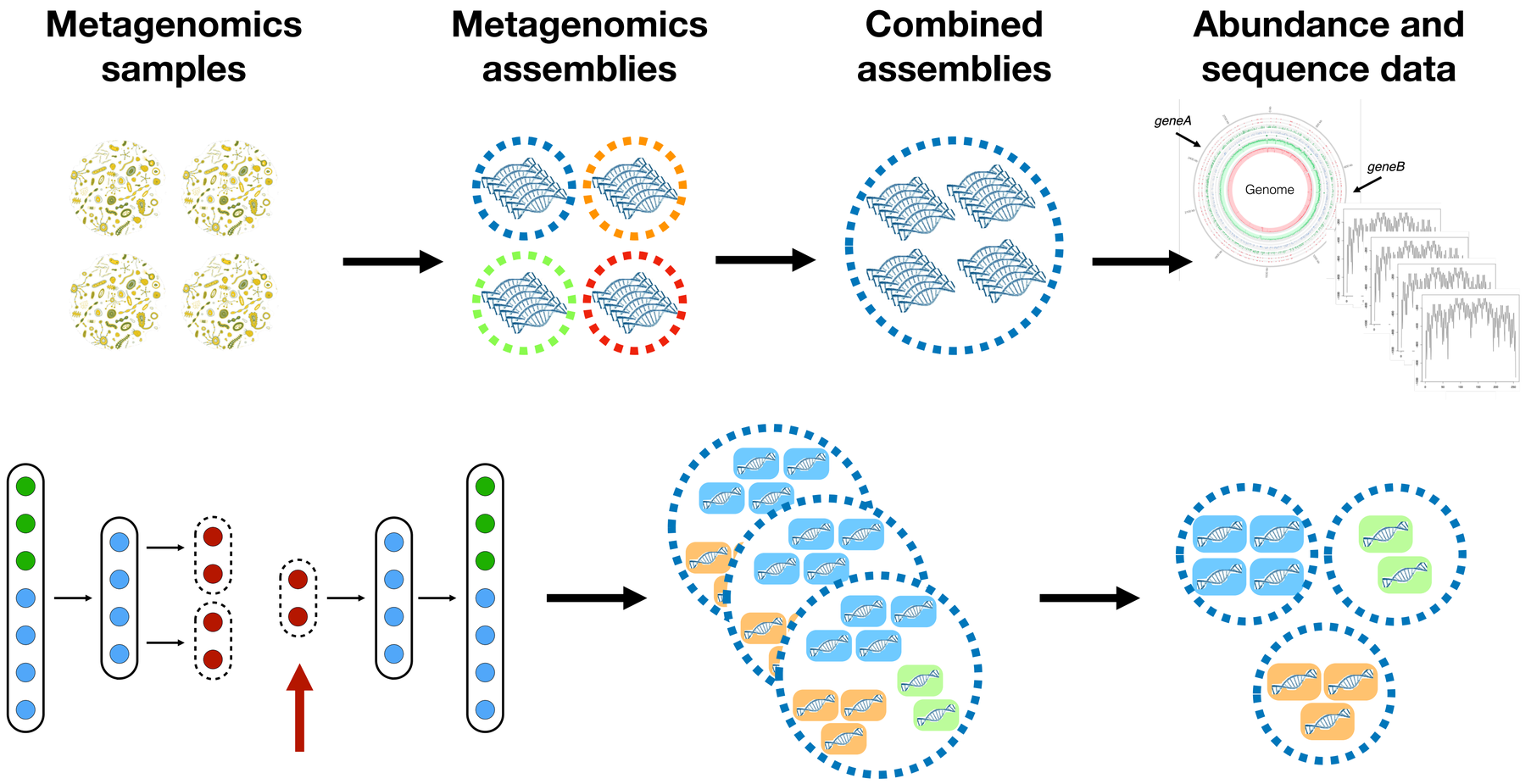
VAMB, dRep, and CheckM

Binning and quality control

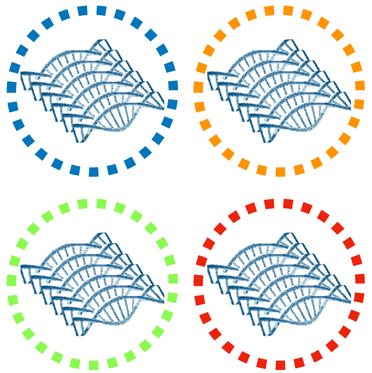
Marie Louise Jespersen, 2021.10.26

VAMB


Metagenomics samples
Metagenomics assemblies
Combined assemblies
Abundance and sequence data
geneA
geneB
Genome
Integrate in Variational Autoencoder
Adaptive clustering to genome clusters
Sample-wise genome bins

Combined assemblies

Metagenomics assembly



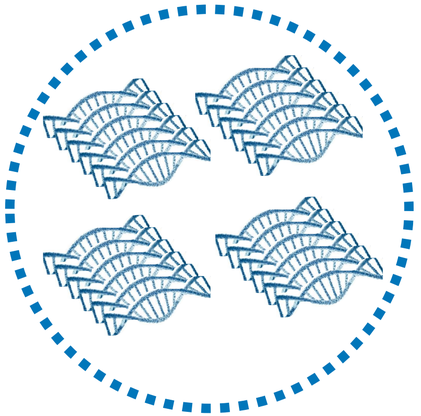
Millions of small pieces of 100bp - 10kbp

"Traditional" approach

Our approach

Redundancy reduced

Easy scalable
Treshold
Chimeric genomes

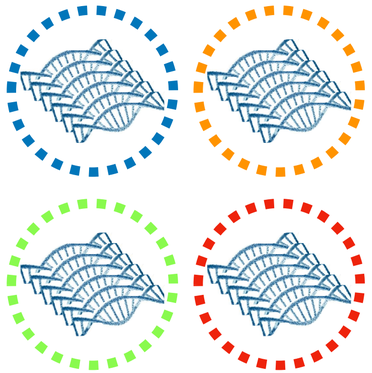


Scalability?
~~*Treshold*~~
~~*Chimeric genomes*~~

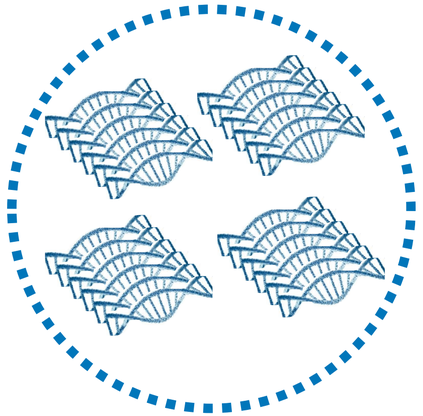
Combined assembly

Global Sewage

Metagenomics assembly



30,171,703 contigs > 2000 bp



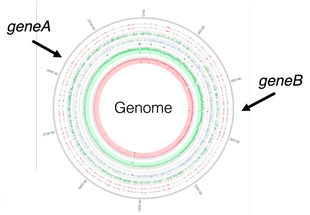
Combined assembly

The data

Combined assembly

Abundance data

Normalize to sum to 1 (prob. distr)

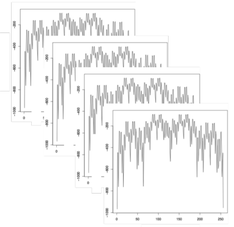


	Sample 1	Sample 2	Sample …	Sample 124
Contig 1	0	0,4	0,6	0
Contig 2	0,45	0	0,05	0,50
Contig 3	0	0,33	0,67	0
Contig 4	0	0	0	1,00
…	…	…	…	…
Contig N	X	Y	Z	x

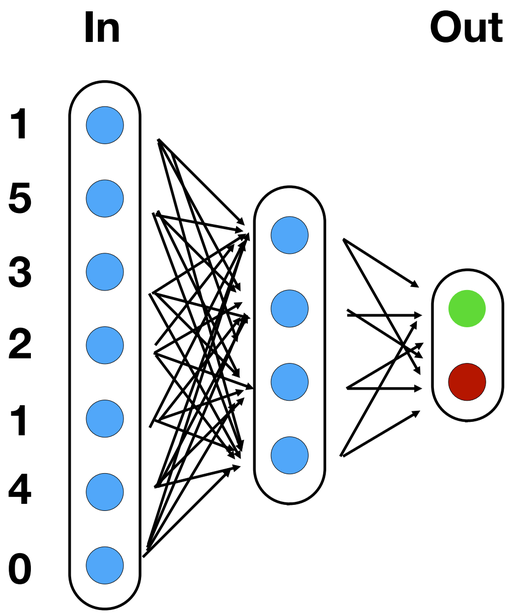
	AAAA	AAAC	…	TTTT
Contig 1	0,08	0,03	0,01	0
Contig 2	0,03	0,02	0,05	0,01
Contig 3	0,08	0,03	0,01	0
Contig 4	0,02	0,03	0,04	0,07
…	…	…	…	…
Contig N	X	Y	Z	x

Sequence data (TNF)

Z-scale normalised per TNF

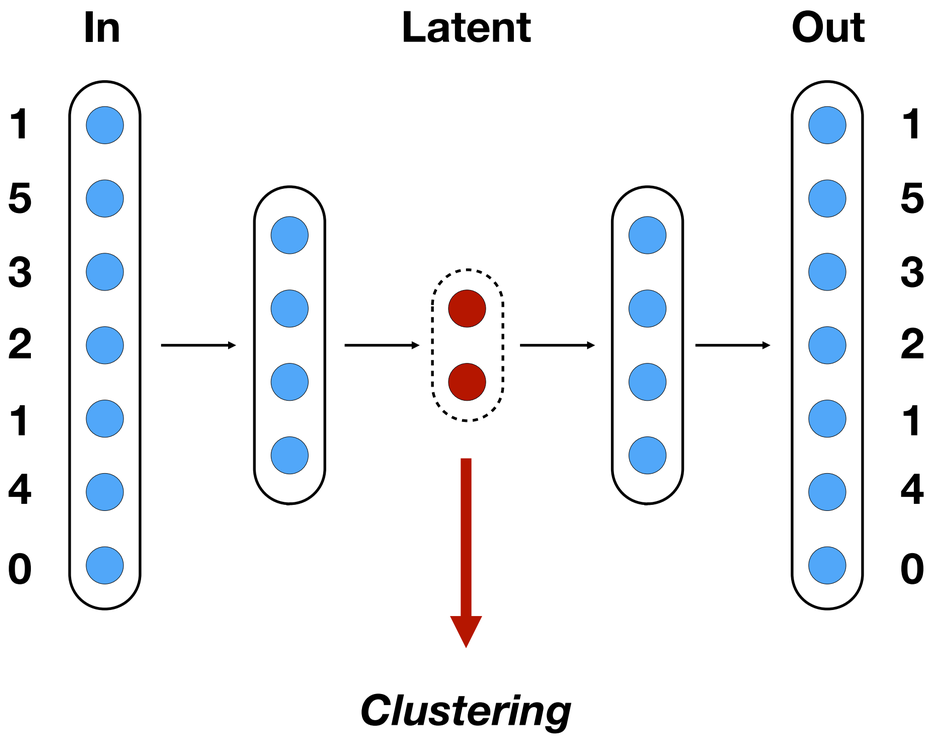


Autoencoders



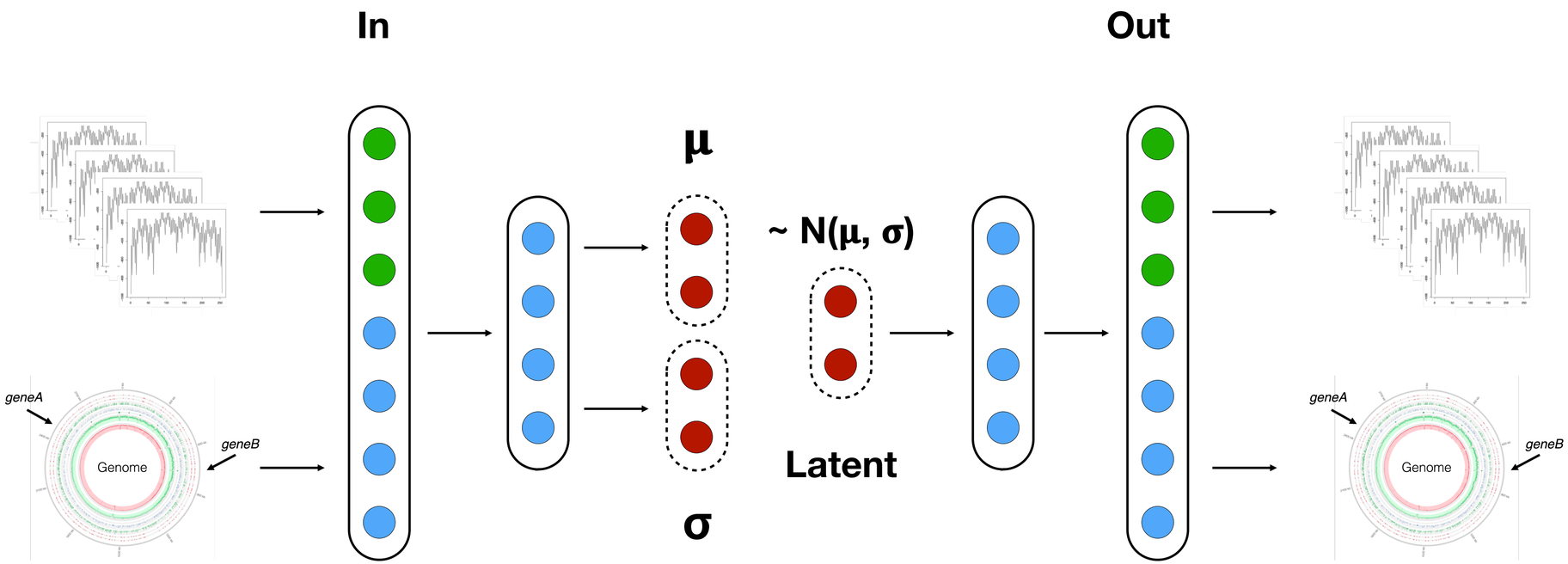
- Type of neural network

Autoencoders

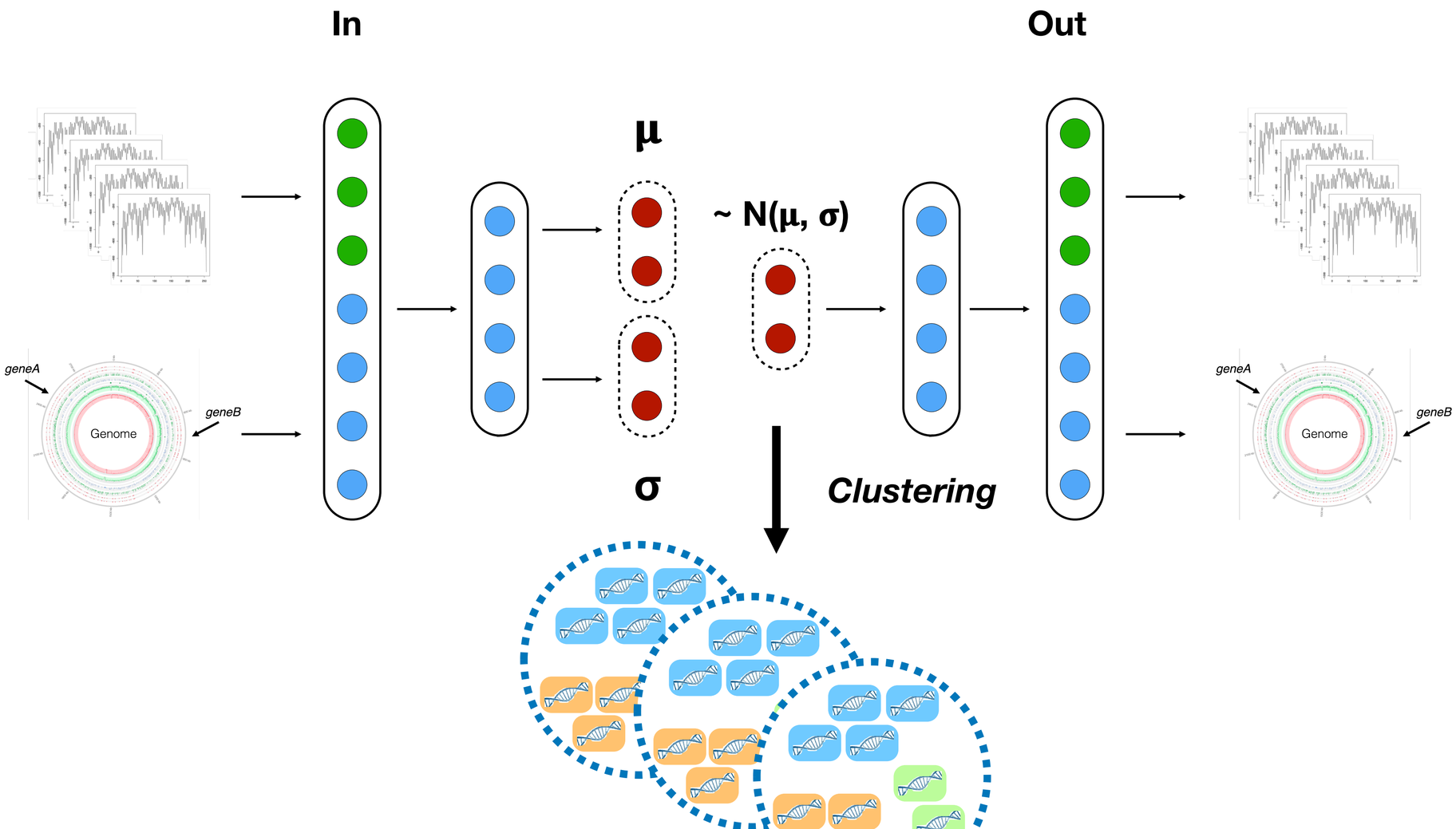


- Type of neural network
- Unsupervised learning
- Minimise difference between input and output (training)
- Force the network to learn a **latent representation**

