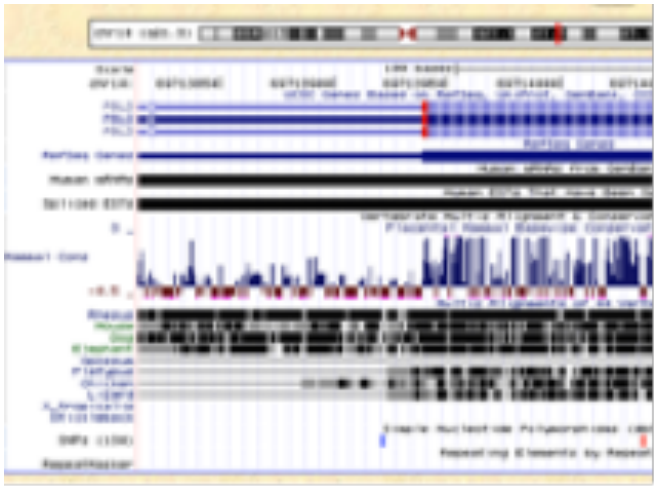
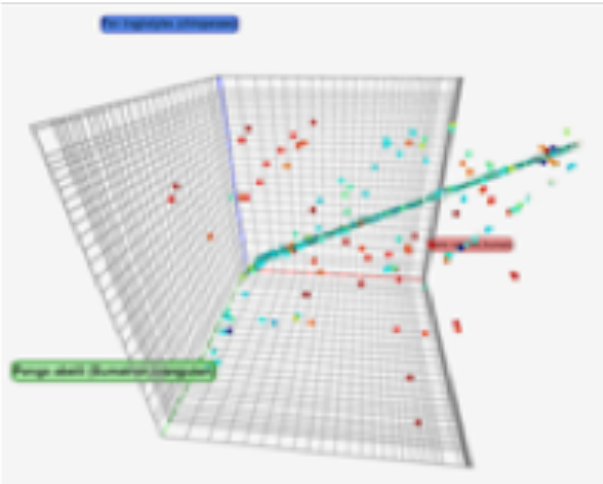
Computational Genomics

Introduction
to
Gene Models and Gene Tables



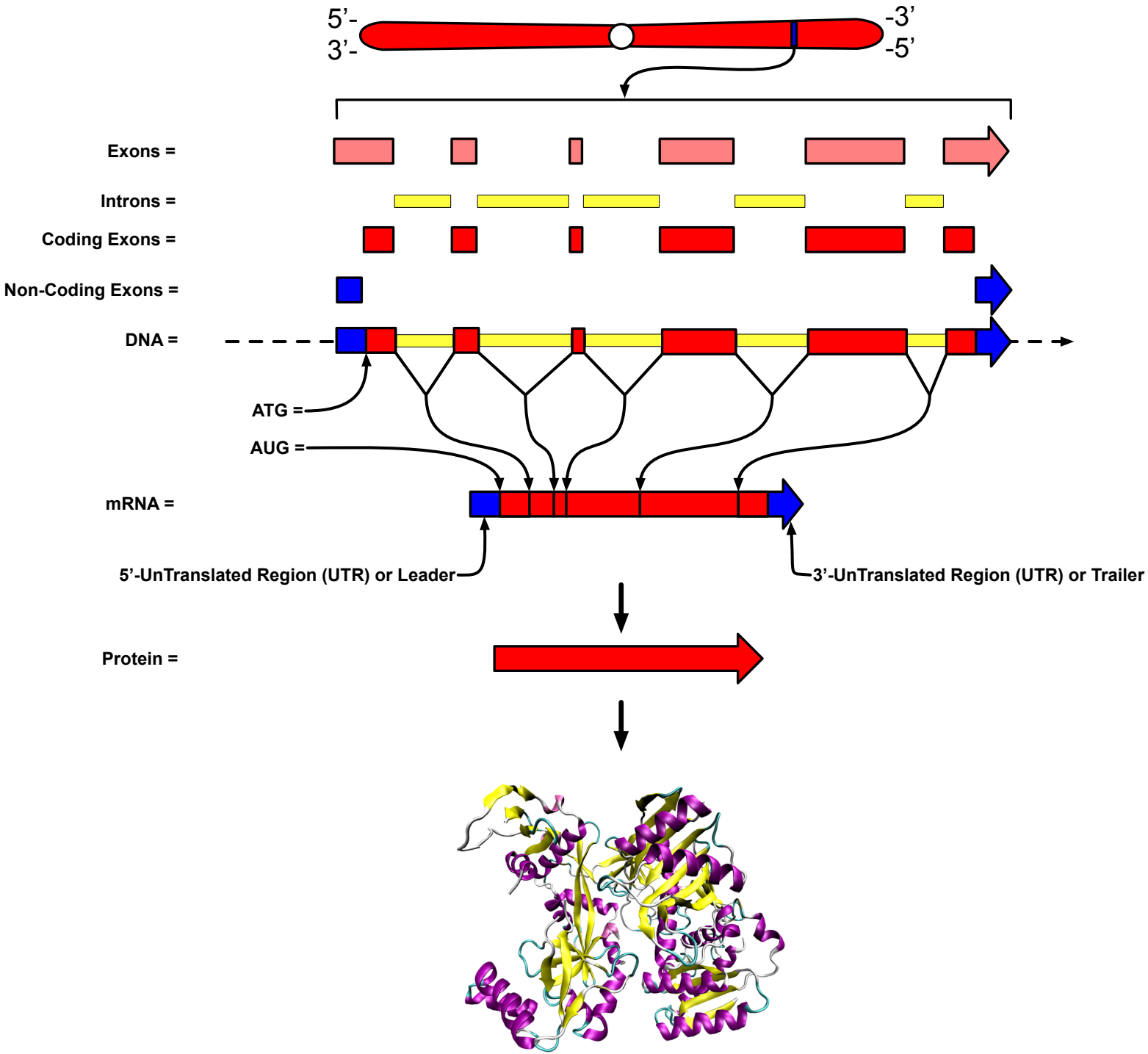


Fasta Format

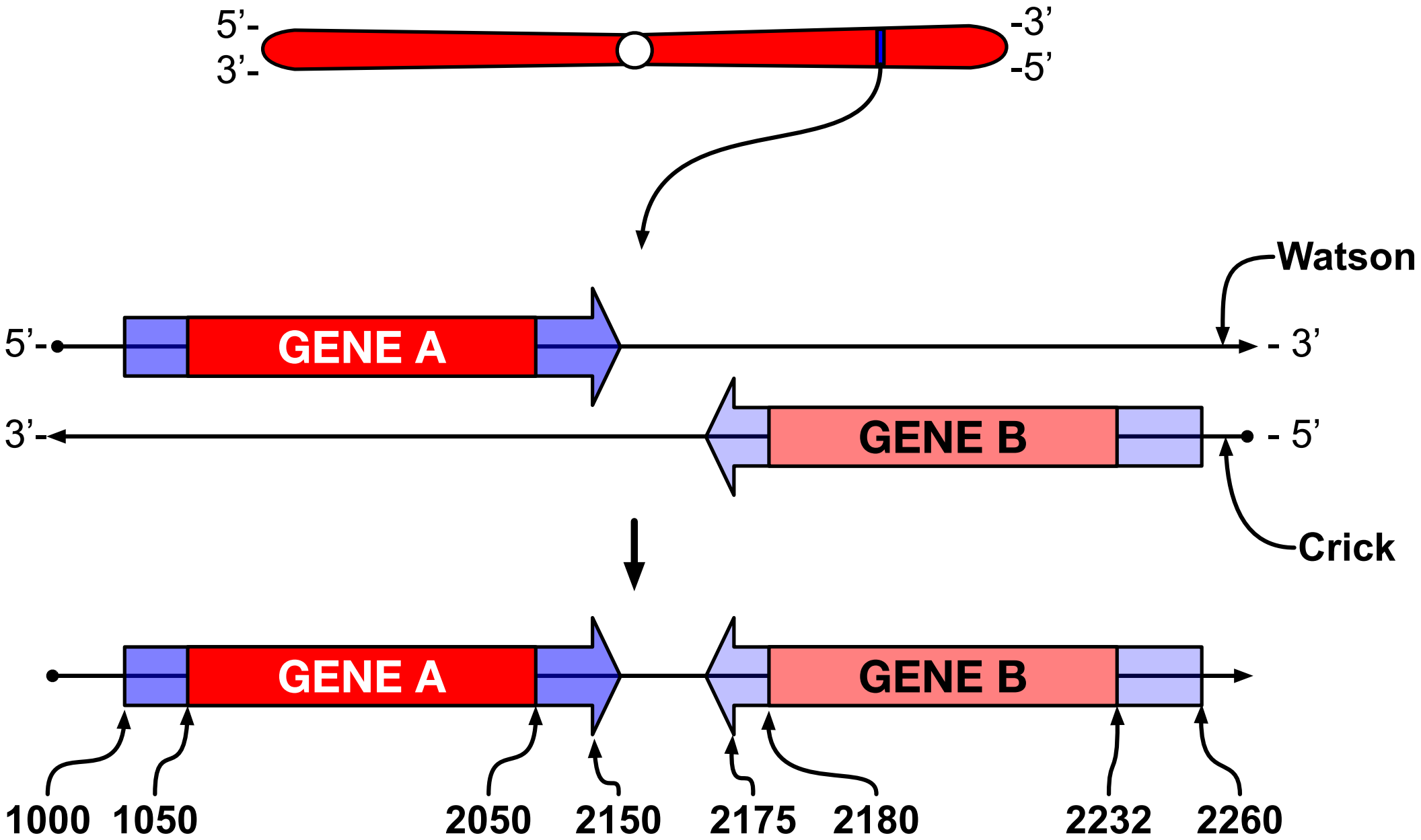
A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length.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

