Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Limited cross-model agreement in predicted RNA effects of complete non-reference human insertions

Yu Zhang
Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University
Da Lin[1]
Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University
Ying Chen[2]
Wenzhou Medical University
Yue Liu[2]
Wenzhou Medical University
Yi Zha[1]
Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University
Wangqiang Feng
healthwz@163.com

Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University

Research Article

Keywords: Human pangenome, Non-reference insertions, Functional genomics, Sequence-to-function models, Regulatory variant interpretation, Prediction reliability

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11548505/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Limited cross-model agreement in predicted RNA effects of complete non-reference human insertions

Yu Zhang[1], Da Lin[1], Ying Chen[2], Yue Liu[2], Yi Zha[1] and Wangqiang Feng[1]*

[1] The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, China.

[2] Wenzhou Medical University, Wenzhou, China.

* Correspondence: Wangqiang Feng; healthwz@163.com. ORCID: 0009-0002-9041-7960.

Abstract

Human pangenomes reveal complete insertion alleles absent from reference assemblies, creating a need to assess consistency in model predictions of their regulatory effects. We evaluated AlphaGenome and Borzoi using a frozen benchmark of 400 naturally occurring human insertions selected independently of model predictions. These insertions were absent at callable homologous loci in GRCh38 and CHM13. Paired alleles retained each complete insertion, and strict RNA effects were calculated over the same homologous target-gene exon bases. Both models provided finite scores for 318/400 loci. Strict cross-model agreement was limited: Spearman ρ was −0.052 (95% confidence interval −0.195 to 0.092), and signs agreed at 155/318 loci (48.7%). Agreement remained limited in retrospective effect-magnitude analyses; the joint upper-quartile subset contained 61 loci, with 45.9% sign concordance and ρ = −0.222. Placement comparisons were more stable, whereas model-member and sequence-context analyses revealed additional sensitivity. A separate, nonmatched comparator of 500 reference-coordinate single-nucleotide variants showed greater agreement (ρ = 0.626; sign concordance 68.2%). Differences in variant class and ascertainment limit this contrast to a descriptive comparison. Gene-preserving recovery in the Graph nonreference sequences expression-association resource yielded 9/33 jointly usable loci, leaving insufficient support for biological calibration. These findings show limited signed RNA-effect agreement within the evaluated scorable insertion subset. Pangenome functional annotation should report coordinate support, scoring coverage, context sensitivity and independent biological calibration separately.

Keywords: Human pangenome; Non-reference insertions; Functional genomics; Sequence-to-function models; Regulatory variant interpretation; Prediction reliability

Introduction

Interpreting the regulatory effects of complete sequences revealed by human pangenomes is a central problem in functional genomics. Haplotype-resolved assemblies recover structural alleles and sequence contexts that are incompletely represented by a single linear reference (Liao et al. 2023; Logsdon et al. 2025). The 1000 Chinese Pangenome extended this catalogue and combined complex-variant expression mapping with epigenomic predictions evaluated against blood chromatin-accessibility measurements (Wang et al. 2026). Graph-based analyses have also linked non-reference sequences to transcript abundance and phenotypic variation (Wu et al. 2024). These studies support the molecular relevance of some non-reference sequences, while the functions of many individual alleles remain unresolved. Sequence reconstruction identifies the complete insertion allele; functional interpretation then requires connecting that allele to a defined molecular target.

Sequence-to-function models offer such a route by predicting molecular readouts from DNA. Borzoi predicts RNA-sequencing coverage across long genomic contexts and derives regulatory-variant scores from paired sequence predictions. Its original fine-mapped expression quantitative trait locus analyses focused on single-nucleotide variants to limit insertion/deletion boundary effects (Linder et al. 2025). AlphaGenome predicts diverse regulatory outputs from approximately one megabase of sequence. Published evaluations report strong performance on variant-effect tasks, including coefficient and direction prediction for insertion/deletion expression quantitative trait loci (Avsec et al. 2026). These evaluations characterize model capabilities within their tested populations and motivate a complementary question. How consistently do the models assign signed RNA effects to complete natural insertions absent from reference assemblies when evaluated against the same biological target?

Personal-genome studies show why this question requires a distinct endpoint. Accurate prediction of expression differences across genes did not consistently translate into accurate prediction of variation among individuals (Sasse et al. 2023; Huang et al. 2023). Personal-genome training in SAGE-net supported prediction for unseen alleles at training genes, but performance did not generalize to unseen alleles at unseen genes (Spiro et al. 2026). Existing pangenome functional analyses also leave agreement for the same complete insertion and homologous gene target unresolved (Wang et al. 2026; Wu et al. 2024). Long, often repetitive insertions displace downstream coordinates, and the exon bases supported by each model can differ (Linder et al. 2025; Logsdon et al. 2025). Comparing reference and alternate alleles requires a shared homologous target and explicit accounting for unavailable scores. Agreement measures consistency between predictions, whereas biological accuracy requires relevant independent molecular measurements. An allele's absence from a reference assembly leaves its potential inclusion in model training data unresolved.

We evaluated cross-model agreement and technical sensitivity for 400 frozen, sequence-resolved human insertions absent from callable homologous loci in GRCh38 and CHM13. We paired complete insertion-containing and empty alleles and scored AlphaGenome and Borzoi RNA effects over shared homologous target-gene exon bases. We then assessed agreement in ranks and effect directions, with placement, context and member analyses examining distinct technical sensitivities. Later exploratory audits examined effect magnitude and scoring coverage. Unscorable loci remained in the frozen coverage denominator. A separate, nonmatched reference-SNV population provided a descriptive comparator, and insertion-associated expression records supported a limited assessment of biological calibration. This sequence-resolved, homology-aware benchmark distinguishes predictive consistency, technical sensitivity, scoring coverage and measured-effect support to inform pangenome functional annotation.

Methods

Study design

We compared AlphaGenome and Borzoi on complete natural insertions absent at callable homologous loci in GRCh38 and CHM13. The primary endpoint was cross-model agreement in signed RNA-density changes over shared target-gene exon bases. The 400-locus core and 100-locus robustness subset were selected before inference, independently of model predictions. Primary and technical sensitivity analyses used the frozen outputs. Later exploratory analyses examined effect magnitude, influence, missingness and eligibility for external calibration, with the original endpoints retained as the primary results. Online Resources 1–3 provide supplementary methods, publication tables and source-level records.

Source datasets

The discovery dataset was the HGSVC3 2024 v1.0 GRCh38 insertion/deletion release, generated with PAV 2.4.0.1 from 65 diploid individuals, comprising 130 haplotypes (Logsdon et al. 2025). This sequence-resolved release was selected before catalogue construction and inference; the scored core used this resource alone, without merging HPRC resources. GRCh38 supplied the reference backbone, and CHM13 supplied a second check for absence at homologous loci (Nurk et al. 2022). GNRS expression associations supplied the separate insertion-calibration panel (Wu et al. 2024). Reference-SNV effects came from GEUVADIS through eQTL Catalogue release 7, QTD000110 (Kerimov et al. 2021; Lappalainen et al. 2013). A later eligibility analysis examined the public SV-eQTL resource of Lin et al. (2026), a medRxiv preprint. Download locations, versions and hashes are listed in Table S11 and Online Resource 3.

Class B insertion eligibility

Structural variant (SV) insertions of at least 50 bp were normalized by removing shared prefix and suffix bases. Complete insertion sequences had to contain only A, C, G or T. Ordered bilateral flanks of at least 2 kb had to align with minimap2 at ≥95% identity and ≥90% coverage. The next-best alignment score had to be <90% of the best (Li 2018). A callable homologous empty allele established locus absence. Missing, partial or ambiguous alignments remained unclassified. The executed GRCh38 anchor check also required exact, fully aligned 2-kb flanks at the insertion boundary. Sequence occurrence elsewhere in the reference was annotated separately from homologous-locus absence.

Class A insertions were absent at the homologous GRCh38 locus but present in CHM13, with ≥95% insertion identity, ≥90% coverage and matching orientation. Class B additionally required a callable empty CHM13 allele and insertion-present and insertion-absent discovery haplotypes. Class C was a population-differentiated Class B subset and did not define a separate primary endpoint. Repeat annotations combined human-lineage RepeatMasker transposable-element intervals based on Dfam 4.0 and Tandem Repeats Finder intervals (Benson 1999; Storer et al. 2021). Discovery allele frequency (AF) was the number of insertion-bearing haplotypes divided by called haplotypes. Local guanine-plus-cytosine (GC) fraction, repeat density, flank mappability, gene density and missingness definitions are specified in Table S1.

Benchmark sampling

Eligible core loci contained complete Class B insertions on canonical chromosomes, with insertion length <262,144 bp and sufficient reference flanks for long-context input. Sampling strata comprised length (50–199, 200–499, 500–1,999, 2,000–5,000 and >5,000 bp), repeat category, discovery AF (<0.01, 0.01–<0.05 and ≥0.05) and transcription start site (TSS) distance (<10, 10–100 and >100 kb). Proportional allocation used floors and largest remainders; ordered identities were sampled with Python random seed 20261010. The 33 GNRS calibration loci were reserved separately. A 67-locus random extension was not pooled into the core, and the fixed 100-locus robustness subset was drawn from the core. No inference outcome affected sampling, and unavailable loci were not replaced. Nearest-TSS assignment with the frozen tie rule selected target genes for the 467 non-calibration loci, whereas GNRS retained its associated genes. Target genes and strands remained fixed throughout all analyses.

