THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 12 Issue 7

REVIEW ARTICLE

Advances in Human Genome Resolution: The Role of Pan-Genomic Strategies and Fine-Tuning Pre-trained Genomic Models

Duo Du[1], Yupeng Zhang[2], Fan Zhong[1*], Lei Liu[1,3*]

[1]School of Basic Medical Sciences and Intelligent Medicine Institute, Fudan University, Shanghai 200032, China.
[2]Institutes of Biomedical Sciences, Fudan University, Shanghai, 200032, China.
[3]Shanghai Institute of Stem Cell Research and Clinical Translation, Shanghai 200120, China.

OPEN ACCESS

PUBLISHED
31 July 2024

CITATION
Du, D., Zhang, Y., et al., 2024. Advances in Human Genome Resolution: The Role of Pan-Genomic Strategies and Fine-Tuning Pre-trained Genomic Models. Medical Research Archives, [online] 12(7). https://doi.org/10.18103/mra.v12i7.5571

DOI
https://doi.org/10.18103/mra.v12i7.5571

ISSN
2375-1924

ABSTRACT

The groundbreaking theory of DNA double helix structure has greatly promoted the development of molecular genetics, shaping and refining the genetic central dogma, thus enabling researchers to explore genotype-phenotype regulation at different levels. In particular, with the continued advancement of third-generation sequencing technology, an increasing number of highly accurate human genomes have been assembled, such as T2T-CHM13 and HG002. These high-quality genome sequences not only provide a more comprehensive human reference sequence, but also enable functional genomics studies within a unified coordinate system. To better explore and resolve the complex genetic information encompassed within human genome sequences, scientists have proposed novel research strategies, involving graphical pan-genome and pre-trained genomic models. The graphical pan-genomes provide population-level high-quality references, revealing the genomic diversity within populations and exploring the sequence complexity of specific regions, such as the KIR immune region. Concurrently, related studies of pre-trained models within the human genome offer new perspectives for interpreting sequence functions and delving into the hidden genetic codes, potentially leading to complete DNA decoding. Overall, graphical pan-genome and pre-trained genomic models represent two crucial strategies in genomics research, which will provide more new insights and make greater breakthroughs in the human genome. Together, these approaches have deepened our understanding of the human genome, fostered the development of bioinformatics ecosystems, and will contribute to the establishment and improvement of the entire field. Therefore, this review focuses on DNA sequencing, human genome assembly, high-quality pan-genome and pre-trained genomic large language models (LLMs), highlighting and summarizing the latest achievements and progress in human genome research, discussing existing challenges and providing future perspectives.

Introduction

The rapid advancements in fields such as molecular biology have significantly contributed to reveal the complex phenotypic regulation within the cellular. Miescher's extraction of nucleic acids in 1869 marked the beginning of a deeper understanding of life's molecular makeup, particularly with the establishment of the DNA double helix model, which has unraveled the mystery of nucleic acid as the primary genetic landscape[1,2]. In eukaryotes, the genetic blueprint primarily comprises nuclear genomic DNA, organellar DNA and cytoplasmic free DNA[3]. The nuclear genome, which is primarily located within the cellular, serves as the main genetic material of the living organisms and is involved in a variety of complex interactions during phenotypic regulation[4,5,6]. Studies have shown that genetic factors not only directly influence individual phenotypes, but also regulate gene expression through mechanisms such as DNA methylation, histone modifications and non-coding RNAs[7] [8]. For this reason, geneticists typically regard the genome sequences as genotypes, with their downstream regulatory expressions and functions are categorized as potential phenotypes. These can be further classified into intermediate molecular phenotypes (RNAs, proteins and metabolites), characteristic phenotypes (influenced by genetic and environmental factors, such as skin color) and clinical phenotypes (physiological and biochemical traits, disease symptoms and responses). A deeper understanding of the genotype-phenotype interplay reveals the complexity of living organisms, offering novel approaches for disease diagnosis, treatment and prevention. Consequently, the resolution and functional interpretation of high-quality genome sequences are particularly crucial for exploring their role in phenotypic shaping.

Subsequently, with the development of sequencing technologies, the Human Genome Project (HGP) was successfully implemented, such as hg38 reference, facilitating the decoding of genetic information and its involvement in the process of phenotypic regulation within living organisms[9]. The genome functional annotations revealed that its sequence contains both coding and non-coding regions, with ~50% of the gene regions and only ~3% consisting of the protein-coding sequences[10]. Notably, most published studies have predominantly focused on protein-coding sequence functionality. In view of this, genomics-related studies can reveal the complexity and regulation of the genetic landscape, including sequence resolution, sequence diversity, gene interactions, non-coding DNA regulation, genome stability and the high-dimensional structure of chromatin. Ultimately these findings will not only deepen our understanding of the genetic material, but also provide novel insights into complex diseases and species evolution, driving advancements in the biomedical field. In recent years, the iterative updates of high-quality human reference genomes have significantly advanced analysis workflows, with sequence alignment-based methods serving as the primary strategy. Although these widely used linear reference sequences can provide a unified reference coordinate system for bioinformatics analysis, their limitations include inadequate representation of population diversity and limited interpretation[11,12].

