AlphaGenome Enhances Personal Gene Expression Prediction but Retains Key Limitations

Li Shen[1,2*]

[1]Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, USA.
[2]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, USA.

Corresponding author(s). E-mail(s): li.shen@mssm.edu;

Abstract

In recent years, numerous genome AI models have been developed to elucidate the relationship between DNA sequence and gene expression. However, these models have faced criticism for their limited accuracy in predicting individual-specific gene expression. AlphaGenome, the current state-of-the-art in genome AI, achieves exceptional performance across a range of sequence-based predictive tasks, but its utility for personal expression prediction has not yet been assessed. In this study, we evaluate AlphaGenome's ability to predict personal gene expression and find that it significantly outperforms its predecessor. Using GTEx data, AlphaGenome improves the prediction of expression direction over Enformer, achieving an odds ratio of 3.0. In some cases, it even reverses previously observed negative correlations into positive ones. Moreover, AlphaGenome demonstrates improved performance for genes with known nonlinear sequence-expression relationships, though it uncovers mechanisms distinct from those identified by tree-based models.

Keywords: gene expression, deep learning, genome AI model, personalized prediction

Introduction

Understanding the relationship between genome sequences and molecular phenotypes–such as chromatin modifications, DNA accessibility, and gene expression–has long been a central focus in genomics. Among these phenotypes, gene expression is arguably the

most extensively studied due to its strong associations with health and disease [1]. Over the past few decades, investigations into the functional link between sequence and expression have largely fallen under the scope of expression quantitative trait locus (eQTL) analysis [2]. Traditional machine learning approaches, including penalized linear regression [3–5] and random forests [6], have been commonly employed in this context.

More recently, deep learning-based genome AI models have emerged as powerful alternatives to traditional methods [7–13], demonstrating remarkable performance in predicting a wide range of molecular phenotypes across previously unseen genomic regions. Among these, Enformer [11] represents a landmark advancement, combining convolutional and transformer layers to capture long-range interactions and outperform prior approaches. At the time of its release, Enformer featured significantly more parameters than its predecessors and required substantially greater computational resources for training. Building on this foundation, AlphaGenome [13] is currently the state-of-the-art (SOTA) genome AI model. It employs a 1 Mb context window and delivers predictions at single-basepair resolution, achieving substantial improvements over existing methods across multiple predictive tasks.

The impressive performance of genome AI models has sparked growing interest in their potential to predict molecular phenotypes–particularly gene expression–at the individual level. Such capabilities hold significant promise for applications in drug target discovery and precision medicine. However, early investigations have yielded disappointing results: many of Enformer's predictions were negatively correlated with observed expression levels [14, 15]. In response, researchers have attempted to fine-tune Enformer and similar models using individual-specific data, yielding modest but encouraging improvements [16, 17].

This study aims to evaluate the potential of AlphaGenome for predicting personal gene expression. Although AlphaGenome was not explicitly trained on individual-level data, its enhanced network architecture and expanded training set make it reasonable to hypothesize that it may outperform its predecessors. To assess its capabilities and limitations, we compare AlphaGenome against its predecessor, Enformer, as well as two classic machine learning methods: Elastic Net and Random Forest. A key advantage of deep learning-based models over traditional approaches lies in their ability to capture complex, nonlinear relationships within data. To investigate this strength, we focus on a subset of genes for which the sequence-expression relationships are likely nonlinear. Additionally, a case study is conducted to examine the marginal effects of genetic variants on AlphaGenome's predictions.

