HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2015 March 01.

Published in final edited form as:

Nat Methods. 2014 September ; 11(9): 938–940. doi:10.1038/nmeth.3038.

Epiviz: interactive visual analytics for functional genomics data

Florin Chelaru[1], **Llewellyn Smith**[1,2,3], **Naomi Goldstein**[1,4], and **Héctor Corrada Bravo**[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA

[2]Department of Mathematics, Williams College, Williamstown, Massachusetts, USA

[3]Department of Computer Science, Williams College, Williamstown, Massachusetts, USA

[4]Dept. of Mechanical Engineering and Materials Science, Washington University in St. Louis, St. Louis, Missouri, USA

Abstract

Visualization is an integral aspect of genomics data analysis where the output of procedures performed in computing environments like Bioconductor is often visualized. Algorithmic-statistical analysis and interactive visualization are usually disjoint but are most effective when used iteratively. We introduce tools that provide this tight-knit integration: Epiviz (http://epiviz.cbcb.umd.edu), a web-based genome browser, and the Epivizr Bioconductor package allowing interactive, extensible and reproducible visualization within a state-of-the-art data analysis platform.

Many analyses in functional genomics including transcriptome analysis, analysis of histone modifications or transcription factor binding using ChIP-seq, and comprehensive microarray or sequencing assays to profile DNA methylation employ powerful computational and statistical tools to preprocess and model data to provide statistical inferences. Visualization of data at each step of these pipelines is essential for its exploratory analysis, characterizing the behavior of the analysis pipeline and making sense of the biological context of results by comparing to other datasets and genomic features. Interactive visualization would make these steps far more efficient, allowing the data scientist to save time, while increasing the impact of these analyses through interactive dissemination.

Advances in web application and visualization frameworks, for example d3.js[1], facilitate development of interactive data visualization tools that are easily deployed through the web. These tools have gradually moved from interactive visualization of fixed data elements to the integration of algorithmic and analytic capabilities[2], thereby accelerating how insights are derived from data. Genome and epigenome browsers are ubiquitous tools, a number of

Corresponding author, Correspondence should be addressed to H.C.B. (hcorrada@umiacs.umd.edu).

Author Contributions. H.C.B. conceived the project. F.C. and H.C.B. designed the project. F.C., L.S., N.G. and H.C.B wrote the Epiviz and Epivizr software. F.C., L.S. and H.C.B. analyzed data. H.C.B. and F.C. wrote the manuscript.

Competing Financial Interests: The authors declare no competing financial interests.

them having adopted modern web-application technologies to provide more efficient visualization and better user interfaces[3,4]. However, none of them yet support integration with computational analysis platforms like Bioconductor.[5] This limits visualization to presentation and dissemination rather than a hybrid tool integrating interactive visualization with algorithmic analysis.

We introduce Epiviz (http://epiviz.cbcb.umd.edu), a web-based genome browsing application (Fig. 1 and Supplementary Fig. 1; a short overview video here: http://youtu.be/099c4wUxozA). It tightly integrates modern visualization technologies with the R-Bioconductor data analysis platform. It implements multiple visualization methods for location-based (e.g., genomic regions of interest with block and line tracks) and feature-based (e.g., exon or transcript-level expression, with scatterplots and heatmaps) data using fundamental, well-established, interactive data visualization techniques[6], not available in web-based genome browsers. For example, since display objects are mapped directly to data elements, Epiviz implements a brushing feature that instantly gives users visual insights of the spatial correlation of multiple datasets. All data-displaying containers are resizable, colors can be mapped dynamically to display objects, and charts can be exported as static image files (pdf, svg, png or postscript).

The design of Epiviz is centered on providing tight-knit integration with computational environments like Bioconductor. The Epivizr Bioconductor package uses WebSocket connections to support two-way communication between Epiviz and interactive R sessions so data in R objects is served in response to requests made by Epiviz. This protocol was implemented over a general Data Provider interface that de-centralizes data storage allowing users to easily integrate external data sources besides R-Bioconductor (including data served by PHP-MySQL and WebSocket connections to other interactive environments like Python, Supplementary. Fig. 2). Epiviz implements a predictive caching strategy to accelerate system response to user-initiated data requests towards any of the integrated data sources

Epiviz integrates human transcriptome data from the Gene Expression Barcode project[7] (Fig. 1). By visualizing these on Epiviz users obtain immediate visual cues on transcriptome state with respect to other genomic features by using the brushing feature. All data sources catalogued by the AnnotationHub Bioconductor package are available for integration as measurements via Epivizr: UCSC genome browser[8], Ensembl[9] and BioMart[10], for instance. Epiviz also allows users to define new data measurements based on integrated measurements using a simple expression language (see Supplementary Note). Epiviz provides persistent URLs for dissemination that replicate both the underlying data, including computed measurements, and the visualization components of shared workspaces.

