BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 4 2012, pages 457–463
doi:10.1093/bioinformatics/btr687

Genome analysis

Advance Access publication December 9, 2011

Identifying small interfering RNA loci from high-throughput sequencing data

Thomas J. Hardcastle*, Krystyna A. Kelly and David C. Baulcombe

Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Small interfering RNAs (siRNAs) are produced from much longer sequences of double-stranded RNA precursors through cleavage by Dicer or a Dicer-like protein. These small RNAs play a key role in genetic and epigenetic regulation; however, a full understanding of the mechanisms by which they operate depends on the characterization of the precursors from which they are derived.

Results: High-throughput sequencing of small RNA populations allows the locations of the double-stranded RNA precursors to be inferred. We have developed methods to analyse small RNA sequencing data from multiple biological sources, taking into account replicate information, to identify robust sets of siRNA precursors. Our methods show good performance on both a set of small RNA sequencing data in Arabidopsis thaliana and simulated datasets.

Availability: Our methods are available as the Bioconductor (`www.bioconductor.org`) package `segmentSeq` (version 1.5.6 and above).

Contact: tjh48@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 31, 2011; revised on November 25, 2011 ; accepted on December 6, 2011

1 INTRODUCTION

Small interfering RNAs (siRNAs) are short RNAs of ~20–25 nt in length cleaved from much longer double-stranded RNA (dsRNA) precursors by Dicer or Dicer-like ribonucleases (Meister and Tuschl, 2004). siRNAs are known to play a substantial role in genetic and epigenetic regulation (Carthew and Sontheimer, 2009), but a full understanding of the mechanisms by which siRNAs themselves are regulated depends, in part, on the characterization of the dsRNAs from which the siRNAs derive. It is clear that, when two siRNAs derive from the same precursor, they must share at least some upstream regulatory factors. As a result, the abundances and functions of the siRNAs derived from the same locus may be correlated, providing increased biological information for downstream analyses.

dsRNAs are highly transient and not easily examined. However, siRNAs that have been stabilized by association with some protein complex [e.g. the RNA-induced silencing complex (Hammond *et al.*, 2000)] are relatively easy to characterize using one of the technologies developed for high-throughput sequencing (Bentley, 2006; Margulies *et al.*, 2005). When siRNAs are derived from the same precursor dsRNA, we expect to see the sequenced reads align to the genome in close proximity to each other and with non-independent abundances. By looking for regions of the genome which show an abundance of reads above background levels, we can establish maps of siRNA *loci* on a genome. These loci approximate the regions of the genome from which the dsRNA precursor is transcribed. Identification of siRNA loci from high-throughput sequencing data must address a number of issues. In sequencing siRNAs, we generally expect to also sequence other classes of small RNAs (sRNAs) with alternative mechanisms of production. These other sRNAs can, if they have been sufficiently characterized, be filtered from the sequenced reads. Alternatively, they can be allowed to remain and be treated in the same way as the siRNA reads, as many other types of sRNA reads will also be generated from some longer precursor, and thus appear as a locus. In the worst case, these sRNA reads will form an additional source of background noise within the data. A degree of background noise is expected in any case, as a result of sequencing errors and the presence of breakdown products from longer molecules such as rRNAs, tRNAs and mRNAs, among other factors. Consequently, not every sequenced read needs be associated with an sRNA locus.

The greatest difficulty in identifying siRNA loci is the problem of accumulation bias; within an siRNA locus, some of the siRNAs will be stabilized by association with some protein complex and be available for sequencing, while others will not. High variation in the coverage of bases within an siRNA locus is therefore expected, with some regions within an locus containing no sequenced reads at all. These accumulation biases within siRNA loci distinguish the data from those found in ChIP-Seq (Johnson *et al.*, 2007) and mRNA-Seq experiments (Wang *et al.*, 2009). Solutions to the problems of peak-calling in ChIP-Seq (Pepke *et al.*, 2009) and transcript discovery in mRNA-Seq (Garber *et al.*, 2011), which in general depend on much more consistent patterns of sequenced reads aligning to the genome, are thus not directly applicable to siRNA locus discovery, and a new set of techniques for the analysis of siRNA loci is required.

A further difficulty is the problem of *multireads* (Mortazavi *et al.*, 2008), sequenced reads that match to multiple places in the genome. The siRNA loci show strong association with repetitive elements such as transposons, and so many siRNA loci match multiple times in the genome. Moreover, whole loci may be replicated exactly (or nearly so) throughout the genome and should all be identified.

Previous attempts to identify siRNA locus maps (Kasschau *et al.*, 2007; Moxon *et al.*, 2008) have taken a relatively naive approach to defining the loci. These methods look for genomic regions in which the number of sequenced reads exceeds some minimum value and there exists no large gap without an aligned sequenced read. The NiBLs algorithm of (MacLean *et al.*, 2010) develops this approach

*To whom correspondence should be addressed.

