Visualizing epigenomic data

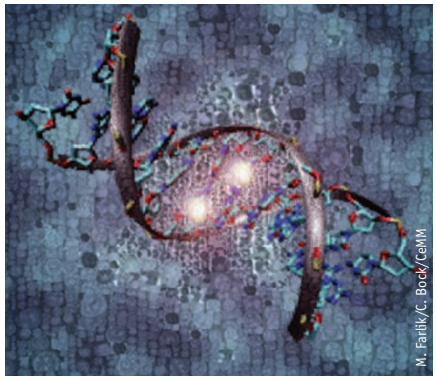
Vivien Marx

Epigenomics researchers are exploring new ways to visualize the many different types of genome marks.

When sketched quickly over coffee, genetic mutations are often drawn as a linear string of nucleotides, and one pen stroke alters one or several of those nucleotides. An epigenomic quick-sketch is more three dimensional (3D). To indicate a methylated site, for example, a drawing might include a dangling pendant representing a chemical group latched onto a DNA base. Another epigenomic drawing might show DNA looping like a tied shoelace to represent interaction between two specific genomic regions.

Such sketches are simplified versions of the kind of complexity with which epigenomic data visualization tool developers must grapple. For example, the DNA regions connected by a loop are actually many thousands of base pairs apart, but they are brought close together by proteins that interact with chromatin, which is composed of DNA and DNA-packaging proteins. Mediating these interactions are epigenomic factors—for example, modified histone proteins and DNA- or chromatin-binding factors.

One way scientists look at epigenomic data for a relatively small genomic region is through genome browsers such as the University of California at Santa Cruz (UCSC) Genome Browser and the Ensembl Genome Browser, says Christoph Bock, who is at the Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM) and who also has an appointment at the Max Planck Institute for Informatics (MPII). His lab is one-third bioinformatics, one-third wet-lab epigenetics and one-third large-scale epigenome sequencing. The design paradigm put to practical use in these browsers is one in which a long string of DNA is visually stretched out and in which epigenetic information is one of many types of relevant annotations, he says. In contrast, the WashU EpiGenome Browser can be used to display only epigenetic data.



Artistic rendition of DNA methylation from the Bock lab, made of thousands of single-cell images.

Ting Wang, who, along with his team, developed the WashU browser, added a metadata heat map alongside the linear display of genomic data. The approach helps researchers to sort and compare hundreds of genomic data sets, he says. The browser also works with the linear approach in such a way as to bring different parts of the genome together. This grouping offers a "gene-set view" as Wang calls it, which can show, for example, just chromatin interactions between two DNA regions. This grouping approach addresses the number one user request he and his colleagues receive, which is to help scientists access a specific data set in the browser.

The US National Institutes of Health (NIH) Roadmap Epigenomics Mapping Consortium recently finished mapping the epigenomes of over 100 cells and tissues (http://www.nature.com/collections/vbqgtr). Scientists can browse the data using a specific and elaborated functionality of the WashU browser, says Wang. "The Roadmap browser uses the engine provided by the WashU browser, but it performs a specific role by automating sample collection and clustering," he says.

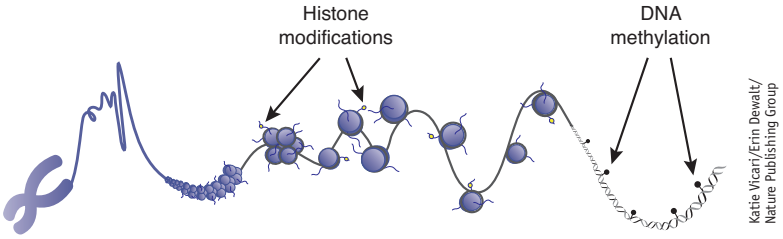
One aspect that Wang values in browser visualization of epigenomic data is the way it allows researchers to do, as he calls it, "visual bioinformatics." That approach goes beyond the more typical means of aligning, slicing and dicing used in genomics analysis when looking at intersections between different data sets. Instead, the WashU team set the browser up so that these slicing-and-dicing actions can be accomplished by dragging, dropping and button clicking.

Wang has a number of WashU browser developments in the works, including expanding the user base and addressing user questions, and he is continuing to work on visualization techniques with new web and database technologies. Two new browser functions are set to be rolled out later this year, he says: a repetitive-element browser and a comparative epigenome browser, which are both visual ways for users to close in on the data they want to analyze.

