Kindel: indel-aware consensus for nucleotide sequence alignments

Bede Constantinides[1] and David L. Robertson[2]

[1]Evolution and Genomic Sciences, University of Manchester, Manchester, UK
[2]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

26 May 2017

Paper DOI: http://dx.doi.org/10.21105/joss.00282
Software Repository: https://github.com/bede/kindel
Software Archive: http://dx.doi.org/10.5281/zenodo.826723

Summary

Kindel is a collection of tools for inferring consensus sequence from an alignment of nucleotide sequences in Sequence Alignment/Map (SAM) format (Li et al. 2009) in the presence of substitutions, insertions and deletions (indels). At regions where reads deviate sufficiently from the reference sequence, partially unaligned sequence context is used to perform local reassembly. In this way, Kindel generates a data-specific reference sequence that maximises overall read-reference similarity. While an elegant streaming approach to consensus inference was implemented in OCOCO (Břinda, Boeva, and Kucherov 2016), like other approaches it fails to reconcile indels.

Kindel was developed for inferring consensus of highly diverse populations of RNA viruses such as hepatitis C and HIV, and is tested with deep sequenced hepatitis C alignments generated by BWA-MEM (Li and Durbin 2009) and Segemehl (Otto, Stadler, and Hoffmann 2014). Furthermore, Kindel may be used to quantify and visualise subconsensus variation in allele frequencies across a reference sequence, facilitating comparison of intrapatient population state among multiple individuals and/or timepoints. Kindel is implemented as a Python 3 package with a command line interface.

References

Břinda, Karel, Valentina Boeva, and Gregory Kucherov. 2016. "Dynamic read mapping and online consensus calling for better variant detection," 1–21. http://arxiv.org/abs/1605.09070.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14): 1754. doi:10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and Samtools." *Bioinformatics* 25 (16): 2078. doi:10.1093/bioinformatics/btp352.

Otto, Christian, Peter F. Stadler, and Steve Hoffmann. 2014. "Lacking Alignments? The Next-Generation Sequencing Mapper Segemehl Revisited." *Bioinformatics* 30 (13): 1837. doi:10.1093/bioinformatics/btu146.

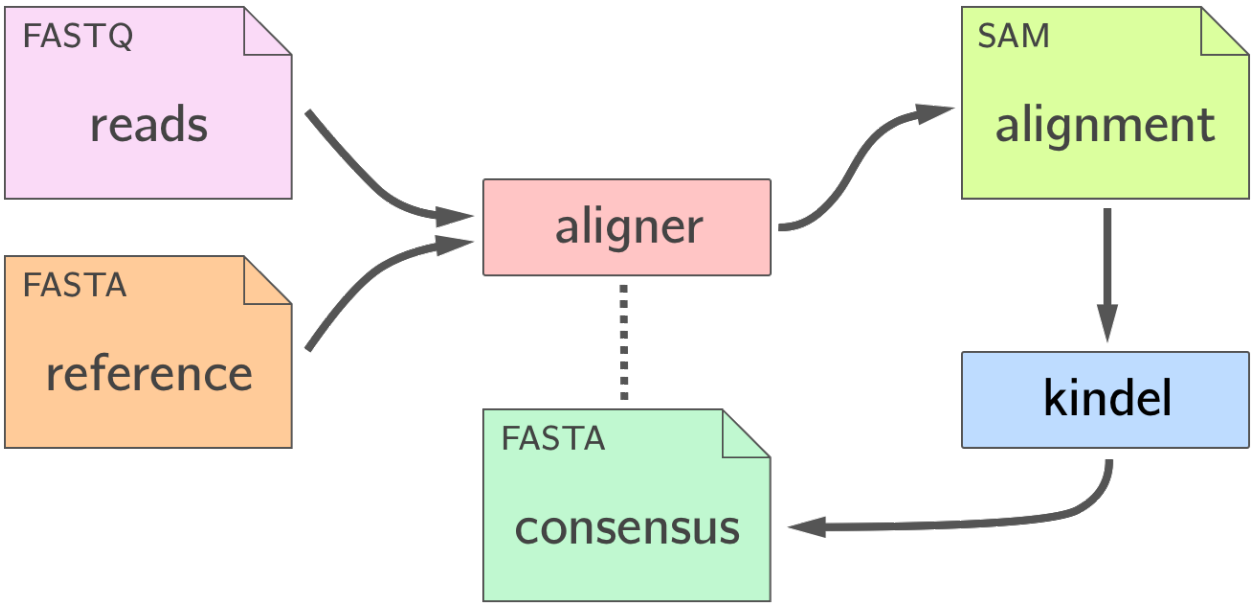

Figure 1: Usage overview.

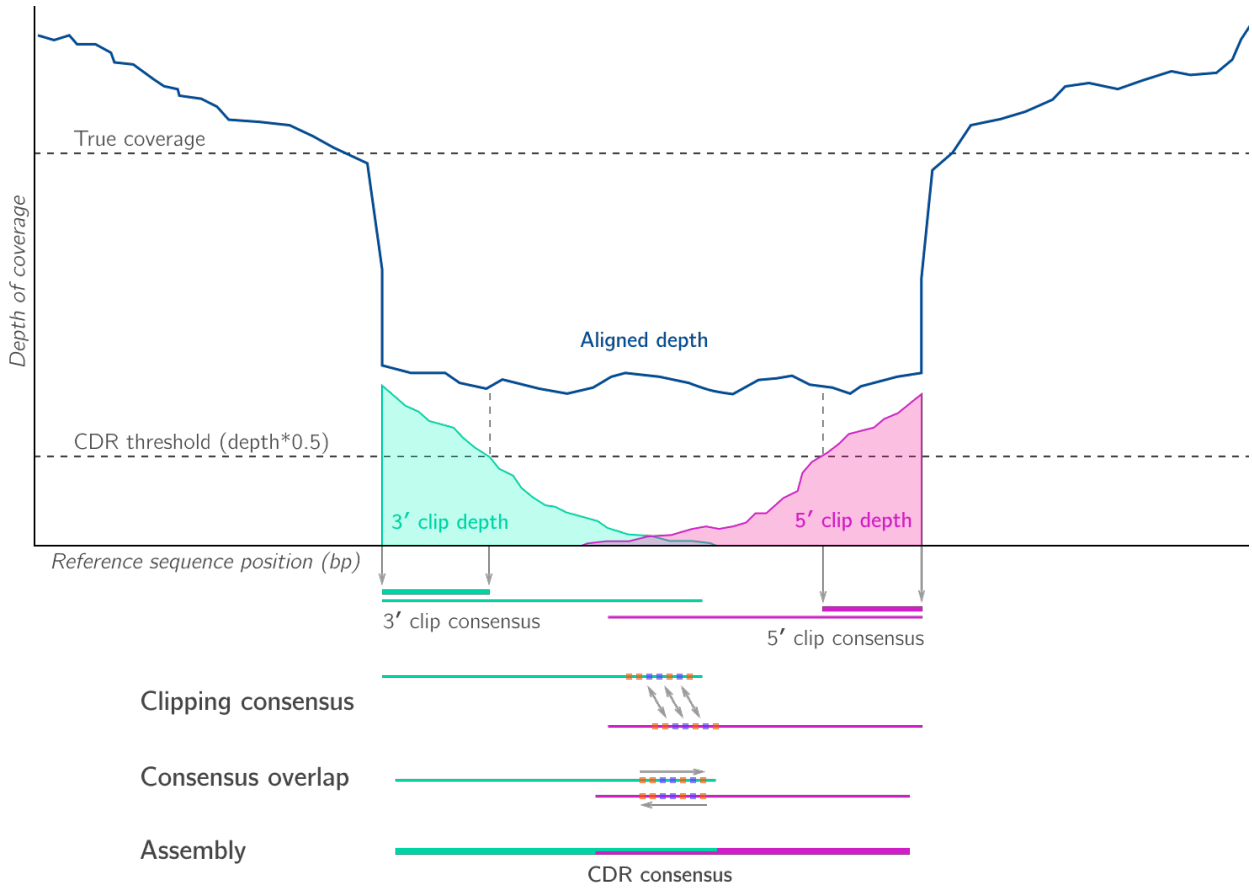

Figure 2: Local reassembly of clip-dominant regions (CDRs). Leveraging partially aligned reads, consensus can be accurately inferred across unrepresentative and poorly covered regions of reference sequence.