The third-generation sequencing technologies have enabled high-quality genomic sequence analysis at the population level, giving rise to pan-genomics as a powerful tool for exploring the genetic information contained within these sequences and enhancing the human reference genome[13]. This provides base-level resolution among populations for a deeper understanding of genetic diversity. Concurrently, the maturation of LLMs technology in natural language processing (NLP) has provided innovative methods to further decipher the genetic information hidden in genomic sequences, facilitating more efficient identification and interpretation of the regulatory mechanisms within genomic regions[14,15,16]. Additionally, the further integration of multi-omics data promises a more comprehensive understanding of genomic regulation in organismal phenotypes. Therefore,

this review will start from the nuclear genome sequencing and sequence assembly, and then further summarize the research progress in graphical pan-genome and pre-trained genomic LLMs in human genome research. Subsequently, it will discuss and envision these developments, with the ultimate goal of strengthening the entire computational biology ecosystem, and will provide a solid foundation for the study of personalized precision medicine and complex diseases.

DNA extraction and sequencing techniques

The isolation and accurate assembly of DNA from cellular is crucial in genome research. Over time, scientists have developed innovative techniques, such as magnetic bead-based and silica membrane-based extraction, which have greatly improved the efficiency and quality of DNA extraction[17,18]. Concurrently, advances in sequencing technologies have facilitated rapid and accurate reconstruction of DNA sequences from the microcellular world. The development of sequencing has progressed from traditional Sanger sequencing to third-generation full-length sequencing (Table 1), gradually achieving high-quality long-read sequencing at the whole-genome level[19,20]. Compared to other sequencing technologies, the hallmark of third-generation sequencing is that it can directly obtain continuous sequences from individual DNA molecules up to tens of thousands of bases[21]. This effectively addresses the challenge posed by complex structures and repeated sequences within genomes. With its reduced cost and improved accuracy, third-generation sequencing is expected to play an increasingly significant role in biomedical research. Moreover, alongside those established sequencing technologies (yielding HiFi reads up to ~30kb), high-throughput sequencing has experienced a surge, particularly in single-cell genome and spatial genomics sequencing. These advancements have enhanced our understanding of cellular heterogeneity and the spatio-temporal specificity of genomic DNA, uncovering cell-specific diversity[22,23]. By integrating diverse sequencing methods, researchers can delve into fundamental life phenomena at different levels, with the insights gained playing a key role in genomics, functional genomics, and cancer genomics.

Table 1: Characterization and comparison of the three main sequencing technologies

	first-generation sequencing	second-generation sequencing	third-generation sequencing
Other names	Sanger sequencing	NGS sequencing	Long-read sequencing
Core principles	Uses labelled dideoxynucleotide triphosphates (ddNTPs); selectively terminates DNA strand extension for sequencing, etc.	Parallel sequencing on a fixed surface; Illumina Bridge Amplification, Roche 454 Emulsion PCR and Ion Torrent Semiconductor technologies, etc.	Real-time single-molecule sequencing technology; nanopore sequencing technology, etc.
Advantages	Highly accurate and reliable; can sequence specific sequence fragments	High throughput; low cost; rapid; low sample requirements and broad applicability	Long-read sequencing; spanning repetitive regions and improving sequence assembly accuracy; identifying DNA epigenetic modifications
Limitations	Low throughput; complex and time-consuming, etc.	Short reads; PCR amplification bias; GC preference in sequencing, etc.	High error rates (HiFi accuracy is high); higher costs; high data storage processing and analysis requirements, etc.
Applications	PCR sequencing commonly used in laboratories; Human Genome Project; genetic disease diagnostics and new drug development, etc.	Large-scale genome sequencing; precision medicine research, etc.	Repeat sequence and structural variation analysis; resolution of low-coverage regions in second-generation sequencing; more accurate gene expression and splicing variation analysis, etc.

