



Framework for Data Curation

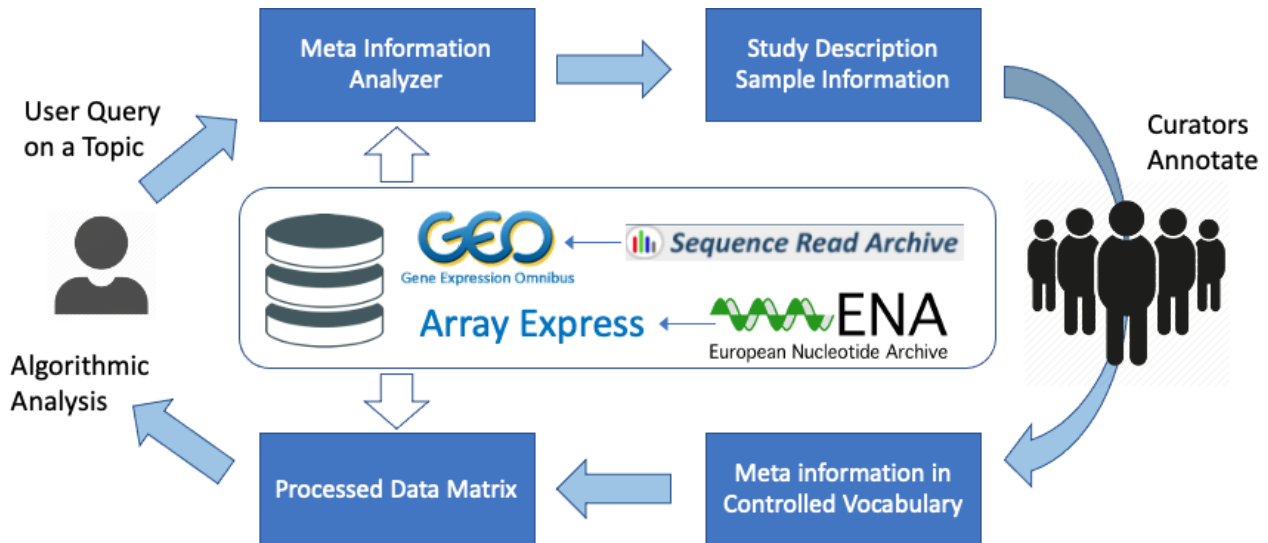
Peng Jiang (peng.jiang@nih.gov)

Table of Content

Introduction	1
User Account	2
Project Management	3
CREATE	4
UPDATE	9
SELECT	9
CURATOR	9
TASK	10
Upload candidate datasets	10
Upload the GEO query download	11
Upload the ArrayExpress query download	12
Upload file by IDs	13
Delete candidate datasets	13
Task assignment to curators	13
Curation Workflow	14
Stage One: Identify relevant datasets	14
Stage Two: Annotate sample conditions	18
Sample tables	18
Assistant function panel	18
Video examples	19
Result management	20
Quality control	20
Important tips on using FDC	22
Reference	23

Introduction

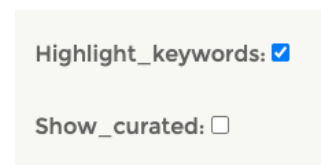
The Framework for Data Curation (FDC) aims to enable researchers to annotate the meta information of datasets in the GEO and ArrayExpress databases to enable automatic algorithmic analysis. Focusing on a research topic, users can input a query result, composing a list of dataset IDs, downloaded from the GEO and ArrayExpress databases. The server will download the meta information of uploaded dataset IDs. Then, curators will annotate the meta-information based on a set of predefined schemes. The annotated sample information will be combined with the processed data matrices from GEO and ArrayExpress databases to enable algorithmic analysis.



Originally, we developed the FDC to create the [CytoSig database](#) for studying human cytokine response (1). Later, we made functions in FDC generally applicable to data curation projects focused on a biological topic. In the following chapters, we will introduce functions in regular fonts with examples from the human cytokine response project in *italic blue*.

User Account

The [new account registration](#) is mostly self-explanatory. The only two fields requiring attention in the user forms are “Highlight_keywords” and “Show_curated”. Selecting the “highlight keywords” checkbox will trigger the keyword highlighting function when annotating datasets (further details in Example 2, Figure 13, Figure 15). We suggest users accept the default checked value. Selecting the “Show_curated” checkbox will enable users to view curated datasets after submitting their annotations, which could be cumbersome to always view lots of datasets in the dataset table (further details in Figure 13, function 6).




Highlight_keywords:

Show_curated:

Figure 1. User Preference

We suggest users accept the default unchecked value unless the curators want to revisit submitted annotations without the help of the project owner, who can always push back a curator submission (Figure 17, function 6).

The user can [update](#) the account profile and annotation preference anytime after the account creation by clicking the user profile icon  on the top right corner. The update form has exactly the same fields as the creation form.

FDC is free to all non-profit users. For commercial users, please read the [software license](#) and apply for a trial account first.

Project Management

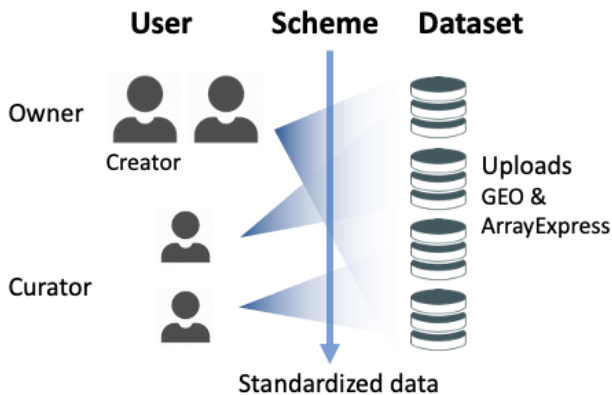


Figure 2. Three core components of a project.

An FDC project has three core components: 1, Users; 2, Annotation Scheme, and 3, Datasets. Users include project owners and curators. The first project owner creates the project and may include more owners to manage the project. Project owners need to upload datasets after querying NCBI GEO and ArrayExpress databases focusing on a biology topic and recruit curators to annotate datasets. Each curator can only access a set of datasets assigned to them by the owner. Owners can also serve as curators and access all uploaded datasets. The annotation scheme includes control vocabulary and rules to make sure that sample annotations follow a standard to enable algorithmic analysis.

The project menu has a few options for project creation and management. When you click the PROJECT, a dropdown menu will show up and all public projects will be presented.

OVERVIEW	CURATION	PROJECT	HELP	CONTACT	PENG_JIANG	LOGOUT
Public projects.		CREATE				
Total 11 rows.		SELECT				
		UPDATE				
ID	Title	Summ.	CURATOR	Fields	Contacts	
0	Human Cytokine Response	This pi in hum	TASK	okine response profiles	Treatment, Condition, Sub Condition, Dose, Duration peng.jiang@nih.gov	
1	Cancer Therapy Response			Group, Type, Sub Type, Therapy, Patient, Age, Gender, Stage	peng.jiang@nih.gov	
3	Human Chemokine Response	This project aims to collect human Chemokine response data.		Treatment, Condition, Sub Condition, Dose, Duration	peng.jiang@nih.gov	
4	Human Signaling, others	Important human signaling molecules, other than cytokine, chemokine, and growth factors.		Treatment, Condition, Sub Condition, Dose, Duration	peng.jiang@nih.gov	
5	Human Growth Factor Response	This project aims to study human growth factor response, excluding cytokines.		Treatment, Condition, Sub Condition, Dose, Duration	peng.jiang@nih.gov	
6	Dexamethasone	Dexamethasone treatment response collection		Treatment, Condition, Sub Condition, Dose, Duration	peng.jiang@nih.gov	
9	test	Test project created by the reviewer.		Treatment, Condition, Sub Condition, Dose, Duration	pengj@alumni.princeton.edu	

Figure 3. Project menu and public projects.

