

CSE 590 Computational Biology

Final Project

Hari Prasath Raman, Karteek Paruchuri, Arvind Ram Anantharam

December 16, 2015

1 Introduction

This project presents rich and interactive dashboards which enables computational biologists to visualize the results of RNA sequencing experiments with ease. Given the result files from different tools, the application provides an interactive interface to do exploratory data analysis. This application also provide features which helps them to gain insights from the data across different plots in the dashboard.

Points in the displayed plots corresponds to a transcript and upon clicking on these points redirects the user to <http://www.ensembl.org/> for more information about the transcript. The graphs are also tweaked and integrated in such a way that, on hovering over a single transcript ID in one plot highlights the same transcript ID across all other plots. The entire results are displayed as table view so that the entries can be searched, sorted by to gain quick insights over the results. The supported tools for this web interface are,

- Kallisto
- Sailfish
- RSEM

This reports is organized as follows,

1. Dashboard -
 - (a) Features
 - (b) Snapshot
2. Technology Stack
3. Getting Started Instructions
4. Future Extensions

2 UI Dashboard

2.1 GC Content and Length vs TPM

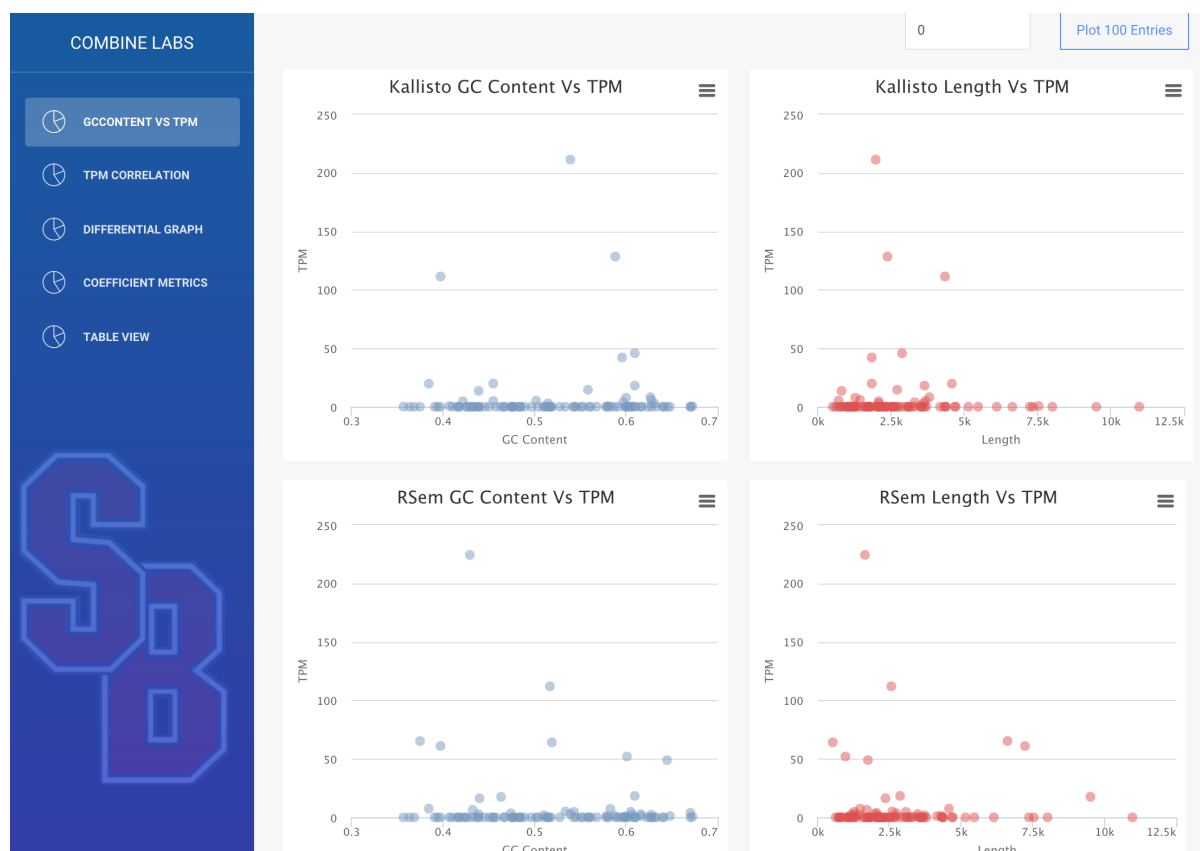


