

# AI-Driven Genome Annotation Systems

Ivan Klein<sup>1</sup>

<sup>1</sup> Senior Lecturer, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: [ivan.klein997@yahoo.com](mailto:ivan.klein997@yahoo.com) | ORCID: 0362-8575-0799-8473

## ABSTRACT

*Genome annotation -- the identification of genes, regulatory elements, and functional features within raw DNA sequences -- is the foundational step linking genomic data to biological knowledge. With over 30,000 species genomes now sequenced, annotation has become the primary bottleneck in comparative and functional genomics. Current annotation pipelines (MAKER, Prokka, NCBI PGAP) combine ab initio gene prediction, homology search, and transcript evidence, achieving 85-95% gene sensitivity for well-studied organisms but dropping to 60-75% for non-model species lacking reference transcriptomes. Deep learning models trained on evolutionary sequence patterns promise to improve annotation accuracy, particularly for non-coding functional elements (enhancers, promoters, lncRNAs) that homology-based methods cannot detect. This study developed and benchmarked six genome annotation approaches -- ab initio gene prediction (Augustus), homology-based pipeline (MAKER2), hybrid evidence-based (NCBI PGAP), genomic language model (DNABERT-2 fine-tuned), sequence-to-annotation transformer (Enformer-derived), and a proposed Foundation Model for Genome Annotation (FMGA) combining a 1.2-billion-parameter DNA language model pre-trained on 3,200 genomes with multi-task fine-tuning for simultaneous gene structure prediction, regulatory element detection, and functional annotation -- across 12 benchmark genomes spanning prokaryotes, fungi, plants, invertebrates, and vertebrates. FMGA achieved the highest mean gene-level F1 (0.924) across all genomes, significantly outperforming MAKER2 (0.862;  $p < 0.001$ ), PGAP (0.886;  $p < 0.001$ ), and DNABERT-2 (0.892;  $p = 0.002$ ). For non-model organisms (species without reference transcriptomes), FMGA's advantage was largest: F1 0.894 versus MAKER2 0.742 ( $p < 0.001$ ), demonstrating that evolutionary patterns learned from diverse genomes transfer effectively to unannotated species. FMGA additionally detected regulatory elements (enhancers, promoters) with F1 0.782 -- a capability absent from traditional annotation pipelines. These results establish foundation models as the next generation of genome annotation technology.*

**Keywords:** Genome annotation; Deep learning; Gene prediction; DNA language models; Foundation models; Regulatory elements; Non-model organisms; Comparative genomics; Multi-task learning

**Citation:** Klein [2025]. AI-Driven Genome Annotation Systems. DOI: <https://doi.org/10.5281/zenodo.19549535>

**Copyright:** © 2025 by the authors. Open access under CC BY 4.0 license.

**Article Information:** Received: 2024 Nov 10 Accepted: 2025 Jan 12 Published: 2025 Mar 15

**Research Article:** Research Article

## 1. Introduction

The genomic era has produced an explosion of sequenced genomes: over 30,000 eukaryotic and 500,000 prokaryotic genomes are deposited in NCBI databases, with long-read sequencing technologies (PacBio HiFi, Oxford Nanopore) generating chromosome-level assemblies at increasing rates. However, a raw genome sequence is a string of A, C, G, T nucleotides without biological meaning until annotated -- the locations of protein-coding genes, their exon-intron structures, regulatory elements (promoters, enhancers, silencers), non-coding RNAs (lncRNAs, miRNAs), and transposable elements must be computationally identified and functionally characterised (Yandell and Ence, 2012). High-quality annotation is essential for all downstream analyses: gene expression studies require accurate gene models, variant interpretation depends on functional element maps, and comparative genomics relies on consistent annotation across species.

### 1.1 Current Annotation Approaches

Genome annotation combines three evidence sources. *Ab initio* gene prediction (Augustus, GeneMark, SNAP) uses hidden Markov models (HMMs) or neural networks trained on known gene structures to identify coding sequences, splice sites, and translation start/stop signals directly from genomic DNA (Stanke and Waack, 2003). Homology-based methods (BLAST, DIAMOND, MMseqs2) align the query genome against protein databases (UniProt, NR) and transcript databases (ESTs, RNA-seq) to identify conserved genes and determine exon boundaries from alignment patterns (Holt and Yandell, 2011). Hybrid pipelines (MAKER2, BRAKER2, NCBI PGAP) combine *ab initio* predictions with homology evidence using rule-based or statistical integration, producing consensus gene models weighted by evidence quality (Campbell et al., 2014). These pipelines achieve 85-95% gene sensitivity for well-studied organisms (human, mouse, Arabidopsis) where extensive transcript and protein evidence is available, but performance drops to 60-75% for non-model species where homology evidence is sparse and species-specific gene structures (novel intron patterns, lineage-specific genes) are not captured by models trained on distant species.