If you are assigned as a curator, but not the owner, you will only see a subset of the menu items.

OVERVIEW	CURATION	PROJECT	HELP	CONTACT
Public projects.		CREATE		
		SELECT		

Figure 4. Project menu for a curator.

CREATE

To create a new project, please click [PROJECT -> CREATE](#) to open the creation form with the following fields. The design of a project should enable curators to annotate standard fields from the sample metadata for downstream algorithmic analysis.

- **ID:** Automatically generated ID for a new project. Cannot edit.
- **Title:** Title of the current project.
- **Description:** Description of your new project. The system administrator will review your new project based on the Title and Description. As long as there is no inappropriate information, we will approve your project.
- **Public:** Checkbox to determine whether or not to show the current project publicly to all users. Please see the example in Figure 3.
- **Fields:** List separated by “,” to enumerate fields to be extracted in the sample metadata table. For example, we listed the following fields for the human cytokine response project to collect experimental conditions across cytokine treatment experiments in cell models:
[Treatment, Model, Sub Condition, Dose, Duration.](#)

Example 1. Fields for annotating cytokine treatment response experiments.

You can also insert new fields or delete existing fields for each individual sample table if you need to collect different fields for a dataset (Figure 16, function 6). For example, different anticancer therapy clinical studies might provide very different clinical information.

- **Keywords:** Keyword patterns in [regular expression](#) forms. Our system will highlight all matched patterns in the dataset and sample tables (Figure 13 and Figure 15). Curators will look at these keywords and determine whether the current dataset is relevant to the study topic. You can use the # symbol to comment on a line. Different patterns can be separated by comma ‘,’ (avoid comma in the regular expression), or a new line. For white space ‘ ’, we will actually match a few more separators, such as “.-”. For example, “IL6”, “IL 6”, “IL.6”, “IL_6”, “IL-6”, will be treated as the same pattern. The platform will only search for isolated keyword patterns starting and ending with either blank space or at the boundary of a text segment. For example, please see our keywords in the human cytokine response project.

```
# Colony stimulating factor
Colony stimulating factor
(G|GM|M) CSF
CSF [1-3]($(?:=[^0-9]))

# Interferon
Type ([1-3]|+) (IFN|Interferon)
(IFN|Interferon) [\alpha\beta\gamma\lambda]
IFN($(?:=[^a-z]))
Interferon

# Interleukin
(IL|Interleukin) [1-9][0-9]?($(?:=[^0-9]))
Interleukin
TSLP
```

LIF(\$\{(?=[^a-z])\}), leukemia inhibitory factor
OSM(\$\{(?=[^a-z])\}), oncostatin M

TNF

Tumor necrosis factor, TNF [α A], *cachectin, TNF*(\$\{(?=[^a-z])\})

lymphotoxin [α A β B]

LT [α A β B]

(*CD27*|*CD30*|*CD40*|*Fas*) (*ligand*|*L*)

4 1BB L

Trail, APO 2 L

OPG L, RANK L

#APRIL

#LIGHT

TWEAK(\$\{(?=[^a-z])\})

BAFF(\$\{(?=[^a-z])\}), *CD257*(\$\{(?=[^0-9])\})

Unassigned

(*TGF*|*Transforming growth factor*[-_]?*[aA β B]*, *Transforming growth factor, TGF*(\$\{(?=[^a-z])\})

Macrophage migration inhibitory factor, MIF(\$\{(?=[^a-z])\})

Example 2. Keywords for highlighting cytokine names

These keywords will trigger highlights in **yellow bold fonts** in the dataset and sample table (Figure 13 and Figure 15). Curators can focus on the sentences containing cytokine names to determine whether the current study contains cytokine treatment data (Example 5, Example 6) or just mentioned cytokine names (Example 7). The curation step will be introduced in further detail later.

- **Keywords_filter:** checkbox to determine whether only include datasets with keywords present in the study description or sample information. The default value is unchecked. In the human cytokine response project, we set it as checked because a cytokine treatment dataset must contain some cytokine names.
- **Processed_filter:** checkbox to determine whether the platform will only show datasets with processed gene expression matrices. If you are only working on human genome-wide transcriptomic studies, please set it as True. Otherwise, please set it as False to include all possible datasets. Currently, FDC only pre-processed genome-wide transcriptomic studies from human samples deposited in NCBI GEO (2), SRA (3), ENA (4), or ArrayExpress (5) databases before February 2020. The metadata of samples in SRA and ENA are included in the GEO and ArrayExpress databases, respectively.
- **Vocabulary:** Controlled vocabulary for the sample table annotation. Ideally, annotated values should follow a defined vocabulary to facilitate automatic downstream analysis. When curators annotate the sample table, the vocabulary will pop up in the dropdown list as hints for selection. However, according to our experience, it is impossible to strictly follow controlled vocabulary in real-world projects. Thus, this function just provides curators an option to use. For example, the controlled vocabularies for the cytokine project are listed below.

Control

GCSF, GMCSF, MCSF

IFNA, IFNB, IFNG, IFNL
IL1A, IL1B, IL1RA, IL2, IL3, IL4, IL5, IL6, IL7, IL9, IL10, IL11, IL12, IL13, IL15, IL16, IL17A, IL17F, IL18, IL19, IL20, IL21, IL22, IL23, IL24, IL25, IL26, IL27, IL28, IL29, IL30, IL31, IL32, IL33, IL34, IL35, IL36A, IL36B, IL36G, IL36RA, IL37, TSLP, LIF, OSM
TNFA, LTA, LTB, CD40L, FASL, CD27L, CD30L, 41BBL, TRAIL, OPGL, APRIL, LIGHT, TWEAK, BAFF
TGFB1, TGFB2, TGFB3, MIF
LPS
PBMC, Monocyte, Macrophage, Fibroblast, T CD4, T CD8, Dendritic, NK, Neutrophil, Lymphocyte

Example 3. Controlled vocabularies for annotating cytokine or cell model names

When annotating the sample table (will introduce in detail later in Figure 15), these vocabularies will pop up in a dropdown list as hints if curators input the first few characters.

ID	title	source name	organism	disease state	cell type	treatment	donor id	description	treatment protocol	Treatment	Condition	Sub Condition	Dose	Duration
GSM1863325	1. PBMC	human blood donor peripheral blood monocyte, untreated	Homo sapiens	healthy	peripheral blood monocytes	none	A	SAM2707860 Processed data file: PBMC_RPKMs.xls	Untreated control or IFN α 1000 U/ml for 6 hours (Sigma).	I IFNA IFNB IFNG IFNL IL1A IL1B IL1RA IL2 IL3 IL4 IL5 IL6 IL7 IL9 IL10	PBMC	A		6 hrs
GSM1863326	2. PBMC	human blood donor peripheral blood monocyte, untreated	Homo sapiens	healthy	peripheral blood monocytes	none	B	SAM2707863 Processed data file: PBMC_RPKMs.xls	Untreated control or IFN α 1000 U/ml for 6 hours (Sigma).		PBMC	B		6 hrs
GSM1863327	3. PBMC	human blood donor peripheral blood monocyte, untreated	Homo sapiens	healthy	peripheral blood monocytes	none	C	SAM2707857 Processed data file: PBMC_RPKMs.xls	Untreated control or IFN α 1000 U/ml for 6 hours (Sigma).		PBMC	C		6 hrs

Figure 5. Dropdown hints of controlled vocabularies after inputting a few characters.

