Human genome and CRISPR

Andrew Valikov

1 Human genome

1.1 Brief history

The Human Genome Project was a 15-year-long, publicly funded project initiated in 1990 with the objective of determining the DNA sequence of the entire euchromatic human genome within 15 years. In May 1985, Robert Sinsheimer organized a workshop to discuss sequencing the human genome, but for a number of reasons the NIH was uninterested in pursuing the proposal. The following March, the Santa Fe Workshop was organized by Charles DeLisi and David Smith of the Department of Energy's Office of Health and Environmental Research (OHER). At the same time Renato Dulbecco proposed whole genome sequencing in an essay in Science. James Watson followed two months later with a workshop held at the Cold Spring Harbor Laboratory.

The fact that the Santa Fe workshop was motivated and supported by a Federal Agency opened a path, albeit a difficult and tortuous one, for converting the idea into a public policy in the United States. In a memo to the Assistant Secretary for Energy Research (Alvin Trivelpiece), Charles DeLisi, who was then Director of the OHER, outlined a broad plan for the project. This started a long and complex chain of events which led to approved reprogramming of funds that enabled the OHER to launch the Project in 1986, and to recommend the first line item for the HGP, which was in President Reagan's 1988 budget submission, and ultimately approved by the Congress. Of particular importance in Congressional approval was the advocacy of Senator Peter Domenici, whom DeLisi had befriended. Domenici chaired the Senate Committee on Energy and Natural Resources, as well as the Budget Committee, both of which were key in the DOE budget process. Congress added a comparable amount to the NIH budget, thereby beginning official funding by both agencies.

Alvin Trivelpiece sought and obtained the approval of DeLisi's proposal by Deputy Secretary William Flynn Martin. This chart was used in the spring of 1986 by Trivelpiece, then Director of the Office of Energy Research in the Department of

Energy, to brief Martin and Under Secretary Joseph Salgado regarding his intention to reprogram $4 million to initiate the project with the approval of Secretary Herrington. This reprogramming was followed by a line item budget of $16 million in the Reagan Administration's 1987 budget submission to Congress. It subsequently passed both Houses. The Project was planned for 15 years.

Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.

In 1990, the two major funding agencies, DOE and NIH, developed a memorandum of understanding in order to coordinate plans and set the clock for the initiation of the Project to 1990. At that time, David Galas was Director of the renamed "Office of Biological and Environmental Research" in the U.S. Department of Energy's Office of Science and James Watson headed the NIH Genome Program. In 1993, Aristides Patrinos succeeded Galas and Francis Collins succeeded James Watson, assuming the role of overall Project Head as Director of the U.S. National Institutes of Health (NIH) National Center for Human Genome Research (which would later become the National Human Genome Research Institute). A working draft of the genome was announced in 2000 and the papers describing it were published in February 2001. A more complete draft was published in 2003, and genome "finishing"work continued for more than a decade.

The $3-billion project was formally founded in 1990 by the US Department of Energy and the National Institutes of Health, and was expected to take 15 years. In addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Australia, China and myriad other spontaneous relationships.

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by U.S. President Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000). This first available rough draft assembly of the genome was completed by the Genome Bioinformatics Group at the University of California, Santa Cruz, primarily led by then graduate student Jim Kent. Ongoing sequencing led to the announcement of the essentially complete genome on April 14, 2003, two years earlier than planned. In May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in Nature.

1.2 State of completion

The project was not able to sequence all the DNA found in human cells. It sequenced only "euchromatic"regions of the genome, which make up 92% of the human genome. The other regions, called "heterochromatic"are found in centromeres and telomeres, and were not sequenced under the project.

The Human Genome Project was declared complete in April 2003. An initial rough draft of the human genome was available in June 2000 and by February 2001 a working draft had been completed and published followed by the final sequencing mapping of the human genome on April 14, 2003. Although this was reported to cover 99% of the euchromatic human genome with 99.99% accuracy, a major quality assessment of the human genome sequence was published on May 27, 2004 indicating over 92% of sampling exceeded 99.99% accuracy which was within the intended goal. Further analyses and papers on the HGP continue to occur.

