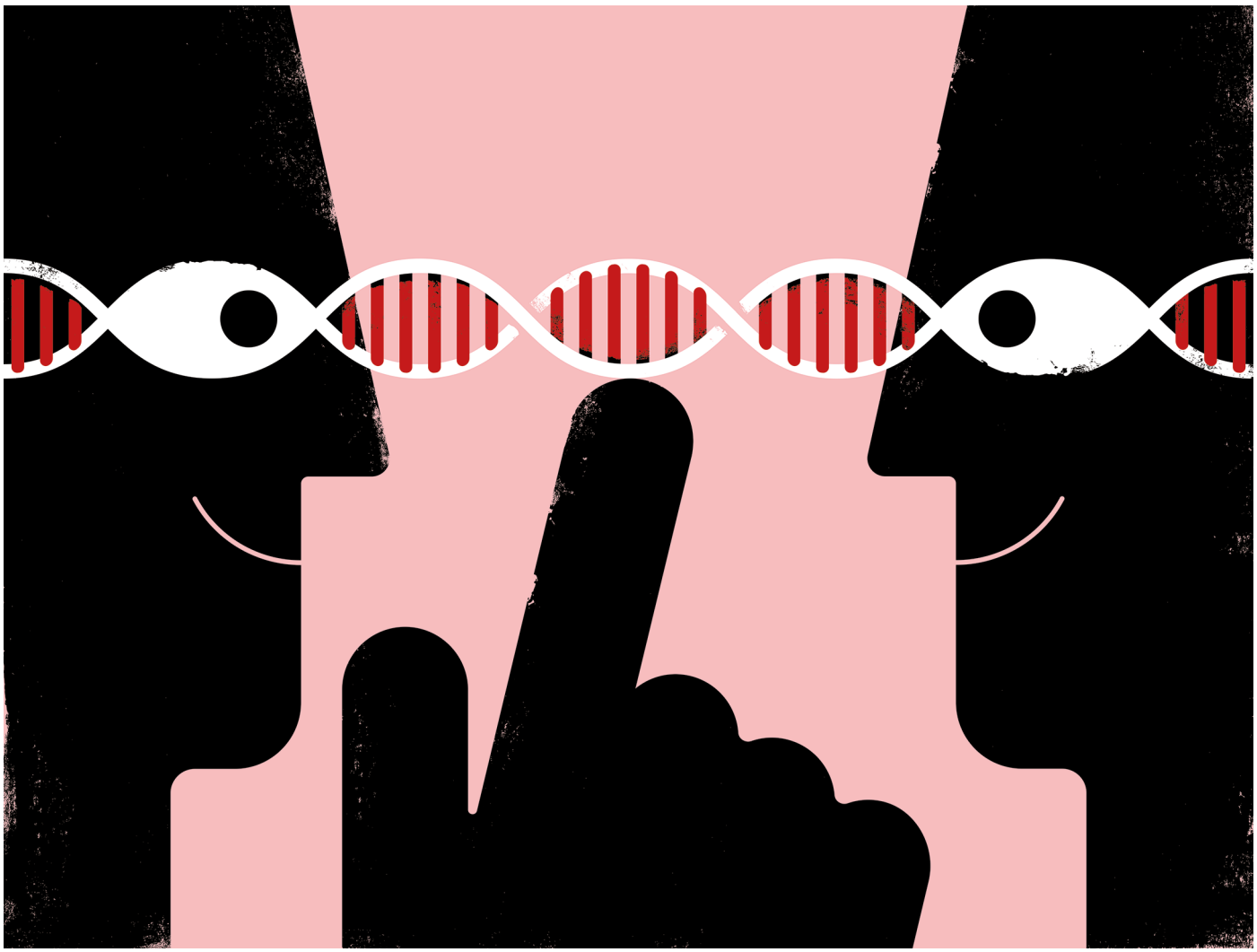

THE PROJECT TWINS

A PICTURE IS WORTH A THOUSAND BASE PAIRS

A small but powerful toolset makes sharing genomic data visualizations straightforward. **By Julian Nowogrodzki**

When Adam Siepel was building algorithms for evolutionary genomics as part of his PhD, he wasn't thinking about visualization. But, as a graduate student in the laboratory of computational biologist David Haussler, at the University of California, Santa Cruz (UCSC), he happened to sit next to the software engineers who were building and maintaining a tool called the UCSC Genome Browser. These engineers helped Siepel to make his algorithms publicly available as a track, or data overlay, that anyone could explore.

Genome browsers are graphical tools that display the genome sequence, usually as a horizontal line. Other sequence-associated data are aligned and stacked above and below that line in 'tracks', for instance to illustrate the relationship between gene expression, DNA modification and protein-binding sites.

Siepel's track identifies sequences that have been retained over evolutionary time; when a user applies it while viewing the alignment of genomic data from two or more species, the track highlights regions that are evolutionarily conserved. Allowing others to use the algorithm to highlight regions of interest in their own data was "probably the single most important thing I did during my PhD", says Siepel, who is now a computational biologist at Cold Spring Harbor Laboratory in New York. Other researchers have used it, for instance, to find mutations associated with diseases and to pinpoint functionally important regions of noncoding RNA molecules.

Today, a growing collection of free and open-source tools exists for sharing such genomic data. Which one is right for you depends on what kind of sharing you want to do: communicating with a collaborator, for instance, requires different software from what you'd use for disseminating data to the broader scientific community.

Whatever the motivation, sharing genomic data broadens its impact, says Siepel. "Almost all of our most-cited papers are supported by browser tracks," he says.

For broad dissemination of genomic data,

Siepel recommends the approach that worked for him: making a track. And he suggests two genome browsers to display them: "UCSC and Ensembl are the leaders," he says.

Dissemination stations

The UCSC Genome Browser and its 'track hubs' – data tracks that are hosted remotely by external teams – can run in a web browser or on the desktop. The desktop version is called Genome Browser in a Box. Users can submit data to be included as a public track registered with UCSC. The team typically accepts only widely useful data (and not, for instance, those limited to a specific disease), and publication in a peer-reviewed journal is a plus. Alternatively, users can build personal track hubs, which involves formatting their genomic data in files of a specific format, indexing those files and making them web-accessible.

The Ensembl genome browser, hosted at the European Molecular Biology Laboratory's European Bioinformatics Institute in Hinxton, near Cambridge, UK, allows users to import data as custom tracks, just as UCSC does; both its own format and the UCSC formats are supported. The Ensembl team has built a searchable Track Hub Registry to make it easy to find relevant track hubs for use with Ensembl or the UCSC Genome Browser.

The UCSC browser accumulates roughly two million hits a day, says Robert Kuhn, associate director of the UCSC project. And several large projects use it to disseminate data. The Genotype-tissue Expression Project, for instance, used the browser to create a track that visualizes as many as 53 tissues from 1,000 donors. The journal *Nucleic Acids Research* requires authors with whole-genome data to create a track hub for reviewers, Kuhn notes, and some authors choose to make them available to readers as well.

If you have basic Unix command-line skills, setting up a UCSC track hub takes just a few hours, says Kuhn. Instructions are available at go.nature.com/2pqkym.

Shareable, embeddable

UCSC and Ensembl allow researchers to share data sets as tracks in a public, centralized database that is controlled by others and available to any user of that database. If your goal is to embed a visualization in a website, or to create a specialized visualization for a paper, other options are available; these include GIVE, JBrowse and IGV. (UCSC and Ensembl can also perform these tasks.)

GIVE is an open-source tool that allows researchers to build custom genome browsers for their labs with little if any programming. According to Xiaoyi Cao, a GIVE developer and a software engineer at Google, there are three ways to host data. One is for researchers to build an entire GIVE instance on their lab server using GIVE-Docker, a pre-packaged version of GIVE that the container engine software Docker can run immediately. Because the data can remain on a private server, they do not have to be visible to the web and thus can be more secure.

