

ANALYSIS OF LARGE-SCALE SEQUENCING OF SMALL RNAS

A. J. OLSON, J. BRENNECKE, A.A. ARAVIN, G.J. HANNON AND R. SACHIDANANDAM

*Cold Spring Harbor Laboratory,
1 Bungtown Road,
Cold Spring Harbor, NY 11724, USA
E-mail: ravi.cshl.work@gmail.com*

The advent of large-scale sequencing has opened up new areas of research, such as the study of Piwi-interacting small RNAs (piRNAs). piRNAs are longer than miRNAs, close to 30 nucleotides in length, involved in various functions, such as the suppression of transposons in germline^{3,4,5}. Since a large number of them (many tens of thousands) are generated from a wide range of positions in the genome, large-scale sequencing is the only way to study them. The key to understanding their genesis and biological roles is efficient analysis, which is complicated by the large volumes of sequence data. Taking account of the underlying biology is also important. We describe here novel analyses techniques and tools applied to small RNAs from germ cells in *D. melanogaster*, that allowed us to infer mechanism and biological function.

1. Introduction

Relatively inexpensive large-scale sequencing has now become readily accessible to the masses, through the efforts of companies such as 454 and Solexa. One drawback of sequences derived from such technologies are the short read lengths, approximately 30 for Solexa and more than a 100 for 454. This does not pose a problem when the sequences can be easily identified in the genome. In fact, for small RNAs, the length is close to their size and hence such sequencing techniques are perfect for their study.

Small RNAs have been discovered to be associated with the Argonaute family of proteins. The Argonaute family is a complex one, and the nomenclature makes it even more confusing¹. The family is further divided into two sub-classes, Argonaute and Piwi. The Argonaute sub-class is involved in the siRNA and miRNA pathways, while the Piwi sub-class is involved in piRNA processing. piRNAs tend to be longer than miRNAs and are more

diverse. Their genesis is not well understood, but our analysis of piRNAs shows that they arise from clusters and are frequently repeat associated, which suggests their role in transposon silencing. Indeed, deletion of certain members of the Piwi sub-class leads to the activation of transposons.

In order to characterize the datasets (we have analyzed three such datasets ^{3,4,5}), as well as understand their biological role, we developed several analysis techniques and tools. Taken individually, the techniques and tools are not novel, but the combination and sequence of steps makes them a novel contribution which will be of use to others in the field. Several other large-scale sequencing projects have been published, but none involves the kind of analysis described here ².

The aim of the experiments described here was to delineate the role of Aubergine and Ago3 in the germline in silencing transposons. We wanted to understand the transposons that are under control of this mechanism and which one of these proteins is involved in targetting the transposons and which one is involved in the maintenance of this silencing. The analysis that we describe here was arrived at by trial and error. We describe the methodology as well as our tools below.

2. Case Study: Analysis of Aubergine and Argonaute-3 associated small RNAs from *D. melanogaster*.

Small RNAs associated with Argonaute-3 and Aubergine in *D. melanogaster* were isolated by immunoprecipitation (IP), using antibodies against the proteins⁴.

The analysis involves sequence processing, mapping and warehousing. In addition, it is very important to allow browsing of the data through user-friendly tools, since the patterns we look for are not readily apparent and the pattern space that has to be searched is immense. We describe our suite of web-based tools for this purpose. The exact implementation of these tools is not very important, as they are standard techniques, but the analyses allowed by the web-based tools is very relevant, since they helped us gain understanding of our datasets.

We first describe the sequence processing and then the tools. We also describe the results we obtained at each step.

2.1. Sequence Processing

The processing of the sequences involves clipping the sequences, mapping them to the genome, using genome annotations to identify the origin of the

small RNAs and warehousing the data. We describe the steps below.

2.1.1. *Clipping*

Adaptors are ligated to the small RNA sequences for amplification and sequencing. It is essential that the parts of the adaptor sequences that are sequenced get identified and clipped. This can involve either exact matching (if the reads are short and high quality) or inexact matching, when the reads are longer and the quality might have degraded towards the end. We use a dynamic programming algorithm that scores in a position dependent manner, allowing for more relaxed matching towards the end and a stricter match towards the beginning of the sequence. This allows cleaning up sequences that do not have any inserts, and also being relatively careful about not removing sequences arbitrarily from the end. The distribution of sizes after clipping suggests the nature of the dataset (if the peak of the distribution is around 22nt, it indicates a library biased towards miRNAs while a peak closer to 30 indicates piRNAs).

2.1.2. *Warehousing the data*

The sequences from the experiment are collapsed to create a unique, non-redundant set, and the multiplicities (the number of times each fragment is sequenced in an experiment) are tracked. We use MySQL's relational database for the storage.

2.1.3. *Mapping to the genome*

Mapping of the small RNAs is an relatively easy problem, compared to mapping mRNAs, since gaps are not expected. In addition, due to the large number of sequences, the ones that do not map exactly to the genome can be ignored. We used a suffix-array based approach to find matches. This is essential in speeding up the processing of the small RNAs and highlights the importance of proper clipping.

Some small RNAs map thousands of times to the genome, while about 10% of the small RNAs map to a unique location on the genome. The unique mappers allow identification of the clusters which are the main sources of the small RNAs.

2.1.4. *Annotating the small RNAs*

The annotations of the underlying genome are used to annotate the small RNAs. The annotation categories are repeats, non-coding RNAs (tRNAs, snoRNAs, miRNAs etc.) and coding mRNAs, both introns and exons. It is essential to get a good reference set of annotations or generate one from curated datasets especially for the non-coding RNAs. Upto 10 mappings of each piRNA are considered for annotation, and a majority rule is used to pick the final annotation. In addition, in case of conflicts, a hierarchy, starting with non-coding RNAs, then repeats, followed by exons and finally introns, is used to pick a unique annotation for the piRNA. The orientation of the piRNA with respect to the underlying annotation is also identified.

Result: The small RNAs in this experiment are predominantly repeat-associated. The Aubergine associated ones are mainly anti-sense to the repeats while the Argonaute-3 associated ones are sense to the repeats.

2.2. *Web-enabled tools*

We built a set of web-based tools to allow exploring the dataset. We describe the function of the tools and the conclusion reached with each tool, but not the exact implementation, since the functionality is important in understanding the nature of the piRNAs while the implementation involves fairly standard techniques. The front-end, which is the starting point of the analysis, is shown in figure 1. The front-end allows filtering the small RNAs by various criteria such as, annotation, multiplicity (number of times the sequence was sampled in the experiment), number of mappings on the genome, and location on the genome. After the filtering, the selected small RNAs can be analyzed for distribution on the genome (by using a genomic viewer that is built into the tool), for the distribution of nucleotides at various positions (by using a tool to generate weight matrices for collections of sequences), for the density distribution on the genome (by specifying a window and step size for the sliding window to a graphing tool built into the tool), and for the correlations between positions on the genome for two sets of small RNAs in a graphical format.

2.3. *Genome View*

We built a viewer to view annotations along with small RNA map positions to allow browsing the genome (Figure 2). The viewer is based on the light-weight genome viewer (lwg^v 7).

DROS 5, SmallRNA Query Page

?

Use The Form Below To Select A File To Upload:

Browse And Select File:
 no file selected

Or, Indicate Options On The Form Below And Click The Button To ...

Title:

Chromosome:

Enhance Map:

Normalize Density?

Averaging Window

Y-Axis Max:

Library:

 Keep Libraries Separate ?

Grade Limits

Density
 Browser View
 Weight Matrix
 Get Sequences

Chromosome Start:

Chromosome End:

Min Map Number:

Max Map Number:

Min Grade:

Max Grade:

Win Size:

Win Step:

Show Chr Gaps

Show Repeats
 Graph Track

Reject:

Select Chromosome From To

Figure 1. Front end of the tool to study small RNAs. This allows the selection of small RNAs by various filters. The filters are annotations (repeats, miRNAs etc.), number of mappings to the genome, the experimental source and by chromosome if necessary. From the selected small RNAs, (i) weight matrices (figure 3), (ii) graphical representation of the density of small RNAs in regions of the genome (figure 5), (iii) a browser view in the form of tracks (figure 2), and (iv) Position correlations between datasets (figure 4) can be generated. Each of these plays a crucial role in the analysis, as explained in the text.

Result Viewing regions of the genome with annotations of repeats and small RNAs confirms the association of the small RNAs with repeats.

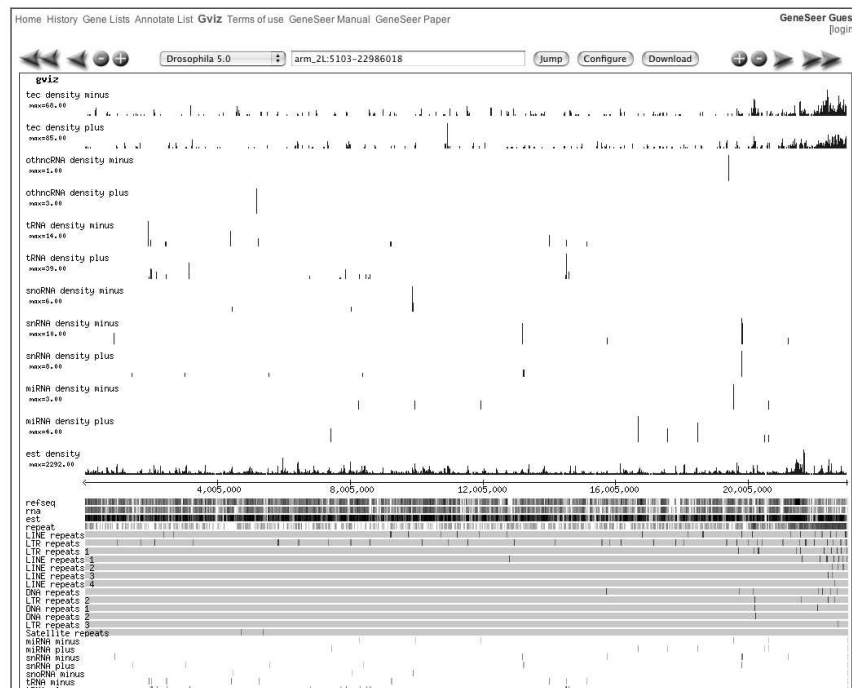


Figure 2. A view of a genomic region showing various features along with the small RNAs in our browser, based on lwgv.

2.4. Nucleotide bias

A standard method of characterizing collections of small sequences is to study the nucleotide bias as a function of position. Our tool generates colored images showing the frequencies of the nucleotides as rectangles, with the height of each rectangle proportional to the frequency (shown in black and white in figure 3).

Result: The Aubergine-associated small RNAs show a T-bias at position 1, which is similar to the one seen in other piRNA sets, while the Ago3-associated ones show a bias for an A at position 10. This suggests the following mechanism. Aubergine uses the small RNAs to target and cleave transposons. The cleavage occurs at position 10, which means there is an A at position 10 of the cleaved sequence from the transposon. The sequence from the transposon gets loaded into Ago3, through an unknown process, which is probably used to target the primary transcript that generates the small RNAs, setting up an amplification cycle, which explains

