Lesson 12

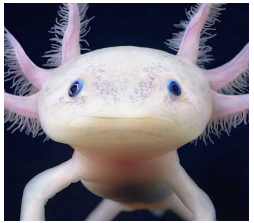
Genome Annotation & Genomic DBs

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:



CGTTGTTGTGATTCCACTCTATTGAGGCATTAACTGATGCGGAA
GGAGATCTGGAATGAACTGGCCTATGTCACAGAAACTGTGCAA
ATACCCAATGTCGTTAGTGTAGGTTCTGACCGATACGTGCTTCG
TTGAGAACTCACAATTTTACAACTGGGGACATAAACCCTACGCC
CATCATCTACTGACGTCCCTGAGGCTCCAGTTCATGTAATGGGA
GAGTATCCGCCGCAAGATCTAGTGCAATGGTGGTATAGTAAGCT
CGTACTGTAGTAGAGGCGACACGGGTAGGATCATCAGTAATAA
GGATAGTGGGAAAGCTCACAGACCACCGCCTATAGG...

Is this useful?

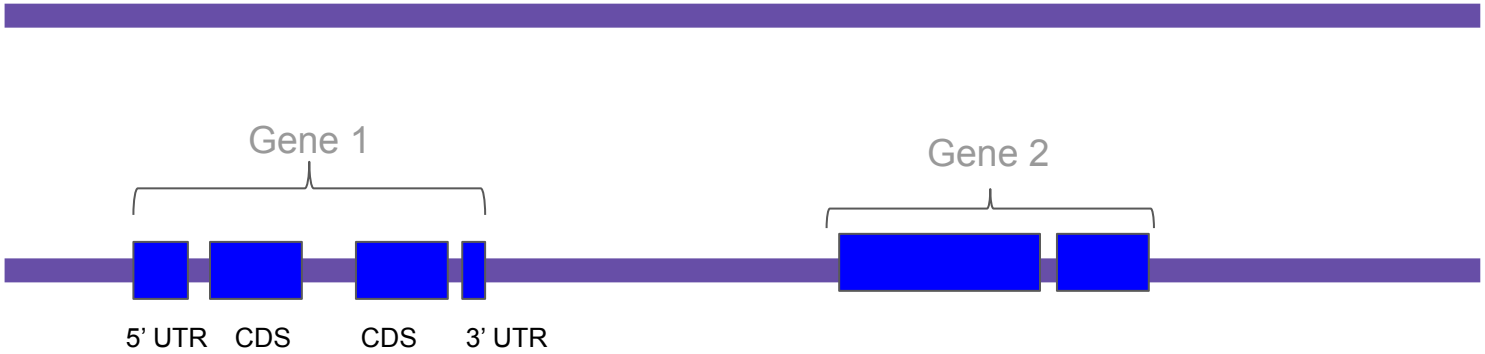
What is genome annotation?

- 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
  - ...

CGTTGTTGTGATTCCACTCTATTGAGGCATTAACTGATGCGGAA
GGAGATCTGGAATGAACTGGCCTATGTCACAGAAACTGTGCAA
ATACCCAATGTCGTTAGTGTAGGTTCTGACCGATACGTGCTTCG
TTGAGAACTCACAATTTTACAACTGGGGACATAAACCCTACGCC
CATCATCTACTGACGTCCCTGAGGCTCCAGTTCATGTAATGGGA
GAGTATCCGCCGCAAGATCTAGTGCAATGGTGGTATAGTAAGCT
CGTACTGTAGTAGAGGCGACACGGGTAGGATCATCAGTAATAA
GGATAGTGGGAAAGCTCACAGACCACCGCCTATAGG...

Structural and Functional annotation

Genome assembly



Structural annotation
Where are the genes?

Functional annotation
What do they do?

