

WebMeV – A Cloud Based Platform for Genomic Analysis

Yaoyu E. Wang

Associate Director, Center for Cancer Computational Biology
Dana-Farber Cancer Institute

Acknowledgements

Dana-Farber Cancer institute
CCCB

John Quackenbush

Software Engineer

Lev Kuznetsov

Antony Partensky

Bioinformatics

Brian Lawney

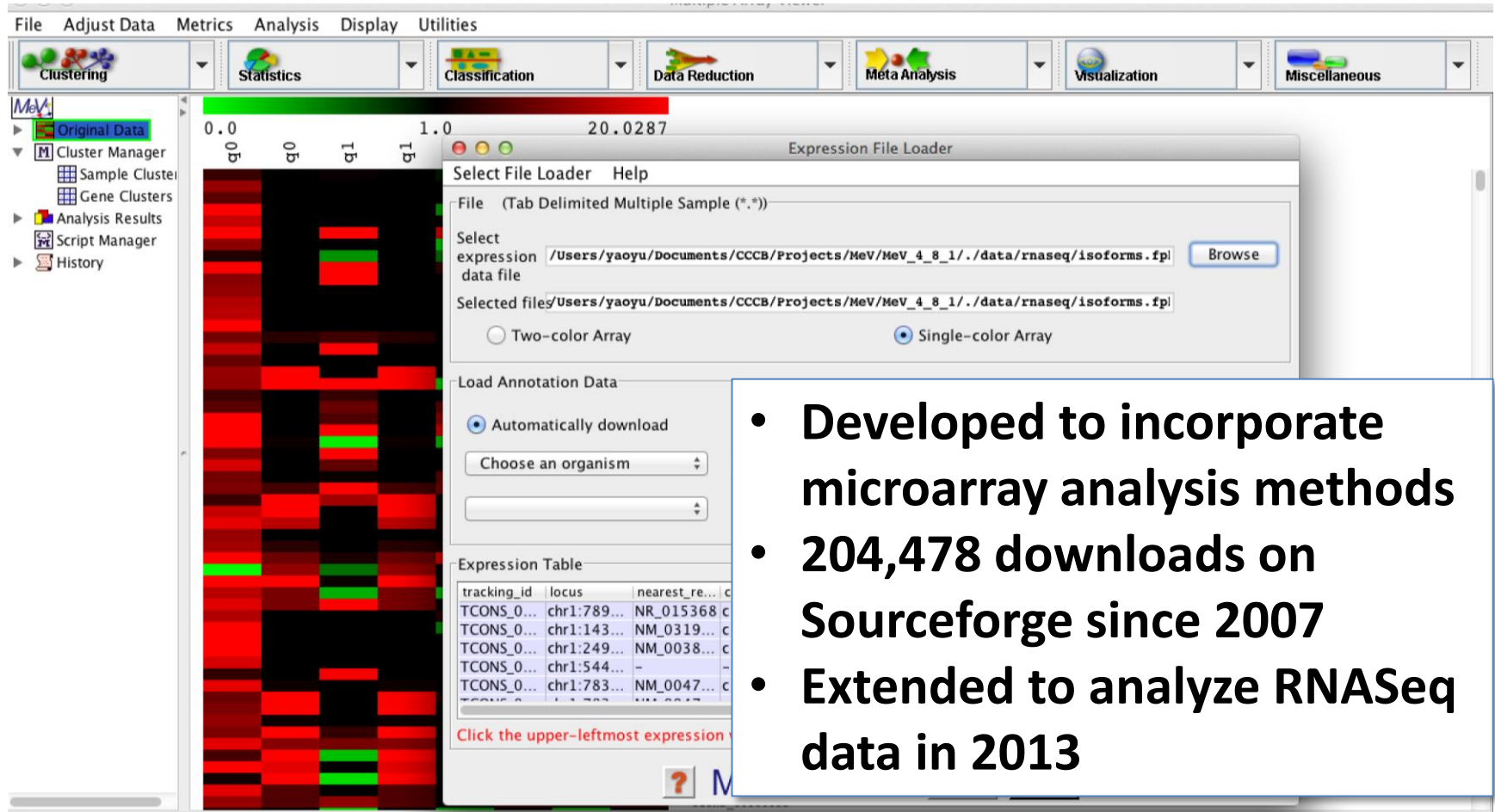
Derrick DeConti

Funding



Standalone Application for Genomic Analysis

MeV – Multi experiment Viewer



File Adjust Data Metrics Analysis Display Utilities

Clustering Statistics Classification Data Reduction Meta Analysis Visualization Miscellaneous

MeV

Original Data

Cluster Manager

Sample Clusters

Gene Clusters

Analysis Results

Script Manager

History

Expression File Loader

Select File Loader Help

File (Tab Delimited Multiple Sample (*.*)

Select expression data file

Selected file: /Users/yaoyu/Documents/CCCB/Projects/MeV/MeV_4_8_1/./data/rnaseq/isoforms.fpl

Two-color Array Single-color Array

Load Annotation Data

Automatically download

Choose an organism

Expression Table

tracking_id	locus	nearest_re...
TCONS_0...	chr1:789...	NR_015368 c
TCONS_0...	chr1:143...	NM_0319... c
TCONS_0...	chr1:249...	NM_0038... c
TCONS_0...	chr1:544...	-
TCONS_0...	chr1:783...	NM_0047... c

Click the upper-leftmost expression v

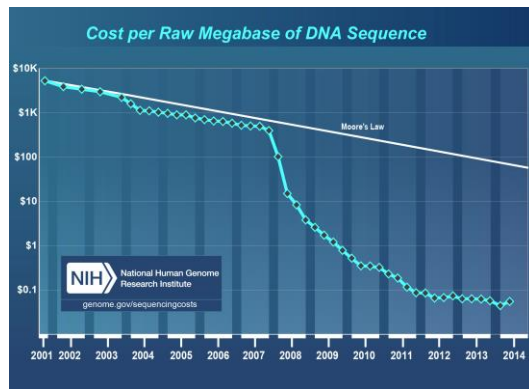
- Developed to incorporate microarray analysis methods
- 204,478 downloads on Sourceforge since 2007
- Extended to analyze RNASeq data in 2013

Standalone Applications Lack Portability and Scalability

- Require maintenance and testing on multiple operating systems
- Application relies heavily on user computing environment
- Developers have limited control over application dependency
- Computing power does not scale with the size of data set
- Require to download datasets onto local machine for analysis

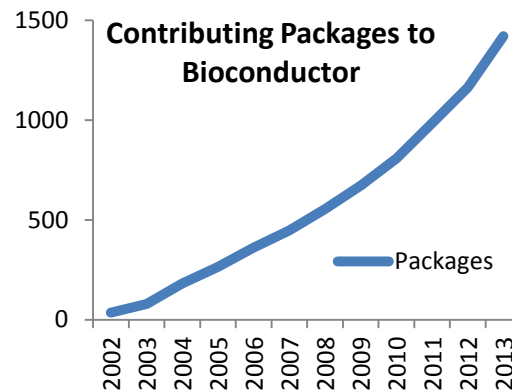
Genomic data and analytical method explosion

Decreasing Cost of NGS Data Generation



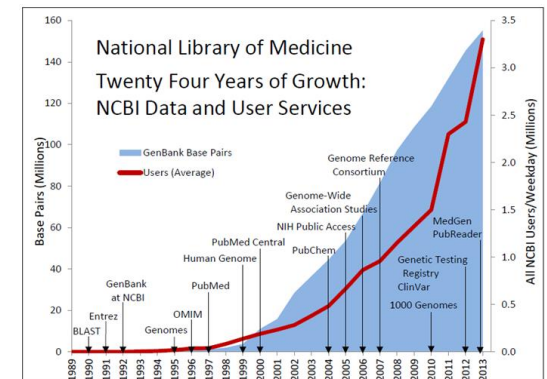
Source: NHGRI

Increasing analytical tool development



Source: Bioconductor Ann Report FY14

Growing amount of public data and data users



Source: NLM Cong. Report FY15

Democratizing data and analytical methods on a common infrastructure is essential

MeV and Genomic Data Consumers



Bioinformaticists/Data Scientists

- Start with raw data (i.e. fastq)
- Process raw data by privately tuned pipelines
- Perform secondary data analysis on self processed data
- Construct secondary analysis pipeline from software packages
- Let data drive scientific hypothesis generation

Translational Scientists

- Start with a specific hypothesis derived from observation
- Select samples/patients of interest for the hypothesis
- Find processed to perform secondary analysis
- Use readily available tools
- Interpret results in the context of initial hypothesis

Aims and Design Principles of MeV

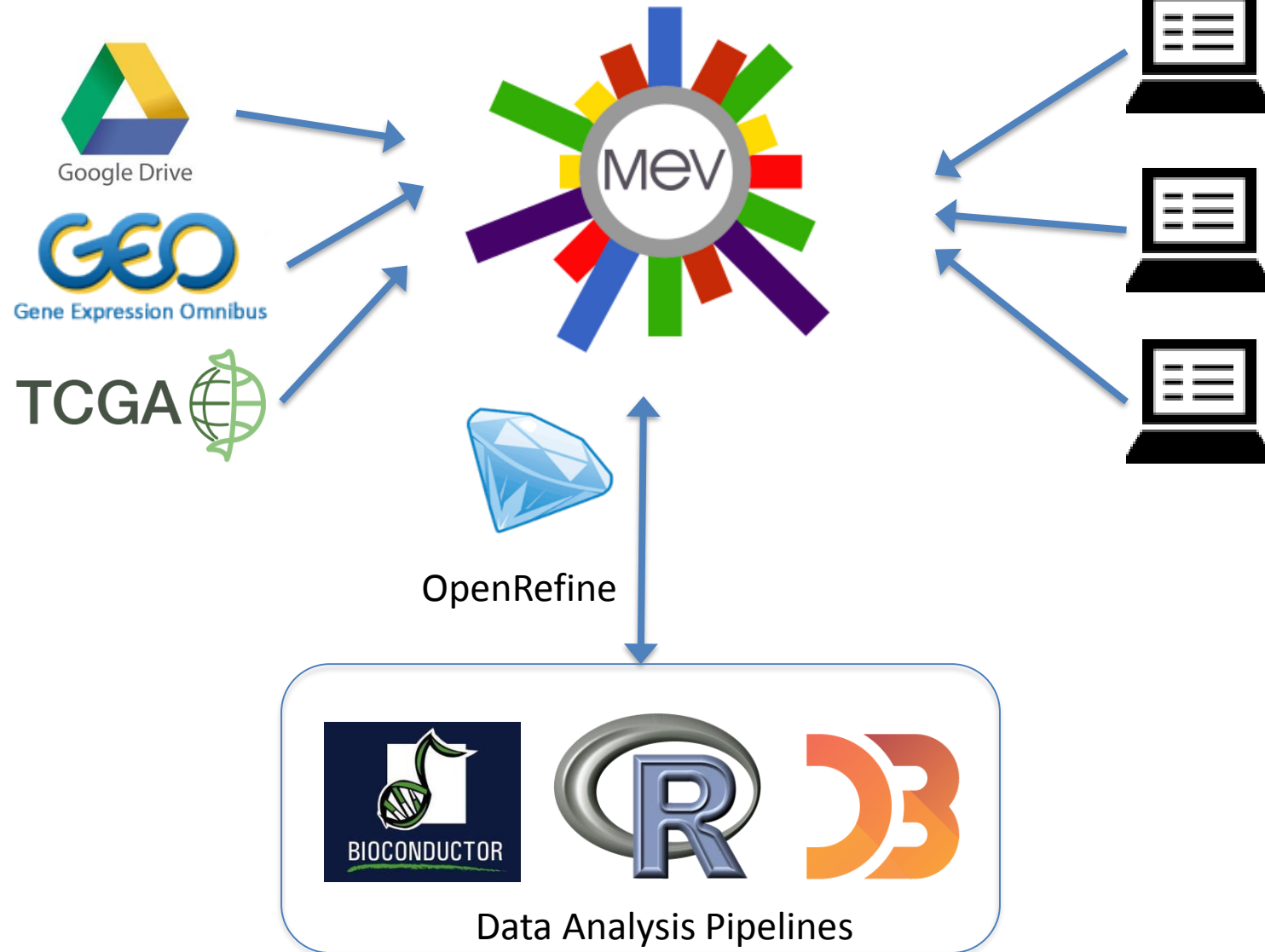
Program Aims:

- As an interface to the wide array of tools available in Bioconductor and through other open-source projects
- Natively integrate large genomic databases
- Support analysis of data emerging from Next Generation sequencing technologies, particularly RNASeq
- Adapt solely on open-source software technology

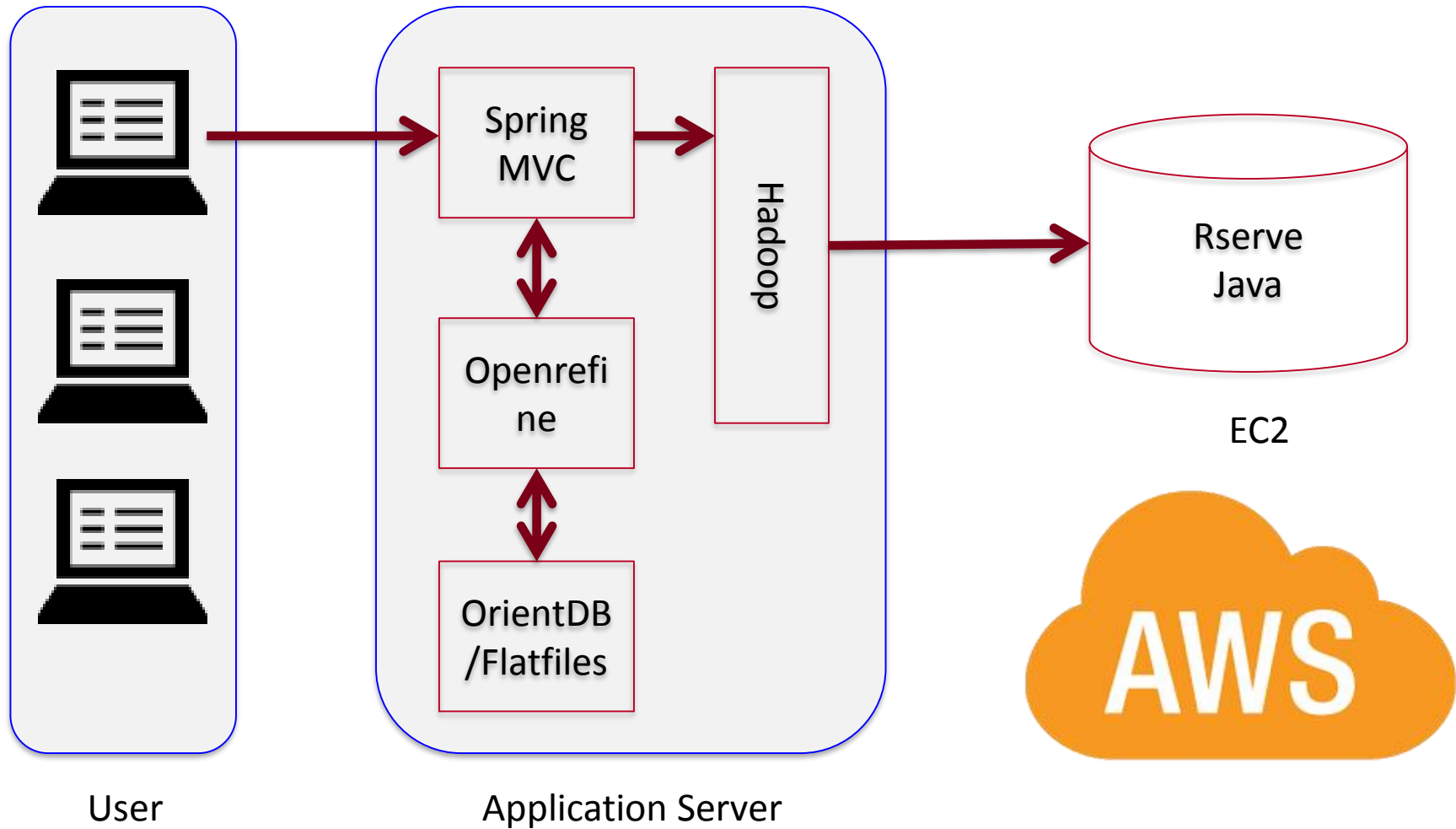
Design Principles:

- Modularized analysis and visualization design for rapid method adaptation
- Interactive result presentation to enhance user exploration
- Provide tools for cohorts stratification, grouping, and selection
- Address questions such as:
 - How my favorite genes vary in the dataset from this paper?
 - How are the phenotypes associated with the differentially expressed pathways

MeV General Workflow



WebMeV Architecture



Implement Rserve client on AWS Compute Node

Pro: Allow for quick development cycle to integrate R/Bioconductor packages

Cons: Difficult to control R versions and dependency for packages. Nightmare for distributed computing and reproducible research

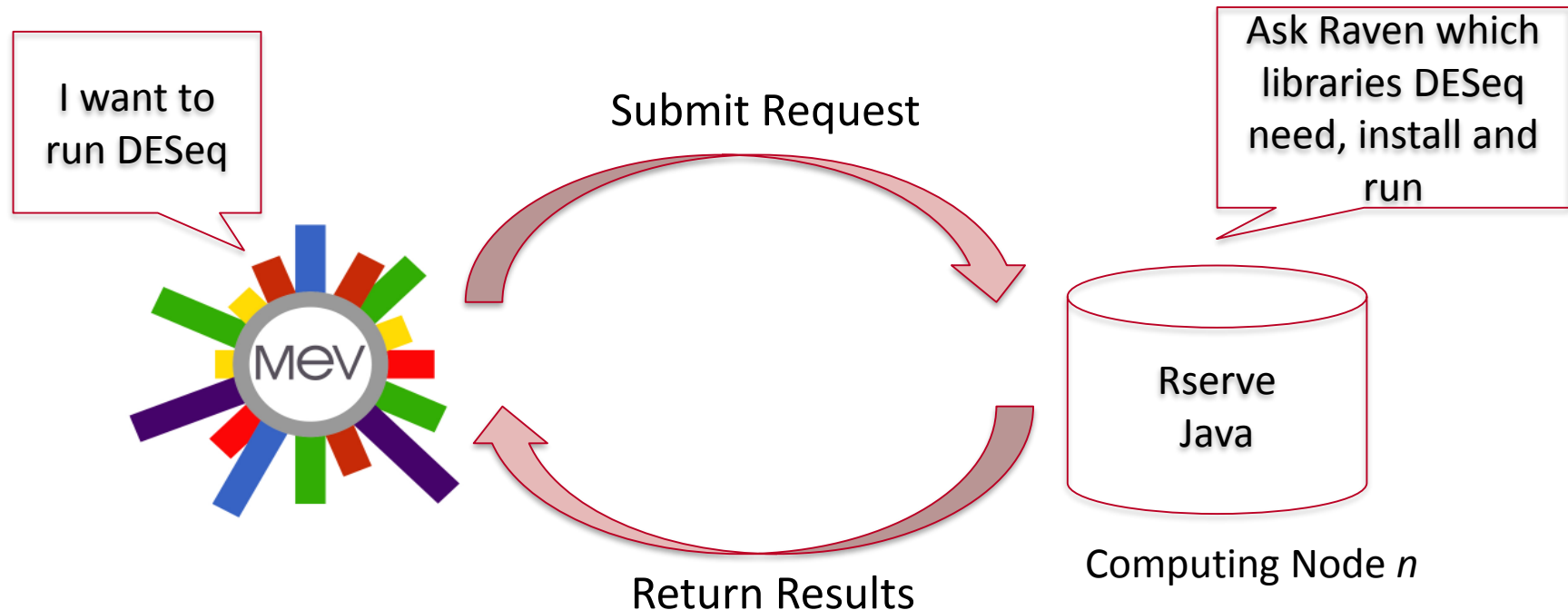
Dependency Injection

Raven: versioned snapshot repository for R, updated daily

- Available: <https://github.com/dfci-cccb/raven>

InjectoR: Dependency injection framework for R.

- Primer at <http://dfci-cccb.github.io/injectoR/>



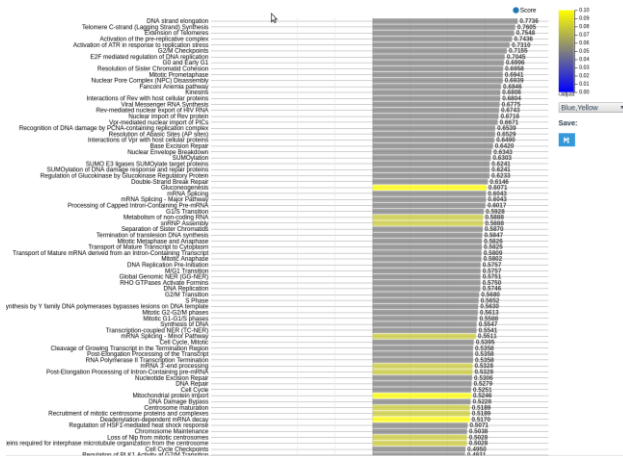
Current Analytical Methods

Analytical Method	R/Bioconductor Packages
Normalization	VST – Variance-stabilizing transformation normalization
	Upperquantile – upperquantile method for RNA-seq read count normalization
	TSS – Total Sum Scaling method for RNA-seq read count normalization
	TMM – Trimmed mean of M-values normalization method for RNA-seq read count normalization
	DESeq – Geometric mean based method for RNA-seq read count normalization as implemented in DESeq
Feature Selection	limma/voom – differential expression analyses for RNA-sequencing and microarray studies using linear model
	edgeR – differential expression analysis for RNA-seq data with normalization
	DESeq – differential expression analysis for RNA-seq data with normalization
Gene Set Approaches	topGO – testing GO terms enrichment while accounting for the topology of the GO graph
	ReactomePA – gene set and pathway enrichment analysis of data by integrating differential expression
Meta analysis	Survival – core survival analysis that performs Kaplan Meir and Cox models

Development cycle for adapting a new R method

- Work flow design
 - i.e. define input and output format
- R method incorporation
 - Takes only days, the least time consuming step
- Result Visualization
 - Mostly done using D3 to be interactive
 - Takes a few days if templates are easy to implement or already exist

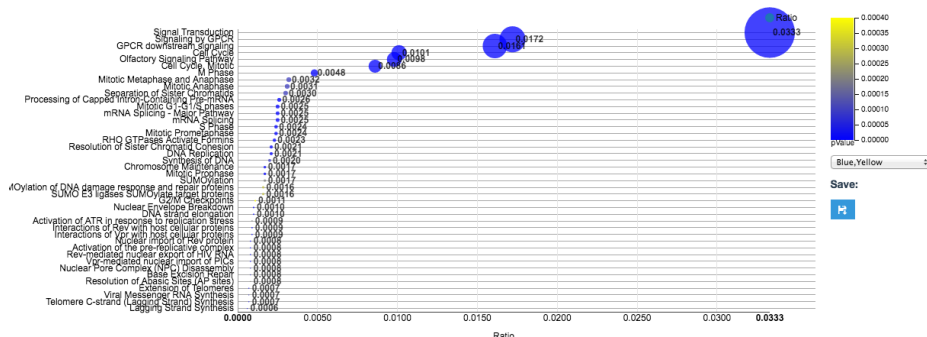
Example Visualization Outputs



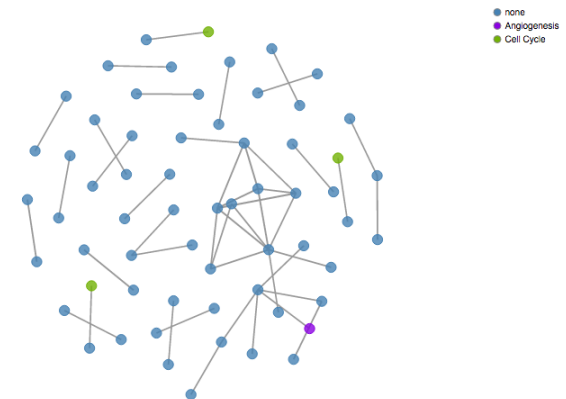
Vertical Barplot



Interactive Scatter plot



Bubble plot



Network Plot

OpenRefine for cohort selection

mev > datasets

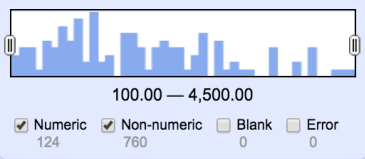
Import TCGA Datasets

BRCA clinical_annotations tsv Import Selected Samples »

Facet / Filter Undo / Redo 3

Refresh Reset All Remove All

days_to_death change reset



100.00 — 4,500.00

☒ Numeric 124 ☒ Non-numeric 760 ☐ Blank 0 ☐ Error 0

histological_type change

7 choices Sort by: name count

- [Not Available] 1
- Infiltrating Ductal Carcinoma 717
- Infiltrating Lobular Carcinoma 82
- Medullary Carcinoma 5
- Mixed Histology (please specify) 34
- Mucinous Carcinoma 6
- Other specify 39

Facet by choice counts

884 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

atic_breast_c	histological_type	her2_immunohistoc	metastatic_breast_c	metastatic_breast_c	days_to_death	pathologic_T	her2_and_cen
ple]	Infiltrating Lobular Carcinoma	2+	[Not Available]	null	[Not Applicable]	T2	[Not Available]
ple]	Infiltrating Ductal Carcinoma	0	[Not Available]	null	[Not Applicable]	T2	[Not Available]
ple]	Infiltrating Ductal Carcinoma	3+	[Not Available]	null	1141	T2b	[Not Available]
ple]	Infiltrating Ductal Carcinoma	1+	[Not Available]	null			
ple]	Medullary Carcinoma	[Not Available]	[Not Available]	null			
ple]	Infiltrating Ductal Carcinoma	[Not Available]	[Not Available]	null			
ple]	Infiltrating Ductal Carcinoma	[Not Available]	[Not Available]	null			
ple]	Infiltrating Ductal Carcinoma	0	[Not Available]	null			
ple]	Infiltrating Ductal	[Not Available]	[Not Available]	null			

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

Next Steps

- Integrate with Cancer Genomics Cloud pilot to streamline TCGA data access
- Refine clinical attribute selection interface
- Integration with VisANT and Cytoscape for network visualization and analysis methods
- Extend data access to other large public domain datasets
- Experiment with Docker container to package analysis

WebMeV Demo

<https://youtu.be/iGQbT1zCOUg>