Applications of machine learning to solve biological puzzles

Carla M. Mann

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons, and the Computer Sciences Commons

Recommended Citation
Mann, Carla M., "Applications of machine learning to solve biological puzzles" (2019). Graduate Theses and Dissertations. 17508. https://lib.dr.iastate.edu/etd/17508

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Applications of machine learning to solve biological puzzles

by

Carla M. Mann

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Drena L. Dobbs, Co-major Professor
Robert Jernigan, Co-major Professor
Carolyn Lawrence-Dill
Maura McGrail
Kris De Brabanter

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2019

Copyright © Carla M. Mann, 2019. All rights reserved.
DEDICATION

There were a number of people who did not participate in the research or writing or lab meetings, but without them this dissertation would not exist and I would be a very different person. Thank you for everything.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xv</td>
</tr>
<tr>
<td>NOMENCLATURE</td>
<td>xvi</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>xix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xxii</td>
</tr>
<tr>
<td><strong>CHAPTER 1. INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Identifying Complex Recognition Signals in Biological Sequences</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Why Machine Learning?</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Specific Aims of This Research</td>
<td>3</td>
</tr>
<tr>
<td>1.4 Organization of This Thesis</td>
<td>3</td>
</tr>
<tr>
<td>References</td>
<td>7</td>
</tr>
<tr>
<td><strong>CHAPTER 2. RNA-PROTEIN INTERACTION PREDICTIONS WIKI PAGE</strong></td>
<td>9</td>
</tr>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Significance</td>
<td>10</td>
</tr>
<tr>
<td>Features</td>
<td>11</td>
</tr>
<tr>
<td>Sequence Based Features</td>
<td>12</td>
</tr>
<tr>
<td>- Sequence composition</td>
<td>12</td>
</tr>
<tr>
<td>- Sequence motifs</td>
<td>14</td>
</tr>
<tr>
<td>- Hydrophobicity and hydrophilicity</td>
<td>14</td>
</tr>
<tr>
<td>Structure-based Features</td>
<td>14</td>
</tr>
<tr>
<td>- Protein secondary structure</td>
<td>15</td>
</tr>
<tr>
<td>- Protein disorder</td>
<td>15</td>
</tr>
<tr>
<td>- RNA secondary structure</td>
<td>15</td>
</tr>
<tr>
<td>Feature Dimensions</td>
<td>16</td>
</tr>
<tr>
<td>Models</td>
<td>16</td>
</tr>
<tr>
<td>- Machine Learning Methods</td>
<td>16</td>
</tr>
<tr>
<td>- Random forests</td>
<td>16</td>
</tr>
<tr>
<td>- Gradient boosting</td>
<td>16</td>
</tr>
<tr>
<td>- Support vector machines</td>
<td>16</td>
</tr>
<tr>
<td>- Neural networks</td>
<td>17</td>
</tr>
<tr>
<td>- Multi-classifier methods</td>
<td>17</td>
</tr>
<tr>
<td>Scoring Systems</td>
<td>17</td>
</tr>
<tr>
<td>Datasets</td>
<td>18</td>
</tr>
<tr>
<td>- Dataset Creation</td>
<td>18</td>
</tr>
<tr>
<td>- Structure-derived datasets</td>
<td>18</td>
</tr>
<tr>
<td>- Datasets from high-throughput experiments</td>
<td>19</td>
</tr>
<tr>
<td>- Non-redundant datasets</td>
<td>19</td>
</tr>
</tbody>
</table>
### Experimentally-validated negative training datasets

Experimentally-validated negative training datasets are used to train machine learning models to predict protein-protein interactions (PPIs). These datasets are important because they provide a benchmark for evaluating the performance of prediction methods. The list includes datasets that have been experimentally validated:
- **RPI2241** structure-derived dataset (RPI-PDB)
- **RPI12252** NPInter-derived dataset (RPI-NPInter*)
- **FMRP119** case-study test dataset (RPI-FMRP)

### Publicly Available Datasets

Publicly available datasets are crucial for validating and testing prediction methods. These datasets are typically used as training and testing sets in the evaluation process.
- **Enzyme Commission** database
- **KEGG** database
- **Reactome** database
- **UniProt** database
- **POSTAR** database
- **NPInter** database
- **GEO** database
- **ENCODE** database
- **Protein Data Bank**
- **Nucleic Acid Database**

### Methods

Methods section includes the following:

- **Introduction**
- **RNA-Protein Interactions Play Important Biological Roles**
- **Examples of Disruptions in Regulatory RNA-Protein Interaction Networks that Lead to Disease**
- **Intrinsic Protein Disorder May Play a Role in Determining RNA-Protein Interaction Specificity**
- **Why Predict RPIs?**
- **Available RPI Prediction Methods**

### Databases of Known Interactions

Databases of known interactions are essential resources for understanding the functional and structural aspects of protein-protein interactions. These databases include:
- **Enzyme Commission**
- **KEGG**
- **Reactome**
- **UniProt**
- **POSTAR**
- **NPInter**
- **GEO**
- **ENCODE**
- **Protein Data Bank**
- **Nucleic Acid Database**

### Structure-based Databases:

Structure-based databases are used to represent the three-dimensional structures of proteins, which are crucial for understanding how these structures interact with other proteins.
- **Enzyme Commission**
- **KEGG**
- **Reactome**
- **UniProt**
- **POSTAR**
- **NPInter**
- **GEO**
- **ENCODE**
- **Protein Data Bank**
- **Nucleic Acid Database**

### Sequence-based Databases:

Sequence-based databases are used to represent the sequences of proteins, which are important for understanding the functional aspects of these interactions.
- **Enzyme Commission**
- **KEGG**
- **Reactome**
- **UniProt**
- **POSTAR**
- **NPInter**
- **GEO**
- **ENCODE**
- **Protein Data Bank**
- **Nucleic Acid Database**

### Feature Representation

Feature representation is the process of converting raw data into a format that can be used by machine learning algorithms. This section covers:
- **Sequence composition**
- **Protein disorder prediction**
- **Protein disorder features**

### Machine Learning Models

Machine learning models are used to predict protein-protein interactions. This section covers:
- **Performance Metrics**
  - True positive rate (sensitivity, recall)
  - False positive rate
  - Specificity
  - Precision
  - F-measure
  - Area under the receiver operating characteristic curve
  - Accuracy
  - Matthews correlation coefficient

### Feature Impact

Feature impact analysis helps in understanding the importance of different features in the prediction of protein-protein interactions.
# Table of Contents

- Feature importance .................................................................................................................. 51
- Results and Discussion ............................................................................................................. 52
  - Performance on the RPI-PDB Training Dataset ..................................................................... 52
  - Parameter Tuning .................................................................................................................... 53
  - Comparison to Other Methods on RPI-NPInter* Dataset ..................................................... 54
  - Characterization of Order and Disorder in the RPI-PDB and RPI-NPInter* Datasets ............. 56
  - Performance on FMRP119 Testing Dataset ......................................................................... 62
  - Contributions of Disordered Features ................................................................................... 67
  - Disorder Feature Importance ............................................................................................... 68
    - Features important for identifying the “Interacting” (positive) class ................................ 69
    - Features important for identifying the “Non-Interacting” (negative) class ......................... 70
  - Feature importance based on Mean Decrease in Accuracy .................................................. 73
  - Gini importance ..................................................................................................................... 74
- Conclusion ............................................................................................................................... 75
- Future Directions ..................................................................................................................... 76
- Availability ............................................................................................................................... 76
- Acknowledgements .................................................................................................................. 77
- References ............................................................................................................................... 77
- Supplemental Material ............................................................................................................ 91
  - Protein Features ..................................................................................................................... 91
  - Supplemental Disorder Feature Tables .................................................................................. 98

# Chapter 4. MEDJED: A MACHINE LEARNING REGRESSION MODEL FOR PREDICTING THE EXTENT OF MICROHOMOLOGY-MEDIATED END JOINING REPAIR IN RESPONSE TO DOUBLE-STRANDED DNA BREAKS

- Abstract ................................................................................................................................... 102
- Introduction & Background ...................................................................................................... 102
  - DNA Double Strand Break (DSB) Repair in Gene Editing ..................................................... 103
  - Predicting Preferred Gene Knockout Sites .............................................................................. 106
  - Motivation for This Study ...................................................................................................... 107
- Materials and Methods .......................................................................................................... 108
  - Dataset ................................................................................................................................. 108
  - Dataset processing ................................................................................................................. 108
  - Training and Validation Datasets .......................................................................................... 109
  - Dataset Effects ...................................................................................................................... 109
  - Features ................................................................................................................................. 109
  - Model development .............................................................................................................. 110
  - Feature selection ................................................................................................................... 110
- Results and Discussion ............................................................................................................ 111
  - Feature Importance .............................................................................................................. 112
  - Dataset Effects ...................................................................................................................... 115
- Conclusion ............................................................................................................................... 117
- Acknowledgements .................................................................................................................. 118
- Supplementary Materials ...................................................................................................... 119
  - Supp. Code 1 ......................................................................................................................... 119
- References ............................................................................................................................... 120
3.3. IncPro - to predict protein binding partner(s) for any known IncRNA .......................... 161
3.4. catRAPID - to predict either individual or transcriptome/proteome wide interactions ...................................................................................... 162
4. Notes ........................................................................................................ 167
Acknowledgments ...................................................................................... 172
References .................................................................................................. 172

APPENDIX B. ROBUST ACTIVATION OF MICROHOMOLOGY-MEDIATED END JOINING FOR PRECISION GENE EDITING APPLICATIONS ................... 176

Abstract ........................................................................................................ 176
Author Summary .......................................................................................... 177
Introduction .................................................................................................. 178
Results .......................................................................................................... 181

MMEJ is an active repair pathway in the genetically unaltered zebrafish embryo .......................................................... 181
Many Bae et al. predicted MMEJ loci are preferentially repaired by NHEJ .................................. 184
Rate of Pattern Score change as a discrimination factor for MMEJ induction in vivo and in vitro ........................................................................ 184
Competition hypothesis predicts new PreMA reagents ........................................... 185
Low competition plus proximity of microhomology arms strongly predicts PreMA reagents: V2 .................................................................................. 189
Mechanism of MMEJ-activation may be conserved in vertebrates ...................... 190
Accessing the PreMA algorithm through MENTHU (MMEJ Knockout Target
Heuristic Utility) .......................................................................................... 192

Discussion ....................................................................................................... 193
Materials and methods .................................................................................. 196

Ethics statement ............................................................................................ 196
Microhomology arms ..................................................................................... 196
Zebrafish husbandry ....................................................................................... 196
DNA oligonucleotide preparation ................................................................... 197
sgRNA expression vector synthesis ................................................................ 197
TALEN synthesis .......................................................................................... 197
In-vitro transcription and RNA preparation ..................................................... 198
CRISPR-Cas9 RNP preparation for microinjections .......................................... 198
TALEN and CRISPR-Cas9 RNA preparation for microinjections ................. 199
Microinjections ............................................................................................... 199
Phenotype scoring .......................................................................................... 199
Zebrafish DNA extraction and assessing mutagenic outcomes ...................... 200
Germline transmission for 5 bp deletion generated by N2B sgRNA #1 ............ 201
Reanalyses of previously published deep sequencing dataset .......................... 201
Cell culture and RNP transfection ................................................................. 202
HEK293T cell DNA extraction and assessing mutagenic outcomes ............... 203
MENTHU .................................................................................................... 204
Statistical analyses ........................................................................................ 205

Supporting information .................................................................................. 205

S1 Note Calculation of Microhomology Fraction .......................................... 216
S2 Note Calculation of Slope Values .............................................................. 216
S3 Note Calculation of Top Microhomology Fraction ........................................ 217
S1 Data Sanger sequencing file used for the study. ............................................. 219
Acknowledgments ................................................................................................. 219
References ............................................................................................................. 219

APPENDIX C. GeneWeld: A METHOD FOR EFFICIENT TARGETED
INTEGRATION DIRECTED BY SHORT HOMOLOGY ........................................ 224

APPENDIX D. GENE SCULPT SUITE USAGE STATISTICS ............................. 225
    MEDJED ........................................................................................................... 225
    GTagHD ......................................................................................................... 227
    MENTHU ....................................................................................................... 228
    Conclusions ................................................................................................... 230
LIST OF FIGURES

Figure 2-1 Illustration of a conjoint triad representation of a protein sequence. 1) *Drosophila* peptide tarsal-less AA protein sequence [UniProt A3RLR1] represented using the 1-letter amino acid code. 2) Alphabet reduction using a 7-letter alphabet based on Muppirala et al. 2011 and Shen et al. 2007. Each amino acid group is represented by a colored block. 3) In a conjoint triad feature representation, each continuous block of three amino acids (p1, p2, etc.) is identified. 4) The frequency of each possible triad (f1, f2, etc.) is calculated by dividing the number of times a triad appears by the length of the sequence (e.g., p1 appears once in the 32-nt long sequence, so the corresponding feature (f27) is 1/32, or 0.03125). ................................................................. 13

Figure 3-1. After the disordered residues in a sequence have been predicted using DISOPRED3, the sequence is divided into quarters, and the percentage of each quarter of the sequence that is predicted to be disordered is then incorporated as features D2, D3, D4, and D5................................................................. 46

Figure 3-2 Example ROC curves. A hypothetical perfect classifier (orange) would have 100% TPR and 0% FPR, regardless of threshold, which gives an AUC of 1. A random (binary) classifier will have 50% TPR and 50% FPR, regardless of threshold, which produces an AUC of 0.5. Better classifiers will not have a large trade-off between TPR and FPR, producing ROC curves with a high TPR when the FPR is low, leading to an AUC closer to 1 (gray). Worse classifiers will not have as high a TPR when the FPR is low, leading to smaller AUCs closer to 0.5 (yellow).................................................................................. 50

Figure 3-3 RPI-PDB 10-fold cross-validation results for three random forest classifiers. The Weka 3.8 random forest classifier with default parameters is in yellow. A parameter-tuned R randomForest classifier is in purple. The parameter-tuned RPIDisorder model is in light blue. The R randomForest method AUC is not reported. ........................................... 53

Figure 3-4 RPI-NPInter validation performance of the Weka random forest classifier with default parameters (yellow), tuned R randomForest classifier (purple), and the tuned RPIDisorder model (blue).................................. 54

Figure 3-5 RPI-NPInter/RPI-NPInter* performance comparison between RPISeq (gray), RPIMotif (red), and RPIDisorder (blue) on the dataset. All three methods were trained on RPI2241. RPIDisorder results do not include a
single interacting instance for which DISOPRED3 did not return results (denoted as RPI-NPInter*). RPISeq and RPIMotif include the full RPI-NPInter dataset.

Figure 3-6 Area Under the Receiver Operating Characteristic Curves (AUC of ROC) for RPISeq (gray), RPIMotif (red), and RPIDisorder (blue). AUCs are 0.87, 0.92, and 0.92 for RPISeq, RPIMotif, and RPIDisorder, respectively. Curve created in R using ggplot2 [Wickham 2016] with code provided courtesy of Kris De Brabanter.

Figure 3-7 Histogram of the proportion of sequence predicted to be disordered for proteins in the RPI-PDB dataset (disorder feature D1). The y-axis is the proportion of all protein sequences within the dataset that have a particular level of sequence disorder.

Figure 3-8 Histogram of the proportion of sequence predicted to be disordered for proteins in the RPI-NPInter* dataset (disorder feature D1). The y-axis is the proportion of all protein sequences within the dataset that have a particular level of sequence disorder.

Figure 3-9 Boxplots of the predicted proportion of disordered residues in regions of RPI2241 proteins (disorder features D6-D15). The protein sequences are divided into tenths, and the proportion of the first tenth, second tenth, etc. which is predicted to be disordered is determined (see Methods and Materials). The boxes represent the 25th, median, and 75th percentiles; the whisker boundaries are 1.5 * the interquartile range above and below the 25th and 75th percentiles.

Figure 3-10 Boxplots of the predicted proportion of disordered residues in regions of RPI12252* proteins (disorder features D6-D15). The protein sequences are divided into tenths, and the proportion of the first tenth, second tenth, etc. which is predicted to be disordered is determined (see Methods and Materials).

Figure 3-11 Boxplots of the proportion of predicted ordered residues within the RPI2241 dataset proteins (gray) and the RPI12252* dataset proteins (green) which are hydrophilic, moderately hydrophilic, hydrophobic, positively charged, negatively charged, and uncharged.

Figure 3-12 Boxplots of the proportion of predicted disordered residues within the RPI2241 dataset proteins (gray) and the RPI12252* dataset proteins (green) which are hydrophilic, moderately hydrophilic, hydrophobic, positively charged, negatively charged, and uncharged.
Figure 3-13 Performance comparison of RPISeq (gray), RPIMotif (red), and RPIDisorder (blue) on the FMRP119 dataset. .......................................................... 66

Figure 3-14 Area Under the Receiver Operating Characteristic Curves (AUC of ROC) for RPISeq (gray), RPIMotif (red), and RPIDisorder (blue). AUCs are 0.86, 0.90, and 0.88 for RPISeq, RPIMotif, and RPIDisorder, respectively. Curves created in R using ggplot2 [Wickham 2016] with code provided courtesy of Kris De Brabanter. ...................................................... 67

Figure 3-15 Comparison of the four importance measures reported by R randomForest: The importance of features in identifying the interacting RNA-protein class (Interacting), the non-interacting RNA-protein class (Non-interacting), the mean decrease in model accuracy (MDA), and the mean decrease in node impurity (Gini) produced by the feature. ........... 68

Figure 3-16 Feature importance (Mean Decrease in Accuracy, %) for identifying interacting RNA-protein pairs. The collective importance of all features for identifying the interacting class is in white. The protein features (P1-P343) are in red. The RNA features (R1-R56) are highlighted in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray. ........................................................................... 71

Figure 3-17 Feature importance for identifying non-interacting RNA-protein pairs. The collective importance of all features for identifying the non-interacting class is in white. The protein features (P1-P343) are highlighted in red. The RNA features (R1-R56) are highlighted in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray. .............................................................................. 72

Figure 3-18 Feature importance ranked by Mean Decrease in Accuracy. The collective importance of all features is in white. The protein features (P1-P343) are in red. The RNA features (R1-R56) are in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray. ................................................................................. 73

Figure 3-19 Feature importance ranked by decrease in node impurity. The collective importance of all features is in white. The protein features (P1-P343) are in red. The RNA features (R1-R56) are in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray. .................................................................................... 74
Figure 4-1 MEDJED performance. On the independent test set, MEDJED achieves a Pearson Correlation Coefficient (PCC) of 81.36%, Mean Absolute Error (MAE) of 10.96%, and Root Mean Square Error (RMSE) of 13.09%. The MEDJED-predicted MMEJ repair proportion (x-axis) is graphed against the observed MMEJ repair proportion (y-axis). .......................... 112

Figure 4-2 Importance measures of the six features in the MEDJED model. The percent increase in mean squared error is to the left (%IncMSE) while the increase in node purity (IncNodePurity) is to the right. The features are (top to bottom) the maximum pattern score, standard deviation of the pattern score, mean microhomology arm length, standard deviation of the microhomology arm length, minimum deletion length, and maximum microhomology arm length for the collection of MMEJ-based deletion patterns at the DSB site. ................................................................. 113

Figure 4-3 Scatterplot of the standard deviation of the Bae et al. [Bae et al. 2014] pattern score plotted against the observed proportion of MMEJ deletions for 89 HeLa cell targets. Pearson correlation coefficient of 56.47%. .............................................................. 115

Figure 4-4 Boxplot of the mean absolute error (MAE), Pearson Correlation Coefficient (PCC), and root mean squared error (RMSE) on test sets. The original dataset was randomly split into training (75%) and testing sets (25%) 5000 times, generating 5000 classifiers with the same parameters as the final MEDJED model; each classifier was then assessed on its respective test dataset. The median MAE was 13.97% with standard deviation of 1.6%. The median PCC was 63.06%, with standard deviation 12.39%. The median RMSE was 16.54% with standard deviation of 1.7%. .............................................................. 116

Figure 5-1 DNA double-strand break (DSB) repair mechanisms. (A) Non-homologous end joining (NHEJ). The DNA DSB ends are bound by the Ku70-Ku80 heterodimer and undergo limited end-resection before DNA polymerases and ligases repair the break. This process may perfectly repair the DSB break, but more frequently introduces short indels (red). (B) Homology-directed repair (HDR). When a DSB is detected, homologous sequences (blue and orange segments), frequently provided by a sister chromatid are used as a template to repair the break (green). The resulting repair is usually precise. (C) Homology-mediated end joining (HMEJ). HMEJ is a catch-all term for repair that utilizes short regions of homology, including MMEJ and SSA. In both MMEJ and SSA, 5'-3' end-resection exposes single-stranded DNA regions, where homologous sections (blue) anneal with one another for repair. The overhanging DNA strands (red) are then
clipped, resulting in a short deletion. MMEJ and SSA are mechanistically similar but distinct pathways, utilizing different protein machinery. MMEJ also utilizes shorter regions of microhomology (~2-25bp) than SSA (>25bp). SSA end-resection can be extensive, so the pathway operates over larger nucleotide distances. ........................................ 125

Figure 5-2 Gene Sculpt Suite (GSS) Architecture. The GSS server uses ShinyProxy (https://www.shinyproxy.io/), to administer the Docker images (solid blue line) for each GSS tool. When a user (blue circle) visits a GSS tool URL, ShinyProxy creates a Docker container (dashed blue line), which essentially is a temporary copy of the Docker image, and allows a user to securely interact within their own container. These containers are temporary, and deleted once a user leaves their URL. A new container is spun up for each unique user. ........................................ 127

Figure 5-3 GeneWeld integration scheme (15). Short homologous sequences from the integration site in the target genome (in blue and orange) are cloned into the flanking regions of the donor plasmid cargo (green). When the cargo is freed from the plasmid, the homologous regions promote the efficient and precise integration of the cargo into the genomic locus using homology-mediated end joining. ........................................ 128

Figure 5-4 MEDJED performance. On a test set of 23 HeLa cell targets from (17), MEDJED achieves a Pearson Correlation Coefficient (PCC) of 85.2%, Mean Absolute Error (MAE) of 10.3%, and Root Mean Square Error (RMSE) of 12.0%. The MEDJED-predicted MMEJ repair proportion (x-axis) is graphed against the observed MMEJ repair proportion (y-axis). .............................................................. 131

Figure 5-5 Example MENTHU output table. Each row corresponds to a single DSB event. The “Target_Sequence” column contains the gRNA or TALEN sequence required to generate the DSB. The “MENTHU_Score” column gives the ratio between the Microhomology-Predictor pattern scores of the top two scoring microhomologies at the site; a DSB site is likely to produce a PreMA if the MENTHU Score is >1.5 (16). “The Frame_Shift” column indicates whether the most frequent expected deletion pattern induces a frameshift. The “Tool_Type” gives the PAM sequence for CRISPR nucleases, and the left arm length/spacer/right arm length combination for TALENs. The “Strand” column indicates whether the “Target_Sequence” occurs on the forward or complement strand. The “Exon_ID” provides the number of the exon in which the DSB site occurs; if no exon information is available, this value is 1. The “DSB_Location” provides the index of the nucleotide to the left of the DSB site within the entire nucleotide sequence. The
“Microhomology” column contains the sequence of the microhomology arms used to generate the deletion. The “PreMA_Sequence” gives the sequence of the predicted predominant repair outcome. The “Context” column (not shown) gives the sequence window used for MENTHU score calculations.

Figure 5-6 Strategy for handling staggered-cutting nucleases. End-resection operates in a 5’-3’ fashion. 5’ overhangs produced by a staggered-cutting nuclease will be removed during the resection phase. The eliminated sequence in the overhangs is thus unavailable for utilization in MMEJ. We can approximate the microhomologies available for use in MMEJ repair by creating a pseudostring DNA sequence made up of the 5’ strand up until the DSB site (orange) concatenated to the 3’ strand (blue). The 5’ overhangs (dashed lines) are effectively removed. This allows staggered DSBs to be treated identically to blunt DSBs, after the 5’ overhangs are removed from the sequence. The “Context” column within the MENTHU results table (see Fig. 5) contains this pseudostring when a staggered-cutting nuclease is chosen.
LIST OF TABLES

Table 2-1 Table of publically available benchmark datasets .......................................... 21
Table 2-2 Table of RNA-protein interaction prediction methods with publically available code .................................................................................................................. 23
Table 2-3 Table of published RNA-protein interaction prediction methods. This table will not be included in the Wiki article ................................................................. 27
Table 3-1: Importance metrics for the R randomForest classifier, ordered in descending order based on Interacting importance, for the top 5 Interacting features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. See “Methods” for description of disorder features. The first three columns are the Mean Decrease in Accuracy (%) for determining the Interacting class, Non-interacting class, and overall MDA, respectively ........................................................................ 71
Table 3-2 Importance metrics for the R randomForest classifier, ordered in descending order based on Non-interacting importance (Nonint), for the top 5 non-interacting features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. RNA conjoint tetrad features start with an “R”. See “Methods” for description of disorder features. .............................................................................................................. 72
Table 3-3 Importance metrics for the R randomForest classifier, ordered in descending order based on overall Mean Decrease in Accuracy, for the top 5 MDA features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. See “Methods” for description of disorder features. .............................................................................................................. 74
Table 3-4 Importance metrics for the R randomForest classifier, ordered in descending order based on Mean Decrease in Node Impurity (MeanDecreaseGini), for the top 5 features. Disordered features are highlighted in blue. RNA conjoint tetrad features start with an “R”. See “Methods” for description of disorder features .............................................................................................................. 75
Supp. Table 4-1 Training and Test set gene IDs .................................................................... 119
# NOMENCLATURE

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Area under the curve</td>
</tr>
<tr>
<td>Cas</td>
<td>CRISPR-associated protein</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered regularly interspersed short palindromic repeats</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSB</td>
<td>Double-strand break</td>
</tr>
<tr>
<td>DSBR</td>
<td>Double-strand break repair</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopedia of DNA Elements</td>
</tr>
<tr>
<td>FPR</td>
<td>False positive rate</td>
</tr>
<tr>
<td>FMRP</td>
<td>Fragile X mental retardation protein</td>
</tr>
<tr>
<td>FXS</td>
<td>Fragile X syndrome</td>
</tr>
<tr>
<td>Gini</td>
<td>Node impurity</td>
</tr>
<tr>
<td>GTagHD</td>
<td>pGTag Homology Designer</td>
</tr>
<tr>
<td>HDR</td>
<td>Homology-directed repair</td>
</tr>
<tr>
<td>HMEJ</td>
<td>Homology-mediated end joining</td>
</tr>
<tr>
<td>HR</td>
<td>Homologous recombination</td>
</tr>
<tr>
<td>IDP</td>
<td>Intrinsically disordered protein</td>
</tr>
<tr>
<td>IDR</td>
<td>Intrinsically disordered region</td>
</tr>
<tr>
<td>IDPR</td>
<td>Intrinsically disordered protein region</td>
</tr>
<tr>
<td>Indel</td>
<td>insertion/deletion</td>
</tr>
<tr>
<td>ISU</td>
<td>Iowa State University</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean absolute error</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>MCC</td>
<td>Matthews correlation coefficient</td>
</tr>
<tr>
<td>MDA</td>
<td>Mean decrease in accuracy</td>
</tr>
<tr>
<td>MEDJED</td>
<td>Microhomology-Evoked Deletion</td>
</tr>
<tr>
<td>MENTHU</td>
<td>Microhomology-mediated End joining</td>
</tr>
<tr>
<td>MMEJ</td>
<td>Microhomology-mediated end joining</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non-coding RNA</td>
</tr>
<tr>
<td>NDB</td>
<td>Nucleic Acid Database</td>
</tr>
<tr>
<td>NHEJ</td>
<td>Non-homologous end joining</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle components analysis</td>
</tr>
<tr>
<td>PCC</td>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>PDB</td>
<td>Proten Data Bank</td>
</tr>
<tr>
<td>piRNA</td>
<td>Piwi-interacting RNA</td>
</tr>
<tr>
<td>PreMA</td>
<td>Predominant MMEJ Allele</td>
</tr>
<tr>
<td>RBD</td>
<td>RNA-binding domain</td>
</tr>
<tr>
<td>RF</td>
<td>Random forest</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean square error</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNP(s)</td>
<td>Ribonucleoprotein(s)</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic (curve)</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>RPIP(s)</td>
<td>RNA-protein interaction partner(s)</td>
</tr>
<tr>
<td>RPIM(s)</td>
<td>RNA-protein interaction motif(s)</td>
</tr>
<tr>
<td>snoRNP</td>
<td>Small nucleolar ribonucleoprotein</td>
</tr>
<tr>
<td>SSA</td>
<td>Single-strand annealing</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>TALEN</td>
<td>Transcription activator-like effector nucleases</td>
</tr>
<tr>
<td>TPR</td>
<td>True positive rate</td>
</tr>
<tr>
<td>ZFN</td>
<td>Zinc finger nuclease</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

First, and foremost, I would like to thank Dr. Drena Dobbs for serving as my major professor and mentor, for her guidance on this journey, and for all of the important lessons she has taught me along the way.

I would also like to thank Drs. Robert Jernigan, Carolyn Lawrence-Dill, Maura McGrail, and Kris De Brabanter for serving on my committee, as well as Dr. Gustavo MacIntosh, for the expertise, advice, and guidance they have given me on this journey. I thank Drs. Dobbs and Jernigan for their editorial contributions to this dissertation.

I would especially like to thank Carolyn for ‘fostering’ me in her lab when I was going through a rough time and for being a wonderful mentor.

I would also like to thank Darwin Campbell, for letting me use his server, showing me the basics of system administration and making all of the websites contained herein possible.

I would also like to thank the members and associates of the Dill-PICL lab (Gokul, Ian, Ming, Scott, Sweta, Colleen, Dennis, Seven, and Nancy), for providing me with emotional, technical, and moral support. I would especially like thank Ian for his valuable conversations and feedback.

Additionally, I would like to thank Jeff Essner and Maura McGrail, and the members of their labs for their absolutely fascinating lab meetings, and their valuable discussions and feedback.

I would like to thank the Ekker and Clark lab groups at the Mayo Clinic, for teaching me the value of effective communication.
Thanks to Gabriel Martínez-Gálvez of the Mayo Clinic for collaborating with me to create MEDJED, and assisting in the development of the MENTHU web tool.

I would also like to thank Jo Anne Powell-Coffman, for being a wonderful faculty mentor during my time in the Preparing Future Faculty Program, and being an amazing GDCB chair.

I also must thank Diane Jepsen, Danise Jones, Carla Harris, and especially Trish Stauble, who are all amazing administrators whose efforts I have benefitted from extensively.

I would also like to acknowledge the Stacked Overflow community and how helpful they can occasionally be.

I would like to thank Dave, John, Ming, and Keting for all the study and brainstorming sessions, late nights, lunches, therapy, and comradery.

I would finally like to thank all my friends and family too numerous to count for at least trying to understand what I do, and for all the support you’ve given me over the years. Consider this work dedicated to you.

Thank you, Auntie Netta, for sharing Commencement with me.

I would like to thank Kitty for emotional support and being the best pillow in the world, and Macy for always being excited to see me, although they’re a cat and a dog, so they might not ever read this.

Thank you, Carolyn, for taking me in at a time when I really, really needed it. Your kindness and generosity over the past few years made it possible for me to finish this dissertation, and I will never be able to thank you enough.
Thank you, Drena, for being an amazing PI, mentor, and friend to me through some of the best and most difficult years of my life. I feel very privileged and honored to be your final graduate student, and I hope I’m sending you out on a high note.

Thank you, Heather, for putting up with me and my shenanigans for the past ten years, and being my best friend. My undergrad years would have been a much lesser experience without you, and I would be a lesser person.

Thank you, Mak, for being utterly amazing and wonderful. You have been truly inspirational to me all these years, and I want to be you when I grow up.

Thank you, Steve, for being the best brother anyone could ask for. All those little things you’ve said and done over the years have meant so much to me, even if – especially if - I’m the only person to see them.

Thank you, Sarah, for all the late nights and encouragement and, for so much of my life, being the giant on whose shoulders I stood.

Thank you, Mom and Dad, for microscopes and science camps and science fairs and tolerating kitchen table science experiments and backyard volcanos. Thank you for the weekly library trips, and reading to me every night. Thank you for always pushing me to do the best I can, and thank you for always encouraging me and telling me that I could be or do anything. Here’s that anything.
ABSTRACT

The era of “big data” has led to the generation of more biological data than any human could hope to process. This flood of data has necessitated the development of computational methods to assist in analysis, and has made it possible to begin to model complex biological systems. Machine learning methods represent one avenue for modeling, and allow for the identification of intricate and often cryptic sequence signals underlying many biological processes.

In this dissertation, I present two machine learning models, RPIDisorder and MEDJED, which were developed to predict RNA-protein interaction partners (RPIPs) and DNA double-strand break (DSB) repair by the microhomology-mediated end joining (MMEJ) pathway, respectively. I also present the Gene Sculpt Suite, a set of freely available web-based software tools for precision gene editing.

RPIDisorder uses signals from protein and RNA sequences (some of which have been previously utilized in published RNA-protein partner prediction methods), but it additionally exploits signal from disordered protein regions to predict interactions with greater specificity than has been possible before. RPIDisorder allows for the prediction of biologically relevant RNA-protein interaction networks, which in turn can assist in the development of clinical interventions for the numerous cancers and neurological and metabolic disorders associated with disruptions in RNA-protein interactions. RPIDisorder is freely available at www.rpidisorder.org.

MEDJED (Microhomology-Evoked Deletion Judication EluciDation) uses signal within and surrounding short stretches of homologous DNA sequence (microhomologies) on either side of an introduced DSB to predict the extent to which a targeted genomic site
will be repaired using the MMEJ pathway. MEDJED is freely available at www.genesculpt.org/medjed/.

The advent of gene editing nucleases including CRISPR/Cas systems, TALENs, and zinc finger nucleases has made it possible to insert, delete, and precisely edit DNA. A great deal of recent research has focused on improving the efficiency and precision of these nucleases by leveraging endogenous DSB repair pathways including non-homologous end joining (NHEJ) and homologous recombination (HR). However, homology-mediated end joining pathways (HMEJ), including MMEJ and single-strand annealing (SSA), provide many advantages over NHEJ and HR. The Gene Sculpt Suite is a set of three web-based tools (GTagHD, MEDJED, and MENTHU) that leverage HMEJ pathways to enhance exogenous DNA knock-in (GTagHD) and produce more efficient and precise gene knock-outs (MEDJED and MENTHU). The Gene Sculpt Suite is freely available at www.genesculpt.org.

Taken together, the results of these studies demonstrate that machine learning models can be valuable for identifying sequence signals that regulate macromolecular recognition, with numerous potential applications in both basic and applied research.
CHAPTER 1. INTRODUCTION

1.1 Identifying Complex Recognition Signals in Biological Sequences

This work addresses two bioinformatics problems, which, superficially, seem very different: i) RNA-protein interaction (RPI) partner prediction, and ii) DNA double-strand break repair (DSB) prediction. RPI partner (RPIP) prediction focuses on determining whether a specific RNA and a specific protein will bind to each other, and thereby influence each other’s structure and function. DNA DSB repair prediction attempts to identify which tool(s) a cell will use to reforge broken DNA.

These two problems are of a great deal of scientific interest. The first three papers describing computational RPIP prediction methodologies [Bellucci et al. 2011, Pancaldi & Bähler 2011, Muppirala et al. 2011], all published in 2011, have been cited collectively 305 times [see Bellucci et al. Altmetrics, Pancaldi & Bähler Altmetrics, Muppirala et al. Altmetrics] as of April 01, 2019. In the eight years since, more than two dozen papers describing RPIP methodologies have been published. This topic is of particular interest due to the crucial roles RPIPs play in regulating gene transcription and translation [Wilusz et al. 2009], post-transcriptional mRNA modification, stabilization, and regulation [Shi & Manley 2015], retrotransposon silencing [Sarkar et al. 2016], the formation of functional ribonucleoprotein complexes including ribosomes, spliceosomes, small nucleolar ribonucleoproteins (snoRNPs) [Bachellerie et al. 2002], and signal recognition particles [Akopian et al. 2013], and dozens of other crucial processes. In addition, numerous diseases including a host of cancers [Darnell et al. 2010, He et al. 2019] and neurological disorders [Darnell et al. 2010, Laneve et al. 2019] including fragile X syndrome, paraneoplastic neurologic syndromes, spinal muscular atrophy [Lukong et al. 2008], and many more

In November 2018, two methods predicting DNA DSB repair sequence outcomes were published in *Nature* and *Nature Biotechnology*; these methods have been collectively cited 15 times as of April 01, 2019, and currently are both among the top 5% of research outputs scored by Altmetric [Shen & Arbab 2018 Altmetrics, Allen & Crepaldi 2018 Altmetrics,]. A collaboration between the Essner, McGrail, Dobbs, Ekker, and Clark labs resulted in the creation of a new methodology (see Appendix B) that leverages biological signal in a DNA sequence to enhance gene editing outcomes by choosing sites likely to use microhomology-mediated end joining (MMEJ) to generate a predominant repair outcome. This paper has been viewed nearly 4,000 times and has been downloaded 2,267 times since it was published in September 2018 [Ata et al. 2018 Altmetrics], as of April 01, 2019. The MENTHU web tool presented in this dissertation (see Chapter 5), which implements this method, has been used 943 times by 481 people since January 25 of 2018 (see Appendix D). Our *unpublished* method, MEDJED, for predicting the extent to which a DSB site will utilize the MMEJ repair pathway (see Chapter 4), has been used 109 times by 65 people since April 1 of 2018 (see Appendix D), as of April 01, 2019.

These seemingly disparate problems, at their most basic level, ask the same question: can we learn or identify signals in biological sequences sufficiently informative to predict macromolecular recognition?

1.2 Why Machine Learning?

With the era of “big data” in full swing, the sheer volume of data in need of analysis requires the development of computational tools to evaluate that information. Machine learning can not only make use of this wealth of data to create models, but also can provide
insight into the signals extracted, and thus allow the generation of testable hypotheses to interrogate the biological mechanisms that employ these signals.

1.3 Specific Aims of This Research

The specific aims of this research are as follows:

1. Determine whether sequence signal from protein disorder can improve RNA-protein interaction prediction by increasing prediction specificity

2. Develop methods for predicting which DNA DSB repair pathway will be utilization after introduction of a DSB at a specific genomic target site with a gene editing nuclease

3. Design and implement freely available, easy-to-use, online and offline tools for predicting RPIPs and DNA DSB repair outcomes

1.4 Organization of This Thesis

This dissertation includes 6 Chapters and 4 Appendices:

Chapter 1 describes the motivation for this research and outlines the specific aims of this dissertation. This chapter does not include a literature review because that is provided in Chapter 2 and Appendix A.

Chapter 2 consists of a Wikipedia-style review entitled “RNA-Protein Interaction Prediction” (Mann CM, Dobbs D (2019) PLoS Genet.) in preparation for submission to the PLoS Genetic’s Topic Pages collection. Topic Pages are peer-reviewed, citable, and PubMed-indexed, and help serve the public by providing high-quality and easily accessible overviews of scientific topics. I was invited to create this Topic Page by PLoS Genetics Topic Pages Collection Editor Thomas Shafee. This chapter describes the features, mathematical models, and datasets used by methods for predicting RNA-protein interaction partners. Drena Dobbs and I came up with the idea for this study. I researched, wrote and
revised the article. This chapter serves as a brief overview of the current status of the RNA-protein partner prediction field.

**Chapter 3** is a manuscript in preparation for submission to *Bioinformatics*, entitled “RPIDisorder: A Method for Predicting RNA-Protein Interaction Partners Using Intrinsic Protein Disorder” (*Mann CM, Walia R, Dobbs D, Bioinformatics*). The preliminary manuscript describes a new machine learning method for predicting RNA-protein interaction partners. Rasna Walia created a dataset used in this study. I conceived the study, carried out the experiments and analysis, created the web tool and webserver, and wrote the manuscript.

**Chapter 4** describes MEDJED (Microhomology Evoked Deletion Judication Elucidation), a machine learning method for predicting the extent to which a DNA DSB site will undergo microhomology-mediated end joining (MMEJ) repair. While this work is likely not of sufficient novelty to be published separately (due to the recent publication of methods that perform functions similar to MEDJED [Shen & Arbab et al. 2018, Allen & Crepaldi et al. 2018], MEDJED is included in the Gene Sculpt Suite (described in **Chapter 5**), which is described in a manuscript under review for inclusion in the annual webserver issue of *Nucleic Acids Research*. This chapter describes the creation and motivation behind MEDJED, implications of results obtained to date, and future directions. Gabriel Martínez-Gálvez collaborated in developing the idea for MEDJED and performed statistical analyses of certain features (not included in Chapter 4). I wrote the entire chapter and performed all of the experiments and analyses described in it.

Ekker SC, Dobbs D (2019) Nucleic Acids Res. Webserver Issue). The version included here has been modified to match the format of this dissertation. This manuscript describes the Gene Sculpt Suite (www.genesculpt.org), a collection of three web tools (GTagHD, MEDJED, and MENTHU) created in conjunction with the McGrail (ISU), Essner (ISU), Clark (Mayo Clinic), and Ekker (Mayo Clinic) labs. Jordan Welker, Wes Wierson, and Maira Almeida, under the supervision of Jeff Essner and Maura McGrail, conceived the GeneWeld methodology, which GTagHD was designed to facilitate. Hiro Ata (Mayo Clinic), under the supervision of Stephen Ekker, conceived the methodology implemented by MENTHU. Gabriel Martínez-Gálvez and I conceived of and performed experiments required to create MEDJED. Gabriel Martínez-Gálvez helped program the TALEN and overhang-related features in MENTHU and contributed to the manuscript. Martínez-Gálvez, Welker, Wierson, Ata, Almeida, Clark, Essner, McGrail, Ekker, and Dobbs edited the manuscript. Dobbs, Essner, McGrail, Ekker, and Clark supervised and guided the research. I created the model used in the MEDJED web tool, performed feature selection, created the dataset used to train MEDJED, wrote most of the MENTHU webserver code, created the code used to analyze the data and wrote most of the manuscript. I am solely responsible for creating, administering, and maintaining the Gene Sculpt Suite webserver, creating and programming the GTagHD web tool, programming the MEDJED web tool, Docker-izing all of the web tools, and designing and writing the documentation and all content for the Gene Sculpt Suite webpages and GitHub READMEs.

Chapter 6 summarizes my contributions to the fields of RNA-protein interaction prediction and gene editing and discusses future directions for these areas of research.
Appendix A consists of an invited peer-reviewed methods chapter (Computational Prediction of RNA-Protein Interactions. Mann CM, Muppirala UK, Dobbs D. (2017) Methods in Molecular Biology: Promoter Associated RNA. Ed. Napoli S.) The manuscript reviews computational methods for predicting RNA-protein interaction partners and instructs readers in how to use these methods. I wrote the instructions for using the described methods and for interpreting their output. Usha Muppirala created figures and assisted in writing the introduction and instructions. Drena Dobbs wrote the introduction and notes.

Appendix B is a published manuscript entitled “Robust Activation of Microhomology-Mediated End Joining for Precision Gene Editing Applications” (Ata H, Ekstrom TL, Martínez-Gálvez G, Mann CM, Dvornikov AV, Schaefbauer KJ, Ma AC, Dobbs D, Clark KJ, Ekker SC (2018) PLoS Genet. 14(9) e1007652.) I created the MENTHU web tool, which implements the methodology described in the manuscript, and assisted in data analysis and processing. I wrote the section of the manuscript describing MENTHU, with input from Gabriel Martínez-Gálvez and Drena Dobbs. Martínez-Gálvez and I also performed validation tests of the MENTHU algorithm.

Appendix D consists of usage statistics for the Gene Sculpt Suite web tools, including page uses and global distribution of users.

