Heat^{2}map: enriched differential gene expression heatmap

Anton Zoubarev

CPSC 533C: Final Report

ABSTRACT

Differential gene expression heatmaps are used to visualize changes in gene expression under different experimental conditions. In some application displaying a p-value associated with each such measurement is important in interpreting the results. We propose a visual encoding based on a traditional differential gene expression heatmap that adds the ability to visualize p-values.

1 INTRODUCTION

Heatmap (matrix visualization) is a popular information visualization technique used in various domains including differential gene expression visualization. We briefly describe the domain, data and tasks performed using this type of heatmap.

1.1 Domain

In a typical gene expression experiment some factor is manipulated (for example, a drug is given) or some factor is observed (for example, age). The snapshots of gene expression levels under these conditions are then acquired using different methods including microarray assays or RNA sequencing.

Differential gene expression is obtained by applying a statistical test on the gene expression level measurements between two sets of samples (baseline (or control) and contrast (or condition)). The output of such test is a direction and a magnitude of change in gene expression, as well as a p-value (a measure of statistical significance) associated with the statistical test. Because of the wide range of possible gene expression levels across genes the magnitude of change (or fold change) is usually log transformed.

1.2 Data

Typical dataset is a table with a row for each gene. The columns correspond to a different comparison (baseline vs. condition). Each cell of this table contains the following attributes: direction of change in gene expression, log transformed magnitude of change and a p-value. These attributes are summarized in the following table:

Attribute	Type	Range
Direction of change	Binary	Up or Down
Magnitude of change	Quantitative	0..5 (most values are within this range)
P-value	Quantitative	0..1

Dataset size is up to a few hundreds of rows and up to a few hundreds of columns.

The data is transformed as follows: direction of change and magnitude are combined to form log fold change attribute which ranges from -5 to 5 (for the majority of the data, values outside of this range are possible). It is further binned into 11 bins. Bins are 1 unit in size except for bins at the edges of the range which are unbounded intervals: '5.5 and more' and '-5.5 and less'. The range of p-values is binned into 5 bins: 0.0001 or less, 0.00001~0.0001, 0.0001~0.001, 0.001~0.01, 0.01~0.05, and anything more than 0.05. These intervals roughly correspond to categories: not significant, somewhat significant, significant and very significant.

1.3 Tasks

A common differential gene expression heatmap is visualized as a matrix with (typically) rows representing genes and columns representing samples or groups of samples for a given condition. The traditional visual encoding is to use colour hue as indication of direction of change (green being down and red being up) while saturation and/or lightness are used to encode magnitude of change.

The tasks performed on this type of heatmap include: communicating results of the experiments, finding differences, similarities, variability between groups of genes or groups of conditions, clustering/sorting/ordering data to find relationships between genes and conditions.

The tasks correspond to questions researchers may ask, such as: how do results of my experiment compare to other similar experiments? Are they surprising or align with existing results? Given results of my experiment are there any other similar experiments? Do experiments that manipulate similar things produce similar results?

When looking at data from multiple experiments, it is important to also show p-values associated with each cell as they can greatly influence interpretation of the results. Different experiments produce results of different quality. Less significant results are not as important as significant ones. However, multiple experiments with relatively high p-values taken as a whole can increase confidence in a certain observation so it is important not to filter them out.

1.4 Desired Characteristics

We have identified following criteria that are necessary to produce useful visualization for our purposes:

1. It should support wide range of zoom levels and rectangular cells. If entire heatmap fits into the screen it greatly simplifies the tasks. So visual encoding that is stable at small cell sizes is preferred. Another issue arises when one of heatmap dimensions is much greater than the other which necessitates squeezing heatmap cells along one direction producing rectangular cells. It is important for visual encoding to work well even when the cells are rectangular (i.e. not square).

2. It should degrade gradually. This is related to performance when zooming out. As the cell size decreases, the less important attributes should get dropped without affecting important ones. Ideally, even with pixel-sized cells the direction of change and approximate measure of its magnitude should be readable.

3. It should be familiar to first time users who have seen differential gene expression heatmaps in the past.

4. It should make cells with low p-values more visible. Traditional heatmaps emphasize high fold change but a subtle change in the expression at a very low p-value is an important result. In other words, the benefit of showing p-value should be two-fold: to enhance interpretation of the results and to ease finding of significant results.

5. It should be colorblind safe.

2 Previous Work

Related work in this area includes work by Clemens Holzhuter et al[1].

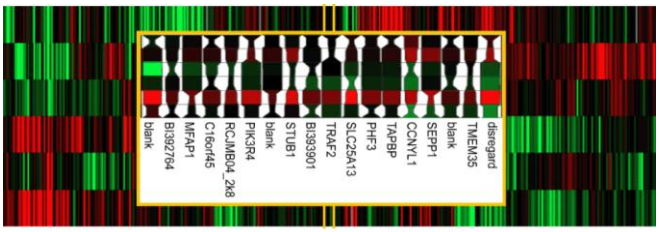

Figure 1: Use of cell size channel to encode p-value on a differential gene expression heatmap.

They use traditional visual encoding for underlying heatmap with p-value encoded by the width of the color band within the cell (the better the p-value the wider the band). [Figure 1]

ArrayExpress[2] shown on Figure 2 uses colour saturation to indicate p-values while using colour hue to indicate direction. The magnitude of change is ignored. The numbers indicate the number of experiments in which differential expression was observed.

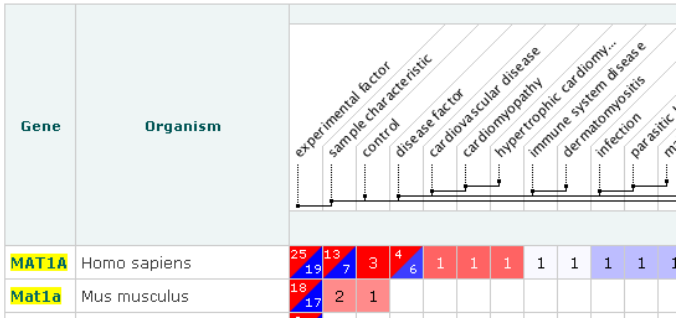

Figure 2: Use of saturation channel to encode p-value in ArrayExpress.

3 Visual Encoding

We address the problem of displaying p-values by proposing a visual encoding that takes it into account. In this section we describe our solution along with justification of some of the design decisions and discussion of some of the strengths and weaknesses of the proposed visual encoding.

3.1 Visual Encoding

The proposed visual encoding is summarized in Figure 3.

We base our encoding on a traditional heatmap layout and colour scheme. This choice is made to ensure that the result looks familiar to a wide range of researches who are accustomed with traditional heatmaps. The only difference being our colour scheme passing through white instead of black as a zero value.

We use a divergent colour scale taken from colorbrewer2.org [5] (which uses combination of lightness and saturation) to visualize log fold change. The colour scale used has perceptual ordering and is colour-blind safe.

P-value is shown using opacity of the black inner rectangle in the center of each cell. Even though it is implemented by varying opacity it uses lightness visual encoding channel.

3.1.1 Channel Interference Issue

This choice of visual encoding channels creates channel separability issue. Since p-value is encoded relative to the cell background the accuracy of p-value reading suffers. Figure 4 demonstrates the accuracy issue: if we look at inner rectangles along diagonals they look similar even though they encode different p-values making it difficult to compare p-values visually.

Figure 4: Some inner rectangles along the diagonals are perceived to be the same even though they encode different p-values.

However, the use of relative visual scale for p-value has the following positive effect. Since we do not want high p-value cells to visually pop-out no matter what background colour (magnitude of change) is, the use of relative scale helps us reduce contrast (first two rows of Figure 3) between inner rectangle and cell background which in turn de-emphasizes high p-value cells preventing them from popping out (criterion 4 in section 1.4).

Further testing can help in determining if this trade-off between the accuracy of judging p-value and reducing pop-out effect for low confidence results is justified.

Another issue that arises at small cell sizes is that inner rectangle colour affects overall cell colour perception. The cell appear darker (greater log fold change) than it should.

3.2 Expressiveness and Effectiveness

We further discuss visual encoding with respect to expressiveness and effectiveness principles. [3]

3.2.1 Expressiveness

Expressiveness principle is satisfied by matching each data type to an appropriate visual encoding channel. Colour scale for log fold change is divergent and has a perceptual ordering. P-value uses inner rectangle's lightness channel which is appropriate for quantitative attribute. It has a perceptual ordering as well.