Allele reconstruction

For an input width W, REF was extracted from GRCh38 around the insertion anchor. ALT joined the same left flank, the complete insertion and the reference right flank; only distal sequence was cropped to retain width W. Baseline inputs centred the insertion at W/2 and followed genomic orientation. Target-gene strand determined RNA-track selection. Sequence length, insertion identity and required exon geometry were verified before scoring. All other bases retained the reference sequence. The pair estimates an isolated insertion effect in a reference-derived context, with other personal-haplotype variants outside its scope.

Homologous-coordinate mapping and RNA effects

Reference bases upstream of the insertion mapped directly to ALT; downstream coordinates shifted by the insertion length. Inserted bases had no REF homolog and were excluded. GENCODE annotations (Harrow et al. 2012) came from UCSC's hg38 v20 tables (UCSC Genome Browser n.d.). Exons assigned to the same gene were merged into a non-overlapping union. The scoring domain contained exonic bases represented in both alleles and every setting required for the comparison, accounting for output cropping and resolution. Borzoi's 32-bp bin boundaries had to lie within supported homologous domains; fractional exon-boundary contributions used piecewise-constant density. AlphaGenome used base-resolution coverage. Float64 integrals were divided by the same retained exon length in both alleles. Signed effects were ALT-minus-REF mean RNA density, aggregated across eligible tracks. This endpoint measures coverage over shared exon bases while retaining model-specific raw scales for rank and sign comparisons.

AlphaGenome inference

AlphaGenome inference used client distribution 0.10.0 and the official human API, with a 1,048,576-bp input, 1-bp RNA outputs and primary ALL_FOLDS predictions (Avsec et al. 2026). Stored request metadata identified tissue, assay family, strand and outputs. The service did not expose a server build or weight SHA256. API failures were recorded separately and retried under the fixed job-control policy. The member analysis used FOLD_0 through FOLD_3 without replacing the primary ALL_FOLDS endpoint. Outputs and derivatives remain subject to the AlphaGenome Output Terms of Use (Google DeepMind n.d.).

Borzoi inference

Borzoi used official human model-0 weights converted to the frozen PyTorch runtime (Linder et al. 2025). The 524,288-bp input yielded 6,144 32-bp bins over 196,608 bp after cropping 163,840 bp from each side. Replicate

0 was primary; replicates 1–3 were used only for member comparisons. All four used test-fold 3 and validation-fold 4. Conversion required all parameters to be populated with no unexpected parameters. CPU float32 evaluation used inference mode, four threads and seed 20261009. RNA unsquashing divided outputs by the track scale and inverted the soft clip. Where applicable, the inverse 0.75 power and division by 32 converted outputs to per-base density. Finiteness and output shape were checked before scoring.

Strict track harmonization and missing scores

The cached exact-family diagnostic used original predictions with matched assay families. Strict harmonization selected shared positive poly(A) and total-RNA families, averaged strand-compatible tracks within families and weighted families equally. Harmonization aligned eligible assay families and homologous scoring domains; the models retained their distinct training distributions, tissue composition, output representations and architectures. Track identities and exclusions are listed in Table S3 and Supplementary Methods. Near-family sensitivity used the compatible GM12878 RNA set. Full-RNA sensitivity averaged eligible human strand-compatible RNA tracks without tissue matching. All scoring definitions were fixed independently of observed agreement.

All 400 core identities were retained. Joint finite coverage was 318/400; the other 82 loci lacked common scorable exon support. Technical failures, unavailable outputs and exact zeros were distinguished in Tables S1 and S11. Rank statistics used finite paired effects. Sign statistics additionally required both effects to be nonzero, following the executed exact-zero rule in place of an earlier protocol tolerance. Every analysis reports its corresponding denominator.

Placement and context analyses

Placement comparisons used offsets of −32,768, 0 and +32,768 bp in the fixed 100-locus subset. Exon support was held constant across placements and alleles, yielding 72 usable loci. Context comparisons required a common intersection across settings and yielded 29 loci. AlphaGenome used native 16,384-, 131,072- and 1,048,576-bp inputs. Borzoi retained centred 16,384- or 131,072-bp sequence within its 524,288-bp input and replaced outside bases with N. Each complete insertion had to remain inside the retained context. Native shorter inputs and N-mask ablation were treated as distinct interventions.

Model-member stability

The member analysis compared four official fitted models while retaining the primary endpoints. Dominant sign fraction was the largest nonzero sign count divided by the valid nonzero member count. Four-member finite summaries included 77 loci per model; all 100 identities and incomplete cases remained in Table S4. Member pooling did not replace AlphaGenome ALL_FOLDS or Borzoi replicate 0.

A later exploratory analysis related the existing GM12878 member effects to strict primary effects at overlapping loci. It used the historical member outputs, without additional strict all-track inference. Gene-cluster intervals and the subgroup containing one jointly unanimous locus are reported in Table S14. Fitted members were treated as model instances, with biological replication assessed separately.

Effect magnitude and influence sensitivity

Retrospective exploratory analyses ranked absolute effects within each model across the 318 strict finite pairs. Separate empirical thresholds defined joint above-P25, above-P50 and above-P75 subsets while retaining model-specific raw scales. Threshold definitions preceded these exploratory calculations and were distinct from the original prespecified endpoints. A 10,000-resample target-gene-cluster bootstrap retained all loci assigned to each sampled gene and refitted thresholds in every resample. Signed-log transformation, 1st/99th-percentile winsorization, single-locus deletion and single-gene deletion examined cached effects, with the primary endpoint retained as the reference. Two-level threshold-stability analysis used 200 outer and 100 inner resamples. Its inclusion probabilities described threshold stability within this dataset and lacked validation as predictive confidence scores. Further specifications are provided in Table S13.

Reference-SNV comparator

The reference comparator was selected before model inference from SNVs with posterior inclusion probability (PIP) ≥0.9 and usable ALT-oriented expression effects. Selection met the primary criterion, leaving the prespecified lower-PIP fallback unused. GRCh38 bases, stable target-gene identifiers, strands and eligible exon support were checked against GENCODE v20 (UCSC Genome Browser n.d.). Unique SNVs were capped at five per gene and ordered using seed 20261011 and a deterministic variant–gene hash, yielding 500 SNVs across 473 genes. REF/ALT pairs differed only by the substituted base. Strict RNA scoring and common-exon integration followed the insertion endpoint framework; measured calibration compared predicted changes with ALT-oriented beta. The panel was selected independently of insertions, without matching on length, repeats, AF, target selection or potential training exposure.

A later feasibility analysis extracted natural reference-present deletions and assessed overlap and balance before inference. A same-gene/length design yielded 89 diagnostic pairs but failed admission requirements. Matched reference-SV inference and its corresponding effect contrast remained unexecuted (Table S16).

GNRS expression calibration

The 33 GNRS insertion–gene pairs and measured insertion-presence directions were fixed before recovery. From 59 eligible transcript-association records, one record per locus was selected before prediction. Records were ordered by ascending beta P, permutation P, nominal P, gene identifier and transcript identifier. Original scores, genes and beta values were retained. Gene-preserving recovery examined legal input geometry and transcript mapping while requiring the same complete insertion and homologous exons. Output-window selection was independent of measured beta, and transcript and scoring sensitivities remained separate from primary recovery. The Borzoi geometry analysis assessed legal placements under frozen annotations; its theoretical bound of 13/33 was distinct from observed joint recovery. Directional calibration required ≥20 jointly finite loci under the original protocol. Estimates from nine recovered loci remain descriptive (Table S8).

External SV-eQTL eligibility audit

The bounded external analysis began with 511 exact catalogue matches to a public GRCh38 insertion VCF and Supplementary Table 4 of Lin et al. (2026), a medRxiv preprint. Qualification required exact sequence and boundary, a target gene, verified insertion-presence effect orientation, relevant expression tissue and legal scoring geometry. It also required independent expression donors and a nonduplicated independent association signal. The original minimum of 20 qualified units was retained. Source versions, scripts, VCF alleles and donor metadata were examined; source-cohort overlap was distinguished from overlap within the actual analysis subset. Independence was assessed at the signal level rather than inferred from counts of association rows or distinct identifier pairs. Unresolved dosage orientation and donor independence prevented admission. The audit therefore stopped before formal prediction or recalibration, with missing direction information left unresolved (Table S17; Online Resource 3).

Statistical analysis

Primary metrics were Spearman rank correlation (ρ) and nonzero sign concordance. Percentile 95% confidence intervals (CIs) used 10,000 bootstrap resamples. Insertion analyses resampled paired loci. Reference analyses resampled target genes as clusters, retaining all variants assigned to each sampled gene. Rank and sign statistics were recomputed within resamples. Exact binomial tests against 0.5 and Clopper–Pearson intervals were retained separately in source tables and did not replace the bootstrap intervals.

The primary nonmatched benchmark difference was reference strict minus insertion strict (TIER1_EXACT). Independent 10,000-resample streams resampled reference target-gene clusters and insertion paired loci. Seeds and the historical exact-family sensitivity contrast are specified in Supplementary Methods. Scores, filters, denominators and tracks remained unchanged. Reference measured-direction nulls used 10,000 target-gene-cluster sign permutations and a plus-one P-value rule; Benjamini–Hochberg correction was applied within the declared family (Benjamini and Hochberg 1995). Table S10 distinguishes executed nulls from unavailable historical matched controls. Exploratory covariate models used L2 logistic regularization C = 1, ridge alpha = 1, 2,000 paired-locus bootstrap resamples, frozen full-core scaling and five seeded cross-validation folds. Table S5

reports the 34-coefficient q-value family and rank diagnostics; features were not selected using agreement outcomes.