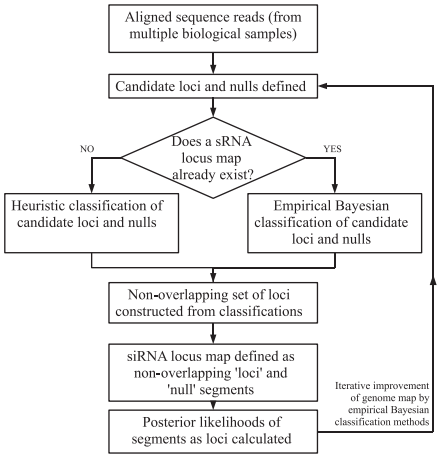

Fig. 1. A flowchart showing the broad outline of the segmentation method for producing an siRNA locus map.

by examining the density of sequenced sRNA reads upon the genome via a graph-theoretic approach whereby sRNA reads are connected if the genomic distance between them lies below some threshold and a region is considered a locus if it is sufficiently connected. Apart from the somewhat arbitrary nature of the thresholds used, a substantial weakness of these approaches is that they allow the analysis of at most two samples, neglecting replicate data. Existing methods are thus unsuitable for robustly defining siRNA loci from large sequencing experiments.

We develop a novel set of methods for constructing siRNA locus maps based on high-throughput sequencing of sRNAs from replicated samples coming from diverse biological conditions. The aim of these methods is to split the genome into non-overlapping *segments*, of which some will be siRNA loci; regions from which in at least some biological condition a dsRNA is expressed and produces siRNAs, and some will be *nulls*; regions in which there is no siRNA locus present in any observed biological condition. The methods are shown to perform well on both simulated and true biological data, and represent a significant step forward in the generation of stable sets of siRNA loci from sequencing data derived from diverse experimental conditions.

2 METHODS

An overview of our methods is given in Figure 1. We begin by defining a set of *candidate loci*; those genomic regions for which it is worthwhile to evaluate whether or not the region is (part of) a true siRNA locus. Our methods also depend on the identification of null regions within candidate loci, and so we also identify the set of *candidate nulls*. We then develop methods to classify the candidate loci and nulls as (part of) *true* loci and nulls within each set of biological replicate samples.

In general, we would prefer to use empirical Bayesian methods to perform these classifications. In order to do this, we must have previously defined a set of siRNA loci and nulls from which to make inferences on the behaviour of the data associated with loci and nulls. This set may be acquired from other sources, but in its absence we will apply heuristic methods to the data and acquire an initial siRNA locus map. We can then use the empirical Bayesian techniques to refine the siRNA locus map.

The methods for classification of the candidate loci and nulls form the core of our method. We develop two distinct approaches to this problem. If no siRNA locus map already exists, we acquire a first classification by applying a heuristic method of low computational cost based on siRNA densities. This approach provides a reasonable first approximation to the siRNA locus map, but fails to discriminate correctly between locus and null regions in some cases, and does not take account of the reproducibility of data within replicate groups. We thus develop empirical Bayesian methods that are able to refine an existing siRNA locus map by estimating the likelihoods that, for each replicate group, a given region forms part of a locus based on the data from all replicates. This approach takes account of the reproducibility of data within each replicate group, but requires an initial siRNA locus map in order to estimate parameters on the distribution of data. We use the heuristic approach to generate this initial map.

The classification of the candidate loci identifies a large set of overlapping regions as true siRNA loci, or parts of loci. These need to be interpreted to form a consensus set of non-overlapping loci. We take an algorithmic approach that identifies a non-overlapping consensus set of loci based on our biological expectations for the behaviour of siRNA loci. We are thus able to *segment* the genome into loci and null segments. We then apply empirical Bayesian methods to all the segments to determine posterior likelihoods that any given segment is truly a locus; this allows us to provide lists of segments ranked by the likelihood that they truly represent a locus. The empirical Bayesian classification of candidate loci and nulls can in theory be iteratively applied to improve performance. However, in practice we find that further applications of the empirical Bayesian methods do not substantially improve performance (data not shown), and given the computational cost of repeatedly applying the methods, this may not be a useful strategy.

The data available for siRNA locus finding consist of a set of sequencing *libraries*, each of which contains a set of sequences defining sRNAs. Each library will belong to a replicate group of samples from biological replicates, and so the samples may be thought of as the set $\{A_1,\cdots,A_m\}$ with a replicate structure defined by $\mathcal{R}=\{R_1,\cdots R_n\}$ where $i\in R_q$ if and only if sample A_i is a member of replicate group q. The sequences from each sample are aligned to a reference genome, creating a set of matches to the reference. In order that data from samples sequenced at different depths may be compared, we define a library scaling factor l_i for the i-th sample by taking the sum of the counts of aligned reads with counts below the 75th percentile of the counts of all aligned reads for that sample (Bullard *et al.*, 2010). This method of determining the library scaling factor filters out very highly expressed reads, helping to stabilize the library scaling factors.

We define a segment starting at base a_j and ending at base b_j as s_j; the length of the segment is then $\lambda_j=b_j-a_j+1$. For each segment, we need also to define a *count* u_{ij} of the number of sRNA associated with this segment for a sample i. The problem of *multireads*, sequenced reads that match to multiple places in the genome, requires consideration here. We address this problem by counting the total number of sequenced reads that match to the genome within this segment from each sample, rather than the total number of matches to this segment. Thus, if a sequenced read appears N times (the *count* of the sequenced read) in the sequenced library i, but within a single segment j matches the genome Q times, it will contribute a total of N, rather than NQ, to the count u_{ij}. However, we allow this read to be counted in each different segment to which it matches, thus allowing the discovery of repeat-associated siRNA loci.