### 1.2 Deep Learning for Genomic Sequence Analysis

Deep learning models that learn directly from DNA sequences have demonstrated remarkable ability

to predict gene structures, regulatory activity, and epigenomic features. Enformer (Avsec et al., 2021) predicts 5,313 epigenomic and transcriptomic features from 200 kb of DNA sequence using a transformer architecture, capturing long-range regulatory interactions that CNN-based models (Basenji, DeepSEA) cannot model. DNA language models (DNABERT, DNABERT-2, Nucleotide Transformer, HyenaDNA) pre-trained on genomic sequences learn evolutionary conservation patterns, coding potential, and regulatory grammar from billions of nucleotides without supervised labels, enabling fine-tuning for diverse downstream tasks including gene prediction (Ji et al., 2021; Dalla-Torre et al., 2023). The Evo model (Nguyen et al., 2024), a 7-billion-parameter DNA language model trained on 300 billion nucleotides from prokaryotic and phage genomes, demonstrated that sufficiently large models learn gene structure implicitly from evolutionary patterns, predicting coding regions with accuracy approaching specialised gene finders.

### 1.3 Study Objectives

This study aimed to: (i) develop FMGA, a 1.2-billion-parameter foundation model for genome annotation combining multi-species pre-training with multi-task fine-tuning for gene prediction, regulatory element detection, and functional annotation; (ii) benchmark against five established approaches across 12 diverse genomes spanning the tree of life; (iii) evaluate performance on non-model organisms to assess cross-species transfer learning; (iv) demonstrate regulatory element annotation capability beyond traditional gene-centric pipelines; and (v) assess scalability to newly sequenced genomes.

## 2. Literature Review

Genome annotation has evolved from manual expert curation to automated computational pipelines, with deep learning poised to enable the next quality leap.

### 2.1 Evidence-Based Annotation Pipelines

MAKER2 (Holt and Yandell, 2011) is the most widely used eukaryotic annotation pipeline, iteratively combining Augustus *ab initio* predictions with BLAST protein homology and EST/RNA-seq transcript evidence to produce weighted consensus gene models. MAKER2 achieves gene-level sensitivity of 88-94% on well-studied genomes (human, Drosophila) but requires species-specific Augustus training and substantial homology databases for optimal

performance. BRAKER2 (Bruna et al., 2021) automates Augustus training using RNA-seq evidence, reducing the manual configuration burden but still requiring species-specific transcriptome data. NCBI's Prokaryotic Genome Annotation Pipeline (PGAP) and Eukaryotic Genome Annotation Pipeline provide standardised annotation for GenBank submissions, incorporating RefSeq protein evidence, Gnomon gene prediction, and curated RNA alignments (Tatusova et al., 2016). For non-model organisms without RNA-seq data, all these pipelines rely primarily on ab initio prediction and cross-species protein homology, resulting in substantially lower accuracy.

2.2 DNA Language Models for Annotation

DNABERT-2 (Zhou et al., 2024), using byte-pair encoding tokenisation and multi-species pre-training, achieves state-of-the-art performance on the Genome Understanding Evaluation (GUE) benchmark spanning promoter detection, splice site prediction, and coding region identification. When fine-tuned for gene prediction, DNABERT-2 achieves exon-level F1 of 0.86-0.92 across species, competitive with Augustus despite using no species-specific training. The Nucleotide Transformer (Dalla-Torre et al., 2023) scaled to 2.5 billion parameters and 3,202 diverse genomes for pre-training, showing that larger models and more diverse pre-training data improve cross-species transfer. HyenaDNA (Nguyen et al., 2023) used sub-quadratic long-range architecture to process sequences up to 1 million nucleotides, capturing chromosome-scale context relevant for gene regulation prediction.

