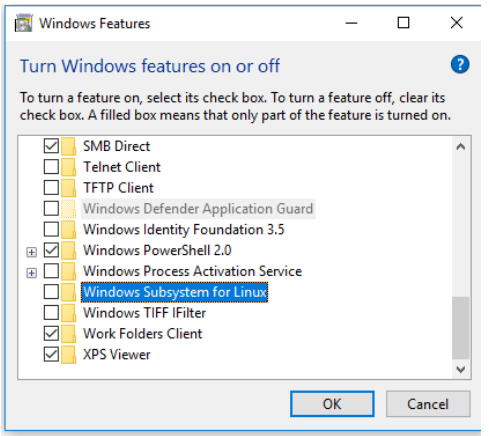
GnuWin

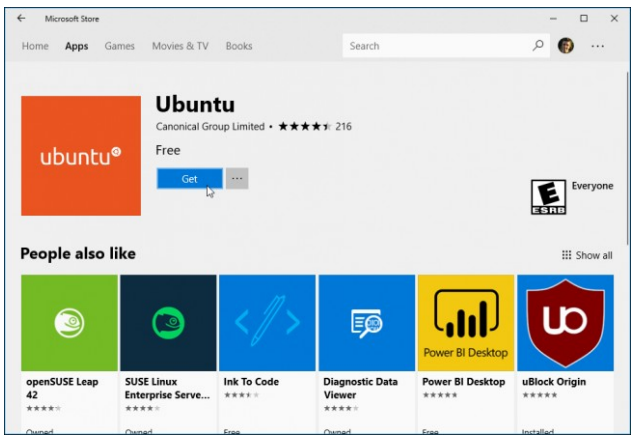
UnxUtils

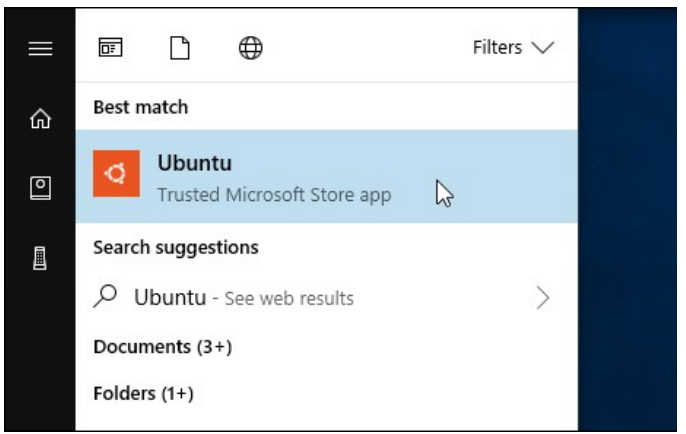
Installing BASH on Windows 10 (64 bits)











Navigating through directories

- `cd` (change directory)
  - Without any parameters goes to home directory.
  - `..` → Upper level directory
  - `.` → Current directory
  - `/` → Root directory
- `ls` (list contents of a directory)
  - Without any parameters shows the content of the current directory
  - `-la` flag shows extended information (date of creation, size...)
- `pwd` (Print Working Directory)
  - Shows the current path

Using quotes

- ' → Single quotes evaluate the content “as is”. No escape or special characters.

- " → Double quotes allow special characters and expressions.

- It is possible to use double quotes inside single quotes and vice versa.

Wildcards

- Wildcards allow to select multiple files or directories
- `*` (Any combination)
  - Single, all elements in the directory
  - In combination with other letters restrict the selection.
    - `*.fas` → Any file ending with .fas
    - `my*` → Any file starting with “my” (myABCD.fas,myFile.txt)
    - `my*File` → Any file starting with “my” and finishing with File (myABCDFile,myFile)(myFile.fas)
- `?` (Any character)
  - Select files where ? Is replaced by exactly one character.
    - `MyFile??.fas` → (myFile01.fas,myFileAB.fas) (myFile1.fas,myFileA.fas)

Create and delete directories

- `mkdir` (make directory)

  -f → Force. Multiple directories can be created at the same time

- `rmdir` (remove directory)

File info and fast visualization

- `wc` (Word Count) → Shows the number of lines, words and characters of a file
  - `wc -l` → Shows the number of lines
  - `wc -w` → Shows the number of words
  - `wc -c` → Shows the number of characters
- `more, less`
  - Shows the contents of a file and loads instantaneously even if the file is huge
  - '`less`' can scroll backwards, but not all Unix implementations contains the '`less`' command
  - Only for visualization

Printing fragments of a file

- `head` → Prints the first lines of file (10 by default)
  - `head -n 5` → Shows the first 5 lines of myFile
- `tail` → Prints the last lines of file (10 by default)
  - `tail -n 5 myFile` → Shows the last 5 lines of myFile
- `cat` (conCATenate) → Print the contents of one or more files
  - `cat myFile myFile2` → prints the concatenated content of myFile and myFile2

Redirections

- > → Prints command output into the specified file, if the file exists it will be overwritten
- >> → Appends output to the specified file
- < → Some programs accept input from the standard input (the keyboard). Left angle redirects the input to a file.

