The Visual Representation of the Human Genome

David Stairs

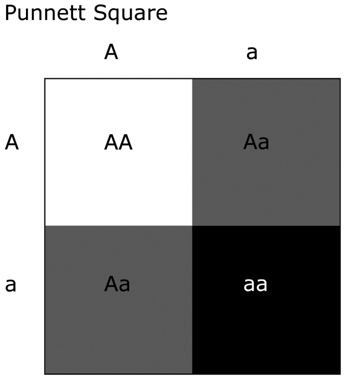
Introduction

On June 26, 2000, before foreign dignitaries and hundreds of distinguished heads of government, business, and science, J. Craig Venter and Francis Collins joined President Bill Clinton to jointly announce the first sequencing of the human genome. As remembered by Venter, "On the great day, happily, all the rivalries were swept aside by everyone's feeling of being part of an historic achievement… As the President, Francis, and I walked together from the hall into the East Room of the White House, a band struck up 'Hail to the Chief,' and we entered to face a standing ovation. Two large plasma screens carried a live video linkup to Downing Street and the British Prime Minister, Tony Blair."[1] Thus is described a moment of relative calm in one of modern science's most contentious rivalries: the one between the National Center for Human Genome Research headed by Collins, and Venter's company, Celera Genomics. As history now remembers it, it was a premature announcement of an only partially completed sequence.

The government program had been launched in the late 1980s, aiming for a completion date of 2005 for sequencing the genome. As the 1990s rolled on, the sprawling public program with several labs spread across America, Britain, and Japan had made mapping a prerequisite to sequencing, but it clearly was proceeding too slowly to reach its goal. Meanwhile, Venter's innovation—a whole genome shotgun assembly sequencing method—eventually proved both its quality and speed on efforts to sequence *Haemophilus influenzae* in 1995 and *Drosophila melanogaster* in 2000. Eventually, the government's program and Celara collaborated to present a partially sequenced genome at the White House.

By 2007, Venter released a complete sequence of his own genome—an amazingly complex visual mapping of all 46 chromosomes. How this map came to be, its predecessors, and its significance to the history of both biology and design are the topics of this paper.

[1] Craig Venter, *A Life Decoded*, (New York: Penguin, 2007), 311-12.



Punnett Square

	A	a
A	AA	Aa
a	Aa	aa

Figure 1
A Punnett square demonstrating heritability of dominant and recessive characteristics after Mendel.

Informatics vs. Infographics

The opening line of Edward Tufte's classic text, *The Visual Display of Quantitative Information,* states that "Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency."[2] The hallmark of communication design is its ability to visually represent data for the purpose of easy comprehension. Here, one might think of how dramatically Henry Beck's 1931 Underground Map for London Transport changed such representations, or of the way graphic user interfaces (GUIs) propelled the PC revolution in the 1990s. Graphically presented data is also important to scientific visualization. These days, most data visualization is computer generated. Unfortunately, the explosion of computer graphics has ushered in such a trivialization of visualized data that a distinction must be made between popular and often rhetorical *infographics,* of the sort one sees in the daily paper, and *informatics,* or the scientific visualization of data.

I take as my point of departure for scientific informatics Dmitri Mendeleev's periodic table of the elements. During the 1860s, several people were trying to make sense of the relationships between the then-known elements. Mendeleev equated the chemical properties of elements with their atomic weights. His visual representation was a simple table, the rows and columns of which are familiar to anyone who today uses the table function of a word processing program. Mendeleev's insightful groupings enabled him to actually *predict* the existence of as yet undiscovered elements. However, visualizing his concept in graphical form didn't require that he have any special skill in art.

Over the years, many variations on Mendeleev's concept have been rendered—some in table form, some in other formats. And these days, of course, we find trivialized variations on the periodic table concept for everything from desserts to video game characters to commodities returns.

