Advanced UNIX

Course: Work with genomic data in Unix

April 2015

Václav Janoušek, Libor Mořkovský

http://ngs-course.readthedocs.org/en/praha-april-2015/

This Session Tasks:

1. How many records in the file
2. Explore in detail the last 'group' column (column 9)
3. Get list of chromosomes (column 1)
4. Get list of features (column 3)
5. Get the number of genes mapping to the assembly in total
6. Get the number of protein coding genes mapping to the assembly
7. Get the number of protein coding genes on chromosomes X and Y
8. Get the number of transcripts of protein coding genes mapping to the assembly
9. Get the gene with the highest number of transcripts
10. Get the gene with the highest number of exons
11. Get the total size (in Mb) of coding sequences
12. Get the longest gene

Task 1: revision of morning skills

- Task 1: Number of records in the GTF file

```
wc -l Mus_musculus.NCBIM37.67.gtf
cat Mus_musculus.NCBIM37.67.gtf | wc -l
```

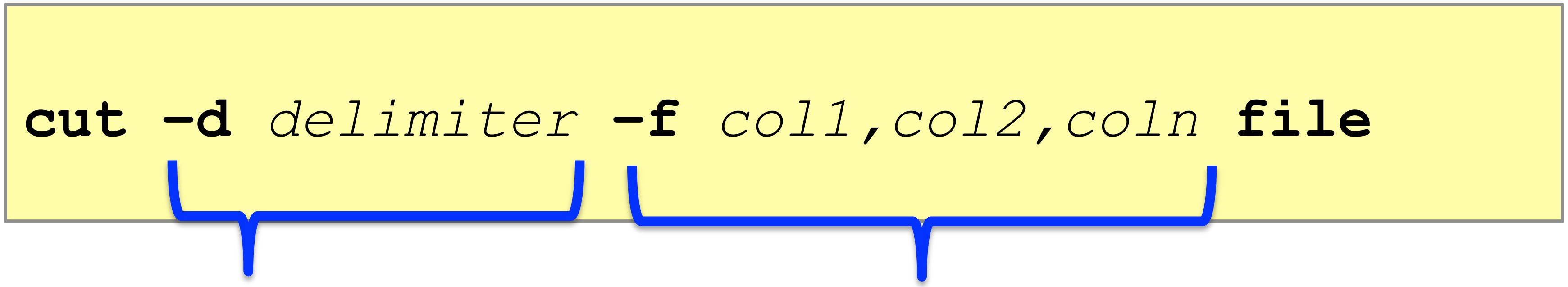
Answer:
1,277,861

Advanced build-in UNIX commands

```
cut
sort
uniq
grep
tr
sed
awk
```

cut

- selection specific portion of data
- use:



```
cut -d delimiter -f col1,col2,coln file
```

Column separator
(TAB as default)

Column(s) aimed to be selected.
Columns number from 1 to n.