Human genome sequence resolution

High-quality resolution of genome sequences in diploid organisms, like humans, is a crucial basic step in understanding their genetic properties. Prior to genome sequence assembly, researchers need to sequence using different libraries and various platforms according to the experimental design, etc. The whole sequencing process includes[24,25,26]: 1) sample preparation, where DNA is extracted from individuals and quality-tested to meet the sequencing platform requirements. 2) library construction, tailored to platform requirements, often fragmenting DNA for second-generation sequencing (adding sequencing adaptors), while third-generation sequencing libraries filter out shorter fragments. 3) sample sequencing, which involves sequencing under specific conditions, such as whole genome or targeted sequencing, mixed-sample analysis or not, and whether the initial DNA requires amplification. 4) data processing, which includes quality control steps such as removing adapters and filtering out low-quality sequences to maintain high quality, as poor quality can significantly affect assembly results.

Genome sequence assembly is conventionally classified into three main assembly levels: primary, pseudo-haplotype and complete haplotype[27,28,29,30,31]. These different levels of sequence resolution are closely related to the sequencing platforms used and the specific research requirements. Briefly, primary assembly is usually the starting point for genome assembly, reconstructing a complete sequence in diploid organisms. However, its inability to accurately distinguish haplotype sequences can potentially lead to loss or misinterpretation of haplotype-specific pathogenic information. Pseudo-haplotype assembly improves on primary assembly by incorporating alternative sequences, aiming to increase resolution and avoiding sequences loss in complex regions as much as possible. Lastly, fully haplotype assembly resolves both parental genetic sequence information with high accuracy, which is crucial for understanding genetic diversity and regulatory mechanisms in complex diseases.

Although the genome assemblies of most species currently remain at the primary and pseudo-haplotype assembly levels, advances in technology will eventually enable more species to have complete haplotype genome sequences[32]. The reference genome hg38, derived from the HGP, provides a critical foundation for understanding human genetic diseases, advancing personalized medicine and developing new drugs, as well as fostering the growth of disciplines such as bioinformatics. The T2T-CHM13, a high-quality homozygous cell line reference genome, which is not only represents an important milestone in genomics field, but also provides a new direction for future research[33]. Meanwhile, fully haplotype-resolved sequences can be resolved by using family lineage information or by combining different long-read sequencing platforms such as Illumina, PacBio, and Oxford Nanopore combined via the Verkko process[34]. Moreover, single platform PacBio HiFi data assembled by hifiasm can also be used to cost-effectively generate high-quality local haplotype sequences[30]. NextPolish2 can further improve assembled sequences quality by reducing overcorrection and haplotype switching errors using HiFi data[35]. However, there is currently no perfect single process currently, and iterative validation with multi-platform data is often required, but there are existing strategies that can well detect repetitive sequences and structural variants (SVs) in genomes[36,37].

High-quality pan-genome reference

The rapid growth of high-quality haplotype genomes has led to a boom in pan-genomic research, presenting researchers are confronted with the challenge of integrating and interpreting vast amounts of genetic data. As a significant field in genomics, pan-genome research aims to comprehensively analyze the genomes of all individuals within a specific species, overcoming the limitations of single reference genomes and providing novel insights into exploring the genetic

complexity within species[13,38]. In particular, the human pan-genome project, by constructing a high-accuracy coordinate system that includes diverse ethnicities, which greatly increased the accuracy and completeness of research. This achievement provides a foundation for disease research, drug development, and precision medicine, providing a more nuanced understanding of population genetics. The high-accuracy pan-genome sequence enables more effective personalized medicine, particularly in the diagnosis and treatment of rare diseases[39, 40]. Meanwhile, the advancement of pan-genomic projects is driving advances in computational ecology and transforming the omics analysis workflow, including the quantification of gene expression, haplotypic expression patterns analysis, and SVs detection in sequencing data[41,42].

Pan-genome analysis originated from microbial research and has since been widely applied in plant and animal fields[43]. This has facilitated our understanding of intraspecific genetic diversity and sequence evolutionary properties. Scientists have explored and proposed various pangenome construction methods, and the most primary strategies include linear and graphical methods, each with specific advantages and applicable scenarios[44, 45, 46]. The construction of a linear pan-genome is similar to the classical genome structure, whereby newly identified sequences are appended to an existing reference. Therefore, the primary task is to identify sample-specific non-reference sequences[47]. Typically, each sample undergoes *de novo* assembly and is aligned to the reference genome to extract long fragments of distinct sequences. Subsequently, these sequences are then combined with the reference genome to construct the pan-genome after de-redundancy[48]. Meanwhile, to reduce the cost of large-scale sample assembly, an alternative approach is to first alignment and identify sample-specific sequencing data (poorly aligned or unaligned reads), which are then assembled and merged with the reference genome after redundancy removal[49]. Given the properties of third-generation sequencing data (Table 1), it is also feasible to directly use sample-specific sequence reads to construct a linear pan-genome after de-redundancy, particularly suitable for populations with small genetic differences. However, the resulting linear pan-genome, which contains only a few large differential sequences is not sufficiently accurate. It lacks single-base resolution, may contain more potential assembly errors or inconsistencies in coordinate assignments among samples.