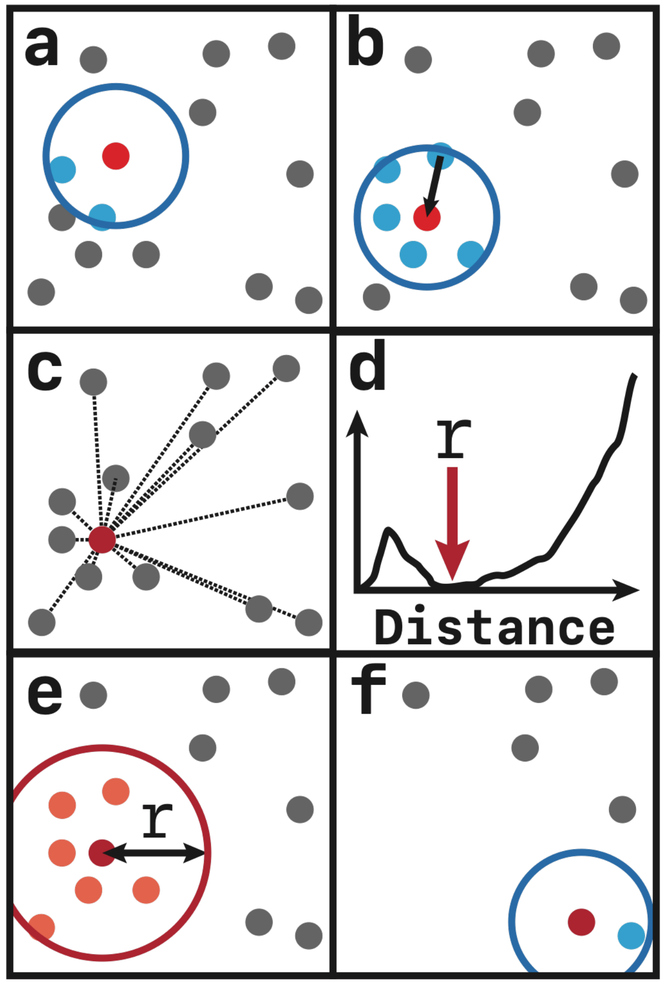
Integration and clustering



Integration and clustering

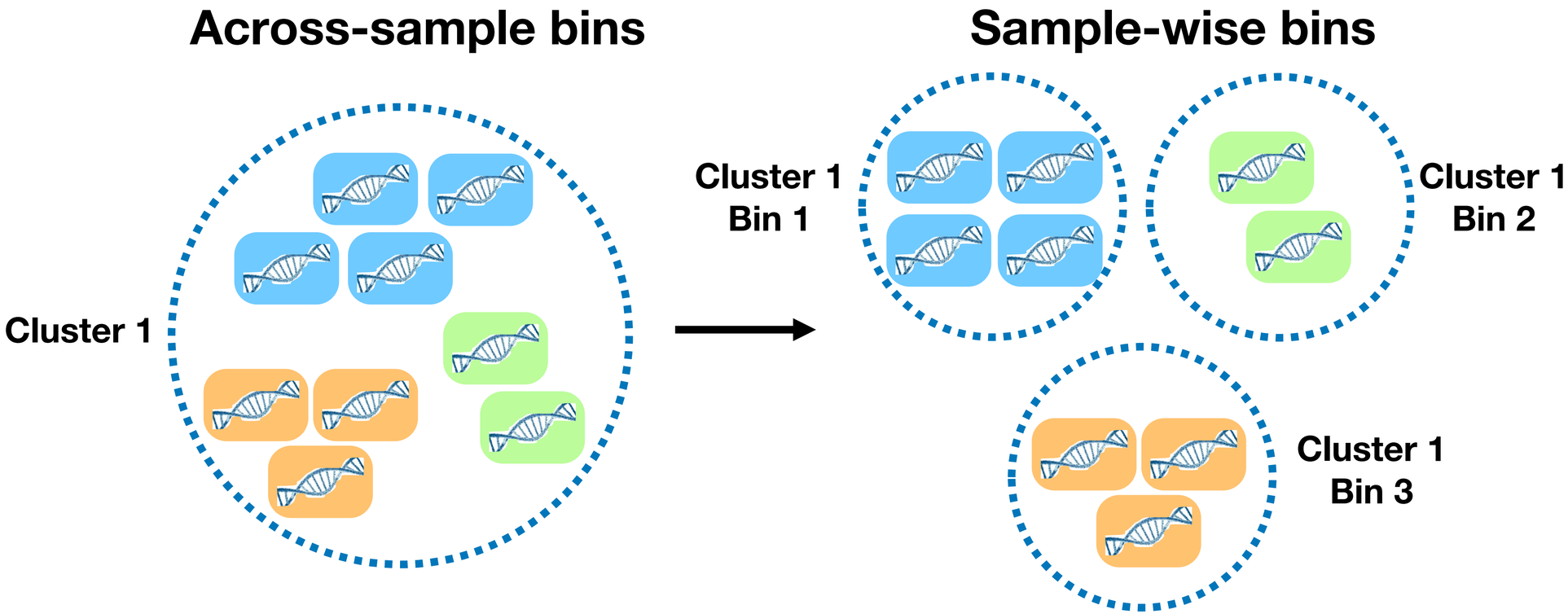


Clustering



- Needs to be **fast** (millions of contigs)
- Sampling based strategy using cosine distance
- Adaptive cluster threshold detection
- GPU Implementation

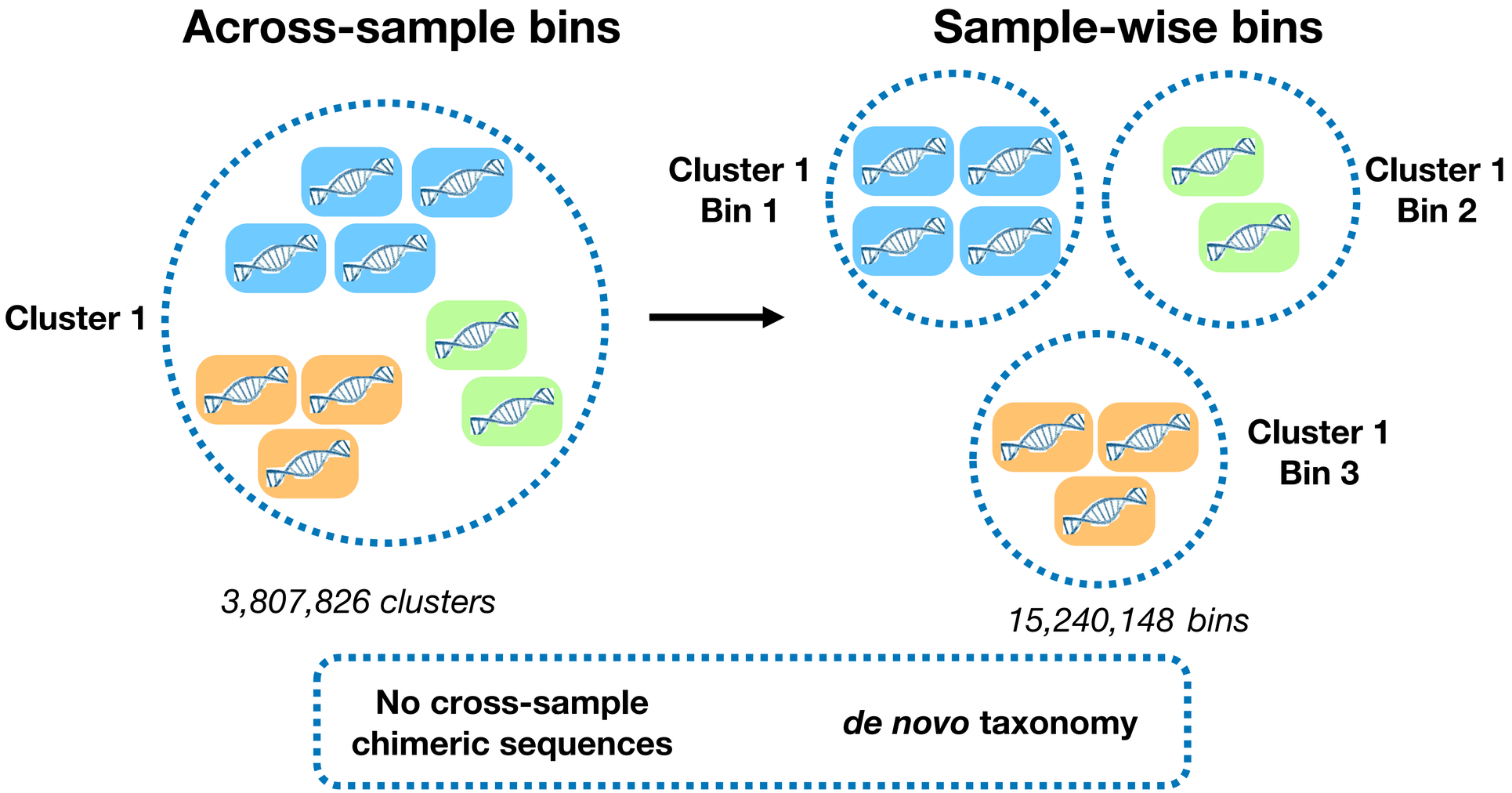
Cluster splitting



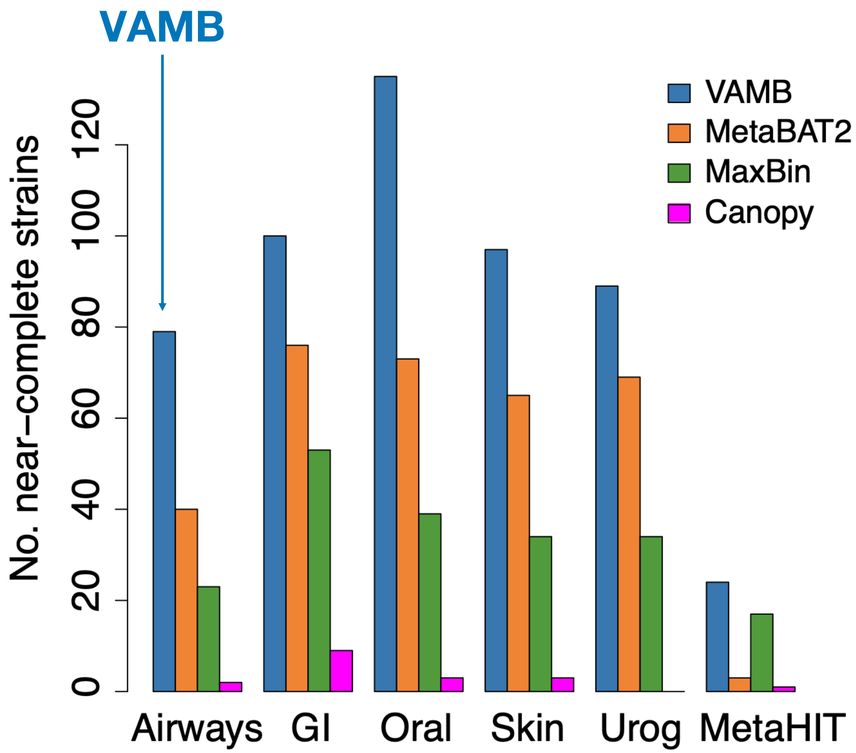
No cross-sample chimeric sequences

de novo taxonomy

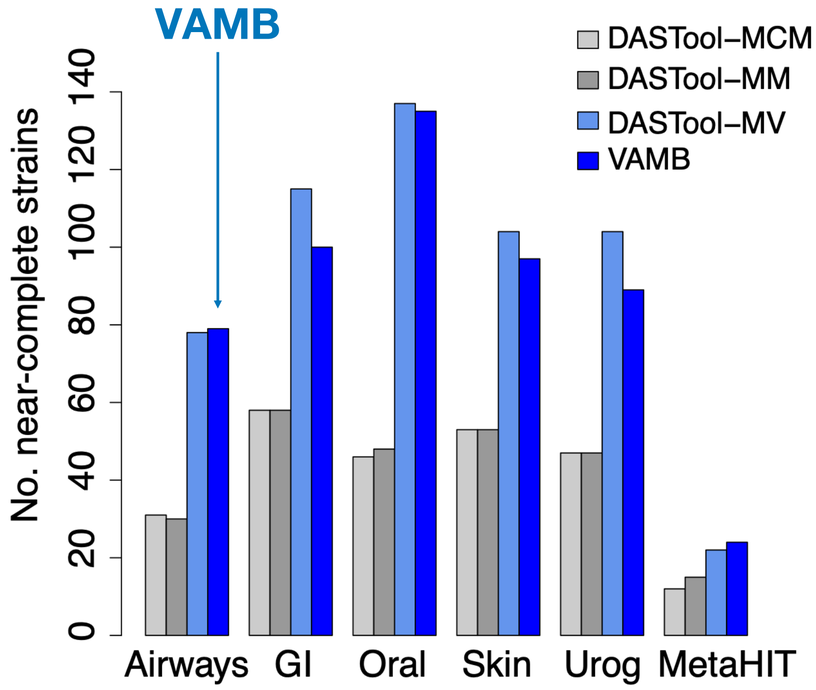
Global Sewage



Benchmark results

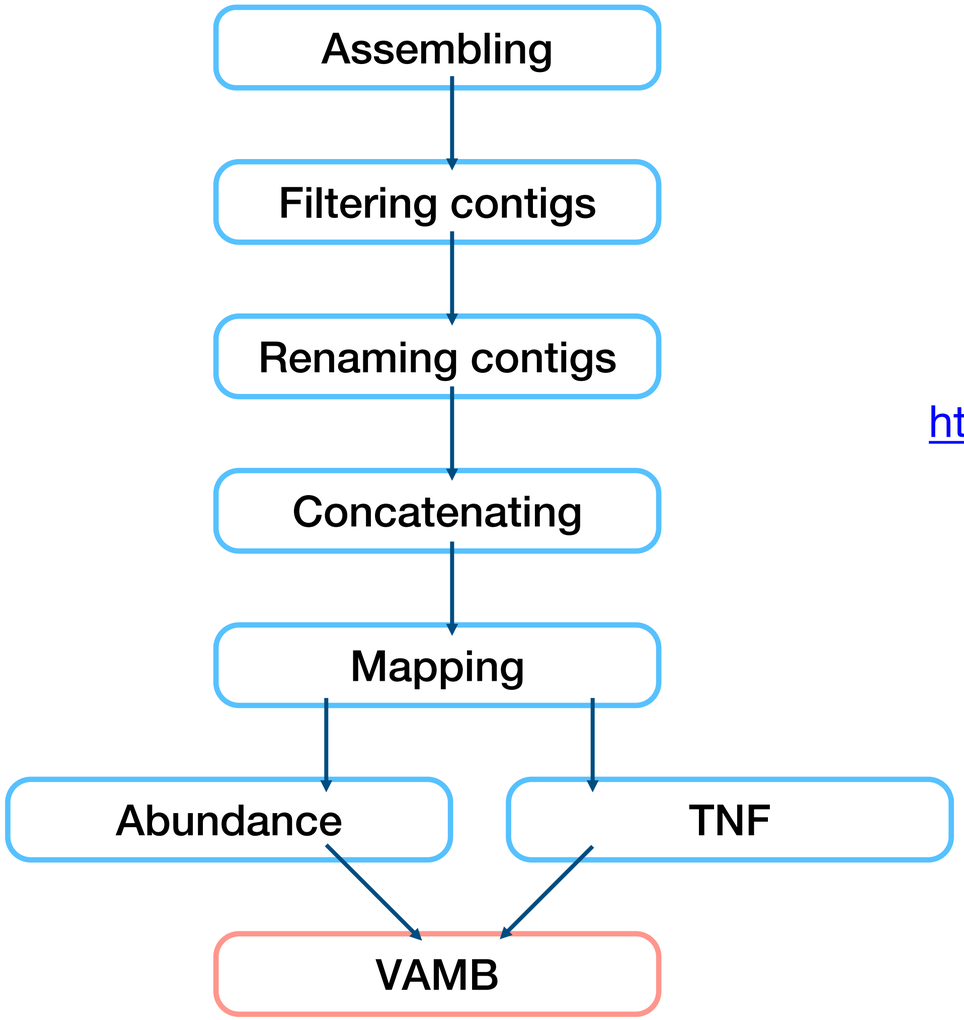


29-98% improvement
compared to state-of-the art



Even better than all other
binners combined

VAMB - input



Workflow at:
https://github.com/RasmussenLab/vamb/tree/master/workflow

VAMB - output

Sample name

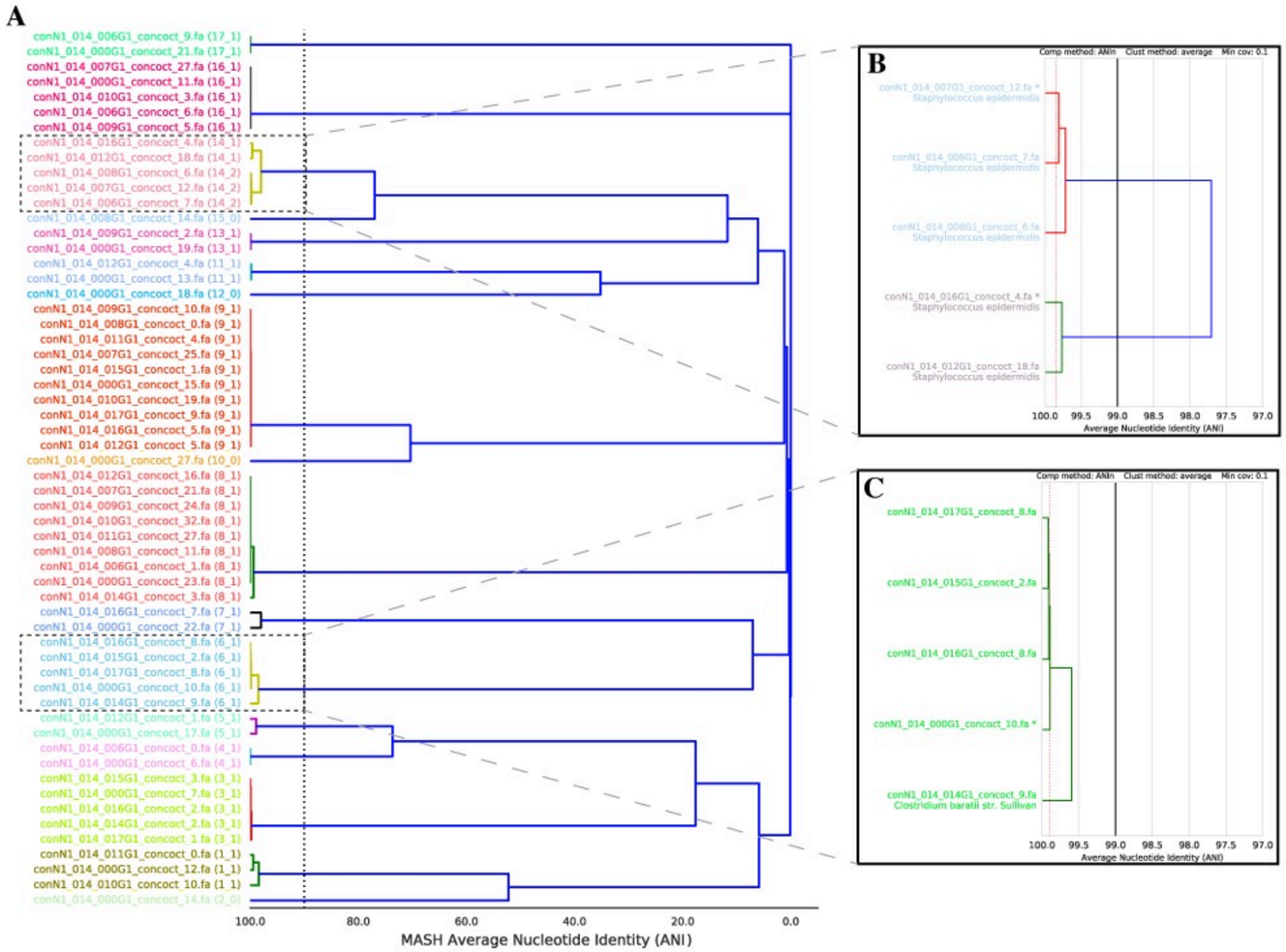
Cluster number

Contig name

DTU_2017_361_1_MG_FI_HE	1
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_17569_length_2692
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_14250_length_2955
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_9854_length_3501
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_33310_length_2024
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_31627_length_2072
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_19631_length_2562
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_32257_length_2053
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_8705_length_3696
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_25419_length_2287
DTU_2017_361_1_MG_FI_HE_NODE_1	DTU_2017_361_1_MG_FI_HE_NODE_33970_length_2006
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_18396_length_2557
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_24903_length_2223
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_13640_length_2936
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_25737_length_2190
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_19313_length_2501
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_18950_length_2524
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_31066_length_2006
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_26764_length_2150
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_28057_length_2104
DTU_2017_604_1_MG_CA_CA_300_NODE_1	DTU_2017_604_1_MG_CA_CA_300_NODE_25192_length_2212

dRep

dRep compare



- Pair-wise comparison of genome sets
- MASH all genomes
- ANI on genomes > 90% MASH ANI

Figure from: https://drep.readthedocs.io/en/latest/overview.html

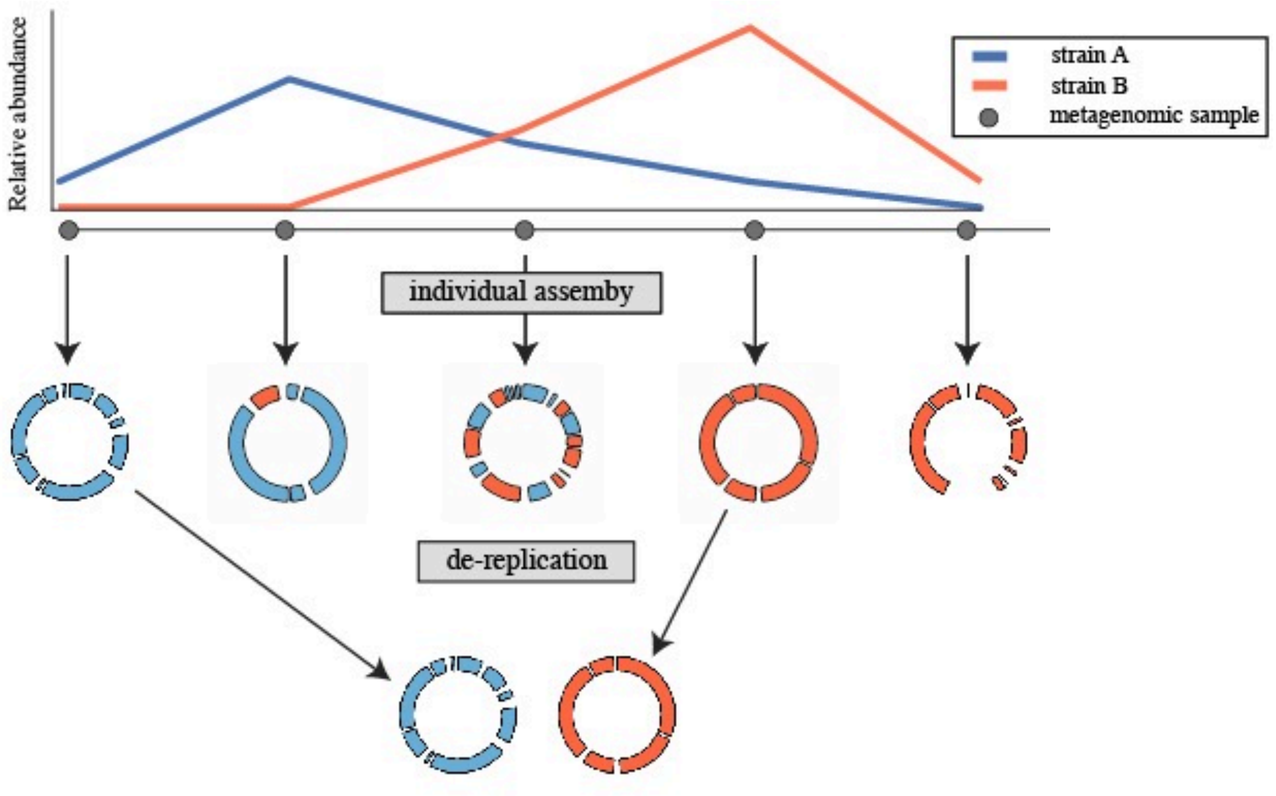
dRep - input

```
#!/bin/sh
### Note: No commands may be executed until after the #PBS lines
### Account information
#PBS -W group_list=cge -A cge
### Job name (comment out the next line to get the name of the script used as the job name)
#PBS -N dRep
### Output files (comment out the next 2 lines to get the job name used instead)
#PBS -e dRep.err
#PBS -o dRep.log
### Only send mail when job is aborted or terminates abnormally
#PBS -m n
### Number of nodes
#PBS -l nodes=1:ppn=40
### Memory
#PBS -l mem=180gb
### Requesting time - format is <days>:<hours>:<minutes>:<seconds> (here, 12 hours)
#PBS -l walltime=10:00:00:00

# Go to the directory from where the job was submitted (initial directory is $HOME)
echo Working directory is $PBS_O_WORKDIR
cd $PBS_O_WORKDIR

#load necessary modules
module unload R gcc
module load anaconda3/2.2.0 tools ngs mash/2.2 mummer/3.23

#run dRep on all metabat and vamb bins
dRep compare dRep_result -p 40 -pa 0.9 -sa 0.95 -g nc_genomes/*.fna

