BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 3 2012, pages 318–323
doi:10.1093/bioinformatics/btr686

Sequence analysis Advance Access publication December 9, 2011

Detection of microRNAs in color space

Antonio Marco* and Sam Griffiths-Jones

Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, M13 9PT Manchester, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Deep sequencing provides inexpensive opportunities to characterize the transcriptional diversity of known genomes. The AB SOLiD technology generates millions of short sequencing reads in color-space; that is, the raw data is a sequence of colors, where each color represents 2 nt and each nucleotide is represented by two consecutive colors. This strategy is purported to have several advantages, including increased ability to distinguish sequencing errors from polymorphisms. Several programs have been developed to map short reads to genomes in color space. However, a number of previously unexplored technical issues arise when using SOLiD technology to characterize microRNAs.

Results: Here we explore these technical difficulties. First, since the sequenced reads are longer than the biological sequences, every read is expected to contain linker fragments. The color-calling error rate increases toward the 3′ end of the read such that recognizing the linker sequence for removal becomes problematic. Second, mapping in color space may lead to the loss of the first nucleotide of each read. We propose a sequential trimming and mapping approach to map small RNAs. Using our strategy, we reanalyze three published insect small RNA deep sequencing datasets and characterize 22 new microRNAs.

Availability and implementation: A bash shell script to perform the sequential trimming and mapping procedure, called SeqTrimMap, is available at: http://www.mirbase.org/tools/seqtrimmap/

Contact: antonio.marco@manchester.ac.uk

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

Received on July 26, 2011; revised on December 6, 2011; accepted on December 7, 2011

1 INTRODUCTION

So-called next-generation sequencing technologies, or deep sequencing, permit the fast and comprehensive analysis of genomes and transcriptomes (Mardis, 2008). The short length of the sequence reads produced is compensated by the capacity to produce millions of reads in a single run. Thus, new strategies to align and assemble highly redundant short sequence reads have been developed over the last few years (Flicek and Birney, 2009; Li and Homer, 2010).

MicroRNAs are endogenous RNA molecules, ∼22 nt in length that repress gene translation (Bartel, 2004). In the past 4 years, the overwhelming majority of novel microRNAs have been identified by deep sequencing. Reads from deep sequencing experiments may contain sequences of the short DNA adapters (termed 'linkers' here) used in the sequencing reaction. Characterization of small RNAs from deep sequencing datasets requires the removal of these linker sequences from the 3′ ends of reads. Illumina/Solexa sequencing has been extensively used to detect microRNAs (Berezikov *et al.*, 2011; Morin *et al.*, 2008; Ruby *et al.*, 2007) and the available data suggest that the 3′ linker sequences are easily detected and removed. For instance, Ruby *et al.* (2007) detected 3′ linker fragments in >82% of the sequenced reads by string matching.

The use of AB SOLiD sequencing to characterize microRNAs is on the rise (Cai *et al.*, 2010; Chen *et al.*, 2010; Goff *et al.*, 2009; Li *et al.*, 2010; Marco *et al.*, 2010). Unlike other technologies, SOLiD machines produce sequences in color space, each color representing a dinucleotide. The rationale behind color space is that, since colors are produced for overlapping dinucleotides (Fig. 1A), each nucleotide is read twice. This is purported to reduce sequencing errors, and to permit better distinction between sequencing errors and polymorphisms (Applied Biosystems, 2008). The characterization of microRNAs in color space produces two specific issues that have so far been overlooked, one associated with each end of the read. First, the read length is longer than the biological sequence, such that the 3′ end of every read derived from a microRNA contains linker sequence that must be removed. However, detecting and removing adapter sequences in color space is not as straightforward as in base space, as we explore in this work. Second, the first color of each read represents the last base of the adapter and the first of the target sequence. The treatment of this color is controversial since different programs keep or remove it. Removal of the first color causes the first base to be lost, whereas retaining the first color may reduce the proportion of reads that map to the genome. The loss of the 5′ nt has critical consequences in the characterization of microRNAs. In this article, we address the issues of using color space sequences to characterize microRNAs and other small RNAs, and provide a simple strategy to easily map color space reads from small RNA libraries to whole genomes.

