Enrichment Analysis of k-mer Composition Enables Identification of Telomeres

Askar Gafurov, Viktória Hodorová, Hana Lichancová, Jozef Nosek, Tomáš Vinař, Broňa Brejová

Faculty of Mathematics, Physics, and Informatics and Faculty of Natural Sciences
Comenius University in Bratislava, Slovakia
{gafurov,brejova}@dcs.fmph.uniba.sk http://compbio.fmph.uniba.sk/

Abstract

Background: Telomeres and repeat-rich subtelomeric regions are often hard to assemble from high-throughput sequencing data, and therefore the exact nature of the telomeric sequences remains unknown in many species.
Results: We have developed a k-mer based sequence analysis method to identify contig ends belonging to telomeric and sub-telomeric regions. Our method uses a combination of long-read and short-read sequencing and compares k-mer composition in reads from untreated DNA to DNA treated with BAL31 nuclease. This enzyme digests ends of DNA molecules and thus creates a depletion of telomeric and sub-telomeric areas.
Conclusions: We have applied our methods to the genome of basidiomycetous yeast *Jaminaea angkorensis* genome. Our approach combining k-mer analysis, BAL31 digestion protocol, and Oxford Nanopore sequencing has improved assembly of repeat-rich subtelomeric regions in this genome.

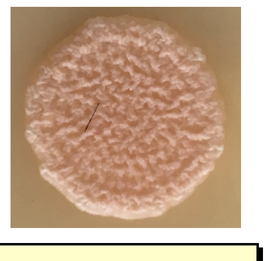
Motivation: Finding telomeres in *Jaminaea angkorensis*

Typical structure of telomeres:

- vertebrates (TTAGGG)*, many insects (TTAGG)*, many plants (TTTAGGG)*
- some fungi (TTAGGG)*, others more variable:
  Saccharomyces cerevisiae (TG$^{2-3}$(TG)$^{1-6}$)*, *Yarrowia lipolytica* (GGACGATTG)*,
  Candida albicans (ACGGATGTCTAACTTCTTGGTGT)*
  Jaminaea angkorensis: fungus from Basidiomycota phylum
  No candidate telomeric repeats similar to the ones above found

J. angkorensis genome assembly:

- Nanopore and Illumina, assembly by Canu (Koren et al., 2017) followed by manual curation resulted in 20 nuclear contigs
- Contig lengths 136kbp–2.8Mbp, total length 20.7Mbp
- Contigs terminated by long tandem repeats, lengths cca 70-170bp
  Are these telomeres / sub-telomeres?