References

http://dx.doi.org/10.1146/annurev-biochem-072711-164732

http://dx.doi.org/10.1038/nbt.4317


http://dx.doi.org/10.1038/nmeth.1611

http://dx.doi.org/10.1016/j.tig.2013.01.004

http://dx.doi.org/10.1016/j.cell.2009.02.011

http://dx.doi.org/10.4143/crt.2010.42.3.125
http://dx.doi.org/10.1016/j.gendis.2019.01.003

http://dx.doi.org/10.3389/fped.2019.00067

http://dx.doi.org/10.1016/j.tig.2008.05.004

http://dx.doi.org/10.1186/1471-2105-12-489

http://dx.doi.org/10.1093/nar/gkr160

http://dx.doi.org/10.1096/fj.201600637RR

http://dx.doi.org/10.1038/s41586-018-0686-x

http://dx.doi.org/10.1101/gad.261974.115

http://dx.doi.org/10.1101/gad.1800909
CHAPTER 2. RNA-PROTEIN INTERACTION PREDICTIONS WIKI PAGE

This chapter is a Wikipedia-style review in preparation for submission to the PLoS Genetic’s Topic Pages collection. Topic Pages are Wikipedia-style articles that are peer-reviewed, citable, and PubMed-indexed. After review at PLoS, they are also submitted to Wikipedia. The original PLoS peer-reviewed page is assigned a DOI and can be viewed in its original format even if the Wikipedia page is later amended. In this article, I describe the features and methods used in a subset of RNA-protein interaction prediction methods. The methods described were selected on the basis of open-source code availability and the quality of the paper describing the method. Thus, catRAPID [Belluci et al. 2011], a high-profile prediction method, is not described here because its code was only recently made available and requires submitting a license request which, amongst other considerations, requires the licensee to provide a tax identification number. Because this software is proprietary and not open source, it is not described in this entry.

Introduction

RNA-protein partner prediction is a field that uses bioinformatics and computational biology techniques to predict physical and/or chemical interactions between specific RNA and protein molecules (RNA-protein interaction “partners”, or RPIPs). RNAs and proteins that interact are often referred to as “binding”, and those that do not interact are referred to as “non-binding.” For recent reviews of RNA-protein interaction prediction methods, see [Si et al. 2015, Ferre et al. 2016, Mann et al. 2017, Zhang SW et al. 2017, Zhang H et al. 2019]. While it is also of interest to predict the amino acids and ribonucleotides directly interacting in an RNA-protein interaction (RPI), we refer to this as the “interface” prediction problem, and methods for predicting these interactions are not discussed here. For recent reviews of
RNA-protein interface prediction methods, see [Si et al. 2015, Xue et al. 2015, Walia et al. 2017, Jung et al. 2019].

Machine learning methods for predicting RNA-protein interactions work because they can “capture” signal for binding through judicious encoding of information about the protein and RNA. Predictive models are created by extracting this information (features) from known RNA-protein interactions and training a mathematical model (or creating a scoring system) to identify novel interactions based on these features. Here we discuss features and models used in open-source RNA-protein interaction prediction methods.

**Significance**

RNA-protein interactions (RPIs) play critical roles in numerous biological processes. These roles include (but are not limited to):

- The structural stabilization of functional RNA-protein complexes, including ribosomes and spliceosomes
- Gene transcription and translation
- Retrotransposon silencing through the action of Piwi-interacting RNAs (piRNA)
- mRNA localization, transport, and degradation

Disruptions in these interactions can therefore lead to a variety of diseases and disorders, including an array of cancers [Darnell 2010], fragile X syndrome [De Boulle et al. 1993], amyotrophic lateral sclerosis (ALS) [Zhao et al. 2018], and dyskeratosis congenita [Mason & Bessler 2011]. Thus, identifying disruptions in RNA-protein interactions provides an avenue for identifying potential new treatments that correct these disruptions.
There are many methods for experimentally characterizing RPIPs, including structural determination methods such as X-ray crystallography [Shi 2014], nuclear magnetic resonance (NMR) spectroscopy [Ke & Doudna 2004, Marion 2013, Yadav & Lukavsky 2016], small-angle x-ray scattering [Kikhney & Svergun 2015, Vestergaard 2016, Chen et al. 2018], cryogenic electron microscopy (cryo-EM) [Bai et al. 2015], and liquid chromatography with tandem mass spectrometry [Tacheny et al. 2013], and assays such as Northwestern blots, gel mobility shift assays, filter retention assays, RIP-ChIP [Keene et al. 2006], and various cross-linking immunoprecipitation (CLIP) [Ule et al. 2003] methods including high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP/CLIP-Seq) [Licatalosi et al. 2008], photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) [Hafner et al. 2010], and individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) [König et al. 2010]. Because it can be time consuming, difficult, and/or expensive to perform these experiments for all proteins and RNAs of interest, computational predictions of such interactions are valuable and can dramatically decrease the search space of biologically relevant potential interactions.

**Features**

Predicting RPIPs requires capturing sufficient information about both the protein and RNA in a way that allows the information to be used as input for machine learning algorithms or scoring systems to generate predictive models. Many different protein and RNA features have been utilized; features commonly used by prediction methods are described here.
**Sequence Based Features**

Structural data for RNA-protein complexes is currently limited, and can be difficult to obtain, so many prediction methods rely on sequence-derived features.

**Sequence composition**

Sequence composition is one of the most commonly used features in prediction models. Because a protein’s primary sequence ultimately determines its tertiary structure, the amino acid sequence of a protein, in conjunction with the RNA sequence to which it binds, contains signal that may be sufficient to predict an RNA-protein interaction. Sequence composition features can capture physical (e.g., size, shape) and chemical (e.g., charge, hydrophobicity) properties of both the protein primary structure and the RNA primary structure, which are important determinants of molecular recognition.

Most machine learning methods require feature vectors of a set length, i.e., each instance used to train the model must have the same features and interactions to be predicted must have the same features. The sequence composition is thus commonly captured using a frequency-based representation that does not vary with sequence length.

Reduced alphabets (see Fig. 2-1.2) are frequently used for representing protein sequences to help limit the size of the feature set to help to alleviate the curse of dimensionality. Common methods for alphabet reduction include defining groups of amino acids based on physical, chemical, or biochemical properties [Muppirala et al. 2011, Akbaripour-Elahabad et al. 2016, Pan et al. 2016, Mann et al. 2019], or groupings based on RNA interaction propensity [Jain et al. 2018].

Sequence composition for both proteins and RNAs has frequently been encoded using conjoint \(k\)-mers (see Fig. 2-1.3). In the conjoint k-mer representation, an \(N \times M\) feature set, \(S\),
is generated, where $N$ is the k-mer length, $M$ is the alphabet size, and $S_{nm}$ is the frequency of occurrence of the corresponding sequence.

Figure 2-1 Illustration of a conjoint triad representation of a protein sequence. 1) Drosophila peptide tarsal-less AA protein sequence [UniProt A3RLR1] represented using the 1-letter amino acid code. 2) Alphabet reduction using a 7-letter alphabet based on Muppirala et al. 2011 and Shen et al. 2007. Each amino acid group is represented by a colored block. 3) In a conjoint triad feature representation, each continuous block of three amino acids (p1, p2, etc.) is identified. 4) The frequency of each possible triad (f1, f2, etc.) is calculated by dividing the number of times a triad appears by the length of the sequence (e.g., p1 appears once in the 32-nt long sequence, so the corresponding feature (f27) is 1/32, or 0.03125)

Many published PRIP prediction methods utilize a conjoint triad representation (shown in Fig. 2-1) for the protein sequence [Muppirala et al. 2011, Akbaripour-Elahabad et al. 2016, Pan et al. 2016, Mann et al. 2019] and a conjoint tetrad representation for the RNA sequence [Muppirala et al. 2011, Akbaripour-Elahabad et al. 2016, Pan et al. 2016, Mann et al. 2019]. Some methods use a conjoint pentad representation for both RNA and protein [Jain...
et al. 2018], based on the number of amino acids and nucleotides typically involved in the binding interfaces in RNA-protein complexes.

**Sequence motifs**

Some methods make use of sequence motifs. In RPIP prediction, these short stretches of ribonucleotide or amino acids are generally extracted from the binding interfaces of structurally characterized RNA-protein complexes, and a binary feature indicating their presence or absence is employed.


**Hydrophobicity and hydrophilicity**

Hydrophobicity and hydrophilicity are important features of RNA-protein interactions and help capture the physical and chemical properties of protein sequences, which may provide signal for discerning partner-specific RNA-protein interactions.

The IncPro method uses hydrophobicity and hydrophilicity as an approximation for van der Waals forces and hydrogen bonding propensities of the amino acids in protein sequences [Lu et al. 2013].

**Structure-based Features**

Although structural data for ribonuleoproteins (RNPs) is limited at present, some methods utilize structure-based features.
**Protein secondary structure**

Protein secondary structure is used to capture potential signal from secondary structural elements such as alpha helices, beta sheets, and coils, based on the premise that these structural elements may reflect propensities for sequence-specific RNA binding. Some methods, e.g., Lu et al. 2013, use Chou-Fasman and/or Deleage-Roux [Deleage & Roux 1987] propensities of amino acids to identify and/or predict these secondary structural elements within the tertiary protein structure.

In cases where experimentally determined protein secondary structure information is not available, secondary structure can be predicted using computational methods.

**Protein disorder**

Intrinsic protein disorder plays an important role in mediating RNA-protein interactions [Järvelin et al. 2016]. RPIDisorder uses information including the location of disorder within protein sequences, as well as the hydrophobicity, hydrophilicity, and charges of the disordered and ordered amino acids in the sequence [Mann et al. 2019].

For proteins where disorder information is not available, disorder must be predicted.

**RNA secondary structure**

As with protein secondary structure, RNA secondary structure is frequently utilized for RPIP prediction under the assumption that if RNA binding is sequence specific, then features that distinguish ribonucleotides must be accessible to proteins for recognition. RNA secondary structure is frequently represented simply by encoding whether ribonucleotides are present in double- or single-stranded regions [Lu et al. 2013].

In cases where RNA secondary structure information is unavailable (i.e., most of the time), RNA secondary structure can be predicted using computational methods. Many RPIP
prediction methods use tools from the ViennaRNA software package to predict RNA secondary structure [Lu et al. 2013].

**Feature Dimensions**

Because prediction models require a set number of features, the feature values used in prediction must not vary with RNA or protein sequence length. This requires that features be “normalized” based on sequence length, use a frequency-based measure, or undergo some sort of vector transformation (e.g., IncPro uses Fourier Transform [Lu et al. 2013]), possibly with truncation, to ensure that every RNA-protein pair has the same number of features.

**Models**

**Machine Learning Methods**

**Random forests**

Random forest classifiers are very commonly used in RPI prediction, in part due to the easily-interpretable feature selection process, which allows for the examination of features important to the model for generating biological hypotheses [Muppirala et al. 2011, Mann et al. 2019].

**Gradient boosting**

Gradient boosting using decision trees has also been applied with success to the RPIP problem [Jain et al. 2018]. In this methodology, a weak classifier is made stronger iteratively by leveraging the residual error in subsequent classifiers to build a stronger model.

**Support vector machines**

Support vector machines (SVM) are frequently used in sequence-based biological prediction problems, e.g., [Muppirala et al. 2011]. While SVMs can separate interacting and non-interacting protein classes effectively, they are not easy to translate back into biological hypotheses.
Neural networks

As deep learning becomes more popular, the use of artificial neural networks for predicting RPIPs has increased.

Stacked auto encoders are a variation on convolutional neural networks. With auto encoders, the input features are encoded and then decoded into output with the goal of minimizing the error between input and output. In RPIP prediction, the encoder is given features derived from known RNA-protein interactions. The encoder is then trained to minimize error from the input through the encoding stages to the decoding stages to the output. Thus, when new instances are introduced, the encoder has been trained to assign weights to important interaction features, and then decodes the input by assigning labels to interacting vs non-interacting classes [Pan et al. 2016]. This method produces results similar to Principal Components Analysis (PCA), which can be used to reduce the complexity of and interpret high-dimensional datasets.

Multi-classifier methods

Some methods use a combination of various algorithms and architectures. While some use logistic regression functions to combine the output of multiple models, others use stacked ensembles, e.g., [Pan et al. 2016], in which a neural network is trained using the output of other prediction models as features.

Scoring Systems

Some RPIP prediction approaches use scoring systems rather than machine learning models.

For example, IncPro uses a matrix-based approach to combine related features, then maps the resulting scores to a 0-100 scale and uses the arithmetic mean to calculate a final score for the interaction [Lu et al. 2013].
Datasets

In order to create reliable prediction models, high quality datasets are required for training models and for testing their efficacy. These datasets are usually derived from existing RPIP structure or sequence databases.

Dataset Creation

Structure-derived datasets

Structure-derived datasets are extracted from three-dimensional structural information available for the relatively small number of RNA-protein complexes for which high resolution 3D structures are available. An RNA and protein are considered to “interact” if they possess heavy atoms (i.e., non-hydrogen atoms) that lie within a certain distance cutoff. Frequently used distances are 5 Å and 8 Å [Muppirala et al. 2011, Jain et al. 2018]. Shorter distances are sometimes employed based on the assumption that a shorter cutoff distance corresponds to a stronger bond, and thus a “stronger” interaction signal; longer distances are chosen to accommodate long-range and/or weaker interactions [Jain et al. 2018].

Structure-derived datasets have several limitations: to date, all published structure-based datasets contain RNA-protein complexes that can be crystallized and thus are considered to be biased towards more stable complexes and structurally ordered proteins and RNAs. Additionally, the technical difficulty, time, and expense involved in crystallizing RNA-protein complexes has limited the size of structure datasets. For example, as of April 04, 2019, the Protein Data Bank (PDB) [Berman et al. 2000], the largest structural database for RNP complexes, contained only 2461 complexes that included both protein and RNA chains (2852 complexes if structures that also contain DNA and DNA/RNA hybrids are included); of these, 1,037 correspond to ribosomes or ribosomal components. Because certain
complexes such as ribosomal and spliceosomal components are heavily represented in these datasets, redundancy removal further reduces the size of these datasets.

Many RNPs in the PDB also have “synthetic” components. For example, a protein may be crystallized in complex with a synthetic bait RNA (e.g., poly-U sequence), so that the resulting complex does not correspond to a naturally occurring interaction. Care must be taken to ensure that the training dataset contains biologically relevant interactions.

Datasets from high-throughput experiments

Other types of experimentally-validated datasets are extracted from published high-throughput experiments for which data are deposited in online databases. NPInter [Yuan et al. 2014] is one such database; the most current release (v3.0) contains 186,433 RNA-protein interactions curated from high- and low-throughput experiments [Hao et al. 2016].

Non-redundant datasets

To avoid biasing datasets towards a particular type of RNA-protein interaction, the generation of many datasets includes a redundancy removal step, in which proteins and RNAs are clustered into groups based on sequence identity, and only a single representative interaction from each group is retained [Muppirala et al. 2011, Jain et al. 2018]. For example, a protein from cluster P1 and an RNA from cluster R1 will be retained, but any subsequent P1- R1 pairings will be discarded. Conversely, a protein from cluster 1 can be included in multiple interactions, as long as the RNAs to which it is paired are from different clusters. Various sequence identity thresholds are used to reduce redundancy; some datasets use very stringent thresholds (e.g., sequence identity cannot be greater than 30% [Muppirala et al. 2011]), whereas others use very lax thresholds (100% identity [Jain et al 2018]).
Experimentally-validated negative training datasets

A few methods (not discussed here) have employed datasets that include negative instances (i.e., pairs of specific RNAs and proteins that do not interact) supported by experimental validation [Livi & Blanzieri 2014, Armaos et al. 2017]. Unfortunately, these datasets are only available upon author request [Armaos et al. 2017], are limited in scope to RNAs that interact with one or two proteins [Livi & Blanzieri 2014], and/or have some of the same problems as computationally-generated “pseudo-negative” datasets [Livi & Blanzieri 2014, Armaos et al. 2017]. “Pseudo-negative” datasets as those generated based on the assumption that because an interaction is not represented in the dataset from which the “positive” interactions are extracted, then that interaction does not occur. Some attempts have been made to use PDB structural data to generate negative datasets that may be “better” than those generated by simply randomly pairing the RNAs and proteins represented in the positive interaction dataset. Some datasets [Suresh et al. 2015] define RNA and protein chains as non-interacting if they come from the same structure, are non-redundant with interacting RNAs and proteins in the same structure, and do not themselves interact [Suresh et al. 2015]; others [Jain et al. 2018] also employ redundancy reduction via sequence identity thresholds when randomly pairing RNAs and proteins from the positive set. The latter approach still has some of the shortcomings of pseudo-negative datasets because they cannot guarantee that the RNA and protein in question do not interact, just that they do not interact in the particular complex examined. As always, the difficulty in generating high-quality negative datasets is that absence of evidence is not evidence of absence.

Publicly Available Datasets

Currently, there are several publicly available “benchmark” datasets (see Table 2-1), although no “gold-standard” positive (or negative) datasets exist at present.
Table 2-1 Table of publically available benchmark datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Derivation</th>
<th>Protein Type</th>
<th>RNA Type</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPI2241 [Muppirala et al. 2011]</td>
<td>Structure-derived</td>
<td>N/A</td>
<td>N/A</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>RPI369 [Muppirala et al. 2011]</td>
<td>Structure-derived</td>
<td>non-ribosomal</td>
<td>N/A</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>RPI1807 [Suresh et al. 2015]</td>
<td>Structure-derived</td>
<td>N/A</td>
<td>non-ribosomal</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>RPI13243 [Muppirala et al. 2011]</td>
<td>non-structure-derived</td>
<td>N/A</td>
<td>mRNA</td>
<td>[Hogan et al. 2008]</td>
</tr>
<tr>
<td>RPI367 [Wang et al. 2013]</td>
<td>non-structure-derived</td>
<td>N/A</td>
<td>ncRNA</td>
<td>NPInter v2.0</td>
</tr>
<tr>
<td>RPI10412 [Pan et al. 2016]</td>
<td>non-structure-derived</td>
<td>N/A</td>
<td>ncRNA</td>
<td>NPInter v2.0</td>
</tr>
</tbody>
</table>
Methods

Because of the lack of gold-standard datasets at present, it is difficult to rigorously compare the prediction performance of different RPIP methods. Often, several methods have been trained on the same benchmark dataset, which is used as a test dataset by other methods and vice versa. Compounding this problem is the tendency for researchers to compare the 10-fold cross-validation results for their own method with results obtained on independent test sets for other methods [Lu et al. 2013, Akbaripour-Elahabad et al. 2016, Pan et al. 2016]. This is problematic because a method trained on a particular dataset will almost always outperform methods that were not trained on that particular dataset simply because the method trained on the dataset is generally tuned to that specific dataset, regardless of the method’s generalizability to other datasets. Thus, results from such flawed evaluation methods tend to over-estimate the performance of the method.

Independent test sets have been appropriately used to compare methods, but these test sets frequently include a very limited number of proteins and/or RNAs. It is thus difficult to gauge whether a method is in fact “the best” in predicting RPIPs or simply has the best performance for a particular RNA and/or protein.

Finally – and unfortunately - many methods are not available as a webserver, and/or code for running the method offline is not freely available, making it impossible to replicate or verify the performance of the method or to rigorously test it against other methods; such methods are not included here. See Table 2-2 for a list of methods for which a webserver is provided or code is publically available.
Table 2-2 Table of RNA-protein interaction prediction methods with publically available code

<table>
<thead>
<tr>
<th>Method Name</th>
<th>RNA Class</th>
<th>Features</th>
<th>Model</th>
<th>Webserver</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPMiner</td>
<td>(lncRNA focused)</td>
<td>sequence composition via conjoint polyads</td>
<td>Stacked Auto Encoder with random forest combined through logistic regression</td>
<td>N/A</td>
<td><a href="https://github.com/xypan1232/IPMiner">https://github.com/xypan1232/IPMiner</a></td>
</tr>
<tr>
<td>rpiCool</td>
<td>all RNAs</td>
<td>combines RNA-binding protein motifs and protein-binding RNA motifs with conjoint polyad representation and reduced alphabet</td>
<td>Random Forest</td>
<td>N/A</td>
<td><a href="http://biocool.ir/rpicool.html">http://biocool.ir/rpicool.html</a></td>
</tr>
<tr>
<td>RPIDisorder</td>
<td>all RNAs</td>
<td>sequence composition via conjoint polyads; predicted disordered protein regions</td>
<td>Random Forest</td>
<td><a href="http://www.rpidisorder.org">www.rpidisorder.org</a></td>
<td><a href="http://www.rpidisorder.org">www.rpidisorder.org</a></td>
</tr>
<tr>
<td>RPISeq</td>
<td>all RNAs</td>
<td>conjoint polyad representation</td>
<td>Support Vector Machines Random Forest</td>
<td><a href="http://pridb.gdcb.iastate.edu/RPISeq/">http://pridb.gdcb.iastate.edu/RPISeq/</a></td>
<td>N/A</td>
</tr>
<tr>
<td>XRPI</td>
<td>all RNAs</td>
<td>conjoint polyad representation</td>
<td>gradient boosting</td>
<td><a href="http://xrpi.ddns.net/">http://xrpi.ddns.net/</a></td>
<td><a href="https://universe.bits-pilani.ac.in/goa/aduri/xRPI">https://universe.bits-pilani.ac.in/goa/aduri/xRPI</a></td>
</tr>
</tbody>
</table>
Databases of Known Interactions

Several valuable databases contain known RNA-protein interactions. These databases can be grouped into structure and sequence-based databases. The structure databases contain 3D structural information (i.e., atomic coordinates) for the protein and RNA in the complex. Sequence-based databases generally contain identifiers or sequences for RNAs and proteins found to bind each other in high-throughput experiments.

Structure-based Databases:

Protein Data Bank

The Protein Data Bank (PDB) [Berman et al. 2000] contains structures for proteins, nucleic acids, and protein-nucleic acid complexes. As of April 4 2019, it included 2,852 RNA-protein complexes. Most of these structures were generated using x-ray crystallography, electron microscopy, or NMR.

Nucleic Acid Database

The Nucleic Acid Database (NDB) [Berman et al. 1992, Coimbatore Narayanan et al. 2014] contains nucleic acid structures, many in complex with proteins or ligands. As of April 4, 2019, it included 2,257 RNA-protein complexes. Most of these structures are generally also represented in the PDB.

Sequence-based Databases:

ENCODE

The ENCODE (Encyclopedia of DNA Elements) database contains the results of more than 40 types of high-throughput assays from more than 14,600 experiments [The Encode Consortium 2012]. While the database consists primarily of DNA binding and transcription assays, it also includes the results from ~700 RNA-binding assays. The results of these assays can be mined for RNA-protein interactions.
**GEO**

The Gene Expression Omnibus database [Edgar et al. 2002] contains the results of high-throughput expression analyses. It contains more than 4,300 data sets, many of which are RNA-binding or protein-binding assays. These datasets can be downloaded and mined for RNA-protein interactions.

**NPInter**

The NPInter v3.0 database contains RNA-RNA, RNA-DNA, and RNA-protein interactions curated from high-throughput datasets and literature mining [Hao et al. 2016]. It contains more than 186,000 RNA-protein interactions. It is possible to download interaction data for specific RNAs, proteins, and organisms. Many datasets used for RPIP prediction have been derived from the NPInter v2.0 database [Yuan et al. 2014], the predecessor to the NPInter v3.0 database [Hao et al. 2016].

**POSTAR2**

The POSTAR2 database (formerly known as CLIPdb and POSTAR) contains high-throughput experimental data specifically focused on RNA-protein interactions from six species [Zhu et al. 2019]. The database is derived from ENCODE RNA- and protein-binding assays, which are processed by POSTAR2 to identify interacting RNAs and proteins. POSTAR2 includes data for more than 280 RNA-binding proteins.

**UniProt**

The UniProt database provides protein sequence, structural, and functional information [The UniProt Consortium 2019]. While UniProt was not created with the goal of collating RNA-protein interactions, it is possible to search the database using Gene Ontology terms corresponding to RNA-binding functions, and thus identify RNA-binding proteins and their RNA-binding partners.
See Also

RNA-binding protein
RNA-binding protein database
DNA-binding protein
Protein-protein interaction
Protein-protein interaction prediction

Methods published in suspected predatory journals or journals run by suspected predatory publishers were not considered. A journal or publisher was considered to be “suspected predatory” if it appeared on Jeffrey Beall’s list (https://beallslist.weebly.com/), the Stop Predatory Journals site (https://predatoryjournals.com), or if it appears to have been removed from Beall’s list due to threatening Beall (MDPI, Frontiers). Methods that do not offer standalone code or whose standalone code was difficult to obtain (e.g., catRAPID [Belluci et al. 2011]) were also excluded.

Thus, while there are >30 published RNA-protein partner prediction methods, this article discusses only RPISeq [Muppirala et al. 2011], lncPro [Lu et al. 2013], IPMiner [Pan et al. 2016], rpiCool [Akbaripour-Elahabad et al. 2016], XRPI [Jain et al. 2018], and RPIDisorder [Mann et al. 2019]. See Table 2-2 for a description of these methods. Please see Table 2-3 for a full list of published methods.
Table 2-3 Table of published RNA-protein interaction prediction methods. This table will not be included in the Wiki article.

<table>
<thead>
<tr>
<th>Method Name</th>
<th>RNA Class</th>
<th>Features</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>catRAPID</td>
<td>IncRNAs</td>
<td>protein secondary structure, RNA secondary structure, van der waals approximation, hydrogen bonding approximation</td>
<td><a href="https://www.nature.com/articles/nmeth.1611">https://www.nature.com/articles/nmeth.1611</a></td>
</tr>
<tr>
<td>catRAPID omics</td>
<td>IncRNAs</td>
<td>catRAPID, catRAPID fragments</td>
<td><a href="https://academic.oup.com/bioinformatics/article/29/22/2928/114358">https://academic.oup.com/bioinformatics/article/29/22/2928/114358</a></td>
</tr>
<tr>
<td>CFRP</td>
<td>ncRNA</td>
<td>sequence composition, geometric and harmonic mean and power operation features of the conjoint n-mer representations</td>
<td><a href="https://www.frontiersin.org/articles/10.3389/fgene.2019.00018/full">https://www.frontiersin.org/articles/10.3389/fgene.2019.00018/full</a></td>
</tr>
<tr>
<td>CTF-ICGR</td>
<td>RPISeq and &quot;chaos game representation&quot;</td>
<td></td>
<td><a href="https://www.tandfonline.com/doi/full/10.1080/21655979.2018.1470721">https://www.tandfonline.com/doi/full/10.1080/21655979.2018.1470721</a></td>
</tr>
<tr>
<td>LPMNI</td>
<td>IncRNAs</td>
<td><a href="https://www.sciencedirect.com/science/article/pii/S1672022916004137">https://www.sciencedirect.com/science/article/pii/S1672022916004137</a></td>
<td></td>
</tr>
<tr>
<td>LPI-ETSLP</td>
<td>IncRNAs</td>
<td><a href="https://pubs.nrc.ca/Content/ArticleLanding/2017/M/BC/7/MX02900#ot23">https://pubs.nrc.ca/Content/ArticleLanding/2017/M/BC/7/MX02900#ot23</a></td>
<td></td>
</tr>
<tr>
<td>LPHN</td>
<td>IncRNAs</td>
<td>RNA similarity network protein-protein interaction network RNA-protein similarity network</td>
<td><a href="https://www.hindawi.com/journals/bmri/2015/671950/">https://www.hindawi.com/journals/bmri/2015/671950/</a></td>
</tr>
<tr>
<td>LPI-NRMLF</td>
<td>IncRNAs</td>
<td><a href="https://doi.org/10.13831/emotarget.21934">https://doi.org/10.13831/emotarget.21934</a></td>
<td></td>
</tr>
<tr>
<td>omiXcore</td>
<td>same as catRAPID signature to identify RNA-binding proteins, then identify interaction regions</td>
<td></td>
<td><a href="https://academic.oup.com/bioinformatics/article/33/19/3104/3868478">https://academic.oup.com/bioinformatics/article/33/19/3104/3868478</a></td>
</tr>
<tr>
<td>Pancaldi and Bahler</td>
<td>mRNAs</td>
<td></td>
<td><a href="https://academic.oup.com/nar/article/38/14/5806/137145">https://academic.oup.com/nar/article/38/14/5806/137145</a></td>
</tr>
<tr>
<td>PRPBU</td>
<td>IncRNAs</td>
<td></td>
<td><a href="https://www.sciencedirect.com/science/article/pii/S0022519318305058">https://www.sciencedirect.com/science/article/pii/S0022519318305058</a></td>
</tr>
<tr>
<td>rpCool</td>
<td></td>
<td>same as catRAPID signature to identify RNA-binding proteins, then identify interaction regions</td>
<td></td>
</tr>
<tr>
<td>RPI-Pred</td>
<td>ncRNA</td>
<td>protein sequence/structure composition via protein blocks combined with reduced alphabet monad representation RNA sequence/secondary composition via RNA structures combined with monad representation</td>
<td></td>
</tr>
<tr>
<td>RPISeq</td>
<td>N/A</td>
<td>conjoint polyad representation</td>
<td></td>
</tr>
<tr>
<td>Wang method</td>
<td>N/A</td>
<td>sequence composition via conjoint polyads</td>
<td></td>
</tr>
<tr>
<td>XRPI</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References

http://dx.doi.org/10.1016/j.jtbi.2016.04.025

http://dx.doi.org/10.1093/bioinformatics/btx361

http://dx.doi.org/https://doi.org/10.1016/j.tibs.2014.10.005

http://dx.doi.org/10.1093/nar/gks1193

http://dx.doi.org/10.1038/nmeth.1611

http://dx.doi.org/10.1016/s0006-3495(92)81649-1


http://dx.doi.org/10.1016/j.cell.2012.04.031

http://dx.doi.org/10.1021/acscombsci.8b00007
http://dx.doi.org/10.1093/nar/gkt980

http://dx.doi.org/10.4143/crt.2010.42.3.125

http://dx.doi.org/10.1038/ng0193-31


http://dx.doi.org/10.1038/nature11247

http://dx.doi.org/10.1093/bib/bbv031

http://dx.doi.org/https://doi.org/10.1016/j.cell.2010.03.009

http://dx.doi.org/10.1093/database/baw057

http://dx.doi.org/10.1371/journal.pbio.0060255
http://dx.doi.org/10.1038/s41598-018-27814-2

http://dx.doi.org/10.1186/s12964-016-0132-3

http://dx.doi.org/10.1371/journal.pcbi.1003075

http://dx.doi.org/10.1002/prot.25639

http://dx.doi.org/https://doi.org/10.1016/j.ymeth.2004.03.027

http://dx.doi.org/10.1038/nprot.2006.47

http://dx.doi.org/10.1016/j.febslet.2015.08.027

http://dx.doi.org/10.1038/nsmb.1838

http://dx.doi.org/10.1002/prot.20607

http://dx.doi.org/10.1038/nature07488
http://dx.doi.org/10.1186/1471-2105-15-123

http://dx.doi.org/10.1186/1471-2164-14-651

http://dx.doi.org/10.1007/978-1-4939-6716-2_8


http://dx.doi.org/10.1074/mcp.O113.030239

http://dx.doi.org/10.1016/j.cancergen.2011.11.002

http://dx.doi.org/10.1186/s12864-016-2931-8

http://dx.doi.org/10.1093/nar/gkr160

http://dx.doi.org/10.1038/nature12311

http://dx.doi.org/10.1016/j.ccell.2016.03.010


http://dx.doi.org/10.3390/ijms161125952


http://dx.doi.org/10.1093/nar/gkv020


http://dx.doi.org/https://doi.org/10.1016/j.jprot.2013.09.011


http://dx.doi.org/10.1126/science.1090095


http://dx.doi.org/10.1093/nar/gky1049


http://dx.doi.org/https://doi.org/10.1016/j.abb.2016.02.029


http://dx.doi.org/10.1007/978-1-4939-6406-2_15


http://dx.doi.org/10.1039/c2mb25292a


http://dx.doi.org/10.1016/j.febslet.2015.10.003

http://dx.doi.org/https://doi.org/10.1016/j.pnmrs.2016.10.001


http://dx.doi.org/10.1093/nar/gkt1057


http://dx.doi.org/10.3390/ijms20061284


http://dx.doi.org/10.2174/1573406413666170510102405


http://dx.doi.org/10.14348/molcells.2018.0243


http://dx.doi.org/10.1093/nar/gky830
CHAPTER 3. RPIDisorder: A METHOD FOR PREDICTING RNA-PROTEIN PARTNERS USING INTRINSIC PROTEIN DISORDER

Abstract

RNA-protein interactions play crucial structural and regulatory roles in biological systems. Disruptions in RNA-protein interactions are associated with a variety of cancers and diseases and identifying RNAs and protein that bind each other offers a pathway for devising therapeutic corrections for these diseases. Computational methods for predicting RNA-protein interactions can aid in identifying relevant biological interactions.

RPIDisorder is a novel method for predicting RNA-protein interaction partners using sequence composition and intrinsic protein disorder features. Intrinsic protein disorder is implicated in sequence-specific RNA-binding, and while disorder information has previously been used to predict RNA-protein interfaces, it has not previously been utilized in predicting RNA-protein binding partners. In this study, we found that disorder information enhances prediction specificity while maintaining a level of sensitivity comparable to other state-of-the-art methods. On an independent dataset of 11,281 interacting and 971 non-interacting RNA-protein partners, RPIDisorder achieved a True Positive Rate (TPR) of 94.3%, Precision 95.6%, Matthews Correlation Coefficient (MCC) 68.4%, AUC 93.5%, and False Positive Rate (FPR) of just 14.4%. RPIDisorder is available online at www.rpidisorder.org, and available for download at https://github.com/Dobbs-Lab/RPIDisorder/.

Introduction

RNA-Protein Interactions Play Important Biological Roles

RNA-protein interactions (RPIs) play critical regulatory, structural, and functional roles in cellular biology [Re et al. 2013]. In addition to their vital role in assembling and stabilizing functional ribonucleoprotein complexes (RNPs) such as ribosomes, spliceosomes,

RPIs also play crucial roles in immune response. The interaction of lincRNA-Cox2 (long intergenic non-coding RNA Cox2) with the RNP A/B and A2/B1 regulates inflammatory response to infection [Carpenter et al. 2013], CRISPR/Cas bacterial immune systems bind to guide RNAs in order to target viral invaders [Wiedenheft et al. 2011], and RNA interference relies on the formation of the RNA-Induced Silencing Complex (RISC) to target pathogen mRNAs for degradation [reviewed in Susi 2007]. Additionally, RPIs play crucial roles in the replication of RNA viruses; e.g., HIV-1 mRNAs cannot be exported out of an infected cell nucleus without the export activity of the Rev protein, which binds to the RNA’s rev response element (RRE) [Fischer et al. 1995].

**Examples of Disruptions in Regulatory RNA-Protein Interaction Networks that Lead to Disease**

Disruption of RNA-protein interaction networks have been implicated in numerous human diseases [Khalil & Rinn 2011]. For example, mutations in RNA-binding proteins
involved in RNA metabolism have been linked to amyotrophic lateral sclerosis (ALS) [Zhao et al. 2018].

In vertebrates, the protein dyskerin binds to telomerase RNA (TERC) at an H/ACA snoRNA-like sequence near its 3’ end. Mutations in the dyskerin PUA RNA-binding domain (and in a region near the N-terminus that folds into close proximity with the PUA domain), as well as mutations in the H/ACA sequence in TERC, can lead to dyskeratosis congenita, a progressive degenerative illness caused by shortening of the telomeres [Mason & Bessler 2011]. The ncRNA Hox Transcript Antisense RNA (HOTAIR) 5’ region binds to the Polycomb Repressive Complex 2 (PRC2) protein, while HOTAIR’s 3’ region binds to the LSD1/CoRST/REST complex; this complex coordinates histone modifications to silence Hox-D locus genes [Tsai et al. 2010]. Disruptions in HOTAIR regulation lead to the silencing of genes controlling suppression of metastasis; high HOTAIR expression levels are strongly associated with metastasis and mortality in breast cancer [Gupta et al. 2010].

Fragile X Mental Retardation Protein (FMRP) binds to a host of mRNAs; meta-analysis of four studies [Ashley et al. 1993, Ashley et al. 1993, Brown et al. 2001, Miyashiro et al. 2003] revealed that FMRP binds ~4% of the transcriptome in mouse brains [Blackwell et al. 2010]). Among these targets, 40 have been summarized by Pasciuto and Bagni [Pasciuto & Bagni 2014], and include the mRNAs encoding FMRP, serine/threonine-protein kinase LMTK1 (gene AATK), amyloid precursor protein (APP), activity-regulated cytoskeleton-associated protein (ARC), calcium/calmodulin-dependent protein kinase type II alpha chain (CAMK2A), postsynaptic density protein 95 (DLG4), gamma-aminobutyric acid receptor subunit delta (GABRD), voltage-gated potassium channel subunit KV3.1 and Kv4.2 (KCNC1 and KCND2), microtubule-associated protein 1B (MAP1B), myelin basic protein
(MBP), and matrix metalloproteinase 9 (MMP9). Every one of these proteins has been linked to neurological disorders, including but not limited to: fragile X syndrome (FXS), autism, schizophrenia, bipolar disorder, Parkinson’s disease, Alzheimer’s disease, and epilepsy, and/or cancer or immune disorders [Pasciuto and Bagni 2014].

A single point mutation in the FMR1 gene coding the KH2 RNA-binding domain of FMRP induces an I304N missense mutation sufficient to cause Fragile X Syndrome (FXS) [DeBoulle et al. 1993], although FXS is most commonly caused by a trinucleotide repeat expansion in the 5’-UTR region, which leads to transcriptional repression of the FMRP protein [Verkerk et al. 1991, Yu et al. 1991, Moore et al. 2004]. Interestingly, the I304N mutation is likely pathogenic because, rather than abrogating FMRP binding to its mRNA target, it appears to prevent FMRP mRNA association with polyribosomes [Feng 1997, Mazroui 2003]. Since FMRP regulates mRNA through suppressing mRNA translation, its inability to associate with polyribosomes leads to runaway translation of FMRP-mRNA regulation targets, and subsequently causes FXS [Phan et al. 2011]. The mRNAs FMRP regulates, in contrast, appear to be determined by interaction of the RGG box in the FMRP C-terminus with complex secondary structures in the target mRNAs (particularly G-rich complex secondary structures) [Ozdilek 2017], including G-quadruplexes [Phan et al. 2011]. Interestingly, FMRP binds the FMRP-mRNA in the purine-rich region encoding the RGG box in a high-affinity interaction that is most likely mediated by the FMRP RGG domain [Schaefferet al. 2001]. Thus, in a truly poetic example of RNA-protein auto-regulation, the binding of the RGG region within the FMRP protein to the RGG encoding region of the FMRP mRNA allows FMRP to suppress the production of additional FMRP.
Intrinsic Protein Disorder May Play a Role in Determining RNA-Protein Interaction Specificity

Intrinsically disordered proteins (IDPs) are proteins that lack any fixed structure [reviewed in Uversky 2017]. Some proteins may be conditionally disordered – that is, they adopt a fixed structure only under certain conditions [Hausrath & Kingston 2017]. Also, proteins can contain intrinsically disordered regions (IDRs), which are stretches of amino acids that lack a fixed structure within an otherwise ordered protein structure [Oldfield & Dunker 2014].

IDRs have previously been established to play important roles in RNA-protein interactions [reviewed in Järvelin et al. 2016], including those whose dysregulation is implicated in disease states [Uversky et al. 2008]. Tompa and Csermely found that IDRs are often found in RNA chaperone proteins, and that loss of these IDRs leads to loss of chaperone function [Tompa & Csermely 2004]. Järvelin et al. collated more than 40 proteins that interface with RNA through an IDR, demonstrating the importance of these regions in RNA-binding [Järvelin et al. 2016], and some RNA-protein interface predictors have utilized protein disorder to great effect [Peng & Kurgan 2015]. Buljan et al. showed that IDPs frequently correspond to non-constitutively expressed exons [Buljan et al. 2012], indicating that alternative splicing of proteins could drastically alter RNA-binding properties and thus enhance their functional diversity [Niklas et al. 2015]. Complexes of RNA with disordered proteins often form membrane-less organelles (RNP bodies), including stress granules and Cajal bodies, and these couplings appear to drive the formation of some liquid-liquid phase separations [Brangwynne et al. 2015]. Castello et al. showed that disordered regions enhance sequence-specific RNA binding [Castello et al. 2012], likely due in part to the conformational flexibility of the disordered protein regions. Additionally, IDPs generally
contain several stretches of charged amino acids; these regions are frequently negatively charged and uncompensated, resulting in a high net charge [Uversky 2019], which may further aid in mediating RPIs. Taken altogether, these advances indicate information from disordered protein regions could therefore provide valuable signal for predicting RNA-protein interactions with enhanced specificity.

**Why Predict RPIs?**

Although there are several methods available for experimentally characterizing RPIs (including high-throughput methods described in Chapter 2), these methods can be time consuming, difficult to implement and scale, and expensive. Most structure-based methods [reviewed in Jones 2016] (including NMR spectroscopy [reviewed in Marion 2013, Yadav & Lukavsky 2016], x-ray crystallography [reviewed in Shi 2014, Ke & Doudna 2004], small-angle x-ray scattering [reviewed in Kikhney & Svergun 2015, Vestergaard 2016, Chen et al. 2018], and cryo EM [reviewed in Bai et al. 2015]) are not high-throughput and thus can’t be scaled effectively. Non-structure-based methods such as RIP-ChIP [Keene et al. 2006], CLIP [Ule et al. 2003] and its variants, liquid chromatography with tandem mass spectrometry [Tacheny et al. 2013], and gel mobility shift assays [Gagnon & Maxwell 2010] are more high-throughput, but even these methods can only identify interactions between a single protein and a transcriptome, or a single RNA and a proteome. The GRCh38.p13 human reference genome assembly contains 20,465 coding genes, 22,229 non-coding genes, and 208,689 gene transcripts [Zerbino et al. 2018]; these numbers imply ~4.7 billion potential RNA-protein interactions in the human interactome alone. The number of proteins to be screened for interactions cannot simply be reduced by eliminating proteins that lack a canonical RNA-binding domain (RBD), as many proteins without a known RBD still bind RNA [Baltz et al. 2012, Castello et al. 2012, Moore et al. 2018].
Thus, the sheer scale of identifying all RNA-protein interaction partners necessitates the development of computational prediction methods. Computational methods can substantially decrease the search space of possible RNA-protein interactions, allowing researchers interested in a particular biological process or disease to focus their investigative efforts. Analysis of the computational models can aid in efforts to generate testable hypotheses to identify underlying biological mechanisms of RNA-protein recognition.

Available RPI Prediction Methods

A plethora of RPI prediction methods are available, including methods for predicting interfacial residues in RNA-protein complexes (the interface prediction problem) [reviewed in Si et al. 2015, Xue et al. 2015, Walia et al. 2017, Jung et al. 2018] and methods for predicting whether or not a given RNA-protein pair interacts (the partner prediction problem) [recently reviewed in Zhang & Fan 2017, Moore & ‘t Hoen [sic] 2019]. This research focuses on the partner prediction problem, for which a Wiki-style review is provided in Chapter 2 (see Table 2-2 for an overview of available methods) and a published Methods in Molecular Biology protocol is provided in Appendix A. In general, existing methods for predicting RNA-protein interaction partners (RPIPs) perform adequately, but there is considerable room for improvement - especially in the specificity of predictions - if these are to be of practical utility in identifying and characterizing RNA-protein interaction networks.

At present, it is difficult to judge which published methods are “best.” Many authors have compared the results of their n-fold cross-validation experiments on a particular dataset to the performance of other methods on the same dataset (which did not benefit from n-fold cross-validation on that same dataset). This type of comparison can be very misleading because a method trained on a particular dataset is expected to outperform a method that has not been trained using that dataset.
Many published methods are specifically focused on predicting ncRNA-protein interactions (including ACCBN [Zhu et al. 2019], catRAPID [Bellucci et al. 2011], catRAPIDomics [Agostini et al. 2013], CFRP [Dai et al. 2019], IPMiner [Pan et al. 2016], IRWNRLPI [Zhao et al. 2018a], IncPro [Lu et al. 2013], LPBNI [Ge et al. 2016], LPI-BNRPA [Zhao et al. 2018b], LPI-ETSLP [Hu et al. 2017], LPI-FKLKRRR [Shen et al. 2019], LPIHN [Li et al. 2015], LPI-NRLMF [Liu et al. 2017], LPLNP [Zhang et al. 2018], and RPI-PRED [Suresh et al. 2015]). While ncRNA interactions are certainly important in many regulatory roles, gene expression is also subject to extensive regulation by both non-specific [Rissland 2017] and sequence-specific mRNA-protein interactions [Donlin-Asp et al. 2017], so methods for predicting interactions of mRNAs with proteins should not be neglected.

