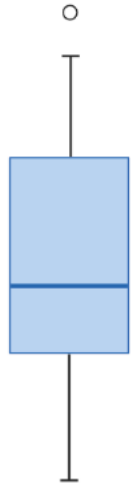
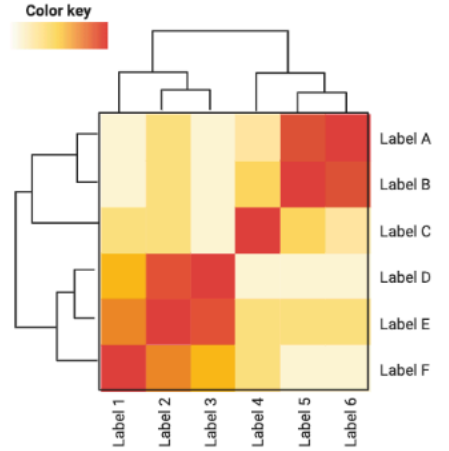
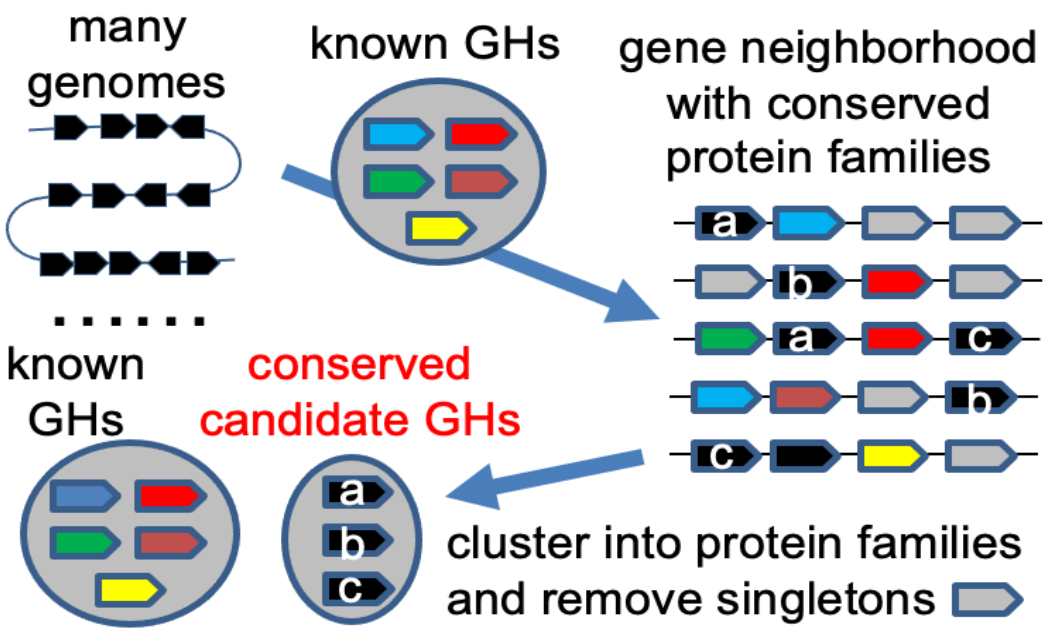
Choose a day and RSVP: https://www.when2meet.com/?15855740-coiPZ

Computational Biology Summer Workshop





- **Free** to anyone with life/agricultural science background
- Organized by **Dr. Yanbin Yin**, Associate Prof of Nebraska Food for Health Center & Food Science & Tech Dept
- Taught by **experienced bioinformatics postdocs and students** of NFHC
- Practical **Bash, Python and R skills** to analyze sequencing data and produce publishable graphs
- **Hands-on project** of (meta)**genome mining** for carbohydrate active enzymes as a case study
- **Date**: **A half day (4h)** of the **3rd week** (20-23) of **June** will be chosen according to a **time survey** (see the top)
- **Location**: In person at **Food Innovation Center 220**



Schedule (June 20 and 23, FIC 111)

Time	Presenter	Topic
12:00-12:30pm	Yanbin	Introduction to the workshop & project
12:30-1:00pm	Tang	UHGG Bacteroides MAGs & stat visualization
1:00-1:45pm	Yuchen	CAZyme & CGC annotation & stat visualization
1:45-2:30pm	Yuchen	Anvio pan-genome analysis & phylogeny visualization
2:30-2:45pm		Break
2:45-3:30pm	Bowen	Pfam annotation & protein seq clustering of nSGs & visualization
3:30-4:15pm	Tang	Glycan substrate prediction & CGC clustering & network visualization
4:15-5:00pm	Jinfang	ML prediction for new GH/PL from conserved nSGs

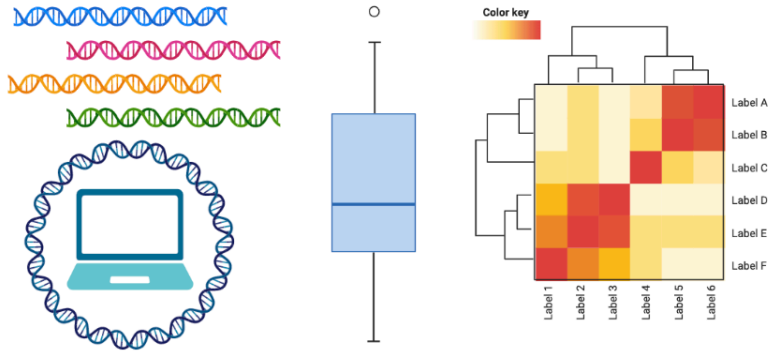
GitHub page: https://github.com/tli14/Workshop_2022_YinLab

acknowledgements

Jerry Arthur Ved

Yuchen Tang Bowen Jinfang

what you need to know



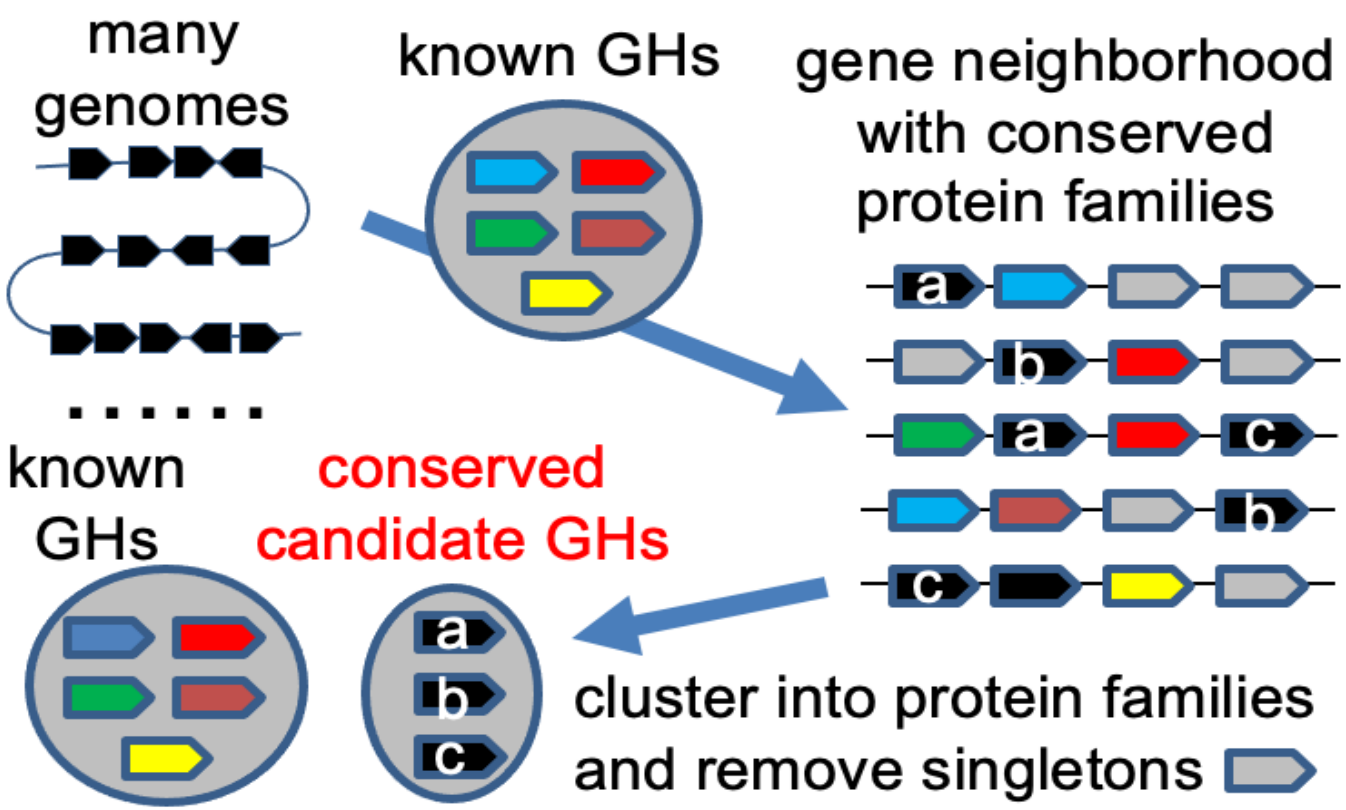
- use holland computing cluster so bring a laptop and have an HCC account
- hands-on activities with a real-world genome mining project
- focus on data visualization during workshop demonstration
- some data computing takes long time, but all intermediate files and codes are provided
- use HCC OnDemand -> R studio + Shell terminal

lower your expectations

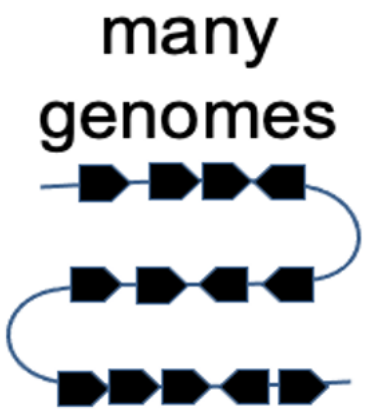
- don’t expect you will understand/learn everything we say
- it is okay if you can’t follow every step (all the codes and files are provided)
- the instructors won’t have time to explain all the codes (it’s a win as long as you can get the right graphs/files)
- questions are welcomed, but we will try to stay on time

about the project

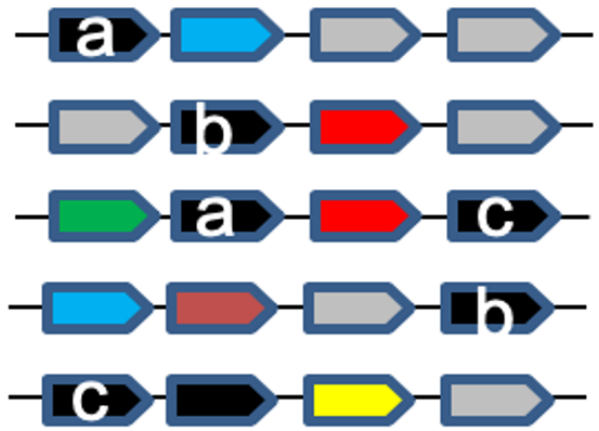
- a new project from a recent proposal
- it will or should work
- input -> genomes,
  output -> new GH predictions
- our focus is data analysis and visualization



steps



- input -> 45 Bacteroides genomes from human gut
- find CAZymes and CGCs (CAZyme gene clusters)
- find null or non-signature genes in CGCs
- cluster nSGs into families



- predict the likelihood of conserved nSG families to be new CAZymes

(meta)genome mining

Targeted search of genomes for specific pathways or genes with important economic values using advanced computational data mining approaches

Review Article | Published: 03 June 2021

Mining genomes to illuminate the specialized chemistry of life

Marnix H. Medema, Tristan de Rond & Bradley S. Moore

Nature Reviews Genetics **22**, 553–571 (2021) | Cite this article

RESEARCH ARTICLE

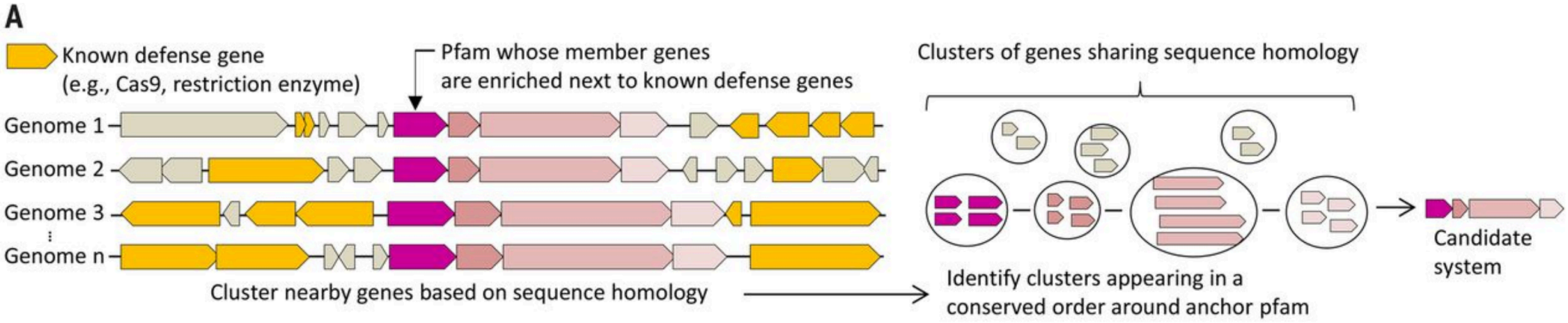
Systematic discovery of antiphage defense systems in the microbial pangenome

Shany Doron*, Sarah Melamed*, Gal Ofir, Azita Leavitt, Anna Lopatina, Mai Keren, Gil Amitai, Rotem Sorek†

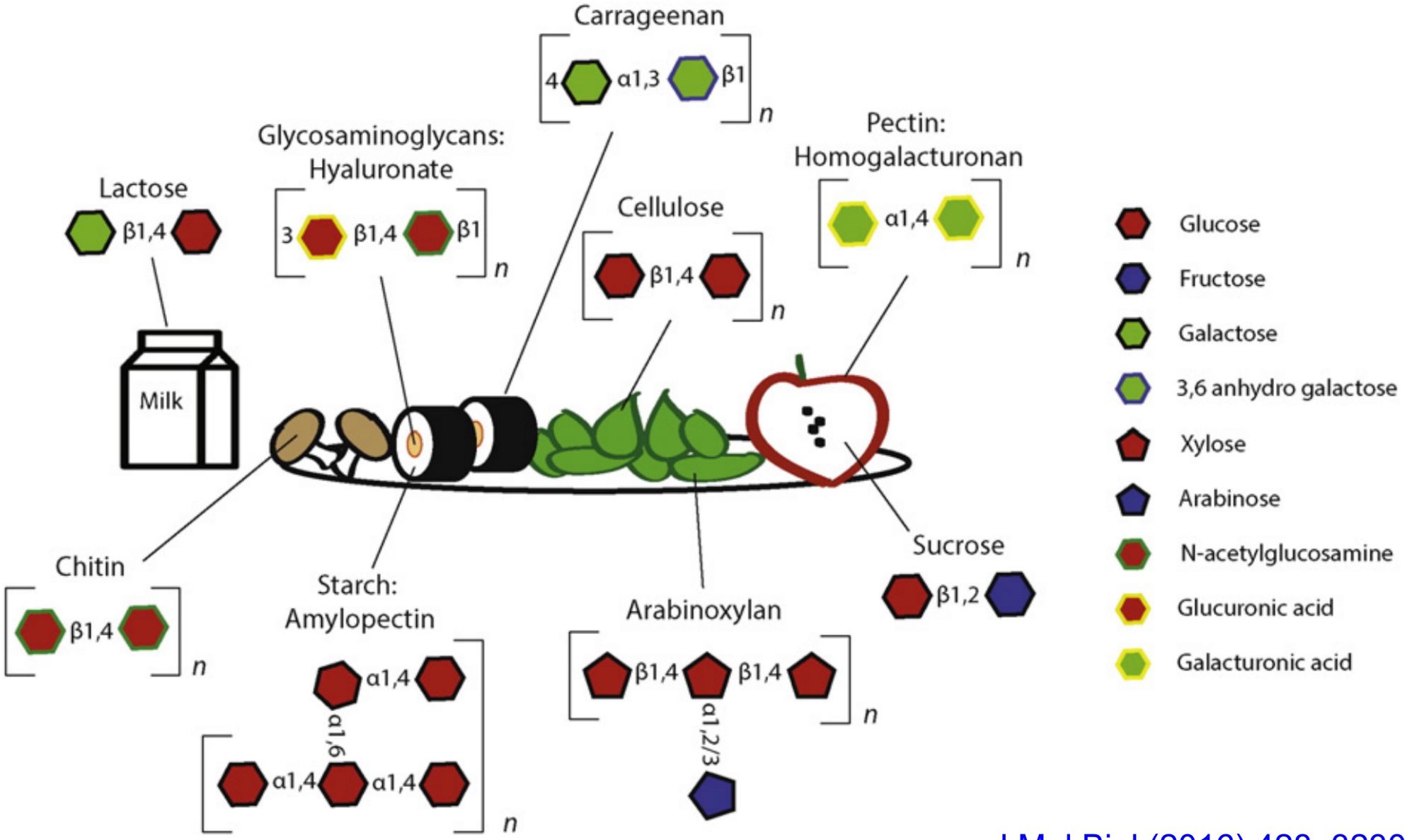
+ See all authors and affiliations

Science 02 Mar 2018:
Vol. 359, Issue 6379, eaar4120
DOI: 10.1126/science.aar4120

Defense island: gene clusters that encode antiphage defense genes



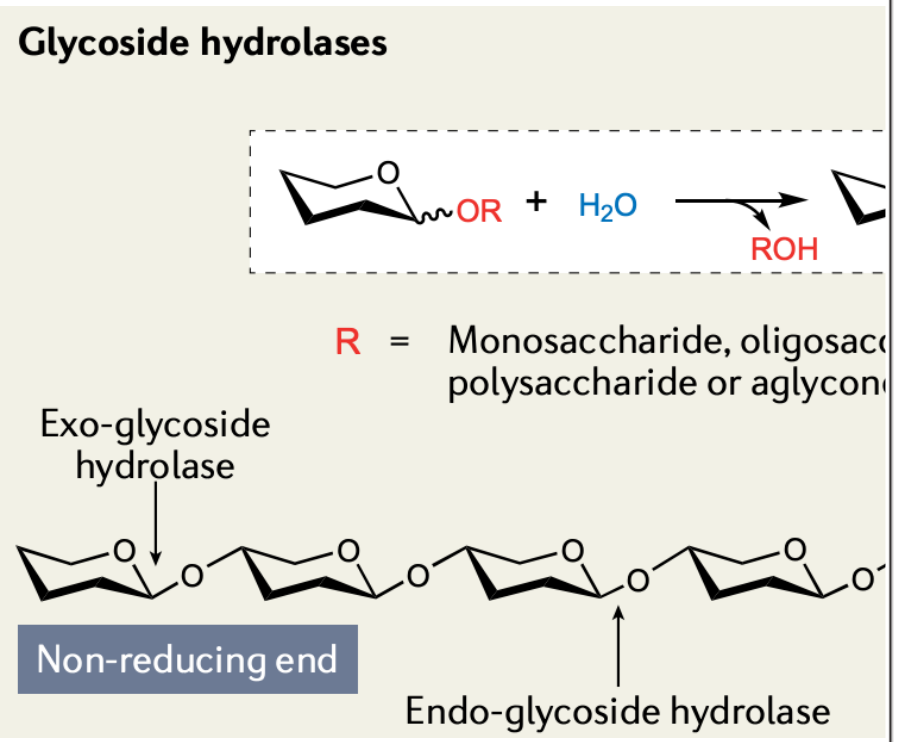
a high diversity of dietary carbohydrates

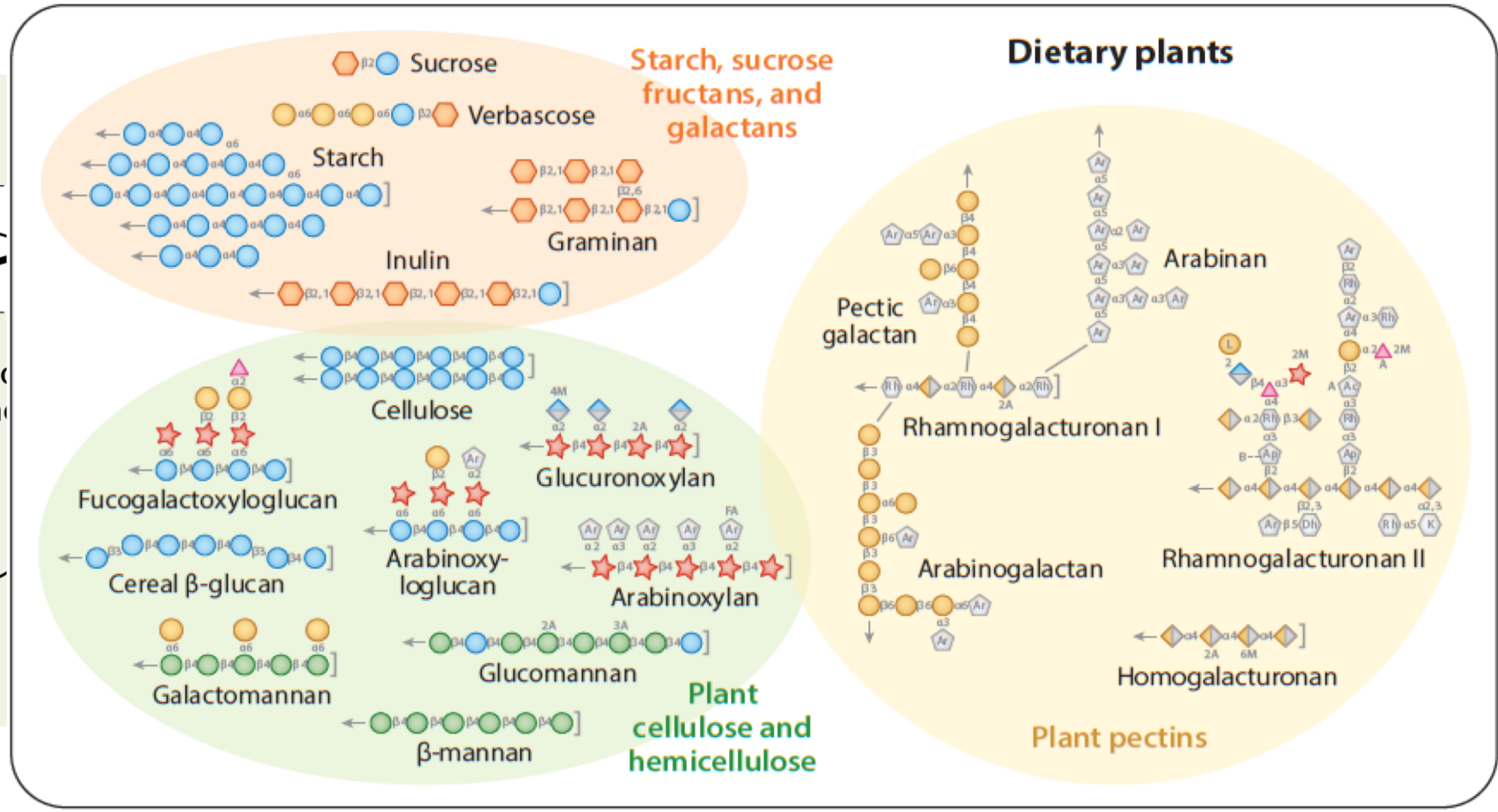


J Mol Biol (2016) 428, 3230–3252

diverse glycosidic linkages exist in the dietary carbs

Nature Reviews Microbiology (2022)