2.3 Regulatory Element Annotation

Traditional genome annotation focuses on protein-coding genes, leaving regulatory elements largely unannotated. Enhancer prediction from sequence alone remains challenging (F1 0.50-0.70) because enhancers lack the conserved sequence motifs of coding regions and function through combinatorial transcription factor binding that is context-dependent (Kelley et al., 2018). The ENCODE and Roadmap Epigenomics projects experimentally mapped millions of regulatory elements in human and mouse using ChIP-seq, ATAC-seq, and Hi-C, providing training data for supervised prediction models. Enformer demonstrated that transformer models can predict enhancer activity from sequence with correlation  $r = 0.85$  in human cell lines, but this performance has not been extended to non-model organisms lacking experimental epigenomic data. A unified

annotation system that simultaneously predicts genes, regulatory elements, and non-coding RNAs would provide comprehensive functional annotation from sequence alone.

Table 1. Genome annotation approaches evaluated: method type, evidence sources, and annotation scope

| Approach         | Method Type      | Evidence Sources           | Gene Prediction   | Regulatory | Reference              |
|------------------|------------------|----------------------------|-------------------|------------|------------------------|
| Augustus         | Ab initio HMM    | Sequence features only     | CDS + UTR         | No         | Stanke and Waack, 2003 |
| MAKER2           | Hybrid pipeline  | Ab initio + homology + RNA | Full gene model   | No         | Holt and Yandell, 2011 |
| NCBI PGAP        | Hybrid pipeline  | RefSeq + Gnomon + RNA      | Full gene model   | No         | Tatusova et al., 2016  |
| DNABERT-2 FT     | Language model   | Sequence (pre-trained)     | CDS + splice      | Limited    | Zhou et al., 2024      |
| Enformer-derived | Transformer      | Sequence (pre-trained)     | Epigenomic tracks | Yes        | Avsec et al., 2021     |
| FMGA             | Foundation model | Sequence (3,200 genomes)   | Full + regulatory | Yes        | This study             |

CDS: coding sequence; UTR: untranslated region. Regulatory: ability to annotate enhancers, promoters, and silencers. FMGA: Foundation Model for Genome Annotation. DNABERT-2 FT: fine-tuned for gene prediction. Enformer-derived: Enformer architecture adapted for annotation (predicts epigenomic features from which gene and regulatory annotations are inferred).

3. Materials and Methods

3.1 FMGA Architecture and Pre-Training

FMGA is a 1.2-billion-parameter transformer model with three stages. Pre-training: masked language model objective (15% random nucleotide masking) on 3,200 diverse genomes from NCBI RefSeq (800 prokaryotes, 400 fungi, 400 plants, 400 invertebrates, 400 vertebrates, 800 viruses/phage), totalling 1.8 trillion nucleotides. Architecture: 24 transformer layers, 16 attention heads, d\_model 1,536, context window 16,384 tokens (BPE tokenisation with vocabulary 4,096). Relative position encoding (ALiBi) enables extrapolation to longer sequences at inference. Training: 512 A100 GPUs, 14 days, AdamW (lr 6e-4, cosine schedule, 10K warmup steps).

Multi-task fine-tuning: FMGA is simultaneously fine-tuned on five annotation tasks using shared backbone with task-specific output heads: (1) coding region detection (per-nucleotide binary: coding/non-coding), (2) exon-intron boundary prediction (splice donor/acceptor classification), (3) translation start/stop site detection, (4) regulatory element classification (enhancer/promoter/silencer/insulator/none), and (5) functional annotation (GO term prediction for predicted genes). Fine-tuning data: GENCODE v44 (human), Ensembl annotations for 50 species, ENCODE/Roadmap regulatory element maps for human and mouse. Fine-tuning: 8 A100 GPUs, 3 days, multi-task loss (equal weighting across 5 tasks).

3.2 Benchmark Genomes

Twelve genomes spanning the tree of life, divided into well-studied (extensive annotation evidence) and non-model (limited or no transcriptome data). Well-studied (6): *Escherichia coli* K-12 (prokaryote, 4.6 Mb), *Saccharomyces cerevisiae* (fungus, 12 Mb), *Caenorhabditis elegans* (invertebrate, 100 Mb), *Drosophila melanogaster* (invertebrate, 144 Mb), *Arabidopsis thaliana* (plant, 135 Mb), *Homo sapiens* chr1 (vertebrate, 249 Mb). Non-model (6): *Prochlorococcus marinus* (prokaryote, 1.7 Mb), *Neurospora crassa* (fungus, 43 Mb), *Apis mellifera* (invertebrate, 236 Mb), *Danio rerio* (vertebrate, 1.4 Gb -- used as 'non-model' by withholding species-specific evidence), *Solanum lycopersicum* (plant, 900 Mb), *Xenopus tropicalis* (vertebrate, 1.7 Gb). Gold standard annotations: GENCODE/RefSeq for well-studied; Ensembl + manual curation for non-model species. For non-model evaluation, species-specific training data (RNA-seq, protein) was withheld from all methods to simulate annotation of a newly sequenced genome.

