Using AnnoTree to get more assignments, faster, in DIAMOND+MEGAN microbiome analysis

Anupam Gautam[1,2], Hendrik Felderhoff[1], Caner Bağci[1,2], Daniel H. Huson[1,2]

1 *Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany.*

2 International Max Planck Research School, "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, Tübingen, Germany.

Introduction

In microbiome analysis, one main approach is to align metagenomic sequencing reads against a protein reference database, such as NCBI-nr, and then to perform taxonomic and functional binning based on the alignments. This approach is embodied, for example, in the standard DIAMOND+MEGAN analysis pipeline, which first aligns reads against NCBI-nr using DIAMOND and then performs taxonomic and functional binning using MEGAN.

Here, we propose the use of the AnnoTree protein database, rather than NCBI-nr, in such alignment-based analyses to determine the prokaryotic content of metagenomic samples. We demonstrate a 2-fold speedup over the usage of the prokaryotic part of NCBI-nr and increased assignment rates, in particular assigning twice as many reads.

Study design

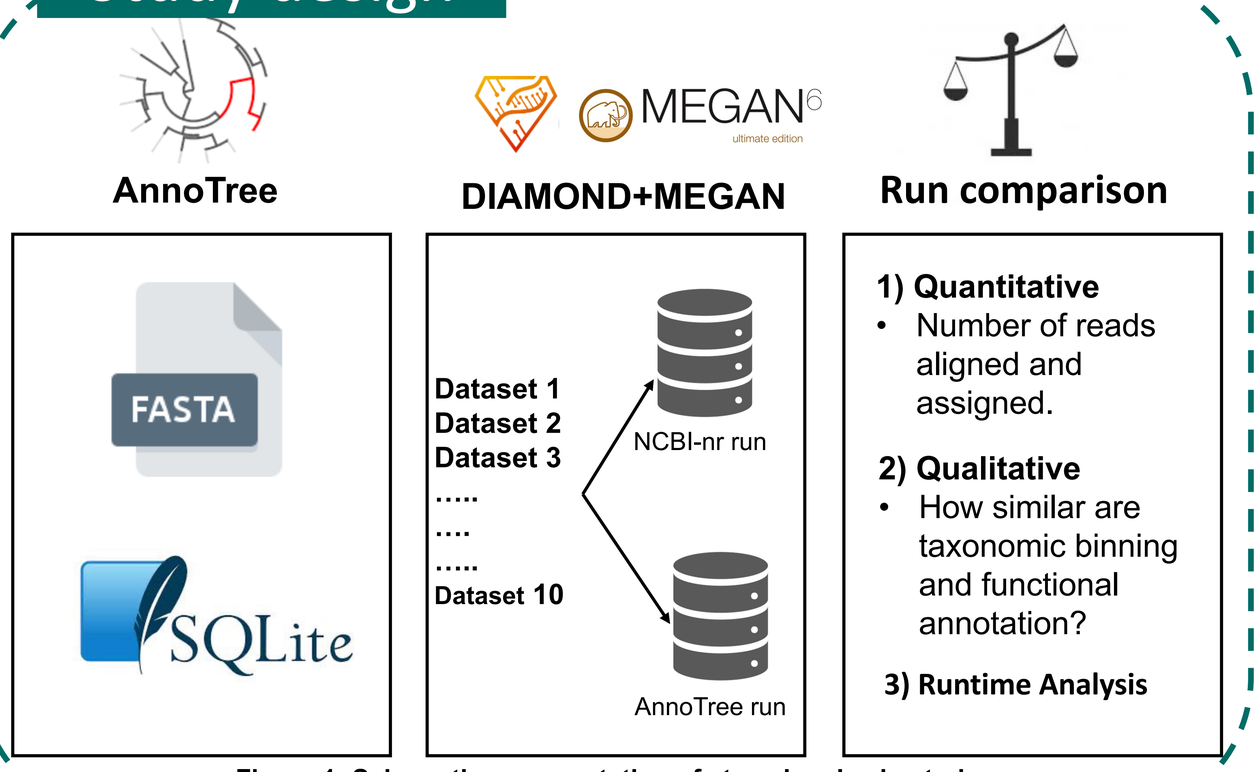

Figure 1: Schematic representation of steps involve in study.

Quantitative comparison

Table 1: Diamond alignment statistics for AnnoTree and NCBI-nr database on 10 dataset.