2 METHODS

2.1 Linker fragments detection

To analyze the nature of the contaminant sequences within reads, we explore the presence of linker fragments at 3′ ends in a *Tribolium* adult library (GEO accession number: GSM639446). We used the cutadapt tool (http://code.google.com/p/cutadapt/) to remove the linker sequences used during the sequencing reaction (5′ linker: CCACTAC GCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT; 3′ linker: CTGCC CCGGGTTCCTCATTCTCTCATCGGCTGCTGTACGGCCAAGGCG). We also mapped all adjacent 20 color length overlapping fragments within the last 30 colors of each read against these linker sequences using Bowtie (Langmead *et al.*, 2009) allowing from 0 to 3 color mismatches.

*To whom correspondence should be addressed.

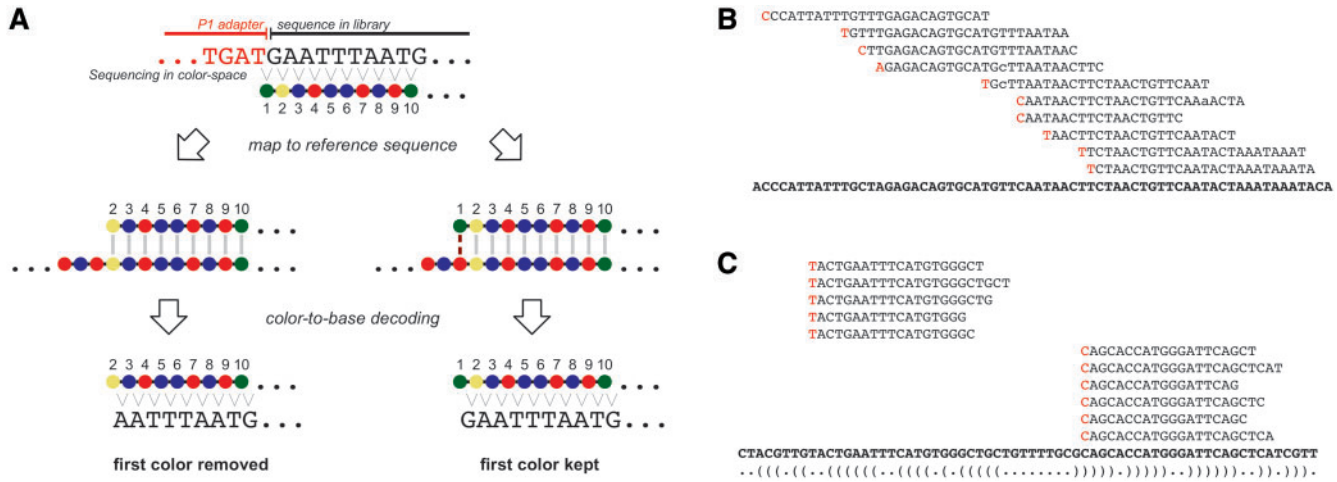

Fig. 1. Effects of color space encoding in the first nucleotide of sequenced reads. (**A**) cDNA sequences are linked to a P1 adapter. The first color produced is determined by the first base of the read and the last of the adapter. If we remove the first color, the first base is lost during the color-to-base decoding. This first base is kept if we do not remove the first color. (**B**) Effect of missing the first nucleotide (red) in sequencing long (fragmented) sequences. (**C**) Effect of missing the first nucleotide in sequencing microRNAs.

