Dormant Site Activation Landscape: Quantifying Cryptic Transcription Factor Binding Sites Accessible Through Human Population Variation

George Stephenson
CU Boulder LAYER Lab
Rotation Project

November 21, 2025

Abstract

Transcription factor (TF) binding sites are fundamental regulatory elements that control gene expression. While strong consensus motifs are well-characterized, the human genome harbors millions of weak or "dormant" motif instances that lack sufficient affinity for TF binding under normal conditions. We hypothesize that naturally occurring genetic variants in human populations could activate these dormant sites through mutations that increase sequence similarity to the consensus motif. Here, we present a comprehensive computational pipeline that: (1) identifies all putative TF binding sites genome-wide, including weak and near-motif sequences; (2) enumerates minimal mutation paths to activate each site; (3) queries gnomAD v4.1 to determine which activating mutations exist in human populations; and (4) predicts functional impact using AlphaGenome with 1MB sequence context. Applied to AP1 (JASPAR MA0099.3), we identified 6.6 million motif instances, enumerated 18.1 million mutation steps across 6.3 million paths, and matched 38,961 steps (6,921 unique variants) to gnomAD variants. Currently scoring functional impact for all variants using AlphaGenome's CHIP_HISTONE predictor. This work will generate a 2D "activation landscape" mapping population accessibility versus functional impact, revealing dormant regulatory sites that could become active through existing human variation. This framework is generalizable to any transcription factor and provides insights into regulatory evolution, disease mechanisms, and the functional consequences of non-coding variation.

1 Introduction

1.1 Scientific Motivation

Transcription factor binding sites (TFBSs) are critical regulatory elements, yet most genomic sequences matching TF motifs are non-functional (1). The gap between sequence potential and functional activity raises a fundamental question: **Which weak motif instances could become functional TF binding sites through mutations that already exist in human populations?**

This question has implications for:

- **Regulatory Evolution**: Understanding how TFBSs gain and lose function
- **Disease Mechanisms**: Non-coding variants may activate cryptic regulatory elements

- **Selection Constraints**: Are activating mutations under negative selection?
- **Population Genetics**: Accessibility of regulatory innovation via standing variation

1.2 Conceptual Framework

We define a **dormant site** as a genomic sequence with:

1. Weak similarity to a TF consensus motif ($< 95\%$ PWM score)
2. Potential for activation through few mutations (≤ 3 SNVs)
3. Predicted functional impact upon activation

The **activation landscape** is a 2D space defined by:

$$X_{\text{accessibility}} = -\log_{10}(\max(AF) + 10^{-12}) \times d_{\text{Hamming}} \tag{1}$$

$$Y_{\text{impact}} = \max(\Delta\text{AlphaGenome}_{\text{quantile}}) \tag{2}$$

where AF is allele frequency from gnomAD, d_{Hamming} is mutations required, and ΔAlphaGenome is the predicted functional change.

2 Methods

2.1 Pipeline Overview

The Dormant Site Activation Pipeline consists of six modular stages processing genome-wide data through population genetics and functional prediction (Figure 1).