Figure 1: Correlation between GC Content vs TPM , Length vs TPM for all tools

The above figure shows the home page of the website - GCContent vs TPM. The home page presents a summarized view in graphs. As we can see the graphs represent the relationship between a tool's GC content and TPM. Another graph which displays relationship between length and TPM. These graphs are shown for the all the tools considered- Kallisto, RSEM, Saifish. The graphs directly render data for 100 records. A user can click on "Plot 100 Entries" to view the data for next 100 records. If a user selects on any of the dots shown in the graph, he'll be redirected to ensemble.org to view further details about that specific transcript. If a user hover over a point, that specific transcript gets highlighted in other graphs.

2.2 TPM vs Truth

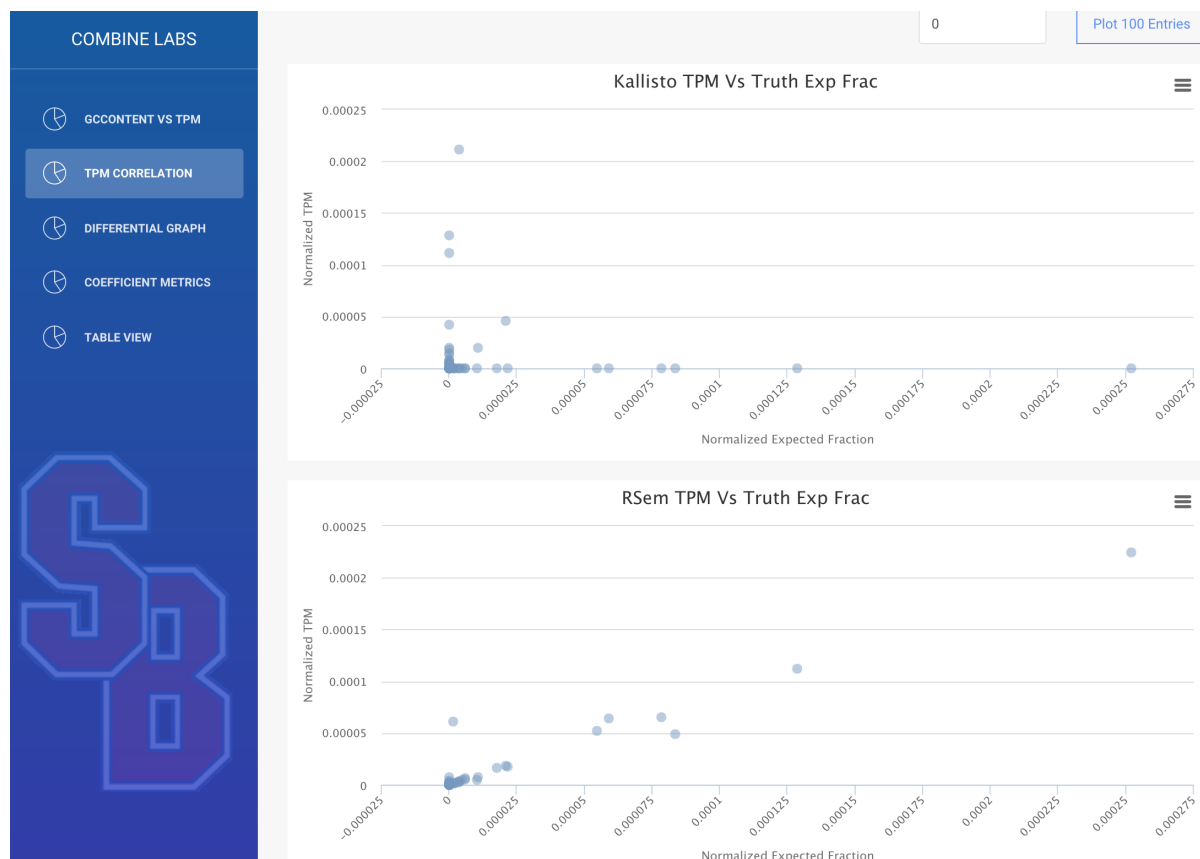


Figure 2: Correlation TPM from all tools vs the Truth value

The TPM Correlation view presents the correlation between the TPMs of all the transcript IDs with its corresponding value from the truth table. The correlation for kallisto, sailfish, rsem are plotted in the same dashboard. If there is high correlation between the truth and the results from different tools, the points will get aligned along the $x = y$ line in the graph. All the TPM values are normalized between 0 to 1 to plot the correlation.

