GENEOGRAPHY

BEN FRY IS TAKING MAPPING INTO NEW TERRITORY WITH HIS PIONEERING VISUALIZATIONS OF GENOMIC DATA. **JANET ABRAMS** HEARS HOW FRY'S WORK IN COMPUTATIONAL INFORMATION DESIGN IS HELPING SCIENTISTS COMPARE THE HUMAN GENOME WITH THOSE OF OTHER SPECIES.

They look like hybrids of architecture, theatrical flats and musical notation: isometric structures whose banded vertical panels are interlaced by grey filaments, criss-crossing as if spun by a slightly crazed spider. In successive frames, the view shifts and the panels start to resemble skyscrapers, or city blocks reimagined by a muted-palette Mondrian.

But in fact these abstract representations are maps of a new kind: Haplotype Maps, visualizations of clusters of nearby genetic variants that tend to occur together along a given stretch of a chromosome. They were produced by Ben Fry as part of his Ph.D. in Media Arts and Sciences, which he completed at Massachusetts Institute of Technology (MIT) in 2004. Fry specializes in visualization of data whose structure and content are undergoing continuous change, with a focus on the burgeoning field of genomics.

Now working at the Broad Institute of MIT and Harvard (formerly the Whitehead Genomic Research Institute) effectively as its first resident information designer, Fry is developing new programming and visualization tools to enable genomics researchers to see what they are discovering in a highly abstract landscape. It is a landscape characterized by dramatic shifts in scale, massive quantities, constantly updated information, and critical variants yielded by tiny modifications in the arrangement of basic units. Sounds like a place you'd like to visit? Could be your own body, since the landscape under consideration sometimes happens to be the Human Genome.

Dr. Eric Lander, director of the Broad Institute, hired Fry upon completion of his doctorate, recognizing that he was charting important new terrain, in which the quantity of data being generated has, thus far, been almost inversely proportional to the quality of information design available to represent it.

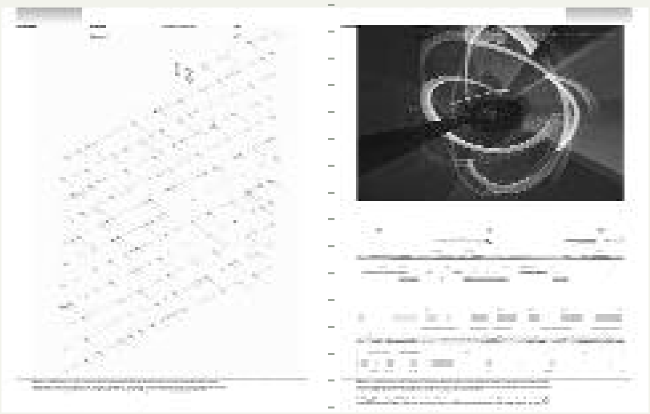
"Biology is undergoing a remarkable revolution right now, from being a laboratory discipline in which people studied their own particular problem, to becoming an information science," Lander explains. "A decade or two ago, a biological scientist typically worked on a specific component: one protein, one gene. It was as if they were studying the earth from ground level, by looking around them.

"But in the last 10 to 20 years, with the advent of genomics, biologists have been able to pull up to the 100,000 feet level and see the entire world of biology in one glance. We now know the whole sequence of the Human Genome, so we're presented with the 2-million genetic variants of the human population, and their correlation in hundreds of people; all the patterns by which genes are turned on and off in diverse tissues in the body; all the mutations that occur in cancer; all the networks and pathways by which signals are sent to the cell. The field is suddenly coming to grips with how to deal with all this data."

Fry became involved with the Broad Institute during his doctoral thesis, when he worked closely with its biologists on ways to represent the intricate pattern of correlation between nearby genetic variations in the Human Genome. "Ben came along and was very rapidly able to develop five or six ways of communicating different and important aspects of the data," says Lander. "There's just no substitute for visualizing data: you see patterns in it that you won't be aware of any other way."

* * *

An alumnus of the MIT Media Lab's Aesthetics and Computation Group (ACG) headed by John Maeda, Fry is clearly influenced by Maeda's emphasis on intellectual "dual processing": to gain admission to this highly selective graduate program, students must demonstrate



gazetteer I 6–7/7 pp. 262–263

Mapping maps

competence not only in graphic design, but also in software design; they develop these skills in tandem. Fry's portfolio includes numerous experimental software studies and, in collaboration with Casey Reas (a fellow ACG alumnus, now assistant professor in the Design|Media Arts program at UCLA), he has developed *Processing*, a new programming environment for learning computational design that has already attracted a tribe of 12,000 alpha-testers.