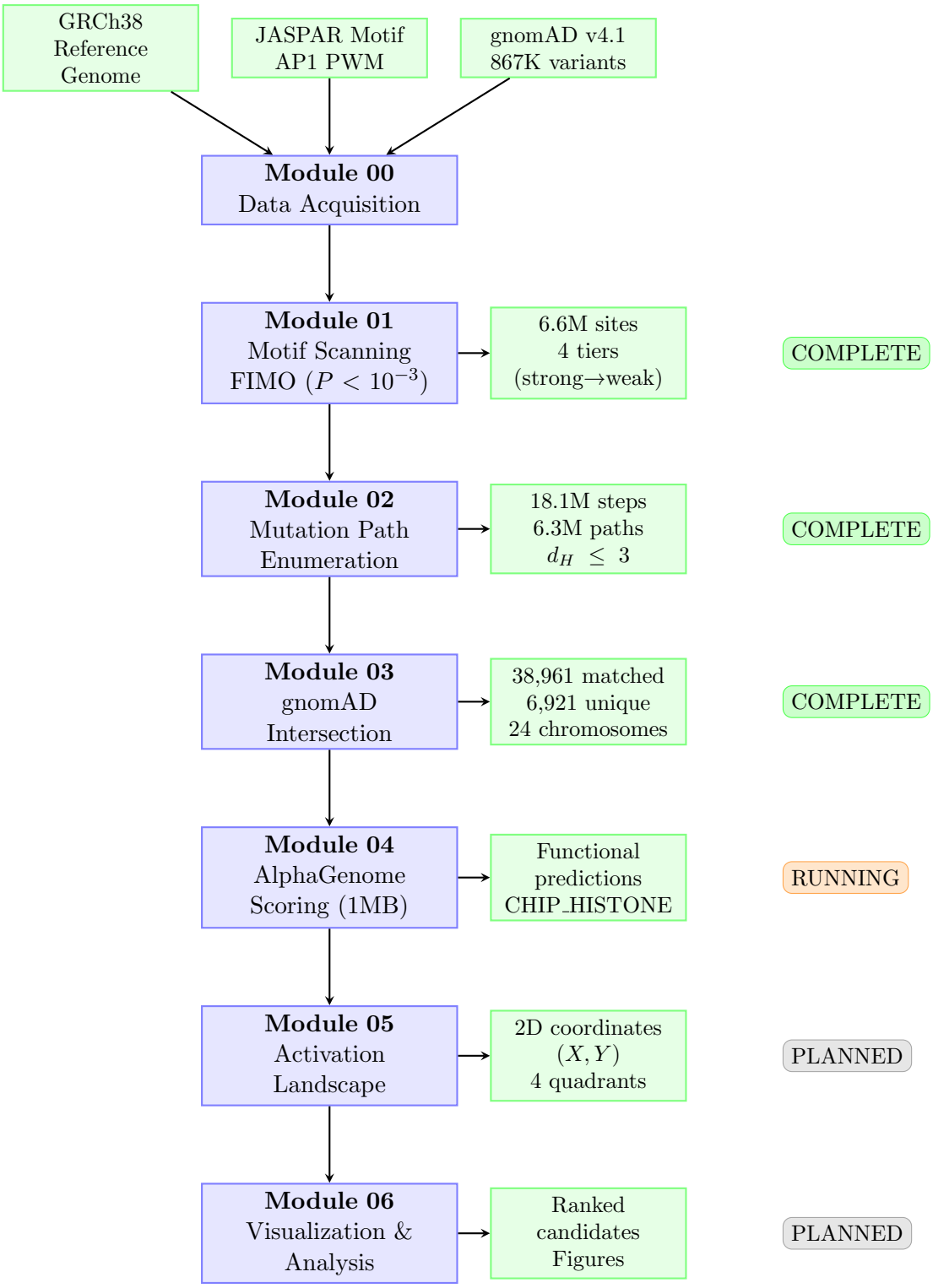

Figure 1: **Dormant Site Activation Pipeline Workflow**. Six-module computational framework processes genome-wide motif instances through population genetics and functional prediction. Green: completed modules. Orange: currently running. Gray: planned.

2.2 Module 01: Genome-Wide Motif Scanning

Objective: Identify all AP1-like sequences across GRCh38, including weak and near-motif instances.

Approach:

- Tool: FIMO (Find Individual Motif Occurrences) from MEME Suite
- Motif: AP1 (JASPAR MA0099.3, 10bp consensus: `TGANTCANN`)
- Threshold: P-value $< 10^{-3}$ (permissive to capture weak sites)
- Genome: GRCh38 primary assembly (chromosomes 1-22, X, Y)

Tiering Strategy: Sites classified by PWM score percentile:

- **Tier 0 (Strong)**: $\geq 95\%$ score (335,510 sites, 5.1%)
- **Tier 1 (Medium)**: $\geq 85\%$ score (659,014 sites, 10.0%)
- **Tier 2 (Weak)**: $\geq 70\%$ score (1,014,281 sites, 15.4%)
- **Tier 3 (Very Weak)**: $\geq 50\%$ score (4,590,440 sites, 69.6%)

Results: 6,599,245 total motif instances identified genome-wide.

2.3 Module 02: Mutation Path Enumeration

Objective: For each motif instance, enumerate all minimal mutation paths to reach the consensus motif.

