Review

Large Language Model as a Promising Framework for the Complete Clinical Interpretation of Human Genetic Variants

Jihun Bhak [†], Dong-Hyun Shin [†], Jongbum Jeon [†], Soobok Joe, Yeonsu Jeon, Hyoungjin Choi, Yoonsung Kwon, Kyungwhan An, Yun Sung Cho, Sungwon Jeon, Haeyoung Jeong [*], Jong Bhak [*]

Posted Date: 11 May 2026

doi: 10.20944/preprints202605.0651.v1

Keywords: large language model; human genetic variants; pangenome; GWAS

 doi:10.20944/preprints202605.0651.v1

Review

Large Language Model as a Promising Framework for the Complete Clinical Interpretation of Human Genetic Variants

Jihun Bhak [1,2,†], Dong-Hyun Shin [1,2,†], Jongbum Jeon [3,†], Soobok Joe [3], Yeonsu Jeon [3], Hyoungjin Choi [1,2], Yoonsung Kwon [1,2], Kyungwhan An [1,2], Yun Sung Cho [4,5], Sungwon Jeon [6], Haeyoung Jeong [3,*] and Jong Bhak [1,2,6,*]

[1] Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[2] Department of Biomedical Engineering, College of Information-Bio Convergence Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[3] Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience & Biotechnology (KRIBB), 125, Gwahak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[4] CG Invites Co., LTD, Seoul, 07802, Republic of Korea

[5] Invites Genomics Co., LTD, Jeju, 63243, Republic of Korea

[6] AgingLab Inc, Ulsan 44919, Republic of Korea

* Correspondence: hyjeong@kribb.re.kr (H.J.); jongbhak@genomics.org (J.B.)

† These authors contributed equally to this work.

Abstract

The number of human genetic variants cataloged in dbSNP has plateaued since 2021, with over ~1.1 billion variants housed. Since the human pangenome reference has enabled the precise identification of even structurally complex variants, capturing the entire spectrum of human genetic variants is almost achievable. However, the clinical impacts of most genetic variants still remain elusive. This is due to limitations in genome-wide association study (GWAS), the standard framework for variant interpretation, which relies solely on statistical assumptions. GWAS cannot interpret low-frequency alleles and capture molecular interactions between variants, hindering its ability to explain complex traits and diseases. Recently, large language models (LLMs) enabled accurate inference of human genetic variants' pathogenicity even without requiring a large sample size or prior annotations by modeling the biological principles encoded within the genome. For instance, Evolutionary Scale Modeling (ESM1b) successfully predicted missense variants in ClinVar, achieving an auROC of up to 0.905. In addition, Evo 2 classified non-coding pathogenic variants in ClinVar with an auROC of 0.987 for single nucleotide variants (SNVs) and 0.971 for non-SNVs. These results suggest that although yet limited to pathogenicity prediction, integrating multiomic and clinical data through LLM will enable the complete clinical interpretation of human genetic variants.

Keywords: large language model; human genetic variants; pangenome; GWAS

Introduction

The ultimate goal of human genetics is to completely understand and efficiently treat diseases [1]. Today, genome assembly and variant discovery are no longer bottlenecks, due to rapid advances in human reference genome construction and sequencing technologies [2,3]. Yet, little progress has been made in genetic analytic methods since the introduction of GWAS. Thus, the greater challenge now lies in accurately deciphering the clinical impacts of genetic variants. However, ongoing large-cohort genome projects, including the UK Biobank [4–6], Trans-Omics for Precision Medicine Program (TOPMed) [7], All of Us Research Program [8–11], Korean Genome Project (KGP) [12,13],

 doi:10.20944/preprints202605.0651.v1

GenomeAsia 100k Project [14], Westlake Biobank for Chinese (WBBC) [15,16], 100,000 Genomes Project [17,18] and BioBigData.Korea (BIKO) still primarily focus on accumulating rather than interpreting whole-genome sequencing data and genetic variant callsets. Consequently, we continue to risk building therapies on insufficient knowledge, overlooking hidden causal variants, and exacerbating population biases in medicine. At this crucial time point, we critically assess the major achievements and persistent limitations in human genetics since the Human Genome Project (HGP) [19,20] and propose LLM as the ultimate framework to resolve these challenges.

Capturing the Full Spectrum of Human Genetic Variants Through the Graph-based Human Pangenome Reference

Constructing a Complete Gapless Human Genome Assembly is No More a Challenge

In 2003, HGP completed the initial human genome by sequencing ~99% of the human euchromatic genome with an accuracy of ~99.99% [19,20]. However, the initial assembly contained 341 gaps and tiling path errors, particularly in complex and repetitive regions [20,21]. To address these limitations, the Genome Reference Consortium (GRC) has been continuously updating the initial human genome, starting with the release of GRCh37 in 2009 [21,22]. This version corrected a few tiling path errors and reduced the number of gaps to 271 [21,23]. Further updates, utilizing methods such as optical mapping and Fluorescence In Situ Hybridization (FISH), led to the release of GRCh38 in 2013, which reduced the total gap length by ~33.31%, from ~239.9Mb (~7.65% of GRCh37 total length) to ~160Mb (~4.98% of GRCh38 total length) [22].

Despite the ongoing updates, GRCh38 has none of its chromosomes fully assembled thus far [24]. Its hundreds of gaps are primarily located in highly repetitive regions, such as pericentromeric areas, the short arms of acrocentric chromosomes, and regions with a high density of segmental duplications, which span from hundreds of kilobases to megabases [25]. These gaps remain mainly due to the inherent limitations of Bacterial Artificial Chromosome (BAC) clone libraries used in HGP [26,27]. The BAC libraries contained clones with insert sizes ranging from ~100 to ~200 kb [19], which were too short to cover the highly repetitive regions [20,26,27]. Furthermore, the BAC libraries were sourced from the diploid genomes of multiple individuals, exacerbating the difficulty of bridging the gaps in assembly process due to structural incompatibilities at their flanks [27]. These shortcomings led the initial human genome to be an incomplete pseudo haploid with a mosaic of haplotypes [19,22,27], which makes achieving the complete gapless assembly of the initial human genome a persisting challenge.

Due to these limitations, in 2022, the Telomere-to-Telomere (T2T) consortium published the first complete gapless human genome assembly (T2T-CHM13) by utilizing an alternative sample, the Complete Hydatidiform Mole 13 (CHM13) cell line [25,27,28]. The genome of CHM13 is homozygous for most alleles, since it is derived from a single sperm that has gone through post-meiotic chromosomal duplication [25]. Although this precluded the assembly of the Y chromosome and did not represent a heterozygous diploid genome, it dramatically simplified the assembly process by eliminating issues of structural incompatibilities and haplotype mosaicism [27]. Furthermore, recent advancements in long-read sequencing technologies, such as PacBio HiFi (up to ~25 kb with < 0.5% error) [29,30] and Oxford Nanopore Technologies' ultra-long read sequencing (N50 > 100 kb, occasionally exceeding 1 Mb) [31,32] enabled the resolution of megabase-scale repeats, making T2T-CHM13 the first complete gapless human genome assembly [27]. With these breakthroughs, complete gapless genome assembly is no longer a challenge, as even heterozygous diploid genomes of living individuals can now be fully constructed and phased by integrating long-range data, such as Hi-C or trio data [2,33,34].

 doi:10.20944/preprints202605.0651.v1

A Single Linear Human Reference Genome Cannot Capture the Entire Diversity of Genetic Variants Across the Global Population

For two decades, the initial human genome has been the reference genomic coordinate for calling and annotating variants, which are critical to understanding the genetic architecture of diseases and phenotypes [35–37]. However, its linear structure, representing only a single sequence, inevitably induces reference bias, where non-reference alleles are either overlooked or miscalled during read mapping and variant calling [38]. To address this limitation, GRCh38 incorporated 261 alternative (ALT) loci across 178 genomic regions, primarily in highly polymorphic regions such as the major histocompatibility complex (MHC) and killer-cell immunoglobulin-like receptor (KIR) loci, overlapping more than 1,120 unique genes [39,40]. This enabled accurate alignment of 23.71% of reads that were initially unmapped to the primary assembly [22]. Nevertheless, the ALT loci collectively span only ~62Mb, constituting ~2% of the human genome [40], thereby providing only partial mitigation of the reference bias.

As a result, GRCh38 remains incomplete in representing the full spectrum of human genetic diversity. Population-specific novel sequences missing in GRCh38 for ethnicities, including East Asians [41–46], Swedes [47,48], and Africans [49], account for up to ~10% (~300 Mb) of the human genome, overlapping 315 distinct protein-coding genes [50] (Table 1). Therefore, underrepresented populations had to develop ethnicity-specific reference genomes [33,34,42,44–47,49,51–58] (Table 2). These reference genomes enhanced the accuracy of variant calling for their respective populations. For instance, constructing a new reference genome by incorporating Swede-specific sequences into GRCh38 led to the correction of 134,851 false-called SNVs and the discovery of 72,157 novel SNVs [47]. Similarly, the number of variants uniquely called using the Korean consensus reference genome (KOREF_C) ranged from 644,075 to 729,834 per Korean individual (48.65% to 50.97% classified as novel based on dbSNP) [42,59], indicating the necessity of using ethnicity-specific reference genomes.

Table 1. Summary of Novel Sequences Identified in Ethnicity-specific Genomes Relative to the Human Reference Genome. GRCh38 was used as the reference genome for population-specific novel sequence identification, unless otherwise specified. a) Each value corresponds to one genome assembly. b) Compared with GRCh37. c) Each value corresponds to one haplotype.

Ethnicity	Number of Samples	Total Length of Ethnicity-specific Novel Sequences (Mb)	Number of Genes Predicted in Ethnicity-specific Novel Sequences	Publication Year	Reference
Egyptian	1 genome assembly + 110 sequencing data	~0.04	-	2020	[49]
African	910 sequencing data	~296	-	2018	[50]
Swede	1,000 sequencing data	~46	-	2019	[48]
	2 genome assemblies	~13.8/~10.6[a)]	-	2018	[47]
Mongolian	1 genome assembly	~26.9[b)]	24	2014	[56]

 doi:10.20944/preprints202605.0651.v1

Chinese	275 genome assemblies	~29.5	188	2019	[41]
	486 sequencing data	~276	53	2021	[43]
	1 haplotype-resolved genome assembly	~12.9/~13.4[c]	246/135[c]	2022	[44]
	1 genome assembly	~12.8	-	2016	[45]
Korean	1 genome assembly + 40 sequencing data	~4.7	-	2016	[42]
Japanese	1 genome assembly	~2.6	-	2019	[46]

Table 2. Summary of Assembly Statistics of Ethnicity-specific Reference Genomes. a) Since KOREF_C is constructed by substituting variants from 40 samples into KOREF_S1, its assembly statistics are those of KOREF_S1. b) Reported as 1.06% of the expected genome size in the original paper. c) Haploid of paternal chromosomes with maternal chromosome X added. d) Median of the entire samples. e) Average of the entire samples.

