BMC Bioinformatics

RESEARCH

Open Access

A statistical approach for identifying primary substrates of ZSWIM8-mediated microRNA degradation in small-RNA sequencing data

Peter Y. Wang[1,2,3] and David P. Bartel[1,2,3*]

*Correspondence:
dbartel@wi.mit.edu

[1] Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, MA 02142, USA
[2] Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[3] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

Background: One strategy for identifying targets of a regulatory factor is to perturb the factor and use high-throughput RNA sequencing to examine the consequences. However, distinguishing direct targets from secondary effects and experimental noise can be challenging when confounding signal is present in the background at varying levels.

Results: Here, we present a statistical modeling strategy to identify microRNAs that are primary substrates of target-directed miRNA degradation (TDMD) mediated by ZSWIM8. This method uses a bi-beta-uniform mixture (BBUM) model to separate primary from background signal components, leveraging the expectation that primary signal is restricted to upregulation and not downregulation upon loss of ZSWIM8. The BBUM model strategy retained the apparent sensitivity and specificity of the previous ad hoc approach but was more robust against outliers, achieved a more consistent stringency, and could be performed using a single cutoff of false discovery rate (FDR).

Conclusions: We developed the BBUM model, a robust statistical modeling strategy to account for background secondary signal in differential expression data. It performed well for identifying primary substrates of TDMD and should be useful for other applications in which the primary regulatory targets are only upregulated or only downregulated. The BBUM model, FDR-correction algorithm, and significance-testing methods are available as an R package at https://github.com/wyppeter/bbum.

Keywords: Differential expression, microRNA, Mixture models, Small-RNA sequencing, RNA-seq

Introduction

Differential expression (DE) analyses seek to identify gene products that change in abundance after either altering a condition or perturbing a regulatory factor. Aiding in these analyses are statistical pipelines, such as DESeq2 [1], edgeR [2], and limma [3], which compare RNA-seq or microarray datasets to identify RNAs with significantly changed levels, after correcting for false discovery rate (FDR) due to multiple

testing using the Benjamini–Hochberg procedure [4]. These pipelines have been invaluable for DE analyses of mRNAs as well as noncoding RNAs.

Noncoding RNAs often subject to DE analyses include the microRNAs (miRNAs), which are small RNAs that direct widespread post-transcriptional repression of metazoan mRNAs [5]. To perform this function, miRNAs associate with the effector protein Argonaute (AGO) to form a complex in which the miRNA specifies which targets are repressed, primarily through base pairing between the seed of the miRNA (miRNA nucleotides 2–7) and a site in the target mRNA [6]. Meanwhile, AGO causes repression, typically by recruiting the cytoplasmic mRNA deadenylation machinery [7].

Most miRNAs are quite stable, with half-lives much greater than those of typical mRNAs, presumably a consequence of their association with AGO, which protects miRNAs from cellular nucleases [8, 9]. However, some miRNAs are more rapidly degraded, and in cells of both mammals and flies, this more rapid degradation appears to be the result of target-directed miRNA degradation (TDMD) [10]. TDMD is a phenomenon in which targets with unusual complementarity cause a conformational change recognized by the ZSWIM8 E3 ubiquitin ligase, which polyubiquitinates the AGO protein, leading to its degradation by the proteasome, thereby exposing the miRNA to degradation by cellular nucleases [10, 11]. Through DE analysis of small-RNA sequencing (sRNA-seq) data acquired after reducing ZSWIM8 in different cell types, more than 40 candidate miRNA substrates of ZSWIM8 have been identified [10].

When identifying candidate substrates of ZSWIM8, standard DE analysis is not sufficient to distinguish between miRNAs that are significantly upregulated due to the primary effect of losing ZSWIM8-mediated TDMD, and those with significantly perturbed expression due to secondary effects, such as transcriptome changes caused by the dysregulation of miRNAs or other changes that might be caused by the loss of ZSWIM8. To exclude miRNAs changing due to secondary effects, Shi et al. [10] use the knowledge that ZSWIM8 mediates degradation of miRNA substrates, which implies that these substrates should undergo only upregulation upon the loss of ZSWIM8. Accordingly, the significance cutoffs (α values) of FDR-adjusted p values from DESeq2 (p_{adj}) are each adjusted down to the most permissive level that excludes all downregulated miRNAs. As a result, these ad hoc adjustments of α values vary widely, ranging from 0.05 to 10^{-7} for different datasets analyzed (Fig. 1A) [10]. Although this approach seems better than classifying any significantly upregulated miRNA as a ZSWIM8 substrate, it has several shortcomings: (1) it is unduly sensitive to outliers among downregulated miRNAs, which can reduce sensitivity; (2) FDRs are inconsistent among experiments and cannot be predetermined; and (3) the FDR and specificity of each analysis depend on the sample size.

Here, we developed a statistical modeling strategy that accounts for varying secondary effects in RNA-seq datasets, thereby enabling primary substrates to be identified at a consistent predetermined statistical stringency. Compared to the previous approach, this strategy achieved more robust results with fewer shortcomings. This strategy should also provide an improved strategy for identifying direct targets of other types of regulatory pathways.