2.2 Sequential trimming and mapping of reads in color space

Sequence datasets were obtained from GEO (http://www.ncbi.nlm.nih.gov/geo/) and SRA (http://www.ncbi.nlm.nih.gov/sra). For each dataset (GEO:GSM639446; GEO:GSM639447; SRA:SRR039230), we mapped the full-length reads, and then sequentially trimmed the last color of the reads and repeated the mapping procedure for all previously unmapped reads, to a minimum read length of 19 colors. At each mapping step, we first removed reads that match a library of known rRNAs and tRNAs for the target species. Ribosomal RNAs (rRNAs) in each target species were extracted from the SILVA database (http://www.arb-silva.de/, release 108). Transfer RNAs were detected in honeybee and beetle genomes using tRNAscan-SE with default parameters (Lowe and Eddy, 1997). The remaining reads were mapped against the reference genome sequence (*Tribolium castaneum* version 3.0 and *Apis mellifera* version 4.0) with Bowtie (Langmead *et al.*, 2009) allowing two color mismatches (-v 2) and retaining all best matches (-a --best --strata). We modified the input files to force Bowtie to keep the first color of the reads by removing the first letter of each sequence in the input file (B.Langmead, personal communication).

2.3 Detection of microRNAs

Reads mapped at color length 20–26 were used for microRNA detection. Reads that map to >5 positions in the genome were removed. We grouped overlapping reads and retrieved flanking regions (−50 to +100 and −100 to +50) from the target genomes. We scanned these genomic fragments for hairpins using RNAfold (Hofacker *et al.*, 1994). We applied a series of filters to the resulting hairpins to detect putative microRNAs based on Marco *et al.* (2010), with minor modifications to increase the stringency of microRNA calls: (i) reads that were sequenced only once were removed prior to microRNA detection; we subsequently require that (ii) at least 10 reads map to the putative arms of the predicted hairpins; (iii) the hairpin folding energy is below −20 kcal/mol; (iv) at least 50% of the nucleotides of one arm of the hairpin pair with nucleotides from the other arm; (v) at least 70% of a minimum of five reads from one arm have the same 5′ end; (vi) putative mature sequences from both arms (so-called miR and miR* sequences) are supported by reads, and the most abundant reads from each arm pair in the predicted hairpin structure across at least 70% of their length; (vii) where a read maps to multiple genomic loci, all loci are annotated as microRNA candidates by criteria 1–6. MicroRNA candidates were manually inspected and compared with previously annotated microRNAs using BLAST (Altschul *et al.*, 1997).

3 RESULTS AND DISCUSSION

3.1 Linker fragments in color space reads

Sequence reads often contain, in their 3′ end, fragments of the linker used during the sequencing reaction. Moreover, when the biological sequences of interest are shorter than the read length, we expect that most reads contain linker sequences. Linkers are typically removed by filtering out 3′ ends that match with known linker sequences. However, in the particular case of SOLiD reads this is problematic. AB SOLiD machines (versions 3 and 4) produce 50 nt long reads. The sequencing error rate rapidly increases toward the 3′ end of the read (Sasson and Michael, 2010). We can estimate the proportion of linker sequences at the 3′ end of the read that contain no sequencing errors. Let e_i be the color call error rate at position i of the read. Then, the proportion of reads with no errors at position i is $(1-e_i)$. Hence, the proportion (P) of reads with no errors at their 3′ end is:

$$P = \prod_{R-r+1}^{R} (1-e_i) \quad (1)$$

where R is the length of the read and r the length of the 3′ end fragment we are testing. Using the error rates described for *Escherichia coli* re-sequencing data in Sasson and Michael (2010), and assuming that the linker accounts for half of the read (25 nt), we estimate that the percentage of reads with no errors in the linker sequence is ∼43%. That means that we could not detect linker fragments in more than half of the reads by simple sequence matching. Moreover, linker fragments can have multiple sequencing errors. Assuming a fixed error rate of 0.033 [average for the last 25 nt of the reads according to Sasson and Michael (2010)], and using the