- Vocabulary_map:** Automatic map to translate terms in sample metadata to standard vocabularies in such format: “*Matching pattern : target vocabulary*”. Our platform will match patterns in case-insensitive regular expressions and replace them with the target vocabulary. Similar to the case of keyword, You can use the # symbol to comment on a line. For white space ‘ ’, we will actually match a few more separators in molecule names, such as “_.”. For example, a rule like “(IFN|Interferon) ([γ G])gamma): IFNG” will map names, such as IFN- γ , interferon_gamma, or Interferon G, all to one standard name IFNG. The map table for the human cytokine response project is listed below.

```

None|Vehicle|DMSO|PBS|Untreated|Untreat|unstimulated|Mock|No stimulation : Control

((G|Granulocyte) (CSF|Colony stimulating factor))|(CSF 3) : GCSF
((GM|Granulocyte macrophage) (CSF|Colony stimulating factor))|(CSF 2) : GMCSF
((M|Macrophage) (CSF|Colony stimulating factor))|(CSF 1) : MCSF

(IFN|Interferon) ([ $\alpha$ A])alpha|alfa)[1-9]?: IFNA
(IFN|Interferon) ([ $\beta$ B])beta): IFNB

```

(IFN\Interferon) ([γ G]\gamma): IFNG
(IFN\Interferon) ([λ L]\lambda)[1-9]?: IFNL
(IL\Interleukin) 28[$\alpha\beta$]? : IFNL
(IL\Interleukin) 29 : IFNL

(IL\Interleukin) 1 ([α A]\alpha|alfa) : IL1A
(IL\Interleukin) 1 ([β B]\beta) : IL1B
(IL\Interleukin) 1 R[α A] : IL1RA
(IL\Interleukin) 2 : IL2
(IL\Interleukin) 3 : IL3
(IL\Interleukin) 4 : IL4
(IL\Interleukin) 5 : IL5
(IL\Interleukin) 6 : IL6
(IL\Interleukin) 7 : IL7
(IL\Interleukin) 9 : IL9
(IL\Interleukin) 10 : IL10
(IL\Interleukin) 11 : IL11
(IL\Interleukin) 12 : IL12
(IL\Interleukin) 13 : IL13
(IL\Interleukin) 15 : IL15
(IL\Interleukin) 16 : IL16
(IL\Interleukin) 17 ([α]\alpha|alfa) : IL17A
(IL\Interleukin) 17 F : IL17F
(IL\Interleukin) 18 : IL18
(IL\Interleukin) 19 : IL19
(IL\Interleukin) 20 : IL20
(IL\Interleukin) 21 : IL21
(IL\Interleukin) 22 : IL22
(IL\Interleukin) 23 : IL23
(IL\Interleukin) 24 : IL24
(IL\Interleukin) (25|17E) : IL25
(IL\Interleukin) 26 : IL26
(IL\Interleukin) 27 : IL27
(IL\Interleukin) 30 : IL27
(IL\Interleukin) 31 : IL31
(IL\Interleukin) 32 : IL32
(IL\Interleukin) 33 : IL33
(IL\Interleukin) 34 : IL34
(IL\Interleukin) 35 : IL35
(IL\Interleukin) 36 ([α A]\alpha|alfa): IL36
(IL\Interleukin) 36 ([β B]\beta): IL36
(IL\Interleukin) 36 ([γ G]\gamma): IL36
(IL\Interleukin) 36 R[α A] : IL36RA
(IL\Interleukin) 37 : IL37

Leukemia inhibitory factor : LIF

Oncostatin M : OSM

(TNF|Tumor necrosis factor) ([αA]alpha|alfa) : TNFA

(LT|Lymphotoxin) ([αA]alpha|alfa) : LTA

(LT|Lymphotoxin) ([βB]beta) : LTB

CD27 (L|ligand) : CD27L

CD30 (L|ligand) : CD30L

CD40 (L|ligand) : CD40L

FAS (L|ligand) : FASL

4 1BB (L|ligand) : 41BBL

(APO 2 (L|ligand))|(TNFSF 10) : TRAIL

(RANK (L|ligand))|(TNFSF 11) : OPGL

(TGF|Transforming growth factor) ([αA]alpha|alfa)[1-9]? : TGFA

(TGF|Transforming growth factor) ([βB]beta) 1? : TGFB1

(TGF|Transforming growth factor) ([βB]beta) 2 : TGFB2

(TGF|Transforming growth factor) ([βB]beta) 3 : TGFB3

lipo poly saccharide : LPS

nanogram per milliliter : ng/m

unit per milliliter : u/m

millimolar : mM

(μM|micromolar) : uM

nanomolar : nM

picomolar : pM

(μ|u|micro)mol/L : uM

(n|nano)mol/L : nM

(p|pico)mol/L : pM

peripheral blood mononuclear cell(s?) : PBMC

monocyte(s?) derived macrophage(s?) : Macrophage

monocyte(s?) derived dendritic cell(s?) : Dendritic

Human Umbilical Vein Endothelial Cell(s?) : HUVEC

Example 4. Mapping rules to standardize cytokine names, dose units, and cell model names

In the assist panel of a sample table, the “**Translate Vocabulary**” button will trigger the conversation of values in the Destination column using the mapping rules defined above (introduced in details later in Figure 16, function 11).

After submitting the project, the system administrator will review the new project creation and approve it if there is no inappropriate content in the project description.

UPDATE

A project owner can update a project by clicking [PROJECT->UPDATE](#) to open the update form with exactly the same fields as the creation form. Typically, the project design after initial creation will have many limitations and the project owner needs to optimize the project design, such as keywords and vocabulary maps, iteratively based on feedback from curators. If there are any changes on the project title and description, which will be displayed publicly, the system administrator will review the changes for approval. Otherwise, the project changes will be effective immediately after submitting. Only project owners, but not curators, can update the project.

SELECT

Project owners need to select an approved project to work on. Similarly, curators need to select a project assigned to work on. After submitting the selection, the interface will jump to the curation module introduced later.

ID	Title	CURATOR	Select
0	Human Cytokine Response	Collect cytokine response profiles in humans.	<input checked="" type="radio"/>
1	Cancer Therapy Response		<input type="radio"/>
3	Human Chemokine Response	This project aims to collect human Chemokine response data.	<input type="radio"/>
4	Human Signaling, others	Important human signaling molecules, other than cytokine, chemokine, and growth factors.	<input type="radio"/>
5	Human Growth Factor Response	This project aims to study human growth factor response, excluding cytokines.	<input type="radio"/>
6	Dexamethasone	Dexamethasone treatment response collection	<input type="radio"/>
9	test	Test project created by the reviewer.	<input type="radio"/>

Figure 6. Project selection

CURATOR

The project owner can click the [PROJECT -> CURATOR](#) menu to open the curator control panel. The left panel (Figure 7, Label 1) can add one curator through the user ID. The project owner may check the “As Owner” to give owner-level permissions to the new user (by default, this option is unchecked). The project owner can remove existing curators from the project by unchecking the Select column and click the Modify button (Figure 7, Label 2).