Here we present RPIDisorder, a method for predicting RNA-protein interaction using both sequence composition and information regarding intrinsically disordered protein regions. We compare the performance of RPIDisorder to RPISeq, a previously published prediction method that uses sequence composition features alone [Muppirala et al. 2011], and to RPIMotif, an unpublished method that outperforms RPISeq [Walia 2014] by exploiting interfacial motifs drawn from the Protein Data Bank (PDB) [Berman et al. 2000], in conjunction with sequence composition features. Compared to RPISeq and RPIMotif, RPIDisorder has comparable performance in detecting interacting RNA-protein partners but is much better at also correctly identifying non-interacting RNA-protein pairs than either RPISeq or RPIDisorder, demonstrating its superior utility for detecting biologically relevant RPIPs. To illustrate this, in addition to benchmark performance evaluations, we describe a case study on the clinically important and highly disordered protein, FMRP (fragile X mental
retardation protein), in which RPIDisorder outperforms both RPISeq and RPIMotif in correctly identifying its interactions with RNA.

**Methods**

**Datasets**

**RPI2241 structure-derived dataset (RPI-PDB)**

For training, we used the RPI2241 structure-derived dataset (RPI-PDB) created by Muppirala et al. [Muppirala et al. 2011]. The RPI-PDB dataset includes a total of 2241 positive interactions derived from 952 protein and 443 RNA interacting chains, which were extracted from 943 unique PDB structures (extracted prior to 2012) [Muppirala et al. 2011]. An RNA and protein chain were considered “interacting” if a ribonucleotide and amino acid possessed atoms within an 8Å cutoff distance of each other. The negative interaction set of 2241 interactions was created by randomly pairing the proteins and RNAs in the positive set and removing any known interacting pairs. These positive and negative datasets have been used as training sets for several RPIP prediction methods, including RPIMotif [Walia 2014]

**RPI12252* NPInter-derived dataset (RPI-NPInter*)**

For validation, we used the RPI12252 high-throughput dataset (RPI-NPInter). This dataset was created by Walia [Walia 2014] and contains 11,281 interacting RNA-protein pairs and 971 non-interacting pairs. The positive (interacting) set was derived from the NPInter v2.0 database [Yuan et al. 2014]. RPI-NPInter, in contrast to RPI-PDB, contains a large number of interactions identified through high-throughput experiments or curated from the literature. The negative examples in the dataset were generated by pairing 971 proteins determined to be non-RNA binding by Kumar et al. [Kumar et al. 2011], with an equal number of non-coding RNAs from the Functional RNA Database (fRNAdb) [Kin et al. 2007].
Results reported here were obtained using RPI-NPInter, except that a single interaction from the interacting set was removed because DISOPRED3 was unable to make predictions on it (the reason for this has not yet been determined.) The dataset lacking this single interaction is designated RPI12252*, or for simplicity and readability, RPI-NPInter*.

**FMRP119 case-study test dataset (RPI-FMRP)**

We created a test dataset for a detailed case study to demonstrate the efficacy of RPIDisorder in predicting interactions for a highly disordered protein, as well as its ability to predict protein interactions not just with ncRNA, but with mRNA as well. FMRP (see Introduction) contains a highly disordered RGG domain which is necessary for highly specific interactions with its regulatory mRNA targets. Thus, FMRP makes for an ideal case study.

The FMRP119 (RPI-FMRP) dataset consists of 75 mRNA 5’-UTR, CDS, or 3’-UTR regions bound by FMRP, and 44 negative examples created by identifying unbound 5’-UTR, CDS, and 3’-UTR regions in mRNAs containing another region bound by FMRP; these interactions were collected from a PAR-CLIP [Hafner et al. 2010] dataset created by Ascano et al. [Ascano et al. 2012] and collated by Anderson et al. [Anderson et al. 2016].

We partnered the 75 mRNA regions obtained from the Ascano study with a truncated FMRP protein corresponding to the amino acid sequence encompassing residues 434-632 of the FMRP Isoform-1 sequence, which is highly disordered [The UniProt Consortium 2019]. The mRNA-binding RGG domain spans residues 527 to 552 in the canonical sequence [Vasilyev et al. 2015]. Using the full-length FMRP sequence leads to predictions of indiscriminate binding, while using the C-terminal region containing the RGG box leads to sequence-specific predictions. The appropriateness of this approach is discussed in Results and Discussions.
Feature Representation

Sequence composition

We utilized 599 features initially proposed by [Muppirala et al. 2011] and subsequently used in numerous studies (e.g., CFRP [Dai et al. 2019], CTF-CGR [Wang et al. 2018], RPITER [Peng et al. 2019], Wang method [Wang et al. 2013]), to encode the sequence composition of the protein and RNA sequences (see Supp. Tables 3-1 and 3-2 for a complete list of these protein and RNA features). In this representation (originally used in protein-protein interaction prediction [Shen et al. 2007]), the protein sequence complexity is diminished by representing it using a reduced alphabet of 7-letters based on dipole moment, sidechain volume, and di-sulfide bonding: \{Ala, Gly, Val\}, \{Ile, Leu, Phe, Pro\}, \{Tyr, Met, Thr, Ser\}, \{His, Asn, Gln, Tpr\}, \{Arg, Lys\}, \{Asp, Glu\}, and \{Cys\} [Shen et al. 2007]. A sliding window of 3 amino acids (representing a conjoint triad) is then applied to the entire reduced sequence, and the frequency of each triad is calculated (see Fig. 2-1 in Chapter 2). The protein is thus represented by a 343-length vector (7 x 7 x 7). The RNA sequence is represented using conjoint tetrads for a vector of length 256 (4 x 4 x 4 x 4). Muppirala et al. chose these lengths because they worked very well; subsequent groups have similarly found these k-mer lengths useful [Muppirala et al. 2011].

Protein disorder prediction

Rather than using structural information present in the PDB to identify IDRs, we used IDRs predicted using DISOPRED3 software [Jones and Cozzetto 2015] in training and testing our model. We chose this approach because proteins and/or protein regions that are disordered in their unbound state may become structured upon binding to RNA. This phenomenon is known to occur in many RNPs, including ribosomal RNA-protein interactions [Draper & Reynaldo 1999] and fragile-X mental retardation protein (FMRP), in
which the highly disordered C-terminus region has been crystallized only in conjunction with RNA [Vasilyev et al. 2015]; even NMR visualization required stabilization with RNA [Phan et al. 2011]. Thus, using structural data in the PDB from RNA-protein complexes may fail to accurately capture disorder information. Additionally, most RNA-protein complexes have no structural data available and future predictions would require predicting protein disorder. The signal captured by DISOPRED3 in the training dataset may translate better to subsequent predictions.

DISOPRED3 software was chosen for predicting protein disorder because it: 1) performed well in multiple Critical Assessment of Protein Structure Prediction (CASP) challenges (ranked 2nd in CASP9 [Monastyrskyy et al. 2011] and CASP10 [Monastyrskyy et al. 2014]), 2) is actively maintained and offline-code is available (which is useful for creating prediction methods), 3) predicts disordered and ordered amino acids with residue-level resolution.

Protein disorder features

We developed a set of features that capture the location of predicted disorder within the protein and the physicochemical properties of the disordered residues. The rationale for using the location of IDRs within the linear sequence is that proteins that bind RNAs in a sequence-specific manner may have IDRs in regions distinct from those proteins that non-specifically bind RNAs. Physicochemical properties of amino acids were used because sequence-specific RNA-protein interactions generally require that amino acids contact the ribonucleotide face rather than the backbone (or bind RNA in the major groove, where it can distinguish between ribonucleotide bases) [Steitz 1999], and thus may have distinct properties compared to proteins which bind RNA through the backbone [Jones 2001].
In total, we defined 27 disorder-based features. The first feature, D1, is simply the percentage of the amino acid sequence predicted to be disordered. Features D2-D15 (14 features) capture the location of disorder within the protein sequence; D2-D5 correspond to the percentage of residues in each quarter of the linear sequence that are predicted to be disordered, while D6-D15 correspond to the percentage of residues in each tenth of the sequence that are predicted to be disordered. (See Fig. 3-1)

An additional 12 features capture the physical and/or chemical properties of the predicted disordered residues: ordPos, ordNeg, and ordNeut, are the percentages of amino acids that are positively or negatively charged, or neutral, respectively; disPos, disNeg, and disNeut, are the percentages predicted to be disordered that are positively or negatively charged, or neutral, respectively; ordPhilic, ordPhobic, ordMod, disPhilic, disPhobic, and disMod, are the proportion of the predicted ordered and disordered residues that are hydrophilic, hydrophobic, and moderately hydrophobic.

**Figure 3-1. After the disordered residues in a sequence have been predicted using DISOPRED3, the sequence is divided into quarters, and the percentage of each quarter of the sequence that is predicted to be disordered is then incorporated as features D2, D3, D4, and D5.**
Machine Learning Models

Machine learning experiments were performed using Weka v3.8 [Frank et al. 2016] and R v3.4.4 implemented in RStudio IDE v1.0.143. Random forests in R were created using the randomForest package [Liaw & Wiener 2002]. We examined several machine learning classifiers in Weka, including Naïve Bayes and logistic regression (not shown), before determining that superior performance was achieved using the Weka random forest classifier, which implements a variant of Breiman’s algorithm [Breiman 2001] that selects features based on information gain [Frank et al. 2016]. Classifiers were selected based on maximization of the specificity while maintaining sensitivity and Matthews Correlation Coefficient (MCC) comparable to available methods (see Performance Metrics).

After training models using several different forest sizes, node depths, and random feature pools (not shown), a model with 500 trees, maximum node depth of 3, and random feature pool of 5 was selected. Training was performed using 10-fold cross-validation on the RPI-PDB dataset, validation was performed on the independent RPI-NPInter* dataset, and a case study was performed on the RPI-FMRP dataset. Random forest classifiers trained using the randomForest package, which implements the Breiman algorithm [Breiman 2001] (and selects features at each node using node impurity) [Liaw & Wiener 2002], were also evaluated. Weka offers finer parameter tuning than the randomForest package and the final Weka model outperforms the best randomForest model, so this is the model that is ultimately used for RPIDisorder.

Performance Metrics

We used RPI-NPInter* as a validation set to tune the RPIDisorder model, and further tested RPIDisorder on the completely independent RPI-FMRP dataset. The metrics utilized are described below; TP = True Positives (correctly identified interacting instances), TN =
True Negatives (correctly identified non-interacting instances), FP = False Positives (non-interacting instances incorrectly predicted to interact), and FN = False Negatives (interacting instances incorrectly predicted to not interact). The reported metrics represent the weighted average of the interacting and non-interacting classes.

**True positive rate (sensitivity, recall)**

The true positive rate (TPR), also known as the sensitivity or recall, represents how well the classifier identifies interacting RNAs and proteins. Values closer to one indicate a better TPR.

\[
\text{True Positive Rate (TPR)/Sensitivity/Recall} = \frac{TP}{TP + FN}
\]

**False positive rate**

The false positive rate represents how bad the classifier is at distinguishing non-interactions from interactions. Values closer to zero indicate a better FPR.

\[
\text{False Positive Rate (FPR)} = 1 - \text{Specificity} = \frac{FP}{FP + TN}
\]

**Specificity**

The specificity represents how good a classifier is at identifying non-interacting instances and is equal to 1-FPR. Values closer to one indicate a better specificity.

\[
\text{Specificity} = 1 - \text{FPR} = \frac{TN}{TN + FP}
\]

Note that machine learning literature frequently defines specificity as “the probability that a positive prediction is correct” [Baldi et al. 2000], i.e., specificity = TP / (TP+FP); we refer to this as “precision”, described below. Here we use “specificity” in the medical statistics sense, where it refers to the ability of a classifier to correctly predict the negative class [Baldi et al. 2000].
Precision

The precision is the proportion of predicted interactions which were actually interactions. Values close to 1 indicate better performance.

\[ \text{Precision} = \frac{TP}{TP + FP} \]

F-measure

The F-measure (aka the F1 score or F-score) is the harmonic mean of precision and recall, and as such is of more use than accuracy for gauging performance on unbalanced datasets. However, as the F-measure does not utilize TN, it is of limited utility for gauging if a classifier is better at distinguishing the non-interacting class. If precision and recall are both perfect (i.e., equal to 1), then the F-measure will also be 1, and thus scores closer to 1 indicate a better classifier.

\[ F - \text{measure} = \left( \frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right) = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \]

Area under the receiver operating characteristic curve

The Area Under the Receiver Operating Characteristic curve (AUC of ROC or simply AUC; see Fig. 3-2) shows the trade-off between the TPR and FPR when using varying “score thresholds” to discriminate the interacting and non-interacting classes. Binary random forest classifiers output the proportion of the decision trees which “voted” to classify an RNA-protein pair as interacting; the score threshold is the proportion of trees which must vote to classify an RNA-protein pair as interacting for it to receive this prediction. E.g., if a threshold of 40% is chosen, then \( \geq 40\% \) of the decision trees in the model must output the “interacting” class. The ROC curve graphs the TPR vs the FPR for these varying thresholds.
The Area Under the Curve (AUC) is closer to 1 for better classifiers; worse classifiers are close to 0.5.

![Example ROC Curves](image)

Figure 3-2 Example ROC curves. A hypothetical perfect classifier (orange) would have 100% TPR and 0% FPR, regardless of threshold, which gives an AUC of 1. A random (binary) classifier will have 50% TPR and 50% FPR, regardless of threshold, which produces an AUC of 0.5. Better classifiers will not have a large trade-off between TPR and FPR, producing ROC curves with a high TPR when the FPR is low, leading to an AUC closer to 1 (gray). Worse classifiers will not have as high a TPR when the FPR is low, leading to smaller AUCs closer to 0.5 (yellow).

**Accuracy**

The *accuracy* is the overall proportion of correctly predicted interactions and non-interactions. Accuracy is very easily skewed in unbalanced datasets (i.e., the number of instances belonging to each class is very unequal) – if a dataset is mostly interacting RNAs and proteins, and a classifier simply predicts the majority class in every instance, it will appear to have a high accuracy, even though such a classifier will only perform well if the population distribution of the classes heavily favors the majority class. Thus, accuracy should never be reported as the sole performance metric. Scores closer to 1 indicate higher accuracy.
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

**Matthews correlation coefficient**

The *Matthews correlation coefficient* [Matthews 1975] (MCC) provides insight into the overall quality of a binary classifier. It represents the correlation between the predicted and observed classes and essentially functions as the Pearson correlation coefficient for binary variables [Boughorbel et al. 2017]. Additionally, the MCC tends to work very well as a metric for performance on unbalanced datasets [Boughorbel et al. 2017]. Values range from -1 to 1; 0 represents no correlation between the predicted and observed classes, and 1 represents perfect performance. Classifiers with a between 0 and 1 therefore have positive correlation between their predictions and reality.

\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}}
\]

**Feature Impact**

**Feature importance**

As Weka v3.8 does not support determining feature importance directly within a created model, I also trained a classifier using the R randomForest package to approximate the feature importance within the Weka classifier.

The randomForest package and Weka implementations of the random forest classifier differ mainly in that while Weka selects features from the random pool at each node using an information gain attribute, the randomForest package selects features on the basis of increase in node purity (Gini) – i.e., how well a feature separates classes at each node. Weka also does not provide a way to determine a feature’s direct importance to a particular model, but randomForest has very robust feature importance tools. I therefore created a disorder model
in randomForest and retrieved the importance features for that model. Although this does not directly show that these disorder features are of the same importance in the Weka model, this approach serves as an approximation of the importance of disordered features in the RPIDisorder model.

Results and Discussion

Performance on the RPI-PDB Training Dataset

We trained several machine learning models on the RPI-PDB training dataset using several different classification methods including Naïve Bayes, logistic regression, and random forest. These models were initially created using the Weka 3.8 default parameter settings. Minimal parameter tuning (not shown) was performed to determine that of these algorithms, the random forest classifier performed best, and further tuning was performed using this classifier.

The default random forest parameters in Weka 3.8 use a forest of 100 trees, do not impose a maximum node depth constraint, and use the following formula to choose the number of random features to examine at each node:

\[
\text{Num. Random Features} = \lfloor \log_2(\text{Num Features}) \rfloor + 1
\]

Since the conjoint triad + conjoint tetrad + disorder feature vector has 626 features, the default Weka 3.8 random forest classifier considered 10 randomly selected features at each node.

After extensive parameter tuning based on performance on the RPI12252* validation set (not shown), the final RPIDisorder model uses the Weka 3.8 random forest classifier with maximum node depth of 3, random feature pool of 5, and a forest of 500 trees (increasing the forest to 1000 trees does not impact performance; data not shown).
I created additional random forest classifiers using the R randomForest package [Liaw & Wiener 2002]. The R randomForest implementation does not allow specification of the maximum node depth but does allow tuning of the number of trees and the random feature pool size. The best randomForest classifier had a forest of 1000 trees and random feature pool size of 2.

The performances of the Weka 3.8 default parameter random forest, tuned R randomForest classifier, and tuned Weka classifier are shown in Fig. 3-3.

Figure 3-3 RPI-PDB 10-fold cross-validation results for three random forest classifiers. The Weka 3.8 random forest classifier with default parameters is in yellow. A parameter-tuned R randomForest classifier is in purple. The parameter-tuned RPIDisorder model is in light blue. The R randomForest method AUC is not reported.

**Parameter Tuning**

Although the default Weka random forest classifier performed best on 10-fold cross-validation and the R randomForest model had comparable metrics (except for MCC), RPIDisorder had a much lower TPR and MCC and higher FPR. This is likely because the lack of a maximum node depth constraint allowed overfitting to the RPI2241 dataset for the
default Weka and R randomForest models. However, the tuned Weka random forest classifier performs far better on the RPI-NPInter* validation set (see Fig. 3-4) than the other methods.

On the RPI-NPInter* validation set, the three methods have comparable TPR, precision, recall, F-measure, and accuracy, largely due to the unbalanced nature of the RPI-NPInter dataset, which has ~10x as many interacting pairs as non-interacting pairs. However, the tuned RPIDisorder classifier has far superior FPR and specificity, which produces a better MCC overall than the other two classifiers.

![RPI12252 Parameter Tuning](image)

**Figure 3-4** RPI-NPInter validation performance of the Weka random forest classifier with default parameters (yellow), tuned R randomForest classifier (purple), and the tuned RPIDisorder model (blue).

**Comparison to Other Methods on RPI-NPInter* Dataset**

Walia created the RPIMotif method, which utilizes the same conjoint triad protein representation and conjoint tetrad RNA representation which RPISeq and RPIDisorder utilize, but additionally scans potential RNA and protein partners for bi-partite interfacial sequence motifs pulled from the interfaces of RNA-protein complex structures in the PDB.
Walia 2014). RPIMotif was also validated on the RPI-NPInter dataset (i.e., the final RPIMotif model was selected based on its performance on this dataset), which makes for a fair performance comparison between RPIMotif and RPIDisorder on this dataset (see Fig. 3-5). Although RPISeq was not validated on RPI-NPInter, Walia tested it on RPI-NPInter, and we include it here for completeness.

![Figure 3-5 RPI-NPInter/RPI-NPInter* performance comparison between RPISeq (gray), RPIMotif (red), and RPIDisorder (blue) on the dataset. All three methods were trained on RPI2241. RPIDisorder results do not include a single interacting instance for which DISOPRED3 did not return results (denoted as RPI-NPInter*). RPISeq and RPIMotif include the full RPI-NPInter dataset. All three methods have comparable TPR, Precision, Recall, F-Measure, and accuracy. However, RPIDisorder has a much lower FPR (14.4%) than either RPISeq (FPR 63%) or RPIMotif (FPR 29%), resulting in a slightly larger MCC (0.684) than RPIMotif (0.64) and a much larger MCC than RPISeq (0.40).

RPIDisorder also has a similar AUC of ROC in comparison with RPIMotif and a better AUC than RPISeq (see Fig. 3-6). RPIDisorder achieves an AUC of 0.92, which is the
same as RPIMotif’s reported AUC on this test set, while RPISeq has a reported AUC of 0.87. The AUC of ROC curves show that RPISeq consistently has a curve below RPIDisorder and RPIMotif, indicating that its high TPR does come at the cost of a higher FPR. RPIMotif’s AUC curve shows that RPIMotif has a better TPR than RPIDisorder up until the FPR hits 0.08%; after this point, RPIDisorder has better TPR and FPR than RPIMotif.

Figure 3-6 Area Under the Receiver Operating Characteristic Curves (AUC of ROC) for RPISeq (gray), RPIMotif (red), and RPIDisorder (blue). AUCs are 0.87, 0.92, and 0.92 for RPISeq, RPIMotif, and RPIDisorder, respectively. Curve created in R using ggplot2 [Wickham 2016] with code provided courtesy of Kris De Brabanter.

Characterization of Order and Disorder in the RPI-PDB and RPI-NPInter* Datasets

The RPI-PDB dataset is a structurally-derived dataset from the PDB [Berman et al. 2000], while the RPI-NPInter* dataset was derived from the NPInter v2.0 database [Yuan et al. 2014], which contains ncRNA-protein interactions identified by high-throughput experiments. Since the RPI-PDB dataset only used RNP structures generated by x-ray
crystallography, it contains only those structures capable of being crystallized, which requires a stable RNP conformation. Thus, this dataset is biased towards ordered structures, as can be seen from the predicted protein disorder frequency in the RPI2241 dataset (Fig. 3-7).

![Frequency of Predicted Sequence Disorder in RPI2241 Training Dataset](image)

**Figure 3-7** Histogram of the proportion of sequence predicted to be disordered for proteins in the RPI-PDB dataset (disorder feature D1). The y-axis is the proportion of all protein sequences within the dataset that have a particular level of sequence disorder.

The RPI-PDB dataset skews heavily towards sequences that have no or low predicted disorder levels (>80% of sequences are predicted to have 20% or fewer of their amino acids be disordered), and no sequences are predicted to be more than 60% disordered. In contrast, the RPI-NPInter* dataset has a very different distribution of predicted disorder levels.
Comparatively few (<40%) of the RPI12252 protein sequences are predicted to have 20% or fewer disordered amino acids, with numerous sequences predicted to be >80% disordered (see Fig. 3-8).

Figure 3-8 Histogram of the proportion of sequence predicted to be disordered for proteins in the RPI-NPInter* dataset (disorder feature D1). The y-axis is the proportion of all protein sequences within the dataset that have a particular level of sequence disorder.

Interestingly, the two datasets do not just have differing levels of sequence disorder – they also, perhaps crucially, differ in where these disordered residues occur. In the RPI-PDB dataset, most of the disordered residues are predicted to occur in N- and C-termini of the protein (see Fig. 3-9), though there are proteins predicted to have high levels of disorder in
the middle of protein sequences as well, indicating possible cases where a disordered protein transitions from disordered to ordered upon RNA binding.

Figure 3-9 Boxplots of the predicted proportion of disordered residues in regions of RPI2241 proteins (disorder features D6-D15). The protein sequences are divided into tenths, and the proportion of the first tenth, second tenth, etc. which is predicted to be disordered is determined (see Methods and Materials). The boxes represent the 25th, median, and 75th percentiles; the whisker boundaries are 1.5 * the interquartile range above and below the 25th and 75th percentiles.

In stark contrast, the RPI-NPInter* dataset sees heavily disordered regions in the middle of proteins, including the 7th, 8th, and 9th sequence 10ths, in addition to the N- and C-termini (see Fig. 3-10). Interestingly, the median disorder level in the RPI-NPInter* regions
corresponding to the 2\textsuperscript{nd} and 5\textsuperscript{th} tenths of the sequence is highly ordered, with \textasciitilde75\% of proteins containing little to no disordered residues in these regions. This may explain why the tenths division worked so well in the model.

Figure 3-10 Boxplots of the predicted proportion of disordered residues in regions of RPI12252* proteins (disorder features D6-D15). The protein sequences are divided into tenths, and the proportion of the first tenth, second tenth, etc. which is predicted to be disordered is determined (see Methods and Materials).

These results indicate that the PDB tends to have highly ordered structures which may be disordered at the N- and C-termini (and it is likely that many RPI-PDB interactions are drawn from structures missing residues from these regions which could not be
crystallized.) The non-structurally determined RPI-NPInter* dataset, in contrast, can include highly disordered sequences incapable of being crystallized which are nonetheless flexible enough to conform to and bind RNA tightly and with specificity [Varadi et al. 2015]. These examinations show that there are large differences in the degree and distribution of disordered residues in PDB complexes vs NPInter complexes.

Interestingly, while the datasets differ in how much and where the disorder occurs, they show a great deal of similarity in the physicochemical properties of both the ordered and disordered residues (see Figs. 3-11 and 3-12).

The predicted ordered residues are nearly identically distributed for RPI-PDB and RPI-NPInter*, with differences only in the proportion of positively charged and neutral residues predicted to be ordered. The RPI-NPInter* dataset has a slightly larger proportion of uncharged residues; this may be due to the propensity of flexible linker groups to be formed from uncharged residues [Chen et al. 2013], thus resulting in fewer ordered PDB examples.

The predicted disordered residues show a much greater variance in distribution, though the medians are still similar for the hydrophilic, moderately hydrophilic, and negatively charged residues, and a small difference between the positively charged medians. Nonetheless, there are large differences in the medians for the hydrophobic and neutral residues, where the RPI12252* dataset has larger proportions than RPI2241. The elevated proportion of neutral residues in the RPI12252* dataset is likely due to the same phenomenon described above, but the hydrophobic elevation is surprising, as hydrophobic regions tend to promote order as the protein folds to bury these regions [Uversky 2011]. It is thus unexpected for the more disordered RPI12252* dataset proteins to (on average) contain more of these residues than the structured RPI2241 dataset. However, the same
hydrophobicity which drives protein folding [Dill 1990] may also encourage disorder-order transition in proteins upon RNA-binding, and the RPI12252* dataset may be capturing RPIs where the protein undergoes a disorder-order transition upon binding, while the RPI2241 dataset may lack a similar proportion of crystallized RNPs.

Altogether, this demonstrates that while there are differences in locations of disordered regions in the RPI2241 and RPI12252* datasets, the physicochemical properties of the amino acids within those regions are similar enough to apply signal learned from the RPI2241 dataset to the RPI12252* dataset.

**Performance on FMRP119 Testing Dataset**

To compare the performance of RPIDisorder, RPISeq, and RPIMotif on an independent test set, we evaluated the performance of the three classifiers using the FMRP119 (RPI-FMRP) dataset, which consists of 119 RNA-protein interactions for the Fragile X Mental Retardation Protein (FMRP). FMRP contains 3-4 RNA-binding domains: at least two (and possibly three) KH domains [Myrick et al. 2015], and an RGG domain located within the heavily disordered region near the C-terminus. The FMRP RGG box is responsible for sequence-specific binding to mRNA targets, particularly targets containing G-quadruplex structural motifs [Darnell et al. 2001]; in fact, specific arginines in the RGG box are important for binding different RNAs [Blackwell et al. 2010]. These qualities make FMRP an ideal case study for demonstrating the practical utility of RPIDisorder for predicting mRNA-protein interactions involving a highly disordered protein.

The RPI-FMRP dataset pairs a truncated FMRP sequence corresponding to the disordered C-terminus region containing the RGG domain responsible for mRNA-binding specificity. While Ascano et al. did not specifically determine if this region was responsible for the mRNA binding activity in this dataset [Ascano et al. 2012], Darnell et al.
Figure 3-11 Boxplots of the proportion of predicted ordered residues within the RPI2241 dataset proteins (gray) and the RPI12252* dataset proteins (green) which are hydrophilic, moderately hydrophilic, hydrophobic, positively charged, negatively charged, and uncharged.
Figure 3-12 Boxplots of the proportion of predicted disordered residues within the RPI2241 dataset proteins (gray) and the RPI12252* dataset proteins (green) which are hydrophilic, moderately hydrophilic, hydrophobic, positively charged, negatively charged, and uncharged.
demonstrated that FMRPs containing the I304N point mutation in the KH2 domain (and the corresponding mutation in the KH1 domain) bind to specific mRNA nearly as well as wildtype FMRP [Darnell et al. 2001], indicating that the KH domains are likely not involved in determining FMRP’s mRNA partners. The RGG domain and C-terminus region appear to be primarily responsible for determining specific FMRP-mRNA interactions [Blackwell et al. 2010, Ozdilek 2017]. Thus, pairing the C-terminal FMRP region with the regions of the mRNA Ascano et al. [Ascano et al. 2012] found it to bind appears to be a valid approach.

In fact, using the full-length FMRP protein sequence including KH domain regions led RPISeq, RPIMotif, and RPIDisorder to all predict indiscriminate binding for both the positive and negative instances for RPI-FMRP (not shown). This is very likely due to the fact that the RPI-PDB dataset, which all the previously mentioned methods were trained on, includes over a thousand ribosomal-RNA/ribosomal-protein interactions, including S3 ribosomal protein paired with the 16S ribosomal RNA – for example, chain C (S3) and chain A (16S rRNA) from PDB structure 2E5L (Thermus thermophilus 30S ribosomal subunit) are paired in the RPI-PDB training dataset. The thermophilus 16S ribosomal subunit is heavily G-rich (35.67% of the sequence is composed of guanines), and the QGRS software (Quadruplex forming G-Rich Sequences) [Kikin et al. 2006], which identifies putative G-quadruplex forming regions in nucleic acid sequences, identifies a whopping 515 (overlapping) potential quadruplex-forming G-rich sequences. This approximates to 1 potential QGRS every 3.37 amino acids in the thermophilus 16S ribosomal subunit. The KH domains in both the thermophilus S3 and FMRP proteins are predicted to be ordered (incidentally, the KH domains in FMRP may actually bind directly to the ribosome to allow FMRP to repress translation [Harigaya & Parker 2014].) Thus, from a sequence and structure
standpoint, the classifiers could very easily have learned from the RPI-PDB dataset that ordered KH domains will bind to G-rich RNA regions, and lack necessary additional KH interactions to predict KH domain binding specificity.

![Figure 3-13 Performance comparison of RPISeq (gray), RPIMotif (red), and RPIDisorder (blue) on the FMRP119 dataset.](image)

Ultimately, on the RPI-FMRP dataset (using the truncated C-terminus region of the FMRP protein), RPIDisorder had a lower TPR (79.0%), but comparable precision, F-measure, and accuracy (78.7%, 78.7%, 79%, respectively; see Fig. 3-13) compared to RPISeq (TPR 88.0%, precision 76.7%, F-measure 78.7%, and accuracy 75.6%) and RPIMotif (TPR 89.3%, precision 78.8%, F-measure 83.6%, accuracy 78.2%). On the other hand, RPIDisorder had a much lower false positive rate (26.4%) than both RPISeq (45.5%) and RPIMotif (40.9%), leading to a marginally better MCC (0.541 vs RPISeq 0.459 and RPIMotif 0.517).

We also generated AUC of ROC curves for this dataset (see Fig. 3-14). Although all the classifiers achieve similar AUCs for this dataset, RPIMotif actually has a slightly larger
AUC than either RPIDisorder or RPISeq, despite the fact that RPIDisorder has a much better specificity/FPR and marginally better MCC and accuracy. This is likely due to the small size of the FMRP dataset and that it is somewhat imbalanced (75 interacting examples and 44 non-interacting examples), leading to a better AUC when achieving a high TPR.

![AUC of ROC on FMRP119 Dataset](image)

Figure 3-14 Area Under the Receiver Operating Characteristic Curves (AUC of ROC) for RPISeq (gray), RPIMotif (red), and RPIDisorder (blue). AUCs are 0.86, 0.90, and 0.88 for RPISeq, RPIMotif, and RPIDisorder, respectively. Curves created in R using ggplot2 [Wickham 2016] with code provided courtesy of Kris De Brabanter.

**Contributions of Disordered Features**

Identifying which of the predicted disorder features and physicochemical properties contribute most to the improved prediction of RPIPs obtained using RPIDisorder could provide insights into mechanisms of specific RNA-protein recognitions. Although it is not possible to extract feature importance directly from the RPIDisorder Weka model, it is
possible to examine feature importance within the R randomForest model, which can (with caution) be used to approximate the importance of the features in the Weka model.

**Disorder Feature Importance**

![Box plot comparison of importance measures](image)

Figure 3-15 Comparison of the four importance measures reported by R randomForest: The importance of features in identifying the interacting RNA-protein class (Interacting), the non-interacting RNA-protein class (Non-interacting), the mean decrease in model accuracy (MDA), and the mean decrease in node impurity (Gini) produced by the feature.

R's randomForest package provides two metrics for identifying a feature’s importance to a random forest model: mean decrease in accuracy (MDA) and mean increase in node impurity (Gini). The MDA is determined by calculating the out-of-bag (OOB) prediction error, then randomly permuting the values of a particular feature in the OOB data instances and computing the difference in error and averaging these values over all trees before
normalizing using the standard deviation of the differences in error [Liaw & Wiener 2018]. The MDA therefore demonstrates the effect a particular feature has on a model’s accuracy; if there is a large decrease in accuracy from randomly permuting the variable, then it must be important in accurate classification. R randomForest can also generate the MDA for class-specific labels (e.g., the importance of a feature for determining the “Interacting” and “Noninteracting” classes) and give an idea of a feature’s importance in determining interacting RNA-protein partners and determining non-interacting RNAs and proteins.

The Gini measures the total increase in node purity (i.e., how well a particular feature “splits” the data between interacting and non-interacting classes) for a feature averaged over the trees in the random forest. If a feature produces a large increase in node purity, this is an indicator that it splits the classes well and is therefore important. This metric can also be determined for class-specific labels.

Feature importance was assessed using the top performing R randomForest classifier, which achieves TPR 87.0%, FPR 12.0%, Specificity 88.0%, Precision 87.8%, F-Measure 87.4%, Accuracy 87.5%, and MCC 0.749 in training on RPI-PDB, and TPR 96.8%, FPR 34.1%, Specificity 65.9%, Precision 97.2%, F-Measure 97%, Accuracy 94.4%, and MCC of 0.610 on RPI-NPInter*. An overview of the results of feature importance assessment is shown in Fig. 3-15. Detailed analyses focused on each class of importance features are provided in the following sections.

**Features important for identifying the “Interacting” (positive) class**

A boxplot of feature importance for all features, conjoint triad protein features, conjoint tetrad features, and disorder features is shown in Fig. 3-16 and the most important features (top 5) are summarized in Table 3-1. Fig. 3-16 shows that the protein features (in red) and disorder features (in gray) were both more important than RNA (in blue) for
identifying the “interacting” (positive) class, with median values of 9.08% for protein and 11.17% for disorder.

As shown in Table 3-1, of the 5 features most important for determining the interacting class as determined by MDA, 3 are disorder features. Of the top 30 features (the top 5% of features overall), nine (29%) are disorder features (see Supp. Table S3-3), even though disorder features make up only 4.2% of the total feature count. The top three features for determining the interacting class were: ordPos, the percentage of residues predicted to be ordered and positively charged; D1, the percentage of the total sequence predicted to be disordered; and D5, the percentage of the final quarter of the sequence predicted to be disordered. Thus, protein and disorder features are more important in identifying the positive/interacting class than RNA features. This contrasts with features important for identifying the non-interacting class (see Fig. 3-17), where the median importance of RNA features (5.67%) is on par with that of protein features (5.27%); and both RNA and protein features are less important than disorder importance (6.38%).

Features important for identifying the “Non-Interacting” (negative) class

Of the top 5 features most important for determining the non-interacting class (as determined by MDA), there are four disordered features (Table 3-2). Disordered features make up 16.1% (5/31) of the top of features for identifying the non-interacting class (see Supp. Table 3-4). Disorder features D1 and D5 are both very important, although ordPos, the percentage of residues predicted to be ordered that are positively charged is only the 5th most important feature, whereas ordNeg, the percentage of predicted ordered residues that are negatively charged is ranked most important. Note that disNeg, the percentage of predicted disordered residues that are negatively charged is also of elevated importance. Because sequence-specific RNA-protein binding may require interaction between the protein and the RNA face,
rather than the backbone, the disposition of the negatively charged residues may be important in mediating these interactions [Jones 2001].

Figure 3-16 Feature importance (Mean Decrease in Accuracy, %) for identifying interacting RNA-protein pairs. The collective importance of all features for identifying the interacting class is in white. The protein features (P1-P343) are in red. The RNA features (R1-R56) are highlighted in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray.

Table 3-1: Importance metrics for the R randomForest classifier, ordered in descending order based on Interacting importance, for the top 5 Interacting features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. See “Methods” for description of disorder features. The first three columns are the Mean Decrease in Accuracy (%) for determining the Interacting class, Non-interacting class, and overall MDA, respectively.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
<tr>
<td>D1</td>
<td>13.66</td>
<td>8.94</td>
<td>14.17</td>
<td>2.96</td>
</tr>
<tr>
<td>D5</td>
<td>13.36</td>
<td>8.84</td>
<td>14.63</td>
<td>2.61</td>
</tr>
<tr>
<td>P251</td>
<td>13.28</td>
<td>8.68</td>
<td>13.97</td>
<td>2.36</td>
</tr>
<tr>
<td>P57</td>
<td>13.10</td>
<td>6.94</td>
<td>13.82</td>
<td>2.59</td>
</tr>
</tbody>
</table>
Figure 3-17 Feature importance for identifying non-interacting RNA-protein pairs. The collective importance of all features for identifying the non-interacting class is in white. The protein features (P1-P343) are highlighted in red. The RNA features (R1-R56) are highlighted in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray.

Table 3-2 Importance metrics for the R randomForest classifier, ordered in descending order based on Non-interacting importance (Nonint), for the top 5 non-interacting features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. RNA conjoint tetrad features start with an “R”. See “Methods” for description of disorder features.

<table>
<thead>
<tr>
<th></th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordNeg</td>
<td>10.78</td>
<td>9.71</td>
<td>12.16</td>
<td>3.42</td>
</tr>
<tr>
<td>D1</td>
<td>13.66</td>
<td>8.94</td>
<td>14.17</td>
<td>2.96</td>
</tr>
<tr>
<td>D5</td>
<td>13.36</td>
<td>8.84</td>
<td>14.63</td>
<td>2.61</td>
</tr>
<tr>
<td>R173</td>
<td>5.98</td>
<td>8.77</td>
<td>9.61</td>
<td>1.98</td>
</tr>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
</tbody>
</table>
Feature importance based on Mean Decrease in Accuracy

The overall MDA for each feature (Fig. 3-18) is fairly similar to the Interacting MDA (Fig. 3-16); in fact, the top five features (ordPos, D5, D1, P251, and P57) are the same between the two measures, though the exact order is slightly different (see Table 3-3). Of the top 5% of most important features as measured by MDA (see Supp. Table S3-5), 9/31 (29%) are disordered features. Between the Interacting, Non-interacting, and overall MDA, disorder features tend to be of high importance for the model.

Figure 3-18 Feature importance ranked by Mean Decrease in Accuracy. The collective importance of all features is in white. The protein features (P1-P343) are in red. The RNA features (R1-R56) are in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray.
Table 3-3 Importance metrics for the R randomForest classifier, ordered in descending order based on overall Mean Decrease in Accuracy, for the top 5 MDA features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. See “Methods” for description of disorder features.

<table>
<thead>
<tr>
<th></th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
<tr>
<td>D5</td>
<td>13.36</td>
<td>8.84</td>
<td>14.63</td>
<td>2.61</td>
</tr>
<tr>
<td>D1</td>
<td>13.66</td>
<td>8.94</td>
<td>14.17</td>
<td>2.96</td>
</tr>
<tr>
<td>P251</td>
<td>13.28</td>
<td>8.68</td>
<td>13.97</td>
<td>2.36</td>
</tr>
<tr>
<td>P57</td>
<td>13.10</td>
<td>6.94</td>
<td>13.82</td>
<td>2.59</td>
</tr>
</tbody>
</table>

Gini importance

Figure 3-19 Feature importance ranked by decrease in node impurity. The collective importance of all features is in white. The protein features (P1-P343) are in red. The RNA features (R1-R56) are in blue. The disorder features (D1-D15, ordPhilic, ordMod, ordPhobic, ordPos, ordNeg, ordNeut, disPhilic, disMod, disPhobic, disPos, disNeg, and disNeut) are in gray.
The mean decrease in node impurity measures how important a feature is for splitting the interacting and non-interacting classes. A larger mean decrease in node impurity means a feature is more important. The disorder features have mean decrease in node impurity of 2.16. However, only one disorder feature (ordNeut) is among the top 5 most important features as measured by Mean Decrease in Gini (see Table 3-4), although the disorder features make up 16% of the top 5% of features as ranked by mean decrease in node impurity (see Supp. Table S3-6).

Table 3-4 Importance metrics for the R randomForest classifier, ordered in descending order based on Mean Decrease in Node Impurity (MeanDecreaseGini), for the top 5 features. Disordered features are highlighted in blue. RNA conjoint tetrad features start with an “R”. See “Methods” for description of disorder features.

<table>
<thead>
<tr>
<th></th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>R6</td>
<td>5.42</td>
<td>5.39</td>
<td>6.08</td>
<td>5.57</td>
</tr>
<tr>
<td>R33</td>
<td>4.44</td>
<td>4.63</td>
<td>5.21</td>
<td>4.30</td>
</tr>
<tr>
<td>R120</td>
<td>5.27</td>
<td>5.47</td>
<td>5.98</td>
<td>4.10</td>
</tr>
<tr>
<td>ordNeut</td>
<td>12.05</td>
<td>7.20</td>
<td>11.72</td>
<td>3.93</td>
</tr>
</tbody>
</table>

Conclusion

RPIDisorder is a machine learning classifier for predicting RNA-protein interactions using sequence composition and predicted disorder information. On the basis of several different performance evaluation metrics, RPIDisorder outperforms the previous state-of-the-art sequence composition-based method, RPISeq [Muppirala et al. 2011]. The performance of RPIDisorder in detecting RPIPs is comparable to that of the motif-based method, RPIMotif [Walia 2014], and it is superior to RPIMotif in correctly identifying non-interacting RNA-protein pairs.

Notably, even though the training dataset (RPI-PDB) was derived from PDB structures and is therefore biased toward proteins that are predominantly structured, it
contains enough disorder-based signal to generate a model that makes reliable predictions on two datasets (RPI-NPInter* and RPI-FMRP) with a much higher proportion of disordered structures. On the other hand, the failure of RPIDisorder, RPIMotif, and RPISeq to predict interactions between the full-length FMRP protein and the RPI-FMRP dataset demonstrates the need for better training datasets.

All of the feature importance results suggest that disorder-based features are important for distinguishing interacting and non-interacting RNA-protein pairings. Thus, RPIDisorder can make highly specific predictions on a broad range of RNPs and RNA-protein interaction networks, including those containing highly disordered proteins, and proteins that interact with diverse ncRNAs and mRNAs.

**Future Directions**

Future work on RPIDisorder will include the generation of a true negative dataset derived from human protein PAR-CLIP experiments deposited in ENCODE [ENCODE Project Consortium 2012] (in progress), to evaluate whether PRIP prediction performance can be further improved. Also, a rigorous evaluation of RPIDisorder will require direct comparison of its performance with methods published recently by other groups, using carefully selected benchmark datasets.

**Availability**

RPIDisorder will soon be available at www.rpidisorder.org, and is available for download at https://github.com/Dobbs-Lab/RPIDisorder.com. All datasets and code necessary for reproducing these results are available at www.rpidisorder.org/reproducible_research/.
Acknowledgements

I would like to thank the Dill-PICL lab for their valuable discussions on the RPIDisorder interface.

