Tutorial for the Exon Ontology website

Table of content

Outline

Step-by-step Guide

1. Preparation of the test-list

2. First analysis step (without statistical analysis)

2.1. The output page is composed of several tabs:

2.2. Link to FasterDB

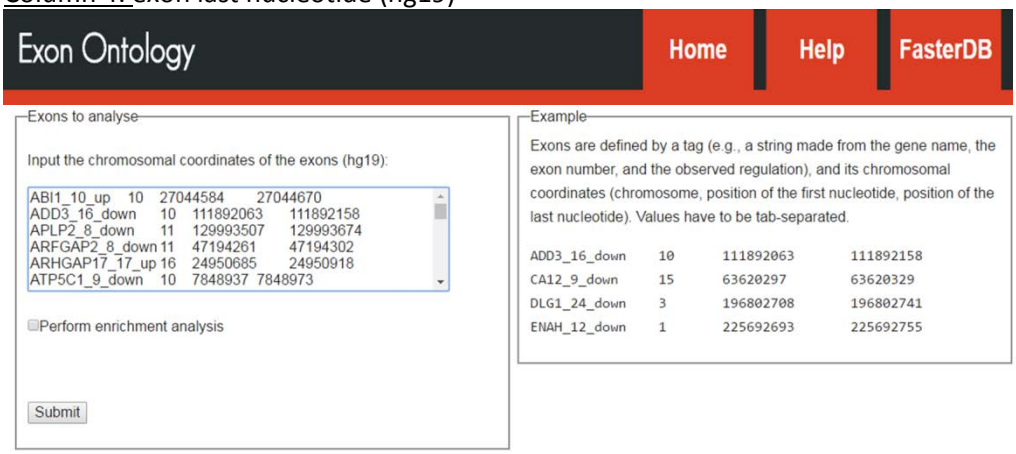
3. Second analysis step with statistical analysis

Outline: After having uploaded the chromosomal coordinates of selected exons in the Exon Ontology search engine, the Exon Ontology web suite provides several tables. The first two tables provide general genomic and protein information on the exons from the test-list, while the “Exon Annotations” table provides information regarding the protein feature(s) (EXONT terms) encoded by each individual exon. The “PTM annotations” table provides information regarding residues encoded by the selected exon set and bearing experimentally validated post-translational modifications. Finally, the “Protein-Protein network” table provides the list of proteins interacting with the products of the genes bearing at least one exon from the selected exon set. In addition, after having selected a control list (First Coding exons, Last Coding exons, Constitutive exons or Alternative exons), the “Functional features” table provides the score, Z-score and statistical significance for each EXONT term when compared to control exons.

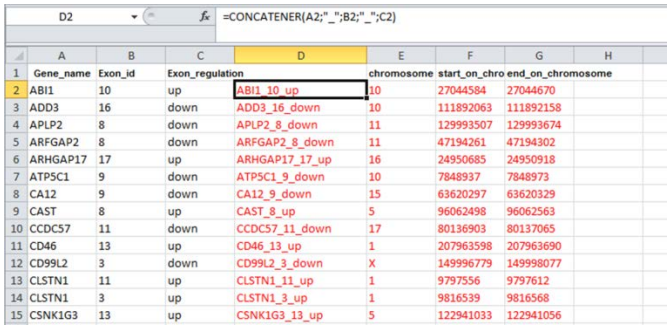
Step-by-step Guide

1. Preparation of the test-list

The Exon Ontology website opens on a window where the user can paste the list of exons to analyze, for example directly from a spreadsheet respecting the following input:
Column 1: name and characteristics of the exon (gene, exon number, regulation…). This column may contain only the gene symbol or any relevant information used to identify the exons (see below).
Column 2: chromosome
Column 3: exon first nucleotide (hg19)
Column 4: exon last nucleotide (hg19)



Below is an example of an spreadsheet that can be used as a template to format the information to download into the Exon Ontology search engine. From the spreadsheet, the red columns and lines are selected and then pasted into the Exon Ontology search engine. The column "D" (e.g., "ABI1_10_up") corresponding to the concatemerization of the gene symbol ("Gene_name"), exon number ("Exon_Id") and regulation ("Exon_regulation") is optional. The user may use only the gene symbol or any relevant information. This step can be easily made using the spreadsheet "CONCATENATE" function as shown below.