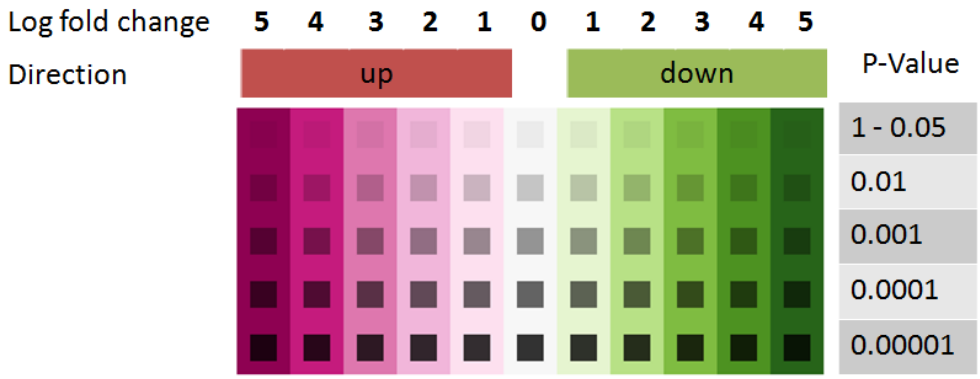

Figure 3: The magnitude of change is gene expression is encoded using green to red colour scale. P-values are encoded by the opacity of inner black rectangle using lightness visual channel.

Attribute's value ranges match the dynamic ranges of encoding channels. However, since p-value uses one of the channels (lightness) used by log fold change some adjustment to the colour scheme may need to be made to reduce the conflict. One way to accomplish this is to use a narrower lightness range to encode the fold change freeing up the rest of the range for p-value.

3.2.2 Effectiveness

An interesting point is that the order of importance of the attributes is not constant across all visualization states. It is clear that statistically significant results should be more prominent visually suggesting that p-value is the most important attribute but at the same time it is not informative if it is shown alone at zoom-out level (small cell size).

In other words, the order of importance of attributes changes depending on the zoom level. For small cell sizes it is: the direction of change, magnitude of change and p-value. For larger cell sizes it is: direction of change with p-value and fold change sharing the second place. We accommodate this change in the ordering by dropping p-value at certain zoom level (Section 3.3.1). The relative importance of p-value can be tweaked by changing relative size of inner rectangle which we further discuss in Section 3.4.

3.3 Behavior

This section briefly describes how visual encoding responds to different zoom levels and going into emphasize p-value mode.

3.3.1 Zooming behavior

We stop drawing inner rectangle after certain size limit is reached. Since at small cell sizes drawing inner rectangle is either not practical or causes interference with cell background colour we drop p-value attribute in favour of preserving accurate representation of a fold change and direction.

3.3.2 Emphasize p-value mode

When cell size is too small or when accurate p-value judgements have to be made the visualization can be redrawn in 'emphasize p-value' mode. In this mode the relative size of inner rectangle is increased and encoding channel is switched to use white to black gradient scale. The transition could be animated (future work). A fragment of the resulting heatmap is shown on Figure 5.

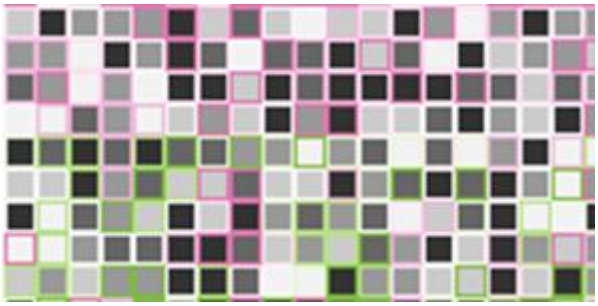

Figure 5: 'Emphasize p-value mode' is achieved by increasing the size of the inner rectangle and changing its visual encoding to use white to black gradient. This allows for more precise comparison of p-values.

3.4 Parameters

Current implementation has following parameters that can be tuned:

- Relative size of the inner rectangle (as a ratio to cell size)
- Relative size of the inner rectangle in the emphasize p-value mode (as a ratio to cell size)
- Cell size at which we should stop drawing inner rectangle (as a sum of cell height and width)

.

3.5 Implementation

The implementation is in JavaScript and does not use any external libraries.

4 Discussion

By doing some ad-hoc testing the visualization seems to hold up to the criteria outlined in Section 1.4 although more formal evaluation is required.

We have compared heatmaps with and without p-value. Heatmaps without p-value are dominated by high magnitude change cells. Adding p-value helps significant but low level expression change cells to be visually prominent.

Figure 6 demonstrates visualization at different zoom levels.

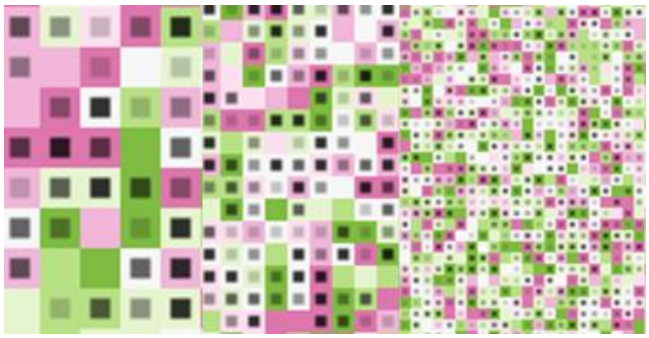

Figure 6: Visualization across multiple zoom levels.