Retrospective gene-cluster sensitivity intervals were reported separately from historical paired-locus intervals. Missingness analyses retained all 400 loci and compared scorable with unavailable groups using fixed feature definitions. Nearest-RefSeq TSS distance was distinguished from distance to the selected GENCODE target gene. Descriptive tests used their recorded Benjamini–Hochberg families, with separate gene-cluster intervals for standardized differences. Mapped exon support was analysed as an availability mechanism and excluded from baseline matching covariates. Exploratory strata and confidence cutoffs were fixed independently of observed agreement.

Software and reproducibility

The inference environment recorded Python 3.12.13, NumPy 2.5.3, SciPy 1.18.1, pandas 3.0.6, matplotlib 3.11.2, PyTorch 2.14.1 and borzoi-pytorch 0.5.1 (Harris et al. 2020; Hunter 2007; Virtanen et al. 2020). Table S11 documents sequence-processing dependencies and runtime provenance. Later analysis environments are identified separately in Online Resource 3. SHA256 manifests link sequence, job, scoring and statistical records. Reproducibility is limited by the unexposed AlphaGenome server-weight fingerprint and separately obtained upstream inputs.

AI assistance

ChatGPT (OpenAI) assisted with code writing, text translation, language polishing and formatting. Human authors retain responsibility for verifying the code, sources, results and interpretation and for the final content of the manuscript.

Results

A sequence-resolved benchmark of complete non-reference human insertions

The core comprised 400 Class B insertions assigned to 387 target genes (Fig. 1; Table 1). Complete insertions were absent at callable homologous loci in both GRCh38 and CHM13. Insertion-present and insertion-absent alleles were observed in discovery haplotypes. Sampling preceded inference, and all unavailable loci remained in the benchmark. The resulting sample represents the eligible catalogue under the specified selection criteria.

The median insertion length was 196.5 bp (interquartile range 87.75–332.25; range 50–38,062), and the median repeat fraction was 0.9826. Repeat categories included 155 VNTR/STR, 124 SINE/Alu, 43 non-repeat, 38 LINE, 30 other-repeat and 10 LTR loci. Joint finite scoring was supported at 318/400 loci (79.5%), spanning 313 genes. The other 82 loci lacked common scorable exon support and remained in all identity and coverage records (Online Resource 2, Table S1).

Table 1 Benchmark characteristics and scoring coverage

Characteristic	Reference-SNV comparator	Insertion core
STUDY DESIGN		
Source / design	Expression-associated, high-PIP SNVs	HGSVC3; outcome-independent stratified sampling
Variant class	Single-base substitution	Complete natural Class B insertion
Frozen variants / target genes	500 / 473	400 / 387
Ascertainment	PIP ≥0.9; unique SNVs; ≤5 per gene	Approved sequence-resolved catalogue; no outcome replacement
SEQUENCE CHARACTERISTICS		
Variant / inserted length	Single-base substitution; net inserted length 0 bp	Complete inserted sequence; median 196.5 bp; IQR 87.75–332.25; range 50–38,062
Inserted-sequence repeat fraction	Not applicable	Median 0.9826; IQR 0.9460–1.0000
Discovery AF availability / definition	Not available	400/400 loci recorded; ALT / called discovery haplotypes
SCORING COVERAGE		
Finite variants / frozen variants	500/500	318/400 (79.5%); 82 unavailable
Genes with finite scores	473	313
BIOLOGICAL EVIDENCE		
Measured expression evidence	ALT-oriented beta for 500 SNVs	Separate GNRS resource: 9/33 jointly usable; below N=20

The populations are nonmatched and differ in variant class and selection. Reference SNVs are single-base substitutions with net inserted length 0 bp; insertion lengths describe complete inserted sequences. Not applicable indicates an undefined characteristic; Not available indicates unsupplied data. AF availability counts loci with a recorded value. Discovery AF is ALT haplotypes divided by called discovery haplotypes and describes the discovery sample. Reference measured-beta comparisons and the separate GNRS recovery use different evidence designs. Finite gene counts refer to genes represented among jointly finite variants. AF, allele frequency; IQR, interquartile range; PIP, posterior inclusion probability.

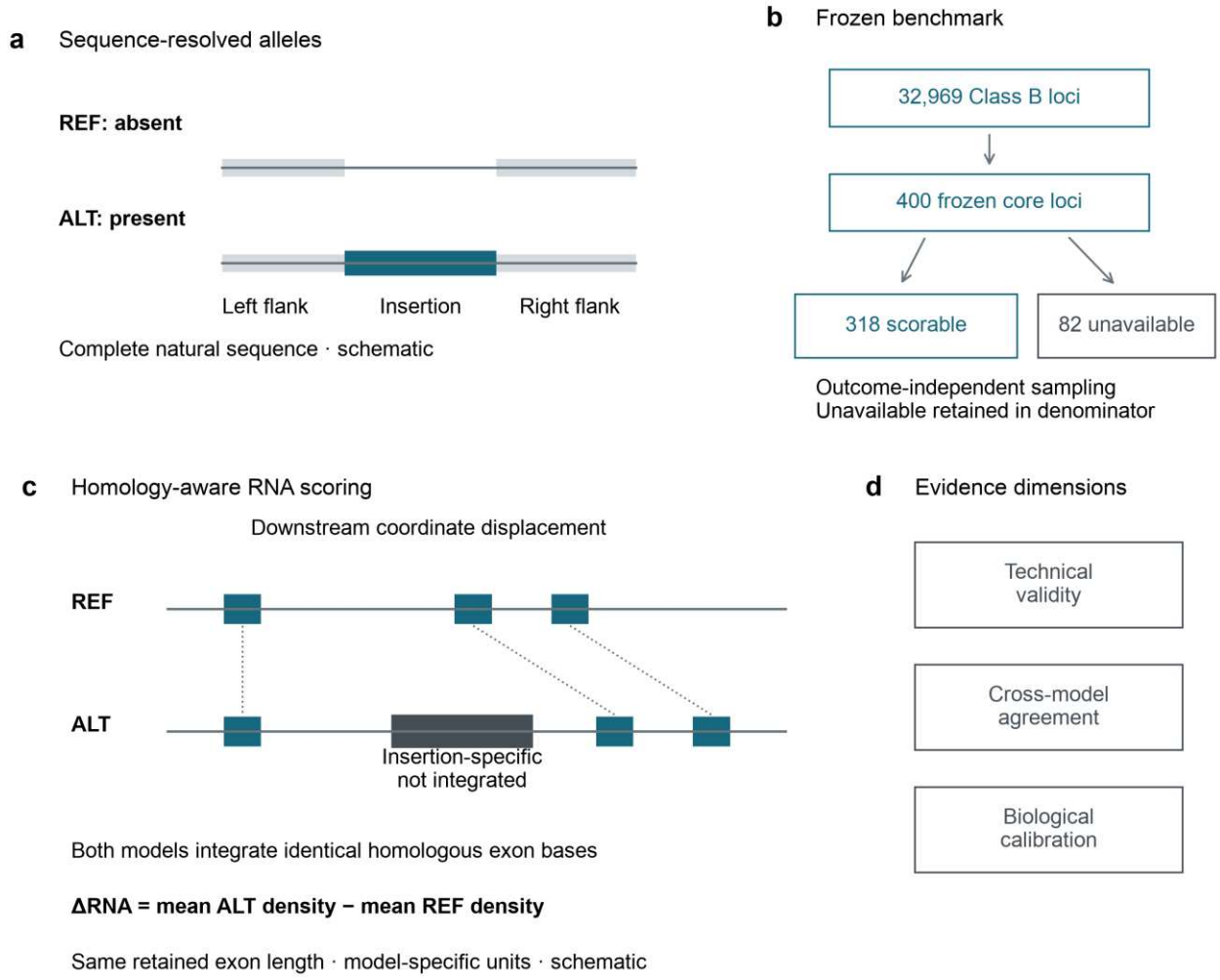


Fig. 1 Benchmark construction and homologous-exon RNA scoring. a Complete insertion-absent REF and insertion-present ALT alleles share homologous flanks. b Sampling independent of model predictions fixes 400 loci from 32,969 approved Class B loci. All 400 remain in the denominator, comprising 318 jointly scorable and 82 unavailable loci. c Schematic mapping links identical target-gene exon bases despite downstream coordinate displacement. Insertion-specific ALT bases without REF counterparts are excluded from the common mask. Each model uses the same retained exon length to calculate per-base mean RNA density; ΔRNA is ALT minus REF in model-specific units. Panels a and c are schematics without real genomic coordinates. d Technical validity, cross-model agreement and biological calibration are separate evidence dimensions, each requiring its own assessment. Counts describe frozen identities and scoring support, with biological validation considered separately

Limited agreement between AlphaGenome and Borzoi

Strict RNA-track harmonization yielded Spearman ρ = −0.052 (95% paired-locus bootstrap confidence interval [CI] −0.195 to 0.092). Sign concordance was 155/318, or 48.7% (95% CI 43.4–54.4%; Fig. 2; Table 2). All 318 finite pairs were nonzero. Original scoring gave ρ = −0.111 (95% CI −0.250 to 0.033) and 149/318 concordant signs, or 46.9% (95% CI 41.5–52.5%). These statistics describe agreement between predicted effects, with biological accuracy assessed separately through measured-effect evidence.

