Massively Parallel Systems I

Mike White
Bio5488
4/1/19
Lectures

• MONDAY: Massively Parallel Reporter Assays (MPRAs)
  • Questions of non-coding function
  • MPRA technologies
  • MPRA applications

• WEDNESDAY: Other Massively Parallel Functional Assays
  • Perturb-seq, pooled CRISPR Screens
  • Deep Mutational Protein Scanning
MPRA Learning Goals

- Understand the main varieties of MPRA technology
- Outline of MPRA data processing
- Know some of the creative applications of MPRA to biological questions
Genomics is a technology driven field
Why Massively Parallel Functional Assays

- Because we can (new technologies to produce large libraries of synthetic DNA sequences)
- More comprehensive data (e.g., test binding site mutations in 1000’s of ChIP peaks)
- Direct functional test of genome elements
Question 1: Which genome sequences are enhancers?

Two numbers:
~8% of the Human Genome is Under Selective Constraint


“The vast majority (80.4%) of the human genome participates in at least one biochemical RNA and/or chromatin associated event in at least one cell type.”

Two Numbers

- ~8% of the human genome is under selective constraint (~1.1 % protein coding)
- 80.4% of the human genome is active in an ENCODE assay in 147 cell types

Which genome sequences encode enhancers?
Genome is packed with potential TF binding sites

“In a 1-kb segment of human DNA it is predicted that a new 7-8 bp protein-binding motif arises, by neutral evolution, on average every 60,000 years.”

Characteristics of an Enhancer

Beware of affirming the consequent!

A classic logical fallacy:

Premise: If A then B
Observation: B is true
Conclusion: A is true

Enhancers exhibit bidirectional transcription. This sequence is bidirectionally transcribed. Therefore this sequence is an enhancer.
Bidirectional Transcription
Not Specific to Enhancers

What evidence would convince you that something is an enhancer?
Question 2: Which genetic variants are causal?
What is good evidence that a non-coding variant affects enhancer function?
We need a method to:

- Test the function of tens of thousands of putative enhancer sequences
- Test functional effect of tens of thousands of putative enhancer variants
Reporter genes!

Non-consensus GLI binding sites in Hedgehog target gene regulation.

Winklmayr M^1, Schmid C, Laner-Plammerg S, Kaser A, Aberger F, Eichberger T, Frischauf AM.
Classic Reporter Gene

Minimal promoter

Cis-regulatory element

CRE

DsRed

Reporter gene
The Barcode

- Minimal promoter
- Cis-regulatory element
- Reporter gene
- Unique DNA Barcode (9bp)
- mRNA
Barcoded Reporter Library & MPRA

... pooled library of $10^5$ or more distinct reporters
Two Technical Problems

• Problem 1: Where does your CRE DNA come from?
  • DNA synthesis
  • Genomic fragments
  • Targeted regulome capture

• Problem 2: How do you read out the reporter?
  • Synthetic barcodes
  • Self-transcribing enhancers
  • Sort-seq (fluorescence/flow cytometry + DNA sequencing)
Problem 1: Source of DNA
DNA Synthesis

Barcoded oligonucleotide library

Minimal promoter + Reporter gene

Cloning step 1

Cloning step 2

Final reporter library
Custom Array Oligo Synthesis

Synthesized DNA

- Advantages?
- Disadvantages?
Sheared Genomic DNA (STARR-seq)

Sheared Genomic DNA (STARR-seq)

Sheared Genomic DNA

- Advantages?
- Disadvantages?
Targeted CRE Capture

1) Design biotinylated bait sequences
2) Pull down & clone targets
Problem 2: Measuring Reporter Activity
The Barcode

Minimal promoter

Cis-regulatory element

Reported gene

Unique DNA Barcode (9bp)

mRNA

CRE

DsRed

BC
RNA-seq on Synthetic Barcodes

Array-synthesized library

Transfection into cells, Sequence BCs

Expression

BC RNA
BC DNA
BC RNA
BC DNA
BC RNA
BC DNA
...
CRE serves as its own barcode

FACS & Sequencing

What We’ve Covered So Far

- Need for an assay to directly test function of non-coding sequences & variants at scale.
- The trick is to use barcodes
- Different ways to construct reporter libraries
- Three primary ways of reading out reporter activity
What are MPRA Limitations?

- Lack of native genomic context
- Multiple DNA copy number
- Mainly done in transfectable cell lines which may not recapitulate enhancer biology
- Potential assay toxicity - plasmids, high reporter gene levels
- Compatibility with reporter/minimal promoter: minimal promoter used may lead to false positives, negatives
RNA-seq on Synthetic Barcodes

Array-synthesized library

Transfection into cells, Sequence BCs

Expression

- BC: RNA
- BC: DNA
- BC: RNA
- BC: DNA
- BC: RNA
- BC: DNA
- ...
MPRA Data Analysis

- See answers on next page
Basic MPRA Data Analysis

1. Match sequencing reads to barcodes. (Discard reads that don’t match.)

2. Normalize barcode RNA (cDNA) by barcode plasmid DNA

3. Filter barcodes (or perhaps CREs) by read count

4. Calculate CRE means over all barcodes

5. Compare CREs - for example, one allele vs other allele
Some MPRA Examples
Example 1:
Do enhancer marks predict reporter activity?

- Test of 1200 cell-type specific ENCODE-predicted enhancers in K562 cells
- How would you design this experiment?
26% of predicted enhancers showed activity above scrambled controls. "Weak" enhancers more likely to be active.
Example 2: Which Non-coding Enhancer Variants Affect Function?

- Test putative 2700 CREs & variants near 75 GWAS tag SNPs
- Goal is to find SNPs that alter CRE activity
- How would you design this experiment?

Cell

Volume 165, Issue 6, 2 June 2016, Pages 1530-1545

Resource

Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits

Jacob C. Ulirsch 1, 2, 6, Satish K. Nandakumar 1, 2, 6, Li Wang 2, Felix C. Giani 1, 2, 3, Xiaolan Zhang 2, Peter Rogov 2, Alexandre Melnikov 2, Patrick McDonel 2, Ron Do 4, Tarjei S. Mikkelsen 2, 5, Vijay G. Sankaran 1, 2, 5, 6, 7, 8, 9
2756 loci in high LD with 75 GWAS hits + controls

synthesize constructs

major allele

minor allele

~14 barcodes

~231,000 distinct plasmids

transfect library

erythroid cells

isolate & sequence DNA/RNA

plasmid DNA counts

mRNA counts

192
350
105
890
370
300
Key Results

4% of tested CREs were highly active
Key Results

• 32 out of 2,756 variants in 23 out of 75 GWAS regions showed statistically significant fold change between major and minor alleles.
Example 3: Directly Measure Enhancer Activity Across Genomes

- STARR-seq to produce ‘tracks’ of enhancer activity across small *Drosophila* genome
- How would you design the experiment?


Genome-wide quantitative enhancer activity maps identified by STARR-seq.

Arnold CD¹, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A.
Key Results
Key Results

- 5499 enhancer activity “peaks” in 169 Mb *Drosophila genome*

- 69% of “strong” peaks in DHS sites
Evaluating MRPAs

- DNA source/library construction. What are the controls?
- Reporter design/barcoding scheme
- Assay readout - does it measure what the paper claims?
- Parameters: Library size, cellular pool, sequencing coverage - how many input barcodes were recovered?
- Reproducibility
Article

Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior

Ryan N. Doan 1, Byoung-il Bae 1, 12, Beatriz Cubelos 2, 3, Cindy Chang 1, Amer A. Hossain 1, Samira Al-Saad 4, Nahit M. Mukaddes 5, Ozgur Oner 6, Muna Al-Saffar 1, 7, Soher Balkhy 8, Generoso G. Gascon 9, The Homozygosity Mapping Consortium for Autism, Marta Nieto 3, Christopher A. Walsh 1, 10, 11, 13 # ✏️
HARs are Neural Enhancers

- Identified constrained intergenic HARS
- Enriched for neural TFBS
- ChIA-PET and Hi-C to map 576 HARs to target genes
- De Novo CNVs in Autism cases involve HARs
- Test 383 patient rare HAR mutations in MPRA in mouse neurospheres
Genome-Integrated MPRA

1) Clone into lentiviral or TE vector

2) Make virus & infect/co-transfect with transposase

Genome-Integrated MPRA

What is the effect of genome position on enhancer activity?

**Step 1:** Make library with >20,000 barcodes, single enhancer, in transposable element vector.

**Step 2:** Integrate into mouse ES cells, grow, map integration locations for each barcode. (How do you map?)

**Step 3:** Perform MRPA - barcode RNA/barcode DNA
Key Results

Domains of chromatin effects on the Mb scale (median 1.23 Mb)

1000-fold expression range from SAME promoter!

Domains correlated with Lamina Associated Domains

![Graph showing correlation between genetic expression and distance from LAD boundary.]

- **log$_2$ (IR expression)**
- **Distance from LAD boundary (kb)**
- **Genes log$_2$ (FPKM + 1)**

Legend:
- mPGK
- tet-Off (no Dox)
- Endogenous genes

LADs
inter-LADs
Take aways

- Genome-integrated MRPAs with transposable element/lentiviral vectors
- Genome position effects are HUGE - 1000-fold expression range of same promoter
- Some correlation with 3D genome structure
- What about many CREs at many genome positions?
MPRA for RNA Stability

[Diagram showing GFP, AAA, and 3' UTR with constant and variable sequences]
Assay

Take *in vitro* transcribed reporters and inject into embryos:

Results

- Model decay with **onset** and **decay rate** parameters
- Classify reporters with model

![Graphs showing decay with model](image)
MPRA of miRNA Sites
Key Results

(a) Fold repression

<table>
<thead>
<tr>
<th>Binding site type</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect match</td>
<td>450</td>
</tr>
<tr>
<td>Bulged</td>
<td>452</td>
</tr>
<tr>
<td>3’ complementary</td>
<td>439</td>
</tr>
<tr>
<td>Seed</td>
<td>447</td>
</tr>
<tr>
<td>Controls</td>
<td>316</td>
</tr>
<tr>
<td>None</td>
<td>54</td>
</tr>
</tbody>
</table>

miRNA
mRNA target

Control sequences
Non-modified contexts
A review:

Alternative splicing:

Human accelerated regions, behavior, genetic variation:

Exhaustive mutagenesis of cis-regulatory sequences:

More miRNAs and RNA binding proteins:

Genome position effect: