frontiers | Frontiers in Science

TYPE Frontiers in Science Viewpoint
PUBLISHED 04 September 2025
DOI 10.3389/fsci.2025.1671650

OPEN ACCESS

EDITED AND REVIEWED BY
Jennifer Graves,
La Trobe University, Australia

*CORRESPONDENCE
Gane Ka-Shu Wong
gane@ualberta.ca

RECEIVED 23 July 2025
ACCEPTED 26 August 2025
PUBLISHED 04 September 2025

CITATION
Wong GK-S. Earth BioGenome Project: the beginning of the end or the end of the beginning?
Front Sci (2025) 3:1671650.
doi: 10.3389/fsci.2025.1671650

Earth BioGenome Project: the beginning of the end or the end of the beginning?

Gane Ka-Shu Wong[1,2]*

[1]Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, AB, Canada, [2]BGI-Research, Shenzhen, China

KEYWORDS

big science, genome, pangenome, biodiversity, phenotype

A Viewpoint on the Frontiers in Science Lead Article

The Earth BioGenome Project Phase II: illuminating the eukaryotic tree of life

Key points

- Reference genomes are insufficient to capture the genetic diversity of a species; we need pangenomes sequenced across cohorts of individuals from each species.
- Integrating phenotypic data with artificial intelligence will be crucial to ensuring that the Earth BioGenome Project generates products or results useful to society.
- The long-term impact of this project will depend on a new generation of scientists trained to perform activities beyond just the generation of reference genomes.

Introduction

> "Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning." Winston Churchill (November 10, 1942)

The Earth BioGenome Project (EBP) aims to generate high-quality reference genome sequences for all named eukaryotic species, an estimated 1.67 million species of plants, animals, fungi, etc. In their lead article for Frontiers in Science (1), Blaxter et al. celebrate past successes and look ahead to the future. Under Phase I, over the last 3 years, EBP-affiliated projects sequenced 1,667 high-quality genomes from a wide diversity of taxa at an average cost of US$28,000 per species. Under Phase II, optimistically within the next 4 years, 150,000 genomes are expected to be sequenced at an even lower cost. Some species will be more difficult, but the authors are confident that the remaining technical challenges

(detailed thoroughly in their manuscript) are surmountable. Among the non-technical geopolitical challenges is their commendable goal of engaging the Global South, where many critical species must be sourced. Regardless of how it unfolds, this ambitious endeavor is destined to become a milestone in science history, much like the Human Genome Project (HGP) a quarter of a century earlier.

HGP–EBP parallels

The HGP provided the first version of the human genome (and a few model organisms), while the EBP seeks to do the same for all other eukaryotic species. Both projects share a sense of completeness: sequencing "every" base pair in a genome versus "every" named eukaryotic species. Both are expensive, costing billions, and justified by their potential impacts not only on science but also on society. For someone like me, whose career spans both eras, the parallels are especially resonant.

In the heyday of the HGP, from the late 1990s to early 2000s, I was a top lieutenant at Maynard Olson's Genome Center at the University of Washington, part of the Department of Molecular Biotechnology (now Genome Sciences) founded by Leroy Hood. Olson and Hood were members of the special committee of the United States National Academy of Sciences that defined the HGP's original goals (2). Hence, I was privy to many of the internal discussions behind key policy decisions, and that experience now informs my views of the EBP. As Mark Twain once said, "History does not repeat itself, but it often rhymes".

Two questions often asked of all "big science" projects are: (i) when is the project finished? and (ii) how do we make this useful? My comments on the former draw on how the HGP made its decisions and what happened after the project was declared complete. As for the latter, there are two interpretations. For the scientists, it may suffice that their data is foundational. But for the public that funds such projects, usefulness is often measured by tangible outcomes, such as new medicines, new crops or improved conservation efforts. Achieving these requires much more than genomics, but as with the HGP, it is in the EBP's interests to facilitate such activities.

When is the project finished?

The HGP originally promised to sequence every base pair to an accuracy of 10^{-4}. Since human genetic variation occurs at a rate of 10^{-3}, this ensures that any observed differences between two sequenced genomes would more likely reflect polymorphisms than experimental errors. An exception was made for highly repetitive heterochromatic regions (e.g., centromeres, telomeres), which at the time could not be cloned, amplified, sequenced, or assembled. No one objected to excluding such regions, as there are few genes embedded within heterochromatin.

When competition arose between the public consortium led by Francis Collins at the National Human Genome Research Institute (NHGRI) and the private effort led by Craig Venter at Celera Genomics, debate flared within the public consortium over whether to abandon the original quality standards. According to Lee Rowen at the Institute for Systems Biology (ISB), which side people supported depended on their outlook in life. Astonished that we raised US$3 billion from the government, some took a "cathedral building" approach, determined to do it right because there would be no second chance. Others, skilled fundraisers, argued that "good enough" data would suffice to justify the expenditure. Being optimists by nature, they also argued that regions too expensive or impossible to sequence with existing technologies could be deferred to future projects, when presumed technological advances would make such regions tractable.

History has largely vindicated the techno-optimist perspective. The absolutist goal of assembling every chromosome into one contiguous sequence was not met by either consortium. The first telomere-to-telomere (T2T) assembly was reported only in 2022, two decades later (3). This was made possible by the invention of single molecule sequencing by Pacific BioSciences of California and Oxford Nanopore Technologies, which provided much longer read lengths (albeit moderate accuracies) capable of sequencing through heterochromatin.

Another question often asked of the HGP was: whose genome was sequenced? No single genome can capture the panoply of sequences found across the human population, and indeed, a Human Pangenome Reference Consortium is now generating haplotype-phased diploid assemblies from a cohort of genetically diverse individuals (4). This too reflects technological advances. Without the million-fold reduction in sequencing costs, largely driven by Illumina, such an effort would have been unthinkable.

Circling back to the EBP, the ultimate goal should be a pangenome for all named eukaryotic species. However, with current technologies, costs remain prohibitive. EBP quality standards (5) (1 Mb contig N50, chromosome-scale scaffolds, and 10^{-4} base pair accuracy) are an acknowledgment of reality. Moreover, some species will be orders of magnitude more expensive due to genome size variation, and many groups will be reluctant to sequence these to the same "high" standards. People will compromise because even imperfect genomes have proven valuable.

A further dimension to the EBP, absent from the HGP, is the impossibility of sampling some species, whether physically or ethically. This should be as acceptable as the HGP's omission of heterochromatin. Other species will be difficult but not impossible to sample. Ironically, the more challenging the acquisition, the stronger the motivation to sequence those samples to the desired standards.

How do we make this useful?

Multi-genome datasets are often published with an evolutionary narrative, in part because computationally derived conclusions can satisfy the reviewers. By contrast, developing a useful product requires phenotypic experiments, which are major projects in their own right and inconceivable on the scale of the EBP. As a result, the enormous value of the sequences being generated will not be fully realized until phenotyping becomes more scalable. Hope lies in

artificial intelligence (AI), which has been astoundingly successful at predicting three-dimensional protein structures, a breakthrough recognized by the 2024 Nobel Prize in Chemistry. We do not yet know how (or if) AI can be extended from predicting simple molecular phenotypes to predicting complex organismal phenotypes. Still, it is a sign of our times that Demis Hassabis, co-developer of the AlphaFold algorithms for protein folding, has declared: "One day […] we can cure all disease with the help of AI […] [m]aybe within the next decade or so" (6).

The success of AI depends on the examples used to train it. For protein folding, the training set was a quarter-million experimentally determined protein structures from the Protein Data Bank (PDB). For phenotyping, the concept of a training set is inherently open-ended, as phenotypes themselves cannot all be defined. Two experiments illustrate this breadth. Both were funded to solve important problems unrelated to AI, one in neuroscience (optogenetics), the other in agriculture.

(i) Optogenetics uses light to activate or inhibit genetically defined classes of neurons in awake behaving animals, including humans. Algae derived channelrhodopsin proteins are expressed heterologously in these animals. The original experiments used a protein from *Chlamydomonas reinhardtii*; but many years later, experimental characterization of channelrhodopsin proteins from 127 phylogenetically diverse algal species uncovered a series of new proteins with improved performance, e.g., red-shifted, faster responses, and higher currents (7). Multiple companies now use these proteins to restore vision in patients with retinitis pigmentosa. In the latest report (8), a woman who had been legally blind can now sit down on the beach and read a book again.

(ii) Modern agriculture suffers from a lack of genetic diversity. Although the germplasm banks abound with useful alleles, crossing them with elite cultivars often transfers unwanted traits along with the desired traits. To unlock this resource, we must first identify the responsible haplotypes. A recent study measured 137 phenotypes in the A. E. Watkins collection of global bread wheat landraces and in recombinant inbred lines generated by crossing selected landraces with an elite cultivar (9). Useful haplotypes included ones conferring yellow rust resistance, higher yield and height, and an alternative dwarfing gene for the Green Revolution. These findings are already being applied in commercial breeding programs.

Biomedical research has the funding to generate incredible resources such as the United Kingdom (UK) Biobank[1], which contains genomic, proteomic, and metabolomic data for half a million human subjects, together with lifelong health records (10). A comparable effort at the scale of the EBP is unrealistic. But is it necessary? In other words, how many phenotypes, and which ones, must be collected across how many species, and which ones, for AI to generalize? Can we bootstrap our way toward a better future by generating abundant phenotypic data now, in the hope that AI will eventually eliminate the need for such experiments?

Epilogue

Scientific progress requires that young scientists build on and improve the achievements of their predecessors. In the case of the HGP, it was a new generation of scientists who created the first T2T assembly, filling in the gaps and correcting errors in the original HGP sequence many years after their predecessors sailed off into the sunset with well-deserved accolades. Similarly, it was a new generation of scientists who constructed the UK Biobank, generating phenotypic data that are empowering the next cycle of discoveries. Long-term success for the EBP will likewise not come from sequence data alone, regardless of its quality. It will depend on the follow-up studies by the next generation of scientists. I urge the EBP to expand their vision to include training young researchers not only in genome sequencing, but also in downstream activities that will translate genomes into knowledge and solutions. If so, the eventual completion of the EBP will be seen as the fitting end to a glorious beginning.

Acknowledgments

The author would like to thank Lee Rowen (Institute for Systems Biology, Seattle, WA, United States), Shifeng Cheng (Agricultural Genomics Institute at Shenzhen, Shenzhen, China), and Lucas Dileo (Broadland Advisors, Cambridge, MA, United States) for their advice on this viewpoint, on the basis of which some modifications were made.

Statements

Author contributions

GW: Conceptualization, Visualization, Software, Investigation, Resources, Funding acquisition, Validation, Formal Analysis, Methodology, Writing – review & editing, Writing – original draft, Data curation, Supervision, Project administration.

Funding

The author declared that no financial support was received for this work and/or its publication.

Conflict of interest

The author is an unpaid advisor for the company BGI-Research, Shenzhen, China, which he played a pivotal role in establishing. The author is a co-inventor along with 4 Massachusetts Institute of Technology scientists from the laboratory of Edward Boyden on the

1 The UK Biobank recorded 1914 papers in 2023 and 2547 papers in 2024.

optogenetics patents implicitly referenced in the paragraph about restoring vision.

Generative AI statement

The author declared that no generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Blaxter M, Lewin HA, Di Palma F, Challis R, da Silva M, Durbin R, et al. The Earth BioGenome Project Phase II: illuminating the eukaryotic tree of life. *Front Sci* (2025) 3:1514835. doi: 10.3389/fsci.2025.1514835

2. United States National Research Council Committee on mapping and sequencing the human genome. Mapping and sequencing the human genome. Washington, DC: National Academies Press (1988). Available at: https://www.ncbi.nlm.nih.gov/books/NBK218249/

3. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science* (2022) 376:44–53. doi: 10.1126/science.abj6987

4. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. *Nature* (2023) 617:312–24. doi: 10.1038/s41586-023-05896-x

5. Earth BioGenome Project Scientific Subcommittee for sequencing and assembly. *Report on Assembly Standards. Version 6.0*. Earth BioGenome Project. Available at: https://www.earthbiogenome.org/report-on-assembly-standards/

6. Scott P. Artificial intelligence could end disease, lead to "radical abundance," Google DeepMind CEO Demis Hassabis says. *CBS News* (2025). Available at: https://www.cbsnews.com/news/artificial-intelligence-google-deepmind-ceo-demis-hassabis-60-minutes-transcript/

7. Klapoetke NC, Murata Y, Kim SS, Pulver SR, Birdsey-Benson A, et al. Independent optical excitation of distinct neural populations. *Nat Methods* (2014) 11:338–46. doi: 10.1038/nmeth.2836

8. Drew L. Restoring vision with optogenetics. *Nature* (2025) 639:S7–9. doi: 10.1038/d41586-025-00656-5

9. Cheng S, Feng C, Wingen LU, Cheng H, Riche AB, Jiang M, et al. Harnessing landrace diversity empowers wheat breeding. *Nature* (2024) 632:823–31. doi: 10.1038/s41586-024-07682-9

10. United Kingdom Biobank. Publications catalog (2025). Available at: https://biobank.ndph.ox.ac.uk/showcase/docs.cgi?id=2