	A	B	C	D	E	F	G	H
1	Gene_name	Exon_id	Exon_regulation		chromosome	start_on_chro	end_on_chromosome	
2	ABI1	10	up	ABI1_10_up	10	27044584	27044670	
3	ADD3	16	down	ADD3_16_down	10	111892063	111892158	
4	APLP2	8	down	APLP2_8_down	11	129993507	129993674	
5	ARFGAP2	8	down	ARFGAP2_8_down	11	47194261	47194302	
6	ARHGAP17	17	up	ARHGAP17_17_up	16	24950685	24950918	
7	ATP5C1	9	down	ATP5C1_9_down	10	7848937	7848973	
8	CA12	9	down	CA12_9_down	15	63620297	63620329	
9	CAST	8	up	CAST_8_up	5	96062498	96062563	
10	CCDC57	11	down	CCDC57_11_down	17	80136903	80137065	
11	CD46	13	up	CD46_13_up	1	207963598	207963690	
12	CD99L2	3	down	CD99L2_3_down	X	149996779	149998077	
13	CLSTN1	11	up	CLSTN1_11_up	1	9797556	9797612	
14	CLSTN1	3	up	CLSTN1_3_up	1	9816539	9816568	
15	CSNK1G3	13	up	CSNK1G3_13_up	5	122941033	122941056	

In summary, users can copy-paste a list of exons in the Exon Ontology search engine, as follows:

ADD3_16_down	10	111892063	111892158
CA12_9_down	15	63620297	63620329
DLG1_24_down	3	196802708	196802741
ENAH_12_down	1	225692693	225692755

Or, as follows:

ADD3	10	111892063	111892158
CA12	15	63620297	63620329
DLG1	3	196802708	196802741
ENAH	1	225692693	225692755

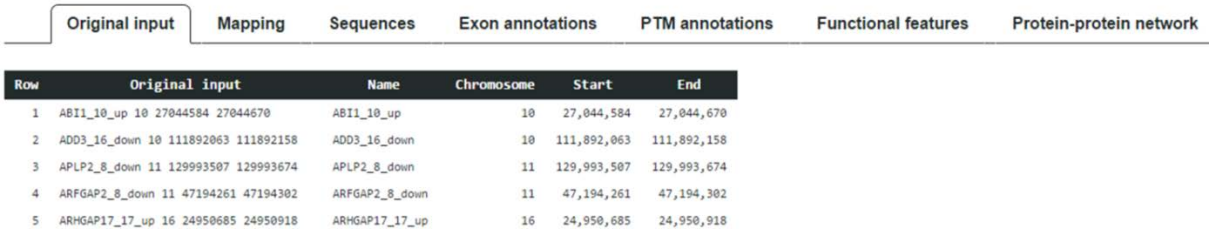
Before to "perform enrichment analysis", we recommend to first "submit" the exon set in order to get information concerning the nature of the submitted exons (i.e., first, internal, last coding exons or alternatively spliced exons). For this end, the user should not select "Perform enrichment analysis" and directly click the "Submit" button.
The analysis may take a few minutes if the list is long.

2. First analysis step (without statistical analysis)

2.1. Output pages:

After submitting the exon set, the Exon Ontology web suite provides 6 tables.

