ADVERSARIAL GENOMIC SEQUENCES COULD EVADE BIOSECURITY SCREENING

Jeyashree Krishnan*
Apart Research
Institute for Computational Biomedicine
Uniklink RWTH Aachen

Ajay Mandyam Rangarajan
Apart Research

Andrea Loehr
Apart Research

Jason Hoelscher-Obermaier
Apart Research

ABSTRACT

Recent biosecurity risk assessments identify biological foundation models as having high misuse potential over the coming decade that lower barriers to develop harmful biological agents, with many tools open source and readily accessible to potential threat actors (Nelson & Rose, 2023; Webster et al., 2025). At the same time, Genomic Foundation Models (GFMs) such as DNABERT-2 (Zhou et al., 2024) and Nucleotide Transformer v2 (Dalla-Torre et al., 2024) could help improve biosecurity defenses; yet their robustness under adversarial attacks remains poorly understood.

To address this gap, we evaluate the susceptibility of GFMs to adversarial mutations. Since genomic inputs are discrete and subject to biological constraints, standard adversarial perturbation models for continuous domains do not directly apply (Kuleshov et al., 2021). We study promoter classification as a representative task and analyze GFM robustness under black-box threat models with nucleotide-level perturbations. Adversarial sequences are generated via genetic algorithms subject to explicit biological constraints, including GC-content bounds, mutation budget limits, and transitions (Ti)/transversions (Tv) bias.

Our analysis reveals that GFMs are vulnerable to perturbation-efficient adversarial attacks requiring only a small number of nucleotide edits. These vulnerabilities are not random: effective adversarial substitutions favor transversions over transitions and concentrate in the 5' regulatory region of promoter sequences. Iterative adversarial training reduces attack success rates from initial values of approximately 40–45%, but substantial attack success of about 20–30% persists after ten rounds of hardening. Our results establish adversarial evaluation of GFMs as a joint biological and machine learning challenge requiring domain-specific threat models and defenses.

1 INTRODUCTION

Artificial intelligence is transforming the biosciences, but this progress brings new biosecurity challenges. As AI-powered tools become more accessible, the gains from new capabilities are accompanied by emerging vulnerabilities, increasing the potential for misuse (Zhang et al., 2025a). Genomic Foundation Models (GFMs) such as DNABERT-2 (Zhou et al., 2024) and Nucleotide Transformer v2 (Dalla-Torre et al., 2024) exemplify this dual-use concern: they achieve strong performance across genomic prediction tasks including promoter and enhancer classification, splice site prediction, and regulatory element identification, yet their robustness against adversarial manipulation remains largely underexplored. These models learn dense representations from large-scale DNA sequence data and are increasingly adopted as backbones for downstream discriminative and generative applications. As GFMs move closer to deployment in high-stakes biological and biosecurity-relevant pipelines, such as sequence screening tools, DNA sequence design assistants, or diagnostic

*Corresponding author: jekrishnan@ukaachen.de

support applications, understanding their robustness under realistic adversarial perturbations becomes critical. If adversaries could reliably induce misclassification or evasion with small, biologically plausible edits, the reliability and safety of GFM-powered systems in real-world applications must be rigorously evaluated (Goodfellow et al., 2016).

We note that the term "biosecurity screening" is used here to refer specifically to GFM-powered classification as components of more comprehensive screening pipelines, rather than to end-to-end screening systems. Our results demonstrate adversarial vulnerabilities in the classification layer; the degree to which these vulnerabilities translate to evasion of full screening workflows depends on additional system-level factors that are outside the scope of this study.

Prior work has demonstrated that AI-driven sequence design can produce sequences that evade screening tools (Wittmann et al., 2024), that DNA injection and synthesis pipelines introduce exploitable attack surfaces (Farbiash & Puzis, 2020), and that jailbreak-style attacks can steer DNA language models toward harmful outputs (Zhang et al., 2025b). Broader assessments of AI-enabled biological risks further emphasize the need for rigorous evaluation of model robustness under realistic threat models (Nelson & Rose, 2023). Recent biosecurity assessment frameworks categorize emerging biotechnology risks and specifically identify genetic modification technologies and biological foundation models as threats expected to mature within the next decade, lowering barriers for adversaries seeking to develop harmful biological agents; notably, the majority of such tools are fully open-source and therefore readily accessible to threat actors (Webster et al., 2025). Consistent with the assessment that no single technology drives risk in isolation, but rather that risk emerges from the convergence and accessibility of tools, our work focuses on a critical vulnerability in GFMs: their susceptibility to minimal adversarial mutations could erode the robustness of downstream genomic classification tasks. The implications extend broadly to biosecurity checkpoints in DNA screening pipelines, including for example the detection of synthetic biology constructs, gain-of-function modifications, or enhanced pathogenic sequences.

Figure 1 summarizes the threat model and evasion mechanism considered in this work. Here, we use promoter sequence classification as a representative genomic classification task and a proxy for more complex classification tasks, such as pathogen classification. It is well suited to serve as an example due to its prevalence in existing benchmarks (Dalla-Torre et al., 2024; Zhou et al., 2024), its relevance for downstream biological decision-making, and the availability of well-characterized assays for experimental validation. Using adversarial sequences, we show that GFMs are vulnerable to certain classes of mutations, revealing failure modes that are not apparent from aggregate accuracy or confidence metrics alone. To mitigate these vulnerabilities, we propose an iterative black-box adversarial training procedure tailored to GFMs.

The central question of this work is: Can genomic foundation models be reliably evaded by an adversary who applies a small number of biologically plausible nucleotide edits to input sequences? We operationalize this question through a concrete task (promoter classification) and a concrete threat model (black-box query access with biologically constrained perturbations), and measure both attack effectiveness and the capacity of iterative adversarial training to mitigate identified vulnerabilities.

The paper is organized as follows. We review related work on genomic foundation models and adversarial robustness in genomics in Section 2. Section 3 describes the task formulation, threat model, attack procedures, and adversarial training framework. Section 4 presents empirical results on attack effectiveness, vulnerability structure, and transferability, followed by discussion and limitations in Section 5 and conclusions in Section 6.

2 Related Work

2.1 Genomic foundation models

The application of large-scale pretrained models to genomics has emerged as a powerful paradigm for learning dense representations from genomic sequences. DNABERT-2 (Zhou et al., 2024) adapts the BERT architecture to genomic sequences using nucleotide-level tokenization and pretraining on large-scale multi-species genomic corpora. The model demonstrates strong performance across diverse downstream tasks, including promoter identification, splice site prediction, and regulatory element classification. Similarly, the Nucleotide Transformer v2 (Dalla-Torre et al., 2024) family

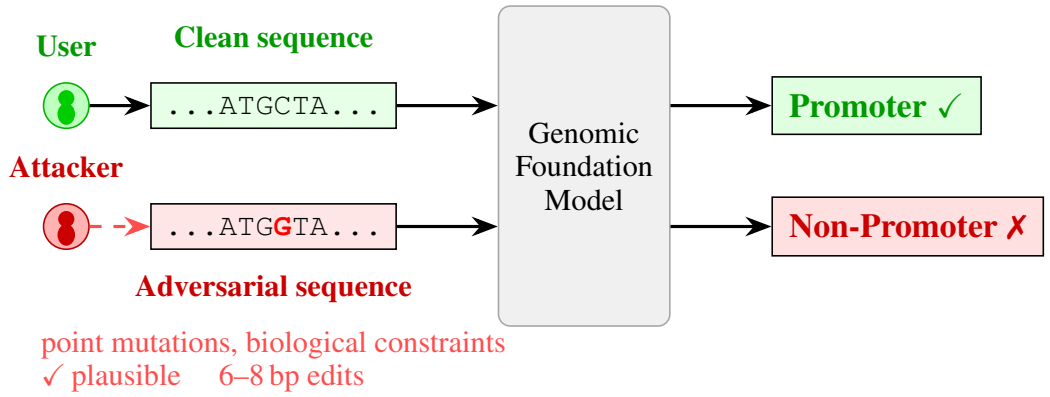

Figure 1: Illustrative threat model: a genomic foundation model (GFM) classifies sequences (e.g., promoter vs. non-promoter). **Benign**: normal classification. **Attacker**: an attacker perturbs the sequence with few nucleotide edits under biological constraints; the adversarial sequence remains plausible but causes misclassification.

provides foundation models at multiple scales (50M to 500M parameters), pretrained on extensive genomic datasets and evaluated for robustness and generalization. These models learn representations that capture sequence-level DNA patterns, regulatory signatures, and evolutionary signals, enabling transfer learning to task-specific genomic prediction problems. These approaches build on foundational work on representation learning and model scaling across domains (Bengio & LeCun, 2007; Hinton et al., 2006; Goodfellow et al., 2016).

2.2 Promoter sequences

Promoter datasets and experimental assays measuring promoter-driven gene expression are well established (Toktay et al., 2022; Alcaraz-Pérez et al., 2008; Fournier et al., 1999), providing a concrete biological context for evaluating model behavior. Recent work has also studied the mutational robustness of promoters from a biological perspective (Fuqua et al., 2025), highlighting that natural promoters exhibit varying degrees of tolerance to nucleotide-level perturbations, a property that computational models should ideally reflect.

2.3 Adversarial robustness in discrete and structured domains

Adversarial robustness has been extensively studied in continuous domains such as computer vision and natural language processing (Goodfellow et al., 2015; Eykholt et al., 2018; Tramèr et al., 2017; Gu et al., 2024). In contrast, genomic sequences pose distinct challenges: inputs are discrete (nucleotides), highly structured, and constrained by biological rules such as codon usage, GC content, and regulatory motifs. Discrete input spaces can admit adversarial examples with minimal edits (Kuleshov et al., 2021), and in genomics, even few-nucleotide perturbations can degrade GFM accuracy (Yoo et al., 2025). However, many existing attacks rely on gradient-based methods or unconstrained perturbations that may violate biological plausibility and therefore fail to reflect realistic threat models.

Recognizing these limitations, several benchmarking frameworks have emerged to evaluate adversarial robustness of GFMs under more domain-aware conditions. FIMBA (Skovorodnikov & Alkhzaimi, 2024) uses feature-importance-based attacks, GenoArmory (Luo et al., 2025) provides a unified evaluation framework for adversarial attacks on GFMs, and SafeGenes (Zhan et al., 2025) focuses on protein foundation models. Together, these efforts underscore the need for adversarial evaluation methods that incorporate biological plausibility constraints, discrete perturbation spaces, and task-relevant threat models, which standard approaches from vision and language domains do not capture.

2.4 ADVERSARIAL TRAINING AND TRANSFERABILITY

Adversarial training, i.e., incorporating adversarial examples into the training set, has been shown to improve robustness in vision and natural language processing, often at the cost of clean accuracy (Goodfellow et al., 2015; Carlini et al., 2019). In genomics, adversarial training remains comparatively underexplored, particularly under biologically constrained threat models. Transferability of adversarial examples across models is a critical concern for threat assessment: if adversarial sequences discovered for one model transfer to another, attackers can optimize against open-weight models and deploy attacks against black-box systems (Tramèr et al., 2017; Gu et al., 2024). Understanding transferability in genomic foundation models is therefore essential for assessing real-world biosecurity risks.

3 METHODOLOGY

3.1 MODELS, TASK, AND DATASET

We study promoter classification formulated as binary classification of fixed-length DNA sequences into promoter versus non-promoter classes. Promoter identification is widely used in genomic benchmarking, is relevant for downstream biological decision-making, and benefits from well-established experimental assays, making it a suitable testbed for robustness analysis.

