

UNIX command line tools for manipulation and analysis of genomic data

#AprendeBioinformáticaEnCasa

YOUR HOME 13/04/2020

Juan Antonio García-Martín

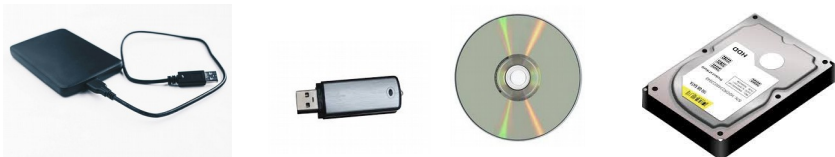
BioinfoGP

Bioinformatics for Genomics and Proteomics



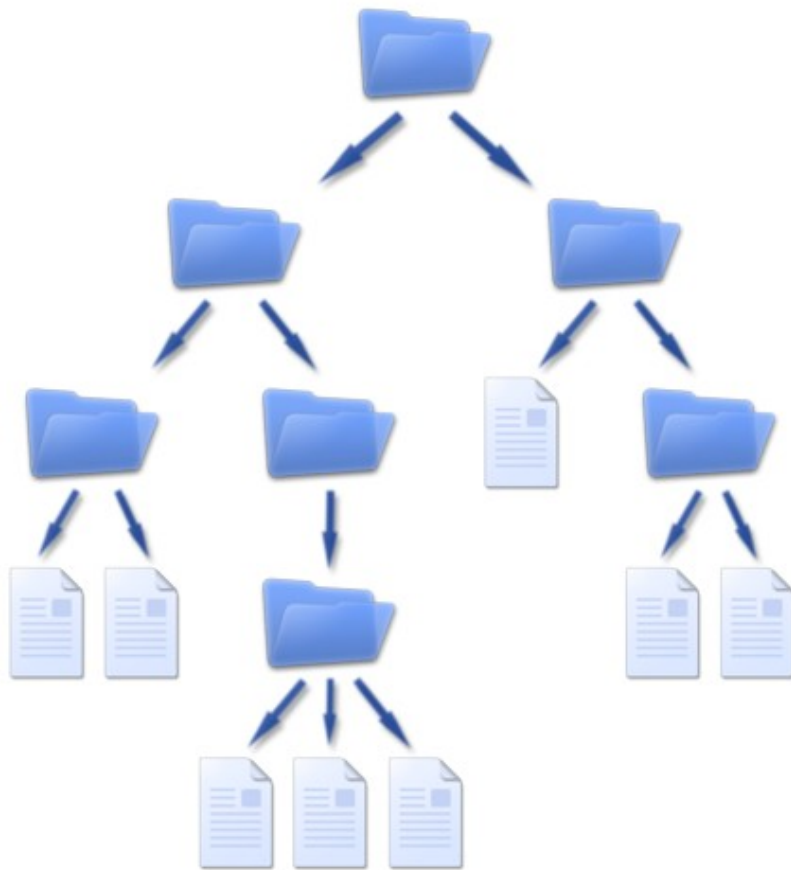
File systems and file paths

File systems



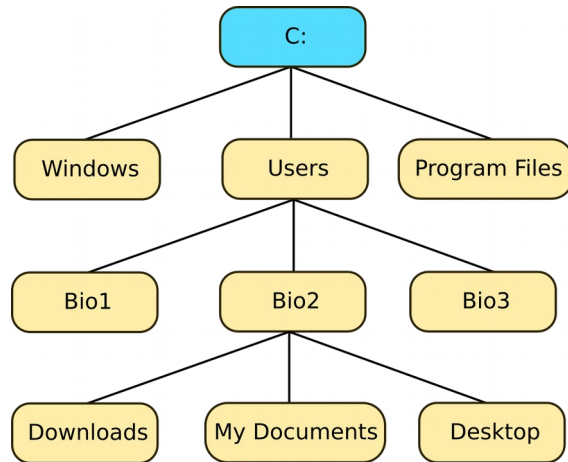
A file system is the way an operative system stores and retrieves data, including:

- Logical drives
- Folders/directories
- Files

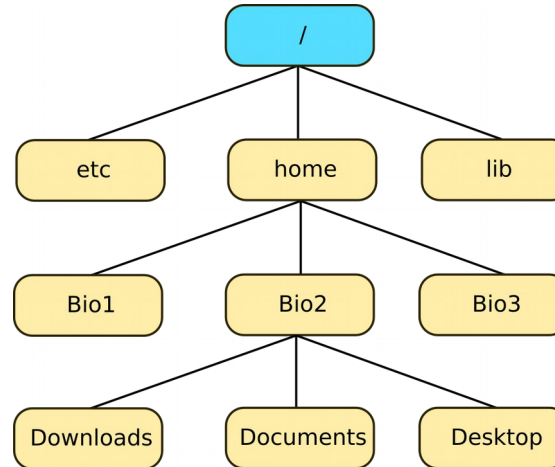


Directory structure by OS

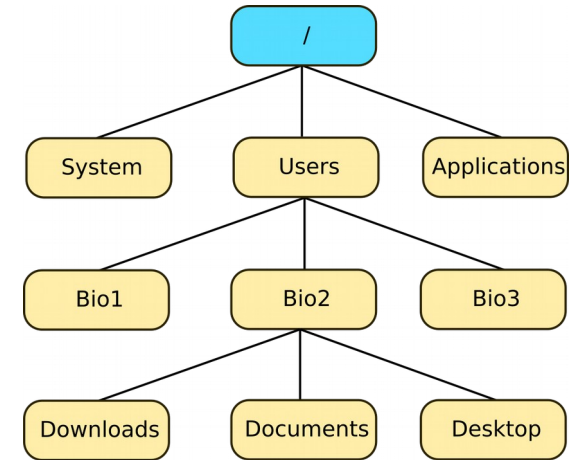
- Directory structure depends on the OS
- The root directory is the lowest level



Windows



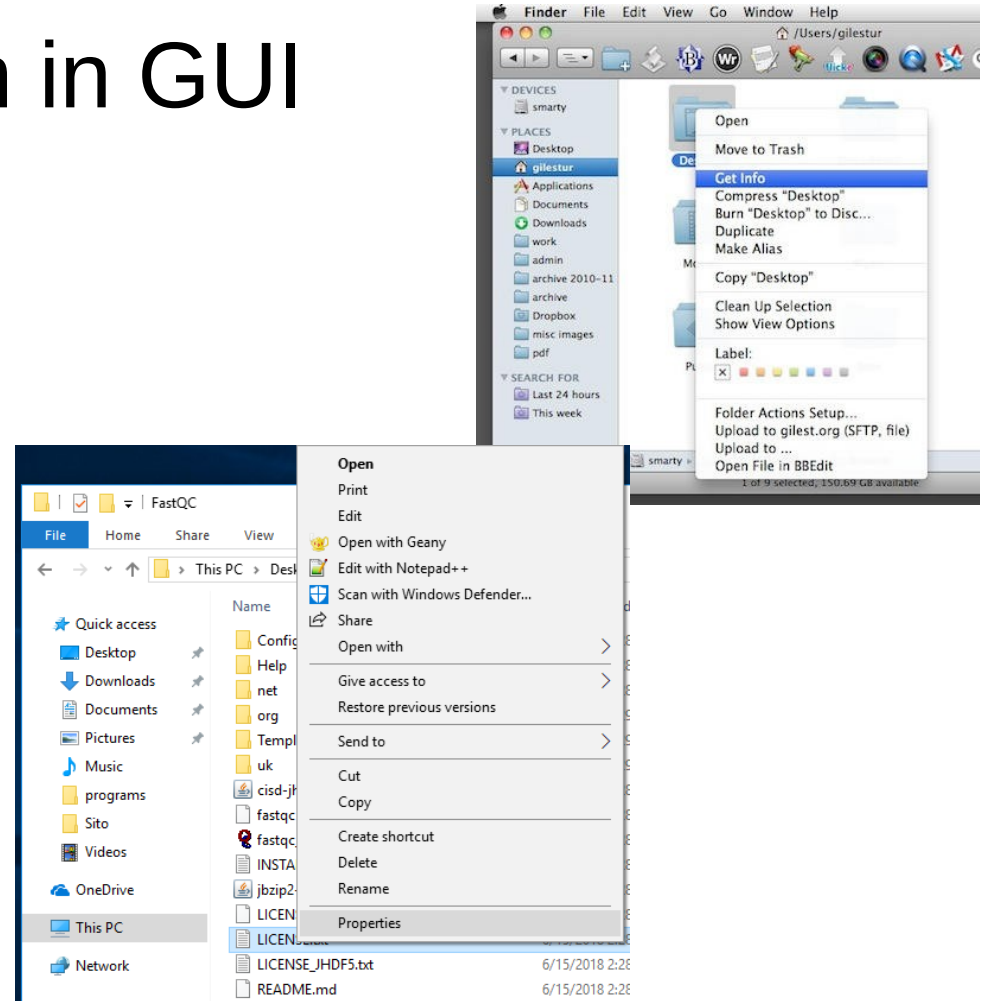
Linux



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.

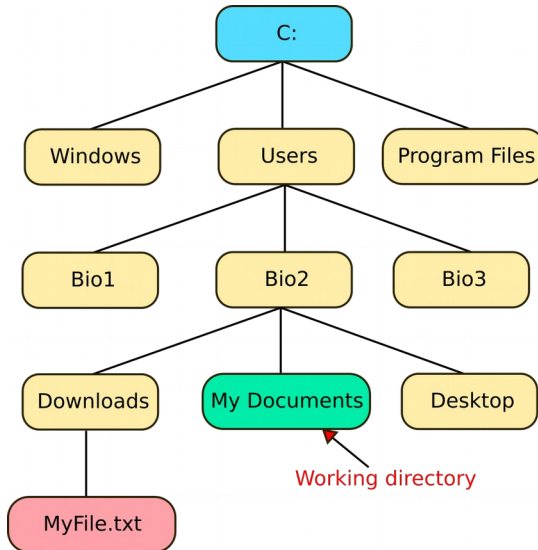
	Separator	Case sensitive
<ul style="list-style-type: none">• Windows<ul style="list-style-type: none">- DriveLetter:\Directory\Filename.extension- e.g.<ul style="list-style-type: none">• C:\Users\Bio2\Downloads\MyFile.txt	\	NO
<ul style="list-style-type: none">• Os X (Mac)<ul style="list-style-type: none">- /Directory/Filename.extension- e.g.<ul style="list-style-type: none">• /Users/Bio2/Downloads/MyFile.txt	/	YES
<ul style="list-style-type: none">• Linux<ul style="list-style-type: none">- /Directory/Filename.extension- e.g.<ul style="list-style-type: none">• /home/Bio2/Downloads/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

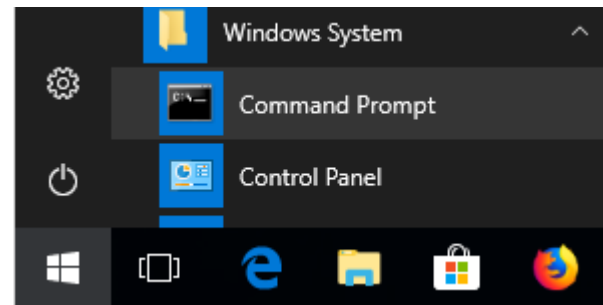
- The working directory is that from where we execute a command or program.
- In a path, a dot . indicates the current directory, and a double dot .. the parent directory.
- Paths can be:
 - Absolute: The full route of the file or directory.
 - Relative: The route of the file or directory relative to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



```
Command Prompt
Microsoft Windows [Version 10.0.17134.112]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\Toño>cd Downloads

C:\Users\Toño\Downloads>dir
Volume in drive C has no label.
Volume Serial Number is 64E2-F881

Directory of C:\Users\Toño\Downloads

06/27/2018  02:49 AM  <DIR>          .
06/27/2018  02:49 AM  <DIR>          ..
11/28/2017  06:45 AM                311,176 Firefox Installer.exe
06/08/2018  12:45 AM            15,377,448 geany-1.33_setup.exe
05/24/2018  08:24 AM  <DIR>          GetGnuWin32
05/24/2018  06:57 AM            3,425,934 GetGnuWin32-0.6.3.exe
11/23/2017  03:09 AM            3,117,648,269 Gorilla_gorilla.gorGor4.dna_rm.toplevel.fa
06/19/2018  01:37 AM  <DIR>          MobaXterm_Portable_v10.6
06/06/2018  05:10 AM            27,266,367 MobaXterm_Portable_v10.6.zip
05/24/2018  06:44 AM  <DIR>          UnxUtils
05/24/2018  06:44 AM            3,365,638 UnxUtils.zip
11/28/2017  08:51 AM            365,069,112 VMware-viclient-all-5.5.0-1281650.exe
              7 File(s)  3,532,463,944 bytes
              5 Dir(s)  19,308,408,832 bytes free

C:\Users\Toño\Downloads>
```

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

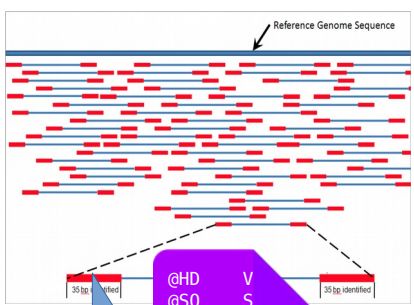
- Images
- Compressed alignments (.bam)
- Executables (.exe)

Biological file types



```
@ SEQNAM
AGCTAGCG
+ SEQNAME
FFBFFIFIFF
@ SEQNAME2
```

FASTQ



```
1 ncbi m
1 ncbi f
1 ncbi gene
1 ncbi mRNA
1 ncbi exon
1 ncbi cds
```

GFF/GTF



```
14 #
15 FASTQ_FILE = vector()
16 ..... SORT DIRECTORIES AND FILES .....
17
18 WORKING_DIR = ""
19 GENOME_FASTA = ""
20
21 FASTQ_FILES = ""
22 FASTQ_FILES[] = ""
23 FASTQ_FILES[] = ""
24 FASTQ_FILES[] = ""
25 FASTQ_FILES[] = ""
26
27 # (add here FASTQ_FILES) Files
28 #
29 #
30 #
31 # Check
32
33 #
34
35 #
36
37 #
38
39 #
40
41 #
42
43 #
44
45 #
46
47 #
48
49 #
50
51 #
52
53 #
54
55 #
56
57 #
58
59 #
60
61 #
62
63 #
64
65 #
66
67 #
68
69 #
70
71 #
72
73 #
74
75 #
76
77 #
78
79 #
80
81 #
82
83 #
84
85 #
86
87 #
88
89 #
90
91 #
92
93 #
94
95 #
96
97 #
98
99 #
100
101 #
102
103 #
104
105 #
106
107 #
108
109 #
110
111 #
112
113 #
114
115 #
116
117 #
118
119 #
120
121 #
122
123 #
124
125 #
126
127 #
128
129 #
130
131 #
132
133 #
134
135 #
136
137 #
138
139 #
140
141 #
142
143 #
144
145 #
146
147 #
148
149 #
150
151 #
152
153 #
154
155 #
156
157 #
158
159 #
160
161 #
162
163 #
164
165 #
166
167 #
168
169 #
170
171 #
172
173 #
174
175 #
176
177 #
178
179 #
180
181 #
182
183 #
184
185 #
186
187 #
188
189 #
190
191 #
192
193 #
194
195 #
196
197 #
198
199 #
200
201 #
202
203 #
204
205 #
206
207 #
208
209 #
210
211 #
212
213 #
214
215 #
216
217 #
218
219 #
220
221 #
222
223 #
224
225 #
226
227 #
228
229 #
230
231 #
232
233 #
234
235 #
236
237 #
238
239 #
240
241 #
242
243 #
244
245 #
246
247 #
248
249 #
250
251 #
252
253 #
254
255 #
256
257 #
258
259 #
260
261 #
262
263 #
264
265 #
266
267 #
268
269 #
270
271 #
272
273 #
274
275 #
276
277 #
278
279 #
280
281 #
282
283 #
284
285 #
286
287 #
288
289 #
290
291 #
292
293 #
294
295 #
296
297 #
298
299 #
300
301 #
302
303 #
304
305 #
306
307 #
308
309 #
310
311 #
312
313 #
314
315 #
316
317 #
318
319 #
320
321 #
322
323 #
324
325 #
326
327 #
328
329 #
330
331 #
332
333 #
334
335 #
336
337 #
338
339 #
340
341 #
342
343 #
344
345 #
346
347 #
348
349 #
350
351 #
352
353 #
354
355 #
356
357 #
358
359 #
360
361 #
362
363 #
364
365 #
366
367 #
368
369 #
370
371 #
372
373 #
374
375 #
376
377 #
378
379 #
380
381 #
382
383 #
384
385 #
386
387 #
388
389 #
390
391 #
392
393 #
394
395 #
396
397 #
398
399 #
400
401 #
402
403 #
404
405 #
406
407 #
408
409 #
410
411 #
412
413 #
414
415 #
416
417 #
418
419 #
420
421 #
422
423 #
424
425 #
426
427 #
428
429 #
430
431 #
432
433 #
434
435 #
436
437 #
438
439 #
440
441 #
442
443 #
444
445 #
446
447 #
448
449 #
450
451 #
452
453 #
454
455 #
456
457 #
458
459 #
460
461 #
462
463 #
464
465 #
466
467 #
468
469 #
470
471 #
472
473 #
474
475 #
476
477 #
478
479 #
480
481 #
482
483 #
484
485 #
486
487 #
488
489 #
490
491 #
492
493 #
494
495 #
496
497 #
498
499 #
500
501 #
502
503 #
504
505 #
506
507 #
508
509 #
510
511 #
512
513 #
514
515 #
516
517 #
518
519 #