The exact-family diagnostic reproduced the original summary. The near-family definition gave the same strict summary estimates, whereas the tissue-unmatched full-RNA definition yielded ρ = −0.089 (95% CI −0.231 to 0.053). Sign concordance under full-RNA scoring was 48.7%, and agreement remained limited across the tested

definitions. The interval containing 50% concordance leaves equivalence to random prediction unresolved. Track identities and all scoring definitions are listed in Online Resource 2, Tables S3 and S12.

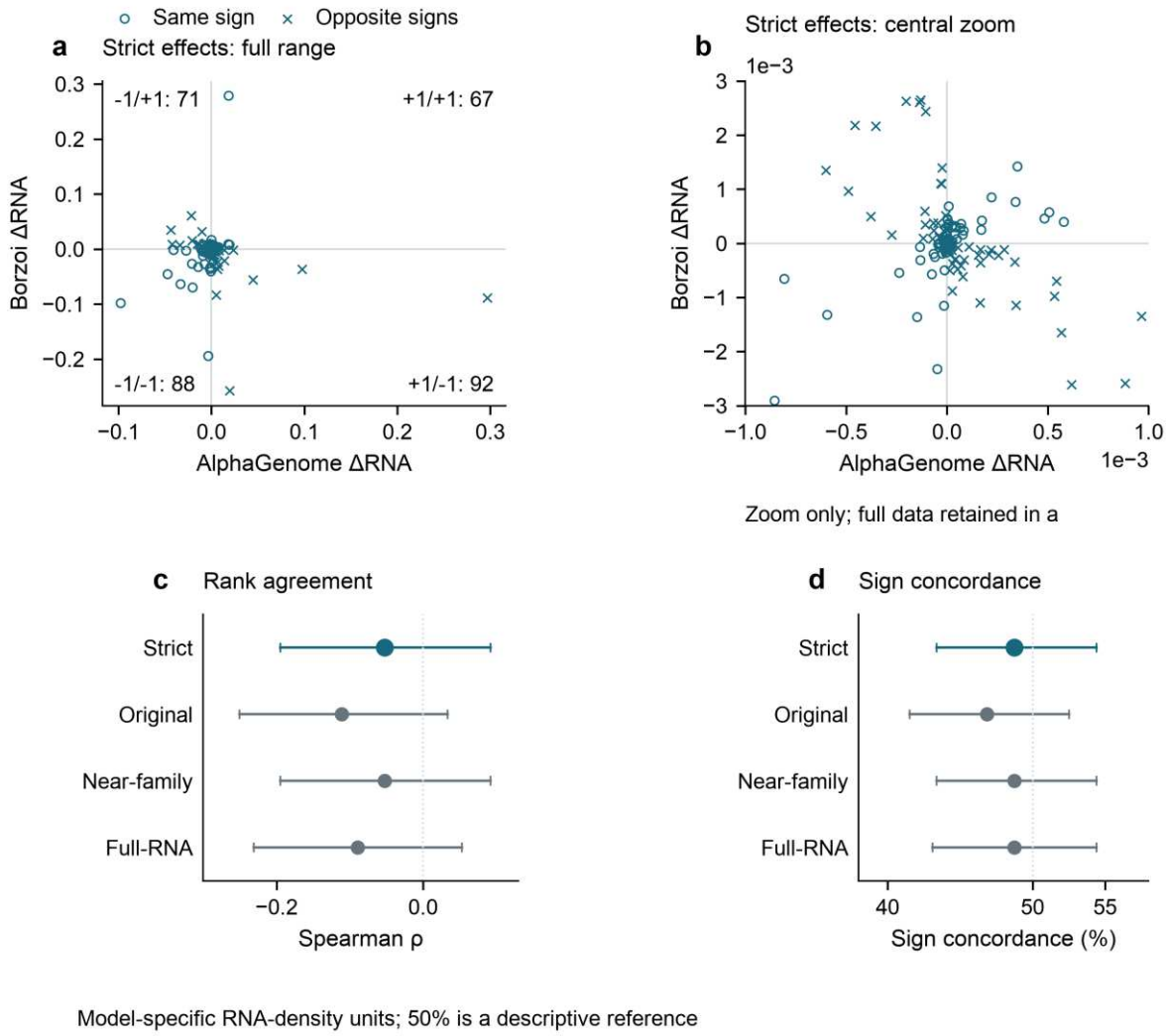


Fig. 2 Strict cross-model agreement and scoring sensitivity. a Full-range strict RNA effects for all 318 finite paired insertions from 400 frozen loci. Open circles denote matching nonzero signs; crosses denote opposite signs. Corner labels give full-dataset counts by horizontal/vertical sign. Exact-zero pairs would form a separate category; all pairs here are nonzero. b Central zoom of the same data (AlphaGenome −0.001 to 0.001; Borzoi −0.003 to 0.003), with all observations retained in a. Each axis retains its model-specific RNA-density units. c Rank agreement and d sign concordance under strict, original, near-family and full-RNA scoring, each with 318 pairs. Points show frozen estimates, with percentile 95% CIs from 10,000 paired-locus bootstrap resamples. The 50% sign guide is a descriptive reference; biological accuracy and equivalence to random prediction require separate assessment. Exploratory magnitude thresholds are reported in Fig. S12 and Table S13

Disagreement persists under effect-magnitude and influence sensitivity analyses

Retrospective exploratory analyses selected loci whose absolute effects exceeded each model's own magnitude percentile. Joint above-P25, above-P50 and above-P75 subsets retained 213, 138 and 61 loci, with sign concordances of 45.54%, 48.55% and 45.90%, respectively (Online Resource 1, Fig. S12; Online Resource 2, Table S13). The above-P75 subset had 28/61 concordant signs and ρ = −0.22247 (95% gene-cluster bootstrap CI −0.50978 to 0.09655); its sign interval was 32.73–60.00%. These subsets retain progressively larger within-model effects without equating the two models' raw scales.

Signed-log transformation preserved strict ranks. A 1st/99th-percentile winsorization sensitivity gave ρ = −0.05281. Every single-locus and single-target-gene deletion retained ρ between −0.06231 and −0.04348. The primary endpoint retained all observations. Gene-cluster bootstrap sensitivity on all 318 loci also retained limited agreement (ρ interval −0.19603 to 0.09261; sign interval 43.26–54.29%). These post hoc analyses assessed sensitivity to effect magnitude and influential observations; predictive confidence thresholds and biological accuracy remained unvalidated (Table S13).

Placement stability differs from model-member and sequence-context sensitivity

The fixed 100-locus robustness subset supported placement comparisons on a common exon intersection at 72 loci. For offsets of −32,768 and +32,768 bp, AlphaGenome's rank correlations with baseline were 0.881 and 0.786, with sign concordances of 91.7% and 88.9%. Borzoi's corresponding values were 0.971 and 0.991, with 98.6% and 97.2% concordance (Fig. 3; Online Resource 2, Table S12). These within-model placement comparisons were more concordant than the cross-model insertion comparison.

Placement and fitted-member comparisons showed different stability patterns. Four finite member effects were available at the same 77/100 loci in both models; 10/77 AlphaGenome loci and 7/77 Borzoi loci had unanimous within-model signs. Median dominant sign fraction was 0.75 in both models. One of the 77 jointly finite loci (1/77) was internally unanimous in each model, with opposite signs between models. These historical GM12878 member effects used an endpoint distinct from strict all-track scoring. The member counts describe fitted models, with biological replication considered separately (Fig. 3; Online Resource 2, Tables S4 and S14).

Context comparisons used 29 loci with fixed common exon support. Relative to its native 1,048,576-bp input, AlphaGenome's 16,384-bp and 131,072-bp inputs gave ρ = 0.056 and 0.500, respectively. Corresponding sign concordances were 48.3% and 69.0%. Borzoi's N-mask restrictions to 16,384 and 131,072 bp within its 524,288-bp input gave ρ = 0.149 and 0.753, respectively. Corresponding sign concordances were 58.6% and 82.8% (Fig. 4). Native shorter inputs and N-mask ablation are distinct interventions. Intervals compare each restricted setting with its native baseline; formal tests between restricted lengths were outside the executed analysis.

