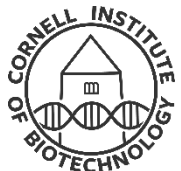


# **Epigenomics data analysis**

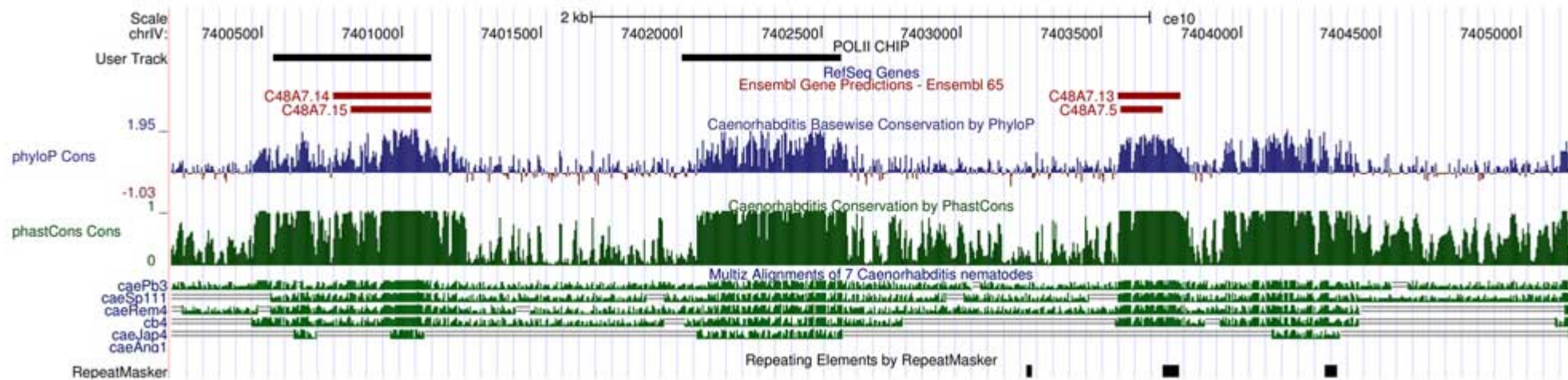
*11/30/2020 – 12/18/2020*

**Qi Sun, William Lai and Jeff Glaubit**

**Bioinformatics Facility & Epigenomics Facility**



# Genome feature annotation & Read enrichment pattern



## Genome features (*gff3 feature types*)

- gene
- CDS
- mRNA
- exon
- five\_prime\_UTR
- three\_prime\_UTR
- rRNA
- tRNA
- ncRNA
- tmRNA
- transcript
- mobile\_genetic\_element
- origin\_of\_replication
- promoter
- repeat\_region

# Genome feature annotation & Epigenomics data analysis

## Week 1

### Genome features

- File formats
- Public databases
- Genome browser

## Week 2

### Peak calling

- Peak calling
- Data QC

## Week 3

### Integrated analysis

- Bedtools
- Deeptools
- Homer

# File formats listed on UCSC web site

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigNarrowPeak table format
- bigWig format
- Chain format
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- Hic format
- Interact and bigInteract format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

# File formats listed on UCSC Genome Browser web site

<https://genome.ucsc.edu/FAQ/FAQformat.html#format3>

## General format

|                      |                      |
|----------------------|----------------------|
| Axt                  | CRAM                 |
| BAM                  | GenePred table       |
| BED                  | GFF                  |
| BED detail           | GTF                  |
| bedGraph             | HAL                  |
| barChart&bigBarChart | Hic                  |
| bigBed               | Interact&bigInteract |
| bigGenePred table    | MAF                  |
| bigPsl table         | Microarray           |
| bigMaf table         | Net                  |
| bigChain table       | Personal Genome      |
| bigNarrowPeak table  | SNP                  |
| bigWig               | PSL                  |
| Chain                | VCF                  |
|                      | WIG                  |

## Encode format

broadPeak  
gappedPeak  
narrowPeak  
pairedTagAlign  
peptideMapping  
RNA elements  
tagAlign

## Sequence

2bit  
fasta  
fastQ  
nib

# GFF2, GFF3 and GTF

9-column tab-delimited text file

Gff2 example:

|      |       |             |     |     |   |   |   |          |
|------|-------|-------------|-----|-----|---|---|---|----------|
| chr1 | Maker | CDS         | 380 | 401 | . | + | 0 | gene_001 |
| chr1 | Maker | CDS         | 501 | 650 | . | + | 2 | gene_001 |
| chr1 | Maker | CDS         | 700 | 707 | . | + | 2 | gene_001 |
| chr1 | Maker | start_codon | 380 | 382 | . | + | 0 | gene_001 |
| chr1 | Maker | stop_codon  | 708 | 710 | . | + | 0 | gene_001 |

| Column | Name                 |
|--------|----------------------|
| 1      | Chromosome/contig    |
| 2      | Source of annotation |
| 3      | Feature type         |
| 4      | Start position       |
| 5      | End position         |

|   |                     |
|---|---------------------|
| 6 | Score               |
| 7 | Strand ( + or -)    |
| 8 | Phase (0,1 or 2)    |
| 9 | Attributes or Group |

# GFF2 vs GTF

Gff2 example:

|      |       | Feature type |     |     |   |   |   | Group    |
|------|-------|--------------|-----|-----|---|---|---|----------|
| chr1 | Maker | CDS          | 380 | 401 | . | + | 0 | gene_001 |
| chr1 | Maker | CDS          | 501 | 650 | . | + | 2 | gene_001 |
| chr1 | Maker | CDS          | 700 | 707 | . | + | 2 | gene_001 |
| chr1 | Maker | start_codon  | 380 | 382 | . | + | 0 | gene_001 |
| chr1 | Maker | stop_codon   | 708 | 710 | . | + | 0 | gene_001 |

## Feature type:

**GFF2:** "CDS" "start\_codon" "stop\_codon" and "exon"

**GTF:** All GFF2 types plus optional types: 5UTR, 3UTR, inter, inter\_CNS, and intron\_CNS

## group:

**GFF2:** Single group attribute, not able to represent multiple splicing forms of the same gene

**GTF:** Two mandatory attributes: **gene\_id** "gene\_001"; **transcript\_id** "gene\_001.1"

# GFF3 vs GTF

GFF3 example:

|      |       |      |      |       |   |   |                                           |
|------|-------|------|------|-------|---|---|-------------------------------------------|
| chr1 | Maker | gene | 1000 | 9000. | + | . | ID=gene00001;Name=EDEN                    |
| chr1 | Maker | mRNA | 1050 | 9000. | + | . | ID=mRNA00001;Parent=gene00001;Name=EDEN.1 |
| chr1 | Maker | exon | 1050 | 1500. | + | . | ID=exon00002;Parent=mRNA00001             |
| chr1 | Maker | exon | 3000 | 3902. | + | . | ID=exon00003;Parent=mRNA00001             |
| chr1 | Maker | CDS  | 1201 | 1500. | + | 0 | ID=cds00001;Parent=mRNA00001              |
| chr1 | Maker | CDS  | 3000 | 3902. | + | 0 | ID=cds00001;Parent=mRNA00001              |

