#AprendeBioinformáticaEnCasa

Ensembl Overview

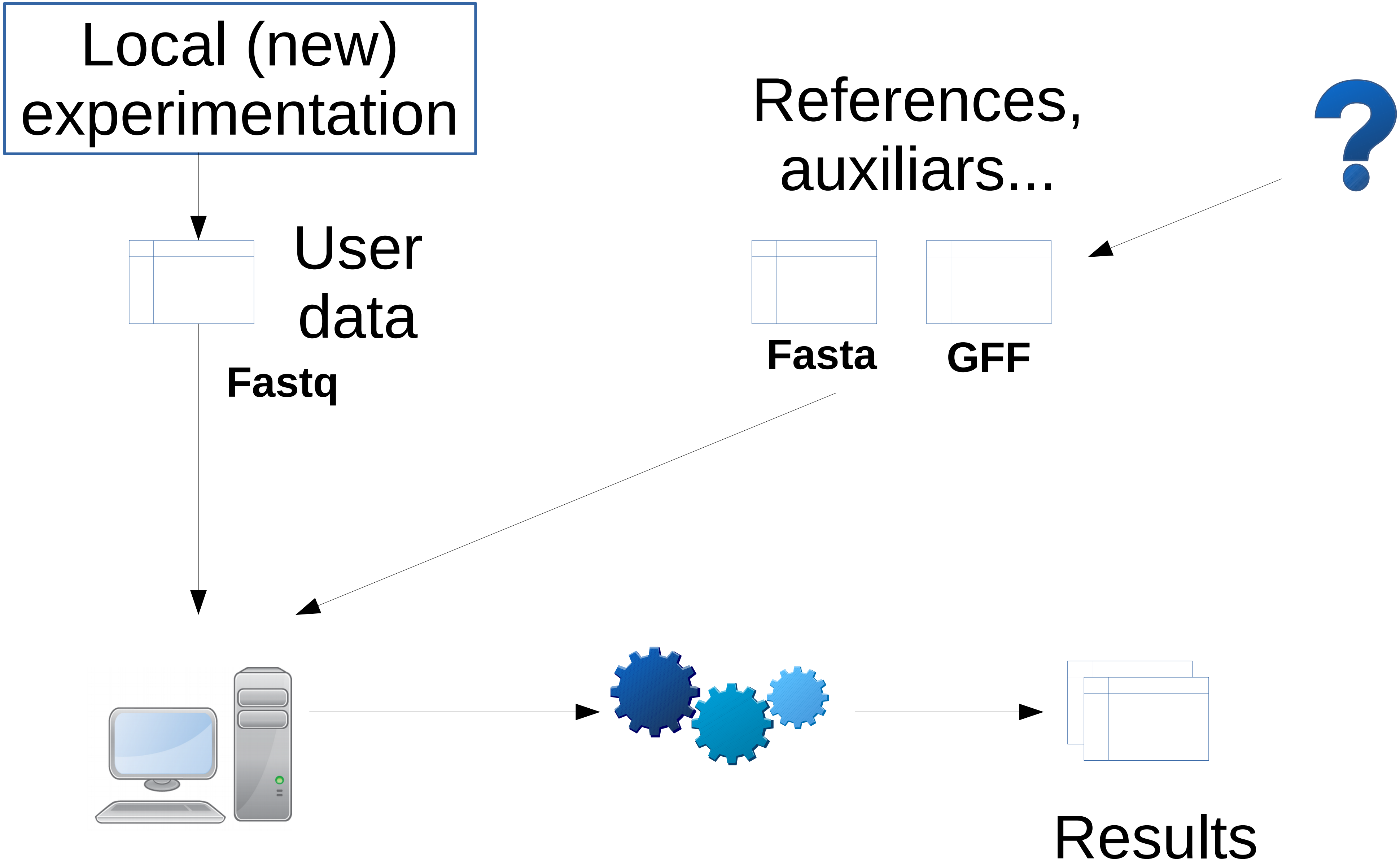
Rafael Torres-Perez

rafael.torres@cnb.csic.es

#QuedateEnCasa 27/04/2020

BioinfoGP

Bioinformatics for Genomics and Proteomics



Local (new) experimentation

User data

Fastq

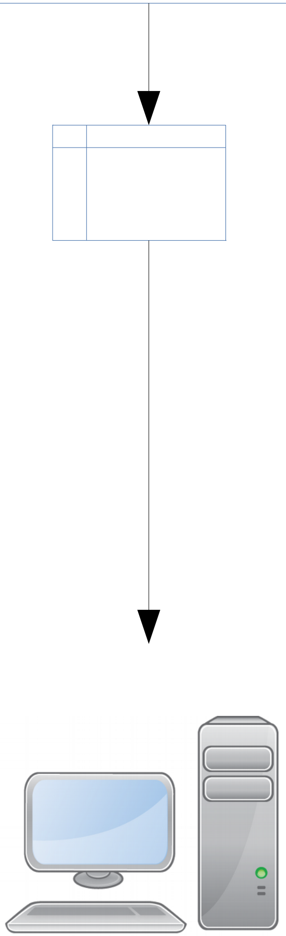
References, auxiliars...

