Performance evaluation of taxonomic classification between different Kraken databases

Fredrik Dyrkell[1], Dimitrios Arnellos[1], Oskar Andersson[1]

[1]1928 Diagnostics AB – Gothenburg (Sweden)

Background

Species identification by means of taxonomic classification is an integral part of computational genomics pipelines. In metagenomics it provides abundance estimations characterising the content, and for single cultures it provides contamination checks and verification of other molecular methods such as MALDI-TOF.

1928 Diagnostics develops a cloud-based platform that analyses WGS sequences to trace outbreaks and perform prospective genomic surveillance. The platform is built to handle all major sequencing platforms and supports 25 different clinically relevant bacterial pipelines with analysis including species identification, high-resolution strain typing, AMR and virulence.

The aim of this project was to benchmark different database versions for species identification and compare results over a large dataset, to evaluate trade-off in performance for bacterial species, while including protozoa and fungi in the database.

Method

From 252 different bioprojects, 2475 single culture isolates were selected and downloaded from ENA, annotated as 138 different species designations. 76 samples were sequenced using Oxford Nanopore, whereas the rest with Illumina. The samples were subsequently analysed by 1928's custom developed identification pipeline, and uses Kraken 2[1] for classification of FASTQ reads in conjunction with Bracken[2] to assign prediction at species level. The database versions under test were:

- k2 pluspf 8gb – 20220908 (pluspf)
- k2 standard 8gb – 20210517 (standard)

The classification results were divided into three distinct classes.

- Samples with a single unique species prediction,
- No unique species, multiple predictions belonging to the same genus
- Different genera in prediction

In the results generating non-unique predictions, classification was performed on species predictions with abundance >1.0%.

Results

The benchmark results for the two databases are shown in Figure 1.

	STANDARD			PlusPF					
	Unique species prediction	Single Genus	Multiple genera	Unique species prediction	Single Genus	Multiple genera	Delta		
Acinetobacter baumannii	53	7	1	57	3	1	4	-4	0
Campylobacter jejuni	86	8	0	88	6	0	2	-2	0
Citrobacter freundii	40	1	60	28	20	53	-12	19	-7
Clostridioides difficile	99	0	0	99	0	0	0	0	0
Enterobacter	108	0	2	107	0	3	-1	0	-1
Enterococcus faecalis	98	0	1	98	0	1	0	0	0
Enterococcus faecium	85	2	0	85	2	0	0	0	0
Escherichia coli	111	0	3	110	0	4	-1	0	1
Klebsiella aerogenes	94	2	6	87	6	9	-7	4	3
Klebsiella oxytoca	59	20	23	60	17	25	1	-3	2
Klebsiella pneumoniae	99	11	7	110	1	6	11	-10	-1
Klebsiella variicola	9	64	28	5	73	23	-4	9	-5
Legionella pneumophila	100	0	0	100	0	0	0	0	0
Listeria monocytogenes	65	0	0	65	0	0	0	0	0
Mycobacterium	91	1	1	86	6	1	-5	5	0
Mycobacteroides abscessus	100	0	1	100	0	1	0	0	0
Nanopore	68	6	2	66	9	1	-2	3	-1
Neisseria meningitidis	51	2	0	51	2	0	0	0	0
Pseudomonas aeruginosa	101	0	2	101	0	2	0	0	0
Salmonella enterica	93	1	4	94	0	4	1	-1	0
Serratia marcescens	28	4	1	23	8	2	-5	4	1
Staphylococcus aureus	116	0	1	116	0	1	0	0	0
Staphylococcus epidermidis	117	1	3	116	1	4	-1	0	1
Streptococcus dysgalactiae	86	15	2	79	22	2	-7	7	0
Streptococcus pneumoniae	100	0	0	100	0	0	0	0	0
Streptococcus pyogenes	45	1	0	46	0	0	1	-1	0
Tail samples	54	16	9	54	19	6	0	3	-3
	2156	**162**	**157**	**2131**	**195**	**149**	**-25**	**33**	**-8**
	2475			2475			-1.01%	1.33%	-0.32%

Figure 1. Prediction difference between Standard and PlusPF databases. "Nanopore" and "Tail samples" consist of species with <=5 samples per species.

Conclusions

Including protozoa and fungi in the Kraken 2 database maintains a high predictive power for bacteria, with only 1.01% fewer unique predictions, while additionally enabling classification of fungal pathogens.

References

1. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. Springer Science and Business Media LLC; 2019.
2. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. PeerJ; 2017. p. e104.

Conflict of interest statement

F.D., D.A. and O.A. are employees of 1928 Diagnostics AB.

Contact

fredrik.dyrkell@1928diagnostics.com