WebMeV: A Cloud Platform for Analyzing and Visualizing Cancer Genomic Data

Yaoyu E. Wang[1], Lev Kutnetsov[1], Antony Partensky[1], Jalil Farid[1], and John Quackenbush[1,2]

Abstract

Although large, complex genomic datasets are increasingly easy to generate, and the number of publicly available datasets in cancer and other diseases is rapidly growing, the lack of intuitive, easy-to-use analysis tools has remained a barrier to the effective use of such data. WebMeV (http://mev.tm4.org) is an open-source, web-based tool that gives users access to sophisticated tools for analysis of RNA-Seq and other data in an interface designed to democratize data access. WebMeV combines cloud-based technologies with a simple user interface to allow users to access large public datasets, such as that from The Cancer Genome Atlas or to upload their own. The interface allows users to visualize data and to apply advanced data mining analysis methods to explore the data and draw biologically meaningful conclusions. We provide an overview of WebMeV and demonstrate two simple use cases that illustrate the value of putting data analysis in the hands of those looking to explore the underlying biology of the systems being studied. *Cancer Res; 77(21); e11–14. ©2017 AACR.*

Introduction

As the cost of sequencing has fallen, the number of large, publicly available genomic datasets has grown dramatically. Projects such as The Cancer Genome Atlas (TCGA; ref. 1), the Genotype-Tissue Expression Project (GTEx; ref. 2), and others have placed large, complex, multiomic data into the public domain. These datasets provide unprecedented opportunities for users to perform complex analyses and re-analyses, to make new discoveries, to search for associations among genes and phenotypes, and to test hypotheses that can then be more fully explored in future experiments. However, technical challenges, such as moving and analyzing large multiomic datasets, and the lack of intuitive and easy-to-use tools for data analysis, have limited broad exploration of the available data, often preventing experimental biologists from directly interacting with the data. WebMeV (http://mev.tm4.org) is an open-source, web-based application designed to take advantage of cloud computing resources to provide users with access to cutting-edge genomic analysis tools, direct access to large public domain datasets, intuitive user interfaces and data visualization, and tools for reproducible research.

Approach

The WebMeV cloud-based architecture is built around an application server that communicates with a computing server deployed on Amazon Web Services (AWS) to provide access to expandable computational power that takes advantage of modern cloud architectures. WebMeV was developed to provide an intuitive, user-friendly interface that allows users to upload their own data or to easily access large public datasets, such as those from TCGA and GTEx. The WebMeV interface also provides users with "click button" access to advanced analytical tools that have been released through Bioconductor (3). An overview of the system and its use to conduct basic analyses is provided in Supplementary Video S1.

Bioconductor

Bioconductor is the largest publicly available collection of open-source genomic data analysis tools. Bioconductor is based primarily on the R programming language with analysis tools distributed as R packages. Although Bioconductor receives contributions from an extensive community of bioinformatics and biostatistics developers, the majority of Bioconductor packages have only command line interfaces. Furthermore, both Bioconductor and R require appropriate installation, configuration, and version control to assure the analytic tools function and provide reproducible results. These factors present significant barriers for many potential users, including basic and translational researchers who are interested in analyzing and exploring complex genomic data, and who are often in the best position to interpret the results. So despite the considerable resource that Bioconductor represents, adoption by noncomputational scientists has been slow. WebMeV was designed to lower those barriers by placing data and appropriate Bioconductor tools into an easy-to-use interface.

The WebMeV application server uses Rserve (https://cran.r-project.org/package=Rserve) to connect through TCP/IP to an active R session running on an AWS cloud-computing instance (Fig. 1A). Every Rserve connection uses a separate workspace and

[1]Center for Cancer Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts. [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.

AACR

Figure 1.
A, An overview of WebMeV's architecture. **B** and **C,** Screenshots of WebMeV showing a simple hierarchical clustering (**B**), and a bubble plot summarizing the results of a gene set enrichment analysis (**C**).

working directory, allowing analyses to be easily distributed from the application server to multiple computing nodes as required, thus leveraging the cloud-computing infrastructure to enhance performance and deliver scalable analytics through the platform. Although WebMeV's cloud-based design was chosen to provide scalability, in reality, the number of computing nodes allocated to the application can affect performance and the number of compute nodes is limited by our funding to support the project. Because we recognize some users may want to run large dedicated analyses, we provide a WebMeV docker container image (https:// hub.docker.com/r/cccb/mev-web/) for users to download and deploy on their own hardware architecture to take advantage of the computing resources to which they have access.