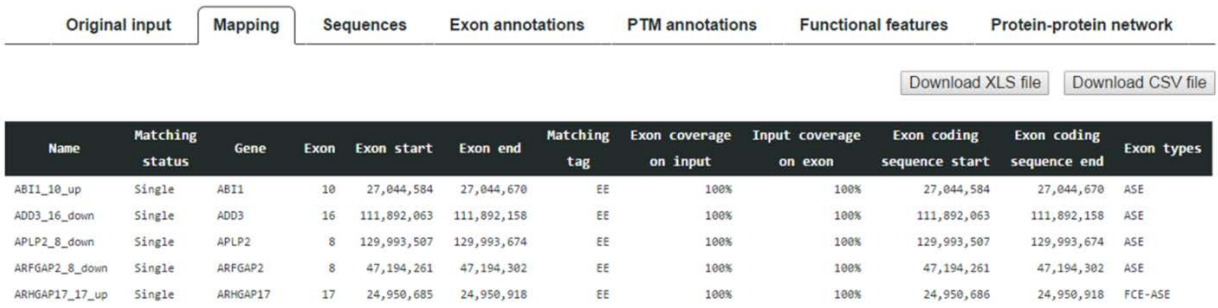
Tab 1. The "Original input" table provides genomic coordinates of the input exon list, and allows the user to verify that the input was correctly formatted.


Original input | Mapping | Sequences | Exon annotations | PTM annotations | Functional features | Protein-protein network

Row	Original input	Name	Chromosome	Start	End
1	ABI1_10_up 10 27044584 27044670	ABI1_10_up	10	27,044,584	27,044,670
2	ADD3_16_down 10 111892063 111892158	ADD3_16_down	10	111,892,063	111,892,158
3	APLP2_8_down 11 129993507 129993674	APLP2_8_down	11	129,993,507	129,993,674
4	ARFGAP2_8_down 11 47194261 47194302	ARFGAP2_8_down	11	47,194,261	47,194,302
5	ARHGAP17_17_up 16 24950685 24950918	ARHGAP17_17_up	16	24,950,685	24,950,918

Tab 2. The "Mapping" table provides exon mapping to FasterDB exons. The "Matching tag" column indicates that the 5' and 3' borders of each tested exon exactly (E) or partially (P or T) match an exon annotated in FasterDB (P indicates that the user sequence is shorter than the FasterDB exon, while T indicates the opposite). It is followed by the coverage (as a percentage) between the tested exon and the FasterDB exon.

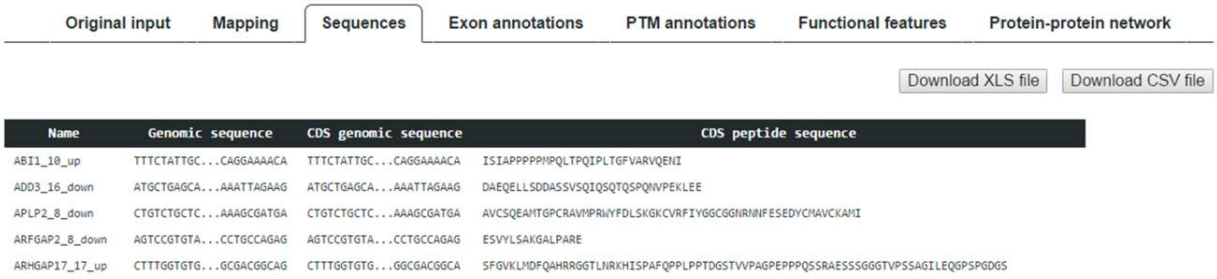
The last column indicates to which category of exons (i.e., "First coding exons" (FCE), "Last coding exons" (LCE), "Alternatively Spliced Exons" (ASE), and/or "Constitutive Exons" (CE)) the exons from the test list belong to according to the FasterDB annotation. Of note, some exons can belong to several categories. For example, some exons may have been classified as FCE and as ASE, depending on their position within different alternative transcripts. This information is important for selecting the most appropriate control for statistical analysis (see "**3. Second analysis step with statistical analysis"**).


