

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	
1	GeneID	pval	adjp	log2fc	Chr	Start	End	Strand	Biotype	Name	Description
2	ENSDARG00000062262	0	0	-3.956870455	23	45958860	45970337	-1	protein_coding	ednrab	endothelin receptor type Ab [Source:ZFIN;Acc:ZDB-GENE-040426-2364]
3	ENSDARG00000045893	2.18E-244	2.14E-240	-3.109541949	25	20260077	20283617	-1	protein_coding	kctd15a	potassium channel tetramerization domain containing 15a [Source:ZFIN;Acc:ZDB-GENE-041010-105]
4	ENSDARG00000101407	2.98E-232	1.95E-228	-3.385493638	23	46240647	46250912	-1	protein_coding	tgm1l4	transglutaminase 1 like 4 [Source:ZFIN;Acc:ZDB-GENE-050913-29]
5	ENSDARG00000094041	3.79E-171	1.86E-167	-2.890598481	19	5450844	5456086	-1	protein_coding	krt17	keratin 17 [Source:ZFIN;Acc:ZDB-GENE-060503-86]
6	ENSDARG00000077467	2.72E-151	1.07E-147	-2.769872183	3	2013057	2022138	1	protein_coding	sox10	SRY (sex determining region Y)-box 10 [Source:ZFIN;Acc:ZDB-GENE-011207-1]
7	ENSDARG00000054616	4.08E-124	1.34E-120	-2.199834566	3	30790404	30795479	1	protein_coding	cldni	claudin i [Source:ZFIN;Acc:ZDB-GENE-010328-9]
8	ENSDARG00000034836	1.23E-115	3.47E-112	-2.275454463	8	7041426	7049959	-1	protein_coding	KRT84	keratin 84 [Source:HGNC Symbol;Acc:HGNC:6461]
9	ENSDARG00000059279	1.79E-107	4.41E-104	-1.350951186	24	8581299	8592141	-1	protein_coding	tfap2a	transcription factor AP-2 alpha [Source:ZFIN;Acc:ZDB-GENE-011212-6]
10	ENSDARG00000027611	1.23E-91	2.68E-88	-1.723492593	9	24412199	24431723	-1	protein_coding	sdpra	serum deprivation response a [Source:ZFIN;Acc:ZDB-GENE-090313-215]
11	ENSDARG00000100190	3.76E-91	7.39E-88	-1.99471938	1	58697974	58709503	-1	protein_coding	ENSDARG00000100190	
12	ENSDARG00000021720	4.89E-70	8.74E-67	-2.351072716	6	40186699	40314609	-1	protein_coding	col7a1	collagen, type VII, alpha 1 [Source:ZFIN;Acc:ZDB-GENE-030131-2427]
13	ENSDARG00000058371	2.68E-67	4.39E-64	-1.442726714	23	10238226	10242488	-1	protein_coding	krt5	keratin 5 [Source:ZFIN;Acc:ZDB-GENE-991110-23]
14	ENSDARG00000090369	9.61E-66	1.45E-62	-1.896096374	3	36146118	36154684	1	protein_coding	zgc:86896	zgc:86896 [Source:ZFIN;Acc:ZDB-GENE-040625-80]
15	ENSDARG00000013310	8.21E-65	1.15E-61	-1.93060886	24	25768121	25843795	1	protein_coding	map3k15	mitogen-activated protein kinase kinase kinase 15 [Source:ZFIN;Acc:ZDB-GENE-091204-246]
16	ENSDARG00000045887	2.70E-64	3.54E-61	-2.246133801	10	38700055	38719558	1	protein_coding	mmp30	matrix metalloproteinase 30 [Source:ZFIN;Acc:ZDB-GENE-060421-5765]
17	ENSDARG00000042021	9.48E-63	1.16E-59	-1.878906506	18	14876157	14891923	-1	protein_coding	mapk12a	mitogen-activated protein kinase 12a [Source:ZFIN;Acc:ZDB-GENE-990415-257]
18	ENSDARG00000031782	7.15E-62	8.27E-59	-1.592129115	17	25463075	25531858	1	protein_coding	aim1a	absent in melanoma 1a [Source:ZFIN;Acc:ZDB-GENE-041008-100]
19	ENSDARG00000092947	7.42E-61	8.11E-58	-1.70802894	19	5445934	5464258	-1	protein_coding	cyt1	type I cyokeratin, enveloping layer [Source:ZFIN;Acc:ZDB-GENE-991008-6]

Why not just use R?

- You can, and probably 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



Hadley Wickham and others at RStudio [CC BY-SA 4.0]

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"' example.tsv | cut -f 1,3,10 | head -10`

```
[iansealy@kevin]~$ awk '$3 < 0.05 && $4 < 1 && $5 == "2"' example.tsv | cut -f 1,3,10 | head -10
ENSDARG00000068177      1.36173476748606e-26      pak2a
ENSDARG00000034080      2.56301537071252e-26      plcd1b
ENSDARG00000020929      1.37406692887583e-18      fam49ba
ENSDARG00000061797      4.13140951350604e-17      zgc:154006
ENSDARG00000052997      2.09283565896043e-16      sema4e
ENSDARG00000024829      1.97492421027613e-15      tnw
ENSDARG00000014522      3.41819488862943e-15      cdh6
ENSDARG00000032317      1.05713357505467e-13      tox
ENSDARG00000102036      4.49641203490701e-12      PTCHD3
ENSDARG00000092225      7.03067067460809e-12      si:dkeyp-13a3.10
```

Example data

- The example data used here can be downloaded from:
 - <https://funcgen2019.buschlab.org/downloads/command-line/example.tsv>
 - <https://funcgen2019.buschlab.org/downloads/command-line/myc-targets.tsv>
- Both files are also available on penelopeprime.
- Mutants homozygous for two neural crest transcription factors are compared to wild-type siblings
(“The gene regulatory basis of genetic compensation during neural crest induction”)
- Columns 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 and description

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 example.tsv`

GeneID	pval	adjp	log2fc	Chr	Start	End	Strand	Biotype	Name	Description								
ENSDARG00000062262			0	0	-3.95687045483116			23	45958860	45970337	-1	protein_coding	ednra	endothelin receptor type A [Source:ZFIN;Acc:ZDB-GENE-040426-2364]				
ENSDARG00000045893			2.17893211051937e-244		2.14308867730133e-240			-3.10954194931744		25	20260077	20283617	-1	protein_coding	kctd15a	potassium channel tetramerization domain containing 15a [Source:ZFIN;Acc:ZDB-GENE-041010-105]		
ENSDARG000000101407			2.97649272721949e-232		1.95168628123782e-228			-3.38549363750172		23	46240647	46250912	-1	protein_coding	tgm1l4	transglutaminase 1 like 4 [Source:ZFIN;Acc:ZDB-GENE-050913-29]		
ENSDARG00000094041			3.79049940272095e-171		1.8640728437731e-167			-2.89059848142129		19	5450844	5456086	-1	protein_coding	krt17	keratin 17 [Source:ZFIN;Acc:ZDB-GENE-060503-86]		
ENSDARG00000077467			2.72042674232468e-151		1.07027028896538e-147			-2.76987218334732		3	2013057	2022138	1	protein_coding	sox10	SRY (sex determining region Y)-box 10 [Source:ZFIN;Acc:ZDB-GENE-011207-1]		
ENSDARG00000054616			4.0826631226122e-124		1.33850110474841e-120			-2.19983456562198		3	30790404	30795479	1	protein_coding	cldn1	claudin 1 [Source:ZFIN;Acc:ZDB-GENE-010328-9]		
ENSDARG00000034836			1.23442424108643e-115		3.46890846377303e-112			-2.27545446343907		8	7041426	7049959	-1	protein_coding	KRT84	keratin 84 [Source:HGNC Symbol;Acc:HGNC:6461]		
ENSDARG00000059279			1.79197114016437e-107		4.40623303727167e-104			-1.35095118552025		24	8581299	8592141	-1	protein_coding	tfap2a	transcription factor AP-2 alpha [Source:ZFIN;Acc:ZDB-GENE-011212-6]		
ENSDARG00000027611			1.22613801411055e-91		2.67992898617429e-88			-1.72349259342463		9	24412199	24431723	-1	protein_coding	sdpra	serum deprivation response alpha [Source:ZFIN;Acc:ZDB-GENE-090313-215]		