Ethnicity	**Name**	**Assembly Method**	**Genome Size (Gb)**	**Contig/Scaffold N50 (Mb)**	**Number of Gaps (Total Gap Length, Mb)**	**Data (Depth)**	**Publication Year**	**Reference**
Ashkenazi Jewish	Ash1 (v1.7)	*De novo* assembly	~2.97	~34.3/~145.1	1,516 (82.9)	Illumina (71x) ONT (23x) PacBio HiFi (29x)	2020	[51,189]
Japanese	JRG (v1)	Integration of novel seqeunces from a *de novo* assembly to	-	-	-	PacBio (101x)	2019	[46]

 doi:10.20944/preprints202605.0651.v1

		GRCh38						
	JG1	Integration of 3 *de novo* assemblies	~3.09	~23.6/~142	473 (~251)	Illumina (55x/59x/57x) Illujmina mate-pair (13x/12x/12x) PacBio (122x/123x/128x) Bionano optical maps (263x/160x/175x)	2021	[52]
Korean	KOREF_C[a)]	*De novo* assembly with variant substitution using 40 sequencing data	~3.12	~0.048/25.3	-(~32.1 Mb)[b]	Illumina (311x) Illumina TruSeq Synthetic Long Read (5.3x) PacBio (10x) OpGen optical maps (240x)	2016	[42,190]
	KOREF_S1(v2.1)[c)]	Haplotype-resolved *de novo* assembly	~2.9	~20/~150.1	-	Illumina (Parental) ONT (235x) PacBio HiFi (38x) Hi-C (294x)	2022	[53,191]
Chinese	HX1	*De novo* assembly	~2.93	~8.3/~22	10,901 (~39.3)	Illumina (143x) PacBio (103x) Bionano optical maps (101x)	2016	[45]
	HJ-H1		~3.07	~28.2/-	427		2022	[44]

 doi:10.20944/preprints202605.0651.v1

	HJ-H2	Haplotype-resolved *de novo* assembly	~2.91	~26/-	390	PacBio Hifi (71x) Hi-C (152x)		
	CN1 (Paternal)	Haplotype-resolved *de novo* assembly	~2.94	Complete Genome	0	Illumina (108x) MGI (97x) MGI (Parental) ONT-UL (79x) PacBio HiFi (69x), PacBio HiFi (Parental) Hi-C (116x)	2023	[33]
	CN1 (Maternal)		~3.04					
	T2T-Yao (Paternal)	Haplotype-resolved *de novo* assembly	~2.92	Complete Genome	0	MGI (278x) MGI (Parental) ONT-UL (336x) PacBio HiFi (92x) Hi-C (584x) Bionano optical mapping	2023	[34]
	T2T-Yao (Maternal)		~3.02					
Chinese (Tujia)	TJ1.p0 (v5)	Haplotype-resolved *de novo* assembly		~14.53/~154.2	907 (~21.5)	Illumina (114x) Illumina mate-pair (39x) Pacbio (165x) Hi-C (118x) Bionano optical mapping (366x) 10x Genomics (135x)	2022	[54]
	TJ1.p1 (v5)		~2.87	~14.53/~155	873 (~19.5)			

 doi:10.20944/preprints202605.0651.v1

Chinese (Tibetan)	ZF1	*De novo* assembly	~2.9	~24.6/ ~58.8	740 (~7.82)	Illumina (100x) PacBio (70x) Hi-C (100x) BioNano optical maps 10x Genomics (100x)	2019	[55]
Mongolian	-	*De novo* assembly	~2.88	~0.056/~7.6	-	Illumina (130.8x)	2014	[56]
Danish	-	*De novo* assembly of 150 samples	~2.83[d]	-/~21[d]	-	Combination of Illumina and Illumina mate-pair (78x)[e]	2017	[57]
Swedish	Swe1	*De novo* assembly	~3.13	~9.5/~49.8	-	PacBio (78.7x) BioNano optical maps (100x)	2018	[47]
	Swe2		~3.1	~8.5/~45.4		PacBio (77.8x) BioNano optical maps (100x)		
Egyptian	EGYPT	*De novo* assembly	~2.82	-	-	Illumina (90x) PacBio (99x) 10x Genomics (80x)	2020	[49]
Saudi Arabian	KSA001 (Paternal)	Haplotype-resolv	~3.03	~133.7/~153.5	17	Illumina (Parental)	2024	[58,192,193]

 doi:10.20944/preprints202605.0651.v1

KSA001 (Maternal)	ed *De novo* Assembly	~3.03	~96.5/154.4	18	ONT-UL (30x) PacBio HiFi (51x) Hi-C (35x)

However, despite their strengths in improving variant calling, ethnicity-specific reference genomes have shown limited usage in subsequent research. For instance, even after constructing the Korean consensus reference genome (KOREF_C) [42], subsequent large-cohort genomic studies (e.g., Korea 1K [12], Korea 4K [13]) from the same research group (Korean Genomics Center, KOGIC) continued to use GRCh38. This limited adoption largely stems from genomic coordinate discrepancies that make variants and discoveries derived from ethnicity-specific references incompatible with major databases (e.g., gnomAD [60], ClinVar [61]) and studies standardized on GRCh37/38.

Even in the post-T2T era, the absence of ethnicity-specific novel sequences and biases in variant calling remain unresolved [33,34]. For example, T2T-CHM13 lacks 429 kb of sequences overlapping 227 genes and 122 regulatory elements identified in the complete gapless Chinese genome (CN1) and causes the omission of 1,871,243 SNVs from 8,869 Chinese individuals [33]. This indicates that, although complete, a single linear reference genome cannot capture the full diversity of human genetic variants across the entire global populations.

Graph-Based Pangenome Reference for Capturing the Entire Spectrum of Human Genetic Variants

The transition from linear to graph-based genome models marks a significant advancement in representing human genetic diversity. Initially, multiple sequence alignment (MSA) was introduced to integrate multiple genome sequences into a single consensus structure that represents both shared sequences and inter-individual variations, laying the groundwork for traditional pangenome studies [62–64]. However, MSA faced limitations due to high computational complexity [65,66], alignment errors in highly polymorphic regions [67], and poor representation of large or non-co-linear structural variants (e.g., inversions, translocations, CNVs) [68]. Later, De Bruijn graph, modeling k-mers as nodes connected and overlapped by edges [69–71], improved efficient and accurate read mapping and variant calling [72], but remained limited in detecting complex structural variants, as performance varies highly depending on the chosen k-mer length [73]. To address these shortcomings, the variation graph models emerged, where nodes denote variants and edges connect them to represent individual's haplotypes [74]. These enabled accurate mapping and calling of even complex variants and structural arrangements such as inversion and duplication [75–77], setting the foundation for modern pangenome references.

Owing to these advancements, in 2023, the Human Pangenome Reference Consortium (HPRC) released the draft human pangenome reference to overcome the limitations of linear reference genomes [78,79]. This reference integrated 47 haplotype-resolved assemblies sourced from 26 global populations into a single variation graph. Compared to GRCh38 which is linear, the human pangenome reference reduced small variant (SNVs and Insertions/Deletions (InDels) <50 bp) discovery errors by ~34% and identified ~64,000 additional small variants per sample. Remarkably, it also doubled the detection of structural variants (SVs >= 50 bp) using short-read sequencing data alone [78]. While the linear human reference genome (GRCh37/38) allows the identification of only ~7,400 [80] - ~9,700 [81] SVs per individual, the human pangenome reference enables the identification of ~16,900 - ~24,900 SVs per individual [78]. Considering that the most accurate method for SV detection, direct comparison of genome assemblies, identifies ~23,000 - ~28,000 SVs per individual [82], these results demonstrate that the human pangenome reference enables the near-complete discovery of the human SV spectrum even with short-read sequencing data.

The human pangenome reference has facilitated the comprehensive and precise discovery of even miscalled and structurally complex variants that have remained poorly characterized for

 doi:10.20944/preprints202605.0651.v1

decades. Although variant calling using ethnicity-specific pangenome references, such as the Chinese [83], Arab [84], and Pacific [85] pangenomes, yet identifies up to ~10 million and ~100 thousand ethnicity-specific small variants and SVs respectively (Table 3), the continuous integration of haplotypes into the human pangenome reference will ultimately enable the complete identification of human genetic variants across the entire global population.

Table 3. Summary Statistics of HPRC and Ethnicity-specific Pangenome References. The number of variants uniquely called using ethnicity-specific pangenome references is counted relative to HPRC. a) Novel sequences not existing in all four reference genomes. MC: Minigraph-Cactus. PGGB: Pangenome Graph Builder. SV: Structural variant.