Original input | Mapping | Sequences | Exon annotations | PTM annotations | Functional features | Protein-protein network

Download XLS file | Download CSV file

Name	Matching status	Gene	Exon	Exon start	Exon end	Matching tag	Exon coverage on input	Input coverage on exon	Exon coding sequence start	Exon coding sequence end	Exon types
ABI1_10_up	Single	ABI1	10	27,044,584	27,044,670	EE	100%	100%	27,044,584	27,044,670	ASE
ADD3_16_down	Single	ADD3	16	111,892,063	111,892,158	EE	100%	100%	111,892,063	111,892,158	ASE
APLP2_8_down	Single	APLP2	8	129,993,507	129,993,674	EE	100%	100%	129,993,507	129,993,674	ASE
ARFGAP2_8_down	Single	ARFGAP2	8	47,194,261	47,194,302	EE	100%	100%	47,194,261	47,194,302	ASE
ARHGAP17_17_up	Single	ARHGAP17	17	24,950,685	24,950,918	EE	100%	100%	24,950,686	24,950,918	FCE-ASE

Tab 3. The “Sequence” tab gives the nucleotidic sequence of each tested exon, the corresponding coding sequence (which can be shorter in the case of a 3’ terminal exon), as well as the most frequently represented aminoacid sequence encoded by the tested exon.



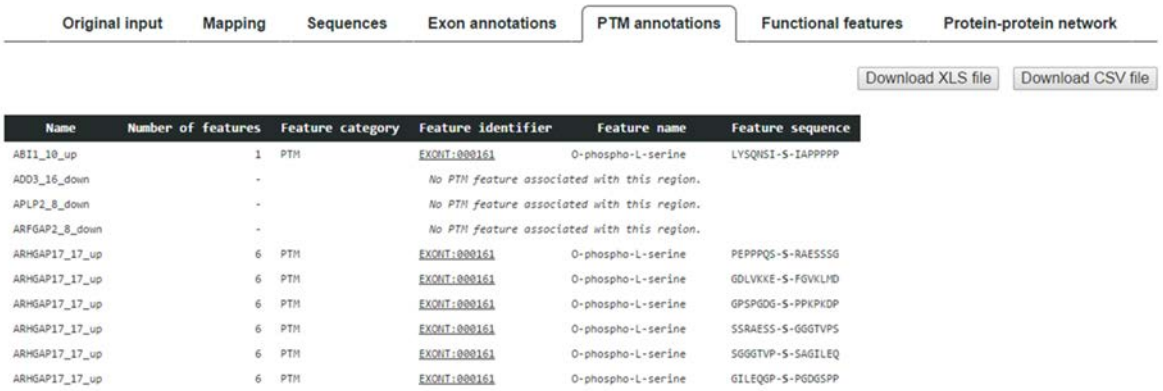
Name	Genomic sequence	CDS genomic sequence	CDS peptide sequence
ABI1_10_up	TTTCTATTGC...CAGGAAAACA	TTTCTATTGC...CAGGAAAACA	ISIAPPPPPMPQLTPQIPLTGFVARVQENI
ADD3_16_down	ATGCTGAGCA...AAATTAGAAG	ATGCTGAGCA...AAATTAGAAG	DAEQELLSDDASSVSQIQSQTQSPQNVPEKLEE
APLP2_8_down	CTGTCTGCTC...AAAGCGATGA	CTGTCTGCTC...AAAGCGATGA	AVCSQEAMTGPCRAVMPRWYFDLSKGKCVRFIYGGCGGNRNNFESEDYCMAVCKAMI
ARFGAP2_8_down	AGTCCGTGTA...CCTGCCAGAG	AGTCCGTGTA...CCTGCCAGAG	ESVYLSAKGALPARE
ARHGAP17_17_up	CTTTGGTGTG...GCGACGGCAG	CTTTGGTGTG...GGCGACGGCA	SFGVKLMDFQAHRRGGTLNRKHISPAFQPPLPPTDGSTVVPAGPEPPPQSSRAESSSGGGTVPSSAGILEQGPSPGDGS

Tab 4. The “Exon annotations” tab provides the list of all features associated with each tested exon. This list can be downloaded as a XLS or CSV file and subsequently managed by the user using a button at the top of the screen. Each line corresponds to one given feature, even if a given exon harbors several features. For each exon, the total number of features is given in Column 2. The other columns give information about the feature’s main category (Catalytic activity, PTM, Receptor activity, Binding, Structure, Transporter, Localization or Unclassified), the feature identifier and the feature name.