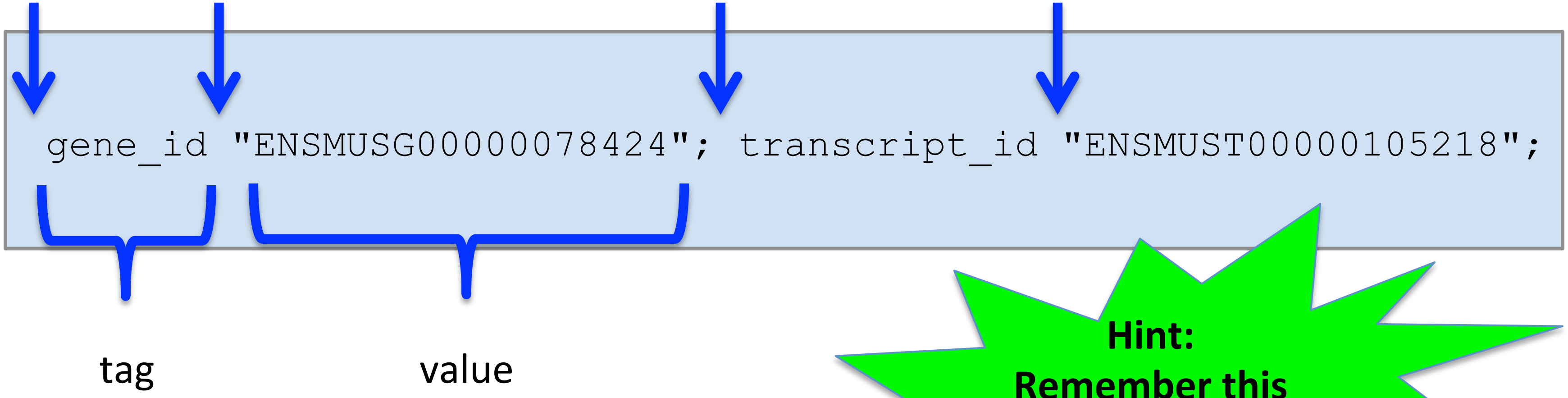
cut

- Task 2: Explore 'group' column (column 9)

```
cut –f 9 Mus_musculus.NCBIM37.67.gtf | less -S
```

'group/attribute' column in the GTF file

- composed of fields delimited by semicolon ';'
- each field composed of 'tag' and 'value' delimited by blank character ' '
- values surrounded by double quotes "value"

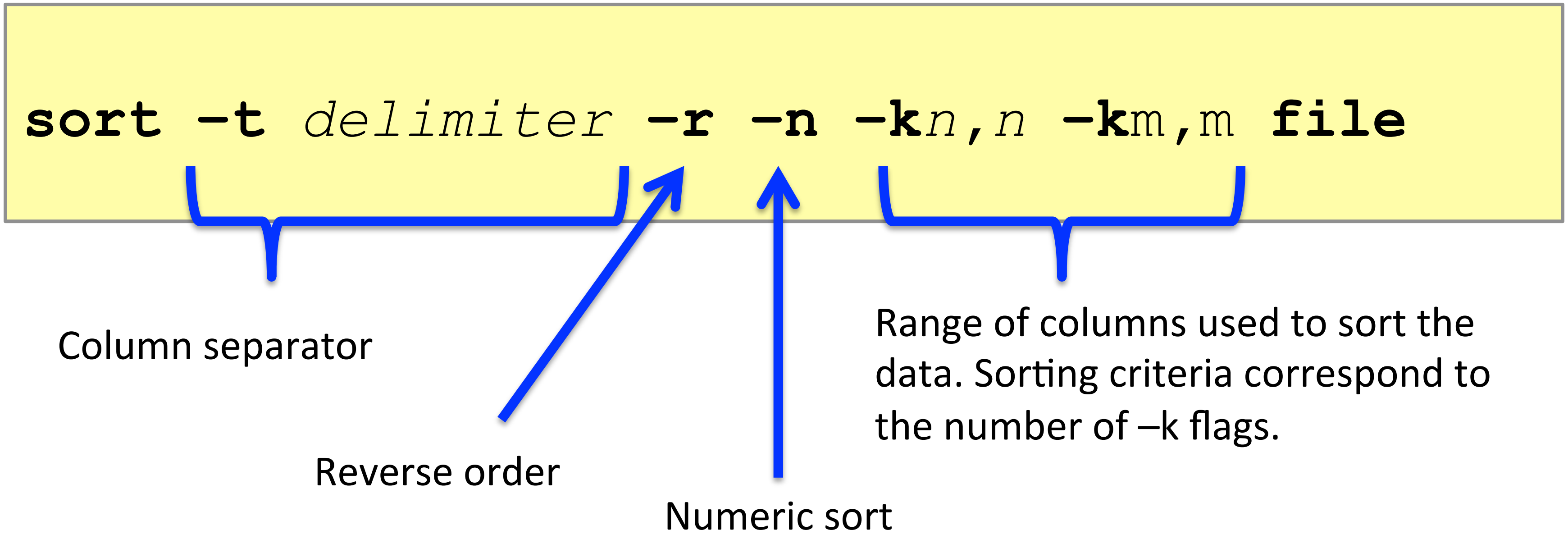


```
gene_id "ENSMUSG00000078424"; transcript_id "ENSMUST00000105218";
```

Hint:
Remember this
- we will use it
later!!!

sort

- sort data
- use:



```
sort -t delimiter -r -n -kn,n -km,m file
```

Column separator

Reverse order

Numeric sort

Range of columns used to sort the data. Sorting criteria correspond to the number of –k flags.

uniq

- returns unique list of records
- data needs to be sorted – usually used along with sort
- use:

```
sort file | uniq -c
```

Returns a column with
counts of unique records

cut|sort|uniq

- Task 3: Get list of chromosomes (column 1)

```
cut –f 1 Mus_musculus.NCBIM37.67.gtf | sort
| uniq
```

cut|sort|uniq

- Task 4: Get list of feature (column 3)

```
cut -f 3 Mus_musculus.NCBIM37.67.gtf | sort
| uniq
```

grep

- pattern specification & matching
- use:

```
grep pattern file # match lines having a pattern
```

```
grep -v pattern file # match lines not having a
pattern
```

`grep`: Regular expressions

- matching string patterns according to certain rules

```
^A # match A at the beginning of line
A$ # match A at the end of line
[0-9] # match numerical characters
[A-Z] # match alphabetical characters
[ATGC] # match A or T or G or C
. # match any character
A* # match A letter 0 or more times
A\{2\} # match A letter exactly 2 times
A\{1,\} # match A letter 1 or more times
A\{1,3\} # match A letter at least 1 times but no
more than 3 times
AATT\|TTAA # match AATT or TTAA
```

grep|cut|cut|sort|uniq|wc

- Task 5: Get the number of genes mapping onto chromosomes

```
grep -v ^NT Mus_musculus.NCBIM37.67.gtf |
cut -f 9 | cut -d ';' -f 1 | sort | uniq |
wc -l
```

Answer:
37,620

grep|grep|cut|cut|sort|uniq|wc

- Task 6: Get the number of protein coding genes mapping onto chromosomes

```
grep –v ^NT Mus_musculus.NCBIM37.67.gtf |
grep protein_coding |
cut –f 9 | cut –d ';' –f 1 | sort | uniq |
wc -l
```

Answer:
22,388

grep|grep|cut|cut|sort|uniq|cut|sort|uniq

- Task 7: Get the number of protein coding genes on chromosomes X and Y

```
grep ^[XY] Mus_musculus.NCBIM37.67.gtf |
grep protein_coding |
cut -f 1,9 | cut -d ';' -f 1 | sort | uniq |
cut -f 1 | sort | uniq -c
```

Answer:
X: 930
Y: 17

grep|grep|cut|cut|sort|uniq|wc

- Task 8: Get the number of transcripts of protein coding genes mapping onto chromosomes

```
grep -v ^NT Mus_musculus.NCBIM37.67.gtf |
grep protein_coding |
cut -f 9 | cut -d ';' -f 2 | sort | uniq |
wc -l
```

Answer:
79,675

tr

- replaces/removes individual characters
- use:

```
tr pattern1 pattern2 file # replace
tr -d pattern file # remove
```

`sed`: Text stream editor

- rich functionality
- matching and replacing more complex patterns
- using regular expressions
- use:

```
sed 's/pattern1/pattern2/g' file
```

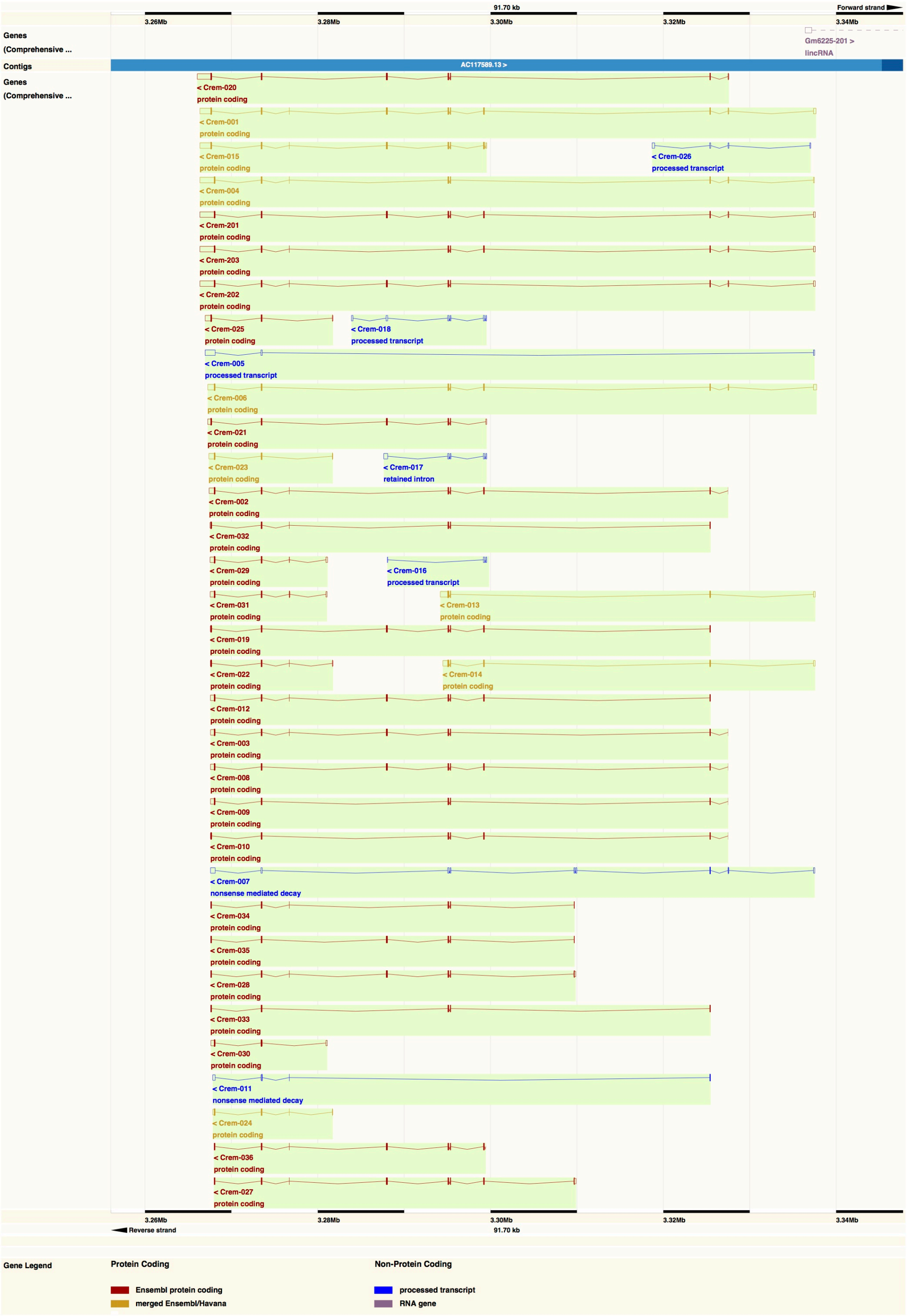
```
grep|grep|cut|cut|tr|sort|uniq|cut|uniq|sed|tr|sort|head
```

- Task 9: Get the gene with the highest number of transcripts

```
grep -v ^NT Mus_musculus.NCBIM37.67.gtf |
grep protein_coding |
cut -f 9 | cut -d " " -f 3,5,9 |
tr -d '";' | sort -k1,1 | uniq |
cut -d ' ' -f 1,3 | uniq -c | sed 's/^ *//' |
tr ' ' "\t" | sort -nr -k1,1 | head
```

Crem gene

- cAMP responsive element modulator
- 41 transcripts



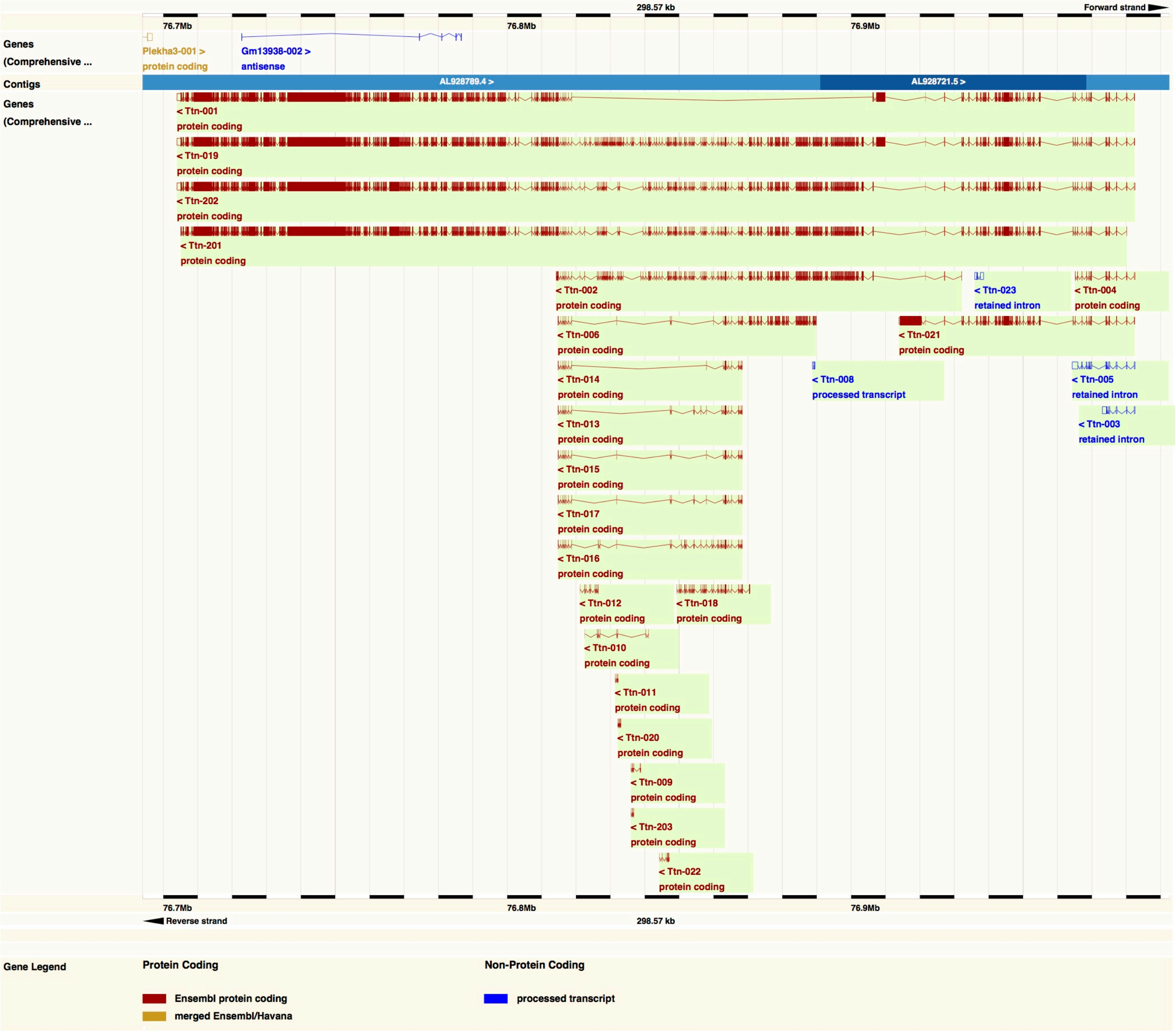
```
grep|grep|cut|cut|tr|sort|uniq|cut|uniq|sed|tr|sort|head
```

- Task 10: Get the gene with the highest number of exons

```
grep -v ^NT Mus_musculus.NCBIM37.67.gtf |
grep protein_coding | grep exon | cut -f 9 |
cut -d " " -f 3,5,9 | tr -d '";' | sort |
uniq -c | sed 's/^ *//g' | tr " " "\t" |
sort -rn -k1,1 | head
```

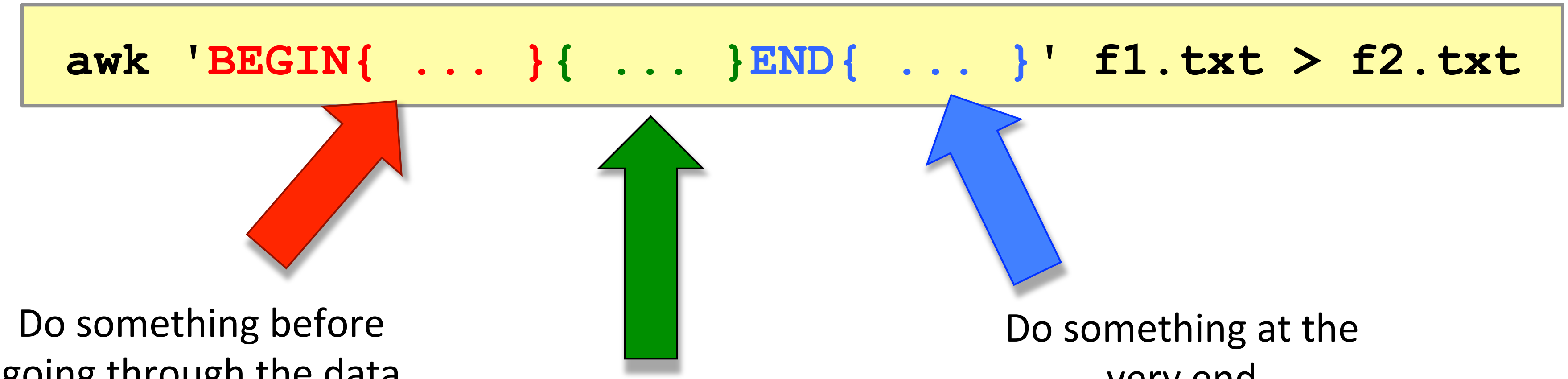
Tnt gene

- Titin
- 695 exons



awk: Scripting in one line

- simple programming language to do a complex data manipulation



```
awk 'BEGIN{ ... }{ ... }END{ ... }' f1.txt > f2.txt
```

Do something before going through the data

Do something while going through the data => for each line/record separately

Do something at the very end

awk: Exercise

- Task 11: Get the total size (in Mb) of coding sequences

```
# BEGIN part
grep CDS Mus_musculus.NCBIM37.67.gtf |
awk -F $'\t' 'BEGIN{OFS=FS;t=0}{...}END{...}'
```

AWK Built-in variables:
FS = Field Separator
OFS = Output Field Separator

awk: Exercise

- Task 11: Get the total size (in Mb) of coding sequences

```
# MIDDLE part
grep CDS Mus_musculus.NCBIM37.67.gtf |
awk -F $'\t' 'BEGIN{OFS=FS;t=0}
{s=$5-$4+1;t+=s}END{...}'
```

awk: Exercise

- Task 11: Get the total size (in Mb) of coding sequences

```
# END part
grep CDS Mus_musculus.NCBIM37.67.gtf |
awk -F $'\t' 'BEGIN{OFS=FS;t=0}{s=$5-$4+1;t+=s}
END{print t/1000000" Mb"}'
```

awk: Exercise

- Task 11: Get the total size (in Mb) of coding sequences

```
# END part
grep CDS Mus_musculus.NCBIM37.67.gtf |
awk -F $'\t' 'BEGIN{OFS=FS;t=0}{s=$5-$4+1;t+=s}
END{print t/1000000" Mb"}'
```

Answer:
78.2556 Mb

grep|grep|cut|cut|sed|sort + awk|sort|head

- Task 12: Get the longest gene

```
grep protein_coding Mus_musculus.NCBIM37.67.gtf | grep
exon | cut -f 1,4,5,9 | cut -d " " -f 1,3 |
sed 's/[";]//g' | sort -k4,4 -k2,2n > exons.bed

< exons.bed awk -F $'\t' 'BEGIN{ OF  =FS }   (    1)
{ gene=$4; chrom=$1; gene_start=$2;    e
else{ if(gene==$4){if(gene_end<=$3){
else{ print gene,chrom,gene_sta
gene_start; gene=$4;chrom=$1;gen
$3; }}}END{print gene, chrom, gen
gene_end-gene_start }' | sort
```

Hint:
Find the start of the first exon and the end of the last exon of a gene and get them on the same line...

grep|grep|cut|cut|sed|sort + awk|sort|head

- Task 12: Get the longest gene

```
grep protein_coding Mus_musculus.NCBIM37.67.gtf |
grep exon | cut -f 1,4,5,9 | cut -d " " -f 1,3 |
tr -d '";' | sort -k4,4 -k2,2n > exons.bed
```

`grep|grep|cut|cut|sed|sort + awk|sort|head`

- Task 12: Get the longest gene

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}...'
```

`grep|grep|cut|cut|sed|sort + awk|sort|head`

- Task 12: Get the longest gene

<u>AWK Built-in variables:</u>

NR = Number Record

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{
    if(NR==1){
        gene=$4; chrom=$1; gene_start=$2; gene_end=$3
    }else{
        if(gene==$4){
            if(gene_end<=$3){gene_end=$3}
        }else{
            print gene,chrom,gene_start,gene_end,gene_end-
gene_start;
            gene=$4; chrom=$1; gene_start=$2; gene_end=$3;
        }
    }
}
END{print gene,chrom,gene_start,gene_end,gene_end-gene_start}'
```

grep|grep|cut|cut|sed|sort + awk|sort|head

- Task 12: Get the longest gene

AWK Built-in variables:

NR = Number Record

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{
    if(NR==1){
        gene=$4; chrom=$1; gene_start=$2; gene_end=$3
    }else{
        if(gene==$4){
            if(gene_end<=$3){gene_end=$3}
        }else{
            print gene,chrom,gene_start,gene_end,gene_end-
gene_start;
            gene=$4; chrom=$1; gene_start=$2; gene_end=$3;
        }
    }
}
END{print gene,chrom,gene_start,gene_end,gene_end-gene_start}'
```

Assign the first gene

`grep|grep|cut|cut|sed|sort + awk|sort|head`

- Task 12: Get the longest gene

<u>AWK Built-in variables:</u>
NR = Number Record

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{
    if(NR==1){
        gene=$4; chrom=$1; gene_start=$2; gene_end=$3
    }else{
        if(gene==$4){
            if(gene_end<=$3){gene_end=$3}
        }else{
            print gene,chrom,gene_start,gene_end,gene_end-
gene_start;
            gene=$4; chrom=$1; gene_start=$2; gene_end=$3;
        }
    }
}
END{print gene,chrom,gene_start,gene_end,gene_end-gene_start}'
```

Assign new end for the same gene

```
grep|grep|cut|cut|sed|sort + awk|sort|head
```

- Task 12: Get the longest gene

AWK Built-in variables:

NR = Number Record

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{
    if(NR==1){
        gene=$4; chrom=$1; gene_start=$2; gene_end=$3
    }else{
        if(gene==$4){
            if(gene_end<=$3){gene_end=$3}
        }else{
            print gene,chrom,gene_start,gene_end,gene_end-
gene_start;
            gene=$4; chrom=$1; gene_start=$2; gene_end=$3;
        }
    }
}
END{print gene,chrom,gene_start,gene_end,gene_end-gene_start}'
```

Next gene; print the previous one

grep|grep|cut|cut|sed|sort + awk|sort|head

AWK Built-in variables:

NR = Number Record

- Task 12: Get the longest gene

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{
    if(NR==1){
        gene=$4; chrom=$1; gene_start=$2; gene_end=$3
    }else{
        if(gene==$4){
            if(gene_end<=$3){gene_end=$3}
        }else{
            print gene,chrom,gene_start,gene_end,gene_end-
gene_start;
            gene=$4; chrom=$1; gene_start=$2; gene_end=$3;
        }
    }
}
END{print gene,chrom,gene_start,gene_end,gene_end-gene_start}'
```

Assign the next gene

```
grep|grep|cut|cut|sed|sort + awk|sort|head
```

AWK Built-in variables:

NR = Number Record

- Task 12: Get the longest gene

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{
    if(NR==1){
        gene=$4; chrom=$1; gene_start=$2; gene_end=$3
    }else{
        if(gene==$4){
            if(gene_end<=$3){gene_end=$3}
        }else{
            print gene,chrom,gene_start,gene_end,gene_end-
gene_start;
            gene=$4; chrom=$1; gene_start=$2; gene_end=$3;
        }
    }
}
END{print gene,chrom,gene_start,gene_end,gene_end-gene_start}
```

Print the last gene

grep|grep|cut|cut|sed|sort + awk|sort|head

- Task 12: Get the longest gene

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{if(NR==1)
{gene=$4; chrom=$1; gene_start=$2; gene_end=$3}
else{if(gene==$4){if(gene_end<=$3){gene_end=$3}}
else{print gene, chrom,gene_start,gene_end,gene_end-
gene_start;gene=$4;chrom=$1;gene_start=$2;gene_end=
$3;}}}END{print gene, chrom, gene_start, gene_end,
gene_end-gene_start}' | sort –rn -k5,5 | head
```

```
grep|grep|cut|cut|sed|sort + awk|sort|head
```

- Task 12: Get the longest gene

```
< exons.bed awk -F $'\t' 'BEGIN{OFS=FS}{if(NR==1)
{gene=$4; chrom=$1; gene_start=$2; gene_end=$3}
else{if(gene==$4){if(gene_end<=$3){gene_end=$3}}
else{print gene, chrom,gene_start,gene_end,gene_end-
gene_start;gene=$4;chrom=$1;gene_start=$2;gene_end=
$3;}}}END{print gene, chrom, gene_start, gene_end,
gene_end-gene_start}' | sort -rn -k5,5 | head
```

Answer: Gm20388 (predicted gene)

That's all for today...