bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Software

Differential Expression Gene Explorer (DrEdGE): A tool for generating interactive online data visualizations for exploration of quantitative transcript abundance datasets

Sophia C. Tintori(1), Patrick Golden(2), Bob Goldstein(1)*

(1) Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
(2) School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
* Correspondence: bobg@unc.edu

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

As the scientific community becomes increasingly interested in and committed to data sharing, there remains a need for tools that facilitate the querying of public data. Mining of RNA-seq datasets, for example, has value to many biomedical researchers, yet such datasets are often effectively inaccessible to non-genomicists, even when the raw data are available. Here we present DrEdGE (dredge.bio.unc.edu), a free tool that allows any researcher to 'dredge' genomics datasets for genes or samples of interest, according to their own conditions. We demonstrate DrEdGE's utility with three examples—human neuronal tissue, mouse embryonic tissue, and a *C. elegans* embryonic series.

Keywords

interactive, data visualization, data sharing, data mining, data access, high throughput sequencing, next generation sequencing, RNA-seq, differential expression

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Background

Data sharing, in the interest of transparency, openness, and reproducibility, is rapidly becoming a standard embraced by the scientific community[1-6]. The practice of data sharing allows colleagues to independently reproduce and verify analyses, and to build future research upon them more reliably. In burgeoning big data fields such as genomics, data sharing enables data mining, allowing others to perform their own analyses on large, multivalent datasets that may have only been used for a small fraction of possible applications at the time of initial publication[7-9].

Currently, the primary method for sharing data is by publishing raw data on databases such as the NCBI Gene Expression Omnibus, the NCBI Sequence Read Archive, DNA Data Bank of Japan, European Nucleotide Archive, Figshare, and Dryad[10-15]. These raw data repositories are critical, as they allow independent analyses to be performed starting from entirely unprocessed material. Also critical are tools for sharing partially processed data in an interactive format. Such tools allow researchers to explore the data without having to reprocess them starting from the rawest form, which can be considerably, often prohibitively, time consuming. Interactive tools for exploring partially processed data can facilitate communication between genomicists and non-genomicist collaborators or colleagues. Without such tools, the data are available but not effectively accessible to most non-genomicist researchers[16,17].

Genome browsers (such as those of UCSC[18], Ensembl[19], and NCBI[20]) represent one such type of interactive tool that allows researchers to explore partially processed data. Users can browse sequencing read density at each nucleotide in the genome from a given experiment, alongside annotation of genomic features such as introns and exons, GC content, repeats, and conservation scores. While this tool is critically useful for visually identifying patterns in the data, the user cannot generate statistical results, and must create his or her own system of organization if he or she wishes to ask biological questions that span more than one locus.

As sequencing costs continue to drop[21] and multiplexing technology becomes more sophisticated[22-24], the rate at which large multivalent datasets are generated is increasing[25]. In turn, the need for tools that make them not only available, but also easily accessible to interested colleagues, continues to grow.

To address this need, we have created the Differential Expression Gene Explorer (DrEdGE, dredge.bio.unc.edu), a web-based interactive data visualization tool that facilitates the exploration of genomics datasets, and the identification of genes or samples of interest. DrEdGE can be used to visualize any number of sample types, and any type of quantitative unit (transcripts, DNA fragments, proteins, etc.). The user can interact with three different data representations on a DrEdGE website—a plot, a table, and a heat map—which each feed into the other representations, creating an iterative workflow that can be cycled through repeatedly for the continuous fine tuning or elaboration of hypotheses. For researchers using our program to create a DrEdGE website for presenting their dataset, the DrEdGE software provides flexibility regarding the type of experiments represented and the statistical methods used. DrEdGE allows genomics researchers to share their datasets for others to access and query, either during collaborations or after publication.

Results

GENERAL PROPERTIES

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

DrEdGE is designed to connect two groups of researchers: those who create, process, and publish genomic datasets, and those interested in asking biological questions of genomic data (who may or may not have specialized computational genomics skills). DrEdGE bridges these two groups by presenting data from the former in a way that makes the data accessible to the latter. In this manuscript we distinguish between these two types of users. The first user—the author—makes an interactive DrEdGE website by importing their dataset. The author can publish this website alongside their data and analyses in a peer-reviewed journal, or he or she can share it privately with colleagues. The second user—the reader—explores the DrEdGE website, querying the data with his or her own biological questions.

We will first describe the DrEdGE Web application, as used by the reader for exploration and analysis of biological data. Then, we will describe a few sample datasets used to make sample DrEdGE websites. Finally, we will describe the process by which the author prepares and uploads a new dataset to create his or her own novel DrEdGE website.

PRIMARY FEATURES OF DREDGE VISUALIZATION

The DrEdGE website displays gene expression data based on two variables: (1) units of genetic material (i.e. cDNA, genome fragments, proteins—for simplicity, and because we showcase RNA-seq experiments in demonstrations below, we will call these units "transcripts"), and (2) sets of experimental replicates (i.e. tissue types, chemical treatments, stages in a time course—we will call these units "treatments") within an experiment.

Three visual elements make up the visualization: (1) a plot displaying relative abundance of transcripts between two treatments, (2) a data table listing transcript abundance measurements and statistics, and (3) a heat map displaying the relative abundances of a transcript across every treatment in the experiment. Results from each data representation can be used as input for another representation, allowing the user to build upon their hypotheses in an iterative fashion.

