

Command Line for Data Filtering

Why not just use Excel?

- You can, but command line is quicker, more flexible and reproducible

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Gene	pval	adjp	log2fc	Chr	Start	End	Strand	Biotype	Name	Description	Cnt_1 count	Cnt_2 count
2	ENSDARG00000031683	4.64E-44	1.05E-39	-1.663131028	20	46552311	46554440	-1	protein_coding	fosab	v-fos FBJ murine osteosarcoma viral oncogene homolog Ab [Source:ZFIN;Acc:ZDB-GENE-031222-4]	555	976
3	ENSDARG00000034503	2.57E-32	2.92E-28	0.838035511	2	48309600	48375342	-1	protein_coding	per2	period circadian clock 2 [Source:ZFIN;Acc:ZDB-GENE-011220-2]	575	760
4	ENSDARG00000104773	4.28E-28	3.24E-24	-0.781054175	3	7654655	7656059	-1	protein_coding	junbb	JunB proto-oncogene, AP-1 transcription factor subunit b [Source:ZFIN;Acc:ZDB-GENE-040426-2666]	794	1066
5	ENSDARG00000055752	2.49E-23	1.42E-19	-0.744744859	14	30636955	30642819	-1	protein_coding	npas4a	neuronal PAS domain protein 4a [Source:NCBI gene;Acc:724016]	682	748
6	ENSDARG00000087440	1.58E-22	7.20E-19	2.067949498	7	22680560	22684292	1	protein_coding	ponzr4	plac8 onzin related protein 4 [Source:ZFIN;Acc:ZDB-GENE-050411-6]	16	12
7	ENSDARG00000059054	2.91E-22	1.02E-18	0.747153832	19	5674907	5687743	1	protein_coding	pkd2b	pyruvate dehydrogenase kinase, isozyme 2b [Source:ZFIN;Acc:ZDB-GENE-040426-939]	5497	5139
8	ENSDARG00000099195	3.13E-22	1.02E-18	-0.792783599	3	33761549	33762726	1	protein_coding	ier2a	immediate early response 2a [Source:NCBI gene;Acc:797137]	447	685
9	ENSDARG00000079227	1.33E-21	3.77E-18	-1.48937022	12	30517386	30523229	-1	protein_coding	plekhs1	pleckstrin homology domain containing, family S member 1 [Source:ZFIN;Acc:ZDB-GENE-080204-50]	234	404
10	ENSDARG00000098761	5.73E-21	1.45E-17	-1.719720219	22	24246213	24248420	-1	protein_coding	rgs2	regulator of G protein signaling 2 [Source:ZFIN;Acc:ZDB-GENE-040718-410]	144	163
11	ENSDARG00000086881	1.53E-18	3.21E-15	-0.860213169	11	31038409	31039533	-1	protein_coding	ier2b	immediate early response 2b [Source:ZFIN;Acc:ZDB-GENE-030131-8244]	422	539
12	ENSDARG00000093494	1.56E-18	3.21E-15	-0.772155926	1	52398205	52400660	1	protein_coding	si:ch211-217k17.9	si:ch211-217k17.9 [Source:ZFIN;Acc:ZDB-GENE-061207-21]	585	809
13	ENSDARG00000032801	3.35E-18	6.33E-15	-1.296632068	10	2715548	2755378	1	protein_coding	grk5	G protein-coupled receptor kinase 5 [Source:ZFIN;Acc:ZDB-GENE-091204-274]	100	140
14	ENSDARG00000023656	4.74E-18	8.29E-15	-9.084933074	22	14093306	14094746	1	protein_coding	he1.1	hatching enzyme 1, tandem duplicate 1 [Source:ZFIN;Acc:ZDB-GENE-021211-3]	202	3
15	ENSDARG00000097826	4.48E-17	7.27E-14	1.191409981	12	39034225	39061186	1	protein_coding	si:dkey-239b22.2	si:dkey-239b22.2 [Source:ZFIN;Acc:ZDB-GENE-131120-8]	145	138
16	ENSDARG00000053761	6.90E-17	1.05E-13	-1.288266969	7	2228276	2234151	1	protein_coding	si:dkey-187j14.4	si:dkey-187j14.4 [Source:ZFIN;Acc:ZDB-GENE-131119-26]	124	188
17	ENSDARG00000070041	1.17E-16	1.67E-13	-2.186973067	12	30528623	30540755	-1	protein_coding	zgc:153920	zgc:153920 [Source:ZFIN;Acc:ZDB-GENE-061013-702]	34	94
18	ENSDARG00000037421	1.95E-16	2.60E-13	-0.565908229	14	21332522	21336385	-1	protein_coding	egr1	early growth response 1 [Source:NCBI gene;Acc:30498]	1968	2827
19	ENSDARG00000077799	1.44E-15	1.82E-12	-0.927137112	23	45579497	45581626	1	protein_coding	egr4	early growth response 4 [Source:ZFIN;Acc:ZDB-GENE-080204-90]	338	383
20	ENSDARG00000089806	1.88E-15	2.25E-12	-7.849496181	12	16634442	16636627	-1	protein_coding	si:dkey-239j18.3	si:dkey-239j18.3 [Source:ZFIN;Acc:ZDB-GENE-121214-39]	88	4
21	ENSDARG00000040277	1.97E-14	2.24E-11	0.514565367	16	25259313	25272633	1	protein_coding	fbxo32	F-box protein 32 [Source:ZFIN;Acc:ZDB-GENE-040426-1040]	1348	1725
22	ENSDARG00000074150	6.14E-13	6.64E-10	-2.113281532	12	46231498	46252062	-1	protein_coding	si:ch211-226h7.5	si:ch211-226h7.5 [Source:ZFIN;Acc:ZDB-GENE-130531-3]	25	84

Why not just use R?

- You can, and maybe should, if you're going to do analysis in R
- But command line can be quicker if you're filtering or reformatting data to use in another tool
- R loads all data into memory first, so can be easier to filter and trim big data on command line before loading into R



Example

- Get the Ensembl ID, adjusted p-value and name of the top 10 most significantly DE genes on chromosome 2 that are down at least two-fold
- `awk '$3 < 0.05 && $4 < 1 && $5 == "2"' Amp.counts.tsv | cut -f 1,3,10 | head -10`

```
$ awk '$3 < 0.05 && $4 < 1 && $5 == "2"' Amp.counts.tsv | cut -f 1,3,10 | head -10
ENSDARG00000034503      2.9240565282158764e-28  per2
ENSDARG00000030896      4.621972579989725e-8   foxq1a
ENSDARG000000101040     1.58428501630858e-7    ccl20a.3
ENSDARG00000035859      1.6758590304348152e-6  angptl4
ENSDARG000000091260     1.9168597484745133e-6  mylk4a
ENSDARG00000017780      4.24216471563672e-4    rorcb
ENSDARG00000079059      0.003337406639336023   retreg1
ENSDARG00000075812      0.004538633443919219   kcnt2
ENSDARG00000040306      0.012767890165001398   otomp
ENSDARG00000078694      0.023409715954031074   phlpp1
```

Example data

- The example data used here can be downloaded from:
 - <https://funcgen2022.buschlab.org/downloads/Amp.counts.tsv>
 - <https://funcgen2022.buschlab.org/downloads/brain-development.txt>
- Both files are also available on penelopeprime
- Embryos exposed to amphetamine compared to unexposed controls
- Columns for `Amp.counts.tsv` are:
Ensembl gene ID (ENSDARG), p-value, adjusted p-value, \log_2 fold change, chromosome, gene start (in bp), gene end (in bp), strand (1 or -1), biotype (e.g. protein coding), name, description, counts and normalised counts

Other commands

- Not covering commands to:
 - View directory contents (`ls`)
 - Change directories (`cd`)
 - Copy files or directories (`cp`)
 - Move or rename files and directories (`mv`)
 - Delete files (`rm`) or directories (`rmdir`)
- To learn about these, see:
<http://korflab.ucdavis.edu/bootcamp.html>
- Or use the `man` command to find out about these or other commands, for example:

```
man cut
```

more

- Look at a file a page at a time using `more` (or `less`)
- e.g. `more Amp.counts.tsv`

Gene	pval	adjp	log2fc	Chr	Start	End	Strand	Biotype	Name	Description	Cnt_1 count	Cnt_2 count	Cnt_3 count	Cnt_4 count	Cnt_6 count	Amp_1 count	Amp_2 count	Amp_3 count	Amp_4 count	Amp_5 count	Amp_6 count																			
											Cnt_1 normalised count	Cnt_2 normalised count	Cnt_3 normalised count	Cnt_4 normalised count	Cnt_6 normalised count	Amp_1 normalised count	Amp_2 normalised count	Amp_3 normalised count	Amp_4 normalised count	Amp_5 normalised count	Amp_6 normalised count																			
ENSDARG0000031683			4.643118070869456e-44		1.0546842697979969e-39	-1.663131028281703					20	46552311	46554440	-1	protein_coding	fosab	v-fos	FBJ murine osteosarcoma viral oncogene homolog Ab [Source:ZFIN;Acc:ZDB-GENE-031222-4]	555	976	720	537	953	215	218	254	209	196	198	611.6631831707907	918.4427089866616	599.6419400743453	539.3186199573113	875.4713170039797	197.61986554691453	216.40538415817159	242.72234032942137	241.6850070736336	207.14127867483734	239.38965893731176
ENSDARG0000034503			2.574560007233878e-32		2.9240565282158764e-28	0.838035510509445					2	48309600	48375342	-1	protein_coding	per2	period circadian clock 2 [Source:ZFIN;Acc:ZDB-GENE-011220-2]	575	760	717	571	737	1174	1340	1325	921	1068	812	633.7050996814498	715.1807979814168	597.1434319907022	573.4654227106606	677.0434004532351	1079.096382102687	1330.198232898853	1266.1696887263122	1065.0329737551033	1128.7086001261546	981.7394093792785	
ENSDARG00000104773			4.280493831142787e-28		3.2410472458136137e-24	-0.7810541746309473					3	7654655	7656059	-1	protein_coding	junbb	JunB proto-oncogene, AP-1 transcription factor subunit b [Source:ZFIN;Acc:ZDB-GENE-040426-2666]	794	1066	946	875	987	558	461	571	467	448	481	875.0640854731673	1003.1351719055136	787.8628823754592	878.7780120347252	906.7053409054859	512.8924882566433	457.62789952714263	545.6474658586599	540.0330062363009	473.46577982819963	581.5476058022574	
ENSDARG0000055752			2.4946351270432096e-23		1.4166409227696627e-19	-0.7447448588293989					14	30636955	30642819	-1	protein_coding	npas4a	neuronal PAS domain protein 4a [Source:NCBI gene;Acc:724016]	682	748	825	578	793	433	405	463	360	334	382	751.6293530134762	703.8884695922366	687.0897230018539	580.4956468069385	728.4876751145392	397.99721758983253	402.0375256149518	442.4426912304019	416.29953371534975	352.98564835405955	461.85277633360147	
ENSDARG0000087440			1.5844414500661688e-22		7.198117507650605e-19	2.0679494980332445					7	22680560	22684292	1	protein_coding	ponzr4	plac8 onzin related protein 4 [Source:ZFIN;Acc:ZDB-GENE-050411-6]	16	12	20	15	25	75	58	78	57	72	65	17.6335332085273	11.292328389180266	16.6567205576207	15.064765920595288	22.966194045225073	68.93716240008646	57.57574440905482	74.53678167596404	65.91409283826371	76.09271461524636	78.58751429760235	
ENSDARG0000059054			2.9112546271422863e-22		1.0170186075775713e-18	0.7471538319065161					19	5674907	5687743	1	protein_coding	pdk2b	pyruvate dehydrogenase kinase, isozyme 2b [Source:ZFIN;Acc:ZDB-GENE-040426-939]	5497	5139	6604	5967	5477	9861	9092	9277	8789	8683	7383	6058.220752954661	4835.939632666449	5500.049128126356	5992.763883212806	5031.4337914279085	9063.858112363368	9025.49427874356	8865.099020614338	10163.490560622804	9176.570013947003	8926.332585526125	