A major challenge in integrating a large number of Bioconductor packages under a single application is management of package-specific dependencies and versions. To solve this problem, we implemented an inversion dependency framework (https://cran. r-project.org/web/packages/injectoR/index.html) with R version repository (https://github.com/dfci-cccb/raven) that decouples Bioconductor package dependencies from the global Rserve

environment. This framework helps WebMeV to tightly control the dependencies being used by defining clear boundaries around each Bioconductor package; this allows for the inclusion of Bioconductor packages with contradicting dependency requirement during implementation.

Data visualization

Another factor that greatly affects application efficiency is data visualization. WebMeV emphasizes bringing genomic data to users through the use of informative and interactive visualizations. WebMeV visualizations were primarily constructed using D3.js (4). D3.js is an open-source JavaScript data visualization library that conforms to SVG, HTML5, and CSS open standards and allows for seamless integration with modern browsers. It also provides software developers the flexibility to manipulate any part of the document object model to produce flexible, dynamic, and interactive genomic data visualization interfaces.

Features

Accessing public datasets

Although WebMeV allows users to upload their own data for analysis, it was also designed to provide intuitive access to and analysis of large, publicly available datasets, allowing exploration of these datasets and the development of new hypotheses. One significant advantage of cloud-based application systems is data can be housed in a central location, such that users can connect and run advanced analyses without having to download the data. This ability to "bring the analytics to the data" allows the analysis of datasets whose size would otherwise be prohibitive. WebMeV provides two routes for deploying and accessing data in the cloud. As the largest repository of gene expression data is the Gene Expression Omibus (GEO; ref. 5), we used the NCBI E-utilities application program interface (API) to access data directly from GEO and to import normalized gene expression and sample annotation data into WebMeV. We also aggregated and curated data from the TCGA, the largest public source of cancer genomic data in cancer, including both RNA sequencing (RNA-Seq) and microarray gene expression data, both of which are stored with annotation in the WebMeV AWS S3 bucket. There are other large expression datasets that users have requested we add, including GTEx (2), which will be included in a future release.

Analysis methods

Because there is a significant collection of gene expression analysis tools in Biocondcutor, the Rserve package integration into WebMeV allowed us to substantially decrease the time and effort associated with deploying new tools. Modularizing both client-end data visualization and data analysis also drastically reduces implementation time. For the initial deployment of WebMeV, we concentrated on incorporating gene expression analysis methods for both microarray and RNA-Seq data. The RNA-Seq methods include a variety of tools for normalization and detection of differentially expressed genes [LIMMA/VOOM (6), edgeR (7), DESeq (8)], for gene set enrichment analysis (topGO, ReactomePA; ref. 9), and tools for survival analysis in a framework that delivers concise, interactive graphical representations of the results (Fig. 1B and C). In addition to gene expression data analysis, the WebMeV interactive data visualization capability enables users to formulate data-driven sample groupings by exploring unsupervised clustering (hierarchical clustering, k-nearest neighbor clustering) and principle component analysis (PCA)

results. Users can interactively select sample groups for additional analysis or comparison using results such as hierarchical clustering tree branches and PCA distribution patterns.

Cohort selection

The basis for most experimental or computational analyses is the identification of distinct cohorts and the comparison of those cohorts to identify differences. WebMeV was designed to allow users to upload data and test-specific hypotheses by defining and comparing cohorts. For example, one could use the TCGA data to ask which genes are differentially expressed between HER2$^+$ and ER$^+$ breast cancer patients. What are the biological pathways that appear active in early-onset prostate cancer patients compared with those who develop the disease later? Answering these questions requires the construction of complex composite queries on high-dimensional clinical attributes to stratify patient samples into specific cohorts that can then be compared to identify significant differences.

To enable such queries on the TCGA, we implemented a cohort selector in WebMeV using OpenRefine (http://openrefine.org). OpenRefine is an open-source application for data cleaning and transformation. OpenRefine has an interface that is similar to a spreadsheet application but behaves like a database in that it allows complex queries and operations on the data. We adapted OpenRefine to support multiple clinical attribute selection methods, such as regular expression, numerical range, and categorical facets, to allow users to easily visualize attribute distribution across cohort for selection. Users can aggregate multiple selections into a composite criterion for complex cohort stratification and perform set operations on defined cohorts to reduce manual curation effort.

