Curse manual

I. Introduction

The Curse suite enables quick and easy construction of expression atlases and co-expression networks from publicly available RNA-Seq experiments. It consists of a web tool named Curse (Curator of Sequencing Experiments) and a portable pipeline named Prose (Processor of RNA-Seq experiments). With Curse, users can browse studies hosted on the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) in order to select relevant RNA-Seq experiments for their compendium. A Prose package can then be downloaded to retrieve and process the raw sequencing data against user-provided transcript sequences. The end result consists of transcript- and gene-level expression atlases for each experiment in the compendium, with the option to construct co-expression networks as well. Curse and Prose are available at: http://bioinformatics.psb.ugent.be/webtools/Curse.

This manual describes the usage and inner workings of Curse. At its core, Curse is an alternative interface to the SRA database. It provides features that enable users to easily and quickly construct compendia of public sequencing experiments for later processing. Curse was created with a focus on RNA-Seq experiments but it can be used for all other types of experiments as well. Only RNA-Seq experiments can be processed with Prose.

Metadata structure

Curse follows the same concepts from SRA to structure data and metadata into studies, experiments, samples, and runs. A study can contain one or more experiments that are all part of a single research effort. The scope of such a study can vary greatly. It can for example consist of merely a control and a treated cell type or it can contain sequencing data from hundreds of different cultivars of multiple plant species. An experiment describes the sequencing of a single sample or a pool of samples. A sample is the biological material taken from a single research subject or source (whole organism, organ, tissue, cell culture, ocean water, ...). A sample pool indicates that multiple samples were physically mixed together prior to sequencing. A run describes the actual sequencing data and contains the output of a single ‘run’ from the sequencing machinery. An experiment can contain multiple runs. Each study, experiment, sample, and run receive their own permanent accession number on SRA.

Replicates and the curation workflow

Replication is a common practice in experimental research including RNA-Seq studies. It is for example required to model biological variation when performing Differential Expression analyses. Two types of replicates exist: Technical replicates are separate experiments that sequenced the same sample or sample pool in order to capture the technical variation of the experimental protocol. Biological replicates are experiments on separate research subjects placed under identical conditions in order to capture biological variation.
No metadata fields exist on SRA for researchers to specify which experiments are replicates or what type of replicates they are. As a consequence, this information is either cryptic, inconsistent, or missing. Your task as a Curse user is to identify the replicates of conditions of interest and place them in replicate groups. Curse offers several features that allow you to easily spot these replicates and the properties that set them apart from other experiments in the study. Assigning experiments to replicate groups automatically adds them to your compendium and forms the basis of the curation workflow. Experiments that are not of interest to you should be added to the blacklist instead.

In Curse, you can give names and descriptions to replicate groups, which will be used to label the experiments in the compendium. The only purpose of replicate groups is the organization and naming of experiments. If you are merely interested in selecting relevant experiments and not in replication, you can simply place each experiment in a separate group. This will not affect the downstream processing of RNA-Seq data with Prose. Curation results are not shared between users or projects.

Something to look out for while curating is that the difference between biological and technical replicates is sometimes vague and often depends on the context of the study. Metadata on SRA regularly contains mistakes against these concepts. For example, experiments on SRA that are technical replicates should be easy to recognize because they should have identical sample accessions. In practice however, the same sample will often be uploaded several times with separate accessions. Sample pools are also often uploaded as if they were single samples and multiple of these sample pools could be mistaken for biological replicates. One could however argue that mixing material from multiple sources eliminates the biological variation that these replicates are supposed to capture.

**Color codes**

In Curse, experiments and studies are consistently placed in three categories, depending on their curation status. Each category is annotated with a color code:

- **Grey** indicates experiments and studies that are pending curation. These have matched your query but have not yet been assigned to a replicate group or to the blacklist.
- **Blue** indicates experiments that were placed in a replicate group. For studies, blue indicates that at least one of its experiments was placed in a replicate group and that any remaining experiments were placed on the blacklist.
- **Black** indicates experiments that were placed on the blacklist. Studies are colored black if all their experiments were placed on the blacklist.

II. Getting started

On the home page of Curse, you are required to create an account to get started, which is necessary to store your compendia and to make sure only you can access them. A username and password is required to create an account. Providing an E-mail address is optional and it will only be used to recover lost passwords.

After logging in, you will be provided with a list of your projects. Each project represents a single compendium. New projects can be created by clicking the ‘NEW PROJECT’ button. Existing projects can be renamed by clicking the name or deleted by clicking the red ‘x’. Clicking a project opens the project page.
### III. Project page

The project page consists of a ‘Query’ panel on top and a ‘Compendium’ panel on the bottom. The Query panel can be used to specify the type of experiments you are looking for. The Compendium panel shows the experiments that are currently in your compendium and provides buttons to export the associated metadata or to download a Prose package to process the RNA-Seq data. To go back to the projects list, click the back button in your browser or in the bottom left of the page.

