1000 Genomes Project Data Biography

Ethan Haque

February 2024

The 1000 Genomes Project dataset is a landmark dataset in the field of genomics whose primary goal was to find common genetic variants with frequencies of at least 1% in the populations studied. In total, the final public dataset contains the genetic information for 2,504 individuals from 26 populations and took from January 2008 to 2015 to complete [3].

The human genome consists of about three billion nucleotides from a four letter alphabet A, T, C, and G with approximately 20,000 protein coding genes. The DNA of any two individuals is remarkably similar with the 1000 Genomes Project finding that a typical individual genome differs from the reference human genome by about 0.6% [3]. As uncompressed plain-text a single human genome can take up three gigabytes of space. Taking into account the minimal variation between genomes, this size can be greatly reduced. Still, sequencing data of this magnitude requires lots of resources to store. Prior to the 1000 Genomes Project, the largest genome sequencing project to date was the Human Genome Project which took over a decade and billions of dollars to sequence a single human genome. The Human Genome Project finished April 14, 2003 [1]. In only five years the 1000 Genomes Project sought to surpass the Human Genome Project by sequencing the genomes of over two thousand individuals from various populations around the world. This lofty goal was made

possible by significant technological improvements in sequencing methods and data analysis, which drastically reduced the cost and time required for genome sequencing, and a large international effort from institutes around the world.

The 1000 Genomes Project was made possible by hundreds of institutions and thousands of researchers. Some familiar names are Harvard, Stanford, MIT, Cornell, Illumina, the Max Planck Institute, Thermo Fisher Scientific, U Chicago, and Oxford to name a few. Because of the distributed nature of the project, it's hard to say exactly how much it cost in total. Early estimates projected the entire project to cost half a billion dollars, but more cost effective sequencing methods were developed while the project was ongoing, which drastically reduced costs. Later estimates projected the total cost to end up around $30 to $120 million dollars in total [7]. Although, the sheer number of people involved in this project makes it hard to quantify just how much work the collective put in and how valuable their time was. A more realistic estimate after the fact would likely be much higher.

An unfortunate reality is that many researchers who dedicated years of their lives to this project do not receive recognition for their contributions. To cite the 1000 Genomes Project, the International Genome Sample Resource (IGSR) who currently maintains the 1000 Genomes Project recommends to cite the final paper in a series of several groups of papers that is titled "A Global Reference for Human Genetic Variation" [2]. The official publication in Nature lists over 200 individual authors as primary authors on the paper, nearly all of which get grouped under the pseudonym *et al.*. This paper is published under the creative commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence, which makes it freely redistributable further complicating the issue of crediting the creators. The physical version of the paper lists, "The 1000 Genomes Project Consortium*" as the author of the paper with the footnote, "Lists of partic-

ipants and their affiliations appear in the online version of the paper." These logistical issues make it nearly impossible to trace the line of credibility and reproducability for the project [3].

The opposite issue to crediting exists for the 1000 Genomes Project as well: the anonymized data the 1000 Genomes Project provides could be de-anonymized. The final dataset is available freely online at `https://www.internationalgenome.org/data` and contains nearly the entire genome for thousands of individuals. Technically, the entire genome for any of these individuals is not available. At the time, it was impossible to construct a full, gapless genome of a human. Neither the 1000 Genomes Project nor the Human Genome Project actually constructed a full human genome. It turns out this extremely difficult, and the first gapless assembly of a human genome was finished in January 2022 [5]. However, these gaps in the 1000 Genomes Project dataset are not a problem for most research purposes. The information contained in the non-gap regions accounts for the majority of an individual's phenotypic information. As such, it is possible to answer many *Guess Who?*-type questions like is this person a male by sex? Does this person have brown eyes? Does this person have type-O blood? These types of questions can quickly narrow down the search space for an individual [6].

Furthermore, by analyzing the genetic variations present in an individual's genome it's possible to construct facial composites that can be combined with other types of metadata and pictures available on social media to associate anonymized genomics data with a specific individual [4]. In many ways, a person's DNA sequence is their ultimate fingerprint. It contains information that is intimately specific to the individual like predisposition to disease and ancestral information that could be used to do harm. Take for example the Chinese government's genocide of Uyghur Muslims and their mass collection of Uyghur

DNA to assist in DNA-based facial reconstruction [8]. This information has already and continues to be used to automate, justify, and intensify racial profiling [9]. Many de-anonymizing attack techniques exist and have been successfully used to de-anonymize genomes available online [6].

At the time of its conception, the 1000 Genomes Project was a monumental and revolutionary idea. The data and associated papers have become some of the most cited resources in the entire field of biology and spurred a decade of innovation [3]. Its success has paved the way for further large-scale genetic studies and has highlighted both the potential and the challenges associated with genomic data. At the time of its creation, genomic data was so expensive and difficult to obtain that relatively little was known about how this type of data could be used to cause systematic harm. As a result, later projects modeled after the 1000 Genomes Project initiated conversations about more ethical considerations and stricter privacy measures to protect individuals' genetic information. While the project has undeniably advanced our understanding of human genetics, it also serves as a reminder of the ongoing need to balance scientific progress with ethical responsibility.

References

[1] URL: `https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project`.

[2] URL: `https://www.internationalgenome.org/faq/how-do-I-cite-IGSR`.

[3] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". en. In: *Nature* 526.7571 (Oct. 2015), pp. 68–74.

[4] Peter Claes, Harold Hill, and Mark D Shriver. *Toward DNA-based facial composites: Preliminary results and validation*. Aug. 2014. URL: https://pubmed.ncbi.nlm.nih.gov/25194685/.

[5] Prabarna Ganguly and Rachael Zisk. *Researchers generate the first complete, gapless sequence of a human genome*. Mar. 2022. URL: https://www.genome.gov/news/news-release/researchers-generate-the-first-complete-gapless-sequence-of-a-human-genome.

[6] Mathias Humbert et al. *De-anonymizing genomic databases using phenotypic traits*. May 2015. URL: https://hal.science/hal-01151960.

[7] Kara Rogers. *1000 genomes project*. URL: https://www.britannica.com/event/1000-Genomes-Project.

[8] Sui-lee Wee and Paul Mozur. *China uses DNA to map faces, with help from the West*. Dec. 2019. URL: https://www.nytimes.com/2019/12/03/business/china-dna-uighurs-xinjiang.html.

[9] Manoush Zomorodi, Katie Monteleone, and Sanaz Meshkinpour. *How facial recognition allowed the Chinese government to target minority groups*. Dec. 2022. URL: https://www.npr.org/2022/12/09/1141627539/how-facial-recognition-allowed-the-chinese-government-to-target-minority-groups.