

Lesson 12

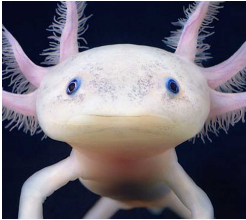
Genome Annotation & Genomic DBs

By the end of this lesson you will...

- Understand the basic concepts of annotation
- Be familiar with two common annotation file formats
 - BED
 - GFF3
- Know how to use the Bedtools software for working with annotations
- Know how to browse and download data from some genomic DBs:
 - SRA/ENA - for raw sequencing data
 - ENSEMBL - for assemblies, annotations, and more

What is genome annotation?

Let's say you have sequenced and assembled a new genome:

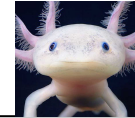


```
CGTTGTTGTGATTCCACTCTATTGAGGCATTAAGTATGCGGAA  
GGAGATCTGGAATGAACTGGCCTATGTCACAGAACTGTGCAA  
ATACCCAATGTCGTTAGTGTAGGTTCTGACCGATACGTGCTTCG  
TTGAGAACTCACAATTTTACAAGTGGGGACATAAACCTACGCC  
CATCATCTACTGACGTCCCTGAGGCTCCAGTTCATGTAATGGGA  
GAGTATCCGCCGCAAGATCTAGTGCAATGGTGGTATAGTAAGCT  
CGTACTGTAGTAGAGGCGACACGGGTAGGATCATCAGTAATAA  
GGATAGTGGGAAAGCTCACAGACCACCGCCTATAGG...
```

Is this useful?

What is genome annotation?

- Labeling of genomic regions/loci
- Detection of genes or other functional sequences
- Assignment of specific functions to loci
- What can we annotate?
 - Protein coding / noncoding genes
 - Gene structure (introns, exons, UTRs)
 - Repetitive elements
 - Transposable elements
 - Regulatory sequences
 - ...



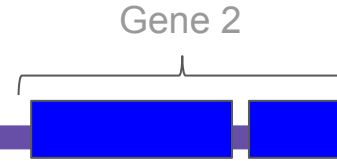
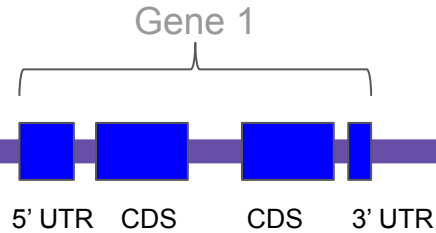
```
CGTTGTTGTGATTCCACTCTATTGAGGCATTAAGTATGCGGAA
GGAGATCTGGAATGAAGTGGCCTATGTACAGAACTGTGCAA
ATACCAATGTCGTTAGTGTAGTTCTGACCGATACGTGCTTCG
TTGAGAACTCACAATTTTACAACTGGGGACATAAACCCCTACGCC
CATCATCTACTGACGTCCCTGAGGCTCCAGTTCATGTAATGGGA
GAGTATCCGCCGCAAGATCTAGTGCAATGGTGGTATAGTAAGCT
CGTACTGTAGTAGAGGCGACACGGGTAGGATCATCAGTAATAA
GGATAGTGGGAAAGCTCACAGACCACCGCCTATAGG...
```

Structural and Functional annotation

Genome assembly

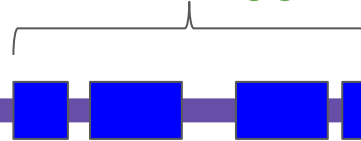


**Structural
annotation**
Where are the
genes?

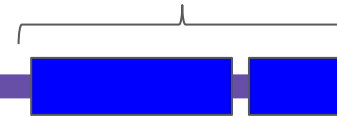


**Functional
annotation**
What do they
do?

DNAse I coding gene



tRNA 7 gene



How can we represent annotations?

- The basic idea - define genomic regions
- Use assembly-specific coordinates
- Add relevant label

Chromosome	Start	End	Label
Chr1	1000	2500	DNAse I gene
Chr1	3000	3500	Predicted gene - unknown function
Chr3	2000	4000	Transposable element

The BED format

- (genome-) **B**rowser **E**xtensible **D**ata
- Simple format for labeling genomic regions
- A TSV text file (table)
- Can have 3-12 columns - usually 4:
 - Chromosome
 - Start
 - End
 - Name (optional)
- Coordinates start from 0
- End coordinate is excluded

ChrI	22230	22552
ChrI	138831	138992
ChrI	160105	160237
ChrI	160238	160574
ChrI	165826	166162
ChrI	182620	182959
ChrI	183142	183474
ChrI	189426	189757
ChrI	209448	209778
ChrII	8848	9092
ChrII	9093	9424
ChrII	9425	9518
ChrII	29644	29975
ChrII	35271	35602
ChrII	35604	35796
ChrII	35851	36221
ChrII	197016	197320
ChrII	197714	198054
ChrII	220895	221036
ChrII	221037	221370
ChrII	226621	226952
ChrII	258670	258976
ChrII	259578	259909
ChrII	265163	265494
ChrII	266178	266256
ChrII	327069	327394
ChrII	327390	327704
ChrII	350459	350751
ChrII	643488	643858
ChrII	643859	644002
ChrII	644362	644603
ChrIII	1179	1429
ChrIII	4073	4322
ChrIII	82700	83036
ChrIII	83055	83194
ChrIII	84069	84294

scaffold_1009275		43	571	Gene9_mRNA1
scaffold_103827	497	719		Gene306_mRNA1
scaffold_103827	719	2761		Gene307_mRNA1
scaffold_103827	80	335		Gene308_mRNA1
scaffold_103827	418	631		Gene309_mRNA1
scaffold_1040613		110	296	Gene123_mRNA1
scaffold_1078619		2521	6561	Gene161_mRNA1
scaffold_1088826		141	1107	Gene71_mRNA1
scaffold_1152430		969	4537	Gene70_mRNA1
scaffold_1164182		2332	4897	Gene142_mRNA1
scaffold_1165675		1215	2448	Gene152_mRNA1
scaffold_1209080		439	1101	Gene337_mRNA1
scaffold_1223358		127	370	Gene42_mRNA1
scaffold_1226651		192	1911	Gene64_mRNA1
scaffold_1282665		956	2350	Gene285_mRNA1
scaffold_129756	303	871		Gene73_mRNA1
scaffold_13509	1456	1639		Gene110_mRNA1
scaffold_13509	1700	2327		Gene111_mRNA1
scaffold_13509	2870	3536		Gene112_mRNA1
scaffold_1374021		118	694	Gene195_mRNA1
scaffold_1403478		1293	1880	Gene331_mRNA1