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

- **G**eneral **F**eature **F**ormat
- A **hierarchical** format for describing genomic features
- A TSV text file
- **Watch out for different GFF versions - it's a mess!**
  - GFF
  - GFF2 ≈ 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=GO:0003674,GO:0005575,GO: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       030W_CDS;orf_classification=Verified
ChrI    SGD     exon    87501   87752   .       [illegible]     exon;orf_classification=Verified
ChrI    SGD     CDS     87501   87752   .       [illegible]     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_ChrI”

- 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:

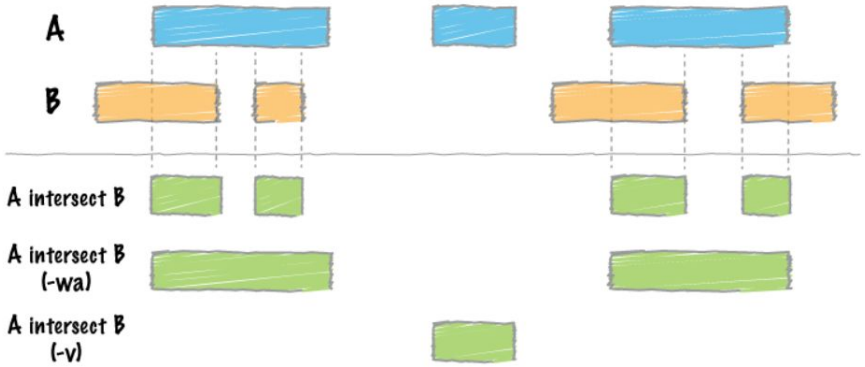
```
$ bedtools <subcommand> -h
```

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.
    genomecov     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 redistrubute 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



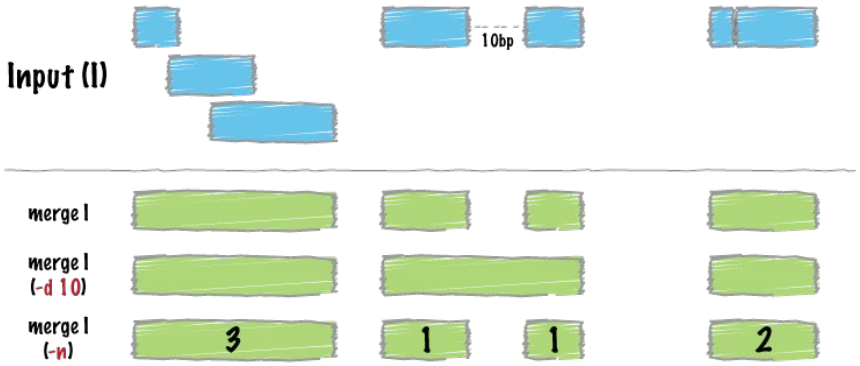

```
$ bedtools intersect -a file1.gff
-b file2.bed > intersect.gff
```