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Gene Models 101

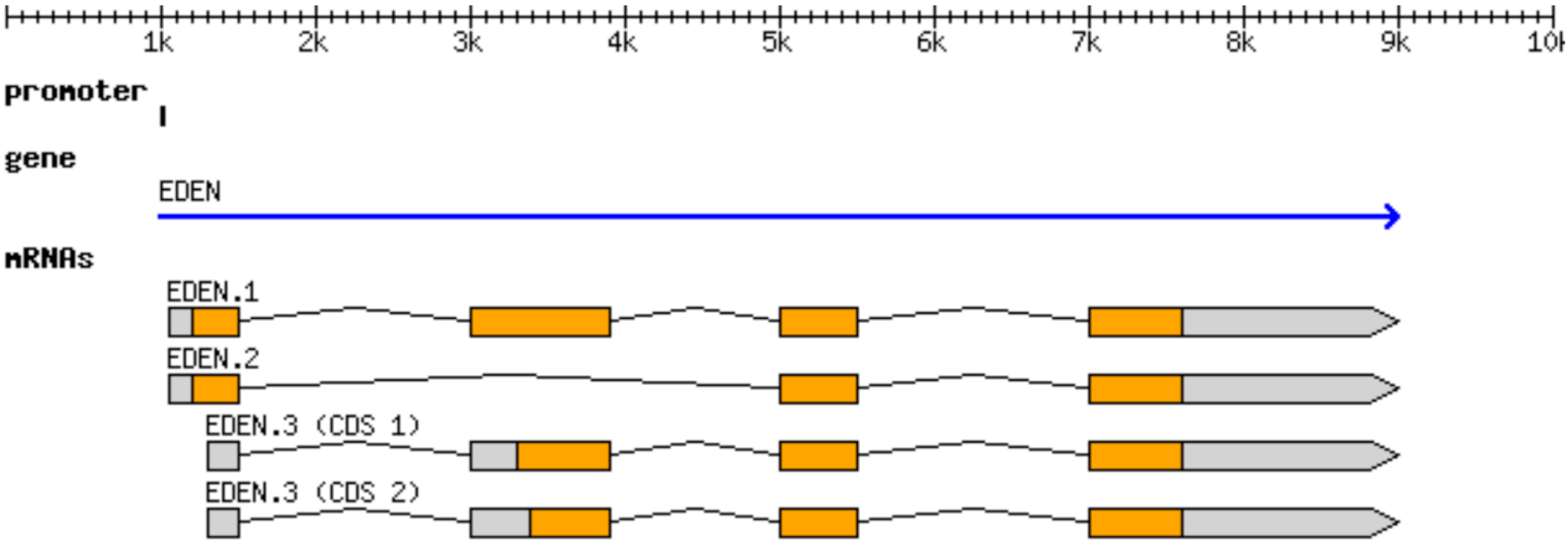


Gene Tables 101



GFF File Format and Gene Models

The Canonical Gene



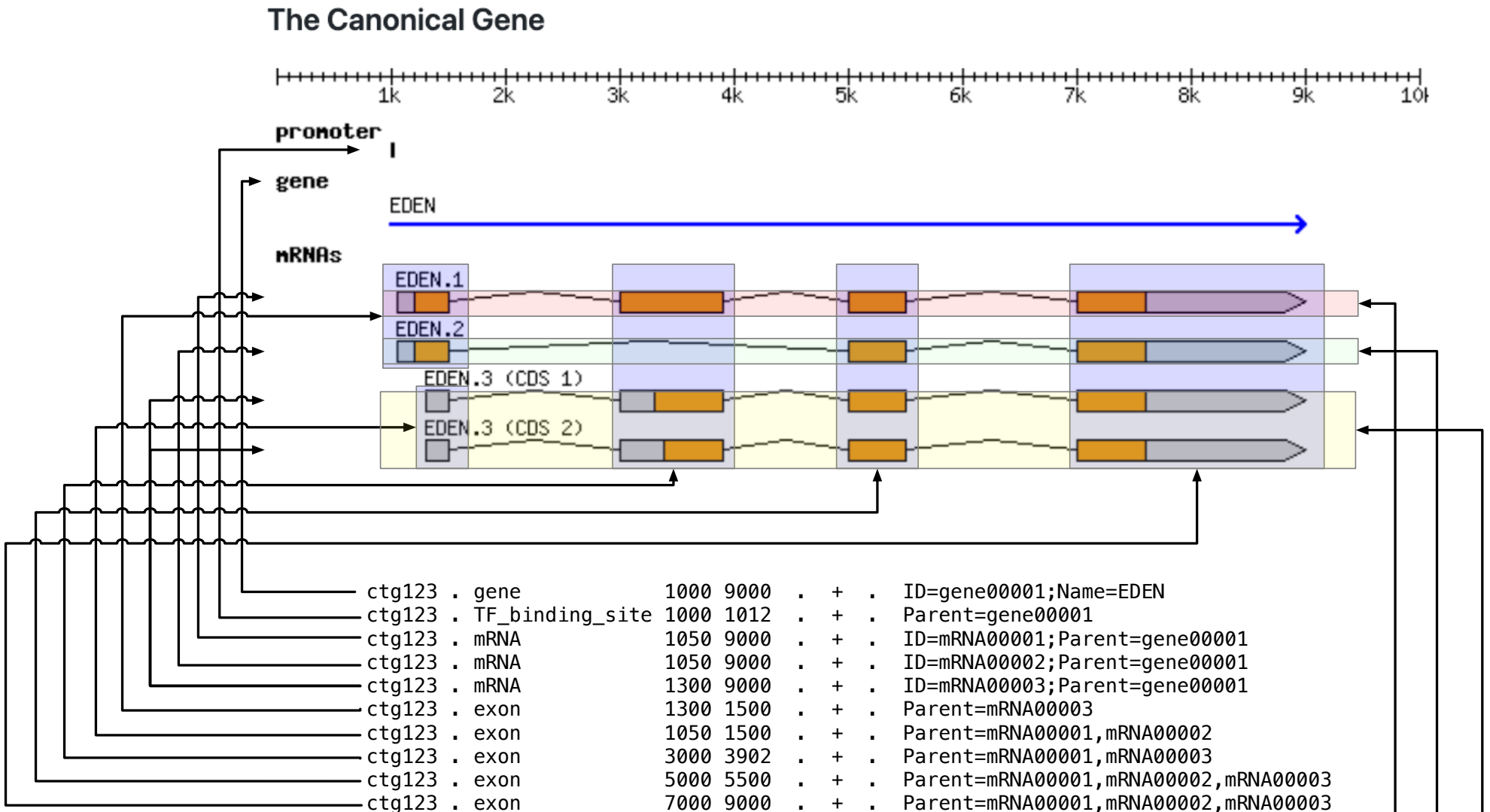
https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

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
```

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

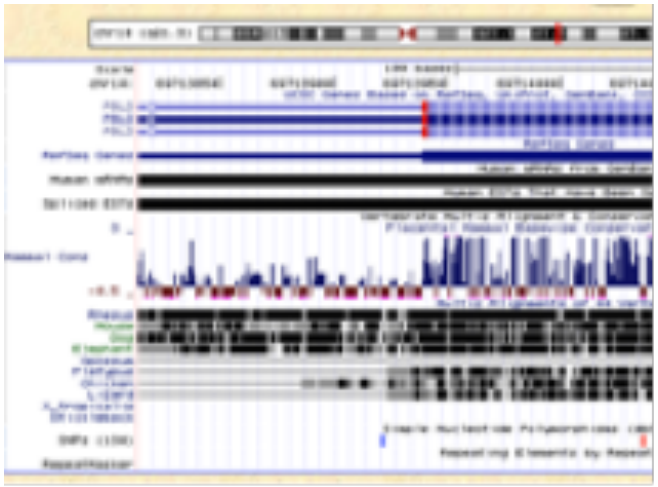
GFF File Format and Gene Models

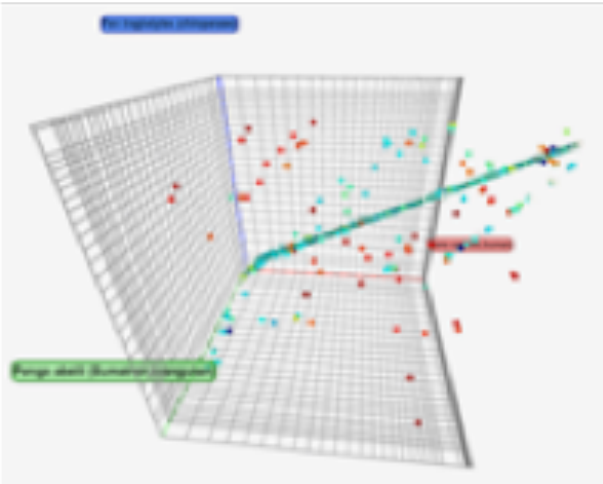



