POINTS OF VIEW

Representing genomic structural variation

Techniques for displaying relations between distant genomic positions.

With a rapidly growing collection of genomes coming from such initiatives as the 1000 Genomes Project, the days of a single reference genome are numbered. Although the genomic sequence between any two human individuals differs only by about 0.1%, there are abundant structural and copy-number variations of different types and sizes. Effective visualization of these genomic variations is required to gain insight into the genetic basis of human health and disease. However, variation data pose new challenges to traditional genome visualization tools, which depend on linear layouts and have difficulty depicting large structural rearrangements.

A structural variant consists of a DNA sequence, typically >1 kilobase, that deviates from a reference sequence in content, order and/or orientation. Depicting such a structural difference requires showing both the variant and reference sequences. The sequence boundaries of a structural variant, so-called 'breakpoints', span a wide range of distances and affect sequence segments of varying size. For example, tandem duplications may involve a localized repetition of only a few kilobases, whereas the breakpoints of translocations are located on nonhomologous chromosome arms and may result in the rearrangement of large genomic chunks. Finding a representation that enables one to track breakpoints across this scale can be challenging. This is exacerbated by the fact that variant genomic fragments can be flipped end to end (inversions), requiring us to also account for their orientation.

A natural solution to depict structural variants is to draw arcs between the breakpoints on a linear layout of the reference genome (**Fig. 1a**). This representation effectively conveys a small number of structural variants spanning similar genomic ranges, but it is impractical for linear genome browsers because it is difficult to display long-range arcs. Using a circular layout, as with a Circos ideogram[1], constrains the distance between any two points, making the display of arcs compact (**Fig. 1b**). However, this design, as with linear layouts, is prone to overplotting; displaying many arcs will give rise to visual clutter.

Although arcs effectively highlight the positions of breakpoints in the reference genome, the order and orientation of these sequences in the variant genome are not explicitly displayed. For example, interpreting that sequence *J* is followed by *K′* and sequence *K* is followed by *J′* in the translocation shown in **Figure 1a**,**b** requires readers to learn the conventions of these graphics. Alternatively, we can directly depict the rearrangement of reference sequences in the variant by using color (**Fig. 1c**). However, color-coding the chromosomes does not capture changes in orientation such as inversions. Another approach that explicitly captures sequence orientation is the dot plot (**Fig. 1d**). The axes of the dot plot correspond to the two genomes being compared, and the points indicate sequence identity. The order and orientation of the sequences in both genomes can be read directly: diagonal lines indicate corresponding sequence segments, and the horizontal offsets highlight reordering. The trade-off for directly depicting the variant sequence as a color-coding or dot plot is that only one variant-reference sequence pair can be expressed at a time.

All of the images presented so far are based on a genomic coordinate system, which heavily emphasizes the distances between breakpoints. It might be more biologically meaningful to focus on the consequences of the breakpoints instead of their genomic arrangement. For example, perhaps we should highlight gene fusions, particularly those whose fused segments are in frame. One way to address these functional questions is to move away from the genome coordinate system and use a different representation, such as a graph, where nodes represent the uninterrupted sequence segments and arrows indicate sequence order (**Fig. 1e**). The layout is then based on maximizing the readability of the connections rather than on preserving the linear order of the genome coordinate. Relevant metadata, such as the presence of an in-frame gene fusion could be emphasized with edge attributes such as color.

As we look for alternative ways to capture the number and diversity of genomic variations, it will be critical to ensure that biologically relevant features are made most noticeable.

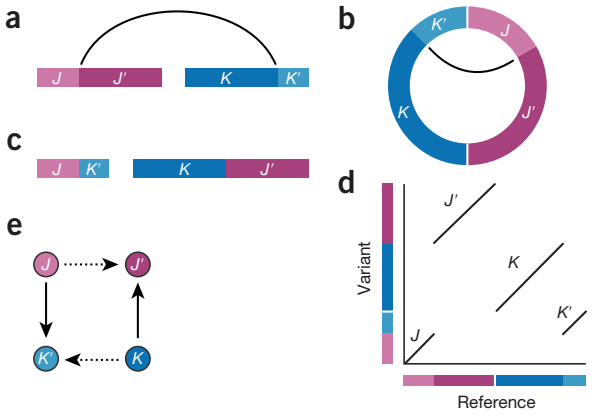

Figure 1 | Representations of a translocation. (**a**,**b**) Linear (**a**) and circular (**b**) reference genome layouts with an arc to depict a translocation between two chromosomes (pink and blue). (**c**) Translocation illustrated as reference-sequence segments with chromosome colors corresponding to those in **a**. (**d**) Dot plot indicating positions of identical sequences in the variant and reference genomes. (**e**) Graph of common sequences (nodes) and their order in variant and reference genomes (solid and dashed arrows, respectively).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Cydney Nielsen & Bang Wong

1. Krzywinski, M. *et al. Genome Res.* **19**, 1639–1645 (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.