OVERVIEW CURATION **PROJECT** HELP CONTACT PENG_JIANG LOGOUT

Assign curators

Human Cytokine Response project

type curator username **Add** ¹

As Owner

Existing curators

Username	ID	Name	Institute	Owner	Select
peng_jiang	1	Peng Jiang	National Cancer Institute, National Institutes of Health	True	<input type="checkbox"/>
lingrui_liu	2	Lingrui Liu	Yale University	False	<input checked="" type="checkbox"/>
zhangy28@nih.gov	6	Yongliang Zhang	NIDDK	False	<input checked="" type="checkbox"/>
Yuzhang@3	7	YU ZHANG	National Cancer Institute	False	<input checked="" type="checkbox"/>
reviewer	13	Reviewer Reviewer	National Cancer Institute	False	<input checked="" type="checkbox"/>
Apurohit	16	Abhilasha Purohit	National Institutes of Health	False	<input checked="" type="checkbox"/>

Modify ²

Figure 7. Project curators

TASK

Upload candidate datasets

The project owner can click the [PROJECT -> TASK](#) menu to open the task control panel, which allows uploading candidate datasets and assigning tasks to curators. To upload datasets (Figure 8, label 1), the project owners should submit a file containing IDs, together with the file title and type.

OVERVIEW CURATION **PROJECT** HELP CONTACT PENG_JIANG LOGOUT

Upload task assignments

Title:

Task_file: No file chosen ¹

File_type:

add (check) or remove (uncheck) uploaded dataset IDs:

Submit

Title	Add	Creator	Time Upload	File
GEO	True	peng_jiang	March 18, 2021, 1:21 p.m.	gds_result.Cytokine.txt
Examples 1	True	peng_jiang	July 1, 2021, 8:51 a.m.	examples.1.txt

Clear History

Curator assignments

Assign task curators. Leave empty if only visible to project owners.

Username	Name	Institute	Include
lingrui_liu	Lingrui Liu	Yale University	<input type="checkbox"/> ³
zhangy28@nih.gov	Yongliang Zhang	NIDDK	<input type="checkbox"/>
Yuzhang@3	YU ZHANG	National Cancer Institute	<input type="checkbox"/>
reviewer	Reviewer Reviewer	National Cancer Institute	<input type="checkbox"/>
Apurohit	Abhilasha Purohit	National Institutes of Health	<input type="checkbox"/>

²

Figure 8. Project task

Currently, FDC only pre-processed metadata from human samples deposited in NCBI GEO (2) or ArrayExpress (5) before February 2020. The meta information of samples in SRA (3) and ENA (4) are included in the GEO and ArrayExpress, respectively. Thus, FDC only processes human datasets from GEO and ArrayExpress deposited before February 2020 and ignores other datasets in the upload. Also, uploading some large task files may cause network error on Google Chrome browser during the first try. Please refresh the page and upload again, or use Firefox or Safari.

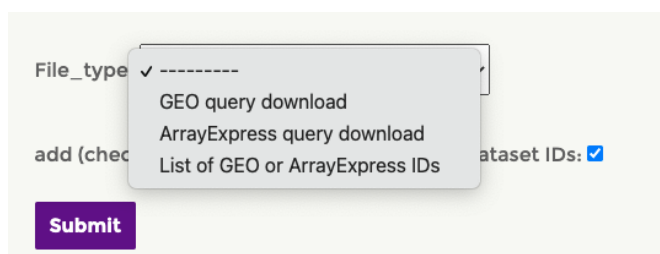


Figure 9. Project task file types. FDC can accept three upload file types: 1, GEO query download; 2, ArrayExpress query download; 3, or List of GEO or ArrayExpress IDs.

Upload the GEO query download

The GEO query download is generated through the [NCBI GEO website](#) based on searching with keywords and condition filters. For example, we input “*interferon OR IFNG OR IFN-G*” as the query on the GEO website (Figure 10, Label 1). Then, we selected humans as the organism by clicking the “Homo sapiens” (Figure 10, Label 2), which will append a query condition in the search box (Figure 10, Label 1). The entry type should be “Series” (Figure 10, Label 3) but NOT the “DataSets”, which represents a small subset of Series whose expression matrix can be pre-processed by NCBI GEO. Since we are only interested in transcriptomic studies in the human cytokine response project, we selected study types, including array or high-throughput sequencing (Figure 10, Label 4).

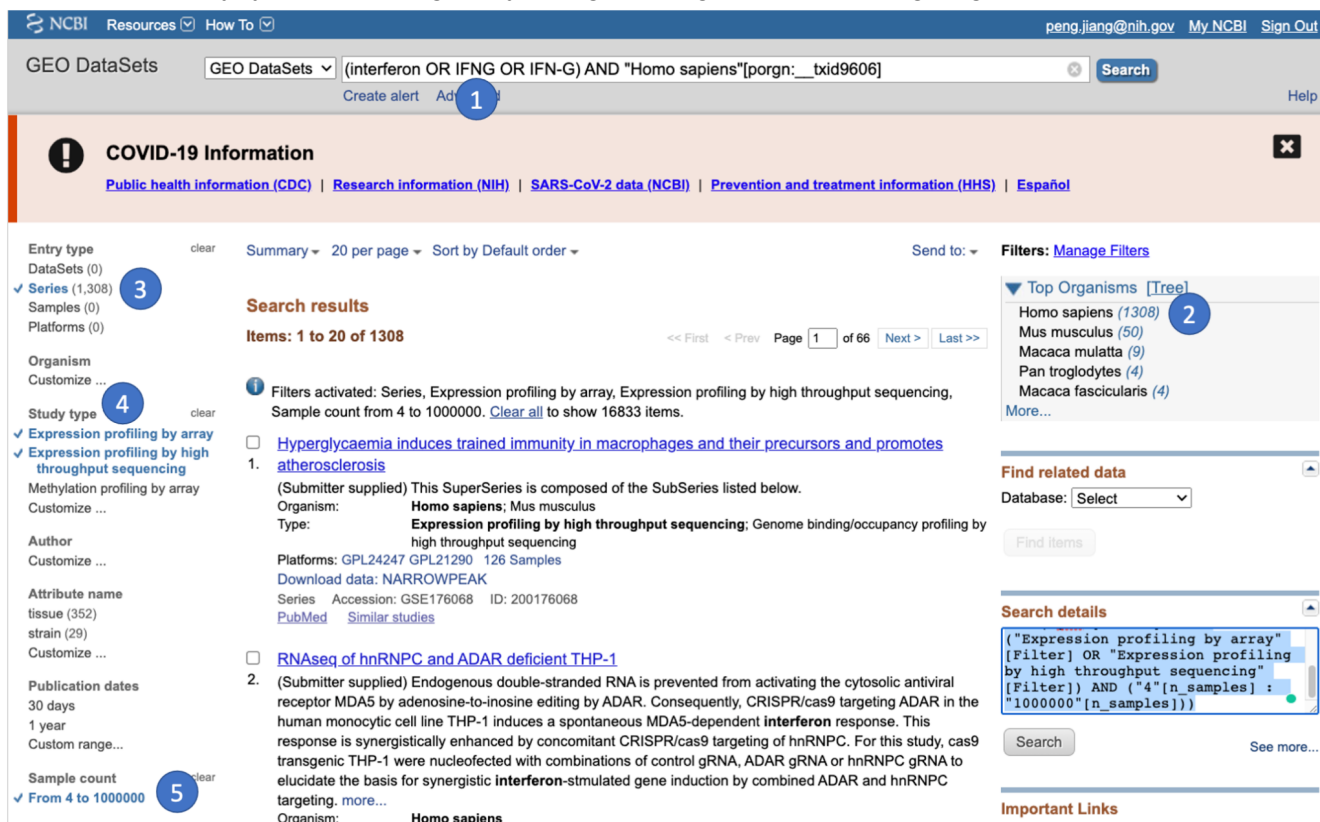


Figure 10. GEO query example.