Another, equally famous example of scientific informatics is the Punnett Square (see Figure 1). After 1900, William Bateson had Gregor Mendel's 1865 paper on plant hybridization translated into English. Working with Bateson, Reginald C. Punnett helped establish genetics at Cambridge, developing his famous square diagram as a visual means of predicting the outcome of a cross between two alleles (i.e., forms of a gene) thereby determining the probability of the offspring's genotype. Both the periodic table of elements and the Punnett-square of pea alleles, known to generations of school children, are examples of graphic diagrams that were generally accepted and expanded upon by subsequent generations of researchers—a form of group consensus that has defined scientific informatics for over a century.

[2] Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 1983),13.

The First Genetic Maps

When I was a 16-year-old biology student at Christian Brothers Academy in Syracuse, NY, Brother George Mason guided us through the experimental method Thomas Hunt Morgan and his assistants had developed at Columbia 60 years earlier. Our subject, of course, was the *Drosophila* fruit fly, and our goal was to trace genetic inheritance through eye color. Curiously, a vision-related characteristic—color blindness—was the first discovered gene-linked human disability.

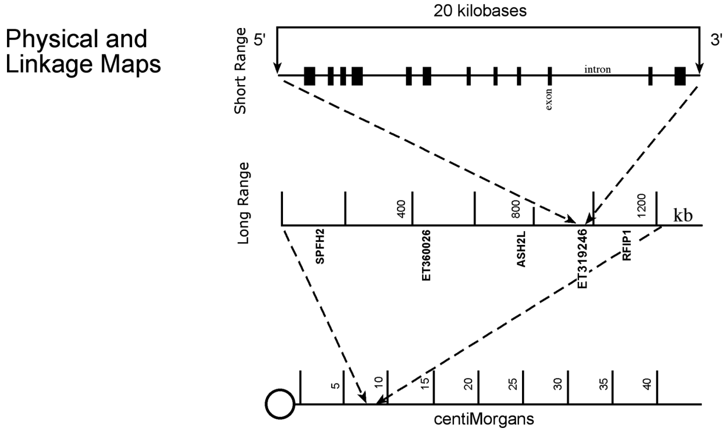
In 1890, Weismann had described the process of meiosis, or how germ cells recombine. Researchers had known of the existence of chromosomes in the cell nucleus, but for a number of years they had been puzzled that sperm and egg cells were haploid, containing only half the number of necessary chromosomes for reproduction. Some biologists were horrified by the idea of Mendel's theory of heredity, suggesting hidden, or recessive, characteristics rather than qualities based on visible characteristics. Then, in 1895, Wilson wrote: "…inheritance may, perhaps, be affected by the physical transmission of a particular chemical compound from parent to offspring."[3] He was referring to DNA, the Rosetta Stone of genetics. By 1900, genes were just beginning to be linked to chromosomes, but the characterization of DNA and the means for observing these sub-microscopic entities were still many years in the future.

By 1902, after researchers increased their observations of meiosis, Walter Sutter not only determined the stages of meiosis, but also made a crucial observation—that chromosomes come in matched pairs, based on size and shape. He called these pairs *homologous* (same proportion) chromosomes. It was later determined that these "homologs" also carry the same genes but that these genes, because of mutations, may vary in form (e.g., eye color). These different forms of a gene were called *alleles.* Moreover, it was observed that these homologs pair during meiosis and exchange genetic material between themselves in a process Morgan called "crossing over." Identifying this exchange of genetic material would be crucial to all future understandings of inheritance.

In 1910, Alfred Sturtevant was an undergraduate student who volunteered in Morgan's lab at Columbia, the famous "fly room." Mendel's 1865 essays on the inherited traits of peas had been rediscovered at the turn of the century by Hugo de Vries and Carl Correns, and only four years had passed since Bateson introduced the term "genetics" at the Royal Horticultural Society's convention in 1906. At that time, the narrow portion of the chromosome, or centromere, was known and would become an important reference point for what followed. One day Sturtevant had a

[3] Alfred Sturtevant, *A History of Genetics* (New York: Harper & Row, 1965),104.