Table 1. Linker fragments detected at 3′ end of sequenced reads using Bowtie

Color mismatches	Reads matching 3′ linkers (%)	Reads matching 5′ linkers (%)
0	15 540 742 (23.17)	1827 (0.00)
1	24 626 724 (36.72)	2745 (0.00)
2	31 983 586 (47.69)	3815 (0.01)
3	38 392 776 (57.24)	5775 (0.01)

binomial distribution, the expected percentage of sequences with at least 2 errors is 20%, and for 3 or more errors is ∼5%. Additionally, >90% of the mature microRNAs registered in miRBase (Kozomara and Griffiths-Jones, 2011) are <25 nt. Hence, these values are likely to be underestimates of the real impact of errors in the 3′ ends of the SOLiD small RNA sequencing reads.

We further explored whether the 3′ end sequences are actually linkers in SOLiD small RNA datasets. We scanned sequence reads from a small RNA library from *T.castaneum* for known 3′ linkers used during the sequencing process using Bowtie (see Section 2.1 for details). We find that ∼23% of reads contain linker fragments with zero mismatches. As we increase the number of color mismatches to 3, we detect linker sequences in up to 57% of the reads (Table 1). As a control, we also mapped against the 5′ linker (P1 adapter) used during the sequencing reaction. As expected, we find virtually no 5′ linker fragments at the 3′ ends of the reads (Table 1). These data suggest that the number of errors in the 3′ linker sequence is higher than that predicted by our model, such that we miss *bona fide* linker fragments. Another possibility is that many reads are chimeric artifacts of small RNAs and other fragments of transcripts. As expected, when we directly remove linkers using the cutadapt tool (see Section 2.1), only 2741 linker sequences are removed, and the number of reads that can be subsequently mapped to the genome is very low (Supplementary Material). Altogether, we conclude that directly filtering for linker sequence is not productive for SOLiD data.

To deal with 3′ end linker sequences of variable length, we propose a strategy based on sequential trimming and mapping. This type of strategy is appropriate for contaminant fragments of unknown length and undetectable origin (Cloonan *et al.*, 2009; Marco *et al.*, 2010). The procedure is as follows. First, we map all reads to a reference genome in color space using Bowtie (Langmead *et al.*, 2009). We trim the last color of only the unmapped reads and repeat the mapping. We sequentially trim one color and re-map, to a minimum read length in our case of 19 colors. Bowtie is particularly amenable to this approach for two main reasons: its speed allows multiple rounds of mapping, and the '--un' option allows easy access to the unmapped reads at each stage. Bowtie is, however, less sensitive to reads with multiple mismatches (David *et al.*, 2011), although this is not a major consideration for microRNA analysis. Bowtie decodes colors to nucleotides using the dynamic programming approach described in (Li and Durbin, 2009); as a consequence, the length of the decoded sequence will be 1 nt shorter than the length in colors of the read. That is, if we trim sequences to a minimum of 20 colors, the minimum nucleotide length will be 19.

We note that RNA2MAP (Applied Biosystems, 2009), the program provided by AB SOLiD, deals with the linker issue in a different way, but using the same principle that we cannot directly remove linker fragments in color space. RNA2MAP maps the reads from 5′ to 3′ by extending an initial aligned seed. During the process, the reads are mapped against 'hypothetical reads' (concatenating genome fragments and linker sequences) in order to detect contaminants (Applied Biosystems, 2009). We compare the two approaches in a later section. RNA2MAP does not prefilter sequences based on quality values. The rationale behind this is that mapping in color space allows the easy identification of errors in color calls. Likewise, we did not prefilter the datasets analyzed in this work. However, we note that a prefiltering step significantly reduces the computational complexity of mapping, with a small impact on sensitivity for microRNA detection (Supplementary Material). This confirms that our mapping approach is also robust to low-quality sequences.

3.2 Mapping the first color of the read

The first color of a read from the SOLiD output is determined by the last nucleotide of the P1 adapter linker and the first nucleotide of the actual read (Fig. 1A). Some mapping algorithms, including Bowtie, remove the first color before mapping, as only half of the information in the first color derives from the sequenced molecule. In this case, during the color-to-base decoding step we lose the first nucleotide of the actual read (Fig. 1A). This may have little or no effect when mapping overlapping reads from a longer transcript (Fig. 1B), but when we map small processed RNAs, we will wrongly detect the beginning of the functional sequences (Fig. 1C). Indeed, the correct definition of the 5′ end of the microRNA is critical for functional analysis. Other programs, such as SHRIMP (Rumble *et al.*, 2009) and BFAST (Homer *et al.*, 2009), keep the first color of the read during the mapping process. We propose keeping this first color but allowing an extra color mismatch to account for the 0.75 probability that the first base encoded in the color (the last base of the 5′ linker) does not match the 5′ flanking base in the genome.

Keeping this first color has an important consequence in the mapping of reads. Since we retrieve all 'best' mapping positions, reads mapping equally well to multiple genome sites may be artificially differentiated by score based on the matching of the 5′ end color. For instance, sequence 0132012 will map better to 0132012 than to 1132012. However, both sequences may map equally well to both positions in base space. On the other hand, if we remove the first color, an alternative problem arises: a read that is unique in the genome may map to many places because the first nucleotide is not taken into account. Bowtie has a new native option to keep the first and last color of the reads but ignore color mismatches in those positions. We reanalyzed the honeybee data using this option. We find that the number of matches per read increases, suggesting a reduction in mapping specificity. The number of microRNAs detected by the downstream analysis falls slightly, suggesting our strategy outperforms the Bowtie option. We, therefore, recommend that the first color is retained for the purpose of microRNA detection. We also suggest that candidate microRNAs detected by deep sequencing (which will likely number only in the hundreds) should be independently mapped to the reference genome, using for example BLAST (Altschul *et al.*, 1997), in order to detect potential copies that escaped our mapping procedure. However, in our experience, this bias is negligible.

Table 2. Reads mapped with the sequential trimming and mapping strategy

Experiment	Description	Total reads	Mapped reads (%)	Small RNA set (19–25 nt) (%)	Expected FP in small RNAs (%)
GEO:GSM639446	*Tribolium* adult RNA library	67 070 132	45 838 144 (68.34)	21 547 990 (32.13)	39 267 (0.06)
GEO:GSM639447	*Tribolium* embryo RNA library	52 620 004	30 382 986 (57.74)	14 120 332 (26.83)	35 162 (0.07)
SRA:SRR039230	*Apis* RNA library	36 796 459	28 909 134 (78.57)	15 204 339 (41.32)	19 059 (0.05)

3.3 Efficient mapping of color space reads

We implemented our integrated sequential trimming and mapping strategy in a simple script (see Section 2.2). This script maps reads using Bowtie, but it modifies the input file so that Bowtie keeps the first color of the read. We have used this approach to reanalyse three published SOLiD datasets. We first consider two *T.castaneum* RNA libraries from adult and embryos, sequenced in our laboratory (Marco *et al.*, 2010). In our previous analyses, we first converted colors to bases directly, and mapped the sequences in base space. We obtained mapping rates to the reference genome of only ∼13% of the adult reads and <4% of the embryonic reads (Marco *et al.*, 2010). When we applied the strategy described here, we mapped ∼68 and 58% for adults and embryos, respectively (Table 2). The direct conversion from color to base space produces a high proportion of artifactual sequences (because a single color mismatch causes errors in every downstream base of the converted sequence). This is the primary explanation for the low number of reads mapped in our previous analyses. We also analyzed a third-party RNA library from honeybee (Chen *et al.*, 2010). In this case, we successfully mapped 79% of the reads. Since our interest is in detecting microRNAs, we consider further only reads mapped at color length 20–26. These sequences account for a large proportion of the total (∼41%).

