Assessing orpheum's efficacy for ORF prediction from prokaryotic metagenomes

Background

- The average bacterial genome is 5 Mbp and encodes 5000 proteins (Land et al. 2015). As of January 2014:
  - The largest bacterial genome in GenBank was *Sorangium cellulosum* strain So0157-2, with 14,782,125 bp encoding 11,599 genes
  - The smallest bacterial genome in GenBank was *Candidatus Nasuia deltocephalinicola* strain NAS-ALF with 112,091 bp encoding 137 genes.
- It is estimated that 88% (40-97%) of the bacterial genome is protein coding (Land et al. 2015).
- It is not currently known how many alternate/overlapping ORFs exist in a bacterial genome.
  - Computational estimates suggest there are 10-400 alternate ORFs per genome (Ardern et al. 2020).
  - Experimental evidence suggest ~100 alternate overlapping ORFs per genome (Zehentner et al. 2020)
- Using 76 isolates in RefSeq, on average 89.2% of the *R. gnavus* genome is coding. While we don't know if CDSs will follow the same percentages in metagenome assembly graph query neighborhoods, this is a good approximate estimate to aim for for recovery of coding sequences.

Data description

The 605 gut microbiome metagenomes analyzed in this repository were originally analysed in the 2020-ibd repository as a meta-cohort of IBD subtypes (CD, UC, and nonIBD). Assembly graph neighborhoods for the query genome *R. gnavus* were extracted with spacegraphcats.

Summary tables

database	alphabet	ksize	reads	coding	non-coding	too short	stop codon
f___Lachnospiraceae	protein	k=10	587221	423522	133886	172	29640
f___Lachnospiraceae	protein	k=6	587221	557529	51	0	29640
f___Lachnospiraceae	protein	k=7	587221	491754	65826	0	29640
p___Firmicutes_A	protein	k=10	587221	437625	119786	169	29640
p___Firmicutes_A	protein	k=7	587221	545883	11697	0	29640
plass_assembly	dayhoff	k=15	587221	542324	8445	6811	29640
plass_assembly	dayhoff	k=17	587221	537445	10097	10038	29640
plass_assembly	protein	k=10	587221	552430	5065	85	29640
plass_assembly	protein	k=7	587221	556196	1384	0	29640
roary_with_megahit_and_isolates	dayhoff	k=11	587221	433832	122855	893	29640
roary_with_megahit_and_isolates	dayhoff	k=13	587221	401952	153666	1962	29640
roary_with_megahit_and_isolates	dayhoff	k=15	587221	392151	158450	6979	29640
roary_with_megahit_and_isolates	dayhoff	k=17	587221	384400	162528	10653	29640
roary_with_megahit_and_isolates	protein	k=10	587221	402173	155236	171	29640
roary_with_megahit_and_isolates	protein	k=6	587221	435704	121876	0	29640
roary_with_megahit_and_isolates	protein	k=7	587221	421721	135860	0	29640
ruminococcusB	protein	k=6	587221	398784	158796	0	29640
ruminococcusB	protein	k=7	587221	386101	171479	0	29640

database	alphabet	ksize	coding	non-coding	too short	stop codon
protein	f___Lachnospiraceae	k=10	72	23	0	5
protein	f___Lachnospiraceae	k=6	95	0	0	5
protein	f___Lachnospiraceae	k=7	84	11	0	5
protein	p___Firmicutes_A	k=10	75	20	0	5
protein	p___Firmicutes_A	k=7	93	2	0	5
dayhoff	plass_assembly	k=15	92	1	1	5
dayhoff	plass_assembly	k=17	92	2	2	5
protein	plass_assembly	k=10	94	1	0	5
protein	plass_assembly	k=7	95	0	0	5
dayhoff	roary	k=11	74	21	0	5
dayhoff	roary	k=13	68	26	0	5
dayhoff	roary	k=15	67	27	1	5
dayhoff	roary	k=17	65	28	2	5
protein	roary	k=10	68	26	0	5
protein	roary	k=6	74	21	0	5
protein	roary	k=7	72	23	0	5
protein	ruminococcusB	k=6	68	27	0	5
protein	ruminococcusB	k=7	66	29	0	5

Controls

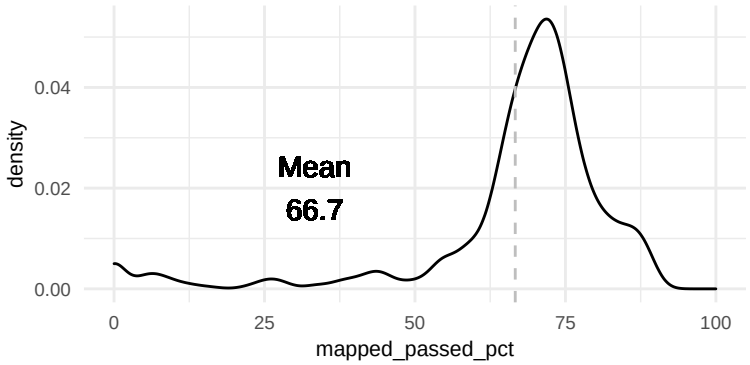
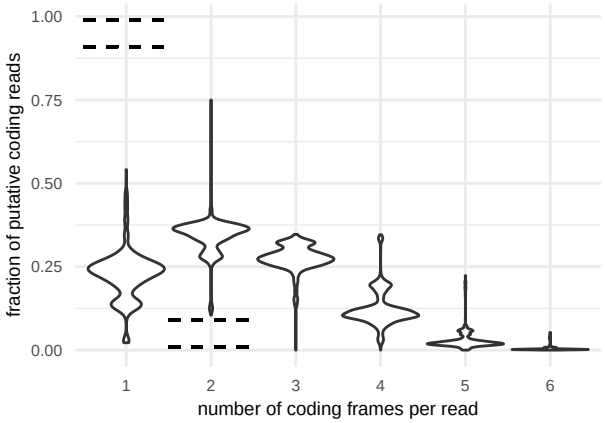


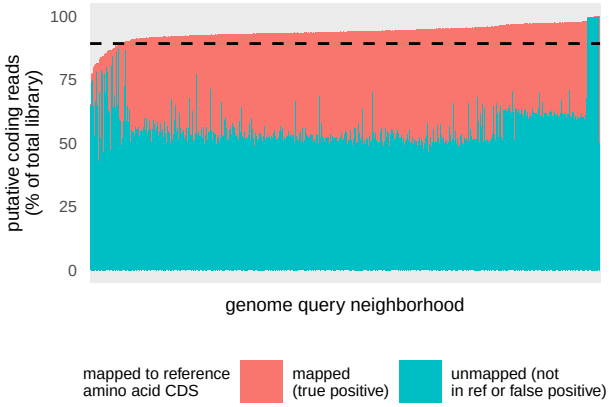
Figure 1: Mean percent mapped reads against AA reference pangenome. On average, **66.7% of reads** mapped against the pagenome reference using the paladin amino acid mapper. This number estimates the lower limit of reads that should be protein coding; at least this many reads should.