Fasta

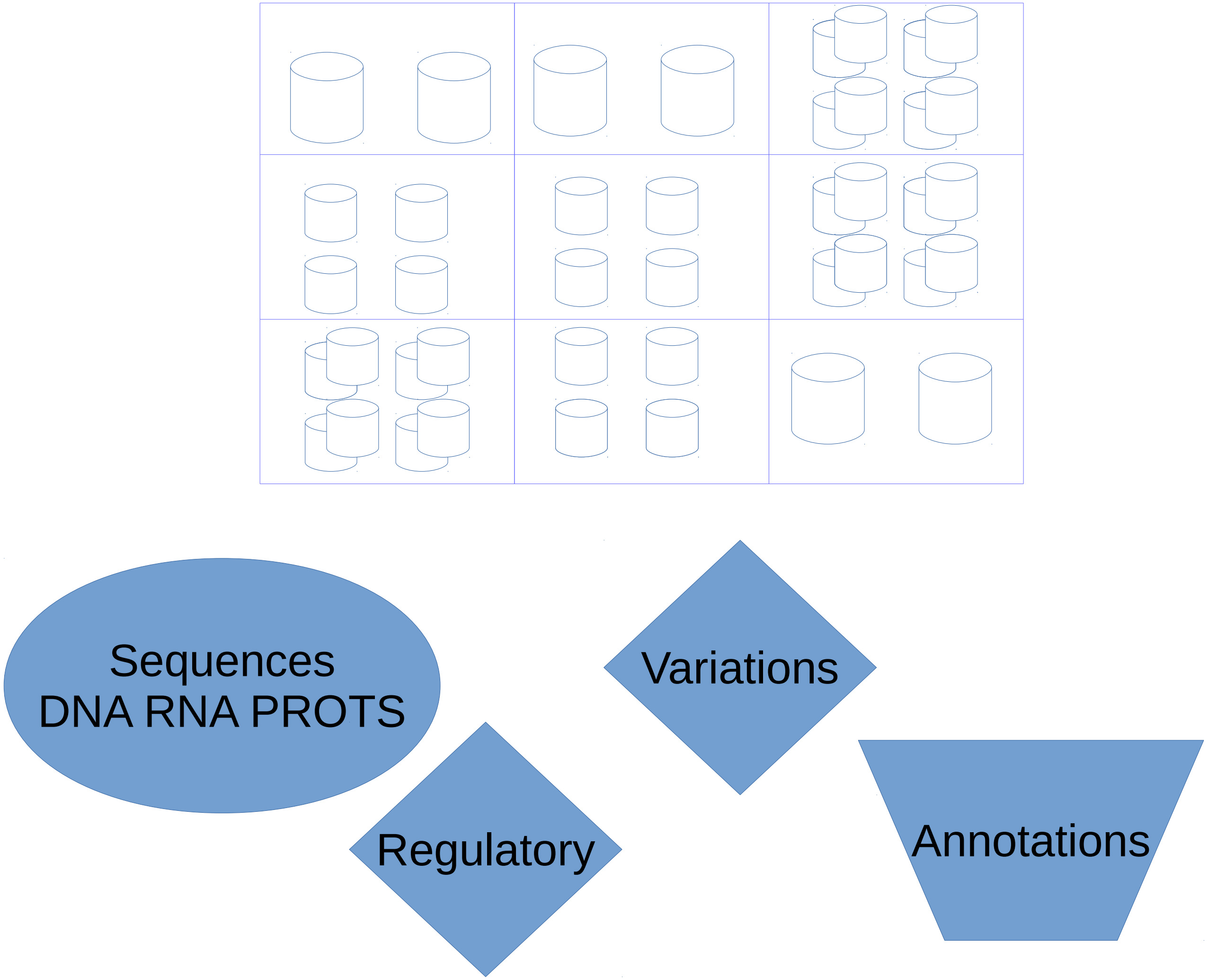
GFF

Results

Local (new) experimentation



Deposited (public) data

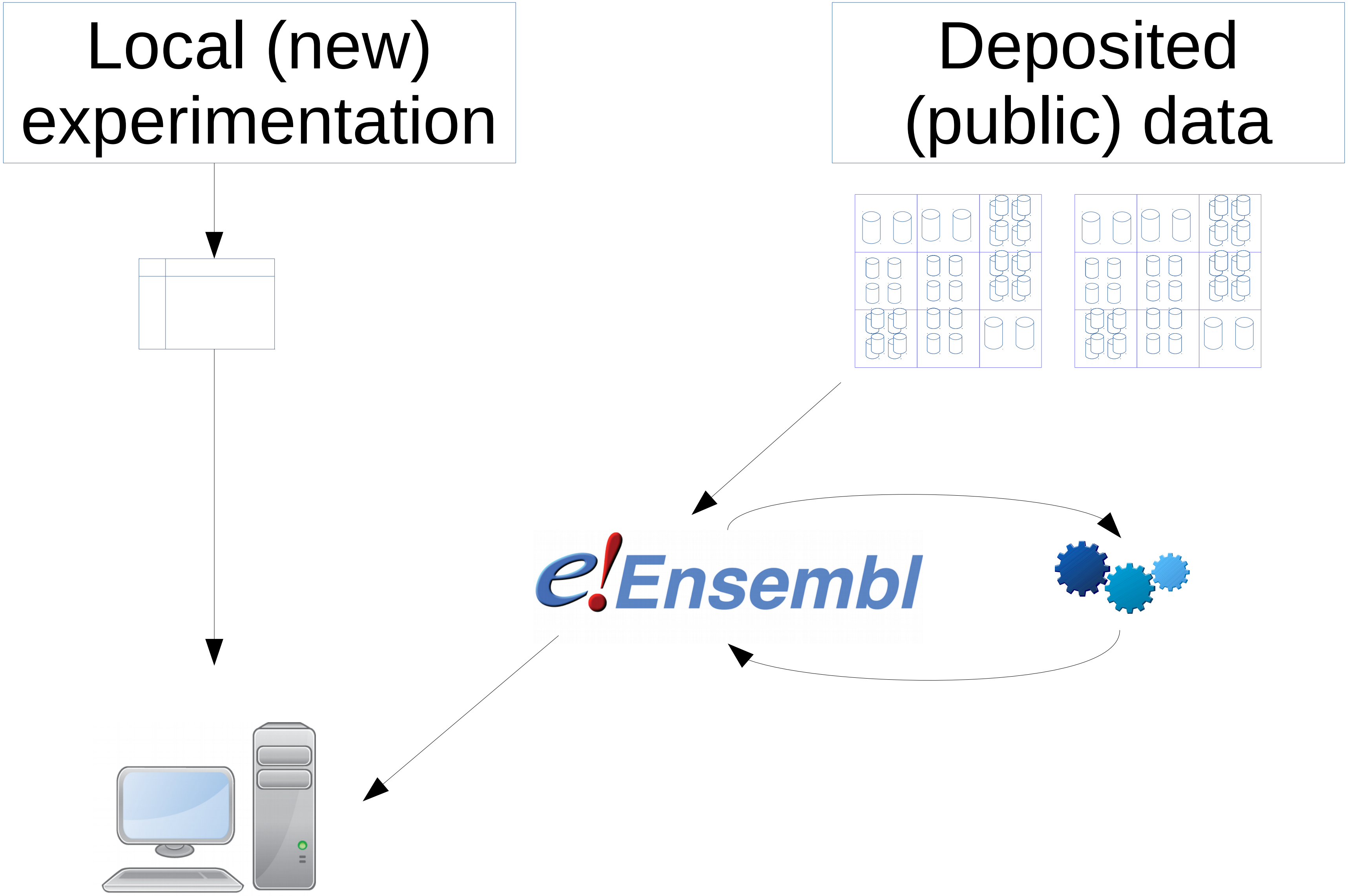


Sequences
DNA RNA PROTS

Variations

Regulatory

Annotations



Local (new) experimentation

Deposited (public) data

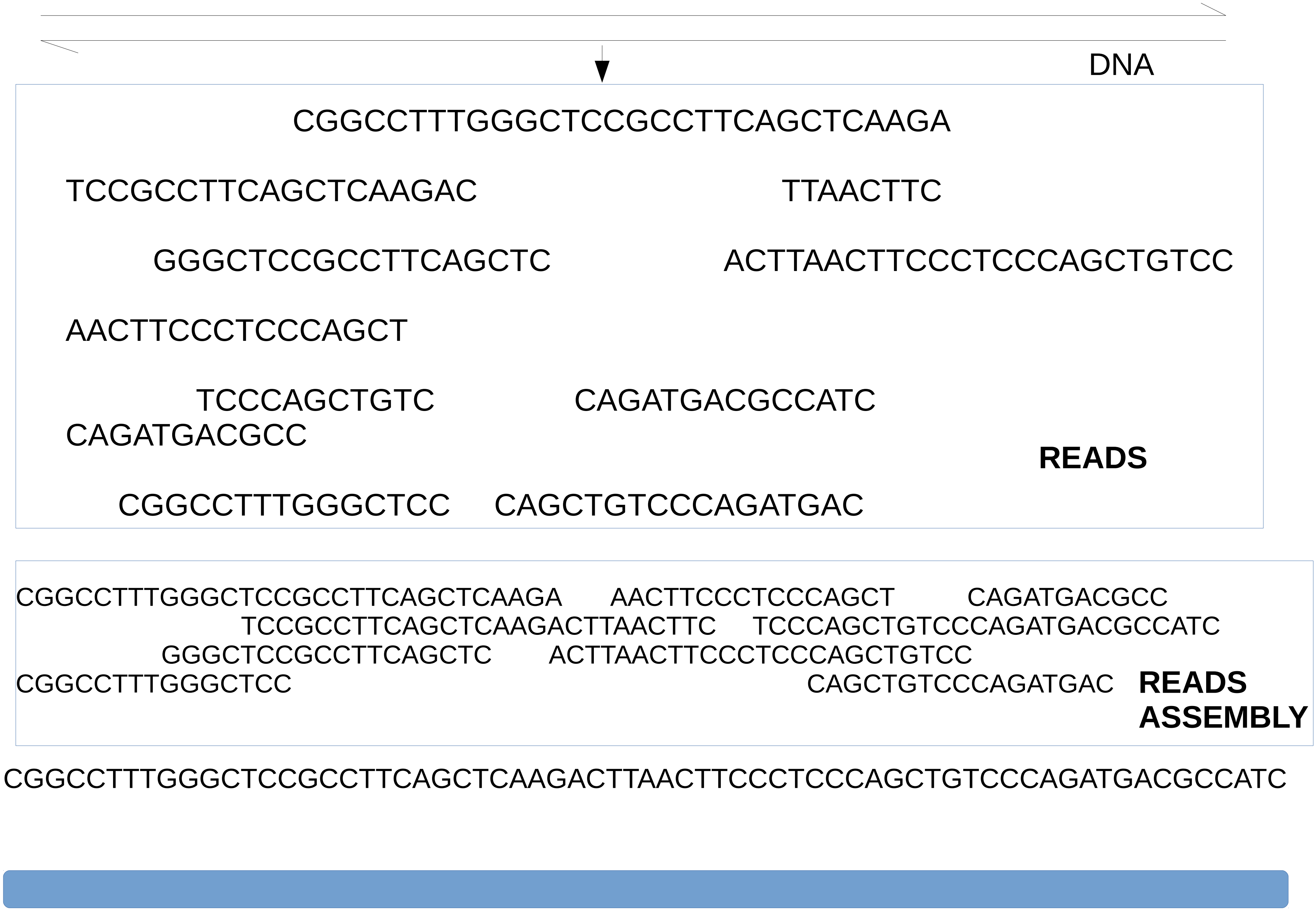
YOUR TASKS TODAY...

