MicroRNA target site identification by integrating sequence and binding information

William H Majoros[1,7], Parawee Lekprasert[1,2,7], Neelanjan Mukherjee[1,3], Rebecca L Skalsky[4], David L Corcoran[1], Bryan R Cullen[4] & Uwe Ohler[1,3,5,6]

High-throughput sequencing has opened numerous possibilities for the identification of regulatory RNA-binding events. Cross-linking and immunoprecipitation of Argonaute proteins can pinpoint a microRNA (miRNA) target site within tens of bases but leaves the identity of the miRNA unresolved. A flexible computational framework, microMUMMIE, integrates sequence with cross-linking features and reliably identifies the miRNA family involved in each binding event. It considerably outperforms sequence-only approaches and quantifies the prevalence of noncanonical binding modes.

miRNAs play an important role in gene regulatory pathways, and the dysregulation of several miRNAs has been implicated in disease[1]. As part of the RNA-induced silencing complex (RISC), miRNAs guide the complex to repress target mRNAs. Immunoprecipitation of Argonaute (AGO) protein family members followed by global profiling of bound RNA has provided an experimental high-throughput approach to map miRNA targets genome wide[2]. In particular, cross-linking and immunoprecipitation (CLIP) of AGO proteins provides specific regions of likely AGO binding and miRNA target sites[3,4].

The resolution of CLIP experiments has been generally insufficient to unambiguously identify the acting miRNA via sequence matches[4]. Recent protocols, including individual nucleotide–resolution CLIP (iCLIP) and photoactivatable ribonucleoside–enhanced CLIP (PAR-CLIP), increased resolution by providing additional signals diagnostic of cross-linked locations[5]. In particular, PAR-CLIP obtains high cross-linking efficiency owing to the presence of the nucleoside analog 4-thiouridine, which leads to characteristic thymine-to-cytosine transitions in the vicinity of protein-mRNA cross-linking[4]. To define a narrow region (typically ~30 nucleotides (nt)) most likely bound by the protein, our previously published PARalyzer software[6] quantifies this T-to-C transition profile relative to the background read density via a kernel density estimator (**Fig. 1a**).

miRNA target sites are characterized by one of several types of seed sequence matches unique to each miRNA family[7] (canonically, a perfect 7-nt complement to positions 2–8 of the miRNA). In AGO-immunoprecipitated libraries, the cross-linking site indicates a radius for RISC-mediated miRNA heteroduplex formation on the mRNA substrate, and this radius restricts the search space for functional seed matches. Furthermore, the PAR-CLIP transition signal is preferentially located directly 5′ of the miRNA seed match on the mRNA[4,6,8], likely because of the biophysical properties of the heteroduplex in functional microribonucleoprotein complexes[7] (**Fig. 1b**). mRNA conservation levels immediately 3′ of peaks indicate that many of these seed matches are selectively maintained (**Fig. 1c**).

Despite continuous improvements in *de novo* computational predictions of target sites, efforts have been fundamentally limited by the brevity of sequence motifs describing functional miRNA-target relationships. The quantitative information in AGO PAR-CLIP data sets motivated us to develop an integrative computational method to identify miRNA target sites at the nucleotide level with high accuracy. It combines the T-to-C transition signal with miRNA seed-pairing and its evolutionary conservation as well as the spatial relation between binding and sequence features. We implemented the model in a general-purpose hidden Markov model (HMM) framework[9] called MUMMIE (multivariate Markov modeling inference engine). We modeled the shape of PAR-CLIP signals with a six-state topology (**Fig. 1d**), in which state 5 expands into a 41-state submodel for detection of seed matches of various types (**Fig. 1e**). The microMUMMIE model is parameterized to preferentially predict seed matches that are long, located closely 3′ to a PAR-CLIP peak and highly conserved. Trade-offs between these features are naturally accounted for during prediction so that suboptimal sites can still be detected, albeit with lower scores (posterior probabilities; **Supplementary Fig. 1**).

Most of the current target predictors identify candidate sites within whole transcripts, usually parameterized on the indirect cumulative effect on mRNA or protein steady-state levels. The availability of *in vivo* direct binding data changes the problem of target prediction into one of assigning the most likely miRNA or miRNAs to each of the observed AGO binding locations. We applied suitable site-level target predictors to cross-linked regions from multiple data sets and evaluated the improvement obtained by the explicit joint model in the new target predictor. Analyses were limited to regions covered by aligned PAR-CLIP

[1]Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, USA. [2]Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina, USA. [3]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. [4]Department of Molecular Genetics and Microbiology, Duke University, Durham, North Carolina, USA. [5]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA. [6]Department of Biology, Humboldt University of Berlin, Berlin, Germany. [7]These authors contributed equally to this work. Correspondence should be addressed to U.O. (uwe.ohler@mdc-berlin.de).