#### The Query Panel

The query panel consists of text boxes and selection lists to search for experiments on SRA. On the left, users can specify the name of the organism they are interested in. This can be the common name or the scientific name. Anything that works on SRA will work here as well. The ‘Keywords’ text box can be used to search for experiments with certain words in their metadata or to search for specific items by using their accession numbers. The selection lists can be used to select the type of experiment you are looking for. For RNA-Seq experiments, it is recommended to select ‘RNA-Seq’ as Strategy and ‘Transcriptome Analysis’ as Study type. The most popular Platform for RNA-Seq experiments is ‘Illumina’. Keep in mind that when researchers forget to fill in these fields while uploading their data, they will default to ‘Other’.

On SRA, queries are represented by a string. When the user constructs a query on Curse, the string that will be sent to SRA is shown below the selection lists. Clicking this string will run the query directly on SRA in a new tab. To run the query through Curse, click the ‘RUN QUERY’ button to the right. Curse will then make use of the [SRA E-utils](https://www.ncbi.nlm.nih.gov/sra/docs/e-utils/) to search the database. To the right of the RUN QUERY button, the ‘reset’ button can be clicked to reset your query to the last one that finished successfully. After a query is completed, this button changes to ‘refresh’ and clicking it will run the same query again. This can be used to include any experiments that were uploaded to SRA since the last time your query was executed. A successful query and its results will be saved automatically by Curse. If you return to the project at a later time, you can pick up where you left off.

When querying, Curse will fetch summaries for all matching experiments from SRA. Depending on the number of matching experiments, this may take some time. If you are querying a popular organism, it is best to use keywords to make your search as precise as possible to limit the number of hits. If more than 50,000 experiments match your query, you will be asked to fine-tune it further.
After the query is completed, the ‘RUN QUERY’ button will be replaced by ‘CURATE’. Clicking the ‘CURATE’ button will open up the curation page (see below). The colored bars above the buttons indicate the number of studies and experiments that are in the compendium (blue), are pending curation (grey), or are blacklisted (excluded from the compendium, black) respectively. Hovering the mouse over the bars will show the number of experiments in each category. For a new project, both bars will be grey.

The Compendium panel
The Compendium panel shows a list of all the experiments you have selected to be in your compendium. It will be empty for a new project. Clicking an experiment in this list shows its metadata on the right-hand side, along with the options to blacklist it or its entire replicate group. The name of the replicate group can be edited by clicking it. The group description can be edited on the right. The experiment aliases are automatically generated based on the group name. The right-hand side also contains buttons to either export the metadata of the current compendium as a tab-delimited file, or to download a Prose package to start processing the RNA-Seq data. In the list, a green checkmark will indicate if this experiment was uploaded to SRA as an RNA-Seq experiment. A red ‘X’ will indicate a non-RNA-Seq experiment, which are not suited for processing with Prose. Hovering over the symbol will show the experiment type.

IV. Curation page
Once your query has completed, you can click ‘CURATE’ to view the matching studies on the curation page. This page is divided into the study list on the left, the experiment list in the center, and the experiment details on the right. When opening the Curation page only the study list will be visible. To go back to the Project page, click the back button in your browser or in the bottom left of the page.
Study List
The study list displays all new studies that match the user’s query, as well as all the studies that have already been curated. A progress bar is shown below the list that indicates the number of studies that are in the compendium (blue), are pending curation (grey), and are blacklisted (black). Clicking one of these colors toggles between hiding and showing the corresponding studies in the list. For a new project, the entire progress bar will be grey.

Individual studies are presented as boxes in the Study list. These boxes contain some basic information about the study. Clicking the study accession on the left in this box opens up a new browser tab to this study on the SRA website. A colored circle in the upper-right corner indicates the status of the study: (orange) the study’s metadata has not yet been retrieved from SRA, (yellow) the study’s metadata is currently being retrieved from SRA, (grey) the study is ready to be curated, (blue) the study is fully curated and has experiments in the compendium, (black) all experiments in this study have been blacklisted, (red) an error has occurred while fetching the metadata from SRA. The progress bar in the bottom indicates the proportions of the study’s experiments that are in the compendium (blue), are pending curation (grey), and are blacklisted (black).

Once a study’s metadata is loaded, hovering over its box in the list rolls out a panel from the bottom of the screen. This details panel contains the full study title, a list of all included organisms, the study abstract, and the design of the study. If the study is linked to a publication, a link to the PubMed page will be provided here as well. Clicking the study box activates this study and loads its experiment metadata in the Experiment List panel. When a study is activated, you can hover the mouse over the collapsed details panel in the bottom to roll it out again.