References

http://dx.doi.org/10.1093/bioinformatics/btt495

http://dx.doi.org/10.1146/annurev-biochem-072711-164732

http://dx.doi.org/10.1093/nar/gkw593

http://dx.doi.org/10.1038/nature11737

http://dx.doi.org/10.1038/ng0793-244


http://dx.doi.org/https://doi.org/10.1016/j.tibs.2014.10.005


http://dx.doi.org/https://doi.org/10.1016/j.molcel.2012.05.021


http://dx.doi.org/10.1038/nmeth.1611


http://dx.doi.org/10.1093/hmg/ddq007


http://dx.doi.org/10.1371/journal.pone.0177678


http://dx.doi.org/10.1038/nphys3532


http://dx.doi.org/10.1023/A:1010933404324


http://dx.doi.org/10.1016/j.molcel.2012.05.039


http://dx.doi.org/10.1002/1873-3468.12621


http://dx.doi.org/10.1038/nature11247

http://dx.doi.org/https://doi.org/10.1016/0092-8674(95)90436-0


http://dx.doi.org/10.1016/j.gpb.2016.01.004

http://dx.doi.org/10.1016/j.febslet.2008.03.004

http://dx.doi.org/10.1016/j.tig.2018.09.006


http://dx.doi.org/10.1038/nature08975


http://dx.doi.org/10.1016/j.cell.2010.03.009


http://dx.doi.org/10.1016/j.molcel.2014.04.027


http://dx.doi.org/10.1007/s00018-017-2558-1


http://dx.doi.org/10.1126/science.1197349


http://dx.doi.org/10.1039/c7mb00290d


http://dx.doi.org/10.1186/s12964-016-0132-3


http://dx.doi.org/10.1093/bioinformatics/btu744


http://dx.doi.org/10.1007/s12551-016-0223-9

http://dx.doi.org/10.1002/prot.25639

http://dx.doi.org/https://doi.org/10.1016/j.ymeth.2004.03.027

http://dx.doi.org/10.1038/nprot.2006.47

http://dx.doi.org/10.1016/j.semcdb.2011.02.016

http://dx.doi.org/10.1016/j.febslet.2015.08.027

http://dx.doi.org/10.1093/nar/gkl253

http://dx.doi.org/10.1093/nar/gkl837

http://dx.doi.org/10.1002/jmr.1061

http://dx.doi.org/10.18632/oncotarget.21934

http://dx.doi.org/10.1186/1471-2164-14-651

http://dx.doi.org/10.1074/mcp.O113.030239

http://dx.doi.org/10.1016/j.cancergen.2011.11.002


http://dx.doi.org/10.1093/hmg/ddg335


http://dx.doi.org/10.1002/prot.23161

http://dx.doi.org/10.1002/prot.24391
http://dx.doi.org/10.1093/brain/awh256

http://dx.doi.org/10.1016/j.gde.2017.11.006

http://dx.doi.org/10.1074/jbc.REV118.004842

http://dx.doi.org/10.1038/nrg3722

http://dx.doi.org/10.1186/1471-2105-12-489

http://dx.doi.org/10.1093/hmg/ddu586

http://dx.doi.org/10.3389/fcell.2015.00008

http://dx.doi.org/10.1146/annurev.bi.53.070184.001003

http://dx.doi.org/10.1146/annurev-biochem-072711-164947

http://dx.doi.org/10.1093/nar/gkx460
http://dx.doi.org/10.1186/s12864-016-2931-8

http://dx.doi.org/10.1016/j.cell.2014.08.035

http://dx.doi.org/10.3390/ijms20051070

http://dx.doi.org/10.1093/nar/gkv585

http://dx.doi.org/10.1038/nsmb.2064

http://dx.doi.org/10.1007/978-1-62703-709-9_23

http://dx.doi.org/10.1016/j.cell.2007.05.022

http://dx.doi.org/10.1002/wrna.1369

http://dx.doi.org/10.1093/emboj/20.17.4803

http://dx.doi.org/10.3389/fgene.2018.00716


http://dx.doi.org/10.1073/pnas.0607879104


http://dx.doi.org/https://doi.org/10.1016/j.cell.2014.10.051


http://dx.doi.org/10.1101/gad.261974.115


http://dx.doi.org/10.3390/ijms161125952


http://dx.doi.org/10.1038/nrm1643


http://dx.doi.org/10.1093/nar/gkv020


Uversky, V.N. (2017) Intrinsic disorder here, there, and everywhere, and nowhere to escape from it. *Cellular and Molecular Life Sciences: CMLS*, **74**, 3065-3067. [http://dx.doi.org/10.1007/s00018-017-2554-5](http://dx.doi.org/10.1007/s00018-017-2554-5)


[http://dx.doi.org/10.1073/pnas.1515737112](http://dx.doi.org/10.1073/pnas.1515737112)


[http://dx.doi.org/https://doi.org/10.1016/j.abb.2016.02.029](http://dx.doi.org/https://doi.org/10.1016/j.abb.2016.02.029)


[http://dx.doi.org/10.1007/978-1-4939-6406-2_15](http://dx.doi.org/10.1007/978-1-4939-6406-2_15)

[http://dx.doi.org/10.1080/21655979.2018.1470721](http://dx.doi.org/10.1080/21655979.2018.1470721)

[http://dx.doi.org/10.1039/c2mb25292a](http://dx.doi.org/10.1039/c2mb25292a)


[http://dx.doi.org/10.1073/pnas.1102716108](http://dx.doi.org/10.1073/pnas.1102716108)

[http://dx.doi.org/10.1101/gad.1800909](http://dx.doi.org/10.1101/gad.1800909)
http://dx.doi.org/10.1021/bi401085h

http://dx.doi.org/10.1016/j.febslet.2015.10.003

http://dx.doi.org/https://doi.org/10.1016/j.pnmrs.2016.10.001


http://dx.doi.org/10.1093/nar/gkt1057

http://dx.doi.org/10.1093/nar/gkx1098

http://dx.doi.org/10.3390/ijms20061284

http://dx.doi.org/10.2174/1573406413666170510102405

http://dx.doi.org/10.14348/molcells.2018.0243
http://dx.doi.org/10.1016/j.omtn.2018.09.020

http://dx.doi.org/10.3389/fgene.2018.00239

http://dx.doi.org/https://doi.org/10.1016/j.neucom.2017.07.065

http://dx.doi.org/10.1186/s12859-018-2586-3
**Protein Features**

Table S3-1 Protein conjoint triads utilized in RPIDisorder. The “Feature Label” is the name given to the feature in the vector. The “Feature” column contains the combination of amino acids whose frequency is represented by the Feature Label. The protein groups are \{A,G,V\}, \{I,L,F,P\}, \{Y,M,T,S\}, \{H,N,Q,W\}, \{R,K\}, \{D,E\}, and \{C\}.

<table>
<thead>
<tr>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>P11</td>
<td>{A,G,V}  {I,L,F,P}  {H,N,Q,W}</td>
<td>P41</td>
<td>{A,G,V}  {D,E}  {D,E}</td>
</tr>
<tr>
<td>P14</td>
<td>{A,G,V}  {I,L,F,P}  {C}</td>
<td>P44</td>
<td>{A,G,V}  {C}  {I,L,F,P}</td>
</tr>
<tr>
<td>P16</td>
<td>{A,G,V}  {Y,M,T,S}  {I,L,F,P}</td>
<td>P46</td>
<td>{A,G,V}  {C}  {H,N,Q,W}</td>
</tr>
<tr>
<td>P18</td>
<td>{A,G,V}  {Y,M,T,S}  {H,N,Q,W}</td>
<td>P48</td>
<td>{A,G,V}  {C}  {D,E}</td>
</tr>
<tr>
<td>P25</td>
<td>{A,G,V}  {H,N,Q,W}  {H,N,Q,W}</td>
<td>P55</td>
<td>{I,L,F,P}  {A,G,V}  {D,E}</td>
</tr>
<tr>
<td>P26</td>
<td>{A,G,V}  {H,N,Q,W}  {R,K}</td>
<td>P56</td>
<td>{I,L,F,P}  {A,G,V}  {C}</td>
</tr>
<tr>
<td>P28</td>
<td>{A,G,V}  {H,N,Q,W}  {C}</td>
<td>P58</td>
<td>{I,L,F,P}  {I,L,F,P}  {I,L,F,P}</td>
</tr>
<tr>
<td>P30</td>
<td>{A,G,V}  {R,K}  {I,L,F,P}</td>
<td>P60</td>
<td>{I,L,F,P}  {I,L,F,P}  {H,N,Q,W}</td>
</tr>
<tr>
<td>Label</td>
<td>Feature</td>
<td>Label</td>
<td>Feature</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>P61</td>
<td>[I,L,F,P] [I,L,F,P] [R,K]</td>
<td>P100</td>
<td>[Y,M,T,S] [A,G,V] [I,L,F,P]</td>
</tr>
<tr>
<td>P63</td>
<td>[I,L,F,P] [I,L,F,P] [C]</td>
<td>P102</td>
<td>[Y,M,T,S] [A,G,V] [H,N,Q,W]</td>
</tr>
<tr>
<td>P65</td>
<td>[I,L,F,P] [Y,M,T,S] [I,L,F,P]</td>
<td>P104</td>
<td>[Y,M,T,S] [A,G,V] [D,E]</td>
</tr>
<tr>
<td>P67</td>
<td>[I,L,F,P] [Y,M,T,S] [H,N,Q,W]</td>
<td>P106</td>
<td>[Y,M,T,S] [I,L,F,P] [A,G,V]</td>
</tr>
<tr>
<td>P68</td>
<td>[I,L,F,P] [Y,M,T,S] [R,K]</td>
<td>P107</td>
<td>[Y,M,T,S] [I,L,F,P] [Y,M,T,S]</td>
</tr>
<tr>
<td>P69</td>
<td>[I,L,F,P] [Y,M,T,S] [D,E]</td>
<td>P108</td>
<td>[Y,M,T,S] [I,L,F,P] [Y,M,T,S]</td>
</tr>
<tr>
<td>P70</td>
<td>[I,L,F,P] [Y,M,T,S] [C]</td>
<td>P109</td>
<td>[Y,M,T,S] [I,L,F,P] [H,N,Q,W]</td>
</tr>
<tr>
<td>P71</td>
<td>[I,L,F,P] [H,N,Q,W] [A,G,V]</td>
<td>P110</td>
<td>[Y,M,T,S] [I,L,F,P] [R,K]</td>
</tr>
<tr>
<td>P72</td>
<td>[I,L,F,P] [H,N,Q,W] [I,L,F,P]</td>
<td>P111</td>
<td>[Y,M,T,S] [I,L,F,P] [D,E]</td>
</tr>
<tr>
<td>P73</td>
<td>[I,L,F,P] [H,N,Q,W] [Y,M,T,S]</td>
<td>P112</td>
<td>[Y,M,T,S] [I,L,F,P] [C]</td>
</tr>
<tr>
<td>P74</td>
<td>[I,L,F,P] [H,N,Q,W] [H,N,Q,W]</td>
<td>P113</td>
<td>[Y,M,T,S] [Y,M,T,S] [A,G,V]</td>
</tr>
<tr>
<td>P75</td>
<td>[I,L,F,P] [H,N,Q,W] [R,K]</td>
<td>P114</td>
<td>[Y,M,T,S] [Y,M,T,S] [I,L,F,P]</td>
</tr>
<tr>
<td>P76</td>
<td>[I,L,F,P] [H,N,Q,W] [D,E]</td>
<td>P115</td>
<td>[Y,M,T,S] [Y,M,T,S] [Y,M,T,S]</td>
</tr>
<tr>
<td>P77</td>
<td>[I,L,F,P] [H,N,Q,W] [C]</td>
<td>P116</td>
<td>[Y,M,T,S] [Y,M,T,S] [H,N,Q,W]</td>
</tr>
<tr>
<td>P79</td>
<td>[I,L,F,P] [R,K] [I,L,F,P]</td>
<td>P118</td>
<td>[Y,M,T,S] [Y,M,T,S] [D,E]</td>
</tr>
<tr>
<td>P80</td>
<td>[I,L,F,P] [R,K] [Y,M,T,S]</td>
<td>P119</td>
<td>[Y,M,T,S] [Y,M,T,S] [C]</td>
</tr>
<tr>
<td>P81</td>
<td>[I,L,F,P] [R,K] [H,N,Q,W]</td>
<td>P120</td>
<td>[Y,M,T,S] [H,N,Q,W] [A,G,V]</td>
</tr>
<tr>
<td>P82</td>
<td>[I,L,F,P] [R,K] [R,K]</td>
<td>P121</td>
<td>[Y,M,T,S] [H,N,Q,W] [I,L,F,P]</td>
</tr>
<tr>
<td>P83</td>
<td>[I,L,F,P] [R,K] [D,E]</td>
<td>P122</td>
<td>[Y,M,T,S] [H,N,Q,W] [Y,M,T,S]</td>
</tr>
<tr>
<td>P84</td>
<td>[I,L,F,P] [R,K] [C]</td>
<td>P123</td>
<td>[Y,M,T,S] [H,N,Q,W] [H,N,Q,W]</td>
</tr>
<tr>
<td>P85</td>
<td>[I,L,F,P] [D,E] [A,G,V]</td>
<td>P124</td>
<td>[Y,M,T,S] [H,N,Q,W] [R,K]</td>
</tr>
<tr>
<td>P86</td>
<td>[I,L,F,P] [D,E] [I,L,F,P]</td>
<td>P125</td>
<td>[Y,M,T,S] [H,N,Q,W] [D,E]</td>
</tr>
<tr>
<td>P87</td>
<td>[I,L,F,P] [D,E] [Y,M,T,S]</td>
<td>P126</td>
<td>[Y,M,T,S] [H,N,Q,W] [C]</td>
</tr>
<tr>
<td>P88</td>
<td>[I,L,F,P] [D,E] [H,N,Q,W]</td>
<td>P127</td>
<td>[Y,M,T,S] [R,K] [A,G,V]</td>
</tr>
<tr>
<td>P89</td>
<td>[I,L,F,P] [D,E] [R,K]</td>
<td>P128</td>
<td>[Y,M,T,S] [R,K] [I,L,F,P]</td>
</tr>
<tr>
<td>P90</td>
<td>[I,L,F,P] [D,E] [I,L,F,P]</td>
<td>P129</td>
<td>[Y,M,T,S] [R,K] [Y,M,T,S]</td>
</tr>
<tr>
<td>P91</td>
<td>[I,L,F,P] [D,E] [C]</td>
<td>P130</td>
<td>[Y,M,T,S] [R,K] [H,N,Q,W]</td>
</tr>
<tr>
<td>P92</td>
<td>[I,L,F,P] [C] [A,G,V]</td>
<td>P131</td>
<td>[Y,M,T,S] [R,K] [R,K]</td>
</tr>
<tr>
<td>P93</td>
<td>[I,L,F,P] [C] [I,L,F,P]</td>
<td>P132</td>
<td>[Y,M,T,S] [R,K] [D,E]</td>
</tr>
<tr>
<td>P94</td>
<td>[I,L,F,P] [C] [Y,M,T,S]</td>
<td>P133</td>
<td>[Y,M,T,S] [R,K] [C]</td>
</tr>
<tr>
<td>P95</td>
<td>[I,L,F,P] [C] [H,N,Q,W]</td>
<td>P134</td>
<td>[Y,M,T,S] [D,E] [A,G,V]</td>
</tr>
<tr>
<td>P96</td>
<td>[I,L,F,P] [C] [R,K]</td>
<td>P135</td>
<td>[Y,M,T,S] [D,E] [I,L,F,P]</td>
</tr>
<tr>
<td>P97</td>
<td>[I,L,F,P] [C] [D,E]</td>
<td>P136</td>
<td>[Y,M,T,S] [D,E] [Y,M,T,S]</td>
</tr>
<tr>
<td>P98</td>
<td>[I,L,F,P] [C] [C]</td>
<td>P137</td>
<td>[Y,M,T,S] [D,E] [H,N,Q,W]</td>
</tr>
</tbody>
</table>
Table S3-1 (continued)

<table>
<thead>
<tr>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>P139</td>
<td>{Y,M,T,S} {D,E} {D,E}</td>
<td>P178</td>
<td>{H,N,Q,W} {R,K} {Y,M,T,S}</td>
</tr>
<tr>
<td>P140</td>
<td>{Y,M,T,S} {D,E} {C}</td>
<td>P179</td>
<td>{H,N,Q,W} {R,K} {H,N,Q,W}</td>
</tr>
<tr>
<td>P141</td>
<td>{Y,M,T,S} {C} {A,G,V}</td>
<td>P180</td>
<td>{H,N,Q,W} {R,K} {R,K}</td>
</tr>
<tr>
<td>P142</td>
<td>{Y,M,T,S} {C} {I,L,F,P}</td>
<td>P181</td>
<td>{H,N,Q,W} {R,K} {D,E}</td>
</tr>
<tr>
<td>P143</td>
<td>{Y,M,T,S} {C} {Y,M,T,S}</td>
<td>P182</td>
<td>{H,N,Q,W} {R,K} {C}</td>
</tr>
<tr>
<td>P144</td>
<td>{Y,M,T,S} {C} {H,N,Q,W}</td>
<td>P183</td>
<td>{H,N,Q,W} {D,E} {A,G,V}</td>
</tr>
<tr>
<td>P145</td>
<td>{Y,M,T,S} {C} {R,K}</td>
<td>P184</td>
<td>{H,N,Q,W} {D,E} {I,L,F,P}</td>
</tr>
<tr>
<td>P146</td>
<td>{Y,M,T,S} {C} {D,E}</td>
<td>P185</td>
<td>{H,N,Q,W} {D,E} {Y,M,T,S}</td>
</tr>
<tr>
<td>P147</td>
<td>{Y,M,T,S} {C} {C}</td>
<td>P186</td>
<td>{H,N,Q,W} {D,E} {H,N,Q,W}</td>
</tr>
<tr>
<td>P148</td>
<td>{H,N,Q,W} {A,G,V} {A,G,V}</td>
<td>P187</td>
<td>{H,N,Q,W} {D,E} {R,K}</td>
</tr>
<tr>
<td>P149</td>
<td>{H,N,Q,W} {A,G,V} {I,L,F,P}</td>
<td>P188</td>
<td>{H,N,Q,W} {D,E} {D,E}</td>
</tr>
<tr>
<td>P150</td>
<td>{H,N,Q,W} {A,G,V} {Y,M,T,S}</td>
<td>P189</td>
<td>{H,N,Q,W} {D,E} {C}</td>
</tr>
<tr>
<td>P151</td>
<td>{H,N,Q,W} {A,G,V} {H,N,Q,W}</td>
<td>P190</td>
<td>{H,N,Q,W} {C} {A,G,V}</td>
</tr>
<tr>
<td>P152</td>
<td>{H,N,Q,W} {A,G,V} {R,K}</td>
<td>P191</td>
<td>{H,N,Q,W} {C} {I,L,F,P}</td>
</tr>
<tr>
<td>P153</td>
<td>{H,N,Q,W} {A,G,V} {D,E}</td>
<td>P192</td>
<td>{H,N,Q,W} {C} {Y,M,T,S}</td>
</tr>
<tr>
<td>P154</td>
<td>{H,N,Q,W} {A,G,V} {C}</td>
<td>P193</td>
<td>{H,N,Q,W} {C} {H,N,Q,W}</td>
</tr>
<tr>
<td>P155</td>
<td>{H,N,Q,W} {I,L,F,P} {A,G,V}</td>
<td>P194</td>
<td>{H,N,Q,W} {C} {R,K}</td>
</tr>
<tr>
<td>P156</td>
<td>{H,N,Q,W} {I,L,F,P} {I,L,F,P}</td>
<td>P195</td>
<td>{H,N,Q,W} {C} {D,E}</td>
</tr>
<tr>
<td>P157</td>
<td>{H,N,Q,W} {I,L,F,P} {Y,M,T,S}</td>
<td>P196</td>
<td>{H,N,Q,W} {C} {C}</td>
</tr>
<tr>
<td>P158</td>
<td>{H,N,Q,W} {I,L,F,P} {H,N,Q,W}</td>
<td>P197</td>
<td>{R,K} {A,G,V} {A,G,V}</td>
</tr>
<tr>
<td>P159</td>
<td>{H,N,Q,W} {I,L,F,P} {R,K}</td>
<td>P198</td>
<td>{R,K} {A,G,V} {I,L,F,P}</td>
</tr>
<tr>
<td>P160</td>
<td>{H,N,Q,W} {I,L,F,P} {D,E}</td>
<td>P199</td>
<td>{R,K} {A,G,V} {Y,M,T,S}</td>
</tr>
<tr>
<td>P161</td>
<td>{H,N,Q,W} {I,L,F,P} {C}</td>
<td>P200</td>
<td>{R,K} {A,G,V} {H,N,Q,W}</td>
</tr>
<tr>
<td>P162</td>
<td>{H,N,Q,W} {Y,M,T,S} {A,G,V}</td>
<td>P201</td>
<td>{R,K} {A,G,V} {R,K}</td>
</tr>
<tr>
<td>P163</td>
<td>{H,N,Q,W} {Y,M,T,S} {I,L,F,P}</td>
<td>P202</td>
<td>{R,K} {A,G,V} {D,E}</td>
</tr>
<tr>
<td>P164</td>
<td>{H,N,Q,W} {Y,M,T,S} {Y,M,T,S}</td>
<td>P203</td>
<td>{R,K} {A,G,V} {C}</td>
</tr>
<tr>
<td>P165</td>
<td>{H,N,Q,W} {Y,M,T,S} {H,N,Q,W}</td>
<td>P204</td>
<td>{R,K} {I,L,F,P} {A,G,V}</td>
</tr>
<tr>
<td>P166</td>
<td>{H,N,Q,W} {Y,M,T,S} {R,K}</td>
<td>P205</td>
<td>{R,K} {I,L,F,P} {I,L,F,P}</td>
</tr>
<tr>
<td>P167</td>
<td>{H,N,Q,W} {Y,M,T,S} {D,E}</td>
<td>P206</td>
<td>{R,K} {I,L,F,P} {Y,M,T,S}</td>
</tr>
<tr>
<td>P168</td>
<td>{H,N,Q,W} {Y,M,T,S} {C}</td>
<td>P207</td>
<td>{R,K} {I,L,F,P} {H,N,Q,W}</td>
</tr>
<tr>
<td>P169</td>
<td>{H,N,Q,W} {H,N,Q,W} {A,G,V}</td>
<td>P208</td>
<td>{R,K} {I,L,F,P} {R,K}</td>
</tr>
<tr>
<td>P170</td>
<td>{H,N,Q,W} {H,N,Q,W} {I,L,F,P}</td>
<td>P209</td>
<td>{R,K} {I,L,F,P} {D,E}</td>
</tr>
<tr>
<td>P171</td>
<td>{H,N,Q,W} {H,N,Q,W} {Y,M,T,S}</td>
<td>P210</td>
<td>{R,K} {I,L,F,P} {C}</td>
</tr>
<tr>
<td>P172</td>
<td>{H,N,Q,W} {H,N,Q,W} {H,N,Q,W}</td>
<td>P211</td>
<td>{R,K} {Y,M,T,S} {A,G,V}</td>
</tr>
<tr>
<td>P173</td>
<td>{H,N,Q,W} {H,N,Q,W} {R,K}</td>
<td>P212</td>
<td>{R,K} {Y,M,T,S} {I,L,F,P}</td>
</tr>
<tr>
<td>P174</td>
<td>{H,N,Q,W} {H,N,Q,W} {D,E}</td>
<td>P213</td>
<td>{R,K} {Y,M,T,S} {Y,M,T,S}</td>
</tr>
<tr>
<td>P175</td>
<td>{H,N,Q,W} {H,N,Q,W} {C}</td>
<td>P214</td>
<td>{R,K} {Y,M,T,S} {H,N,Q,W}</td>
</tr>
<tr>
<td>P176</td>
<td>{H,N,Q,W} {R,K} {A,G,V}</td>
<td>P215</td>
<td>{R,K} {Y,M,T,S} {R,K}</td>
</tr>
<tr>
<td>P177</td>
<td>{H,N,Q,W} {R,K} {I,L,F,P}</td>
<td>P216</td>
<td>{R,K} {Y,M,T,S} {D,E}</td>
</tr>
</tbody>
</table>
Table S3-1 (continued)

<table>
<thead>
<tr>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
</tr>
</thead>
</table>
Table S3-1 (continued)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>P295</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P320</td>
</tr>
<tr>
<td>P296</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P321</td>
</tr>
<tr>
<td>P297</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P322</td>
</tr>
<tr>
<td>P298</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P323</td>
</tr>
<tr>
<td>P299</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P324</td>
</tr>
<tr>
<td>P300</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P325</td>
</tr>
<tr>
<td>P301</td>
<td>{C}</td>
<td>{A,G,V}</td>
<td>P326</td>
</tr>
<tr>
<td>P302</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>A,G,V</td>
</tr>
<tr>
<td>P303</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>I,L,F,P</td>
</tr>
<tr>
<td>P304</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>Y,M,T,S</td>
</tr>
<tr>
<td>P305</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>H,N,Q,W</td>
</tr>
<tr>
<td>P306</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>R,K</td>
</tr>
<tr>
<td>P307</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>D,E</td>
</tr>
<tr>
<td>P308</td>
<td>{C}</td>
<td>{I,L,F,P}</td>
<td>C</td>
</tr>
<tr>
<td>P309</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>A,G,V</td>
</tr>
<tr>
<td>P310</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>I,L,F,P</td>
</tr>
<tr>
<td>P311</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>Y,M,T,S</td>
</tr>
<tr>
<td>P312</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>H,N,Q,W</td>
</tr>
<tr>
<td>P313</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>R,K</td>
</tr>
<tr>
<td>P314</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>D,E</td>
</tr>
<tr>
<td>P315</td>
<td>{C}</td>
<td>{Y,M,T,S}</td>
<td>C</td>
</tr>
<tr>
<td>P316</td>
<td>{C}</td>
<td>{H,N,Q,W}</td>
<td>A,G,V</td>
</tr>
<tr>
<td>P317</td>
<td>{C}</td>
<td>{H,N,Q,W}</td>
<td>I,L,F,P</td>
</tr>
<tr>
<td>P318</td>
<td>{C}</td>
<td>{H,N,Q,W}</td>
<td>Y,M,T,S</td>
</tr>
<tr>
<td>P319</td>
<td>{C}</td>
<td>{H,N,Q,W}</td>
<td>H,N,Q,W</td>
</tr>
</tbody>
</table>
RNA Features

Table S3-2 RNA conjoint tetrads utilized in RPIDisorder. The “Label” is the name given to the feature in the vector. The “Feature” column contains the combination of amino acids whose frequency is represented by the Label.

<table>
<thead>
<tr>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>AAAA</td>
<td>R31</td>
<td>AUGC</td>
<td>R61</td>
<td>AGGA</td>
<td>R91</td>
<td>UUCC</td>
</tr>
<tr>
<td>R2</td>
<td>AAAU</td>
<td>R32</td>
<td>AUGG</td>
<td>R62</td>
<td>AGGU</td>
<td>R92</td>
<td>UUCG</td>
</tr>
<tr>
<td>R3</td>
<td>AAAC</td>
<td>R33</td>
<td>ACAA</td>
<td>R63</td>
<td>AGGC</td>
<td>R93</td>
<td>UUGA</td>
</tr>
<tr>
<td>R4</td>
<td>AAAG</td>
<td>R34</td>
<td>ACAU</td>
<td>R64</td>
<td>AGGG</td>
<td>R94</td>
<td>UUGU</td>
</tr>
<tr>
<td>R5</td>
<td>AAUA</td>
<td>R35</td>
<td>ACAC</td>
<td>R65</td>
<td>UAAA</td>
<td>R95</td>
<td>UUGC</td>
</tr>
<tr>
<td>R6</td>
<td>AAUU</td>
<td>R36</td>
<td>ACAG</td>
<td>R66</td>
<td>UAAU</td>
<td>R96</td>
<td>UUGG</td>
</tr>
<tr>
<td>R7</td>
<td>AAUC</td>
<td>R37</td>
<td>ACUA</td>
<td>R67</td>
<td>UAAC</td>
<td>R97</td>
<td>UCAA</td>
</tr>
<tr>
<td>R8</td>
<td>AAUG</td>
<td>R38</td>
<td>ACUU</td>
<td>R68</td>
<td>UAAG</td>
<td>R98</td>
<td>UCAU</td>
</tr>
<tr>
<td>R9</td>
<td>ACAA</td>
<td>R39</td>
<td>ACUC</td>
<td>R69</td>
<td>UAUA</td>
<td>R99</td>
<td>UCAC</td>
</tr>
<tr>
<td>R10</td>
<td>AACU</td>
<td>R40</td>
<td>ACUG</td>
<td>R70</td>
<td>UAUU</td>
<td>R100</td>
<td>UCAG</td>
</tr>
<tr>
<td>R11</td>
<td>AACC</td>
<td>R41</td>
<td>ACCA</td>
<td>R71</td>
<td>UAUC</td>
<td>R101</td>
<td>UCUA</td>
</tr>
<tr>
<td>R12</td>
<td>AACG</td>
<td>R42</td>
<td>ACCU</td>
<td>R72</td>
<td>UAUG</td>
<td>R102</td>
<td>UCUU</td>
</tr>
<tr>
<td>R13</td>
<td>AAGA</td>
<td>R43</td>
<td>ACCC</td>
<td>R73</td>
<td>UACA</td>
<td>R103</td>
<td>UCUC</td>
</tr>
<tr>
<td>R14</td>
<td>AAGU</td>
<td>R44</td>
<td>ACCG</td>
<td>R74</td>
<td>UACU</td>
<td>R104</td>
<td>UCCG</td>
</tr>
<tr>
<td>R15</td>
<td>AAGC</td>
<td>R45</td>
<td>ACGA</td>
<td>R75</td>
<td>UACC</td>
<td>R105</td>
<td>UCCA</td>
</tr>
<tr>
<td>R16</td>
<td>AAGG</td>
<td>R46</td>
<td>ACGU</td>
<td>R76</td>
<td>UACG</td>
<td>R106</td>
<td>UCCU</td>
</tr>
<tr>
<td>R17</td>
<td>AUAA</td>
<td>R47</td>
<td>ACGC</td>
<td>R77</td>
<td>UAGA</td>
<td>R107</td>
<td>UCCC</td>
</tr>
<tr>
<td>R18</td>
<td>AUAU</td>
<td>R48</td>
<td>ACGG</td>
<td>R78</td>
<td>UAGU</td>
<td>R108</td>
<td>UCCG</td>
</tr>
<tr>
<td>R19</td>
<td>AUAC</td>
<td>R49</td>
<td>AGAA</td>
<td>R79</td>
<td>UAGC</td>
<td>R109</td>
<td>UCGA</td>
</tr>
<tr>
<td>R20</td>
<td>AUAG</td>
<td>R50</td>
<td>AGAU</td>
<td>R80</td>
<td>UAGG</td>
<td>R110</td>
<td>UCGU</td>
</tr>
<tr>
<td>R21</td>
<td>AUUA</td>
<td>R51</td>
<td>AGAC</td>
<td>R81</td>
<td>UUAA</td>
<td>R111</td>
<td>UCGC</td>
</tr>
<tr>
<td>R22</td>
<td>AUUU</td>
<td>R52</td>
<td>AGAG</td>
<td>R82</td>
<td>UUAU</td>
<td>R112</td>
<td>UCGG</td>
</tr>
<tr>
<td>R23</td>
<td>AUUC</td>
<td>R53</td>
<td>AGUA</td>
<td>R83</td>
<td>UUAC</td>
<td>R113</td>
<td>UGAA</td>
</tr>
<tr>
<td>R24</td>
<td>AUUG</td>
<td>R54</td>
<td>AGUU</td>
<td>R84</td>
<td>UUAG</td>
<td>R114</td>
<td>UGUA</td>
</tr>
<tr>
<td>R25</td>
<td>AUCA</td>
<td>R55</td>
<td>AGUC</td>
<td>R85</td>
<td>UUUA</td>
<td>R115</td>
<td>UGAC</td>
</tr>
<tr>
<td>R26</td>
<td>AUCU</td>
<td>R56</td>
<td>AGUG</td>
<td>R86</td>
<td>UUUU</td>
<td>R116</td>
<td>UGGU</td>
</tr>
<tr>
<td>R27</td>
<td>AUCC</td>
<td>R57</td>
<td>AGCA</td>
<td>R87</td>
<td>UUUC</td>
<td>R117</td>
<td>UGUA</td>
</tr>
<tr>
<td>R28</td>
<td>AUUG</td>
<td>R58</td>
<td>AGCU</td>
<td>R88</td>
<td>UUUG</td>
<td>R118</td>
<td>UGUU</td>
</tr>
<tr>
<td>R29</td>
<td>AUGA</td>
<td>R59</td>
<td>AGCC</td>
<td>R89</td>
<td>UUCA</td>
<td>R119</td>
<td>UGUC</td>
</tr>
<tr>
<td>R30</td>
<td>AUGU</td>
<td>R60</td>
<td>AGCG</td>
<td>R90</td>
<td>UUCU</td>
<td>R120</td>
<td>UGUG</td>
</tr>
</tbody>
</table>
Table S3-2 (continued)

<table>
<thead>
<tr>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
<th>Label</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>R121</td>
<td>UGCA</td>
<td>R156</td>
<td>CUCG</td>
<td>R191</td>
<td>CGGC</td>
<td>R226</td>
<td>GCAU</td>
</tr>
<tr>
<td>R122</td>
<td>UGCU</td>
<td>R157</td>
<td>CUGA</td>
<td>R192</td>
<td>CGGG</td>
<td>R227</td>
<td>GCAC</td>
</tr>
<tr>
<td>R123</td>
<td>UGCC</td>
<td>R158</td>
<td>CUGU</td>
<td>R193</td>
<td>GAAA</td>
<td>R228</td>
<td>GCAG</td>
</tr>
<tr>
<td>R124</td>
<td>UGCG</td>
<td>R159</td>
<td>CUGC</td>
<td>R194</td>
<td>GAAU</td>
<td>R229</td>
<td>GCUA</td>
</tr>
<tr>
<td>R125</td>
<td>UGGA</td>
<td>R160</td>
<td>CUGG</td>
<td>R195</td>
<td>GAAC</td>
<td>R230</td>
<td>GCUU</td>
</tr>
<tr>
<td>R126</td>
<td>UGUA</td>
<td>R161</td>
<td>CCAA</td>
<td>R196</td>
<td>GAAG</td>
<td>R231</td>
<td>GCUC</td>
</tr>
<tr>
<td>R127</td>
<td>UGGC</td>
<td>R162</td>
<td>CCAU</td>
<td>R197</td>
<td>GAUA</td>
<td>R232</td>
<td>GCUG</td>
</tr>
<tr>
<td>R128</td>
<td>UGGG</td>
<td>R163</td>
<td>CCAC</td>
<td>R198</td>
<td>GAUU</td>
<td>R233</td>
<td>GCCA</td>
</tr>
<tr>
<td>R129</td>
<td>CAAA</td>
<td>R164</td>
<td>CCAG</td>
<td>R199</td>
<td>GAUC</td>
<td>R234</td>
<td>GCCU</td>
</tr>
<tr>
<td>R130</td>
<td>CAAU</td>
<td>R165</td>
<td>CCUA</td>
<td>R200</td>
<td>GAUG</td>
<td>R235</td>
<td>GCCG</td>
</tr>
<tr>
<td>R131</td>
<td>CAAC</td>
<td>R166</td>
<td>CCUC</td>
<td>R201</td>
<td>GACA</td>
<td>R236</td>
<td>GCCG</td>
</tr>
<tr>
<td>R132</td>
<td>CAAG</td>
<td>R167</td>
<td>CCUC</td>
<td>R202</td>
<td>GACU</td>
<td>R237</td>
<td>GCGA</td>
</tr>
<tr>
<td>R133</td>
<td>CAUA</td>
<td>R168</td>
<td>CCUG</td>
<td>R203</td>
<td>GACC</td>
<td>R238</td>
<td>GCGU</td>
</tr>
<tr>
<td>R134</td>
<td>CAUU</td>
<td>R169</td>
<td>CCCA</td>
<td>R204</td>
<td>GACG</td>
<td>R239</td>
<td>GCGC</td>
</tr>
<tr>
<td>R135</td>
<td>CAUC</td>
<td>R170</td>
<td>CCCU</td>
<td>R205</td>
<td>GAGA</td>
<td>R240</td>
<td>GCGG</td>
</tr>
<tr>
<td>R136</td>
<td>CAUG</td>
<td>R171</td>
<td>CCCC</td>
<td>R206</td>
<td>GAGU</td>
<td>R241</td>
<td>GGAA</td>
</tr>
<tr>
<td>R137</td>
<td>CACA</td>
<td>R172</td>
<td>CCCG</td>
<td>R207</td>
<td>GAGC</td>
<td>R242</td>
<td>GAUA</td>
</tr>
<tr>
<td>R138</td>
<td>CACU</td>
<td>R173</td>
<td>CCGA</td>
<td>R208</td>
<td>GAGG</td>
<td>R243</td>
<td>GGAC</td>
</tr>
<tr>
<td>R139</td>
<td>CACC</td>
<td>R174</td>
<td>CCGU</td>
<td>R209</td>
<td>GAAC</td>
<td>R244</td>
<td>GAGG</td>
</tr>
<tr>
<td>R140</td>
<td>CACG</td>
<td>R175</td>
<td>CCGG</td>
<td>R210</td>
<td>GAAC</td>
<td>R245</td>
<td>GGUA</td>
</tr>
<tr>
<td>R141</td>
<td>CAGA</td>
<td>R176</td>
<td>CCGG</td>
<td>R211</td>
<td>GUAC</td>
<td>R246</td>
<td>GGUU</td>
</tr>
<tr>
<td>R142</td>
<td>CAGU</td>
<td>R177</td>
<td>CGAA</td>
<td>R212</td>
<td>GUAG</td>
<td>R247</td>
<td>GGUC</td>
</tr>
<tr>
<td>R143</td>
<td>CAGC</td>
<td>R178</td>
<td>CGAU</td>
<td>R213</td>
<td>GUUA</td>
<td>R248</td>
<td>GGUG</td>
</tr>
<tr>
<td>R144</td>
<td>CAGG</td>
<td>R179</td>
<td>CGAC</td>
<td>R214</td>
<td>GUUC</td>
<td>R249</td>
<td>GGCU</td>
</tr>
<tr>
<td>R145</td>
<td>CUAA</td>
<td>R180</td>
<td>CGAG</td>
<td>R215</td>
<td>GUUC</td>
<td>R250</td>
<td>GGCU</td>
</tr>
<tr>
<td>R146</td>
<td>CUAA</td>
<td>R181</td>
<td>CGUA</td>
<td>R216</td>
<td>GUUG</td>
<td>R251</td>
<td>GGCC</td>
</tr>
<tr>
<td>R147</td>
<td>CUAC</td>
<td>R182</td>
<td>CGUU</td>
<td>R217</td>
<td>GUCA</td>
<td>R252</td>
<td>GGCG</td>
</tr>
<tr>
<td>R148</td>
<td>CUAU</td>
<td>R183</td>
<td>CGUC</td>
<td>R218</td>
<td>GUUC</td>
<td>R253</td>
<td>GGGG</td>
</tr>
<tr>
<td>R149</td>
<td>CUUA</td>
<td>R184</td>
<td>CGUG</td>
<td>R219</td>
<td>GUCC</td>
<td>R254</td>
<td>GGGU</td>
</tr>
<tr>
<td>R150</td>
<td>CUUU</td>
<td>R185</td>
<td>CGCA</td>
<td>R220</td>
<td>GUCG</td>
<td>R255</td>
<td>GGGC</td>
</tr>
<tr>
<td>R151</td>
<td>CUUC</td>
<td>R186</td>
<td>CGCU</td>
<td>R221</td>
<td>GUGA</td>
<td>R256</td>
<td>GGGG</td>
</tr>
<tr>
<td>R152</td>
<td>CUUG</td>
<td>R187</td>
<td>CGCC</td>
<td>R222</td>
<td>GUGU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R153</td>
<td>CUCA</td>
<td>R188</td>
<td>CGCG</td>
<td>R223</td>
<td>GUGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R154</td>
<td>CUCU</td>
<td>R189</td>
<td>CGGA</td>
<td>R224</td>
<td>GUGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R155</td>
<td>CUCC</td>
<td>R190</td>
<td>CGGU</td>
<td>R225</td>
<td>GCAA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Disorder Feature Tables

Table S3-3 Importance metrics for the R randomForest classifier, ordered in descending order based on Interacting importance, for the top 5% Interacting features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. See “Methods” for description of disorder features. The first three columns are the Mean Decrease in Accuracy (%) for determining the Interacting class, Non-interacting class, and overall MDA, respectively.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
<tr>
<td>D1</td>
<td>13.66</td>
<td>8.94</td>
<td>14.17</td>
<td>2.96</td>
</tr>
<tr>
<td>D5</td>
<td>13.36</td>
<td>8.84</td>
<td>14.63</td>
<td>2.61</td>
</tr>
<tr>
<td>P251</td>
<td>13.28</td>
<td>8.68</td>
<td>13.97</td>
<td>2.36</td>
</tr>
<tr>
<td>P57</td>
<td>13.10</td>
<td>6.94</td>
<td>13.82</td>
<td>2.59</td>
</tr>
<tr>
<td>P11</td>
<td>13.05</td>
<td>6.00</td>
<td>13.28</td>
<td>1.67</td>
</tr>
<tr>
<td>P201</td>
<td>12.97</td>
<td>6.29</td>
<td>12.67</td>
<td>2.14</td>
</tr>
<tr>
<td>P152</td>
<td>12.79</td>
<td>4.75</td>
<td>12.03</td>
<td>1.86</td>
</tr>
<tr>
<td>P2</td>
<td>12.78</td>
<td>6.71</td>
<td>13.10</td>
<td>2.57</td>
</tr>
<tr>
<td>P225</td>
<td>12.75</td>
<td>7.78</td>
<td>12.67</td>
<td>2.35</td>
</tr>
<tr>
<td>P101</td>
<td>12.73</td>
<td>7.68</td>
<td>13.35</td>
<td>2.16</td>
</tr>
<tr>
<td>P62</td>
<td>12.71</td>
<td>6.92</td>
<td>12.18</td>
<td>2.05</td>
</tr>
<tr>
<td>P8</td>
<td>12.63</td>
<td>5.32</td>
<td>11.80</td>
<td>2.12</td>
</tr>
<tr>
<td>P127</td>
<td>12.59</td>
<td>7.60</td>
<td>12.91</td>
<td>2.04</td>
</tr>
<tr>
<td>P61</td>
<td>12.58</td>
<td>7.71</td>
<td>12.54</td>
<td>2.43</td>
</tr>
<tr>
<td>P106</td>
<td>12.39</td>
<td>6.01</td>
<td>12.27</td>
<td>1.62</td>
</tr>
<tr>
<td>P85</td>
<td>12.37</td>
<td>5.64</td>
<td>11.92</td>
<td>2.16</td>
</tr>
<tr>
<td>P29</td>
<td>12.31</td>
<td>8.30</td>
<td>13.11</td>
<td>2.06</td>
</tr>
<tr>
<td>P12</td>
<td>12.30</td>
<td>7.59</td>
<td>12.08</td>
<td>2.34</td>
</tr>
<tr>
<td>ordPhillic</td>
<td>12.56</td>
<td>6.71</td>
<td>12.20</td>
<td>3.38</td>
</tr>
<tr>
<td>P106</td>
<td>12.39</td>
<td>6.01</td>
<td>12.27</td>
<td>1.62</td>
</tr>
<tr>
<td>disNeut</td>
<td>12.37</td>
<td>5.64</td>
<td>11.92</td>
<td>2.16</td>
</tr>
<tr>
<td>P257</td>
<td>12.22</td>
<td>5.41</td>
<td>12.19</td>
<td>1.73</td>
</tr>
<tr>
<td>P13</td>
<td>12.27</td>
<td>7.66</td>
<td>12.90</td>
<td>2.00</td>
</tr>
<tr>
<td>P5</td>
<td>12.22</td>
<td>7.98</td>
<td>13.37</td>
<td>2.11</td>
</tr>
<tr>
<td>P212</td>
<td>12.16</td>
<td>5.59</td>
<td>13.03</td>
<td>1.80</td>
</tr>
<tr>
<td>P90</td>
<td>12.06</td>
<td>6.30</td>
<td>12.44</td>
<td>1.59</td>
</tr>
<tr>
<td>P90</td>
<td>12.05</td>
<td>7.20</td>
<td>11.72</td>
<td>3.93</td>
</tr>
<tr>
<td>disMod</td>
<td>12.30</td>
<td>6.89</td>
<td>13.23</td>
<td>2.14</td>
</tr>
<tr>
<td>P275</td>
<td>12.26</td>
<td>7.62</td>
<td>12.61</td>
<td>1.59</td>
</tr>
<tr>
<td>P13</td>
<td>12.27</td>
<td>7.66</td>
<td>12.90</td>
<td>2.00</td>
</tr>
<tr>
<td>P5</td>
<td>12.22</td>
<td>7.98</td>
<td>13.37</td>
<td>2.11</td>
</tr>
<tr>
<td>P4</td>
<td>12.22</td>
<td>5.41</td>
<td>12.19</td>
<td>1.73</td>
</tr>
<tr>
<td>P32</td>
<td>12.21</td>
<td>6.35</td>
<td>12.37</td>
<td>1.48</td>
</tr>
<tr>
<td>P212</td>
<td>12.16</td>
<td>5.59</td>
<td>13.03</td>
<td>1.80</td>
</tr>
<tr>
<td>P90</td>
<td>12.06</td>
<td>6.30</td>
<td>12.44</td>
<td>1.59</td>
</tr>
<tr>
<td>ordNeut</td>
<td>12.05</td>
<td>7.20</td>
<td>11.72</td>
<td>3.93</td>
</tr>
<tr>
<td>D6</td>
<td>11.99</td>
<td>7.45</td>
<td>12.22</td>
<td>2.66</td>
</tr>
<tr>
<td>P85</td>
<td>11.91</td>
<td>5.88</td>
<td>11.82</td>
<td>2.02</td>
</tr>
</tbody>
</table>
Table S3-4 Importance metrics for the R randomForest classifier, ordered in descending order based on Non-interacting importance (Nonint), for the top 5% of Non-interacting features. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. RNA conjoint tetrad features start with an “R”. See “Methods” for description of disorder features.

<table>
<thead>
<tr>
<th></th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordNeg</td>
<td>10.78</td>
<td>9.71</td>
<td>12.16</td>
<td>3.42</td>
</tr>
<tr>
<td>D1</td>
<td>13.66</td>
<td>8.94</td>
<td>14.17</td>
<td>2.96</td>
</tr>
<tr>
<td>D5</td>
<td>13.36</td>
<td>8.84</td>
<td>14.63</td>
<td>2.61</td>
</tr>
<tr>
<td>R173</td>
<td>5.98</td>
<td>8.77</td>
<td>9.61</td>
<td>1.98</td>
</tr>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
<tr>
<td>P251</td>
<td>13.28</td>
<td>8.68</td>
<td>13.97</td>
<td>2.36</td>
</tr>
<tr>
<td>R208</td>
<td>4.82</td>
<td>8.63</td>
<td>8.51</td>
<td>2.13</td>
</tr>
<tr>
<td>R41</td>
<td>6.22</td>
<td>8.62</td>
<td>9.87</td>
<td>2.12</td>
</tr>
<tr>
<td>R189</td>
<td>5.88</td>
<td>8.35</td>
<td>7.81</td>
<td>1.98</td>
</tr>
<tr>
<td>P29</td>
<td>12.31</td>
<td>8.30</td>
<td>13.11</td>
<td>2.06</td>
</tr>
<tr>
<td>R190</td>
<td>6.01</td>
<td>8.23</td>
<td>7.82</td>
<td>2.44</td>
</tr>
<tr>
<td>P16</td>
<td>10.74</td>
<td>8.05</td>
<td>10.86</td>
<td>2.07</td>
</tr>
<tr>
<td>P5</td>
<td>12.22</td>
<td>7.98</td>
<td>13.37</td>
<td>2.11</td>
</tr>
<tr>
<td>R113</td>
<td>5.36</td>
<td>7.97</td>
<td>7.94</td>
<td>3.50</td>
</tr>
<tr>
<td>P33</td>
<td>11.73</td>
<td>7.94</td>
<td>11.75</td>
<td>2.28</td>
</tr>
<tr>
<td>disNeg</td>
<td>11.89</td>
<td>7.94</td>
<td>12.83</td>
<td>2.13</td>
</tr>
<tr>
<td>P58</td>
<td>9.43</td>
<td>7.92</td>
<td>10.55</td>
<td>2.52</td>
</tr>
<tr>
<td>P36</td>
<td>11.35</td>
<td>7.89</td>
<td>11.49</td>
<td>2.79</td>
</tr>
<tr>
<td>P79</td>
<td>9.45</td>
<td>7.79</td>
<td>9.83</td>
<td>2.19</td>
</tr>
<tr>
<td>P225</td>
<td>12.75</td>
<td>7.78</td>
<td>12.67</td>
<td>2.35</td>
</tr>
<tr>
<td>R206</td>
<td>4.35</td>
<td>7.78</td>
<td>6.84</td>
<td>1.68</td>
</tr>
<tr>
<td>R254</td>
<td>6.25</td>
<td>7.78</td>
<td>8.43</td>
<td>2.57</td>
</tr>
<tr>
<td>R196</td>
<td>6.04</td>
<td>7.72</td>
<td>7.77</td>
<td>2.23</td>
</tr>
<tr>
<td>P61</td>
<td>12.58</td>
<td>7.71</td>
<td>12.54</td>
<td>2.43</td>
</tr>
<tr>
<td>P99</td>
<td>11.23</td>
<td>7.71</td>
<td>11.75</td>
<td>2.20</td>
</tr>
<tr>
<td>P82</td>
<td>10.20</td>
<td>7.69</td>
<td>10.18</td>
<td>2.59</td>
</tr>
<tr>
<td>P101</td>
<td>12.73</td>
<td>7.68</td>
<td>13.35</td>
<td>2.16</td>
</tr>
<tr>
<td>R47</td>
<td>4.47</td>
<td>7.67</td>
<td>6.10</td>
<td>1.72</td>
</tr>
<tr>
<td>P100</td>
<td>11.21</td>
<td>7.66</td>
<td>12.21</td>
<td>2.29</td>
</tr>
<tr>
<td>P13</td>
<td>12.27</td>
<td>7.66</td>
<td>12.90</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Table S3-5 Importance metrics for the R randomForest classifier, ordered in descending order based on overall Mean Decrease in Accuracy, for the top 5% of features as determined by MDA. Disordered features are highlighted in blue. Protein conjoint triad features start with a “P”. See “Methods” for description of disorder features.

```
<table>
<thead>
<tr>
<th></th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
<tr>
<td>D5</td>
<td>13.36</td>
<td>8.84</td>
<td>14.63</td>
<td>2.61</td>
</tr>
<tr>
<td>D1</td>
<td>13.66</td>
<td>8.94</td>
<td>14.17</td>
<td>2.96</td>
</tr>
<tr>
<td>P251</td>
<td>13.28</td>
<td>8.68</td>
<td>13.97</td>
<td>2.36</td>
</tr>
<tr>
<td>P57</td>
<td>13.10</td>
<td>6.94</td>
<td>13.82</td>
<td>2.59</td>
</tr>
<tr>
<td>P5</td>
<td>12.22</td>
<td>7.98</td>
<td>13.37</td>
<td>2.11</td>
</tr>
<tr>
<td>P101</td>
<td>12.73</td>
<td>7.68</td>
<td>13.35</td>
<td>2.16</td>
</tr>
<tr>
<td>P11</td>
<td>13.05</td>
<td>6.00</td>
<td>13.28</td>
<td>1.67</td>
</tr>
<tr>
<td>disMod</td>
<td>12.30</td>
<td>6.89</td>
<td>13.23</td>
<td>2.14</td>
</tr>
<tr>
<td>P29</td>
<td>12.31</td>
<td>8.30</td>
<td>13.11</td>
<td>2.06</td>
</tr>
<tr>
<td>P2</td>
<td>12.78</td>
<td>6.71</td>
<td>13.10</td>
<td>2.57</td>
</tr>
<tr>
<td>P108</td>
<td>11.66</td>
<td>7.50</td>
<td>13.08</td>
<td>1.71</td>
</tr>
<tr>
<td>P212</td>
<td>12.16</td>
<td>5.59</td>
<td>13.03</td>
<td>1.80</td>
</tr>
<tr>
<td>P127</td>
<td>12.59</td>
<td>7.60</td>
<td>12.91</td>
<td>2.04</td>
</tr>
<tr>
<td>P13</td>
<td>12.27</td>
<td>7.66</td>
<td>12.90</td>
<td>2.00</td>
</tr>
<tr>
<td>disNeg</td>
<td>11.89</td>
<td>7.94</td>
<td>12.83</td>
<td>2.13</td>
</tr>
<tr>
<td>P201</td>
<td>12.97</td>
<td>6.29</td>
<td>12.67</td>
<td>2.14</td>
</tr>
<tr>
<td>P225</td>
<td>12.75</td>
<td>7.78</td>
<td>12.67</td>
<td>2.35</td>
</tr>
<tr>
<td>P30</td>
<td>11.82</td>
<td>6.24</td>
<td>12.65</td>
<td>2.07</td>
</tr>
<tr>
<td>P275</td>
<td>12.27</td>
<td>6.72</td>
<td>12.61</td>
<td>1.59</td>
</tr>
<tr>
<td>P61</td>
<td>12.58</td>
<td>7.71</td>
<td>12.54</td>
<td>2.43</td>
</tr>
<tr>
<td>disPhillic</td>
<td>11.82</td>
<td>7.05</td>
<td>12.53</td>
<td>2.34</td>
</tr>
<tr>
<td>P90</td>
<td>12.06</td>
<td>6.30</td>
<td>12.44</td>
<td>1.59</td>
</tr>
<tr>
<td>P4</td>
<td>11.37</td>
<td>6.13</td>
<td>12.42</td>
<td>1.67</td>
</tr>
<tr>
<td>P55</td>
<td>11.64</td>
<td>6.28</td>
<td>12.40</td>
<td>1.81</td>
</tr>
<tr>
<td>P32</td>
<td>12.21</td>
<td>6.35</td>
<td>12.37</td>
<td>1.48</td>
</tr>
<tr>
<td>P106</td>
<td>12.39</td>
<td>6.01</td>
<td>12.27</td>
<td>1.62</td>
</tr>
<tr>
<td>D6</td>
<td>11.99</td>
<td>7.45</td>
<td>12.22</td>
<td>2.66</td>
</tr>
<tr>
<td>P100</td>
<td>11.21</td>
<td>7.66</td>
<td>12.21</td>
<td>2.29</td>
</tr>
<tr>
<td>ordPhillic</td>
<td>12.56</td>
<td>6.71</td>
<td>12.20</td>
<td>3.38</td>
</tr>
<tr>
<td>P257</td>
<td>12.22</td>
<td>5.41</td>
<td>12.19</td>
<td>1.73</td>
</tr>
</tbody>
</table>
```
Table S3-6 Importance metrics for the R randomForest classifier, ordered in descending order based on node impurity (MeanDecreaseGini), for the top 5% of features as determined by Gini. Disordered features are highlighted in blue. RNA conjoint tetrad features start with an “R”. See “Methods” for description of disorder features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Interacting</th>
<th>Nonint</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>R6</td>
<td>5.42</td>
<td>5.39</td>
<td>6.08</td>
<td>5.57</td>
</tr>
<tr>
<td>R33</td>
<td>4.44</td>
<td>4.63</td>
<td>5.21</td>
<td>4.30</td>
</tr>
<tr>
<td>R120</td>
<td>5.27</td>
<td>5.47</td>
<td>5.98</td>
<td>4.10</td>
</tr>
<tr>
<td>ordNeut</td>
<td>12.05</td>
<td>7.20</td>
<td>11.72</td>
<td>3.93</td>
</tr>
<tr>
<td>R150</td>
<td>5.73</td>
<td>6.26</td>
<td>6.59</td>
<td>3.89</td>
</tr>
<tr>
<td>R31</td>
<td>5.24</td>
<td>6.02</td>
<td>6.30</td>
<td>3.84</td>
</tr>
<tr>
<td>R242</td>
<td>5.36</td>
<td>7.05</td>
<td>6.29</td>
<td>3.80</td>
</tr>
<tr>
<td>R49</td>
<td>4.62</td>
<td>7.19</td>
<td>6.90</td>
<td>3.74</td>
</tr>
<tr>
<td>ordPos</td>
<td>14.46</td>
<td>8.70</td>
<td>14.92</td>
<td>3.69</td>
</tr>
<tr>
<td>R113</td>
<td>5.36</td>
<td>7.97</td>
<td>7.94</td>
<td>3.50</td>
</tr>
<tr>
<td>R115</td>
<td>4.53</td>
<td>4.78</td>
<td>4.99</td>
<td>3.47</td>
</tr>
<tr>
<td>R106</td>
<td>5.88</td>
<td>5.58</td>
<td>6.61</td>
<td>3.46</td>
</tr>
<tr>
<td>R194</td>
<td>5.16</td>
<td>4.68</td>
<td>5.67</td>
<td>3.43</td>
</tr>
<tr>
<td>R34</td>
<td>5.21</td>
<td>3.17</td>
<td>5.33</td>
<td>3.43</td>
</tr>
<tr>
<td>ordNeg</td>
<td>10.78</td>
<td>9.71</td>
<td>12.16</td>
<td>3.42</td>
</tr>
<tr>
<td>R64</td>
<td>5.71</td>
<td>5.47</td>
<td>7.26</td>
<td>3.38</td>
</tr>
<tr>
<td>ordPhillic</td>
<td>12.56</td>
<td>6.71</td>
<td>12.20</td>
<td>3.38</td>
</tr>
<tr>
<td>ordPhobic</td>
<td>11.55</td>
<td>7.31</td>
<td>11.64</td>
<td>3.36</td>
</tr>
<tr>
<td>R202</td>
<td>4.75</td>
<td>4.07</td>
<td>5.41</td>
<td>3.33</td>
</tr>
<tr>
<td>R102</td>
<td>5.06</td>
<td>3.88</td>
<td>5.42</td>
<td>3.30</td>
</tr>
<tr>
<td>R230</td>
<td>5.11</td>
<td>4.55</td>
<td>5.47</td>
<td>3.27</td>
</tr>
<tr>
<td>R14</td>
<td>5.23</td>
<td>6.80</td>
<td>6.72</td>
<td>3.26</td>
</tr>
<tr>
<td>R54</td>
<td>4.71</td>
<td>6.11</td>
<td>5.45</td>
<td>3.24</td>
</tr>
<tr>
<td>R96</td>
<td>4.67</td>
<td>4.52</td>
<td>5.24</td>
<td>3.22</td>
</tr>
<tr>
<td>R65</td>
<td>6.16</td>
<td>5.10</td>
<td>6.78</td>
<td>3.22</td>
</tr>
<tr>
<td>R5</td>
<td>4.53</td>
<td>3.96</td>
<td>4.73</td>
<td>3.20</td>
</tr>
<tr>
<td>R21</td>
<td>5.00</td>
<td>6.03</td>
<td>6.02</td>
<td>3.19</td>
</tr>
<tr>
<td>R105</td>
<td>5.76</td>
<td>2.79</td>
<td>5.81</td>
<td>3.18</td>
</tr>
<tr>
<td>R117</td>
<td>5.00</td>
<td>6.55</td>
<td>5.64</td>
<td>3.18</td>
</tr>
<tr>
<td>R222</td>
<td>4.72</td>
<td>6.49</td>
<td>6.07</td>
<td>3.15</td>
</tr>
</tbody>
</table>
CHAPTER 4. MEDJED: A MACHINE LEARNING REGRESSION MODEL FOR PREDICTING THE EXTENT OF MICROHOMOLOGY-MEDIATED END JOINING REPAIR IN RESPONSE TO DOUBLE-STRANDED DNA BREAKS

Abstract

The development of precise DNA editing nucleases that induce double-strand breaks (DSBs) - including zinc finger nucleases, TALENs, and CRISPR/Cas systems - has revolutionized gene editing and genome engineering. Endogenous DNA DSB repair mechanisms can be leveraged to improve editing efficiency and precision. While the non-homologous end joining (NHEJ) and homologous recombination (HR) DNA DSB repair pathways have already been the subject of a great deal of investigation, an alternative pathway, microhomology-mediated end joining (MMEJ), remains relatively unexplored. However, the MMEJ pathway's ability to produce reproducible and efficient deletions in the course of repair makes it an ideal pathway for use in gene knockouts.

MEDJED, (Microhomology Evoked Deletion Judication Elucidation) is a random forest machine learning-based method for predicting the extent to which the site of a targeted DNA DSB will be repaired using the MMEJ repair pathway. On an independent test set of 24 HeLa cell DSB sites, MEDJED achieved a Pearson Correlation Coefficient (PCC) of 81.36%, Mean Absolute Error (MAE) of 10.96%, and Root Mean Square Error (RMSE) of 13.09%. This performance demonstrates MEDJED's value as a tool for researchers who wish to leverage MMEJ to produce efficient and precise gene knock outs.

Introduction & Background

One of the most important applications of gene editing technology is the creation of gene knockouts for understanding gene function; as of April 01, 2019, more than 11,800 articles indexed in PubMed Central contain the word “knockout” in their title, and the
overwhelming majority describe or investigate phenotypes associated with gene knockouts [PubMed Central search].

In addition to basic research on gene function, gene knockouts are of critical importance in drug discovery and clinical applications. The creation of gene knockouts that recapitulate human disease phenotypes in model organisms, such as mice or zebrafish, allows for rapid and large-scale small molecule and chemical library screening to identify potential therapeutic agents [Parng et al. 2002, Dinda & Baraban 2015, Zuberi & Lutz 2016, Kithcart & MacRae 2017, Gehrig et al. 2018]. Gene knockouts have promising clinical applications in gene therapies to directly treat diseases. For example, tri-nucleotide repeat extension disorders (e.g., Huntington’s chorea [Shin et al. 2016]) could potentially be treated by knocking out the defective protein or excising repeated regions [Cox et al. 2015]. Several ongoing clinical studies are underway to examine the efficacy of knocking out the CCR5 gene (which produces the receptor that macrophage-tropic HIV-1 uses to infect CD4+ T cells) in patient T cells [Case Western Reserve, City of Hope Medical].

The ability to induce a double-strand break (DSB) in a gene using targeted nucleases such as CRISPR/Cas9 systems [Mali et al. 2013], transcription activator-like effector nucleases (TALENs) [Boch 2011], or zinc finger nucleases (ZFNs) [Kim et al. 1996], [reviewed in Chandrasegaran & Carroll 2016] and then leverage endogenous DNA repair pathways to knockout the gene through the process of repair, is highly desirable.

**DNA Double Strand Break (DSB) Repair in Gene Editing**

In the early days of gene editing, knockouts were generated through the introduction of a DNA fragment or construct containing a large region of sequence homology shared with the targeted endogenous DNA locus, but modified such that the targeted gene would become non-functional or deleted (e.g., by swapping the targeted gene for a drug-resistance gene)
Successful homology-directed repair (HDR) at the homology site resulted in the incorporation of the desired modification into the endogenous gene locus, leading to a targeted knockout or knock-down. Unfortunately, this approach is highly inefficient [Vasquez et al. 2001]. Poly-allelic organisms frequently have un-edited alleles after this process; additionally, the homologous recombination (HR) repair pathway is active only during the S and G2 phases of the cell cycle, and even during those phases it is out-competed by other DNA repair mechanisms such that it typically constitutes only ~0.1% of repair events in mouse-embryo derived stem cells [Thomas & Capecchi, 1987 Cell].

With the discovery and development of DSB-inducing gene editing nucleases, including ZFNs, TALENs, and CRISPR/Cas9 enzymes, an alternative repair mechanism called non-homologous end joining (NHEJ) has gained popularity. NHEJ, in contrast to HR, is highly efficient at repairing DSBs. This is largely due to the fact that NHEJ directly ligates the broken DNA ends together without relying on a homologous template, or, indeed, any sort of quality control check [Pannunzio et al. 2018]. In the context of precision gene editing, NHEJ has been considered a largely error-prone, haphazard process that results in random short deletions and/or insertions (indels). Although NHEJ can correctly repair a DSB, it allows for repeated re-cutting of a target site until an error large enough to render the site un-targetable by the nuclease occurs, leading to the “error-prone” nature of the pathway [Betermier et al. 2014]. Recent studies, however, suggest that not only is NHEJ not random [van Overbeek et al. 2016, Allen & Crepaldi et al. 2018, Shen & Arbab et al. 2018], but the created indels are generally reproducible and, in fact, somewhat predictable [Allen & Crepaldi et al. 2018, Shen & Arbab et al. 2018]. Even though these repair outcomes are predictable, a variety of indels may be generated at a particular locus [Allen & Crepaldi et al. 2018].
2018], and in-frame indels are not conducive to generating knockouts. Also, in model organisms such as zebrafish, founders are frequently mosaic, requiring a cumbersome outcrossing process to generate stable, non-mosaic knockout lines [Ata et al. 2018]. Finally, due to strict import/export controls, international distribution of such stable lines is expensive and time-consuming, leading many researchers to attempt to recreate already established lines within their own laboratories, adding to the amount of time research groups spend creating models rather than studying a gene of interest [Ekker et al. 2017 (personal communication)]. Therefore, it is highly desirable to create lines in a manner that can be quickly, efficiently, and inexpensively reproduced with high fidelity.

In contrast to both the inefficient but precise HR pathway and the efficient but imprecise NHEJ pathway, the microhomology-mediated end joining (MMEJ) repair pathway seems almost tailor-made for creating reproducible gene knockouts. In the MMEJ mechanism, a DSB is repaired in four basic stages: 1) limited (<200 bp) 5’ to 3’ end-resection of the broken DNA strands, resulting in 3’ single-stranded DNA overhangs, 2) base pairing of newly-exposed short regions (~2-25 nts) of sequence homology (microhomology arms) within the single-stranded overhangs, 3) excision of overhanging 3’ sequence proximal to the DSB site, and 4) ligation and repair of the DNA duplex [Seol et al. 2018]. By definition, MMEJ repair (in contrast to NHEJ, which can create insertions) always induces a short deletion corresponding to the intervening sequences between microhomology arms, as well as one of the arms itself (this is the excision step described in (3) above). Additionally, MMEJ-based repairs appear to produce less mosaicism than NHEJ-based repairs, and are highly reproducible [Allen & Crepaldi et al. 2018]. Thus, if MMEJ could be reliably induced, a gene knockout experiment could be designed to induce a DSB at a desired target site,
resulting in an out-of-frame deletion with enhanced probability of a successful gene knockout.

**Predicting Preferred Gene Knockout Sites**

Bae et al. [Bae et al. 2014] first devised a method for choosing preferred DSB sites likely to produce out-of-frame insertions and deletions by developing a scoring system to determine whether a site is located near microhomology arms capable of producing frameshift mutations upon repair. The Bae “Microhomology Finder” method works by enumerating all possible microhomology arms (≥2 bp in length) within a specified distance of a DSB site, and then determining the deletion pattern that would be produced by each pair of microhomology arms (i.e., the sequence of the repaired DSB if those microhomology arms were utilized). For each potential deletion pattern, a *pattern score* is calculated according to the following formula:

\[ e^{-\left(\frac{\text{deletion length}}{20}\right)} \times (\text{microhomology index}) \]

where the *deletion length* is the number of nucleotides that would be deleted in the course of a repair event using a particular set of microhomology arms, divided by an empirically determined weighting factor of 20, and the *microhomology index* is the length of the microhomology arm added to the number of G and C residues it contains (e.g., the microhomology “CAGCCT” would have a microhomology index of 10, because it is 6 nucleotides long and contains 1 G and 3 Cs).

In the Bae approach, the *pattern scores* for all potential deletions associated with a particular target site are summed into the *microhomology score*, and the final *out-of-frame score* is determined by calculating the fraction of the microhomology score that corresponds to out-of-frame deletion patterns. A target site is recommended for gene knockouts if the out-
of-frame score is >66% (i.e., 2/3 of the total microhomology score comes from out-of-frame deletion patterns).

The Microhomology Finder tool tends to produce high-efficiency knockouts [Bae et al. 2014, Ekker et al. 2017 (personal communication)], and the out-of-frame score for a particular DSB site tends to correlate very well with the observed proportion of out-of-frame indels.

**Motivation for This Study**

The Bae results suggested that it may be possible (and is certainly desirable) to predict the actual repair outcome for a particular site, instead of just whether or not a site is likely to be a “good” site for generating gene knockouts. Thus, the ultimate goal of the current study was to create a machine learning model to predict the sequence outcomes of DSB repair, i.e., the probability of obtaining each of the potential MMEJ deletion patterns for a particular DSB site (this idea was conceived before it was demonstrated that NHEJ-based repair is non-random). Due to limitations imposed by the size of the dataset available at the time we initiated this study (see Materials and Methods), we decided to first tackle a smaller problem: Can we predict the *extent* to which a particular DSB site will be repaired using the MMEJ repair pathway – that is, the proportion of deletions observed that correspond to an MMEJ-deletion pattern – and thus assist users in picking targets likely to successfully create gene knockouts?

Towards that end, we developed a machine learning model, *MEDJED* (Microhomology-Evoked Deletion Judication Elucidation), which predicts the proportion of deletions for a DSB site that match any of that site’s possible MMEJ-based repairs. MEDJED is freely available online at www.genesculpt.org/medjed/. MEDJED’s source code is
available for download at GitHub (https://github.com/Dobbs-Lab/MEDJED/), and the tool can also be used as a Docker image (https://hub.docker.com/r/cmmann/medjed/).

**Materials and Methods**

**Dataset**

Bae et al. kindly provided a dataset of 92 DNA sequences targeted by CRISPR/Cas9 nucleases for DSB [Bae et al. 2014]. All DSB targets were generated in HeLa cells. The data included the gRNA used to target the DNA sequence, and sequencing read data for each target. The read data had already undergone quality control and was de-duplicated, to generate a list of the unique amplicon sequences for each target along with a read count for the amplicon.

**Dataset processing**

I wrote an R script called *Method for Assorting Amplicons Tidily*, or MAAT, to sort the recovered amplicons for each target site and characterize features including the number of wildtype reads, single base insertions and deletions, deletions matching an MMEJ deletion pattern, deletions not matching an MMEJ deletion pattern, and total deletions (see Supp. Code 1).

Because I was not provided with a list of wildtype sequences until much later, I designed MAAT to also attempt to identify the canonical wildtype sequence and any SNP variants. It does this by identifying the longest sequence(s) with the largest read count, and then identifying single base pair differences between the ‘canonical’ sequence and other amplicons of identical length. If wildtype sequences are known, MAAT can use these sequences instead of attempting to identify the wildtype sequence from the read data.

During this process, three genes (RP3, CCR5, FLT3) were discarded due to low read counts and/or problems with the proto-spacer motif.
Unfortunately, MAAT does not (yet) identify cases in which an imperfect microhomology arm match was used. It is possible for longer microhomology arms to tolerate 1-2 bp mismatches between the arms [Allen] – e.g., ‘GCCACCG’ might pair with ‘CGGGGGC’, even though the center nucleotide is mismatched. Although this does not appear to occur frequently and requires long microhomology arms when it does occur [Allen & Arbab et al. 2018], it is possible that MAAT missed some of these instances in the Bae et al. dataset.

**Training and Validation Datasets**

We randomly divided the data into a training set of 75% of the data (66 gene targets) and a validation set of the remaining 23 targets. The IDs of the genes in the training and testing sets is provided in **Supp. Table 1**.

**Dataset Effects**

To determine the effect of the dataset on the model, we also randomly sampled 75% of the data, trained a classifier on these data, and then tested performance on the remaining 25%. We did this 5000 times and examined the variance in performance metrics (see **Fig. 4**).

**Features**

Based on the study performed by Bae et al. [Bae et al. 2014] and various features our group determined might be predictive, we generated an initial input vector of 29 features. These features included several “aggregate” features – i.e., features that aggregated information from all potential MMEJ repair patterns for a particular target. These aggregate features included the calculated minimum, maximum, mean, median, and standard deviation of: 1) the number of nucleotides deleted (deletion length), 2) the length of the
microhomology arm (microhomology arm length), 3) the GC content (number of guanines and cytosines divided by the total number of nucleotides) of the microhomology arm (microhomology GC content), 4) the number of nucleotides between the 3’ end of the 5’ microhomology arm and the DSB site (distance to cut site), and 5) the Bae et al. pattern score, for a total of 25 features. We also assessed an additional four features: 1) the GC content of the wildtype DNA sequence surrounding the DSB site, 2) the number of deletion patterns producing an out-of-frame deletion divided by the total number of deletion patterns, 3) the Bae microhomology score, and 4) the Bae out-of-frame score.

Model development

We trained several classifiers using various classification algorithms including Naïve Bayes, logistic regression, generalized linear model, and random forest (not shown). Of these, the random forest classifier worked best. We used the R randomForest package to create the model and perform feature selection.

Feature selection

Feature importance was assessed using the R randomForest package. Due to the small size of the dataset relative to the dimension of features assessed, we performed feature selection to improve the model. The randomForest package assesses two measures of importance for regression models: 1) Mean increase in squared error (%IncMSE), which assesses the effect of randomly permuting a feature’s value on the error of the out-of-bag instances, and 2) the decrease in node impurity (IncNodePurity, assessed using the residual sum of squares) caused by splitting the data using a particular feature (averaged over all the trees in the random forest), which assesses how “close” using a particular feature gets the regression model to the observed value [Liaw & Wiener 2002]. If randomly permuting a
feature produces a large increase in MSE, or if it produces a large decrease in node impurity, then the feature is important to the model’s accuracy.

Feature selection was performed by assessing feature importance, discarding features with negative %IncMSE (that is, randomly permuting these features improved model performance) and IncNodePurity of 0. A new model was created from the remaining features, and this process repeated iteratively until no features had negative %IncMSE and IncNodePurity of 0.

Results and Discussion

The best-performing model, hereafter referred to as “MEDJED”, uses 6 features, a forest of 5000 trees, random feature pool of 2, and imposes a maximum node constraint of nine nodes per tree. On the training set, it achieved a Pearson correlation coefficient (PCC) of 81.36%, mean absolute error (MAE) of 10.96%, and root mean squared error (RMSE) of 13.09%. On the test set (which does not overlap with the training set), MEDJED achieved PCC of 85.20%, MAE of 10.26%, and RMSE of 12.02% (see Fig. 4-1). The “zeroR” prediction method, which “learns” the mean value of the training data and predicts that value for the test data, achieved an MAE of 18.98%, RMSE of 20.99%, and can’t produce a PCC due to the lack of a standard deviation.
Figure 4-1 MEDJED performance. On the independent test set, MEDJED achieves a Pearson Correlation Coefficient (PCC) of 81.36%, Mean Absolute Error (MAE) of 10.96%, and Root Mean Square Error (RMSE) of 13.09%. The MEDJED-predicted MMEJ repair proportion (x-axis) is graphed against the observed MMEJ repair proportion (y-axis).

**Feature Importance**

After feature selection, six features were included in MEDJED: the maximum pattern score, the standard deviation of the pattern score, the mean microhomology length, the standard deviation of the microhomology length, the minimum deletion length, and the maximum microhomology length (see Fig 4-2). The maximum pattern score is likely most important because it represents the “strength” of the strongest microhomology within range of a DSB site. Ata et al. [Ata et al. 2018] found that if: i) the deletion pattern with the highest score at a DSB site has a pattern score ≥ 1.5 times larger than the pattern score of the next highest-scoring deletion, and ii) the induced deletion is ≤ 5 bp, the DSB site tends to produce a single, predominant MMEJ repair outcome called a PreMA (predominant MMEJ allele), which makes up ≥ 50% of the total repair outcomes. It is possible that MEDJED has learned
signal from these PreMA sites, especially given that the minimum deletion length was also among the important features. Interestingly, the minimum deletion length has no correlation on its own with the observed MMEJ outcomes (PCC 0.83%), although it does negatively correlate with the maximum pattern score (PCC 29.47%) as expected, because the maximum pattern score for a DSB site is frequently generated by deletion patterns producing short deletions.

Figure 4-2 Importance measures of the six features in the MEDJED model. The percent increase in mean squared error is to the left (%IncMSE) while the increase in node purity (IncNodePurity) is to the right. The features are (top to bottom) the maximum pattern score, standard deviation of the pattern score, mean microhomology arm length, standard deviation of the microhomology arm length, minimum deletion length, and maximum microhomology arm length for the collection of MMEJ-based deletion patterns at the DSB site.
Interestingly, the standard deviation of the pattern score is also an important feature. Ata et al. formulated a “competition hypothesis” regarding Bae pattern scores and outcomes observed in gene editing experiments - i.e., if there are several microhomologies of similar “strength”, represented by similar pattern scores, then MMEJ appears to have trouble “deciding” which microhomology to use, and an alternative pathway might then repair the DSB instead [Ata et al. 2018]. If this is indeed the case, if the standard deviation of the pattern scores at a particular site is smaller (representing similar pattern scores), then we would expect to see a decrease in the MMEJ at that site.

I plotted the standard deviation of the pattern score for each target site against the observed proportion of MMEJ-based deletions, and found a PCC of 56.47% (see Fig. 4-3), which shows that increasing standard deviation of the pattern scores at a particular deletion site correlates moderately well with its observed MMEJ deletion outcomes, thus providing support for the Ata et al. competition hypothesis.

The correlation between the standard deviation of the microhomology arm lengths and repair outcomes (PCC 41.53%) lends further support to the competition hypothesis: a longer microhomology arm leads to a larger pattern score, and so if there is large variation in microhomology arm lengths, this implies large variation in microhomology arm strengths, and thus less competition for a single predominant MMEJ outcome.

The remaining features are the mean and maximum microhomology arm lengths, which correlate moderately well with the observed MMEJ repair outcomes (PCC 40.00% and 33.04%, respectively.) The importance of the mean and maximum microhomology arm lengths is likely due to MMEJ favoring long microhomologies for base pairing [Allen & Crepaldi et al. 2018, Shen & Arbab et al. 2018].
Figure 4-3 Scatterplot of the standard deviation of the Bae et al. [Bae et al. 2014] pattern score plotted against the observed proportion of MMEJ deletions for 89 HeLa cell targets. Pearson correlation coefficient of 56.47%.

**Dataset Effects**

To determine if the model’s performance was a result of a fortunate split between examples in the training and testing set, I randomly divided the HeLa dataset into training and test sets 5000 times and then trained and tested random forest classifiers with MEDJED’s parameters (see Fig. 4-4). The mean absolute error (MAE) for these 5000 trees ranged from 8.73% to 19.80%, with median 13.97% and mean of 14.02% compared to the final MEDJED model’s MAE of 10.26%. The root mean square error (RMSE) of the 5000 trees ranged from 10.57% to 22.43%, with median 16.54% and mean 16.56%, compared to MEDJED’s RMSE
of 12.02%. The Pearson Correlation Coefficients of the 5000 trees ranged from just 11.20% to 91.03%, compared to MEDJED’s PCC of 85.20%.

This demonstrates that while the final MEDJED model may benefit from some effects of the random split of the training and test data, the model is not a statistical outlier.

Figure 4-4 Boxplot of the mean absolute error (MAE), Pearson Correlation Coefficient (PCC), and root mean squared error (RMSE) on test sets. The original dataset was randomly split into training (75%) and testing sets (25%) 5000 times, generating 5000 classifiers with the same parameters as the final MEDJED model; each classifier was then assessed on its respective test dataset. The median MAE was 13.97% with standard deviation of 1.6%. The median PCC was 63.06%, with standard deviation 12.39%. The median RMSE was 16.54% with standard deviation of 1.7%.
Conclusion

MEDJED is a machine learning regression model for predicting the extent to which a DSB site will undergo microhomology-mediated end joining. MEDJED is capable of predicting the extent of MMEJ-repair utilization in sequence deletion outcomes, and is available online at www.genesculpt.org/medjed/, and can be downloaded at https://github.com/Dobbs-Lab/medjed/. It is also available as a Docker image at https://hub.docker.com/r/cmmann/medjed.

In November 2018 (several months after the development of MEDJED), a group from the Wellcome Sanger Institute [Allen & Crepaldi et al. 2018] and a group including scientists from the Broad Institute, the Massachusetts Institute of Technology, and Harvard Medical School [Shen & Arbab et al. 2018] both published methods, called FORECasT and inDelphi, respectively, for predicting the sequence of repair outcomes at DSB sites. The FORECasT method was demonstrated to achieve near the theoretical limit of prediction accuracy (i.e., the level of reproducibility between gene editing experiments) [Allen & Crepaldi et al. 2018]. FORECasT was trained using ~40,000 DSB sites constructed for the study [Allen and Crepaldi et al. 2018]; inDelphi was trained with ~4000 DSBs similarly generated for use in the study [Shen & Arbab et al. 2018]. The existence of these tools makes the further development of MEDJED to predict the sequence of repair outcomes unnecessary.

What is of far more interest, however, is the potential MEDJED offers for understanding the biological mechanisms of DNA DSB repair. FORECasT requires enumerating all possible deletions and short insertions (≤2 bp, within 3bp of the cut site), then calculating over 2,000 features for each possible sequence outcome [Allen & Crepaldi et al. 2018]. inDelphi must also enumerate potential sequence outcomes, then uses microhomology lengths and GC content (when applicable), and deletion lengths as input to
three separate neural networks [Shen & Arbab et al. 2018]. In contrast, MEDJED requires just 6 features calculated from only potential microhomologies (rather than enumerating all possible outcomes) in order to make predictions, and as a random forest regression model, it is possible to analyze and interpret these features to generate hypotheses (or provide support for a hypothesis, as with the competition hypothesis.)

**Acknowledgements**

We would like to acknowledge and thank the Dill-PICL Lab group, especially Darwin Campbell, for their assistance in bug-testing and hosting early versions of MEDJED, as well as valuable discussions. We would also like to thank the Essner, McGrail, Ekker, and Clark labs for valuable discussions and bug testing.
Supplementary Materials

Supp. Code 1

See https://github.com/Dobbs-Lab/MAAT

Supp. Table 4-1 Training and Test set gene IDs

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL1</td>
<td>BMX</td>
</tr>
<tr>
<td>ACK</td>
<td>CCK4</td>
</tr>
<tr>
<td>ALK</td>
<td>EGFR</td>
</tr>
<tr>
<td>ARG</td>
<td>EphA1</td>
</tr>
<tr>
<td>AXL</td>
<td>EphA6</td>
</tr>
<tr>
<td>BLK</td>
<td>EphB1</td>
</tr>
<tr>
<td>BRK</td>
<td>EphB2</td>
</tr>
<tr>
<td>CSK</td>
<td>EphB4</td>
</tr>
<tr>
<td>CTK</td>
<td>FES</td>
</tr>
<tr>
<td>DDR1</td>
<td>HER4/ErbB4</td>
</tr>
<tr>
<td>DDR2</td>
<td>JAK2</td>
</tr>
<tr>
<td>EphA10</td>
<td>KDR</td>
</tr>
<tr>
<td>EphA2</td>
<td>KIT</td>
</tr>
<tr>
<td>EphA3</td>
<td>LCK</td>
</tr>
<tr>
<td>EphA4</td>
<td>LMR1</td>
</tr>
<tr>
<td>EphA5</td>
<td>MUSK</td>
</tr>
<tr>
<td>EphA7</td>
<td>MER</td>
</tr>
<tr>
<td>EphA8</td>
<td>PDGFRa</td>
</tr>
<tr>
<td>EphB3</td>
<td>ZAP70</td>
</tr>
<tr>
<td>EphB6</td>
<td>TEC</td>
</tr>
<tr>
<td>FAK</td>
<td>TIE2</td>
</tr>
<tr>
<td>FER</td>
<td>TNK1</td>
</tr>
<tr>
<td>FGFR1</td>
<td>TYRO3</td>
</tr>
</tbody>
</table>
References

http://dx.doi.org/10.1038/nbt.4317

http://dx.doi.org/10.1371/journal.pgen.1007652

http://dx.doi.org/10.1038/nmeth.3015

http://dx.doi.org/10.1371/journal.pgen.1004086

http://dx.doi.org/10.1038/nbt.1767


http://dx.doi.org/10.1016/j.jmb.2015.10.014


http://dx.doi.org/10.1038/nm.3793

http://dx.doi.org/10.1523/eneuro.0068-15.2015

http://dx.doi.org/10.3389/fped.2018.00183

[http://dx.doi.org/10.1016/j.jacbts.2017.01.004](http://dx.doi.org/10.1016/j.jacbts.2017.01.004)


[http://dx.doi.org/10.1126/science.1232033](http://dx.doi.org/10.1126/science.1232033)

[http://dx.doi.org/10.1074/jbc.TM117.000374](http://dx.doi.org/10.1074/jbc.TM117.000374)

[http://dx.doi.org/10.1089/154065802761001293](http://dx.doi.org/10.1089/154065802761001293)


[http://dx.doi.org/10.1038/s41586-018-0686-x](http://dx.doi.org/10.1038/s41586-018-0686-x)

[http://dx.doi.org/10.1093/hmg/ddw286](http://dx.doi.org/10.1093/hmg/ddw286)


http://dx.doi.org/10.1016/j.molcel.2016.06.037

http://dx.doi.org/10.1073/pnas.111009698

http://dx.doi.org/10.1093/ilar/ilw021
CHAPTER 5. THE GENE SCULPT SUITE: A SET OF TOOLS FOR GENOME EDITING

Carla M. Mann$^{1,2}$†*, Gabriel Martínez-Gálvez$^3$†, Jordan M. Welker$^2$, Wesley A. Wierson$^2$, Hirotaka Ata$^3$, Maira P. Almeida$^2$, Karl J. Clark$^4$, Jeffrey J. Essner$^{2,*}$, Maura McGrail$^2$, Stephen C. Ekker$^{3,4,*}$, Drena Dobbs$^{1,2}$

1 Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA, 50011, USA
2 Genetics, Development and Cell Biology Department, Iowa State University, Ames, IA, 50011, USA
3 Department of Physiology and Biomedical Engineering, The Mayo Clinic, Rochester, MN, 55905, USA
4 Department of Biochemistry and Molecular Biology, The Mayo Clinic, Rochester, MN, 55905

* To whom correspondence should be addressed.
†Co-first authors

DOI: 10.1093/nar/gk405

Abstract

The discovery and development of DNA-editing nucleases (Zinc Finger Nucleases, TALENs, CRISPR/Cas systems) has given scientists the ability to precisely engineer or edit genomes as never before. Several different platforms, protocols, and vectors for precision genome editing are now available, leading to the development of supporting web-based software. Here we present the Gene Sculpt Suite, which comprises three tools: 1) GTagHD, which automatically designs and generates oligonucleotides for use with the GeneWeld knock-in protocol; 2) MEDJED, a machine learning method, which predicts the extent to
which a double-stranded DNA break site will utilize the microhomology-mediated repair pathway; and 3) MENTHU, a tool for identifying genomic locations likely to give rise to a single predominant microhomology-mediated end joining allele (PreMA) repair outcome. All tools in the Gene Sculpt Suite are freely available for download under the GPL v3.0 license and can be run locally on Windows, Mac, and Linux systems capable of running R and/or Docker. The Gene Sculpt Suite is also freely available online at www.genesculpt.org.

Introduction

Recent additions to the gene editing toolbox include methods for identification of off-target sites (1, 2), strategies for improving nuclease specificity (3), and the expansion of nuclease targeting capabilities (4-7). Other approaches have focused on DNA double-strand break (DSB) repair by increasing the efficiency of homology directed repair (HDR)/homologous recombination (HR) or enhancing the precision of the non-homologous end joining (NHEJ) DNA repair pathway (8) (see Fig. 1A, Fig. 1B). However, relatively little work has been done to leverage homology-mediated end joining (HMEJ) pathways (Fig. 1C), including microhomology-mediated end joining (MMEJ) and single-strand annealing (SSA), and their potential to enhance the efficiency, precision, and reproducibility of gene-editing experiments.

Gene knock-in research has focused on increasing the frequency of HDR/HR-based DSB repair to precisely integrate DNA cargo into a genomic locus, e.g., by modifying the Cas9 protein (9) or inhibiting NHEJ (10). However, these methods can be difficult to implement and can be highly inefficient, with only a few successful knock-ins per hundreds of attempts. In addition, HR is almost completely inhibited during the G1 phase of the cell cycle (11), which inhibits targeted integration in post-mitotic cells and decreases gene-editing knock-in efficiencies in embryos. Much of the recent research on enhancing gene knockouts
Figure 5-1 DNA double-strand break (DSB) repair mechanisms. (A) Non-homologous end joining (NHEJ). The DNA DSB ends are bound by the Ku70-Ku80 heterodimer and undergo limited end-resection before DNA polymerases and ligases repair the break. This process may perfectly repair the DSB break, but more frequently introduces short indels (red). (B) Homology-directed repair (HDR). When a DSB is detected, homologous sequences (blue and orange segments), frequently provided by a sister chromatid are used as a template to repair the break (green). The resulting repair is usually precise. (C) Homology-mediated end joining (HMEJ). HMEJ is a catch-all term for repair that utilizes short regions of homology, including MMEJ and SSA. In both MMEJ and SSA, 5’-3’ end-resection exposes single-stranded DNA regions, where homologous sections (blue) anneal with one another for repair. The overhanging DNA strands (red) are then clipped, resulting in a short deletion. MMEJ and SSA are mechanistically similar but distinct pathways, utilizing different protein machinery. MMEJ also utilizes shorter regions of microhomology (~2-25bp) than SSA (>25bp). SSA end-resection can be extensive, so the pathway operates over larger nucleotide distances. Has focused on NHEJ. This pathway has been thought to repair DNA DSBs in an apparently random and inherently error-prone fashion through the introduction of short indels. Recent work has demonstrated that these errors are not necessarily random and are frequently reproducible (12 - 14). Although there are now methods for predicting repair profiles (12,13), DSB sites that rely heavily on NHEJ - as opposed to MMEJ - often lead to highly mosaic DSB repair profiles, i.e., they do not display a single predominant repair outcome (12).
In contrast, the Gene Sculpt Suite (GSS) tools (GTagHD (15), MEDJED, and MENTHU (16)) leverage HMEJ, a catch-all term for repair methods such as MMEJ and SSA, which utilize short regions of sequence homology to repair DSBs. GTagHD aids researchers in implementing the GeneWeld protocol, which leverages HMEJ repair to introduce targeted knock-ins with efficiencies much higher than previously reported (15). MMEJ repairs frequently have highly predictable outcomes based on the ‘strength’ of the microhomology regions present (17). The relative strengths of these homologies can be used to identify predominant MMEJ allele (PreMA) reagents, i.e., nuclease that target sites likely to result in a single MMEJ-based deletion composing >50% of all repair outcomes (16). MENTHU and MEDJED are GSS tools designed to assist researchers in identifying PreMA reagents (16) and assessing the MMEJ potential of potential target sites, respectively.

