POINTS OF VIEW

Plotting symbols

Choose distinct symbols that overlap without ambiguity and communicate relationships in data.

Scatter plots require us to visually assemble data point symbols into patterns so that we can understand the relationship between the variables. Symbols can therefore have a large impact on figure legibility and clarity. Well-chosen symbols mitigate the effects of data occlusion and maintain the visual independence of different data categories.

In plots with one data category, the primary concern is to minimize data occlusion caused by overlapping symbols. Here the open circle is the best choice. In contrast with other common geometric shapes (such as squares, triangles and diamonds), the intersection of a circle with another circle does not form an image of itself (that is, another circle) (**Fig. 1**). The benefit of the open form is that overlapping instances build up regions of denser ink on the page, which can be a practical substitute for density maps.

Multiple data categories should be encoded with distinct symbols that form strong visual boundaries (**Fig. 2a**). Symbols that have similar appearances can be easily missed on first inspection, especially in regions where symbols overlap. Insufficient symbol contrast can make

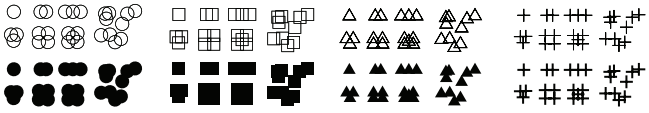

Figure 1 | The hollow circle is a flexible and robust plotting symbol.

it difficult to identify each data category. The most common shapes in plots—polygons—blend and lack distinctiveness. Luckily, user studies in symbol discrimination offer guidance for putting together a versatile symbol set[1,2].

If there is clear and simple distinction between data categories, it may be possible to use the first letter in the category name as a plotting symbol (**Fig. 2b**). This practice makes decoding figures easier because

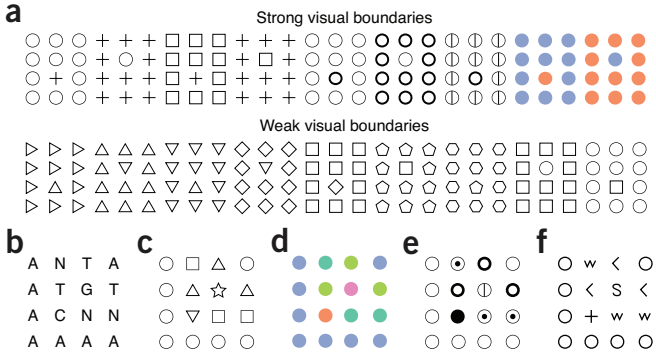

Figure 2 | Symbol diversity can be achieved by varying shape, fill or color. (**a**) Symbols that contrast with one another make good combinations. (**b**) Letters simplify legend lookups, but many appear the same (such as C/G, B/R/P and E/F/H). (**c**) Shapes are powerful discriminators—but beware that, for a given width, they may appear to have different sizes owing to differences in areas. (**d**–**f**) Color is one of the differentiators (**d**). For black-and-white applications, vary the fills for low data densities (**e**) and use texture symbols when overlap is high (**f**).

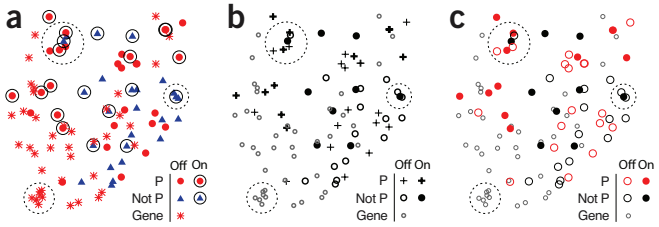

Figure 3 | Symbols should encode natural hierarchies in data to simplify legend lookup and help reveal patterns. (**a**–**c**) The choice of encoding three different gene types in **a** is nonintuitive[5] (for example, transcribed state is shown by a circular outline, repeating a shape already in use), and symbols overlap awkwardly (dotted regions). (**b**,**c**) Alternative symbol sets in black and white (**b**) and color (**c**).

the reader does not have to repeatedly refer to the legend, as long as the letters are visually distinct (for example, H, Q and X[2]).

Care should be taken that the shapes of plotting symbols appear to be the same size and have the same degree of complexity. For example, the five-pointed star draws considerably more attention than other symbols of the set in **Figure 2c** and may therefore bias readers to assign its category undue importance.

When available, color is a highly effective discriminator (**Fig. 2d**), but it should be used judiciously—its salience diminishes as the number of hues increases. Good color choices for data categories are the qualitative Brewer palettes (http://colorbrewer2.org/). These have been selected for their desirable perceptual properties. In the event that your communication will be reproduced in black and white, we suggest using symbols with a variety of fills for data sets with low overlap (**Fig. 2e**). When the density of data points is high, choose highly distinct symbols (**Fig. 2f**) that form strong visual boundaries[3].

Often the categories of data points fall into natural hierarchies. For example, the data points could represent genes classified by type (such as 'gene', 'nonprocessed pseudogene' or 'processed pseudogene') and their transcription state ('off' or 'on') (**Fig. 3a**). If, within these categories, one state is deemed more relevant (for example, transcribed as opposed to nontranscribed), assign the symbols to reflect this hierarchy. Map salience to relevance[4] by using symbols with greater visual weight (fill and/or color) to distinguish and elevate important data (**Fig. 3b**). The use of a single color is effective at isolating a single variable (**Fig. 3c**). Use less prominent symbols for data that are less relevant (such as reference data included for context).

When there is a large number of symbols, it may be difficult to discriminate among them no matter how well they are chosen. If your plot has more than six or seven categories, consider presenting the data in several panels with each showing a few data categories—a technique known as small multiples.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Bang Wong

1. Cleveland, W.S. & McGill, R. *J. Am. Stat. Assoc.* **79**, 807–822 (1984).
2. Lewandowsky, S. & Spence, I. *J. Am. Stat. Assoc.* **84**, 682–688 (1989).
3. Cleveland, W.S. *Elements of Graphing Data* 2nd edn. (Hobart Press, 1994).
4. Wong, B. *Nat. Methods* **8**, 889 (2011).
5. Zheng, D. *Genome Res.* **17**, 839–851 (2007).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at the Johns Hopkins University School of Medicine.