MA plot - The reader begins his or her investigation by selecting two different treatments to compare (Figure 1A). Treatments are selected from either a vector graphic representing the dataset, or a simple dropdown menu in which all treatments are listed. The two treatments selected will then populate the rest of the visualization, starting with the MA (log ratio *M* by mean average *A*) plot, which describes differential abundance (in RPKM or otherwise normalized values) between the two treatments for all transcripts (Figure 1B). These plots are commonly used to visualize how abundant a transcript is and how specifically it is enriched in one treatment over another. Each square on this plot represents one or more transcripts, with the size and darkness of each square scaled to the number of transcripts it represents. Transcripts in the top half are more abundant in the treatment selected above the plot, and those on the bottom half are more abundant in the treatment selected below. Transcripts to the right of the plot have a greater average abundance than those to the left. The reader selects transcripts in an area of the MA plot by brushing (clicking and dragging) with the mouse, and the transcripts selected in this way then populate further elements of the visualization.

For example, if the reader is exploring mouse tissue transcriptome data (available at http://dredge.bio.unc.edu/mouse-embryonic-tissue/, and described in further detail below) and wishes to know what forebrain transcripts are more abundant at embryonic day 15.5 than two days earlier, he or she selects “forebrain || embryonic day 15.5” above the MA plot and

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

"forebrain || embryonic day 13.5" below the plot. Points at the top of the plot represent transcripts with higher abundance on day 15.5 than day 13.5, and overall abundance is highest in points furthest to the right within the plot. The reader can filter out transcripts by P-value of differential abundance by sliding the p-value cutoff selector. The reader can use the mouse to highlight transcripts in a region of the plot for closer examination or for tracking (or "watching") through new treatment comparisons. Watched transcripts can be added or removed via buttons above the data table, which allows the reader to curate a list of transcripts that fit multiple criteria spanning several pairwise comparisons.

Data table – Transcripts highlighted on the MA plot appear in the table to the right (Figure 1C). This table displays the fold change values between the two treatments selected, P-values for fold change, and transcript abundance values for each treatment. For example, if the reader has selected a swath of transcripts from the region of the MA plot representing higher abundance at embryonic day 15.5 than embryonic day 13.5, each of those selected transcripts will populate the table. The reader can then sort the transcripts by name, average abundance, ratio of abundances between the two treatments, or P-value. Abundance levels for each treatment are available both as mean and median, which allows some insight into the distribution of data across replicates.

The reader can curate a list of transcripts of note by adding them to a "watched" list in the data table. These watched transcripts will persist in the table, highlighted by red dots, when no area is brushed and when new treatments are selected for comparison. To add transcripts one-by-one to the watched list, the reader can either click the arrow next to a selected transcript name on the table, or enter transcripts manually in a search bar. To add groups of transcripts to the watched list, the user can upload a comma delineated file with one transcript name on each line, or the user can click "watch selected," which will add all transcripts currently brushed over on the plot. At any point the reader can export their current list of watched transcripts.

As the user hovers over or clicks rows in the data table, the visualization updates to display information about the transcript that row represents. First, a red circle appears in the MA plot around that transcript's position. Second, a heat map diagram is populated under the table, showing the relative abundance of the transcript for all treatments within the dataset.

Heat map – The heat map (Figure 1D) graphically represents abundances of a single transcript in all experimental treatments. This feature allows a reader to examine the experiment-wide context for a transcript that was selected based on just a pairwise comparison. If the heat map reveals that another treatment besides the two selected for the MA plot is of interest, the reader can click on that treatment's icon to replace the top or bottom treatment in the MA plot.

In the hypothetical example mentioned above, the user may click on a transcript in the table and generate an experiment-wide heat map. The heat map may reveal that the transcript is also highly abundant in the liver at embryonic day 14.5. Curious to see what other forebrain-enriched transcripts also become enriched in the liver at this time, the user might repopulate the MA plot by clicking on the "liver || embryonic day 14.5" icon for the top of the plot, and the "liver || embryonic day 12.5" icon for the bottom of the plot. This allows the reader to see which of the already-highlighted transcripts are also enriched in the liver at embryonic day 14.5 when compared to two days earlier. Heat maps maybe be shown as a grid of boxes (as in Figure 1E), as illustrated icons (as in Figure 1F), or both.

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Iterative analysis - Each of the three data representations—the MA plot, the table, and the heat map—can be further investigated in one of the other representations: A transcript with an interesting differential abundance in the MA plot can be selected to populate the table, a transcript with interesting statistical values in the table can be selected to generate an experiment-wide heat map, and treatments with an interesting transcript abundance in the context of the whole experiment can be selected to generate a new MA plot. Curated lists of transcripts of interest can be generated, added to, and pruned continuously, using the buttons shown in Figure 1G. The export and import functions (Figure 1H) allow analyses to be continued across datasets and experiments, if transcript names remain consistent across experiments.

EXAMPLE DATASETS FROM HUMAN, MOUSE, AND WORM

To demonstrate the utility of DrEdGE we have generated three sample DrEdGE websites using data mined from published human, mouse, and worm RNA-seq studies. These examples showcase a range of options for heat map graphics (Figure 2A-E), each designed to suit the different sizes and biological features of the dataset. All of the required input files for these websites are provided in the supplementary materials, and can be referenced as templates for researchers generating their own DrEdGE websites of similar design, if desired.

Although we have generated these examples from mined data, we strongly encourage the lab generating and analyzing original data to create a DrEdGE website themselves. This will ensure that the website is created with the most thorough understanding of the idiosyncrasies of the data, and the statistics they require.