getfasta


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

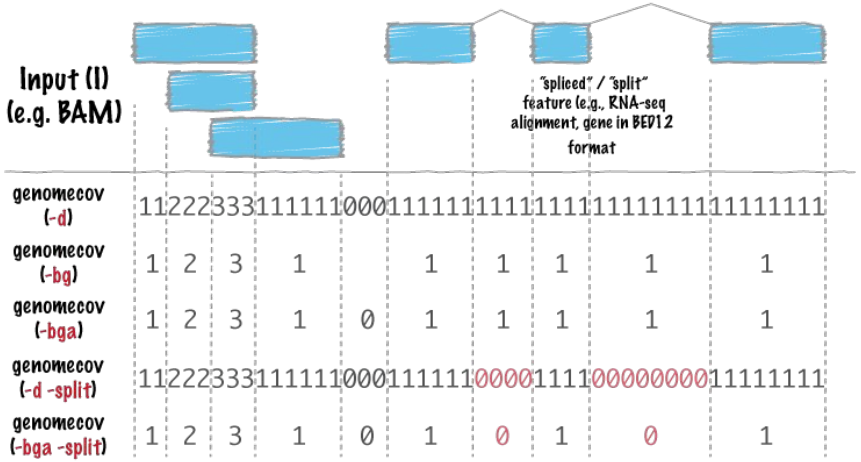
Bedtools - useful commands

Merge




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

genomecov



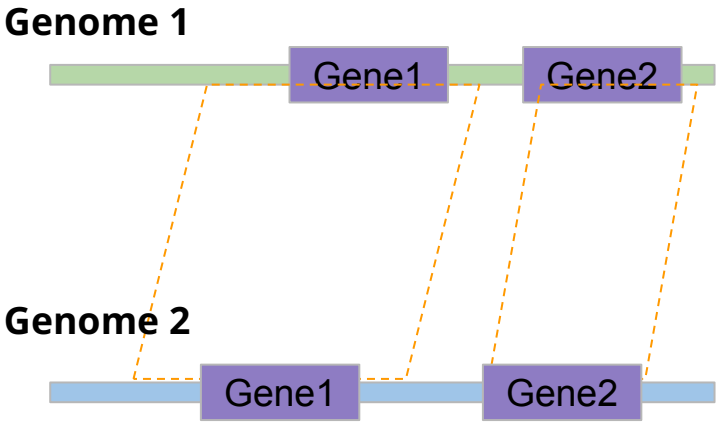

```
$ bedtools genomecov -ibam file.bam
-bg >read_counts.bedGraph
```