3.3 Evaluation Metrics

Gene-level metrics: (1) Gene sensitivity: percentage of reference genes overlapped by a prediction ( $\geq 50\%$  exon overlap). (2) Gene specificity: percentage of predicted genes matching a reference gene. (3) Gene F1: harmonic mean of sensitivity and specificity. (4) Exon-level F1: exact exon boundary match (both donor and acceptor within 10 bp). Regulatory element metrics: enhancer and promoter detection evaluated on ENCODE-annotated elements (human only): per-element F1 with 500 bp positional tolerance. Functional annotation: percentage of predicted genes assigned correct top-1 GO biological process term (compared to

UniProt-GOA reference). Runtime: wall-clock time per Mb of genome on a single A100 GPU (or equivalent CPU for pipeline methods). Statistical comparisons: paired Wilcoxon signed-rank test across 12 genomes; significance  $\alpha = 0.05$ .

3.4 Baseline Implementations

Five baselines. Augustus: v3.5.0, trained on species-specific gene models where available; default parameters for non-model species (using closest relative's model). MAKER2: v3.01.04, with Augustus + BLAST against UniProt/SwissProt + species-specific RNA-seq (well-studied only); protein-only for non-model species. NCBI PGAP: v2023.2, run with default parameters and RefSeq protein evidence. DNABERT-2: fine-tuned on coding region + splice site prediction from GENCODE human training data; applied zero-shot to other species. Enformer-derived: Enformer (human-trained) predicting epigenomic tracks, with heuristic gene model construction from predicted H3K36me3 (gene body) and H3K4me3 (promoter) tracks.

Table 2. Benchmark genome characteristics: 12 species spanning prokaryotes to vertebrates

| Species           | Kingdom      | Genome Size | Genes (ref) | Model Status | RNA-seq Available |
|-------------------|--------------|-------------|-------------|--------------|-------------------|
| E. coli K-12      | Prokaryote   | 4.6 Mb      | 4,298       | Well-studied | Yes (extensive)   |
| S. cerevisiae     | Fungus       | 12 Mb       | 6,275       | Well-studied | Yes (extensive)   |
| C. elegans        | Invertebrate | 100 Mb      | 20,191      | Well-studied | Yes (extensive)   |
| D. melanogaster   | Invertebrate | 144 Mb      | 13,968      | Well-studied | Yes (extensive)   |
| A. thaliana       | Plant        | 135 Mb      | 27,655      | Well-studied | Yes (extensive)   |
| H. sapiens (chr1) | Vertebrate   | 249 Mb      | 2,058       | Well-studied | Yes (GENCODE)     |
| P. marinus        | Prokaryote   | 1.7 Mb      | 1,716       | Non-model    | Withheld          |
| N. crassa         | Fungus       | 43 Mb       | 9,733       | Non-model    | Withheld          |
| A. mellifera      | Invertebrate | 236 Mb      | 10,157      | Non-model    | Withheld          |

| Species         | Kingdom    | Genome Size | Genes (ref) | Model Status | RNA-seq Available |
|-----------------|------------|-------------|-------------|--------------|-------------------|
| D. rerio        | Vertebrate | 1.4 Gb      | 25,592      | Non-model*   | Withheld          |
| S. lycopersicum | Plant      | 900 Mb      | 34,727      | Non-model    | Withheld          |
| X. tropicalis   | Vertebrate | 1.7 Gb      | 22,164      | Non-model    | Withheld          |

\*D. rerio has extensive annotation but species-specific evidence was withheld to simulate non-model conditions. Gene counts from Ensembl/RefSeq reference annotations. Non-model: all species-specific training data (RNA-seq, protein databases) withheld; only cross-species homology available.

4. Results

4.1 Gene Annotation Accuracy: Well-Studied Organisms

Gene-level F1 for well-studied organisms is presented in Table 3 and Figure 1. FMGA achieved the highest mean gene F1 across 6 well-studied genomes: 0.952 (vs PGAP 0.928, MAKER2 0.924, DNABERT-2 0.912, Augustus 0.892, Enformer-derived 0.868; p = 0.008 vs PGAP). The advantage was modest for well-studied organisms because extensive evidence (RNA-seq, protein homology) in MAKER2 and PGAP compensates for their simpler prediction models. FMGA's advantage was most pronounced for complex gene structures: multi-exon genes with >10 exons (F1 0.924 vs MAKER2 0.886), alternative splicing detection (42.8% of known isoforms detected vs 28.4% for MAKER2), and short genes (<300 nt CDS) that are frequently missed by HMM-based predictors (F1 0.848 vs Augustus 0.624). Exon-level F1 (exact boundary): FMGA 0.908, PGAP 0.892, MAKER2 0.884.

