POINTS OF VIEW

Representing the genome

The choice of visual representation of the linear genome is guided by the question being asked.

Many genomics techniques produce measurements that have both a value and a position on a reference genome. The genome coordinate provides a natural ordering to these data values and is the organizing principle driving how we commonly display and navigate genomic data today. A popular plotting approach is to arrange the linear genome coordinate along the *x* axis and express the data value range on the *y* axis (**Fig. 1a**). This conventional representation is limiting. By using other organizational frameworks we can better extract the information of interest and make sense of its patterns.

The genomes of many model organisms are large and pose a considerable display challenge. For example, the human genome is over 3 billion bases long. Using a 1-point line (0.014 inch thick) to represent each base of chromosome 1 would require a sheet of paper over 50 miles long. The initial human genome research article[1] uses extensive roll folds to create condensed 'chromosome maps' (**Fig. 1b**).

One way to build a condensed overview is to divide the genome into equally sized chunks and report a summary value for each. This works well if the features are large and exceed the chunk size. But if the features are much smaller than the divisions, important information will often be obscured. This is why images that capture large swaths of the genome provide poor overviews of relatively small features such as genes. The interactive zoom of genome browsers addresses this problem by enabling researchers to inspect the genome at different scales. By zooming in, chunk sizes can be made ever smaller thereby increasing our ability to resolve compact features of interest.

An approach to creating a more meaningful overview is to isolate only the features of interest. By removing the intervening portions of the genome, we bring the relevant signals together for effective side-by-side comparisons while preserving the linear genomic context (**Fig. 2a**). The University of California Santa Cruz Cancer Genomics Browser enables researchers to limit the display to a set of genes, for example, those belonging to specific biological pathways. The result is a balance between overview and detail.

Another strategy is to leave the genome intact and maximize the amount that is displayed. For example, the genome can be arranged according to space-filling curves such those described by mathematicians Giuseppe Peano and David Hilbert[2] (**Fig. 2b**). This presentation has the advantage of representing adjacent positions with adjacent pixels; however, some distortions are unavoidable, and some proximal pixels will correspond to distant loci. Although this method uses space efficiently, it restricts the view to a single data set and requires the same limiting summarization across genomic chunks as described for linear overviews.

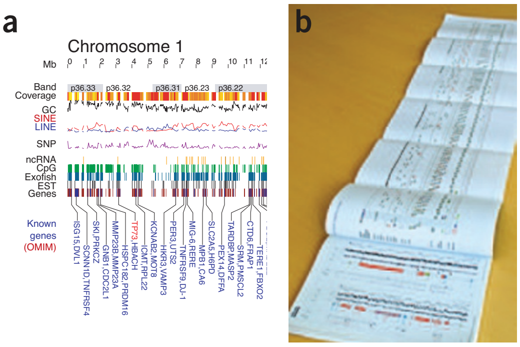

Figure 1 | The immense scale of genomic space. (**a**) Example of data features from human chromosome 1 (reprinted from ref. 1). (**b**) Roll fold showing chromosome maps from the initial human genome publication[1].

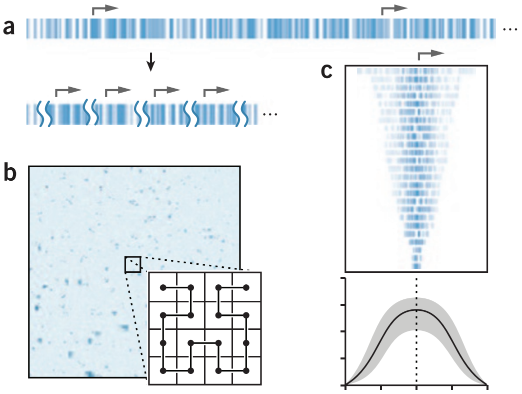

Figure 2 | Different ways to display genomic data. (**a**) Accordion view with transcriptional start sites (arrows) intact and intervening sequence removed. (**b**) Hilbert curve display of data across a chromosome. (**c**) Stack of regions from **a** centered on transcriptional start sites with hypothetical summary statistics plotted.

In displaying genomic data we are faced with a trade-off between focusing on data features in isolation and seeing them in context. There are times when the spatial arrangement of features along the genome is of little interest and the genomic ordering can be abandoned altogether. Regions of interest can be extracted and stacked vertically along common reference points, such as transcriptional start sites (**Fig. 2c**). This allows them to be sorted using various metrics to reveal patterns. Summary statistics complement the considerable amount of data that are typically displayed with this approach.

These techniques do not account for the three-dimensional packaging of the genome. As we better understand how the genome folds, we will likely change our approach to organizing and accessing genomic data. For example, open and closed chromatin have been observed to occupy different spatial compartments. Grouping data by whether they coincide with one or the other of these states may prove more informative than an arrangement on the linear genomic coordinate.

Next month, we will examine the challenges posed by comparing data from multiple experiments as we move from looking at features along the genome's *x* axis to information spanning the *y* axis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Cydney Nielsen & Bang Wong

1. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
2. Anders, S. *Bioinformatics* **25**, 1231–1235 (2009).

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.