***Note**: The feature identifier is an hypertext link that redirects the user to a web page providing information about the nature of each feature (corresponding computational tool, original database, references, etc…).*

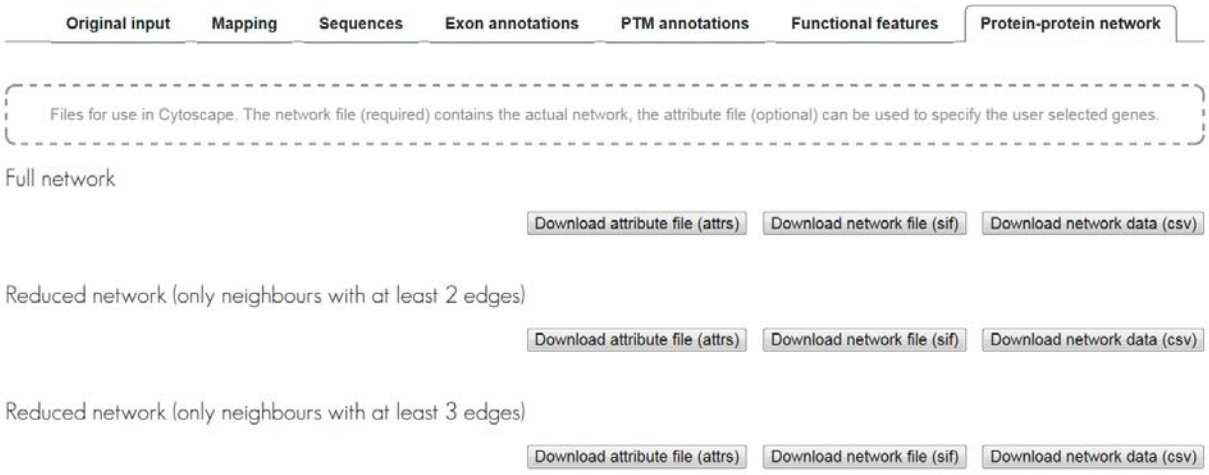
Name	Number of features	Feature category	Feature identifier	Feature name
ABI1_10_up	4	Binding	EXONT:001885	Protein binding
ABI1_10_up	4	Structure	EXONT:000024	Compositionally_biased_region_of_peptide
ABI1_10_up	4	Structure	EXONT:000074	Intrinsically_unstructured_polypeptide_region
ABI1_10_up	4	PTM	EXONT:000161	O-phospho-L-serine
ADD3_16_down	1	Structure	EXONT:000074	Intrinsically_unstructured_polypeptide_region
APLP2_8_down	4	Catalytic activity	EXONT:003180	Proteinase inhibitor I2, Kunitz metazoa

Tab 5. The "PTM" tab provides specific details for each annotated post-translational modification predicted from the list of tested exons. This includes the modified aminoacid residue and its flanking aminoacids (7 aminoacids on each side). Again, the list of PTMs can be downloaded as a CSV file, allowing to easily recover the list of PTMs. This is useful for example to search for a potential consensus aminoacid sequence corresponding to a known protein kinase site.



Name	Number of features	Feature category	Feature identifier	Feature name	Feature sequence
ABI1_10_up	1	PTM	EXONT:000161	O-phospho-L-serine	LYSQNSI-S-IAPPPPP
ADD3_16_down	-		*No PTM feature associated with this region.*		
APLP2_8_down	-		*No PTM feature associated with this region.*		
ARFGAP2_8_down	-		*No PTM feature associated with this region.*		
ARHGAP17_17_up	6	PTM	EXONT:000161	O-phospho-L-serine	PEPPPQS-S-RAESSSG
ARHGAP17_17_up	6	PTM	EXONT:000161	O-phospho-L-serine	GDLVKKE-S-FGVKLMD
ARHGAP17_17_up	6	PTM	EXONT:000161	O-phospho-L-serine	GPSPGDG-S-PPKPKDP
ARHGAP17_17_up	6	PTM	EXONT:000161	O-phospho-L-serine	SSRAESS-S-GGGTVPS
ARHGAP17_17_up	6	PTM	EXONT:000161	O-phospho-L-serine	SGGGTVP-S-SAGILEQ
ARHGAP17_17_up	6	PTM	EXONT:000161	O-phospho-L-serine	GILEQGP-S-PGDGSPP