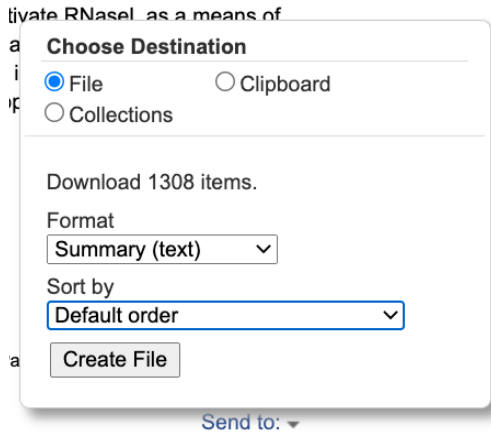


Figure 11. GEO query download. After setting these parameters, please download the GEO query file at the bottom left corner of the page by clicking “Send to” with parameters on the left. Then, you will get a file named “gds_result.txt”, which you can upload to FDC as “GEO query download” (Figure 9).

Upload the ArrayExpress query download

Besides NCBI GEO, [ArrayExpress](#) is the other website to search for candidate datasets. We input the same query keywords shown in Figure 10 (Figure 12, Label 1) and select a few parameters in the “Filter search results” panel (Figure 12, Label 2). The organism should be “Homo sapiens”. The experiment type should be RNA assay to focus on transcriptomic studies. The “ArrayExpress data only” checkbox should be checked to exclude data imported from NCBI GEO by ArrayExpress because the import did not include all GEO datasets. Finally, please click the button “Export table in Tab-delimited format” to download the query result (Figure 12, Label 3), which you can upload to FDC as “ArrayExpress query download” (Figure 9).

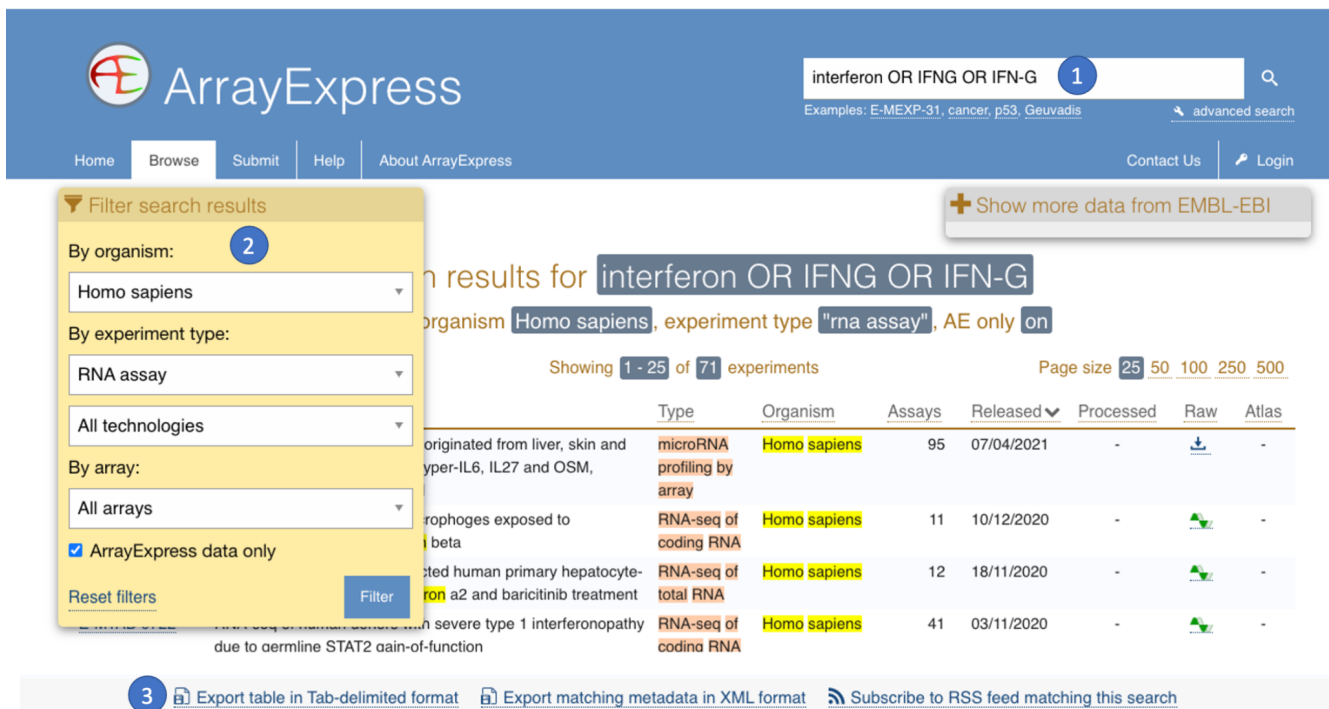


Figure 12. ArrayExpress query download

Upload file by IDs

Besides uploading the query file downloaded from GEO or ArrayExpress, the project owner can also upload a list of dataset IDs separated by newlines (the last option in Figure 9, example on the left).

[GSE72502](#)
[GSE37624](#)
[E-MTAB-9255](#)

Delete candidate datasets

Besides uploading candidate datasets, project owners can also delete previously uploaded datasets by submitting the same file explained above but selecting the checkbox “add (check) or remove (uncheck) uploaded dataset IDs” in Figure 8. Datasets that are already annotated by a curator will not be removed in this function.

Task assignment to curators

The “curator assignments” panel (Figure 8, Label 3) can control curators who can access the uploaded datasets for annotation. First, project owners can access all candidate datasets. By default, curators cannot access any uploaded datasets. If project owners want a particular set of curators to work on a set of candidate datasets, please check the “Include” column on the right panel before submitting the file on the left panel.

Curation Workflow

After creating the project, the project owner should add curators or serve as curators by themselves (Figure 7). After registering user accounts, curators should ask project owners to add them into a project. Curators should use Google Chrome for the best annotation functions. Firefox and Safari also work, although the regular expression functions will be compromised. This platform **does NOT work on Microsoft IE**.

The curation procedure has two stages: 1, identification of relevant datasets; and 2, the annotation of sample information. In stage one, curators should identify datasets related to a specific study topic by reading the title, summary, description, and sample tables. In step two, for each relevant dataset, the curators should annotate the sample metadata table and create standardized columns. Before any curation work, please select a project to work on through the menu **PROJECT -> SELECT**.

Stage One: Identify relevant datasets

In the dataset selection table, you will determine which datasets are relevant to the project topic. Each row represents a candidate dataset. Please review their study design and sample annotation tables (if necessary) to determine the Yes or No status of selection.

The screenshot shows a web interface for dataset curation. At the top, there are navigation tabs: OVERVIEW, CURATION (selected), PROJECT, HELP, CONTACT, and a user profile for PENG_JIANG with a LOGOUT button. The main heading is 'Candidate datasets of project 0 : Human Cytokine Response'. Below this, a progress bar labeled 'Parse table' is shown with a green bar and a '1' callout. The text 'Total 2516 rows.' is followed by a '2' callout. The table has columns: ID, Title, Summary, Design, Count, Status, and Comment. The first row (GSE64282) has a '5' callout in the Count column and a '6' callout in the Status column. A blue arrow points from a '4' callout in the Summary column to the text 'G-CSF, IL-3 and GM-CSF' in the Design column. The second row (GSE39017) has a '16' callout in the Count column. The third row (E-MEXP-2878) has a '4' callout in the Count column. At the bottom, there are buttons for 'Export to CSV' (9) and 'Submit' (10). A '7' callout is at the bottom right of the table area.

Figure 13. Dataset table.

1. Progress bar for parsing data. If the current project contains many candidate datasets, parsing their study design information to highlight keywords (function 4) will take some time. Please wait until the process finishes and the progress bar disappears. If this step takes too long, you can uncheck the “Highlight_keywords” flag (Figure 1) in your user profile to disable this function.
2. Table header with multiple functions. Using the mouse cursor, you can click each column to sort in ascending or descending order and resize the width by pulling on the column boundary. You can also click and hold on each column and drag it to a different position.

