

Abstract

Recent genomic foundation models (GFMs) based on large language model architectures enable contextual understanding of DNA sequences but remain limited in patient-level cancer analysis. We developed **DNACHUNKER**, a lightweight GFM using an H-net-based learnable tokenization method optimized for long genomic sequences. Embeddings from DNACHUNKER are aggregated via a transformer-based model and combined with copy-number variation (CNV) features to generate patient-level representations. Applied to ~3.1k cancer whole-genome sequences, this approach achieves accurate stratification by **cancer type**, homologous recombination deficiency (**HRD**), and prediction analysis of microarray 50 (**PAM50**) subtypes, demonstrating the potential of **foundation-model-driven genomics** for precision oncology.

Main Designs

1. GFM Training: How can we train GFM to perform better at mutation sequence embedding
2. Embedding Aggregation: How to aggregate embedding to bridge between GFM and clinical applications

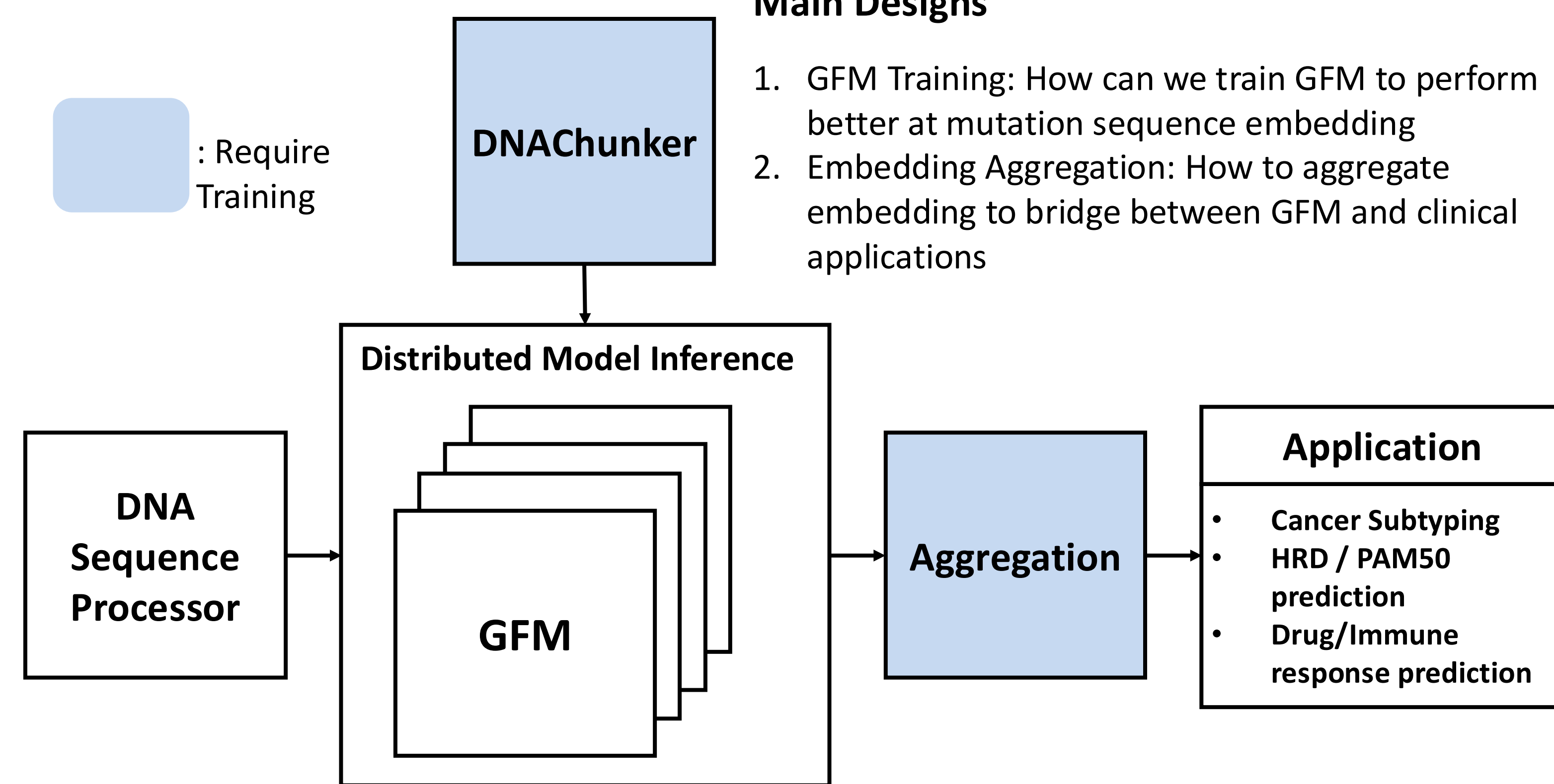


Figure 1. Overview of developing Cancer-GFM. Embedding vectors generated from each mutation by DNACHUNKER are integrated through a transformer-based aggregation model to produce a unified, sample-level representation.

Methods

Chunking with dynamic boundaries

DNACHUNKER adaptively segments DNA sequences into variable-length chunks using boundary probabilities p_t :

$$p_t = \left(1 - \frac{(q_t)^T k_{t-1}}{\|q_t\| \cdot \|k_{t-1}\|}\right), \quad q_t = W_{enc,q} \hat{x}_t, \quad k_t = W_{enc,k} \hat{x}_t$$

Positions with $p_t > 0.5$ form a chunk; $p_t < 0.5$ marks a boundary. Low-information regions are compressed, while functional regions are preserved at high resolution via a **two-stage Caduceus encoder** capturing bidirectional genomic context.