Figure 2
Linkage Map.



flash of insight. As he described it, "I suddenly realized that the variations in the strength of linkage already attributed by Morgan to difference in the spatial separation of the gene offered the possibility of determining sequence in the linear dimensions of a chromosome. I went home and spent most of the night in producing the first chromosome map."[4]

Sturtevant's insight was of a *spatial* nature; he had noticed that the frequency with which two genes recombine relates to their distance from one another on their chromosome. This insight generated the need to visualize that distance diagrammatically. Genes with greater physical separation are likelier to mix during meiosis, while those closer together are more likely to be inherited together, as if "linked." Greater distance equated to an increased chance of recombination frequency. Using recombined frequencies, Sturtevant drew a linear diagram, thus creating the first "linkage map" showing relative gene location.

A linkage map unit, called a centimorgan in honor of Morgan, equals a 1% (1 in 100) recombination frequency, based on the recombination rate of same-site homologous chromosomes during meiosis. A modern example of Sturtevant's mapping technique appears in Figure 2.

The relative locations of genes and gene markers are first represented as a small fragment, 20,000 base pairs of the gene ET319246. This gene is itself one of several genes on the longer range 1.2 million base pair fragment pulled from the linkage map. The visual technique of representing increased orders of resolution by linear nesting, as illustrated here, is an example of how Sturtevant's idea was later expanded.

From Sturtevant's initial insight and mapping, research continued to develop; according to Sturtevant, "The first major undertaking after 1913 was the mapping of the new genes as they became available. Here again, while we all took part, it was Bridges who did most of the spadework, and who gradually accumulated and organized the data to produce the maps; and with the

[4] Peter J. Russell, Stephen L. Wolfe, Paul E. Hertz, Cecie Starr, *Biology, the Dynamic Science* (San Francisco: Brooks/Cole, 2008), 259.

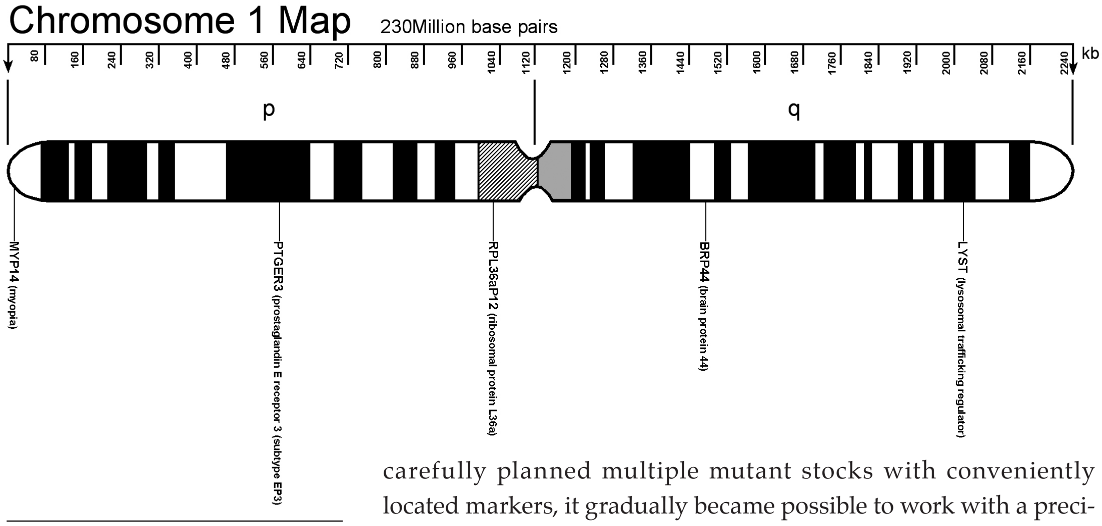


Figure 3
Chromosome Map.

carefully planned multiple mutant stocks with conveniently located markers, it gradually became possible to work with a precision that was heretofore impossible with any other material."[5]