- Table row selector. Please move your mouse a little beneath the header to trigger the row selector. If you only want to work on a subset of rows, you can input keywords in the box beneath each column. Advanced users can input regular expressions by using “/pattern/” or “/pattern/i” for case insensitive queries. The result submission (function 9) will only include selected rows.

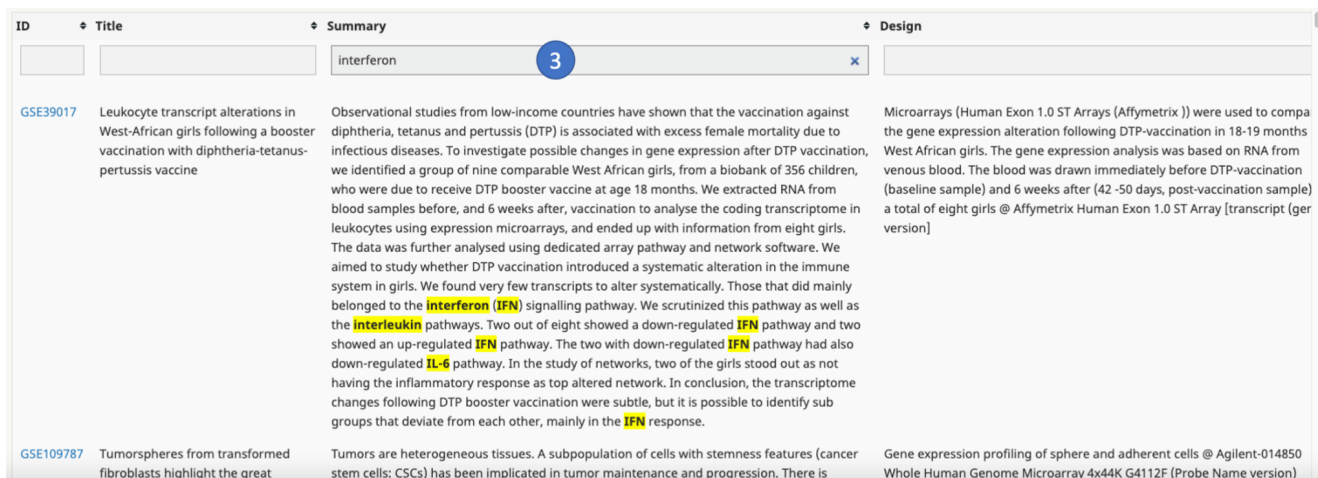
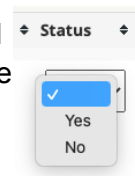


Figure 14. Dataset table row selector

- Keyword highlighting. If you have input keywords in the project setup (Example 2), the webpage will display all matches. This step will trigger the parsing progress bar (function 1).
- Sample count with link to the sample table. The number of samples in each dataset. You can click the number and open the sample table (Figure 15). For many cases, you have to look at the sample table together with the study design to determine the relevance of a dataset. If a study is relevant, you need to curate the sample annotation table in Stage two.
- Dataset status. Annotate the relevance of a dataset as Yes, No, or Blank (not discussed). Once you are ready (even without annotating all datasets), please scroll down the whole page to the bottom and click the “Submit” button (function 10). In the next round of annotation, a dataset with blank annotation will be listed again; and curated datasets will not show up unless you check the “Show_curated” preference (Figure 1).
- Scrollbars for the dataset table vertically and horizontally.
- Comment box. This field might be hidden initially due to the large size of the dataset table. Please use either the horizontal scrollbar or mouse to slide the table to the left to reveal this column. Curators could write some comments for each dataset and submit (function 10) it together with your status annotation (function 6).
- Export the current table to a CSV file.
- Submit your annotation. If successful, you should see a confirmation window and click OK. You don’t have to determine the relevance of all datasets (function 6). A dataset with blank annotation will show up when you open the dataset table again. If the row selector (function 3) is involved, only rows passed the keyword filter will be submitted.



We will show a few examples from the human cytokine response project.

ID	Title	Summary	Design	Count	Status
GSE37624	Transcriptional response of confluent human umbilical vein endothelial cells to stimulation with the related cytokines interleukin-1b or interleukin-33 .	The similar response of endothelial cells to exogenous IL-33 or IL-1β prompted us to compare the genome-wide transcription profile of confluent human umbilical vein endothelial cell (HUVEC) cultures after 4 hours exposure to IL-33 or IL-1β . Analysis of these data revealed a striking similarity in the transcriptional response to the two cytokines.	Pooled HUVEC from 10 donors were seeded 10.000 cells/cm2 and cultured for 4 days before stimulation for 4 hours with IL-1β 0.5 ng/ml or IL-33 50 ng/ml. Total RNA was isolated using the RNeasy Mini Kit (Qiagen) according to instructions of the manufacturer. Microarrays were performed using an Illumina Human6 v2 Expression Beadchip, and the raw data was pre-processed by Illumina's BeadStudio software V2. The expression service was provided by Helse Sør-Øst/University of Oslo Microarray Core Facility, a member of the Norwegian Microarray Consortium supported by the functional genomics program (FUGE) at the Research Council of Norway. @ Illumina human-6 v2.0 expression beadchip	6	Yes

Example 5. GSE37624 as a relevant dataset. *From keyword highlights in the title, summary, and design, we can clearly see that this dataset is about IL1B and IL33 treatment.*

ID	Title	Summary	Design	Count	Status
GSE69602	Ribosome profiling analysis of Dengue Virus	Dengue virus is an + strand RNA virus. We have carried our infections of human cells with Dengue and analyzed the translation, replication, and localization of the Dengue RNA. This allowed for clear definition of the life cycle of the Dengue virus inside a host cell. We also assessed the host response to Dengue virus, finding that a large fraction of the translational response is due to Interferon function.	Translational and transcriptional analysis of the cellular response to Dengue virus infection @ Illumina HiSeq 2500 (Homo sapiens)	116	Yes

Example 6. GSE69602 as a relevant dataset. *From the summary, we see a highlight of Interferon, which cannot confirm this dataset as interferon treatment. Therefore, we opened the sample table below by clicking the Count number "116".*

ID	title	source name	organism	cell line	protocol	compartment	analysis	replicate	strain	description
	seq (Rep 1)									dengueExpression-fractionated.xlsx
GSM1704573	Uninfected cytosol mRNA-seq (Rep 2)	Huh7 cells	Homo sapiens	Huh7 cells	Uninfected	cytosol	mRNA-seq	Rep 2	none	processed data file: dengueExpression-fractionated.xlsx
GSM1704574	Uninfected ER mRNA-seq (Rep 2)	Huh7 cells	Homo sapiens	Huh7 cells	Uninfected	ER	mRNA-seq	Rep 2	none	processed data file: dengueExpression-fractionated.xlsx
GSM1704559	Interferon treatment cytosol mRNA-seq (Rep 1)	Huh7 cells	Homo sapiens	Huh7 cells	Interferon treatment	cytosol	mRNA-seq	Rep 1	DENV2 (M29095.1)	processed data file: dengueExpression-fractionated.xlsx
GSM1704560	Interferon treatment ER mRNA-seq (Rep 1)	Huh7 cells	Homo sapiens	Huh7 cells	Interferon treatment	ER	mRNA-seq	Rep 1	DENV2 (M29095.1)	processed data file: dengueExpression-fractionated.xlsx

From the title and protocol columns, we can see that this dataset profiled interferon treatment, although the type is unclear. Later, when we annotated the interferon type in the sample table, we opened the original publication and found that the type is IFNB.