Forty-two whole embryo C. elegans developmental time points- Total RNA sequencing data from whole *C. elegans* embryos were mined from Boeck et al, 2016[26] to create a DrEdGE website, available at dredge.bio.unc.edu/c-elegans-timecourse. In this study, RNA abundance was reported for forty-two time points from 28 to 737 minutes after the two-cell stage. These 42 time points were generated from three independent time courses composed of 9, 12, and 21 timepoints each. Development progressed at varying speeds in each of the three time courses, and so the researchers used a Bayesian statistical model to expand or contract the timeline of each time course to match the others. Each of the final 42 reconstructed timepoints is defined by two replicates, with each replicate used in two consecutive time points. For any given time point, one of the two replicates is shared with the previous timepoint, and the other replicate is shared with the following timepoint.

Because these samples were numerous and have a linear relationship to each other, we graphically represented them in the heat map using box icons in a custom arrangement, as shown in Figure 2B.

Nine neuronal tissues from humans- Human total RNA sequencing data were mined from the ENCODE consortium[4]. We selected nine neuronal tissues—camera-type eye, cerebellum, diencephalon, frontal cortex, occipital lobe, parietal lobe, spinal cord, temporal lobe, and tibial nerve—to generate a DrEdGE website, which is available at dredge.bio.unc.edu/human-neuronal-tissue. All libraries were originally prepared and sequenced by the Gingeras lab for the ENCODE consortium. Each treatment (tissue type, in this case) consisted of 4-8 replicates in total from 2-4 individuals between the ages of 19 weeks and 54 years. All tissues were sampled from an equal representation of males and females, with the exception of the temporal lobe,

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

cerebellum, and camera-type eye tissues, which were all taken exclusively from embryonic females. All samples were released as part of the ENCODE project on June 30th, 2014, with the exception of the tibial nerve samples, which were released on February 18, 2016.

Because the number of treatments used was relatively small, and because the treatments have a spatial and anatomical relationship to each other, we created custom SVG icons to illustrate each tissue type in the heat map data representation, as shown in Figure 2C.

Forty embryonic tissues from mice- Mouse polyadenylated RNA sequencing data were mined from the ENCODE consortium[4] to create a DrEdGE website, which is available at dredge.bio.unc.edu/mouse-embryonic-tissue. Sixteen tissue or cell types were selected: embryonic facial prominence, forebrain, heart, hindbrain, intestine, kidney, limb, liver, midbrain, neural tube, skeletal muscle tissue, spleen, stomach, C2C12 cells, C3H10T1/2 cells, and C3H myocyte cells from C2C12. Each tissue or cell type was sampled on a subset of the following eight developmental timepoints: embryonic day 10.5, 11.5, 12.5, 13.5, 14.5, 15.5, 16.5, and postnatal day 0. Forty "treatments" (or sample types) in total were used, with 2-6 replicates per tissue and 1-2 replicates per cell culture, which were released between May 7, 2012 and August 11, 2016. All samples were prepared and sequenced by the Wold lab for the ENCODE consortium.

Because these samples are numerous we used the default box icons, rather than illustrations, to represent them in the heat map. Because the samples can each be defined by two factors—time point and tissue type—we organized the icons in a custom 2D grid, with each row describing a developmental stage and each column describing a tissue type, as shown in Figure 2D. In some instances, it may be appropriate to show treatments defined in space *and* time with illustrative icons, as in the example from Tintori et al 2016 shown Figure 2E[27].

USAGE STEPS FOR GENERATING A DREDGE WEBSITE

We designed the DrEdGE software such that no coding is necessary to create a DrEdGE website, if the author is starting from normalized transcript abundance data. If starting from sequence files (.fasta or .fastq), the author will have to align reads to the genome and process the alignments him- or herself to generate normalized transcript count files.

Input files - The DrEdGE software builds a DrEdGE website from five required components, and four optional components, as shown in Figure 3 (and described in full detail in Supplementary File 2): (1) a table of transcript abundance counts, (2) a table describing the experimental design, (3) a directory of pairwise comparison tables, (4) minimum and maximum values from the comparison tables within that directory, and (5) a JSON file describing the experimental design. The author can provide only the first two components and use R scripts from the DrEdGE package to generate the last three components. Alternatively, if the author has specific statistical methods they prefer to use, the author can generate all five components him- or herself, so long as the final formats are correct.

There are four optional components the DrEdGE software can additionally accept. These include (6) a table of transcript synonyms, or alternate names, (7) a base URL for searching for transcripts on an external organism-specific database, (8) a custom grid scheme for organizing treatments in a 2D heat map grid, and (9) custom SVG icons representing treatments in an illustrated heat map.

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Processing input files to generate a DrEdGE website – To process the 5-9 components described above into a DrEdGE website, the author must download the DrEdGE package from dredge.bio.unc.edu, unzip the package, and host this DrEdGE directory locally. The index.html file within that directory will present instructions for how to upload each of these components, perform a test run, and save the project once assembly is complete. Saving the project will create a new project.json file, which must be added to the DrEdGE directory. The author may then upload the directory to the Web to share their DrEdGE website privately with collaborators, or publicly as part of a publication.

Storage space required to host DrEdGE website – The final size of the DrEdGE directory will depend primarily on the pairwise comparison tables. If the number of transcripts evaluated (t) is known, and the number of treatments (n) in the experiment is known, the size of the DrEdGE directory can be estimated with the following formula: Size = (t) * (n^2) * 15 bytes. The three samples presented in this publication range in size from 40 to 500 MB.

