

MageComet

Vincent Xue

August 11, 2011

1 Purpose

MageComet is a web application designed for quick annotation and manipulation of MAGE-TAB files. The webapp features tools that allow curators to easily edit MAGE-TAB documents, without spending excessive time and effort formatting files to MAGE-TAB specifications. MageComet's goal is to reduce the amount of time editing MAGE-TAB documents by automating tasks commonly encountered during curation. This automation allows curators to focus more on the biological data presented instead of spending time formatting the document.

2 Loading Files

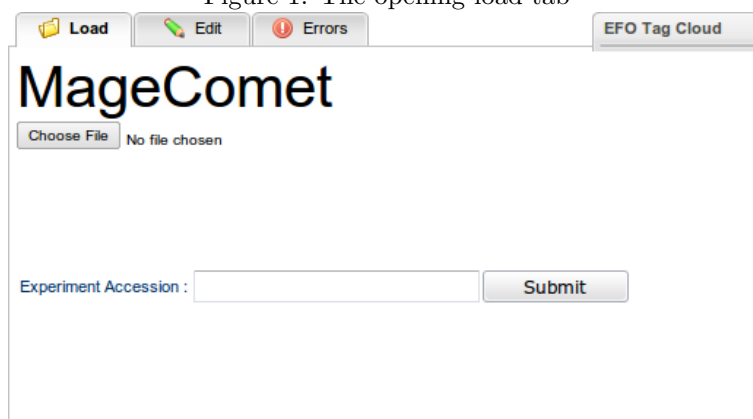
There are two ways to begin editing a set of MAGE-TAB files.

Direct Loading

Direct loading is used when MAGE-TAB documents are locally stored on the client machine. There are two files required to proceed, namely the SDRF, and IDF files which are usually appended by "sdrf.txt" and "idf.txt" respectively. Though the MAGE-TAB specification does not mandate that SDRF and IDF files have these suffixes, MageComet uses these suffixes during load, and will not proceed without them.

When starting the MageComet webapp, the user will be presented with the "Load Tab". This can be seen in Figure 1.

Figure 1: The opening load tab



To load the MAGE-TAB files to the server, the user must click the "Choose File" button. A popup will appear and the user can select either the SDRF file or the IDF file to upload first. The same "Choose File" button must be clicked after the first file has loaded. When both files have successfully been loaded, the screen should look like Figure 2.

Figure 2: Snapshot after loading IDF and SDRF files

ArrayExpress FTP

To load a file via the ArrayExpress FTP service the user can use the form directly below the "Choose File" button. The user can simply type in the experiment accession and click submit, which will automatically fetch and load the IDF and SDRF files.

3 Filtering Data

Once the files have been loaded the user can proceed to the "Edit" tab. If loading has been successful, the page should resemble Figure 3.

Figure 3: Edit after successful load

The first tool that is visible to the user is the filter and replace tool as shown in Figure 4. This tool is very similar to the search and replace tool in excel, but it is designed to be more specific for editing columns of data.

There are 5 components of this tool that a user can customize.

A Filter Column - This is the column that has a trait of interest. Usually this is the "Description" column.

B Logic - This dropdown box determines the logic a user wants to apply on the filter column. It contains a list of items such "equals", "contains", "does not contain" and more.

C Filter Value - This is the value the user will filter on.

Parts A, B, and C contain the logic for filtering.

Example: A user wants to filter for all the rows that contain female in the description column. The values for the A, B, and C would be A:Description, B:Contains, C:Female

D Target Column - The column whose value will be set

E New Value - The value that the target column will be set to.

Parts D and E contain the logic for replacing.

Example: A user wants to replace all the values in the column Characteristics[sex] with female. The values for D and E would be D:Characteristic[sex], E:female.

Note: The value for D must be a column that already exists

Figure 4: Filter and Replace

4 Column Manipulation

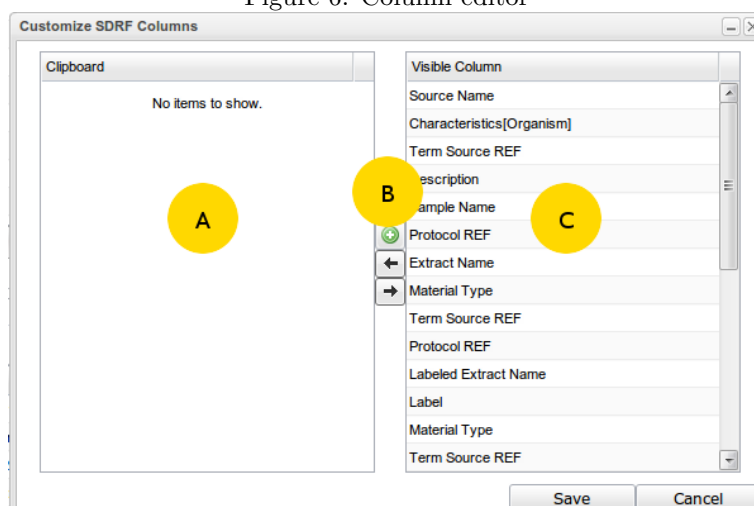
All column manipulation must be performed from the "Column Editor". The button to activate this editor is shown in Figure 5 and the popup editor is shown in Figure 6.

In this column editor, all of the rows are represented as rows. The left side of the editor designated by **A** is the clipboard. This section of the popup acts as a scratch buffer where columns are discarded to. In addition, new columns also appear in this clipboard before they are placed.

The right side of the editor designated by **C** is the representation of the SDRF columns. This section shows what columns are in the SDRF and in what order it appears.

Figure 5: Column editor button

Figure 6: Column editor



Adding Columns

To add a column, click the green plus sign below **B**. This will add a row to the clipboard **A**.

Removing Columns

To remove a column, drag the desired column from section **C** to section **A**. In addition, a user can also select the column from section **C** and click the ← button.

Reordering Columns

To reorder column orders, drag and drop the row to the desired position.

Renaming Columns

To rename a column, double click the row.

5 Extracting Data

The extract tool is a feature that splits row values formatted with delimiters. If all of the rows in a column are formatted similarly, this tool allows the user to separate a single column into many. An example of splittable column is seen in Figure 7.

In this example the "Description" column should be separated into new columns for organism part, disease stage, sex, age, set, and disease state. Automatic splitting is difficult and avoided because a single column could have many different combinations of delimiters. In this example alone, there are commas, semicolons, colons, and text delimiters in a single row. However, because this format is consistent throughout the column, the extract feature implemented in MageComet works well in these situations.

Figure 7: Example of a splittable column

Description
peripheral blood, n/a. uveitis status: No uveitis; gender: Female; age: 57; set: 1; tissue: Blood; disease state: Sarcoidosis
peripheral blood, n/a. uveitis status: Uveitis; gender: Female; age: 59; set: 1; tissue: Blood; disease state: Sarcoidosis
peripheral blood, n/a. uveitis status: No uveitis; gender: Male; age: 32; set: 1; tissue: Blood; disease state: Control
peripheral blood, n/a. uveitis status: Uveitis; gender: Male; age: 61; set: 1; tissue: Blood; disease state: Axial Spondyloarthritis
peripheral blood, n/a. uveitis status: Uveitis; gender: Male; age: 28; set: 1; tissue: Blood; disease state: Sarcoidosis
peripheral blood, n/a. uveitis status: Uveitis; gender: Male; age: 51; set: 1; tissue: Blood; disease state: Axial Spondyloarthritis
peripheral blood, n/a. uveitis status: Uveitis; gender: Male; age: 35; set: 1; tissue: Blood; disease state: Axial Spondyloarthritis
peripheral blood, n/a. uveitis status: No uveitis; gender: Female; age: 36; set: 1; tissue: Blood; disease state: Control
peripheral blood, n/a. uveitis status: Uveitis; gender: Female; age: 33; set: 2; tissue: Blood; disease state: Axial Spondyloarthritis
peripheral blood, n/a. uveitis status: Uveitis; gender: Male; age: 68; set: 1; tissue: Blood; disease state: Axial Spondyloarthritis

The extract tool is located in the same panel as the filter tool. By clicking on the "Extract" tab, the user can activate the panel that should resemble Figure 8. Like the filter tool, the extract tool has 5 components which the user must fill in.

- A** From Column - This is the column that is splittable. Usually this is the "Description" column.
- B** Left Input - This field should be filled in with the text that is left of the value that should be extracted. The input however should be **unique** as it will only match the first input found.
- C** Right Input - This field should be filled in with the text that is right of the value that should be extracted.
- D** Type Column - This field is the target column type. The possible choices are "Clipboard", "Characteristic", "Factor Value", or "Both".
- E** Column Name- This field is the target column name. This is the value that will fit in between the brackets [].

Figure 8: Extract

Filter

Extract

Input the character surrounding the value that will be extracted into a new column. To indicate the start of the row use "*" and to indicate the end of a row use "\$".

From : Left : Right : Type : New Column :

Sample Extract:

Key	Source Name	Characteristics[Organism]	Term Source REF	Description
1	GSE8192GSM201162	Homo sapiens	mo	HeLa shRNA- 30min 1: This is the sample of HeLa-shRNA cells without expression of short hairpin RNA, 30 min incubation with ActD. Strain: HeLa; Gender: Female
2	GSE8192GSM201162	Homo sapiens	mo	HeLa shRNA+ 120min 2: This is the sample of HeLa-shRNA cells with expression of short hairpin RNA, 120 min incubation with ActD. Strain: HeLa; Gender: Female
3	GSE8192GSM201194	Homo sapiens	mo	HeLa shLuc+ 120min 2: This is the sample of HeLa-shLuc cells with expression of short hairpin RNA, 120 min incubation with ActD. Strain: HeLa; Gender: Female
4	GSE8192GSM201163	Homo sapiens	mo	HeLa shRNA- 0min 3: This is the sample of HeLa-shRNA cells without expression of short hairpin RNA. Strain: HeLa; Gender: Female
5	GSE8192GSM201165	Homo sapiens	mo	HeLa shRNA+ 0min 3: This is the sample of HeLa-shRNA cells with expression of short hairpin RNA. Strain: HeLa; Gender: Female
6	GSE8192GSM201185	Homo sapiens	mo	HeLa shLuc- 120min 1: This is the sample of HeLa-shLuc cells without expression of short hairpin RNA, 120 min incubation with ActD. Strain: HeLa; Gender: Female
7	GSE8192GSM201146	Homo sapiens	mo	HeLa shRNA+ STV 1: This is the sample of HeLa-shRNA cells with expression of short hairpin RNA, cultured 24 hrs in serum-depleted medium. Strain: HeLa; Gender: Female

When the user completes values for **B** and **C**, the section designated by **F** will show the sample values that will be extracted.

More Examples

The following lists the input values for **B** and **C** that will extract the targeted value successfully for Figure 7.

Left	Right	Sample Extract Row 1
^	,	peripheral blood
status:	;	No uveitis
gender:	;	Female
set:	;	1
tissue:	;	Blood
state:	\$	Sarcoidosis

Table 1: Sample inputs and outputs for Figure 7

The following lists the input values for **B** and **C** that will extract the targeted value successfully for Figure 9.

Figure 9: Another example of a splittable column (intermediate difficulty)

Description
HeLa shRHAU- 30min 1: This is the sample of HeLa-shRHAU cells without expression of short hairpin RNA, 30 min incubation with ActD. Strain: HeLa; Gender: Female
HeLa shRHAU+ 120min 2: This is the sample of HeLa-shRHAU cells with expression of short hairpin RNA, 120 min incubation with ActD. Strain: HeLa; Gender: Female
HeLa shLuc+ 120min 2: This is the sample of HeLa-shLuc cells with expression of short hairpin RNA, 120 min incubation with ActD. Strain: HeLa; Gender: Female
HeLa shRHAU- 0min 3: This is the sample of HeLa-shRHAU cells without expression of short hairpin RNA. Strain: HeLa; Gender: Female
HeLa shRHAU+ 0min 3: This is the sample of HeLa-shRHAU cells with expression of short hairpin RNA. Strain: HeLa; Gender: Female
HeLa shLuc- 120min 1: This is the sample of HeLa-shLuc cells without expression of short hairpin RNA, 120 min incubation with ActD. Strain: HeLa; Gender: Female
HeLa shRHAU+ STV 1: This is the sample of HeLa-shRHAU cells with expression of short hairpin RNA, cultured 24 hrs in serum-depleted medium. Strain: HeLa; Gender: Female

