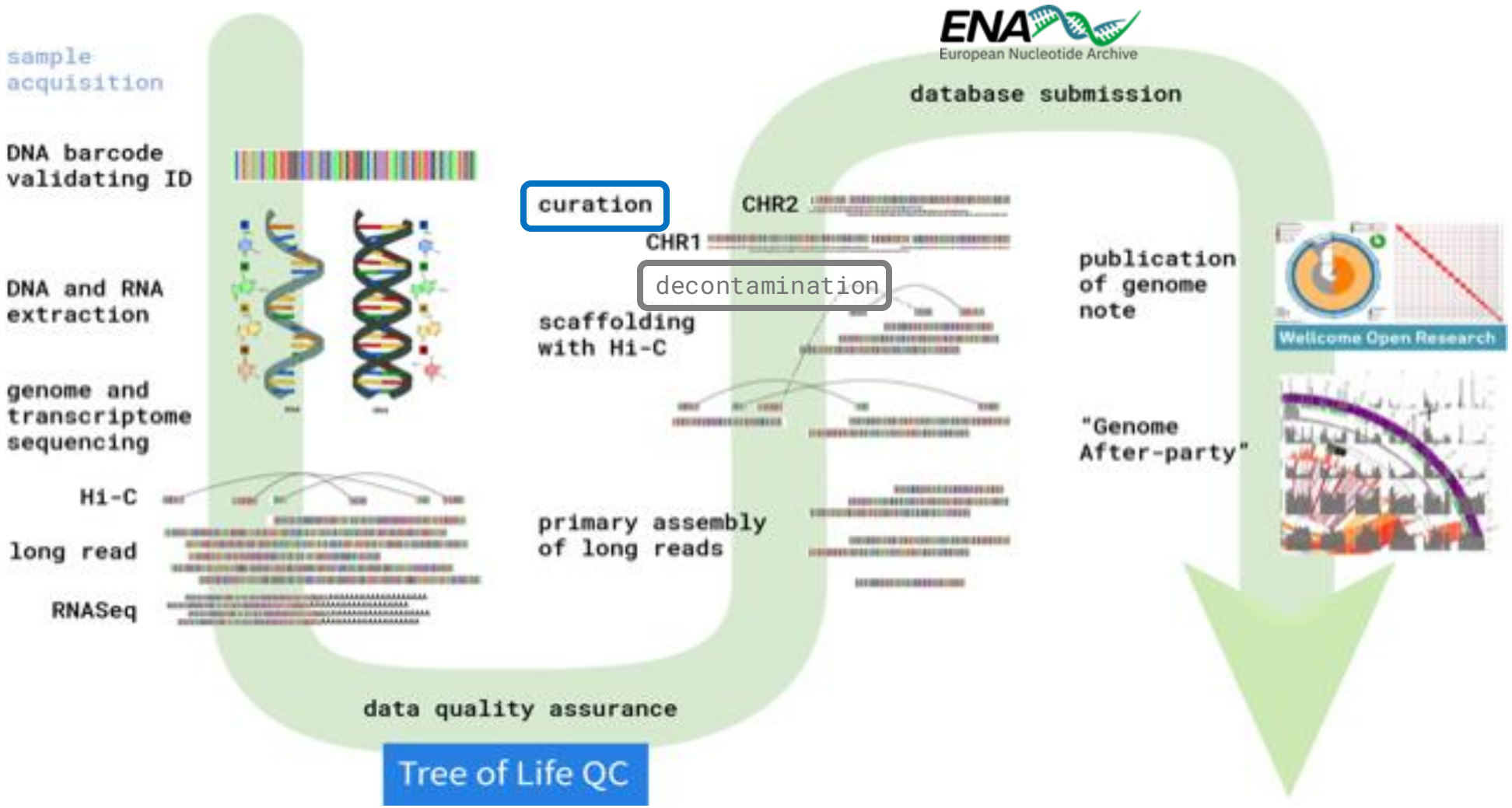
Session 2.2: Decontaminate your assembly before curation

Genome Reference Informatics Team (GRIT)
Wellcome Sanger Institute - Tree of Life

The Tree of Life genome factory



Why do we need to decontaminate our assemblies?

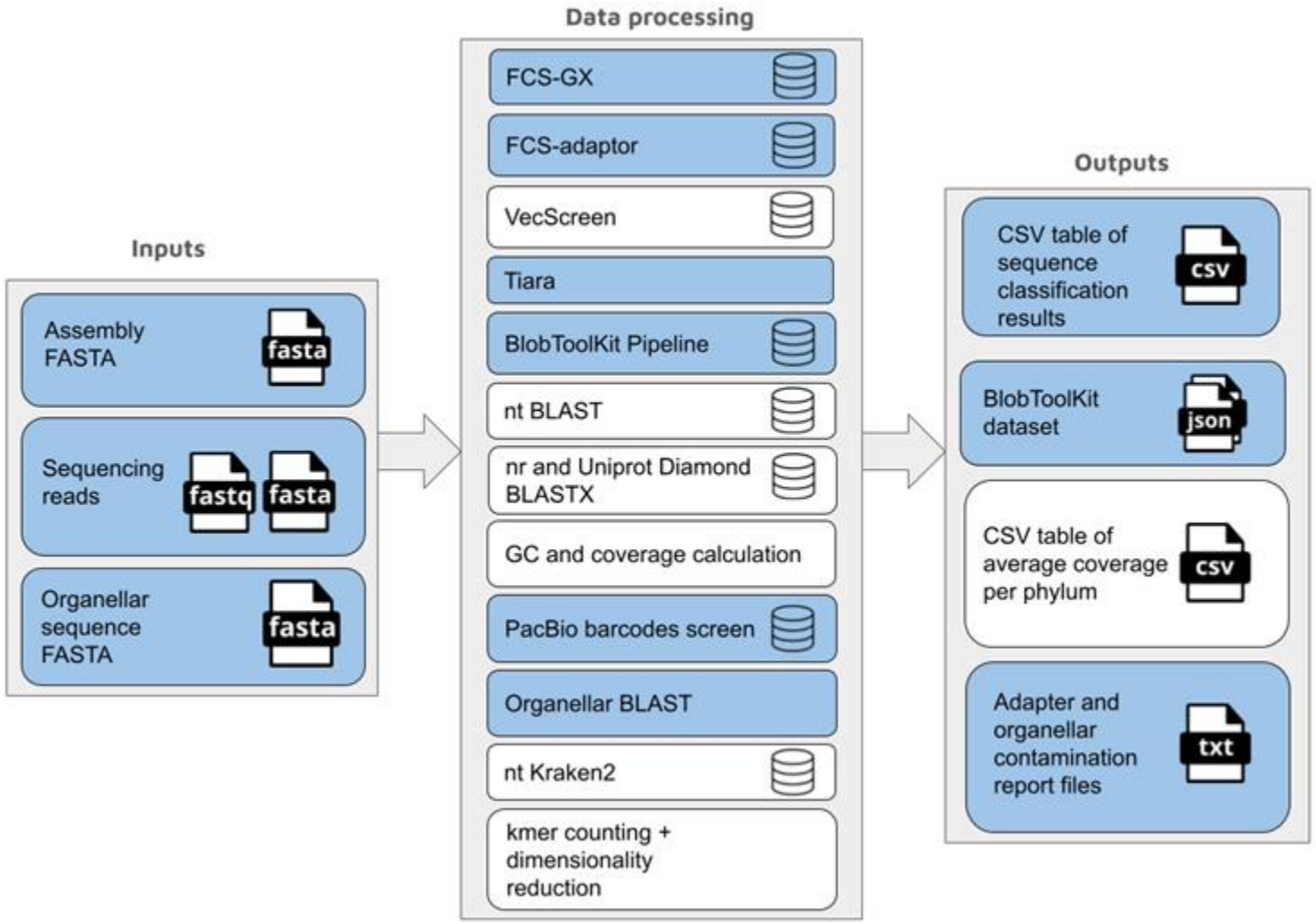
Cobionts/contamination disturb curation

Important not to submit contamination to public databases

Cause errors to genome downstream analysis

A pipeline for detecting cobionts and contaminants in genome assemblies

- The pipeline is called ASCC (Assembly Screen for Cobionts and Contaminants)
- The pipeline runs multiple tools for classifying sequences in the assembly and then merges the results
- The pipeline consists of Nextflow and Python scripts and Singularity images with software dependencies
- It is run with all genomes that are submitted for curation in the Tree of Life
  - The pipeline components that are used in the standard pre-curation runs are coloured blue in the diagram
  - The BlobToolKit pipeline runs its own processes for BLAST, Diamond, coverage and GC calculation
- A newer version of this pipeline that is nf-core compatible is undergoing development



FCS-GX

FCS-GX detects contamination from foreign organisms in genome sequences

FCS-GX operates in six main steps:

1. Repeat and low-complexity sequence masking
2. Alignment to reference database using GX aligner
3. Alignment refinement with high-scoring taxa matches
4. Classifying sequences to assign taxonomic divisions
5. Generating contaminant cleaning actions
6. Clean the genome

https://github.com/ncbi/fcs/wiki/FCS-GX-quickstart

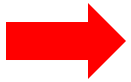
Run on a cluster

470 Gb disk space for database files
512 Gb shared memory

Docker or Singularity
FASTA genome

```
python3 ./fcs.py screen genome \
--fasta FCS_combo_test.fa \
--out-dir ./gx_out/ \
--gx-db /my_tmpfs/gxdb \
--tax-id 4932
```

Report



```
fcs_gx_report.txt contamination summary:
----------------------------------------
                                  seqs       bases
                                ------ -----------
TOTAL                            405      404339
-----                           ------ -----------
prok:g-proteobacteria            202      201923
anml:primates                    201      200894
virs:eukaryotic viruses            1        1000
anml:nematodes                     1         522

-------------------------------------------------

fcs_gx_report.txt action summary:
---------------------------------
                                  seqs       bases
                                ------ -----------
TOTAL                            405      404339
-----                           ------ -----------
EXCLUDE                          401      400522
FIX                                2        1922
TRIM                               2        1895

-------------------------------------------------
```

FCS-adaptor

FCS-adaptor detects adaptor and vector contamination in genome sequences

FCS-adaptor operates in three main steps:

1. BLAST alignment to reference database
2. Generate contaminant cleaning actions
3. Clean the genome

https://github.com/ncbi/fcs/wiki/FCS-adaptor-quickstart

```
./run_fcsadaptor.sh \
 --fasta-input FCS_combo_test.fa \
 --output-dir ./outputdir \
--euk
```

Report

#accession	length	action	range	name
seq_00001	230276	ACTION_TRIM	1..58	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00002	813242	ACTION_TRIM	1..58	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00003	316678	ACTION_TRIM	1..58	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00004	1531991	ACTION_TRIM	1..58	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00005	576932	ACTION_TRIM	1..58	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00006	270219	ACTION_TRIM	100001..100058	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00007	1090998	ACTION_TRIM	100001..100058	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00008	562701	ACTION_TRIM	100001..100058	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00009	439946	ACTION_TRIM	100001..100058	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00010	745809	ACTION_TRIM	100001..100058	CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer
seq_00018	522	ACTION_EXCLUDE		CONTAMINATION_SOURCE_TYPE_ADAPTOR:NGB00360.1:Illumina PCR Primer

Tiara

Approach for identification of eukaryotic sequences in the metagenomic data

https://github.com/ibe-uw/tiara?tab=readme-ov-file

- Archaea
- Bacteria
- Prokarya
- Eukarya
- Unknown
- Organelle

- Mitochondria
- Plasmid
- Unknown

Conda installation

```
tiara \
-i sample_input.fasta \
-o out.txt
```

The sequences in the fasta file should be at least 3000 bases long

> 1000 bp are not recommended

Reports:

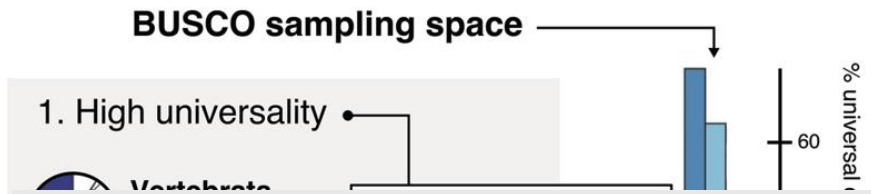
tiara_out.txt

SCAFFOLD_40	eukarya	n/a
SCAFFOLD_41	unknown	n/a
SCAFFOLD_42	prokarya	n/a
SCAFFOLD_43	eukarya	n/a
SCAFFOLD_44	unknown	n/a
SCAFFOLD_45	bacteria	n/a
SCAFFOLD_46	prokarya	n/a
SCAFFOLD_47	eukarya	n/a
SCAFFOLD_48	prokarya	n/a
SCAFFOLD_49	eukarya	n/a
SCAFFOLD_50	eukarya	n/a
SCAFFOLD_51	archaea	n/a
SCAFFOLD_52	prokarya	n/a
SCAFFOLD_53	unknown	n/a
SCAFFOLD_54	unknown	n/a
SCAFFOLD_55	prokarya	n/a
SCAFFOLD_56	bacteria	n/a

log_tiara_out.txt

```
First iteration statistics:
        archaea: 4
        bacteria: 41
        eukarya: 72
        prokarya: 20
        unknown: 24
```

BUSCO



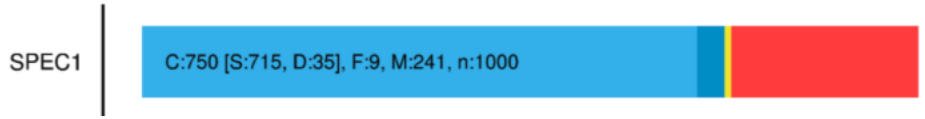

```
busco \
-i [SEQUENCE_FILE] \
-m [MODE] [genome, proteins, transcriptome]
```

GNU nano 6.2 full_table.tsv

```
# BUSCO version is: 5.7.1
# The lineage dataset is: insecta_odb10 (Creation date: 2020-09-10, number of genomes: 75, number of BUSCOs: 1367)
# Busco id      Status  Sequence        Gene Start      Gene End        Strand  Score   Length  OrthoDB url     Description
47at50557       Complete        SUPER_2 10678816        10779066        +       5641.4  3748    https://www.orthodb.org/v10?query=47at50557     Immunoglobulin-li>
547at50557      Complete        SUPER_16        3259017 3323910 +       3212.7  2303    https://www.orthodb.org/v10?query=547at50557    protein unc-80 homolog
755at50557      Complete        SUPER_6 8972264 9005574 +       3285.0  2174    https://www.orthodb.org/v10?query=755at50557    SRCR-like domain
767at50557      Complete        SUPER_15        11152708        11068556        -       3168.9  2175    https://www.orthodb.org/v10?query=767at50557    WD40-repe>
888at50557      Missing
966at50557      Complete        SUPER_26        1790620 1715357 -       1925.2  1616    https://www.orthodb.org/v10?query=966at50557    leucine-rich repeat serin>
983at50557      Complete        SUPER_9 2695240 2734518 +       3839.6  1825    https://www.orthodb.org/v10?query=983at50557    pre-mRNA-processing-splicing fact>
1451at50557     Complete        SUPER_26        5451462 5484363 +       1880.3  1940    https://www.orthodb.org/v10?query=1451at50557   small subunit processome >
1593at50557     Complete        SUPER_16        3735218 3760367 +       821.3   799     https://www.orthodb.org/v10?query=1593at50557   cadherin-89D
1621at50557     Complete        SUPER_8 8894874 8859355 -       3076.7  2237    https://www.orthodb.org/v10?query=1621at50557   GPCR, family 2, extracellular hor>
1859at50557     Complete        SUPER_10        5176171 5188751 +       1790.1  1097    https://www.orthodb.org/v10?query=1859at50557   Zinc finger C2H2 superfam>
1997at50557     Complete        SUPER_5 8408424 8435395 +       2091.7  1551    https://www.orthodb.org/v10?query=1997at50557   protein sidekick isoform X1
2098at50557     Complete        SUPER_17        2294828 2314902 +       1334.3  856     https://www.orthodb.org/v10?query=2098at50557   pecanex-like protein 1 is>
2101at50557     Complete        SUPER_10        8915532 8907477 -       1424.7  1296    https://www.orthodb.org/v10?query=2101at50557   serine/threonine-protein >
2110at50557     Complete        SUPER_8 1566852 1502237 -       1234.4  1272    https://www.orthodb.org/v10?query=2110at50557   proteasome-associated protein ECM>
2168at50557     Complete        scaffold_37     139503  164496  +       1214.1  910     https://www.orthodb.org/v10?query=2168at50557   DNA polymerase
2187at50557     Complete        SUPER_5 16212387        16227901        +       2806.2  1685    https://www.orthodb.org/v10?query=2187at50557   eIF-2-alpha kinas>
```

Marker genes present in at least 90% of the species in a given lineage

Single copy in 90% of those species



Installation

Conda

Docker

↓

Recommended

Manually

↓

Many dependencies

ASCC runs all BUSCO datasets

ASCC (Assembly Screen for Cobionts and Contaminants)