- Nivel genoma
  - Obtener el genoma de referencia de especie X (.fasta)
  - Obtener las anotaciones de la especie X (.gff3, .gtf)
  - Otros ficheros genómicos: variaciones, regulación… (.gff3 , .tsv)
- Nivel gen
  - Obtener la secuencia de un tránscrito T (.fa)
  - Obtener la secuencia de exones, etc. de un tránscrito T (.fa)
- Nivel intermedio (personalizado)
  - Obtener un *conjunto de anotaciones* interesantes de un *conjunto de genes* de interés (.tsv, .html…)
  - Obtener secuencias de un conjunto de genes de interés (.fasta)

What we have in Ensembl

- **Genomes**
- **Genes**
- **Transcripts**
- **Exons, introns, CDS…**
- Proteins
- Regulatory regions (promotors...)
- Variants (SNP, Indels...)
- **Functional annotations (Gene Ontology...)**
- Homology relationships
- ...more (depending on the species)

- Assembly of genomes:
  - Contigs
  - Scaffolds
  - Chromosomes



DNA

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA

TCCGCCTTCAGCTCAAGAC TTAACTTC

GGGCTCCGCCTTCAGCTC ACTTAACTTCCCTCCCAGCTGTCC

AACTTCCCTCCCAGCT

TCCCAGCTGTC CAGATGACGCCATC

CAGATGACGCC

READS

CGGCCTTTGGGCTCC CAGCTGTCCCAGATGAC

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA AACTTCCCTCCCAGCT CAGATGACGCC

TCCGCCTTCAGCTCAAGACTTAACTTC TCCCAGCTGTCCCAGATGACGCCATC

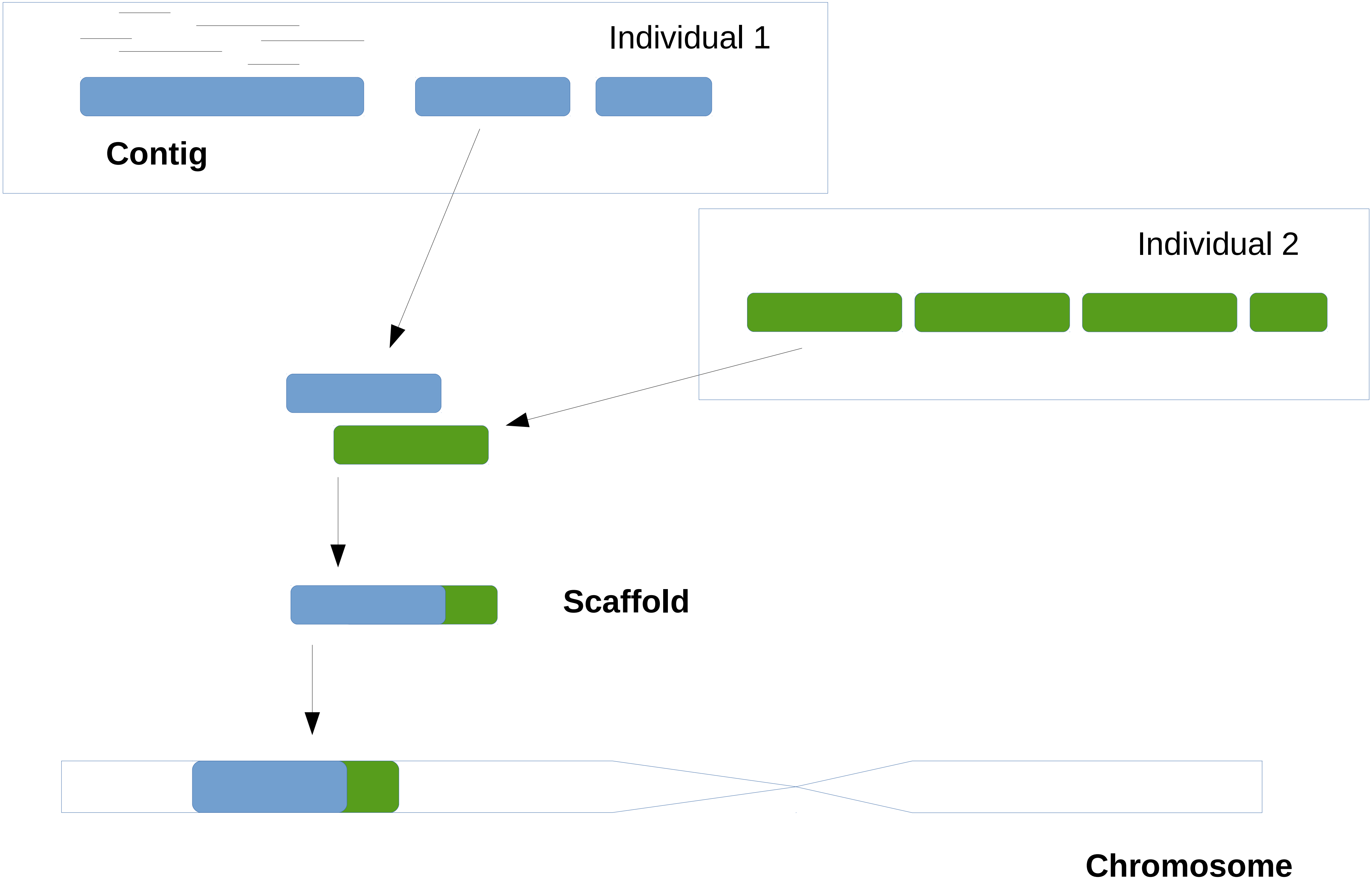
GGGCTCCGCCTTCAGCTC ACTTAACTTCCCTCCCAGCTGTCC

