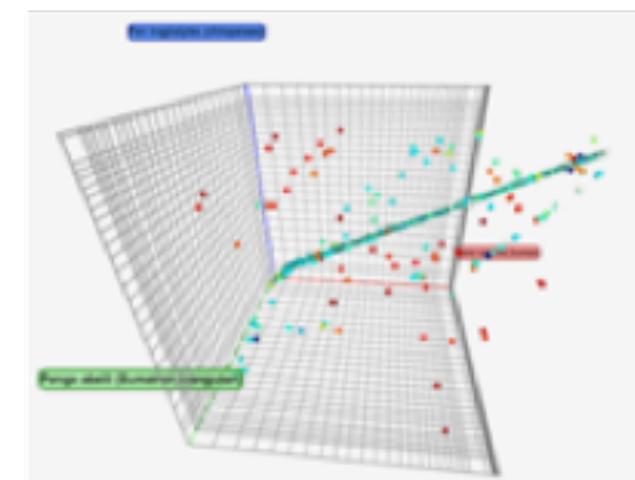
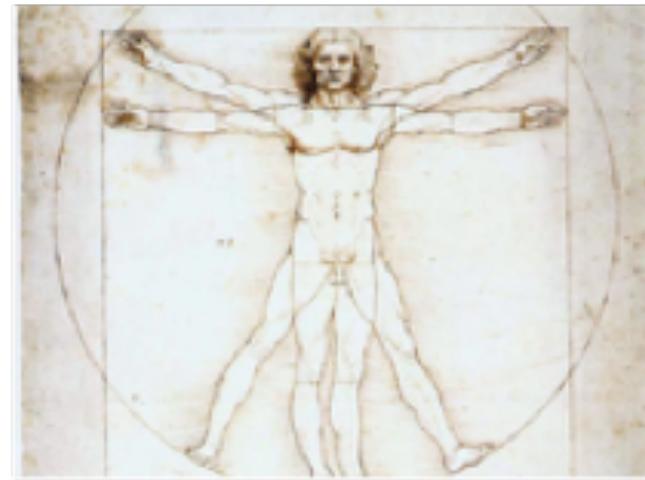
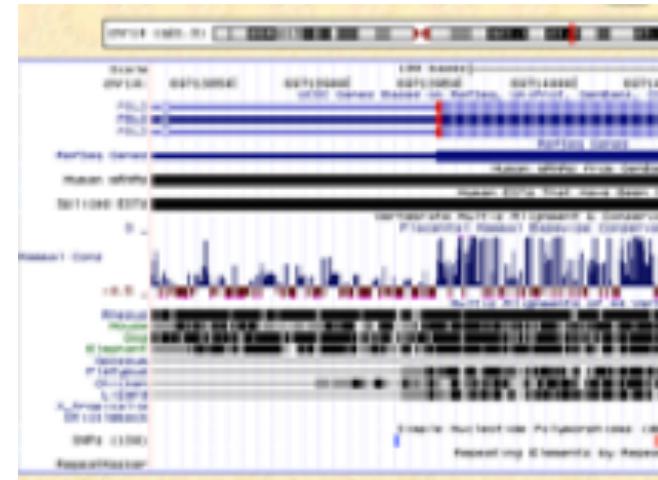


# Computational Genomics

## Data Retrieval II



# Data Retrieval

## Retrieving Data From ENSEMBL

### FTP

Show 10 entries															Show/hide columns		Filter						
★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM/BigWig						
Y	<a href="#">Human <i>Homo sapiens</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	<a href="#">GVF ↗</a>	<a href="#">VCF ↗</a>	<a href="#">VEP ↗</a>	<a href="#">Regulation ↗ (GFF)</a>	<a href="#">Regulation data files ↗</a>	<a href="#">BAM/BigWig ↗</a>										
Y	<a href="#">Mouse <i>Mus musculus</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	<a href="#">GVF ↗</a>	<a href="#">VCF ↗</a>	<a href="#">VEP ↗</a>	<a href="#">Regulation ↗ (GFF)</a>	<a href="#">Regulation data files ↗</a>	<a href="#">BAM/BigWig ↗</a>										
Y	<a href="#">Zebrafish <i>Danio rerio</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	<a href="#">GVF ↗</a>	<a href="#">VCF ↗</a>	<a href="#">VEP ↗</a>	-	-	<a href="#">BAM/BigWig ↗</a>										
	<a href="#">Abingdon island giant tortoise <i>Chelonoidis abingdonii</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	<a href="#">BAM/BigWig ↗</a>										
	<a href="#">African ostrich <i>Struthio camelus australis</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	<a href="#">BAM/BigWig ↗</a>										
	<a href="#">Agassiz's desert tortoise <i>Gopherus agassizii</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	<a href="#">BAM/BigWig ↗</a>										
	<a href="#">Algerian mouse <i>Mus spretus</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	-										
	<a href="#">Alpaca <i>Vicugna pacos</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	-										
	<a href="#">Alpine marmot <i>Marmota marmota marmota</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	<a href="#">BAM/BigWig ↗</a>										
	<a href="#">Amazon Molly <i>Poecilia formosa</i></a>	<a href="#">FASTA ↗</a>	<a href="#">EMBL ↗</a>	<a href="#">GenBank ↗</a>	<a href="#">GTF ↗</a> <a href="#">GFF3 ↗</a>	<a href="#">TSV ↗</a> <a href="#">RDF ↗</a> <a href="#">JSON ↗</a>	<a href="#">MySQL ↗</a>	-	-	<a href="#">VEP ↗</a>	-	-	<a href="#">BAM/BigWig ↗</a>										
Showing 1 to 10 of 311 entries															<a href"="">&lt;&lt;</a>	<a href"="">&lt;</a>	<a href"="">1</a>	<a href"="">2</a>	<a href"="">3</a>	<a href"="">4</a>	<a href"="">5</a>	<a href"="">&gt;</a>	<a href="">&gt;&gt;</a>