These early maps were approximations. They didn't show precise distances between genes, but only relative positions. During the 1920s, Sturtevant did determine that *Drosophila's* genes were in linear order by proving that closely related species had similar mutations, as a result of crossing over, that were allelic and therefore probably identical. His proposals of linear arrangement resulted in the evolution from physical maps through linkage maps to our classic conception of chromosome maps, or *ideograms.*

In the chromosome map shown in Figure 3, a much larger distance is represented because chromosomes contain many genes. The centromere is pictured at the middle of Chromosome 1. Five of the approximately 4,000 genes that have been mapped on this chromosome are shown in their relative locations. Linkage suggests that, during the crossing over that occurs at meiosis, the genes at greater distance from the center are likelier to exchange material.

Molecular Biology, and Beyond

In the years before and after World War II, scientific attention was turning from high-energy physics, which had been the darling discipline of science for decades, to biology. Max Delbruck, a German physicist, moved to California in 1937 to pursue his interest in biology at Caltech. He began to study the viruses of bacteria, or "phages," and was an early proponent of applying mathematics to make quantitative predictions in biological experiments. His course in bacteriophage genetics at Cold Spring Harbor and his promotion, with Salvador Luria, of the "Phage Group" served as an inspiration to many young scientists in the early days of molecular biology.

Two of these newcomers, James Watson and Francis Crick, met at the Cavendish Laboratory of Cambridge University. Their search for the physical structure of DNA, so colorfully described in Watson's 1968 book, *The Double Helix,* actually relates events that

[5] Sturtevant, *A History of Genetics,* 53-54.

took place 15 years earlier. In 1952, they embarked on their now legendary undertaking. Basing their suppositions on earlier research by Maurice Wilkins, Watson and Crick struggled to understand the "tetranucleotide." Competing with a team led by Linus Pauling, and aided by the brilliant X-ray diffraction spectroscopy of Rosalind Franklin, the two young scientists eventually correctly modeled the molecule's helical structure.

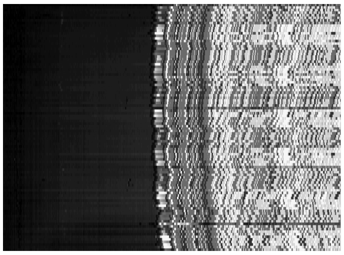


Figure 4
Automated Sequencing Array. Each of the 98 horizontal lines represents one capillary tube DNA sequence. The banding at the center is the result of an attached end sequence. The four bases, Adenine, Cytosine, Guanine, and Thymine, are represented by four distinct hues: green, yellow, red, and blue. (courtesy of Jiping Wang, Central Michigan University).

This work, informing decades of subsequent research, did not directly advance genomic mapping techniques. Rather, in unlocking DNA's structural mysteries, Watson and Crick aided research leading to the rise of molecular engineering as we know it. Crick's "central dogma"—that DNA translates to RNA, which organizes protein synthesis—led to efforts not only to locate human genes, but also to discover for what proteins these genes coded. The confluence of discoveries leading inexorably toward whole genome sequencing was made possible by other researchers building on Watson and Crick's work. During the 1950s, for example, Margaret Oakley Dayhoff pioneered the first computerized databanks of DNA sequences. Her 1965 book, the *Atlas of Protein Sequence and Structure,* contained data on all of the 65 then-known protein sequences.[6]

Less than a decade after Watson and Crick published their landmark 1953 paper on the structure of DNA, two Philadelphia researchers, Peter Nowell and David Hungerford, noticed that the blood cells of patients with chronic myelogenous leukemia (CML) had an unusually tiny chromosome. In 1973, Janet D. Rowley concluded the chromosome was the result of a translocation of material between chromosomes 9 and 22. At a time when the genetic underpinnings of disease were unclear, the discovery of this abnormality—dubbed the Philadelphia Chromosome—marked the first time a specific genetic defect was linked to a cancer, paving the way for drugs that targeted the defect and turned this rare leukemia into a manageable disease.