The graphs display the correlation for the first 100 transcript IDs. The offset can be changed by entering the offset value in the text box to plot the next 100 entries from the given offset. The graphs directly render data for 100 records. A user can click on "Plot 100 Entries" to view the data for next 100 records. All the above discussed features - hover highlight, redirection to ensemble.org etc are present in this view as well.

2.3 Differential Graph

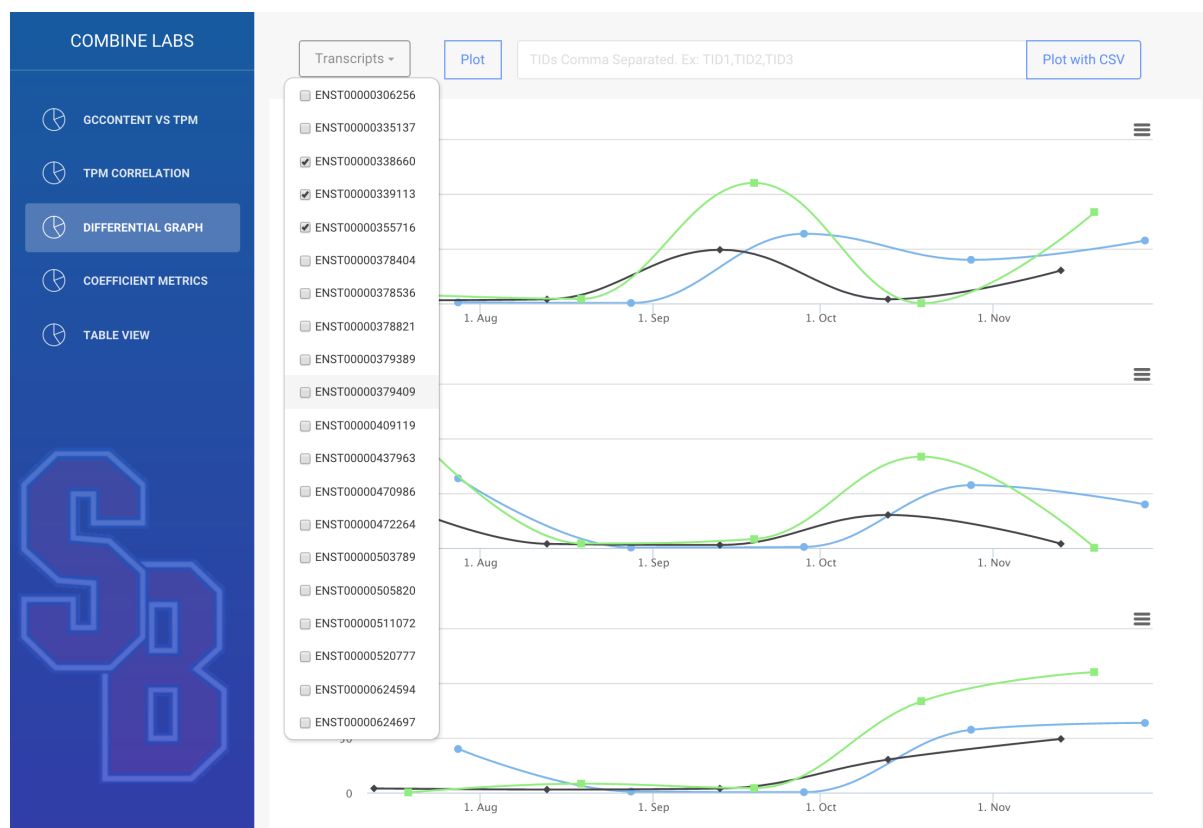


Figure 3: TPM vs Time across experiments for set of Transcript IDs selected from drop-down

The differential graph view compares different transcripts over time. It shows the TPM of the chosen transcript IDs behavior across different experiments. Let us consider the evaluation of regeneration in starfish experiment. We need to find the transcripts responsible for regeneration. This can be identified by watching the transcripts in different portions of the starfish at the same time frame. For this, we might need to identify the pattern of multiple transcripts responsible for regeneration at the same time. This differential data is visualized in the same graph.