RECEIVED 2 JANUARY; ACCEPTED 1 MAY; PUBLISHED ONLINE 26 MAY 2013; DOI:10.1038/NMETH.2489

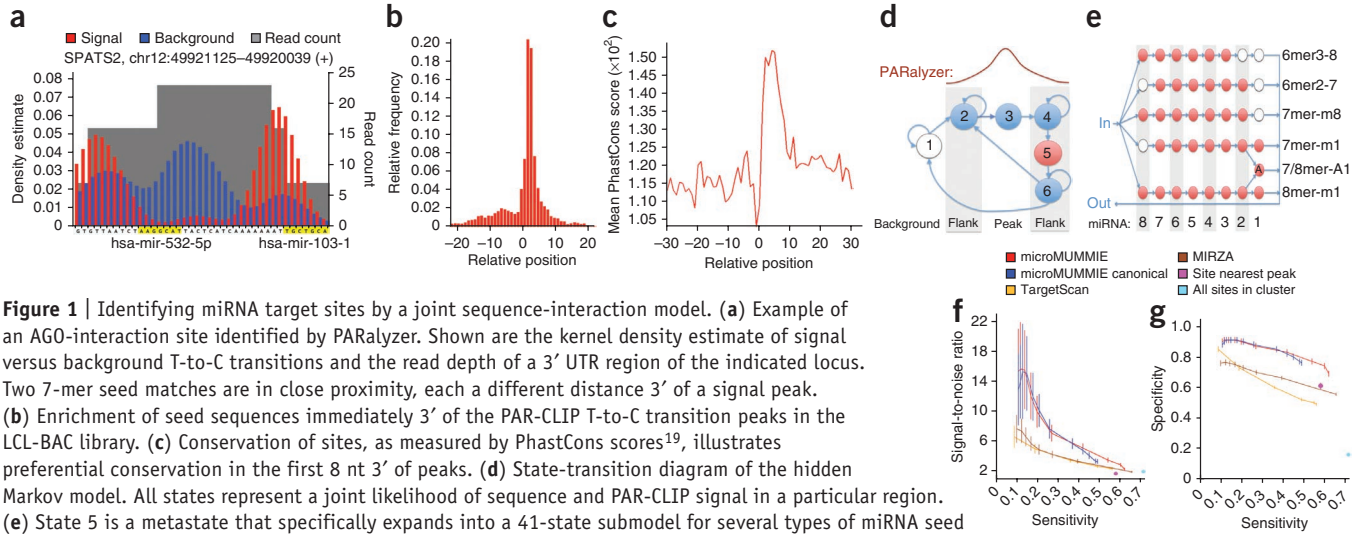

Figure 1 | Identifying miRNA target sites by a joint sequence-interaction model. (**a**) Example of an AGO-interaction site identified by PARalyzer. Shown are the kernel density estimate of signal versus background T-to-C transitions and the read depth of a 3′ UTR region of the indicated locus. Two 7-mer seed matches are in close proximity, each a different distance 3′ of a signal peak. (**b**) Enrichment of seed sequences immediately 3′ of the PAR-CLIP T-to-C transition peaks in the LCL-BAC library. (**c**) Conservation of sites, as measured by PhastCons scores[19], illustrates preferential conservation in the first 8 nt 3′ of peaks. (**d**) State-transition diagram of the hidden Markov model. All states represent a joint likelihood of sequence and PAR-CLIP signal in a particular region. (**e**) State 5 is a metastate that specifically expands into a 41-state submodel for several types of miRNA seed matches. (**f**) Sensitivity-SNR trade-off for microMUMMIE, TargetScan, MIRZA and two baselines, based on the 100 top-expressed miRNAs in the LCL-BAC cell line. Explicit expression information was not used by any predictor (see also **Supplementary Fig. 2**). (**g**) Sensitivity-specificity trade-off as in **f**. Error bars indicate ±1 s.d.

reads (termed groups); with this restriction, methods were much more accurate than those applied without the CLIP information (**Supplementary Fig. 2**). Because the identity of the miRNA in bound AGO complexes is unknown, we assessed the relative prediction accuracy via signal-to-noise ratio (SNR) based on randomized shuffling of miRNA sequences (enrichment of complementary miRNA seeds over decoy sequences among the predictions; Online Methods).

Specifically, we compared microMUMMIE to several alternatives. One was the latest release of TargetScan[10], a state-of-the-art *de novo* miRNA target site predictor that uses conservation evidence as well as various sequence context scores but makes no explicit use of CLIP information. Another was MIRZA[11], which implements a new biophysical model of target recognition that is based on context-specific RNA duplex formation estimated from high-quality PAR-CLIP data. We contrasted these model-based approaches with two baseline strategies for identifying miRNA target sites from PAR-CLIP data: (i) choosing the seed match nearest to each PARalyzer peak and (ii) predicting all seed matches in each PARalyzer cluster (the maximal interval in which T-to-C transition rate exceeds background expectation[6]). All approaches used the 100 most highly expressed miRNAs in the cell line, which delivered the best trade-off between SNR and sensitivity (**Supplementary Fig. 3**).