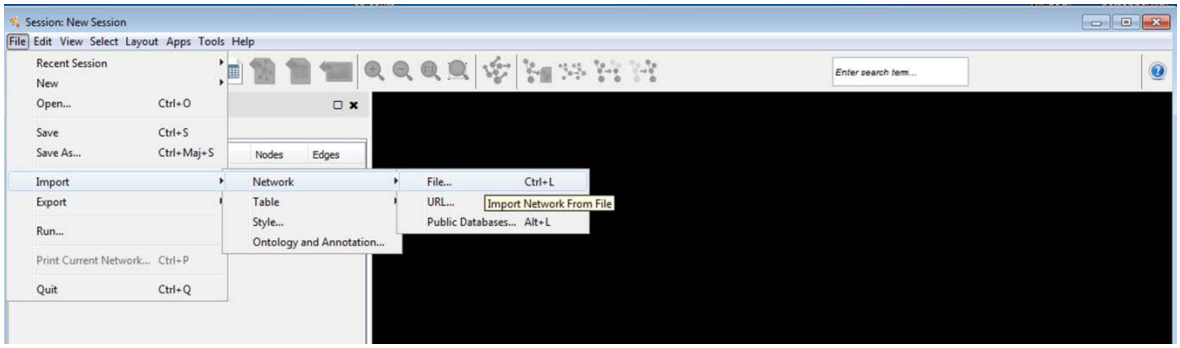
Tab 6. The "Protein-Protein network" tab allows to get the list of proteins interacting with the products of the genes bearing at least one exon from the selected exon set.



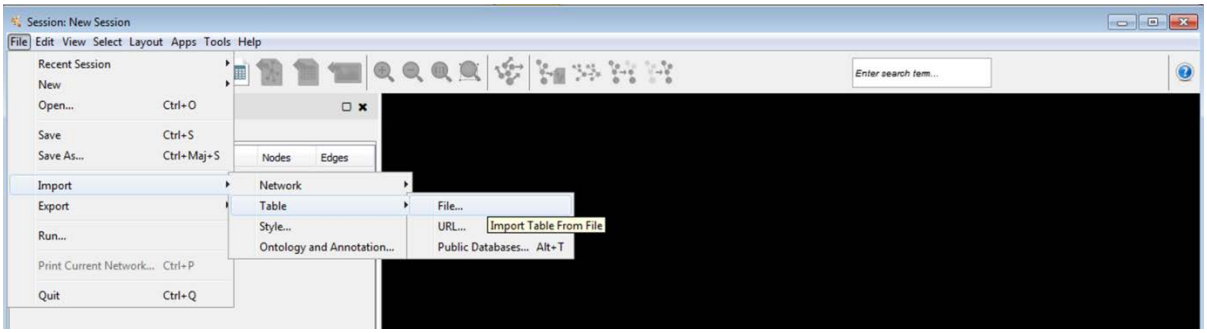
Three files can be downloaded. The ".csv" file ("network data") gives to the user the list of all partners retrieved from the Intact database for each of the genes bearing at least one exon from the user's exon list. The ".sif" file ("network file") can be directly used to build an interaction network in Cytoscape, an open source platform for complex network analysis (http://www.cytoscape.org/). Finally, the ".attrs" file ("attribute file") is useful in Cytoscape to annotate the proteins of the built network that are from the "test list".

To ease the interpretation of the network, which can be extremely complex if the number of tested exons is high, it is possible from the same Exon Ontology tab to obtain the same kind of information, but with a reduced list of protein partners. In this case, single-edged interactions ("neighbours with at least 2 edges") or single- and double-edged interactions ("neighbours with at least 2 edges") have been removed.