In his Ph.D. thesis, Fry drew an analogy between genomic cartography and conventional cartography, which synthesizes illustration, information design and statistics, and uses technological tools for implementation. "It was a useful way to frame things. In doing this new stuff on computer, people tend to presume it's completely different than what anyone else has done. But when I went to look for books and resources, I found there isn't nearly as much good material in the interactive work as in cartography." Over centuries, cartographic skills have been honed for tackling complexity and establishing hierarchy, to present rich information in a limited space. "Then the question is: how do you make cartography dynamic and what happens when you apply it to genetics?" In his Ph.D. dissertation, Fry writes:

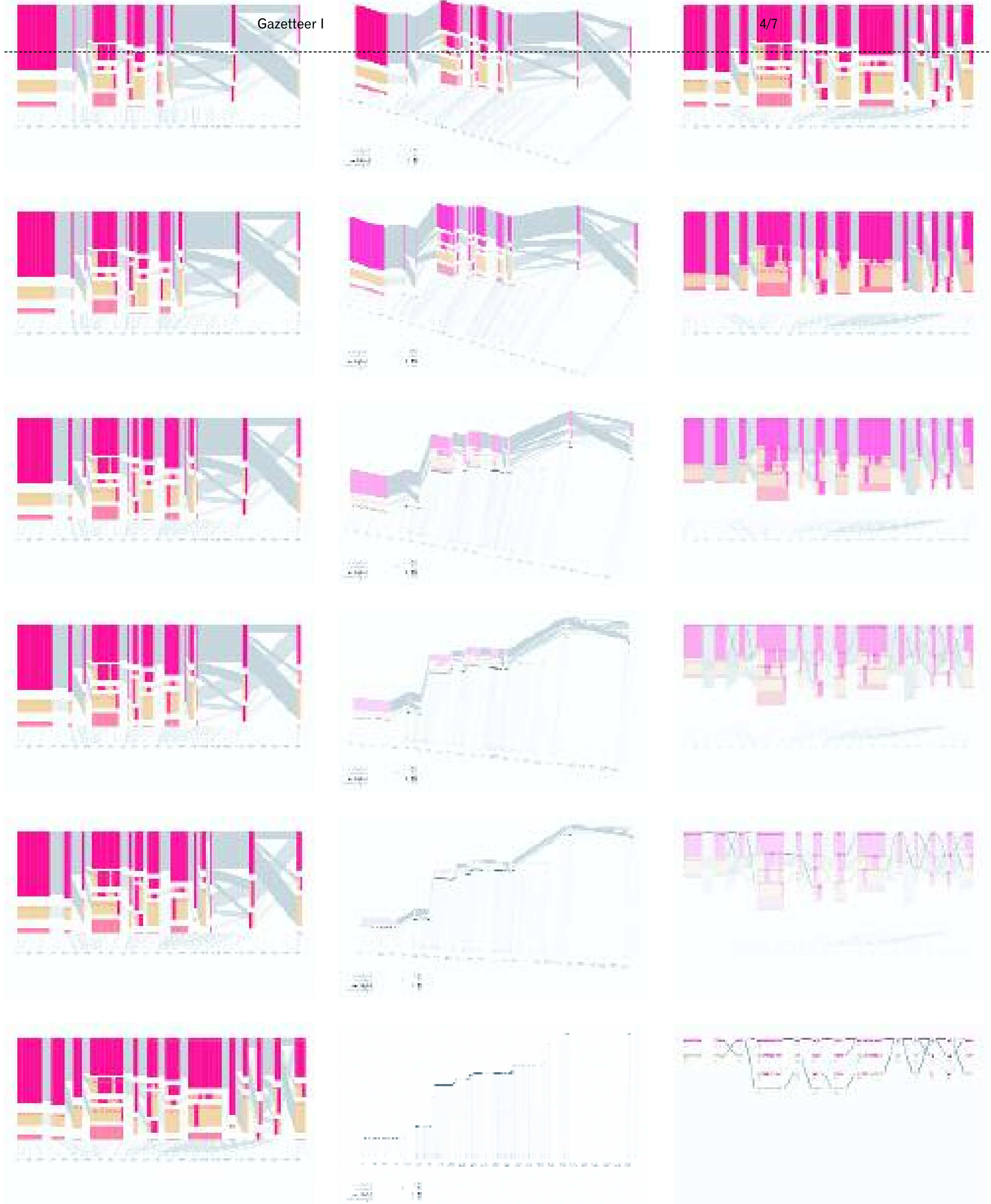
<http://acg.media.mit.edu/people/fry/>

> Rather than focusing on a final outcome, genomic cartography should be a flexible process around a dynamic diagram that can handle change. To attain this flexibility, software presentation methods must be grounded in variability—the positioning of elements must be semi-automatic, based on rules set by the cartographer. Rather than simply applying rules of graphic design, the designer must be able to abstract some of these rules, and implement them as active software elements.

