CellNet web application tutorial

How to use the CellNet web application at http://cellnet.hms.harvard.edu to analyze your gene expression data

Introduction

To use our website to analyze your expression data, first, you need to upload you raw expression data and a sample annotation table, and select several parameters. Once you have pressed the 'Run CellNet' button, your files will begin to upload. Once all of the files have been uploaded to our server, your analysis will be queued for processing on Harvard Medical School's Orchestra compute cluster. At this point, a new page will be displayed that lists the progress of the analysis. Typically, the longest part of the run is uploading the data. Once the raw data is uploaded, executing CellNet will take several minutes and you will be notified by email with instructions to download the analysis results. The results include normalized data, classification scores, GRN establishment scores, network influence scores, and corresponding figures. More information about the results is listed at the end of this tutorial, on the website's FAQ, and in the paper describing the platform (Cahan et al. Cell 2014).

Filling out the 'Run CellNet' form

There are 8 steps to be completed on **cellnet.hms.harvard.edu/run/.** Each step asks for a specific input, and all are required, except for '*Step 3. What column in your sample table indicates which samples (if any) are replicates?*' which is optional. Below is a guide to each step:

> **Step 1. Upload your raw expression data**.
> You may upload uncompressed raw expression files (e.g. .CEL files for Affymetrix data). However, to reduce the upload time, we suggest that you compress all of your data files based on the same array platform and species using a compression utility. zip, gzip, or bzip compression formats are accepted. **The maximum upload size of all files together is 128MB**. If your dataset is larger, you can use the CellNet R package locally to analyze your data, or you can break up your data set into smaller pieces and analyze them separately.

> **Step 2. Upload a table that describes your data**
> CellNet needs to know how (and if) your samples are grouped into replicates, and the preferred order of the samples. Please make and upload a sample annotation file. The format of the sample annotation table must include the following columns:
> - sample_id
> - sample_name
> - description1 (specifies experimental grouping of samples)
> - file_name (must exactly correspond to the CEL files' names of your raw expression data (see Step 1) above)
>
> To make a sample annotation file you can edit our example file, which is linked to from the 'Run CellNet' page, in any text editor. If you use MS Excel, be sure to 'Save as' in the '.csv' format.

> **Step 3. What column in your sample table indicates which samples (if any) are replicates?**
> In the example above, this is 'description1'. This field is optional. It is used in two ways. First, the order of the samples in the sample table is how the samples will be ordered in all of the results files. Second, the sample table is used when displaying GRN statuses. If there are replicates, then the GRN status bar plot will show the mean GRN status of each experimental group and standard deviation bars. If left blank, then each sample will be displayed in the GRN status bar plot.

> **Step 4. Name this analysis**.

Please provide a brief name for this analysis. The result files, which will be available for download will include this name.

Step 5. What species was profiled?
Select either mouse or human. Only after a species as been selected may you select the microarray platform in Step 6.

Step 6. What microarray platform was used?
Please select the microarray platform that was used to profile your samples. Only Affymetrix platforms are allowed for the web application. Illumina users can use CellNet R package. You must choose a species in Step 5 before you can select a platform. Selecting the wrong platform will produce an error. The microarray platform information can be found from the Affymetrix support website (http://www.affymetrix.com/support/mas/index.affx#1_1) or from the NIH GEO database (http://www.ncbi.nlm.nih.gov/geo/).

Step 7. What is the target tissue or cell type?
Please select the tissue or cell type that was the intended end point of the cell engineering study. For example, if you were reprogramming fibroblasts to pluripotent stem cells, you would choose 'embryonic stem cells' here. This is used to calculate the Network Influence Score, which prioritizes candidate transcriptional regulators according to their dysregulation and their importance to the target cell type gene regulatory network. You must choose a species (Step 5) and platform (Step 6) before you can select a target tissue or cell type.

Step 8. Notification via Email Address
Email address, used only to notify you of the URL where your completed analysis may be downloaded.

After pressing the 'Run CellNet' button and your data has uploaded, you will see a page that looks like:

When the analysis is done, you will receive an email that contains a link to your results, as well as a guide to interpret your results. The results are compressed into a single file (.zip), and can be uncompressed on most platforms by double-clicking on the results files.

Interpreting your results

We highly recommend that you read Cahan et al. Cell 2014, which describes in detail how CellNet was made and several example analyses of cell engineering experiments. In brief, CellNet compares your expression data to a large compendium of expression data sets to determine the extent to which cell and tissue specific gene regulatory networks (GRNs) were established. The results file names are prepended with the analysis name that you entered in Step 4. The three main outputs of CellNet included in your results are:

1. **Cell and tissue type classification.** This is a data table in which each row represents one cell or tissue type, and each column represents one of your samples. The file name will be analysis name (which you entered in Step 4) _classRes.csv. The values represent the classification score, and reflect the probability that a sample is indistinguishable from a cell or tissue type by gene expression. We typically represent this as a heatmap, on a black-> green->yellow scale (0.0->0.5->1.0), where the rows and columns are ordered as in the data table. This heatmap, and all other plots are provided in PDF format, named [date of analysis_plots.pdf]. The panels are easy to import into graphics programs such as Adobe Illustrator for re-formatting.
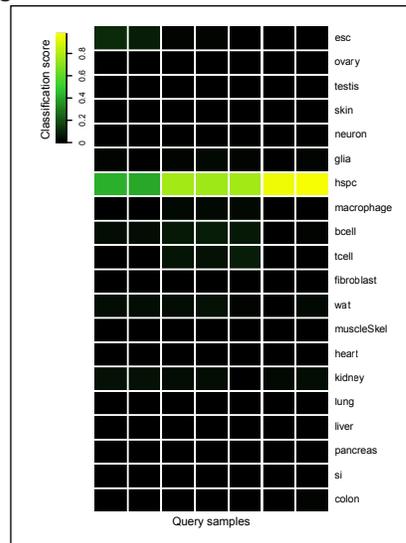
2. **GRN status.** This is a data table in which each row represents one GRN associated with a particular cell or tissue type, and each column represents one of your samples. The values represent the extent to which the GRN is established to a level equivalent to that seen in the associated cell or tissue type. The file is named grnScores.csv in your results. We represent this as a bar plot, were replicates are combined into one bar (colored light blue), and the training data of the starting and target cell types are also shown as points of reference (dark blue).
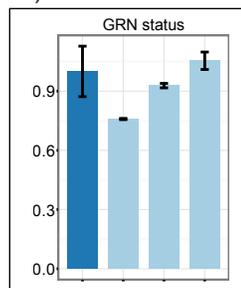
3. **Network influence score.** The Network influence score is computed for each transcriptional regulator in the target cell type GRN according to the extent that it is either too highly or too lowly expressed in your terminal sample. The NIS also integrates other information into scoring transcriptional regulators, including the extend to which predicted target genes are dysregulated, the number of transcriptional

targets, and the expression level of the regulator in the target cell type. Positive values indicate that the regulator is too highly expressed, and negative values indicate that the regulator is too lowly expressed. The results include a file named NIS.csv that includes the NIS scores of all transcriptional regulators of the target cell type GRN in all of the samples. In the paper, we visualized the NIS as a bar plot. In your results, we represent the NIS as a heatmap, and only display the 50 most lowest scoring transcriptional regulators.
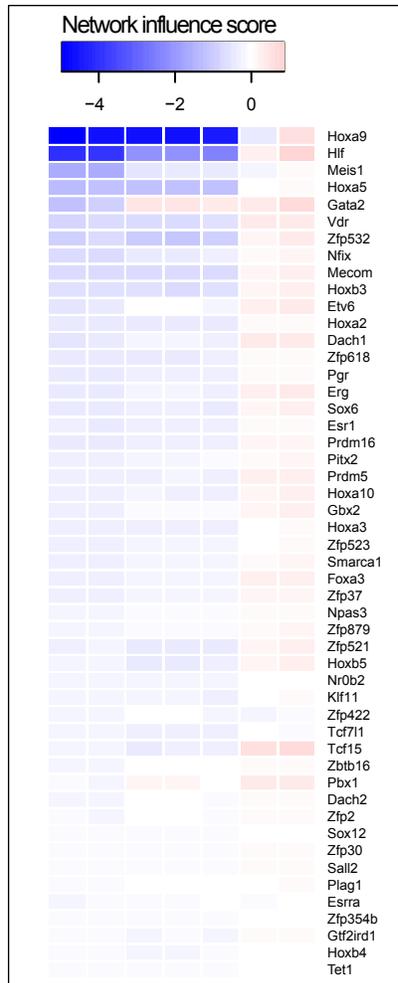
4. **Other output.** We also include a log file that lists the steps in your analysis, and expNorm.csv, which contains your normalized data.

Problems analyzing your data?

There are several issues that could impede the analysis of your data:

- Data upload: Were all of the files listed in your sample table selected for upload, or included in the compressed file? Did you include at least two samples?
- Sample annotation file: We strongly recommend that you use the provided template file to make your own sample annotation table. It is crucial that, if you make it in Microsoft Excel, that you save the file as a .csv.
- Platform: Does the platform that you selected in Step 6 match the platform that was used to generate all of your data files.

If you have addressed all of these potential issues and are still having problems, please visit the CellNet web application user group at https://groups.google.com/forum/#!forum/cellnet_webapp.