Bioinformatics, 36(8), 2020, 2581–2583
doi: 10.1093/bioinformatics/btz972
Advance Access Publication Date: 3 January 2020
Applications Note

Genome analysis

Differential Expression Gene Explorer (DrEdGE): a tool for generating interactive online visualizations of gene expression datasets

Sophia C. Tintori [1,†], **Patrick Golden** [2] **and Bob Goldstein** [1,*]

[1]Department of Biology and [2]School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

*To whom correspondence should be addressed.
[†]Present address: Department of Biology, Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA

Associate Editor: Jonathan Wren

Received on June 12, 2019; revised on December 26, 2019; editorial decision on December 26, 2019; accepted on January 3, 2020

Abstract

Summary: Differential Expression Gene Explorer (DrEdGE) is a web-based tool that guides genomicists through easily creating interactive online data visualizations, which colleagues can query according to their own conditions to discover genes, samples or patterns of interest. We demonstrate DrEdGE's features with three example websites generated from publicly available datasets—human neuronal tissue, mouse embryonic tissue and *Caenorhabditis elegans* whole embryos. DrEdGE increases the utility of large genomics datasets by removing technical obstacles to independent exploration.

Availability and implementation: Freely available at http://dredge.bio.unc.edu.

Contact: bobg@email.unc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Data sharing, in the interest of transparency, openness and reproducibility, is becoming increasingly embraced by the scientific community (Gewin, 2016; Nosek *et al.*, 2015). In burgeoning big data fields such as genomics, sharing also enables data mining, which can lead to new discoveries. Currently, the primary method for sharing data is to publish raw data on databases (e.g. GEO) (Edgar *et al.*, 2002). While such access to raw data is critical, it is also critical to share partially processed data in an interactive format. This allows collaborators and colleagues to explore data without starting from the rawest form, which can be considerably, often prohibitively, time consuming. Without such tools, the data may be *available*, but they are not effectively *accessible* to most researchers, especially non-genomicists.

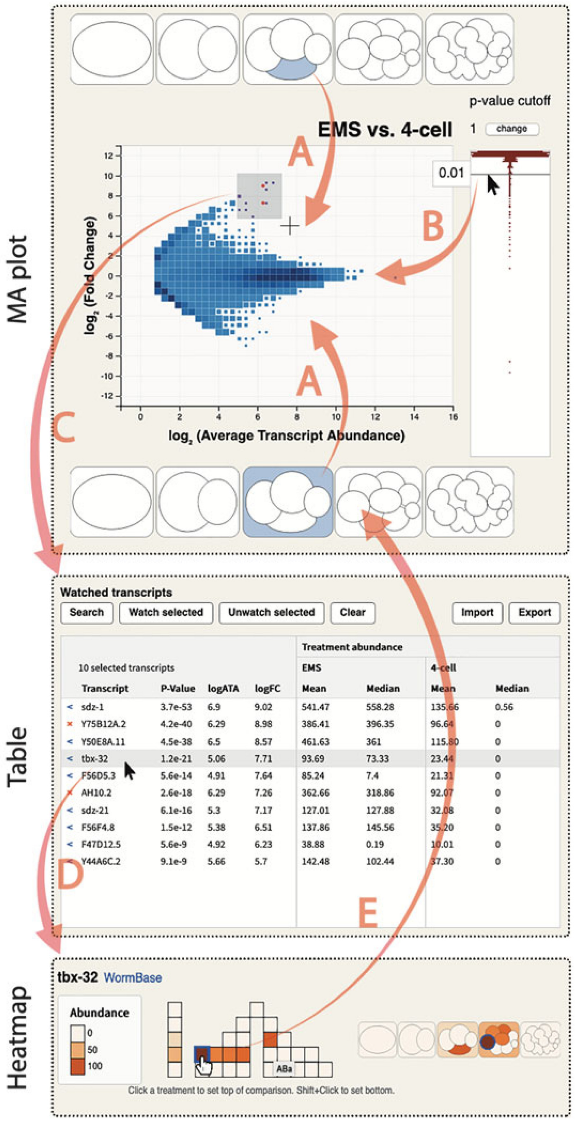
Genome browsers represent one such type of interactive tool displaying partially processed data (Kent *et al.*, 2002; Spudich *et al.*, 2007). Genome browsers allow users to visually identify patterns, but they cannot generate statistical results, nor easily address biological questions that span more than one locus. As the rate at which large, multivalent datasets are generated increases, the need for tools that make these datasets accessible and minable continues to grow.

To address this need, we have created the Differential Expression Gene Explorer (DrEdGE), a web-based interactive data visualization tool. DrEdGE allows genomics researchers to share datasets in an accessible, queryable format, which colleagues can then use to identify genes or samples of interest. DrEdGE's flexible design can be used to visualize any type of quantitative unit (transcripts, DNA fragments, proteins, etc.) using any statistical model for differential expression. The user interacts with three data representations—a differential expression plot, a statistical table and an experiment-wide heatmap—which each provide input into the other representations, creating an iterative workflow that can be cycled through repeatedly for continuous fine tuning or elaboration of hypotheses (Fig. 1). Videos demonstrating the features and configuration of a DrEdGE website are available at http://vimeo.com/dredge.

2 Results

2.1 Primary features of a DrEdGE visualization

On a DrEdGE website, gene expression is defined by two variables: (i) units of genetic material (e.g. transcripts, genome fragments, proteins—for simplicity we will call these units 'transcripts') and (ii) sets of experimental replicates (e.g. tissue types, chemical treatments, stages in a time course—we will call these units 'treatments') within an experiment. Three elements make up the visualization.



Watched transcripts

10 selected transcripts

				Treatment abundance			
				EMS		4-cell	
Transcript	P-Value	logATA	logFC	Mean	Median	Mean	Median
< sdz-1	3.7e-53	6.9	9.02	541.47	558.28	135.66	0.56
× Y75B12A.2	4.2e-40	6.29	8.98	386.41	396.35	96.64	0
< Y50E8A.11	4.5e-38	6.5	8.57	461.63	361	115.80	0
< tbx-32	1.2e-21	5.06	7.71	93.69	73.33	23.44	0
< F56D5.3	5.6e-14	4.91	7.64	85.24	7.4	21.31	0
× AH10.2	2.6e-18	6.29	7.26	362.66	318.86	92.07	0
< sdz-21	6.1e-16	5.3	7.17	127.01	127.88	32.08	0
< F56F4.8	1.5e-12	5.38	6.51	137.86	145.56	35.20	0
< F47D12.5	5.6e-9	4.92	6.23	38.88	0.19	10.01	0
< Y44A6C.2	9.1e-9	5.66	5.7	142.48	102.44	37.30	0

tbx-32 WormBase

Fig. 1. Interactivity between elements of DrEdGE visualization. (**A**) The user selects two treatments to compare in an MA plot, which shows differential expression of transcripts. (**B**) A slider filters out transcripts by *P*-value. (**C**) The user selects transcripts from the MA plot to populate a table of statistical values. Curated tables can be saved, added to, pruned or cleared, through multiple analyses. (**D**) Transcripts selected in the table will populate a heatmap showing relative transcript abundance across all treatments. (**E**) Treatments selected from the heatmap can be used to set the parameters for the MA plot. Sample data are from Tintori *et al.* (2016), and are available at http://dredge.bio.unc.edu/c-elegans-transcriptional-lineage

2.1.1 MA plot

The user begins an investigation by selecting two treatments to compare. The MA plot (log ratio *M* by mean average *A*) becomes populated based on this selection, describing the ratio of transcripts' expression levels between the two treatments (y-axis) and their average abundance (x-axis). Transcripts selected from this plot will populate the data table.

2.1.2 Data table

The table displays (and can be sorted by) transcript name, ratio of transcript abundance between two treatments, *P*-value and average transcript abundance. The user can curate a list of transcripts by adding them to a 'watched' list, either by selecting from the plot, typing into the search bar or uploading a list. Adding or removing transcripts from the table during sequential pairwise comparisons allows the user to curate a list that satisfies multiple criteria.