Further down the line Wang sees a personal genomics and a cancer browser. "But nothing beats my wish to develop 'data viz' in epigenomics into a video game," he says. Investigators will, for example, be able to jump into the data display, yank out a long noncoding RNA or blow up a nucleosome. "Research should be fun," he says.

Expanding tools to include analysis

Tool developers have to keep evolving the capabilities of their software in this burgeoning field. Traditionally, says Massachusetts Institute of Technology researcher Manolis Kellis, biologists used browsers to visually explore raw data and mentally take note of features such as surprising genomic regions. As data become large and composed of different types as is typical in epigenomics, on-the-fly computation becomes necessary.



Epigenomic data visualization tool developers must contend with plenty of complexity.

Computation and statistical reasoning do not suffer from the biases that scientists—and humans more generally—have, says Kellis.

A new generation of data visualization tools are emerging, he says, that combine statistics, computation, human intuition and interactive analysis. The user can ask questions and answer them in an interactive way through visualization and computation. Although the tools may be browser-like and show tracks, they are not displaying raw data but rather computationally interpreted data. They can—for example, as in his software ChromHMM—reveal patterns of epigenomic marks that represent a chromatin state.

Héctor Corrada Bravo and Florin Chelaru of the University of Maryland developed Epiviz to visualize genomic and epigenomic data[1]. Epiviz is a hybrid visualization and analysis tool, designed from the start to offer visualization and to integrate analysis tools, says Corrada Bravo. Epivis has a browser and it also has a module that integrates Bioconductor and its many statistical and graphical packages for high-throughput genomic data analysis. The software framework Bioconductor uses the R statistical programming language.

Corrada Bravo has new visualizations in development for Epiviz, as well as extended support for users' sequencing data. Originally, Epiviz integrated sequencing data that are summarized over genomic regions. The researchers plan on adding extensions that allow users to directly integrate unsummarized sequencing read data aligned to a reference genome, he says. The team is also on the lookout for new ways for users to visualize and explore nonadjacent regions of the epigenome.

As data sets and sample sizes grow, visualization tools have to accommodate the trend. The large sample sizes are one reason Corrada Bravo and his team developed Epiviz to be tightly integrated with R/Bioconductor. Dealing with large data sets means that users need to combine data-reduction approaches through statistical and computational methods. And it requires scalable visualization methods.

With R/Bioconductor, scientists avoid the dreaded "out-of-memory" notification when filtering data sets calls for more memory than is available. For example, they can process data with multiple computing units by leveraging the parallel processing facilities built into R/Bioconductor, says Corrada Bravo. Many data-reduction operations can happen behind the scenes, and the appropriate statistical summaries can be visualized.

As Chelaru explains, the more he became acquainted with the data and their visualization, the more he realized that the software needed to incorporate computational tools that let users transform data for exploratory, interactive analysis. Chelaru recently defended his PhD thesis and will be a postdoctoral fellow in Kellis's lab at the Massachusetts Institute of Technology in the fall.

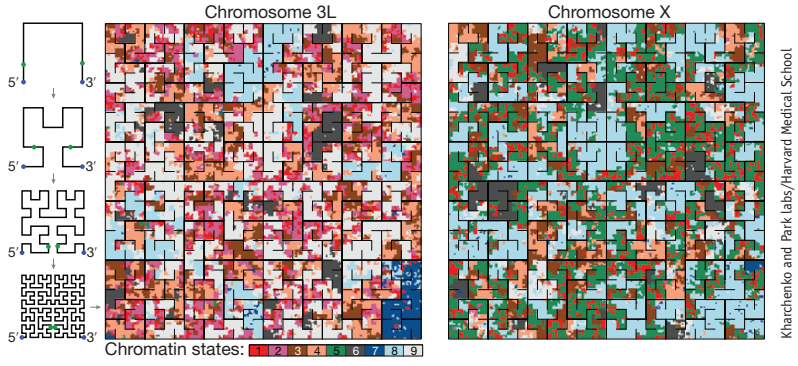
Large and larger projects

Both epigenomics and computational epigenomics have benefited from large research consortia, says Bock. These initiatives include the Human Epigenome Project; the High-throughput Epigenetic Regulatory Organisation In Chromatin (HEROIC) consortium; the NIH Roadmap Epigenomics Mapping Consortium; the International Epigenome Consortium, which aims to sequence 1,000 human epigenomes; the European Blueprint consortium; and DEEP, the German epigenome program. Despite their focus on cell lines, Bock also counts the Encyclopedia of DNA Elements (ENCODE) and model organism ENCODE (modENCODE) projects in this series. Given all of these efforts, he sees the bioinformatics processing and visualization of single epigenome tracks as "essentially a solved problem," he says.

