

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).

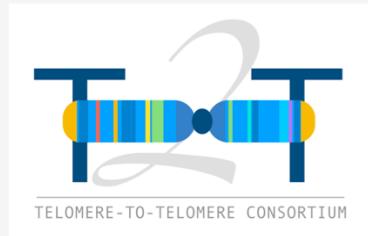
For example, Genome Reference Consortium human 38 patch 14 (**GRCh38.p14**)

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.

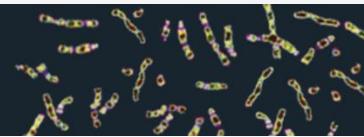
Contribute towards the **Human Pangenome Project**, to generate 350 high quality gapless human genome sequences from ancestrally diverse people.

Reference genomes

The 1000 Genome Project

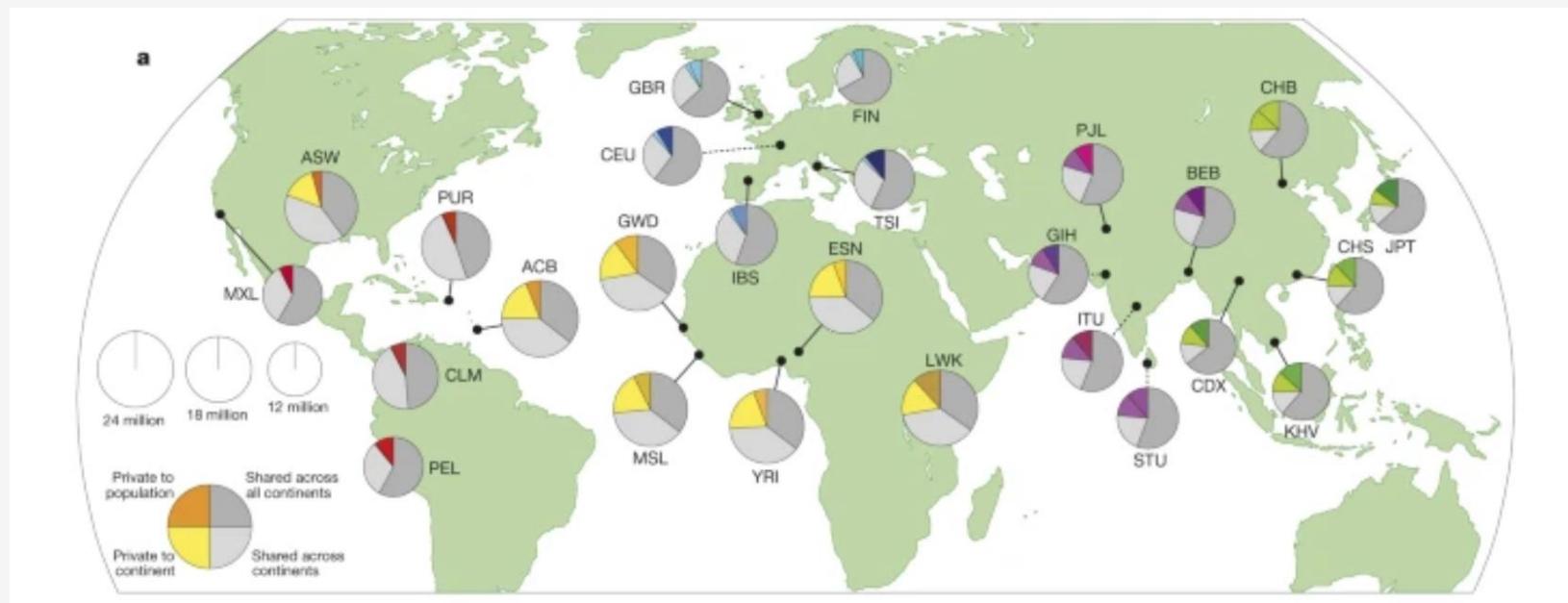
IGSR: The International Genome Sample Resource

Supporting open human variation data



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

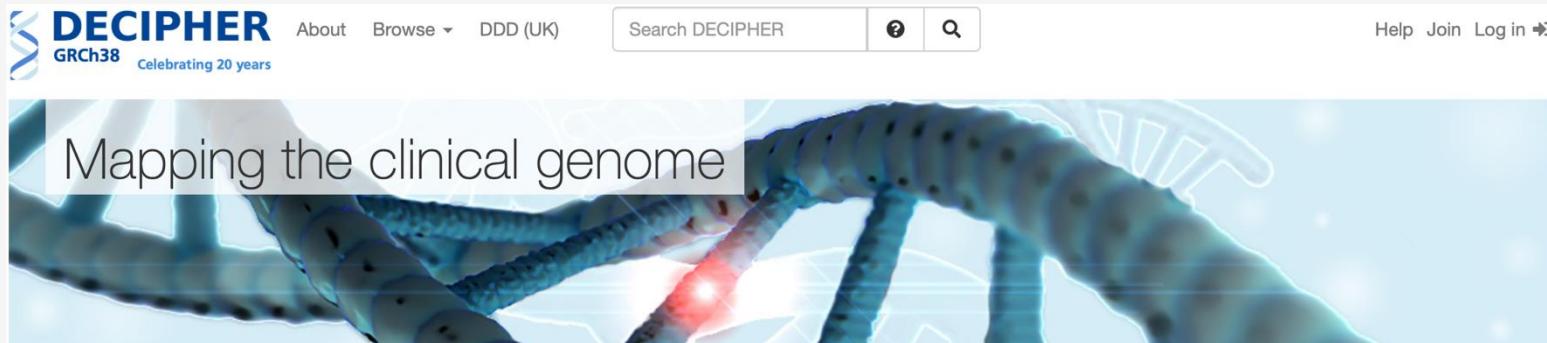
Genomes of 2,504 individuals from **26 global populations** were studied. The consortium identified **>88 million variants** (SNPs, InDels and structural variants). Variants are associated with molecular and disease phenotypes.



Nature 526, 68–74 (2015).

Reference genomes

DECIPHER



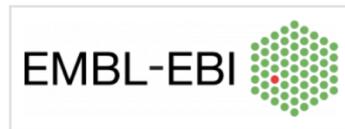
DECIPHER (DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources) is an interactive web-based database which incorporates a suite of tools designed to aid the **interpretation of genomic variants found in rare diseases**.

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.

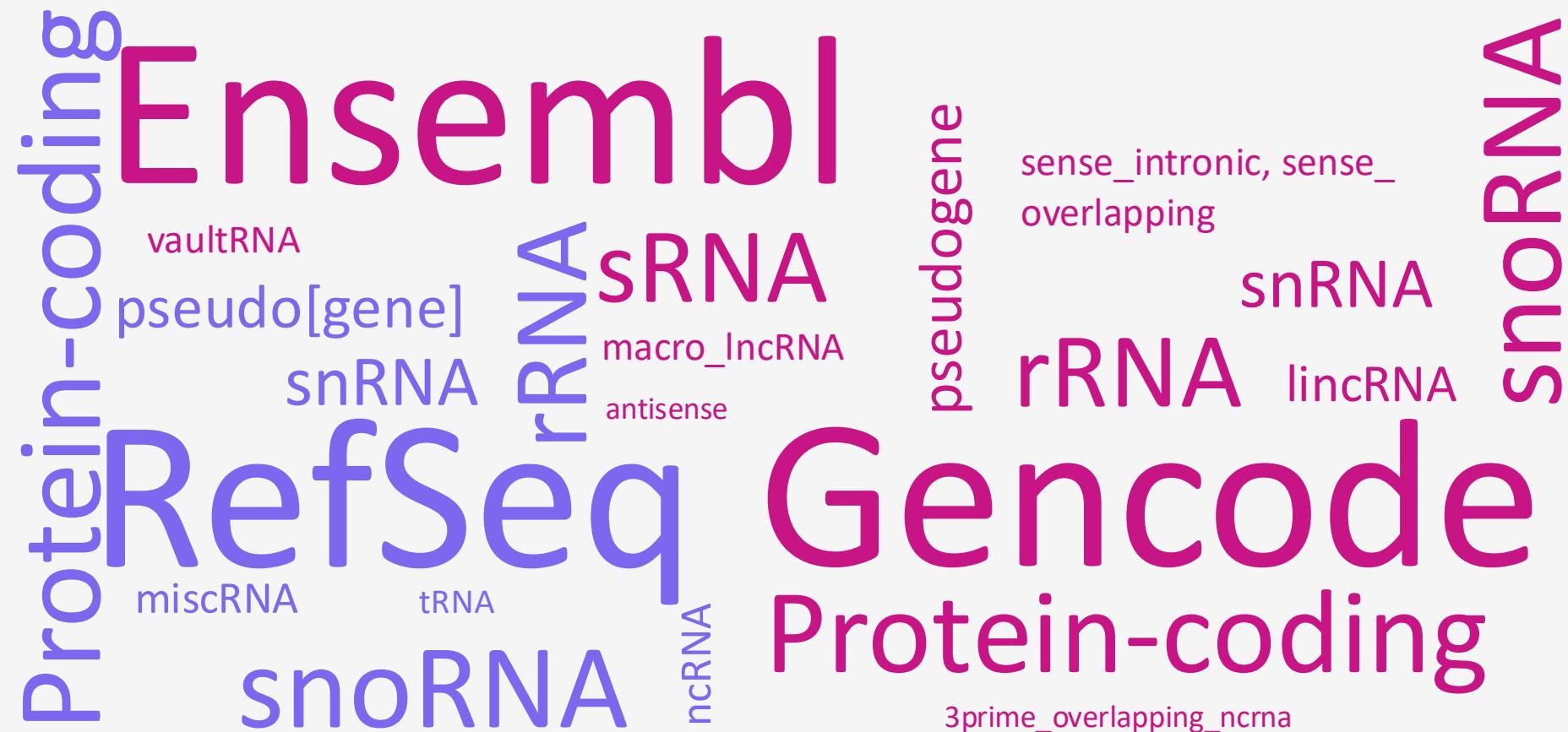
MANE displays:

- 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.

The screenshot shows the Ensembl homepage with a dark blue header. The header includes the Ensembl logo, a search bar with placeholder text "Search all species...", and links for "Login/Register". Below the header, there are four main tool sections: "Tools", "BioMart > Export custom datasets from Ensembl with this data-mining tool", "BLAST/BLAT > Search our genomes for your DNA or protein sequence", and "Variant Effect Predictor > Analyse your own variants and predict the functional consequences of known and unknown variants". A large search bar is centered on the page, with a blue arrow pointing to the input field. Below the search bar, there's a section for "All genomes" with a dropdown menu for selecting a species, and sections for "Favourite genomes" (Human, GRCh38.p14; Mouse, GRCm39; Zebrafish, GRCz11) and "Pig breeds" (Pig reference genome and 12 additional breeds). To the right, there's a "Ensembl Release 112 (May 2024)" section with a list of updates and a "More release news" link. Below that is the "Ensembl Rapid Release" section with a "Go" button and a link to "Rapid Release news". At the bottom, there are six smaller tool boxes: "Compare genes across species" (with a diagram of two organisms), "Find SNPs and other variants for my gene" (with a sequence snippet: GTTATACATTC CTTAAAGTCTT CTTCATAATTGT GCAACATTTCC), "Gene expression in different tissues" (with a tissue sample image), "Retrieve gene sequence" (with a sequence snippet: GCGCTGACTTCCTCGGGTGGT GGGCTGCTTGCGGGCGAGC GGGCTGCTTGCGGGCGAGC ACGGGACAGATTGGTGGAG CACCTCTGAGACGGTTT GCGCAAGTCGAAGCGTGGCG), "Find a Data Display" (with arrows pointing to TABLE, HEATMAP, SEQUENCE, and PIE CHART), and "Use my own data in Ensembl" (with a small chart image).

Genome Browsers

Ensembl example

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

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

New Search

Restrict category to:

Gene	548
Transcript	316
Variant	1485
Phenotype	60
GeneTree	88
GenomicAlignment	66437
ProbeFeature	7
Protein Domain	945

apoe 

69886 results match apoe

[APOE \(Human Gene\)](#)
ENSG00000130203 19:44905791-44909393:1
Apolipoprotein E [Source:HGNC Symbol;Acc:HGNC:613]

[APOLIPOPROTEIN E; APOE](#) [*107741] (MIM gene record; description: APOLIPOPROTEIN E; **APOE**) is an external reference matched to Gene ENSG00000130203
Variant table • Phenotypes • Location • External Refs. • Regulation • Orthologues • Gene tree

[APOE carrier status \(Human Phenotype\)](#)
Human Phenotype
APOE carrier status.

Restrict species to:

Human	2365
Mouse	9733
Zebrafish	579
Abingdon island giant tortoise	8
African green monkey	7
African ostrich	1
... 268 more species ...	