Results

AlphaGenome Significantly Outperforms Its Predecessor in Predicting Personal Gene Expression

The GTEx database [18], which includes RNA-seq data from 953 individuals across 50 tissues, was used as a benchmark to evaluate each model's ability to predict personal gene expression. The normalized expression matrix, downloaded from the GTEx website, contains 42,372 genes. After filtering for sparsity, 377,857 gene-tissue pairs

remained for model training and evaluation. Both Elastic Net and Random Forest were trained and evaluated separately for each gene using nested 10-fold cross-validation. Based on the results from Elastic Net, 300 genes were randomly selected to span a range of coefficients of determination (R2) across tissues–from negative values to greater than 0.5. AlphaGenome was originally trained to predict population-averaged expression levels for all GTEx tissues; thus, its predictions for individual sequences on the selected 300 genes were used directly. Since Enformer does not natively output GTEx tissue predictions, we followed AlphaGenome's approach: Enformer's frozen embeddings were used to train predictors for population-averaged expression across all GTEx tissues, and these predictors were then applied to individual sequences. For fair comparison across models, Pearson correlation was used as the evaluation metric, as it does not require scale-matching between predicted and observed values. For details of data preprocessing, model training and evaluation, see Section S1.

Figure 1a presents the distribution of Pearson correlation coefficients across all gene-tissue pairs for the four methods compared. As expected, both AlphaGenome and Enformer exhibit lower correlations than Elastic Net and Random Forest, as neither deep learning model was trained on individual-level data. Notably, AlphaGenome outperforms Enformer on average, with a median correlation higher by 0.07 and an even greater difference at the upper quartile. Among all gene-tissue pairs, AlphaGenome yields 2,459 positive and 971 negative correlations, while Enformer yields 1,557 positive and 1,873 negative correlations. This is particularly striking given that AlphaGenome was not trained on personal expression data.

To further assess their relative performance, head-to-head comparisons were conducted between AlphaGenome and each of the other three methods across all gene-tissue pairs. AlphaGenome significantly outperforms Elastic Net in 218 gene-tissue pairs (Figure 1b) and Random Forest in 63 gene-tissue pairs (Figure 1c). When compared to Enformer, AlphaGenome shows a clear advantage: 1,374 gene-tissue pairs have significantly higher correlations than Enformer, while only 430 favor Enformer, resulting in a winning ratio of 3.2.

To gain a more intuitive understanding of AlphaGenome's improvement over Enformer, the top 15 genes from each model–ranked by the difference in Pearson correlation in lung tissue–are shown in Figure 1e. For AlphaGenome, the top-performing genes exhibit dramatic reversals in correlation direction; for example, CUTALP improves from a correlation of -0.81 to $+0.82$. In contrast, the improvements among Enformer's top genes are comparatively modest. These results highlight AlphaGenome's substantial advantage in predicting personal gene expression, despite not being explicitly trained for this task.

AlphaGenome Captures Nonlinear Dependencies with Mechanisms Distinct from Tree-Based Methods

Elastic Net is a linear regression model that predicts gene expression as a weighted sum of genetic variants. In contrast, Random Forest is an ensemble decision tree-based method capable of modeling nonlinear relationships, including higher-order interactions among variants. Notably, Random Forest performs more comparably

