IMPERIAL

wellcome connecting science

Guided computational practical:

Genes and Genomes

Masue Marbiah, PhD

07/10/2024

Overview

1. Reference Genomes
2. Genome Annotation
3. Genome Browsers
4. Gene Structure
5. CRISPR Design Process and Tools
6. Other sgRNA Design Tools

Reference Genomes

Reference genomes
Genome Reference Consortium

Consortium goals:

- Maintain reference genomes
  - Correct regions in the genome that are currently misrepresented
    - To close as many genomic gaps as possible
      - Scientific community can report loci in need of review

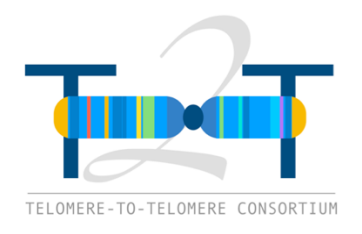
Genome Reference Consortium releases include major and minor updates (patches).

For example, Genome Reference Consortium human 38 patch 14 (**GRCh38.p14**)

Equivalent gene builds for **mouse**, rat, zebrafish and chicken are also available.

https://www.ncbi.nlm.nih.gov/grc

Reference genomes
Telomere-to-Telomere Consortium



Telomere-to-Telomere (T-2-T) Consortium **completely sequenced the entire human genome** (no gaps) in 2022.

High quality telomere-to-telomere assemblies from diploid human genomes.

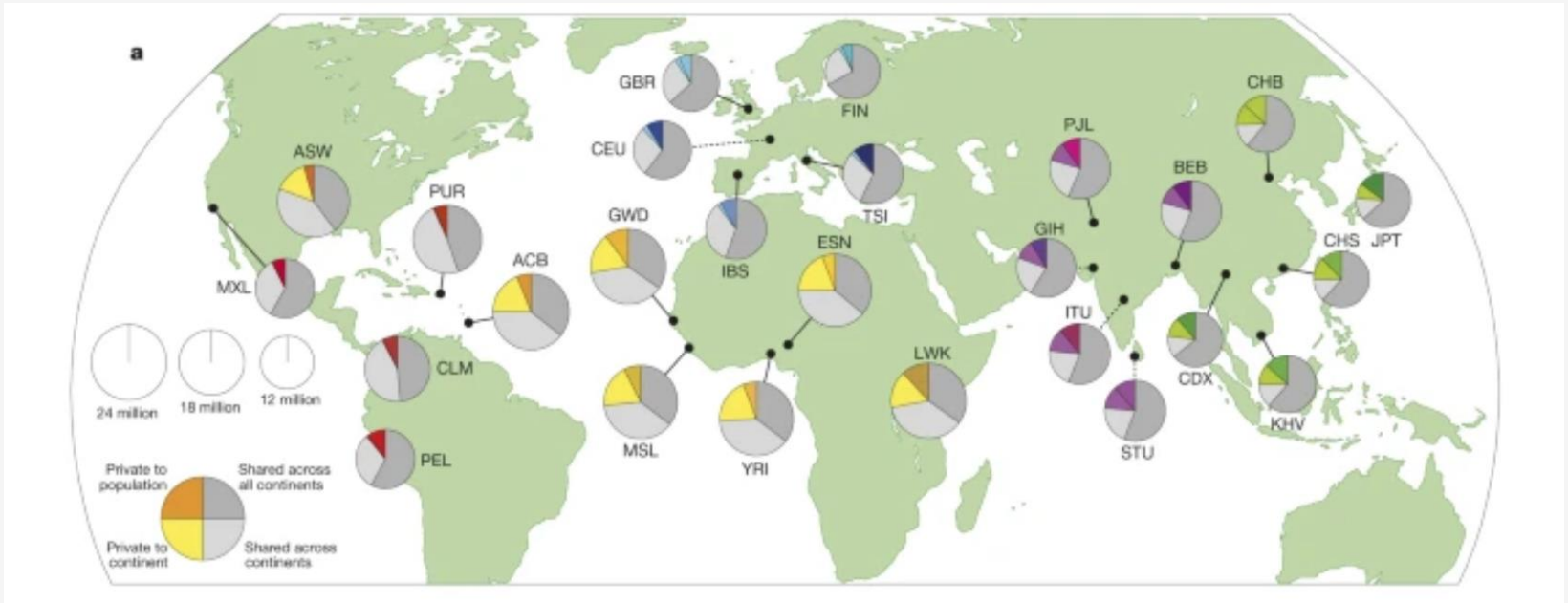
Y-chromosome was fully sequenced for the first time in 2023 providing genetic insight into how it influences male development and health.

Contribute towards the **Human Pangenome Project**, to generate 350 high quality gapless human genome sequences from ancestrally diverse people.

Reference genomes
The 1000 Genome Project

The goal of the 1000 Genomes Project was to find common genetic variants with frequencies of at least 1% in the populations studied.

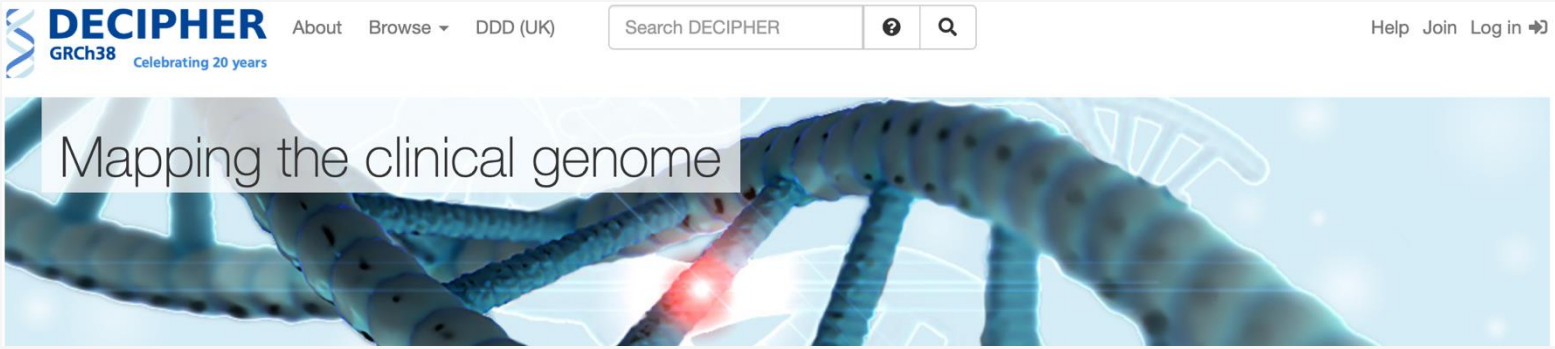
Genomes of 2,504 individuals from **26 global populations** were studied. The consortium identified **>88 million variants** (SNPs, InDels and structural variants). Variants molecular and disease phenotypes.



Nature 526, 68–74 (2015).

Reference genomes

DECIPHER



DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources) is an interactive web-based database which incorporates a suite of tools designed to aid the **interpretation of genomic variants found in rare diseases.**

Data from **over 270 centres** reporting on **>36,000 cases** are publicly available with patient consent.

This dataset includes the DDD (Deciphering Developmental Disorders) study, which has recruited ~ **14,000 children with severe undiagnosed developmental disorders,** and their parents from around the UK and Ireland.

Reference genomes
Mouse Genome Project

The Mouse Genomes Project is an ongoing effort to **catalogue all forms of genetic variation between the common laboratory mouse strains** and to produce high quality annotated reference genomes for the key strains.

>50 strains including BALB/c and C57BL/6

Project focuses on:

- Short-sequencing reads of many lab strains to identify variants relative to C57BL/6
  - Generate reference genomes and strain specific annotations of the commonly used strains.

Genome Annotation

Gene Annotation
Databases

GENCODE

High quality reference gene annotation and experimental validation for human and mouse genomes, based on manual annotation that includes Ensembl annotation.

Data is primarily based on mRNA, Expressed Sequence Tags (EST) and protein evidence that is aligned to a reference genome.

GENCODE Accessions are:

	Human	Mouse
Gene (G)	ENSG00000012345	ENSMUSG00000012345
Transcript (T)	ENST00000012345	ENSMUST00000012345
Exon (E)	ENSE00000012345	ENSMUSE00000012345
Protein (P)	ENSP00000012345	ENSMUSP00000012345

Gene Annotation
Databases

RefSeq (transcript-centric)

RefSeq transcript and protein records are generated by several processes including:

- Computation (mRNA XM_123456; Protein XP_123456; RNA XR_123456
  - Manual curation (mRNA NM_123456; Protein NP_123456; RNA NR_123456)

Annotation includes 392 eukaryotic genomes.

Ensembl (genome-centric)

Computational alignment-based annotation, including cDNAs, proteins and RNA-seq reads for over 70 different vertebrate species.

MANE (Matched Annotation between NCBI and Ensembl)

Collaborative project that aims to provide a minimal set of **matching RefSeq and Ensembl transcripts** of human protein-coding genes.

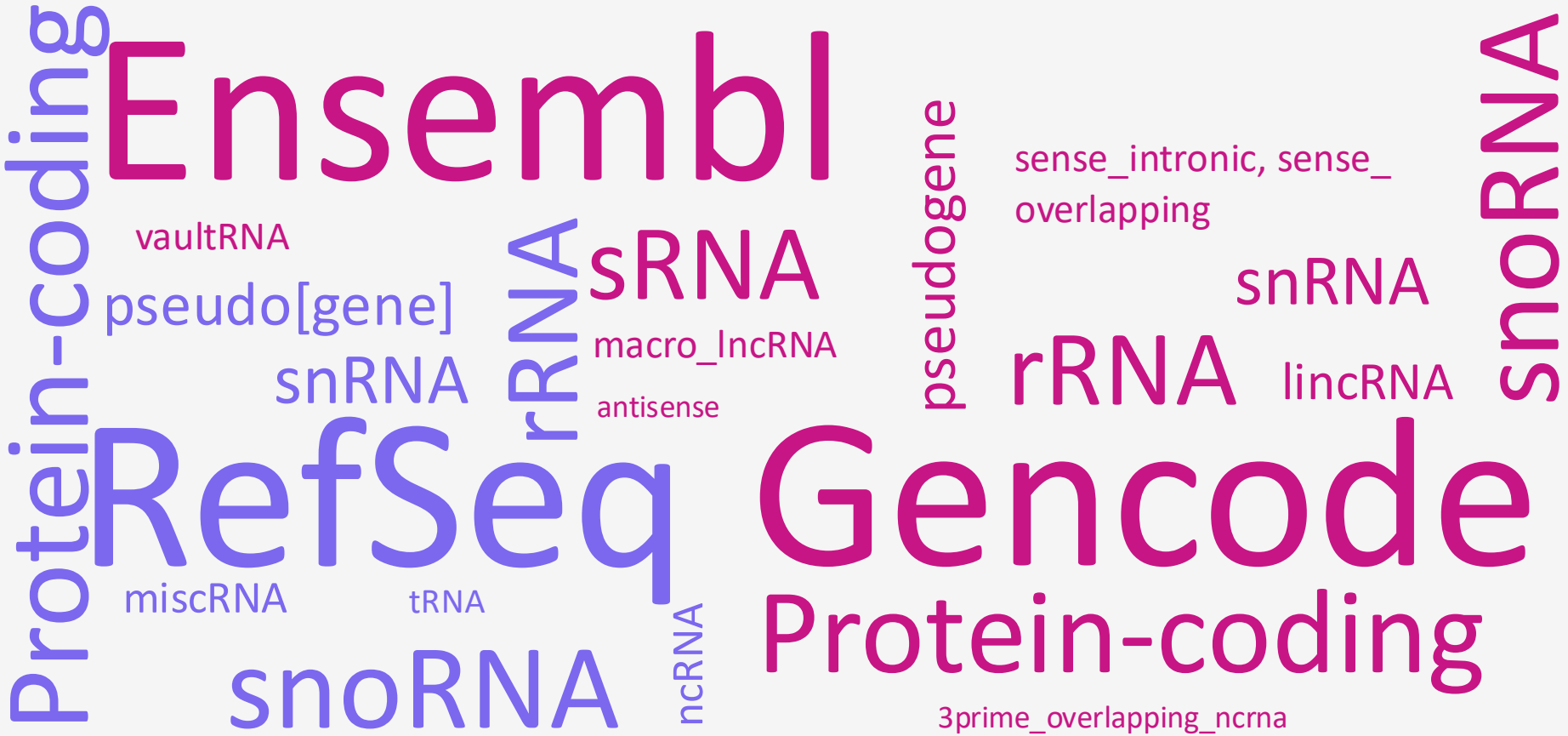
MANE displays:

- MANE Select: One high quality representative transcript per protein-coding gene. Supported empirically.
  - MANE Plus Clinical: Additional transcripts with relevant “Pathogenic (P)” or ”Likely pathogenic (LP)” variants.
    - MANE: All other matched transcripts.

Gene Annotation

Biotypes

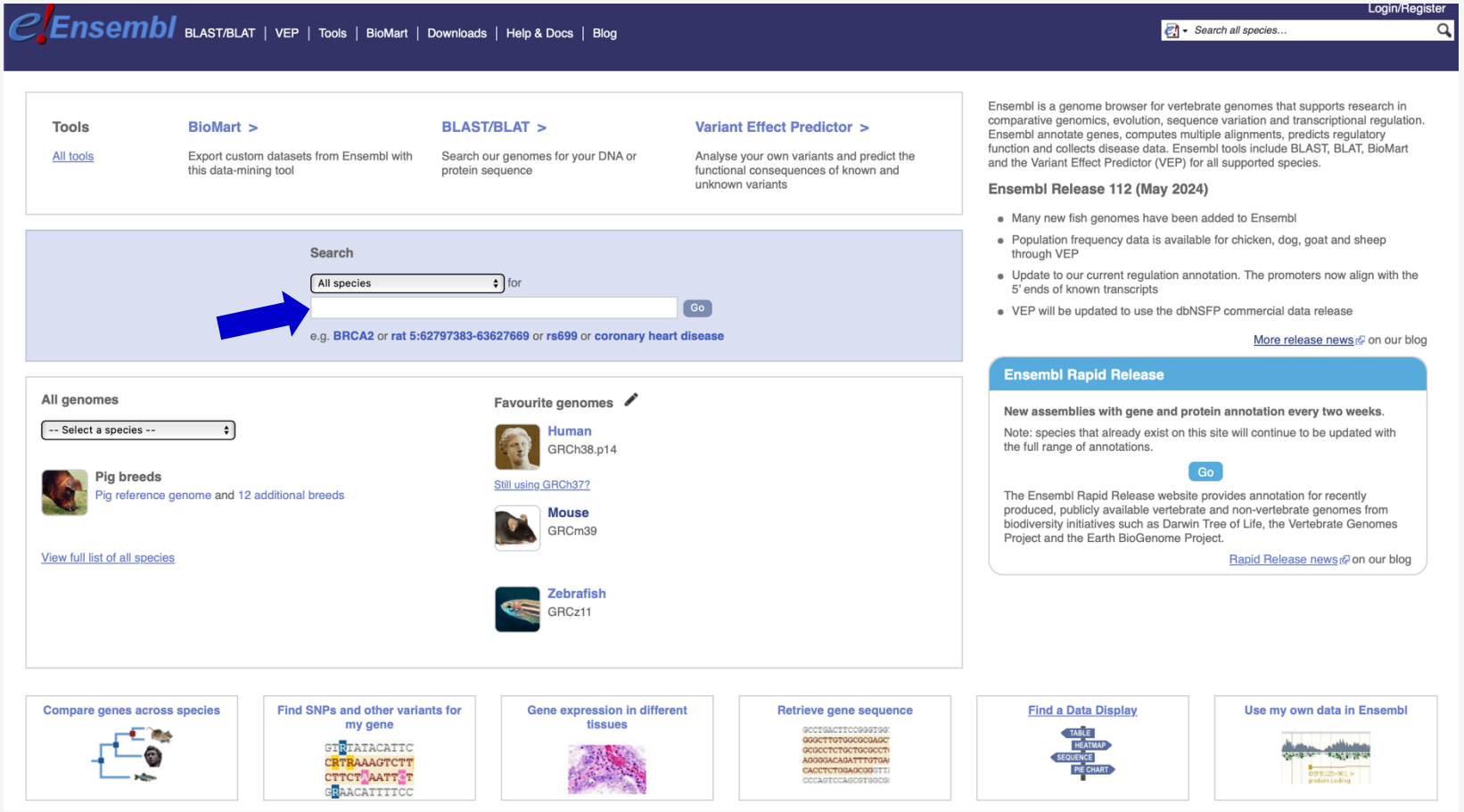
terms used to classify genes or transcripts.