Left	Right	Sample Extract Row 1
^	[0-9]	HeLa shRHAU-
[+]	min	30
gender:	;	Female
RNA,	min	30
Strain:	;	HeLa
Gender:	\$	Female

Table 2: Sample inputs and outputs for Figure 9

6 Tag Cloud

The tag cloud is a feature that helps curators identify important biological information text-mined from the IDF and SDRF text. To open the window, click on the button designated by Figure 10. A window should pop up that resembles Figure 11.

In the tagcloud representation of the text, each item represents an EFO ontology term that has been mined from 2 documents. The size of an item corresponds to where the term came from. The smallest size, indicates that only the IDF mentions the word. The medium sized text indicates that the text was mined from the SDRF, and the largest sized text indicates that both the IDF and SDRF mention the ontology term.

Adding characteristics via tagcloud

A sub feature of the tag cloud is the ability to add a characteristic term to all rows in a document. This feature is useful when some vital information is mentioned in the experiment description but is not mentioned in the SDRF document. By clicking on the term in the tag cloud, a new popup will show up, giving the user the option of adding the selected term to the SDRF. Figure 12.

Figure 10: Default tagcloud position

The screenshot shows the 'EFO Tag Cloud' window. At the top, there are tabs for 'Edit', 'Errors', and 'EFO Tag Cloud'. Below the tabs is a table with 6 columns. The first column contains terms, and the others contain metadata. A red arrow points to a small icon in the top right corner of the table area. Below the table, there is a 'Filter' section with a 'Source Name' dropdown, a 'Filter' button, and a 'Replace' button. A detailed popup is visible on the right side, showing the description for 'RNAi of human RHAU'.

1	2	3	4	5	6
RNAi of human RHAU-					
me]	The DEXH-box RNA he				
	unknown_experiment_1 RNAi profiling by array				
irm Source REF	EFO				
sReleaseDate]	2008-06-16				
:ession]	GSE8192				
ORE]	4				
sAccession]	E-GEOD-8192				
imeStamp_Version]	2010-08-08 01:24:49 L				
me					
pe					
rm Source REF					

RNAi of human RHAU (RNA helicase-associated with AU-rich element) is AU-rich element-mediated mRNA degradation. The finding mRNA degradation occurs in cytoplasm, prompted us to c localized throughout the nucleoplasm with some concentr Transcriptional arrest altered its localization to nucleolar c p72, suggesting that RHAU is involved in transcription-rel global gene expression either transcriptionally or posttrans prepared from RHAU-depleted HeLa cell lines, measuring ActinomycinD-chase. We found that most transcripts who show changes in their half-lives, suggesting the involve that RHAU has dual functions involved in synthesis and d Experiment Overall Design: HeLa-shRNAU and HeLa-shL the 4th day of dox-treatment, 4 clones of each cell line wei For the starvation experiment; medium was replaced with RNA. The ActD-chase experiment was done on the 6th ds was added to the medium at 5 ÅÅ/g/ml and total RNA was Samples for time 0, representing the total amount RNA in t in triplicates, whereas ActD-treated samples for the mRNA/ using the RNeasy kit from QIAGEN (Hombrechtikon, Swit each replicate was reverse transcribed and labelled using

Figure 11: The TagCloud Window

The screenshot shows the 'EFO Tag Cloud' window. It has a title bar 'EFO Tag Cloud' and two tabs: 'Weight By Location' and 'Highlight Mode'. The main area displays a list of terms in a tag cloud format. The terms are: control, labelling, protocol, gene, assay, submitter, submitted, irritable bowel syndrome, software, ulcerative colitis, transcription profiling by array, RNA, protocol parameter, Copenhagen, protein, DNA, transcription profiling, role, data transformation, cell type, array, descending colon, Homo sapiens, array design, organism part, whole organism, organism, colitis.