It run multiple cobiont/contaminant detection tools in parallel and then combine the results and look for consensus in the results.

```
bsub -n1 -q basement -R"span[hosts=1]" -o <Tol_ID>_ascc_minimal.o -e
<Tol_ID>_ascc_minimal.e -M5000 -R 'select[mem>5000] rusage[mem=5000]'
"/software/team311/ea10/ascc_latest/cobiontcheck/ascc.py \
<path_to_your_assembly.fa> \
--static_config_path
/lustre/scratch123/tol/teams/grit/mh6/ascc_logs/static_settings.config \
--pacbio_reads_path <path_to_your_pacbio_reads> \
--assembly_title <Tol_ID> \
--sci_name <'species_scientific_name'> \
--taxid <species_taxonomic_ID> \
--steps tiara coverage fcs-gx fcs-adaptor create_btk_dataset btk_busco
autofilter_assembly \
--threads 24 \
--pipeline_run_folder <output_directory_path>"
```

BlobTool Kit (BTK)

https://grit-btk.tol.sanger.ac.uk

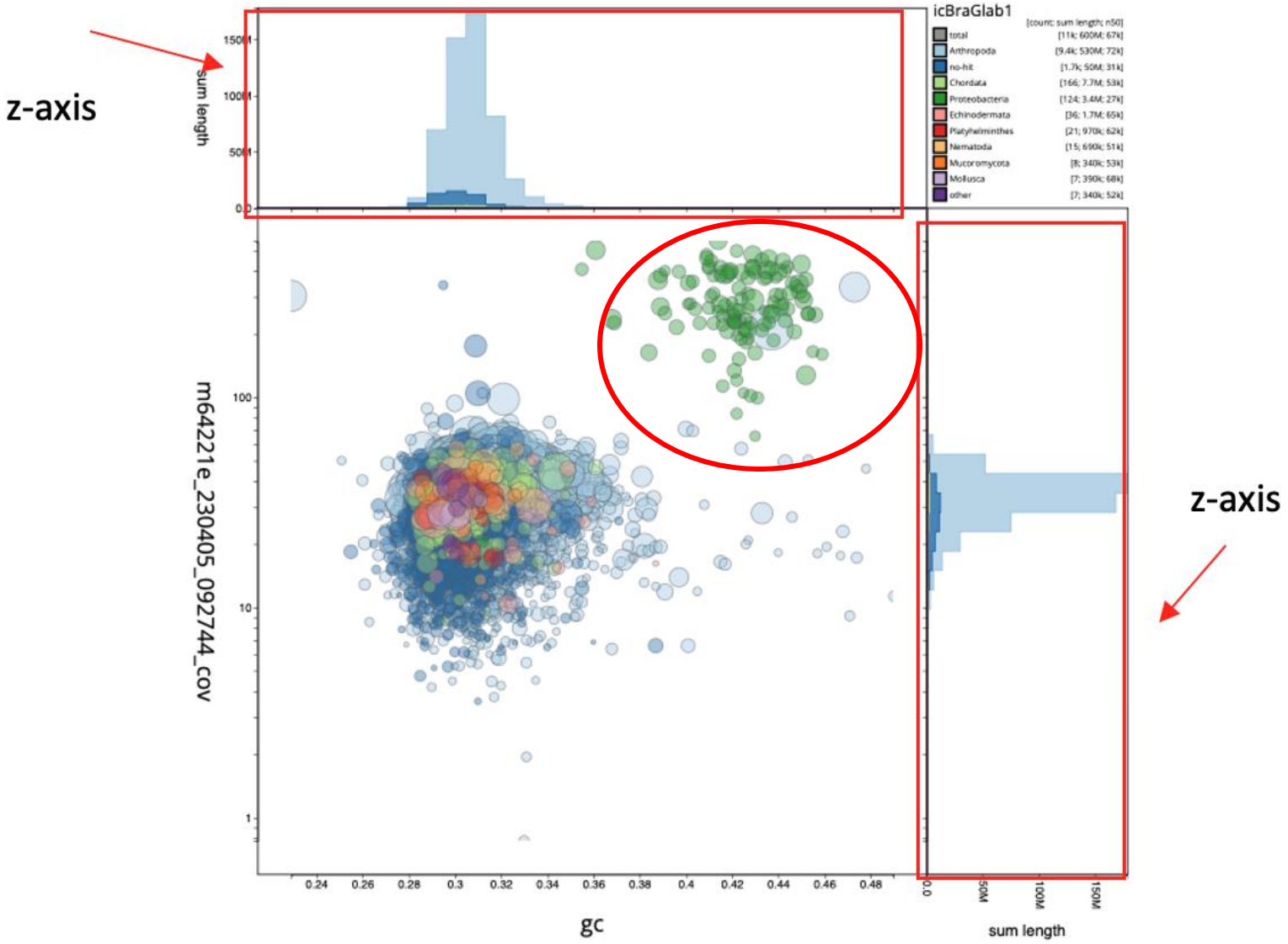
Main menu

Selecting a dataset

How to filter your BTK results?

BlobTool Kit (BTK)

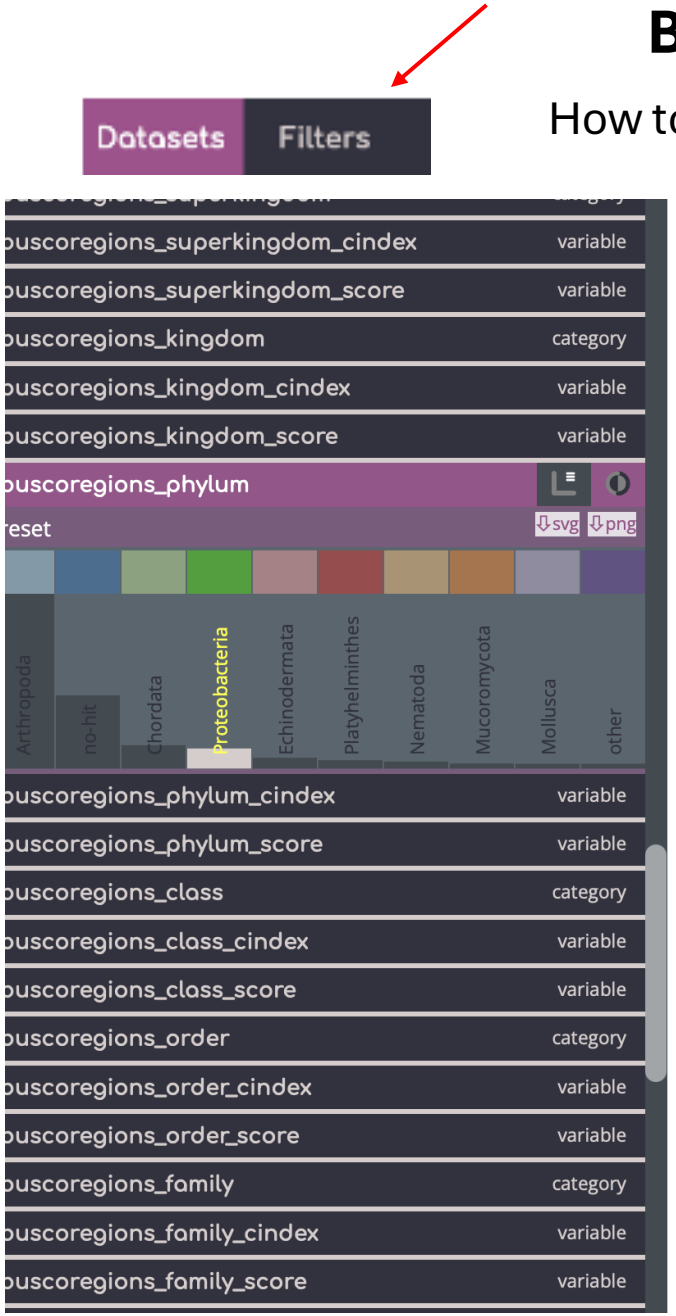
How to filter your BTK results?

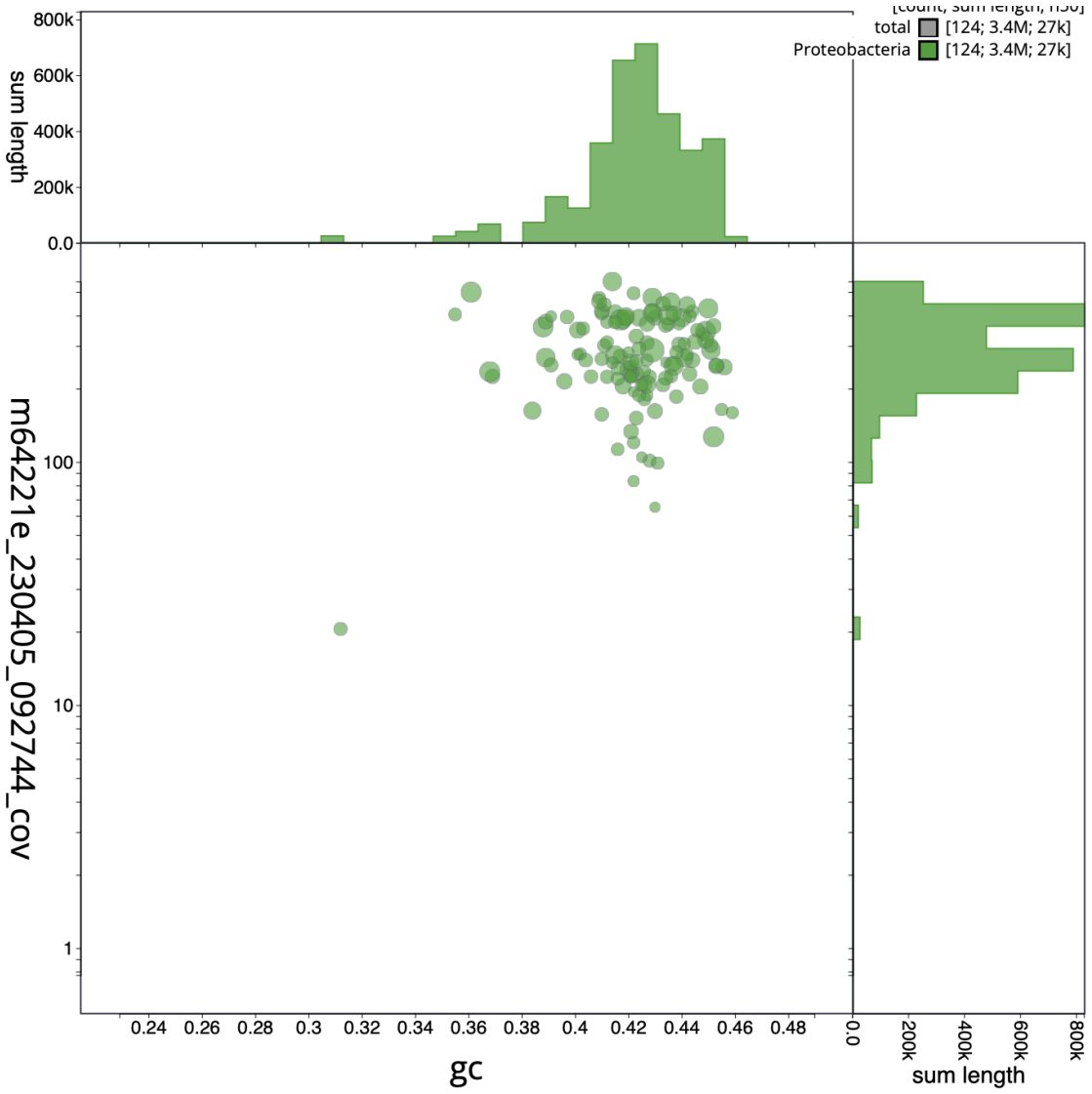


Blob plot of base coverage in m64221e against GC proportion for scaffolds in assembly icBraGlab1. *More*

BlobTool Kit (BTK)

How to filter your BTK results?





BlobTool Kit (BTK)

How to filter your BTK results?

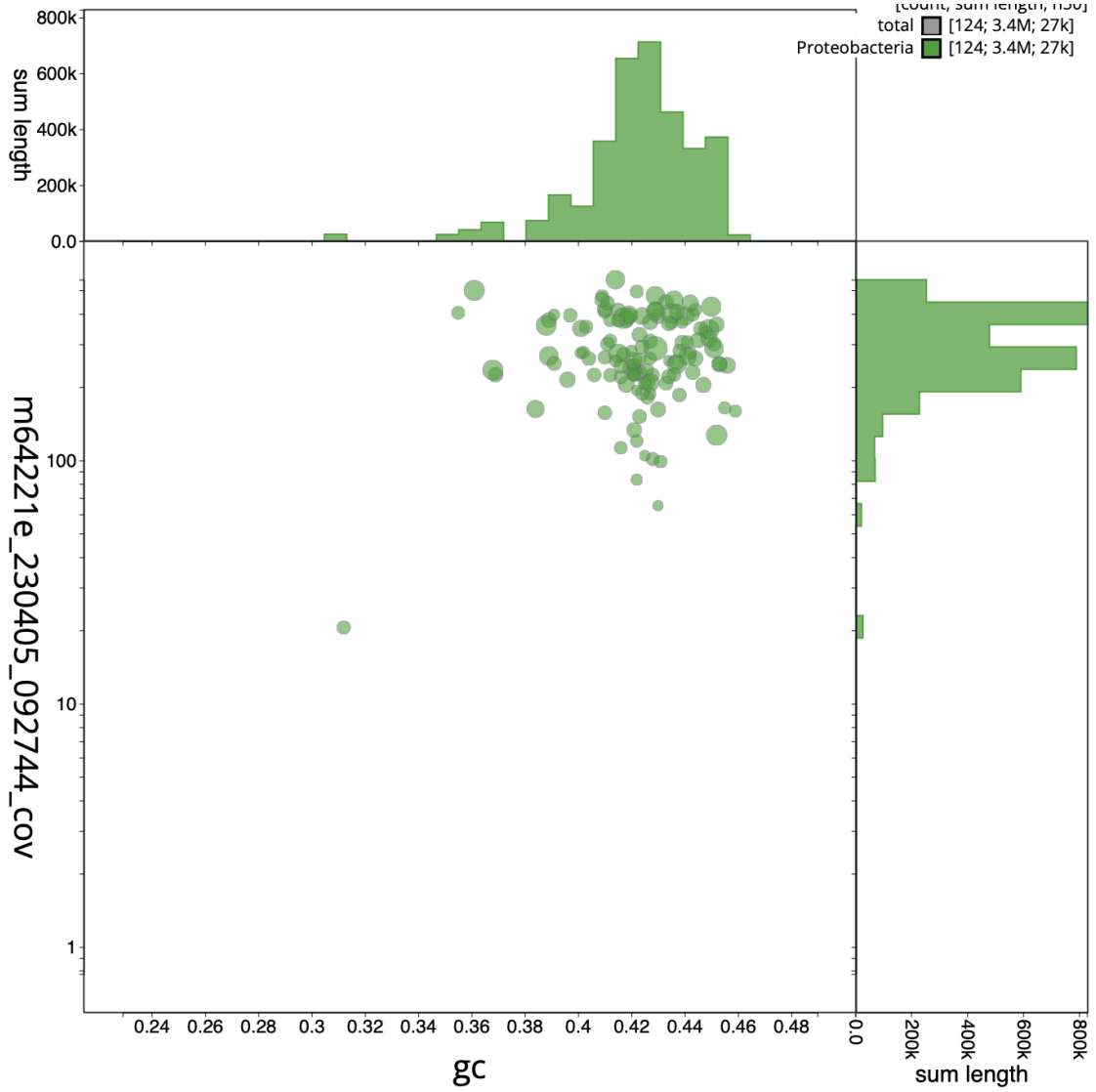
Removing bacteria

blob busco cumulative detail report snail table

csv

	#	ID	Length	GC	m64221e 230405_0... Coverage	buscoregions phylum	tiara
			Variables			Categories	
☐	843	atg000884l_1	27,450	0.439	304.647	Proteobacteria	bacteria
☐	1150	atg001207l_1	29,226	0.423	228.319	Proteobacteria	bacteria
☐	1488	atg001556l_1	24,451	0.428	224.34	Proteobacteria	bacteria
☐	1831	atg001916l_1	33,371	0.42	239.488	Proteobacteria	bacteria
☐	2230	atg002339l_1	30,148	0.456	245.605	Proteobacteria	bacteria
☐	2994	atg003164l_1	40,355	0.449	344.579	Proteobacteria	bacteria
☐	3090	atg003270l_1	22,448	0.426	235.724	Proteobacteria	bacteria
☐	3394	atg003601l_1	38,385	0.45	428.365	Proteobacteria	bacteria
☐	3407	atg003615l_1	26,815	0.443	230.149	Proteobacteria	bacteria
☐	4251	atg004536l_1	28,903	0.447	204.059	Proteobacteria	bacteria
☐	4311	atg004603l_1	25,269	0.421	246.035	Proteobacteria	bacteria
☐	4431	atg004736l_1	22,354	0.459	159.485	Proteobacteria	bacteria
☐	4557	atg004874l_1	22,454	0.427	262.347	Proteobacteria	bacteria
☐	4561	atg004878l_1	26,667	0.415	414.329	Proteobacteria	bacteria
☐	4625	atg004947l_1	28,223	0.43	161.886	Proteobacteria	bacteria
☐	4724	atg005058l_1	26,207	0.427	309.064	Proteobacteria	bacteria
☐	4769	atg005112l_1	40,733	0.361	499.657	Proteobacteria	bacteria
☐	4837	atg005193l_1	23,039	0.441	306.255	Proteobacteria	bacteria
☐	5208	atg005611l_1	37,045	0.451	288.739	Proteobacteria	bacteria
☐	5367	atg005783l_1	23,803	0.435	368.617	Proteobacteria	bacteria
☐	5387	atg005804l_1	26,648	0.417	272.141	Proteobacteria	bacteria