Experimental approach: BAL31 digestion

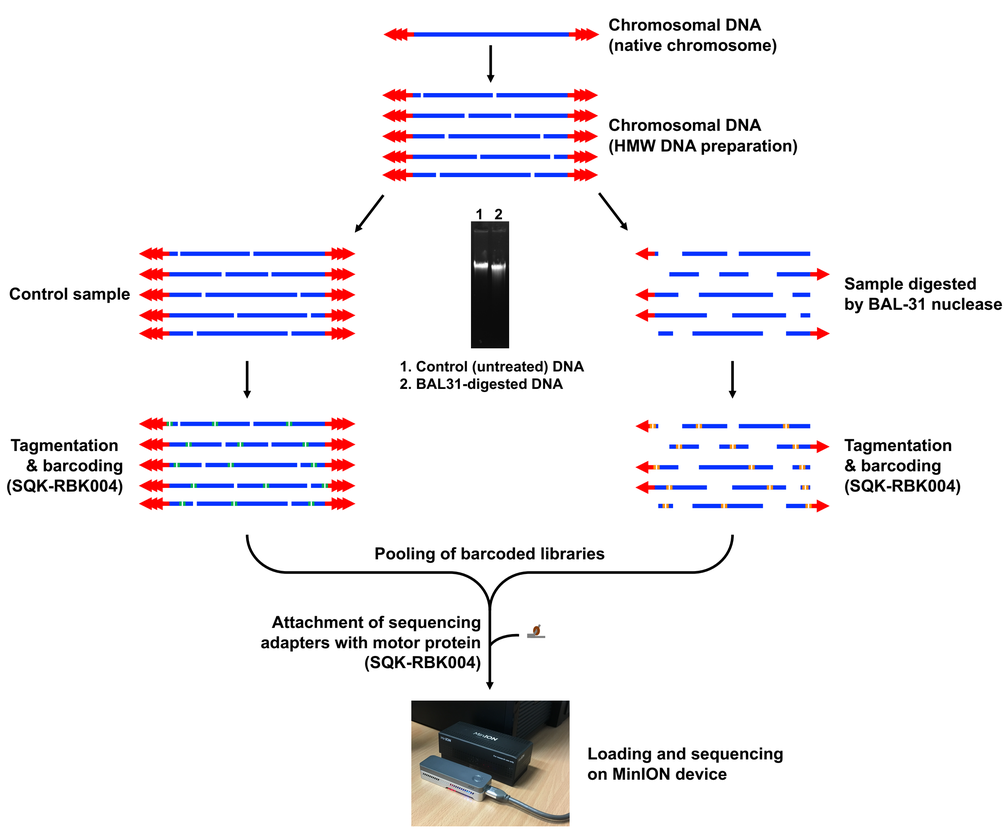


Figure 1: *Digestion of choromosome ends followed by sequencing of control and treated sample for differential analysis. Telomeric regions should be depleated in the treated sample compared to the control. Protocol inspired by Peška et al. (2015)*

Naive approach through read alignment does not work

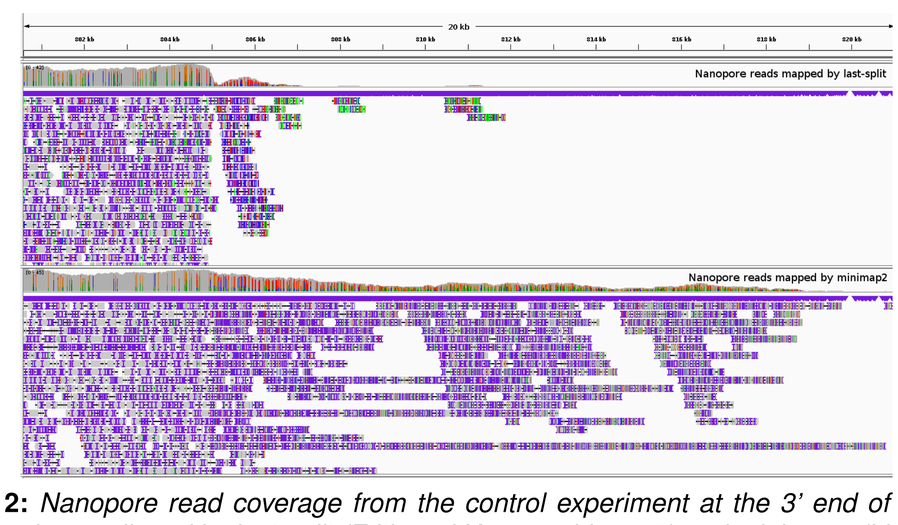


Figure 2: *Nanopore read coverage from the control experiment at the 3' end of contig12 when reads are aligned by last-split (Frith and Kawaguchi, 2015) and minimap2 (Li, 2018)*

Differential analysis of k-mer abundance

1. Estimate k-mer abundances in control and treated sample using Jellyfish (Marçais and Kingsford, 2011) (in our experiments $k = 21$)
2. Filtering: remove unique k-mers and k-mers not present in Illumina (likely representing sequencing and base calling errors)