The graphical pan-genome construction represents a relatively novel strategy for the visualization of population genetic diversity through a structural graph containing all possible sequence variants[50, 51].These methods are more suitable for sequence alignment and variant resolution at the single-base genome level in genetically diverse populations. Minigraph-Cactus, a reference-based method, efficiently handles complex variants in large datasets by constructing a sequence variation graph in which differential bases are represented as nodes connected by edges, forming paths that represent complete sequences, clearly demonstrating genome associations and differences at the single-base level resolution[52]. PGGB is another "all to all" alignment tool among long sequences, generating accurate genome graphs representing the genetic diversity of a population within a single graph[53]. Compared to the other methods, Minigraph-Cactus and PGGB are currently the most mature process, yet they still face challenges in accurately representing and constructing sequences in repetitive regions[54]. As computational power and sequencing technologies continue to advance, graphical pan-genomes will facilitate a deeper understanding of genomic diversity in large-scale studies, being expected to play a significant role in interpreting biodiversity and underlying genetic mechanisms.

Currently, the Human Pangenome Project is primarily promoted by the Human Pangenome Reference Consortium (HPRC) and the Chinese Pangenome Consortium (CPC). The HPRC project selected diverse samples from different ethnicities,

geographical locations and populations for multi-platform sequencing, and ultimately constructed an initial comprehensive human pan-genome database, which was then subjected to rigorous data cleaning and complex assembly processes[55]. This resource provides valuable high-quality sequence for studying human genetic diseases and developing personalized nucleic acid vaccines. The CPC focuses on filling the gap in genetic information for East Asians by pinpointing specific ethnic and regional genomic data, thereby deepening our understanding of Chinese population genetics[56]. The principal challenges and future directions of current pan-genome research include[57,58]: 1) Technological innovation to continuously improve sequencing technologies and analytical methods to improve the integrity of the genome. 2) Extensive international collaboration and data sharing. 3) Integration of interdisciplinary domain knowledge to interpret sequence functions. 4) Translation of research findings into clinical practice for disease risk assessment, drug development and rare disease treatment. 5) Further refinement of bioinformatics tools to better understand the genetic basis of genome evolution and phenotypic diversity in populations.

Pre-trained genomic large language models

The pre-trained LLMs have revolutionized NLP research, and their impact has recently extended to the genomics field, with notable achievements. These models can be classified into two categories: Transformer-based models such as DNABERT[59], LOGO[60], Nucleotide Transformer[61], GenSLM[62], GENA-LM[63] and GenomicLLM[64], as well as those based on other language modeling frameworks like HyenaDNA[65] and Mamba[66], etc. Compared to other language framework models, Transformer-based models are currently only able to understand the context of up to 4k tokens (~0.0013% of the human genome) due to the limitation of the secondary scaling using the attention mechanisms, leading to a limited understanding of long-range genome interactions. As for Transformer-based models, DNABERT is the first BERT-based pre-trained model of DNA sequences. It outperforms CNN, RNN, and LSTM networks in modeling DNA language, capturing global information and transferring it to various downstream tasks. The optimal 6-mers were discovered in experiments and adopted by subsequent researchers. DNABERT2, an improved version of DNABERT, which overcomes length limitations by using linear bias attention instead of learned positional embedding, reducing time and memory consumption while improving performance[67]. In addition, it also employs Byte Pair Encoding (BPE) strategy for DNA sequence analysis, overcoming the limitations of *K*-mers tokenization, which in turn benefits from the computational efficiency of non-overlapping tokenization. LOGO, a lightweight genome ALBERT model using reference genome hg19, excels in sequence labeling tasks such as promoter identification, enhancer-promoter interaction prediction, and chromatin state prediction, etc. Nucleotide Transformer, utilizing different nucleic acid databases and *K*-mers tokenization for cross-species representation learning, improves accurate molecular phenotype prediction even in resource-constrained scenarios, bridging the gap between genetic information and observable traits. GENA-LM, employing BPE tokenization for input lengths up to 3kb, is pretrained on T2T-CHM13 genome sequences, demonstrating strong performance in various genomic downstream tasks. GenSLMs, trained on 110 million prokaryotic genome sequences and fine-tuned on 1.5 million SARS-CoV-2 DNA sequences, which can facilitate accurately and rapidly variant detection, accelerating the identification of novel COVID-19 variants. GenomicLLM, using a nano-LLaMA2 network, enables better understanding of mixed corpora containing sequence and non-sequence inputs (textual information from human gene annotation), enabling a wider range of applications including classification, regression and generation tasks.