The creation of visual tools for expressing ideas is not a matter of merely garnishing scientific findings, argues Fry. "I try to explain [to the scientists] that rather than seeing the visual thing as the icing put on top, to ornament a paper, it's much closer to the writing they're doing. You couldn't get away with a writing style of the kind that they use in the images; you'd get laughed out of the review process. So they recognize they need to have a certain degree of skill. Even among the 'priesthood', they need to be able to talk to each other. You can only move things along in a field as fast as you can communicate."

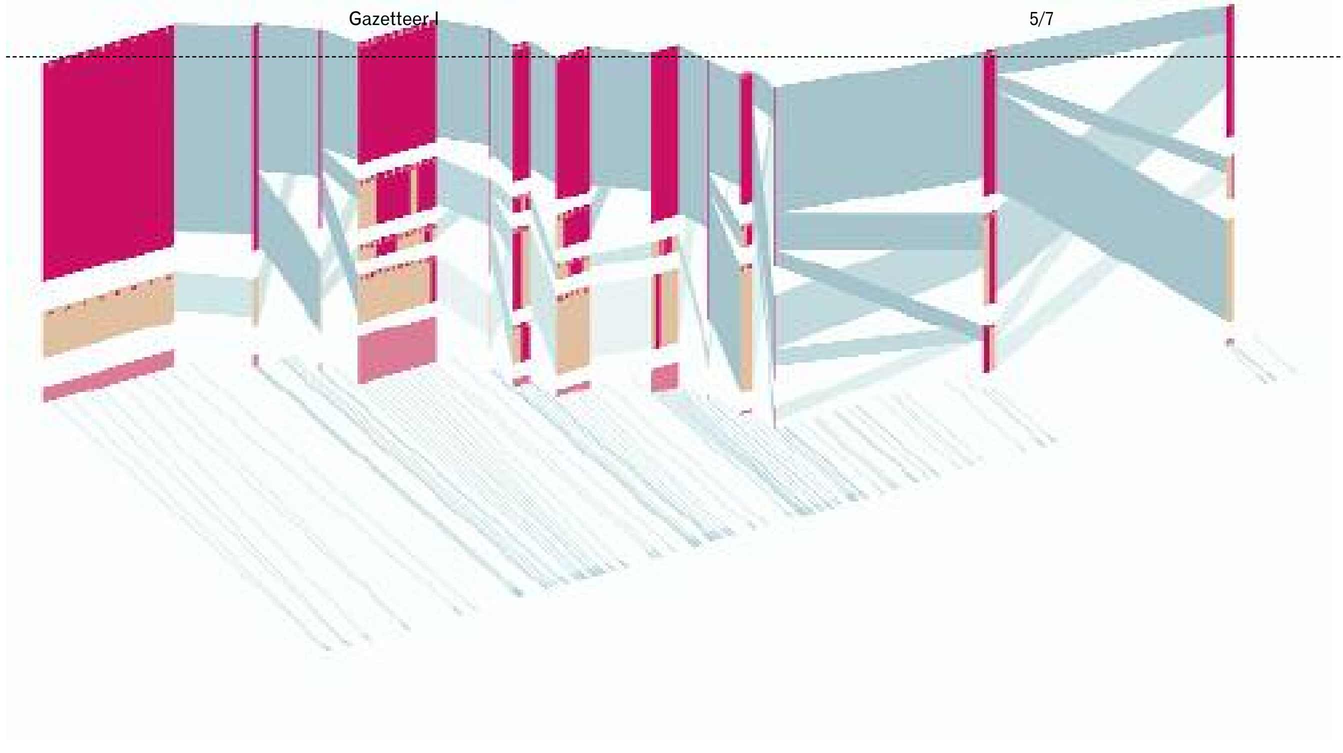
Standard genomics representations are barely decipherable, unless one is trained to read them. Produced using a small range of software visualization tools made by scientists, they suffer from lack of awareness of basic principles of information design; they are, at best, cluttered, dense and confusing; at worst, actively misleading. "How do you even start *looking* at one of these images?" Fry asks. "If you look at an enormously complex map of a city, you get some immediate understanding of where the streets are, what its features are. That's what's missing from this complex diagramming."

