

Accepted Manuscript

Graphic Depiction of Bioinformatics Data

Agustin Calatroni, M.A. M.S., Jeremy J. Wildfire, M.S.

PII: S0091-6749(17)31019-9

DOI: [10.1016/j.jaci.2017.05.043](https://doi.org/10.1016/j.jaci.2017.05.043)

Reference: YMAI 12884

To appear in: *Journal of Allergy and Clinical Immunology*

Received Date: 31 August 2016

Revised Date: 10 May 2017

Accepted Date: 17 May 2017

Please cite this article as: Calatroni A, Wildfire JJ, Graphic Depiction of Bioinformatics Data, *Journal of Allergy and Clinical Immunology* (2017), doi: 10.1016/j.jaci.2017.05.043.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Graphic Depiction of Bioinformatics Data

Agustin Calatroni, M.A. M.S.,¹ Jeremy J. Wildfire, M.S.¹

Affiliations:

¹ Rho Federal Systems Division, Chapel Hill, NC

Corresponding Author:

Agustin Calatroni, M.A. M.S.
6330 Quadrangle Drive
Chapel Hill, NC 27517
Telephone: (919) 408-8000
Fax: (919) 408-0999
Email: agustin_calatroni@rhoworld.com

Word Count: 1384

Introduction

The advent of whole-genome DNA sequencing and multiple -omic technologies has led to a big data revolution in health research^{1,2}. All of these technologies generate massive amounts of data, and many are relevant to Allergy and Immunology, including transcriptomics, proteomics, metabolomics, epigenomics, microbiomics and exposomics. The interdisciplinary field of bioinformatics has formed in response to this influx of big data. Bioinformatics combines computer science, statistics, and biology to analyze and interpret biological data. Although bioinformatics is a technical field, visualization remains a central tool for understanding the complex relationships found in these data sets. This paper provides a review of current bioinformatics data visualization best practices through a series of case studies using publically available mechanistic data from a recent oral immunotherapy clinical trial³.

While statistical graphics have been in use since at least the 1600s⁴, work to understand links between visualization and human perception and to create visualization “best practices” is much more recent.⁵⁻⁶ Modern data visualization combines techniques developed across diverse fields including psychology (perception, attention), color theory, statistics and computer science (human-computer interaction, big data storage, and processing). This multi-disciplinary approach to visualization serves as a cornerstone for understanding bioinformatics data.

Bioinformatics Visualization Techniques

Fundamentals

Visualizing bioinformatics data allows researchers to gain insight into their data that is not obvious from simple descriptive statistics (Supplemental Figure 1). Much can be accomplished by iteratively adding layers to a simple chart (Supplemental Figure 2) using foundational data visualization techniques such as data and visual encodings (Supplemental Figure 3), color theory (Supplemental Figures 4 and 5), plot types and data-to-ink ratio. New visualization techniques specifically for the big data sets used in bioinformatics are also constantly being developed and refined⁷. Some common bioinformatics visualization techniques are shown in Figure 1 and are discussed below. Code for the figures is available at <https://github.com/RhoInc/JACI-BioinformaticsGraphics>

Visualization in Allergy and Immunology

The basic techniques described above can be used to great effect in allergy and immunology research. As a case study, we downloaded the publically available datasets from the Oral Immunotherapy for Treatment (OIT) of Egg Allergy in Children study³ through the National Institutes for Allergy and Infectious Diseases (NIAID) data sharing site, ImmPort and created a series of data visualizations using the study data. We used the basic data visualization techniques described above to create figures showing the changes in cytokine levels by treatment arm (Supplemental Figure 6) that matched the findings from Burke et al. showing that OIT reduces CD63+ Basophil levels when compared to placebo³. Data from the same study is used throughout this report to demonstrate other visualization techniques.

Dealing with Overplotting

Overlapping points, also known as overplotting, are very common when visualizing large data sets (Supplemental Figure 7). Simple data transformations (e.g. logarithmic scales) should always be considered when visualizing data, but transformations alone don't always resolve this issue. Visualization techniques such as alpha blending (reducing the opacity of individual points), jittering (adding small amounts of random variation to points) and binning (grouping similar points) are useful tools for dealing with overplotting, as are visualizations using summary statistics. Boxplots, histograms and density plots can provide an overview of univariate distributions and bagplots (an extension of boxplots), contour plots and hexagonal binning can summarize bivariate relationships more effectively than a plot showing every data point. It can be especially informative to combine multiple approaches in the same graphic (Figure 1C and Supplemental Figure 7, Panel F).

Faceting

Visualizations often need to be expanded to include more than 2 variables of interest. Information may be visually encoded in many ways, but only some encodings scale well to large data sets. For example, adding different symbols or detailed annotations to a plot with thousands of points will not accomplish much, given a large amount of data displayed. On the other hand, expanding a chart to include multiple panels, also known as faceting, is generally an effective

method of incorporating multivariate data in visualization and can be used in combination with any number of other techniques (Supplemental Figure 7, Panel G).

Color

Color encoding is one of the most common visualization techniques, but color can reveal patterns when used effectively or hide them when used poorly. Different color palettes are appropriate for sequential data (light colors for low data, dark for high data), diverging data (light colors for mid-range data, high-contrast dark colors for low- and high-range data) and qualitative data (colors designed to give maximum visual difference between classes).

ColorBrewer⁸ provides a series of color palettes for all types of color encodings, including color-blind and photocopy friendly palettes, and serves as an excellent entry point when selecting colors (Supplemental Figures 4 and 5).

Scatterplot Matrices and Heat Maps

A scatterplot matrix is a two-dimensional grid showing each possible pair of associations for a series of variables. These matrices are flexible tools that work well in bioinformatics analyses. Depending on analysis objectives, these matrices can either be customized to incorporate diverse statistical methodologies and visualization techniques (Supplemental Figure 8) or streamlined to present thousands of comparisons in a heat map (Figure 1B and Supplemental Figure 9).

High-Dimensional Data

Dealing with overplotting and incorporating several variables into a visualization using faceting and thoughtful use of color is valuable, but is of limited use in data with hundreds or thousands of variables. Data at this scale, also known as high-dimensional data, is common in bioinformatics analysis. These high-dimensional data present complex statistical issues and require methods that combine applied statistics and computer science. Novel custom data visualizations, such as genome explorers⁹ and circos diagrams¹⁰ showing massive numbers of bivariate associations in a circular layout are constantly being developed and refined in conjunction with these new methods.

Dimension reduction

Dimension reduction techniques, such as Principal Components Analyses (PCA) or Metric Multidimensional Scaling (MDS) can also be used to reduce high-dimensional data to its most salient features. The main idea of these techniques is to transform the original multivariate data into a set of new smaller dimensions, which can be visualized using standard techniques. It is important to normalize variables before using these techniques. Biplots, an exploratory visualization allowing both samples and variables to be displayed, are particularly useful. Points that are close to each other represent observations with similar values, and highly correlated variables point in the same direction (or opposite for negative correlations) while uncorrelated variables are at right angles to each other (Figure 1D and Supplemental Figure 10).

Network diagrams

Bioinformatics data are often represented as networks. A network diagram is composed of nodes (or vertices) representing their features and edges for their connections. Given the sizes and complexity of networks, techniques that automate their creation are important. For example, a force-directed layout applies simulated attractive and repulsive physical properties to nodes and edges to determine optimal positions for each element of the visualization (Supplemental Figure 11).

Interactive Graphics

As research and communication methodologies have moved from paper-based to digital, visualizations are no longer required to be static, nor limited to the constraints of print. As a result, interactive data visualization has become a compelling complement to all types of static visualization. Interactive graphics are often implemented in web pages that pair traditional statistical graphics with simple graphical user interfaces that allow for real-time data exploration. An interactive graphic that starts with a summary display (overview) and then allows for the removal of uninteresting items (filtering) while focusing on data points and trends of interest (zooming and details-on-demand) gives researchers the ability to investigate complex data with intuitive tools¹¹. This pattern mirrors biomedical research in general and scales well to bioinformatics data. Until recently specialized technical knowledge was needed to configure interactive graphics for bioinformatics data, but many standalone packages that provide simple ways to configure and share sophisticated interactive graphics have been released in recent years,

significantly lowering the bar for entry (Supplemental Table 1). Many compelling medical journals increasingly encourage the development of web-only, online supplemental material in parallel with traditional manuscripts.

Conclusion

Bioinformatics data is too complex and nuanced to rely on one-size-fits-all visualization solution, but a set of foundational tools and techniques is rapidly emerging (Supplemental Table 1). Creating informative visual representations of bioinformatics data requires a well-rounded approach that combines foundational data visualization principals, custom visualizations created for specific biologic processes, and new visualization techniques merging traditional statistical principles with emerging technology.

Figure 1 –Bioinformatics graphics for Allergy and Clinical Immunology including a paneled plot with multiple enhancements (Panel A), a heat map (Panel B), a contour plot showing flow cytometry results (Panel C) and the results of the principal component analysis (Panel D). Additional details for each panel are provided in the online supplement (See supplemental figures 6, 9, 7, and 10 for panels A, B, C, and D, respectively). All graphics use publically available data from Burks et al. Source code are available at <https://github.com/RhoInc/JACI-BioinformaticsGraphics>.

A. Clinical Comparison

B. Heat map

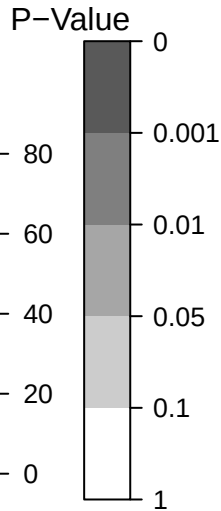
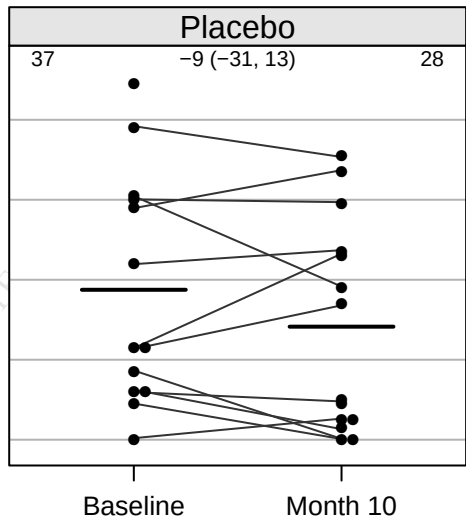
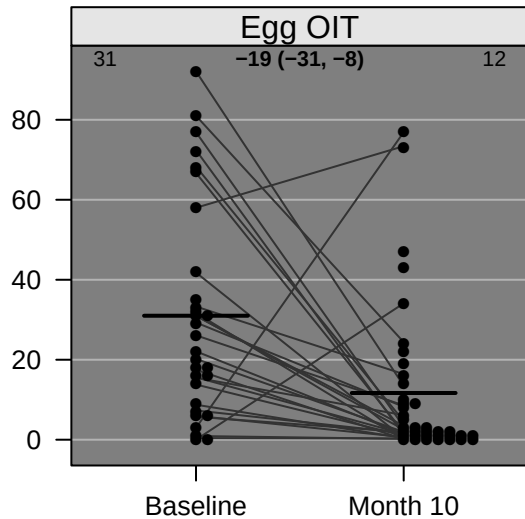
C. Cytometry data with contours

