Preprocessing genomic data

HackBio

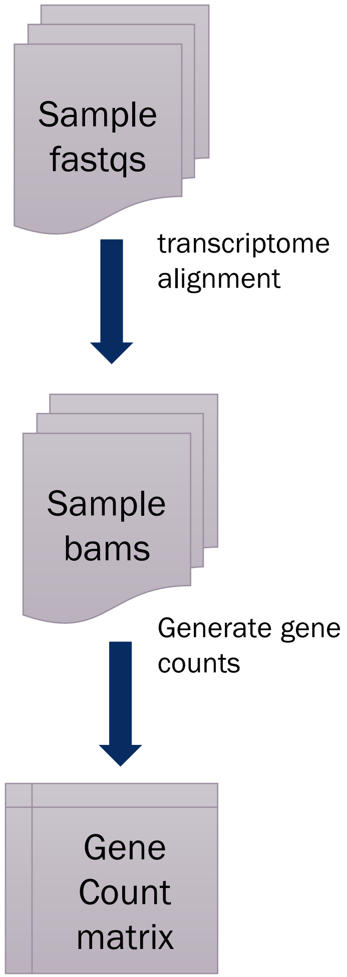
Melyssa Minto

West Lab, Duke Neurobiology

Computational Biology and Bioinformatics

Transcriptomics pre-proccesing workflow

Preprocessing



Sample fastqs

transcriptome alignment

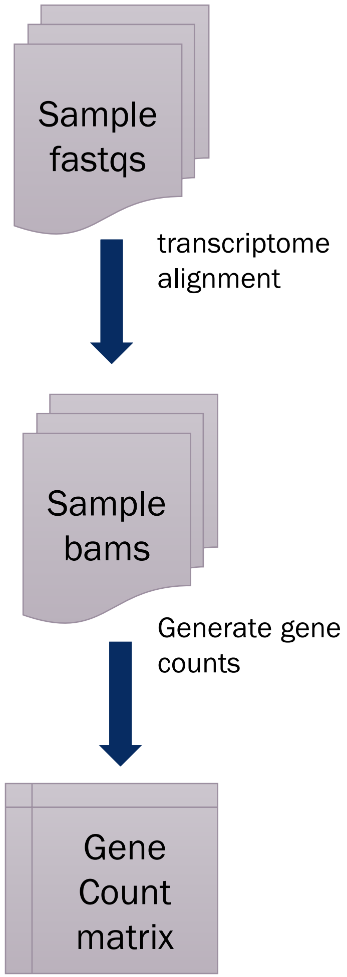
Sample bams

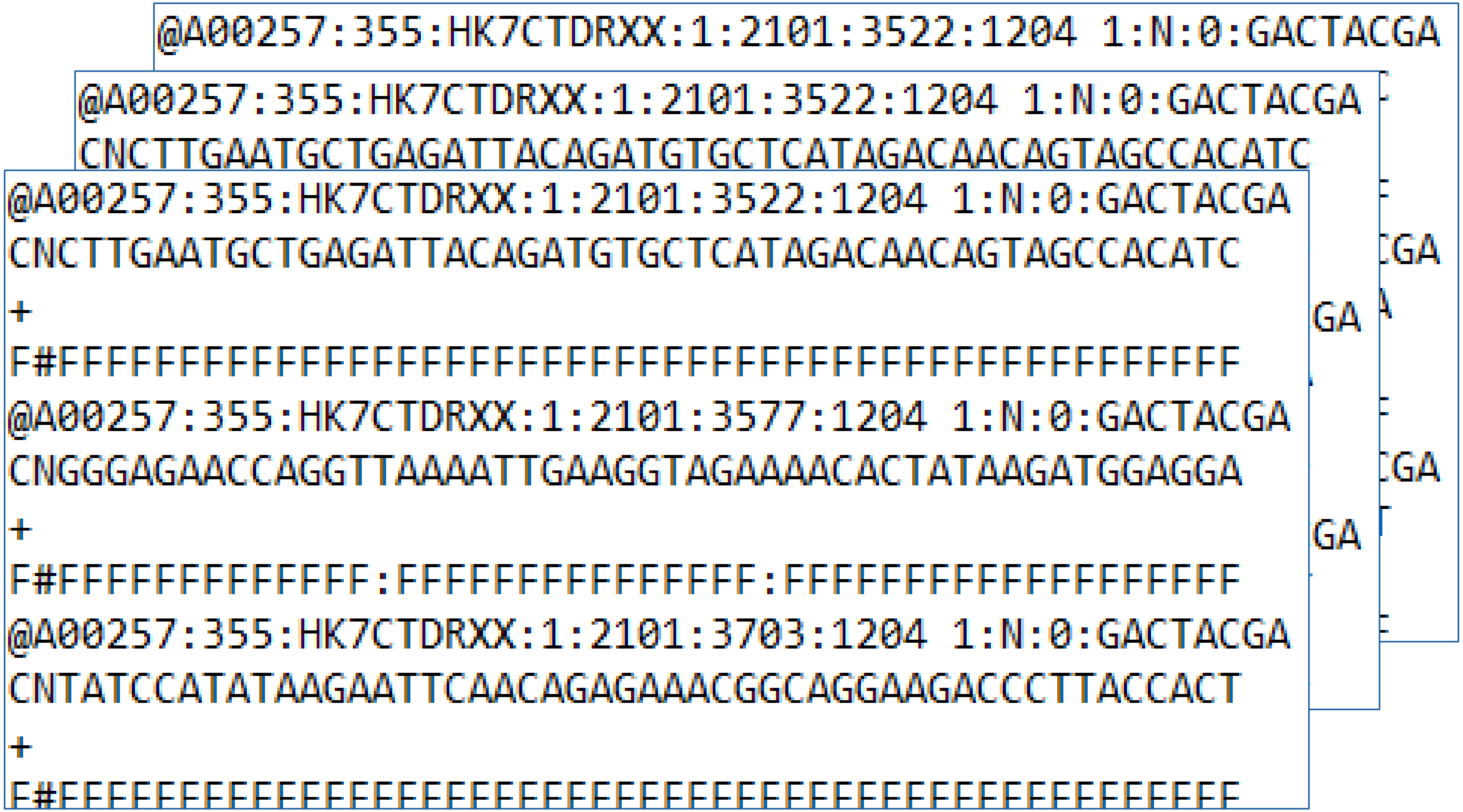
Generate gene counts

Gene Count matrix

Transcriptomics pre-proccesing workflow

Preprocessing

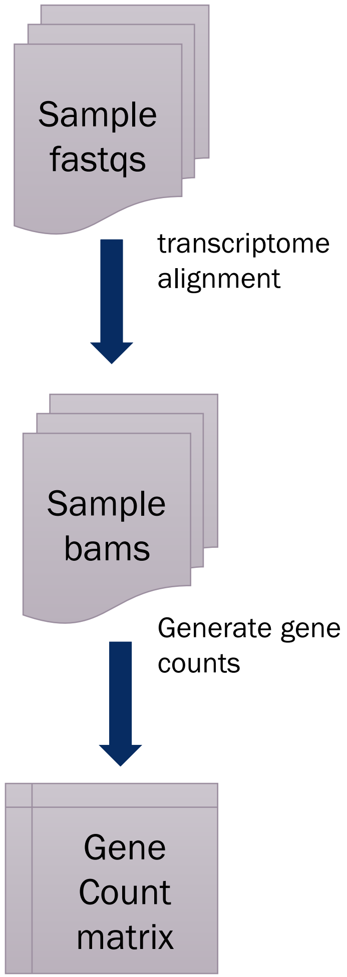




```
@A00257:355:HK7CTDRXX:1:2101:3522:1204 1:N:0:GACTACGA
CNCTTGAATGCTGAGATTACAGATGTGCTCATAGACAACAGTAGCCACATC
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3577:1204 1:N:0:GACTACGA
CNGGGAGAACCAGGTTAAAATTGAAGGTAGAAAACACTATAAGATGGAGGA
+
F#FFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3703:1204 1:N:0:GACTACGA
CNTATCCATATAAGAATTCAACAGAGAAACGGCAGGAAGACCCTTACCACT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Transcriptomics pre-proccesing workflow

Preprocessing



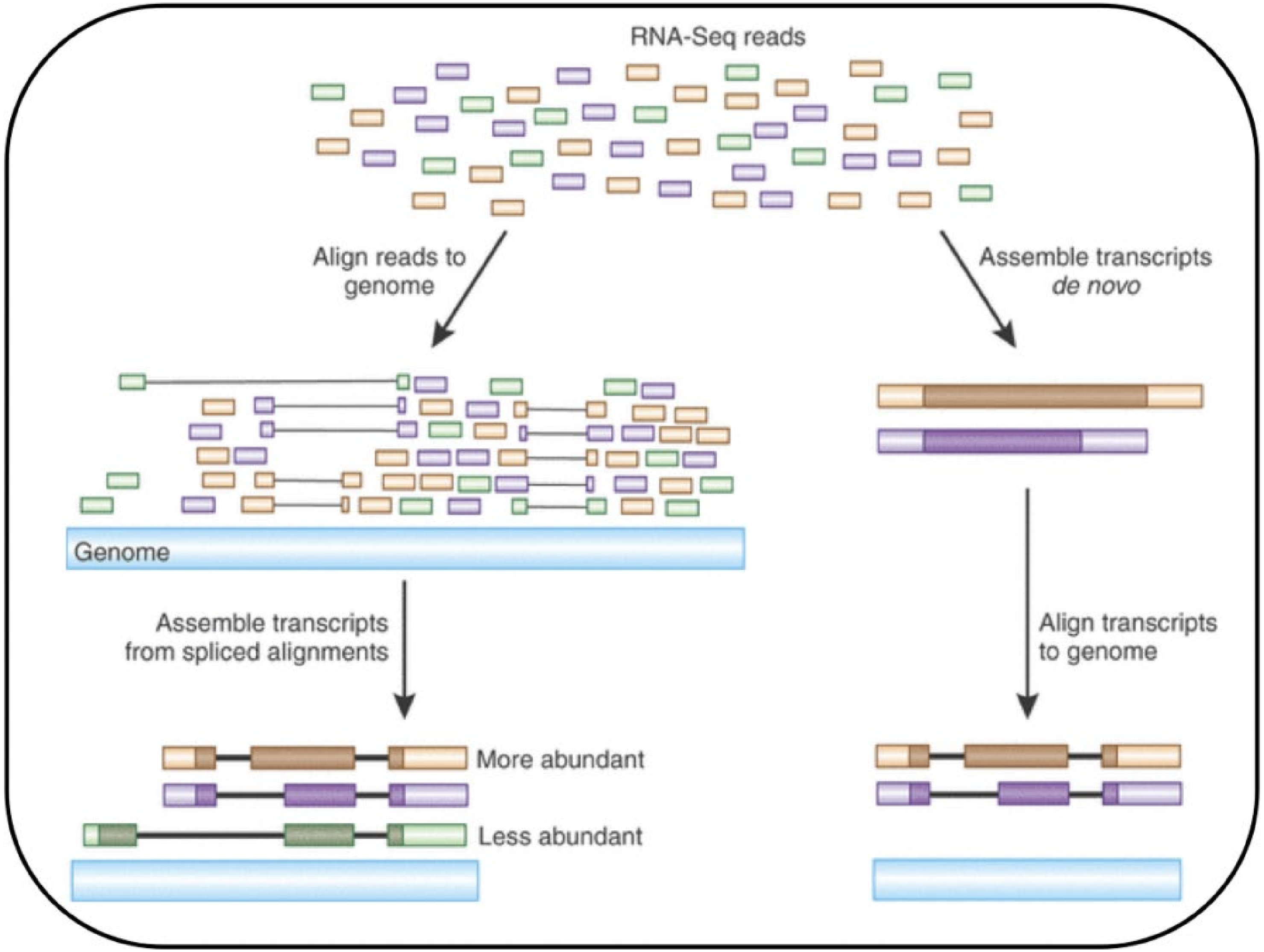
Sample fastqs

transcriptome alignment

Sample bams

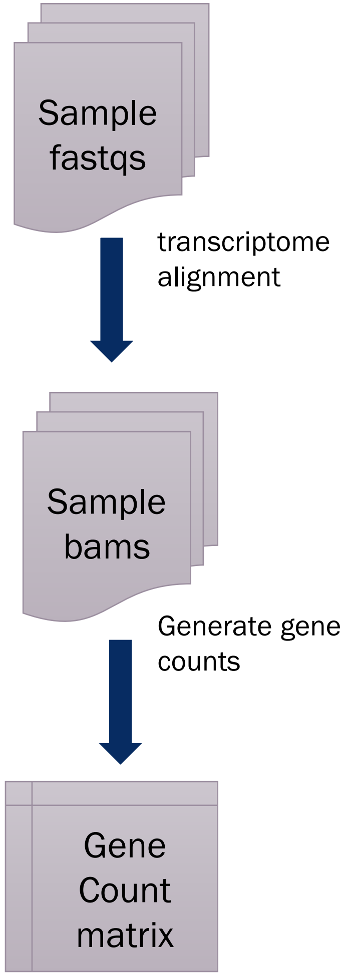
Generate gene counts

Gene Count matrix


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Transcriptomics pre-proccesing workflow

Preprocessing



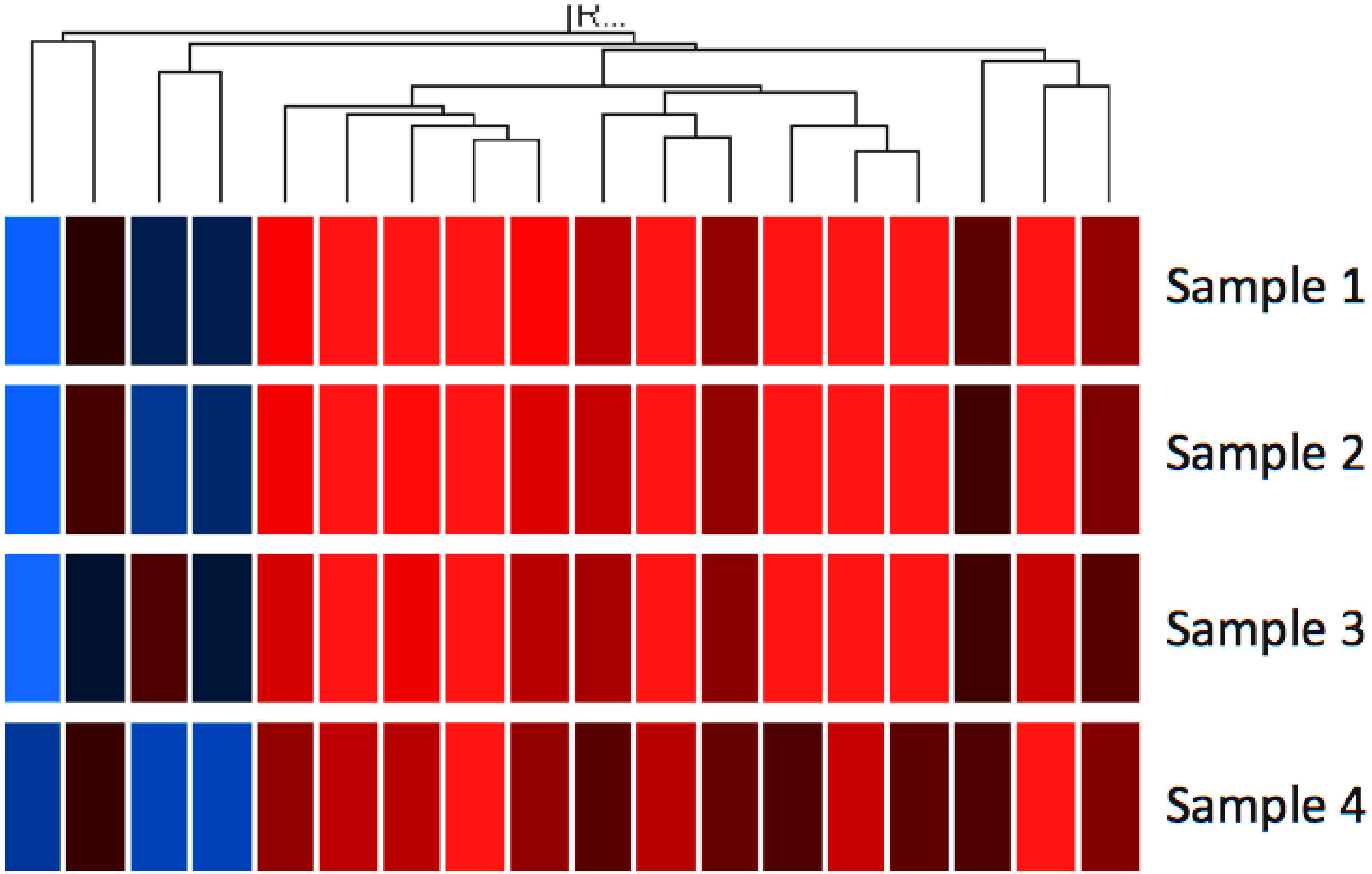
Sample fastqs

transcriptome alignment

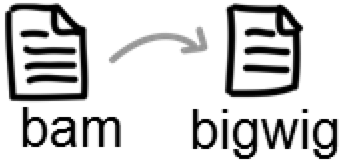
Sample bams

Generate gene counts

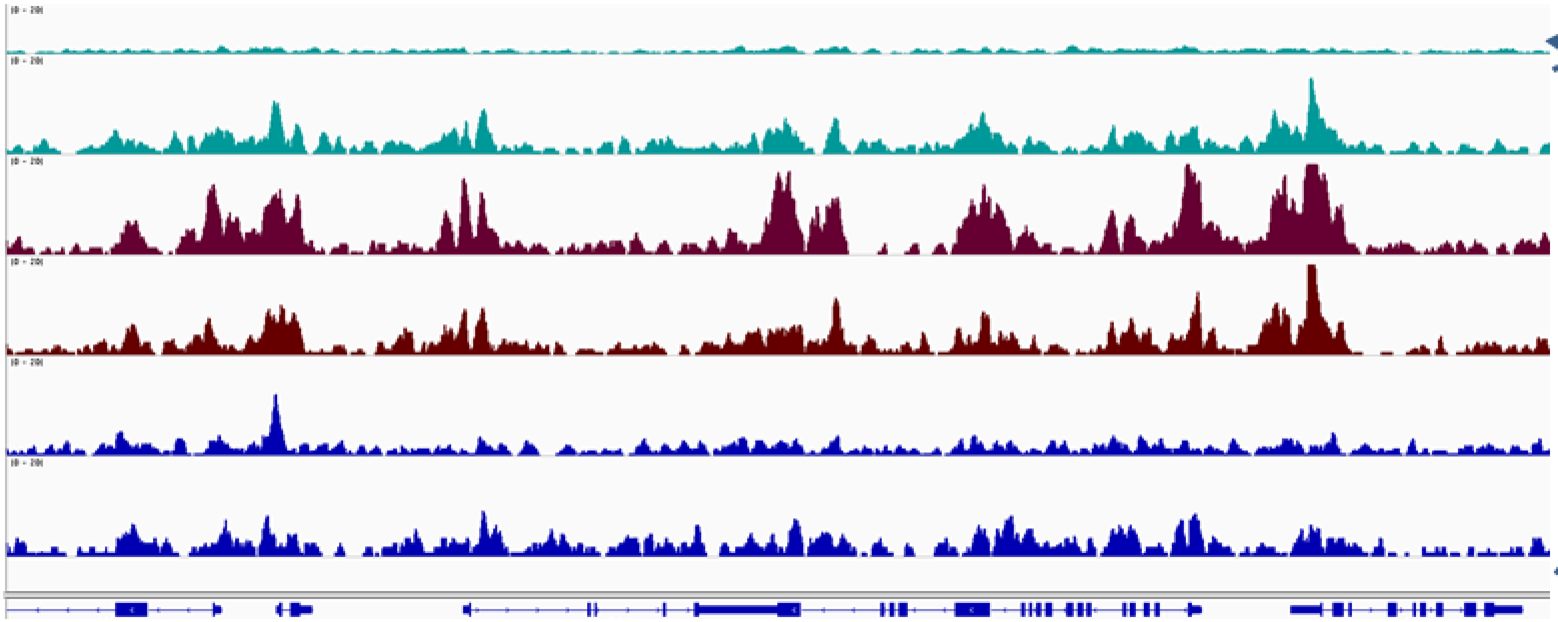
Gene Count matrix



Visualizing aligned genomic data



for visualizing continuous data, e.g. in the UCSC Genome Browser or IGV, bigWig files come in really handy



input

different ChIP-seq samples

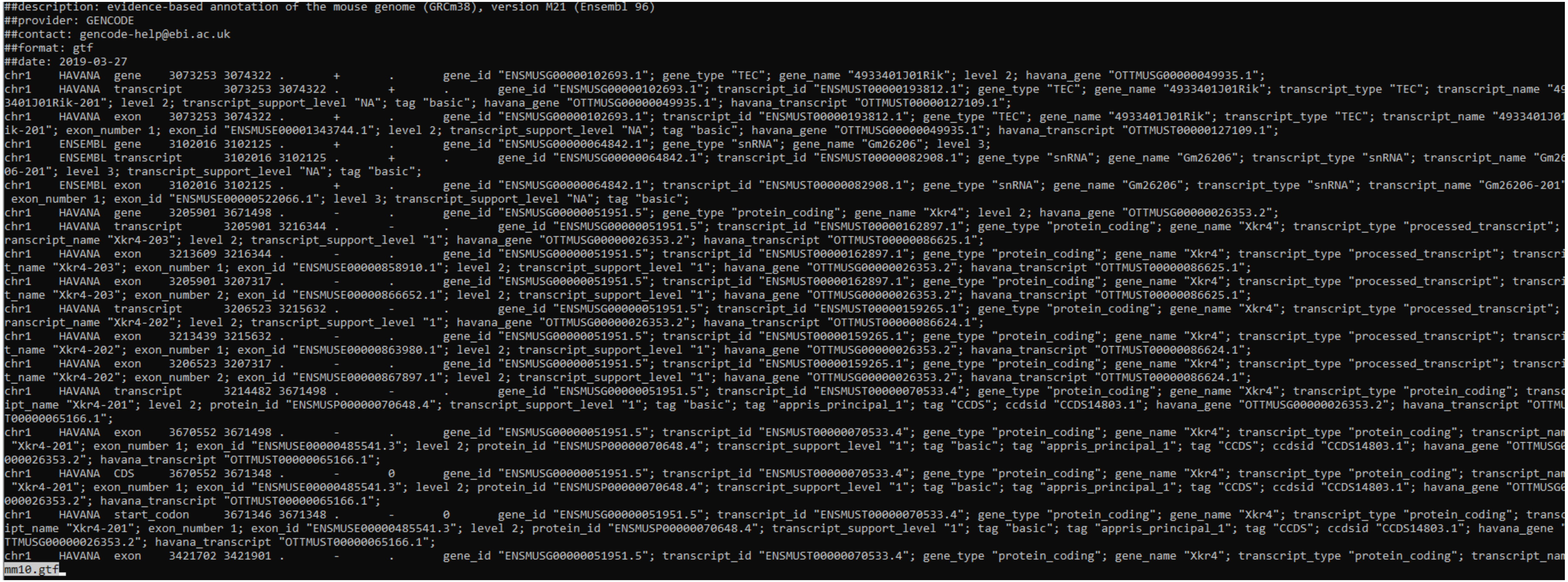
remember that there are 2 deepTools for bam → bigWig conversion:

- ***bamCoverage:*** *for individual files (like those shown here)*
- ***bamCompare:*** *to normalize two files to each other*

Quality Control of Alignments

- Mapping logs
- samtools flagstat <bam>
- Read depth

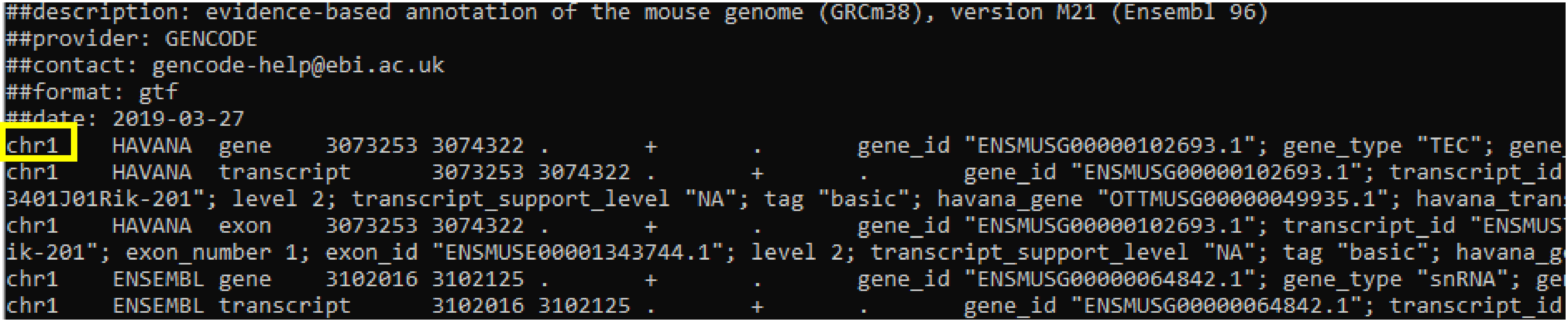
Extracting reads – what is a gtf file?



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_name "4933401J01Rik"; level 2; havana_gene "OTTMUSG00000049935.1";
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcript_name "49
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcript_name "4933401J01
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; gene_name "Gm26206"; level 3;
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_name "Gm26206"; transcript_type "snRNA"; transcript_name "Gm26
06-201"; level 3; transcript_support_level "NA"; tag "basic";
chr1    ENSEMBL exon    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_name "Gm26206"; transcript_type "snRNA"; transcript_name "Gm26206-201"
 exon_number 1; exon_id "ENSMUSE00000522066.1"; level 3; transcript_support_level "NA"; tag "basic";
chr1    HAVANA  gene    3205901 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; gene_type "protein_coding"; gene_name "Xkr4"; level 2; havana_gene "OTTMUSG00000026353.2";
chr1    HAVANA  transcript      3205901 3216344 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript";
ranscript_name "Xkr4-203"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  exon    3213609 3216344 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-203"; exon_number 1; exon_id "ENSMUSE00000858910.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  exon    3205901 3207317 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-203"; exon_number 2; exon_id "ENSMUSE00000866652.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  transcript      3206523 3215632 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript";
ranscript_name "Xkr4-202"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  exon    3213439 3215632 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-202"; exon_number 1; exon_id "ENSMUSE00000863980.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  exon    3206523 3207317 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-202"; exon_number 2; exon_id "ENSMUSE00000867897.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  transcript      3214482 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transc
ipt_name "Xkr4-201"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMU
T00000065166.1";
chr1    HAVANA  exon    3670552 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_nam
 "Xkr4-201"; exon_number 1; exon_id "ENSMUSE00000485541.3"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG
000026353.2"; havana_transcript "OTTMUST00000065166.1";
chr1    HAVANA  CDS     3670552 3671348 .       -       0       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_nam
 "Xkr4-201"; exon_number 1; exon_id "ENSMUSE00000485541.3"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG
000026353.2"; havana_transcript "OTTMUST00000065166.1";
chr1    HAVANA  start_codon     3671346 3671348 .       -       0       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transc
ipt_name "Xkr4-201"; exon_number 1; exon_id "ENSMUSE00000485541.3"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "
TTMUSG00000026353.2"; havana_transcript "OTTMUST00000065166.1";
chr1    HAVANA  exon    3421702 3421901 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_nam
mm10.gtf
```

Extracting reads – what is a gtf file?

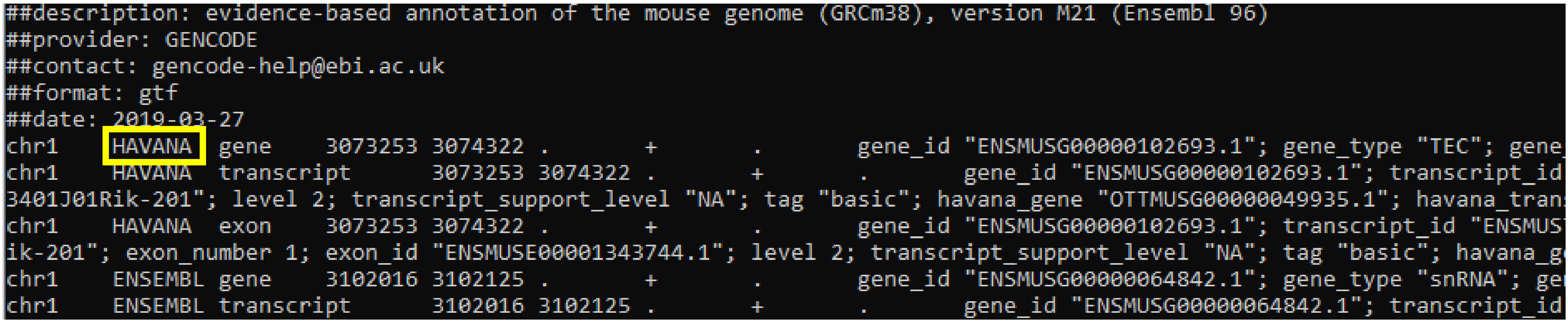
1. **Sequence Name**
2. Source of Annotation
3. Feature
4. Start
5. End
6. Score
7. Strand
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

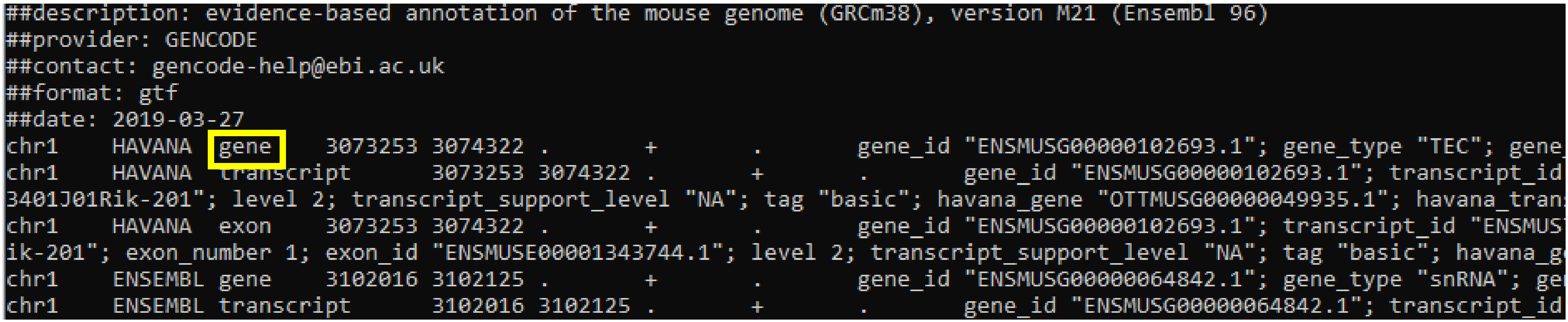
1. Sequence Name
2. **Source of Annotation**
3. Feature
4. Start
5. End
6. Score
7. Strand
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

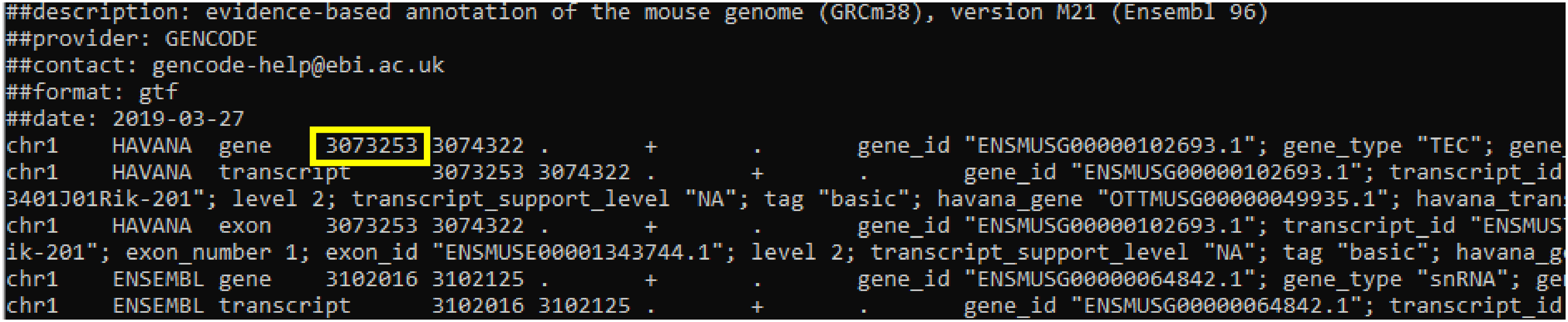
1. Sequence Name
2. Source of Annotation
3. **Feature**
4. Start
5. End
6. Score
7. Strand
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

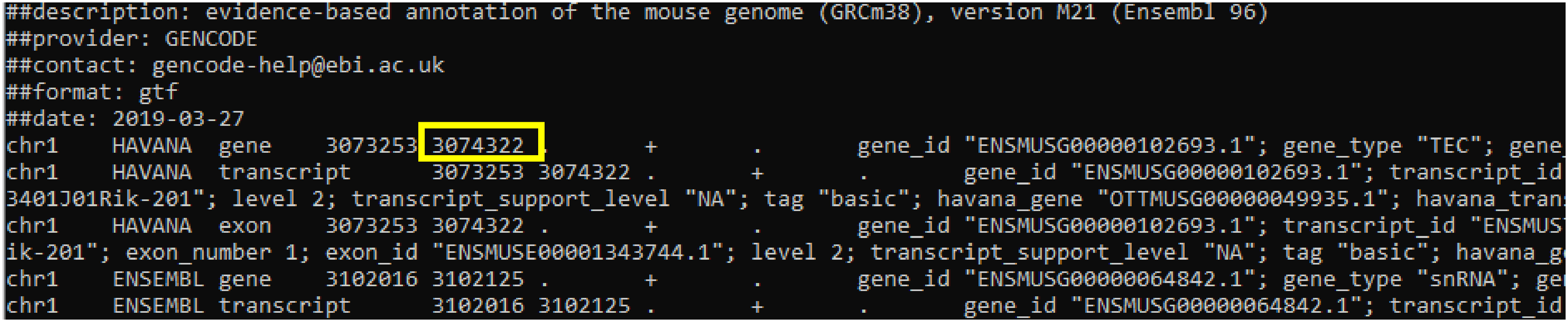
1. Sequence Name
2. Source of Annotation
3. Feature
4. **Start**
5. End
6. Score
7. Strand
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

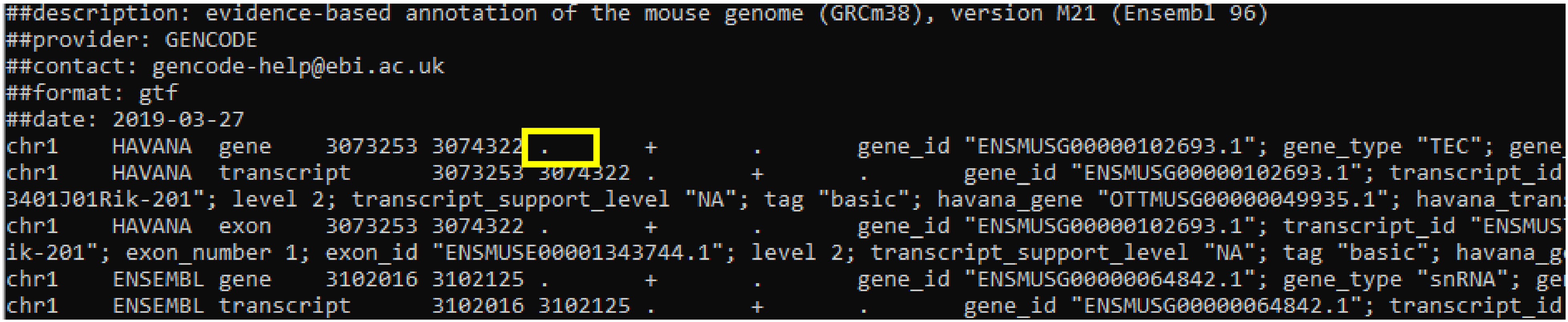
1. Sequence Name
2. Source of Annotation
3. Feature
4. Start
5. **End**
6. Score
7. Strand
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

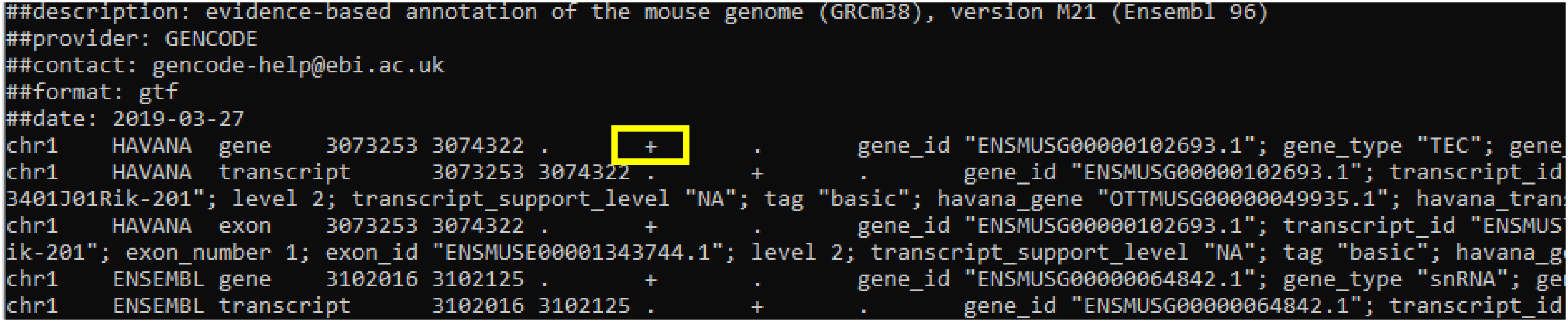
1. Sequence Name
2. Source of Annotation
3. Feature
4. Start
5. End
6. **Score**
7. Strand
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

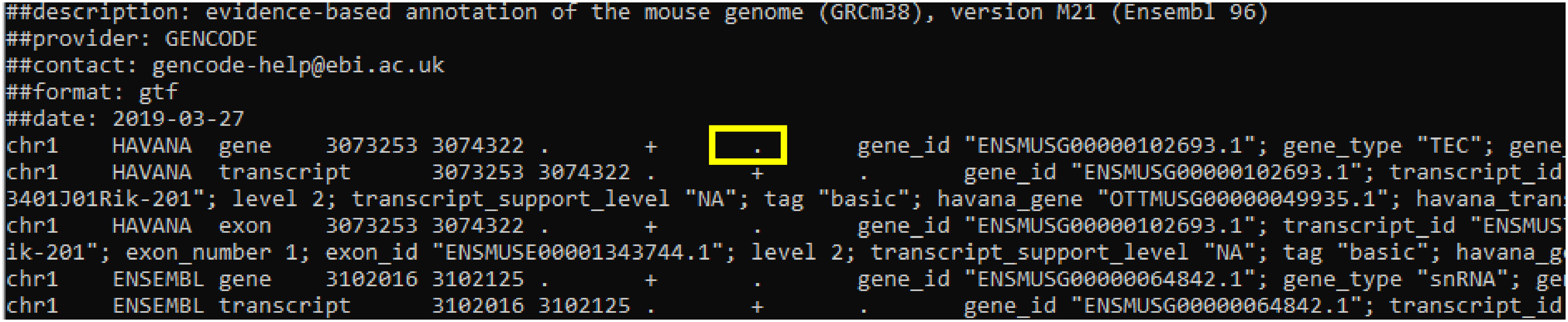
1. Sequence Name
2. Source of Annotation
3. Feature
4. Start
5. End
6. Score
7. **Strand**
8. Frame
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

Extracting reads – what is a gtf file?

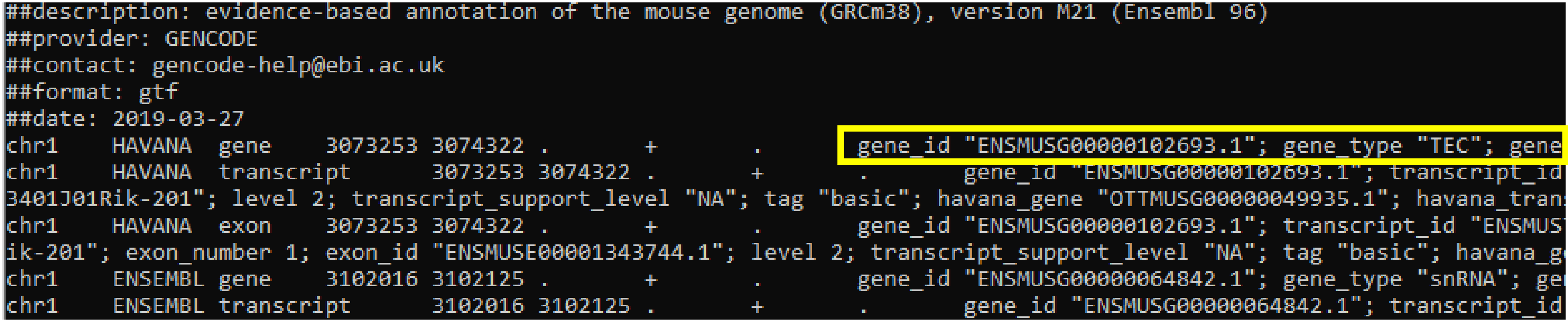
1. Sequence Name
2. Source of Annotation
3. Feature
4. Start
5. End
6. Score
7. Strand
8. **Frame**
9. Attribute



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

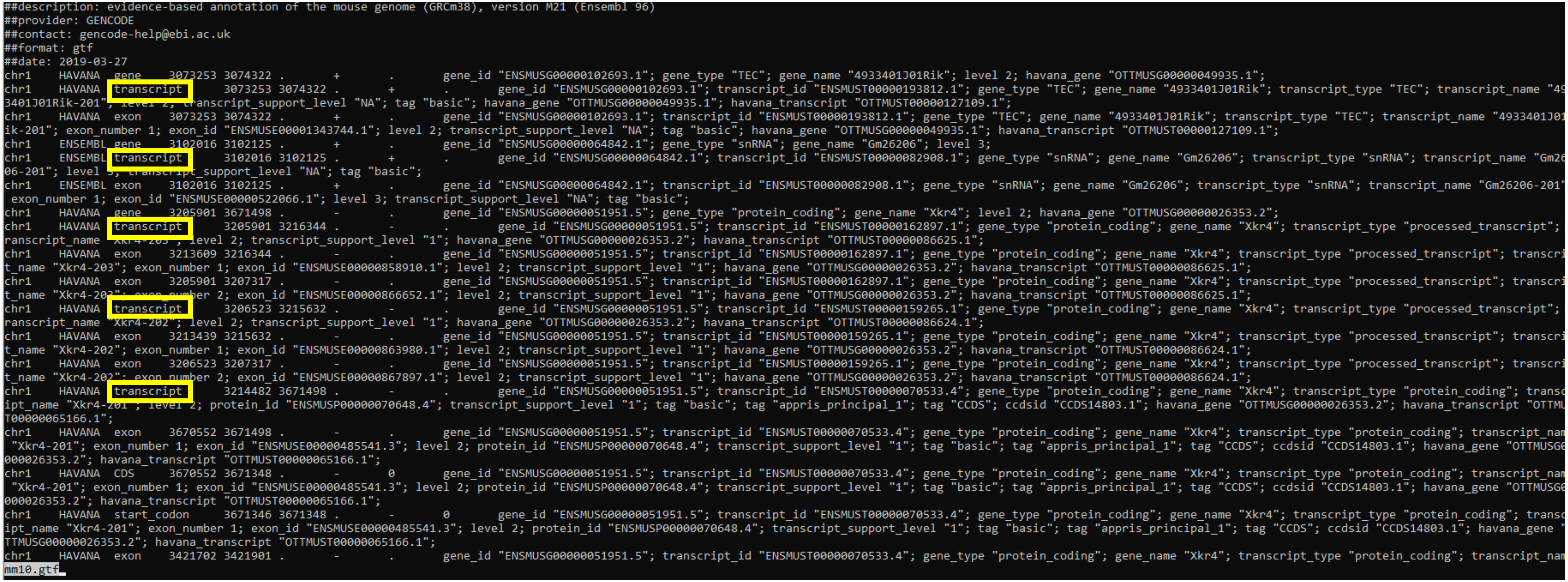
Extracting reads – what is a gtf file?

1. Sequence Name
2. Source of Annotation
3. Feature
4. Start
5. End
6. Score
7. Strand
8. Frame
9. **Attribute**



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_tran
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUS
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_g
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; ge
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id
```

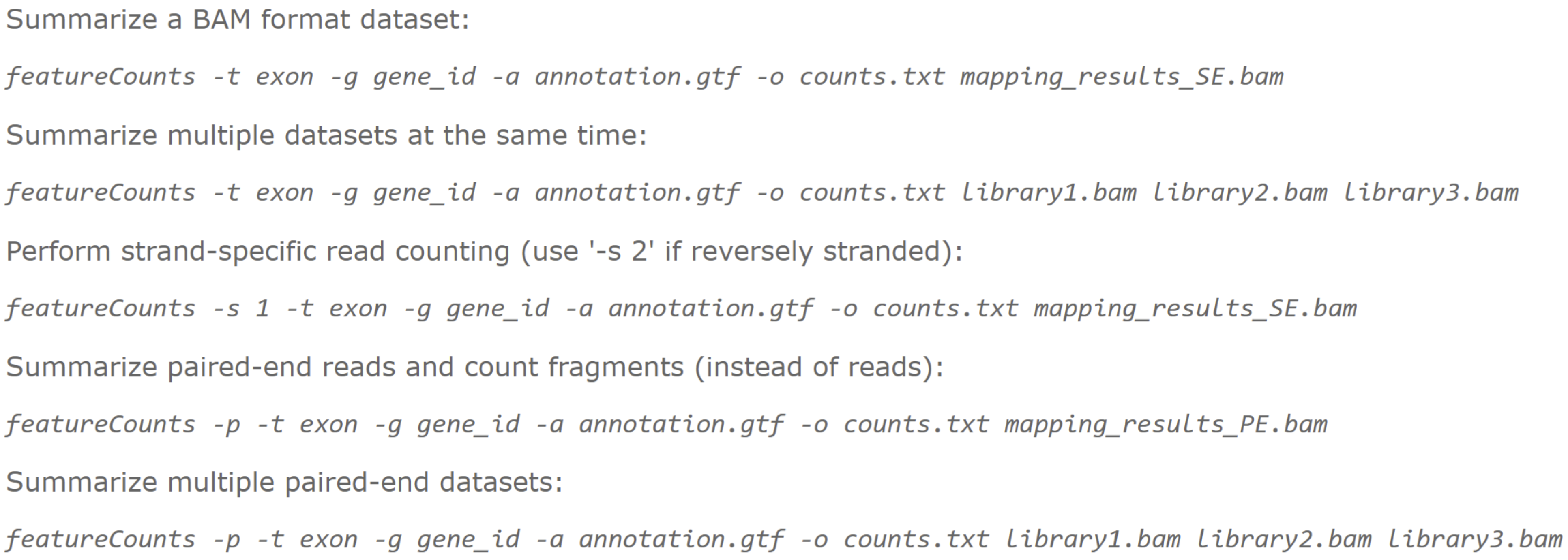
Extracting reads – what is a gtf file?



```
##description: evidence-based annotation of the mouse genome (GRCm38), version M21 (Ensembl 96)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2019-03-27
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_name "4933401J01Rik"; level 2; havana_gene "OTTMUSG00000049935.1";
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcript_name "49
3401J01Rik-201"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcript_name "4933401J01
ik-201"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; gene_name "Gm26206"; level 3;
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_name "Gm26206"; transcript_type "snRNA"; transcript_name "Gm26
06-201"; level 3; transcript_support_level "NA"; tag "basic";
chr1    ENSEMBL exon    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_name "Gm26206"; transcript_type "snRNA"; transcript_name "Gm26206-201"
 exon_number 1; exon_id "ENSMUSE00000522066.1"; level 3; transcript_support_level "NA"; tag "basic";
chr1    HAVANA  gene    3205901 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; gene_type "protein_coding"; gene_name "Xkr4"; level 2; havana_gene "OTTMUSG00000026353.2";
chr1    HAVANA  transcript      3205901 3216344 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript";
ranscript_name "Xkr4-203"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  exon    3213609 3216344 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-203"; exon_number 1; exon_id "ENSMUSE00000858910.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  exon    3205901 3207317 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-203"; exon_number 2; exon_id "ENSMUSE00000866652.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  transcript      3206523 3215632 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript";
ranscript_name "Xkr4-202"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  exon    3213439 3215632 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-202"; exon_number 1; exon_id "ENSMUSE00000863980.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  exon    3206523 3207317 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcri
t_name "Xkr4-202"; exon_number 2; exon_id "ENSMUSE00000867897.1"; level 2; transcript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  transcript      3214482 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transc
ipt_name "Xkr4-201"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMU
T00000065166.1";
chr1    HAVANA  exon    3670552 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_nam
 "Xkr4-201"; exon_number 1; exon_id "ENSMUSE00000485541.3"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG0
000026353.2"; havana_transcript "OTTMUST00000065166.1";
chr1    HAVANA  CDS     3670552 3671348 .       -       0       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_nam
 "Xkr4-201"; exon_number 1; exon_id "ENSMUSE00000485541.3"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG0
000026353.2"; havana_transcript "OTTMUST00000065166.1";
chr1    HAVANA  start_codon     3671346 3671348 .       -       0       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transc
ipt_name "Xkr4-201"; exon_number 1; exon_id "ENSMUSE00000485541.3"; level 2; protein_id "ENSMUSP00000070648.4"; transcript_support_level "1"; tag "basic"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "
TTMUSG00000026353.2"; havana_transcript "OTTMUST00000065166.1";
chr1    HAVANA  exon    3421702 3421901 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_nam
mm10.gtf
```

Extracting reads – different tools

- Subreads feature count
- Htseq
- RSEM



Summarize a BAM format dataset:

```
featureCounts -t exon -g gene_id -a annotation.gtf -o counts.txt mapping_results_SE.bam
```

Summarize multiple datasets at the same time:

```
featureCounts -t exon -g gene_id -a annotation.gtf -o counts.txt library1.bam library2.bam library3.bam
```

Perform strand-specific read counting (use '-s 2' if reversely stranded):

```
featureCounts -s 1 -t exon -g gene_id -a annotation.gtf -o counts.txt mapping_results_SE.bam
```

Summarize paired-end reads and count fragments (instead of reads):

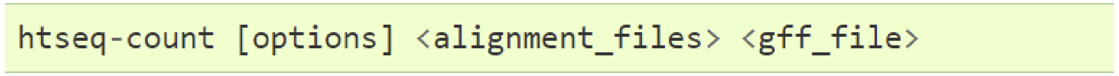
```
featureCounts -p -t exon -g gene_id -a annotation.gtf -o counts.txt mapping_results_PE.bam
```

Summarize multiple paired-end datasets:

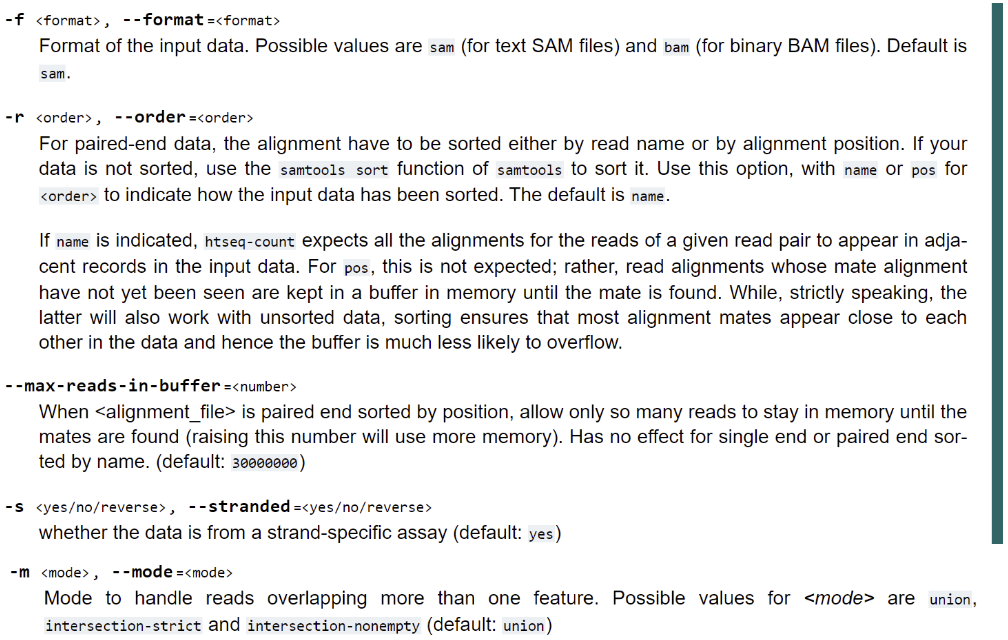
```
featureCounts -p -t exon -g gene_id -a annotation.gtf -o counts.txt library1.bam library2.bam library3.bam
```

Extracting reads – different tools

- Subreads feature count
- Htseq
- RSEM



```
htseq-count [options] <alignment_files> <gff_file>
```



-f <format>, **--format**=<format>

Format of the input data. Possible values are `sam` (for text SAM files) and `bam` (for binary BAM files). Default is `sam`.

-r <order>, **--order**=<order>

For paired-end data, the alignment have to be sorted either by read name or by alignment position. If your data is not sorted, use the `samtools sort` function of `samtools` to sort it. Use this option, with `name` or `pos` for `<order>` to indicate how the input data has been sorted. The default is `name`.

If `name` is indicated, `htseq-count` expects all the alignments for the reads of a given read pair to appear in adjacent records in the input data. For `pos`, this is not expected; rather, read alignments whose mate alignment have not yet been seen are kept in a buffer in memory until the mate is found. While, strictly speaking, the latter will also work with unsorted data, sorting ensures that most alignment mates appear close to each other in the data and hence the buffer is much less likely to overflow.

--max-reads-in-buffer=<number>

When <alignment_file> is paired end sorted by position, allow only so many reads to stay in memory until the mates are found (raising this number will use more memory). Has no effect for single end or paired end sorted by name. (default: `30000000`)

-s <yes/no/reverse>, **--stranded**=<yes/no/reverse>

whether the data is from a strand-specific assay (default: `yes`)

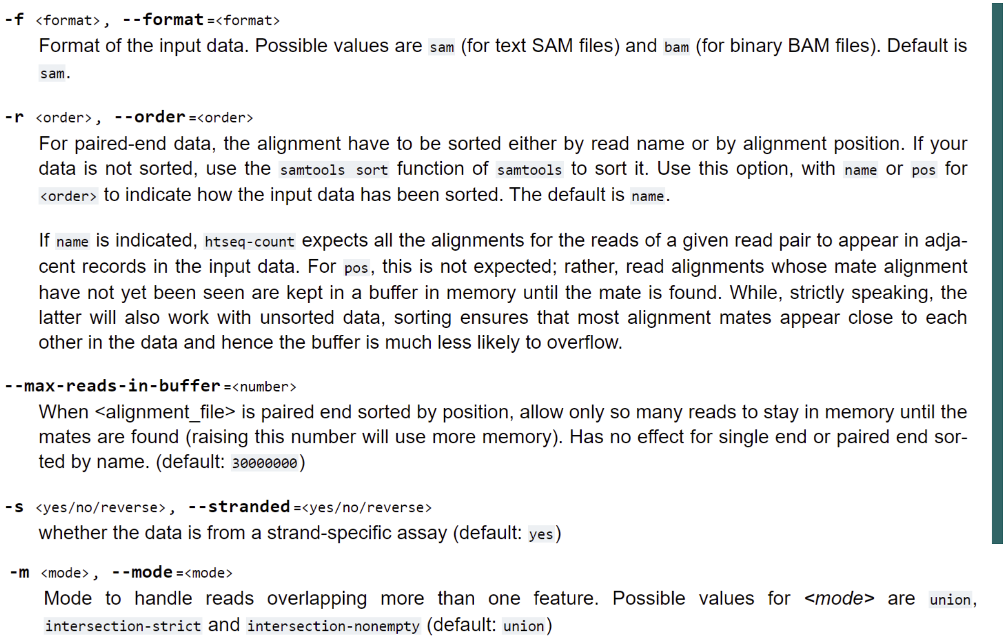
-m <mode>, **--mode**=<mode>

Mode to handle reads overlapping more than one feature. Possible values for *<mode>* are `union`, `intersection-strict` and `intersection-nonempty` (default: `union`)

Extracting reads – different tools

- Subreads feature count
- Htseq
- RSEM