Alternatively, labs can submit a list of URLs that point to the data sets they want to include, and GIVE will build the database for them, no programming required. The data can be in any of multiple formats, including those for gene-expression and protein-binding data. The resulting database will be based at a GIVE instance, or mirror, hosted by the University of California, San Diego (UCSD). And according to Sheng Zhong, the UCSD computational biologist who heads the GIVE team, it takes just two to three minutes to set up.

The third option is to include your data in the public GIVE data hub. Researchers submit their metadata to an online form, and the GIVE developers will let them know if their data have been selected.

"We have all these wonderful new data types, and we have to figure out how to visualize and combine them."

JBrowse can run either in a browser or on the desktop. Ian Holmes, JBrowse's lead developer and a computational biologist at the University of California, Berkeley, says that he designed the tool to be responsive, intuitive and accessible for non-coders. With the desktop version, users can load data directly from their computer, Holmes says. The browser version requires an index file that tells the browser where to find the relevant data; it also requires data to be web-visible (for example, in the cloud or on a lab server). The JBrowse community has compiled a repository of about 50 plugins "that significantly enriches the visualization", says Holmes. One example allows users to see all DNA methylation results in a single track.

Several specialized genome browsers and databases use JBrowse for data visualization; among these are the cancer genome browser COSMIC and VEuPathDB, a genomic database for pathogens and disease vectors. David Beare, a computational biologist at the Wellcome Sanger Institute in Hinxton, says that COSMIC uses JBrowse in part because "it was faster and more responsive, and certainly more intuitive" than other genome-browser options available. The VEuPathDB database developers found that JBrowse "was most amenable to our own active development" of plugins, says Omar Harb, a microbiologist at the University of Pennsylvania in Philadelphia, and director of scientific outreach and education for VEuPathDB.

JBrowse has also been used to build a collaborative annotation tool called Web Apollo, which allows multiple researchers to simultaneously annotate the same data in real time, as in Google Docs.

Get a Broad view

IGV is a genome browser maintained by UCSD and the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. Available in desktop, browser-based and embeddable Javascript versions, IGV can generate QR codes (square barcodes) for specific data visualizations; for example, for inclusion on a poster. "IGV is always run locally on a user's computer. There is no notion of 'uploading' data or saving sessions to a central IGV server hosted by us," says Helga Thorvaldsdóttir, a software engineer at the Broad Institute. That also makes the system compatible with restricted data.

Jim Robinson, IGV's lead developer at UCSD, says that the browser is fast and easy to use. "Most users can learn the basics in a half hour or less," Robinson says. And the tool has racked up more than 7,000 citations, Thorvaldsdóttir says. At the Memorial Sloan Kettering Cancer Center in New York City, researchers have used IGV to visually check the genomic variants of patients whose cancer they sequence, says Robinson.

Prospective users of these tools can find plentiful educational resources online, including video tutorials. The UCSC Genome Browser has two archived and searchable listservs, or electronic mailing lists: one for website and data questions, the other for queries on setting up and maintaining Genome Browser mirrors. JBrowse users can ask questions on Github or on the software's open instant-messaging channel, but Holmes suggests contacting the developers directly. "We have some developers who really like getting feedback from users," he says.

And that includes suggestions for handling new challenges. Despite their utility, genome browsers are still mostly built on a fundamental assumption: that genomic data are best displayed in a linear format. But that doesn't work so well for some kinds of information, including interactions between distant genomic regions, and evolutionary relationships, says Siepel. Some researchers, such as Maria Nattestad, a bioinformatician at Google in Palo Alto, California, have built niche tools for tackling these issues. Nattestad built a tool called Ribbon to better visualize long read alignments, for instance: these can snarl up other browsers because they often align with more than two places in genome.

"We have all these wonderful new data types, and we have to figure out how to visualize and combine them," says Nattestad. "It keeps me up at night in the best way."

Julian Nowogrodzki is a journalist in the Boston area of Massachusetts.