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



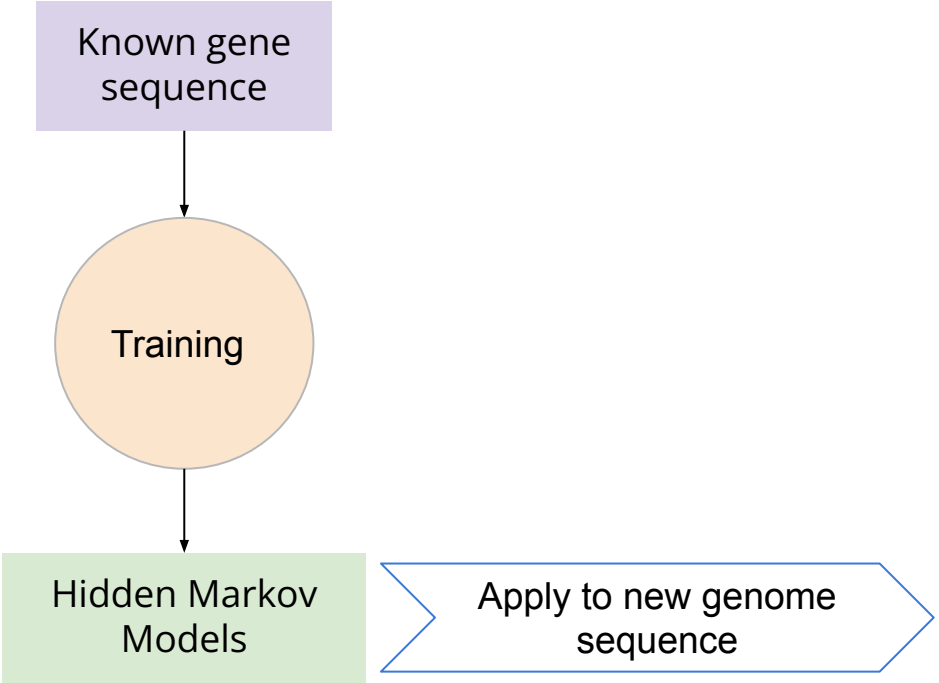
ORF finders

- 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

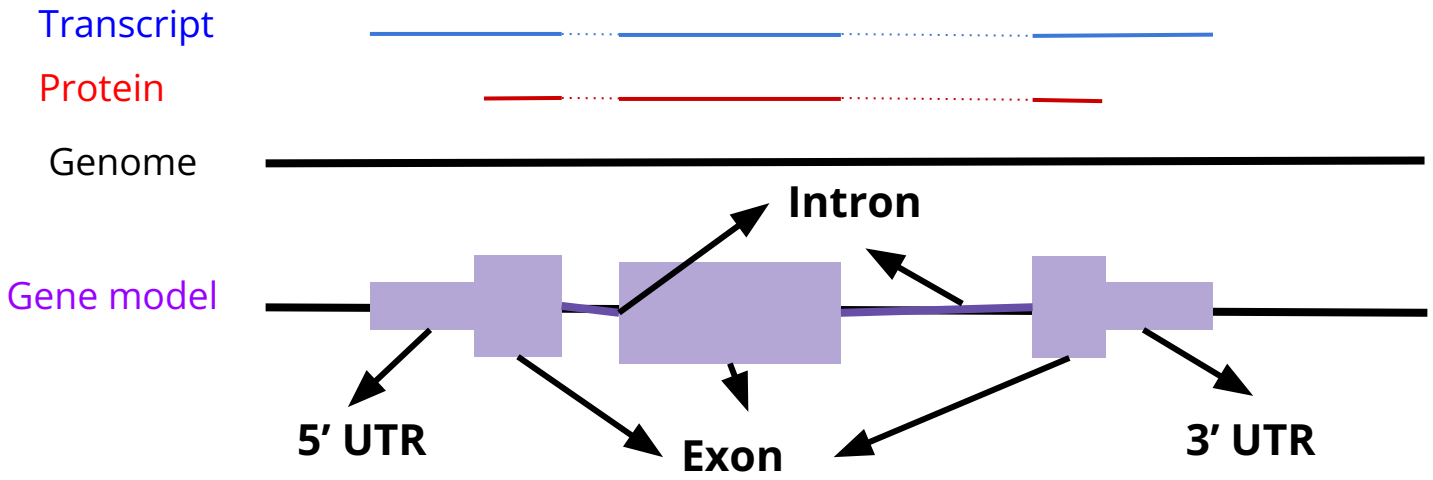


CGTTGTTGTGATTCCACTCTATTGAGGCATTAACTGATGCGGAA
GGAGATCTGGAATGAACTGGCCTATGTCACAGAAACTGTGCAA
ATACCCAATGTCGTTAGTGTAGGTTCTGACCGATACGTGCTTCG
TTGAGAACTCACAATTTTACAACTGGGGACATAAACCCTACGCC
CATCATCTACTGACGTCCCTGAGGCTCCAGTTCATGTAATGGGA
GAGTATCCGCCGCAAGATCTAGTGCAATGGTGGTATAGTAAGCT
CGTACTGTAGTAGAGGCGACACGGGTAGGATCATCAGTAATAA
GGATATAGGGAAAGCTCACAGACCACCGCCTATAGG...