APOE-201 (Human Transcript)
ENST00000252486 19:44905796-44909393:1
Apolipoprotein E [Source:HGNC Symbol;Acc:HGNC:613].
Location • External Refs. • cDNA seq. • Exons • Variant table • Protein seq. • Population • Protein summary

APOE-203 (Human Transcript)
ENST00000434152 19:44905812-44909025:1
Apolipoprotein E [Source:HGNC Symbol;Acc:HGNC:613].
Location • External Refs. • cDNA seq. • Exons • Variant table • Protein seq. • Population • Protein summary

APOE-202 (Human Transcript)
ENST00000425718 19:44906360-44908954:1
Apolipoprotein E [Source:HGNC Symbol;Acc:HGNC:613].
Location • External Refs. • cDNA seq. • Exons • Variant table • Protein seq. • Population • Protein summary



Genome Browsers

Ensembl example

Click on the Apoe gene entry to open the ‘Gene Tab’

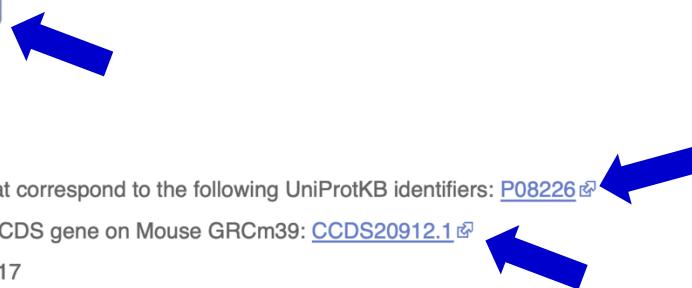
Gene: Apoe

Gene: Apoe ENSMUSG00000002985

Description	apolipoprotein E [Source:MGI Symbol;Acc: MGI:88057]
Gene Synonyms	Apoe
Location	Chromosome 7: 19,430,034-19,433,113 reverse strand. GRCm39:CM001000.3
About this gene	This gene has 11 transcripts (splice variants), 219 orthologues , 3 paralogues and is associated with 190 phenotypes .
Transcripts	Show transcript table

Summary

Name	Apoe (MGI Symbol)
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: P08226
CCDS	This gene is similar to a CCDS gene on Mouse GRCm39: CCDS20912.1
Ensembl version	ENSMUSG00000002985.17
Gene type	Protein coding
Annotation method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article .



Click on “Show transcript table”

Genome Browsers

Ensembl example

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

The screenshot shows a transcript table from the Ensembl genome browser. The columns include Transcript ID, Name, bp, Protein, Biotype, CCDS, UniProt Match, and Flags. A tooltip is displayed over the 'Ensembl Canonical' flag in the Flags column, providing a detailed explanation of what it means for a transcript to be canonical. The tooltip text is as follows:

A single transcript chosen for a gene which is the most conserved, most highly expressed, has the longest coding sequence and is represented in other key resources, such as NCBI and UniProt. This is defined in detail on http://www.ensembl.org/info/genome/genebuild/canonical_transcripts.html

Show/hide columns (1 selected)	Name	bp	Protein	Biotype	CCDS	UniProt Match	Flags
ENSMUST00000174064.9	Apoe-206	1408	311aa	Protein coding	CCDS20912	P08226 Q3TXU4	Ensembl Canonical GENCODE basic APPRIS P1 TSL:1
ENSMUST00000173739.8	Apoe-205	1221	311aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST0000003066.16	Apoe-201	1128	311aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000174355.8	Apoe-209	1104	311aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000172983.8	Apoe-204	817	232aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000174144.8	Apoe-207	816	231aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000174710.2	Apoe-210	514	71aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000174191.2	Apoe-208	472	71aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000172808.2	Apoe-203	456	146aa	Protein coding	CCDS20912	P08226 Q3TXU4	GENCODE basic APPRIS P1 TSL:1
ENSMUST00000167646.9	Apoe-202	728	No protein	Protein coding CDS not defined		-	TSL:1
ENSMUST00000207525.2	Apoe-211	897	No protein	Retained intron		-	TSL:NA

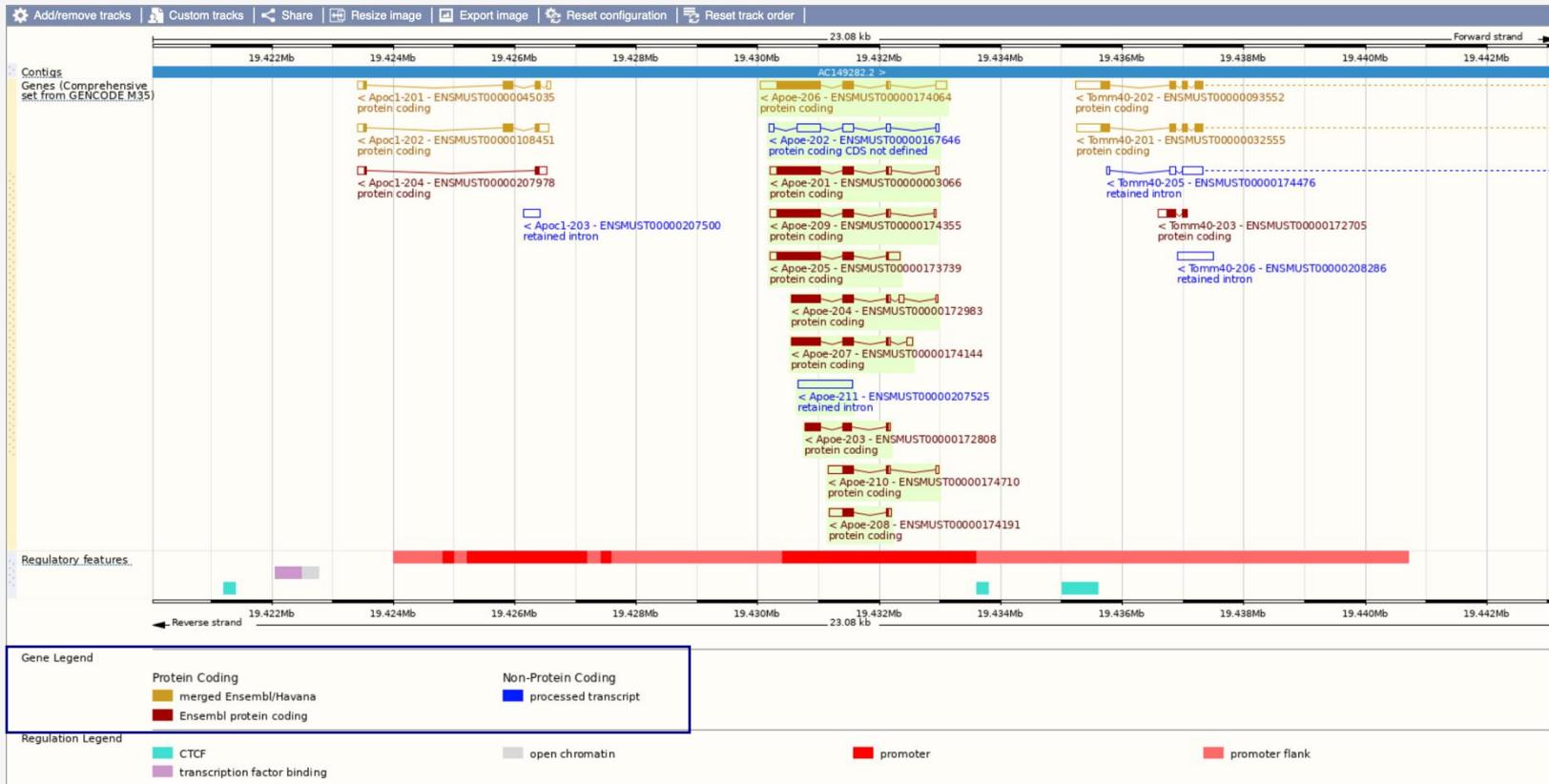
Position the cursor over a flag to view more information.

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

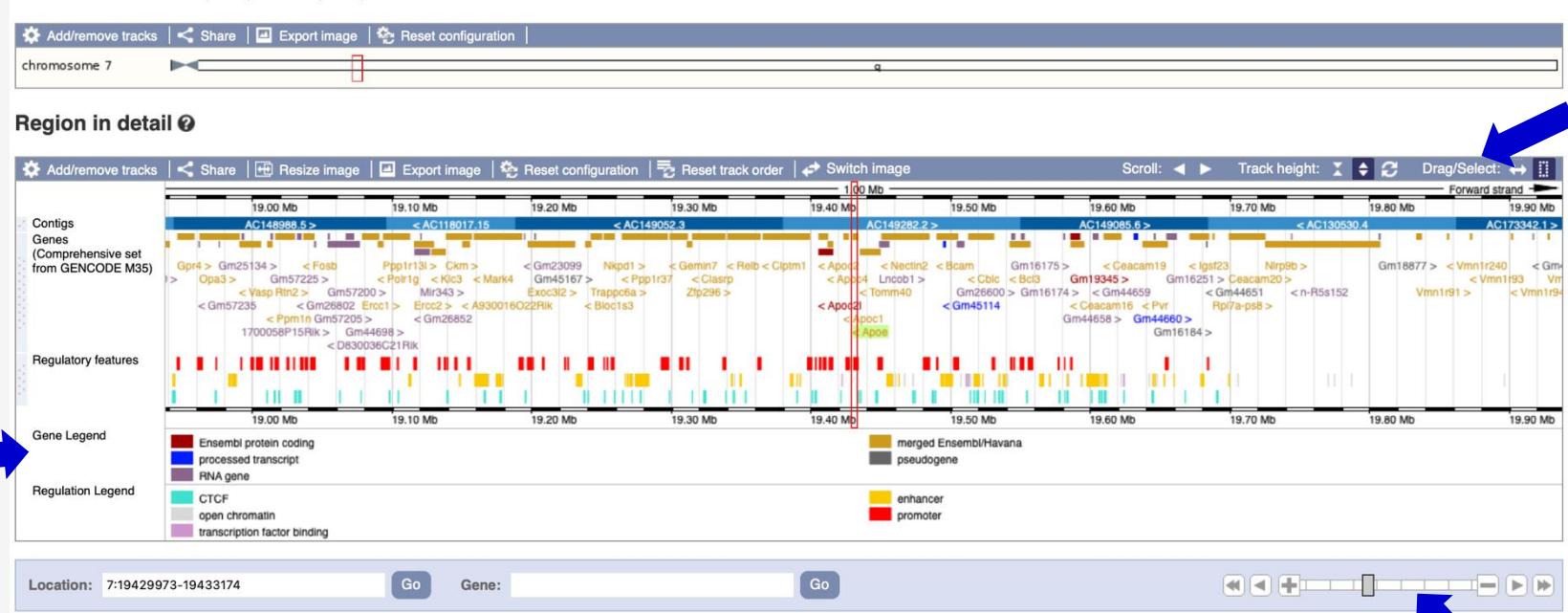
Click on 'Region in detail'.