The user can select from list of available transcripts or he can enter the transcript IDs by comma separation in the text box. The comma separation entry is shown in the next figure. For example, ENST00000335137, ENST00000306256, ENST00000338660

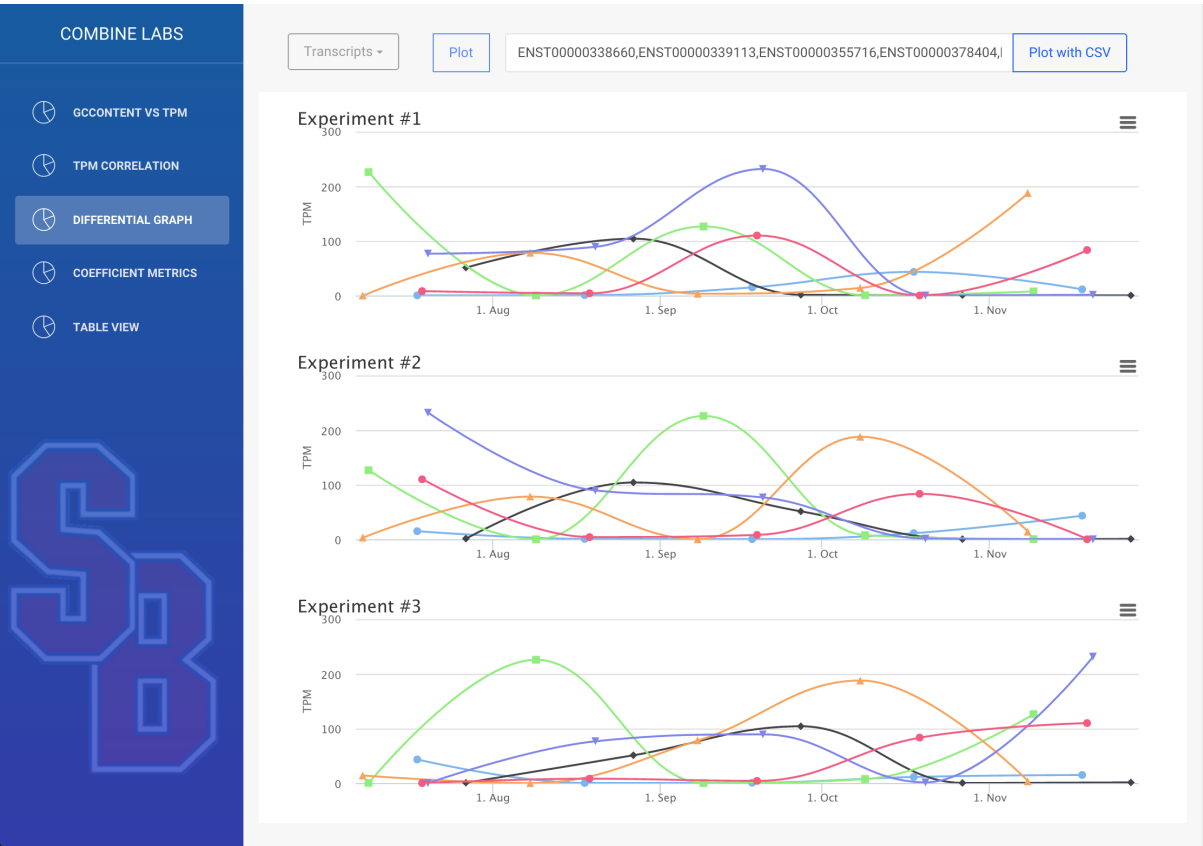