Dataset	Total no of reads	Reads with DIAMOND alignments				Ratio
		AnnoTree (no.)	%	NCBI-nr (no.)	%	
River1	646,178	410,118	63.5	406,913	63.0	1.0
River2	129,753,222	90,535,941	69.8	88,403,713	68.1	1.0
Seagrass	98,260,754	36,053,215	36.7	33,717,202	34.3	1.1
Skin	22,827,626	13,403,495	58.7	14,122,490	61.9	0.9
Stool	33,214,614	29,132,562	87.7	30,101,313	90.6	1.0
Soil	97,595,185	10,992,188	11.3	7,264,223	7.4	1.5
Thermal Pools	52,908,626	15,751,382	29.8	16,625,446	31.4	0.9
Bioreactor1	99,998,110	73,151,916	73.1	72,806,515	72.8	1.0
Bioreactor2	44,258,996	36,608,649	82.7	37,477,641	84.7	1.0
Bioreactor3*	694,827	613,958	88.4	616,536	88.7	1.0
Total	580,158,138	306,653,424	52.86	301,541,992	51.98	1.02

Table 2: MEGAN Assignment statistics for different classifications on 10 dataset.

Classification	AnnoTree run			NCBI-nr run			
	Assigned	% of R	% of AI	Assigned	% of R	% of AI	Ratio
NCBI taxonomy	305,150,157	52.6	99.5	297,539,333	51.3	98.7	1.0
GTDB taxonomy	303,770,449	52.4	99.1	282,269,816	48.6	93.6	1.1
EC	78,874,545	13.6	25.7	76,552,285	13.2	25.4	1.0
eggNOG	95,932,149	16.5	31.3	87,131,284	15.0	28.9	1.1
InterPro	142,250,858	24.5	46.4	143,885,580	24.8	47.7	1.0
KEGG	209,371,499	36.1	68.3	123,130,673	21.2	40.8	1.7
SEED	102,452,692	17.7	33.4	100,615,086	17.3	33.4	1.0

Qualitative comparison

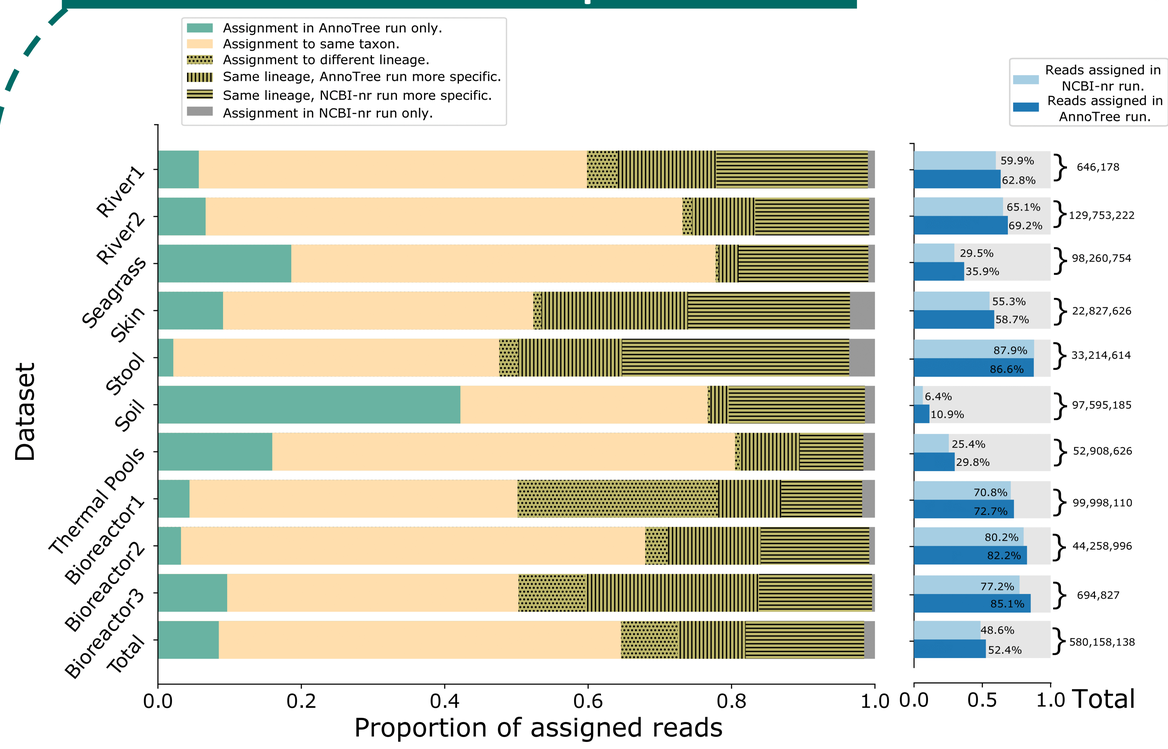

Figure 2: Detailed assignment of reads to the GTDB Taxonomy.