Experimental DB1: PLASS assembly

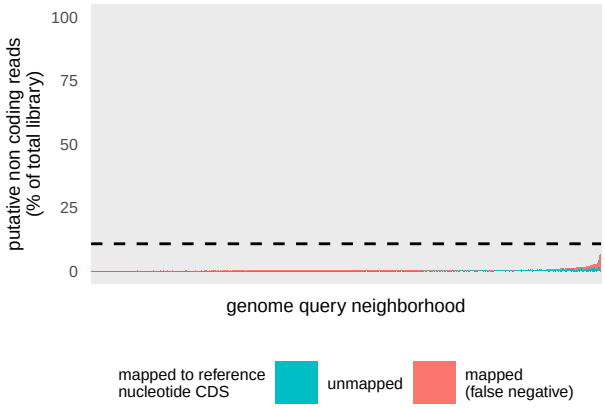
db = plass_assembly, alphabet = protein, k = 7



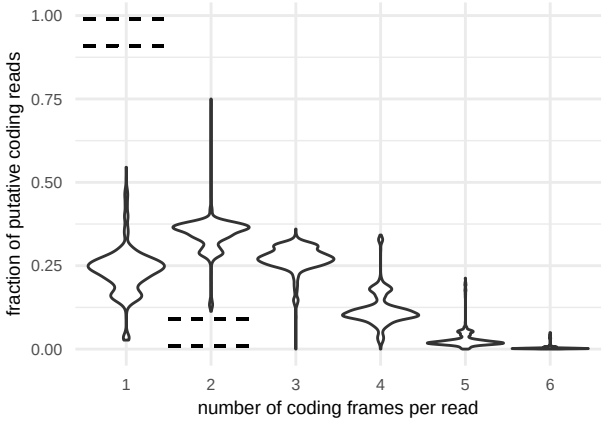
db = plass_assembly, alphabet = protein, k = 7



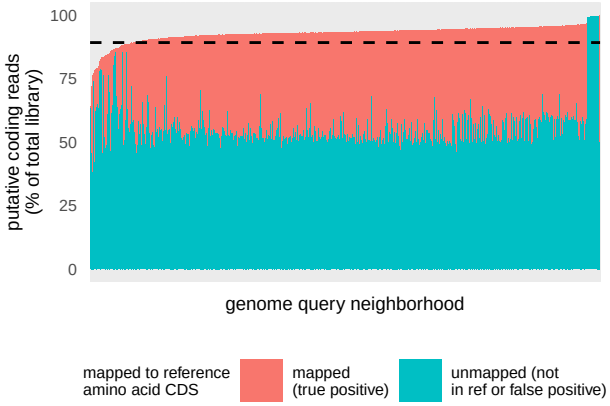
db = plass_assembly, alphabet = protein, k = 7



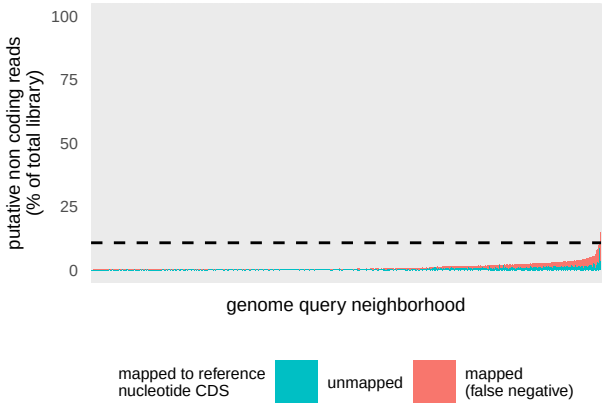
db = plass_assembly, alphabet = protein, k = 10



db = plass_assembly, alphabet = protein, k = 10

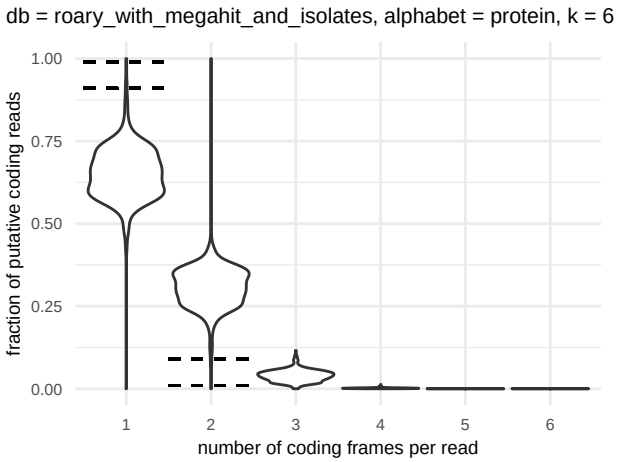


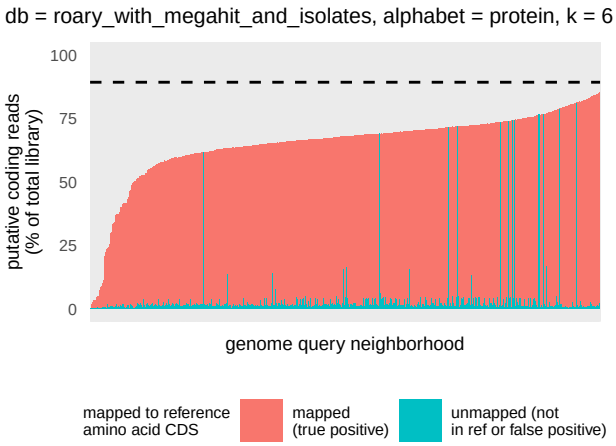
db = plass_assembly, alphabet = protein, k = 10

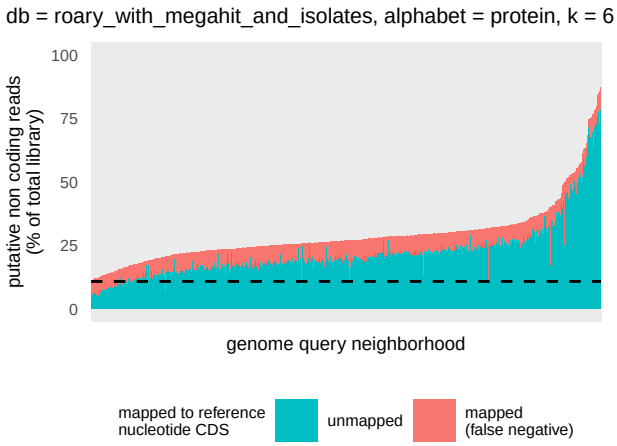


Experimental DB2: Reference pangenome

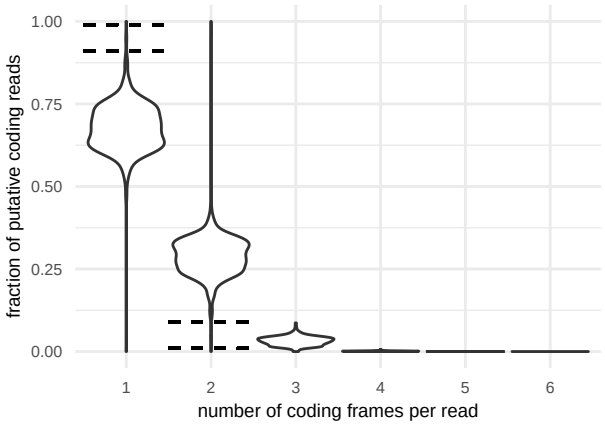
- The reference pangenome is ~37k prokka predicted protein sequences from 76 *R. gnavus* isolates from RefSeq and megahit assemblies of R. gnavus sgc nbhds



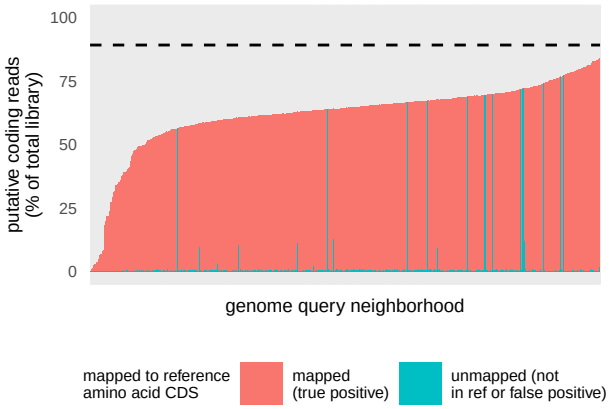




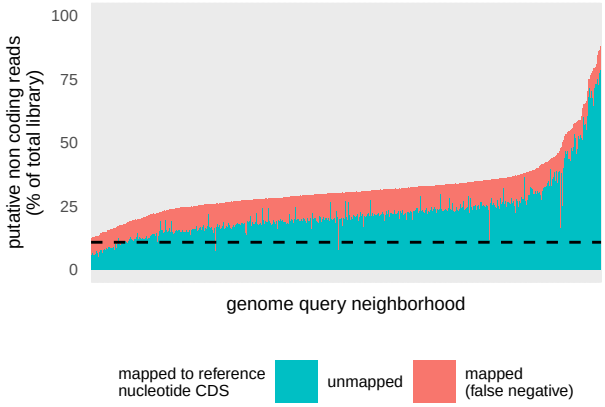
db = roary_with_megahit_and_isolates, alphabet = protein, k = 7



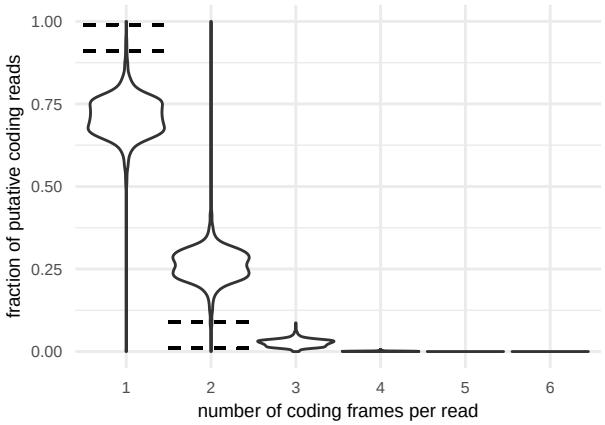
db = roary_with_megahit_and_isolates, alphabet = protein, k = 7



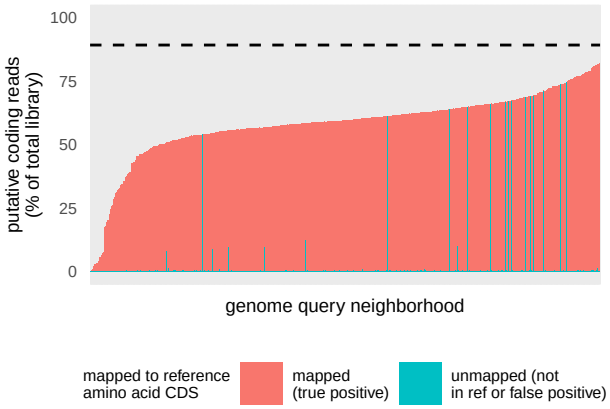
db = roary_with_megahit_and_isolates, alphabet = protein, k = 7