On a PAR-CLIP data set from a lymphoblastoid cell line infected with Epstein-Barr virus (LCL-BAC)[12], our model achieved higher sensitivity, specificity and SNR than other approaches (**Fig. 1f**,**g**). Thus, the joint model has a greater ability to discriminate real from decoy target sites (has a higher SNR), and it does so by predicting fewer randomized decoy sites within those CLIP groups (has higher specificity). TargetScan limits predictions to the three 'canonical' seed matches (7mer-A1, 7mer-m8 and 8mer-A1), whereas our model predicts seven types (6mer3-8, 6mer2-7, 7mer-m8, 7mer-m1, 7mer-A1, 8mer-A1 and 8mer-m1; see also **Supplementary Fig. 4**). MIRZA's energy-based model scores the whole duplex and is therefore agnostic to specific seed matches. Although MIRZA reached a higher sensitivity than microMUMMIE, the largest contribution by far of 'noncanonical' targets appeared to involve the 6-mer seeds included in microMUMMIE. The baseline models exhibited competitive sensitivity, but at very low SNR or specificity levels. We confirmed all of the observed performance trends with results on additional PAR-CLIP data sets prepared by micrococcal nuclease (MNase) digestion[13] in place of the double RNase T1 digestion used in the original protocol. (**Supplementary Fig. 5** and **Supplementary Table 1**).

To make results as comparable as possible, we used TargetScan's branch-length score as conservation evidence[10], which is unavailable for 6-mer seed matches. This modification effectively biases our model against predicting 6-mer sites when the model is parameterized to weight conservation highly, and it helped improve specificity and SNR at the low-sensitivity end of the spectrum (**Supplementary Fig. 2**). Increasing the number of miRNAs represented in our model beyond the 100 most highly expressed ones progressively degraded performance (**Supplementary Fig. 3**), a result illustrating the complementary benefit of accounting for miRNA expression in addition to RISC binding information. Examples of several previously or newly experimentally validated microMUMMIE predictions[12] involving different seed types are given in **Supplementary Figures 6** and **7**.

To directly address the question of whether predicted targets were due to the presence of the specified miRNA, we contrasted the PAR-CLIP signal at LCL-BAC predicted sites with PAR-CLIP data from three additional Epstein-Barr virus–infected cell lines, each of which has one viral miRNA deleted (BHRF1-1, BHRF1-2 and BHRF1-3). The PAR-CLIP signal in the deletion lines was nearly absent at the locations of predictions involving a BHRF miRNA in the original cell line (**Fig. 2a**,**b** and **Supplementary Fig. 8**). For BHRF1-2 and BHRF1-3, but not BHRF1-1, the loss of targeting in the corresponding deletion line was accompanied by consistent mRNA upregulation of microMUMMIE-predicted targets, as measured by Illumina sequencing (**Fig. 2c** and Online Methods). Predicted target mRNAs are often regulated by multiple miRNAs, and downregulation mediated by individual