To facilitate exploratory data analysis, we implemented updating, filtering and subsetting operations on R objects that immediately update their visualization on Epiviz. Epivizr also supports interactive exploratory browsing: users can navigate in order through a ranked list of genomic regions of interest, for example, regions of differentially expressed genes from an RNAseq experiment obtained from packages like DESeq. Epiviz features performance optimizations that map multiple data objects to aggregate visual objects (Supplementary

Figs. 3 and 4). Using these optimizations, Epiviz displays full exon-level RNA-seq data from chromosome 11 as a scatter plot (~12,000 data points) in 150 milliseconds.

Epiviz provides a powerful and flexible extension system through Chart (for extensions with user-provided d3.js visualizations) and Data Provider (for integration of data sources) interfaces. Epivizr also provides direct support for data types defined in the Bioconductor infrastructure[11], used in many of its software packages[12,13], supporting interactive visualization directly for packages that extend its data types. Users can integrate data in SAM or BAM files in their visualizations through this infrastructure.

We illustrate the power of truly interactive visual computing using an integrative analysis of DNA methylation and exon-level expression data in colon cancer. Loss of methylation in large, gene-poor, domains associated with heterochromatin and nuclear lamina binding is an early and consistent event in colon tumorgenesis[14]. We replicate the analysis in Hansen et al.[14] for a chromosome 11 region (Fig. 1, the persistent Epiviz workspace can be accessed at http://epiviz.cbcb.umd.edu/?ws=cDx4eNK96Ws) allowing us to interactively inspect the overlap of regions of methylation loss in colon cancer and partially methylated domains (PMDs) reported in fibroblast[15]. We observed that multiple cancer types show similar expression patterns within these hypomethylation blocks (Supplementary Fig. 5) where genes are silent in normal tissues and activated in tumors. We also inferred long blocks of methylation difference in colon cancer using the minfi package[16] in Bioconductor from Illumina HumanMethylation450k beadarray data from the Cancer Genome Atlas project[17]. We used Epivizr to visually analyze the overlap of detected blocks in the TCGA samples using the 450k beadarray and the colon cancer blocks reported by Hansen et al.[14]. We found that the 450k blocks displayed high overlap with sequencing blocks (Figure 2). The method used in minfi for the 450k array ignores methylation measurements in CpG islands by design, so that long blocks of methylation change would span across CpG islands. The algorithm in Hansen et al. did not use this design, so blocks are frequently punctuated by CpG islands. Using Epivizr confirmed that the minfi procedure works as expected (Supplementary Fig. 6).

We next obtained exon-level RNAseq data from the Cancer Genome Atlas (TCGA) project[17]. RNAseq data can be referenced by genomic location (exon-level coverage, counting the number of fragments aligned to a specific exon), and by feature, e.g., transcript, or gene expression. The multi-perspective organization of Epiviz is designed for this type of analyses. We integrated this data using Epivizr and created an MA plot based on exon-level expression using the computed measurements feature on the Epiviz web application (Fig. 2). We observed the association between higher expression in cancer, now at exon-level, and hypo-methylation blocks for specific genes—the *MMP* gene family (Fig. 2). Note that the MA transformation could also be applied on the R session, demonstrating the flexibility and power of a hybrid statistical analysis environment integrated with a modern, powerful visualization tool.

We further analyzed the correlation between exon-level expression and DNA methylation. To support visualization for this analysis, we created a track-based visualization for continuous measurements of exon-level expression. We used the Epiviz Chart API to

include JavaScript files defining the new d3.js visualization hosted on GitHub Gist and are loaded into Epiviz from there. Using this we defined metadata and rendering code for the new exon-level expression visualization track. An overview visualization of the data confirmed the observation that hypo-methylated blocks are gene-poor[14] (Supplementary Fig. 7) and that exons in both normal and cancer tissues tend to be globally silenced within blocks, consistent with their association with heterochromatin, while exons outside blocks tend to be expressed (Supplementary Fig. 8).