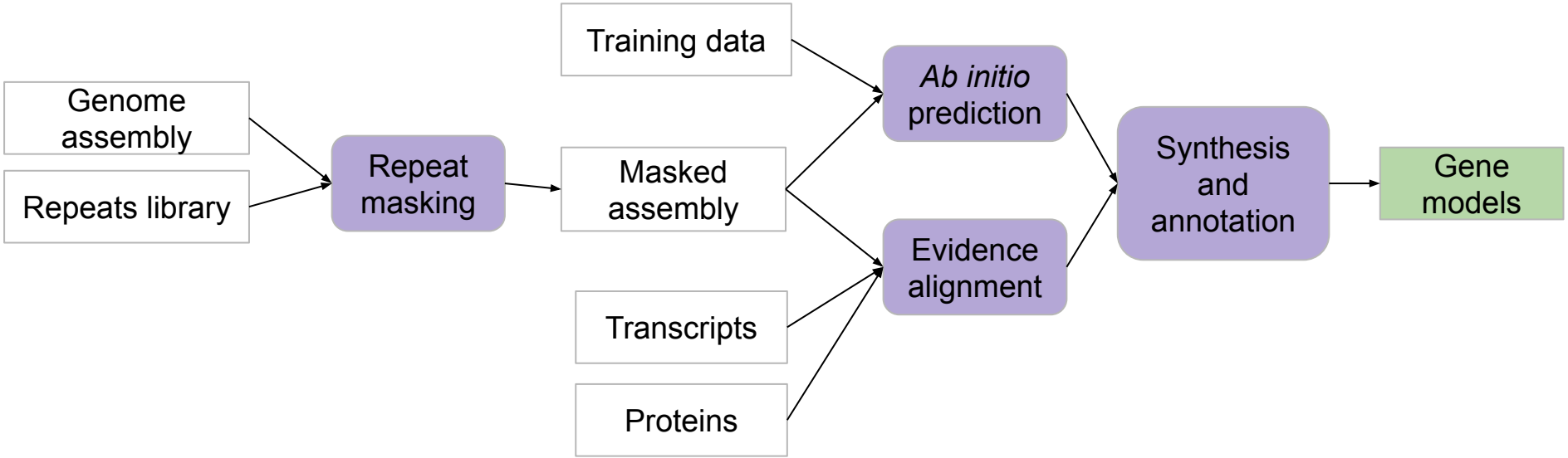
Evidence-based annotation

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

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

- Multiple files can be bundled together into one tar archive

```
# 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

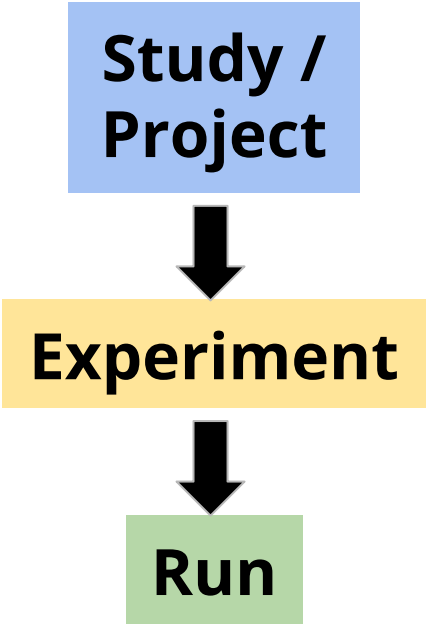
- Maintained by NCBI (and others)
- As of 8.6.2021, contains **~5.44 x 10^{16}** bases (= 5,440 Tb) of reads from ~72,000 species
- Contains multiple types of reads
  - Short/long reads
  - DNA/RNA-seq reads
- Most common repository for depositing NGS reads
- Data are stored in base-call format (.sra)
- 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

ENA
European Nucleotide Archive

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 results page

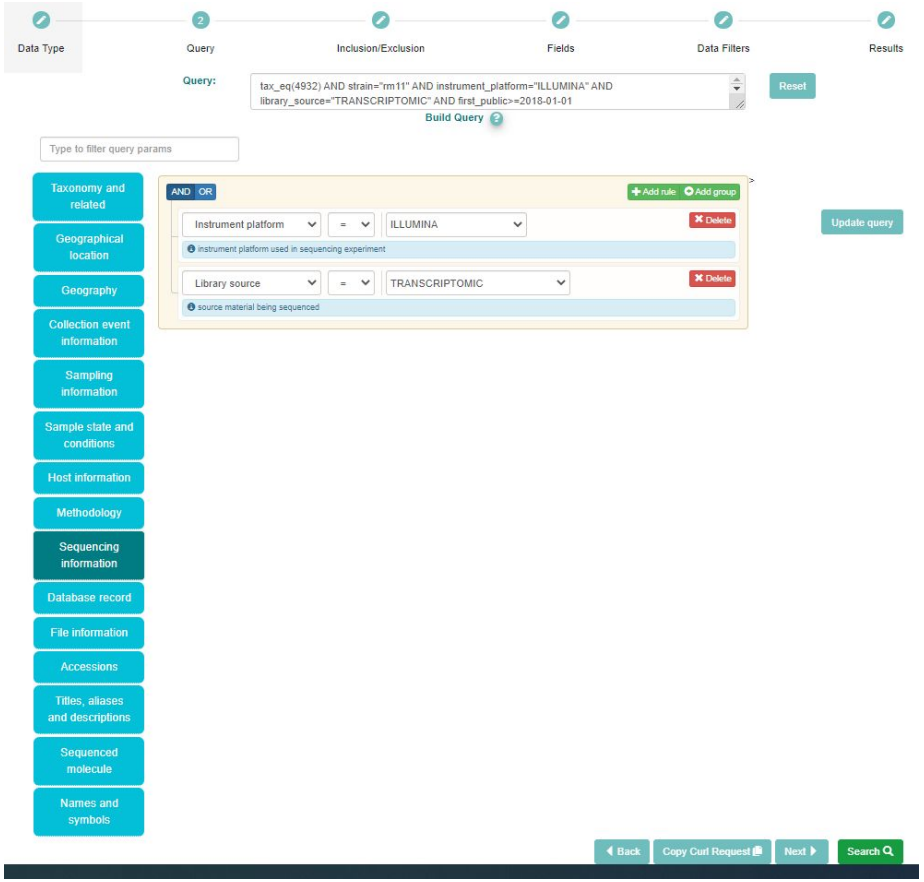
Download All

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	SRR727948...fastq.gz SRR727948...fastq.gz
PRJNA454805	SAMN09229632	SRX4183309	SRR7279482	195583	Solanum lycopersicum var. cerasiforme	Esther052516_BGV006904	SRR727948...fastq.gz SRR727948...fastq.gz
PRJNA454805	SAMN09229595	SRX4183308	SRR7279483	4084	Solanum pimpinellifolium	Esther120315_BGV006148	SRR727948...fastq.gz SRR727948...fastq.gz
PRJNA454805	SAMN09229594	SRX4183307	SRR7279484	195583	Solanum lycopersicum var. cerasiforme	Esther120315_BGV005912	SRR727948...fastq.gz SRR727948...fastq.gz
PRJNA454805	SAMN09229602	SRX4183306	SRR7279485	195583	Solanum lycopersicum var. cerasiforme	Esther120315_BGV006232	SRR727948...fastq.gz SRR727948...fastq.gz