**Results**

**Availability and Implementation**

The GSS server is hosted on an Amazon Web Services Elastic Compute Cloud Ubuntu 16.04 LTS instance. Each tool was built in R (https://www.r-project.org/) using RStudio (https://www.rstudio.com/) and is an RShiny (https://shiny.rstudio.com/) application contained in a Docker (https://www.docker.com) image using the Open Analytics r-base image (https://hub.docker.com/r/openanalytics/r-base). When a user visits a GSS tool URL, ShinyProxy (https://www.shinyproxy.io) spins up a new container from that tool’s Docker image; the user can then securely interact within the confines of their container until they close their browser page (Fig. 2). ShinyProxy releases and deletes the container one minute after the browser connection has closed. This allows users to securely interact with the server in their own virtual environments.
Figure 5-2 Gene Sculp Suite (GSS) Architecture. The GSS server uses ShinyProxy (https://www.shinyproxy.io/), to administer the Docker images (solid blue line) for each GSS tool. When a user (blue circle) visits a GSS tool URL, ShinyProxy creates a Docker container (dashed blue line), which essentially is a temporary copy of the Docker image, and allows a user to securely interact within their own container. These containers are temporary, and deleted once a user leaves their URL. A new container is spun up for each unique user.

Each tool in the Suite is also available for download via GitHub (https://github.com/Dobbs-Lab) and as a Docker image through Docker Hub (https://hub.docker.com/u/cmmann). These tools can be run locally on Windows, Linux, and Mac operating systems capable of running R v3.5.2 or later and/or Docker v18.06.1-ce or later. All tools are available at www.genesculpt.org, which also includes links to the GitHub and Docker Hub repositories.

**GTagHD**

GTagHD (pGTag Homology Designer) designs oligonucleotides for use with the GeneWeld protocol ((15); see Fig. 3). GeneWeld uses short sections of sequence homology
between a plasmid donor and a genomic locus to efficiently and precisely integrate the plasmid cargo into the specified locus, with minimal disruption to surrounding DNA. For additional details regarding the GeneWeld technology and its advantages over previous integration methods see Wierson et al. (15).

Figure 5-3 GeneWeld integration scheme (15). Short homologous sequences from the integration site in the target genome (in blue and orange) are cloned into the flanking regions of the donor plasmid cargo (green). When the cargo is freed from the plasmid, the homologous regions promote the efficient and precise integration of the cargo into the genomic locus using homology-mediated end joining.

**Input**

GTagHD takes the genomic integration site with surrounding DNA sequence and a user-specified length of sequence homology between the plasmid donor and integration site as input. Users input the genomic locus as a pasted DNA sequence or GenBank, RefSeq, or Ensembl ID. The gRNA sequence used to target the integration site is input as the 20-nucleotide guide (with no PAM sequence). GTagHD assumes a Cas9-like DSB will be generated 3 bp upstream of the PAM sequence, allowing flexibility in the choice of CRISPR
nuclease in targeting the genomic locus. We have developed two plasmid series for use with the GeneWeld protocol, and although we strongly recommend using these plasmids with GTagHD, the tool also supports custom plasmids and cargos, which require the gRNA sequence for freeing the cargo from the custom plasmid as the only additional input.

**Processing**

GTagHD identifies the integration site using the provided genomic gRNA sequence. GTagHD checks to ensure that this gRNA appears exactly once within the provided genomic DNA, but does not check for off-target sites within the rest of the genome; several tools (including CRISPRscan (18)) are available for this purpose. GTagHD extracts the user-specified length of homologous sequence surrounding the integration site, automatically adds additional nucleotides to repair frameshifts caused by the DSB, adds restriction enzyme sites for cloning into the plasmid, accounts for custom plasmid gRNAs (if provided), and performs additional plasmid-series dependent processing.

**Output**

GTagHD outputs four oligonucleotide sequences: 5' “forward”, 5' “reverse”, 3' “forward”, and 3' “reverse”. The oligonucleotides sequences can be downloaded as a text file and are ready-to-order. The synthetic oligonucleotides can be easily cloned into a plasmid vector. If a user chooses to use a plasmid from the GeneWeld series, they can also download automatically-generated plasmid maps containing their incorporated oligonucleotides in A Plasmid Editor (ApE) format, which is compatible with the GenBank format (gb).

**Comparison to other methods**

The *GeneWeld* protocol was inspired by the PITCh protocol (19 - 20), which is also available for designing knock-in construct guides (http://www.mls.sci.hiroshima-
u.ac.jp/smg/PITChdesigner/index.html). However, GTagHD has a few features that may make it more convenient for users than the PITCh designer 2.0 web tool (21).

First, users can submit GenBank, RefSeq, and Ensembl IDs to specify their genomic locus, instead of copying and pasting whole sequences as in PITCh 2.0. When using an ID, GTagHD can automatically identify and repair frameshifts created by the DSB site to maintain the correct codon and keep the original sequence in frame and intact. PITCh 2.0 requires users to manually specify the reading frame and corrects frameshifts by inserting “Cs” or by deleting a codon entirely, thus altering the original genomic sequence.

Second, GTagHD identifies the DSB integration site in the genomic sequence from user-provided gRNA, and does not require users to manually scroll through the sequence to identify the location, as in PITCh 2.0.

Finally, GTagHD does not require any information about the plasmid vector beyond (possibly) the gRNA sequence used to free the cargo, whereas PITCh 2.0 requires sequence context from the insert.

**MEDJED**

MEDJED (Microhomology Evoked Deletion Judication EluciDation) is a random forest machine learning-based method for predicting the extent to which a DSB site will undergo MMEJ repair. MEDJED was trained on 66 and tested on 23 CRISPR Cas9 sites in HeLa cells acquired from Bae et al. (17). As shown in Fig. 4, when comparing the predicted proportion of MMEJ-based deletions against the observed proportion of MMEJ-based deletions on an independent test set, MEDJED achieved a correlation coefficient of 85.2%, mean absolute error (MAE) of 10.3%, and root mean square error (RMSE) of 12.0%.
Figure 5-4 MEDJED performance. On a test set of 23 HeLa cell targets from (17), MEDJED achieves a Pearson Correlation Coefficient (PCC) of 85.2%, Mean Absolute Error (MAE) of 10.3%, and Root Mean Square Error (RMSE) of 12.0%. The MEDJED-predicted MMEJ repair proportion (x-axis) is graphed against the observed MMEJ repair proportion (y-axis).

**Input**

MEDJED takes a pasted DNA sequence between 20 and 200 nucleotides in length as input and assumes the DSB occurs in the exact middle of the sequence.

**Processing**

MEDJED assess the strengths of all microhomologies present, utilizing features including the minimum deleted sequence length, the maximum, mean, and standard deviation of the microhomology arm lengths, and the maximum and standard deviation of the Microhomology-Predictor pattern score (17). These features are input into the MEDJED regression model.
Output

MEDJED returns a prediction of the proportion of deletion repair outcomes at the provided site expected to result from MMEJ-based repair. It also outputs the values of the six features used in predicting the MMEJ-based repair proportion, as well as a table of all the MMEJ-based deletion outcomes for the targeted site. These outputs can be downloaded individually or collectively as a zip file.

Comparison to other methods

The Microhomology-Predictor (http://www.rgenome.net/mich-calculator/, (17)), on which MEDJED is partially based, calculates an “out-of-frame” score for choosing DSB sites likely to generate out-of-frame deletions; if the score is above 66, the site is recommended for generating gene knockouts. Microhomology-Predictor does not, however, predict the extent of MMEJ at a particular site, and while the out-of-frame score tends to correlate closely with the observed proportion of out-of-frame repairs, it is not a probability of such events occurring.

inDelphi (https://indelphi.giffordlab.mit.edu/, (13)) and FORECasT (Favoured Outcomes of Repair Events at Cas9 Targets, https://partslab.sanger.ac.uk/FORECasT, (12)) both predict expected “repair profiles” at a DSB site – that is, they enumerate all possible repair outcomes for a particular site (within a limited sequence window), and compute the probability of each outcome. inDelphi is notably feature-rich and offers the option to predict probabilities in different cell types; however, determining the probability of MMEJ-based repair for a particular site requires additional calculations on the part of the user. FORECasT, while simple to use, does not output an intuitive human-readable result, requiring users to perform remapping of each outcome to calculate the predicted proportion of MMEJ repair.
MENTHU

MENTHU (Microhomology-mediated End joining kNockout Target Heuristic Utility) identifies sites likely to have a predominant microhomology-mediated end joining allele (PreMA) repair outcome (16). MENTHU expands on the Microhomology-Predictor tool algorithm (17), which produces a “pattern score” for each possible MMEJ-based deletion within a sequence. This score is based on the length, GC content, and deleted sequence length expected to be produced by the microhomology, with a higher score corresponding to a “stronger” microhomology. MENTHU evaluates the ratio between the two highest scoring deletions as a surrogate for relative competitiveness between microhomology sites in recruiting the MMEJ machinery, in order to identify “low competition” sites where a single microhomology pairing is likely to be predominant. For additional details, see Ata et al. (16).

Input

MENTHU takes a user-specified CRISPR or TALEN nuclease and a target DNA region as input. Users can choose from a list of CRISPR nucleases or can specify custom nucleases by providing a Protospacer Adjacent Motif (PAM) sequence, distance between DSB site and PAM, and length of 5’ overhangs (for nucleases producing sticky-end DSBs, like Cas12a). The genomic DNA target can be specified by pasting a DNA sequence or a GenBank, RefSeq, or Ensembl ID. MENTHU also allows users to specify exons to increase search speed and biological relevance of the results.

Processing

MENTHU scans the input DNA for selected nuclease target sites. For each matching site, MENTHU identifies all microhomology pairings within an 80 bp window centered at the DSB site and scores them according to the algorithm employed by Microhomology-
Predictor (17). MENTHU then identifies sites in which the highest scoring predicted deletion has ≤ 5 intervening nucleotides between the microhomology arms in the wildtype sequence and calculates the quotient between its pattern score and the next highest scoring microhomology. This ratio is the MENTHU score.

**Output**

MENTHU outputs a table of likely PreMA reagents in descending order of MENTHU score (Fig. 5). The table consists of ten columns. The “Target_Sequence” provides the gRNA or TALEN sequence needed to induce a DSB at a particular site. The “MENTHU_Score” column contains the computed MENTHU score. The “Frame_Shift” column indicates whether the PreMA deletion generates a frameshift. The “Tool_Type” provides the PAM sequence, in the case of CRISPR nucleases, and the length of the arms and spacer in the case of TALEN inputs. The “Strand” column indicates whether the Target_Sequence matches the forward or complement strand. The “Exon_ID” gives the exon in which the Target_Sequence site occurs, while the “DSB_Location” gives the position of the nucleotide directly to the left of the DSB site. The “Microhomology” column gives the sequence of the microhomology producing the deletion. The “PreMA_Sequence” column shows the top predicted MMEJ deletion sequence (PreMA) for the site. The “Context” column (not shown) gives the “wildtype” sequence corresponding to the PreMA region. The table is searchable, sortable, and can be downloaded in CSV format. Targets can be filtered to show only recommended sites (with MENTHU score ≥ 1.5). By default, all sites for which the top MMEJ deletion has <5 bp between microhomology arms in wildtype sequence are shown, although the results can be filtered to show only recommended sites (MENTHU score ≥ 1.5). Targets can also be filtered to display only T7-compatible gRNAs.
Figure 5-5 Example MENTHU output table. Each row corresponds to a single DSB event. The “Target_Sequence” column contains the gRNA or TALEN sequence required to generate the DSB. The “MENTHU_Score” column gives the ratio between the Microhomology-Predictor pattern scores of the top two scoring microhomologies at the site; a DSB site is likely to produce a PreMA if the MENTHU Score is $>1.5$ (16). “The Frame_Shift” column indicates whether the most frequent expected deletion pattern induces a frameshift. The “Tool_Type” gives the PAM sequence for CRISPR nucleases, and the left arm length/spacer/right arm length combination for TALENs. The “Strand” column indicates whether the “Target_Sequence” occurs on the forward or complement strand. The “Exon_ID” provides the number of the exon in which the DSB site occurs; if no exon information is available, this value is 1. The “DSB_Location” provides the index of the nucleotide to the left of the DSB site within the entire nucleotide sequence. The “Microhomology” column contains the sequence of the microhomology arms used to generate the deletion. The “PreMA_Sequence” gives the sequence of the predicted predominant repair outcome. The “Context” column (not shown) gives the sequence window used for MENTHU score calculations.
Comparison to other methods

The Microhomology-Predictor tool (17), FORECasT (12), and inDelphi (13) all assist users in choosing sites for gene knockout. However, MENTHU has several key features that may make it more convenient for some users. MENTHU utilizes the Pattern Score devised by Bae et al. and used in the Microhomology-Predictor tool (17). As previously described, the Microhomology–Predictor uses the Pattern Score to identify sites likely to produce a frameshift (and by extension, gene knockout). In contrast, MENTHU uses the ratio between Pattern Scores for various MMEJ-based deletion patterns to approximate “competition” between available microhomologies for use by the MMEJ repair machinery (16). This “competition score” is then used to reduce mosaicism in repair outcomes. Microhomology-Predictor does not offer any insights into the level of mosaicism in repair outcomes. In addition, users can scan for only Cas9 NGG sites, whereas MENTHU has been validated using TALENs and offers the ability to search for a wide variety of PAMs.

MENTHU provides several conveniences over FORECasT. The web interface for FORECasT does not allow for automatic analysis of multiple DSB sites along a sequence. It also only supports NGG PAMs; if a non-NGG PAM is of interest, it must be manually specified by its numeric location in the sequence. In contrast, MENTHU scans an input sequence for any targets matching one or more user-specified PAMs or TALENs automatically. In addition, while the FORECasT web interface outputs the predicted repair outcome probabilities for the single specified target site, the downloadable output of the tool consists of a machine-readable file containing a code specifying the deletion, rather than the actual sequence. Thus, while the ability of FORECasT to predict the sequence outcomes for a given DSB is useful, the current web tool is of limited utility for users who wish to locate those sites.
In contrast, inDelphi’s web interface is very feature-rich and accepts any Cas9-like PAM. The “single” mode allows users to manually scan for PAM sites in five different cell lines and then outputs the likely mutation probability profile for each. inDelphi outputs additional information including the predicted frameshift probabilities, the predicted distribution of 1 bp insertions and of deletions up to 60 bp in length, the “precision” (the expected proportion of the most prevalent mutation outcome for a given DSB), a “microhomology strength” score, and the frameshift frequency, in addition to detailed information regarding the predicted outcomes.

inDelphi can also be run in batch mode, allowing users to access features in “single” mode for every potential DSB site along an input sequence. Additionally, users can ask inDelphi to recommend gRNAs likely to produce a specified genotypic outcome, which MENTHU does not currently perform. However, this mode is limited to Cas9-like outcomes and pasted input DNA sequences only. inDelphi’s “gene” mode offers the “batch” mode treatment for precomputed human (hg38) and mouse (mm10) genes for SpCas9 only. In contrast, MENTHU has been validated in zebrafish models, and can perform expanded scanning within a gene or genomic region of interest based on accession ID, allowing for greater flexibility in target site scanning.

Unlike FORECasT and inDelphi, MENTHU has been validated for TALEN platforms and supports scanning for PreMA TALEN sites. Additionally, while none of these tools (including MENTHU) have been validated for enzymes that generate staggered-DSBs, such as Cas12a/Cpf1, MENTHU can provide predictions for these sites based on our current understanding of MMEJ repair machinery (Fig 6).
138

Figure 5-6 Strategy for handling staggered-cutting nucleases. End-resection operates in a 5'-3' fashion. 5' overhangs produced by a staggered-cutting nuclease will be removed during the resection phase. The eliminated sequence in the overhangs is thus unavailable for utilization in MMEJ. We can approximate the microhomologies available for use in MMEJ repair by creating a pseudostring DNA sequence made up of the 5' strand up until the DSB site (orange) concatenated to the 3' strand (blue). The 5' overhangs (dashed lines) are effectively removed. This allows staggered DSBs to be treated identically to blunt DSBs, after the 5' overhangs are removed from the sequence. The “Context” column within the MENTHU results table (see Fig. 5) contains this pseudostring when a staggered-cutting nuclease is chosen.

Ultimately, the intended functionality of MENTHU is different from that of inDelphi and FORECasT, which are designed to predict full mutational profiles resulting from specific DSBs. In contrast, MENTHU aims to identify target sites that are likely to result in a particular outcome. Genome engineers will find a more detailed description of editing outcomes in inDelphi and FORECasT, but more accessible targeting recommendations in MENTHU for a wider variety of nucleases and input DNA sequences.

**Discussion**

The tools in the Gene Sculp Suite are designed to empower researchers to deploy MMEJ-based gene editing, which allows them to focus their efforts on the editing repair outcomes for functional genomics and gene therapy applications. They also enable users to
accurately design HMEJ-based targeted gene integration vectors by helping them design oligonucleotides to implement the highly efficient GeneWeld strategy for creating knock-in mutations, which has been reported to yield ~50% germline transmission rates (16).

All tools in the GSS are under active development. Additional GeneWeld plasmid series are nearing completion (J Welker and J Essner, personal communication), and we will add tools for these to GTagHD as they are developed. Work to further improve MENTHU performance in targeting intronic sequences and to validate MENTHU performance for editing with Cas12a systems is underway. We are also using MENTHU to investigate the frequency and occurrence of PreMA alleles (15) in various genomes and producing genome browser tracks to display precomputed PreMA sites for the entire human genome.

**Data Availability**

The Gene Sculpt Suite is freely available online through www.genesculpt.org.

project.org/web/packages/rhandsontable), Biostrings
(https://rstudio.github.io/DT), jsonlite (https://rdrr.io/cran/jsonlite, (26)), httr (https://cran.r-
project.org/web/packages/httr), and Bioconductor (https://bioconductor.org, (27)). All of
these packages are freely available, and code to quickly install them is included in GSS
installation instructions on GitHub.

Plasmid maps for GeneWeld plasmids are available through GTagHD’s webpage.
GeneWeld plasmids are available at AddGene: https://www.addgene.org/Jeffrey_Essner/.

Acknowledgement

We would like to thank Carolyn Lawrence-Dill, Darwin Campbell, Scott Zarecor,
Kokulapalan Wimalanathan, Mingze He, and Ian Braun of the Dill Plant Informatics and
Computation Lab (Dill-PICL) for their assistance in bug- and stress-testing the GSS
webserver. We would especially like to thank Darwin Campbell for hosting previous
versions of this server and assisting in the transition to Amazon Web Services.

Funding

This work was supported by the National Institutes of Health [R24 OD020166].
Funding for open access charge: National Institutes of Health.

Conflict of Interest

Iowa State University and The Mayo Clinic have filed for patent protection for the
GeneWeld targeted knock-in technology. Wesley Wierson, Jeffrey Essner, Maura McGrail,
Karl Clark and Stephen Ekker have financial interests in LIFEngine Technologies Inc., a
licensee of the GeneWeld technology.
Author Contributions

Mann CM – concept, writing, editing, programming (GTagHD, MEDJED, MENTHU), algorithm development (MEDJED), system administration; Martínez-Gálvez G – writing, editing, figures, programming (MEDJED, MENTHU), algorithm development (MEJDED); Welker JM – editing, algorithm development (GTagHD), bug testing; Wierson, WA – algorithm development (GTagHD), bug testing; Ata H – algorithm development (MENTHU); Almeida MP – algorithm development (GTagHD); Clark KJ supervision; Essner JE – supervision, editing; McGrail M – supervision, editing; Ekker SC – supervision, editing; Dobbs, D – supervision, editing.

References


CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Summary of Contributions

In the course of my dissertation research, I created two novel software tools, RPIDisorder and MEDJED, and implemented two additional web tools, GTagHD [Wierson & Welker et al. 2019] and MENTHU [Ata et al. 2018], designed to facilitate new methodologies for gene editing. All of these tools are freely available online (www.rpidisorder.org and www.geneculpt.org), in addition to being available for download through GitHub and DockerHub (in the case of the Gene Sculpt Suite tools GTagHD, MEDJED, and MENTHU.) Each tool is thoroughly documented and posted with detailed instructions for installation, use, and trouble-shooting. The Gene Sculpt Suite is already heavily used by a global community (see Appendix D), even though only one of the tools in the suite had been published before this dissertation was completed.

RPIDisorder and MEDJED represent the first steps in two distinct avenues of fascinating research. The remainder of this chapter will discuss future directions for these tools, and future development plans for the Gene Sculpt Suite.

6.2 Future Directions: RPIDisorder

RPIDisorder introduces several new directions for future research. In the short term, I plan to compare the performance of RPIDisorder to that of other available methods for predicting RNA-protein interaction partners using several published datasets. I also intend to investigate the efficacy of RPIDisorder in recapitulating human RNA-protein interaction networks using a new “true negative” dataset of my own creation. I will also investigate the effects of using different reduced alphabets and conjoint k-mer representations proposed by [Suresh et al. 2015] and [Jain et al. 2018] on prediction performance. In addition, there are
several potential approaches to improving the impact of disorder-based features, such as testing different disorder prediction methods and identifying new ways to encode disorder information.

I am currently generating a “true negative” dataset of RNA-protein pairs that do not interact, derived from high-throughput CLIP-seq experiments deposited in ENCODE [The Encode Consortium 2012] for several hundred proteins. Pre-processing was performed by POSTAR2 [Zhu et al. 2019]. I am generating true negative or “non-interacting” pairs by randomly pairing RNAs and proteins not found among the true positive or “interacting” pairs (from the same cell line). The logic here is that if protein A and RNA 1 are found to interact in HeLa cells, and protein B and RNA 2 also interact in HeLa cells, but a pairing of protein A with RNA 2 is not identified, then protein A and RNA 2 represent a “true” negative interaction, because protein A and RNA 2 should have had the opportunity to interact, but did not interact sufficiently to be cross-linked and immunoprecipitated. It is important to note that a failure to detect an interaction does not guarantee that protein A and RNA 2 do not interact. It is possible that a protein and RNA do not interact under the physiological conditions of a particular experiment, but can interact under other conditions. For example, the presence of disordered protein-RNA complexes in nuclear stress bodies and stress granules [Spector 2006] implies that some RPIs may occur only under stress conditions. Despite this caveat, this approach should be superior to randomly pairing proteins and RNAs drawn from a database with no consideration of their cellular or temporal expression patterns. Once complete, this new “true negative” dataset of RNA-protein pairs will be valuable due to the sheer number of interactions it will contain. Currently, the dataset includes ~200 proteins and thousands of transcripts mapping to every human chromosome across ten cell lines.
I intend to continue developing RPIDisorder and disorder-based RNA-protein interaction prediction after leaving Iowa State. My long-term goal is to teach Bioinformatics at a small liberal arts college. Further exploration of the influence of disorder on RNA-protein recognition would serve as the foundation for a highly relevant and engaging undergraduate research program.

6.3 Future Directions: MEDJED

While the MEDJED chapter (see Chapter 4) in its current form is likely not of sufficient novelty for publication, it represents a step forward in identifying the sequence determinants of DSB repair pathways. The features identified as important for MEDJED predictions lend further support to the competition hypothesis for MMEJ-mediated repairs proposed by Ata et al. [Ata et al. 2018]. The fact that DSB repair outcomes, including non-homologous end joining deletions and (apparently) insertions, are reproducible [Allen & Crepaldi et al. 2018] implies that DSB repair is non-random, and suggests that DNA sequence and/or structural determinants influence which repair pathway is activated. The large datasets made public by Allen et al. and Shen et al. [Allen & Crepaldi et al. 2018, Shen & Arbab et al. 2018] provide a wealth of data that can be mined to identify those determinants.

6.4 Future Directions: The Gene Sculpt Suite

I intend to continue developing the Gene Sculpt Suite (GSS). At the time I created it, it represented the very best of my coding abilities. Since then, I have gained the knowledge and to create a much more stable and less cumbersome tool than the current implementation. I will be working on that project in the weeks before I leave Iowa State.

Several additional plasmid series for CRISPR-based precision gene editing are under development by the McGrail and Essner groups at Iowa State. These plasmids require
different processing methods than the pGTag and pPRISM plasmids currently handled by the GTagHD web tool. Therefore, in addition to stabilizing the GSS webserver, I will develop an updated version of GTagHD, to be called OCYRIS-C (Oligonucleotide deCYner for Rigorous Integration of Specialized Cassettes), which will handle oligonucleotide design for all of these plasmid series. The new name will also bring the oligonucleotide designer tool in line with the Egyptian-deity themed names of Medjed (a mysterious deity which appears in two scrolls and is otherwise unmentioned in the Egyptian pantheon) and Menthu (an Egyptian god of competition and war).

References

http://dx.doi.org/10.1038/nbt.4317

http://dx.doi.org/10.1371/journal.pgen.1007652

http://dx.doi.org/10.1038/nature11247

http://dx.doi.org/10.1038/s41598-018-27814-2

http://dx.doi.org/10.1038/s41586-018-0686-x

http://dx.doi.org/10.1016/j.cell.2006.11.026

http://dx.doi.org/10.1093/nar/gkv020


https://doi.org/10.1101/431627


http://dx.doi.org/10.1093/nar/gky830
APPENDIX A. COMPUTATIONAL PREDICTION OF RNA-PROTEIN INTERACTIONS

Originally published as a chapter in Methods in Molecular Biology, vol 1543 (2017):

Promoter Associated RNA (Humana Press, New York, NY)

DOI: 10.1007/978-1-4939-6716-2_8

Carla M. Mann, Usha K. Muppirala, Drena Dobbs

Abstract

Experimental methods for identifying protein(s) bound by a specific promoter-associated RNA (paRNA) of interest can be expensive, difficult and time-consuming. This chapter describes a general computational framework for identifying potential binding partners in RNA-protein complexes or RNA-protein interaction networks. Protocols for using three web-based tools to predict RNA-protein interaction partners are outlined. Also, Tables listing additional web servers and software tools for predicting RNA-protein interactions, as well as databases that contain valuable information about known RNA-protein complexes and recognition sites for RNA-binding proteins are provided. Although only one of the tools described, lncPro, was designed expressly to identify proteins that bind lncRNAs (including paRNAs), all three approaches can be applied to predict potential binding partners for both coding and non-coding RNAs (ncRNAs).

Key Words:

RNA-protein interactions; computational prediction; RPISeq; catRAPID; lncPRO; RNA-protein databases; machine learning; ncRNA.
Introduction

Our understanding of RNA-protein interactions has advanced dramatically over the past decade due to exciting advances in experimental technologies for identifying binding partners in RNA-protein complexes and RNA-protein interaction networks [1,2]. These include high-throughput CHIP and RNASeq based methods that can identify RNAs bound by specific proteins in vivo (reviewed in [3-5]) proteomics methods that can identify RNA binding proteins, their target RNAs, and their RNA binding sites on a genome-wide scale [6-9] and integrated biochemical and bioinformatics approaches that can identify the specific recognition sequences for RNA binding proteins [10]. A major motivation for these studies has been the search for cellular and molecular functions for non-coding RNAs (ncRNAs), many of which have been shown to play important roles in disease as well as normal development [11]. In particular, promoter-associated RNAs (paRNAs), the focus of this volume, not only regulate transcription [12], but also serve as epigenetic modulators that affect cellular differentiation (e.g., [13], protein localization [14], and gene regulation [15]. Genetic regulation by paRNAs or other ncRNAs is often mediated through interactions of the RNA with specific RNA-binding proteins; thus, identifying the binding partner of a newly discovered paRNA (or any ncRNA) can provide important clues to its function [16].

Despite the technical advances mentioned above, the experimental time, effort, and expense required to identify biologically relevant protein binding partners for a specific RNA (or vice versa) has created a demand for computational methods that can predict the most likely binding partners in RNA-protein complexes and/or identify novel candidate interaction partners in RNA-protein interaction networks. The goal of the chapter is to provide step-by-step protocols to assist molecular biologists and other experts in accessing and utilizing available computational resources that provide access to existing information about specific
RNA-protein interactions, as well as software for predicting potential RNA-protein binding partners when experimental information is not available. For additional background and details regarding these and other published approaches, we refer the reader to recent reviews [17,18]. The methods outlined in this chapter are generally applicable to any RNA, coding or non-coding, small or large; thus, they can be valuable for quickly identifying potential protein binding partners for any specific paRNA.

In this chapter, we focus on currently available web-based computational tools for partner prediction, i.e., predicting which protein binds to a specific RNA of interest in an RNA-protein complex or RNA-protein interaction network. Several available tools are also capable of predicting the converse, i.e., which RNA(s) bind to a specific protein of interest. Software and servers for interface prediction, i.e., predicting which specific amino acid residues and/or ribonucleotides are involved in recognition and binding are not described here, but have been reviewed elsewhere [19-21].

The protocol involves two major steps (illustrated in Figure 1):

Step 1: Determine whether experimental data regarding the binding partner(s) of the query ncRNA or putative RNA-binding protein are already available. This step is described in Methods, Section 3.1, which outlines strategies for exploiting available online databases and servers (provided in Table 1 below) that focus on ncRNA or RNA-protein interactions, or provide sequence and/or structural data regarding RNA-protein complexes.

Step 2: If known or potential interaction partners cannot be identified using available resources, or if the user wishes to identify additional potential binding partners, use
one (or preferably, all three) of the following web-based tools for predicting RNA-protein interaction partners:

- **RPISeq** (Section 3.2) - a machine learning-based approach developed by our group [22], which requires only sequence information for the RNA(s) and protein(s) of interest. This method was not specifically designed for predicting partners of promoter-associated RNAs (or ncRNAs), but it can readily predict these interactions.

- **IncPro** (Section 3.3) - a method developed by the group of Tingting Li [23], specifically for predicting the likelihood that a specific long non-coding RNA (IncRNA) interacts with one or more candidate protein sequences.

- **catRAPID** (Section 3.4) – a suite of programs developed by the group of Gian Gaetano Tartaglia, including algorithms for estimating the binding propensity for individual RNA-protein pairs (catRAPID graphic) [24] and a server for large scale interactome predictions, e.g., for the interaction of a single RNA with an entire proteome or a single protein with an entire transcriptome (catRAPID omics) [25].
Figure A 1 Flowchart for identifying potential RNA-protein interaction partners

The user is strongly encouraged to try all 3 web-based tools because the underlying algorithms and datasets used for training and performance evaluation are different in each case. Direct performance comparisons of the methods on various benchmark datasets indicate that the methods have different strengths and weaknesses [17,18,26]. The user should, of course, interpret all prediction results with caution: although each of these tools has been
shown to perform well “on average” in predicting RNA-protein interactions, a highly accurate prediction for any given RNA-protein pair cannot be guaranteed.

2. Materials

2.1 Databases of experimentally validated RNA-protein complexes and interactions

Before making computational predictions, the user is advised to search for existing experimental evidence regarding the specific RNA-protein interaction of interest, both in published literature and in relevant specialized databases. At present, the availability of information regarding validated RNA-protein interactions is increasing rapidly as new experimental data are incorporated into web-based resources. These include databases containing evidence for physical or genetic interactions obtained from both “low” and “high” throughput experiments (e.g., NPInter or PRD), as well as databases of high resolution structural information regarding both the components and the interfaces in RNA-protein complexes (e.g., PDB or NDB). Table 1 provides an alphabetical listing of several valuable databases that contain information about RNA-protein complexes and interaction networks. A suggested strategy for utilizing selected resources from this list is provided in Section 3.1 below.

2.2. Servers and Software for predicting RNA-protein partners

At present there are only a few published methods for predicting partners in RNA-protein complexes or interaction networks. Sections 3.2 - 3.5 focus on three published methods (RPSeq, catRAPID, IncPro) that are freely available on web-based servers. Table 2 below includes, in addition, several published approaches for which a web server is not yet available, but software is freely available for download.
Table A 1 Databases of RNA-Protein Interactions and Interfaces

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IntAct [40]</td>
<td><a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a></td>
<td>Manually curated molecular interactions, including comprehensive data about their source experiments</td>
</tr>
<tr>
<td>NDB [30,31]</td>
<td><a href="http://ndbserver.rutgers.edu/">http://ndbserver.rutgers.edu/</a></td>
<td>Nucleic acid and DNA/RNA-protein complex structures, including derived data for nucleic acids</td>
</tr>
<tr>
<td>PDB [29]</td>
<td><a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a></td>
<td>Experimentally determined three-dimensional structures</td>
</tr>
<tr>
<td>RBPDB [41]</td>
<td><a href="http://rbpdb.ccbr.utoronto.ca/">http://rbpdb.ccbr.utoronto.ca/</a></td>
<td>Experimental data on binding preferences and specificities of RBPs</td>
</tr>
<tr>
<td>RPIntDB</td>
<td><a href="http://pridb.gdcb.iastate.edu/RPISeq/">http://pridb.gdcb.iastate.edu/RPISeq/</a></td>
<td>RPIs from databases and high-throughput experiments in literature</td>
</tr>
</tbody>
</table>

Note: PRIDB [39], which was developed by our group, is not included in this list of recommended databases because it is no longer maintained.
Table A.2: Computational Methods for Predicting RNA-Protein Interaction Partners

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Dataset</th>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lncPro</td>
<td>322 interacting and non-interacting pairs of RNA-protein chains from 18 RNA-protein complexes from PDB</td>
<td>Physicochemical properties consisting of RNA secondary structure propensities, hydrogen-bonding propensities, and van der Waals interaction propensities</td>
<td>Propensities are calculated for the protein and RNA sequence and fed through a complex equation to generate a probability score ranging from 0-100. (<a href="http://bioinfo.bjmu.edu.cn/lncpro/">http://bioinfo.bjmu.edu.cn/lncpro/</a>)</td>
</tr>
<tr>
<td>catRAPID</td>
<td>7,409 interacting pairs from 858 RNA-protein complexes from PDB</td>
<td>Physicochemical properties including secondary structure propensities, hydrogen-bonding propensities, and van der Waals interaction propensities</td>
<td>Propensities are calculated for each amino acid and ribonucleotide to generate an interaction profile (<a href="http://service.tartaglialab.com/page/catrapid_group">http://service.tartaglialab.com/page/catrapid_group</a>)</td>
</tr>
<tr>
<td>RPISeq</td>
<td>2,241 interacting pairs from 943 RNA-protein complexes from PRIDB</td>
<td>Sequence composition of proteins, represented as conjoint triads, and RNAs, represented as tetrads</td>
<td>Protein and RNA sequences encoded sequence-composition-based features are used to train SVM and RF classifiers (<a href="http://pridb.gdcb.iastate.edu/RPISeq">http://pridb.gdcb.iastate.edu/RPISeq</a>)</td>
</tr>
<tr>
<td>Wang et al.</td>
<td>RPI 2241 generated by Muppirala et al. &amp; 367 interacting pairs from NPInter</td>
<td>Sequence composition of protein and RNA</td>
<td>Input to NB and ENB classifiers is a combination of protein triads and RNA triad features similar to those used in RPISeq</td>
</tr>
</tbody>
</table>
3. Methods

3.1. Search the literature and databases for existing experimental evidence

Before using computational approaches to predict potential interactions, the user should search published literature and existing databases (e.g., http://omictools.com/ [27]) for experimental evidence regarding interactions involving the RNA or protein of interest (see Note 1). If the sequence of interest corresponds to a protein or RNA of unknown function, potential homologs can be identified via a BLAST search. As outlined below, both the original query sequence and its homologs can be used to search databases of known RNA-protein interactions, such as those listed in Table 1.

1. Run the sequence or sequences through NCBI’s BLAST server, available at http://blast.ncbi.nlm.nih.gov/Blast.cgi [28] or use similar genomics resources
elsewhere (see Note 2). BLAST, or Basic Local Alignment Search Tool, finds highly similar sequences in the NCBI or ENSEMBL databases. If the query sequence has been previously identified and/or analyzed, clues to its involvement in specific RNA-protein interactions may be found in the NCBI “Gene” or “Protein” pages corresponding to the sequence (see Note 3). If the query sequence itself is not available in one of the NCBI databases, potential homologs identified by BLAST can be used as a starting point for subsequent searches in the databases listed in Steps 2 – 5 below.

2. Query NPInter, available at http://www.bioinfo.org/NPInter/index.htm {Yuan, 2014 #198}. NPInter v3.0 is the largest curated database of experimentally validated biomolecular interactions involving ncRNAs extracted from the literature. NPInter currently contains more than 900,000 ncRNA interactions, including interactions with proteins, as well as with DNA and both ncRNAs and mRNAs. Note, however, that tRNA and rRNA interactions are excluded (see Note 4). NPInter includes data from 22 different organisms and provides an integrated USCS Genome Browser to assist the user in locating binding sites in the human, mouse and yeast genomes. The database is searchable by molecule name, molecule type, or database ID and provides access to software and servers, including IncPro (described Section 3.3 below) for predicting lncRNA-protein interactions and the iRNA server for predicting RNA-RNA interactions (see Note 5).

3. Query the Protein-RNA Interaction Database (PRD), available at http://pri.hgc.jp/ {Fujimori, 2012 #413}. The PRD is smaller than NPInter, containing 10,817 experimentally validated RNA-protein interactions, but is valuable because it includes
both rRNA and tRNA interactions, which are excluded from NPInter. The database offers flexible keyword searches (see Note 6).

4. In every case, the user should search the **Protein Data Bank (PDB)**, available at www.rcsb.org [29]. The PDB contains over 1,600 three-dimensional structures of RNA-protein complexes determined using experiments such as X-ray crystallography, nuclear magnetic resonance (NMR) imaging, and cryo-electron microscopy. The PDB has a powerful search engine that allows the database to be queried in a variety of ways, e.g., by RNA or protein name, sequence, or GO terms. The PDB also provides excellent structure visualization tools as well as links to valuable third-party resources for visualizing and analyzing the structures of macromolecules (see Note 7).

5. In the same vein, the **Nucleic Acid Database (NDB)**, available at http://ndbserver.rutgers.edu [30,31] is another valuable resource that focuses on experimentally determined three-dimensional structures of nucleic acids, including DNA-protein and RNA-protein complexes. Although the NDB contains only a subset of structures in the PDB, NDB makes it easier for the user to focus on structures that contain RNA-RNA, RNA-protein or RNA-drug interactions. Also, the NDB provides convenient access to a wide variety of tools and software specifically designed for analyzing RNA sequences and structures (see Note 8).

### 3.2. RPISeq – to predict binding partner(s) for any known RNA or protein sequence

The RPISeq web server implements the RPISeq method developed by Muppirala *et al.* [22]. RPISeq uses two types of machine learning classifiers, Random Forest (RF) and Support Vector Machine (SVM), to predict RNA-protein interactions using only sequence...
information. RPISeq can be used to predict the interaction probabilities of any type of RNA (not just ncRNAs) with any protein(s) of known sequence (see Note 9).

1. Access the RPISeq web server available at http://pridb.gdcb.iastate.edu/RPISeq/.
   A stand-alone version of RPISeq is also freely available (see Note 10).

2. For single RNA – single protein predictions: The simplest function of the RPISeq server is to predict whether a specific known RNA interacts with a specific known protein. In this case, the user must enter the protein and RNA sequences (in FASTA format) in the appropriate text boxes on the homepage and click “Submit.”

3. The RPISeq results, which are typically returned a few seconds after submission, include a display of the submitted input sequences along with the interaction probabilities predicted using both the RF and SVM classifiers. A probability greater than 0.50 is usually considered to be a positive prediction, although more stringent thresholds can be chosen.

4. For single RNA – multiple protein predictions: To predict the interaction probabilities for a single RNA with multiple potential protein partners, go to http://pridb.gdcb.iastate.edu/RPISeq/batch-rna.html.

5. Enter the RNA sequence and click “Choose File” to upload a file of protein sequences in FASTA format (limited to 100 Mb) (see Note 10). Click “Submit.”

6. The results are returned as a table listing the interaction probabilities for the input RNA with each protein in the FASTA input file. Probabilities greater than 0.50 are usually considered to be positive predictions. The results may take several minutes to display, depending on the number of protein sequences submitted.
7. For **multiple RNA – single protein predictions**: To predict the interaction probabilities for a single protein with multiple potential RNA partners, go to http://pridb.gdcb.iastate.edu/RPISeq/batch-prot.html.

8. Enter the protein sequence in the text box and click “Choose File” to upload a file of RNA sequences in FASTA format (limited to 100 Mb) (see **Note 10**). Click “Submit.”

9. The results are returned as a table listing the interaction probabilities for the input protein with each RNA sequence in the FASTA input file. Again, probabilities greater than 0.50 are usually considered to be positive predictions. The results may take several minutes to display if a large number of RNAs were submitted.

### 3.3. IncPro - to predict protein binding partner(s) for any known IncRNA

The IncPro web server implements the IncPro method developed by Lu et al. [23]. IncPro is designed to predict whether a specific long non-coding RNA (lncRNA) interacts with one or more user-provided protein sequences. The method uses the hydrogen bonding and van der Waals propensities of the RNA and protein sequences, in addition to the predicted secondary structure of the RNA, to calculate the probability that a specific lncRNA and protein will interact with one another (see **Note 11**).

1. Access the IncPro server homepage at bioinfo.bjmu.edu.cn/Incpro/.
   A stand-alone version of IncPro is also freely available for download (see **Note 12**).

2. For **single RNA – multiple protein predictions**: On the IncPro homepage, click on the “Predict” tab. The IncPro server takes as input a single RNA sequence in FASTA format and a file of multiple protein sequences in FASTA format.
3. Enter the RNA sequence and click “Choose File” to upload a file of protein sequences in FASTA format. Click “Submit.”

4. The results are returned as a table listing the interaction probabilities for the input IncRNA sequence with each protein sequence in the FASTA input file. Probabilities greater than 0.50 are usually considered to be positive predictions.

3.4. catRAPID - to predict either individual or transcriptome/proteome wide interactions

The catRAPID suite of RNA-protein interaction predictors includes methods for predicting the interaction propensity for individual RNA and protein partners (catRAPID graphic) [24]); identifying segments of RNA and protein that most likely participate in forming the RNP interface (catRAPID fragments); estimating the interaction strength of an RNA-protein pair in comparison to a reference set (catRAPID strength); and identifying the most probable interactions between a specific protein and a complete transcriptome, or between a specific RNA and a complete proteome, for eight model organisms [25]). Additional catRAPID modules can predict pairs of co-expressed proteins and RNAs in human tissues (catRAPID express [32]) and whether a protein is likely to bind RNA (catRAPID signature [33]) (see Note 13).

1. The suite website is located at http://service.tartaglialab.com/page/catrapid_group; extensive documentation and tutorials for all tools in the catRAPID suite are provided at: http://service.tartaglialab.com/static_files/shared/tutorial.html.

2. For **single RNA – single protein predictions**: Navigate to the catRAPID group page at http://service.tartaglialab.com/page/catrapid_group and select "catRAPID graphic."
3. Enter the protein and RNA sequences (in FASTA format) in the text boxes provided. Note: The input protein sequence must be between 50 - 750 amino acids in length; the input RNA sequence must be between 50 - 1200 ribonucleotides in length.

4. If desired, enter a name in the optional “Submission label” box and enter your email address in the optional “Email address” box for notification when results are available. Click “Submit.”

5. The results are returned as a graphical “heat map” representing the interaction score for individual amino acid and ribonucleotide pairs; this interaction score is derived from the interaction propensity, which is also reported. The interaction scores range from -3 to +3, with higher values (red) corresponding to a higher probability of interaction. catRAPID graphic also generates a discriminative power (DP) score, which is a confidence metric; a DP score greater than 50%, coupled with a positive interaction score, indicates that an interaction is likely to occur. A DP score greater than 75% is a very high-confidence prediction. The results page also provides the server-generated ID for the job, the date and timestamp for the run, links for downloading the protein and RNA sequences submitted by the user, and the interaction heat map in .png format.

6. For **single RNA - proteome predictions**: Navigate to the catRAPID group page at http://service.tartaglialab.com/page/catrapid_group and select "catRAPID omics."

7. Select the second option: "catRAPID omics [transcript vs. nucleotide-binding proteome]" to open the submission page. The first text box is an optional box for a submission label, which should include the RNA name or other identifier for easy
identification of results. Enter the query RNA sequence (which must be longer than 50 ribonucleotides) in FASTA format.

