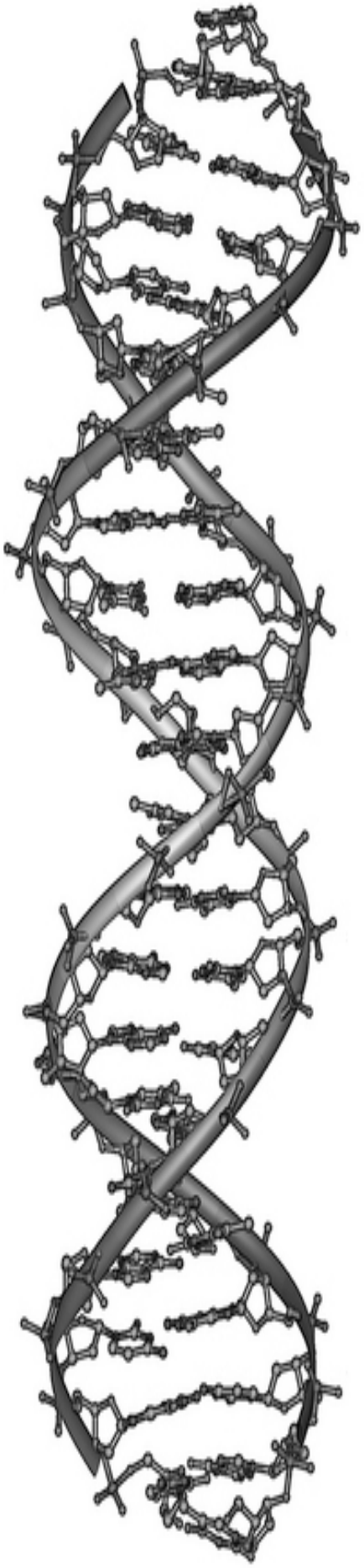


credit: Michael Ströck (Wikimedia Commons).

Our Genomic Future

A decade after the Human Genome Project

By Nolan Kamitaki

Imagine a world where parents selectively screen embryos to handpick the most viable children and companies discriminate potential employees using their DNA sequences. What sounds like science fiction (and served as a central point of the 1997 movie *Gattaca*) is becoming a reality, relating political and ethical concerns related to information privacy. Following the first published sequence of the human genome in 2001, one of the dominant trends in the last decade has been the emerging personalized medicine market. While much personalized medicine up to this point has consisted of measuring protein levels to determine an optimal course of action for the patient, the advent of faster and cheaper genome sequencing means that people may soon be able to easily acquire custom-designed medicines. With life expectancy predicted to increase by more than a decade by 2050 due to improvements in medicine, perhaps the only thing more exciting than our promising future is the tumultuous story of how we came to the first sequence of the human genome, a feat that was accomplished less than a decade ago (1).

First Steps of the Genome Project

The first complete genome to be sequenced was that of a virus. The genetic code of bacteriophage MS2 was read by Walter Fiers and his team from Belgium in 1976 (2). It was not long before others would improve upon their methods. In 1977, Frederick Sanger introduced shotgun sequencing. Nicknamed the "Sanger method," this Nobel Prize-winning technique involves breaking up long genomes into readable fragments, which are pieced back together based on overlapping ends (3). In order to match the increased complexity of sorting randomized shotgun pieces, the process of genome sequencing emerged into the field of bioinformatics, with computational automation becoming a necessity in most biological labs.

In the 1980s, it was first proposed that the human genome might be a feasible long term target for sequencing. Innumerable benefits would include not only intimate knowledge of our genetic and intergenic composition, but also improved pharmaceuticals and retrospective analysis of our ancestry. A conference held in Santa Fe in 1986 initiated the series of meetings of the Department of Energy (DOE) that would jumpstart the Human Genome Project (HGP) (4). Finally, in 1990, the National Institutes of Health and the DOE presented their plan to Congress, embarking on what would become a fifteen year-long scientific effort, arguably one of the largest and most intensive

of any scientific event in history (5).

Even though organisms like the yeast and *C. elegans* had had their genomes sequenced by the late 1990s, the sequencing of the human genome was a rather gargantuan step up (6, 7). While all three are eukaryotic organisms, the human genome is about 30 times larger than that of *C. elegans* and 250 times larger than that of yeast.

The method proposed by the HGP consortium was termed "hierarchical shotgun sequencing." DNA was divided into 150 megabase-long pieces before being inserted into bacterial artificial chromosomes (BACs) (8). After placing these BACs in *E. coli,* they could be reproduced and distributed to scientists around the world. Genome reassembly would then occur by shotgun-sequencing each fragment contained in a BAC and then aligning them together to generate the original genome. This method would not only allow for distribution of sequencing duties to separate research institutions, but also for a higher level of quality than previous methods. At the time, few would have guessed that from this ambitious beginning would emerge a scientific rivalry, the likes of which had not been seen since the space race between the United States and the Soviet Union.

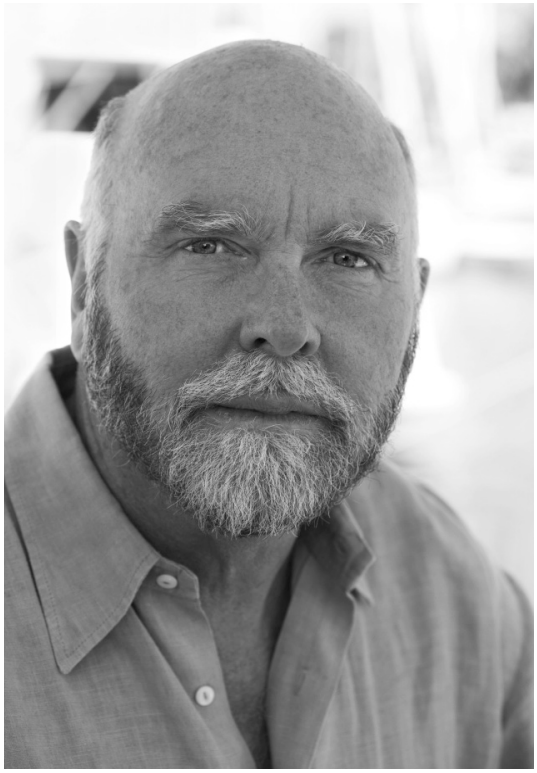
credit: NIH, http://www.genome.gov/

▲ **Figure 1.** Researcher working with a DNA sequencer.

The Private Effort: Craig Venter

Recounting the long background of sequencing the human genome without mentioning the private industry's foray would be telling only half the story. Craig Venter, a name now commonly associated with genetics, began his career as a medic in the Vietnam War. There, he experienced the fine line separating life and death and gained a newfound appreciation for the human health. Returning to the United States for college, he soon became involved in medical research, eventually rising through the ranks to become an mRNA researcher at NIH. During his time there, he first touched upon the ethical controversy of gene patenting, a topic that would follow him throughout his research in genetics.

When Venter heard of the HGP, he sent a proposal in 1991 to James Watson (of Watson-Crick DNA fame) about what he predicted would be a more efficient way to sequence the human genome: sequencing only the parts of the genome which code for genes. Expressed Sequence Tags (ESTs), the tools Venter proposed to Watson, are created by taking all the mRNA in a given cell and making cDNA copies of each (8). These cDNAs are sequenced at both ends to produce "tags" and are then used as markers to find the gene from which they transcribed. Using ESTs to identify and map genes to the genome was very cost-effective and quick, allowing for the discovery of many more genes than traditional methods would permit. This was viewed by much of the scientific community as a rushed, haphazard way of going about the problem because it would ignore most of the genome but "discover" numerous genes for which no function could be identified. It also struck a sour note with many researchers because it could be used to gobble up gene patents, locking them away in a process many regard as ethically comparable to patenting a law of nature. When Watson rejected the approach, Venter departed from NIH and entered the biotech industry, financed by venture capitalist Wallace Steinberg (9).



credit: article by Liza Gross (Wikimedia Commons)

▲ **Figure 2.** Craig Venter

credit: NIH, http://www.genome.gov/

▲ **Figure 3.** Artistic illustration of personalized medicine.

The approach favored by Craig Venter differed largely from the public side of the HGP. Where the public effort sought careful, high quality sequencing through painstaking detail, Venter was more interested in efficiency. Relying more on the computational power of algorithms and improving processing speed to reassemble broken genome bits like puzzle pieces, Venter's whole genome shotgun sequencing method was regarded by many as a crude and less accurate rendition of sequencing procedures. Furthermore, this private method relied on using data released by the public HGP as a scaffold upon which new data could be arranged, inciting another source of criticism.

By establishing The Institute for Genomic Research (TIGR) and Human Genome Sciences (HGSI), Craig Venter set out to sequence several genomes before approaching the sequence within his own cells. TIGR proved its worth in 1995 by sequencing the *H. influenzae* bacteria, the first organism to be sequenced (10). Venter formed a new company in 1998, claiming that he could sequence the human genome within three years. Named "Celera" after the Latin word for speed, the team used the whole genome shotgun sequencing method to tackle the fruit fly genome, finishing in three months (8). With 2001 set as the target for completion of the human genome by Celera, the HGP changed their goals to have a "working draft" also completed by that time.

Finishing the First Draft

This race to be the first to publish the initial draft of the human genome gained fuel through public interest. As the world listened with bated breath, President Bill Clinton announced on June 25, 2000, that the first survey of the human genome had been completed through efforts from both public and private ventures (11). In reality, Celera actually beat the public effort in terms of speed, informally announcing completion on April 6, 2000 (9). After a series of meetings, Venter and his team agreed to wait until June for an official proclamation and to publish their findings at the same time as the public effort. Indeed, the two versions emerged as side-by-side journal articles in special issues of *Science* and *Nature* on February 15 and 16, 2001 (8). It took the President and public pressure to broker a truce between the competing groups.

The initial working drafts of the human genome from both the public and the private publications were far from complete. Besides not covering approximately 10% of the genome, these first efforts contained somewhere in the neighborhood of ~250,000 gaps in the sequence (12, 13). Although the genome published in 2001 was an impressive display of technological advances made in both the public and private sector, it was far from finished. The HGP consortium proceeded to sequence a much higher quality genome, with the advantage of using bacterial artificial chromosomes becoming apparent. Certain regions of the genome containing recently duplicated segments were hard to reconstruct using Celera's whole-genome shotgun method. By 2004, a draft of the human genome containing more than 99% of the genome and only 341 gaps was published (14). While 2001 will likely always be remembered as the date of original publication, we have to keep in mind that, for such an enormous project, increasing the level of accuracy is still a work in progress today. With improving sequence accuracy, variation between any two cohorts can be studied in much more detail, potentially leading to future medical discoveries.

As Eric Lander, a leader of the Human Genome Project and co-chair of Obama's Council of Advisors on Science and Technology, wrote in a recent review of the first working draft, "the per-base cost of DNA sequencing has plummeted by ~100,000-fold over the past decade, far outpacing Moore's law of technological advance in the semiconductor industry" (15). Today, a possible end goal of a hundred dollar genome doesn't sound too far off and

may be approachable by the end of this decade. From a purely scientific standpoint, this will allow researchers to not only find much rarer variants scattered across the genome, but to also explain the heritability of patterns in common variants. From a medical perspective, this reduction in cost will allow genome sequencing to become a regular tool used in diagnostics to screen for diseases and drug efficacy. The last decade of research was spent improving the first working draft; hopefully, the years to come will usher in a new era of increased understanding of genetics and medical genomics.

"The per-base cost of DNA sequencing has plummeted by ~100,000-fold over the past decade, far outpacing Moore's law of technological advance in the semiconductor industry."

Sequencing of the human genome is certainly one of the greatest scientific and technological feats of our generation, but fully comprehending what it means and how we can reap its benefits will take much more time than it took to generate the sequence itself. Companies that proclaimed a genetic sequence as a complete solution to illness misunderstood the fact that many more studies will have to be done before reliable associations can be drawn between genetic test results and medical conclusions. Perhaps the fantasized era of personalized drug synthesis will be realized in the next decade.

Applications and the Future of Medical Genomics

Doctors can still make personalized risk assessments based purely off the portion of heritability researchers have already identified in genetic variation. In other words, the correlations we can draw between genetic mutations and physical effects are useful, despite our lack of understanding of exactly how they work. This is the basis of a type of genetic epidemiology study known as a genome-wide association study (GWAS) (16): single nucleotide polymorphisms (SNPs), or point mutations, are compared between a sample population with or without a given disease to see if any SNPs are associated with a particular disease group.

To more readily understand the distribution of this variation, researchers formed the International HapMap (Haplotype Map) Project, an effort to exploit the fact that nearby SNPs are often highly correlated, as adjacent pieces of DNA will undergo less recombination than pieces farther apart (17). A small set of SNPs can be identified as predictors for disease if found to be correlated with physiological differences through a GWAS. Despite the relative efficiency of these sorts of naïve epidemiological discoveries, many more patients will have to be sequenced before rare mutations can be identified and linked to a specific illness.

Lastly, in an area of perhaps more anthropological interest, genome sequencing from individuals around the globe has given mankind a peek into its own past over the last several thousand years. For instance, genetic evidence suggests that the original group of humans to migrate out of Africa went through several population mixing events, including with Neanderthals. Indeed, both Europeans and Asians display signs of inheriting 1 to 4% of the Neanderthal genome, which African populations lack (18). On a smaller scale, people who order personal genetic testing kits can now trace their lineage back within much more recent generations to determine ancestral origins. Besides revealing powerful evidence for population migration patterns, genome sequencing can also reveal areas currently undergoing positive evolutionary selection.

As we celebrate advances made by countless organizations towards first sequencing and then understanding our own genetic code, developments since 2001 have provided a simultaneously sobering and exciting reality. While sequencing the genome did not lead to many immediate rewards outside of an improved purely biological understanding, it opened the door to a world of discoveries waiting to be unearthed. Indeed, we now have much more genetic information than manpower or computational ability to completely process it. The true benefit of the Human Genome Project will be how it revolutionizes human health; the number of associations with diseases and novel gene pathways researchers have identified in the last few years would suggest an abundance of medical advances lying just beyond today's horizon. H

—Nolan Kamitaki '14 is a prospective Molecular and Cellular Biology concentrator in Mower Hall.

References

1. Cox, L. (2009) We Will Live Longer in 2050, Study Predicts. Available from: http://abcnews.go.com/Health/ActiveAging/humans-live-longer-2050-scientists-predict/story?id=9330511.
2. Fiers, W., et al. (1976) Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene. Nature. 260(5551): 500-507.
3. Sanger, F., et al. (1978) Nucleotide Sequence of Bacteriophage [phi]X174. Journal of Molecular Biology. 125(2): 225-246.
4. DeLisi, C. (2008) Meetings that changed the world: Santa Fe 1986: Human genome baby-steps. Nature. 455: 876-877.
5. Fink, L. (1990) Five-Year Plan Goes to Capitol Hill. Human Genome Program, U.S. Department of Energy. Human Genome News. 2.
6. Goffeau, A., et al. (1996) Life with 6000 Genes. Science. 274(5287): 546-567.
7. (1998) The C. elegans Sequencing Consortium, Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 282(5396): 2012-2018.
8. Programs, U.S.D.o.E.G. genomics.energy.gov. Available from: http://genomics.energy.gov/.
9. Venter, J.C., A Life Decoded: My Genome: My Life. 2008: Penguin.
10. Fleischmann, R.D., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 269(5223): 496-512.
11. House, T.W. (2000) President Clinton Announces the Completion of the First Survey of the Entire Human Genome: Hails Public and Private Efforts Leading to This Historic Achievement, O.o.t.P. Secretary, Editor.
12. The International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. Nature, 2001. 409: p. 860-921.
13. Venter, J.C., et al, The Sequence of the Human Genome. Science, 2001. 291(5507): p. 1304-1351.
14. (2004) The International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature. 431: 931-945.
15. Lander, E.S. (2011) Initial impact of the sequencing of the human genome. Nature. 470: 187-197.
16. Lander, E.S. (1996) The New Genomics: Global Views on Biology. Science. 274(5287): 536-539.
17. (2005) The International HapMap Consortium, A haplotype map of the human genome. Nature. 437: 1299-1320.
18. Green, R.E., et al. (2010) A Draft Sequence of the Neandertal Genome. Science. 328(5979): 710-722.