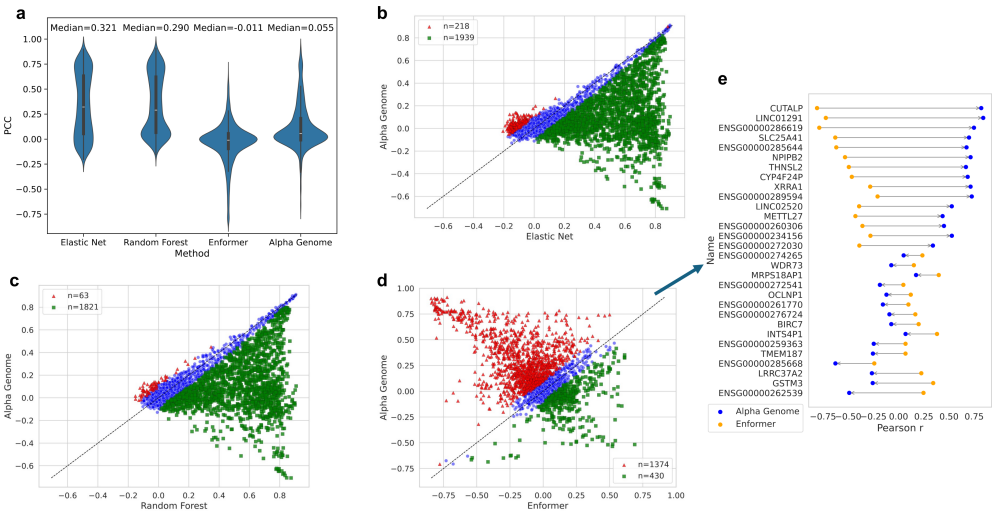

Fig. 1 Accuracy of personal gene expression prediction for four comparing methods. **a** Violin plots of Pearson correlations of the four methods across all gene-tissue pairs. **b-d** Scatter plots of AlphaGenome vs. Elastic Net, Random Forest and Enformer. Red triangles represent gene-tissue pairs that AlphaGenome is significantly better, while green squares represent gene-tissue pairs that other methods are significantly better. The statistical significance is determined from 1000 bootstrap samples. **e** Dumbbell plots of the top 15 genes from AlphaGenome and Enformer based on the difference of Pearson correlation.

to AlphaGenome than Elastic Net does, exhibiting substantially fewer genetissue pair losses relative to AlphaGenome (63 vs. 218). Because Random Forest can capture nonlinear genotypeexpression relationships, its improved performance reduces AlphaGenomes relative advantage. This pattern supports the hypothesis that AlphaGenome is more likely to outperform Elastic Net when the underlying sequenceexpression relationship is nonlinear.

Based on this rationale, we used Random Forest as a proxy for nonlinearity to identify genetissue pairs in which nonlinear modeling provides a measurable benefit. Specifically, we identified 6,295 genetissue pairs in which Random Forest significantly outperformed Elastic Net (Figure 2a), representing approximately 1.7% of all evaluated pairs. While it is possible that other gene-tissue pairs exhibit nonlinear relationships that Random Forest fails to capture–but that AlphaGenome could–this is considered unlikely, and running AlphaGenome across all 377,857 gene-tissue pairs is computationally impractical. From the filtered set, the top 300 genes (based on differences in correlation with Elastic Net) were selected to compare AlphaGenome with Elastic Net and Random Forest. AlphaGenome outperformed Elastic Net in 99 genetissue pairs but did not outperform Random Forest in any (Figure 2b). This outcome may be partially biased, as the selected gene-tissue pairs were pre-filtered based on Random Forest's performance.

To gain deeper insight into how AlphaGenome outperforms Elastic Net on genes with nonlinear relationships, the top-ranking gene, ABI3, was selected as a case study. Model predictions in cultured fibroblast cells were compared to observed expression values across 953 individuals, as shown in Figure 2c. Both AlphaGenome and Random

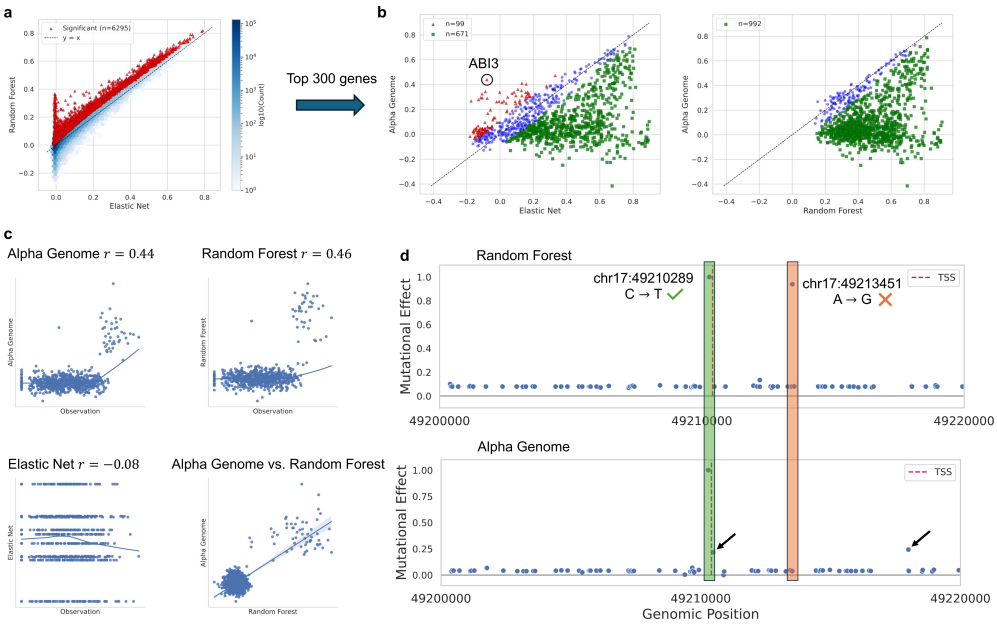

Fig. 2 AlphaGenome vs. Random Forest and Elastic Net in nonlinearity-filtered gene-tissue pairs. **a** Scatter plot of Random Forest vs. Elastic Net to identify nonlinear gene-tissue pairs. **b** Scatter plots of AlphaGenome vs. Elastic Net and Random Forest for top 300 nonlinear genes. **c** Scatter plots of the comparing model's predictions vs. observation. **d** Marginal effects of genetic mutations for Random Forest and AlphaGenome in a 20 Kb region around the transcriptional start site of ABI3.

Forest achieved comparable correlations with the observed data (0.44-0.46), whereas Elastic Net produced a slightly negative correlation. Despite the similarity in overall correlation, AlphaGenome and Random Forest made distinct predictions: while both successfully classified individuals into high- and low-expression groups, their predictions within each group showed minimal correlation, suggesting that the two models capture different underlying patterns.

In silico mutagenesis was performed on all identified variants within a 1 Mb window centered on the transcriptional start site (TSS) of ABI3 to assess their marginal effects on predictions from Random Forest and AlphaGenome (details are provided in Section S1). A 20 Kb region surrounding the TSS–where the strongest effects were observed–is shown in Figure 2d. Both models identified a C $\rightarrow$ T mutation at chr17:49,210,289 as having the most significant effect. However, the two models diverged in their assessment of other variants. Random Forest identified an A $\rightarrow$ G mutation at chr17:49,213,451 as important, while AlphaGenome showed minimal response to this variant. Conversely, AlphaGenome highlighted two other mutations with moderate effects (normalized impact ≈ 0.2) that were not detected by Random Forest. These differences suggest that Random Forest and AlphaGenome capture distinct nonlinear relationships between sequence variation and gene expression for the same gene.

Discussion

Is bigger always better? For researchers in the field of large language models (LLMs), the answer is often a resounding yes, as the scaling law has consistently held true [19]. We are currently witnessing the rise of extreme-scale LLMs–such as GPT-4 [20] and Grok 4 [21]–that continue to grow in size and capability. A similar trend may be unfolding in genomics, as the declining cost of sequencing paves the way for personal genomes to be generated at unprecedented scale [22].

AlphaGenome, with its 1 Mb context window and single-basepair resolution, represents the current state-of-the-art in genome AI. It has demonstrated exceptional accuracy in predicting molecular phenotypes across multiple modalities using the reference genome. As shown in this study, AlphaGenome also improves the prediction of individual-level gene expression–despite never being trained on personal data. This performance gain likely stems from advances in both model architecture and training data: single-basepair resolution increases sensitivity to SNVs and indels, the most common types of genetic variation; the expanded context window enables the model to integrate long-range regulatory interactions; and multimodal training offers a more holistic understanding of the transcriptional regulatory code.

Despite its advancements, AlphaGenome's performance still lags behind that of classic machine learning models trained directly on personal-level data. Currently, training AlphaGenome on such data is infeasible, as DeepMind has only released APIs for model inference and explicitly prohibits fine-tuning AlphaGenome outputs for downstream tasks [23]. However, it is anticipated that future models of this kind will be developed, enabling deeper insights into the relationship between genome sequences and molecular phenotypes–not only across genomic regions, but also at the individual level.

