Independent Evaluation of AlphaGenome

A Comprehensive Research Protocol

Consortium Research Protocol v1.0

Document Information	
Prepared for:	International Statistical and Population Genetics Consortium
Date:	February 2026
Target Publication:	Nature Methods / Nature Genetics / Genome Biology
Version:	1.0

Executive Summary

This protocol outlines a comprehensive, systematic independent evaluation of AlphaGenome, the unified deep learning model for regulatory variant effect prediction published by Google DeepMind (Nature, January 2026). Our consortium of 20 researchers will conduct evaluations across multiple dimensions not fully explored in the original publication.

Key Focus Areas:

- Cross-ancestry generalizability across 5+ global populations
- Novel disease cohort applications and clinical utility
- Comparison with emerging models (Evo 2, Nucleotide Transformer, HyenaDNA)
- Clinical utility assessment for rare disease diagnosis
- Computational reproducibility and uncertainty quantification

Part I: Study Design Framework

1.1 Evaluation Objectives

Objective	Priority	Novel Contribution
Cross-ancestry variant effect prediction	High	Original paper primarily used European ancestry data
Rare disease diagnostic utility	High	Clinical application not systematically evaluated
Comparison with post-publication models	High	Evo 2, Nucleotide Transformer v2, etc.
Long-read sequencing integration	Medium	ONT/PacBio structural variant prediction
Pharmacogenomic variant interpretation	Medium	Drug response prediction
Non-human primate transferability	Medium	Cross-species validation for preclinical
Single-cell context predictions	Medium	Cell-type specificity beyond bulk data
Computational benchmark	High	Runtime, memory, reproducibility

1.2 Consortium Structure (20 Members)

Work Package	Members	Focus Area
WP1: Track Prediction	4	Expression, chromatin, epigenomic track validation
WP2: Variant Effect Prediction	6	eQTL, sQTL, caQTL evaluation
WP3: Cross-Ancestry	4	Multi-ancestry QTL, population-specific variation
WP4: Disease Applications	3	Mendelian disease, complex trait fine-mapping
WP5: Benchmarking	3	Model comparison, computational reproducibility

1.3 Timeline (12 Months)

Phase	Duration	Activities
Phase 1	Months 1-2	Data acquisition, pipeline setup, API access

Phase 2	Months 3-6	Primary evaluations across all WPs
Phase 3	Months 7-9	Secondary analyses, cross-WP integration
Phase 4	Months 10-11	Manuscript preparation
Phase 5	Month 12	Internal review, submission

Part II: Data Resources

2.1 Expression QTL Datasets

Dataset	Ancestry	Tissues	Samples	Access
eQTLGen Phase 2	European	Blood	>30,000	Public
GENOA	African American	Blood, lymphocytes	1,263	dbGaP
MESA	Multi-ethnic	Monocytes	1,264	dbGaP
GTEx v10	Multi-ethnic	54 tissues	~1,000	dbGaP
BioVU eQTL	Multi-ethnic	Multiple	>100,000	Collaboration
Japanese eQTL	East Asian	Blood	10,000+	Collaboration
INTERVAL	European	Blood	3,301	EGA
BLUEPRINT	European	Immune cells	197	EGA

2.2 Biobank GWAS & PheWAS Resources

Resource	Population	Phenotypes	Samples	Key Features
FinnGen R12	Finnish	2,400+	500,000+	Summary stats, fine-mapping, founder enrichment
UK Biobank	European	7,000+	500,000	Exome + imputed, deep phenotyping
AZ PheWAS	Multi-ethnic	1,500+	450,000+	AstraZeneca internal + UKB, rare variants
Genebass	European	4,500+	400,000	Gene-burden tests, exome-wide associations
Biobank Japan	Japanese	200+	200,000+	East Asian GWAS, fine-mapping
All of Us	Multi-ethnic	2,000+	300,000+	Diverse ancestry, WGS available

2.3 Variant Annotation & Population Databases

Resource	Content	Size	Key Use Cases

gnomAD v4	Population frequencies	800K exomes, 76K genomes	AF filtering, constraint (pLI, LOEUF)
ClinVar (2026)	Clinical interpretations	>2.5M submissions	Pathogenic/benign classification
OMIM	Gene-disease relationships	>16,000 genes	Mendelian disease annotation
HGMD Professional	Disease mutations	>350K variants	Literature-curated pathogenic
dbSNP b156	Variant catalog	>1B variants	rsID mapping, validation
ClinGen	Gene/variant curation	>1,500 genes	Expert-curated validity

2.4 Integrated Genetics Platforms

Platform	Data Types	Key Features	Evaluation Use
Open Targets Genetics	GWAS, eQTL, chromatin	L2G scores, coloc, fine-mapping	Variant-to-gene validation
Open Targets Platform	Drug targets, diseases	Target tractability, safety	Clinical relevance scoring
GWAS Catalog	Published associations	>500K associations	Benchmark variant sets
PhenoScanner v2	Cross-phenotype lookup	GWAS + eQTL + pQTL	Pleiotropy analysis
Ensembl VEP	Variant annotation	Consequence prediction	Functional annotation

