Focus on Computer Resources

Cancer Research

A Galaxy Implementation of Next-Generation Clustered Heatmaps for Interactive Exploration of Molecular Profiling Data

Bradley M. Broom[1], Michael C. Ryan[2], Robert E. Brown[2], Futa Ikeda[2], Mark Stucky[2], David W. Kane[3], James Melott[1], Chris Wakefield[1], Tod D. Casasent[1], Rehan Akbani[1], and John N. Weinstein[1,4]

Abstract

Clustered heatmaps are the most frequently used graphics for visualization of molecular profiling data in biology. However, they are generally rendered as static, or only modestly interactive, images. We have now used recent advances in web technologies to produce interactive "next-generation" clustered heatmaps (NG-CHM) that enable extreme zooming and navigation without loss of resolution. NG-CHMs also provide link-outs to additional information sources and include other features that facilitate deep exploration of the biology behind the image. Here, we describe an implementation of the NG-CHM system in the Galaxy bioinformatics platform. We illustrate the algorithm and available computational tool using RNA-seq data from The Cancer Genome Atlas program's Kidney Clear Cell Carcinoma project. *Cancer Res; 77(21); e23–26. ©2017 AACR.*

Introduction

A heatmap is a two-dimensional visual representation of data in which numerical values of points are represented by a range of colors. Heatmaps have been used in a variety of fields from environmental science to financial analysis to geology. In the early 1990s, we introduced clustered heatmaps (CHM) into "omic" biology, initially for pharmacogenomic analysis (1), then for integrated visualization of genomic, transcriptomic, proteomic, pharmacologic, and functional data (2). They have since become the ubiquitous way to visualize patterns in molecular profiling data, for example, from microarrays and sequencing technologies (3–9). The essence of the graphic is that clustering both rows and columns brings like together with like to produce patches of color that correspond to patterns in the data. Algorithms other than clustering (e.g., rank ordering of rows and/or columns) can also be used. CHMs have appeared in many thousands of publications, but in the context of biology, they have been presented as static images or images with only modest interactive character (9).

What we wanted, in contrast, was a dynamically interactive CHM environment. Accordingly, we invoked a tiling technology to produce "next-generation" clustered heatmaps (NG-CHM) with the following interactive capabilities, among many others:

- Extreme zooming without loss of resolution for drill-down into large data matrices.
- Fluent navigation.
- Link-outs from labels or pixels to a variety of pertinent annotation resources, including GeneCards, PubMed, the Gene Ontology, Google, and cBioPortal.
- Annotation with pathway data.
- Flexible real-time recoloring.
- Capture of all metadata necessary to reproduce any chosen state of the map, even months or years later.
- High-resolution graphics that meet the requirements of all major journals.

Galaxy (10) is a widely used open-source bioinformatics platform with a web interface that enables researchers to perform many data analyses without programming. It provides an extensive collection of computational tools from many scientific domains.

Results

We have created NG-CHM builder and visualization tools for the Galaxy platform. Supplementary Video S1 demonstrates the key interactive capabilities of NG-CHMs, and Supplementary Video S2 demonstrates NG-CHMs within Galaxy. See http://bioinformatics.mdanderson.org/main/NG-CHM-V2:Overview for detailed information about the NG-CHM system, as well as download instructions, additional videos, and a detailed user guide. We have made the NG-CHM tool available as both (i) a standalone Docker image running Galaxy with the NG-CHM tool preloaded; and (ii) a separate tool that can be installed in one's own instance of Galaxy.

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [2]In Silico Solutions, Falls Church, Virginia. [3]Hobsons, Cincinnati, Ohio. [4]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Authors: Bradley M. Broom, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030. Phone: 713-792-2617; Fax: 713-563-4242; E-mail: BMBroom@mdanderson.org; and John N. Weinstein, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030. Phone: 713-563-9296; Fax: 713-563-4242; E-mail: jweinste@mdanderson.org