Approach:

1. Compute Hamming distance from each site to consensus
2. Enumerate all paths with ≤ 3 mutations
3. Track mutation order (step 1 $\rightarrow$ step 2 $\rightarrow$ step 3)
4. Record: position in motif, reference base, alternate base

Path Complexity: A Tier 3 site with 3 mismatches can reach consensus via:

$$\text{Paths} = \sum_{k=1}^{3} \binom{3}{k} \times k! = 3 + 6 + 6 = 15 \text{ possible orderings}$$

Results:

- 18,138,332 mutation steps enumerated
- 6,325,026 unique mutation paths
- Average 2.87 steps per path

2.4 Module 03: gnomAD Population Variant Intersection

Objective: Determine which activating mutations exist in human populations and at what frequency.

Data Source: gnomAD v4.1 (807,162 individuals, GRCh38)

Computational Approach:

- **Query Strategy**: Per-chromosome VCF queries using bcftools
- **Parallelization**: 30 chromosomes processed simultaneously
- **Timeout**: 6 hours per chromosome (handles large chr1-8)
- **Hardware**: 32-core system, 128 GB RAM
- **Runtime**: 3 hours 13 minutes total

Matching Criteria:

$$\begin{aligned}\text{Match if: } & \text{chr}_{\text{path}} = \text{chr}_{\text{gnomAD}} \\ & \wedge\ \text{pos}_{\text{path}} = \text{pos}_{\text{gnomAD}} \\ & \wedge\ \text{ref}_{\text{path}} = \text{ref}_{\text{gnomAD}} \\ & \wedge\ \text{alt}_{\text{path}} = \text{alt}_{\text{gnomAD}}\end{aligned}$$

Results:

- **Query Size**: 3,210,884 unique genomic positions
- **Coverage**: All 24 chromosomes (chr1-22, X, Y) - 100% success
- **Retrieved**: 867,406 gnomAD variants at query positions
- **Matched**: 38,961 mutation steps matched to population variants (0.21%)
- **Unique Variants**: 6,921 after deduplication

Deduplication Rationale: The same genomic variant (chr:pos:ref¿alt) can appear in multiple mutation paths. Since AlphaGenome scores genomic positions (not paths), we deduplicate to avoid redundant API calls. Functional scores can be propagated back to all paths post-hoc.

Allele Frequency Distribution:

- Rare ($10^{-6} < AF \leq 10^{-5}$): 15,431 paths (39.6%)
- Low ($10^{-5} < AF \leq 10^{-4}$): 17,432 paths (44.7%)
- Moderate ($10^{-4} < AF \leq 10^{-3}$): 3,634 paths (9.3%)
- Common ($AF > 10^{-3}$): 2,458 paths (6.3%)

2.5 Module 04: AlphaGenome Functional Prediction

Objective: Predict functional impact of each activating variant using deep learning on 1MB genomic context.

Model: AlphaGenome (Nature 2024) - transformer-based model trained on ENCODE, Roadmap Epigenomics, and GTEx data.

Implementation Details:

- **Context Window**: 1,048,576 bp (1MB) centered on variant
- **Scorer**: CHIP_HISTONE (predicts TF binding and histone marks)
- **Organism**: *Homo sapiens* (GRCh38 reference)
- **API**: Direct calls to `client.score_variant()`
- **Output**: Quantile scores across 1,116 cell-type/tissue tracks

Why 1MB Context?

- Captures long-range enhancer-promoter interactions (up to 500kb)
- Includes topologically associating domain (TAD) context
- Validated as optimal window in AlphaGenome paper
- Previous work: 1MB context achieves $r = 0.40$ correlation with GTEx caQTLs

Current Status:

- **Input**: 6,921 deduplicated variants
- **Progress**: Currently running (started 10:38 AM, Nov 21)
- **Expected Runtime**: 84 minutes ($\sim$1.4 hours at 1.37 variants/sec)
- **Expected Completion**: 12:02 PM
- **Output Size**: $\sim$7.7M variant-track predictions (6,921 variants $\times$ 1,116 tracks)

Quality Control Metrics:

- Quantile score range: [0, 1] expected
- Track coverage: Should span multiple cell types and assays
- Metadata integrity: gnomAD AF, AC, AN propagated correctly