Go to [Region in Detail](#) for more tracks and navigation options (e.g. zooming)

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

Chromosome 7: 19,436,431-19,439,632



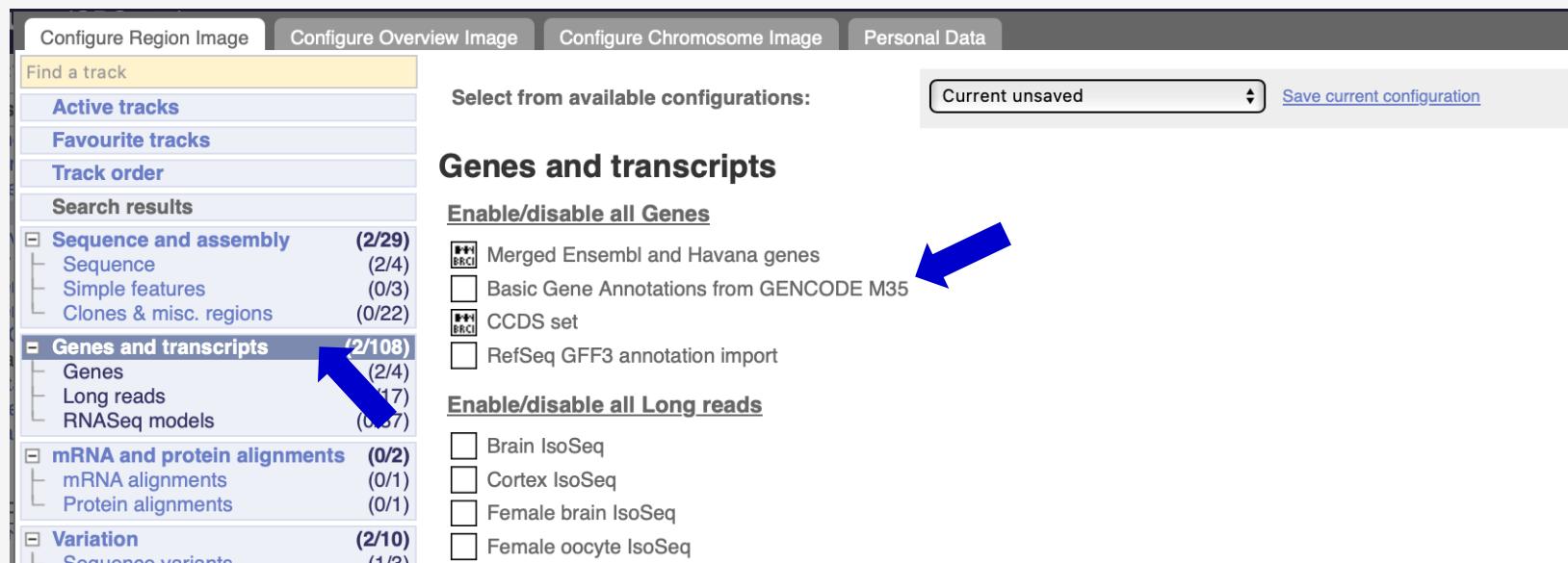
Genome Browsers

Ensembl example

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

 Configure this page

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



The screenshot shows the 'Configure this page' interface for Ensembl. The left sidebar has a 'Find a track' section and a tree view of available configurations:

- Sequence and assembly (2/29)
 - Sequence (2/4)
 - Simple features (0/3)
 - Clones & misc. regions (0/22)
- Genes and transcripts (2/108)**
 - Genes (2/4)
 - Long reads (1/17)
 - RNASeq models (0/37)
- mRNA and protein alignments (0/2)
 - mRNA alignments (0/1)
 - Protein alignments (0/1)
- Variation (2/10)
 - Sequence variants (1/3)

The main area shows configuration options for 'Genes and transcripts':

Enable/disable all Genes

- Merged Ensembl and Havana genes
- Basic Gene Annotations from GENCODE M35
- CCDS set
- RefSeq GFF3 annotation import

Enable/disable all Long reads

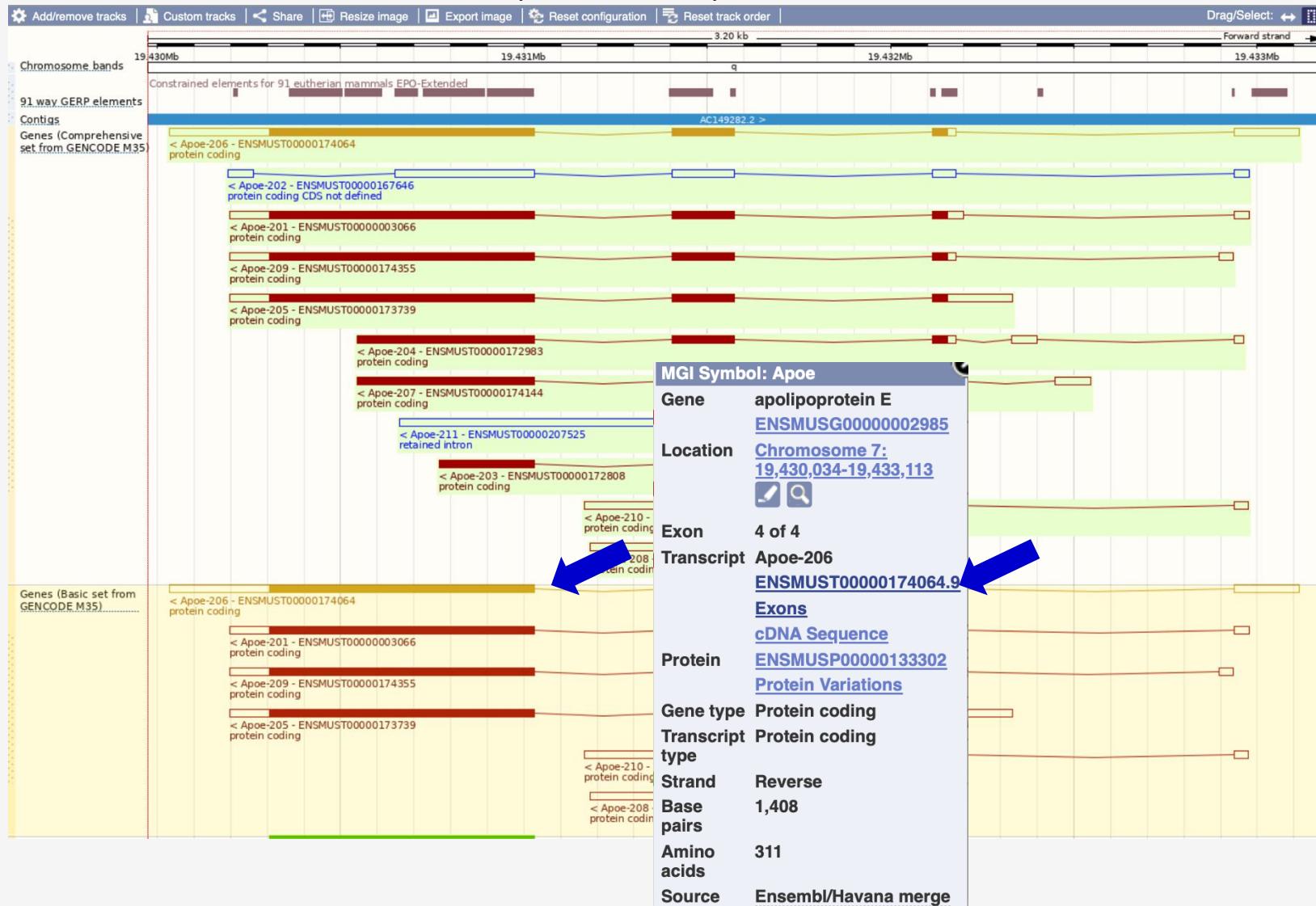
- Brain IsoSeq
- Cortex IsoSeq
- Female brain IsoSeq
- Female oocyte IsoSeq

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.

Transcript-based displays

- Summary
- Sequence
 - Exons
 - cDNA
 - Protein
- Protein Information
 - Protein summary
 - Domains & features
 - Variants
 - PDB 3D protein model
 - AlphaFold predicted model
- Genetic Variation
 - Variant table
 - Variant image
 - Population comparison
 - Comparison image
- External References
 - General identifiers
 - Oligo probes
 - Supporting evidence
- ID History
 - Transcript history
 - Protein history

Transcript: ENSMUST00000174064.9 Apoe-206

Description apolipoprotein E [Source:MGI Symbol;Acc:[MGI:88057](#)]

Gene Synonyms Apoe

Location Chromosome 7: 19,430,034-19,433,113 reverse strand.

About this transcript This transcript has 4 exons, is annotated with 14 domains and features, is associated with 655 variant alleles and maps to 259 oligo probes.

Gene This transcript is a product of gene [ENSMUSG0000002985.17](#) [Show transcript table](#)

Summary

Export image |

< Apoe-206 - ENSMUST00000174064
protein coding

Reverse strand ————— 3.08 kb

Statistics Exons: 4, Coding exons: 3, Transcript length: 1,408 bps, Translation length: 311 residues

Uniprot This transcript corresponds to the following Uniprot identifiers: [P08226](#)

CCDS This transcript is a member of the Mouse CCDS set: [CCDS20912](#)

Transcript Support Level (TSL) TSL:1

Version ENSMUST00000174064.9

Type Protein coding

Annotation Method Transcript where the Ensembl genebuild transcript and the Havana manual annotation have the same sequence, for every base pair. See [article](#).

GENCODE basic gene This transcript is a member of the [Gencode basic](#) gene set.

Ensembl release 112 - May 2024 © EMBL-EBI

Permanent link - [View in archive site](#)

Genome Browsers

Ensembl example

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

Exons

Download sequence

Exons/ Introns Translated sequence Flanking sequence Intron sequence UTR

Variants 3 prime UTR 5 prime UTR Frameshift Missense Stop gained Synonymous

Markup loaded

Show/hide columns Filter

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
1	ENSMUSE00000902063	19,433,113	19,432,936	-	-	178gttgcggggaggggacgggggtacaaggcatcaaactcaccc TTTCCTGCCCCCTGCTGTGAAGGGGAGAGACAACCCGCCCTCGTGACAGGGGCCTGGCA CAGCCGGCCCTAGCTTGAAGGGGGGGAGCAGGGGAGTCCATATAATTGACCGTC TGGGATCCGATCCCCCTGCTGACAGCCCTGAGGCTAAAGGACTTGTTCGAAGAGGCTG
2	Intron 1-2	19,432,935	19,432,179			757	gtaaagacaagtgggtgggatt.....tcaaagacaattttccctccgca ACTGGCCAATGAAATGCGAAAGATGAAGGCTGTGTCCTGCTGTTGGTGTGATGCG TGACAG
2	ENSMUSE00001380292	19,432,178	19,432,113	-	1	66	
3	Intron 2-3	19,432,112	19,431,573			540	gtatggagcaaggacttgcgtgg.....ccagccttaaacttactctacacag GATGCCCTAGCCGAGGGAGGGCGGGTGCAGATCAGCTCGAGTGGCAAAGAACCAAC
3	ENSMUSE00000231320	19,431,572	19,431,404	1	2	169	CCTGGAGCAGGCCCTGAAACCGTTCTGGGATTACCTGGCTGCGTCAAGGCT ACCAGGTCAGGAAGAAGTCAAGACTCAGACAAAGAACGTACACAAAGAACGTAC
4	Intron 3-4	19,431,403	19,431,029			375	gttgatgttcacgttt GGCACATGATGGAGGACACTATGAGAAAGTAAGGCTTACAAAAGGAGCGCGAAC GCTGGGTCCTGGCGAGGAGAGACGGGGCAGCTGGGGAGAGAGAGTGCAAGGGCAAC
4	ENSMUSE00000931635	19,431,028	19,430,034	2	-	995	GGCCGCACTCGGGCGACATGGAGGGATCTACGGCAACGACTCGGGCAGTACCCCAACGA GGTGCAACACCATGCTGGCCAGAGCAGAGGATACGGGGGGGCTCTCCACACACT GCGCAAGATGCCAACCGCTTGTGCGGATGCGCGTCCCGCGATCTGCAAGGCGCT GTACAAGGCAAGGGCACCGGAAGGGCGGGCGGTGACTGCCAAGCTGAGGGCT GGGGCTCTGGTGGAGCACAGGTGCCCGCCACCGCAACCTAGGGCTGGGGCCCA GCCCTGGCGGATGCCCGGAGGCTTGGTGAUCGCGATCGAGGGGGCTGAGGAAGT GGGAACCAAGGGCCCGTGAAGGGCGCTAGAGAGTGGAGACATGGAGGGGTGGCTC CAAGATGGAGAACAGACCCAGCAAAATACCGCTGGAGGGAGATCTCCAGGCCGCT CAAGGCTGTTGGAGCCAAATGGAAGACATGCGCAGTGGCAAACTGTGAGGAA GAAGATACAGGCTCTGGCTTACACAGCTCATACACCCAGTCCTCAAGAGAATCA ATGAGATACCTCTCCCTGCTTCGAACATCATATCCAGCCAGGTGGCCCTGTC AGCACCTCTGGCCCTCTGGTGGCCCTGCTTAATAAAGATCTCCGGACATCTGA GTCTCTGTGAGTATTCCAAATCAGCTTCAGCTGAGTATTGTTTTTGCTTACCTAG CACACATTCCATGGCCCTGTCACTATCTGTAGAGGGAGGTGGTTTTGAGCAATAGA GAAGCTTAGGACCTAACACATAAAAGAAACAGTGGT
	3' downstream sequence						atccatccactgagccacgccccacagccccctcaactggggattctaggcag.....

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.