# Data Retrieval

## Retrieving Data From ENSEMBL

BioMart



# Data Retrieval

## Retrieving Data From ENSEMBL

BioMart for Data Mining

Ensembl's BioMart

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## What is BioMart?

A tool in your browser to:

- Export data with no programming required
- Build queries with a few mouse clicks
- Generate custom data tables and sequence files

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## Training materials



- Ensembl training materials are protected by a CC BY license:  
[creativecommons.org/licenses/by/4.0/](http://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](http://ensembl.org/info/about/publications.html)

# Data Retrieval

## Retrieving Data From ENSEMBL

BioMart for Data Mining



# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## Why use BioMart?

For things that would be time consuming/difficult with the Ensembl browser:

- Query multiple things at once:
  - ID conversions
  - Gene locations
  - Download sequences
- Export large amounts of data

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining



[BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#)

[Login/Register](#)

[Search all species...](#)



[HMMER](#) | [BLAST](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Documentation](#) | [Website help](#)

[Login/Register](#)

[Search Ensembl Protists...](#)



# Where can I find BioMart?

[ensembl.org/biomart/martview](http://ensembl.org/biomart/martview)  
[protists.ensembl.org/biomart/martview](http://protists.ensembl.org/biomart/martview)

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

For which genomes is BioMart available?

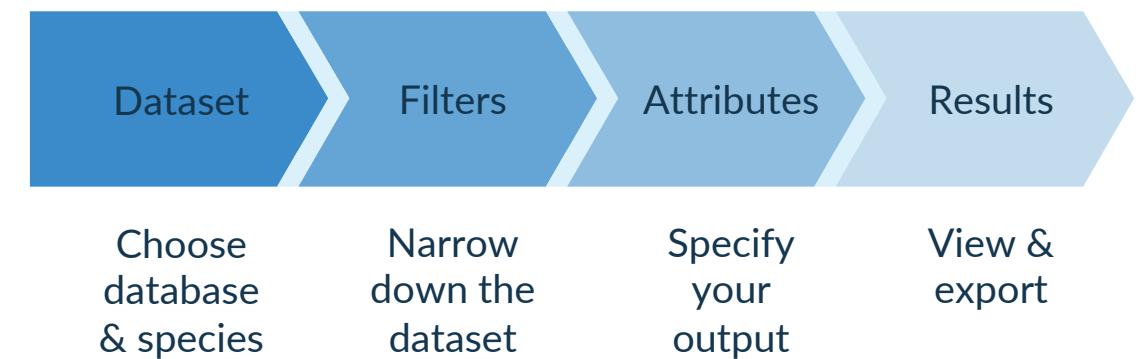
- *e!Ensembl*  
(some exceptions)
- *e!EnsemblFungi*  
(some exceptions)
- *e!EnsemblMetazoa*  
(some exceptions)
- *e!EnsemblPlants*  
(some exceptions)
- *e!EnsemblProtists*  
(some exceptions)

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

How do I use  
BioMart?  
The four steps



# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## How do I use BioMart?