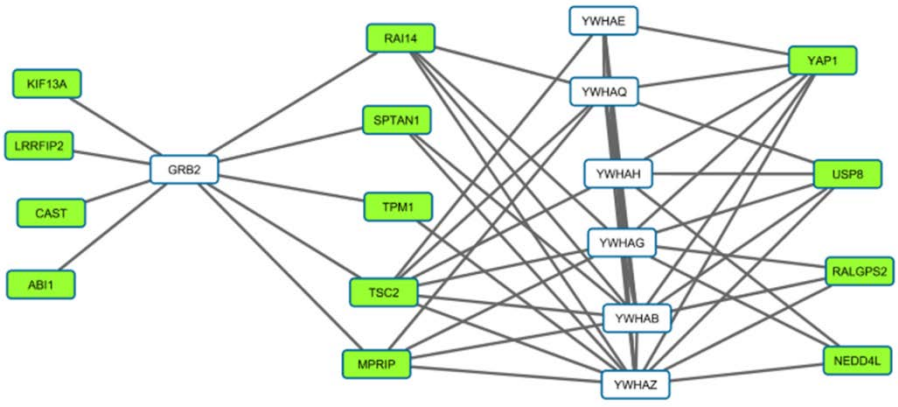
To build an interaction network in Cytoscape, upload the ".sif" file ("Import", "Network", "File") as shown on the screenshot.



To annotate the proteins of the built network that are from the "test list" in Cytoscape, upload the ".attrs" file ("Import", "Table", "File") as shown on the screenshot.



A network generated by Cytoscape is shown below. Green proteins belong to the test-list.



2.2. Link to FasterDB

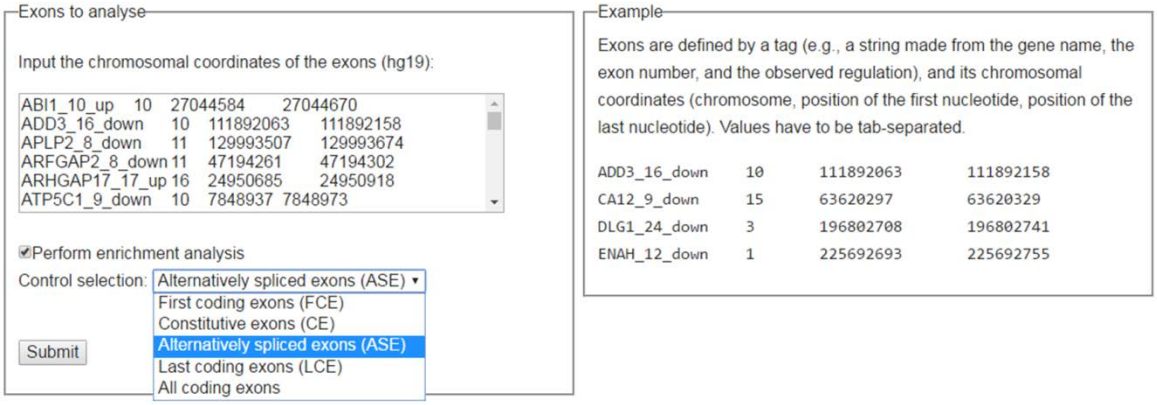
The Exon Ontology web page has a link to redirect the user to the FasterDB database (Mallinjoud *et al*., 2014), in which we have integrated Exon Ontology-related features. The user can enter the name of his/her favourite gene and get access to a series of information related to alternative splicing regulation of the gene transcripts, as described previously (Mallinjoud *et al*., 2014).
Below the scheme of the gene's global structure are several tabs that correspond to Exon Ontology main functional categories (Catalytic activity, Receptor, Transporter, Binding, Localization motifs, Structural elements, PTM or Unclassified). Clicking on any of these tabs will give instant access to detailed information about the functional features annotated in FasterDB for each exon of the gene. A mark under a given exon indicates the presence of the corresponding feature.