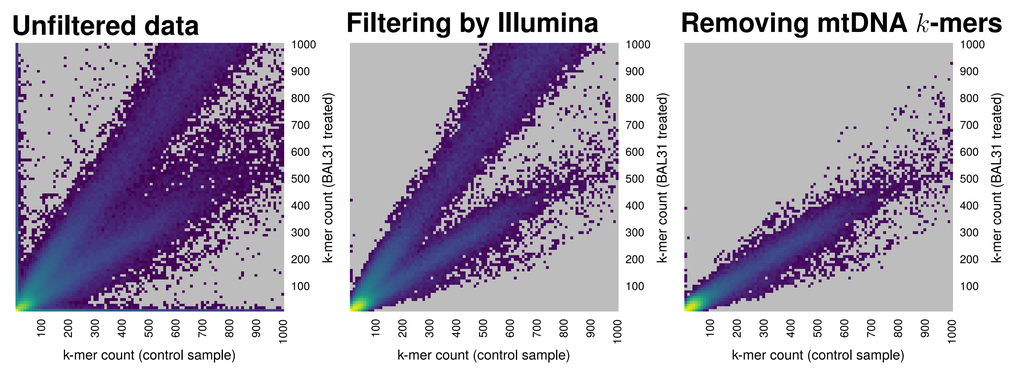


Figure 3: *Abundance of k-mers in control vs. treated sample.*

3. Finding k-mers depleted in treated sample: for each k-mer use Fisher's exact test on the number of occurrences in treated and control sample; keep k-mers with $P < 0.025$
4. Mapping: map depleted k-mers back to the genome; find 1kb windows with at least 50% bases covered by depleted k-mers

Results

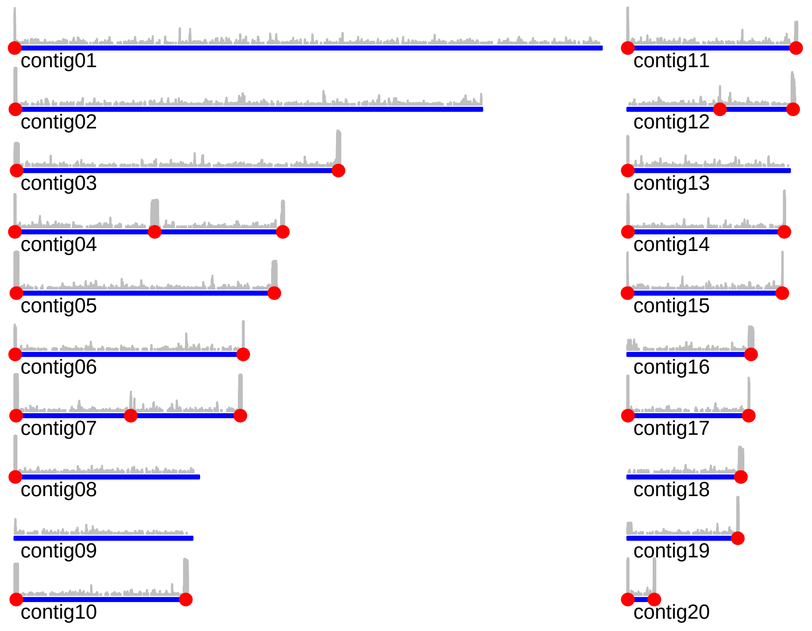


Figure 4: *Differential analysis confirms 30 out of 40 contig ends. Coverage of windows by depleted k-mers is shown in gray, regions with significantly depleted windows are highlighted in red.*

Future plans and open questions

- More complex probabilistic models for k-mer differential analysis
- Further experimental testing of putative telomeres
- Biological/evolutionary mechanisms?
- Examination of atypical k-mers
- Tandem repeats remain difficult for many bioinformatics tasks

Acknowledgments

This work was funded by the Slovak Research and Development Agency grants APVV-14-0253, APVV-18-0239, and APVV-15-0022 and grants from VEGA 1/0684/16 (BB) and 1/0458/18 (TV).

References

Frith, M. C. and Kawaguchi, R. (2015). Split-alignment of genomes finds orthologies more accurately. *Genome biology*, 16(1):106.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

Peška, V., Fajkus, P., Fojtová, M., Dvořáčková, M., Hapala, J., Dvořáček, V., Polanská, P., Leitch, A. R., Sýkorová, E., and Fajkus, J. (2015). Characterisation of an unusual telomere motif (TTTTTTAGGG)$_n$ in the plant *Cestrum elegans* (Solanaceae), a species with a large genome. *The Plant Journal*, 82(4):644–654.