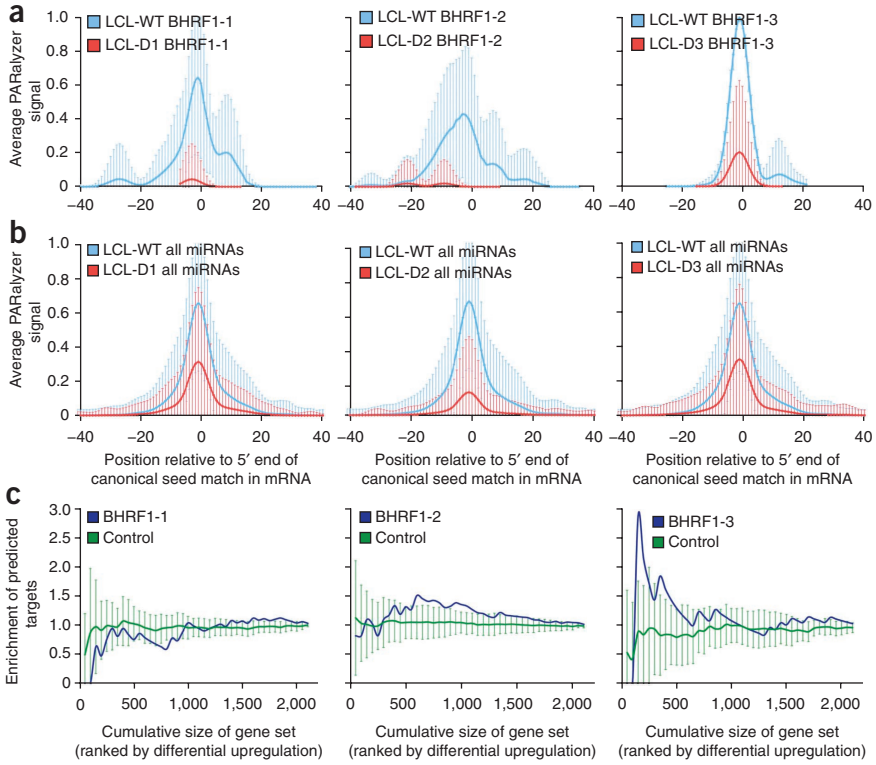

Figure 2 | Validation of predicted sites and their impact on expression. (**a**) Loss of PAR-CLIP signal at LCL-BAC predicted target sites for BHRF1-1, BHRF1-2 and BHRF1-3 (blue) in the corresponding deletion lines (LCL-BAC-D1, LCL-BAC-D2 and LCL-BAC-D3, respectively; red) (23 predicted target sites for BHRF1-1, 52 for BHRF1-2, and 10 for BHRF1-3). (**b**) Control results: signal difference for predicted target sites in LCL-BAC (blue) versus deletion lines (red) for all but the miRNA of interest (2,159 predicted targets for the D1 control, 2,082 for the D2 control and 2,172 for the D3 control). Binding loss of BHRF1-1 ($P = 0.0007$) and BHRF1-2 ($P = 0.0329$) were statistically significant compared to the control; loss of BHRF1-3 was consistent but not significant owing to the small number of sites ($P = 0.1714$; Wilcoxon rank-sum test). Predictions were obtained using the 100 top-expressed miRNAs, with the microMUMMIE PAR-CLIP signal variance parameter set to 0.01. All predicted targets were aligned across the beginning of miRNA seed matches in the mRNAs. (**c**) Impact of site loss on steady-state mRNA expression between LCL-BAC and each of the deletion lines based on RNA-seq data. The enrichment of genes with BHRF target sites among the top gene sets, ranked by differential upregulation in a deletion line compared to LCL-BAC, is contrasted with the enrichment of control miRNAs with similar numbers of predicted targets. Error bars indicate ±1 s.d.

BHRF1-1 sites may not be apparent in mRNA steady-state levels; alternatively, the effect of BHRF1-1 may not be mediated at the RNA level. microMUMMIE-based observations agreed with MIRZA-predicted targets (**Supplementary Fig. 9**).

As shown in previous studies, the presence of a perfect seed match cannot explain all CLIP groups[4,6,14]. In the LCL-BAC data set, the sensitivity of microMUMMIE did not exceed 80% at the most lenient settings; a naive scan for 6-mer seed matches to the 100 top-expressed miRNAs did not exceed 84% (**Supplementary Fig. 10**). This motivated us to investigate the extent to which binding with an imperfect seed match might explain remaining CLIP groups. We focused on a recently described class of imperfect targets, namely, 'bulge' sites, which have been suggested to explain as many as 25% of all the high-throughput sequencing–CLIP (HITS-CLIP) clusters for miR-124 (ref. 14). We adapted our model to sites in which one of the mRNA residues pairs with the miRNA only during an initial annealing step but becomes unpaired in the final duplex, forming a bulge in the mRNA. In addition to the previously proposed pivot pairing between mRNA positions 5 and 6, we investigated potential bulges between positions 4 and 5 and between positions 3 and 4. Bulge-site predictions showed distinctly lower SNR profiles, and when we combined all three pivot locations, predictions covered only ~15–20% of orphan groups (~1–2% of total groups; **Supplementary Table 2** and **Supplementary Figs. 11–13**).

Our findings illustrate the advantage of making explicit use of deep-sequencing data over *de novo* sequence analysis to find seed matches in read clusters for identifying miRNA target sites active in a specific condition. By jointly modeling multiple relevant variables—including evolutionary conservation, residue transition rates, spatial positioning and sequence composition—microMUMMIE permits finer discrimination between true and decoy sites. There is room for further improvement via incorporation of additional evidence including structural accessibility and relative or absolute production or steady-state levels of miRNAs or mRNAs. The framework provides flexibility in the ability not only to incorporate additional covariates but also to choose the desired compromise between sensitivity and specificity and to define additional classes of seed matches (such as centered sites[15] or compensatory 3′ binding[16]). However, in our comparative evaluation, genuinely imperfect seed sites appeared to account for a small fraction of total targets only. The remaining 20% of AGO clusters may reflect low-affinity transitory binding, experimental noise or AGO targets that do not involve miRNAs[17].

By using complementary information sources, we have arrived at a performance level for which the question of miRNA targeting is no longer a binary one but one of quantitative impact of a miRNA on a transcript (**Supplementary Fig. 6**; this idea has specifically been pursued in the concurrently developed MIRZA). For this type of assessment, however, multiple genomic data sets have to be generated; recent sequence-only target predictors remain more generally applicable[18], albeit at the cost of substantial reductions in accuracy. Finally, distinct transition patterns are also observed for other post-transcriptional regulators such as Quaking and Pumilio[4], suggesting the utility of our integrated modeling approach for RNA-binding protein target identification in general.

Source code and executables for microMUMMIE are available at http://www.genome.duke.edu/labs/ohler/research/MUMMIE/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene expression omnibus: GSE41437 (PAR-CLIP data for LCL-BAC, LCL-BAC-D1 and LCL-BAC-D3); GSE46611 (PAR-CLIP data for LCL-BAC-D2 and RNA-seq data for all four LCLs).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by grants from the US National Science Foundation (MCB-0822033) and US National Institutes of Health (NIH R01-GM104962) to U.O. and by awards from the NIH (R01-AI067968; R01-DA030086) to B.R.C. Contributions by R.L.S. were additionally supported by a Duke Center for AIDS Research small grant award (P30-AI064518). We thank S. Grosswendt and M. Piechotta for critical reading of a manuscript draft.

