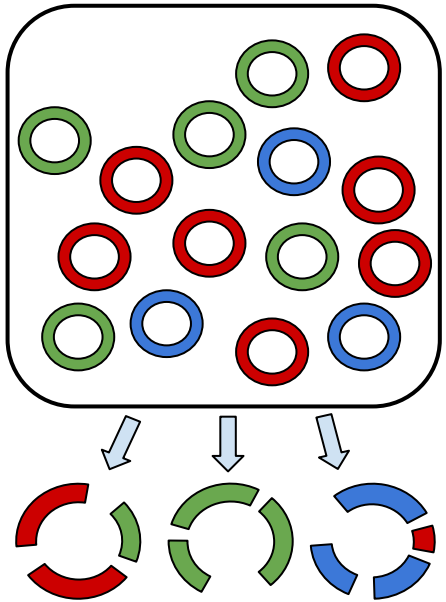
Estimating differential abundance profiles for metagenomic series binning

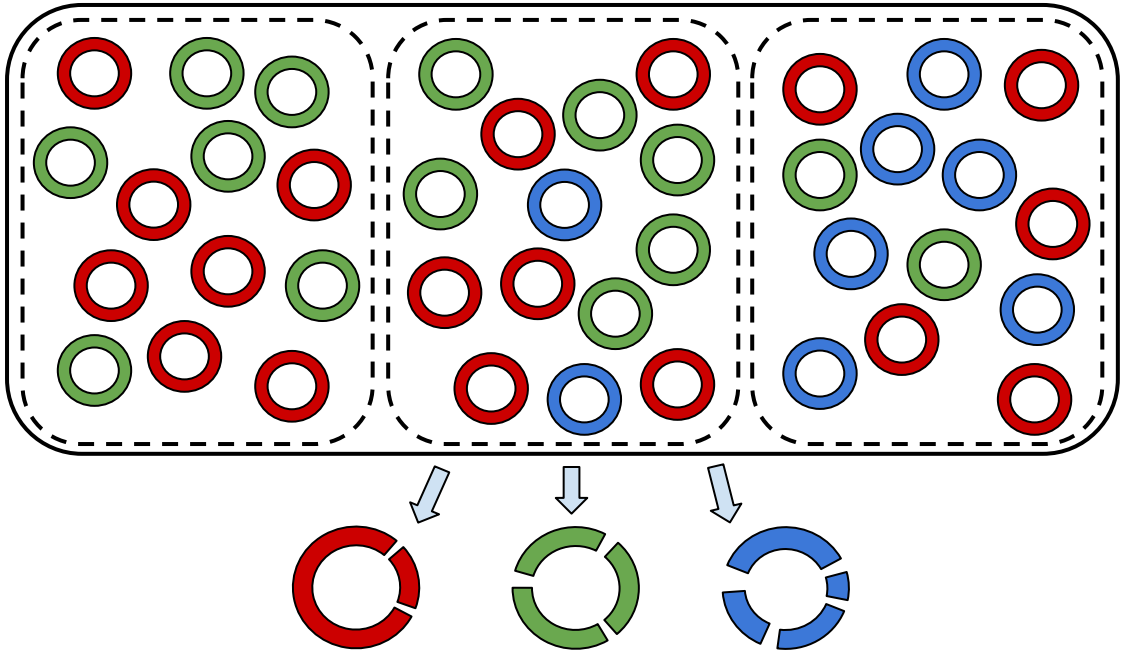
Krivonosova Kristina

Yuri Gorshkov, Sergey Nurk

Metagenome-assembled genomes



Assembly of a single metagenome

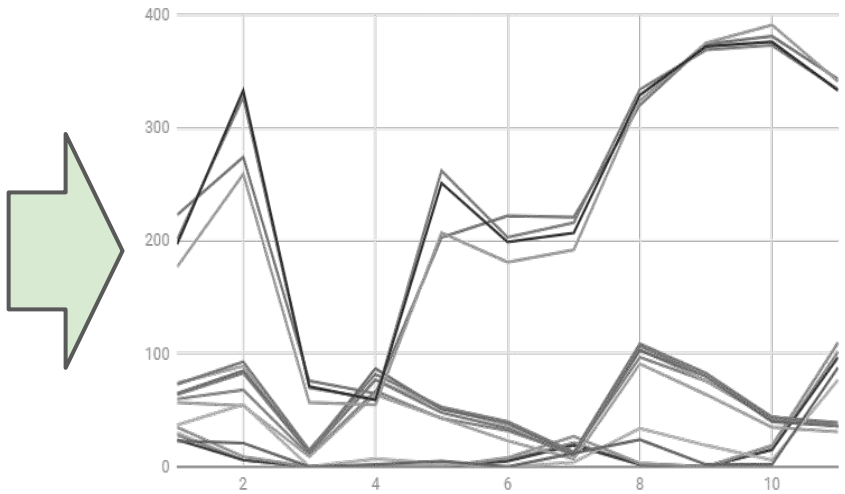


Assembly of a time series

Coverage profiles

Contig coverage values across time-points form the **coverage matrix**

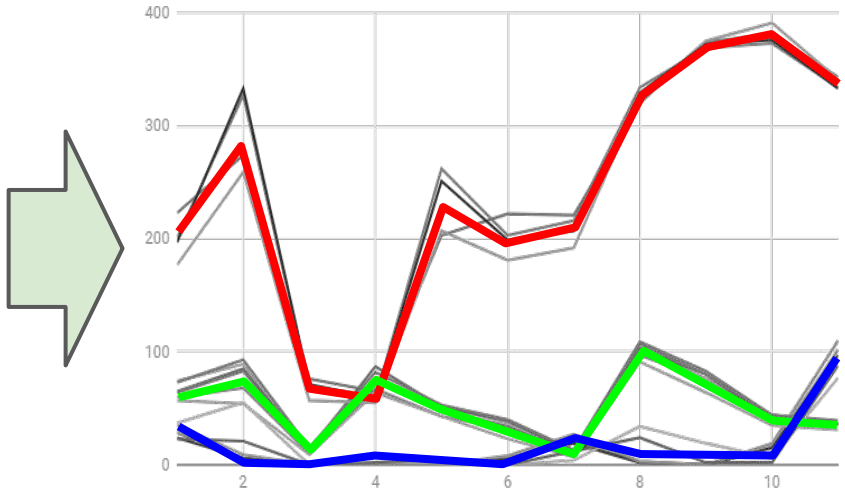
NODE_1_length_778649	201	327	70	60	262	203	216	334	369	373	334
NODE_2_length_765432	223	274	76	65	203	222	221	320	374	381	343
NODE_3_length_455494	30	6	0	0	0	6	21	3	0	17	110
NODE_4_length_443322	28	6	0	0	0	6	21	2	0	15	102
NODE_5_length_397531	24	6	0	0	0	5	19	2	0	15	97
NODE_7_length_279507	35	9	0	0	0	8	27	3	0	19	110
NODE_10_length_267494	60	68	10	67	43	32	12	97	76	42	36
NODE_11_length_260333	74	89	13	77	52	37	13	104	76	44	36
NODE_23_length_59427	65	85	13	87	51	39	11	103	80	40	36
NODE_25_length_50127	64	83	12	77	48	34	10	107	80	44	36
NODE_37_length_40540	73	93	15	82	53	40	13	109	83	44	39
NODE_32_length_45131	57	54	9	64	43	23	7	91	64	35	31
NODE_43_length_35572	37	55	0	7	3	0	4	34	19	6	77
NODE_58_length_19476	23	21	0	2	5	0	12	24	2	2	88
NODE_60_length_7786	177	259	57	55	207	181	192	324	375	391	341
NODE_62_length_7023	197	333	71	59	251	199	207	329	372	376	333