In terms of other language framework models, researchers have introduced the HyenaDNA

genomics model, a modification of the implicit convolutional large language model Hyena on the genome task, which rivals attention mechanisms in its ability to process long contexts and offers lower time complexity. This model is pre-trained on sequences up to 1 million tokens at single nucleotide resolution, which significantly addressing the limitations of context length and single nucleotide resolution in existing genomic models, achieving state-of-the-art performance across multiple genomic tasks. Building upon this progress, scientists recently have developed the Evo foundational model using the StripedHyena architecture, which integrates attention mechanisms with data-driven convolution operators. This model encompassing DNA, RNA and protein sequences, effectively simulates the genetic central dogma and demonstrates capabilities in both prediction and generation tasks spanning molecular to genomic scales[68]. Additionally, the Mamba model, which is based on linear time series modeling with selective state spaces, exhibits superior pre-training quality and downstream performance compared to Hyena and Transformers, and exhibits improved performance as context length increases. Despite the high computational costs, pre-trained genomic models, trained on large amounts of unlabeled DNA sequences, can be directly fine-tuned for specific downstream tasks.

The emerging field of pre-trained DNA models faces several limitations and challenges in their application to downstream tasks. These problems stem from the scarcity of high-quality datasets, specialized model frameworks, and task-specific biomedical corpora with meaningful significance. Advances in sequencing technologies, particularly third-generation long-read sequencing, are increasing the availability of high-quality genomic data. However, there is a need to incorporate more sophisticated model frameworks, such as human feedback reinforcement learning[69], to facilitate biologists' optimization of models, and lightweight models that minimize pre-training costs should be prioritized. There is a need for a biologically meaningful corpus consisting of a pre-training phase and model evaluation, with DNA corpora constructed by researchers using biomedical knowledge can improve a model's comprehension of the human genome. Furthermore, the development of high-quality datasets linking genomes to human phenotypic traits is crucial for assessing the generalization capabilities of pre-trained DNA models, a field that remains underdeveloped.

Prospects for pan-genomic large language models

Artificial intelligence (AI), symbolizing "silicon-based" life, has made significant progress in LLMs, accelerating the development of artificial general intelligence (AGI). Notably, the deployment of AI generated content (AIGC), such as ChatGPT, is having a profound impact on our lifestyles (Figure 1A). In contrast, as representatives of "carbon-based" life, biological intelligent agents interact with their environment and maintain corresponding homeostasis over long-term species evolution, collectively forming intelligent systems of living organisms. The genomic DNA, located in the cell nucleus, acts as a central regulatory mechanism, utilizing sophisticated strategies like the genetic central dogma to maintain the functioning of the entire intelligent system (Figure 1B).

Genomic DNA, the fundamental component of biological intelligence systems, has a complex sequence, a sophisticated genetic blueprint and a regulatory network. Over the years, human scientists have focused primarily on only protein coding sequences, but the results of these studies remain limited. Approximately 97% of the genome sequence may harbors hidden "dark matter," and the structure and sequence of DNA may exhibit spatiotemporal specificity within the nucleus. To accelerate the exploration of this genomic "dark matter" as well as truly decipher and interpret the genetic blueprint (the Life 2.0 era, with the completion of the HGP marking the end of Life 1.0), current AI foundation models (entering the AI

2.0 era, with AI 1.0 marked by computer simulations of human intelligence) can be used to accelerate this process. The organic combination between human scientists' exploratory mindset (employing global reasoning from local insights, etc.) and computational intelligence (utilizing recursion-based divide-and-conquer strategies, etc.) will facilitate the deciphering of this genetic blueprint through approaches such as reinforcement learning (agents), and ultimately push human civilization into higher dimensions in a secure manner, as marked by the synthesis of independent life forms. Throughout this process, we should be guided by these questions --"Who are we? Where do we come from? Where are we going? How can we survive better?" and explore the origins and endpoints of life using the 'human in the loop' strategy.