We estimated the number of expected false positive mappings that passed our criteria. Calculating the frequency of a given word in a genome is a known problem that often requires the use of distribution approximations (Robin and Schbath, 2001). We used a simpler approach to estimate the number of reads that map by chance to the genome, assuming no sequence bias composition and no effect of overlapping words. Consider the probability that a read maps at exactly one position in the genome by chance, that is a function of the number of mapping positions (approximately the length of the genome, L, multiplied by the number of colors) and the number of potential different sequences (4^l, where l is the read length). Hence, the average number of sequences mapped by chance to the genome in 5 or fewer sites (since we discarded reads mapping to >5 positions in our analysis), M, is given by:

$$M \approx N \sum_{i=1}^{5} \left[\frac{4L}{4^l} \right]^i \quad (2)$$

where N is the number of sequences to be mapped in each step. As shown in Table 2, the number of expected false positives (expected number of sequences mapped by chance) is, in all cases, <0.1%.

Chen *et al.* (2010) used RNA2MAP to analyze their honeybee dataset. RNA2MAP first maps the reads to known microRNAs at miRBase, and the remaining reads are mapped against the reference genome. In order to compare our results with those published previously, we followed a similar approach with our pipeline. When mapping to previously known microRNAs, we observed that both mapping strategies yielded similar results (Fig. 2A). However, we observed that for some microRNAs we map many more sequences than RNA2MAP. For example, the authors detected 546 reads that map to mir-1, while our strategy successfully mapped >300 000 sequences. We note that there are two identical copies of mir-1 in the honeybee genome: ame-mir-1-1 and ame-mir-1-2. We also map many more reads than the previous study for other multiple copy sequences, for example mir-92b and mir-87 (Fig. 2A). It is important to keep reads with multiple matches in the genome, since they may map to real microRNAs. Additionally, our sequential trimming strategy clearly outperforms RNA2MAP-based mapping when mapping into the reference genome (Fig. 2B). We conclude that the ability of Bowtie to deal with multiple mapped reads outperforms that of RNA2MAP, and our sequential trimming strategy permits the removal of highly degraded linkers that escape RNA2MAP.

Reads mapped to the genome by our sequential trimming approach can be used to detect microRNAs with in-house built tools, or can easily be converted to input files for popular microRNA detection tools such as miRDeep (Friedlander *et al.*, 2008) for animal microRNAs or miRCat (Moxon *et al.*, 2008) for plants.

Using the mapping results of our sequential trimming procedure, we reanalyze the honeybee small RNA dataset for new microRNAs. Using strict detection criteria (see Section 2.3), we detected eight new microRNAs, one of them (ame-mir-2765) with known homologs in other species (Table 3). All but one have a relatively low number of reads (Table 3). We also reanalyzed our own *T.castaneum* datasets (Marco *et al.*, 2010). We detected 14 additional new microRNAs that escaped our earlier annotation (Table 3, Supplementary File 1 in Supplementary Material). Four out of the 14 new microRNAs in *Tribolium* (tca-mir-6007, tca-mir-6016, tca-mir-927b, tca-mir-9e) were detected because of our improved strategy to characterize microRNAs (as described in Section 2.3). However, the other 10 detections were only possible in color-space, mostly because low abundance reads from one of the arms did not map in base space due to sequencing errors.