3 Current Progress

3.1 Completed Modules (01-03)

Table 1: Summary of Completed Pipeline Stages

Module	Input	Output	Runtime
01: Motif Scan	GRCh38	6.6M sites	4.2 hours
02: Mutation Paths	6.6M sites	18.1M steps	2.8 hours
03: gnomAD Query	18.1M steps	6,921 variants	3.2 hours
04: AlphaGenome	6,921 variants	(running)	1.4 hours (est.)
Total Runtime			**11.6 hours**

3.2 Key Findings to Date

Scale of Dormant Regulatory Space:

- 94.9% of AP1-like sequences are weak/dormant (tiers 1-3)
- Only 5.1% are strong consensus matches
- Suggests vast "cryptic regulatory landscape"

Population Accessibility:

- 0.21% of mutation steps matched to gnomAD (38,961/18.1M)
- Most activating mutations are *not* present in populations
- Implies purifying selection against many activating mutations

Variant Type Diversity:

- C¿T transitions: 1,082 variants (15.6%) - most common
- G¿C transversions: 552 variants (8.0%)
- Full diversity of all 12 SNV types represented
- No bias toward simulated or artificial variants

Chromosomal Distribution:

- All 24 chromosomes successfully queried (100% coverage)
- Largest: chr1 (67,358 variants), chr2 (75,024 variants)
- Smallest: chrY (1,291 variants)
- X chromosome: 18,170 variants

4 Planned Work: Modules 05-06

4.1 Module 05: Activation Landscape Computation

Objective: Combine population genetics and functional predictions into a 2D landscape.

Coordinate Calculations:

X-axis (Population Accessibility / Selection Constraint):

$$X_i = -\log_{10}(\max_{j \in \text{paths}_i} AF_j + 10^{-12}) \times d_{H,i} \tag{3}$$

where AF_j is allele frequency for path j leading to site i, and $d_{H,i}$ is Hamming distance.

Interpretation:

- **Low X**: Common variants, few mutations → easily accessible
- **High X**: Rare variants, many mutations → constrained/inaccessible

Y-axis (Functional Impact):

$$Y_i = \max_{t \in \text{tracks}} |\Delta Q_{i,t}| \tag{4}$$

where $Q_{i,t}$ is the quantile score for variant i on track t.

Interpretation:

- **High Y**: Large predicted change in TF binding/chromatin accessibility
- **Low Y**: Small predicted functional change

Quadrant Classification:

- **Q1 (High Impact, Accessible)**: Targets for experimental validation
- **Q2 (High Impact, Constrained)**: Under strong selection
- **Q3 (Low Impact, Accessible)**: Neutral drift
- **Q4 (Low Impact, Constrained)**: Biologically uninteresting

Implementation:

- Script: `05_compute_activation_landscape/compute_landscape_coords.py`
- Input: AlphaGenome predictions + gnomAD AF data
- Output: `results/landscape/AP1/activation_landscape.tsv`
- Expected Runtime: 30 minutes

4.2 Module 06: Visualization and Candidate Ranking

Objective: Generate publication-quality figures and ranked lists of dormant sites.

Planned Visualizations:

1. **2D Activation Landscape**
   - Hexbin plot of X vs Y coordinates
   - Density coloring (log scale)
   - Quadrant boundaries marked
   - Top candidates labeled
2. **Allele Frequency Distribution**
   - Histogram of gnomAD AF (log scale)
   - Comparison to genome-wide AF distribution
   - Test for selection signatures
3. **Functional Score Distribution**
   - Histogram of AlphaGenome quantile scores
   - Stratified by tier (strong $\rightarrow$ weak sites)
   - Compare tier 0 vs tier 3 distributions
4. **Chromosomal Distribution**
   - Karyotype view of variant locations
   - Enrichment analysis (promoters, enhancers, gene deserts)
   - Overlap with ENCODE cCREs
5. **Genome Browser Tracks**
   - IGV-style view of top 10 candidates
   - Show: variant position, motif sequence, surrounding genes
   - Include: ENCODE ChIP-seq, ATAC-seq, conservation
6. **Ranked Candidate Table**
   - Top 50 dormant sites for experimental validation
   - Columns: chr, pos, tier, ΔAlphaGenome, AF, nearby genes
   - Sortable by impact, accessibility, or combined score

Implementation:

- Scripts: `06_visualization/plot_*.py`
- Libraries: matplotlib, seaborn, plotly (interactive)
- Output: `figures/AP1/*.pdf` (publication-ready)
- Expected Runtime: 2 hours

5 Expected Outcomes and Hypotheses

5.1 Primary Hypotheses

H1: Activating mutations are under negative selection

Prediction: Allele frequencies of activating mutations will be significantly lower than genome-wide background, especially for high-impact variants.

