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]University of Veterinary Medicine, [2]Eötvös Loránd University, Budapest, Hungary

tothadrienngreta@gmail.com

Introduction

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 (NGS) 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 strongly diverging 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. Our goal was to have quantitative information of the FP by comparing some aligment/classification tools using the same reference sequence database.

Materials and Methods

To analyse the amount of the incorrect read classifications artificial short reads were generated from the genome of different eukaryotic organisms (*Caenorhabditis elegans*, *Chrysemys picta*, *Danio rerio*, *Equus caballus*, *Falco cherrug*, *Homo sapiens*, *Hydra vulgaris*, *Orcinus orca*, *Ornithorhynchus anatinus*, *Pan paniscus*, *Phascolarctos cinereus*, *Phincodon typus*, *Sus scrofa*, *Xenopus laevis*, *Xenopus tropicalis*, *Zea mays*). For all genomes various lengths of reads (bp: 150, 200, 250, 300, 350, 400) were generated by bbmap (Bushnell, 2014) not allowing any error in the reads. The sequences generated fully covered the original genome. Exonic, intronic and intergenic sequences of all the genomes involved in the project were separated and brought under the same short read simulation.

The simulated reads were aligned and taxonomically classified by using DIAMOND-MEGAN6 (Buchfink et al., 2015; Huson et al., 2016), KAIJU (Menzel et al., 2016) and KRAKEN2 (Wood et al., 2019). As common reference sequence collection the standard KRAKEN database (RefSeq of archea, bacteria, viruses and *Homo sapiens*) was used. For KRAKEN it was built on nucleotide and protein sequences as well. Based on the KRAKEN protein sequences databases were created for KAIJU and DIAMOND too. The default settings were applied for all tools in alignment and taxon classification. The domain level taxon classification of the reads was calculated in R (R Core Team, 2019) and expressed as the proportion of the total classified reads per syntetic sample.

Results

Out of all our findings, the read classification results of two species genome's can be observed below. The proportions of the read classifications presented on Figure 1 and 2 for *Sus scrofa* and *Zea mays*, respectively. The minimal falsely classified read numbers by alignment/classification tool and by genome parts are presented in Table 1.