4 CONCLUSIONS

Our exploration of the consequences of detecting microRNAs in color space can be summarized in two recommendations, independent of the particular software used for small RNA mapping. First, be aware of how your program of choice is dealing with the first color of the read. Second, do not rely on simple pattern match approaches to remove linker sequences. The sequential trimming approach discussed here, or seed mapping and extension, are likely to result in significantly higher mapping rates. We do not recommend

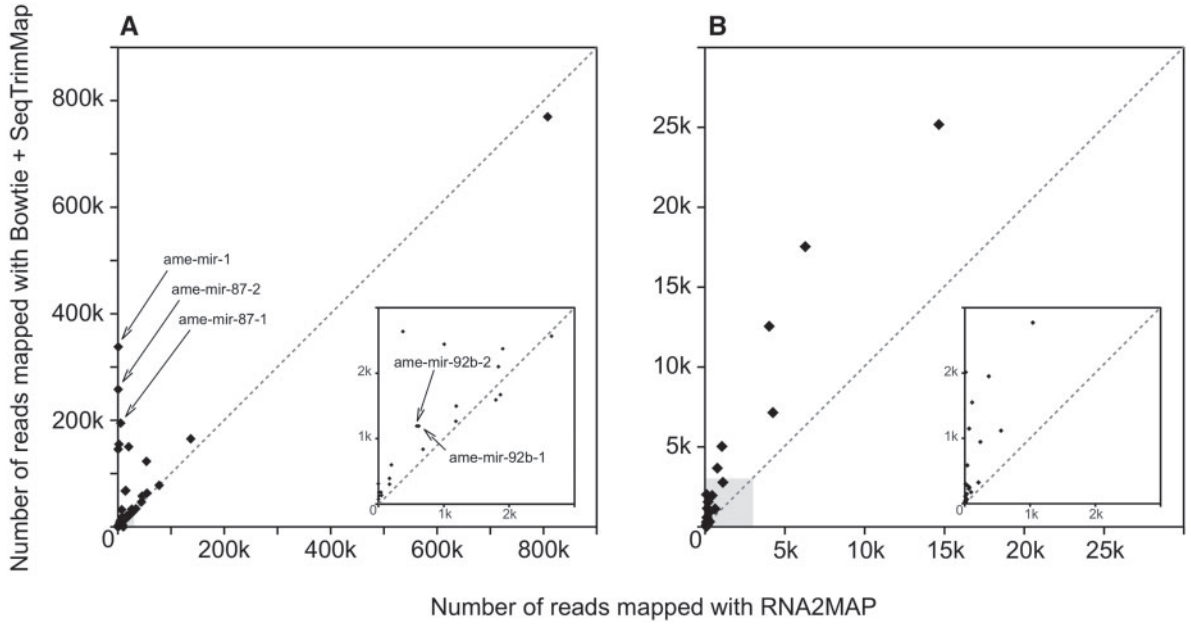

Fig. 2. Comparison of reads mapped with RNA2MAP and a sequential trimming strategy. (**A**) Reads mapped using both strategies to known microRNAs in miRBase. (**B**) Reads mapped for both strategies to newly discovered microRNAs by Chen *et al.* (2010). The inset in both graphs shows a zoomed view of the shaded area.

Table 3. Novel microRNAs discovered in *Apis mellifera* and *T.castaneum*

Name	Chr	Str	Start	End	Reads
ame-mir-6000	LG11	−	1 144 1637	11 441 734	48
ame-mir-6001	LG13	−	2 650 488	2 650 555	2973
ame-mir-6002	LG16	−	2 944 352	2 944 431	15
ame-mir-6003	LG2	−	705 902	7 059 571	35
ame-mir-6004	LG5	+	7 573 377	7 573 464	26
ame-mir-6005	LG6	−	6 509 945	6 510 107	180
ame-mir-6006	LG9	−	62 686	62 773	19
ame-mir-2765	LG9	+	5 203 815	5 203 903	162
tca-mir-6007	CHR1	+	8 492 377	8 492 463	867
tca-mir-6008	CHR2	+	14 782 252	14 782 347	27
tca-mir-6009	CHR2	+	186 871	186 960	44
tca-mir-6010	CHR2	+	11 831 158	11 831 242	33
tca-mir-6011	CHR3	−	31 219 647	31 219 723	41
tca-mir-6012	CHR3	−	9 600 253	9 600 425	258
tca-mir-6013	CHR4	+	11 124 335	11 124 402	17
tca-mir-6014	CHR4	+	3 369 897	3 369 978	46
tca-mir-6015	CHR4	+	11 485 945	1 1486 036	23
tca-mir-6016	CHR7	−	17 010 368	17 010 442	57
tca-mir-6017	CHR7	−	10 450 348	10 450 502	252
tca-mir-6018	CHR8	−	247 709	247 801	42
tca-mir-927b	CHR9	−	16 099 288	16 099 383	224
tca-mir-9e	CHR9	−	11 062	11 172	3696