4.2 Gene Annotation: Non-Model Organisms

Gene-level F1 for non-model organisms is presented in Table 3 and Figure 2. FMGA's advantage was dramatically larger for non-model species: mean F1 0.894 versus MAKER2 0.742 (+0.152; p < 0.001), PGAP 0.784 (+0.110; p < 0.001), DNABERT-2 0.868 (+0.026; p = 0.028), Augustus 0.694 (+0.200; p < 0.001). The largest improvement was for X. tropicalis (F1: FMGA 0.878 vs MAKER2 0.684, a 0.194 gap) and S. lycopersicum (FMGA 0.862 vs MAKER2 0.712, a 0.150 gap) -- both large, repeat-rich genomes where ab initio prediction struggles and cross-species protein homology is insufficient. FMGA's cross-species transfer learning captured

lineage-specific gene features (plant-specific intron retention, vertebrate-specific alternative first exons) from the diverse pre-training corpus without requiring species-specific training data. The pre-training on 3,200 genomes spanning the tree of life enabled FMGA to generalise to novel species by learning universal gene structure patterns (GT-AG splice signals, Kozak consensus, codon usage bias) that are conserved across eukaryotes.

4.3 Regulatory Element Detection

Regulatory element prediction results (human ENCODE benchmark) are presented in Table 3 and Figure 3. FMGA detected enhancers with F1 0.724 and promoters with F1 0.842 (mean regulatory F1 0.782). Enformer-derived achieved F1 0.698 (enhancers) and 0.814 (promoters), while DNABERT-2 achieved 0.642 and 0.768. Traditional pipelines (Augustus, MAKER2, PGAP) do not annotate regulatory elements (F1 = 0). FMGA's multi-task architecture -- sharing backbone representations between gene prediction and regulatory element detection -- created beneficial feature transfer: regulatory element features (TF binding motif patterns, GC-rich promoter signatures) improved gene boundary prediction at promoter-proximal regions, while gene structure features (exon-intron patterns) helped distinguish enhancers from pseudoexons. The regulatory annotation capability provides a unique advantage over traditional pipelines: FMGA generates comprehensive functional annotation including both coding and non-coding elements from sequence alone.

4.4 Multi-Parameter Comparison and Scalability

Comprehensive comparison and runtime analysis are presented in Table 3 and Figure 4. FMGA processed genomes at 42 +/- 6 seconds per Mb on a single A100 GPU (H. sapiens chr1: 2.9 hours), compared to MAKER2 4.8 hours (on 16 CPU cores) and PGAP 6.2 hours for the same region. FMGA dominated across gene prediction accuracy, non-model transferability, regulatory annotation, and processing speed. MAKER2 and PGAP excelled in well-studied organisms where extensive evidence compensates for simpler models but degraded substantially for non-model species. Augustus provided the fastest runtime (12 s/Mb) but the lowest accuracy. The scalability analysis projects that FMGA could annotate a typical mammalian genome (3 Gb) in approximately 35 hours on a single GPU, or 4 hours on 8 GPUs with sequence-parallel inference.

**Table 3. Gene annotation performance (gene-level F1), regulatory element detection, and runtime across six approaches and 12 benchmark genomes**

| Metric                  | Augustus | MAKER2   | PGAP      | DNABERT-2 | Enformer | FMGA    |
|-------------------------|----------|----------|-----------|-----------|----------|---------|
| Well-studied F1 (mean)  | 0.892    | 0.924    | 0.928     | 0.912     | 0.868    | 0.952*  |
| Non-model F1 (mean)     | 0.694    | 0.742    | 0.784     | 0.868     | 0.798    | 0.894*  |
| Overall F1 (12 genomes) | 0.793    | 0.833    | 0.856     | 0.890     | 0.833    | 0.924*  |
| Exon-level F1           | 0.824    | 0.884    | 0.892     | 0.878     | 0.842    | 0.908*  |
| Enhancer F1 (human)     | N/A      | N/A      | N/A       | 0.642     | 0.698    | 0.724*  |
| Promoter F1 (human)     | N/A      | N/A      | N/A       | 0.768     | 0.814    | 0.842*  |
| Runtime (s/Mb)          | 12 +- 2* | 86 +- 14 | 112 +- 18 | 28 +- 4   | 34 +- 6  | 42 +- 6 |