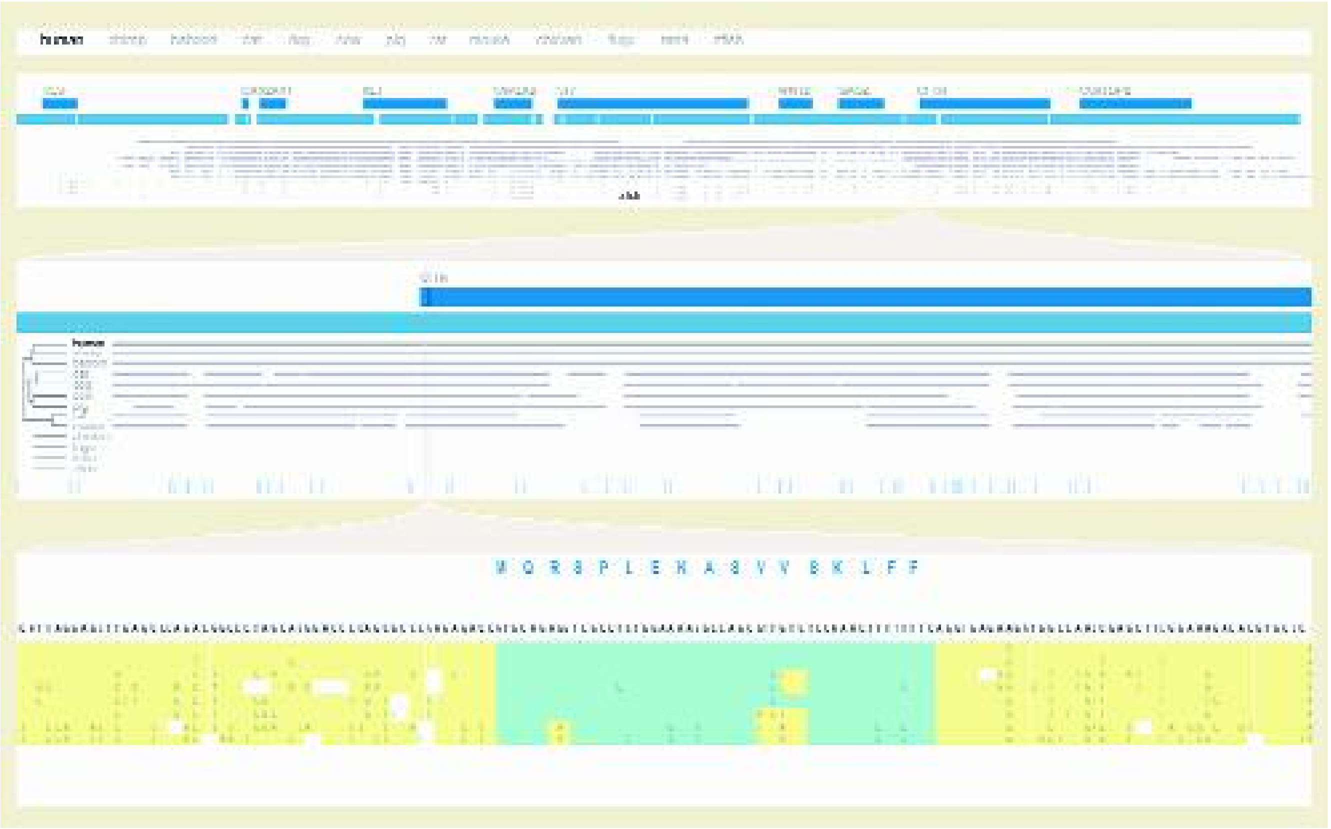
Fry's work in genomic visualization has so far tackled several different levels, from single chromosomes (human chromosomes 13, 14, 20, 21 and 22) to an entire genome, with side excursions into contexts-of-use, such as a prototype PDA *Genome Browser* that would enable scientists to read data on hand-held devices. He has tried out various aesthetic approaches to mapping massive amounts of known data, while also depicting the *terra incognita* of genes not yet identified, from the deep-space planetary orbit of his 2002 *Genome Valence* (a 3D view of BLAST, the algorithm most commonly used for genome searches) to the architectonic lattice of *Chromosome 13* (done in 2001). These studies are marked by a certain visual astringency: a translation, perhaps, of the Bauhaus design legacy inherited via Maeda's Media Lab predecessor, Muriel Cooper, inflected by Maeda's Japanese-inspired purism.



Ben Fry *Haplotype Maps*, 2004 / frames from an interactive application comparing genomic data in a group of 500 people. The difference between the genomes of two individuals can mostly be traced to single nucleotide polymorphisms (SNPs: single-letter changes occurring once per 1000 letters of genetic code). Related SNPs ("haplotypes") are shown grouped together; colors within each block represent percentages of each grouping; gray lines show how often given letter sequences are found in adjacent blocks.

