Comparative Analysis of Deep Learning Models for Predicting Causative Regulatory Variants

Gaetano Manzo[1], Kathryn Borkowski[1,2] and Ivan Ovcharenko[1,*]

[1]Computational Biology Branch, Division of Intramural Research, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, USA.,

[2]Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH, USA.

[*]Corresponding author

Email: ovcharen@nih.gov

Abstract

Motivation: Genome-wide association studies (GWAS) have identified numerous noncoding variants associated with complex human diseases, disorders, and traits. However, resolving the uncertainty between GWAS association and causality remains a significant challenge. The small subset of noncoding GWAS variants with causative effects on gene regulatory elements can only be detected through accurate methods that assess the impact of DNA sequence variation on gene regulatory activity. Deep learning models, such as those based on Convolutional Neural Networks (CNNs) and transformers, have gained prominence in predicting the regulatory effects of genetic variants, particularly in enhancers, by learning patterns from genomic and epigenomic data. Despite their potential, selecting the most suitable model is hindered by the lack of standardized benchmarks, consistent training conditions, and performance evaluation criteria in existing reviews.

Results: This study evaluates state-of-the-art deep learning models for predicting the effects of genetic variants on enhancer activity using nine datasets stemming from MPRA, raQTL, and eQTL experiments, profiling the regulatory impact of 54,859 SNPs across four human cell lines. The results reveal that CNN models, such as TREDNet and SEI, consistently outperform other architectures in predicting the regulatory impact of SNP. However, hybrid CNN-transformer models, such as Borzoi, display superior performance in identifying causal SNPs within a linkage disequilibrium block. While fine-tuning enhances the performance of transformer-based models, it remains insufficient to surpass CNN and hybrid models when evaluated under optimized conditions.

Introduction

Genome-wide association studies (GWAS) have revealed that around 95% of disease-associated genetic variants occur in non-coding regions of the human genome, with causative variants commonly affecting regulatory elements that modulate gene expression [1, 2, 3]. These regulatory variants can profoundly affect phenotypes and alter disease susceptibility via dysregulation of their target genes [4].

Deep learning models have emerged as powerful tools for predicting the regulatory effects of genetic variants, particularly in enhancers. These methods leverage high-throughput genomic and epigenomic data to learn DNA sequence patterns associating sequence features with regulatory activity. For instance, convolutional neural networks (CNNs) have been used to identify variants that disrupt transcription factor binding or chromatin accessibility, providing insights into their potential phenotypic impact (e.g., DeepSEA [5], SEI [6], and TREDNet [7]). Similarly, modern, large transformer-based models can very accurately predict cell-type-specific regulatory effects, such as DNA methylation changes and fine-mapping of disease-associated loci; these primarily include DNABERT series [8, 9], Nucleotide Transformer series [10, 11], and Enformer [12].

Selecting the most suitable model for detecting the regulatory effects of genetic variants remains a significant challenge despite several surveys offering detailed overviews of the deep learning ecosystem in this domain [13, 14, 15]. These surveys have highlighted the unique strengths and limitations of various models. However, they lack a unified

framework for assessment. Specifically, existing reviews often fail to benchmark models on a standardized dataset, train or fine-tune them under consistent conditions, and evaluate their performance using uniform criteria. Furthermore, there is a fundamental difference between benchmarking models on regulatory regions versus regulatory variants, which is rarely addressed. While regulatory region analyses focus on identifying broader functional elements, regulatory variant assessments require evaluating the impact of specific sequence alterations within these regions, presenting distinct challenges and opportunities for model evaluation.

In this study, we evaluate state-of-the-art deep learning models to predict the effects of genetic variants on enhancer activity in the human genome. Our approach involved curating and integrating nine datasets derived from diverse experimental methodologies, including massively parallel reporter assay (MPRA), reporter assay quantitative trait loci (raQTL), and expression quantitative trait loci (eQTL) studies. These datasets encompass 54,859 single-nucleotide polymorphisms (SNPs) in enhancer regions across four human cell lines.

We tackle three tasks: predicting log2-fold changes in enhancer activity, classifying SNPs by regulatory impact, and identifying causal SNPs within LD blocks. We also assess how data variability affects model performance, emphasizing the importance of data quality. Finally, we illustrate how enhancer detection models can be effectively leveraged for variant effect assessment, highlighting their practical utility in interpreting genomic data and prioritizing candidate regulatory variants.

To ensure a robust and thorough assessment, we fine-tuned 22 deep learning models for each cell line, systematically exploring a broad spectrum of architectural designs

and parameter configurations. Our results indicate that CNN models outperform more "advanced" architectures, such as transformers, on causative regulatory variant detection. However, fine-tuning significantly boosts the performance of transformer-based architectures, revealing their potential to surpass CNNs under optimized conditions.

Results

Comparison of Enhancer Variant Prediction Models

We assessed several state-of-the-art deep learning models for their ability to predict the effects of genetic variants on enhancer activity in humans. The selected models differ significantly, not only in terms of architecture and parameters but also in the type of data used for training or pre-training. More importantly, these models were originally designed for different purposes: some, like Nucleotide Transformer [10] and DNABERT2 [9], were developed for a broad range of tasks, while others, such as Geneformer [13] and TREDNet [7], were tailored for more specific applications. Crucially, this means we are not only comparing model architectures but also the underlying training datasets, which introduces additional complexity when benchmarking their performance. To further disentangle these effects, we also evaluate how different datasets impact the predictions of the same model. Despite these differences, both model categories have been adapted to predict the effects of enhancer variants (see Materials and Methods). Our evaluation encompasses a range of architectural and parameter configurations across model families, focusing on 54,859

single-nucleotide polymorphisms located in enhancer regions from nine datasets originating from four cell lines: K562, HepG2, NPC, and HeLa.

We performed regression analyses between model predictions and experimental log2-fold changes for variant effects in enhancers across the human genome sequence (Pearson Correlation Fig. 1, Spearman Correlation Fig. S1, Materials and Methods). CNN-based models, such as TREDNet and SEI, demonstrate the best alignment with experimentally recorded variant effects, achieving Pearson correlation coefficients of 0.297 and 0.276, respectively, while the most accurate transformer-based model, Nucleotide Transformer v2, scores only at 0.105, well below the level of statistical uncertainty. These findings highlight the effectiveness of CNNs in recognizing spatial and structural genomic patterns essential for predicting enhancer variant effects. Convolutional layers identify genomic features, such as local motifs and epigenetic markers, that influence genetic variant effects [16].

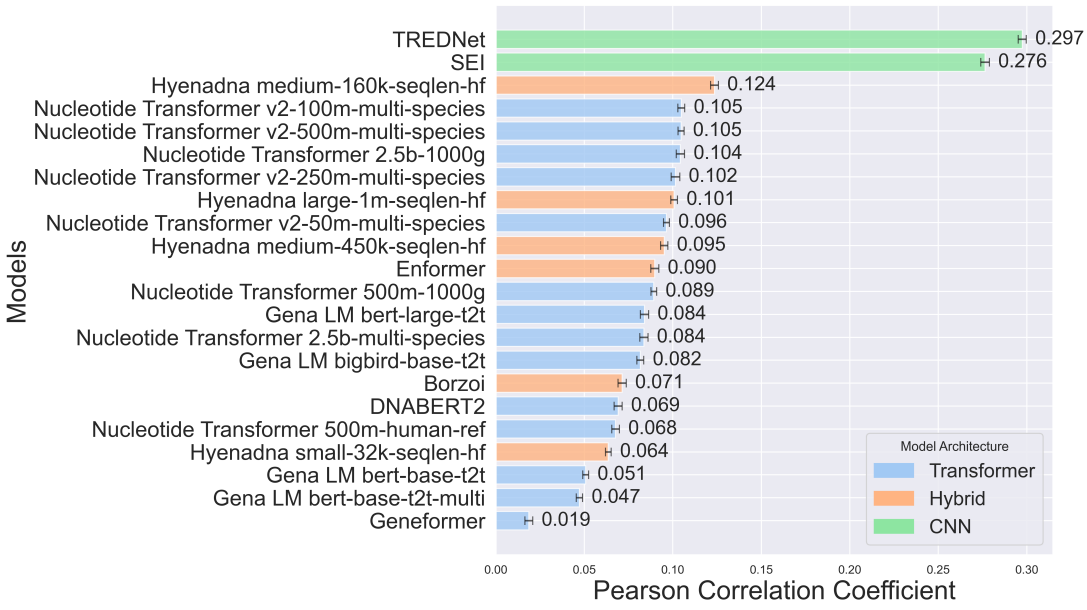

Fig 1: Pearson correlation coefficients between model predictions and experimental log2-fold difference for enhancer variant effects across the human genome. Bar colors denote model architectures (CNN: green, transformer: blue, and hybrid: orange). All correlations have p-values < 0.05, and error bars show variance.

This capability is crucial for capturing the complex genomic relationships underlying gene regulation and enhancer activity.

In contrast, transformer-based models fine-tuned for specific cell lines show lower correlation values (0.019–0.105). Despite their ability to model long-range relationships and complex sequences, they struggle to detect the subtle genomic patterns linked to regulatory variant effects.

This challenge likely stems from the complexity of genomic data, lower granularity in feature detection compared to CNNs, and the extensive data required to fully exploit transformer capabilities.

To bridge the gap between local pattern recognition and long-range dependency modeling, hybrid models combine components such as CNNs, transformers, and LSTMs. Models like the HyenaDNA series and Borzoi show moderate correlations, with HyenaDNA medium-160k-seqlen-hf achieving the highest among hybrids at 0.124. The integration of diverse components enhances performance compared to transformer-based models while optimizing computational costs relative to CNN-based models. However, despite their architectural versatility, these hybrid models do not surpass the performance of CNN-based models, highlighting a trade-off between flexibility and focused efficiency (Fig. 1, Fig. S1).

We acknowledge that our classification groups together all models that are not purely CNN- or transformer-based, despite significant architectural differences—particularly in the case of HyenaDNA, which diverges substantially from attention-based models. Nonetheless, our primary objective is to establish a broad guideline for enhancer variant prediction models, rather than to define rigid model categories. The classification task

142 provides a deeper understanding of model performance by evaluating true positive and
143 true negative rates in predicting variants that upregulate or downregulate gene
144 expression (Fig. 2). Among 54,859 experimental SNPs, approximately half are linked
145 to upregulation and half to downregulation.
146

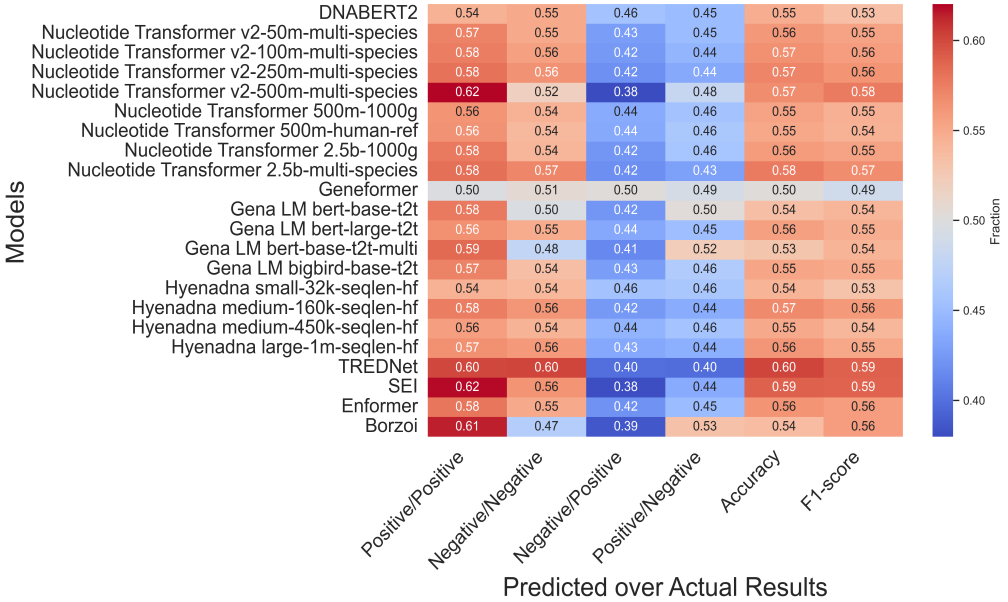

Fig 2: Heatmap of model variant predictions (Predicted) versus experimental values (Actual Results). The color intensity represents the fraction of values predicted as positive (negative) relative to the experimental positive (negative) values. Red indicates higher fractions (desired in the first two columns from the left), while blue indicates lower fractions (desired in the follow two columns). Accuracy and F1-score are indicated in the last two columns.

147

148 CNN-based models, like TREDNet and SEI, excel in predicting variant effects.
149 TREDNet achieves balanced performance with correctly predicted upregulation (0.60)
150 and downregulation (0.60) rates, while SEI leads in upregulation detection (0.62) but
151 struggles with downregulation (0.56). Transformer-based models show mixed results.
152 Nucleotide Transformer v2-500m-multi-species matches SEI's upregulation prediction
153 rate (0.62) but underperforms in downregulation (0.52). DNABERT2 performs steadily
154 (0.54, 0.55), while Geneformer shows the weakest outcomes. Borzoi achieves a strong

upregulation prediction rate (0.61) but falters in downregulation (0.47). Enformer shows moderate, balanced results (0.58, 0.55), while HyenaDNA remains consistent but unremarkable.

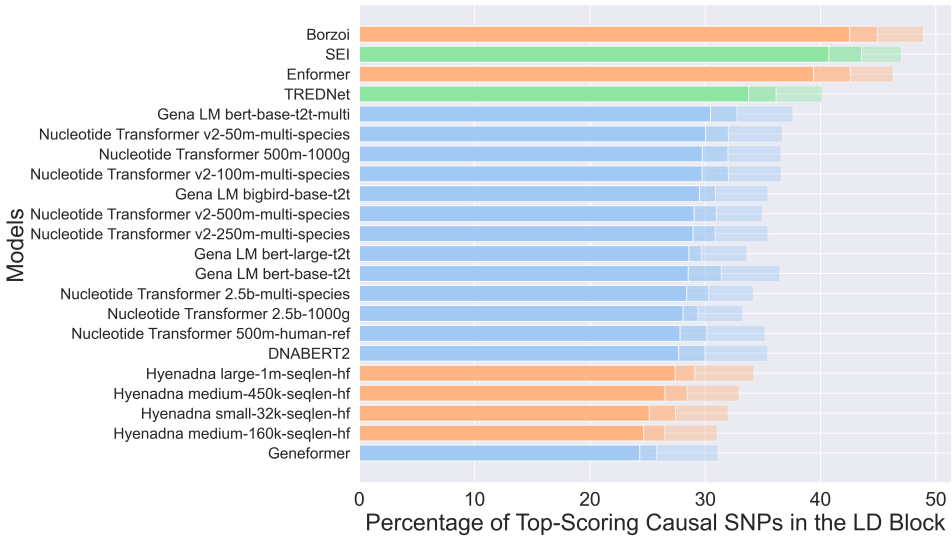

Fig 3: Percentage of top-scoring causal SNPs predicted within the Linkage Disequilibrium (LD) block by various models. Models are grouped by architecture: CNN (green), transformer (blue), and hybrid (orange). Bars are sorted by top-1 performance, with decreasing transparency indicating lower ranks (top-2 and top-3 results are progressively faded compared to top-1).

A crucial aspect of each model is its ability to identify causal SNPs within their respective Linkage Disequilibrium (LD) blocks. Essentially, this is the most important function of variant classification models, which can be directly applied to the resolution of causal variants within large blocks of associated variants identified in GWAS studies. To investigate this, we curated a dataset of 10,098 causal SNPs specific to the HepG2 cell line. For each causal SNP, we identified associated LD-block SNPs with an r^2-value above 0.8 within a 500 kbp window, resulting in 263,286 SNPs. With this set, we can directly assess each method's ability to retrieve the causal SNP within this ~26:1 pool of associated SNPs. For each identified SNP, we generated two sequences of 1

169 kbp: one containing the SNP and the other containing the reference nucleotide. In both
170 cases, the SNP or reference nucleotide were positioned at the center of the sequence.
171 These sequences were then input into several models to assess their ability to predict
172 causal SNPs by calculating the log2 foldchange between the scores of the alternative
173 sequence (containing the SNP) and the reference sequence (Methods and Materials
174 sections). This procedure was repeated for all LD-block associated SNPs and the log2
175 fold-change score of the known causal SNP was compared to its associated variants.
176 Finally, we computed the percentage of causal SNPs within an LD-block correctly
177 predicted as causal based on their score being either the highest ("top-1" test) or within
178 the top two and three highest scores ("top-2" and "top-3" tests).
179 The Borzoi model achieved the highest accuracy, correctly identifying 42.5% of causal
180 SNPs, followed by the SEI, Enformer, and TREDNet models, with 40.7%, 39.4%, and
181 33.8% in the top1 test, respectively, (Fig. 3). The causal variant detection rate increased
182 substantially in the top-2 and top-3 test with Borzoi (44.9%, 48.9), SEI (43.6%, 47.0%),
183 Enformer (42.4%, 46.2%), and TREDNet (36.1%, 40.1%). Among the transformer-
184 based models, the Gena LM bert-based-t2t-multi demonstrated the best performance
185 (top-1 30.4%, top-2 32.7%, and top-3 37.5%), marginally surpassing other Nucleotide
186 Transformer and Gena LM variants, which achieved accuracies between 28.0% and
187 30.0% in top-1 test. The HyenaDNA models, categorized as hybrid architectures,
188 displayed moderate performance, with accuracies ranging from 24.3% to 26.5% in top-
189 1 test, and larger sequence lengths, such as those used in the HyenaDNA large model,
190 led to improved results. The Geneformer model, however, had the lowest accuracy,
191 identifying 24.3% of causal SNPs in top-1 test. Overall, hybrid and CNN architectures

192 outperformed transformer-based models in this application, underscoring their
193 suitability for tasks involving causal SNP detection.
194 Overall, models like TREDNet and SEI consistently outperform others in both
195 regression analysis and classification tasks, showcasing their strength in capturing local
196 genomic patterns critical for regulatory variant effect prediction. Hybrid models, such
197 as Borzoi and Enformer, offer balanced performance, bridging the gap between local
198 pattern recognition and long-range dependency modeling but fail to surpass CNNs in
199 overall accuracy. Yet, these models excel at accurate detection of causal variants within
200 LD-blocks of associated SNPs. Transformer-based models, despite their potential to
201 model complex sequences and long-range interactions, struggle to match the
202 granularity of CNNs and hybrids.
203
204 **Impact of Certainty in Experimental Results**
205 The results of experimental assays of regulatory variants depend on the selected degree
206 of certainty in separating significant and insignificant variant effects. To investigate the
207 impact of experimental data significance on correlation with modeling results, we
208 analyzed the top models for each architecture type: TREDNet (CNN), HyenaDNA
209 (hybrid), and Nucleotide Transformer (transformer). The experimental data was binned
210 using the following significance thresholds, $p < 0.5$, $p < 0.1$, $p < 5 \cdot 10^{-5}$, $p < 10^{-5}$.
211
212 The results demonstrate that models align better with experimental findings when
213 applied to variants with higher statistical significance. For instance, data points with
214 greater experimental confidence show increased clustering in the first and third
215 quadrants, representing cases where both experimental and predicted results are either

positive (upregulation) or negative (downregulation) (Fig. 4, panels A-C, Figures S2, S3a, S3b, S3c, and S3d).

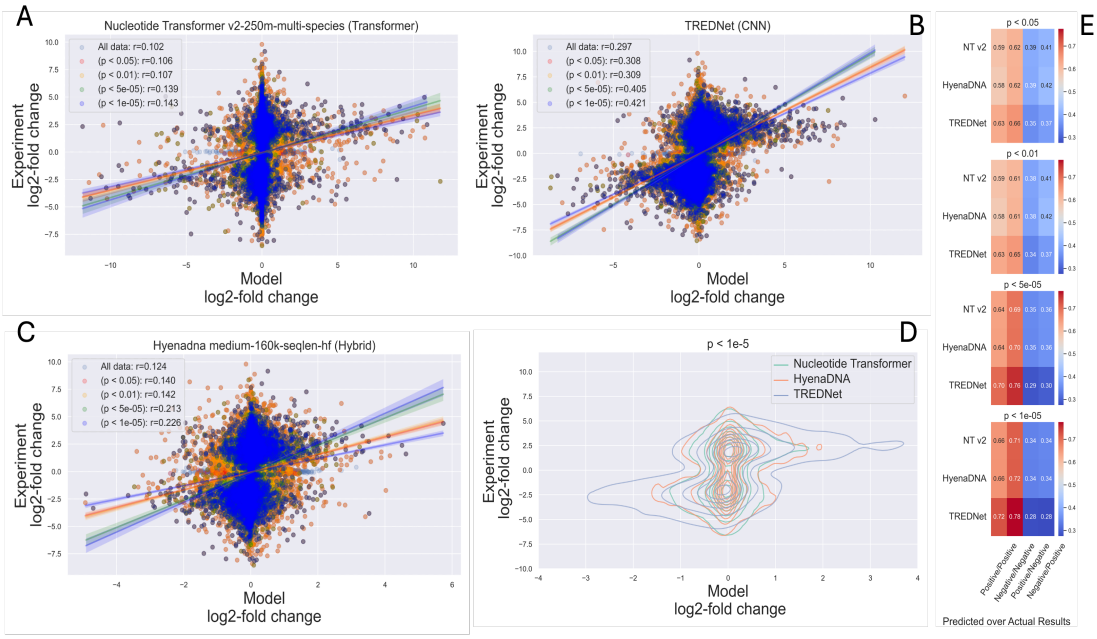

Fig 4: Model performance relative to experimental data significance, using top models from each architecture category. Panels A-C show scatter plots for TREDNet, HyenaDNA, and Nucleotide Transformer, with the correlation between model predictions and experimental results as significance improves. Panel D compares model predictions for highly significant variants. Panel E presents heatmaps of classification metrics at different p-value thresholds, highlighting performance variations in identifying positive/negative outcomes across

Among the evaluated architectures, HyenaDNA exhibited the greatest improvement in predictive accuracy with the increase of the statistical significance of experimental results (from all data 0.124 to 0.226 with $p < 10^{-5}$, Fig. 4, panel C). However, TREDNet consistently displayed superior overall performance, with a higher density of points aligning with experimental outcomes (from all data 0.298 to 0.421 with $p < 10^{-5}$, Fig. 4, panel B). This trend is further illustrated in the density plots, where TREDNet exhibits increased clustering in the first and third quadrants, demonstrating stronger correlations between model predictions and experimental values compared to both hybrid and transformer models (Fig. 4, panel D). Moreover, the correlation coefficients

229 across varying significance thresholds highlight TREDNet's robustness, achieving the
230 highest values under all conditions. True positives and true negatives increase with the
231 significance of the experimental results (positive/positive 0.63, negative/negative 0.66,
232 with $p < 0.05$; positive/positive 0.72, negative/negative 0.78, with $p < 10^{-5}$), with the
233 hybrid model and the transformer model following in performance (Fig. 4, panel E).
234 The rest of models' performance relative to experimental data significance can be found
235 in Fig. S2, and Figures S3a, S3b, S3c, and S3d.
236 Given the strong cell-line specificity of enhancers (Wu and Huang, 2024), determining
237 whether model performance depends on the type of cell line used is essential for
238 understanding their predictive capabilities and limitations. For this, we evaluated
239 whether a model fine-tuned for a specific cell line demonstrates superior performance
240 compared to models from other architectures, such as CNNs, which often rely on
241 generalization rather than cell-line-specific adaptation. Indeed, transformer-based
242 models, pre-trained on extensive datasets and fine-tuned for targeted tasks, can offer
243 both flexibility and accuracy, making them well-suited for capturing the unique
244 regulatory landscape of cell lines. In contrast, CNN-based models, such as TREDNet
245 and SEI, though highly effective for specific tasks, could require additional steps, such
246 as retraining or the addition of specialized layers, to adapt their outputs for cell-line-
247 specific predictions. However, the results indicate that TREDNet and SEI remain the
248 most effective models for predicting variant effects across different cell lines, even after
249 model fine-tuning (Pearson Correlation Table 1, Spearman Correlation Sup Table 1,
250 Dataset-wise Sup Table 2 and Sup Table 3). For example, the Nucleotide Transformer
251 v2-250m-multi-species model aligns with experimental results in K562 and HepG2
252 (0.1998 and 0.11) but performs less effectively in NPC and HeLa cells. This variability

could stem from differences in pre-training datasets—such as multi-species versus human-specific data—and the fine-tuning approach.

Table 1: Pearson correlation coefficient for various deep learning models across four cell lines: K562 (19321 SNPs), HepG2 (16255 SNPs), NPC (14042 SNPs), and HeLa (5241 SNPs). Standard Error is reported in brackets. Bold and underline styles denote the top and second-highest correlations per cell line, respectively.

Models	Cell Line			
	K562 (19321 SNPs)	**HepG2 (16255 SNPs)**	**NPC (14042 SNPs)**	**Hela (5241 SNPs)**
DNABERT2	0.086 (0.039)	0.096 (0.191)	-0.004 (0.007)	0.120 (0.140)
NT v2-50m-ms	0.147 (0.058)	0.103 (0.036)	0.020 (0.003)	0.074 (0.047)
NT v2-100m-ms	0.152 (0.065)	0.127 (0.050)	0.016 (0.003)	0.042 (0.027)
NT v2-250m-ms	0.166 (0.064)	0.110 (0.020)	0.023 (0.003)	0.030 (0.018)
NT v2 500m-ms	0.199 (0.036)	0.114 (0.007)	0.028 (0.003)	0.036 (0.011)
NT 500m1000g	0.123 (0.084)	0.116 (0.023)	0.001 (0.002)	0.022 (0.018)
NT 500m-h-ref	0.149 (0.139)	0.119 (0.038)	0.013 (0.001)	0.067 (0.043)
NT 2.5b-1000g	0.147 (0.052)	0.112 (0.021)	0.005 (0.003)	0.087 (0.105)
NT 2.5b-m-s	0.153 (0.055)	0.139 (0.027)	0.027 (0.001)	0.066 (0.027)
Geneformer	0.005 (0.198)	0.027 (0.160)	-0.002 (0.138)	-0.007 (0.057)
Gena LM-base	0.077 (0.042)	0.050 (0.033)	0.024 (0.002)	0.029 (0.227)
Gena LM large	0.117 (0.044)	0.088 (0.036)	0.030 (0.002)	0.069 (0.216)
Gena LM b-multi	0.077 (0.051)	0.062 (0.024)	0.010 (0.001)	0.026 (0.061)
Gena LM bigbird	0.136 (0.055)	0.099 (0.029)	0.021 (0.002)	0.045 (0.015)
Hyenadna 32k	0.084 (0.071)	0.098 (0.061)	0.002 (0.002)	0.031 (0.013)
Hyenadna 160k	0.148 (0.053)	0.142 (0.040)	-0.007 (0.003)	0.046 (0.018)
Hyenadna 450k	0.077 (0.074)	0.134 (0.050)	-0.007 (0.003)	0.055 (0.015)
Hyenadna 1mf	0.118 (0.056)	0.136 (0.059)	0.001 (0.003)	0.046 (0.013)
TREDNet	**0.315 (0.025)**	**0.314 (0.018)**	**0.075 (0.003)**	0.076 (0.063)
SEI	0.297 (0.022)	0.298 (0.029)	0.073 (0.003)	**0.280 (0.008)**
Enformer	0.059 (0.059)	0.141 (0.110)	0.008 (0.080)	0.245 (0.027)
Borzoi	0.033 (0.016)	0.125 (0.025)	0.023 (0.004)	-0.066 (0.005)

Notably, the hybrid model Enformer, which integrates features from transformers and CNNs, performs particularly well in HeLa cells (0.245), ranking second only to SEI (0.279). As observed in the previous test, CNN-based models demonstrate superior performance, showcasing robustness across cell-line variants. We also assess the impact of the certainty of experimental results, SEI, which does not require training or fine-tuning for specific cell lines, outperforms fine-tuned models. TREDNet, likely benefiting from its second-phase training tailored to the specific cell line, achieves the best overall performance, further highlighting its ability to leverage cell-line-specific information effectively.

Enhancer Detection Models for Variant Effect Assessment

Models designed to predict variant effects rely primarily on the change in confidence when detecting the enhancer region after introducing the variant. Next, we investigated whether improving the accuracy of enhancer sequence detection enhances predictions of variant effects. While SEI, Borzoi, and Enformer do not require fine-tuning, TREDNet, despite being a CNN-based model, incorporates a two-phase training process. The second phase can be considered fine-tuning [7].

Transformer-based architectures, particularly Nucleotide Transformer models pre-trained on multi-species datasets, emerge as strong performers in enhancer classification (Fig. S4). Models such as Nucleotide Transformer v2-250m-multi-species and Nucleotide Transformer 2.5b-multi-species consistently demonstrate high accuracy, as indicated by their auROC and auPRC metrics (auROC: 0.967, 0.963; auPRC: 0.958, 0.951). However, the correlation between the models' outputs and the MPRA experimental results, which measure the fold-change effect between reference and alternative sequences, remains relatively low (ranging from 0.112 to 0.15). SEI

stands out as the top performer in detecting enhancer activity, achieving the highest auROC (0.974) and auPRC (0.971) scores among all models. It is also the second best for the correlation between model outputs and experimental results (0.295). TREDNet demonstrates strong performance with a higher correlation (0.318), while balancing high auROC (0.872) and auPRC (0.814) scores. The HyenaDNA models show competitive classification metrics, with the Hyenadna medium-160k-seqlen-hf and Hyenadna large-1m-seqlen-hf versions attaining notable auROC (0.933, 0.904) and auPRC (0.921, 0.893) scores.

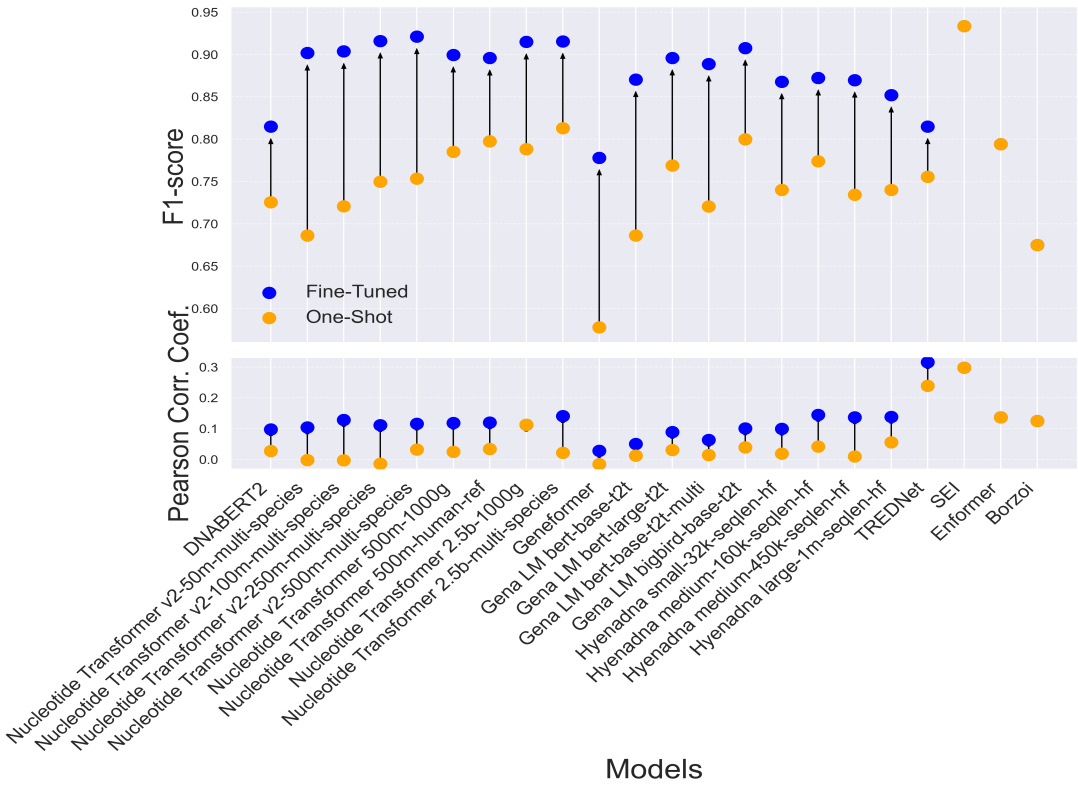

***Fig 5**: Comparison of F1 score (top) and Pearson Correlation Coefficient (bottom) between fine-tuned (blue dots) and one-shot (orange dots) models across various architectures, evaluated on Dataset 3, HepG2 cell line, (14182 SNPs). Arrows indicate the changes in performance between fine-tuned and one-shot implementations for each model.*

However, their Pearson correlation coefficients are moderate (0.152, 0.151), indicating that while these models classify enhancer activity effectively, further optimization is needed to improve their predictive accuracy for variant-specific effects. Borzoi performs the weakest in detecting enhancer regions (auROC: 0.672, auPRC: 0.65), while Geneformer fares poorly in detecting variants (auROC: 0.027). Yet, this weak enhancer detection performance doesn't translate in inaccurate detection of causal variants, as Borzoi outperforms all other methods in this critical test.

Fine-tuned models consistently outperform their one-shot counterparts, achieving F1 scores between 0.85 and 0.92, compared to 0.70 to 0.80 for one-shot models (Fig. 5, top). This improvement in performance is reflected in the correlation between model predictions and experimental results, particularly in detecting enhancer variants (Fig. 5, bottom). The performance enhancement is consistent across most architectures, with CNN-based models showing remarkable results even without fine-tuning. However, SEI, which is available only in its 'one-shot' implementation, achieves a high F1 score and correlation, surpassing many fine-tuned transformer models. Additionally, some transformer-based models, such as Nucleotide Transformer 2.5b-1000g, show no improvement in detecting enhancer variants.

Discussion

Our evaluation of deep learning models highlights the strengths and limitations of different architectures in predicting the effects of genetic variants on enhancer activity. CNN-based models, such as TREDNet and SEI, outperform transformer-based and hybrid architectures in both regression and classification tasks. Notably, Borzoi, a

316 hybrid model, excels in causal SNP prediction, leveraging both local and global
317 genomic features critical for regulatory variant effect prediction.
318
319 **CNNs-, Transformers-, and Hybrid-based Models**
320 TREDNet and SEI achieve superior performance in regression (Fig. 1) and
321 classification tasks (Fig. 2). TREDNet demonstrates balanced predictive capability for
322 both upregulated and downregulated variants, while SEI excels in detecting upregulated
323 variants. CNNs effectively capture local genomic patterns like motifs and epigenetic
324 markers, making them well-suited for enhancer activity analysis. Their simpler
325 architecture yields high Pearson correlation coefficients (0.297 for TREDNet, 0.276 for
326 SEI). However, their inability to model long-range dependencies limits performance in
327 tasks requiring broader sequence context.
328 Transformers, including the Nucleotide Transformer and DNABERT 2, capture long-
329 range dependencies but struggle with fine-grained regulatory signals. They
330 underperform in enhancer variant effect prediction, with accuracies between 28.0% and
331 30.0% in causal SNP identification (Fig. 3). These models require substantial training
332 data, and their performance could improve with expanded datasets or refined focus on
333 regulatory signals.
334 Hybrid models, such as HyenaDNA and Borzoi, integrate CNNs, transformers, and
335 LSTMs to balance local and long-range dependencies. While HyenaDNA shows
336 moderate correlation with experimental results, Borzoi outperforms other models in
337 causal SNP detection (Fig. 3), making it valuable for GWAS analysis. However, hybrid
338 models do not consistently surpass CNNs in overall predictive accuracy and should be
339 assessed based on task-specific requirements.

340

341 **Impact of Experimental Data Quality and Model Generalization**

342 The reliability of model predictions is closely tied to the quality of experimental data.
343 Model alignment with experimental results improves as statistical confidence in variant
344 effects increases (Fig. 4). TREDNet, for instance, maintains high predictive accuracy
345 even under varying significance thresholds, emphasizing the robustness of CNN-based
346 models. These findings suggest that enhancing experimental rigor and statistical
347 refinement can reduce uncertainty in variant effect assessments, leading to more
348 reliable predictions.

349 In addition to data quality, model generalization across different cell lines is a critical
350 factor. Transformer models fine-tuned for specific cell lines can capture regulatory
351 landscapes but often underperform compared to CNNs. For instance, the Nucleotide
352 Transformer v2-250m-multi-species achieves reasonable results in K562 and HepG2
353 but struggles in NPC and HeLa, likely due to pre-training limitations. In contrast, CNN
354 models generalize well across cell lines without requiring fine-tuning, further
355 demonstrating their robustness in variant effect prediction.

356 Enhancer detection models, such as the Nucleotide Transformer v2-250m-multi-
357 species, achieve high classification accuracy (auROC: 0.967, auPRC: 0.958), yet their
358 predictions show weak correlation with experimental results (Fig. S4). CNN-based
359 models, particularly SEI, perform well in both enhancer detection and variant effect
360 prediction. While fine-tuning benefits transformer models, SEI, despite its 'one-shot'
361 implementation, still outperforms many fine-tuned transformers (Fig. 5). These results
362 highlight the advantages of CNN architectures in both enhancer activity analysis and

363 cross-cell-line generalization, reinforcing their suitability for regulatory variant effect
364 prediction.
365
366 In conclusion, our findings suggest that SEI and TREDNet are the most effective
367 models for predicting the effects of genetic variants on enhancer activity. Hybrid
368 models like Borzoi demonstrate strong potential in extracting causal SNPs from GWAS
369 data, showcasing their value in applications that require the integration of both local
370 and global sequence contexts. Transformer-based models, on the other hand, are
371 particularly effective at identifying enhancer regions due to their ability to capture long-
372 range dependencies, although they still face challenges with fine-grained genomic
373 features and often require substantial data for optimal performance. These results
374 highlight the strengths of different model architectures for various aspects of regulatory
375 variant prediction and underscore the need for further optimization to enhance their
376 performance. Future work should aim to improve feature detection in transformer
377 models and refine the flexibility of hybrid architectures to better capture both local and
378 long-range genomic dependencies.
379

380 ## Materials and Methods

381 **Data Pre-Processing**

382 The datasets from HepG2, K562, NPC, and HeLa cell lines underwent a standardized
383 preprocessing pipeline to ensure consistency and feature extraction for downstream
384 analysis. Two main pipelines were applied: one for training and testing enhancer
385 classifier models, such as Nucleotide Transformers and DNABERT, and another for

mutagenesis experiments. For the enhancer classification pipeline, 1kb sequences were extracted for both enhancer (positive) and control regions.

To maintain consistency across models, all nucleotides were capitalized, as some models are case-sensitive. Unknown nucleotides marked as "N" were removed from the sequences. The datasets were then split into training, validation, and test sets, and tokenized using the corresponding model's tokenizer with a maximum length of 512 tokens. Labels were generated to distinguish between positive and control sequences.

In the mutagenesis experiment pipeline, the reference and alternative nucleotide locations were extracted for each row of the dataset. For each nucleotide of interest, a 1kb sequence was generated with the target nucleotide positioned in the center. Chromosome and position information, along with reference and alternative alleles, were also extracted. Log2 activity metrics, such as fold-change values or effect sizes, were calculated to assess regulatory activity. Additionally, signed p-values were derived from raw or adjusted p-values to capture both the statistical significance and directionality of the regulatory effects.

The preprocessing step ensured that variations in dataset structure did not interfere with downstream analysis. To further enhance consistency, hg19 genomic coordinates were used for variant mapping, and enhancer activity metrics were uniformly applied across datasets. Differences in expression or activity levels between experimental conditions, such as mutant versus control or modern versus archaic sequences, were computed when applicable. Ultimately, all datasets were converted into a standardized format that included chromosome information, positions, reference and alternative alleles, log2 activity ratios, signed p-values, raw p-values, and other relevant regulatory features.

410

411

412 **Model Adaptation and Fine-tuning**

413 The objective of fine-tuning is to adapt a model originally designed as an enhancer
414 classifier to assess the impact of SNPs on enhancer activity. By introducing variant
415 sequences, we evaluate how nucleotide changes affect the model's score. To achieve
416 this, we fine-tuned pre-trained DNA language models, such as Geneformer (initially
417 built for scRNA-seq data), as enhancer classifiers. Sequences were preprocessed as
418 described earlier, with a length of 1 kbp, balancing genomic context and model
419 compatibility. This choice reflects the average enhancer length of 400 bp.

420 Most models were loaded from the Hugging Face Hub using AutoTokenizer and
421 AutoModelForSequenceClassification. Exceptions, such as TREDNet, Enformer, SEI,
422 and Borzoi, were sourced from GitHub and required custom pipelines for preprocessing
423 and inference. Fine-tuning leveraged the Hugging Face transformers library for flexible
424 training and evaluation. Models like HyenaDNA, Geneformer, Gena-LM,
425 DNABERT2, and Nucleotide Transformers utilized the Trainer API. We fine-tuned the
426 models on four biosample-specific datasets: HeLa (BioS2), neural progenitor cells
427 (NPC, BioS45), HepG2 (BioS73), and K562 (BioS74). Sequences were cleaned to
428 remove ambiguous bases (Ns) and split into training, validation, and test sets. Model-
429 specific tokenizers processed sequences with a maximum length of 512 tokens. For
430 non-fine-tuned models like SEI and Enformer, the same 1 kbp sequence length was
431 used with padding as needed.