At the same time, the intelligence of AI is often being assessed by human standards, but this assessment typically reflects the distant associations derived from the extensive instant storage capacity, model parameters, and associated databases, rather than the essence of intelligence. Although scientists commonly compare neurons to neural networks, the inherent limitations and biases in the understanding of the human brain make it an imperfect reference for AI development. Furthermore, most real-world data is artificial and carries a human subjective bias. Conversely, genomic DNA, as the fundamental genetic material of living organisms (the code left behind by the 'gods'[70]), plays an objective role in shaping the phenotype of life. High-quality genome sequence data, which is both clean and informative, provides fertile ground for AI innovation. By directly simulating or generating these "godlike" codes, we may uncover a novel approach to AI, potentially mitigating human bias in scientific inquiry. Directly simulating and generating codes left by "god" with AI itself is a method of generating intelligence. While the natural cognition of human scientists may be lost due to the subjective bias, the codes left by "god" will persist with the continuation of "carbon-based" life. In future research, we will build on the concept that fully decoding the DNA genetic blueprint could be a way to accelerate the realization of AGI, with the aim of contributing to AI 2.0 and ultimately paving the way towards genuine AI.

Overall, the accumulated human pan-genome DNA datasets represent an excellent application scenario for AI 2.0, which play a central role in unravelling the essence of biological intelligence and driving advancements in genuine AI technologies. Meanwhile, improving the understanding of genomic DNA data (spatio-temporal specificity) and enabling models to learn this knowledge is crucial for the whole process, which requires the collaboration of experts from various fields including health data scientists, AI specialists, bioinformaticians, evolutionary biologists, physicists, mathematicians, and natural philosophers. This field is currently booming with recent advancements. For instance, the Evo model can generate multimodal life language sequences (DNA, RNA, and proteins) within genomes. Recently, a novel "Embed-Search-Align" framework for DNA sequence alignment has been developed that is comparable in performance to traditional algorithms such as Bowtie and BWA-Mem[71]. Besides, DNA sequence language models have also been used to predict genome-wide variant effects, cis-regulatory regions, DNA-protein interaction, DNA methylation and splice sites, etc[72].

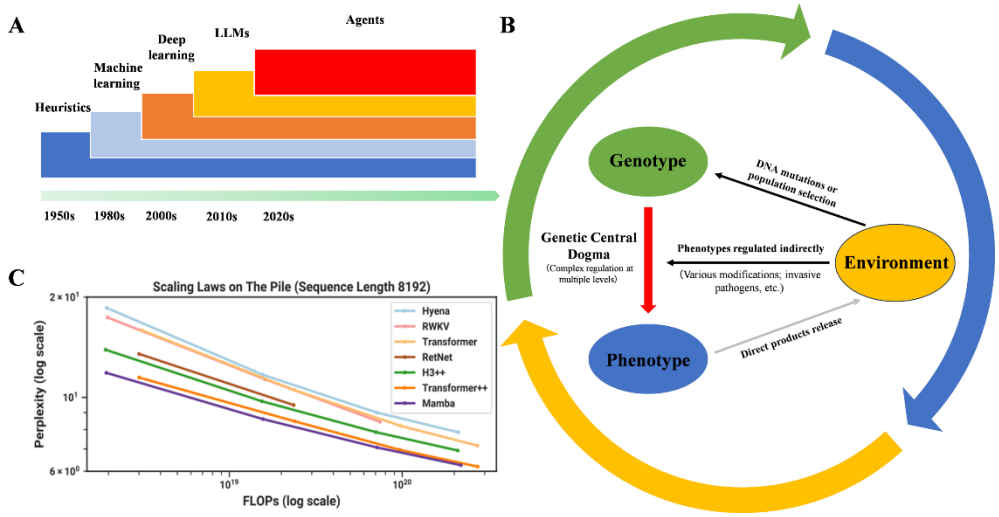

Figure 1 Artificial Intelligence and Biological Intelligence Systems
(A) The evolution of artificial intelligence technology. (B) The concept of lifeform intelligent system. (C) Comparison of the new energy of different large model frameworks[66].

Limitations

As the core content of bioinformatics analysis, human genome sequence research has always attracted much attention. In order to construct a comprehensive omics analysis workflow, this review primarily focuses on two aspects: human pan-genomes and the fine-tuning of pre-trained genomics LLMs. It focuses primarily on the analysis and interpretation at the genome sequence level, but there is a notable absence of discussion on phenotype-related topics. Furthermore, given the current limited availability of haplotype genome datasets, there is a lack of discussion of potential biases in downstream analysis related to human population genetics, e.g., the pre-trained of existing linear genomes may have bias and thus affect the downstream analysis task. In addition, the pan-genomic studies' content did not include a comparative analysis of their findings against human reference sequences, and the fine-tuning component of the pre-trained genomic models lacked a systematic comparison with other deep learning models. In short, this paper provides a preliminary overview of the human pan-genome and the fine-tuning of pre-trained genomic LLMs, with many details still to be discussed.

Conclusion

The resolution and functional interpretation of the human genome sequence is central to biomedical research, so this study reviews the relevant aspects from the following three aspects: 1) Sequencing technologies and human genome assembly, we provide an overview of sequencing technology development and its comparison, comparing different levels of human genome sequence resolution methods at the present stage, and summarized the different strategies and their sequence resolution effects. 2) For pan-genome construction, we outline the pan-genome construction methods, particularly focusing on the mainstream graphical pan-genome construction strategies and the current progress of this project. 3) For human genome sequence pre-training, we classify them according to whether they follow the Transformer framework or not, and summarize the latest progress and advantages in the corresponding domains. In conclusion, this review provides a detailed discussion and outlook for the above researches, with the expectation that these methods will be more effective in mining and deciphering the genetic blueprint of the genome. Meanwhile, we also point out the direction of genome-NLP or genome-LLM using pan-genome, which will further explore new research paradigms. In particular, the adoption of a multi-species pre-training strategy aims to exploit evolutionarily conserved genomic information, thereby improving the modelling of its underlying grammar[73] and ultimately contribute to the realization of AGI.

Conflict of Interest Statement:

The authors report no conflicts of interest pertaining to this work.

Funding Statement:

This work was supported by the Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai.

Acknowledgements:

None.

References:

1. Dahm R. Friedrich Miescher and the discovery of DNA. *Dev Biol* 278, 274-288 (2005).

2. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737-738 (1953).

3. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5, 123-135 (2004).

4. Mefford HC. Genotype to phenotype-discovery and characterization of novel genomic disorders in a "genotype-first" era. *Genet Med* 11, 836-842 (2009).

5. Orgogozo V, Morizot B, Martin A. The differential view of genotype-phenotype relationships. *Front Genet* 6, 179 (2015).

6. Raben TG, Lello L, Widen E, Hsu SDH. From Genotype to Phenotype: Polygenic Prediction of Complex Human Traits. *Methods Mol Biol* 2467, 421-446 (2022).

7. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7, 29-59 (2006).

8. Komili S, Farny NG, Roth FP, Silver PA. Functional specificity among ribosomal proteins regulates gene expression. *Cell* 131, 557-571 (2007).

9. Hood L, Rowen L. The Human Genome Project: big science transforms biology and medicine. *Genome Med* 5, 79 (2013).

10. Hatje K, Muhlhausen S, Simm D, Kollmar M. The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *Bioessays* 41, e1900066 (2019).

11. An assembly line for an improved human reference genome. *Nature*, (2022).

12. O'Leary K. Diversifying the 'reference' genome. *Nat Med* 29, 2972 (2023).

13. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nat Rev Genet* 21, 243-254 (2020).

14. Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J. Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol* 25, 83 (2024).

15. Naveed H, *et al.* A Comprehensive Overview of Large Language Models. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv230706435N (2023).

16. Tang L. Large models for genomics. *Nat Methods* 20, 1868 (2023).

17. Ayoib A, Hashim U, Gopinath SCB, Md Arshad MK. DNA extraction on bio-chip: history and preeminence over conventional and solid-phase extraction methods. *Appl Microbiol Biotechnol* 101, 8077-8088 (2017).

18. Kloten V, *et al.* Liquid biopsy in colon cancer: comparison of different circulating DNA extraction systems following absolute quantification of KRAS mutations using Intplex allele-specific PCR. *Oncotarget* 8, 86253-86263 (2017).

19. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet* 34, 666-681 (2018).

20. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17, 333-351 (2016).

21. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 21, 30 (2020).

22. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 24, 695-713 (2023).

23. Park J, *et al.* Spatial omics technologies at multimodal and single cell/subcellular level. *Genome Biol* 23, 256 (2022).

24. Hess JF, *et al.* Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol Adv* 41, 107537 (2020).

25. Ekblom R, Wolf JB. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl* 7, 1026-1042 (2014).

26. Pasquali F, *et al.* Application of different DNA extraction procedures, library preparation protocols and sequencing platforms: impact on

sequencing results. *Heliyon* 5, e02745 (2019).

27. Li H, Durbin R. Genome assembly in the telomere-to-telomere era. *ArXiv*, (2023).

28. Dominguez Del Angel V, *et al.* Ten steps to get started in Genome Assembly and Annotation. *F1000Res* 7, (2018).

29. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36, 2896-2898 (2020).

30. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18, 170-175 (2021).

31. DeRaad DA, *et al.* De novo assembly of a chromosome-level reference genome for the California Scrub-Jay, Aphelocoma californica. *J Hered* 114, 669-680 (2023).

32. Rhie A, *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737-746 (2021).

33. Nurk S, *et al.* The complete sequence of a human genome. *Science* 376, 44-53 (2022).

34. Rautiainen M, *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* 41, 1474-1482 (2023).

35. Hu J, Wang Z, Liang F, Liu S-L, Ye K, Wang D-P. NextPolish2: A Repeat-aware Polishing Tool for Genomes Assembled Using HiFi Long Reads. *Genomics, Proteomics & Bioinformatics*, (2024).