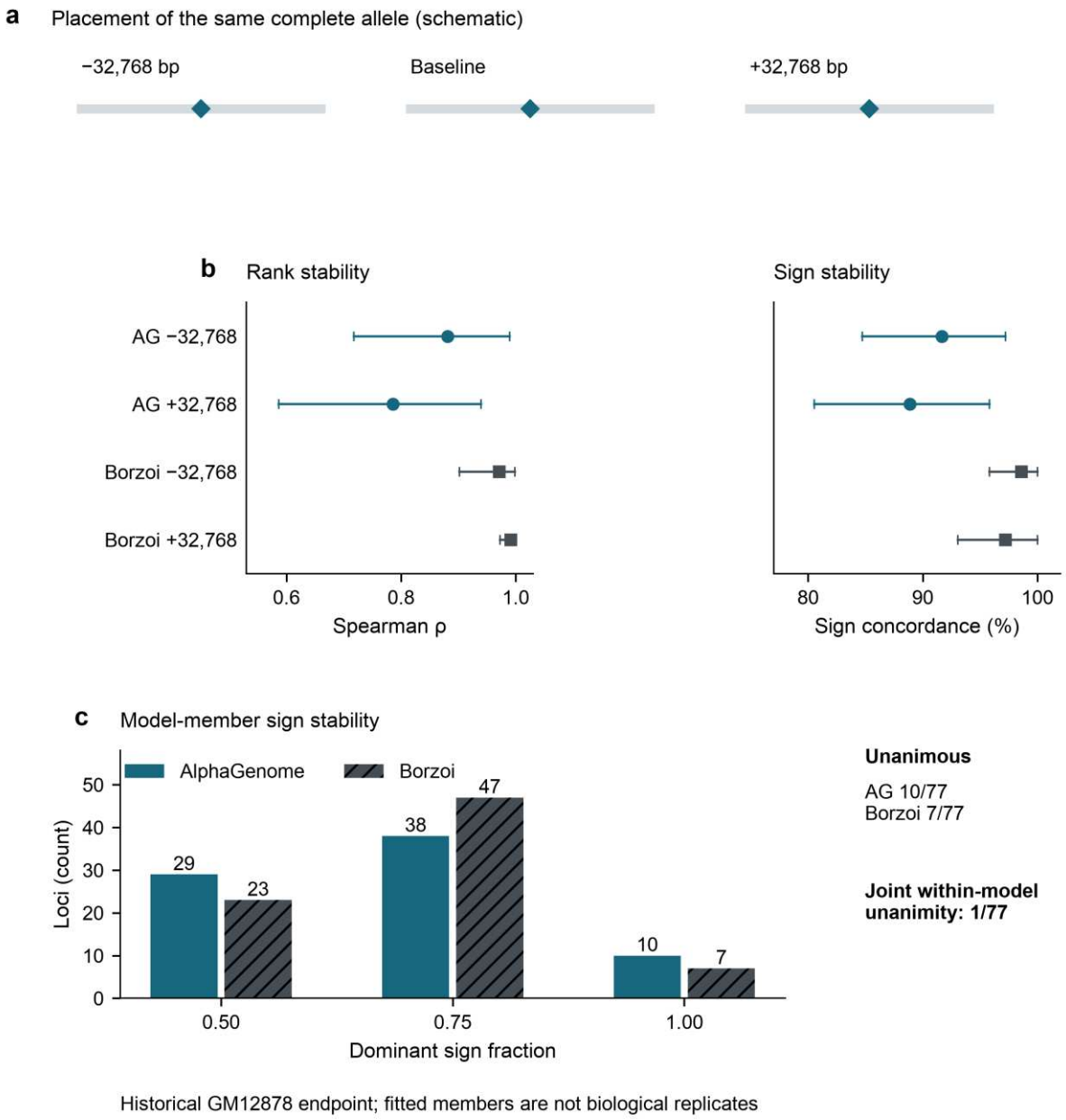


Fig. 3 Within-model placement and fitted-member stability. a Schematic baseline and prespecified −32,768-bp and +32,768-bp placements of the same complete allele. b Rank and sign stability relative to baseline, with 72/100 common-support loci per model and shift. Points show frozen estimates, with percentile 95% CIs from 10,000 paired-locus bootstrap resamples. AG denotes AlphaGenome. c Counts of dominant sign fractions across four finite fitted members at 77/100 loci per model. AlphaGenome counts at 0.50, 0.75 and 1.00 are 29, 38 and 10; Borzoi counts are 23, 47 and 7. The finite sets coincide, yielding joint N = 77. One of these loci (1/77) is internally unanimous in each model, with opposite signs between models. Bars show descriptive counts without newly estimated CIs. These historical GM12878 effects use an endpoint distinct from strict all-track member inference. Fitted members describe model variation and are not independent biological replicates. Unavailable identities remain in the source records

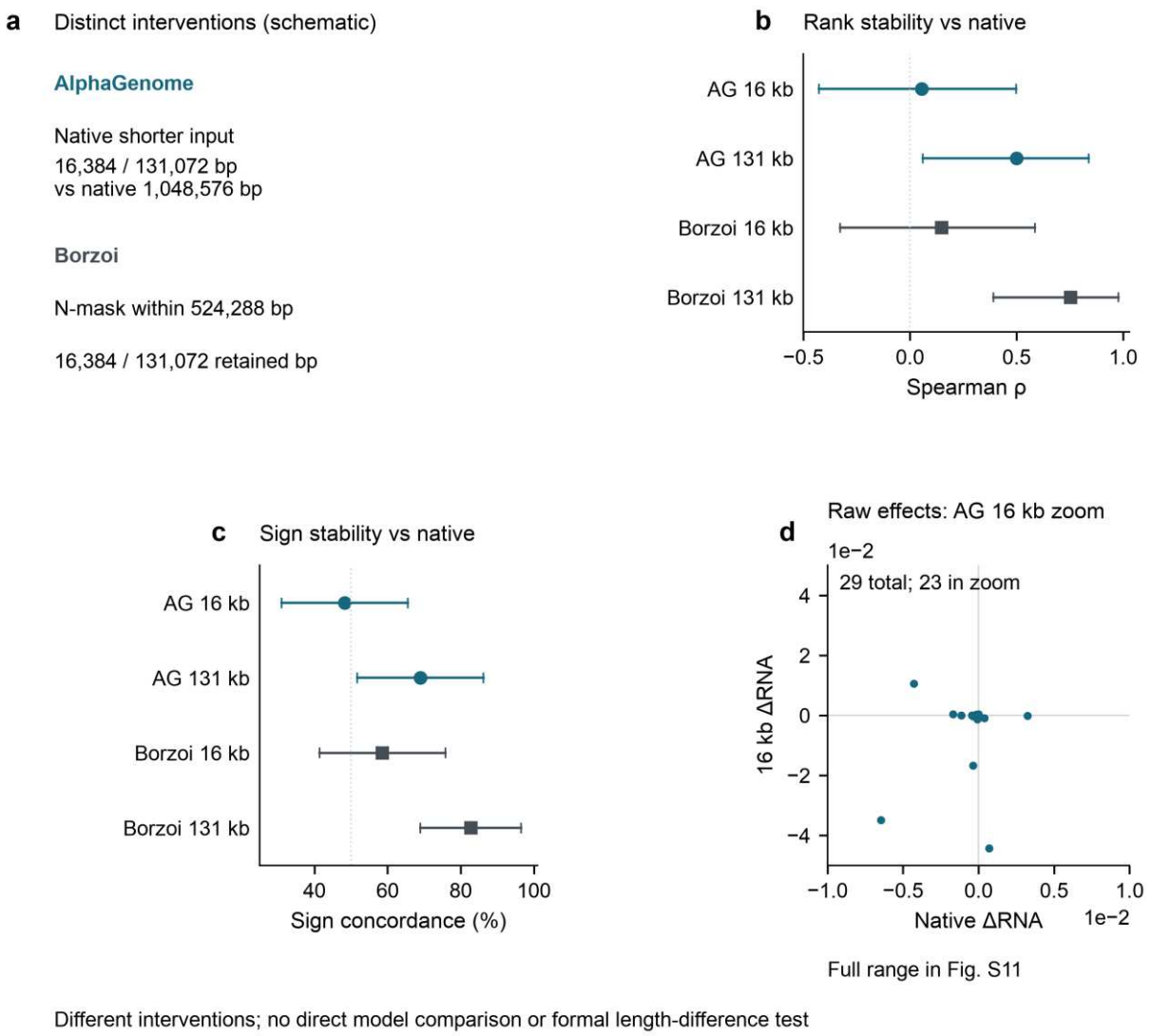
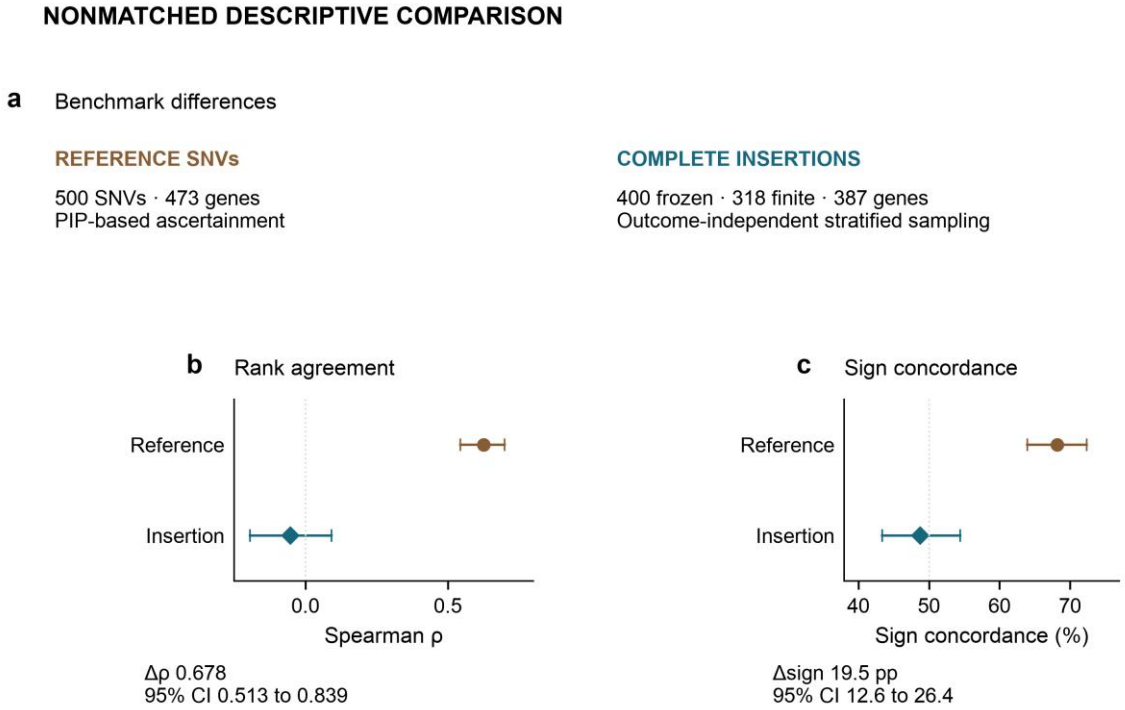