3. Second analysis step with statistical analysis

As indicated above, to perform statistical analysis the user must select the appropriate background control list. It is thus necessary to go back to the "Home" page, to select "Perform enrichment analysis" as detailed below. To select the control list(s), the user must first analyze the downloaded "Mapping" table (see above, Tab 2).

If the user's test list contains a mixture of "ASE" (alternatively spliced exons) and "CE" (constitutive exons) in the downloaded table, we recommend to perform parallel analyses using either "Alternative spliced exons" or "Constitutive exons" as controls.

If the user's test list contains also a large number of "FCE" (First coding exons) and/or "LCE" (Last coding exons), we recommend to generate separated exon lists for further analyses.



Once the "Background selection" has been made, click on the "Submit" button.

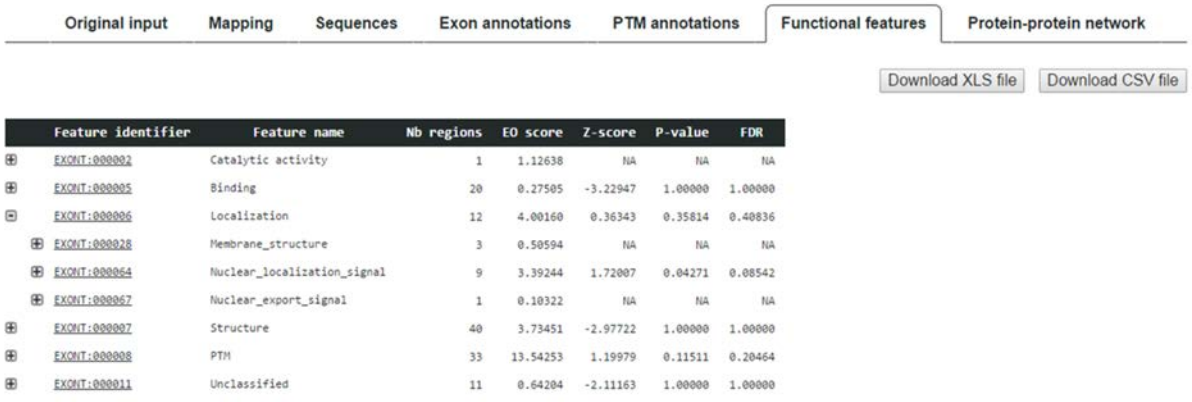
The statistical analysis may take a few minutes if the list is long.

The user must next click on the "Functional feature" tab.

The "Functional features" tab first allows a quick view of the statistical representation for each of the general categories of features in the list of tested exons. For each category, the total number of identified regions is given, followed by statistical parameters that include a global score, a Z-score, a P-value and a False Discovery Rate (FDR), as described in the paper.

From the same tab, the user can move down in the Exon Ontology tree by clicking on the "+" buttons, which gives access to the different sub-categories and to their corresponding statistical representation. The complete list of features can be downloaded as a XLS or CSV file.

***Note**: This is important to remember that even if a general feature category does not exhibit any enrichment or impoverishment compared to the control set of exons, this is not necessarily true for each of its specific sub-categories.*



Feature identifier	Feature name	Nb regions	EO score	Z-score	P-value	FDR
EXONT:000002	Catalytic activity	1	1.12638	NA	NA	NA
EXONT:000005	Binding	20	0.27505	-3.22947	1.00000	1.00000
EXONT:000006	Localization	12	4.00160	0.36343	0.35814	0.40836
EXONT:000028	Membrane_structure	3	0.50594	NA	NA	NA
EXONT:000064	Nuclear_localization_signal	9	3.39244	1.72007	0.04271	0.08542
EXONT:000067	Nuclear_export_signal	1	0.10322	NA	NA	NA
EXONT:000007	Structure	40	3.73451	-2.97722	1.00000	1.00000
EXONT:000008	PTM	33	13.54253	1.19979	0.11511	0.20464
EXONT:000011	Unclassified	11	0.64204	-2.11163	1.00000	1.00000