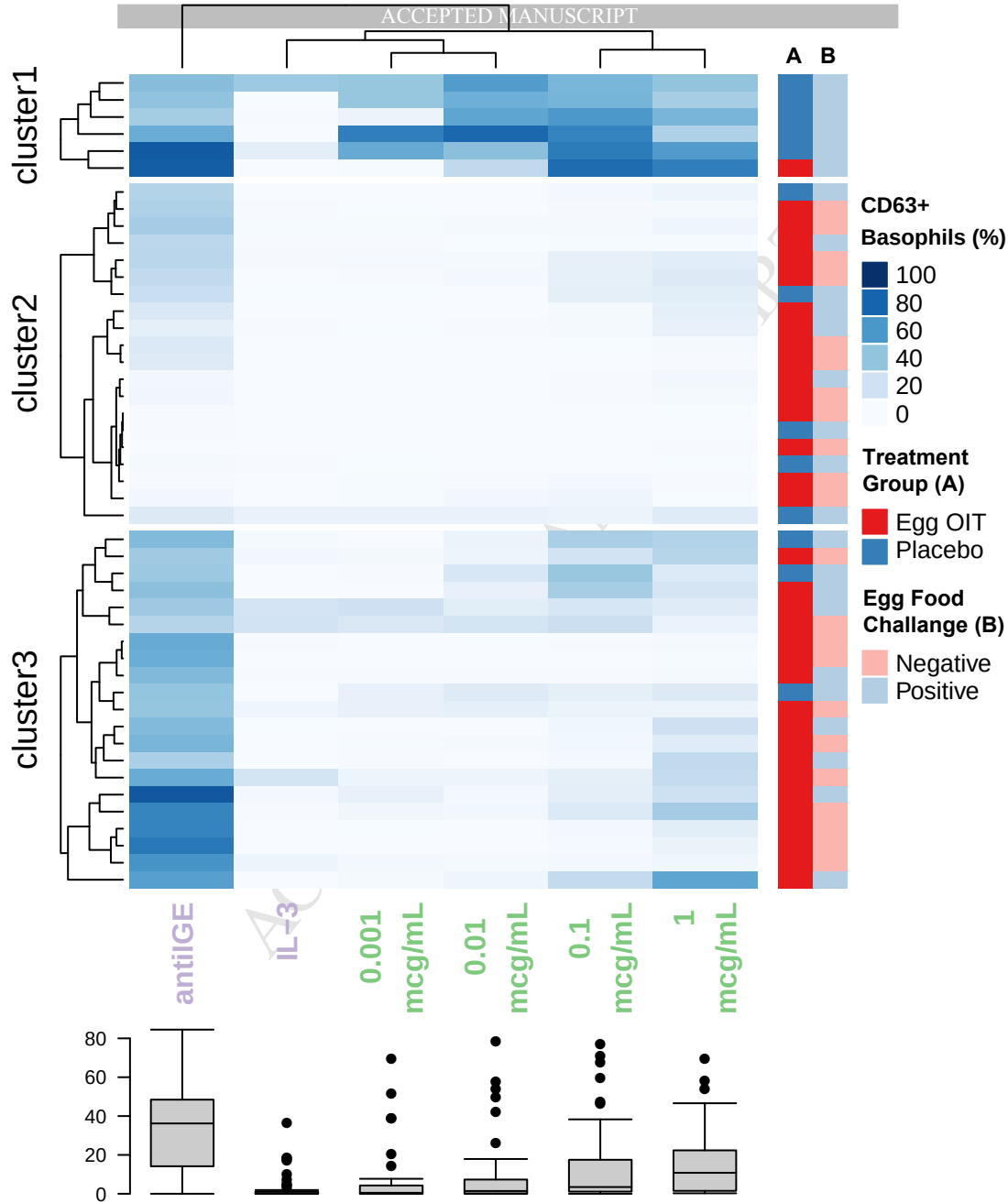
D. Principal Component Analysis (PCA)

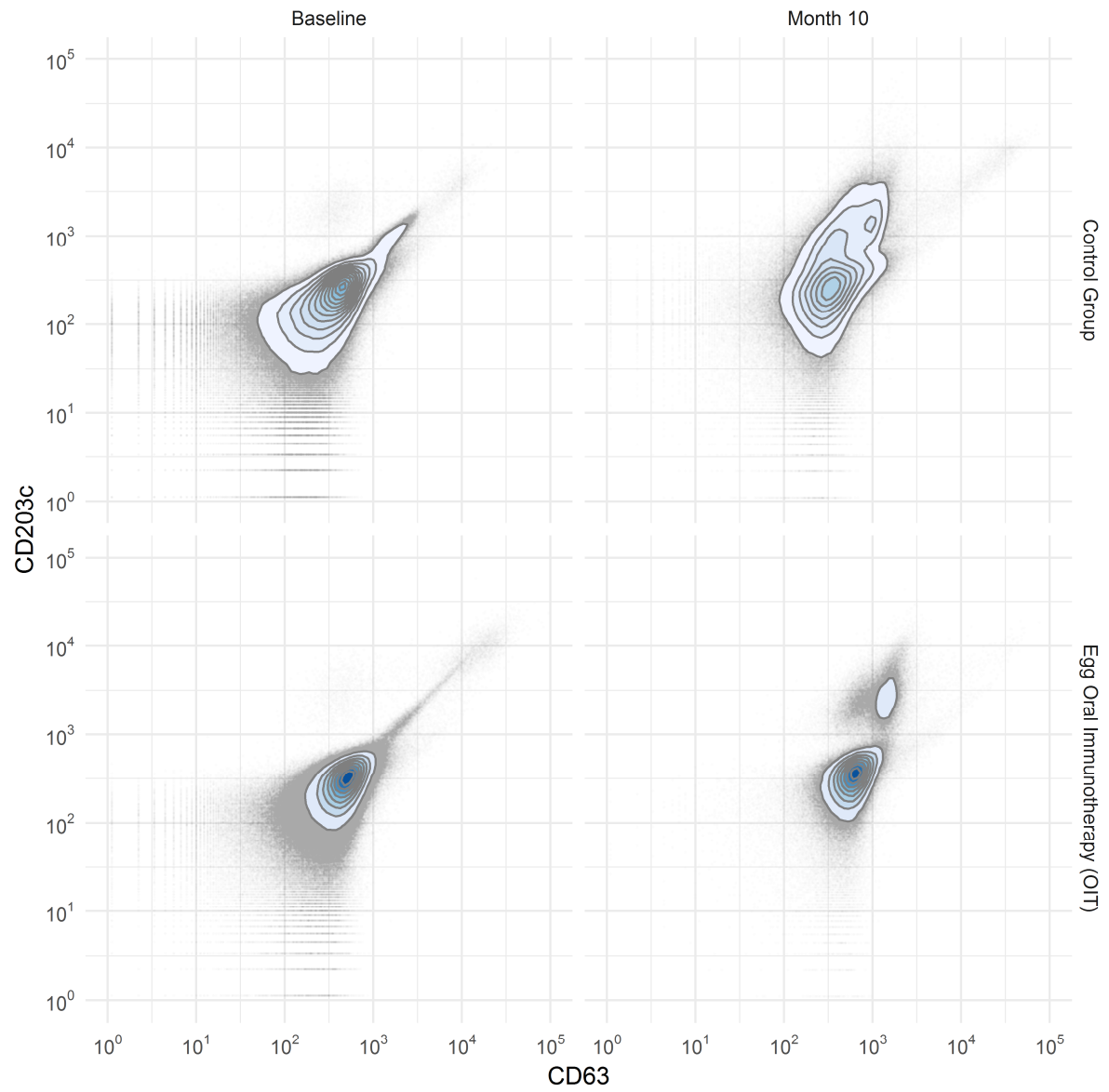
References

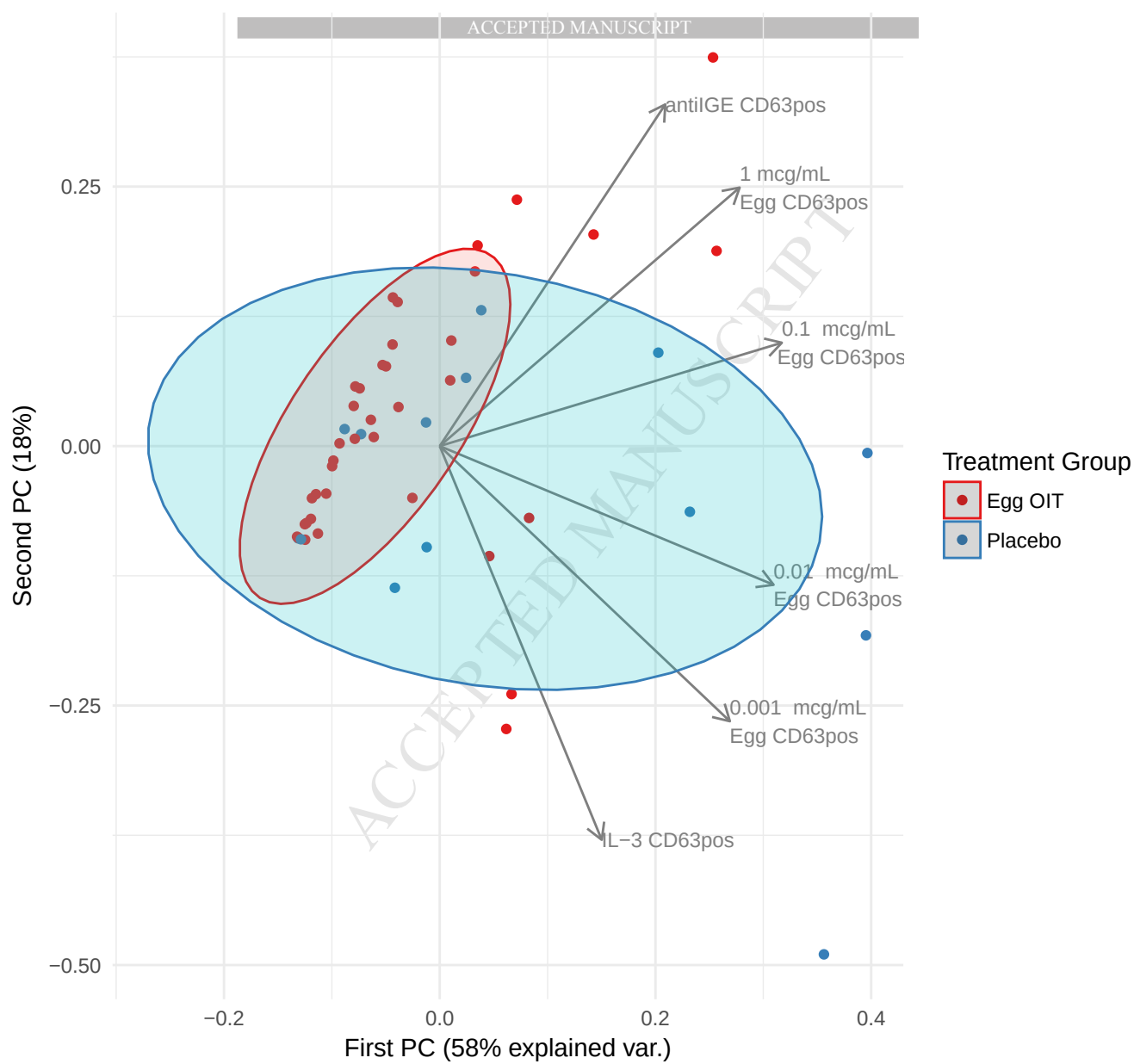
- 1 - Wang X, Baumgartner C, Shields DC, Deng HW, Beckmann JS, editors. Application of Clinical Bioinformatics. Springer; 2016 Mar 31.
- 2 - Ye SQ, editor. Big data analysis for bioinformatics and biomedical discoveries. Chapman and Hall/CRC; 2015 Dec 22.
- 3 - Burks AW, Jones SM, Wood RA, Fleischer DM, Sicherer SH, Lindblad RW, et al. Oral immunotherapy for treatment of egg allergy in children. *New England Journal of Medicine*. 2012 Jul 19;367(3):233-43.
- 4 - Friendly M. Milestones in the history of data visualization: A case study in statistical historiography. In *Classification—the Ubiquitous Challenge 2005* (pp. 34-52). Springer Berlin Heidelberg.
- 5 - Tufte ER, McKay SR, Christian W, Matey JR. Visual explanations: images and quantities, evidence and narrative. *Computers in Physics*. 1998 Mar 1;12(2):146-8.
- 6 - Few S. Show me the numbers: Designing tables and graphs to enlighten. Analytics Press; 2012 Jun 1..
- 7 - Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*. 2004 Sep 15;5(10):R80.
- 8 - Harrower M, Brewer CA. ColorBrewer. org: an online tool for selecting colour schemes for maps. *The map reader: Theories of mapping practice and cartographic representation*. 2011 Apr 6:261-8.
- 9 - Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, et al: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*. 2010 Jan 21:19-0..
- 10 - Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al: an information aesthetic for comparative genomics. *Genome research*. 2009 Sep 1;19(9):1639-45.
- 11 - Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on* 1996 Sep 3 (pp. 336-343). IEEE.

% CD63 Expression
1 mcg/mL Egg









Graphic Depiction of Bioinformatics Data**Supplemental Figures**

Agustin Calatroni, M.A. M.S.,¹ Jeremy J. Wildfire, M.S.¹

Affiliations:

¹ Rho Federal Systems Division, Chapel Hill, NC

Corresponding Author:

Agustin Calatroni, M.A. M.S.
6330 Quadrangle Drive
Chapel Hill, NC 27517
Telephone: (919) 408-8000
Fax: (919) 408-0999

Email: agustin_calatroni@rhoworld.com

Supplemental Figure 1: Data table and paneled scatter plot for Anscombe's quartet¹. Anscombe's quartet¹, created by statistician Francis Anscombe, demonstrates the value of visualizing data by demonstrating how four very different data distributions can have the same descriptive statistics (mean, variance and correlation). Source code available [here](#).

