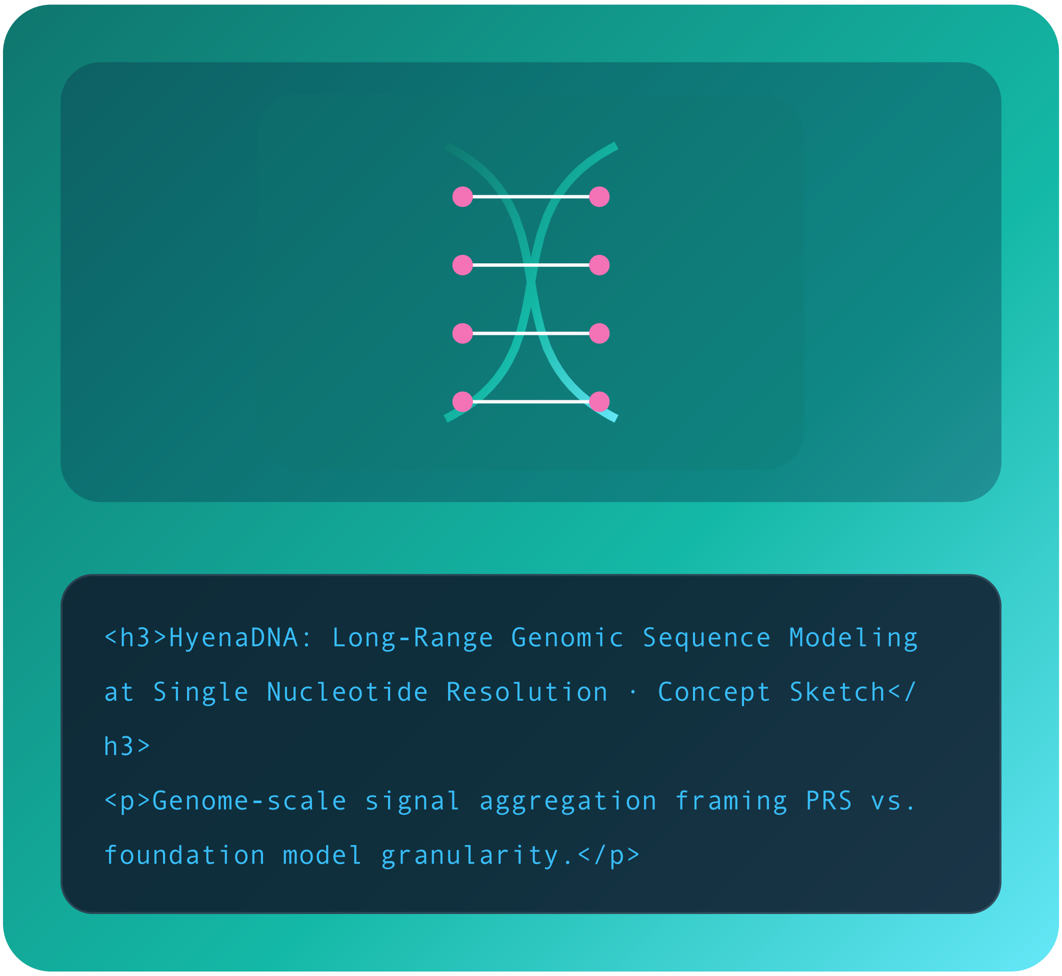

HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution

Authors: Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, Christopher Ré

Year: 2023

Venue: 37th Conference on Neural Information Processing Systems (NeurIPS)

1. Classification

- **Domain Category:**
  - **Genomics FM.** HyenaDNA is a genomic foundation model pretrained on the human reference genome, designed to capture long-range dependencies in DNA sequences at single nucleotide resolution.
- **FM Usage Type:**
  - **Core FM development.** The paper introduces a new family of DNA foundation models based on Hyena operators (implicit convolutions) that enable context lengths up to 1 million tokens, a 500× increase over previous dense attention-based models.
- **Key Modalities:**
  - Single-modality DNA sequence (human reference genome; single nucleotide-level tokens, no k-mer aggregation).