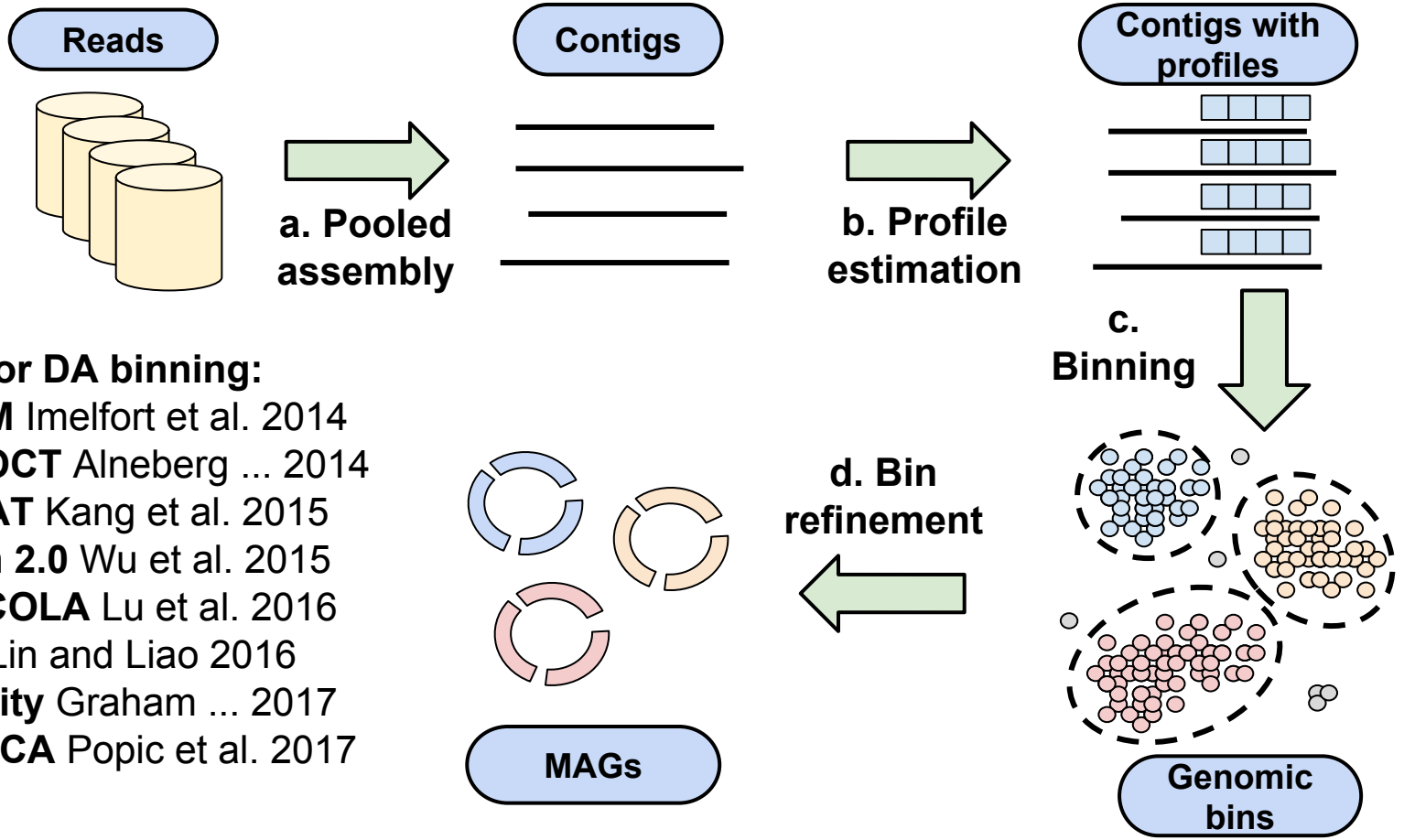
Coverage profiles

Fragments with similar coverage profiles are likely to originate from the same genome

NODE_1_length_778649	201	327	70	60	262	203	216	334	369	373	334
NODE_2_length_765432	223	274	76	65	203	222	221	320	374	381	343
NODE_3_length_455494	30	6	0	0	0	6	21	3	0	17	110
NODE_4_length_443322	28	6	0	0	0	6	21	2	0	15	102
NODE_5_length_397531	24	6	0	0	0	5	19	2	0	15	97
NODE_7_length_279507	35	9	0	0	0	8	27	3	0	19	110
NODE_10_length_267494	60	68	10	67	43	32	12	97	76	42	36
NODE_11_length_260333	74	89	13	77	52	37	13	104	76	44	36
NODE_23_length_59427	65	85	13	87	51	39	11	103	80	40	36
NODE_25_length_50127	64	83	12	77	48	34	10	107	80	44	36
NODE_37_length_40540	73	93	15	82	53	40	13	109	83	44	39
NODE_32_length_45131	57	54	9	64	43	23	7	91	64	35	31
NODE_43_length_35572	37	55	0	7	3	0	4	34	19	6	77
NODE_58_length_19476	23	21	0	2	5	0	12	24	2	2	88
NODE_60_length_7786	177	259	57	55	207	181	192	324	375	391	341
NODE_62_length_7023	197	333	71	59	251	199	207	329	372	376	333