Epiviz is the first system to provide tight integration between a state-of-the-art analytics platform and a modern, powerful, integrative visualization system for functional genomics. Infrastructure from the core Bioconductor team and hundreds of contributed packages are used in a large number of projects analyzing data that ranges from expression microarrays to next-generation sequencing. The development of interactive visualization tools based on the Bioconductor infrastructure immediately supports a number of widely used, state-of-the-art methods for a) ChIPseq where iterative visualization of data and results of peak-calling algorithms is necessary; b) RNA-seq analyes where both location-based coverage and feature-based expression levels are required; c) methylation analyses using where location-based analysis at multiple genomic scales is important. By supporting interactive visualization of fundamental data structures provided by Bioconductor, developers of new methods using this framework can immediately benefit from the powerful, extensible visualizations of Epiviz. The Galaxy platform[18] provides integration of analysis and visualization, but targets a different type of interaction as that provided by Epiviz Bioconductor defines data structures that allow direct interactive and exploratory data manipulation that is immediately reflected in the Epiviz visualization environment. Integration of analysis and visualization in Galaxy is geared toward pipeline workflows. Epiviz is an extensible platform that may incorporate Canvas-based graphics[19]in the future, while targeting integration with interactive data environments beyond extensibility capabilities available in current browsers[4].

Online Methods

Annotation Data

We obtained annotation data from the UCSC genome browser for hg19. PMDs were obtained from Lister et al., generated from bisulfite sequencing in fibroblast cells[15]. DNAm data and hypomethylation block regions were obtained from Hansen et al.[14]. Affymetrix hgu133plus2 expression data was obtained from the Gene Expression Barcode project[7].

Illumina HumanMethylation450k beadarray data

IDAT files for 17 normal colon and 34 colon tumor samples were obtained from the TCGA project[17]. All processing was performed using the minfi Bioconductor package. Data was preprocessed and normalized using the standard Illumina method, hypomethylation block finding was performed using the method in minfi.

RNA-seq data

Raw count tables at the exon level were obtained for 3 normal colon and 37 colon tumor samples from the TCGA project[17]. Counts were normalized for library size using the DESeq method[12]. Exon annotation using UCSC ids were included by the TCGA project.

Software

Analyses were performed using Bioconductor packages minfi (v. 1.8.9) and epivizr (v. 1.3.3). The Epiviz web application is hosted at http://epiviz.cbcb.umd.edu, the Epivizr Bioconductor package is available through the Bioconductor project. JavaScript files defining exon-level expression visualization tracks are available as Github Gists (http://gist.github.com/11279474 and http://gist.github.com/11279449). Open source code for all components is available in the Epiviz project github page: http://github.com/epiviz, API descriptions and other documentation for Epiviz is available online at http://epiviz.cbcb.umd.edu/help.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Bioconductor core team and the Bioinformatics and Computational Biology Department of Genentech Research and Early Development for helpful suggestions and comments. This work was partially supported by NIH grants R01 HG006102 to H.C.B. and F.C., R01 HG005220-03 to H.C.B, an undergraduate internship sponsorship from the Illumina Corporation to L.S and support from Genentech.

References

1. Bostock M, Ogievetsky V, Heer J. IEEE Trans. Visual. Comput. Graphics. 2011; 17:2301–2309.
2. Stolte C, Tang D, Hanrahan P. Commun. ACM. 2008; 51:75–84.
3. Lister R, et al. Cell. 2008; 133:523–536. [PubMed: 18423832]
4. Zhou X, et al. Nature Methods. 2011; 8:989–990. [PubMed: 22127213]
5. Gentleman RC, et al. Genome Biol. 2004; 5:R80. [PubMed: 15461798]
6. Yi JS, Kang YA, Stasko J, Jacko J. IEEE Trans. Visual. Comput. Graphics. 2007; 13:1224–1231.
7. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA. Nucleic Acids Res. 2011; 39:D1011–D1015. [PubMed: 21177656]
8. Karolchik D, et al. Nucleic Acids Res. 2008; 36:D773–D779. [PubMed: 18086701]
9. Hubbard TJP, et al. Nucleic Acids Res. 2009; 37:D690–D697. [PubMed: 19033362]
10. Durinck S, et al. Bioinformatics. 2005; 21:3439–3440. [PubMed: 16082012]
11. Lawrence M, et al. PLoS Comput Biol. 2013; 9:e1003118. [PubMed: 23950696]
12. Anders S, Huber W. Genome Biol. 2010; 11:R106. [PubMed: 20979621]
13. Paulson JN, Stine OC, Bravo HC, Pop M. Nature Methods. 2013; 10:1200–1202. [PubMed: 24076764]
14. Hansen KD, et al. Nat Genet. 2011; 43:768–775. [PubMed: 21706001]
15. Lister R, et al. Nature. 2009; 462:315–322. [PubMed: 19829295]
16. Aryee MJ, et al. Bioinformatics. 2014; 30:1363–1369. [PubMed: 24478339]
17. Cancer Genome Atlas Network. Nature. 2012; 487:330–337. [PubMed: 22810696]
18. Goecks J, et al. BMC Genomics. 2013; 14:397. [PubMed: 23758618]