1	9999941	9999946	7	
1	9999946	9999966	8	
1	9999966	9999982	9	
1	9999982	9999983	10	
1	9999983	9999984	11	
1	9999984	9999997	12	
1	9999997	10000001		13
1	10000001		10000003	14
1	10000003		10000006	15
1	10000006		10000008	14
1	10000008		10000010	15
1	10000010		10000011	16
1	10000011		10000014	17
1	10000014		10000015	18
1	10000015		10000016	19
1	10000016		10000017	18
1	10000017		10000026	20
1	10000026		10000029	21
1	10000029		10000031	22
1	10000031		10000033	21
1	10000033		10000034	20
1	10000034		10000038	19
1	10000038		10000040	18
1	10000040		10000041	19
1	10000041		10000042	18
1	10000042		10000045	17
1	10000045		10000046	19
1	10000046		10000047	20
1	10000047		10000056	21
1	10000056		10000057	22
1	10000057		10000067	23
1	10000067		10000069	22

The GFF format

- **General Feature Format**
- A **hierarchical** format for describing genomic features
- A TSV text file
- **Watch out for different GFF versions - it's a mess!**
 - GFF
 - GFF2 \approx GTF
 - GFF3

The GFF3 format

- Always starts with the line:
`##gff-version 3`
- 9 mandatory columns
- Coordinates start from 1
- End coordinate is included

	Name	Description
1	seqid	Chromosome/scaffold name
2	source	Program or DB that created the annotation
3	type	Feature type (gene, exon, CDS...)
4	start	Start position on chromosome/scaffold
5	end	End position on chromosome/scaffold
6	score	Some score we assign to the feature (or .)
7	strand	+/- (or .)
8	phase	0,1 or 2 - indicating the coding frame (or .)
9	attributes	Additional information about the feature

GFF3 - example

```
##gff-version 3
ChrI   SGD   chromosome   1       230218   .       .       .       ID=chrI;dbxref=NCBI:NC_001133;Name=chrI
ChrI   SGD   telomere       1       801     .       -       .       ID=TEL01L;Name=TEL01L;Note=Telomeric%20region%20on%20the%2
ChrI   SGD   X_element      337     801     .       -       .       ID=TEL01L_X_element;Name=TEL01L_X_element;dbxref=SGD:S0000
ChrI   SGD   X_element_combinatorial_repeat 63      336     .       -       .       ID=TEL01L_X_element_combinatorial_repeat;N
ChrI   SGD   telomeric_repeat 1       62      .       .       .       ID=TEL01L_telomeric_repeat;Name=TEL01L_telomeric_r
ChrI   SGD   gene          335     649     .       +       .       ID=YAL069W;Name=YAL069W;Ontology_term=G0:0003674,G0:0005575,G0:000
ChrI   SGD   mRNA          335     649     .       +       .       ID=YAL069W_mRNA;Name=YAL069W_mRNA;Parent=YAL069W
ChrI   SGD   exon          335     649     .       +       0       Parent=YAL069W_mRNA;Name=YAL069W_exon;orf_classification=Dubious
ChrI   SGD   CDS           335     649     .       +       0       Parent=YAL069W_mRNA;Name=YAL069W_CDS;orf_classification=Dubious
ChrI   SGD   ARS           707     776     .       .       .       ID=ARS102;Name=ARS102;Alias=ARSI-1;Note=Autonomously%20Replicating
ChrI   SGD   gene          87286   87752   .       +       .       ID=YAL030W;Name=YAL030W;gene=SNC1;Alias=SNC1,SNAP%20receptor%20SNC
ChrI   SGD   mRNA          87286   87752   .       +       .       ID=YAL030W_mRNA;Name=YAL030W_mRNA;Parent=YAL030W
ChrI   SGD   exon          87286   87387   .       +       0       Parent=YAL030W_mRNA;Name=YAL030W_exon;orf_classification=Verified
ChrI   SGD   CDS           87286   87387   .       +       0       YAL030W_CDS;orf_classification=Verified
ChrI   SGD   exon          87501   87752   .       +       .       YAL030W_exon;orf_classification=Verified
ChrI   SGD   CDS           87501   87752   .       +       .       YAL030W_CDS;orf_classification=Verified
```

What Linux command
should we use to
extract features of type
“gene” located on
chromosome II?

Viewing BED and GFF files in IGV



Common pitfalls - BED/GFF3 files

- Annotation file doesn't match genome assembly version
- Different chromosome names
 - "Chr1" vs. "1"
 - "ChrI" vs. "S288C_Chrl"
- BED - shifted coordinates (remember it starts from 0)

Bedtools - working with annotation files

- Can read BED, GFF and BAM files
- Contains many useful features
- Use:

```
$ bedtools <subcommand> -h
```

to get help

```
(NGS_course_new) [taungs@hpcssd ~]$ bedtools
bedtools: flexible tools for genome arithmetic and DNA sequence analysis.
usage: bedtools <subcommand> [options]

The bedtools sub-commands include:

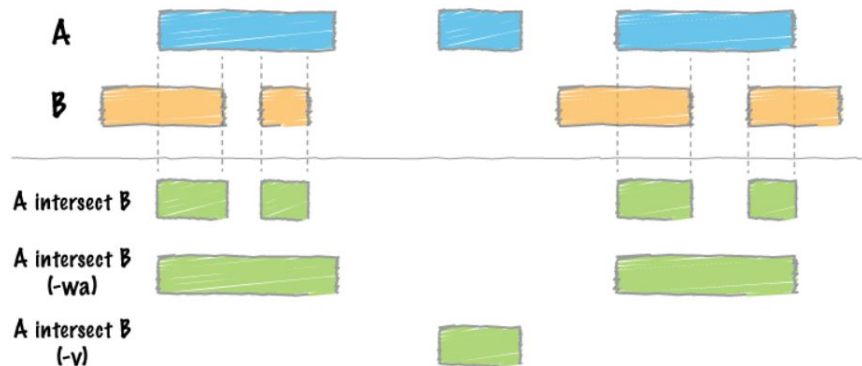
[ Genome arithmetic ]
  intersect      Find overlapping intervals in various ways.
  window         Find overlapping intervals within a window around an interval.
  closest        Find the closest, potentially non-overlapping interval.
  coverage       Compute the coverage over defined intervals.
  map            Apply a function to a column for each overlapping interval.
  genomcov      Compute the coverage over an entire genome.
  merge          Combine overlapping/nearby intervals into a single interval.
  cluster        Cluster (but don't merge) overlapping/nearby intervals.
  complement     Extract intervals _not_ represented by an interval file.
  shift          Adjust the position of intervals.
  subtract       Remove intervals based on overlaps b/w two files.
  slop           Adjust the size of intervals.
  flank          Create new intervals from the flanks of existing intervals.
  sort           Order the intervals in a file.
  random         Generate random intervals in a genome.
  shuffle        Randomly redistribute intervals in a genome.
  sample         Sample random records from file using reservoir sampling.
  spacing        Report the gap lengths between intervals in a file.
  annotate       Annotate coverage of features from multiple files.

[ Multi-way file comparisons ]
  multiinter     Identifies common intervals among multiple interval files.
  unionbedg      Combines coverage intervals from multiple BEDGRAPH files.

[ Paired-end manipulation ]
  pairtobed      Find pairs that overlap intervals in various ways.
  pairtopair     Find pairs that overlap other pairs in various ways.
```

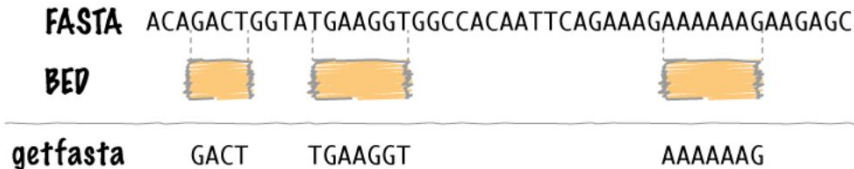
Bedtools - useful commands

intersect



```
$ bedtools intersect -a file1.gff  
-b file2.bed > intersect.gff
```

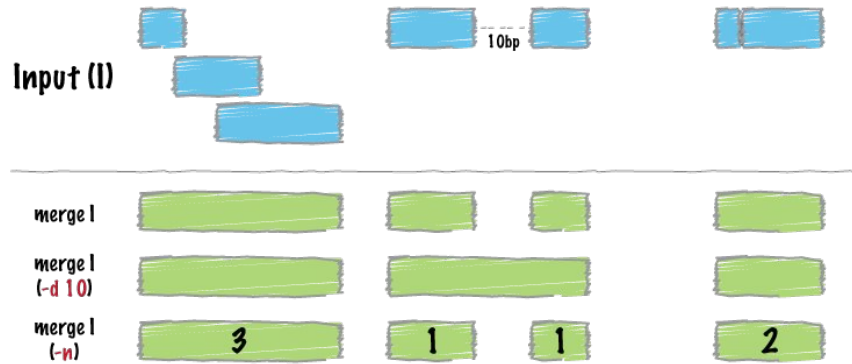
getfasta



```
$ bedtools getfasta -fi  
genome.fasta -bed regions.bed >  
regions.fasta
```