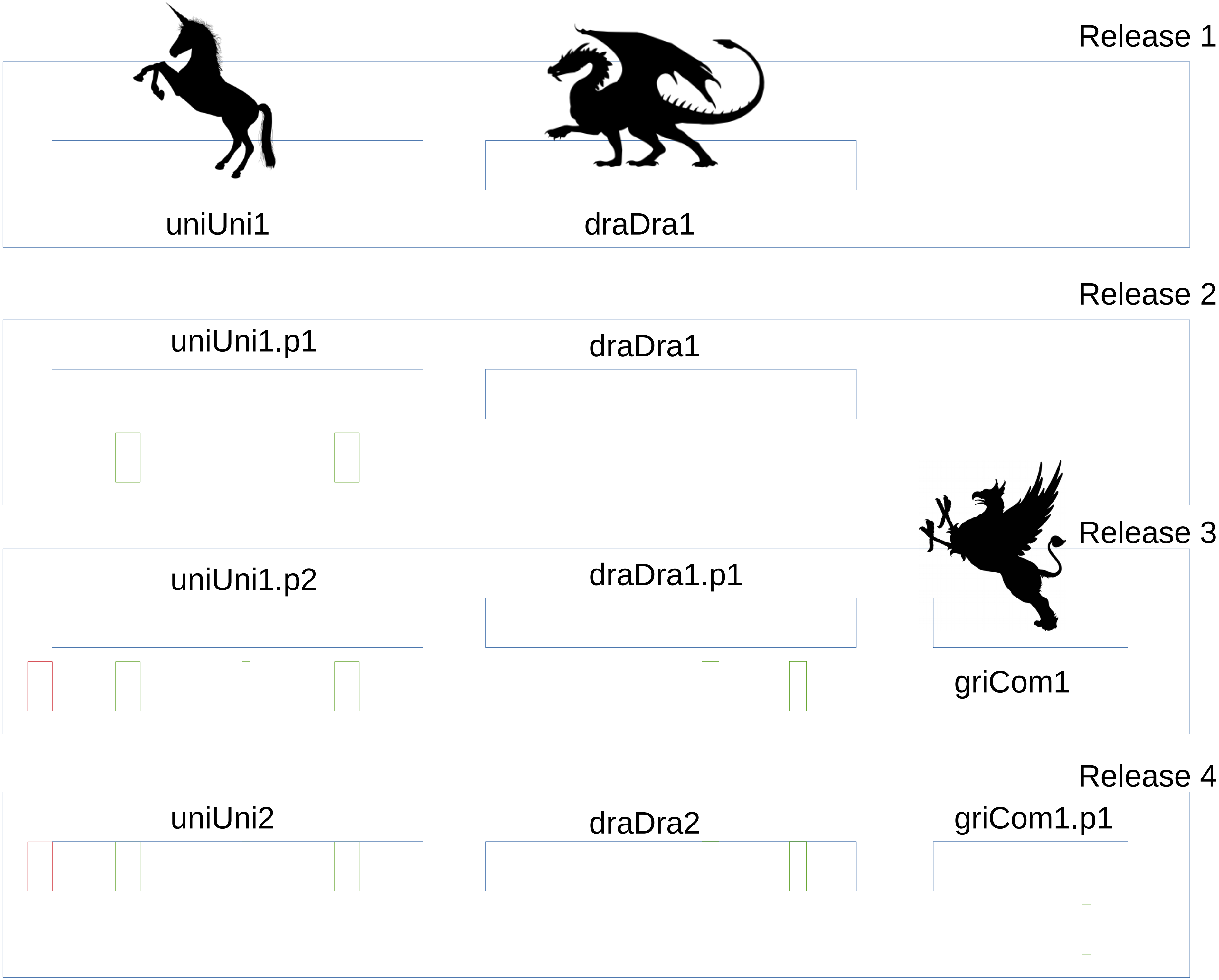
CGGCCTTTGGGCTCC CAGCTGTCCCAGATGAC

READS ASSEMBLY

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATC

“CONTIG”