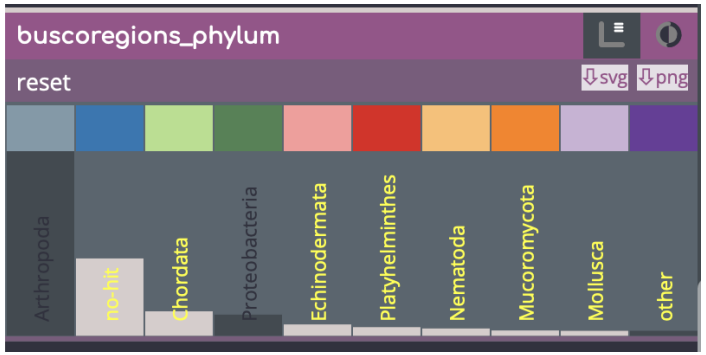
buscoregions_family_score variable



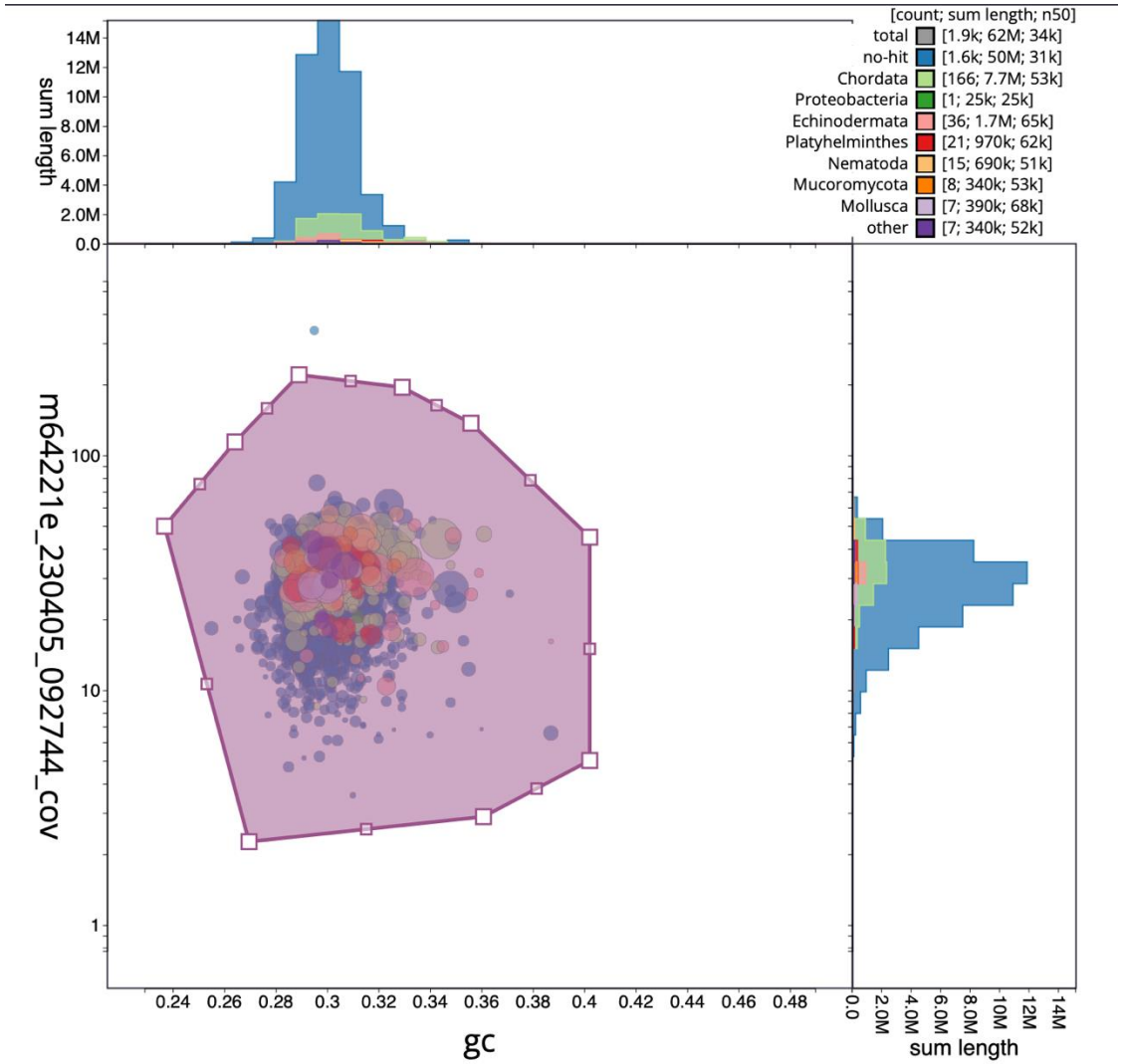
BlobTool Kit (BTK)

How to filter your BTK results?

Turn Arthropoda and bacteria off

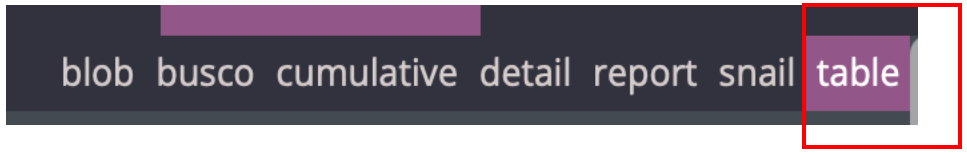


Select scaffolds with overlapping distribution to target species

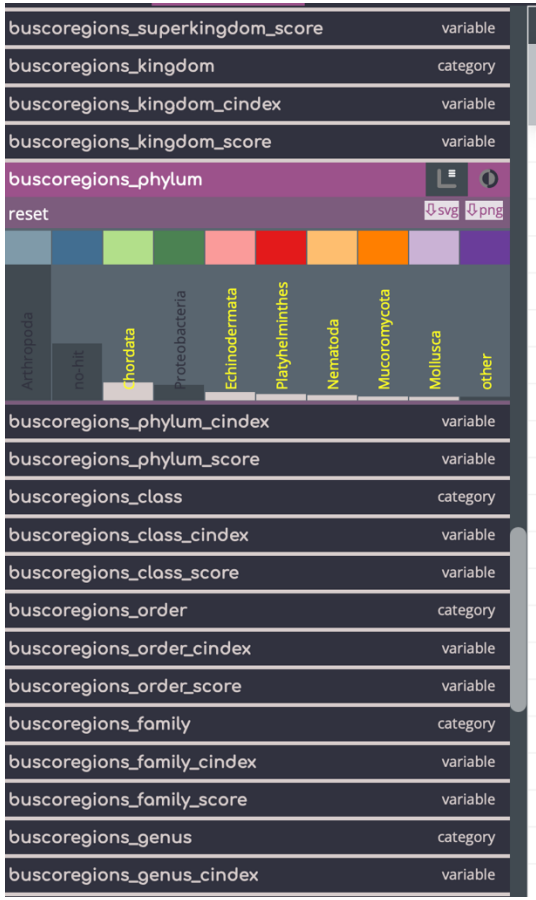


BlobTool Kit (BTK)

How to filter your BTK results?



Turn off ‘no-hit’



Removing eukarya

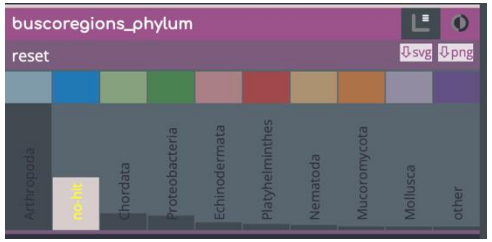
	Variables				Categories		csv
#	ID	Length	GC	m64221e 230405_0... Coverage	buscoregions phylum	tiara	
68	atg000070l_1	94,902	0.328	35.827	Chordata	eukarya	
87	atg000091l_1	42,900	0.307	29.833	Chordata	eukarya	
110	atg000115l_1	78,872	0.301	38.931	Platyhelminthes	eukarya	
327	atg000341l_1	44,315	0.317	17.217	Platyhelminthes	eukarya	
412	atg000431l_1	35,603	0.303	33.388	Mollusca	eukarya	
456	atg000476l_1	193,811	0.302	40.505	Chordata	eukarya	
517	atg000543l_1	107,416	0.292	28.526	Echinodermata	eukarya	
556	atg000583l_1	73,959	0.319	33.593	Chordata	eukarya	
572	atg000600l_1	46,400	0.296	32.728	Chordata	eukarya	
659	atg000690l_1	120,238	0.304	26.654	Chordata	eukarya	
703	atg000734l_1	66,117	0.301	36.927	Platyhelminthes	eukarya	
773	atg000809l_1	195,453	0.302	32.474	Echinodermata	eukarya	
831	atg000871l_1	146,186	0.296	28.968	Chordata	eukarya	
898	atg000944l_1	59,648	0.294	26.704	Mollusca	eukarya	
922	atg000969l_1	36,770	0.286	41.402	Platyhelminthes	eukarya	
931	atg000978l_1	39,096	0.32	28.708	Chordata	eukarya	
949	atg000996l_1	72,322	0.304	36.699	Chordata	eukarya	
1010	atg001061l_1	76,596	0.301	35.376	Chordata	eukarya	
1114	atg001170l_1	86,311	0.319	32.675	Echinodermata	eukarya	
1120	atg001176l_1	46,174	0.315	40.04	Chordata	eukarya	
1127	atg001183l_1	108,627	0.334	30.294	Echinodermata	eukarya	

