UFIT Workshop 2023 Data Analysis- Additional Information and helpful links

The following documents provide helpful information as you all begin looking into the data from UFIT. This should supplement the material discussed in the data analysis portion of the workshop. As always, please don't hesitate to reach out to me (allina.bennett@ufl.edu) with any questions and I'll make sure to forward to those on our team who are able to help. 😊

Table of contents:

Helpful Links, including OneDrive folders

Please notify Allina if you do not have access to these folders. All permissions should be updated with participant's information.

All data for groups 1-4: ONT_seq_output_and_pipeline

- → Contains pipeline (scripts for adapters, kraken, mash, etc)
- → This was originally shared via email on March 15, 2023

Additional scripts and README.txt including steps for taxonomic classification: 2023_UFIT_pipeline_Jose

- → Contains demonstration analysis by group (demo_output_by_teams)

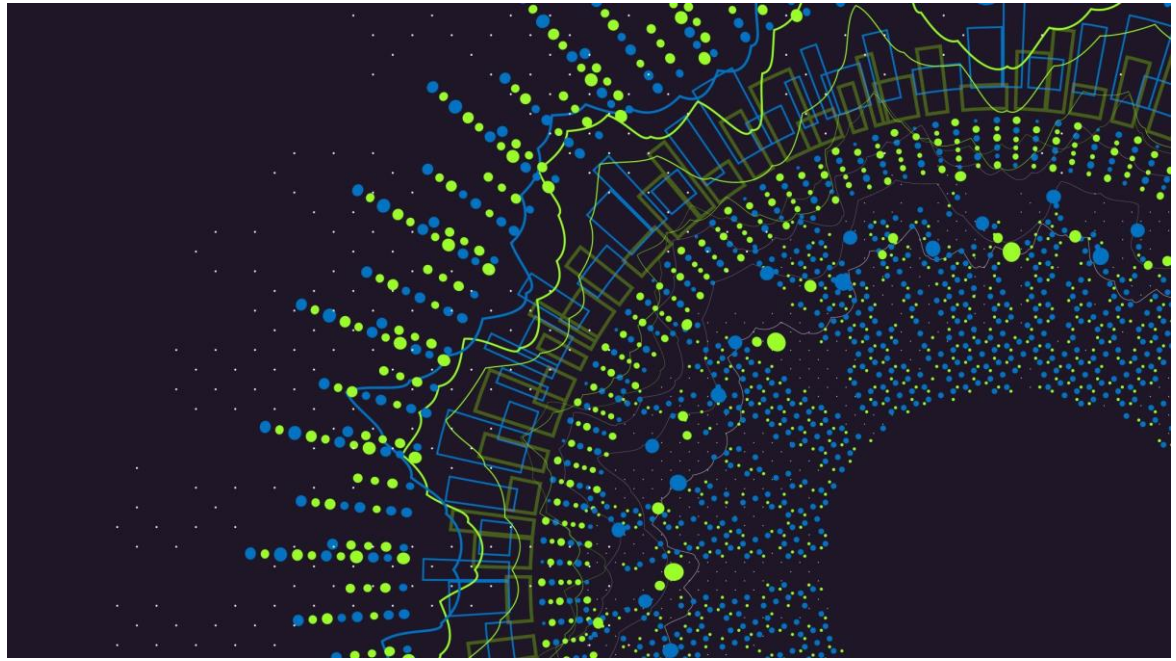
Helpful online resources:

- → Detailed instructions for Mash 2.0: Publications — Mash 2.0 documentation
  - Tutorial for Mash: https://mash.readthedocs.io/en/latest/tutorials.html
- → Detailed instructions for Kraken2: kraken2/MANUAL.markdown at master · DerrickWood/kraken2 · GitHub
  - Script for Kraken database: kraken_db.sh

April 24, 2023

Pei-Ling Yu

Workflow for Kraken2 and Mash



Kraken2

Required files

- kraken_4_fungi.sh: bash script
- Filtered reads file (.fastq)
- Database: customized or standard database (Instruction: Manual · DerrickWood/kraken2 Wiki · GitHub)

- If you are the user of HiPerGator, follow the instruction below to print the page of module usage. (copy the command after “$”))

- 1) Module load kraken
  - $ ml kraken
- 2) Print the page of module usage
  - $ kraken2 –help

- Please refer to the source page of the supercomputer service of your institute.