### 1. Dataset

Dataset

- Define the database you want to search with your filters:
  - Genes
  - Variation
  - Regulation
  - Mouse strains (genes)
- Define the species

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## How do I use BioMart? 2. Filters



Define a (large) set of genes/variants by combination of filters, e.g.:

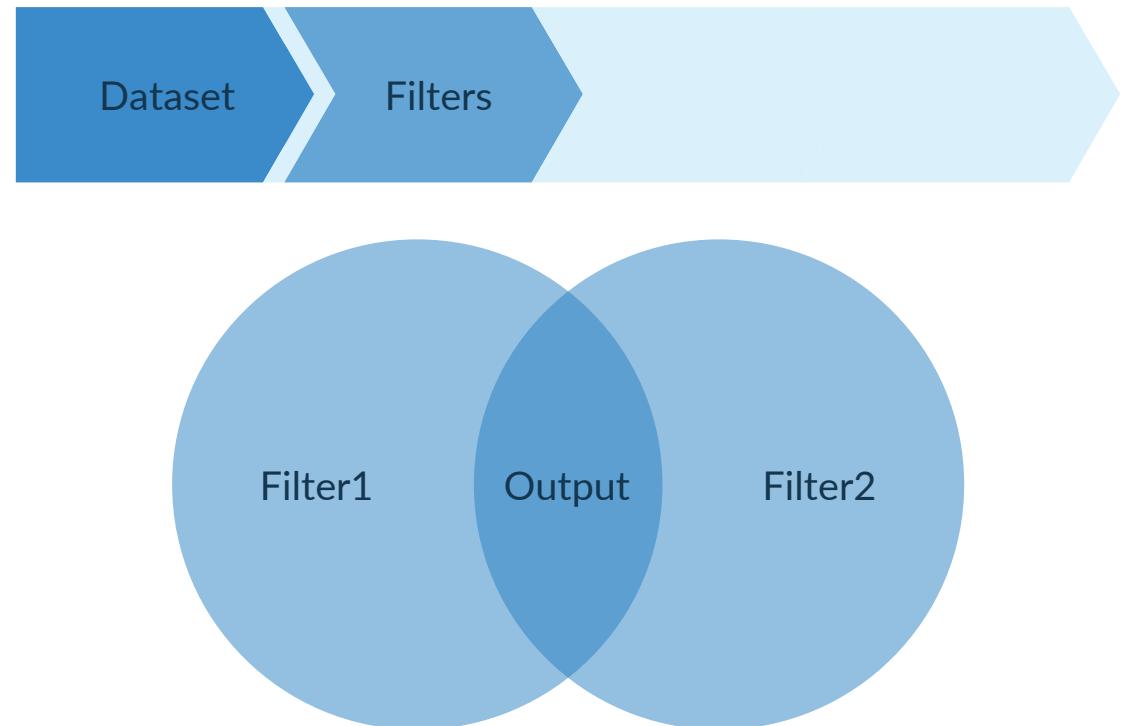
- Region
- List of IDs
- Function (GO term)
- Phenotype

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

How do I use  
BioMart?  
2. Filters



# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining



## How do I use BioMart? 3. Attributes

Define the data you want to export (your output), e.g.:

- IDs
- Features
- Variants
- Orthologues/paralogues
- Sequences

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## How do I use BioMart? 4. Results



View and export data table/sequence in a number of formats:

- html
- csv
- tsv
- xls
- fasta

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

**biomaRt:**  
Bioconductor  
package for  
BioMart

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data using the R statistical programming language

- Easy installation:

```
source("http://bioconductor.org/biocLite.R")
biocLite("biomaRt")
```

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart for Data Mining

## More information

Kinsella *et al.*

**Ensembl BioMarts: a hub for data retrieval across taxonomic space**

*Database: the Journal of Biological Databases and Curation* (2011)

[europepmc.org/abstract/MED/21785142](https://europepmc.org/abstract/MED/21785142)

Smedley *et al.*

**BioMart – biological queries made easy**

*BMC Genomics* (2009) 10:22

[europepmc.org/abstract/MED/19144180](https://europepmc.org/abstract/MED/19144180)

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Go to BioMart and Retrieve: **Human: ENSP00000467141**

The screenshot shows the BioMart interface. At the top, there are navigation buttons: New, Count, Results, URL, XML, Perl, and Help. On the left, there are sections for Dataset, Filters, and Attributes. The Dataset section shows 'Ensembl Genes 105' selected. The Filters section contains the query 'Human genes (GRCh38.p13)'. The Attributes section lists 'Gene stable ID' and 'Transcript stable ID'. At the bottom, another Dataset section shows '[None Selected]'. The main area is currently empty, indicating no results have been retrieved.

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Go to BioMart and Retrieve: **Human: ENSP00000467141**

The screenshot shows the BioMart search interface. On the left, there's a sidebar with sections for Dataset (Human genes (GRCh38.p13)), Filters (Protein stable ID(s) [e.g. ENSP00000000233]: [ID-list specified]), and Attributes (Gene stable ID, Transcript stable ID). The main area has tabs at the top: New, Count, Results, URL, XML, Perl, and Help. A message says "Please restrict your query using criteria below" and "If filter values are truncated in any lists, hover over the list item to see the full text". The search form includes fields for REGION, GENE, and other filters like microarray probes. The GENE section has a checked checkbox for "Input external references ID list [Max 500 advised]" and a file input field containing "ENSP00000467141". Other sections include "Limit to genes (microarray probes/probesets)" and "Input microarray probes/probesets ID list [Max 500 advised]".