Transcript: ENSMUST00000174064.9 Apoe-206

Description	apolipoprotein E [Source:MGI Symbol;Acc: MGI:88057]
Gene Synonyms	Apoe
Location	Chromosome 7: 19,430,034-19,433,113 reverse strand.
About this transcript	This transcript has 4 exons , is annotated with 14 domains and features , is associated with 655 variant alleles and maps to 259 oligo probes .
Gene	This transcript is a product of gene ENSMUSG0000002985.17 Show transcript table

Protein summary

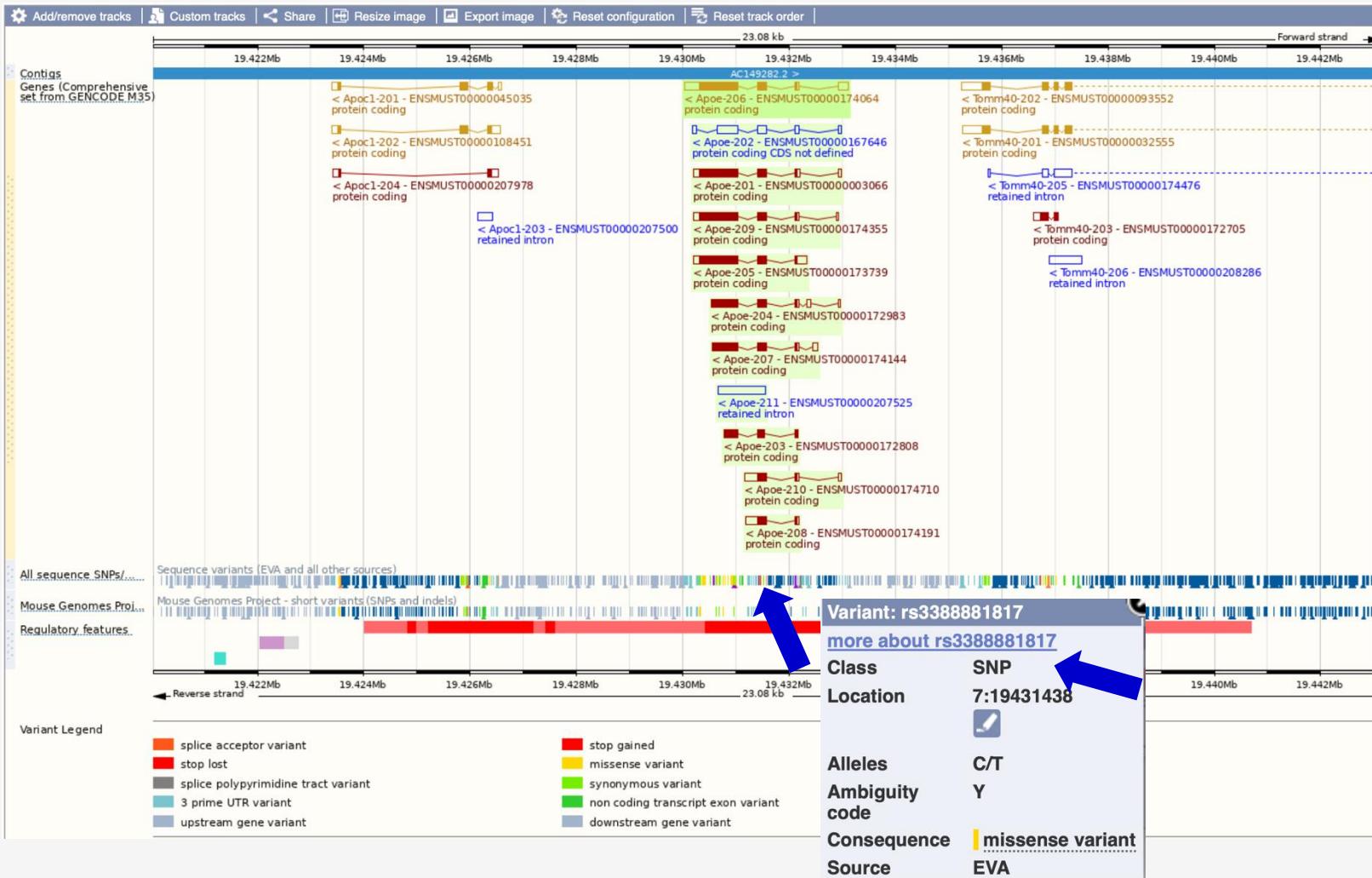
Protein domains for ENSMUSP00000133302.2



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).

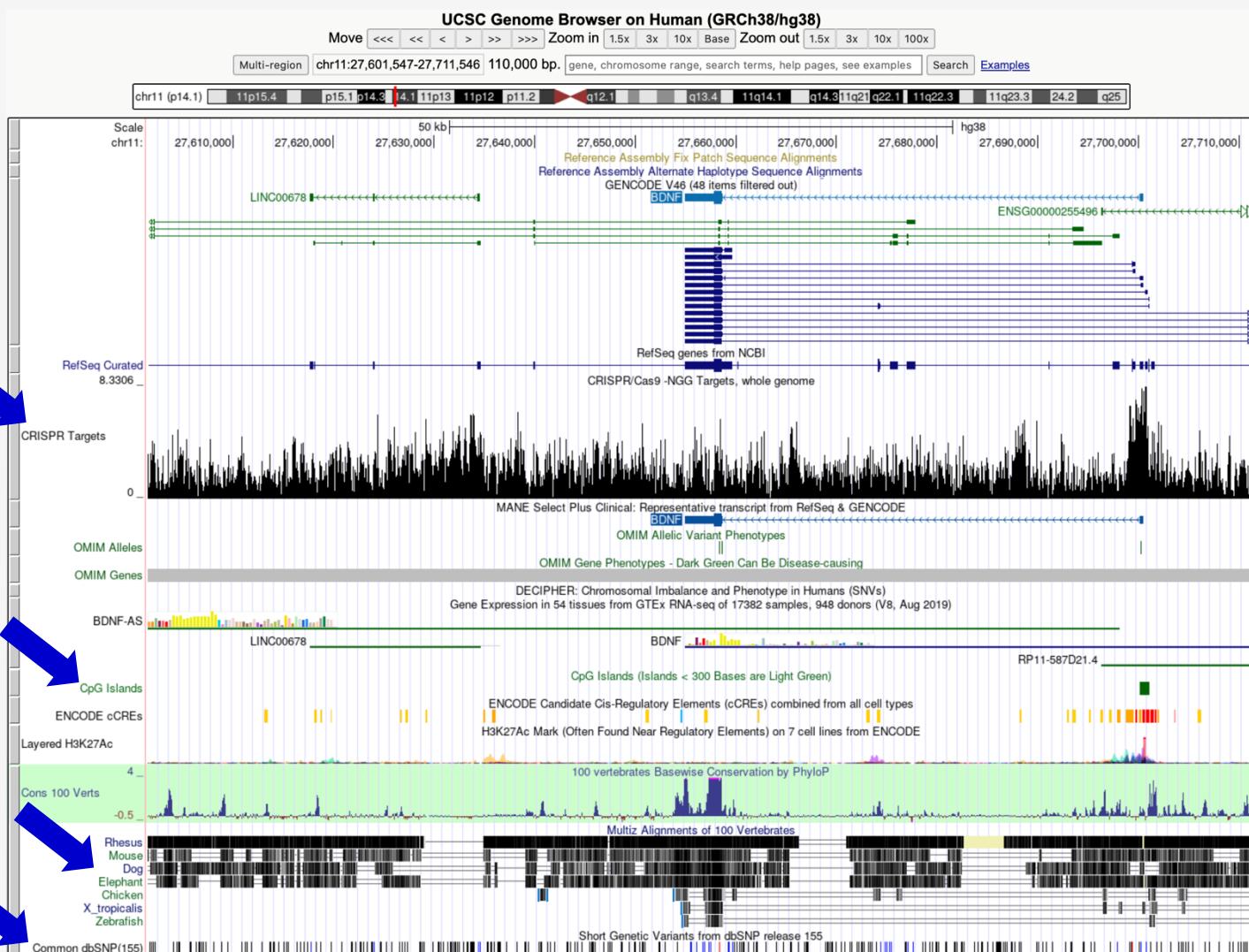
https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_cpgplot

The screenshot shows the EMBOSS CpGPlot web interface. At the top, there is a header with the title "EMBOSS CpGplot" and a sub-section "Sequence Statistics (seqstats)". Below the header, there is a navigation bar with links: "Job Dispatcher", "Help & Privacy", "Your Jobs", "Input form" (which is highlighted in blue), and "Feedback". The main content area has a sub-header "Identify and plot CpG islands in nucleotide sequence(s.)". There are two input fields: "Input sequence" (with a "Choose File" button) and "Paste your sequence here - or use the example sequence" (with a text area). Below these fields are buttons: "Choose File" (disabled), "Use the example", "Clear sequence", and "More example inputs". At the bottom left, there is a "Parameters" section with a "More options" dropdown. At the bottom right, there is a "Submit" section with a "Title" field containing the text "EMBOSS CpGplot's job".

