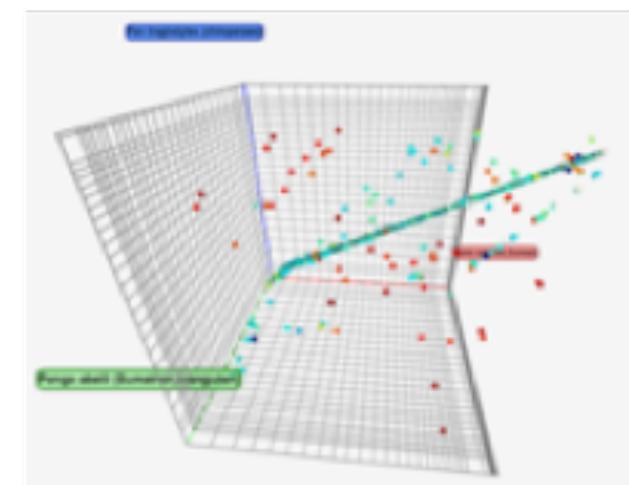


Computational Genomics

Introduction to Gene Models and Gene Tables

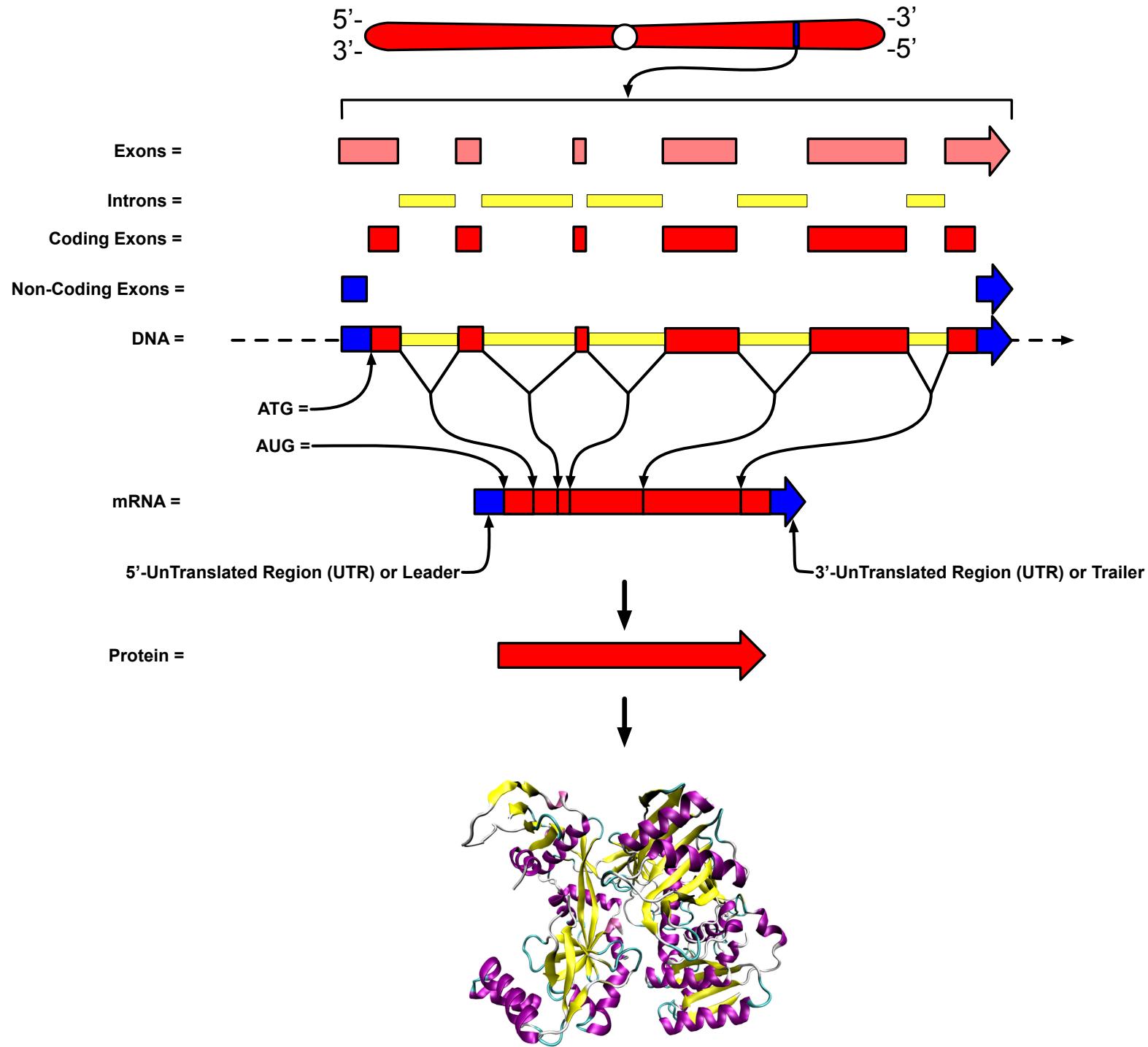


Fasta Format

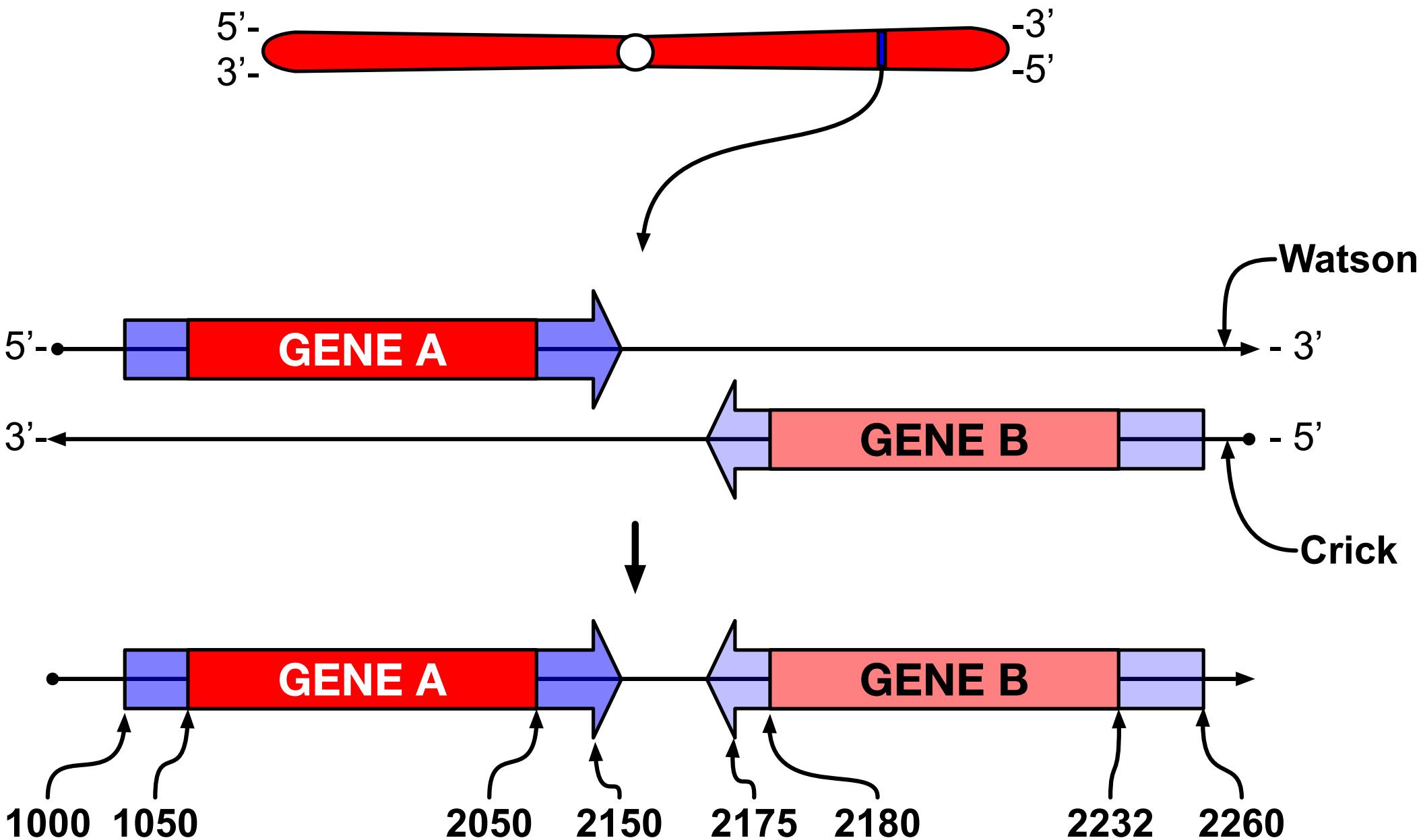
A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPSEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Gene Models 101

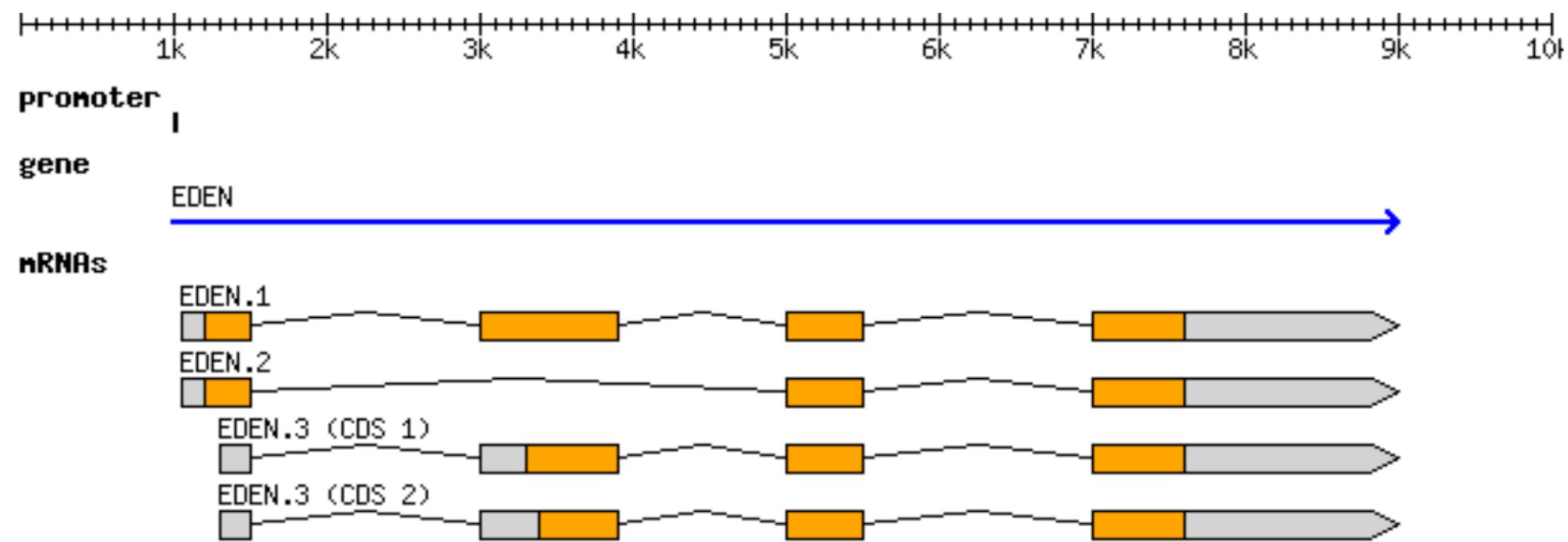


Gene Tables 101



GFF File Format and Gene Models

The Canonical Gene

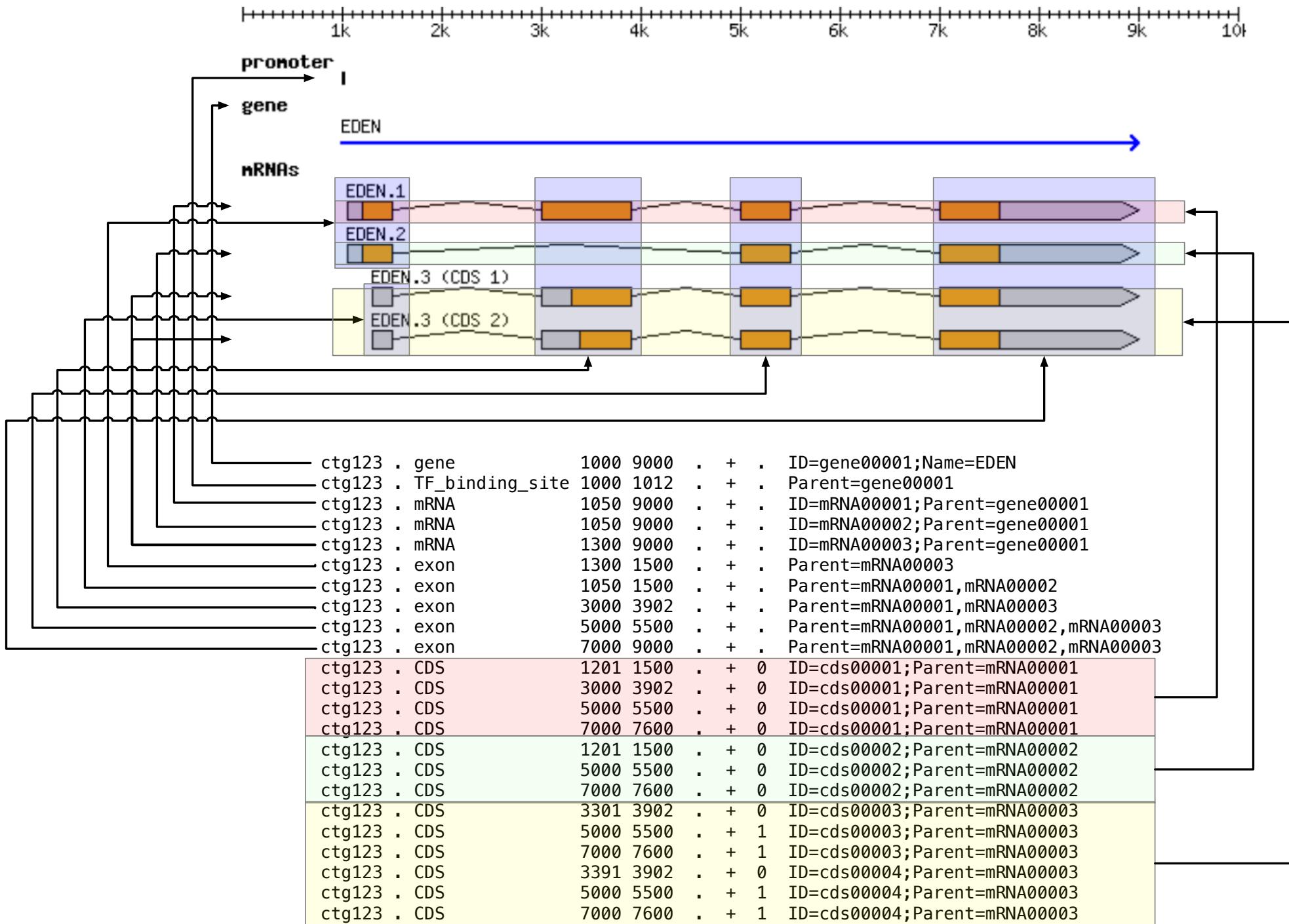


GFF File Format and Gene Models

```
0 ##gff-version 3.1.26
1 ##sequence-region ctg123 1 1497228
2 ctg123 . gene      1000 9000 . + . ID=gene00001;Name=EDEN
3 ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
4 ctg123 . mRNA     1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5 ctg123 . mRNA     1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6 ctg123 . mRNA     1300 9000 . + . ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7 ctg123 . exon    1300 1500 . + . ID=exon00001;Parent=mRNA00003
8 ctg123 . exon    1050 1500 . + . ID=exon00002;Parent=mRNA00001,mRNA00002
9 ctg123 . exon    3000 3902 . + . ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon   5000 5500 . + . ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon   7000 9000 . + . ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS    1201 1500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS    3000 3902 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS    5000 5500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS    7000 7600 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS    1201 1500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS    5000 5500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS    7000 7600 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS    3301 3902 . + 0 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS    5000 5500 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS    7000 7600 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS    3391 3902 . + 0 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS    5000 5500 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS    7000 7600 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

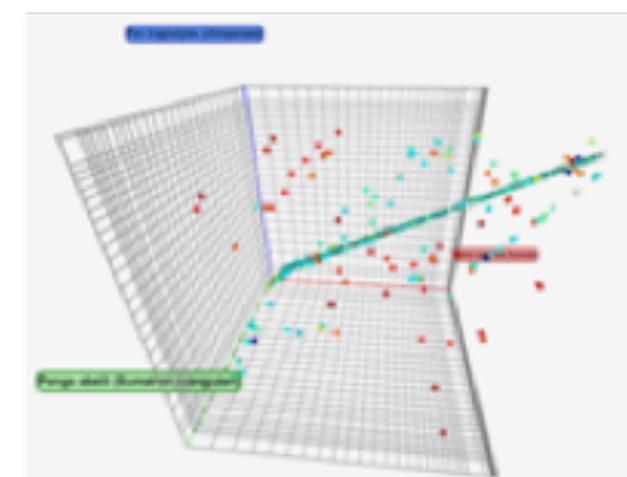
GFF File Format and Gene Models

The Canonical Gene



Computational Genomics

Introduction to Genome Browsers ENSEMBL



Training materials



- Ensembl training materials are protected by a CC BY license:
creativecommons.org/licenses/by/4.0/
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:
ensembl.org/info/about/publications.html

Exploring the Ensembl genome browser

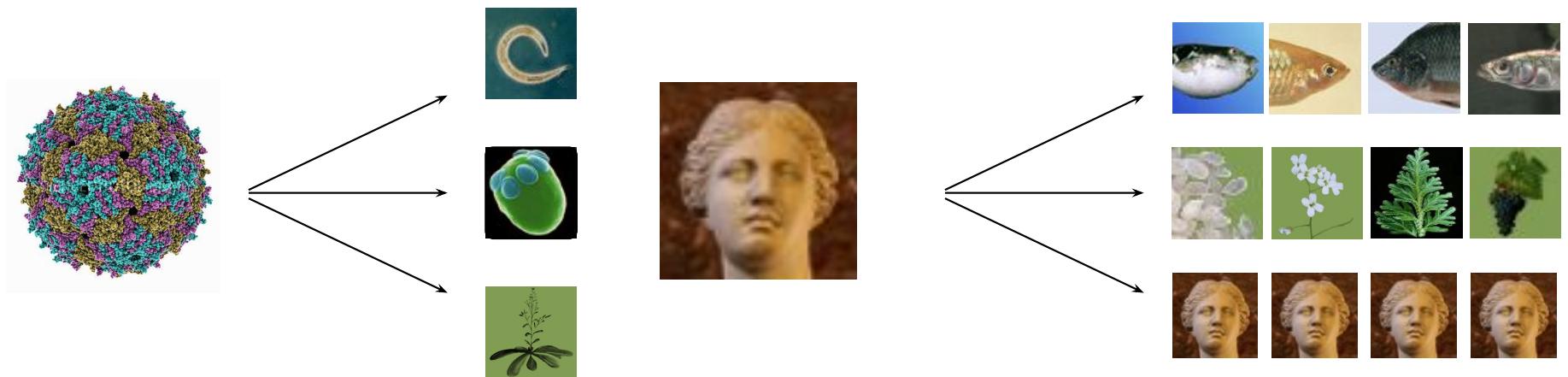
The screenshot shows the Ensembl homepage. At the top, there's a search bar with the placeholder "Search: All species" and a "Go" button. Below the search bar, there's a "Browse a Genome" section with icons for Human (GRCh37), Mouse (GRCm38), Zebrafish (Danio), and a "Popular genomes" dropdown. There's also a "All genomes" link and a note about other species available. To the right of this is a "What's New in Release 73 (September 2013)" section with links to updated patches, a new variation citation page, and a VEP output upload feature. Below this is a "Latest blog posts" section with links to specific blog entries. The main content area features several cards: "ENCODE data in Ensembl" (with ENCODE and VEP logos), "Find SNPs and other variants for my gene" (with a DNA sequence card), "Retrieve gene sequence" (with a sequence card), "Compare genes across species" (with a brain and eye icon), "Use my own data in Ensembl" (with a brain and eye icon), and "Learn about a disease or phenotype" (with an eye icon). A yellow sidebar on the right contains a "Did you know...?" section with a "FAQs" button.

Introduction

Why do we need genome browsers?

1977: 1st genome to be sequenced (5 kb)

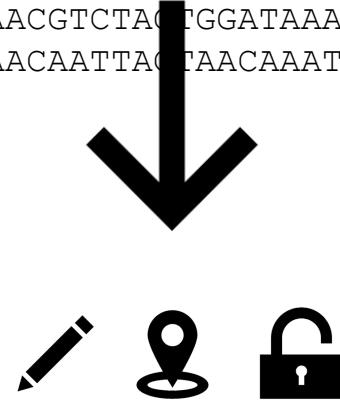
2004: finished human sequence (3 Gb)



CGGCCTTGGGCTCCGCCTCAGCTCAAGACTTAACCTCCCTCCCAGCTGTCCCAGATGACGCCATCTGAAATTCTGGAA
ACACGATCACTTTAACGGAATTGCTGTTGGGAAGTGTGTTACAGCTGCTGGCACGCTGTATTGCCTTAAGC
CCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGATTACCAAGCAGCGTGGCTCTAAGTGGAGCCCTGTCCCC
ACTAGCCACGCGTCAGCTGGTAGCGTATTGAAACTAAATCGTATGAAAATCCTCTCTAGTCGCACTAGCCACGTTCG
AGTGCTTAATGTGGCTAGTGGCACCGGTTGGACAGCACAGCTGTAAAATGTTCCCACAGTAAGCTGTTACCGTTC
CAGGAGATGGGACTGAATTAGAACAAATTCCAGCGCTCTGAGTTACCTCAGTCACATAATAAGGAATGCAT
CCCTGTGTAAGTGCATTTGGTCTCTGTTGAGACTTACCAAGCATTGGAGGAATATCGTAGGTAAAATGCCTA
TTGGATCCAAAGAGAGGCCAACATTGGAAATTAAAGACACGCTGCAACAAAGCAGGTATTGACAAATTATATAAC
TTTATAAATTACACCGAGAAAGTGTGTTCTAAAAAATGCTGCTAAAACCAGTACGTACAGTGTGCTTAGAACCAA
ACTGTTCTTATGTGTATAAATCCAGTTAACACATAATCATCGTTGCAGGTTAACCATGATAAATATAGAACGTCT
AGTGGATAAAGAGGAAACTGGCCCCTGACTAGCAGTAGGAACAATTACTAACAAATCAGAACATTAGTTACTTATGG
CAGAAGTTGTCACCTTTGGTTCAGTACTCCTATACTCTTAAAATGATCTAGGACCCCCGGAGTGCTTGTATG
TAGCTTACCATATTAGAAATTAAACTAAGAATTAAAGGCTGGCGTGGCTCACGCCTGTAATCCCAGCAGCTGGGA
GGCGAGGTGGCGGATCACTGAGGCCAGAAGTTGAGACCAGCCTGGCCAACATGGTAAACCCATTCTACTAAAAAT
ACAAAAAAATGTGCTGCGTGGTGCCTGTAATCCCAGCTACACGGGAGGTGGAGGCAGGAGAACGCTTGAACCC
TGGAGGCAGAGGTGCAGTGAGCCAAGATCATGCCACTGCACTCTAGCCTGGCCACATAGCATGACTCTGCTCAAAACAA
ACAAACAAACAAAAACTAAGAATTAAAGTTAAACTTAAAGCTAACAGTGGAGTGGAAATAGTTTACATTGCACTGGCT
TCTTAGGAAAATAACTTTGAAAACAAGTGGAGTGGAAATAGTTTACATTGCACTGGCTCTTTAATGTCTGGCTAAAT
AGAGATAGCTGGATTCACTTATCTGTGCTAATCTGTTATTGGTAGAAGTATGTGAAAAAAATTAAACCTCACGTTGAAA
AAAGGAATATTAAATAGTTCACTTGGTATTTCAGTTACTTTGGTATTTCCTGTACTTGCATAGATTTCAAAGATCTAATAGAT
ATACCATAGGTCTTCCATGTCGAACATCATGCAGTGATTATTGGAAGATAGTGGTGTCTGAATTATAACAAAGTTCC
AAATATTGATAAATTGCATTAAACTATTAAAATCTCATTCAATTAAACCAACCATGGATGTCAGAAAAGTCTTTAAGAT
TGGTAGAAATGAGCCACTGGAAATTCTAATTTCATTGAAAGTCACATTGTCATTGACAACAAACTGTTCTTGC
AGCAACAAGATCACTCATTGATTGTGAGAAAATGTCTACCAAATTATTAAGTTGAAATAACTTGTCAAGCTGTTCTTC
AAGTAAAATGACTTTCATTGAAAAAATTGCTTGTCAAGATCACAGCTAACATGAGTGCTTTCTAGGCAGTATTGTACT
TCAGTATGCAGAAGTGCTTATGTATGCTTCTATTGTCAGAGATTATTAAAAGAAGTGCTAAAGCATTGAGCTTCGAAA
TTAATTTCAGTGCCTCATTAGGACATTCTACATTAAACTGGCATTATTACTATTATTAAACAAGGACACTCAGTG
GTAAGGAATATAATGGCTACTAGTATTAGTTGGTGCCTAGCCATAACTCATGCAAATGTGCCAGCAGTTACCCAGCAT
CATCTTGCACTGTTGATACAAATGTCAACATCATGAAAAAGGGTTGAAAAAAGGAATATTAAATAGTTCACTTT

What is Ensembl?

```
AGTGCTTAATGTGGCTAGTGGCACCGGTTGGACAGCACAGCTGTAA  
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA  
CTGAATTAGAACAAACAAATTTCAGCGCTCTGAGTTTACCT  
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTGGTCT  
TCTGTTTGCAGACTTACCAAGCATTGGAGGAATATCGTAGGT  
AAAAATGCCTATTGGATCCAAGAGAGGCCAACATTTTGAAATT  
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTATATAACT  
TTATAAAATTACACCGAGAAAGTGTCTAAAAAAATGCTTGCTAAAA  
ACCCAGTACGTACAGTGTGCTTAGAACCATAAACTGTTCTTATG  
TGTGTATAAATCCAGTTAACACATAATCATCGTTGCAGGTTAAC  
ACATGATAAAATATAGAACGTCTACTGGATAAAGAGGAAACTGGCCCC  
TTGACTAGCAGTAGGAACAATTACTAACAAATCAGAACGATTAATGT
```



Ensembl annotates and maps genomic features from genome sequences

What is Ensembl?

```
AGTGCTTAATGTGGCTAGTGGCACCGGTTGGACAGCACAGCTGTAA  
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA  
CTGAATTAGAACAAATTTCAGCGCTCTGAGTTTACCT  
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTGGTCT  
TCTGTTTGCAGACTTACCAAGCATTGGAGGAATATCGTAGGT  
AAAAATGCCTATTGGATCCAAGAGAGGCCAACATTTTGAAATT  
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTATATAACT  
TTATAAATTACACCGAGAAAGTGTCTAAAAAATGCTTGCTAAAA  
ACCCAGTACGTACAGTGTGCTTAGAACCATAAACTGTTCTTATG  
TGTGTATAAATCCAGTTAACACATAATCATCGTTGCAGGTTAAC  
ACATGATAAAATATAGAACGTCTACTGGATAAAGAGGAAACTGGCCCC  
TTGACTAGCAGTAGGAACAATTACTAACAAATCAGAACGATTAATGT
```



Ensembl is an ‘added value resource’ bringing together information from a wide range of other databases in a single site

What is Ensembl?

```
AGTGCTTAATGTGGCTAGTGGCACCGGTTGGACAGCACAGCTGTAA  
AATGTTCCCACCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA  
CTGAATTAGAACAAATTTCAGCGCTCTGAGTTTACCT  
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTGGTCT  
TCTGTTTGCAGACTTACCAAGCATTGGAGGAATATCGTAGGT  
AAAAATGCCTATTGGATCCAAGAGAGGCCAACATTTTGAAATT  
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTATATAACT  
TTATAAATTACACCGAGAAAGTGTCTAAAAAATGCTTGCTAAAA  
ACCCAGTACGTACAGTGTGCTTAGAACCATAAACTGTTCTTATG  
TGTGTATAAATCCAGTTAACACATAATCATCGTTGCAGGTTAACCC  
ACATGATAAAATATAGAACGTCTACTGGATAAAGAGGAAACTGGCCCC  
TTGACTAGCAGTAGGAACAATTACAAACAAATCAGAACGATTAATGT
```



U.S. National Library of Medicine

NCBI

Genome Data Viewer

[www.ncbi.nlm.nih.gov/
genome/gdv/](http://www.ncbi.nlm.nih.gov/genome/gdv/)



www.ensembl.org



www.genome.ucsc.edu

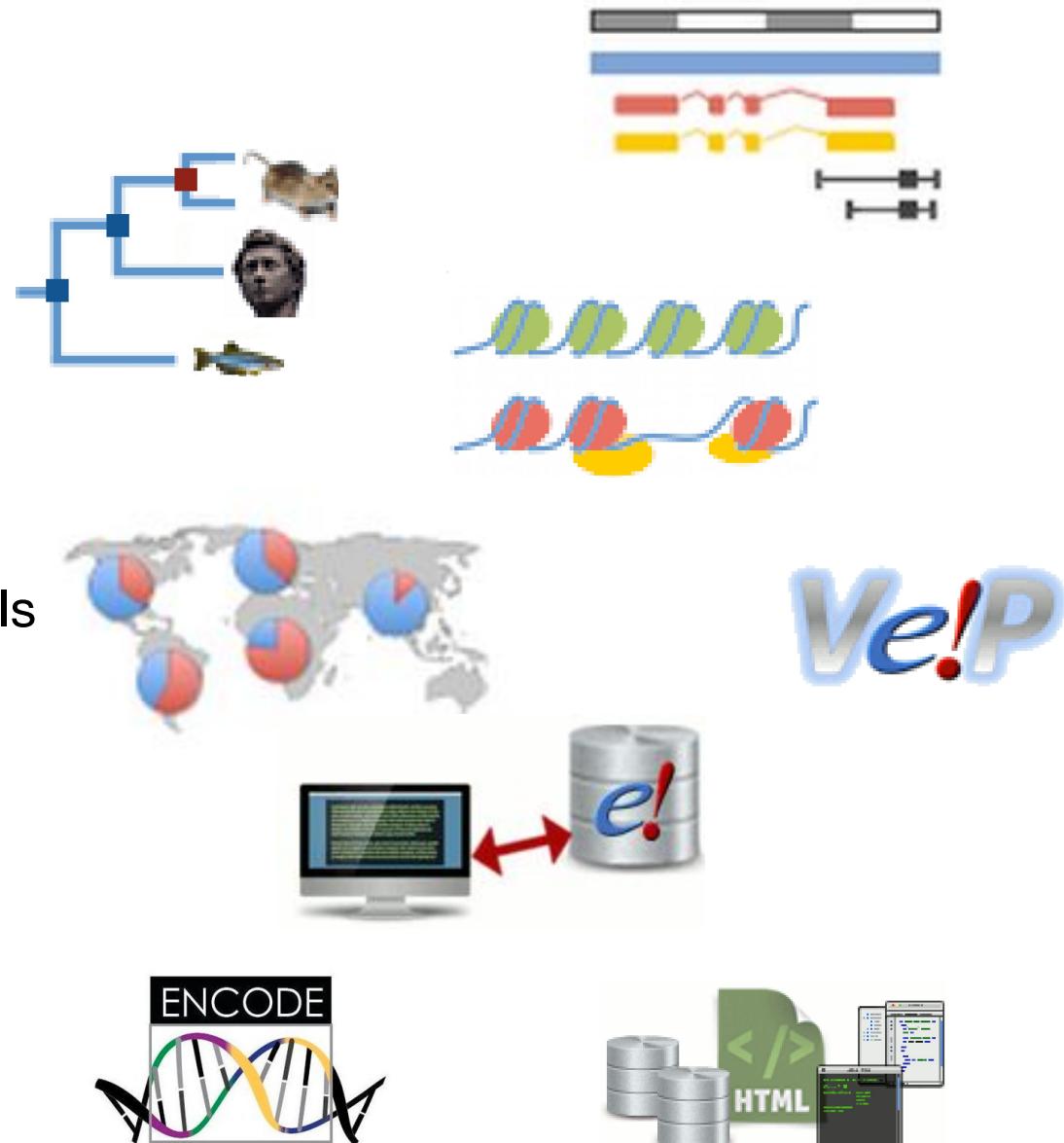


www.ensemblgenomes.org

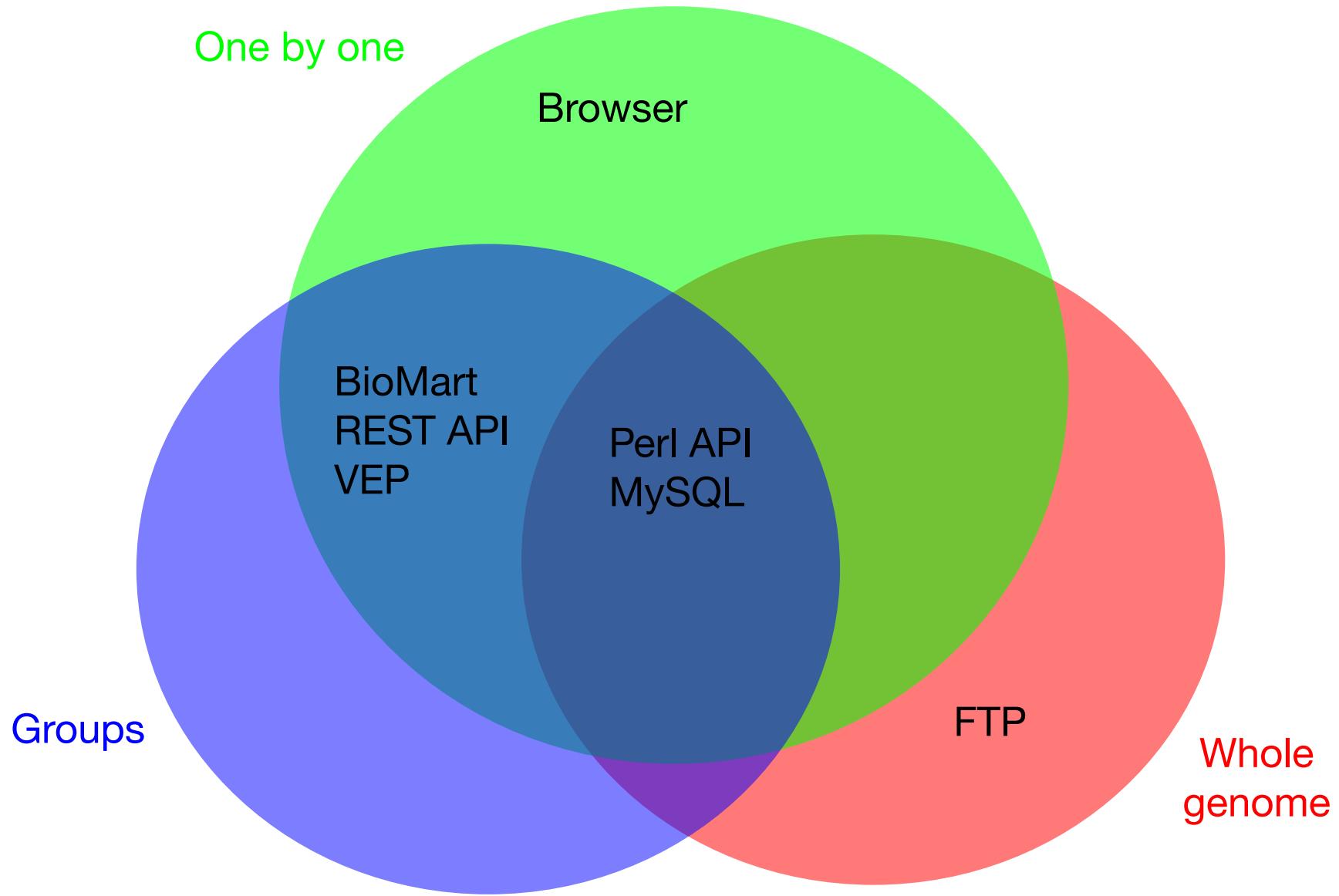


Ensembl features

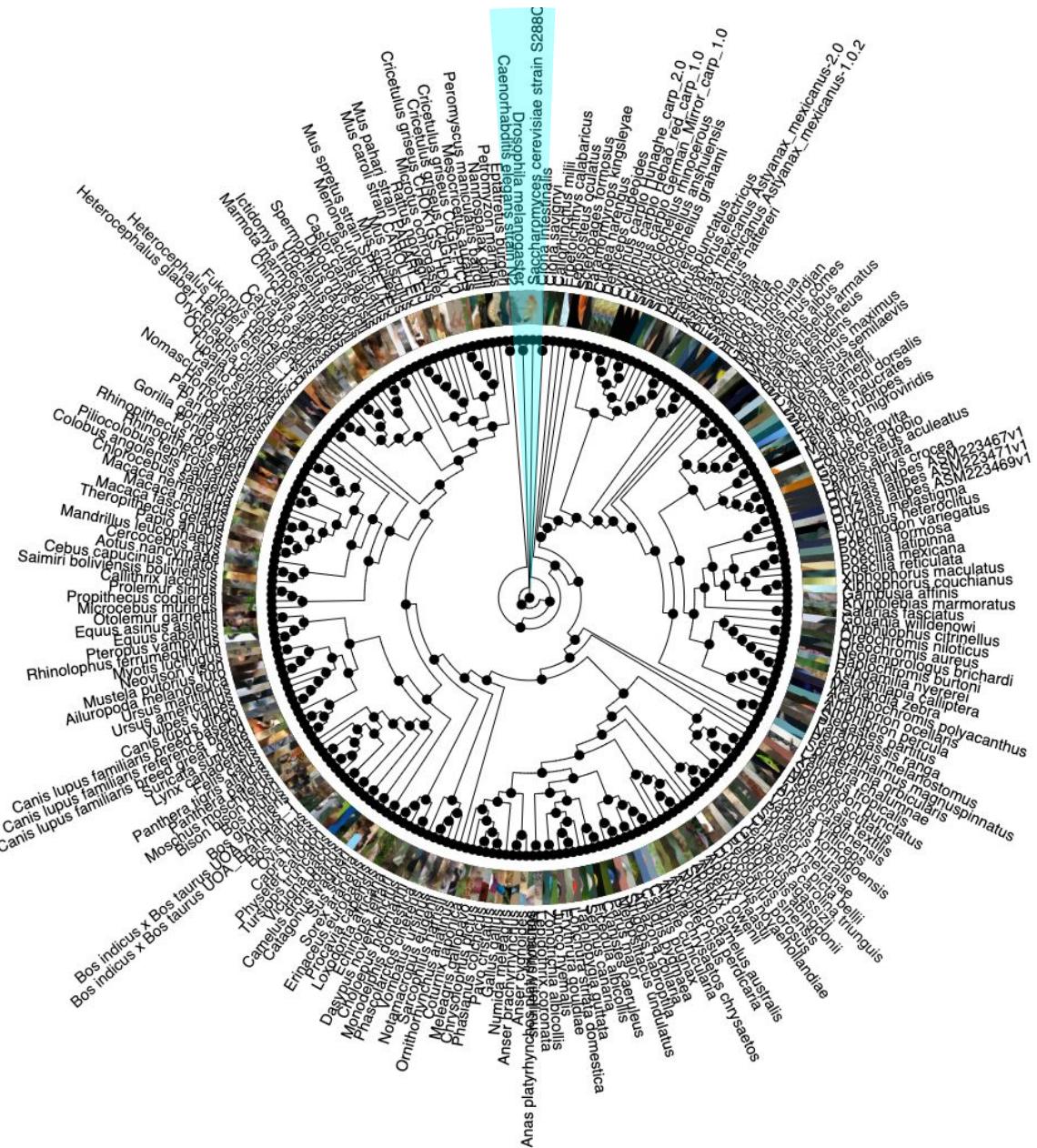
- Gene builds for >300 species
- Gene trees
- Regulatory build (ENCODE)
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source



Access scales



Vertebrate species in Ensembl



Non-vertebrates on Ensembl genomes

EnsemblBacteria | BLAST | More | Search Ensembl Bacteria species...

Help & Documentation Species List

Find a Species

Ensembl Bacteria Species

Bacillus collection 78 genomes

- Bacillus amyloliquefaciens** European Nucleotide Archive
- Bacillus anthracis A0248** European Nucleotide Archive
- Bacillus anthracis Ames** European Nucleotide Archive
- Bacillus anthracis CDC 684** European Nucleotide Archive
- Bacillus anthracis Sterne** European Nucleotide Archive
- Bacillus cereus 03BB102** European Nucleotide Archive
- Bacillus cereus 172560W** European Nucleotide Archive
- Bacillus cereus 05/201** European Nucleotide Archive
- Bacillus cereus AH1271** European Nucleotide Archive
- Bacillus cereus AH1272** European Nucleotide Archive
- Bacillus cereus AH187** European Nucleotide Archive
- Bacillus cereus AH603** European Nucleotide Archive
- Bacillus cereus AH820** European Nucleotide Archive
- Bacillus cereus ATCC 10876** European Nucleotide Archive
- Bacillus cereus ATCC 14579** European Nucleotide Archive
- Bacillus cereus B4264** European Nucleotide Archive
- Bacillus cereus BD RD-ST24** European Nucleotide Archive
- Bacillus cereus F65185**
- Bacillus cereus G9842**
- Bacillus cereus MM3**

Bacteria

EnsemblProtists | BLAST | More | Search Ensembl Protists species...

Help & Documentation Species List

Find a Species

Ensembl Protists Species

Alveolata

- Plasmodium berghei** GenoDB | Plasmodium berghei ANKA
- Plasmodium knowlesi** Wellcome Trust Sanger Institute | Plasmodium knowlesi
- Toxoplasma gondii** ToxoDB | Toxoplasma gondii
- Plasmodium vivax** The Institute for Genomic Research | Plasmodium vivax
- Tetrahymena thermophila** The Institute for Genomic Research | Tetrahymena thermophila SB210

Amoebozoa

- Dictyostelium discoideum** DictyBase | Dictyostelium discoideum
- Entamoeba histolytica** AmoebaDB | Entamoeba histolytica HM-1:IMSS

Stramenopiles

- Albugo laibachii** The Sainsbury Laboratory | Albugo laibachii Nc14
- Phytophthora infestans** BROAD | Phytophthora infestans
- Pythium ultimum** Pythium Genome Database | Pythium ultimum

Protists

EnsemblFungi | BLAST | More | Search Ensembl Fungi species...

Help & Documentation Species List

Find a Species

Ensembl Fungi Species

Copadiales

- Mycosphaerella graminicola** JGI | Mycosphaerella graminicola IP023

Eurotiales

- Aspergillus clavatus** CADRE | Aspergillus clavatus
- Aspergillus fumigatus1163** CADRE | Aspergillus fumigatus1163 A1163
- Aspergillus flavus** CADRE | Aspergillus flavus
- Aspergillus nidulans** CADRE | Aspergillus nidulans FGSC A4
- Aspergillus terreus** CADRE | Aspergillus terreus
- Aspergillus niger** CBS 513.89
- Fusarium oxysporum** Broad Institute | Fusarium oxysporum 4287
- Gibberella zeae** Broad Institute | Gibberella zeae PH-1
- Trichoderma virens** JGI | Trichoderma virens C 8

Fungi

EnsemblMetazoa | BLAST | More | Search Ensembl Metazoa species...

Help & Documentation Species List

Find a Species

Ensembl Metazoa Species

Diptera

- Aedes aegypti** VectorBase | Aedes aegypti
- Anopheles darlingi** European Nucleotide Archive | Anopheles darlingi
- Anopheles gambiae** VectorBase | Anopheles gambiae
- Culex quinquefasciatus** VectorBase | Culex quinquefasciatus
- Drosophila ananassae** FlyBase | Drosophila ananassae
- Drosophila erecta** FlyBase | Drosophila erecta
- Drosophila grimshawi** FlyBase | Drosophila grimshawi
- Drosophila melanogaster** FlyBase | Drosophila melanogaster
- Drosophila mojavensis** FlyBase | Drosophila mojavensis
- Drosophila persimilis** FlyBase | Drosophila persimilis
- Drosophila pseudodorsalis** FlyBase | Drosophila pseudodorsalis
- Drosophila sechellia** FlyBase | Drosophila sechellia
- Drosophila simulans** FlyBase | Drosophila simulans
- Drosophila virilis** FlyBase | Drosophila virilis
- Drosophila willistoni** FlyBase | Drosophila willistoni
- Drosophila yakuba** FlyBase | Drosophila yakuba

Metazoa

EnsemblPlants | BLAST | More | Search Ensembl Plants species...

Help & Documentation Species List

Find a Species

Ensembl Plants Species

Liliopsida

- Brachypodium distachyon** Brachypodium.org | Brachypodium distachyon (L.) Beauvois
- Oryza glaberrima** JGI | Oryza glaberrima
- Sorghum bicolor** JGI | Sorghum bicolor BT
- Hordeum vulgare** MSU | Hordeum vulgare
- Oryza sativa** MSU | Oryza sativa Nipponbare (Japonica rice)
- Zea mays** MaizeSequence | Zea mays
- Musa acuminata** CIB | Musa acuminata Double haploid Pahang (DH-Pahang)
- Oryza sativa Indica Group** HIS | Oryza indica 93-11 (Indica rice)
- Oryza brachyantha** QSE | Oryza brachyantha
- Setaria Italica** JGI | Setaria italica

eudicotyledons

- Arabidopsis lyrata** JGI | Arabidopsis lyrata
- Glycine max** PGSC | Glycine max
- Arabidopsis thaliana** TAIR | Arabidopsis thaliana
- Populus trichocarpa** JGI | Populus trichocarpa
- Solanum tuberosum** PGSC | Solanum tuberosum
- Vitis vinifera** Genoscope | Vitis vinifera

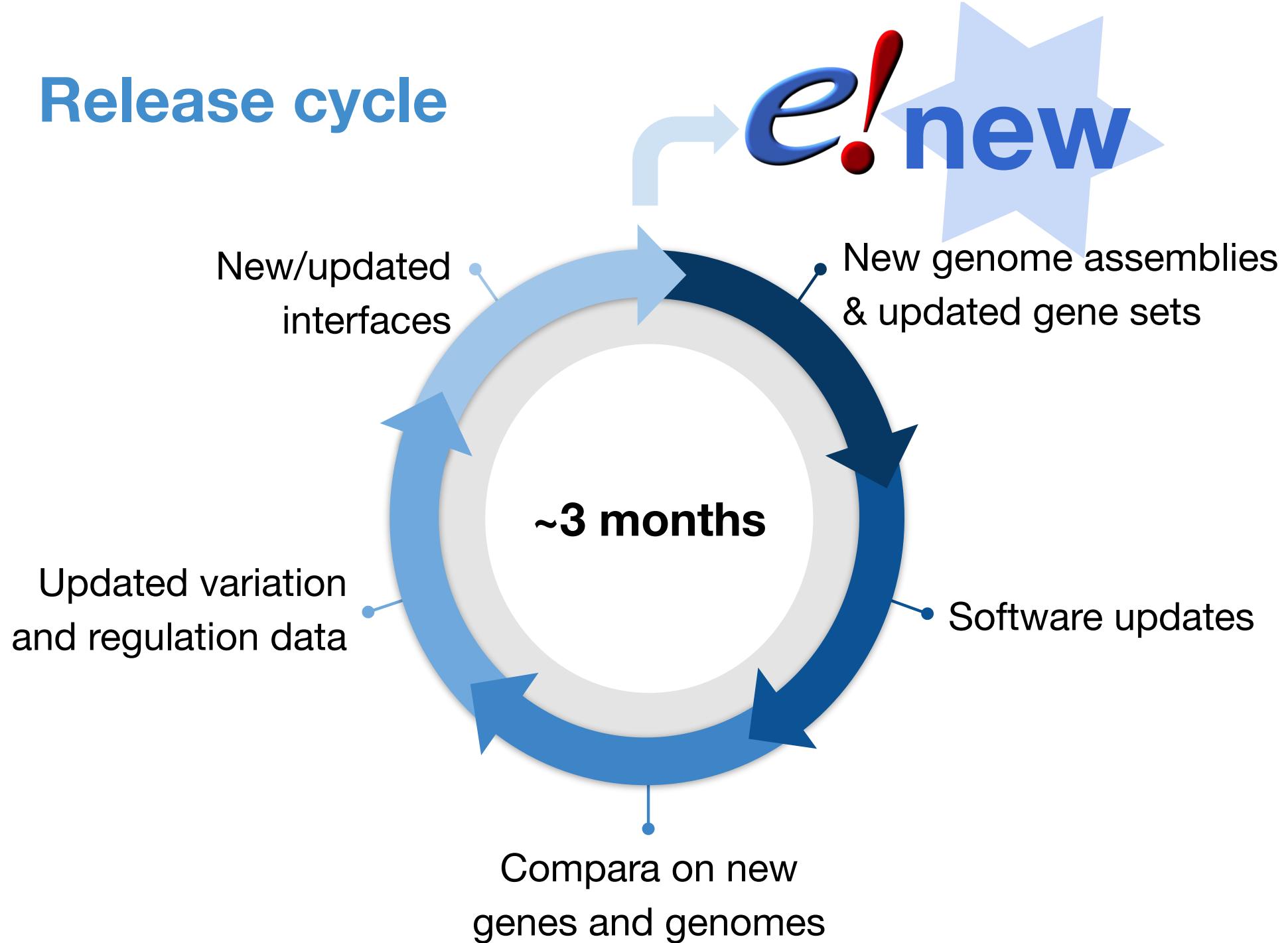
Plants

www.ensemblgenomes.org

Ensembl and Ensembl Genomes

	Ensembl	Ensembl Genomes
Released	2000	2009
Species	Vertebrates (fly, worm and yeast as outgroups)	Non-vertebrates (protists, plants, fungi, metazoa, bacteria)
Annotation	by Ensembl	in collaboration with the scientific communities
URL	www.ensembl.org	www.ensemblgenomes.org

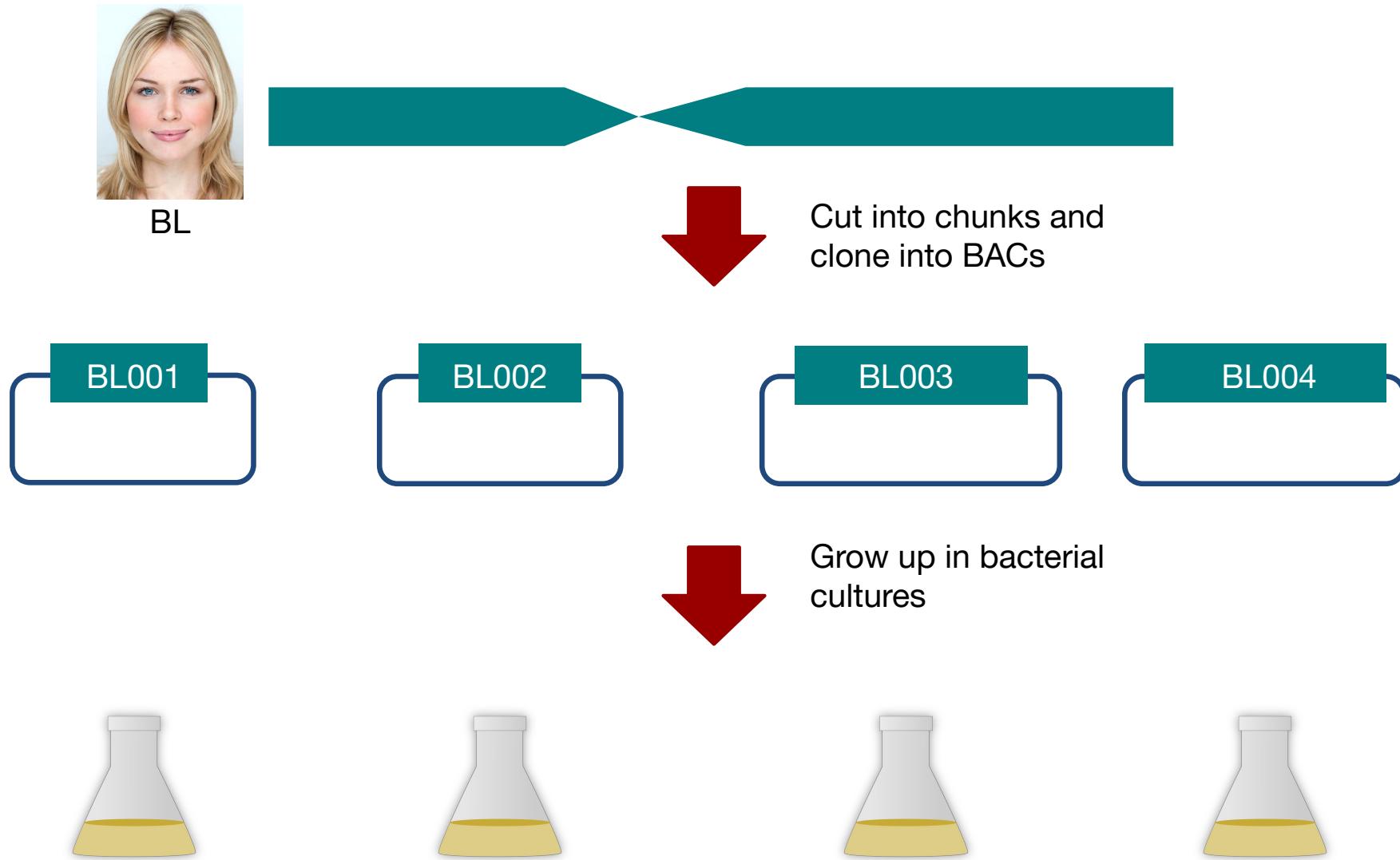
Release cycle



Ensembl Rapid Release

- Released every two weeks ✓
- Genome with gene annotation only ✓
- BLAST ✓
- No BioMart ✗
- No gene trees/homologues ✗
- No variation ✗

Cloning into BACs



Making a contig

Sequence reads

CGGCCTTGCGCTTCAGCTCAAGA		
CAGCTGTCCCAGATGAC	ACTTAACTCCCTCCCAGCTGTCC	
GGGCTCCGCCTTCAGCTC	AACTCCCTCCCAGCT	TCCCAGCTGTCCCAGATGACGCCATC
CGGCCTTGCGCTCC	CAGATGACGCC	TCCGCCTTCAGCTCAAGACTTAACCTC

Match up overlaps

CGGCCTTGCGCTTCAGCTCAAGA	AACTCCCTCCCAGCT	CAGATGACGCC
TCCGCCTTCAGCTCAAGACTTAACCTC	TCCCAGCTGTCCCAGATGACGCCATC	
GGGCTCCGCCTTCAGCTC	ACTTAACTCCCTCCCAGCTGTCC	
CGGCCTTGCGCTCC		CAGCTGTCCCAGATGAC

Contig

CGGCCTTGCGCTTCAGCTCAAGACTTAACCTCCCTCCCAGCTGTCCCAGATGACGCCATC

Contigs to scaffolds

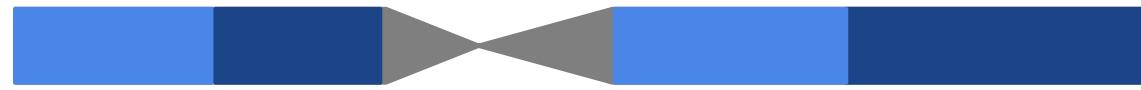
BACs from different individuals assembled together with overlaps
Tilepath



Overlaps trimmed to give **contigs**. A run of contigs with no gaps is a **scaffold**.



Genetic maps are used to assemble scaffolds into a chromosome

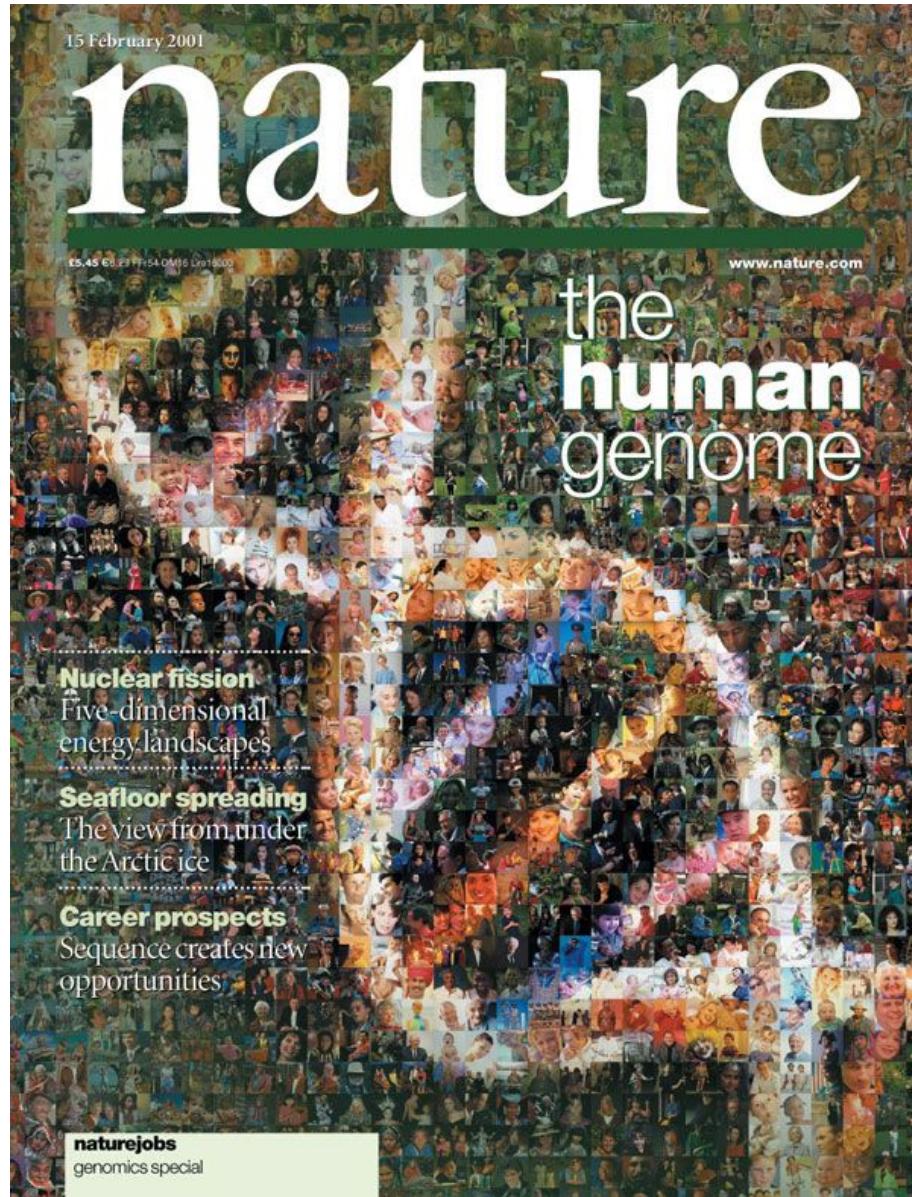


Tilepath

Contig/
scaffold

Chromosome

Genome assemblies



BL



AL

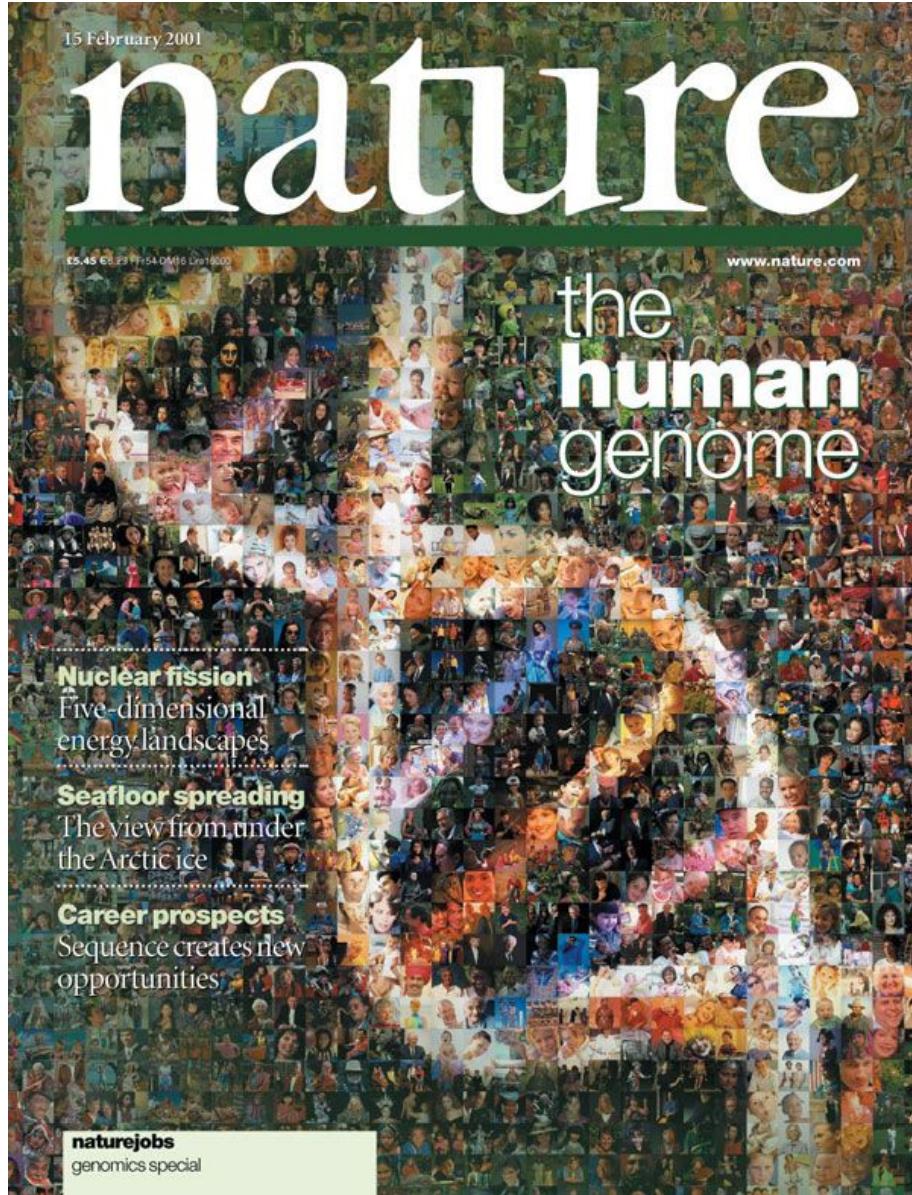


CM



IM

Genome contigs



BL



AL



CM



IM

BL001

AL002

CM003

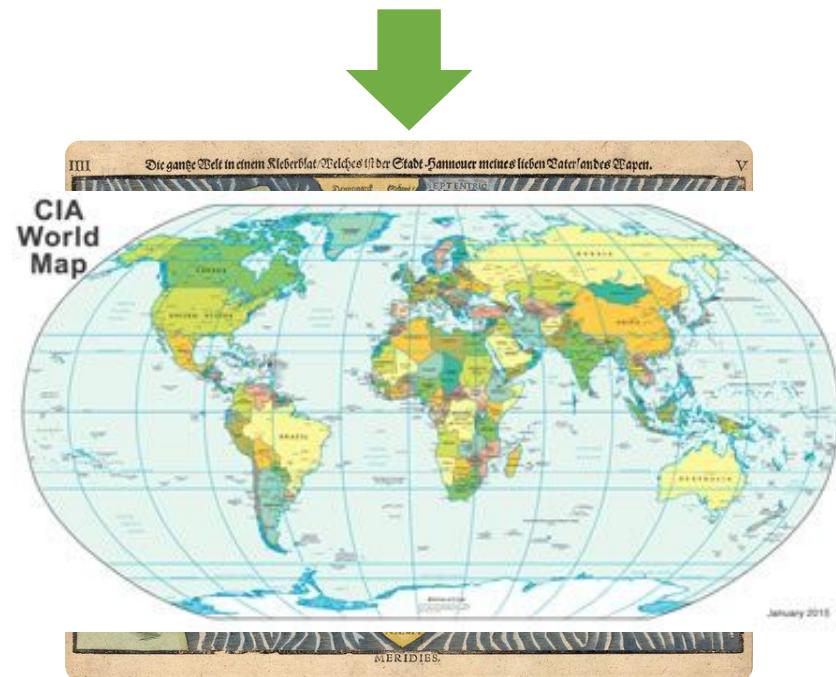
IM004

Genome assemblies

Genome
“DNA within a cell”



Genome assembly
Representation of a genome
Contains errors and gaps
Coordinate system

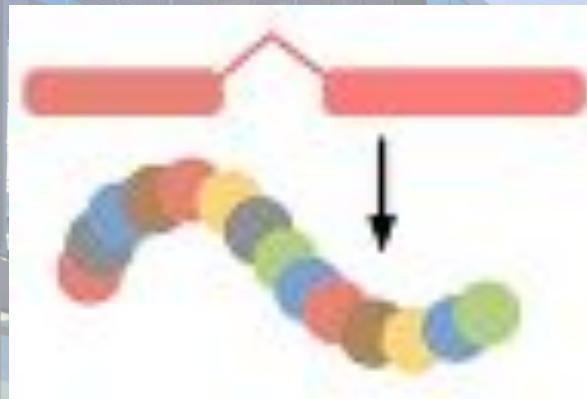


Human genome assemblies

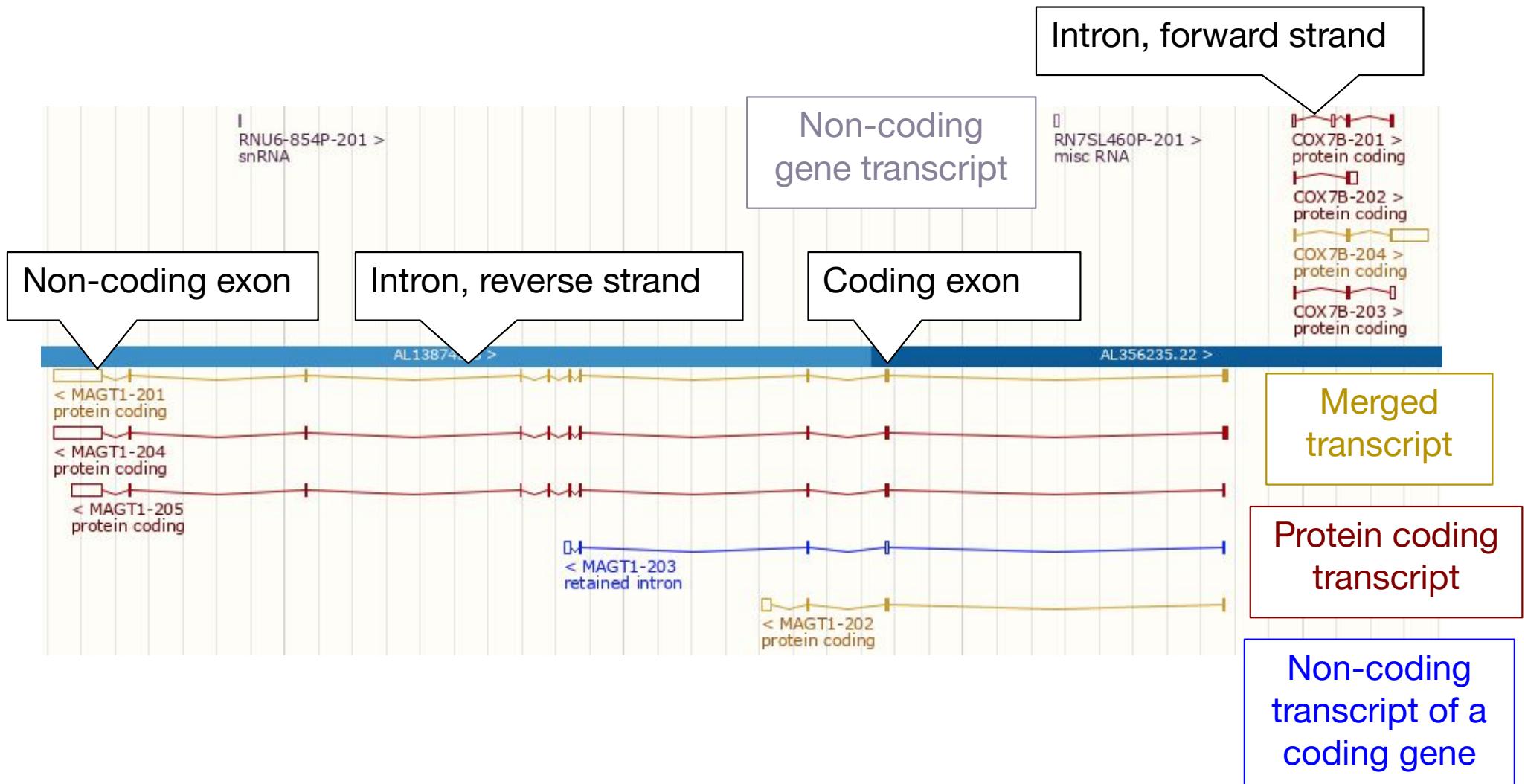
- GRCh38 (aka hg38)
• Many rare/private alleles replaced.
• www.ensembl.org
• Most up-to-date and supported
- GRCh37 (aka hg19)
• Some large gaps
• grch37.ensembl.org
• Limited data and software updates
• Still the preferred genome of the clinical community
- NCBI36 (aka hg18)
• Many gaps
• ncbi36.ensembl.org
• No longer updated



Genes and Transcripts



Gene views

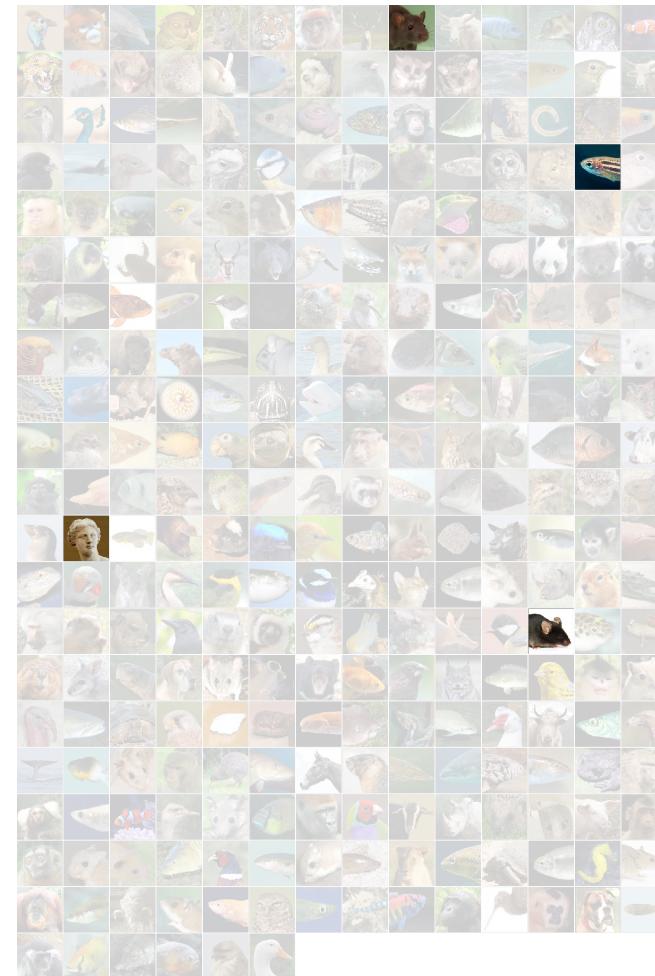


Automatic and manual annotation

Automatic annotation



Manual annotation

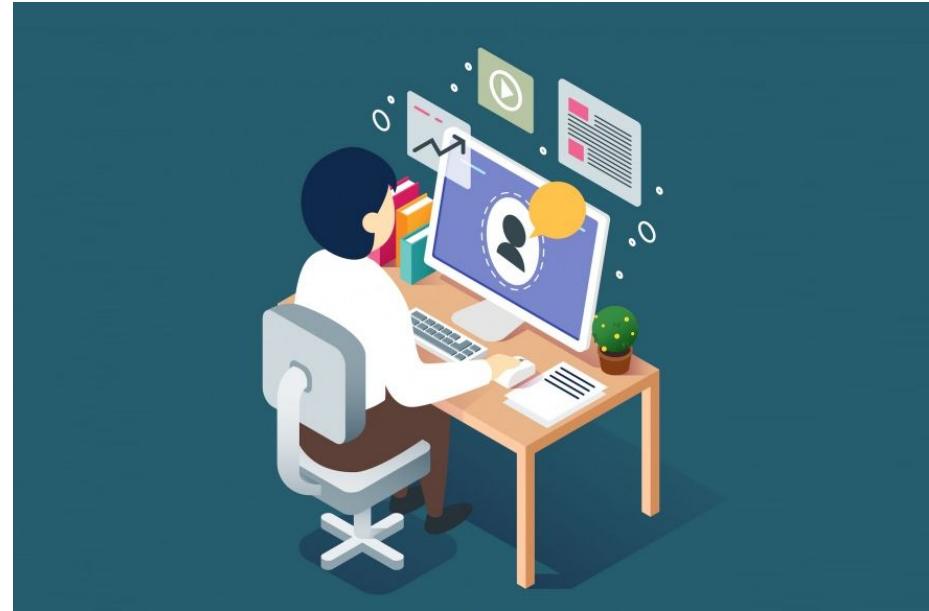


Automatic and manual annotation

Automatic annotation

- computational pipeline
 - one genome = two weeks

Manual annotation



- human annotator
 - one genome = several years

Annotation sources

Automatic annotation

- INSDC
 - cDNAs
 - ESTs
 - RNA-seq
- Protein sequence databases
 - Swiss-Prot: manually curated
 - TrEMBL: unreviewed translations

Manual annotation

- RNA-seq transcriptome data
 - Illumina short read
 - Oxford Nanopore long read
 - PacBio long read
- Transcript structure data
 - Introns
 - CAGE transcription start sites
 - PolyA-Seq transcription ends
- Mass Spec protein data
- publications

Manual vs automatic

- Manual annotation is more comprehensive
 - More transcripts per gene, especially non-coding transcripts
 - More genes overall, especially non-coding genes
- More biotypes
 - e.g. polymorphic pseudogene, NMD, non-stop decay, stop codon read-through
- Manual annotation can be more accurate for difficult to annotate features such as:
 - UTRs
 - Splice sites
 - Single exon transcripts
 - Exceptions, such as immunoglobulins, stop codon readthroughs



GENCODE == Ensembl human and mouse genes

GENCODE is the default gene set used by major projects such as:

- gnomAD/ ExAC: Exome Aggregation Consortium, Genome Aggregation
- GTEx: Genotype-Tissue Expression
- Decipher
- 100,000 Genomes Project, Genomics England
- ENCODE: Encyclopedia of DNA Elements
- TCGA: The Cancer Genome Atlas
- ICGC: International Cancer Genome Consortium
- Roadmap: NIH Roadmap Epigenomics Mapping Consortium
- Blueprint: Blueprint Epigenome
- 1000 Genomes Project
- HCA: Human Cell Atlas

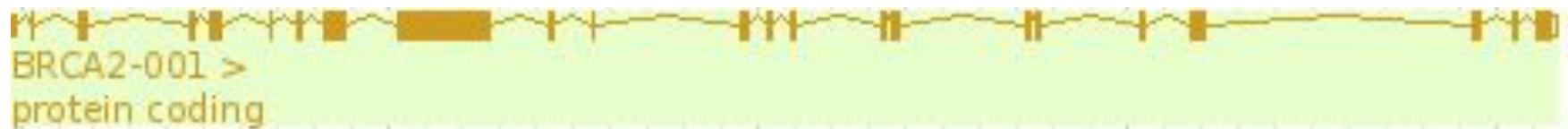


Golden transcripts

- Identical annotation



- Higher confidence and quality



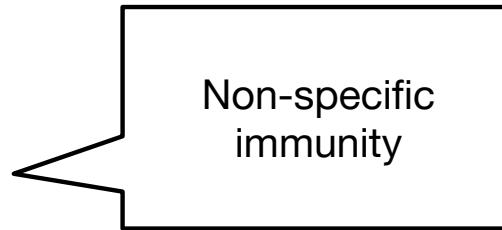
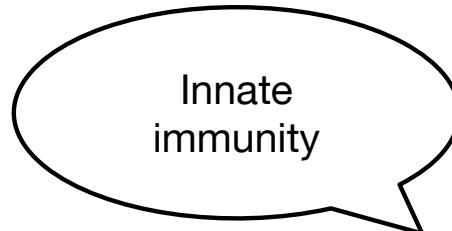
Ensembl stable IDs

- ENSG#####.# Ensembl Gene ID
- ENST#####.# Ensembl Transcript ID
- ENSP#####.# Ensembl Peptide ID
- ENSE#####.# Ensembl Exon ID
- For non-human species a suffix is added:
MUS (*Mus musculus*) for mouse ENSMUSG###
DAR (*Danio rerio*) for zebrafish: ENSDARG###

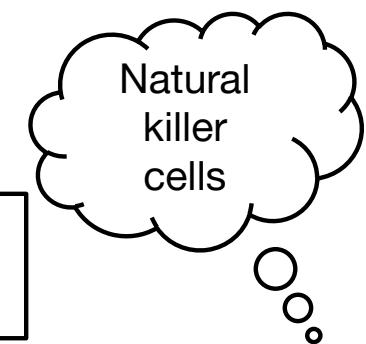
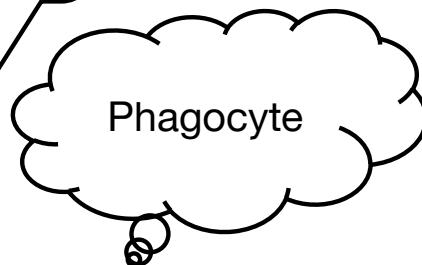
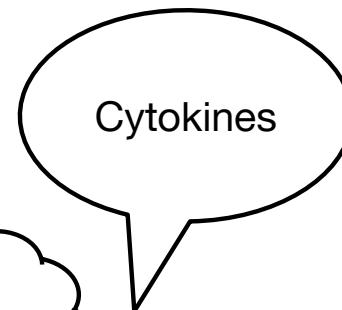
http://www.ensembl.org/info/genome/stable_ids/index.html

Why Gene Ontology (GO)?