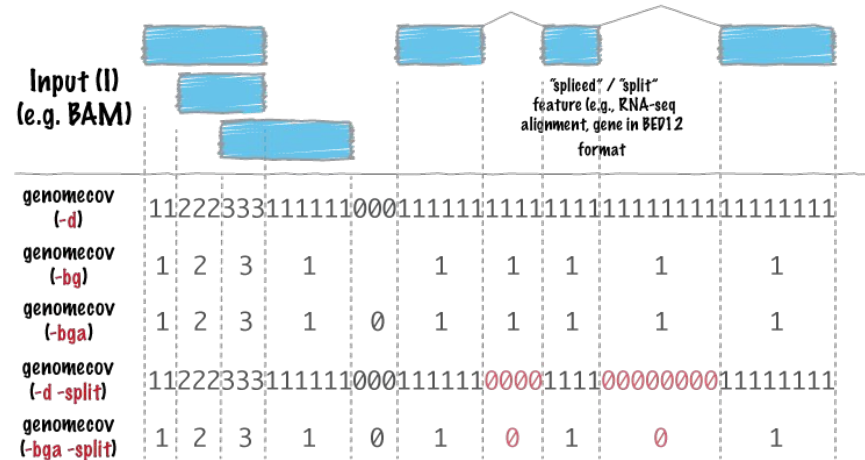
Bedtools - useful commands

Merge



```
$ bedtools merge -i
file.bed > merged.bed
```

genomecov



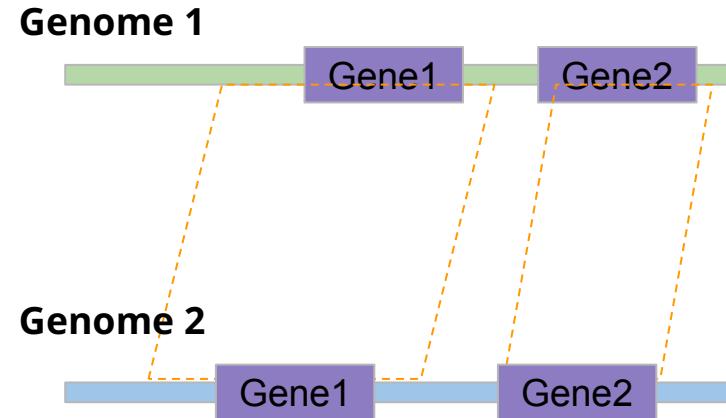
```
$ bedtools genomecov -ibam file.bam  
-bg >read_counts.bedGraph
```


How genome annotations are created?

1. Annotation lift-over
2. ORF finders
3. Ab-initio gene prediction
4. Evidence-based annotation
5. Combine everything - annotation pipelines
6. Manual curation

Annotation lift-over

- Transform gene coordinates from one assembly to another
- Relevant for:
 - Different assembly of the same genome
 - Different individual from the same species
 - Closely-related species



ORF finders

- Look for Open Reading Frames
- Mainly relevant for prokaryotes
- High ratio of false positives

CGGATCTCGGATATGAGACCACTC ... ATTTAGATCGCTAGTAGACCCACATA

Ab initio gene prediction

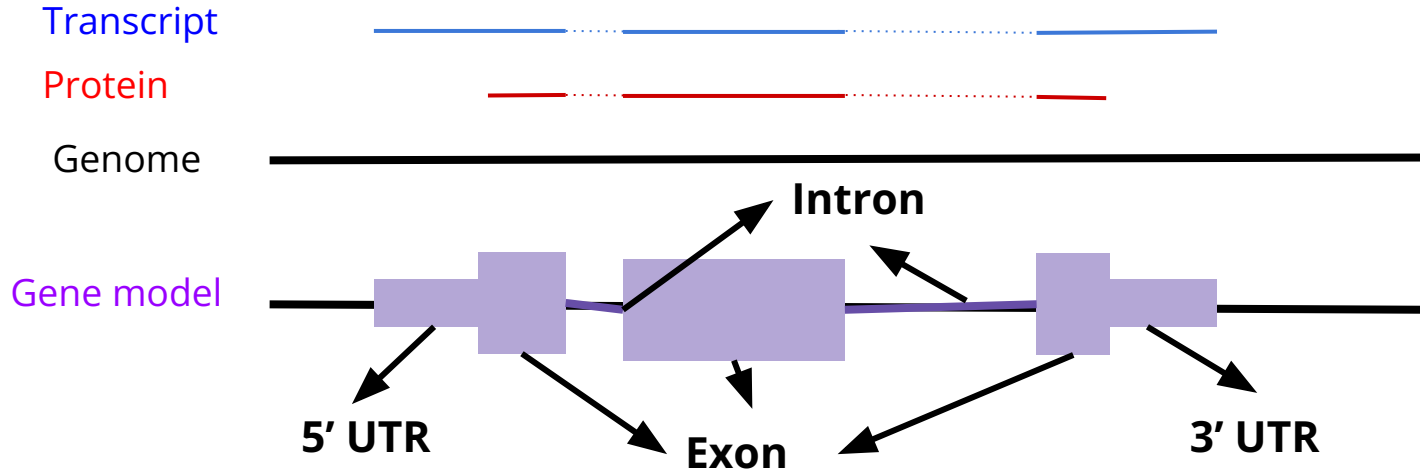
Use mathematical models to describe gene structures