Genome Browsers

Other browsers

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

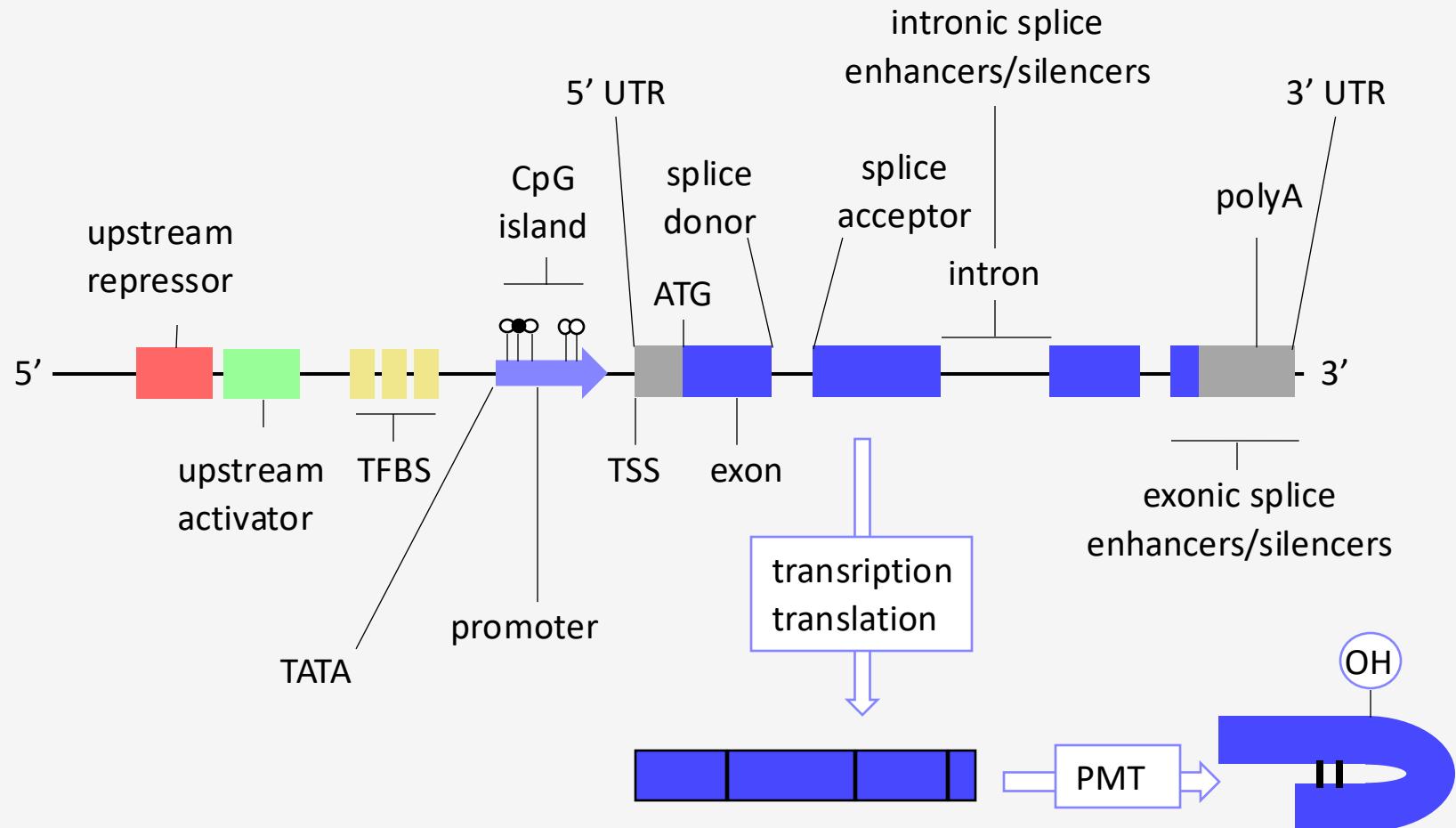


Gene Structure

Gene Structure

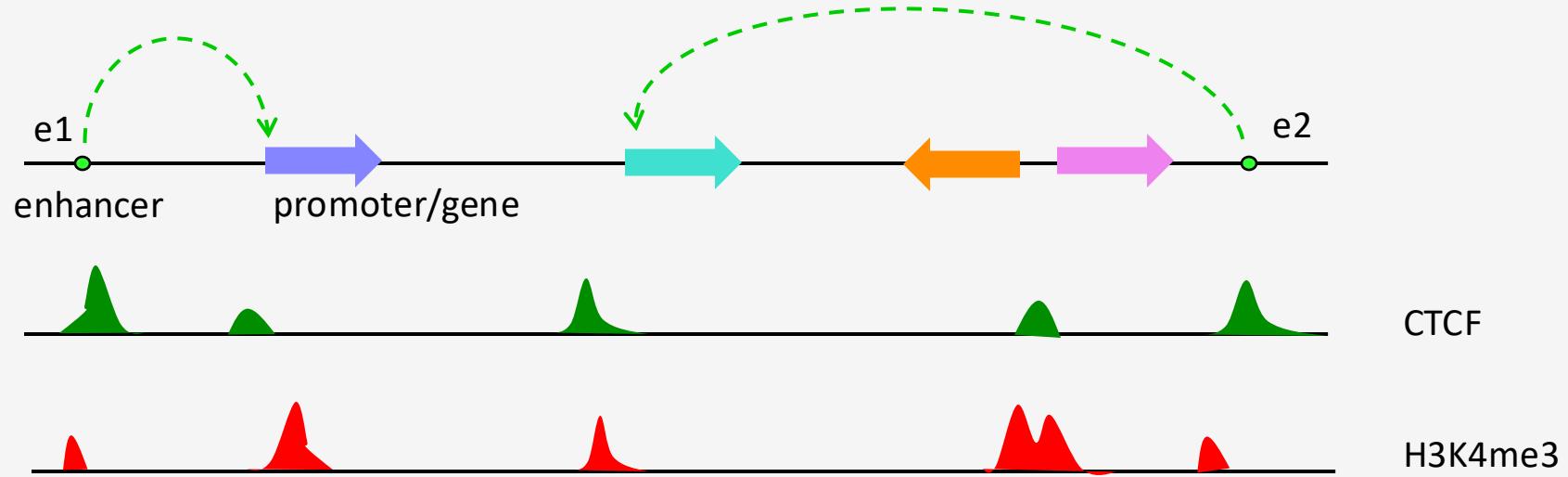
What is a gene?

Protein coding or non-protein coding DNA sequences.

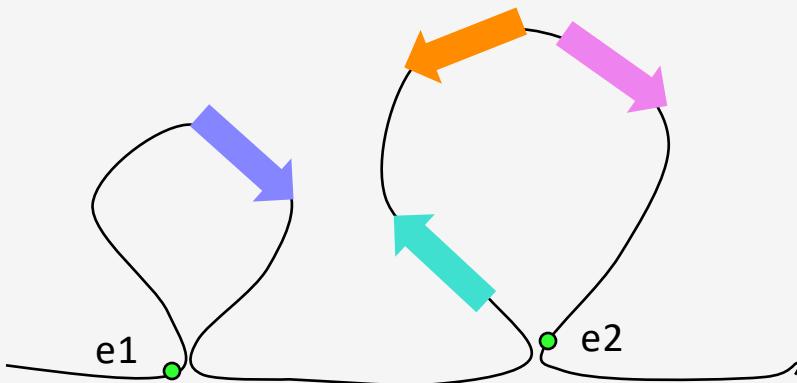


Gene Structure Enhancers and Repressors

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



1D locations are really 3D interactions

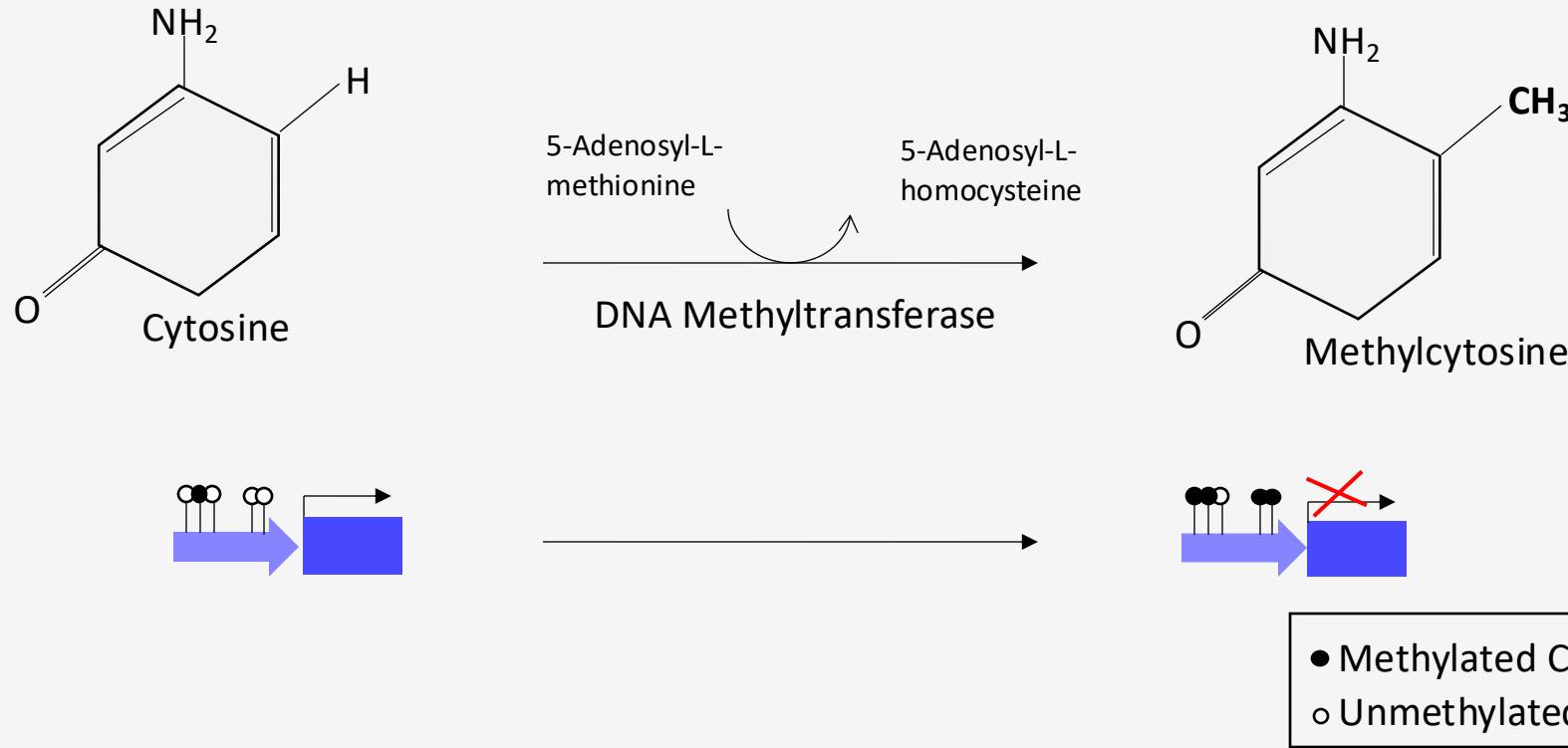


Gene Structure

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)

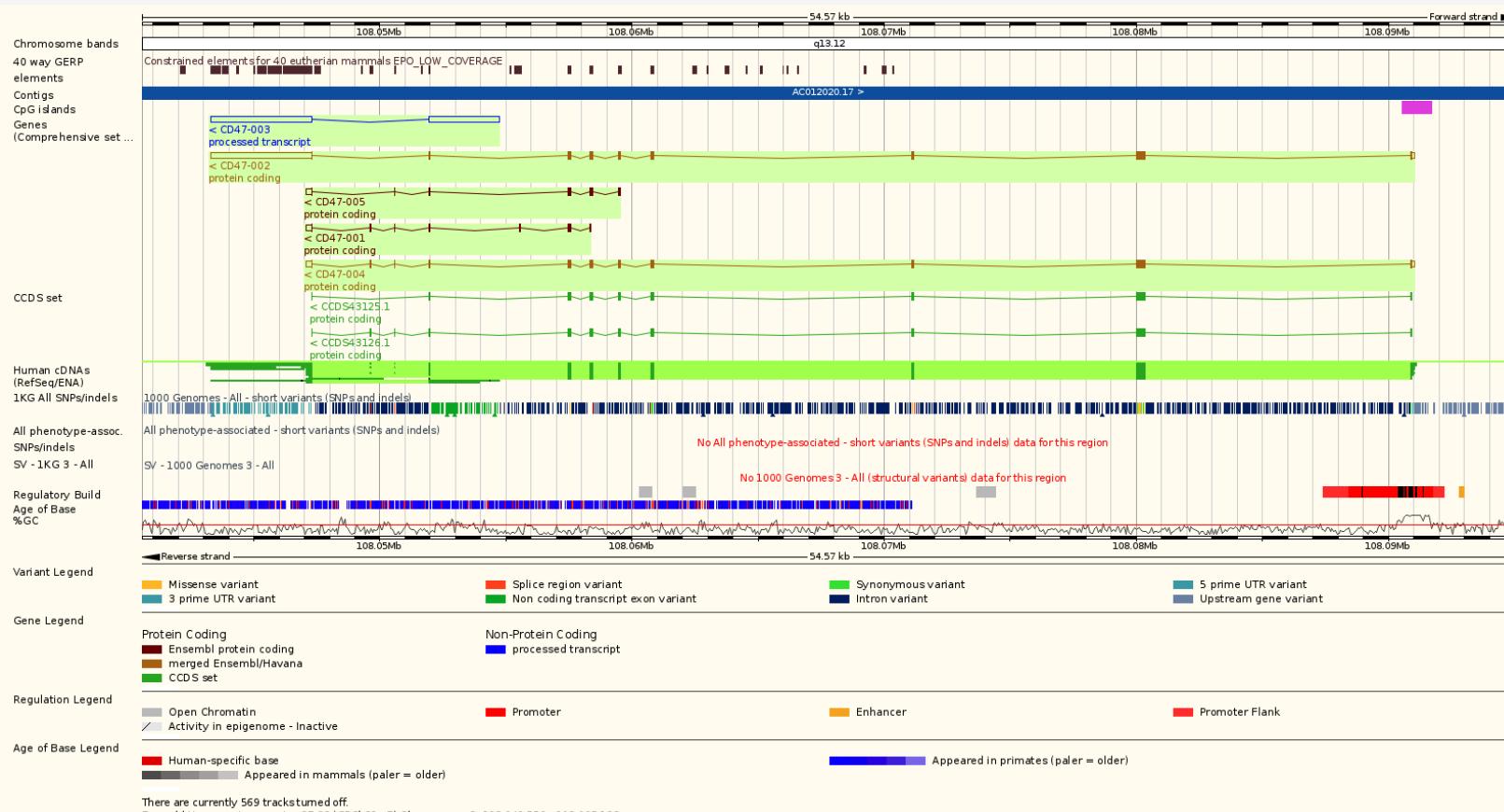
Different 5' UTR isoforms associated with

- 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.

