Massively Parallel versus High Throughput?
Recap of Monday

- Pooled assays of synthetic constructs vs genome-wide measurements of cellular state

- Technologies made feasible by sequence barcodes, low cost of DNA sequencing and synthesis

- No standard experimental or computational methods - wide variety of approaches
Recap of Monday

• Need for an assay to directly test function of non-coding sequences & variants *at scale*.

• The trick is to use barcodes

• Three main ways to construct reporter libraries

• Three primary ways to read out reporter activity
How would you do these experiments?

- RNA-seq for every yeast knockout strain
- ID target genes by deleting 500 enhancers and measuring expression
- Mutagenize ever base pair of a super enhancer & measure effect on reporter gene
- Measure drug sensitivity after knockdown of 500 tumor suppressors
Pooled CRISPR Assays

Why?

- Complement limitations of MPRA
- Combine perturbation with genome-wide expression readout
In Class Answers: Elements needed for a pooled CRISPR Assay

- A library of guide RNAs
- Method to design gRNAs that are specific for your intended targets
- Cell line expressing Cas9 and possibly a genomic ‘landing pad’ for the gRNA library
- A reporter gene or other method (single cell sequencing) to read out the effect of the gRNA
- Means to get 1 gRNA per cell
- A way to link the gRNA identity with the reporter gene readout (usually a barcode)
Example 1: Mapping Regulatory DNA by *In Vivo* Mutagenesis

High-throughput mapping of regulatory DNA


*Nature Biotechnology* **34**, 167–174 (2016)  Download Citation
**Aim:** Identify regulatory sequence for a gene by *in vivo* mutagenesis

**Method:** Identify regulatory sequence for a gene by *in vivo* mutagenesis with 1000’s of gRNAs

What problems have to be solved to make the method work?
Multiplexed Editing Regulatory Assay (MERA)
“Surprisingly, we observed a novel class of genomic elements downstream of Tdgf1 that did not coincide with any known markers of regulatory activity…”

We designated such elements that do not contain any of these markers as unmarked regulatory elements (UREs). UREs were often over 1 kb in length and produced a loss of GFP comparable to that induced by some distant enhancers.”
Perturb-seq

Cell

Volume 167, Issue 7, 15 December 2016, Pages 1853-1866.e17

Resource

Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens

Atray Dixit 1, 2, 9, Oren Parnas 1, 9, 10, Biyu Li 1, Jenny Chen 1, 2, Charles P. Fulco 1, 4, Livnat Jerby-Aronon 1, Nemanja D. Marjanovic 1, 3, Danielle Dionne 1, Tyler Burks 1, Raktima Raychowdhury 1, Britt Adamson 5, Thomas M. Norman 5, Eric S. Lander 1, 4, 6, Jonathan S. Weissman 5, 7, Nir Friedman 1, 8, Aviv Regev 1, 6, 7, 11

See also companion papers:
Cell, Volume 167, Issue 7, 15 December 2016, Pages 1867-1882.e21
Cell, Volume 167, Issue 7, 15 December 2016, Pages 1883-1896.e15
**Aim:** Identify factors in LPS immune response by genetic screen

**Method:** Combine pooled CRISPR library with single cell RNA-seq

What problems have to be solved to make the method work?
<table>
<thead>
<tr>
<th>Cell type</th>
<th>sgRNA pool</th>
<th>Total cells</th>
<th>Time points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse BMDC</td>
<td>Transcription factors (67 guides)</td>
<td>70,000</td>
<td>0 and 3 hr post-LPS</td>
</tr>
<tr>
<td>Human K562</td>
<td>Transcription factors (46 guides)</td>
<td>104,000</td>
<td>7 and 13 days</td>
</tr>
<tr>
<td>Human K562</td>
<td>Cell cycle regulators (36 guides)</td>
<td>26,000</td>
<td>7 days</td>
</tr>
</tbody>
</table>
\[
\log(\text{Cells}) + 1 = \begin{bmatrix}
0 & 8 & 0 & 1 & 0 & 2 & \ldots & 0 \\
1 & 2 & 0 & 0 & 1 & 1 & \ldots & 0 \\
0 & 7 & 0 & 1 & 0 & 0 & \ldots & 1 \\
\vdots \\
0 & 5 & 0 & 1 & 0 & 0 & \ldots & 0 \\
\end{bmatrix}
\begin{bmatrix}
Y \\
\text{Expression matrix}
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 & \ldots & -0.1 & \ldots \\
0 & 1 & 0 & \ldots & 0.3 & \ldots \\
0 & 0 & 1 & \ldots & 0.2 & \ldots \\
\vdots \\
0 & 0 & 0 & \ldots & -0.2 & \ldots \\
\end{bmatrix}
\begin{bmatrix}
X \\
\text{Design matrix}
\end{bmatrix}
= \begin{bmatrix}
\beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \ldots & \beta_{1,G} \\
\beta_{2,1} & \beta_{2,2} & \beta_{2,3} & \ldots & \beta_{2,G} \\
\beta_{3,1} & \beta_{3,2} & \beta_{3,3} & \ldots & \beta_{3,G} \\
\vdots \\
\beta_{C,1} & \beta_{C,2} & \beta_{C,3} & \ldots & \beta_{C,G} \\
\end{bmatrix}
\begin{bmatrix}
\beta \\
\text{Coefficient (regulatory) matrix}
\end{bmatrix}
\begin{bmatrix}
\text{Covariates} \\
\text{Genes}
\end{bmatrix}
\begin{bmatrix}
\text{Signature decomposition} \\
\text{Inference}
\end{bmatrix}
\begin{bmatrix}
\text{Cell features} \\
\text{Design of experiments}
\end{bmatrix}
\begin{bmatrix}
\text{Inference} \\
\text{Interpretation}
\end{bmatrix}
\]
Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation

Daniel R Fuentes, Tomek Swigut, Joanna Wysocka
Stanford University School of Medicine, United States
**Aim:** Identify regulatory impact of a human-specific LTR on gene expression

**Method:** Simultaneously modulate ~700 LTRs with CRISPRi or CRISPRa and CARGO
CRISPRi

Block transcription initiation

Effective for both NT and T strands

Chimeric Array of gRNA Oligos (CARGO)
Epigenetic State of LTRs

![Diagram showing epigenetic states of LTRs with different modifications and conditions.](image-url)
LTR-Regulated Genes in Stem Cells

C. LTR5HS-regulated genes expressed in naïve hESC

D. LTR5HS-regulated genes expressed in human preimplantation epiblast
Perturb-seq + ATAC-seq

Resource

Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks

Adam J. Rubin 1, 9, Kevin R. Parker 1, 2, 9, Ansuman T. Satpathy 1, 3, 9, Yanyan Qi 1, Beijing Wu 1, Alvin J. Ong 1, 2, Maxwell R. Mumbach 1, 2, 4, Andrew L. Ji 1, Daniel S. Kim 1, Seung Woo Cho 1, 2, Brian J. Zarnegar 1, William J. Greenleaf 4, 5, 6, Howard Y. Chang 1, 2, 4, 7 8, Paul A. Khavari 1, 8, 10 9 10
A Perturb-ATAC

Cas9+ cells

Lentiviral sgRNA library

FACS sgRNA+ cells

Single-cell capture

ATAC-seq

RT sgRNA
sgRNA + Cell Barcodes

A Guide Barcode Amplification

- Single Cell
- Puromycin Resistance
- P2A
- mCherry
- 22nt Barcode
- M13 Primer Site
- WPRE
- mRNA → RT
- cDNA → PCR 1
- First Amplicon → PCR 2
- Second Amplicon → PCR 3
- Final Amplicon → Sequencing Read

Lentivirus sgRNA Plasmid

- mCherry
- GBC
- WPRE
- Illumina Adapter
- Cell Index 1
- Read 1 Start Site
- Read 2 Start Site
- Cell Index 2
- Illumina Adapter
Computational Pipeline

Sequence

↓

Align Reads to Guide Barcodes

↓

Count Guide Reads

Guide Counts Table

<table>
<thead>
<tr>
<th></th>
<th>Guide 1</th>
<th>Guide 2</th>
<th>Guide 3</th>
<th>⋮</th>
<th>Guide n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell 1</td>
<td>10000</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
<td>0</td>
</tr>
<tr>
<td>Cell 2</td>
<td>0</td>
<td>7000</td>
<td>4000</td>
<td>⋮</td>
<td>0</td>
</tr>
<tr>
<td>Cell 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
<td>8000</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>Cell n</td>
<td>2000</td>
<td>0</td>
<td>5000</td>
<td>⋮</td>
<td>0</td>
</tr>
</tbody>
</table>

Normalize Sequencing Depth by Plate

↓

Filter Cells

↓

Assign Guide Calls

Guide Calls

<table>
<thead>
<tr>
<th></th>
<th>Guide 1</th>
<th>Guide 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell 1</td>
<td>SPI1</td>
<td>(N/A)</td>
</tr>
<tr>
<td>Cell 2</td>
<td>EZH2</td>
<td>RELA</td>
</tr>
<tr>
<td>Cell 3</td>
<td>EBER2</td>
<td>(N/A)</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>Cell n</td>
<td>SPI1</td>
<td>RELA</td>
</tr>
</tbody>
</table>
Testing Genetic Interactions

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100%</td>
</tr>
<tr>
<td>X</td>
<td>80%</td>
</tr>
<tr>
<td>A</td>
<td>60%</td>
</tr>
<tr>
<td>X</td>
<td>70%</td>
</tr>
</tbody>
</table>
Gene x Gene x Environment interactions
Double CRISPRi Knockdown

- gRNA 2 (from query)
- URA3
- gRNA 1 (from pool)

25 gRNA2 interaction targets
782 gRNA1 targets

gRNA2 barcode
Fitness Assay

1. Sequence original pool of yeast strains
2. Grow (in multiple conditions in parallel)
3. Sequence strain pool after growth
4. Compare input pool to output pool
Key Results

- Only 3.6% of gRNA pairs showed strong condition specific interaction.
- BUT interaction correlations across conditions helped suggest functions for uncharacterized genes.
Coding Sequences
What’s the goal of GWAS?
Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek, Konrad J. Karczewski [...] Exome Aggregation Consortium

No GWAS will ever be larger enough!
Variants of Unknown Significance (VUS)

“[With current cancer panels], the probability of detecting a VUS is higher than the probability of detecting a pathogenic variant.”


Prevalence of pathogenic variants and variants of unknown significance in patients at high risk of breast cancer: A systematic review and meta-analysis of gene-panel data.

van Marcke C¹, Collard A², Vikkula M³, Duhoux FP⁴.
Commentary

Variant Interpretation: Functional Assays to the Rescue

Lea M. Starita 1 ᵃ, ᵉ, Nadav Ahituv ², ³, Maitreya J. Dunham ¹, Jacob O. Kitzman ⁴, ⁵, Frederick P. Roth ⁶, ⁷, ⁸, ⁹, Georg Seelig ¹⁰, ¹¹, Jay Shendure ¹, ¹², Douglas M. Fowler ¹, ¹³ ᵃ, ᵉ
The Current Vision
Caveats?

• Proper cell type?

• Hard to test translocations, large indels, CNVs

• Non-cell autonomous effects

• Interactions, epistasis

• What function do you choose?
Caveats to Functional Assays

- Many proteins have more than one function
- Assays only feasible in cell lines that may be wrong context
- Assays don’t capture epistasis between variants
- Assays don’t necessarily reflect in vivo function
- Can do SNPs, not CNVs
- Noisy data
- Scale of the problem is still too large

Starita, et al., AJHG, Volume 101, Issue 3, 7 September 2017, Pages 315-325
Accurate classification of BRCA1 variants with saturation genome editing

Tested 96.5% of all possible SNVs in 13 (of 23) exons

Found 400 non-functional missense mutations

Found 300 mutations that affect expression
Per RING and BRCT exon:

- HAP1 cells
- Genome editing
- Cell survival

**SNV library + Cas9/gRNA**

**Cas9, multiplex repair**

All possible SNVs

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>gDNA</th>
<th>mRNA</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNV 1</td>
<td>No</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>SNV 2</td>
<td>Yes</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>SNV 3</td>
<td>Yes</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>SNV 4</td>
<td>No</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>
Method

An extended set of yeast-based functional assays accurately identifies human disease mutations

Song Sun,1,2,3,4,5 Fan Yang,1,2,3,4 Guihong Tan,1,2 Michael Costanzo,1,2 Rose Oughtred,6 Jodi Hirschman,6 Chandra L. Theesfeld,6 Pritpal Bansal,1,2,3,4 Nidhi Sahni,7,8,10 Song Yi,7,8 Analyn Yu,1,2,3,4 Tanya Tyagi,1,2,3,4 Cathy Tie,4 David E. Hill,7,8 Marc Vidal,7,8 Brenda J. Andrews,1,2 Charles Boone,1,2 Kara Dolinski,6 and Frederick P. Roth1,2,3,4,7,9
Surrogate genetics platform to identify functional variation

[Diagram showing genotypes and phenotypes with variant 1, variant 2, variant 3, and variant N, and their corresponding phenotypes for individuals 1 to N.]
• 101 disease, 78 non-disease human variants in 22 genes (~200? human genes would be good candidates for this assay)

• Simple assay scoring (growth rate or semi-quantitative estimate of growth on solid media)
Multiplex assessment of protein variant abundance by massively parallel sequencing


Premise: Protein stability is a good functional assay
VAMP-Seq
(Variant abundance by massively parallel sequencing)

1. Create variants by site-saturation mutagenesis (variable oligos with inverse PCR on a WT plasmid template)
2. Variants fused to GFP, sequence barcodes cloned in
3. Integrate library at single locus in HEK293 cells
4. Sort-seq: Sort cells on GFP levels, sequence sorted bins
Always Assess Reproducibility

<table>
<thead>
<tr>
<th>score1</th>
<th>score2</th>
<th>score3</th>
<th>score4</th>
<th>score5</th>
<th>score6</th>
<th>score7</th>
<th>score8</th>
</tr>
</thead>
<tbody>
<tr>
<td>r: 0.637</td>
<td>r: 0.640</td>
<td>r: 0.566</td>
<td>r: 0.459</td>
<td>r: 0.488</td>
<td>r: 0.625</td>
<td>r: 0.654</td>
<td>r: 0.645</td>
</tr>
</tbody>
</table>

PTEN
Key Takeaways

• Simple assays with simple scoring metrics do reasonably well classifying known pathogenic vs benign

• Not enough data yet to assess true clinical usefulness of these assays

• Relatively affordable - synthesis of all AA substitutions of a 300 aa is protein is ~$13,000
Final Thoughts on Massively Parallel Systems

- Increasingly common and important in genomics
- No standard experimental or statistical methods yet - each work should be judged on reproducibility, good experimental design, controls, reasonable statistical arguments
- Single cell-seq data poses analysis challenges
- Anyone can do them!
The Current Vision