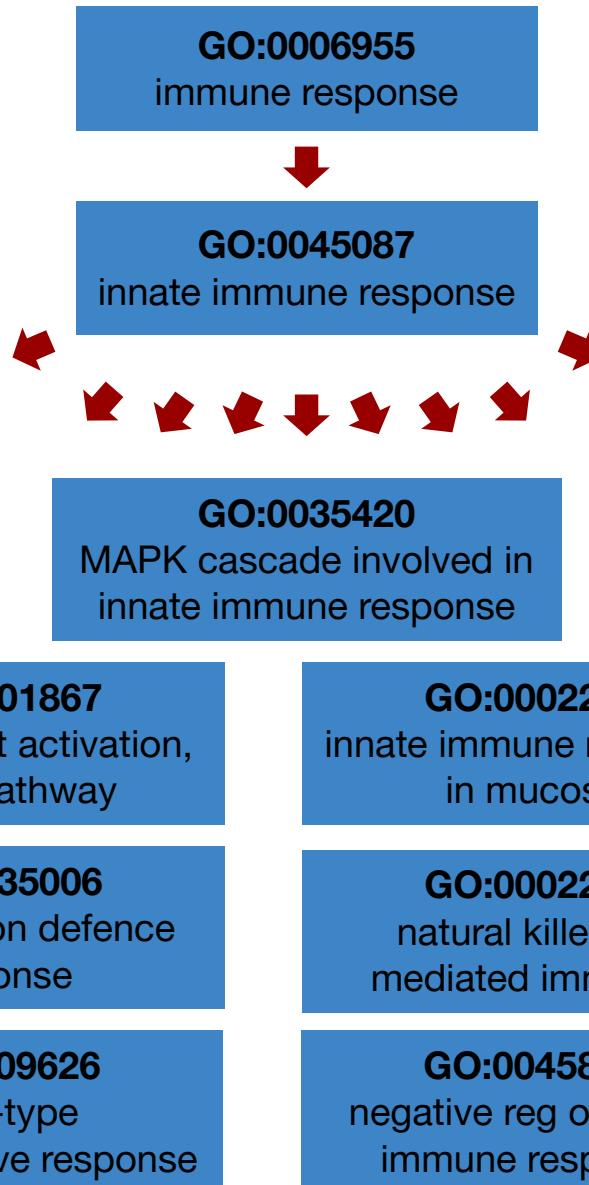
Multiple terms for
the same thing



Gene descriptions
too specific



GO terms are hierarchical



GO:0006957
complement activation,
alternative pathway

GO:0042381
hemolymph coagulation

GO:0034342
response to type II
interferon

GO:0034340
response to type I
interferon

GO:0045089
positive reg of innate
immune response

GO:0009682
induced systemic
resistance

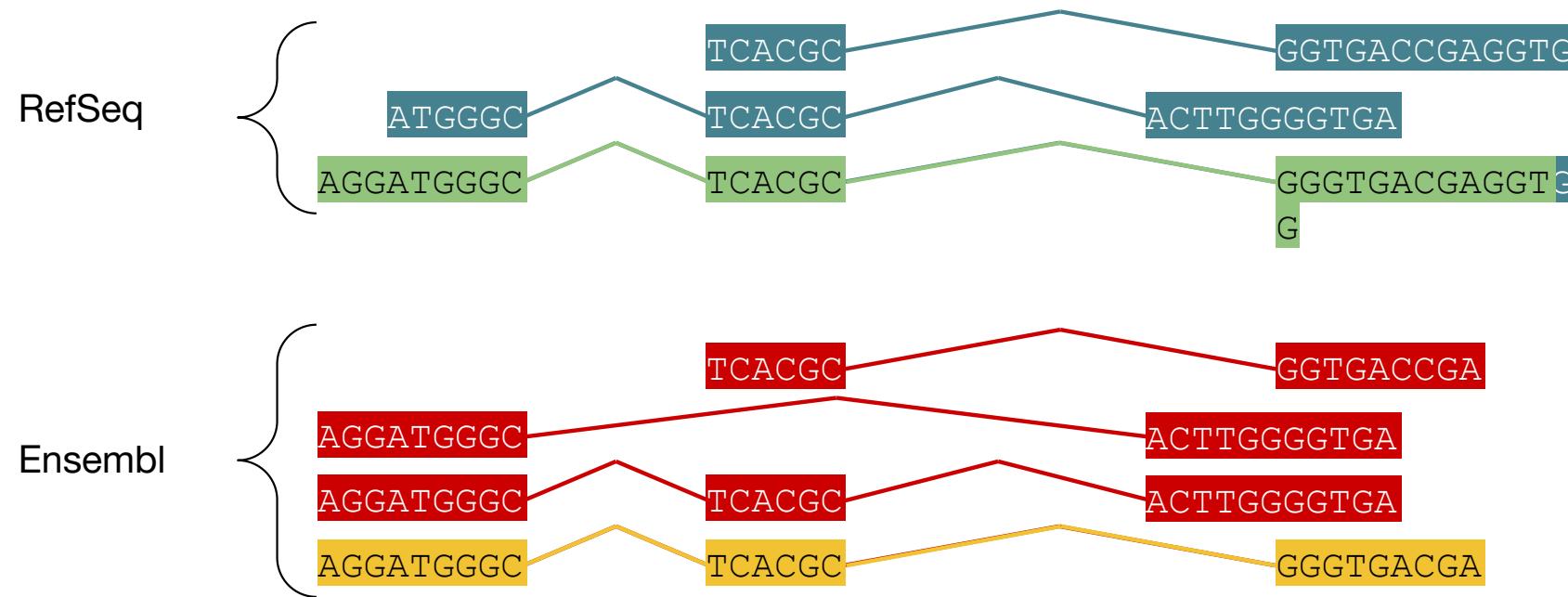
GO:0009814
defence response,
incompatible interaction

GO:0009616
virus induced gene
silencing

GO:0034341
response to
interferon-gamma

GO:0045088
regulation of innate
immune response

MANE Select

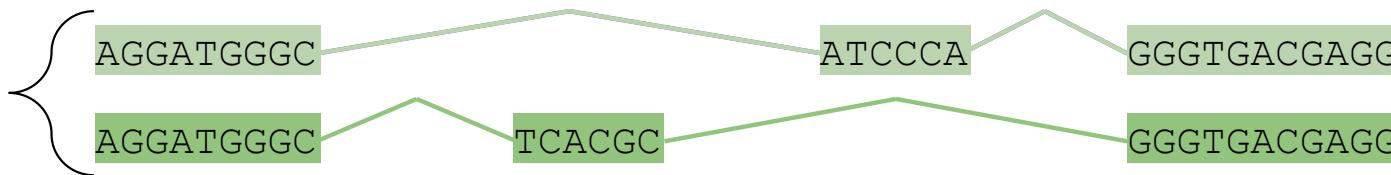


AACTAAGAATTAAAGGATGGCGTGGTGGCTACGCCTGTAATCCCAGCACTGGGTGACCGAGGTGGCGGATCACTTGA
GG

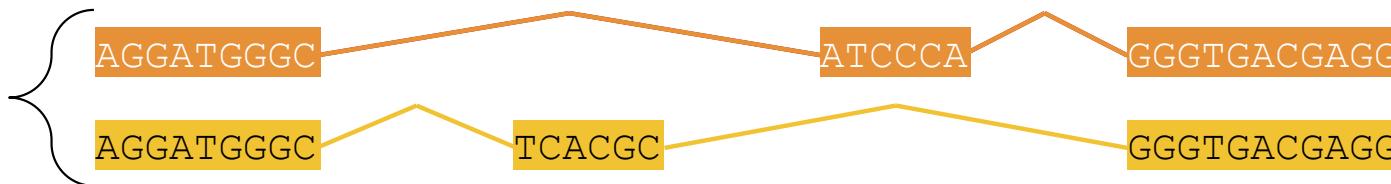
Select = most clinically relevant, based on expression
levels, clinically important genetic variation and
conservation between species

MANE Plus Clinical

RefSeq



Ensembl



AACTAAGAATTAAAGGATGGCGTGGTGGCTC ACGCCTGTAATCCA GCAC TTGGGTGACCGAGGTGGCGGATCACTTGA
GG

Clinically important variants

Canonical transcript