- You can also work on your local computer. Please follow the instruction here: [Manual · DerrickWood/kraken2 Wiki · GitHub]

```
[plyu@login1 20230208_4_samples_enriched]$ ml kraken
[plyu@login1 20230208_4_samples_enriched]$ kraken2 -help
Usage: kraken2 [options] <filename(s)>

Options:
  --db NAME               Name for Kraken 2 DB
                          (default: none)
  --threads NUM           Number of threads (default: 1)
  --quick                 Quick operation (use first hit or hits)
  --unclassified-out FILENAME
                          Print unclassified sequences to filename
  --classified-out FILENAME
                          Print classified sequences to filename
  --output FILENAME       Print output to filename (default: stdout); "-" will
                          suppress normal output
  --confidence FLOAT      Confidence score threshold (default: 0.0); must be
                          in [0, 1].
  --minimum-base-quality NUM
                          Minimum base quality used in classification (def: 0,
                          only effective with FASTQ input).
  --report FILENAME       Print a report with aggregrate counts/clade to file
  --use-mpa-style         With --report, format report output like Kraken 1's
                          kraken-mpa-report
  --report-zero-counts    With --report, report counts for ALL taxa, even if
                          counts are zero
  --report-minimizer-data With --report, report minimizer and distinct minimizer
                          count information in addition to normal Kraken report
  --memory-mapping        Avoids loading database into RAM
  --paired                The filenames provided have paired-end reads
  --use-names             Print scientific names instead of just taxids
  --gzip-compressed       Input files are compressed with gzip
  --bzip2-compressed      Input files are compressed with bzip2
  --minimum-hit-groups NUM
                          Minimum number of hit groups (overlapping k-mers
                          sharing the same minimizer) needed to make a call
                          (default: 2)
  --help                  Print this message
```

Modify script

- Open “kraken_4_fungi.sh” using [NANO text editor]():

  $ nano kraken_4_fungi.sh

- Areas that are pointed by arrows or boxes are need to be changes accordingly.
- Ctrl+X to close/save the text file.

```
[plyu@login1 kraken2]$ nano kraken_4_fungi.sh
[plyu@login1 kraken2]$
```

```
 GNU nano 2.3.1                File: kraken_4_fungi.sh

#!/bin/sh
#SBATCH --account=jeremybrawner
#SBATCH --qos=jeremybrawner
#SBATCH --job-name=k_test
#SBATCH --mail-type=END,FAIL
#SBATCH --mail-user=plyu@ufl.edu
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --mem=200gb
#SBATCH --time=72:00:00
#SBATCH --output=k_test_%j.out
pwd; hostname; date

ml kraken
kraken2 --db /blue/jeremybrawner/share/kraken_fungi_local_db/fungi_local --quick --use-names \
--output kraken_results_A01 /blue/jeremybrawner/plyu/20230208_4_samples_enriched/Barcode_A01_i5_1000bp.fastq \
--report A01_report  --threads 8

                                                [ Read 18 lines ]
^G Get Help  ^O WriteOut  ^R Read File ^Y Prev Page ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is  ^V Next Page ^U UnCut Text ^T To Spell
```

Location of database

Input: fileted reads

Kraken report name

Kraken output directory

Execute the script

```
[plyu@login1 kraken2]$ sbatch kraken_4_fungi.sh
Submitted batch job 62394306
[plyu@login1 kraken2]$ squeue -u plyu
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          62394306 hpg-milan   k_test     plyu  R       2:51      1 c0713a-s25
[plyu@login1 kraken2]$
```

- ***sbatch*** submits a batch script to Slurm.

  $ sbatch kraken_4_fungi.sh

- ***squeue***: view information about jobs located in the Slurm scheduling queue

  $ squeue –u plyu

Let’s check the outputs

```
[plyu@login1 kraken2]$ head -n 25 A01_report
  4.01  2287    2287    U       0       unclassified
 95.99  54689   0       R       1       root
 95.99  54689   0       R1      131567    cellular organisms
 95.99  54689   0       D       2759        Eukaryota
 95.99  54689   0       D1      33154         Opisthokonta
 95.99  54689   3688    K       4751            Fungi
 85.25  48573   4263    K1      451864            Dikarya
 66.00  37605   8       P       4890                Ascomycota
 65.96  37581   185     P1      716545                saccharomyceta
 64.68  36850   94      P2      147538                  Pezizomycotina
 62.81  35789   859     P3      716546                    leotiomyceta
 58.98  33607   149     P4      715989                      sordariomyceta
 58.14  33125   239     C       147550                        Sordariomycetes
 56.16  31999   46      C1      222543                          Hypocreomycetidae
 55.54  31642   369     O       5125                              Hypocreales
 49.13  27994   10      F       5129                                Hypocreaceae
 48.16  27442   6       G       5543                                  Trichoderma
 48.12  27416   0       G1      2600217                                 unclassified Trichoderma
 48.11  27413   27413   S       2809032                                   Trichoderma sp. MLT1J1
  0.00  2       2       S       2694992                                   Trichoderma sp. TW21990_1
  0.00  1       1       S       2717280                                   Trichoderma sp. TAM-2020a
  0.01  4       4       S       398673                                  Trichoderma gamsii
  0.01  3       3       S       654480                                  Trichoderma cornu-damae
  0.00  2       2       S       1195189                                 Trichoderma gracile
  0.00  2       2       S       500994                                  Trichoderma pleuroti
[plyu@login1 kraken2]$
```