Fig. 4 Sensitivity to distinct context interventions. a Schematic interventions: AlphaGenome uses native 16,384/131,072-bp inputs compared with 1,048,576 bp; Borzoi N-masks outside retained 16,384/131,072-bp regions within 524,288 bp. b Rank and c sign stability compare each condition with its native baseline at 29/100 loci with fixed common support. AG denotes AlphaGenome; circles and squares identify AlphaGenome and Borzoi, respectively. Points and lines show frozen estimates and percentile 95% CIs from 10,000 paired-locus bootstrap resamples, with native self-identities omitted. d Central viewport of all 29 AlphaGenome 16,384-bp/native effect pairs: native −0.01–0.01, shortened −0.05–0.05; 23 are visible. Full-range data and all other comparisons are in Fig. S11; model-specific ΔRNA units remain untransformed. The interventions are treated as categorical settings; dose–response, between-model significance and differences between shortened lengths were not formally tested

Reference-coordinate variants provide a nonmatched descriptive comparison

The separate comparator contained 500 high-posterior-inclusion-probability reference-coordinate single-nucleotide variants (SNVs) spanning 473 genes. Strict cross-model agreement was $\rho = 0.626$ (95% target-gene-cluster bootstrap CI 0.544–0.699), with 341/500 concordant signs (68.2%; 95% CI 63.9–72.4%). Reference-minus-insertion differences were $\Delta\rho = 0.678$ (95% CI 0.513–0.839) and 19.5 percentage points in sign concordance (95% CI 12.6–26.4; Fig. 5). The SNVs and insertions differed in variant class, length, repeats, ascertainment, target selection and possible training exposure. The contrast describes differences between benchmark populations, leaving the contribution of reference absence unresolved. Full-range reference-SNV effects are retained in Online Resource 1, Fig. S13.

Against ALT-oriented reference expression-association beta, AlphaGenome had ρ = 0.522 and 72.4% sign concordance; Borzoi had ρ = 0.397 and 65.0% concordance (Online Resource 2, Table S12). Gene-cluster sign-permutation analyses gave P and adjusted q = 0.000100 at their 10,000-permutation resolution. A reference-present structural-variant feasibility analysis yielded 89 diagnostic same-gene/length matches, below its initial target of 100 and with unresolved comparability. Matched reference-SV inference remained unexecuted, leaving these 89 diagnostic matches insufficient for a controlled structural-variant contrast (Online Resource 2, Table S16).



d Comparator nonexchangeability

Feature	Reference SNVs	Insertions
Variant class	Single-base substitution	Complete structural allele
Length definition	1-base substitution; net 0 bp	Inserted: 50–38,062 bp
Inserted repeat fraction	Not applicable	Median 0.9826
Ascertainment	PIP ≥0.9; capped per gene	Stratified, outcome-independent
Target selection	Expression-associated gene	Nearest-TSS assignment
Training exposure	Not fully established	Not fully established

Fig. 5 Nonmatched descriptive reference-SNV comparison. a The populations differ in variant class, ascertainment and target selection. The comparator contains 500 reference SNVs across 473 genes; the core contains 400 frozen insertions across 387 genes, including 318 finite pairs. b Rank agreement and c sign concordance, with frozen percentile 95% CIs. Circles identify reference SNVs and diamonds identify insertions. Reference intervals resample target-gene clusters, retaining assigned variants; insertion intervals resample paired loci. Descriptive differences use independent 10,000-resample reference gene-cluster and insertion paired-locus streams with the frozen seeds. Differences are Δρ = 0.678 (95% CI 0.513 to 0.839) and Δsign = 19.5 percentage points (12.6 to 26.4). The 50% guide is descriptive. d Verified comparator characteristics, with unresolved training exposure left unassigned. The nonmatched comparison describes agreement across the selected populations; reference-absence causality and accuracy loss remain unresolved. Full-range reference scores are in Fig. S13; insertion scores are in Fig. 2. Measured-beta calibration is reported in Table S12

Biological calibration and scoring coverage remain constrained

The separate Graph nonreference sequences (GNRS) panel contained 33 frozen insertion–gene candidates. Original scoring yielded seven jointly usable loci; recovery preserving the gene, complete insertion and measured direction increased this to 9/33. The theoretical Borzoi geometry bound of 13/33 described possible geometric support, separately from observed joint recovery. The nine usable loci remained below the original N = 20 minimum for directional calibration (Fig. 6; Online Resource 2, Table S8). Their model-specific estimates remained descriptive, with insufficient evidence to rank biological accuracy.

The external SV-eQTL eligibility analysis confirmed exact inserted sequence and boundary at 511 catalogue loci, representing 758 candidate SV–gene combinations. Certification required verified dosage-to-beta orientation and independence of the actual analysis donors; both remained unresolved. Admission therefore stopped with no new independent, direction-resolved calibration unit certified (Online Resource 1; Online Resource 2, Table S17; Online Resource 3).

Score availability also depended on genomic features. Scorable and unavailable loci had median nearest-RefSeq transcription start site (TSS) distances of 26,243 and 206,047 bp, respectively. The standardized difference in log distance was −1.721 (95% gene-cluster CI −2.045 to −1.455); GC and repeat composition also differed (Fig. 6; Table S15). Agreement estimates apply to the 318 output-supported loci. Extending them to the unavailable 82 or the entire Class B catalogue requires assumptions about this feature-dependent coverage.

Table 2 Core agreement and nonmatched descriptive differences

Analysis	N analyzed / frozen	Rank statistic (95% CI)	Directional statistic (95% CI)	Evidence category
Insertion strict	318/400	ρ -0.052 (-0.195 to 0.092)	48.7% (43.4 to 54.4)	Frozen primary comparison
Insertion original	318/400	ρ -0.111 (-0.250 to 0.033)	46.9% (41.5 to 52.5)	Frozen historical scoring
Joint above-P75	61/318 scorable; 61/400 frozen	ρ -0.222 (-0.510 to 0.097)	45.9% (32.7 to 60.0)	Retrospective exploratory
Reference-SNV strict	500/500	ρ 0.626 (0.544 to 0.699)	68.2% (63.9 to 72.4)	Nonmatched comparator
Reference − insertion Δρ	500 vs 318 (400 frozen)	Δρ 0.678 (0.513 to 0.839)	Not applicable	Nonmatched descriptive contrast
Reference − insertion Δsign	500 vs 318 (400 frozen)	Not applicable	Δ 19.5 pp (12.6 to 26.4)	Nonmatched descriptive contrast

Parentheses give frozen percentile 95% CIs. Spearman rank correlation is denoted by ρ, and Δρ denotes a difference in correlations. Δsign is a difference in percentage points (pp), distinct from a concordance percentage. Primary and historical insertion intervals use 10,000 paired-locus resamples; reference intervals use target-gene clusters. P75 is retrospective exploratory and uses gene-cluster resampling with threshold refitting; its denominator is 61/318 scorable and 61/400 frozen loci. The two nonmatched differences use independent reference gene-cluster and insertion paired-locus streams and provide descriptive population contrasts. Placement, context and measured-beta results are in Table S12; GNRS eligibility is in Table S8, and additional magnitude/gene-cluster sensitivities are in Table S13. Independent biological calibration remains unavailable, and unperformed or unknown quantities retain their recorded statuses

