EXPLORING THE POTENTIAL OF GENETIC VARIATION AND ZYGOSITY IN DNA LANGUAGE MODELS

Ali Saadat, Jacques Fellay
School of Life Sciences
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{ali.saadat, jacques.fellay}@epfl.ch

ABSTRACT

Advancements in DNA language models (DNA-LMs) have improved phenotype prediction from DNA sequences, yet the roles of zygosity and genetic variation (GV) remain underexplored. In this study we quantify their effects on gene expression prediction as an example of variation-sensitive phenotype, showing that baseline models benefit from zygosity- and GV-aware encoding, while DNA-LMs struggle to utilize them. These findings underscore the need for integrating biologically meaningful features like zygosity and GV in DNA-LM pretraining to better capture genetic diversity and improve variant interpretation.

1 INTRODUCTION

Advances in DNA language models (DNA-LMs) have revolutionized the interpretation of genetic sequences (Benegas et al., 2025; Sanabria et al., 2024), leveraging natural language processing (NLP) architectures to predict features such as gene expression and variant effects (Avsec et al., 2021; Benegas et al., 2023). Pretrained models like the Nucleotide-Transformer (NT) (Dalla-Torre et al., 2024) use masked token prediction to capture sequence patterns but often overlook two critical factors: genetic variation (GV) and zygosity.

GV reflects genetic diversity, while zygosity describes allele variation at a locus. In diploid organisms, alleles can be homozygous (identical) or heterozygous (different), influencing dominant and recessive traits (Wilkie, 2018; Saadat & Fellay, 2024b). Current DNA-LMs, primarily trained on reference genomes, fail to account for population-level diversity and individual zygosity, limiting their performance in variation-sensitive tasks. While simpler models can encode zygosity effectively, transformers (Vaswani et al., 2017) lack explicit mechanisms to incorporate this information.

This study investigates the impact of GV and zygosity on gene expression prediction, an example of variation-sensitive phenotypes, using DNA-LMs and baseline convolutional neural networks (CNNs) (O'Shea & Nash, 2015). We demonstrate that CNNs benefit from GV- and zygosity-aware encodings, whereas DNA-LMs require further optimization to fully leverage this information.

By evaluating the roles of GV and zygosity in training and inference, we propose strategies to enhance DNA-LM pretraining for better capturing genetic diversity. These advancements aim to improve DNA-LM performance in phenotype prediction and deepen our understanding of how genetic variation shapes biological outcomes.

2 METHODS

2.1 DATA

We used the Geuvadis dataset(Lappalainen et al., 2013), which includes phased whole-genome sequencing (WGS) and log-scaled gene expression data from lymphoblastoid cell lines (LCLs) of 455 individuals. The human reference genome (hg38), lacking variation and zygosity information, served as a baseline, with matched expression data generated by averaging gene expression values across individuals.

2.2 STUDY DESIGN

To assess the impact of GV and zygosity, we designed a task to predict gene expression from DNA sequences. Using 1,000 base pair sequences centered on transcription start sites (TSS) for all protein-coding genes, models were trained on Geuvadis samples with individual-specific gene expression as outputs. For the reference genome, input sequences were paired with averaged gene expression values (Figure 1).

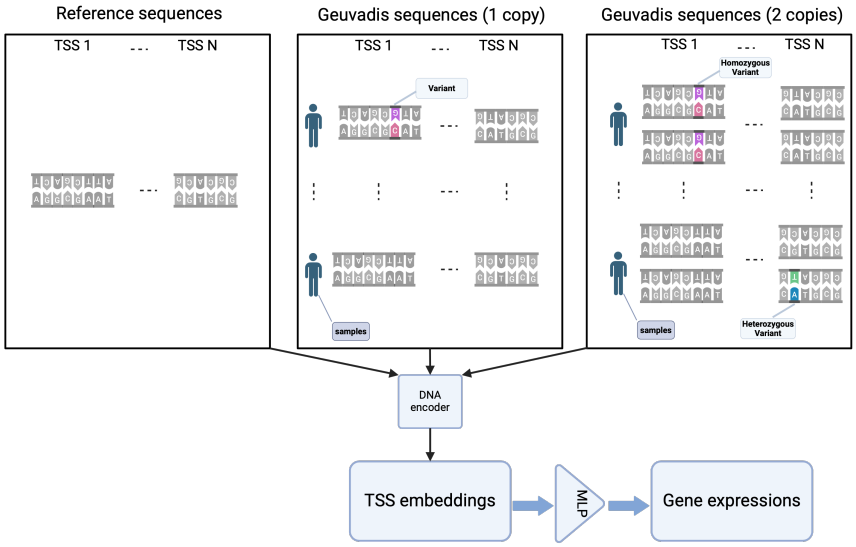


Figure 1: Study overview: 1000 bp DNA sequences around TSS are embedded using a DNA encoder, then used to predict average (reference) or individual (Geuvadis) gene expression via a feedforward neural network.

2.3 DNA ENCODERS

We employed three DNA encoders: a CNN without pretraining, the Nucleotide-Transformer trained on the reference genome (NTREF), and the Nucleotide-Transformer trained on diverse genomes from the 1000 Genomes Project (NT1000G). The CNN architecture consisted of three 1D convolutional layers with filter sizes of 4, 8, and 12, each using a kernel size of 5 and padding of 2. Both NTREF and NT1000G models included 500 million parameters.

For the CNN, reference or single-copy Geuvadis genomes were one-hot encoded, while two-copy Geuvadis genomes with zygosity information used additive genotype encoding, representing heterozygous alleles as 1 and homozygous as 2 (Figure 2).

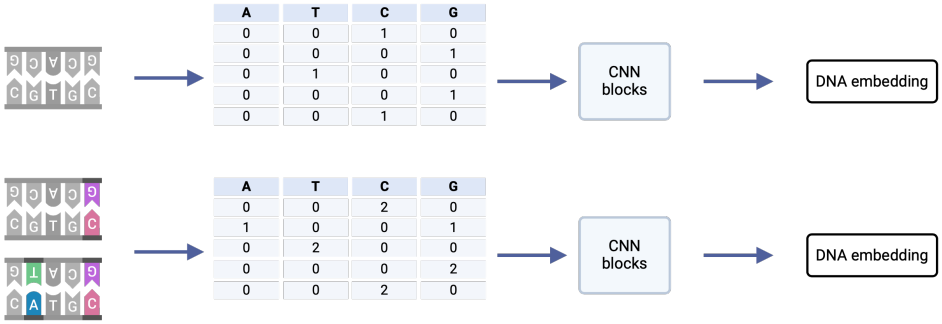


Figure 2: To train the CNN models, DNA sequences are first encoded using either one-hot encoding (when using reference or 1 copy of Geuvadis genomes) or additive genotype encoding (when using 2 DNA copies).

For NTREF and NT1000G, DNA sequences were passed through pretrained models to generate embeddings. Zygosity was incorporated by concatenating embeddings of positive and negative DNA strands. (Figure 3).

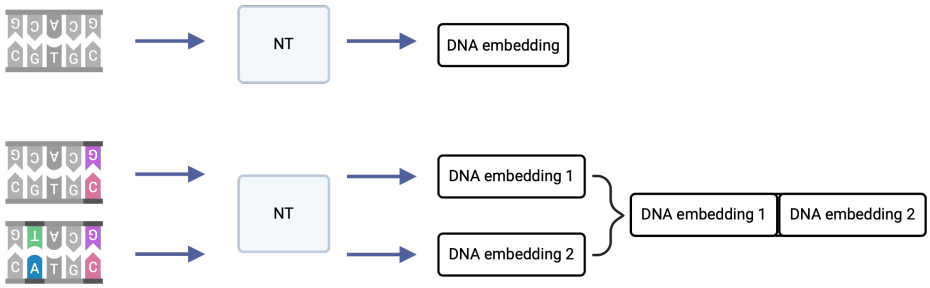


Figure 3: For the NT-based, we passed the DNA sequences through the respective pretrained models to generate DNA embeddings. When using 2 DNA copies, we concatenated the embeddings.

2.4 Model training and evaluation

We split the data by chromosomes: chromosome 8 for testing, chromosome 9 for validation, and the rest for training. The model used a multilayer perceptron (MLP) with a 64-unit hidden layer, trained for up to 100 epochs with early stopping to prevent overfitting. Prediction accuracy was evaluated using the Pearson correlation between predicted and observed gene expression levels.

3 Results

3.1 Including GV and zygosity information during training

Figure 4 presents the Pearson correlation of all models on the unseen test set. CNN performance improved when trained on sequences with one DNA copy (containing GV) or two DNA copies (including GV and zygosity). In contrast, NT-based models showed the opposite trend: incorporating GV or zygosity during training reduced performance compared to training on reference sequences. Additionally, NTREF models outperformed NT1000G models, consistent with findings from previous studies (Dalla-Torre et al., 2024).

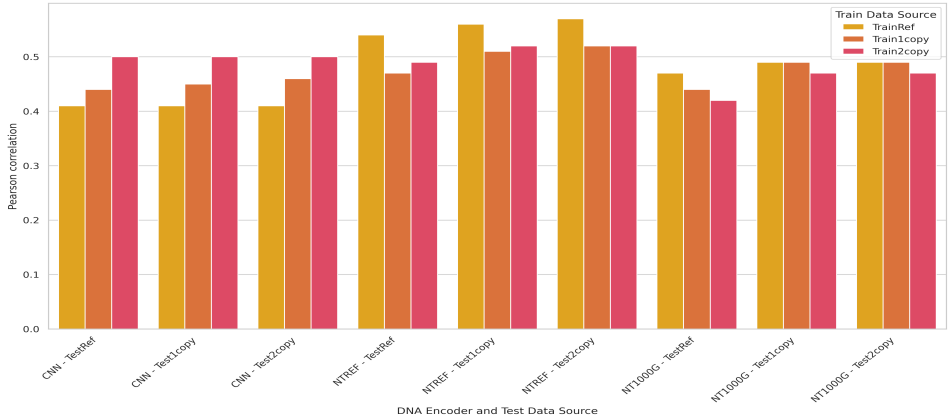


Figure 4: Pearson correlation coefficients of various models on the unseen test set. The X-axis represents the models and the test sequences used, while colors indicate the source of the training sequences.

3.2 Including GV and zygosity information during inference

Figure S1 illustrates model performance when incorporating GV or zygosity information during inference. For both CNN- and NT-based models, including GV or zygosity consistently improved performance compared to using reference sequences alone.

4 Discussion

Our study highlights the importance of GV and zygosity in DNA sequence models, comparing baseline CNNs with pretrained DNA-LMs. CNNs effectively utilized zygosity through additive genotype encoding, where heterozygous alleles are encoded as 1 and homozygous as 2. This simple encoding improved gene expression predictions, showing that simple models can benefit from GV- and zygosity-aware features. We observed that CNNs can match the performance of DNA-LMs, especially when zygosity information was included, suggesting simpler models can compete in specific tasks with well-designed inputs.

We observed that DNA-LMs pretrained on reference genomes (NTREF) consistently outperformed those trained on diverse genomes (NT1000G), likely because pretraining tasks like masked token prediction emphasize reference patterns over genetic variation. We also found that DNA-LMs performed better during inference when variation and zygosity were included, despite not leveraging these features during training. This reveals the untapped potential of DNA-LMs to handle variation-sensitive tasks when properly optimized.

Future improvements should focus on pretraining tasks that explicitly incorporate GV, zygosity, and population allele frequencies to enable DNA-LMs to better capture genetic diversity. Such advancements could greatly enhance their performance in predicting variation-sensitive phenotypes, paving the way for more robust genomics applications (Consens et al., 2025; Saadat & Fellay, 2024a; 2025).

References

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, October 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL http://dx.doi.org/10.1038/s41592-021-01252-x.

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44), October 2023. ISSN 1091-6490. doi: 10.1073/pnas.2311219120. URL http://dx.doi.org/10.1073/pnas.2311219120.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S. Song. Genomic language models: opportunities and challenges. *Trends in Genetics*, January 2025. ISSN 0168-9525. doi: 10.1016/j.tig.2024.11.013. URL http://dx.doi.org/10.1016/j.tig.2024.11.013.

Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. Transformers and genome language models. *Nature Machine Intelligence*, 7(3):346–362, March 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01007-9. URL http://dx.doi.org/10.1038/s42256-025-01007-9.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, November 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z. URL http://dx.doi.org/10.1038/s41592-024-02523-z.

Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013. ISSN 1476-4687. doi: 10.1038/nature12531. URL http://dx.doi.org/10.1038/nature12531.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. URL https://arxiv.org/abs/1511.08458.

Ali Saadat and Jacques Fellay. Dna language model and interpretable graph neural network identify genes and pathways involved in rare diseases. In *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pp. 103–115. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.langmol-1.13. URL http://dx.doi.org/10.18653/v1/2024.langmol-1.13.

Ali Saadat and Jacques Fellay. Proteome-wide prediction of mode of inheritance and molecular mechanism underlying genetic diseases using structural interactomics, 2024b. URL https://arxiv.org/abs/2410.17708.

Ali Saadat and Jacques Fellay. From mutation to degradation: Predicting nonsense-mediated decay with nmdep, 2025. URL https://arxiv.org/abs/2502.14547.

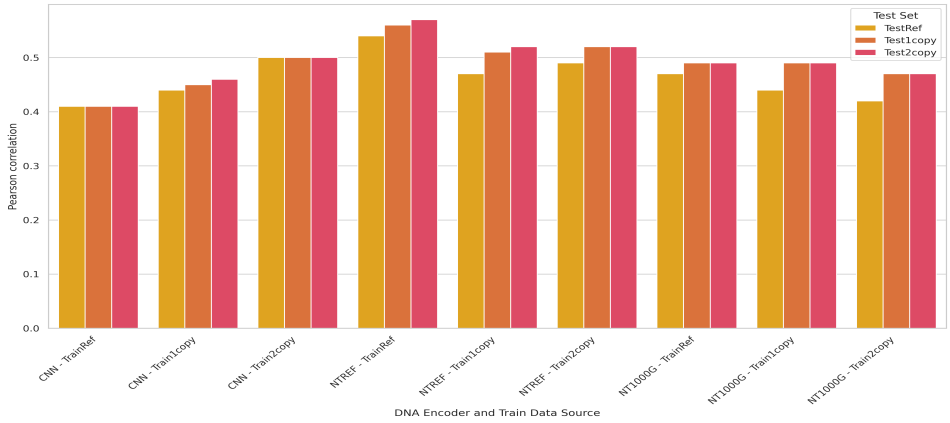
Melissa Sanabria, Jonas Hirsch, Pierre M. Joubert, and Anna R. Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, July 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00872-0. URL http://dx.doi.org/10.1038/s42256-024-00872-0.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Andrew OM Wilkie. Dominance and recessivity, April 2018. URL http://dx.doi.org/10.1002/9780470015902.a0005475.pub2.

A Appendix

A.1 Supplementary figures



Supplementary Figure S1: Pearson correlation coefficients of different models on the unseen test set. The X-axis represents the models and their training sequences, while colors indicate the source of the testing sequences.