- 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 example.tsv`
- Will only show columns 1 (ID), 3 (adjusted p-value), 4 ($\log_2$ fold change) and 10 (name)
- -f is short for “fields”

ENSDARG00000086405	0.999562330126382	-0.000179405736027379	mterf2
ENSDARG00000039325	0.999609448070624	0.000131801057957922	pmepa1
ENSDARG00000057899	0.999700666108247	-0.000108106081918277	zgc:114081
ENSDARG00000105305	0.999700666108247	-0.000110968264902105	si:dkeyp-79b7.12
ENSDARG00000034146	0.999700666108247	7.28817147782303e-05	ccnk
ENSDARG00000071678	0.999700666108247	9.41692151597997e-05	atcaya
ENSDARG00000045827	0.999700666108247	9.1766658712139e-05	lyrm5b
ENSDARG00000093977	0.999700666108247	9.06145711124573e-05	si:ch211-220f16.1
ENSDARG00000012829	0.999700666108247	-8.79815865901993e-05	asah2
ENSDARG00000040177	0.999924882564427	2.90539977491282e-05	rgs16
ENSDARG00000089875	0.999924882564427	3.09165380424641e-05	si:ch211-226o13.2
ENSDARG00000097737	0.999965845994847	1.20374630567276e-05	si:dkey-188h10.3
ENSDARG00000029215	0.999965845994847	2.91914137791411e-06	ube2z

Pipe

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

```
GeneID  adjp    log2fc  Name
ENSDARG00000062262      0      -3.95687045483116      ednrab
ENSDARG00000045893      2.14308867730133e-240    -3.10954194931744      kctd15a
ENSDARG00000101407      1.95168628123782e-228    -3.38549363750172      tgm1l4
ENSDARG00000094041      1.8640728437731e-167     -2.89059848142129      krt17
ENSDARG00000077467      1.07027028896538e-147    -2.76987218334732      sox10
ENSDARG00000054616      1.33850110474841e-120    -2.19983456562198      cldni
ENSDARG00000034836      3.46890846377303e-112    -2.27545446343907      KRT84
ENSDARG00000059279      4.40623303727167e-104    -1.35095118552025      tfap2a
ENSDARG00000027611      2.67992898617429e-88     -1.72349259342463      sdpra
ENSDARG00000100190      7.39097494565812e-88     -1.99471938023915      ENSDARG00000100190
ENSDARG00000021720      8.73859902301245e-67     -2.3510727158879      col7a1
ENSDARG00000058371      4.38675578819498e-64     -1.44272671351009      krt5
```

column

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

GeneID	adjp	log2fc	Name
ENSDARG00000062262	0	-3.95687045483116	ednrab
ENSDARG00000045893	2.14308867730133e-240	-3.10954194931744	kctd15a
ENSDARG00000101407	1.95168628123782e-228	-3.38549363750172	tgm1l4
ENSDARG00000094041	1.8640728437731e-167	-2.89059848142129	krt17
ENSDARG00000077467	1.07027028896538e-147	-2.76987218334732	sox10
ENSDARG00000054616	1.33850110474841e-120	-2.19983456562198	cldni
ENSDARG00000034836	3.46890846377303e-112	-2.27545446343907	KRT84
ENSDARG00000059279	4.40623303727167e-104	-1.35095118552025	tfap2a
ENSDARG00000027611	2.67992898617429e-88	-1.72349259342463	sdpra
ENSDARG00000100190	7.39097494565812e-88	-1.99471938023915	ENSDARG00000100190
ENSDARG00000021720	8.73859902301245e-67	-2.3510727158879	col7a1
ENSDARG00000058371	4.38675578819498e-64	-1.44272671351009	krt5

head

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

```
[iansealy@kevin]~$ cut -f1,3-4,10 example.tsv | head -10 | column -t
```

