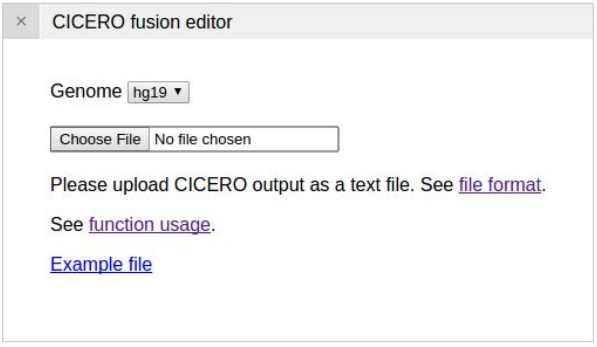
Review CICERO fusion gene results with Fusion Editor

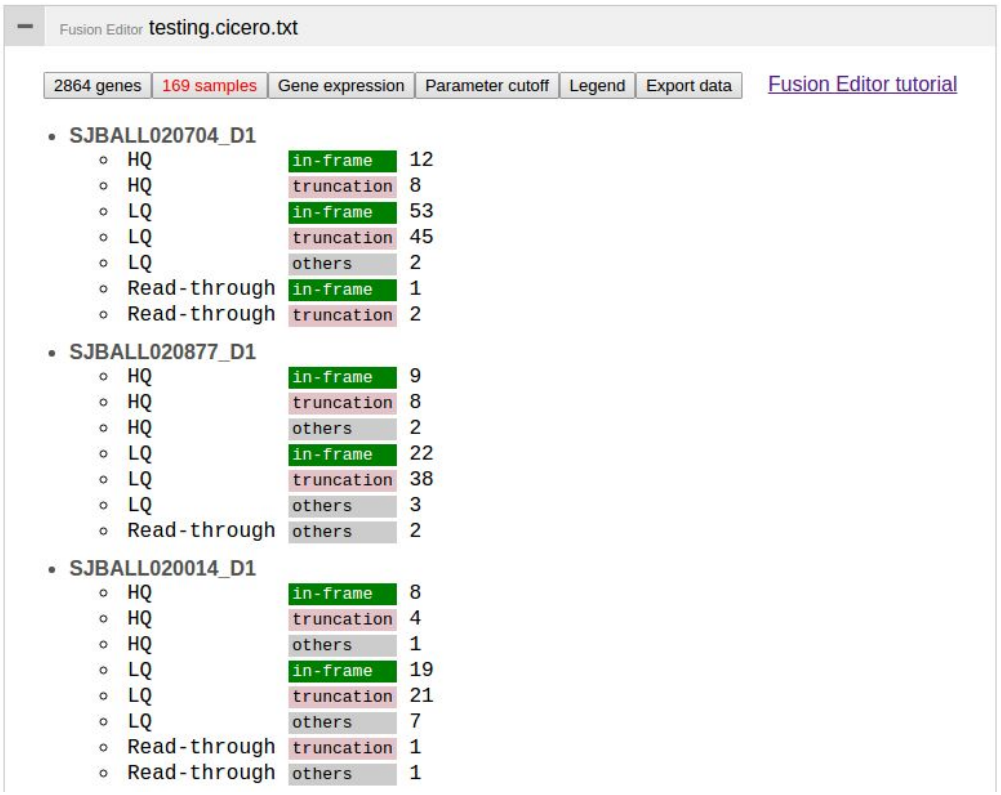
Loading files to the editor

To start, click on the “ProteinPaint” button from top left, and click the “Fusion Editor” from the list of applications:

Then, the file uploading box appears:



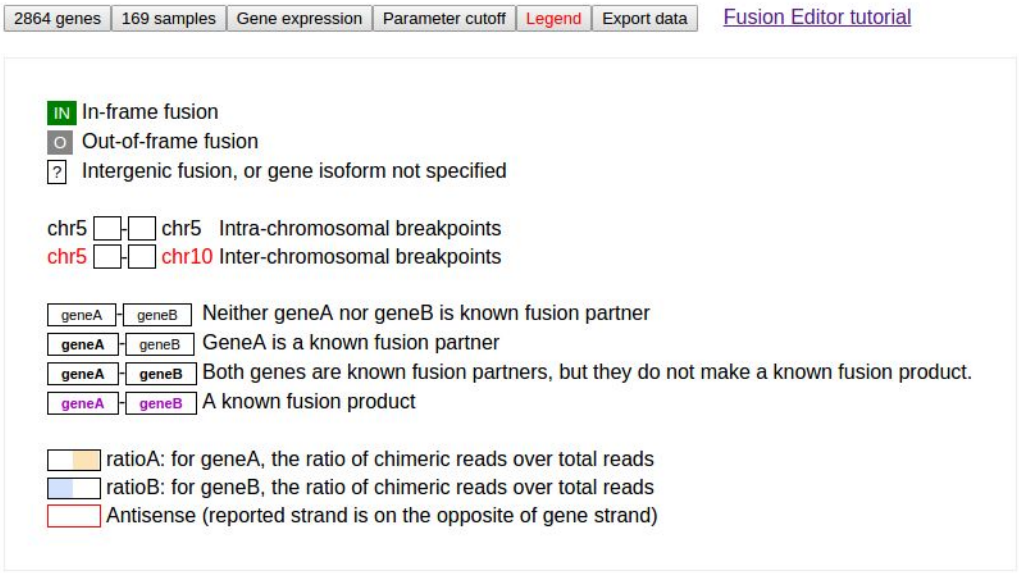
Load a CICERO output file to the editor, a new panel appears, displaying a row of buttons on top, and a summary section for each sample.



A row of buttons appears on the top. The buttons can be used to show/hide different components. By default the sample listing is displayed, and the sample button is in red as an indication. Click on the sample button to hide the sample listing.

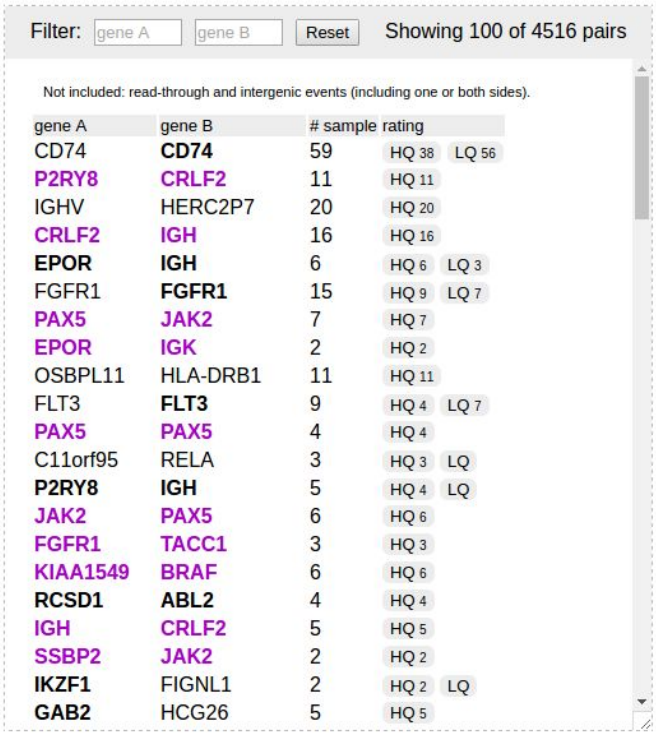
Fusion editor legend

Click the button Legend to view the legend:



Browsing by genes

Click on the gene button to show a table of unique fusions/gene pairs, ranked with most recurrent pairs on top:

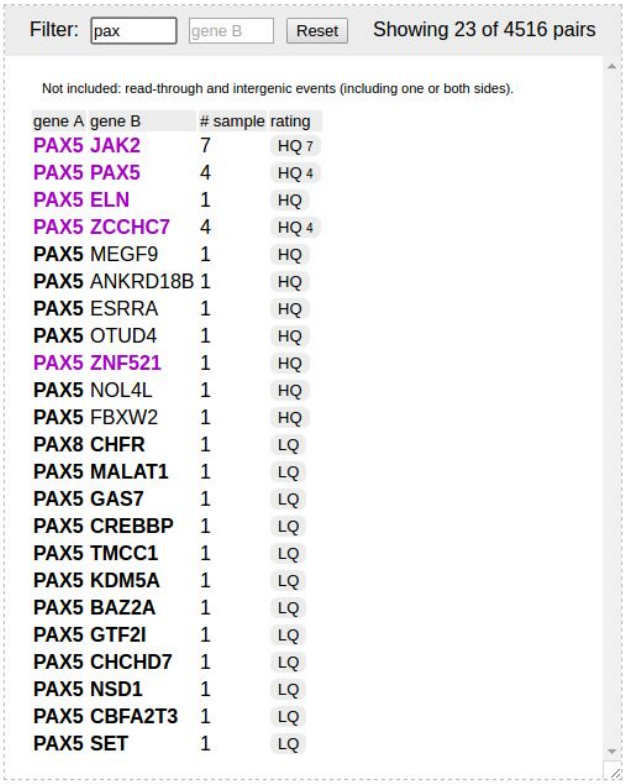


By default the table shows first 100 fusions. **Following are not shown in the table**:

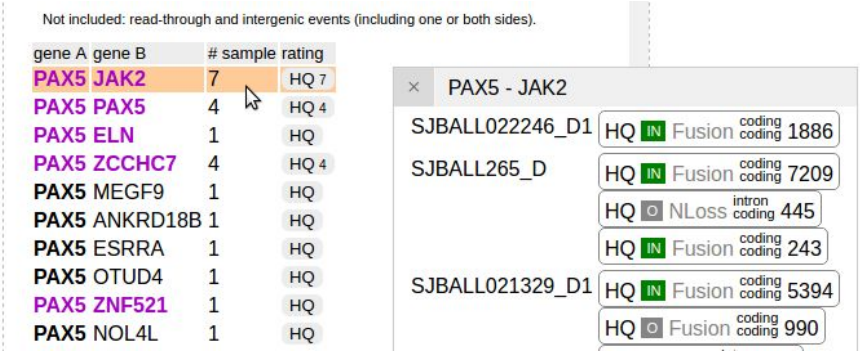
- Intergenic fusions, with either one or both intergenic breakends
- Read-through events

Bold font face and magenta color represents known fusion events, as explained in the legend.

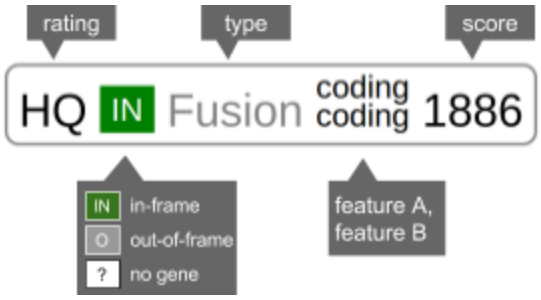
To filter this table and find interested genes, type in gene name as either **gene A** or **gene B** into the respective text boxes, the table will only show matching fusion events:



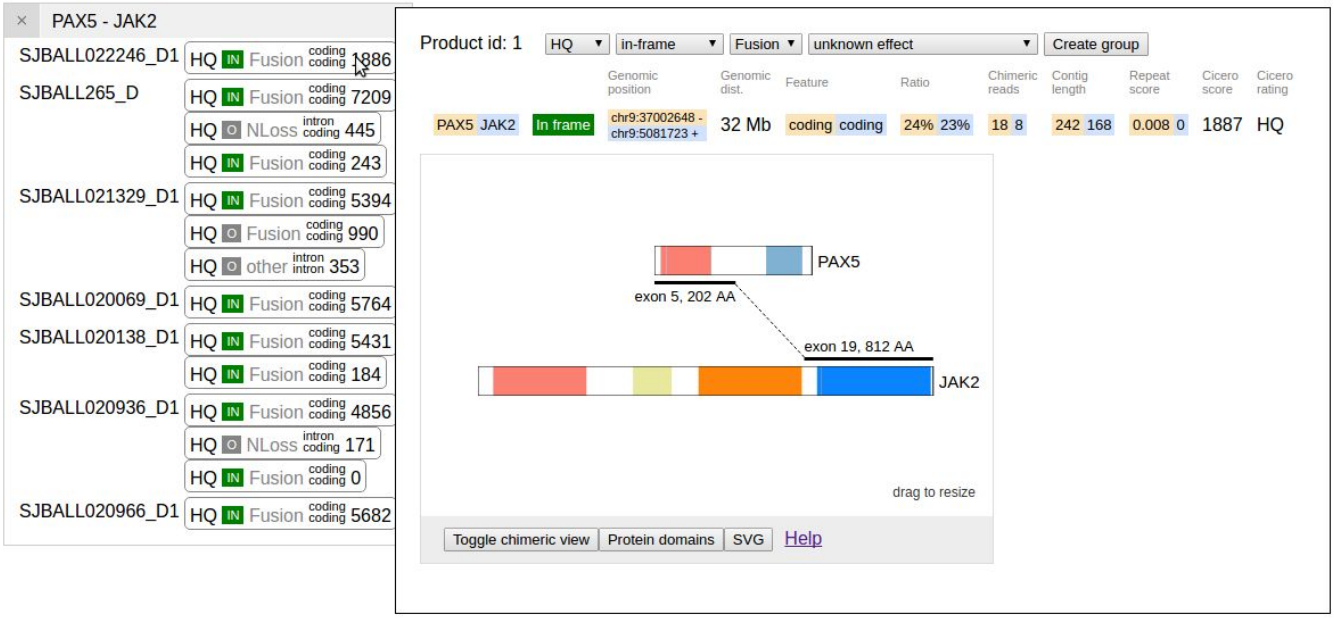
To view the fusion in all samples detected with this event, click on the table row. Following screenshot shows PAX5-JAK2 fusions from 7 samples:



Each fusion product is summarized using a set of vital metrics:



To view an interactive graphical display of a fusion event, and make edits in the meantime, mouse over a fusion product summary. The display is described in the following section:



User can also click on a fusion product to view all the details, explained in the following section:

sv_ort	>
coverageA	609
coverageB	337
qposA	242
qposB	242
contig	GAGCGGGTGTGTGACAATGACA ...»
type	ITX
medal	4
sv_refseqA_coding_base_number	601,601,601,601,277,60 ...»
sv_refseqA_last_coding_base_number	
sv_refseqB_coding_base_number	2435,2435,2435,2435,24 ...»
sv_refseqB_last_coding_base_number	3399,3399,3399,3399,33 ...»
sv_AA	ERVCDNDTVPSVSSINRIIRTK ...»
sv_desc	PAX5:NM_001280547[1,5] ...»
sv_processing_exception	
sv_general_info	geneB_transcripts=NM_0 ...»
sv_interstitial_AA	,,,,,,VSTGSVTQVSSVSTDS ...»
sv_frame_index	0,0,0,0,0,0,0,0,0,0,0

in-frame
frame code: 1

gene	isoform	gene position	exon	anchor	contig AA	contig bp
PAX5	NM_016734	codon: 202	5	perfect	ends at 80	ends at 242
JAK2	NM_004972	codon: 812	19	tuple	starts at 81	starts at 243

Browsing by samples

Click the sample button to show a full list of samples. In the list, samples with more in-frame fusions are moved towards the top.

In each sample the fusion events are assigned into 4 rating grades:

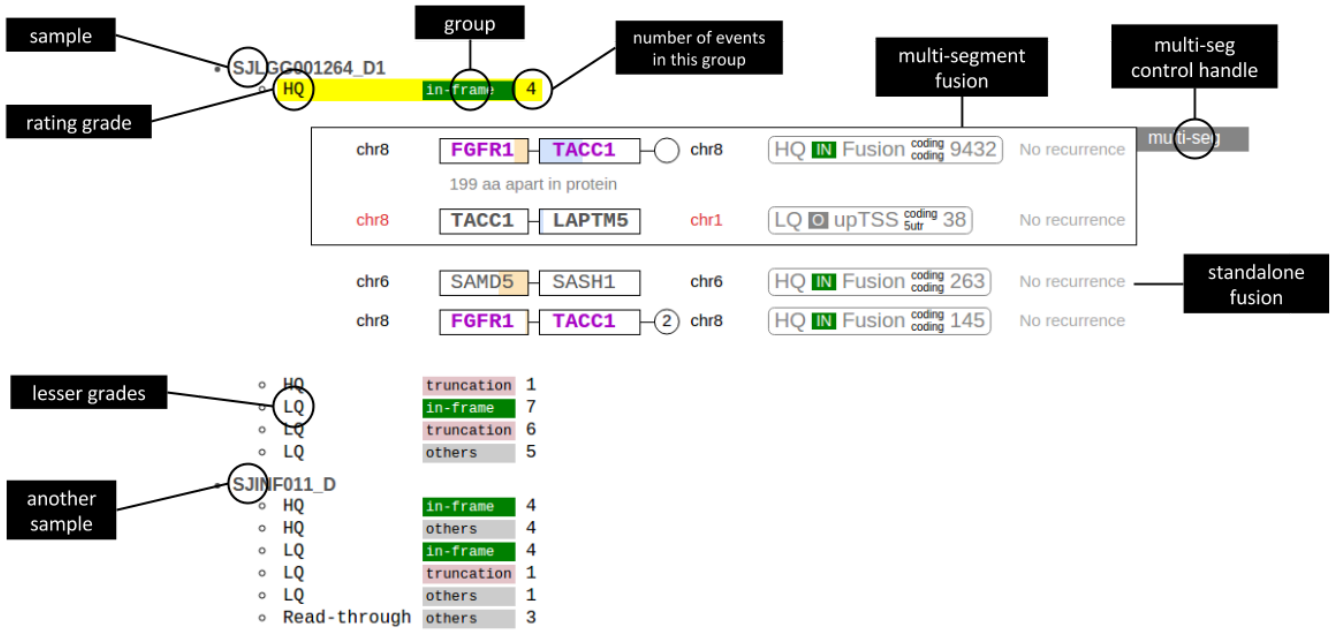
1. HQ
2. LQ
3. read-through
4. bad

Each grade is further separated into at most 3 sub-grades:

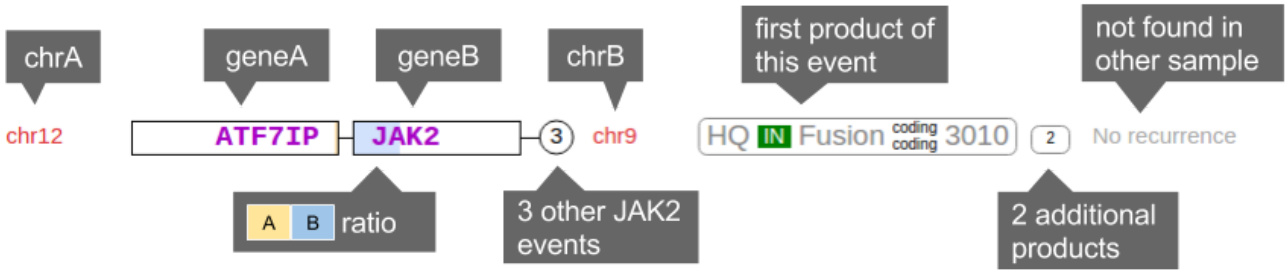
1. **In-frame fusion events**, this include any events provided with an in-frame RefSeq isoform pair (from geneA and geneB).
2. **Truncation**, this include:
    a. events with either “NLoss” (N-terminus loss) or “CLoss” (C-terminus loss), in such cases one end of the break point is inside a gene while the other is in intergenic region
    b. a gene-to-gene fusion but without any in-frame RefSeq isoform pair
3. **Others**, events that are neither in-frame or truncation

In the display, each row is a sub-grade. The number of fusion events is indicated in the end of each row.

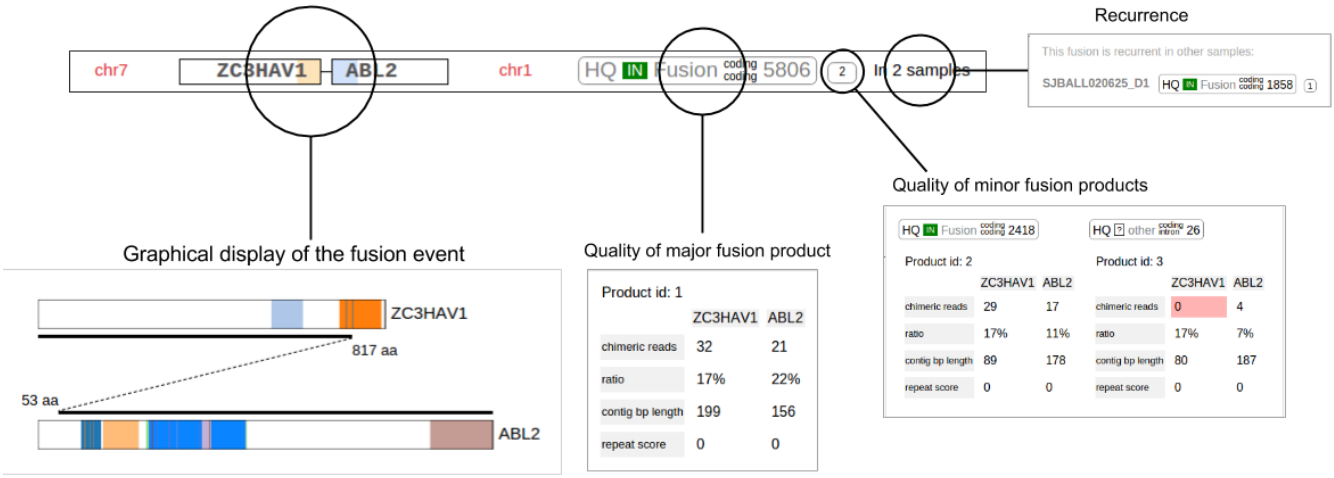
Click on a row to show the fusion events. The row will turn yellow upon clicking. Following is an illustrated example:



Fusion events display style

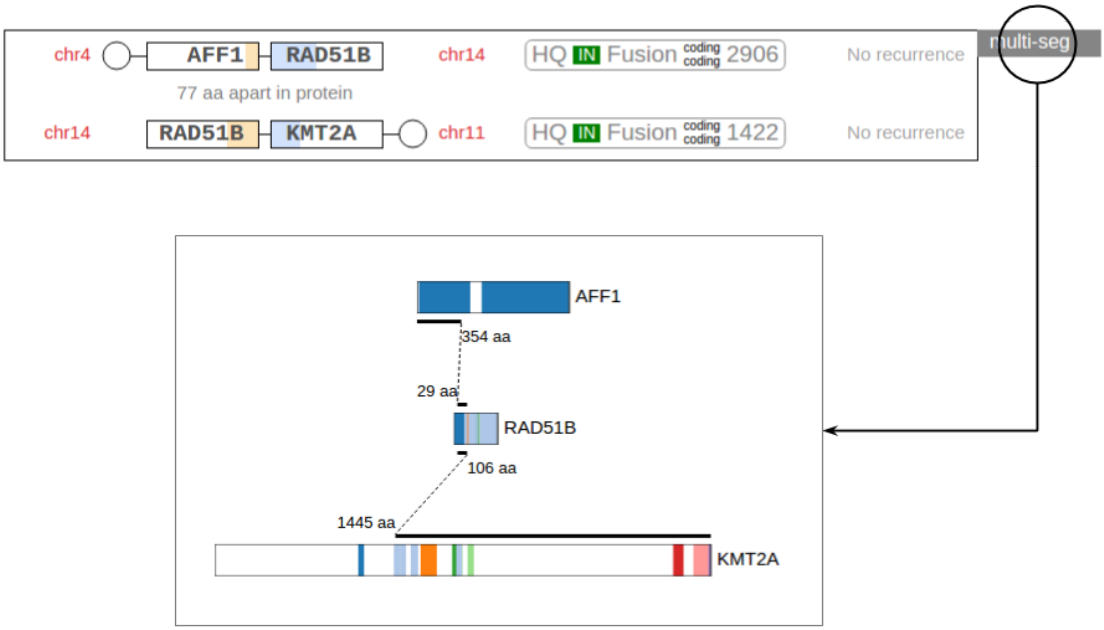


Mouse over various parts of the fusion display for different information:



For description on the graphical rendering of the fusion events, [view this tutorial](#).