Discussion

As the genomic era ushers in larger and larger datasets, data sharing is becoming a standard practice that (1) helps ensure reproducibility, and (2) enables data mining. Processing raw data into a format that can be mined takes a considerable, and often prohibitive, investment of time and specialized skills, raising the barrier to data mining. This barrier can be bypassed, and data mining promoted, if data are published in a partially processed, interactive format. Few tools are currently available for genomicists to share their data with others. Here, we have presented a program that creates interactive data visualization websites where researchers can explore genomics datasets by comparing transcript abundances between samples, surveying experiment-wide transcript abundance patterns, and filtering or sorting by specific statistics. This tool allows genomics researchers to share their work with other interested parties, either during collaboration or during data mining after publication.

For simplicity, we have limited our demonstration of DrEdGE's utility to RNA-seq experiments. DrEdGE's flexible design, however, allows users to visualize any two-dimensional numeric dataset, including comparative proteomics, population diversity maps, or even non-biological datasets.

Data accessibility- The decreasing cost of genetic sequencing has led to large sequencing projects in which dozens or hundreds of samples are sequenced in a single experiment. The resulting datasets contain many more potential insights than can possibly be described in a single report, and are therefore useful for data mining by other groups. The DrEdGE visualization tool presented here makes such data available and accessible for interested parties to query for insights relevant to their own work.

Statistical flexibility- Each sequencing experiment has unique qualities, whether due to idiosyncrasies of the genome being studied or the technique being used, that must be taken into account during analysis. For this reason, we decided not to build a statistical engine into DrEdGE. We provide an R script that processes a transcript abundance table into pairwise comparison tables using edgeR (statistical details in Supplementary File 2 and in Robinson et al.,

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2010[28]), but we invite authors to provide their own pairwise comparison values using their preferred statistical method if preferred. This flexibility extends DrEdGE's utility to authors with specific statistical requirements or philosophies, as well as to future applications, when unforeseeable practices for analyzing genomic data may become standard.

Iterative analysis- DrEdGE's iterative analysis design allows readers to select data points of interest from each data representation to investigate further in the next representation. The reader can curate a list of transcripts through multiple analyses. This allows the reader to ask sophisticated biological questions of the data, and generate statistically supported results. This differs from current interactive genomics data visualization websites, such as genome browsers, because comparative biological questions can be asked, tested, refined, and tested again continuously. Exporting curated lists of transcripts from one experiment and importing it to another experiment also allows for integrated analyses across different datasets or types of assays.

Low technical overhead- The potential for users to use statistical methods of their choice has two benefits. First, it means that DrEdGE consists solely of static files, making it easier to host and maintain than a typical dynamic Web application powered by a database and server-side program (written in e.g. PHP, Python, Perl)[29]. Second, it allows DrEdGE to fit within diverse data analysis pipelines. Different groups may have different methods for calculating statistical analyses, which DrEdGE is able to accommodate.

Conclusions

We have presented DrEdGE, a tool for generating interactive visualizations of large genomics datasets. These visualizations allow interested parties to independently explore data by generating their own MA plots, transcript abundance statistics tables, and experiment-wide heat maps, and to modify lists of transcripts of interest through multiple analyses. This tool fills the gap in data sharing practices between raw data on databases (that require a substantial amount of time and expertise to process) and a handful of completed analyses in publications (that are static and cannot be queried further). Our hope is that this will increase the utility of sequencing studies by removing technical obstacles that prevent interested parties from exploring multivalent datasets.

Methods

Implementation and availability: DrEdGE is a browser-based application powered by client-side JavaScript. The entire application is bundled into one HTML file and one JavaScript file, using the JavaScript compiler Browserify[30] and the automation tool GNU Make[31]. All of the pages in the application are built and rendered using the React library[32]. Interactive data visualizations are created using the D3.js library[33]. All code is freely available on GitHub (github.com/ptgolden/dredge) under the GNU Affero General Public License (AGPLv3).

Sourcing of sample datasets- RNA-seq data for humans, mice, and worms were collected from ENCODE[4] or Boeck et al., 2016[26]. Human experiment accession numbers: ENCSR000AFO, ENCSR000AEW, ENCSR000AEX, ENCSR000AEY, ENCSR000AFD, ENCSR000AFE,

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ENCSR000AFJ, ENCSR858QEL, ENCSR648OSR, ENCSR796HLX, ENCSR272UNO, ENCSR000AFH. Mouse experiment accession numbers: ENCSR809VYL, ENCSR851HEC, ENCSR823VEE, ENCSR636CWO, ENCSR160IIN, ENCSR970EWM, ENCSR752RGN, ENCSR691OPQ, ENCSR284YKY, ENCSR727FHP, ENCSR597UZW, ENCSR526SEX, ENCSR420QTO, ENCSR848GST, ENCSR537GNQ, ENCSR173PJN, ENCSR347SQR, ENCSR830IVQ, ENCSR648YEP, ENCSR448MXQ, ENCSR867YNV, ENCSR826HIQ, ENCSR096STK, ENCSR457RRW, ENCSR992WBR, ENCSR307BCA, ENCSR908JWT, ENCSR343YLB, ENCSR557RMA, ENCSR719NAJ, ENCSR337FYI, ENCSR115TWD, ENCSR667TOX, ENCSR946HWC, ENCSR579FCW, ENCSR290RRR, ENCSR178GUS, ENCSR000AHY, ENCSR000AHZ, ENCSR000AHX, ENCSR000AIA. Counts were compiled into transcript count abundance tables and experimental design tables in R. Transcript synonym tables for humans, mice, and worms were generated from the HUGO Gene Nomenclature Committee's Biomart[34,35], Mouse Genome Informatics[36], and Wormmine[37], respectively.