2.5 Clinical Variant Resources

Resource	Variant Type	Count	Access
ClinVar (2026 release)	Pathogenic/Benign/VUS	>2.5M submissions	Public FTP
OMIM	Gene-phenotype	>16,000 genes	API/License
HGMD Professional	Disease mutations	>350K	License required
Deciphering Developmental Disorders	De novo variants	~10K	EGA
SPARK autism	De novo + inherited	~15K families	SFARI Base
ClinGen Expert Panels	Curated assertions	>20K variants	Public

2.6 GWAS Fine-Mapping Resources

Resource	Traits	Credible Sets	Method
Open Targets Genetics v8	>5,000	>50,000	SuSiE + FINEMAP
FinnGen R12 Fine-mapping	2,400+	>15,000	SuSiE, founder LD
GWAS Catalog fine-mapping	>1,000	>20,000	Multiple methods
Biobank Japan	200+	>5,000	FINEMAP
UKBB (Weissbrod)	94	~2,000	PolyFun + SuSiE
Genebass credible sets	4,500+	Gene-level	Burden test

Part III: Evaluation Strategies

3.1 Database-Specific Evaluation Approaches

Database	Evaluation Strategy	Key Metrics
FinnGen	Finnish founder variant prioritization; enriched rare variants	Fine-mapping PIP correlation
AZ PheWAS	Cross-phenotype pleiotropy; rare variant effects	Effect size prediction
Genebass	Gene-level burden test validation; LoF variant effects	Burden score correlation
Open Targets	L2G score comparison; variant-to-gene validation	L2G vs AG score concordance
gnomAD	Constraint metric integration; rare variant filtering	pLI/LOEUF enrichment
ClinVar	Pathogenic vs benign classification; VUS prioritization	auROC, auPRC
OMIM	Mendelian gene regulatory variant discovery	Sensitivity for known genes

3.2 FinnGen-Specific Analyses

FinnGen provides unique opportunities due to Finnish founder population enrichment for rare variants:

- Founder variant effect prediction: Test on variants enriched 10-100x in Finland
- Fine-mapping validation: Compare AlphaGenome scores to SuSiE PIPs
- Loss-of-function variant prioritization: Predict regulatory LoF effects
- Phenome-wide association: Correlate predicted effects with PheWAS results

3.3 Open Targets Integration

Open Targets Genetics provides variant-to-gene (V2G) assignments that can benchmark AlphaGenome:

- L2G (Locus-to-Gene) score comparison: Does AlphaGenome improve gene assignment?
- Colocalization validation: Compare eQTL-GWAS coloc with AlphaGenome predictions
- Credible set prioritization: Rank variants within Open Targets credible sets
- Drug target validation: Evaluate predictions for therapeutically relevant genes

3.4 gnomAD Constraint Integration

gnomAD constraint metrics provide complementary information to sequence-based predictions:

- Constraint-stratified evaluation: Performance in high-pLI vs low-pLI genes

- LOEUF correlation: Do AlphaGenome regulatory scores correlate with constraint?
- Rare variant allele frequency: Predict effects for gnomAD rare variants
- Regional constraint (RMC): Evaluate within-gene regional differences

3.5 ClinVar Clinical Validation

Systematic evaluation on ClinVar variants for clinical utility assessment:

- Pathogenic vs Benign: Non-coding regulatory variant classification (auROC target >0.85)
- VUS prioritization: Rank variants of uncertain significance by predicted effect
- Expert panel concordance: Compare to ClinGen expert-curated assertions
- Submission recency: Evaluate on 2024-2026 submissions (post-training data)

3.6 OMIM Mendelian Disease Analysis

OMIM provides curated gene-disease relationships for Mendelian conditions:

- Regulatory variant discovery: Identify non-coding variants in OMIM disease genes
- Promoter mutation prediction: Evaluate known promoter pathogenic variants
- Enhancer mutation detection: Long-range regulatory variant effects
- Tissue-specificity validation: Match predicted expression to disease manifestation

3.7 Genebass Rare Variant Analysis

Genebass exome-wide associations provide gene-level validation:

- Burden test weighting: Use AlphaGenome scores to weight rare variant tests
- Regulatory burden: Extend burden tests to non-coding regulatory variants
- Splice variant contribution: Validate splicing predictions with exome data
- Gene-level effect correlation: Compare predicted vs observed gene effects

3.8 Track Prediction Evaluation (WP1)

This work package focuses on validating AlphaGenome's ability to predict functional genomic tracks across independent datasets not used in training. Key analyses include:

- ENCODE4 RNA-seq validation on new cell types
- Single-cell expression prediction using Human Cell Atlas pseudo-bulk
- Condition-specific expression (stimulus-response, disease vs control)
- ENCODE4 chromatin accessibility (ATAC-seq, DNase-seq)
- CUT&Tag; histone modification data (higher resolution than ChIP)
- Micro-C 3D chromatin architecture (higher resolution than Hi-C)

3.9 Variant Effect Prediction (WP2)

Systematic benchmarking of eQTL, sQTL, and chromatin QTL effect prediction across ancestries and tissues. Primary metrics include:

Metric	Use Case	Interpretation
Sign auROC	Direction prediction	Discrimination of up vs down regulation
Spearman ρ	Effect size correlation	Rank correlation with observed effects
Causality auPRC	Causal variant prioritization	Precision-recall for fine-mapping

3.10 Cross-Ancestry Evaluation (WP3)

A critical evaluation gap: the original paper primarily used European ancestry data. Our consortium will systematically evaluate performance across:

Population	Data Source	Sample Size	Specific Tests
European (EUR)	GTEx, INTERVAL	>30,000	Baseline comparison
African (AFR)	GENOA, MESA	>2,000	AFR-specific eQTLs
East Asian (EAS)	Biobank Japan	>10,000	EAS-specific variants
South Asian (SAS)	UK Biobank	>5,000	SAS regulatory elements
Hispanic (AMR)	MESA	>1,000	Admixed population effects

Part IV: Novel Evaluation Strategies

These analyses represent unique contributions not covered in the original publication:

Strategy	Description	Novelty
Therapeutic Oligo Design	siRNA/ASO target site validation	First therapeutic application test
Allele-Specific Expression	Within-individual ASE validation	Controls for trans-effects
Cancer Somatic Mutations	Non-coding driver identification	TCGA/PCAWG integration
CRISPR Tiling Screens	Base-pair resolution validation	Gasperini/Fulco datasets
Personal Genome Prediction	Individual-specific predictions	Not evaluated in paper
Developmental Dynamics	Fetal vs adult regulation	Training data gap
Uncertainty Quantification	Calibration analysis	Clinical reliability
Adversarial Testing	Systematic failure modes	Edge case identification
Foundation Model Comparison	Evo 2, NT, HyenaDNA	Emerging alternatives
Rare Variant Burden	Weighted association tests	GWAS integration
Multi-Omics Consistency	Cross-modality validation	Internal coherence
Cross-Species Transfer	NHP/mouse evaluation	Preclinical relevance

Part V: Systematic Model Comparison

5.1 Models to Evaluate

Model	Type	Input Length	Outputs
AlphaGenome	Multi-modal	1 Mb	7,000+ tracks
Borzoi	Multi-modal	500 kb	7,000+ tracks
Enformer	Multi-modal	200 kb	5,000+ tracks
Evo 2	Foundation	1 Mb	Embeddings
Nucleotide Transformer v2	Foundation	12 kb	Embeddings
HyenaDNA	Foundation	1 Mb	Embeddings
DNABERT-2	Foundation	512 bp	Embeddings
SpliceAI	Splicing	10 kb	Splice scores
Pangolin	Splicing	5 kb	Splice scores
ChromBPNet	Accessibility	2 kb	Chromatin

5.2 Comparison Framework

- Head-to-head track prediction on standardized test intervals
- Variant effect prediction with multiple scoring strategies per model
- Computational efficiency (inference time, memory, GPU requirements)
- Ensemble approaches: when does combining models improve performance?
- Statistical significance testing with paired tests and bootstrap CIs

Part VI: Deliverables

6.1 Primary Manuscript

Target Journal: Nature Methods or Genome Biology

Figure	Content
Figure 1	Study design and data overview
Figure 2	Track prediction benchmarks on independent data
Figure 3	eQTL/sQTL prediction performance by ancestry
Figure 4	Clinical variant classification
Figure 5	Model comparison across tasks
Figure 6	Novel application results (therapeutic, personal genomes)
Figure 7	Computational benchmarks and failure modes

6.2 Supplementary Resources

- Complete benchmark results tables and per-variant predictions
- Code repository (GitHub) with all evaluation scripts
- Standardized benchmark datasets for community use
- Interactive web portal for exploring results
- Preprint on bioRxiv upon completion

Part VII: Budget Estimate

Category	Item	Estimated Cost
Computational	Cloud GPU (inference)	$15,000
Computational	Cloud storage	$2,000
Computational	API costs (if applicable)	$5,000
Data	Data access fees	$3,000
Publication	Publication costs (OA)	$5,000
Travel	Conference presentation	$5,000
Subtotal	Direct costs	$35,000
Personnel	In-kind contribution	$325,000
Total		$360,000

Part VIII: Risk Assessment

Risk	Likelihood	Impact	Mitigation
API access limitations	Medium	High	Request research access; local deployment
Computational costs	Medium	Medium	Cloud credits; optimize batch processing
Dataset access delays	Medium	Medium	Start applications early; backup datasets
Model updates during study	Low	High	Version lock; document any changes
Negative results	Medium	Low	Pre-register; frame as informative

Success Metrics

1. 10+ independent datasets evaluated across all work packages
2. 5+ ancestry groups analyzed with statistical comparison
3. Head-to-head comparison with 8+ competing models
4. Clinical utility quantified (auROC for pathogenic variant classification)
5. Computational reproducibility documented with version-locked environment
6. Actionable recommendations for practitioners published
7. Code and data released under open-source license for community use
8. Manuscript submitted to high-impact journal within 12 months

---

Document generated: February 05, 2026

Version 1.0 | For Consortium Internal Use