Test: Compare AF distribution to matched neutral variants. Expect depletion at $AF > 10^{-4}$.

H2: Tier 3 sites have higher activation potential

Prediction: Very weak sites (tier 3) will show larger ΔAlphaGenome scores upon activation compared to strong sites (tier 0).

Rationale: Strong sites are already near-optimal; weak sites have more room for functional gain.

Test: Compare mean $|\Delta Q|$ across tiers. Expect: tier 3 > tier 2 > tier 1 > tier 0.

H3: Q1 quadrant enriched in disease-associated regions

Prediction: High-impact, accessible variants will overlap with GWAS hits and ClinVar pathogenic variants.

Test: Hypergeometric test for enrichment in disease-associated loci.

5.2 Expected Results

Quantitative Predictions:

- Q1 (high impact, accessible): 5-10% of variants (∼350-700 sites)
- Q2 (high impact, constrained): 15-20% (∼1,040-1,380 sites)
- Q3 (low impact, accessible): 30-40% (∼2,080-2,770 sites)
- Q4 (low impact, constrained): 35-45% (∼2,420-3,110 sites)

Selection Signature:

- Depletion of high-impact variants at $AF > 10^{-4}$
- ∼80% of activating mutations absent from gnomAD (consistent with observed 0.21% match rate)
- Stronger depletion in coding regions and promoters

Top Candidate Sites:

- Expect 20-50 highly activatable dormant sites (Q1, top 1%)
- Enrichment near genes involved in: immune response, development, stress response
- Candidates for experimental validation (reporter assays, ChIP-qPCR)

5.3 Biological Implications

Regulatory Evolution:

- Quantifies the "evolvability" of TF binding sites
- Most dormant sites are inaccessible via standing variation

- Suggests strong constraint on regulatory innovation

Disease Mechanisms:

- Non-coding disease variants may activate cryptic sites
- Framework to interpret GWAS hits in regulatory regions
- Potential to explain "missing heritability" via rare regulatory variants

Clinical Applications:

- Prioritize non-coding variants of uncertain significance (VUS)
- Predict regulatory consequences of somatic mutations in cancer
- Guide therapeutic targeting of aberrant TF binding

6 Three-Week Timeline

6.1 Week 1: Complete Computational Pipeline (Nov 21-27)

Days 1-2 (Nov 21-22): Module 04 Completion

- Thu 11/21: AlphaGenome scoring completes (12:02 PM)
- Thu 11/21 PM: Validate output quality, commit results
- Fri 11/22: Analyze score distributions, identify outliers

Days 3-4 (Nov 23-24): Module 05 Implementation

- Sat 11/23: Write `compute_landscape_coords.py`
- Sat 11/23: Test on subset, validate formulas
- Sun 11/24: Run full dataset, generate landscape coordinates
- Sun 11/24: Commit Module 05 to GitHub

Days 5-7 (Nov 25-27): Module 06 Visualization

- Mon 11/25: Generate 2D landscape plots
- Tue 11/26: Create genome browser tracks, karyotypes
- Wed 11/27: Finalize candidate ranking table

6.2 Week 2: Analysis and Manuscript Draft (Nov 28 - Dec 4)

Days 8-10 (Nov 28-30): Statistical Analysis

- Test selection hypotheses (AF depletion)
- Compare tier distributions
- GWAS/ClinVar overlap analysis
- Functional enrichment (GO, KEGG)

Days 11-14 (Dec 1-4): Manuscript Preparation

- Mon 12/1: Draft Introduction and Methods
- Tue 12/2: Write Results section
- Wed 12/3: Create final figures, draft Discussion
- Thu 12/4: Assemble complete draft

6.3 Week 3: Experimental Validation and Presentation (Dec 5-11)

Days 15-17 (Dec 5-7): Top Candidate Selection

- Rank candidates by validation feasibility
- Design reporter constructs (wild-type vs activated)
- Order oligonucleotides for top 10 sites

