BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 123–124
doi:10.1093/bioinformatics/btp601

Sequence analysis

MicroRazerS: rapid alignment of small RNA reads

Anne-Katrin Emde[1,2,*,†], Marcel Grunert[3,*,†], David Weese[1], Knut Reinert[1] and Silke R. Sperling[3]

[1]Department of Computer Science, Free University of Berlin, Takustr. 9, [2]International Max Planck Research School for Computational Biology and Scientific Computing and [3]Group Cardiovascular Genetics, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

Received on June 19, 2009; revised on September 27, 2009; accepted on October 14, 2009

Advance Access publication October 29, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Deep sequencing has become the method of choice for determining the small RNA content of a cell. Mapping the sequenced reads onto their reference genome serves as the basis for all further analyses, namely for identification and quantification. A method frequently used is Mega BLAST followed by several filtering steps, even though it is slow and inefficient for this task. Also, none of the currently available short read aligners has established itself for the particular task of small RNA mapping.

Results: We present MicroRazerS, a tool optimized for mapping small RNAs onto a reference genome. It is an order of magnitude faster than Mega BLAST and comparable in speed with other short read mapping tools. In addition, it is more sensitive and easy to handle and adjust.

Availability: MicroRazerS is part of the SeqAn C++ library and can be downloaded from http://www.seqan.de/projects/MicroRazerS.html.

Contact: emde@inf.fu-berlin.de; grunert@molgen.mpg.de

1 INTRODUCTION

MicroRNAs (miRNAs) are short, single-stranded RNA molecules, ranging from 19 to 25 nt in length, which regulate expression of target genes and thereby play an essential role in many biological processes. Until now only a limited number of small RNAs have been characterized in depth (Kawaji and Hayashizaki, 2008). With the invention of high-throughput sequencing technologies (e.g. Solexa/Illumina), we are now able to explore genomes and RNA transcriptomes with unprecedented depth of coverage, thereby enabling comprehensive insight into the miRNA content of a cell. For functional annotation of small RNAs, the reads resulting from deep sequencing have to be mapped to the reference genome. By determining the number of reads that map to annotated miRNA genes, the abundance of known miRNAs can be measured. Inspecting clusters of non-annotated but mapped sequences has great potential for detecting novel miRNAs or other small non-coding RNAs. Although specific short read mapping software exists, recent large-scale studies (Friedländer *et al.*, 2008; Morin *et al.*, 2008) have used the less sensitive and very time-consuming Mega BLAST algorithm (Zhang *et al.*, 2000). This is due to the special requirements of small RNA read mapping. Usually, a high quality 5′ end with an exactly matching seed sequence and trailing mismatches at the 3′ end is expected. As small RNAs may be shorter than the sequenced reads, the sequencing process can reach into the adapter. As a consequence, the 3′ ends of the reads may contain variable lengths of adapter sequence causing mismatches in the read-to-reference alignment. If the adapter sequence is known, the 3′ ends can be trimmed, but this process is imperfect and complicated by the presence of sequencing errors occurring especially at the 3′ end.

A common strategy is to search for the longest possible prefix-match of each read, i.e. the longest contiguous match starting at the first read base. Mega BLAST aligns all reads to the genome with a minimum word size. The output then needs to be filtered for matches meeting the above criteria. This means discarding all matches with <100% identity in the 5′ seed sequence and afterwards only retaining the longest match(es) for each read (Friedländer *et al.*, 2008; Morin *et al.*, 2008). The resulting set of matches usually constitutes only a small fraction of the original Mega BLAST output. This strategy is unnecessarily slow and unhandy. To our knowledge, there is no short read aligner that directly implements this strategy. However, tools employing similar strategies exist, like the recently developed BWT-based aligners SOAP2 (Li *et al.*, 2009) and Bowtie (Langmead *et al.*, 2008), which allow the user to set a minimum 5′ seed length.

We therefore developed a read mapping tool specifically tailored to the needs of short RNA mapping. It is robust to possible adapter sequence at the 3′ end of a read and requires no adapter trimming. It can map millions of reads within a few minutes and is not only much easier to handle than Mega BLAST, but also more sensitive, especially in the presence of sequencing errors and SNPs. Moreover, no extensive filtering after mapping is required.

2 ALGORITHM

MicroRazerS is a special version of the general purpose short read mapping tool RazerS (Weese *et al.*, 2009) and is implemented within the C++ library SeqAn (Döring *et al.*, 2008). It is based on a q-gram counting strategy that builds an index over the reads and uses an implementation of the Swift filter algorithm (Rasmussen *et al.*, 2006) to scan over the reference and efficiently filter regions containing possible read matches (see Weese *et al.*, 2009, for detailed information). These regions are identified by a certain minimal

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Evaluation of small RNA mapping tools

	MicroRazerS	BLAST	SOAP2	Bowtie
Running time (min)	24	194	6	5
Building index (min)	–	–	84	206
Output size (GB)	0.1	8.6	6.8	0.7
Memory usage (GB)	3.4	1.4	8.3	2.3
Unique sequences aligned	1 319 218	891 215	1 318 504	1 184 590
Mappable reads	7 743 516	7 001 832	7 742 266	7 410 239
Reads annotated as miRNA	5 819 189	5 746 588	5 819 184	5 667 027
MiRNAs (read count >150)	381 101	372 96	381 101	372 99