A. Table

1		2		3		4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

B. Scatterplot Figure

Supplemental Figure 2: Sample workflow for visualizing bioinformatics data. It can be challenging to separate signal from noise when working with bioinformatics big data. There is no one-size-fits-all solution, so we use an iterative process, guided by both clinical expertise and data visualization best practices, to create informative data visualizations²⁻¹⁰. Generally speaking, we start with very simple visualizations of the raw data (Step 1) and then layer on appropriate visual encodings and data transformations to better highlight features of the data that are not immediately apparent in large bioinformatics data sets (Step 2). We may then add additional data dimensions via a group-wise comparison (Step 3) or add annotations highlighting the most relevant features of the data (Step 4). Finally, additional data dimensions can be incorporated with faceting (Step 5) or via interactive data visualization (Step 6). Here we show the step-by-step evolution of a graphic, from the initial simple univariate dot plot to a complex set of multi-panel histograms. The authors have published many displays in JACI using this process ([link](#)) and the remaining figures in this online supplement were created using a similar process. Source code for the figures below and additional step-by-step examples are available [here](#).

ACCEPTED MANUSCRIPT

Supplemental Figure 3: Overview of visual encodings for data. These displays provide an overview of the ways data can be encoded in a data visualization. Panel A describes some of the most common ways to encode information visually, and Panel B shows the relative strength of each encoding for quantitative (i.e. length – 160 cm), ordinal (i.e: quality of meat - grade A, AA, AAA, ...) and nominal (i.e: fruits - apples, oranges, ...) data domains. Displays are from Nathan Yau¹¹ via Lucy Park¹² and are based on research from Cleveland¹³, Wilkinson¹⁴, Bertin¹⁵ and others.

A. Data Encoding

B. Visual encodings by data type

Supplemental Figure 4: Color Brewer palettes for effective color selection. Color Brewer¹⁶ provides optimized color palettes for data visualization. Sequential Palettes are optimized for ordered data that progress from low to high. Lightness steps dominate the look of these schemes, with light colors for low data values to dark colors for high data values. Qualitative Palettes do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes. Qualitative schemes are best suited to representing nominal or categorical data. Diverging palettes put equal emphasis on mid-range critical values and extremes at both ends of the data range. The critical class or break in the middle of the legend is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues. Source code available [here](#).

A. Sequential palettes

B. Qualitative palettes

C. Diverging palettes

Supplemental Figure 5 - An Interactive Color Brewer palettes explorer tool. This interactive tool from the R tmaptools package¹⁷ can adjust the number of colors and their contrast. Selected palettes can be tested for different types of color blindness including deuteranopia (reduced the sensitivity to green light, the most common), protanopia (reduces the sensitivity to red light) and tritanopia (reduced sensitivity to blue light, extremely rare). Overall, color blindness affects 8% of men and 0.5% of women. Source code available [here](#).

Supplemental Figure 6: Visualization Case Study for Oral Immunotherapy. This series of visualizations demonstrates the step-wise process for visualizing data related to allergy and clinical immunology. The series uses publicly available CD63+ Basophils (%) data from the Oral Immunotherapy for Treatment (OIT) of Egg Allergy in Children¹⁸. Panel A shows the different level of CD63+ Basophils (%) for 0.1 µg of egg extract. Here we plot all the data but avoid overplotting of individual points by moving them horizontally. Furthermore, we facilitate the interpretation of results by adding mean lines and using different panels for each treatment arm. Panel B adds lines connecting the paired data for each participant (baseline and month 10 visit) thereby stressing the changes in outcome within each arm. While the overall trend towards decreased CD63 in the treatment arm is still clear, the lines show that a few participants in the Egg OIT arm had their values increase after treatment as seen by the positive slopes. Because annotations are an important layer in any chart, we added the means for each group at each time point and their difference with 95% confidence interval. Finally, the panels are shaded based on the statistical significance of the group comparison, with darker cells indicating lower p-values. Panel C adds an additional dimension to the chart by showing additional panels for other concentrations. When working with multidimensional data, we suggest using the lattice framework^{2,3}. By using the same axes between all the panels, it becomes easy to compare the results of the treatment with different extracts. Clearly, the treatment is significant across all the extracts, but the effect is smallest at the lowest concentration (difference of 4%). Source code for all figures available [here](#).

A.

B.

Supplemental Figure 7 - Methods for Addressing Overplotting using Egg Oral Immunotherapy data.

Overplotting occurs when elements of a figure are drawn on top of each other. Here we look at flow cytometry data from the OIT Egg trail¹⁸. Panel A presents a scatterplot of 225,000 points with no transformation. Overplotting remains an issue even after transformation (Panel B). Reducing the opacity of the points (alpha blending, Panel C) and inserting marginal distributions using a histogram (Panel D) address the issue with varying degrees of success. However, binning (grouping the points in to hexagons) the raw data (Panel E) and contour plots (Panel F) more clearly show the underlying bivariate distribution. Finally, we compare the effect of the OIT intervention over time for 2 participants using multiple panels in the lattice framework (Panel G). Although there are different profiles at baseline for the two participants, there is a clear change in the treated participant while no change was observed in the placebo participant. Source Code for all figures is available [here](#).

A.
C.
E.

B.

D. Scatterplot w/ marginal distributions

F.

ACCEPTED MANUSCRIPT

Supplemental Figure 8 - Sample Scatterplot Matrix with Egg Oral Immunotherapy data. Scatterplot matrix visualizations show a series of pairwise relationships that can be customized to meet many common needs in bioinformatics research. Details and custom analyses can be incorporated into each pane of the matrix to give a detailed description; furthermore, they can be arranged by ordering the variables accordingly to their associations. The figure below illustrates this approach using the Egg OIT¹⁸ mechanistic data at baseline. We highlight the correlations and 95% CI between variables (upper triangle, bold for those with statistical significance), marginal distributions using a kernel density plots (diagonal blue lines) to visualize the distribution of the variable over a continuous interval, and scatterplots with LOESS (Local Polynomial Regression) regression (lower triangle red lines) are added to each panel. We also order columns (and thus also the associated rows) of the display according to the correlation between the variables (variables with stronger correlations will appear next to each other) by using a hierarchical clustering approach and include the associated dendrogram (showing in the top margin) in the figure. The source code is available [here](#).

Supplemental Figure 9: Sample Heatmap with Egg Oral Immunotherapy data. This figure shows a heatmap^{19, 20} using the Egg OIT¹⁸ mechanistic data at Month 10. The heat map shows the 47 participants and 6 variables. We have heavily annotated the figure using a sequential color scale, the treatment group and the outcome of the egg food challenge. Furthermore, we have added the marginal distributions of each variable using a boxplot and reorganized the columns and rows using a hierarchical cluster analysis so that participants and variables that are most similar are closest together. The results of these clusters are annotated using marginal dendrograms. Cluster three is of particular interest; it shows a group of 5 placebo participants and 1 OIT participant with a positive egg challenge, and strong responses to the mechanistic measurements. The source code is available [here](#).

Supplemental Figure 10 – Demonstration of Dimension Reduction Techniques with Egg Oral

Immunotherapy data. Principal Component Analysis (PCA) is a multivariate technique that summarizes systematic patterns of variation in the data²¹. In bioinformatics data, where large numbers of variables are measured, PCA is used to simplify complex datasets by reducing the number of dimensions measured. This dimension reduction helps to reveal the most salient structures in both observations and variables. PCA achieves this by transforming the observed variables into a set of new variables, the principal components (PCs), which are uncorrelated and explain the variation in the dataset. This example shows PCA results of CD63+ Basophils collected at the 10 Month visit in the egg OIT study¹⁸. The display shows the relationship of two PCs (plotted on the x and y-axis) with two different data domains. First, the relationship between the PCs and the participants (points) and treatment groups (circles) is shown. The positions of the circles show that the participants in the Egg OIT group clustered more closely together, especially for the first principle component. The second domain shows the variables used to calculate the PCs. The arrows (also known as eigenvectors) represent the variable used to calculate the PCs. The length and direction of the arrows represent the strength of the relationship between a given variable and the PCs. The length represents the magnitude of the relationship while the direction indicates to what degree the variable is related to each PC. Arrows for correlated variables are either side-by-side or diametrically opposed (for negative correlations) while the arrows for independent variables form right angles. In our example, the patterns of expression of CD63 appeared to be similar in antiIGE, 1 & 0.1 mcg/ml Egg in contrast to those IL-3 and to a lesser degree 0.01 & 0.001 mcg/mg Egg. The source code is available [here](#).

Supplemental Figure 11 - Network Diagram. Network diagrams are useful for visualizing many common bioinformatics data types (e.g. metabolic, protein interaction and gene regulatory networks). This example shows a network analysis of cytokine-by-stimulant data measured longitudinally from the Urban Environment and Childhood Asthma (URECA) study²². The map uses the correlation matrix among the 164 variables as the proximity matrix to create the network. More highly correlated variables are closer to each other. Correlations greater than 0.26 (95th percentile) in absolute value are represented by lines. The age at which the cytokine responses were assessed is color-coded. The source code is available [here](#).

ACCEPTED MANUSCRIPT

Supplemental Figure 12 - Examples of Interactive Visualization for Bioinformatics data. An interactive visualization is a powerful approach for exploring bioinformatics big data with wide-ranging applications. New open frameworks for web-based interactive data visualization²³⁻²⁶ facilitate linked displays that allow users to zoom into points of special interest, which is a simple, but powerful technique. Many interactive visualization libraries targeting specific questions related to bioinformatics have been created in recent years, including the two shown below used for visualizing and exploring gene expression data. Panel A shows an interactive data explorer created using the Glimma R package²⁷. The right panel shows a volcano plot summarizing expression data for a selected gene by showing posterior odds of differential expression vs. log Fold Change. Clicking a point shows the raw sample data for a single gene in the panel to the left. The table beneath the plots shows additional summary information and allows the user to search for a particular gene. Panel B shows a snapshot of a visualization created using the Degust R package. That web-based tool allows users to visualize and explore RNA-seq differential gene-expression data²⁸. The tool provides several types of visualization (parallel coordinate, Mean-Average, MDS plot and heat maps) with built-in interactive features including real-time filters and access to linked Kegg pathway analyses for genes of interest.

A.

B.

Supplemental Table 1: Visualization R packages that we frequently used during our analyses

Package	Title	Description	Author	Learn More
lattice	Trellis Graphics for R	A powerful and elegant high-level data visualization system inspired by Trellis graphics, with an emphasis on multivariate data. Lattice is sufficient for typical graphics needs, and is also flexible enough to handle most nonstandard requirements.	Deepayan Sarkar	Lattice: Multivariate Data Visualization with R - Figures and Code http://lmdvr.r-forge.r-project.org/figures/figures.html
latticeExtra	Extra Graphical Utilities Based on Lattice	Building on the infrastructure provided by the lattice package, this package provides several new high-level functions and methods, as well as additional utilities such as panel and axis annotation functions.	Felix Andrews	http://latticeextra.r-forge.r-project.org/
RColorBrewer	ColorBrewer Palettes	Provides color schemes for maps (and other graphics) designed by Cynthia Brewer	Erich Neuwirth	http://colorbrewer2.org
colorspace	Color Space Manipulation	Carries out mapping between assorted color spaces including RGB, HSV, HLS, CIEXYZ, CIELUV, HCL (polar CIELUV), CIELAB and polar CIELAB. Qualitative, sequential, and diverging color palettes based on HCL colors are provided along with an interactive palette picker (with either a Tcl/Tk or a shiny GUI)	Achim Zeileis	https://cran.r-project.org/web/packages/colorspace/vignettes/hcl-colors.pdf
dichromat	Color Schemes for Dichromats	Collapse red-green or green-blue distinctions to simulate the effects of different types of color-blindness	Thomas Lumley	
ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.	Hadley Wickham	http://ggplot2.tidyverse.org/
scales	Scale Functions for Visualization	Graphical scales map data to aesthetics, and provide methods for automatically determining breaks and labels for axes and legends.	Hadley Wickham	https://github.com/hadley/scales
ggExtra	Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements	Collection of functions and layers to enhance 'ggplot2'. The main function is ggMarginal(), which can be used to add marginal histograms/boxplots/density plots to 'ggplot2' scatterplots.	Dean Attali	https://github.com/daattali/ggExtra http://www.ggplot2-exts.org/ggiraph.html
ggfortify	Data Visualization Tools for Statistical Analysis Results	Unified plotting tools for statistics commonly used, such as GLM, time series, PCA families, clustering and survival analysis. The package offers a single plotting interface for these analysis results and plots in a unified style using 'ggplot2'.	Masaaki Horikoshi	https://github.com/sinhrks/ggfortify
plotly	Create Interactive Web Graphics via 'plotly.js'	Easily translate 'ggplot2' graphs to an interactive web-based version and/or create custom web-based visualizations directly from R. Once uploaded to a 'plotly' account, 'plotly' graphs (and the data behind them) can be viewed and modified in a web browser	Carson Sievert	https://plot.ly/r/
qgraph	Graph Plotting Methods, Psychometric Data Visualization and	Can be used to visualize data as networks as well as provides an interface for visualizing weighted graphical models.	Sacha Epskamp	http://sachaepskamp.com/qgraph