Another limitation of the present study is that model evaluation was restricted to 300 randomly selected genes, which may introduce sampling bias. This design choice was primarily driven by practical constraints: the AlphaGenome server imposes limits on the number of functional calls permitted from a single IP address. Evaluating 300 genes required nearly one week of continuous processing, rendering genome-wide evaluation impractical within a reasonable timeframe.

References

[1] Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**, 184–194 (2009).

[2] Mackay, T. F. C., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects **10**, 565–577. URL https://www.nature.com/articles/nrg2612. Number: 8 Publisher: Nature Publishing Group.

[3] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data **47**, 1091–1098. URL https://www.nature.com/articles/ng.3367. Publisher: Nature Publishing Group.

[4] Zhang, Z., Bae, Y. E., Bradley, J. R., Wu, L. & Wu, C. SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification **13**, 6336. URL https://www.nature.com/articles/s41467-022-34016-y. Publisher: Nature Publishing Group.

[5] Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies **48**, 245–252. URL https://www.nature.com/articles/ng.3506. Publisher: Nature Publishing Group.

[6] Stephan, J., Stegle, O. & Beyer, A. A random forest approach to capture genetic effects in the presence of population structure **6**, 7432. URL https://www.nature.com/articles/ncomms8432. Publisher: Nature Publishing Group.

[7] Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* **12**, 931–934 (2015).

[8] Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks **31**, 107663. URL https://www.sciencedirect.com/science/article/pii/S2211124720306161.

[9] Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks **28**, 739–750.

[10] Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics* **50**, 1171–1179 (2018).

[11] Avsec, . *et al.* Effective gene expression prediction from sequence by integrating long-range interactions **18**, 1196–1203. URL https://www.nature.com/articles/s41592-021-01252-x. Number: 10 Publisher: Nature Publishing Group.

[12] Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation **57**, 949–961. URL https://www.nature.com/articles/s41588-024-02053-6. Publisher: Nature Publishing Group.

[13] Avsec, . *et al.* AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. URL https://www.biorxiv.org/content/10.1101/2025.06.25.661532v1. Pages: 2025.06.25.661532 Section: New Results.

[14] Huang, C. *et al.* Personal transcriptome variation is poorly explained by current genomic deep learning models **55**, 2056–2059. URL https://www.nature.com/articles/s41588-023-01574-w. Number: 12 Publisher: Nature Publishing Group.

[15] Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings **55**, 2060–2064. URL https://www.nature.com/articles/s41588-023-01524-6. Number: 12 Publisher: Nature Publishing Group.

[16] Drusinsky, S., Whalen, S. & Pollard, K. S. Deep-learning prediction of gene expression from personal genomes. URL https://www.biorxiv.org/content/10.1101/2024.07.27.605449v1. Pages: 2024.07.27.605449 Section: New Results.

[17] Rastogi, R., Reddy, A. J., Chung, R. & Ioannidis, N. M. Fine-tuning sequence-to-expression models on personal genome and transcriptome data. URL https://www.biorxiv.org/content/10.1101/2024.09.23.614632v1. Pages: 2024.09.23.614632 Section: New Results.

[18] Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project **45**, 580–585. URL https://www.nature.com/articles/ng.2653. Number: 6 Publisher: Nature Publishing Group.

[19] Kaplan, J. *et al.* Scaling laws for neural language models. URL http://arxiv.org/abs/2001.08361. 2001.08361 [cs].

[20] OpenAI *et al.* GPT-4 technical report. URL http://arxiv.org/abs/2303.08774. 2303.08774 [cs].

[21] xAI. Grok 4 | xAI. URL https://x.ai/news/grok-4.

[22] Shendure, J. *et al.* Dna sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

[23] DeepMind. AlphaGenome API additional terms of service. URL https://deepmind.google.com/science/alphagenome/terms.

[24] Danecek, P. *et al.* Twelve years of samtools and bcftools. *GigaScience* **10**, giab008 (2021).

[25] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

Acknowledgements

This work was supported in part through the Minerva computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences.

Funding

None declared.

Competing Interests

None declared.

