Published in final edited form as:

Biometrics. 2011 December ; 67(4): 1206–1214. doi:10.1111/j.1541-0420.2010.01570.x.

Dynamic linear model for the identification of miRNAs in next-generation sequencing data

W. Evan Johnson*,

Brigham Young University, Provo, UT 84602, USA

Noah C. Welker, and

University of Utah, Salt Lake City, UT 84112, USA

Brenda L. Bass

University of Utah, Salt Lake City, UT 84112, USA

Summary

Next generation sequencing technologies are poised to revolutionize the field of biomedical research. The increased resolution of these data promise to provide a greater understanding of the molecular processes that control the morphology and behavior of a cell. However, the increased amounts of data require innovative statistical procedures that are powerful while still being computationally feasible. In this research, we present a method for identifying small RNA molecules, called miRNAs, which regulate genes by targeting their mRNAs for degradation or translational repression. In the first step of our modeling procedure, we apply an innovative dynamic linear model that identifies candidate miRNA genes in high-throughput sequencing data. The model is flexible and can accurately identify interesting biological features while accounting for both the read count, read spacing, and sequencing depth. Additionally, miRNA candidates are also processed using a modified Smith-Waterman sequence alignment that scores the regions for potential RNA hairpins, one of the defining features of miRNAs. We illustrate our method on simulated data sets as well as on a small RNA *Caenorhabditis elegans* data set from the Illumina sequencing platform. These examples show that our method is highly sensitive for identifying known and novel miRNA genes.

Keywords

Bayesian methods; Dynamic linear model; Markov Chain Monte Carlo; miRNA prediction; Smith-Waterman algorithm; Solexa/Illumina sequencing

1. Introduction

Within each cell of an organism, instructions for the proper function of the cell are carried from the DNA by messenger RNA (mRNA) molecules that are translated into proteins outside the cell nucleus. Recently, biologists have discovered several classes of non-protein coding RNAs (ncRNAs) that also play important roles in the proper function of a cell. One of the most popular classes of ncRNAs are microRNAs (miRNAs), which function to regulate the abundance of certain mRNAs in the cell (Ambros, 2004; Bartel, 2009). In

*evan@stat.byu.edu.

Supplementary Materials

Web Appendices referenced in Sections 1, 3.2, 4.1, 5, 6.1, and 6.1.2 are available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

animal cells, miRNAs target mRNA molecules by binding to the 3' end of the mRNA and by recruiting a protein complex that silences the mRNA by degrading it or preventing it from being translated into a protein (Ambros, 2004; Bartel, 2009). Since their discovery a decade and a half ago, miRNAs have emerged as one of the most important and widely studied regulatory mechanisms in animals and plants (Bartel, 2004; Cullen, 2004). miRNAs have been determined to play important roles in metabolism, immunity, cancer, and in specific cell types such as the heart, liver, bone marrow and stem cells (Bushati and Cohen, 2007; Bartel, 2004). Therefore, identifying miRNAs and their targets is an important problem facing biomedical researchers today.

In animal cells, miRNA genes originate from the transcription of small regions of DNA, typically several hundred nt in length. Once transcribed, the RNA from these genes typically fold into a shape that resembles a common hairpin. This hairpin folding is often referred to as the *secondary structure* of the RNA. Both ends of the miRNA hairpin are then cleaved off by cellular enzymes resulting in a ~21–23 nt double-stranded RNA molecule (Lee *et al.*, 2003). Finally, the two strands of the RNA molecule are separated, one strand becoming the mature miRNA and the other strand, called the miRNA* or *star sequence*, is degraded. The hairpin structure and the miRNA* are important characteristics that aid researchers in identifying novel miRNAs. A more detailed description of miRNA processing is included in the Web Appendix A.

1.1 Sequence-based approaches to miRNA discovery

The earliest computational approaches for identifying novel miRNAs were based on a sequence analysis approaches (Lim *et al.*, 2003a,b; Lai *et al.*, 2003; Bentwich, 2005). These methods typically started with conserved non-coding genomic regions that could potentially form hairpin loops or resemble known miRNA precursors. This led to the discovery of several hundred miRNAs, many of which have been experimentally validated. However, these methods lack sensitivity, particularly for miRNAs that are not conserved across species or do not resemble other known miRNAs. In addition, these sequence analysis approaches are prone to false positives.

In total, computational approaches combined with more conventional approaches such as cloning and Sanger sequencing have only led to the discovery of ~100–200 validated miRNAs in animals such as human, rat, mice and *C. elegans*. The actual number of miRNAs in these organisms is still being debated with estimates ranging from ~150 in *C. elegans* to ~700 in *H. sapiens* (miRBase version 12.0)

1.2 Approaches based on next-generation sequencing

The most recent attempts at profiling miRNAs have utilized the deep-sequencing capabilities of next generation sequencing technologies such as the Solexa platform from Illumina, Inc (San Diego, CA), the 454 platform from Roche (Branford, CT) or the SOLiD platform from Applied Biosystems (Foster City, CA). These platforms can produce hundreds of thousands or millions of short sequencing 'reads' in a single run, creating many gigabytes of data. These new sequencing methods have led to a massive increase in the quantity of data produced, which in turn leads to more accurate results than traditional sequencing methods (Schendure *et al.*, 2008). In most applications, this can lead to fast and accurate results with base pair resolution.

Many researchers have been highly successful in searching for small RNAs in animal genomes using next generation sequencing (Ruby *et al.*, 2006; Berezikov *et al.*, 2006; Girard *et al.*, 2007; Pak and Fire, 2007; Brennecke *et al.*, 2007; Tarasov *et al.*, 2007; Morin *et al.*, 2008). These researchers typically mapped reads to the genome or transcriptome and looked

for small RNA sequences that appeared frequently in the data and were consistent with the typical profile of miRNAs. In addition, many of these studies filtered candidates using computational methods and properties of miRNAs (Ruby *et al.*, 2006; Berezikov *et al.*, 2006). However, none of these researchers produced a complete, publicly available algorithm for predicting miRNAs in high-throughput sequencing data.

Recently, Friedlander *et al.* (2008) presented the first publicly available algorithm for analyzing high-throughput sequencing data for predicting miRNAs, called miRDeep. The miRDeep algorithm computes the secondary RNA structure of each read in an attempt to identify RNA hairpins. A filtering step discards any candidates that are inconsistent with known miRNA properties and each remaining miRNA candidate is scored. The filtering steps along with the many different elements incorporated into the scoring function makes miRDeep a highly specific method. However, the gain in specificity leads to a loss in sensitivity as this method struggles to find miRNAs that are present at low levels compared to background, do not exactly fit the filtering criteria, or are not conserved across different organisms.

Here we present a rigorous statistical approach for predicting miRNAs from high-throughput sequencing data. The reads are first mapped to the genome using a probabilistic mapping algorithm (Clement *et al.*, 2010). The mapped read counts for each genomic location are then analyzed using a novel exponential dynamic linear model that identifies candidate miRNA regions. For this application the model parameters are very intuitive, giving base-level estimates of the read ‘hit’ frequency. In addition, this model is well-suited for next-generation sequencing applications as it accounts for both the read counts and the spacing between reads, allowing the model to accurately identify regions in genomic repeats as well as unique genomic regions.

Next, the method post-processes the putative miRNA regions using a sequence analysis method based on a modified version of the Smith-Waterman local alignment algorithm (Nussinov *et al.*, 1978; Smith and Waterman, 1981; Eddy, 2004). Our implementation incorporates the estimated read frequency into the sequence folding algorithm, increasing the accuracy of our predictions. We show that our algorithm works well at identifying and characterizing known miRNAs as well as predicting novel ones. We conclude our research with a simulation study that explores the finite-sample properties of the algorithm and compares with the miRDeep approach.

2. Preprocessing the reads

Henceforth, we will focus the details of our method on data generated from the Solexa/Illumina platform, although all the methods and algorithms below can be naturally applied to data from other platforms with little or no modification. The sequencing ‘reads’ were processed using the standard Solexa pipeline to produce intensity values, quality scores and to call bases. The next step is to map the reads back to their genomic location, for which we applied a probabilistic alignment algorithm, such as GNUMAP (Clement *et al.*, 2010) or NovoAlign (<http://www.novocraft.com/>). These methods use the Illumina quality scores to map the reads back to the genome, as well as produce a posterior probability of alignment for reads that map back to multiple locations in the genome. Such an approach has been shown to outperform other mapping algorithms in the presence of high sequencing errors and in genomic repeat regions (Clement *et al.*, 2010). Due to this probabilistic mapping approach, we expect to observe a large number of small probability reads in repeat regions and a smaller number of high probability reads in unique regions. This means that both the read probabilities and the read densities are both informative to downstream analyses. For

this reason we incorporate both the read probabilities and the read frequencies into our miRNA identification method.

3. Identifying putative miRNAs

3.1 Dynamic linear model

Once the reads are mapped to the genome and the sum of the read probabilities are calculated for all the genomic locations, we use a dynamic linear model (DLM) to identify regions of interest. A DLM is very a flexible tool that is commonly used in time series analysis (West and Harrison, 1999). DLMs are also referred to as continuous-state *Hidden Markov Models* or *state space models*. We selected a DLM for its flexibility as it will allow us to very accurately estimate the very small miRNA regions. In the following paragraphs we construct a novel Bayesian DLM based on a correlated gamma process that can accommodate the small RNA sequencing data. The objective of this model is to estimate the dynamic ‘hit rate’ (the rate at which the reads map back to a given location) for each location across all the genome. The genomic regions of interest are the regions with the highest hit rates.

3.2 MicroRNA identification model

To formally define our model, first let $X_1, \dots, X_N$ be the summed read matches ordered by genomic location, and let Y_i (for $i = 1, \dots, N - 1$) be the genomic distance between X_i and X_{i+1} . Figure 1 is an illustration of the relationships between the X_i and Y_i in the miRNA identification model. Letting $Exp(\mu)$ represent an Exponential distribution with expected value $1/\mu$, we assume that

$$Y_i \sim Exp(\mu_i) \text{ and } X_i \sim Exp(\tau\eta_i),$$

where $1/\mu_i$ is the expected distance between the read hits at locations i and $i + 1$, $1/\eta_i$ is the expected number of read hits at location i , and τ is a parameter that depends on the sequencing depth (total number of RNA fragments sequenced) of the Illumina sequencing run.

Following similar arguments as in the correlated gamma process prior for the Cox Model of Mezzetti and Ibrahim (2000) (see also Ibrahim *et al.* (2001)), we can induce a conjugate DLM over on these data by assuming the following model structure:

$$\begin{aligned} \mu_i | \eta_{i+1} &\sim Gamma(\alpha, \alpha\eta_{i+1}), \text{ and} \\ \mu_{i+1} | \eta_{i+1} &\sim Gamma(\alpha, \alpha\mu_{i+1}), \end{aligned}$$

where α is a smoothing parameter and $Gamma(\alpha, \beta)$ is a Gamma distribution with expected value α/β . A comparison of the correlation structure induced by these assumptions is compared with the correlation structure of a standard DLM in Figure 2. We assign $Gamma(a_\alpha, b_\alpha)$ prior for α and a $Gamma(a_\tau, b_\tau)$ as a prior distribution for τ . Given these assumptions, the complete conditionals for the μ_i s ($i = 1, \dots, N-1$), the η_i s ($i = 2, \dots, N-1$), and τ are as follows:

$$\begin{aligned} \mu_i | y_i, \eta_i, \eta_{i+1}, \alpha &\sim Gamma(2\alpha + 1, y_i + \alpha(\eta_i + \eta_{i+1})), \\ \eta_i | x_i, \mu_{i-1}, \mu_i, \alpha &\sim Gamma(2\alpha + 1, \tau x_i + \alpha(\mu_{i-1} + \mu_i)), \\ \tau | x_1, \dots, x_N, \eta_1, \dots, \eta_N &\sim Gamma\left(n + a_\tau + 1, \sum_{i=1}^n x_i \eta_i + b_\tau\right), \end{aligned}$$

and $\eta_1|\mu_1 \sim \text{Gamma}(\alpha + \tau, \tau x_1 + \alpha \mu_1)$ and $\eta_N|\mu_{N-1} \sim \text{Gamma}(\alpha + \tau, \tau x_N + \alpha \mu_{N-1})$. The complete conditional distribution of α , denoted by $\pi(\alpha)$, is rather complex and is not a known form, as seen below:

$$\pi(\alpha) \propto \frac{1}{\Gamma(\alpha)^{2n}} \alpha^{2n\alpha} \prod_{i=1}^{n-1} \mu_i^{2\alpha-1} \prod_{i=1}^n \eta_i^{2\alpha-1} e^{-\alpha(\sum_{i=1}^{n-1} \eta_i \mu_i + \sum_{i=2}^n \mu_{i-1} \eta_i)}.$$

An Exponential likelihood for the X_i s is appropriate because the summed read probabilities are a positive continuous variable with a high density for low values and strong right-skewness. The Y_i s represent a ‘waiting time’ distribution for which the Exponential is a natural approximation (note that the Y_i s are technically discrete). In addition, one major advantage of our Exponential-Exponential likelihood designation is that it leads to a conjugate DLM and which naturally integrates the X_i s and Y_i s in an intuitive way. Further justification and empirical evidence of the appropriateness of our designations can be found in Web Appendix C.

To further explain the need for including τ in the model, we note that as sequencing depth increases, the reads are more likely map to locations with reads already aligned (not increasing the number of X_i s) as opposed to mapping to new locations to create a new X_i . This is due to the fact that the RNA sample being sequenced has disproportionate levels of different fragments, so the random sampling of a new fragment from the population is likely to map to the same location as previous reads, especially as the sequencing depth increases.

The model framework and the τ parameter allows the data to be ‘binned’ at a courser level while only mildly affecting the results (see simulations below). This is particularly useful in that the miRNA data could be binned into windows of 10–15 nt without strongly impacting the sensitivity of the method. In addition, if applying this method to a different type of sequencing data set with larger fragment size, say an mRNA sequencing experiment or a ChIP sequencing experiment, the reads can be binned at 50–100 nt without a loss in accuracy or much precision. Binning will cut down the number of data points and will speed up the algorithm substantially.

3.3 Parameter estimation

Estimation will be conducted using a Gibbs Sampling algorithm that successively samples from the conditional distributions above, and uses a Metropolis-Hastings step for estimating α . Interesting regions can be selected based on the marginal distributions of the η_i or μ_i . In addition, we have formulated an ECM algorithm (Meng and Rubin, 1993) that estimates the posterior modes and calls interesting regions based on the conditional distributions evaluated at the posterior modes. The ECM algorithm was designed for cases where the size of data restricts the ability to apply the computationally intensive MCMC sampling approach.

4. Post-processing the miRNA regions

After the putative miRNA genes have been identified using the DLM, the significant regions are post-processed to determine whether each candidate region fits other characteristics of miRNAs. In particular, assuming that each short significant region represents a miRNA, we extract the DNA sequences 80 base pairs upstream and downstream from the region. We then perform a sequence folding to the region to determine the likelihood that the region folds into a hairpin structure. The size of 80 base pairs was selected as it is the approximate size of the hairpin.

We have adapted the Smith-Waterman alignment algorithm (Nussinov *et al.*, 1978; Smith and Waterman, 1981; Eddy, 2004) to identify the optimal sequence folding. We applied two sequence alignment algorithms, the first of which applies a standard scoring method designed to investigate the region's likelihood of forming a hairpin structure. The second algorithm integrates the sequencing data into scoring the region, which biases the sequence analysis to favor conformations that include highest scoring sequences in a region, since the mature miRNA should be within the highest scoring segment of the region, and because the miRNA must be part of the predicted hairpin loop.

4.1 MicroRNA folding algorithm

The Smith-Waterman algorithm is a dynamic programming algorithm that is guaranteed to find an optimal local sequence alignment. In order to use such an algorithm to find the optimal folding of an RNA sequence of length M , and let a_i denote the i th element of the sequence, we define two $M \times M$ matrices, and denote them by $\mathbf{H}$ and $\mathbf{T}$. Also, we let $H_{i,j}$ and $T_{i,j}$ be the i th elements of $\mathbf{H}$ and $\mathbf{T}$, respectively. The following recursive algorithm is used to fill in the elements of the matrix:

1. For $i = 1, \dots, M$, set $H_{i0} = H_{0i} = 0$ and $T_{i0} = T_{0i} = 4$.
2. For $\{(i, j) : i + j < M\}$, set

$$H_{i+1,j+1} = \max \begin{cases} H_{i,j} + s(a_i, a_{M-j}) \\ H_{i,j+1} + \text{gap}_{i,M-j+1} \\ H_{i+1,j} + \text{gap}_{i+1,M-j} \\ 0, \end{cases} \quad (1)$$

and let $T_{i+1,j+1}$ take on the value of either 1, 2, 3, or 4, corresponding to which of the four choices/row above resulted in the maximum.

In Equation 1 above, $\text{gap}_{i,j}$ is the penalty for matching a_i with a gap in the alignment in lieu of matching it with a_j , and $s(a_i, a_{M-j})$ is a base scoring function. For the standard scoring function, we set

$$s(a_i, a_j) = \begin{cases} \text{gap}_{i,j} = -1, & \\ 1, & \text{if } a_i \equiv a_j, \text{ and} \\ -1, & \text{otherwise,} \end{cases}$$

where $a_i \equiv a_j$ implies that a_i and a_j are complementary bases (i.e. $A \equiv T$, $T \equiv A$, $C \equiv G$, and $G \equiv C$).

Once the elements of the dynamic programming matrices $\mathbf{H}$ and $\mathbf{T}$ are calculated, the optimal sequence alignment is obtained by using the following 'traceback' algorithm:

1. Identify (i, j) such that $H_{i,j} = \max_{k,l} H_{k,l}$ (for $k + l < M$).
2. While $T_{i,j} \neq 4$ repeat the following:
 - a. If $T_{i,j} = 1$ match a_i with a_j in the sequence alignment. Set $(i, j) = (i - 1, j - 1)$.
 - b. If $T_{i,j} = 2$ match a_i with a gap in the sequence alignment. Set $(i, j) = (i - 1, j)$.
 - c. If $T_{i,j} = 3$ match a_j with a gap in the sequence alignment. Set $(i, j) = (i, j - 1)$.

This results in the optimal sequence alignment, and the miRNA predicted targets are scored and ranked according to their alignment scores.

We note that the folding algorithm presented here differs slightly from a standard Smith-Waterman algorithm in order to accommodate RNA folding. The major difference lies in restricting $i + j < M$ as well as the complementary base scoring function (as opposed to similarity scoring in the Smith-Waterman). A likelihood justification of the RNA folding algorithm as well as a more detailed description of the differences between the folding algorithm and Smith-Waterman algorithm are given in Web Appendices E and F.

4.2 Integrating sequencing data into region scoring

We also present a novel alternative scoring based on the results from the sequencing data and DLM. In order to define our sequence-based scoring function and the gap penalties, we first obtain the hit frequencies estimated from the DLM. Let q_j be the genomic position of η_j from the DLM and let p_i be the genomic position of the i th element of the region's sequence. Define

$$s_i = \begin{cases} 1/(\widehat{\tau}\widehat{\eta}_j), & \text{if } \exists j \text{ such that } p_i = q_j, \text{ or} \\ \widehat{\mu}_j/\widehat{\tau}, & \text{for } j \text{ such that } p_i \in (q_{j-1}, q_j). \end{cases}$$

where $\widehat{\tau}$, $\widehat{\eta}_j$, and $\widehat{\mu}_j$ are the posterior modes of τ , η_j , and μ_j , respectively, estimated from the DLM. The gap and scoring functions are defined as follows:

$$s(a_i, a_j) = \begin{cases} \text{gap}_{i,j} = -s_i - s_j, & \\ s_i + s_j, & \text{if } a_i \equiv a_j, \text{ and} \\ -s_i - s_j, & \text{otherwise,} \end{cases}$$

where $a_i \equiv a_j$ is defined as above. This results in the optimal sequence alignment weighted by the sequencing analysis results from the DLM. The miRNA predicted targets are scored and ranked according to their alignment scores.

5. Application to *C. elegans* sequencing data

We applied our method to a small RNA sequencing experiment conducted on *C. elegans* using the Solexa platform. RNA was extracted from whole, mixed stage nematodes and small RNA molecules ~15–30 nt in length were processed following the standard Illumina protocol. Following sequencing, the data were processed using the standard Solexa pipeline and quality measures were obtained. For our results below, we focus our analysis on Chromosome I (ChrI). The GNUMAP (Clement *et al.*, 2010) algorithm was used to map the reads to ChrI. We deleted any genomic location with a summed read value of less <0.01, as we assumed that these were measuring noise. In all, for the ~15 million bases that comprise ChrI, ~1 million reads were mapped to 49,870 unique genomic locations and were imported into the DLM.

We used the following diffuse prior values: $a_\alpha = b_\alpha = 1$ and $a_\tau = 1$, $b_\tau = 10$. For α , our prior assumes *a priori* that, when smoothing the DLM, the neighboring parameters have equal weight as a single data point. Our prior for τ represents our prior belief that $\tau < 1$, or more specifically that $E(\tau) \approx .1$ or the X s are on average 10 times $1/Y$. However, it is clear from our sensitivity analysis (see Web Appendix G) that these prior values had virtually no effect on the analysis due to the very large sample size of the data.

We used the Gibbs Sampling approach described above to estimate our parameters, setting the number of iterations equal to 25,000. Traceplots indicate that the model ‘burned-in’ very quickly and explored the posterior distribution well. In addition, we applied the gibbsit algorithm (<http://lib.stat.cmu.edu/general/gibbsit>) based on the ideas of Raftery and Lewis (1995). This approach uses a ratio calculated by the sum of burn-in iterations plus the number of iterations after burn-in divided by the size of a pilot sample. Raftery and Lewis (1995) recommend that ratio values be lower than 5. The α and τ parameters did not achieve this ratio, but did so if we thinned the iterations by a factor of 7 (keeping only every 7th iteration) or more. However, thinning this produced no difference in the putative regions called by the method, so we elected to use the initial 25,000 iterations to increase the precision of our posterior estimates. Computation time was between 24–36 hours on a single processor core, but the program required a large amount of memory to complete the analysis (~10–15 GB RAM).

The posterior mean of the smoothing parameter α was estimated to be 1.68 with a 95% credible interval of (1.64, 1.72). This means that the proceeding and flanking μ_i s and τ_i s for the τ_i s and μ_i s (respectively) had a relative weight of 0.771 in influencing the location parameter, compared to a relative weight of 0.229 for the data observation at position i . The τ parameter had a posterior mean of 0.0310 with a 95% credible interval of (0.0302, 0.0317). The median μ value was 0.010 (mean is 0.13, but is highly skewed due a few very large μ_i values), indicating that there was roughly 1 read per 100 base pair on average.

Regions were selected based on the marginal credible intervals for $1/\tau\eta_i$. All genomic positions (with mapped reads) were ranked according to the lower bound of their 95% marginal credible interval. The top 5% of the positions were selected, and any positions within 80 base pairs were combined into one region. For the data in this example, this cutoff was 6.38, which means that any position with a lower credible bound greater than 6.38 reads per base, will be further evaluated for miRNA properties. For each position/region, we extracted the DNA sequence 80 base pairs upstream and downstream of the region. Since most miRNA hairpins are less than 80 base pairs, this is a reasonable size for sequence analysis. This procedure produced a list of 88 putative regions on ChrI that were short enough (<250 base pairs) to contain a miRNA gene. One of these regions, for the known miRNA, *mir-2*, is shown in Figure 3(a) with the read counts, DLM smoothed hit frequency and credible interval lower bound, and the cutoff.

The two sequence folding algorithms were applied to the data. After applying the standard scoring method, the sequences were ranked according to the $\max_{i,j} H_{i,j}/M$ where the $H_{i,j}$ are the elements of the dynamic programming matrix and M is the region’s sequence length. This folding procedure was designed to check for the ability of each region to form a hairpin structure. Of the 88 regions found by the DLM, we selected the top 24 regions that showed strong aptitude for forming a hairpin structure.

The second sequence folding algorithm was based on the sequencing-data dependent gap and scoring functions described in Section 4.2. Due to the high bias associated with locations with high numbers of reads, this algorithm did not give a good indicator for a region’s affinity to form a hairpin. However, this scoring approach is very beneficial for identifying the mature miRNA and its complementary miRNA* sequence. Figure 4 is an illustration of these results for the miRNA, *mir-1*, which was one of the top 24 regions.

For comparison purposes, we also applied the miRDeep algorithm (Friedlander *et al.*, 2008) to these data using default settings and read processing as suggested by the authors. From this dataset, miRDeep produced a list of 6 putative miRNA genes.

5.1 Comparison with known miRNAs

We compared our results and the miRDeep results with the set of annotated miRNAs obtained from mirBase (Griffiths-Jones, 2004). There are currently 21 annotated miRNAs on ChrI. Our method found 9 (42%) of these regions and miRDeep found only 6 (29%). Of the miRNAs found by our method and not by miRDeep, one region had a small number of mapped reads and the other two had a large number of reads and an usually large number of miRNA* reads. Of the 11 regions not found by either method, 2 of them had strong miRNA profiles but barely missed the cutoff for the original 88 region list. Three of the regions had very few mapped reads, and 7 of the regions had absolutely no reads mapped back to them, so they were completely undetectable in these data. Two of these 11 were previously predicted by miRDeep in another data set (Friedlander *et al.*, 2008).

The sequence alignments for these 9 miRNAs were considerably strong and very accurate. The alignment algorithm incorporating the sequencing data was useful for identifying the precise locations of the mature miRNA. Figure 4 is an illustration of this result for *mir-1*. This figure illustrates the high sensitivity of the method for identifying the mature miRNA. The sequencing data based folding identified a fold that was within 1 base on either side of the miRNA. Many of the other known miRNAs showed similar results.

5.2 Predicted miRNAs

Of the top 24 regions selected, 9 of them were known annotated miRNAs obtained from mirBase (Griffiths-Jones, 2004). Of the remaining 15 regions, we identified 4 outstanding miRNA candidates that exhibited strong hairpins and had visible possible miRNA* sequences nearby. The other 11 regions either did not have a strong candidate miRNA* sequence or showed high numbers of repeated elements (runs of A's or T's) so they were not considered strong candidates for biological validation. The miRDeep results did not contain any miRNA predictions.

6. Simulation Studies

We conducted two simulation studies. The first study, designed to emulate a typical high quality next-generation sequencing dataset, is designed to compare the performance of the proposed method with miRDeep as well as evaluate the performance of the RNA folding algorithms. The second simulation is designed to evaluate how accurately the DLM can estimate the true parameter values, the relationship between μ and η , how well the DLM identifies miRNAs in regions with few reads, and how the resolution of the data impacts the results.

6.1 Simulation study 1

We generated 500 simulated next-generation sequencing datasets to evaluate our sequencing data-based folding algorithm as well to compare our approach with the miRDeep algorithm. We randomly generated reads from known miRNAs, miRNA*s, and from background (at 5 different background coverage levels). In total, each of the 500 datasets contained between 2,000 and 7,500 simulated reads (precise details on the generation of these reads can be found in Web Appendix H). The reads were mapped to the entire *C. elegans* ChrI using the GNUMAP algorithm. We applied our approach as well as MiRDeep.

6.1.1 Comparison with miRDeep—We compared the DLM with miRDeep to evaluate which method could most effectively identify the true miRNA spike-ins. For our first comparison, we took the 25 highest scoring regions from both methods and identified how many of the 21 miRNA spike-ins were present. Interestingly, for both methods the spike-in level did not have much effect on the ability of the algorithm to identify the miRNAs.

Overall across all 500 datasets, the 70% of the spike-in regions were identified by the DLM compared to only 13% identified by miRDeep. When accounting for the background level, the DLM varied between 41% and 86%, while miRDeep varied between 1% and 27%. These results are summarized in Table 1.

6.1.2 Evaluation of the sequence folding algorithms—We also evaluated the performance of the sequence folding algorithms on the 500 datasets. We focus on three comparisons here, namely the alignment strength as measured by the proportion of the bases in the folded hairpin that match the correct complementary base, the proportion of the mature miRNA that is correctly folded into the putative hairpin, and the percentage of the miRNA is correctly aligned to the miRNA*. For the alignment strength comparison, the standard scoring algorithm matched 79.5% of the bases with a complementary base, whereas with the sequence-based alignment 80.9% of the bases were matched. For the next comparison, we calculated the percentage of the mature miRNA captured in the folded hairpin by each algorithm. For the standard scoring algorithm, 89.7% of the miRNA bases were within the hairpin while in the sequence-based scoring algorithm 92.8% of the miRNA bases were in the hairpin. Finally, we observed that the standard alignment algorithm only matched 74.5% of the miRNA to the miRNA* sequence, whereas the sequence-based approach matched 85.2% of miRNA to the miRNA*. While the first metrics indicates only a small benefit for utilizing sequencing data into RNA folding, the second and third indicate a strong benefit for the the sequence-based approach. We also considered other metrics for evaluating the folding algorithms. Details can be found in Web Appendix H.

6.2 Simulation study 2

We conducted a second simulation study to evaluate how accurately the DLM can estimate the true parameter values, the relationship between μ and η , how well the DLM identifies miRNAs in regions with few reads, and how the resolution of the data impacts the results. We set the background μ_i to be 1/100. We also spiked in $n \times r = 60$ small ‘miRNA’ regions, ranging in length from $n = 1, \dots, 6$ consecutive genomic locations with $\mu_i = 2^r/100$ for $r = 1, \dots, 10$. We set $\eta_i = (\mu_{i-1} + \mu_i)/2$ and $\tau = 1/10$. We then generated data $y_i \sim \text{Exp}(\mu_i)$ and $x_i \sim \text{Exp}(\tau\eta_i)$ for $i = 1, \dots, N = 1,430$. Based on these parameter values, the background region reads will be spaced on average 100nt apart and will occur at an average level of 0.1 hits. In the spike-in regions, the average hit levels were 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, and 102.4 reads per base (with the average read spacing equal to 10/hit level). Prior values for a_τ , b_τ , a_α , and b_α were selected as in the previous section. We generated 100 datasets for evaluating the DLM in this study.

Averaged across the 100 datasets, the model estimated the posterior mean of τ to be 0.13 with a 95% credible interval of (0.12, 0.15). The slight bias is acceptable due to the fact that the α and τ are highly related, so a slight smoothing bias is expected. The smoothing parameter, α had a posterior mean of 1.9, with a 95% credible interval of (1.7, 2.1). The median value over all the μ_i s was 0.013 which was very close to the expected result.

6.2.1 Identification of regions with few reads—The DLM was also assessed based on its ability to recover the spike-in regions from the simulation data. For this section, we generated 100 data sets and applied our DLM. The top 60 regions were selected from each data set using the lower bound of the 95% credible interval as in Section 5 of this work. Table 2(a) shows the results for this comparison. Regions with only 2–4 times as many reads as the background (0.1 reads per base) needed to be in longer regions to be detected, although some were found. Also, regions with length $n = 1$ needed to be spiked-in at $2^5/10 = 3.2$ reads per base to be reliably identified with high probability. This indicates that the DLM is highly sensitive and can identify miRNA regions with only a small number of reads.

6.2.2 Resolution of the data—One interesting feature of the DLM model is in its ability to produce similar results when the data resolution is coarsened. To illustrate this idea, we binned our data into groups of 100 bases, and then fit the DLM. Figure 3(c), shows that consistent results are obtained between the two data sets, although the courser data have lower resolution and less noise. We also selected regions and compared them with the truth as above. Table 2(b) summarizes these results. According to these results, the DLM is much more accurate for larger *bin size*. This means that the loss in resolution has the trade off of increasing the accuracy of the method.

7. Discussion

Next generation sequencing technologies are providing researchers with increased power to identify interesting biological features relating to the function of a cell. In this research, we presented a sensitive method for identifying miRNA genes from next generation sequencing data. Our method applied a novel dynamic linear model (DLM) based on a correlated gamma process, and the model is developed to have a conjugate form for straight-forward estimation. This model is very flexible and able to accurately identify miRNA genes with only a small number of reads mapped to its location. In addition, we applied an RNA folding method to post-process the candidate miRNA regions from the DLM. This folding method is able to very accurately identify the miRNA and miRNA* sequences within a few bases.

We would like to reiterate that although this paper focuses on the Solexa/Illumina sequencing platform, the methods for identifying miRNA should be naturally applicable without modification to data from other next generation sequencing methods. In addition, the DLM is applicable for the identification of interesting biological features in other applications of sequencing. For example, the DLM could be used to identify regions of protein occupancy in ChIP-seq data or to measure expression or identify new transcripts, transcription start sites, or splice junctions in RNA-seq data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the following individuals for helpful insights and discussions related to this work: David Nix, Brad Cairns, William Christensen, Mark Clement, Shannon Neeley, Quinn Snell and Raphael Gottardo. This work was supported by funds from NIH to W.E.J (HG005692) and B.L.B. (GM067106).

References

- Ambros V. The functions of animal microRNAs. *Nature*. 2004; 431:350–355. [PubMed: 15372042]
- Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*. 2004; 116:281–297. [PubMed: 14744438]
- Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009; 136:215–233. [PubMed: 19167326]
- Berezikov E, et al. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* 2006; 38:1375–1377. [PubMed: 17072315]
- Bentwich I. Prediction and validation of miRNAs and their targets. *FEBS Lett.* 2005; 579:5904–5910. [PubMed: 16214134]
- Brennecke J, et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*. 2007; 128:1089–1103. [PubMed: 17346786]
- Bushati N, Cohen SM. microRNA Functions. *Annu. Rev. Cell Dev. Biol.* 2007; 23:175–205. [PubMed: 17506695]

- Clement NL, Snell Q, Clement MJ, Hollenhorst PC, Purwar J, Graves BJ, Cairns BR, Johnson WE. The GNUMAP Algorithm: Unbiased Probabilistic Mapping of Oligonucleotides from Next-Generation Sequencing. *Bioinformatics*. 2010; 26(1):38–45. [PubMed: 19861355]
- Cullen BR. Transcription and Processing of Human microRNA Precursors. *Molecular Cell*. 2004; 16:861–865. [PubMed: 15610730]
- Durbin, R.; Eddy, S.; Krogh, A.; Mitchinson, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press; 1998.
- Eddy SR. How do RNA folding algorithms work? *Nature Biotechnology*. 2004; 22:1457–1458.
- Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*. 2008; 26:407–415.
- Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*. 2006; 442:199–202. [PubMed: 16751776]
- Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Research*. 2004; 32:D109–D111. (Database Issue). [PubMed: 14681370]
- Ibrahim, JG.; Chen, MH.; and Sinha, D. *Bayesian Survival Analysis*. Springer; 2001. p. 72-75.
- Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of *Drosophila* microRNA genes. *Genome Biology*. 2003; 4:1–20.
- Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN. The nuclear RNase III Drosha initiates microRNA processing. *Nature*. 2003; 425:415–419. [PubMed: 14508493]
- Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of *Caenorhabditis elegans*. *Genes and Dev*. 2003; 17:991. [PubMed: 12672692]
- Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. Vertebrate microRNA genes. *Science*. 2003; 299:1540. [PubMed: 12624257]
- Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*. 1993; 80:267–278.
- Mezzeti, M.; Ibrahim, JG. Technical Report. Department of Biostatistics, Harvard School of Public Health; 2000. Bayesian inference for the Cox model using correlated gamma process priors.
- Miranda KC, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*. 2006; 126:1203–1217. [PubMed: 16990141]
- Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu A-L, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*. 2008; 18:610–621. [PubMed: 18285502]
- Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. Algorithms for loop matchings. *SIAM Journal of Applied Mathematics*. 1978; 35:68–82.
- Pak J, Fire A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science*. 2007; 315:241–244. [PubMed: 17124291]
- Raftery, AE.; Lewis, SM. The number of iterations, convergence diagnostics and generic Metropolis algorithms. In: Gilks, WR.; Spiegelhalter, DJ.; Richardson, S., editors. *Practical Markov Chain Monte Carlo*. London, U.K: Chapman and Hall; 1995.
- Ruby JG, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*. 2006; 127:1193–1207. [PubMed: 17174894]
- Smith TF, Waterman MS. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*. 1981; 147:195–197. [PubMed: 7265238]
- Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnology*. 2008; 26:1135–1145.
- Tarasov V, et al. Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. *Cell Cycle*. 2007; 6:1586–1593. [PubMed: 17554199]
- West, M.; Harrison, J. *Bayesian forecasting and dynamic models*. Springer; 1999.

Figure 1.

Relationship between Y_i , the distance between ordered read i and $i+1$, and X_i which is the number of reads that map to genomic position i . Note that the read spacing and read counts are informative for identifying interesting genomic regions. Also note the inverse relationship between X and Y , *i.e* when X increases, Y should decrease.

Figure 2.

Comparison of the correlation structure of a standard dynamic linear model (top) with the miRNA region finding model (bottom). Note that there is no difference in the actual structure of the two models, only that the expected values of X , η are inverses of the expected values of Y , μ .

(a)

(b)

(c)

Figure 3.

(a) For the annotated miRNA *mir-2*, the figure contains a genomic view of the data, DLM smoothed hit frequency and lower bound, and the cutoff used to select the 88 initial miRNA candidates. The extremely high peak on the left is the mature miRNA and the small peaks on the right is the miRNA* sequence. (b) Relationship between the posterior means of the μ and η parameters. Notice that they are almost mirror images of each other minus a small amount of variation due to the differences/randomness in the x and y data points. From this plot, it is clear that one can either use the μ or the η parameter for inference. (c) Comparison of the base resolution data and the data binned into 100 base regions. It is clear the the

binned data results in less noisy results, however, it comes at the cost of accurate predictions of miRNA regions.

As given in Bartel (2004):

```

*****
3' UGGUGGGGCAAGAUGUAUGAAGAAAUGUAA-GGUGUAGGUA  U  AA
   |||  |||  ||  |||||  |||||  |||  |  ||  A
5' ACCGUACCGAGCUGCAUACUCCUUACAUGCCCAUA-CUAU  U  UA

```

Standard scoring:

```

*****
3' G-UGGGGCAAGAUGUAUGAAGAAAUGUAAGG-UAUAGGUA  U  AA
   |  |  |  ||  |||||  |||||  ||  |||  |  ||  A
5' CGA----G--CUGCAUACUCCUUACAUGCCCAUA-CUAU  C  AU

```

Scoring based on sequencing data:

```

*****
3' UGGUGGGGCAAGAUGUAUGAAGAAAUGUAAGG-UGUA  AAU  GGU
   ||  |||||  |||||  ||  |||  A
5' ACCGUACCGAGCUGCAUACUCCUUACAUGCCCAUAC  UAU  UAU

```

Figure 4.

Sequence folding for the *mir-1* miRNA. The *** gives the actual location of the mature miRNA. The Sequencing weighted data gives a highly accurate location for the mature miRNA and the miRNA* sequences.

Table 1

Results from simulation study 1. We generated 500 simulated miRNA spike-in sequencing datasets at five different background levels (Note that '1X' coverage means that the number of bases in the reads is equal to the number of bases in the region). Below are the average proportion of spike-in miRNAs that are identified in the top 25 scoring region from each method. Notice that the performance of miRDeep decreases as the background level increases.

Background Coverage	DLM	miRDeep
0.063X	0.41	0.27
0.125X	0.62	0.20
0.250X	0.86	0.14
0.500X	0.82	0.05
1.000X	0.78	0.01

Results from simulation study 2. 60 regions were spiked in at an average concentration of $2^{r+1}/10$ reads per base-pair for $r = 0, \dots, 9$ and at lengths of $n = 1, \dots, 6$. (a) For each of 100 randomly generated data sets, the top 60 scoring regions from the DLM were obtained and the proportion of times the region appeared in the top 60 list is given in the table. Regions with only 0.2 or 0.4 reads per base pair (background is 0.1 reads per base) needed to be in longer regions to be detected, although some were found with high probability. Regions with $n = 1$ needed to be spiked-in at only $2^5=10 = 3.2$ to be highly identifiable. (b) Results with the data bin size = 100. Comparing these results with (a) indicates that the DLM is more accurate for higher bin size, although this results in a lower resolution of the regions.

Table 2

(a)		$n=1$	2	3	4	5	6
	$r=0$	0.07	0.15	0.26	0.35	0.52	0.57
	1	0.20	0.60	0.80	0.85	0.92	0.98
	2	0.45	0.84	0.97	0.98	1.00	1.00
	3	0.74	0.98	0.99	1.00	1.00	1.00
	4	0.78	1.00	1.00	1.00	1.00	1.00
	5	0.94	1.00	1.00	1.00	1.00	1.00
	6	0.98	1.00	1.00	1.00	1.00	1.00
	7	1.00	1.00	1.00	1.00	1.00	1.00
	8	0.99	1.00	1.00	1.00	1.00	1.00
	9	1.00	1.00	1.00	1.00	1.00	1.00
(b)		$n=1$	2	3	4	5	6
	$r=0$	0.21	0.37	0.59	0.66	0.78	0.82
	1	0.55	0.78	0.95	0.98	0.99	1.00
	2	0.77	0.93	0.98	1.00	1.00	1.00
	3	0.92	0.99	1.00	1.00	1.00	1.00
	4	0.95	1.00	1.00	1.00	1.00	1.00
	5	1.00	1.00	1.00	1.00	1.00	1.00
	6	0.99	1.00	1.00	1.00	1.00	1.00
	7	1.00	1.00	1.00	1.00	1.00	1.00
	8	1.00	1.00	1.00	1.00	1.00	1.00
	9	1.00	1.00	1.00	1.00	1.00	1.00