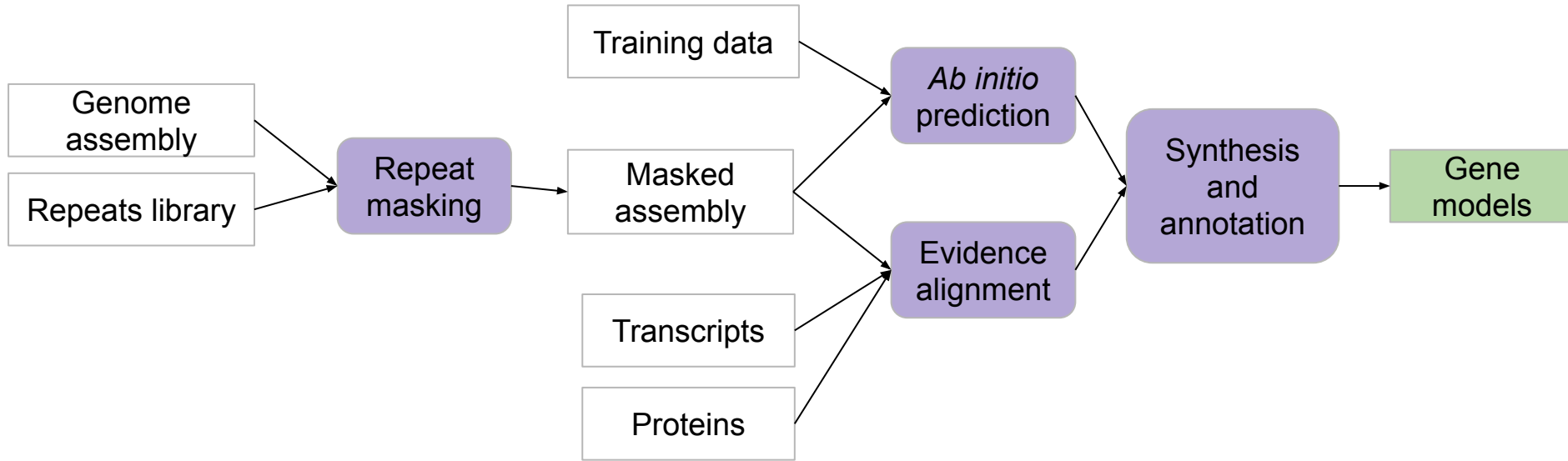
Evidence-based annotation

Transcript and protein sequences can be used as gene evidence

1. Map evidence to genome sequence
2. Infer gene location and structure

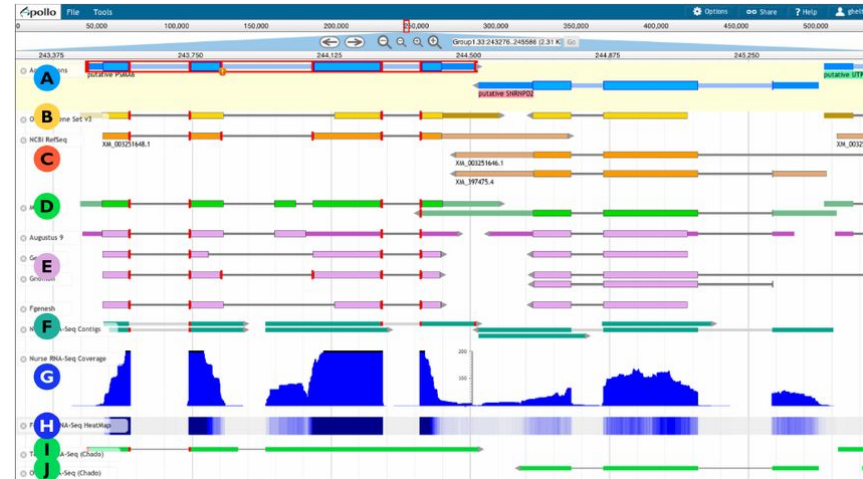


A typical annotation pipeline



Manual curation

- Manual inspection of gene models vs. evidence
- Manually fixing models when needed



Still noisy...

Even with best evidence and predictions -

Automatic annotation is error-prone!

Both missing predictions and false predictions

Manual curation is too expensive - most genes are never curated



What about functional annotation?

- Usually based on **homology**
- Similarity to known proteins
- Presence of protein domains
- Assignment to a gene family
- Main methods:
 - Best BLAST hit
 - InterproScan



Genomic DBs

Downloading data with wget

- The `wget` command allows you to download files from HTTP/HTTPS/FTP URLs
- Input: URL address
- By default, downloads to `./<original file name>`
- To specify another location - use `-O`

```
wget  
"ftp://ftp.ensemblgenomes.org/pub/bacteria/release-47/gff3/bacteria_4  
5_collection/abiotrophia_defectiva_atcc_49176/README" -O dir/new_file
```

- You may need to use the `--no-check-certificate` flag, but only for trusted websites!

Working with compressed data - tar/gzip

- Files can be compressed (zipped) to save space
- Multiple files can be bundled together into one tar archive

```
# compress a file
gzip file.txt
# decompress file
gzip -d file.txt.gz
```

```
# tar multiple files
tar -cvf my.tar file1 file2 file3
# tar a directory
tar -cvf my.tar dir/
# untar an archive
tar -xvf my.tar
```

- In many cases, the tar archive will also be compressed - .tar.gz / .tgz

```
# extract and decompress file
tar -zxvf file.tar.gz
```

SRA - Sequence Read Archive

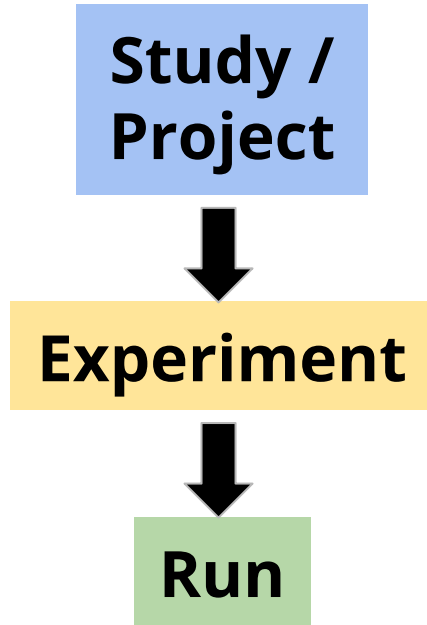
- Maintained by NCBI (and others)
- As of 8.6.2021, contains **$\sim 5.44 \times 10^{16}$** bases (= 5,440 Tb) of reads from $\sim 72,000$ species
- Contains multiple types of reads
 - Short/long reads
 - DNA/RNA-seq reads
- Most common repository for depositing NGS reads
- Data are stored in base-call format (.sra)
- To download data in FastQ format - use [SRA toolkit](#)

ENA - European Nucleotide Archive

- Maintained by the European Bioinformatics Institute
- Synchronized with SRA
- Stores reads as FastQ files
- Provides an easy and fast interface to the data
- <https://www.ebi.ac.uk/ena>



SRA/ENA data organization and accession IDs



Description	Prefix	Example
A study or sequencing project	SRP	SRP056687
A sequencing experiment	SRX	SRX972180
A specific sequencing run	SRR	SRR1945497

Searching ENA by accession IDs

Data availability

Raw genome and RNA-Seq reads have been deposited into the National Center for Biotechnology Information Sequence Read Archive under accession codes [SRP150040](#), [SRP186721](#) and [SRP172989](#), respectively. The nonreference genome sequences and annotated genes of the tomato pan-genome and SNPs called from the RIL population are available via the Dryad Digital Repository (<https://doi.org/10.5061/dryad.m463f7k>).

Search 

Examples: histone, RN000065

View 

Search results page

 [Download All](#)

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Library Name	FASTQ FTP
PRJNA454805	SAMN09229594	SRX4183310	SRR7279481	195583	Solanum lycopersicum var. cerasiforme	Esther052516_BGV005912	<input type="checkbox"/> SRR727948...fastq.gz
							<input type="checkbox"/> SRR727948...fastq.gz
PRJNA454805	SAMN09229632	SRX4183309	SRR7279482	195583	Solanum lycopersicum var. cerasiforme	Esther052516_BGV006904	<input type="checkbox"/> SRR727948...fastq.gz
							<input type="checkbox"/> SRR727948...fastq.gz
PRJNA454805	SAMN09229595	SRX4183308	SRR7279483	4084	Solanum pimpinellifolium	Esther120315_BGV006148	<input type="checkbox"/> SRR727948...fastq.gz
							<input type="checkbox"/> SRR727948...fastq.gz
PRJNA454805	SAMN09229594	SRX4183307	SRR7279484	195583	Solanum lycopersicum var. cerasiforme	Esther120315_BGV005912	<input type="checkbox"/> SRR727948...fastq.gz
							<input type="checkbox"/> SRR727948...fastq.gz
PRJNA454805	SAMN09229602	SRX4183306	SRR7279485	195583	Solanum lycopersicum var. cerasiforme	Esther120315_BGV006232	<input type="checkbox"/> SRR727948...fastq.gz
							<input type="checkbox"/> SRR727948...fastq.gz

ENA advanced search

Filter results by:

- Organism
- Experiment properties
- Sequencing properties
- Date

The screenshot displays the ENA advanced search interface. At the top, a progress bar shows five steps: Data Type, Query, Inclusion/Exclusion, Fields, Data Filters, and Results. The 'Query' step is active, showing a query box with the text: `tax_eq(4932) AND strain="m11" AND instrument_platform="ILLUMINA" AND library_source="TRANSCRIPTOMIC" AND first_public>=2018-01-01`. Below the query box is a 'Build Query' button. To the right of the query box are 'Reset' and 'Update query' buttons. Below the query box is a text input field labeled 'Type to filter query params'. To the left of the query builder is a vertical list of filter categories: Taxonomy and related, Geographical location, Geography, Collection event information, Sampling information, Sample state and conditions, Host information, Methodology, Sequencing information, Database record, File information, Accessions, Titles, aliases and descriptions, Sequenced molecule, and Names and symbols. The 'Sequencing information' category is highlighted. The query builder itself has a dropdown for 'AND OR' and two rules. The first rule is 'Instrument platform' equals 'ILLUMINA' with a 'Delete' button. The second rule is 'Library source' equals 'TRANSCRIPTOMIC' with a 'Delete' button. At the bottom right, there are buttons for 'Back', 'Copy Curl Request', 'Next', and 'Search'.

Query: `tax_eq(4932) AND strain="m11" AND instrument_platform="ILLUMINA" AND library_source="TRANSCRIPTOMIC" AND first_public>=2018-01-01`

Build Query ?

Type to filter query params

AND OR

Instrument platform = ILLUMINA ✕ Delete

Instrument platform used in sequencing experiment

Library source = TRANSCRIPTOMIC ✕ Delete

source material being sequenced

Back Copy Curl Request Next Search

Downloading data from ENA

- **Option 1** - direct download from FTP using wget

```
wget "ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR194/005/SRR1945435/SRR1945435_1.fastq.gz"
wget "ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR194/005/SRR1945435/SRR1945435_2.fastq.gz"
gzip -d *.gz
```

- **Option 2** - use [Kingfisher](#)
 - Requires a simple setup
 - Can enhance download speed by 60 fold
 - Very useful for large data volumes



```
kingfisher get -r ERR1739691 -m ena-ascp
gzip -d *.gz
```



- Ensembl is a web portal for genomic data
- Contains data for thousands of organisms from all kingdoms
 - Genome sequences (assembly)
 - Annotation (structural/functional)
 - Transcripts and proteins
 - Variation (SNPs/SVs)
 - Comparative genomics resources
- Provides easy data access
- Contains a web genome browser
- Offers various tools like BLAST/BLAT search

The Ensembl websites



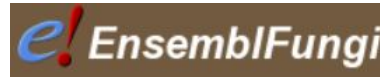
<https://www.ensembl.org>

**Vertebrates
(including human)**



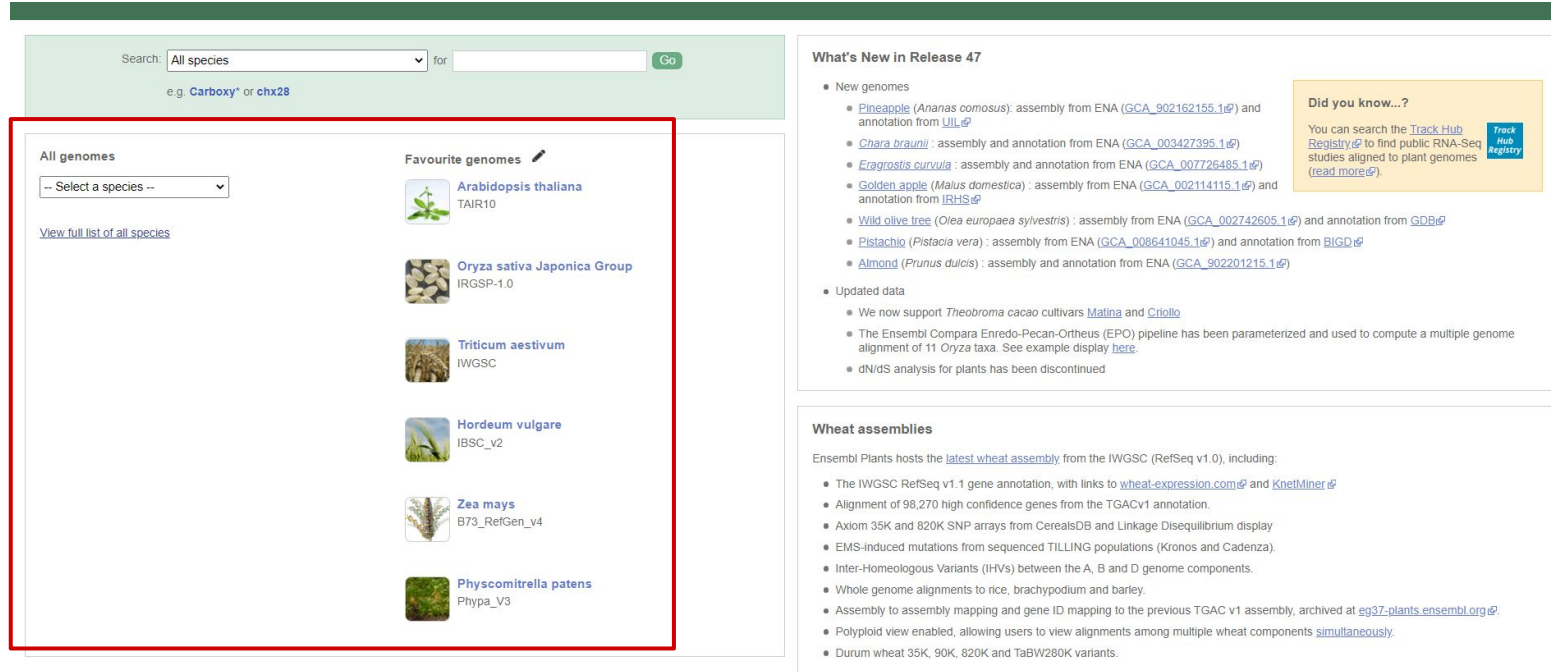
<http://ensemblgenomes.org>

Anything else



Downloading genomic data from Ensembl

1. Find your species of interest



The screenshot displays the Ensembl Plants website interface. At the top, there is a search bar with the text "Search: All species for" and a "Go" button. Below the search bar, there is a link "e.g. Carboxy* or chx28".

The main content area is divided into two columns. The left column, titled "All genomes", contains a dropdown menu labeled "-- Select a species --" and a link "View full list of all species". The right column, titled "Favourite genomes", lists several species with their respective logos and names:

- Arabidopsis thaliana** (TAIR10)
- Oryza sativa Japonica Group** (IRGSP-1.0)
- Triticum aestivum** (IWGSC)
- Hordeum vulgare** (IBSC_v2)
- Zea mays** (B73_RefGen_v4)
- Physcomitrella patens** (Phyba_V3)

On the right side of the page, there is a section titled "What's New in Release 47" which lists new genomes and updated data. Below this, there is a section titled "Wheat assemblies" which lists various wheat assemblies and their features.

What's New in Release 47

- New genomes
 - Pineapple** (*Ananas comosus*): assembly from ENA ([GCA_902162155.1](#)) and annotation from [JGI](#)
 - Chara braunii**: assembly and annotation from ENA ([GCA_003427395.1](#))
 - Eragrostis curvula**: assembly and annotation from ENA ([GCA_007726485.1](#))
 - Golden apple** (*Malus domestica*): assembly from ENA ([GCA_002114115.1](#)) and annotation from [IRIS](#)
 - Wild olive tree** (*Olea europaea sylvestris*): assembly from ENA ([GCA_002742605.1](#)) and annotation from [GDB](#)
 - Pistachio** (*Pistacia vera*): assembly from ENA ([GCA_008641045.1](#)) and annotation from [BIGD](#)
 - Almond** (*Prunus dulcis*): assembly and annotation from ENA ([GCA_902201215.1](#))
- Updated data
 - We now support *Theobroma cacao* cultivars [Mating](#) and [Criollo](#)
 - The Ensembl Comparo Enredo-Pecan-Ortheus (EPO) pipeline has been parameterized and used to compute a multiple genome alignment of 11 *Oryza* taxa. See example display [here](#).
 - dN/dS analysis for plants has been discontinued

Did you know...?

You can search the [Track Hub Registry](#) to find public RNA-Seq studies aligned to plant genomes ([read more](#)).

Wheat assemblies

Ensembl Plants hosts the [latest wheat assembly](#) from the IWGSC (RefSeq v1.0), including:

- The IWGSC RefSeq v1.1 gene annotation, with links to [wheat-expression.com](#) and [KnetMiner](#)
- Alignment of 98,270 high confidence genes from the TGACv1 annotation.
- Axiom 35K and 820K SNP arrays from CerealsDB and Linkage Disequilibrium display
- EMS-induced mutations from sequenced TILLING populations (Kronos and Cadenza).
- Inter-Homologous Variants (IHVs) between the A, B and D genome components.
- Whole genome alignments to rice, brachypodium and barley.
- Assembly to assembly mapping and gene ID mapping to the previous TGAC v1 assembly, archived at [eg37-plants.ensembl.org](#).
- Polyploid view enabled, allowing users to view alignments among multiple wheat components [simultaneously](#).
- Durum wheat 35K, 90K, 820K and TaBW280K variants.