ENA advanced search

Filter results by:

- Organism
- Experiment properties
- Sequencing properties
- Date



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

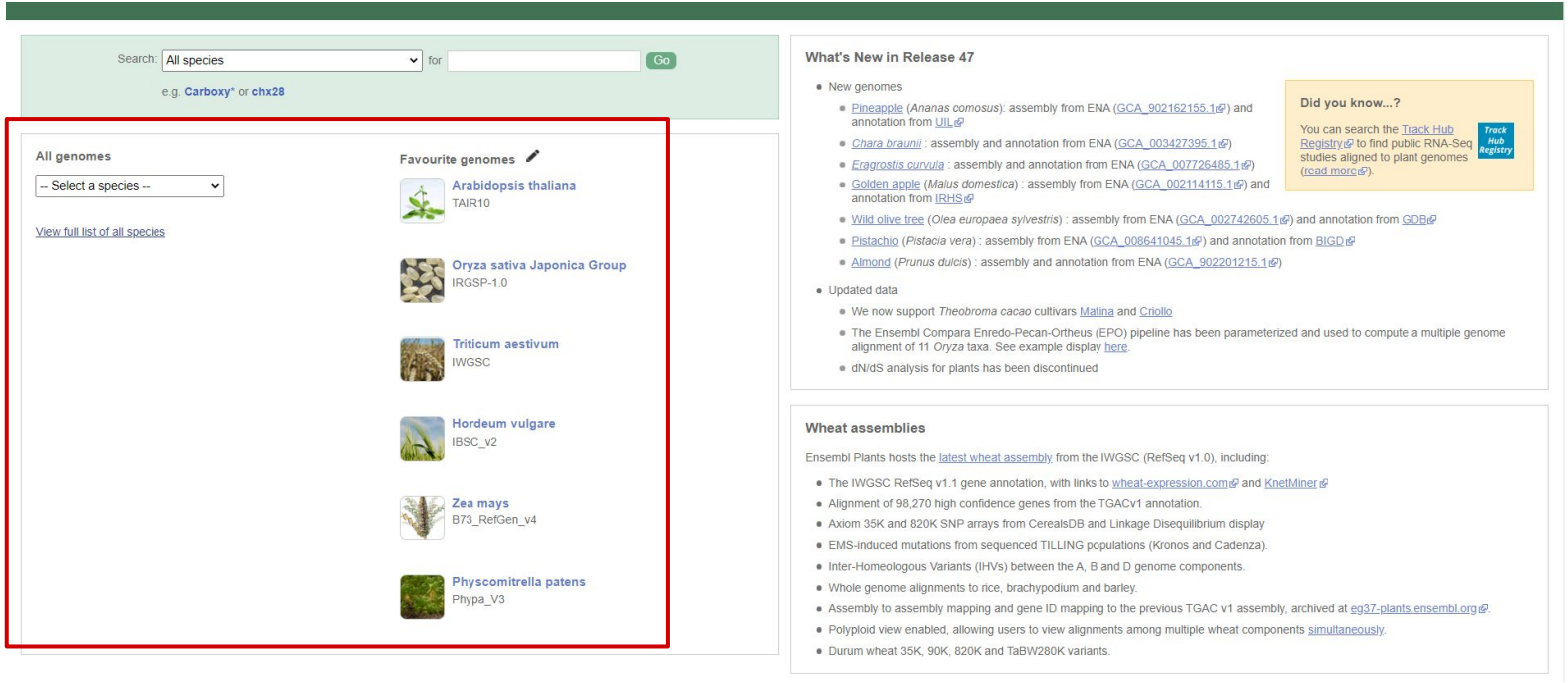
Vertebrates (including human)

http://ensemblgenomes.org

Anything else

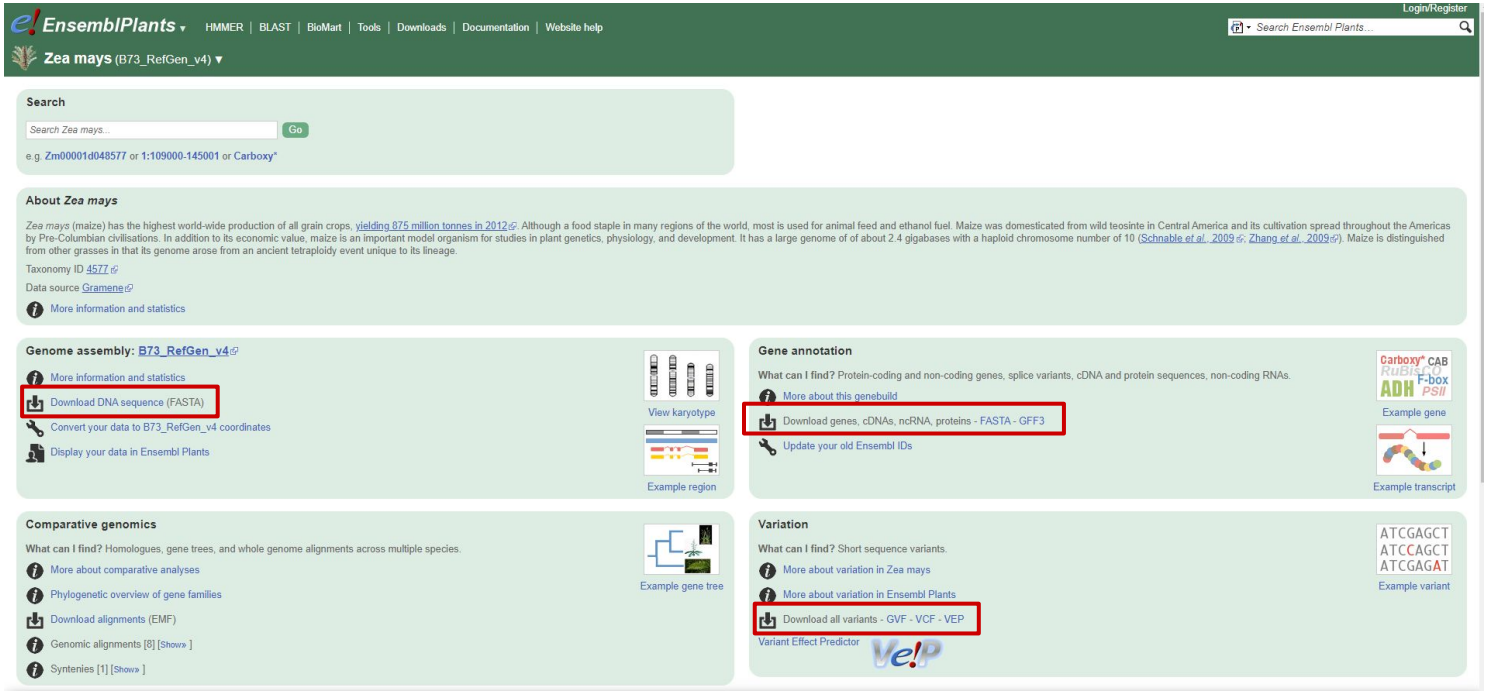
Downloading genomic data from Ensembl

1. Find your species of interest



Downloading genomic data from Ensembl

2. Choose the data you are interested in



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

FASTA

Full length transcripts

Transcripts without UTR

proteins

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

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