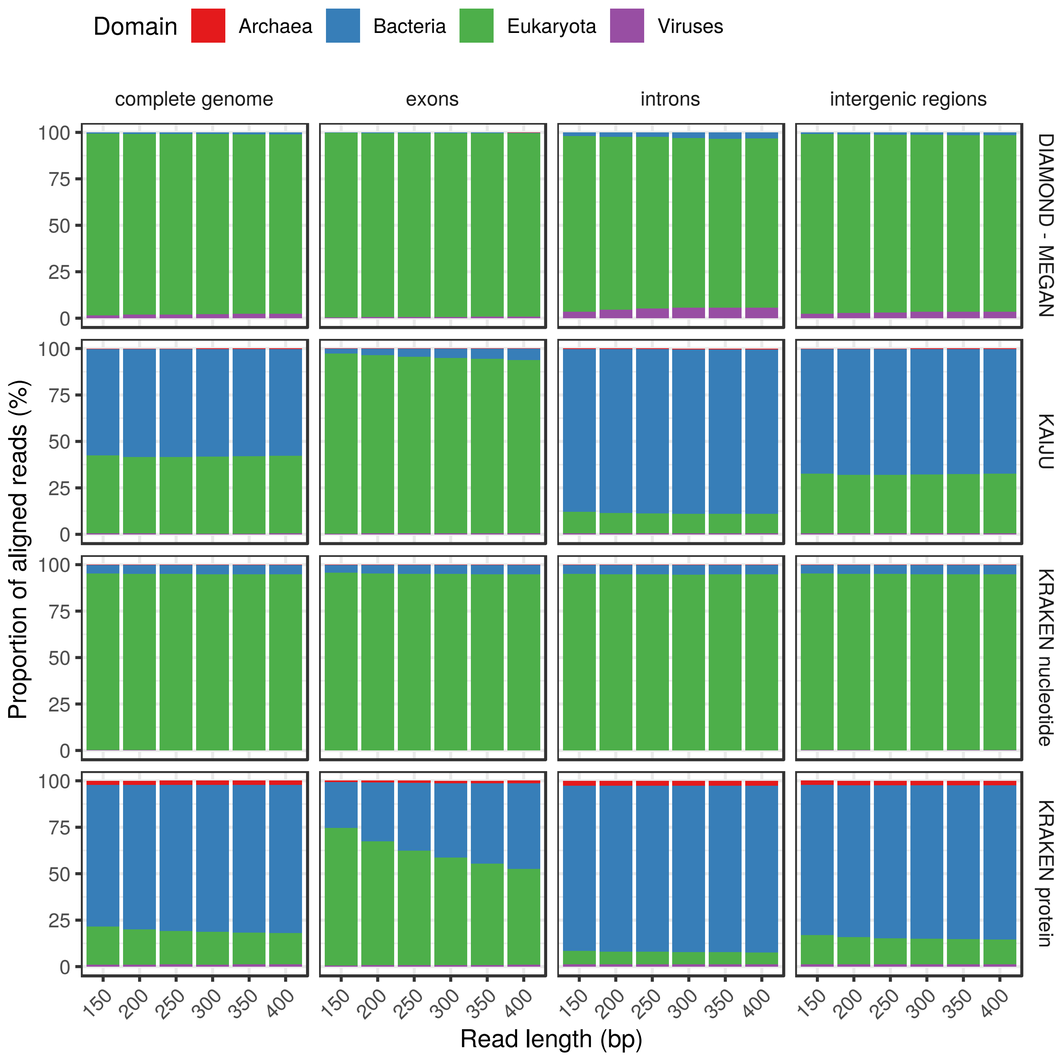

Figure 1: The domain level proportions of classified reads generated from the *Sus scrofa* genome.

Conclusions

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 proportion of FP microbial assignments.