We evaluate two genomic foundation models (GFMs): DNABERT-2 (Zhou et al., 2024), containing 117M parameters, and the Nucleotide Transformer v2-250m-multi-species model (Dalla-Torre et al., 2024), containing 250M parameters. Throughout this paper, we use "NT v2" and "Nucleotide Transformer v2" to refer exclusively to this 250M-parameter variant. Both models are large-scale pretrained encoders trained on multi-species genomic corpora and designed to capture sequence-level and regulatory patterns in DNA.

For both models, we train a classifier on top of the pretrained encoder; in the iterative adversarial training procedure we retrain only the classifier head with unfrozen base model to improve robustness.

We use the promoter classification dataset from the Eukaryotic Promoter Database (EPD) (Périer et al., 2000) and released as part of the Nucleotide Transformer benchmark (Dalla-Torre et al., 2024). The dataset consists of labeled promoter and non-promoter sequences and is publicly available on Hugging Face.[1] Further information about the training set and class split can be found in Section 3.5.

3.2 BLACK-BOX ADVERSARIAL ATTACK

Our approach assumes an adversary with no access to model internals, gradients, or training data, but with the ability to query model outputs such as class probabilities or confidence scores. This setting closely mirrors realistic deployment scenarios for GFM-powered genomics tools. Adversarial sequences are generated using iterative black-box optimization procedures via genetic algorithms. We use a genetic algorithm with a population size of 100, crossover rate of 0.9, and mutation rate of 0.3 to propose discrete nucleotide-change events under biological plausibility rules: GC-content bounds, mutation-budget limits, Ti/Tv bias, and position-dependent target-window constraints. The genetic algorithm runs for a fixed 200 generations per attack attempt with no early stopping criterion, to make a consistent attack effort across all sequences. These constraints ensure that generated sequences remain close to the original data and avoid biologically implausible artifacts. The fitness combines confidence drop (and prediction flip) and penalty for number of edits and for violating biological constraints.

Adversarial sequence generation is subject to explicit biological plausibility constraints implemented in the genetic algorithm: we limit the total number of perturbations (max_perturbations) (1000 Genomes Project Consortium, 2015) and score each candidate sequence with a biological plausibility score based on enabled constraint checks. In the implementation used in this work, the biological plausibility components include (i) a GC-content deviation constraint (Saxonov et al., 2006), (ii) a motif-preservation score over a fixed list of short regulatory motifs

[1] https://huggingface.co/collections/InstaDeepAI/nucleotide-transformer

Table 1: Training and optimization hyperparameters.

Hyperparameter	Baseline training	Adversarial retraining
Optimizer	AdamW	AdamW
Encoder	Frozen (train classifier head only)	Unfrozen (full-model fine-tuning)
Learning rate	1×10^{-3}	2×10^{-5}
Weight decay	0.01	0.01
LR schedule	Linear warmup + decay	Linear warmup + decay
Warmup steps	500	500
Batch size	16	16
Epochs per training run	3	3
Max sequence length	300	300
Seed	42	42

when motif preservation is enabled, and (iii) a transition-preference score that penalizes transversions relative to transitions when transition preference is enabled (Neininger et al., 2019). This biological score is incorporated into the GA fitness via a penalty term rather than as a hard rejection rule.

The biological plausibility score $s_{\mathrm{bio}} \in [0, 1]$ is computed in the code as the arithmetic mean of the enabled constraint components:

$$s_{\mathrm{bio}}(x, x') = \frac{1}{K} \sum_{k=1}^{K} s_k(x, x'), \quad (1)$$

where (for the active configuration) K includes: GC validity $s_{\mathrm{gc}} \in \{0, 1\}$ given by $|\mathrm{GC}(x') - \mathrm{GC}(x)| \leq \delta$ (with $\delta = 0.05$ in our config), motif preservation $s_{\mathrm{motif}} \in [0, 1]$ computed from the relative preservation of counts of short motifs (e.g., TATA, CAAT) when enabled, and a transition-preference component $s_{\mathrm{ts}} \in [0, 1]$ that assigns higher scores to transition substitutions than to transversions and averages over the edited positions (with $s_{\mathrm{ts}} = 1$ when no substitutions occur).

3.3 Defense Method

We employ an iterative adversarial training procedure tailored to GFMs. After an initial baseline evaluation of each model using the unmutated ("clean") test set, we (1) regenerate adversarial examples in each iteration by re-running the black-box genetic algorithm against the current checkpoint on a held-out attack pool, filtering the attack pool to sequences that the current model initially predicts with the true label; (2) augment the training set by accumulating the newly generated successful adversarial examples together with previously accumulated successes, selecting for each successful attack the top-N generations by wrong-class confidence and deduplicating by sequence; and (3) re-evaluate both clean accuracy and attack success.

Crucially, entirely new adversarial examples are generated fresh at each iteration by running the genetic algorithm against the most recently retrained model checkpoint, so the attacker adapts to an evolving defender rather than attacking a static one. Training data for iteration k consists of the original training set augmented with the cumulative union of all successful adversarial examples from iterations 1 through k, deduplicated by sequence.

We repeat this regeneration-and-retraining procedure for 10 iterations per model, so the attacker always optimizes against the most recently retrained checkpoint (adapting to a moving defender rather than attacking a static model). We evaluate robustness on the held-out test set within this biologically constrained GA threat model; robustness to different attacker strategies or perturbation regimes is left for future work. Figure 2 provides an overview of the iterative adversarial training pipeline.

Full training and optimization hyperparameters are provided in Table 1.

Training and optimization hyperparameters. Table 1 summarizes the training and optimization hyperparameters used for baseline training and iterative adversarial training.