Multi-segment fusions



ProteinPaint will automatically connect fusion events into multi-segment fusions if the 3’ breakend of the first fusion event is upstream of the 5’ breakend of the second fusion event.

Rules of joining two events (1 and 2):

- Event 1 chr B is the same as event 2 chr A
- Event 1 strand B is the same as event 2 strand A
- If both events are genic:

    - Event 1 gene B and event 2 gene A must share the same RefSeq gene isoform
    - Event 1 breakpoint position B must be upstream of event 2 breakpoint position A with respect to the transcription direction of shared gene isoform
- If both events are intergenic:
    - Event 1 breakpoint B must be upstream of event 2 breakpoint A with respect of the shared strand
- Chimeric read counts are all above 0, and show less than 5 fold difference

Other notes:

- Both gene fusion and intergenic fusions can be joined into multi-segment fusions.
- The joined fusion events will be represented together in a black-lined box. Category assignment is based on the highest rating and existence of in-frame events in the group.
- Between the two connected events, a text displays the inferred genomic or protein sequence distance between the two break points.
- ITD (internal tandem duplication) are not joined.
- Click on the black button on right to get the graphical display of multi-segment event.
- When exporting data, multi-segment fusions will be recorded.

Reciprocal fusions

Reciprocal fusions are a pair of fusion events generated by the balanced translocation between two genes. ProteinPaint automatically detects such A-B and B-A fusions and show them in a dotted-lined box.

When exporting data, reciprocal fusions will not be recorded. Contrary to the solid box of multi-segment fusion, the dotted box is suggest that it is “inconsequential”.

Gene expression data

The fusion genes are detected from RNA-seq results, from which the gene expression data can be readily derived. When displayed alongside fusion events, they provide useful clues for deciding if a fusion is true or not.

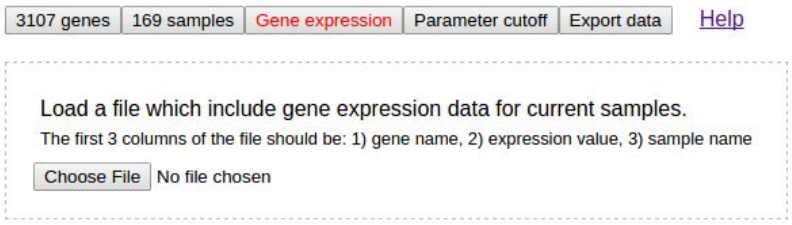
ProteinPaint supports uploading the gene-level expression data after a CICERO result file has been loaded.

Gene-level expression data is in a tabular text file with columns 1-3 must be gene name, expression value (e.g. FPKM), and sample name.

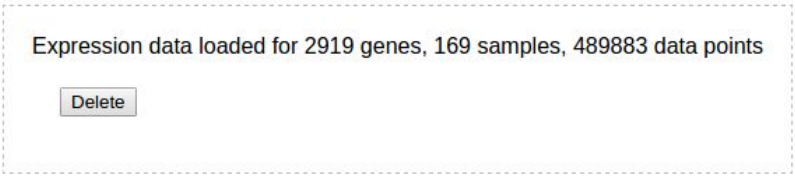
Notes:

- Arbitrary genes and samples can be provided in the expression data file. But user needs to filter the file to retain only relevant data and reduce the file to a size that user's web browser can manage.
- Genes that are not involved in any fusion should be excluded from the expression file, since their expression data will never be displayed, and can help reduce file size.
- Samples without any fusions are suggested to be included. This can provide a reference to be compared against samples with fusion genes.

Click the button Gene expression to show the file uploading interface:

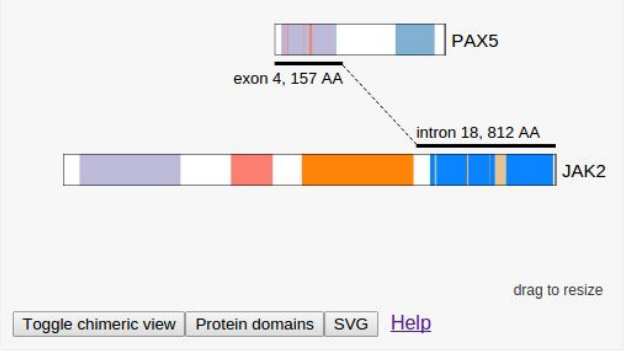


If the file is loaded successfully, a message like below will replace the uploading interface:



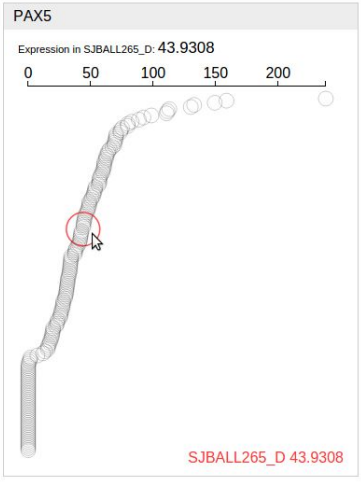
For reference, the file used in this example is 14 MB in size and is manageable by the web browser. To get this file, go to ….

Go to the gene or sample browsing panel and view a fusion event, one or more new panels will be displayed to show the expression of each gene in all available samples.



The expression data is displayed in a consistent style with the ProteinPaint display. Each circle represents a sample, it is shifted to right by a distance scaled by the gene's expression value relative to the cohort maximum. Samples are ranked according to the expression value. The red circle highlights the current sample.

Mouse over to see the sample name and value printed on the bottom right corner:



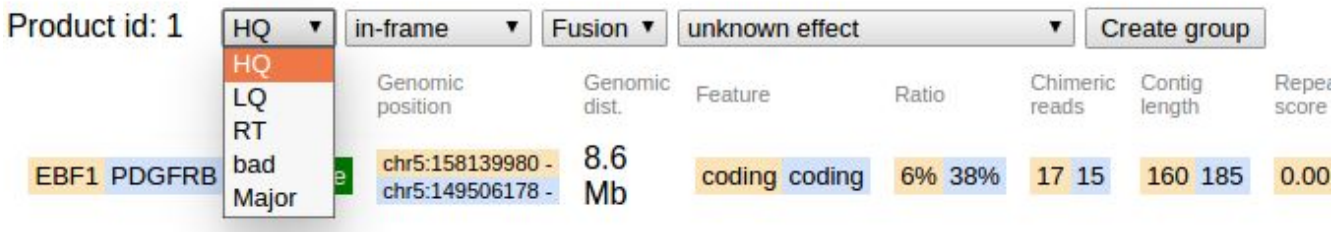
Drag on the text at the bottom right corner to resize the expression graph.

Manual editing

Editing individual events

Use the 4 drop-down menus on top to edit following attributes regarding an event:

- Rating grade
    - See “Exporting data” for the explanation on the “Major” category
- In/out of frame
- Fusion type
- Fusion effect



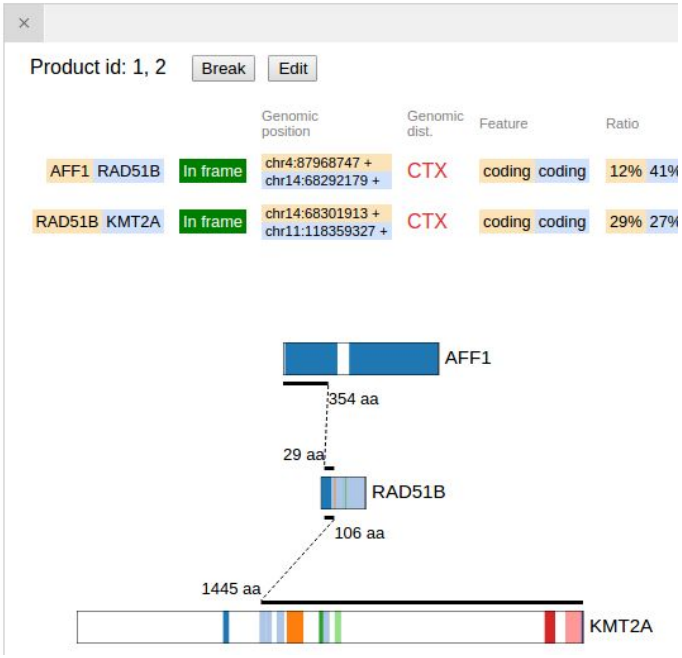
Fusion product id

ProteinPaint provides features for user to edit certain aspects of the fusion data. The “fusion product id” is designed to enable the editing. To see the id of a fusion product, mouse over a fusion event and find the id on top of the display panel:

Product id is continuous integers, it starts from 1 in each sample.

Editing a multi-segment fusion event

Click on the black flag on the right of the multi-segment fusion to get a panel showing editing options on the top, along with the graphical display:



To break the group, click “Break”. All fusion products from this group will now be shown standalone.

To edit, click “Edit”. In place of the button it now shows a text box:

The value “1,2” refers to the id of two fusion products in this group. By entering a third id “3”, a new fusion product can be joined to the group by the order of 1-2-3. Separate ids by comma.

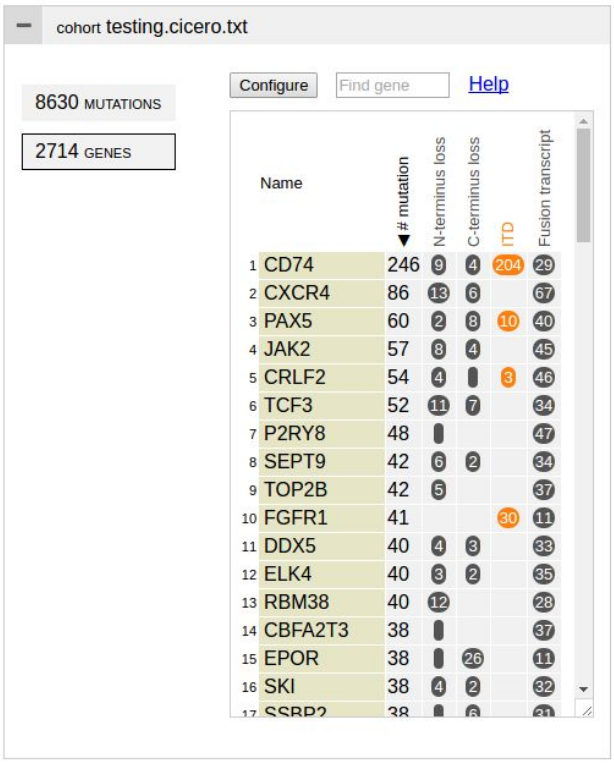
In this way, user can add or drop fusion products, and reorder the products within a group.

Creating a multi-segment group from a single product

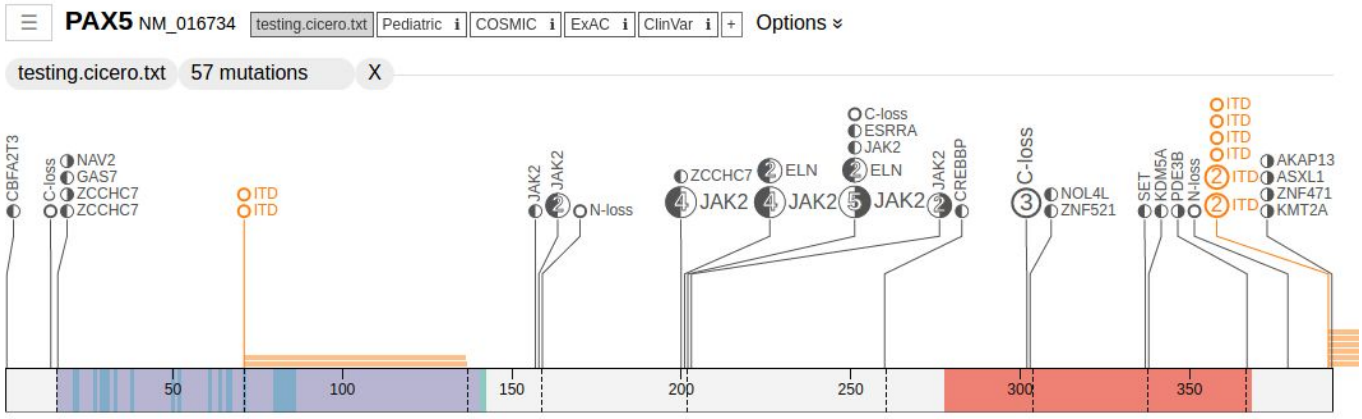
Mouse over any fusion product to see the “Create group” button on top of the new panel, as seen in Section 6.1. Click this button to get a text box containing the id of current product. Proceed to enter new id and create group.

Viewing results in ProteinPaint

Click on the button “Explort data” from the top row, then click a button “view in ProteinPaint”, a new panel like below will appear:

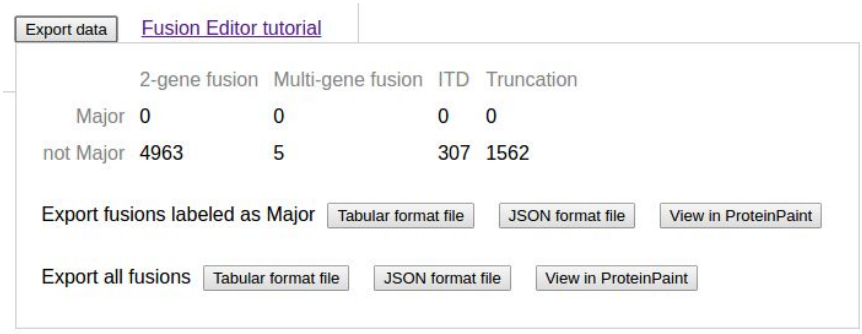


The panel shows a ranked gene list by number of fusion events, and the number of various types of fusions for each gene. Click on a gene name to see the fusion events on this gene:



Exporting data to text files

Click the button “Export data” to show the following menu:



The menu displays a small table summarizing number of fusion events available for export.

The first row of buttons will only use the fusion events marked as “Major”. Refer to the “Manual editing” section to see how to select and label an event as “Major”.

The second row of buttons will work with all the fusion events.

Click the button “Tabular format file” will yield a text file nearly identical to the input file, with a few more columns appended for describing the grouping of individual events into multi-segment events.

Click the button “JSON format file” will yield a JSON/text mixed format file. It is for loading into a SQLite database for querying by gene or isoform names. Such database can be one of the “queries” of a dataset in ProteinPaint. See tutorial here (TODO).

The table schema is:

```
CREATE TABLE rna_fusion (
     sample_name character varying(50) NOT NULL,
     genes text null,
     isoforms text null,
     fusions json not null
);
CREATE INDEX fusiongene on rna_fusion(genes);
CREATE INDEX fusionisoform on rna_fusion(isoforms);
```