**GFF3:** Two mandatory attributes: **ID** and **Parent**

**GTF:** Two mandatory attributes: **gene\_id** “gene0001”; **transcript\_id** “mRNA00001”



# **gffread, a tool to convert between GFF3 and GTF**

GFF3 -> GTF format

```
gffread -E -T -o output.gtf input.gff3
```

GTF -> GFF3 format

```
gffread -E -G -o output.gff3 input.gtf
```

# Lost information when converting from GFF3 to GTF

## GFF3

|      |       |      |      |       |   |   |                                           |
|------|-------|------|------|-------|---|---|-------------------------------------------|
| chr1 | Maker | gene | 1000 | 9000. | + | . | ID=gene00001;Name=EDEN                    |
| chr1 | Maker | mRNA | 1050 | 9000. | + | . | ID=mRNA00001;Parent=gene00001;Name=EDEN.1 |
| chr1 | Maker | exon | 1050 | 1500. | + | . | ID=exon00002;Parent=mRNA00001             |

## GTF

|      |       |      |      |       |   |   |                                                                          |
|------|-------|------|------|-------|---|---|--------------------------------------------------------------------------|
| chr1 | Maker | exon | 1050 | 1500. | + | . | gene_id "gene0001"; transcript_id "mRNA00001";<br>gene_name "AT1G01010"; |
|------|-------|------|------|-------|---|---|--------------------------------------------------------------------------|

gffread -E -G -o output.gff3 input.gtf

mRNA feature created from  
first exon

Extract mRNA and protein sequences from the genome using *gffread*

```
gffread input.gff3 -g input.fasta -w transcript.fasta -x protein.fasta
```

- Chromosome names should match between files;
- “Chr1”, “chr1” and “1” are different

## Other features in gff files:

```
Chr1  Araport11  lnc_RNA 11101 11372 . + . ID=AT1G03987.1;Parent=AT1G03987;Name=AT1G03987.1
Chr3  Araport11  tRNA 2918413 2918486 74 - . ID=AT3G09505.1;Parent=AT3G09505;Note=tRNA-Arg
Chr2  Araport11  transposable_element 2782038 2785706 . + .ID=AT2TE12270;Name=AT2TE12270;Alias=ATCOPIA75
```

```
zcat ara.gff.gz | cut -f3 | sort | uniq
```

|                          |                           |
|--------------------------|---------------------------|
| CDS                      | pseudogenic_exon          |
| antisense_RNA            | pseudogenic_tRNA          |
| antisense_lncRNA         | pseudogenic_transcript    |
| exon                     | rRNA                      |
| five_prime_UTR           | snRNA                     |
| gene                     | snoRNA                    |
| lnc_RNA                  | tRNA                      |
| mRNA                     | three_prime_UTR           |
| miRNA                    | transcript_region         |
| miRNA_primary_transcript | transposable_element      |
| ncRNA                    | transposable_element_gene |
| protein                  | transposon_fragment       |
| pseudogene               | uORF                      |

# BED

3-column format

|      |      |      |
|------|------|------|
| chr1 | 1000 | 5000 |
| chr2 | 2000 | 6000 |

12-column format

|      |      |      |       |   |   |      |      |   |   |          |        |
|------|------|------|-------|---|---|------|------|---|---|----------|--------|
| chr1 | 1000 | 5000 | gene1 | . | + | 1000 | 5000 | 0 | 2 | 567,488, | 0,3512 |
| chr2 | 2000 | 6000 | gene2 | . | - | 2000 | 6000 | 0 | 2 | 433,399, | 0,3601 |

| Column | Name              |
|--------|-------------------|
| 1      | Chromosome/contig |
| 2      | Start position    |
| 3      | End position      |
| 4      | Name              |
| 5      | Score             |

|     |                               |
|-----|-------------------------------|
| 6   | Strand ( + or -)              |
| 7-9 | thickStart, thickEnd, itemRgb |
| 10  | blockCount                    |
| 11  | blockSizes                    |
| 12  | blockStarts                   |

# BED ↔ GFF3

## GFF3

|      |       |      |      |       |   |   |                                           |
|------|-------|------|------|-------|---|---|-------------------------------------------|
| chr1 | Maker | gene | 1000 | 9000. | + | . | ID=gene00001;Name=EDEN                    |
| chr1 | Maker | mRNA | 1050 | 9000. | + | . | ID=mRNA00001;Parent=gene00001;Name=EDEN.1 |
| chr1 | Maker | exon | 1050 | 1500. | + | . | ID=exon00002;Parent=mRNA00001             |

## BED

|      |     |      |   |   |   |
|------|-----|------|---|---|---|
| chr1 | 999 | 9000 | . | . | + |
|------|-----|------|---|---|---|

- Most software ignore the last 6 columns of BED

Coordinate system for a 100-bp sequence)

### **BED, bam**

0-start, half-open (0-based):      first nucleotide: 0;      last nucleotide: 100

### **GFF, GTF, vcf, genbank, sam, blast output**

1-start, fully-closed (1-based):      first nucleotide: 1;      last nucleotide: 100

# Start/end positions of stranded DNA sequence:

BED

Plus strand

|      |   |     |   |   |   |
|------|---|-----|---|---|---|
| chr1 | 0 | 100 | . | . | + |
|------|---|-----|---|---|---|

Minus strand

|      |   |     |   |   |   |
|------|---|-----|---|---|---|
| chr1 | 0 | 100 | . | . | - |
|------|---|-----|---|---|---|

BLAST  
output

Plus strand

|      |   |     |
|------|---|-----|
| chr1 | 1 | 100 |
|------|---|-----|

Minus strand

|      |     |   |
|------|-----|---|
| chr1 | 100 | 1 |
|------|-----|---|

Transcript start site in BED file:

Gene on plus strand gene: start position  
(add 1)

Gene on minus strand: end position

# Stranded read position in SAM file format (1-based)

Coor

ref

12345678901234

5678901234567890123456789012345

AGCATGTTAGATAA\*\*GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1

TTAGATAAAGGATA\*CTG

+r002

aaaAGATAA\*GGATA

+r003

gcctaAGCTAA

+r004

ATAGCT.....TCAGC

-r003

ttagctTAGGC

-r001/2

CAGCGGCAT

Leftmost alignment position on reference

@HD

VN:1.6

S0:coordinate

@SQ

SN:ref

LN:45

r001

99

ref

7

30

8M2I4M1D3M

=

37

39

TTAGATAAAGGATACTG

\*

r002

0

ref

9

30

3S6M1P1I4M

\*

0

0

AAAAGATAAGGATA

\*

r003

0

ref

9

30

5S6M

\*

0

0

GCCTAAGCTAA

\*

SA:Z:ref,29,-,6H5M,17,0;

r004

0

ref

16

30

6M14N5M

\*

0

0

ATAGCTTCAGC

\*

r003

2064

ref

29

17

6H5M

\*

0

0

TAGGC

\*

SA:Z:ref,9,+,5S6M,30,1;

r001

147

ref

37

30

9M

=

7

-39

CAGCGGCAT

\*

NM:i:1



# Use LINUX “awk” to convert file format

GFF3

|      |       |      |      |      |   |   |   |                        |
|------|-------|------|------|------|---|---|---|------------------------|
| chr1 | Maker | gene | 1000 | 9000 | . | + | . | ID=gene00001;Name=EDEN |
|------|-------|------|------|------|---|---|---|------------------------|