Name	Number of Individuals	Total Number of Nodes/Edges (m) (Length, Gb)	Total Length of Novel Sequences (Mb) (Reference genome used for comparison)	Total Number of Ethnicity-specific Variants	Publication Year	Reference
Human Pangenome Reference (HPRC)	47	MC: ~85.6/~118.4 (3.32) PGGB: ~110.0/154.8 (~8.42)	~119 (CHM13)	-	2023	[78]
Chinese Pangenome (CPC)	58	MC: ~64.5/~89.6 (~3.38)	~189 (GRCh38)	small variants: ~5.9m SVs: ~34.2k	2023	[83]
Arab Pangenome Reference (APR)	43	MC: ~72.3/~99.7 (3.31)	~101 (GRCh38, CHM13, HPRC, CPC)[a)]	small variants: ~10.7m SVs: ~108.7k	2023	[84]
Pacific Ancestry Pangenome Reference	23	MC: ~47.0/~65.0 (~3.22)	~31 (HPRC)	small variants: ~3.4m SVs: ~4.7k	2024	[85]

Genetic Variants: Vastly Cataloged but Poorly Interpreted

Large-Cohort Population Genetic Studies Have Discovered Massive Amounts of Genetic Variants

Advancements in sequencing technologies opened the era of large-cohort population genetic studies. Large-cohort genome projects including the Personal Genome Project (PGP) [86–88], 1,000 Genomes Project (1KGP) [89–92], All of Us Research Program [8–11], UK Biobank [4–6], and TOPMed [7] have collected numerous amounts of human genomic data, which enabled the massive accumulation of genetic variants (Table 4). For instance, 1KGP [91] identified ~88 million variants (~84.7 million SNVs, ~3.6 million InDels, and ~60 thousand SVs) from 2,504 individuals, while TOPMed [7] discovered ~410 million genetic variants (~381 million SNVs and 29 million InDels) from 53,831 individuals, incorporating ~40 million and ~323 million additional variants (SNVs and InDels) into dbSNP, respectively. Corresponding to the massive accumulation of genetic variants, dbSNP (established by the National Center for Biotechnology Information (NCBI) in 1998 [93]) catalogs ~1.1 billion unique small variants, including SNVs, multi-nucleotide variants (MNVs), InDels, short tandem repeats (STRs), and retrotransposable element insertions as of build 156 [59]. Similarly, gnomAD [60] (founded by the Exome Aggregation Consortium (ExAC) in 2014 [94]) includes ~786.6

 doi:10.20944/preprints202605.0651.v1

million SNVs and ~122.6 million InDels as of version 4. Regarding SVs, gnomAD houses ~1.2 million SVs and ~66.9 thousand rare (site frequency < 1%) exome CNVs [95], while dbVar [96] (launched by National Center for Biotechnology Information (NCBI) in 2010 [97]) catalogs ~38.4 million SVs spanning ~8.2 million genomic regions as of May 2025 [98].

Table 4. Summary of Large-cohort Genome Projects. a) Number of variants identified from 3,552 samples of the project. b) Number of variants identified from 4,810 samples of the project. c) Number of biallelic variants identified. d) Number of variants identified from 3,781 samples of the project. SNV: Single nucleotide variant. InDel: Insertion and deletion. SV: Structural variant.

Project	Lead Institute (Principal Investigators)	Number of Samples	Target Final Sample Size	Number of Variants Identifie d	Start–End Date	References
Harvard Personal Genome Project	Harvard, US (Dr. George Church)	3,083 (as of access date: May 26, 2025)	100,000	-	2005 -	[86–88,194]
1,000 Genomes Project	Wellcome Sanger Institute, UK, Beijing Genomics Institute, China, National Human Genome Research Institute (NHGRI), US	2,504	-	~84.7m SNVs ~3.6m InDels ~60m SVs	2008 - 2015	[89–92]
All of Us Research Program	National Institute of Health (NIH), US (Dr. Josh Denny)	414,840 (as of October 2023)	1,000,000	~1.4b SNVs and InDels	2018 -	[8–11,195,196]
UK Biobank	UK Biobank, UK (Dr. Rory Collins)	490,640 (as of December 2023)	500,000	~1b SNVs ~97.2m InDels ~1.9m SVs	2006 -	[4–6,197,198]
Trans-Omics for Precision Medicine Program	National Heart, Lung, and Blood Institute (NHLBI), US (Dr. Albert Smith, Mr. Alastair	~206,000 (as of data freeze 9)	-	~781m SNVs ~62m InDels	2014 -	[7,199–201]

 doi:10.20944/preprints202605.0651.v1

	Thomson, Dr. Weiniu Gan)					
Korean Genome Project	Korean Genomics Center (KOGIC), Korea (Dr. Jong Bhak)	4,157 (as of April 2024)	10,000	~64.3m SNVs 8.8m InDels	2016 -	[12,13]
National Bio Bigdata Project (Pilot Project of BioBigData. Korea)	Korea Research Institute of Bioscience and Biotechnology (KRIBB), Korea, Korea National Institute of Heath (KNIH), Korea (Dr. Seonyong Kim, Dr. Heonyong Park)	15,000	15,000	~422.3.6m SNVs, ~ 38.3 InDels, ~ 12,629 SVs[a)]	2020-2022	This study
BioBigData. Korea	BioBigData.Korea (BIKO), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Korea (Dr. Rong-Min Baek, Dr. Haeyoung Jeong)	4,000 (as of June 2025)	340,000	-	2024 -	This study
Simons Genome Diversity Project	Harvard, US (Dr. David Reich)	278	300	~34.4m SNVs ~2.1m InDels	-	[202]
Tohoku Medical Megabank Project	Tohoku Medical Megabank Organization, Tohoku University, Japan (Dr. Masayuki Yamamoto)	100,000 (as of June, 2024)	150,000	~45.9m SNVs ~6.3, InDels[a)]	2011 -	[203–210]

 doi:10.20944/preprints202605.0651.v1

China Metabolic Analytics Project	Shanghai Jiao Tong University, China (Dr. Guang Ning)	10,588 (as of April 2020)	1,100,000	~136.8m SNVs ~10.7m InDels	-	[211,212]
Westlake BioBank for Chinese	Westlake University, China (Dr. Hou-Feng Zheng)	10,376 (as of May 2022)	100,000	~74.1m SNVs ~7.4m InDels	-	[15,16,213,214]
100,000 Genomes Project	Genomics England (Dr. Mark Caulfield)	~100,000	100,000	-	2012 – 2018	[17,18,215,216]
Singapore National Precision Medicine Project	Precision Health Research, Singapore (Dr. Patrick Tan)	10,323 (as of January 2023)	1,000,000	~89.2m SNVs ~9.1m InDels[b)]	2017 -	[217–219]
GenomeAsia 100K Project	Nanyang Technological University, Singapore (Dr. Stephan Schuster)	1,739 (as of December 2019)	100,000	~63.2m SNVs ~3.8m InDels	2016 -	[14,220]
Genome of Netherlands	University of Groningen (Dr. Cisca Wijmenga)	769 (as of June 2014)	1,000	~20.4m SNVs[c)] ~1.2m InDels[c)] ~27.5k SVs	-	[221–223]
UK 10K Project	Wellcome Sanger Institute (Dr. Richard Durbin)	8,963	10,000	~42m SNVs[d)] ~3.5m InDels[d)] ~18.7k SVs[d)]	2010 - 2013	[224–227]
Egypt Genome Project	Egypt Center for Research and Regenerative Medicine (Dr. Khaled Amer)	-	100,000	-	2022 -	[228,229]
SweGen Project	Uppsala University (Dr. Ulf Gyllensten)	1,000	-	~29.2 m SNVs	-	[230]

 doi:10.20944/preprints202605.0651.v1

	~3.8m InDels ~8.6m SVs

The accumulation of small variants in dbSNP derived from GRCh38 has plateaued since 2021 [59], which is just six years after the completion of the 1KGP. Although the discovery of SVs has progressed relatively slowly compared to small variants due to their difficulty in identification, the human pangenome reference has now enabled precise SV calling even with short-read sequencing data. As a result, with the availability of large-scale genome data from existing cohorts, the complete discovery of the full spectrum of human genetic variants is expected to be completed soon.

Clinical Interpretation of Genetic Variants Through GWAS

The final goal of large-cohort population genetic studies is interpreting the clinical impacts of genetic variants on diseases and phenotypes. GWAS (first conducted in 2003 by Dr. Tanaka [99–101]) has been the standard approach for this purpose. Collectively, GWASs have expanded our understanding of genetic variants. For instance, the Korean Genome Project discovered 2,635 significant associations between 2,324 variants and 34 clinical traits [13]. Likewise, the FinnGen project identified 275 significant associations between 235 loci and 15 diseases [102].

Such clinical interpretations of genetic variants through GWASs have contributed to the medical field. One notable example is the repurposing of Ustekinumab, a monoclonal antibody preventing the interaction of Interleukin-12 and Interleukin-23 with their receptors originally for psoriasis therapeutics [103,104]. A study identified a nonsynonymous SNV in the Interleukin-23 Receptor (*IL23R)* gene impacting Crohn's disease through GWAS [105]. Building on this discovery, a series of research studies revealed the efficacy of Ustekinumab for Crohn's disease [106–108]. This line of research ultimately led to the official approval of Ustekinumab as a treatment for Crohn's disease [104].

GWAS Cannot Lead to the Complete Clinical Interpretation of Genetic Variants

Despite extensive GWASs, the clinical effects of most genetic variants are still elusive. This is primarily because GWAS has limited applicability to most genetic variants. GWAS requires a large sample size for reliable variant interpretation with sufficient statistical power [109]. However, almost all (more than 99%) missense variants exhibit minor allele frequency (MAF) below 0.5% [110]. Also, ~50% of SNVs and InDels found in enhancers, promoters, open chromatin regions, and 5′ and 3′ untranslated regions consist of singletons [7]. Moreover, in a large-cohort study analyzing ~150,000 genomes, ~620.8 million (~96.4%) of the ~643.7 million SNVs and InDels had frequency below 0.1%, including ~293.1 million (~45.5%) singletons [4], indicating that even such a large dataset does not provide a sufficient sample size for most variants to be interpreted through GWAS.

In addition, GWAS struggles to interpret variants with small impacts due to its stringent significance threshold [111]. Most human complex traits result from cumulative effects of numerous genetic variants, each exerting only a small influence. Thus, GWAS can interpret only a small fraction of the genetic variants contributing to a given phenotype, leading to the missing heritability problem [112]. For instance, variants annotated through GWAS account for only ~1.5% [113] and ~10.5% [114] of the variation in BMI and height, respectively, despite the heritabilities of these traits being ~47% - ~90% [115], and 85% - 91% [116]. In contrast, incorporating thousands of common SNVs increases the explained variance to ~16.5%, and ~44.8% for these traits [117]. This underscores the limited applicability of GWAS to not only rare variants but also many common variants.

Finally, GWAS misinterprets the clinical impacts of many variants. For example, Wünnemann et al. could only validate 42 variants as contributors to coronary artery disease using CRISPR perturbation, out of 1,998 sentinel and proxy variants annotated through GWASs [118]. This is because GWAS interprets each genetic variant independently. Genetic variants determine human

 doi:10.20944/preprints202605.0651.v1

phenotype through complicated interactions. Thus, the clinical effect of one variant can be amplified or abolished by another variant [119,120]. For instance, although an individual's *OCA2* gene contains the alleles for brown eyes, an SNV in intron 86 of the *HERC2* gene can override this causing the individual to have non-brown eyes [121]. Applying GWAS in this case may lead to the misinterpretation of the brown eye alleles as determinants of different eye colors. This shows that GWAS not only exhibits limited applicability to genetic variants but even misinterprets the clinical impacts of genetic variants.

The root of all GWAS limitations lies in its inability to account for the biological mechanisms of genetic variants. Genetic variants collectively shape human biology by forming and regulating DNA elements that define genomic contexts and interact in highly complex, interdependent ways [122,123]. Deciphering the hidden logic encoded within the human genome is therefore a prerequisite for fully interpreting the clinical impacts of genetic variants. However, as GWAS cannot capture these underlying interdependencies, regardless of how many additional samples are sequenced and variants are identified through large-scale genome projects, achieving the complete clinical interpretation of genetic variants through GWAS will remain challenging.

Large Language Models for the Complete Interpretation of Human Genetic Variants

Deciphering the Hidden Principles Encoded in the Genome for Precise Interpretation of Genetic Variants

Deciphering the biological principles encoded within the genome—the fundamental rules governing the arrangement and interaction of its sequence elements—is the only way to completely interpret the functions of every known and yet-to-be-discovered genetic variant. Analogous to human languages [124], genomic and protein sequences are built from a finite set of letters (four nucleotide bases in DNA; twenty amino acids in proteins), whose specific combinations form distinct functional elements (in proteins, motifs and domains; in the genome, promoter, enhancers, regulatory elements, etc)—analogous to a lexicon and morphology [125–129]. The order, arrangement, and composition of these elements (syntax) dictate their functional roles (semantics) [130]— for example, conferring a protein's 3D-structure and enzymatic functions [131], directing chromatin remodeling [132], modulating the timing and magnitude of gene expression [133–136]. Importantly, genetic variants perturb this encoded logic by altering the sequence and structure of the elements, leading to shifts in their functional roles, and consequently affecting clinical outcomes and phenotypes [137–139]. Thus, decoding the genome's underlying logic (grammar) is the only definitive approach—surpassing the persistent constraints of sample size and traditional methodologies—to completely unveil the biological and clinical implications of all genetic variants.

Recently, LLMs have emerged as the most promising solution for decoding the long-unresolved hidden principles of genomes. Central to LLMs are attention mechanisms (especially, multi-head self-attention), which accurately capture context-dependent and higher-order interactions among all elements within a sequence. Specifically, self-attention computes each element's importance (attention weight) by evaluating its relationships with other elements, dynamically adjusting these weights based on the surrounding context. Together, multi-head structure runs multiple self-attention processes in parallel, uncovering distinct interaction patterns and hierarchical structures within the sequence [140–144]. At the lower level, it reveals functional motifs and elements derived from nucleotide bases and amino acids; at the higher level, it captures functional implications based on their organizations [145–149].

Building on these attention-based architectures, LLMs effectively leverage self-supervised learning to uncover the hidden logic encoded within unlabeled sequences by predicting masked or subsequent sequence segments based on their surrounding or prior contexts [142,150]. This approach is especially well-suited to genomics, given the availability of millions of protein and genomic sequences—spanning thousands of species—that encode a structured grammar shaped through billions of years of evolution [151,152]. Indeed, all contemporary genomes descend from the last

 doi:10.20944/preprints202605.0651.v1

universal common ancestor (LUCA) [153], preserving deeply conserved and divergent patterns that reveal how functional motifs and elements—such as enzymatic domains, promoters, enhancers, and regulatory elements—have been progressively assembled, recombined, and diversified across lineages to drive the molecular and phenotypic diversity observed in modern species [154–157]. Through self-supervised pretraining on a vast collection of these sequences, LLMs autonomously learn conserved motifs, co-occurrence patterns, and the hierarchical organization of functional elements without relying on any explicit annotations [151,152,158–161].

Accurate Pathogenicity Prediction of Coding Variants Through Proteomic Language Models

Coding variants determine human diseases and phenotypes by altering the structures and interactions of proteins. However, computationally modeling protein structures remained a half-century challenge, largely due to the complex higher-order interactions among residues involved in protein folding [162–164]. Recently, AlphaFold2 (developed by Google DeepMind in 2021), a proteomic deep learning model, broke new ground by predicting protein structures up to an experimental-level accuracy (a GDT_TS score > 90) [165] for the first time [166]. AlphaFold2 (developed by Google DeepMind in 2021) attained a median Global Distance Test (GDT) of 92.4 and a median Cα root-mean-square deviation (RMSD) of 0.96 Å at the 14th Critical Assessment of Structure Prediction (CASP14) [166,167], far surpassing the < 2–3 Å structural error threshold required for reliable biological interpretation and drug discovery [162]. Strikingly, while AlphFold2 relies on structural information, proteomic language models can accurately predict protein structures using only sequence data, without any MSAs for training. For instance, ESMfold (developed by MetaAI in 2023) achieved a remarkable Template Modeling score (TM-score) [166,168] of 0.83 at Continuous Automated Model EvaluatiOn (CAMEO) [169], nearly comparable to AlphaFold's 0.88 [161]. Moreover, Recurrent Geometric Network 2 (RGN2) outperformed AlphaFold2 in predicting the structures of orphan proteins with low homology to proteins in existing structural databases (average ΔdRMSD of 0.65 Å and ΔGDT_TS of 5.34 against AlphaFold2) [170,171].

By modelling amino acid co-variation, conserved motifs and their interactions that underlie protein folds [160,172–174], proteomic language models even precisely identify the pathogenicity of missense variants. For instance, ESM1b, without any variant-specific training, cataloged the pathogenicity of all possible ~450 million missense variants, including even previously unannotated ones, across 42,336 human protein isoforms with an auROC of 0.905 for ClinVar annotated variants [160,175]. Furthermore, unlike most variant effect prediction (VEP) methods, it effectively predicted the pathogenicity of complex alterations such as in-frame InDels (includes deletion–insertion combinations (DelIns) here) and SVs [175]. For InDels in ClinVar, ESM1b, although unsupervised, attained an auROC of 0.874, outperforming Combined Annotation–Dependent Depletion (CADD), the widely used supervised VEP model [176,177], which achieved an auROC of 0.835. When restricted to DelIns, which are highly complex, ESM1b attained an auROC of 0.887, far surpassing CADD's 0.671 [175]. These capabilities of protein language models to infer protein structures and variant pathogenicity directly from sequence alone, open new opportunities for the clinical interpretation of coding variants.

Accurate Pathogenicity Prediction of Noncoding Variants Through Genomic Language Models

Interpreting the clinical impacts of non-coding variants—whose functional consequences vary depending on whether they disrupt promoters, enhancers, other regulatory elements—necessitates the accurate identification of these regulatory features. Recent breakthroughs in genomic language models such as DNABERT-2 [178] and Nucleotide Transformer (NT) identify regulatory elements by merely pre-training on hundreds of multi-species genomes without any biological label. Surprisingly, NT accurately distinguishes diverse genomic elements, including enhancers, UTRs, introns, exons, promoters, and CTCF binding motifs, with accuracy exceeding 0.78 [146]. This shows that genomic language models successfully capture the syntax, lexicon, and morphology of the genome.

 doi:10.20944/preprints202605.0651.v1

By capturing the hidden logic of the genome, genomic language models infer the pathogenic impacts of genetic variants without any variant-specific supervision. For instance, NT achieved auROCs of 0.8 and 0.7 for pathogenicity classification on ClinVar and Human Gene Mutation Database (HGMD) [179], respectively [146]. Furthermore, Genomic Pre-trained Network-MSA (GPN-MSA), pre-trained with whole-genome MSA of 100 vertebrate genomes, performed better than CADD with an auROC of 0.969 compared to 0.963 in predicting the pathogenicity of ClinVar variants [180]. It also achieved more than twice the accuracy when classifying somatic missense variants frequently observed in the Catalogue Of Somatic Mutations In Cancer (COSMIC) [180,181]. Furthermore, GPN-MSA achieved up to a fourfold higher auROC than the conventional VEP for regulatory variants residing in promoters, enhancers, UTRs, and ncRNAs that are implicated in Mendelian disorders (OMIM) [180,182].

In 2025, Evo 2, the state-of-the-art genomic language model pre-trained on over 128,000 genomes of multi-species unprecedentedly, further enabled the precise pathogenicity prediction of even non-SNVs [152]. For non-SNV coding variants, where models like AlphaMissense [183] and GPN-MSA [180] fall short, Evo 2 achieved an auROC of 0.918 for pathogenicity prediction based on zero-shot classification. Similarly, for noncoding variants, Evo 2 showed superior predictability for both SNVs and non-SNVs, achieving an auROC of 0.987 for SNVs and 0.971 for non-SNVs [152].

Conclusion

Despite showing remarkable performance in predicting the pathogenicity of variants by modelling the fundamental principles and architectures of the genome, current genomic and proteomic language models cannot directly serve as a superior alternative to GWAS, nor can they immediately realize the complete clinical interpretation of genetic variants. This is because such pathogenicity predictions only yet reflect genetic variants' impact on organismal fitness—such as disruptions to conserved structure or function—rather than specific clinical traits or diseases. Furthermore, considering that the clinical impacts of genetic variants vary depending on cellular and environmental context (e.g., cell type, developmental stages, metabolic conditions, and external stimuli) [184], present LLMs are insufficient for completely interpreting the clinical impacts of genetic variants.

Nevertheless, LLM will ultimately enable the complete clinical interpretation of genetic variants since LLM can integrate genomic, multiomic and clinical data. By integrating these data, LLM can precisely model how genetic variants determine diseases and clinical traits through multiomic layers across diverse biological states and environmental conditions. In fact, by leveraging supervised and multi-task learning with multiomic data, recent LLM-like transformer-based deep learning models have demonstrated unprecedented performance in capturing the cascading effects of genetic variants across gene regulations—such as chromatin accessibility, transcription, splicing, and translation—linking sequence variation to downstream molecular phenotypes. For instance, sequence-to-function models like Enformer [185] and Borzoi [186] achieved near-experimental accuracy in predicting variant impacts on gene expression. By multi-task learning with multiomic data such as Cap Analysis of Gene Expression sequencing (CAGE-seq), RNA-seq, Chromatin Immunoprecipitation sequencing (ChIP-seq), DNase I Hypersensitive Sites sequencing (DNase-seq), and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), these models decoded the regulatory mechanisms of noncoding regions, identifying enhancer-promoter interactions and functional impacts of genetic variants that drive cell-type-specific and tissue-specific transcriptional shifts, including alternative isoform usage and polyadenylation. To further improve the predictability for inter-individual variability, models like Performer [187] and UKBioFormer [188] (fine-tuned extensions of Enformer) incorporated paired genome–RNA-seq data or large-cohort genomic data (e.g., genetic variants from the UK Biobank), respectively, thereby enhancing the prediction of personalized gene expression. Altogether, these results suggest that although challenges remain, precisely modeling the complex biology of humans through integrating genomic, multiomic, and

 doi:10.20944/preprints202605.0651.v1

clinical data within an LLM framework will ultimately enable the complete clinical interpretation of genetic variants.

Data Availability: No new data were generated in this study.

Acknowledgments: We thank all participants of the Korean Genome Project and BioBigData.Korea. We thank all members of the Korean Genomics Center and Korea Bioinformation Center.

Author Contributions: Jihun Bhak (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Dong-Hyun Shin (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jongbum Jeon (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Soobok Joe (Writing—review & editing), Yeonsu Jeon (Writing—review & editing), Hyoungjin Choi (Writing—review & editing), Yoonsung Kwon (Writing—review & editing), Kyungwhan An (Writing—review & editing), Yun Sung Cho (Writing—review & editing), Sungwon Jeon (Writing—review & editing), Haeyoung Jeong (Writing—review & editing, Supervision), Jong Bhak (Writing—review & editing, Supervision).

Funding: This study was supported by BioBigData.Korea (RS-2024-00438566).

Conflicts of Interest: J. Bhak is the founder of AgingLab. S. Jeon is the CEO of AgingLab. Y. Cho is an employee of CG Invites Co., LTD and Invites Genomics Co., LTD.

References

1. Claussnitzer, M., Cho, J.H., Collins, R. et al. A brief history of human disease genetics. *Nature*. 2020; 577: 179–189. https://doi.org/10.1038/s41586-019-1879-7
2. Li, H. and Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat Rev Genet*. 2024; 25: 658–670. https://doi.org/10.1038/s41576-024-00718-w
3. Shendure, J., Balasubramanian, S., Church, G.M. et al. DNA sequencing at 40: past, present and future. *Nature*. 2017; 550: 345–353. https://doi.org/10.1038/nature24286
4. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature*. 2022; 607: 732–740. https://doi.org/10.1038/s41586-022-04965-x
5. Bycroft, C., Freeman, C., Petkova, D. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018; 562: 203–209. https://doi.org/10.1038/s41586-018-0579-z
6. Li, S., Carss, K.J., Halldorsson, B.V. et al. Whole-genome sequencing of half-a-million UK Biobank participants. *medRxiv*. 2023; https://doi.org/10.1101/2023.12.06.23299426
7. Taliun, D., Harris, D.N., Kessler, M.D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021; 590: 290–299. https://doi.org/10.1038/s41586-021-03205-y
8. The All of Us Research Program Investigators. The "All of Us" Research Program. *N Engl J Med*. 2019; 381: 668–676. https://doi.org/10.1056/NEJMsr1809937
9. The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature*. 2024; 627: 340–346. https://doi.org/10.1038/s41586-023-06957-x
10. Ramirez, A.H., Gebo, K.A. and Harris, P.A. Progress With the All of Us Research Program: Opening Access for Researchers. *JAMA*. 2021; 325: 2441–2442. https://doi.org/10.1001/jama.2021.7702
11. Ramirez, A.H., Sulieman, L., Schlueter, D.J. et al. The All of Us Research Program: Data quality, utility, and diversity. *Patterns*. 2022; 3: 100570. https://doi.org/10.1016/j.patter.2022.100570
12. Jeon, S., Bhak, Y., Choi, Y. et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv*. 2020; 6: eaaz7835. https://doi.org/10.1126/sciadv.aaz7835
13. Jeon, S., Choi, H., Jeon, Y. et al. Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups. *Gigascience*. 2024; 13: https://doi.org/10.1093/gigascience/giae014
14. McGonigle, I. and Schuster, S.C. Global science meets ethnic diversity: Ian McGonigle interviews GenomeAsia100K Scientific Chairman Stephan Schuster. *Genet Res*. 2019; 101: e5. https://doi.org/10.1017/S001667231800006X

 doi:10.20944/preprints202605.0651.v1

15. Cong, P.K., Bai, W.Y., Li, J.C. et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat Commun*. 2022; 13: 2939. https://doi.org/10.1038/s41467-022-30526-x
16. Zhu, X.W., Liu, K.Q., Wang, P.Y. et al. Cohort profile: the Westlake BioBank for Chinese (WBBC) pilot project. *BMJ Open*. 2021; 11: e045564. https://doi.org/10.1136/bmjopen-2020-045564
17. The 100,000 Genomes Project Pilot Investigators. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med*. 2021; 385: 1868–1880. https://doi.org/10.1056/NEJMoa2035790
18. Sosinsky, A., Ambrose, J., Cross, W. et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med*. 2024; 30: 279–289. https://doi.org/10.1038/s41591-023-02682-0
19. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409: 860–921. https://doi.org/10.1038/35057062
20. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431: 931–945. https://doi.org/10.1038/nature03001
21. Church, D.M., Schneider, V.A., Graves, T. et al. Modernizing reference genome assemblies. *PLoS Biol*. 2011; 9: e1001091. https://doi.org/10.1371/journal.pbio.1001091
22. Schneider, V.A., Graves-Lindsay, T., Howe, K. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017; 27: 849–864. https://doi.org/10.1101/gr.213611.116
23. Genome Reference Consortium. Human Genome Assembly GRCh37. https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37, (26 May 2025, date last accessed).
24. Genome Reference Consortium. Human Genome Assembly GRCh38.p14. https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38.p14, (26 May 2025, date last accessed).
25. Miga, K.H., Koren, S., Rhie, A. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. 2020; 585: 79–84. https://doi.org/10.1038/s41586-020-2547-7
26. Eichler, E.E., Clark, R.A. and She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet*. 2004; 5: 345–354. https://doi.org/10.1038/nrg1322
27. Nurk, S., Koren, S., Rhie, A. et al. The complete sequence of a human genome. *Science*. 2022; 376: 44–53. https://doi.org/10.1126/science.abj6987
28. Steinberg, K.M., Schneider, V.A., Graves-Lindsay, T.A. et al. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res*. 2014; 24: 2066–2076. https://doi.org/10.1101/gr.180893.114
29. Hon, T., Mars, K., Young, G. et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data*. 2020; 7: 399. https://doi.org/10.1038/s41597-020-00743-4
30. Wenger, A.M., Peluso, P., Rowell, W.J. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019; 37: 1155–1162. https://doi.org/10.1038/s41587-019-0217-9
31. Jain, M., Koren, S., Miga, K.H. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018; 36: 338–345. https://doi.org/10.1038/nbt.4060
32. Payne, A., Holmes, N., Rakyan, V. et al. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*. 2019; 35: 2193–2198. https://doi.org/10.1093/bioinformatics/bty841
33. Yang, C., Zhou, Y., Song, Y. et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res*. 2023; 33: 745–761. https://doi.org/10.1038/s41422-023-00849-5
34. He, Y., Chu, Y., Guo, S. et al. T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. *Genomics Proteomics Bioinformatics*. 2023; 21: 1085–1100. https://doi.org/10.1016/j.gpb.2023.08.001
35. Richards, S., Aziz, N., Bale, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17: 405–424. https://doi.org/10.1038/gim.2015.30
36. Halim-Fikri, H., Syed-Hassan, S.R., Wan-Juhari, W.K. et al. Central resources of variant discovery and annotation and its role in precision medicine. *Asian Biomed (Res Rev News)*. 2022; 16: 285–298. https://doi.org/10.2478/abm-2022-0032

 doi:10.20944/preprints202605.0651.v1

37. Harrow, J., Frankish, A., Gonzalez, J.M. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012; 22: 1760–1774. https://doi.org/10.1101/gr.135350.111
38. Ballouz, S., Dobin, A. and Gillis, J.A. Is it time to change the reference genome? *Genome Biol*. 2019; 20: 159. https://doi.org/10.1186/s13059-019-1774-4
39. Church, D.M., Schneider, V.A., Steinberg, K.M. et al. Extending reference assembly models. *Genome Biol*. 2015; 16: 13. https://doi.org/10.1186/s13059-015-0587-3
40. Jager, M., Schubach, M., Zemojtel, T. et al. Alternate-locus aware variant calling in whole genome sequencing. *Genome Med*. 2016; 8: 130. https://doi.org/10.1186/s13073-016-0383-z
41. Duan, Z., Qiao, Y., Lu, J. et al. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol*. 2019; 20: 149. https://doi.org/10.1186/s13059-019-1751-y
42. Cho, Y.S., Kim, H., Kim, H.M. et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun*. 2016; 7: 13637. https://doi.org/10.1038/ncomms13637
43. Li, Q., Tian, S., Yan, B. et al. Building a Chinese pan-genome of 486 individuals. *Commun Biol*. 2021; 4: 1016. https://doi.org/10.1038/s42003-021-02556-6
44. Yang, X., Zhao, X., Qu, S. et al. Haplotype-resolved Chinese male genome assembly based on high-fidelity sequencing. *Fundam Res*. 2022; 2: 946–953. https://doi.org/10.1016/j.fmre.2022.02.005
45. Shi, L., Guo, Y., Dong, C. et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun*. 2016; 7: 12065. https://doi.org/10.1038/ncomms12065
46. Nagasaki, M., Kuroki, Y., Shibata, T.F. et al. Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. *Hum Genome Var*. 2019; 6: 27. https://doi.org/10.1038/s41439-019-0057-7
47. Ameur, A., Che, H., Martin, M. et al. De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. *Genes*. 2018; 9: https://doi.org/10.3390/genes9100486
48. Eisfeldt, J., Martensson, G., Ameur, A. et al. Discovery of Novel Sequences in 1,000 Swedish Genomes. *Mol Biol Evol*. 2020; 37: 18–30. https://doi.org/10.1093/molbev/msz176
49. Wohlers, I., Kunstner, A., Munz, M. et al. An integrated personal and population-based Egyptian genome reference. *Nat Commun*. 2020; 11: 4719. https://doi.org/10.1038/s41467-020-17964-1
50. Sherman, R.M., Forman, J., Antonescu, V. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*. 2019; 51: 30–35. https://doi.org/10.1038/s41588-018-0273-y
51. Shumate, A., Zimin, A.V., Sherman, R.M. et al. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol*. 2020; 21: 129. https://doi.org/10.1186/s13059-020-02047-7
52. Takayama, J., Tadaka, S., Yano, K. et al. Construction and integration of three de novo Japanese human genome assemblies toward a population-specific reference. *Nat Commun*. 2021; 12: 226. https://doi.org/10.1038/s41467-020-20146-8
53. Kim, H.S., Jeon, S., Kim, Y. et al. KOREF_S1: phased, parental trio-binned Korean reference genome using long reads and Hi-C sequencing methods. *Gigascience*. 2022; 11: https://doi.org/10.1093/gigascience/giac022
54. Lou, H., Gao, Y., Xie, B. et al. Haplotype-resolved de novo assembly of a Tujia genome suggests the necessity for high-quality population-specific genome references. *Cell Syst*. 2022; 13: 321–333 e326. https://doi.org/10.1016/j.cels.2022.01.006
55. Ouzhuluobu, He, Y., Lou, H. et al. De novo assembly of a Tibetan genome and identification of novel structural variants associated with high-altitude adaptation. *Natl Sci Rev*. 2020; 7: 391–402. https://doi.org/10.1093/nsr/nwz160
56. Bai, H., Guo, X., Zhang, D. et al. The genome of a Mongolian individual reveals the genetic imprints of Mongolians on modern human populations. *Genome Biol Evol*. 2014; 6: 3122–3136. https://doi.org/10.1093/gbe/evu242
57. Maretty, L., Jensen, J.M., Petersen, B. et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*. 2017; 548: 87–91. https://doi.org/10.1038/nature23264
58. Kulmanov, M., Tawfiq, R., Liu, Y. et al. A reference quality, fully annotated diploid genome from a Saudi individual. *Sci Data*. 2024; 11: 1278. https://doi.org/10.1038/s41597-024-04121-2
59. Phan, L., Zhang, H., Wang, Q. et al. The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res*. 2025; 53: D925–D931. https://doi.org/10.1093/nar/gkae977

 doi:10.20944/preprints202605.0651.v1

60. Chen, S., Francioli, L.C., Goodrich, J.K. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*. 2024; 625: 92–100. https://doi.org/10.1038/s41586-023-06045-0
61. Landrum, M.J., Lee, J.M., Riley, G.R. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014; 42: D980–985. https://doi.org/10.1093/nar/gkt1113
62. Vernikos, G., Medini, D., Riley, D.R. et al. Ten years of pan-genome analyses. *Curr Opin Microbiol*. 2015; 23: 148–154. https://doi.org/10.1016/j.mib.2014.11.016
63. Tettelin, H., Masignani, V., Cieslewicz, M.J. et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. 2005; 102: 13950–13955. https://doi.org/10.1073/pnas.0506758102
64. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief Bioinform*. 2018; 19: 118–135. https://doi.org/10.1093/bib/bbw089
65. Wang, L. and Jiang, T. On the complexity of multiple sequence alignment. *J Comput Biol*. 1994; 1: 337–348. https://doi.org/10.1089/cmb.1994.1.337
66. Edgar, R.C. and Batzoglou, S. Multiple sequence alignment. *Curr Opin Struct Biol*. 2006; 16: 368–373. https://doi.org/10.1016/j.sbi.2006.04.004
67. Ranwez, V. and Chantret, N., N., *Phylogenetics in the Genomic Era*. No commercial publisher | Authors open access book, 2020.
68. Katoh, K. and Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010
69. Holley, G., Wittler, R. and Stoye, J. Bloom Filter Trie: an alignment-free and reference-free data structure for pan-genome storage. *Algorithms Mol Biol*. 2016; 11: 3. https://doi.org/10.1186/s13015-016-0066-8
70. Idury, R.M. and Waterman, M.S. A new algorithm for DNA sequence assembly. *J Comput Biol*. 1995; 2: 291–306. https://doi.org/10.1089/cmb.1995.2.291
71. Pevzner, P.A., Tang, H. and Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*. 2001; 98: 9748–9753. https://doi.org/10.1073/pnas.171285098
72. Limasset, A., Cazaux, B., Rivals, E. et al. Read mapping on de Bruijn graphs. *BMC Bioinformatics*. 2016; 17: 237. https://doi.org/10.1186/s12859-016-1103-9
73. Iqbal, Z., Caccamo, M., Turner, I. et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*. 2012; 44: 226–232. https://doi.org/10.1038/ng.1028
74. Garrison, E. and Guarracino, A. Unbiased pangenome graphs. *Bioinformatics*. 2023; 39: https://doi.org/10.1093/bioinformatics/btac743
75. Garrison, E., Siren, J., Novak, A.M. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 2018; 36: 875–879. https://doi.org/10.1038/nbt.4227
76. Rakocevic, G., Semenyuk, V., Lee, W.P. et al. Fast and accurate genomic analyses using genome graphs. *Nat Genet*. 2019; 51: 354–362. https://doi.org/10.1038/s41588-018-0316-4
77. Eggertsson, H.P., Kristmundsdottir, S., Beyter, D. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun*. 2019; 10: 5402. https://doi.org/10.1038/s41467-019-13341-9
78. Liao, W.W., Asri, M., Ebler, J. et al. A draft human pangenome reference. *Nature*. 2023; 617: 312–324. https://doi.org/10.1038/s41586-023-05896-x
79. Wang, T., Antonacci-Fulton, L., Howe, K. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature*. 2022; 604: 437–446. https://doi.org/10.1038/s41586-022-04601-8
80. Collins, R.L., Brand, H., Karczewski, K.J. et al. A structural variation reference for medical and population genetics. *Nature*. 2020; 581: 444–451. https://doi.org/10.1038/s41586-020-2287-8
81. Byrska-Bishop, M., Evani, U.S., Zhao, X. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 2022; 185: 3426–3440 e3419. https://doi.org/10.1016/j.cell.2022.08.004
82. Ebert, P., Audano, P.A., Zhu, Q. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021; 372: https://doi.org/10.1126/science.abf7117
83. Gao, Y., Yang, X., Chen, H. et al. A pangenome reference of 36 Chinese populations. *Nature*. 2023; 619: 112–121. https://doi.org/10.1038/s41586-023-06173-7

 doi:10.20944/preprints202605.0651.v1

84. Nassir, N., Almarri, M.A., Kumail, M. et al. A draft Arab pangenome reference. *bioRxiv*. 13 July 2024, preprint: not peer reviewed https://doi.org/10.1101/2024.07.09.602638
85. Littlefield, C., Lazaro-Guevara, J.M., Stucki, D. et al. A Draft Pacific Ancestry Pangenome Reference. *bioRxiv*. 09 August 2024, preprint: not peer reviewed https://doi.org/10.1101/2024.08.07.606392
86. Church, G.M. The personal genome project. *Mol Syst Biol*. 2005; 1: 2005 0030. https://doi.org/10.1038/msb4100040
87. Ball, M.P., Bobe, J.R., Chou, M.F. et al. Harvard Personal Genome Project: lessons from participatory public research. *Genome Med*. 2014; 6: 10. https://doi.org/10.1186/gm527
88. Lunshof, J.E., Bobe, J., Aach, J. et al. Personal genomes in progress: from the human genome project to the personal genome project. *Dialogues Clin Neurosci*. 2010; 12: 47–60. https://doi.org/10.31887/DCNS.2010.12.1/jlunshof
89. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467: 1061–1073. https://doi.org/10.1038/nature09534
90. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491: 56–65. https://doi.org/10.1038/nature11632
91. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015; 526: 68–74. https://doi.org/10.1038/nature15393
92. Siva, N. 1000 Genomes project. *Nat Biotechnol*. 2008; 26: 256. https://doi.org/10.1038/nbt0308-256b
93. National Center for Biotechnology Information. About dbSNP. https://www.ncbi.nlm.nih.gov/snp/docs/about/, (26 May 2025, date last accessed).
94. The Genome Aggregation Database. About gnomAD. https://gnomad.broadinstitute.org/about, (26 May 2025, date last accessed).
95. The Genome Aggregation Database. What's in gnomAD. https://gnomad.broadinstitute.org/stats, (26 May 2025, date last accessed).
96. Lappalainen, I., Lopez, J., Skipper, L. et al. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res*. 2013; 41: D936–941. https://doi.org/10.1093/nar/gks1213
97. National Institutes of Health. NCBI launches the Database of Genomic Structural Variations. https://www.nih.gov/news-events/news-releases/ncbi-launches-database-genomic-structural-variations, (26 May 2025, date last accessed).
98. National Center for Biotechnology Information. dbVar Variant Summary. https://www.ncbi.nlm.nih.gov/dbvar/content/var_summary/, (26 May 2025, date last accessed).
99. Ozaki, K., Ohnishi, Y., Iida, A. et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat Genet*. 2002; 32: 650–654. https://doi.org/10.1038/ng1047
100. Ikegawa, S. A short history of the genome-wide association study: where we were and where we are going. *Genomics Inform*. 2012; 10: 220–225. https://doi.org/10.5808/GI.2012.10.4.220
101. Thomas, D.C., Haile, R.W. and Duggan, D. Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet*. 2005; 77: 337–345. https://doi.org/10.1086/432962
102. Kurki, M.I., Karjalainen, J., Palta, P. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023; 613: 508–518. https://doi.org/10.1038/s41586-022-05473-8
103. Savage, L.J., Wittmann, M., McGonagle, D. et al. Ustekinumab in the Treatment of Psoriasis and Psoriatic Arthritis. *Rheumatol Ther*. 2015; 2: 1–16. https://doi.org/10.1007/s40744-015-0010-2
104. Reay, W.R. and Cairns, M.J. Advancing the use of genome-wide association studies for drug repurposing. *Nat Rev Genet*. 2021; 22: 658–671. https://doi.org/10.1038/s41576-021-00387-z
105. Duerr, R.H., Taylor, K.D., Brant, S.R. et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*. 2006; 314: 1461–1463. https://doi.org/10.1126/science.1135245
106. Sandborn, W.J., Feagan, B.G., Fedorak, R.N. et al. A randomized trial of Ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with moderate-to-severe Crohn's disease. *Gastroenterology*. 2008; 135: 1130–1141. https://doi.org/10.1053/j.gastro.2008.07.014
107. Sandborn, W.J., Gasink, C., Gao, L.L. et al. Ustekinumab induction and maintenance therapy in refractory Crohn's disease. *N Engl J Med*. 2012; 367: 1519–1528. https://doi.org/10.1056/NEJMoa1203572

