Review Article

The Sequence of the Human Genome — A Contemporary Review

Hao Chen [1]*, Ying Wu [2]

1. Sun Yat-sen University Zhongshan School of Medicine, Guangzhou, China.
2. West China School of Medicine Sichuan University, Chengdu, China.

* Correspondence: haochen.med.cn@gmail.com

Abstract: The 2001 Science publication by Venter and colleagues, reporting Celera Genomics' whole-genome shotgun assembly, marked a pivotal moment in genomics. Together with the International Human Genome Sequencing Consortium (IHGSC) draft, it redefined strategies for large-scale sequencing and catalyzed broad advances in biomedical research. Celera's assembly-centric approach provided early insights into gene content, repetitive elements, and polymorphism, while simultaneously provoking debate over accuracy, completeness, and data-access models. This review situates the Celera draft in its historical context, compares its methodology and findings with the IHGSC effort, and evaluates both its immediate and long-term impacts on sequencing technology, data sharing, and medical genomics. We further trace how subsequent developments—including next-generation sequencing, long-read technologies, telomere-to-telomere assemblies, and emerging pangenome projects—have addressed limitations of the 2001 drafts. Finally, we consider the enduring legacy of the Celera paper, highlighting its scientific contributions and the lessons it offers for future public–private partnerships in genomic science.

Keywords: human genome, Celera Genomics, whole-genome shotgun, Venter 2001, Human Genome Project

1. INTRODUCTION

In February 2001, the scientific community witnessed a watershed moment: the near-simultaneous publication of two high-profile draft human genome sequences. The publicly funded International Human Genome Sequencing Consortium (IHGSC) published its results in Nature, while Celera Genomics, a private biotechnology company led by J. Craig Venter, published its work in science. Although both groups sought the same end—a reference human genome sequence—their philosophies, methodologies, and data-sharing stances diverged sharply [1]. The juxtaposition of these two efforts catalyzed the genomic era, sparking debates about science, commerce, and collaboration, while accelerating downstream research and applications in medicine, technology, and computational biology. This review focused on Celera's 2001 Science contribution, "The sequence of the human genome," which presented a draft genome assembled using whole-genome shotgun (WGS) sequencing. This approach, bold in scope, represented a radical departure from the hierarchical clone-by-clone strategy employed by the IHGSC. Celera's methodology, findings, and the controversy surrounding its release illustrate both the promise and the limitations of early genome sequencing efforts [2]. The Human Genome Project (HGP) began in 1990 as a large-scale international public effort. Its guiding principle was to produce a high-quality reference genome through map-based sequencing, where large bacterial artificial chromosome (BAC) clones were ordered along chromosomes and then sequenced in a hierarchical fashion. This strategy emphasized accuracy, reproducibility, and free public data release. By the late 1990s, Celera Genomics proposed an alternative path. Rather than a

hierarchical BAC-by-BAC approach, Celera advocated whole-genome shotgun sequencing: randomly fragmenting the genome into smaller pieces, sequencing these pieces using Sanger technology, and relying on computational algorithms to reassemble them into contigs and scaffolds. Proponents argued that WGS would be faster and cheaper, while skeptics cautioned that the complexity of the human genome—with its vast repeats, segmental duplications, and low-complexity regions—would overwhelm available computational methods. Celera's approach was not just scientific but also philosophical [3]. The company planned to monetize access to its assembly and annotations by offering subscription-based licensing to pharmaceutical companies and academic institutions. This challenged the HGP's open-data ethos and ignited fierce debate about ownership of fundamental biological knowledge. Celera's strategy involved generating a large volume of Sanger reads from different fragment libraries. Random shearing of genomic DNA produced overlapping fragments of varying sizes, which were cloned into plasmids, fosmids, and BACs. Paired-end sequencing of these clones provided linking information to order contigs and bridge repetitive elements. Assembly posed the greatest computational challenge [4]. The Celera Assembler, a then state-of-the-art algorithm, integrated repeat-masking strategies, overlap-layout-consensus methods, and hierarchical scaffolding to produce draft contigs. Deep coverage was critical: the project produced approximately 27 million reads, achieving ~5.4-fold coverage of the genome. Paired-end information from large-insert clones was essential for spanning repeats and scaffolding across difficult regions. Even with these innovations, Celera's assembly had limitations. Highly repetitive sequences, centromeric regions, and long segmental duplications were poorly resolved. Gaps remained, and the assembly was less contiguous than the HGP's map-based draft in some regions. Nonetheless, the WGS approach succeeded to an extent many had doubted possible, demonstrating that large complex genomes could be tackled without clone-by-clone sequencing.

2. PRINCIPAL FINDINGS REPORTED BY CELERA

Celera's landmark 2001 Science publication fundamentally altered genomic science by presenting the first private-sector draft of the human genome sequence assembled through a whole-genome shotgun (WGS) strategy [2]. One of the most striking outcomes of this work was its estimate of the number of human protein-coding genes. For decades prior, speculation placed the count as high as 80,000–120,000 genes, a reflection of the assumption that organismal complexity required vast protein diversity. Celera's draft, alongside the publicly funded International Human Genome Sequencing Consortium (IHGSC) release in Nature [4], revised this view downward to ~26,000–38,000 protein-coding genes in their early assessments. This revelation, echoed by the IHGSC's ~30,000 figure, initiated a paradigm shift: complexity in humans could not be explained solely by gene number but rather by regulatory networks, post-transcriptional processes such as alternative splicing, and the functional roles of noncoding sequences [6]. Current estimates now suggest ~19,000–20,000 canonical protein-coding genes [5], affirming that Celera's downward revision was prescient, even if somewhat inflated relative to contemporary annotations. The realization that humans possess fewer protein-coding genes than previously assumed redirected scientific focus toward the regulatory genome. Advances in epigenomics, single-cell transcriptomics, and long-read sequencing have since demonstrated that alternative splicing, RNA editing, small RNAs, and enhancer-promoter interactions are central determinants of cellular diversity and organismal complexity [3]. In hindsight, the Celera and IHGSC findings catalyzed the post-genomic era by compelling researchers to look beyond protein-coding regions, ushering in initiatives such as ENCODE and GT Ex, which systematically annotated noncoding elements and their tissue-specific activity. The intellectual legacy of this transition continues to shape genomic medicine, where noncoding variants are increasingly implicated in common complex diseases. Another critical contribution of the Celera draft was its handling of repetitive DNA. The human genome is dominated by repetitive elements, including retrotransposons such as long interspersed nuclear

elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeat (LTR) retroelements, together accounting for approximately half the genome [7]. Celera's assembly confirmed this abundance and acknowledged the challenges such sequences posed to WGS assembly given the short Sanger read lengths available at the time. Highly identical segmental duplications and structurally complex regions proved especially difficult, leaving gaps and mis assemblies that persisted in both Celera and public consortium drafts. These limitations underscored the necessity of future technological advances. Indeed, it was not until the adoption of long-read sequencing (PacBio HiFi and Oxford Nanopore) and complementary scaffolding technologies (e.g., Hi-C, optical mapping) that near-complete assemblies of these repeat-rich regions were achieved, culminating in the Telomere-to-Telomere (T2T) Consortium's complete human genome in 2022 [8]. Retrospectively, Celera's inability to resolve repeats was not a shortcoming of methodology but rather a technological constraint—one that shaped the trajectory of sequencing innovation for two decades. Celera's early attempts at cataloging human genetic variation also represent a foundational milestone. By sequencing multiple individuals and integrating multi locus comparisons, Celera reported one of the first systematic catalogs of single nucleotide polymorphisms (SNPs) and structural variation in humans. These efforts, although limited in scope and biased by donor representation, anticipated the central role of genetic variation in disease biology and precision medicine. At the time, however, the underrepresentation of population diversity and the challenges of resolving repetitive and segmentally duplicated regions meant that large fractions of structural and non-European variation remained invisible. Contemporary research has addressed these shortcomings through large-scale initiatives such as the 1000 Genomes Project, the Genome Aggregation Database (gnomAD), and more recently, the Human Pangenome Reference Consortium (HPRC), which leverages long-read sequencing to capture structural variants and haplotype diversity across global populations [9]. These efforts have revealed that structural variation contributes more nucleotide differences between individuals than SNPs and has profound consequences for disease susceptibility, drug response, and evolutionary adaptation [2]. The limitations of Celera's WGS assembly also sparked debates about reference genome philosophy that remain relevant today. Celera's assembly was a mosaic derived from several individuals, while the IHGSC draft emphasized a single haploid composite from a smaller donor pool. Both strategies underrepresented population diversity and left unresolved regions that are now known to be functionally important, such as centromeres, segmental duplications, and structurally polymorphic loci [10]. In response, the field has progressively moved toward pan genomic representations—graph-based references that incorporate haplotypic and structural variation across ancestries rather than privileging a single linear reference [5]. In this sense, the Celera draft's multi-individual approach anticipated contemporary shifts toward inclusive references, though its proprietary data access model was antithetical to the open-access ethos that ultimately prevailed. From a historical perspective, Celera's 2001 publication exemplifies the complex interplay between scientific innovation, commercial interests, and public policy. Celera's business model—restricting access to assemblies under subscription or licensing—provoked widespread concern about the privatization of genomic information, particularly when juxtaposed against the HGP's commitment to immediate, unrestricted public release. These tensions accelerated the adoption of open data norms codified in the Bermuda Principles and later reinforced by the Fort Lauderdale and Toronto agreements [2]. The consensus that genomic data should be freely accessible has since become foundational to biomedical research, with Celera's legacy serving as a cautionary tale about the risks of restricting foundational resources. Ironically, while Celera's assembly was technologically groundbreaking, its influence on scientific culture was equally profound: it highlighted the necessity of public data infrastructures to maximize downstream innovation. Two decades later, the impact of Celera's contribution can be evaluated in light of subsequent advances. On one hand, many of its limitations—gaps in repetitive regions, underrepresentation of population diversity, and inflated gene counts—were resolved only through collective international efforts and newer

sequencing platforms. On the other hand, Celera's demonstration that whole-genome shotgun sequencing was feasible at the scale of the human genome was transformative, paving the way for next-generation sequencing (NGS) and modern assembly strategies. Today's genomes, whether generated for clinical diagnostics, population studies, or synthetic biology, are built upon the methodological lineage inaugurated by Celera's audacious application of WGS. Without this proof of concept, the acceleration of sequencing innovation and its integration into clinical medicine might have been delayed significantly.

3. IMMEDIATE SCIENTIFIC AND SOCIAL IMPACT

The release of the draft human genome sequences—Celera's and the public consortium's—was transformative not only for scientific knowledge but also for research infrastructure and translational applications. The immediate utility of draft assemblies was apparent: they provided the substrate for the first genome browsers, which allowed researchers to navigate sequence data interactively, and for the development of gene annotation pipelines that mapped coding and noncoding features onto the genome. These resources rapidly became foundational tools across biology and medicine. In parallel, SNP arrays and early catalogs of common variants enabled large-scale association studies, laying the groundwork for genome-wide association studies (GWAS) that emerged later in the decade. Functional genomics resources, including expression arrays and cross-species comparative genomics, flourished as researchers now had reference coordinates for mapping experimental data. Celera's data—delivered through commercial licensing—stimulated a wave of biotech ventures, proprietary databases, and diagnostic tool development [11]. Meanwhile, the IHGSC's insistence on open, immediate release of data (a principle later enshrined in the "Bermuda Principles") established community expectations that have since defined genomics: rapid, unrestricted access to foundational sequence data. This divergence between commercial and public-access philosophies generated intense debate in 2001, yet ultimately shaped a hybrid ecosystem in which industry and academia coevolved, with public reference data serving as the bedrock for innovation. However, the limitations of both drafts were quickly evident [12]. Chromosome-level analyses and targeted finishing projects highlighted gaps, mis assemblies, and regions of uncertain accuracy—especially within repetitive and structurally complex domains such as centromeres, telomeres, and segmental duplications. These shortcomings underscored the challenges of assembling the human genome with Sanger-era read lengths and the necessity of iterative validation and gap-filling. As sequencing technologies advanced, the field converged on more robust pipelines for finishing, quality assessment, and community curation, processes that remain critical in modern reference genome projects.

4. TECHNOLOGICAL TRAJECTORY SINCE 2001: ADDRESSING THE DRAFTS' LIMITATIONS

The mid-2000s ushered in a second major inflection point with the introduction of massively parallel short-read sequencing platforms, most notably Illumina. These technologies slashed sequencing costs by several orders of magnitude, enabling population-scale resequencing, genome-wide association studies (GWAS), transcriptomics, and epigenomics on a scale that would have been unimaginable in 2001. However, short-read data exacerbated challenges in de novo assembly and structural variant detection. Reads of only 50–150 bp could not reliably resolve long repeats, segmental duplications, or complex rearrangements, reinforcing the reliance on reference-guided approaches and leaving many structurally complex regions inaccessible [13]. The emergence of long-read sequencing technologies—Pacific Biosciences' circular consensus (HiFi) reads and Oxford Nanopore Technologies' nanopore sequencing—addressed many of these limitations by producing read lengths spanning thousands to hundreds of thousands of bases. These longer sequences enabled far more contiguous assemblies, better characterization of structural variation, and the resolution of repeat-rich regions that had resisted previous efforts. The culmination of these advances was the Telomere-to-Telomere (T2T)

Consortium's 2022 release of a complete, gapless human genome assembly (T2T-CHM13). This milestone filled hundreds of unresolved gaps from the 2001 drafts, corrected mis assemblies, and revealed hundreds of millions of additional bases, novel genes, and unexpected structural complexities [14]. The T2T reference highlighted the degree to which technological progress could overcome the inherent limitations of Sanger-based WGS. At the same time, the field has increasingly recognized the limitations of a single, linear reference genome in capturing the breadth of human genomic diversity. Building on lessons from Celera's early acknowledgment of population variation, the Human Pangenome Reference Consortium (HPRC) and similar initiatives have moved toward graph-based, multi-assembly references. By integrating sequences from diverse ancestries, these pangenome frameworks aim to reduce reference bias, improve variant calling across populations, and provide more equitable foundations for clinical genomics. Collectively, these projects extend and correct both the public and Celera drafts, ensuring that the reference framework for genomics reflects not only technological completeness but also global human diversity [15].

5. LONG-TERM TRANSLATIONAL IMPACTS

The 2001 drafts—both Celera's and the public consortium's—did more than establish a scientific milestone; they provided the coordinate systems and gene catalogs that became the foundation of modern medical genomics. These references enabled the development of exome sequencing, which rapidly proved transformative in the molecular diagnosis of Mendelian disorders. Over time, the shift from exome to whole-genome sequencing expanded diagnostic yield, uncovering causal variants in noncoding and structural regions that early assemblies had left incomplete. More recent gapless assemblies and pangenome initiatives have further improved diagnostic sensitivity, particularly for variants in complex or previously inaccessible loci. Reference genomes also catalyzed a wide spectrum of translational applications. Pharmacogenomic allele catalogs, tumor–normal sequencing pipelines in oncology, and large-scale population biobanks all depend critically on reliable reference sequences and representative variant data. The iterative refinement of reference assemblies and variant catalogs has improved the accuracy of variant interpretation, reduced reference bias, and expanded the scope of clinically actionable findings. In this way, the trajectory from Celera's draft genome to today's pan genomic references illustrates how foundational infrastructure not only accelerated discovery but also reshaped clinical practice by enabling increasingly precise, equitable, and data-driven decision-making.

6. CONCLUSION

The impact of the draft human genomes on medicine and biotechnology cannot be overstated. The coordinate systems and gene catalogs established in 2001 provided the essential scaffolding for exome sequencing, clinical resequencing, and precision diagnostics. Whole-exome and whole-genome sequencing became routine tools for diagnosing rare Mendelian disorders, while reference-guided approaches enabled tumor-normal sequencing in oncology, pharmacogenomics, and the establishment of large-scale biobanks. Improvements in reference assemblies and variant catalogs over time have enhanced the accuracy of variant interpretation and reduced sources of bias in clinical practice. Thus, even as the limitations of the original drafts became apparent, their utility as frameworks for clinical and translational research was profound. Perhaps the most enduring legacy of Celera's Science paper is its dual technical and societal significance. Technically, it demonstrated that large-scale genomic projects could be executed rapidly and with unconventional strategies, inspiring subsequent projects that combined ambition with computational innovation. Societally, it highlighted the importance of openness, collaboration, and equitable access in realizing the full benefit of genomic resources. The debates over data sharing and commercialization that surrounded Celera's publication continue to echo in contemporary discussions about biobank governance, genomic privacy, and the balance of public and private interests in biomedical research. In summary, "The sequence of the human genome" by

Venter and colleagues was a transformative contribution to biology. It advanced an alternative, computation-driven route to genome assembly, validated the feasibility of rapid large-scale sequencing, and set the stage for decades of downstream innovation. While subsequent advances in sequencing technology and assembly strategies have extended, corrected, and diversified the genomic landscape, the historical importance of the 2001 Celera paper endures. Its legacy lies not only in its technical demonstration but also in its role as a catalyst for discussions about openness, validation, and inclusivity—principles that remain essential for maximizing the scientific and clinical benefits of genomics.

REFERENCES

[1] International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921.

[2] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … & Zhu, X. (2001). The sequence of the human genome. Science, 291(5507), 1304–1351.

[3] Miao Cai (2025). Phase 1/2 Trial of Lentiviral Gene Therapy with Reduced-Intensity Conditioning For Sickle Cell Disease. Journal of Medical Innovations, 4(07):483-488.

[4] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., … & Phillippy, A. M. (2022). The complete sequence of a human genome. Science, 376(6588), 44–53.

[5] Human Pangenome Reference Consortium. (2023). A draft human pangenome reference. Nature, 617, 312–326.

[6] venter, J. C., Smith, H. O., & Adams, M. D. (2015). The sequence of the human genome. *Clinical chemistry*, *61*(9), 1207-1208.

[7] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., ... & Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44-53.

[8] Durrant, M. G., Fanton, A., Tycko, J., Hinks, M., Chandrasekaran, S. S., Perry, N. T., ... & Hsu, P. D. (2023). Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nature biotechnology*, *41*(4), 488-499.

[9] Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., ... & Taipale, J. (2022). Sequence determinants of human gene regulatory elements. *Nature genetics*, *54*(3), 283-294.

[10] Wei, W., Schon, K. R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., ... & Chinnery, P. F. (2022). Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature*, *611*(7934), 105-114.

[11] Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., ... & Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978), 344-354.

[12] Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., ... & Ren, B. (2021). A single-cell atlas of chromatin accessibility in the human genome. *Cell*, *184*(24), 5985-6001.

[13] Keller, E. F. (2022). Nature, nurture, and the human genome project. In *The Ethics of Biotechnology* (pp. 335-354). Routledge.

[14] Edith Ahmadu (2025). Early and Periodic Screening, Diagnostic, and Treatment (EPSDT): A Critical Analysis of Medicaid's Mandate for Children and Adolescents. Dinkum Journal of Medical Innovations, 4(02):58-62.

[15] Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., ... & Pierrot, T. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, *22*(2), 287-297.