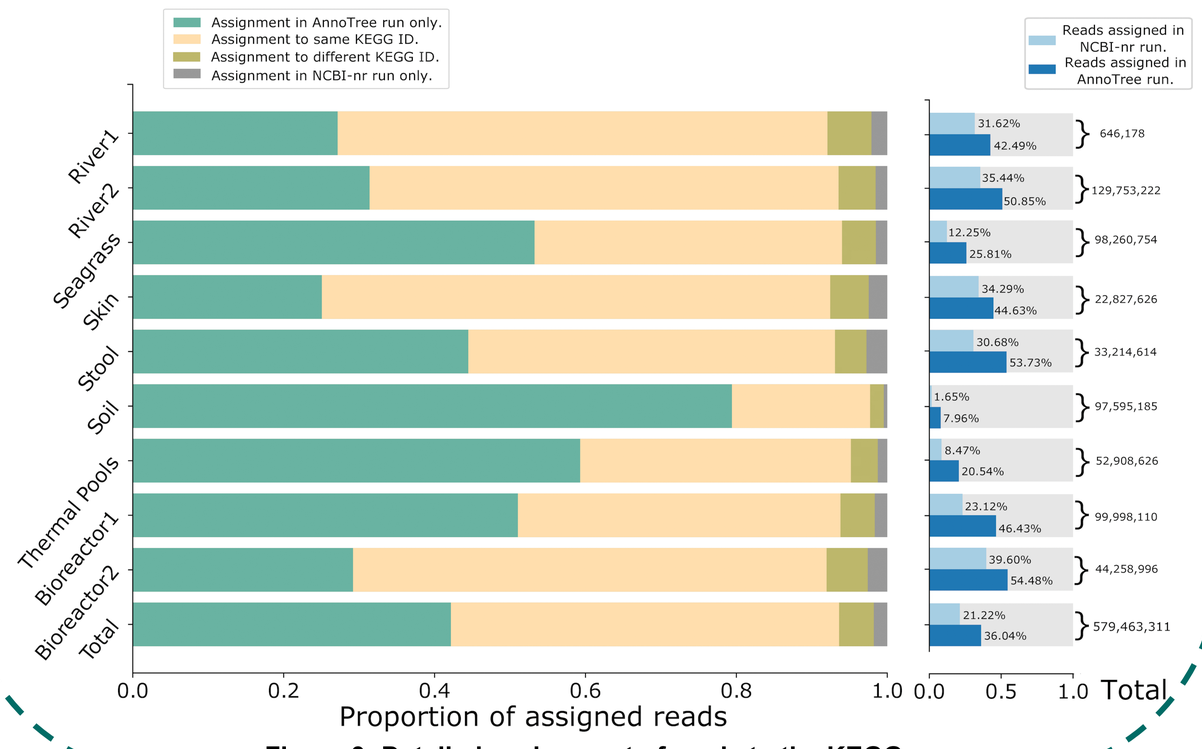

Figure 3: Detailed assignment of reads to the KEGG.

Runtime analysis

Table 3: Run Time analysis for Annotree run and NCBI-nr run.

DIAMOND			MEGANIZER			DIAMOND+MEGAN		
NCBI-nr	AnnoTree	Ratio	NCBI-nr	AnnoTree	Ratio	NCBI-nr	AnnoTree	Ratio
125,288 min	61,443 min	2.0	2,241 min	2,404 min	0.9	127,529 min	63,847 min	2.0

Conclusion

- AnnoTree is only 1/4 the size of the full NCBI-nr.
- Similar alignment and assignment rate.
- Twice as many reads assigned to KEGG.
- 2-fold speed up.

Publication and Data Availability

- Gautam, Anupam, Hendrik Felderhoff, Caner Bağci, and Daniel H. Huson. "Using AnnoTree to get more assignments, faster, in DIAMOND+ MEGAN microbiome analysis." ***Msystems*** 7, no. 1 (2022): e01408-21.

- MEGAN-AnnoTree Download Page: