Analysis of Variant Effect Data

Mike White
Bio5488
4/7/21
MAVE: Multiplex Assay of Variant Effect
Be able to critically read reports of MAVE analysis and evaluate:

- **Aims of the method** (predict molecular or clinical impact of untested variants, quantify effects of assayed variants from complex data, etc.)

- **What is being quantified or modeled?** What is the output of the model?

- **What data or features are the model trained on?**
Not Today’s Learning Goals

• Details of machine learning/regression/computation

Check out Computational Molecular Biology, Machine Learning, Statistics/Bioinformatics courses
Outline

1. Protein assays
2. MPRAs
Why not just a lookup table?
Why do we need sophisticated models for MAVE data?

Don’t we just measure all variant effects?

- Predict effect of variants in proteins that haven’t yet been assayed
- Assay itself is not a direct readout of phenotype: noise, non-linearities, etc.
- Even the biggest DMS only sparsely samples mutation space - you want to predict more combinations of variants than you can measure
Big Themes of MAVE analysis

- MAVE data is sequencing data, not a direct measure of the function of interest. Models connect data to function.

- To achieve practical goal of predicting variant effects, MAVE approaches aim to understand/model general, \textit{quantitative} sequence-to-function relationships.
Example 1: Predicting Effects of Protein Variants

Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data

Vanessa E. Gray 1, Ronald J. Hause 1, Jens Luebeck 1, Jay Shendure 1,2, Douglas M. Fowler 1,3,4,
Rationale for the approach

1. Existing predictors of variant effect are categorical, not quantitative, and are trained on imperfect clinical annotations.

2. Accurate clinical prediction should be founded on accurate molecular prediction (MAVE data).

3. Large-scale mutagenesis data may reveal general patterns that determine when variants are deleterious.
1. Collect and normalize MAVE data for multiple proteins (10)

2. Annotate each residue with a set of features (chemical properties, conservation, etc.)

3. Train model (decision tree ensemble) to predict variant effect from residue features.

Procedure
Models trained on individual proteins

Global predictor did less well than other models for predicting *clinical* effect, but did well on predicting MAVE results.
Amino acid features used in the model
Takeaways

• Idea is to build good predictor of molecular effects as path to improve clinical prediction.

• Presumes that variant effects can be predicted from amino acid and protein ‘features’.

• Requires MAVE data plus curated features
Example 2: Modeling Assay Effects

MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect

Ammar Tareen, Anna Posfai, William T. Ireland, David M. McCandlish, Justin B. Kinney


This article is a preprint and has not been certified by peer review [what does this mean?].
Idea: MAVE data need to be modeled to get accurate quantification of variant effect.
MAVE as Noisy Measurement of Genotype-Phenotype Map

Goal is to model both G-P map and measurement process
What Should a Genotype-Phenotype Map Look Like?

1. Additive (all positions in a sequence contribute independently)
2. Neighbor (interaction terms between neighboring positions)
3. Pairwise (each position interacts with every other position)
4. Black Box (roll your own relationship function)
Discrete vs Continuous Measurements

• Type of data influences how the model is fit.

• Barcode RNA-seq is continuous

• Sort-seq expression values are limited by the number of sorted bins.
Contrast MAVE-NN with ENVISION

- Predict amino acid variant effects globally (ENVISION) vs model genotype-phenotype map of the sequences tested (MAVE-NN)

- Input data: Additional ‘features’ to describe amino acids (ENVISION) vs sequence and assay results only (MAVE-NN)
Genotype-Phenotype Map of GB1

GB1 = B1 domain of protein G, assayed for IgG binding
Modeling the Non-linear Relationship Between Assay Output and Modeled Phenotype
Takeaways

- MAVE output is sequencing data, not a direct measure of the property of interest.
- Modeling gets you from sequencing data to the property of interest.
Example 3: Understanding Epistasis

RESEARCH ARTICLE

Inferring the shape of global epistasis

Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin

PNAS August 7, 2018 115 (32) E7550-E7558; first published July 23, 2018; https://doi.org/10.1073/pnas.1804015115
Global epistasis:

“Mutations may act additively on some underlying, unobserved trait, and this trait is then transformed via a nonlinear function to the observed phenotype as a result of subsequent biophysical and cellular processes.”
Translation:

Global epistasis means mutations act *non-additively* on phenotype (epistasis), but this may be due to *additive* effects on some underlying trait.

Local epistasis means specific pairwise interactions between residues.
Phenotype

Mutations move additively along x-axis

Unobserved molecular trait

Mutations move additively along x-axis
Model of global epistasis

Unobserved additive trait

\[ \phi = \beta_0 + \sum_{i}^{L} \beta_{i,a_i} \]

Sum of variant effects

Predicted phenotype

\[ y = g(\phi) + \varepsilon \]

Unobserved non-linear map between additive trait and phenotype
GB1 binding to IgG
MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M. Starita, Anthony T. Papenfuss, Frederick P. Roth, Douglas M. Fowler & Alan F. Rubin

Genome Biology 20, Article number: 223 (2019) | Cite this article

3720 Accesses | 7 Citations | 43 Altmetric | Metrics
MPRA Data Analysis

MPRAnalyze: statistical framework for massively parallel reporter assays

Tal Ashuach, David S. Fischer, Anat Kreimer, Nadav Ahituv, Fabian J. Theis & Nir Yosef

Genome Biology 20, Article number: 183 (2019) | Cite this article

4967 Accesses | 6 Citations | 14 Altmetric | Metrics
MPRA Quantified as RNA/DNA

Array-synthesized library

<table>
<thead>
<tr>
<th>CRE</th>
<th>DsRed</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expression

<table>
<thead>
<tr>
<th>BC</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>DNA</td>
</tr>
<tr>
<td>BC</td>
<td>RNA</td>
</tr>
<tr>
<td>BC</td>
<td>DNA</td>
</tr>
</tbody>
</table>

Transfection into cells, Sequence BCs
Problems with barcode RNA/DNA ratios

- Ratio of two noisy measurements
- When read counts are low, small differences in reads lead to very large changes in ratio
- Errors across measurements of multiple barcodes can be propagated to estimate variance, but common averaging approaches fail to take advantage of this.

If you want maximize the statistical power of your experiment, then you need a good estimate of variance!
Modeling Approach

External Covariates

Plasmids

DNA counts

Transcripts

RNA counts

Latent

Observed

\( \alpha \)
Consider observed reads as sampling from a distribution

\[ \hat{d} \sim \text{Gamma} \left( k, b \right) \]

\[ \hat{r} | \hat{d} \sim \text{Poisson} \left( \alpha \hat{d} \right) \]

Estimated ‘latent’ DNA reads

Estimated ‘latent’ RNA reads

Combine to get negative binomial for estimated RNA reads:

\[ \hat{r} \sim NB \left( \mu = \frac{\alpha \cdot k}{\beta}, \psi = k \right) \]
How well does it work?

Model tr. rate
Digression: How many barcodes per element in an MPRA library?

included in each construct with no sequence-induced activity. We examined the variance of the estimates on these sets. In the Kwasnieski dataset, the limited number of barcodes ($n=4$) is mitigated by high counts per barcode (Fig. 1a), leading to all estimates having similarly low variance. In the barcode-rich datasets ($n\geq90$), the mean ratio is expectedly [20] the most variable, with $\alpha$ being the most consistent in the Inoue-Kircher datasets
Reproducibility

4 barcodes per CRE

$R^2 = 0.995$

>90 barcodes per CRE

$R^2 = 0.876$
How many barcodes?

• More barcodes means bigger library, less coverage, not all constructs transfected as efficiently, noisier per-barcode measurements.

• BUT fewer barcodes means less built-in replication, thus less robustness against a noisy system, synthesis errors, etc.

• Common practice: More barcodes to test variants, fewer to test sets of CREs
Significance testing MPRA Variants

How would you calculate a p-value for a single reporter gene compared with a variant?

Reporter Expression

<table>
<thead>
<tr>
<th>CRE</th>
<th>DsRed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CRE</th>
<th>DsRed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>
Significance testing MPRA Variants - What happens when you scale up?

- With more tests, more opportunity for false positives.
- Failure of assumptions for significance testing can generate *many* false positives.
- BUT: More opportunity generate robust estimates of variance by leveraging large number of measurements.
Better estimates of variance give you better power to reliably detect variant effects.

**MPRAscore: robust and non-parametric analysis of massively parallel reporter assays**

Abhishek Niroula, Ram Ajore, Björn Nilsson  


**Published:** 29 July 2019  
**Article history**
Takeaways

- Scale of MPRAs can be leveraged to improve statistical tests for variant effects.
- As with protein DMS assays, MPRAs produces sequencing data, which can be modeled for better quantification and significance testing.
Lecture Summary

- Aim of MAVE is to understand quantitative relationship between sequence and function.

- A key premise: To effectively predict pathogenicity of variants at scale, it is important to have models of the molecular effects of variants on function.

- MAVE datasets are large, complex, and consist of sequencing data, thus modeling the data is an effective way to quantify variant effects.
For a deeper dive

Massively Parallel Assays and Quantitative Sequence–Function Relationships

*Annual Review of Genomics and Human Genetics*
Vol. 20:99-127 (Volume publication date August 2019)
First published as a Review in Advance on May 15, 2019
https://doi.org/10.1146/annurev-genom-083118-014845

Justin B. Kinney and David M. McCandlish
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; email: jkinney@cshl.edu, mccandlish@cshl.edu