Genome analysis

Extending assembly of short DNA sequences to handle error

William R. Jeck[1,*], Josephine A. Reinhardt[1], David A. Baltrus[1], Matthew T. Hickenbotham[2], Vincent Magrini[2], Elaine R. Mardis[2], Jeffery L. Dangl[1,3] and Corbin D. Jones[1,3]

[1]Department of Biology, University of Carolina—Chapel Hill, Chapel Hill, NC 27599, [2]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108 and [3]Carolina Center for Genome Sciences, University of Carolina—Chapel Hill, Chapel Hill, NC 27599, USA

ABSTRACT

Inexpensive *de novo* genome sequencing, particularly in organisms with small genomes, is now possible using several new sequencing technologies. Some of these technologies such as that from Illumina's Solexa Sequencing, produce high genomic coverage by generating a very large number of small reads (~30 bp). While prior work shows that partial assembly can be performed by *k*-mer extension in error-free reads, this algorithm is unsuccessful with the sequencing error rates found in practice. We present VCAKE (Verified Consensus Assembly by K-mer Extension), a modification of simple *k*-mer extension that overcomes error by using high depth coverage. Though it is a simple modification of a previous approach, we show significant improvements in assembly results on simulated and experimental datasets that include error.

Availability: http://152.2.15.114/~labweb/VCAKE

Contact: william.jeck@gmail.com

1 INTRODUCTION

Rapid and inexpensive sequencing of genomes is now possible via new high-throughput sequencing technologies that produce large numbers of small sequencing reads (Sundquist *et al.*, 2007). Theoretical and computational work suggests that *de novo* assembly of small genomes is possible using these technologies (Warren *et al.*, 2007; Whiteford *et al.*, 2005). For example, the SSAKE assembler can assemble a viral genome using error-free short read sequences. However, the two leading short read sequencing technologies, Illumina and 454 Life Sciences, show appreciable error rates in their reads (Bentley, 2006). A viable *de novo* assembler using these technologies must account for this error. With a purely greedy approach to *k*-mer extension, erroneous bases will be incorporated into the assembly with a frequency equal to the error rate. We exhibit that even low error rates result in significantly poorer assembly for SSAKE.

Our algorithm, VCAKE (Verified Consensus Assembly by K-mer Extension), makes significant improvements in handling error. VCAKE extends seed sequences using a *k*-mer extension method much like SSAKE, where sequences are efficiently found from a hash table. Rather than using greedy extension or simply using Q values, which are too poorly assigned to result in adequate assembly, VCAKE considers all reads overlapping with the seed sequence. VCAKE extends the seed sequence one base at a time using the most commonly represented base from these matching reads, provided that the set of reads passes certain conditions.

Using actual and simulated data, we show that VCAKE behaves well with short reads with and without error. Our results imply that complete assembly of viral genomes and partial assembly of bacterial genomes may be possible using presently available short read technology even with prevalent error and in taxa far diverged from any reference.

2 METHODS

2.1 Material

Whole genome sequences for SARS-TOR2 and *Pseudomonas syringae* pv. tomato str. DC3000 (DC3000) came from GenBank (AY274119 and AE016853, respectively). Random 30mer simulated reads were generated at different coverage levels for each whole genome sequence. For those simulations including errors, each base of the sequences was randomly and independently changed to another base with a fixed probability. Three lanes of Solexa sequencing of DC3000 whole DNA were also provided (Baltrus *et al.*, unpublished data) and used for test assembly. All lanes combined were 202 884 352 bases in 6 340 136 reads, making a coverage depth of 31.7*x*.

2.2 VCAKE algorithm

Initially, the VCAKE assembly process is nearly identical to SSAKE, except that two multi-FASTA files separately populate *bin* and *set* hash tables from a pool of reads. Divergence from the SSAKE method occurs during extension of the seed sequences from *set*. VCAKE finds all exact *k*-mer matches of the 3′ end of the sequence up to a user-defined minimum *n*. The first 11 bases of the *k*-mer are used to efficiently search *bin* and all returned keys are checked against the remainder of the *k*-mer. Those matching perfectly are pushed into an array a number of times equal to the values keyed by that sequence in *bin*. As in SSAKE, the *bin* hash table consists of a treed hash table keyed by the first 11 bases of the read (or its reverse complement) followed by the sequence itself, with the value containing the number of appearances of that read or its reverse

---

*To whom correspondence should be addressed.

Table 1. Results of VCAKE and SSAKE assemblies of simulated and actual sequencing data on an Itanium2 1.3 GHz processor with 96 GB memory

Organism	Program	Coverage (x)	Error rate (%)	Run time (s)	N50 (bp)	N75 (bp)	Genome covered (%)	Largest contig (bp)	Runtime options
SARS-TOR2	SSAKE	50	0	9.29	29 715	29 715	100	29 715	$-s$ 2 $-m$ 15
SARS-TOR2	VCAKE	50	0	54.2	29 715	29 715	100	29 715	$-t$ 1 $-v$ 1 $-n$ 20 $-m$ 15
SARS-TOR2	SSAKE	200	5	204	0	0	0	0	$-s$ 2 $-m$ 15
SARS-TOR2	VCAKE	200	5	4815	28 332	28 332	100	28 332	$-t$ 12 $-c$.5 $-v$ 5 $-n$ 20 $-m$ 15
DC3000	SSAKE	50	0	2666	0	0	46.5	103 158	$-s$ 1 $-m$ 20
DC3000	VCAKE	50	0	38 585	5563	2845	98.86	31 446	$-v$ 1 $-n$ 20
DC3000	SSAKE	100	1	13 237	67	0	67.6	669	$-s$ 1 $-m$ 20
DC3000	VCAKE	100	1	118 021	7494	3875	98.89	34 229	$-n$ 23 $-m$ 20 $-v$ 5 $-t$ 12
DC3000	SSAKE	3 lanes Solexa	~3%	5796	0	0	31	219	$-s$ 1 $-m$ 20 $-l$ 30
DC3000	VCAKE	3 lanes Solexa	~3%	28 338	249	84	97.5	2749	$-n$ 19 $-m$ 16 $-v$ 10 $-t$ 5

Performance of VCAKE and SSAKE on a virus and a bacteria using real and simulated input. All data shown are for contigs generated that were longer than a single read and matched the relevant reference genome perfectly. The N50 and N80 figures mark the maximum number of base pairs for which all contigs greater than or equal to that threshold covered 50% or 80% of the genome, respectively. Solexa sequencing data was used without filtering, and showed ~3% error.

complement, allowing efficient searching. Having reached the minimum k-mer length n, if the total matching sequence occurrences (repetition of reads included) is less than a user-defined value, t, then the algorithm will proceed further. In this case, VCAKE extracts all k-mer matches up to a lower user-defined length m. If t reads are still not found, then k-mer matches up to user-defined length, e, are considered. This last group may have one mismatch in overlaps with the k-mer after the first 11 bases. This last procedure halts when t total sequence occurrences have been found or the minimum overlap, e, is reached. If a given sequence matches two different k-mer frames, then contig extension on that side is terminated.

The array of matching sequences is then considered. Each sequence offers a 'vote' for the first overhanging base called by that read. The votes are totaled and the base exceeding a threshold of representation, c, is added to the contig. However, the contig will be terminated if the number of reads calling the second most common base exceeds the user-defined threshold, v. This value, v, represents the number of occurrences at which a base call can be considered an indication of duplication of the sequence elsewhere in the genome, rather than due to sequencing error. The user may also define a number of found reads, x, at which the algorithm terminates the contig. This is used when high representation of reads from a given sequence suggests repetitive sequence in the genome. Regardless of contig termination, all sequences retrieved that occur completely within the contig, are deleted both from *bin* and *set*.

Extension proceeds, one base at a time, until no matches are found or the contig is terminated for one of the other reasons above. The contig is then reverse complemented and extension of the opposite end is performed by the same method. The program finally outputs the contig to a file in multi-FASTA format and begins again with an unused seed from *set*.

3 RESULTS

Both SSAKE and VCAKE assemblies of reads without error at $50x$ coverage reproduced 100% of the original SARS-TOR2 genome perfectly (Table 1). In assembly of reads with error, SSAKE showed no coverage by contigs perfectly matching the original genome. VCAKE, in contrast, showed

100% coverage by perfectly matching contigs, of which the largest was 28 332 bp.

For assembly of error-free simulated DC3000 reads at $50x$ coverage, SSAKE produced perfect contigs covering 46.5% of the genome with a longest contig of 103 158 bp. VCAKE showed superior coverage of 98.9%, but a shorter longest contig (31 446 bp). Assembly of reads with an error rate of 1% at higher coverage showed unambiguously superior assembly by VCAKE for those parameters analyzed. VCAKE had a longest contig of 34 229 bp as opposed to SSAKE's 669 bp, and VCAKE covered 98.9% of the genome where SSAKE covered 67.7%.

Using actual Solexa data, VCAKE also showed unambiguously improved assembly, covering 97.5% of the genome as opposed to 31.0% for SSAKE. Of the contigs perfectly matching the reference, VCAKE produced the longest contig (2749 bp versus 219 bp for SSAKE assembly) and a higher N50 of 249 bp. SSAKE assembly of the Solexa data showed no N50, as contigs perfectly matching the genome covered <50% of the reference.

4 CONCLUSION

VCAKE can partially assemble a *de novo* genome from short reads in a range of genomic contexts, error rates and sequencing coverage. VCAKE shows significant improvements compared to SSAKE in situations with error. Our analysis suggests that viral genomes are tractable for *de novo* assembly using current short read sequencing technology with VCAKE. We also believe that VCAKE may be a step towards the assembly of larger bacterial genomes from short reads, particularly with the development of superior base calling or paired end technology.

REFERENCES

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.

Sundquist,A. *et al.* (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE*, **2**, e484.

Warren,R.L. *et al.* (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500–501.

Whiteford,N. *et al.* (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.*, **33**, e171.