```
htseq-count [options] <alignment_files> <gff_file>
```



-f <format>, **--format**=<format>
Format of the input data. Possible values are `sam` (for text SAM files) and `bam` (for binary BAM files). Default is `sam`.

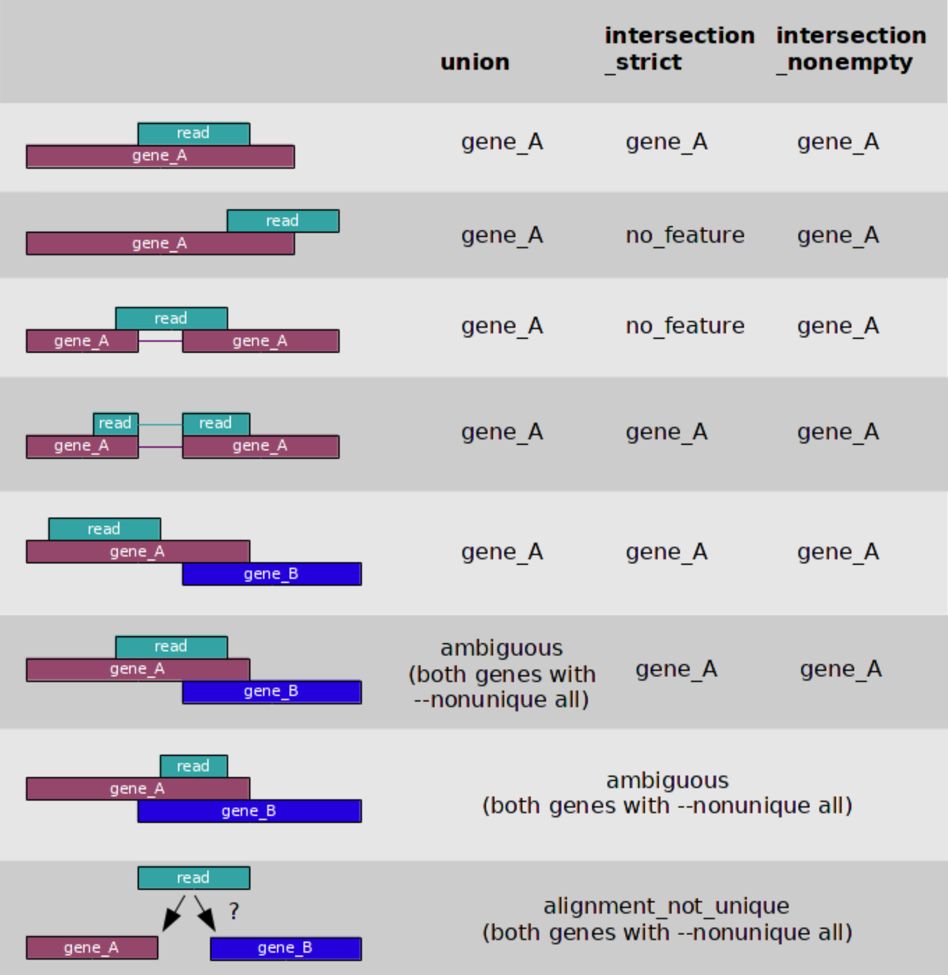
-r <order>, **--order**=<order>
For paired-end data, the alignment have to be sorted either by read name or by alignment position. If your data is not sorted, use the `samtools sort` function of `samtools` to sort it. Use this option, with `name` or `pos` for `<order>` to indicate how the input data has been sorted. The default is `name`.

If `name` is indicated, `htseq-count` expects all the alignments for the reads of a given read pair to appear in adjacent records in the input data. For `pos`, this is not expected; rather, read alignments whose mate alignment have not yet been seen are kept in a buffer in memory until the mate is found. While, strictly speaking, the latter will also work with unsorted data, sorting ensures that most alignment mates appear close to each other in the data and hence the buffer is much less likely to overflow.

--max-reads-in-buffer=<number>
When <alignment_file> is paired end sorted by position, allow only so many reads to stay in memory until the mates are found (raising this number will use more memory). Has no effect for single end or paired end sorted by name. (default: `30000000`)

-s <yes/no/reverse>, **--stranded**=<yes/no/reverse>
whether the data is from a strand-specific assay (default: `yes`)

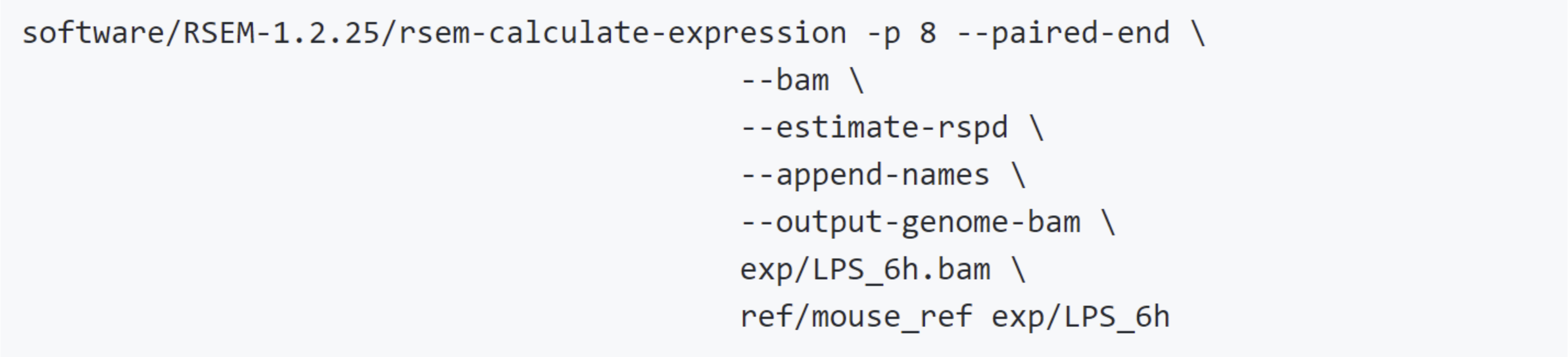
-m <mode>, **--mode**=<mode>
Mode to handle reads overlapping more than one feature. Possible values for *<mode>* are `union`, `intersection-strict` and `intersection-nonempty` (default: `union`)



	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)		
read ? gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

Extracting reads – different tools

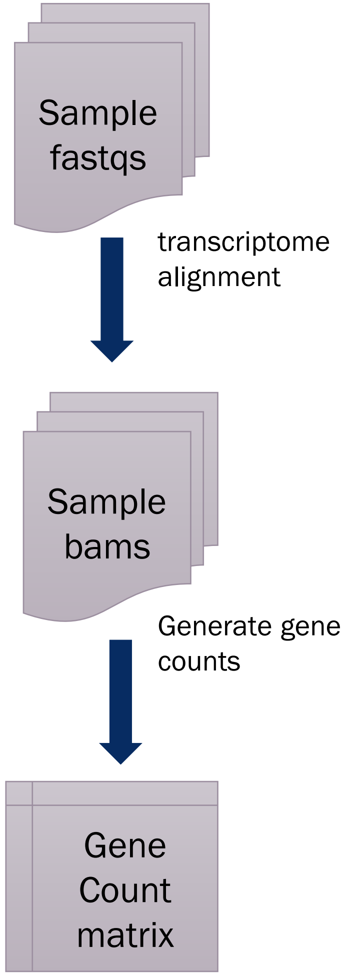
- Subreads feature count
- Htseq
- RSEM



```
software/RSEM-1.2.25/rsem-calculate-expression -p 8 --paired-end \
                                        --bam \
                                        --estimate-rspd \
                                        --append-names \
                                        --output-genome-bam \
                                        exp/LPS_6h.bam \
                                        ref/mouse_ref exp/LPS_6h
```

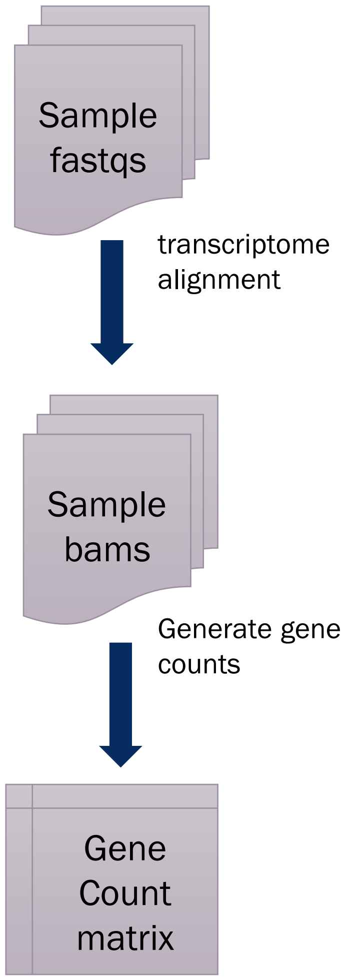
Transcriptomics pipeline/workflow

Preprocessing



Transcriptomics pipeline/workflow

Preprocessing



Sample fastqs

transcriptome alignment

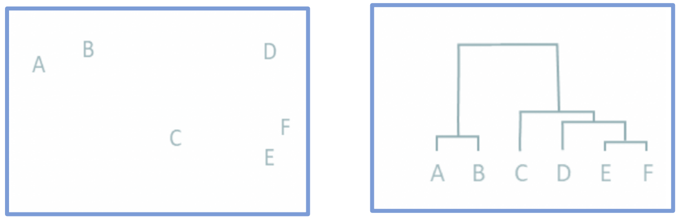
Sample bams

Generate gene counts

Gene Count matrix

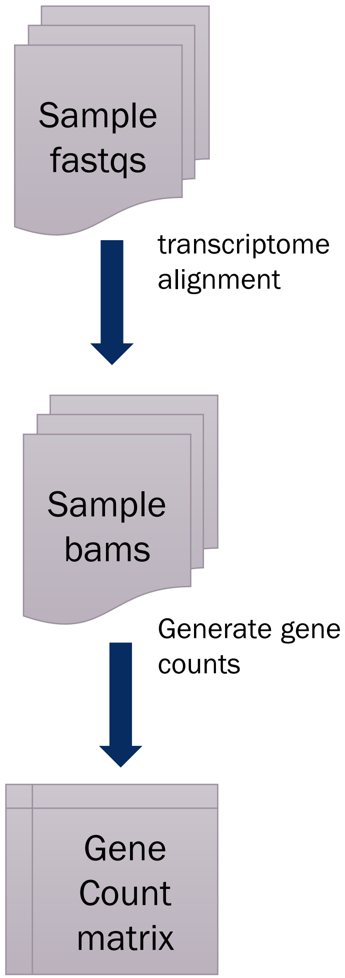
Analyses

Clustering



Transcriptomics pipeline/workflow

Preprocessing



Sample fastqs

transcriptome alignment

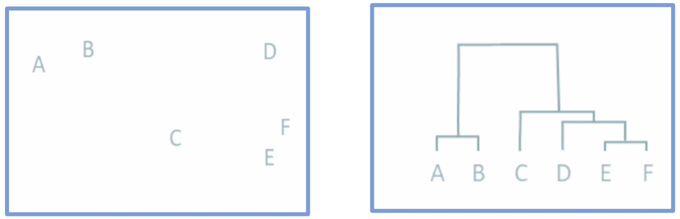
Sample bams

Generate gene counts

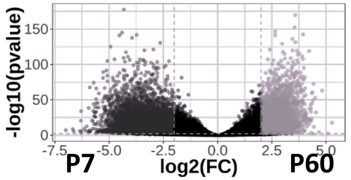
Gene Count matrix

Analyses

Clustering



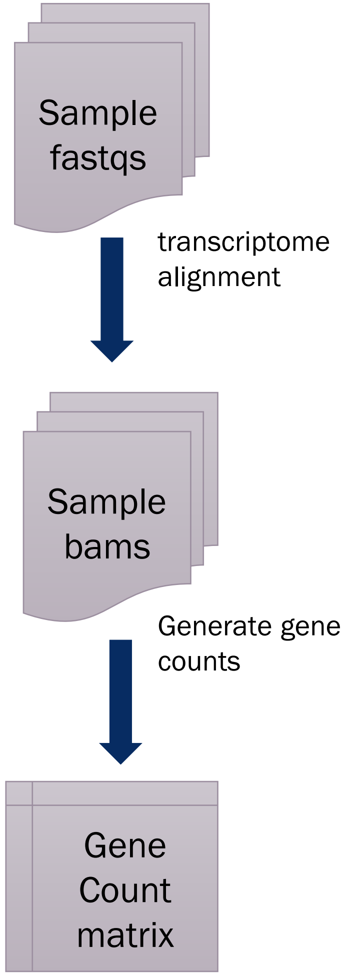
Differential Expression



DESeq2 limma Glimma edgeR

Transcriptomics pipeline/workflow

Preprocessing



Sample fastqs

transcriptome alignment

Sample bams

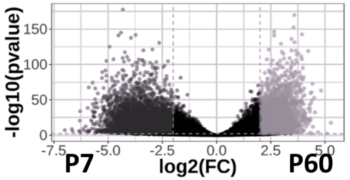
Generate gene counts

Gene Count matrix

Analyses

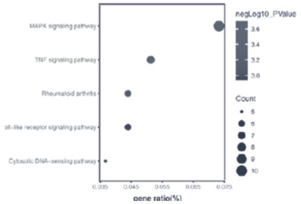
Clustering

Differential Expression



DESeq2 limma Glimma edgeR

Functional Enrichment



GSEA Gene Set Enrichment Analysis

MetaCore

GREAT

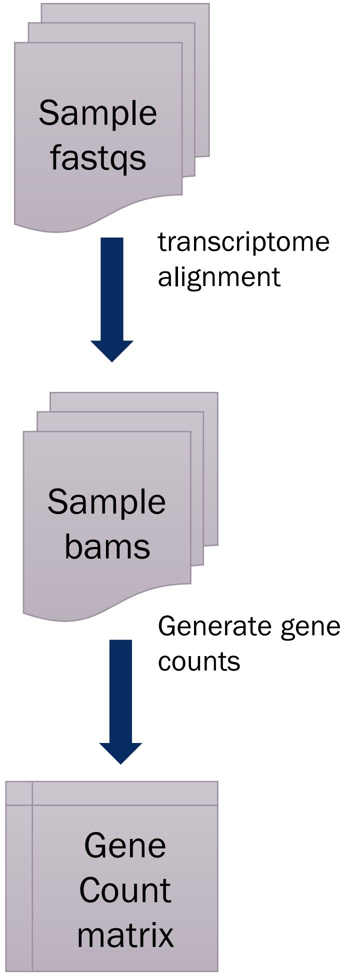
KEGG Kyoto Encyclopedia of Genes and Genomes

reactome

DAVID Bioinformatics Database

Transcriptomics pipeline/workflow

Preprocessing



Sample fastqs

transcriptome alignment

Sample bams

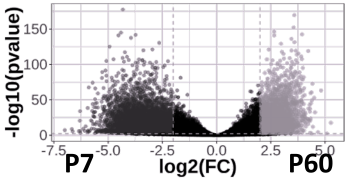
Generate gene counts

Gene Count matrix

Analyses

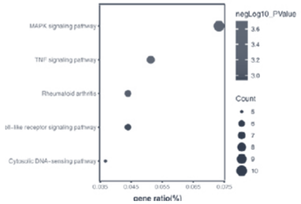
Clustering

Differential Expression



DESeq2 limma Glimma edgeR

Functional Enrichment



GSEA Gene Set Enrichment Analysis

MetaCore

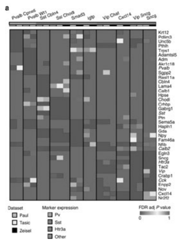
KEGG Kyoto Encyclopedia of Genes and Genomes

reactome

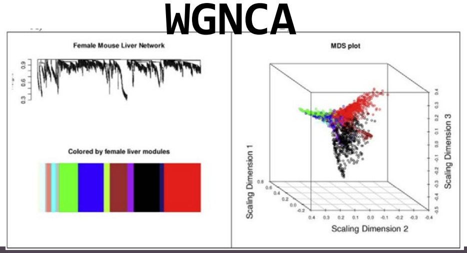
GREAT

DAVID BIOINFORMATICS DATABASE

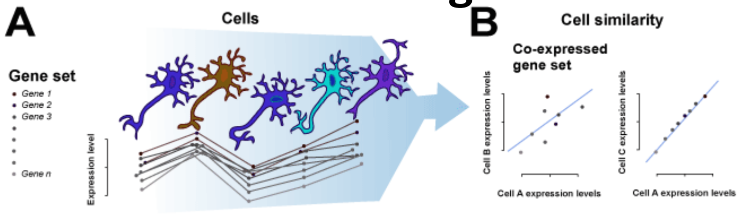
Coregulated Gene Expression



WGNCA



MetaNeighbor


A
Gene set
Gene 1
Gene 2
Gene 3
Gene n
Expression level
Cells
B
Cell similarity
Co-expressed gene set
Cell B expression levels
Cell C expression levels
Cell A expression levels
Cell A expression levels