Quantitative Biology 2021, 9(4): 361–365
https://doi.org/10.15302/J-QB-021-0277

FEATURE

Mapping genetic variations in the first assembled human genome

Jinghui Zhang*

Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105-3678, USA
* Correspondence: jinghui.zhang@stjude.org

Received August 21, 2021; Revised September 4, 2021; Accepted September 6, 2021

On June 26, 2000, President Bill Clinton held a ceremony at the White House announcing the completion of the human genome sequencing. Finally, the leaders of the two competing groups, Dr. Francis Collins from the National Institute of Health, and Dr. J. Craig Venter from Celera Genomics, shook their hands. As a scientist who had been actively involved in both the public and private sides of human genome sequencing, I can still feel the excitement of being a contributor of this landmark scientific project even after 20 years. Nowadays, I can also attest the significance of this achievement as the reference human genome sequence is being used every day by my research team to find cures for pediatric cancers. In this feature I would like to present a foot soldier's account on assembling and analyzing the sequence of the human genome.

FROM WET BENCH TO THE GENOMES AND GENETIC VARIATIONS

My passion for biological research was sparked in elementary school after I finished reading "A Path to Conquer Pathogens", a 1978 book published by China Children's Press & Publication Group. I was fascinated by the stories such as invention of vaccine against smallpox by Edward Jenner and discovery of penicillin by Alexander Fleming. Working in a laboratory to find cures for diseases would become the dream of my life. Years later, I was saddened to realize that this dream was fading away because I did not possess the "golden fingers" required to perform the intricate wet-lab experiments in college and graduate school (Fudan University in Shanghai and the University of Connecticut, respectively). While suffering through the disappointing results of my benchwork, I excelled at computer science (CS) classes, beating my classmates in both coding and data analysis projects. Planning my escape from the wet bench, I showed up at the door of Dr. Jim Ostell, Chief of the Software Engineering Branch at National Center for Biotechnology Information (NCBI), imploring him to be my Ph.D. advisor. After some pondering, Jim agreed to take me in. Thus in 1992, I dried my hands and became the first graduate student at NCBI with a goal of analyzing sequence data at genome scale, an unexplored area of NCBI at the time.

Escherichia coli, a bacterium used as the preferred experimental model for biological research, was the first organism I worked on. In 1992, the only complete genomes were bacterial phages in the range of tens or hundreds of kilobases, and the 4.6-megabase (Mb) genome of *E. coli* was considered the first mountain to climb in the pursuit of sequencing the full genome of an organism that could live on its own. Using a collection of gene-based *E. coli* sequences generated by individual research labs around the world, I started to build a genome sequence map comprised of gaps and sequenced regions. By performing computational prediction of restriction enzyme digestion sites on these sequences, I could order the sequences to the *E. coli* genome by aligning them to a restriction enzyme-based physical map generated by Yuji Kohara [1]. This work, coupled with a genome browser for dynamic exploration of this

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

incomplete *E. coli* genome, provided sufficient credentials for me to earn my Ph.D. in 1994 [2].

As DNA is the genetic code for all species, I stayed on at NCBI as a postdoctoral fellow focusing on more advanced species such as human. This was a timely transition — major sequencing centers for the Human Genome Project (HGP) had begun direct data deposition to NCBI as they were committed to the Bermuda Principles: *i.e.*, making human sequencing data available within 24 hours of data generation to encourage research and development and maximize the HGP's impact. I became a regular at the annual Human Genome Sequencing Meeting at the Cold Spring Harbor Laboratory (Fig. 1), geeking out on sequence analysis during the day and square dancing (led by David Cox from Stanford) at night.

Figure 1. With Stephanie Chissoe, a postdoctoral fellow at Washington University, at the Cold Spring Harbor Genome Sequence Meeting in 1995. Dr. Francis Collins (middle) passed by in front of us.

Genome annotation was a prerequisite for effective use of HGP sequences, and the researchers were split into the two camps of evidence-based versus model-based approaches. David Lipman, our Director, was a firm believer in the evidence-based approach. Partial cDNA sequences, in the form of expressed sequence tag (EST) [2], became a rich resource for performing evidence-based gene annotation. Working with K. M. Chao, a postdoctoral fellow at Webb Miller's lab at Pennsylvania State University, we developed several algorithms for aligning cDNA to genomic sequence [3,4]. These algorithms, when combined with the network BLAST API in a new software PowerBlast [5] that I wrote, enabled rapid annotation of raw human sequences as the latest release of the EST data was made accessible via the Internet. As PowerBLAST became popular, I was introduced to gene hunters such as Richard Gibbs (the Director of Human Genome Sequencing Center at Baylor College of Medicine) and Jeff Trent (then the National Human Genome Research Institute's Intramural Scientific Director) from whom I learned how to use genetic variations to discover genes involved in Mendelian diseases and complex diseases such as diabetes.

I was instantaneously awestruck by the world of genetic variations — in this world, I could work with computers instead of Petri dishes to search for disease cures, rekindling my childhood dream. More excitingly, I could already discern genetic variations, in the form of mismatches and small indels in the alignment of EST data to the HGP sequences, from the output of PowerBlast. After hearing a speech by Dr. Allen Roses where he advocated using genetic variations to make "right medicine for the right patients", I heard and answered the call, joining his newly formed Genetics Directive at Glaxo Wellcome (later GlaxoSmithKline) in 1998.

At Glaxo we focused on implementing pharmacogenomics in drug development, a novel approach pioneered by Allen who had discovered association of the APOE4 variant associated with Alzheimer's disease in 1993 [6]. DNA variations were discovered by analyzing sequencing data generated from 8 volunteers, and genotyping assays were developed to profile patients and controls for statistical assessment of disease-related association. While in the throes of completing DNA variation maps for selected loci relevant for Alzheimer, diabetes, and adverse drug responses, I was contacted by Celera Genomics, a biotech company headed by Dr. J. Craig Venter who initiated a privately funded effort to sequence the human genome in 2000. I found it impossible to resist the opportunity to develop the first genome-wide map of single-nucleotide polymorphisms (SNPs) using the shotgun sequencing data generated from multi-ethnic donors for the Celera genome project. So I moved back to the Washington DC metro area — but this time, I would be on the opposite side of my former NIH colleagues as a private-public race towards completing the human genome sequence began to unfold.

SEQUENCING THE HUMAN GENOME: SHOTGUN VERSUS CLONE-BY-CLONE

The determination of which sequencing approaches were best suited for assembling the human genome was at the core of the private/public debate and race. The clone-by-clone approach used by the public HGP was comprised of two steps involving generating and mapping clones containing 100 – 200 kb DNA fragment of the human genome, followed by separate shotgun sequencing of selected clones. By contrast, the whole-

genome shotgun approach used by Celera would generate libraries of different insert sizes (2 kb, 10 kb, and 50 kb) for shotgun sequencing, and the resulting fragments would be assembled computationally. The concept for human whole-genome shotgun sequencing (WGSS) was first proposed by Jim Weber and Eugene Myers in 1997 [7], where they demonstrated its feasibility through computational simulation. However, the academic research community including HGP did not view this approach favorably due to concerns over the quality of resulting reference sequence map as well as the computational resources required to perform assembly. Jim Weber visited NCBI in 1997, and I remember vividly how few were convinced by his bold vision.

When Celera announced that it would use WGSS to tackle the entire human genome, Eugene took the opportunity to turn his simulation to reality and became the VP responsible for developing the assembler. True to his academic roots, Eugene insisted on working in a cubicle along with his team. Amongst them was Granger Sutton who had developed the assembler to put together the WGSS reads for the 1.6 Mb genome of *Haemophilus influenza* in 1995 [8], making *H. flu* the first complete whole genome of a free-living organism (much to the chagrin of those of us who had toiled over the genome of *E. coli*). To provide the computing resources required for assembly of the 27 million human WGSS reads, a server farm comprised of 440 Compaq Computer Corporation Alpha CPUs was constructed, featuring a control room reminiscent of Star Trek.

DEVELOPING A GENOME-WIDE SNP MAP

The DNA library of the public HGP was constructed from individuals from Buffalo, New York. By contrast, the DNA pool used for Celera human genome sequencing was deliberately multi-ethnic including one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians. Therefore, the shotgun reads generated from this ethnically-diverse DNA library would be a fertile ground for discovery of novel genetic polymorphisms. In 2000, high-quality SNPs in the public domain were largely contributed by The SNP Consortium (TSC), an international collaboration of academic centers, pharmaceutical companies including Glaxo and a private foundation, with a goal of discovering and releasing at least 300,000 human SNPs. At the time Celera's shotgun sequencing was completed, TSC had identified 148,459 SNPs [9].

While annotation and assembly at Celera were tackled by large teams, SNP discovery was by contrast a miniscule 2-person operation, comprised of myself and Andy Clark, a renowned population geneticist working as a part-time consultant. Not all sequence variants were genetic polymorphisms, in fact errors from sequencing and assembly could account for the vast majority of variations. I had pored over many trace chromatograms at Glaxo to study the discrepancy between *in silico* detection and experimental validation of candidate SNPs, and recognized that, unlike the conventional wisdom, sequencing errors were not randomly distributed, and this could not be all laid at the feet of read quality scores quality. Recognizing these patterns, I was able to develop computational filters to identify and remove sequencing errors. Assembly error would cause paralogous variants (*i.e.,* variation between different regions of the same genome) to be misclassified as genetic variations, a major issue for highly repetitive regions. This problem could not be resolved by simply setting a threshold based on variant density, as my Glaxo experience made me keenly aware that the human genome contained highly polymorphic regions with variant density as high as paralogous variations. It should also be noted that mis-assembly of paralogous regions was an inherent problem in the upstream assembly process: this issue fundamentally affected variant detection, but exerted far less impact on gene annotation.

I struggled with this problem, feeling as if I had been fumbling in a dark alley for days, growing increasingly desperate. One day, while driving in a thunderstorm, I suddenly realized that I could construct haplotypes by phasing the variants from the same individual using the read-pairs from the same fragment. If we detected more than two haplotypes per individual, this could be a good indicator of paralogous variants. This haplotype filtering became a critical step in distinguishing high-density SNPs from paralogous variants. To perform variant annotation, I modified PowerBlast to incorporate publicly available full length mRNA data resources. A genome-wide analysis of the full pipeline would require processing jobs in batch — this was also my first experience in submitting thousands of jobs to a computer cluster for parallelized analysis.

When all the data were analyzed, curated, and organized, we wound up with close to 3 million variants. Soon Craig Venter appeared at my cubicle, and I showed him variants in donor A in several interesting regions — it was rumored that Craig was donor A, the Caucasian male whose genome had the highest coverage. The first region that I selected was APOE, a gene that I examined in detail while working at Glaxo. I told him that donor A did not have APOE4, the risk allele for Alzheimer's Disease (AD). Instead, donor A had an allele that was protective of AD but might pose higher risk for heart disease. Upon hearing my report,

Craig told me that I could write a manuscript focused on genetic variation for the journal *Science*.

While getting ready to work on our variant paper, Andy and I received the disappointing news that the company intended to patent the variants and incorporate them into a commercial product, so I would have to work on developing a subscription web portal for pharmaceutical companies. We would have to come up new ways of discovering variants for scientific publication. Therefore, I resorted to finding variants by comparing the Celera sequence with the public HGP sequence. In the final accounting, I excluded the paralog/repeat-prone regions identified by donor-specific haplotype analysis. In the end, we reported the profile of 2.1 million SNPs discovered by this approach in a chapter "A Genome-Wide Examination of Sequence Variations" in the Celera genome paper published in *Science* [10]. All variants were deposited to dbSNP, more than doubling all the publicly available SNPs at the time.

EPILOGUE

It took the president of United States, Bill Clinton, to call the race between Celera and HGP [11] a "tie" in 2000, and the leaders from both sides, Dr. Francis Collins for HGP and Craig for Celera, finally shook hands (Fig. 2). The race resulted in the completion of the HGP draft genome sequence 3 years ahead of schedule, a major benefit for those whose research would be accelerated by the availability of human genome sequence. Considering my mission to build a SNP map from non-disease donors accomplished, I joined the National Cancer Institute (NCI). I took this step on the advice of Dr. Samuel Broder, a former NCI Director who was the senior VP at Celera at the time, to continue my journey in trying to find genetic variants and disease cures. At NCI, I learned about acquired somatic mutations in cancer and how drugs such as Gleevec could be developed to specifically kill the tumor cells harboring these mutations [12]. In 2009, I finally found my own target, an activating mutation affecting the kinase JAK2 [13], present only in high-risk childhood leukemias, and targetable by JAK inhibitors (*e.g.*, ruxolitinib), which were developed initially for non-cancer blood diseases such as polycythemia vera. Motivated by a strong desire to see how my discovery would work in clinical care, I joined St. Jude Children's Research Hospital in 2010, diving deep into pediatric cancers. There are currently four clinical trials ongoing targeting the JAK/STAT pathway in pediatric leukemia [14], and I have developed the computational tools to analyze the whole-genome sequencing data of every pediatric patient at St. Jude, as part of their standard clinical care [15,16]. And so my childhood dream indeed came true, just by way of computers rather than Petri dishes.

Figure 2. Francis Collins (left) and Craig Venter (right) shook hands in 2001 at a press conference. Original photo from the web site: https://www.genome.gov/human-genome-project/20th-anniversary-of-landmark-human-genome-project-publications.

COMPLIANCE WITH ETHICS GUIDELINES

The author Jingjui Zhang declares that she has no conflict of interests.

OPEN ACCESS

REFERENCES

1. Kohara, Y., Akiyama, K. and Isono, K. (1987) The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell, 50, 495−508
2. Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993) dbEST–database for "expressed sequence tags". Nat. Genet., 4, 332−333
3. Chao, K. M., Zhang, J., Ostell, J. and Miller, W. (1995) A local alignment tool for very long DNA sequences. Comput. Appl. Biosci., 11, 147−153
4. Chao, K. M., Zhang, J., Ostell, J. and Miller, W. (1997) A tool for aligning very similar DNA sequences. Comput. Appl. Biosci., 13, 75−80
5. Zhang, J. and Madden, T. L. (1997) PowerBLAST: a new

network BLAST application for interactive or automated sequence analysis and annotation. Genome Res., 7, 649−656

6. Strittmatter, W. J., Weisgraber, K. H., Huang, D. Y., Dong, L. M., Salvesen, G. S., Pericak-Vance, M., Schmechel, D., Saunders, A. M., Goldgaber, D. and Roses, A. D. (1993) Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. Proc. Natl. Acad. Sci. USA, 90, 8098−8102
7. Weber, J. L. and Myers, E. W. (1997) Human whole-genome shotgun sequencing. Genome Res., 7, 401−409
8. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 269, 496−512
9. Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. and Lander, E. S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature, 407, 513−516
10. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001) The sequence of the human genome. Science, 291, 1304−1351
11. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al*. (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860−921
12. Druker, B. J., Sawyers, C. L., Kantarjian, H., Resta, D. J., Reese, S. F., Ford, J. M., Capdeville, R. and Talpaz, M. (2001) Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. N. Engl. J. Med., 344, 1038−1042
13. Mullighan, C. G., Zhang, J., Harvey, R. C., Collins-Underwood, J. R., Schulman, B. A., Phillips, L. A., Tasian, S. K., Loh, M. L., Su, X., Liu, W., *et al*. (2009) JAK mutations in high-risk childhood acute lymphoblastic leukemia. Proc. Natl. Acad. Sci. USA., 106, 9414−9418
14. Harvey, R. C. and Tasian, S. K. (2020) Clinical diagnostics and treatment strategies for Philadelphia chromosome-like acute lymphoblastic leukemia. Blood Adv., 4, 218−228
15. Rusch, M., Nakitandwe, J., Shurtleff, S., Newman, S., Zhang, Z., Edmonson, M. N., Parker, M., Jiao, Y., Ma, X., Liu, Y., *et al*. (2018) Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat. Commun., 9, 3962
16. Newman, S., Nakitandwe, J., Kesserwan, C. A., Azzato, E. M., Wheeler, D. A., Rusch, M., Shurtleff, S., Hedges, D. J., Hamilton, K. V., Foy, S. G., *et al*. (2021) Genomes for kids: The scope of pathogenic mutations in pediatric cancer revealed by comprehensive DNA and RNA sequencing. Cancer Discov., candisc. 1631. 2020