1.3 Techniques

The process of identifying the boundaries between genes and other features in a raw DNA sequence is called genome annotation and is in the domain of bioinformatics. While expert biologists make the best annotators, their work proceeds slowly, and computer programs are increasingly used to meet the high-throughput demands of genome sequencing projects. Beginning in 2008, a new technology known as RNA-seq was introduced that allowed scientists to directly sequence the messenger RNA in cells. This replaced previous methods of annotation, which relied on inherent properties of the DNA sequence, with direct measurement, which was much more accurate. Today, annotation of the human genome and other genomes relies primarily on deep sequencing of the transcripts in every human tissue using RNA-seq. These experiments have revealed that over 90% of genes contain at least one and usually several alternative splice variants, in which the exons are combined in different ways to produce 2 or more gene products from the same locus.

The genome published by the HGP does not represent the sequence of every individual's genome. It is the combined mosaic of a small number of anonymous donors, all of European origin. The HGP genome is a scaffold for future work in identifying differences among individuals. Subsequent projects sequenced the genomes of multiple distinct ethnic groups, though as of today there is still only one "reference genome."

1.4 Findings

Key findings of the draft (2001) and complete (2004) genome sequences include:

There are approximately 22,300 protein-coding genes in human beings, the same range as in other mammals. The human genome has significantly more segmental duplications (nearly identical, repeated sections of DNA) than had been previously suspected. At the time when the draft sequence was published fewer than 7% of protein appeared to be vertebrate specific.

1.5 Accomplishment

The first printout of the human genome to be presented as a series of books, at the Wellcome Collection, London. The Human Genome Project was started in 1990 with the goal of sequencing and identifying all three billion chemical units in the human genetic instruction set, finding the genetic roots of disease and then developing treatments. It is considered a megaproject because the human genome has approximately 3.3 billion base-pairs. With the sequence in hand, the next step was to identify the genetic variants that increase the risk for common diseases like cancer and diabetes.

It was far too expensive at that time to think of sequencing patients' whole genomes. So the National Institutes of Health embraced the idea for a "shortcut which was to look just at sites on the genome where many people have a variant DNA unit. The theory behind the shortcut was that, since the major diseases are common, so too would be the genetic variants that caused them. Natural selection keeps the human genome free of variants that damage health before children are grown, the theory held, but fails against variants that strike later in life, allowing them to become quite common. (In 2002 the National Institutes of Health started a $138 million project called the HapMap to catalog the common variants in European, East Asian and African genomes.)

The genome was broken into smaller pieces; approximately 150,000 base pairs in length. These pieces were then ligated into a type of vector known as "bacterial artificial chromosomes or BACs, which are derived from bacterial chromosomes which have been genetically engineered. The vectors containing the genes can be inserted into bacteria where they are copied by the bacterial DNA replication machinery. Each of these pieces was then sequenced separately as a small "shotgun"project and then assembled. The larger, 150,000 base pairs go together to create chromosomes. This is known as the "hierarchical shotgun"approach, because the genome is first broken into relatively large chunks, which are then mapped to chromosomes before being selected for sequencing.

Funding came from the US government through the National Institutes of Health

in the United States, and a UK charity organization, the Wellcome Trust, as well as numerous other groups from around the world. The funding supported a number of large sequencing centers including those at Whitehead Institute, the Wellcome Sanger Institute (then called The Sanger Centre) based at the Wellcome Genome Campus, Washington University in St. Louis, and Baylor College of Medicine.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) served as an important channel for the involvement of developing countries in the Human Genome Project.

1.6 Applications

The sequencing of the human genome holds benefits for many fields, from molecular medicine to human evolution. The Human Genome Project, through its sequencing of the DNA, can help us understand diseases including: genotyping of specific viruses to direct appropriate treatment; identification of mutations linked to different forms of cancer; the design of medication and more accurate prediction of their effects; advancement in forensic applied sciences; biofuels and other energy applications; agriculture, animal husbandry, bioprocessing; risk assessment; bioarcheology, anthropology and evolution. Another proposed benefit is the commercial development of genomics research related to DNA based products, a multibillion-dollar industry.

The sequence of the DNA is stored in databases available to anyone on the Internet. The U.S. National Center for Biotechnology Information (and sister organizations in Europe and Japan) house the gene sequence in a database known as GenBank, along with sequences of known and hypothetical genes and proteins. Other organizations, such as the UCSC Genome Browser at the University of California, Santa Cruz, and Ensembl present additional data and annotation and powerful tools for visualizing and searching it. Computer programs have been developed to analyze the data, because the data itself is difficult to interpret without such programs. Generally speaking, advances in genome sequencing technology have followed Moore's Law, a concept from computer science which states that integrated circuits can increase in complexity at an exponential rate. This means that the speeds at which whole genomes can be sequenced can increase at a similar rate, as was seen during the development of the above-mentioned Human Genome Project.

1.7 Interpretation

The work on interpretation and analysis of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues

for advances in medicine and biotechnology. Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as Myriad Genetics, started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including breast cancer, hemostasis disorders, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biologists. For example, a researcher investigating a certain form of cancer may have narrowed down their search to a particular gene. By visiting the human genome database on the World Wide Web, this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, and diseases associated with this gene or other datatypes. Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of cellular processes, it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of evolution. In many cases, evolutionary questions can now be framed in terms of molecular biology; indeed, many major evolutionary milestones (the emergence of the ribosome and organelles, the development of embryos with body plans, the vertebrate immune system) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest relatives (the primates, and indeed the other mammals) are expected to be illuminated by the data in this project.

The project inspired and paved the way for genomic work in other fields, such as agriculture. For example, by studying the genetic composition of Tritium aestivum, the world's most commonly used bread wheat, great insight has been gained into the ways that domestication has impacted the evolution of the plant. Which loci are most susceptible to manipulation, and how does this play out in evolutionary terms? Genetic sequencing has allowed these questions to be addressed for the first time, as specific loci can be compared in wild and domesticated strains of the plant. This will allow for advances in genetic modification in the future which could yield healthier,

more disease-resistant wheat crops.

2 CRISPR Cas9

2.1 General information

CRISPR is a family of DNA sequences in bacteria. The sequences contain snippets of DNA from viruses that have attacked the bacterium. These snippets are used by the bacterium to detect and destroy DNA from similar viruses during subsequent attacks. These sequences play a key role in a bacterial defense system, and form the basis of a technology known as CRISPR/Cas9 that effectively and specifically changes genes within organisms.

The CRISPR/Cas system is a prokaryotic immune system that confers resistance to foreign genetic elements such as those present within plasmids and phages that provides a form of acquired immunity. RNA harboring the spacer sequence helps Cas (CRISPR-associated) proteins recognize and cut exogenous DNA. Other RNA-guided Cas proteins cut foreign RNA. CRISPRs are found in approximately 40% of sequenced bacterial genomes and 90% of sequenced archaea.

CRISPR is an abbreviation of Clustered Regularly Interspaced Short Palindromic Repeats. The name was minted at a time when the origin and use of the interspacing subsequences were not known. At that time the CRISPRs were described as segments of prokaryotic DNA containing short, repetitive base sequences. In a palindromic repeat, the sequence of nucleotides is the same in both directions. Each repetition is followed by short segments of spacer DNA from previous exposures to foreign DNA (e.g., a virus or plasmid). Small clusters of cas (CRISPR-associated system) genes are located next to CRISPR sequences.

A simple version of the CRISPR/Cas system, CRISPR/Cas9, has been modified to edit genomes. By delivering the Cas9 nuclease complexed with a synthetic guide RNA (gRNA) into a cell, the cell's genome can be cut at a desired location, allowing existing genes to be removed and/or new ones added.

CRISPR/Cas genome editing techniques have many potential applications, including medicine and crop seed enhancement. The use of CRISPR/Cas9-gRNA complex for genome editing was the AAAS's choice for breakthrough of the year in 2015. Bioethical concerns have been raised about the prospect of using CRISPR for germline editing.

2.2 Embryos gene-editing

Scientists reported selectively altering genes in viable human embryos for the first time this year. For nearly five years, researchers have been wielding the molecular scissors known as CRISPR/Cas9 to make precise changes in animals' DNA. But its use in human embryos has more profound implications, researchers and ethicists say. «We can now literally change our own species», says Mildred Solomon, a bioethicist and president of the Hastings Center, a bioethics research institute in Garrison, N.Y.

CRISPR/Cas9 is a bacterial immune system turned into a powerful gene-editing tool. First described in 2012, the editor consists of a DNA-cutting enzyme called Cas9 and a short piece of RNA that guides the enzyme to a specific spot that scientists want to edit. Once the editing machinery reaches its destination, Cas9 cleaves the DNA. Cells can repair the break by gluing the cut ends back together, or by pasting in another piece of DNA. Scientists have developed variations of the editor that make other changes to DNA without cutting, including one version described in October that performs a previously impossible conversion of one DNA base into another.

Whether scientists should use CRISPR/Cas9's power to create gene-edited babies is a matter of heated debate. Until March, the battles were mostly academic because previous attempts to edit human embryos were done in embryos that would never develop into a baby. But in March, Lichun Tang of China's Beijing Proteome Research Center and colleagues reported using CRISPR/Cas9 to correct disease-causing mutations in a small number of viable human embryos. Other groups posted separate reports of CRISPR/Cas9 repair in viable human embryos in August and October.

A 5-day-old human embryo is usually composed of about 200 cells in a hollow ball configuration called a blastocyst. Embryos edited to remove the OCT4 gene fail to make normal blastocysts. Rather than edit out problems, a different group led by developmental biologist Kathy Niakan of the Francis Crick Institute in London snipped DNA with CRISPR/Cas9 to intentionally create mutations in human embryos for the first time. Those embryos were used to study the role of a gene important in development. Together, the studies illustrate that the gene-editing technology can make a variety of changes in human DNA that would last a lifetime and stretch across generations. It's the relative ease and permanence that have many people worried that CRISPR/Cas9 could lead to new classes of genetically enhanced people and discrimination against others born with uncorrected genetic diseases. Taken to extremes, that discrimination could extend to people whose parents chose not to (or didn't have the means to) genetically alter their children's athletic power, intellectual ability or other characteristics.

In February, a panel of ethicists and other experts convened by the U.S. National Academies of Sciences, Engineering and Medicine warned against using CRISPR to

enhance health or other traits. But the panel said using human gene editing to correct diseases, in certain circumstances, could be permissible.

No babies have been born with changes made by CRISPR/Cas9 or any other gene-editing technology. But it could be only a matter of time. «I would not be surprised if there were a CRISPR-modified baby somewhere in the world in the next couple of years,» said CRISPR pioneer Jennifer Doudna of the University of California, Berkeley on October 26 in San Francisco at the World Conference of Science Journalists. Doudna said she does not support using CRISPR/Cas9 to make gene-edited babies.

2.3 Genetic rewrite

Base editors fuse enzymes that can alter the chemical structure of DNA bases to a "dead"version of CRISPR/Cas9 that doesn't cut DNA. The first base editors developed in 2016 use cytidine deaminase to change C-G base pairs to T-A pairs. A new base editor uses DNA adenine deaminase to make the opposite change, transforming A-T base pairs to G-C pairs.

The study soon came under criticism from other researchers. "The evidence for fixing is not there,"says Dieter Egli, a developmental biologist at Columbia University. Mitalipov and colleagues have not presented enough data to support their interpretation, he says. Egli and colleagues posted their criticism online August 28 at bioRxiv.org. "The conclusion that the correction occurred is, at best, premature,"Egli says. "At worst, it might be false."

Mitalipov and Oregon colleague Paula Amato say they have submitted more evidence to support their claim to Nature and hope to publish the data soon. The finding needs to be replicated by other groups, Amato says, but "at the moment, we stand by our conclusions."

It may also be possible to fix genetic mutations without any cutting. By dulling Cas9's blades, researchers led by David Liu of Harvard University have developed "base editors". The enzyme can grasp DNA but not slice through it. The researchers attached other enzymes that chemically change one DNA base into another. DNA bases are the information-carrying part of the DNA molecule and are often represented by A, C, T and G. In 2016, Liu described a base editor that transforms a C into a T.

But before any type of human embryo editing can be used in the clinic, it must be as safe and effective as existing embryo screening methods. Today, doctors working with embryos created through in vitro fertilization can extract a few cells for genetic testing in a process called preimplantation genetic diagnosis, or PGD. Embryos that

don't have mutations can be transferred to a woman's uterus to establish a pregnancy. In some rare cases, couples may not produce any healthy embryos. Future gene editing might help these couples have a healthy biological child. For other couples, gene editing could increase the number of healthy embryos available by fixing some that would otherwise be thrown away, Amato says. "If it's shown to be just as safe and effective as PGD, I'd say, 'Why not use it?'"