BED

|      |     |      |   |   |   |
|------|-----|------|---|---|---|
| chr1 | 999 | 9000 | . | . | + |
|------|-----|------|---|---|---|

Awk command:

```
awk 'BEGIN { OFS = "\t" } {if ($3=="gene") print $1, $4-1, $5, ".", ".", $7}' input.gff3 > output.bed
```

# SAM/BAM file formats

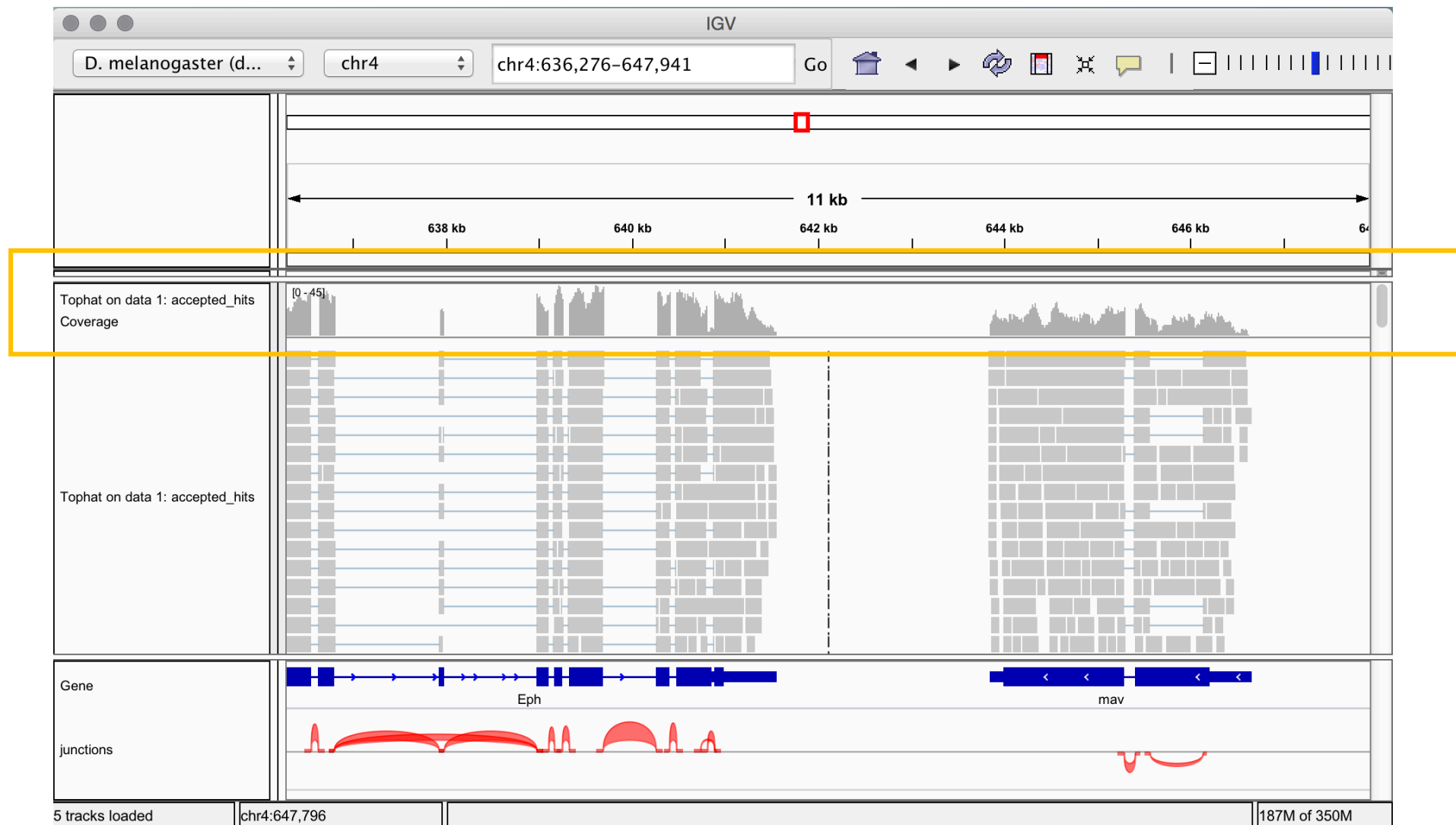
## READ alignment to the reference genome

```
HWUSI-EAS525_0042_FC:6:23:10200:18582#0/1    16    1    10    40    35M    *    0    0
AGCCAAAGATTGCATCAGTTCTGCTGCTATTTTCCT    agafgfaffcfdf[fdcffcggggccfdffagggg    MD:Z:35 NH:i:1 HI:i:1
NM:i:0 SM:i:40 XQ:i:40 X2:i:0
HWUSI-EAS525_0042_FC:3:28:18734:20197#0/1    16    1    10    40    35M    *    0    0
AGCCAAAGATTGCATCAGTTCTGCTGCTATTTTCCT    hghhghhhhhhhhhhhhhhhhhhhghhhhhghhhfhhh    MD:Z:35
NH:i:1 HI:i:1 NM:i:0 SM:i:40 XQ:i:40 X2:i:0
HWUSI-EAS525_0042_FC:3:94:1587:14299#0/1    16    1    10    40    35M    *    0    0
AGCCAAAGATTGCATCAGTTCTGCTGCTATTTTCCT    hfhghhhhhhhhhhhghhhhhhhhhhhhhhhhhhhhg    MD:Z:35
NH:i:1 HI:i:1 NM:i:0 SM:i:40 XQ:i:40 X2:i:0
```