AUTHOR CONTRIBUTIONS

W.H.M. implemented the modeling framework and designed the models, P.L. and W.H.M. performed the computational experiments, P.L. investigated bulge sites, and N.M. and D.L.C. configured and ran the PARalyzer pipeline and provided guidance on analyzing its outputs. P.L., W.H.M. and U.O. analyzed the data. R.L.S., B.R.C. and N.M. designed and performed wet-lab experiments, and W.H.M., P.L. and U.O. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Jiang, Q. *et al. Nucleic Acids Res.* **37**, D98–D104 (2009).
2. Hendrickson, D.G., Hogan, D.J., Herschlag, D., Ferrell, J.E. & Brown, P.O. *PLoS ONE* **3**, e2126 (2008).
3. Chi, S.W., Zang, J.B., Mele, A. & Darnell, R.B. *Nature* **460**, 479–486 (2009).
4. Hafner, M. *et al. Cell* **141**, 129–141 (2010).
5. Sugimoto, Y. *et al. Genome Biol.* **13**, R67 (2012).
6. Corcoran, D.L. *et al. Genome Biol.* **12**, R79 (2011).
7. Bartel, D.P. *Cell* **136**, 215–233 (2009).
8. Baltz, A.G. *et al. Mol. Cell* **46**, 674–690 (2012).
9. Juang, B.H. & Rabiner, L.R. *Technometrics* **33**, 251–272 (1991).
10. Friedman, R.C., Farh, K.K.-H., Burge, C.B. & Bartel, D.P. *Genome Res.* **19**, 92–105 (2009).
11. Khorshid, M., Hausser, J., Zavolan, M. & van Nimwegen, E. *Nat. Methods* **10**, 253–255 (2013).
12. Skalsky, R.L. *et al. PLoS Pathog.* **8**, e1002484 (2012).
13. Kishore, S. *et al. Nat. Methods* **8**, 559–564 (2011).
14. Chi, S.W., Hannon, G.J. & Darnell, R.B. *Nat. Struct. Mol. Biol.* **19**, 321–327 (2012).
15. Shin, C. *et al. Mol. Cell* **38**, 789–802 (2010).
16. Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. *PLoS Biol.* **3**, e85 (2005).
17. Leung, A.K. *et al. Nat. Struct. Mol. Biol.* **18**, 237–244 (2011).
18. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. *Genome Biol.* **11**, R90 (2010).
19. Siepel, A. *et al. Genome Res.* **15**, 1034–1050 (2005).

ONLINE METHODS

PAR-CLIP data generation. Multiple Argonaute 2 PAR-CLIP libraries from two different human cell line sources (LCLs and HEK293) were used in this study. Established lymphoblastoid cell lines (LCL-BAC, LCL-BAC-D1, LCL-BAC-D2, LCL-BAC-D3) from the same donor were infected with variations of EBV B95-8 Bacmid. LCL-BAC-D1, LCL-BAC-D2 and LCL-BAC-D3 (the 'deletion lines') were infected with EBV miRNA-deficient viruses lacking miR-BHRF1-1, miR-BHRF1-2 and miR-BHRF1-3 expression, respectively[20]. All LCLs were maintained in RPMI 1640 medium supplemented with 15% FBS and 10 μg/mL gentamicin (GIBCO) and analyzed 3–5 months post-infection. PAR-CLIP data for LCL-BAC, LCL-BAC-D1 and LCL-BAC-D3 were previously published by us[12] (GEO GSE41437); PAR-CLIP data for LCL-BAC-D2 were newly generated following the same protocol (GEO GSE46611).

LCL-BAC libraries were prepared using the original PAR-CLIP protocol[4], which includes double RNase T1 digestion steps to fragment cross-linked RNA before sequencing. This has raised concerns that identified binding events may exhibit biases due to sequence preferences of the digestion enzymes[13,14]. We therefore additionally analyzed previously published PAR-CLIP data from HEK 293 cells, two biological replicates involving MNase digestion of isolated RNA in place of double T1 digestion ("MNase A," GEO GSM714646; "MNase B," GSM714647)[11]. See **Supplementary Table 1** for library and annotation metrics for all PAR-CLIP libraries.

For miRNA target prediction we considered only genomic intervals annotated as 3′ untranslated regions (UTRs) in Ensembl v.58 (ref. 21); for genes with multiple annotated transcripts (isoforms), we kept the one with the longest 3′ UTR (measured in bases of mature transcript). For the LCL-BAC data set, the miRNA set used for prediction is based on small RNA deep-sequencing data of the same culture[12]. For the external MNase data sets, we used the miRNA set of the same cell line that was reported in a previous study[4].