Example use case 1

WebMev was designed for use in a variety of different scenarios. As a first example, consider a simple experiment involving measuring gene expression using RNA-Seq to compare gene expression in three treated samples with that in a set of three control samples. We will assume that the researcher starts with the typical data produced by an RNA-Seq experiment: a tab-separated gene-level raw count matrix file (a file containing a set of genes and their RNA-Seq read counts in each sample, with tabs between column entries). The researcher can upload the file to WebMeV and first normalize across all samples. The normalized gene expression distribution can be explored using PCA, and experimental and control groups to be contrasted can be defined on the basis of the groupings visualized in the PCA plots. The researcher can also perform unsupervised hierarchical clustering to define sample groups, or simply separate samples into groups based on treatment status. The groups defined by the user can be compared to identify differentially expressed genes using a method such as VOOM (6). The differentially expressed genes can then be further filtered and saved as gene list for pathway enrichment analysis. A video showing a step-by-step walkthrough of this example is available at https://www.youtube.com/watch?v=6aK4t2vXcg4&feature=youtu.be.

Example use case 2

WebMeV also allows users to download and analyze public datasets, such as those from TCGA. Consider a researcher is interested in comparing gene expression between breast cancer patients differing in HER2, PR, and ER status using all the available breast invasive carcinoma (TCGA-BRCA) RNA-Seq data.

The researcher can launch WebMeV, select TCGA data, identify, and import the full level 3 TCGA-BRCA dataset. Data and metadata from TCGA are uploaded, and the researcher can use the text facet function in the sample set selector to stratify the cohort based on hormone receptor status. Once individual subcohorts are defined, the researcher can use the sample sets operation to construct subcohorts with composite criteria, such as patients with triple-negative status. Cohorts or subcohorts can be compared to identify differentially expressed genes and to perform a meta-analysis. The gene expression pattern of each sample group can be visualized and explored, including using PCA. A video demonstrating this example is available at https://youtu.be/W48Zhouzdr4.

Conclusions

In genomics and computational biology, there is too often a divide between those who generate the data and those who analyze it. A physician scientist treating patients with ovarian cancer will approach the analysis of ovarian cancer gene expression data with a very different mindset than a computational biologist or biostatistician and, if able to follow his instincts through hands-on analysis, may well reach different insights than if working through an intermediary quantitative scientist. WebMeV provides a resource that emphasizes making large-scale RNA-Seq and other genomic datasets accessible to a wide range of potential users. In particular, WebMeV helps to assure that basic, clinical, and translational scientists have access to datasets, such as those from the TCGA and other public projects in a tool that allows exploration of the data to develop and test hypotheses, and to extend those explorations using own data. Although we recognize the importance of having quantitative scientists, trained in computational biology, bioinformatics, or biostatistics, involved in analyzing data, scientists who understand cancer have unique intuition if followed and can lead to new discoveries.

Disclosure of Potential Conflicts of Interest

J. Quackenbush is a co-founder of Genospace, LLC, which was acquired by the Sarah Cannon Cancer Institute in January 2017. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: Y.E. Wang, L. Kutnetsov, A. Partensky, J. Farid, J. Quackenbush
Development of methodology: Y.E. Wang, L. Kutnetsov, A. Partensky, J. Farid, J. Quackenbush
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): Y.E. Wang
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Y.E. Wang, J. Quackenbush
Writing, review, and/or revision of the manuscript: Y.E. Wang, J. Quackenbush
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Y.E. Wang, L. Kutnetsov, A. Partensky, J. Farid, J. Quackenbush
Study supervision: Y.E. Wang, J. Quackenbush
Other (software development): L. Kutnetsov, A. Partensky
Other (software development and design): J. Farid

References

1. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
2. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–60.
3. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:1.
4. Cottam JA, Lumsdaine A, Wang P. Abstract rendering: out-of-core rendering for information visualization. Wong PC, Kao DL, Hao MC, Chen C, editors. IS&T/SPIE Electronic Imaging. Bellingham, WA: International Society for Optics and Photonics; 2013;9017:90170K.
5. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41:D991–5.
6. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47–7.
7. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.
8. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
9. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 2016;12:477–9.