The major challenge now, in Bock's view, is to find ways to move much faster and more routinely through data—and, he jests, without depending on one-off heroic efforts of a bioinformatics postdoctoral fellow. The plan is to help researchers go directly from data analysis and visualization to biological insight and biomedical impact. He and his team focus on medical epigenomics. He thinks epigenetic drugs and epigenetic biomarkers can be expected to play a major role in personalized medicine. "DNA methylation is by far the most useful epigenetic mark for clinical diagnostics because it is stable and easy to analyze in clinical samples," he says.

This interest in DNA methylation led him and a team of MPII PhD students to develop the RnBeads software[2] to analyze DNA methylation in large cohorts. RnBeads is about developing and exploring hypotheses computationally without needing the guiding hand of a bioinformatician at every step, even though it is a deep dive into massive data sets.

These large data sets collected from consortia, various epigenome-wide association studies and cancer biomarker projects pose problems for genome browsers, says Bock. In the UCSC Genome Browser, researchers look at, for example, 10-kilobase chunks of the epigenome, but there are 300,000 such regions across the genome, he says. "Given a typical screen size, you could scroll for 150 kilometers from chromosome 1 to



This 2D representation of chromatin data compresses data all along the chromosome into a grid, here using a Hilbert curve. The chromatin states are color coded.

chromosome Y," says Bock. "We clearly need graphical discovery tools that help us home in on where the action is in the epigenome."

There are several approaches underway to address these issues, and Bock points to work by Kellis and others, who reduce complexity through data segmentation, such as with ChromHMM[3]; through imputation with ChromImpute[4]; or through aggregation with the help of a tool called epilogos (http://compbio.mit.edu/epilogos/) developed by Wouter Meuleman, a postdoctoral fellow in the Kellis lab. All of these approaches compress the data and highlight biological information that can again be plotted alongside the genome in a genome browser, says Bock.

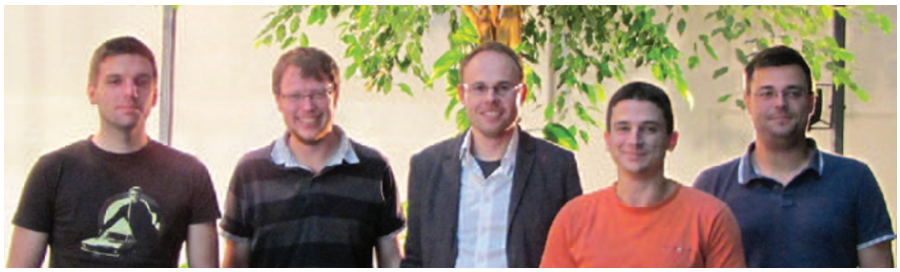
As Meuleman explains, epilogos aims to help users rapidly visualize hundreds or even thousands of epigenomes so as to prioritize genome regions of interest or to find epigenomic signatures. A paper about the software is in preparation in the Kellis lab.

Going wide, deep, small

As Harvard Medical School researcher Peter Park explains, now that thousands of epigenetic profiles are publicly available, the browser view gets crowded quickly when users explore the data. A more refined interface is needed to see what profiles are available and to allow appropriate tracks to be selected, he says. Browser tracks are, in his view, not sufficient for a meaningful exploration of the data.

To present information relating to the chromatin landscape in the fruit fly, Park and colleagues developed a data display approach to show specific chromatin features. To do this, they fold the chromosome geometrically[5]. "We provide a compact form of visualization for a browser track by folding it into a two-dimensional figure," says Park. They developed an online web application (http://compbio.med.harvard.edu/flychromatin/) to browse the folded views and the linear annotation side by side.

This highly compressed representation of chromatin states shows data all along the chromosome without having a linear visualization extend beyond the width of a display screen, explains Nils Gehlenborg, a postdoctoral fellow in Park's group who was not involved with the paper. The visualization applies a so-called Hilbert curve, which optimizes the mapping such that, on average, regions close to each other in the 1D sequence will be close in the 2D representation.



P. Lutsik

Epigenomics data visualization developers need to embrace teamwork, says Christoph Bock. From left to right, the developers of RnBeads and EpiExplorer: Peter Ebert, Fabian Müller, Christoph Bock, Yassen Assenov and Konstantin Halachev (not shown: Pavlo Lutsik).

With the software tool—EpiExplorer—developed in his lab, Bock[6] and his team are focusing on how to allow users to interact with epigenomic data in real time and, Bock says, "not just region by region as in a genome browser, but on a truly genome-wide scale." The approach uses elements of an Internet search engine to plow through epigenomic data.