19. Miller CA, Anthony J, Meyer MM, Marth G. Bioinformatics. 2013; 29:381–383. [PubMed: 23172864]

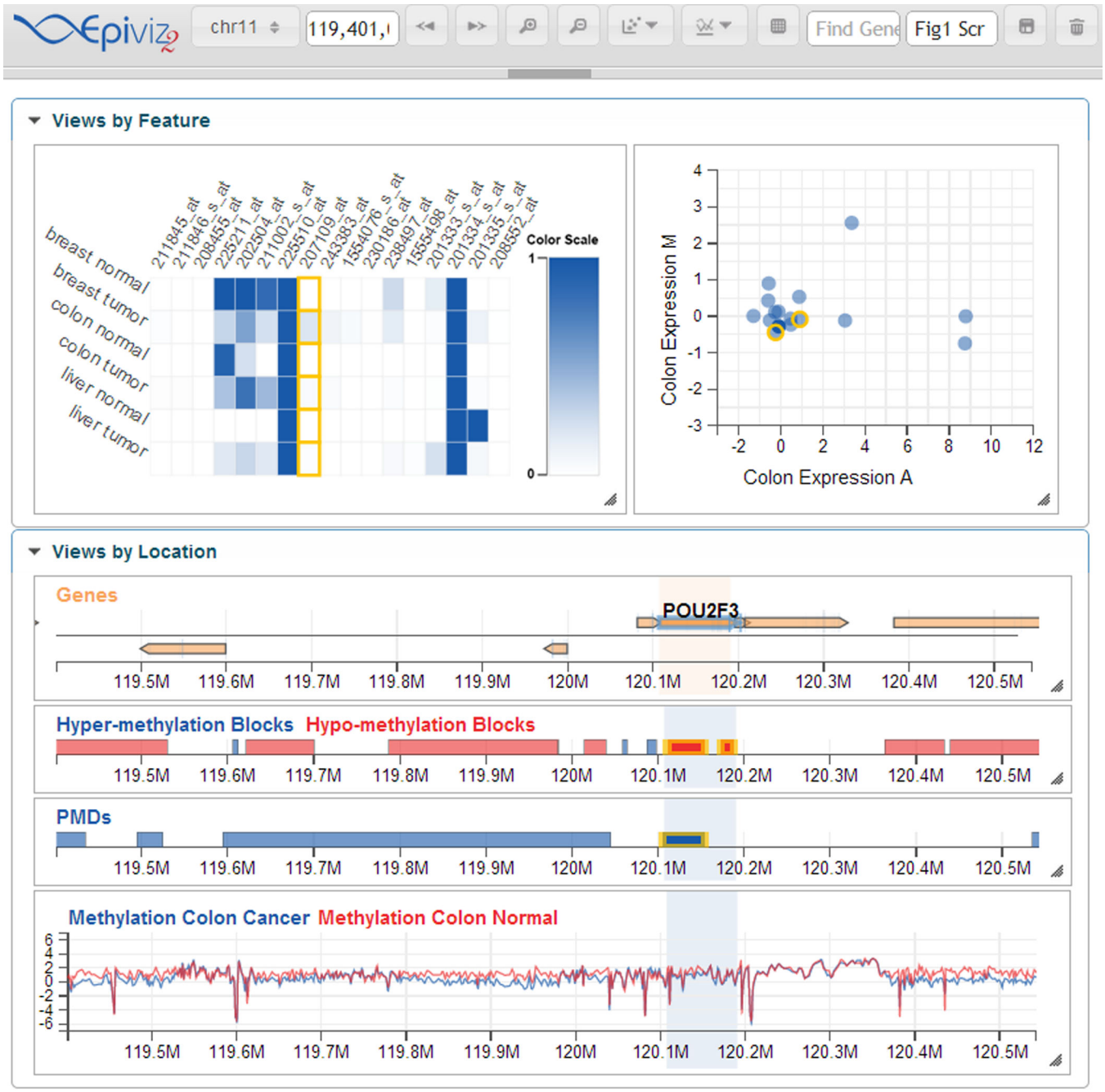

Figure 1.

Colon cancer methylome visualization using Epiviz. Long regions of methylation changes in colon cancer (Hypo- and Hyper-methylation blocks) are shown along with the smoothed base-pair resolution data (Methylation Colon Cancer and Normal) used to define them. Colon gene expression data on an MA plot (top right) shows genes within the viewing region that are differentially expressed. Data from the gene expression barcode shows transcriptome state across multiple tissues (top left). Highlighted region shows the brushing

feature linking all charts by spatial location. This workspace can be accessed at http://epiviz.cbcb.umd.edu/?ws=cDx4eNK96Ws.