When a tag cloud item is clicked, a popup will appear, providing some useful information about the term. It usually provides a description near the top, which is pulled from the EFO ontology and the term source number.

The user can choose to add a characteristic column, term source ref column, or a term source number column to the SDRF, depending on how granular the curation is. The input field designated by **A** in Figure 12 is the value that will be placed in the brackets [].

Figure 12: Adding a characteristic via tagcloud

The screenshot shows a web form titled "Term: descending colon". It contains a description of the term and several checkboxes for adding characteristics. A yellow circle with the letter 'A' is placed over the "Column Title" input field.

Term: descending colon

Description:
The portion of the colon between the left colic flexure and the sigmoid colon at the pelvic brim; the portion of the descending colon lying in the left iliac fossa is sometimes called the iliac colon.

☒ Add Term as Value to All Records

☐ New Characteristic Column Title :

☐ Term Source REF Ontology : EFO

☐ Term Source Number Accession Number : EFO_0000845

Save Cancel

Highlighting text via tagcloud

Another sub feature of the tag cloud is to highlight text. This feature can be accessed by clicking the "Highlight Mode" tab in Figure 11. The same cloud will appear, but clicking on a term will cause it to highlight on the page. This is demonstrated by Figure 13.

Figure 13: Tagcloud highlight feature

The screenshot shows the EFO Tag Cloud interface. On the left, there's a sidebar with 'IDF' and 'SDRF' sections. The main area displays a list of terms with columns for Key, Source Name, Material Type, Characteristics, and Description. A tag cloud is overlaid on the right, showing various terms like 'control', 'labelling', 'protocol', 'gene', 'assay', 'submitter', 'submitted', 'irritable bowel syndrome', 'software', 'ulcerative colitis', 'transcription profiling by array', 'RNA', 'protocol parameter', 'Copenhagen', 'protein', 'DNA', 'transcription profiling role', 'data transformation', 'cell type', 'array', 'descending colon', 'Homo sapiens', 'array design', 'organism part', 'whole organism', 'organism', and 'colitis'. The word 'gene' is highlighted in yellow in the tag cloud.

7 EFO Search Box

The EFO search box is a convenience feature implemented into MageComet. If the user wants to confirm that an EFO term exists in the EFO ontology, the EFO search box can be used as shown in Figure 14.

The search field autocompletes based on the query text and displays 3 terms that match the query. If an item is in parentheses as shown with human, the value in parentheses is the standard name for the synonym found. If the user wants to copy the standard ontology name, pressing enter while over the term will fill in the field, making it copyable.

Figure 14: EFO searchbox

The screenshot shows the EFO search box interface. A search field contains the text 'hum'. A dropdown menu is open, showing three suggestions: 'Human - (Homo sapiens)', 'human embryonic kidney cell - (HEK293)', and 'Human Gingival Fibroblasts - (HGF)'. The interface includes buttons for 'Confirm Factor Values', 'Export IDF', 'Export SDRF', and 'Revalidate'. Below the search field, there's a table with columns for 'ng by array', 'transcription profiling by array', and 'transcription profiling by array'. The table has two rows, both with 'EFO' in the second and third columns.

8 Adding Factor Values

When the user has finished extracting the factor values in the SDRF document, the IDF document must be updated to reflect the changes. The IDF document can be updated via the "Confirm Factor Value" button as shown in Figure 15.

Once clicked a window will appear that shows all of the Factor Values in the SDRF. Figure 16. After filling in the corresponding Factor Value Types and clicking save, the IDF document will have the correct values automatically inserted.

