Genome analysis

Assembling millions of short DNA sequences using SSAKE

René L. Warren*, Granger G. Sutton1, Steven J. M. Jones and Robert A. Holt

British Columbia Cancer Agency, Genome Sciences Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada and 1J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

Received on October 6, 2006; revised on November 15, 2006; accepted on December 5, 2006
Advance Access publication December 8, 2006
Associate Editor: Alex Bateman

ABSTRACT
Summary: Novel DNA sequencing technologies with the potential for up to three orders magnitude more sequence throughput than conventional Sanger sequencing are emerging. The instrument now available from Solexa Ltd, produces millions of short DNA sequences of 25 nt each. Due to ubiquitous repeats in large genomes and the inability of short sequences to uniquely and unambiguously characterize them, the short read length limits applicability for de novo sequencing. However, given the sequencing depth and the throughput of this instrument, stringent assembly of highly identical sequences can be achieved. We describe SSAKE, a tool for aggressively assembling millions of short nucleotide sequences by progressively searching through a prefix tree for the longest possible overlap between any two sequences. SSAKE is designed to help leverage the information from short sequence reads by stringently assembling them into contiguous sequences that can be used to characterize novel sequencing targets.

Availability: http://www.bcgsc.ca/bioinfo/software/ssake
Contact: rwarren@bcgsc.ca

1 INTRODUCTION

High-throughput DNA sequencing instrumentation capable of producing tens of millions of short (~25 bp) sequences (reads) is becoming available (Bennett, 2004). The two most striking attributes of this technology, the large read depth and short sequence length, make it suitable for re-sequencing applications where a known reference sequence is used as a template for alignment. However, the ability to decode novel sequencing targets, such as unsequenced genomes or metagenomic libraries is limited. Twenty-five mers are far more ubiquitous than Sanger-size reads (500–1000 bp) in any given genome. Since the sequence complexity increases by a factor 4 for every base added, the likelihood of observing redundant sequences increases dramatically with decreased read length for sequences shorter than 20 bp. The read length needed to achieve maximal uniqueness varies depending on the genome being sequenced, its size and repeat content (Whiteford et al., 1977), coronavirus SARS TOR2 (Marra et al., 2003), bacteria Haemophilus influenzae (Fleischmann et al., 1995) genomes and on 40 million 25mers from the whole-genome shotgun (WGS) sequence data from the Sargasso sea metagenomics project (Venter et al., 2004). Our results indicate that SSAKE could be used for complete assembly of sequencing targets that are 30 kb in length (e.g. viral targets) and to cluster millions of identical short sequences from a complex microbial community.

2 METHODS

2.1 Material

The PhiX174, SARS TOR2 and H.influenzae genomes were downloaded from GenBank (GenBank identifier J02482,AY274119 and L42023, respectively). All possible 25mers were extracted from both strands for these genomes. Sequences were selected at random to simulate up to 400x read coverage for the viral genomes and up to 100x read coverage for H.influenzae. Forty million 25mers were selected at random from the Sargasso Sea WGS metagenomics data obtained from the Venter Institute (https://research.venterinstitute.org/sargasso/).

2.2 SSAKE algorithm

DNA sequences in a single multi fasta file are read in memory, populating a hash table keyed by unique sequence reads with values representing the number of occurrences of that sequence in the set. A prefix tree is used to organize the sequences and their reverse-complemented counterparts by their first eleven 5’ end bases. The DNA prefix tree is used to limit the K-mer search and 3’ read Extension (SSAKE) program cycles through sequence data stored in a hash table, and progressively searches through a prefix tree for the longest possible k-mer between any two sequences. We ran the algorithm on simulated error-free 25mers from the bacteriophage PhiX174 (Sanger et al., 1977), coronavirus SARS TOR2 (Marra et al., 2003), bacteria Haemophilus influenzae (Fleischmann et al., 1995) genomes and on 40 million 25mers from the whole-genome shotgun (WGS) sequence data from the Sargasso sea metagenomics project (Venter et al., 2004). Our results indicate that SSAKE could be used for complete assembly of sequencing targets that are 30 kb in length (e.g. viral targets) and to cluster millions of identical short sequences from a complex microbial community.

*To whom correspondence should be addressed.
when a $k$-mer matches the $5'$ end of more than one sequence read ($\sim s$). This leads to shorter contigs, but minimizes sequence misassemblies. The second is to stop the extension when a $k$-mer is smaller than a user-set minimum word length ($m$). SSAKE outputs a log file with run information along with two multi fasta files, one containing all sequence contigs constructed and the other containing the unassembled sequence reads.