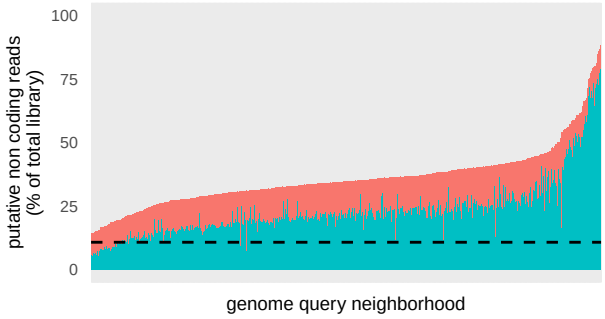
db = roary_with_megahit_and_isolates, alphabet = protein, k = 10



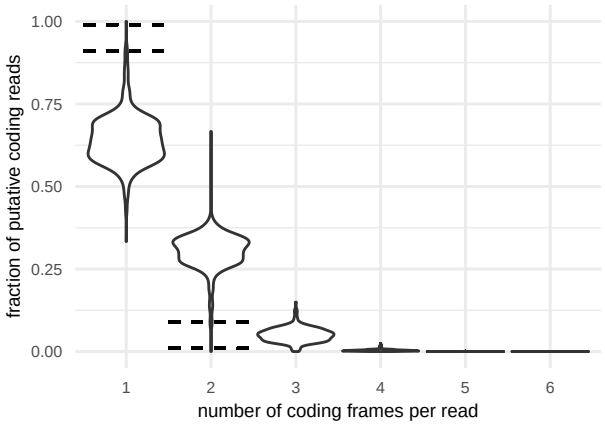
db = roary_with_megahit_and_isolates, alphabet = protein, k = 1

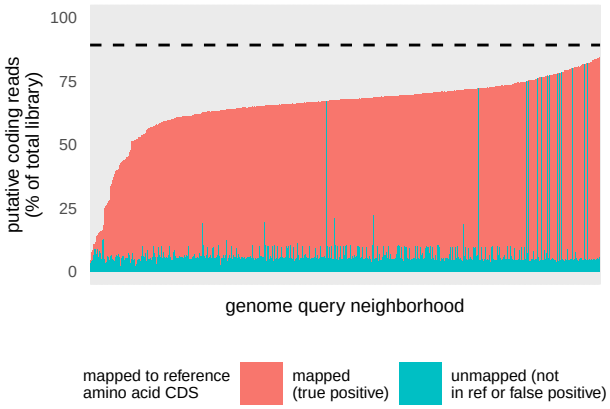


db = roary_with_megahit_and_isolates, alphabet = protein, k = 10

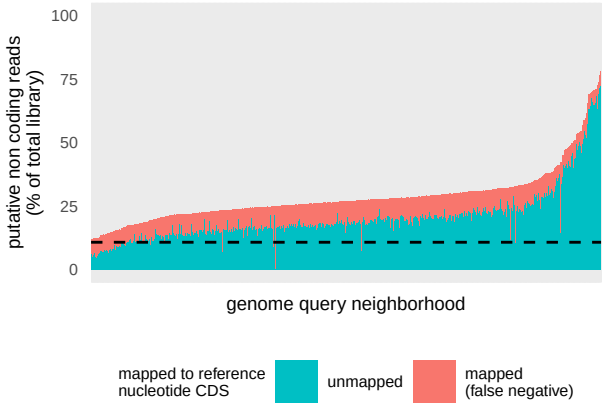


db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 11db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 1

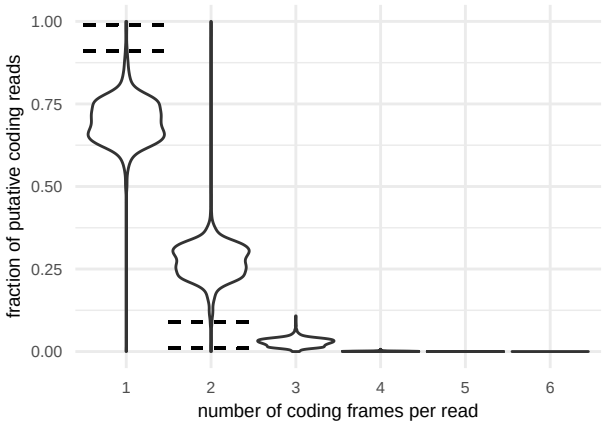


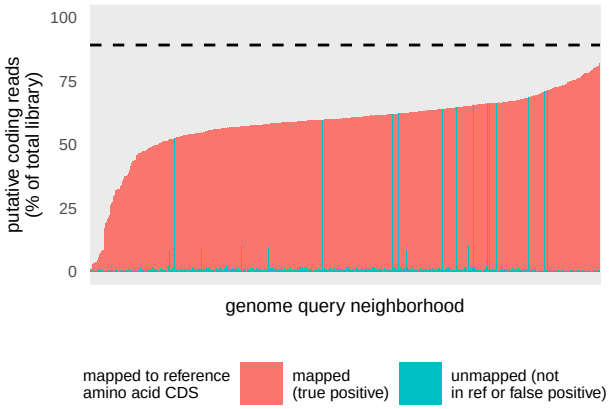


db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 11

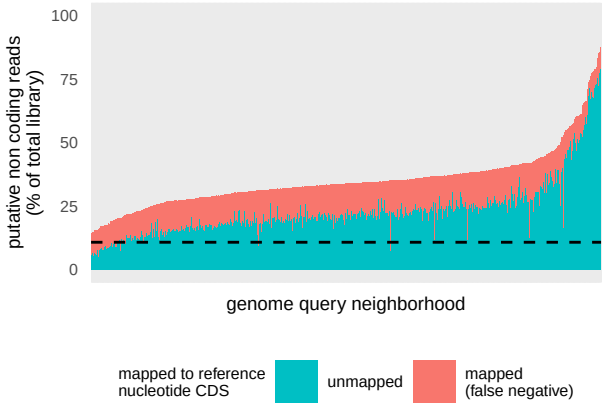


db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 13db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 1

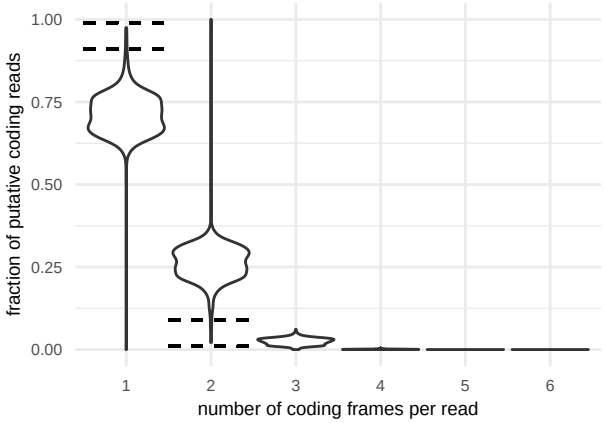




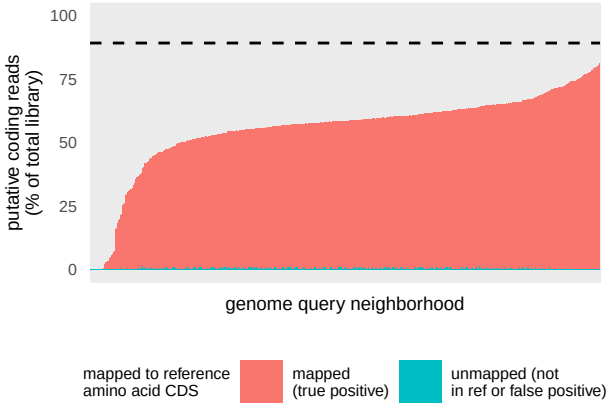
db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 13



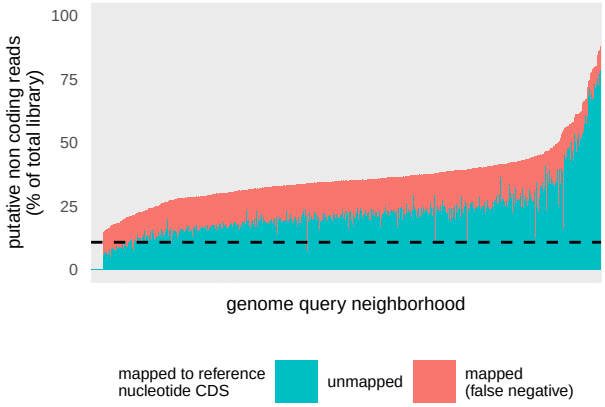
db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 15



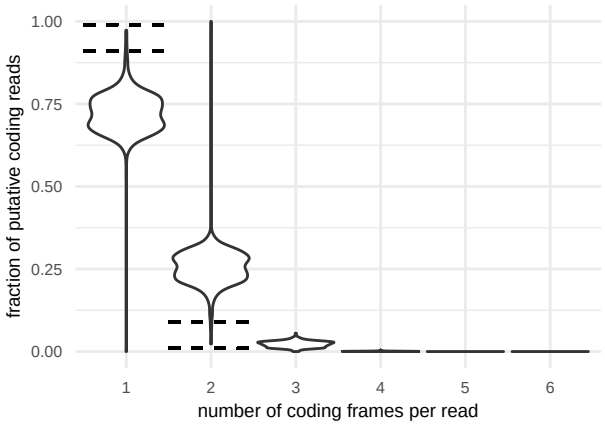
db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 1

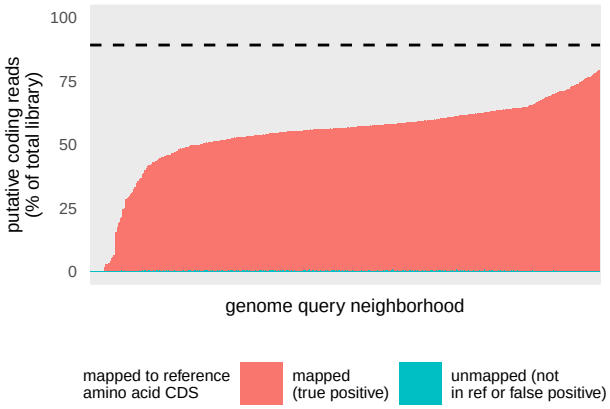


db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 15

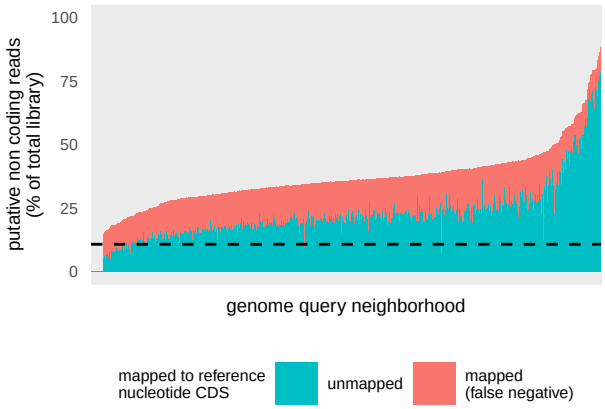


db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 17db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 1