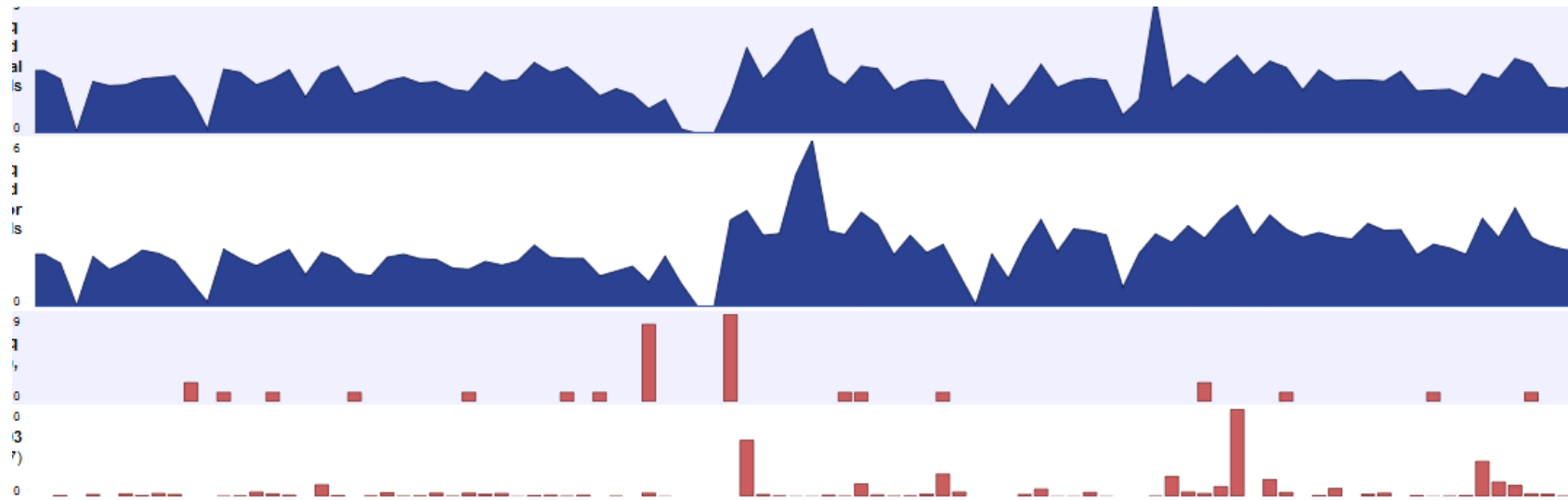
Convert to bed file:

| Chr | start | end | name | score | strand |
|-----|-------|-----|------|-------|--------|
| 1   | 9     | 44  | .    | 40    | -      |
| 1   | 9     | 44  | .    | 40    | -      |
| 1   | 9     | 44  | .    | 40    | -      |

# Wiggle (wig) file format:



# Sparse vs dense data profile



# Wiggle (wig), BedGraph, and bigwig file format:

## Fixed step wig (1-based)

```
fixedStep chrom=chr3 start=400601 step=100 span=5  
11  
22  
33
```

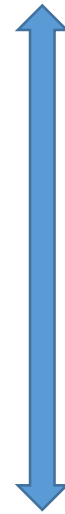
## Variable step wig (1-based)

```
variableStep chrom=chr3 span=5  
400601 11  
400701 22  
400801 33
```

## BedGraph (0-based)

```
chr3    400600  400605  11  
chr3    400700  400705  22  
chr3    400800  400805  33
```

Good for dense data



Good for sparse data

## BigWig file format replacing the Wig format:

- Compressed and indexed binary format;
- For both sparse and dense data;
- Recommended format for genome browser;

# Software for file conversion:

## **Bedtools genomecov**

Bam => bedgraph

## **Deeptools bamCoverage**

Bam => bedgraph or bigwig

## **Kent Utility:**

Conversion between wig, bigwig,

# Public genomic databases

NCBI Refseq : NCBI gene annotation

Ensembl:        Ensembl gene annotation

UCSC:            Integrated with UCSC genome browser

\* All three databases use the same genome assembly



<https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

## Index of /genomes/refseq

## Download from NCBI Refseq FTP site

| Name                                                   | Last modified    | Size |
|--------------------------------------------------------|------------------|------|
| Parent Directory                                       | -                | -    |
| <a href="#">archaea/</a>                               | 2020-09-23 20:29 | -    |
| <a href="#">bacteria/</a>                              | 2020-09-23 20:29 | -    |
| <a href="#">fungi/</a>                                 | 2020-09-23 20:29 | -    |
| <a href="#">invertebrate/</a>                          | 2020-09-23 20:29 | -    |
| <a href="#">plant/</a>                                 | 2020-09-23 20:29 | -    |
| <a href="#">protozoa/</a>                              | 2020-09-23 20:29 | -    |
| <a href="#">vertebrate_mammalian/</a>                  | 2020-09-23 20:29 | -    |
| <a href="#">vertebrate_other/</a>                      | 2020-09-23 20:29 | -    |
| <a href="#">viral/</a>                                 | 2020-09-23 20:29 | -    |
| <a href="#">README.txt</a>                             | 2020-09-02 16:26 | 43K  |
| <a href="#">assembly_summary_refseq.txt</a>            | 2020-09-23 20:28 | 701  |
| <a href="#">assembly_summary_refseq_historical.txt</a> | 2020-11-27 06:05 | 14   |

- Genome fasta
- GFF3
- GTF
- Protein sequence
- Transcript sequence

# Download from ENSEMBL FTP

Show 10 entries

Show/hide columns

| ★ | Species                                     | DNA (FASTA)           | cDNA (FASTA)          | CDS (FASTA)           | ncRNA (FASTA)         | Protein sequence (FASTA) | Annotated sequence (EMBL) | Annotated sequence (GenBank) | Gene sets                                   | Other annotations                                                  | Whole databases       | Variation (GVF)     | Variation (VCF)     | Variation (VEP)     | Regulation                 |
|---|---------------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------|---------------------------|------------------------------|---------------------------------------------|--------------------------------------------------------------------|-----------------------|---------------------|---------------------|---------------------|----------------------------|
| Y | Mouse<br>Mus musculus                       | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> | <a href="#">Regulation</a> |
|   | Algerian mouse<br>Mus spretus               | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>129/SvJ<br>Mus musculus            | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>AJ<br>Mus musculus                 | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>AKR/J<br>Mus musculus              | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>BALB/cJ<br>Mus musculus            | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>C3H/HeJ<br>Mus musculus            | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>C57BL/6N<br>Mus musculus           | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>CAST/EiJ<br>Mus musculus           | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>CAST/EiJ<br>Mus musculus castaneus | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |
|   | Mouse<br>CB17<br>Mus musculus               | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a> | <a href="#">FASTA</a>    | <a href="#">EMBL</a>      | <a href="#">GenBank</a>      | <a href="#">GTF</a><br><a href="#">GFF3</a> | <a href="#">TSV</a><br><a href="#">RDF</a><br><a href="#">JSON</a> | <a href="#">MySQL</a> | <a href="#">GVF</a> | <a href="#">VCF</a> | <a href="#">VEP</a> |                            |

Showing 1 to 10 of 22 entries (filtered from 310 total entries)