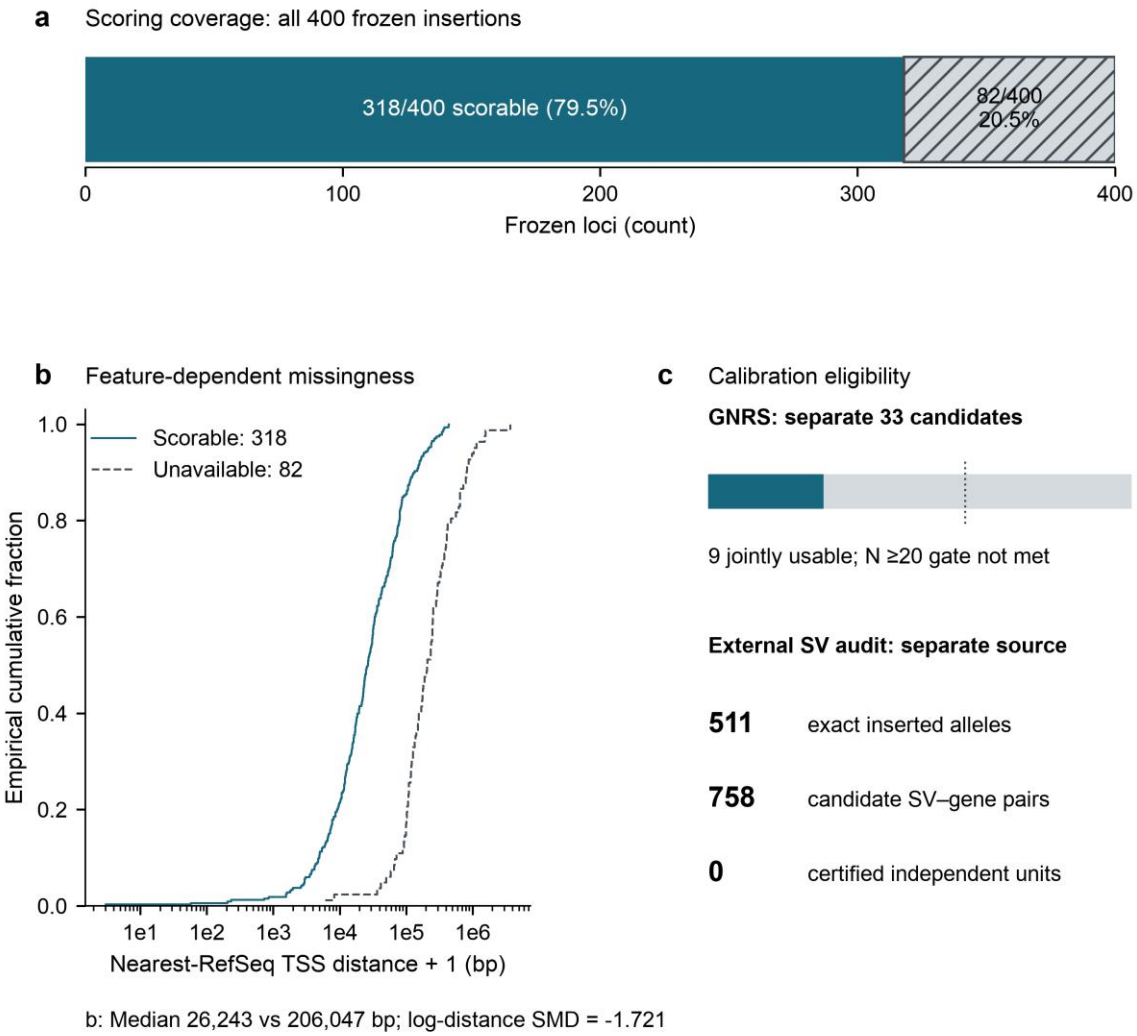


Fig. 6 Scoring coverage, feature-dependent missingness and calibration eligibility. a The full 400-locus denominator comprises 318 scorable (79.5%) and 82 unavailable (20.5%) insertions. b Empirical cumulative distributions use all 318 and 82 frozen nearest-RefSeq TSS distances, displayed as distance plus one on a log axis. Frozen medians are 26,243 and 206,047 bp; the reported log-distance standardized mean difference (SMD) is −1.721. The panel displays the frozen descriptive results, without additional confidence intervals or missingness tests. This covariate differs from distance to the selected GENCODE target gene. c Separate eligibility sources: GNRS yields 9/33 jointly usable candidates below the original N ≥20 gate. The external audit yields 511 exact inserted alleles, 758 candidate SV–gene combinations and zero newly certified independent calibration units. These sources are separate panels; sequence matches establish identity, with independent expression validation and prediction outcomes assessed separately. An additional 89 diagnostic reference-SV matches left matched-SV inference unexecuted. Source records preserve unavailable and unknown states

Discussion

Technical reproducibility and predictive agreement are distinct properties of sequence-based functional annotation. Complete allele reconstruction and homologous-exon scoring provided an auditable comparison, and placement perturbations were comparatively stable. AlphaGenome and Borzoi nevertheless showed limited signed RNA-effect agreement across jointly scorable insertions. Agreement remained limited under strict RNA harmonization and the tested magnitude restrictions. The findings characterize cross-model consistency for this insertion population and motivate separate consideration of allele properties, model differences and evaluation design. Identifying biological accuracy or a causal contribution of reference absence requires additional evidence.

Several computational explanations remain plausible. The insertions contain extensive repetitive sequence, and structural alleles can alter both sequence content and surrounding genomic context (Ebert et al. 2021; Logsdon et al. 2025). Length, repeat composition and sequence complexity may each affect prediction; their individual contributions remain unresolved in this benchmark. AlphaGenome is trained on multiple regulatory modalities, whereas Borzoi is trained to predict RNA coverage (Avsec et al. 2026; Linder et al. 2025). RNA harmonization aligned the scoring target while preserving differences in architecture, training distributions, resolution, output representation and biological conditions. Context-dependent regulatory interactions offer a possible explanation for sensitivity to surrounding sequence (Toneyan and Koo 2024). Placement changes relocate sequence within a window while largely preserving available information. Context restriction changes that information, whereas switching fitted members changes learned parameters. These distinct interventions help explain how placement stability can coexist with context and member sensitivity; the responsible regulatory elements remain unidentified. Only one of the 77 jointly finite loci was internally unanimous in each historical member panel, and the models predicted opposite signs there. DNA-language-model benchmarks and regulatory-genomics syntheses also emphasize dependencies on training objectives and evaluation tasks (Alfisi et al. 2026; Nagai et al. 2026). The former evaluated zero-shot ClinVar SNV scores, an endpoint distinct from supervised RNA prediction. These studies support examining the contributing factors separately, with architecture considered alongside the other differences. Near-zero predictions can produce sensitive signs, yet the retrospective joint P75 subset retained limited agreement. Its 61 loci and wide confidence interval limit precision. This filter remains exploratory, and agreement at more extreme prediction quantiles remains unresolved.

Published regulatory-variant benchmarks for Borzoi and AlphaGenome provide a complementary perspective on the limited agreement observed here (Avsec et al. 2026; Linder et al. 2025). Those studies evaluate predictions against measurements in selected variant populations, with AlphaGenome explicitly including indel eQTL comparisons. Our endpoint compares models on complete natural insertions under shared homologous scoring support. Personal-genome studies similarly distinguish cross-gene expression prediction from variation among individuals (Huang et al. 2023; Sasse et al. 2023). SAGE-net improves prediction of unseen alleles at training genes, but performance does not generalize to unseen alleles at unseen genes in its reported experiments (Spiro et al. 2026). A non-peer-reviewed blood-eQTL benchmark reports stronger measured-effect alignment and inter-model convergence for fine-mapped than nominal associations, with estimates varying by ancestry (Sun et al. 2026). Another non-peer-reviewed study reports widespread underestimation of AlphaGenome effects (Drusinsky and Pollard 2026). It also reports greater directional support among SNV predictions in approximately the top 0.1% of prominence. Prominence ranks effects relative to neighbouring background variants, whereas our P75 intersection uses each model's insertion-effect distribution. The filters, populations and accuracy endpoints address different questions and should be interpreted within those settings. Genome representation can also affect measured functional signals and downstream analyses (Macias-Velasco et al. 2026). That comparison concerns genome choice, a different estimand from cross-model agreement for fixed alleles. Together, these studies characterize task-specific model capabilities and clarify the distinct contribution of evaluating complete-insertion agreement.

The methodological contribution is a sequence-resolved, homology-aware evaluation of the same complete alleles across models. Full insertion-containing and empty alleles retain natural structural variation, while integration compares identical homologous target-gene exon bases. This shared scoring domain keeps the biological target constant despite downstream coordinate displacement. Existing pangenome studies provide molecular associations and functional-prediction evaluations for non-reference sequences (Wang et al. 2026; Wu et al. 2024). Our evaluation adds explicit separation of sequence validity, scoring coverage, technical sensitivity, predictive agreement and biological calibration for the same complete alleles. All 400 frozen identities remain in the coverage denominator, including 82 without common scorable exon support. Outcome-independent inclusion makes coverage loss visible and preserves difficult loci regardless of their predictions. The contribution lies in the evaluation design and its application to existing models.

For pangenome annotation workflows, this separation suggests three complementary evidence layers. Sequence and coordinate evidence should document complete REF/ALT alleles, syntenic anchors, target-gene assignment, homologous exon support and the scorable fraction. Predictive evidence should report rank and sign agreement alongside placement, context and member sensitivity. Biological evidence requires tissue-relevant measurements, exact effect-allele orientation and independent samples. Stable predictions can share biases, while

unstable predictions can still concern functional alleles. Discordant loci can motivate independent experiments, and their functional status should be evaluated using molecular evidence. Pathogenicity assessment likewise requires evidence beyond predictive agreement. Reporting these layers separately would make annotations easier to interpret while keeping the status of biological support explicit. The recommendations concern evidence reporting; their use as a reliability score or clinical decision rule would require external validation.

Several limitations define the evidence needed to extend these findings. The 500 reference SNVs differ from insertions in class, length, repeat composition and ascertainment, leaving the causal contribution of reference absence unresolved. The 89 diagnostic structural-variant matches yielded no admitted scored control set; a balanced reference-present SV comparator remains necessary. Coverage is selective: 318/400 loci were jointly scorable, and unavailable loci differed in TSS distance and other features. Agreement estimates therefore apply to loci with supported scoring geometry, motivating broader coverage and explicit assessment of missingness. Nearest-TSS assignment for non-calibration loci may miss true regulatory targets; independently supported eQTLs, enhancer–gene links or perturbations would strengthen target assignment. GNRS supplied 9/33 jointly usable records, below the original N=20 calibration gate. The external audit found 511 exact insertion matches and 758 candidate SV–gene combinations, with no newly certified independent calibration units. These counts describe sequence identity and eligibility; inference stopped before model evaluation. Establishing biological accuracy requires documented counted dosage, actual analysis participants, donor independence and independent expression-effect estimates. Our homologous-exon RNA-density endpoint excludes insertion-specific transcription without a REF homolog and does not fully represent changes in splicing or chromatin organization. Those molecular readouts require separately defined prediction and validation tasks. Examples include splice prediction (Jaganathan et al. 2019) and sequence-based prediction of three-dimensional genome folding (Fudenberg et al. 2020). Reference-derived counterfactual inputs also omit other personal-haplotype variants, limiting inference about interactions among cis variants. Member estimates used 77 finite loci per model, and context analyses involved small subsets with different interventions across architectures. The historical GM12878 member endpoint also differs from strict all-track scoring. Fitted members describe model variation, with independent biological replication still needed. Unavailable server-weight fingerprints and incompletely known training exposure further restrict reproducibility and attribution. Future studies should combine sequence-resolved personal haplotypes, balanced structural comparators and independently measured molecular effects. This evidence would connect auditable prediction comparisons to biological interpretation across defined tissues and targets.