Genome Browsers

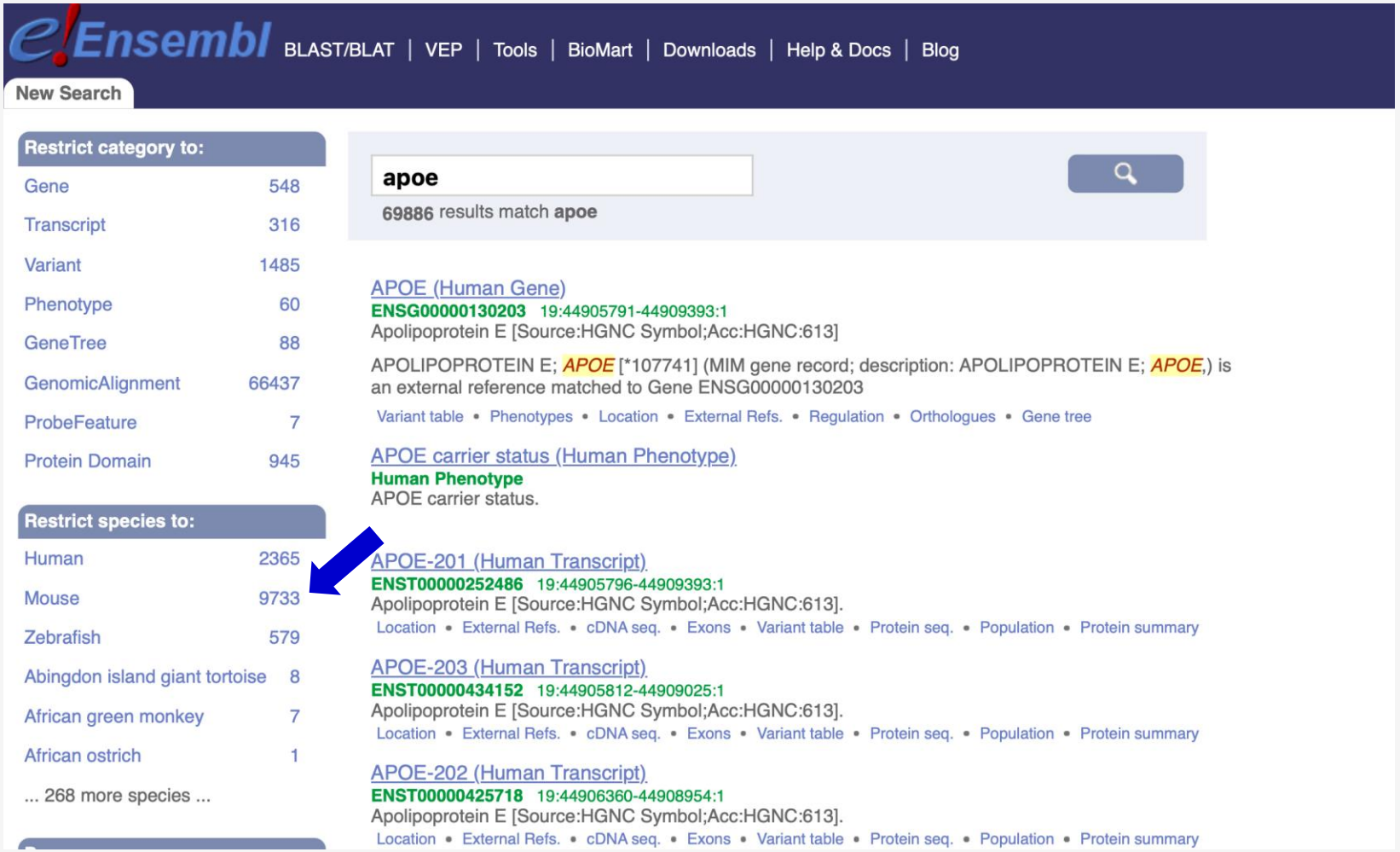
Genome Browsers
Ensembl example

Go to http://www.ensembl.org and search for the Apoe gene.



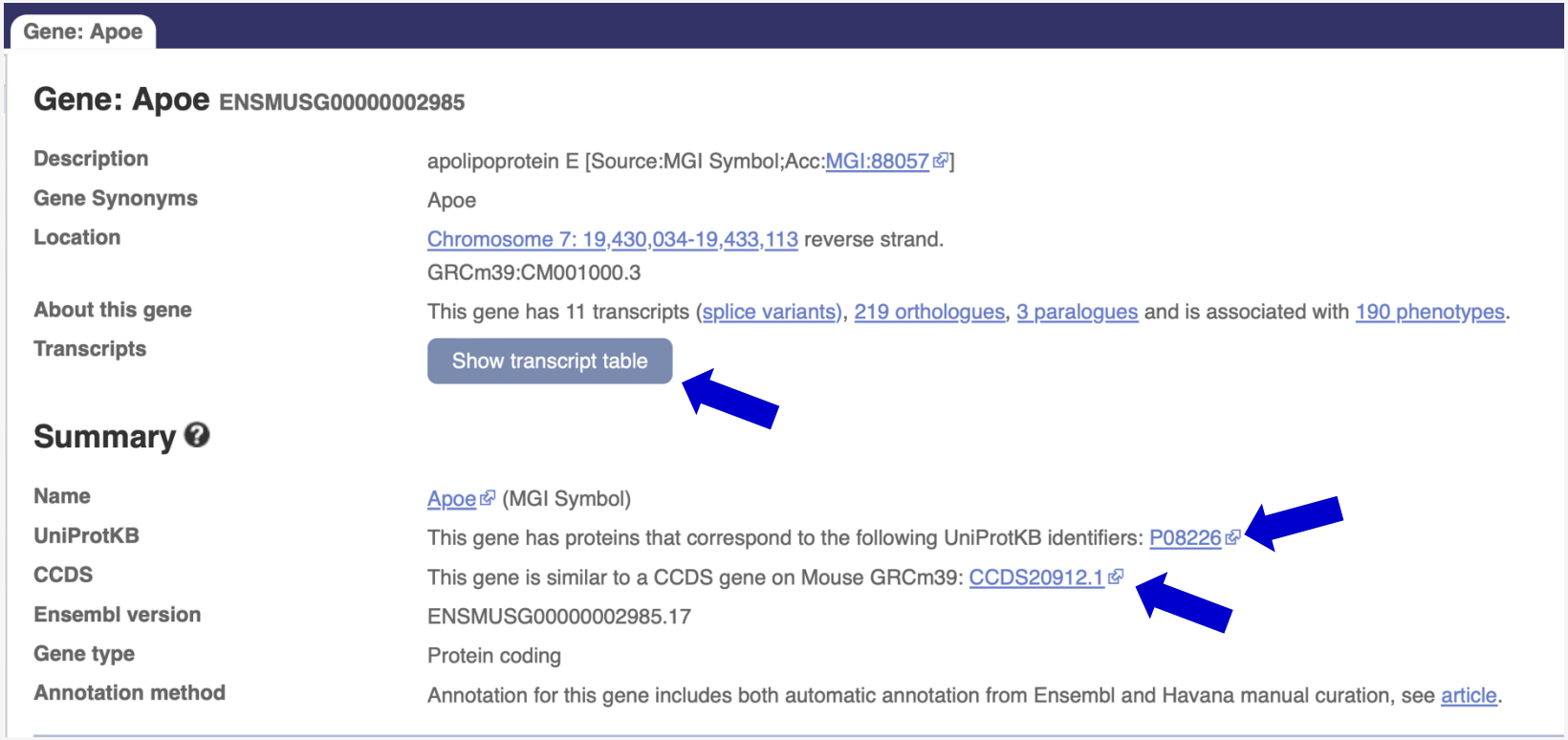
Genome Browsers
Ensembl example

The results table lists ”Apoe” found in all species, including entries for gene and transcripts.



Genome Browsers
Ensembl example

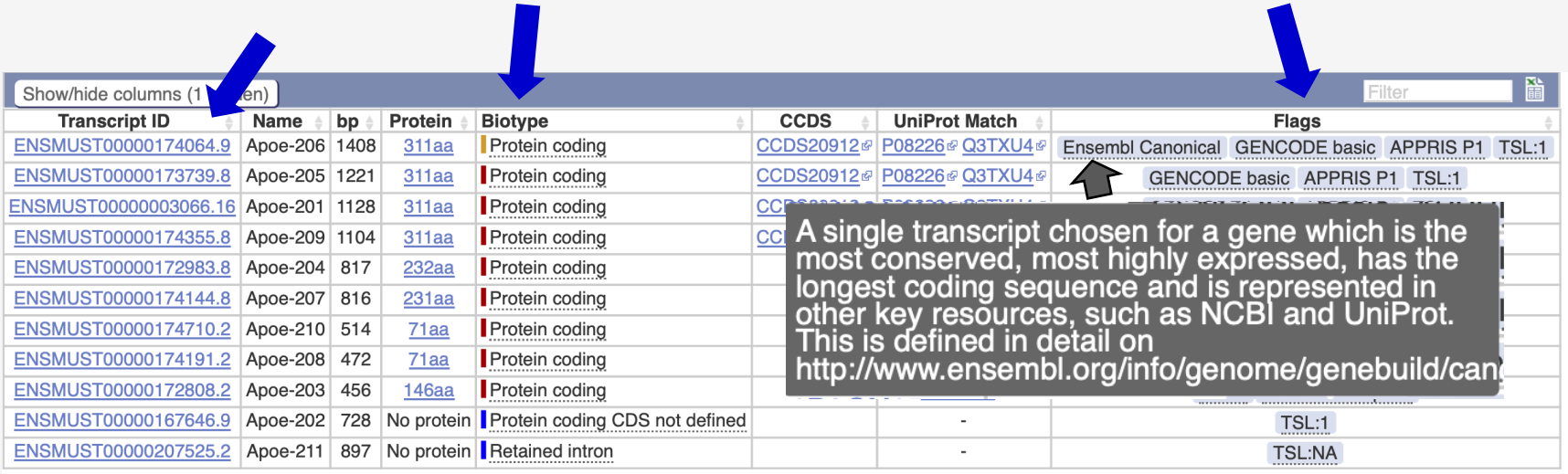
Click on the Apoe gene entry to open the 'Gene Tab'



Click on "Show transcript table"

Genome Browsers
Ensembl example

This allows you to view details such as identifiers, biotypes and transcript “Flags”.

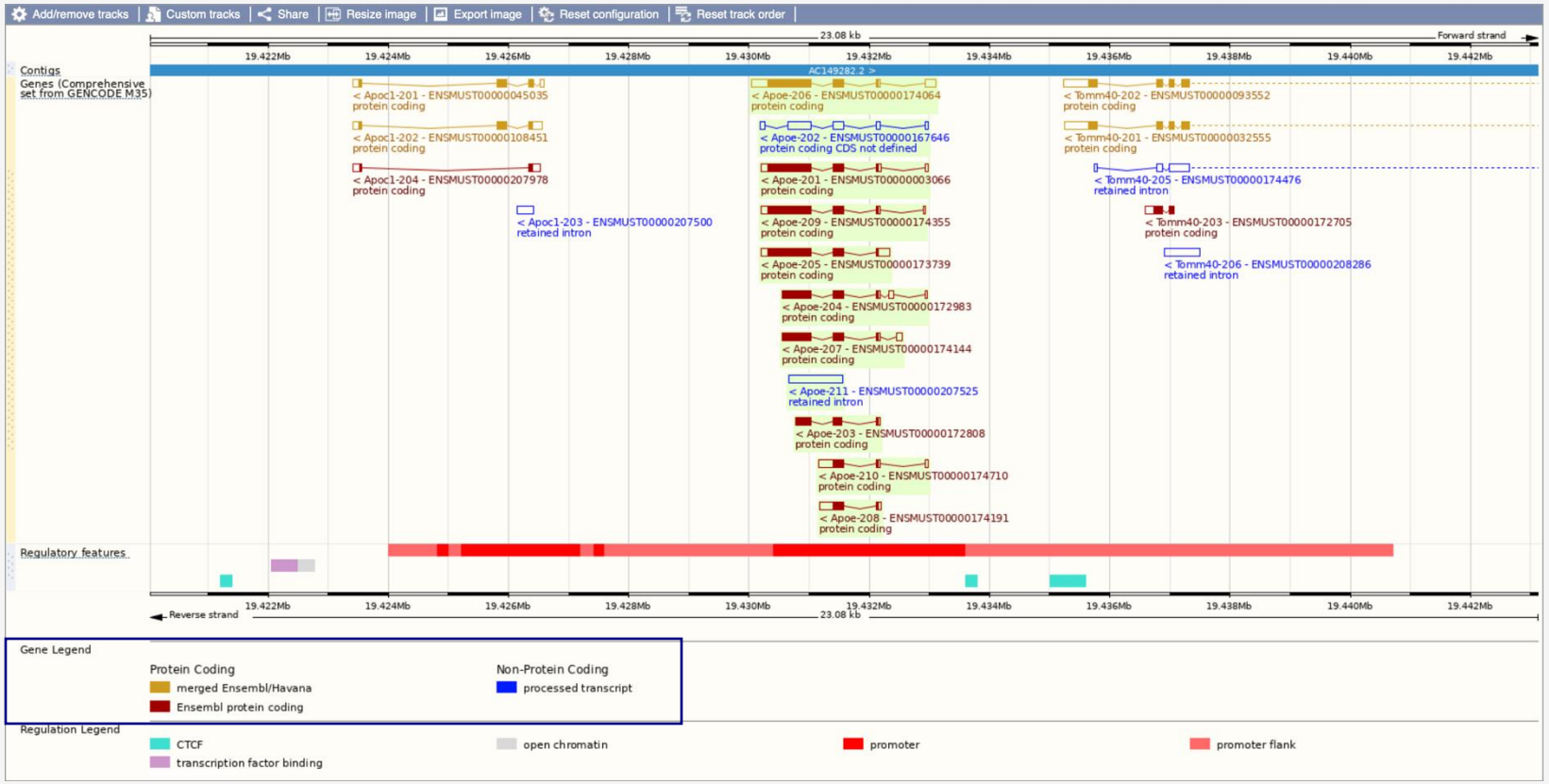


Position the cursor over a flag to view more information.

Close the transcript table and view the graphical display of all transcripts within the genomic context.

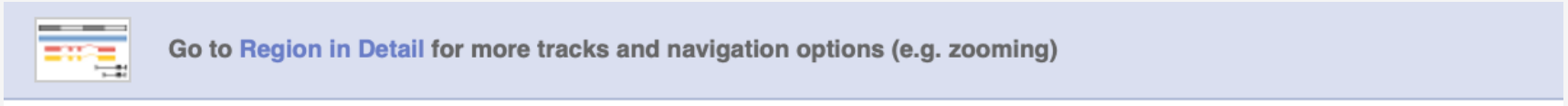
Genome Browsers
Ensembl example

Transcripts are displayed in different colours to indicate their origin.



Genome Browsers
Ensembl example

Click on ‘Region in detail’.

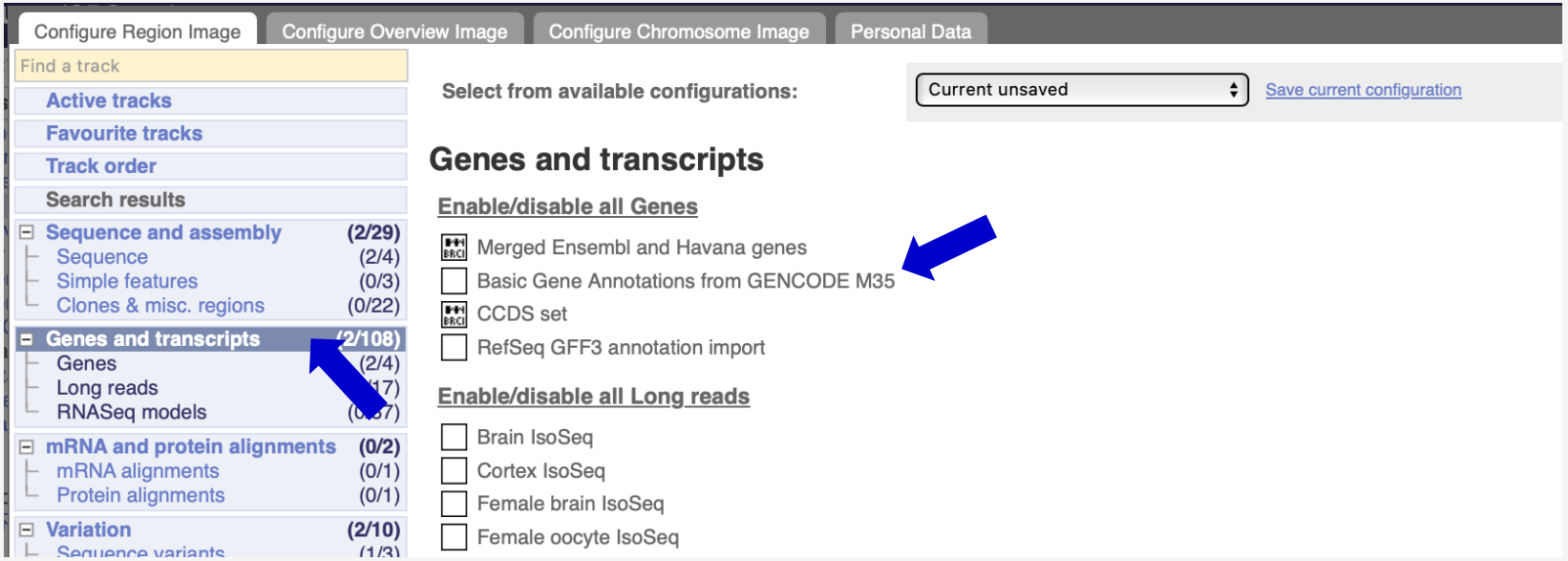


Ensembl displays the **GENCODE comprehensive gene set by default** (shown below), which includes all annotated transcripts.

Genome Browsers
Ensembl example

To display the GENCODE Basic gene set, first select “Configure this page” from the side menu.

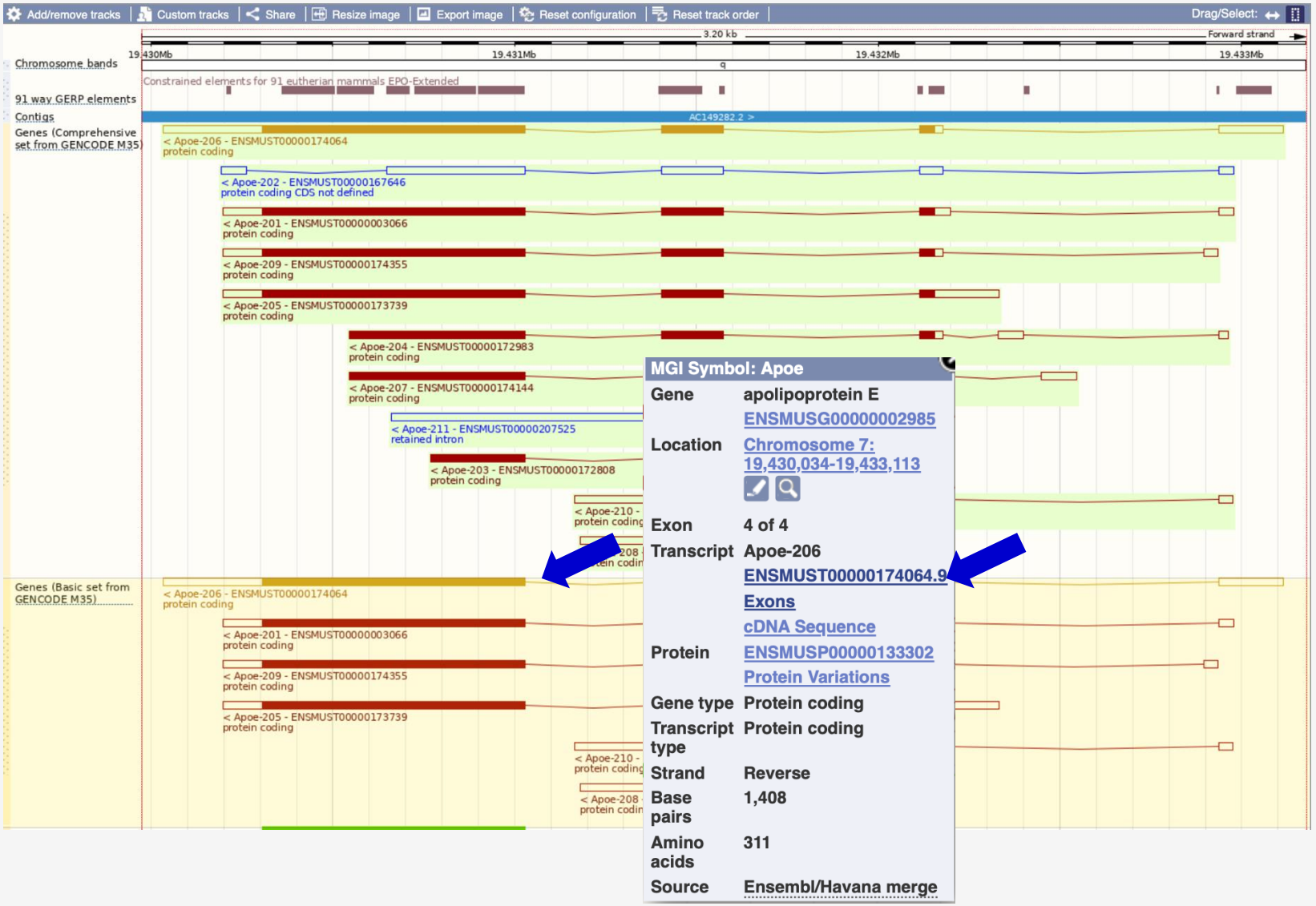
A new tab opens, and you can select desired options from the menu.



The Basic gene set only shows the main isoforms.

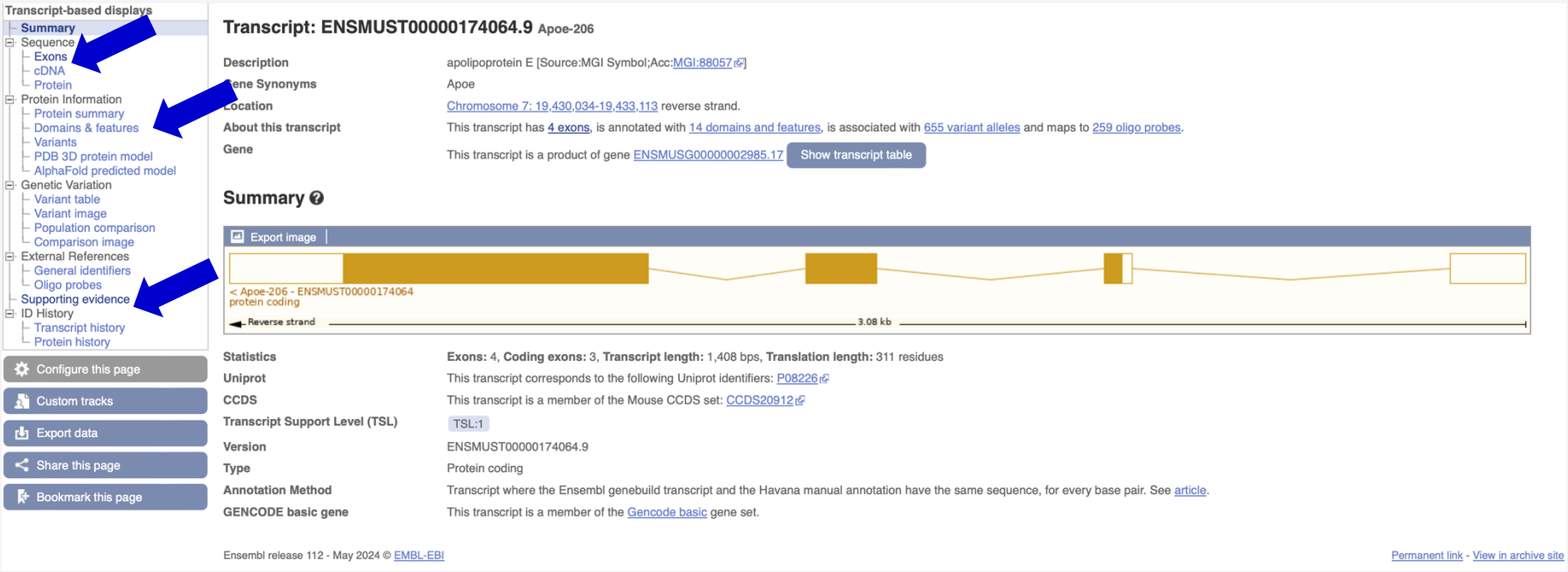
Genome Browsers
Ensembl example

View information about the transcript, exons and protein.



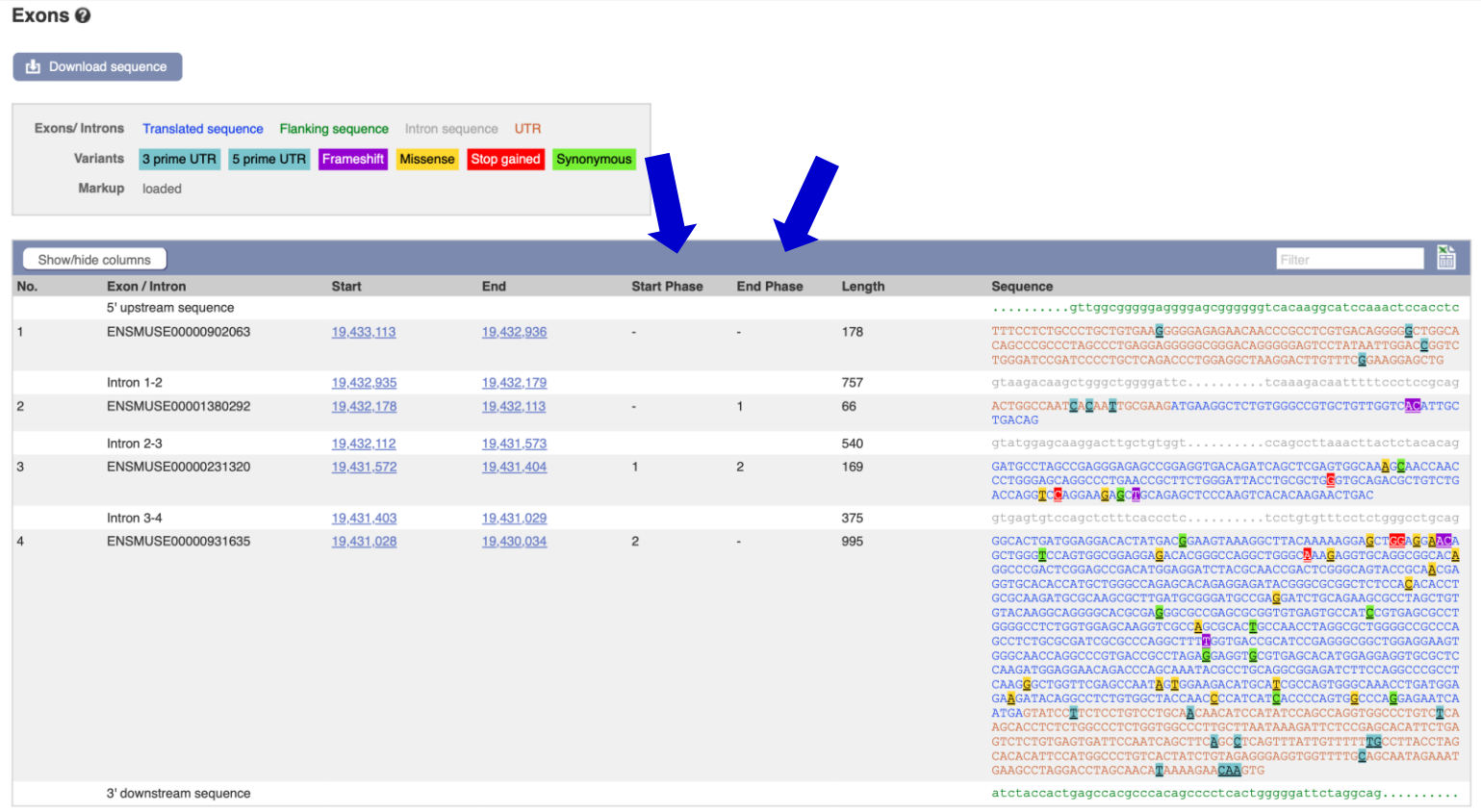
Genome Browsers
Ensembl example

From this view, use the side menu to view data such as supporting evidence, exon information and protein domain.



Genome Browsers
Ensembl example

Exon summary displays UTRs, CDS and common variants that are colour coded.

