Stemformatics: A user-friendly database of well-curated biological data

1. Introduction to Stemformatics

Stemformatics [8] is a collaboration platform between the stem cell and bioinformatics community.

We were motivated by the plethora of exciting cell models in the public and private domains, and the realisation that for many biologists these were easily inaccessible. We wanted a fast, user-friendly way to find and visualise interesting genes in these exemplar stem cell datasets. With Stemformatics, anyone can find data from leading stem cell laboratories in a format easy to search, visualise and export. No prior bioinformatics knowledge is required. Stemformatics is aimed at bioinformaticians and biologists wishing to collaborate. It is designed and maintained with the intention to remove many of the friction points common to such collaborations.

2. Stemformatics data statistics

Currently we have 141 datasets publicly available and 114 private datasets

<table>
<thead>
<tr>
<th>Dataset status</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUBLIC</td>
<td>141</td>
</tr>
<tr>
<td>PRIVATE</td>
<td>114</td>
</tr>
</tbody>
</table>

3. Data processing pipeline

We process microarray and RNA-Seq data.

- User request submitted
- Quality assessment of experimental design
- Raw data and sample information obtained
- Data enters pipeline
- Normalisation
- Quality control checks
- Online annotation
- Quality control checks
- Visualisation in Stemformatics

Figure 1. A snapshot of the Stemformatics data analysis pipeline.

The Stemformatics team performs various steps to ensure data reliability. To illustrate, of the 463 dataset requests we have received to date, only 141 have been made public on Stemformatics.

4. Visualisation tools

Stemformatics implements a series of tools that allow users to easily:

(A) Identify similar gene expression trends

<table>
<thead>
<tr>
<th>Gene</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZBTB7A</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>NLRP9</td>
<td>4.71E-02</td>
</tr>
<tr>
<td>PEX5L</td>
<td>7.00E-02</td>
</tr>
</tbody>
</table>

(B) Visualise amplitude of change in expression between samples

<table>
<thead>
<tr>
<th>Sample 1 Value</th>
<th>Sample 2 Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probe ID</td>
<td>Change</td>
</tr>
<tr>
<td>SAMPLE_1</td>
<td>SAMPLE_2</td>
</tr>
<tr>
<td>X</td>
<td>5.00E-02</td>
</tr>
<tr>
<td>7.00E-02</td>
<td></td>
</tr>
<tr>
<td>0.00E+00</td>
<td></td>
</tr>
</tbody>
</table>

(C) Obtain a measure of similarity for expression between samples

Figure 2. The heatmap shows expression patterns for the CLEC gene family with respect to sample type [7].

Samples are presented on the X axis, genes on the Y axis. Genes and samples are grouped by similarity. Red cells indicate that the gene for a given sample is highly expressed. Blue cells indicate low expression.

Figure 3. Results of StringDB [7] summarise known protein interactions with Th17-related ZBTB7A.

(B) Pathway analysis (InnateDB [8], not shown)

(C) Genome browser (UCSC)

Figure 4. A screenshot of the UCSC Genome Browser [4]. It is both highly visual and interactive, allowing users to browse-annotated genomes.

5. Linkout to public resources

Stemformatics also supports direct linkage of user data to public resources:

(A) Network analysis (StringDB)

(B) Pathway analysis (InnateDB [8], not shown)

(C) Genome browser (UCSC)

Figure 5. Yeast [1] allows users to visualise normalised gene expression data across multiple independent datasets.

(B) Compare multiple gene expression levels

Figure 6. Expression levels of genes involved in the caffeine metabolism pathway [2, 3] are shown.

(C) View multiple datasets at once (not shown)

As with most objects in Stemformatics, these images and their underlying data can be exported and/or shared via email with a click.

6. Special Features

Additional special features in Stemformatics:

(A) Compare gene expression across datasets

(B) Visualise amplitude of change in expression between samples

(C) Obtain a measure of similarity for expression between samples

7. Why bioinformaticians find Stemformatics useful

We have a repository of highly curated datasets with standardised formats. We also offer users the ability to batch download; selecting and downloading multiple independent datasets at once.

All our methods and quality control plots are accessible online through Stemformatics.

8. Future Directions

Stemformatics will continue to add new datasets and analyses that are of interest to the stem cell community. It is possible that this platform will evolve to support additional model organisms.

9. References & Acknowledgements

Figure 7. A screenshot of the Stemformatics data analysis pipeline.

The Stemformatics team performs various steps to ensure data reliability. To illustrate, of the 463 dataset requests we have received to date, only 141 have been made public on Stemformatics.