```

dRep - output

```
(base) [maloj@g-12-l0002 Binning_vamb]$ cd dRep_result/
(base) [maloj@g-12-l0002 dRep_result]$ ll
total 160
drwxrwxr-x 5 maloj cge  91 Sep 16  2020 data
drwxrwxr-x 3 maloj cge 121 Oct 18 11:47 data_tables
drwxrwxr-x 2 maloj cge   0 Sep 16  2020 dereplicated_genomes
drwxrwxr-x 2 maloj cge 196 Sep 16  2020 figures
drwxrwxr-x 2 maloj cge  98 Sep 16  2020 log
(base) [maloj@g-12-l0002 dRep_result]$ cd data_tables/
(base) [maloj@g-12-l0002 data_tables]$ ll
total 3309752
-rw-rw-r-- 1 maloj cge     815524 Sep 16  2020 Bdb.csv
-rw-rw-r-- 1 maloj cge     481793 Sep 16  2020 Cdb.csv
-rw-rw-r-- 1 maloj cge 2787455810 Sep 16  2020 Mdb.csv
-rw-rw-r-- 1 maloj cge   23171334 Sep 16  2020 Ndb.csv
drwxrwxr-x 3 maloj cge        183 Oct 18 11:47 tmp
(base) [maloj@g-12-l0002 data_tables]$ head Cdb.csv
genome,secondary_cluster,threshold,cluster_method,comparison_algorithm,primary_cluster
DTU_2017_417_3_MG_NG_ZA_c.metabin.63.fna,1_1,0.05000000000000044,average,ANImf,1
DTU_2017_417_3_MG_NG_ZA_c.vambbin.27102.fna,1_1,0.05000000000000044,average,ANImf,1
DTU_2017_539_1_MG_HU_BU_235.metabin.109.fna,2_1,0.05000000000000044,average,ANImf,2
DTU_2017_539_1_MG_HU_BU_235.vambbin.58560.fna,2_1,0.05000000000000044,average,ANImf,2
DTU_2017_445_1_MG_GB_NC.metabin.74.fna,3_1,0.05000000000000044,average,ANImf,3
DTU_2017_609_1_MG_FI_TU_305.metabin.78.fna,3_1,0.05000000000000044,average,ANImf,3
DTU_2017_609_1_MG_FI_TU_305.vambbin.17301.fna,3_1,0.05000000000000044,average,ANImf,3
DTU_2017_457_1_MG_KH_PP.metabin.42.fna,4_1,0.05000000000000044,average,ANImf,4
DTU_2017_495_1_MG_KH_PP_191.metabin.8.fna,4_1,0.05000000000000044,average,ANImf,4
(base) [maloj@g-12-l0002 data_tables]$
```

dRep de-replication



- Compares genomes to avoid redundancy and identify best representative

Figure from: https://drep.readthedocs.io/en/latest/overview.html

dRep de-replication

$$A*Completeness-B*Contamination+C*(Contamination*(strainheterogeneity/100))+D*log(N50)+E*log(size)+F*(centrality-Sani)$$

Where A-F are command-line arguments with default values of 1, 5, 1, 0.5, 0, and 1, respectively.

Completeness, Contamination, and strain heterogeneity are provided by the user or calculated with checkM.

N50 is a measure of how big the pieces are that make up the genome.

size is the total length of the genome.

Centrality is a measure of how similar a genome is to all other genomes in it’s cluster.

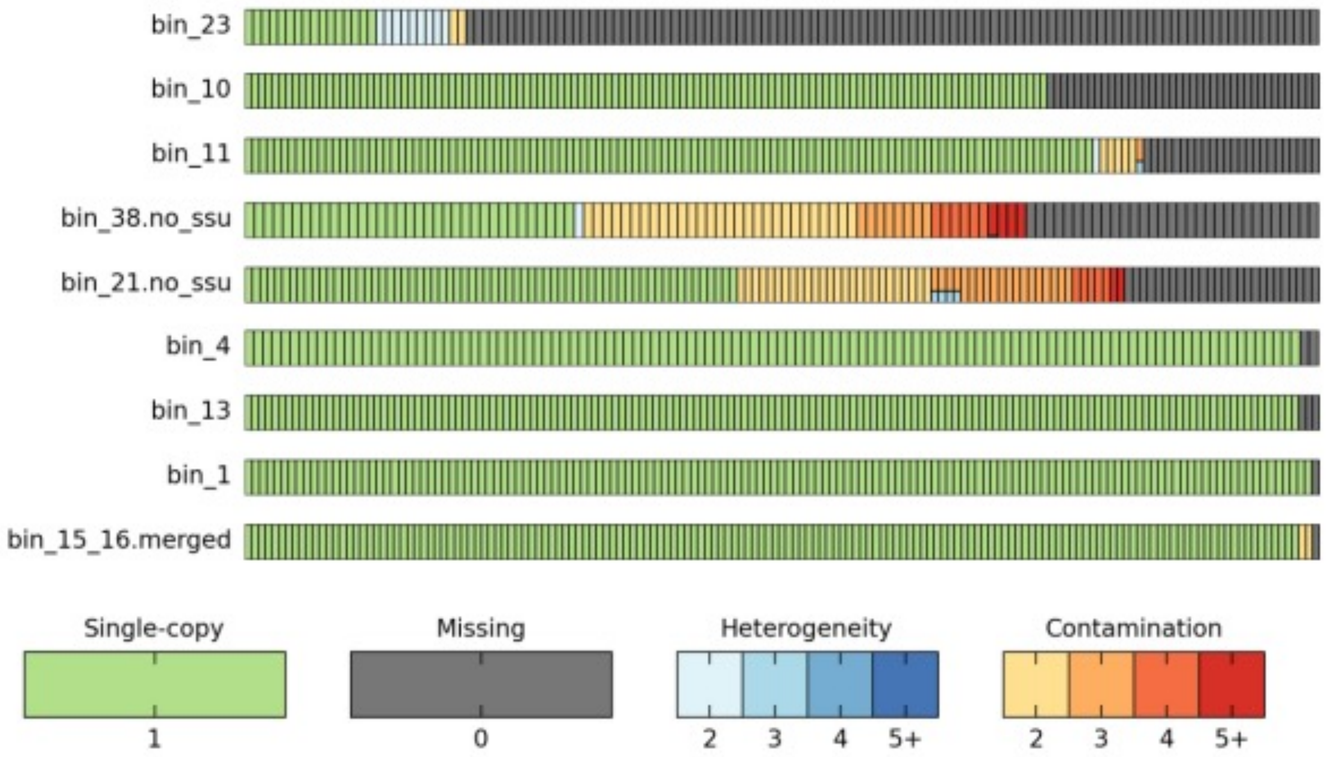
Text from: https://drep.readthedocs.io/en/latest/overview.html

dRep output

```
(base) [maloj@g-12-l0002 999_1]$ cd data_tables/
(base) [maloj@g-12-l0002 data_tables]$ ll
total 824
-rw-r--r-- 1 maloj cge  1419 Jan 20  2021 Bdb.csv
-rw-r--r-- 1 maloj cge   769 Jan 20  2021 Cdb.csv
-rw-r--r-- 1 maloj cge  1854 Jan 20  2021 Chdb.csv
-rw-r--r-- 1 maloj cge   664 Jan 20  2021 genomeInfo.csv
-rw-r--r-- 1 maloj cge   664 Jan 20  2021 genomeInformation.csv
-rw-r--r-- 1 maloj cge  7034 Jan 20  2021 Mdb.csv
-rw-r--r-- 1 maloj cge 10890 Jan 20  2021 Ndb.csv
-rw-r--r-- 1 maloj cge   506 Jan 20  2021 Sdb.csv
-rw-rw-r-- 1 maloj cge   346 Jan 21  2021 Sdb.nonredundant.csv
-rw-r--r-- 1 maloj cge    87 Jan 20  2021 Wdb.csv
-rw-r--r-- 1 maloj cge   331 Jan 20  2021 Widb.csv
(base) [maloj@g-12-l0002 data_tables]$ head genomeInfo.csv
genome,completeness,contamination,strain_heterogeneity,length,N50
DTU_2017_539_1_MG_HU_BU_235_metabin_52.fna,90.86,3.15,55.56,2608720,15968
DTU_2017_539_1_MG_HU_BU_235_vambbin_2922.fna,91.94,2.62,57.14,2727821,15863
DTU_2018_671_1_MG_HU_BU_359_metabin_15.fna,93.28,0.54,0.0,2635428,56192
DTU_2018_671_1_MG_HU_BU_359_vambbin_2922.fna,94.35,0.54,0.0,2758953,48885
DTU_2018_811_1_MG_HU_BU_495_metabin_66.fna,93.01,2.93,18.75,2917234,23851
DTU_2018_811_1_MG_HU_BU_495_vambbin_2922.fna,93.01,1.75,75.0,2758723,24618
DTU_2018_922_1_MG_BJ_TO_606_metabin_5.fna,95.43,1.61,33.33,2753212,60124
DTU_2018_923_2_MG_BJ_TO_606_re_1_metabin_75.fna,94.89,1.61,33.33,2733445,103723
(base) [maloj@g-12-l0002 data_tables]$ head Sdb.csv
genome,score
DTU_2017_539_1_MG_HU_BU_235_metabin_52.fna,78.96176526197165
DTU_2017_539_1_MG_HU_BU_235_vambbin_2922.fna,82.4372606621162
DTU_2018_671_1_MG_HU_BU_359_metabin_15.fna,92.954837244945
DTU_2018_671_1_MG_HU_BU_359_vambbin_2922.fna,93.9945878097604
DTU_2018_811_1_MG_HU_BU_495_metabin_66.fna,81.09812829620331
DTU_2018_811_1_MG_HU_BU_495_vambbin_2922.fna,87.76812638367528
DTU_2018_922_1_MG_BJ_TO_606_metabin_5.fna,90.3061369330644
DTU_2018_923_2_MG_BJ_TO_606_re_1_metabin_75.fna,89.88455053473032
(base) [maloj@g-12-l0002 data_tables]$ ll ../dereplicated_genomes/
total 3560
-rw-rw-r-- 1 maloj cge 2810855 Sep  3  2020 DTU_2018_671_1_MG_HU_BU_359_vambbin_2922.fna
(base) [maloj@g-12-l0002 data_tables]$
```

CheckM

CheckM



- Quality of genomes
  - Heterogeneity: AAI≥90%
  - Contamination: AAI<90%

Figure from: http://ecogenomics.github.io/CheckM/manual/checkm_manual.pdf

CheckM

Criterion	Description
Finished (SAG/MAG)	
Assembly quality[a]	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion[b]	>90%
Contamination[c]	<5%
Medium-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	≥50%
Contamination[c]	<10%
Low-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	<50%
Contamination[c]	<10%
This is a compressed set of genome reporting standards for SAGs and MAGs. For a complete list of mandatory and optional standards, see Supplementary Table 1.	

[a]Assembly statistics include but are not limited to: N50, L50, largest contig, number of contigs, assembly size, percentage of reads that map back to the assembly, and number of predicted genes per genome.

[b]Completion: ratio of observed single-copy marker genes to total single-copy marker genes in chosen marker gene set.

[c]Contamination: ratio of observed single-copy marker genes in ≥2 copies to total single-copy marker genes in chosen marker gene set.

- MIMAG:*
  - assembly quality
  - genome completeness
  - Genome contamination

* Minimum Information about a Metagenome-Assembled Genome

Table from: Bowers, R., Kyrpides, N., Stepanauskas, R. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35,** 725–731 (2017).

CheckM - input

```
#!/bin/sh
### Note: No commands may be executed until after the #PBS lines
### Account information
#PBS-W group_list=co_23260 -A co_23260
### Job name (comment out the next line to get the name of the script used as the job name)
#PBS -N checkM
### Output files (comment out the next 2 lines to get the job name used instead)
#PBS -e checkM.err
#PBS -o checkM.log
### Only send mail when job is aborted or terminates abnormally
#PBS -m n
### Number of nodes
#PBS -l nodes=1:ppn=20
### Memory
#PBS -l mem=90gb
### Requesting time - format is <days>:<hours>:<minutes>:<seconds> (here, 10 days)
#PBS -l walltime=10:00:00:00
### The node reserved for course 23260
#PBS advres=co_23260.1633

# Go to the directory from where the job was submitted (initial directory is $HOME)
echo Working directory is $PBS_O_WORKDIR
cd $PBS_O_WORKDIR

# Load all required modules for the job
module load anaconda2/4.0.0 prodigal/2.6.2 hmmer/3.1b2 pplacer/1.1.alpha17

# set input and output
GROUP_DIR="/home/projects/co_23260/data/groups/group_test" # change to own group directory
IN_DIR="$GROUP_DIR/metabat2/" # check that you have the same drectory structure
OUT_DIR="$GROUP_DIR/checkm/"

# Run whole checkm workflow
checkm lineage_wf -f $OUT_DIR/checkm.txt -t 20 --tab_table -x fa $IN_DIR $OUT_DIR
```

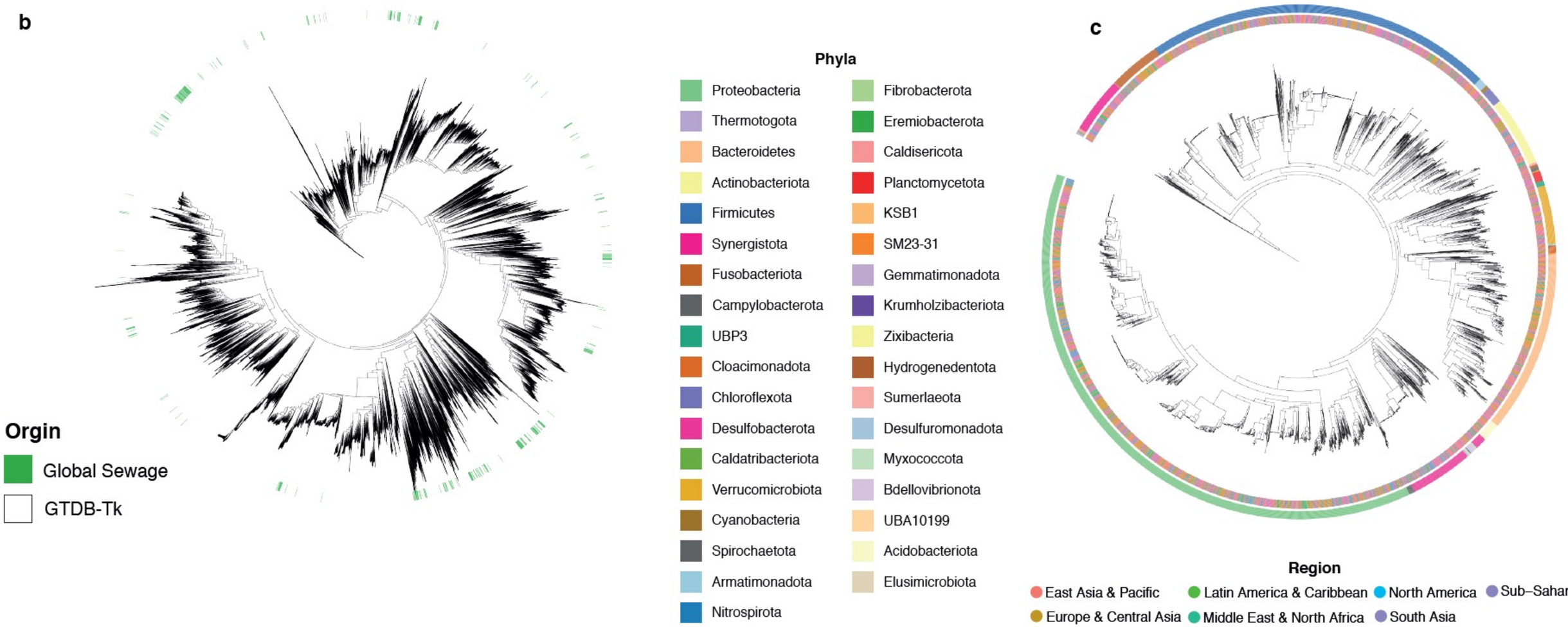
CheckM - output

Bin Id	Marker lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination	Strain heterogeneity
DTU2019_MG_483_1084_NODE_1008	k__Bacteria (UID203)	5449	103	57	0	82	21	0	0	0	100.00	12.28	95.24
DTU2019_MG_479_1293_NODE_175	f__Lachnospiraceae (UID1286)	57	420	207	0	402	14	4	0	0	100.00	7.25	53.85
DTU2019_MG_477_323_NODE_99	root (UID1)	5656	56	24	0	0	55	1	0	0	100.00	104.17	0.00
DTU2019_MG_477_323_NODE_197	p__Euryarchaeota (UID3)	148	188	125	0	188	0	0	0	0	100.00	0.00	0.00
DTU2019_MG_476_322_NODE_708	p__Bacteroidetes (UID2605)	350	314	208	0	314	0	0	0	0	100.00	0.00	0.00
DTU2019_MG_477_323_NODE_1409	o__Selenomonadales (UID1024)	64	334	167	1	328	5	0	0	0	99.98	1.70	0.00
DTU2019_MG_481_1306_NODE_847	o__Bacteroidales (UID2621)	198	427	260	1	425	0	1	0	0	99.62	0.77	0.00
DTU2019_MG_476_322_NODE_782	p__Bacteroidetes (UID2605)	350	314	208	1	311	2	0	0	0	99.52	0.96	0.00
DTU2019_MG_476_322_NODE_724	o__Bacteroidales (UID2621)	198	427	260	2	422	3	0	0	0	99.42	0.64	33.33
DTU2019_MG_483_1084_NODE_176	f__Bifidobacteriaceae (UID1458)	77	464	220	2	456	5	1	0	0	99.39	2.73	50.00
DTU2019_MG_482_1048_NODE_39	c__Gammaproteobacteria (UID4201)	1164	271	170	2	268	1	0	0	0	99.38	0.29	0.00
DTU2019_MG_478_329_NODE_39	c__Gammaproteobacteria (UID4201)	1164	271	170	2	267	2	0	0	0	99.38	0.88	50.00
DTU2019_MG_477_323_NODE_39	c__Gammaproteobacteria (UID4201)	1164	271	170	2	268	1	0	0	0	99.38	0.29	0.00
DTU2019_MG_480_1305_NODE_501	f__Bifidobacteriaceae (UID1462)	65	476	217	4	469	3	0	0	0	99.35	0.71	66.67
DTU2019_MG_481_1306_NODE_1393	o__Clostridiales (UID1212)	172	263	149	1	258	4	0	0	0	99.33	1.68	75.00
DTU2019_MG_479_1293_NODE_49	o__Clostridiales (UID1212)	172	263	149	1	262	0	0	0	0	99.33	0.00	0.00
DTU2019_MG_479_1293_NODE_1068	o__Clostridiales (UID1212)	172	263	149	1	262	0	0	0	0	99.33	0.00	0.00
DTU2019_MG_477_323_NODE_811	o__Clostridiales (UID1212)	172	263	149	1	261	1	0	0	0	99.33	0.67	0.00
DTU2019_MG_477_323_NODE_6039	o__Clostridiales (UID1212)	172	263	149	1	258	4	0	0	0	99.33	2.35	100.00
DTU2019_MG_477_323_NODE_142	o__Clostridiales (UID1342)	32	340	177	2	338	0	0	0	0	99.32	0.00	0.00
DTU2019_MG_481_1306_NODE_577	o__Clostridiales (UID1120)	304	250	143	1	241	8	0	0	0	99.30	1.86	12.50
DTU2019_MG_476_322_NODE_555	o__Bacteroidales (UID2657)	160	492	269	6	486	0	0	0	0	99.21	0.00	0.00
DTU2019_MG_481_1306_NODE_5325	c__Clostridia (UID1118)	387	223	124	2	220	1	0	0	0	99.19	0.81	0.00
DTU2019_MG_477_323_NODE_5325	c__Clostridia (UID1118)	387	223	124	2	220	1	0	0	0	99.19	0.81	0.00
DTU2019_MG_477_323_NODE_130	c__Clostridia (UID1118)	387	223	124	1	221	1	0	0	0	99.19	0.27	0.00
DTU2019_MG_477_323_NODE_555	o__Bacteroidales (UID2657)	160	492	269	9	482	1	0	0	0	99.14	0.12	0.00
DTU2019_MG_480_1305_NODE_712	f__Lachnospiraceae (UID1286)	57	420	207	14	401	5	0	0	0	99.09	1.14	40.00
DTU2019_MG_476_322_NODE_97	p__Bacteroidetes (UID2605)	350	314	208	2	311	1	0	0	0	99.04	0.48	100.00
DTU2019_MG_477_323_NODE_124	o__Bacteroidales (UID2654)	163	486	266	16	464	6	0	0	0	99.01	0.98	66.67
DTU2019_MG_481_1306_NODE_28151	o__Clostridiales (UID1212)	172	263	149	3	260	0	0	0	0	98.99	0.00	0.00

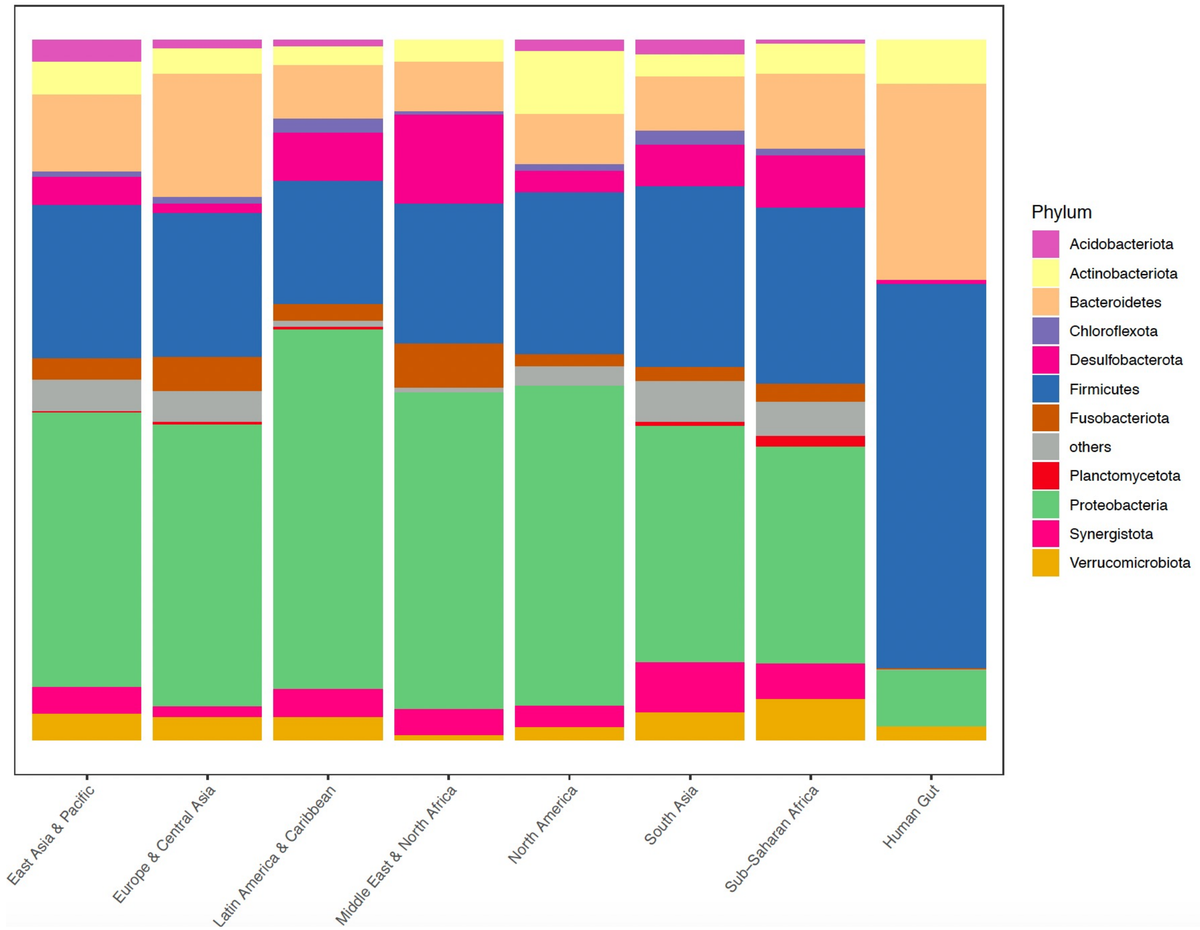
Global Sewage - results

Bins/MAGs/strains		Clusters/Species
15,240,148	30 mio. contigs 757 samples ↓ VAMB	*3,807,826*
20,785	↓ Filtering	-
3,353	↓ CheckM	-
-	↓ dRep	*894*

Global Sewage - results



Global Sewage - results



Questions?