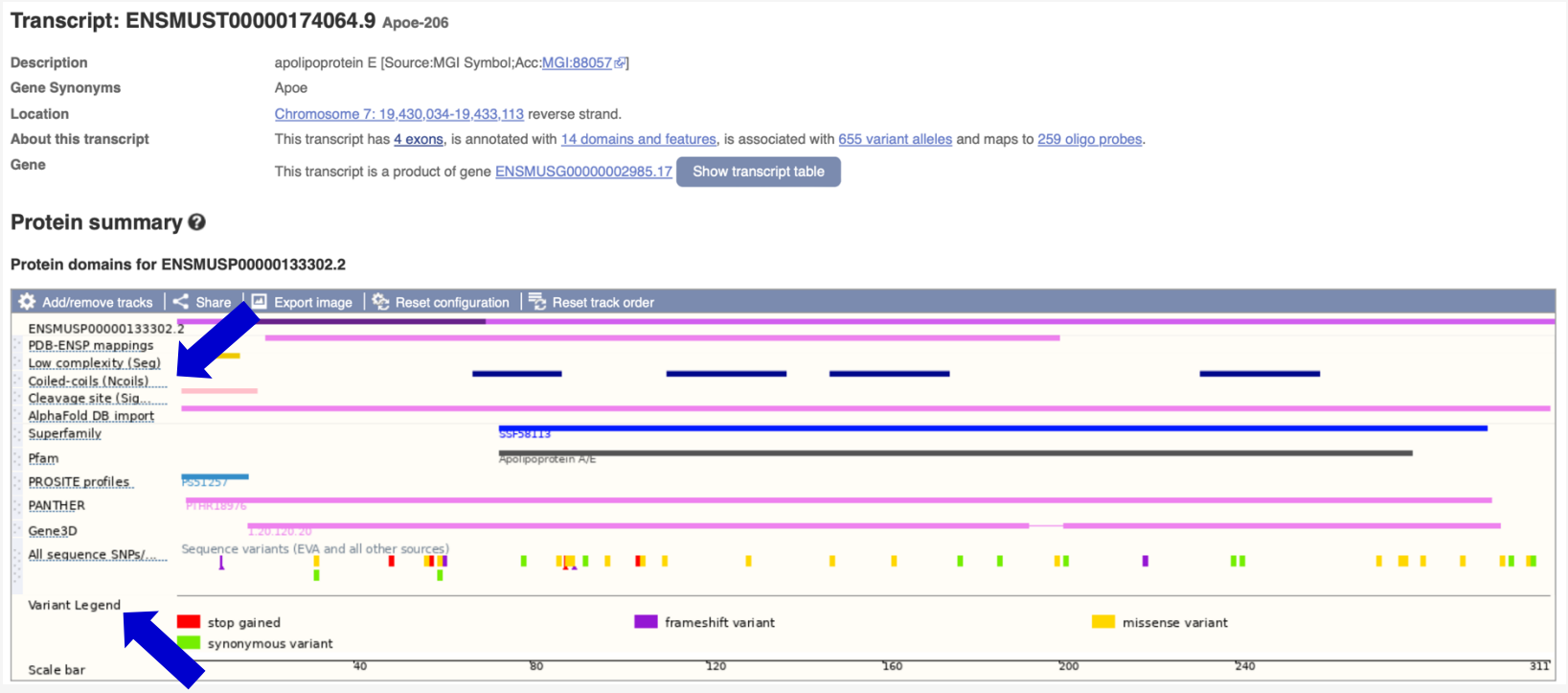


No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						.....gttggcgggggaggggagcgggggtcacaaggcatccaaactccacctc
1	ENSMUSE00000902063	19,433,113	19,432,936	-	-	178	TTTCCTCTGCCCTGCTGTGAAGGGGGAGAGAACAACCCGCCTCGTGACAGGGGGCTGGCA CAGCCCGCCCTAGCCCTGAGGAGGGGGCGGGACAGGGGGAGTCCTATAATTGGACCGGTC TGGGATCCGATCCCCTGCTCAGACCCTGGAGGCTAAGGACTTGTTTCGGAAGGAGCTG
	Intron 1-2	19,432,935	19,432,179			757	gtaagacaagctgggctggggattc.....tcaaagacaatttttccctccgcag
2	ENSMUSE00001380292	19,432,178	19,432,113	-	1	66	ACTGGCCAATCACAATTGCGAAGATGAAGGCTCTGTGGGCCGTGCTGTTGGTCACATTGC TGACAG
	Intron 2-3	19,432,112	19,431,573			540	gtatggagcaaggacttgctgtggt.....ccagccttaaacttactctacacag
3	ENSMUSE00000231320	19,431,572	19,431,404	1	2	169	GATGCCTAGCCGAGGGAGAGCCGGAGGTGACAGATCAGCTCGAGTGGCAAACGAACCAAC CCTGGGAGCAGGCCCTGAACCGCTTCTGGGATTACCTGCGCTGCGTGCAGACGCTGTCTG ACCAGGTCCAGGAAGAGCTGCAGAGCTCCCAAGTCACACAAGAACTGAC
	Intron 3-4	19,431,403	19,431,029			375	gtgagtgtccagctctttcaccctc.....tcctgtgtttcctctgggcctgcag
4	ENSMUSE00000931635	19,431,028	19,430,034	2	-	995	GGCACTGATGGAGGACACTATGACGGAAGTAAAGGCTTACAAAAAGGAGCTGGAGGAAGA GCTGGGTCCAGTGGCGGAGGAGACACGGGCCAGGCTGGGCAAAGAGGTGCAGGCGGCACA GGCCCGACTCGGAGCCGACATGGAGGATCTACGCAACCGACTCGGGCAGTACCGCAACGA GGTGCACACCATGCTGGGCCAGAGCACAGAGGAGATACGGGCGCGGCTCTCCACACACCT GCGCAAGATGCGCAAGCGCTTGATGCGGGATGCCGAGGATCTGCAGAAGCGCCTAGCTGT GTACAAGGCAGGGGCACGCGAGGGCGCCGAGCGCGGTGTGAGTGCCATCCGTGAGCGCCT GGGGCCTCTGGTGGAGCAAGGTCGCCAGCGCACTGCCAACCTAGGCGCTGGGGCCGCCCA GCCTCTGCGCGATCGCGCCCAGGCTTTTGGTGACCGCATCCGAGGGCGGCTGGAGGAAGT GGGCAACCAGGCCCGTGACCGCCTAGAGGAGGTGCGTGAGCACATGGAGGAGGTGCGCTC CAAGATGGAGGAACAGACCCAGCAAATACGCCTGCAGGCGGAGATCTTCCAGGCCCGCCT CAAGGGCTGGTTCGAGCCAATAGTGGAAGACATGCATTCGCCAGTGGGCAAACCTGATGGA GAAGATACAGGCCTCTGTGGCTACCAACCCCATCATCACCCCAGTGGCCCAGAGAATCA ATGAGTATCCTTCTCCTGTCCTGCAACAACATCCATATCCAGCCAGGTGGCCCTGTCTTCA AGCACCTCTCTGGCCCTCTGGTGGCCCTTGCTTAATAAAGATTCTCCGAGCACATTCTGA GTCTCTGTGAGTGATTCCAATCAGCTTCAGCTCAGTTTATTGTTTTTTGCCTTACCTAG CACACATTCCATGGCCCTGTCACTATCTGTAGAGGGAGGTGGTTTTCGAGCAATAGAAAT GAAGCCTAGGACCTAGCAACATAAAAGAACAAGTG
	3' downstream sequence						atctaccactgagccacgcccacagcccctcactgggggattctaggcag.....

Exon start and end phases (reading frames) are listed.

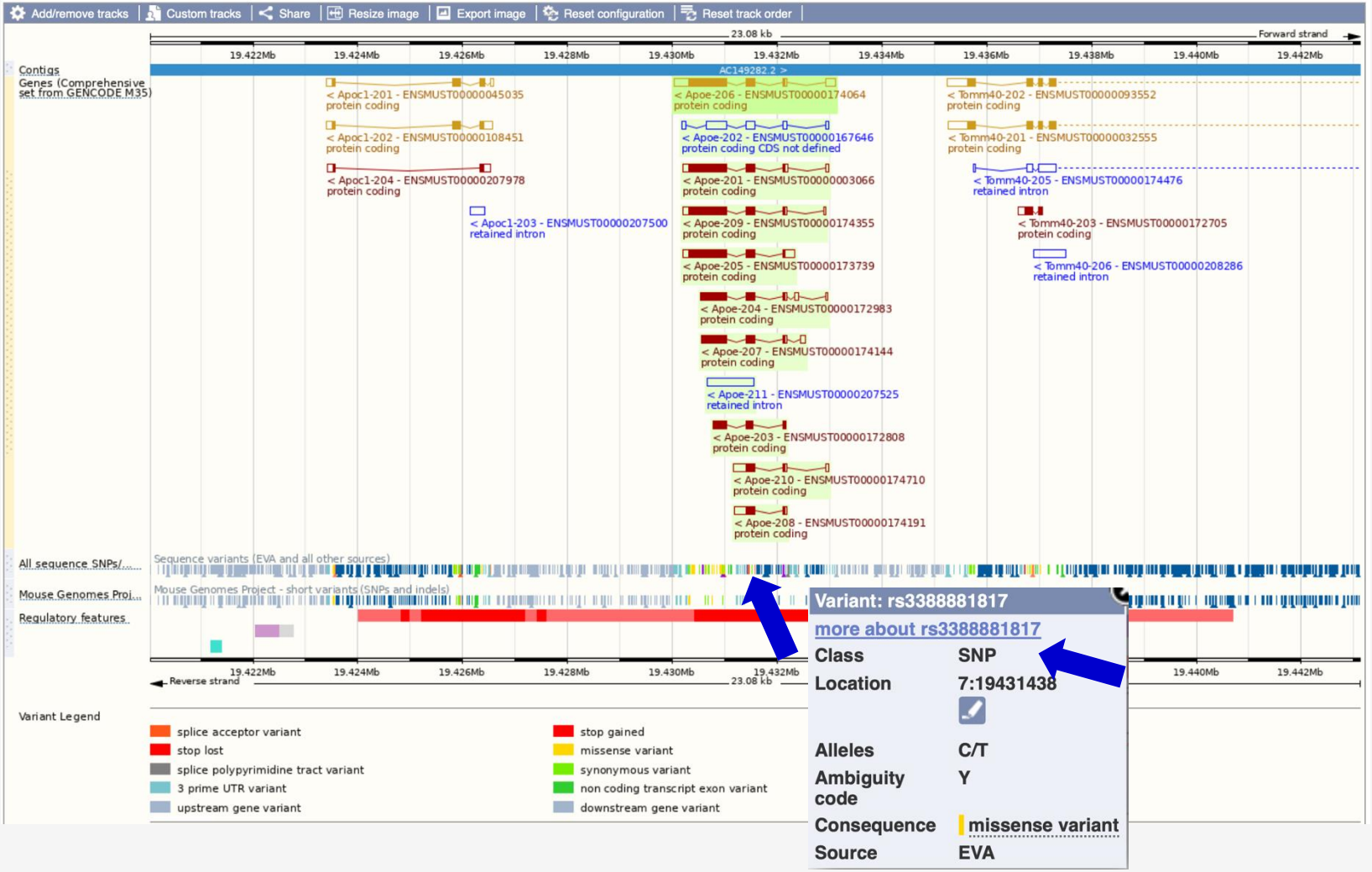
Genome Browsers
Ensembl example

Protein summary displays protein domains, signal peptides and transmembrane regions relative to the coding exons.



Genome Browsers
Ensembl example

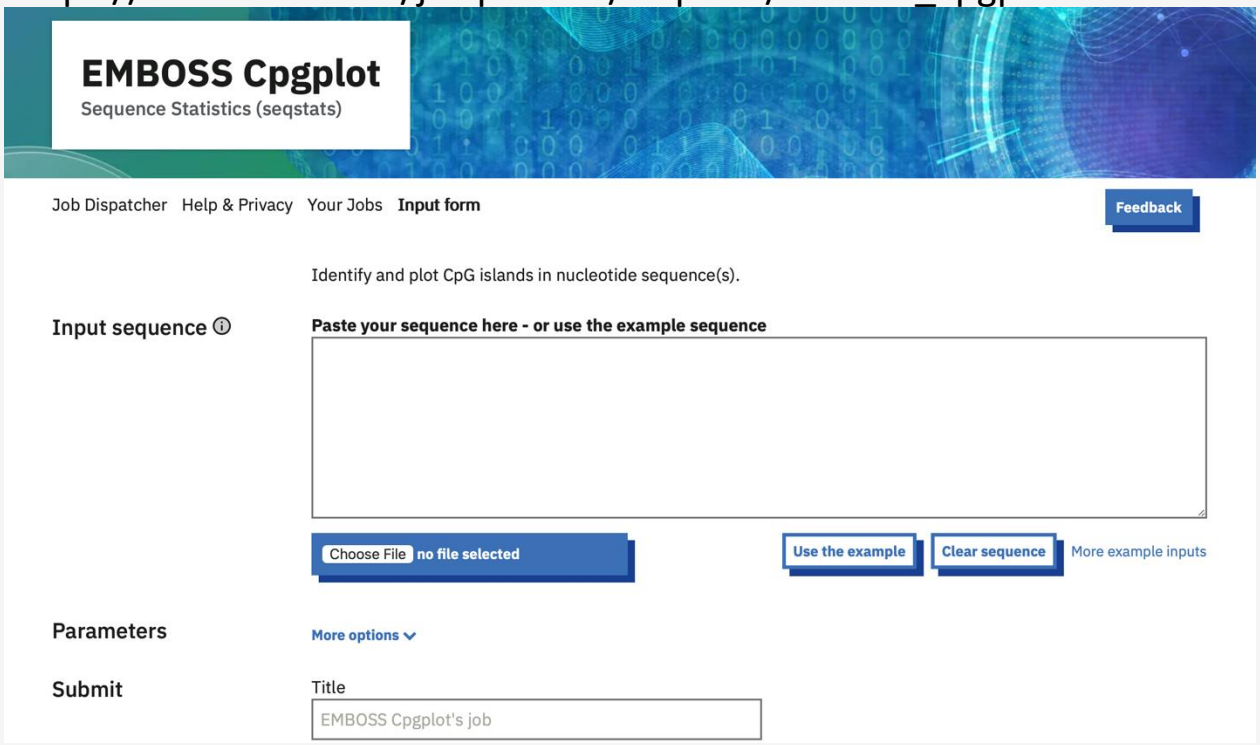
Variation data in Ensembl is a colour coded annotation that denotes the consequence of the change, e.g. frameshift.



Genome Browsers
Ensembl example

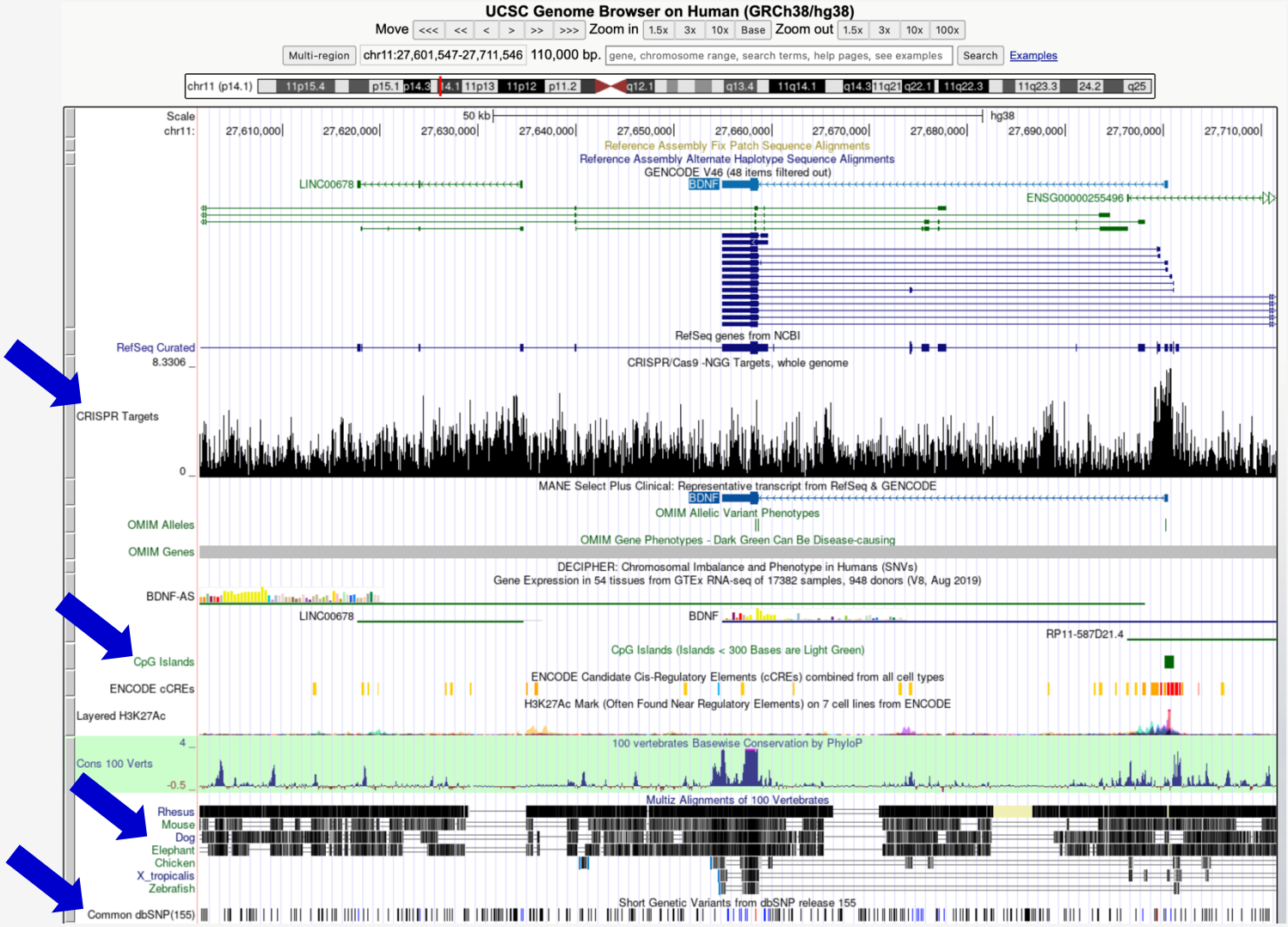
Export data for use in other software (e.g. EMBOSS CpGPlot or CRISPR sgRNA design tool).