8. The user is then presented with nine radial buttons under the heading: "Which library would you like to analyze?" Select the proteome of the model organism of interest--ideally, the source organism for the RNA query sequence (or its closest relative)--from the 8 organism libraries available (see Note 14). The user can also create a custom proteome library (see Note 15).

9. The user is then asked whether she/he would like to use nucleic acid binding domains. The default "No" option will query all proteins in the proteome (<750 amino acids long), regardless of whether they possess a recognized RNA-binding domain. The "Yes" option will specifically examine and make predictions between the query RNA and proteins in the selected proteome that possess an RNA-binding domain.

10. Select whether predictions should be made against RNA-binding proteins only or against both RNA- and DNA-binding proteins, and whether disordered proteins should be included in the prediction; the latter is recommended because disordered protein regions frequently bind RNA.

11. A valid email address can be entered into the next text box for notification of when the job is completed. The user should then click "Submit query."

12. catRAPID omics results may take a few hours to be generated, depending on the size of the selected proteome. The results page contains a section summarizing the input parameters and a pie chart showing the distribution of rankings (on a scale of 1-3) of the possible RNA-protein interactions. Red indicates a likely interaction; orange
indicates a moderately likely interaction, and yellow indicates an unlikely interaction.

The next section consists of a 9-column table (see Fig. 2), in which the interactions are listed in order of highest to lowest scores (first column). The "Ranking" displayed in the last column is a metric of the probability of interaction: 3 stars indicates a strong interaction probability; 0 stars indicates an unlikely interaction.

Protein-RNA pairs with both a high star ranking (3) and high discriminative power (>75%) are predicted to interact. High star rankings with low discriminative power correspond to low-confidence predictions; as an example, note that in Fig. 2, the highest-scoring interaction has a very low discriminative power and is therefore considered to be unreliable (see Note 16).

<table>
<thead>
<tr>
<th>#</th>
<th>Protein ID</th>
<th>RNA ID</th>
<th>Z-score</th>
<th>Discriminative Power (%)</th>
<th>Interaction Strength (%)</th>
<th>Domain</th>
<th>Motif</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PCBP3_MOUSE</td>
<td>RupA_1</td>
<td>0.98</td>
<td>10</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>3 stars</td>
</tr>
<tr>
<td>2</td>
<td>SKL_MOUSE</td>
<td>RupA_1</td>
<td>1.93</td>
<td>99</td>
<td>98</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>3</td>
<td>SIN2_MOUSE</td>
<td>RupA_1</td>
<td>1.50</td>
<td>99</td>
<td>99</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>4</td>
<td>HDX_MOUSE</td>
<td>RupA_1</td>
<td>1.46</td>
<td>96</td>
<td>96</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>5</td>
<td>SSRP1_MOUSE</td>
<td>RupA_1</td>
<td>1.36</td>
<td>98</td>
<td>93</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>6</td>
<td>RNPA_MOUSE</td>
<td>RupA_1</td>
<td>1.28</td>
<td>98</td>
<td>94</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>7</td>
<td>RTF1_MOUSE</td>
<td>RupA_1</td>
<td>1.25</td>
<td>98</td>
<td>91</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>8</td>
<td>RFZ2_MOUSE</td>
<td>RupA_1</td>
<td>1.15</td>
<td>98</td>
<td>89</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
<tr>
<td>9</td>
<td>ZMAT1_MOUSE</td>
<td>RupA_1</td>
<td>1.12</td>
<td>97</td>
<td>97</td>
<td>yes</td>
<td>no</td>
<td>2 stars</td>
</tr>
</tbody>
</table>

Figure A 2 Example of catRAPID omics results pages

13. For **single-protein - transcriptome predictions**: Navigate to the catRAPID group page at http://service.tartaglialab.com/page/catrapid_group and select "catRAPID omics."

14. On the subsequent page, select the first option: "catRAPID omics [transcript vs. coding and non-coding transcriptome]," which will open a submission page. The first text box is an optional box for a submission label, which should include the query
protein name for easy identification of results. Enter the query protein sequence
(which must be longer than 50 amino acids) in FASTA format.

15. The user is then presented with nine radial buttons under the heading "Which library
would you like to analyze?" Select the transcriptome of the model organism of
interest, or its closest relative. The user can generate a custom transcriptome library,
if desired (see Note 15).

16. The user is then asked whether she/he would like to use nucleic acid binding
domains. If “No” is selected, the query protein must be <750 amino acids in length. In
this case, the method will utilize the whole protein sequence instead of focusing on
RNA- or DNA-binding domains. The "Yes" option will reveal two new sets of radial
buttons; the first group allows the user to specify whether only RNA-binding domains
should be selected, or whether both RNA- and DNA-binding domains should be
examined. The second group allows the user to include predicted disordered regions
in the query protein as part of the calculation (recommended).

17. The last group of radial buttons allows the user to specify whether she/he wishes to
query partners from the coding (mRNA) or non-coding (tRNA, rRNA, ncRNA, etc.)
transcriptome. A valid email address can be entered for notification of job
completion. The user should then click "Submit query."

18. catRAPID omics results may take a few hours to be generated, depending on the size
of the selected transcriptome. The results page is identical to that returned for the
RNA vs. proteome predictions described in Step 12 above, and results are interpreted
the same way (see Note 15).
4. Notes

1. At present, none of the available computational tools for predicting RNA-protein interaction partners report whether experimental evidence for a specific interaction is available (i.e., even when an interaction partner is known, the software will make a prediction, which may or may not correspond to the experimentally validated interaction partner). Thus, as a first step, the user should always search published literature (via search engines such as NCBI/PubMed (http://www.ncbi.nlm.nih.gov/) or Google Scholar (http://scholar.google.com) and relevant databases (see Section 3.1) for existing experimental data regarding a specific RNA-protein interaction of interest. In addition to the resources described in Section 3.1 and Table 1, many additional valuable databases and servers that provide extensive information regarding in vivo RNA-protein complexes, RNA binding proteins and their recognition sites, RNA-protein complexes and RNA-protein interaction networks are becoming available. OMICtools (http://omictools.com) provides an extensive and up-to-date directory of these resources [27]).

2. According to OmicsTools (http://omictools.com)[27], the ENA (European Nucleotide Archive) Sequence Search tool (http://www.ebi.ac.uk/ena) [34], hosted by the EMBL-EBI, is a “nucleotide search tool which is far faster than BLAST for large datasets, with only a marginal loss in search sensitivity” (http://omictools.com/ena-sequence-search-s2042.html).

3. If the query sequence corresponds to a known protein or RNA, the NCBI “Gene” database (http://www.ncbi.nlm.nih.gov/gene) is an excellent starting point for investigating whether potential binding partners of the query have been previously identified. (Tip: Because proteins and RNAs from humans are usually better
annotated than those from other organisms, valuable information can be obtained by visiting the Gene page for the human homolog of a query sequence.) On the sidebar of the Gene full report page, the Table of Contents may include links to Pathways from BioSystems (for a protein query), and Interactions (for both protein and RNA queries). In addition, the General Gene Information link provides a list of GO annotations, such as “RNA binding,” or under the GO Component heading, a list of specific macromolecular complexes with which the protein or RNA has been associated. Finally, the GenRifs section can provide direct access to the most relevant literature regarding RNAs bound by the query protein.

4. Because NPInter [35] specifically excludes protein-ncRNA interactions that involve tRNA or rRNA, the user interested in such interactions should consult the PRD (see Section 3.1.3), as well as the PDB and NDB (see Section 3.1.4) because these databases contain many such interactions.

5. NPInter [35] provides tools for: i) BLASTing a given protein or RNA sequence against every RNA or protein sequence in NPInter (http://www.bioinfo.org/NPInter/blast/blast_link.cgi); ii) predicting whether or not a specific lncRNA-protein interaction is likely, using IncPro (http://www.bioinfo.org/NPInter/IncPro.htm) (see Section 3.3); and iii) predicting whether two specific RNAs are likely to interact, using RIsearch (http://www.bioinfo.org/NPInter/RIssearch.htm).

6. The PRD [36] contains 10,817 documented physical interactions between RNA and proteins extracted from BioGRID [37], IntAct [37] and the PDB [29], including many interactions that involve tRNA or rRNA.
7. The PDB Advanced Search
(http://www.rcsb.org/pdb/search/advSearch.do?search=new) is a powerful tool that allows the user to BLAST a sequence of interest against all structures in the database, to identify GO annotations, citations in publications, etc. In addition, the PDB offers several built-in visualization tools
(http://www.rcsb.org/pdb/secondary.do?p=v2/secondary/visualize.jsp - RCSBviewer) as well as links to additional resources and software for analyzing macromolecular structures

8. The NDB [30,31] focuses on structures that contain either DNA or RNA and provides links to many valuable RNA sequence and structure analysis tools
(http://ndbserver.rutgers.edu/ndbmodule/services/index.html) as well as software for identifying RNA motifs and for predicting secondary and tertiary structures of RNA molecules (http://ndbserver.rutgers.edu/ndbmodule/services/softwares.html).

9. RPISeq [22] consists of Random Forest (RF) and Support Vector Machine (SVM) machine-learning classifiers that predict the probability of interaction between an RNA and a protein based solely on their primary sequences. In this method, RNA sequences are encoded as normalized frequencies of RNA tetrads, and protein sequences are encoded using a conjoint triad feature (CTF) method originally proposed by Shen et al. for predicting protein-protein interactions [38]. Based on the propensity of the observed conjoint triads to bind the observed RNA tetrads, RPISeq outputs the probability that the submitted RNA and protein will interact. In performance evaluation experiments using 10-fold cross-validation on RPI2241 (a
non-redundant dataset including 2241 RNA-protein pairs derived from PRIDB [39]), the RPISeq SVM classifier achieved an accuracy of 87.1%, and the Random Forest classifier achieved an accuracy of 89.6%. Additional performance metrics and comparisons with other methods are provided in Muppirala et al. [22,17].

10. The RPISeq web server is currently capable of returning predictions for up to 100 sequences (or up to 100 Mb) in a single run. For larger datasets, a stand-alone version of RPISeq is available upon request to the author (instructions available at http://pridb.gdcb.iastate.edu/RPISeq/contact.php).

11. IncPro [23] encodes potentially interacting IncRNA and protein sequences as feature vectors of identical dimensions (based on secondary structure, hydrogen-bonding, and van der Waal’s interaction propensities observed in 41 RNP complexes from the PDB) and uses matrix multiplication to generate an interaction score for each RNA-protein pair. The algorithm was trained and tested on a dataset of 726 non-redundant RNA-protein pairs extracted from 18 complexes in the PDB that contain RNAs longer than 100 nts. In 4-fold cross-validation experiments, the method obtained a Discriminative Power (DP) value of 90.3%. Additional performance metrics and comparisons with other methods (including catRAPID) are provided in [23].

12. A stand-alone version of IncPro is available at:


13. The original catRAPID graphic algorithm [24] generates predictions using interaction profiles of the query protein and RNA sequences, which are based on several physicochemical properties, including predicted secondary structure, hydrogen bonding, and van der Waals interaction propensities. On a non-redundant
dataset of 858 RNA-protein complexes from the PDB, the reported discriminative power was 78%. Additional performance metrics and comparisons with other methods are provided in [18].

14. Currently 8 proteome libraries are available, from: *C. elegans*, zebrafish, fruit fly, human, mouse, brown rat, yeast, and western clawed frog. The user also has the option of submitting a custom sequence library.

15. To generate a custom library, the user should select the ninth radial button in the "library" section. This will cause a text box to appear along with a link to generate a library. Select the red-highlighted "Generate" text to navigate to the library generation tool. (Note that simply clicking on the link will open the library generation tool in the current tab, which may cause data loss. It is highly recommended that the user right-click or command-click the link to open the library generation tool in a new window.)

This takes the user to a library submission page, where she/he can label the library (with a descriptive name, including the source organism) and submit an email address for notification purposes. Select the button to upload a file of FASTA formatted sequences (≤ 500 sequences), click "Submit query" at the bottom of the page, and wait for the library to finish processing. The user will be provided with an ID reference for the library.

16. The first column of the catRAPID omics result table (Fig. 2) lists the numerical rankings assigned to the protein-RNA interactions, from most to least probable. The first row of the table corresponds to the most highly ranked RNA-protein pairing. The second column contains the ID for protein being analyzed (with a clickable link to the protein's ENSEMBL entry). The third column contains the ID for the query RNA
sequence (linked to its sequence). The fourth column contains the normalized interaction propensity (Z-score), with higher values indicating a more likely interaction. The fifth column contains the discriminative power (%) score. The sixth column contains the interaction strength, which is an indicator of the specificity of the reaction; a low value for the interaction strength may indicate that the protein binds the RNA nonspecifically. The seventh column indicates whether the protein possesses a known RNA-binding domain; the eight indicates whether the RNA has any recognized protein-binding motifs. The ninth column is a "star ranking" of the results. The "star rank" of an interaction is a value from 0 to 3, calculated based on three criteria: i) whether the protein has an RNA-binding domain, or both a DNA-binding domain and a disordered region: if both are present, 1 is added to the star rank; if the protein has only a DNA-binding domain or only a disordered region, 0.5 is added to the star rank. If the protein has neither an RNA- or DNA-binding domain and no disordered regions, 0 is added to the rank score; ii) whether the protein has any RNA-binding motifs: if so, 1 is added to the score, and 0 otherwise; (iii) the predicted interaction propensity (which is normalized on a scale of 0 to 1) and added to the scores from (i) and (ii).

Acknowledgments

This work was supported by NIH grant GM066387 and a Presidential Initiative for Interdisciplinary Research (PIIR) award from Iowa State University to DD. We thank Rasna Walia for valuable discussions and suggestions.

References


APPENDIX B. ROBUST ACTIVATION OF MICROHOMOLOGY-MEDIATED END JOINING FOR PRECISION GENE EDITING APPLICATIONS

Hirotaka Ata, Thomas L. Ekstrom, Gabriel Martinez-Galvez,* Carla M. Mann,* Alexey V. Dvornikov, Kyle J. Schaefbauer, Alvin C. Ma, Drena Dobbs, Karl J. Clark, Stephen C. Ekker

*These authors contributed equally


It was published in PLoS Genetics 14(9), September 2018.

The MENTHU web tool was designed to facilitate the methodology described here.

This paper is available at: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007652

Gabriel Martinez-Galvez and I wrote the section of the manuscript describing the MENTHU web tool, and performed the computational validation studies. A detailed description of contributions to the web server’s development is provided in Chapter 1.1.

Abstract

One key problem in precision genome editing is the unpredictable plurality of sequence outcomes at the site of targeted DNA double stranded breaks (DSBs). This is due to the typical activation of the versatile Non-homologous End Joining (NHEJ) pathway. Such unpredictability limits the utility of somatic gene editing for applications including gene therapy and functional genomics. For germline editing work, the accurate reproduction of the identical alleles using NHEJ is a labor intensive process. In this study, we propose
Microhomology-mediated End Joining (MMEJ) as a viable solution for improving somatic sequence homogeneity in vivo, capable of generating a single predictable allele at high rates (56% ~ 86% of the entire mutant allele pool). Using a combined dataset from zebrafish (Danio rerio) in vivo and human HeLa cell in vitro, we identified specific contextual sequence determinants surrounding genomic DSBs for robust MMEJ pathway activation. We then applied our observation to prospectively design MMEJ-inducing sgRNAs against a variety of proof-of-principle genes and demonstrated high levels of mutant allele homogeneity. MMEJ-based DNA repair at these target loci successfully generated F0 mutant zebrafish embryos and larvae that faithfully recapitulated previously reported, recessive, loss-of-function phenotypes. We also tested the generalizability of our approach in cultured human cells. Finally, we provide a novel algorithm, MENTHU (http://genesculpt.org/menthu/), for improved and facile prediction of candidate MMEJ loci.

We believe that this MMEJ-centric approach will have a broader impact on genome engineering and its applications. For example, whereas somatic mosaicism hinders efficient recreation of knockout mutant allele at base pair resolution via the standard NHEJ-based approach, we demonstrate that F0 founders transmitted the identical MMEJ allele of interest at high rates. Most importantly, the ability to directly dictate the reading frame of an endogenous target will have important implications for gene therapy applications in human genetic diseases.

**Author Summary**

New gene editing tools precisely break DNA at pre-defined genomic locations, but cells repair these lesions using diverse pathways that often lead to unpredictable outcomes in the resulting DNA sequences. This sequence diversity in gene editing outcomes represents an important obstacle to the application of this technology for human therapies. Using a
vertebrate animal as a model system, we provide strong evidence that we can overcome this obstacle by selectively directing DNA repair of double-stranded breaks through a lesser-described pathway termed Microhomology-mediated End Joining (MMEJ). Unlike other, better-understood pathways, MMEJ uses recurring short sequence patterns surrounding the site of DNA breakage. This enables the prediction of repair outcomes with improved accuracy. Importantly, we also show that preferential activation of MMEJ is compatible with effective gene editing. Finally, we provide a simple algorithm and software for designing DNA-breaking reagents that have high chance of activating the MMEJ pathway. We believe that the MMEJ-centric approach to be broadly applicable for a variety of gene editing applications both within the laboratory and for human therapies.

**Introduction**

Programmable nucleases such as TALEN (Transcription Activator-like Effector Nuclease) and CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) systems have enabled a new era of scientific research [1, 2]. Instead of relying on knock-down models or expensively outsourced knock out lines, laboratories across the world now have tools with which to generate indels (insertions and deletions) of varying sizes on the gene(s) of interest. However, DNA Double-strand Break (DSB) repairs largely result in diverse sequence outcomes owing to the unpredictable nature of the most commonly used Non-homologous End Joining (NHEJ) pathway [3, 4] (Fig 1). This significantly confounds experimental readouts as knock-out cell lines often harbor more than just one desired frameshift mutation. In the case of model organisms such as zebrafish (*Danio rerio*), the F0 founders are genetically mosaic, warranting a complex and time-consuming series of outcrossing to establish molecularly defined lines before any biological questions can be addressed [5, 6].
Figure 1 MMEJ is a unique DSB repair pathway that results in highly efficient and highly stereotyped mutagenesis.

DSB by conventionally designed Programmable Nucleases typically proceeds through a versatile yet unpredictable classical non-homologous end joining (NHEJ) pathway. As a result, a rather diverse cohort of mutant alleles are generated, making the subsequent selection process labor intensive to enrich for the allele of interest. The resulting genetic composition of the specific loci are often complex, requiring careful molecular characterization of each allele. Efficient activation of microhomology-mediated end joining (MMEJ) pathway, on the other hand, can greatly limit allelic diversity and enable the intentional generation of a particular deletion allele of interest at a high rate. Consequently, the downstream applications become more streamlined with facile generation of homozygous frameshift allele in diploid cells.

https://doi.org/10.1371/journal.pgen.1007652.g001
In contrast to NHEJ, the MMEJ (Microhomology-mediated End Joining) DNA repair pathway utilizes a pair of locally available direct sequence repeats on both sides of a DSB that are apposed, annealed and extended [7–10]. As such, DSB repair outcomes are highly stereotyped (Fig 1), resulting in deletion of the intervening sequence as well as one of the repeats. Consequentially, there is an increasing interest in utilizing MMEJ for precision genome engineering applications [11–14]. To date, however, effective harnessing of this pathway remains challenging due to the paucity of genetic and mechanistic understanding [8].

Bae et al. [14] developed a sequence-based scoring system to estimate the frequency of MMEJ-associated deletions induced by DSBs in human cells. While this improved the predictability of MMEJ activation, the DSB repair outcomes tended to consist of a heterogeneous population of multiple MMEJ alleles. In this study, we sought to improve upon the existing algorithm with the goal of developing tools to more reliably predict target loci that would be predisposed to generate a more homogeneous mutant allele population through MMEJ. We demonstrate the feasibility and utility of such reagent design on the molecular level (i.e., DNA repair outcomes) and on the physiological level (i.e., F0 phenotype). We further demonstrate that our approach can be applied to generating highly homogeneous MMEJ alleles in cultured human cells, suggesting our findings may be broadly translatable to multiple model systems. We believe our approach can inform and benefit applications such as rapid phenotype-genotype correlation in F0 animals, with an eye toward applications in human gene therapy and facilitation of resource sharing & recreation of various cell and animal lines on a global scale.
Results

MMEJ is an active repair pathway in the genetically unaltered zebrafish embryo

Prior works examining MMEJ activation in vertebrate organisms primarily focused on *in vitro* models [8–10, 14–18]. Initial analyses using a targeted knock-in strategy suggested that MMEJ was operational in the zebrafish embryo, though the efficiency of these MMEJ outcomes was rather modest [13]. Importantly, while previous studies reported incidental identification of several zebrafish genomic loci that repaired preferentially through MMEJ when using programmable nucleases [19, 20], no consortium–small or large–of genomic loci that repair primarily through NHEJ vs MMEJ has been compiled. To this end, we examined the repair outcomes of previously designed TALEN and CRISPR-Cas9 genomic reagents (*S1 Table*). The plurality of custom enzymes induced diverse sequence outcomes, consistent with the idea that NHEJ is being used as the primary DNA repair pathway at these loci. However, a few reagents induced sequence outcomes satisfying the following criteria, suggesting that MMEJ was the preferred pathway: 1) most predominant mutant allele is the top predicted allele by the Bae et al. algorithm [14], 2) most predominant mutant allele comprises ≥ 50% of the total mutant allele population, and 3) mutagenic efficiency > 20%. For the purpose of this study, a programmable nuclease satisfying all these criteria is referred to as a Predominant MMEJ Allele (PreMA) reagent. Three sticky-end generating TALEN (*chrd, mitfa #4 & surf1*) and two blunt-end generating CRISPR-Cas9 (*surf1 & tyr #2*) reagents fell into this category (*S1 Table, Fig 2A, Fig 3A*).
PreMA TALEN reagent can be used to recapitulate previously reported loss-of-
*chrd*-function phenotype in 1 dpf F0, injected larvae.

**A.** Top—Wildtype *chrd* sequence with TALEN binding sites annotated in teal. The dotted red boxes are MH arms predicted to be used most frequently. Raw sequence alignment of the whole PCR amplicon demonstrates that the majority of reads are the expected 7 bp deletion allele. Bottom—summary data from subcloning analyses. 50% of the mutant allele recovered were of the predicted MH allele. **B.** Previously reported *chrd* loss-of-function phenotype was successfully recapitulated using this TALEN pair. Phenotype severity was graded by the degree of Intermediate-Cell-Mass expansion in the tail and by the reduced head size by 1 dpf. Box plot demonstrating phenotypic penetrance is provided with each experiment denoted by a unique marker shape. N = 3 biological and technical replicates. At least 29 injected animals were scored in each experiment.

https://doi.org/10.1371/journal.pgen.1007652.g002

Injecting the *chrd* TALEN pair (37.5 pg/arm) resulted in characteristic *chrd* loss of function phenotypes: Intermediate-Cell-Mass expansion and a smaller head by 1 day post-fertilization [21] (1 dpf; Fig 2B). Median penetrance for Moderate and Severe phenotypes was 15.8% and 20.0%, respectively (Fig 2B, S2 Table). Strong MMEJ activation by this TALEN pair was confirmed by subcloning analysis (Fig 2A)– 16/32 recovered mutant reads corresponded to the top predicted 7 bp deletion allele. Similarly, perturbing tyr gene with a CRISPR-Cas9 reagent recapitulated a previously reported, loss of melanin production phenotype, observable by 2 dpf [22] (Fig 3B). Ribonucleoprotein (RNP) delivery at the dose...
Figure 3 PreMA sgRNA against tyr can be used to recapitulate loss-of-melanophore phenotype in 2 dpf, injected F0 larvae.

**A.** Top–Wildtype tyr sequence with the #2 sgRNA target site annotated in green. The dotted red boxes are MH arms predicted to be used most frequently. Raw sequence alignment of the whole PCR amplicon demonstrates that the majority of reads are the expected 4 bp deletion allele. Bottom–summary data from subcloning analyses. 88% of the mutant allele recovered were of the predicted MH allele. **B.** Previously reported tyr-loss-of-function phenotype was successfully recapitulated using this CRISPR-Cas9. Phenotype severity was graded by the loss of retinal pigmentation. Partial loss of retinal pigmentation was considered a Weak phenotype, whereas complete loss of pigmentation in one or both eyes were considered Moderate and Severe phenotypes, respectively. Box plot demonstrating phenotypic penetrance is provided with each experiment denoted by a unique marker shape. N = 3 biological and technical replicates. At least 12 injected animals were scored in each experiment.

https://doi.org/10.1371/journal.pgen.1007652.g003

of 300 pg tyr #2 sgRNA and 660 pg Cas9 resulted in Moderate and Severe loss of pigmentation phenotypes in 22.7% and 50.0% of embryos respectively (Fig 3B, S2 Table). Subcloning analysis showed 21/24 (88%; Fig 3A) of resulting alleles contained a 4 bp deletion consistent with a strong MMEJ activation by this CRISPR-Cas9. Together with the chrd TALEN results, these data support that MMEJ can be an effective repair pathway in F0 embryos at some genomic loci, irrespective of programmable nucleases used.
Many Bae et al. predicted MMEJ loci are preferentially repaired by NHEJ

A subset of these zebrafish reagents described above was prospectively designed using the Bae et al. algorithm (S1 Table). This algorithm calculates the strength of each pair of microhomology arms (i.e., Pattern Score) according to the length and GC content of each pair, as well as the length of the intervening sequence. The additive sum of all the possible Pattern Scores is then returned as Microhomology Score. This latter score was found to have positive correlation with the rate of MMEJ activation in HeLa cells [14]. All fourteen prospectively designed reagents had a Microhomology Score of at least 4000—a median score found on human BRCA1 gene. However, only four of these reagents induced majority MMEJ outcomes as judged by the Microhomology Fraction (S1 Table, S1 Note). We therefore retrospectively analyzed the repair outcomes of these reagents to identify additional factor(s) that would enhance predictability of MMEJ induction.

Rate of Pattern Score change as a discrimination factor for MMEJ induction in vivo and in vitro

Intriguingly, when the pattern score values clustered closely to one another (i.e., a flatter Slope Value as calculated according to S2 Note), this was indicative of an unfavorable target for MMEJ activation in zebrafish embryos. Conversely, loci at which Pattern Scores dropped precipitously (i.e., a steeper Slope Value) were good candidates of MMEJ activation in vivo (p = 0.0048; S1 Fig). Based on these observations, we hypothesized that locally available microhomology pairs are in direct competition with one another such that overabundance of these pairs is a negative predictor of MMEJ activation. In other words, MMEJ activation is more favorable at loci with one or two predominant microhomology pair(s) (Low Competition loci) rather than many strong microhomology pairs (High Competition loci).
To determine whether the zebrafish-based hypothesis was generalizable to human cells (HeLa), we re-analyzed the deep sequencing dataset used to generate the Bae et al. algorithm [14]. Available results from 90 genomic loci were sorted alphabetically by the names of target genes then divided into two groups: first 50 and the remaining 40. The first group was then used for a retrospective, correlative analysis while the latter was used for an analysis compatible with a prospective study design. Outcomes from the first 50 targets showed a correlation similar to that observed in zebrafish; higher Microhomology Fractions generally correlated with low Slope Values (p = 0.00001; S2A Fig). This correlation was lost when microhomology arms of 2 bp were included in the analysis (p = 0.2644; S2B Fig); accordingly, microhomology arms of less than 3 bp were excluded from subsequent analyses. The remaining 40 targets were then binned into High, Medium and Low Competition groups based on quartile distribution of the Slope Value (S2C Fig). In agreement with our Competition Hypothesis, the median Microhomology Fraction was significantly higher in the Low Competition group than in the High Competition group (0.300 vs 0.105, p = 0.011; S2D Fig).

**Competition hypothesis predicts new PreMA reagents**

Based on this Competition Hypothesis, we designed 20 Low Competition sgRNA targets across 9 genes and analyzed the DSB repair outcomes (S3 Table). Slope Values smaller than -40 was used as the cut-off for Low Competition, as 3 out of 4 previously designed zebrafish targets produced majority MMEJ outcomes in this range (S1 Table and S1 Fig). For initial assessments, we used TIDE (Tracking Indels by DEcomposition) analysis—a chromatogram analyzing tool that estimates proportions of length varying mutant alleles present in a pool of mixed alleles [23]–which revealed that 5 of these
sgRNAs against 3 genes (mtg1, tdgf1, ttn.2 #1, ttn.2 #2, and ttn.2 N2B #1) were in the PreMA class. These results were subsequently confirmed by subcloning analyses (S3 Table).

Perturbation of tdgf1 (alternatively known as One-eyed Pinhead) causes aberrant, “pinhead” morphology and cyclopia as judged by reduced forebrain protrusion by 1 dpf [24] (Fig 4B). RNP injections of CRISPR-Cas9 at the dose of 300 pg sgRNA and 660 pg Cas9 resulted in highly homogeneous DSB repair outcomes, generating the top-predicted 4bp allele in 28 of 39 clones analyzed (Fig 4A). Aberrant head morphology alone was classified as Weak whereas that in combination with varying degrees of forebrain protrusion was classified as Moderate or Severe phenotypes. Median penetrance for Moderate and Severe morphology was 21.8% and 11.4% (Fig 4B, S2 Table), consistent with the subcloning results.

![Figure 4A](image)

**Figure 4** Prospectively designed PreMA reagent against *tdgf1* can be used to reproduce gross developmental defect in 1 dpf, injected F0 larvae.
A. Top—Wildtype *tdgf1* sequence with sgRNA target site annotated in orange. The dotted red boxes are MH arms predicted to be used most frequently. Raw sequence alignment of the whole PCR amplicon demonstrates that the majority of reads are the expected 4 bp deletion allele. Bottom—summary data from subcloning analyses. 72% of the mutant allele recovered were of the predicted MH allele. B. Previously reported *tdgf1* loss-of-function phenotype was successfully recapitulated using this CRISPR-Cas9. Phenotype severity was graded by the “pinhead” morphology and cyclopia. Pinhead morphology alone was classified as Weak, whereas Moderate and Severe phenotypes also presented with varying degrees of cyclopia judged by the distance of forebrain protrusion. In the Severe class, the forebrain does not separate the eyes, and they are fused together. Box plot demonstrating phenotypic penetrance is provided with each experiment denoted by a unique marker shape. N = 4 with 3 biological and 4 technical replicates. At least 42 injected animals were scored in each experiment.

https://doi.org/10.1371/journal.pgen.1007652.g004

We next explored whether these PreMA reagents are useful for recapitulating a more subtle phenotype beyond aberrant gross morphologies observed in the *tdgf1* mutants. Splice blockade at the N2B exon of *ttm.2* gene by a synthetic morpholino oligonucleotide was previously reported to reduce the cardiac contractility by ~70% on 2 dpf [25], phenocopying the pickwickm171 mutation [26]. RNP delivery at the dose of 300 pg *ttm.2* N2B #1 sgRNA + 660 pg Cas9 resulted in reduction of the shortening fraction to a comparable degree (Fig 5B). Importantly, RNP delivery of NHEJ-inducing *ttm.2* N2B #2 sgRNA at the same dose only resulted in a more attenuated phenotype, despite it targeting the same exon and having comparable activity (Fig 5; S4 Table). Due to the high editing efficiency, animals injected with these doses of *ttm.2* N2B #1 RNP were not viable in post larval phases. For this reason, animals injected at the lower dose of 75 pg sgRNA + 165 pg Cas9 protein were raised to adulthood. Two F0 founders were successfully outcrossed to wildtype zebrafish. Heterozygous offspring were identified using the dsDNA heteroduplex-cleaving Surveyor assay [27], and the transmission of the top predicted 5 bp deletion allele was confirmed from both founders by subcloning analyses (S3 Fig).
Figure 5 PreMA reagent against *ttn.2* N2B results in specific reduction of shortening fraction in 2 dpf F0 zebrafish.

**A.** *Top*—Wildtype *ttn.2* sequence at the N2B exon with sgRNA target site annotated in red. The dotted red boxes are MH arms predicted to be used most frequently. Raw sequence alignment of the whole PCR amplicon demonstrates that the majority of reads are the expected 5 bp deletion allele. *Bottom*—summary data from subcloning analyses. 86% of the mutant allele recovered were of the predicted MH allele. **B.** Previously reported *pickwick* phenotype was successfully recapitulated using this CRISPR-Cas9. 2 dpf zebrafish were immobilized in 3% methylcellulose for live recording of cardiac functions. Whereas injections with Cas9 only (660 pg), N2B #1 sgRNA only (300 pg), or *tyr* #2 sgRNA RNP (300 pg sgRNA + 660 pg Cas9) did not result in changes in shortening fraction at this age, MMEJ-inducing RNP injection targeting N2B #1 (300 pg sgRNA + 660 pg Cas9) resulted in a specific reduction in shortening fraction by 78.4%. In contrast, NHEJ-inducing RNP injection targeting N2B #2 (300 pg sgRNA + 660 pg Cas9) resulted in attenuated effects on shortening fraction (53.3% reduction), despite similarly high edit efficiency. Each data point represents an individual animal scored with the shape of the marker denoting unique experiment. N ≥ 3 biological and technical replicates, except for N2B #2 where N = 2. At least 5 injected animals were scored in each experiment. P-values calculated by Wilcoxon’s Each Pair Calculation (adjusted for multiple comparisons).

https://doi.org/10.1371/journal.pgen.1007652.g005

We also designed an sgRNA against exon 13 of *ttn.2* (*ttn.2* #2 sgRNA), expected to produce a 12 bp deletion allele as a proof-of-principle for in-frame gene correction (Fig 6A).
bp deletion allele in 72.7% of the clones. While the injected animals presented with mild cardiac edema evident by 2 dpf (median rate: 50.0%; Fig 6B, S2 Table), unlike the N2B #1 sgRNA CRISPR-Cas9 injected animals, these were viable to adult age.

Figure 6 PreMA reagent can be used for in-frame gene alteration.

A. Top—Wildtype tt1.2 sequence with sgRNA target site annotated in red. The dotted red boxes are MH arms predicted to be used most frequently. Raw sequence alignment of the whole PCR amplicon demonstrates that the majority of reads are the expected 12 bp deletion allele. Bottom—summary data from subcloning analyses. 73% of the mutant allele recovered were of the predicted MH allele.

B. 2 dpf zebrafish larvae injected with tt1.2 #2 sgRNA RNP (300 pg sgRNA + 660 pg Cas9) grossly appear normal with the exception of mild cardiac edema. Median penetrance was 50%. N = 3 biological and technical replicates. At least 9 injected animals were scored in each experiment.

https://doi.org/10.1371/journal.pgen.1007652.g006

**Low competition plus proximity of microhomology arms strongly predicts PreMA reagents: V2**

These data implicate the utility of PreMA reagents for various applications that require precision gene editing. However, sgRNA design based on the Competition Hypothesis only yielded 5 PreMA reagents out of 20 that were tested (S3 Table, S3 Note).

While this represented an improvement over the initial approach solely relying on the Microhomology Score (1 out of 14; S1 Table), we sought to further fine-tune the predictability for the PreMA targets. To this end, we pooled the results from all the
programmable nucleases described above (S1 and S3 Tables) and seven Medium ~ High Competition sgRNAs designed as controls based on the Competition Hypothesis (S4 Table). In so doing, we noted that PreMA outcomes were only observed if the two arms of the top predicted microhomology were separated by no more than 5 bp. We subsequently identified the second parameter: high ratio (≥ 1.5) of the Pattern Scores between the top predicted and second predicted MMEJ alleles for a given locus (Fig 7). Seven out of eight reagents that satisfied both of these parameters were PreMA. Of the nine reagents that satisfied the first parameter but not the second, two were PreMA. All the other thirty reagents that failed to meet the first parameter failed to induce the top predicted MMEJ allele strongly. Most importantly, all the failed cases (i.e., incorrect predictions according to the original Competition Hypothesis) can be explained using our revised approach (Competition Hypothesis V2; Fig 7C). The Version 2 also captured three PreMA reagents that would have been missed by the original Competition Hypothesis alone, and one PreMA reagent that would have been missed by the Microhomology Score alone. Similar trends were observed using independently collected, previously published deep sequencing dataset from zebrafish [28] and HeLa cells [14] (S4 Fig).

**Mechanism of MMEJ-activation may be conserved in vertebrates**

To test the generalizability of our findings, we prospectively designed 11 sgRNAs against the human genome (S5 Table) and delivered as RNPs to HEK293T cells. Of the 5 active guides cutting above 20% efficiency, DSBs induced by GJB2 #1 and #2 guides resulted in more homogeneous repair outcomes (Fig 8A and 8B) than any of the 92 guides tested by Bae et al. (S4B Fig) [14]. DSBs at AAVS1 #2 and MYO7A #3, on the other hand, repaired primarily through 1bp indels, consistent with the report by Bae et al. using HeLa cells. Intriguingly, the second most prevalent class of repair at these loci was the top
A. Outlier plot summarizing repair outcomes from 47 genomic targets using TALEN and CRISPR-Cas9. Close proximity of top predicted MH arms (Groups 3 and 4) appears to be the primary determinant for generating PreMA type outcomes as no target from Groups 1 and 2 had Top MH Fraction exceeding 0.5. When the top predicted allele had at least 50% higher Pattern Score than the second predicted allele (Groups 2 and 4), it was a strong indicator for inducing MMEJ-class repairs. B. Top Definition for each of the 4 groups used in Panel A. Each and every zebrafish genomic locus was segmented into these categories. Pattern scores were derived using RGEN online tool. Bottom P-values calculated by Wilcoxon’s Each Pair Calculation (adjusted for multiple comparisons). C. Graphical representation of each group detailed in Panel A. Groups 1 and 2 are prone to activate NHEJ-type outcomes, presumably because the yet-identified MMEJ factor fails to localize to suitable microhomology arm pairs, limited by how far apart these arms are. Group 4 is most suitable for strong MMEJ activation because it satisfies the proximity requirement AND the relative strength requirement. The latter may aid in the kinetics of the yet-unidentified MMEJ factor binding to the microhomology arms. Our data suggest that Group 3 is an intermediate group in terms of MMEJ activation. Perhaps extragenetic factors, such as cell cycle and epigenetic status may determine how favorable the loci are for MMEJ inductions. 

https://doi.org/10.1371/journal.pgen.1007652.g007
predicted MMEJ allele (Fig 8C and 8D), as identified by subcloning analyses. We thus conclude that the specific trigger for efficient MMEJ-activation may be conserved in vertebrate organisms, albeit with nuances that are yet to be elucidated.

Figure 8 Competition hypothesis V2 targets trigger primary repair by MMEJ in HEK293T cells.

A & B. Top—Wildtype human GJB2 sequences with sgRNA target sites annotated. The dotted red boxes denote the top predicted MH arms. Summary TIDE analysis outcomes are also presented showing ~45% Top MH Fractions for GJB2 #1 and #2 sgRNA. Red bar indicates the predicted deletion allele. Calculations for Adjusted Prevalence conform to calculations for Top MH Fractions detailed in S3 Note. Bottom—summary data from subcloning analyses for GJB2 #1 sgRNA (A) and #2 sgRNA (B).

C & D. Top—Wildtype human AAVS1 and MYO7A sequences with sgRNA target sites annotated. The dotted red boxes denote the top predicted MH arms. Bottom—summary data from subcloning analyses for AAVS1 #2 sgRNA (A) and MYO7A #3 sgRNA (B).

https://doi.org/10.1371/journal.pgen.1007652.g008

Accessing the PreMA algorithm through MENTHU (MMEJ kNockout Target Heuristic Utility)

The broad potential utility of this updated PreMA Algorithm for MMEJ prediction led us to develop a web-based automated analysis tool called MENTHU (http://genesculpt.org/menthu/). The tool can also be downloaded and installed on a local
computer ([www.github.com/Dobbs-Lab/menthu](http://www.github.com/Dobbs-Lab/menthu)). MENTHU accepts a user-specified DNA sequence and targeting scheme as input, and outputs recommended CRISPR gRNA target sites that are predicted to result in PreMA type outcomes. We validated the accuracy and functionality of MENTHU against select gRNA sites used in this study using whole exonic sequences as inputs ([S6 Table](#)); importantly, the software identified novel PreMA candidate loci against *surf1* and *tdgf1* where only Group 3 gRNA loci had been found by previous methods. Finally, we conducted a preliminary assessment to examine the prevalence of PreMA loci and found roughly 10% prevalence of such loci among all possible NGG PAM on human *CSF2* as well as zebrafish *tp53* genes ([S7 Table](#)).

**Discussion**

To date, precision genome engineering is limited by the ability to predictably, efficiently, and reproducibly induce the identical sequence alterations in each and every cell. Here, we demonstrate the feasibility and utility of creating allelic consistency by an MMEJ-centric approach for designing programmable nucleases. While the precise cellular components of the molecular machinery involved in MMEJ remain incompletely understood [8], we provide evidence that we can enrich for MMEJ events by strictly sequence-based queries.

We also demonstrate that MMEJ predominant repairs do not operate at the cost of overall mutagenic efficiency; median edit efficiency for PreMA reagents was 91.4% in zebrafish. As genetically unaltered wildtype zebrafish were used throughout the study, we have no reason to believe that NHEJ should have failed at any tested loci. This is in contrast to the proposal that MMEJ is a back-up pathway to NHEJ [7, 8, 16, 17, 29]. Our findings, on the other hand, are compatible with a previous report wherein MMEJ-specific factors such as PolQ are abundantly expressed in embryonic zebrafish [20]. Interestingly, maternally zygotic
PolQ mutant embryos failed to repair DSB at two out of three MMEJ loci, leading to premature deaths [20]. The third locus—which preferentially used a 2 bp microhomology and exhibited more heterogeneous DSB repair outcomes—was able to be repaired at a measurable rate, though significantly less so than in WT embryos. Thus NHEJ and MMEJ may be non-competing, parallel processes with unique triggers.

Based on the data presented here, we speculate that there is a reaction-limiting factor for MMEJ that is involved in identifying compatible microhomology pairs on both sides of the DNA double stranded break. In the case of abundantly available local microhomology pairs, this factor may fail to localize to a single suitable pair, thus rejecting the MMEJ activation. As end-resection is required for MMEJ and not for NHEJ [9, 17, 18], this yet identified factor may be the deciding factor for committing DSB repair through one End Joining pathway to another. This view is similar to a recent report wherein CtIP/Artemis dependent limited end resection was a key trigger for a slow-kinetic Lig1/3 independent NHEJ event that frequently utilized Microhomology to repair a reporter plasmid [30]. In our analysis, the primary driver of this decision making process is the proximity of 2 microhomology arms, further aided by the lack of competing microhomology arms.

Successful deployment of the PreMA reagents makes it possible to directly dictate the reading frame or to do in-frame gene manipulations on endogenous targets. Even assuming a somewhat modest outcome of 50% edit efficiency in which 50% of the mutant allele pool is of the desirable allele, more than 10% of the cell population will be homozygous for this desired allele. Conversely, many real-life gene editing applications would require only one of the diploid copies to be corrected. In these settings under the same assumptions, just 11
viable cells are needed to achieve 95% confidence for establishing the right clone, bringing the idea of precision molecular surgery closer to reality.

Our present study expands upon the current state-of-art understanding for MMEJ activation and demonstrates the ability to prospectively design robustly active PreMA reagents in-vivo. We also provide evidence that this 2-component approach may be broadly applicable beyond zebrafish; testing of the true generalizability of our approach will be facilitated by our web-based application, MENTHU (http://genesculpt.org/menthu/). Importantly, MENTHU allows users to flexibly define a PAM sequence and the cut site (in nts from PAM) so as to accommodate potential future variants of the CRISPR system. Active investigations are underway to accommodate alternative or more lax PAM requirements, such as the case with xCas9—a recently described variant of Cas9 that may function efficiently on an NG PAM [31]. As MMEJ-based loci are inherently restricted to genomic locations that leverage endogenous sequence contexts, availability of more flexible programmable nucleases will become the key for broadening the utility of PreMA reagents.

We provide strong evidence to support the utility of the MMEJ-centric approach beyond phenotype-genotype correlations in F0 animals. We envision this approach to be useful for: 1) studying the effects of homozygous gene knock-out in culture cells (as opposed to more common, compound heterozygous loss-of-function cell lines), 2) rapid small molecule screening in F0 animals as a complimentary approach to studying in germline mutant animals, 3) globally sharing and reproducing gene knock-out cell and animal lines, 4) pathway dissection for MMEJ, and finally, 5) human gene therapy.
Materials and methods

Ethics statement

The animal studies were conducted following guidelines and standard procedures established by the Mayo Clinic Institutional Animal Care and Use Committee (Mayo IACUC). The Mayo IACUC approved all protocols involving live vertebrate animals (A23107, A 21710 and A34513).

Microhomology arms

For the purpose of this study, microhomology is defined as any endogenous direct sequence repeats of ≥ 3 bp surrounding a DSB site. ≤ 2 bp direct sequence repeats were not considered sufficient substrates of MMEJ activation based on our initial analyses of the DSB repair outcomes by previously designed programmable nucleases. Correlation for Microhomology Fraction vs the Slope Value was tangentially stronger when only ≥ 3 bp arms were considered ($r^2 = 0.382$ vs $r^2 = 0.353$; S1 Fig) in zebrafish, whereas the correlation was lost when 2 bp arms were considered in HeLa cells ($r^2 = 0.339$ vs $r^2 = 0.034$; S2 Fig).

Zebrafish husbandry

All zebrafish (*Danio rerio*) were maintained in accordance with protocols approved by the Institutional Animal Care and Use Committee at Mayo Clinic. Zebrafish pairwise breeding was set up one day before microinjections and dividers were removed the following morning. Following microinjections, the fertilized eggs were transferred to Petri dishes with E3 media [5 mM NaCl, 0.17 mM KCl, 0.33 mM CaCl$_2$, and 0.33 mM MgSO$_4$ at pH 7.4] and incubated at 28.5 °C. All subsequent assays were conducted on fish less than 3 dpf, with the exception of assessing for germline transmission. In this case, injected founders were raised to adulthood per the standard zebrafish husbandry protocol.
**DNA oligonucleotide preparation**

All of the oligonucleotides used for this study were purchased from IDT (San Jose, CA). Upon arrival, they were reconstituted into 100μM suspensions in 1x TE and stored at -20 °C until use.

**sgRNA expression vector synthesis**

pT7-gRNA was a gift from Wenbiao Chen (Addgene plasmid # 46759). Given that the minimum requirement for the T7 promoter is a single 5’ G, the GG start on this vector was mutagenized via site-directed mutagenesis (SDM) to accommodate GA, GC, GT starts, using Forward and Reverse primers given (S8 Table). Platinum Pfx DNA Polymerase (Invitrogen 11708013. Carlsbad, CA) was used for 20 cycles of PCR amplification with the Tm of 60 °C and extension time of 3 minutes. DpnI (NEB R0176. Ipswich, MA) was subsequently added to reaction prior to transforming DH5α cells. The target sequence was cloned in as previously described, with the exception of conducting oligo annealing and T4 ligation (NEB M0202. Ipswich, MA) in 2 separate steps. In each case, transformed cells were cultured with Carbenicillin, and plasmids were purified with Plasmid Mini Kit (Qiagen 12123. Hilden, Germany).

**TALEN synthesis**

TALEN constructs were generated using the FusX kit (Addgene # 1000000063) as previously described [32]. In short, RCIscript-GoldyTALEN was linearized with BsmBI (NEB R0580. Ipswich, MA) along with 6 triplet RVD (Repeat-Variable Diresidue) plasmids. Subsequently, they were ligated together in one reaction by a modified Golden-Gate Assembly. Blue-White colony screening with X-Gal/IPTG, colony PCR and finally pDNA sequencing were done to ascertain the correct assembly.
**In-vitro transcription and RNA preparation**

pT3TS-nCas9n (a gift from Wenbiao Chen: Addgene plasmid # 46757) was linearized with XbaI (NEB R0145. Ipswich, MA), whereas TALEN constructs were linearized with SacI-HF (NEB R3156. Ipswich, MA) and sgRNA vector with BamHI-HF (NEB R3136. Ipswich, MA). Tyr sgRNA #2—a construct made in the Essner Lab—was linearized with HindIII (NEB R0104. Ipswich, MA). RNA was made using T3 mMessage mMACHINE kit (Ambion AM1348. Foster City, CA) or HiScribe T7 High Yield RNA synthesis kit (NEB E2040. Ipswich, MA) according to manufacturer’s protocols with the addition of RNA Secure to the reaction (Ambion AM7010. Foster City, CA). To purify RNA, phenol-chloroform extraction was performed using Acid Phenol, Chloroform, and MaXtract High Density Tubes (Qiagen 129046. Hilden, Germany). RNA was then precipitated with Isopropanol at -20 °C, pelleted, air dried and resuspended into nuclease free water. The quality and quantity of RNA were ascertained by using a Nanodrop spectrophotometer and running aliquot on agarose gel. Each batch of RNA was aliquoted into small single use tubes and stored at -80 °C until the morning of microinjections.

**CRISPR-Cas9 RNP preparation for microinjections**

sgRNA was thawed on ice in the morning of microinjections. This was then diluted to the concentration of 300 ng/μL in Duplex Buffer [100 mM KCH₃COO, 30 mM HEPES at pH 7.5]. Appropriate folding of sgRNA was induced by heating it to 95 °C for 5 minutes and gradually cooling the solution to room temperature. Equal volumes of sgRNA and 0.66 mg/mL Alt-R S.p. Cas9 Nuclease 3NLS (IDT 1074181. San Jose, CA) in Cas9 Working Buffer [20 mM HEPES, 100 mM NaCl, 5 mM MgCl₂, 0.1 mM EDTA at pH 6.5] were mixed and incubated at 37 °C for 10 minutes. RNP solutions were subsequently kept on ice until immediately before use.
**TALEN and CRISPR-Cas9 RNA preparation for microinjections**

RNA was thawed on ice in the morning of microinjections. TALEN mRNA was diluted to working concentrations in the range of 12.5 ng/μL to 100 ng/μL in Danieau solution [58 mM NaCl, 0.7 mM KCl, 0.4 mM MgSO₄, 0.6 mM Ca(NO₃)₂, 5.0 mM HEPES at pH 7.6]. sgRNA and nCas9n mRNA were mixed and diluted to the final concentrations of 150 ng/μL and 100 ng/μL, respectively, in Danieau solution. These were all kept on wet ice until immediately before use.

**Microinjections**

Microinjections were carried out as previously described [33]. In short, 1-cell stage fertilized embryos were harvested and aligned on an agarose plate with E3 media. In the case of CRISPR-Cas9 reagents, either 1 or 2 nL was delivered to the cell. In the case of TALEN reagents, 1 ~ 3 nL was delivered to the yolk mass. They were then transferred to Petri dishes in E3 media for incubation at 28.5 °C. Dead and/or nonviable embryos were counted and removed each subsequent morning.

**Phenotype scoring**

Each experiment was conducted in at least a technical triplicate and a biological duplicate. Detailed outcomes are provided in **S4 Table**. Gross phenotypes were scored visually on either 1 dpf or 2 dpf using a standard dissecting microscope. Subsequently, representative pictures were taken with Lightsheet Z.1 (Zeiss 2583000135. Oberkochen, Germany). Shortening Fractions were scored as previously reported [34]. In short, live 2 dpf larvae were immobilized and positioned in 3% methylcellulose. An Amscope camera (MU1403. Irvine, CA) mounted on a Leica Microscope (M165. Wetzlar, Germany) was used to capture a 15 second clip of the beating heart at 66 fps. These clips were subsequently used to measure the distance of the long axis along the ventricle at maximum dilation and
maximum contraction using ImageJ software [35]. Shortening Fraction was calculated as below:

\[
\text{Shortening Fraction} = 100 \times \left(1 - \frac{\text{Distance at Maximum Shortening}}{\text{Distance at Maximum Dilation}}\right)
\]

Shortening Fractions from 5 cycles were averaged for each animal.

**Zebrafish DNA extraction and assessing mutagenic outcomes**

Typically, 8 uninjected wildtype fish and 8 injected fish were randomly collected without prior screening for phenotype. Chorion was predigested with 1 mg/mL Pronase at room temperature as needed. 1 ~ 3 dpf animals were then sacrificed for individual DNA extractions in 100 mM NaOH for 15 minutes at 95 °C. Equal volumes of 8 fish DNA from the same condition were then mixed and used as templates for PCR with either MyTaq (Bioline BIO-21108. London, UK), Phusion (NEB M0530. Ipswich, MA), or KOD (EMD Millipore 71085. Burlington, MA) polymerases per manufacturer’s protocols. The PCR amplicon was resolved on agarose gel, gel extracted with either Monarch DNA Gel Extraction Kit (NEB T1020. Ipswich, MA) or QiaEx II Gel Extraction Kit (Qiagen 20021. Hilden, Germany), and subsequently sent out for sequencing. The chromatograms from both uninjected and injected amplicons were used for TIDE analysis [23]. Alternatively, purified amplicons were used for subcloning analysis with either Topo-TA Cloning Kit (Thermo Fisher Scientific 451641. Waltham, MA) or StrataClone PCR Cloning Kit (Agilent 240205. Santa Clara, CA) per manufacturer’s protocols. Resultant white to pale blue colonies by Blue-White screening were subjected to colony PCR with M13F and R primers, using MyTaq polymerase. Once successful amplification was confirmed on agarose gel, these amplicons were sent out for sequencing either with M13F, M13R or endogenous gene target primers.
Germline transmission for 5 bp deletion generated by N2B sgRNA #1

RNP containing N2B sgRNA #1 was prepared at 4x diluted dose as described above. Following microinjections, viable fish were raised to sexual maturity. Both F0 founders were attempted to out cross successfully mated and produced viable embryos. DNA was extracted from all viable embryos on 1 dpf, and individual DNA was used as template for PCR amplification using MyTaq Polymerase. Once the thermocycling ran to completion, the amplicons were melted by heating to 95 °C and re-annealed by a gradual step-wise cooling. Surveyor assay [27] was conducted per the manufacturer’s protocol (IDT 706025. San Jose, CA), and the results were analyzed by resolving the post-digest amplicons on agarose gel. Amplicons from 4 heterozygous offspring each were subcloned, and 5 colonies each were sent for Sanger Sequencing to confirm successful transmission of the 5 bp deletion allele.

Reanalyses of previously published deep sequencing dataset

For zebrafish dataset, sgRNA screen SRA files were obtained from NCBI’s Short Read Archive (Accession: PRJNA245510) [28]. These files were converted to the fastq format with fastq-dump command using—split-spot function under SRA Toolkit (NCBI. Bethesda, MD). The fastq files were then uploaded onto Cas-Analyzer (http://www.rogenome.net/cas-analyzer/) and analyzed with Comparison range of 25 ~ 40 and Minimum frequency of 1 [36]. Following number of reads were recorded: total, total mutant, total top predicted allele. A top predicted allele was allowed to be included so long as the read contained no more than 2 polymorphisms on the analysis window AND the polymorphisms did not fall on the microhomology arms. Subsequently, the calculated mutagenic efficiency was plotted against the reported efficiency ($r^2 = 0.306$). Of 122 targets designed by Gangnon, et al, following were excluded to arrive to the 34 targets that were used for analysis presented in S4 Fig Panel A: non-NGG targets (36 loci), targets that did not
align to WT consensus sequence (GRCz11; 8 loci), targets with total recovered read counts less than 1% of expected (7 loci), high rate of permutation outside of the target site (1 locus), targets that did not have good agreements between calculated and reported (i.e., fell beyond 99% Confidence Interval; 10 loci), targets that had less than 5% calculated AND reported mutagenic efficiencies (26 loci).

The HeLa cell dataset [14] was obtained from Dr. Kim in the form of excel spreadsheet with aligned sequence outputs +/- 25 bp of the predicted cut site. Following number of reads were recorded: total, total mutant, total top predicted allele with 2 bp microhomology, and total top predicted allele with 3 bp or longer microhomology. As with zebrafish dataset, top predicted allele was allowed to be included so long as the read contained no more than 2 polymorphisms on the analysis window AND the polymorphisms did not fall on the microhomology arms. Of the 92 targets, following were removed to arrive to 74 targets that were used for analyses presented in S2 Fig and S4 Fig Panel B: targets with total recovered read counts less than 1% of expected (2 loci), and targets that had less than 20% mutagenic efficiency (16 loci). There were no targets with non-NGG PAM, no alignment against consensus sequence, nor a high rate of permutation outside of the predicted cut site.

Cell culture and RNP transfection

HEK293T cell line was purchased from ATCC (Manassas, VA) and maintained in DMEM (Invitrogen. Carlsbad, CA) with 10% Fetal Bovine Serum (Sigma. St. Louis, MO). DAPI stain was used to check for mycoplasma contamination.

RNP transfection was conducted as follows in a 48-well format using Lipofectamine CRISPRMAX reagent (Invitrogen CMAX00015. Carlsbad, CA). In vitro transcribed sgRNA was diluted to 2 μM concentration in Duplex Buffer. Secondary structure was induced by heating it to 95 °C for 5 minutes and gradually cooling it to room temperature. 3.0 μL of
sgRNA was then complexed with 3.0 μL of 2 μM Alt-R S.p. Cas9 Nuclease V3 (IDT 1081058. San Jose, CA) in 42.8 μL OPTI-MEM (Life Technologies. Carlsbad, CA) and 1.2 μL Cas9 Plus Reagent. This mixture was incubated for 5 minutes at 25 °C. 2.4 μL of CRISPRMAX reagent and 47.6 μL OPTI-MEM was then added to the RNP, transferred to empty wells, and further incubated for 20 minutes at 25 °C. 200 μL cell suspension at 400,000 cells / mL in complete medium were subsequently added to each well. The dosing of RNP was consistent for all targets except for both GJB2 targets wherein 1 μM each of sgRNA and Cas9 protein was used.

**HEK293T cell DNA extraction and assessing mutagenic outcomes**

HEK293T cells were harvested 24 hour post transfection for gDNA extraction using DNeasy Blood & Tissue Kit (Qiagen 69506. Hilden, Germany). 20 ng of gDNA was used as a template for PCR with KOD polymerase per manufacturer’s protocol. The PCR amplicon was resolved on agarose gel, gel extracted with Monarch DNA Gel Extraction Kit and subsequently sent out for sequencing. The chromatograms from both uninjected and injected amplicons were used for TIDE analysis [23].

For sgRNAs that showed > 20% activity by TIDE, single A overhang was added to the 3’ end of purified amplicons by incubating them with MyTaq polymerase at 72 °C for 15 minutes. They were then used for subcloning analysis with StrataClone PCR Cloning Kit. 96 resultant white to pale blue colonies by Blue-White screening were subjected to colony PCR with endogenous primers using MyTaq polymerase. Once successful amplification was confirmed on agarose gel, these amplicons were subjected to T7E1 assay [27]. Briefly, 2.5 μL each of colony PCR amplicon and wildtype amplicon were heteroduplexed in 1x NEB 2.0 Buffer (25 μL). This was incubated for 15 minutes at 37 °C with 0.5 μL T7 Endonuclease I (NEB m3020. Ipswich, MA) and 4.5 μL dH2O. The digested amplicon was resolved on 2%
agarose gel. Number of colony PCR-positive clones and digest positive clones are reported in S5 Table. Some of the digest positive clones were then sent for sequencing to ascertain the nature of mutation.

2 targets (CSF2 #1 and MYO7A #4) that did not meet the 20% edit efficiency cutoff nonetheless produced statistically significant aberrant sequence peaks by TIDE analysis (p < 0.001). Summary outcomes for Top MH Fraction calculation based on estimated allelic prevalence is given in S5 Table.

MENTHU

We developed a software tool, MENTHU (MMEJ kNockout Target Heuristic Utility), to automate calculations required to implement the 2-component PreMA strategy: 1) identification of top predicted microhomology arms separated by ≤ 5 bp of intervening sequence, and 2) identification of “low competition” target sites (i.e., with a #1-ranked to #2-ranked Pattern Score ratio ≥ 1.5). We designed MENTHU to first compute two of same sequence-based parameters (Pattern Score and Microhomology Score) used in the algorithm of Bae et al., (which are computed online by the RGEN online tool, http://www.rgenome.net).

To do so, we used R [37] to re-implement and modify the original Python source code provided in S3 Fig of the original publication [14]. The MENTHU webserver operates under R version 3.4.1 and RShiny [38] v1.0.5. The MENTHU code was built through RStudio [39] v1.1.442. Details regarding specific R package versions, complete documentation and a full downloadable version of MENTHU for local installation are provided at www.github.com/Dobbs-Lab/menthu/. MENTHU v2.0 can be freely accessed online at http://genesculpt.org/menthu/.

To preliminarily assess the abundance of PreMA loci, MENTHU was locally run to screen the sequences of two genes: human colony stimulating factor 2 (CSF2; Gene ID--
and zebrafish tumor protein p53 (*tp53*; Gene ID– 30590). MENTHU was run twice on each gene: exonic target screen and whole gene target screen. A custom R script was used to mine the MENTHU results in a .csv format to determine both the amounts of total targetable sites by spCas9 (i.e., total number of unique cut sites with NGG PAM on either strand) and the subset of those predicted to be PreMA.

**Statistical analyses**

All of the statistical analyses were carried out using JMP software (SAS Institute, Cary, NC). In all instances, p-values were calculated assuming non-Gaussian Distributions. Wilcoxon Each Pair calculation was used for multiple group comparisons with adjusted p-values.

**Supporting information**

Figure S1 Overabundance of Microhomology arms is a negative predictor of MMEJ activation in zebrafish.

A Box plot showing the distribution of *Slope Values* across 19 zebrafish genomic targets. BScatter plot of *MH Fraction* against *Slope Value*, focused only on microhomology arms of ≥ 3 bp. Linear fit with 95% Confidence Interval (shade) is shown. $r^2 = 0.382$, p = 0.0048. C Scatter plot of *MH Fraction* against *Slope Value* including 2 bp microhomology arms. Linear fit with 95% Confidence Interval (shade) is shown. $r^2 = 0.353$, p = 0.0073. *Pattern Scores* and *Microhomology Scores* were derived using RGEN online tool (http://www.rgenome.net).

https://doi.org/10.1371/journal.pgen.1007652.s001
Figure S2 Overabundance of Microhomology arms is a negative predictor of MMEJ activation in HeLa cell.

**A** Scatter plot of MH Fraction against Slope Value, focused only on microhomology arms of $\geq 3$ bp using the first 50, alphabetically sorted HeLa cell targets. Linear fit with 95% Confidence Interval (shade) is shown. $r^2 = 0.339$, $p = 0.0001$. **B** Scatter plot of MH Fraction against Slope Value including microhomology arms of 2 bp using the first 50, alphabetically sorted HeLa cell targets. Linear fit with 95% Confidence Interval (shade) is shown. $r^2 = 0.034$, $p = 0.2644$. **C** Box plot showing the distribution of Slope Values across the first 50, alphabetically sorted HeLa cell targets. **D** Box plot showing the MH Fractions for High and Low competition sites amongst the remaining 40 HeLa cell targets, focused only on microhomology arms of $\geq 3$ bp. $p = 0.011$. Targets with $< 20\%$ overall edit efficiency were excluded in all panels. Pattern Scores and Microhomology Scores were derived using RGEN online tool (http://www.rgenome.net).

https://doi.org/10.1371/journal.pgen.1007652.s002
Figure S3 Microhomology allele generated by *ttm.2* N2B sgRNA #1 is germline transmitted.

Agarose gel showing PCR amplicon post Surveyor digest. 752 bp band is the whole amplicon. The expected cleavage products due to mutations at the CRISPR site are denoted by yellow arrowheads. The red asterisk denotes positive digest band due to a background T -> A SNP at position 389 from the 5’ end of the amplicon. Heterozygous animals are bolded and underlined. Genotypes of the first 4 heterozygous progenies from each founder were ascertained by subcloning analyses.

https://doi.org/10.1371/journal.pgen.1007652.s003
Figure S4 Fitting Competition Hypothesis Version 2 using independently collected zebrafish dataset and HeLa cell dataset.

A) Outlier plot summarizing independently collected repair outcomes from 34 zebrafish targets. All three Group 4 targets as well as some Group 3 targets yielded PreMA outcomes, validating our own training dataset. Importantly, none of Groups 1 and 2 targets were of this class. B) Outlier plot summarizing repair outcomes from 90 genomic targets using CRISPR-Cas9. Similar to the findings in zebrafish, close proximity of the top predicted MH arms (Groups 3 and 4) appears to be the primary determinant for utilizing this MH pair efficiently. When the top predicted allele had at least 50% higher Pattern Score than the second predicted allele (Groups 2 and 4), median Top MH Fractions trended higher compared to Group 1 and 3, respectively. P-values calculated by Wilcoxon’s Each Pair Calculation (adjusted for multiple comparisons). Targets with < 20% overall edit efficiency were excluded from analysis. Pattern Scores were derived using RGEN online tool (http://www.rgenome.net).

https://doi.org/10.1371/journal.pgen.1007652.s004
Table S1 List and summary mutagenic outcomes of TALEN and CRISPR-Cas9 reagents that were designed primarily using the Bae et al. algorithm [14].

Underlined & italicized bases in sgRNA sequence denote mismatched bases due to the promoter requirement. Pattern Scores and Microhomology Scores were derived using RGEN online tool (http://www.rgenome.net). MH: Microhomology, SC: Subcloning. * Reagents prospectively designed according to Bae et al. algorithm [14]. † No raw sequencing data were available. However, the outcome had been compiled into a table prior to conception of this study. ‡ Injected with sgRNA and Cas9 mRNA (150 pg and 100 pg, respectively). ^ Gift from Wenbiao Chen (addene # 46761).

https://doi.org/10.1371/journal.pgen.1007652.s005
Table S2 Summary gross phenotyping outcomes from PreMA reagent injections.

For *tdgf1*, Experiments 1a and 1b correspond to technical replicates using WT 1 as reference, uninjectected control. *chrD* and *tdgf1* phenotypes were scored on 1 dpf, whereas *tyr*, *ttn.2 N2B*, *ttn.2* phenotypes were scored on 2 dpf.

https://doi.org/10.1371/journal.pgen.1007652.s006
List and summary sequence outcomes of Low Competition sgRNA that were designed around the Competition Hypothesis. Underlined & Italicized bases in gRNA sequence denote mismatched bases due to the promoter requirement. Pattern Scores and Microhomology Scores were derived using RGEN online tool. MH: Microhomology, SC: Subcloning, TIDE: Tracking Indels by DEComposition. †Injected RNP at the dose of 115 pg sgRNA and 245 pg Cas9 due to poor viability at higher doses.

![Table S3 List and summary sequence outcomes of Low Competition sgRNA that were designed around the Competition Hypothesis.](https://doi.org/10.1371/journal.pgen.1007652.s007)
Table S4 List and summary sequence outcomes of Medium ~ High Competition sgRNA that were designed around the Competition Hypothesis.

Underlined & Italicized bases in sgRNA sequence denote mismatched bases due to the promoter requirement. Pattern Scores and Microhomology Scores were derived using RGEN online tool (http://www.rgenome.net). MH: Microhomology, SC: Subcloning, TIDE: Tracking Indels by Decomposition. † Injected RNP at the dose of 115 pg sgRNA and 245 pg Cas9 due to poor viability at higher doses.

https://doi.org/10.1371/journal.pgen.1007652.s008

<table>
<thead>
<tr>
<th>sgRNA</th>
<th>gRNA Sequence</th>
<th>MH Score</th>
<th>Slope value</th>
<th>Top Predicted Allele</th>
<th>%2 Allele</th>
<th>Injection Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pattern Score</td>
<td>Microhomology</td>
<td>Deletion length</td>
<td>Pattern Score</td>
</tr>
<tr>
<td>kita #1</td>
<td>GCCTACACATAACCCAGA</td>
<td>6513.9</td>
<td>-36.9</td>
<td>461.6</td>
<td>CCC</td>
<td>11 bp</td>
</tr>
<tr>
<td>kita #2</td>
<td>GAATCCCAGAGAGGTCCG</td>
<td>6879.1</td>
<td>-36.9</td>
<td>461.6</td>
<td>CCC</td>
<td>11 bp</td>
</tr>
<tr>
<td>frpec #2</td>
<td>GTGACGACACGTGTTT</td>
<td>4429.1</td>
<td>-24.5</td>
<td>335.0</td>
<td>CAC</td>
<td>8 bp</td>
</tr>
<tr>
<td>frpec #3</td>
<td>GGGATATTAGACACACATA</td>
<td>5405.3</td>
<td>-38.2</td>
<td>377.6</td>
<td>GACAC</td>
<td>15 bp</td>
</tr>
<tr>
<td>frpec #4</td>
<td>GGCGACACATCCAGATTT</td>
<td>3814.2</td>
<td>-18.3</td>
<td>179.6</td>
<td>ACT</td>
<td>16 bp</td>
</tr>
<tr>
<td>mprv17 #2</td>
<td>GCTGGAACCTTTACCCTTC</td>
<td>4305.1</td>
<td>-36.1</td>
<td>389.5</td>
<td>CCT</td>
<td>5 bp</td>
</tr>
<tr>
<td>ttn.2 #3†</td>
<td>GTCTAAAGGAGATGCGGT</td>
<td>5994.2</td>
<td>-16.5</td>
<td>274.5</td>
<td>CCA</td>
<td>12 bp</td>
</tr>
<tr>
<td>ttn.2 N2B #2</td>
<td>GGCACACACAGAACAGCTC</td>
<td>3547.5</td>
<td>-16.5</td>
<td>229.6</td>
<td>CCAGA</td>
<td>25 bp</td>
</tr>
</tbody>
</table>
Table S5 List and summary sequence outcomes of human CRISPR-Cas9 targets that were designed around the Competition Hypothesis V2.

Underlined & Italicized bases in sgRNA sequence denote mismatched bases due to the T7 promoter requirement. For loci wherein mutagenic efficiency and/or Top MH Fraction was calculated based on subcloning results, number of mutant/top predicted allele colonies are given in numerator and the total number of colonies analyzed are given in the denominator. Pattern Scores and Microhomology Scores were derived using RGEN online tool (http://www.rgenome.net).

https://doi.org/10.1371/journal.pgen.1007652.s009

<table>
<thead>
<tr>
<th>sgRNA</th>
<th>gRNA Sequence</th>
<th>MH Score</th>
<th>Top Predicted Allele</th>
<th>#2 Allele</th>
<th>RNP Transfection Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pattern Score</td>
<td>Micrhomology</td>
<td>Deletion length</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TIDE</td>
<td>SC - T7E1</td>
<td>Method</td>
</tr>
<tr>
<td>AAV51 #1</td>
<td>GCTGATGAGAACCTGGAAGG</td>
<td>5761.2</td>
<td>370.5</td>
<td>YGG</td>
<td>6 bp</td>
</tr>
<tr>
<td>AAV51 #2</td>
<td>GATCCCTCTGAAGCAGCAGG</td>
<td>5558.7</td>
<td>402.0</td>
<td>GCC</td>
<td>8 bp</td>
</tr>
<tr>
<td>AAV51 #3</td>
<td>GGCGGCGATGCTGGCCGCAGG</td>
<td>8178.8</td>
<td>469.0</td>
<td>GCC</td>
<td>8 bp</td>
</tr>
<tr>
<td>AAV51 #4</td>
<td>GGCGGAGATGCGCCGCCGG</td>
<td>7592.5</td>
<td>516.6</td>
<td>CGG</td>
<td>3 bp</td>
</tr>
<tr>
<td>CSF2 #1</td>
<td>GCTGGGCCGGGCGGAGCGG</td>
<td>10775.6</td>
<td>670.0</td>
<td>CGCC</td>
<td>8 bp</td>
</tr>
<tr>
<td>GUB2 #1</td>
<td>GCAGATGAGAAGATGGAAT</td>
<td>4506.2</td>
<td>389.5</td>
<td>TGG</td>
<td>5 bp</td>
</tr>
<tr>
<td>GUB2 #2</td>
<td>GCTCCACATCCGCTAT</td>
<td>6362.6</td>
<td>423.0</td>
<td>GCC</td>
<td>7 bp</td>
</tr>
<tr>
<td>MYO7A #1</td>
<td>GCTCCCTCCCGGCTGGACTTG</td>
<td>4744.5</td>
<td>335.0</td>
<td>CAC</td>
<td>8 bp</td>
</tr>
<tr>
<td>MYO7A #2</td>
<td>GGCGATGCTGGAGGAGT</td>
<td>4805.5</td>
<td>352.5</td>
<td>TCC</td>
<td>7 bp</td>
</tr>
<tr>
<td>MYO7A #3</td>
<td>GCGAGTCTTCAGCTGAC</td>
<td>5605.3</td>
<td>352.5</td>
<td>CTG</td>
<td>7 bp</td>
</tr>
<tr>
<td>MYO7A #4</td>
<td>GACCAGGCTCTGACTCAF</td>
<td>5605.3</td>
<td>352.5</td>
<td>CTG</td>
<td>7 bp</td>
</tr>
</tbody>
</table>
Table S6 Example MENTHU output from select CRISPR-Cas9 targets used in this study, focusing only on out-of-frame mutations.

The output was obtained by using the entire target exon sequence with 40 bp intronic sequence each on both 5’ and 3’ ends. The MENTHU output provides a 3’ NGG PAM sequence for each gRNA targets (italicized and underlined). MENTHU gRNA outputs that matched the target sequences used in this study are bolded. Criteria 1 and 2 refer to 1) if top predicted microhomoology arm is separated by 5 bp or less, and 2) if the ratio of top to second predicted Pattern Scores is at least 1.5. MENTHU is programmed to terminate calculations if the target site is negative for Criterion 1. As a result, no gRNA sequence output is obtained for chr#1 and mitfa #2. Importantly, in two instances (surf1 and tgf#1) where we only had Group 3 reagents, novel candidate PreMA sites were identified. * Result obtained by adjusting the value for Criterion 2 to 1.0 as these were Group 3 guides that, by definition, does not satisfy Criterion 2 of 1.5 or higher. † in-frame mutation by the experimental design. ‡ 16 other candidate loci identified on this 3771 bp exon; only a partial list for alternate loci is given. § 16 other candidate loci identified on this 822 bp exon; only a partial list for alternate loci is given.

https://doi.org/10.1371/journal.pgen.1007652.s010

<table>
<thead>
<tr>
<th>sgRNA</th>
<th>gRNA sequence: this study</th>
<th>gRNA sequence: MENTHU</th>
<th>Grouping: this study</th>
<th>Grouping: MENTHU</th>
<th>Additional PreMA loci identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr#1</td>
<td>GGCTGGATGGCCAGGCGC</td>
<td>N/A</td>
<td>Group 1</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>kppoc #2</td>
<td>GTCAAGCGACGAGCTT</td>
<td>GAAGTCCTACGACAGGGATTTTT</td>
<td>Group 4</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>mitfa #4</td>
<td>G2AGCCCGCCAGCTTGA</td>
<td>N/A</td>
<td>Group 2</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>mtg1</td>
<td>G2AGGCCATTGGAGAGAG</td>
<td>CTTGAGGCGAGGACAGGAGA</td>
<td>Group 3</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>surf1</td>
<td>G2TCAGAGACCCCTGGCTTC</td>
<td>GTCCAGAGATCCCAGCTTCCTCT</td>
<td>Group 3</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>tgf#1</td>
<td>G2AGCCGCTTACCTTGCTG</td>
<td>GAGGGAGGCTTTACGCTTCTG</td>
<td>Group 3</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>tnf2 #1</td>
<td>G2AGTTAAGTTGAGAG</td>
<td>CTGGTTTGAAGTGAGGACGAGA</td>
<td>Group 4</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>tnf2 #2</td>
<td>G2AGATGGGTCGCTCTTG</td>
<td>AGGGAGATGGGTCGCTCTCTG</td>
<td>Group 4</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>tnf2 N26 #1</td>
<td>G2ATTGCCAACCAACGAGAC</td>
<td>TGGATTGCCACACACAGGAC</td>
<td>Group 4</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>tyr #1</td>
<td>G2GCCCACGATATCCTACCT</td>
<td>GAGGCGCAGATATCCTACCTCA</td>
<td>Group 4</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>
Table S7 Preliminary analyses on the prevalence of PreMA loci reveal that about 10% of the CRISPR-Cas9 targetable loci on both human CSF2 and zebrafish tp53 genes fall in this category.

This holds true for both at the gene and exonic levels. As expected, roughly two thirds of the PreMA reagents are predicted to induce preferentially out-of-frame mutations.

https://doi.org/10.1371/journal.pgen.1007652.s011

<table>
<thead>
<tr>
<th>Gene</th>
<th>Target</th>
<th>Forward</th>
<th>Reverse</th>
<th>Polymerase</th>
<th>Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>TALEN</td>
<td>sgRNA #1</td>
<td>CCACCCAGACATACAACACACAAC</td>
<td>CGGAAGCTGGTACGTACCTTTGAGG</td>
<td>MyTelq</td>
<td>R</td>
</tr>
<tr>
<td>sgRNA #2</td>
<td>AAGCTTCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #3</td>
<td>SGACGCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #4</td>
<td>CAGGCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #5</td>
<td>TCCACCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #6</td>
<td>CCACCCAGACATACAACACACAAC</td>
<td>CGGAAGCTGGTACGTACCTTTGAGG</td>
<td>MyTelq</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>sgRNA #7</td>
<td>CAGGCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #8</td>
<td>CAGGCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #9</td>
<td>CAGGCTAGTACCCATTGTAACGCA</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>urod</td>
<td>sgRNA #1</td>
<td>GCCAATGGGGTGGGCTGCTGGG</td>
<td>MyTelq</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>sgRNA #2</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #3</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #4</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #5</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #6</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #7</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #8</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgRNA #9</td>
<td>GACGCTTCATCCGATAGG</td>
<td>MyTelq</td>
<td>F</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S8 List of primers used in this study.

All the primer sequences are provided in 5’ -> 3’ order. For urod Reverse primer, M13F primer sequence was added at the 5’ end of the endogenous target sequence (bolded and italicized). For SDM primers, intended point mutation is indicated by bold and italic. * No endogenous primer was used to sequence the genomic loci of interest.

https://doi.org/10.1371/journal.pgen.1007652.s012
**S1 Note Calculation of Microhomology Fraction**

1) When the mutagenic outcomes were assessed by subcloning, the Microhomoloy Fraction was calculated according to the formula below:

\[ \text{MH Fraction} = \frac{\text{(# of mutant colonies harboring 3bp microhomology or longer)}}{\text{(# of total mutant colonies recovered)}} \]

2) When the mutagenic outcomes were assessed by TIDE analysis, the Microhomology Fraction was not calculated.

3) For HeLa cell data, Microhomology Fraction was calculated as below, discarding any alleles with allele frequency of < 0.1%.

\[ \text{MH Fraction} = \frac{\text{(# of mutant reads harboring 3bp microhomology or longer)}}{\text{(# of total mutant reads recovered)}} \]

**S2 Note Calculation of Slope Values**

The Slope Values were calculated for each target locus as follows:

1) Input 80bp endogenous gene sequence flanking the predicted DSB site into the Microhomology-Predictor (http://www.rgenome.net/mich-calculator/) \cite{Bae, 2014 #1}.

   a. In the case of CRISPR-Cas9 reagents, phosphodiester bond between the 3\textsuperscript{rd} and 4\textsuperscript{th} base pairs distal to the PAM was chosen as the presumptive DSB site. Subsequently, 40bp each on both sides of this break site was used as the input sequence.

   b. In the case of TALEN reagents with even number of bases in the Spacer region, the phosphodiester bond between the 5’ and 3’ halves of the Spacer was chosen as the presumptive DSB site. Subsequently, 40bp each on both sides of this break site was used as the input sequence.
c. In the case of TALEN reagents with odd number of bases in the Spacer region, the center-most base that bridges 5’ and 3’ halves was identified. Subsequently, a 79bp sequence containing this center-most base and 39bp each on both sides of this base was used as the input sequence.

2) Ranked the top 10 candidates by the Pattern Score in a descending order.
   a. In the dataset wherein only microhomology arms of 3bp or greater were considered, candidate mutant alleles harboring only 2bp microhomology arms were omitted from further analysis.
   b. In the dataset wherein microhomology arms of 2bp or greater were considered, no candidate mutant alleles were omitted.

3) Plotted the Pattern Scores against the numerical rank on Scatter Plot using Microsoft Excel
   a. Drew simple linear regression
   b. The “a” in the fitted line formula $y = a \times x + b$ is the Slope Value

4) The steeper (i.e. larger absolute value for a) the slope, the lower influence there is from the competing locally available microhomology arms. For zebrafish, the cutoff used for Low Competition sites was -40.

5) The flatter (i.e. values closer to 0 for a) the slope, the more influence there is from the competing locally available microhomology arms. For zebrafish, the cutoff used for High Competition sites was -20.

S3 Note Calculation of Top Microhomology Fraction

1) When the mutagenic outcomes were assessed by subcloning, the Top Microhomology Fraction was calculated according to the formula below:
Top MH Fraction = \frac{(\text{of colonies with mutation corresponding to top predicted allele with } 3\text{bp microhomology or longer})}{(# \text{ of total mutant colonies recovered})}

2) When the mutagenic outcomes were assessed by TIDE analysis, the Top Microhomology Fraction was calculated according to the formula below:

Top MH Fraction = \frac{\left(\% \text{ mutat population corresponding with the same number of bases deleted as top predicted allele with } 3\text{bp microhomology or longer}\right)}{(100\% - \% 0\text{bp change population})}

e.g) *ttna* sgRNA #1 (See below for TIDE output)

Top MH Fraction = \frac{29.7\%}{(100\% - 56.2\%)} = 0.678

3) For HeLa cell data, Microhomology Fraction was calculated as below, discarding any alleles with allele frequency of < 0.1%.

Top MH Fraction = \frac{(\text{of reads with mutation corresponding to top predicted allele with } 3\text{bp microhomology or longer})}{(# \text{ of total mutant reads recovered})}
S1 Data Sanger sequencing file used for the study.

Whole amplicon sequencing outcomes are deposited as .ab1 files for chr d TALEN and tyr #2, tdgf1, ttn.2 N2B #1, ttn.2 #2 sgRNA targets. Other sequencing outcomes, including those used for subcloning analyses are provided in the .fastq formats.

https://doi.org/10.1371/journal.pgen.1007652.s016

Acknowledgments

We would like to acknowledge Melissa McNulty for TALEN synthesis and data curation, Mark Urban for zebrafish husbandry and help with microinjections, Patrick Blackburn for designing chr d #1 sgRNA, William Gendron for his assistance with human cell work, Mayo Clinic Zebrafish Facility and staff for their support, as well as Bryce Bergene and Mayo’s Creative Studio their help with figures. We also thank the Jin-Soo Kim group for developing the Microhomology-Predictor CRISPR RGEN Tool, for making source code freely available, and for sharing the deep sequencing output from their HeLa cell experiments. We thank Wesley A Wierson and Jeffrey J Essner for the tyr #2 gRNA, and the research groups of Dr. Essner and Dr. Maura McGrail for valuable discussions and feedback and for MENTHU server testing, and Carolyn Lawrence-Dill and her group, especially Darwin Campbell, for valuable discussions and hosting services for MENTHU.

References


APPENDIX C. GeneWeld: A METHOD FOR EFFICIENT TARGETED INTEGRATION DIRECTED BY SHORT HOMOLOGY

Wesley A. Wierson, Jordan M. Welker, Maira P. Almeida, Carla M. Mann, Dennis A. Webster, Trevor J. Weiss, Melanie E. Torrie, Macy K. Vollbrecht, Merrina Lan, Kenna C. McKeighan, Zhitao Ming, Alec Wehmeier, Christopher S. Mikelson, Jeffrey A. Haltom, Kristen M. Kwan, Chi-Bin Ghien, Darius Balciunas, Stephen C. Ekker, Karl J. Clark, Beau R. Webber, Branden Moriarity, Staci L. Solin, Daniel F. Carlson, Drena L. Dobbs, Maura McGrail, Jeffrey J. Essner

This manuscript is in preparation for submission to Development.

It describes the protocol GTagHD was designed to facilitate.

A pre-print of this paper is available at:

https://www.biorxiv.org/content/10.1101/431627v1

I wrote the portion of the manuscript describing the GTagHD web tool. I created and programmed the entirety of the GTagHD web tool.
APPENDIX D. GENE SCULPT SUITE USAGE STATISTICS

I implemented Google Analytics tracking for each Gene Sculpt Suite web tool. Google Analytics (GA) collects a (terrifyingly) large amount of data about each user visiting the web tools and is able to localize users based on their IP addresses and Internet Service Providers, in addition to identifying unique users through browser cookies. GA is adept at filtering out bot traffic and can differentiate between pageviews and sessions. A pageview implies that a user visited a page but did not interact with it. A session is the period of time in which a user is actively engaged with the page (i.e., using the tool.)

I began tracking MENTHU in January of 2018 in conjunction with the expected release of a pre-print in March. MEDJED and GTAGHD were tracked starting in April of 2018. Here I present the GA usage statistics for each tool as of April 4, 2019, prior to the publication of all three methods in the NAR webserver issues (July 2019).

MEDJED

MEDJED is an unpublished method, which has been presented at conferences but is not available as a pre-print, and as such experiences less traffic than GTAGHD or MENTHU. MEDJED has 169 pageviews and 114 sessions, initiated by 69 users from the United States, Canada, Germany, Singapore, Spain, Hong Kong, Israel, and Japan (see Fig. I-1). Those users are engaged with MEDJED for, on average, 3:51m.

Within the United States, MEDJED is heavily utilized by users in Ames, but also has users in locations including (but not limited to) Ashburn, VA (possibly from the HHMI Janelia Research Campus), Ithaca, New York (from the Cornell University network), and St. Paul Minnesota (University of Minnesota).
Figure I-1 Country-of-origin of MEDJED users. The overwhelming majority of MEDJED users are located in the United States.

MEDJED does not experience high traffic, but it does experience relatively consistent traffic with an average of 14 sessions per month (see Fig I-2).

Figure I 2 MEDJED users per day between April 01 2018 and April 04 2018.
GTagHD

GTagHD is also unpublished, but the protocol it is used in, GeneWeld, was posted as a pre-print on bioRxiv on Oct 3, 2018:
https://www.biorxiv.org/content/biorxiv/early/2018/10/03/431627

GTagHD has recorded 839 sessions from 384 users in 21 different countries and territories including the United States, Canada, Germany, China, the United Kingdom, Singapore, Japan, Austria, Switzerland, the Netherlands, Belgium, South Korea, Norway, Australia, Spain, Israel, Morocco, Argentina, Hungary, Serbia, and Sweden (see Fig. I-3). Although 71 of the 287 US-based users are located in Ames (GTagHD is used in Iowa State University undergraduate classes that employ pGTag plasmids), the remaining users are spread over at least 28 states (including Iowa). 19 users (with a total of 42 sessions between them) trace back to the Longwood Medical and Academic Area Network in Boston, Massachusetts, which serves, amongst other institutions, Harvard’s Medical School and teaching hospitals, the Dana-Farber Cancer Institute, Massachusetts College of Pharmacy and Health Sciences, and Boston Children’s Hospital; another 7 users accessed the site from Cambridge (home to Harvard and MIT.) Another 15 users initiating 30 sessions trace back to Davis, California.

All told, users are engaged with GTagHD for an average of 2:10 minutes per session. GTagHD is used, on average, 71 times per month (see Fig. I-4).
Figure I-3 Country-of-origin of GTagHD users. GTagHD users hail from 21 different countries around the world.

Figure I-4 GTagHD users per day from Apr 01 2018 to Apr 04 2018

**MENTHU**

MENTHU was published as part of a *PloS Genetics* paper by Ata et al. in Oct 2018, but was posted online as a pre-print in March of 2018. As a published method, it experiences fairly heavy traffic from 481 users located in 28 countries and territories including the US, China, Japan, Austria, France, Australia, the UK, Canada, Germany, Hong Kong, the
Netherlands, Spain, Sweden, South Korea, Switzerland, Hungary, Belgium, India, Iran, Russia, Argentina, Brazil, Colombia, Ireland, Mexico, Norway, New Zealand, and Singapore (see Fig. I-5).

Of the 285 users located in the US, the majority (93) originate from Ashburn, VA, where they have a total of 190 sessions. Only 40 users are located in Ames (though they are responsible for 141 sessions among them); other users are spread across 26 states in cities including Iowa City, IA, Coralville, IA, Boston, MA, New York City, NY, St. Paul, MN, Chicago, IL, Baltimore, MD, Berkeley, CA, Bethesda, MD, Minneapolis, MN, Madison, WI.

MENTHU’s users have logged 943 sessions between them, lasting on average 2:40 (see Fig. I-6).
Figure I-6 MENTHU daily users between Jan 20, 2018 and Apr 04, 2019.

Conclusions

These usage statistics demonstrate that the Gene Sculpt Suite is actively and widely used not only at Iowa State but (literally) around the world. Even though MEDJED and GTagHD are as-yet unpublished, they have been used (collectively) by more than 400 people, more than 900 times. MENTHU, as a published method, has seen even greater use. This level of use demonstrates that the tools in the Gene Sculpt Suite are of sufficient interest and utility to justify continued development of the Gene Sculpt Suite webserver.