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

SSAKE 3.2.1

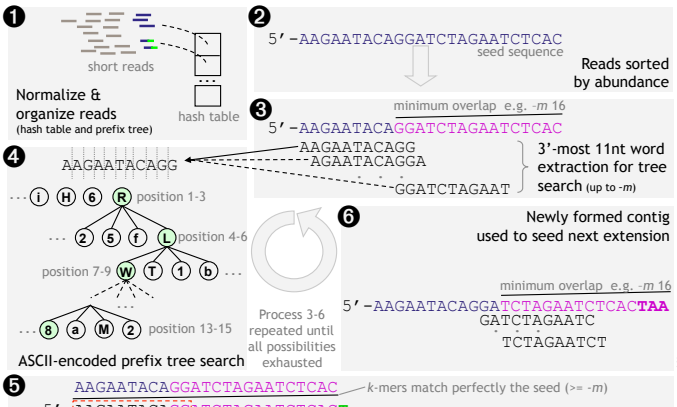
René L Warren and Robert A Holt

Canada's Michael Smith
Genome Sciences Centre

Introduction

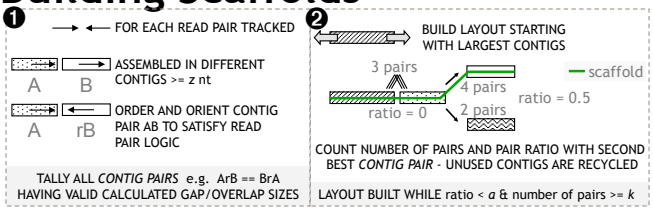
With SSAKE, we demonstrated that *de novo* assembly of large contigs using millions of short (25 bp) error free reads is feasible (Warren *et al.* 2006). This work introduced the use of a prefix tree to organize short sequence reads for efficient *k*-mer search. The SSAKE data representation and general 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 subsequently published short read assemblers 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, ASCII-encoded prefix tree search and contig end-bases adjustments and discuss potential research applications.

Algorithm



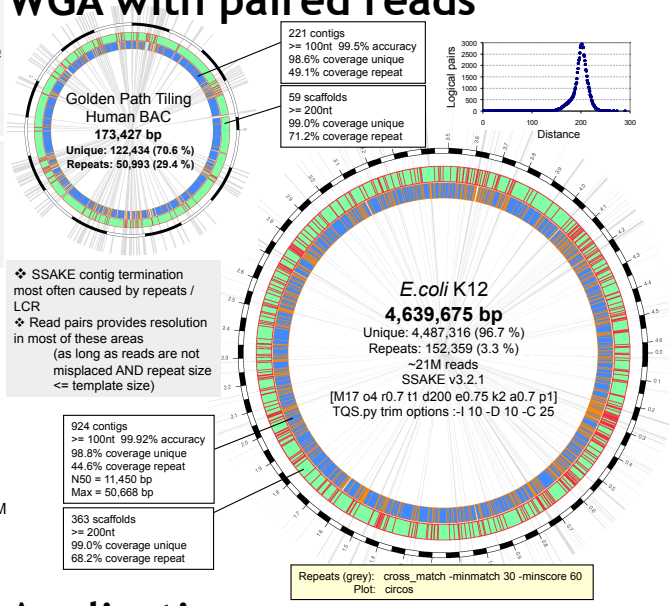
We have implemented in the current version of SSAKE a published approach for handling error-rich sequencing data. In essence, all overhanging bases of reads aligning perfectly to a seed sequence are considered for extension, using a majority-rule approach for building a consensus sequence of the overhanging bases, similar to 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.

Building Scaffolds



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

WGA with paired reads



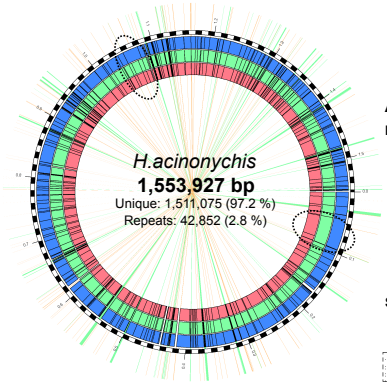
❖ SSAKE contig termination most often caused by repeats / LCR
❖ Read pairs provides resolution in most of these areas (as long as reads are not misplaced AND repeat size <= template size)

WGA with end-error correction

A bacterial genome in 299 contigs, assembled in 40min. on 2.0GHz AMD Opteron with 8GB RAM

End error correction	-t 0	-t 1	-t 2
# Contigs >= 100nt	329	317	299
N50 contig length (nt)	8503	9390	9961
Largest Contig	40.0 kbp	49.5 kbp	49.5 kbp
Accuracy*	99.96%	99.96%	99.95%
%Coverage* (contigs)	95.8% (319)	96.2% (308)	95.0% (289)
Mean contig size (nt)	4661	4834	5122

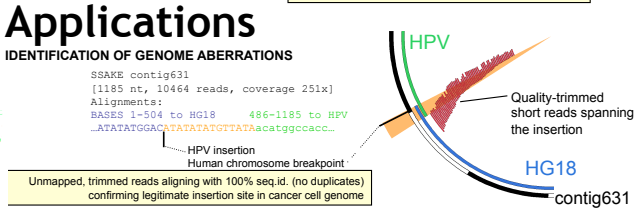
*well-assembled contigs >= 95% sequence identity



Showing SSAKE v3.2.1 -t 2 and -t 0 contigs (blue and red rectangles, respectively) aligning to *Helicobacter acinonychis*. Illumina sequences quality-trimmed with TQS.py (-t 10 -d 10 -c 25 -l 36) and the 5.7M remaining unpaired reads assembled with SSAKE (v3.2 -m 16 -o 3 -r 0.7)

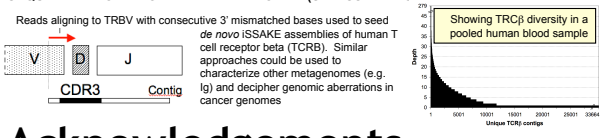
Applications

IDENTIFICATION OF GENOME ABERRATIONS



Unmapped, trimmed reads aligning with 100% seq.id. (no duplicates) confirming legitimate insertion site in cancer cell genome

SEQUENCE-PROFILING T CELL REPERTOIRE (SEE iSSAKE POSTER)



Reads aligning to TRBV with consecutive 3' mismatched bases used to seed *de novo* iSSAKE assemblies of human T cell receptor beta (TCRB). Similar approaches could be used to characterize other metagenomes (e.g. Ig) and decipher genomic aberrations in cancer genomes

Acknowledgements

Funding

GenomeBritishColumbia

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

Enjoy SSAKE responsibly!