We used a query dataset of ∼2.4 M non-redundant read sequences (length 36 bp) representing a total of ∼9.3 M reads. Using MicroRazerS the parameters were set as follows: -m 20 (maximum number of best matches), -pa (purge ambiguous reads having more than 20 equally best hits) and -sL 16 (seed length). A seed length of 16 bp (100% identity) was used for all mapping tools. Searching for miRNAs with a length between 19–25 nt, we found reads with a minimal length of 16 nt to be good seeds for read mapping. In the case of MicroRazerS, we allowed no mismatch in the read prefix. For SOAP2, we allowed 20 mismatches in one read but only exact matches in the seed part of read. For Bowtie, a quality cutoff '-e 500' was used (which corresponds to allowing 20 mismatches, as each base quality in all reads was set to Phred score quality 25). The resulting alignments except those from MicroRazerS were filtered to get the best (longest) hits with at most 20 positions in the human genome.

number t of q-grams, short subsequences of length q which are shared between read and reference. Filter efficiency is determined by the parameters q and t. For read-reference alignments that are not allowed to have gaps, i.e. if only Hamming distance mapping is considered, filter sensitivity can be strongly increased by using gapped q-grams (subsequences containing 'don't care' positions). MicroRazerS employs this gapped q-gram method in conjunction with the Swift parallelogram filter to detect with 100% sensitivity all read matches for which a read prefix of length s contains 0 or 1 mismatch. We call s the seed length. Seed matches are extended to the right (3′ end) until the first mismatch is encountered. MicroRazerS then guarantees to find for each read the match that has (i) the lowest number of mismatches in the seed and (ii) can be extended furthest to the right. If multiple best matches exist, all of them are detected. Note that the balance between speed and sensitivity can be controlled by the recognition rate parameter (default 100). The higher the recognition rate the more sensitive is MicroRazerS. The lower the recognition rate the faster runs the tool.

MicroRazerS supports seed length values from 10 to 26, a parameter that can be adjusted via the command line. Allowing an error in the seed can be switched on and off. If multiple best matches exist, a user-defined maximum number of hits is reported, optionally discarding all reads having more best hits than this number.

3 RESULTS AND DISCUSSION

Small RNAs were isolated from human normal heart RNA and prepared for Solexa sequencing. Deep sequencing of the small RNA library produced 9 286 222 sequenced single-end reads of 36 bases in length, yielding 2 402 361 unique (i.e. non-redundant) read sequences. These unique reads were mapped to the human genome (NBCI build 36.1) using Mega BLAST, SOAP2, Bowtie and MicroRazerS.

The mapping results of all programs are shown in Table 1. The running time was measured on an AMD Opteron 2384 with 32 GB memory running a 64-bit Linux system. In our test setting, MicroRazerS is nine times (170 min) faster than Mega BLAST and 20 min slower than SOAP2 or Bowtie. However, SOAP2 took 84 min and Bowtie 206 min to build a BWT index for the human reference genome. Moreover, Mega BLAST and SOAP2 produce huge output files that need to be filtered, taking in both cases ∼30 min of post-processing and decreasing output file size down to a similar size as observed for MicroRazerS output.

To annotate the sequence reads with known miRNAs, we checked for overlaps with positions annotated by the miRBase database (release 13.0). Of note, MicroRazerS is able to map a higher number of reads than all other programs. While in this dataset almost no differences in miRNA predictions between SOAP2 and MicroRazerS were observed, the slightly lower sensitivity of SOAP2 could lead to missing miRNA measurement in other datasets.

An additional feature of MicroRazerS is its -sE option that allows to map reads with at most one error in the seed sequence. Especially if one is interested in finding SNPs or miRNAs at low abundance where robustness toward sequencing errors might be crucial, the 100% identity criterium has to be dropped. Indeed, we observe that a higher number of reads can be annotated as miRNAs when one error is allowed. Using these options, MicroRazerS mapped 97% of all unique sequences to the human genome representing 99% of the total reads, resulting in 414 known miRNAs.

In conclusion, the results suggest that MicroRazerS can substantially facilitate the profiling and discovery of miRNAs obtained from high-throughput sequencing.

ACKNOWLEDGEMENTS

We gratefully acknowledge Ilona Dunkel for small RNA library preparation and the German Heart Center Berlin for providing the sample material.

Funding: European Community's Sixth Framework Program contract ('HeartRepair') LSHM-CT-2005-018630.

Conflict of Interest: none declared.

REFERENCES

Döring,A. *et al.* (2008) SeqAn - an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

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

Kawaji,H. and Hayashizaki,Y. (2008) Exploration of small RNAs. *PLoS Genet.*, **4**.

Langmead,B. *et al.* (2008) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**.

Li,R. *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Morin,R. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.

Rasmussen,K. *et al.* (2006) Efficient q-gram filters for finding all epsilon-matches over a given length. *J. Comput. Biol.*, **13**, 296–308.

Weese,D. *et al.* (2009) RazerS - fast read mapping with sensitivity control. *Genome Res.*, **19**, 1646–1654.

Zhang,Z. *et al.* (2000) A greedy algorithm for aligning dna sequences. *J. Comput. Biol.*, **7**, 203–214.