```
ctg123 . gene            1000 9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012  .  +  .  Parent=gene00001
ctg123 . mRNA            1050 9000  .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 . mRNA            1050 9000  .  +  .  ID=mRNA00002;Parent=gene00001
ctg123 . mRNA            1300 9000  .  +  .  ID=mRNA00003;Parent=gene00001
ctg123 . exon            1300 1500  .  +  .  Parent=mRNA00003
ctg123 . exon            1050 1500  .  +  .  Parent=mRNA00001,mRNA00002
ctg123 . exon            3000 3902  .  +  .  Parent=mRNA00001,mRNA00003
ctg123 . exon            5000 5500  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000 9000  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201 1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             3000 3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             5000 5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             7000 7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             1201 1500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             5000 5500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             7000 7600  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             3301 3902  .  +  0  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             5000 5500  .  +  1  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             7000 7600  .  +  1  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             3391 3902  .  +  0  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             5000 5500  .  +  1  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             7000 7600  .  +  1  ID=cds00004;Parent=mRNA00003
```

Computational Genomics

Introduction to Genome Browsers
ENSEMBL





Training materials

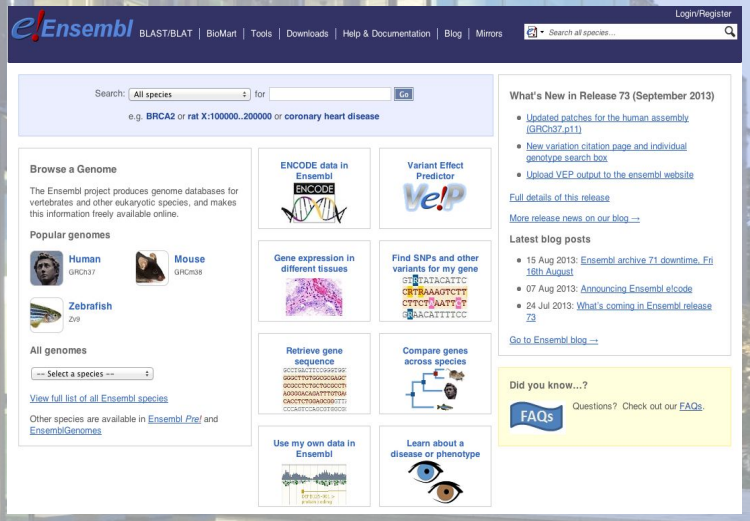
- Ensembl training materials are protected by a CC BY license:

  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



EMBL-EBI

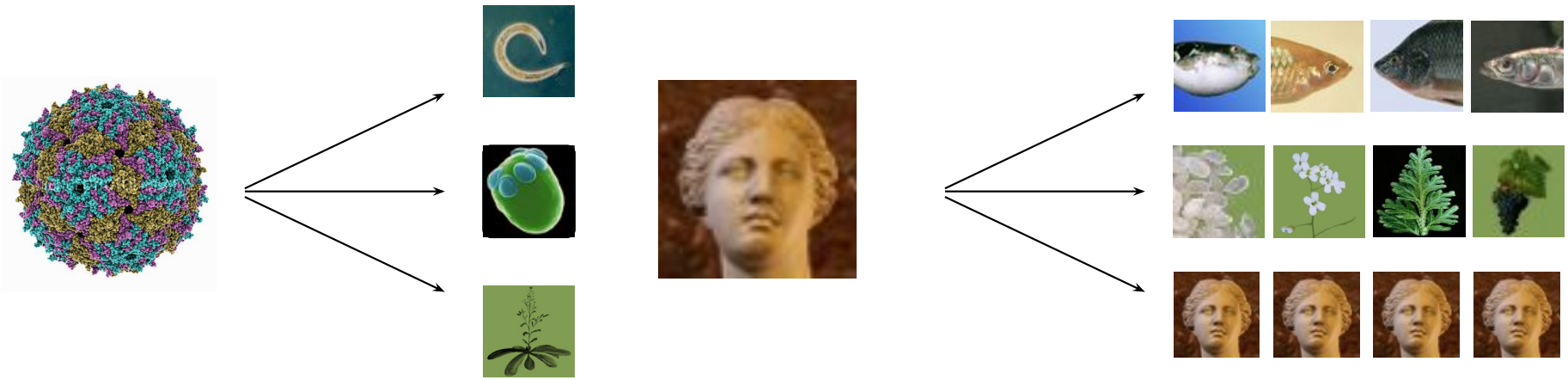
Ensembl

Introduction

Why do we need genome browsers?

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

2004: finished human sequence (3 Gb)



CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATCTGAAATTTCTTGGAA
ACACGATCACTTTAACGGAATATTGCTGTTTTGGGGAAGTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTTACTTAAGC
CCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGGAATTACCAGGCGGCGTTGGTCTCTAACTGGAGCCCTCTGTCCCC
ACTAGCCACGCGTCACTGGTTAGCGTGATTGAAACTAAATCGTATGAAAATCCTCTTCTCTAGTCGCACTAGCCACGTTTCG
AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAAAATGTTCCCATCCTCACAGTAAGCTGTTACCGTTC
CAGGAGATGGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCTCAGTCACATAATAAGGAATGCAT
CCCTGTGTAAGTGCATTTTGGTCTTCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAAAAATGCCTA
TTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTTTTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAAC
TTTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAAACCCAGTACGTCACAGTGTTGCTTAGAACCATAA
ACTGTTCCTTATGTGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACCACATGATAAATATAGAACGTCT
AGTGGATAAAGAGGAAACTGGCCCCTTGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGTTACTTTATGG
CAGAAGTTGTCCAACTTTTTGGTTTCAGTACTCCTTATACTCTTAAAAATGATCTAGGACCCCCGGAGTGCTTTTGTTTATG
TAGCTTACCATATTAGAAATTTAAAACTAAGAATTTAAGGCTGGGCGTGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGA
GGCCGAGGTGGGCGGATCACTTGAGGCCAGAAGTTTGAGACCAGCCTGGCCAACATGGTGAAACCCTATCTCTACTAAAAAT
ACAAAAAATGTGCTGCGTGTGGTGGTGCGTGCCTGTAATCCCAGCTACACGGGAGGTGGAGGCAGGAGAATCGCTTGAACCC
TGGAGGCAGAGGTTGCAGTGAGCCAAGATCATGCCACTGCACTCTAGCCTGGGCCACATAGCATGACTCTGTCTCAAAACAA
ACAAACAAACAAAAAACTAAGAATTTAAAGTTAATTTACTTAAAAATAATGAAAGCTAACCCATTGCATATTATCACAACAT
TCTTAGGAAAAATAACTTTTTGAAAACAAGTGAGTGGAATAGTTTTTACATTTTTGCAGTTCTCTTTAATGTCTGGCTAAAT
AGAGATAGCTGGATTCACTTATCTGTGTCTAATCTGTTATTTTGGTAGAAGTATGTGAAAAAAAATTAACCTCACGTTGAAA
AAAGGAATATTTTAATAGTTTTCAGTTACTTTTTGGTATTTTTCCTTGTACTTTGCATAGATTTTTCAAAGATCTAATAGAT
ATACCATAGGTCTTTCCCATGTCGCAACATCATGCAGTGATTATTTGGAAGATAGTGGTGTTCTGAATTATACAAAGTTTCC
AAATATTGATAAATTGCATTAAACTATTTTAAAAATCTCATTCATTAATACCACCATGGATGTCAGAAAAGTCTTTTAAGAT
TGGGTAGAAATGAGCCACTGGAAATTCTAATTTTCATTTGAAAGTTCACATTTTGTCATTGACAACAAACTGTTTTCCTTGC
AGCAACAAGATCACTTCATTGATTTGTGAGAAAATGTCTACCAAATTATTTAAGTTGAAATAACTTTGTCAGCTGTTCTTTC
AAGTAAAAATGACTTTTCATTGAAAAAATTGCTTGTTCAGATCACAGCTCAACATGAGTGCTTTTCTAGGCAGTATTGTACT
TCAGTATGCAGAAGTGCTTTATGTATGCTTCCTATTTTGTCAGAGATTATTAAAAGAAGTGCTAAAGCATTGAGCTTCGAAA
TTAATTTTTACTGCTTCATTAGGACATTCTTACATTAAACTGGCATTATTATTACTATTATTTTTAACAAGGACACTCAGTG
GTAAGGAATATAATGGCTACTAGTATTAGTTTGGTGCCACTGCCATAACTCATGCAAATGTGCCAGCAGTTTTACCCAGCAT
CATCTTTGCACTGTTGATACAAATGTCAACATCATGAAAAAGGGTTGAAAAAAGGAATATTTTAATAGTTTTCAGTTACTTT

What is Ensembl?

AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAA
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA
CTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCT
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCT
TCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGT
AAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTT
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAACT
TTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAA
ACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCTTATG
TGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACC
ACATGATAAATATAGAACGTCTACTGGATAAAGAGGAAACTGGCCCC
TTGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGT

Ensembl annotates and maps genomic features from genome sequences

What is Ensembl?

AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAA
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA
CTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCT
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCT
TCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGT
AAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTT
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAACT
TTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAA
ACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCTTATG
TGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACC
ACATGATAAATATAGAACGTCTAC[illegible]GGATAAAGAGGAAACTGGCCCC
TTGACTAGCAGTAGGAACAATTAC[illegible]AACAAATCAGAAGCATTAATGT

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

What is Ensembl?

