Large Language Models for Genomics

Bernardo P. de Almeida, Thomas Pierrot

InstaDeep

b.dealmeida@instadeep.com, t.pierrot@instadeep.com

Abstract

In the era of ChatGPT, we hear the words Language Models, Large Language Models, Foundation Models everywhere. This trend has already arrived to genomics, existing now several DNA foundation models that can be easily adapted to solve a diverse set of prediction tasks. While it is clear now that supervised learning and Deep Learning strongly impacted the field in the past years and expanded significantly our in-silico capabilities, it looks like we are entering a new paradigm shift with such language models that promise to expand our capabilities also here in genomics. Thus it is important to engage the community in a joint discussion about what these models actually are and can already do, their weaknesses, and what the next generations of models could look like. This opinion paper is a starting point for this discussion.

Decoding the Human Genome

The human genome sequence provides the underlying code for human biology. Despite 20 years of study since it was first sequenced, our understanding of how the genome sequence encodes the cellular programs and development of different cell types, as well as the role of genetic variants in disease is far from complete. Models that can "read" the genome of each individual and predict the different regulatory layers and cellular processes across cell types hold the promise to better understand, prevent and treat diseases, and to achieve personalized medicine based on each individual's genetic profile (Gunter and Green 2023).

Recent advances in genomics research and sequencing technologies have resulted in an explosion of molecular data. This wealth of data has enabled comprehensive characterisations of the functional activity of the genome across hundreds of human and mouse cell types, as well as for other species (ENCODE 2012; Roadmap Epigenomics 2015). The confluence of such large-scale datasets, coupled with significant advancements in deep learning and large language models (LLMs), is bringing us closer to the long-sought aspiration of unraveling the language of molecular biology.

A New Paradigm with Deep Learning

Machine learning algorithms are designed to automatically detect patterns in data and have early been applied to genomics large-scale data (Libbrecht and Noble 2015). However, the performance of such methods is highly dependent on the quality of features extracted from the data and has shown limited capability. A transformative paradigm shift has emerged in recent years with the advent of deep neural networks, obviating the necessity for predefined features by embedding the computation of features into the automated learning process. The application of deep learning in genomics has heralded a revolutionary transformation in the way we analyze the human genome. Since their first applications in genomics in 2015 that deep neural networks (mostly based on CNNs) have been applied to predict diverse molecular phenotypes and have emerged as the leading type of predictive models in genomics (Eraslan et al. 2019; Yue et al. 2023). Through the interpretation of the features learned by such models, new biological features have also been discovered about each of the individual gene regulation layers (Avsec et al. 2021b; de Almeida et al. 2022; Janssens et al. 2022; Linder et al. 2022; Agarwal and Kelley 2022).

However, some problems such as predicting the expression of every gene in different cell types from sequence alone remain unsolved. This is a challenging task as gene expression is influenced by regulatory elements located very far from the target gene, and it involves multiple layers of gene regulation such as transcription, splicing, termination/polyadenylation, and RNA stability. Enformer (Avsec et al. 2021a) achieved a good improvement in gene expression prediction over previous models (Kelley et al. 2018; Zhou et al. 2018) by combining convolutional and transformer layers to effectively integrate information from up to 100 kb away in the genome. However, Enformer is still limited at capturing the causal effects of distal regulatory elements on expression and their sequence features (Karollus, Mauermeier, and Gagneur 2023) or the effect of personal variants in gene expression (Sasse et al. 2023). The limited performance in this and other tasks is most probable due to the small quantity of available regulatory and gene expression data, and could benefit from leveraging existent unlabeled genomic data such as the genomes sequenced of human cohorts.

From Supervised to Self-Supervised learning

Labeled data scarcity remains one of the major challenges in most scientific fields and is even more present in genomics given the time and cost associated with performing such assays. However, the development of modern sequencing techniques in association with their continuously decreasing cost has led to a significant increase in the amount of raw genome data, with genomes available for thousands of species and individuals (The 1000 Genomes Project 2015). This motivates the need to exploit such data to build better models.

Self-supervised learning imposed itself as the solution in NLP to face a similar challenge, and subsequently in other fields such as computer vision and proteomics (Ericsson et al. 2022). A common strategy to learn from unlabeled sequential and discrete data, where we would typically refer to each discrete entity as a token - think about words in English or Nucleotides for DNA - is to pre-train the model to reconstruct masked tokens in a sequence or predict the next token from left to right. These training strategies, often applied with Transformers, were shown to learn general representations of the input data that can be then leveraged to solve supervised tasks even in low-data regimes. These models are often referred to as foundation - as they build general knowledge about the data they are trained on. They are also commonly referred to as language models (LM) - or often even large language models (LLMs) as performance was shown to increase with the model size - as if one assumes tokens to be the base constituent of a language, then these models implement a model of such language (Naveed et al. 2023).