Annu. Rev. Microbiol (2017) 71:349–69

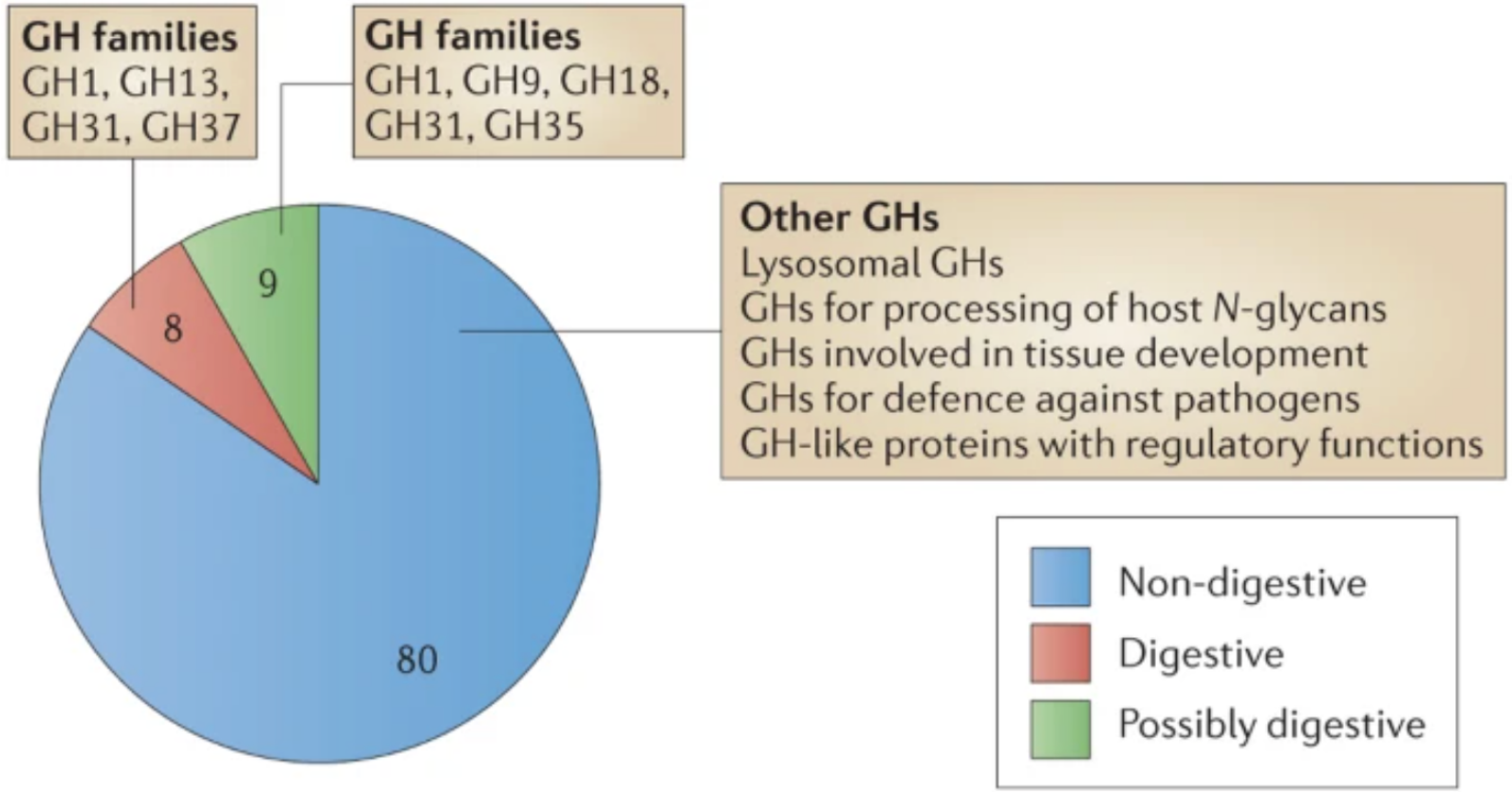
diverse glycosidic linkages need various CAZymes to break

http://www.cazy.org/

- GlycosylTransferases (GTs): **115**
- **Glycoside Hydrolases (GHs): 172**
- Polysaccharide Lyases (PLs): **42**
- Carbohydrate Esterases (CEs): **19**
- Auxiliary Activities (AAs): **17**
- Carbohydrate-Binding Modules (CBMs): **89**

Cantarel B. et al. 2009, Nucleic Acids Res
Lombard V. et al. 2014, Nucleic Acids Res

Human encodes 17 food digesting GHs



NATURE REVIEWS | **MICROBIOLOGY,** doi:10.1038/nrmicro3050, Kaoutari, 2013

gut bacteria dedicate > 6% of their genes to CAZymes

Bacterium	Total CAZymes	GH	GT	PL	CE	Total CBMs
Bacteroides thetaiotaomicron VPI-5482	386	263	87	16	20	31
B. xylanisolvens XB1A*	349	224	81	22	22	26
B. vulgatus ATCC-8482	279	177	78	7	17	18
B. fragilis 638R	223	138	78	1	6	26
Roseburia intestinalis XB6B4*	175	115	46	0	14	11
Butyrivibrio fibrisolvens 16/4*	115	75	37	0	3	31
Ruminococcus champanellensis 18P13*	87	54	12	9	12	34
Bifidobacterium adolescentis ATCC15703	94	54	37	0	3	6

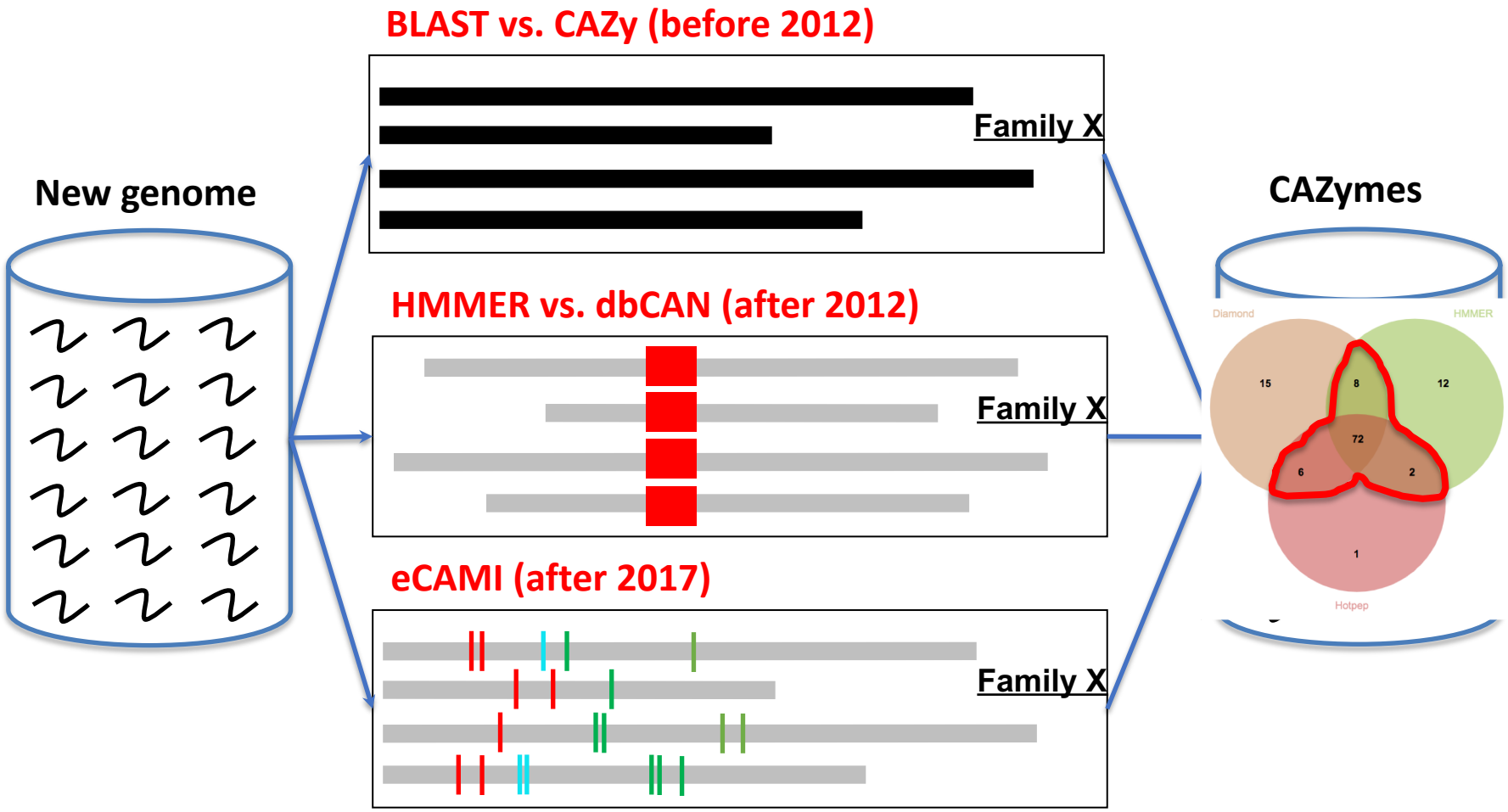
Gut Microbes 3:4, 289-306; 2012

1000 (species) x 100 (genes) = 100,000 CAZymes

dbCAN2 combines three tools for CAZyme genome mining

Tanner

Le

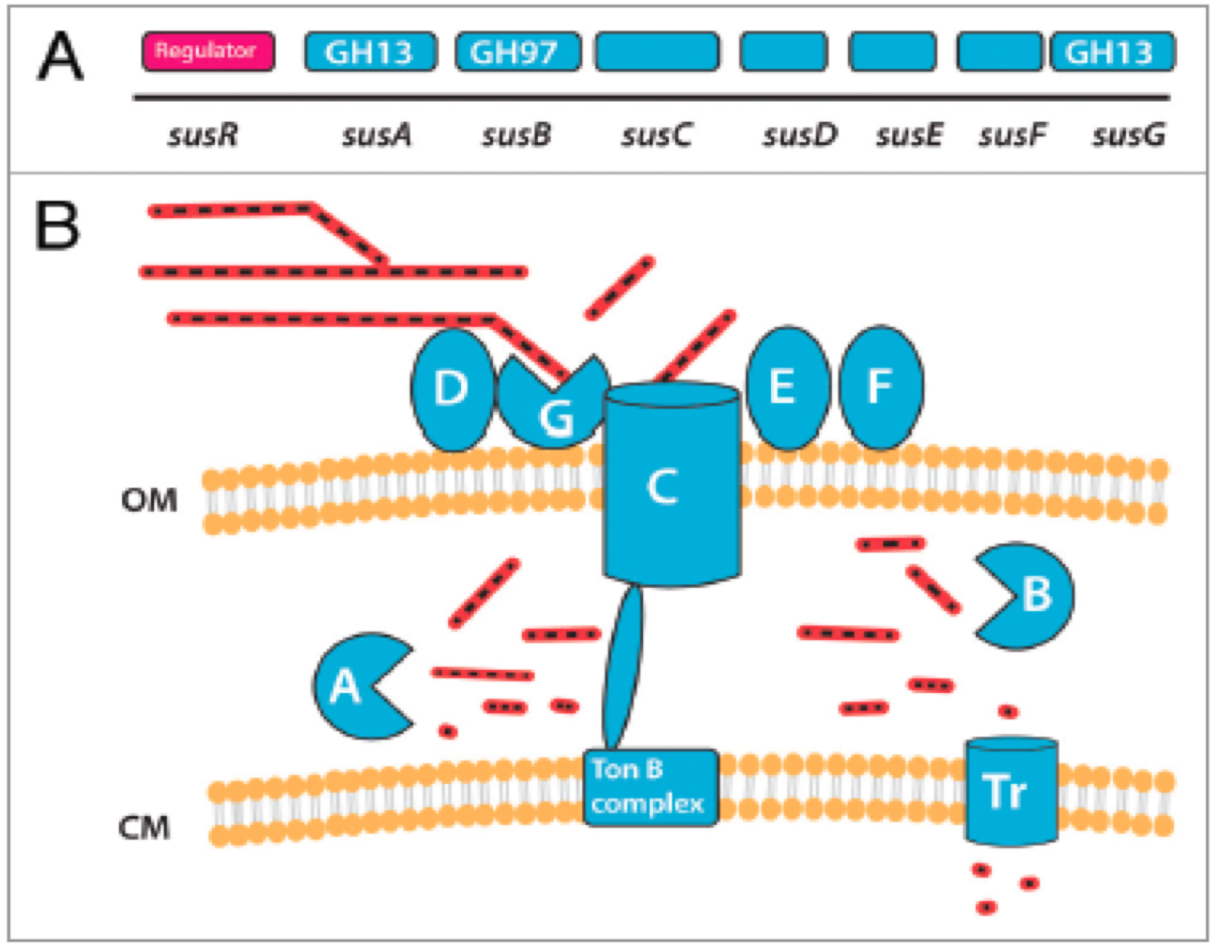


dbCAN: *Yin et al, Nucleic acids research 2012*

eCAMI: *Xu J et al., Bioinformatics, 2020*

PUL: polysaccharide utilization loci

Sus in Bacteroides thetaiotaomicron



Gut Microbes 3:4, 289-306; 2012

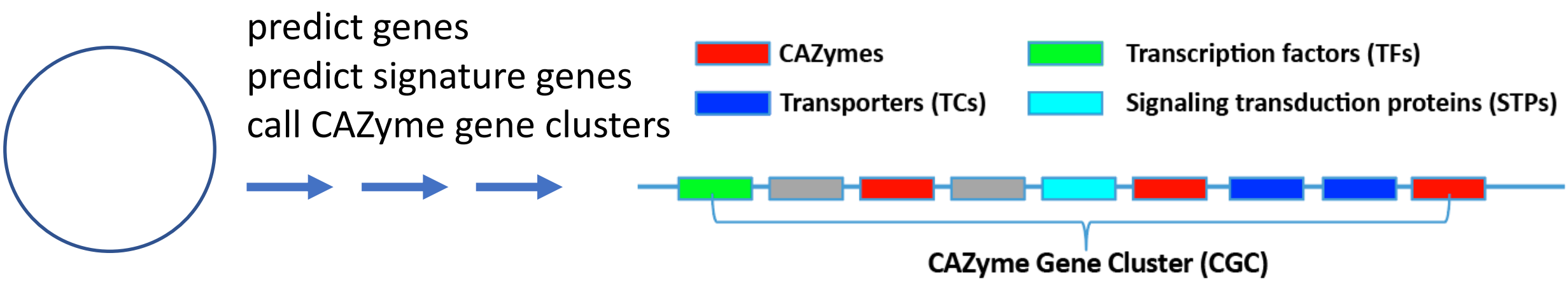
Species	Number of PULs[a]	Habitat
Bacteroides thetaiotaomicron	90	Gut
Bacteroides ovatus	118	Gut
Bacteroides fragilis	60	Gut
Bacteroides xylanisolvens	100	Gut
Proteinphilum acetatigenes	77	Wastewater sludge
Chitinophaga pinensis	106	Soil
Chitinophaga niabensis	153	Soil
Cytophaga hutchinsonii	2	Soil
Flavobacterium johnsoniae	40	Soil/freshwater
Prevotella melaninogenica	25	Upper respiratory tract
Prevotella salivae	32	Mouth
Prevotella ruminicola	24	Rumen
Zobellia galactanivorans	61	Marine

Environmental Microbiology Reports (2021)

CGC-Finder: predict CAZyme gene clusters from (meta)genomes

Huang et al., Nucleic Acids Res 2018

Le



The distance threshold:

how many non-signature genes are inserted between two adjacent signature genes

Summary

dbCAN-PUL

a very high diversity of carbs in our diet (other environment too)

dbCAN -> automated CAZyme genome mining

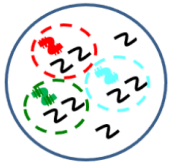
dbCAN-PUL -> literature derived PULs with known carb substrates

complex carb degrading genes form gene clusters in genomes

CGC-Finder can mine for these CGCs (without substrates)

https://bcb.unl.edu/dbCAN2
https://github.com/linnabrown/run_dbcan
https://bcb.unl.edu/dbCAN_PUL
https://github.com/yinlabniu/eCAMI
https://bcb.unl.edu/dbCAN_seq/
https://bcb.unl.edu/plantcazyme/
https://github.com/linnabrown/run_dbcan
https://bcb.unl.edu/dbCAN_PUL
https://github.com/yinlabniu/eCAMI
https://bcb.unl.edu/dbCAN_seq/
https://bcb.unl.edu/plantcazyme/
https://bcb.unl.edu/pHMM-Tree/source/

