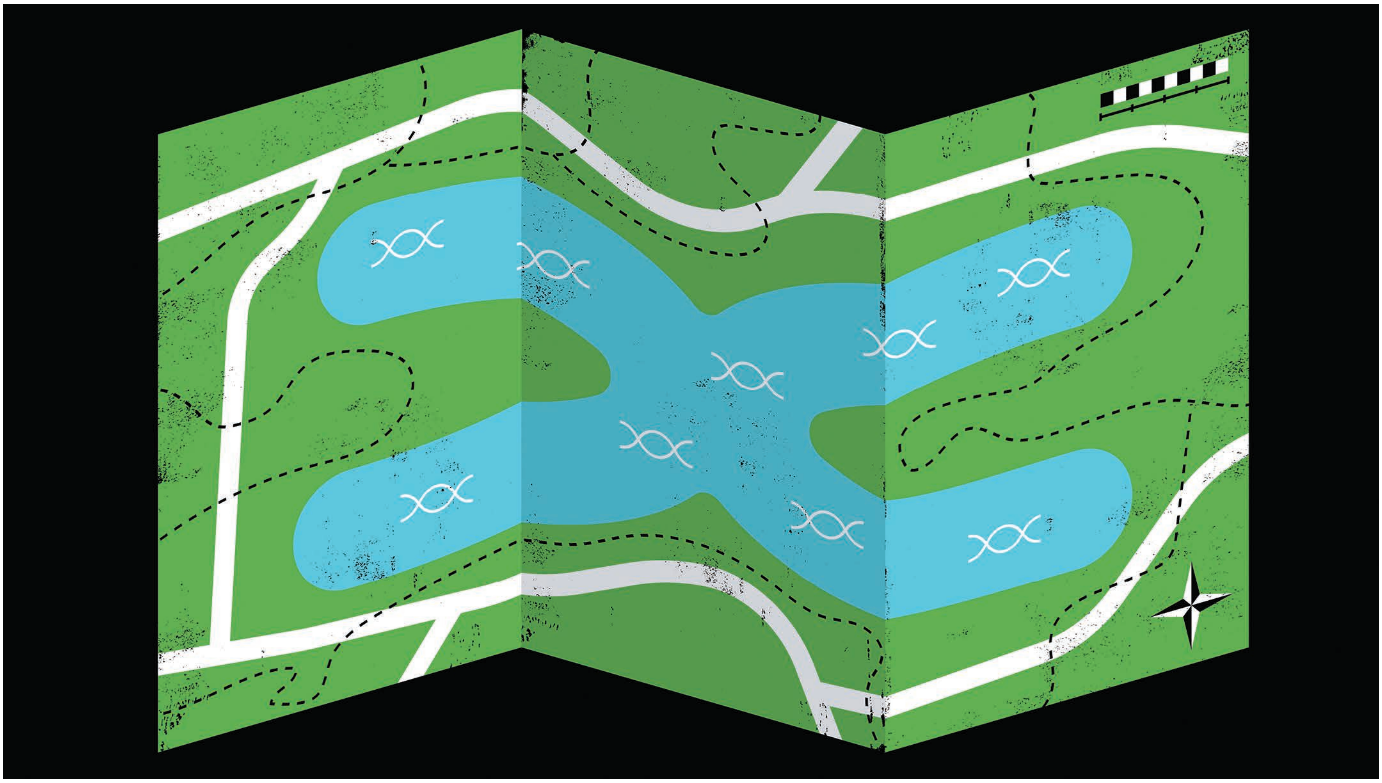
TOOLBOX

PLOT A COURSE THROUGH THE GENOME

Inspired by Google Maps, a suite of tools is allowing researchers to chart the complex conformations of chromosomes.

ILLUSTRATION BY THE PROJECT TWINS



BY JEFFREY M. PERKEL

Chromatin does much more than just keep DNA neat and tidy. This complex of genomic DNA and protein assumes many different structures and conformations, which can affect the expression of the genes wrapped around it. In certain conformations, two sequences that are far apart in the linear DNA might actually be located next to each other and influence each other's activity; in other conformations, they might be far apart.