520
521 #
522
523 #
524
525 #
526
527 #
528
529 #
530
531 #
532
533 #
534
535 #
536
537 #
538
539 #
540
541 #
542
543 #
544
545 #
546
547 #
548
549 #
550
551 #
552
553 #
554
555 #
556
557 #
558
559 #
560
561 #
562
563 #
564
565 #
566
567 #
568
569 #
570
571 #
572
573 #
574
575 #
576
577 #
578
579 #
580
581 #
582
583 #
584
585 #
586
587 #
588
589 #
590
591 #
592
593 #
594
595 #
596
597 #
598
599 #
600
601 #
602
603 #
604
605 #
606
607 #
608
609 #
610
611 #
612
613 #
614
615 #
616
617 #
618
619 #
620
621 #
622
623 #
624
625 #
626
627 #
628
629 #
630
631 #
632
633 #
634
635 #
636
637 #
638
639 #
640
641 #
642
643 #
644
645 #
646
647 #
648
649 #
650
651 #
652
653 #
654
655 #
656
657 #
658
659 #
660
661 #
662
663 #
664
665 #
666
667 #
668
669 #
670
671 #
672
673 #
674
675 #
676
677 #
678
679 #
680
681 #
682
683 #
684
685 #
686
687 #
688
689 #
690
691 #
692
693 #
694
695 #
696
697 #
698
699 #
700
701 #
702
703 #
704
705 #
706
707 #
708
709 #
710
711 #
712
713 #
714
715 #
716
717 #
718
719 #
720
721 #
722
723 #
724
725 #
726
727 #
728
729 #
730
731 #
732
733 #
734
735 #
736
737 #
738
739 #
740
741 #
742
743 #
744
745 #
746
747 #
748
749 #
750
751 #
752
753 #
754
755 #
756
757 #
758
759 #
760
761 #
762
763 #
764
765 #
766
767 #
768
769 #
770
771 #
772
773 #
774
775 #
776
777 #
778
779 #
780
781 #
782
783 #
784
785 #
786
787 #
788
789 #
790
791 #
792
793 #
794
795 #
796
797 #
798
799 #
800
801 #
802
803 #
804
805 #
806
807 #
808
809 #
810
811 #
812
813 #
814
815 #
816
817 #
818
819 #
820
821 #
822
823 #
824
825 #
826
827 #
828
829 #
830
831 #
832
833 #
834
835 #
836
837 #
838
839 #
840
841 #
842
843 #
844
845 #
846
847 #
848
849 #
850
851 #
852
853 #
854
855 #
856
857 #
858
859 #
860
861 #
862
863 #
864
865 #
866
867 #
868
869 #
870
871 #
872
873 #
874
875 #
876
877 #
878
879 #
880
881 #
882
883 #
884
885 #
886
887 #
888
889 #
890
891 #
892
893 #
894
895 #
896
897 #
898
899 #
900
901 #
902
903 #
904
905 #
906
907 #
908
909 #
910
911 #
912
913 #
914
915 #
916
917 #
918
919 #
920
921 #
922
923 #
924
925 #
926
927 #
928
929 #
930
931 #
932
933 #
934
935 #
936
937 #
938
939 #
940
941 #
942
943 #
944
945 #
946
947 #
948
949 #
950
951 #
952
953 #
954
955 #
956
957 #
958
959 #
960
961 #
962
963 #
964
965 #
966
967 #
968
969 #
970
971 #
972
973 #
974
975 #
976
977 #
978
979 #
980
981 #
982
983 #
984
985 #
986
987 #
988
989 #
990
991 #
992
993 #
994
995 #
996
997 #
998
999 #
1000
1001 #
1002
1003 #
1004
1005 #
1006
1007 #
1008
1009 #
1010
1011 #
1012
1013 #
1014
1015 #
1016
1017 #
1018
1019 #
1020
1021 #
1022
1023 #
1024
1025 #
1026
1027 #
1028
1029 #
1030
1031 #
1032
1033 #
1034
1035 #
1036
1037 #
1038
1039 #
1040
1041 #
1042
1043 #
1044
1045 #
1046
1047 #
1048
1049 #
1050
1051 #
1052
1053 #
1054
1055 #
1056
1057 #
1058
1059 #
