POINTS OF VIEW

Heat maps

Heat maps are useful for visualizing multivariate data but must be applied properly.

Heat maps represent two-dimensional tables of numbers as shades of colors. This is a popular plotting technique in biology, used to depict gene expression and other multivariate data. The dense and intuitive display makes heat maps well-suited for presentation of high-throughput data. Hundreds of rows and columns can be displayed on a screen. Heat maps rely fundamentally on color encoding and on meaningful reordering of the rows and columns. When either of these components is compromised, the utility of the visualization suffers.

Using color to represent numbers in a table is an old idea; an example is from 1873 by the French economist Toussaint Loua (**Fig. 1a**)[1]. Color is a relative medium and can be unreliable when used to represent discrete values. Whereas one can be strict in translating a number to a color, the resulting color may not be perceived as intended; the same color may look different depending on the color of neighboring cells (see August 2010 column)[2]. Data visualization relies on communicating with images, and the discordance between what we 'should' see and what we 'actually' see needs to be considered in designing and selecting effective representations.

Heat maps are typically used to show a range of values, and designing an appropriate color map is essential to highlight one or both ends of that spectrum. A divergent color gradient defined by three hues (for example, from blue to white to red) will make the low and high ends of the range visually distinct. In contrast, a gradient created by varying the lightness of a single hue is effective at highlighting one extreme. A grayscale with range of 10–90% black works well as a linear color map. Avoid red-green as a color combination because it limits accessibility to information for colorblind individuals.

When used with suitable color scales, clustering can dramatically affect our ability to see structure in heat maps. After rows and columns are arranged according to similarity, previously undetectable
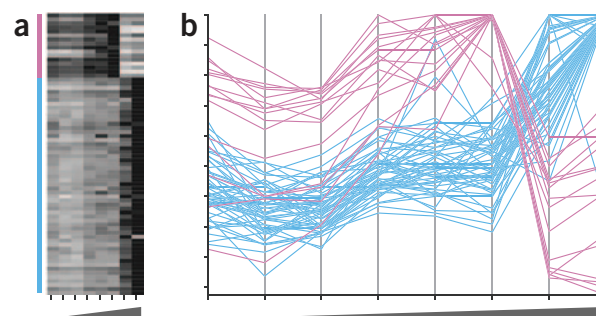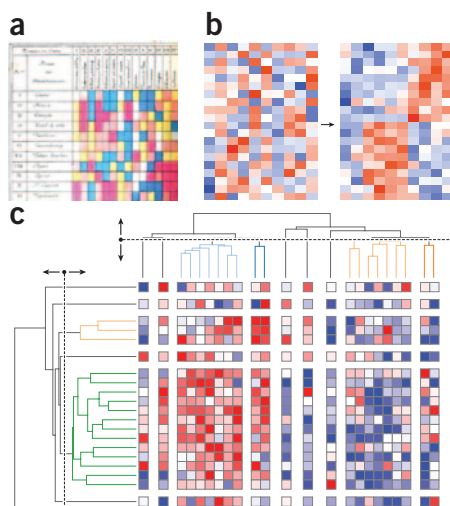
Figure 2 | Parallel coordinate plots. (**a**) Gene expression data shown for two groups of profiles (blue and purple). (**b**) The data from **a** with each row plotted as a profile and each column as a vertical axis.

patterns can become obvious (**Fig. 1b**). Hierarchical clustering is one technique for reordering matrices that creates several display challenges. First, because there are 2^{n-1} possible arrangements for n rows or columns related by a cluster tree, a static heat map is only one of many possible outcomes. Second, clustering creates useful relationship information captured in the cluster tree typically displayed on the sides of the matrix. The linear ordering may require that some distantly related rows or columns be placed next to one another, thus obscuring the relationships reflected in the cluster tree. GENE-E is software from the Broad Institute (http://www.broadinstitute.org/cancer/software/GENE-E/) with the ability to impart the useful information from the periphery to the matrix (**Fig. 1c**). These 'gap maps' enable one to quickly hone in on color blocks that are deemed to be most related by hierarchical clustering.

Heat maps in which both rows and columns are clustered create blocks of similarly colored cells that are easy to spot. However, when data with inherent ordering of columns are visualized as heat maps (for example, those from time series or dose-response studies), clustering is only applied to the rows. With these types of data it is necessary to understand how the fluctuations in color sequence across a row relate to time or concentration. In such cases an effective plotting alternative is the parallel coordinate plot (**Fig. 2**). The reliance on spatial encoding not only enables more accurate reading of absolute values, complex trends are easier to understand as captured by an undulating profile graph than with color. Parallel coordinate plots are particularly well suited for highlighting small discrepancies between samples. As these parallel coordinate plots layer information, graphing data with more than a few dozen profiles will make it difficult to distinguish profiles.

Next month, we will look at high-dimensional data display and explore how additional information can be added to networks and heat maps.

Nils Gehlenborg & Bang Wong

1. Loua, M.T. *Atlas Statistique de la Population de Paris* (Imprimerie et Librairie de L'Ecole Centrale. Paris, France, 1873).
2. Wong, B. *Nat. Methods* **7**, 3 (2010).

Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute. Bang Wong is the creative director of the Broad Institute and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Figure 1 | Heat maps. (**a**) An example of a colored table from ref. 1. (**b**) Clustering brings like next to like items to reveal patterns in the data. (**c**) Adding gaps according to the hierarchical cluster tree helps emphasize relationships in the matrix.