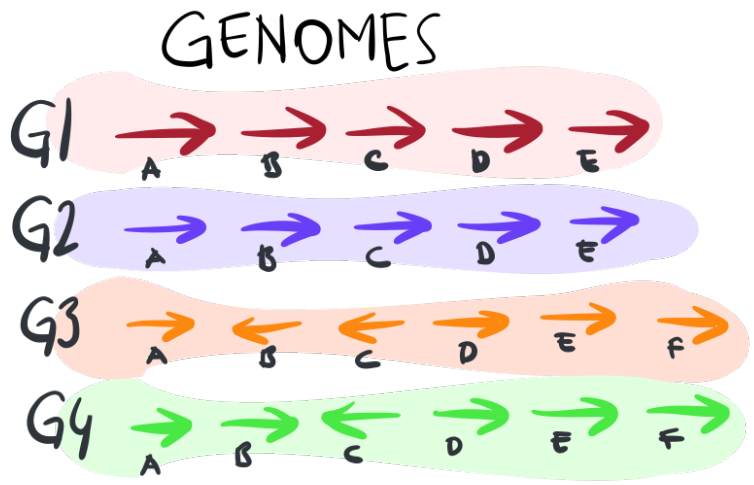
dbCAN-seq

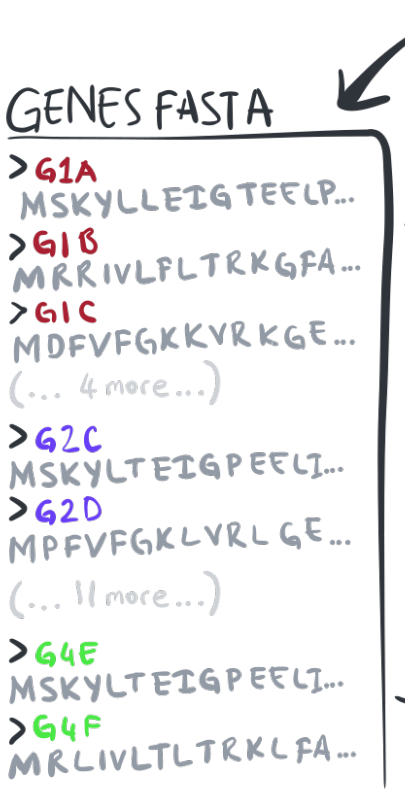


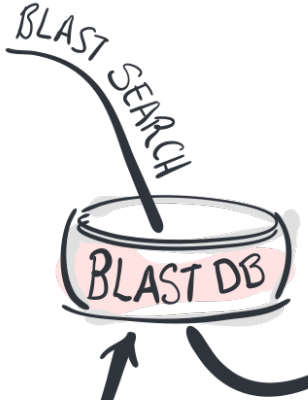
dbCAN-sub
Subfamily -> EC -> substrates

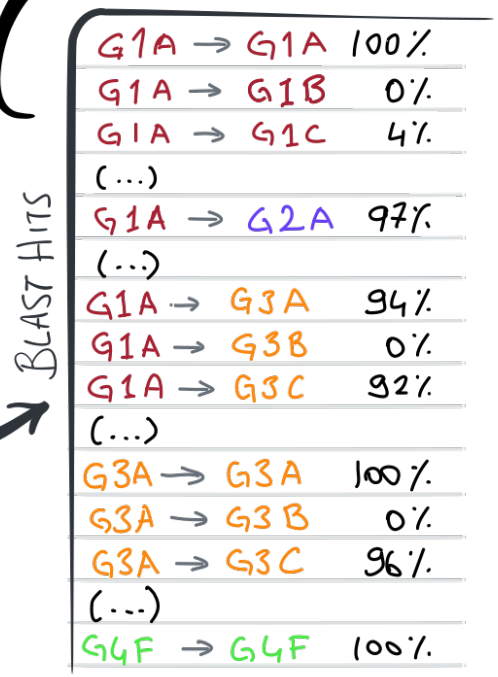
eCAMI

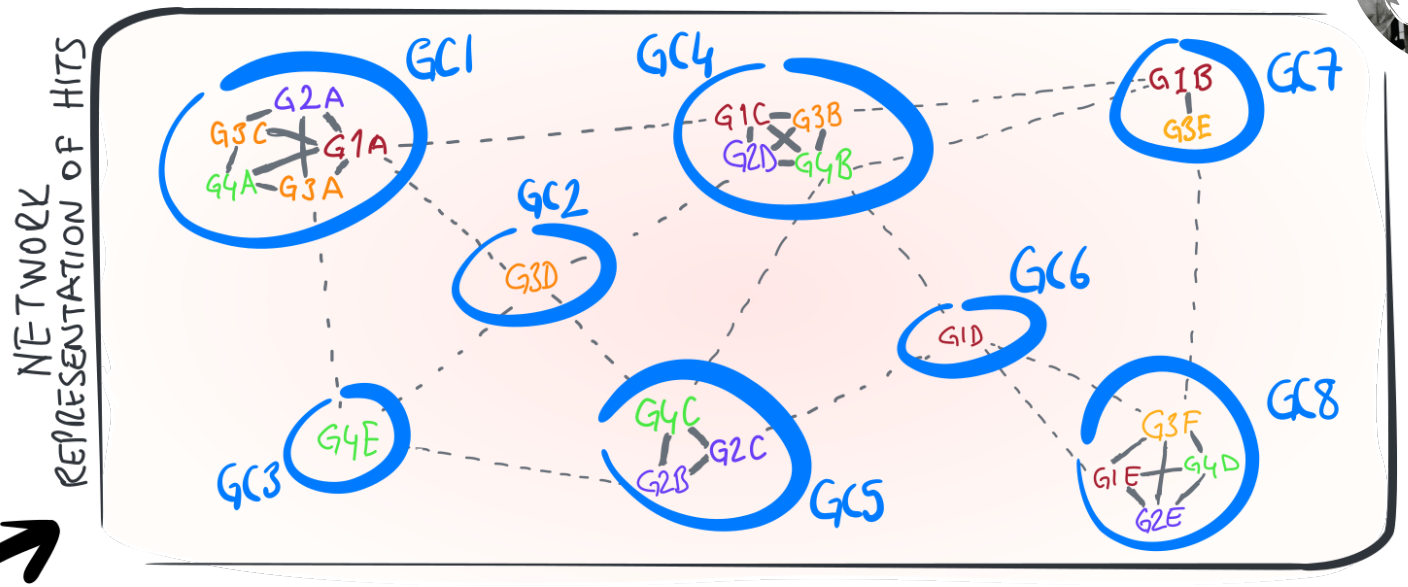
pan-genomics

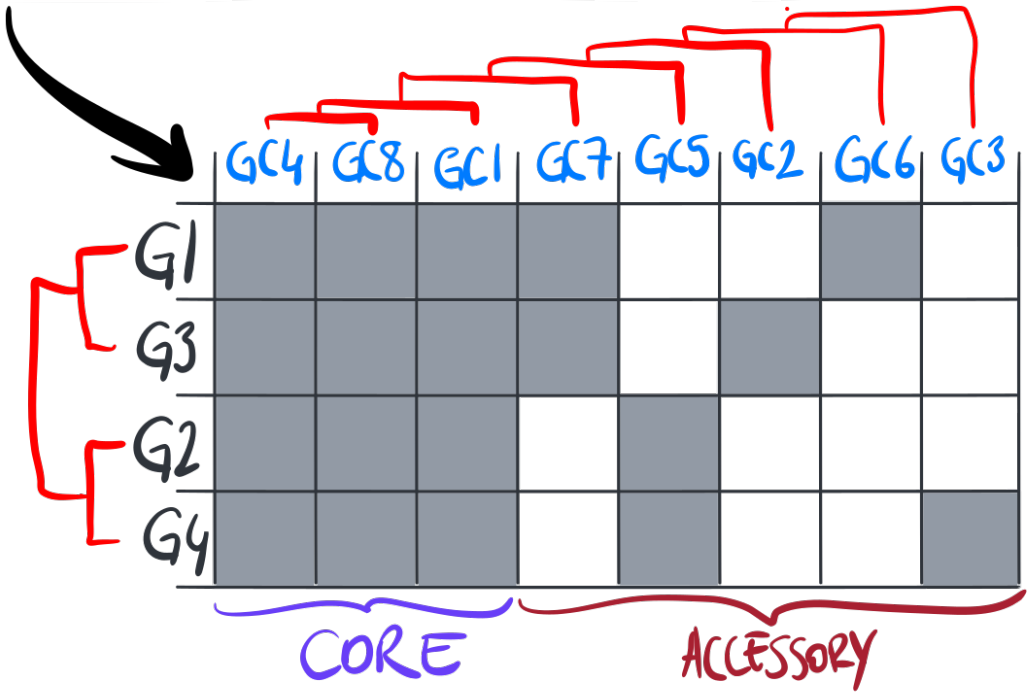




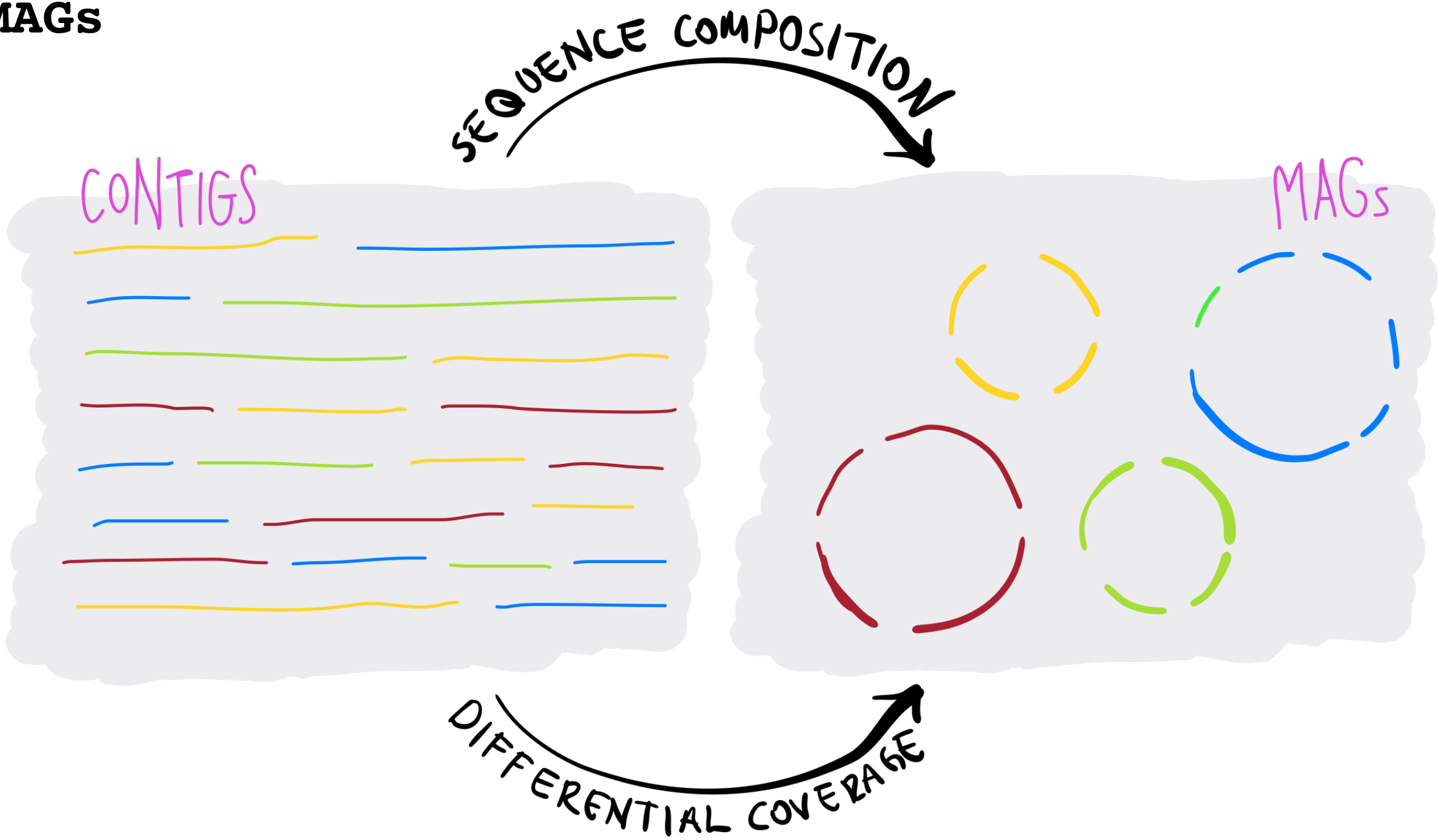








MAGs


SEQUENCE COMPOSITION
CONTIGS
MAGs
DIFFERENTIAL COVERAGE

Human microbiome

Cell

CIBIO

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Graphical Abstract

EBI

Article | Open Access | Published: 11 February 2019

A new genomic blueprint of the human gut microbiota

Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn

Nature **568**, 499–504 (2019) | Cite this article

HBC, CGR

Article | Open Access | Published: 13 March 2019

New insights from uncultivated genomes of the global human gut microbiome

JGI

Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard & Nikos C. Kyrpides

Nature **568**, 505–510 (2019) | Cite this article

Forster, S. C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).

Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).

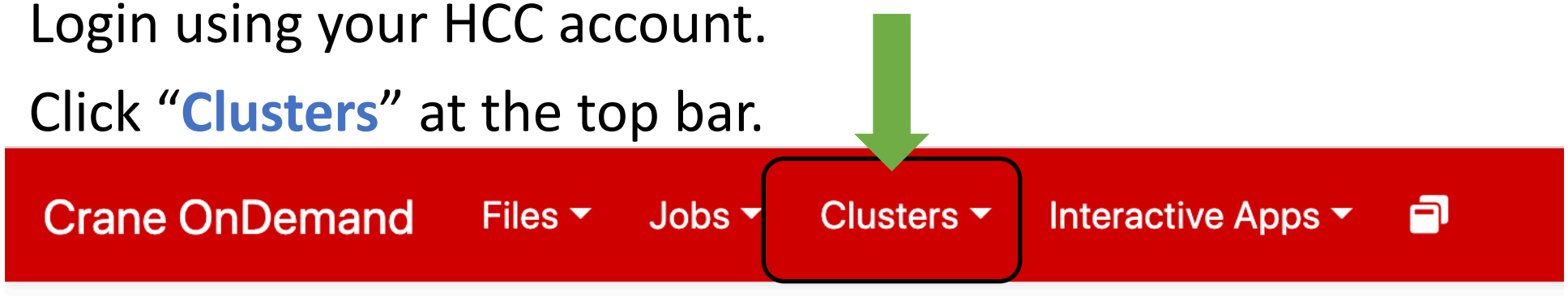
Schedule (June 20 and 23, FIC 111)

Time	Presenter	Topic
12:00-12:30pm	Yanbin	Introduction to the workshop & project
12:30-1:00pm	Tang	UHGG Bacteroides MAGs & stat visualization
1:00-1:45pm	Yuchen	CAZyme & CGC annotation & stat visualization
1:45-2:30pm	Yuchen	Anvio pan-genome analysis & phylogeny visualization
2:30-2:45pm		Break
2:45-3:30pm	Bowen	Pfam annotation & protein seq clustering of nSGs & visualization
3:30-4:15pm	Tang	Glycan substrate prediction & CGC clustering & network visualization
4:15-5:00pm	Jinfang	ML prediction for new GH/PL from conserved nSGs

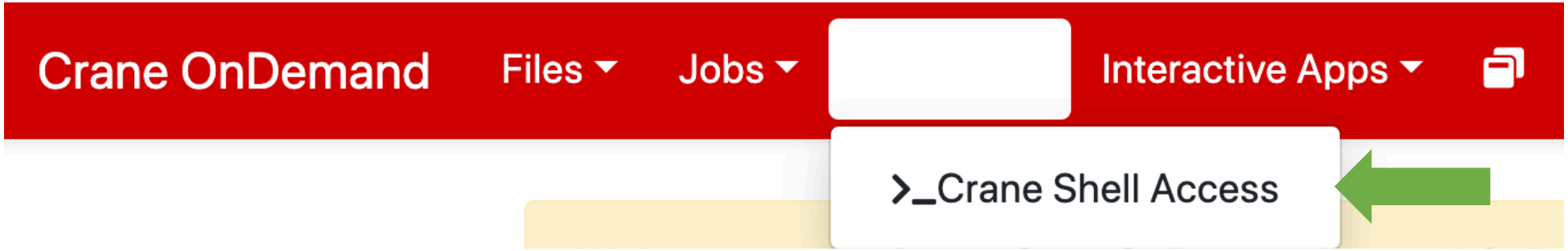
How to Use Crane OnDemand for the Workshop ?

❖Crane OnDemand

1. Click or copy the link to your browser: https://crane-ood.unl.edu/pun/sys/dashboard
2. Login using your HCC account.
3. Click "**Clusters**" at the top bar.



4. Select and click the "**_Crane Shell Access**"



❖Copy folder and open tutorial HTML

Step 1: Change directory from HOME to WORK: cd $WORK

```
[tangli@login.crane ~]$ cd $WORK
```

Step 2: Copy all materials to your working directory:

```
cp -r /work/yinlab/tangli/workshop_2022 . &
cp -r /work/yinlab/tangli/workshop_2022_backup . &
```

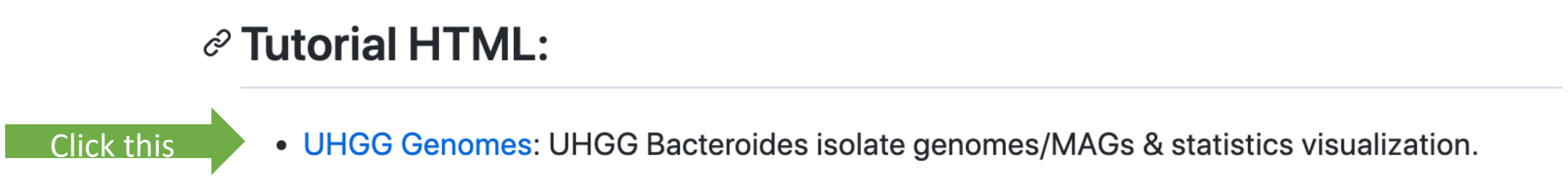
If failed, try:

```
cp -r /tmp/workshop_2022 . &
cp /tmp/workshop_2022_backup.tar.gz . &
tar -xvf workshop_2022_backup.tar.gz &
```

❖Copy folder and open tutorial HTML

Step 3: Click or copy this link to your browser to open GitHub page: https://github.com/tli14/Workshop_2022_YinLab.

Step 4: Open tutorial HTML file in GitHub page.



Step 5: Follow the tutorial HTML. Copy and paste each cmd to your "Crane shell" and run it.

UHGG Genomes

1. Select Bacteroides genomes from UHGG

In [1]:
```
# Change your directory to UHGG_genomes folder
cd $WORK/workshop_2022/UHGG_genomes
```

In [2]:
```
# Check your working directory
pwd
```

```
/work/yinlab/tangli/workshop_2022/UHGG_genomes
```

In [3]:
```
# Select all genomes from Bacteroides genus.
cat genomes-all_metadata.tsv | grep "s__Bacteroides" > Bacteroides_metadata.tsv
```

In [4]:
```
# Covert dos to unix format.
dos2unix Bacteroides_metadata.tsv
```

```
dos2unix: converting file Bacteroides_metadata.tsv to Unix format...
```

In [5]:
```
# Select representative genome ids.
cat genomes-all_metadata.tsv | grep "s__Bacteroides" | cut -f 14 | sort | uniq > spe_list.txt
```

In [6]:
```
# Select representative genomes from file.
python extract_info.py Bacteroides_metadata.tsv spe_list.txt
```

❑ Part 1

UHGG Genomes

Unified Human Gastrointestinal Genome (UHGG)

- A total of 289,232 prokaryotic genomes from the human gut microbiome were clustered into 4,744 species representatives.

MGnify Overview Submit data Text search Sequence search Browse data API About Help Login

Biome	Accession	Length	Num. of genomes	Completeness	Contamination	Type	Taxonomy	Last Updated
	MGYG000000001	3219617	4	98.59	0.7	Isolate	GCA-900066495 sp902362365	12/7/2021
	MGYG000000002	4433109	359	99.37	0	Isolate	Blautia_A faecis	12/7/2021
	MGYG000000003	3229518	1181	100	0	Isolate	Alistipes shahii	12/7/2021
	MGYG000000004	3698896	25	98.66	0.22	Isolate	Anaerotruncus colihominis	12/7/2021
	MGYG000000005	3930428	2	99.3	0	Isolate	Terrisporobacter glycolicus_A	12/7/2021

Unified Human Gastrointestinal Genome (UHGG)

- The "genomes-all_metadata.tsv" was downloaded from this web link: http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0

- **Bacteroides:** one of the most important genus in human gut microbiome; CAZyme and PULs are well-studied for this genus.

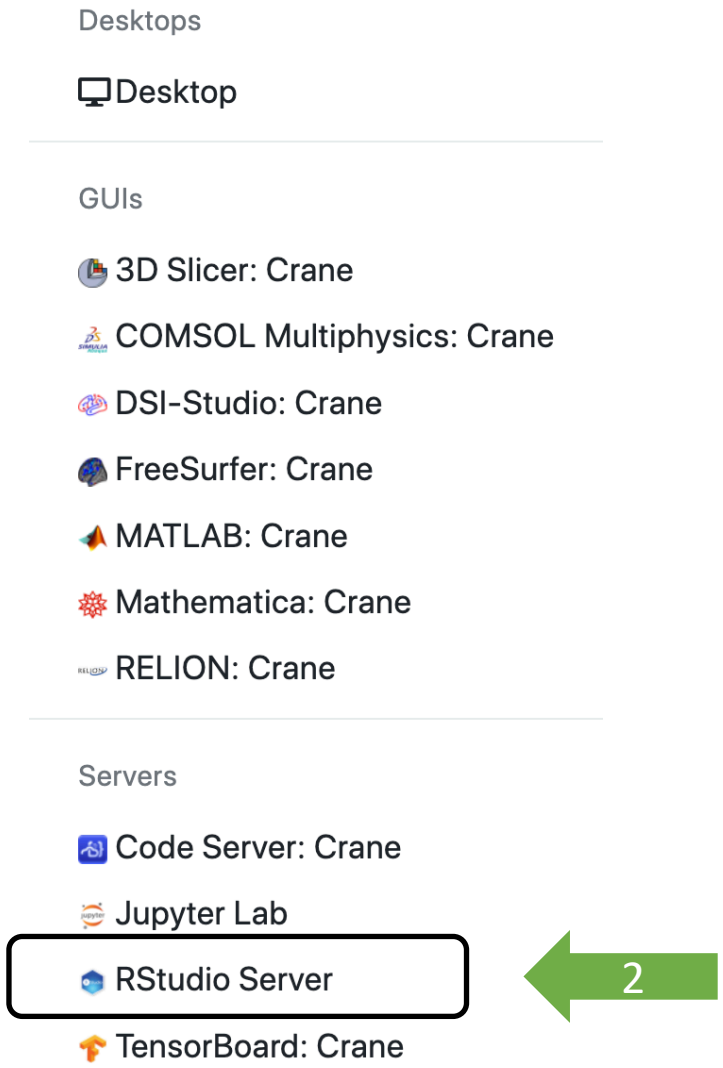
❖ 45 Bacteroides genomes from UHGG

1. Select representative species genomes in Bacteroides genus (n=45).
2. Download genomes.
3. Collect metadata information: genome id, species name, genome types, genome length, protein count, contig N50, completeness, country and continent.
4. Visualize metadata information for 45 genomes.

Use Crane OnDemand Rstudio server:

1. Click "**Interactive Apps**" at the top bar.

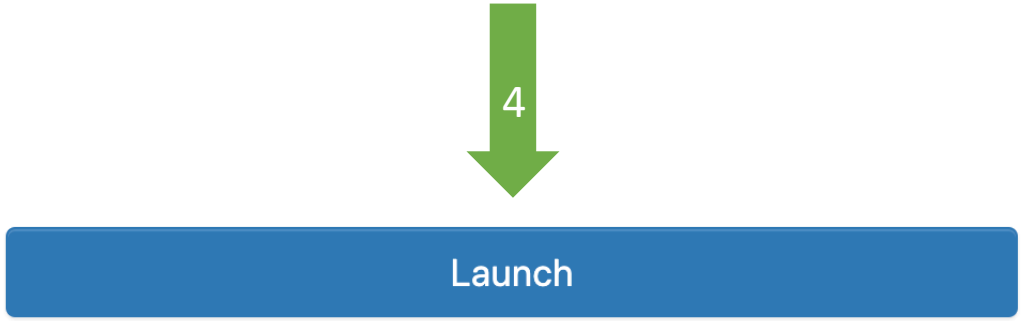
2. Click "**Rstudio Server**" to start the Rstudio.

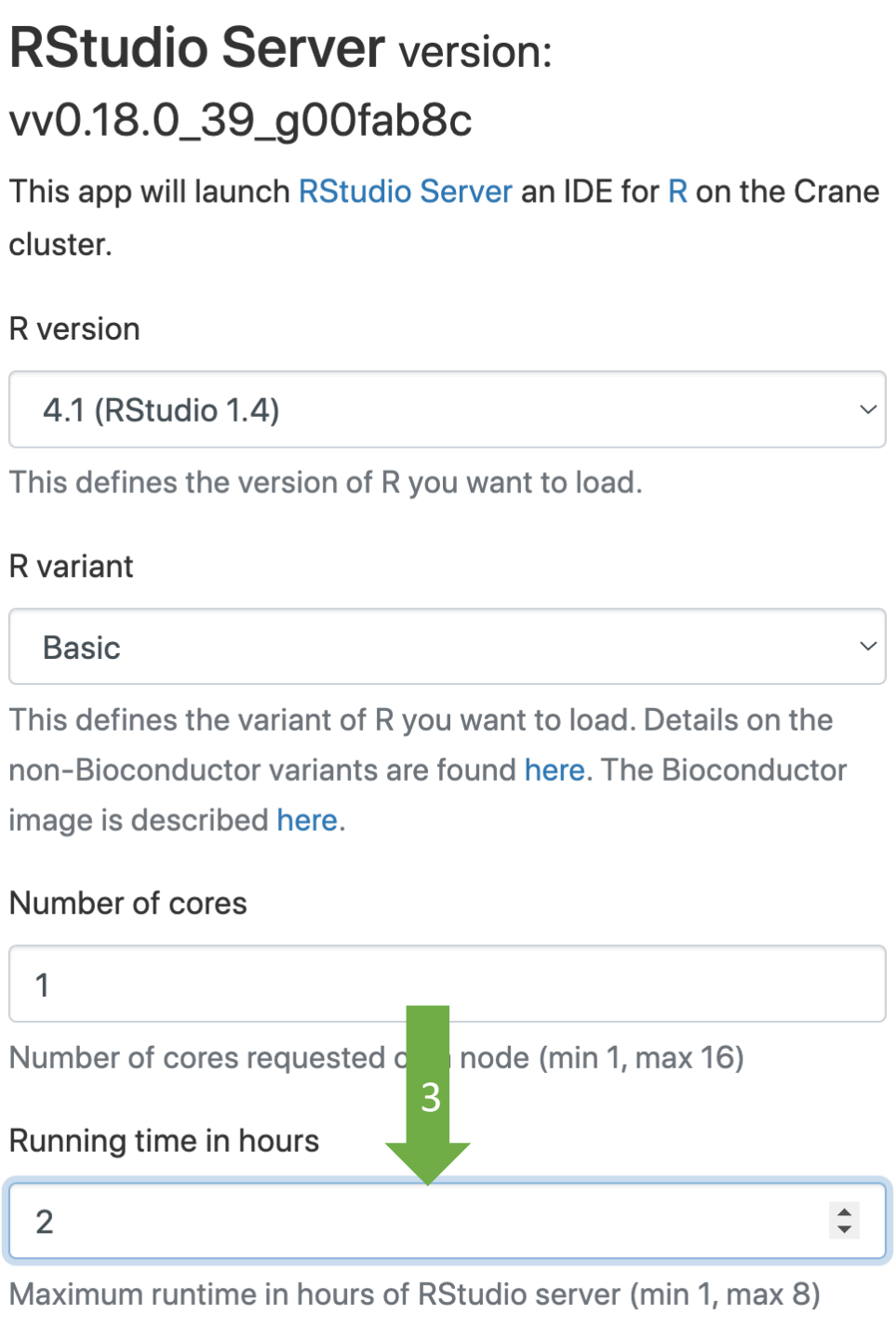


Use Crane OnDemand Rstudio server:

3. Set running time to **1 or 2 hrs**.

4. Click "**Launch**" at the bottom of the page.





RStudio Server (41599959) 1 node | 1 core | Running

Host: c3706.crane.hcc.unl.edu

Delete

Created at: 2022-06-16 11:36:46 CDT

Time Remaining: 23 minutes

Session ID: ea8c9741-301d-4848-8e88-b31e99087717

Importing modules with R

Import module function

```
source(file.path(Sys.getenv("LMOD_PKG"), "init/R"))
```

Load system module

```
module("load", "blast")
```

Load custom module

```
module("use", "~/local/share/lmodfiles")
module("load", "MyModuleA")
```

® Login to RStudio Server

6

6. Click to login.

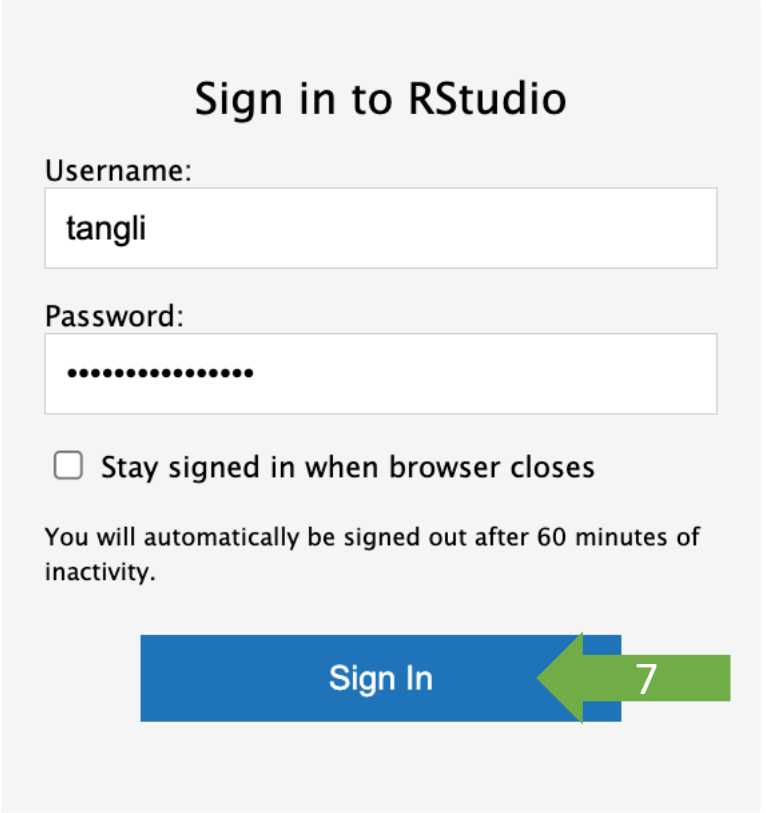
To login, use the following credentials:

Username: tangli

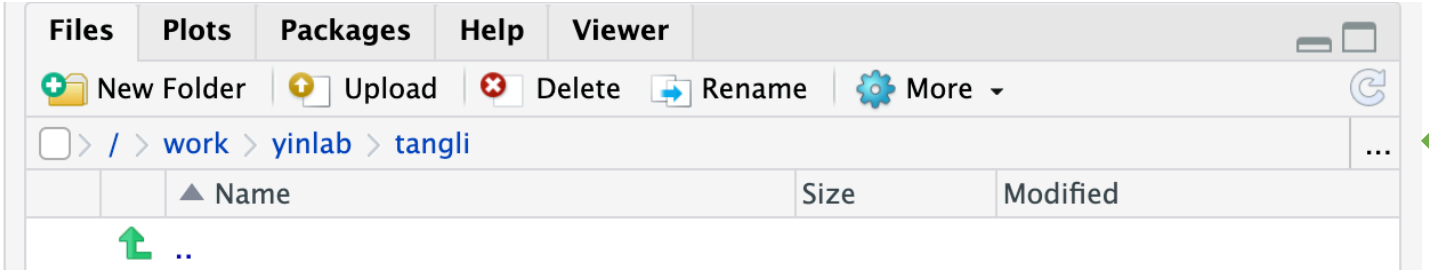
Password: Click to copy

5

5. Click to copy the password.

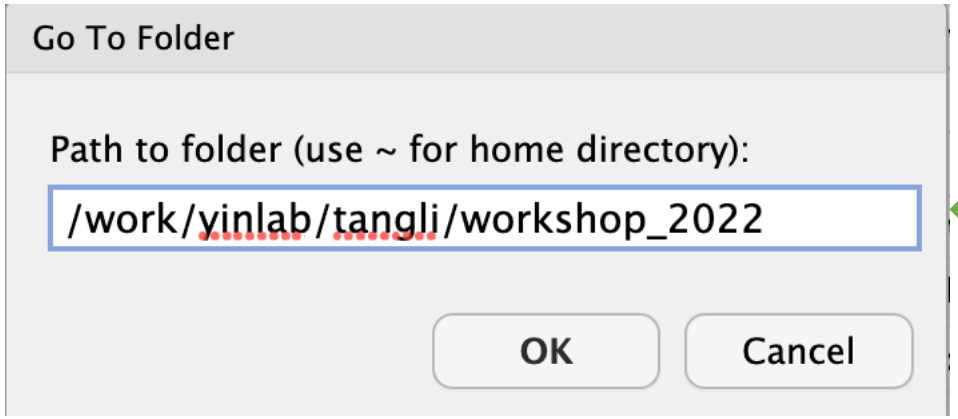


7. Login using username and copied password.



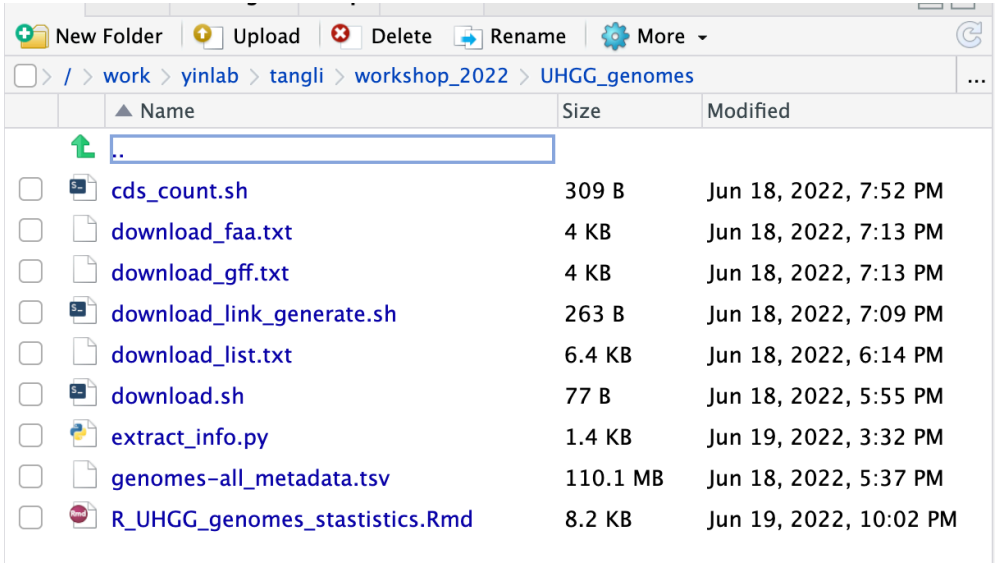
8

8.Click "..."to change the directory



9

9. Enter your working directory path: /work/your_group/your_name/workshop_2022

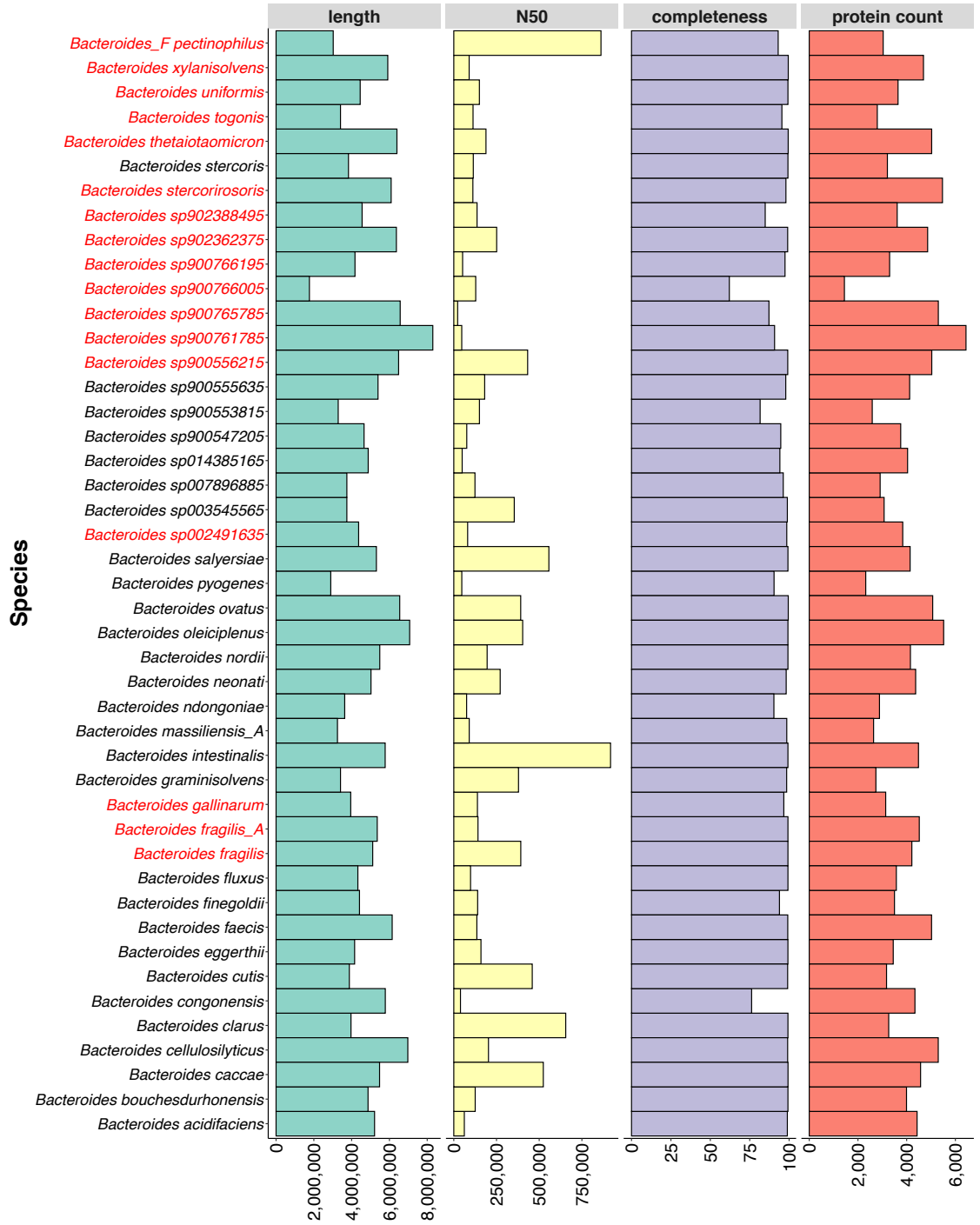


10

10. Open "UHGG_genomes / R_UHGG_genomes_stastistics.Rmd"

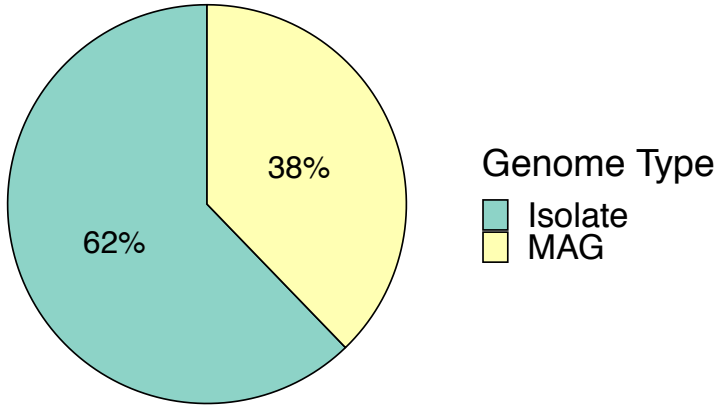
❖ The Overall Statistics of 45 genomes

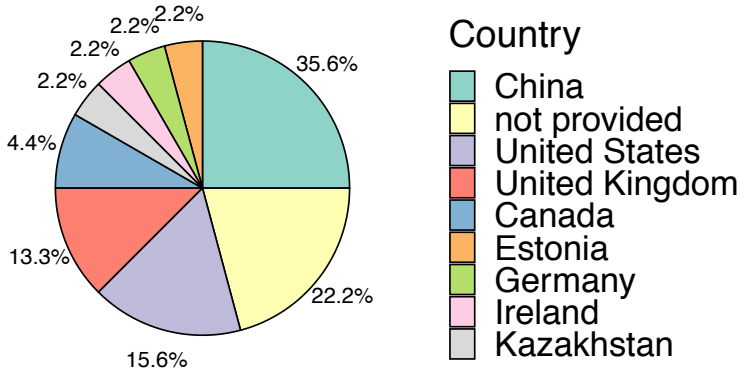
- Genomes: 17 MAGs (red) and 28 isolates (back).
- Average Length: ~4.87 Mb.
- Completeness: 99.46% - 62.07% (only 5 genomes < 90%)
- Protein count: 6,436 - 1,444.

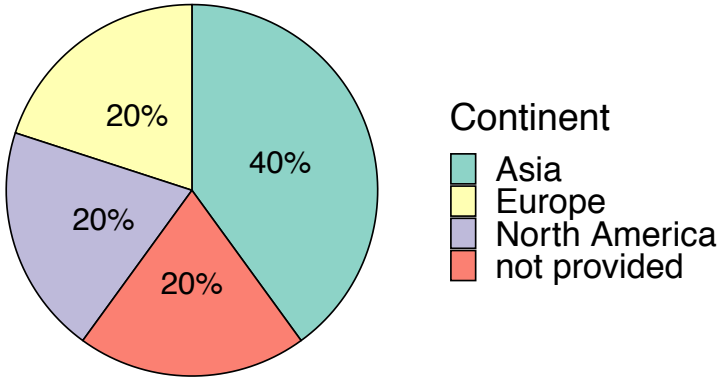


❖ The Overall Statistics of 45 genomes

- More isolate genomes than MAGs.
- Most are from China, United States and United Kingdom.
- Mainly from Asia.

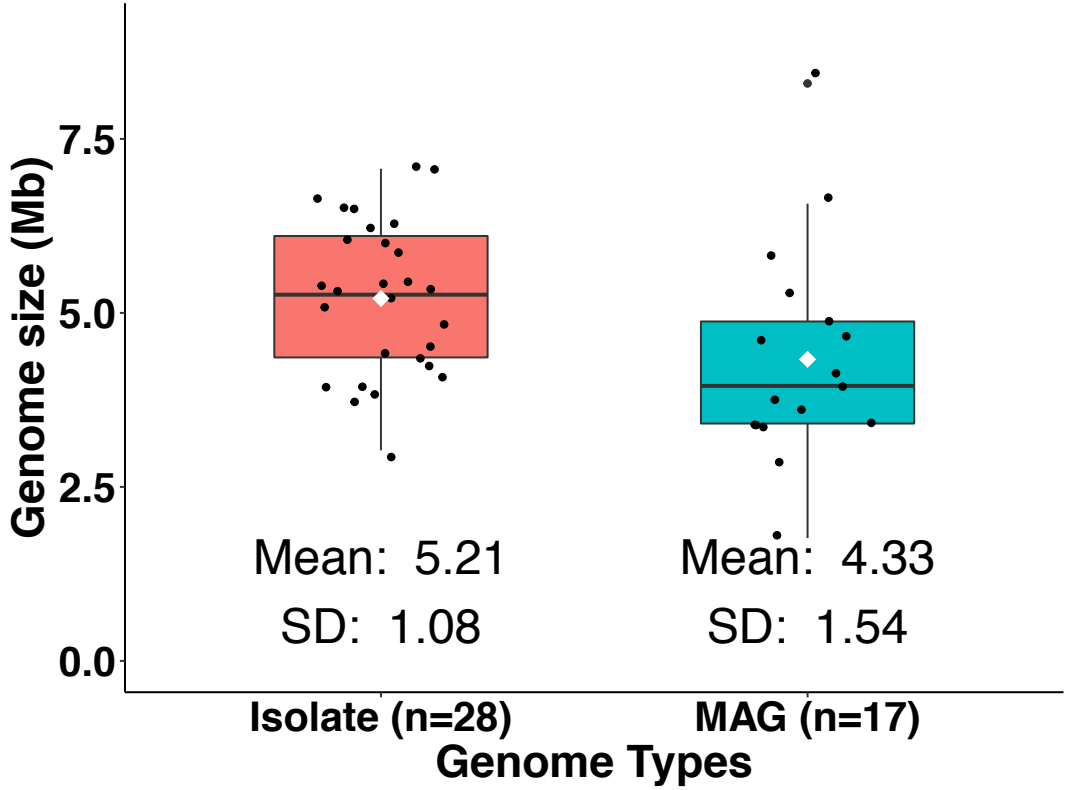


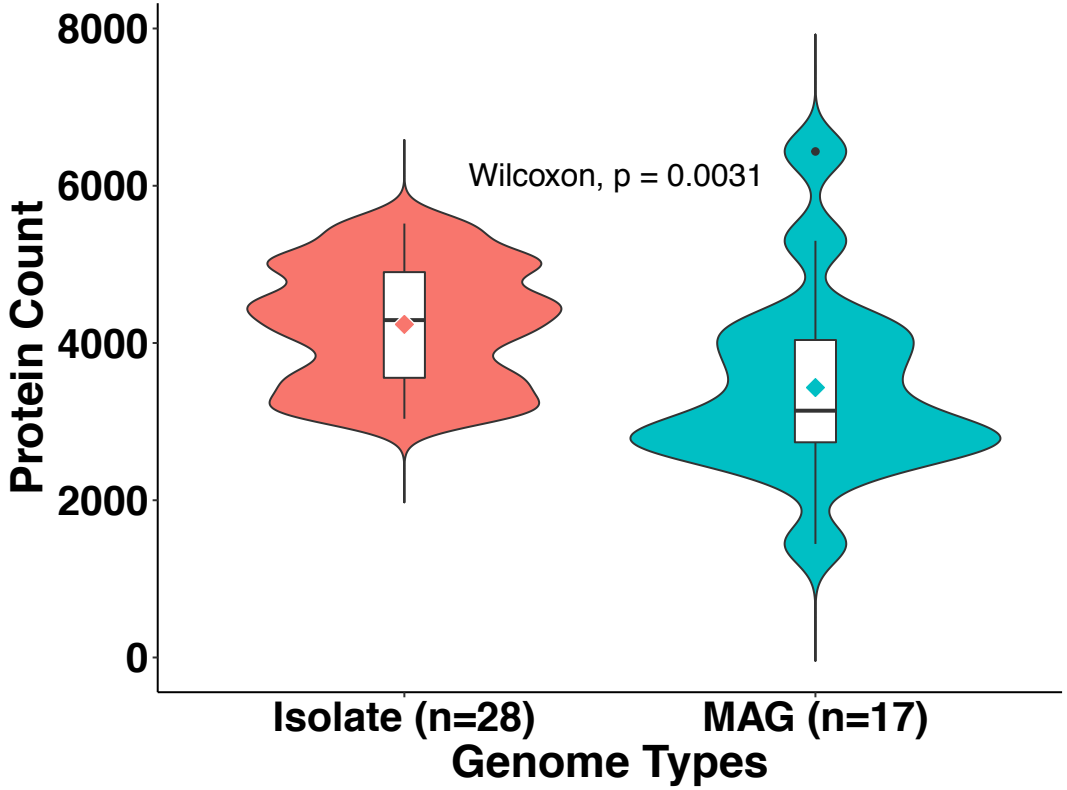




❖ Group comparison for 45 genomes.

- Isolate genomes have larger average genome size and more proteins than MAGs.





❑ Part 2

CAZyme and CGC analyses

❖ Run dbCAN3 for CAZyme and CGC annotation

Github: https://github.com/linnabrown/run_dbcan
Online server: https://bcb.unl.edu/dbCAN2/

- Install and run on terminal or submit sequence to online server

- dbCAN3 output files

```
cgc.gff
cgc.out
cgc_standard.out
diamond.out
eCAMI.out
hmmer.out
overview.txt
stp.out
tf-1.out
tf-2.out
tp.out
uniInput
```

- cgc_standard.out

CGC#	Gene Type	Contig ID	Protein ID	Gene Start	Gene Stop		Direction	Protein Family
CGC1	CAZyme	MGYG000001313_17	MGYG000001313_00026	8246	10834	-	GH3	
CGC1	TC	MGYG000001313_17	MGYG000001313_00027	11063	12571	-	2.A.50.2.1	
CGC2	TC	MGYG000001313_17	MGYG000001313_00354	371036	371953	-	2.A.4.7.9	
CGC2	null	MGYG000001313_17	MGYG000001313_00355	372128	374287	+	null	
CGC2	null	MGYG000001313_17	MGYG000001313_00356	374329	374808	+	null	
CGC2	CAZyme	MGYG000001313_17	MGYG000001313_00357	374866	377136	-	PL6|PL6_1	
CGC2	null	MGYG000001313_17	MGYG000001313_00358	377167	378744	-	null	
CGC2	TC	MGYG000001313_17	MGYG000001313_00359	378753	380198	-	8.A.46.1.2	
CGC2	TC	MGYG000001313_17	MGYG000001313_00360	380217	383324	-	1.B.14.6.1	
CGC2	null	MGYG000001313_17	MGYG000001313_00361	383604	383711	-	null	
CGC2	CAZyme	MGYG000001313_17	MGYG000001313_00362	383793	385955	+	PL17|PL17_2	
CGC2	null	MGYG000001313_17	MGYG000001313_00363	385987	386331	+	null	
CGC2	TC	MGYG000001313_17	MGYG000001313_00364	386337	387803	+	2.A.1.14.25	
CGC3	TC	MGYG000001313_17	MGYG000001313_00377	400867	401637	+	3.A.1.15.15	
CGC3	null	MGYG000001313_17	MGYG000001313_00378	401795	405139	+	null	
CGC3	CAZyme	MGYG000001313_17	MGYG000001313_00379	405162	405776	+	CE4	
CGC4	CAZyme	MGYG000001313_17	MGYG000001313_00393	416047	417153	-	GH105	

❖ CAZyme and CGC analyses

- Count the number of CAZymes and CGCs in each genome

```
genome_id,taxa,num_genes,num_CAZyme,num_cgc,num_genes_in_cgc,num_cazyme_in_cgc,num_TC_in_cgc,num_TF_in_cgc,
MGYG000000013,Bacteroides sp902362375,4861,459,110,883,348,247,23,25,240,0.094425,0.18165
MGYG000000029,Bacteroides finegoldii,3500,309,76,590,231,161,13,13,172,0.0882857,0.168571
MGYG000000054,Bacteroides acidifaciens,4428,276,67,425,156,129,8,10,122,0.0623306,0.0959801
MGYG000000057,Bacteroides sp002491635,3839,212,58,375,141,112,9,3,110,0.0552227,0.0976817
MGYG000000098,Bacteroides bouchesdurhonensis,3994,233,59,334,118,116,5,9,86,0.0583375,0.0836254
MGYG000000105,Bacteroides clarus,3263,225,59,454,157,139,18,6,134,0.0689549,0.139136
MGYG000000196,Bacteroides thetaiotaomicron,5027,404,96,728,257,208,25,19,219,0.080366,0.144818
```

(taxa_number_count_and_percentage.csv)

- CAZyme family analyses

```
genome_id,taxa,gene_id,CAZyme_family,CAZyme_class
MGYG000000013,Bacteroides sp902362375,MGYG000000013_00026,GH29,GH
MGYG000000013,Bacteroides sp902362375,MGYG000000013_00028,GH16,GH
MGYG000000013,Bacteroides sp902362375,MGYG000000013_00030,GH76,GH
MGYG000000013,Bacteroides sp902362375,MGYG000000013_00037,GH92,GH
MGYG000000013,Bacteroides sp902362375,MGYG000000013_00039,GH76,GH
MGYG000000013,Bacteroides sp902362375,MGYG000000013_00040,GH130,GH
```

(taxa_cazyme_family.csv)

- Visualization by R, go to RStudio!

R Markdown file for making plots:

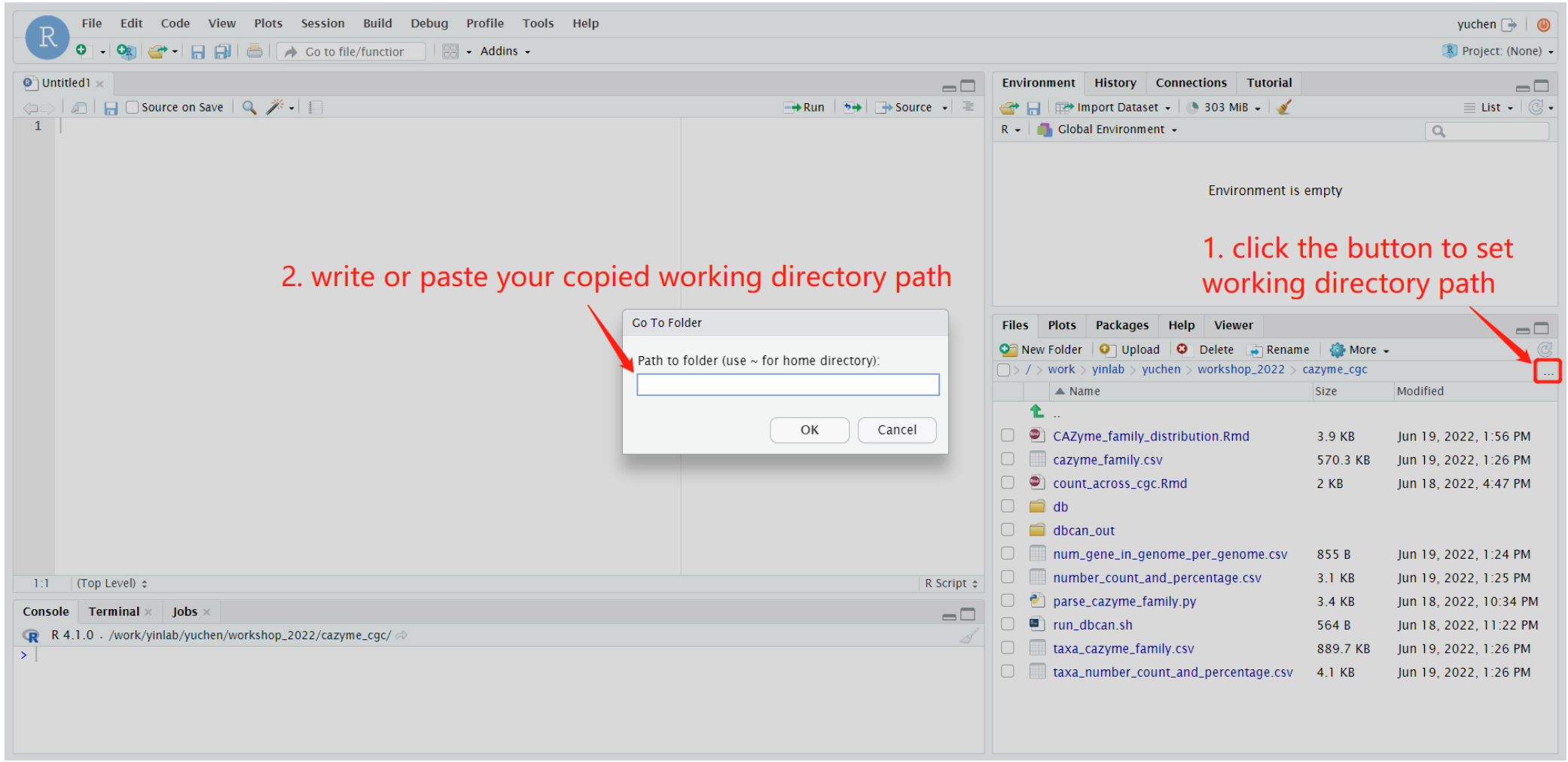
count_across_cgc.Rmd

CAZyme_family_distribution.Rmd

❖ Visualization in RStudio

R Markdown file for making plots:
count_across_cgc.Rmd
CAZyme_family_distribution.Rmd

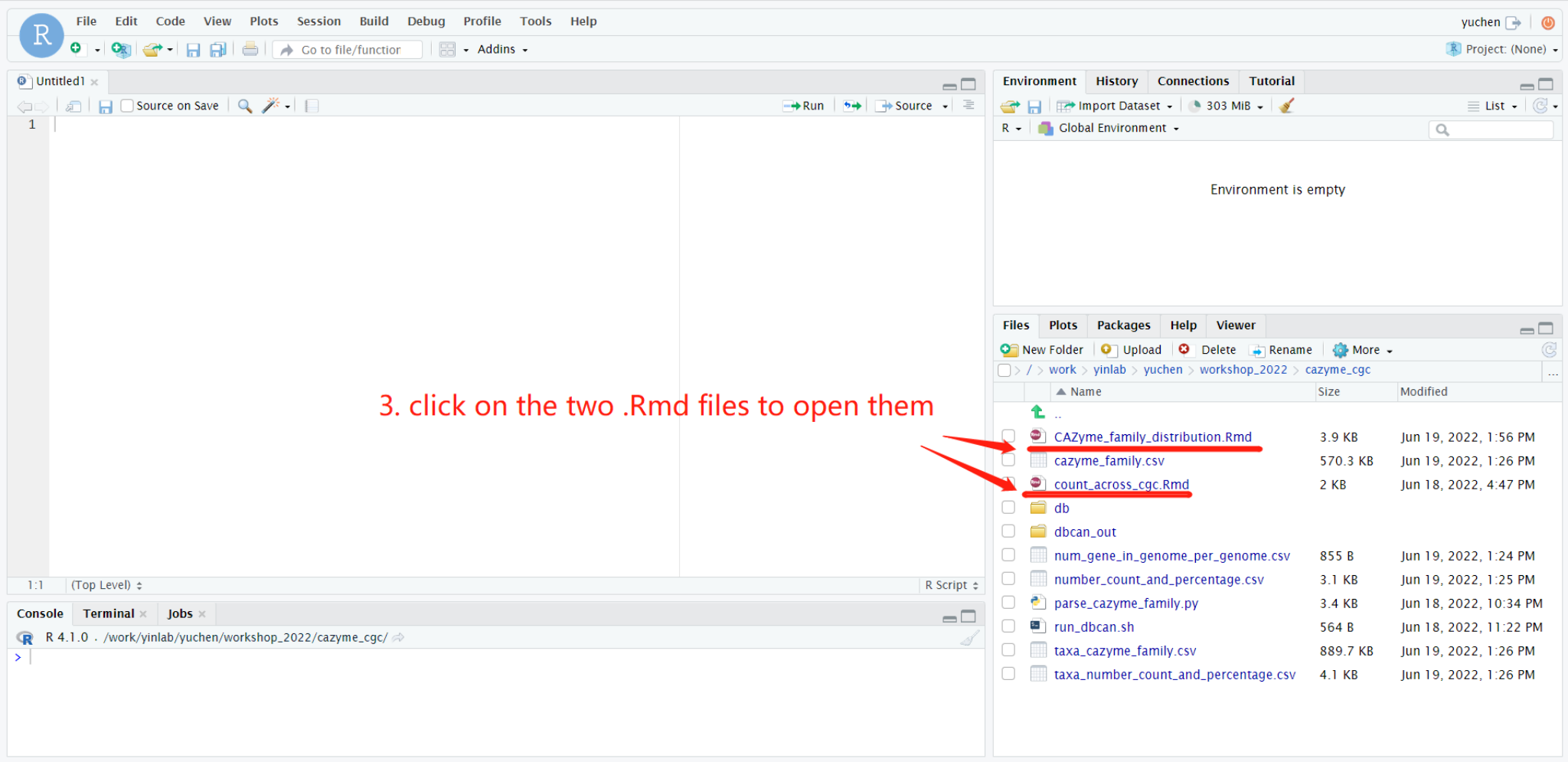
- Set working directory to cazyme_cgc/



❖ Visualization in RStudio

- Set working directory to cazyme_cgc/

R Markdown file for making plots:
count_across_cgc.Rmd
CAZyme_family_distribution.Rmd

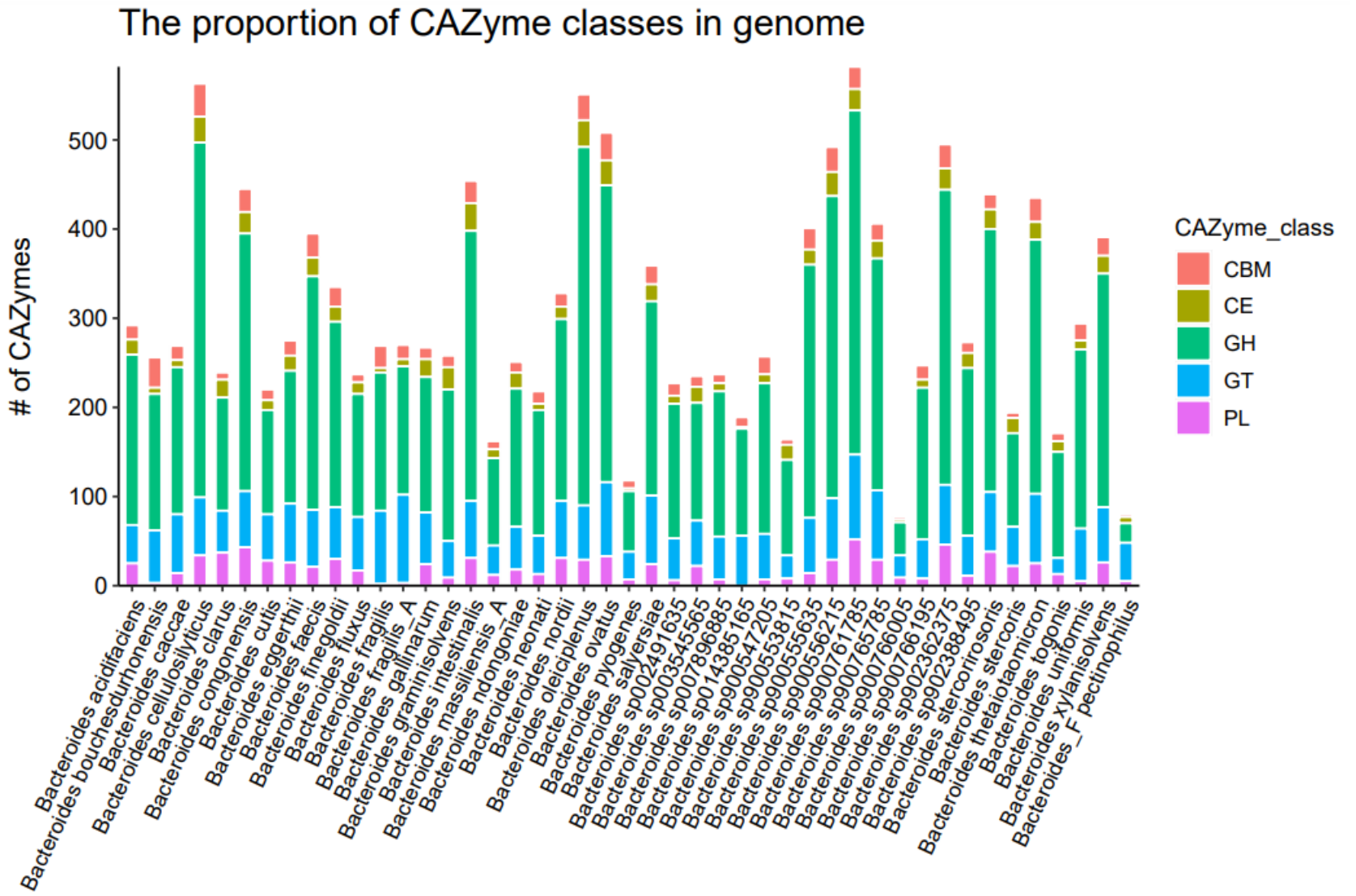


- Plots of CAZyme and CGC analyses

- The proportion of CAZyme classes in genome (cazy_class_stackbar.pdf)

- The total number of CAZymes varies across genomes.

- GH account for the largest proportion in most of genomes

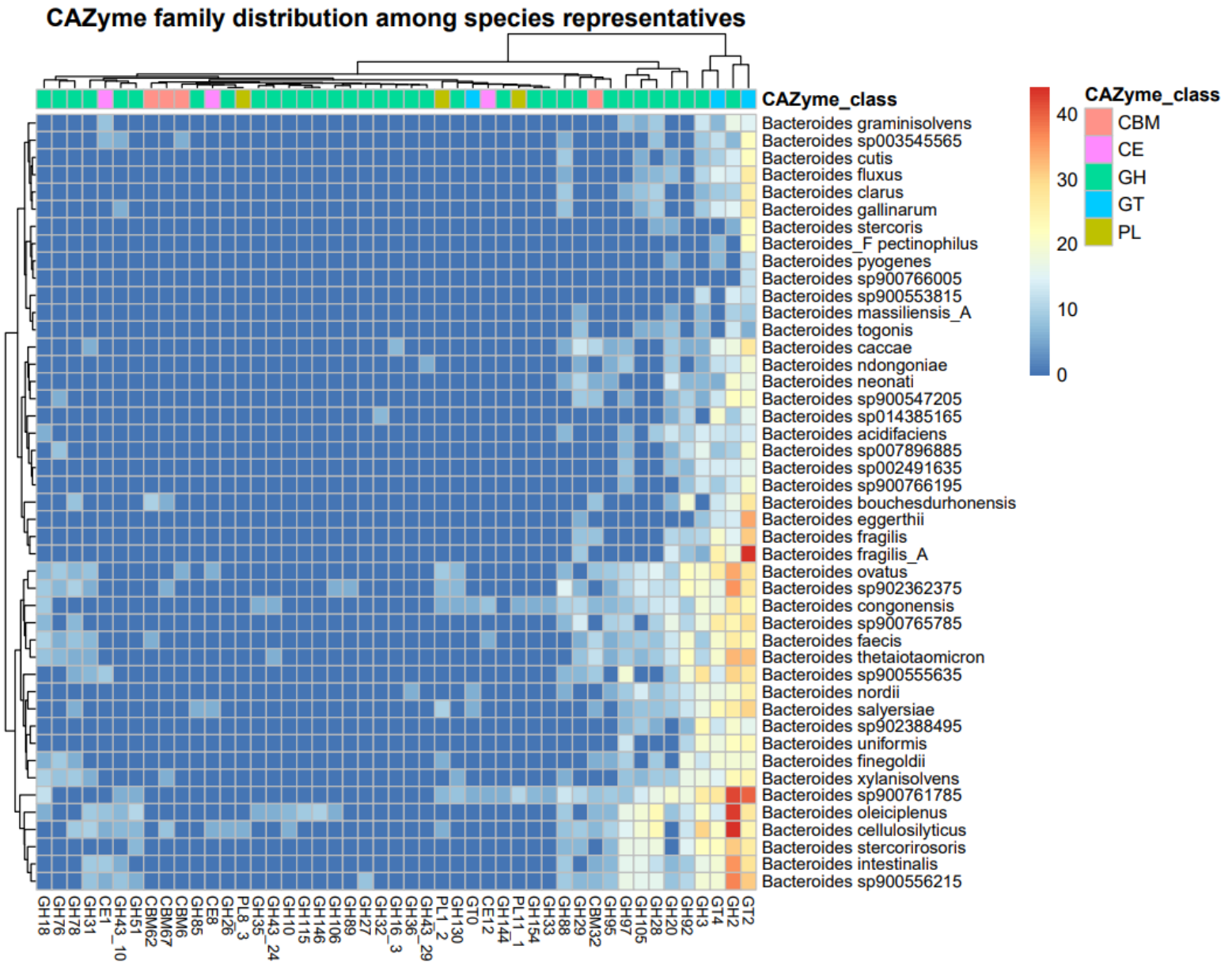


- Plots of CAZyme and CGC analyses

- CAZyme family distribution among spp representatives (cazy_distr_hm.pdf)

- The number of CAZyme families in one genome > 5

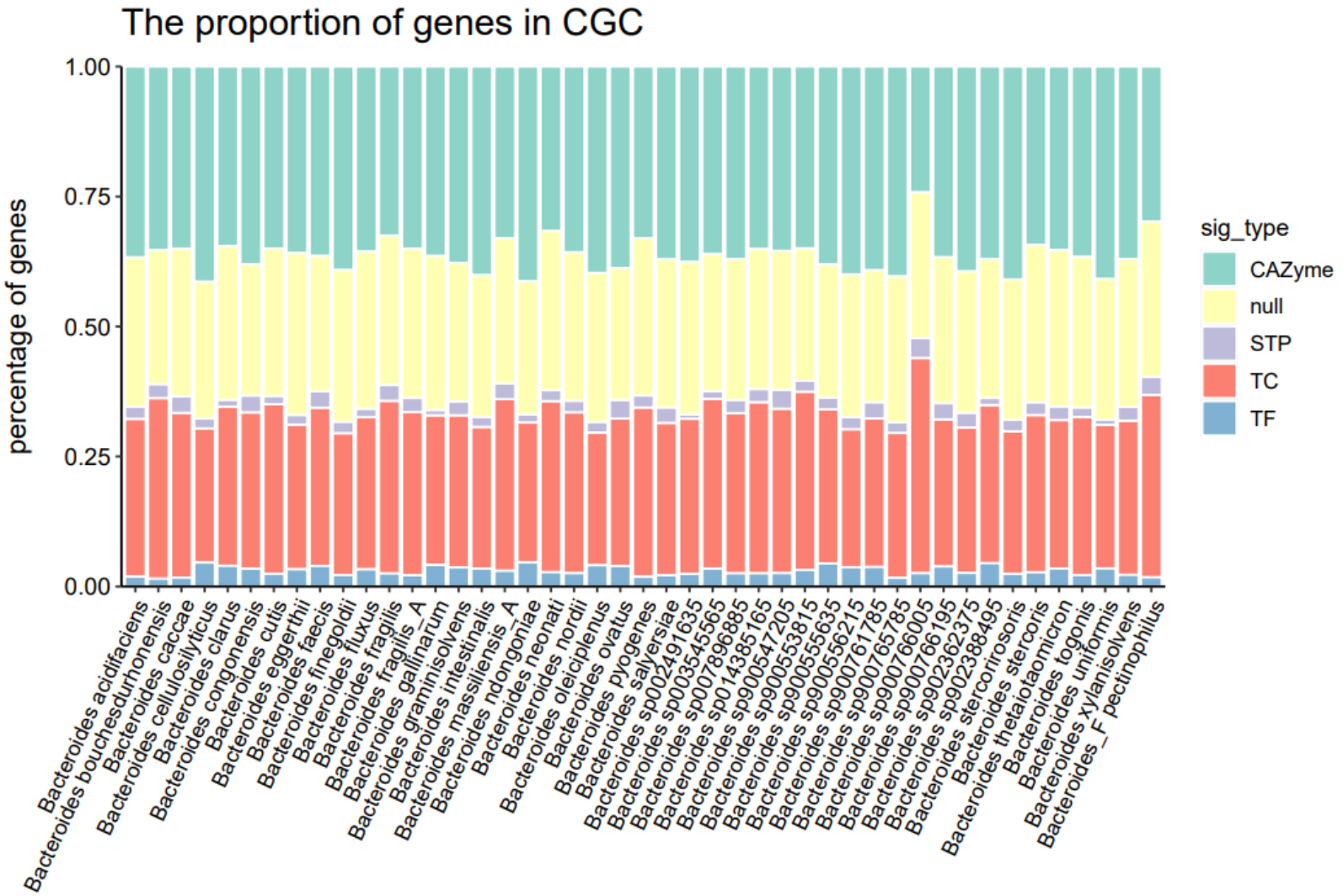
- Some CAZyme e.g. GT2, GH3 are very common and abundant



❖ Plots of CAZyme and CGC analyses

- The proportion of genes in CGC (prop_sig_cgc_stackbar.pdf)

- Null genes account for a relatively large proportion within CGC in each genome



❑ Part 3

Pan-genomics by anvi’o

Phylogeny analysis based on SCG alignment by anvi'o

❖ Pangenomics by anvi'o

- Make anvio contig database and import dbCAN annotation

function_import_table/

```
MGYG000000013_import.CAZymeinCGC.tsv
MGYG000000013_import.CAZyme.tsv
MGYG000000013_import.cgc.tsv
MGYG000000029_import.CAZymeinCGC.tsv
MGYG000000029_import.CAZyme.tsv
MGYG000000029_import.cgc.tsv
```

45_Bacteroides_spp_rep_seq/

```
MGYG000000013_ext_gene_call.txt
MGYG000000013.fna
MGYG000000013.fna.fixed.fa
MGYG000000013.fna.fixed.fa.db
MGYG000000013.fna.fixed.fa.report.txt
```

- Make Pan-genome by anvi'o

Pan_OUT/

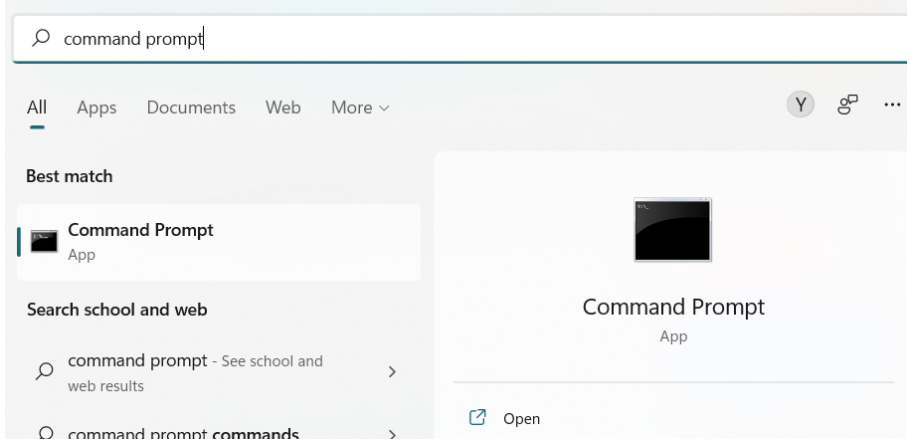
```
combined-aas.fa
combined-aas.fa.unique
combined-aas.fa.unique.dmnd
combined-aas.fa.unique.names
diamond-search-results.txt
diamond-search-results.txt.unique
log.txt
mcl-clusters.txt
mcl-input.txt
UHGG_Bacteroides_Pangenome-PAN.db
```

- Visualize anvi'o pan-genome on the interactive page [open local computer terminal]

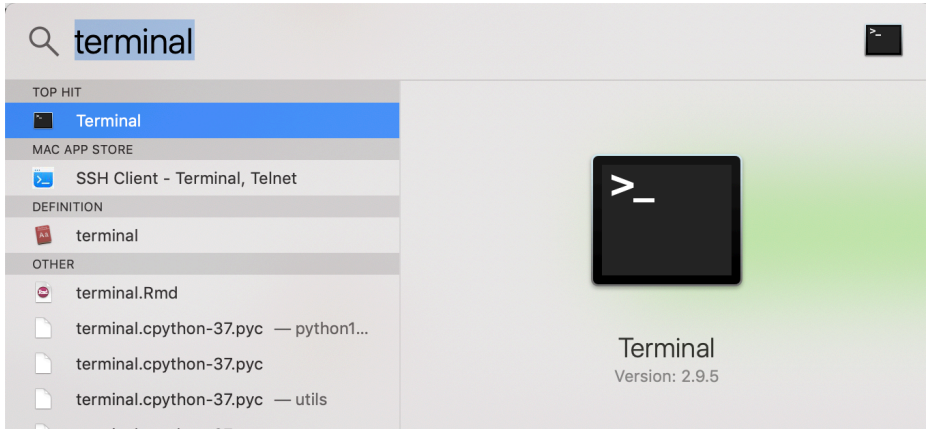
- Visualize Anvi'o Pan-genome on the interactive page

- Find your local computer terminal and open it

Windows: "Command Prompt"



Mac: "Terminal"



- Type in following command lines on your local computer terminal

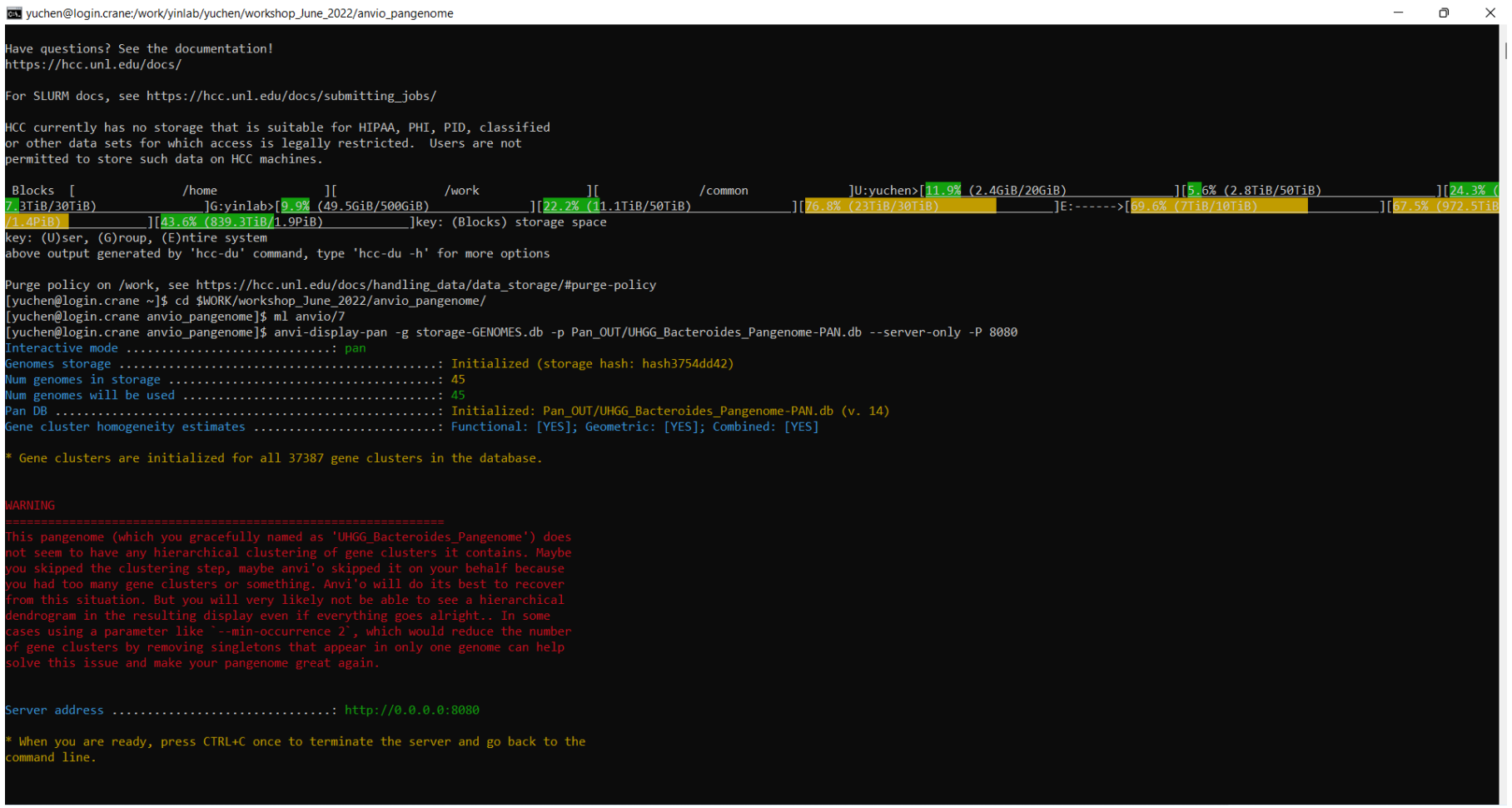
ssh -L 8080:localhost:8080 yourusername@crane.unl.edu
cd $WORK/workshop_2022/pan_genomics/
ml anvio/7
anvi-display-pan -g storage-GENOMES.db -p Pan_OUT/UHGG_Bacteroides_Pangenome-PAN.db --server-only -P 8080

- Visualize Anvi’o Pan-genome on the interactive page

- If succeed, your terminal should look like the picture.

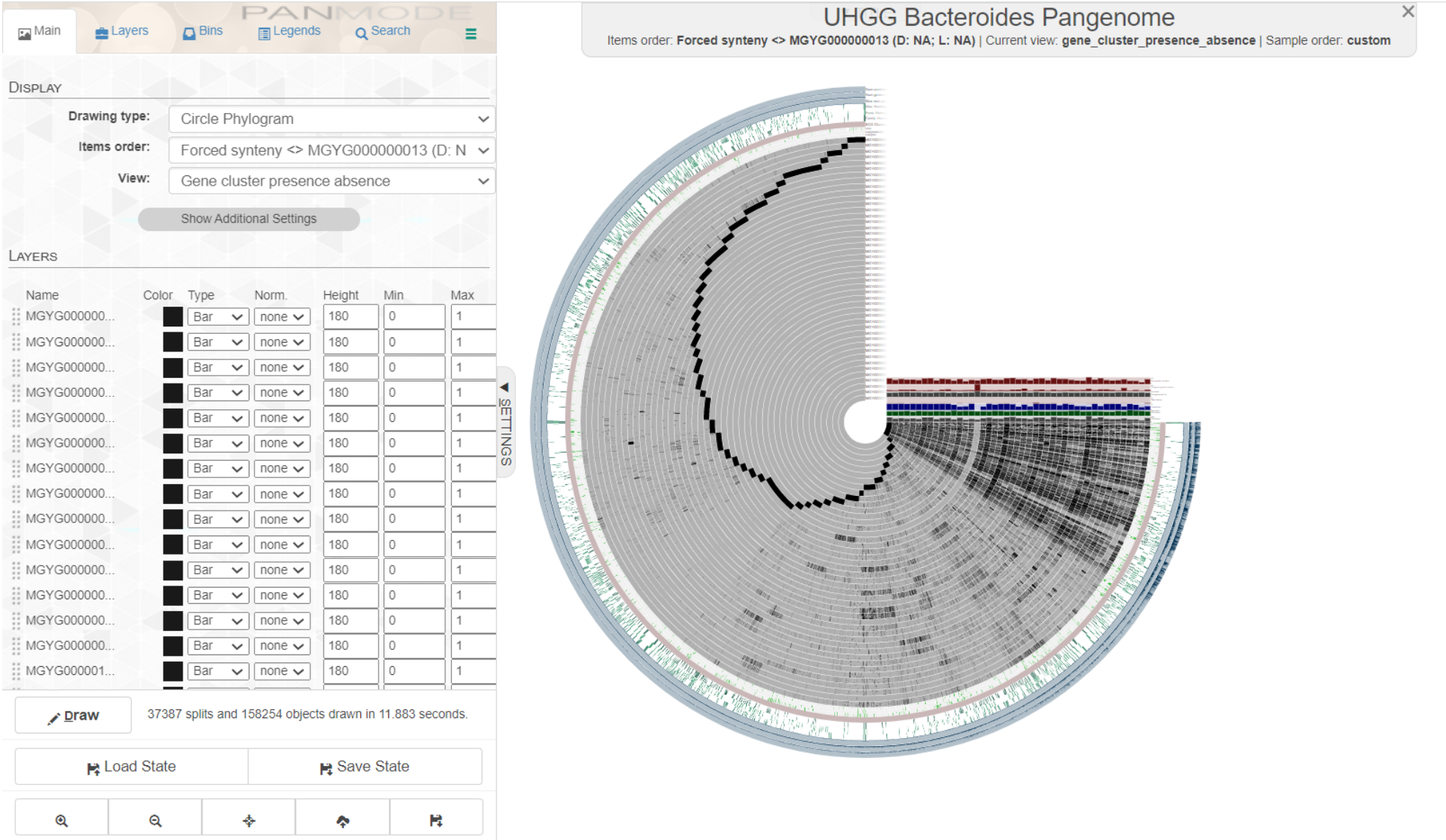
- The server port 8080 may be occupied by others sometimes, just try 8079, 8078, ……

- Then open your web browser, go to the web address: http://localhost:8080



- Visualize Anvi'o Pan-genome on the interactive page

http://localhost:8080



❖ Phylogeny analysis based on SCG alignment by anvi'o

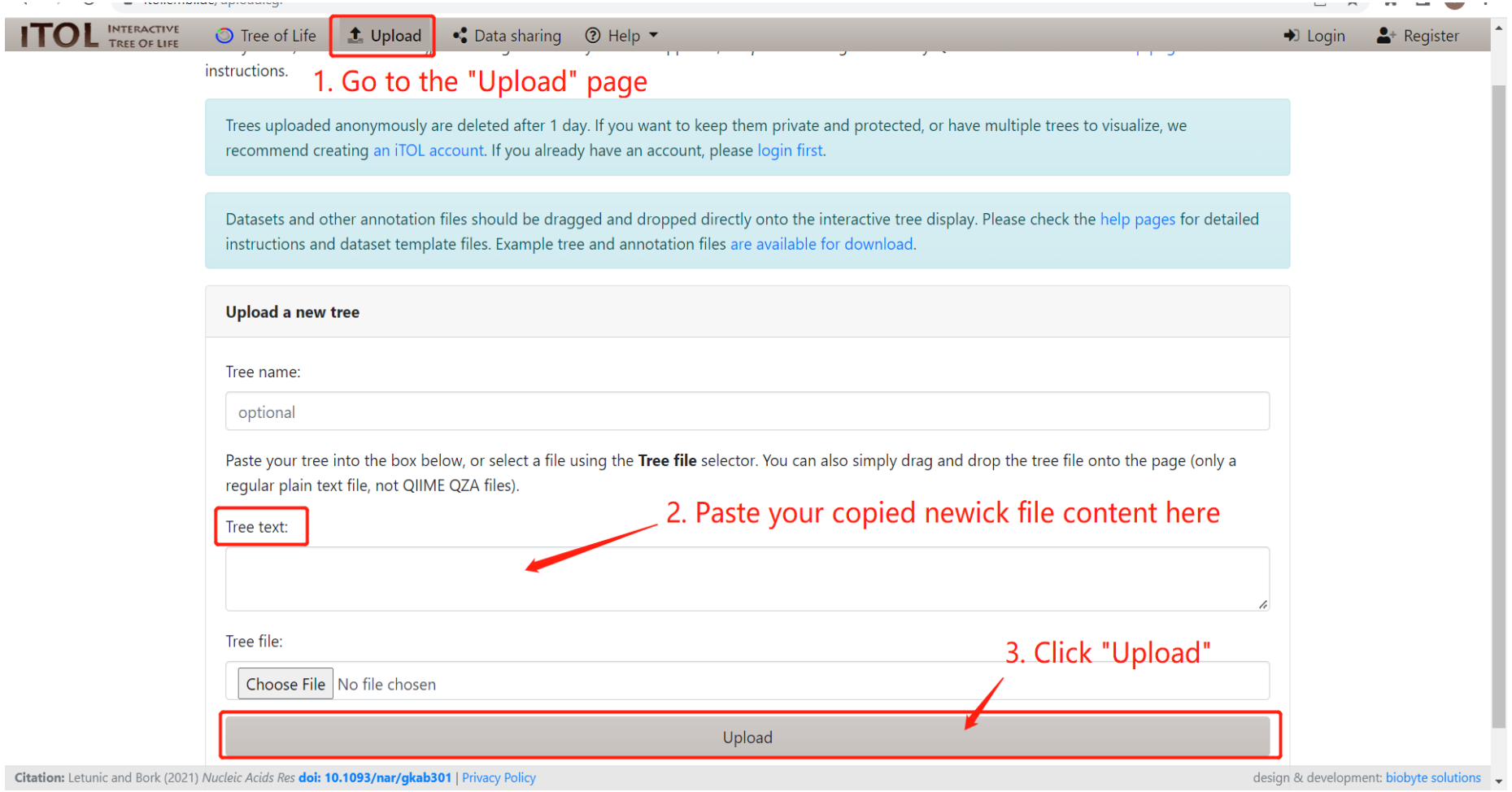
- Go back to HCC terminal
- We will use anvi’o to generate a newick.txt for phylogenetic tree

```
(run_dbcan) [yuchen@login.crane pan_genomics]$ cat newick.txt
(MGYG000002549:0.015023773,((MGYG000003312:0.021494242,(MGYG000000196:0.008507761,MGYG000002281:0.009175631)0.999:0.012779289)0.996:0.009641698,(MGYG000001345:0.003831312,(MGYG000000013:0.003602951,MGYG000001378:0.002567293)0.984:0.005407553)0.941:0.004537657)0.943:0.003923698,(((MGYG000004676:0.051835106,(MGYG000000098:0.034560858,(MGYG000000788:0.071961528,(((((MGYG000002455:0.002640053,(MGYG000000211:0.006274879,MGYG000002470:0.004358557)0.834:0.002597252)0.994:0.008482269,(MGYG000001422:0.009183000,MGYG000002273:0.010145464)0.749:0.004064178)1.000:0.037045714,((MGYG000002621:0.043142922,(MGYG000004185:0.069371549,(MGYG000000652:0.025559704,MGYG000004019:0.028981078)1.000:0.024859256)0.446:0.013813632)1.000:0.036605579,((MGYG000002300:0.022813087,(MGYG000001661:0.021020557,(MGYG000001313:0.014841693,(MGYG000000224:0.009268363,(MGYG000000105:0.004949125,MGYG000003681:0.011658614)0.756:0.002647424)0.925:0.004490447)0.803:0.003743051)0.992:0.010010838)1.000:0.016094775,((MGYG000002561:0.029287662,MGYG000004899:0.026386030)1.000:0.027768028,(MGYG000001370:0.029671377,(MGYG000000057:0.011723391,MGYG000001346:0.008485561)1.000:0.021607571)0.361:0.003765829)0.855:0.008243670)0.336:0.006981293)0.988:0.014388175)0.816:0.013028626,((MGYG000003363:0.042999564,MGYG000003367:0.070880730)0.542:0.019528936,MGYG000001314:1.014654859)0.912:0.028759842)1.000:0.044405561,((MGYG000003922:0.042838532,(MGYG000000265:0.016956626,MGYG000001433:0.025507493)0.974:0.008483314)0.954:0.009441221,((MGYG000001461:0.066088533,MGYG000003064:0.140523921)0.612:0.018415075,(MGYG000000236:0.009649412,MGYG000001337:0.004815503)1.000:0.021025872)1.000:0.029198831)0.999:0.020251736)0.999:0.021705428)0.997:0.016878840)0.992:0.011869038)0.177:0.004451508,MGYG000000029:0.016381082)0.946:0.005172100,((MGYG000000675:0.005789374,MGYG000003252:0.004341234)0.983:0.004701886,(MGYG000000054:0.007425716,MGYG000003351:0.002118179)0.955:0.002922030)0.999:0.008351133)0.891:0.003050315);
(run_dbcan) [yuchen@login.crane pan_genomics]$
```

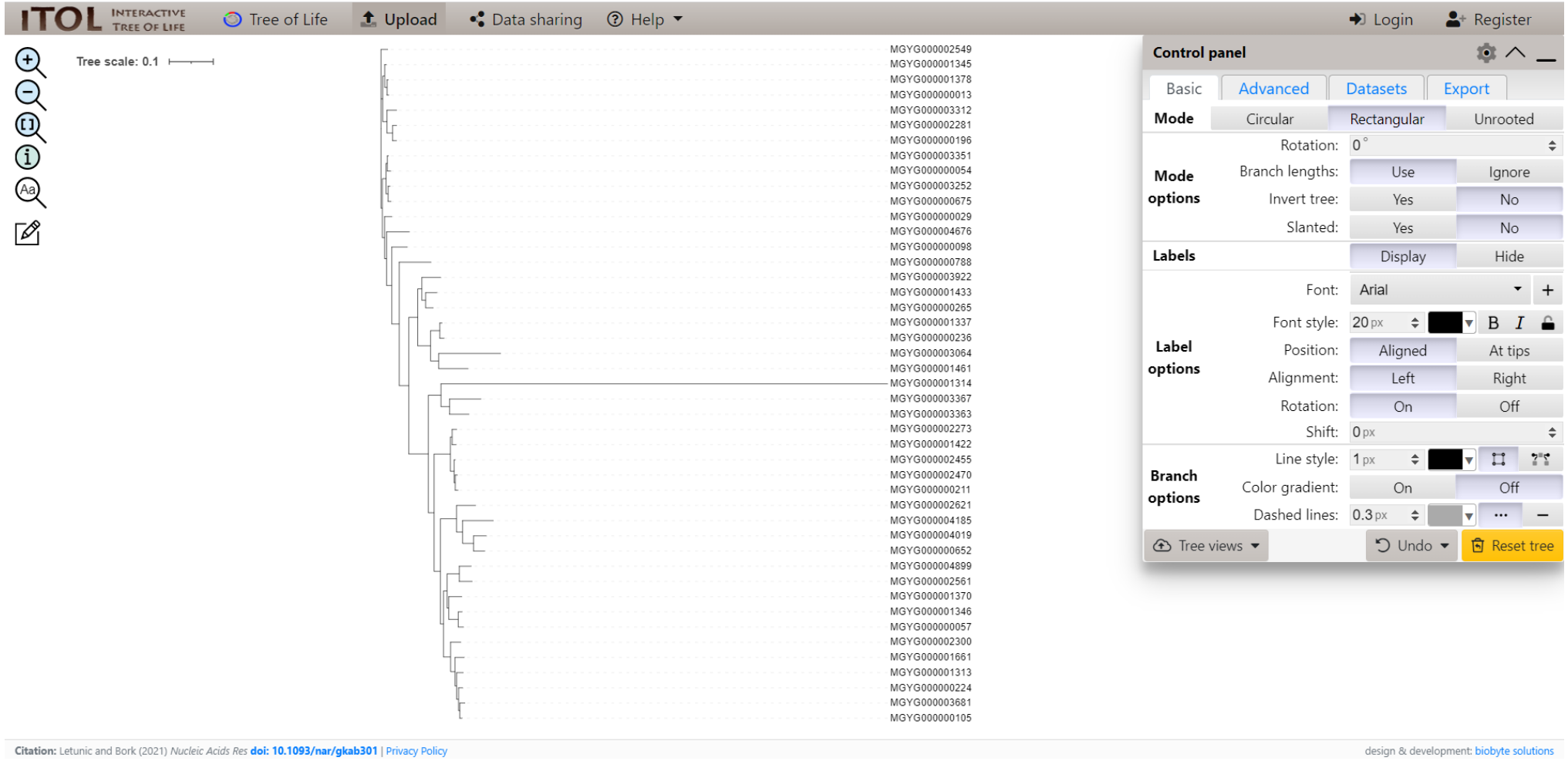
Copy the content

❖ Phylogeny analysis based on SCG alignment by anvi'o

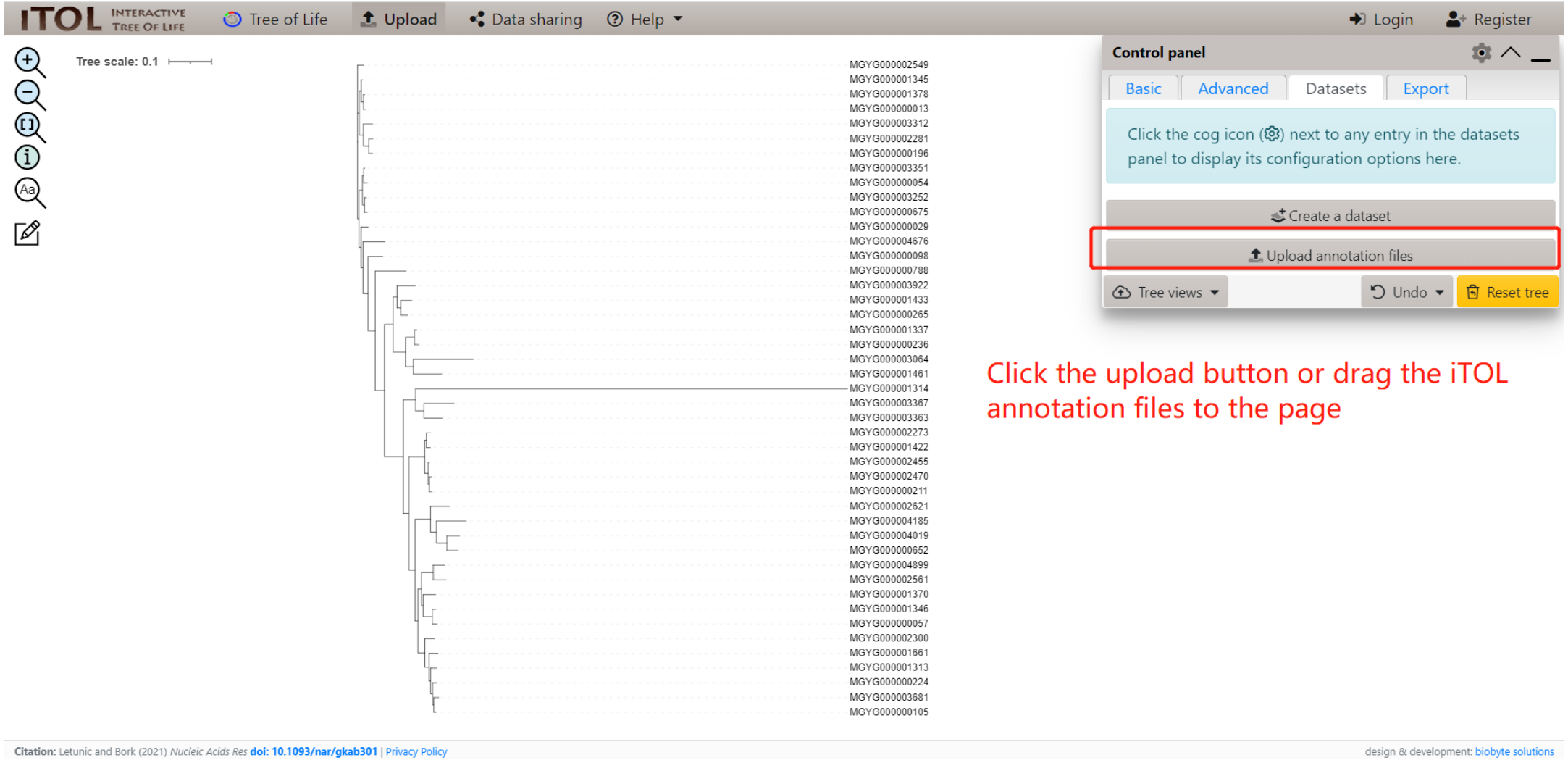
- Open iTOL web page (https://itol.embl.de/upload.cgi)
- Upload newick content for tree visualization



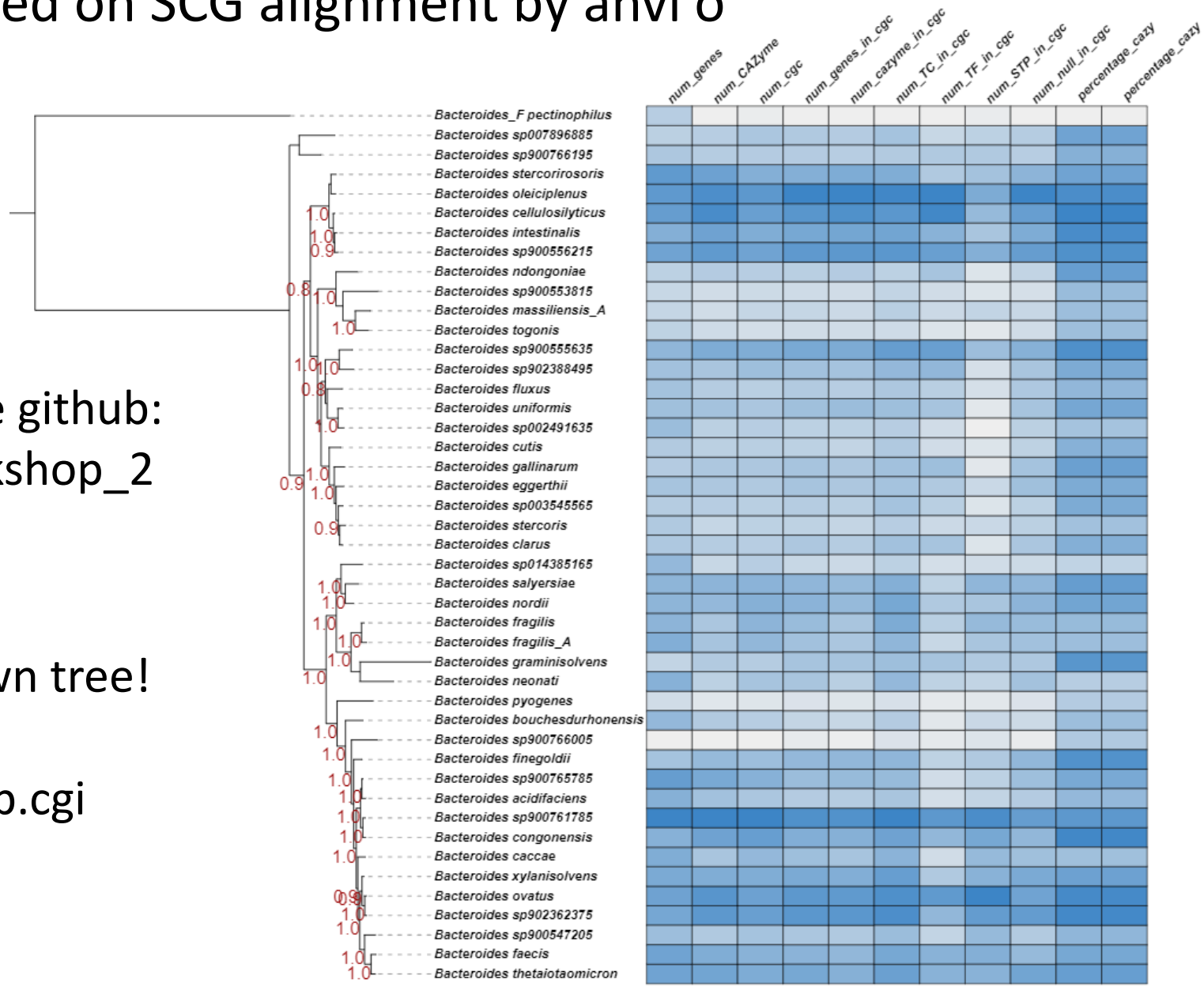
❖ Phylogeny analysis based on SCG alignment by anvi'o



❖ Phylogeny analysis based on SCG alignment by anvi'o



❖ Phylogeny analysis based on SCG alignment by anvi'o



- The annotation files are on the github: https://github.com/tli14/Workshop_2022_YinLab

- You can also customize your own tree!

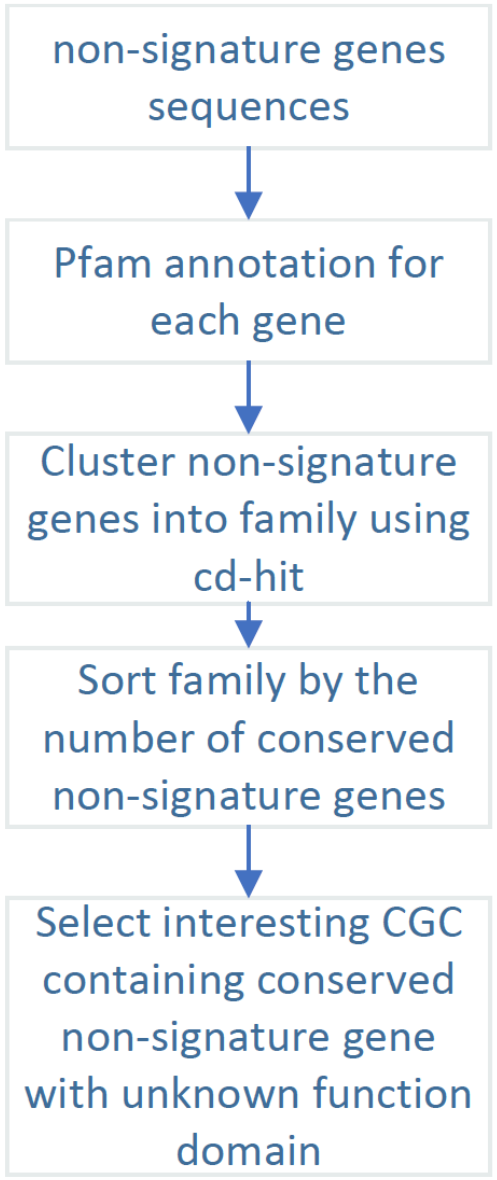
  The iTOL help page:

  https://itol.embl.de/help.cgi

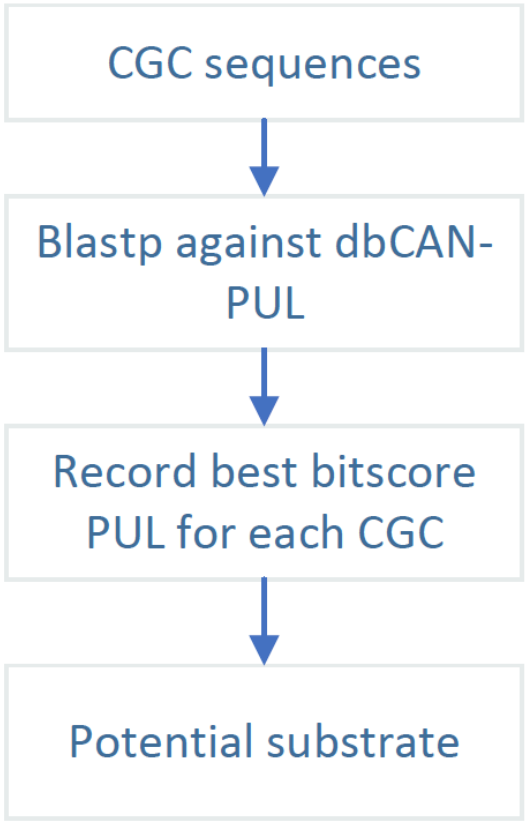
❑ Part 4

Pfam annotation of non-signature genes

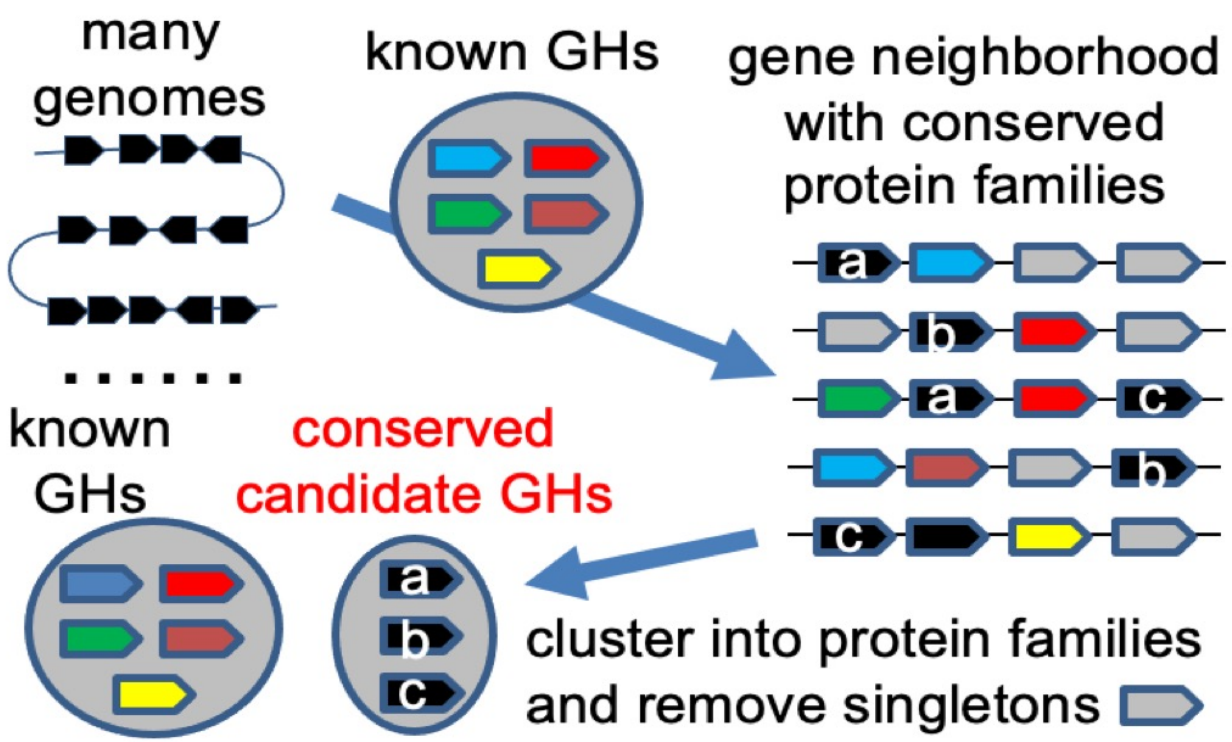
Pipeline of conserved non-signature genes



Pipeline of substrate prediction



Pipeline of conserved non-signature genes



Script Name: Prepare_null_gene.sh

Script Function: Extract all non-signature gene sequences from dbCAN result.

- Input 1: all the fasta sequence (**uniInput**)
- Input 2: cgc predicted by dbCAN (**cgc_standard.out**)
- Output: non-signature gene protein sequences (**non-signature.pep**)

Script Name: pfam_paralla_job.sh

Script Function: Domain annotation of non-signature genes to get possible function

- Input 1: **non-signature** protein sequences
- Input 2: Pfam database.
- Output: domain annotation

Script Name: cd-hit.sh

cd-hit -i .pep -o non-signature_40.pep -c 0.4 -M 0 -T 32 -n 2

Part1 Function:

- Input: non-signature protein sequences
- Output: clustered non-signature protein at 40% identity

–c 0.4 : global identity = 0.4
–M 0: Memory usage not limited
–T 32: CPU 32
-n 2: word size for global identity

Part2 Function: Annotate the non-signature gene clusters with pfam.

Pipeline of Substrate Prediction

Script Name: Prepare_CGC_seq.sh

Script Function: Extract all CGC sequences from dbCAN result.