1060
1061 #
1062
1063 #
1064
1065 #
1066
1067 #
1068
1069 #
1070
1071 #
1072
1073 #
1074
1075 #
1076
1077 #
1078
1079 #
1080
1081 #
1082
1083 #
1084
1085 #
1086
1087 #
1088
1089 #
1090
1091 #
1092
1093 #
1094
1095 #
1096
1097 #
1098
1099 #
1100
1101 #
1102
1103 #
1104
1105 #
1106
1107 #
1108
1109 #
1110
1111 #
1112
1113 #
1114
1115 #
1116
1117 #
1118
1119 #
1120
1121 #
1122
1123 #
1124
1125 #
1126
1127 #
1128
1129 #
1130
1131 #
1132
1133 #
1134
1135 #
1136
1137 #
1138
1139 #
1140
1141 #
1142
1143 #
1144
1145 #
1146
1147 #
1148
1149 #
1150
1151 #
1152
1153 #
1154
1155 #
1156
1157 #
1158
1159 #
1160
1161 #
1162
1163 #
1164
1165 #
1166
1167 #
1168
1169 #
1170
1171 #
1172
1173 #
1174
1175 #
1176
1177 #
1178
1179 #
1180
1181 #
1182
1183 #
1184
1185 #
1186
1187 #
1188
1189 #
1190
1191 #
1192
1193 #
1194
1195 #
1196
1197 #
1198
1199 #
1200
1201 #
1202
1203 #
1204
1205 #
1206
1207 #
1208
1209 #
1210
1211 #
1212
1213 #
1214
1215 #
1216
1217 #
1218
1219 #
1220
1221 #
1222
1223 #
1224
1225 #
1226
1227 #
1228
1229 #
1230
1231 #
1232
1233 #
1234
1235 #
1236
1237 #
1238
1239 #
1240
1241 #
1242
1243 #
1244
1245 #
1246
1247 #
1248
1249 #
1250
1251 #
1252
1253 #
1254
1255 #
1256
1257 #
1258
1259 #
1260
1261 #
1262
1263 #
1264
1265 #
1266
1267 #
1268
1269 #
1270
1271 #
1272
1273 #
1274
1275 #
1276
1277 #
1278
1279 #
1280
1281 #
1282
1283 #
1284
1285 #
1286
1287 #
1288
1289 #
1290
1291 #
1292
1293 #
1294
1295 #
1296
1297 #
1298
1299 #
1300
1301 #
1302
1303 #
1304
1305 #
1306
1307 #
1308
1309 #
1310
1311 #
1312
1313 #
1314
1315 #
1316
1317 #
1318
1319 #
1320
1321 #
1322
1323 #
1324
1325 #
1326
1327 #
1328
1329 #
1330
1331 #
1332
1333 #
1334
1335 #
1336
1337 #
1338
1339 #
1340
1341 #
1342
1343 #
1344
1345 #
1346
1347 #
1348
1349 #
1350
1351 #
1352
1353 #
1354
1355 #
1356
1357 #
1358
1359 #
1360
1361 #
1362
1363 #
1364
1365 #
1366
1367 #
1368
1369 #
1370
1371 #
1372
1373 #
1374
1375 #
1376
1377 #
1378
1379 #
1380
1381 #
1382
1383 #
1384
1385 #
1386
1387 #
1388
1389 #
1390
1391 #
1392
1393 #
1394
1395 #
1396
1397 #
1398
1399 #
1400
1401 #
1402
1403 #
1404
1405 #
1406
1407 #
1408
1409 #
1410
1411 #
1412
1413 #
1414
1415 #
1416
1417 #
1418
1419 #
1420
1421 #
1422
1423 #
1424
1425 #
1426
1427 #
1428
1429 #
1430
1431 #
1432
1433 #
1434
1435 #
1436
1437 #
1438
1439 #
1440
1441 #
1442
1443 #
1444
1445 #
1446
1447 #
1448
1449 #
1450
1451 #
1452
1453 #
1454
1455 #
1456
1457 #
1458
1459 #
1460
1461 #
1462
1463 #
1464
1465 #
1466
1467 #
1468
1469 #
1470
1471 #
1472
1473 #
1474
1475 #
1476
1477 #
1478
1479 #
1480
1481 #
1482
1483 #
1484
1485 #
1486
1487 #
1488
1489 #
1490
1491 #
1492
1493 #
1494
1495 #
1496
1497 #
1498
1499 #
1500
1501 #
1502
1503 #
1504
1505 #
1506
1507 #
1508
1509 #
1509
```

```
function
for(i=1
myVar=f(i)
set(myVar)
}
```