Figure 4: TPM vs Time across experiments for set of Transcript IDs given as CSV

2.4 Coefficient Metrics

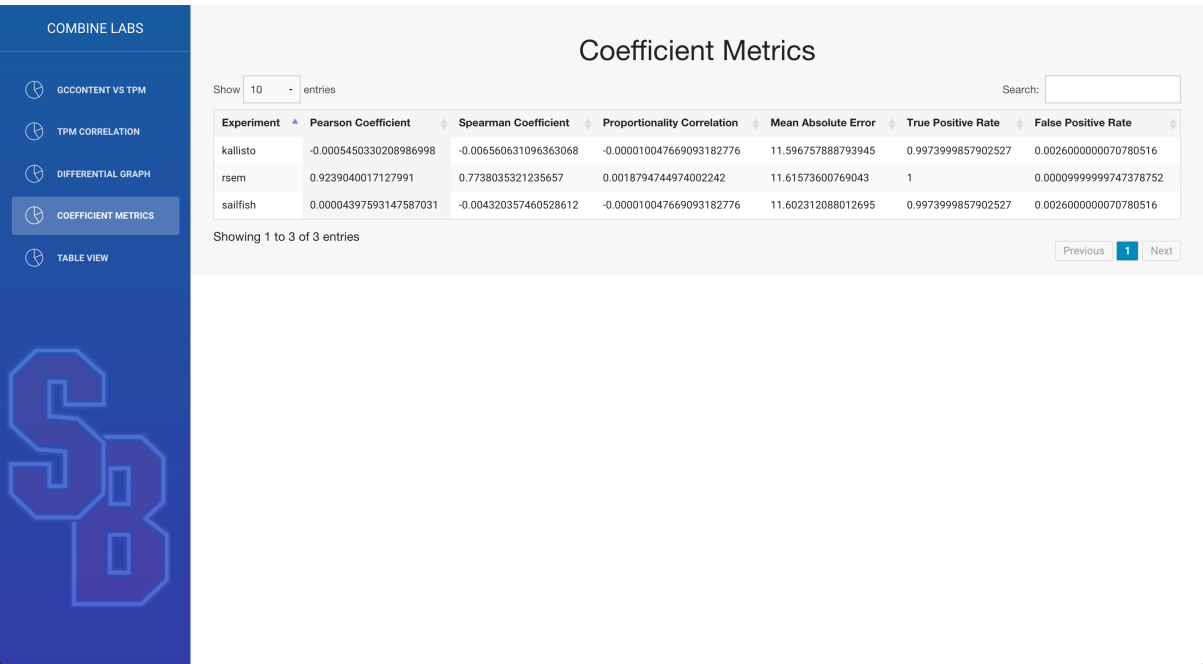


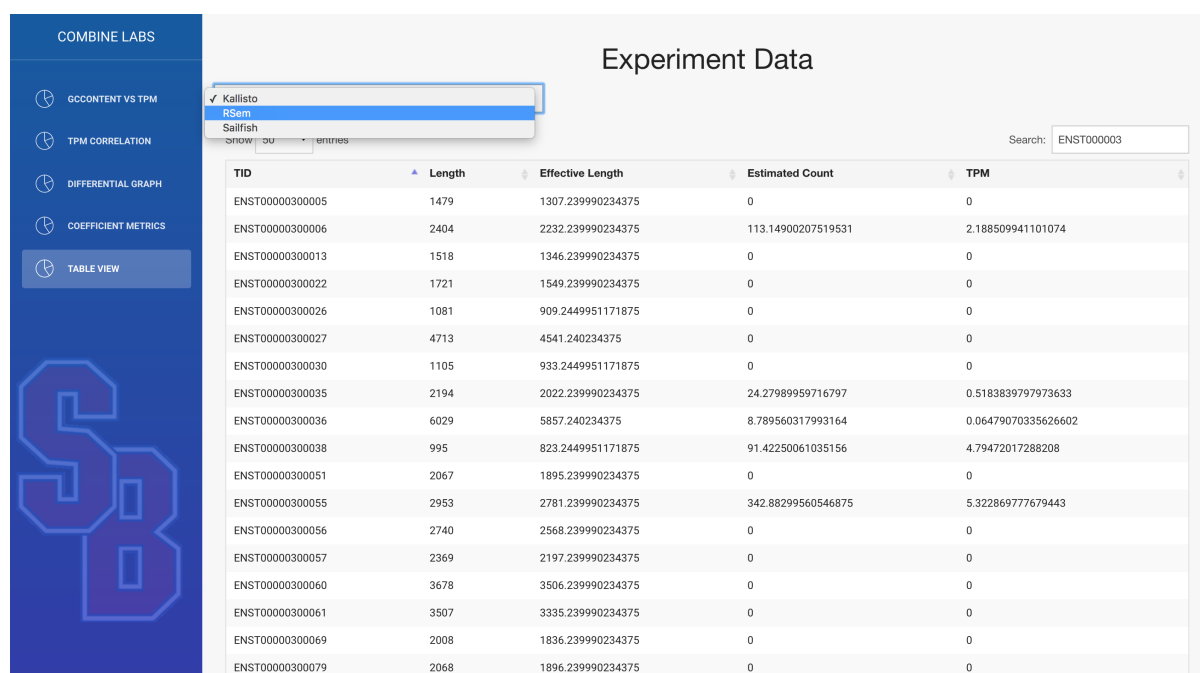
Figure 5: Table showing different correlation metrics from results of Kallisto, RSEM, Sailfish