- Press **Enter** to get another line
- Press **Space** or **PgDn** to see the next page
- Press **b** or **PgUp** to go back a page
- Press **q** to quit

cut

- View specific columns of your file using `cut`
- e.g. `cut -f1,3-4,10 Amp.counts.tsv`
- Will only show columns 1 (ID), 3 (adjusted p-value), 4 (\log_2 fold change) and 10 (name)
- -f is short for “fields”

Gene	adjp	log2fc	Name
ENSDARG00000031683	1.0546842697979969e-39	-1.663131028281703	fosab
ENSDARG00000034503	2.9240565282158764e-28	0.838035510509445	per2
ENSDARG000000104773	3.2410472458136137e-24	-0.7810541746309473	junbb
ENSDARG000000055752	1.4166409227696627e-19	-0.7447448588293989	npas4a
ENSDARG000000087440	7.198117507650605e-19	2.0679494980332445	ponzr4
ENSDARG000000059054	1.0170186075775713e-18	0.7471538319065161	pdk2b
ENSDARG000000099195	1.0170186075775713e-18	-0.7927835991597645	ier2a
ENSDARG000000079227	3.769207162775586e-18	-1.4893702196945517	plekhs1
ENSDARG000000098761	1.4460258466667767e-17	-1.7197202185116396	rgs2
ENSDARG000000086881	3.2117430800886777e-15	-0.8602131689498111	ier2b
ENSDARG000000093494	3.2117430800886777e-15	-0.772155925667385	si:ch211-217k17.9
ENSDARG000000032801	6.334736320467128e-15	-1.2966320682914954	grk5
ENSDARG000000023656	8.289616663669282e-15	-9.08493307391428	he1.1
ENSDARG000000097826	7.271082588264714e-14	1.1914099813769465	si:dkey-239b22.2
ENSDARG000000053761	1.045518015896552e-13	-1.288266969193545	si:dkey-187j14.4
ENSDARG000000070041	1.6653254390037004e-13	-2.1869730666865665	zgc:153920
ENSDARG000000037421	2.599300812048074e-13	-0.5659082294197473	egr1

Pipe

- Can join two commands with a **pipe** |
- e.g. `cut -f1,3-4,10 Amp.counts.tsv | more`
- The output of the **cut** command becomes the input of the **more** command

```
Gene      adjp      log2fc  Name
ENSDARG00000031683  1.0546842697979969e-39  -1.663131028281703  fosab
ENSDARG00000034503  2.9240565282158764e-28  0.838035510509445  per2
ENSDARG000000104773  3.2410472458136137e-24  -0.7810541746309473  junbb
ENSDARG000000055752  1.4166409227696627e-19  -0.7447448588293989  npas4a
ENSDARG000000087440  7.198117507650605e-19  2.0679494980332445  ponzr4
ENSDARG000000059054  1.0170186075775713e-18  0.7471538319065161  pdk2b
ENSDARG000000099195  1.0170186075775713e-18  -0.7927835991597645  ier2a
ENSDARG000000079227  3.769207162775586e-18  -1.4893702196945517  plekhs1
ENSDARG000000098761  1.4460258466667767e-17  -1.7197202185116396  rgs2
ENSDARG000000086881  3.2117430800886777e-15  -0.8602131689498111  ier2b
ENSDARG000000093494  3.2117430800886777e-15  -0.772155925667385  si:ch211-217k17.9
ENSDARG000000032801  6.334736320467128e-15  -1.2966320682914954  grk5
ENSDARG000000023656  8.289616663669282e-15  -9.08493307391428  he1.1
ENSDARG000000097826  7.271082588264714e-14  1.1914099813769465  si:dkey-239b22.2
ENSDARG000000053761  1.045518015896552e-13  -1.288266969193545  si:dkey-187j14.4
ENSDARG000000070041  1.6653254390037004e-13  -2.1869730666865665  zgc:153920
ENSDARG000000037421  2.599300812048074e-13  -0.5659082294197473  egr1
```

column

- Can format column data tidily using `column`
- e.g. `cut -f1,3-4,10 Amp.counts.tsv | column -t | more`
- `-t` is short for “table”

Gene	adjp	log2fc	Name
ENSDARG00000031683	1.0546842697979969e-39	-1.663131028281703	fosab
ENSDARG00000034503	2.9240565282158764e-28	0.838035510509445	per2
ENSDARG000000104773	3.2410472458136137e-24	-0.7810541746309473	junbb
ENSDARG00000055752	1.4166409227696627e-19	-0.7447448588293989	npas4a
ENSDARG00000087440	7.198117507650605e-19	2.0679494980332445	ponzr4
ENSDARG00000059054	1.0170186075775713e-18	0.7471538319065161	pdk2b
ENSDARG00000099195	1.0170186075775713e-18	-0.7927835991597645	ier2a
ENSDARG00000079227	3.769207162775586e-18	-1.4893702196945517	plekhs1
ENSDARG00000098761	1.4460258466667767e-17	-1.7197202185116396	rgs2
ENSDARG00000086881	3.2117430800886777e-15	-0.8602131689498111	ier2b
ENSDARG00000093494	3.2117430800886777e-15	-0.772155925667385	si:ch211-217k17.9
ENSDARG00000032801	6.334736320467128e-15	-1.2966320682914954	grk5
ENSDARG00000023656	8.289616663669282e-15	-9.08493307391428	he1.1
ENSDARG00000097826	7.271082588264714e-14	1.1914099813769465	si:dkey-239b22.2
ENSDARG00000053761	1.045518015896552e-13	-1.288266969193545	si:dkey-187j14.4

head

- Can truncate data using head
- e.g. `cut -f1,3-4,10 Amp.counts.tsv | head -10 | column -t`
- Gives top 10 lines of output
- Change the number to get a different number of lines

```
$ cut -f1,3-4,10 Amp.counts.tsv | head -10 | column -t
Gene                adjp                log2fc              Name
ENSDARG00000031683  1.0546842697979969e-39  -1.663131028281703  fosab
ENSDARG00000034503  2.9240565282158764e-28  0.838035510509445   per2
ENSDARG000000104773 3.2410472458136137e-24  -0.7810541746309473  junbb
ENSDARG00000055752  1.4166409227696627e-19  -0.7447448588293989  npas4a
ENSDARG00000087440  7.198117507650605e-19   2.0679494980332445   ponzr4
ENSDARG00000059054  1.0170186075775713e-18  0.7471538319065161   pdk2b
ENSDARG00000099195  1.0170186075775713e-18  -0.7927835991597645  ier2a
ENSDARG00000079227  3.769207162775586e-18   -1.4893702196945517  plekhs1
ENSDARG00000098761  1.4460258466667767e-17  -1.7197202185116396  rgs2
```

tail

- Can also truncate data using `tail`
- e.g. `cut -f1,3-4,10 Amp.counts.tsv | tail -10 | column -t`
- Gives last 10 lines of output
- Change the number to get a different number of lines

```
$ cut -f1,3-4,10 Amp.counts.tsv | tail -10 | column -t
ENSDARG00000117705  NA  NA  BX927244.2
ENSDARG00000117712  NA  NA  CABZ01086755.1
ENSDARG00000117742  NA  NA  F0904898.6
ENSDARG00000117746  NA  NA  BX897747.1
ENSDARG00000117774  NA  NA  BX571809.3
ENSDARG00000117785  NA  NA  AL807749.2
ENSDARG00000117797  NA  NA  BX005302.1
ENSDARG00000117798  NA  NA  BX001047.1
ENSDARG00000117811  NA  NA  BX545917.7
ENSDARG00000117815  NA  NA  FP236157.8
```

AWK

- AWK is a powerful command line tool used for text processing
- Can filter based on a specific column
- `awk -F"\t" '$4 > 0' Amp.counts.tsv | more`
(get “up” genes)
- `awk -F"\t" '$5 == "2"' Amp.counts.tsv | more`
(get genes on chromosome 2)
- `-F"\t"` tells AWK that the file is delimited with tabs
- `$4` is the 4th column; `==` checks for equality (whereas `=` indicates assignment)

```
ENSDARG00000034503 2.574560007233878e-32 2.9240565282158764e-28 0.838035510509445 2 48309600 48375342 -1 protein_coding per2 period circadian clock 2 [Source:ZFIN;Acc:ZD
B-GENE-011220-2] 575 760 717 571 737 1174 1340 1325 921 1068 812 633.7050996814498 715.1807979814168 597.1434319907022 573.4654227106606 677.
0434004532351 1079.096382102687 1330.198232898853 1266.1696887263122 1065.0329737551033 1128.7086001261546 981.7394093792785
ENSDARG00000030896 7.52863682410829e-11 4.621972579989725e-8 -0.7985946819383261 2 720779 722170 -1 protein_coding foxq1a forkhead box Q1a [Source:ZFIN;Acc:ZDB-GENE-070424-74] 179
211 240 130 232 84 124 110 93 110 101 197.2751527703992 198.55677417641968 199.88064669144842 130.56130464515917 213.12628073968867 77.2
0962188809685 123.09297080556549 105.11597415841082 107.5440462097987 116.25275843995973 122.11290683165903
ENSDARG000000101040 2.859594350369878e-10 1.58428501630858e-7 -2.0331427128516757 2 45191049 45193014 1 protein_coding ccl20a.3 chemokine (C-C motif) ligand 20a, du
plicate 3 [Source:ZFIN;Acc:ZDB-GENE-081022-193] 26 36 26 16 39 6 0 4 8 8 12 28.654491463856864 33.876985167540795 21.653736724906913 16.0
69083648634976 35.82726271055111 5.514972992006918 0 3.822399060305848 9.251100749229995 8.454746068360707 14.508464178018896
ENSDARG00000035859 3.836437137689209e-9 1.6758590304348152e-6 0.4745609888648301 2 24536762 24542518 1 protein_coding angptl4 angiotensin-like 4 [Source:ZFIN;Acc:ZDB-GEN
E-041111-222] 896 863 1201 849 1001 1432 1458 1304 987 1206 988 987.4778596775288 812.1066166552141 1000.2360694851232 852.6657511056933 919.56640957
08119 1316.2402207589844 1447.3350922138263 1246.1020936597065 1141.3545549362507 1274.5529698053767 1194.5302173235557
```

Filtering on multiple columns with AWK

- Could just pipe two AWK commands together:

```
awk -F"\t" '$4 > 0' Amp.counts.tsv | awk -F"\t" '$5 == "2"' | more
```

- But can combine terms with &&:

```
awk -F"\t" '$4 > 0 && $5 == "2"' Amp.counts.tsv | more
```

```
ENSDARG00000034503 2.574560007233878e-32 2.9240565282158764e-28 0.838035510509445 2 48309600 48375342 -1 protein_coding per2 period circadian clock 2 [Source:ZFIN;Acc:ZD
B-GENE-011220-2] 575 760 717 571 737 1174 1340 1325 921 1068 812 633.7050996814498 715.1807979814168 597.1434319907022 573.4654227106606 677.
0434004532351 1079.096382102687 1330.198232898853 1266.1696887263122 1065.0329737551033 1128.7086001261546 981.7394093792785
ENSDARG00000035859 3.836437137689209e-9 1.6758590304348152e-6 0.4745609888648301 2 24536762 24542518 1 protein_coding angptl4 angiopoietin-like 4 [Source:ZFIN;Acc:ZDB-GEN
E-041111-222] 896 863 1201 849 1001 1432 1458 1304 987 1206 988 987.4778596775288 812.1066166552141 1000.2360694851232 852.6657511056933 919.56640957
08119 1316.2402207589844 1447.3350922138263 1246.1020936597065 1141.3545549362507 1274.5529698053767 1194.5302173235557
ENSDARG00000091260 4.556919498904852e-9 1.9168597484745133e-6 0.43618182007037065 2 394166 424925 1 protein_coding mylk4a myosin light chain kinase family, member 4a [Source:ZFIN;Acc
:ZDB-GENE-120824-2] 360 456 566 407 514 597 586 594 481 628 519 396.7544971918643 429.1084787888501 471.3851917806659 408.7573153121522 472.
1849495698275 548.7398127046882 581.7135555811401 567.6262604554184 556.2224325474534 663.6975663663155 627.4910756993172
ENSDARG00000017780 1.9609390056872356e-6 4.24216471563672e-4 0.2881630249967304 2 49417900 49454451 1 protein_coding rorcb RAR-related orphan receptor C b [Source:ZFIN
;Acc:ZDB-GENE-040724-215] 1893 2265 2743 2257 2462 2798 2751 2757 2296 2607 2334 2086.2673977338864 2131.4269834577753 2284.4692244776793 2266.745112185571
2261.710789573765 2571.8157386058924 2730.8771184363763 2634.588552315806 2655.0659150290085 2755.190375027046 2821.896282624675
```

Combining AWK and cut

- `awk -F"\t" '$4 > 0' Amp.counts.tsv | cut -f1-3,10 | column -t | more`

Gene	pval	adjp	Name
ENSDARG00000034503	2.574560007233878e-32	2.9240565282158764e-28	per2
ENSDARG00000087440	1.5844414500661688e-22	7.198117507650605e-19	ponzr4
ENSDARG00000059054	2.9112546271422863e-22	1.0170186075775713e-18	pdk2b
ENSDARG00000097826	4.481406834061457e-17	7.271082588264714e-14	si:dkey-239b22.2
ENSDARG00000040277	1.969136601114675e-14	2.236446894715992e-11	fbxo32
ENSDARG00000020876	2.635143664325429e-12	2.394291533406085e-9	pdk2a
ENSDARG00000099186	3.0174161517054074e-12	2.6361772263841666e-9	slc1a5
ENSDARG00000094198	9.873623557087298e-12	7.23481803545929e-9	CR354556.1
ENSDARG00000099555	1.1974939254669367e-11	8.242749853630747e-9	foxo1a
ENSDARG00000100826	3.16740548704351e-11	2.116106342299804e-8	hif1a1
ENSDARG00000058285	5.4037342231985244e-11	3.409606191109846e-8	cpt1b
ENSDARG00000074526	1.208157401208195e-9	6.237112583737307e-7	zbtb16b
ENSDARG00000102558	1.8680631201135703e-9	9.224576907256467e-7	pde6ha
ENSDARG00000069946	2.3051172906822488e-9	1.090848734538485e-6	itga6b
ENSDARG00000045768	2.658651018742464e-9	1.232474650831328e-6	cry1a

Reordering columns

- `awk -F"\t" '$4 > 0' Amp.counts.tsv | cut -f1,10,3 | column -t | more`
- Note that name column is 3rd, not 2nd, as requested - can't reorder columns with cut

```
Gene          adjp          Name
ENSDARG0000034503 2.9240565282158764e-28 per2
ENSDARG0000087440 7.198117507650605e-19 ponzr4
ENSDARG0000059054 1.0170186075775713e-18 pdk2b
ENSDARG0000097826 7.271082588264714e-14 si:dkey-239b22.2
ENSDARG0000040277 2.236446894715992e-11 fbxo32
ENSDARG0000020876 2.394291533406085e-9  pdk2a
ENSDARG0000099186 2.6361772263841666e-9  slc1a5
ENSDARG0000094198 7.23481803545929e-9   CR354556.1
ENSDARG0000099555 8.242749853630747e-9  foxo1a
ENSDARG00000100826 2.116106342299804e-8  hif1a1
ENSDARG0000058285 3.409606191109846e-8  cpt1b
ENSDARG0000074526 6.237112583737307e-7  zbtb16b
ENSDARG00000102558 9.224576907256467e-7  pde6ha
ENSDARG0000069946 1.090848734538485e-6  itga6b
ENSDARG0000045768 1.232474650831328e-6  cry1a
```


Replacing cut with AWK

- `awk -F"\t" '$4 > 0 { print $1 "\t" $10 "\t" $2 "\t" $3 "\t" $4 }' Amp.counts.tsv | column -t | more`
(can change order of columns, but can't do ranges)
- `"\t"` indicates a tab should be printed

Gene	Name	pval	adjp	log2fc
ENSDARG00000034503	per2	2.574560007233878e-32	2.9240565282158764e-28	0.838035510509445
ENSDARG00000087440	ponzr4	1.5844414500661688e-22	7.198117507650605e-19	2.0679494980332445
ENSDARG00000059054	pdk2b	2.9112546271422863e-22	1.0170186075775713e-18	0.7471538319065161
ENSDARG00000097826	si:dkey-239b22.2	4.481406834061457e-17	7.271082588264714e-14	1.1914099813769465
ENSDARG00000040277	fbxo32	1.969136601114675e-14	2.236446894715992e-11	0.5145653669246746
ENSDARG00000020876	pdk2a	2.635143664325429e-12	2.394291533406085e-9	0.4230653213635479
ENSDARG00000099186	slc1a5	3.0174161517054074e-12	2.6361772263841666e-9	0.36506359299074265
ENSDARG00000094198	CR354556.1	9.873623557087298e-12	7.23481803545929e-9	0.5368545572513594
ENSDARG00000099555	foxo1a	1.1974939254669367e-11	8.242749853630747e-9	0.40927581346248776
ENSDARG000000100826	hif1a1	3.16740548704351e-11	2.116106342299804e-8	0.3813334401548895
ENSDARG00000058285	cpt1b	5.4037342231985244e-11	3.409606191109846e-8	0.41574265769828317
ENSDARG00000074526	zbtb16b	1.208157401208195e-9	6.237112583737307e-7	0.5406284577533254
ENSDARG000000102558	pde6ha	1.8680631201135703e-9	9.224576907256467e-7	0.5053666683932402
ENSDARG00000069946	itga6b	2.3051172906822488e-9	1.090848734538485e-6	0.5197625731527228
ENSDARG00000045768	cry1a	2.658651018742464e-9	1.232474650831328e-6	0.2891972239228979
ENSDARG00000056511	arr3a	2.8166643208909532e-9	1.27961060098076e-6	0.3832728509279693
ENSDARG00000035859	angptl4	3.836437137689209e-9	1.6758590304348152e-6	0.4745609888648301
ENSDARG000000101849	adipor2	4.470575317447273e-9	1.9160211006757513e-6	0.2527385193558007

sort

- Reorder data using `sort`
- `cut -f1-4,10 Amp.counts.tsv | sort -k5 | more`
- `-k5` means sort by the 5th column (name)

```
ENSDARG00000074221 0.4902499074462968 0.8068331047929754 0.05993427211789899 ABCA7
ENSDARG00000099178 0.08926367721406397 0.46536808771027627 0.9736178388501474 ABCD2
ENSDARG00000059587 0.13280449231351354 0.5294323440947504 0.1382552267743402 ABR
ENSDARG00000098174 0.81345545632802 0.9449975214160297 -0.06913537695241816 ACADSB
ENSDARG00000056478 0.9382919844835393 0.9839936485477191 -0.006294613763243476 ACAP2
ENSDARG00000024602 0.07768268586401583 0.4474441636609983 -0.2731666039671127 ACBD3
ENSDARG00000086314 0.6377163009669158 0.8781356557022001 -0.2899640060599673 ACKR2
ENSDARG00000054534 0.0018577735814168537 0.08406240418701959 0.5598335481644736 ACOT12
ENSDARG00000059503 0.9979383698254681 0.9995937742830112 3.5866570215769726e-4 ACSF3
ENSDARG00000057911 0.029077033790526827 0.3121720244447629 -0.19946095230852298 ACTC1
ENSDARG00000086172 0.10436470932076417 0.48980255624404095 0.36420130689853286 ACVR1C
ENSDARG00000116062 0.7241280288072605 0.912845783581604 -0.1450471496961243 ADAM12
ENSDARG00000102478 0.0043942546326026285 0.13344317376947687 0.8337613325376052 ADAMTS7
ENSDARG00000075899 0.9849733719234435 0.9952800192700316 -0.002113091253950258 ADGRL2
ENSDARG00000090624 0.8765609531985548 0.9653806014598366 0.009904849786200228 ADGRL3
ENSDARG00000035573 0.8885223764951126 0.9691618736726122 -0.013276926073133785 AK6
```