```
AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAA
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA
CTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCT
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCT
TCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGT
AAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTT
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAACT
TTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAA
ACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCTTATG
TGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACC
ACATGATAAATATAGAACGTCTACTGGATAAAGAGGAAACTGGCCCC
TTGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGT
```

NIH U.S. National Library of Medicine | NCBI

Genome Data Viewer

www.ncbi.nlm.nih.gov/genome/gdv/

e!Ensembl

www.ensembl.org

UCSC

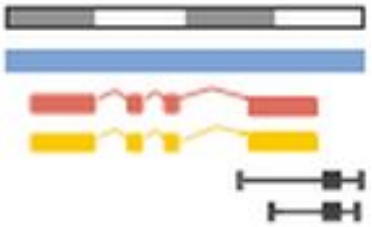
www.genome.ucsc.edu

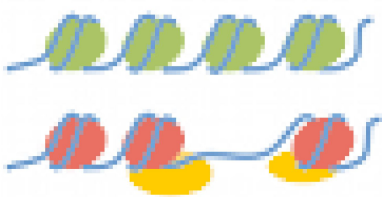
e!EnsemblGenomes

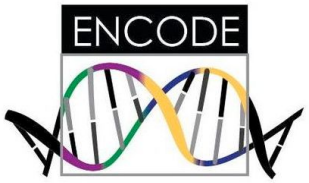
www.ensemblgenomes.org

Ensembl features

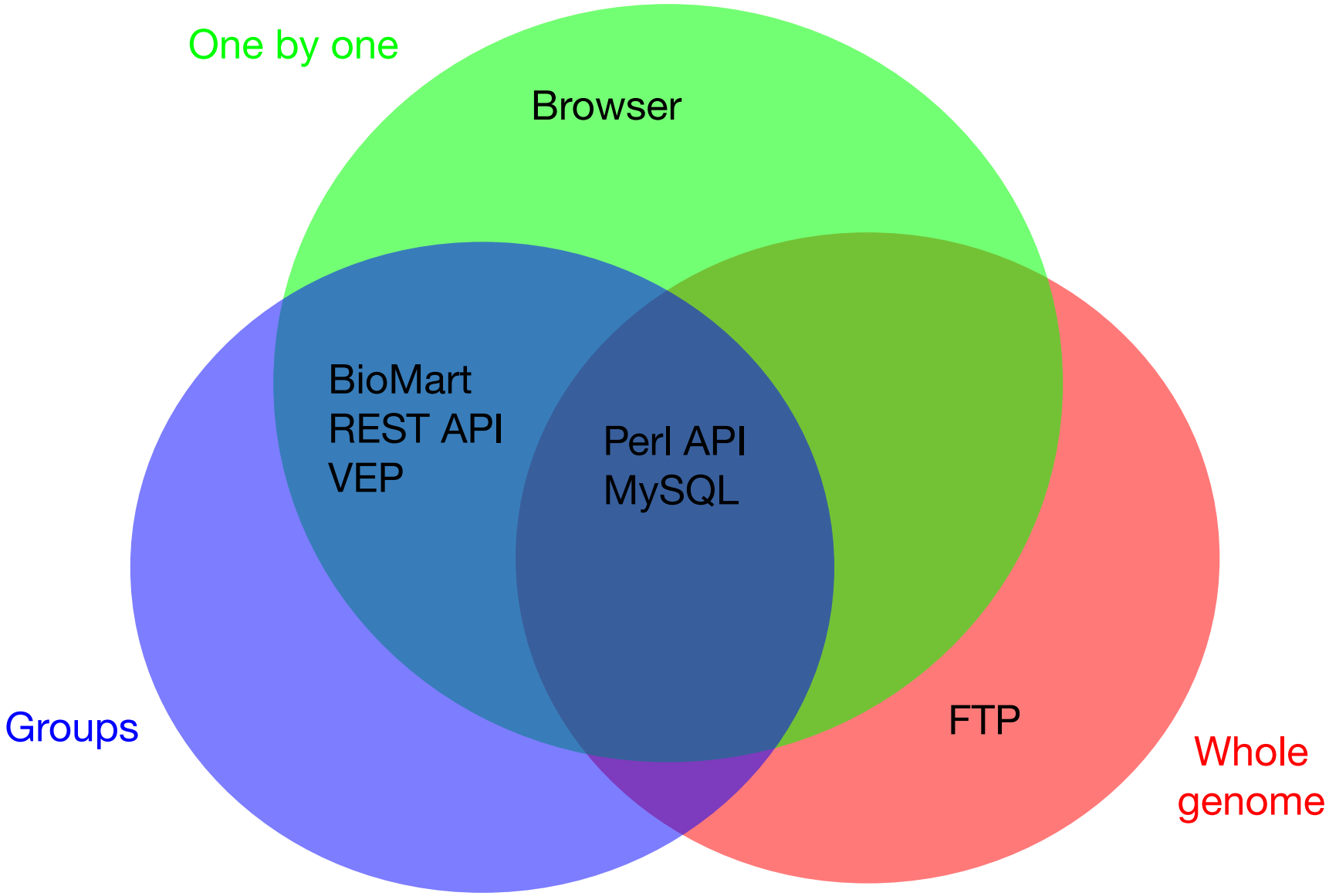
- Gene builds for >300 species
- Gene trees
- Regulatory build (ENCODE)
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source



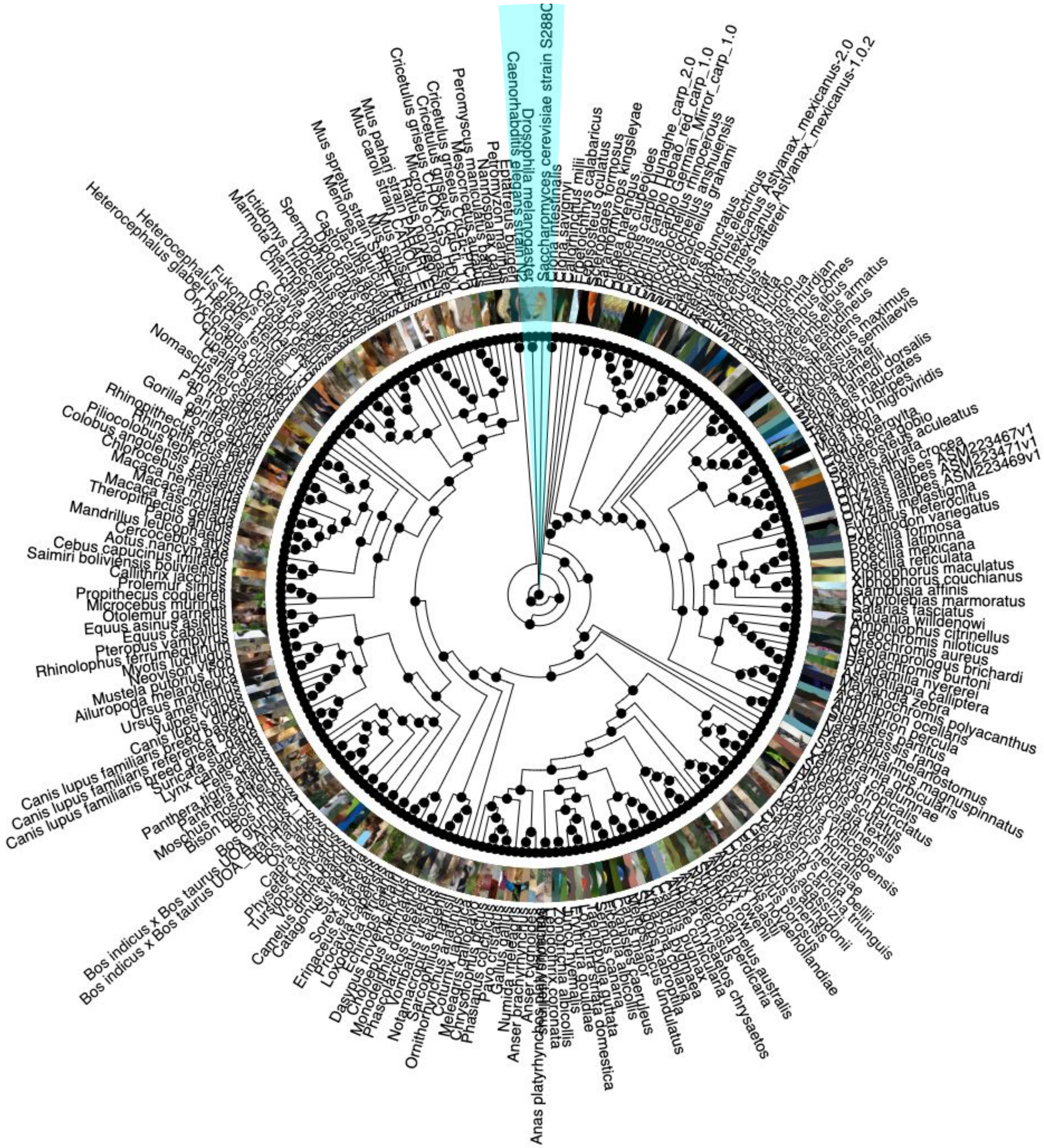




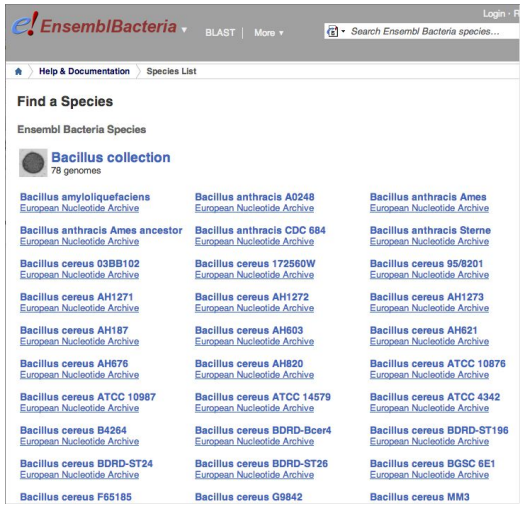
Access scales



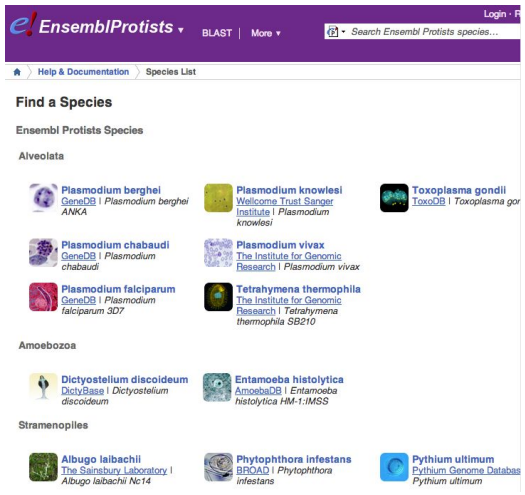
Vertebrate species in Ensembl



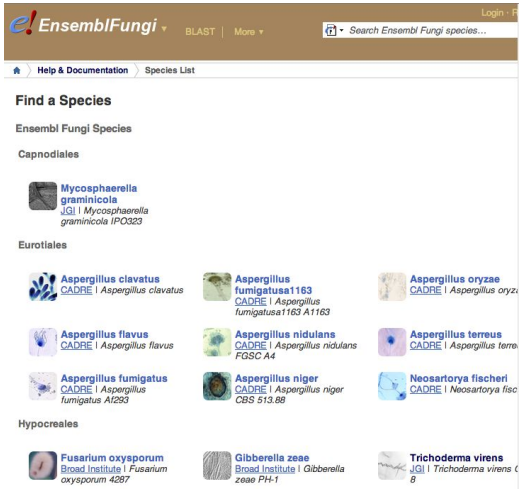
Non-vertebrates on Ensembl genomes



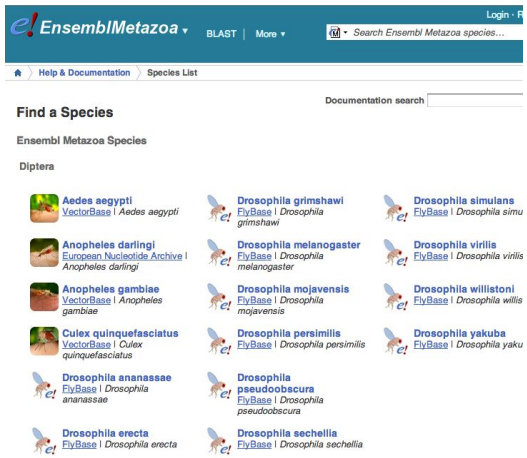
Bacteria



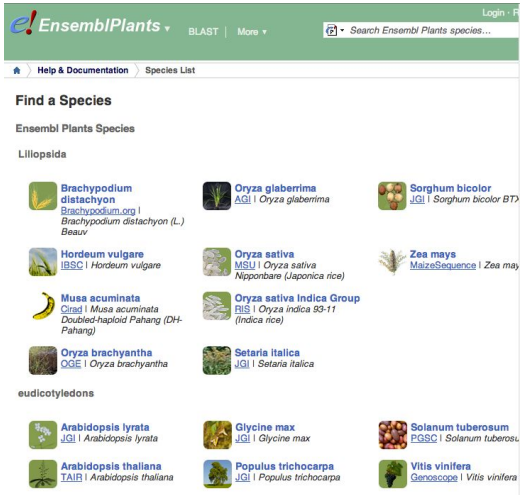
Protists



Fungi



Metazoa



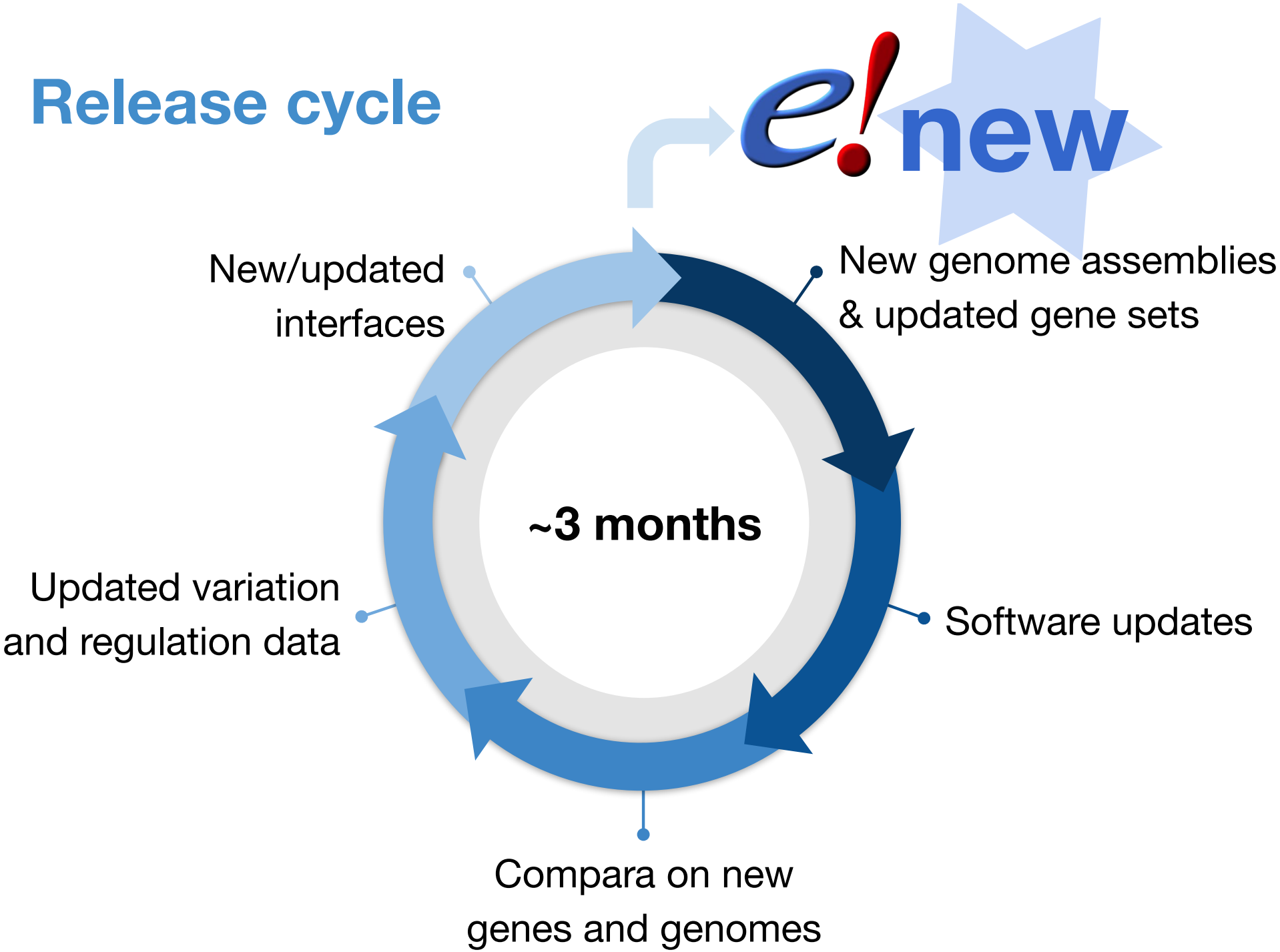
Plants

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

BL

Cut into chunks and clone into BACs

Grow up in bacterial cultures

Making a contig

Sequence reads