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure Legends

Figure 1: Layout of DrEdGE visualization. A) The user can select two treatments to compare, using icons or a drop column. B) An MA plot shows differential expression between two treatments. C) A table shows statistical values of transcripts highlighted from the MA plot. D) A heat map shows relative transcript abundance across all treatments for a transcript selected from the table, graphically represented as E) a grid of square icon, F) an illustration, or both. G) A curated list of transcripts can be saved, added to, pruned, or cleared, through multiple analyses. H) Lists of transcripts can be imported or exported to save, share with colleagues, or interface with other methods and analyses in the user's analytical pipeline.

Figure 2: Graphical options for heat map. A) Default box icons in a default compact grid. B) Box icons in a custom linear arrangement. C) Box icons custom-arranged in a grid. D) SVG illustrated icons. E) SVG illustrated icons (from Tintori et al. 2016)[27].

Figure 3: Pipeline of data input and products.

Supplementary File 1: Materials used to generate the three example DrEdGE websites

Supplementary File 2: Descriptions of each input file for generating a DrEdGE website

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Author Contributions

Project was conceived of by ST. Designed by ST and PG with input from BG. Javascript code written by PG, R code written by ST. Data mining and processing by ST. Manuscript written by ST with input from PG and BG.

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015,348(6242):1422-5.
2. Gewin V. Data sharing: An open mind on open data. Nature. 2016;529:117–119.
3. Bioarxiv. http://www.bioarxiv.org/ (2013).
4. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012,489(7414):57-74.
5. The Cancer Genome Atlas. http://cancergenome.nih.gov/ (2006).
6. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015,526;68–74.
7. Grabowski P, Rappsilber J. A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight? Trends in Biochemical Sciences. 2018,1521.
8. Huttenhower C, Hofmann O. A Quick Guide to Large-Scale Genomic Data Mining. PLoS Comput Biol. 2010,6(5):e1000779.
9. Cheng PF, Dummer R, Levesque MP. Data mining The Cancer Genome Atlas in the era of precision cancer medicine. Swiss Medicine Weekly. 2015;145:w14183.
10. Edgar R, Domrachev M, Lash AE. NCBI GEO - Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002,30(1):207-10.
11. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. Nucleic Acids Research. 2011,39: D19–D21.
12. Toribio AL, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. European Nucleotide Archive in 2016. Nucleic Acids Research. 2017,45;D32–D36.
13. Figshare. https://figshare.com/ (2011).
14. Dryad. http://datadryad.org/ (2008).
15. Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T. DNA data bank of Japan (DDBJ) progress report. Nucleic Acids Research. 2016,44;D51–D57.
16. Markowetz F. All biology is computational biology. PLoS Biol. 2017,15(3):e2002050.
17. Longo DL, Drazen JM. Data Sharing. New England Journal of Medicine. 2016,374;3.
18. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Research. 2002,12(6):996-1006.
19. Spudich G, Fernández-Suárez XM, Birney E. Genome Browsing with Ensembl: a practical overview. Brief Funct Genomic Proteomic. 2007,6: 202-219.
20. NCBI Genome Data Viewer. https://www.ncbi.nlm.nih.gov/genome/gdv/.
21. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). 2018. www.genome.gov/sequencingcostsdata. Accessed December 6, 2018.
22. Wong KH, Jin Y, Moqtaderi Z. Multiplex Illumina Sequencing Using DNA Barcoding. Current Protocols in Molecular Biology. 2013,7.11.1-7.11.11.
23. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161,1187–1201.

bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

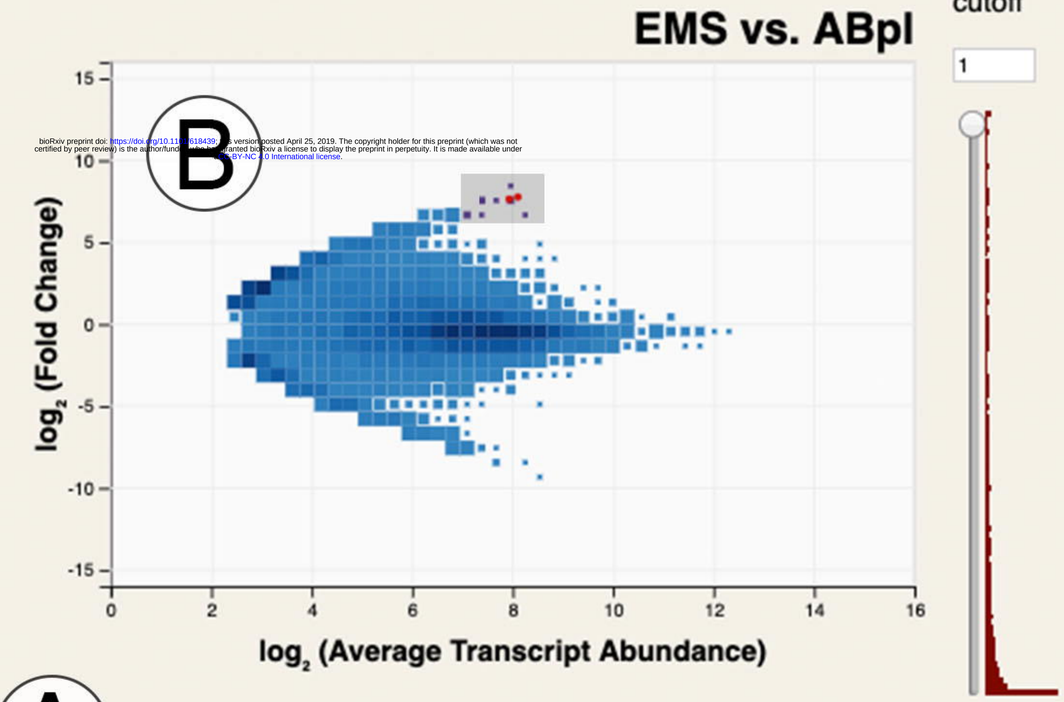
24. Hashimshony T, Senderovich N, Avita G, Klochendler A, de Leeuw Y, Anavy1 L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, Dor Y, Regev A, Yanai I. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biology. 2016,17:77.
25. Karsch-Mizrachi I, Takagi T, Cochrane G on behalf of the International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. Nucleic Acids Research. 2018,46(D1):D48-D51.
26. Boeck ME, Huynh C, Gevirtzman L, Thompson OA, Wang G et al. The time-resolved transcriptome of C. elegans. Genome Research. 2016,26:1441–1450.
27. Tintori SC, Osborne Nishimura E, Golden P, Lieb JD, Goldstein B. A transcriptional lineage of the early *C. elegans* embryo. Developmental Cell. 2016;38(4):430-44.
28. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
29. Hedstrom M. Digital preservation: A time bomb for digital libraries. Computers and the Humanities. 1998;31:189-202.
30. Browserify. http://browserify.org/ (2011).
31. GNU Make. https://www.gnu.org/software/make/.
32. React. https://reactjs.org/.
33. D3. https://d3js.org/ (2010).
34. Yates B, Braschi B, Gray K, Seal R, Tweedie S, Bruford E. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017;45(D1):D619-625.
35. HUGO Gene Nomenclature Committee at the European Bioinformatics Institute. HUGO Mart. https://biomart.genenames.org/martform/#!/default/HGNC?datasets=hgnc_gene_mart. Accessed August 21, 2018.
36. MGI Data and Statistical Reports. Mouse Genome Informatics. 1996. http://www.informatics.jax.org/downloads/reports/index.html. Accessed August 15, 2018.
37. Wormmine. Wormbase. http://intermine.wormbase.org/tools/wormmine/begin.do. Accessed August 10, 2015.

Transcriptional lineage with gastrulation stage

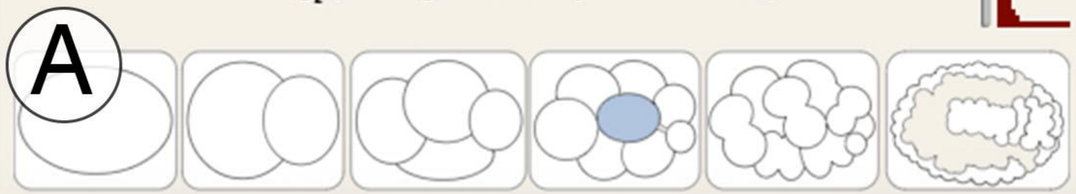
Open dataset | Menu

p-value cutoff

1



bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.



Watched transcripts

Search | Watch selected | Unwatch selected | Clear | Import | Export

G H

C

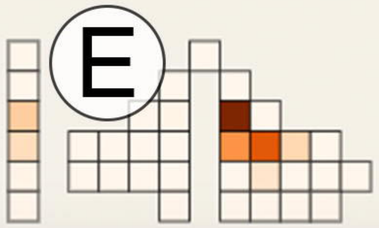
11 selected transcripts

Transcript	P-Value	logATA	logFC ▲	Treatment abundance: EMS Mean	EMS Median	ABpl Mean	ABpl Median
Y75B12A.2	0.000	7.99	8.18	386.41	396.35	0.00	0.00
spsb-2	0.000	8.11	7.76	413.31	460.15	0.49	0.50
swd-2.2	0.000	7.81	7.65	331.54	310.36	0.29	0.00
AH10.2	0.000	7.93	7.63	362.66	318.86	0.44	0.00
F58E6.6	0.000	7.45	7.62	264.17	251.60	0.00	0.00
szy-4	0.000	7.50	7.51	271.05	293.45	0.13	0.00
T08A9.6	0.000	7.07	7.22	210.26	90.75	0.00	0.00
unc-101	0.000	7.04	6.81	211.13	108.05	0.34	0.00
pup-2	0.000	7.36	6.44	241.99	274.72	1.28	0.89
dkf-2	0.000	7.00	6.37	179.38	178.51	0.81	0.00
try-1	0.000	8.22	6.35	438.92	371.90	3.43	0.54