doi: 10.1158/0008-5472.CAN-17-0318

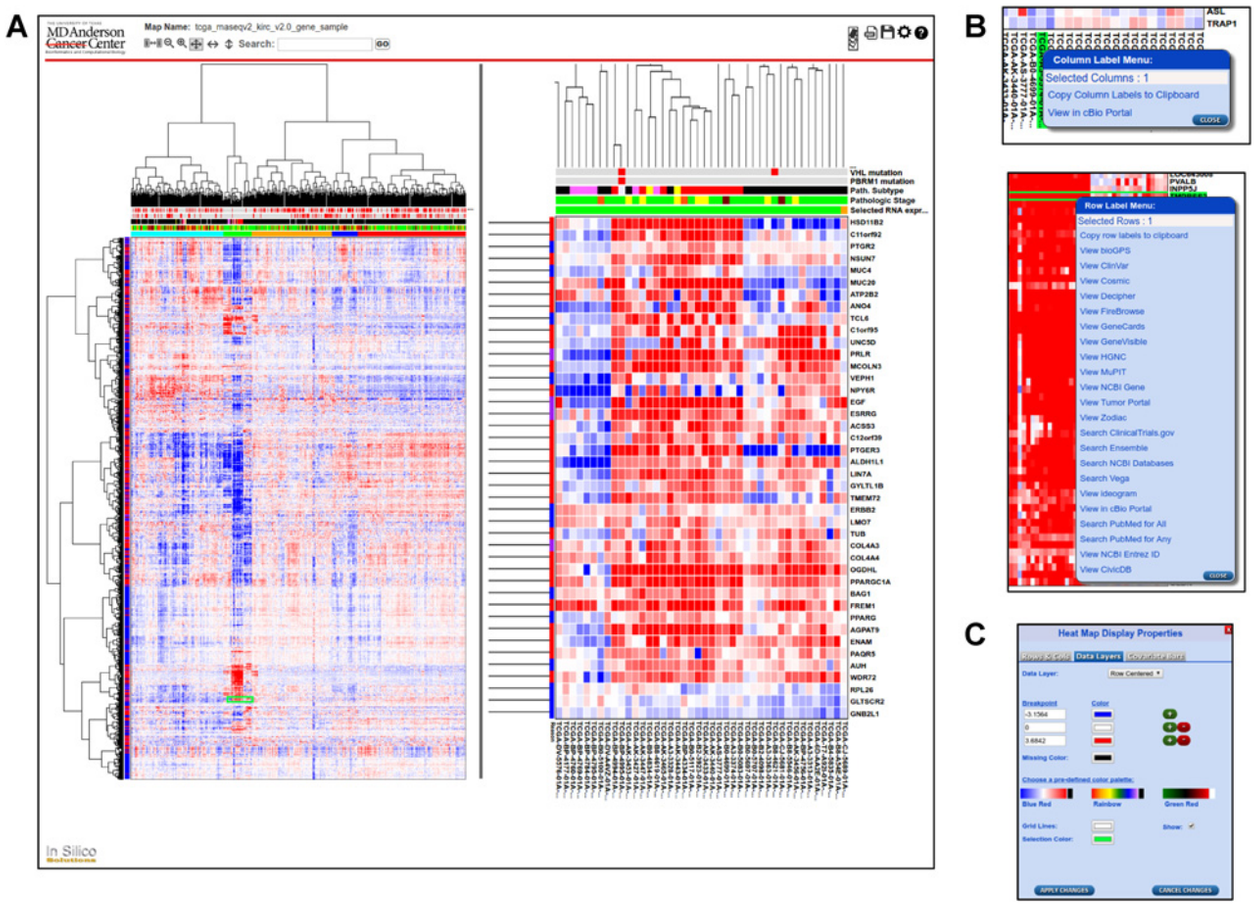

Figure 1.

Screenshots of the NG-CHM viewer. **A,** Screenshot of the viewer showing RNA-seq data from the TCGA project. Left, overview of the entire heatmap; right, zoomed view of the highlighted region. The location of the zoomed view is indicated on the left by a green rectangle. The bar at the top contains interactive control elements, including controls for zooming, changing zoom mode, searching, generating a PDF, saving a local copy, and changing map options. **B,** Context-specific menus, which are displayed when the user right-clicks on a label. Top, link-outs for TCGA sample identifiers; bottom, link-outs for gene symbols. **C,** Dialog for recoloring on the fly. The user can create a color map manually or choose a predefined color palette. The user can also choose the colors used for grid lines and the selection.

This section introduces the NG-CHM Viewer through an illustrative application to mRNA expression [RNA sequencing (RNA-seq)] data from The Cancer Genome Atlas (TCGA) Kidney Clear Cell Cancer (KIRC) project (11). Our aim is not to describe all available features or to provide a user guide (which can be found at the website), but rather to indicate the range of capabilities for exploratory research.