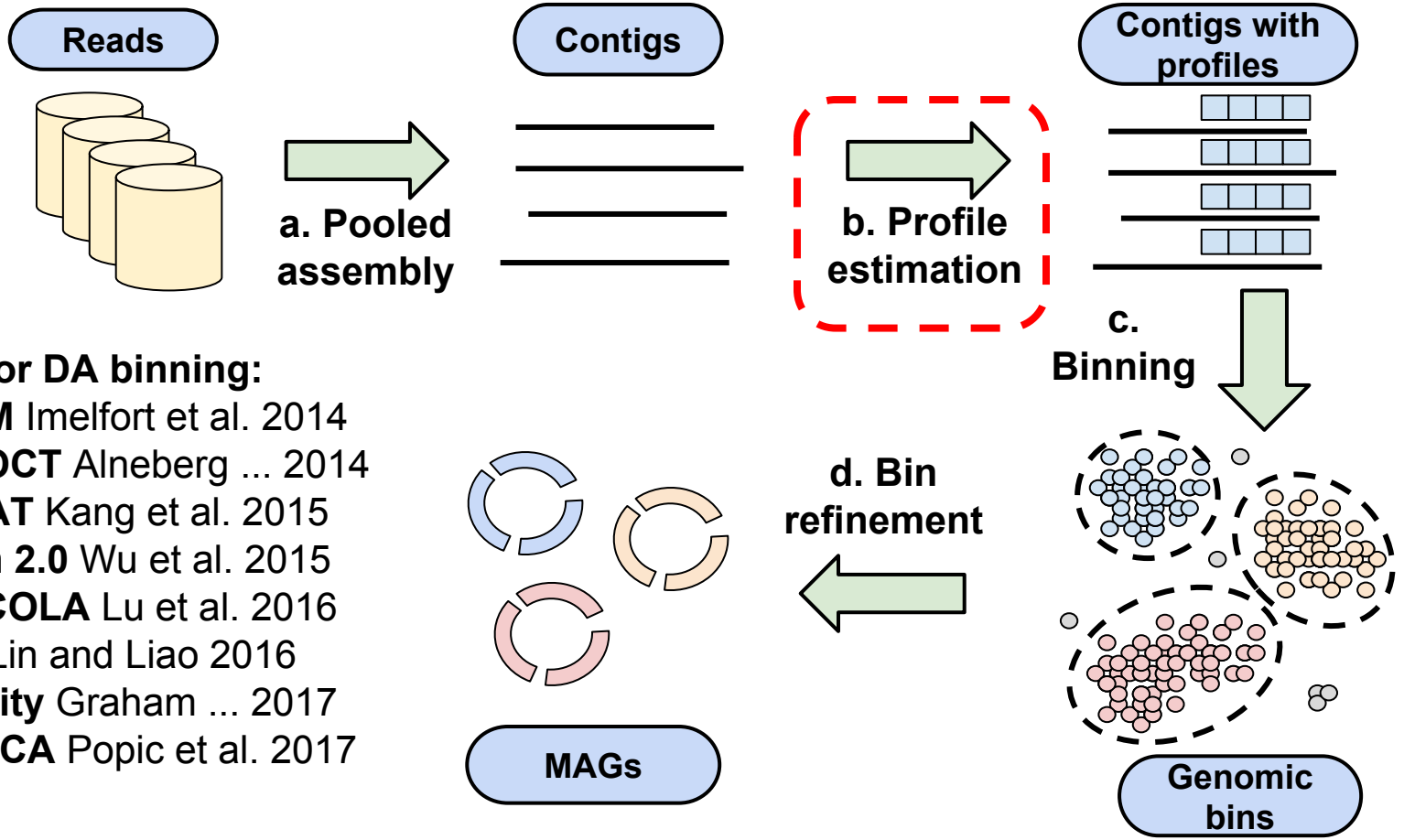


Conventional workflow



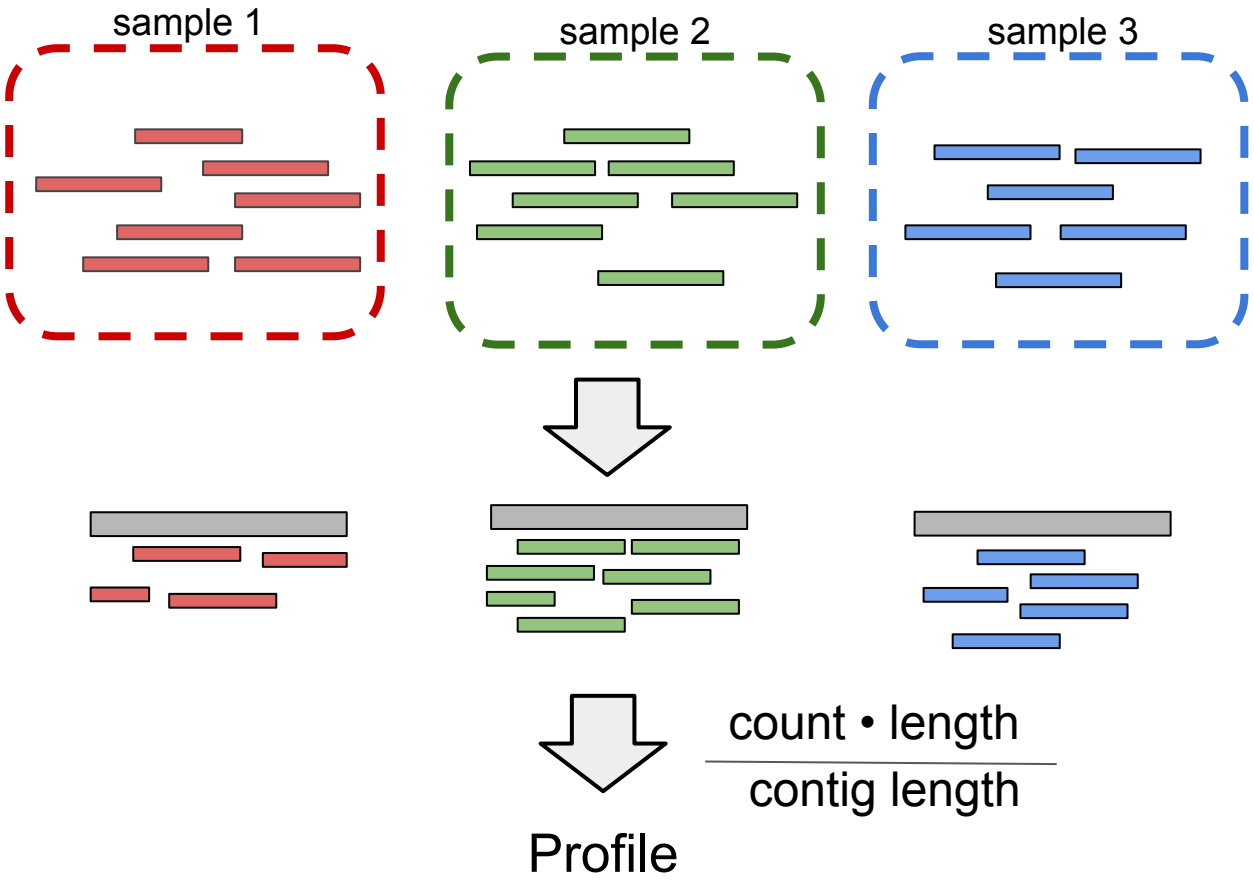
Tools for DA binning:
GroopM Imelfort et al. 2014
CONCOCT Alneberg ... 2014
MetaBAT Kang et al. 2015
MaxBin 2.0 Wu et al. 2015
COCACOLA Lu et al. 2016
MyCC Lin and Liao 2016
BinSanity Graham ... 2017
GATTACA Popic et al. 2017

Conventional workflow



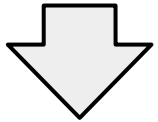
Tools for DA binning:
GroopM Imelfort et al. 2014
CONCOCT Alneberg ... 2014
MetaBAT Kang et al. 2015
MaxBin 2.0 Wu et al. 2015
COCACOLA Lu et al. 2016
MyCC Lin and Liao 2016
BinSanity Graham ... 2017
GATTACA Popic et al. 2017

Reads coverage profile



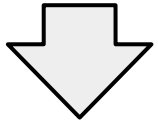
Kmers coverage profile

Sample 1 Sample 2 Sample 3



Contig_N

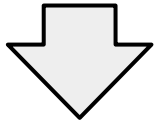
Kmer	Sample1	Sample2	Sample3
AAA..	4	14	3
TAC..	2	10	3
CAA..	3	13	4



Profile

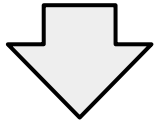
Kmers coverage profile

sample 1 sample 2 sample 3



Contig_N

Kmer	Sample1	Sample2	Sample3
AAA..	4	14	3
TAC..	2	10	3
CAA..	3	13	4



median

Profile

Goals

Improve approximation of kmers coverage profiles

Tasks

- Try different estimators for kmer profiles.
- Try different tools for binning.