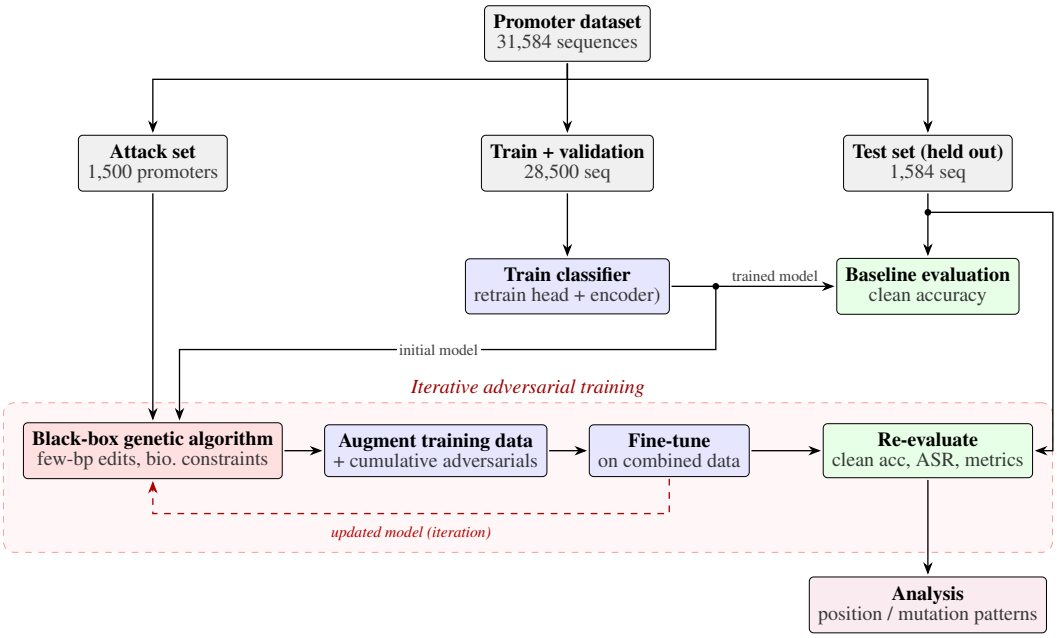

Figure 2: Adversarial evaluation and iterative adversarial training pipeline: data splits keep attack set (promoters only) separate from test to avoid leakage; baseline eval, then iterate over black-box GA on attack set, combine original + cumulative adversarials, fine-tune, re-evaluate (clean accuracy, ASR, perturbations), and optionally analyze position/mutation patterns and generate robustness plots.

An important open question is whether robustness gains from adversarial training against genetic algorithm attacks generalize to other attack families. Our iterative training procedure hardens the model against a specific class of black-box optimization attacks, but alternative attack strategies such as gradient-based methods (e.g., FGSM (Goodfellow et al., 2015)), feature-importance-guided perturbations (e.g., FIMBA (Skovorodnikov & Alkhzaimi, 2024)), or codon-level and backtranslation-based transformations (Yoo et al., 2025) may exploit different model weaknesses. The low cross-model transferability observed in Section 4.3 provides indirect evidence that different attack surfaces exist, since perturbations effective against one model's decision boundary are largely ineffective against another's. Evaluating robustness under diverse attack families and establishing whether adversarial training against one attack type confers partial robustness to others is a necessary direction for future work. Recent unified benchmarks such as GenoArmory (Luo et al., 2025) provide infrastructure for such multi-attack evaluations.

The flowchart explicitly reflects iterative retraining with cumulative adversarial augmentation and re-evaluation across rounds.

3.4 Model Descriptions

DNABERT-2. DNABERT-2 (Zhou et al., 2024) is an efficient foundation model for multi-species genome understanding. It uses a BERT-style encoder pretrained on genomic sequences at the nucleotide level and is designed for downstream tasks such as promoter identification, splice site prediction, and regulatory element classification. The model we evaluate has 117M parameters. We use the publicly released checkpoint and train a classifier on top of the encoder; during iterative adversarial training we retrain the classification head with the encoder for promoter classification.

Nucleotide Transformer v2 (250M). The Nucleotide Transformer v2 (Dalla-Torre et al., 2024) is a family of foundation models for human genomics that are pretrained on large-scale genomic corpora and evaluated for robustness and generalization across multiple prediction tasks. The family includes variants at different scales (e.g., 50M, 100M, 250M parameters). In this work we use the 250M-parameter multi-species variant (Nucleotide Transformer v2-250m-multi-species). Among

the publicly available Nucleotide Transformer v2 variants, the 250M model offers the best trade-off between capacity and efficiency and achieves strong performance on standard genomic benchmarks. As with DNABERT-2, during iterative adversarial training we retrain the classification head with a unfrozen encoder for promoter classification.

3.5 Promoter Dataset

Experiments in this work use the promoter_all dataset from the nt_new_data_promoter repository[2]. Both train and test data share the same structure:

- **Columns:** `sequence` (300 bp DNA), `name` (genomic coordinates, e.g. `chr:start--end|id`), `label` (0 = non-promoter, 1 = promoter), `task` (`promoter_all`).
- **Sequence length:** 300 nucleotides per sequence.

Train set: 25,650 sequences (13,455 non-promoter, 12,195 promoter).

Validation set: 2,850 sequences (1,495 non-promoter, 1,355 promoter).

Attack set: 1,500 sequences (promoters only), held out from training and used exclusively for generating adversarial examples.

Test set: 1,584 sequences (792 non-promoter, 792 promoter; balanced). This held-out set is used for all reported accuracy and evaluation metrics and is never used for attack generation or training.

4 Results

4.1 Iterative adversarial training

The baseline classification performances of the models correspond to iteration 0 of the test accuracy panel in Figure 3 i.e. 88% and 86% for DNABERT-2 and Nucleotide Transformer v2 respectively. Clean test accuracy remains stable across iterations for both models, staying within the range of approximately 86–90%, indicating that adversarial training does not compromise performance on unperturbed sequences. The clean accuracy measures of these models correspond to a Matthews Correlation Coefficient (MCC) of 0.75 for DNABERT-2 and 0.74 for NTv2 and match with the results reported in Dalla-Torre et al. (2024). Figure 3 presents the results of iterative adversarial training for DNABERT-2 and Nucleotide Transformer v2 across multiple robustness metrics.

Attack success rate (ASR) is measured as the number of sequences that are misclassified by the model in the 200 generations of the genetic algorithm (Holland, 1992) in each iteration of adversarial training. ASR is initially high, at approximately 40–45%, and is similar for both models. Further, iterative adversarial training improves robustness without degrading clean accuracy, but does not eliminate vulnerability. As adversarial training progresses, attack success rate fluctuates but exhibits an overall downward trend during iterations, stabilizing around 20–30% by tenth iteration. Detailed results about the number of successful attacks can be found in Table 2.

Biological plausibility scores remain consistently between 0.9–0.98, indicating that successful adversarial sequences continue to satisfy biological constraints such as nucleotide composition and sequence structure.

The average number of perturbations required for successful attacks remains close to the imposed budget of eight nucleotide substitutions. This corresponds to 2.7% of base pairs being modified. This indicates that attacks remain perturbation-efficient despite adversarial training, and that observed robustness improvements do not arise from forcing large or unrealistic sequence changes.

Both DNABERT-2 and Nucleotide Transformer v2 exhibit highly comparable robustness across all iterations, with similar attack success rate trajectories and perturbation efficiency. Under the considered black-box genetic attack paradigm, this similarity suggests that differences in model size

[2] `https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark`

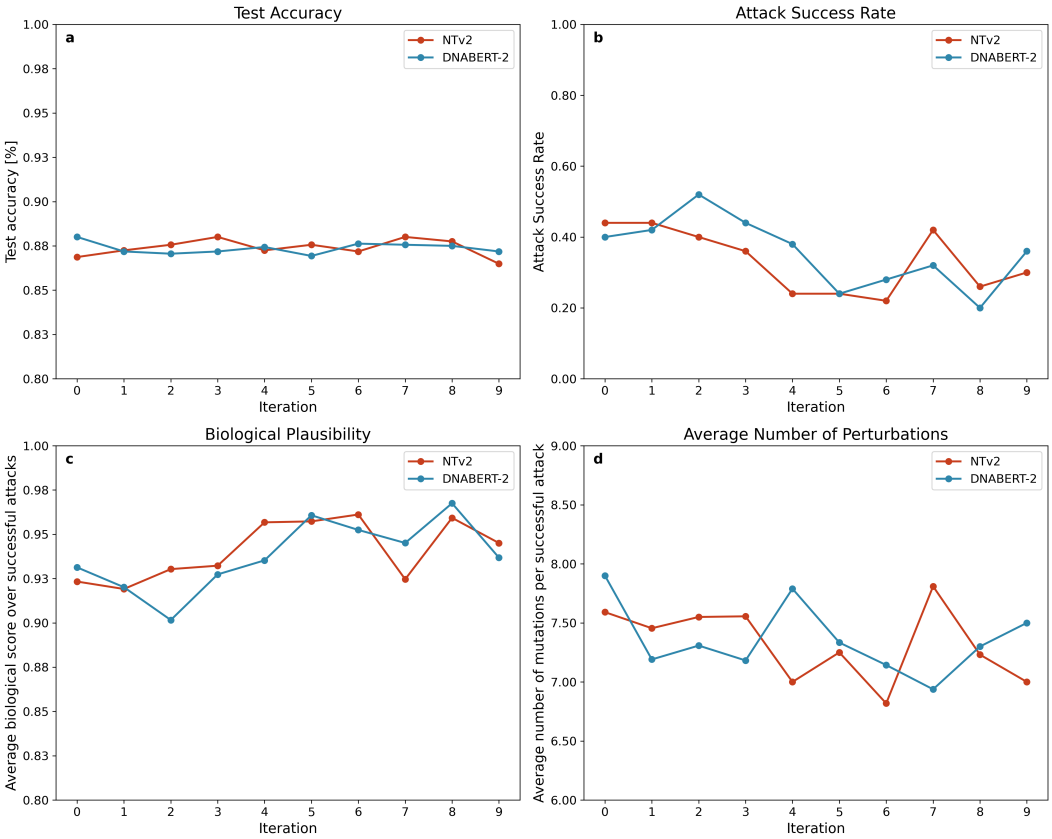

Figure 3: Iterative adversarial training: DNABERT-2 and Nucleotide Transformer v2 on NT promoter dataset. Test accuracy (a), attack success rate (b), average biological score (c), and average perturbations per successful attack (d) over 10 iterations.

and architecture alone do not confer a clear robustness advantage, pointing instead to factors such as training data distribution or representation learning objectives as potentially more influential.

4.2 Mutation type and regional analysis

Beyond measuring attack success, we introduce a static computational analysis of adversarial perturbations to characterize the structure of model vulnerabilities. We analyze only successful attacks, defined as cases in which the model prediction flips between promoter and non-promoter classes, and compare adversarially generated sequences across DNABERT-2 and Nucleotide Transformer v2 to identify systematic patterns in mutation behavior.

Across successful adversarial examples, transversions (e.g., A↔C, G↔T) account for a larger fraction of effective mutations than transitions (A↔G, C↔T) (Figure 4). This trend is observed consistently for both DNABERT-2 and Nucleotide Transformer v2. Transversions also exhibit higher average occurrences per sequence, with comparable variability across models (Figure 5). Together, these results indicate that, within successful attacks, GFMs are more susceptible to transversion mutations than to transitions. Mutation-type statistics for successful adversarial attacks are summarized in Table 2.

Position-level analysis reveals a strong regional bias in effective perturbations. The first 100 base pairs (bp) of the input sequence (positions 1–100) contain the 64.3% and 55.1% of successful adversarial mutations for DNABERT-2 and Nucleotide Transformer v2 models, respectively, while positions 101-300 account for the remaining mutations (Figure 6).

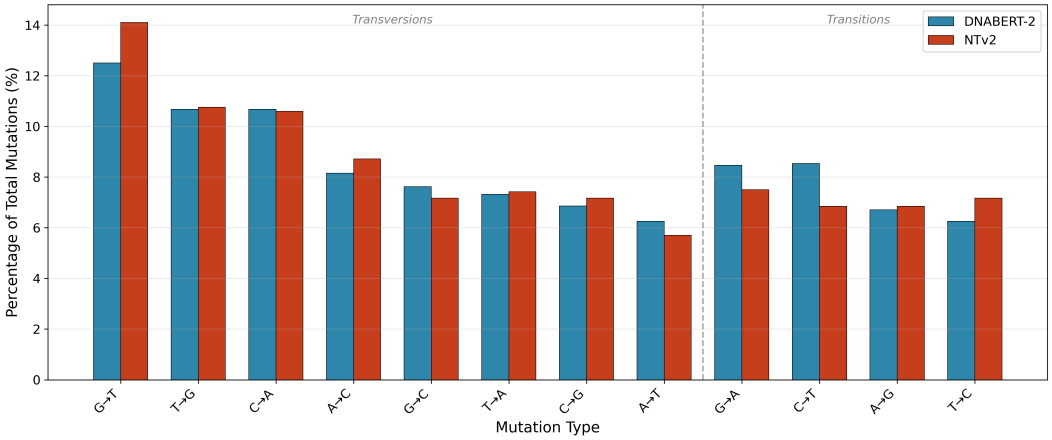

Figure 4: Mutation type frequencies comparison between Nucleotide Transformer v2 and DNABERT-2 across 10 adversarial training iterations (successful attacks only). Transversions (left of dashed line) are sorted by total count, followed by transitions (right).

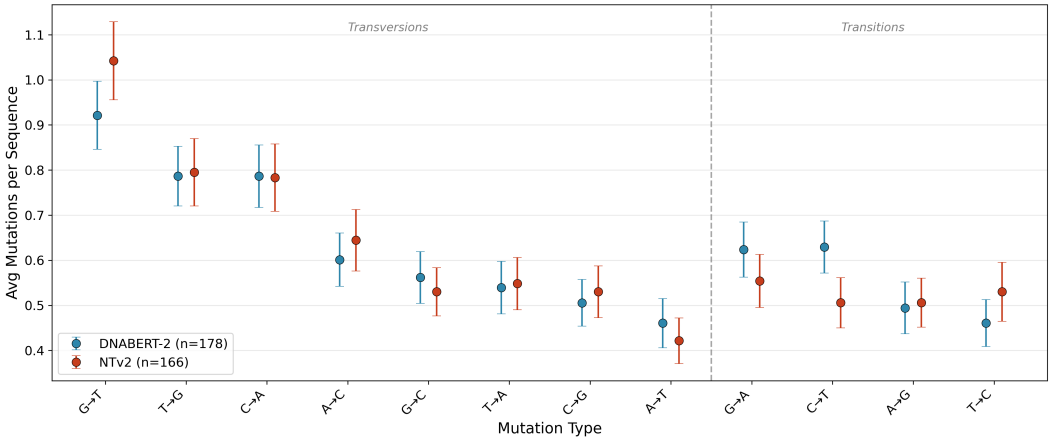

Figure 5: Average mutations per sequence by mutation type for DNABERT-2 and Nucleotide Transformer v2 (successful attacks only). Error bars show 95% CI from 1000 bootstrap resamples (successful sequences resampled with replacement).

Together, these analyses show that DNABERT-2 and Nucleotide Transformer v2 exhibit highly similar vulnerability profiles, including consistent concentration of adversarial perturbations in the 5' region and a shared susceptibility to transversions over transitions. At the same time, subtle differences in effective substitution patterns suggest that specific failure modes can vary across architectures. Importantly, these structured vulnerabilities are not apparent from aggregate accuracy or confidence metrics alone.

4.3 Transferability of adversarial examples across GFMs

We conducted a preliminary investigation of whether adversarial examples crafted for one GFM transfer to another, which has important implications for threat assessment. High transferability would imply that a single attack could compromise multiple screening systems, and that attackers could optimize adversarial sequences on open-weight models to target proprietary deployments.

To evaluate transferability, we train fresh instances of each model using identical configurations and test adversarial examples that successfully flip predictions on one model against the other. Specifically, we evaluate 178 successful DNABERT-2 adversarial examples on Nucleotide Transformer v2, and 166 successful Nucleotide Transformer v2 adversarial examples on DNABERT-2. We observe

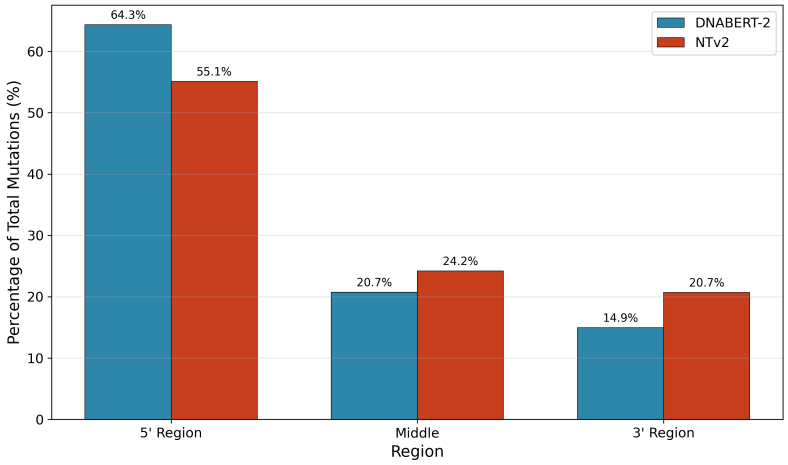

Figure 6: Distribution of mutation location in successful attacks as observed in DNABERT-2 and NTv2 across 10 adversarial training iterations.

Table 2: Summary of successful adversarial attacks and mutation-type statistics for DNABERT-2 and Nucleotide Transformer v2 (NT v2). Statistics are computed over successful attacks only, defined as cases in which the model prediction flips between promoter and non-promoter classes.

Metric	**DNABERT-2**	**NT v2**
Number of successful attacks	178	166
Total transitions	393	348
Total transversions	919	879
Transition percentage (%)	29.9	28.4
Transversion percentage (%)	70.0	71.6
Average transitions per successful adversarial sequence	2.21	2.10
Average transversions per successful adversarial sequence	5.16	5.30

low transferability between models in both directions: only 2.8 % of DNABERT-2 adversarial examples transfer to Nucleotide Transformer v2 (5 out of 178), while 5.4 % of Nucleotide Transformer v2 examples transfer to DNABERT-2 (9 out of 166).

These preliminary results indicate that, despite similar vulnerability profiles in terms of mutation types and regional sensitivities (Section 4.2), the two models exhibit largely non-overlapping adversarial failure modes. Adversarial perturbations that exploit one model's weaknesses are typically ineffective against the other, suggesting that the learned decision boundaries differ substantially across architectures.

5 Discussion

This work proposes a principled computational framework for evaluating the adversarial robustness of GFMs under biologically constrained and black-box threat models. Here, we combined black-box genetic attacks, static analysis of adversarial perturbations, and iterative adversarial training. Our analysis reveals that both models are vulnerable to perturbation-efficient adversarial attacks. However, high plausibility scores alone do not guarantee that the generated sequences correspond to genuinely adversarial effects in real biological systems, particularly in the absence of experimental validation. Incorporating more biologically grounded scoring functions and empirical validation would be necessary to establish real-world adversarial impact, and is left to future work.

The structure of successful adversarial perturbations reveals consistent and biologically meaningful patterns. Across both DNABERT-2 and Nucleotide Transformer v2, models are more susceptible

to transversions, compared to transitions, accurately reflecting biological mutation patterns in the human genome (Neininger et al., 2019). While both models show similar overall mutation type preferences, some differences emerge: in the current setting Nucleotide Transformer v2 exhibits higher rates of G→T transversions, while DNABERT-2 shows elevated rates of C→T and G→A transitions. Mutations in the 5' region of the sequences are more successful, a pattern that is expected given the dataset construction (promoter activity encoded in the first 70 bases). The similarity of the vulnerability profiles across both GFM architectures suggests that, under the considered threat model, robustness is shaped more strongly by the training data and the classification task than by model size or architectural differences alone.

ASR reduces from an initial value of 40-45% to a value between 20-30%. Although this improves robustness moderately, successful attacks remain frequent even after 10 rounds of adversarial training, highlighting the persistent effectiveness of black-box optimization attacks. This indicates that such attacks remain effective despite adversarial training, raising concerns for the deployment of GFMs in high-stakes genomic and biosecurity settings.

Despite shared vulnerability patterns, adversarial examples show very low transferability between architectures. Attacks crafted for one model rarely succeed against the other, indicating that DNABERT-2 and Nucleotide Transformer v2 learn distinct decision boundaries even when trained on the same task.

An important limitation of this analysis is that adversarial attacks are evaluated at the level of the downstream classification task, rather than directly probing the internal representations of the foundation models themselves. As a result, the observed lack of transferability may reflect differences in the task-specific classifier heads, differences in how representation in each GFM is utilized, or a combination of both. Disentangling attacks on the underlying foundation model from attacks on the classifier head would require a more fine-grained analysis of representation-level vulnerabilities, which we leave to future work. Despite this limitation, the observed model specificity has practical implications for defense. Robustness improvements learned by one architecture may not automatically generalize to others, suggesting that ensemble-based screening systems that combine multiple genomic foundation models could offer increased resilience to black-box adversarial attacks.

Our findings highlight that adversarial risk in genomics is both structured and model-dependent, and that evaluating robustness requires joint consideration of biological constraints, task structure, and realistic attacker capabilities.

6 Conclusions

This work presents an evaluation of genomic foundation models (GFMs) that integrates biological structure, adversarial machine learning, and biosecurity considerations. We analyze the robustness of DNABERT-2 and Nucleotide Transformer v2 under biologically constrained black-box attacks on sequences used in a binary classification task and show that both models remain vulnerable to adversarial manipulation despite adversarial training.

We adopt a threat model reflecting realistic deployment scenarios in which inputs are biologically constrained and model internals may be inaccessible to attackers, even when open-source implementations exist. Within this framework, we identify structured vulnerability patterns that are not apparent from aggregate accuracy or confidence metrics alone.

Although adversarial sequences were constrained to biological plausibility, this does not guarantee functional validity. To determine whether successful perturbations correspond to real biological effects rather than model-specific artifacts, future work will include targeted wet-lab validation by synthesizing selected wild-type and adversarial promoter sequences. Additional directions include incorporating richer biologically grounded scoring functions, extending evaluations to other genomic tasks and datasets, exploring alternative black-box attack strategies, and studying transferability across architectures.

Overall, we position adversarial robustness in genomics as a joint biological and machine learning challenge, laying groundwork for safer and more reliable deployment of GFMs in biosecurity-relevant and broader biological applications.

ACKNOWLEDGMENTS

We thank Jacob Haimes for helpful feedback and editing suggestions on our manuscript. We also thank Marie Lopez and Tessa Alexanian for valuable discussions on the biosecurity implications of this work. This project originated at the CBRN AI Risks Research Sprint organized by Apart Research in September 2025. The continued development of this work was supported by the Apart Research fellowship program.

REFERENCES

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, September 2015. ISSN 1476-4687. doi: 10.1038/nature15393. URL http://dx.doi.org/10.1038/nature15393.

Francisca Alcaraz-Pérez, Victoriano Mulero, and María L Cayuela. Application of the dual-luciferase reporter assay to the analysis of promoter activity in zebrafish embryos. *BMC Biotechnology*, 8(1), October 2008. ISSN 1472-6750. doi: 10.1186/1472-6750-8-81. URL http://dx.doi.org/10.1186/1472-6750-8-81.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston (eds.), *Large Scale Kernel Machines*. MIT Press, 2007. URL http://yann.lecun.com/exdb/publis/pdf/bengio-lecun-07.pdf.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. URL https://arxiv.org/abs/1902.06705.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, November 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z. URL http://dx.doi.org/10.1038/s41592-024-02523-z.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634. IEEE, June 2018. doi: 10.1109/cvpr.2018.00175. URL http://dx.doi.org/10.1109/CVPR.2018.00175.

Dor Farbiash and Rami Puzis. Cyberbiosecurity: DNA injection attack in synthetic biology, 2020. URL https://arxiv.org/abs/2011.14224.

Bénédicte Fournier, Annie Gravel, David C. Hooper, and Paul H. Roy. Strength and regulation of the different promoters for chromosomal ß-lactamases of klebsiella oxytoca. *Antimicrobial Agents and Chemotherapy*, 43(4):850–855, April 1999. ISSN 1098-6596. doi: 10.1128/aac.43.4.850. URL http://dx.doi.org/10.1128/AAC.43.4.850.

Timothy Fuqua, Stepan Denisov, Mato Lagator, and Andreas Wagner. Strong promoters are mutationally robust, 2025. URL http://dx.doi.org/10.1101/2025.10.20.683477.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, Xiaochun Cao, and Philip Torr. A survey on transferability of adversarial examples across deep neural networks, 2024. URL https://arxiv.org/abs/2310.17626.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006. ISSN 1530-888X. doi: 10.1162/neco.2006.18.7.1527. URL http://dx.doi.org/10.1162/neco.2006.18.7.1527.

John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, April 1992. ISBN 9780262275552. doi: 10.7551/mitpress/1090.001.0001. URL http://dx.doi.org/10.7551/mitpress/1090.001.0001.

Volodymyr Kuleshov, Evgenii Nikishin, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. Quantifying and understanding adversarial examples in discrete input spaces, 2021. URL https://arxiv.org/abs/2112.06276.

Haozheng Luo, Chenghao Qiu, Yimin Wang, Shang Wu, Jiahao Yu, Zhenyu Pan, Weian Mao, Haoyang Fang, Hao Xu, Han Liu, Binghui Wang, and Yan Chen. Genoarmory: A unified evaluation framework for adversarial attacks on genomic foundation models, 2025. URL https://arxiv.org/abs/2505.10983.

Kerstin Neininger, Tobias Marschall, and Volkhard Helms. Snp and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome. *PLOS ONE*, 14(4):e0214816, April 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0214816. URL http://dx.doi.org/10.1371/journal.pone.0214816.

Cassidy Nelson and Sophie Rose. Understanding AI-facilitated biological weapon development, October 2023. URL http://dx.doi.org/10.71172/nm7j-qzt1.

Rouaïda Cavin Périer, Viviane Praz, Thomas Junier, Claude Bonnard, and Philipp Bucher. The eukaryotic promoter database (epd). *Nucleic Acids Research*, 28(1):302–303, January 2000. ISSN 1362-4962. doi: 10.1093/nar/28.1.302. URL http://dx.doi.org/10.1093/nar/28.1.302.

Serge Saxonov, Paul Berg, and Douglas L. Brutlag. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417, January 2006. ISSN 1091-6490. doi: 10.1073/pnas.0510310103. URL http://dx.doi.org/10.1073/pnas.0510310103.

Heorhii Skovorodnikov and Hoda Alkhzaimi. FIMBA: Evaluating the robustness of AI in genomics via feature importance adversarial attacks, 2024. URL https://arxiv.org/abs/2401.10657.

Yagmur Toktay, Bengisu Dayanc, and Serif Senturk. Engineering and validation of a dual luciferase reporter system for quantitative and systematic assessment of regulatory sequences in chinese hamster ovary cells. *Scientific Reports*, 12(1), April 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-09887-2. URL http://dx.doi.org/10.1038/s41598-022-09887-2.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples, 2017. URL https://arxiv.org/abs/1704.03453.

Toby Webster, Richard Moulange, Barbara Del Castello, James Walker, Sana Zakaria, and Cassidy Nelson. Global risk index for AI-enabled biological tools: Summary assessment and methods report, September 2025. URL http://dx.doi.org/10.71172/wjyw-6dyc.

Bruce J. Wittmann, Svetlana P. Ikonomova, Fernanda Piorino, David J. Ross, Samuel W. Schaffter, Olga Vasilyeva, Eric Horvitz, James Diggans, Elizabeth A. Strychalski, Sheng Lin-Gibson, and Geoffrey J. Taghon. Experimental evaluation of AI-driven protein design risks using safe biological proxies, December 2024. URL https://www.biorxiv.org/content/10.1101/2024.12.02.626439v1.

Hyunwoo Yoo, Haebin Shin, Kaidi Xu, and Gail Rosen. Exploring adversarial robustness in classification tasks using dna language models, 2025. URL https://arxiv.org/abs/2409.19788.

Huixin Zhan, Clovis Barbour, and Jason H. Moore. SafeGenes: Evaluating the adversarial robustness of genomic foundation models, 2025. URL https://arxiv.org/abs/2506.00821.

Zaixi Zhang, Souradip Chakraborty, Amrit Singh Bedi, Emilin Mathew, Varsha Saravanan, Le Cong, Alvaro Velasquez, Sheng Lin-Gibson, Megan Blewett, Dan Hendrycs, Alex John London, Ellen Zhong, Ben Raphael, Adji Bousso Dieng, Jian Ma, Eric Xing, Russ Altman, George Church, and Mengdi Wang. Generative AI for biosciences: Emerging threats and roadmap to biosecurity, 2025a. URL https://arxiv.org/abs/2510.15975.

Zaixi Zhang, Zhenghong Zhou, Ruofan Jin, Le Cong, and Mengdi Wang. GeneBreaker: Jailbreak attacks against DNA language models with pathogenicity guidance, 2025b. URL https://arxiv.org/abs/2505.23839.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome, 2024. URL https://arxiv.org/abs/2306.15006.