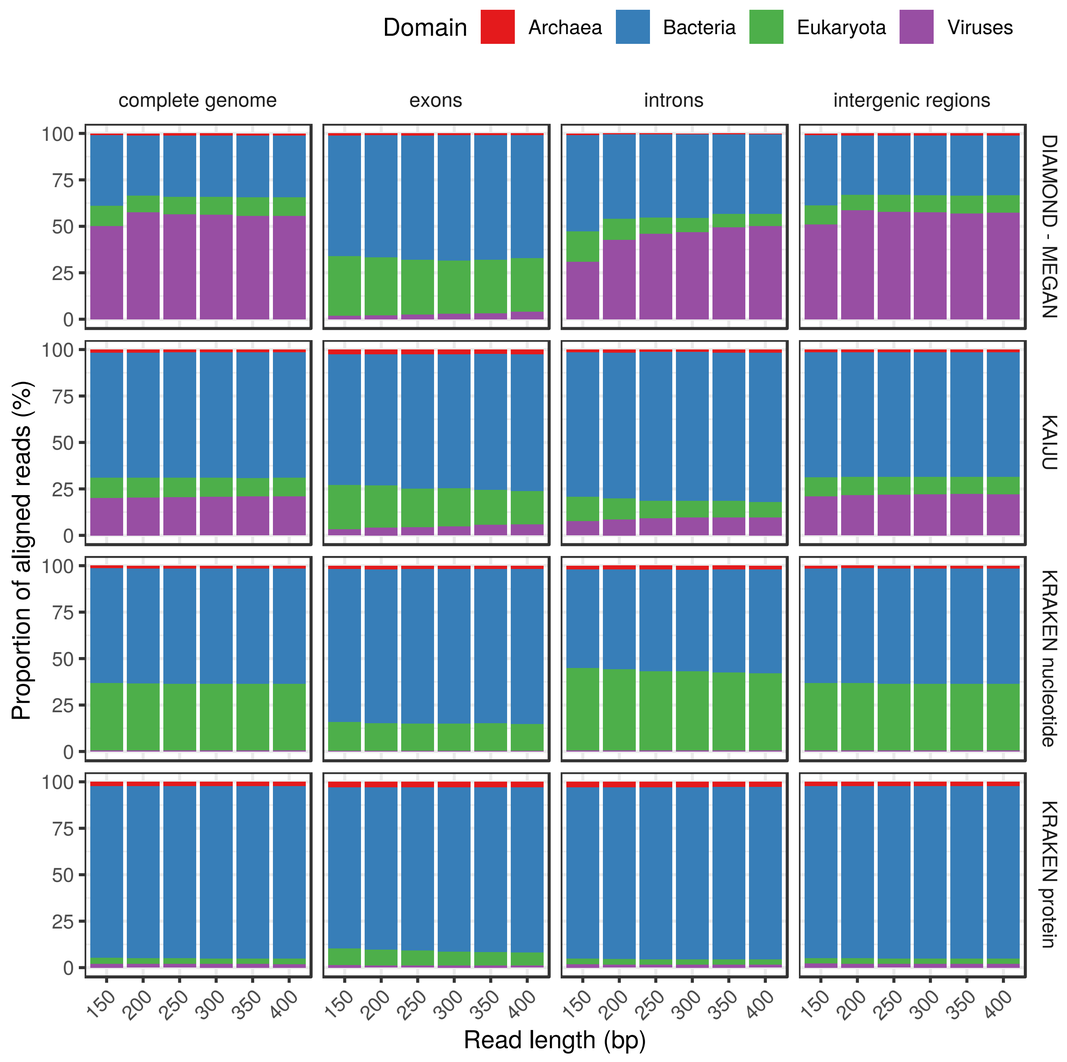

Figure 2: The domain level proportions of classified reads generated from the *Zea mays* genome.

Table 1: The minimal number of the reads classified microbial despite those were generated from eukaryotic genomes.

Origin of reads	Tool	Number of reads					
		Sus scrofa			*Zea mays*		
		Archaea	Bacteria	Viruses	Archaea	Bacteria	Viruses
complete genome	DIAMOND - MEGAN	0	1,114	3,109	555	26,268	34,539
	KAIJU	1,091	281,012	2,502	1,155	52,633	15,727
	KRAKEN nucleotide	4,952	136,070	4,450	4,848	218,801	2,232
	KRAKEN protein	29,649	1,079,527	16,810	24,255	937,416	21,838
exons	DIAMOND - MEGAN	0	85	283	156	10,870	326
	KAIJU	48	2,746	261	270	8,414	536
	KRAKEN nucleotide	198	5,917	149	510	26,239	102
	KRAKEN protein	924	33,343	759	1,455	44,047	627
introns	DIAMOND - MEGAN	0	400	827	23	1,728	1,035
	KAIJU	587	134,662	892	45	3,065	316
	KRAKEN nucleotide	2,704	70,358	1,709	342	9,763	112
	KRAKEN protein	15,335	547,050	8,277	1,329	42,915	843
intergenic regions	DIAMOND - MEGAN	0	1,553	4,699	1,026	45,494	61,498
	KAIJU	1,522	421,402	3,863	1,979	91,465	28,783
	KRAKEN nucleotide	7,110	193,320	6,810	8,838	390,181	4,020
	KRAKEN protein	42,558	1,556,629	24,142	42,862	1,686,047	39,986

Acknowledgements

The project is supported by the European Union and co-financed by the European Social Fund (No. EFOP-3.6.3-VEKOP-16-2017-00005), ÚNKP-19-2 New National Excellence Program of the Ministry for Innovation and Technology and has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 643476 (COMPARE).

References

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59.

Bushnell, B. (2014). Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, 12(6).

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7(1):1–9.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome biology*, 20(1):257.