36. Jung H, *et al.* Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Comput Biol* 16, e1008325 (2020).

37. Zhang L, Zhou X, Weng Z, Sidow A. De novo diploid genome assembly for genome-wide structural variant detection. *NAR Genom Bioinform* 2, lqz018 (2020).

38. Singh V, Pandey S, Bhardwaj A. From the reference human genome to human pangenome: Premise, promise and challenge. *Front Genet* 13, 1042550 (2022).

39. Cohen ASA, *et al.* Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* 24, 1336-1348 (2022).

40. Groza C, *et al.* Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun* 15, 657 (2024).

41. Sibbesen JA, *et al.* Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat Methods* 20, 239-247 (2023).

42. Zheng Z, *et al.* A sequence-aware merger of genomic structural variations at population scale. *Nat Commun* 15, 960 (2024).

43. Liu K, *et al.* Pan-Genome Analysis of TIFY Gene Family and Functional Analysis of CsTIFY Genes in Cucumber. *Int J Mol Sci* 25, (2023).

44. Eisenstein M. Every base everywhere all at once: pangenomics comes of age. *Nature* 616, 618-620 (2023).

45. Tao Y, Zhao X, Mace E, Henry R, Jordan D. Exploring and Exploiting Pan-genomics for Crop Improvement. *Mol Plant* 12, 156-169 (2019).

46. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. *Nat Plants* 6, 914-920 (2020).

47. Li R, *et al.* Recovery of non-reference sequences missing from the human reference genome. *BMC Genomics* 20, 746 (2019).

48. Duan Z, *et al.* HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol* 20, 149 (2019).

49. Sherman RM, *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* 51, 30-35 (2019).

50. Liu Y, Tian Z. From one linear genome to a graph-based pan-genome: a new era for genomics. *Sci China Life Sci* 63, 1938-1941 (2020).

51. Outten J, Warren A. Methods and Developments in Graphical Pangenomics. *J Indian Inst Sci* 101, 485-498 (2021).

52. Hickey G, *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol*, (2023).

53. Garrison E, *et al.* Building pangenome graphs. *bioRxiv*, (2023).

54. Andreace F, Lechat P, Dufresne Y, Chikhi R. Comparing methods for constructing and representing human pangenome graphs. *Genome Biol* 24, 274 (2023).

55. Liao WW, *et al.* A draft human pangenome reference. *Nature* 617, 312-324 (2023).

56. Gao Y, *et al.* A pangenome reference of 36 Chinese populations. *Nature* 619, 112-121 (2023).

57. Abondio P, Cilli E, Luiselli D. Human Pangenomics: Promises and Challenges of a Distributed Genomic Reference. *Life (Basel)* 13, (2023).

58. Wang T, *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 604, 437-446 (2022).

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

60. Yang M, *et al.* Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res* 50, e81 (2022).

61. Dalla-Torre H, *et al.* The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, 2023.2001.2011.523679 (2023).

62. Zvyagin M, *et al.* GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *bioRxiv*, (2022).

63. Fishman V, *et al.* GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences. *bioRxiv*, 2023.2006.2012.544594 (2023).

64. Liu H, Zhou S, Chen P, Liu J, Huo K-G, Han L. Exploring Genomic Large Language Models: Bridging the Gap between Natural Language and Gene Sequences. *bioRxiv*, 2024.2002.2026.581496 (2024).

65. Nguyen E, *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv230615794N (2023).

66. Gu A, Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv231200752G (2023).

67. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv230615006Z (2023).

68. Nguyen E, *et al.* Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv*, 2024.2002.2027.582234 (2024).

69. Sun H. Reinforcement Learning in the Era of LLMs: What is Essential? What is needed? An RL Perspective on RLHF, Prompting, and Beyond. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv231006147S (2023).

70. Weigmann K. The code, the text and the language of God. When explaining science and its implications to the lay public, metaphors come in handy. But their indiscriminant use could also easily backfire. *EMBO Rep* 5, 116-118 (2004).

71. Holur P, *et al.* Embed-Search-Align: DNA Sequence Alignment using Transformer Models. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv230911087H (2023).

72. Liu J, Yang M, Yu Y, Xu H, Li K, Zhou X. Large language models in bioinformatics: applications and perspectives. Preprint at https://ui.adsabs.harvard.edu/abs/2024arXiv240104155L (2024).

73. Consens ME, *et al.* To Transformers and Beyond: Large Language Models for the Genome. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv231107621C (2023).

