POINTS OF VIEW

Managing deep data in genome browsers

Techniques are at hand for taming the ever-growing number of data tracks.

Obtaining genome-scale data has never been easier. In addition to sequencing genomes, biologists now routinely profile epigenomes, transcriptomes and proteomes. There are exciting opportunities to better understand genome regulation by integrating diverse data types into unified views. Visualization facilitates data interpretation, but designing meaningful visual depictions of these data is a challenge.

Most genome browsers arrange data from different experiments vertically and align them to a reference coordinate. This arrangement of stacked data rows, or 'tracks', facilitates comparisons between diverse data types. However, as the number of tracks grows, it becomes increasingly difficult to see all of the data and to find meaningful patterns (**Fig. 1**). Because different data types warrant different graphical representations, the process of displaying disparate data creates a high degree of visual complexity. The ability to reorder and color-code tracks helps to organize information, but researchers urgently need ways to manage the overwhelming depth of genomic data.

There are several strategies available to reduce this visual complexity. With each there is a trade-off between gaining a meaningful overview and losing data details. Finding the balance depends on the resolution at which the data need to be analyzed. Two popular approaches to dealing with the track depth in genome browsers are (i) compaction, which preserves the original data but presents them in a more succinct and graphically economical way, and (ii) summarization, which replaces the original data with an abridged view.

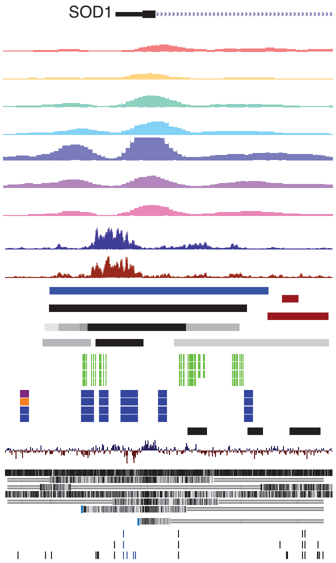

Figure 1 | Genome-scale data as depicted by the University of California Santa Cruz Genome Browser[1].

Compaction is a practical approach to reclaim valuable screen space. The most straightforward compaction technique is to make each track of a browser shorter. A more extensive approach, however, is to coalesce multiple tracks into a single row (**Fig. 2a**). The University of California Santa Cruz Genome Browser[1] uses transparency to overlay so-called 'wiggle' tracks. These histograms displaying dense continuous data are common in genome browsers and their characteristic shapes can be highly informative. Placing the histograms in front of one another gives them a shared *y* axis that makes comparing the shapes and heights of peaks easier than having the profiles arranged in separate and vertically stacked tracks. The drawback with overlaid histograms is that some data is obscured. Furthermore, deciphering constituent tracks in the overlay can be nontrivial because of color mixing.

Heat maps provide another form of compaction (**Fig. 2b**). In this approach, peak heights are depicted as value intensities in which taller peaks produce darker bands. Although this representation takes up less space, it can be difficult to evaluate quantitative information from intensity alone. Heat maps are best suited for distinguishing broad value ranges, such as the highs from the lows. Employing a divergent color gradient can help emphasize the extremes.

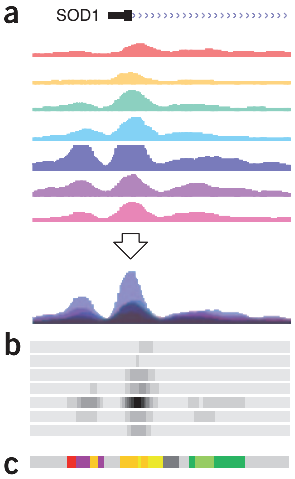

Figure 2 | Examples of reduced visual complexity. (**a**) Individual histogram tracks are made partially transparent and collapsed into a single track. (**b**) A heat-map view replaces peak heights with color intensity and requires less display space. (**c**) Summarization of data vertically into biologically meaningful categories.

Unlike compaction, summarization provides higher-level reasoning about the data at the expense of data details. When data is presented as it is collected—as one track per experiment—the resulting number of tracks can be overwhelming, making it difficult to find relevant trends. Summarization involves computing metrics across experiments to create a novel portrayal of the data (**Fig. 2c**). For example, the metric could be a simple average or a more domain-specific value, such as chromatin state inferred from combinations of chromatin modifications[2]. With the details hidden, researchers can focus on global trends and more readily prioritize points in the data that warrant deeper inspection.

Compaction and summarization are both required to tackle the challenges posed by the ever-growing genome browser track stack. Although the examples presented in this column focus on data from sequencing-based technologies, the principles of compaction and summarization generalize to other data types. There is great potential for innovation in the development of new summarization methods. However, these abstractions are unlikely to replace primary data altogether; rather, the more verbose track displays will be shown as a second layer of information. This would require genome browsers to support a hierarchy of summary tracks with distinct sub-tracks showing the original data.

Cydney Nielsen & Bang Wong

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).
2. Ernst, J. & Kellis, M. *Nat. Methods* **9**, 215–216 (2012).

Cydney Nielsen is a Canadian Institutes of Health Research and Michael Smith Foundation for Health Research postdoctoral fellow at the British Columbia Cancer Agency in Vancouver. 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.