New Count Results URL XML Perl Help

**Dataset**  
Human genes (GRCh38.p13)

**Filters**  
Protein stable ID(s) [e.g. ENSP00000000233]: [ID-list specified]

**Attributes**  
Gene stable ID  
Transcript stable ID

**Dataset**  
[None Selected]

Please restrict your query using criteria below  
(If filter values are truncated in any lists, hover over the list item to see the full text)

**REGION:**

**GENE:**

Limit to genes (external references)... With BioGRID Interaction data, The General Repository for...

Only  
 Excluded

Input external references ID list [Max 500 advised]

Protein stable ID(s) [e.g. ENSP00000000233]  
ENSP00000467141

Choose File No file chosen

Limit to genes (microarray probes/probesets)... With AFFY HC G110 probe ID(s)

Only  
 Excluded

Input microarray probes/probesets ID list [Max 500 advised]

AFFY HC G110 probe ID(s) [e.g. 737\_at]

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Select: Peptide

New Count Results URL XML Perl Help

**Dataset 1 / 68005 Genes**  
Human genes (GRCh38.p13)

**Filters**  
Protein stable ID(s) [e.g.  
ENSP00000000233]: [ID-list  
specified]

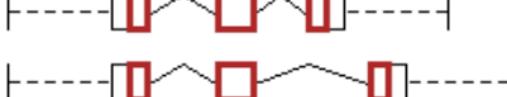
**Attributes**  
Peptide  
Gene stable ID  
Gene stable ID version  
Transcript stable ID  
Transcript stable ID version

**Dataset**  
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following [FAQ](#)

Features       Variant (Germline)  
 Structures       Sequences  
 Homologues (Max select 6 orthologues)

SEQUENCES:  
Sequences (max 1)  
  
 Unspliced (Transcript)       5' UTR  
 Unspliced (Gene)       3' UTR  
 Flank (Transcript)       Exon sequences  
 Flank (Gene)       cDNA sequences  
 Flank-coding region (Transcript)       Coding sequence  
 Flank-coding region (Gene)       Peptide

**Upstream flank**  
 Upstream flank

**Downstream flank**

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Select: **Gene stable ID, Transcript stable ID and Protein stable ID**

New Count Results      URL XML Perl Help

**Dataset 1 / 68005 Genes**  
Human genes (GRCh38.p13)

**Filters**  
Protein stable ID(s) [e.g. ENSP00000000233]: [ID-list specified]

**Attributes**  
Peptide  
Gene stable ID  
Transcript stable ID  
Protein stable ID version

**Dataset**  
[None Selected]

**HEADER INFORMATION:**

**Gene Information**

Gene stable ID  
 Gene stable ID version  
 Gene description  
 Gene name  
 Source of gene name  
 Chromosome/scaffold name  
 Gene start (bp)

Gene end (bp)  
 Gene type  
 Version (gene)  
 UniParc ID  
 UniProtKB/Swiss-Prot ID  
 UniProtKB/TrEMBL ID

**Transcript Information**

CDS start (within cDNA)  
 CDS end (within cDNA)  
 5' UTR start  
 5' UTR end  
 3' UTR start  
 3' UTR end  
 Transcript stable ID  
 Transcript stable ID version  
 Protein stable ID

Protein stable ID version  
 Transcript type  
 Version (transcript)  
 Version (protein)  
 Strand  
 Transcript start (bp)  
 Transcript end (bp)  
 Transcription start site (TSS)  
 Transcript length (including UTRs and CDS)

**Exon Information**

CDS Length  
 Start phase

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Rename Download as: **Hs\_ENSP00000467141.pep.fa**

**New** **Count** **Results**      **URL** **XML** **Perl** **Help**

**Dataset** 1 / 68005 Genes  
Human genes (GRCh38.p13)  
**Filters**  
Protein stable ID(s) [e.g.  
ENSP00000000233]: [ID-list  
specified]  
**Attributes**  
Peptide  
Gene stable ID  
Transcript stable ID  
Protein stable ID version  
  