The coefficient metrics table shows a summarized view of various metrics on data from each tool. The metrics are computed with comparison against truth table data. The TPM value from results of each tool and the truth value are used to compute the following metrics:

- Pearson Coefficient
- Spearman Coefficient
- Proportionality Coefficient
- Mean Absolute Error
- True Positive Rate
- False Positive Rate

The metrics are computed using the scikit-learn python package.

2.5 Table View



TID	Length	Effective Length	Estimated Count	TPM
ENST000003000005	1479	1307.239990234375	0	0
ENST000003000006	2404	2232.239990234375	113.14900207519531	2.188509941101074
ENST000003000013	1518	1346.239990234375	0	0
ENST000003000022	1721	1549.239990234375	0	0
ENST000003000026	1081	909.2449951171875	0	0
ENST000003000027	4713	4541.240234375	0	0
ENST000003000030	1105	933.2449951171875	0	0
ENST000003000035	2194	2022.239990234375	24.27989959716797	0.5183839797973633
ENST000003000036	6029	5857.240234375	8.789560317993164	0.06479070335626602
ENST000003000038	995	823.2449951171875	91.42250061035156	4.79472017288208
ENST000003000051	2067	1895.239990234375	0	0
ENST000003000055	2953	2781.239990234375	342.88299560546875	5.322869777679443
ENST000003000056	2740	2568.239990234375	0	0
ENST000003000057	2369	2197.239990234375	0	0
ENST000003000060	3678	3506.239990234375	0	0
ENST000003000061	3507	3335.239990234375	0	0
ENST000003000069	2008	1836.239990234375	0	0
ENST000003000079	2068	1896.239990234375	0	0

Figure 6: Table showing entire results of Kallisto, RSEM, Sailfish from database

The table view is a visual representation of data available from each tool. The user can select required tool from the dropdown and visualize that particular data. He can sort the visible data by different columns. A search box is also to enable user to narrow down his results. The search functions for the TranscriptID column only. He can select from the dropdown how many records he would like to visualize in a single view.

3 Tech Stack

- Python
- Flash Web Framework
- Scikit-Learn Library
- PostgreSQL Database
- Bootstrap UI
- Highcharts
- Bootstrap Data Tables

4 Getting Started with Tool Development

1. Python Installation
<https://www.python.org/downloads/>
2. Flask Installation
<http://flask.pocoo.org/docs/0.10/installation/>
3. scikit-learn Installation
<http://scikit-learn.org/stable/install.html>
4. PostgreSQL Installation
<http://www.postgresql.org/download/>
5. pg8000 python-postgreSQL driver
<https://pypi.python.org/pypi/pg8000>

Use the *cb_psql_cmds.txt* file to create database tables and import the experiment. The code should be modified to point to the local database. This code change should be done in *corelcoeff.py* and *server.py*. Use the files in the *scripts/* folder to convert, compute and store the results into the database

- *fasta_parser.py*
- *corelcoeff.py*

Then start the server using
python *server.py* and visit <http://localhost:5000> to visualize the dashboard

5 Future Extensions

1. Upload the results of the experiments to plot graphs based on the uploaded data
2. Auto detect the tool name from the uploaded result file.
3. Host the application in the cloud and expose it to the Computational Biologist community.