Package	Title	Description	Author	Learn More
	Graphical Model Estimation			
hexbin	Hexagonal Binning Routines	Binning and plotting functions for hexagonal bins. Now uses and relies on grid graphics and formal (S4) classes and methods.	Edzer Pebesma	https://github.com/edzer/hexbin https://cran.r-project.org/web/packages/hexbin/vignettes/hexagon_binning.pdf
directlabels	Direct Labels for Multicolor Plots	An extensible framework for automatically placing direct labels onto multicolor 'lattice' or 'ggplot2' plots. Label positions are described using Positioning Methods which can be re-used across several different plots. There are heuristics for examining "trellis" and "ggplot" objects and inferring an appropriate Positioning Method.	Toby Dylan Hocking	http://directlabels.r-forge.r-project.org/
<u>ComplexHeatmap</u>	make complex heatmaps as well as self define annotation graphics	Complex heatmaps are efficient to visualize associations between different sources of data sets and reveal potential structures. Here the ComplexHeatmap package provides a highly flexible way to arrange multiple heatmaps and supports self-defined annotation graphics.	Zuguang Gu	https://github.com/jokergoo/ComplexHeatmap https://academic.oup.com/bioinformatics/article-abstract/32/18/2847/1743594/Complex-heatmaps-reveal-patterns-and-correlations
dendextend	Extending 'Dendrogram' Functionality in R	Offers a set of functions for extending 'dendrogram' objects in R, letting you visualize and compare trees of 'hierarchical clusterings'. You can (1) Adjust a tree's graphical parameters - the color, size, type, etc of its branches, nodes and labels. (2) Visually and statistically compare different 'dendrograms' to one another.	Tal Galili	https://github.com/talgilili/dendextend
corrgram	Plot a Correlogram	Calculates correlation of variables and displays the results graphically. Included panel functions can display points, shading, ellipses, and correlation values with confidence intervals.	Kevin Wright	https://github.com/kwstat/corrgram
scatterplot3d	3D Scatter Plot	Plots a three dimensional (3D) point cloud.	Uwe Ligges	https://cran.r-project.org/web/packages/scatterplot3d/vignettes/s3d.pdf
shiny	Web Application Framework for R	Makes it incredibly easy to build interactive web applications with R. Automatic "reactive" binding between inputs and outputs and extensive prebuilt widgets make it possible to build beautiful, responsive, and powerful applications with minimal effort.	Winston Chang	http://shiny.rstudio.com
<u>trelliscopejs</u>	Create and Navigate Large Multi-Panel Visual Displays	Trelliscope is a scalable, flexible, interactive approach to visualizing data. The trelliscopejs R package provides methods that make it easy to create a Trelliscope display specification for the Trelliscope JavaScript library <u>trelliscopejs-lib</u> . High-level functions are provided for creating displays from within dplyr (via summarise()) or ggplot2 (via facet_trelliscope()) workflows. Low-level functions are also provided for creating new interfaces.	Ryan Hafen	https://github.com/hafen/trelliscopejs
d3Network	Tools for creating D3 JavaScript network, tree, dendrogram, and	This packages is intended to make it easy to create D3 JavaScript network, tree, dendrogram, and Sankey graphs from R using data frame	Christopher Gandrud	http://christophergandrud.github.io/d3Network/

Package	Title	Description	Author	Learn More
	Sankey graphs from R			
d3heatmap	Interactive Heat Maps Using 'htmlwidgets' and 'D3.js'	Create interactive heat maps that are usable from the R console, in the 'RStudio' viewer pane, in 'R Markdown' documents, and in 'Shiny' apps. Hover the mouse pointer over a cell to show details, drag a rectangle to zoom, and click row/column labels to highlight.	Joe Cheng	https://github.com/rstudio/d3heatmap https://cran.r-project.org/web/packages/d3heatmap/vignettes/Introduction.html
pairsD3	D3 Scatterplot Matrices	Creates an interactive scatterplot matrix using the D3 JavaScript library. See http://d3js.org/ for more information on D3	Garth Tarr	https://github.com/garthtarr/pairsD3/
scatterD3	D3 JavaScript Scatterplot from R	Creates 'D3' 'JavaScript' scatterplots from 'R' with interactive features : panning, zooming, tooltips, etc.	Julien Barnier	https://github.com/juba/scatterD3
DiagrammeR	Create Graph Diagrams and Flowcharts Using R	Create graph diagrams and flowcharts using R.	Richard Iannone	http://rich-iannone.github.io/DiagrammeR/
RCircos	Circos 2D Track Plot	A simple and flexible way to generate Circos 2D track plot images for genomic data visualization is implemented in this package. The types of plots include: heatmap, histogram, lines, scatterplot, tiles and plot items for further decorations include connector, link (lines and ribbons), and text (gene) label. All functions require only R graphics package that comes with R base installation.	Hongen Zhang	https://cran.r-project.org/web/packages/RCircos/vignettes/Using_RCircos.pdf https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-244
heatmap3	An Improved Heatmap Package	An improved heatmap package. Completely compatible with the original R function 'heatmap', and provides more powerful and convenient features.	Shilin Zhao	https://cran.r-project.org/web/packages/heatmap3/vignettes/vignette.pdf https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S10-P16
eulerr	Area-Proportional Euler Diagrams	If possible, generates exactly area-proportional Euler diagrams, or otherwise approximately proportional diagrams using numerical optimization. A Euler diagram is a generalization of a Venn diagram, relaxing the criterion that all interactions need to be represented.	Johan Larsson	https://cran.r-project.org/web/packages/eulerr/vignettes/Introduction.html
UpSetR	A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets	Creates visualizations of intersecting sets using a novel matrix design, along with visualizations of several common set, element and attribute related tasks	Jake Conway	https://cran.r-project.org/web/packages/UpSetR/README.html http://biorxiv.org/content/early/2017/03/25/120600
Degust	An interactive web tool for visualizing Differential Gene Expression data	Take the time to digest and appreciate your Differential Gene Expression data	<u>David R. Powell</u>	http://victorian-bioinformatics-consortium.github.io/degust/
Glimma	*interactive HTML graphics	This package generates interactive visualisations for analysis of RNA-sequencing data using output from limma, edgeR or DESeq2 packages in an HTML page. The interactions are built on top of the popular static representations of analysis results in order to provide additional information.	Shian Su	https://bioconductor.org/packages/release/bioc/html/Glimma.html https://github.com/Shians/Glimma

Supplemental References

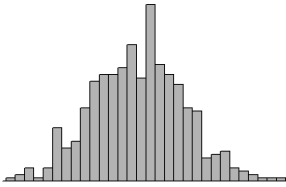
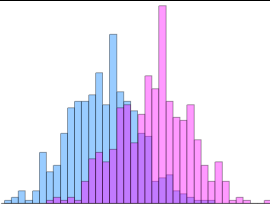
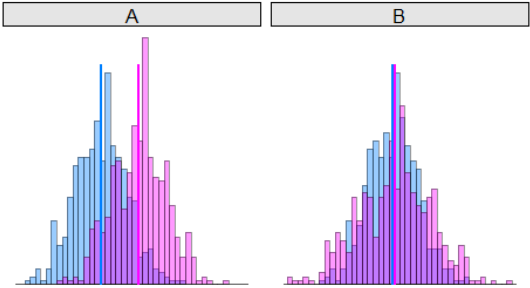
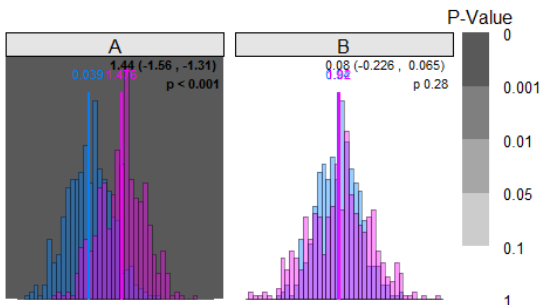
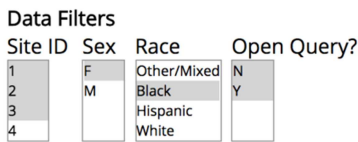
1. Anscombe FJ. Graphs in statistical analysis. *The American Statistician*. 1973 Feb 1;27(1):17-21.
2. Sarkar D. *Lattice: multivariate data visualization with R*. Springer Science & Business Media; 2008 Feb 15.
3. Wickham H. *ggplot2: elegant graphics for data analysis*. Springer New York. 2009;1(2):3.
4. Wong B. Visualizing biological data: data visualization is increasingly important, but it requires clear objectives and improved implementation. *Nature Methods*. 2012 Dec 1;9(12):1131-2.
5. Kelleher C, Wagener T. Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*. 2011 Jun 30;26(6):822-7.
6. Chen CH, Härdle WK, Unwin A, editors. *Handbook of data visualization*. Springer Science & Business Media; 2007 Dec 18.
7. Friendly M. A brief history of data visualization. *Handbook of data visualization*. 2008:15-56.
8. Fry BJ. *Computational information design* (Doctoral dissertation, Massachusetts Institute of Technology).
9. Kirk A. *Data Visualization: a successful design process*. Packt Publishing Ltd; 2012.
10. Heiberger RM, Holland B. *Statistical analysis and data display: an intermediate course with examples in R*. Springer; 2015 Dec 23.
11. Yau N. *Data points: Visualization that means something*. John Wiley & Sons; 2013 Mar 25.
12. Park, L. Visualization – Courses [Internet] Seoul National University [cited 2017 Mar 1] Available from <https://www.lucypark.kr/courses/2015-dm/visualization.html>
13. Cleveland WS. *Visualizing data*. Hobart Press; 1993 Sep 1.
14. Wilkinson L. *The grammar of graphics*. Springer Science & Business Media; 2006 Jan 28.
15. Bertin J. *Semiology of graphics: diagrams, networks, maps*.
16. Harrower M, Brewer CA. ColorBrewer. org: an online tool for selecting colour schemes for maps. *The map reader: Theories of mapping practice and cartographic representation*. 2011 Apr 6:261-8.
17. Martijn Tennekes .tmaptools: Thematic Map Tools. [Internet] R package, version 1.2. [cited 2017 Mar 1] Available from <https://CRAN.R-project.org/package=tmaptools>
18. Burks AW, Jones SM, Wood RA, Fleischer DM, Sicherer SH, Lindblad RW, Stablein D, Henning AK, Vickery BP, Liu AH, Scurlock AM. Oral immunotherapy for treatment of egg allergy in children. *New England Journal of Medicine*. 2012 Jul 19;367(3):233-43.
19. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016 Sep 15;32(18):2847-9.
20. Wilkinson L, Friendly M. The history of the cluster heat map. *The American Statistician*. 2009 May 1;63(2):179-84.
21. Tang Y, Horikoshi M, Li W. ggfortify: unified interface to visualize statistical results of popular R packages. *The R Journal*. 2016 Dec 1;8(2):478-89.
22. – Gern JE, Calatroni A, Jaffee KF, Lynn H, Dresen A, Cruikshank WW, Lederman HM, Sampson HA, Shreffler W, Bacharier LB, Gergen PJ. Patterns of immune development in urban preschoolers with recurrent wheeze and/or atopy. *Journal of Allergy and Clinical Immunology*. 2017 Jan 13.
23. Bostock M, Ogievetsky V, Heer J. D³ data-driven documents. *IEEE transactions on visualization and computer graphics*. 2011 Dec;17(12):2301-9.
24. Murray S. *Interactive data visualization for the Web*. " O'Reilly Media, Inc."; 2013 Mar 11.
25. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R, 2015. URL [http://CRAN.R-project.org/package= shiny](http://CRAN.R-project.org/package=shiny). R package version 0.11. 2015 Oct.
26. Sievert,C., Parmer,C., Hocking,T., Chamberlain,S., Ram,K. et al. (2016) plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 4.5.6.