PAR-CLIP data processing and significance testing. PAR-CLIP data for all cell lines were processed by PARalyzer[6] to produce sequences of continuous values. PARalyzer smoothes the sequence of discrete read count values via kernel density estimation, and the resulting continuous values represent relative frequencies of T-to-C transitions. This sequence is referred to as the "signal track" for use in microMUMMIE. In practice, the PARalyzer signal track for a complete UTR will typically be sparse; i.e., 0 everywhere except at and around an occupied binding site. These nonzero regions, which correspond to scaffolds of aligned PAR-CLIP reads, are termed "groups." When binding sites are close together, their scaffolds may coalesce, resulting in one group for multiple occupied sites. PARalyzer therefore defines individual "interaction sites" by identifying maximal intervals in which the T-to-C transition signal is higher than the background; we term these intervals "clusters." Clusters therefore occur within groups, and they generally provide for a higher resolution of RBP target sites. However, the clusters' utility in pinpointing miRNA target sites in AGO libraries is complicated owing to lack of T-to-C transition signal within the miRNA-mRNA duplex site: the signal 'peak' generally does not overlap the actual miRNA seed match, and the PARalyzer software therefore provides an option to 'pad' and increase the size of AGO clusters. Here, other than for the baseline seed-match comparison, we do not explicitly use PARalyzer clusters; instead, we run microMUMMIE on the complete UTR, after normalization of the signal for individual groups. Normalization of groups ensures that the maximum signal value for each group is 1.0, and it helps to control for differing binding affinities, sequencing biases and miRNA/mRNA abundance.

To validate library-specific predictions, we compared the PAR-CLIP signal at predicted 'true' sites, for a miRNA expressed in LCL-BAC, against the signal in one of the LCL-BAC-D1/2/3 libraries not expressing the miRNA of interest (**Fig. 2a**). The significance of the signal difference was assessed by a nonparametric test, in which we first ranked target sites of all 100 top-expressed miRNAs by their difference in signal and quantified the enrichment of true targets at the top, i.e., among the most-changed sites. This implicitly accounts for the overall difference in signal observed across all targets due to incomplete saturation of the deep-sequencing libraries (**Fig. 2b**). Specifically, we started from the smoothed T-to-C transition frequencies as computed by PARalyzer, i.e., the signal track including normalization as described above. To represent a binding signal, we summed over frequencies in a local area (±3 nt) around a consistent reference location: in this case, the mRNA coordinate across from miRNA position 8, i.e., the beginning of the longest possible seed match. As a test statistic, we used the fraction of the binding signal detected in the deletion line compared to the signal in the wild type. Significance was assessed by the Wilcoxon rank-sum test.

Quantification of mRNA level changes. Paired-end strand-specific RNA-seq was performed for all four established LCLs (GEO GSE46611). Total RNA was collected with TRIzol (Life Technologies), and libraries were prepared using ScriptSeq v2 kit and Ribo-Zero (Epicentre). Biological replicates were performed for LCL-BAC-D1. RNA-seq data were aligned and quantified using RSEM with reference genome and transcript annotation[22]. Differential analysis was performed using EBseq within RSEM[23]. Real fold change (RealFC value from EBseq) for genes with median $\log_2(\text{TPM} + 1)$ of >5 across all samples were used for evaluating differential expression of microMUMMIE or MIRZA predictions; the same sets of genes were used to assess loss of binding events in the deletion libraries. See **Supplementary Table 3** for specific analysis parameters and quality assessments.

For assessing the enrichment of predicted targets of a specific miRNA among the top differentially regulated gene sets, genes were ranked by mRNA expression fold change (on the basis of the data for LCL-BAC vs. the corresponding deletion line) and grouped into sets of increasing size. For each top differentially regulated set, an enrichment of predicted targets was computed as a ratio of a fraction of predicted targets in the top set and a fraction of total number of predictions in the full gene set. Target sets of miRNAs with a similar number of microMUMMIE predictions (between roughly half to twice as many sites) were used to quantify background expression changes of presumably unaffected targets.

Luciferase assay. 293T cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS (FBS) and antibiotics. miRNA expression plasmids contain ~200 nt of the primary miRNA cloned from genomic DNA into pLCE at

XhoI/XbaI. Expression of the miRNA was confirmed using indicator assays as previously described[12]. Luciferase reporter assays were carried out as previously described[12]. Briefly, 293T cells were cotransfected in 24-well plates with 1 μg miRNA expression vector (pLCE-miRNA), 12 ng pL-CMV-GL3-3′ UTR and 12 ng pL-CMV-RLUC (*Renilla* internal control) using FuGENE 6 (Promega) according to manufacturer's instructions. Lysates were harvested 48–72 h post-transfection and assayed for luciferase activity using the dual luciferase reporter assay kit (Promega).

Modeling software. Binding-site models were implemented within the open-source software MUMMIE (multivariate Markov modeling inference engine). This system permits user-specified model topologies to be designed for a wide array of potential bioinformatic applications. The software includes command-line routines for both parameter estimation ("training") and decoding ("prediction"). The following sections describe the models and algorithms used to obtain predictions for this study.

Models. Prediction of target sites from binding and sequence data is achieved via hybrid continuous-discrete multivariate hidden Markov models. A Markov model is a probabilistic generative model for sequential data with a fixed set of discrete states $Q = \{q_0, q_1, \ldots, q_{N-1}\}$. We here consider discrete-time models, in which observations are available for $T = \{0, 1, \ldots, L-1\}$ positions in a biological sequence $S = s_0, s_1, \ldots, s_{L-1}$, which the model is assumed to have emitted. Emitted sequence elements may take continuous values; for multivariate models, each sequence element is a vector of continuous values (each emitted by a separate "emission channel"). In hybrid continuous-discrete models, each emission channel is either discrete or continuous. Thus, for each vector in the emission sequence, each component of that vector will be a value that is either a symbol drawn from a discrete alphabet or a numerical value. For example, an RNA emission channel would accept symbols from the set {A,C,G,U}, an "aligned read-count" channel would accept non-negative integers denoting numbers of reads aligned at a given genomic position, and a "probability of being unpaired" track would accept real values between 0 and 1.