- To view the log file:

  $ cat k_test_JOBID.out

- To view the first 5 line of kraken output:

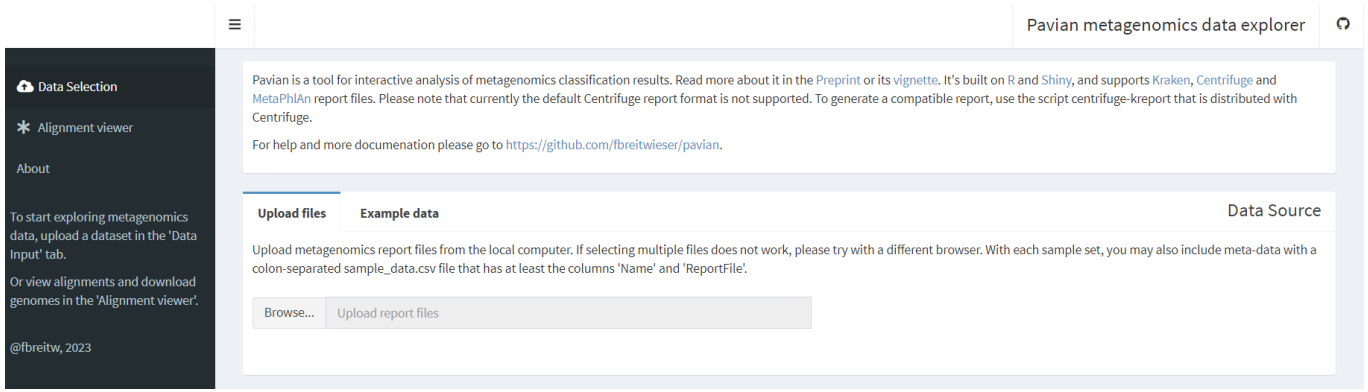
  $ head –n 5 kraken_results_A01

- To view the first 25 line of kraken report (human readable):

  $ head –n 25 A01_report

To visualize the output on Pavian metagenomic data explorer

- Navigate yourself to [Pavian (shinyapps.io)]