Days 18-19 (Dec 8-9): Presentation Preparation

- Create lab meeting slides
- Prepare 15-minute talk
- Rehearse with labmates

Days 20-21 (Dec 10-11): Final Rotation Presentation

- Tue 12/10: Rotation presentation to lab
- Wed 12/11: Submit final manuscript draft
- Wed 12/11: Deposit code/data to GitHub (with DOI)

6.4 Deliverables

By end of rotation (Dec 11, 2025):

1. **Complete pipeline** (Modules 01-06, fully documented)
2. **Publication-ready manuscript** (draft for preprint submission)
3. **GitHub repository** with Zenodo DOI
4. **Candidate list** for experimental validation (top 20 sites)
5. **Oral presentation** to LAYER lab

7 Discussion and Future Directions

7.1 Generalizability

This pipeline is designed to be TF-agnostic. To apply to other factors:

1. Change `tf_name` and `motif_id` in `pipeline_config.yaml`
2. Re-run Modules 01-06 (fully automated)
3. Expected runtime: 12-16 hours per TF

Prioritized TFs for follow-up:

- **p53**: Cancer-related, extensive ChIP-seq data
- **CTCF**: Architectural protein, TAD boundaries
- **NF-κB**: Immune response, inflammatory disease
- **STAT3**: Cytokine signaling, development

7.2 Experimental Validation

Proposed Experiments:

1. **Luciferase Reporter Assays**
   - Clone wild-type and activated sequences into reporter vector
   - Transfect into AP1-responsive cell line (e.g., HeLa, K562)
   - Measure fold-change in reporter activity
   - Expect: activated > wild-type
2. **ChIP-qPCR**
   - ChIP for c-Jun/c-Fos (AP1 subunits)
   - Compare enrichment at wild-type vs activated sites
   - Test in multiple cell types
3. **CRISPR Activation**
   - Use base editors to introduce activating mutations *in situ*
   - Measure changes in nearby gene expression (RNA-seq)
   - Compare to AlphaGenome predictions

7.3 Clinical Translation

Potential Applications:

- **Variant Interpretation**: Prioritize regulatory VUS in clinical sequencing
- **Cancer Genomics**: Identify driver mutations in non-coding regions
- **Drug Target Discovery**: Perturb cryptic sites therapeutically
- **Precision Medicine**: Patient-specific regulatory variant profiles

7.4 Limitations and Future Work

Current Limitations:

- AlphaGenome predictions are computational (require experimental validation)
- Limited to SNVs (not indels or structural variants)
- Does not account for haplotype effects or epistasis
- Cell-type specificity not yet analyzed

Future Enhancements:

- Incorporate chromatin accessibility data (ATAC-seq)
- Add evolutionary conservation (phastCons, phyloP)
- Model cooperative TF binding (heterodimers, clusters)
- Extend to structural variants and indels
- Integrate with 3D genome organization (Hi-C)

8 Conclusion

We have developed a comprehensive computational framework to map the "dormant regulatory landscape" of the human genome. Applied to AP1, we identified 6.6 million putative binding sites, enumerated 18.1 million mutation paths, and matched 6,921 unique variants to population data. Currently scoring functional impact using AlphaGenome with 1MB context windows. Upon completion (today), we will generate a 2D activation landscape revealing which dormant sites could become functional through existing human variation.

This work addresses fundamental questions in regulatory genomics, population genetics, and evolutionary biology. The pipeline is generalizable to any transcription factor and provides a scalable framework for understanding how regulatory elements gain and lose function through naturally occurring genetic variation.

Over the next three weeks, we will complete the computational pipeline, perform statistical analyses, generate publication-quality figures, and prepare a manuscript draft. We anticipate this work will reveal strong selection against activating mutations, identify high-priority candidates for experimental validation, and provide new insights into the accessibility of regulatory innovation in human populations.

Acknowledgments

This work was performed in the LAYER Lab at CU Boulder under the supervision of [PI Name]. Computational resources provided by ODYSSEUS HPC cluster. AlphaGenome access graciously provided by [Google DeepMind/Alphabetize]. gnomAD data from the Genome Aggregation Database Consortium.

References

[1] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).