The model M emits a biological sequence S as follows. Beginning in a special silent state q_0, at each time step t the model chooses a state to transition into by drawing a new state $x_{t+1} \in Q$ from a transition probability distribution, $P(x|x_t)$, which is conditional on the current state, x_t. Upon entering the new state, it draws a random vector $\mathbf{y} \in V$ from the emission distribution $P(\mathbf{y}|x_{t+1})$, where V is the set of all possible emission vectors for the model. After emitting the vector's components into the respective emission channels, the model chooses the next transition according to $P(x|x_t)$. If state q_0 is chosen, the machine terminates and S is taken to be the complete emission; otherwise, it continues its run by alternating between transitions and emissions until state q_0 is eventually chosen. Note that q_0 never makes emissions.

Because the emission distribution is conditional only on the current state, each emission is conditionally independent of all other emissions, given the associated state. This is the case of 0th-order emissions. A model with higher-order emissions may condition each emission on events (transitions or emissions) progressively further into the past. For the present work we consider only higher-order discrete emissions, in which the emission of a symbol may be conditioned on some number of immediately preceding symbols emitted into the same channel; this permits us to model seed matches for many different miRNAs while using a relatively small number of states (see below).

For miRNA target prediction, we designed a 47-state Markov model. The model is a hybrid continuous-discrete multivariate hidden Markov model with three emission channels. The first channel is the PARalyzer signal representing the relative abundance of T-to-C transitions in aligned CLIP scaffolds. This channel is continuous and modeled via a four-component Gaussian mixture. The second channel, branch-length score (BLS), provides a continuous measure of conservation; it was computed using a script from the TargetScan package. The third channel comprises the genomic DNA sequence corresponding to the mRNA and is modeled using a 7th-order Markov chain. Channels are prealigned so that for each position in a UTR we have 1 nt and two continuous values.

The model topology is depicted schematically in **Figure 1d**. State 1 models the background regions between PAR-CLIP groups. State 3 models the peak of the PAR-CLIP group, and states 2, 4, 5 and 6 model the flanking regions around the peak. State 5 is a metastate[24] that expands into the 41-state submodel shown in **Figure 1e**; this metastate models the actual target site of the miRNA seed. Our default model detects seven types of miRNA binding: 6mer3-8, 6mer2-7, 7mer-m8, 7mer-m1, 7mer-A1, 8mer-A1 and 8mer-m1.

The model in **Figure 1d** preferentially visits state 5 (the miRNA binding site) a short distance 3′ of a PAR-CLIP peak (which is generally matched by state 3); this preference reflects the strong enrichment of miRNA seed matches 3′ of peaks (**Fig. 1b**). In consequence, model predictions also show a strong preference for this location relative to the peak (**Supplementary Fig. 1**).

Figure 1b suggests a small fraction of sites that are located 5′ of the peak, and the model predicts a fraction of these 5′ sites. This illustrates the flexibility of this class of models: when the most promising site occurs 5′ of the peak, the decoding algorithm will permit state 3 to be visited earlier in the sequence so that state 5 can match the putative target site. Likewise, when no promising seed match is present, the decoding algorithm may choose to not predict any binding site for the PAR-CLIP group. This flexibility is a function of the model parameters—in particular, the variance of the emission distribution for the PAR-CLIP signal: for low variances (~0.01) the model predicts many sites (typically one per PAR-CLIP group), whereas for high variances (~1.0) it predicts fewer sites. The different points in **Figure 1f,g** correspond to different settings of the PARalyzer signal variance, with a high variance parameter corresponding to high specificity and high SNR and a low variance parameter corresponding to high sensitivity. (The following variances were used for these plots: 1.5, 1.25, 0.75, 0.5, 0.375, 0.25, 0.20, 0.15, 0.1, 0.075, 0.05, 0.01 and 0.005.)

To model bulge sites with bulges at specific positions complementary to the seed, we used the same metamodel structure shown in **Figure 1d**, but simplified the submodel for state 5 to include only a single linear chain of seven states to match the six seed residues plus the bulge position embedded within the seed.

Inference algorithms. Prediction of sites can be carried out in MUMMIE using Viterbi decoding, posterior decoding or a combination of the two.

doi:10.1038/nmeth.2489

Viterbi decoding identifies the single most probable path through the HMM states for emitting the multivariate sequence. Any occurrence of metastate 5 in that most probable path is interpreted as a predicted miRNA binding site; the exact sequence of states within the metastate dictates the type of binding (for example, 6mer2-7, 7mer-m1, etc.), and the matching nucleotides in the emitted sequence identify the miRNA family (seed sequences may or may not be unique to a given miRNA within a family; when multiple miRNAs share a predicted seed, all are listed as the potential binding miRNA).