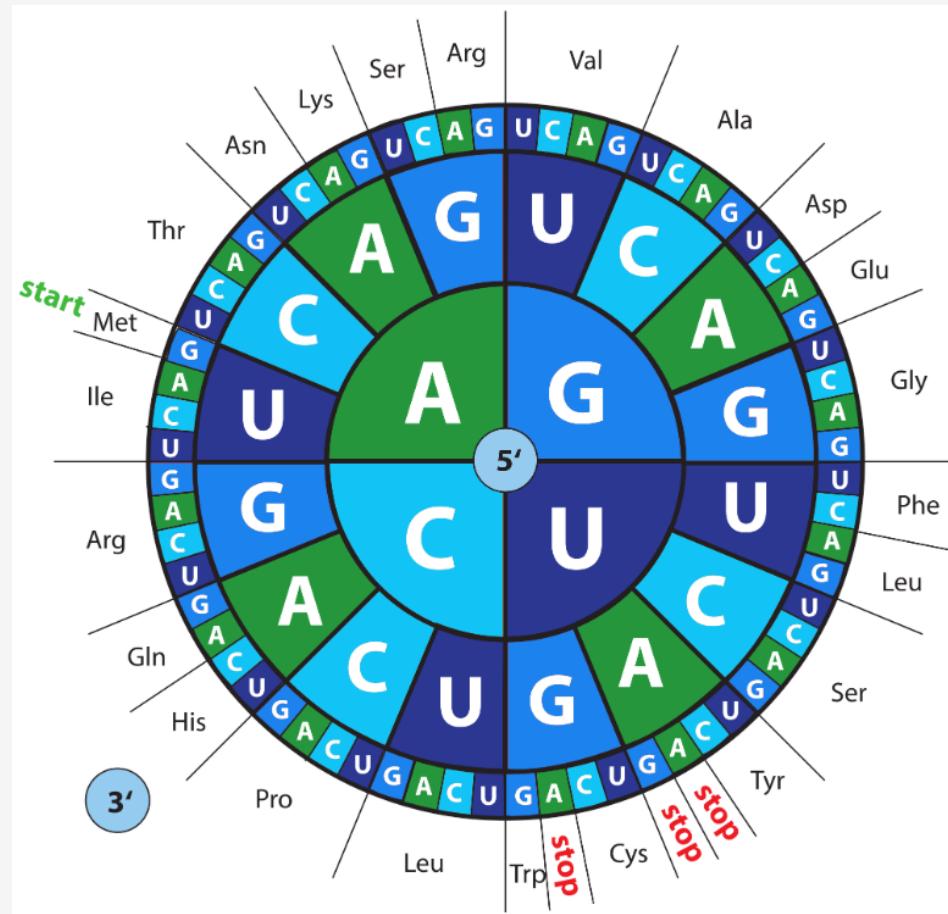
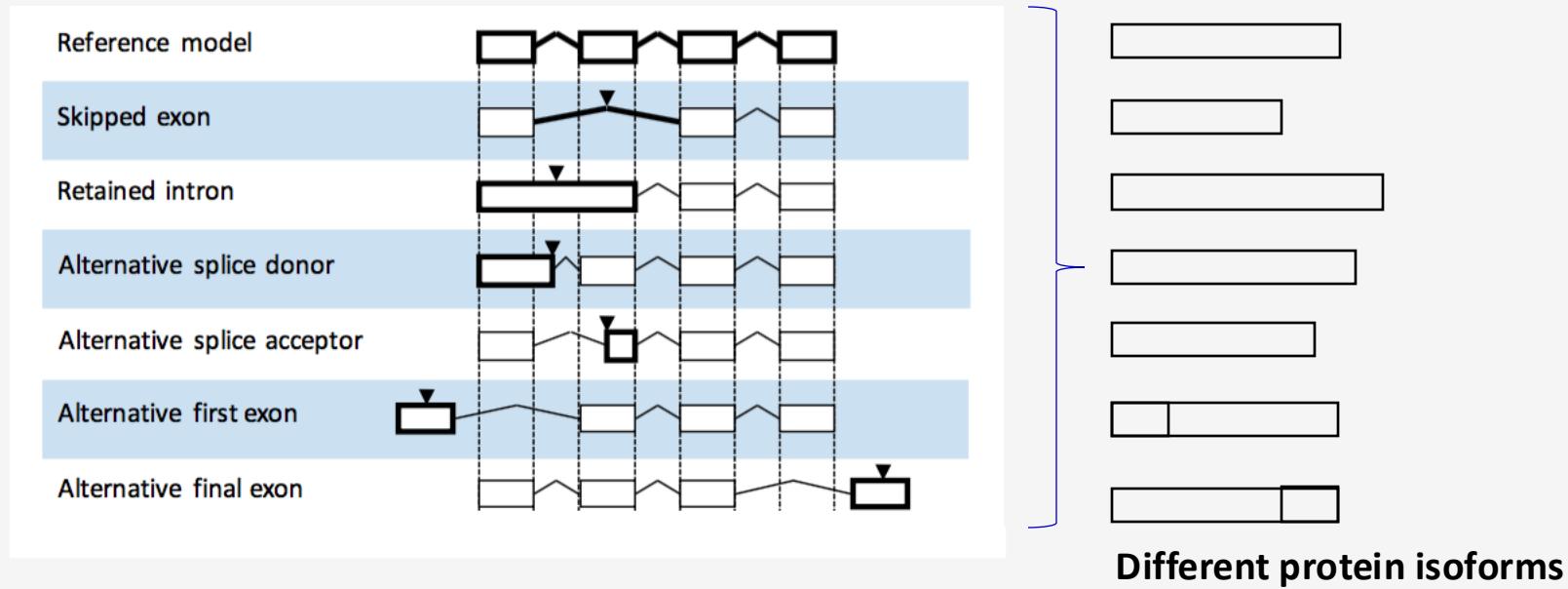


Image taken from Eurofins

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.



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

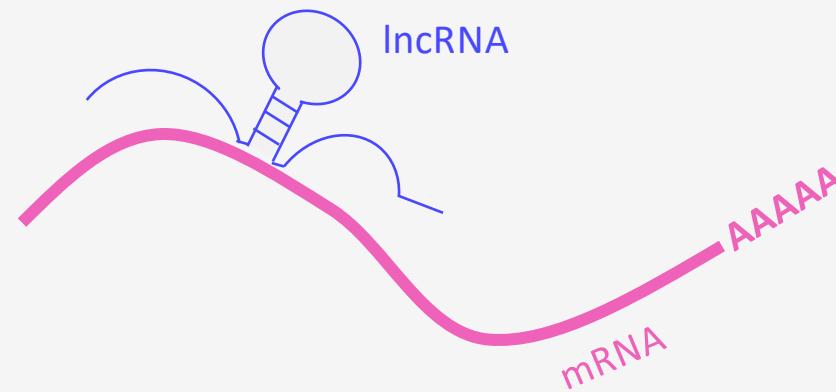
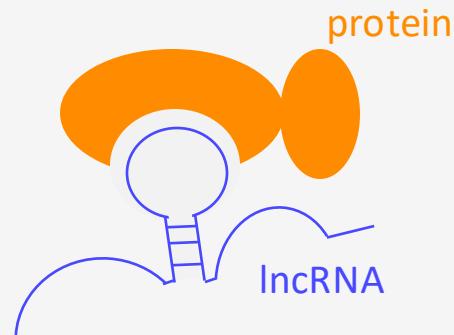
Translation re-initiation

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.

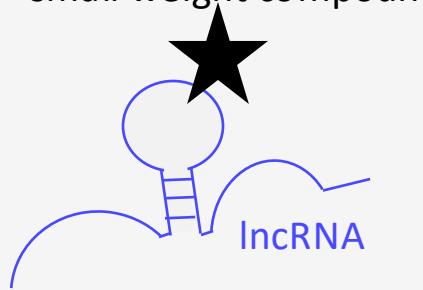
Translation re-initiation is likely when a STOP codon <35 aa away from initiating methionine. It is also determined by the rate at which the translation initiation complex dissociates.

Gene Structure

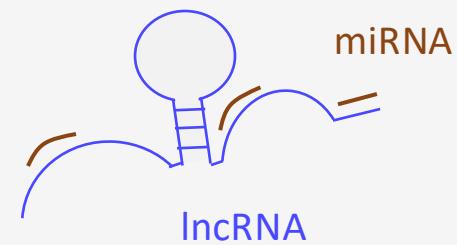
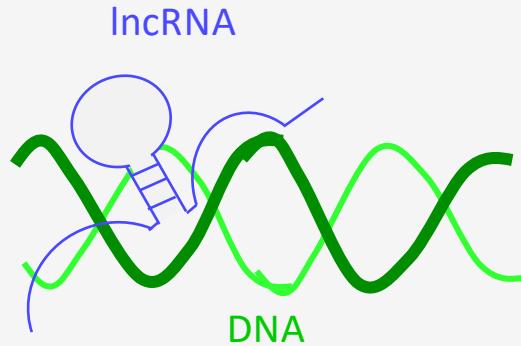
Long non-coding RNA (lncRNA)



small weight compounds



lncRNA interactome



CRISPR Design Process and Tools

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

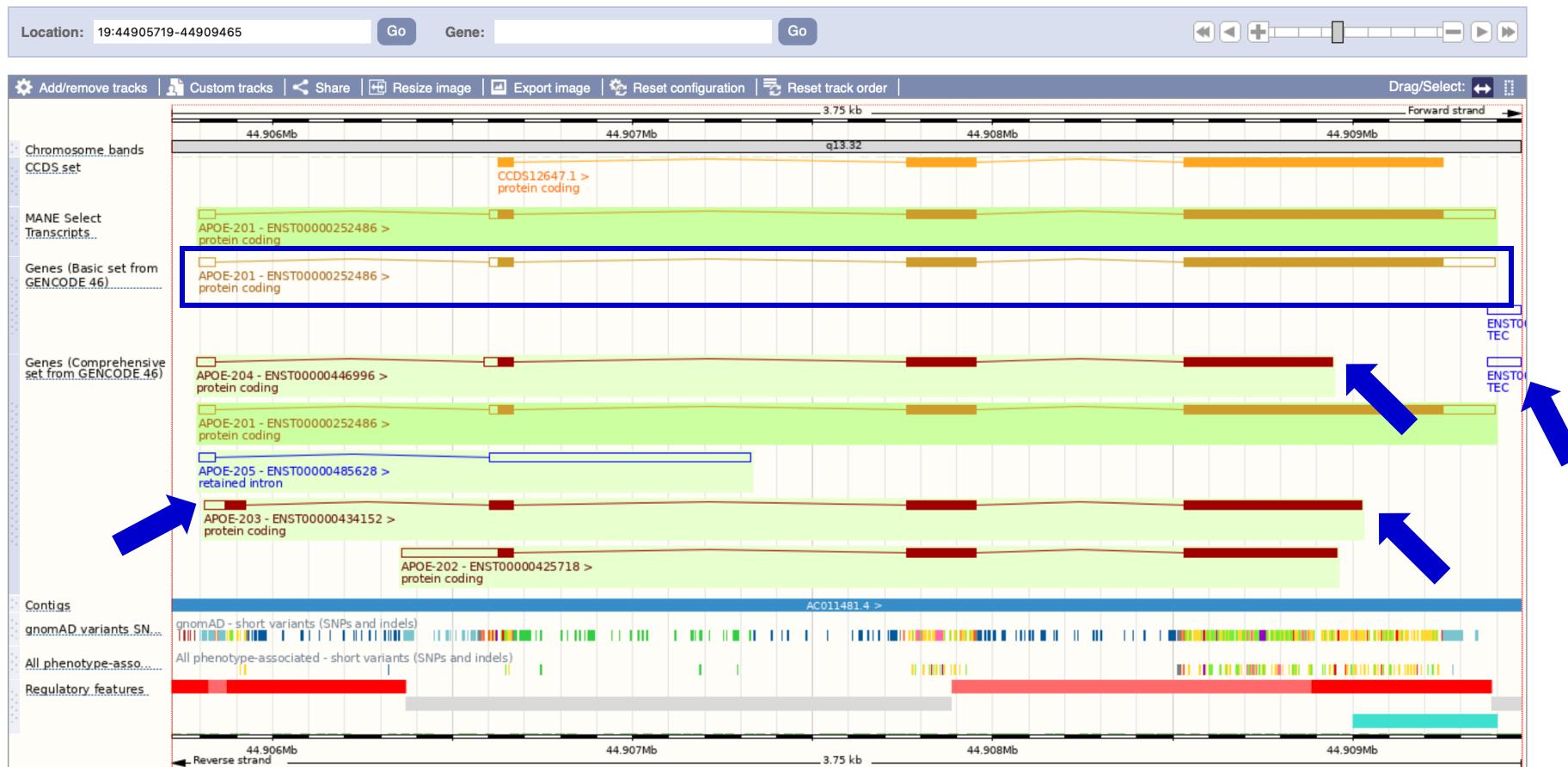
Are there alternative **translation start sites**?

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

Which transcript is most representative?

CRISPR Design Process

Basic vs Comprehensive gene sets.



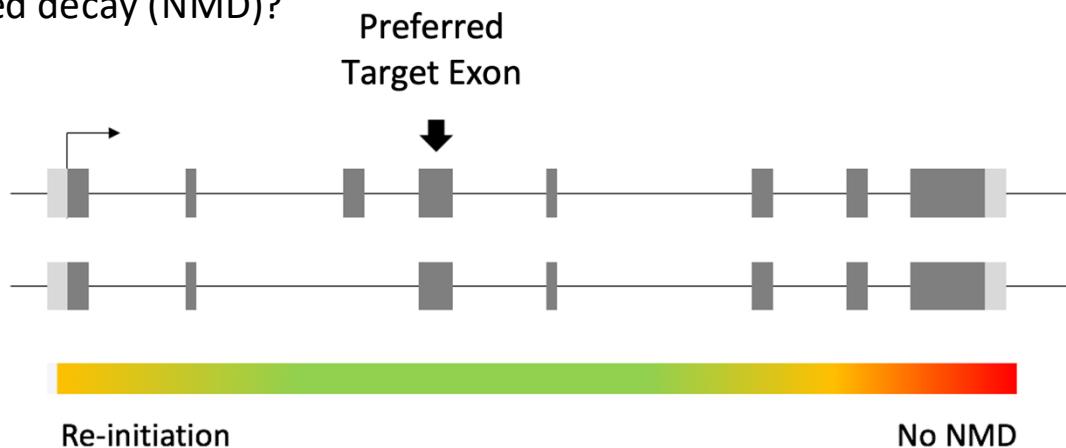
CRISPR Design Process and Tools

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.

