BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

SSAKE 3.0

Improved speed, accuracy and contiguity

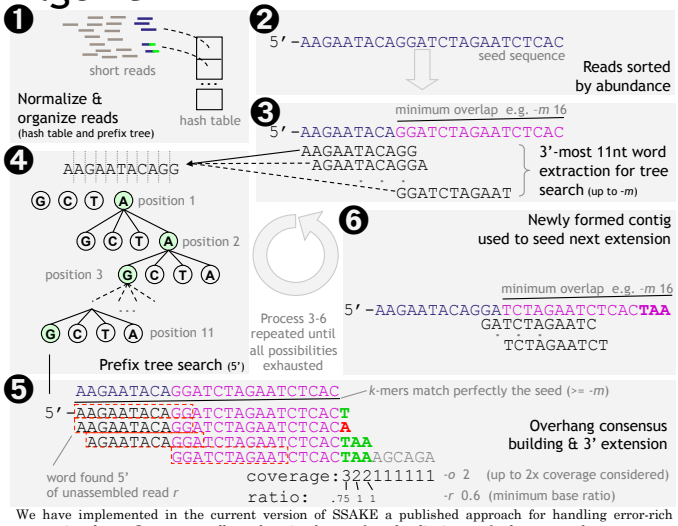
René L. Warren and Robert A. Holt

Canada's Michael Smith
Genome Sciences Centre

Introduction

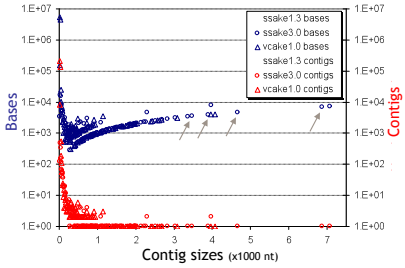
With SSAKE, we demonstrated that *de novo* genome assembly of large contigs using millions of short (25 bp) error free reads is feasible (Warren *et al.* 2006). This work introduced the use of a prefix tree to organize short sequence reads for efficient *k*-mer search and faster assembly speeds. The SSAKE data representation and general algorithm outline with its 3' extension feature is an efficient approach for handling short read data of the type produced by massively parallel sequencing platforms (e.g. Illumina, ABI SOLiD) that has provided the foundation for all subsequently published short read assemblers, including VCAKE and SHARCGS (Jeck *et al.* 2007, Dohm *et al.* 2007). Here we present further improvement made to the SSAKE algorithm, including paired-end read support for building scaffolds.

Algorithm



We have implemented in the current version of SSAKE a published approach for handling error-rich sequencing data. In essence, all overhanging bases of reads aligning perfectly to a seed sequence are considered for extension, using a majority-rule approach for building a consensus sequence of the overhanging bases, much like VCAKE (Jeck *et al*, 2007). However, the SSAKE implementation yields assembly speeds 3 to 5 times faster. SSAKE 3.0 also outperformed VCAKE in contig accuracy and sequence coverage of a reference Human BAC sequence by well-assembled contigs. Compared to the initial release of SSAKE, the current version produces more contiguous and accurate assemblies using real massively parallel sequencing data. (*table below*)

Performance



Using SSAKE 3.0, 435,000 quality-trimmed Illumina sequences providing 75-fold coverage of a Human BAC assembled in little over four minutes and yielded 368 contigs covering 97% of the BAC with 99.3% sequence identity. These contigs comprised 86% of all sequence reads generated. The rest of the reads consisted of linker sequences, low complexity reads and reads crippled with errors.

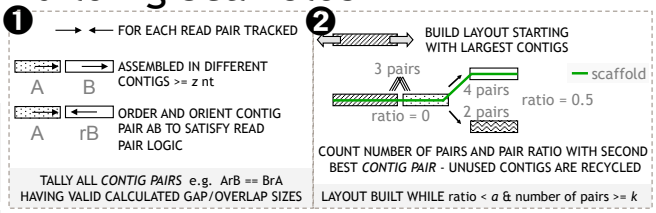
(left) Contig size distribution between SSAKE 1.3, 3.0 and VCAKE 1.0 shows more contiguous assemblies with SSAKE 3.0

De novo assemblies of 435K quality-trimmed Illumina sequences† from a Human BAC

Assembler	Run time (mm:ss)	# Contigs 1. total 2. aligned*	Mean size (nt)	Largest contig (nt)	N50 length (nt)	Coverage* 1. % total 2. % unique¶	Accuracy* (%)
SSAKE 1.3 -m 16 -t 0 –s 0	2:26	12,923 2,684	262	2,768	316	96.6% 99.3%	98.4%
VCAKE § -m 16 –c 0.6 –k 20 other options defaulted	12:50	3,158 246	544	4,078	914	76.7% 94.4%	99.5%
SSAKE 3.0 -m 16 -r 0.6 –p 0 other options defaulted	4:06	1,375 368	484	7,078	1,145	91.8% 96.7%	99.3%

† Trimmed using TQS.py -l 35 -c 20 -t 20 -d 20
§ Calls to STDOUT removed to speed execution
¶ Repeats identified/masked using cross_match (Green P. *www.phrap.org*)
* Contigs 100 nt and up aligning to the reference BAC with 90 % sequence identity or more

Building Scaffolds

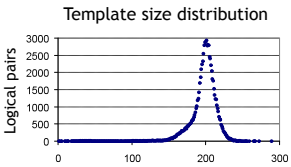


Read pairing within contigs can be used to assess contig assembly quality (*below*)

PAIRING STATS - Human BAC Sequencing

Paired-end reads sequenced	669,038
Missing reads from contigs >= 50nt (-z)	312,461
Reads in contigs < 50 nt	1,545
Assembled reads	355,032
Assembled pairs	177,516
Satisfied in distance/logic within contigs	77,521
Unsatisfied in distance within contigs	106
Unsatisfied pairing logic within contigs	35
Satisfied in distance/logic within a contig pair	47,728
Unsatisfied in distance within a contig pair	52,126

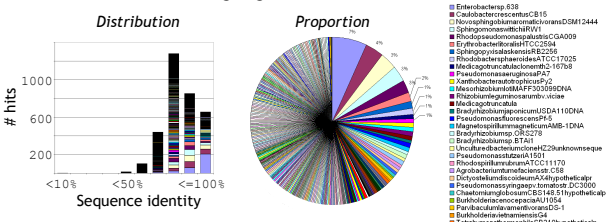
Pairs satisfied: 71% Unsatisfied: 29%



When 670,000 x 27nt paired-end reads co-assembled with 435,000 x 20-35nt unpaired Illumina reads at run-time, SSAKE yielded 75 accurate scaffolds 1 kbp or larger that covered 91% of the unique, non-repeated portion of the Human BAC sequence with 99.5% accuracy. The rest of the genome was covered by scaffolds shorter than 1000 nt. Supplementing 435K unpaired sequences with 150% more reads did not improve individual sequence contiguity, but rather improved the long-range assembly contiguity by further ordering and orientating 366 contigs into 75 large scaffolds.

Applications

Soil SSAKE contig alignments to *nt*



(*Above*) SSAKE assembled ~5M short reads generated from a soil metagenomics sample in 52min. on 2x2.0GHz AMD Opteron 8GB RAM. Contigs were aligned to genbank-*nt* to identify bacterium living in this complex environment and promote the discovery of molecules involved in secondary metabolites biosynthesis.

(*not shown*) ~6.1M cow transcriptome reads assembled with SSAKE in a proof-of-concept study aimed at evaluating the feasibility of short (36bp) reads for transcriptome profiling. With defaults, SSAKE yielded 2,820 contigs 100bp or larger (N50=163 bp). The two largest contigs (~1.1 kbp ea.) aligned to *Bos Taurus* mitochondrial cytochrome c oxidase and NADH dehydrogenase with 91% and 90% identity.

SSAKE now handles error-rich data sets. It does so, quickly and accurately, while maximizing contig length. To our knowledge, it is the first short read assembler to use paired-end reads for scaffolding. SSAKE can be used for *de novo* assembly of single targets or complex DNA, including metagenomes and transcriptomes, to assist in gene and transcript discovery.

Acknowledgements

Funding

GenomeBritishColumbia

RAH is a Michael Smith Foundation for Health Research Scholar

References

Dohm et al. 2007. Genome Res. 17:1697-706
Jeck et al. 2007. Bioinformatic. epub nov07
Warren et al. 2007. Bioinformatics. 23:500-1
epub dec06