Download

Index of /pub/release-101

[parent directory]

|             | Name            | Size   | Date modified    |
|-------------|-----------------|--------|------------------|
| <div></div> | README          | 6.5 kB | 8/18/20, 12:36:0 |
| <div></div> | assembly_chain/ |        | 7/11/20, 6:31:0  |
| <div></div> | bamcov/         |        | 8/19/20, 1:10:0  |
| <div></div> | bed/            |        | 7/13/20, 7:30:0  |
| <div></div> | compara/        |        | 7/13/20, 7:34:0  |
| <div></div> | data files/     |        | 8/19/20, 1:10:0  |

## Index of /pub/release-101

 [\[parent directory\]](#)

|                                                                                     | Name                                     | Size    | Date modified        |
|-------------------------------------------------------------------------------------|------------------------------------------|---------|----------------------|
|    | README                                   | 6.5 kB  | 8/18/20, 12:36:00 PM |
|    | assembly_chain/                          |         | 7/11/20, 6:31:00 PM  |
|    | bamcov/                                  |         | 8/19/20, 1:10:00 PM  |
|    | bed/                                     |         | 7/13/20, 7:30:00 AM  |
|    | compara/                                 |         | 7/13/20, 7:34:00 AM  |
|    | data_files/                              |         | 8/19/20, 1:10:00 PM  |
|    | embl/                                    |         | 7/27/20, 5:43:00 AM  |
|    | emf/                                     |         | 7/13/20, 7:30:00 AM  |
|    | fasta/                                   |         | 7/10/20, 2:34:00 PM  |
|    | genbank/                                 |         | 7/26/20, 10:38:00 PM |
|    | gff3/                                    |         | 7/28/20, 7:22:00 PM  |
|    | gtf/                                     |         | 7/24/20, 6:08:00 PM  |
|    | json/                                    |         | 8/15/20, 1:02:00 PM  |
|    | maf/                                     |         | 7/13/20, 7:30:00 AM  |
|    | mysql/                                   |         | 8/11/20, 8:13:00 AM  |
|    | ncbi_blast/                              |         | 7/10/20, 1:51:00 PM  |
|    | new_genomes.txt                          | 1.5 kB  | 8/3/20, 10:47:00 AM  |
|  | rdp/                                     |         | 8/17/20, 11:08:00 AM |
|  | regulation/                              |         | 6/9/20, 9:45:00 AM   |
|  | removed_genomes.txt                      | 35 B    | 8/3/20, 10:47:00 AM  |
|  | renamed_genomes.txt                      | 44 B    | 8/3/20, 10:47:00 AM  |
|  | solr_src/                                |         | 8/10/20, 6:05:00 AM  |
|  | species_EnsemblVertebrates.txt           | 47.1 kB | 8/3/20, 10:47:00 AM  |
|  | species_metadata_EnsemblVertebrates.json | 2.7 GB  | 8/3/20, 11:02:00 AM  |
|  | species_metadata_EnsemblVertebrates.xml  | 1.5 GB  | 8/4/20, 7:38:00 AM   |
|  | summary.txt                              | 748 B   | 8/3/20, 10:47:00 AM  |
|  | tsv/                                     |         | 7/13/20, 7:30:00 AM  |
|  | uniprot_report_EnsemblVertebrates.txt    | 39.8 kB | 8/3/20, 10:47:00 AM  |
|  | updated_annotations.txt                  | 63 B    | 8/3/20, 10:47:00 AM  |
|  | updated_assemblies.txt                   | 102 B   | 8/3/20, 10:47:00 AM  |
|  | variation/                               |         | 7/17/20, 6:23:00 AM  |
|  | virtual_machine/                         |         | 5/26/20, 8:00:00 PM  |
|  | xml/                                     |         | 7/13/20, 7:30:00 AM  |

## Index of /pub/release-101/gff3/

 [\[parent directory\]](#)

|                                                                                       | Name                              | Size | Date modified        |
|---------------------------------------------------------------------------------------|-----------------------------------|------|----------------------|
|    | acanthochromis_polyacanthus/      |      | 8/6/20, 4:02:00 PM   |
|    | accipiter_nisus/                  |      | 8/17/20, 12:59:00 PM |
|    | ailuropoda_melanoleuca/           |      | 8/17/20, 12:44:00 PM |
|    | amazona_collaria/                 |      | 8/6/20, 3:54:00 PM   |
|    | amphilophus_citrinellus/          |      | 8/17/20, 12:45:00 PM |
|    | amphiprion_ocellaris/             |      | 8/17/20, 12:46:00 PM |
|    | amphiprion_percula/               |      | 8/17/20, 12:54:00 PM |
|    | anabas_testudineus/               |      | 8/17/20, 12:58:00 PM |
|    | anas_platyrhynchos/               |      | 8/17/20, 12:28:00 PM |
|    | anas_platyrhynchos_platyrhynchos/ |      | 8/6/20, 3:47:00 PM   |
|    | anas_zonorhyncha/                 |      | 8/17/20, 12:34:00 PM |
|    | anolis_carolinensis/              |      | 8/17/20, 12:46:00 PM |
|    | anser_brachyrhynchus/             |      | 8/17/20, 11:32:00 AM |
|    | anser_cygnoides/                  |      | 8/17/20, 12:59:00 PM |
|    | aotus_nancymae/                   |      | 8/17/20, 12:29:00 PM |
|    | apteryx_haastii/                  |      | 8/17/20, 12:55:00 PM |
|  | apteryx_owenii/                   |      | 8/17/20, 12:58:00 PM |
|  | apteryx_rowi/                     |      | 8/17/20, 12:47:00 PM |
|  | aquila_chrysaetos_chrysaetos/     |      | 8/17/20, 1:00:00 PM  |
|  | astatotilapia_calliptera/         |      | 8/17/20, 12:28:00 PM |
|  | astyanax_mexicanus/               |      | 8/17/20, 12:46:00 PM |
|  | astyanax_mexicanus_pachon/        |      | 8/17/20, 12:43:00 PM |
|  | athene_cunicularia/               |      | 8/17/20, 12:46:00 PM |
|  | balaenoptera_musculus/            |      | 8/6/20, 3:52:00 PM   |
|  | betta_splendens/                  |      | 8/6/20, 3:56:00 PM   |
|  | bison_bison_bison/                |      | 8/17/20, 12:57:00 PM |
|  | bos_grunniens/                    |      | 8/17/20, 12:43:00 PM |
|  | bos_indicus_hybrid/               |      | 8/17/20, 12:37:00 PM |

# UCSC table browser

## Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description and the [User's Guide](#) for general information and sample queries. For more complex queries, you may use the [MySQL server](#). To examine the biological function of your set through annotation enrichments, send the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be found from the [Sequence and Annotation Downloads](#) page.