Figure 15: Confirm factor value button

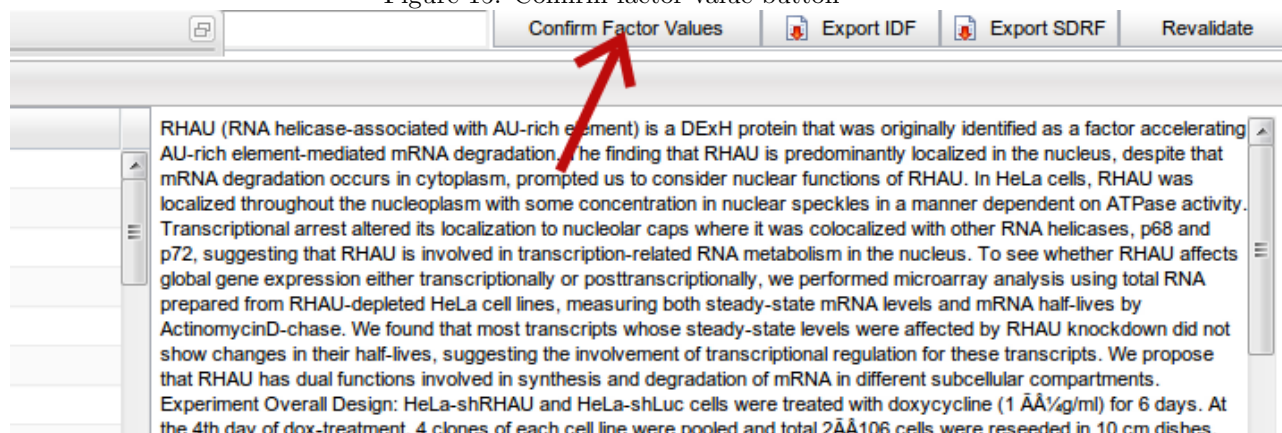
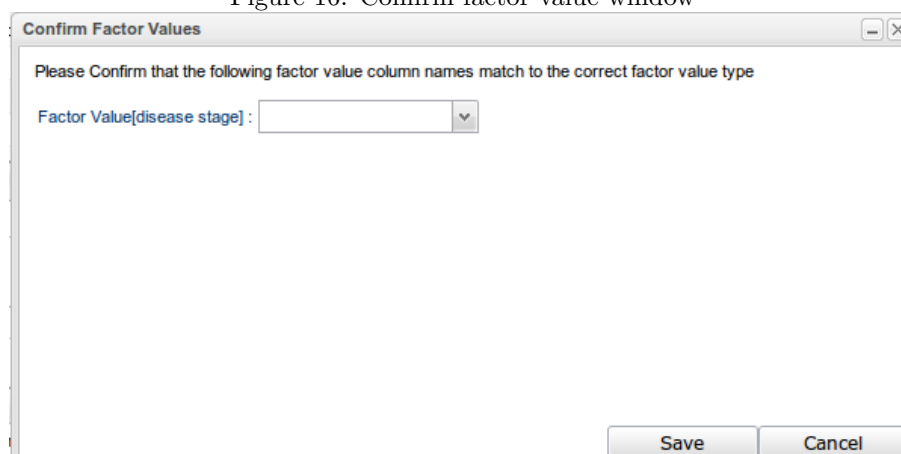


Figure 16: Confirm factor value window



9 Revalidation/Validation

MageComet also has implementation of the Limpopo validators. A user can see all the of the errors by clicking on the Errors tab as shown in Figure 17.

In the errors tab, the user can selectively view errors, warnings, missing information, and revalidate the current MAGE-TAB documents after changes have been made.

Figure 17: Errors button

The screenshot shows the EFO Tag Cloud interface. At the top, there are three buttons: 'Load', 'Edit', and 'Errors'. The 'Errors' button is highlighted with a red arrow. Below the buttons, there is a section for 'IDF' (Investigation Design Factor) and a section for 'SDRF' (Source Design Factor). The 'IDF' section contains a table with fields and values. The 'SDRF' section contains a filter and extract interface.

Field	1	2	3	4	5
Investigation Title	RNAi of human RHAU-				
Comment[Submitted Name]	The DEXH-box RNA he				
Experimental Design	unknown_experiment_ RNAi profiling by array				
Experimental Design Term Source REF	EFO				
Comment[ArrayExpressReleaseDate]	2008-06-16				
Comment[SecondaryAccession]	GSE8192				
Comment[AEMIAMESCORE]	4				
Comment[ArrayExpressAccession]	E-GEOD-8192				
Comment[MAGETAB TimeStamp_Version]	2010-08-08 01:24:49 L				
Experimental Factor Name					
Experimental Factor Type					
Experimental Factor Term Source REF					

Below the IDF section, there is a section for 'SDRF' (Source Design Factor). It includes a 'Filter' button and an 'Extract' button. There is also a 'Match All' button and a 'Filter' button. The 'Filter' button is highlighted with a red arrow.

Figure 18: Errors tab

The screenshot shows the EFO Tag Cloud interface with the 'Errors' tab selected. The 'Errors' tab displays a table of errors. The table has columns for Code, Type, Message, Comment, Line, and Column. There are two error entries.

Code	Type	Message	Comment	Line	Column
1015	validation error	Missing info in IDF: 'the following fields have no data'	IDF date tag Date Of Experiment is missing	-1	-1
24	validation error	"A required IDF tag had a null value"	Error: At least one Email address must be provided in	-1	-1

10 Exporting

The final stage in editing is exporting the changes. To save a file locally, the user can click either "Export" button as designated by Figure 19.

Figure 19: Export buttons

The screenshot shows a software interface with a top toolbar containing buttons: 'Confirm Factor Values', 'Export IDF', 'Export SDRF', and 'Revalidate'. Two red arrows point to the 'Export IDF' and 'Export SDRF' buttons. Below the toolbar is a table with columns 5 and 6. Column 6 contains a long text paragraph about RHAU (RNA helicase-associated with AU-rich element). At the bottom of the interface, there is a section with a label 'nn:', a dropdown menu, a 'Value:' label, and a 'Replace' button. An 'Edit Columns' button is also visible on the right side of the table area.

Confirm Factor Values Export IDF Export SDRF Revalidate

5	6
	<p>RHAU (RNA helicase-associated with AU-rich element) is a DExH protein that was originally identified as a factor accelerating AU-rich element-mediated mRNA degradation. The finding that RHAU is predominantly localized in the nucleus, despite that mRNA degradation occurs in cytoplasm, prompted us to consider nuclear functions of RHAU. In HeLa cells, RHAU was localized throughout the nucleoplasm with some concentration in nuclear speckles in a manner dependent on ATPase activity. Transcriptional arrest altered its localization to nucleolar caps where it was colocalized with other RNA helicases, p68 and p72, suggesting that RHAU is involved in transcription-related RNA metabolism in the nucleus. To see whether RHAU affects global gene expression either transcriptionally or posttranscriptionally, we performed microarray analysis using total RNA prepared from RHAU-depleted HeLa cell lines, measuring both steady-state mRNA levels and mRNA half-lives by ActinomycinD-chase. We found that most transcripts whose steady-state levels were affected by RHAU knockdown did not show changes in their half-lives, suggesting the involvement of transcriptional regulation for these transcripts. We propose that RHAU has dual functions involved in synthesis and degradation of mRNA in different subcellular compartments. Experiment Overall Design: HeLa-shRHAU and HeLa-shLuc cells were treated with doxycycline (1 μg/ml) for 6 days. At the 4th day of dox-treatment, 4 clones of each cell line were pooled and total 2\times10⁶ cells were reseeded in 10 cm dishes. For the starvation experiment; medium was replaced with serum-free DMEM on the 5th day, 24 h before the collection of RNA. The ActD-chase experiment was done on the 6th day of doxycycline treatment. For the mRNA decay experiment; ActD was added to the medium at 5 μg/ml and total RNA was collected at 0, 30, 60, 90, and 120 min after the addition of ActD. Samples for time 0, representing the total amount RNA in the normal condition, and samples from starved cells were analyzed in triplicates, whereas ActD-treated samples for the mRNA decay study were analyzed in duplicates. Total RNA was isolated using the RNeasy kit from QIAGEN (Hombrechtikon, Switzerland). Experiment Overall Design: Total RNA (5 μg) from each replicate was reverse transcribed and labelled using the Affymetrix 1-cycle labelling kit according to manufacturer's instructions. Purified cDNA (20 μg/l) was generated by heating with reverse transcriptase for 45 min at 42 $^{\circ}$C.</p>

nn : Value : Replace

Edit Columns