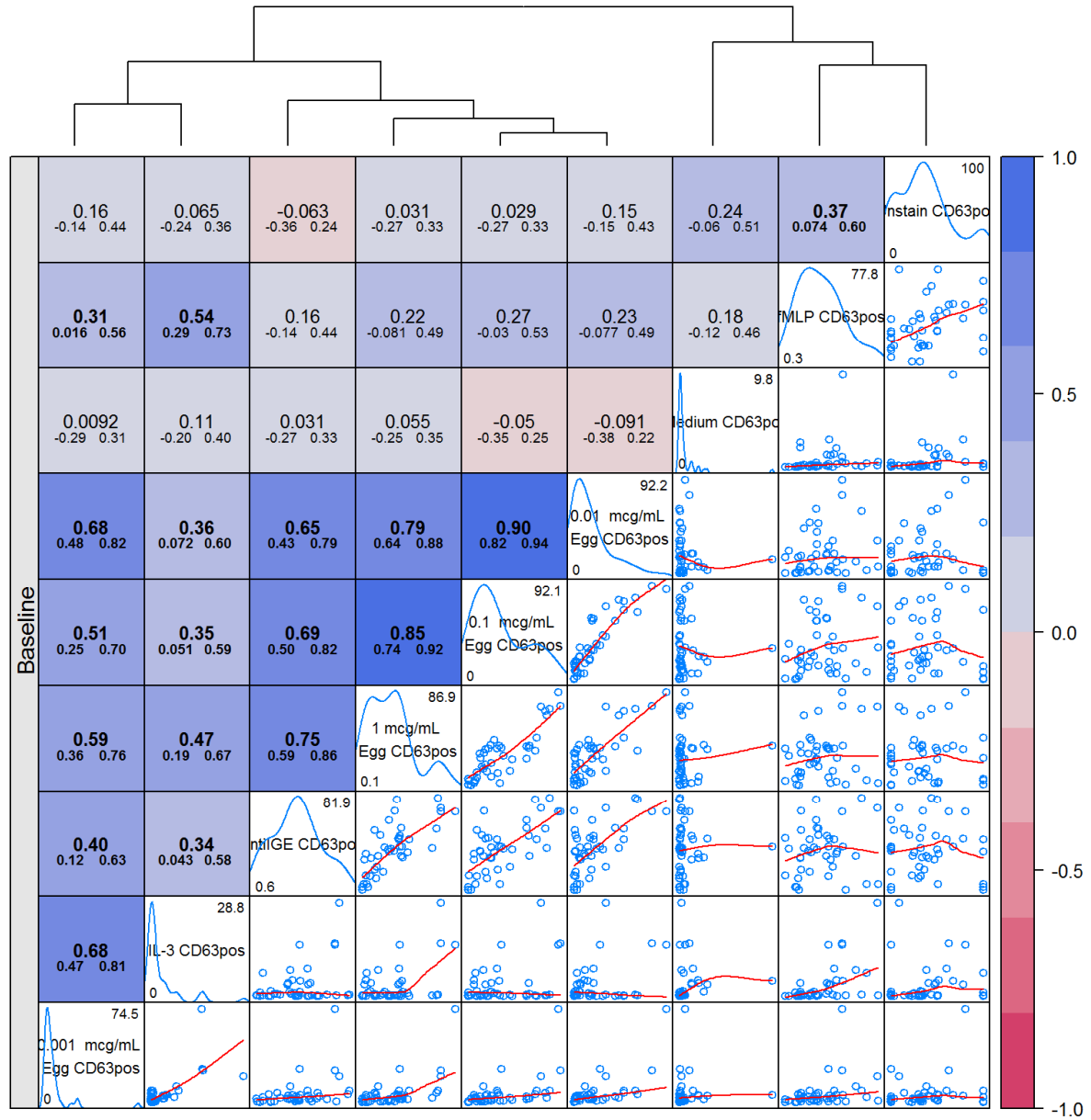
27. Su S, Law CW, Ah-Cann C, Asselin-Labat ML, Blewitt M, Ritchie M. Glimma: interactive graphics for gene expression analysis. bioRxiv. 2017 Jan 1:096107.

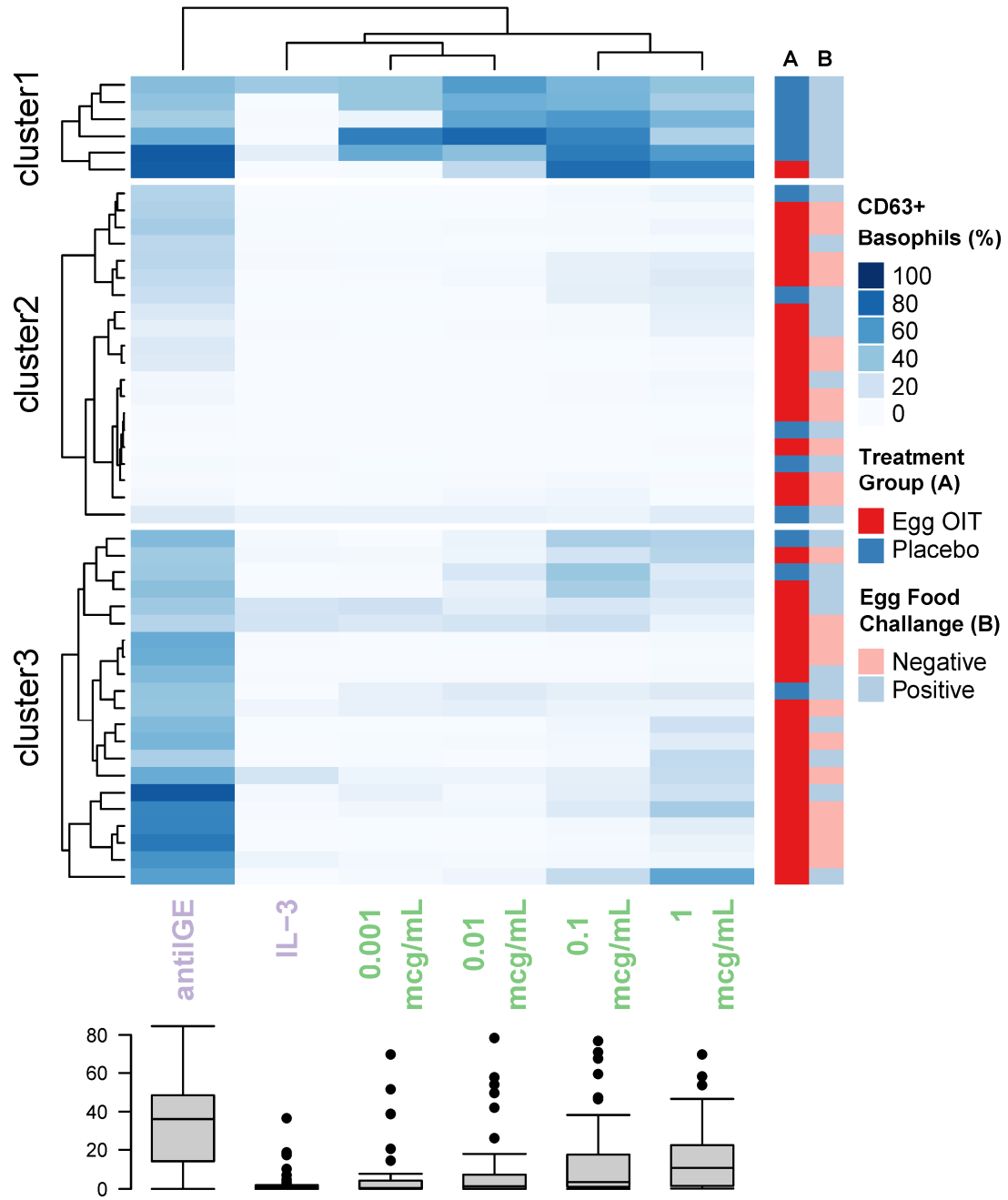
28. Powell, DR. Degust : Perform RNA-seq analysis and visualization. Victorian Bioinformatics Consortium, Monash University [cited 2017 Mar 1] Available at <http://vicbioinformatics.com/degest/index.html>