ID	Title	Summary	Design	Count	Status
GSE118951				9	No
GSE118951	A novel CD4+ T cell population expanded in SLE blood provides B cell help through IL10 and succinate	A better understanding of the mechanisms involved in human plasma cell differentiation will accelerate therapeutic target identification in autoantibody-mediated diseases such as Systemic Lupus Erythematosus (SLE). Here, we describe a novel CXCR5- CXCR3+ PD1hi CD4+ T cell 'helper' population distinct from follicular helper T cells (Tfh) and expanded in blood and inflamed kidneys of SLE patients. Upon activation, these cells express IFN γ and high levels of IL10. Additionally, they accumulate high amounts of mitochondrial ROS (mtROS) as the result of reverse electron transport (RET) fueled by succinate. These cells provide potent help to B cells through the synergistic effect of IL10 and succinate. Cells with similar phenotype and function are generated in vitro upon priming naive CD4+ T cells with oxidized mitochondrial DNA (Ox mtDNA)-activated plasmacytoid dendritic cells (pDCs) in a PD1-dependent manner. Our results provide a novel mechanism for the initiation and/or perpetuation of extrafollicular humoral responses in SLE.	9 total samples; 3 groups of 3 biological replicates: control group Th0, co-culture group CpGA-pDC, and co-culture group Ox mtDNA-pDC @ Illumina NextSeq 500 (Homo sapiens)		No

Example 7. GSE118951 as a non-relevant dataset, dataset table entry. *Although the title and summary mentioned a few cytokines, these descriptions do not indicate that the current dataset is about cytokine treatment. We also clicked the Count number “9” to open the sample table below, and did not see any evidence of cytokine treatment.*

ID	title	source name	organism	cell source	cell type	conditions	replicate	description
GSM3351822	CD4+Tcells_Th0_I	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	Th0	I	cDNA3271_0_I
GSM3351823	CD4+Tcells_Th0_II	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	Th0	II	cDNA3272_0_II
GSM3351824	CD4+Tcells_Th0_III	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	Th0	III	cDNA3273_0_III
GSM3351825	CD4+Tcells_CpGA-pDC_I	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	CpGA-pDC	I	cDNA3274_A_I
GSM3351826	CD4+Tcells_CpGA-pDC_II	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	CpGA-pDC	II	cDNA3275_A_II
GSM3351827	CD4+Tcells_CpGA-pDC_III	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	CpGA-pDC	III	cDNA3276_A_III
GSM3351828	CD4+Tcells_Ox mtDNA-pDC_I	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	Ox mtDNA-pDC	I	cDNA3277_N_I
GSM3351829	CD4+Tcells_Ox mtDNA-pDC_II	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	Ox mtDNA-pDC	II	cDNA3278_N_II
GSM3351830	CD4+Tcells_Ox mtDNA-pDC_III	CD4+Tcells	Homo sapiens	in vitro	CD4+Tcells	Ox mtDNA-pDC	III	cDNA3279_N_III

Stage Two: Annotate sample conditions

Once the curator decides a dataset to be relevant to the topic of interest, the curator needs to analyze the sample table and extract destination columns in standardized vocabulary to enable algorithmic analysis. To open the sample table, please click the count numbers in the dataset table for each dataset entry (Figure 13, function 5).

Sample tables

The sample tables have two parts, including source columns (Figure 15, region 1) and destination columns (Figure 15, region 2). The source columns include metadata extracted from NCBI GEO or ArrayExpress. The destination columns include input boxes for value extraction over user-defined fields (Example 1). Curators can either manually input values or utilize the assist functions introduced below. Many features in the sample table are the same as those in the dataset table (Figure 13) in Stage one, such as the column filters, keyword highlighting with progress bars, scroll bars, table headers, and result submission.

ID	title	source name	organism	protocol	cell type	description	Treatment	Condition	Sub Condition	Dose	Duration
GSM923359	HUVEC_IL-1_4hours_rep1	Confluent HUVEC, stimulated with IL-1b 0.5ng/ml for 4 hours	Homo sapiens	stimulated with IL-1b 0.5ng/ml for 4 hours	HUVEC pool of 10 donors	HUVEC pool of 10 donors stimulated with IL-1b 0.5ng/ml for 4 hours	IL1B	HUVEC		0.5ng/ml	4h
GSM923360	HUVEC_IL-33_4hours_rep1	Confluent HUVEC, stimulated with IL-33 50ng/ml for 4 hours	Homo sapiens	stimulated with IL-33 50ng/ml for 4 hours	HUVEC pool of 10 donors	HUVEC pool of 10 donors stimulated with IL-33 50ng/ml for 4 hours	IL33	HUVEC		50ng/ml	4h
GSM923361	HUVEC_medium_rep2	Confluent HUVEC, given medium for 4 hours (control)	Homo sapiens	control	HUVEC pool of 10 donors	HUVEC pool of 10 donors control_replicate2	Control	HUVEC			4h
GSM923362	HUVEC_medium_rep1	Confluent HUVEC, given medium for 4 hours (control)	Homo sapiens	control	HUVEC pool of 10 donors	HUVEC pool of 10 donors control_replicate1	Control	HUVEC			4h

Figure 15. Sample table

Assistant function panel

In this panel, curators can select columns and apply a set of automatic transformations.

1 History « Undo Redo »

2 Column Source Destination

3a Copy

4 Clear

5 Delete Source

6 Delete Destination

6 Add new column name

Paste From Clipboard 3b

7 Join Front Join Back

8 Append Front e.g., hour Append Back

9 Split Front e.g., rep Split Back

10 Replace From To

11 Translate Vocabulary

Figure 16. Sample table

1. Undo or Redo actions from the assist function panel. **This function might be slow for a large table**; thus, please either wait for a while or avoid using this function.
2. Column selection. All assist functions are targeting a Destination Column. Some functions, including Copy (function 3a) and Join (function 8), will involve a Source Column.
3. Copy content to the Destination column.
 - a. From the Source column.
 - b. From the Clipboard, triggered by Ctrl + V or Command + V (Mac OS). You can either paste a whole column copied from an Excel table or a text content to all cells in the destination column. The checkbox will be unselected automatically after pasting to prevent unexpected further pasting.
4. Clear all content in the destination column.
5. Delete the source or destination column. Sometimes, a source column may be irrelevant to the annotation and contain repetitive contents at every row. Or the curator may want to remove a destination column. Please select the column (function 2) and delete it with this function.
6. Add a new destination column.
7. Join the source column to the front or back of the destination column with the separator in the text box.
8. Append the word in the text box to the front or back of every cell in the destination column.
9. Split a substring from the front or back of the destination column. In each cell, the content will be split by the text in the input box. Advanced users can input regular expressions by using `"/pattern/"` or `"/pattern/i"` for case insensitive matches.
10. Replace substrings (text input box "From") in the destination column to a target string (text input box "To"). Advanced users can input regular expressions in the "From" field by using `"/pattern/"` or `"/pattern/i"` for case insensitive matches.
11. Translate contents to standardized vocabularies in the destination column. In the project setting, the user may define a vocabulary map from non-standard texts to their standard names in controlled vocabularies (Example 4). Be cautious that the automatic translation might misinterpret the content, and human proofreading is always necessary.

Please note that if a row selector is applied on the sample table (similar to Figure 14 on the dataset table), the assist function transformations will only be effective on selected rows.

We will show a few examples of applying these assist functions in sample table annotation from the human cytokine response project.