**clade:** Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

**group:** Genes and Gene Predictions **track:** GENCODE v32 [add custom tracks](#) [track hub](#)

**table:** knownGene [describe table schema](#)

**region:** ☐ genome ☒ position chr3:53,494,611-53,813,733 [lookup](#) [define regions](#)

**identifiers (names/accessions):** [paste list](#) [upload list](#)

**filter:** [create](#)

**intersection:** [create](#)

**correlation:** [create](#)

**output format:** GTF - gene transfer format (limited) [all fields from selected table](#) [selected fields from primary and related tables](#) [sequence](#) [GTF - gene transfer format \(limited\)](#) [CDS FASTA alignment from multiple alignment](#) [BED - browser extensible data](#) [custom track](#) [hyperlinks to Genome Browser](#)

**Note:** Table Browser will return gene identifiers or symbols. If you are looking for genomic coordinates, use in genome analysis pipelines, check the [bigZ](#).

Send output to ☐ [keep output in browser](#)

**output file:**

**file type returned:**

## Ensembl Biomart:

## Features:

Dataset

Mouse genes (GRCm38.p6)

Filters

[None selected]

Attributes

Gene stable ID

Gene stable ID version

Transcript stable ID

Transcript stable ID version

Dataset

[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following [FAQ](#)

☒ Features

☐ Structures

☐ Homologues (Max select 6 orthologues)

☐ Variant (Germline)

☐ Sequences

GENE:

Ensembl

☒ Gene stable ID

☒ Gene stable ID version

☒ Transcript stable ID

☒ Transcript stable ID version

☐ Protein stable ID

☐ Protein stable ID version

☐ Exon stable ID

☐ Gene description

☐ Chromosome/scaffold name

☐ Gene start (bp)

☐ Gene end (bp)

☐ Strand

☐ Karvotvbe band

☐ Transcript support level (TSL)

☐ GENCODE basic annotation

☐ APPRIS annotation

☐ Gene name

☐ Source of gene name

☐ Transcript name

☐ Source of transcript name

☐ Transcript count

☐ Gene % GC content

☐ Gene type

☐ Transcript type

☐ Source (gene)

☐ Source (transcript)

# Ensembl Biomart:

## Ortholog in other species:

|                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <div>Dataset</div> <div>Mouse genes (GRCm38.p6)</div> <div>Filters</div> <div>[None selected]</div> <div>Attributes</div> <div>Gene stable ID</div> <div>Gene stable ID version</div> <div>Transcript stable ID</div> <div>Transcript stable ID version</div> | <div>Please select columns to be included in the output and hit 'Results' when ready</div> <div>Missing non coding genes in your mart query output, please check the following <a href="#">FAQ</a></div> <div><div><input type="radio"/> Features</div><div><input type="radio"/> Variant (Germline)</div></div> <div><div><input type="radio"/> Structures</div><div><input type="radio"/> Sequences</div></div> <div><input checked="" type="radio"/> Homologues (Max select 6 orthologues)</div> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

GENE:

ORTHOLOGUES [A-E]:

Abingdon island giant tortoise Orthologues

☐ Abingdon island giant tortoise gene stable ID

☐ Abingdon island giant tortoise gene name

☐ Abingdon island giant tortoise protein or transcript stable ID

☐ Abingdon island giant tortoise chromosome/scaffold name

☐ Abingdon island giant tortoise chromosome/scaffold start (bp)

☐ Abingdon island giant tortoise chromosome/scaffold end (bp)

☐ Query protein or transcript ID

☐ Last common ancestor with Abingdon island giant tortoise

☐ Abingdon island giant tortoise homology type

☐ %id. target Abingdon island giant tortoise gene identical to query gene

☐ %id. query gene identical to target Abingdon island giant tortoise gene

☐ Abingdon island giant tortoise Gene-order conservation score

☐ Abingdon island giant tortoise orthology confidence [0 low, 1 high]

Agassiz's desert tortoise Orthologues

☐ Agassiz's desert tortoise gene stable ID

☐ Agassiz's desert tortoise gene name

☐ Agassiz's desert tortoise protein or transcript stable ID

☐ Last common ancestor with Agassiz's desert tortoise

☐ Agassiz's desert tortoise homology type

☐ %id. target Agassiz's desert tortoise gene identical to query gene

# Matching Gene/Transcript ID

## General databases

**NCBI**

**Ensembl**

**UCSC**

**JGI**

## Species specific databases

**Drosophila: Flybase**

**C. elegans: Wormbase**

**Yeast: SGD**

**Maize: MaizeGDB**

**Arabidopsis: TAIR**

Ensembl BioMart provides IDs  
from many different databases

### External References (max 3)

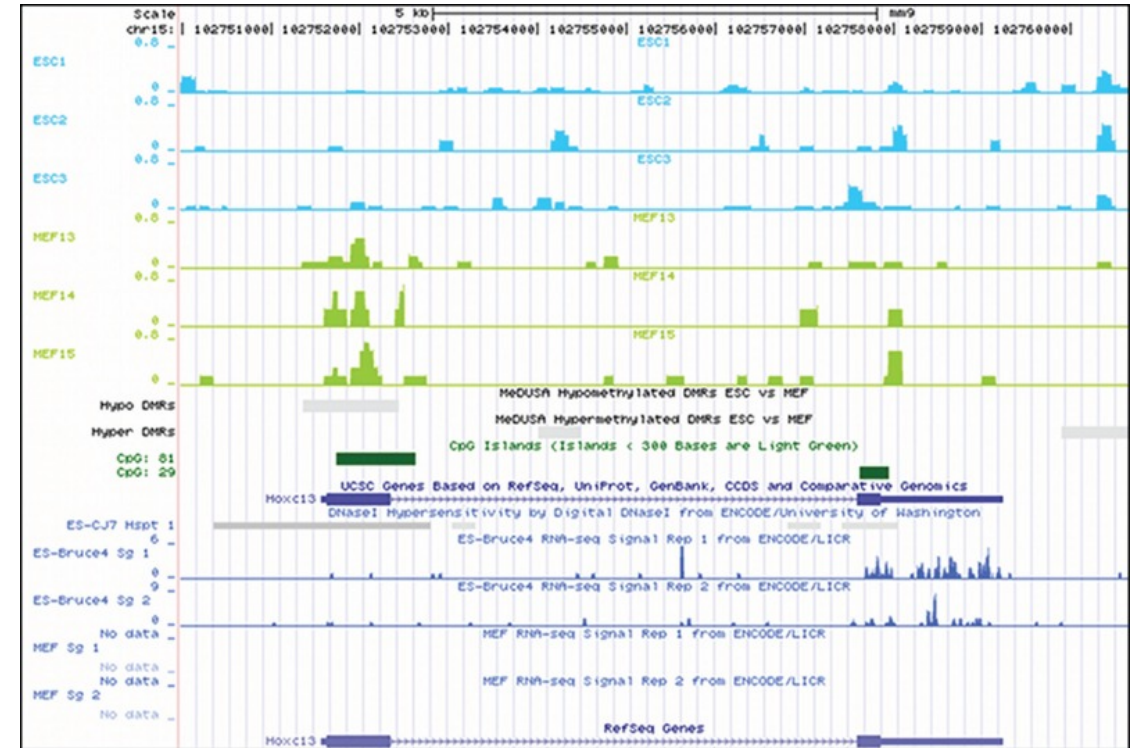
- |                                                              |                                                      |
|--------------------------------------------------------------|------------------------------------------------------|
| <input type="checkbox"/> CCDS ID                             | <input type="checkbox"/> PDB ID                      |
| <input type="checkbox"/> ChEMBL ID                           | <input type="checkbox"/> Reactome ID                 |
| <input type="checkbox"/> Clone-based (Ensembl) gene ID       | <input type="checkbox"/> Reactome gene ID            |
| <input type="checkbox"/> Clone-based (Ensembl) transcript ID | <input type="checkbox"/> RefSeq mRNA ID              |
| <input type="checkbox"/> EntrezGene transcript name ID       | <input type="checkbox"/> RefSeq mRNA predicted ID    |
| <input type="checkbox"/> European Nucleotide Archive ID      | <input type="checkbox"/> RefSeq ncRNA ID             |
| <input type="checkbox"/> Expression Atlas ID                 | <input type="checkbox"/> RefSeq ncRNA predicted ID   |
| <input type="checkbox"/> GeneDB ID                           | <input type="checkbox"/> RefSeq peptide ID           |
| <input type="checkbox"/> Havana gene ID                      | <input type="checkbox"/> RefSeq peptide predicted ID |
| <input type="checkbox"/> Havana transcript ID                | <input type="checkbox"/> RFAM ID                     |
| <input type="checkbox"/> Havana translation ID               | <input type="checkbox"/> RNACentral ID               |
| <input type="checkbox"/> HGNC ID                             | <input type="checkbox"/> Transcript name ID          |
| <input type="checkbox"/> HGNC symbol                         | <input type="checkbox"/> UCSC Stable ID              |
| <input type="checkbox"/> INSDC protein ID                    | <input type="checkbox"/> UniParc ID                  |
| <input type="checkbox"/> KEGG Pathway and Enzyme ID          | <input type="checkbox"/> UniProtKB Gene Name symbol  |
| <input type="checkbox"/> MEROPS - the Peptidase Database ID  | <input type="checkbox"/> UniProtKB Gene Name ID      |
| <input type="checkbox"/> MGI description                     | <input type="checkbox"/> UniProtKB/Swiss-Prot ID     |
| <input type="checkbox"/> MGI symbol                          | <input type="checkbox"/> UniProtKB/TrEMBL ID         |
| <input type="checkbox"/> MGI ID                              | <input type="checkbox"/> Vega gene ID                |
| <input type="checkbox"/> MGI transcript name ID              | <input type="checkbox"/> Vega transcript ID          |
| <input type="checkbox"/> miRBase accession                   | <input type="checkbox"/> Vega translation ID         |
| <input type="checkbox"/> miRBase ID                          | <input type="checkbox"/> WikiGene description        |



# Genome browsers



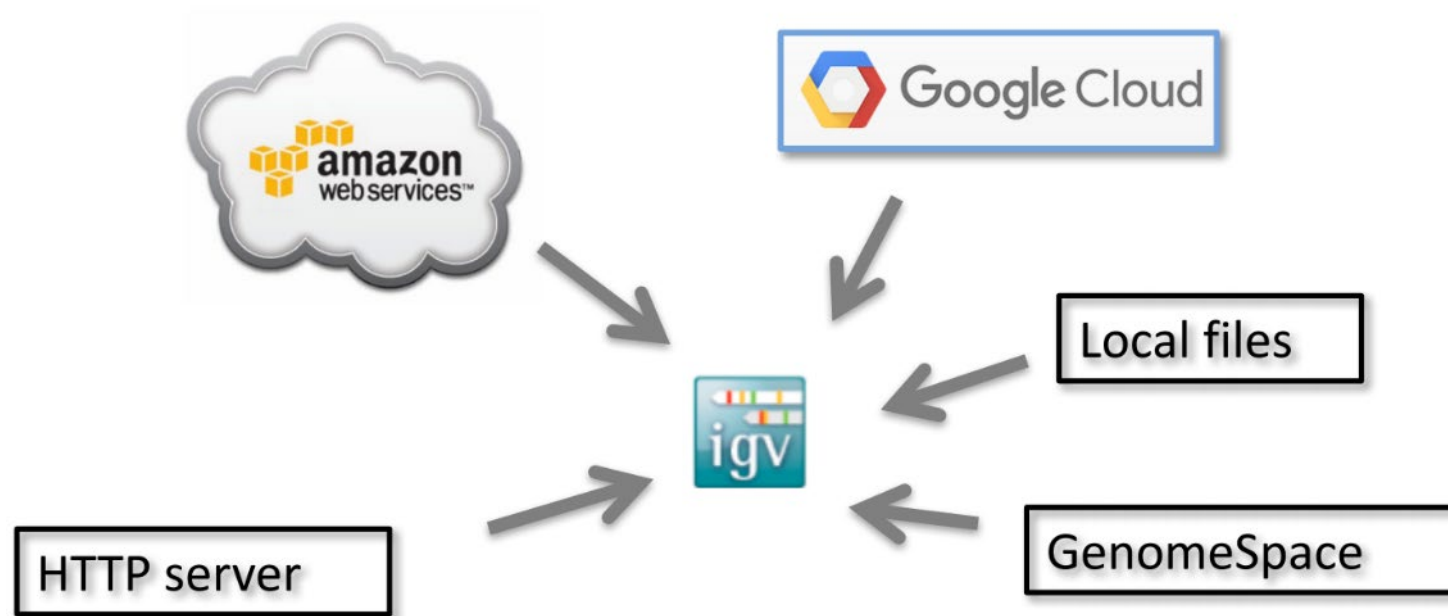
**IGV – desktop application**



**UCSC Genome browser  
– web based**

# IGV – Integrated Genome Viewer

**Software running on your laptop/desktop**

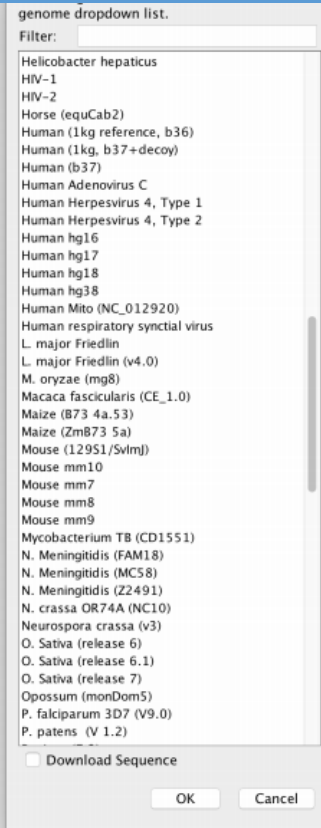
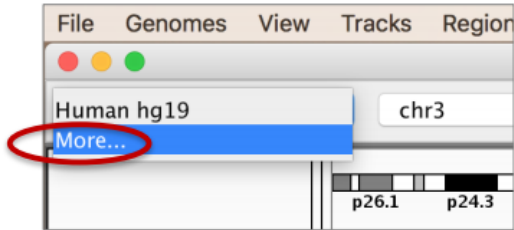