Figure 1A shows an NG-CHM for TCGA level 3 RNA-seq data (log-transformed) for 3,538 highly variable and/or cancer-interesting genes (rows) across 534 KIRC samples (columns). The left-hand panel shows the entire dataset. Immediately above the data matrix are a number of "covariate" bars (several for genes frequently mutated in KIRC and one each for pathologic subtype, pathologic stage, and mRNA cluster to which the sample was assigned). In general, a large number of discrete- and/or continuous-valued covariate bars can be included. Dendrograms produced by any hierarchical clustering method can be displayed or the data can be ordered in any user-defined way.

For this NG-CHM, we divided the samples into five clusters (bottom covariate bar) based on the dendrogram. The second (green) cluster shows strong enrichment for samples with chromophobe-like pathology. The chromophobes (defined histologically by tumor cells with well-defined borders, voluminous cytoplasm, and finely reticular, low-grade nuclei) constitute approximately 5% of kidney cancers.

Analogous NG-CHMs can be generated for many other types of data including miRNA expression, protein expression, DNA methylation, DNA copy number change, and mutational status. The objective is to illuminate patterns of similarity and difference in the data, then to drill down to explore the biology behind those patterns.

Overall patterns can be identified in the global view to the left in Fig. 1A. In this example, a prominent feature is the patch of highly overexpressed (red) genes that define the small (green) chromophobe cluster in molecular terms. However, individual pixels and labels are too small to be distinguished, hence the need for the zoomed-in view shown in the right-hand panel of Fig. 1A. A green rectangle in the left-hand panel shows the current position of the zoomed view.

The NG-CHM system provides a rich assortment of view controls to set the zoomed view easily and rapidly to the desired focus. Of note, Ribbon View mode enables one axis to zoom, while the other shows all rows or columns.

NG-CHMs can have multiple layers, each with its own color map, and the user can toggle rapidly among the layers. The example NG-CHM includes layers for different data normalization methods.

NG-CHMs provide context-dependent menus for selected row(s), column(s), or matrix element(s). Some menu entries (e.g., "copy to clipboard") are standard across all heatmaps; others are specific to the type of data in the map. The row labels in Fig. 1A are gene symbols, so the specific row menu entries (Fig. 1B) include links to gene-specific resources, such as GeneCards. The column labels are TCGA sample identifiers, so the column menu includes the specific entry View in cBioPortal (12). NG-CHMs based on different row or column types have different menu entries, for instance, miRNA identifiers link to miRBase.

Menu entries can also operate on groups of rows or columns. For instance, the gene-specific View Ideogram menu entry displays the locations of all selected genes on an interactive ideogram view of the genome.

For quickly locating a specific label, the search box can be used to select and highlight any matching rows and/or columns. The adjacent arrow buttons enable zooming to the search results.

The Color Scheme dialog (Fig. 1C, bottom right) enables the user to choose a predefined color scheme or create a new one on the fly.

The NG-CHM system enables the user to generate a high-resolution PDF suitable for publication in any of the major journals. The PDF can include the global view, current zoomed view, or both, as well as legends for the covariate bars.

The user can save a copy of the NG-CHM to his/her computer. Any changes made, for instance, changes to the color scheme, will be included in the saved copy. Fully interactive copies can be viewed in a standalone version of the NG-CHM Viewer and easily shared with colleagues.

In our Docker image, the Galaxy NG-CHM Builder is available in the Heat Map collection on the left of the Galaxy window (Supplementary Fig. S1). Once selected, the Builder displays an input form in which to specify the name, description, dataset(s), clustering options, and covariates to be included in the heatmap. After that information has been supplied, clicking the Execute button submits a Galaxy job to build the map, and a job summary is displayed. When the job has completed, clicking on the Visualize icon at the bottom of the job summary displays the NG-CHM in the Galaxy window. Clicking on the Download icon saves a copy to one's computer.

Discussion

NG-CHMs enable the user to visualize and explore patterns in large data matrices by fluent zooming and navigation. Context-dependent menu entries enable the user to address a wide range of questions about the data or information behind the data. We have illustrated how the NG-CHM system can be used to explore the results of a molecular profiling experiment using as an example a gene versus sample NG-CHM for KIRC gene expression data from TCGA. In the process, we explored a cluster of overexpressed genes in a small group of chromophobe-like samples.

The NG-CHM system provides a unique, fluently navigable visual environment tightly coupled to an extensive and extensible collection of context-dependent link-outs to external resources.

We demonstrated the NG-CHM system using a "standard" gene versus sample CHM. However, the system fully supports data from other domains, as well as many other forms of heatmap, for example, gene versus gene and sample versus sample NG-CHMs, maps that correlate data from different platforms, Gene Ontology maps, and maps in which one axis consists of clinical or histopathologic image features. NG-CHMs provide a wide range of interactive capabilities, including the ability to produce publication-quality images.

We have also developed a variety of tools for building and viewing NG-CHMs outside of Galaxy. For instance, we have a standalone version of the NG-CHM Viewer and an R package for building NG-CHMs.

Future enhancements to the Galaxy NG-CHM system will include a tool for modifying an existing NG-CHM. We also plan to make the content-specific menu entries easier to customize, and we plan to port additional NG-CHM features to the Galaxy environment.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: B.M. Broom, M.C. Ryan, R.E. Brown, F. Ikeda, M. Stucky,
J. Melott, C. Wakefield, T.D. Casasent, J.N. Weinstein
Development of methodology: B.M. Broom, M.C. Ryan, R.E. Brown, J. Melott, C. Wakefield, R. Akbani, J.N. Weinstein
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): B.M. Broom, T.D. Casasent
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): B.M. Broom, M. Stucky, T.D. Casasent, R. Akbani, J.N. Weinstein
Writing, review, and/or revision of the manuscript: B.M. Broom, M.C. Ryan, R.E. Brown, C. Wakefield, T.D. Casasent, R. Akbani, J.N. Weinstein
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): B.M. Broom, F. Ikeda, D.W. Kane, J. Melott, C. Wakefield, T.D. Casasent, R. Akbani, J.N. Weinstein
Study supervision: B.M. Broom, J.N. Weinstein

Acknowledgments

This research was supported in part by the Michael & Susan Dell Foundation (honoring Lorraine Dell) and by the Mary K. Chapman Foundation.

Grant Support

This work was supported by grants from "Next Generation" clustered Heat Maps for fluent, interactive exploration of omic data (1U24 CA199461 01, NIH/NCI), an Integrative Pipeline for Analysis & Translational Application of TCGA Data grant (GDAC; 3U24 CA143883, NIH/NCI), Batch Effects in Molecular Profiling Data on Cancers: Detection, Quantification, Interpretation, and Correction (1U24CA210949, NIH/NCI), Integrated Analysis of Protein Expression Data from the Reverse Phase Protein Array (RPPA) Platform (1U24CA210950, NIH/NCI), A Proteomics and Metabolomics Core Facility at the University of Texas MD Anderson Cancer Center (RP130397, CPRIT), and MD Anderson Cancer Center support grant P30 CA016672 (the Bioinformatics Shared Resource).

Received February 2, 2017; revised July 19, 2017; accepted September 19, 2017; published online November 1, 2017.

References

1. Weinstein JN, Myers T, Buolamwini J, Raghavan K, van Osdol W, Licht J, et al. Predictive statistics and artificial intelligence in the U.S. National Cancer Institute's Drug Discovery Program for Cancer and AIDS. Stem Cells 1994;12:13–22.
2. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, et al. An information-intensive approach to the molecular pharmacology of cancer. Science 1997;275:343–9.
3. Myers TG, Anderson NL, Waltham M, Li G, Buolamwini JK, Scudiero DA, et al. A protein expression database for the molecular pharmacology of cancer. Electrophoresis 1997;18:647–53.
4. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.
5. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24:227–35.
6. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236–44.
7. Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res 2003;13:703–16.
8. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, et al. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments and with application to studies of Common Variable Immune Deficiency (CVID). BMC Bioinformatics 2005;6:168.
9. Weinstein JN. A postgenomic visual icon. Science 2008;319:1772–3.
10. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 2010;11:R86.
11. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499:43–9.
12. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.

A Galaxy Implementation of Next-Generation Clustered Heatmaps for Interactive Exploration of Molecular Profiling Data

Bradley M. Broom, Michael C. Ryan, Robert E. Brown, et al.

Cancer Res 2017;77:e23-e26.