S1 Supplementary Methods

Data Preprocessing

For RNA-seq data, normalized expression values (V10) were downloaded from the GTEx website, with no additional processing applied. The tissue-specific expression matrix for each gene was constructed separately by excluding tissues that more than 50% of the expression values across individuals were missing.

For personal genomes, phased whole-genome sequencing data (V9) were used. The genotype matrices used for Elastic Net and Random Forest were constructed as follows. First, for each gene, variants within a 1 Mb window centered on the transcriptional start site (TSS) were extracted from phased variant call format (VCF) files separately for each haplotype. Second, each variant was encoded in binary form (0/1) per haplotype and concatenated to generate individual-level numerical feature vectors for downstream model training. Both SNVs and indels were included, and variants with a minor allele frequency below 1% were filtered out. For AlphaGenome and Enformer, the haplotype sequences were generated using the *consensus* command from the bcftools[24] package. Both haplotypes were used as input to the deep learning models, and the resulting outputs were averaged to produce the final predictions.

Use AlphaGenome and Enformer for Personal Expression Prediction

For AlphaGenome, a 1 Mb window centered on the transcriptional start site (TSS) of each gene was used as input. The model outputs predictions over the same window at single-basepair resolution. For each gene, predicted expression values from the TSS to the transcriptional end site (TES) were averaged to represent that gene's expression level. AlphaGenome's half-window size of 500 Kb is sufficient to span the full gene body (TSS to TES) for the vast majority of genes; only 422 genes–approximately 0.7% of all GTEx genes–exceed 500 Kb in length.

Enformer, by contrast, requires training to predict GTEx tissue-level expression. The same train-validation-test split as used in the Borzoi [12] model was adopted, where each GTEx gene was assigned to a set based on the genomic region closest to its TSS. A 196,608 bp window centered on the TSS was extracted from the reference genome and used as input to the pretrained Enformer to obtain an embedding. As with AlphaGenome, predicted values from TSS to TES were averaged to represent gene-level expression. The target values were defined as the average expression across all GTEx individuals for each tissue.

The position-averaged embedding vectors were then used to train a Ridge regressor to predict tissue-specific expression levels. The regularization strength was selected from 30 evenly spaced values in log space between 1 and 5. The optimal model was chosen based on R2 performance on the validation set. For personal-level expression prediction, individual genome sequences were fed into Enformer to generate embeddings, which were then input into the trained Ridge regressor to produce tissue-specific expression predictions.

Use Elastic Net and Random Forest for Personal Expression Prediction

The genotype matrices were used as input to Elastic Net and Random Forest. Model performance was evaluated using 10-fold nested cross-validation. Within the inner loop, Elastic Net hyperparameters were optimized using 5-fold cross-validation to select the regularization parameter from 30 candidate α values. Random Forest models were trained using default hyperparameter settings. All analyses were conducted using the scikit-learn package [25].

Choosing 300 Genes that Span Different R2 Values

The R2 values from 10-fold cross-validation using Elastic Net were used as a reference to assess gene predictability. Genes were grouped into 10 bins based on their R2 values across tissues: $(-\infty, 0]$, $(0, 0.01]$, $(0.01, 0.02]$, $(0.02, 0.05]$, $(0.05, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$, $(0.4, 0.5]$, and $(0.5, 1]$. From each bin, 30 genes were randomly selected, resulting in a set of 300 genes (Supplemental Table 1) that span a wide range of predictability levels.

In silico mutagenesis

In silico mutagenesis (ISM) was performed to estimate the marginal effect of individual genetic mutations on AlphaGenome's output. Each mutation was introduced into a reference sequence spanning 1 Mb centered on the TSS, and the resulting predicted expression was compared to the prediction from the unmodified reference sequence. The difference between these two predictions was taken as the mutation's effect.

For Random Forest, the reference expression was obtained using a feature vector of all zeros. Each mutation was then individually activated (set to one) to compute the corresponding alternative expression. To enable direct comparison between AlphaGenome and Random Forest, the marginal effects for all mutations were rescaled to the [0, 1] range.