Table of Contents

Introduction to Unix - awk and Makefiles

Welcome to the Programming for Evolutionary Biology workshop!!

Giovanni M. Dall'Olio and Alvaro Perdomo-Sabogal, 03/03/2019. All materials available here: https://github.com/dalloliogm/evop2019/archive/master.zip (https://github.com/dalloliogm/evop2019/archive/master.zip)

In this fourth part we will use the **awk** command to excplore the contents of files, and some basic regular expressions.

Press space or down key to continue.

Working with tabular files: Awk

The **awk** command allows to search and manipulate tabular files from the command line.

Imagine it as the equivalent of Excel/Calc for the command line. It allows to do search on specific columns of a file, to do numerical operations, or to change the order of the columns.

The advantage of a command-line tool over graphical software is that the memory footprint is much lower. So you can access and modify large files in a fraction of the time that it would take with Excel.

Example of tabular file: the GFF3 format

The file genes/chr8.gff contains an example of file in the GFF3 format:

In [14]:

```
head genes/chr8.gff
```

```
##gff-version 3
##source-version refgene 1.28.10
##date 2016-09-08
##genome-build .        hg19
chr8    refgene gene    18248755        18258723        .       +
.       gene_id=10;symbol=NAT2;;ID=10
chr8    refgene gene    100549014       100549089       .       -
.       gene_id=100126309;symbol=MIR875;;ID=100126309
chr8    refgene gene    144895127       144895212       .       -
.       gene_id=100126338;symbol=MIR937;;ID=100126338
chr8    refgene gene    145619364       145619445       .       -
.       gene_id=100126351;symbol=MIR939;;ID=100126351
chr8    refgene gene    91970706        91997485        .       -
.       gene_id=100127983;symbol=C8orf88;;ID=100127983
chr8    refgene gene    74332309        74353753        .       +
.       gene_id=100128126;symbol=STAU2-AS1;;ID=100128126
```

As you can see it is a tab-separated file, which we could easily read in Excel or Calc.

The **GFF** (General Feature Format) format specifications are defined here (https://genome.ucsc.edu/FAQ/FAQformat.html#format3), but in short:

- the **col1**, **col4** and **col5** contain the chromosome name and genomic coordinates (start and end),
- the **col2** describes the tool or resource that generated the annotation,
- the **col3** describe the type of feature (e.g. gene, transcript, exon, TF binding site, Histone Acetylation mark, etc...
- the **col9** column contains several fields, separated by a semicolon

Basic AWK syntax: filters

The basic AWK syntax is the following:

** awk 'filters {print statements}' filename**

Awk is quite smart at recognizing the field separator, and by default assumes they are separated by tabs.

Each column of the file can be referred to with the dollar sign followed by the number of column.

For example $2 refers to the second column, and so on.

The following code filters all the lines belonging to chromosome 8, between the coordinates 100000 and 200000:

In [15]:

```
awk '$1=="chr8" && $4>100000 && $5<200000 ' genes/chr8.gff
```

```
chr8    refgene gene    182200  197339  .       +       .       gene_i
d=169270;symbol=ZNF596;;ID=169270
chr8    refgene gene    116086  117024  .       -       .       gene_i
d=441308;symbol=OR4F21;;ID=441308
chr8    refgene gene    158345  182318  .       -       .       gene_i
d=644128;symbol=RPL23AP53;;ID=644128
```

Exercise

Can you print all the lines between 5000000 and 10000000 ?

In [16]:

```
awk '$4 > 5000000 && $5 < 10000000 ' genes/chr8.gff
```

```
chr8    refgene gene    7143733 7212876 .       -       .       gene_i
d=100128890;symbol=FAM66B;ID=100128890
chr8    refgene gene    7215498 7220490 .       -       .       gene_i
d=100131980;symbol=ZNF705G;ID=100131980
chr8    refgene gene    7812535 7866277 .       +       .       gene_i
d=100132103;symbol=FAM66E;ID=100132103
chr8    refgene gene    7783859 7809935 .       +       .
_________
chr8    refgene gene    6261077 6264069 .       -       .
/ Cows in \
chr8    refgene gene    7272385 7274354 .       -       .
| the     |
chr8    refgene gene    7946463 7946611 .       -       .
\ Genome! /
chr8    refgene gene    6602685 6602765 .       +       .
---------
chr8    refgene gene    8905955 8906028 .       +       .
\   ^__^
chr8    refgene gene    6602689 6602761 .       -       .
\  (oo)\_______
chr8    refgene gene    6693076 6699975 .       +       .
(__)\       )\/\
chr8    refgene gene    8559666 8561617 .       +       .
||----w |
chr8    refgene gene    9182561 9192590 .       +       .
||      |
chr8    refgene gene    8175258 8239257 .       -       .       gene_i
d=157285;symbol=SGK223;ID=157285
chr8    refgene gene    9757574 9760839 .       -       .       gene_i
d=157627;symbol=LINC00599;ID=157627
chr8    refgene gene    6835171 6856724 .       -       .       gene_i
d=1667;symbol=DEFA1;ID=1667
chr8    refgene gene    6793345 6795786 .       -       .       gene_i
d=1669;symbol=DEFA4;ID=1669
chr8    refgene gene    6912829 6914259 .       -       .       gene_i
d=1670;symbol=DEFA5;ID=1670
chr8    refgene gene    6782216 6783598 .       -       .       gene_i
d=1671;symbol=DEFA6;ID=1671
chr8    refgene gene    6728097 6735529 .       -       .       gene_i
d=1672;symbol=DEFB1;ID=1672
chr8    refgene gene    7752199 7754237 .       +       .       gene_i
d=1673;symbol=DEFB4A;ID=1673
chr8    refgene gene    6844700 6866346 .       -       .       gene_i
d=170949;symbol=DEFT1P;ID=170949
chr8    refgene gene    7353368 7366833 .       +       .       gene_i
d=245910;symbol=DEFB107A;ID=245910
chr8    refgene gene    6357175 6420784 .       -       .       gene_i
d=285;symbol=ANGPT2;ID=285
chr8    refgene gene    8086092 8102387 .       +       .       gene_i
d=286042;symbol=FAM86B3P;ID=286042
chr8    refgene gene    6666041 6693166 .       -       .       gene_i
d=389610;symbol=XKR5;ID=389610
chr8    refgene gene    7829183 7830775 .       -       .       gene_i
d=392188;symbol=USP17L8;ID=392188
chr8    refgene gene    7189909 7191501 .       +       .       gene_i
d=401447;symbol=USP17L1;ID=401447
```

```
chr8    refgene gene    9760898 9760982 .       -       .       gene_i
d=406907;symbol=MIR124-1;ID=406907
chr8    refgene gene    7413660 7431920 .       -       .       gene_i
d=441317;symbol=FAM90A7P;ID=441317
chr8    refgene gene    7627106 7628835 .       +       .       gene_i
d=441328;symbol=FAM90A10P;ID=441328
chr8    refgene gene    6808248 6809121 .       -       .       gene_i
d=449491;symbol=DEFA8P;ID=449491
chr8    refgene gene    6816811 6817683 .       -       .       gene_i
d=449492;symbol=DEFA9P;ID=449492
chr8    refgene gene    6825663 6826635 .       -       .       gene_i
d=449493;symbol=DEFA10P;ID=449493
chr8    refgene gene    7669242 7673238 .       -       .       gene_i
d=503614;symbol=DEFB107B;ID=503614
chr8    refgene gene    6565878 6619021 .       +       .       gene_i
d=55326;symbol=AGPAT5;ID=55326
chr8    refgene gene    7194637 7196229 .       +       .       gene_i
d=645402;symbol=USP17L4;ID=645402
chr8    refgene gene    7833915 7835507 .       -       .       gene_i
d=645836;symbol=USP17L3;ID=645836
chr8    refgene gene    7705402 7721319 .       +       .       gene_i
d=653423;symbol=SPAG11A;ID=653423
chr8    refgene gene    9599182 9599278 .       +       .       gene_i
d=693182;symbol=MIR597;ID=693182
chr8    refgene gene    6886123 6887011 .       -       .       gene_i
d=724068;symbol=DEFA11P;ID=724068
chr8    refgene gene    6873391 6875823 .       -       .       gene_i
d=728358;symbol=DEFA1B;ID=728358
chr8    refgene gene    6264113 6501140 .       +       .       gene_i
d=79648;symbol=MCPH1;ID=79648
chr8    refgene gene    8993764 9009152 .       -       .       gene_i
d=79660;symbol=PPP1R3B;ID=79660
chr8    refgene gene    9413445 9639856 .       +       .       gene_i
d=8658;symbol=TNKS;ID=8658
chr8    refgene gene    8860314 8890849 .       +       .       gene_i
d=90459;symbol=ERI1;ID=90459
chr8    refgene gene    8641999 8751131 .       -       .       gene_i
d=9258;symbol=MFHAS1;ID=9258
```

Awk: printing columns and doing operations

Awk also allows to print only specific columns, and do algebraic operations on them.

Remember that each column can be referred as 1,2, $3, etc...

For example the following code prints the first column, and the sum of the fourth and third. We can pipe the output to head or less, to make it easier to visualize:

In [17]:

```
awk '{print $1, $5-$4}' genes/chr8.gff | head
```

```
##gff-version 0
##source-version 0
##date 0
##genome-build 0
chr8 9968
chr8 75
chr8 85
chr8 81
chr8 26779
chr8 21444
```

Notice how this also prints the headers of the file. We can exclude these by adding a grep condition:

In [18]:

```
awk '{print $1, $5-$4, $9}' genes/chr8.gff | grep -v '^#' |  head
```

```
chr8 9968 gene_id=10;symbol=NAT2;;ID=10
chr8 75 gene_id=100126309;symbol=MIR875;;ID=100126309
chr8 85 gene_id=100126338;symbol=MIR937;;ID=100126338
chr8 81 gene_id=100126351;symbol=MIR939;;ID=100126351
chr8 26779 gene_id=100127983;symbol=C8orf88;;ID=100127983
chr8 21444 gene_id=100128126;symbol=STAU2-AS1;;ID=100128126
chr8 12197 gene_id=100128338;symbol=FAM83H-AS1;;ID=100128338
chr8 1835 gene_id=100128627;symbol=CDC42P3;;ID=100128627
chr8 3282 gene_id=100128750;symbol=RBPMS-AS1;;ID=100128750
chr8 69143 gene_id=100128890;symbol=FAM66B;ID=100128890
grep: write error: Broken pipe
```

Exercise (difficult)

Starting from the previous command, can you extract the gene symbol into a separate column?

Hints: pipe an additional awk statement after the first. Use the -F option to specify a different field separator.

In [19]:

```
awk '{print $1, $5-$4, $9}' genes/chr8.gff | grep -v '^#' | awk -F';' '{print $1, $2
```

```
chr8 9968 gene_id=10 symbol=NAT2
chr8 75 gene_id=100126309 symbol=MIR875
chr8 85 gene_id=100126338 symbol=MIR937
chr8 81 gene_id=100126351 symbol=MIR939
chr8 26779 gene_id=100127983 symbol=C8orf88
chr8 21444 gene_id=100128126 symbol=STAU2-AS1
chr8 12197 gene_id=100128338 symbol=FAM83H-AS1
chr8 1835 gene_id=100128627 symbol=CDC42P3
chr8 3282 gene_id=100128750 symbol=RBPMS-AS1
chr8 69143 gene_id=100128890 symbol=FAM66B
```

AWK: searching by regular expressions

Awk can also be used to search by regular expression.

For example, the following code will print all the lines in which the symbol starts with "MIR":

In [20]:

```
awk '$9 ~ /symbol=MIR/ {print $0}' genes/chr8.gff
```

```
chr8    refgene gene    100549014       100549089       .       -
.       gene_id=100126309;symbol=MIR875;;ID=100126309
chr8    refgene gene    144895127       144895212       .       -
.       gene_id=100126338;symbol=MIR937;;ID=100126338
chr8    refgene gene    145619364       145619445       .       -
.       gene_id=100126351;symbol=MIR939;;ID=100126351
chr8    refgene gene    65285775        65295842        .       +
.       gene_id=100130155;symbol=MIR124-2HG;;ID=100130155
chr8    refgene gene    128972879       128972941       .       +
.       gene_id=100302161;symbol=MIR1205;;ID=100302161
chr8    refgene gene    10682883        10682953        .       -
.       gene_id=100302166;symbol=MIR1322;;ID=100302166
chr8    refgene gene    129021144       129021202       .       +
.       gene_id=100302170;symbol=MIR1206;;ID=100302170
chr8    refgene gene    129061398       129061484       .       +
.       gene_id=100302175;symbol=MIR1207;;ID=100302175
chr8    refgene gene    128808208       128808274       .       +
.       gene_id=100302185;symbol=MIR1204;;ID=100302185
chr8    refgene gene    145625476       145625559       .       -
.       gene_id=100302196;symbol=MIR1234;;ID=100302196
chr8    refgene gene    113655722       113655812       .       +
.       gene_id=100302225;symbol=MIR2053;;ID=100302225
chr8    refgene gene    27743556        27743633        .       -
.       gene_id=100422828;symbol=MIR4287;;ID=100422828
chr8    refgene gene    29814788        29814864        .       -
.       gene_id=100422876;symbol=MIR3148;;ID=100422876
chr8    refgene gene    28362633        28362699        .       -
.       gene_id=100422903;symbol=MIR4288;;ID=100422903
chr8    refgene gene    96085142        96085221        .       +
.       gene_id=100422964;symbol=MIR3150A;;ID=100422964
chr8    refgene gene    104166842       104166917       .       +
.       gene_id=100422992;symbol=MIR3151;;ID=100422992
chr8    refgene gene    12584746        12584808        .       +
.       gene_id=100500838;symbol=MIR3926-2;;ID=100500838
chr8    refgene gene    27559194        27559276        .       +
.       gene_id=100500858;symbol=MIR3622A;;ID=100500858
chr8    refgene gene    12584741        12584813        .       -
.       gene_id=100500870;symbol=MIR3926-1;;ID=100500870
chr8    refgene gene    27559190        27559284        .       -
.       gene_id=100500871;symbol=MIR3622B;;ID=100500871
chr8    refgene gene    96085139        96085224        .       -
.       gene_id=100500907;symbol=MIR3150B;;ID=100500907
chr8    refgene gene    117886967       117887039       .       -
.       gene_id=100500914;symbol=MIR3610;;ID=100500914
chr8    refgene gene    42751340        42751418        .       -
.       gene_id=100616115;symbol=MIR4469;;ID=100616115
chr8    refgene gene    94928250        94928347        .       -
.       gene_id=100616169;symbol=MIR378D2;;ID=100616169
chr8    refgene gene    29920258        30108213        .       -
.       gene_id=100616190;symbol=MIR548O2;;ID=100616190
chr8    refgene gene    92217713        92217786        .       +
.       gene_id=100616245;symbol=MIR4661;;ID=100616245
chr8    refgene gene    124228028       124228103       .       -
.       gene_id=100616260;symbol=MIR4663;;ID=100616260
chr8    refgene gene    143257700       143257779       .       +
.       gene_id=100616268;symbol=MIR4472-1;;ID=100616268
chr8    refgene gene    144815253       144815323       .       -
```

```
.       gene_id=100616318;symbol=MIR4664;;ID=100616318
chr8    refgene gene    101394991       101395073       .       +
.       gene_id=100616451;symbol=MIR4471;;ID=100616451
chr8    refgene gene    62627347        62627418        .       +
.       gene_id=100616484;symbol=MIR4470;;ID=100616484
chr8    refgene gene    103137660       103137743       .       +
.       gene_id=100847001;symbol=MIR5680;;ID=100847001
chr8    refgene gene    131020580       131020699       .       -
.       gene_id=100847051;symbol=MIR5194;;ID=100847051
chr8    refgene gene    81153624        81153708        .       +
.       gene_id=100847056;symbol=MIR5708;;ID=100847056
chr8    refgene gene    75460778        75460852        .       +
.       gene_id=100847058;symbol=MIR5681A;;ID=100847058
chr8    refgene gene    75460785        75460844        .       -
.       gene_id=100847091;symbol=MIR5681B;;ID=100847091
chr8    refgene gene    9760898 9760982 .       -       .       gene_i
d=406907;symbol=MIR124-1;ID=406907
chr8    refgene gene    65291706        65291814        .       +
.       gene_id=406908;symbol=MIR124-2;;ID=406908
chr8    refgene gene    135812763       135812850       .       -
.       gene_id=407030;symbol=MIR30B;;ID=407030
chr8    refgene gene    135817119       135817188       .       -
.       gene_id=407033;symbol=MIR30D;;ID=407033
chr8    refgene gene    22102475        22102556        .       -
.       gene_id=407037;symbol=MIR320A;;ID=407037
chr8    refgene gene    75512101        75670587        .       +
.       gene_id=441355;symbol=MIR2052HG;;ID=441355
chr8    refgene gene    14710947        14711019        .       -
.       gene_id=494332;symbol=MIR383;;ID=494332
chr8    refgene gene    41517959        41518026        .       -
.       gene_id=619554;symbol=MIR486-1;;ID=619554
chr8    refgene gene    1765397 1765473 .       +       .       gene_i
d=693181;symbol=MIR596;;ID=693181
chr8    refgene gene    9599182 9599278 .       +       .       gene_i
d=693182;symbol=MIR597;ID=693182
chr8    refgene gene    10892716        10892812        .       -
.       gene_id=693183;symbol=MIR598;;ID=693183
chr8    refgene gene    100548864       100548958       .       -
.       gene_id=693184;symbol=MIR599;;ID=693184
chr8    refgene gene    145019359       145019447       .       -
.       gene_id=724031;symbol=MIR661;;ID=724031
```

Last exercise!

Calculate the lenght of the gene POU5F1B.

Find the Gene whose gene_id is equal to that number.

In [21]:

```
awk '$9 ~ /POU5F1B/ {print $5-$4}' genes/chr8.gff
```

```
1584
```

In [22]:

```
awk '$9 ~ /gene_id=1584/ {print $0}' genes/chr8.gff
```

```
chr8    refgene Good_Job!       143953773       143961236       .
-       .       gene_id=1584;symbol=CYP11B1;;ID=1584
```

Bonus: Makefiles

Let's have a look at the file called Makefile in the unix_intro directory:

In [24]:

```
cd ..
head Makefile
```

```
# This is a Makefile, which will be explained later in the course.
# Please don't look at it yet :-)

publish: slides_bash commit
	echo "convert the slides to pdf, commit, and push to github"
	git push


test_exercises: start help ignorecase multiplefiles
generate_exercises: generate_grep generate_awk
```

Press space or the down key to continue

Defining pipelines with Makefiles

Makefiles are a basic way to define pipelines of shell commands.

Nowadays there are more sophisticated tools available, but most of these are based on Makefiles.

A Makefile is a collection of "rules".

Each of these rules follows this basic syntax is:

```
target: prerequisites
    commands to execute
```

As you can see in the Makefile included, most of the rules allow to regenerate the exercise files, or to execute some commands without having to type them everytime.

For example, the rule "testrule" is associated to two echo commands.

How to run Makefile rules

To execute a rule in the Makefile, simply type:

```
make [name of the rule]
```

For example:

In [25]:

```
make testrule
```

```
echo this is a Makefile rule
this is a Makefile rule
echo You can associate it to as many commands you want
You can associate it to as many commands you want
```

The program "make" will automatically detect any file named "Makefile" in the current directory, and execute any rule with the specific name.

Rules can also be nested together. For example the two rules "test_exercises" and "generate_exercises" at the beginning of the file are a way to call several other rules together.

Dinner time!!