**Sequence analysis**

**MicroRazerS: rapid alignment of small RNA reads**

Anne-Katrin Emde1,2,*,†, Marcel Grunert3,*,†, David Weese1, Knut Reinert1, and Silke R. Sperling3

1Department of Computer Science, Free University of Berlin, Takustr. 9, 2International Max Planck Research School for Computational Biology and Scientific Computing and 3Group Cardiovascular Genetics, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnesstr. 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 been characterized in depth (Kawaji and Hayashizaki, 2008). 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
Evaluation of small RNA mapping tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Running time (min)</th>
<th>Building size (index size)</th>
<th>Memory usage (GB)</th>
<th>Unique sequences aligned</th>
<th>Mappable reads</th>
<th>Reads annotated as miRNA</th>
<th>MiRNAs (read count &gt;150)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MicroRazerS</td>
<td>24</td>
<td>–</td>
<td>3.4</td>
<td>1 319 218</td>
<td>7 743 516</td>
<td>5 819 189</td>
<td>381</td>
</tr>
<tr>
<td>BLAST</td>
<td>194</td>
<td>–</td>
<td>8.6</td>
<td>891 215</td>
<td>7 001 832</td>
<td>5 746 588</td>
<td>372</td>
</tr>
<tr>
<td>SOAP2</td>
<td>6</td>
<td>–</td>
<td>6.8</td>
<td>1 318 504</td>
<td>7 742 266</td>
<td>5 819 184</td>
<td>381</td>
</tr>
<tr>
<td>Bowtie</td>
<td>5</td>
<td>–</td>
<td>8.3</td>
<td>1 184 590</td>
<td>7 410 239</td>
<td>5 667 027</td>
<td>372</td>
</tr>
</tbody>
</table>

We used a query dataset of ~2.4 M non-redundant read sequences (length 50 bp) representing a total of ~9.3 M reads. Using MicroRazerS the parameters were set as follows: -m (seed length) 20, -t (read length) 16, -a (max mismatches) 100, -u (unique reads) 1, -sL 20 (seed length). A seed length of 16 bp (100% identity) was used for all mapping tools. Searching for miRNAs with a length between 19–27 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 our 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.

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 ‘$e$’ 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