```
                    CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA
  CAGCTGTCCCAGATGAC      ACTTAACTTCCCTCCCAGCTGTCC
GGGCTCCGCCTTCAGCTC                    TCCCAGCTGTCCCAGATGACGCCATC
                  AACTTCCCTCCCAGCT
   CGGCCTTTGGGCTCC                    TCCGCCTTCAGCTCAAGACTTAACTTC
                    CAGATGACGCC
```

Match up overlaps

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA   AACTTCCCTCCCAGCT    CAGATGACGCC
            TCCGCCTTCAGCTCAAGACTTAACTTC  TCCCAGCTGTCCCAGATGACGCCATC
       GGGCTCCGCCTTCAGCTC   ACTTAACTTCCCTCCCAGCTGTCC
CGGCCTTTGGGCTCC                          CAGCTGTCCCAGATGAC
```

Contig

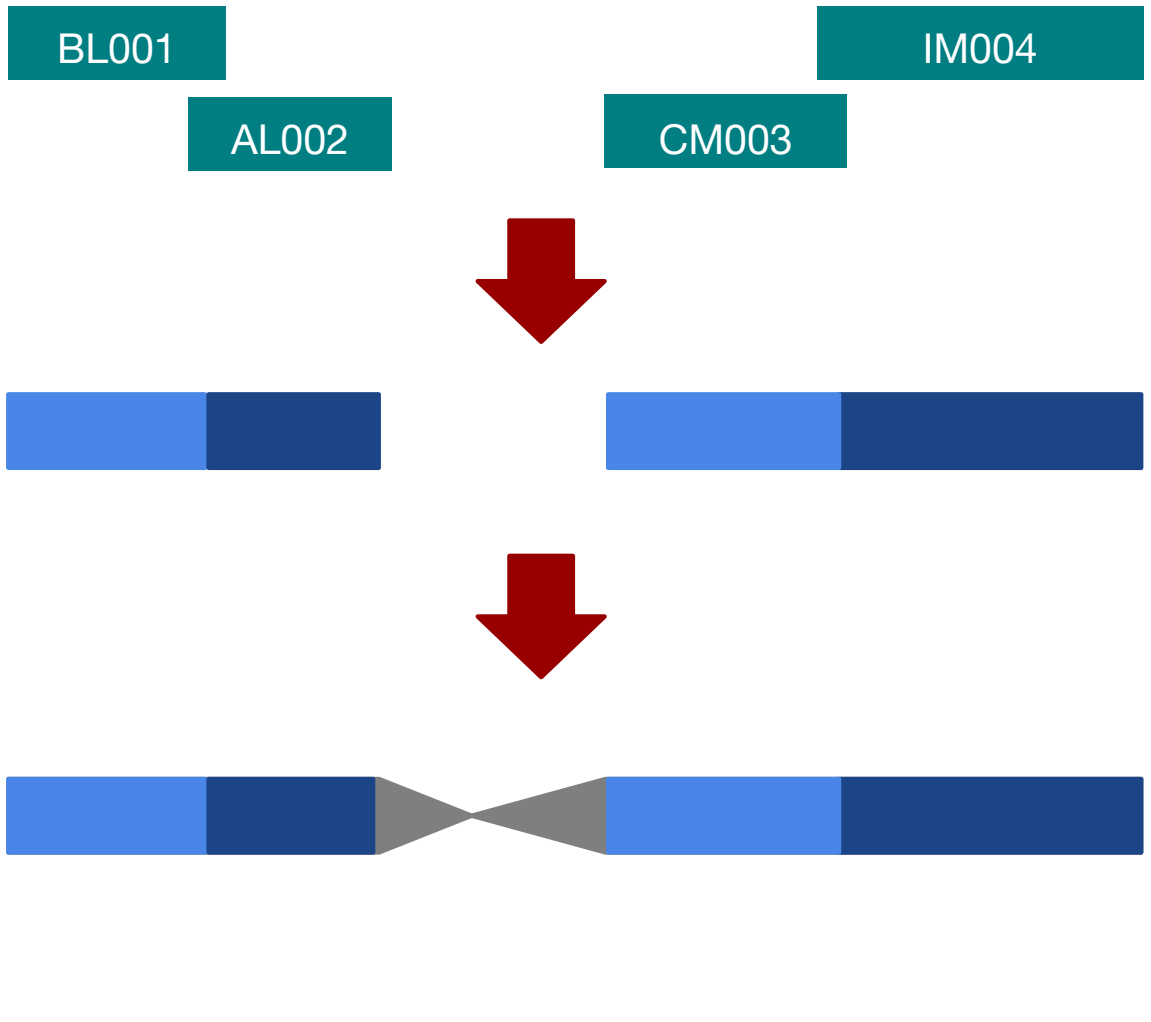
```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATC
```