Conclusions

Complete non-reference human insertions showed limited cross-model signed RNA-effect agreement in the evaluated scorable subset. The pattern persisted across the tested scoring, magnitude and influence sensitivity analyses. Placement, member and context comparisons captured distinct technical properties. Functional annotation of these alleles should report homologous-coordinate support, scoring coverage and uncertainty alongside the status of independent biological calibration.

Supplementary information

Online Resource 1 contains Supplementary Methods and Results and Figures S1–S13. Online Resource 2 contains Tables S1–S18 with identities, full-precision scores, technical sensitivities and audit summaries. Online Resource 3 contains source-level numerical and eligibility records, figure sources, version evidence, scripts and reproducibility manifests. Together, these resources provide the publication supplements and their supporting provenance.

Statements and Declarations

Funding

This research received no specific funding.

Competing interests

The authors declare that they have no competing interests.

Author contributions

Yu Zhang: Conceptualization, methodology, software, data curation, formal analysis, visualization, and writing: original draft. Da Lin: Methodology, software, data curation, formal analysis, and writing: review and editing. Ying Chen: Investigation, data curation, validation, and writing: review and editing. Yue Liu: Investigation, validation, visualization, and writing: review and editing. Yi Zha: Investigation, validation, and writing: review and editing. Wangqiang Feng: Conceptualization, supervision, project administration, resources, and writing: review and editing.

Data availability

Frozen numeric inputs and historical summaries are publicly archived in OFFREF-REG v1.0.1 at https://zenodo.org/records/23366652 and https://github.com/seefreewind/offref-reg/releases/tag/v1.0.1 (OFFREF-REG 2026). This archive predates the subsequent computational audits and the present supplementary package. Online Resources 2 and 3 provide the publication tables and later source-level audit records used here. HGSVC3 2024 v1.0, GNRS accessions PRJCA007976, GVM000324 and GWHBHSK00000000, and eQTL Catalogue release 7 QTD000110 are identified in the provenance inventory. Original third-party sequences, donor-level genotypes/expression, annotation databases, credentials and model weights are available through their providers under the applicable access conditions. AlphaGenome-derived scores, summaries and figures remain subject to the AlphaGenome Output Terms of Use. Users agree to those terms by using this information: https://deepmind.google.com/science/alphagenome/output-terms (Google DeepMind n.d.). Modifications comprise homologous-exon integration, paired-allele differencing, track harmonization, statistical summaries and visualization. Original-code licensing applies alongside upstream output and data conditions.

Code availability

Original analysis and figure code are available under MIT in OFFREF-REG release v1.0.1 (OFFREF-REG 2026). Source scripts and manifests for the present figures and later computational analyses are supplied in Online Resource 3. The historical public archive contains frozen numeric inputs; complete upstream datasets and third-party models remain available through their respective providers. Reproducing inference requires official dependencies, model access and separately obtained upstream inputs. The strict nonmatched comparison retains the original bootstrap specification. Matched reference-SV inference and external calibration stopped at their eligibility gates and remained unexecuted.

Ethics approval

The Ethics Committee of The Second Affiliated Hospital of Wenzhou Medical University determined that this secondary analysis of public genomic and expression-association resources was exempt from ethics review. No ethics approval number was issued.

Clinical trial number: not applicable.

Consent to participate

This secondary study did not recruit participants or collect new human samples. The institutional ethics determination described above covered the secondary public-data analysis. Consent and data-use conditions for the original cohorts remain those established by their source studies.

Consent for publication

Not applicable. This study reports secondary analyses of public resources and does not include identifiable individual participant information.

Acknowledgements

None.

References

Alfisi I, Ciapi F, Baragli M, Magi A (2026) Benchmarking DNA foundation models for zero-shot variant effect prediction shows the importance of context, training, and architecture. Genome Biol. https://doi.org/10.1186/s13059-026-04238-0

Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T et al. (2026) Advancing regulatory variant effect prediction with AlphaGenome. Nature 649:1206-1218. https://doi.org/10.1038/s41586-025-10014-0

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 57:289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573-580. https://doi.org/10.1093/nar/27.2.573

Drusinsky S, Pollard KS (2026) Causal variant underestimation is a major overlooked driver of sequence-to-function model underperformance. bioRxiv [Preprint; not peer reviewed]. https://doi.org/10.64898/2026.09.08.750172

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ et al. (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372:eabf7117. https://doi.org/10.1126/science.abf7117

Fudenberg G, Kelley DR, Pollard KS (2020) Predicting 3D genome folding from DNA sequence with Akita. Nat Methods 17:1111-1117. https://doi.org/10.1038/s41592-020-0958-x

Google DeepMind (n.d.) AlphaGenome Output Terms of Use. https://deepmind.google.com/science/alphagenome/output-terms. Accessed 28 October 2026

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D et al. (2020) Array programming with NumPy. Nature 585:357-362. https://doi.org/10.1038/s41586-020-2649-2

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res 22:1760-1774. https://doi.org/10.1101/gr.135350.111

Huang C, Shuai RW, Baokar P, Chung R, Rastogi R, Kathail P et al. (2023) Personal transcriptome variation is poorly explained by current genomic deep learning models. Nat Genet 55:2056-2059. https://doi.org/10.1038/s41588-023-01574-w

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. Comput Sci Eng 9:90-95. https://doi.org/10.1109/mcse.2007.55

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell 176:535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L et al. (2021) A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet 53:1290-1299. https://doi.org/10.1038/s41588-021-00924-w

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506-511. https://doi.org/10.1038/nature12531

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G et al. (2023) A draft human pangenome reference. Nature 617:312-324. https://doi.org/10.1038/s41586-023-05896-x

Lin J, Gustafson JA, Wertz J, Sui Y, Yoo D, Porubsky D et al. (2026) A high-resolution human pangenome structural variant resource for improved disease association. medRxiv [Preprint; not peer reviewed]. https://doi.org/10.64898/2026.08.21.26361050

Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR (2025) Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nat Genet 57:949-961. https://doi.org/10.1038/s41588-024-02053-6

Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J et al. (2025) Complex genetic variation in nearly complete human genomes. Nature 644:430-441. https://doi.org/10.1038/s41586-025-09140-6

Macias-Velasco JF, Zhuo X, Tomlinson C, Belter EA, Kremitzki M, Albracht D et al. (2026) Benchmarking genome choice in functional genomics analyses. Nat Commun 17:6837. https://doi.org/10.1038/s41467-026-73663-3

Nagai M, Murphy AE, Rizzo K, Koo PK (2026) Toward generalizable and interpretable AI in regulatory genomics. Nat Genet 58:1787-1802. https://doi.org/10.1038/s41588-026-02670-3

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A et al. (2022) The complete sequence of a human genome. Science 376:44-53. https://doi.org/10.1126/science.abj6987

OFFREF-REG (2026) OFFREF-REG: frozen sequence-to-function reliability audit. Version 1.0.1. Zenodo. https://zenodo.org/records/23366652. Accessed 28 October 2026

Sasse A, Ng B, Spiro AE, Tasaki S, Bennett DA, Gaiteri C et al. (2023) Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. Nat Genet 55:2060-2064. https://doi.org/10.1038/s41588-023-01524-6

Spiro AE, Tu X, Sheng Y, Sasse A, Hosseini R, Chikina M et al. (2026) A scalable approach to investigating sequence-to-function predictions from personal genomes. Nat Methods 23:1308-1312. https://doi.org/10.1038/s41592-026-03124-8

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12:2. https://doi.org/10.1186/s13100-020-00230-y

Sun X, Mews M, Wheeler NR, Benchek P, Gu T, Gomez L et al. (2026) Evaluating sequence-to-function deep learning models for ancestry-stratified regulatory variant effect prediction using multi-ancestry blood eQTLs. bioRxiv [Preprint; not peer reviewed]. https://doi.org/10.64898/2026.06.22.730889

Toneyan S, Koo PK (2024) Interpreting cis-regulatory interactions from large-scale deep neural networks. Nat Genet 56:2517-2527. https://doi.org/10.1038/s41588-024-01923-3

UCSC Genome Browser (n.d.) GENCODE v20 annotation tables for hg38. https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/wgEncodeGencodeCompV20.txt.gz; https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/wgEncodeGencodeAttrsV20.txt.gz. Accessed 28 October 2026

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261-272. https://doi.org/10.1038/s41592-019-0686-2

Wang Y, Duan Z, Chen D, Shi D, Ding Y, Wang Z et al. (2026) The 1000 Chinese Pangenome empowers medical and population genetics. Nature 654:121-130. https://doi.org/10.1038/s41586-026-10315-y

Wu Z, Li T, Jiang Z, Zheng J, Gu Y, Liu Y et al. (2024) Human pangenome analysis of sequences missing from the reference genome reveals their widespread evolutionary, phenotypic, and functional roles. Nucleic Acids Res 52:2212-2230. https://doi.org/10.1093/nar/gkae086

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource2.xlsx
- OnlineResource3.zip
- OnlineResource1.docx