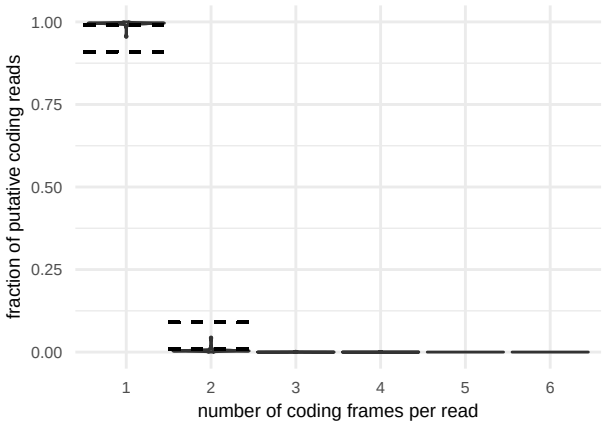


db = roary_with_megahit_and_isolates, alphabet = dayhoff, k = 17

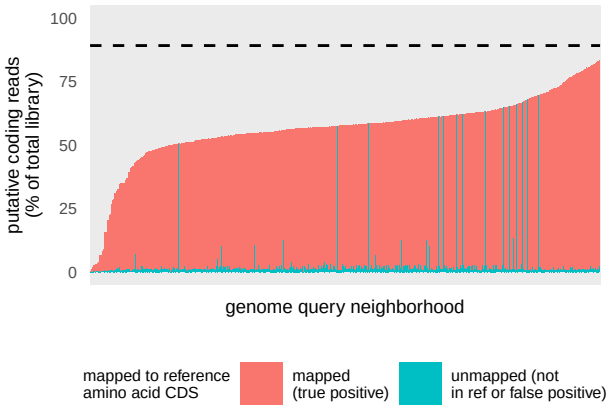


Experimental DB3: GTDB ruminococcusB

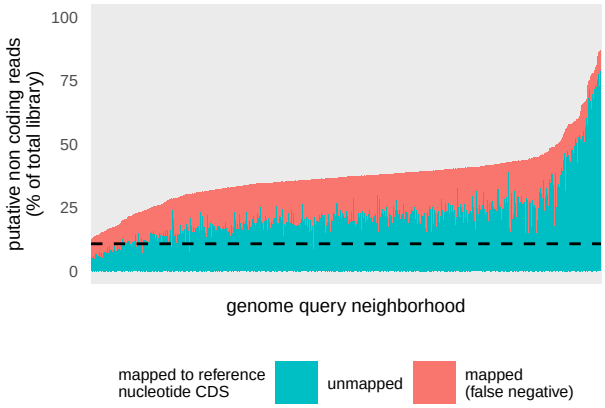
db = ruminococcusB, alphabet = protein, k = 6



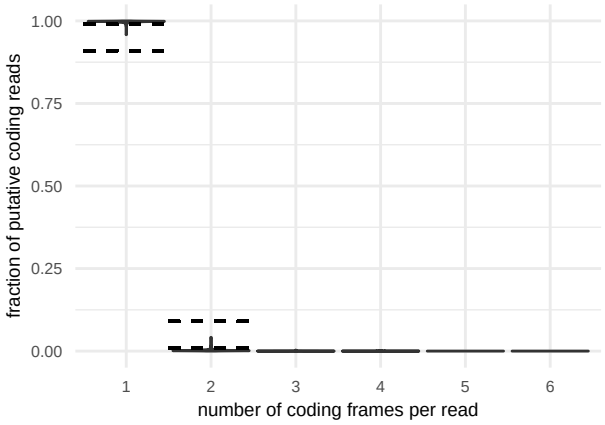
db = ruminococcusB, alphabet = protein, k = 6



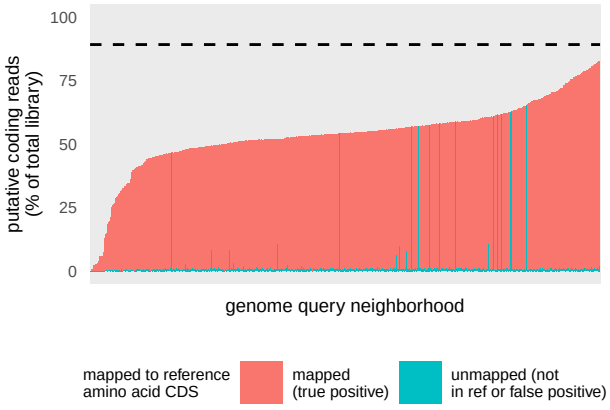
db = ruminococcusB, alphabet = protein, k = 6



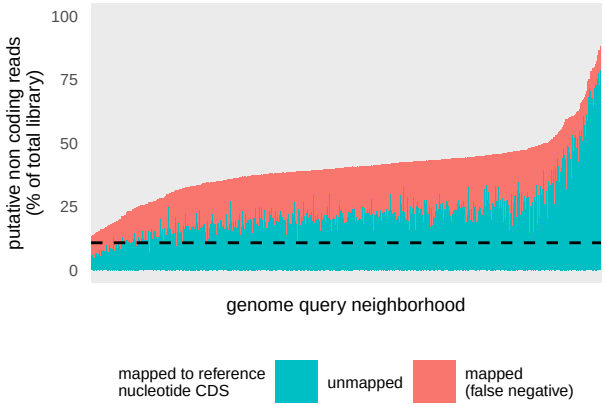
db = ruminococcusB, alphabet = protein, k = 7



db = ruminococcusB, alphabet = protein, k = 7

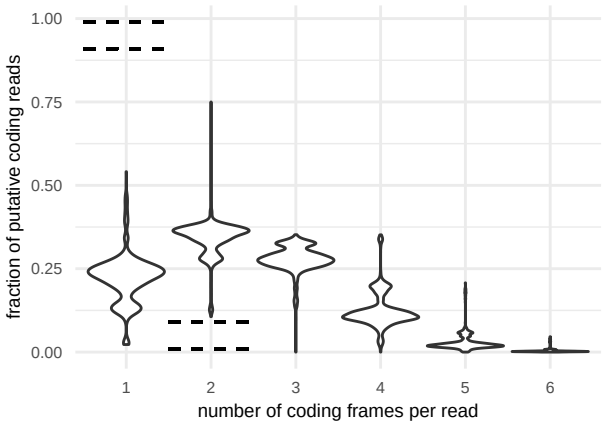


db = ruminococcusB, alphabet = protein, k = 7

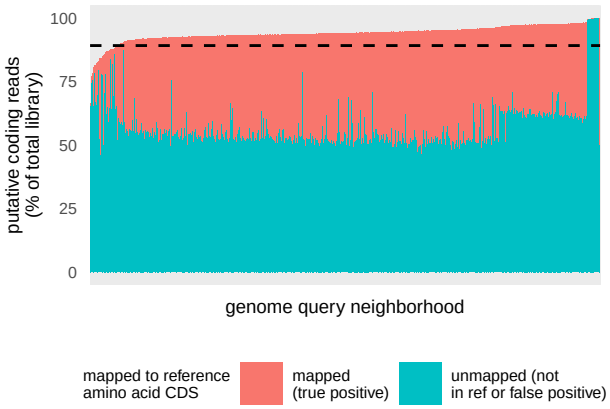


Experimental DB4: GTDB family f___Lachnospiraceae

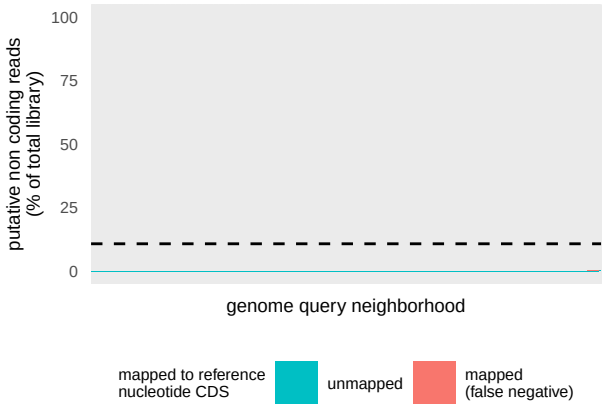
db = f__Lachnospiraceae, alphabet = protein, k = 6



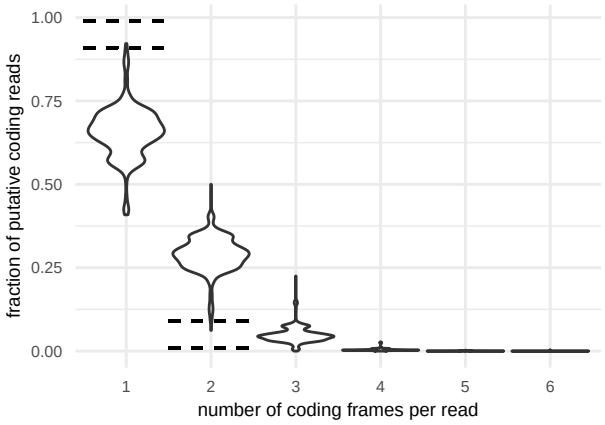
db = f__Lachnospiraceae, alphabet = protein, k = 6



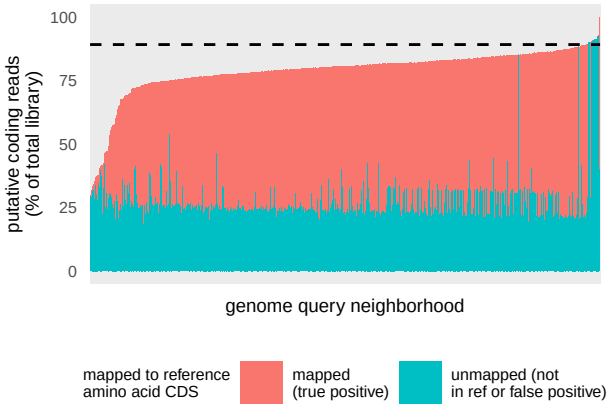
db = f__Lachnospiraceae, alphabet = protein, k = 6



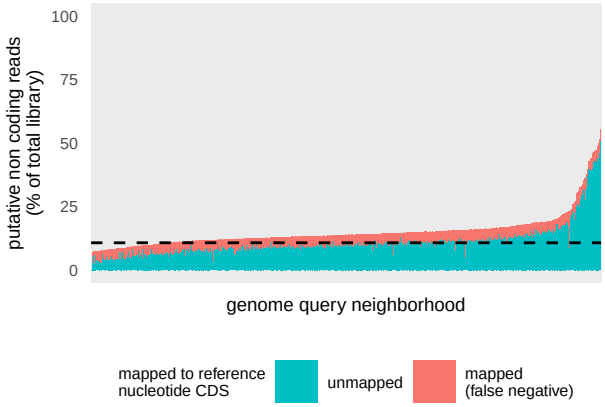
db = f__Lachnospiraceae, alphabet = protein, k = 7



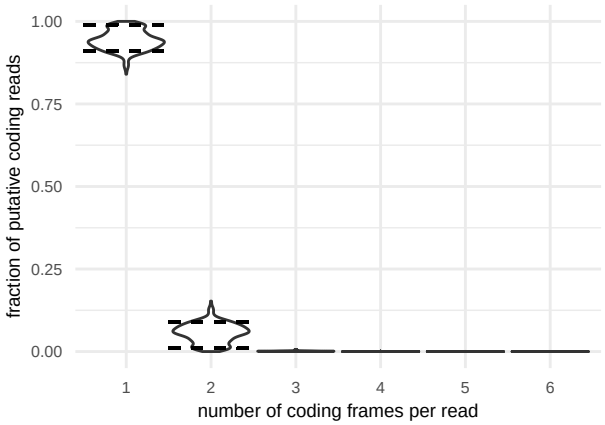
db = f__Lachnospiraceae, alphabet = protein, k = 7



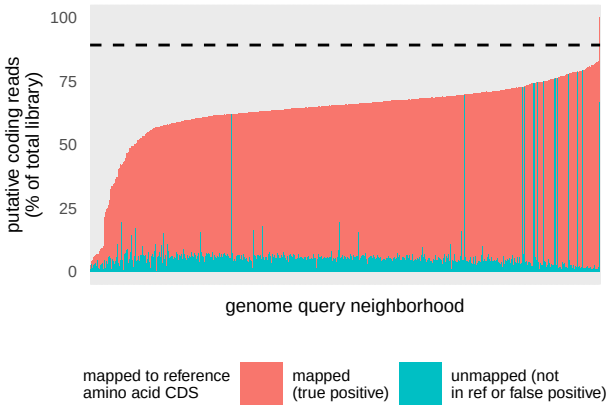
db = f__Lachnospiraceae, alphabet = protein, k = 7



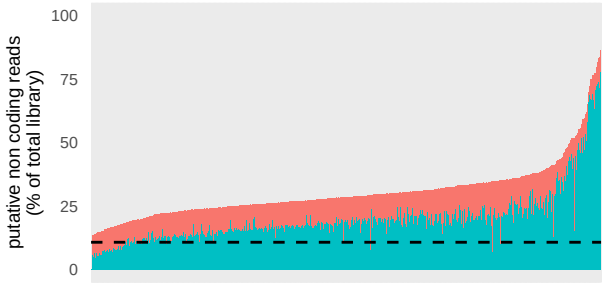
db = f__Lachnospiraceae, alphabet = protein, k = 10