Erez Aiden was a graduate student at the Massachusetts Institute of Technology in Cambridge when he co-developed a technology that, for the first time, revealed the landscape of chromosome folding on a genomic scale. Hi-C details the DNA loops and structural domains that influence gene expression, and can even help to piece together complex genomes. The data take the form of 2D matrices detailing chromatin contacts, but in 2009, Aiden had no easy way to explore them. So, he improvised.

"I would simply print out Hi-C matrices at multiple resolutions and I would use up hundreds of pages of paper," he recalls. "I would find the biggest conference table I could and I'd just array printed pieces of paper in front of me in order to be able to see a big chunk of the matrix."

"It was a great interface," Aiden says. Still, he concedes, a more environmentally sustainable — and sharable — approach was required.

The result was Juicebox, a Java-based desktop application that provides Google Maps-style exploration of chromatin-interaction data. It allows researchers to zoom from the genome level down to small structural features.

Released in 2014, Juicebox has been downloaded some 14,000 times, Aiden says, and a browser-based version launched this year. Juicebox is just one of a range of freely available programs for exploring 2D interaction data: some focus on relatively narrow chromosomal loci, whereas others enable genome exploration. A growing subset infers 3D structures from 2D matrices. But all reflect the growing richness of chromatin-interaction data sets, not to mention the influence of funding initiatives such as the 4D Nucleome Project.

"Because [the data] have become so complex, visualization just became a lot more important," says Peter Park, a bioinformatician at Harvard Medical School in Boston, Massachusetts.

The University of California, Santa Cruz (UCSC) Genome Browser is one of the ▶

most popular portals for exploring genomic data. Like most genome browsers, it renders sequence data as a linear array of letters decorated with epigenetic features, such as histone modifications and methylation sites, displayed in 1D 'tracks'.

Hi-C, however, generates 2D matrices. The technology identifies sequences that are far apart in the linear DNA sequence but close neighbours in 3D space. "You look at a pair of positions in the genome, and it tells you often they bump into one another," Aiden explains. Typically, those data are rendered as heat maps, with colour intensity reflecting the interaction frequency between two points.

Aiden and his co-developers, including James Robinson of the University of California, San Diego (UCSD), took inspiration from Google Maps, in which users can seamlessly zoom from the global to the street level. The entire data set is massive, but Google doesn't deliver it all at once. Instead, the software "divides the world into tiles at different resolution", says Robinson. At any one time, users view just a handful of tiles, which are organized to make adjacent tiles easy to fetch. "As long as you can always get to the four you're looking at quickly, you can support an interactive map," he says.

Similarly, Juicebox 'hic' files store precomputed tile sets for each possible chromosomal pair at multiple resolutions. A look-up table speeds access by allowing the software to retrieve data without having to search. As a result, Juicebox users can seamlessly explore an entire genome's worth of interactions, and then zoom in to view fine-scaled features.

Users can access any of several hundred precomputed contact maps that the Aiden lab has made publicly available, or view their own. They can overlay those data with standard browser tracks, such as gene locations or histone marks, either from their own studies or from public repositories. Binding sites for the DNA-binding protein CTCF, for instance, highly correlate with chromosomal loops. And they can flag and record features of interest.

GENOMES IN SYNC

HiGlass, a web-based tool launched in March by biomedical informatician Nils Gehlenborg at Harvard Medical School, also provides a Google Maps-like experience. As with Juicebox, researchers can import genomic tracks to help make sense of what they're seeing, but HiGlass also allows them to open multiple HiGlass views in one browser window and synchronize them so that they always display the same region. That way, Gehlenborg says, researchers can compare chromatin conformations across different conditions or experiments. "We are enabling the investigators and the analysts to generate new hypotheses," he says. (The browser-based version of Juicebox also allows multiple synchronized views per window, Aiden says; users of the desktop Juicebox app can synchronize maps across different windows, but not in a single display.)

Gehlenborg's team has established a HiGlass server for exploring publicly available data. Researchers who need to analyse custom data sets must install the software locally; a Docker container is available for that purpose.