2.1.3 Heatmap

As the user hovers over transcripts in the data table, the heatmap diagram is populated, showing the transcript's abundance in all treatments of the dataset. The reader can click on any treatment in the heatmap to generate a new MA plot, displaying new data about the transcripts.

2.2 Example datasets

To demonstrate DrEdGE's utility, we have generated three sample websites using human neuronal tissue, mouse embryonic tissue and whole *Caenorhabditis elegans* embryo datasets (Boeck *et al.*, 2016; The ENCODE Project Consortium, 2012). These can be accessed from the DrEdGE homepage. Details about the source data, including all files used for DrEdGE configuration, can be found in Supplementary Materials. Although these examples use publicly available datasets, we strongly encourage the lab generating original data to create a DrEdGE website themselves. This will ensure that the website is created with the most thorough understanding of the data and the statistics that suit the dataset.

2.3 Requirements for configuration

DrEdGE websites can be configured quickly (<20 min) and require three elements as input: a transcript count matrix, a metadata table and a folder of differential expression tables. We provide a generic script for generating the third element (differential expression tables) from the first two, but users are welcome to upload tables generated by any preferred method. Details on how to configure a DrEdGE page are provided on the DrEdGE website, accompanying instructional videos and Supplementary Figures S1 and S2.

The size of the DrEdGE directory is primarily based on the number of transcripts (t) assayed and the number of treatments (n) in the experiment, and can be estimated by the equation: $(t) \times (n^2) \times 15$ bytes. Implementation details and comparisons between DrEdGE and other existing data sharing software can be found in Supplementary Materials.

3 Conclusion

As the genomic era ushers in larger and larger datasets, there is a growing need for tools that facilitate mining and collaboration. Here, we have presented DrEdGE, a program that allows genomicists to easily create and publish interactive data visualization websites. On a DrEdGE website, users can compare differential transcript abundance between samples, survey experiment-wide transcript abundance patterns and filter or sort results by specific statistics.

The potential for authors to use statistical methods of their choice allows DrEdGE to (i) consist solely of static files, making it easier to host and maintain than a typical dynamic Web application and (ii) fit within diverse data analysis pipelines. Each experiment requires unique statistical considerations, whether due to idiosyncrasies of the genome, molecular technique or philosophy of the researcher in a rapidly changing field. DrEdGE is able to accommodate this.

DrEdGE helps fill the gap in data sharing practices between raw data on databases (that require a substantial amount of time and expertise to process) and completed analyses in publications (that are few, static and cannot be queried further). By removing the technical obstacles that prevent exploration by interested parties, DrEdGE increases the utility and impact of large genomics datasets.

Acknowledgements

The authors thank Dan McKay, Jonathan Seguin, Max Boeck, David King, Ann Wehman, Kira Heikes and our editor and reviewers for their thoughtful feedback on versions of the manuscript.

Funding

This work was supported by grants to B.G. from the National Institutes of Health [R01 GM083071, R35 GM134838]; the National Science Foundation [IOS 1557432]; and to S.C.T. from the Damon Runyon Cancer Research Foundation [19-1211].

Conflict of Interest: none declared.

References

Boeck,M.E. *et al.* (2016) The time-resolved transcriptome of *C. elegans*. *Genome Res.*, **26**, 1441–1450.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Gewin,V. (2016) Data sharing: an open mind on open data. *Nature*, **529**, 117–119.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Nosek,B.A. *et al.* (2015) Promoting an open research culture. *Science*, **348**, 1422–1425.

Spudich,G. *et al.* (2007) Genome browsing with Ensembl: a practical overview. *Brief. Funct. Genomic. Proteomic.*, **6**, 202–219.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Tintori,S.C. *et al.* (2016) A transcriptional lineage of the early *C. elegans* embryo. *Dev. Cell*, **38**, 430–444.