https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_cpgplot



Genome Browsers
Other browsers

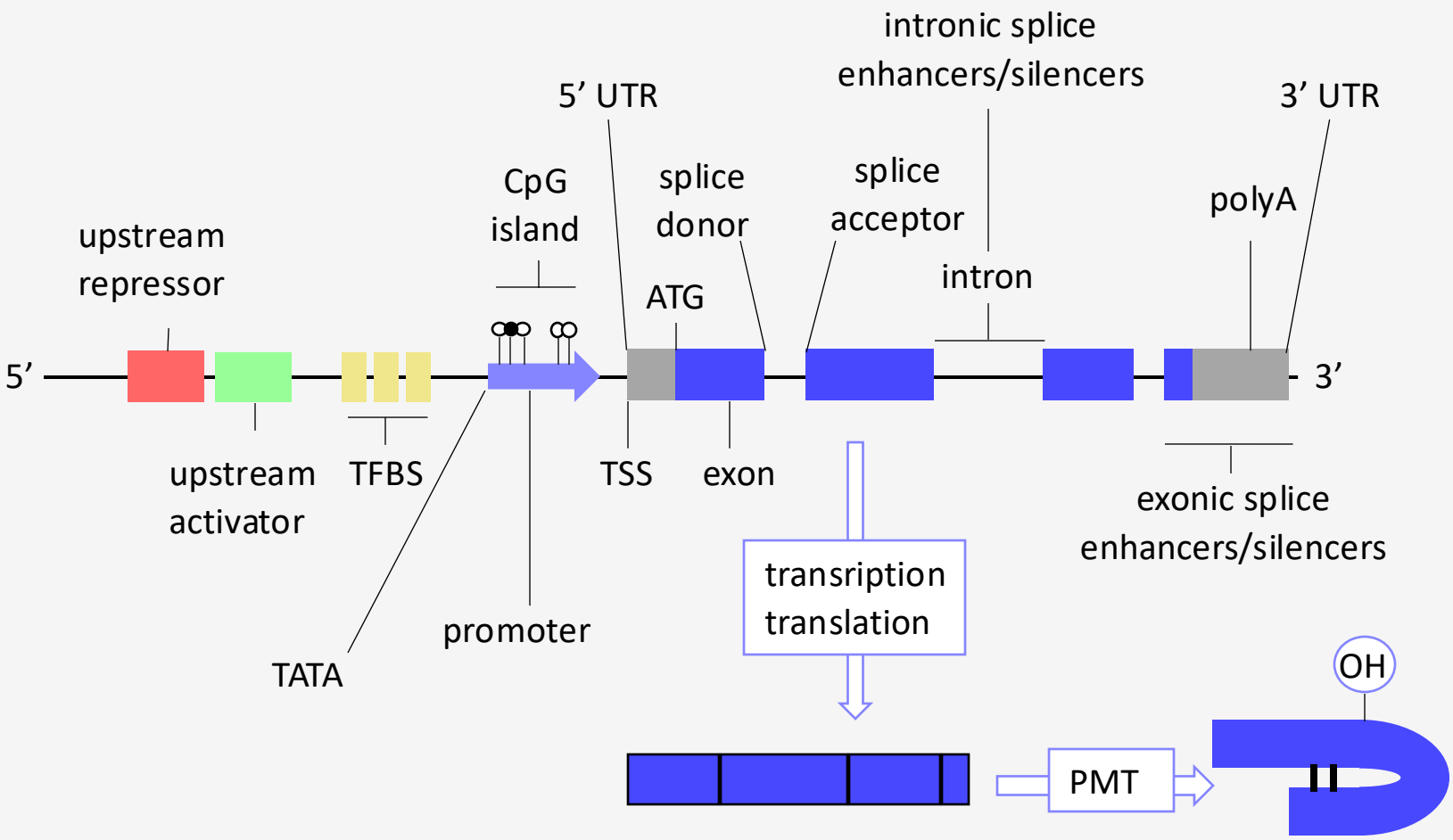
UCSC genome browser: http://genome.ucsc.edu



Gene Structure

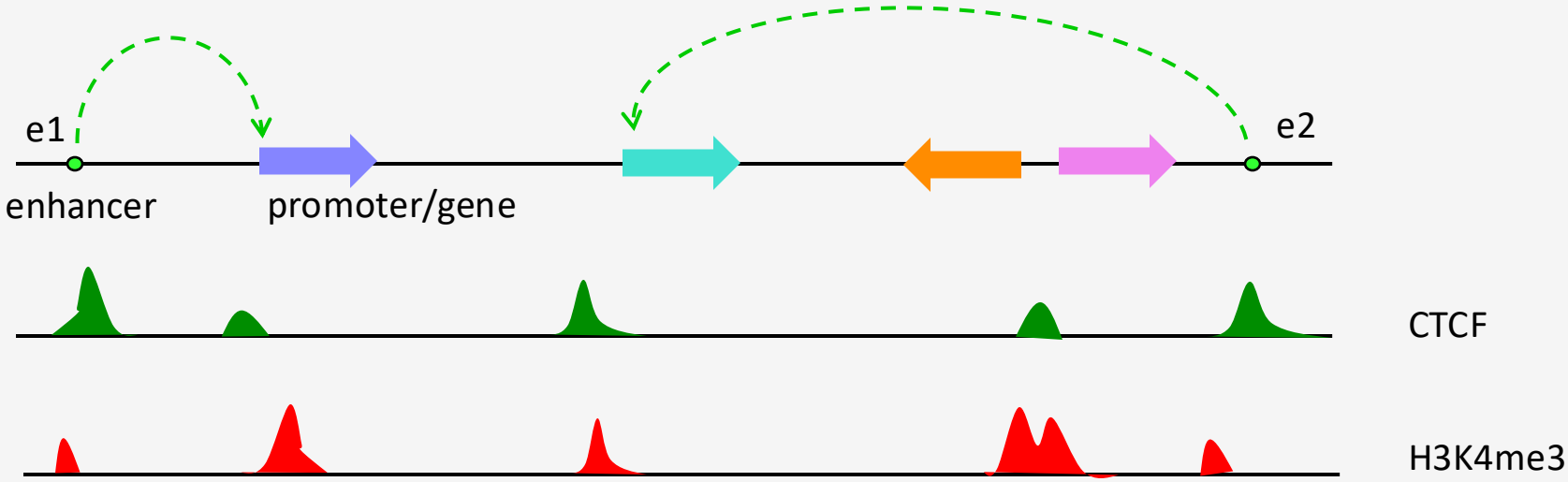
Gene Structure
What is a gene?

Protein coding or non-protein coding DNA sequences.

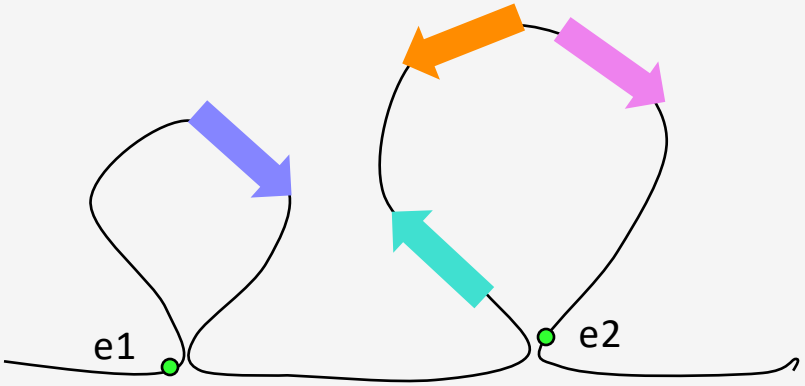


Gene Structure
Enhancers and Repressors

Analysis of chromatin interactions reveals details of regulatory domains and promoter interactions.

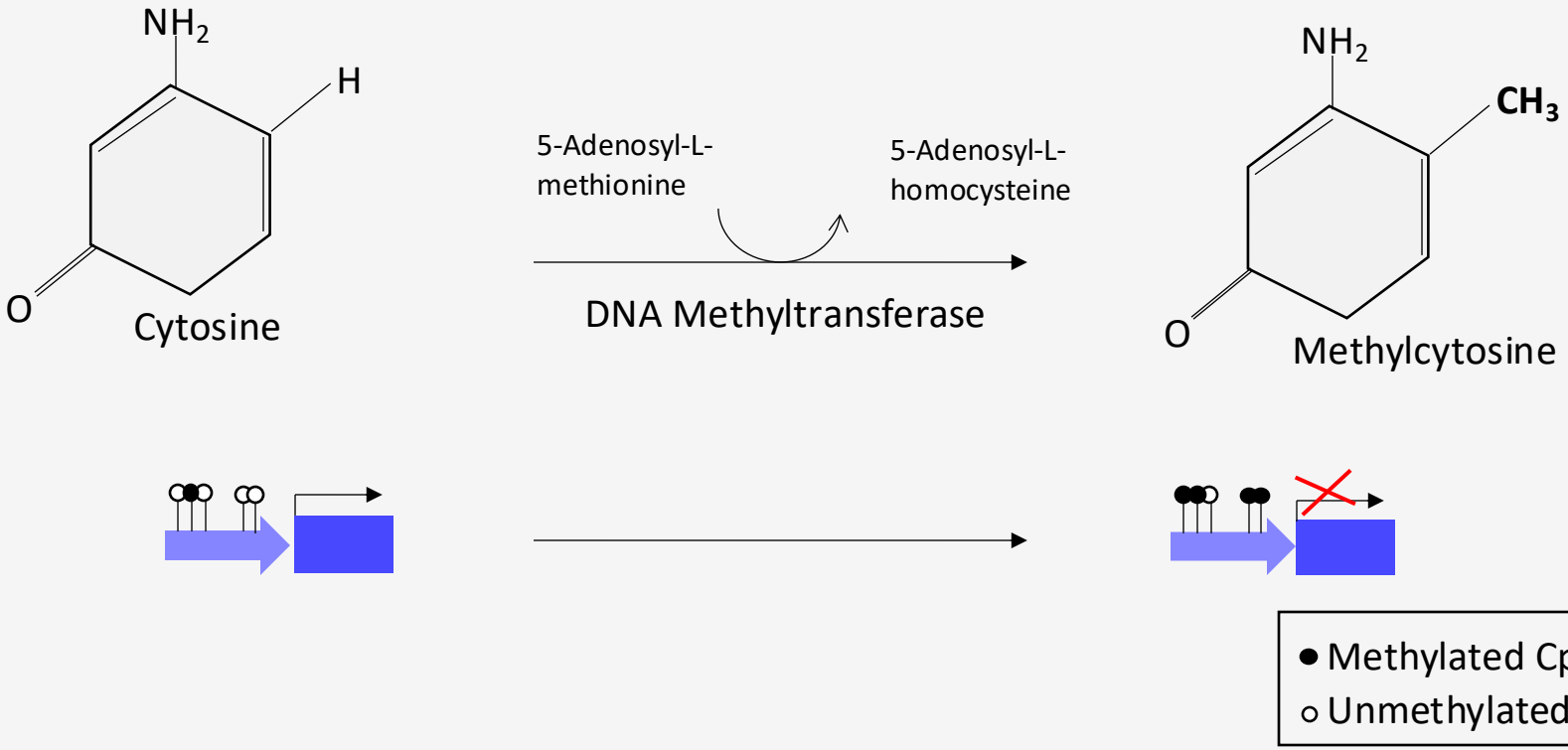


1D locations are really 3D interactions



Gene Structure
CpG islands and DNA methylation

CpG islands (un-methylated) are distinct from other CpG dinucleotides (methylated).
Often found near promoter regions ethylation of CpG island typically inactivates gene expression.



CpG finder algorithm is optimized for human, which can penalize other species.

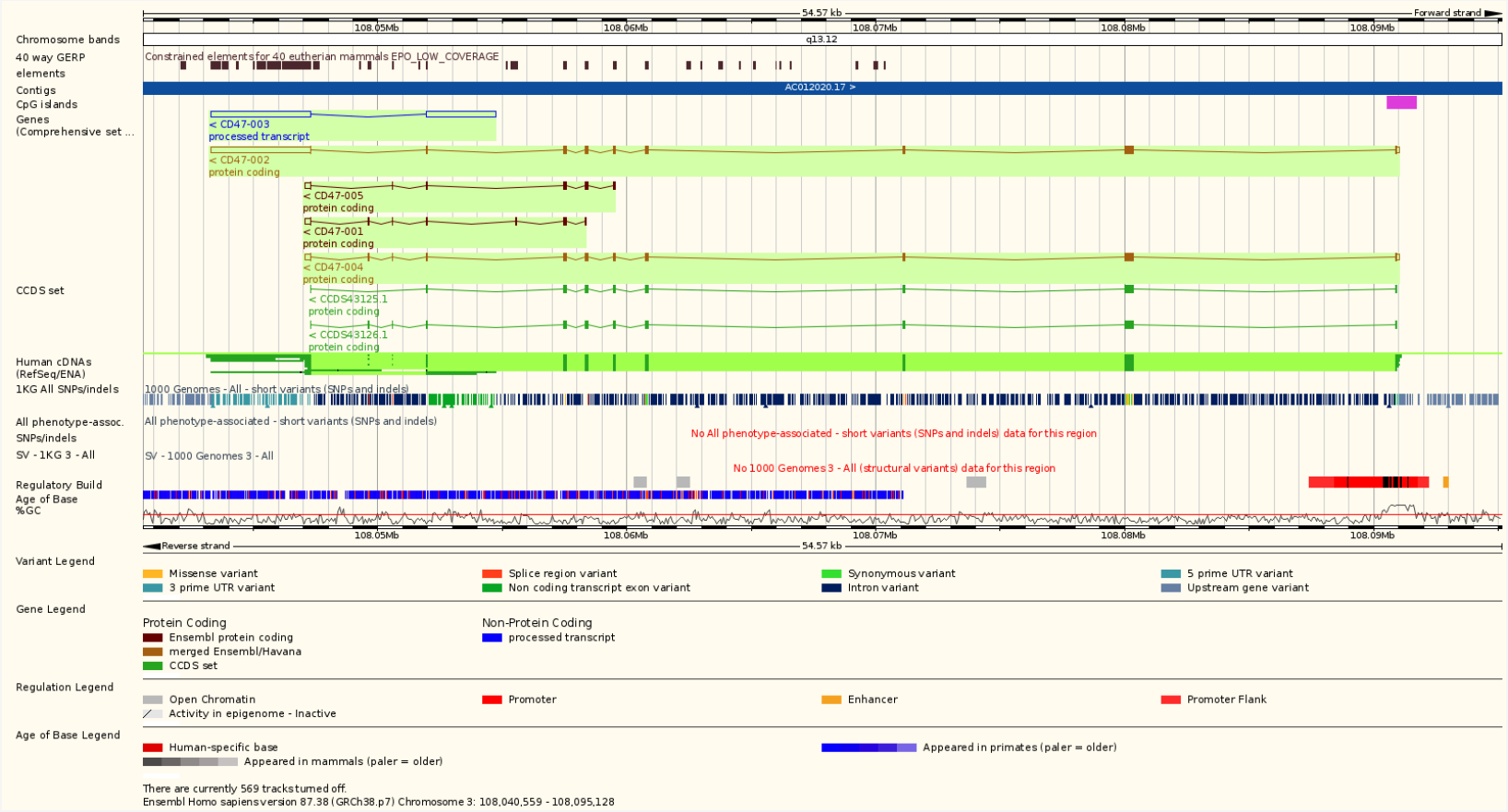
Experimental methylation data is more accurate.

Gene Structure
Untranslated regions (UTR)

Different 5’ UTR isoforms associated with
- Promoters (e.g. tissue specificity)
  - Transcription start sites (TS)
    - Translation start sites

Differences in 3’ UTR associated with
- Alternative polyadenylation (ApA)
  - mRNA stability, localisation, expression and regulation

Human CD47 molecule



Gene Structure
Transcription Start Site

GENCODE TSS positions coincide with the first nucleotide position of transcripts
- often based on extent of evidence, rather than experimental validation.

Large scale experimental data provided by the ENCODE and Fantom5 consortiums
- ENCODE contains Cap Analysis of Gene Expression (CAGE) and RAMPAGE data
- Fantom5 provides CAGE-seq peaks identifying TSS in different tissues and cell types for human and mouse

- Ongoing efforts to improve TSS annotation, with associated experimental data.

Gene Structure
CDS and Exon

Coding sequence (CDS) defines translated regions.
Different transcripts can have a different CDS for the same gene.
Every three bases of a CDS defines a codon for a specific tRNA.
Provides 3-frames of translation for forward or reverse strand.