SOURCE

```
@HD V
@SQ S
@RG PL:ILLUM
1:497:R:-272
1:497:R:-272
1:497:R:-272
```

SAM/BAM

```
function
for(i=1
myVar=f(i)
set(myVar)
}
```

SOURCE

```
#CHROM PO
20 14 . G A
20 173 . T A
20 196 . A G
```

VCF/BCF

```
> 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

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

BED



(Optional)

Sequence files (FASTA)

```
> SEQ
AGCTAGCG
CGTACAGCGT
GCTACTGCGA
GCTATCGCAT
```

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';;
CGACACGCGTCCGCTTTGTGACTTCACCATGGCGTACCGCGGCCAGGGCCAAAAGGTGCA
GAAGGTGATGGTGCAGCCCATCAACCTTATCTTCAGATACTTGCAAAATAGATCTCGAAT
TCAGGTGTGGCTGTATGAACAAGTGAATATGCGGATAGAAGGTTGTATTATTGGCTTTGA
TGAGTACATGAACCTCGTATTAGATGATGCAGAAGAGATTCATTCTAAAACAAAGTCAAG
AAAACAAC TGGGTCGGATCATGCTAAAAGGAGATAATATTACTCTGCTCCAAAGTGTTC
CAACTAGCAGTGAATGGTGAAGTCTGTAGGATGTTGAGAAGACCCCTTGAGCGTGTTTAA
AGATGTCTGCTGCAACCTGCATTTACTCAACTTGTTTTACTTGCACATTATTATTAGGTG
ACAATAAATGCTGTAGGAAGTTTTTGT
```

Header starts with a right angle >

No empty lines between header and sequence.

Sequences can be in one or multiple lines

```
>ENA|AL022645|AL022645.1 Mus musculus cDNA clone 528-1F10 5';;
CGACACGCGTCCGCTTTGTGACTTCACCATGGCGTACCGCGGCCAGGGCCAAAAGGTGCA
GAAGGTGATGGTGCAGCCCATCAACCTTATCTTCAGATACTTGCAAAATAGATCTCGAAT
TCAGGTGTGGCTGTATGAACAAGTGAATATGCGGATAGAAGGTTGTATTATTGGCTTTGA
TGAGTACATGAACCTCGTATTAGATGATGCAGAAGAGATTCATTCTAAAACAAAGTCAAG
AAAACAAC TGGGTCGGATCATGCTAAAAGGAGATAATATTACTCTGCTCCAAAGTGTTC
CAACTAGCAGTGAATGGTGAAGTCTGTAGGATGTTGAGAAGACCCCTTGAGCGTGTTTAA
AGATGTCTGCTGCAACCTGCATTTACTCAACTTGTTTTACTTGCACATTATTATTAGGTG
ACAATAAATGCTGTAGGAAGTTTTTGT
```

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
CGTGTCAAAATTTGCTTGTACACTCTGCCAAGGCATCATTCCTAGCCTTT
+
BBBFFFFFFFFFFFFFFFFIIIIIIFFIIIIIIFFIIIIIIFFIFFBFFIFIFFFFFFFIF
+SRR3214175.20 HISEQ01:422:HABGFADXX:1:1101:2196:2034/1
GGGCTGGCACGGAGCAGAGTGATTTCGTAGAGCAGAGGGTCAGCTCCCTGG
+
BBBFFFFFFFFFFFFFFFFIIIIIIFFIIIIIIFFIIIIIIIBFFFFFFFFIIFII
+SRR3214175.30 HISEQ01:422:HABGFADXX:1:1101:2768:2070/1
CTTATTTGCTGAGAACAGTAAGATCTGCAGGCAGTTTCAGATGAAACAGGC
+
BBBFFFFFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIFIIFF
+SRR3214175.40 HISEQ01:422:HABGFADXX:1:1101:3177:2234/1
AATCATTTGAGAGGTTTTGCCATAATTATTACATGAGGCATAGAAAGTTTG
+
BBBFFFFFFFFFFFFI<FFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
+SRR3214175.50 HISEQ01:422:HABGFADXX:1:1101:3661:2200/1
CAAAAGTATGTCCTTCTAGTTCCGTGTGGCCTTTGTCAGCATGGACAGCAT
+
BBBFFFFFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIFII
```

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.

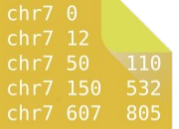
Basic formats: Tables



- 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 origin 1049-1668	http://wishart.biology.ualberta.ca/cgview/
forward	1	2657	2963	origin_of_replication	f1	f1 origin 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)



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

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

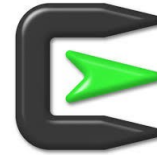


vmware®

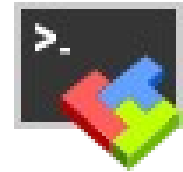


VirtualBox

Unix environment emulation



CygWin

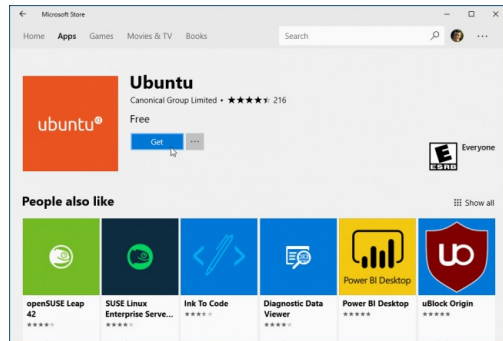


MobaXterm



MinGW + MSSYS

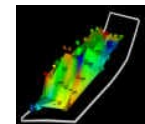
Linux subsystem for Windows



Unix ports for Windows

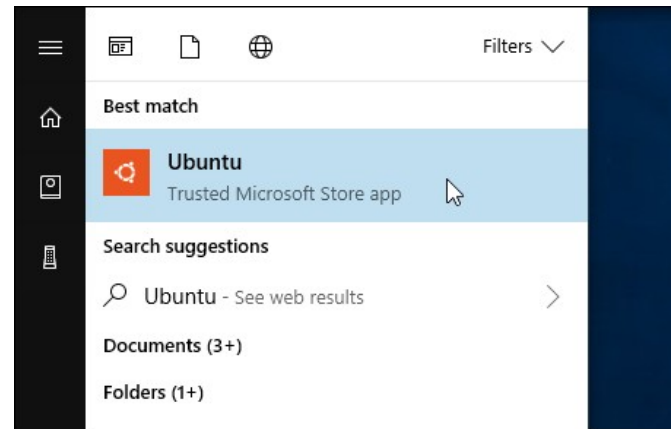
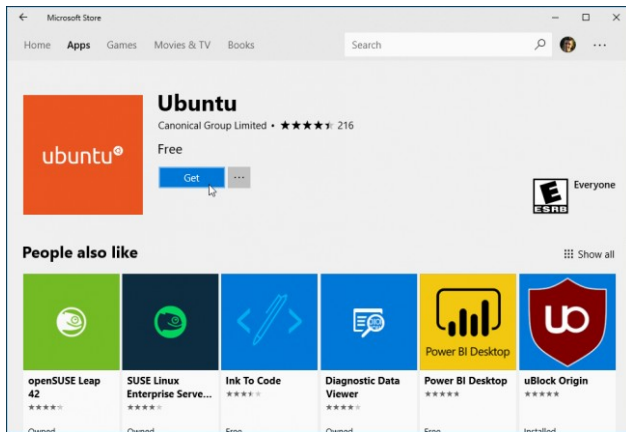
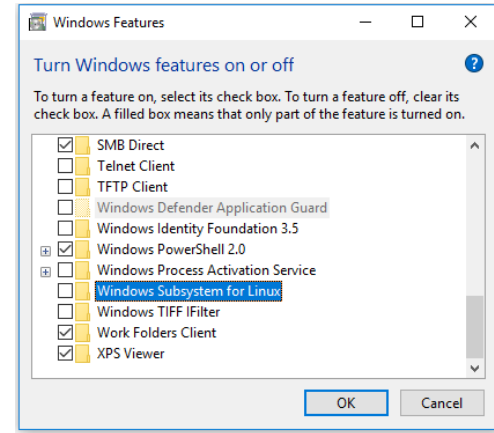
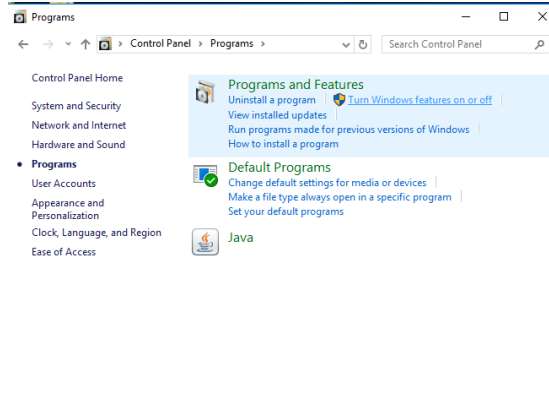
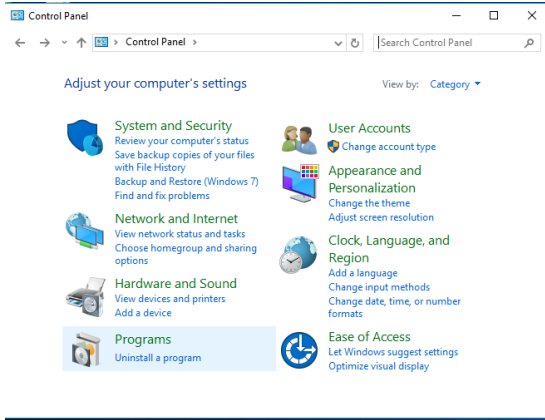
GnuWin

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 compressed 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 adjacent, often used in combination of `sort`
 - `-c` → prefix lines by 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 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.

```
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.
```

`man paste`