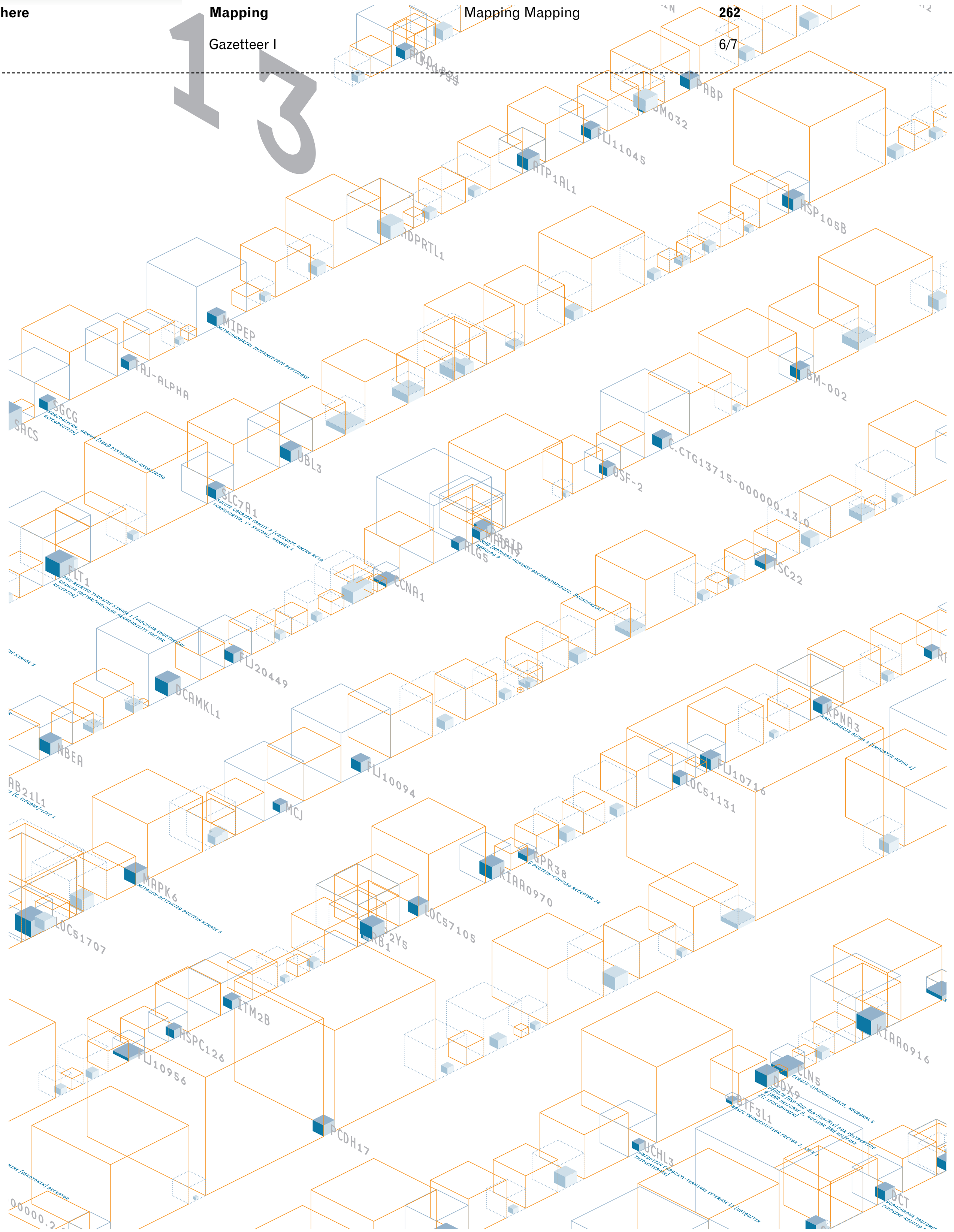
Each of the three columns, above, shows a different type of transition, from top to bottom. (left): SNPs shift from even spacing to expanded spacing, to make individual letters more easily legible, while thin grey lines maintain a connection to the SNPs' true positions along the nucleotide scale on the horizontal axis. (center): from isometric 3D to aerial 2D, corresponding to the LDU (Linkage Disequilibrium Units) plot, shown in lowest frame. (right): from graphical to quantitative representation / see "Geneography," pp. 264–267.



