POINTS OF VIEW

Binning high-resolution data

Limitations in print resolution and visual acuity impose limits on data density and detail.

The size of features in genomic data sets span many orders of magnitude, and it is a challenge to draw elements in a figure small enough to preserve detail but large enough to be visible. In a previous column[1], strategies were identified to present genomic data in context[2,3]. This month we look at methods to bin high-density information and provide guidelines for the minimum size of elements in a figure.

Visual acuity imposes stricter limits than output resolution. A common unit of length in print is the point (pt; 1 pt = 1/72 inch). The resolving power of the eye is about 1/4 pt at a distance of 30 cm, and many journals impose a 1/4-pt or 1/2-pt minimum line width for figures. Although it is possible to discern 1/4-pt lines that are 1/4 pt apart (**Fig. 1a**), such fine detail can overwhelm the eye. We suggest lines at least 1/2 pt in width that are no closer together than 3/4 pt (**Fig. 1b**).

A size of at least 1 pt is needed to resolve the color of small elements, and to comfortably assess differences in adjacent heights (**Figs. 1** and **2**). When 1/2-pt line widths are used for axes and grids, a 1-pt line thickness for data traces is suggested, and symbols in line plots should be no smaller than 3 pt (**Fig. 1**). In any context, data traces should use symbols no finer than 1.5 pt on a 1/2-pt line. For scatter plots of high density, when large points can occlude each other, or if outliers are shown in a distinct visual channel, data points can be as small as 1 pt.

These requirements inform the extent of binning required for dense data tracks. **Figure 2** demonstrates the visibility of binned data for bins of 1/4 to 2 pt. Finding local maxima is relatively easy even with 1/4-pt bins, but judging the average, assessing variability and discerning minima are difficult with bins smaller than 1 pt. Histograms are preferred over heat maps, except where space is an issue—heat maps can be more compact and effective for sparse data (track d, **Fig. 2**). We suggest not binning data into more than ~250 intervals for one-column figures (3.5 inches wide) or ~500 intervals for two-column figures (7.2 inches). This corresponds roughly to 1 pt in print, 4 pixels on a high-resolution screen or 2 pixels on a typical LCD projector. The limit on bin size reduces detail and smoothes out variation—for example, a full-page figure of human chromosome 1 requires bins of 500 kb (~50 times the average gene size). One can mitigate this by encoding central tendency (median, average), extrema (minimum, maximum) and spread (s.d., interquartile range) (**Fig. 3**), or by highlighting global extrema or outliers (track d in **Fig. 2** and tracks e–g in **Fig. 3**).

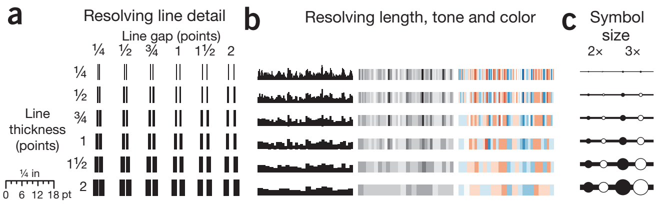

Figure 1 | Visual-acuity limits impose a minimum size on elements. (**a**) Lines thinner than 1/2 pt cannot be comfortably resolved if less than 1/2 pt apart. (**b**) Differences in tone, length and color are difficult to judge for elements smaller than 1 pt. (**c**) Data points should be at least three times the width of their line. White circles are shown with a 1/4-pt outline.

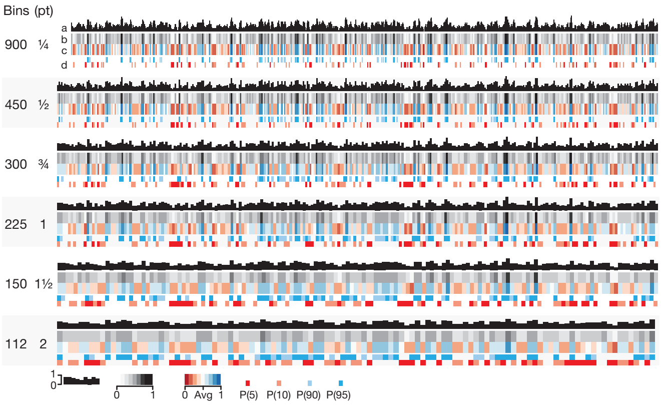

Figure 2 | Each set of tracks shows the same simulated coverage of a sixfold sequencing process differing only in how the data are binned (1, 2, 3, 4, 6 and 8 values into 900, 450, 300, 225, 150 and 112 bins, respectively). Bin sizes range from 1/4 to 2 pt. The coverage average across each bin is shown as a histogram (a) and heat map sampling the nine-color gray sequential Brewer palette (b). Coverage relative to the average coverage is shown as a heat map with a red–blue diverging Brewer palette (c). Bins with values at least as extreme as the 5th, 10th, 90th or 95th percentile (P(5)–P(95), respectively) of the full data set are marked in shades of red and blue according to the key at the bottom (d).

COMPETING FINANCIAL INTERESTS

The authors declared no competing financial interests.

Martin Krzywinski

1. Nielsen, C. & Wong, B. *Nat. Methods* **9**, 423 (2012).
2. Anders, S. *Bioinformatics* **25**, 1231–1235 (2009).
3. Nielsen, C. *et al. Genome Res.* **22**, 2262–2269 (2012).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.

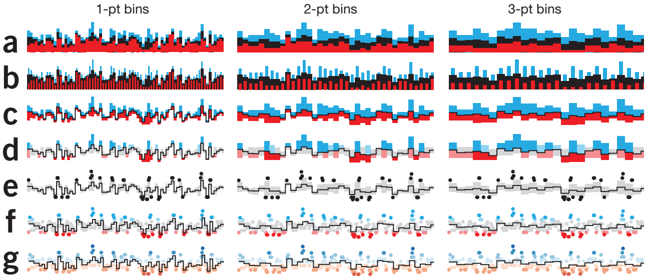

Figure 3 | Aggregate statistics about central tendency, extrema and variation can be quantitatively encoded using multiple overlapping traces, shown here for the data in **Figure 2** using 1-, 2- and 3-pt bins. (a,b) Bin minimum (red), average (black) and maximum (blue). (c) Like a and b, except that minimum and maximum bin values extend from the average. The average is shown using a 3/8-pt line. (d) Like c, but with extrema values within the 10th–90th percentiles of global data shown in gray. (e) Bin s.d. (gray) with individual values (1-pt circles) in the bottom or top 5th percentile (black points). (f) As in b, but all individual values are shown, with those within the 10th–90th percentiles of global data in gray, and those in the top 10th percentile encoded in color like in d. (g) Individual values with *z*-scores encoded with a nine-color red–blue diverging Brewer palette.