GeneID	adjp	log2fc	Name
ENSDARG00000062262	0	-3.95687045483116	ednrab
ENSDARG00000045893	2.14308867730133e-240	-3.10954194931744	kctd15a
ENSDARG00000101407	1.95168628123782e-228	-3.38549363750172	tgm1l4
ENSDARG00000094041	1.8640728437731e-167	-2.89059848142129	krt17
ENSDARG00000077467	1.07027028896538e-147	-2.76987218334732	sox10
ENSDARG00000054616	1.33850110474841e-120	-2.19983456562198	cldni
ENSDARG00000034836	3.46890846377303e-112	-2.27545446343907	KRT84
ENSDARG00000059279	4.40623303727167e-104	-1.35095118552025	tfap2a
ENSDARG00000027611	2.67992898617429e-88	-1.72349259342463	sdpra

tail

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

```
[iansealy@kevin]~$ cut -f1,3-4,10 example.tsv | tail -10 | column -t
ENSDARG000000105305  0.999700666108247  -0.000110968264902105  si:dkeyp-79b7.12
ENSDARG00000034146  0.999700666108247  7.28817147782303e-05   ccnk
ENSDARG00000071678  0.999700666108247  9.41692151597997e-05   atcaya
ENSDARG00000045827  0.999700666108247  9.1766658712139e-05    lyrm5b
ENSDARG00000093977  0.999700666108247  9.06145711124573e-05   si:ch211-220f16.1
ENSDARG00000012829  0.999700666108247  -8.79815865901993e-05  asah2
ENSDARG00000040177  0.999924882564427  2.90539977491282e-05   rgs16
ENSDARG00000089875  0.999924882564427  3.09165380424641e-05   si:ch211-226o13.2
ENSDARG00000097737  0.999965845994847  1.20374630567276e-05   si:dkey-188h10.3
ENSDARG00000029215  0.999965845994847  2.91914137791411e-06   ube2z
```

AWK

- AWK is a powerful command line tool used for text processing
- Can filter based on a specific column
- `awk -F"\t" '$4 > 0' example.tsv | more`
(get up genes)
- `awk -F"\t" '$5 == "2"' example.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)

```
ENSDARG00000068177      4.63810835349325e-29    1.36173476748606e-26    -0.711981591430234      2      36626449      36654420      1      protein_coding pak2a  p21 pr
otein (Cdc42/Rac)-activated kinase 2a [Source:ZFIN;Acc:ZDB-GENE-021011-2]
ENSDARG00000034080      8.85999924805305e-29    2.56301537071252e-26    -0.994112437235758      2      21729842      21780493     -1      protein_coding plcd1b  phosph
olipase C, delta 1b [Source:ZFIN;Acc:ZDB-GENE-030131-9435]
ENSDARG00000020929      8.24258541036797e-21    1.37406692887583e-18    -0.997651870118026      2      32232054      32279069     -1      protein_coding fam49ba family
with sequence similarity 49, member Ba [Source:ZFIN;Acc:ZDB-GENE-030131-9537]
ENSDARG00000061797      2.91935296821382e-19    4.13140951350604e-17    -1.23764563292536      2      24730691      24746710     -1      protein_coding zgc:154006
zgc:154006 [Source:ZFIN;Acc:ZDB-GENE-061013-134]
ENSDARG00000052997      1.62779653205706e-18    2.09283565896043e-16    -0.831936360058288      2      49834750      49869451      1      protein_coding sema4e  semaph
orin 4e [Source:ZFIN;Acc:ZDB-GENE-990715-7]
ENSDARG00000024829      1.68668225980575e-17    1.97492421027613e-15    -1.17379354701359      2      35526961      35584559     -1      protein_coding tnw     tenasc
in W [Source:ZFIN;Acc:ZDB-GENE-990415-262]
ENSDARG00000014522      3.02356723410869e-17    3.41819488862943e-15    -0.611627666394501      2      28393410      28542893     -1      protein_coding cdh6    cadher
in 6 [Source:ZFIN;Acc:ZDB-GENE-050320-92]
```