Downloading genomic data from Ensembl

2. Choose the data you are interested in

EnsemblPlants | [HMMER](#) | [BLAST](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Documentation](#) | [Website help](#) | [Login/Register](#)

Zea mays (B73_RefGen_v4) ▾

Search

Search Zea mays... [Go](#)

e.g. [Zm00001d048577](#) or [1:109000-145001](#) or [Carboxy*](#)

About Zea mays

Zea mays (maize) has the highest world-wide production of all grain crops, [yielding 875 million tonnes in 2012](#). Although a food staple in many regions of the world, most is used for animal feed and ethanol fuel. Maize was domesticated from wild teosinte in Central America and its cultivation spread throughout the Americas by Pre-Columbian civilisations. In addition to its economic value, maize is an important model organism for studies in plant genetics, physiology, and development. It has a large genome of about 2.4 gigabases with a haploid chromosome number of 10 ([Schmable et al., 2003](#); [Zhang et al., 2003](#)). Maize is distinguished from other grasses in that its genome arose from an ancient tetraploidy event unique to its lineage.

Taxonomy ID [4571](#)

Data source [Gramene](#)

[More information and statistics](#)

Genome assembly: B73_RefGen_v4

[More information and statistics](#)

[Download DNA sequence \(FASTA\)](#)

[Convert your data to B73_RefGen_v4 coordinates](#)

[Display your data in Ensembl Plants](#)

[View karyotype](#)

[Example region](#)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

[More about this genebuild](#)

[Download genes, cDNAs, ncRNA, proteins - FASTA - GFF3](#)

[Update your old Ensembl IDs](#)

[Example gene](#)

[Example transcript](#)

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

[More about comparative analyses](#)

[Phylogenetic overview of gene families](#)

[Download alignments \(EMF\)](#)

[Genomic alignments \[8\] \[Show\]](#)

[Synteries \[1\] \[Show\]](#)

[Example gene tree](#)

Variation

What can I find? Short sequence variants.

[More about variation in Zea mays](#)

[More about variation in Ensembl Plants](#)

[Download all variants - VCF - VEP](#)

Variant Effect Predictor

[Example variant](#)

Downloading genomic data from Ensembl

3. Copy FTP link and download (wget → gzip)

Genome assembly

All chromosomes

Hard repeat
masked

Soft repeat
masked

	Name	Size	Date Modified
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.1.fa.gz	8.8 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.2.fa.gz	5.8 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.3.fa.gz	6.8 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.4.fa.gz	5.4 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.5.fa.gz	7.8 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.Mt.fa.gz	112 kB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.chromosome.Pt.fa.gz	47.2 kB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz	34.8 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.1.fa.gz	7.4 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.2.fa.gz	4.5 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.3.fa.gz	5.5 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.4.fa.gz	4.3 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.5.fa.gz	6.5 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.Mt.fa.gz	103 kB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.chromosome.Pt.fa.gz	24.3 kB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_rm.toplevel.fa.gz	28.4 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.1.fa.gz	9.2 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.2.fa.gz	6.0 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.3.fa.gz	7.1 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.4.fa.gz	5.6 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.5.fa.gz	8.2 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.Mt.fa.gz	118 kB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.chromosome.Pt.fa.gz	50.6 kB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	Arabidopsis_thaliana.TAIR10.dna_sm.toplevel.fa.gz	36.3 MB	3/4/20, 1:13:00 PM
<input type="checkbox"/>	CHECKSUMS	1.5 kB	3/8/20, 4:01:00 PM
<input type="checkbox"/>	README	4.9 kB	3/4/20, 1:13:00 PM

Downloading genomic data from Ensembl

3. Copy FTP link and download (wget → gzip)







Genome annotation

GFF3

Name	Size	Date Modified
 Arabidopsis_thaliana.TAIR10.47.chromosome.1.gff3.gz	2.4 MB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.chromosome.2.gff3.gz	1.4 MB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.chromosome.3.gff3.gz	1.8 MB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.chromosome.4.gff3.gz	1.4 MB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.chromosome.5.gff3.gz	2.1 MB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.chromosome.Mt.gff3.gz	11.7 kB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.chromosome.Pt.gff3.gz	10.3 kB	3/8/20, 1:26:00 PM
 Arabidopsis_thaliana.TAIR10.47.gff3.gz	9.1 MB	3/8/20, 1:27:00 PM
 CHECKSUMS	520 B	3/8/20, 3:59:00 PM
 README	11.0 kB	3/5/20, 11:57:00 PM

Full length
transcripts

FASTA

Name	Size	Date Modified
 cdna/		3/8/20, 5:02:00 PM
 cds/		3/8/20, 5:02:00 PM
 dna/		3/8/20, 5:02:00 PM
 dna_index/		3/8/20, 5:02:00 PM
 ncrna/		3/8/20, 5:04:00 PM
 pep/		3/8/20, 5:04:00 PM

Transcripts without UTR

proteins

Bulk and advanced downloads

If you want to download lots of data, check out:

A. Ensembl Biomart

- A web-based tool for querying and exporting Ensembl data

B. Bioconductor BiomaRt package

- An R package for accessing the Biomart from within R

The screenshot displays the Ensembl Biomart web interface. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar is located on the right. Below the navigation bar, there are tabs for New, Count, and Results. The main content area is divided into two panels. The left panel, titled 'Dataset', shows 'Mouse genes (GRCm38.p6)' and 'Filters' with 'Chromosome/scaffold: 6' and 'Start: 1' to 'End: 100000'. The 'Attributes' section lists 'Gene stable ID', 'Gene stable ID version', 'Transcript stable ID', and 'Transcript stable ID version'. The right panel shows the export options: 'Export: all results to' (File), 'Email notification to' (empty field), 'View' (10 rows as HTML), and 'Unique results only' (checked). A 'Go' button is present at the bottom right.

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search all species...

New Count Results

URL XML Perl Help

Dataset
Mouse genes (GRCm38.p6)
Filters
Chromosome/scaffold: 6
Start: 1
End: 100000
Attributes
Gene stable ID
Gene stable ID version
Transcript stable ID
Transcript stable ID version
Dataset
[None Selected]

Export: all results to File TSV Unique results only Go

Email notification to

View 10 rows as HTML Unique results only

Gene stable ID Gene stable ID version Transcript stable ID Transcript stable ID version