Contigs to scaffolds

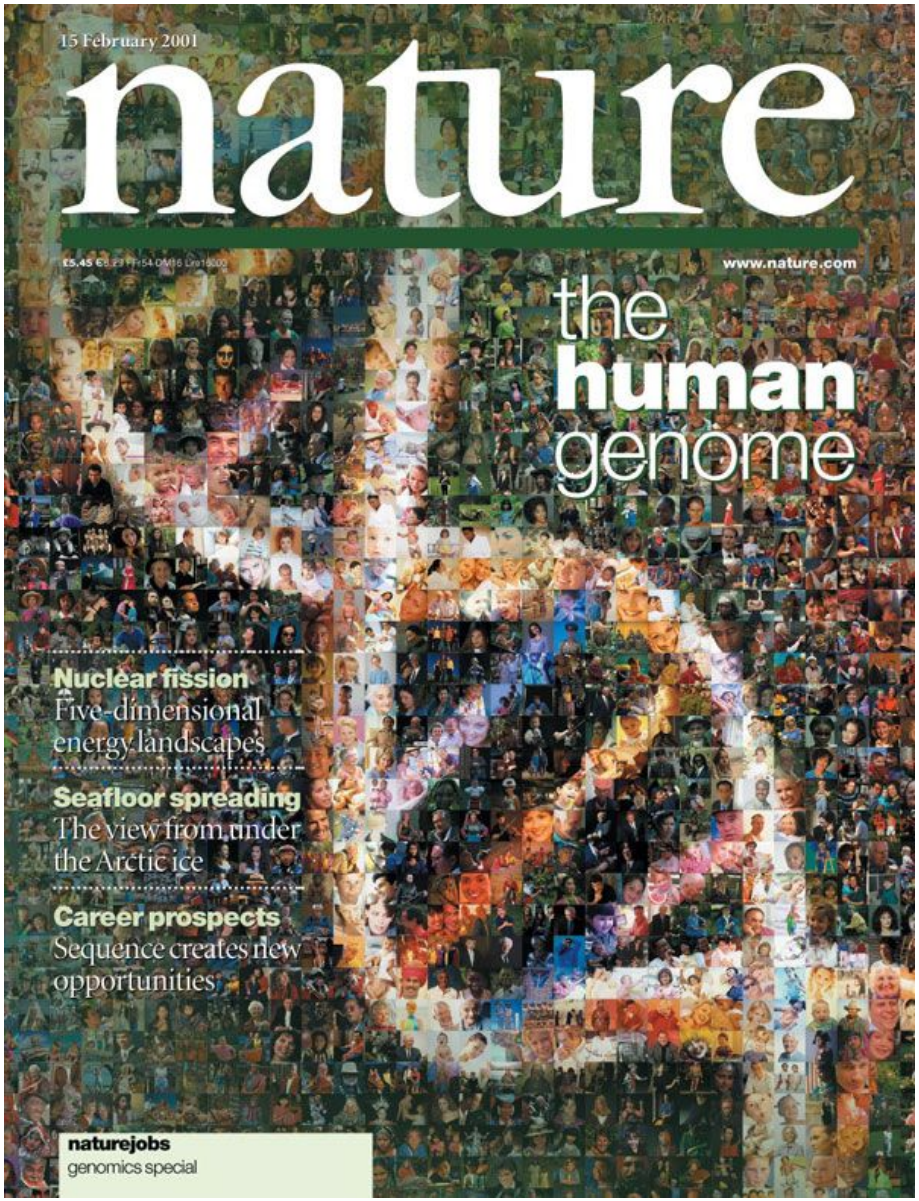
BACs from different individuals assembled together with overlaps Tilepath

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



Genome assemblies



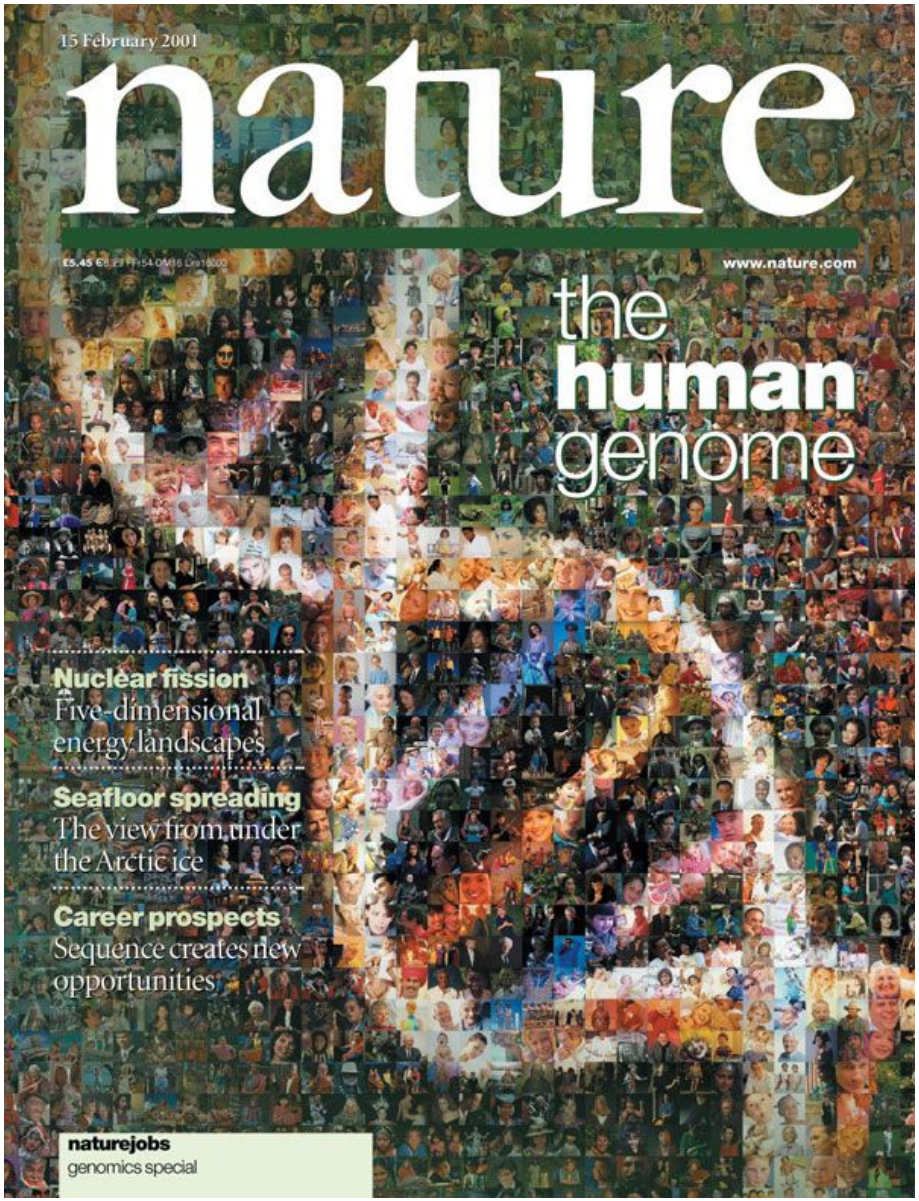
BL

AL

CM

IM

Genome contigs



BL

AL

CM

IM

BL001
AL002
CM003
IM004

Genome assemblies

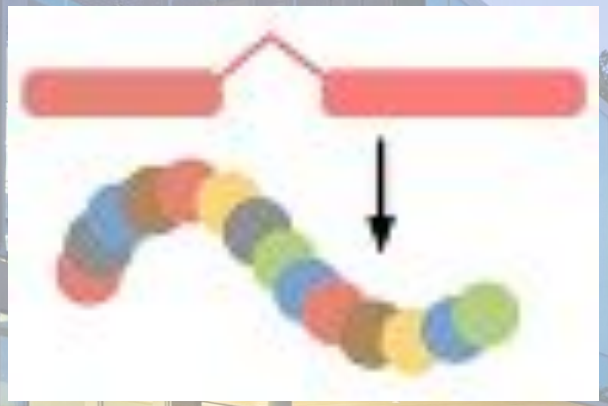
Genome
"DNA within a cell"

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

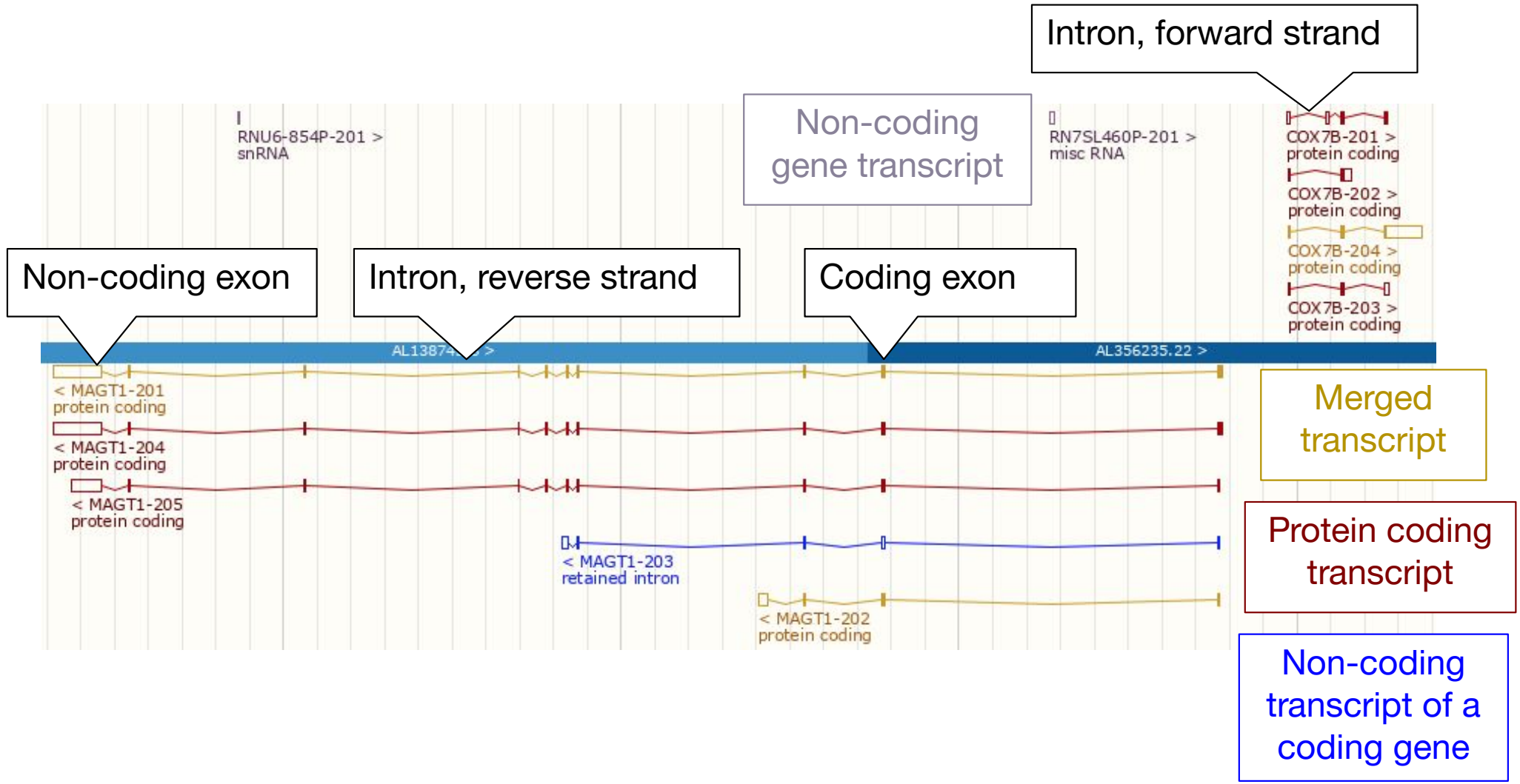
Genes and Transcripts



EMBL-EBI

Ensembl

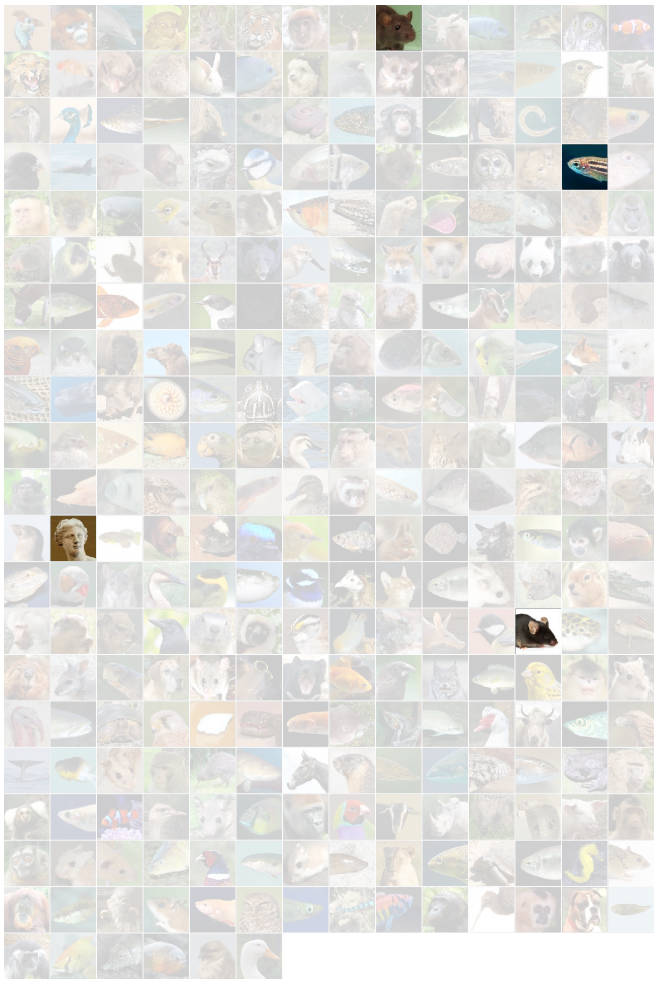
Gene views


Intron, forward strand
Non-coding gene transcript
RNU6-854P-201 >
snRNA
RN7SL460P-201 >
misc RNA
COX7B-201 >
protein coding
COX7B-202 >
protein coding
COX7B-204 >
protein coding
COX7B-203 >
protein coding
Non-coding exon
Intron, reverse strand
Coding exon
AL356235.22 >
< MAGT1-201
protein coding
< MAGT1-204
protein coding
< MAGT1-205
protein coding
< MAGT1-203
retained intron
< MAGT1-202
protein coding
Merged transcript
Protein coding transcript
Non-coding transcript of a coding gene

Automatic and manual annotation

Automatic annotation

Manual annotation



Automatic and manual annotation

Automatic annotation

- computational pipeline
- one genone = 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:

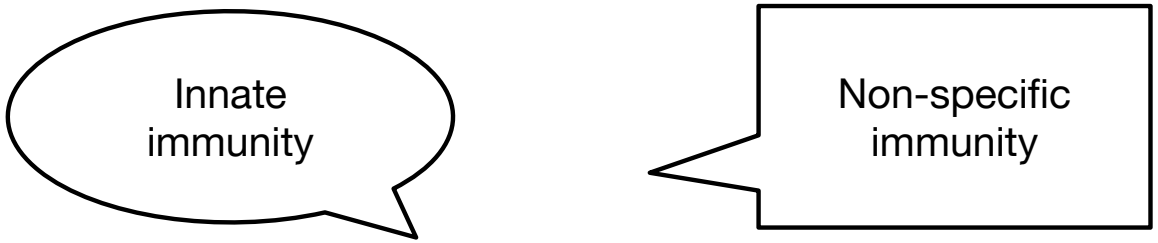
  MUS (*Mus musculus*) for mouse ENSMUSG###

  DAR (*Danio rerio*) for zebrafish: ENSDARG###

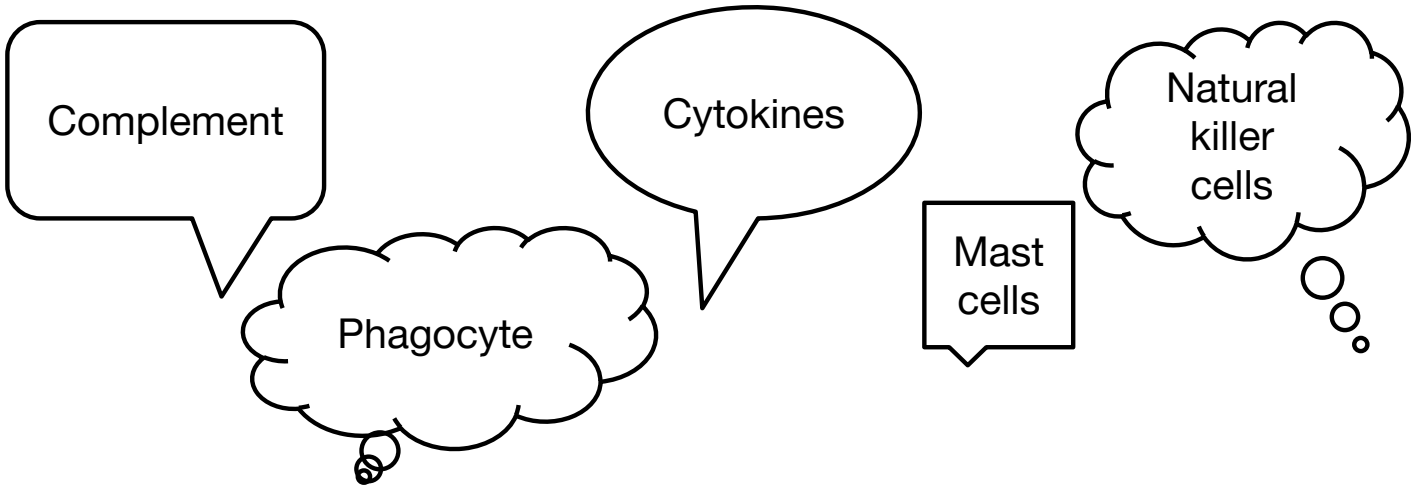
http://www.ensembl.org/info/genome/stable_ids/index.html

Why Gene Ontology (GO)?