Both Juicebox's web version and HiGlass allow users to create sharable URLs that point to specific views of the data — a feature that Aiden calls his software's "killer app". Suppose a user notices that a genomic structure overlaps perfectly with particular 1D track, he says. "You just take that URL, copy it, and you can tweet it. And all the people who receive the tweet can just click on it and boom! They get the exact same configuration that you had now on their computers as well."

Two other visualization options, the 3D Genome Browser and the WashU EpiGenome Browser, provide more localized views. Users select a locus of interest and the browsers display contacts in the area.

> *"We are enabling the investigators and the analysts to generate new hypotheses."*

Whereas Juicebox and HiGlass render heat maps as squares divided diagonally into two mirror images, these browsers show heat maps as triangles — that is, half of the square, without its mirror image. "We cut down the half that is redundant information," says genome biologist Bing Ren at UCSD. (The WashU browser can also display contact data as arcs connecting linked regions.)

That change may sound trivial, but according to Feng Yue of Pennsylvania State University in Hershey, who developed his first 3D Genome Browser prototype as a postdoctoral researcher with Ren, it makes it easier to identify functional regions. The 3D Genome Browser, for instance, allows its users to align heat maps from two species, one atop the other, to assess the evolutionary conservation of folding architecture. A 'virtual-4C' mode allows users to query Hi-C data sets for sequences interacting with a specific genomic locus, providing a window into interactions between gene-regulatory regions.

Another option is GIVE, released by bioengineer Sheng Zhong and his colleagues at UCSD. This allows researchers to incorporate a fully functional genome browser, including a 2D contact data viewer, into their personal or lab web pages with just a few lines of HTML code. Researchers can thus share data with colleagues, publish it alongside their manuscripts, or explore it themselves — all with about 20 minutes' work, says Zhong.

Francesco Ferrari, a computational biologist at the FIRC Institute of Molecular Oncology in Milan, Italy, visualizes his Hi-C data using the R programming language and the Bioconductor software library. These text-based programs lack the interactivity of other software, but because the team already runs data analysis using R and Bioconductor, Ferrari explains, "it's just more convenient" to use them for visualization as well. The Bioconductor package HiTC provides Hi-C visualization tools, as does the Python library HiCPlotter.

GOING 3D

Ultimately, 2D contact matrices imply 3D structure. After all, if two regions interact, they are probably in close physical proximity. Increasingly, some researchers are using their 2D data to compute and visualize 3D structures directly.

Csilla Várnai, a postdoc at the Babraham Institute in Cambridge, UK, helped to produce the 3D models for a single-cell Hi-C study earlier this year (T. Nagano *et al. Nature* **547,** 61–67; 2017). She used a generic molecular modelling package called Gromacs to model a chromosome as a string of beads — each representing about 100 kilobases — and then asked it to fold, using the Hi-C contacts as 'constraints' on that process.

Other packages have been designed specifically to model chromatin structure. Chrom3D, developed by bioinformatician Jonas Paulsen at the University of Oslo blends Hi-C data with information on proximity to the nuclear envelope to model the position of chromosomes in the nucleus. "That matters a lot for gene regulation," Paulsen explains. Genes near the nuclear periphery tend to be repressed, whereas more centrally located genes are usually active. Another tool, TADkit, from Marc Martí-Renom and Mike Goodstadt at the National Center for Genomic Analysis–Center for Genomic Regulation in Barcelona, Spain, allows users to view 3D chromosome models alongside the corresponding 2D heat map and 1D tracks. Selecting a feature in one representation highlights overlapping features in the others.

It remains to be seen what insights such 3D representations can provide that 2D heat maps cannot, especially as most Hi-C data sets represent millions of cells, rather than a single structure. Leonid Mirny, a bioinformatician at the Massachusetts Institute of Technology, likens the resulting data to averaging a batch of photographs to determine what a typical person looks like. "It's not going to be actually representative of anyone whom you take pictures of," he says. Also unclear is which tool, if any, will emerge as the de facto standard for genome visualization. Debate on that front is ongoing, says Zhong.

When it comes to genome biology, says Ren, visualization is key. Analytical tools are based on statistics, he explains; sometimes they miss things, and sometimes they detect features that aren't there. "At the end of the day, nothing replaces looking at the data yourself." ■

Jeffrey M. Perkel *is technology editor at* Nature.