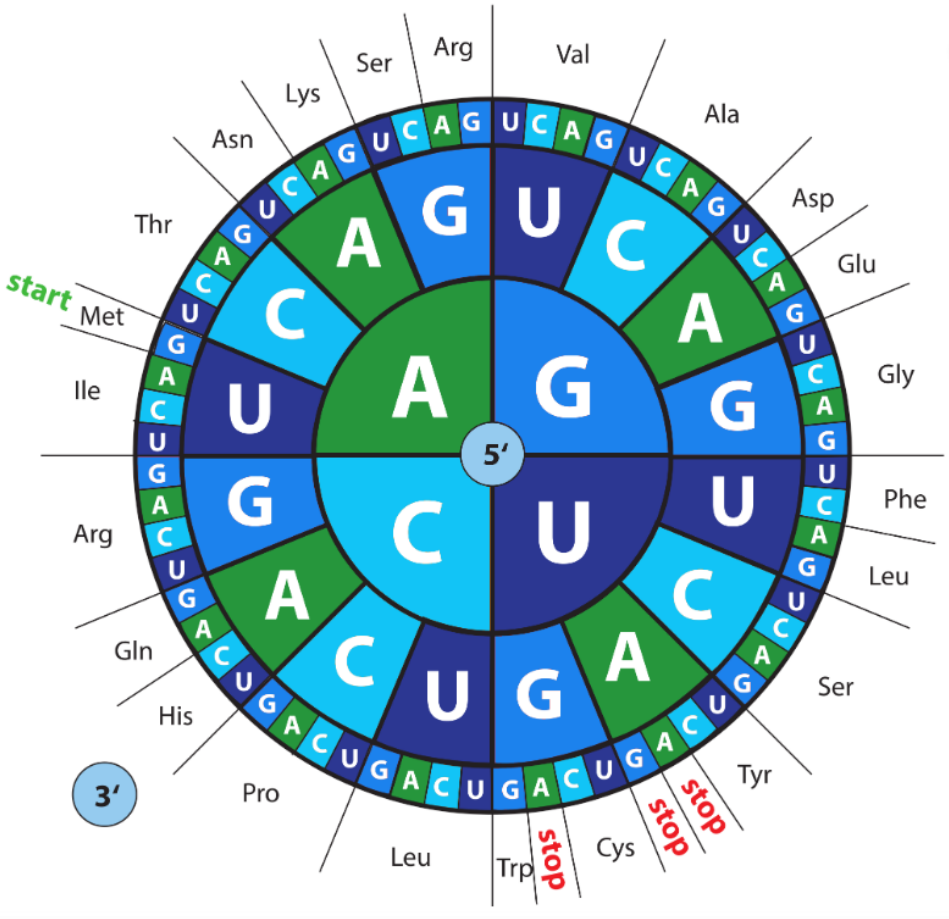
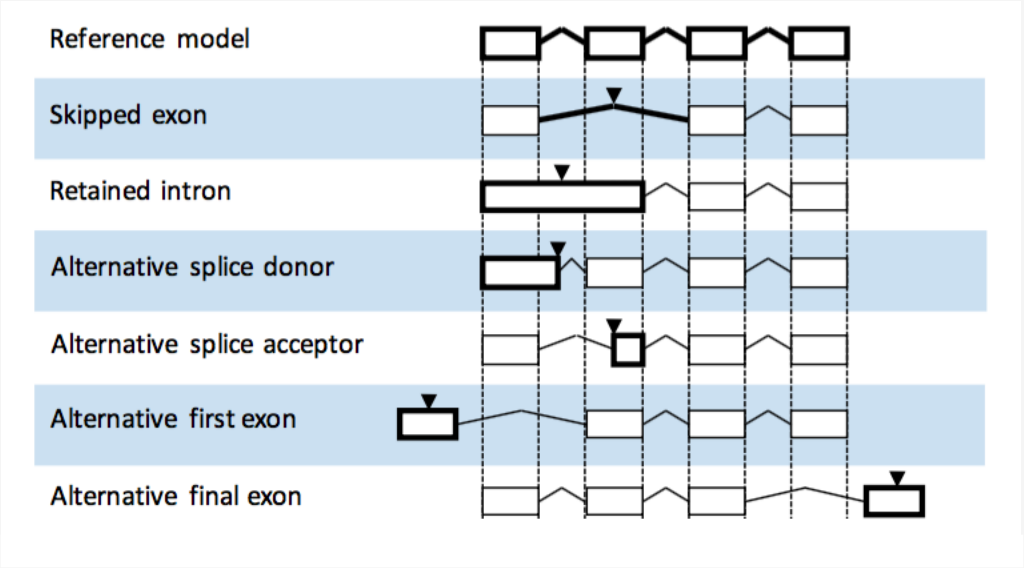


Image taken from Eurofins

Gene Structure
Alternative splicing

Most genes are subject to alternative splicing, which can include alternative first or last exons, exon skipping or variation in splice acceptor/donor sites.



Different protein isoforms

Alternative promoters and transcription start sites (TSS) can also give rise to isoforms.

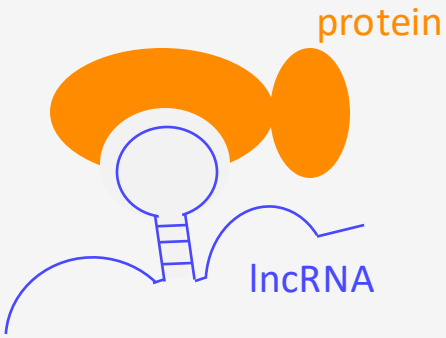
If alternative splicing introduces a premature stop codon, the aberrant transcript is degraded by the **Nonsense-mediated decay (NMD) pathway**.

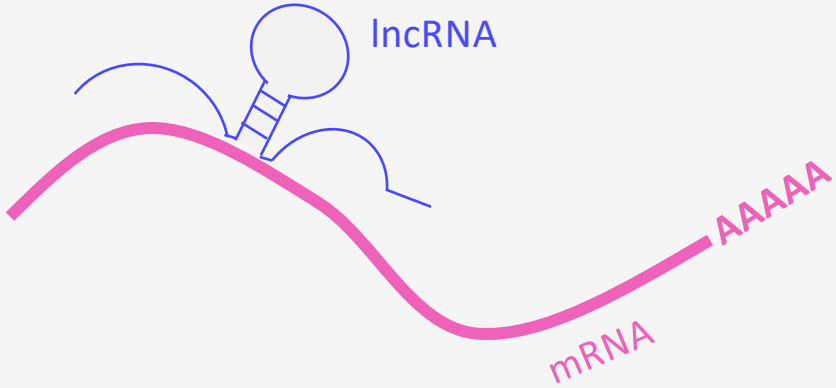
Gene Structure
Translation re-initiation

Occurs when a ribosome encounters a termination codon soon after the initiation codon, allowing the ribosome to progress through short ORFs found in the 5’ UTR.

Translation re-initiation is likely when a STOP codon <35 aa away from initiating methionine. It is also determined by the rate at which the translation initiation complex dissociates.

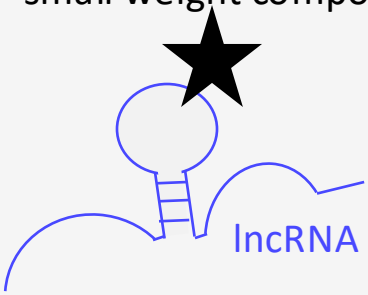
Gene Structure
Long non-coding RNA (lncRNA)

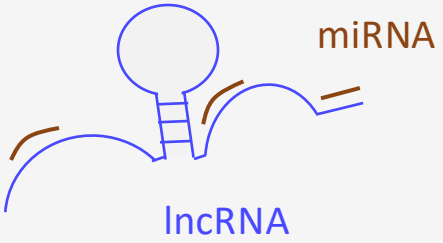


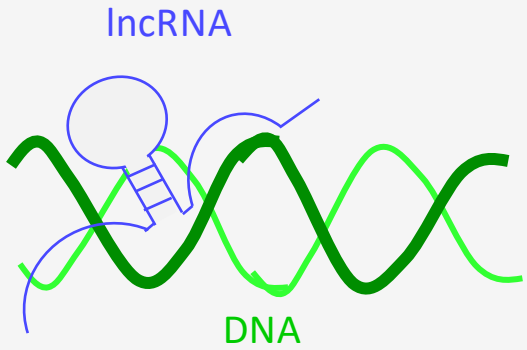


lncRNA interactome

small weight compounds







CRISPR Design Process and Tools

Are there any **other genes or genomic elements** in close proximity?

Are there alternative first or last **exons**?

Step 1: Gene structure and location

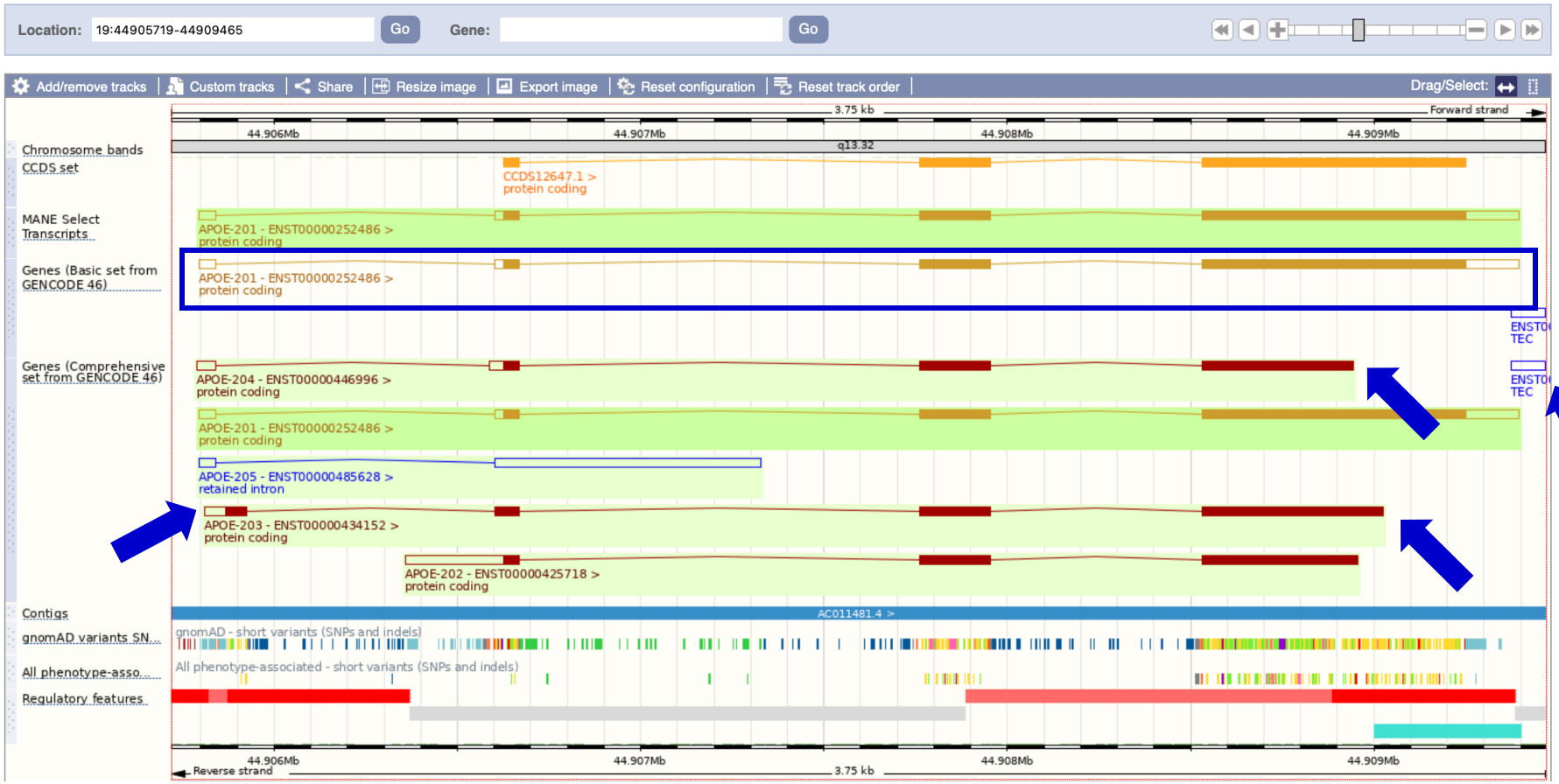
Are there alternative **translation start sites**?

Definition of the **gene start and end**?

Which transcript is most representative?

CRIPSR Design Process

Basic vs Comprehensive gene sets.

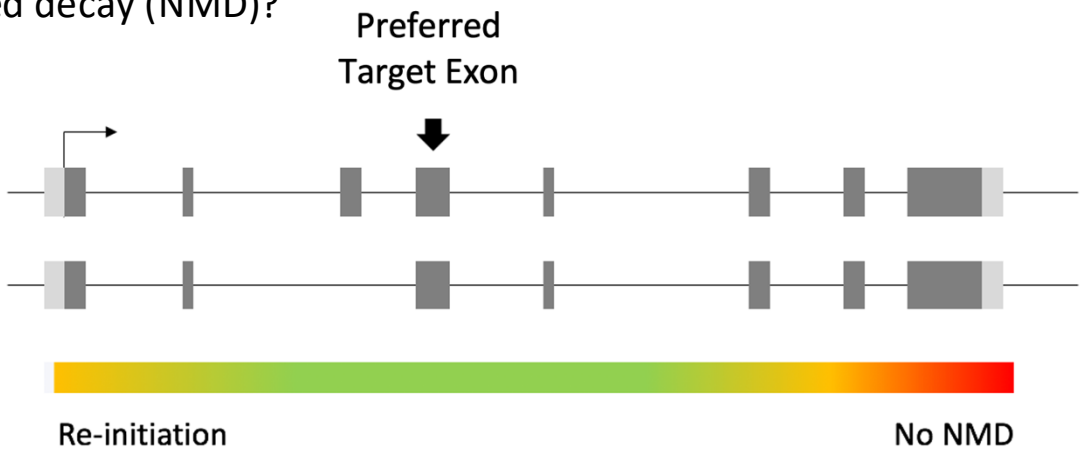


Which exons are **common** to all transcript isoforms?

Which exon(s) encode protein **domains**??

Step 2: Identify target region(s)

Will the mutant allele avoid translation re-initiation and be susceptible to nonsense-mediated decay (NMD)?



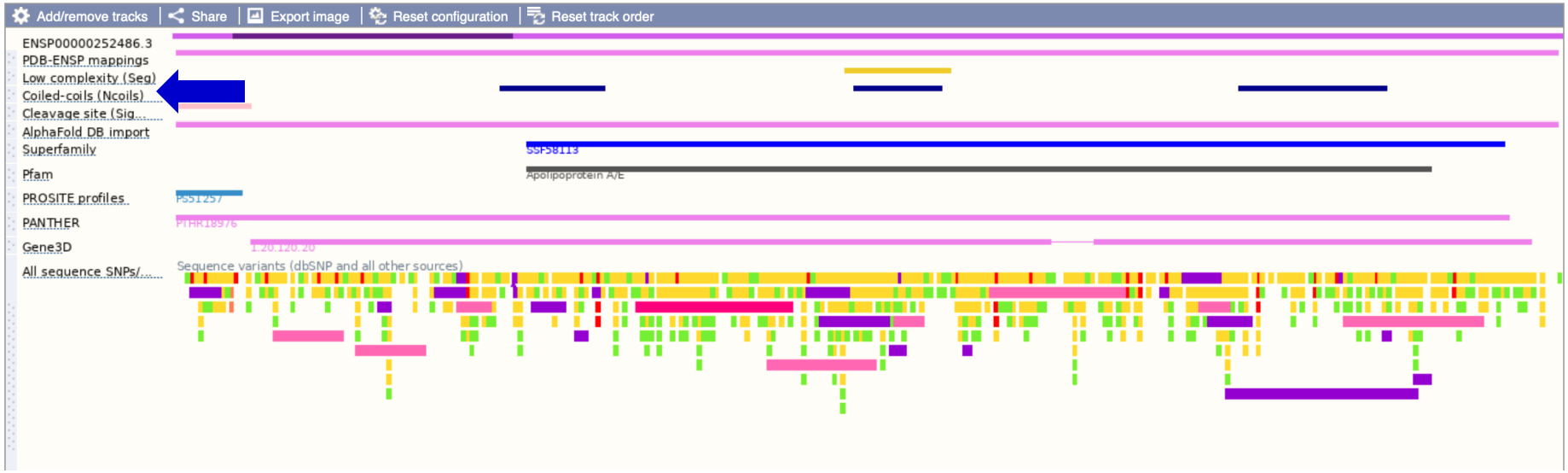
CRISPR Design Process and Tools

Open the transcript in a new window.