Datasets

infant gut:

- 11 samples
- 10 references

perchlorate soil

- 5 samples
- 48 references

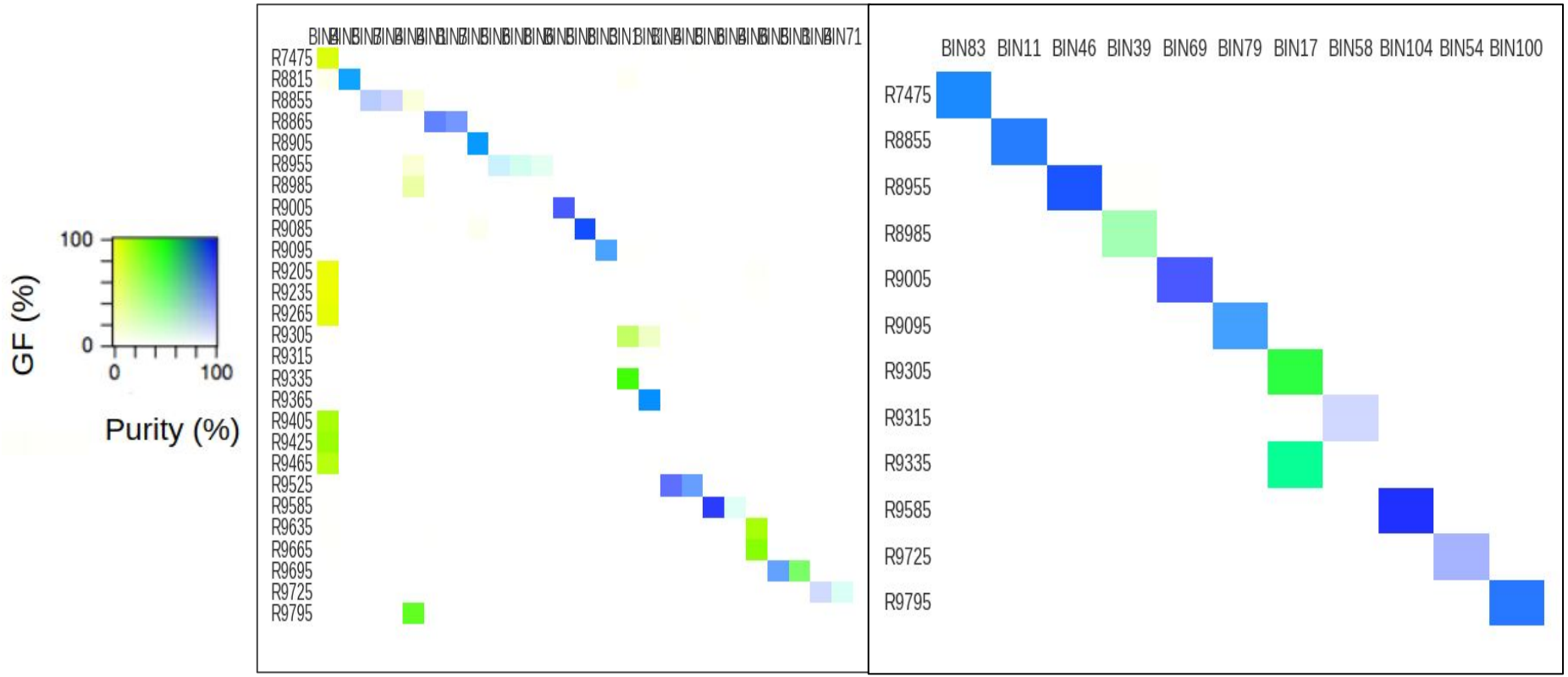
mock strains:

- 20 samples
- 19 references

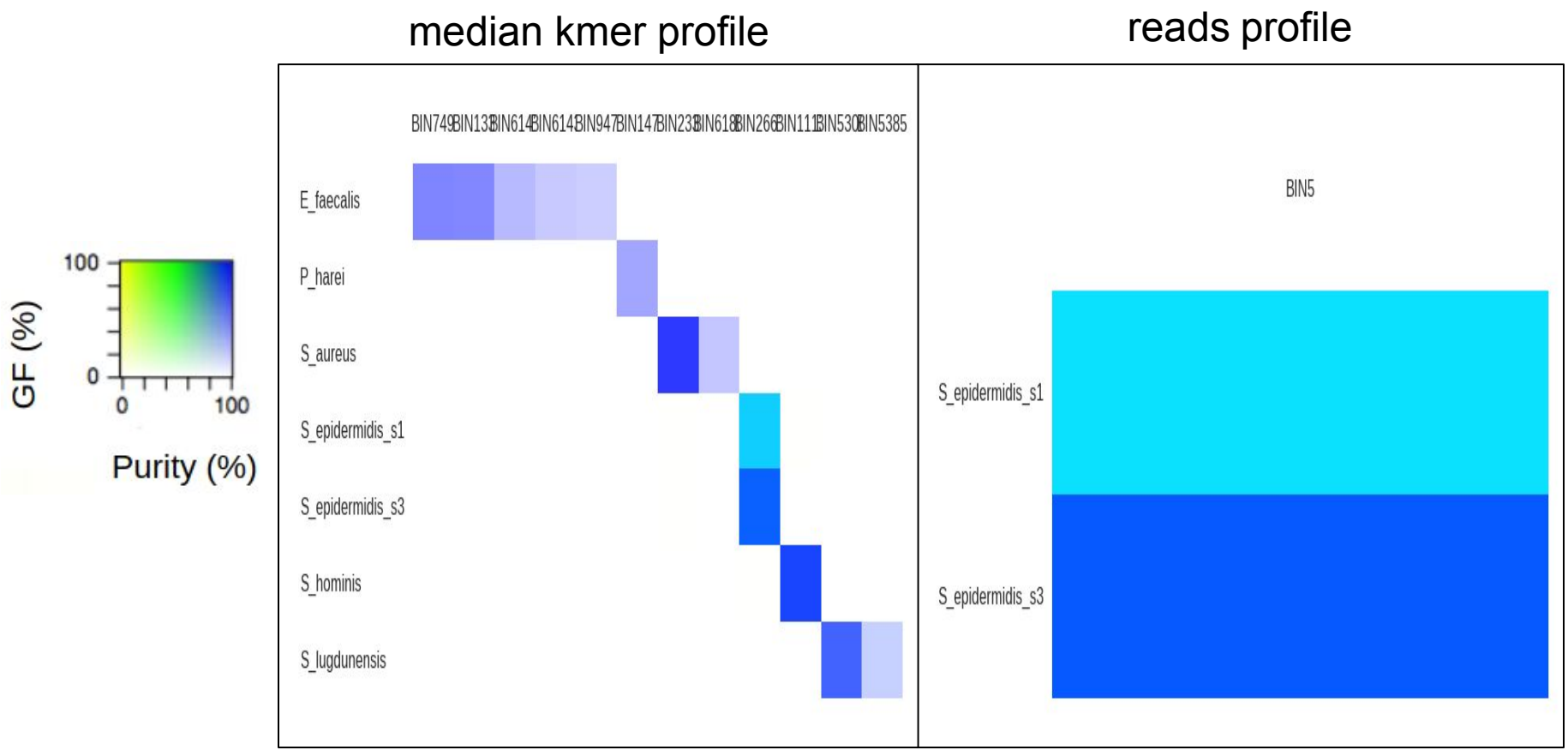
Perchlorate dataset (Before)

median kmer profile

Reads profile



Infant gut dataset (Before)

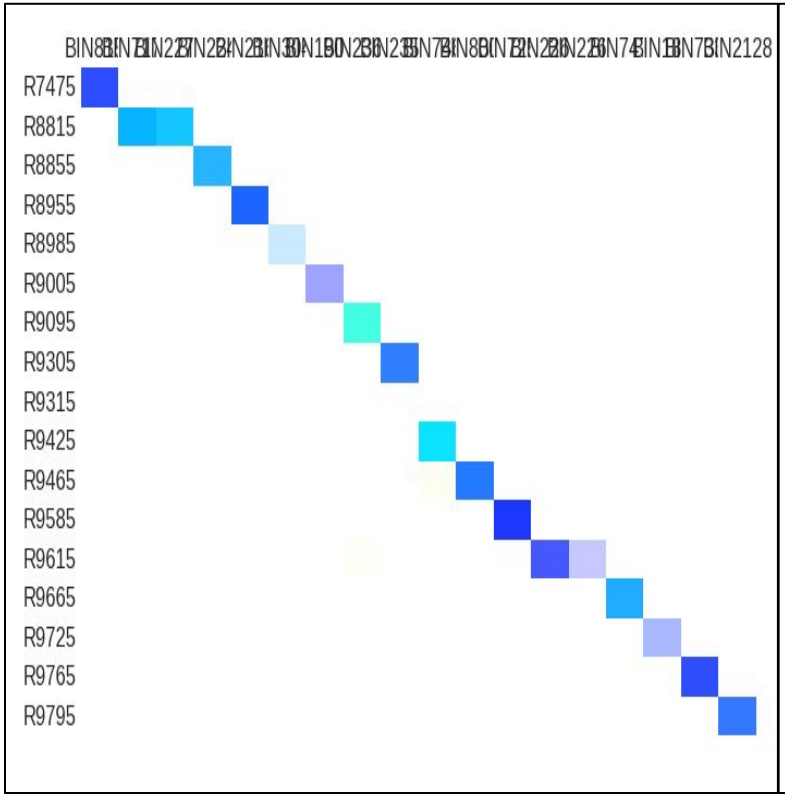


Winsorized mean

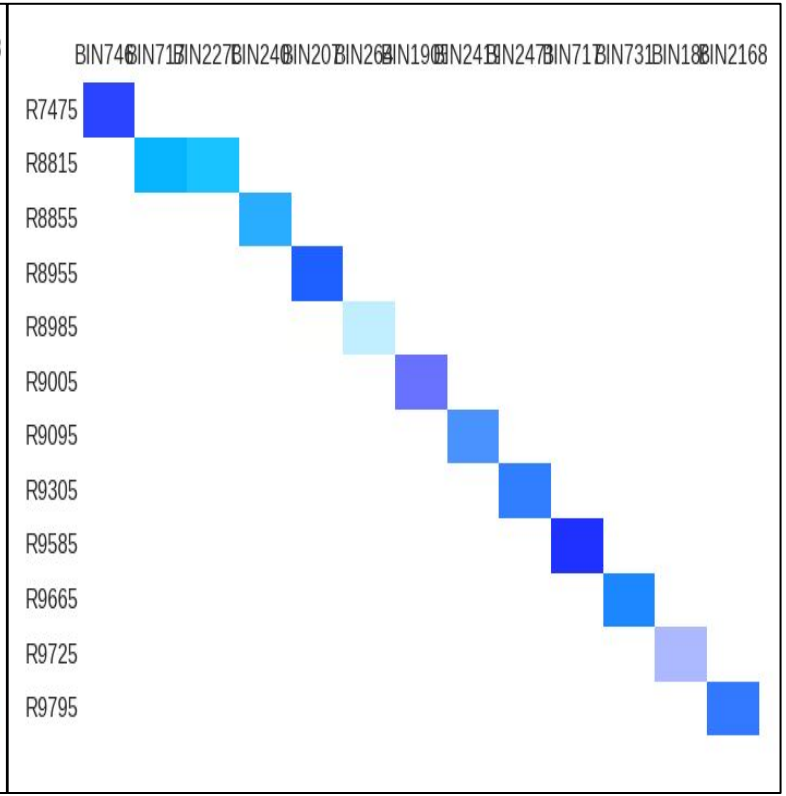
$$\frac{\overbrace{x_2 + x_2} + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + \overbrace{x_9 + x_9}}{10}.$$

Perchlorate dataset (After)

winsorized mean kmer profile



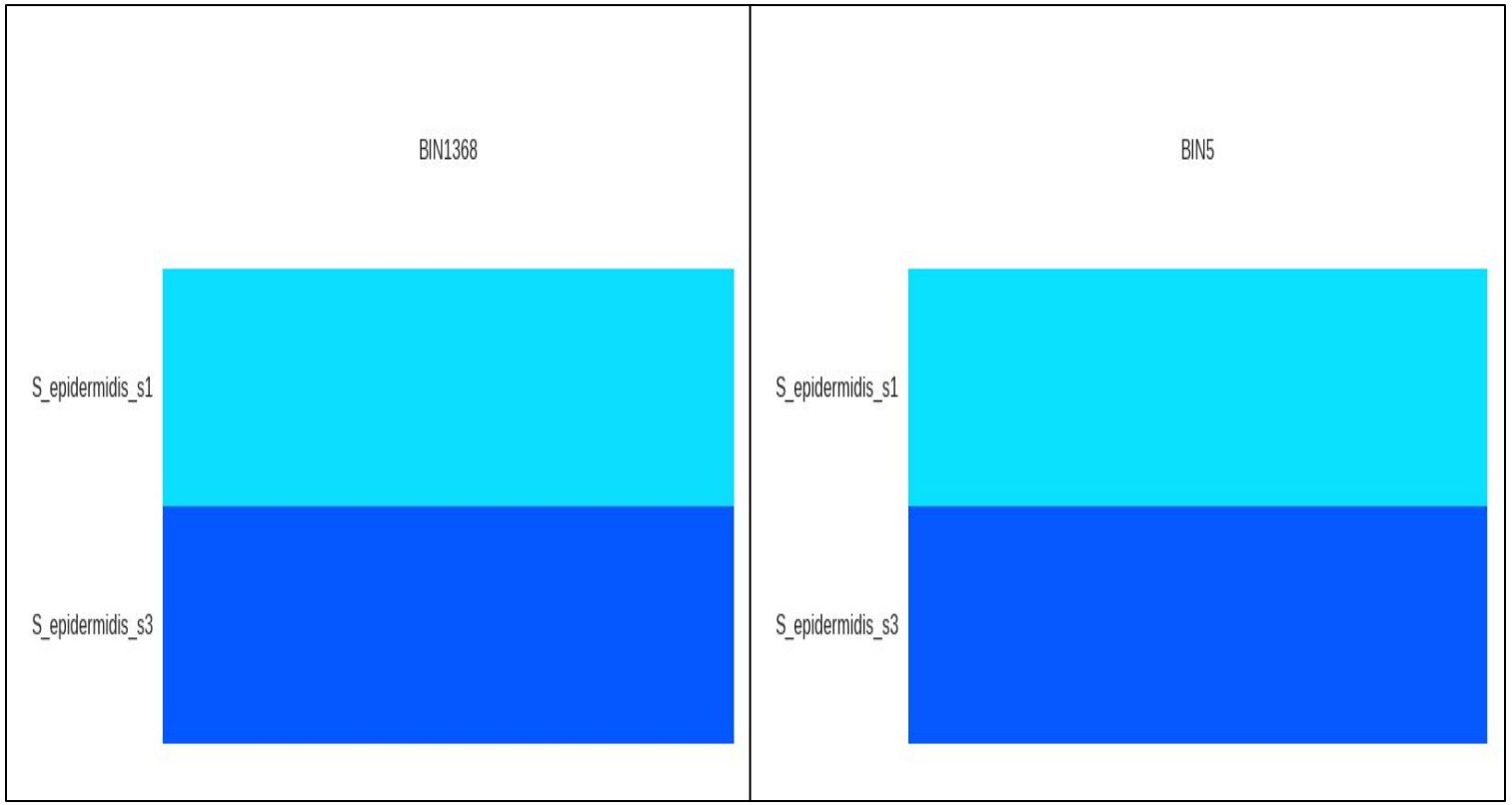
Reads profile



Infant gut dataset (After)