432 Key hyperparameters included a batch size of 8, a learning rate of 1e-5, and a maximum
433 of 200 epochs. Early stopping with a patience of 5 epochs was applied to prevent

overfitting, with progress logged every 500 steps. Fine-tuned models were saved in output directories named by model checkpoint and biosample ID and uploaded to the Hugging Face Hub for reproducibility. GPU memory was cleared after each session for resource efficiency. This workflow allowed us to adapt general-purpose DNA language models for cell-type-specific regulatory prediction while maintaining consistency with non-fine-tuned models. Training on NIH BioWulf node clusters with A100 GPUs enabled efficient handling of large datasets, such as the largest set (~130,000 sequences for positive and control enhancer regions), which completed in hours using 8 A100 GPUs.

Validation and Post-Processing

During the fine-tuning process, the models were optimized to distinguish enhancer sequences from control sequences. In contrast, the validation process focused on evaluating the difference between the reference enhancer sequence and the alternative sequence, where the middle nucleotide was mutated. The difference in output scores between these two sequences, obtained from the fine-tuned model, allowed for the estimation of the variant's effect on the enhancer. If the output scores for both sequences were identical, the variant had no effect. Conversely, a change in output scores indicated either up- or down-regulation.

The prediction process involved generating regulatory effect predictions for each dataset by calculating the log2 ratio of alternative to reference allele probabilities for each variant. This metric provided a standardized approach to assess the relative regulatory effects of alternative alleles compared to reference alleles. Since each fine-tuned model produced logits for each class (enhancer or not), predictions for both

458 reference and alternative alleles were processed by first calculating margin logits,
459 which represent the difference in predicted class probabilities. These logits were then
460 transformed into probabilities using the sigmoid function. The log2 ratio of these
461 probabilities was computed to quantify the regulatory differences between the two
462 alleles. Additionally, precomputed predictions stored in pickle files were loaded for
463 specific models and datasets. The predictions for reference and alternative alleles were
464 used to compute a ratio, representing the alternative-to-reference prediction ratio. This
465 ratio was subsequently log2-transformed to ensure consistency in the output format.
466 Additional preprocessing steps were applied to specific datasets as needed, such as
467 loading specialized prediction files.
468 The prediction pipeline was applied across all datasets by iterating through experiments
469 defined in a metadata table. For each experiment, data were preprocessed, and
470 predictions were either computed or loaded accordingly. The results were compiled into
471 a unified dictionary containing log2 prediction ratios for all variants across
472 experiments. As in the fine-tuning process, the validation of variant/reference
473 sequences was performed using 1 kb sequences, ensuring consistency in genomic
474 context and input length across models. Several metrics were employed to evaluate
475 model performance, including the Area Under the Curve (AUC), Precision-Recall
476 Curve (PRC), F1-score, and correlation coefficients such as Pearson and Spearman.
477 These metrics provided a comprehensive assessment of model performance,
478 encompassing both classification accuracy and the correlation between predicted and
479 observed values.
480
481 **Deep Learning Models**

482 This study utilizes several deep learning models designed for genomic sequence
483 analysis, with a particular focus on applications to the Homo sapiens genome. The
484 models represent the forefront of innovation in deep learning for genomics, leveraging
485 diverse architectures and pre-training paradigms to handle the complexity of human
486 genomic data. To maintain focus on models explicitly designed for the human genome,
487 this study excluded models like EVO and GPN, which are primarily aimed at other
488 species or biological domains.

Model	Architecture	Training Details	Applications
DNABERT 2 [9]	Transformer-based with ALiBi and Flash Attention	Pre-trained using Byte Pair Encoding (BPE) tokenization; compact representation of sequences; optimized for computational efficiency and scalability.	Sequence classification, motif discovery, variant effect prediction across species.
Nucleotide Transformer [10, 11]	Transformer-based (50M–2.5B parameters)	Pretrained on multi-species genomes with 6-mer tokenization; trained on 300B tokens using Adam optimizer with warmup and decay schedules; supports long-range genomic dependencies.	Comparative genomics, functional annotation, variant effect prediction.
Geneformer [13]	Transformer-based with rank-based analysis	Pretrained on single-cell transcriptome datasets using masked gene prediction; optimized for noise resilience and stability in single-cell data.	Gene network dynamics analysis, single-cell state classification.
GENA-LM [17]	BERT-base, BERT-large, BigBird variants	Pretrained on genomic sequences using sparse attention mechanisms (BigBird); supports long-range sequence modeling and local feature detection; memory augmentation techniques applied for scalability.	Local feature detection, long-range genomic dependency modeling.
Enformer [12]	Convolution + Transformer hybrid	Pretrained on DNA sequences up to 200,000bp using convolutional blocks for spatial reduction and transformer layers for long-range interactions; predicts over 5,000 epigenetic features.	Gene expression prediction, regulatory element analysis.
HyenaDNA [18]	Runtime-scalable models	Pretrained on human reference genome with extended sequence lengths (32,000–1,000,000bp); optimized for runtime scalability using AWS HealthOmics and SageMaker infrastructure.	Long-range genomic interaction analysis.
Borzoi [19]	Multi-layer regulatory model	Trained to predict RNA-seq coverage directly from DNA sequences; integrates multiple layers of regulatory predictions; fine-tuned for tissue-specific applications.	Cell- and tissue-specific RNA-seq coverage prediction, cis-regulatory pattern detection.
SEI [6]	Residual dilated convolutional layers	Trained on one-hot encoded sequences (4,096bp) using residual paths for multi-scale pattern detection; incorporates global integration layers for long-range dependencies; balances computational efficiency with expressiveness.	Variant effect prediction on cis-regulatory activity across chromatin profiles.

TREDNet [7]	Two-phase CNN architecture	Trained on ENCODE and NIH Roadmap Epigenomics datasets; uses CNNs to predict epigenomic signals (e.g., histone modifications) and refine enhancer predictions; supports in silico saturated mutagenesis for variant prioritization.	Enhancer prediction, variant prioritization.

Training Data

We constructed a dataset of positive and control enhancer sequences from four human cell lines: HepG2, Hela, K562, and Neural Progenitor Cells (NPC). These cell lines were selected for their relevance in genomic research and the availability of comprehensive epigenomic data, such as DNase-Seq and histone modification profiles, sourced from the ENCODE project using genome annotations for hg19 and hg38.

Positive enhancer sequences were defined through a two-step process. First, open chromatin regions were identified using DNase-Seq data. These regions were then intersected with H3K27ac histone modification signals, a hallmark of active enhancers. To refine the dataset, we excluded non-enhancer regions such as exonic regions (coding sequences), promoter regions (near transcription start sites), and low-confidence regions listed in the ENCODE Blacklist (e.g., repetitive sequences).

Specific blacklisted peaks from UCSC browser tables (wgEncodeDacMapabilityConsensusExcludable) were also removed.

Control sequences were selected using a similar filtering process but specifically excluded regions overlapping both DNase-Seq peaks and H3K27ac signals to ensure they represented non-active genomic regions. Both positive and control sequences underwent identical exclusion criteria to eliminate confounding genomic features such as exons and promoters.

509 All sequences were standardized to 1 kb in length to balance sufficient coverage of
510 regulatory regions with computational efficiency. This size is widely used in enhancer
511 studies as it captures relevant regulatory elements while avoiding extraneous genomic
512 features. The dataset includes an equal number of positive and control sequences for
513 each cell line: HepG2 (14,062 each), NPC (12,466 each), K562 (19,968 each), and Hela
514 (31,247 each). Metadata from the ENCODE project was used to retrieve biosample IDs
515 and associated DNase-Seq and histone modification data. Reference files provided
516 coordinates for exons, blacklist regions, and promoters, ensuring only valid enhancer
517 or control regions were included in the final dataset.
518
519 **Validation Data**
520 Datasets 1 and 2 were sourced from the study by [20], which utilized the Survey of
521 Regulatory Elements (SuRE) reporter assay to identify reporter assay quantitative trait
522 loci (raQTLs)—SNPs influencing regulatory element activity. For K562 cells, 19,237
523 raQTLs were identified with an average allelic fold change of 4.0-fold, while HepG2
524 cells yielded 14,183 raQTLs with a higher average fold change of 7.8-fold. Most
525 raQTLs were cell-type-specific, showing limited overlap between the two cell lines.
526 Dataset 3 was derived from [21] and focuses on a ~600 bp enhancer region at the
527 SORT1 locus (1p13). This region includes SNP rs12740374, which creates a C/EBP
528 binding site that influences SORT1 expression. The enhancer was tested in HepG2 cells
529 using a luciferase reporter assay. Data for flipped sequences were excluded due to the
530 orientation-independent nature of enhancers, and deletion datasets were omitted due to
531 computational model constraints, which focus on single-nucleotide variants.

532 Dataset 4, from [22] used MPRA to investigate 2,756 variants linked to red blood cell
533 traits in K562 cells. The study identified 84 highly significant SNPs with functional
534 effects, enriched in open chromatin regions shared between K562 cells and primary
535 human erythroid progenitors (HEPs). This dataset provides insights into erythroid-
536 specific regulatory activity.
537 Dataset 5, generated by [23], includes MPRA results for 284 SNPs tested in HepG2
538 cells. These SNPs, located within regions of high linkage disequilibrium identified
539 through eQTL analyses, offer a detailed map of allele-specific regulatory effects
540 relevant to liver-specific gene regulation and genotype-phenotype associations.
541 Dataset 6, from [24], includes lentiMPRA results for 14,042 single-nucleotide variants
542 fixed or nearly fixed in modern humans but absent in archaic humans (e.g.,
543 Neanderthals). Tested across three cell types—embryonic stem cells, neural progenitor
544 cells, and fetal osteoblasts—the study identified 1,791 sequences with regulatory
545 activity, including 407 with differential expression between modern and archaic
546 humans. This dataset focuses on NPC data.
547 Datasets 7, 8, and 9, from [25], include MPRA results for three coding exons—SORL1
548 exon 17, TRAF3IP2 exon 2, and PPARG exon 6—tested in HeLa cells. Respectively,
549 these datasets cover 1,962, 1,614, and 1,665 SNPs. The study explored enhancer
550 activity within these coding exons (eExons), finding that ~12% of mutations altered
551 enhancer activity by at least 1.2-fold, with ~1% causing changes of less than 2-fold.
552 This version eliminates repetition while maintaining all critical details about the
553 datasets and their relevance.
554

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Funding

This research was supported by the Division of Intramural Research of the National Library of Medicine, National Institutes of Health; 1-ZIA-LM200881-12 (to I.O.).

References

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nat Rev Methods Primers. 2021 Dec;1(1):59.

2. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.

3. Knight JC. Approaches for establishing the function of regulatory genetic variants involved in disease. Genome Med. 2014 Oct 31;6(10):92.

4. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015 Apr;16(4):197–212.

5. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015 Oct;12(10):931–4.

6. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022 Jul 11;54(7):940–9.

7. Hudaiberdiev S, Taylor DL, Song W, Narisu N, Bhuiyan RM, Taylor HJ, et al. Modeling islet enhancers using deep learning identifies candidate causal variants at loci associated with T2D and glycemic traits. Proc Natl Acad Sci USA. 2023 Aug 29;120(35):e2206612120.

8. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021 Aug 9;37(15):2112–20.

9. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv. 2023;

10. Dalla-Torre H, Gonzalez L, Mendoza Revilla J, Lopez Carranza N, Henryk Grywaczewski A, Oteri F, et al. The nucleotide transformer: building and evaluating robust foundation models for human genomics. BioRxiv. 2023 Jan 15;

11. de Almeida BP, Dalla-Torre H, Richard G, Blum C, Hexemer L, Gelard M, et al. SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models. BioRxiv. 2024 Mar 15;

12. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021 Oct 4;18(10):1196–203.

13. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023 Jun;618(7965):616–24.

14. Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. To Transformers and Beyond: Large Language Models for the Genome. arXiv. 2023;

15. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics. 2022 Jul 25;16(1):26.

16. Yue T, Wang Y, Zhang L, Gu C, Xue H, Wang W, et al. Deep learning for genomics: from early neural nets to modern large language models. Int J Mol Sci. 2023 Nov 1;24(21).

17. Fishman V, Kuratov Y, Petrov M, Shmelev A, Shepelin D, Chekanov N, et al. GENA-LM: A Family of Open-Source Foundational Models for Long DNA Sequences. BioRxiv. 2023 Jun 13;

18. Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. arXiv. 2023 Nov 14;

19. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. BioRxiv. 2023 Sep 1;

20. van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, et al. High-throughput identification of human SNPs affecting regulatory element activity. Nat Genet. 2019 Jul;51(7):1160–9.

21. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019 Aug 8;10(1):3583.

620 22. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic
621 functional dissection of common genetic variation affecting red blood cell traits. Cell.
622 2016 Jun 2;165(6):1530–45.

623 23. Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, et al.
624 Massively parallel quantification of the regulatory effects of noncoding genetic
625 variation in a human cohort. Genome Res. 2015 Aug;25(8):1206–14.

626 24. Weiss CV, Harshman L, Inoue F, Fraser HB, Petrov DA, Ahituv N, et al. The cis-
627 regulatory effects of modern human-specific variants. eLife. 2021 Apr 22;10.

628 25. Birnbaum RY, Patwardhan RP, Kim MJ, Findlay GM, Martin B, Zhao J, et al.
629 Systematic dissection of coding exons at single nucleotide resolution supports an
630 additional role in cell-specific transcriptional regulation. PLoS Genet. 2014 Oct
631 23;10(10):e1004592.