In 1970, Hamilton Smith purified the first restriction enzymes. These "molecular scissors" were used to cut DNA at specific sites. Herbert Boyer and Paul Berg used these enzymes to splice viral and bacterial DNA in 1972, creating the first recombinant molecule. Boyer's company, Genentech, founded in 1976, went on to synthetically manufacture human growth hormone and insulin.

Frederick Sanger in the United Kingdom and Walter Gilbert in the United States independently developed new techniques for sequencing DNA using gel electrophoresis—a technique later superseded by automated sequencing (see Figure 4). At this point,

[6] Molly Fitzgerald-Hayes and Frieda Reichsman, *DNA and Biotechnology* (Burlington, MA: Academic Press, 2009), 150.

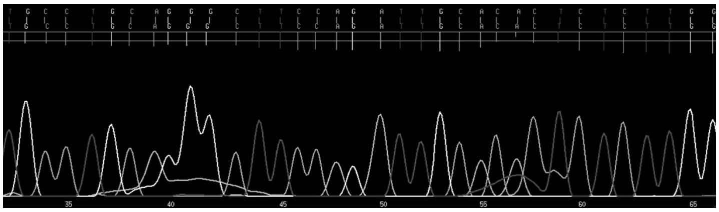


Figure 5
A chromatogram, or trace file, is what researchers actually look at when reading a sequence array. Amplitudes of each of the four base pairs—A, T, C, and G— are represented by a different color. (courtesy of Eric Linton, Central Michigan University).

peeking behind the veil of nature to read the specific order of the amino acids that code human proteins finally became possible. In 1977, PhiX174 became the first organism to be fully sequenced. Still largely a matter of guessing, identifying the location of genes by custodial "tagging"—first with radiological substances and then with fluorescent dye tags—eventually led to automatic sequence reading (see Figure 5). In these early "BI" (before Internet) days, gene sequencing was usually accompanied by experimental characterization of the gene's biochemical function; so few genes had been sequenced, an effort was made to understand how they worked. By 1980, when the U.S. Supreme Court overturned a lower court ruling allowing genetically modified organisms to be patented, the age of molecular engineering had truly arrived.

Computational Genomics

In the mid-1980s, the Department of Energy (DOE) was actively involved in genomics. Responsible for management of the nation's high-energy labs at Los Alamos and Lawrence Livermore, DOE scientists were interested in radiation's effect on genes. Breaking up of DNA into sequenced tag sites (STSs) (i.e., long sequences of tagged DNA) and expressed sequence tags (ESTs) (i.e., short DNA copies made from the ends of messenger RNA (mRNA)) had begun to lead to automated sequencing using the raw computational power of algorithms that were able to analyze large-volume throughput of sequence data. DOE scientists and engineers—modelers of the chaos of nuclear explosions—oversaw some of the nation's most formidable computing power, and today they still run the Joint Genome Institute (www.jgi.doe.gov/).

GenBank, which began in 1982 as a Los Alamos project and is now housed at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/), was moved to the National Institutes of Health in Bethesda, MD, at the beginning of the government-sponsored Human Genome Project. James Watson, who joined as the project's first director from 1988 until 1992, had by that time headed up the research institute at Cold Spring Harbor for 20 years, and research there had uncovered the genetic roots of many human diseases, including cancer. GenBank was designed as a

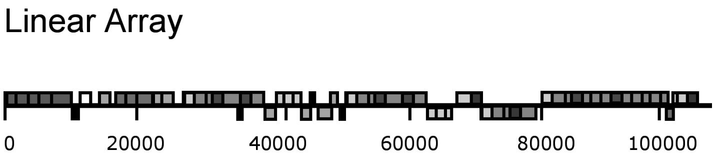


Figure 6
Linear representation of a bacterial gene measured in 20,000 base pair increments. Genes are color-coded. Top half runs 5'-3', bottom half is reversed.

repository for any published genomic sequences, but because the rate at which they were published was limited by techniques then available, the archive was small and intensely studied.

