Taxon classification on protein or nucleotide database

Adrienn Gréta Tóth[1], **István Csabai**[2], **Eszter Krikó**[1], **Róbert Farkas**[1], **László Makrai**[1], **Norbert Solymosi**[1, 2]

[1]Univeristy of Veterinary Medicine Budapest, [2]Eötvös Loránd University

tothadrienngreta@gmail.com

Abstract

In clinical metagenomics, microbial profiling and pathogen identification are both important fields of use. In both areas, the taxon classification of the next generation sequencing short reads can be performed using numerous tools and databases. Some alignment tools used before the reads' assignment to their taxa are nucleotide-based, while others are protein-sequence-based. Analysing animal gut microbiomes, we found that various types of databases can produce different results. After protein-based classification, much more short reads get classified as bacterial hits. Examining this excess, it seemed that the difference can come from false-positive (FP) assignments. Investigating the reasons, we found that the presence of eukaryotic genomes increased the number of results classified FPs in the different NGS samples. We generated artificial datasets simulating short reads from various eukaryotic genomes managing their exonic, intronic and intergenic regions separately. Classifying these reads on protein and nucleotide databases, we found a typical means of misclassification. Nucleotide based assignment of the eukaryotic reads had lower rates of FPs for each region compared to the protein-based assignments, where we had a higher proportion of FP microbial hits. The excess of FPs was considerably higher in reads from the intronic or intergenic parts of the genome compared to the exonic regions. Although in metagenome analyses the host genome is filtered out before taxon classification, other eukaryotic genome components (e.g. feed) might remain among the short reads. Our findings suggest that in samples with eukaryotic contamination, the nucleotide-based taxon classification decreases the FP microbial hits.