CRISPR Design Process and Tools

Protein summary

Protein domains for ENSP00000252486.3

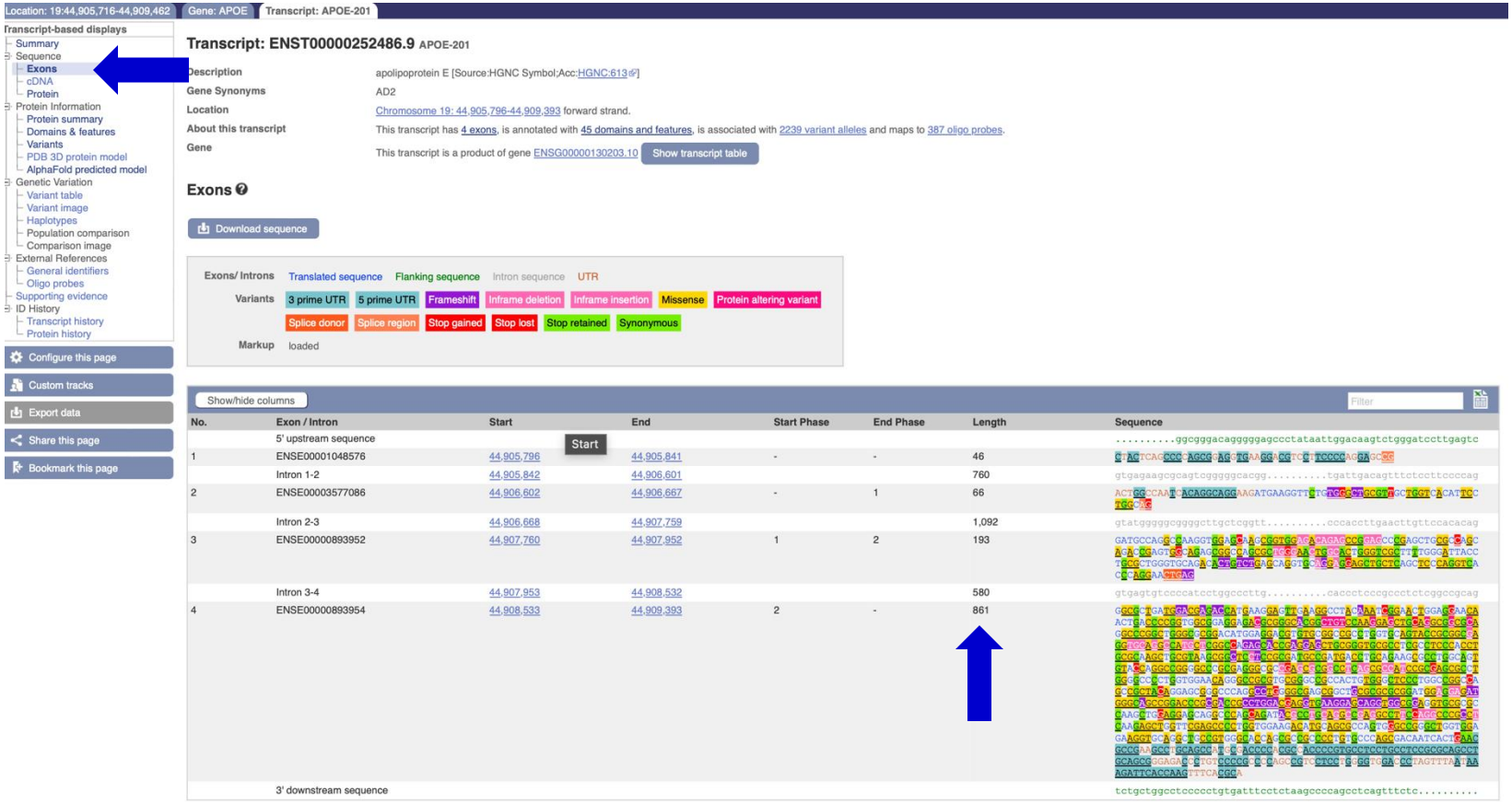


Other features

Show All entries Show/hide columns

Feature type	Start	End
Seg	154	178
SignalP	1	18
Alphafold	1	317
Ncoils	75	99
Ncoils	156	176
Ncoils	244	278

CRISPR Design Process and Tools



The first coiled-coil (aa 75-99) is in exon 3.

Low off-target scores and high on-target scores.

Avoid sgRNA with less than two mismatches.

Step 3: Select suitable gRNAs

Select sgRNA >50bp from the splice sites, to avoid disruption of splice junctions.

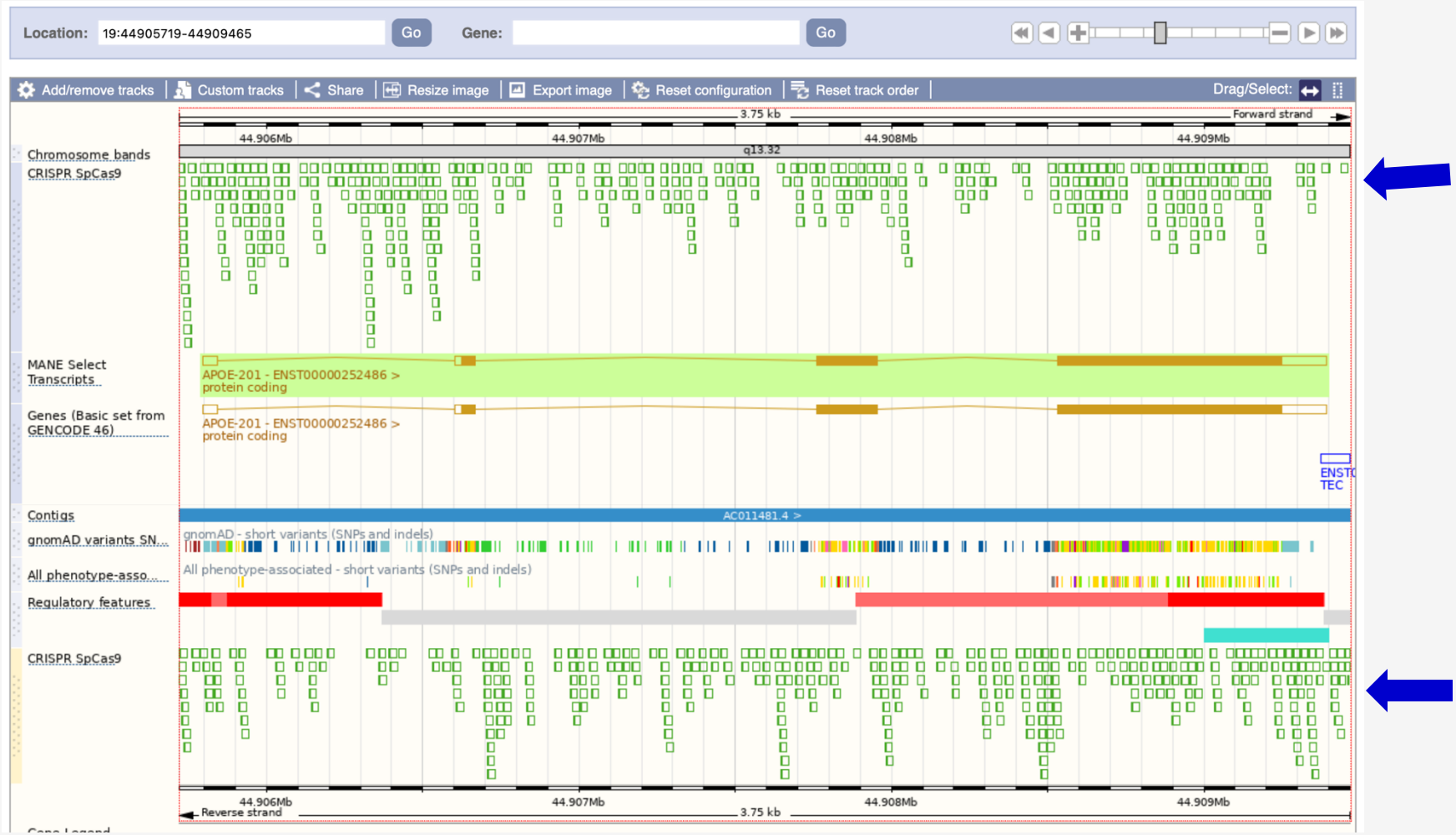
If there are no suitable gRNAs, **go back to step 2 and identify a new target region**.

CRISPR Design Process and Tools

Go back to the location tab and add a track to the “region in detail”.

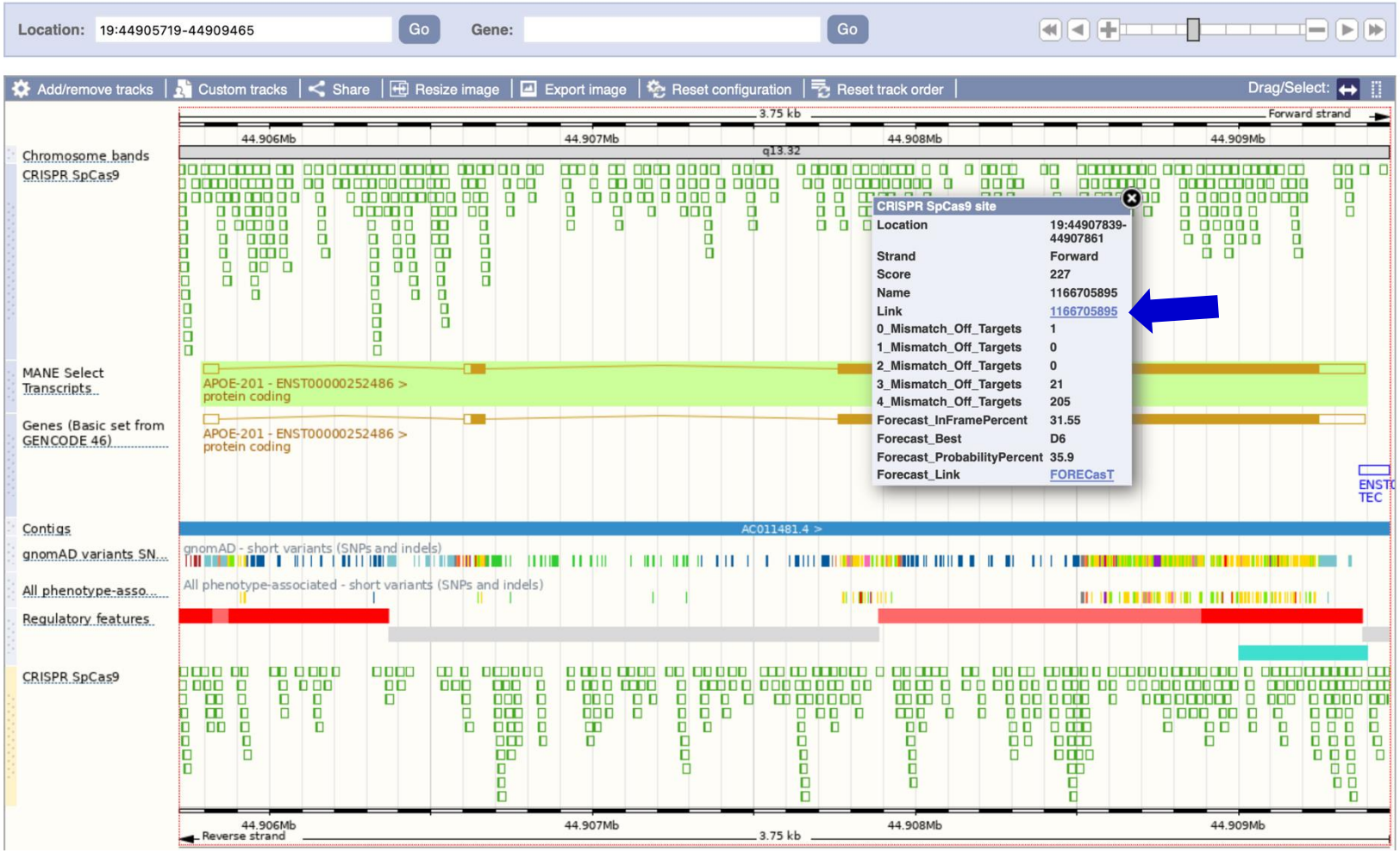
CRISPR Design Process and Tools

Guides targeting the forward strand are displayed on the top, and those targeting the reverse strand as shown below.



CRISPR Design Process and Tools

Information and links to other sources are available for each guide.



CRISPR Design Process and Tools

Individual CRISPR Summary

ID: 1166705895

View in Genome Browser

Species	Human (GRCh38)
Location	19:44907839-44907861
Sequence	CCAGCGCTGGGAACTGGCAC TGG
Strand	+
Crispr in exon?	Yes
Crispr in intron?	No

Off-Target Counts	All	Exonic	Intronic	Intergenic
Total	227	31	103	93
Summary	{0:1, 1:0, 2:0, 3:21, 4:205}	{0:1, 1:0, 2:0, 3:6, 4:24}	{0:0, 1:0, 2:0, 3:8, 4:95}	{0:0, 1:0, 2:0, 3:7, 4:86}

Found 5 Related Crispr Pairs

Show Crispr Pairs

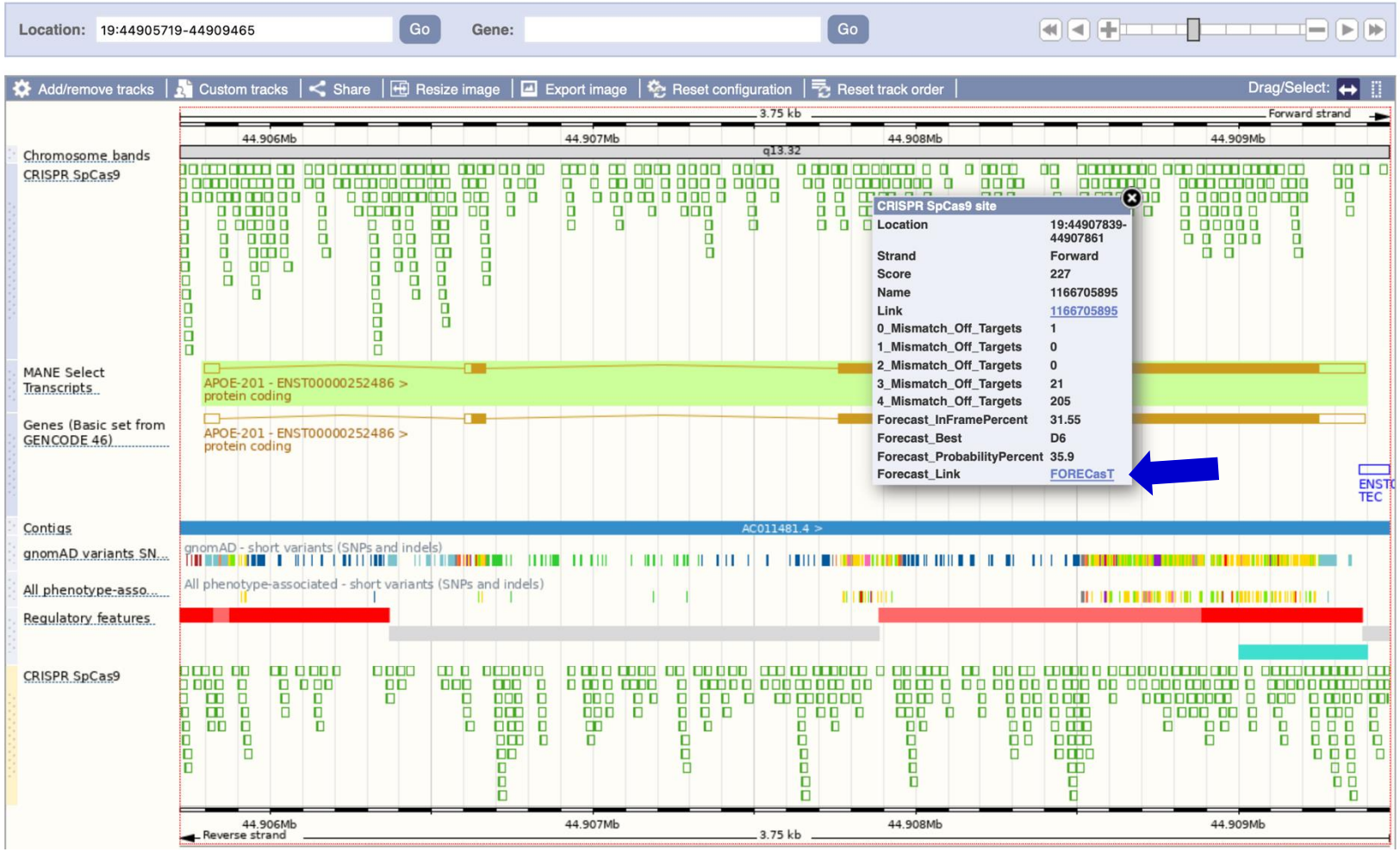
Off-Target Sites

Note: the row highlighted in blue is the original CRISPR

WGE ID	Location	Sequence	Mismatches	Strand	Type
1166705895	Original CRISPR	CCAGCGCTGGGAACTGGCAC TGG			Exonic
1166705895	19:44907839-44907861	CCAGCGCTGGGAACTGGCAC TGG	0	+	Exonic
1183733406	22:39630637-39630659	GCAGAGCCGGGAACTGGCAC AGG	3	+	Intronic
1021787321	7:24164871-24164893	GCAGCTCTAGGAACTGGCAC TGG	3	+	Intergenic
1001791760	5:174463776-174463798	CCACCCCTGGCAACTGGCAC AGG	3	+	Intergenic
1119608739	14:76043865-76043887	CCAGGGCTTGCAACTGGCAC TGG	3	-	Intronic

