Analysis of Variant Effect Data II

Mike White
Bio5488
4/13/22
Learning Goals Yet Again

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?**
Big Themes

- Modeling MPRA data
- Predictions of variant effect from evolutionary/population variation/genomic annotation data
Database of MAVE results with visualization tools

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
About

MaveDB is a public repository for datasets from Multiplexed Assays of Variant Effect (MAVEs), such as those generated by deep mutational scanning (DMS) or massively parallel reporter assay (MPRA) experiments.

MaveDB is open-source, released under the AGPLv3 license.

MaveDB is hosted by the Fowler Lab in the Department of Genome Sciences at the University of Washington. It is supported and developed by the University of Washington, the Walter and Eliza Hall Institute of Medical Research, and the Brotman Baty Institute.

Citation

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


News

[2021-11-29]:
MaveDB v2.1.0 has launched! See the GitHub release notes for details.

Highlights

Organisms
- Homo sapiens
- Other - genome not listed
- Saccharomyces cerevisiae

Target genes
- HSP90
- VIM-2 with p.Met1_Phe2insGly
- alpha-synuclein

Keywords
- MPRA
- barcode sequencing
- regression
Modeling Reporter Gene Assays
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

Expression

RNA
DNA
Problems with MPRA Measurements

- Measure is a ratio of counts - can be unstable
- Multiple barcodes tagging same CRE: how should we account for them?
- Transfection of complex libraries across biological replicates can introduce significant variation
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!
MPRAnalyze Basic Idea:
Assume Linear Relationship

\[ RNA_{\text{molecules}} = \alpha \times DNA_{\text{molecules}} \]

Goal is to estimate \( \alpha \), the transcription rate.
Modeling Approach

- **External Covariates**
- **Plasmids**
  - DNA counts
- **Transcripts**
  - RNA counts


diagram showing the flow among Plasmids, Transcripts, DNA counts, and RNA counts with an external covariate influence.
DNA Sampling

Estimated

Plasmids

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

Observed

DNA counts
RNA Estimate

\[ \tilde{r} | \tilde{d} \sim \text{Poisson} \left( \alpha \tilde{d} \right) \]
Consider observed reads as sampling from a distribution

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
Reproducibility

4 barcodes per CRE

$R^2 = 0.995$

>90 barcodes per CRE

$R^2 = 0.876$
Better estimates of variance give you better power to reliably detect variant effects

MPRA score: robust and non-parametric analysis of massively parallel reporter assays
Abhishek Niroula, Ram Ajore, Björn Nilsson

Bioinformatics, Volume 35, Issue 24, 15 December 2019, Pages 5351–5353,
https://doi.org/10.1093/bioinformatics/btz591
Published: 29 July 2019  Article history ▼
Significance testing MPRA Variants

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

![Diagram showing reporter expression levels]

<table>
<thead>
<tr>
<th>Reporter Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
</tr>
<tr>
<td>85</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
Multiple Oligos Covering A SNP

Genome

Oligos w. ref. allele

Oligos w. alt. allele
1. Calculate RNA/DNA for each barcode

2. Average over barcodes for one sequence window

3. Calculate difference in means between alleles for one seq. window

4. Calculate mean allele difference weighted by variance (see paper for details - this is the trick)

5. Perform significance testing by permutation testing
Comparison of p values by different methods
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.
Variant Effect Prediction from Epigenetic and Evolutionary Information
Predicting variant effects without MAVE data:

- Functional genomic annotation
- Evolutionary conservation
- Human population diversity
A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko, Melissa J Hubisz, Ilan Gronau & Adam Siepel

Nature Genetics 47, 276–283 (2015) | Cite this article

14k Accesses | 147 Citations | 184 Altmetric | Metrics
Functional genomic data

Clustering by functional genomic fingerprint
Estimation of a probability of fitness consequences (fitCons) score per cluster.

Mapping of fitCons scores to the genome.
fitCons score = probability site is under selection
fitCons Predictions Across the Genome
Support Vector Machine Predictor

- Machine learning approach trained on positive versus negative sequences (e.g. regulatory sequences vs random DNA)
- Predictions are from sequence alone
- Primarily non-coding regulatory variants
Support Vector Machine Classifiers

A method to predict the impact of regulatory variants from DNA sequence

Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion & Michael A Beer

*Nature Genetics* 47, 955–961 (2015) | Cite this article

23k Accesses | 247 Citations | 102 Altmetric | Metrics
Positive training set

Putative regulatory sequences

Negative training set

Random genomic sequences

gkm-SVM training

Regulatory sequence vocabulary

<table>
<thead>
<tr>
<th>All unique 10-mers</th>
<th>SVM weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGACTCATT</td>
<td>3.275</td>
</tr>
<tr>
<td>ATGAGTCATC</td>
<td>3.147</td>
</tr>
<tr>
<td>ATCATGTGAC</td>
<td>3.091</td>
</tr>
<tr>
<td>GTGATGTCATC</td>
<td>2.992</td>
</tr>
<tr>
<td>ACGAGAACA</td>
<td>0.0002</td>
</tr>
<tr>
<td>ACTATAACCA</td>
<td>0.0001</td>
</tr>
<tr>
<td>ATTTGCTAG</td>
<td>0.0001</td>
</tr>
<tr>
<td>GGATAAAATA</td>
<td>-0.0001</td>
</tr>
<tr>
<td>CAGGTGTGAG</td>
<td>-1.171</td>
</tr>
<tr>
<td>ATTCACCTG</td>
<td>-1.183</td>
</tr>
<tr>
<td>ACACACCTGT</td>
<td>-1.253</td>
</tr>
<tr>
<td>ATCCAGGGTG</td>
<td>-1.282</td>
</tr>
</tbody>
</table>
Sequence variant(s) of interest

WT

...TCTCTGCAACAAAGACAGA...

Variant

...TCTCTGCAAACAAAGACAGA...

deltaSVM calculation

<table>
<thead>
<tr>
<th>WT 10-mer</th>
<th>Weight</th>
<th>Variant 10-mer</th>
<th>Weight</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCTCTGCAAC</td>
<td>0.012</td>
<td>TCTCTGCAA</td>
<td>0.298</td>
<td>0.286</td>
</tr>
<tr>
<td>CTCTGCAACA</td>
<td>0.082</td>
<td>CTCTGCAAA</td>
<td>0.104</td>
<td>0.021</td>
</tr>
<tr>
<td>TCTGCAACAA</td>
<td>0.280</td>
<td>TCTGCAAAA</td>
<td>0.114</td>
<td>-0.166</td>
</tr>
<tr>
<td>CTGCAACAAA</td>
<td>0.330</td>
<td>CTGCAA AAAA</td>
<td>-0.029</td>
<td>-0.359</td>
</tr>
<tr>
<td>TGCAACAAAG</td>
<td>0.441</td>
<td>TGCAAAAAAAG</td>
<td>-0.025</td>
<td>-0.466</td>
</tr>
<tr>
<td>GCAACAAAGA</td>
<td>0.784</td>
<td>GCAAAAAAGA</td>
<td>0.008</td>
<td>-0.776</td>
</tr>
<tr>
<td>CAAACAGAC</td>
<td>1.031</td>
<td>CAAA AAAAGAC</td>
<td>0.109</td>
<td>-0.922</td>
</tr>
<tr>
<td>AACAAAGACAA</td>
<td>0.545</td>
<td>AAAA AGACCA</td>
<td>-0.453</td>
<td>-0.998</td>
</tr>
<tr>
<td>ACAAGACAG</td>
<td>0.671</td>
<td>AAAA AGACAG</td>
<td>-0.516</td>
<td>-1.187</td>
</tr>
<tr>
<td>CAAAGACAGA</td>
<td>-0.036</td>
<td>AAAA AGACAG</td>
<td>-0.478</td>
<td>-0.442</td>
</tr>
</tbody>
</table>

deltaSVM = -5.007
<table>
<thead>
<tr>
<th>Ref 10-mer</th>
<th>Weight</th>
<th>Alt 10-mer</th>
<th>Weight</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTGGAAATCC</td>
<td>-0.033</td>
<td>TTGGAAATCT</td>
<td>-0.470</td>
<td>-0.437</td>
</tr>
<tr>
<td>TGGAATCCC</td>
<td>2.209</td>
<td>TGGAATCT</td>
<td>0.254</td>
<td>-1.955</td>
</tr>
<tr>
<td>GGAATCCC</td>
<td>5.574</td>
<td>GGAATCTC</td>
<td>1.179</td>
<td>-4.395</td>
</tr>
<tr>
<td>GAAATCCCCA</td>
<td>2.265</td>
<td>GAAATCTCCA</td>
<td>-0.082</td>
<td>-2.347</td>
</tr>
<tr>
<td>AAATCCCAAG</td>
<td>1.541</td>
<td>AAATCTCCAG</td>
<td>-0.220</td>
<td>-1.762</td>
</tr>
<tr>
<td>AATCCCTCAT</td>
<td>0.941</td>
<td>AATCTCCAGT</td>
<td>-0.139</td>
<td>-1.080</td>
</tr>
<tr>
<td>ATCCCTCAAGT</td>
<td>0.228</td>
<td>ATCTCCAGTT</td>
<td>-0.025</td>
<td>-0.254</td>
</tr>
<tr>
<td>TCCTCAGTTT</td>
<td>-0.217</td>
<td>TCCTCAGTTT</td>
<td>-0.781</td>
<td>-0.564</td>
</tr>
<tr>
<td>CCCCAGTTTA</td>
<td>0.223</td>
<td>CTCCAGTTTA</td>
<td>-0.028</td>
<td>-0.251</td>
</tr>
<tr>
<td>CCCCAGTTTAT</td>
<td>-0.429</td>
<td>TCCAGTTTAT</td>
<td>-0.515</td>
<td>-0.086</td>
</tr>
</tbody>
</table>

deltaSVM = -13.131
Effect sizes of DNaseI sensitivity QTLs versus deltaSVM score.
Causal SNPs Stand Out
Main Points

• Scale of MPRAs means data are effectively analyzed with statistical modeling (compared with one at a time reporter assays)

• Models often try to capture both biology and assay effects, frequently by estimating latent (unobserved) variables

• Prediction from sequence, conservation, genome annotation leverage extensive data

• There is no one predictor to rule them all
Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN

Reading Tips

• Don’t worry too much about biological details

• Focus on input data, variant effect prediction strategy, model output.

• How well did it work?
Questions?

mawhite@wustl.edu