 doi:10.20944/preprints202605.0651.v1

108. Feagan, B.G., Sandborn, W.J., Gasink, C. et al. Ustekinumab as Induction and Maintenance Therapy for Crohn's Disease. *N Engl J Med*. 2016; 375: 1946–1960. https://doi.org/10.1056/NEJMoa1602773
109. Hong, E.P. and Park, J.W. Sample size and statistical power calculation in genetic association studies. *Genomics Inform*. 2012; 10: 117–122. https://doi.org/10.5808/GI.2012.10.2.117
110. Wu, Y., Li, R., Sun, S. et al. Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet*. 2021; 108: 1891–1906. https://doi.org/10.1016/j.ajhg.2021.08.012
111. Yang, J., Benyamin, B., McEvoy, B.P. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010; 42: 565–569. https://doi.org/10.1038/ng.608
112. Brandes, N., Weissbrod, O. and Linial, M. Open problems in human trait genetics. *Genome Biol*. 2022; 23: 131. https://doi.org/10.1186/s13059-022-02697-9
113. Speliotes, E.K., Willer, C.J., Berndt, S.I. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*. 2010; 42: 937–948. https://doi.org/10.1038/ng.686
114. Lango Allen, H., Estrada, K., Lettre, G. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010; 467: 832–838. https://doi.org/10.1038/nature09410
115. Elks, C.E., den Hoed, M., Zhao, J.H. et al. Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol (Lausanne)*. 2012; 3: 29. https://doi.org/10.3389/fendo.2012.00029
116. Macgregor, S., Cornes, B.K., Martin, N.G. et al. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum Genet*. 2006; 120: 571–580. https://doi.org/10.1007/s00439-006-0240-z
117. Yang, J., Manolio, T.A., Pasquale, L.R. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*. 2011; 43: 519–525. https://doi.org/10.1038/ng.823
118. Wunnemann, F., Fotsing Tadjo, T., Beaudoin, M. et al. Multimodal CRISPR perturbations of GWAS loci associated with coronary artery disease in vascular endothelial cells. *PLoS Genet*. 2023; 19: e1010680. https://doi.org/10.1371/journal.pgen.1010680
119. Phillips, P.C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*. 2008; 9: 855–867. https://doi.org/10.1038/nrg2452
120. Nagel, R.L. Epistasis and the genetics of human diseases. *C R Biol*. 2005; 328: 606–615. https://doi.org/10.1016/j.crvi.2005.05.003
121. White, D. and Rabago-Smith, M. Genotype-phenotype associations and human eye color. *J Hum Genet*. 2011; 56: 5–7. https://doi.org/10.1038/jhg.2010.126
122. Kim, S. and Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol Cell*. 2023; 83: 373–392. https://doi.org/10.1016/j.molcel.2022.12.032
123. Atkinson, T.J. and Halfon, M.S. Regulation of gene expression in the genomic context. *Comput Struct Biotechnol J*. 2014; 9: e201401001. https://doi.org/10.5936/csbj.201401001
124. Harris, Z.S. Distributional Structure. *WORD*. 1954; 10: 146–162. https://doi.org/10.1080/00437956.1954.11659520
125. Yu, L., Tanwar, D.K., Penha, E.D.S. et al. Grammar of protein domain architectures. *Proc Natl Acad Sci U S A*. 2019; 116: 3636–3645. https://doi.org/10.1073/pnas.1814684116
126. Nagy, G. and Nagy, L. Motif grammar: The basis of the language of gene expression. *Comput Struct Biotechnol J*. 2020; 18: 2026–2032. https://doi.org/10.1016/j.csbj.2020.07.007
127. Searls, D.B. The Linguistics of DNA. *American Scientist*. 1992; 80: 579–591.
128. Searls, D.B. The language of genes. *Nature*. 2002; 420: 211–217. https://doi.org/10.1038/nature01255
129. Ji, S. The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Ann N Y Acad Sci*. 1999; 870: 411–417. https://doi.org/10.1111/j.1749-6632.1999.tb08916.x
130. Kwon, J.J., Pan, J., Gonzalez, G. et al. On knowing a gene: A distributional hypothesis of gene function. *Cell Syst*. 2024; 15: 488–496. https://doi.org/10.1016/j.cels.2024.04.008
131. Ofer, D., Brandes, N. and Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J*. 2021; 19: 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022
132. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods*. 2006; 3: 17–21. https://doi.org/10.1038/nmeth823

 doi:10.20944/preprints202605.0651.v1

133. Rister, J. and Desplan, C. Deciphering the genome's regulatory code: the many languages of DNA. *Bioessays*. 2010; 32: 381–384. https://doi.org/10.1002/bies.200900197
134. Wray, G.A., Hahn, M.W., Abouheif, E. et al. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*. 2003; 20: 1377–1419. https://doi.org/10.1093/molbev/msg140
135. Dynan, W.S. Modularity in promoters and enhancers. *Cell*. 1989; 58: 1–4. https://doi.org/10.1016/0092-8674(89)90393-0
136. Arnone, M.I. and Davidson, E.H. The hardwiring of development: organization and function of genomic regulatory systems. *Development*. 1997; 124: 1851–1864. https://doi.org/10.1242/dev.124.10.1851
137. Friedman, M.J., Wagner, T., Lee, H. et al. Enhancer-promoter specificity in gene transcription: molecular mechanisms and disease associations. *Exp Mol Med*. 2024; 56: 772–787. https://doi.org/10.1038/s12276-024-01233-y
138. Horton, C.A., Alexandari, A.M., Hayes, M.G.B. et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science*. 2023; 381: eadd1250. https://doi.org/10.1126/science.add1250
139. Erceg, J., Saunders, T.E., Girardot, C. et al. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLoS Genet*. 2014; 10: e1004060. https://doi.org/10.1371/journal.pgen.1004060
140. Vaswani, A., Shazeer, N., Parmar, N. et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017; 30:
141. Boshar, S., Trop, E., de Almeida, B.P. et al. Are genomic language models all you need? Exploring genomic language models on protein downstream tasks. *Bioinformatics*. 2024; 40: https://doi.org/10.1093/bioinformatics/btae529
142. Devlin, J., Chang, M.-W., Lee, K. et al. Bert: pre-training of deep bidirectional transformers for language understanding. *NACCL*. 2019; 1: 4171–4186.
143. Solan, Z., Horn, D., Ruppin, E. et al. Unsupervised learning of natural languages. *Proc Natl Acad Sci U S A*. 2005; 102: 11629–11634. https://doi.org/10.1073/pnas.0409746102
144. Radford, A., Wu, J., Child, R. et al. Language models are unsupervised multitask learners. 2019;
145. Ji, Y., Zhou, Z., Liu, H. et al. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*. 2021; 37: 2112–2120. https://doi.org/10.1093/bioinformatics/btab083
146. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods*. 2025; 22: 287–297. https://doi.org/10.1038/s41592-024-02523-z
147. de Almeida, B.P., Dalla-Torre, H., Richard, G. et al. SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models. *bioRxiv*. 15 March 2024, preprint: not peer reviewed https://doi.org/10.1101/2024.03.14.584712
148. Hwang, Y., Cornman, A.L., Kellogg, E.H. et al. Genomic language model predicts protein co-regulation and function. *Nat Commun*. 2024; 15: 2880. https://doi.org/10.1038/s41467-024-46947-9
149. Vig, J., Madani, A., Varshney, L.R. et al. BERTology meets biology: interpreting attention in protein language models. *bioRxiv*. 13 July 2020, preprint: not peer reviewed https://doi.org/10.1101/2020.06.26.174417
150. Radford, A., Narasimhan, K., Salimans, T. et al. Improving language understanding by generative pre-training. 2018;
151. Hayes, T., Rao, R., Akin, H. et al. Simulating 500 million years of evolution with a language model. *Science*. 2025; 387: 850–858. https://doi.org/10.1126/science.ads0018
152. Brixi, G., Durrant, M.G., Ku, J. et al. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*. 21 February 2025, preprint: not peer reviewed https://doi.org/10.1101/2025.02.18.638918
153. Woese, C. The universal ancestor. *Proc Natl Acad Sci U S A*. 1998; 95: 6854–6859. https://doi.org/10.1073/pnas.95.12.6854
154. Villar, D., Berthelot, C., Aldridge, S. et al. Enhancer evolution across 20 mammalian species. *Cell*. 2015; 160: 554–566. https://doi.org/10.1016/j.cell.2015.01.006

 doi:10.20944/preprints202605.0651.v1

155. Moody, E.R.R., Alvarez-Carretero, S., Mahendrarajah, T.A. et al. The nature of the last universal common ancestor and its impact on the early Earth system. *Nat Ecol Evol*. 2024; 8: 1654–1666. https://doi.org/10.1038/s41559-024-02461-1
156. Andrews, G., Fan, K., Pratt, H.E. et al. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science*. 2023; 380: eabn7930. https://doi.org/10.1126/science.abn7930
157. Bejerano, G., Pheasant, M., Makunin, I. et al. Ultraconserved elements in the human genome. *Science*. 2004; 304: 1321–1325. https://doi.org/10.1126/science.1098119
158. Elnaggar, A., Heinzinger, M., Dallago, C. et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell*. 2022; 44: 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381
159. Nguyen, E., Poli, M., Durrant, M.G. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*. 2024; 386: eado9336. https://doi.org/10.1126/science.ado9336
160. Rives, A., Meier, J., Sercu, T. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*. 2021; 118: https://doi.org/10.1073/pnas.2016239118
161. Lin, Z., Akin, H., Rao, R. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023; 379: 1123–1130. https://doi.org/10.1126/science.ade2574
162. Dill, K.A. and MacCallum, J.L. The protein-folding problem, 50 years on. *Science*. 2012; 338: 1042–1046. https://doi.org/10.1126/science.1219021
163. Dill, K.A., Ozkan, S.B., Shell, M.S. et al. The protein folding problem. *Annu Rev Biophys*. 2008; 37: 289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558
164. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science*. 1973; 181: 223–230. https://doi.org/10.1126/science.181.4096.223
165. Kovalevskiy, O., Mateos-Garcia, J. and Tunyasuvunakool, K. AlphaFold two years on: Validation and impact. *Proc Natl Acad Sci U S A*. 2024; 121: e2315002121. https://doi.org/10.1073/pnas.2315002121
166. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021; 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2
167. Kryshtafovych, A., Schwede, T., Topf, M. et al. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins*. 2021; 89: 1607–1617. https://doi.org/10.1002/prot.26237
168. Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins*. 2004; 57: 702–710. https://doi.org/10.1002/prot.20264
169. Haas, J., Barbato, A., Behringer, D. et al. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*. 2018; 86: 387–398. https://doi.org/10.1002/prot.25431
170. Michaud, J.M., Madani, A. and Fraser, J.S. A language model beats alphafold2 on orphans. *Nat Biotechnol*. 2022; 40: 1576–1577. https://doi.org/10.1038/s41587-022-01466-0
171. Chowdhury, R., Bouatta, N., Biswas, S. et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol*. 2022; 40: 1617–1623. https://doi.org/10.1038/s41587-022-01432-w
172. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J. et al. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. 2017; 35: 128–135. https://doi.org/10.1038/nbt.3769
173. Gobel, U., Sander, C., Schneider, R. et al. Correlated mutations and residue contacts in proteins. *Proteins*. 1994; 18: 309–317. https://doi.org/10.1002/prot.340180402
174. Zhang, Z., Wayment-Steele, H.K., Brixi, G. et al. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc Natl Acad Sci U S A*. 2024; 121: e2406285121. https://doi.org/10.1073/pnas.2406285121
175. Brandes, N., Goldman, G., Wang, C.H. et al. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet*. 2023; 55: 1512–1522. https://doi.org/10.1038/s41588-023-01465-0
176. Kircher, M., Witten, D.M., Jain, P. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014; 46: 310–315. https://doi.org/10.1038/ng.2892
177. Rentzsch, P., Witten, D., Cooper, G.M. et al. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019; 47: D886–D894. https://doi.org/10.1093/nar/gky1016