A first search result delivers a large number of genomic regions, and each further step restricts the search until a user finds what he or she is hunting. That might be, for example, all blood-specific enhancers that are epigenetically altered in leukemia and located next to a cell-cycle gene, says Bock. "All of this is happening live and interactively, without any need to carry out complex analysis workflows."

Behind the scenes, EpiExplorer also draws on text search technology. Each regulatory region in the genome gets a kind of micro–web page. Fast search algorithms filter the region's data sets, and then the results are visualized.

The initial EpiExplorer server (http://epiexplorer.mpi-inf.mpg.de/) at MPII has been a proof-of-concept prototype. Now Bock and his team are working with Paul Flicek and colleagues at the European Bioinformatics Institute (EBI) to engineer a new version of EpiExplorer that can dig into data from the European Blueprint project or the International Human Epigenome Consortium.

As Flicek explains, genomics has been moving from a relatively flat description of the genome with genes, exons and other regions to a much more complex, nuanced description of chromatin activity. Displaying such high-dimensional data sets on a screen requires "some astute user experience (UX) design," he says.

Pioneering efforts in this area included the Galaxy Project's interactive data analyses and the extensions of the Genomic HyperBrowser that offered easy access to advanced statistics, says Flicek. EBI engineers are advising the Bock team for the next version of EpiExplorer in preparation for a possible deployment at the EBI, and he and his team are happy about the progress to date, says Flicek.

"Ultimately, excellent UX requires incessant trial and error, and all these projects are contributing fresh ideas," says Flicek. He and his team also keep tabs on 'big data' visualization tools outside of biology, which are growing by leaps and bounds. For epigenomics, he envisions an ecosystem of independent computational services that work hand in hand and also cross-pollinate one another with ideas.

Research needs will change and shift as epigenomics continues to mature. One area of interest to Bock is single-cell epigenomics. He and his team have published a way to reconstruct and visualize epigenetic cell state. The idea, he says, is to obtain DNA methylation fingerprints from a large number of single cells and to use single-cell data to directly reconstruct epigenome landscapes for cellular differentiation and cancer. An interactive visualization tool that does so is in the works in his lab.

Another interesting question, says Bock, is how to handle variation and uncertainty in epigenomic data visualization in a way that facilitates biological discovery. He wants to expand his concept of reference epigenome corridors that can be used to test whether a given cell line can be considered epigenetically equivalent to a high-quality embryonic stem cell. The corridor is a quantitative approach Bock has developed to establish the distribution of DNA methylation and gene expression in a reference set of cells. It can be used to overlap epigenome data such as DNA methylation for many different biological replicates of the same cell type in order to define the epigenetic variability for each position in the genome, he says.

This reference epigenome corridor can then be used to identify biological replicates that are strong outliers compared to the reference and to identify in which genomic regions they deviate. The method helps with quality control—for example, for experiments with pluripotent cells—and it can be generalized for single cells, with single cells taking the place of biological replicates that give rise to the corridor. In this fashion, says Bock, scientists can build so-called composite methylomes that fully account for single-cell variability in a given cell type.

The user's eye

Researchers and data visualization tool developers must continually consider the user perspective. To help provide answers to the questions that biologists might ask, says Park, more algorithms need to be integrated into these tools. Those questions might be about which factors bind to the promoter of a specific gene or about the cell types in which a given enhancer is active. For the time being, he says, most tools assist with answering only the question of what the epigenetic profiles look like around the gene that a scientist specifies.

As WashU's Wang explains, when he tests a new tool, he can tell whether it has been developed with users—biologists—in mind and with an understanding of the types of questions these investigators want to address. In other words, he says, "it takes more than a software engineer to develop a bioinformatics tool."