Chr, Chromosome/linkage group; Str, strand; start, first nucleotide position; end, last nucleotide position; reads: total number of reads.

cropping the read sequences to fixed length since this will create both false positives and false negatives. For example, a 50 nt read that maps to the genome is very unlikely to be a microRNA, yet cropping to 25 nt before mapping could cause an erroneous microRNA annotation. Deep sequencing was originally devised for high coverage of relatively long sequences. The application of such techniques to the detection of small RNAs highlights new issues and biases. Characterizing these issues is crucial for the development of more efficient and biologically congruent mapping tools.

ACKNOWLEDGEMENTS

We would like to thank Ana Kozomara, Matthew Ronshaugen, Ian Donaldson, Max Haussler, Dave Gerrard, Aaron Webber and three anonymous reviewers for helpful comments and stimulating discussion, and Franziska Bonath and Leo Zeef for testing the bash script. We also thank Ben Langmead for detailed discussion of Bowtie functionality.

Funding: Biotechnology and Biological Sciences Research Council (BB/G011346/1); University of Manchester (fellowship to S.G.-J.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 338–402.

Applied Biosystems (2008) Principles of Di-base Sequencing and the Advantages of Color Space Analysis in the SOLiD System. *Report 139AP 10-01*.

Applied Biosystems (2009) *SOLiD3 System Application Documentation Small RNA Analysis Tool*. Available: http://solidsoftwaretools.com/gf/

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 28–97.

Berezikov,E. *et al.* (2011) Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome Res.*, **21**, 203–215.

Cai,Y. *et al.* (2010) Novel microRNAs in silkworm (Bombyx mori). *Funct. Integr. Genomics*, 10, 405–415.

Chen,X. *et al.* (2010) Next-generation small RNA sequencing for microRNAs profiling in the honey bee Apis mellifera. *Insect Mol. Biol.*, **19**, 799–805.
Cloonan,N. *et al.* (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. *Bioinformatics*, **25**, 2616–2615.
David,M. *et al.* (2011) SHRiMP2: sensitive yet practical short read mapping. *Bioinformatics*, **27**, 1011–1012.
Flicek,P. and Birney,E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Methods*, **6**, S6–S12.
Friedlander,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
Goff,L.A. *et al.* (2009) Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors. *PLoS One*, **4**, e7192.
Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Mon. Chem.*, **125**, 167–188.
Homer,N. *et al.* (2009) Local alignment of two-base encoded DNA sequence. *BMC Bioinformatics*, **10**, 175.
Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics*, **11**, 473–483.
Li,S.-C. *et al.* (2010) Discovery and characterization of medaka miRNA genes by next generation sequencing platform. *BMC Genomics*, **11** (Suppl. 4), S8.
Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
Marco,A. *et al.* (2010) Functional shifts in insect microRNA evolution. *Genome Biol. Evol.*, **2**, 686–696.
Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.
Morin,R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.
Moxon,S. *et al.* (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, **24**, 2252–2253.
Robin,S. and Schbath,S. (2001) Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comput. Biol.*, **8**, 349–359.
Ruby,J.G. *et al.* (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.*, **17**, 1850–1864.
Rumble,S.M. *et al.* (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, **5**, e1000386.
Sasson,A. and Michael,T.P. (2010) Filtering error from SOLiD output. *Bioinformatics*, **26**, 849–850.