CRISPR Design Process and Tools

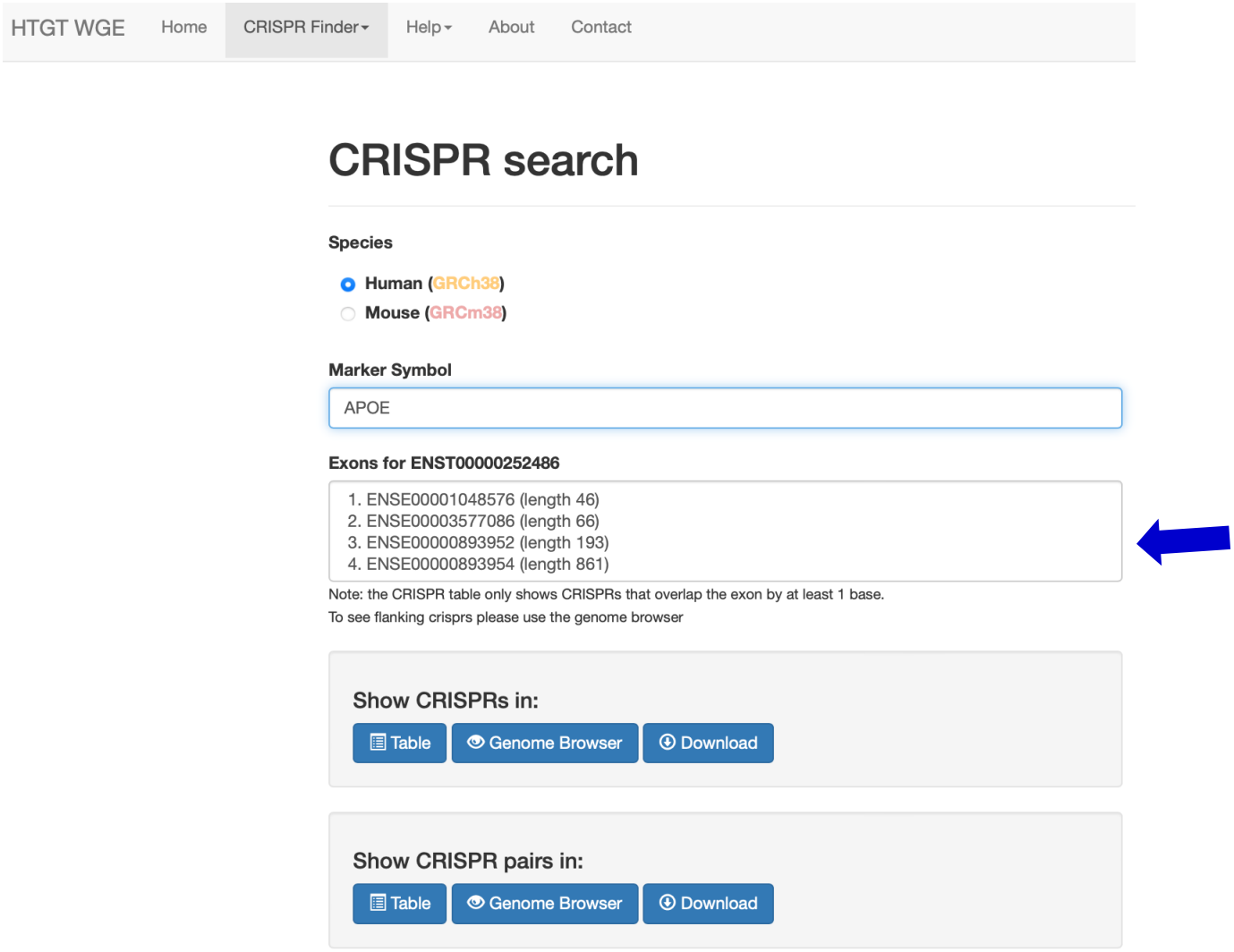
Information and links to other sources are available for each guide.



Nat Biotechnol 37, 64–72 (2019)

CRISPR Design Process and Tools

Generate a list of sgRNAs rather than selecting them individually.



CRISPR Design Process and Tools

Exon ID	ID	Location	Seq	Off targets
ENSE00000893952	1166705873	19:44907738-44907760	ACCTTGAACTTGTTCCACACAGG	0: 1 1: 0 2: 0 3: 13 4: 93
	1166705874	19:44907739-44907761	CCTTGAACTTGTTCCACACAGGA	0: 1 1: 0 2: 0 3: 16 4: 137
	1166705875	19:44907746-44907768	CTTGTTCCACACAGGATGCCAGG	0: 1 1: 0 2: 1 3: 16 4: 189
	1166705877	19:44907752-44907774	CCACACAGGATGCCAGGCCAAGG	0: 1 1: 0 2: 4 3: 34 4: 290
	1166705876	19:44907752-44907774	CCACACAGGATGCCAGGCCAAGG	0: 1 1: 0 2: 1 3: 30 4: 251

Required for confirming correct targeting or frameshifts

If there are no unique genotyping primers available, then go back to the previous stages and select an alternative gRNA and/or target region

Step 4: Design genotyping primers

https://www.ncbi.nlm.nih.gov/tools/primer-blast/

Some gRNA design tools have a built-in primer design tool.

CRISPR Design Process and Tools

Primers for target on one template | Primers common for a group of sequences

Retrieve recent results | Publication | Tips for finding specific primers

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

Range Clear

From To

Forward primer

Reverse primer

Or, upload FASTA file Choose File no file selected

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Clear

Use my own reverse primer (5'->3' on minus strand) Clear

	Min	Max
PCR product size	70	1000
# of primers to return	10	

	Min	Opt	Max	Max T_m difference
Primer melting temperatures (T_m)	57.0	60.0	63.0	3

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span No preference

Exon junction match

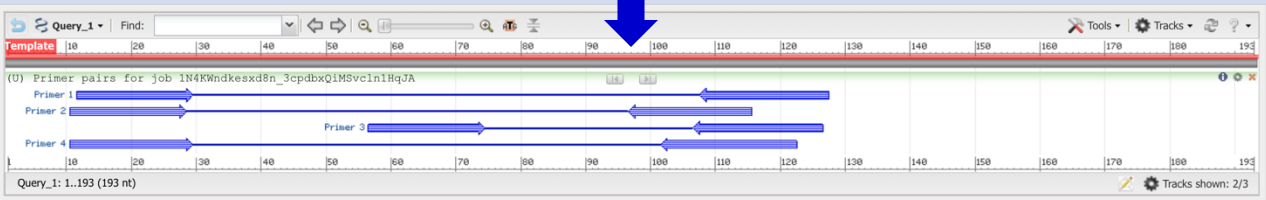
Min 5' match	Min 3' match	Max 3' match
7	4	8

Minimal and maximal number of bases that must anneal to exons at the 5' or 3' side of the junction

Intron inclusion Primer pair must be separated by at least one intron on the corresponding genomic DNA

Intron length range

Graphical view of primer pairs



Detailed primer reports

Download primer pairs

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AAGGTGGAGCAAGCGGTG	Plus	18	12	29	60.28	61.11	2.00	0.00
Reverse primer	CAGCGCAGGTAATCCCAAAA	Minus	20	127	108	58.83	50.00	4.00	0.00
Product length	116								

Other sgRNA Design Tools

sgRNA Design Tools

Other tools

HTGT WGE | Home | CRISPR Finder | Help | About | Contact

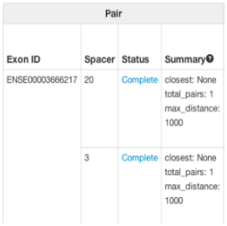
Wellcome Sanger Institute Genome Editing (WGE) is a website that provides tools to aid with genome editing of human and mouse genomes

CRISPR Finder Ensembl for Human Ensembl for Mouse

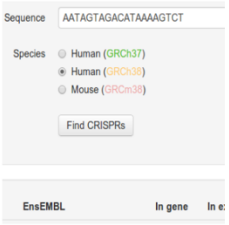
The CRISPR Finder will show CRISPR sites (paired or single) in and around genes. You can ask the finder to score the pairs for potential off-target sites, and browse individual and paired CRISPR sites using the Genoverse genome browser tool. We also provide the ability to find CRISPRs in genomic sequence or by gRNA:

Find CRISPRs in our genome browser:

Find CRISPRs by gene using our table:

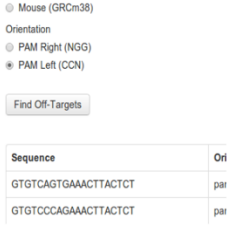


Find CRISPRs by 20bp gRNA:



Find CRISPRs in genomic sequence:

Find off-targets by sequence:

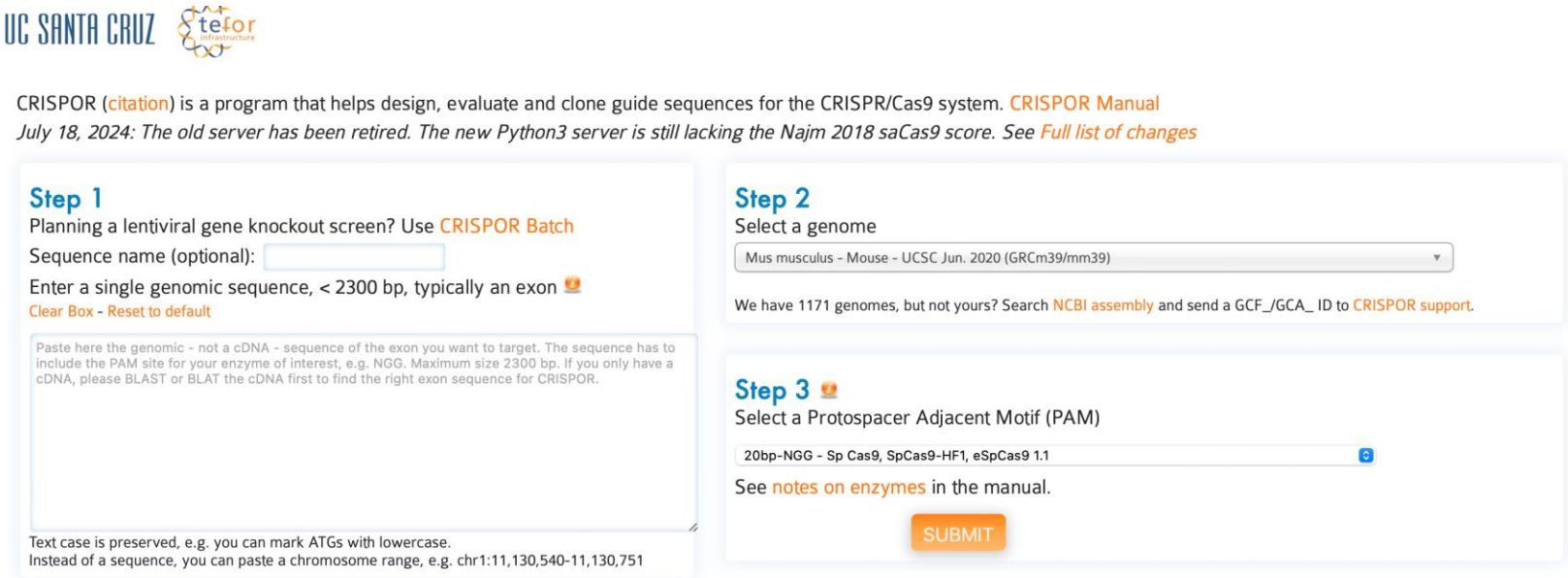


If you use this site in your research, please cite:
WGE: A CRISPR database for genome engineering. Alex Hodgkins; Anna Farne; Sajith Perera; Tiago Grego; David J. Parry-Smith; William C. Skarnes; Vivek Iyer (Bioinformatics 2015) doi:10.1093/bioinformatics/btv308

wellcome sanger institute

sgRNA Design Tools

Other tools



sgRNA Design Tools

Other tools

Home Instructions Scoring About Updates Submissions Contact FAQ

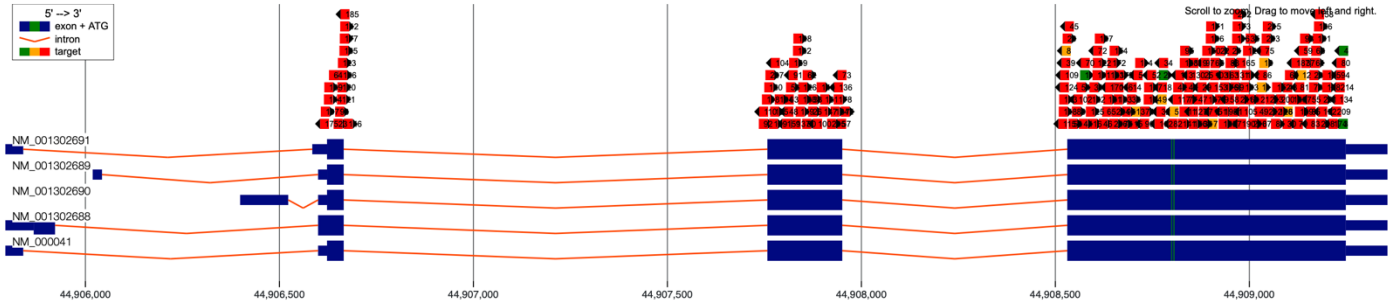
CHOPCHOP

Target: APOE (RefSeq/ENSEMBL/gene ID or genomic coordinates.)
In: Homo sapiens (hg38/GRCh38) (Add new species.)
Using: CRISPR/Cas9 (Change default PAM and guide length in Options.)
For: knock-out (Presets can be adjusted in Options.)

Paste Target | Options | Reset Options

Find Target Sites!

APOE



Download results: Please select one

View in UCSC genome browser

Rank	Target sequence	Genomic location	Strand	GC content (%)	Self-complementarity	MM0	MM1	MM2	MM3	Efficiency
1	GGCCTACAAATCGGAACTGGAGG	chr19:44908566	+	55	2	0	0	0	0	63.59
2	ACCTGCGCAAGCTGCGTAAGCGG	chr19:44908769	+	60	1	0	0	0	0	54.43
3	GGCGTTCAGTGATTGTCGCTGGG	chr19:44909233	-	55	0	0	0	0	0	50.48
4	CGGCGTTCAGTGATTGTCGCTGG	chr19:44909234	-	60	0	0	0	0	0	28.62
5	TCTGCAGGTCATCGGCATCGCGG	chr19:44908797	-	60	0	0	0	0	1	70.56

Pair	Left primer coordinates	Left primer	Left primer Tm	Left primer off-targets	Right primer coordinates	Right primer	Right primer Tm	Right primer off-targets	Pair off-targets	Product size
1	chr19:44907696-44907718	TAGGTACTAGATGCCTGGACGG	60.5	0	chr19:44907941-44907962	GACACTCACCTCAGTTCCTGG	59.7	0	0	266

IMPERIAL

Thank you

Guided computational practical: Genes and Genomes
07/10/2024