Much of the work at that point was theoretical because of the paucity of data, says Owen Smith, of the University of Maryland, who worked with Craig Venter in the early 1990s. “There were people asking genomic questions at the bench…like trying to figure out the amount of DNA content in a cell, the number of chromosomes, the rate of mutation, and thinking about what a genome was with a chalkboard.”[7] People were thinking about DNA in electrical terms, like a signal in a wire, and researching various physical means for measuring its content. “There was a point in history,” he says, in which “basically every sequence in the public archive had real meaning, as in it was intensively studied with bench experiments. Now all that data is an endangered species because it is swamped out by the high throughput data coming from genome projects.”[8]

This perspective is corroborated by Gary Zweiger, who in *Transducing the Genome* writes:

> In the early days of molecular genetics, genes were identified on the basis of their function, but when sequences began gushing into the databases like water from a hydrant, designations of function began to lag. Currently, there are many thousands of these orphaned genes, poor un-christened protein-coders that are nonetheless rich in concealed information.[9]

In those “early days,” a variety of approaches were used to visually represent genomes. As had been the case for decades, some researchers, following Sturtevant, used the linear model (see Figure 6). This model was not a matter of chance. DNA has a natural *directional linearity,* replicating with what scientists call a 5′ (5-prime) to 3′ (3-prime) polarity, or from the phosphate end to the hydroxide end. According to Eric Linton, a biologist at Central Michigan University, the convention of linear representation arises out of a consensus among scientists for purposes of comparative research. A graphic that follows the inherent logic of the substance represented is more likely to be universally accepted and gradually improved.

Graphic notation of genetic information also can often be found in a circular format. Because bacterial chromosomes naturally assume a circle, representing their genomes in circular form

[7] Personal correspondence (March 4, 2011).

[8] Ibid.

[9] Gary Zweiger, *Transducing the Genome* (New York: McGraw-Hill, 2001), 140.

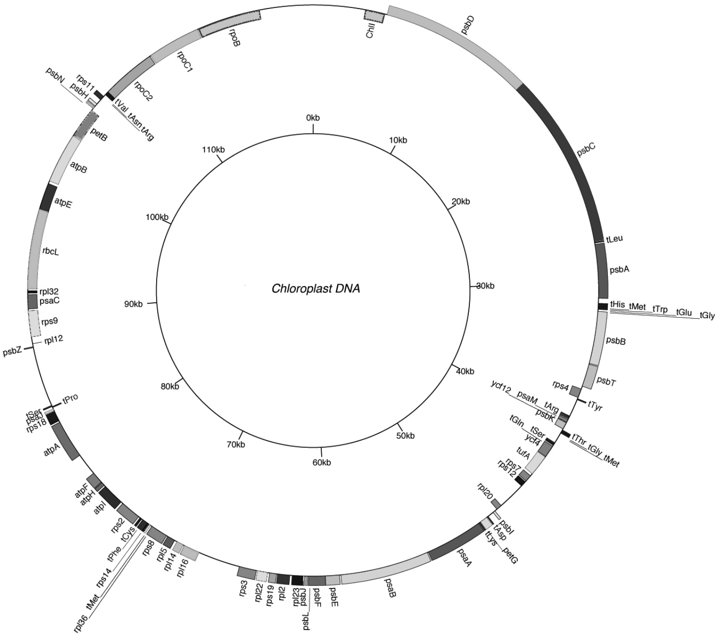


Figure 7
Chloroplast genome (courtesy of Eric Linton, Central Michigan University).

is logical. In Figure 7, for instance, a double-stranded representation of the *chloroplast* genome shows the positive 5′ strand on the outside winding clockwise and the negative 3′ strand on the inside winding counter-clockwise. In both cases, all the currently known genes are represented.

Bioinformatics: The New Means for Representing the Human Genome

On February 16, 2001, nearly eight months after the historic announcement at the White House, a scientific paper with 274 co-authors, titled "The Sequence of the Human Genome," finally appeared in *Science.* A companion article representing the research findings of the public program had appeared the previous day in the British journal, *Nature.* Closely reasoned, generously footnoted, and profusely illustrated with tables and figures, the *Science* paper, with Craig Venter listed as lead author, was an immediate sensation. Citing 134 preceding articles, Venter's Celera *Science* article has itself been cited 5,056 times and, like Watson and Crick's article before it, has gone on to immortality in the annals of scientific literature.

Sequencing the genome in the lightning-fast time of approximately nine months, Venter's *whole genome shotgun assembly sequencing,* a method that essentially shredded and reassembled the DNA of five voluntary individuals using banks of powerful automatic sequencing devices, was an attempt at the then-inconceivable task of creating a high-resolution map that not only located gene-dense regions on human chromosomes, but also described gaps in the sequencing process.

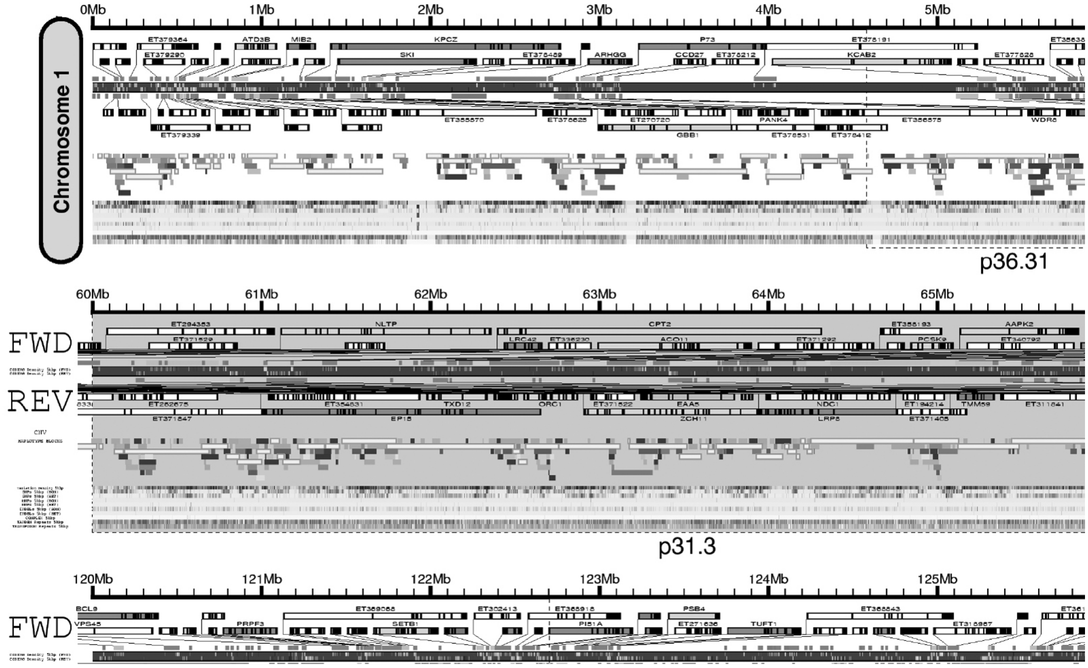


Figure 8
Micro-detail of Chromosome 1 on the Venter map, complete with nested, labeled known genes. Forward or + polarity phosphate is 5'-3'; reverse is 3'-5'. The rule at top is in million, or mega base pairs. (reproduced by permission, J. Craig Venter Institute, under a Creative Commons license from PLoS Biology)

Although the process of transcribing human genes and expressing proteins is fiendishly elaborate, researchers were surprised to find that it is quite a bit simpler than they had imagined. The number of human genes is now known to hover around 25,000, although scientists once thought they might find as many as 150,000; these genes also are governed by fewer mutations than once thought possible. These variations, known as single-nucleotide polymorphisms (SNPs), effect changes in gene expression among people. Code readers discovered that "less than 1% of SNPs affect protein function, resulting in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins."[10]