Filtering on multiple columns with AWK

- Could just pipe two AWK commands together:

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

- But can combine terms with &&:

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

ENSDARG00000032317	1.13392905463136e-15	1.05713357505467e-13	0.409655914635066	2	22324851	22447193	1	protein_coding	tox	thymoc
yte selection-associated high mobility group box [Source:ZFIN;Acc:ZDB-GENE-070912-181]										
ENSDARG00000038868	1.33538727177764e-13	1.02610949309133e-11	0.581261532543583	2	29992886	29996423	1	protein_coding	eng2b	engrai
led homeobox 2b [Source:ZFIN;Acc:ZDB-GENE-980526-40]										
ENSDARG00000011166	2.515284866271e-12	1.68293090491214e-10	0.955601549885026	2	29493850	29501875	-1	protein_coding	cahz	carbon
ic anhydrase [Source:ZFIN;Acc:ZDB-GENE-980526-39]										
ENSDARG00000009922	1.73129044361296e-10	9.53955583089933e-09	0.570054706931402	2	10387510	10394667	-1	protein_coding	dmbx1a	dience
phalon/mesencephalon homeobox 1a [Source:ZFIN;Acc:ZDB-GENE-020117-1]										
ENSDARG00000054454	2.98755526857787e-10	1.59263413789147e-08	0.35165583572309	2	40102182	40170399	-1	protein_coding	epha4a	eph re
ceptor A4a [Source:ZFIN;Acc:ZDB-GENE-001207-7]										
ENSDARG00000031372	5.64089832243052e-10	2.87466608550598e-08	0.442448492823539	2	26647993	26788869	1	protein_coding	efna2a	ephrin
-A2a [Source:ZFIN;Acc:ZDB-GENE-990415-66]										

Combining AWK and cut

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

GeneID	pval	adjp	Name
ENSDARG00000098949	2.55735156723272e-59	2.39550774662071e-56	mslna
ENSDARG00000079543	4.64888943404216e-41	2.34482830915495e-38	dpys
ENSDARG00000101613	1.46907296449864e-31	4.89798886180556e-29	si:dkey-238i5.2
ENSDARG00000037362	8.54754892382638e-28	2.36815260395195e-25	scube2
ENSDARG00000010770	1.51256683420274e-22	2.83368592339068e-20	sox19a
ENSDARG00000000212	1.3204435366822e-21	2.34004007297978e-19	krt97
ENSDARG00000040266	2.48854724428456e-20	3.88509625732712e-18	sox19b
ENSDARG00000095896	7.26000907429691e-19	9.84907851727548e-17	pou3f3b
ENSDARG00000008540	3.1505427709026e-18	3.92242574977374e-16	sox21b
ENSDARG00000042032	1.25416668420929e-17	1.52288350895561e-15	pou3f3a
ENSDARG00000035735	2.39709429769081e-17	2.79013265857254e-15	gsx1
ENSDARG00000086645	2.9255003056043e-17	3.35892930804784e-15	hs3st3b1b

Reordering columns

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

GeneID	adjp	Name
ENSDARG00000098949	2.39550774662071e-56	mslna
ENSDARG00000079543	2.34482830915495e-38	dpys
ENSDARG00000101613	4.89798886180556e-29	si:dkey-238i5.2
ENSDARG00000037362	2.36815260395195e-25	scube2
ENSDARG00000010770	2.83368592339068e-20	sox19a
ENSDARG00000000212	2.34004007297978e-19	krt97
ENSDARG00000040266	3.88509625732712e-18	sox19b
ENSDARG00000095896	9.84907851727548e-17	pou3f3b
ENSDARG00000008540	3.92242574977374e-16	sox21b
ENSDARG00000042032	1.52288350895561e-15	pou3f3a
ENSDARG00000035735	2.79013265857254e-15	gsx1
ENSDARG00000086645	3.35892930804784e-15	hs3st3b1b

Replacing cut with AWK

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

GeneID	Name	pval	adjp	log2fc
ENSDARG00000098949	mslna	2.55735156723272e-59	2.39550774662071e-56	2.15193206980885
ENSDARG00000079543	dpys	4.64888943404216e-41	2.34482830915495e-38	1.7882009831906
ENSDARG00000101613	si:dkey-238i5.2	1.46907296449864e-31	4.89798886180556e-29	1.17099436067348
ENSDARG00000037362	scube2	8.54754892382638e-28	2.36815260395195e-25	0.702111405099621
ENSDARG00000010770	sox19a	1.51256683420274e-22	2.83368592339068e-20	0.520693336841397
ENSDARG00000000212	krt97	1.3204435366822e-21	2.34004007297978e-19	1.01050292091677
ENSDARG00000040266	sox19b	2.48854724428456e-20	3.88509625732712e-18	0.64066524668934
ENSDARG00000095896	pou3f3b	7.26000907429691e-19	9.84907851727548e-17	0.523017660639776
ENSDARG00000008540	sox21b	3.1505427709026e-18	3.92242574977374e-16	0.610438334797731
ENSDARG00000042032	pou3f3a	1.25416668420929e-17	1.52288350895561e-15	0.630453223905055
ENSDARG00000035735	gsx1	2.39709429769081e-17	2.79013265857254e-15	0.834111463466981
ENSDARG00000086645	hs3st3b1b	2.9255003056043e-17	3.35892930804784e-15	0.469061850694741
ENSDARG00000100398	pax7a	4.96157006549066e-17	5.3922124175838e-15	0.747678223566135

sort

- Reorder data using `sort`
- `cut -f1-4,10 example.tsv | sort -g -k4 | more`
- `-g` means sort numerically
- `-k4` means sort by the 4th column ($\log_2$ fold change)

```
[iansealy@kevin]~$ cut -f1-4,10 example.tsv | sort -g -k4 | more
GeneID  pval    adjp    log2fc  Name
ENSDARG00000062262      0      0      -3.95687045483116      ednrab
ENSDARG00000101407      2.97649272721949e-232  1.95168628123782e-228  -3.38549363750172      tgm1l4
ENSDARG00000045893      2.17893211051937e-244  2.14308867730133e-240  -3.10954194931744      kctd15a
ENSDARG00000094041      3.79049940272095e-171  1.8640728437731e-167   -2.89059848142129      krt17
ENSDARG00000077467      2.72042674232468e-151  1.07027028896538e-147  -2.76987218334732      sox10
ENSDARG00000021720      4.88661426735483e-70   8.73859902301245e-67   -2.3510727158879      col7a1
ENSDARG00000034836      1.23442424108643e-115  3.46890846377303e-112  -2.27545446343907      KRT84
ENSDARG00000045887      2.69888352962943e-64   3.53931586075604e-61   -2.24613380054078      mmp30
ENSDARG00000054616      4.0826631226122e-124   1.33850110474841e-120  -2.19983456562198      cldni
ENSDARG00000056938      8.51137458213534e-53   6.43950959250709e-50   -2.08512100355591      kera
```