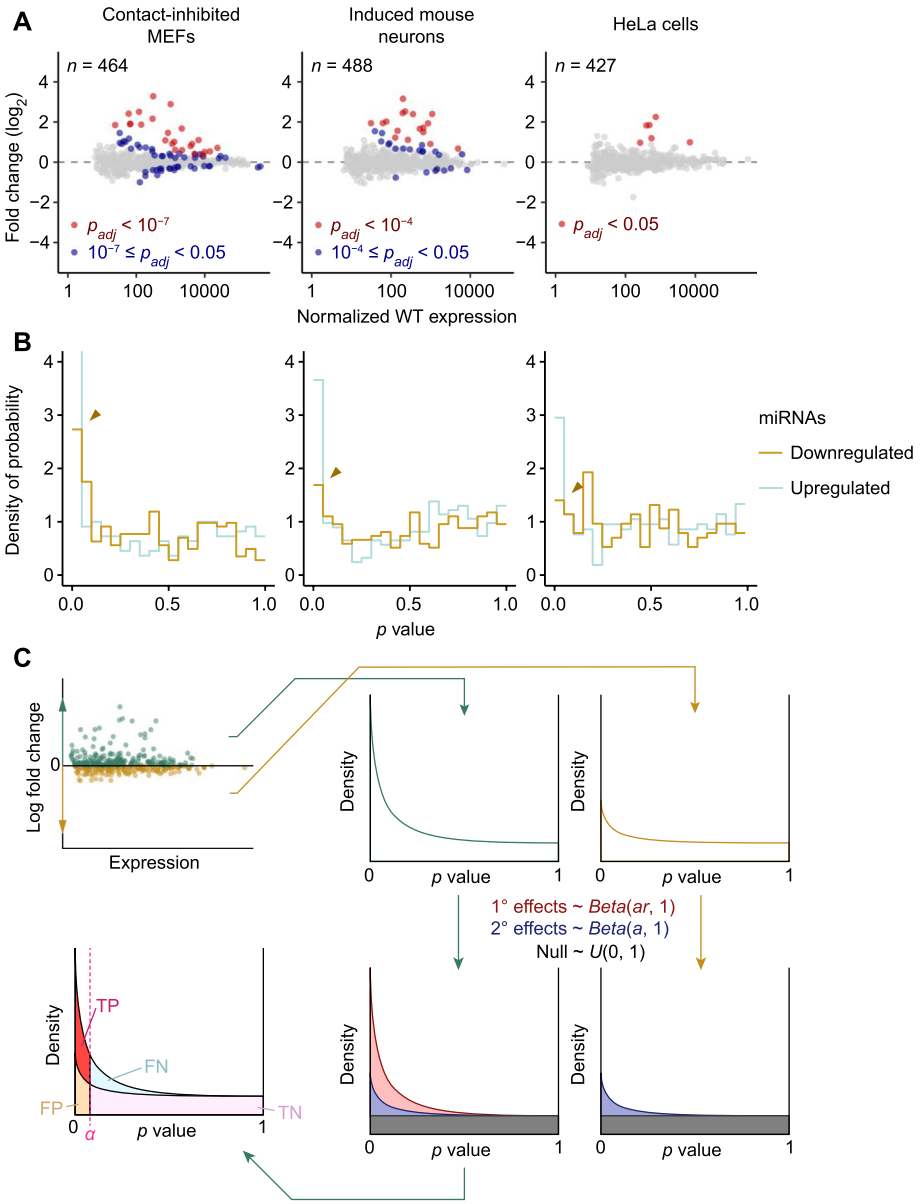

Fig. 1 Statistical modeling of secondary effects in DE datasets. **A** Representative plots of fold changes in miRNA levels observed upon *ZSWIM8* knockout, as measured by small-RNA sequencing. Points for miRNAs with significant upregulation based on the cutoffs defined by Shi et al. [10] are colored in red. Points for miRNAs that did not meet the adjusted cutoffs but would be significant based on a common cutoff of 0.05 are colored in blue. The number of miRNAs and passenger strands quantified in each dataset is indicated. **B** Histograms of raw *p* values for upregulated and downregulated miRNAs and passenger strands analyzed in datasets of **A**. The peak near $p = 0$, which indicates true signal above null, is indicated by an orange caret for downregulated miRNAs in each dataset. **C** BBUM modeling of *p* values from DE datasets. *p* values corresponding to upregulated and downregulated miRNAs are fit in parallel to the BBUM model in which downregulated points are fit to a distribution missing the beta component for primary effects. Based on the fitted model, expected FDR and other statistics can be calculated for any cutoff α (*TP* True positives; *FN* False negatives; *FP* False positives; *TN* True negatives)

Results

Significant signal among downregulated miRNAs implies secondary effects

In published datasets, a standard p_{adj} cutoff at $\alpha = 0.05$ was suitable for identifying primary substrates in some contexts, such as HeLa cells [10]. However, in other contexts, such as contact-inhibited mouse embryonic fibroblasts (MEFs) or induced mouse neurons (iMNs), the same cutoff would have classified many miRNAs with changes that seemed no different than background as primary substrates (Fig. 1A). These differences observed between datasets, which motivated the use of varying stringency of α values, were presumably due to varying levels of secondary effects. To evaluate this idea, we examined the downregulated miRNAs, reasoning that because loss of ZSWIM8 should only cause upregulation of primary miRNA substrates, any signal among downregulated miRNAs in excess of that expected by chance implied the existence of true secondary effects. To search for this signal, we examined *p*-value distributions. A set of data points drawn from a null distribution is expected to have a uniform distribution of *p* values, ranging from 0 to 1, as the cumulative fraction of points called as false positives should equal to α for all values of α. Thus, a significant signal should manifest as a peak of enriched density near $p = 0$ [12]. For each of the three datasets of Fig. 1A, the distribution of raw *p* values from DESeq2 was examined, looking separately at the results for upregulated and downregulated miRNAs. As expected for datasets that included ZSWIM8 substrates, distributions for upregulated miRNAs peaked near $p = 0$ (Fig. 1B). In addition, for the two datasets that required a more stringent α value, the distributions for downregulated miRNAs also peaked near $p = 0$, albeit at a level lower than that observed for upregulated miRNAs (Fig. 1A, B). These results supported the idea that some miRNAs were truly downregulated upon the loss of ZSWIM8, likely as a result of secondary effects, and the idea that contexts with stronger secondary effects required stronger adjustments of stringency.

Statistical modeling of *p* values enables the separation of primary and secondary effects

If secondary effects acted symmetrically, causing miRNAs to increase as well as decrease (at similar frequency and similar magnitudes), then the excess in the peak near $p = 0$ observed for upregulated miRNAs compared to that observed for downregulated miRNAs should correspond to the density of true primary substrates of ZSWIM8 (Fig. 1C). Accordingly, we developed a statistical strategy to separate the components of the *p*-value distribution to better classify the primary substrates and the secondary effects. Our strategy made three assumptions: (1) primary effects were stronger than secondary effects, (2) secondary effects were approximately symmetrical between upregulated and downregulated data points, and (3) primary effects caused upregulation and never downregulation. Previous studies have described the successful use of a beta-uniform mixture (BUM) distribution model and its variants to model *p*-value distributions between 0 and 1 [13–15]. In these studies, the uniform component represents the distribution of null data points, while the beta component represents the characteristic peak near $p = 0$. Building upon these concepts, we implemented a modified mixture distribution model, which we call the bi-beta-uniform mixture (BBUM) model, to describe our *p* values. The BBUM distribution contains two beta components [$Beta(ar, 1)$ and $Beta(a,$

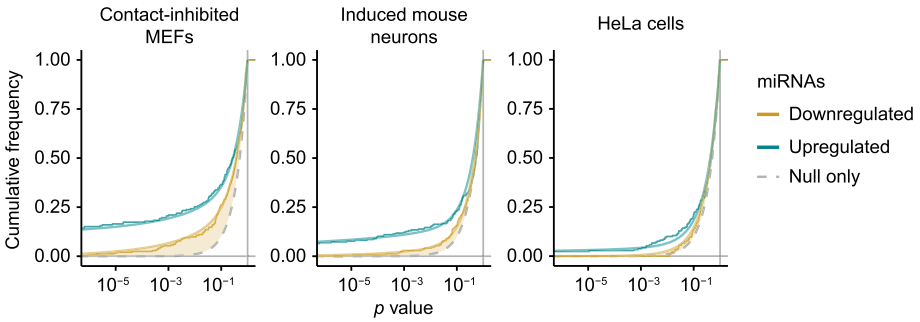

Fig. 2 Fit of empirical *p* values to the BBUM model. Plotted for each dataset are empirical cumulative distributions of *p* values for upregulated miRNAs (turquoise) and downregulated miRNAs (orange), with the respective cumulative distributions of the fitted BBUM model overlaid as smooth lines. The uniform null distribution is shown as a gray dashed line. Deviation between the distribution for downregulated miRNAs and the null distribution is shaded in light orange; a greater deviation at the lower tail indicates a greater excess in density near $p = 0$, which corresponds to a more substantial contribution of the beta component for secondary effects in the BBUM model

1)], instead of one, which respectively correspond to the primary and secondary effects, followed by a similar uniform component [$U(0, 1)$] for the null distribution (Fig. 1C). The *p* values were fit to this mixture model, with the downregulated data points fit to a model that lacked the primary-effect component (Fig. 1C).

The model faithfully captured the distributions of *p* values from both halves of each dataset, especially for the *p*-value density near 0 (Fig. 2). As expected, datasets that required more stringent α values for p_{adj} (Fig. 1A) were modeled with more pronounced beta components for secondary effects, as indicated by the greater deviation between the null distribution and the model fit for downregulated points (Fig. 2, orange shading). These results indicated that for these datasets the model was able to separate primary and secondary effects.

Significance testing after BBUM adjustment predicts direct substrates of TDMD

Because the model was able to represent the distribution density that corresponded specifically to the primary effects, the expected FDR could be calculated at any desired cutoff among the *p* values considered, which we defined as the BBUM-FDR-adjusted *p* value (p_{BBUM}). Choosing a p_{BBUM} cutoff of 0.05, we reanalyzed the datasets from mammalian and fly cell lines from Shi et al. [10] to identify ZSWIM8 primary substrates. Across all nine datasets examined, the proposed primary substrates identified using the BBUM strategy largely matched those identified previously, while imposing a consistent, predetermined FDR cutoff (Fig. 3A). Out of the 75 instances classified as significant at this cutoff, four were newly classified as significant. Three of these four involved miRNAs that were either also proposed to be ZSWIM8 substrates in other contexts or related to another proposed ZSWIM8 substrate, which supported the idea that these four miRNAs included true ZSWIM8 substrates (Fig. 3A, Additional file 1: Table S1). This idea was further reinforced by analysis of the passenger strands of these candidate miRNAs. During miRNA biogenesis, the pre-miRNA hairpin is processed into a miRNA duplex containing the miRNA paired to its passenger strand. When this duplex associates with AGO, the miRNA strand is retained,

and the passenger strand is ejected and rapidly degraded. Because TDMD acts upon mature AGO–miRNA complexes, the miRNA strand and not the passenger strand should be affected by the loss of ZSWIM8 [10]. Indeed, each of the four newly significant miRNAs were upregulated upon the loss of ZSWIM8 without significant change in their passenger strands, as observed for other miRNAs predicted to be ZSWIM8 substrates (Fig. 3B).

Three other edge cases proposed to be ZSWIM8 substrates by the previous method were not identified at $p_{BBUM} < 0.05$ when using the BBUM model (Fig. 3A, Additional file 1: Table S1). One of these was miR-7-5p in clonal ZSWIM8 knockout cells. This known TDMD substrate [16] was not sufficiently upregulated in knockout K562 cells to reach significance over the relatively high background variability of this dataset (Fig. 3A). Nonetheless, statistical significance was readily achieved for miR-7-5p in K562 cells when using datasets in which ZSWIM8 was knocked down using CRISPRi, which led to much lower background variability than observed with clonal knockout cells (Fig. 3A). Likewise, in a dataset analyzing a different ZSWIM8-knockout line (Han et al., 2020), miR-7-5p upregulation met the significance threshold using the BBUM approach (Fig. 3C).

Thus, on the whole, using the BBUM model, candidate primary substrates of ZSWIM8 were identified while implementing a consistent and predetermined FDR confidence value across all cellular contexts examined, without noticeably sacrificing the apparent sensitivity or specificity of the previous approach. We attribute this success to the ability of the BBUM model to adjust the varying spread of background *p* values to a consistent range (Fig. 3D).

BBUM correction applies a consistent statistical stringency

To benchmark the performance of our approach, we randomly generated simulated datasets of *p* values containing varying levels of primary and background secondary signal under the BBUM distribution. We compared the empirical FDR of our BBUM strategy, using $p_{BBUM} < 0.05$ as the significance cutoff, against that of the method used previously by Shi et al. [10]. The BBUM approach had a mean FDR of 0.050 ± 0.002 (95% confidence interval (CI)) (Fig. 4A). The Shi et al. method produced a comparable mean FDR

(See figure on next page.)

Fig. 3 Identification of candidate ZSWIM8-sensitive miRNAs using the BBUM method. **A** Plots of $\log_2$ fold changes in miRNA levels observed upon loss of ZSWIM8 in all mammalian and fly cell lines examined in Shi et al. [10]. Two datasets were derived from CRISPRi knockdown (KD) of *ZSWIM8* in K562 cells, using one of two different guide RNAs (A or B). All other datasets were derived from knockout (KO) of either *ZSWIM8* or *Dora*, the Drosophila *ZSWIM8* ortholog. Points for miRNAs that met the common p_{BBUM} significance cutoff of 0.05 are in red, with *n* indicating the number passing this cutoff, shown as fraction of the total number of miRNAs and passenger strands quantified. Points for passenger strands of these miRNAs are in cyan, if the passenger strands were both annotated and observed above the expression threshold. Points for miRNAs with classifications differing from that of previous work [10] are outlined in black ($n = 7$). Points for miR-7-5p, a known TDMD substrate [16], are labeled. **B** Different effects of the loss of ZSWIM8 on miRNAs and their passenger strands. Fold changes in miRNA levels are shown for two datasets with newly significant miRNAs. Only miRNAs found significant by the BBUM method and whose passenger strands were quantified in the dataset are shown, with each miRNA paired with its passenger strand(s). Points for newly significant miRNAs are outlined and labeled. **C** Plot of fold changes in miRNA levels observed upon knockout of *ZSWIM8* in K562 cells by Han et al. [11]. Points for two downregulated outliers (hsa-miR-221-3p, hsa-miR-222-3p) are shown in magenta. Otherwise this panel is as in **A**. **D** Volcano plots of raw and BBUM-FDR-adjusted *p* values. Colors are as in **A**. The significance cutoff at 0.05 is shown as orange dashed lines for p_{BBUM}

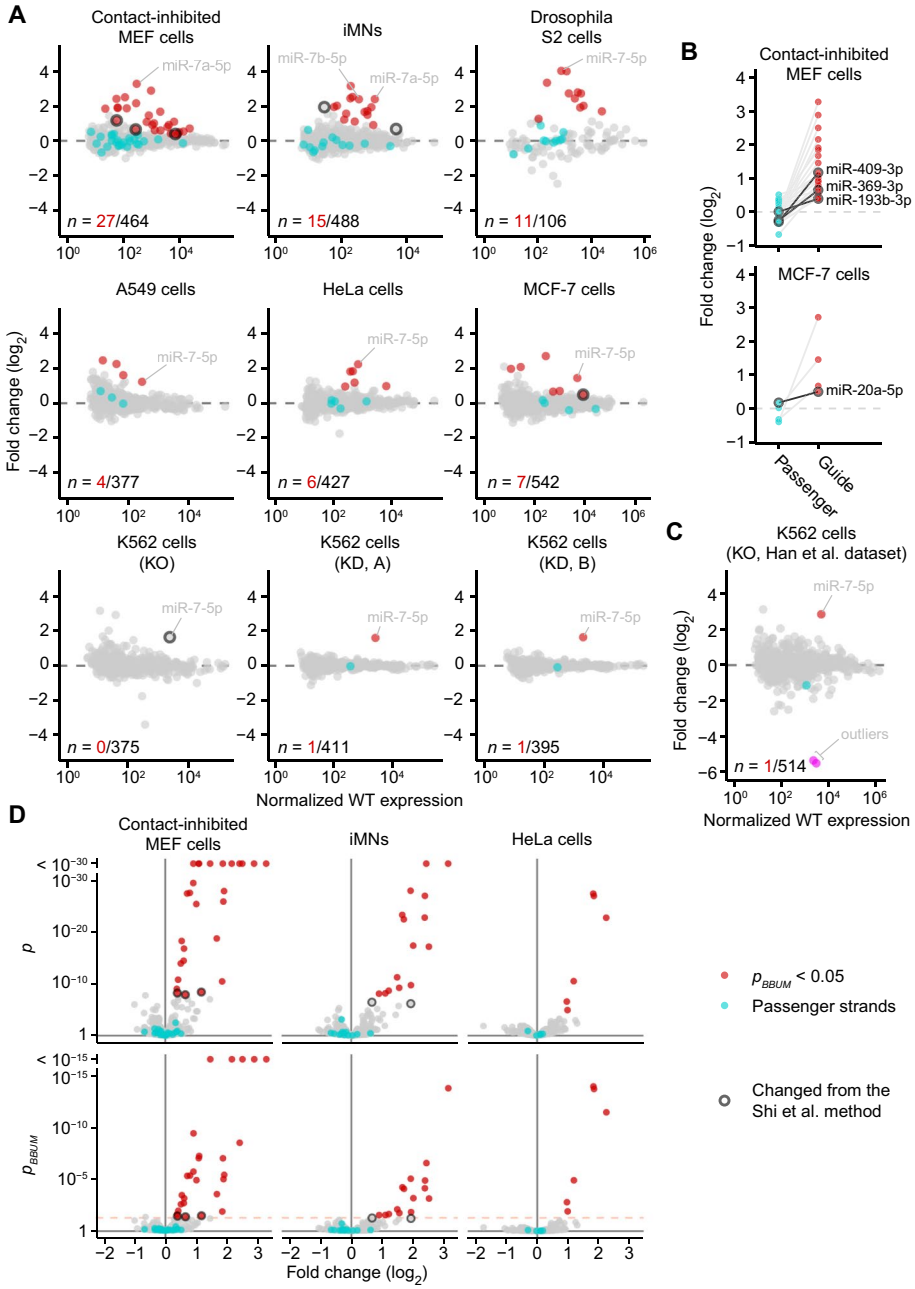

Fig. 3 (See legend on previous page.)

of 0.036 ± 0.002 but was less constant, as measured by the coefficient of variation (CV) of the FDR, which was 2.47 ± 0.11 for the previous method, compared to 1.55 ± 0.08 for the BBUM method. Thus, BBUM correction produced an accurate and consistent FDR when evaluated using simulated datasets.

BUM models are reported to be sensitive to outliers with extremely small *p* values due to the asymptotic behavior at zero of the likelihood function of the type of beta distribution used [14]. Indeed, we found that adding artificial extreme outliers to downregulated miRNAs in either empirical or simulated datasets could cause the

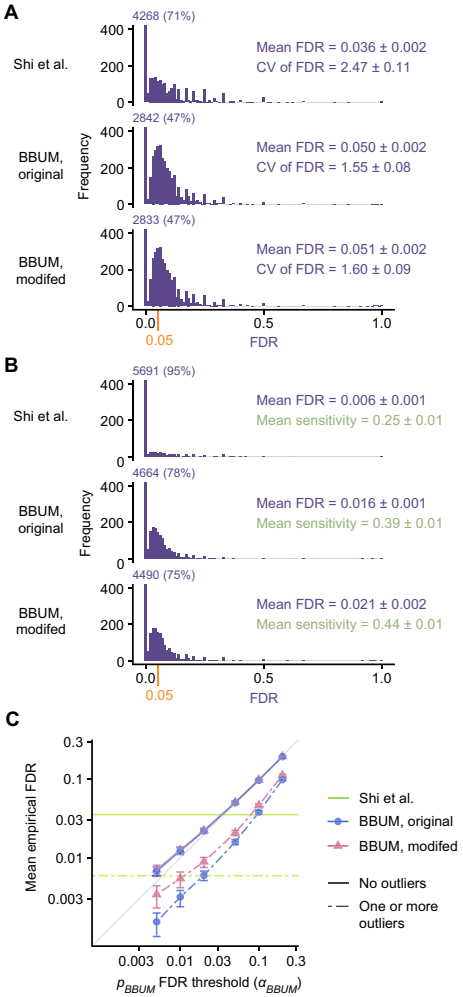

Fig. 4 Benchmarking of correction procedures and significance-testing methods. **A** Histograms of the empirical FDR for 6000 simulated *p* value datasets with no outliers. Results are plotted for the extreme-value method used by Shi et al., the original BBUM procedure without outlier trimming at $p_{BBUM} < 0.05$, and the modified BBUM procedure with automatic outlier trimming at $p_{BBUM} < 0.05$. The frequency of the density near 0, which exceeded the *y*-axis scale, is indicated alongside its fraction among all trials for each histogram. The FDR cutoff of the BBUM methods at 0.05 is indicated on the x-axis in orange. The mean FDR of all simulation trials is indicated on the right with corresponding 95% CIs by bootstrapping. The CV of FDR is similarly indicated. **B** As in **A**, but for 6000 simulations with at least one programmed outlier. The mean FDR and mean sensitivities are indicated with corresponding 95% CIs by bootstrapping. **C** The mean empirical FDR of significance-testing methods at different p_{BBUM} significance thresholds. Results are plotted for all six threshold values tested. If the empirical FDR matched the predetermined FDR cutoff, the point would lie on the gray diagonal line. Error bars indicate 95% CIs by bootstrapping. Because the Shi et al. method does not allow quantitative adjustments of significance thresholds, results are shown as green horizontal lines with the mean empirical FDR values as the *y*-intercepts

BBUM procedure to overcorrect for secondary effects. The previous approach by Shi et al. [10] was even more prone to overcorrection, with a single extreme outlier able to cause all of the upregulated miRNAs to be designated as background. The influence of outliers was also illustrated in the Han et al. [11] dataset for K562 cells. Two extreme outliers within this dataset prevented any miRNAs to be designated as

primary ZSWIM8 substrates when using the approach of Shi et al. [10] and severely weakened the significance of miR-7-5p when using the BBUM model (Fig. 3C).

To mitigate the effects of outliers, we developed a conservative outlier detection method that used the fitted r parameter of the BBUM model to identify and trim probable outliers from downregulated data points. The performance of the modified BBUM procedure was similarly benchmarked using simulated p-value datasets, with and without added outliers. In datasets without added outliers, the modified BBUM procedure did not have significantly increased mean FDR (0.051 ± 0.002) or CV of FDR (1.60 ± 0.09) when compared to the original BBUM procedure (Fig. 4B). In datasets with added outliers, the modified BBUM procedure was somewhat improved over the unmodified procedure, with mean FDR increasing from 0.016 ± 0.001 for the original BBUM procedure to 0.021 ± 0.002 for the modified procedure and mean sensitivity increasing from 0.39 ± 0.01 to 0.44 ± 0.01 (Fig. 4B). Importantly, both the original and the modified BBUM procedures were less sensitive to outliers than the previous method of Shi et al. [10], which had a mean FDR of 0.006 ± 0.001 and mean sensitivity of 0.25 ± 0.01 in the presence of one or more outliers (Fig. 4B). Moreover, the modified BBUM procedure successfully identified and trimmed the two outliers in the K562 dataset (Fig. 3C), as well as any artificial outliers we added to other empirical datasets.

The ad hoc method by Shi et al. [10] provides a fixed stringency for each dataset. In contrast, the BBUM method allows the stringency to be quantitatively tuned by choosing the significance cutoff for p_{BBUM} to suit the needs of the experiment or hypothesis at hand. Therefore, we extended our benchmarking analysis to a range of possible significance cutoffs for p_{BBUM}, and assessed the accuracy of the BBUM method at each cutoff α_{BBUM} in our simulations. When no outliers were present, both the original and the modified BBUM methods faithfully achieved results at the predetermined FDR at all α_{BBUM} values tested (Fig. 4C). When one or more outliers were present, the modified BBUM method partially mitigated the overcorrection of secondary effects by the original BBUM method at all α_{BBUM} values tested, especially when α_{BBUM} was small, which was the condition in which overcorrection was most severe in the original BBUM method (Fig. 4C). In fact, the overcorrection of the modified BBUM procedure, tested across a wide range of α_{BBUM} values, was no worse than that seen at $\alpha_{BBUM} = 0.05$, where the mean empirical FDR was 0.021 ± 0.002 (Fig. 4B, C). In no cases did this conservative outlier detection method adjust the FDR to above the intended stringency (Fig. 4C). Hence, the BBUM procedure performed robustly in both ideal and non-ideal scenarios and performed significance testing with improved consistency and flexibility, as well as lower sensitivity to outliers.

BBUM correction can be applied to other RNA-seq datasets

Encouraged by the success of our approach for analyzing sRNA-seq datasets examining the effects of ZSWIM8 knockout/knockdown on miRNA levels, we applied the approach to another type of experimental dataset. MicroRNAs invariably reduce the levels of their primary regulatory targets [5, 17]. Thus, RNA-seq datasets examining the changes in mRNAs observed upon introducing a miRNA seemed appropriate for BBUM

analysis, in that the primary effects were expected to be in one direction, whereas the secondary effects and background variability were expected to be symmetrically distributed between both up- and downregulation. To test whether BBUM might help identify mRNAs most likely to be directly targeted, we analyzed 29 RNA-seq datasets in which a miRNA was transfected into either HeLa or HEK293FT cell lines [18]. As observed for the sRNA-seq results, a standard DESeq2 analysis using $\alpha = 0.05$ as the cutoff for p_{adj} classified many RNAs as differentially expressed following miRNA transfection—some downregulated but others upregulated (Fig. 5A). For most datasets, BBUM FDR correction robustly identified mRNAs that were the most significantly downregulated as top candidates for the primary targets of these miRNAs (Fig. 5A, B). In two out of the 29 datasets (miR-139 and miR-144 in HeLa cells), the BBUM approach failed to identify any direct-target candidates because no data points passed our false-discovery cutoff of 0.05 for p_{BBUM} (Fig. 5; Additional file 1: Fig. S1). Across the remaining 27 datasets, the proportion of primary and secondary effects appeared to span a wide range, and the BBUM analysis helped to characterize these differences. For some datasets (e.g. miR-122 in HEK293FT cells), primary effects appeared to far exceed secondary effects. Indeed, for 11 datasets, no secondary effects were detectable, despite a clear signal for primary target repression (with 19–91 genes classified as significantly repressed). In contrast, three datasets had hundreds of genes called as significantly influenced through secondary effects, far exceeding the number of genes under significant primary effects (Fig. 5; Additional file 1: Fig. S1). Despite the broadly varying levels of secondary effects, the BBUM approach consistently adjusted *p* values to the same level of stringency across the 29 datasets tested (Fig. 5B, C). Thus, the approach can be applied to differential expression datasets with widely varying levels of secondary effects.

Discussion

Our BBUM method will help to more rigorously identify miRNAs subject to TDMD and shows promise for detecting the relative contributions of primary and secondary effects when analyzing miRNA-mediated regulation. We suspect it will also be useful in other analyses in which the primary effect of a regulatory phenomenon causes changes in one direction, whereas secondary effects and background variability cause changes in either direction. Other potential uses include mRNA analyses identifying the targets of transcriptional inhibitors or proteomic analyses identifying the targets of ubiquitin ligases or other degradation phenomena.

We envision some cases in which the use of our BBUM approach would be limited: (1) If there is an overwhelming degree of secondary effects, the signal-to-noise ratio might not be sufficient for the primary and secondary signal to be separated, leading to low sensitivity. (2) If the data are not normalized to spike-ins, and a substantial number of genes significantly change in one direction, the assumption that the background data points are symmetrical about a fold-change of zero might not be fully satisfied. For example, a mild bias towards downregulation was observed for our dataset examining the effect of ZSWIM8/Dora in S2 cells, in which more than 10% of miRNAs were significantly upregulated (Fig. 3A). Nonetheless, our method was empirically robust against this effect.

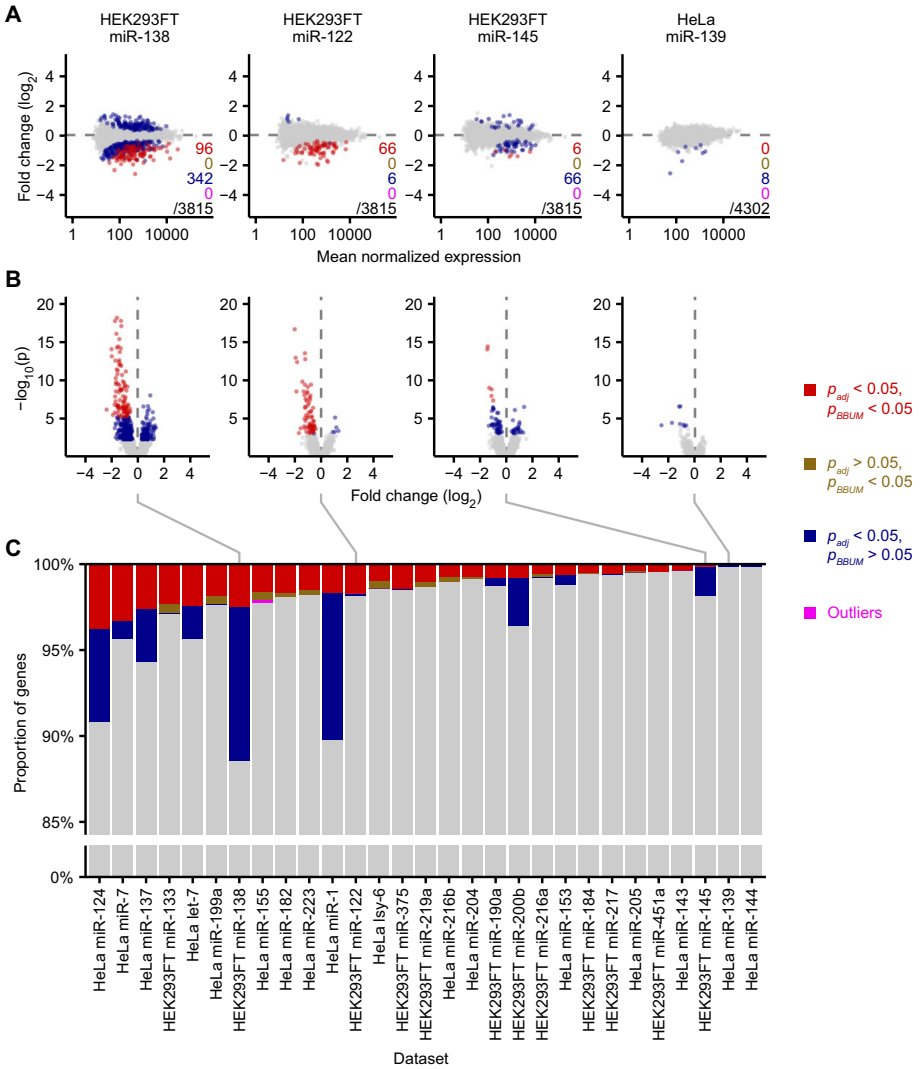

Fig. 5 Application of the BBUM correction method to RNA-seq datasets measuring the effects of transfecting a miRNA into either HeLa or HEK293FT cells. **A** Plots of $\log_2$ fold changes in mRNA levels observed upon transfection of the indicated miRNA duplexes. Four representative datasets out of a total of 29 are shown (Additional file 1: Fig. S1). Points for mRNAs that met both the significance cutoff by the BBUM method and the default Benjamini–Hochberg method (p_{adj}) are colored red. Those that met the cutoff for the BBUM method but not p_{adj} are colored brown. Those that met the cutoff for p_{adj} but not the BBUM method are colored blue; these points roughly correspond to genes subject to significant secondary effects. Outliers identified by the BBUM model are colored in magenta. Number of points belonging to each category and the total number of points plotted are indicated for each plot. **B** Volcano plots of raw *p* values as a function of mRNA $\log_2$ fold change for the same datasets analyzed in **A**. Colors and order are as in **A**. Red data points falling beyond the plotted range are omitted. **C** Proportions of mRNAs in each category for each of the 29 datasets analyzed (Additional file 1: Fig. S1). Colors are as in **A** and **B**. Datasets are arranged in descending order of the number of data points called as significant after BBUM FDR adjustment. Bars corresponding to the four representative datasets shown in **A** and **B** are indicated

Although this work applied our method to datasets downstream of DESeq2, other common differential expression analysis pipelines, such as edgeR and limma, should also be compatible. As the BBUM correction is agnostic to the source of *p* values, we expect that any *p*-value datasets with similar behaviors and constraints should also be

equally applicable, as was the case for our simulated p values. Combining BBUM correction with current statistical procedures can allow the significance testing of more specific hypotheses and a more standardized significance cutoff across different datasets and experiments.

Methods

Analysis of sRNA-seq results

All sRNA-seq datasets analyzed contained three replicates each for the wildtype and the knockdown/knockout conditions, with two exceptions: datasets for wildtype MCF7 cells and ZSWIM8-knockout iMNs had only two replicates. Read counts for the Shi et al. datasets were downloaded directly from published data, while the K562 dataset by Han et al. was reanalyzed using an identical pipeline. Sequencing reads were matched to the first 19 nt of mature miRNA sequences downloaded from TargetScan7 [19] to generate read counts for each annotated miRNA and its passenger strand. We note that the miRNA sequences of TargetScan include a set of "other miRBase annotations," which includes most of the miRNA passenger strands as well as hundreds of false-positive annotations [20, 21]. Shi et al. [10] had included this set of other miRBase annotations because it contained most of the miRNA passenger strands, which serve as useful internal standards for secondary effects. We opted to use the same list of miRNA annotations as Shi et al. so that we could better compare our results to those of Shi et al. Including these false-positive annotations was not expected to affect our analysis for two reasons: (1) most of the false-positive annotations had too few reads to pass our expression threshold for downstream analysis, and (2) the false-positives that passed our expression threshold were not expected to be sensitive to ZSWIM8 loss. Indeed, examination of all the ZSWIM8-sensitive miRNAs identified by either our new procedure or the previous procedure of Shi et al. confirmed that they were each bona-fide miRNAs, as annotated in MirGeneDB [20].

After mapping to miRNA annotations, changes and p values were calculated using DESeq2 (v1.32.0) [1], using the default Wald test method, without the lfcShrink() function, and with independentFiltering = FALSE in results(). All subsequent analyses were filtered to consider only miRNAs that had at least five matched reads in each replicate of each treatment of each dataset. To conduct BBUM FDR adjustment, the modified BBUM method was used with all default settings in the BBUM_DEcorr function of the bbum R package from this work. Upregulated miRNAs were used as the signal set and downregulated miRNAs were used as the background set. Passenger strands were as annotated in TargetScan7. If no passenger strand was annotated for a miRNA, or if the annotated passenger strand did not meet the read-count cutoff, its fold-change was not plotted.

Significance testing by the method of Shi et al. [10]

To call changes as significant by the method of Shi et al. [10], the default FDR-adjusted p values calculated by DESeq2 using the Benjamini–Hochberg procedure (p_{adj}) were used. The p_{adj} threshold for significance was chosen as the most permissive value out of a defined sequence of canonical critical values (0.05, 0.01, 0.001, 0.0001,…) that excluded

all downregulated miRNAs in the dataset. For example, in iMNs, the strongest downregulated miRNA had a p_{adj} value of 0.000262, and thus the threshold for that dataset was adjusted to 0.0001.

The BBUM statistical model of p values

The p values from the DE experiments can be reasonably modeled as a random variable X following the BBUM distribution:

$$\begin{aligned} X &\sim BBUM(\lambda, a, \theta, r) \\ &= \theta \cdot Beta(ar, 1) + (1-\theta)(1-\lambda) \cdot Beta(a, 1) + (1-\theta)\lambda \cdot U(0, 1); \end{aligned}$$

the probability density function (PDF) of the model is defined as:

$$f(p|\lambda, a, \theta, r) = \theta ar p^{ar-1} + (1-\theta)(1-\lambda) a p^{a-1} + (1-\theta)\lambda;$$

likewise, the cumulative distribution function (CDF) is defined as:

$$F(p|\lambda, a, \theta, r) = \theta p^{ar} + (1-\theta)(1-\lambda) p^{a} + (1-\theta)\lambda p,$$

for $0 < p \leq 1$,$0 < \lambda < 1$, $0 < a < 1$, $0 < \theta < 1$, and $0 < r < 1$. λ represents the fraction of null distribution density over all density except the primary signal, θ represents the fraction of primary signal distribution density over all density, and a describes the shape of the secondary signal peak. r describes the ratio between the shape parameters of the primary and secondary signal components, such that the shape of the primary signal peak is described by ar. The PDF asymptotes to infinity as p approaches 0, and monotonically decreases as p goes from 0 to 1.

BBUM model fitting and parameter estimation

Given a set of p values $\boldsymbol{p}$, the BBUM distribution model was fit to $\boldsymbol{p}$ using a modified maximum likelihood estimation (MLE) method, which varies the parameter values until the calculated (log-)likelihood function value is at its maximum. While varying a shared set of parameters, p values associated with upregulated miRNAs, $\boldsymbol{p}_+$, were fit to the full BBUM function, whereas p values associated with downregulated miRNAs, $\boldsymbol{p}\_$, were fit to the same BBUM function with θ fixed at 0 and r fixed at 1, which corresponded to a zero component for primary effects (Fig. 1C). The total log-likelihood function value ℓ was used as the maximization target for MLE fitting and was defined as the sum of log-likelihood values of the two halves:

$$\ell(\lambda, a, \theta, r|\boldsymbol{p}) = \sum \log\left(f\left(\boldsymbol{p}_+|\lambda, a, \theta, r\right)\right) + \sum \log\left(f\left(\boldsymbol{p}_-|\lambda, a, 0, 1\right)\right).$$

p values of 0 may appear in datasets as a result of underflow, due to the computational approximation of very small values. Due to the asymptotic behavior of the likelihood function when any p values in $\boldsymbol{p}$ equals to 0, any p values smaller than $10\times$ the machine limit in R (.Machine$double.xmin, which was 2.23×10^{-308} in the implementation of this work) were adjusted to $10\times$ the machine limit to avoid this issue.

Fitting was performed using the optim() function in R using the default Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for a maximum of 200 iterations. Unless

otherwise stated, parameters λ, a, and r were bounded to (0, 1). θ was bounded to (0, $1-2\alpha_{BBUM}$), where α_{BBUM} was the critical threshold of BBUM-adjusted p values used for significance testing, to prevent the local minimum near $\theta=1$ where all upregulated points would be regarded as primary signal when there is very low or no primary signal.

The four parameters were bounded by transforming the values using the logit (log-odds) function. Given a parameter x with lower bound x_{lb} and upper bound x_{ub}, the logit function transforms a value as:

$$\text{logit}(x) = \log \frac{x - x_{lb}}{x_{ub} - x}.$$

For each dataset, fitting was initiated at each of six sets of fixed initial parameter values (Additional file 1: Table S2). The initial values for θ were linearly rescaled to its bounds. Out of the six attempts, the successfully converged solution with the highest total log-likelihood function value was chosen as the final solution.

Outlier detection and trimming

Due to the asymptotic behavior of the likelihood function at zero, BBUM model fitting can be prone to overcorrection for secondary effects when outliers with extremely low p values among downregulated miRNAs are present. To mitigate this, we developed and implemented a conservative method for outlier detection and trimming. The data were first preliminarily fit to the model with a wider bound for r at (0, 10). When very strong outliers were present among downregulated miRNAs, the algorithm would converge to a value of $r>1$, which implied a stronger secondary effect than primary effect, violating the assumption of our model and suggesting the existence of unexpectedly strong signal in the background. An increasing number of downregulated miRNAs with the strongest p values were then iteratively trimmed as outliers until the algorithm converged to a value of $r<1$, unless the condition was not met after trimming 5% of or 10 downregulated miRNAs, whichever was lower. BBUM FDR correction of p values using this outlier trimming method is specifically referred to as the modified BBUM method in this work, and is the default in the BBUM_corr and BBUM_DEcorr functions in the bbum R package from this work.

Significance testing

The expected FDR level of falsely calling either null or secondary signal data points as the primary signal can be calculated at any given cutoff for raw p values of upregulated data points. We employed a strategy for adjusting p values that resembled the one that DESeq2 adopts for the Benjamini–Hochberg procedure [1]. For every raw p value of an upregulated miRNA, we calculated the expected FDR value at that value using the BBUM model and denoted it as the BBUM-FDR-adjusted p value (p_{BBUM}) for significance testing. Thus, to control for FDR at a pre-determined cutoff, such as $\alpha_{BBUM}=0.05$, all points with $p_{BBUM}<0.05$ would be called as significant. The expected FDR and the value of p_{BBUM} were thus calculated as

$$p_{BBUM} = FDR_{BBUM}(p) = 1 - \frac{\theta p^{ar}}{F(p|\lambda, a, \theta, r)}.$$

Simulation of p values and benchmarking

To benchmark strategies for significance calling and correction, p values with primary and secondary effects were simulated using the BBUM model. For each simulation, the total number of p values, n, was generated from a uniform distribution, and the number of upregulated points was determined through a binomial distribution:

$$\begin{aligned} N &\sim U(200, 1000) \\ N_{p_+} &\sim B(n, 0.5) \\ n_{p_-} &= n - n_{p_+}. \end{aligned}$$

The values of p were then modeled as a random variable X, which followed a compound distribution of the upregulated and downregulated halves, X_{p+} and X_{p-}, respectively:

$$X \sim \frac{n_{p_+} X_{p_+} + n_{p_-} X_{p_-}}{n}.$$

For each half of the dataset, p values were simulated under respective BBUM models. If programmed outliers were simulated among downregulated points, the outliers were simulated as a separate beta component similarly to the primary signal, where

$$\begin{aligned} X_{p+} &\sim \frac{n_{p+,\,null} X_{p+,\,null} + n_{p+,\,2^\circ} X_{p+,\,2^\circ} + n_{p+,\,1^\circ} X_{p+,\,1^\circ}}{n_{p+}} \\ X_{p-} &\sim \frac{n_{p-,\,null} X_{p-,\,null} + n_{p-,\,2^\circ} X_{p-,\,2^\circ} + n_{p-,\,outliers} X_{p-,\,outliers}}{n_{p-}} \end{aligned}$$

The proportions of the mixture components were drawn from binomial distributions using the values of λ, θ, and θ', with θ' representing the θ parameter for outliers, where

$$\begin{aligned} N_{i,\,1^\circ} &\sim B(n_i, \theta) \\ N_{i,\,outliers} &\sim B\left(n_i, \theta'\right) \\ N_{i,\,null} &\sim B\left(n_i - n_{i,\,1^\circ} - n_{i,\,outliers}, \lambda\right) \\ n_{i,\,2^\circ} &= n_i - n_{i,\,1^\circ} - n_{i,\,outliers} - n_{i,\,null} \end{aligned}$$

Each component of the BBUM model was modeled as previously described, where

$$\begin{aligned} X_{null} &\sim U(0, 1) \\ X_{2^\circ} &\sim Beta(a, 1) \\ X_{1^\circ} &\sim Beta(ar, 1) \\ X_{outliers} &\sim Beta\left(ar', 1\right). \end{aligned}$$

For each simulation, the BBUM parameters were randomly generated from uniform or exponential distributions over reasonable expected ranges of values:

For λ: $\Lambda \sim U(0.1, 0.9)$
For a: $A \sim U(0.1, 0.9)$
For θ: $\log_{10} \Theta \sim U(-1.5, -0.5)$; range of $\Theta \approx [0.03, 0.3]$
For r: $\log_{10} R \sim U(-1.5, -0.5)$; range of $R \approx [0.03, 0.3]$
For θ': $\log_{10} \Theta' \sim U(-2.5, -1.5)$; range of $\Theta' \approx [0.003, 0.03]$
For r': $\log_{10} R' \sim U(-2.0, -1.0)$; range of $R' \approx [0.01, 0.1]$.

Only simulations with at least three data points under primary effects were accepted, to allow sufficient true hit data points for benchmarking. For simulations with outliers, only simulations with at least one outlier were accepted.

True and false positives and negatives were identified by comparing the significance calling of every point to the BBUM distribution component that the point belonged to. Benchmarking statistics, such as FDR and sensitivity, were then calculated using the following equations (Fig. 1C):

$$FDR = \frac{FP}{FP + TP}; Sensitivity = \frac{TP}{FN + TP}$$

For each simulation, 6000 simulated datasets were generated. Confidence intervals for mean and CV statistics were generated by ordinary bootstrapping using the boot package in R. Datasets were resampled 3000 times, and nonparametric 95% confidence intervals were defined by the empirical bootstrap confidence intervals, using the "basic" method of the boot.ci function. Based on the central limit theorem, confidence intervals were presented as the mean deviation of the lower and upper intervals from the sample mean.

Artificial outliers for datasets

To assess the potential impact of outliers on different adjustment methods using empirical datasets from Shi et al. [10], we added to each dataset one to five extreme downregulated outliers with raw p values at 10^{-300}.

Analysis of RNA-seq datasets

Read counts per protein-coding transcript were directly obtained from published RNA-seq datasets [18]. Briefly, these RNA-seq samples had been prepared from HeLa or HEK293FT cell lines transfected with respective miRNA duplexes using RNAiMAX (ThermoFisher, 13778150), and the libraries were prepared using NEXTFLEX Rapid Directional RNA-Seq Kit with poly(A)-selection beads (PelkinElmer, NOVA-5138-07). Reads were aligned to the human genome (reference assembly hg19) using STAR v2.2, and read counts were calculated using htseq-count. Twelve miRNAs were transfected in duplicate in HEK293FT cells, and 17 miRNAs were transfected in duplicate in HeLa cells.

To calculate differential expression of each mRNA upon miRNA transfection, the mRNA read counts observed following miRNA transfection were compared with those observed after the transfection of each of the other miRNAs in that cell line. Using these other datasets as the reference accounted for the global changes caused by miRNA transfection, including the effects of competition with endogenous miRNAs for loading into AGO

proteins [18]. Mean fold changes and p values were calculated using DESeq2 (v1.32.0), using the default Wald test method, without the lfcShrink() function, and with independentFiltering = FALSE in results(). Benjamini-Hochberg-adjusted p values (p_{adj}) were calculated as default by DESeq2. Genes were filtered for having at least 5 read counts across all samples from the corresponding cell line. BBUM FDR-adjusted p values (p_{BBUM}) were calculated using the modified BBUM approach with all default settings in the BBUM_DEcorr function, and using downregulated mRNAs as the signal set and upregulated mRNAs as the background set.

Abbreviations

TDMD	Target-mediated microRNA degradation
BBUM	Bi-beta uniform mixture
FDR	False discovery rate
DE	Differential expression
CI	Confidence interval

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05306-z.

Additional file 1: Table S1. miRNAs in the Shi et al. datasets identified as ZSWIM8-sensitive. **Table S2.** Raw initial parameter values for BBUM model fitting. **Figure S1.** Application of the BBUM correction method to RNA-seq datasets measuring the effects of transfecting a miRNA into either HeLa or HEK293FT cells. **A** Plots of $\log_2$ fold changes in mRNA levels observed upon transfection of the indicated miRNA duplexes, showing all 29 datasets analyzed. Colors are as in Figure 5. **B** Volcano plots of raw p values as a function of mRNA $\log_2$ fold change, showing all 29 datasets analyzed. Colors are as in Figure 5

Acknowledgements
We thank E. R. Kingston, C. Y. Shi, and others in the Bartel laboratory for helpful discussions and feedback.

Author contributions
PYW conceived the approach, designed and implemented the model, and analyzed the data, with input from DPB. Both authors wrote the manuscript. Both authors read and approved the final manuscript.

Funding
Open Access funding provided by the MIT Libraries. This work was supported by a grant from the NIH (GM118135). D.P.B. is an investigator of the Howard Hughes Medical Institute.

Availability of data and materials
All data analyzed during this study are included in these published articles [10, 11, 18] and their supplementary information files. The BBUM fitting and significance-calling algorithm is available as an R package at https://github.com/wyppeter/bbum. Other code used for this work, including data analyses, simulations, and data visualization, is available at https://github.com/wyppeter/BBUM-TDMD_2022.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
D.P.B. has equity in Alnylam Pharmaceuticals, where he is a co-founder and advisor. P.Y.W. declares that he has no competing interests.

Received: 18 February 2022 Accepted: 25 April 2023
Published online: 12 May 2023

References

1. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

2. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
3. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
5. Bartel DP. Metazoan MicroRNAs. Cell. 2018;173:20–51.
6. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215–33.
7. Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet. 2015;16:421–33.
8. Kingston ER, Bartel DP. Global analyses of the dynamics of mammalian microRNA metabolism. Genome Res. 2019;29:1777–90.
9. Reichholf B, Herzog VA, Fasching N, Manzenreither RA, Sowemimo I, Ameres SL. Time-resolved small RNA sequencing unravels the molecular principles of MicroRNA homeostasis. Mol Cell. 2019;75:756-768.e7.
10. Shi CY, Kingston ER, Kleaveland B, Lin DH, Stubna MW, Bartel DP. The ZSWIM8 ubiquitin ligase mediates target-directed microRNA degradation. Science. 2020;370:eabc9359.
11. Han J, LaVigne CA, Jones BT, Zhang H, Gillett F, Mendell JT. A ubiquitin ligase mediates target-directed microRNA decay independently of tailing and trimming. Science. 2020;370:eabc9546.
12. Donahue RMJ. A note on information seldom reported via the P value. Am Stat. 1999;53:303–6.
13. Allison DB, Gadbury GL, Heo M, Fernández JR, Lee C-K, Prolla TA, et al. A mixture model approach for the analysis of microarray gene expression data. Comput Stat Data Anal. 2002;39:1–20.
14. Markitsis A, Lai Y. A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. Bioinformatics. 2010;26:640–6.
15. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics. 2003;19:1236–42.
16. Kleaveland B, Shi CY, Stefano J, Bartel DP. A network of noncoding regulatory RNAs acts in the mammalian brain. Cell. 2018;174:350-362.e17.
17. Eisen TJ, Eichhorn SW, Subtelny AO, Bartel DP. MicroRNAs cause accelerated decay of short-tailed target mRNAs. Mol Cell. 2020;77:775-785.e8.
18. McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science. 2019;366:eaav1741.
19. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;4:e05005.
20. Fromm B, Domanska D, Høye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, et al. MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Res. 2020;48:D132–41.
21. Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. 2010;24:992–1009.

Publisher's Note