BlobTool Kit (BTK)

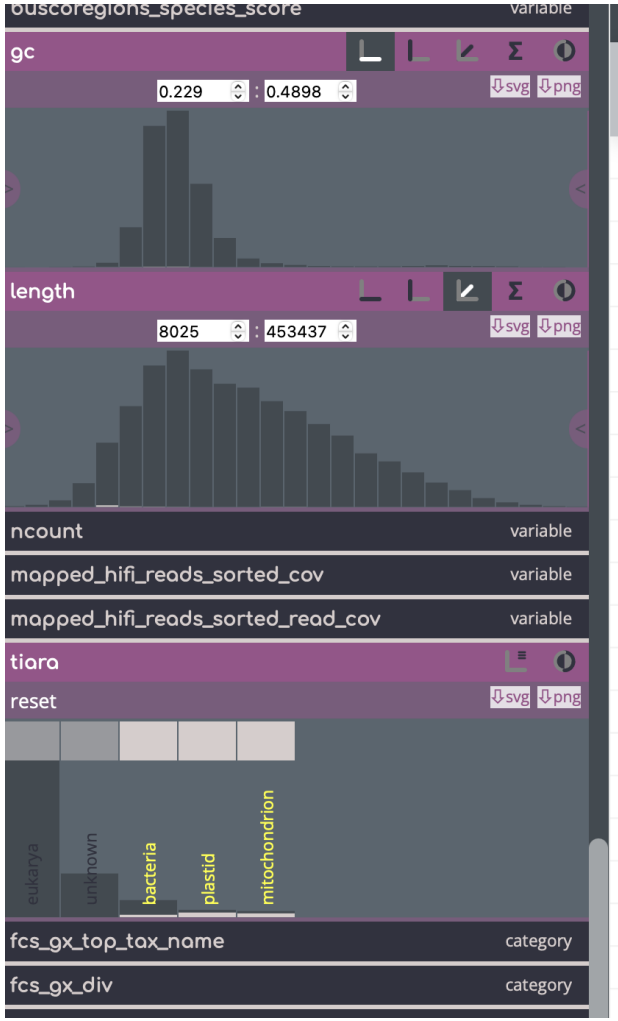
How to filter your BTK results?

Should be removed

Turn on ‘no-hit’ only



In Tiara:
Turn on eukarya and unknown →



	#	ID	Length	GC	m64221e 230405_0... Coverage	buscoregions phylum	tiara
			Variables			Categories	csv
	4128	atg004406l_1	51,239	0.309	174.949	no-hit	bacteria
	3787	atg004033l_1	60,367	0.31	103.994	no-hit	bacteria
	8021	atg009011l_1	27,215	0.306	53.308	no-hit	bacteria
	10088	atg012737l_1	19,966	0.278	51.09	no-hit	plastid
	3046	atg003224l_1	15,545	0.311	44.044	no-hit	plastid
	8598	atg009825l_1	16,591	0.31	40.137	no-hit	mitochondrion
	4733	atg005067l_1	17,076	0.283	37.52	no-hit	plastid
	2261	atg002371l_1	16,932	0.316	36.962	no-hit	plastid
	8217	atg009284l_1	17,974	0.309	28.514	no-hit	plastid
	9570	atg011447l_1	22,053	0.304	27.946	no-hit	mitochondrion
	7233	atg008006l_1	22,140	0.296	24.434	no-hit	mitochondrion
	8639	atg009886l_1	15,552	0.285	23.562	no-hit	plastid
	4280	atg004571l_1	15,922	0.297	23.098	no-hit	mitochondrion
	8555	atg009756l_1	19,054	0.297	20.651	no-hit	plastid
	7261	atg008040l_1	15,398	0.301	17.36	no-hit	plastid
	9897	atg012144l_1	16,248	0.306	15.761	no-hit	mitochondrion
	7683	atg008566l_1	16,661	0.296	15.045	no-hit	plastid
	9633	atg011585l_1	15,310	0.285	14.553	no-hit	plastid
	5295	atg005708l_1	18,370	0.297	14.439	no-hit	mitochondrion
	10289	hap_ptg000572l_1_1	15,897	0.302	14.34	no-hit	mitochondrion

Scaffolds which are ‘unknown’ by Tiara, ‘no-hit’ by BUSCO and share distribution with target → BLASTn

BlobTool Kit (BTK)

Visualizing the data in btk_dataset folder

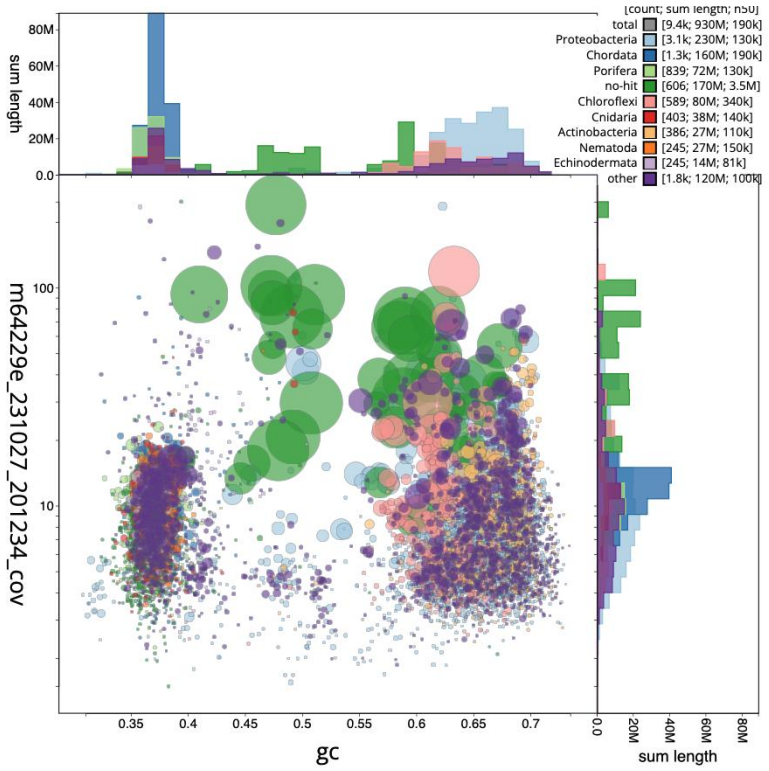
Complementary approach: by BP reads coverage

odRhoOdor1

Datasets | Filters | Lists | Settings | Summary | Help | About

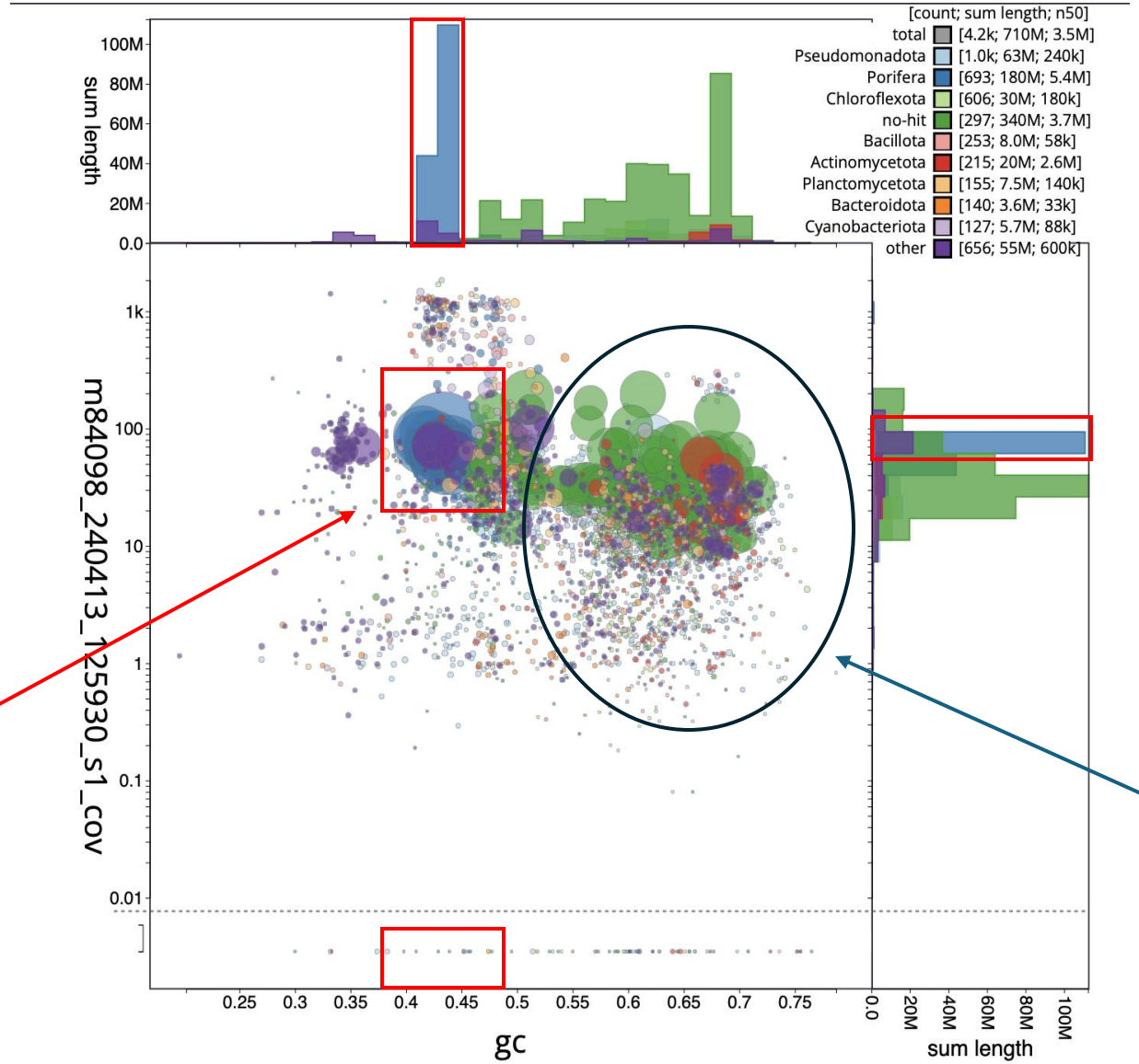
buscogenes_genus_cindex variable
buscogenes_genus_score variable
buscogenes_species category
buscogenes_species_cindex variable
buscogenes_species_score variable
buscoregions_superkingdom category
buscoregions_superkingdom_cindex variable
buscoregions_superkingdom_score variable
buscoregions_kingdom category
buscoregions_kingdom_cindex variable
buscoregions_kingdom_score variable
buscoregions_phylum
buscoregions_phylum_cindex variable
buscoregions_phylum_score variable
buscoregions_class category
buscoregions_class_cindex variable
buscoregions_class_score variable
buscoregions_order category
buscoregions_order_cindex variable
buscoregions_order_score variable
buscoregions_family category
buscoregions_family_cindex variable
buscoregions_family_score variable
buscoregions_genus category
buscoregions_genus_cindex variable

			Variables		Categories		
#	ID	Length	GC	m64229e 231027_201234 Coverage	buscoregions phylum	tiara	fcs gx_action
0	ptg000001l_1	3,143,044	0.621	48.699	no-hit	bacteria	REVIEW
1	ptg000002c_1	4,712,121	0.584	65.497	no-hit	bacteria	NA
2	ptg000003l_1	22,161	0.35	4.693	no-hit	eukarya	NA
3	ptg000004c_1	6,927,729	0.589	30.57	no-hit	bacteria	NA
4	ptg000005l_1	49,677	0.401	4.063	Uroviricota	bacteria	NA
5	ptg000006l_1	537,450	0.655	28.929	Proteobacteria	bacteria	REVIEW
6	ptg000007l_1	214,416	0.668	15.593	Proteobacteria	bacteria	REVIEW
7	ptg000008c_1	4,331,260	0.633	118.628	Chloroflexi	bacteria	NA
8	ptg000009l_1	47,020	0.642	6.801	Chloroflexi	bacteria	NA
9	ptg000010l_1	202,664	0.623	9.76	Proteobacteria	bacteria	EXCLUDE
10	ptg000011l_1	56,915	0.638	6.773	Proteobacteria	bacteria	REVIEW
11	ptg000013l_1	114,401	0.373	12.214	Platyhelminthes	eukarya	NA
12	ptg000014l_1	98,088	0.41	51.654	Firmicutes	bacteria	NA
13	ptg000015l_1	14,557	0.379	3.403	no-hit	eukarya	NA
14	ptg000016l_1	47,290	0.38	11.025	Arthropoda	eukarya	NA
15	ptg000017l_1	167,878	0.38	13.687	Chordata	eukarya	NA
16	ptg000018l_1	119,928	0.365	10.017	Cnidaria	eukarya	NA
17	ptg000019l_1	6,504,112	0.508	29.649	no-hit	bacteria	NA
18	ptg000020l_1	19,742	0.358	6.729	no-hit	eukarya	NA
19	ptg000021l_1	166,766	0.383	6.593	Nematoda	eukarya	NA
20	ptg000022l_1	422,751	0.611	18.755	Chloroflexi	bacteria	REVIEW
21	ptg000023l_1	41,742	0.712	12.684	Proteobacteria	bacteria	REVIEW
22	ptg000024l_1	136,394	0.372	10.316	Echinodermata	eukarya	NA
23	ptg000025l_1	725,255	0.684	21.905	Proteobacteria	bacteria	REVIEW
24	ptg000026l_1	30,999	0.369	6.452	Porifera	eukarya	NA
25	ptg000027l_1	44,928	0.657	10.081	Proteobacteria	bacteria	REVIEW
26	ptg000028l_1	1,295,866	0.626	48.929	Proteobacteria	bacteria	NA
27	ptg000029l_1	371,116	0.675	22.815	Proteobacteria	bacteria	REVIEW



Microbial-rich species - High amount of cobionts

Rhabdastrella globostellata



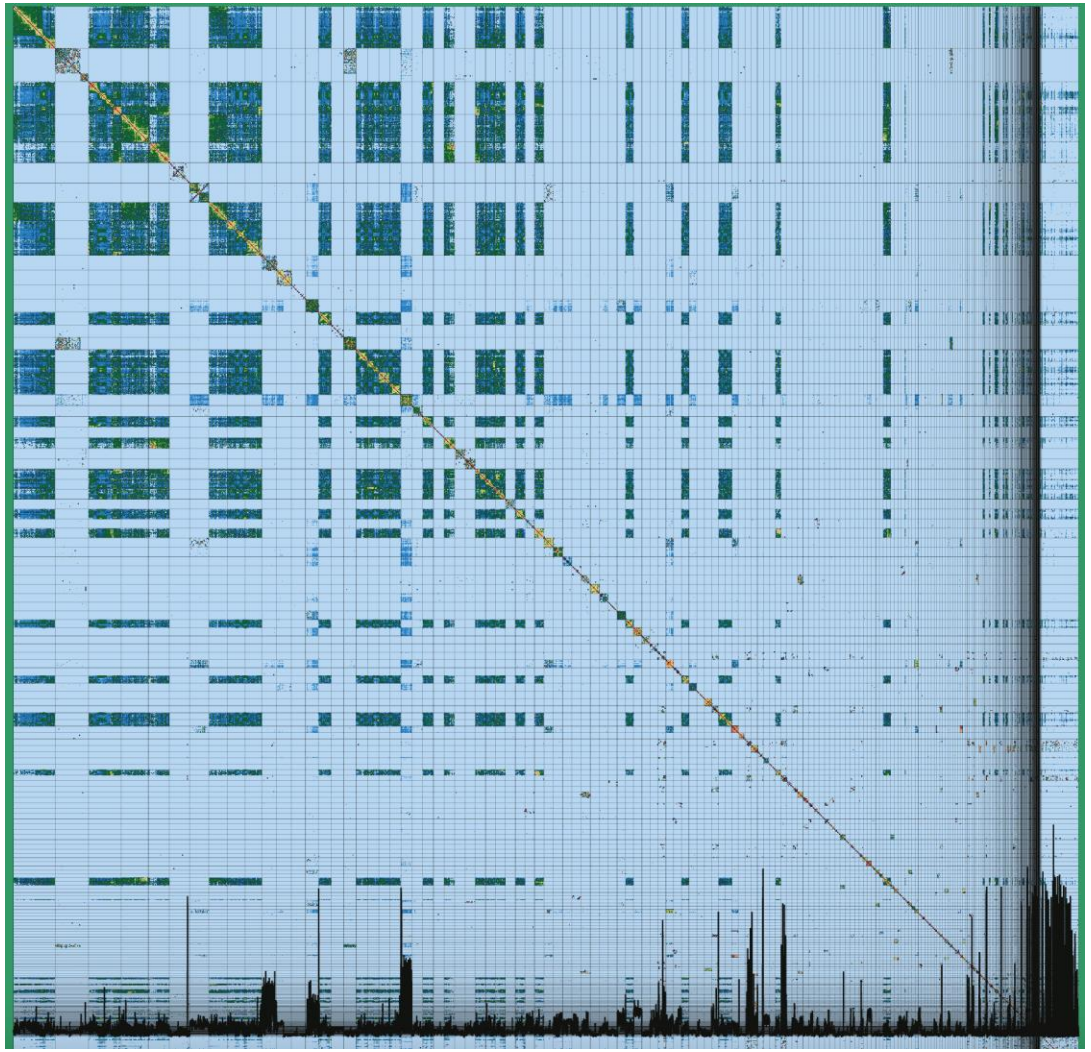
Dark blue: sponge

Other colours: Cobionts

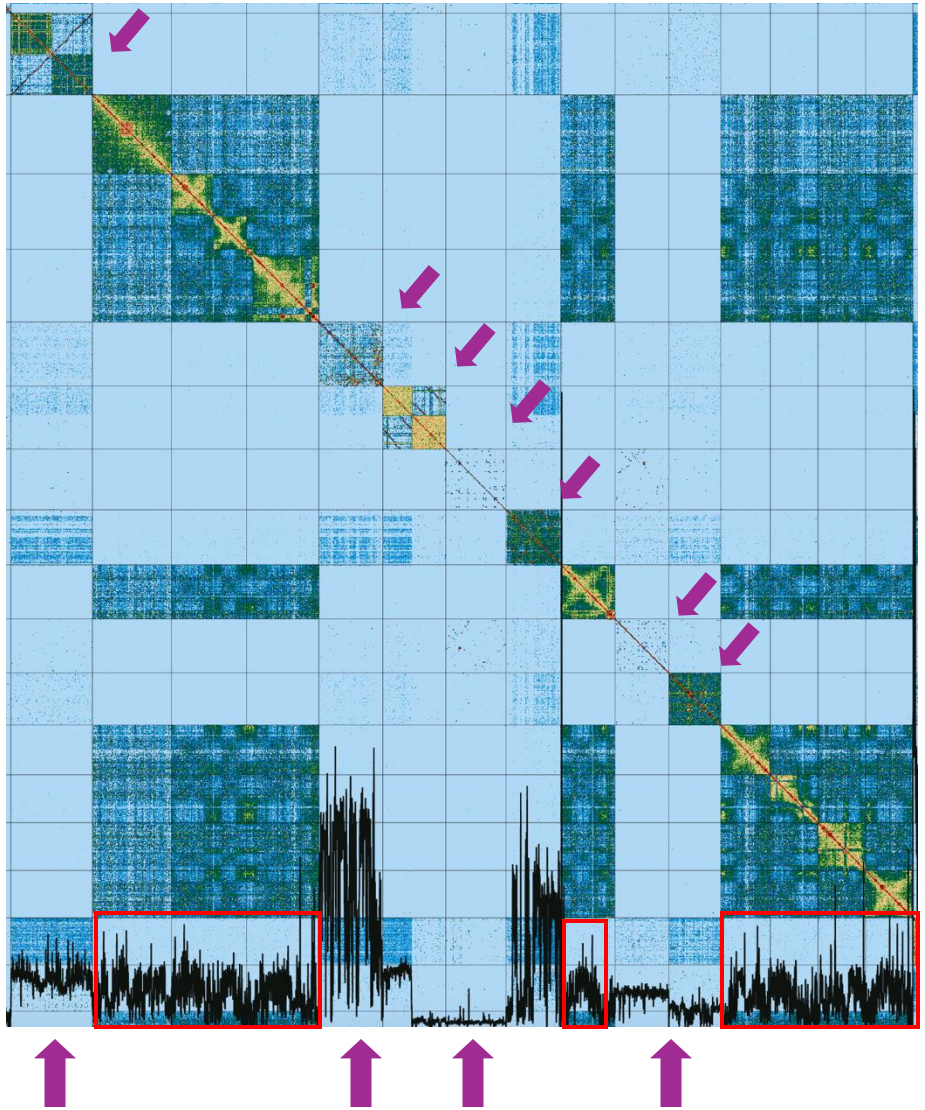
Microbial-rich species - High amount of cobionts

Rhabdastrella globostellata

Before curation



Zoom-in



Microbial-rich species - High amount of cobionts

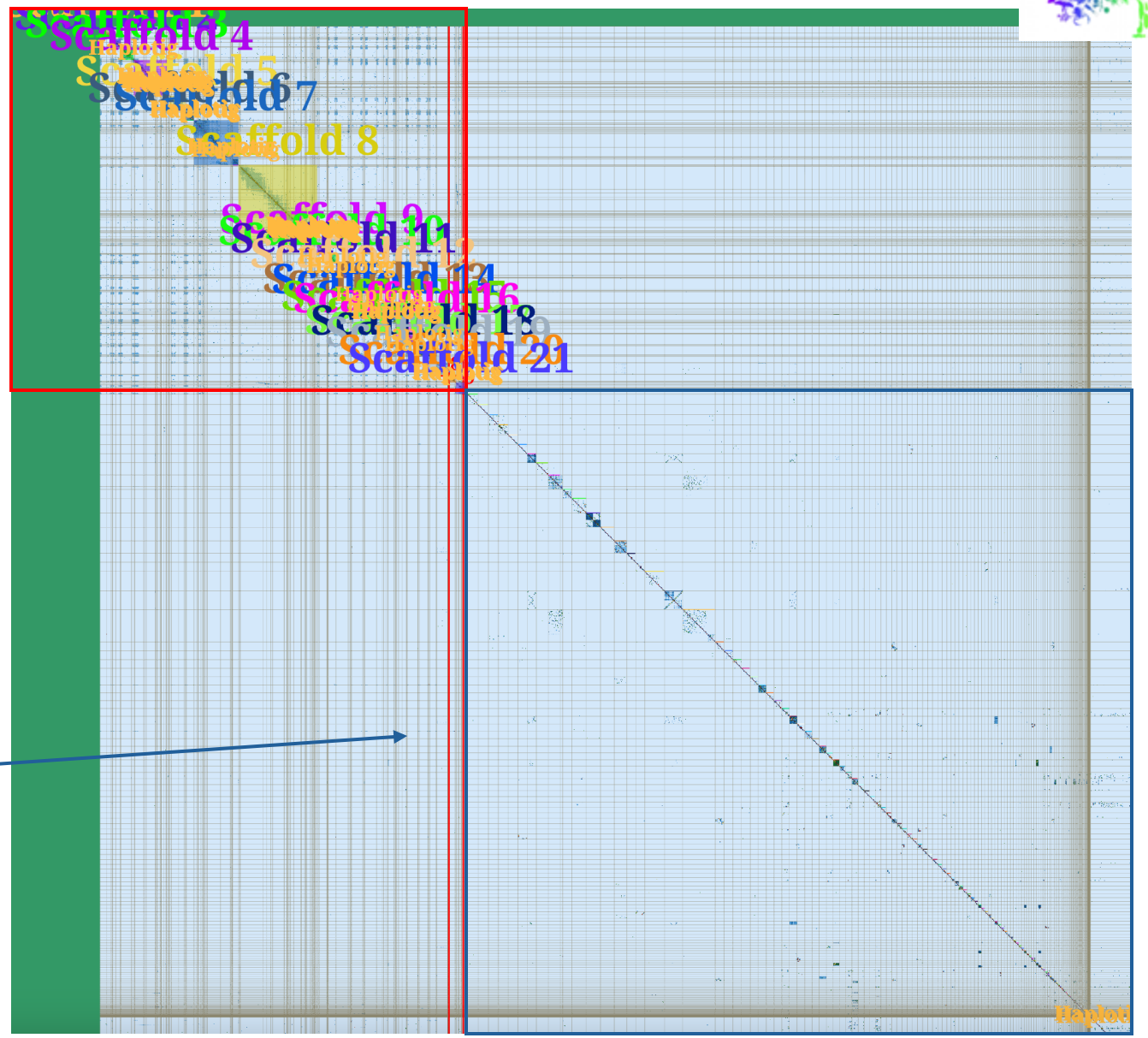
Rhabdastrella globostellata

After curation

Sponge genome

Difference in the background
Telomeres lighting up

Cobionts to be removed
during decon
(> 60%)



Hands-on

- https://github.com/csantos-alvess/Physalia-Manual-Genome-Curation/blob/main/Session2.2.md