Datasets

We trained the **Cancer Aggregation Model** using whole-genome sequencing (WGS) datasets:

- **PCAWG**: 2,050 samples across 21 cancer types.
- **CUBRICS**: 1,053 breast cancer **WGS** and **whole transcriptome** samples generated by **INOCRAS**, annotated with detailed clinical outcomes and molecular subtype labels (e.g., PAM50).
- **TCGA-BRCA**: Whole-genome sequencing of TCGA-BRCA samples, used for external validation.

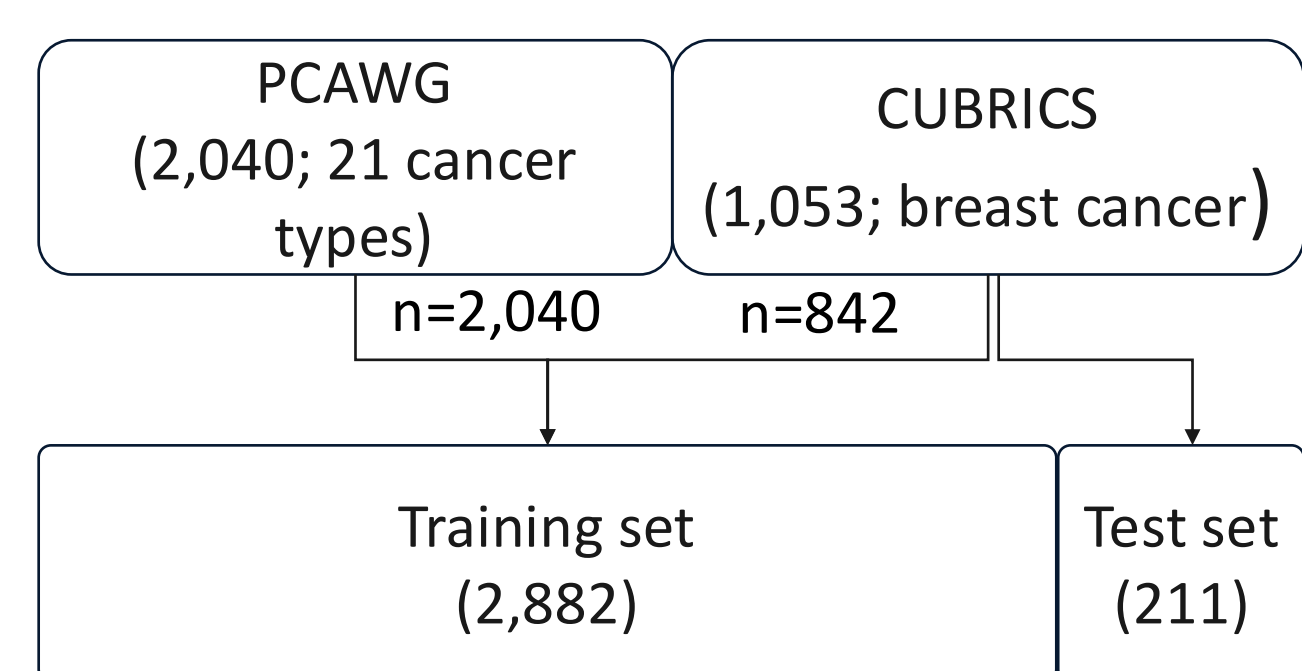


Figure 4. Data structure of training data.

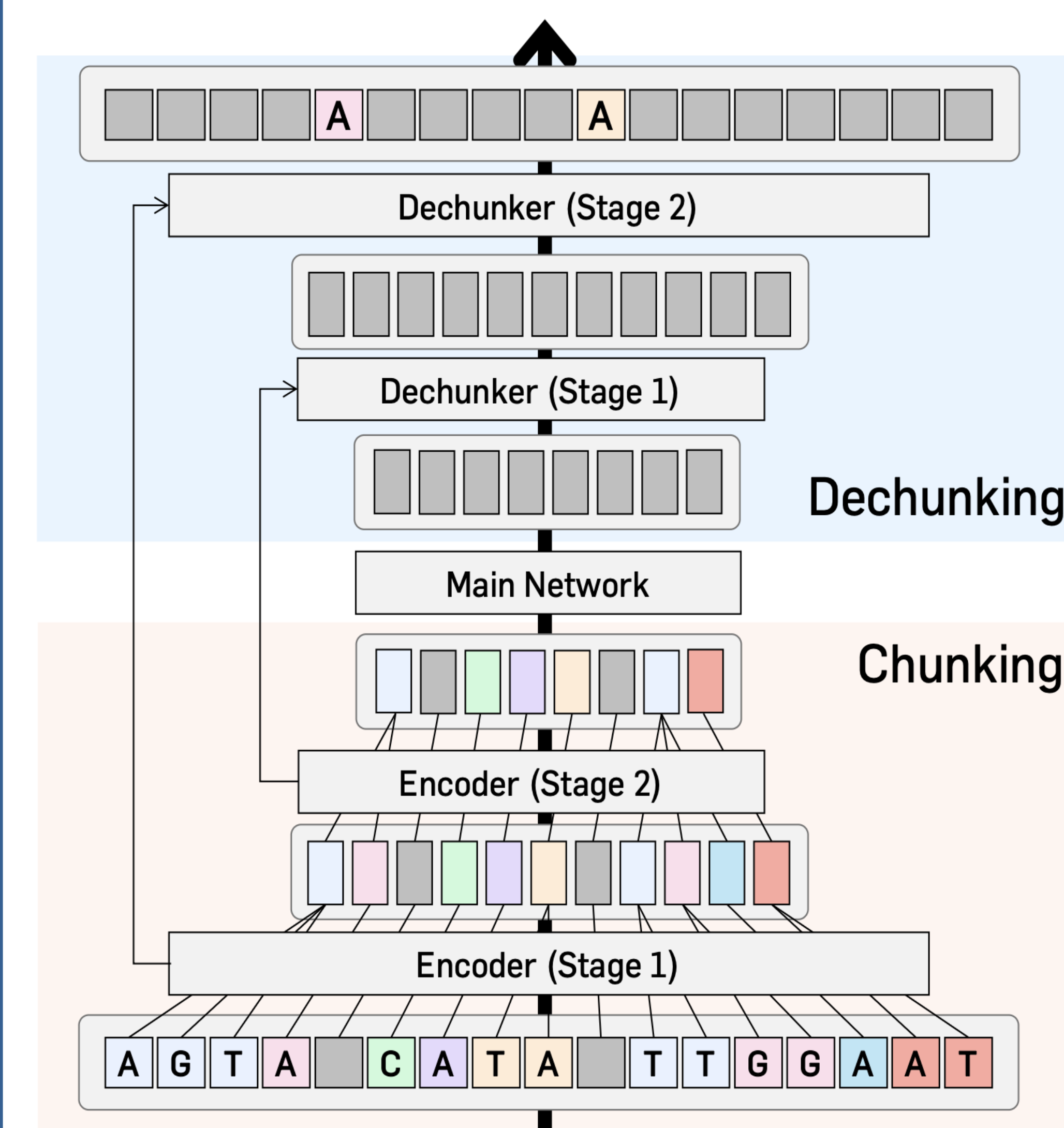
Whole genome sequencing

We performed tumor-normal WGS pair using CancerVisionTM¹. Tumors were sequenced at ~40x coverage, while normal were ~20x. To assess HRD, we used our proprietary algorithm by combining HRD-associated features, such as mutational signatures of point mutations, and copy number changes.

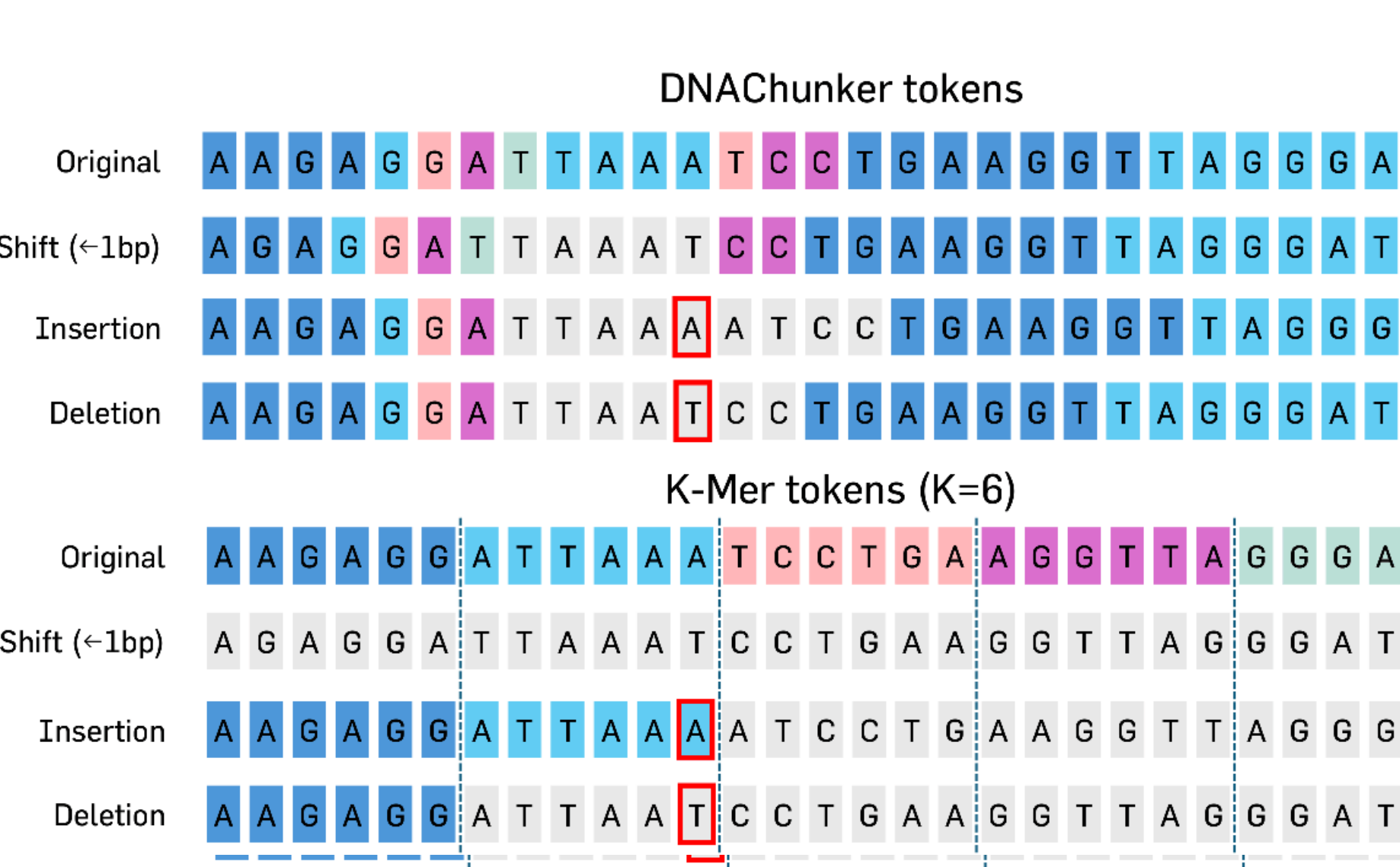
DNACHUNKER²: Learnable Tokenization For DNA Language Models

DNACHUNKER adopts the **dynamic chunking mechanism** of H-net³ (1), allowing it to segment DNA sequences into variable-length chunks. This adaptive tokenization provides two key advantages: (1) robustness to nucleotide-level shifts and mutations, and (2) finer representation of functionally important regions. Trained on the human reference genome (HG38) and evaluated on the Nucleotide Transformer and Genomic Benchmarks datasets, **DNACHUNKER achieves performance comparable to the state-of-the-art GENERator⁴** (2) with 1.2 billion parameters—while using only 156 million parameters.

(a) Model architecture of DNACHUNKER



(b) Robustness of shift or mutation (H313 gene)



(c) Distribution of chunk size of chromosome 21 and 22

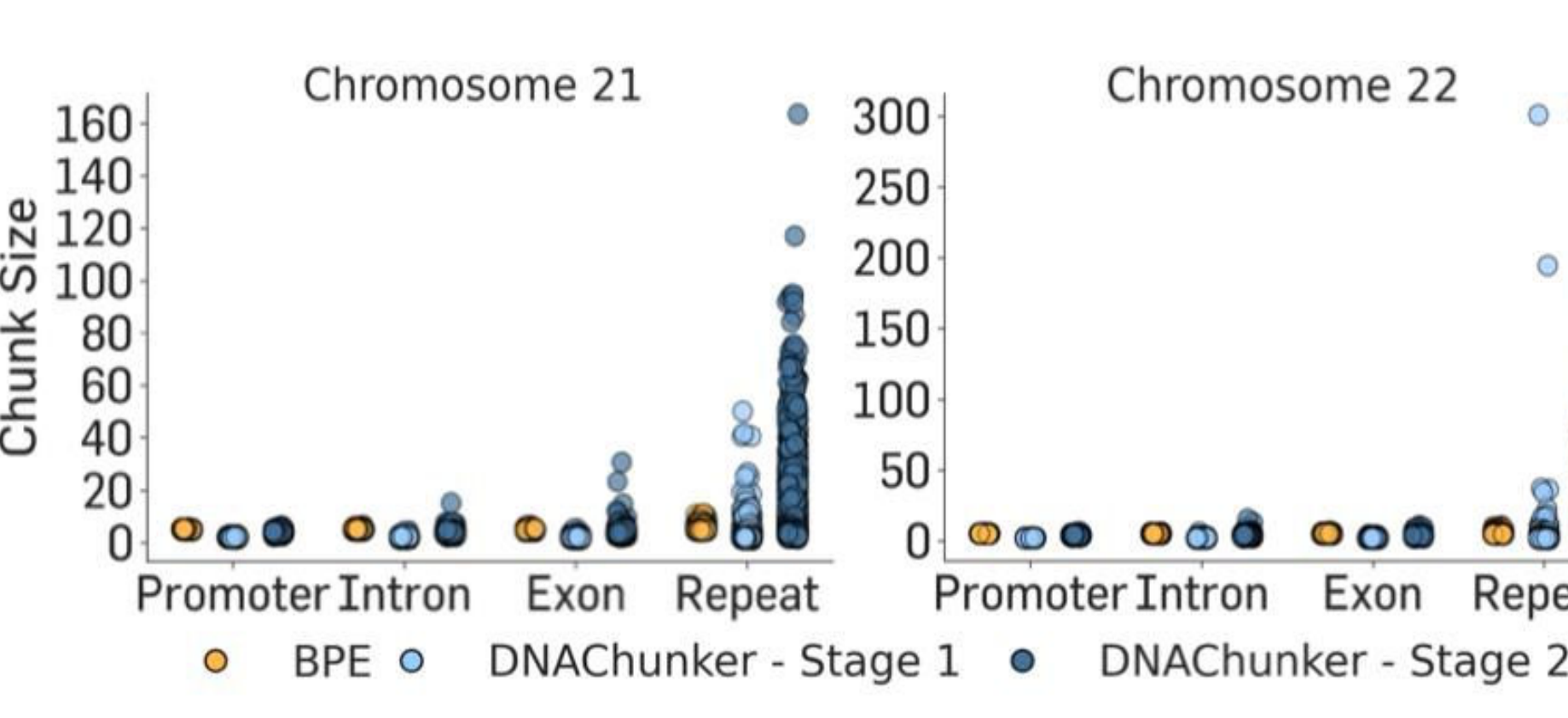


Figure 2. Architecture, tokenizer robustness, and distribution of chunk size. (a) Overview of the DNACHUNKER architecture. (b) The tokenizer remains consistent under nucleotide shifts or mutations, preserving token identities (colors). (c) DNACHUNKER adaptively encodes functional regions (promoter, exon, intron) with fine resolution and compresses repetitive, non-functional regions with larger chunks.

Table 1: Nucleotide Transformer Benchmark. The reported values represent the Matthews Correlation Coefficient (MCC; mean ± standard error) averaged over 10-fold cross-validation.

	Enformer (252M)	DNABERT-2 (117M)	HyenaDNA (55M)	NT-multi (2.5B)	NT-v2 (500M)	Caduceus-Ph (8M)	Caduceus-PS (8M)	GROVER (87M)	GENERator (1.2B)	DNACHUNKER (172M)
Histone Markers										
H3	0.724 ± 0.018	0.785 ± 0.012	0.781 ± 0.015	0.793 ± 0.013	0.788 ± 0.010	0.794 ± 0.012	0.772 ± 0.022	0.768 ± 0.008	0.806 ± 0.005	0.817 ± 0.011
H3K14ac	0.284 ± 0.024	0.515 ± 0.009	0.608 ± 0.020	0.538 ± 0.009	0.538 ± 0.015	0.564 ± 0.013	0.596 ± 0.018	0.548 ± 0.020	0.605 ± 0.008	0.711 ± 0.021
H3K36me3	0.345 ± 0.019	0.591 ± 0.005	0.614 ± 0.014	0.618 ± 0.011	0.618 ± 0.015	0.590 ± 0.018	0.611 ± 0.018	0.563 ± 0.017	0.657 ± 0.007	0.677 ± 0.003
H3K4me1	0.291 ± 0.016	0.512 ± 0.008	0.512 ± 0.008	0.541 ± 0.005	0.544 ± 0.009	0.468 ± 0.015	0.487 ± 0.029	0.461 ± 0.018	0.553 ± 0.009	0.631 ± 0.009
H3K4me2	0.207 ± 0.021	0.333 ± 0.013	0.455 ± 0.028	0.324 ± 0.014	0.302 ± 0.020	0.332 ± 0.034	0.341 ± 0.016	0.403 ± 0.042	0.424 ± 0.013	0.599 ± 0.011
H3K4me3	0.156 ± 0.022	0.353 ± 0.021	0.550 ± 0.015	0.408 ± 0.011	0.437 ± 0.028	0.490 ± 0.042	0.528 ± 0.033	0.458 ± 0.022	0.512 ± 0.009	0.660 ± 0.045
H3K9me3	0.498 ± 0.013	0.615 ± 0.010	0.669 ± 0.014	0.623 ± 0.010	0.623 ± 0.010	0.641 ± 0.028	0.682 ± 0.018	0.626 ± 0.026	0.670 ± 0.011	0.731 ± 0.012
H3K9ac	0.415 ± 0.020	0.545 ± 0.009	0.586 ± 0.021	0.547 ± 0.011	0.567 ± 0.020	0.575 ± 0.024	0.564 ± 0.018	0.581 ± 0.015	0.612 ± 0.006	0.678 ± 0.007
H4	0.735 ± 0.023	0.797 ± 0.008	0.763 ± 0.012	0.808 ± 0.007	0.795 ± 0.008	0.788 ± 0.010	0.799 ± 0.010	0.769 ± 0.017	0.815 ± 0.008	0.813 ± 0.012
H4ac	0.275 ± 0.022	0.465 ± 0.013	0.564 ± 0.011	0.492 ± 0.014	0.502 ± 0.025	0.548 ± 0.027	0.585 ± 0.018	0.530 ± 0.017	0.592 ± 0.015	0.687 ± 0.027
Average MCC (†)	0.393	0.551	0.610	0.569	0.571	0.579	0.606	0.571	0.625	0.701
Regulatory Annotation										
Enhancer	0.454 ± 0.029	0.525 ± 0.026	0.520 ± 0.031	0.545 ± 0.028	0.561 ± 0.029	0.522 ± 0.024	0.511 ± 0.026	0.516 ± 0.018	0.580 ± 0.015	0.558 ± 0.011
Enhancer Type	0.312 ± 0.043	0.423 ± 0.018	0.403 ± 0.056	0.444 ± 0.022	0.444 ± 0.036	0.403 ± 0.028	0.410 ± 0.026	0.433 ± 0.029	0.477 ± 0.017	0.519 ± 0.005
Promoter All	0.910 ± 0.004	0.945 ± 0.003	0.919 ± 0.003	0.951 ± 0.004	0.952 ± 0.002	0.937 ± 0.002	0.941 ± 0.003	0.926 ± 0.004	0.962 ± 0.002	0.967 ± 0.003
Promoter NonTATA	0.910 ± 0.006	0.944 ± 0.003	0.919 ± 0.004	0.969 ± 0.003	0.952 ± 0.003	0.935 ± 0.007	0.940 ± 0.002	0.925 ± 0.006	0.962 ± 0.001	0.971 ± 0.007
Promoter TATA	0.920 ± 0.012	0.911 ± 0.011	0.881 ± 0.020	0.919 ± 0.008	0.933 ± 0.009	0.895 ± 0.010	0.903 ± 0.010	0.891 ± 0.009	0.948 ± 0.008	0.961 ± 0.015
Average MCC (†)	0.701	0.750	0.728	0.766	0.768	0.738	0.741	0.738	0.786	0.796
Splice Site Annotation										
Splice Acceptor	0.772 ± 0.007	0.909 ± 0.004	0.935 ± 0.005	0.973 ± 0.002	0.973 ± 0.004	0.918 ± 0.017	0.907 ± 0.015	0.912 ± 0.010	0.981 ± 0.002	0.969 ± 0.013
Splice Site All	0.831 ± 0.012	0.950 ± 0.003	0.917 ± 0.006	0.974 ± 0.005	0.974 ± 0.005	0.935 ± 0.011	0.953 ± 0.005	0.919 ± 0.005	0.976 ± 0.001	0.968 ± 0.030
Splice Donor	0.813 ± 0.015	0.927 ± 0.003	0.894 ± 0.013	0.974 ± 0.002	0.977 ± 0.007	0.912 ± 0.009	0.930 ± 0.010	0.888 ± 0.012	0.979 ± 0.001	0.960 ± 0.007
Average MCC (†)	0.805	0.929	0.915	0.974	0.975	0.922	0.930	0.906	0.978	0.965
Total Average MCC (†)	0.547	0.669	0.694	0.690	0.693	0.680	0.697	0.673	0.728	0.772
Total Average Rank (‡)	9.67	6.72	6.00	4.83	4.56	6.33	5.61	7.22	2.06	1.67

Table 2: Genomic Benchmark. The reported values represent accuracy (mean ± standard error) averaged over 5-fold cross-validation.

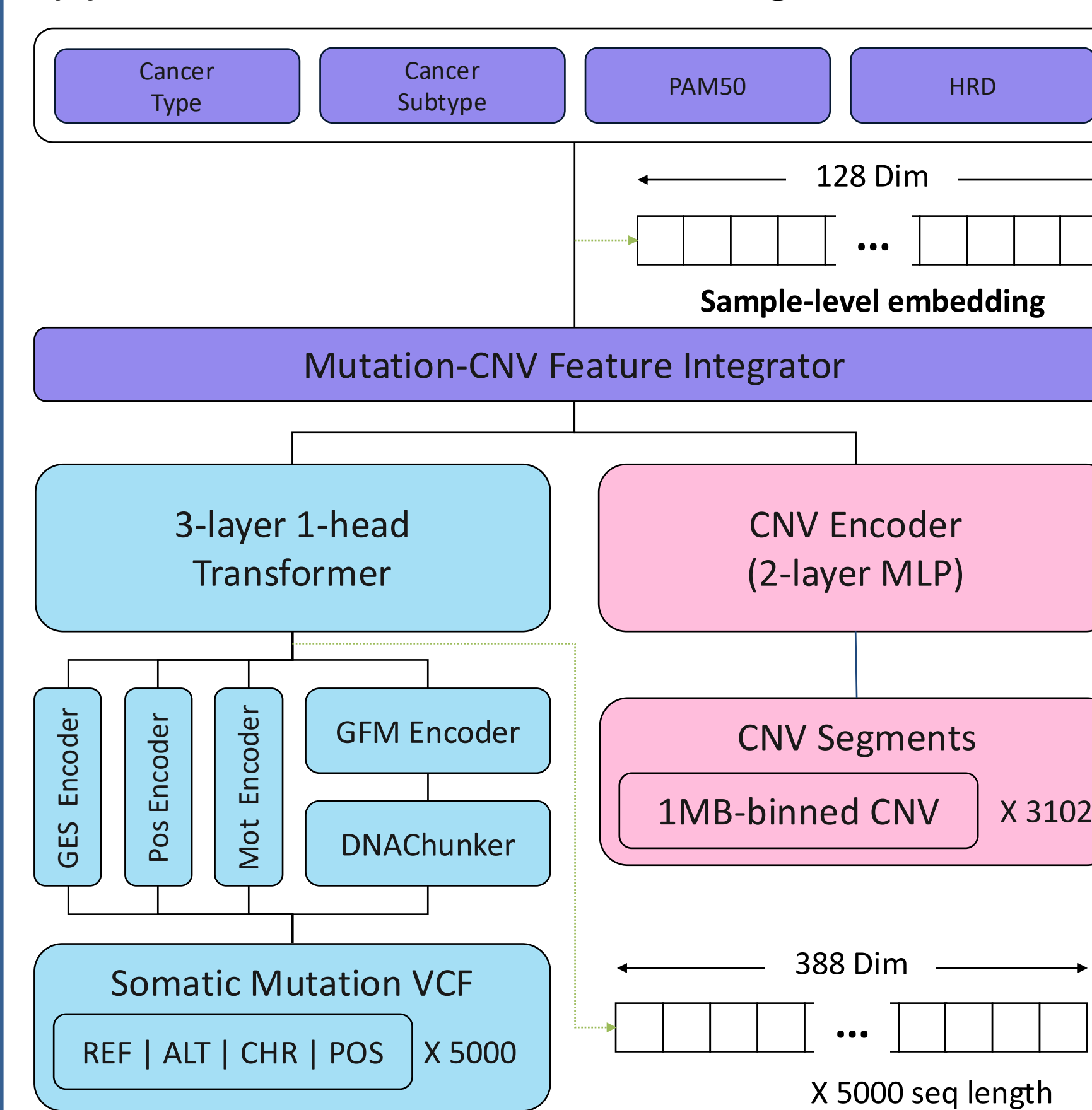
	DNABERT-2 (117M)	HyenaDNA (55M)	NT-v2 (500M)	Caduceus-Ph (8M)	Caduceus-PS (8M)	GROVER (87M)	GENERator (1.2B)	GENERator-All (1.2B)	DNACHUNKER (172M)
Coding vs. Intergenic	0.951 ± 0.002	0.902 ± 0.004	0.955 ± 0.001	0.933 ± 0.001	0.944 ± 0.002	0.919 ± 0.002	0.963 ± 0.000	0.959 ± 0.001	0.955 ± 0.012
Drosophila Enhancers Stark	0.774 ± 0.011	0.770 ± 0.016	0.797 ± 0.009	0.827 ± 0.010	0.816 ± 0.015	0.761 ± 0.011	0.821 ± 0.005	0.768 ± 0.015	0.779 ± 0.021
Human Enhancers Cohn	0.758 ± 0.005	0.725 ± 0.009	0.756 ± 0.006	0.747 ± 0.003	0.749 ± 0.003	0.738 ± 0.003	0.763 ± 0.002	0.754 ± 0.006	0.761 ± 0.011
Human Enhancers Ensembl	0.918 ± 0.003	0.901 ± 0.003	0.921 ± 0.004	0.924 ± 0.002	0.923 ± 0.002	0.911 ± 0.004	0.917 ± 0.002	0.912 ± 0.002	0.922 ± 0.007
Human Ensembl Regulatory	0.874 ± 0.007	0.932 ± 0.001	0.941 ± 0.001	0.938 ± 0.004	0.941 ± 0.001	0.897 ± 0.001	0.928 ± 0.001	0.926 ± 0.001	0.935 ± 0.005
Human non-TATA Promoters	0.957 ± 0.008	0.894 ± 0.023	0.932 ± 0.006	0.961 ± 0.003	0.961 ± 0.002	0.950 ± 0.005	0.958 ± 0.001	0.955 ± 0.005	0.962 ± 0.001
Human OCR Ensembl	0.806 ± 0.003	0.774 ± 0.004	0.813 ± 0.001	0.825 ± 0.004	0.826 ± 0.003	0.789 ± 0.002	0.823 ± 0.002	0.812 ± 0.003	0.810 ± 0.007
Human vs. Worm	0.977 ± 0.001	0.958 ± 0.004	0.976 ± 0.001	0.975 ± 0.001	0.975 ± 0.001	0.966 ± 0.001	0.980 ± 0.000	0.978 ± 0.001	0.969 ± 0.001
Mouse Enhancers Ensembl	0.865 ± 0.014	0.756 ± 0.010	0.855 ± 0.018	0.788 ± 0.028	0.826 ± 0.021	0.742 ± 0.025	0.871 ± 0.015	0.784 ± 0.027	0.874 ± 0.020
Average Acc (†)	0.876	0.846	0.883	0.880	0.885	0.853	0.892	0.872	0.885
Average Rank (‡)	5.11	8.22	4.17	3.89	3.33	8.11	2.89	5.44	3.29

*Best result are bold; second best are underlined.

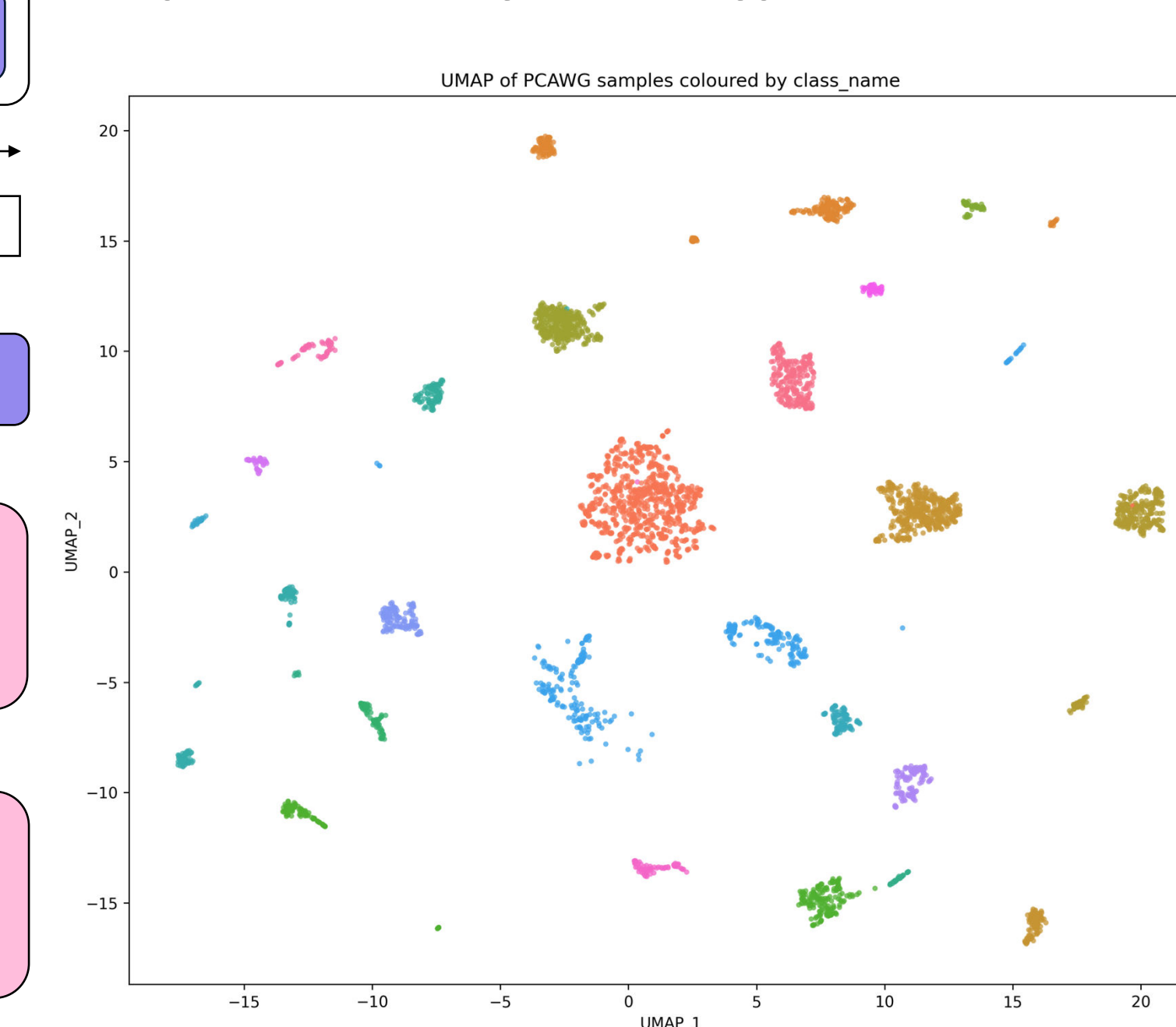
Cancer Aggregation Model: Multimodal Sample-Level Representation for Clinical Applications

The **Cancer Aggregation Model** integrates GFM-derived mutation embeddings with 1 Mb-binned CNV features to generate a unified, patient-level representation. Unsupervised clustering reveals clear separation by **HRD** and **PAM50** subtypes, while multi-task learning achieves high accuracy across cancer type, HRD, and PAM50 classification.

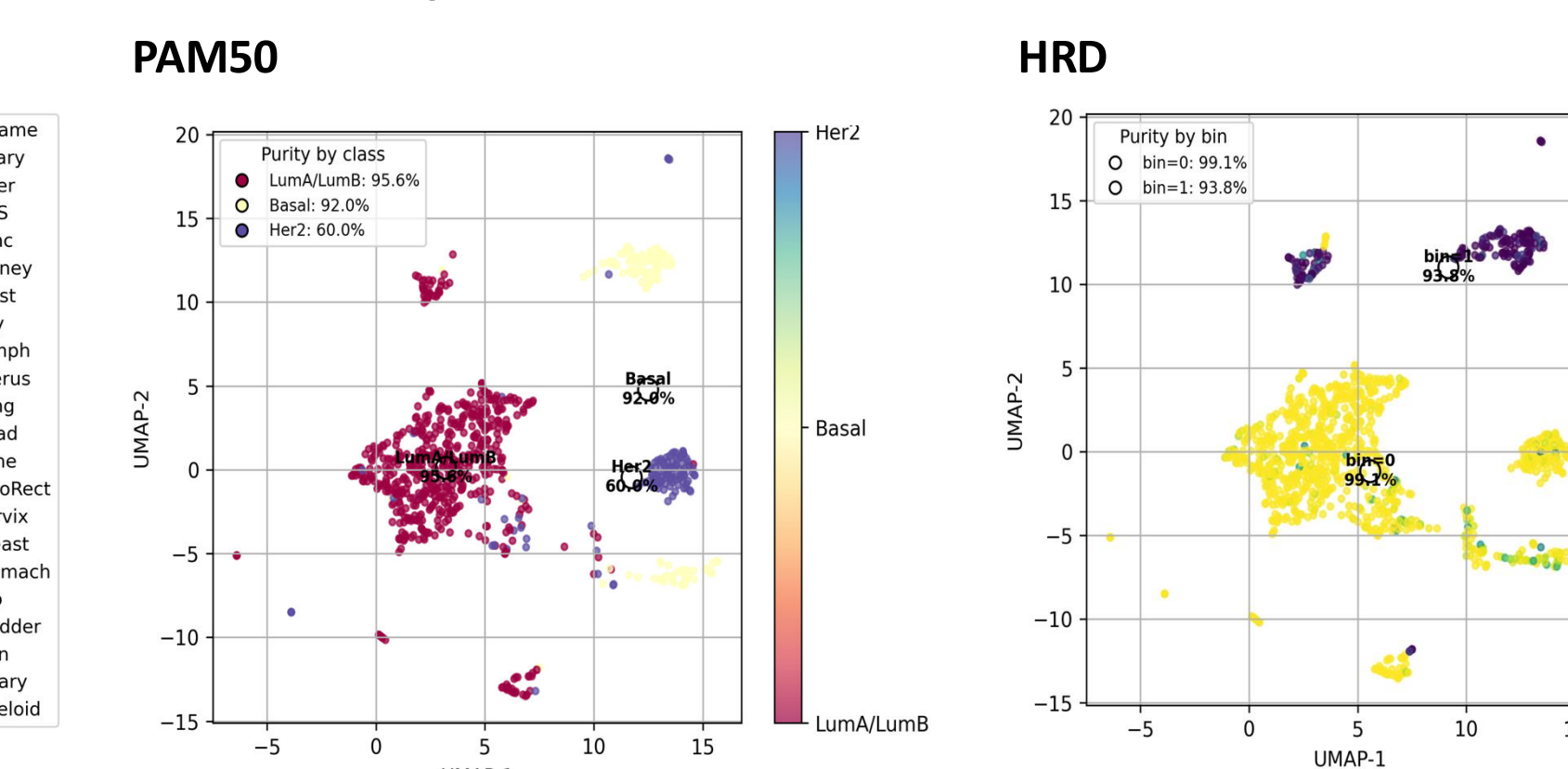
(a) Model architecture and Data Design



(b) Sample-level embedding visualization of PCAWG samples colored by cancer type.



(c) Sample-level embedding visualization of breast cancer samples



(d) Performance (Accuracy) and external validation

	Cancer type prediction	HRD	PAM50	n
CUBRICS	99.05%	98.10%	84.05%	211
TCGA-BRCA	96.89%	92.83%	N/A	837

Figure 3. Overview and Evaluation of the Cancer Aggregation Model. (a) Model architecture. (b, c) UMAPs of PCAWG and breast cancer datasets showing clear clustering according to each feature (cancer type, PAM50, and HRD). (d) Classification performance on breast cancer and TCGA-BRCA cohorts.

Discussion

- Built a **SOTA Genomic Foundation Model (GFM)** with learnable tokenization.
- **Cancer Aggregation Model** enables **DNA-based PAM50** subtyping, replacing RNA dependency.
- Demonstrates the potential of **AI-driven precision oncology**.
- Applicable to **primary origin of malignancy of unknown origin, survival, and treatment response prediction**.

References

1. Lee, S. et al. Target-enhanced whole-genome sequencing (TE-WGS) shows clinical validity equivalent to commercially available targeted oncology panel. *medRxiv* (2023). doi:10.1101/2023.12.20.23300156
2. Kim, T. et al. DNACHUNKER: Learnable tokenization for DNA language models. *arXiv* (2024). doi:10.48550/arXiv.2401.03019
3. Hwang, S., Wang, B. & Gu, A. Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv* (2025). doi:10.48550/arXiv.2507.07955
4. Wu, W. et al. GENERator: A long-context generative genomic foundation model. *arXiv* (2025). doi:10.48550/arXiv.2502.07272

Contact

Jonghoon Lee
Inocras Inc.
Email:
jonghoonlee@inocras.com
Website:
<https://inocras.com/>