The ultimate purpose of the research, of course, is not merely to delineate a particular method. The location and enumeration of human genes is crucial to understanding how genes function and how the genome evolves. In addition, qualitative observations, including the fact that proteins with disease associations belong to duplicated segments of genetic code, have advanced our understanding of the role that genes play in human illness. The 2007 "Diploid Genome Sequence of J. Craig Venter" (see Figure 8) resulted in the visual representation of a high-resolution map for all of Venter's 46 chromosomes, in both the 5' and 3' directions. It utilized what are called *scaffolds,* or long segments of nucleotides clearly defined by location. These scaffolds were further defined by a process of extremely detailed, linear nested, information overlap, resulting in a remarkable compression of

[10] J. Craig Venter et al., "The Sequence of the Human Genome," *Science* 291 (2001): 1330.

visually represented data that was nearly able to visually approximate the coded representation of the submicroscopic amino acids themselves. This information exists in printed form, in a wall poster 40 inches wide by 60 inches high, and it also is available as a high-resolution 88MB PDF that allows for a 10x zoom ratio.

This map visually represents the entire three-billion-base genome, each chromosome in mega-base pairs with its introns and exons. The exons are the areas of the genome that contain genetic coding, and they are shown here with their nested pull-outs of currently known genes. Venter's sequence was hailed by *Design Observer* on September 10, 2007, as "…one of the most complex single infographics ever created"—a left-handed compliment at best, given that it represents the culmination of a century of genetic *informatics.*

Reading and effectively interpreting the highly specialized information in this map takes a trained eye, but even with a rudimentary understanding of the science, viewers can begin to understand it. It reads left to right and top to bottom in a linear order, reminiscent of Sturtevant's early physical maps. I've cross-checked it against NCBI's Map Viewer for Chromosome 1—a fascinating exercise in comprehending relational complexity.[11]

As the gaps in the genome have been filled, our knowledge of ourselves has grown exponentially. The partial sequence released in 2001 has become much more complete, as demonstrated by Venter's 2007 visualization, and new discoveries proceed apace. With this growth in complexity comes new challenges. Genetic mapping is so much more detailed than it was a century ago that Sturtevant (1891-1970), who lived until the dawn of molecular biology, would have been amazed. Once an area dominated by generalists, genomics has evolved into an area of high specialization. With the marriage of life sciences and computation, statisticians and computer scientists work side by side with cytologists and geneticists to carry on the work once done by undergrad lab assistants. The efforts of the big pharmaceutical companies to mine data in pursuit of the next generation of drugs continues; meanwhile, the constantly improving database is online and available for both researchers and informed amateurs to access and update.

As the study of genetics yields new data, as computational genomics provides tools for new medical and scientific diagnoses, and as new methods for treating age-old human maladies are further developed, such revelations of nature must be illustrated by ever deepening, more detailed, and subtle forms of visual information. Such *bioinformatics* are often pioneered by scientists seeking to represent complex ideas in visual form. Sometimes, as with Mendeleev's table, an idea sticks and is adopted by subsequent generations. If such visualizations are, in Edward Tufte's words,

[11] See http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?TAXID=9606&CHR=1&MAPS=ugHs,genes,genec[5988.93%3A8650.78]-r&QSTR=100359407[gene_id]&QUERY=uid%28-2141675409%29&CMD=UP (accessed March 15, 2012).

"…communicated with clarity, precision, and efficiency…," later generations of researchers will adopt, revise, and improve them. The history of the development of genomic graphics is rich—and it will continue to become richer. As visual creatures, we can learn effectively from images, graphs, and diagrams, no matter how complex the data set.

Acknowledgement
The author wishes to especially thank Dr. Eric Linton of Central Michigan University for his assistance with this essay.