- Input 1: **uniInput** all the fasta sequence
- Input 2: **cgc_standard.out**, CGCs predicted by dbCAN
- Output: all CGC sequences

Script Name: blastp.sh

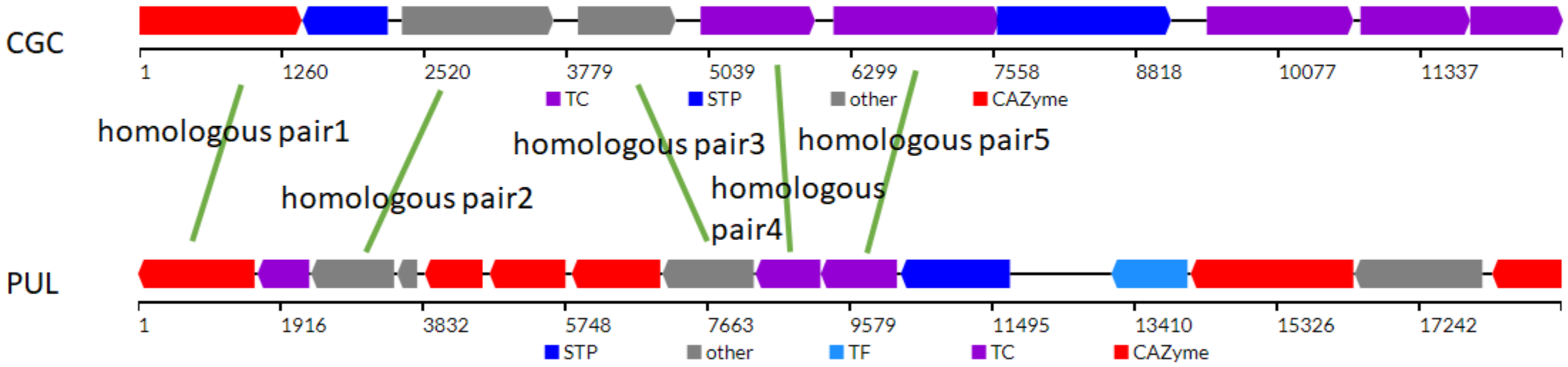
Script Function: CGC database search against with PUL database to find homologous

- Input 1: CGC sequences
- Input 2: dbCAN-PUL database, download from dbCAN-PUL server
- Output: homologous hits

Script Name: substrate.sh

Script Function: Deal with blastp result to infer potential substrate

- Input 1: blastp outputs (Bacteroides.CGC.PUL.blastp)
- Input 2: CGC predictions (cgc_standard.out)
- Input 3: **dbCAN-PUL.xlsx**
- Output: CGC with substrate



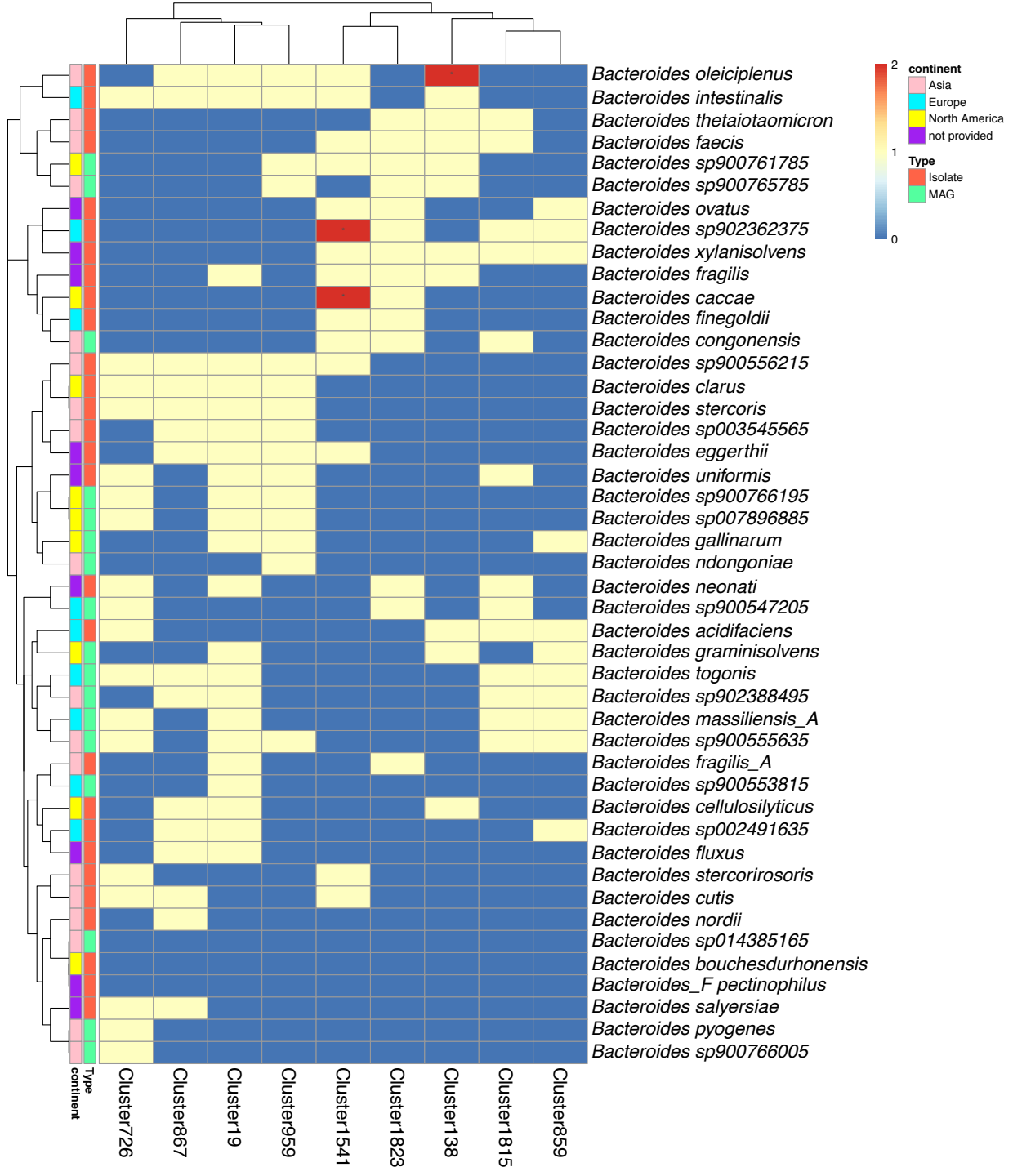
❖ Pfam annotation for non-signature genes (nSGs)

- Two Pfam families (PF07980.14 and PF14322) can be found in 44/45 genomes → only absent in *Bacteroides_F pectinophilus*.

- A total of 25 Pfam families can be found in > 30 genomes.

❖ Pfam unannotated protein clusters for nSGs

- The nSGs were clustered into protein clusters.
- Some protein clusters without Pfam annotation were found in > 10 genomes, which might be conserved nSGs in Bacteroides genus.



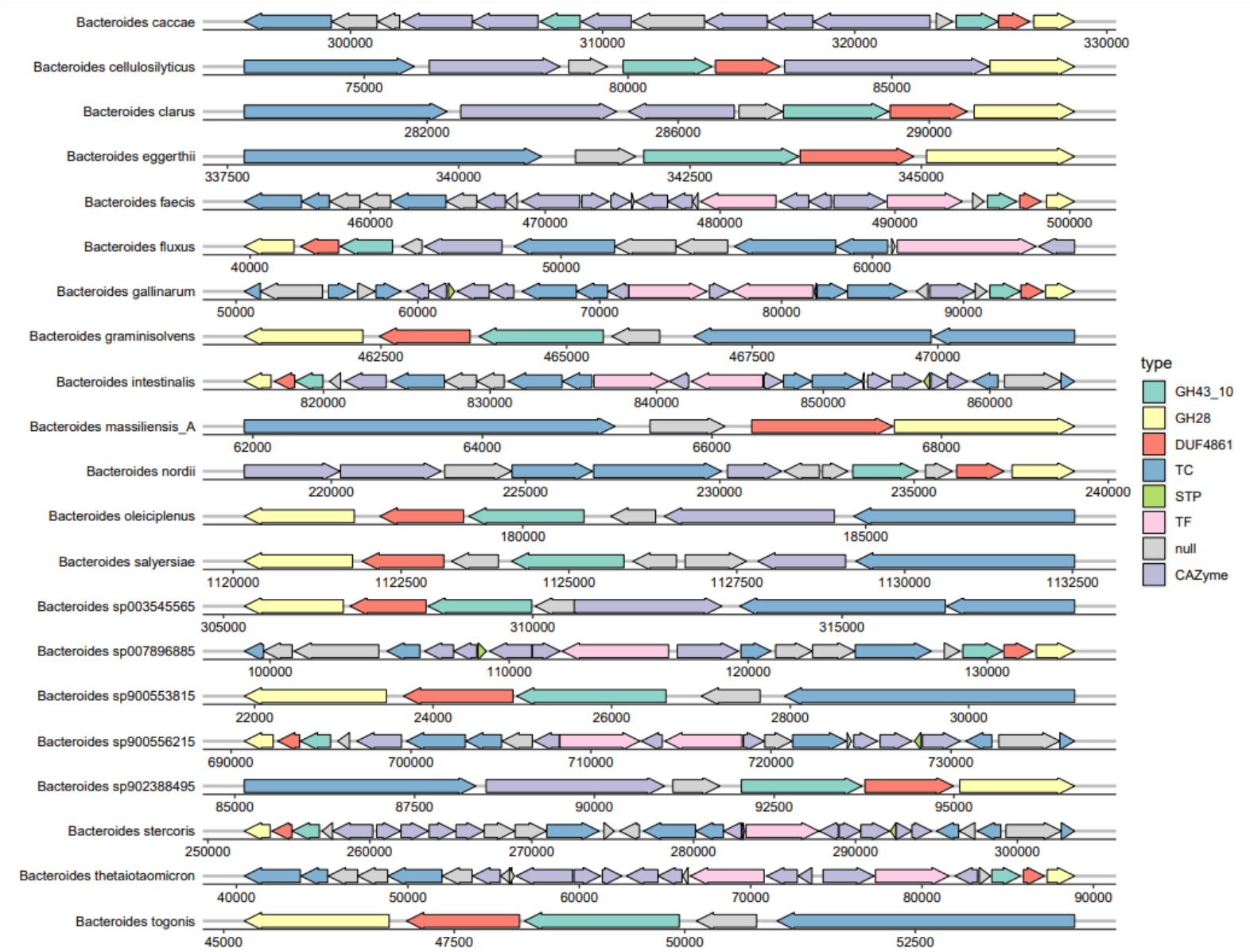
❖ CGC synteny plot with conserved nSG family highlighted

- The proportion of genes in CGC (cgc_synteny.pdf)

- Null gene DUF4861 without pfam function annotated is conserved in 21 of 45 genomes

- CGC containing DUF4861 is also conserved across genomes, with GH43_10 and GH28

- Haven't ordered by phylogeny



❑ Part 5

CGC clustering

❖ CGCs clustering

1. Prepare Input Table

Genome_ID	CGC_ID	sig_gene_seq	low_level_substr	Species
MGYG000000013	MGYG000000013_1ICGC1	1.B.14,GH29,GH16,GH76	galactomannan	Bacteroides sp902362375
MGYG000000013	MGYG000000013_1ICGC2	1.B.14,HTH_AraC,GH92,GH76,GH1…	alpha-mannan	Bacteroides sp902362375
MGYG000000013	MGYG000000013_1ICGC3	GH18,8.A.46,1.B.14	host glycan	Bacteroides sp902362375
MGYG000000013	MGYG000000013_1ICGC4	GH13,GH97,1.B.14	starch	Bacteroides sp902362375
MGYG000000013	MGYG000000013_1ICGC6	2.A.21,GT2,2.A.1,SIS,GH115,GH95,…	arabinoxylan	Bacteroides sp902362375

2. Train and Test Subsets

70% for CGC clustering

30% for mapping training

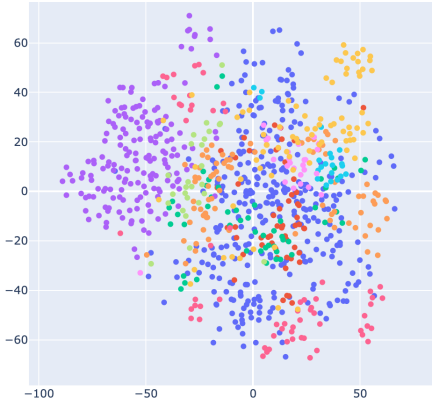
3. Convert words to vectors

```
0: '1.A.1',
1: '1.A.11',
2: '1.A.13',
3: '1.A.22',
4: '1.A.23',
5: '1.A.26',
```

4. CGC clustering

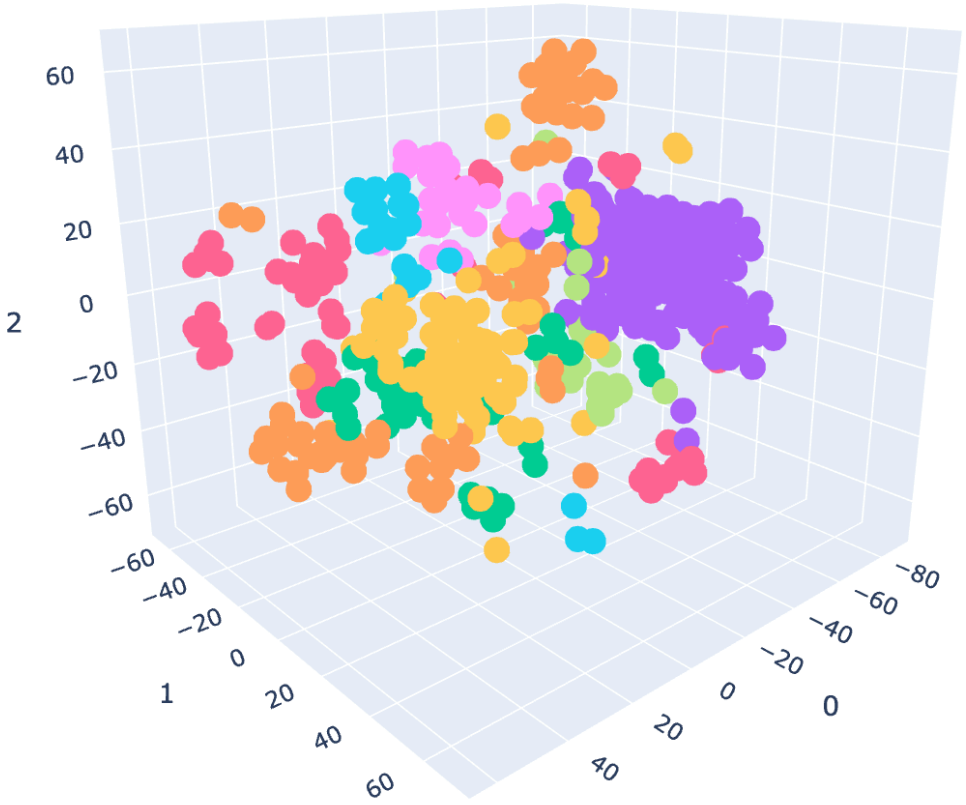
	cluster_id	number of samples
4	80	58
39	81	45
7	1	44
24	141	38
42	5	35

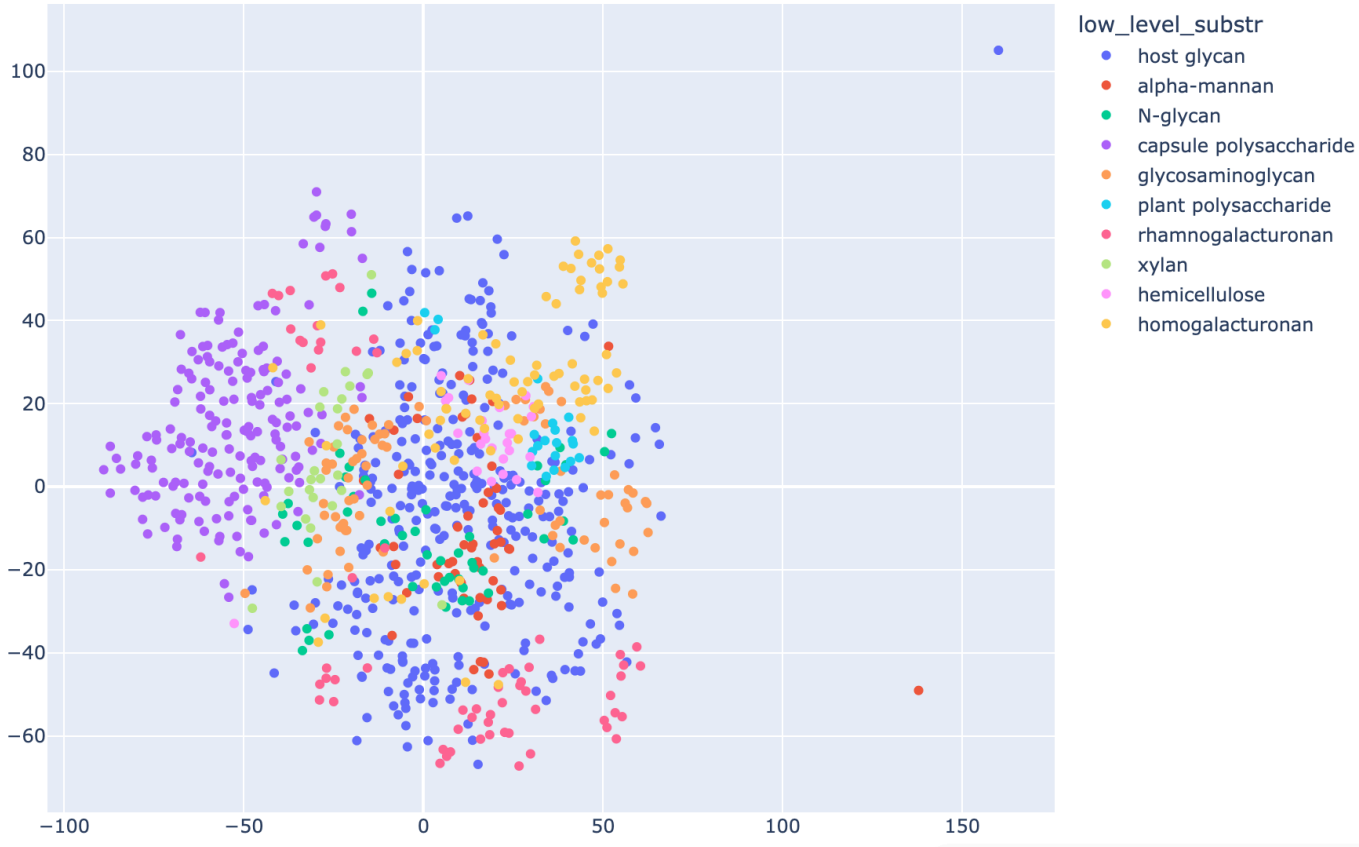
5. Visualization



❖ CGCs clustering results

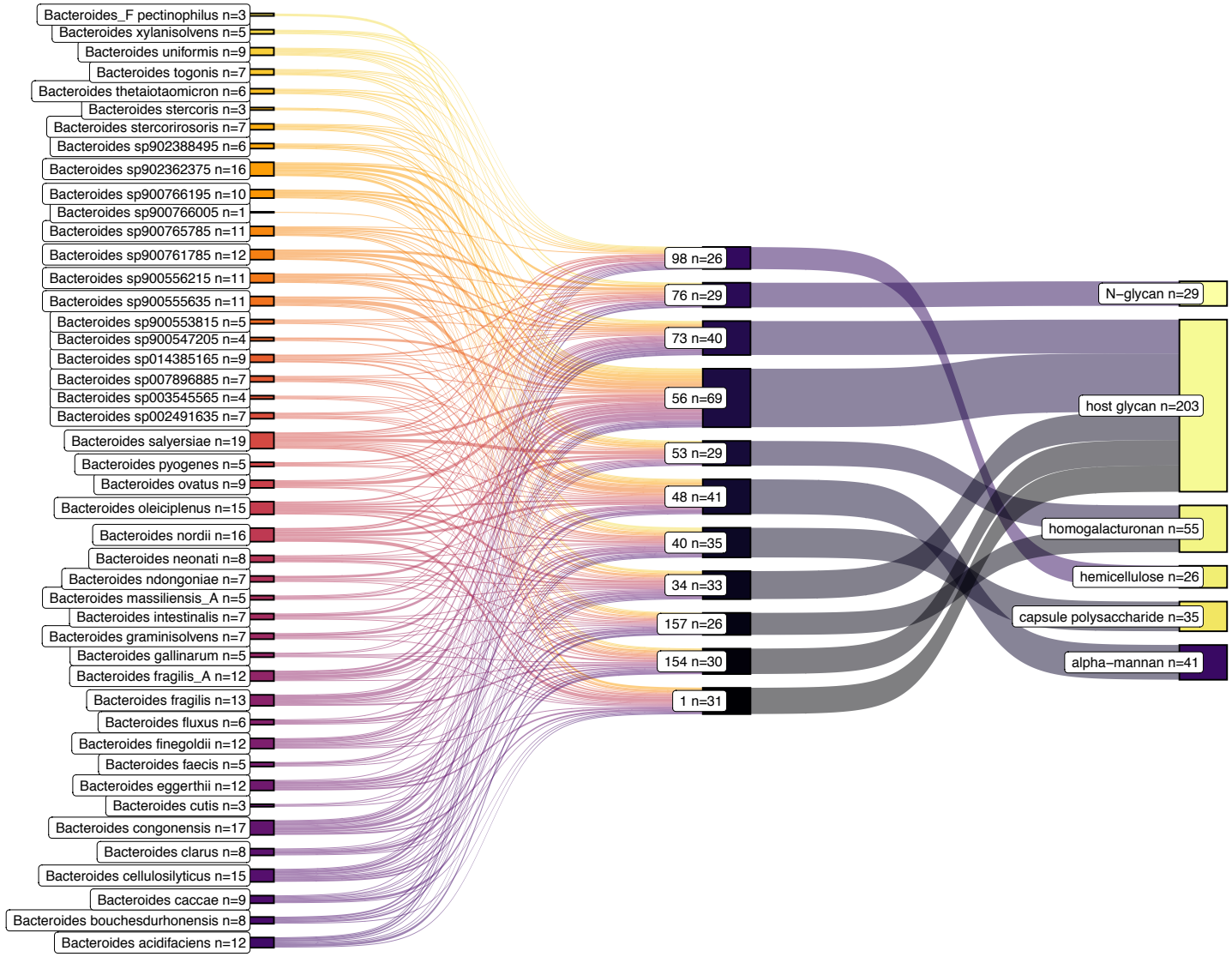
- T-distributed Stochastic Neighbor Embedding (**TSNE**) plot: visualize high-dimensional data.
- The top 10 CGC families and their substrates were found.





❖ CGCs clustering results

- **Sankey plot**: allows to show study flows.
- 11 CGC families contain at least 25 CGC sequences.
- Host glycan is the most abundant substrate for these CGC families.




Thank You!