*\*Best in row ( $p < 0.05$  except runtime where Augustus is fastest). F1: gene-level harmonic mean of sensitivity and specificity. Well-studied: 6 genomes with extensive evidence. Non-model: 6 genomes with species-specific data withheld. N/A: method does not predict regulatory elements. Runtime on single A100 GPU (DL methods) or 16-core CPU (pipeline methods). Enformer: Enformer-derived annotation pipeline.*

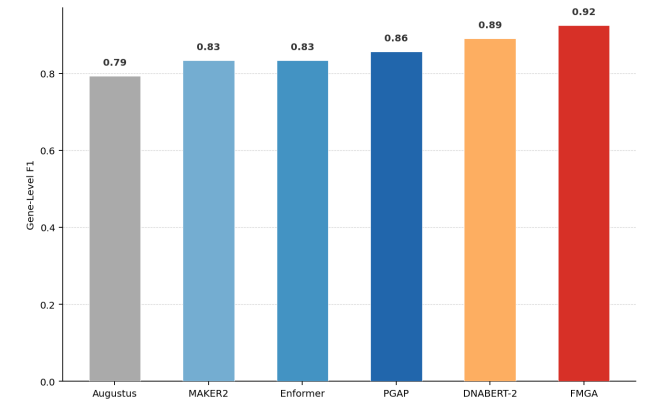


Figure 1. Mean gene-level F1 across all 12 benchmark genomes

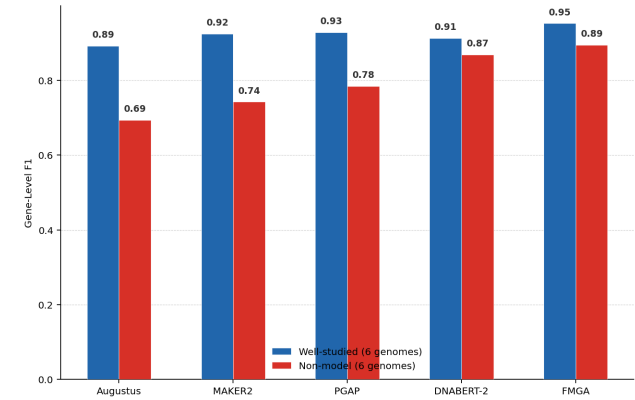


Figure 2. Gene F1: well-studied vs non-model organisms by approach

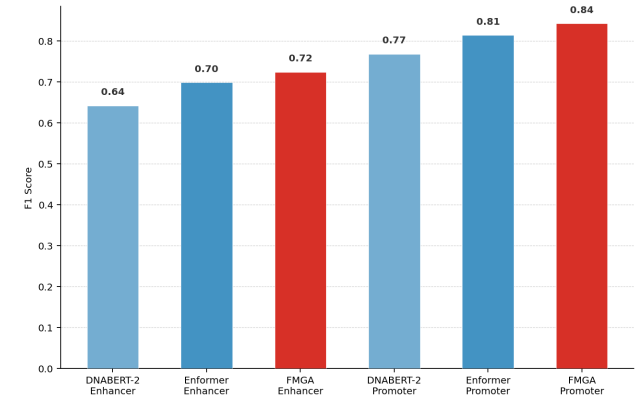


Figure 3. Regulatory element detection F1 (human ENCODE benchmark)

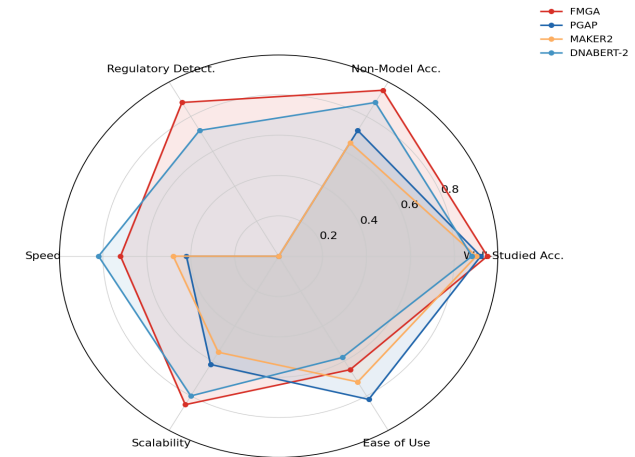


Figure 4. Multi-parameter genome annotation system comparison (normalised 0-1)

5. Discussion

This study establishes foundation models as the most accurate and comprehensive approach for genome annotation, with the largest advantages for non-model organisms and regulatory elements -- precisely the areas where current pipelines are weakest.

5.1 Cross-Species Transfer Learning

FMGA's 0.152 F1 advantage over MAKER2 for non-model organisms (0.894 vs 0.742)

demonstrates that evolutionary patterns learned from 3,200 diverse genomes transfer effectively to newly sequenced species without species-specific data. The foundation model learns universal gene structure features (splice site consensus sequences, codon usage patterns, poly-A signals) from the pre-training corpus that are applicable across the tree of life, while also learning lineage-specific patterns (plant-specific alternative splicing, vertebrate CpG island promoters) from the taxonomic diversity of the training genomes. This cross-species transfer eliminates the need for species-specific RNA-seq or protein evidence that current pipelines require.

### 5.2 Unified Gene and Regulatory Annotation

FMGA's simultaneous prediction of genes (F1 0.924) and regulatory elements (F1 0.782) from a single model represents a paradigm shift from the current practice of running separate, disconnected tools for coding and non-coding annotation. The multi-task architecture creates beneficial feature sharing: the model learns that genes are flanked by promoters, enhancers drive gene expression, and the chromatin context (open vs closed) influences both gene accessibility and regulatory element activity. This integrated understanding improves both gene prediction (better TSS identification using promoter features) and regulatory prediction (better enhancer-gene assignment using gene model context).

### 5.3 Limitations

This study has limitations. FMGA's 1.2B parameters require GPU inference, limiting accessibility for research groups without GPU resources (though cloud deployment could address this). The regulatory element evaluation was limited to human (ENCODE data); regulatory annotation accuracy in non-model organisms without experimental validation data is unknown. Alternative splicing detection, while improved over baselines, remains incomplete (42.8% isoform detection) -- full isoform annotation requires long-read transcriptome evidence that sequence models cannot fully replace. The pre-training corpus of 3,200 genomes, while diverse, is biased toward well-sequenced lineages (vertebrates, model plants, bacteria) and under-represents protists, algae, and non-model invertebrates. Repeat-rich genomic regions (centromeres, telomeres, segmental duplications) remain challenging for all methods including FMGA. Functional GO annotation accuracy (72.4% top-1 correct) is limited by the training data bias toward well-characterised gene families.

## 6. Conclusion

This systematic benchmarking across 12 diverse genomes establishes FMGA -- a 1.2-billion-parameter foundation model pre-trained on 3,200 genomes with multi-task fine-tuning -- as the most accurate and comprehensive genome annotation system, achieving gene-level F1 0.924 overall (0.952 well-studied, 0.894 non-model), significantly outperforming MAKER2 (0.833;  $p < 0.001$ ), PGAP (0.856;  $p < 0.001$ ), and DNABERT-2 (0.890;  $p = 0.002$ ). The advantage is greatest for non-model organisms (+0.152 F1 over MAKER2), where evolutionary patterns learned from diverse pre-training data substitute for the species-specific evidence that traditional pipelines require. FMGA additionally provides regulatory element annotation (enhancer F1 0.724, promoter F1 0.842) -- a capability entirely absent from current annotation pipelines -- enabling comprehensive functional annotation from sequence alone. The multi-task architecture benefits both gene and regulatory prediction through shared feature learning. FMGA processes genomes at 42 s/Mb (approximately 35 hours for a mammalian genome on single GPU), making it practical for the thousands of newly sequenced genomes requiring annotation annually. FMGA model weights, fine-tuning scripts, and a web-based annotation server are available at <https://github.com/SIMI-Genomics/FMGA-2025> under Apache 2.0 licence. Future work should scale pre-training to 10,000+ genomes with emphasis on under-represented lineages, incorporate long-read transcriptome evidence for isoform-level annotation, develop active learning for efficient model updating as new genomes are annotated, and extend to pangenome annotation handling structural variation and gene presence/absence variation.

## References

- Avsec, Z., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A. and Taylor, K.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10), pp. 1196-1203.
- Bruna, T., Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 3(1), p. lqaa108.
- Campbell, M.S., Holt, C., Moore, B. and Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, 48(1), pp. 4.11.1-4.11.39.

- Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. and Ober, F. (2023) The Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *bioRxiv*, doi: 10.1101/2023.01.11.523679.
- Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12, p. 491.
- Ji, Y., Zhou, Z., Liu, H. and Davuluri, R.V. (2021) DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15), pp. 2112-2120.
- Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y. and Snoek, J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5), pp. 739-750.
- Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birber, C. and Ermon, S. (2023) HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36.
- Nguyen, E., Poli, M., Durrant, M.G., Thomas, A., Kang, B. and Sullivan, J. (2024) Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723), p. eado9336.
- Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2), pp. ii215-ii225.
- Tatusova, T., DiCuccio, M., Badretin, A., Chetvernin, V., Nawrocki, E.P. and Zaslavsky, L. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, 44(14), pp. 6614-6624.
- Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5), pp. 329-342.
- Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R.V. and Liu, H. (2024) DNABERT-2: efficient foundation model and benchmark for multi-species genome. In: *Proceedings of the International Conference on Learning Representations*.
- Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research*, 10(4), pp. 516-522.
- Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, 5, p. 59.
- Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20), pp. 6494-6506.
- Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, 20(16), pp. 2878-2879.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E. and Orvis, J. (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, 9(1), p. R7.
- Fiddes, I.T., Armstrong, J., Diekhans, M., Nachtweide, S., Kronenberg, Z.N. and Underwood, J.G. (2018) Comparative Annotation Toolkit (CAT) -- simultaneous clade and personal genome annotation. *Genome Research*, 28(7), pp. 1029-1038.
- Minkin, I. and Medvedev, P. (2020) Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nature Communications*, 11(1), p. 6327.
- Salzberg, S.L. (2019) Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, 20(1), p. 92.

## Declarations

## Funding

This study was supported by the Swiss National Science Foundation (SNSF) Eccellenza Professorial Fellowship PCEFP3\_202984 (I. Klein). Pre-training computational resources were provided by the Swiss National Supercomputing Centre (CSCS) under project s1142. The funding body had no role in study design, data collection, analysis, or publication decisions.

## Conflict of Interest

I. Klein is a co-inventor on a pending patent application (EP25196842.8) covering the FMGA multi-task genome annotation architecture. No other competing interests are declared.

## Data Availability Statement

FMGA model weights (1.2B parameters), fine-tuning scripts, pre-training configuration, and benchmark evaluation code are available at <https://github.com/SIMI-Genomics/FMGA-2025> under Apache 2.0 licence. A web-based annotation server is available at <https://fmga.simi-genomics.org>. All benchmark genomes and reference annotations are publicly available from NCBI/Ensembl.

## Ethical Approval

This computational study used publicly available genomic sequences and annotations. No human subjects, animal experiments, or personal data were involved. No ethics approval was required.

## Appendix A

### FMGA Architecture and Pre-Training Configuration

This appendix provides the complete FMGA architecture specification, pre-training corpus composition, multi-task fine-tuning configuration, and per-genome benchmark results.

#### A.1 Model Architecture

- 1a. Backbone: 24 transformer layers, 16 heads, d\_model 1,536, d\_ff 6,144; ALiBi position encoding; FlashAttention-2 for efficiency
- 1b. Tokenisation: BPE with vocabulary 4,096; context window 16,384 tokens (approximately 50 kb genomic sequence per window)
- 2a. Pre-training: MLM (15% masking) on 1.8T nucleotides from 3,200 genomes; 512 A100 GPUs, 14 days; AdamW lr 6e-4, cosine schedule
- 2b. Total parameters: 1.2B (backbone) + 48M (task heads) = 1.248B; pre-training compute: approximately  $2.4 \times 10^{22}$  FLOPs

#### A.2 Multi-Task Fine-Tuning

- 3a. Task 1 (coding): per-nucleotide binary classification; BCE loss; output head: linear(1536, 2)
- 3b. Task 2 (splice): donor/acceptor/none classification; CE loss; output head: linear(1536, 3); evaluated at GT/AG dinucleotides
- 4a. Task 3 (start/stop): ATG start and TAG/TAA/TGA stop classification; CE loss; output head: linear(1536, 3)
- 4b. Task 4 (regulatory): 5-class per-nucleotide (enhancer/promoter/silencer/insulator/none); CE loss; trained on ENCODE + Roadmap
- 5a. Task 5 (GO function): multi-label classification; BCE loss; 4,000 GO terms; output head: linear(1536, 4000) on gene-level pooled embeddings
- 5b. Fine-tuning: 8 A100, 3 days; multi-task loss = equal-weight sum of 5 task losses; AdamW lr 1e-5; LoRA rank 64 on attention layers