**Dataset**  
[None Selected]

Export all results to **File**  **FASTA**  Unique results only  
Email notification to   
View **10** rows as **FASTA**  Unique results only

```
>ENSG00000155657|ENST00000589042|ENSP00000467141.1
MTTQAPFTQPLQLQSVVVLEGSTATFEAHISGFPVPEVSWFRDGQVISTSTLPGVQISFSD
GRAKLTIPAVTKANSGRYSLKATNGSGQATSTAELLVKAETAPPNFVQRLQSMTVRQGSQ
VRLQVRVTGIPPTPVVKFYRDGAEIQSSLDFQISQEGDLYSLLIAEAYPEDSGTYSVNATN
SVGRATSTAELLVQGEEEVPAKKTKTIVSTAQISESRQTRIEKKIEAHFDARSIAATVEMV
IDGAAGQQLPHKTTPRIPPKPKRSRSPPPSIAAKAQLARQQSPSPIRHSPSPVRHVRAPT
PSPVRSVSPAARISTSPIRSVRSPLLMRKTQASTVATGPEVPPPWKQEGYVASSSEAEMR
ETTLTTSTQIRTEERWEGRYGVQEVTIISGAAGAAASVSASASYAAEAVATGAKEVKQDA
DKSAAVATVVAAVDMARVREPVISAVEQTAQRTTTAVHIQPAQEQRKEAKTAVTKVV
VAADKAKEQELKSRTKEVITTKQEQMHTHEQIRKETEKTTFVPKVVISAAKAKEQETRIS
EEITKKQKQVTQEAIROQETEITAASMVVVATAKSTKLETVPGAQEETTTQDQMHLSYEK
IMKETRKTVVPKVIVATPKVKEQDLVSRGREGITTKREQVQITQEKMRAEKTALSTIA
VATAKAKEQETILRTRETMATRQEIQIQTGHKVDVGKKAEEAVATVVAAVDQARVREPREP
GHLEESYAQOTTLEYGYKERISAQVAEPPQRPASEPHVVPKAVKPRVIQAPSETHIKTT
DQKGMISSQIKTTDLTTERLVHVDKRPTASPHFTVSKISVPKTEHGYEASIAGSAIA
TLQKELSATSSAQKITKSVKAPTVKPSETRVRAEPTPLQFPFADTPDTYKSEAGVEVKK
EVGVSITGTTVREERFEVLHGREAKVTETARVPAPVEIPVTPPTLVSGLKNVTVIEGESV
TLECHISGYPSPVTWYREDYQIESSIDFQITFQSGIARLMIREAFAEDSGRFTCSAVNE
AGTVSTSCYLAQVSEEFEKETTAVTEKFTTEEKRFVESRDVVMTDTSLEEQAGPGEPA
APYFITKPVVQKLVEGGSVVFGCQVGGNPKPHVYWKSGVPLTTGYRYKVSYNKQTGECK
LVIISMTFADDAGEYTIVVRNKHGETSASASLLEEADYLLMKSQQEMLYQTOVTAFFVQEP
```

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- From the Header we extract the following table of relationships:

>ENSG00000155657|ENST00000589042|ENSP00000467141.1

- Now, we will proceed to extract: ENSG00000155657

The screenshot shows the BioMart interface for retrieving data from the ENSEMBL dataset. The left sidebar lists the dataset as "Human genes (GRCh38.p13)" and provides filter and attribute options. The main panel displays a query form with the following details:

**Dataset:** Human genes (GRCh38.p13)

**Filters:**

- Gene stable ID(s) [e.g. ENSG00000000003]: [ID-list specified]

**Attributes:**

- Gene stable ID
- Transcript stable ID
- Protein stable ID version
- Unspliced (Gene)

**Query Criteria:**

Please restrict your query using criteria below  
(If filter values are truncated in any lists, hover over the list item to see the full text)

**REGION:**

**GENE:**

- Limit to genes (external references)...  
With BioGRID Interaction data, The General Repository for  
 Only  
 Excluded
- Input external references ID list [Max 500 advised]  
Gene stable ID(s) [e.g. ENSG00000000003]  
ENSG00000155657  
Choose File No file chosen
- Limit to genes (microarray probes/probesets)...  
With AFFY HC G110 probe ID(s)  
 Only  
 Excluded
- Input microarray probes/probesets ID list [Max 500 advised]  
AFFY HC G110 probe ID(s) [e.g. 737\_at]

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- From the Header we extract the following table of relationships:

>ENSG00000155657|ENST00000589042|ENSP00000467141.1

- Now, we will proceed to extract the: **Unspliced (Gene)**

**New** **Count** **Results**      **URL** **XML** **Perl** **Help**

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following [FAQ](#)

**Features**       **Variant (Germline)**  
 **Structures**       **Sequences**  
 **Homologues (Max select 6 orthologues)**

**SEQUENCES:**

**Sequences** (max 1)



Unspliced (Transcript)       5' UTR  
 Unspliced (Gene)       3' UTR  
 Flank (Transcript)       Exon sequences  
 Flank (Gene)       cDNA sequences  
 Flank-coding region (Transcript)       Coding sequence  
 Flank-coding region (Gene)       Peptide

**Upstream flank**  
 Upstream flank

**Downstream flank**

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- From the Header we extract the following table of relationships:

>ENSG00000155657|ENST00000589042|ENSP00000467141.1

- Select: **Gene stable ID, Transcript stable ID and Protein stable ID**

[New](#) [Count](#) [Results](#) [URL](#) [XML](#) [Perl](#) [Help](#)

**Dataset**  
Human genes (GRCh38.p13)

**Filters**  
Protein stable ID(s) [e.g.  
ENSP00000000233]: [ID-list  
specified]

**Attributes**  
Gene stable ID  
Transcript stable ID  
Protein stable ID version  
Unspliced (Gene)

**Dataset**  
[None Selected]

**HEADER INFORMATION:**

**Gene Information**

Gene stable ID  
 Gene stable ID version  
 Gene description  
 Gene name  
 Source of gene name  
 Chromosome/scaffold name  
 Gene start (bp)

Gene end (bp)  
 Gene type  
 Version (gene)  
 UniParc ID  
 UniProtKB/Swiss-Prot ID  
 UniProtKB/TrEMBL ID

**Transcript Information**

CDS start (within cDNA)  
 CDS end (within cDNA)  
 5' UTR start  
 5' UTR end  
 3' UTR start  
 3' UTR end  
 Transcript stable ID  
 Transcript stable ID version  
 Protein stable ID

Protein stable ID version  
 Transcript type  
 Version (transcript)  
 Version (protein)  
 Strand  
 Transcript start (bp)  
 Transcript end (bp)  
 Transcription start site (TSS)  
 Transcript length (including UTRs and CDS)

**Exon Information**

CDS Length  
 CDS start

Start phase  
 End phase

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Rename Download as: **Hs\_ENSG00000155657.dna.fa**

**New** **Count** **Results**      **URL** **XML** **Perl** **Help**

**Dataset**  
Human genes (GRCh38.p13)

**Filters**  
Gene stable ID(s) [e.g.  
ENSG00000000003]: [ID-list  
specified]

**Attributes**  
Gene stable ID  
Transcript stable ID  
Protein stable ID version  
Unspliced (Gene)

**Dataset**  
[None Selected]

Export all results to **File**  **Go**  FASTA  Unique results only

Email notification to

View **10** rows as **FASTA**  Unique results only

```
>ENSG00000155657|ENST00000589042;ENST00000591111;ENST00000342175;ENST0000035
GTTCCAGTTCTGCTGAGACACAACCTCCCTGGGAAGCCTCCTGACCCATCAAGTCCAGA
GTAGATGGTCCTGCCATGTGCACTGTGCAGCACCTGCAGTGCCTGCCAGCCACGCCCTGAATA
TACACAAAGGAATCGCAGGTTGAGCATCAGTCTCCTCTGGCTAGGCTGTAAGAACCTTA
TCAGCAGGATTCATGACAGCCCTGTTGTCTTGTATTGCCAGTACCTAACAGACACAGTCAC
ACTCAGTAAACTCTTTGAATTGAATTGTTGCCAAAGGTAACAAAGATAAACCTAGCTT
GTTTTCTCCCCTTTAGAGGGAGGAGAGGGAAATGTAAAAAGACAAAGATGATTCTC
CCTTCTTATAGACAGGAGTCAATCCGAATCAGGCCCTAACATGAAGCACTCATGAAAT
CTCTCACCTCATCTGCAAAACTTGTGTGAGAACCCCTCCAGACCCTCTGGCTAGACCCTG
ATTGAAAATTGAATTAAAAATAGATGCTGGCTCCACGGACTAAAATAATAGTAAGGAG
GAAAGAGGAAAGGTCCAGGATTGTTGTTCTGGATCAATATTGTTCAACATTAAAGT
GTTCCAACATTAAAACATATGTACTTAGGCATAAATCATCATCCGGTATGAGGAACGT
GATGCCTCCACTCTATGACTAAATCCAAACTGTCAAAACCAGAATTATGTACTTCTT
TAGAACACTACGTAAATGATCAATATTAAATTGTAGCTGGGCTCAGTAGTCTGGCTGT
GAACAGATCTATTGAAACCAGTATTCAACAAACTTGTAAATATAAGAAAGCCAAA
TCCAAGAACCTTAAGAGTATTGAATTCCAAATGAAGAAAGGAAAGAATATCTTATTGT
TCTGCTTATGTTGTCTTATTCACTGACCAATTCTTACTAAAATCACCCTGAAAGCT
TTGAAACTCCTGTTGTATGCACAGCATTAAAGTGCAGCCATGAAATAAGTTAGTTGAA
ACAAAATAAAATATCTCCTCCAGTATGCTCTGATATTGTGTAATGTAATGCCAAGGC
ATGACAGACTGTTCCATTATAAAAGCGGAATCTTGAGTTGCAATTGGAAAGCTAAA
GATTGCGTTGACTTAAATGAAACTGCTCTTTTAAAGCAAAACAAATTCATTCTCTG
```

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- From the Header we extract the following table of relationships:

>ENSG00000155657|ENST00000589042|ENSP00000467141.1

- Now, we will proceed to extract: ENST00000589042

The screenshot shows the BioMart interface with the following details:

- Header:** URL, XML, Perl, Help.
- Dataset:** Human genes (GRCh38.p13).
- Filters:**
  - Transcript stable ID(s) [e.g. ENST000000000233]: [ID-list specified]
  - Attributes:
    - Gene stable ID
    - Transcript stable ID
    - Protein stable ID version
    - Unspliced (Gene)
- Query Parameters:**
  - REGION:** (checkboxes for 'Limit to genes' and 'With BioGRID Interaction data')
  - GENE:**
    - Limit to genes (external references)...
      - Only
      - Excluded
    - Input external references ID list [Max 500 advised]  
ENST00000589042  
Choose File No file chosen
    - Limit to genes (microarray probes/probesets)...
      - Only
      - Excluded
    - Input microarray probes/probesets ID list [Max 500 advised]  
AFFY HC G110 probe ID(s) [e.g. 737\_at]

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- From the Header we extract the following table of relationships:

>ENSG00000155657|ENST00000589042|ENSP00000467141.1

- Now, we will proceed to extract the: **cDNA sequences**

**New** **Count** **Results**      **URL** **XML** **Perl** **Help**

**Dataset**  
Human genes (GRCh38.p13)

**Filters**  
Transcript stable ID(s) [e.g. ENST00000000233]: [ID-list specified]

**Attributes**  
Gene stable ID  
Transcript stable ID  
Protein stable ID version  
cDNA sequences

**Dataset**  
[None Selected]

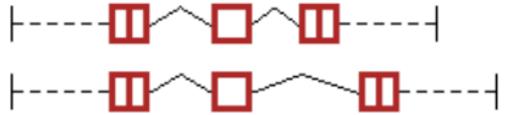
Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following [FAQ](#)

Features       Variant (Germline)  
 Structures       Sequences  
 Homologues (Max select 6 orthologues)

**SEQUENCES:**

**Sequences** (max 1)

|------

Unspliced (Transcript)  
 Unspliced (Gene)  
 Flank (Transcript)  
 Flank (Gene)  
 Flank-coding region (Transcript)  
 Flank-coding region (Gene)

5' UTR  
 3' UTR  
 Exon sequences  
 cDNA sequences  
 Coding sequence  
 Peptide

**Upstream flank**  
 Upstream flank

**Downstream flank**

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- From the Header we extract the following table of relationships:

>ENSG00000155657|ENST00000589042|ENSP00000467141.1

- Now, we will proceed to extract the: **cDNA sequences**

The screenshot shows the BioMart interface for retrieving data from the ENSEMBL database. The left sidebar lists datasets: Human genes (GRCh38.p13), Transcript stable ID(s) [e.g. ENST00000000233], and Attributes (Gene stable ID, Transcript stable ID, Protein stable ID version, cDNA sequences). The main panel shows the 'Dataset' section selected, with [None Selected] chosen. The 'Attributes' section is expanded, showing checkboxes for various gene, transcript, and exon information. The 'Gene Information' section includes checked boxes for Gene stable ID and Gene start (bp), and unchecked boxes for Gene end (bp), Gene type, Version (gene), UniParc ID, UniProtKB/Swiss-Prot ID, and UniProtKB/TrEMBL ID. The 'Transcript Information' section includes checked boxes for Transcript stable ID and Protein stable ID version, and unchecked boxes for Transcript type, Version (transcript), Version (protein), Strand, Transcript start (bp), Transcript end (bp), Transcription start site (TSS), and Transcript length (including UTRs and CDS). The 'Exon Information' section includes unchecked boxes for CDS Length and CDS start.

New Count Results URL XML Perl Help

**Dataset**  
Human genes (GRCh38.p13)

**Filters**  
Transcript stable ID(s) [e.g. ENST00000000233]: [ID-list specified]

**Attributes**

Gene stable ID  
Transcript stable ID  
Protein stable ID version  
cDNA sequences

**Dataset**  
[None Selected]

**HEADER INFORMATION:**

**Gene Information**

Gene stable ID  
 Gene stable ID version  
 Gene description  
 Gene name  
 Source of gene name  
 Chromosome/scaffold name  
 Gene start (bp)  
 Gene end (bp)  
 Gene type  
 Version (gene)  
 UniParc ID  
 UniProtKB/Swiss-Prot ID  
 UniProtKB/TrEMBL ID

**Transcript Information**

CDS start (within cDNA)  
 CDS end (within cDNA)  
 5' UTR start  
 5' UTR end  
 3' UTR start  
 3' UTR end  
 Transcript stable ID  
 Transcript stable ID version  
 Protein stable ID  
 Protein stable ID version  
 Transcript type  
 Version (transcript)  
 Version (protein)  
 Strand  
 Transcript start (bp)  
 Transcript end (bp)  
 Transcription start site (TSS)  
 Transcript length (including UTRs and CDS)

**Exon Information**

CDS Length  
 CDS start  
 Start phase  
 End phase

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Rename Download as: **Hs\_ENST00000589042.rna.fa**

Screenshot of the BioMart interface for retrieving data from ENSEMBL.

The left sidebar shows the following filters:

- Dataset:** Human genes (GRCh38.p13)
- Filters:** Transcript stable ID(s) [e.g. ENST00000000233]: [ID-list specified]
- Attributes:** Gene stable ID, Transcript stable ID, Protein stable ID version, cDNA sequences
- Dataset:** [None Selected]

The main panel includes the following controls:

- Export all results to: File  Go
- Email notification to:
- View: 10 rows as FASTA  Unique results only

The FASTA sequence output is as follows:

```
>ENSG0000155657|ENST00000589042|ENSP00000467141.1
GAGCAGTCGTGCATTCCCAGCCTCGCCTGGGTAGGGATTGCATAGAAAAGCAAAACT
ACACAGTCTTGACTGTGTAGTTTGTAGGATTAGAGGCTCACCGATTATGTCGGA
GATGGTCAGAAAAACCAACTCTCCATAGGACGTCTTCAGAAGCAACCTTGGGCTTAGT
CCCACCCTTTTAGGCACTCTTGAGAAATCAGAGTGCCTAGAAAGATGACAACCTCAAGCA
CCGACGTTTACGCCGTTACAAAGCCTGTGGTACTGGAGGGTAGTACCGCAACCTTT
GAGGCTCACATTAGTGGTTTCCAGTTCTGAGGTGAGCTGGTTAGGGATGGCCAGGTG
ATTTCCACTTCACTCTGCCCGCGTGCAGATCTCCTTAGCGATGGCCGCGCTAAACTG
ACGATCCCCGCCGTGACTAAAGCCAACAGTGGACGATATCCCTGAAAGCCACCAATGGA
TCTGGACAAGCGACTAGTACTGCTGAGCTCTCGTGAAGCTGAGACAGCACCACCAAC
TTCGTTAACGACTGCAGAGCATGACCGTGAGACAAGGAAGCCAAGTGAGACTCCAAGTG
AGAGTGACTGGAATCCCTACACCTGTGGTAAGTTCTACCGGGATGGAGCCGAAATCCAG
AGCTCCCTGATTCAAATTTCACAAGAAGGCACCTCTACAGCTTACTGATTGAGAA
GCATACCCCTGAGGACTCAGGGACCTATTCACTAAATGCCACCAATAGCGTTGGAAGAGCT
ACTTCGACTGCTGAATTACTGGTTCAAGGTGAAGAAGAAGTACCTGCTAAAAAGACAAAG
ACAATTGTTCGACTGCTCAGATCTCAGAATCAAGACAAACCGAATTGAAAAGAAGATT
GAAGCCCACCTTGATGCCAGATCAATTGCAACAGTTGAGATGGTCATAGATGGTGGCGCT
GGGCAACAGCTGCCACATAAAACACCTCCAGGATTCTCCGAAGCCAAAGTCAAGATCC
CCAACACCAACCGTCTATTGCTGCCAAAGCACAGCTGGCTGGCAGCAGTCCCCATCGCCC
ATAAGACACTCCCCCTCCCCGGTCAGACACGTGCGGGCACCGACCCCATCTCGGTAGG
TCCGTGTCTCCAGCAGCAAGAATCTCCACATCCCCCATCAGGTCTGGTCTCCATTG
```

# Data Retrieval

## Retrieving Data From ENSEMBL

### BioMart

- Go to BioMart and Do the same for:

Mouse: **ENSMUSP00000097561**

- Name the resulting files as:
  - **Mn\_ENSMUSP00000097561.pep.fa**
  - **Mm\_ENSMUSG00000051747.dna.fa**
  - **Mn\_ENSMUST00000099981.rna.fa**

# Data Retrieval

## Retrieving Data From Uniprot

- Go to: [Uniprot](#)
- Search for:  
**danio rerio ribosomal protein 40s 60s**
- In a different tab search for:  
**danio rerio ribosomal protein 40s 60s AND reviewed:yes**
- Download Results (Follow instructions)  
**Fasta/Canonical/Compressed**  
**Fasta/Canonical & Isoform/Compressed**
- Upload results into Kaiser
- Rename Datasets:  
**DrRibosomalProtCan**  
**DrRibosomalProtCanIso**
- Check File Type  
Go to ‘Edit Atributes’ > “Datatype”  
Make sure it is set to ‘fasta’  
Save