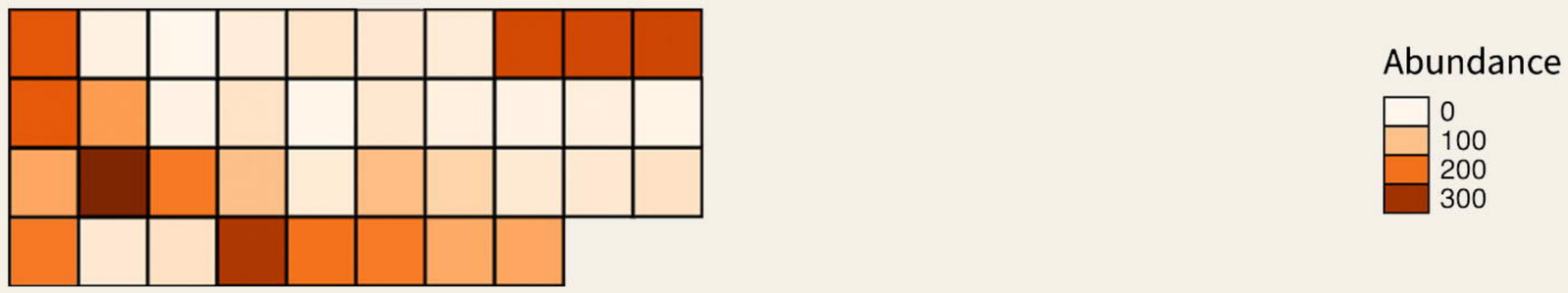
AH10.2

D

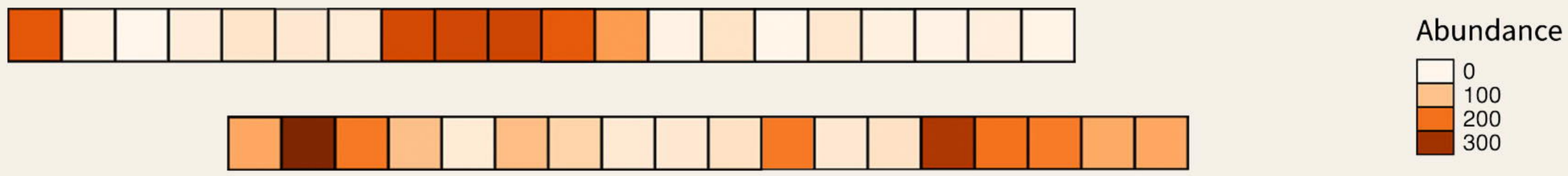
Abundance
0
100
200
300



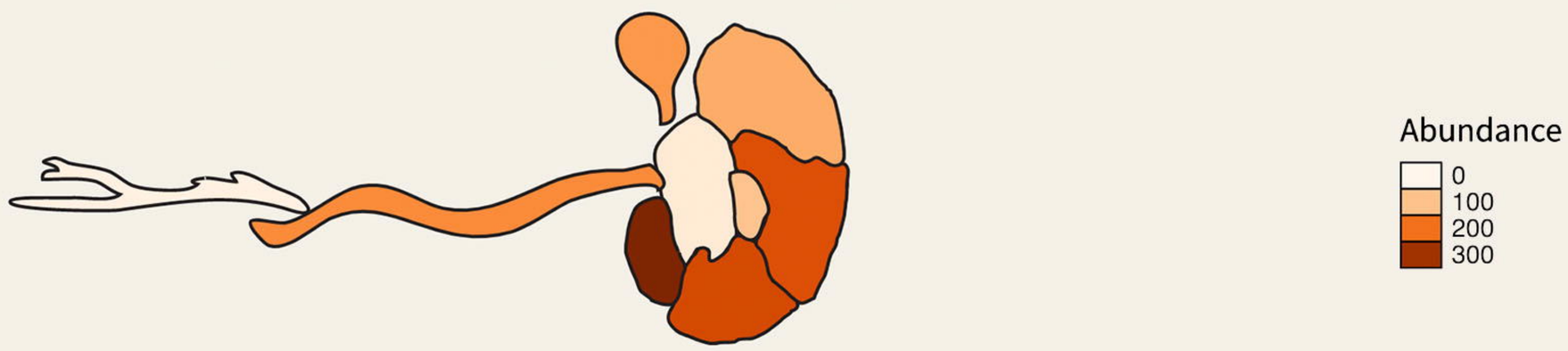
A



B

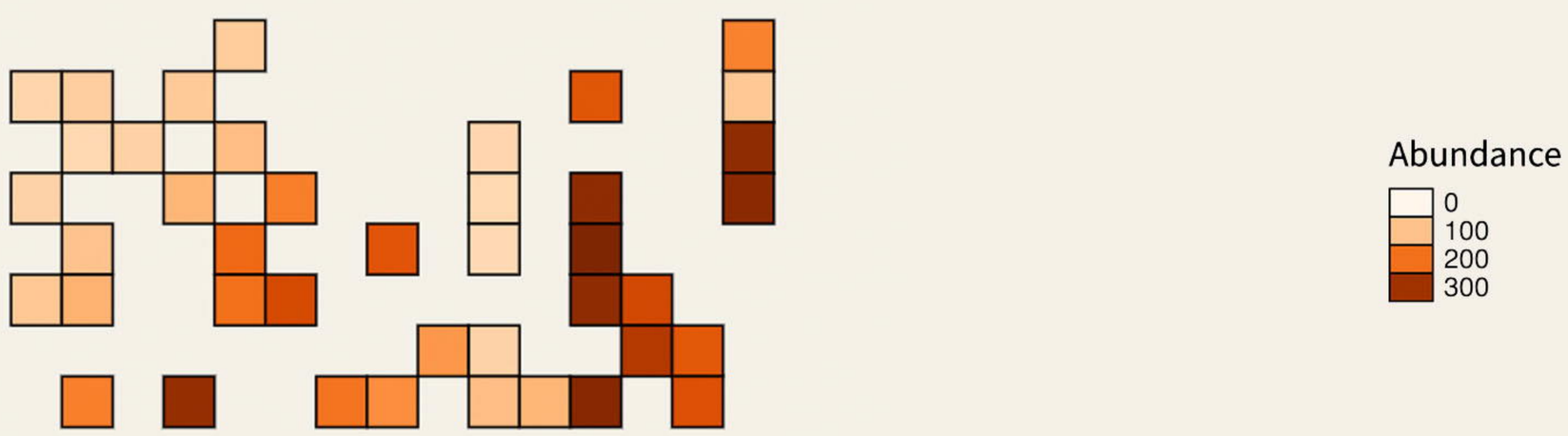


C

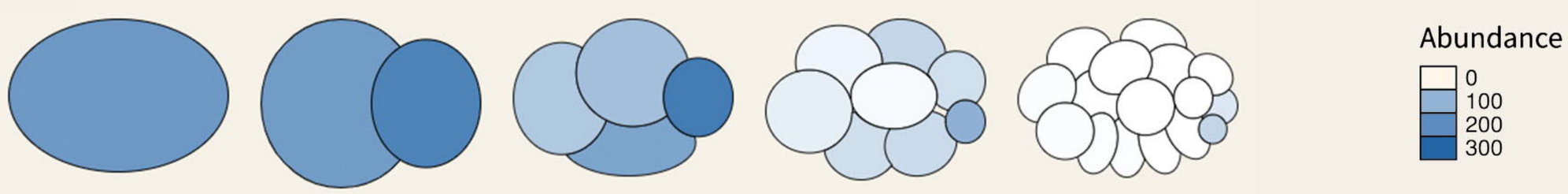