### 3 RESULTS

SSAKE assembly of 4208 PhiX174 reads took 0.84 s on a single 2.2 GHz two dual-core CPU AMD Opteron™ computer with 4 GB RAM and yielded a single contig bearing 100% sequence identity (sum of identical base matches between two sequences divided by the contig length) with the PhiX174 genome (Table 1). On the same hardware, we were able to assemble the SARS-associated coronavirus de novo into a single contig having 99.91% sequencing identity with the genome. The read coverage needed to achieve this was 20 times higher than for PhiX174. Increased coverage was needed to insure only one valid path could be taken to assemble all reads. Assembly of *H.influenzae* reads was impaired by the presence, in the genome, of 28 perfectly repeated segments ranging in size from 70 to 5723 bases and 29 766 repeated 25mers. At best, we were able to assemble 7.3 million sequence reads into 284 contigs equal or larger than 75 bp and totaling 1.78 Mb. Of these contigs, 241 showed single, unique, full-length alignments to *H.influenzae*, and covered 1007 kb (54.62% of the genome) with 99.43% sequence identity. The remaining 43 contigs totaled 776 kb and all incorporated $k$-mers that mapped to repeats, causing broken alignments between the contigs and the genome.

Forty million 25mers generated at random from Sargasso Sea genome shotgun Sanger-reads (Venter et al., 2004) were assembled using $m=16 \sim s=0$, SARS using $m=15 \sim s=0$, *H.influenzae* using $m=16 \sim s=1$ and Sargasso Sea using $m=16 \sim s=0$.

<table>
<thead>
<tr>
<th>Species (size bp)</th>
<th>Input random 25mers</th>
<th>Coverage</th>
<th>Run time (s)</th>
<th>Contig N50 length (bp)</th>
<th>Genome covered (%)</th>
<th>Mean sequence identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhiX-174 (5386)</td>
<td>4208</td>
<td>20</td>
<td>0.84</td>
<td>5382</td>
<td>99.92</td>
<td>100</td>
</tr>
<tr>
<td>SARS TOR2 (29 751)</td>
<td>476016</td>
<td>400</td>
<td>45.13</td>
<td>29 744</td>
<td>99.98</td>
<td>99.91</td>
</tr>
<tr>
<td><em>H.influenzae</em> (1 830 023)$^a$</td>
<td>7 316 203</td>
<td>100</td>
<td>580.53</td>
<td>22 230</td>
<td>54.62</td>
<td>99.43</td>
</tr>
<tr>
<td>Sargasso Sea metagenome</td>
<td>40 000 000</td>
<td>NA</td>
<td>9.2E + 4</td>
<td>423</td>
<td>NA</td>
<td>92.29</td>
</tr>
</tbody>
</table>

Assembly of 40 M Sargasso Sea 25mers was done on a single 4x 1.4 GHz AMD Opteron™ CPU with 32 GB RAM.

**Table 1.** Short read assembly of PhiX174, SARS TOR2 and *H.influenzae* genomes using SSAKE on a single 2x 2.2 GHz dual-core AMD Opteron™ CPU with 4 GB RAM.

CONCLUSION

We have shown that with high-sequencing depth, short sequences can be used for *de novo* assembly of small DNA targets (e.g. viral genomes) that are up to 10’s of kb in length. For larger and more complex sequencing targets, such as bacterial genomes, short reads can be rapidly and stringently assembled into contigs that accurately represent the non-repetitive portion of the genome. It is clear that the best approach for *de novo* sequencing of targets more complex than viral genomes will likely involve some combination of Sanger reads and assembled short reads. For metagenomics, our simulation involving 40 M short reads from the Sargasso Sea WGS data indicate that these types of reads can be used to produce conservative contigs in a robust and tractable manner, while minimizing probabilistic errors. As a stringent, efficient assembly tool SSAKE is expected to have broad application in *de novo* sequencing.

ACKNOWLEDGEMENTS

The authors thank Martin Krzywinski for his insights on efficient $k$-mer search. S.J.MJ and R.A.H. are Michael Smith Foundation for Health Research Scholars. Funding to pay the Open Access publication charges for this article was provided by the British Columbia Cancer Agency.

Conflict of Interest: none declared.

REFERENCES