Winsorized mean profile

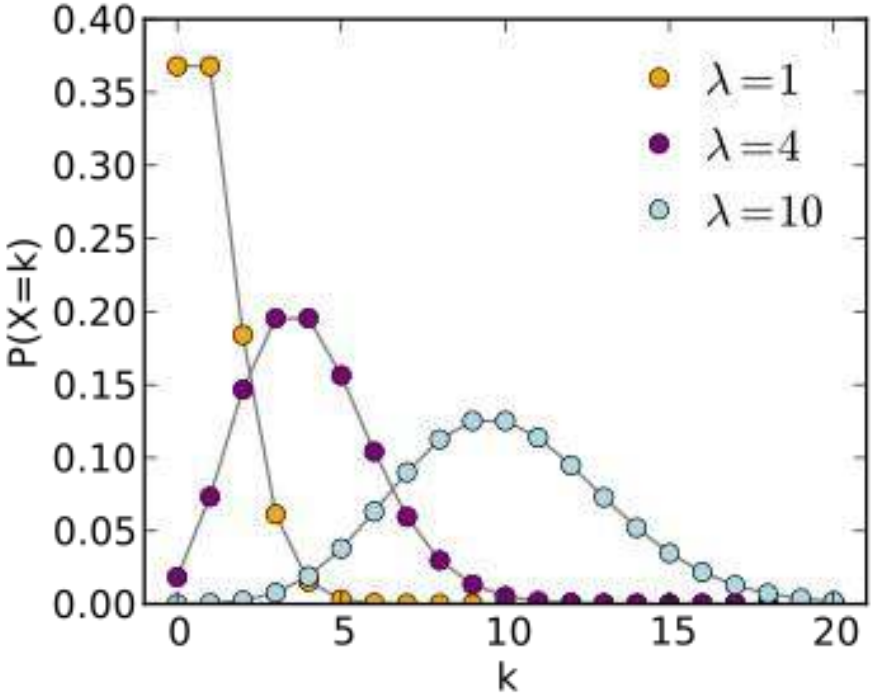
reads profile



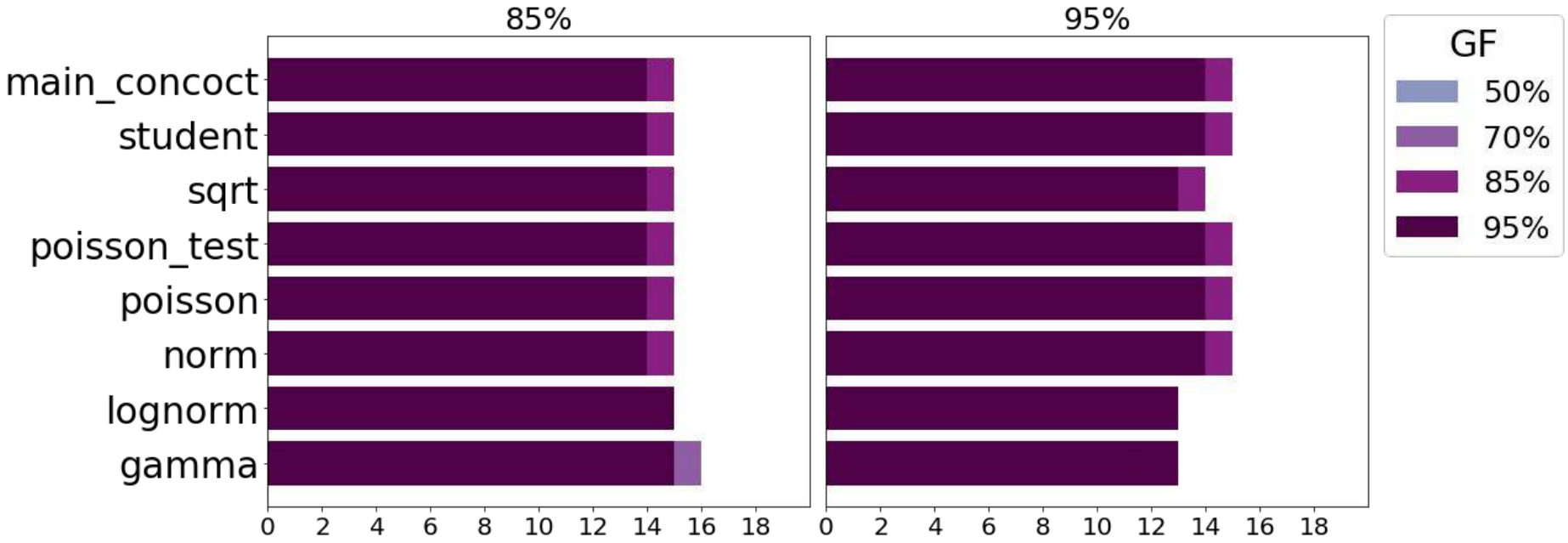
Weighted mean

$$\frac{\sum_i w_i\, x_i}{\sum_i w_i} \quad \Rightarrow \quad \frac{\sum_i P(x_i;\Theta) x_i}{\sum_i P(x_i;\Theta)}$$

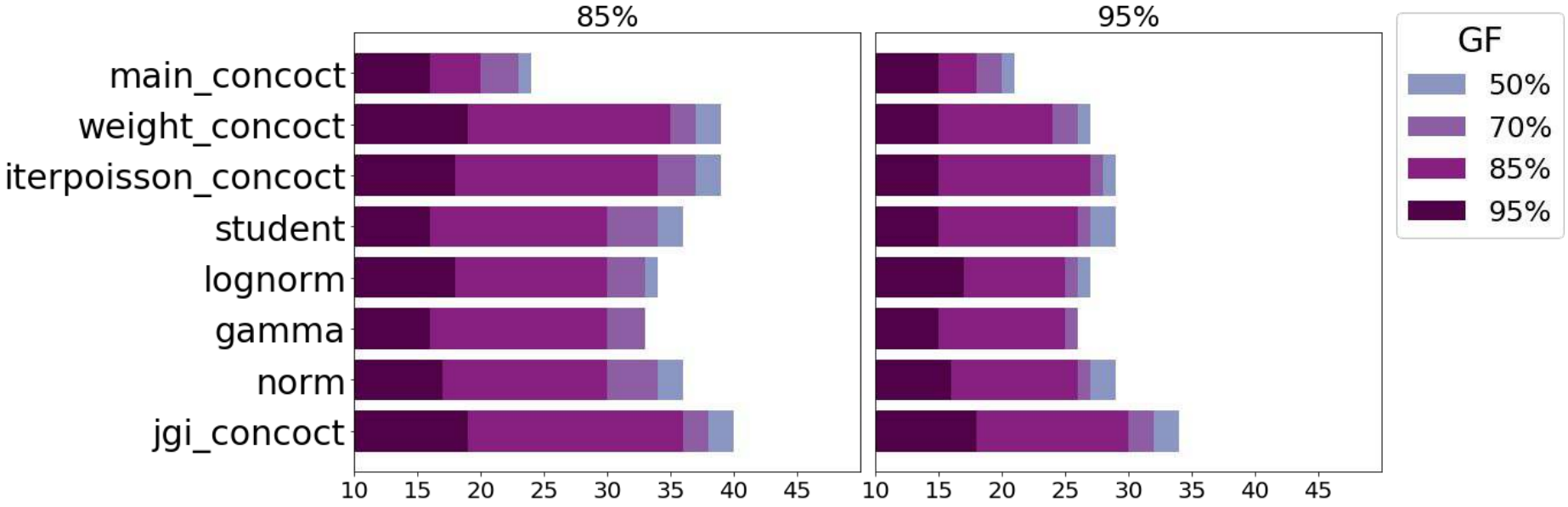
Poisson distribution



Mock stains

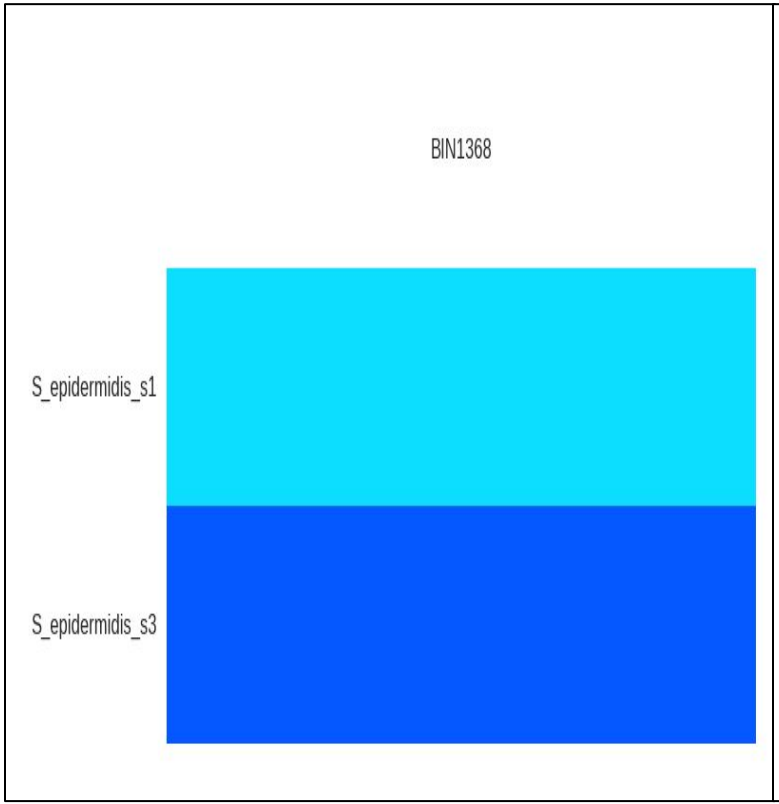

85%
95%
main_concoct
student
sqrt
poisson_test
poisson
norm
lognorm
gamma
0
2
4
6
8
10
12
14
16
18
GF
50%
70%
85%
95%

Perchlorate dataset

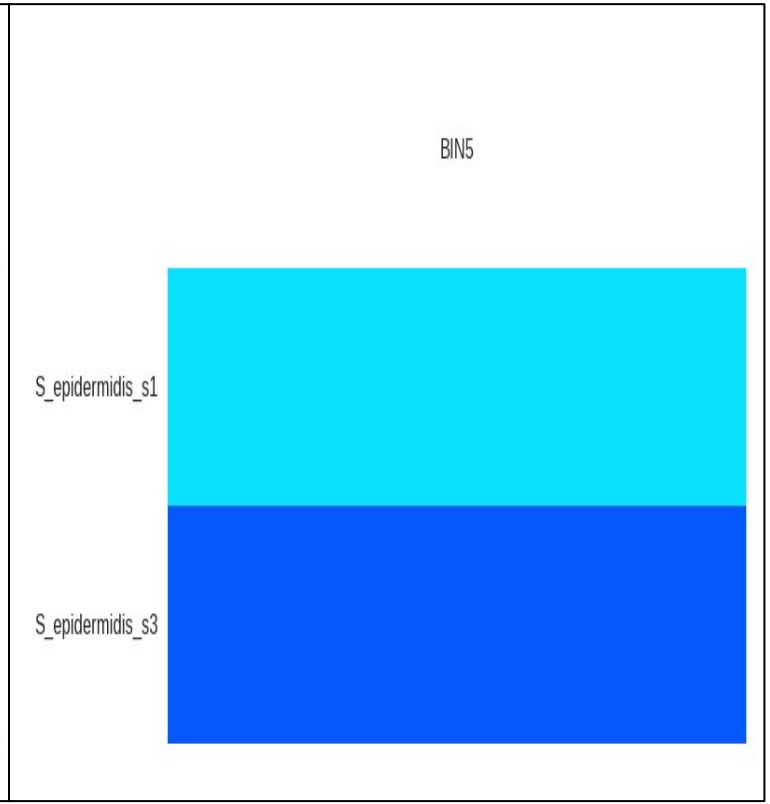


Infant gut dataset (After)

poisson iter weighted mean profile

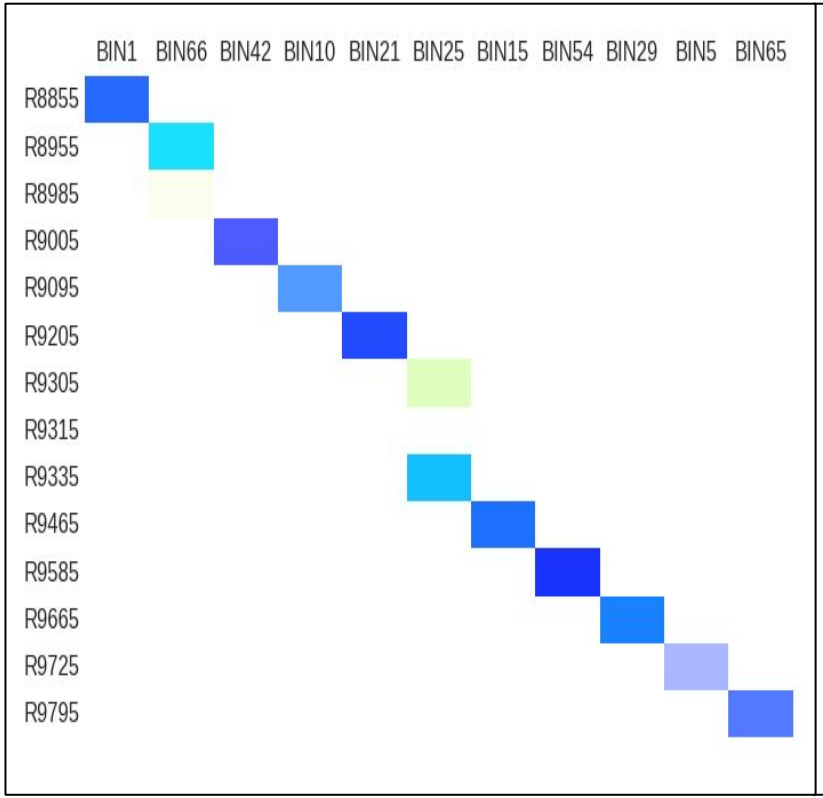


reads profile

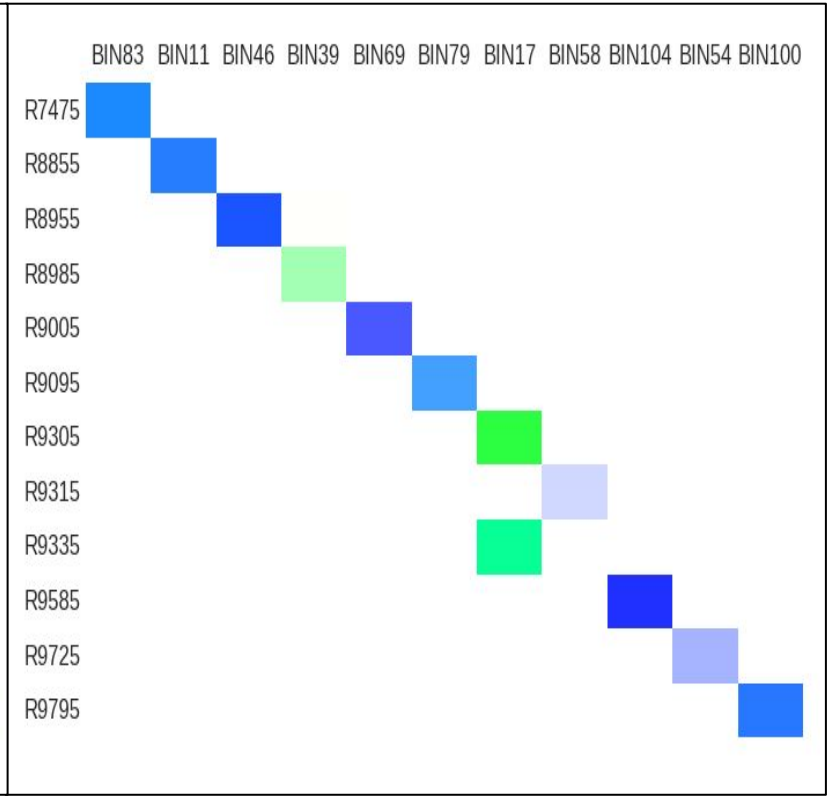


Perchlorate dataset (After)