---

2. Executive Summary

This paper introduces HyenaDNA, a genomic foundation model that addresses two critical limitations of previous DNA language models: short context windows (512-4k tokens, <0.001% of human genome) and loss of single nucleotide resolution due to k-mer tokenization. Building on Hyena—a language model architecture using implicit convolutions that scales sub-quadratically—HyenaDNA achieves context lengths of up to 1 million tokens at single nucleotide resolution, enabling modeling of very long genomic regions (e.g., entire genes, regulatory domains) while preserving fine-grained resolution critical for detecting single nucleotide polymorphisms (SNPs). The model uses a decoder-only architecture with Hyena operators (long convolutions with data-controlled gating) pretrained on the human reference genome using next nucleotide prediction. Key innovations include a sequence length warm-up scheduler that gradually increases context during training (reducing training time by 40% and improving accuracy by 7.5 points at 450k length), and soft prompting techniques for downstream adaptation that leverage the extended context window. On downstream benchmarks, HyenaDNA achieves state-of-the-art performance on 12 of 18 Nucleotide Transformer tasks and 7 of 8 GenomicBenchmarks tasks, despite using 1500× fewer parameters (1.6M vs. 2.5B) and 3200× less pretraining data (1 human genome vs. 3202 genomes) compared to Nucleotide Transformer v2. The model also demonstrates novel capabilities enabled by long context, including in-context learning for species classification and effective handling of ultralong-range tasks. This work shows how architectural innovations (sub-quadratic operators) can unlock new capabilities (long-range modeling at fine resolution) that were previously impossible with attention-based models, and demonstrates the value of full-stack recipe development (architecture + training + adaptation) for foundation models.

---

3. Problem Setup and Motivation

- **Scientific / practical problem**
  - Genomic sequences are extremely long (human genome is 3.2B nucleotides) with long-range dependencies spanning 100k+ nucleotides (e.g., enhancer-promoter interactions, chromatin organization).
  - Many genomic tasks require both long-range context (to capture regulatory interactions) and single nucleotide resolution (to detect SNPs, mutations, fine-scale regulatory elements).
  - Previous DNA foundation models face a fundamental trade-off:
    - **Short context:** Attention-based models (DNABERT, Nucleotide Transformers) are limited to 512-4k tokens due to quadratic scaling, capturing <0.001% of the genome.

        - **Loss of resolution:** K-mer tokenization aggregates nucleotides into "words," losing single nucleotide resolution where subtle variations (SNPs) can have profound biological effects.
- **Why this is hard**
    - **Quadratic attention scaling:**
        - Transformers scale as $O(L^2)$ in sequence length, making 1M+ token contexts computationally infeasible.
        - Sparse attention and linear attention approximations trade expressivity for efficiency.
    - **Single nucleotide vs. long-range trade-off:**
        - Character-level modeling preserves resolution but produces very long sequences (1M+ tokens for 1M bp).
        - K-mer tokenization reduces sequence length but loses fine-grained information.
    - **Training stability at ultralong sequences:**
        - Directly training on 200k+ token sequences causes gradient variance issues and training instability.
    - **Downstream adaptation:**
        - Standard fine-tuning doesn't leverage the extended context window effectively.
        - Need new paradigms for adapting long-context models to downstream tasks.

---

4. Data and Modalities

- **Datasets used**
    - **Pretraining:**
        - Human reference genome (HG38/GRCh38), processed as contiguous sequences.
        - Total scale: sequences up to 1M nucleotides (tokens) in length.

- **Downstream evaluation:**
  - **GenomicBenchmarks:** 8 regulatory element classification tasks (enhancers, promoters, coding vs. intergenic), sequence lengths 200-4,776 bp.
  - **Nucleotide Transformer benchmark:** 18 tasks (enhancers, promoters, histone modifications, splice sites), lengths 200-600 bp.
  - **Species classification:** Novel long-range task requiring 1M+ context to distinguish species.
  - **Chromatin profile prediction:** 919-way multi-task predicting TF binding, DHS, histone marks.

- **Modalities**
  - Single modality: **DNA sequence** at single nucleotide resolution (A, C, G, T, N).
  - Outputs vary by task: binary/multi-class labels, continuous profiles, expression levels.
- **Preprocessing / representation**
  - **Character-level tokenization:**
    - Each nucleotide (A, C, G, T) is a token (vocabulary size: 4, plus special tokens for padding, separation, unknown).
    - No k-mer aggregation or BPE tokenization; preserves single nucleotide resolution.
  - **Sequence length warm-up:**
    - Training starts at L_1=64 tokens, doubles at each stage (64 → 128 → 256 → … → 1M).
    - Global batch size kept constant, so each stage processes more tokens per iteration.

5. Model / Foundation Model

- **Model Type**
  - **Decoder-only autoregressive model** based on Hyena operators (implicit convolutions with data-controlled gating).
  - Architecture: stack of Hyena blocks (Hyena operator + feed-forward network), similar to Transformer decoder but with attention replaced by Hyena operators.
- **Is it a new FM or an existing one?**
  - **New FM.** HyenaDNA is a ground-up redesign of genomic foundation models using Hyena operators instead of attention, enabling unprecedented context lengths at single nucleotide resolution.
- **Key components and innovations**
  - **Hyena operator:**
    - Replaces self-attention with long convolutions parameterized implicitly via neural networks.
    - Structure: $H(x_1, x_2)v = D_{x_2} T_h D_{x_1} v$, where:
      - T_h is a Toeplitz matrix from a learnable long convolution filter h (produced by neural network γ_θ).
      - D_{x_1}, D_{x_2} are element-wise gating matrices controlled by input projections.
    - Time complexity: $O(L \log^2 L)$ vs. $O(L^2)$ for attention.
  - **Sequence length warm-up scheduler:**
    - Gradually increases sequence length during training (64 → 128 → … → 1M).
    - Reduces training time by 40% and improves accuracy by 7.5 points at 450k length.
    - Acts as both stability mechanism and implicit batch size warm-up.
  - **Soft prompting for downstream adaptation:**
    - Injects learnable prompt tokens (up to 32k) directly into input sequence.

    - Only prompt parameters are optimized; pretrained model weights frozen.
    - Enables competitive performance without standard fine-tuning.
  - **Single nucleotide resolution:**
    - No k-mer tokenization; each nucleotide is a token.
    - Enables detection of SNPs and fine-scale regulatory elements.
- **Training setup**
  - **Pretraining objective:** Next nucleotide (token) prediction (autoregressive language modeling).
  - **Model sizes:** 2-8 layers, 128-256 hidden dimensions, context lengths 1024 to 1M tokens.
  - **Efficiency:** At 1M tokens, HyenaDNA is 160× faster than Transformer with Flash Attention.
  - **Gradient checkpointing:** Reduces memory footprint by 3× on sequences >160k.

---

6. Multimodal / Integration Aspects (If Applicable)

- **Not applicable.** HyenaDNA is a unimodal foundation model focused exclusively on DNA sequences. The long-context capabilities could potentially enable integration with other modalities (e.g., epigenomic tracks, expression data) in future work, but this is not explored in the paper.

---

7. Experiments and Results

Main findings

- **State-of-the-art performance with far fewer parameters:**
    - On Nucleotide Transformer benchmark: SotA on 12 of 18 tasks using 1.6M parameters vs. 2.5B for NT v2-2.5B (1500× fewer).
    - On GenomicBenchmarks: SotA on 7 of 8 tasks, with improvements up to +20 accuracy points on enhancer identification.
    - Pretraining data: 1 human genome vs. 3202 genomes for NT (3200× less data).
- **Long-range capabilities:**
    - **Species classification:** Effectively solves task by increasing context to 1M tokens (no downsampling needed).
    - **Chromatin profiles:** Competitively performs 919-way multi-task prediction against larger sparse-attention BigBird Transformer.
- **Single nucleotide resolution benefits:**
    - Outperforms k-mer-based models (DNABERT) on tasks requiring fine-scale resolution.
    - Enables detection of SNPs and single-nucleotide regulatory elements.
- **Training efficiency:**
    - Sequence length warm-up reduces training time by 40% at 450k length.
    - 160× faster than Transformer at 1M tokens (forward + backward pass).
- **In-context learning:**
    - Soft prompting enables adaptation to new tasks without fine-tuning.
    - Performance improves with more prompt tokens (up to 32k tested), approaching fine-tuning performance.

Ablation studies

- **Sequence length warm-up:**
  - Without warm-up: training is slower and less stable at long sequences.
  - With warm-up: 40% faster training, 7.5 point accuracy improvement at 450k.
- **Context length vs. perplexity:**
  - Longer context improves pretraining perplexity (better next-token prediction).
  - However, for models too shallow, perplexity can degrade at very long sequences (inflection points).
- **Soft prompting:**
  - More prompt tokens (2 → 32k) improve performance, saturating near fine-tuning baseline.
  - K-shot demonstrations (few-shot learning) less effective than soft prompting for this model.

Key insights

- **Long context enables new capabilities:**
  - Species classification requires 1M+ context to distinguish sequences; HyenaDNA is the first model to handle this without downsampling.
- **Single nucleotide resolution matters:**
  - Preserving fine-grained resolution is critical for tasks like enhancer identification and variant effect prediction.
- **Efficiency unlocks scale:**
  - Sub-quadratic scaling ($O(L \log^2 L)$) makes 1M token contexts feasible, enabling new applications.

8. Strengths and Limitations

Strengths

- **Unprecedented context length:**
  - 1M tokens at single nucleotide resolution is a 500× increase over previous dense attention models.
- **Preserves fine-grained resolution:**
  - Character-level tokenization enables detection of SNPs and single-nucleotide regulatory elements.
- **Computational efficiency:**
  - 160× faster than Transformers at 1M tokens, enabling practical training and inference.
- **Strong performance with minimal resources:**
  - Achieves SotA with 1500× fewer parameters and 3200× less pretraining data than Nucleotide Transformer v2.
- **Full-stack recipe:**
  - Provides architecture, training strategies (warm-up), and adaptation methods (soft prompting) as a complete package.
- **Novel capabilities:**
  - Enables in-context learning and ultralong-range tasks previously impossible.

Limitations

- **Still smaller than some baselines:**
  - 1.6M parameters is very small; larger HyenaDNA models might achieve even better performance.
- **Limited pretraining data:**
  - Only trained on human genome; multi-species pretraining (like NT) might improve generalization.

- **No RC-equivariance:**
  - Doesn't explicitly encode reverse-complement symmetry (unlike Caduceus); relies on data augmentation if needed.
- **In-context learning is limited:**
  - DNA vocabulary is small (4 nucleotides), making pure in-context learning challenging; requires soft prompting or instruction tuning.
- **Training stability:**
  - Even with warm-up, very long sequences (1M+) can be challenging to train; warm-up schedule needs careful tuning.

Open questions / future directions

- **Scaling laws:**
  - How does performance scale with model size, pretraining data, and context length?
- **Multi-species pretraining:**
  - Would training on multiple species (like NT) improve performance?
- **RC-equivariance integration:**
  - Can HyenaDNA be combined with RC-equivariant architectures (like Caduceus) for even better performance?
- **Generative capabilities:**
  - Can HyenaDNA generate long genomic sequences? How does it compare to Evo 2 or GENERator?
- **Interpretability:**
  - What long-range patterns does HyenaDNA learn? Can we interpret the learned representations?

9. Context and Broader Impact

Relation to other work

- **Compared to Nucleotide Transformer (Dalla-Torre et al., 2023):**
  - NT uses attention with 6-mer tokenization, limited to ~1k tokens.
  - HyenaDNA uses Hyena operators with character-level tokens, enabling 1M tokens.
  - HyenaDNA achieves similar or better performance with 1500× fewer parameters.
- **Compared to DNABERT-2 (Zhou et al., 2024):**
  - DNABERT-2 uses BPE tokenization and attention, limited context.
  - HyenaDNA preserves single nucleotide resolution and enables much longer contexts.
- **Compared to Caduceus (Schiff et al., 2024):**
  - Caduceus uses Mamba SSMs with RC-equivariance for long-range modeling.
  - HyenaDNA uses Hyena operators without explicit RC-equivariance but achieves longer contexts (1M vs. 100k+).
- **Compared to Evo 2 (Brixi et al., 2025):**
  - Evo 2 uses StripedHyena 2 (multi-hybrid architecture) for generative DNA modeling at 1M context.
  - HyenaDNA is an earlier, simpler architecture that demonstrates long-context capabilities for discriminative tasks.
- **Connection to Hyena (Poli et al., 2023):**
  - HyenaDNA adapts the Hyena architecture (designed for language) to genomics, showing the generality of sub-quadratic operators.

Broader scientific and practical impact

- **Enables new genomic applications:**
  - Long-context modeling opens possibilities for whole-gene, whole-chromosome, or even whole-genome analysis.

    - Single nucleotide resolution enables fine-scale variant effect prediction and regulatory element identification.
- **Demonstrates value of architectural innovation:**
    - Shows how sub-quadratic operators (Hyena) can unlock capabilities impossible with attention-based models.
- **Provides practical recipe:**
    - Full-stack approach (architecture + training + adaptation) makes it easier for others to build long-context genomic models.
- **Influences future work:**
    - Evo 2 and other recent models build on similar principles (long-context, sub-quadratic operators).

Open questions for future research

- **How to scale further?**
    - Can we model entire genomes (3.2B nucleotides) with current architectures?
- **Multi-species pretraining:**
    - Would training on diverse species improve generalization?
- **RC-equivariance:**
    - How to combine long-context capabilities with architectural RC-equivariance?
- **Generative modeling:**
    - Can HyenaDNA-style architectures generate long, biologically valid sequences?
- **Interpretability:**
    - What long-range patterns do these models learn? Can we extract biological insights?

10. Key Takeaways

1. **Architectural innovation unlocks new capabilities:**
   Sub-quadratic operators (Hyena) enable context lengths (1M tokens) that are computationally infeasible with attention ($O(L^2)$ scaling).

2. **Resolution vs. context trade-off is real:**
   K-mer tokenization reduces sequence length but loses single nucleotide resolution; character-level modeling preserves resolution but requires efficient architectures for long sequences.

3. **Training strategies matter:**
   Sequence length warm-up is crucial for stable training at ultralong sequences, reducing training time and improving accuracy.

4. **Efficiency enables scale:**
   Being 160× faster than Transformers at 1M tokens makes previously impossible applications feasible.

5. **Fewer parameters can be enough:**
   HyenaDNA achieves SotA with 1500× fewer parameters than Nucleotide Transformer, showing that architecture and training matter more than raw parameter count.

6. **Full-stack recipe development:**
   Don't just propose an architecture; provide training strategies, adaptation methods, and evaluation protocols as a complete package.

7. **Long context enables new tasks:**
   Extended context windows unlock capabilities like species classification and ultralong-range regulatory element prediction that weren't possible before.

8. **In-context learning is possible in genomics:**
   Soft prompting enables adaptation to new tasks without fine-tuning, though it requires careful design due to small vocabulary.

9. **Single nucleotide resolution matters:**
   Preserving fine-grained resolution is critical for detecting SNPs and fine-scale regulatory elements that k-mer models miss.

10. **This is foundational work:**
    HyenaDNA demonstrates that long-context, fine-resolution genomic modeling is possible, influencing subsequent work like Evo 2 and showing the path forward for genomic foundation models.