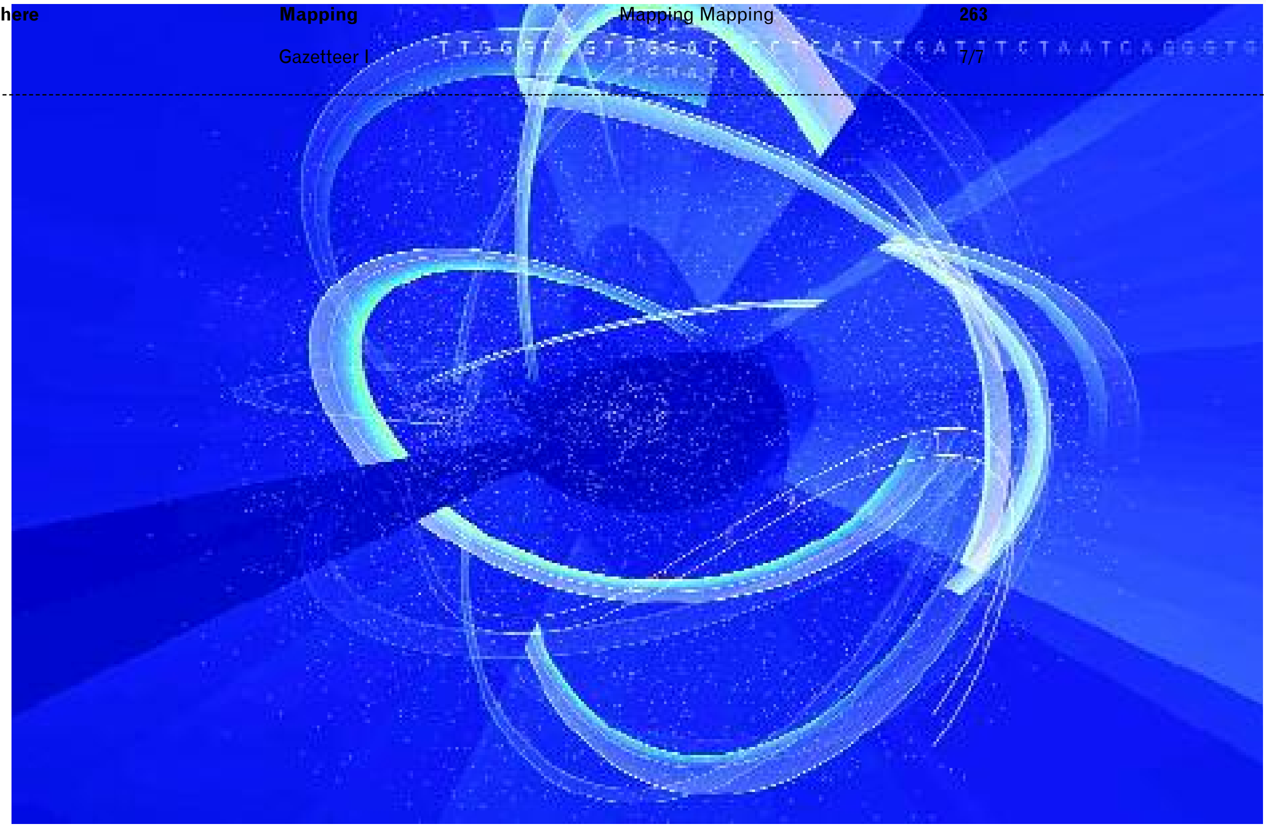


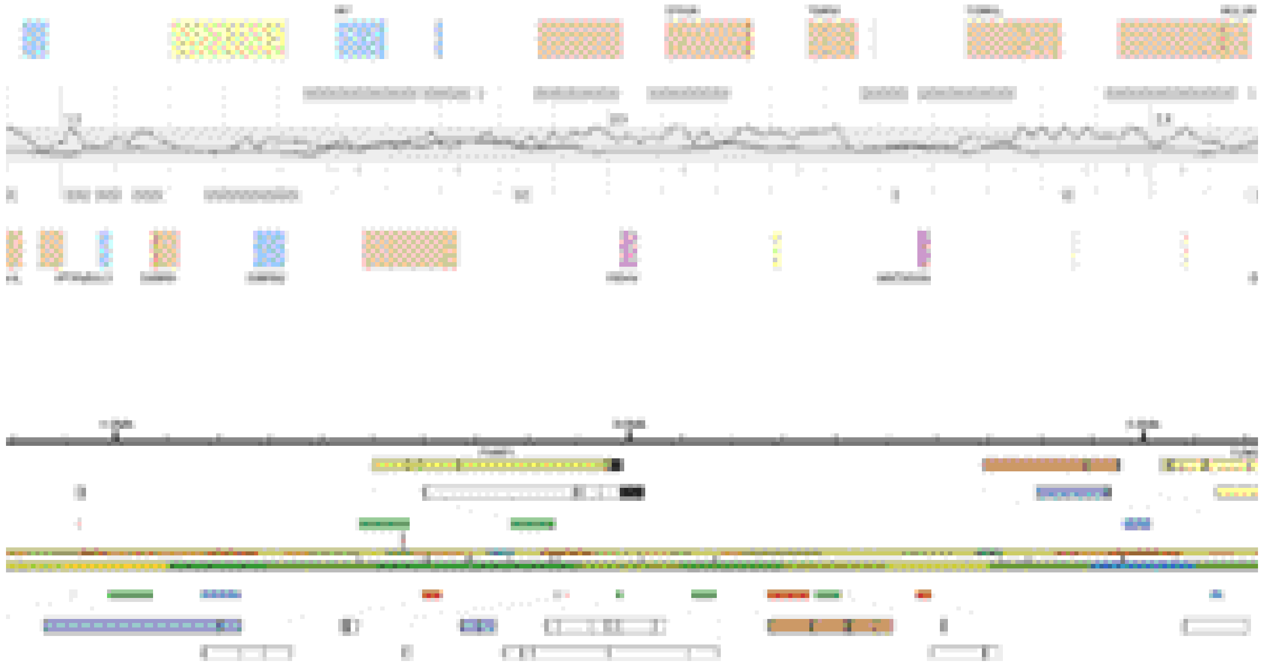
Ben Fry *Isometric block*, 2004 / (top): enlarged frame from interactive application comparing genomic data in a group of 500 people. In this view, each block is offset slightly along the z-axis, so that the gray lines depicting the transitions between blocks (representing related SNPs) can be seen more clearly. A "false" 3D isometric projection is employed that allows the data to be shown while preserving the linear scaling of the nucleotide scale in the horizontal axis / see "Geneography," pp. 264–267.