Location: 19:44,905,716-44,909,462 Gene: APOE Transcript: APOE-201

Transcript-based displays

- Summary
- Sequence
 - Exons
 - cDNA
 - Protein
- Protein Information
 - Protein summary
 - Domains & features
 - Variants
 - PDB 3D protein model
 - AlphaFold predicted model
- Genetic Variation
 - Variant table
 - Variant image
 - Haplotypes
 - Population comparison
 - Comparison image
- External References
 - General identifiers
 - Oligo probes
 - Supporting evidence
- ID History
 - Transcript history
 - Protein history

Summary

Export image

APOE-201 - ENST00000252486 > protein coding

Statistics

Exons: 4, Coding exons: 3, Transcript length: 1,166 bps, Translation length: 317 residues

This MANE Select transcript contains [ENSP00000252486](#) and matches to [NM_000041.4](#) and [NP_000032.1](#).

This transcript corresponds to the following Uniprot identifiers: [P02649](#).

This transcript is a member of the Human CCDS set: [CCDS12647](#).

TSL:1

ENST00000252486.9

Version

Type

Protein coding

Annotation Method

Transcript where the Ensembl genebuild transcript and the Havana manual annotation have the same sequence, for every base pair. See [article](#).

This transcript is a member of the [Gencode basic](#) gene set.

Configure this page

Custom tracks

Export data

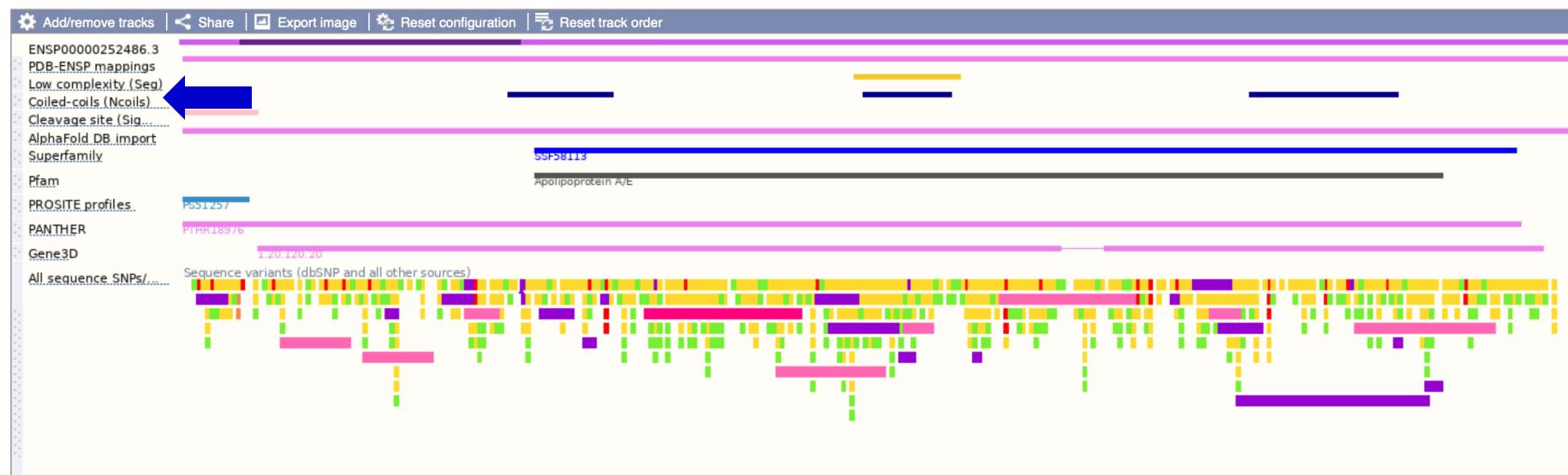
Share this page

Bookmark this page

CRISPR Design Process and Tools

Protein summary

Protein domains for ENSP00000252486.3



Other features

Show	All	entries	Show/hide columns
Feature type			Start
Seg			154
SignalP			1
Alphafold			1
Ncoils			75
Ncoils			156
Ncoils			244

CRISPR Design Process and Tools

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

CRISPR Design Process and Tools

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”.

The screenshot shows the 'Configure Region Image' tab selected in the top navigation bar. The main content area is titled 'Genome targeting'. On the left, there's a sidebar with a tree view of genomic data categories. The 'Genome targeting' category is expanded, showing a single track named 'CRISPR SpCas9'. A blue arrow points to this track. Another blue arrow points to the 'Genome targeting' section in the sidebar. The right side of the interface includes a 'Key' section with icons for track style, external data, forward strand, reverse strand, favourite track, and track information. There are also notes about external tracks and URL-based tracks.

Configure Region Image Configure Overview Image Configure Chromosome Image Personal Data

Find a track

Active tracks
Favourite tracks
Track order
Search results

CRISPR SpCas9

Select from available configurations: Current unsaved

Genome targeting

Looking for more data? Search the [Trackhub Registry](#) for external sources of annotation

Key

- Track style
- External External data
- Forward strand
- Custom User-added track
- Reverse strand
- Favourite track
- Track information

Please note that the content of external tracks is not the responsibility of the Ensembl project.

URL-based tracks may either slow down your ensembl browsing experience OR may be unav

Sequence and assembly (3/29)
Sequence (2/4)
Simple features (1/3)
Clones & misc. regions (0/22)

Genes and transcripts (2/108)
Genes (2/4)
Long reads (0/17)
RNASeq models (0/87)

mRNA and protein alignments (0/2)
mRNA alignments (0/1)
Protein alignments (0/1)

Variation (2/10)
Sequence variants (1/3)
Failed variants (0/1)
Phenotype annotations (0/2)
Structural variants (1/4)

Regulation (1/166)
Regulatory features (1/1)
Activity by Cell/Tissue (0/161)
DNA methylation (0/2)
Other regulatory regions (0/2)

Comparative genomics (1/29)
Multiple alignments (0/4)
Conservation regions (1/4)
BLASTz/LASTz alignments (0/21)

Genome targeting (0/1)

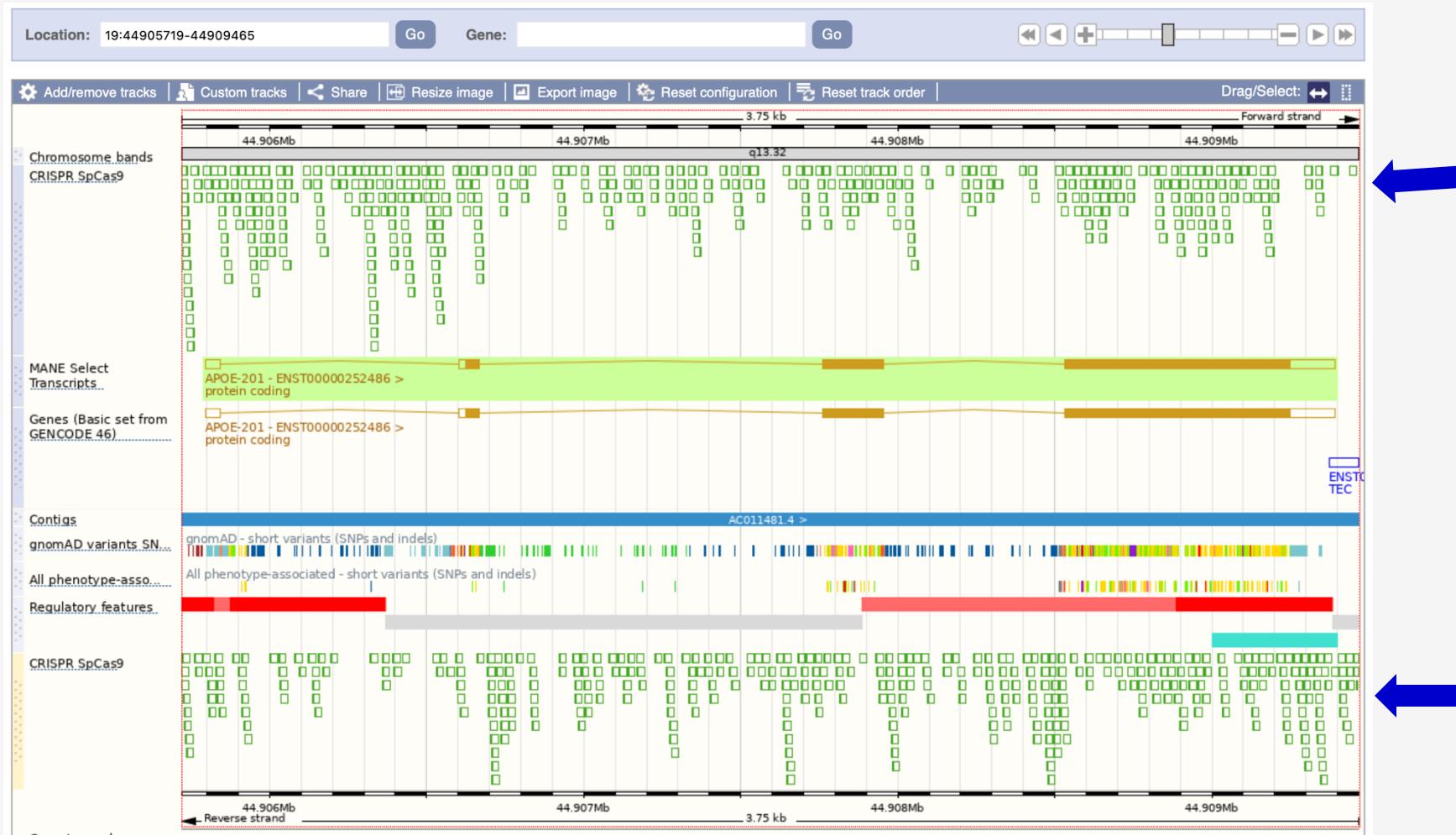
Oligo probes (0/26)
Repeat regions (0/12)

Information and decorations (13/16)

Display options

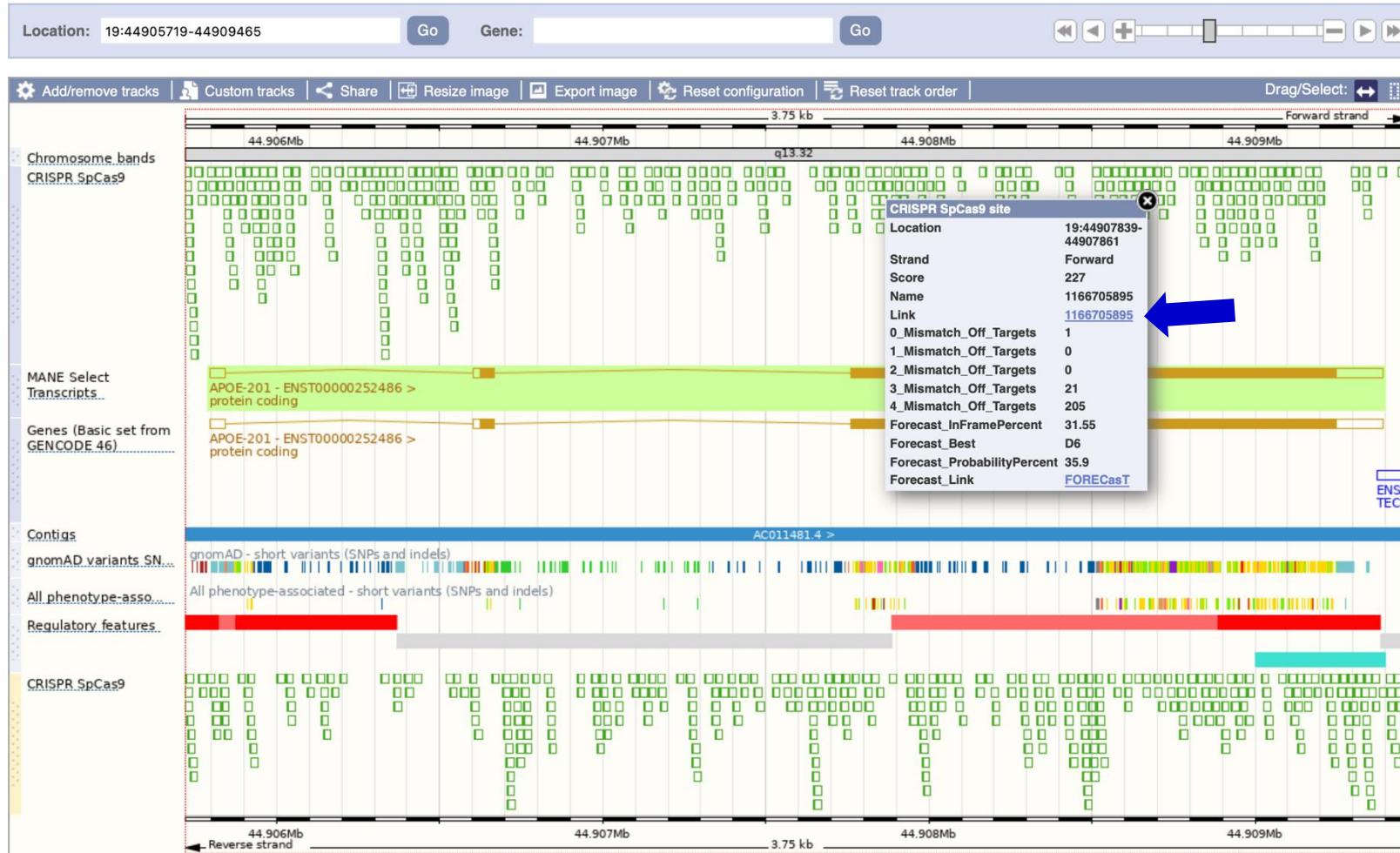
CRISPR Design Process and Tools

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

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

[View in Genome Browser](#)



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](#)

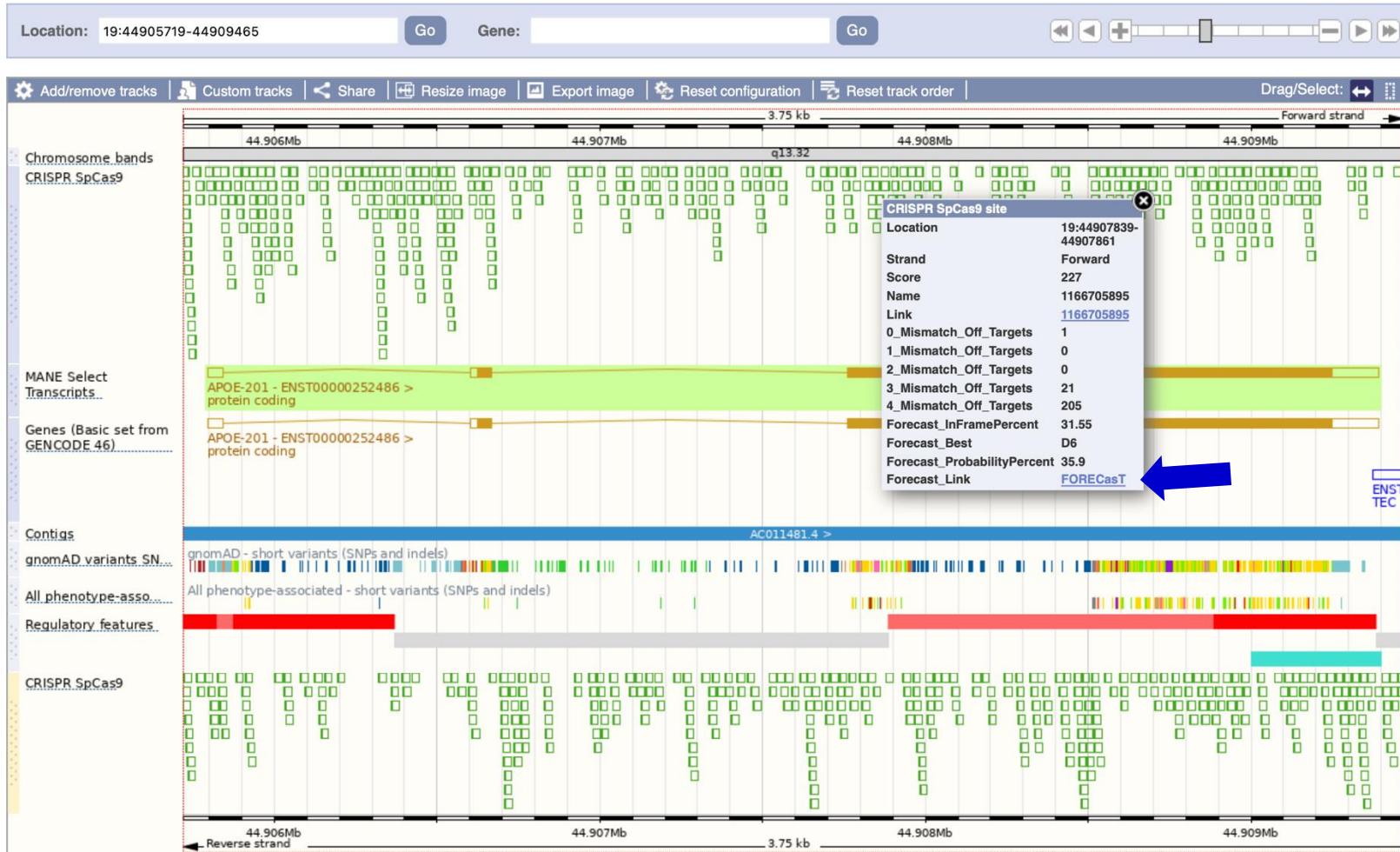
Off-Target Sites

Note: the row highlighted in blue is the original CRISPR

WGE ID	Location	Sequence	Mismatches <small>i</small>	Strand	Type
1166705895	Original CRISPR	CCAGCGCTGGGAACCTGGCAC TGG			Exonic
1166705895	19:44907839-44907861	CCAGCGCTGGGAACCTGGCAC TGG	0	+	Exonic
1183733406	22:39630637-39630659	GCAGAGCCGGAACCTGGCAC AGG	3	+	Intronic
1021787321	7:24164871-24164893	GCAGCTCTAGAACCTGGCAC TGG	3	+	Intergenic
1001791760	5:174463776-174463798	CCAC CCC CTGGCAACTGGCAC AGG	3	+	Intergenic
1119608739	14:76043865-76043887	CCAG GGCTTG CAACTGGCAC TGG	3	-	Intronic

CRISPR Design Process and Tools

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



CRISPR Design Process and Tools

Generate a list of sgRNAs rather than selecting them individually.

HTGT WGE

Home

CRISPR Finder

Help

About

Contact

CRISPR search

Species

- Human (GRCh38)
- Mouse (GRCm38)

Marker Symbol

APOE

Exons for ENST00000252486

1. ENSE00001048576 (length 46)
2. ENSE00003577086 (length 66)
3. ENSE00000893952 (length 193)
4. ENSE00000893954 (length 861)

Note: the CRISPR table only shows CRISPRs that overlap the exon by at least 1 base.

To see flanking crisprs please use the genome browser



Show CRISPRs in:

Table Genome Browser Download

Show CRISPR pairs in:

Table Genome Browser Download

CRISPR Design Process and Tools

Exon ID	ID	Location	Seq	Off targets
ENSE0000893952	1166705873	19:44907738-44907760	ACCTTGAACCTGTTCCACACAGG	0: 1 1: 0 2: 0 3: 13 4: 93
	1166705874	19:44907739-44907761	CCTTGAACCTGTTCCACACAGGA	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

CRISPR Design Process and Tools

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](#)

Or, upload FASTA file [Choose File](#) no file selected

Range [?](#) [Clear](#)

Forward primer From

Reverse primer To

Primer Parameters

Use my own forward primer (5'→3' on plus strand)

 [?](#) [Clear](#)

Use my own reverse primer (5'→3' on minus strand)

 [?](#) [Clear](#)

PCR product size

Min 70 Max 1000

of primers to return

 10

Opt Max Max T_m difference 3 [?](#)

Primer melting temperatures (T_m)

Min 57.0 Opt 60.0 Max 63.0 Max T_m difference 3 [?](#)

Exon/intron selection

Exon junction span

No preference [?](#)

Exon junction match

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

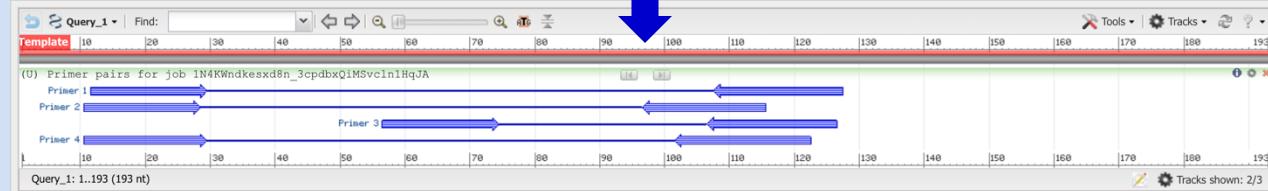
Intron inclusion

Minimal and maximal number of bases that must anneal to exons at the 5' or 3' side of the junction [?](#)

Intron length range

Primer pair must be separated by at least one intron on the corresponding genomic DNA [?](#)

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	AAGGTGAGCAAGCGGTG	Plus	18	12	29	60.28	61.11	2.00	0.00
Reverse primer	CAGCGCAGGTAATCCAAA	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:

Pair		
Exon ID	Spacer	Status
ENSE00003666217	20	Complete
		closest: None total_pairs: 1 max_distance: 1000

Find CRISPRs by 20bp gRNA:

Sequence: AATAGTAGACATAAAAGTCT

Species: Human (GRCh37) Human (GRCh38) Mouse (GRCm38)

Find CRISPRs

EnsEMBL	In gene	In e...
---------	---------	---------

Find CRISPRs in genomic sequence:

Sequence: IAAGGAATGTTCCC**A**ATTAGTAGACATAAAAGT**C**TCG

Search Again

Crisp ID	EnsEMBL	In g...
1106710403	13:32325087-32325109	No
1106710404	13:32325088-32325110	No
1106710405	13:32325110-32325132	No

Find off-targets by sequence:

Orientation: Mouse (GRCm38) PAM Right (NGG) PAM Left (CCN)

Find Off-Targets

Sequence	Ori...
GTGTCAGTGAAACTTACTCT	par
GTGCCCCAGAAACTTACTCT	par

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

Copyright (c) 2019 Genome Research Limited Your use of this site indicates your agreement to the GNU AGPLv3 licence



sgRNA Design Tools

Other tools



CRISPOR ([citation](#)) is a program that helps design, evaluate and clone guide sequences for the CRISPR/Cas9 system. [CRISPOR Manual](#)

July 18, 2024: The old server has been retired. The new Python3 server is still lacking the Najm 2018 saCas9 score. See [Full list of changes](#)

Step 1

Planning a lentiviral gene knockout screen? Use [CRISPOR Batch](#)

Sequence name (optional):

Enter a single genomic sequence, < 2300 bp, typically an exon

[Clear Box](#) - [Reset to default](#)

Paste here the genomic - not a cDNA - sequence of the exon you want to target. The sequence has to include the PAM site for your enzyme of interest, e.g. NGG. Maximum size 2300 bp. If you only have a cDNA, please BLAST or BLAT the cDNA first to find the right exon sequence for CRISPOR.

Text case is preserved, e.g. you can mark ATGs with lowercase.

Instead of a sequence, you can paste a chromosome range, e.g. chr1:11,130,540-11,130,751

Step 2

Select a genome

We have 1171 genomes, but not yours? Search [NCBI assembly](#) and send a GCF_/_GCA_ ID to [CRISPOR support](#).

Step 3

Select a Protospacer Adjacent Motif (PAM)

See [notes on enzymes](#) in the manual.

SUBMIT

sgRNA Design Tools

Other tools

Home Instructions Scoring About Updates Submissions Contact FAQ



Target: APOE In: Homo sapiens (hg38/GRCh38) Using: CRISPR/Cas9 For: knock-out

RefSeq/ENSEMBL/gene ID or genomic coordinates. Add new species.

Paste Target Options Reset Options

Find Target Sites!

APOE

5' → 3'
exon + ATG
intron
target

NM_001302691
NM_001302689
NM_001302690
NM_001302688
NM_000041

44,906,000 44,906,500 44,907,000 44,907,500 44,908,000 44,908,500 44,909,000

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	GGCCTACAAATCGGAACCTGGAGG	chr19:44908566	+	55	2	0	0	0	0	63.59
2	ACCTGCGCAAGCTCGTAAAGCGG	chr19:44908769	+	60	1	0	0	0	0	54.43
3	GGCGTTCACTGATTGTCGCTGG	chr19:44909233	-	55	0	0	0	0	0	50.48
4	CGCGGTTCACTGATTGTCGCTGG	chr19:44909234	-	60	0	0	0	0	0	28.62
5	TCTGCAAGTCATCGGCATCGCGG	chr19:44908797	-	60	0	0	0	0	1	70.56

Pair	Left primer coordinates	Left primer	Left primer Tm	Left primer 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	GACACTCACCTCAGTTCCCTGG	59.7	0	0	266

IMPERIAL

Thank you

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