Retrieving metadata from SRA
Curse keeps a cache of study and experiment metadata in its database. When a study is included for the first time in the session of a user, all the associated metadata will be retrieved from SRA and will be stored permanently by the Curse server. This will be visualized by the status of the study changing from orange, to yellow, to grey. For smaller studies this will take mere seconds but for larger studies this may take some time. Once the study metadata has been fetched for a single user however, it will be available much faster to all users in the future. Hovering over a study with the mouse will give it the highest priority to retrieve its metadata.

Note that Curse was built to accommodate RNA-Seq studies that rarely exceed 500 experiments. On SRA, studies of other types can be found with thousands of experiments, such as large-scale genome sequencing projects. Curse can still load and display these, but the usefulness of the features and the performance of the tool will suffer. If the number of experiments in a study exceeds 500, this number will be underlined in red in the study list.
Experiment List
When a study is clicked in the Study List, its experiments will be displayed in the Experiment List. Similar to the Study List, each experiment in the Experiment List is represented as a box of metadata. Most of the information you will need to curate the study will be presented in these boxes.

Text-mining: Metadata alignment
The three most important pieces of metadata you need to interpret the properties of an experiment are the experiment title, the sample title, and the sample attributes. On the SRA website, this information is scattered on separate pages. In Curse, it is displayed compactly in the experiment list. Very often, experiments from the same study will receive experiment- and sample titles that share the same structure. Similarly, experiments will often have the same attribute fields. Certain words in the titles and attributes will however differ between experiments, such as ‘wild type’ vs. ‘mutant’ or ‘treated’ vs. ‘control’. These words are vital in recognizing replicates and in interpreting the properties of the experiments. Curse implements a type of multiple-sequence alignment of the experiment metadata in order to identify these important words (See Addendum). Based on these alignments, Curse provides three key features to the user:

1. Automatic highlighting of important words. The words that do not match between the experiments in a study are highlighted in blue. Hovering over such a word will highlight matching and mismatching words from other experiments in green and red respectively.

2. Automatic sorting of experiments. When loading the experiment list, Curse will attempt to sort the experiments to place replicates adjacent to each other. Sometimes, this sorting will fail to accurately group replicates, but users can fix this by clicking on the highlighted words to re-sort them. Curse will then re-order the experiments alphabetically or numerically on that word, preserving the previous order as much as possible.

3. Visualization of distances. Based on the alignment, distances can be calculated between experiments. These distances are visualized by empty space between the experiment boxes. This places replicates tightly together and will visually separate them from the rest of the experiments.

These features rely on the completeness and accuracy of the available metadata. Following the ‘garbage in, garbage out’ principle, missing or inconsistent information, mistakes in the metadata, or inconsistent data entry by the original uploader can cause the alignment to fall short for some studies.

Curating experiments
The user’s task is to curate experiments by either placing them on the blacklist or by selecting them for the compendium. In order to select an experiment for the compendium, they need to be assigned to a replicate group, which should contain both technical and biological replicates. Each group gets a number ranging from 1 to 999. This group number is shown in the group selection display in the right hand side of the experiment box. Experiments pending curation and experiments on the blacklist will display a grey ‘-’ and a black ‘B’, respectively.
After clicking a grouped experiment, its group can be given a name and a description in the Experiment Details panel. Once all experiments in the study are assigned to a group or are placed on the blacklist, the study will be considered to be fully curated and its status color will change from grey to blue (or black if all experiments were blacklisted). Experiments that were assigned to a group will be displayed in the Compendium panel on the Project page. There, group names and descriptions can be edited as well. To avoid loss of progress, all changes to group assignment or naming are automatically saved on the Curse server.

There are multiple ways to assign experiments to groups or to the blacklist:

1. **Group display buttons** - The ‘+’ and ‘-’ buttons above and below the Group selection display can be used to change the group number. Experiments pending curation and on the blacklist are in groups zero and minus one respectively.

2. **Dragging** - Clicking the empty space of an experiment box and dragging over other experiments will assign all dragged experiments to the same replicate group. Doing the same by right-clicking and dragging will place them on the blacklist instead.

3. **Right clicking** - Right clicking in the empty space of an experiment box places that experiment on the blacklist. Right clicking on a highlighted word places all experiments with a matching word on the blacklist as well. Right clicking on the study box in the Study List places all experiments in the study on the blacklist.

4. **Quick-group widget** - (see below)

When a study is opened, it is possible that some of the experiments are automatically blacklisted. These are experiments that belong to the study, but did not match the query the user has provided. For example, experiments that do not match the species or the sequencing strategy.

**The Quick-group widget**

The quick-group widget in the lower-right corner of the page can be used to change the curation status of multiple experiments of the study. It consists of three spheres that represent all experiments that have been grouped (blue), are pending curation (grey), and have been blacklisted (black). These spheres can be dragged into one another with the mouse to move all experiments from one group to another:

![Quick-group widget](image)

Six different operations are available through this widget by dragging...

- **... BLUE into GREY** resets all grouped experiments back to pending curation
- **... BLUE into BLACK** places all grouped experiments on the blacklist
- **... GREY into BLUE** automatically groups all experiments that were pending curation
- **... GREY into BLACK** places all experiments that were pending curation on the blacklist
- **... BLACK into BLUE** automatically groups all blacklisted experiments
- **... BLACK into GREY** resets all blacklisted experiments back to pending curation

The automatic assignment to groups uses the distances between the experiment metadata, which is represented by the empty spaces between experiment boxes, and depends on accurate sorting of the experiments. The threshold Curse uses to decide which experiments are similar enough to be replicates can be adjusted by dragging the slider on the left side of the widget up or down. While adjusting the slider, the suggested groups will be shown with red borders around the experiment boxes.
Visual cues for replicate groups
While curating, visual cues are applied to highlight the grouping of replicates. Blacklisted experiments will receive a grey background. Grouped experiments will be presented with a transparent blue border around them and a red line above and below them. The red lines draw attention to groupings that separate two experiments that were found to be highly similar by Curse. If experiments that are not adjacent to each other are placed in the same replicate group, their group number will turn red. This avoids erroneously addition experiments to an existing group instead of creating a new one. These visual cues depend on the correct sorting of the experiments.

Experiment details
Clicking an experiment in the Experiment list opens up additional metadata in the Experiment details panel. The experiment accession, title, species, sequencing strategy, library source, and library selection info are shown in the top. The experimental protocol, if one is available, is shown below. If the experiment was assigned to a replicate group, the name and description of that group can be edited in the text fields at the bottom.

If you intend to use Prose to process RNA-Seq data, keep in mind that file names will need to be created from the group names. Avoid illegal characters such as ‘/’. If no name is provided, the group will receive a generic name like ‘group_3’. Based on these group names, the experiment aliases are generated. These can be seen in the Compendium panel on the Project page and will be used to track the experiments in the Prose pipeline.
V. Exporting metadata

Once your compendium has been completed, the associated metadata can be exported from the Project page. The exported file will be a tab-delimited table file that can be opened in Excel. It contains the following columns of information for each experiment in your compendium:

1. studyAccession: The SRA accession number of the study
2. experimentAccession: The SRA accession number of the experiment
3. sampleAccession: The SRA accession number of the sample
4. runAccessions: A comma separated list of SRA accession numbers of the runs
5. experimentAlias: The alias of the experiment, generated from the group name
6. groupName: The name of the replicate group. This can be edited in Curse
7. groupDescription: The description of the replicate group. This can be edited in Curse
8. strategy: The sequencing strategy (e.g. ‘RNA-Seq’)
9. platform: The sequencing platform (e.g. ‘Illumina HiSeq 2500’)
10. layout: ‘SINGLE’ for single-end or ‘PAIRED’ for paired-end
11. reads: The total number of reads summed over all runs
12. species: The scientific name of the species
13. studyTitle: The title of the study
14. experimentTitle: The title of the experiment
15. sampleTitle: The title of the sample
16. attributes: A ‘;’-separated list of attribute key-value pairs

This metadata file is required by Prose to download and process the RNA-Seq data of the experiments in your compendium. Even if you do not intend to use Prose and want to run a pipeline of your own, the run accessions in this file can be used to download the data from SRA by using the sra-toolkit.

Alternatively, you can download a Prose package to process RNA-Seq experiments in your compendium in order to construct expression atlases and co-expression networks. See the Prose manual for more information.

VI. Addendum - Metadata alignment algorithm

The text alignment algorithm of Curse supports its most useful features. The algorithm is based on a Smith-Waterman alignment, frequently used in local sequence alignment of DNA. Here, words in the description are aligned as if they were nucleotides in a sequence. Mismatches and gaps receive a penalty in the alignment. Smith-Waterman can however only align two sequences to each other and is by default not suited for the alignment of multiple sequences. To make this possible, Curse starts by aligning the descriptions of the first two experiments to each other. Based on the alignment result, a new composite description is created that includes all gaps and mismatches. Additional descriptions are then iteratively aligned to and incorporated into this composite description. Within this composite, words that mismatch or span gaps are identified and marked as important. These are then highlighted in the visualization. The composite also allows Curse to quickly calculate pairwise distances between two experiments by counting the mismatches and gaps between them. These distances are visualized as a physical separation of the experiment boxes with a black spacer.

During alignment, if an experiment is found to be too dissimilar to the composite, it will not be incorporated into it but will instead be used to create a new composite to which the other sequences can align. This is preferred to adding it to the existing composite, since it would introduce too many gaps and mismatches, which would mark almost every word as important.