Decoding “Y”: Machine Learning Algorithms Led to the First Fully Sequenced Human Y Chromosome in History

Isaac Yoo

Three years ago, in one of the biggest breakthroughs in genetics, the human X chromosome was decoded, opening countless doors for gene research and therapy (*NIH researchers generate complete human X chromosome sequence*, 2020). However, the genomic configuration of the Y chromosome, half of the DNA of all of the men on this planet, lacked comprehensive sequencing (*World Bank Open Data*, n.d.).

Until this year.

Just three months ago, in a historic moment, a team of researchers finally fully sequenced the human Y chromosome, revealing all 62,460,029 base pairs.

To provide some background, deoxyribonucleic acid (better known as DNA), is commonly referred to as the “blueprint” of our body, and this double helix determines the function of various life processes in our body, such as protein production and cell division. Furthermore, the structure of our DNA and the sequence of critical base pairs dictates the function of each gene segment.

Our DNA has four nitrogenous bases (adenine, thymine, guanine, and cytosine) and each single helix of DNA has a sequence that these bases are ordered in. In the common DNA model, these bases form the “bridge” aspect that connects the two helixes together and each of these bases will form bonds with another base present on the other helix, creating the bridge. However, they will only form bonds with one specific other base: for example, adenine will only form a bond with thymine, and guanine will only form a bond with cytosine. These bonds between these nitrogenous bases are referred to as “base pairs”, and the order of these base pairs in DNA define

its function and how our body reads it. Additionally, chromosomes, structures that store DNA, bundle up this genetic material for vital life processes such as mitosis (cell division) and reproduction. Consequently, being able to read the sequence of these base pairs in chromosomes is critical to understanding our DNA.

Unfortunately, the Y chromosome was acknowledged by many to be a notoriously difficult sequence. The presence of complex repeating structures and complicated palindromes and duplications, made fully sequencing this chromosome a daunting task (Hallast et al., 2023). But over the past two decades, countless studies have slowly chipped away at this chromosome, which led to our previous best model: GRCh38. With a seemingly strong reference of about 30,000,000 base pairs, this model at the time seemed rather serviceable (Rhie et al., 2023). Unfortunately for the GRCh38 model, as I might have previously mentioned, 30,000,000 base pairs is less than half of the true size of the Y chromosome.

In a scientific breakthrough by Rhie et al., through an intensive analysis of telomeres (essentially a protective “cap” or sequence of base to prevent chromosomal degradation), not only did this new model essentially double the number of base pairs we considered, but further corrected and fixed previous errors found in GRCh38 (2023). We now know that our former knowledge of the Y chromosome was incomplete.

However, despite the strong biological and genetic techniques used, it would be entirely fair to claim that this achievement would not have been possible without the recent advancements in machine learning and artificial intelligence.

Perhaps the best way to represent the relationship between the biomedical techniques utilized in this study and the critical artificial intelligence needed to fully decode the Y chromosome is best summed up by stating that if this telomeric analysis retrieved and found

groups of base pairs, artificial intelligence, and machine learning interpreted them. For example, given a batch of numerous sequences, how can we predict which sequences are stable, and which are more comparatively unstable and will likely form alternative structures not commonly found in nature? This distinction is not only valuable to have a more complete genomic data set but can have significant clinical ramifications as well, especially in gene therapy. The team of researchers knew the answer to this question. Machine learning. This study used a machine learning algorithm to predict relative stability and quantified stability with a value referred to as the Quadron Score that allowed them to differentiate structures that could form different sequences and those that could not, offering a more accurate and thorough representation of the Y chromosome (Rhie et al., 2023). Many more examples, such as using machine learning to weed out false positives and repeats in structures, have unraveled this difficult chromosome.

Ultimately, this breakthrough in fully sequencing the Y chromosome will cause a significant ripple effect, opening new opportunities for research, treatments, and studies for decades to come. And in this historic, landmark moment of science, to many, one message is made clear.

Machine learning and artificial intelligence is the future of biomedicine.

References:

Hallast, P., Ebert, P., Loftus, M., Yilmaz, F., Audano, P. A., Logsdon, G. A., Bonder, M. J., Zhou, W., Höps, W., Kim, K., Li, C., Hoyt, S. J., Dishuck, P. C., Porubský, D., Tsetsos, F., Kwon, J. Y., Zhu, Q., Munson, K. M., Hasenfeld, P., . . . Lee, C. (2023). Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature*, *621*(7978), 355–364. https://doi.org/10.1038/s41586-023-06425-6

NIH researchers generate complete human X chromosome sequence. (2020, July 23). National Institutes of Health (NIH). https://www.nih.gov/news-events/news-releases/nih-researchers-generate-complete-human-x-chromosome-sequence

Rhie, A., Nurk, S., Čechová, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N., Chin, C., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., . . . Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978), 344–354. https://doi.org/10.1038/s41586-023-06457-y

World Bank Open Data. (n.d.). World Bank Open Data. https://data.worldbank.org/indicator/SP.POP.TOTL.MA.ZS