Genomics experienced the emergence of such models very recently (Ji et al. 2021; Zhou et al. 2023; Nguyen et al. 2023; Benegas, Batra, and Song 2023; Dalla-Torre et al. 2023; Fishman et al. 2023), a bit later than proteomics for instance where these models are now more mature (Jumper et al. 2021; Baek et al. 2021; Lin et al. 2023). Typically, such models consider single nucleotides or *k*-mers, akin to DNA "words", with *k* ranging from 3 to 6 as tokens. Some works also used learnable encoding techniques, such as byte pair encoding, to build a more complex "vocabulary". An interesting discussion to have here is whether DNA - that is often referred to as the "language of life" - should be considered as a language at the same extent than Human languages, where nucleotide tokens can be considered as words and chunks of genomes as sentences, or whether this tokenization is a simple computational trick to apply the latest self-supervised technique on that type of data where token sizes and complexity are just adapted to optimize the trade-off between sequence length and vocabulary size without having actually any "true meaning". While we highlight the fact that it wouldn't be the first time that one would anthropomorphize models and over-interpret what they actually are, we leave this discussion to the reader. What is for sure true is that these models are building strong representations of DNA data despite any supervision.

DNABERT (Ji et al. 2021) was the first published model of that sort, BERT referring to that training technique that predicts masked tokens. It was trained on the reference human genome, using windows of 512bp. The authors showed that after its self-supervised training - that we refer to as pre-training - the model can be quickly finetuned to predict promoters, splice sites and transcription factor binding sites, highlighting the versatility of the learned representations. A few months later, the Nucleotide Transformer (NT) paper (Dalla-Torre et al. 2023) was released with the first attempt to build a systematic study and benchmark to build foundation models for DNA, testing models of varying sizes up to 2.5B parameters and pre-trained on diverse genomes from different individuals and species using also a BERT protocole. The authors also showed that during pre-training such models learn to focus their attention on crucial genomic elements and sequence motifs. Both DNABERT and the NT released a v2 of their models (Zhou et al. 2023; Dalla-Torre et al. 2023) that are smaller, thus faster, while showing higher performance on the downstream tasks by leveraging the multi-species dataset introduced in the original NT study and multiple implementation tricks coming from the recent NLP literature.

Additional models pre-trained on DNA sequences include GPN, GENA-LM and Genslms (Benegas, Batra, and Song 2023; Fishman et al. 2023; Zvyagin et al. 2022), while for RNA sequences there are the SpliceBERT (Chen et al. 2023) and BigRNA (Celaj et al. 2023). GPN-MSA was also introduced recently to leverage aligned sequences from related species (MSA) to improve performance for prediction of deleterious coding and non-coding variants in a zero-shot manner (Benegas et al. 2023).

What these models can/cannot already do?

We now reached a stage where multiple strong DNA foundation models have been developed, trained on up to 6 trillion nucleotides and were evaluated on dozens of classification and regression genomics tasks, where the NT-v2 seems to be dominating after fine-tuning (Dalla-Torre et al. 2023). A point worth mentioning is that the scaling laws observed in these recent model developments are very encouraging for the field. It was demonstrated that dozens of billions of parameters are required in NLP for the models to actually express their potential (Hoffmann et al. 2022; Touvron et al. 2023a,b). In computer vision however, recent works showed that only a few billions parameters maximum are enough (Zhai et al. 2022). Similarly, in genomics, it seems that a few hundreds of millions might be the sweet spot, which implies that such models can be developed much faster and at a lower computational cost (Dalla-Torre et al. 2023).

However, such models still fail to significantly outperform specialized methods when a large amount of labeled data is available. Good examples would be works such as DeepSEA/Sei for chromatin predictions (Zhou and Troyanskaya 2015; Chen et al. 2022), SpliceAI for splice site prediction (Jaganathan et al. 2019) or DeepSTARR for enhancer activity prediction (de Almeida et al. 2022). Models such as NT-v2 were shown to be able to reach only similar performance. These tasks have in common that they exhibit datasets much larger than the other typical downstream tasks where these foundation models excel. Think about hundreds of thousands / millions of labeled inputs versus only thousands / dozen of thousands of inputs. We believe that the key to an improved performance also for high-data tasks is to be able to achieve transfer between them. We discuss this point in the next section.

Current foundation models remain also limited by the size of the nucleotide sequences they can process. The NT-v2 models can handle sequences of length 12kb (Dalla-Torre et al. 2023), going up to 36kb for the GENA-LM models (Fishman et al. 2023). This limitation originates from the quadratic scaling with the input sequence length of the memory and compute time required for Transformer models. This hinders the application of such models to more relevant tasks in the genomics community such as gene expression prediction or precise and accurate variant effect prediction which require modeling long-range effects that can happen up to hundreds of thousands of base pairs away. Recently, a new collection of Foundation models dumped HyenaDNA were developed to tackle this challenge (Nguyen et al. 2023). These models rely on a novel architecture that leverages fourier transforms to avoid the quadratic scaling. However, such models haven't demonstrated their capabilities to solve complex tasks over very large sequences of nucleotides yet. In addition, recent works showed that their performance in short-range tasks decreases when the training input size increases (Dalla-Torre et al. 2023). As such, there are still many opportunities to develop the next generation of models that can make accurate predictions from full genes and maybe chromosomes, including their regulatory elements. As the field of NLP observed recently the appearance of many techniques that managed to successfully increase drastically the input length of Transformer models to process hundreds of thousands of tokens at the same time (Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020; Ding et al. 2023; Guu et al. 2020), we are hopeful that the field of genomics will experience a similar trend over the next few years.

Another key aspect to be discussed over the next few years is the interpretability of these models. As this topic is bigger than genomics, we leave that conversation for future work.

What will the next generations look like?

DNA foundation models are now at a stage similar to the one NLP was in 2018, before the "T5 moment" (Raffel et al. 2020). Models can produce accurate representations but they lack transfer capabilities. In NLP this was achieved through sequence-to-sequence models and formulating all tasks as text-to-text tasks. This consequently led to the "GPT moment" (Brown et al. 2020), where researchers observed that scaling these text-to-text models allowed new capabilities to emerge such as in zero-shot settings, which means solving new tasks without showing examples. Current DNA foundation models are incapable of such things. While several models showed that computing distances over the produced sequence representations can correlate well with interesting effects such as mutation deleteriousness (Cheng et al. 2023; Dalla-Torre et al. 2023; Benegas et al. 2023; Benegas, Batra, and Song 2023) - calling this zero-shot as well - they fail to produce actual transfer between tasks. In other words, current models can exploit all the raw genome data available to become stronger but they fail later to connect labeled tasks as well as to achieve transfer and emergent capabilities. This is mainly due to the current structure of these models. As framing all tasks as text-to-text solved this issue in NLP (Raffel et al. 2020), it is therefore very tempting to do the same in genomics.

However, DNA as such is not a general enough medium to do so. While all the tasks of interest in NLP can be expressed with English, all the tasks of interest in genomics cannot be expressed solely with nucleotides. Think about producing gene annotations or predicting gene expression using only a nucleotide-based alphabet. To achieve this, current approaches will need to expand their vocabulary - whatever this means, being a human interpretable one or not - to unify all tasks of interest within the same framework. This vocabulary should also be an opportunity to provide the models more context such as the cell or tissue type or the position of the DNA sequence within the genome. This will allow to blend in the same models not only all sequenced genomes but also all existing annotations and assays that have been performed - such as RNA-seq, ATAC-seq and others. This prospect is very exciting and we foresee it will be the key to unlocking completely new capabilities and genomics insights with these types of models.

Conclusions and Future Perspectives

Deciphering the language of biology, how to go from DNA to gene expression and to proteins, and linking it to human health is closer than ever. Foundation models might be the new paradigm that we were missing to help make sense of all existent labeled and unlabeled data to create more general models with emerging capabilities. Taking advantage of the large-scale biobanks containing genetic and health information of thousands of individuals (Halldorsson et al. 2022), these models will help associate genomic variant information and other molecular data with human health information.

To achieve that we will need to rethink existing foundation models over two main axes: (1) increasing their reception field up to at least 1M nucleotides to process full genes with their regulatory elements. These models should also be able to focus their attention at both low and large scale to solve tasks on varying lengths of sequencing, from detecting single-nucleotide genetic variants to predicting gene expression; (2) enabling transfer between tasks through building sequence-to-sequence models as well as a unified vocabulary that encompasses DNA and all needing quantities to express the tasks of interest. Only through such changes can these models keep improving and integrating all the labeled data available to develop the capabilities required to aid in our understanding of biology and medicine.

We believe that these developments should be in line with the main goals in the field for the next years: (1) more accurate prediction of the different molecular profiles, and in particular gene expression, from sequence; (2) understand the mechanisms and causality; (3) predict the impact of genetic variants on function and disease, improving polygenic risk scores and genetic diagnosis; (4) the design of synthetic sequences with specific regulatory properties (Taskiran 2022), with important implications for cell and gene therapy; (5) ultimately providing new tools and therapeutics for personalized medicine. These are challenging goals and will require close collaboration between biologists/geneticists and AI/-computer scientists.

References

Agarwal, V.; and Kelley, D. R. 2022. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biology*, 23: 245.

Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; and Kelley, D. R. 2021a. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203.

Avsec, Ž.; Weilert, M.; Shrikumar, A.; Krueger, S.; Alexandari, A.; Dalal, K.; Fropf, R.; McAnany, C.; Gagneur, J.; Kundaje, A.; et al. 2021b. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3): 354–366.

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; and Baker, D. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557): 871–876.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Benegas, G.; Albors, C.; Aw, A. J.; Ye, C.; and Song, Y. S. 2023. GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction. *bioRxiv*.

Benegas, G.; Batra, S. S.; and Song, Y. S. 2023. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44): e2311219120.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Celaj, A.; Gao, A. J.; Lau, T. T.; Holgersen, E. M.; Lo, A.; Lodaya, V.; Cole, C. B.; Denroche, R. E.; Spickett, C.; Wagih, O.; Pinheiro, P. O.; Vora, P.; Mohammadi-Shemirani, P.; Chan, S.; Nussbaum, Z.; Zhang, X.; Zhu, H.; Ramamurthy, E.; Kanuparthi, B.; Iacocca, M.; Ly, D.; Kron, K.; Verby, M.; Cheung-Ong, K.; Shalev, Z.; Vaz, B.; Bhargava, S.; Yusuf, F.; Samuel, S.; Alibai, S.; Baghestani, Z.; He, X.; Krastel, K.; Oladapo, O.; Mohan, A.; Shanavas, A.; Bugno, M.; Bogojeski, J.; Schmitges, F.; Kim, C.; Grant, S.; Jayaraman, R.; Masud, T.; Deshwar, A.; Gandhi, S.; and Frey, B. J. 2023. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv*.

Chen, K.; Zhou, Y.; Ding, M.; Wang, Y.; Ren, Z.; and Yang, Y. 2023. Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. *bioRxiv*.

Chen, K. M.; Wong, A. K.; Troyanskaya, O. G.; and Zhou, J. 2022. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7): 940–949.

Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L. H.; Zielinski, M.; Sargeant, T.; Schneider, R. G.; Senior, A. W.; Jumper, J.; Hassabis, D.; Kohli, P.; and Žiga Avsec. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664): eadg7492.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; Richard, G.; Skwark, M.; Beguir, K.; Lopez, M.; and Pierrot, T. 2023. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*.

de Almeida, B. P.; Reiter, F.; Pagani, M.; and Stark, A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, 54(5): 613–624.

Ding, J.; Ma, S.; Dong, L.; Zhang, X.; Huang, S.; Wang, W.; Zheng, N.; and Wei, F. 2023. LongNet: Scaling Transformers to 1,000,000,000 Tokens. arXiv:2307.02486.

ENCODE. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74.

Eraslan, G.; Žiga Avsec; Gagneur, J.; and Theis, F. J. 2019. Deep learning: new computational modelling techniques for genomics. *Nature Review Genetics*, 20: 389–403.

Ericsson, L.; Gouk, H.; Loy, C. C.; and Hospedales, T. M. 2022. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3): 42–62.

Fishman, V.; Kuratov, Y.; Petrov, M.; Shmelev, A.; Shepelin, D.; Chekanov, N.; Kardymon, O.; and Burtsev, M. 2023. GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences. *bioRxiv*, 2023–06.

Gunter, C.; and Green, E. D. 2023. To boldly go: Unpacking the NHGRI's bold predictions for human genomics by 2030. *The American Journal of Human Genetics*, 110(11): 1829–1831.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv:2002.08909.

Halldorsson, B. V.; Eggertsson, H. P.; Moore, K. H. S.; Hauswedell, H.; Eiriksson, O.; Ulfarsson, M. O.; Palsson, G.; Hardarson, M. T.; Oddsson, A.; Jensson, B. O.; Kristmundsdottir, S.; Sigurpalsdottir, B. D.; Stefansson, O. A.; Doruk Beyter, G. H.; Tragante, V.; Gylfason, A.; Olason, P. I.; Zink, F.; Asgeirsdottir, M.; Sverrisson, S. T.; Sigurdsson, B.; Gudjonsson, S. A.; Sigurdsson, G. T.; Halldorsson, G. H.; Sveinbjornsson, G.; Norland, K.; Styrkarsdottir, U.; Magnusdottir, D. N.; Snorradottir, S.; Kristinsson, K.; Sobech, E.; Jonsson, H.; Geirsson, A. J.; Olafsson, I.;

Jonsson, P.; Pedersen, O. B.; Erikstrup, C.; Brunak, S.; Ostrowski, S. R.; Consortium, D. G.; Thorleifsson, G.; Jonsson, F.; Melsted, P.; Jonsdottir, I.; Rafnar, T.; Holm, H.; Stefansson, H.; Saemundsdottir, J.; Gudbjartsson, D. F.; Magnusson, O. T.; Masson, G.; Thorsteinsdottir, U.; Helgason, A.; Jonsson, H.; Sulem, P.; and Stefansson, K. 2022. The sequences of 150,119 genomes in the UK Biobank. *Nature*, 607(7920): 732–740.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jaganathan, K.; Panagiotopoulou, S. K.; McRae, J. F.; Darbandi, S. F.; Knowles, D.; Li, Y. I.; Kosmicki, J. A.; Arbelaez, J.; Cui, W.; Schwartz, G. B.; et al. 2019. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3): 535–548.

Janssens, J.; Aibar, S.; Taskiran, I. I.; Ismail, J. N.; Gomez, A. E.; Aughey, G.; Spanier, K. I.; Rop, F. V. D.; González-Blas, C. B.; Dionne, M.; Grimes, K.; Quan, X. J.; Papasokrati, D.; Hulselmans, G.; Makhzami, S.; Waegeneer, M. D.; Christiaens, V.; Southall, T.; and Aerts, S. 2022. Decoding gene regulation in the fly brain. *Nature*, 601(7894): 630–636.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Karollus, A.; Mauermeier, T.; and Gagneur, J. 2023. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1): 56.

Kelley, D. R.; Reshef, Y. A.; Bileschi, M.; Belanger, D.; McLean, C. Y.; and Snoek, J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genomics Research*, 28: 739–750.

Libbrecht, M. W.; and Noble, W. S. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6): 321–332.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; and Rives, A. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Linder, J.; Koplik, S. E.; Kundaje, A.; and Seelig, G. 2022. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biology*, 23: 232.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2023. A Comprehensive Overview of Large Language Models. arXiv:2307.06435.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Roadmap Epigenomics. 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539): 317–330.

Sasse, A.; Ng, B.; Spiro, A.; Tasaki, S.; Bennett, D. A.; Gaiteri, C.; Jager, P. L. D.; Chikina, M.; and Mostafavi, S. 2023. How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks? *bioRxiv*.

The 1000 Genomes Project. 2015. A global reference for human genetic variation. *Nature*, 526(7414): 68–74.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yue, T.; Wang, Y.; Zhang, L.; Gu, C.; Xue, H.; Wang, W.; Lyu, Q.; and Dun, Y. 2023. Deep Learning for Genomics: From Early Neural Nets to Modern Large Language Models. *International Journal of Molecular Sciences*, 24(21): 15858.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33: 17283–17297.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12104–12113.

Zhou, J.; Theesfeld, C. L.; Yao, K.; Chen, K. M.; Wong, A. K.; and Troyanskaya, O. G. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8): 1171–1179.

Zhou, J.; and Troyanskaya, O. G. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10): 931–934.

Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; and Liu, H. 2023. DNABERT-2: Efficient Foundation Model and

Benchmark For Multi-Species Genome. *arXiv preprint arXiv:2306.15006*.

Zvyagin, M.; Brace, A.; Hippe, K.; Deng, Y.; Zhang, B.; Bohorquez, C. O.; Clyde, A.; Kale, B.; Perez-Rivera, D.; Ma, H.; Mann, C. M.; Irvin, M.; Pauloski, J. G.; Ward, L.; Hayot-Sasson, V.; Emani, M.; Foreman, S.; Xie, Z.; Lin, D.; Shukla, M.; Nie, W.; Romero, J.; Dallago, C.; Vahdat, A.; Xiao, C.; Gibbs, T.; Foster, I.; Davis, J. J.; Papka, M. E.; Brettin, T.; Stevens, R.; Anandkumar, A.; Vishwanath, V.; and Ramanathan, A. 2022. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *bioRxiv*.