 doi:10.20944/preprints202605.0651.v1

178. Zhou, Z., Ji, Y., Li, W. et al. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *ArXiv*. 26 June 2023, preprint: not peer reviewed https://doi.org/10.48550/arXiv.2306.15006
179. Stenson, P.D., Ball, E.V., Mort, M. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*. 2003; 21: 577–581. https://doi.org/10.1002/humu.10212
180. Benegas, G., Albors, C., Aw, A.J. et al. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nat Biotechnol*. 2025; https://doi.org/10.1038/s41587-024-02511-w
181. Tate, J.G., Bamford, S., Jubb, H.C. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019; 47: D941–D947. https://doi.org/10.1093/nar/gky1015
182. Amberger, J.S., Bocchini, C.A., Scott, A.F. et al. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. 2019; 47: D1038–D1043. https://doi.org/10.1093/nar/gky1151
183. Cheng, J., Novati, G., Pan, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023; 381: eadg7492. https://doi.org/10.1126/science.adg7492
184. Sinnott-Armstrong, N., Fields, S., Roth, F. et al. Understanding genetic variants in context. *Elife*. 2024; 13: https://doi.org/10.7554/eLife.88231
185. Avsec, Z., Agarwal, V., Visentin, D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021; 18: 1196–1203. https://doi.org/10.1038/s41592-021-01252-x
186. Linder, J., Srivastava, D., Yuan, H. et al. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat Genet*. 2025; 57: 949–961. https://doi.org/10.1038/s41588-024-02053-6
187. Drusinsky, S., Whalen, S. and Pollard, K.S. Deep-learning prediction of gene expression from personal genomes. *bioRxiv*. 27 July 2024, preprint: not peer reviewed https://doi.org/10.1101/2024.07.27.605449
188. Rastogi, R., Reddy, A.J., Chung, R. et al. Fine-tuning sequence-to-expression models on personal genome and transcriptome data. *bioRxiv*. 25 September 2024, preprint: not peer reviewed https://doi.org/10.1101/2024.09.23.614632
189. National Center for Biotechnology Information. Genome assembly Ash1.7. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_011064465.1/, (26 May 2025, date last accessed).
190. National Center for Biotechnology Information. Genome assembly KOREF1.0. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_001712695.1/, (26 May 2025, date last accessed).
191. National Center for Biotechnology Information. Genome assembly KOREF_S1v2.1. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_020497085.1/, (26 May 2025, date last accessed).
192. National Center for Biotechnology Information. Genome assembly ASM3717755v1. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037177555.1/, (26 May 2025, date last accessed).
193. National Center for Biotechnology Information. Genome assembly ASM3717763v1. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037177635.1/, (26 May 2025, date last accessed).
194. Personal Genome Project. Public genetic data. https://my.pgp-hms.org/public_genetic_data/statistics, (26 May 2025, date last accessed).
195. All of Us RESEARCH PROGRAM. All of Us Research Program Staff. https://allofus.nih.gov/about/who-we-are/nih-all-us-research-program-staff, (26 May 2025, date last accessed).
196. All of Us Research Hub. Data Browser. https://databrowser.researchallofus.org/, (26 May 2025, date last accessed).
197. UK biobank. About us. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us, (26 May 2025, date last accessed).
198. UK biobank. Celebrating 15 years of UK Biobank. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/our-impact/celebrating-15-years-of-uk-biobank, (26 May 2025, date last accessed).
199. Global Alliance for Genomics & Health. Trans-Omics for Precision Medicine (TOPMed). https://www.ga4gh.org/driver_project/trans-omics-for-precision-medicine-topmed/, (26 May 2025, date last accessed).
200. National Heart, L., and Blood Institute,. Trans-Omics for Precision Medicine (TOPMed) Program. https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program, (26 May 2025, date last accessed).
201. National Heart, L., and Blood Institute,. Trans-Omics for Precision Medicine. https://topmed.nhlbi.nih.gov/, (26 May 2025, date last accessed).

 doi:10.20944/preprints202605.0651.v1

202. Mallick, S., Li, H., Lipson, M. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016; 538: 201–206. https://doi.org/10.1038/nature18964
203. Nagasaki, M., Yasuda, J., Katsuoka, F. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun*. 2015; 6: 8018. https://doi.org/10.1038/ncomms9018
204. Tadaka, S., Katsuoka, F., Ueki, M. et al. 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. *Hum Genome Var*. 2019; 6: 28. https://doi.org/10.1038/s41439-019-0059-5
205. Fuse, N., Sakurai-Yageta, M., Katsuoka, F. et al. Establishment of Integrated Biobank for Precision Medicine and Personalized Healthcare: The Tohoku Medical Megabank Project. *JMA J*. 2019; 2: 113–122. https://doi.org/10.31662/jmaj.2019-0014
206. Ogishima, S., Nagaie, S., Mizuno, S. et al. dbTMM: an integrated database of large-scale cohort, genome and clinical data for the Tohoku Medical Megabank Project. *Hum Genome Var*. 2021; 8: 44. https://doi.org/10.1038/s41439-021-00175-5
207. Hozawa, A., Tanno, K., Nakaya, N. et al. Study Profile of the Tohoku Medical Megabank Community-Based Cohort Study. *J Epidemiol*. 2021; 31: 65–76. https://doi.org/10.2188/jea.JE20190271
208. TOHOKU MEDICAL MEGABANK ORGANIZATION. Organization and Members. https://www.megabank.tohoku.ac.jp/english/about/member/, (26 May 2025, date last accessed).
209. TOHOKU MEDICAL MEGABANK ORGANIZATION. The Whole Genome Sequence of 100,000 Japanese General Population Has Completed -Largest in Asian Population, One of the World's Leading-. https://www.megabank.tohoku.ac.jp/english/the-whole-genome-sequence-of-100000-japanese-general-population-has-completed-largest-in-asian-population-one-of-the-worlds-leading/, (26 May 2025, date last accessed).
210. TOHOKU MEDICAL MEGABANK ORGANIZATION. Mission and Outline. https://www.megabank.tohoku.ac.jp/english/about/outline/, (26 May 2025, date last accessed).
211. Cao, Y., Li, L., Xu, M. et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res*. 2020; 30: 717–731. https://doi.org/10.1038/s41422-020-0322-9
212. mBiobank. China Metabolic Analytics Project. http://www.mbiobank.com/info/, (26 May 2025, date last accessed).
213. Westlake BioBank for Chinese (WBBC). Research Team. https://wbbc.westlake.edu.cn/research.html, (26 May 2025, date last accessed).
214. Westlake BioBank for Chinese (WBBC). Home. https://wbbc.westlake.edu.cn/index.html, (26 May 2025, date last accessed).
215. Genomics England. The journey to 100,000 genomes. https://www.genomicsengland.co.uk/news/journey-to-100000-genomes, (26 May 2025, date last accessed).
216. Genomics England. 100,000 Genomes Project. https://www.genomicsengland.co.uk/initiatives/100000-genomes-project, (26 May 2025, date last accessed).
217. Wu, D., Dou, J., Chai, X. et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell*. 2019; 179: 736–749 e715. https://doi.org/10.1016/j.cell.2019.09.019
218. Wong, E., Bertin, N., Hebrard, M. et al. The Singapore National Precision Medicine Strategy. *Nat Genet*. 2023; 55: 178–186. https://doi.org/10.1038/s41588-022-01274-x
219. Bellis, C., Kolle, G., Yong, J. et al. National Scale Genomic Engine for Precision Medicine: Singapore PRECISE-SG100K Experience. *bioRxiv*. 15 March 2025, preprint: not peer reviewed https://doi.org/10.1101/2025.03.13.642552
220. GenomeAsia 100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 2019; 576: 106–111. https://doi.org/10.1038/s41586-019-1793-z
221. Boomsma, D.I., Wijmenga, C., Slagboom, E.P. et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet*. 2014; 22: 221–227. https://doi.org/10.1038/ejhg.2013.118
222. GENOME of the NETHERLANDS. The Genome of the Netherlands. https://www.nlgenome.nl/menu/main/home, (26 May 2025, date last accessed).
223. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014; 46: 818–825. https://doi.org/10.1038/ng.3021

 doi:10.20944/preprints202605.0651.v1

224. The UK 10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature*. 2015; 526: 82–90. https://doi.org/10.1038/nature14962
225. Muddyman, D., Smee, C., Griffin, H. et al. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med*. 2013; 5: 100. https://doi.org/10.1186/gm504
226. Kaye, J., Hurles, M., Griffin, H. et al. Managing clinically significant findings in research: the UK10K example. *Eur J Hum Genet*. 2014; 22: 1100–1104. https://doi.org/10.1038/ejhg.2013.290
227. UK10K. UK10K Consortium Membership. https://www.uk10k.org/consortium.html, (26 May 2025, date last accessed).
228. Amer, K., Soliman, N.A., Soror, S. et al. Egypt Genome: Towards an African new genomic era. *J Adv Res*. 2025; 71: 415–427. https://doi.org/10.1016/j.jare.2024.06.003
229. Elmonem, M.A., Soliman, N.A., Moustafa, A. et al. The Egypt Genome Project. *Nat Genet*. 2024; 56: 1035–1037. https://doi.org/10.1038/s41588-024-01739-1
230. Ameur, A., Dahlberg, J., Olason, P. et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet*. 2017; 25: 1253–1260. https://doi.org/10.1038/ejhg.2017.130