WebMeV: A Robust Platform for Intuitive Genomic Data Analysis (U24)

ITCR Teleconference
Oct 4. 2019

Yaoyu E. Wang
John Quackenbush
Harvard T.H. Chan School of Public Health
Brief History of WebMeV

• Started as a stand-alone application for microarray analysis (MeV)

• Funded by NCI ITCR program under U01 to migrate from stand-alone to cloud-based application as WebMeV

• Again funded by NCI ITCR program under U24 to expand its functionalities
Overall Goal

To help assure that analytical access to large public data is democratized so that scientists and physicians can test hypotheses by directly interacting with the data in a way that is not limited by their available computational resources and in a system that helps ensure their research is reproducible.
Welcome to the Web MEV

WebMEV (Multiple Experiment Viewer) is a cloud-based application supporting analysis, visualization, and stratification of large genomic data, particularly for RNASeq and microarray data.

With WebMEV platform, you can:
- Perform RNASeq Analysis - Perform differential expression analysis using RNASeq raw count data to draw biological insights.
- Access to Public Domain Data - Directly search and pull TCGA and GEO gene expression and sample attribute data in addition to private data for analysis.
- Stratify Cohorts using Clinical Attributes - Perform complex cohort stratification using sophisticated regular expression, facet filter and set operations.
- Web MEV is being built to meet the challenge of exploring large public genomic data set and Next Generation Sequencing data with intuitive graphical interface for analysis. MEV is a free and open-source cloud service platform that does not require log in to use.
- Where to start - A series of videos tutorials are available on YouTube and accessible after entering the WebMEV example files for count matrix and sample attribute are available here.
- Customized Analysis - Complex bioinformatics analysis outside the scope of WebMEV functionalities can contact the Center for Cancer Computational Biology at Dana-Farber Cancer Institute as consulting project. Click here for more information.
Private Data and Public Data Retrieval

Upload a dataset
Use the Upload button to upload a tab-delimited (tsv) data file. The first row is assumed to contain headers and the first column contains the feature names (i.e. gene names). Here is an example RNASeq read count file and an example of metadata file (metadata.txt) that can be loaded once entered into the platform.

MeV Dataset Archive
You can download the entire MeV dataset as a ZIP archive by clicking the icon next to the dataset name. This archive will contain the original data as well as any analyses and annotations.

Note on Data Storage
MeV does not store your data on the server beyond the duration of your session. Once your session times out - all your data is deleted from the server.

Your dataset and analyses are saved to the browser’s local storage on your computer. You will find them under “Offline Datasets” section below “Your Datasets”.

You may activate an offline dataset for viewing or further analysis by clicking the icon next to the offline dataset.

Please note, the amount of disk space available to save offline datasets is determined by your browser. If this disk quota is exceeded the browser will delete the offline datasets.

To ensure your MeV dataset is saved indefinitely you may download it to your machine as a zip archive.

Tutorials
Please see the videos in the Web MeV Tutorials playlist.
Keyword Search for GEO Import
Interactive Data Visualization for Transcriptomic Data and Analysis Results
Interactive Data Visualization for Transcriptomic Data and Analysis Results
Interactive Data Visualization for Transcriptomic Data and Analysis Results
### Cohort Construction Interface

#### Import TCGA Datasets

**BRCA clinical_annotations.tsv**

<table>
<thead>
<tr>
<th>Facet / Filter</th>
<th>Undo / Redo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refresh</td>
<td>Reset All</td>
</tr>
<tr>
<td>Reset All</td>
<td>Remove All</td>
</tr>
</tbody>
</table>

#### View Details
- View cohort details
- View aggregate statistics
- View value distribution

#### Actions:
- Filter data to analyze for selected cohort
- Search by self define facets
- Build composite phenotypes
- Build cohort sets

### BRCA Clinical Annotations Table

<table>
<thead>
<tr>
<th></th>
<th>884 rows</th>
<th>Show as: rows records Show: 5 10 25 50 rows</th>
<th>BRCA clinical Annotations tsv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### BRCA Clinical Annotations

- **Days to Death**
  - Numeric: 124
  - Non-numeric: 700
- **Histological Type**
  - Infiltrating Lobular Carcinoma
  - Infiltrating Ductal Carcinoma
  - Medullary Carcinoma
- **Other specify**
  - Mixed Histology (please specify)
  - Mucinous Carcinoma

#### Table Example

<table>
<thead>
<tr>
<th></th>
<th>histological_type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Infiltrating Lobular Carcinoma</td>
</tr>
<tr>
<td>2</td>
<td>Infiltrating Ductal Carcinoma 0</td>
</tr>
<tr>
<td>3</td>
<td>Infiltrating Ductal Carcinoma 3+</td>
</tr>
<tr>
<td>4</td>
<td>Infiltrating Ductal Carcinoma 1+</td>
</tr>
<tr>
<td>5</td>
<td>Medullary Carcinoma</td>
</tr>
<tr>
<td>6</td>
<td>Infiltrating Ductal Carcinoma</td>
</tr>
<tr>
<td>7</td>
<td>Infiltrating Ductal Carcinoma</td>
</tr>
<tr>
<td>8</td>
<td>Infiltrating Ductal Carcinoma</td>
</tr>
<tr>
<td>9</td>
<td>Infiltrating Ductal Carcinoma</td>
</tr>
</tbody>
</table>

#### Table Columns
- histological_type
- days_to_death
- Other specify

#### Table Rows
- BRCA clinical Annotations tsv
- BRCA clinical Annotations tsv
Manage and Annotate Cohort Sets
Specific Aim 1: We will maintain, support, and extend the WebMeV infrastructure, providing it to the community as a sustainable, open-source, cloud-based web application for the analysis of large high-dimensional genomic data, particularly Next-Generation Sequence (NGS) data.

- Refactor Source Code
- Optimize Data Visualization
- Containerization
- Bioconductor/Python Integration
- Community Outreach
Specific Aim 2: We will develop a drag-and-drop pipeline for RNA-Seq raw data mapping and normalization, allowing users to easily move from raw fastq sequence data to analysis of patterns of gene expression.

Figure 4. a. Basic WebMeV architecture for sequence data analysis connecting to CNAP. Users will be able to upload fastq files by linking a Dropbox folder, process, and further analyze results using WebMeV.
**Specific Aim 3:** We will integrate a wide range of methods for systems biology into WebMeV centered on gene network inference and analysis and develop new interactive network-based displays for gene networks.

**Figure 7.** LIONESS estimate individual sample network

**Figure 8.** ALPACA modules associated with angiogenic ovarian tumors.