poisson iter weighted mean profiles



reads profiles

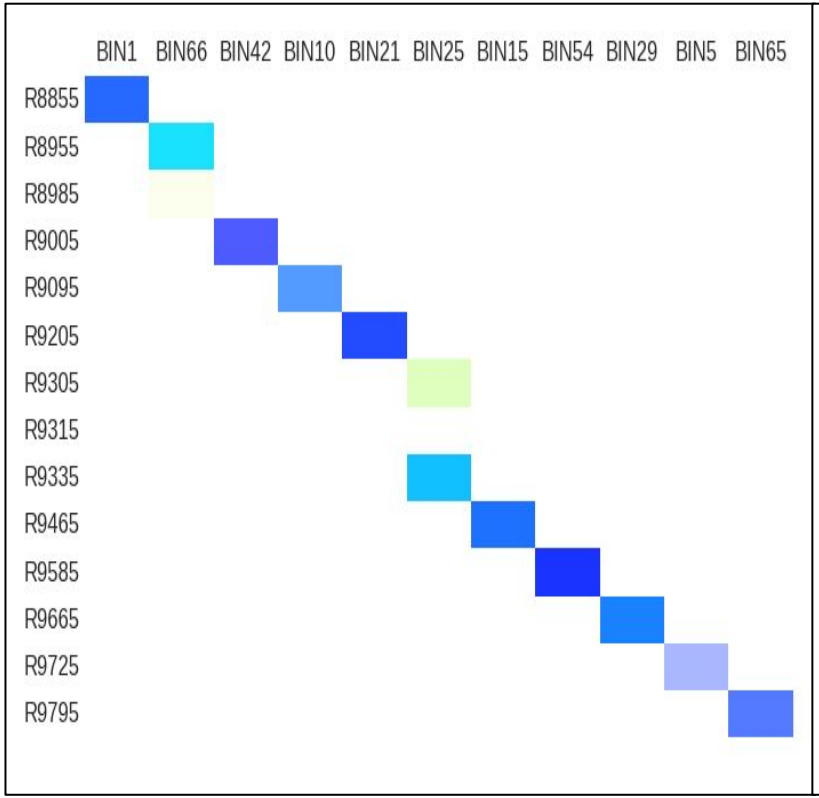


Results

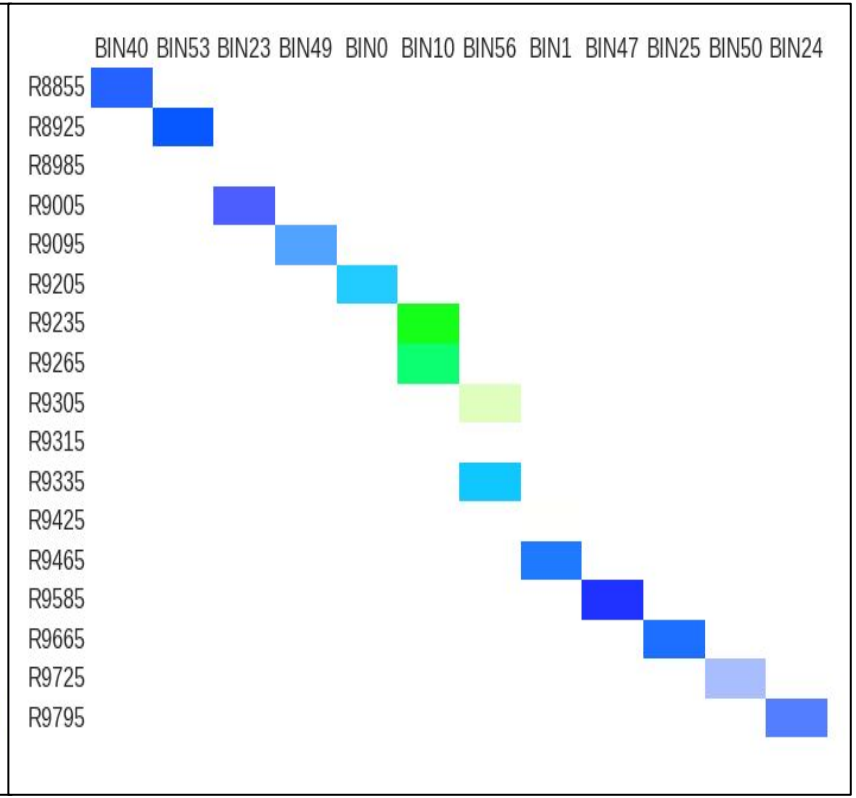
- winsorized mean + Metabat 2
- poisson iter weighted mean + CONCOCT

Спасибо за внимание

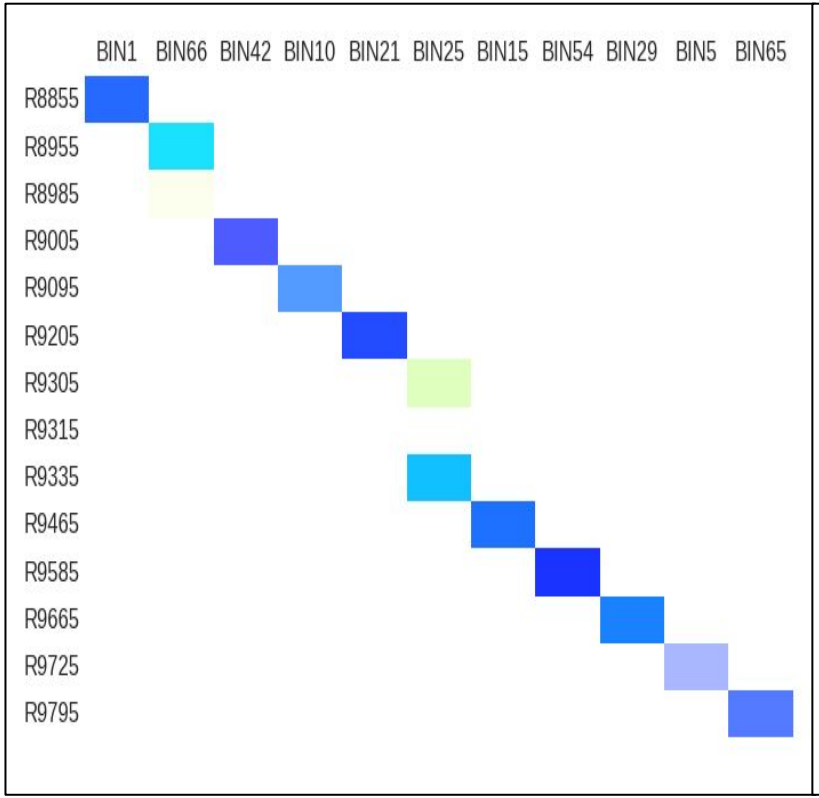
iter poisson



norm



iter poisson



student