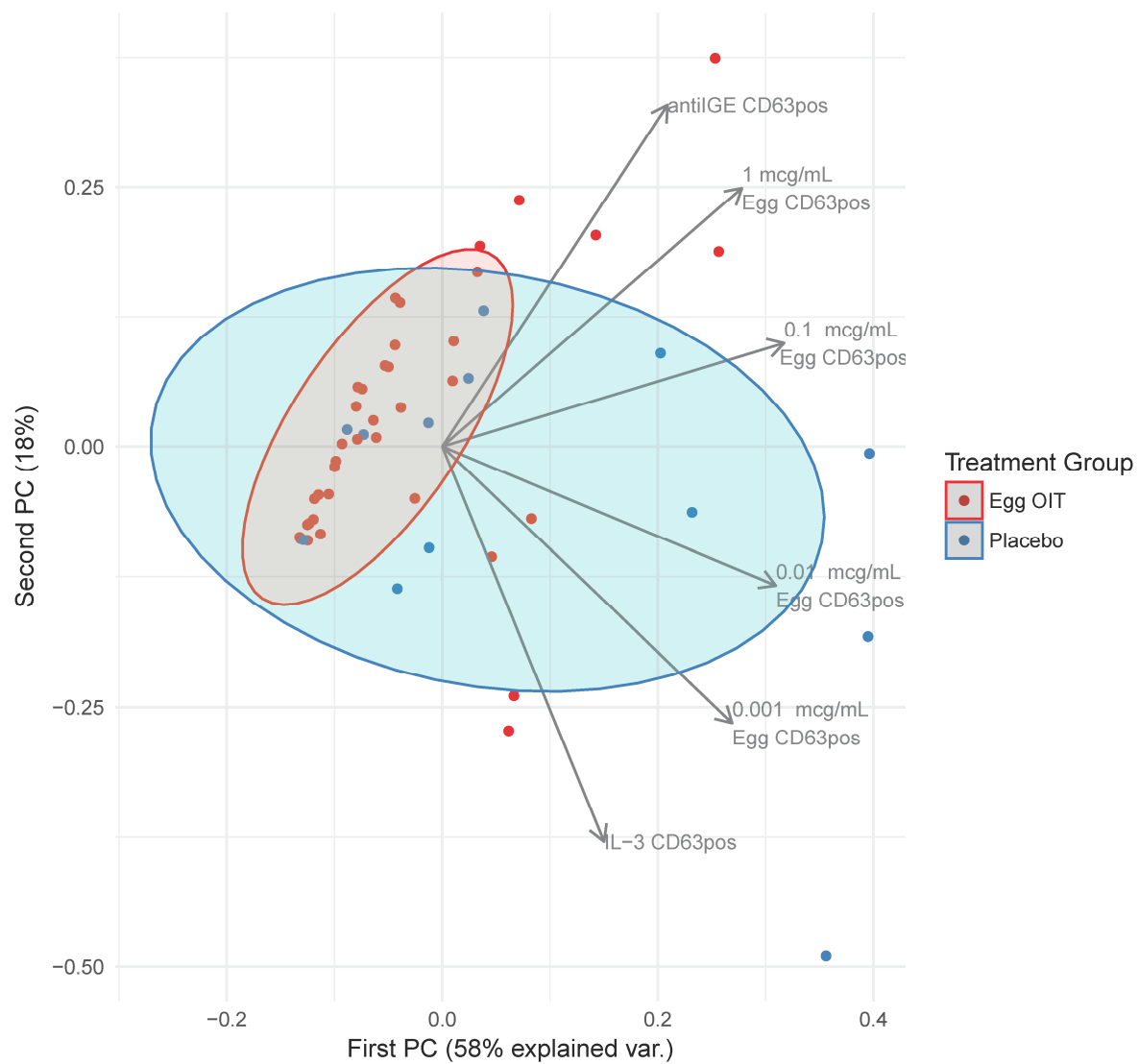
ACCEPTED MANUSCRIPT

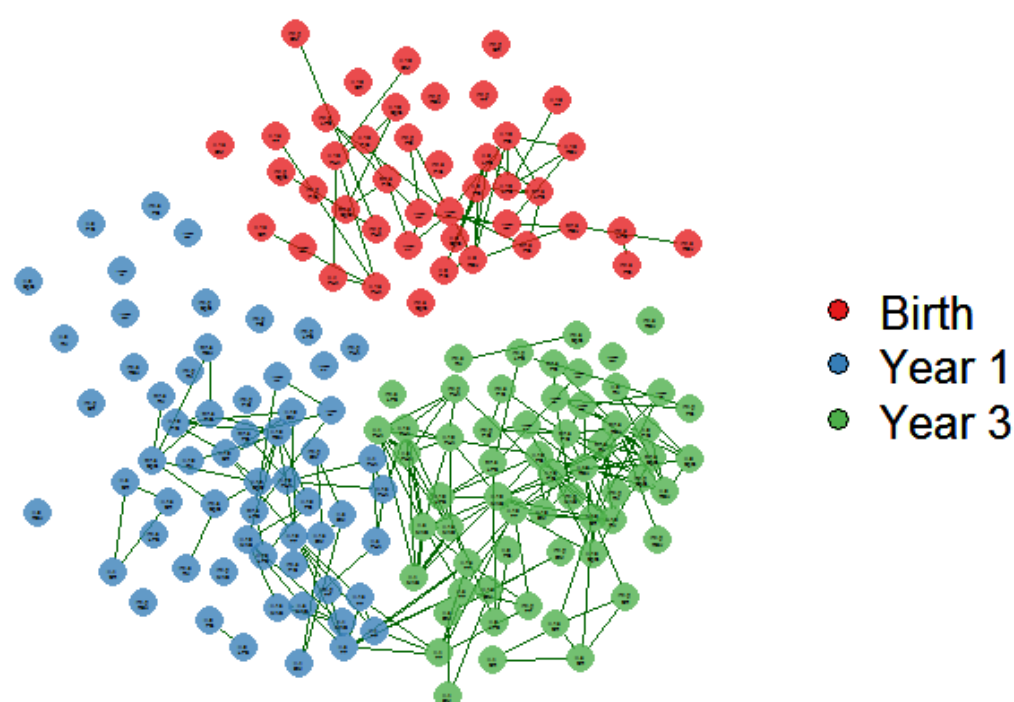
ACCEPTED MANUSCRIPT

Steps	Considerations	Mechanism	Examples
1	Start with raw data	Dot plot	
2	Aggregation Summarization Transformation	Histogram Density plot Boxplot Log, square root	
3	Comparison	Groupings	
4	Multivariate	Facets Lattice	
5	Annotation	Axis Legends Color Schema Statistical Test Reorder	
6	Interactivity	Explore	

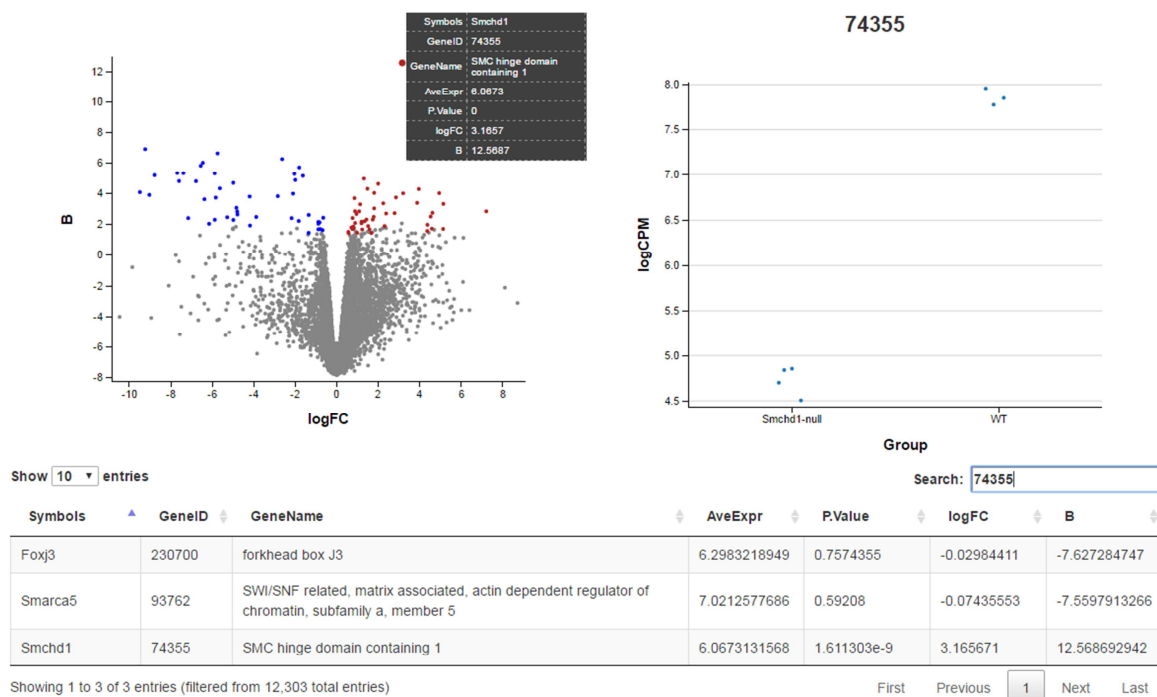






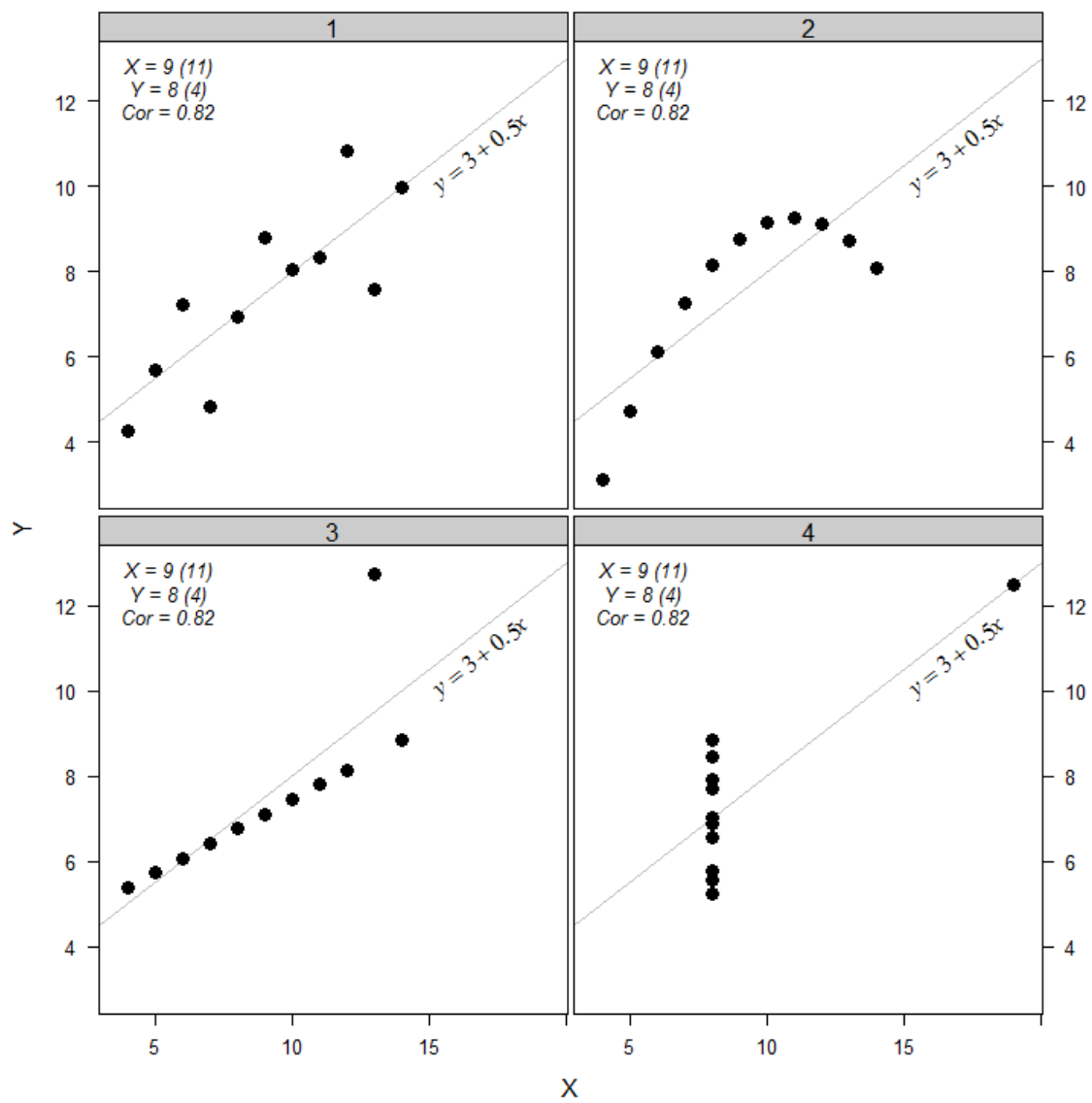


a.



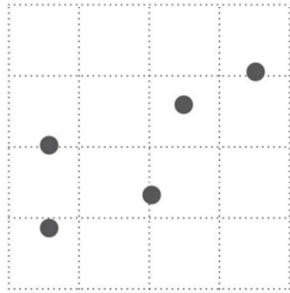
b.





Position

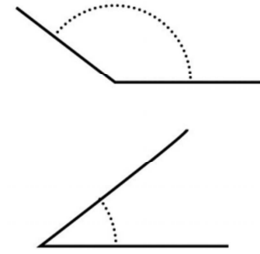
Where in space the data is

**Length**

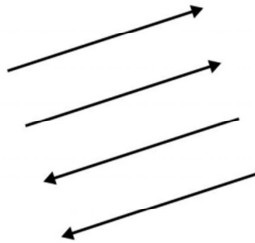
How long the shapes are

**Angle**

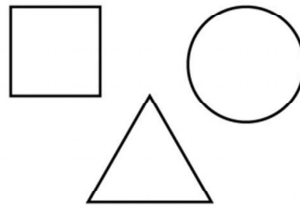
Rotation between vectors

**Direction**

Slope of a vector in space

**Shapes**

Symbols as categories

**Area**

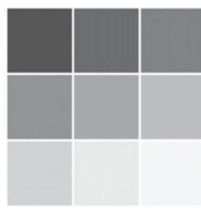
How much 2-D space

**Volume**

How much 3-D space

**Color saturation**

Intensity of a color hue

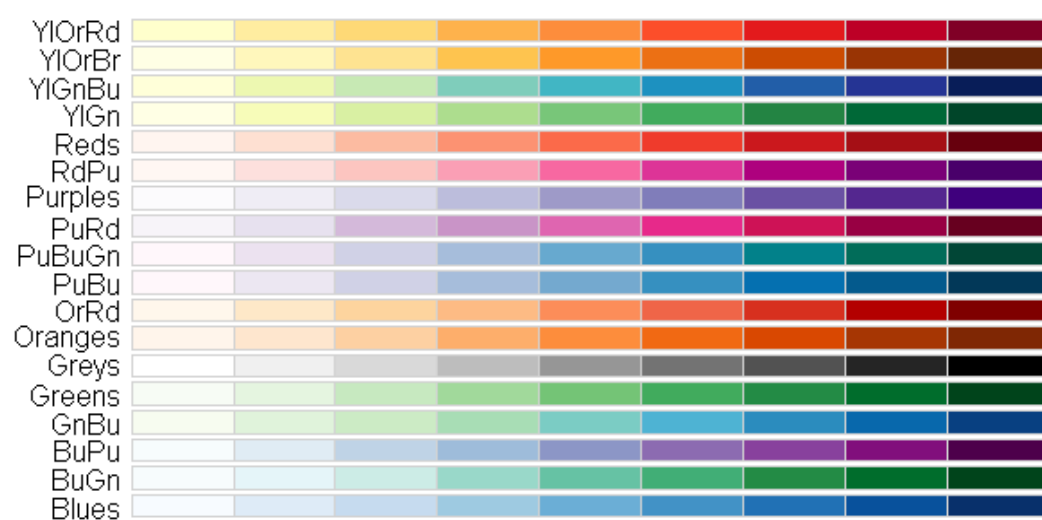
**Color hue**

Usually referred to as color

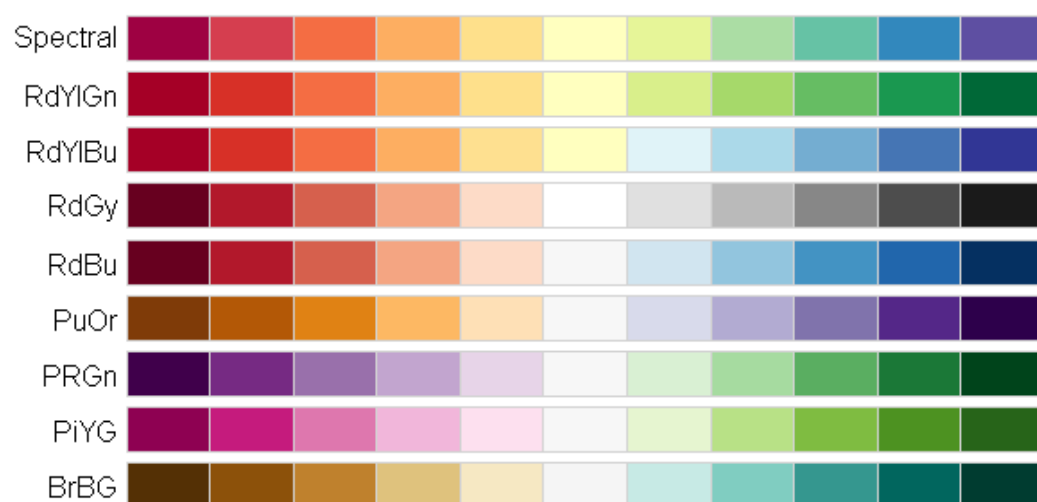
**FIGURE 3-3** Visual cues

Visual encoding by data type

	Quantitative		Ordinal		Nominal	
More Accurate ↑ ↓ Less Accurate	Position		Position		Position	
	Length		Density		Hue	
	Angle		Saturation		Density	
	Slope		Hue		Saturation	
	Area		Length		Shape	
	Density		Angle		Length	
	Saturation		Slope		Angle	
	Hue		Area		Slope	
	Shape		Shape		Area	

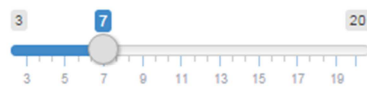




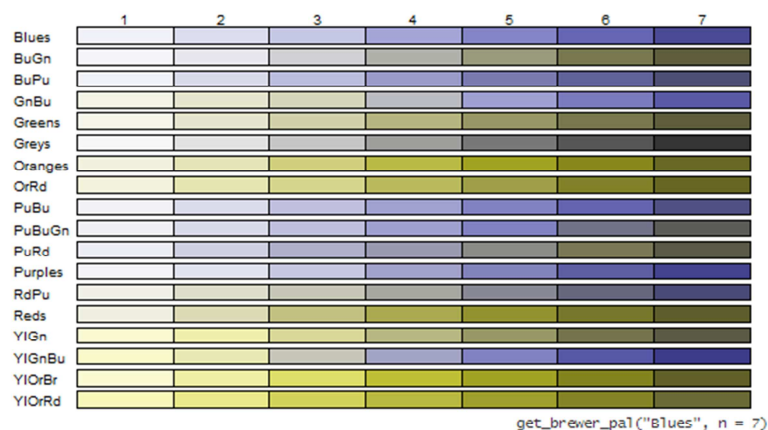
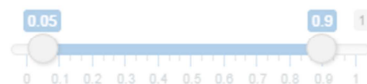


Sequential

Number of colors

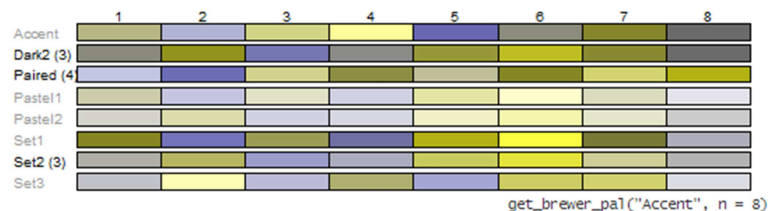


Contrast range

☒ Automatic

Categorical

Number of colors

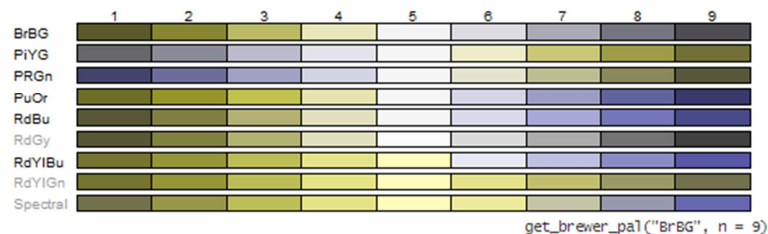
☒ Stretch

Diverging

Number of colors



Contrast range

☒ Automatic

Options

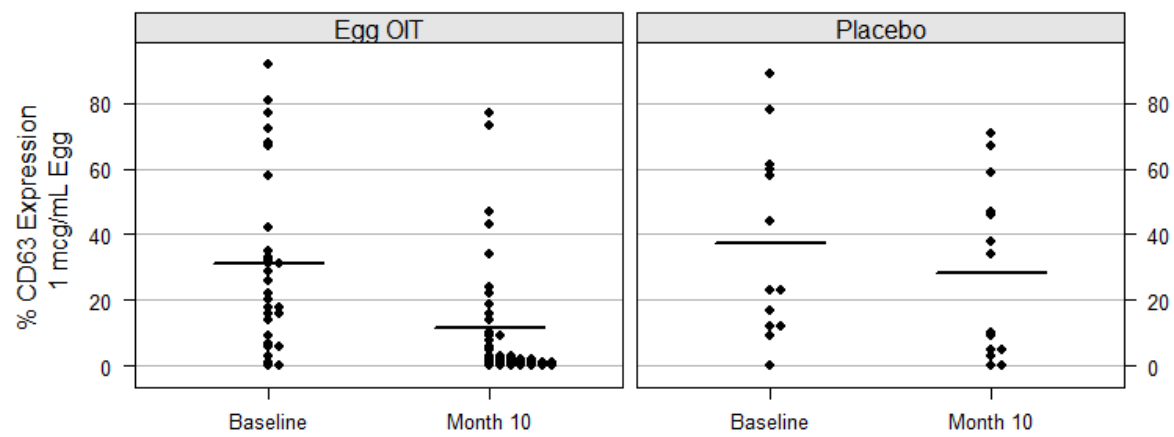
☐ Print color values

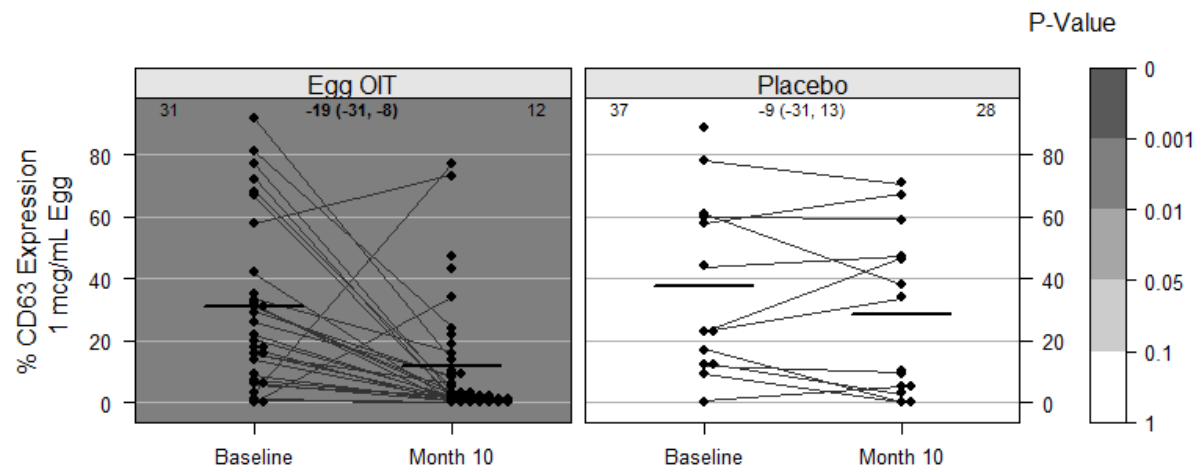
Code generator

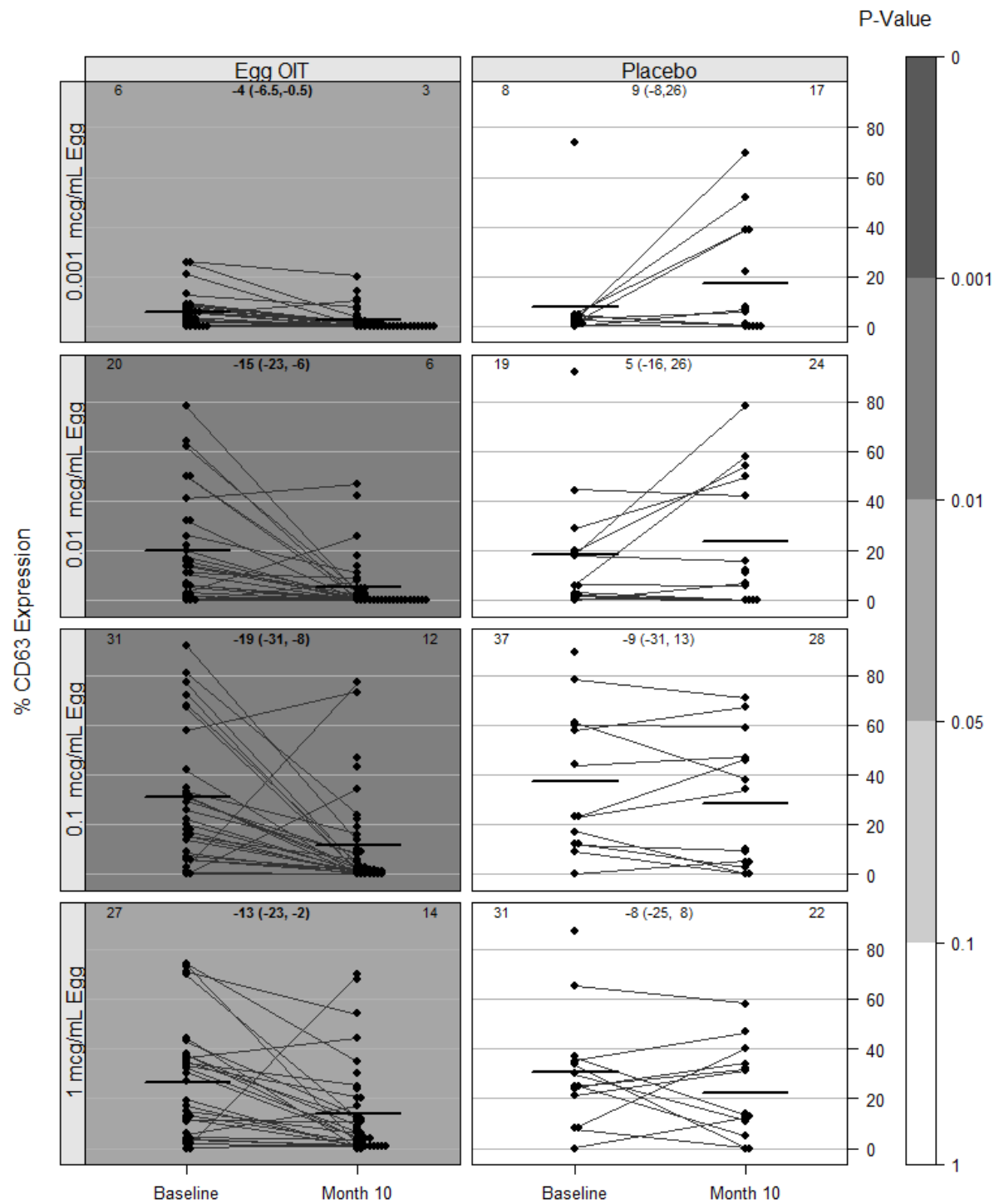
- ☒ Direct code
- ☐ tmap layer function code