Video examples

Example 8. [GSE37624 dataset: Youtube URL or Download Video. Try it by yourself here](#)

Example 9. [GSE77808 dataset: Youtube URL or Download Video. Try it by yourself here](#)

Result management

After curators submit their annotations, the project owner can proofread results by clicking **CURATION -> RESULTS**. In this panel, the project owner can either validate that the result is correct or return problematic results to curators to revisit. Most functions in the result table, such as headers with sorter and row selectors, are the same as the dataset table (Figure 13). We will only focus on unique functions numbered below.

Quality control

ID	Title	Count	Curator	Category	Processed Data	Time	Comment	Validated	Revisit
GSE17301	The effect of IFN α on human CD8 T cells_with other concomitant signals	15	Yuzhang@3	Yes	GSE17301.MicroArray.HG-U133A_2.processed.gz	March 23, 2020, 9:43 a.m.	What if there are two kinds of IFN α ?	<input type="checkbox"/>	<input type="checkbox"/>
GSE23935	Gene responses to TGF-beta receptor inhibition in glioblastoma	22	Yuzhang@3	No	GSE23935.MicroArray.HG-U133_Plus_2.processed.gz	March 22, 2020, 9:55 p.m.	TGFB receptor inhibitor will be considered?	<input type="checkbox"/>	<input type="checkbox"/>
GSE105094	RNA sequencing of HCT116 and HKE3 colorectal cancer cell lines before and after stimulation with TGF-alpha	36	Yuzhang@3	No	GSE105094.RNASeq.SRP120173_GRCh38.processed.gz	March 17, 2020, 2:29 p.m.	TGFA is not our interest?	<input type="checkbox"/>	<input type="checkbox"/>
GSE107295	Bone density loci identified by genome-wide association studies segregate a lineage-specific PU.1-dependent gene regulatory network in osteoclasts [HsMmMicroarray]	8	Yuzhang@3	No	GSE107295.MicroArray.HG-U133_Plus_2.processed.gz , GSE107295.MicroArray.Mouse430_2.processed.gz	March 17, 2020, 4:40 p.m.	RANKL is highlighted, but it is not in the table list	<input type="checkbox"/>	<input type="checkbox"/>
GSE12124	In vitro response of fibroblasts isolated from patients with immunodeficiencies	72	Yuzhang@3	No	GSE12124.MicroArray.GPL6106.processed.gz	March 21, 2020, 1:30 p.m.	Patients with specific genes deficiency will be considered?	<input type="checkbox"/>	<input type="checkbox"/>

Export to CSV Submit Download

Figure 17. Result table

1. Count number with link to the curator's annotation. Clicking the number in the Count column will open the sample annotation submitted by each curator. If the curator only submitted the relevance of each dataset (Figure 13, function 6) but not the sample annotation (Figure 15), the Count number will not have the link.
2. Dataset relevance annotation by the curator. This column shows the dataset relevance annotation in the dataset table (Figure 13, function 6). If sample annotation (Figure 15) is also submitted, clicking the status will trigger the download of sample annotation in a text file.
3. Processed Data by FDC. Gene expression matrices processed by FDC from the GEO, ArrayExpress, SRA, and ENA databases. Please note that FDC cannot automatically process many gene expression data due to reasons, such as decayed files, unknown platforms, etc. Therefore, if necessary and also for non-transcriptomic studies, users can go to each database and process datasets that FDC cannot extract automatically.
4. Comment from the curators.
5. Validated status. If the project owner agrees with the curator annotation, please check this box and submit (function 7).
6. Revisit. If the project owner finds any problems in the annotation, please check this box and submit (function 7). The curator will see this dataset as unannotated and revisit it again.

7. Submit. Clicking this button will submit the “Validated” or “Revisit” decision from the project owner.
8. Download. After clicking the download button, the FDC will send a text file with URLs of annotated sample metadata and transcriptomic matrices (if FDC can automatically process the data). For example, the following excerpt is from the human cytokine response project. Some datasets, such as GSE58613, may have metadata annotated by multiple curators. Thus, the project owner needs to decide the priority.

```
https://curate.ccr.cancer.gov/download/Curation/0/1/GSE58613.meta/  
https://curate.ccr.cancer.gov/download/Curation/0/7/GSE58613.meta/  
https://curate.ccr.cancer.gov/download/Curation/0/7/GSE89970.meta/  
https://curate.ccr.cancer.gov/download/GEO/Data/GSE58613/GSE58613.MicroArray.HG-U133A_2.  
processed.gz/  
https://curate.ccr.cancer.gov/download/GEO/Data/GSE89970/GSE89970.RNASeq.SRP093727_G  
RCh38.processed.gz/
```

Example 10. Excerpt of download file from the human cytokine response project.

After getting the result file with the URL list, you can download all files with "wget" or python "urllib", and develop an automatic analysis script. Here is [an example python program](#) to download files and process data into differential expression profiles upon cytokine treatment in the human cytokine response project.

Important tips on using FDC

Data collection projects, making knowledge rediscovery from public datasets, are cost-effective strategies for biological research in the big-data era. The framework for data curation (FDC) aims to help researchers to accomplish their customized data integration projects. Through our previous experience, we recommend a few tips for successfully organizing such a project.

1. **Annotation scheme design.** For a data curation project, any human intervention could be time consuming. Therefore, project owners should make sure that the dataset relevance determination (Figure 13) and sample table annotation (Figure 15) are the only two steps requiring human curation. If any downstream analysis cannot be automated with a program and requires additional manual work, the project owner should redesign the annotation scheme, including the annotation field (Example 1), controlled vocabulary (Example 3), and vocabulary maps (Example 4).
2. **Curator training.** Although we try to write the FDC tutorial as comprehensive as possible, a customized training manual for each project is still very helpful because our tutorial is for general purpose and each individual project may have its unique requirements. For the human cytokine response project, we wrote a [training manual for curators](#), which could serve as an example to organize similar curator training. The project owner should also prepare a few exercises, in which the sample annotation tasks could cover most representative scenarios. For example, here are a few datasets that we provided to curators for exercise.

GSE72502	GSE78193
GSE37624	GSE69602
GSE77808	E-MTAB-6300
GSE73313	GSE18686
GSE58613	GSE2770

Example 11. Exercise datasets from the human cytokine response project.

3. **Use regular expressions.** To automate annotations as much as possible, advanced users might find the convenience of using [regular expressions](#), utilized by FDC in many components. The learning of regular expressions for people without programming backgrounds might be challenging initially, but the gain will be significant eventually. There are many [tutorials](#) for beginners. Please install the [Atom editor](#) for exercise.
4. **Selection of curators.** Eventually, the most important factor of a successful curation project is that the curators should care about the quality of annotations. Typically among many curators, one or two of them will finish the majority of high-quality annotations. They are either the ones who will use these annotations later or people who really care about their performance.
5. **Iterative cycles.** Typically, the annotation scheme is not perfect at the very beginning. Also, curators will not be experienced initially. Therefore, the project owner and curators should give each other feedback iteratively while the project is moving forward. The project owner should optimize the project design and annotation scheme iteratively until minimal effort is required from the curators with assist functions.

Reference

1. Jiang P, Zhang Y, Ru B, Yang Y, Vu T, Rohit P, et al. Systematic Investigation of Cytokine Signaling Activity at the Tissue and Single-Cell Level. *Nature Methods*.
2. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41:D991–5.
3. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive [Internet]. *Nucleic Acids Research*. 2011. page D19–21. Available from: <http://dx.doi.org/10.1093/nar/gkq1019>
4. Amid C, Alako BTF, Balavenkataraman Kadhivelu V, Burdett T, Burgin J, Fan J, et al. The European Nucleotide Archive in 2019. *Nucleic Acids Res*. 2020;48:D70–6.
5. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*. 2007;35:D747–50.