2.1 Principles of the empirical Bayesian analysis

We begin by establishing the principles of an empirical Bayesian analysis of these data, adapting our previous work on differential expression analyses of count data (Hardcastle and Kelly, 2010). Both the classification of the candidate loci and nulls, and the evaluation of the posterior likelihoods of the siRNA locus map, depend on an empirical Bayesian analysis of the data and so we establish the principles in a fairly general form.

Suppose that we have two models for the data in replicate group q; M_L^q and M_N^q. These model the number of sRNA sequences u_{ij} that align within a given segment s_j under the condition that s_j is a locus or a null in that replicate group. We want to evaluate the posterior likelihoods of M_L^q and M_N^q given the observed data for segment j. We wish to estimate the posterior

likelihoods of the models for each segment. To do this, we must first acquire a distribution for the parameters of the data given in the model. Given a pre-existing siRNA locus map, we can construct prior distributions for the data contained within the loci by taking some sampling Θ_q of the regions classified as loci. For each of these regions, we find parameters defining the mean and dispersion of the data, giving a distribution of these parameters for data within loci. If we have some previous estimate of the likelihood of this region as a locus, we are able to refine this distribution further. For any given region, we are then able to assess the likelihood of the data associated with that region given that the region is a locus. Similarly, we can construct a distribution on the parameters for data within nulls, and assess the likelihood of the data associated with any region given that the region is a null. Given priors on the models, we can then calculate posterior likelihoods for each model. We formalize this approach below.

Estimation of posterior likelihoods: suppose that we have some model M describing the number of small RNAs that will align within a segment for all the samples within a particular replicate group q. The posterior likelihood of a generic model M given the observed data pertaining to segment s_j is given by

$$\mathbb{P}(M \mid \{u_{ij} : i \in R_q\}; l_i, \lambda_j) = \frac{\mathbb{P}(\{u_{ij} : i \in R_q\}; l_i, \lambda_j \mid M)\mathbb{P}(M)}{\mathbb{P}(\{u_{ij} : i \in R_q\}; l_i, \lambda_j)} \quad (1)$$

The primary challenge in calculating this posterior likelihood is that of finding $P_{qj} \equiv \mathbb{P}(\{u_{ij} : i \in R_q\}; l_i, \lambda_j \mid M)$, the likelihood of the data under the model M. We assume that for a given replicate group R_q and segment s_j the *counts* u_{ij} are distributed independently negative binomially with a dispersion ϕ_j and a mean proportional to both the library scaling factor l_i of the sample and to the length of the segment λ_j, with constant of proportionality μ_{qj}.

Distribution of parameters: now suppose that we have some distribution θ_q on the values (μ_q, ϕ). Then we can evaluate the likelihood of the count data as

$$P_{qj} = \int \mathbb{P}(\{u_{ij} : i \in R_q\}; l_i, \lambda_j \mid \mu_q, \phi)\mathbb{P}(\mu_q, \phi \mid \theta_q)\mathrm{d}\theta_q$$

If we have a set of values Θ_q that are sampled from the distribution of θ_q, then we can derive the approximation

$$P_{qj} \approx \frac{1}{|\Theta_q|} \sum_{\{\mu_{qk}, \phi_k\} \in \Theta_q} \mathbb{P}(\{u_{ij} : i \in R_q\}; l_i, \lambda_j \mid \mu_{qk}, \phi_k) \quad (2)$$

We can acquire the set of values Θ_q by sampling a random set of non-overlapping regions from some appropriately chosen set of segments. This set of segments should conform in some way with the model for which a distribution is being sought; for example, if our model is that of loci, then we will want to sample a set of segments which we believe to represent loci.

Estimation of parameters from sample: for a specific segment s_j, we have the observed number of matches u_{ij} of sRNAs within this segment from each sample i. We assume that the number of matches in each sample are independently negative binomially distributed, with the mean number of matches for each sample i given by $\hat{\mu}_{qj} l_i \lambda_j$ whenever $i \in R_q$, and that the dispersion of data is ϕ_j for all samples. We then take a quasi-likelihood approach to estimate these parameters for the observed data. We first define $\hat{\mu}_{qj} = \left\langle \left\{ \frac{u_{ij}}{l_i \lambda_j} : i \in R_q \right\} \right\rangle$, and then choose ϕ_j such that

$$\sum_q \sum_{i \in R_q} u_{ij} \log\left[\frac{u_{ij}}{l_i \hat{\mu}_{qj}}\right] - (u_{ij} + \phi_j^{-1}) \log\left[\frac{u_{ij} + \phi_j^{-1}}{l_i \hat{\mu}_{qj} + \phi_j^{-1}}\right] = \frac{n-1}{2}$$

where n is the number of samples. Using this value for ϕ_j, we can then re-estimate the values $\hat{\mu}_{qj}$ via maximum likelihood methods. We then iterate on our estimations of ϕ_j and $\hat{\mu}_{qj}$ until convergence. Then $\{\hat{\mu}_{qj}, \phi_j) \in \Theta_q$ for each q. By repeating these calculations for a large number of sampled j, we acquire the sets Θ_q for each q.