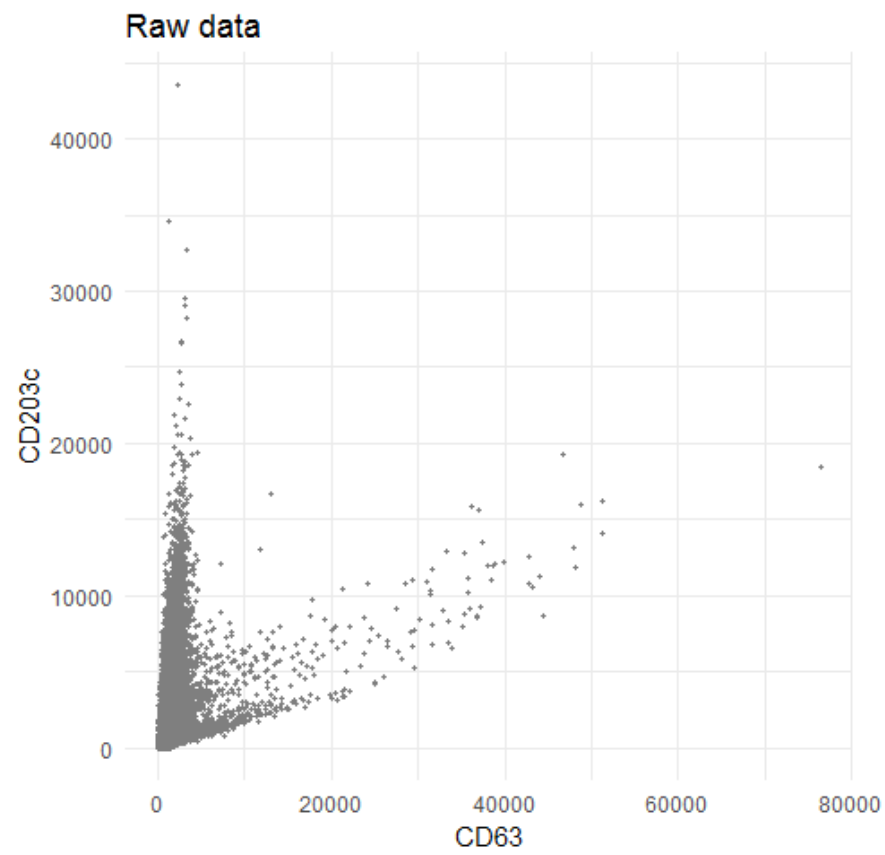
Color blindness simulator

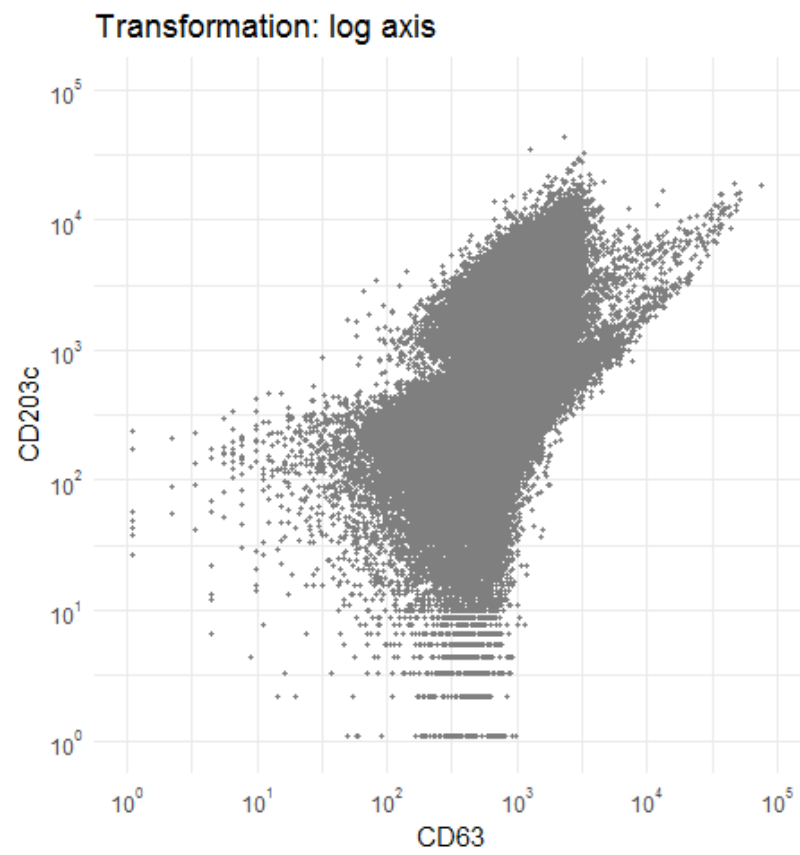
- ☐ Normal
- ☒ Deuteranopia
- ☐ Protanopia
- ☐ Tritanopia

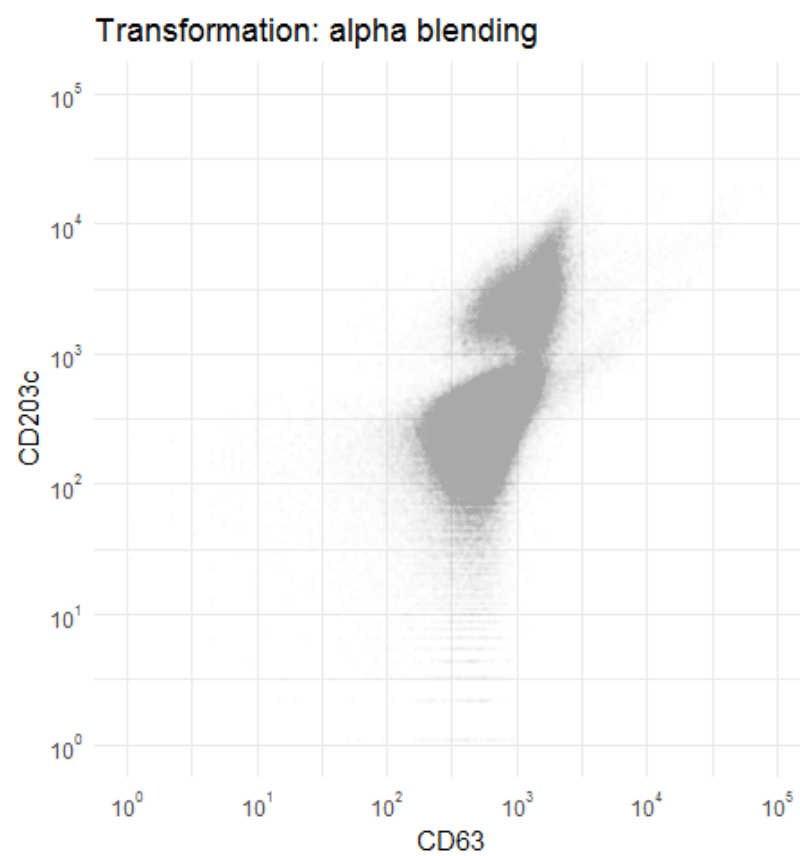




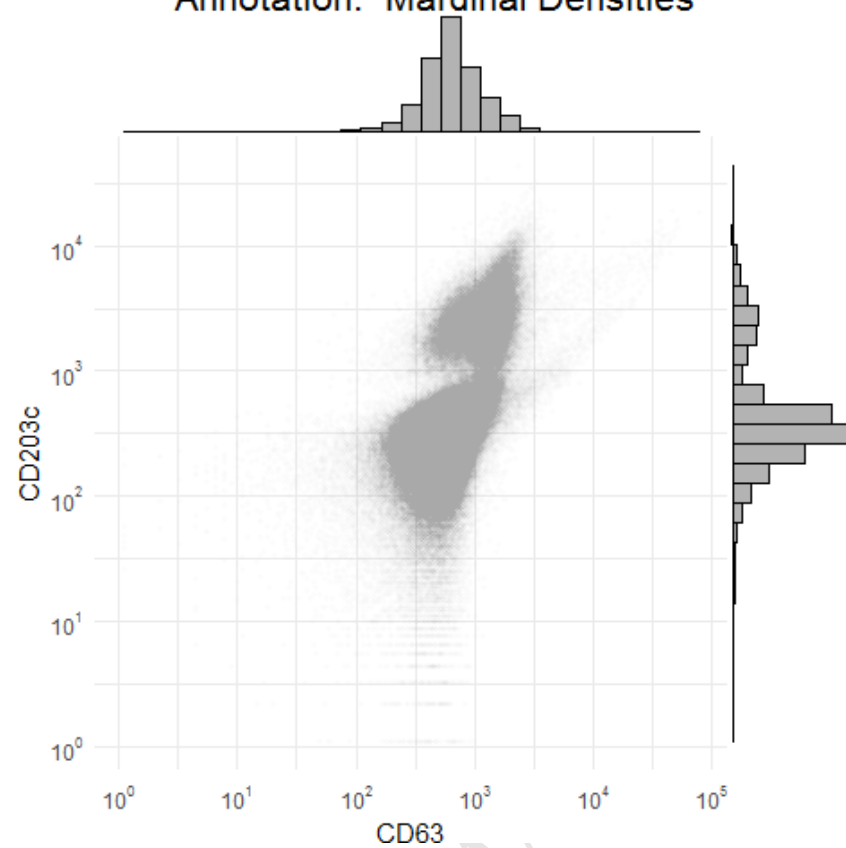


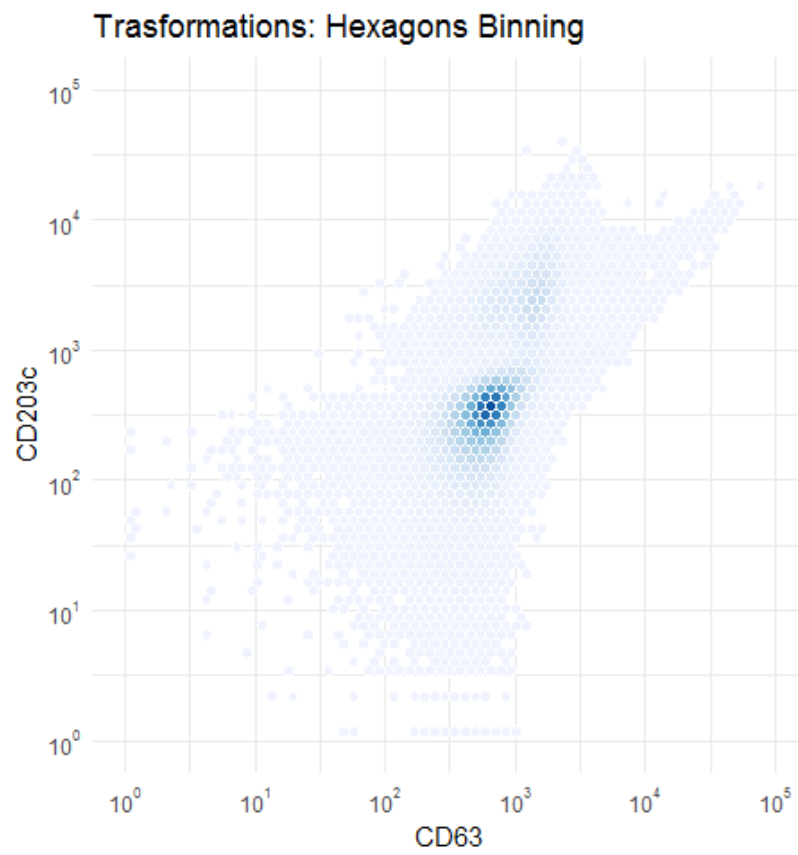


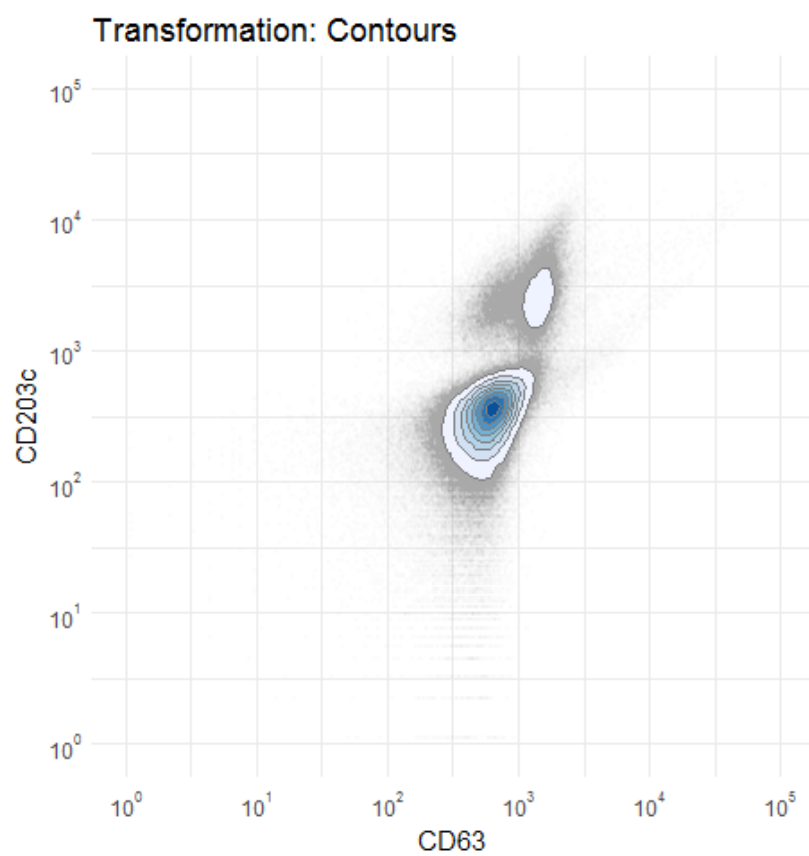




Annotation: Marginal Densities







Expand: Facets