db = f__Lachnospiraceae, alphabet = protein, k = 10

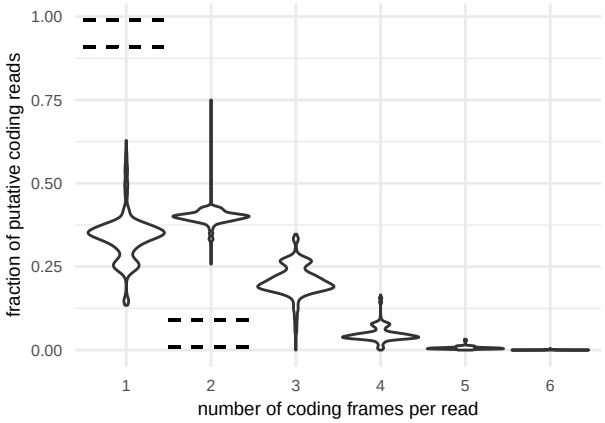


db = f__Lachnospiraceae, alphabet = protein, k = 10

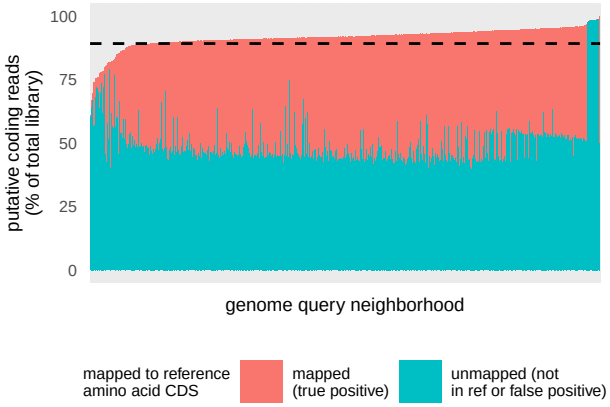


Experimental DB5: GTDB family p___Firmicutes_A

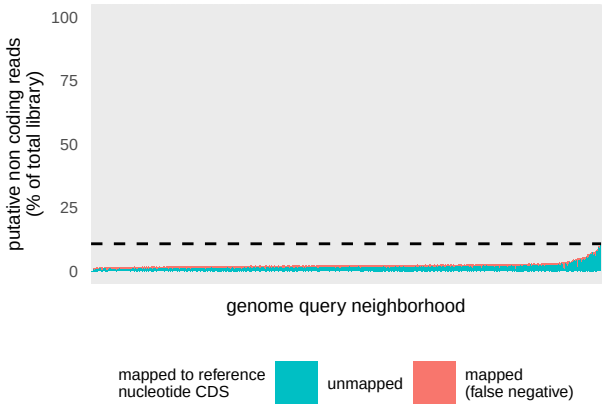
db = p__Firmicutes_A, alphabet = protein, k = 7



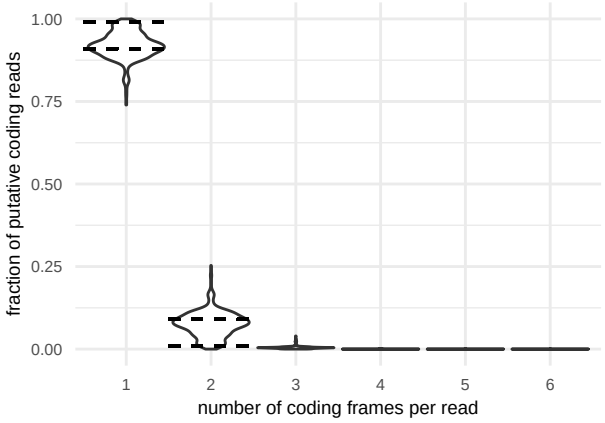
db = p__Firmicutes_A, alphabet = protein, k = 7



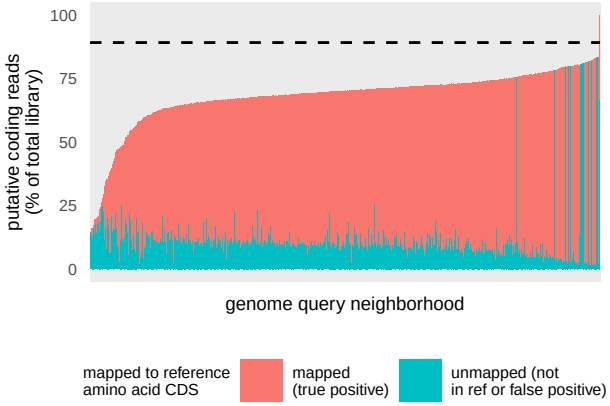
db = p__Firmicutes_A, alphabet = protein, k = 7



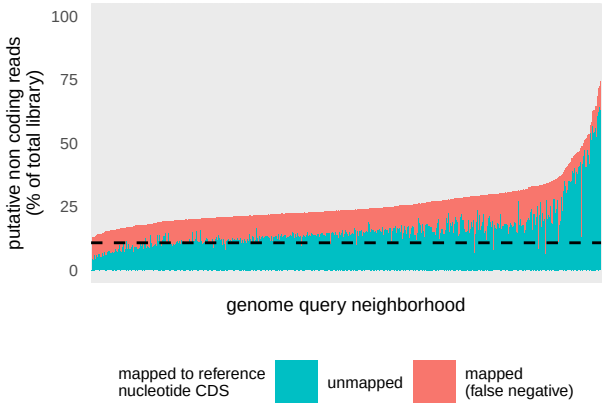
db = p__Firmicutes_A, alphabet = protein, k = 10



db = p__Firmicutes_A, alphabet = protein, k = 10



db = p__Firmicutes_A, alphabet = protein, k = 10



Conclusions

- PLASS is too promiscuous of an assembler to use to generate the reference DB; it leads to too many off-ORF predictions for reads.
- The reference pangenome seems to work pretty well, it marginally outperforms just mapping reads with paladin, and all the coding reads map to the CDS in the pangenome.

TODO

- GTDB reps/GTDB as db
- Search for shine dalgarno sequences in non-coding reads (consensus seq AGGAGG). Although SD seqs are only 8 bp ahead of AUG (start codon), so may not be in majorit of seq. May be worth looking into other conserved intergenic seqs.

References

- Land et al. 2015: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4361730/
- Ardern et al. 2020: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7457138/
- Zehentner et al. 2020: https://www.biorxiv.org/content/10.1101/2020.11.18.388249v1.full.pdf

Supplementary figure panels