Individual 1
Contig
Individual 2
Scaffold
Chromosome



Release 1

uniUni1

draDra1

Release 2

uniUni1.p1

draDra1

Release 3

uniUni1.p2

draDra1.p1

griCom1

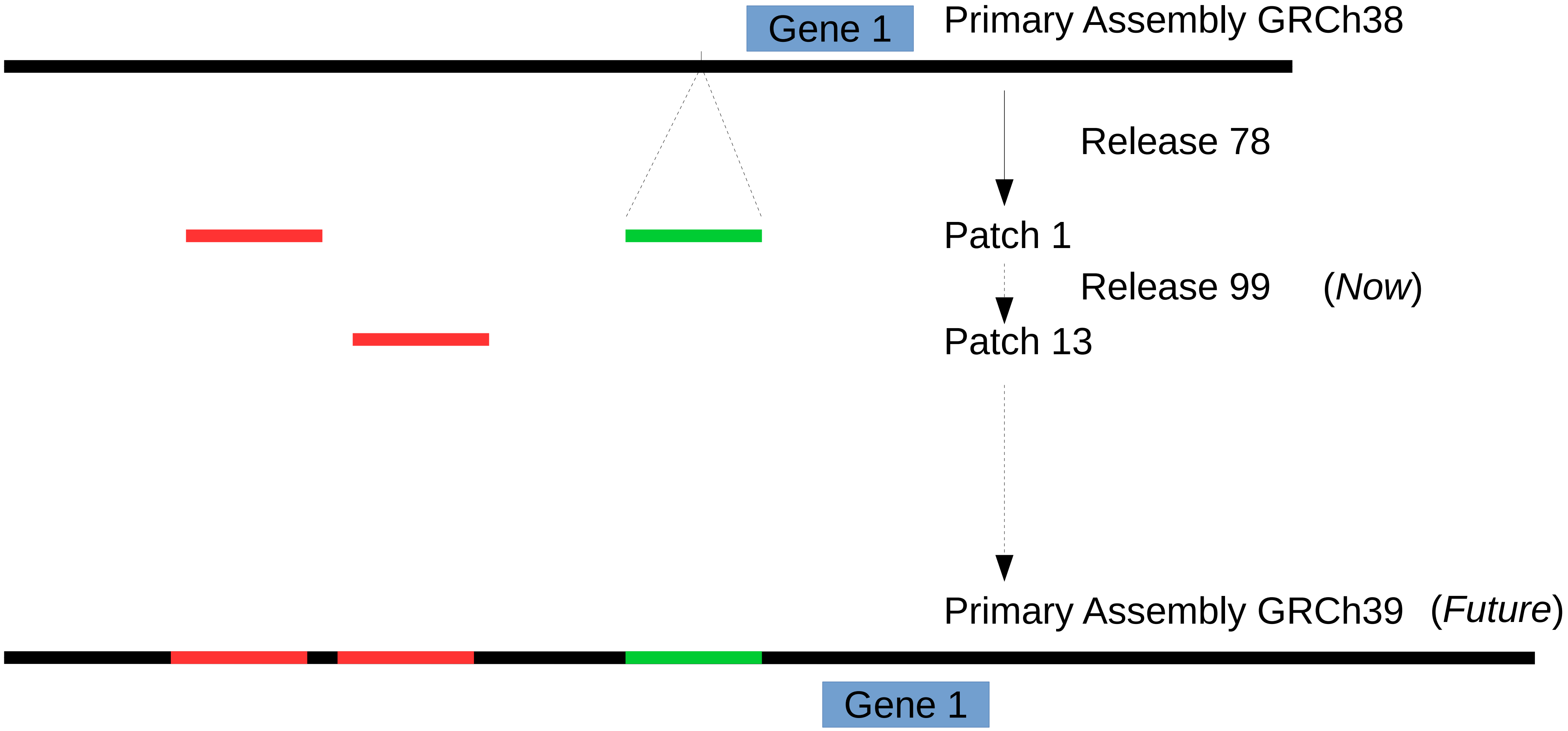
Release 4

uniUni2

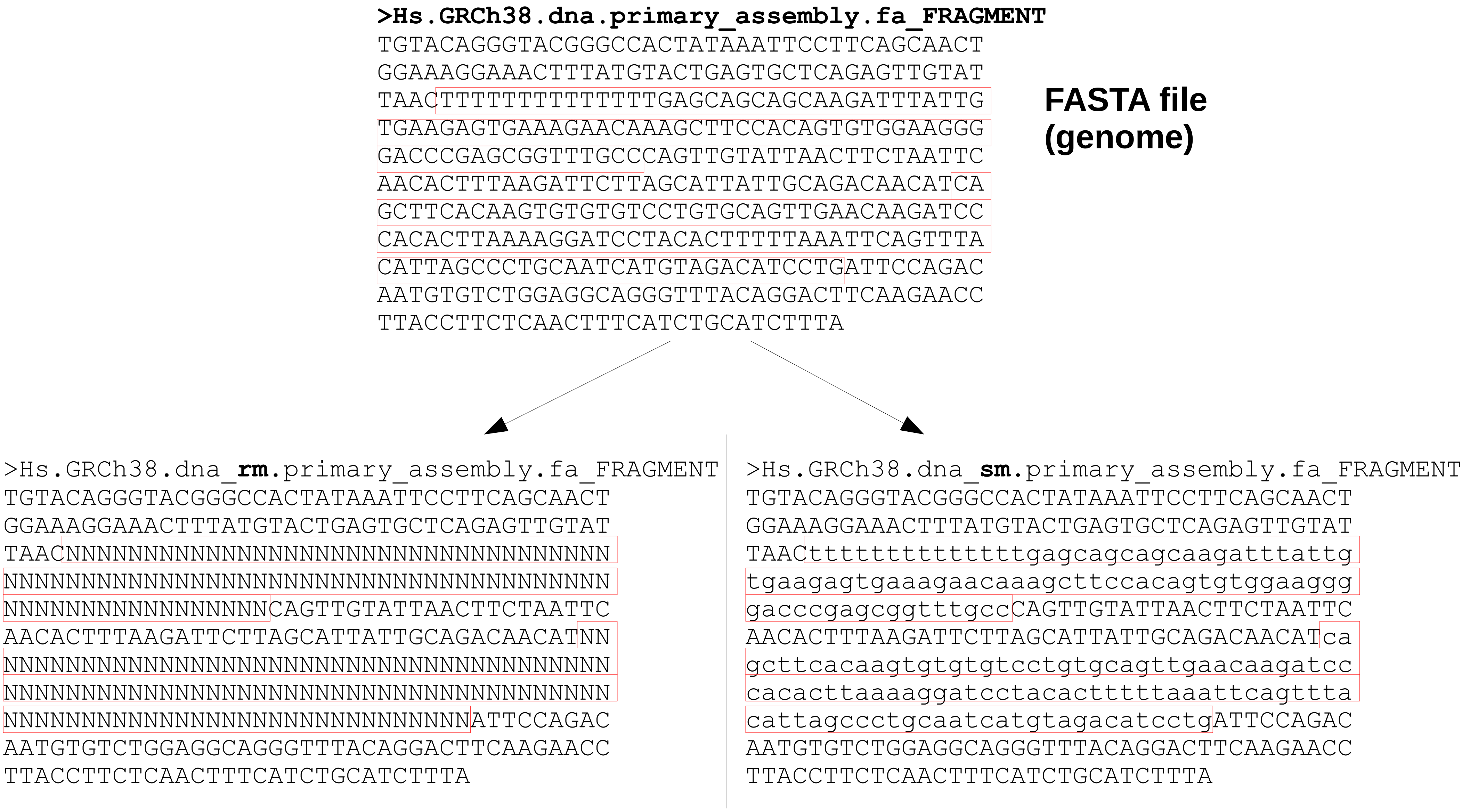
draDra2

griCom1.p1

Coordinates change from assembly to assembly version



Masking the zones of Low Complexity in the genome:
FASTA files “rm” and “sm” in Ensembl FTP



>Hs.GRCh38.dna.primary_assembly.fa_FRAGMENT

```
TGTACAGGGTACGGGCCACTATAAATTCCTTCAGCAACT
GGAAAGGAAACTTTATGTACTGAGTGCTCAGAGTTGTAT
TAACTTTTTTTTTTTTTTGAGCAGCAGCAAGATTTATTG
TGAAGAGTGAAAGAACAAAGCTTCCACAGTGTGGAAGGG
GACCCGAGCGGTTTGCCCAGTTGTATTAACTTCTAATTC
AACACTTTAAGATTCTTAGCATTATTGCAGACAACATCA
GCTTCACAAGTGTGTGTCCTGTGCAGTTGAACAAGATCC
CACACTTAAAAGGATCCTACACTTTTTAAATTCAGTTTA
CATTAGCCCTGCAATCATGTAGACATCCTGATTCCAGAC
AATGTGTCTGGAGGCAGGGTTTACAGGACTTCAAGAACC
TTACCTTCTCAACTTTCATCTGCATCTTTA
```

FASTA file (genome)

>Hs.GRCh38.dna_**rm**.primary_assembly.fa_FRAGMENT

```
TGTACAGGGTACGGGCCACTATAAATTCCTTCAGCAACT
GGAAAGGAAACTTTATGTACTGAGTGCTCAGAGTTGTAT
TAACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNCAGTTGTATTAACTTCTAATTC
AACACTTTAAGATTCTTAGCATTATTGCAGACAACATNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCCAGAC
AATGTGTCTGGAGGCAGGGTTTACAGGACTTCAAGAACC
TTACCTTCTCAACTTTCATCTGCATCTTTA
```

Hard masked (rm) sequence

>Hs.GRCh38.dna_**sm**.primary_assembly.fa_FRAGMENT

```
TGTACAGGGTACGGGCCACTATAAATTCCTTCAGCAACT
GGAAAGGAAACTTTATGTACTGAGTGCTCAGAGTTGTAT
TAACttttttttttttttgagcagcagcaagatttattg
tgaagagtgaaagaacaaagcttccacagtgtggaaggg
gacccgagcggtttgccCAGTTGTATTAACTTCTAATTC
AACACTTTAAGATTCTTAGCATTATTGCAGACAACATca
gcttcacaagtgtgtgtcctgtgcagttgaacaagatcc
cacacttaaaaggatcctacactttttaaattcagttta
cattagccctgcaatcatgtagacatcctgATTCCAGAC
AATGTGTCTGGAGGCAGGGTTTACAGGACTTCAAGAACC
TTACCTTCTCAACTTTCATCTGCATCTTTA
```

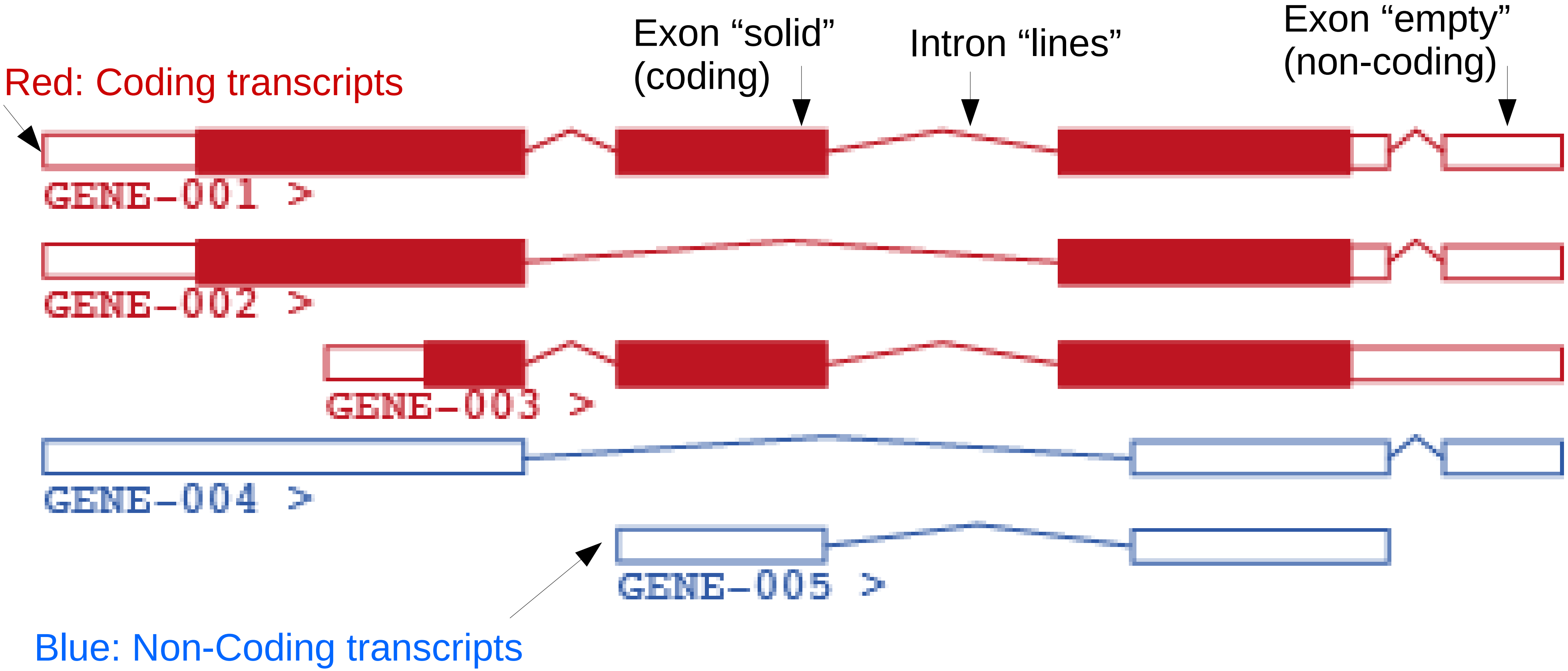
Soft masked (sm) sequence

Correspondance between FASTA reference and GFF3 (or GTF) annotations file

```
zcat <your-path>/Homo_sapiens.GRCh38.99.chr.gff3.gz | less
```

```
.
.
##sequence-region   9 1 138394717
##sequence-region   MT 1 16569
##sequence-region   X 1 156040895
##sequence-region   Y 2781480 56887902
#!genome-build Ensembl GRCh38.p13
#!genome-version GRCh38
#!genome-date 2013-12
.
.
```

TRANSCRIPTS REPRESENTATION IN ENSEMBL



Red: Coding transcripts

Exon “solid” (coding)

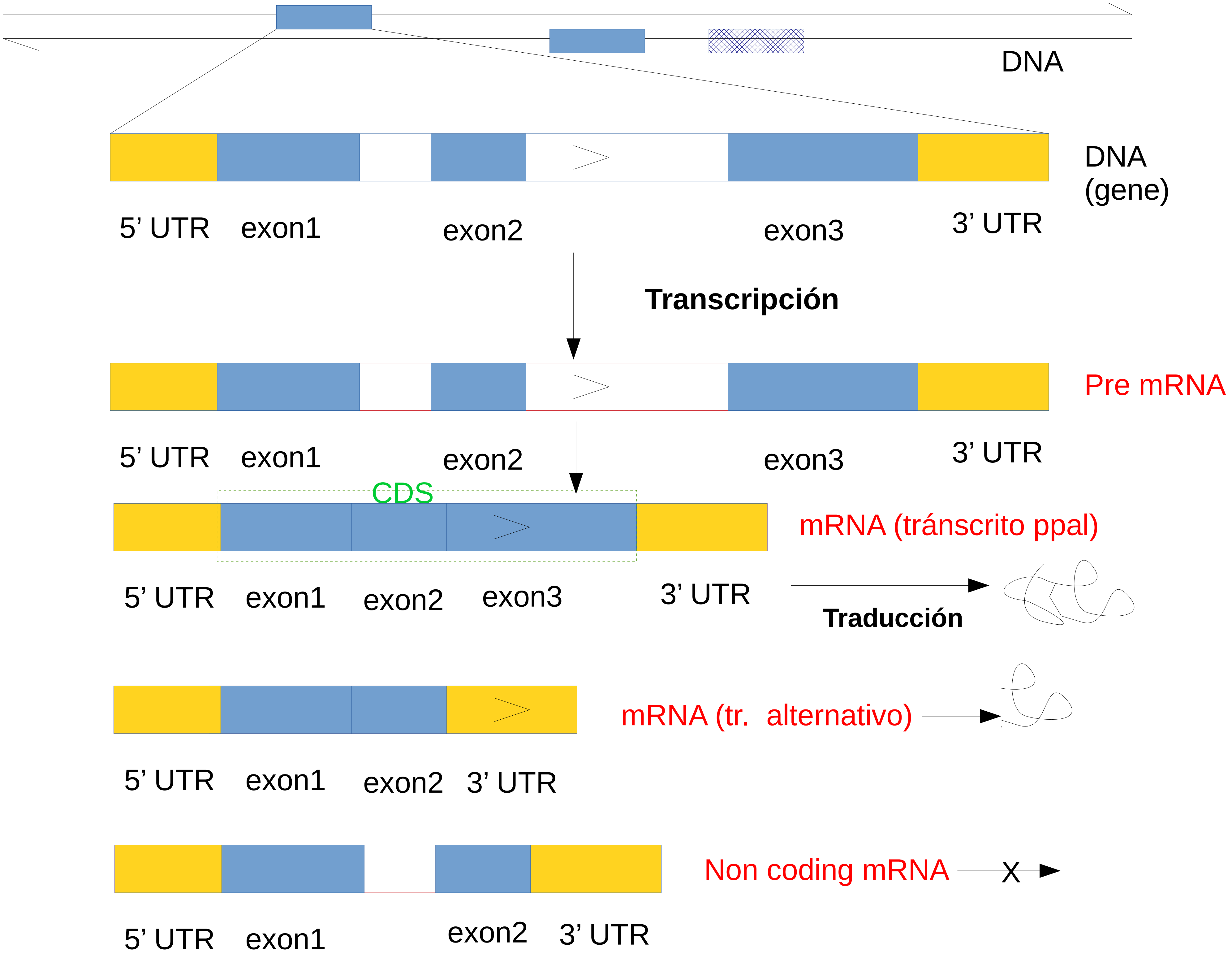
Intron “lines”

Exon “empty” (non-coding)

Blue: Non-Coding transcripts

Choosing the Transcript to use (Criteria)

- 1. **MANE** Select: Complete transcript (coding and UTR) matches RefSeq and it has been selected by Ensembl and RefSeq as the most biologicallyrelevant transcript
- 2. **APPRIS** principal isoform: The major isoform(s) from combining protein structural information, functionally important residues and evidence from cross-species alignments.
- 3. **GENCODE** Basic: Only the “complete” transcripts (where a gene has complete transcripts)
- 4. **Transcript support level**: Scored 1-5 for quality, where 1 is the best
- 5. **CCDS**: Matching coding sequence with RefSeq
- 6. **Golden transcripts**: Matching annotation from Ensembl and Havana annotation



DNA

DNA (gene)

5’ UTR exon1 exon2 exon3 3’ UTR

Transcripción

Pre mRNA

5’ UTR exon1 exon2 exon3 3’ UTR

CDS

mRNA (tránscrito ppal)

5’ UTR exon1 exon2 exon3 3’ UTR

Traducción

mRNA (tr. alternativo)

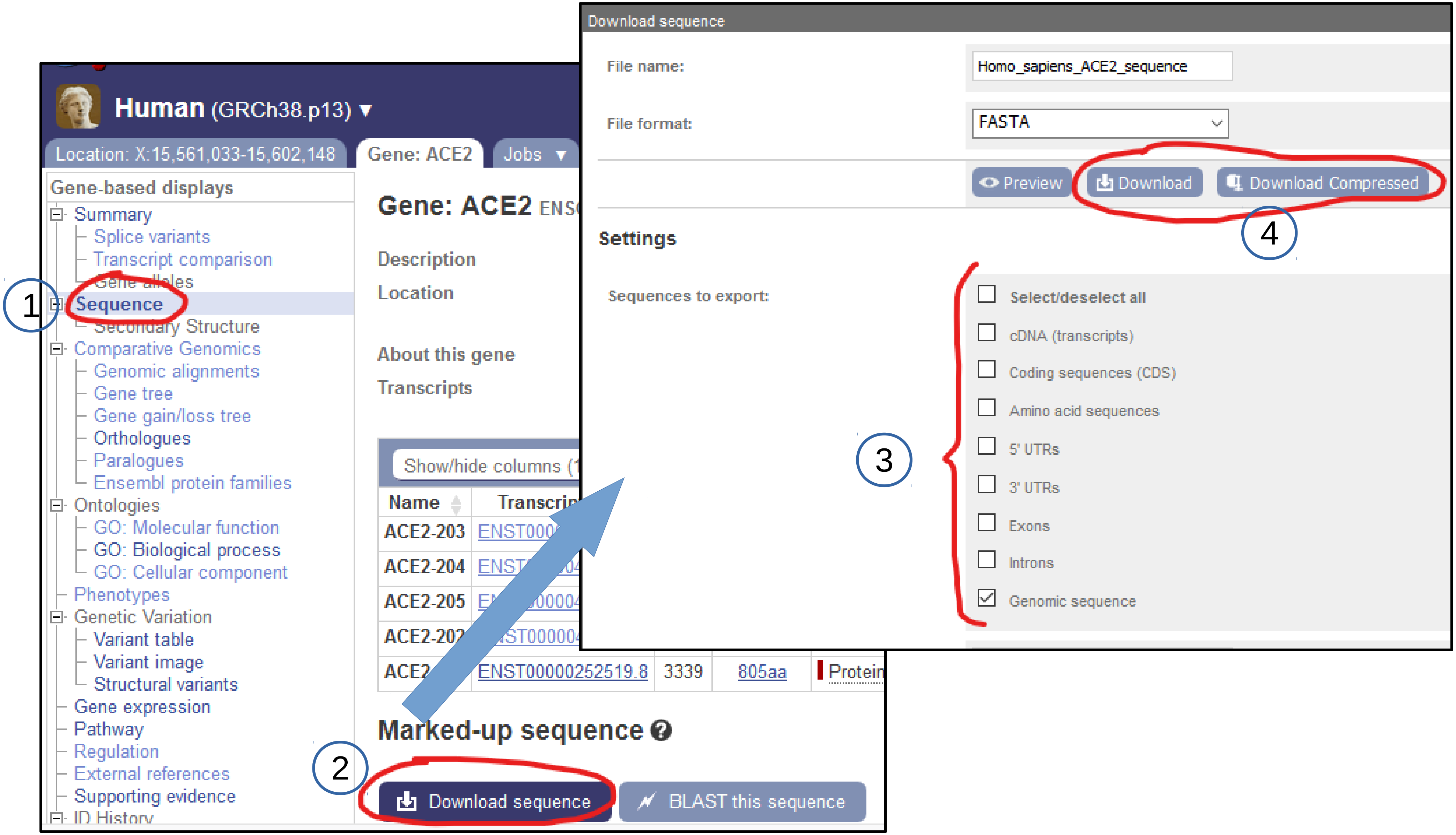
5’ UTR exon1 exon2 3’ UTR

Non coding mRNA

X

5’ UTR exon1 exon2 3’ UTR

Downloading a gene sequence in Ensembl Browser



Gene: ACE2 ENS

Description

Location

About this gene

Transcripts

Name	Transcrip			
ACE2-203	ENST000			
ACE2-204	ENST			
ACE2-205	E 0000			
ACE2-202	ST00000			
ACE2	ENST00000252519.8	3339	805aa	Protein

Marked-up sequence

File name: Homo_sapiens_ACE2_sequence

File format: FASTA

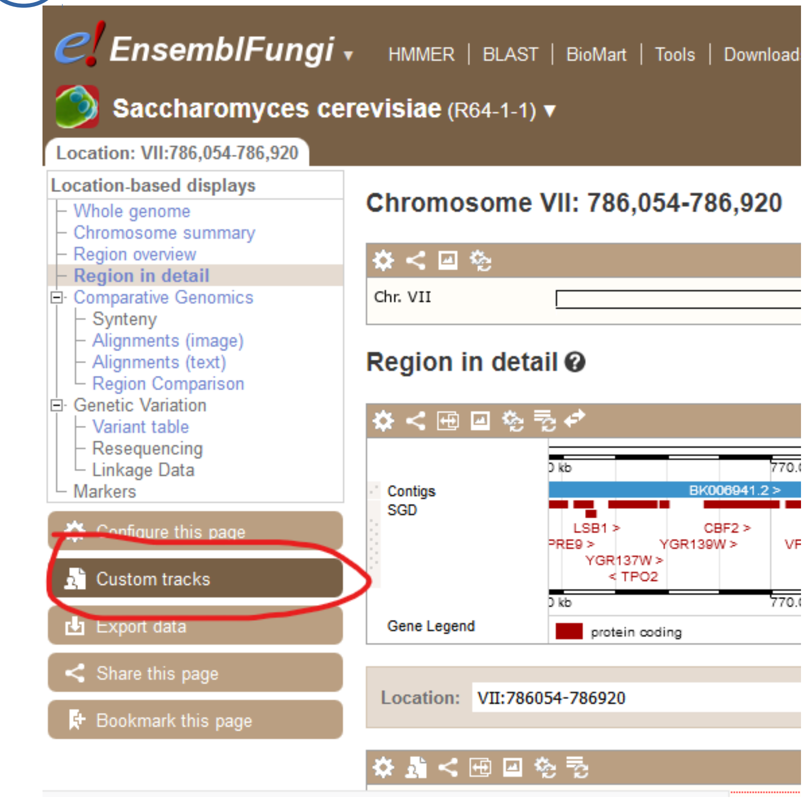
Settings

Sequences to export:

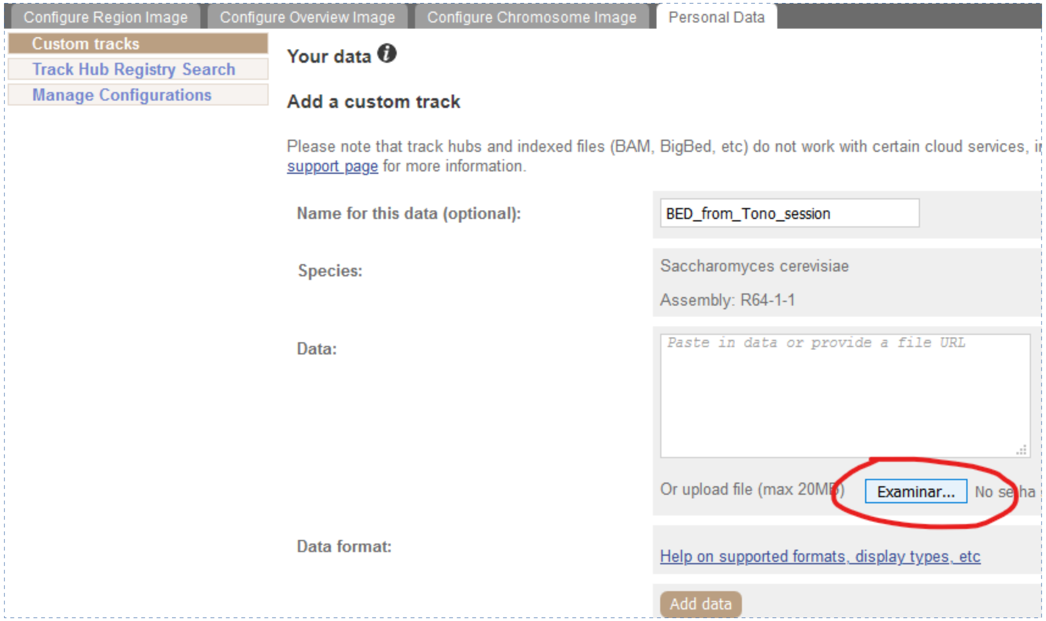
- □ Select/deselect all
- □ cDNA (transcripts)
- □ Coding sequences (CDS)
- □ Amino acid sequences
- □ 5' UTRs
- □ 3' UTRs
- □ Exons
- □ Introns
- ☑ Genomic sequence

Loading a Custom Track in Ensembl Browser (I)

1



2



Please note that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services, i support page for more information.