More sort

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

```
[iansealy@kevin]~$ cut -f1-4,10 example.tsv | sort -r -g -k4 | more
ENSDARG00000098949      2.55735156723272e-59      2.39550774662071e-56      2.15193206980885      mslna
ENSDARG00000079543      4.64888943404216e-41      2.34482830915495e-38      1.7882009831906 dpys
ENSDARG000000101613     1.46907296449864e-31      4.89798886180556e-29      1.17099436067348      si:dkey-238i5.2
ENSDARG00000071724      3.66228530281467e-16      3.58412010903818e-14      1.11774161911335      ankha
ENSDARG000000052644     6.90812682992255e-15      6.06650727104493e-13      1.08640911310048      ca10a
ENSDARG00000000212     1.3204435366822e-21      2.34004007297978e-19      1.01050292091677      krt97
ENSDARG00000011166     2.515284866271e-12      1.68293090491214e-10      0.955601549885026      cahz
ENSDARG000000053858     4.47449020267656e-12      2.91449327075665e-10      0.955472570805111      cripl
ENSDARG000000103277     2.36974721866481e-11      1.42119809568157e-09      0.919061275611138      cyp24a1
ENSDARG000000091715     1.48775561310227e-10      8.29054976355094e-09      0.889231512508225      si:dkey-162h11.2
ENSDARG000000073814     9.44548579404264e-10      4.56516341657525e-08      0.854606347717019      pax1b
```

grep

- Extract data by search term using `grep`
- `grep ENSDARG000000068567 example.tsv`
- `grep shh example.tsv`

```
[iansealy@kevin]~$ grep shh example.tsv
ENSDARG000000068567      0.0815394303019272      0.335346463196573      -0.117374977553236      7      40603939      40611263      1      protein_coding  shha      sonic
hedgehog a [Source:ZFIN;Acc:ZDB-GENE-980526-166]
ENSDARG000000038867      0.378717792849533      0.718989845445756      -0.10682399807432      2      30067008      30071899      -1      protein_coding  shhb      sonic
hedgehog b [Source:ZFIN;Acc:ZDB-GENE-980526-41]
```

More grep

- Extract using list of search terms
- `grep -f myc-targets.tsv example.tsv | more`
- `-f` is short for “file”

```
ENSDARG00000007241      1.27358519946464e-08    5.197654452006e-07    -0.383442390814686    2      32033038      32035153      1      protein_coding mycb      v-myc
avian myelocytomatosis viral oncogene homolog b [Source:ZFIN;Acc:ZDB-GENE-040426-780]
ENSDARG000000104760      2.57975211425462e-05    0.000568904751563932  -0.407873211753083    16     35641749      35692535      1      protein_coding MAP3K6    mitoge
n-activated protein kinase kinase kinase 6 [Source:HGNC Symbol;Acc:HGNC:6858]
ENSDARG000000101482      0.000337757242625277    0.00545486265983729  -0.259014859645494    5      13370084      13503972      1      protein_coding hk2      hexoki
nase 2 [Source:ZFIN;Acc:ZDB-GENE-040426-2017]
ENSDARG00000004576      0.000468254237031145    0.00710937579176955    0.189445946580249    17     43420215      43437859     -1      protein_coding plk4      polo-l
ike kinase 4 (Drosophila) [Source:ZFIN;Acc:ZDB-GENE-030619-14]
ENSDARG00000006219      0.0730799474958417      0.312304215736613     -0.0687372326767907    11     3940108 3951648 1      protein_coding gnl3      guanine nucleotide bin
ding protein-like 3 (nucleolar) [Source:ZFIN;Acc:ZDB-GENE-030131-616]
ENSDARG000000074807      0.0803377228020051      0.331930969384214     0.103333070126629    16     27725383      27743307     -1      protein_coding tbrg4     transf
orming growth factor beta regulator 4 [Source:ZFIN;Acc:ZDB-GENE-091020-8]
```