Figure 7: Visualization with cells squeezed horizontally and vertically.

Figure 7 shows visualization with rectangular cells.

4.1 Search Process

By experimenting with various visual encodings some of which are described in the Appendix I have made a few observations.

I have found that symmetry and constancy of mark's position within the cell are very important factors contributing to how heatmap will look as a whole. Specifically, encodings using tilt or angle, as well as encodings using position of the mark within the cell produce heatmaps with visual artifacts that distort heatmap matrix layout.

With the added constraint of symmetry most visual encoding channels become unavailable limiting possible number of choices.

4.2 User Feedback

We have integrated the proposed visual encoding with a slightly different colour scale into a test version of Gemma[4] (a web-based system that aggregates and analyses thousands of gene expression experiments and allows performing differential gene expression searches on them). The result is shown on Figure 8.

Some early feedback is positive so far. One suggestion was that having 5 bins for p-values might not be necessary and that 3 would be sufficient with an interactive way to select ranges for these bins. This can alleviate issues with accuracy of perception of p-values described previously.

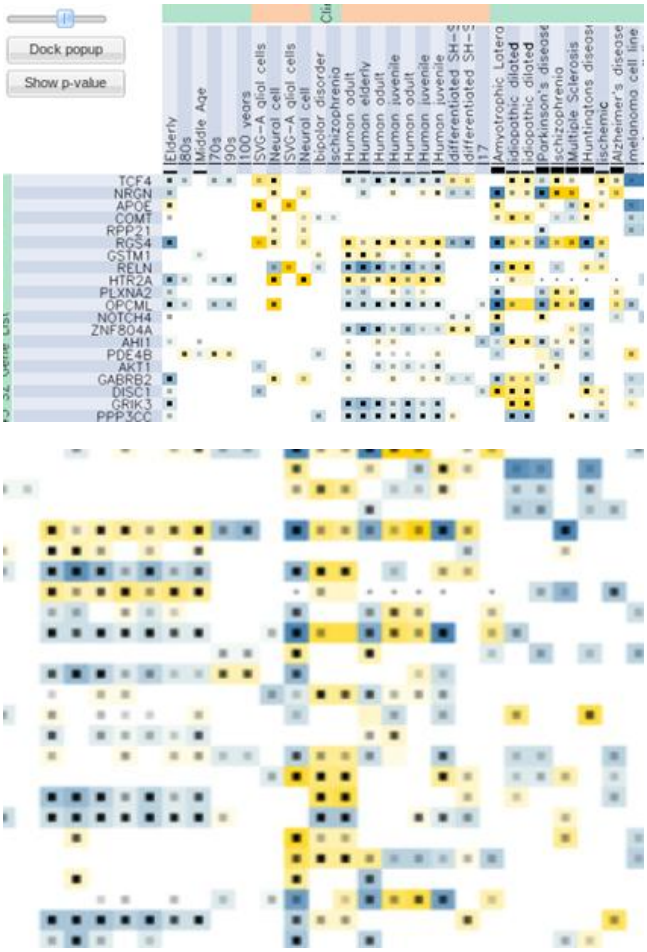

Figure 8: A similar to proposed encoding is implemented in Gemma. Cells with low p-values pop-out visually (solid inner rectangle) even if the magnitude of change is small (white cell background).

5 Future Work

Future work focuses on evaluation and parameter tuning. The main question is: does the encoding enrich the heatmap and does it enrich it in the right way. Another question is: how does it compare with other solutions.

5.1 Evaluation

I would like to design a study to evaluate this visual encoding. The questions to ask are what to measure and what to compare it with. It would be interesting to see how this solution behaves at different zoom levels compared to Clemens Holzhuter et al [1] solution. Another good candidate to compare with is two heatmaps side by side each encoding different attribute.

Also, it would be interesting to see how it compares with a simple filtering by p-value approach that just hides cell bellow certain threshold. It could very well be a better solution to the problem of making certain classes of cell more visually salient.

5.2 Parameter Tuning

It is interesting to see how adjusting parameters influences the visualization. There could be some optimal settings that can be found that produce the best results. It may also be necessary to add more parameters.

5.3 Missing Features

One important omission is the lack of handling of missing values. I plan to represent those with a little 'x' mark inside the cell.

6 Lessons Learned

If had to this project all over again I would focus more on analysis and evaluation of solutions that might work rather than on a long and exhaustive search of the solution space.

This project is also lacking deeper research into related work. I would spend a significantly more time on placing my solution within the context of previous work.

I could have also benefited from having a working prototype displaying real data as soon as possible.

References

[1] Holzh, Clemens. Enriched Heatmaps for Visualizing Uncertainty in Microarray Data.

[2] Alvis Brazma et al. ArrayExpress—a public repository for microarray gene expression data at the EBI.

[3] Jock Mackinlay. "Automating the Design of Graphical Presentations of Relational Information." ACM Trans. on Graphics (TOG) 5:2 (1986), 110–141.

[4] Not published yet.

[5] Brewer, Cynthia A., http://www.ColorBrewer2.org, accessed December 1st.

7 Appendix

Figures 9 to 12 show different visual encoding choices to encode p-value: line position (Figure 9), cell area (Figure 10) and inner rectangle size (Figure 11). The encoding on Figure 12 uses pseudo-3D. The pyramid pointing up or down encodes direction of change with the lightness channel encoding p-value.

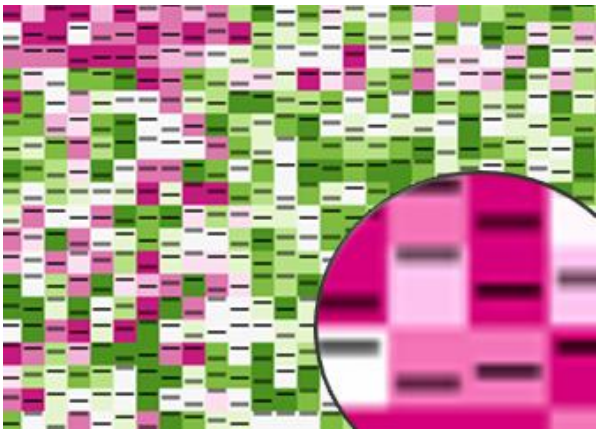

Figure 9: Using line position to encode p-value. The higher the line the better is the p-value.

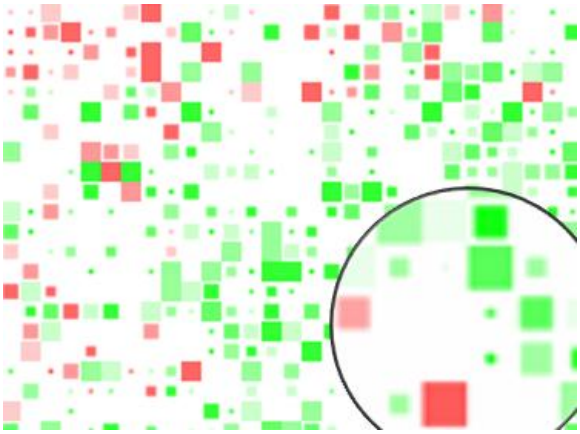

Figure 10: Using area to encode p-value.

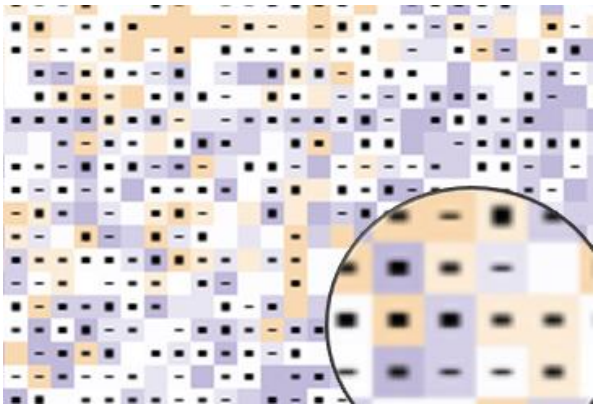

Figure 11: Using inner rectangle area to encode p-value

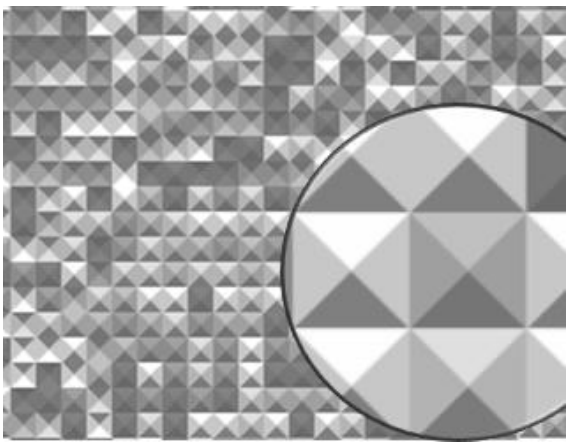

Figure 12: The pyramid pointing away the screen (up) or into the screen (down) encodes the direction of change. Lightness channel is used to encode p-value.