More sort

- `cut -f1-4,10 Amp.counts.tsv | sort -r -g -k4 | more`
- `-g` means sort numerically
- `-r` means reverse the order

```
ENSDARG00000055809    0.015475047955822855    NA    3.291646764751294    vtg2
ENSDARG00000092421    0.12755867561335754    NA    3.1505960159969315    si:ch1073-394i4.1
ENSDARG00000096024    0.10070173812049037    NA    3.032874286456962    si:ch211-178j18.4
ENSDARG00000032637    0.025302618854766215    NA    2.865105028020855    si:dkey-261m9.12
ENSDARG00000070132    0.032400250118289835    NA    2.8583244608938028    aste1a
ENSDARG00000105234    0.06184304848118969    NA    2.7613028575893215    BX255920.1
ENSDARG00000094925    0.0741417823817791    NA    2.7591835876573607    BX248322.1
ENSDARG00000089852    0.02089351851593089    NA    2.7295381506308836    si:dkey-51d8.1
ENSDARG00000099020    0.051606931575944835    NA    2.7202371505338925    CT025874.1
ENSDARG00000092303    0.013348407485646967    NA    2.6933365833458685    si:dkey-11o15.8
ENSDARG00000098020    0.21072824055774922    NA    2.690545084315753    BX005129.5
ENSDARG00000096584    0.5016741076192502    NA    2.6886247767706344    BX004779.2
ENSDARG00000112149    0.23832152690242808    NA    2.5940826373675265    CABZ01080023.2
ENSDARG00000117684    0.08623243510674931    NA    2.5911550750900894    BX908744.1
```

grep

- Extract data by search term using `grep`
- `grep ENSDARG00000068567 Amp.counts.tsv`
- `grep shh Amp.counts.tsv`

```
$ grep shh Amp.counts.tsv
ENSDARG00000068567 0.45075051918264025 0.7858468066032446 -0.05630380531678281 7 40884012 40893295 1 protein_coding shha sonic hedgehog signaling molecule a [Source:
NCBI gene;Acc:30269] 340 432 417 351 350 374 345 388 332 295 269 374.71258068120517 406.52382201048954 347.2926236263916 352.5155225419298 321.
526716633151 343.76664983509784 342.476410709033 370.77270884966725 383.9206810930448 311.7687612708011 325.2314053239236
ENSDARG00000038867 0.8088626405568975 0.9440013462459302 0.029186073400437888 2 30050541 30055432 -1 protein_coding shhb sonic hedgehog signaling molecule b [Source:
NCBI gene;Acc:30444] 81 77 131 102 103 92 95 108 82 100 76 89.26976186816947 72.45910716390671 109.10151965241559 102.44040826004796 94.6
207194663273 84.56291921077273 94.30509860103807 103.20477462825791 94.82378267960745 105.68432585450884 91.88693979411967
```

More grep

- Extract using list of search terms
- `grep -f brain-development.txt Amp.counts.tsv | more`
- `-f` is short for “file”

```
ENSDARG0000040649 2.0693015551829967e-7 5.8029857809854036e-5 0.4594452789031625 25 314608 348784 -1 protein_coding prickle1a prickle homolog 1a [Source:ZFIN;Acc:ZDB-GENE-030724-5] 240 246 341 261 308 341 379 404 314 336 328 264.5029981279095 231.49273197819545 283.997085507433 262.126927018358 282.9435106371729 313.4342983790598 376.22770915572033 386.0623050908907 363.1057044072773 355.0993348711497 396.5646875325165
ENSDARG0000054879 4.6761471974167624e-4 0.03319333862166305 -0.28695600548579464 12 25600685 25603341 1 protein_coding six3b SIX homeobox 3b [Source:ZFIN;Acc:ZDB-GENE-990415-128] 402 534 555 427 471 406 364 398 344 355 285 443.04252186424844 502.50861331852184 462.22399547397447 428.8436698729459 432.68309581204034 373.17983912580144 361.3374304292406 380.3287065004319 397.7973322168898 375.1793567835064 344.5760242279488
ENSDARG0000043531 0.0010306737293043567 0.05794988554739718 -0.2059555522857944 20 15552657 15554606 1 protein_coding jun Jun proto-oncogene, AP-1 transcription factor subunit [Source:ZFIN;Acc:ZDB-GENE-030131-7859] 1151 1389 1687 1375 1454 1320 1170 1206 1126 1059 843 1268.5122951884327 1307.0870110476158 1404.994379035306 1380.936876054568 1335.71384567029 1213.2940582415217 1161.4417406654163 1152.4533166822132 1302.0924304541218 1119.1970107992488 1019.2196085058274
ENSDARG00000100558 0.0020098629255510415 0.0884768146393254 -0.2857672079342749 14 14841685 14846166 1 protein_coding slbp stem-loop binding protein [Source:ZFIN;Acc:ZDB-GENE-030131-9686] 226 224 264 194 230 169 191 178 149 181 158 249.07365657044812 210.79012993136496 219.86871136059327 194.83763923969906 211.28898521607067 155.33840594152818 189.60288245050813 170.09675818361023 172.30175145440865 191.288629796661 191.02811167724877
```


WC

- `wc` stands for "word count"
- Count number of lines returned
- `wc -l brain-development.txt`
- `grep -f brain-development.txt Amp.counts.tsv | wc -l`
- `-l` is short for "lines"

```
$ wc -l brain-development.txt
    327 brain-development.txt
$ grep -f brain-development.txt Amp.counts.tsv | wc -l
    274
```

Redirecting to a file

- `grep hox Amp.counts.tsv > hox.txt`
- `more hox.txt`
- Take care - file will be overwritten without any warning if it already exists

```
$ grep hox Amp.counts.tsv > hox.txt
$ more hox.txt
ENSDARG0000054033      6.896475144500434e-4      0.043514842474257596      -0.33468620327317333      12      27141140      27142834      1      protein_coding hoxb1b homeobox B1b [Source:NCBI gene;Acc:30374]
      127      153      171      124      164      109      102      123      108      109      92      139.96616984268545      143.9771869620484      142.414960767657      124.53539827692106      150.65823293667648
      100.188676021459      101.25389534006193      117.53877110440483      124.88986011460493      115.19591518141465      111.23155869814487
ENSDARG0000059280      0.0030554189936227547      0.10829032882017198      -0.1912091212105787      9      1937049 1959917 -1      protein_coding hoxd3a homeobox D3a [Source:NCBI gene;Acc:30349]      473      573
      599      445      523      480      433      498      386      380      344      521.2913254770883      539.2086805833577      498.86878070074005      446.92138897766023      480.4527794261085      441.19783936
05534 429.83271257104724      475.8886830080781      446.36561115034726      401.60043824713364      415.90930643654167
ENSDARG0000056819      0.02185410597016362      0.27955207143871863      -0.285575275597263      16      20915319      20917584      1      protein_coding hoxa9b homeobox A9b [Source:NCBI gene;Acc:58048]
      121      144      168      143      157      139      119      144      93      89      81      133.3535948894877      135.50794067016318      139.9164526840139      143.61743510967509      144.22769860401345
      127.76354098149359      118.12954456340559      137.60636617101054      107.5440462097987      94.05905001051288      97.93213320162754
ENSDARG0000056023      0.03509183190691942      0.337729967517132      -0.16325094218410652      3      23677351      23682436      1      protein_coding hoxb9a homeobox B9a [Source:NCBI gene;Acc:30344]
      206      269      286      217      260      243      212      220      192      191      158      227.031740059789      253.13636139079097      238.19110397397603      217.93694698461184      238.84841807034076
      223.35640617628016      210.4492726675797      210.23194831682164      222.02641798151987      201.8570623821119      191.02811167724877
```


Appending to a file

- `head -1 Amp.counts.tsv > hox.txt`
- `grep hox Amp.counts.tsv >> hox.txt`
- `more hox.txt`

```
$ head -1 Amp.counts.tsv > hox.txt
$ grep hox Amp.counts.tsv >> hox.txt
$ more hox.txt
Gene      pval      adjp      log2fc Chr      Start  End      Strand  Biotype Name      Description      Cnt_1 count      Cnt_2 count      Cnt_3 count      Cnt_4 count      Cnt_6 count      Amp_1 count      Amp_2 count      Amp_
3 count Amp_4 count      Amp_5 count      Amp_6 count      Cnt_1 normalised count      Cnt_2 normalised count      Cnt_3 normalised count      Cnt_4 normalised count      Cnt_6 normalised count      Amp_1 normalised count      Amp_2 normal
ised count      Amp_3 normalised count      Amp_4 normalised count      Amp_5 normalised count      Amp_6 normalised count
ENSDARG00000054033      6.896475144500434e-4      0.043514842474257596      -0.33468620327317333      12      27141140      27142834      1      protein_coding      hoxb1b      homeobox B1b [Source:NCBI gene;Acc:30374]
127      153      171      124      164      109      102      123      108      109      92      139.96616984268545      143.9771869620484      142.414960767657      124.53539827692106      150.65823293667648
100.188676021459      101.25389534006193      117.53877110440483      124.88986011460493      115.19591518141465      111.23155869814487
ENSDARG00000059280      0.0030554189936227547      0.10829032882017198      -0.1912091212105787      9      1937049      1959917      -1      protein_coding      hoxd3a      homeobox D3a [Source:NCBI gene;Acc:30349]      473      573
599      445      523      480      433      498      386      380      344      521.2913254770883      539.2086805833577      498.86878070074005      446.92138897766023      480.4527794261085      441.19783936
05534      429.83271257104724      475.8886830080781      446.36561115034726      401.60043824713364      415.90930643654167
ENSDARG00000056819      0.02185410597016362      0.27955207143871863      -0.285575275597263      16      20915319      20917584      1      protein_coding      hoxa9b      homeobox A9b [Source:NCBI gene;Acc:58048]
121      144      168      143      157      139      119      144      93      89      81      133.3535948894877      135.50794067016318      139.9164526840139      143.61743510967509      144.22769860401345
127.76354098149359      118.12954456340559      137.60636617101054      107.5440462097987      94.05905001051288      97.93213320162754
ENSDARG00000056023      0.03509183190691942      0.337729967517132      -0.16325094218410652      3      23677351      23682436      1      protein_coding      hoxb9a      homeobox B9a [Source:NCBI gene;Acc:30344]
206      269      286      217      260      243      212      220      192      191      158      227.031740059789      253.13636139079097      238.19110397397603      217.93694698461184      238.84841807034076
223.35640617628016      210.4492726675797      210.23194831682164      222.02641798151987      201.8570623821119      191.02811167724877
```

Thank You

Any Questions?