Ben Fry *Genome Comparison*, 2005 / (bottom): prototype interface for comparing the genomes of various species, including human, mouse, chicken and zebrafish, to identify areas of genetic similarity, here examining a region including the CFTR gene, which is tied to Cystic Fibrosis. Three different scales of data are shown in three linked horizontal bands that show, top to bottom, 1.8 megabases (1.8 million A, C, G, T letters), 50 kilobases (50,000 letters) and 150 base pairs (150 letters) / see "Geneography," pp. 264–267.



Ben Fry *Chromosome 13*, 2001 (detail) / map of a single chromosome from the human genome. Yellow wireframe boxes signify gaps between genes, areas where the letters are considered "junk DNA." Blue areas indicate where genes exist, and the blue wireframe boxes are proportional in size to the start- and end-point of the gene / see "Geneography," pp. 264–267.





Ben Fry *Genome Valence*, 2002 / (top): a visual representation of the BLAST algorithm, commonly used for genome searches, allowing comparison between genomes of three different organisms.

Ben Fry *gff2ps*, 2003 / (middle): section of the human genome as represented in Fry's redesign of *gff2ps*, the software program used to read GFF, a file format used for annotating gene sequences. (bottom): **Josep Francesc Abril Ferrando** and **Roderic Guigo Serra**, original design of *gff2ps* showing same section of human genome / see "Geneography," pp. 264–267.

In one virtuoso section of his dissertation research, Fry painstakingly unpacks the map of the Human Genome as represented in *gff2ps*—the software program created to read General Feature Format (GFF), a standard file format for annotating genomic sequences. He demonstrates how poor use of basic visual elements—such as color, line thickness, sizing and spacing, category indication, and alignment of data tracks—conspire to render this widely-used diagram far less legible than it might be. He then proposes a clean-up and a way to implement the improved design as a software tool (developed using *Processing*) that could be applied to various genomic data sets.

Part of the challenge in reading such diagrams is that one needs a basic grasp of the terminology (arcane, at least to a layperson) used to describe the different categories of genomic information. First there are the building blocks of a genome, the set of bases represented by the letters A, C, G and T, which stand for the lengthy names of chemicals. C, for example, is the abbreviation for Cytosine, which is short for “deoxycytidine triphosphate.” Chromosomes are made up of DNA, each a long polymer molecule composed of many such bases. The shortest human chromosome, Chromosome 22, comprises about 50 million letters; the longest, Chromosome 1, about 250 million. A genome—the collective name for a particular organism’s complete set of chromosomes—may be composed of anything from a few thousand letters to billions of letters. To give a sense of scale, the 3.1-billion letters that make up the 24 unique chromosomes comprising human DNA would extend more than five thousand miles if printed out as a single line of 12-point Times Roman.

* * *

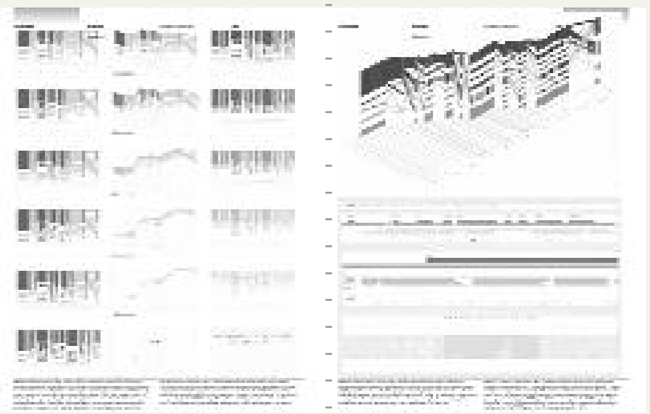
At the Broad Institute, Fry takes part in weekly group meetings with its team of computational biologists and has clearly learned to “speak biology” fluently enough to participate meaningfully in their discussions. “Our biggest issue,” Fry explains, “is how to compare the genomes of different species in a really clear picture. So we can say: ‘Here’s an evolutionary tree, here’s the data for all the animals.’ If you look for the differences between the human genome and the mouse genome, the mouse has over ten times as many olfactory genes as the human. By contrast, we share the HOX region of genes—which is involved in the development of our limbs and structural core—with a striking number of species. Right now, scientists are looking at the area around CFTR, the gene tied to Cystic Fybrosis: a region of 1.7 million letters in the human genome. The task is to find that region in the chimp, dog, mouse, fruit fly, zebrafish, etc., and compare with all those organisms.”

Fry is currently creating interactive maps for browsing individual genomes and comparing those of different species, as part of the International Haplotype Mapping project, a $100-million research venture involving six countries, in which the Broad Institute is a major partner. The project attempts to show, at a glance, variations between people, families and populations. Fry’s interactive visualizations—which allow data comparisons from multiple viewpoints—have already proved “incredibly useful in helping to digest the data,” according to Dr. Eric Lander, who explains: “These kinds of visual representations become ‘memes’—cultural units that spread very rapidly when they click for people.”

Most of the variation between the genomes of two individuals comes down to Single Nucleotide Polymorphisms (SNPs, pronounced “snips”), single-letter changes that occur once every thousand or so letters of genetic code. Fry’s Haploptype mapping visualization takes data for 103 SNPs in a population of 500 different people, and offers a quick visual way to grasp patterns in connections between sets of SNPs, and adjacent clusters.

For example, in each frame of the sequence shown in the left-hand column on p. 260 (“even to expanded spacing”), the block at the far left indicates that some 70 percent of the study population has the same haplotype block structure (shown in dark red). The colors in each row indicate one of only two possible variations for each SNP, with the most common in dark red, and less common in pale orange; the salmon pink layer at the bottom indicates variations that occur in less than five percent of the population.

Fry’s interactive application allows the user to move between alternate views of the same data, prioritizing different aspects as needed. Shuttling between them enables “a tight coupling between the qualitative—useful for an initial impression and getting a ‘feel’ for that data—and the quantitative—necessary for determining specific frequencies of haplotypes of interest for specific study,” according to Fry.



Mapping maps

gazetteer I 4–5/7 pp. 260–261

When the important thing is to be able to read the individual letters on a chromosome, a view can be selected that gives each letter (or SNP) equal spacing; a thin grey line links to a real scale at the base of the diagram, showing its true position along the chromosome. By clicking a button, the diagram shifts to show SNP columns once again in their correct proportions, giving a sense of their pattern of relatedness.

The definition of a block is still under debate, so Fry’s software lets users modify the mathematical parameters that determine block boundaries, by tweaking the algorithm via an on-screen slider. Block transitions can be emphasized by shifting the view from 2D to 3D, and offsetting the blocks along the z-axis. For those who like their data quantitative, another view shows just raw letters and their respective percentages.

Given the exponential scale of the data being navigated, it might seem tempting to devise a means to travel visually from millions of letters down to the individual letters of a genome. But as Fry points out, “What gets missed in *Powers of Ten*-style continuous zooms are the plateaus along the way where interesting things are happening. You need to design for each of those plateaus, where you see very different phenomena, at the relevant scale. When you look at the 1.7 million letters around the CFTR gene, there are seven different genes to consider within that region, and the first thing to figure out is just where they start and stop. At the 250,000-letter scale, there are other kinds of things to consider. At the 150-letter scale, you can read all the letters in one go.” In Fry’s recent *Genome Comparison* chart (2005)—a prototype interface for comparing the genomes of several species, from human to zebrafish—these exponentially different scales of data are presented in linked horizontal bands.

He makes an analogy to designing type for different scales of reading. “When you take a 12-point font and blow it up, it’s not the same as a typeface designed at 72-point. It’s like the difference between a headline face and a book face, optically.”

* * *

There are pluses and minuses to being a pioneer in a field, Fry notes. “The downside is that I don’t have many people to compete with. At conferences, there isn’t a visualization section, but I suspect that’s going to change.” (One of the few other designers working in this area is David Small, who also earned his Ph.D. at the MIT Media Lab, and is working on another user interface design commission for the Broad Institute.) Fry anticipates building a team with another designer and software engineers, to expand the Institute’s representation capabilities.

Meanwhile, its director, Dr. Eric Lander, feels that “some of this just involves a creative wandering and exploration that has an aspect of individual creativity. I want to give Ben the chance to turn the data over in his mind, take various people’s problems, and come up with really unusual, quirky and insightful ways to visualize them.”

Asked how he chooses what to work on, Fry says “it’s a combination of me speculating and asking the scientists: ‘What do you care about at this level?’ Then saying to them, ‘Actually, you need *this* level, and *this* set of features.’ The scale of the projects is very big. You go region by region, half-a-million letters of code at a time.”