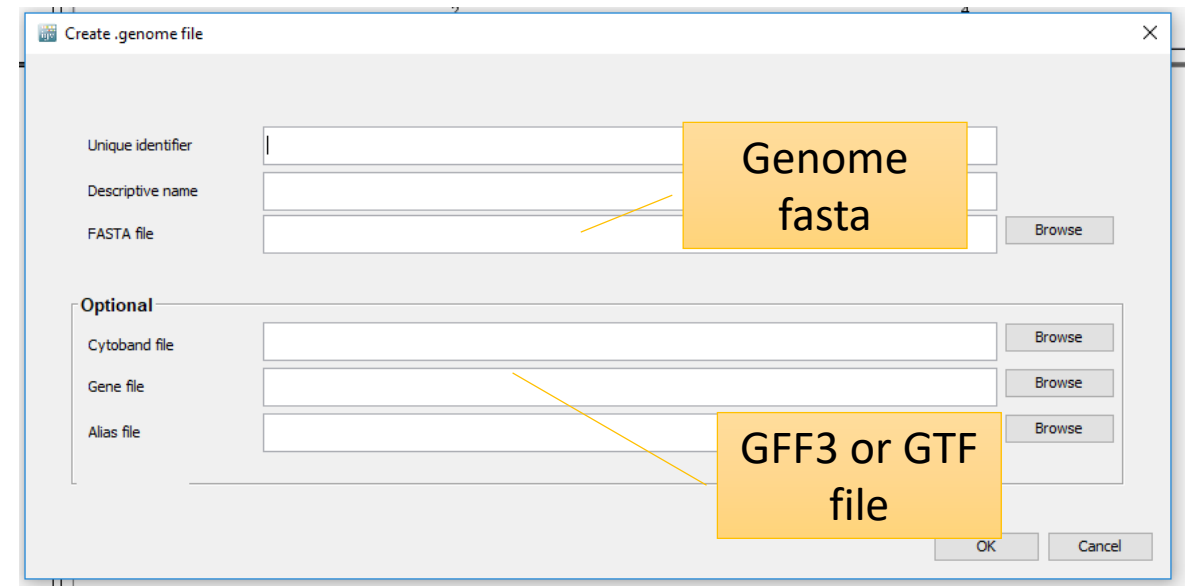
## Select pre-loaded reference genome

To get more genomes from the IGV server:

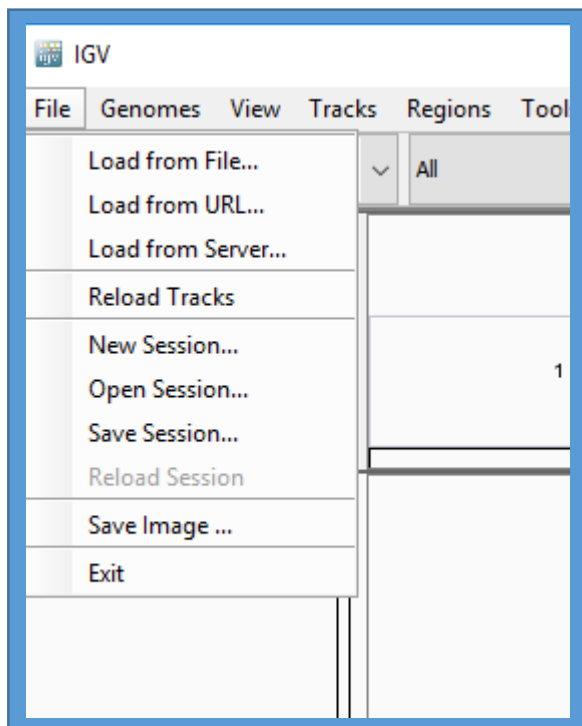
- (1) Select **More** from the genome menu to see the complete list



## Create your own reference genome db



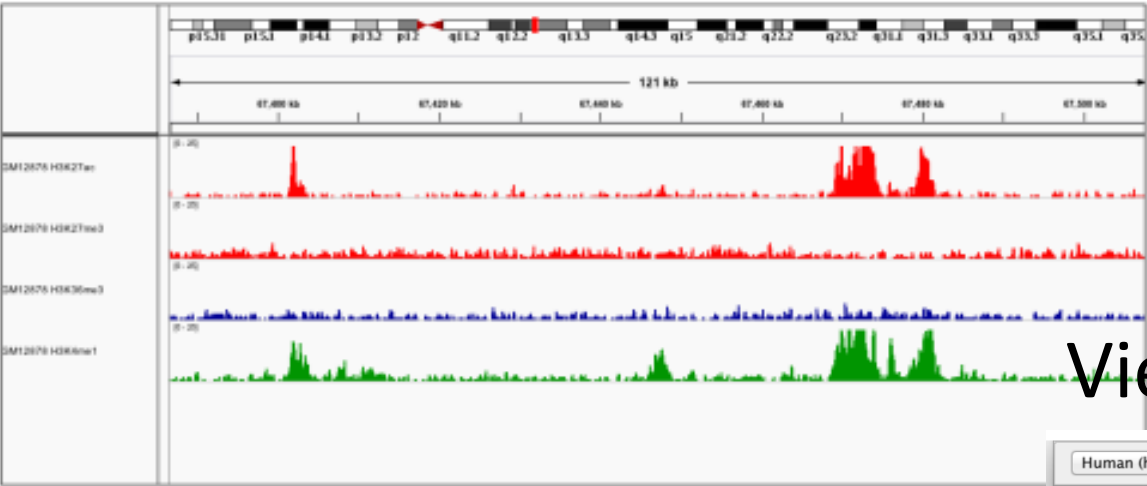
# Load data track



IGV supports many different file formats

- [BAM](#)
- [BED](#)
- [BedGraph](#)
- [bigBed](#)
- [bigWig](#)
- [Birdsuite Files](#)
- [broadPeak](#)
- [CBS](#)
- [CN](#)
- [Custom File Formats](#)
- [Cytoband](#)
- [FASTA](#)
- [GCT](#)
- [genePred](#)
- [GFF/GTF](#)
- [GISTIC](#)
- [Goby](#)
- [GWAS](#)
- [IGV](#)
- [LOH](#)
- [MAF \(Multiple Alignment Format\)](#)
- [MAF \(Mutation Annotation Format\)](#)
- [Merged BAM File](#)
- [MUT](#)
- [narrowPeak](#)
- [PSL](#)
- [SAM](#)
- [Sample Info \(Attributes\) file](#)
- [SEG](#)
- [SNP](#)
- [TAB](#)
- [TDF](#)
- [Track Line](#)
- [Type Line](#)
- [VCF](#)
- [WIG](#)
- [chrom.sizes](#)

# View multiple tracks



View large data files, not need to upload