The larger the size of the sample used to generate the set Θ_q, the more closely this set will estimate θ_q and the more accurate the estimation of the posterior likelihoods will be. However, increasing the sample size beyond a certain point increases the computational time required by the analyses without substantially affecting the estimated posterior likelihoods. In our analyses of siRNA-seq data, together with previous work in the analysis of sequencing data (Hardcastle and Kelly, 2010) we have found empirically that a sample size of 10^5 is sufficient to give stable and accurate results, and this value is used throughout the analyses performed here.

Weighting the distribution: we can refine the approximation given in Equation (2) further by weighting the calculated parameters taken. If we know that the likelihood of a particular sampled segment s_k conforming to our model M is w_{qk}, then we can derive the approximation

$$P_{qj} \approx \frac{1}{|\sum w_k|} \sum_{\{\mu_{qk}, \phi_k\} \in \Theta_q} w_j \mathbb{P}(\{u_{ij} : i \in R_q\}; l_i, \lambda_j \mid \mu_{qk}, \phi_k) \quad (3)$$

where $\{\mu_{qk}, \phi_k\} \in \Theta_q$ are the parameters derived from s_k.

We can thus use either Equation (2) (if no likelihoods are available) or Equation (3) (if likelihoods are available) to estimate the likelihood of the data given some model M.

The key element of the estimation of the posterior likelihood of the model M is the acquisition of the set Θ_q by sampling from some set of segments known to conform with the model M. Since we wish to find posteriors for the likelihood that a region is either a locus or a null, this means that we need to have sets of loci and nulls from which to sample our data. Before we can apply the empirical Bayesian approach, we therefore need an existing map defining loci and null regions. We can acquire this through heuristic methods, and then improve our estimation of the loci using the empirical Bayesian approach.

2.2 Identification of candidate loci and nulls

In order to identify true loci and nulls, we first need a set of *candidate* loci and nulls to which we apply our classifiers. In theory, we can take an exhaustive approach and consider every possible start and end within each chromosome. For computational reasons, however, this is impractical, and it is necessary to place restrictions on the set of candidate loci. We begin by limiting the set of candidate loci to those that begin at the start site of some sRNA match in at least one dataset, and end at the end site of some (not necessarily different) match. We further restrict the set by requiring any candidate locus that overlaps with a match to completely contain that match; thus, if two matches overlap then they must be within the same locus. A final limitation is that no candidate locus contain a region within that locus containing no matches whose length exceeds some limit Λ.

We next define the set of candidate null segments in terms of the candidate loci. Four classes of these null segments exist: those acquired by considering the empty regions between loci, those consisting of each candidate locus extended into the empty space to the left, those consisting of each candidate locus extended into the empty space to the right and those consisting of each candidate locus extended into the empty space to both the left and right.

2.3 Methods for classification of loci and nulls

Heuristic classification: we can carry out an initial classification of the segments into loci and nulls using heuristic methods. For a replicate group R_q and segment s_j we can consider the mean density of sRNAs (measured in terms of RKPM, or reads per kilo base per million sequenced reads) that align within the segment;

$$D_{qj} = \frac{10^9}{|R_q|} \sum_{i \in R_q} \frac{u_{ik}}{l_i \lambda_j}$$

The simplest approach to classification of segments is to define some cutoff δ on the density of sRNAs within a segment that implies the presence of a locus. However, under this classification scheme, very short empty segments will be classified as nulls, which is unlikely to be appropriate.

A refinement to this scheme is thus to define some minimum gap ζ for null regions. The classification scheme then becomes

$$\begin{aligned} \textit{locus} \text{ if } & D_{qj} > \delta \\ \textit{null} \text{ if } & D_{qj} \leq \delta \wedge Z_{qj} > \zeta \\ \textit{unclassified} \text{ if } & D_{qj} \leq \delta \wedge Z_{qj} \leq \zeta \end{aligned}$$

Empirical Bayesian classification: now suppose that we have some siRNA locus map consisting of a set of non-overlapping segments $\mathcal{S}$ for which we have already defined for each segment s_j and each replicate group q the posterior likelihood p_{qj} of the segment representing a locus. The acquisition of these likelihoods for each segment is discussed below. We can then build an empirical distribution for the parameters of the data for the candidate loci and nulls based on this genomic map, and thus refine our classification of the loci and null segments.

We first determine whether some candidate locus can be classified as a locus. A candidate locus should be classified as a locus if it lies within a true locus. Conversely, a candidate locus should not be classified as a locus if it lies within a true null region. We therefore seek to establish the posterior likelihood that a candidate locus lies within some true locus, given the observed data. We can calculate this posterior likelihood for each candidate locus and each replicate group using Equation (1), together with the approximations defined by Equation (3).

Suppose that for a replicate group q, we have two models for the behaviour of the data within a candidate locus. These models are slightly different from those described by M_L^q and M_N^q in that they look at the distribution of data within the candidate loci, rather than the loci themselves. Accumulation biases within loci can lead to candidate loci with low densities of small RNA alignment. Conversely, localized background noise can lead to some candidate loci within a null region having relatively high densities of small RNA alignment. We, therefore, have two models $M_{L'}$, in which the candidate locus exists within some true locus, and $M_{N'}$, in which the candidate locus exists within a null region.

We derive the set Θ_q for these models by taking a non-overlapping random sample (without replacement) from the set of candidate loci that lie wholly within some segment of the siRNA locus map. Suppose that the candidate locus l_k lies within some segment s_j. Then in order to establish the likelihood of the data under model $M_{L'}$ [via Equation (3)], the weighting w_{qk} of the corresponding member of Θ_q is p_{qj}, the likelihood that the containing segment s_j is a locus. Conversely, in order to establish the likelihood of the data under model $M_{N'}$, the weighting w_{qk} is $1-p_{qj}$. This allows us to establish the posterior likelihood of a candidate locus forming part of a true locus via Equation (1). In order to classify a candidate loci as a locus, and thus acquire an unambiguous siRNA locus map, we choose some minimum value on the posterior likelihood for a candidate locus to be classified as a locus. This minimum value might reasonably be anything >0.5 (so that the candidate locus is more likely to be a locus than not); in practice, we find a likelihood of 0.9 ensures that only the high confidence loci are considered, and this value is used in the analyses presented in this article.

We next need to discover which of the candidate nulls that lie within the regions we classify as loci are truly nulls. A true locus cannot contain a null region, and so if a segment classified as a locus contains a null, we will need to split that segment into two or more separate loci.

In classifying the candidate nulls, we assume that they can be modelled either as a true locus or a true null, and so we use the original models M_L^q and M_N^q. We can then establish posterior likelihoods on the candidate nulls as before, deriving the set Θ_q by taking a sample from the set of segments $\mathcal{S}$. To evaluate the likelihood of the data given the model M_L^q via Equation (3), we weight each member of Θ_q by the likelihood p_{qj} that the corresponding segment j is a locus. Similarly, we evaluate the likelihood of the data given the model M_N^q by weighting each member of Θ_q by the likelihood $1-p_{qj}$ that the corresponding segment is a null.

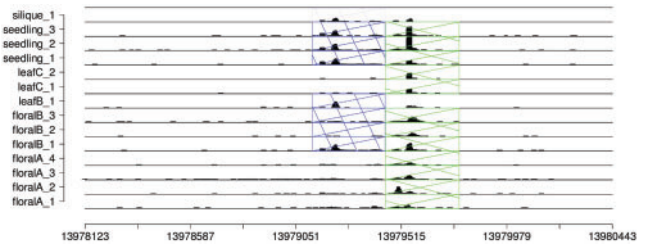

Fig. 2. sRNA densities on a region on chromosome 1 of *Arabidopsis thaliana*. Two adjacent loci are identified based on different patterns of abundance of sRNAs in different replicate groups. Colour intensity indicates the posterior likelihood of a locus being present for a given biological sample. Green and blue colourings are used to allow nearby loci to be distinguished from one another.

2.4 Constructing the siRNA locus map

Given a set of overlapping candidate loci and nulls, let us suppose we have classified these regions as true loci or nulls within each replicate group, allowing the possibility that our classifier will be unable to determine whether a segment is either a locus or a null in some cases. We use these classifications of the candidate loci and nulls to define the siRNA locus map.

Null exclusion: we begin with the assumption that a true locus cannot contain a null region within it. Suppose that some candidate locus l_r is classified as a locus in replicate groups Ψ_r. If there exists some candidate null n_s contained within the candidate locus l_r that is classified as a null in one or more of the replicate groups Ψ_r, we discard the locus l_r.

We then rank the loci that remain after excluding those that contain null regions by the number of replicate groups in which they are classified as a locus, settling ties by choosing the longer locus. We then filter these segments by choosing all those segments that do not overlap with some higher ranking segment to obtain a non-overlapping set of loci.

This process allows us to separate loci based on the pattern of differential expression of the loci between two or more replicate groups. Figure 2 shows how this can occur; two loci are identified with no (or a small) gap between them due to the identification of very different patterns of abundance of siRNAs between the replicate groups. These different patterns of siRNA abundance suggest that two or more siRNA loci may be present at this location; an inference that would be impossible to draw without the simultaneous analysis of data from multiple biological sources. Further examples of such regions are shown in Supplementary Figure S1.

Locus extension: we next wish to identify the nulls as the regions between the filtered loci. However, in some cases the filtration step can lead to the case where two loci are sufficiently close together that the region between them should not be classified as a null, but instead incorporated into one or both of the loci. This prevents large numbers of very short nulls from distorting the maps.

Suppose that we have two loci l_r and l_t, separated by a candidate null n_s, where the locus l_r is classified as a locus in replicate groups Ψ_r and the locus l_t is classified as a locus in replicate groups Ψ_t. If n_s is classified as a null in at least one replicate group from both Ψ_r and Ψ_t, then we confirm n_s as a null. However, if n_s is not classified as a null in any replicate group in Ψ_r, but is classified as a null in some replicate group in Ψ_t, then we extend the locus l_r to cover the region n_s. Similarly, if n_s is not classified as a null in any replicate group in Ψ_t, but is classified as a null in some replicate group in Ψ_r, then we extend the locus l_t to cover n_s. If n_s is not classified as a null in any replicate group in either Ψ_r and Ψ_t, then we extend both l_r and l_t (proportionally to their length) to cover the region n_s.

Given some method for classifying regions as loci and nulls for each replicate group, we can thus acquire a set of non-overlapping loci $\mathcal{L}$ and a set of non-overlapping nulls $\mathcal{N}$.

2.5 Likelihoods of loci

Suppose we have a map defining some set of loci $\mathcal{L}$, and a set of nulls $\mathcal{N}$. The union of these sets is a non-overlapping set of segments $\mathcal{S}$ that covers the entire genome. However, for some of these loci, individual replicate groups may not be expressed. Moreover, we believe that some of the loci, or nulls, are incorrectly called, particularly in the heuristic analyses. We thus seek to establish posterior likelihoods for each replicate group q that a segment s_j conforms to a locus model M_L^q or a null model M_N^q based on an empirical Bayesian analysis of the data.

We first take a random sample of the loci in order to construct a set $\Theta_q^{(L)}$ from which we can assess the likelihood of the data given the model M_L^q [Equation (2)]. Similarly, a sample of the nulls allows the construction of the set $\Theta_q^{(N)}$ from which we assess the likelihood of the data given the model M_N^q. Equation (1) then gives us the posterior likelihood of each segment as a locus or a null.

Given the posterior likelihoods of each segment as a locus or a null in replicate group q, we then take a random sample from the set of segments $\mathcal{S}$ to acquire the set Θ_q. By using the posterior likelihoods of each member of the sample being a locus as the weightings w_{qk}, we re-evaluate the likelihoods of the data for each segment given the model M_L^q [Equation (3)]. Similarly, by weighting each member of the sample by the posterior likelihoods of being a null, we re-evaluate the likelihoods of the data for each segment given the model M_N^q. This allows the posterior likelihoods of each of the two models M_L^q and M_N^q to be re-evaluated. We can then iterate on the estimation of the posterior likelihoods until we achieve convergence.

3 RESULTS

We test our methods using data from 14 samples (Gene Expression Omnibus accession number GSE31211) of *Arabidopsis thaliana* in which sRNAs from a variety of tissues have been sequenced. There are six distinct *replicate groups*: sets of sequenced samples that can reasonably be identified as replicates. Within a replicate group, we expect to see replication of loci, that is, for a locus to be positively identified it should appear consistently within all members of that replicate group. However, a locus need not appear in all replicate groups, as some siRNA loci will be expressed only in specific tissue types, or due to technical effects, be found only in some library preparations. The samples, and the methods used to process them, are described in Supplementary Materials.

We next take the maps derived from the biological data and use them to estimate parameters for simulation studies on which we can compare the methods. The empirical Bayesian approach is shown to offer substantial improvements over the heuristic methods in both biological and simulated data.

3.1 Analyses of siRNA loci from *A.thaliana*

The sequencing data contains reads mapping to 5 038 063 unique locations. From this mapping, 8 405 236 candidate loci are identified (using $\Lambda = 100$). The heuristic method, with a minimum RKPM [reads per kilobase per million mapped reads (Mortazavi *et al.*, 2008)] of 2000 required for a locus to be identified, and a maximum gap permitted within a locus of 100 bp (Section 2), 11 031 loci are discovered, covering ∼1.41 Mb in total. When the posterior likelihoods of the genomic segments are calculated, we find that 10 718 segments have a posterior likelihood of >50% of being a locus in at least one replicate group, and that these segments cover 1.40 Mb. If we consider only loci with a >90% likelihood of being a locus, we find 7368 loci, covering 1.04 Mb.

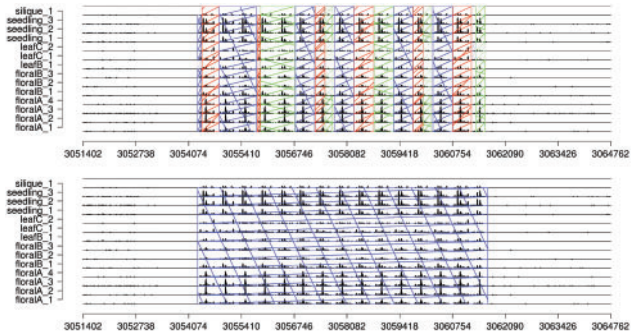

Fig. 3. sRNA densities on a (repetitive) region on Chromosome 4 of *Arabidopsis*. The heuristic methods (above) separate the region into multiple small siRNA 'loci', while the empirical Bayesian methods call the region as a single 'locus'. Colour intensity indicates the posterior likelihood of a locus being present for a given biological sample. Different colours are used to allow nearby loci to be distinguished from one another.

Using the siRNA locus map from this heuristic analysis as a basis for the emprical Bayesian methods, we find 9229 loci, covering ∼1.73 Mb in total. Examining the posterior likelihoods of the genomic segments, we find that 9199 segments have a posterior likelihood >50% of being a locus in at least one replicate group, and that these segments again cover 1.73 Mb, while 7485 segments have a posterior likelihood >90% of being a locus in at least one replicate group, covering 1.39 Mb.

Comparison of the siRNA locus maps demonstrates that the heuristic method tends to identify sets of short loci that lie close together upon the genome. In contrast, the empirical Bayesian method is often able to identify these regions as part of some larger locus, giving fewer but longer loci. Figure 3 and Supplementary Figure S3 show examples of this; a region of the genome appears to be over-segmented using the heuristic methods, in which many short loci are identified within a moderately short region. The empirical Bayesian method instead identifies a single locus in this region.

The loci discovered can in some cases be verified. Mature microRNAs (miRNAs), while produced by a different mechanism (Bartel, 2004) to siRNAs, can be found in sRNA-seq experiments and have been extensively studied. The miRNA loci appear in the sequenced sRNA data as short, dense regions of sequenced reads, qualitatively different to the majority of siRNA loci. Methods specifically designed for the discovery of miRNA loci (Friedländer *et al.*, 2008; Yang and Li, 2011) that take into account hairpin folding and sRNA distribution, are of course to be preferred to a purely density based approach. Nevertheless, the methods developed here for detecting siRNA loci are able to detect many miRNAs simply on the basis of sRNA densities. Of the 180 miRNA loci identified in The Arabidopsis Information Resource (version 10) (Swarbreck *et al.*, 2008), the empirical Bayesian methods identify segments overlapping with 112 of the miRNA loci at a likelihood of >0.9 in at least one of the replicate groups. The heuristic methods identify segments overlapping with 103 of the miRNA loci at a likelihood of >0.9 in at least one of the replicate groups. These are high proportions given that different miRNAs are expressed at different stages of development and in different tissues.

The best-studied candidates for siRNA loci are the TAS loci (Vaucheret, 2005), from which sRNA production is induced from

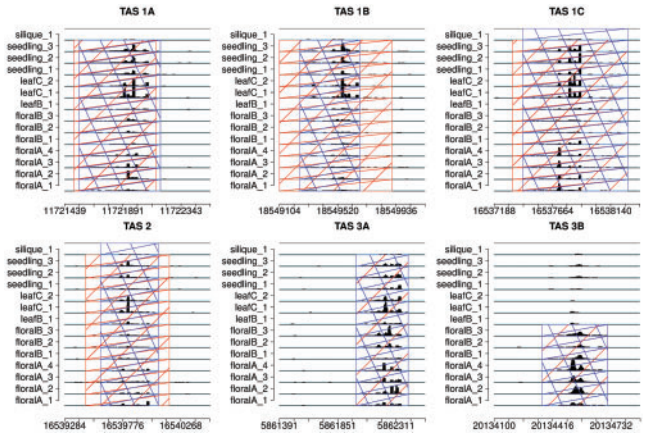

Fig. 4. Loci identified in the *Arabidopsis* TAS regions. Blue regions are the coordinates of loci identified by heuristic methods; red regions the loci identified by empirical Bayesian methods. The intensity of the coloured regions indicates the calculated posterior likelihoods of a locus being present for a given biological sample.

non-coding RNAs via miRNA cleavage. Of the six TAS loci in *A.thaliana* for which exact genomic coordinates are known, all are identified by both the heuristic and empirical Bayesian methods (Fig. 4), although the heuristic method in general identifies a smaller region than the empirical Bayesian method within the defined TAS loci coordinates.

3.2 Comparison of locus finding methods on simulated data

We next apply our methods to simulated data for which we know with certainty the complete set of true positives. The simulation of these data is described in Supplementary Materials.

We first consider the ability of the methods to discriminate between locus and null regions. We rank the discovered segments by the maximum likelihood of their being a locus in one of the two replicate groups, and, for the top n segments, examine the extent to which each of them is contained within some true locus. If more than some proportion p of the segment exists within a single true locus, then we call this a true positive, otherwise, we consider it a false positive. Figure 5a shows receiver operating characteristic curves for these analyses for heuristic and empirical Bayesian loci for various proportions p. The loci identified by empirical Bayesian methods identify a lower proportion of false positives for high proportions of true positives; moreover, nearly all true positives have been identified with posterior likelihoods >50% in at least one of the replicate groups.

We also assess the methods on their ability to accurately define the true loci. We do this by again ranking the discovered segments by the maximum likelihood of being a locus in one of the two replicate groups, but now counting as true positives the highest rank segment which contains some proportion p of a previously undiscovered true locus, and does not extend the boundaries of this locus (in total) by some limit l. Figure 5b shows the capacity of the methods to define the true loci, requiring at least 90% of the true locus to be identified, with various options upon the limit l. Where a low extension of the boundaries of the true loci is permitted, this implies that the locus detection methods must more precisely define the boundaries of

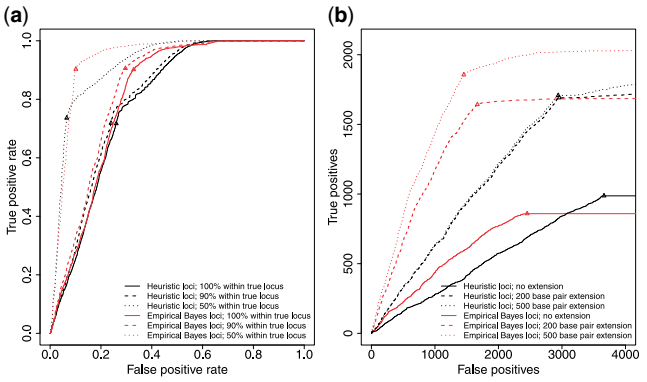

Fig. 5. Performance indicators of the two methods on simulated data. In both cases, the loci are ranked by their maximum likelihood of being a locus in any replicate group. Triangles indicate where this maximum likelihood falls <50% in the discovered loci. (**a**) Receiver operating characteristic curves for the loci discovered from simulated data. A locus is counted as a true positive if more than some proportion p is contained within a true locus, otherwise it is considered a false positive. (**b**) True positives for the 'discovery' of simulated loci; a segment is called a true positive if it covers >90% of a true locus and exceeds the boundaries of that locus by no more than some limit l.

the locus. The heuristic method is able to perfectly define ($l = 0$) the boundaries of more loci than the empirical Bayesian methods, but at a cost of substantially increased false positives. When the conditions on defining boundaries are relaxed, the empirical Bayesian method is able to detect more true positives with far fewer false positives than the heuristic method.

4 DISCUSSION

We have developed a set of methods for the discovery of sRNA loci from high-throughput sequencing. Two elements form the key to our approach; a method for classifying any given region as either a locus or a null region in each replicate group, and an algorithmic approach for combining these classifications for overlapping regions to generate a locus map. We propose two methods of classification; a heuristic approach based on the average density of sRNA reads within a candidate locus for a given replicate group and an empirical Bayesian approach that can refine the output from the heuristic approach by assessing the posterior likelihood of a region as a locus for each replicate group. The empirical Bayesian methods are considerably more computationally intensive than the heuristic methods. Approximately, one thousand processor-hours were required to compute the siRNA locus map for the wild-type samples described here, although the methods are readily parallelizable and the run-time can thus be much shorter than this in practice. Despite the increased computational resources required, the empirical Bayesian approach appears to offer a number of advantages over the heuristic approach. In analyses on true biological data, the empirical Bayesian approach identifies more miRNA-associated regions than the heuristic approach, and more accurately defines the boundaries of the TAS loci. In comparisons on simulated data, the empirical Bayesian methods identify more true positives than do the heuristic methods, with fewer false positives being detected. The empirical Bayesian methods also appear better placed to deal with the problem of accumulation bias than the

heuristic methods, which may split what appears to be a single locus into many separate loci.

It is clear that improvements might be made in any individual case by adjusting the parameters of the heuristic method. However, it is not possible in advance to know what these parameters should be, nor to readily compare siRNA locus maps produced with different parameterizations. Since the empirical Bayesian method depends on the map produced by the heuristic methods, we might also expect the performance of the empirical Bayesian method to vary depending on the parameters used to characterize the heuristic map. Supplementary Figures S4 and S5 show how the performance of the methods on the simulated data alters with different choices for the minimum RKPM and maximum gap used to define the heuristic map. The performance of the heuristic method does show substantial variation according to the choice of parameters; however, the performance of the empirical Bayesian methods is much more robust. Given a sufficiently poor choice of parameters for the heuristic method, the empirical Bayesian method will show a degradation in performance as the initial map will be too far from the truth to allow reasonable inferences to be made on the distribution of the underlying parameters of the data. Nevertheless, the empirical Bayesian method consistently shows an improvement over the heuristic method regardless of the initial map, suggesting that this approach may be usefully used on any existing siRNA locus map. Given a reasonable first approximation to the siRNA locus map, the empirical Bayesian approach offers the advantage that the majority of the parameters involved are estimated from the data; thus, factors such as background noise and reproducibility of results are automatically accounted for.

Our methods are a substantial step forward in locus finding from high-throughput sequencing data. Comparisons with the NiBLs algorithm (MacLean *et al.*, 2010), the most recent existing method for detecting small RNA loci demonstrate that both the heuristic and empirical Bayesian methods offer substantial improvements in specificity (Supplementary Fig. S3). This is not surprising, as the incorporation of replicate data allows much greater certainty in distinguishing between true loci and background noise. Existing methods for siRNA locus detection, which allow the analysis of only one or, at most, two samples, are unlikely to show similar performance to the methods developed here. Crucially, our methods incorporate replicate data from any number of replicate groups, and, as such, are able to define robust siRNA locus maps from large datasets. This represents a significant advance over previous methods, which were unable to take into account data from replicate groups or to assess data from more than one or two samples.

We envisage that the application of these methods to sufficiently large datasets will allow the definition of robust siRNA locus maps, which will be made publicly available for use in downstream analyses, avoiding the computational costs involved in regenerating maps for each study and allowing easier comparison of results between experiments. Methods for easily incorporating new data into existing locus maps are also being developed. Although we have not explored the possibility here, we are hopeful that these methods may also be useful on other high-throughput sequencing data which produce broad, flat regions rather than peaks, for example, chromatin modification ChIP-Seq data.

Funding: European Commission Seventh Framework Programme grant number (233325 to T.J.H.).

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Carthew,R.W. and Sontheimer,E.J. (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.

Friedländer,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Hammond,S.M. *et al.* (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, **404**, 293–296.

Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kasschau,K.D. *et al.* (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol.*, **5**, e57.

MacLean,D. *et al.* (2010) Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *BMC Bioinformatics*, **11**, 93.

Margulies,M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Moxon,S. *et al.* (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, **24**, 2252–2253.

Pepke,S. *et al.* (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6** (11 Suppl.), S22–S32.

Swarbreck,D. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.

Vaucheret,H. (2005) MicroRNA-dependent trans-acting siRNA production. *Science STKE*, **2005**, pe43.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.*, **10**, 57–63.

Yang,X. and Li,L. (2011) miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, **27**, 2614–2615.