Introduction

Main Objectives

Materials and Methods

Mathematical Section

Results

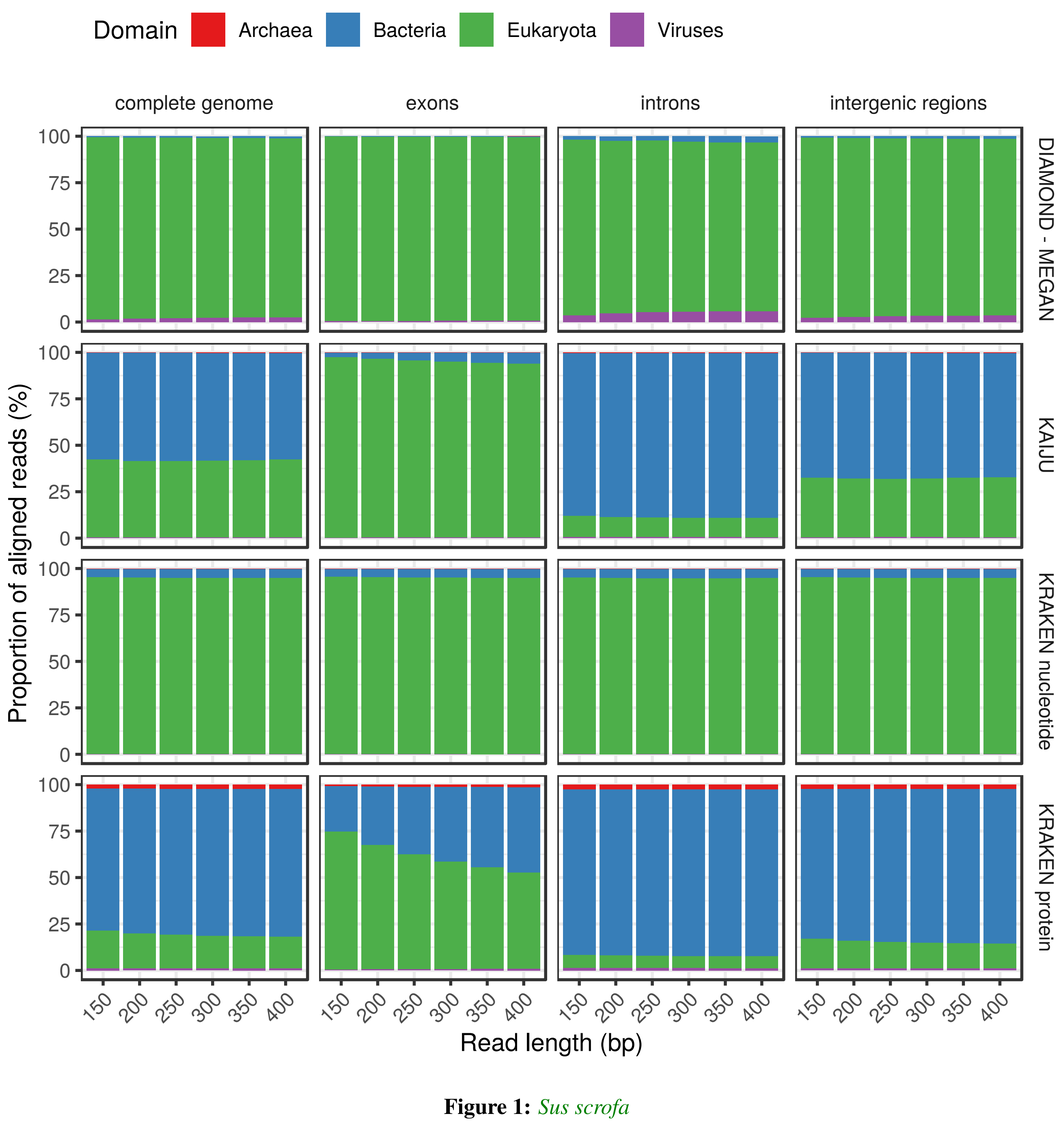

Figure 1: *Sus scrofa*

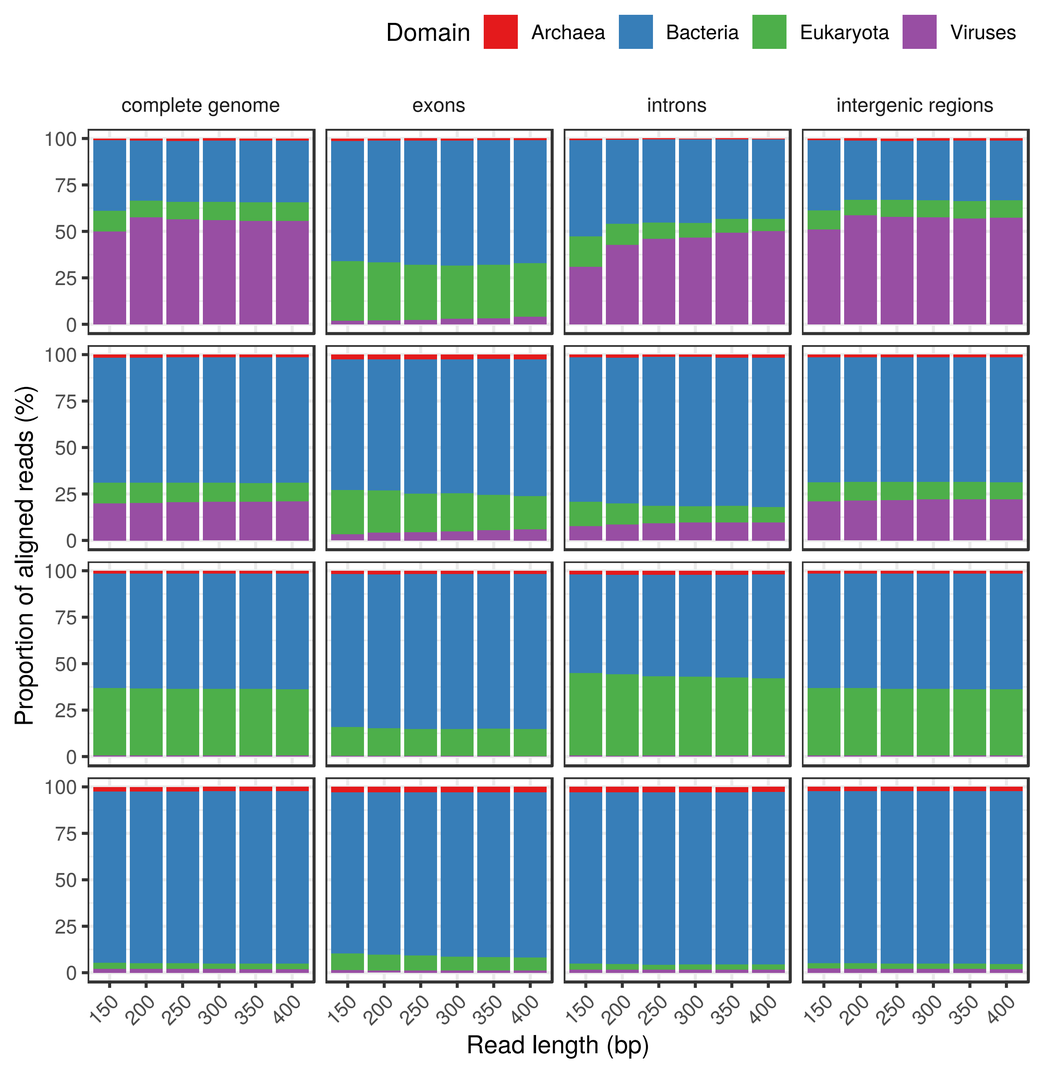

Figure 2: *Zea mays*

Conclusions

Forthcoming Research

Acknowledgements

Take a picture
to download