Even more grep

- Exclude lines that match a search term
- `grep -v protein_coding example.tsv`
- `grep -v GeneID example.tsv`
- `-v` is short for “inVert”

```
ENSDARG00000032077      0.993495145486912      0.997921406806299      0.000977771296364643      8      20532536      20548497      1      unprocessed_pseudogene      mfsd12
b      major facilitator superfamily domain containing 12b [Source:ZFIN;Acc:ZDB-GENE-080305-1]
ENSDARG00000096338      0.994840031123885      0.99836280621833      0.000684356149280859      3      28564920      28568014      -1      processed_transcript      si:dke
y-20j1.4      si:dkey-20j1.4 [Source:ZFIN;Acc:ZDB-GENE-120214-35]
ENSDARG00000096996      0.995973918605368      0.998644225811287      0.000591271189117843      24      17259051      17261947      -1      lincRNA      si:dkey-117j14.7      s
i:dkey-117j14.7 [Source:ZFIN;Acc:ZDB-GENE-131120-91]
ENSDARG00000099925      0.996105213795442      0.998644225811287      0.000616252165872914      11      24342963      24343498      1      processed_transcript      si:dke
y-27b23.3      si:dkey-27b23.3 [Source:ZFIN;Acc:ZDB-GENE-141216-227]
ENSDARG00000105439      0.997739510911278      0.999263475339125      -0.000364213561967776      7      5907437      5914756      -1      lincRNA      si:dkey-23a13.3      si:dkey-23a13.3 [Sourc
e:ZFIN;Acc:ZDB-GENE-160113-8]
ENSDARG00000080718      0.998780664055026      0.999492010104615      0.000175815337682424      MT      6038      6107      1      Mt_tRNA      AC024175.7
ENSDARG00000097737      0.999920112579403      0.999965845994847      1.20374630567276e-05      3      55223549      55224870      1      lincRNA      si:dkey-188h10.3      s
i:dkey-188h10.3 [Source:ZFIN;Acc:ZDB-GENE-030131-376]
```

WC

- `wc` stands for "word count"
- Count number of lines returned
- `wc -l myc-targets.tsv`
- `grep -f myc-targets.tsv example.tsv | wc -l`
- `-l` is short for "lines"

```
[iansealy@kevin]~$ wc -l myc-targets.tsv
57 myc-targets.tsv
[iansealy@kevin]~$ grep -f myc-targets.tsv example.tsv | wc -l
56
```

Redirecting to a file

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

```
[iansealy@kevin]~$ grep hox example.tsv > hox.txt
[iansealy@kevin]~$ more hox.txt
ENSDARG00000091029      4.47809533925821e-10    2.34902969116129e-08    0.708800277760977      14      16759662      16763059      -1      protein_coding  phox2bb paired
-like homeobox 2bb [Source:ZFIN;Acc:ZDB-GENE-050407-3]
ENSDARG00000096956      5.1520364446755e-05     0.00105021460003328    0.485050363804754      23      35982622      35989431      -1      antisense      hoxc10a
ENSDARG00000008174      0.000100423601082204    0.00189217687441382    0.318658284190981      3       23638350      23640594      1       protein_coding  hoxb1a  homeob
ox B1a [Source:ZFIN;Acc:ZDB-GENE-990415-101]
```

Appending to a file

- `head -1 example.tsv > hox.txt`
- `grep hox example.tsv >> hox.txt`
- `more hox.txt`

```
[iansealy@kevin]~$ head -1 example.tsv > hox.txt
[iansealy@kevin]~$ grep hox example.tsv >> hox.txt
[iansealy@kevin]~$ more hox.txt
GeneID pval adjp log2fc Chr Start End Strand Biotype Name Description
ENSDARG00000091029 4.47809533925821e-10 2.34902969116129e-08 0.708800277760977 14 16759662 16763059 -1 protein_coding phox2bb paired
-like homeobox 2bb [Source:ZFIN;Acc:ZDB-GENE-050407-3]
ENSDARG00000096956 5.1520364446755e-05 0.00105021460003328 0.485050363804754 23 35982622 35989431 -1 antisense hoxc10a
ENSDARG00000008174 0.000100423601082204 0.00189217687441382 0.318658284190981 3 23638350 23640594 1 protein_coding hoxb1a homeob
ox B1a [Source:ZFIN;Acc:ZDB-GENE-990415-101]
```

Thank You

Any Questions?