Massively Parallel Characterization of Transcriptional Regulatory Elements

Authors: Vikram Agarwal et al.
Journal: Nature
DOI: 10.1038/s41586-024-08430-9
Keywords: cis-regulatory elements (cCREs), lentiMPRA, enhancers, promoters, transcription, gene regulation, machine learning, CNN, transformers

Summary

This study presents a large-scale functional analysis of **cis-regulatory elements (cCREs)** using an improved **lentivirus-based massively parallel reporter assay (lentiMPRA)** across three cell types (HepG2, K562, WTC11). The authors tested over **680,000 sequences** to analyze promoter/enhancer function and train **sequence-based models** for predicting cCRE activity and variant effects.

🔬 Key Findings

1️⃣ Functional Characterization of cCREs

- **41.7% of tested sequences exhibited regulatory activity**.
- **Promoters vs. Enhancers**:
  - Promoters show **strong orientation dependence** and act as **universal "on switches"**.
  - Enhancers have **weaker orientation bias** but **greater tissue specificity**.
- **MPRA measurements correlate with endogenous gene expression** but do not fully explain cell-type-specific regulation.

🤖 AI Models for Predicting cCRE Activity

The study benchmarks **sequence-based deep learning models** against **biochemical models** for predicting enhancer/promoter activity.

Models Tested

Model	Type	Architecture	Key Features	Performance (Pearson r)
MPRALegNet	CNN	EfficientNetV2-inspired	Optimized conv layers, pooling	**0.83**
MPRAnn	CNN	Standard CNN	Baseline sequence model	0.79
EnformerMPRA	CNN + Transformer	Enformer backbone	Uses 5,313 biochemical features + regression	0.81
SeiMPRA	CNN + Transformer	Sei framework	21,907 biochemical features + regression	**0.87**
Biochemical Model	LASSO Regression	Non-sequence-based	Uses epigenomic features	0.72

Findings

- **Sequence-based models outperform biochemical models**.
- **MPRALegNet (CNN)** achieves the best performance relative to model complexity.
- **Transformer-based models (EnformerMPRA, SeiMPRA) achieve the best overall prediction power** but are **computationally expensive**.
- The **best models reach experimental reproducibility levels**, meaning further improvement would require more training data.

🔍 Predicting Variant Effects & Fine-Mapping

1️⃣ Model Predictions of Regulatory Variants

- **Tested on GWAS SNPs and allele-specific regulatory variants**.
- **MPRALegNet and EnformerMPRA successfully predict enhancer disruptions**.
- **Key predicted loss-of-function (LOF) SNPs:**
  - **RBM38 (rs2426715, rs376911010, rs737092)**
  - **LMO2 enhancer (rs75395676)**
- **Performance on allele-specific binding (ASB) datasets:**
  - **EnformerMPRA and MPRALegNet predictions align well with ChIP–seq and ATAC-seq validated ASVs.**
  - **Odds ratio > 2.1 in all tested cases**, indicating strong enrichment for true regulatory variants.

2️⃣ Variant Effect Predictions vs. Experimental Saturation Mutagenesis

- **PKLR enhancer MPRA dataset** used to validate predictions.
- **MPRALegNet identified key TF binding sites** and predicted **effect sizes of mutations**.
- Correlation with **real MPRA data**:
  - **PKLR (K562): r = 0.65**
  - **SORT1 (HepG2): r = 0.49**
  - **LDLR (HepG2): r = 0.66**
  - **F9 (HepG2): r = 0.51**
- **Similar performance to Enformer**, but MPRALegNet is **200x smaller in size**, making it more efficient for genome-wide inference.

🎯 Key Machine Learning Insights

1️⃣ Universal vs. Cell-Specific TF Motifs

- **Universal motifs (found in all cell types)**:
  - **KLF-related**
  - **ETS-related**
  - **CTCF (context-dependent activation/repression)**
- **Cell-specific motifs**:
  - **HepG2**: **HNF4A/G** (hepatic function)
  - **K562**: **GATA-TAL1 dimer** (hematopoietic regulation)
  - **WTC11**: **POU5F1-SOX2** (pluripotency)

2️⃣ Combinatorial Effects of TF Binding Sites

- **Homotypic TFBS effects**:

    - Most TFs follow **log-additive activation patterns**.
    - **Some TFs (e.g., STAT, ETS) show saturation effects at high binding site dosages.**
- **Heterotypic TFBS interactions**:
    - **Super-multiplicative activation**: ATF3/FOS–JUN + FOXD2
    - **Sub-multiplicative repression**: HNF4A/G + NFYA/C
    - **MPRALegNet accurately models these interactions**.

3️⃣ Cross-Cell-Type Generalization

- **MPRALegNet was trained on only 3 cell types but generalizes well**.
- **Supports the use of cell-type-agnostic models for variant effect predictions.**
- **Enformer and SeiMPRA outperform on fine-mapping tasks**, but **MPRALegNet is more efficient**.

🏆 Implications

- **High-throughput lentiMPRA is a scalable method to functionally map regulatory elements.**
- **Deep learning models enable accurate, genome-wide enhancer/promoter activity prediction.**
- **MPRALegNet is an efficient alternative to transformer-based models for regulatory genomics.**
- **Fine-mapping efforts for GWAS hits can benefit from these models to prioritize causal variants.**
- **Future work should integrate single-cell epigenomics and train on primary tissues.**

📌 Next Steps

- **Expand to more cell types & disease-relevant tissues.**
- **Integrate MPRA with single-cell chromatin and transcriptomic data.**
- **Validate regulatory SNPs in functional assays (e.g., CRISPR screens).**
- **Develop genome-wide variant effect predictors based on trained models.**

📝 Related Notes

- MPRA Techniques
- Deep Learning in Genomics
- Regulatory Variant Fine-Mapping
- Transformer Models for Functional Genomics
- Transcription Factor Combinatorial Effects

This version provides a **detailed breakdown of the machine learning aspects**, making it useful for **cross-referencing with other AI/genomics notes** in Obsidian. Let me know if you want to **highlight specific aspects further**!