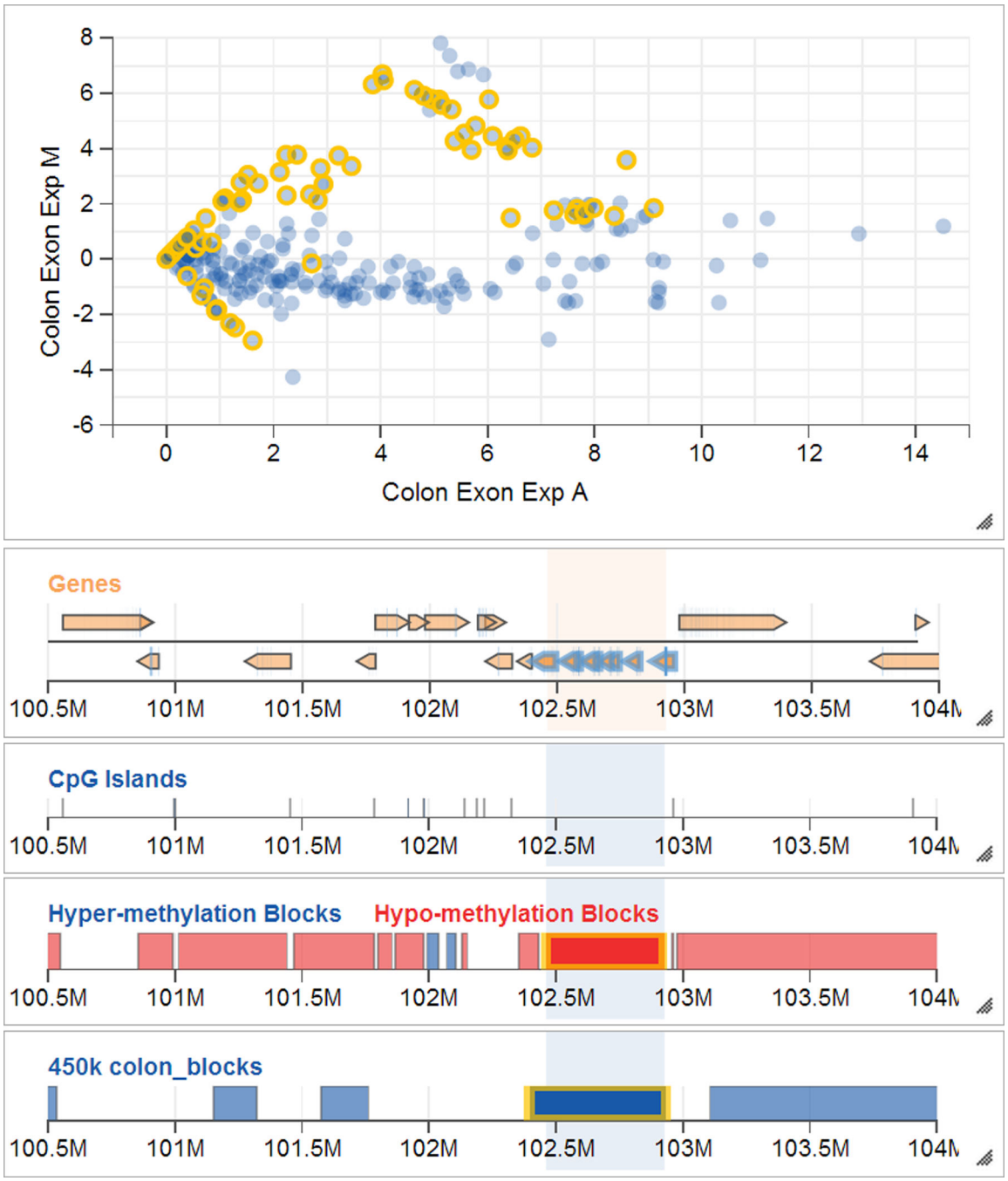

Figure 2.

Integrative analysis of Illumina HumanMethylation450k data and exon-level RNAseq data using Epivizr. Regions of hypomethylation blocks obtained from TCGA data using the 450k array (bottom track) shown along with regions obtained from sequencing data (Hansen et al.) on independent samples. An MA plot (top) of exon-level RNA-seq data from the TCGA project over the same region (the MA transformation was obtained using the computed measurements tool in the Epiviz UI).