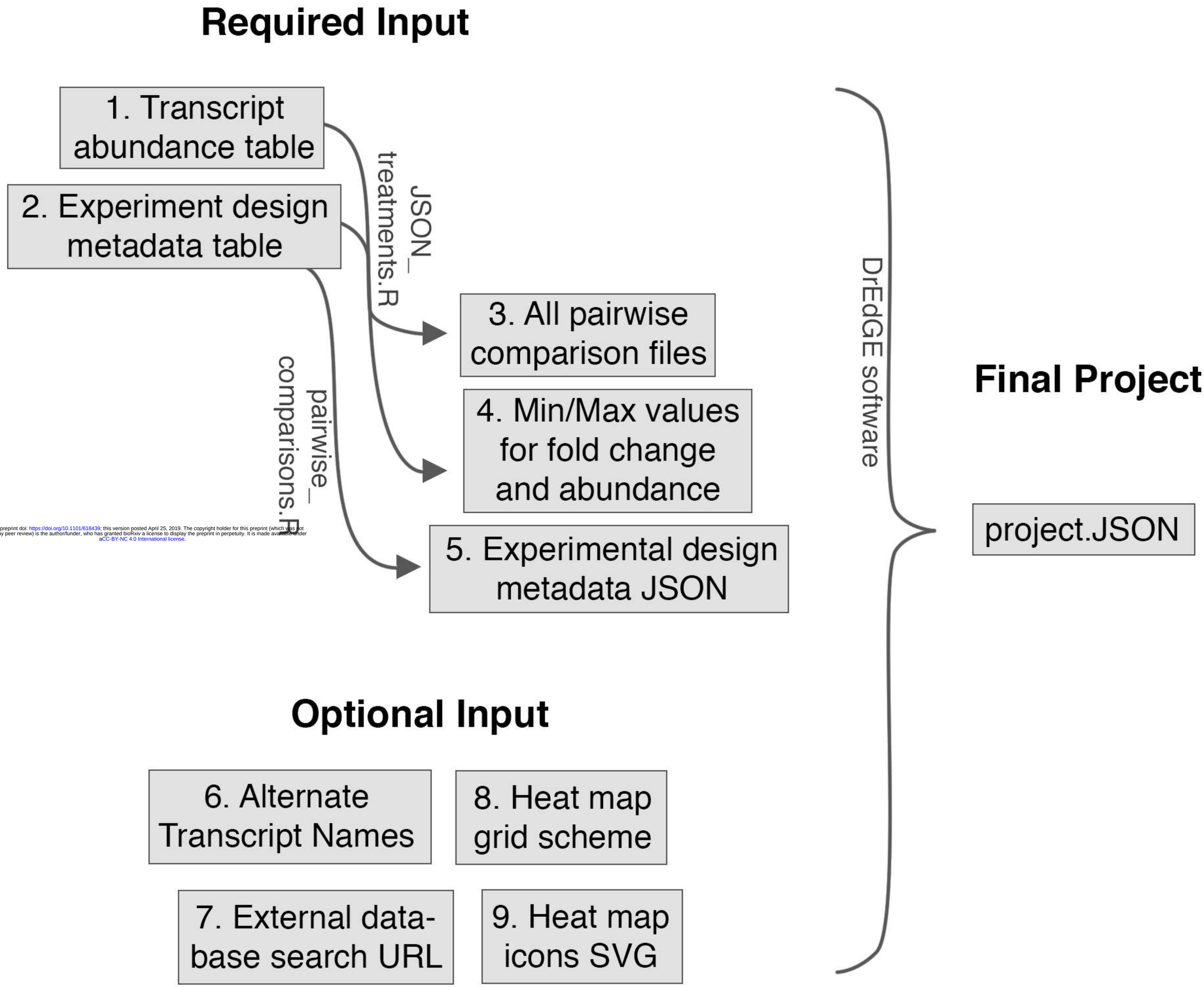
bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

D



E





bioRxiv preprint doi: https://doi.org/10.1101/618439; this version posted April 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.