In contrast to Viterbi decoding, posterior decoding uses the well-known forward and backward algorithms[24] to compute the posterior probability $P(q,k)$ that state q was active at position k when the sequence was generated by the HMM. In addition to providing this standard functionality, microMUMMIE also computes for each 8-bp interval the posterior probability that a miRNA binds within that 8-bp interval, via any of the modeled binding modes (i.e., by summing the posteriors for each binding mode for that miRNA within that 8-bp interval).

Finally, MUMMIE supports a combination of posterior and Viterbi decoding via the posterior Viterbi algorithm[25], in which emission and transition probabilities in the Viterbi algorithm are replaced with posterior probabilities. This decoding algorithm was used for the results presented here.

Model training. Considering all states modeling the actual miRNA binding site to be foreground states and all others to be background states, training of the background states was accomplished by applying expectation maximization (EM) to entire UTRs, with the following exceptions. For the peak state (state 3), the PARalyzer signal mean was set to 1.0 and its variance was arbitrarily set to 0.01; for the flanking states (states 2, 4 and 6) the signal mean was set to 0.5 and the variance to 0.01. These mean values were chosen to have these states match the respective regions of an idealized PARalyzer signal profile at a binding site. The mean and variance of the conservation track were set to 0 and 0.01, respectively, for all background states.

For the foreground states, the nucleotide emission distribution was set so as to permit only the training miRNA seeds to be emitted; PARalyzer signal mean was set to 0.5 and variance to 0.01 as in the background flank states; conservation mean was set to 3.0. Nucleotide sequence emissions for flanking states were trained via EM on sequences within PARalyzer groups, and for the peak state they were trained on the collection of single nucleotides at each PARalyzer peak (**Supplementary Table 4**). Variance of the conservation track was trained on a held-out set of training data by observing the resulting sensitivity-SNR curves at different settings and selecting the value that produced the curve with the highest overall SNR. All covariances were set to 0. The Markov order of nucleotide emissions varied from 0 to 7 along the length of a seed, so as to enforce that only valid miRNA seeds may be emitted, as per the training set.

Bulge-site analyses. We investigated different locations for the pivot nucleotide in bulge-site pairing. A bulge seed match is a perfect seed match to miRNA positions 2–7 with a bulge nucleotide insertion that can pair with the pivot nucleotide. Similarly to the canonical perfect seed-match prediction, we computed SNR of bulge-site predictions. Here we excluded the bulge seed matches that share a 6-mer with any human or EBV miRNA.

Additionally, for bulge type 2 (positions 4 and 5) evaluation and prediction, we excluded bulge seed matches to original miRNAs with the same nucleotide at positions 5 and 6, as its pairing in the nucleation step could extend one more position and thus provide for the same bulge seed match as in type 1 (positions 5 and 6). Similarly, seeds for bulge type 3 (positions 3 and 4) did not include bulge seeds that extended to bulge type 2.

Prediction accuracy evaluation and comparison with other predictors. We computed SNR ratios by comparing numbers of nonshuffled and shuffled sites among a set of predictions. Shuffled miRNAs were created via a random process that preserves the dinucleotide frequencies observed in real miRNAs, as described previously[26]. These were then screened to ensure that their expected seed-match frequencies in PAR-CLIP groups did not differ by more than 15% from the same statistic for the real (nonshuffled) miRNAs. The SNR was averaged over 1,000 runs in which we sampled one shuffled miRNA for each original (nonshuffled) miRNA; state 5 of the model was then trained on the resulting set of shuffled and nonshuffled seeds so that a single decoding run of the model could predict any of the seeds in the training set.

We used TargetScan release 6.1 (ref. 10) to compute TargetScan predictions, "context+" scores and branch-length scores. A 23-way alignment of 3′ UTRs was used, and miRNAs were evaluated individually. We computed SNR at varying context+ score thresholds for the TargetScan sensitivity-SNR plots.

MIRZA[11] was executed in its default "noupdate" mode and optionally provided with cell type–specific expression levels of miRNAs. Mirroring the example of its developers, inputs to MIRZA were 51-nt sequences centered on the peak T-to-C conversion site as identified by PARalyzer as well as 21-nt mature miRNA sequences. MIRZA's "target frequency" score was used to rank predictions when computing sensitivity, specificity and SNR.

20. Feederle, R. *et al. J. Virol.* **85**, 9801–9810 (2011).
21. Flicek, P. *et al. Nucleic Acids Res.* **38** (suppl. 1), D557–D562 (2010).
22. Li, B. & Dewey, C.N. *BMC Bioinformatics* **12**, 323 (2011).
23. Leng, N. *et al. Bioinformatics* **29**, 1035–1043 (2013).
24. Majoros, W.H. *Methods for Computational Gene Prediction* (Cambridge Univ. Press, 2007).
25. Fariselli, P., Martelli, P.L. & Casadio, R. *BMC Bioinformatics* **6** (suppl. 4), S12 (2005).
26. Lekprasert, P., Mayhew, M. & Ohler, U. *PLoS ONE* **6**, e20622 (2011).