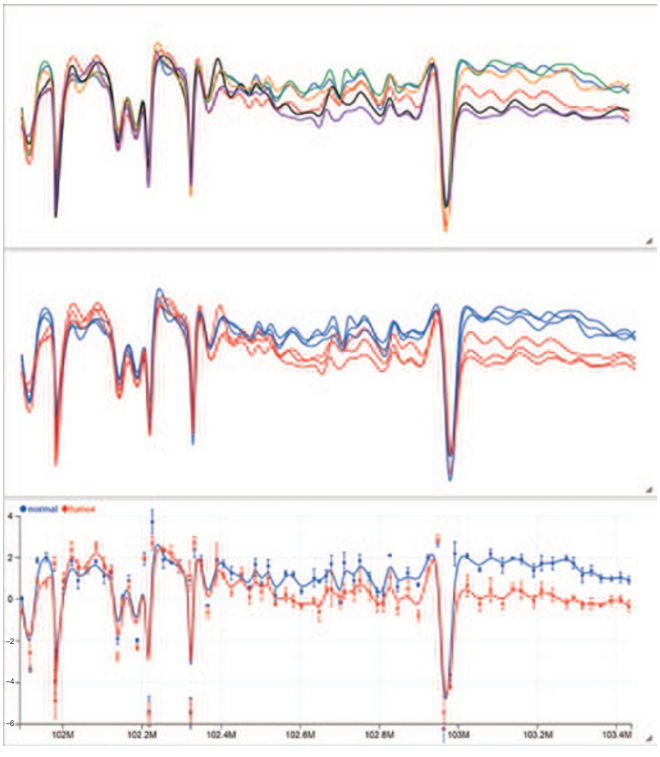
Bock tells his team and others to develop tools that are fast and that do not make the user wait. These tools should also be fault tolerant and make it easy to go back to the previous step if something has failed. "And why aren't there any 'undo' buttons in bioinformatics web services?" he asks.

In Park's view, a data visualization tool may seem inviting at first. But, he says, he finds that all too often the tools crash when he first tries to use them. In his view, "journals and reviewers must do extensive testing and insist on more stability of the software before accepting the paper about the tool."

Human imperfection is an important factor when developing or using tools for data visualization. The human eye is a great pattern recognition machine, says Bock. "Use it, use it extensively, but don't trust it," he says. "The eye sees interesting patterns, and our brain finds plausible biological explanations even in entirely random data." This phenomenon is also called apophenia.

Martin Krzywinski, a data visualization researcher at the BC Cancer Agency, is equally wary of this human penchant to see patterns. "We're distracted by the first and most obvious pattern," he says. Browsers need to be equipped with ways to set apart and to compact rows that are identical because the data sets are getting so large that it becomes impossible to look at everything. Instead, he wants to direct the gaze and the analysis to fine-grained differences.

These differences that need to be visualized are not just the epigenetic modification of a single-nucleotide polymorphism, says Krzywinski. Rather, it is more challenging and important to see the many ways in which differences can stack in these data sets, he says. For example, genomes from individuals show distinct somatic mutations. Researchers want to explore the different epigenetic modifications across the regions of the genome where those mutations lie.



F. Chelaru/Corrada Bravo lab

Interactive visualization of methylation levels in six colon tissue samples (top) helps users to differentiate tumor from other tissues (middle) and then to aggregate data into two series (bottom).

Beyond visualization

Visualization is just a start, says Bock, because observations need to be backed by statistical methods. He recommends that researchers always write papers about data visualization tools with the assumption that Reviewer 3 is going to be an expert statistician.

A challenging issue is the potential for data visualization to skew data analysis. As Wang explains, it is best to not think of these tools as black boxes. Rather, scientists should understand how the data are transformed and displayed. "Always understand the null hypothesis, the behavior of input control, and how the data were generated in the first place," he says. "Data display is just a means to help investigators sort through data easier and faster, while perhaps having some fun."

In working on epigenomic data visualization, says Flicek, scientists will want to keep in mind that the collected data are still an incomplete view of an individual cell's epigenome. Part of the explanation is that assays such as chromatin immunoprecipitation require a population of cells. Another aspect, he says, is that technology does not currently enable measurement of all possible chromatin or epigenetic marks with the same level of accuracy.

Flicek says what is also still lacking is an effective 3D resolution of cells that would help scientists understand all of the fine-scale epigenetic interactions. Despite the limitations, epigenomic data and their visualization "are incredibly useful," he says, "but it is important to remember that there is still more than meets the eye."

1. Chelaru, F., Smith, L., Goldstein, N. & Corrada Bravo, H. *Nat. Methods* **11**, 938–940 (2014).
2. Assenov, Y. *Nat. Methods* **11**, 1138–1140 (2014).
3. Ernst, J. & Kellis, M. *Nat. Methods* **9**, 215–216 (2012).
4. Ernst, J. & Kellis, M. *Nat. Biotechnol.* **33**, 364–376 (2015).
5. Kharchenko, P.V. *et al. Nature* **471**, 480–485 (2011).
6. Halachev, K., Bast, H., Albrecht, F., Lengauer, T. & Bock, C. *Genome Biol.* **13**, R96 (2012).

Vivien Marx is technology editor for *Nature* and *Nature Methods* (v.marx@us.nature.com).