Archive files

- `gzip/gunzip` → Compress and uncompress files using gzip
- `zip/unzip` → Compress and uncompress files using zip
- `bzip2/bunzip2` → Compress and uncompress files using bgzip
- `tar` → Pack or unpack files with tar archiver
  - `-x` extract
  - `-u` create
  - `-z` compresed file
  - `-v` verbose mode
  - `-f` force overwrite

Concatenating commands

- `|` → The pipe symbol allows to concatenate commands, using the output of the first command as input for the second.
- Example:
  - `cat myFile1 myFile2 | sort > myOrderedFile`

Sorting and removing duplicates

- `sort` → Sorts lines of a text file
  - `-u` → remove duplicates
  - `-r` → reverse order
  - `-n` → numeric sort
  - `-h` → human readable numbers sort

- `uniq` → Reports or omit repeated lines
  - Lines must be adacent, often used in combination of sort
  - `-c` → prefix lines ny the number of occurrences
  - `-d` → print duplicate lines

Compare files

- `diff` (Differences) → Compare two files and print lines that differ

- `comm` (Compare sorted files) → Compare two files and print elements unique to file 1, file 2 of common
  - `-1` → suppress column 1 (lines unique to FILE1)
  - `-2` → suppress column 2 (lines unique to FILE2)
  - `-3` → suppress column 3 (lines that appear in both files)

Managing plain text tables

- `cut` → Prints columns of a file separated by delimiters
  - `-d 'delimiter'` → Specify delimiter, can be one or more characters
  - `-f` → Specify the fields to print separated by commas
  - Example:

  `cut -d ',' -f2,5,7 myFile.csv` → Prints columns columns 2,5 and 7 of the comma separated file myFile.csv

- `paste` → Vertically concatenates files
  - `-d 'delimiter'` → Specify output delimiter
  - If the selected files do not have the same number of lines, paste output file will have the same lines as the longest one.

Selecting specific lines of a file

- `grep` → Prints those lines in a file containing a specific word or pattern
  - `-v` → print non-matching lines
  - `-A n` → print n lines after the match
  - `-B n` → print n lines before the match
  - `-f` → read matching sequence from a file (allows multiple matches)
  - `-o` → print only the matching region
  - `-P` → use perl-like regular expressions (does not work in OsX, but regular expressions are allowed)

Modifying file contents

- `tr` → Translate or delete characters in a file.
  - `tr` does not accept a file as parameter, so redirections or pipes are necessary.
  - `tr a b < myFile` → replace all occurrences of a to b in myFile
  - `cat myFile | tr [abc] [def]` → replace all occurrences of a,b or c by d,e or f respectively.
  - `tr -d 'soft' < myFile` → delete all instances of 'soft'
- `sed` → (Stream EDitor) Is a more complex tool for filtering and replacing text.
  - The most common and basic usage is `sed 's/text/replacement/g' myFile` where 's' indicates that we are using the substitution tool and 'g' that is global (otherwise it will only replace the first occurrence)
  - `sed 's/my/your/g'` → replace all occurrences of 'my' by 'your'.
  - `sed 's/U//'` → delete the first U in the file

Basic regular expressions

- `[ ]` → Search for any character enclosed, a ^ at the beginning negates the expression
  - `[ACGU]` → Any line with A,C,G or U
  - `[^ACGU]` → Any line with **no** A,C,G or U
- `( )` → Enclose groups of characters
- `–` → Range of letters
  - `[0-9]` → Any line with a number
  - `[a-z]` → Any line with a lowercase letter
- `+` → At least 1 occurrence of the previous element
  - `[ACGU]+` → Any line with one or more A,C,G or U
- `*` → 0 or many occurrences of the previous element
  - `B([ACGU]*)C` → Any line with B and C with zero or more occurrences of A,C,G or U between them

Basic regular expressions

- `^` → Marks beginning of the line
  - `^>` → Any line that starts with > (like FASTA headers)
- `$` → Marks the end of a line
  - `[ACGU]$` → Any line that ends with A,C,G or U
  - `^[ACGU]+$` → Any line composed only by a combination A,C,G and/or U. Useful for parsing RNA sequences
- `\t` → Tab
- `\n` → New line
- `\` → Escape special characters `“,',(,),+,-,[,],$,^,\,|`

Learn more about your commands

- man (manual) → Shows a complete description of a command: usage,options, flags and output.

`man paste`

PASTE(1) User Commands

NAME
paste - merge lines of files

SYNOPSIS
paste [OPTION]... [FILE]...

DESCRIPTION
Write lines consisting of the sequentially corresponding lines from each FILE, separated by TABs, to standard output. With no FILE, or when FILE is -, read standard input.

Mandatory arguments to long options are mandatory for short options too.

-d, --delimiters=LIST
reuse characters from LIST instead of TABs

-s, --serial
paste one file at a time instead of in parallel

--help display this help and exit

--version
output version information and exit

AUTHOR
Written by David M. Ihnat and David MacKenzie.