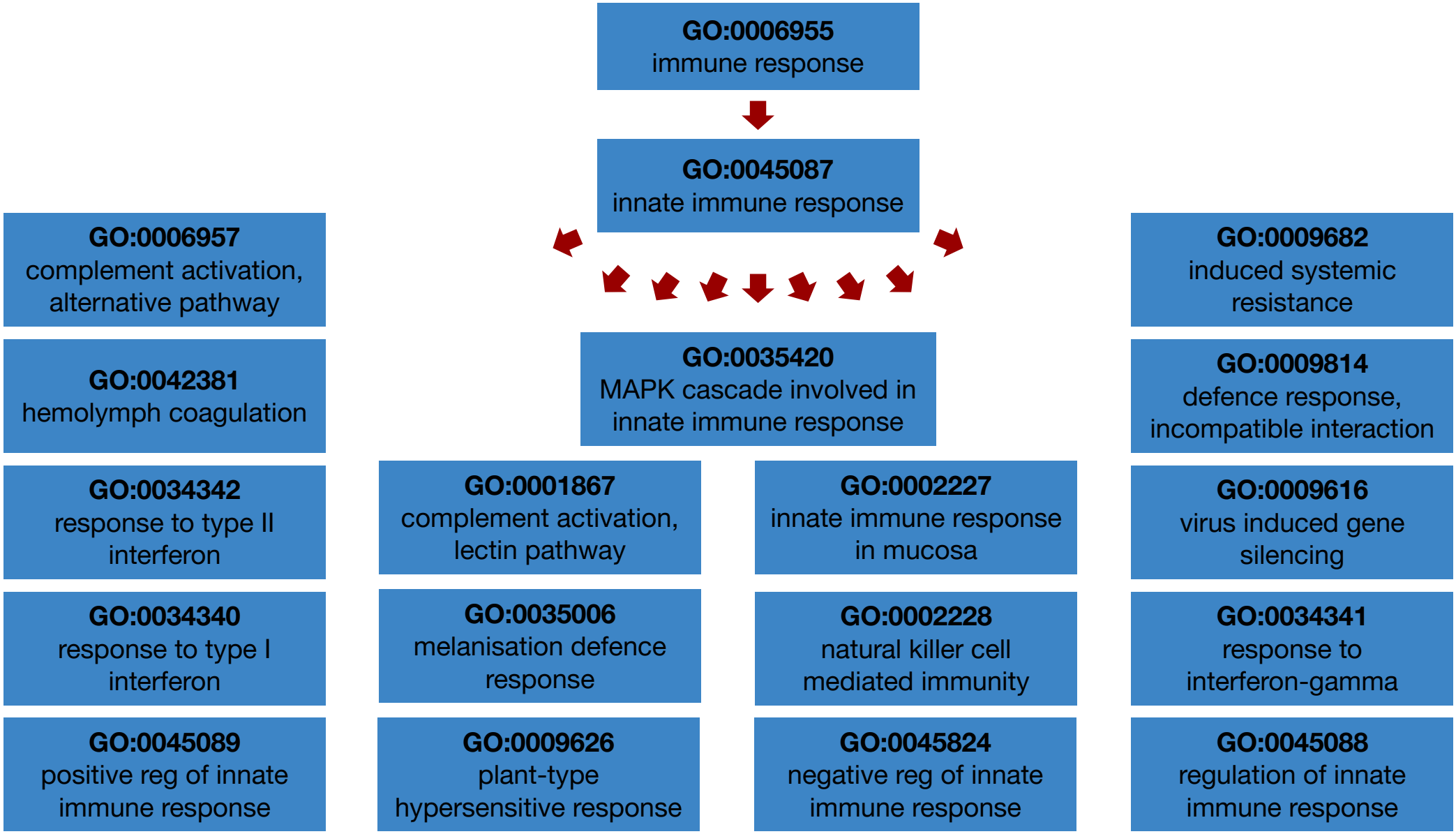
Multiple terms for the same thing



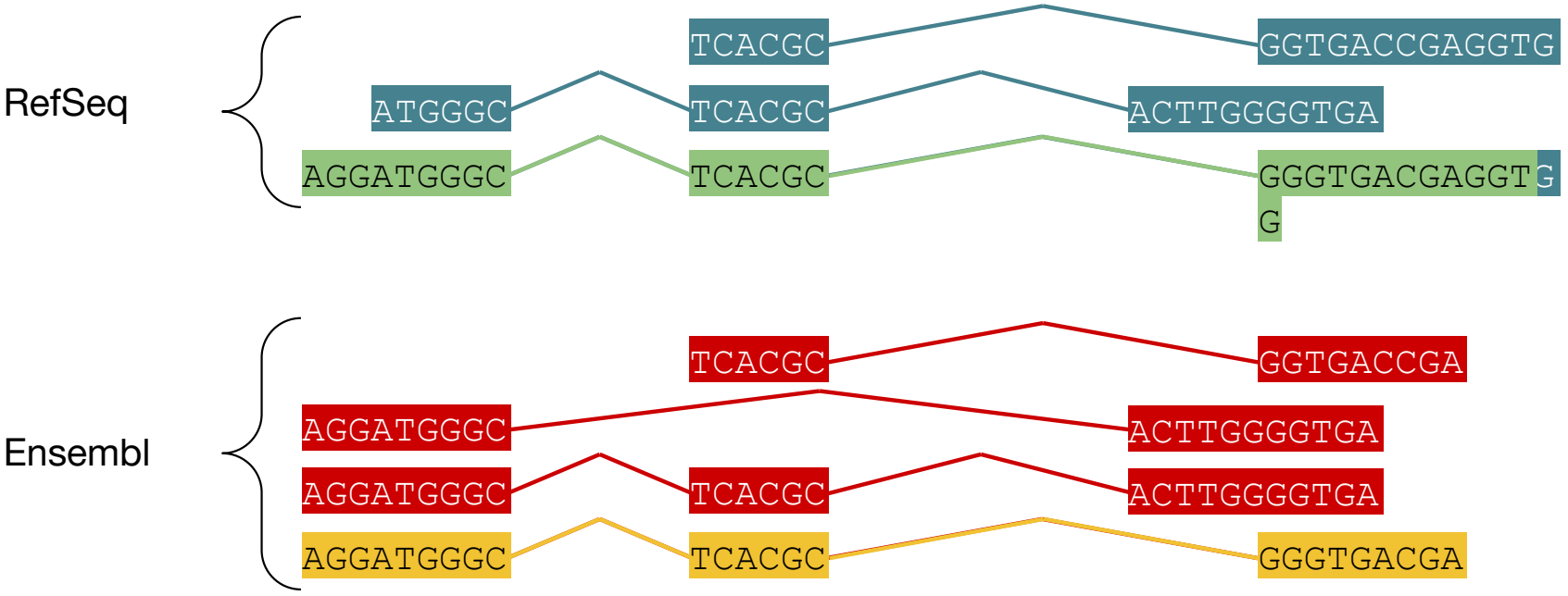
Gene descriptions too specific



GO terms are hierarchical



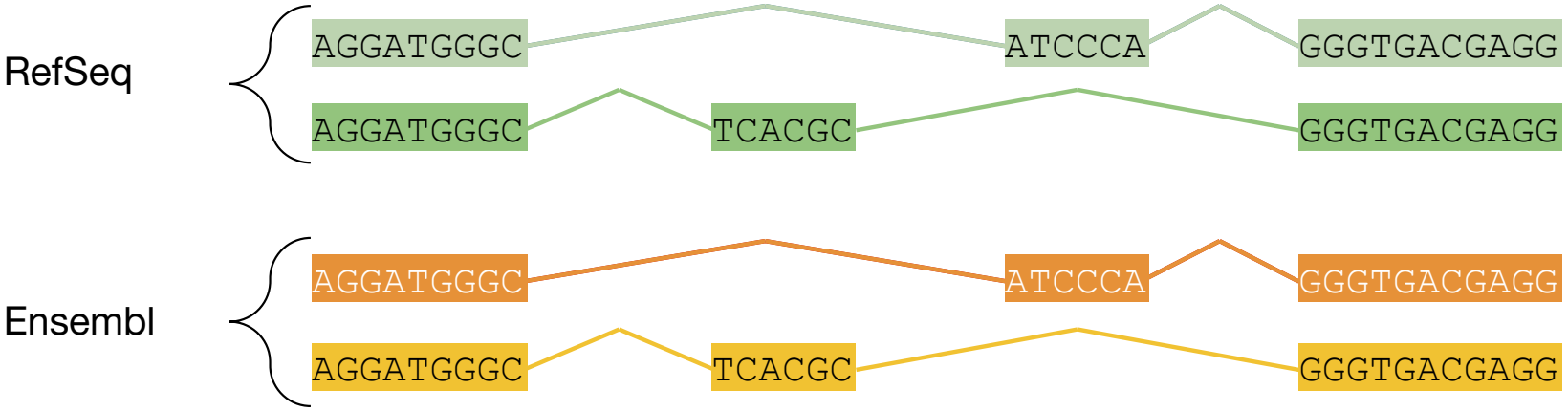
MANE Select



AACTAAGAATTTAAGGATGGGCGTGGTGGCTCACGCCTGTAATCCCAGCACTTGGGGTGACCGAGGTGGGCGGATCACTTGAGG

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

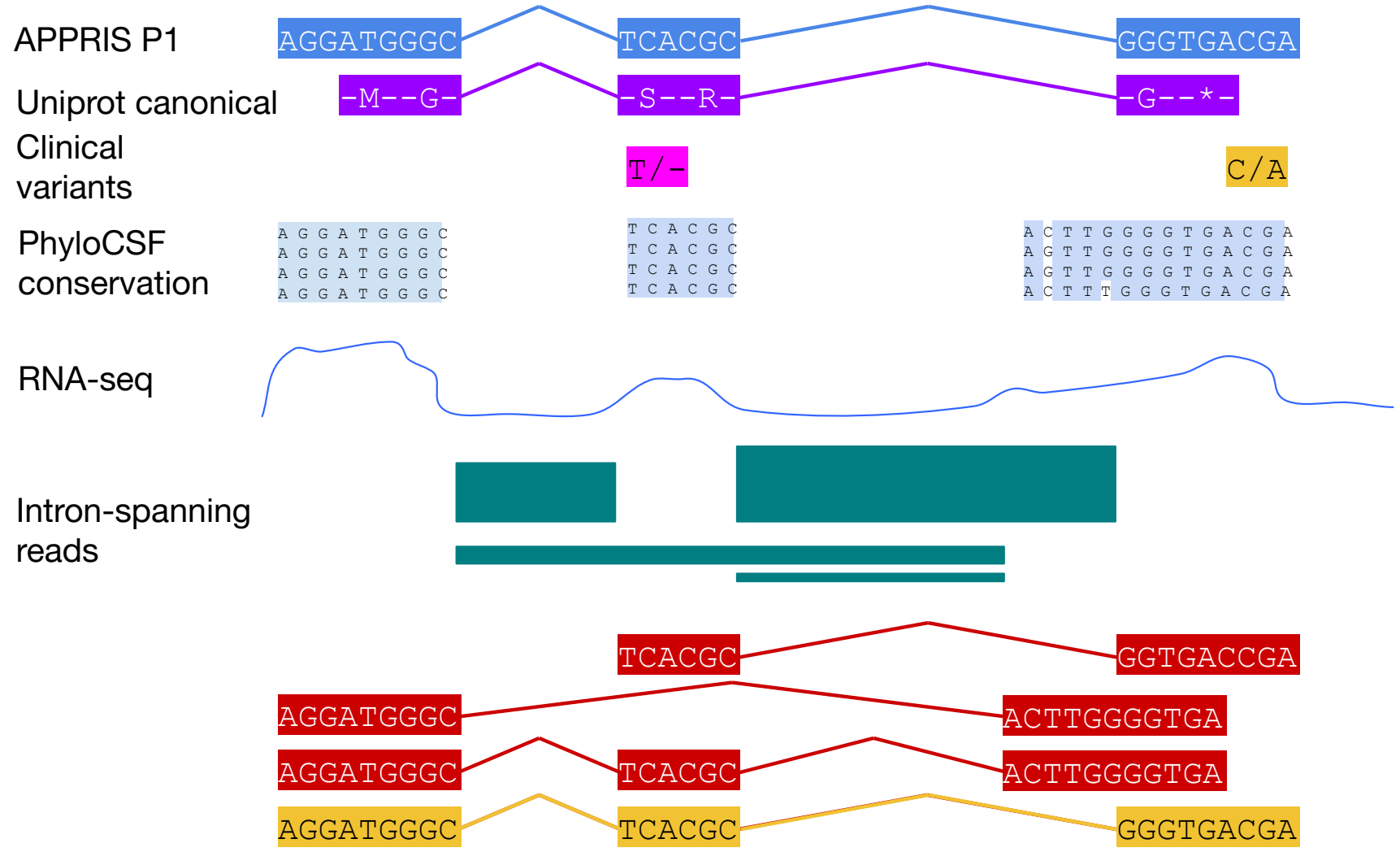
MANE Plus Clinical



AACTAAGAATTTAAGGATGGGCGTGGTGGCTC ACGCCTGTAATCCCA GCACTTGGGGTGACCGAGGTGGGCGGATCACTTGA
GG

Clinically important variants

Canonical transcript



AACTAAGAATTTAAGGATGGGCGTGGTGGCTCACGCCTGTAATCCCAGCACTTGGGGTGACCGAGGTGGGCGGATCACTTGA
GG