Name for this data (optional): BED_from_Tono_session

Species: Saccharomyces cerevisiae

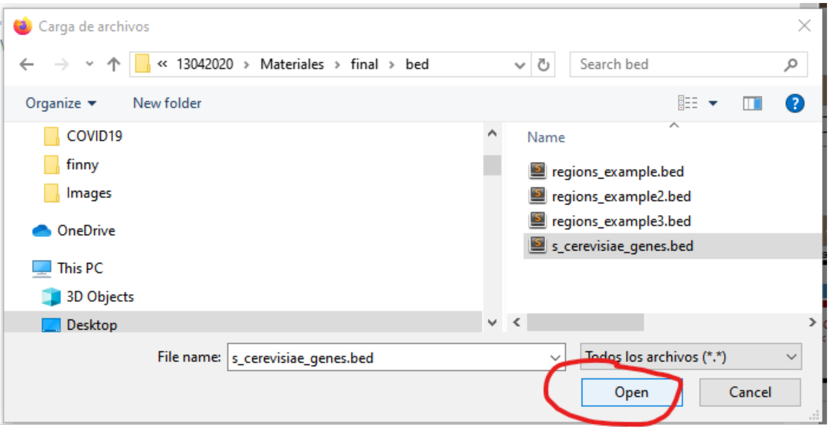
Assembly: R64-1-1

Data: Paste in data or provide a file URL

Or upload file (max 20MB) Examinar... No se ha

Data format: Help on supported formats, display types, etc

3



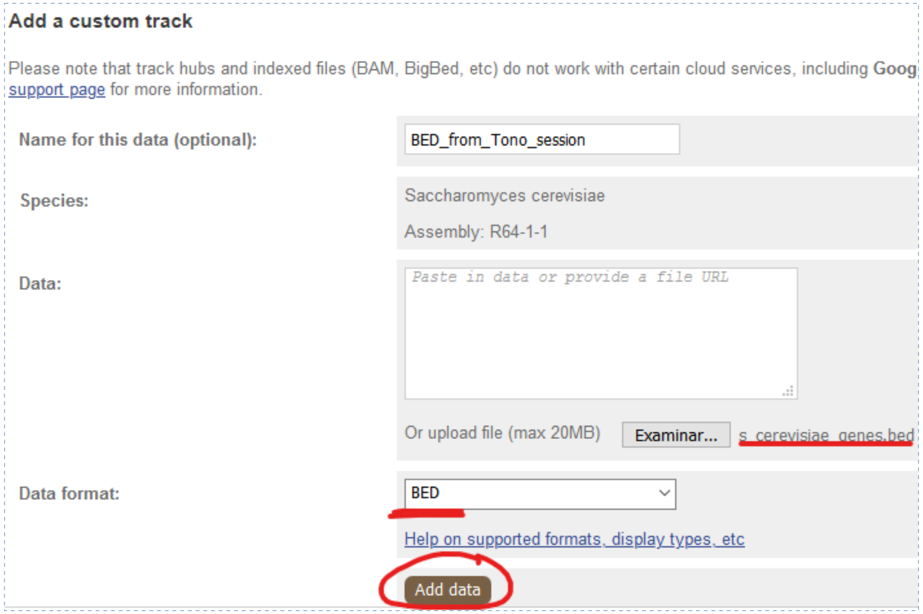
File name: s_cerevisiae_genes.bed

Todos los archivos (*.*)

Loading a Custom Track in Ensembl Browser (II)

3

4



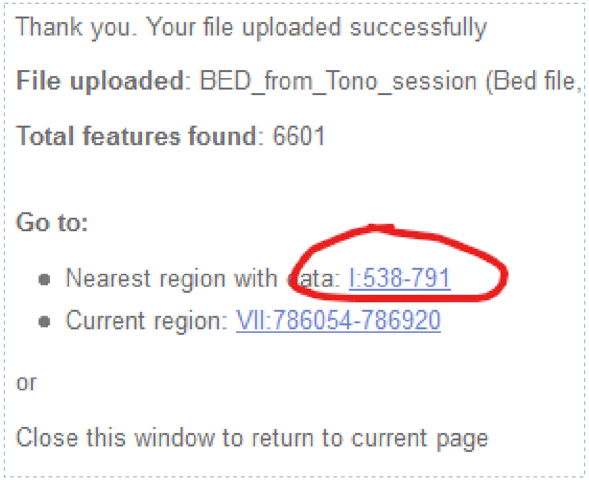
Add a custom track

Please note that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services, including **Goog**
support page for more information.

Name for this data (optional):	BED_from_Tono_session
Species:	Saccharomyces cerevisiae Assembly: R64-1-1
Data:	Paste in data or provide a file URL Or upload file (max 20MB) Examinar... s_cerevisiae_genes.bed
Data format:	BED Help on supported formats, display types, etc

Add data

5



Thank you. Your file uploaded successfully

File uploaded: BED_from_Tono_session (Bed file,

Total features found: 6601

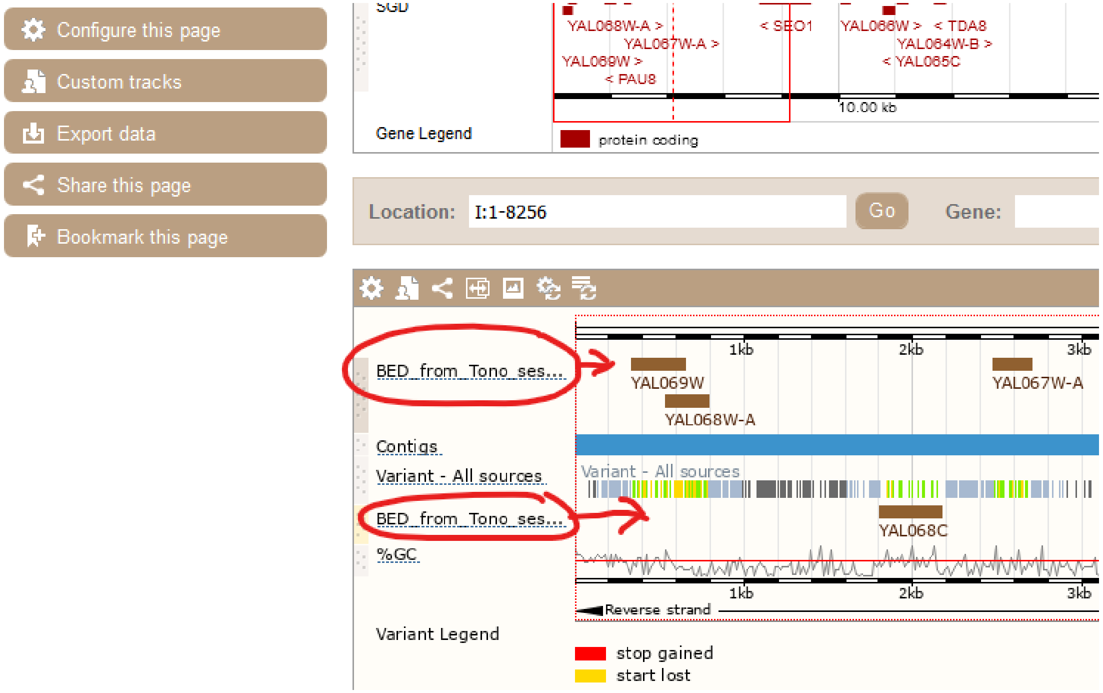
Go to:

- Nearest region with data: I:538-791
- Current region: VII:786054-786920

or

Close this window to return to current page

6



Location: I:1-8256 Go Gene:

Take home recommendations (I):

1. You will be sure of the version of the assembly (FASTA) to use you use or you are given (GRCh38? 37? species? strain?). Coordinates don’t match between assemblies…

2. You will match or check the matching between the FASTA file, the GFF3/GTF. (Note: Do FASTA and GFF share the same number and name of chromosomes?)

3. You will match or check the matching between GFF3/GTF and BAM file, VCF files…

4. BioMart: choose the design of the table beforehand. Remember: the features order you select is the columns order you get.

Take home recommendations (II):

5. Choose a limited set of attibutes for your BioMart table. Too many attributes, less understable. Study beforehand what is it needed (avoid “just in case”).

6. But…don’t forget to include the IDs of genes, transcripts, variants, GO terms, etc. present in the table (Names/Descriptions are not enough).

7. Think beforehand the best method to retrieve the data. If you need to deal with a lot of genes/variations or it is not defined, download the entire genomic files (i.e. **FTP**). If you need a short list of genes (less than 500 for instance) and you have a clear idea of the features you need, **BioMart** is your tool. For a very short list of genes or regions in-deep study, **Ensembl browser** is your tool.