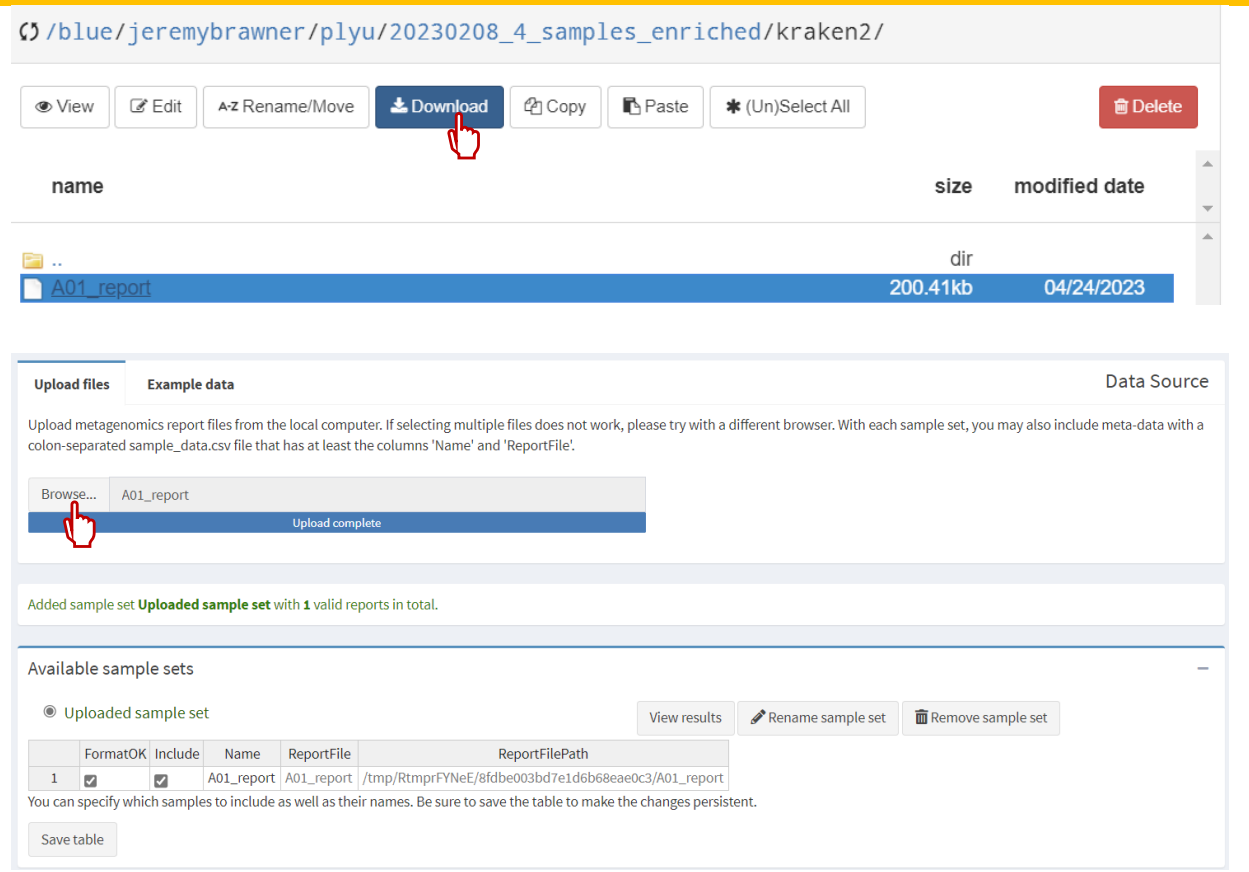
Upload kraken output file

- Download a file from a server to your desktop using SSH:

  $ scp your_username@remotehost:pathy_to_your_file /local/dir

  Or download through OnDemand:

- Upload the output file to Pavian



Data Input

Uploaded sample set

Results Overview

Sample

Comparison

Alignment viewer

About

Bookmark state ...

Generate HTML report ...

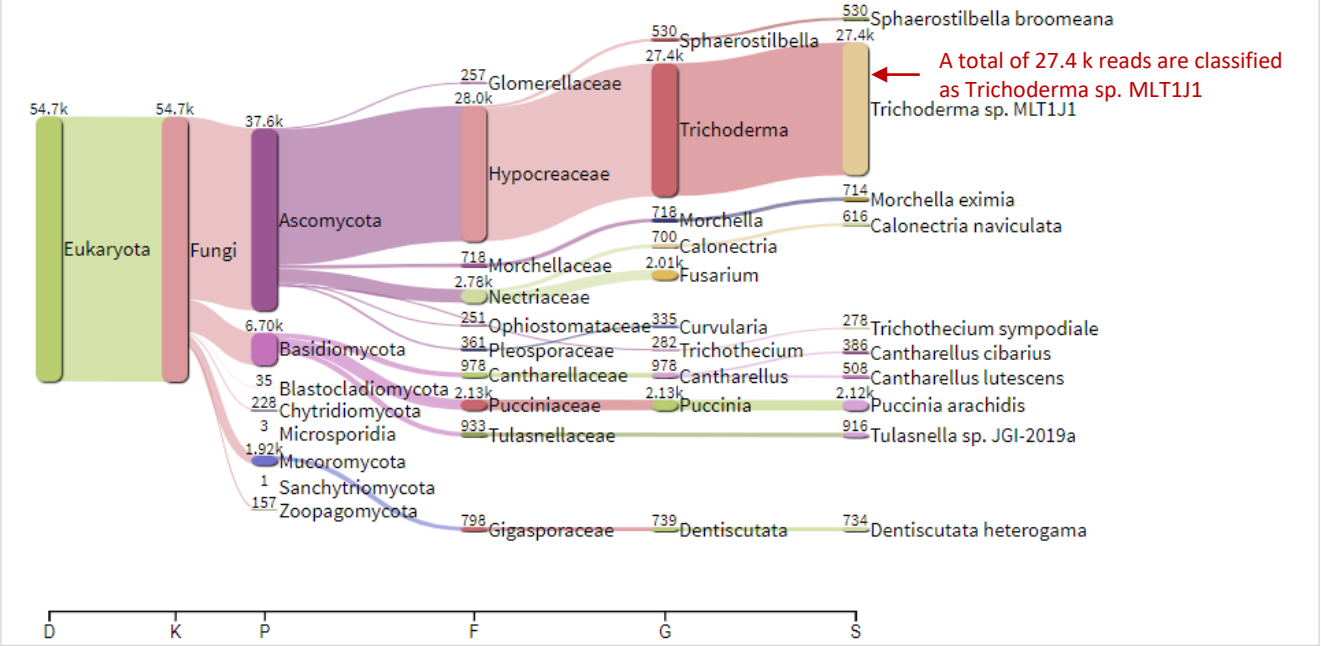
@fbreitw, 2023

Select sample A01_report

Filter taxa Chordata artificial sequences Other

Sankey visualization | **Table** | **Text**

Hover over a node to see the abundance of the taxon in other samples. Configure Sankey ...



Save Network

Across samples

Need more than one sample in sample

Extract FASTA files classified to certain taxa

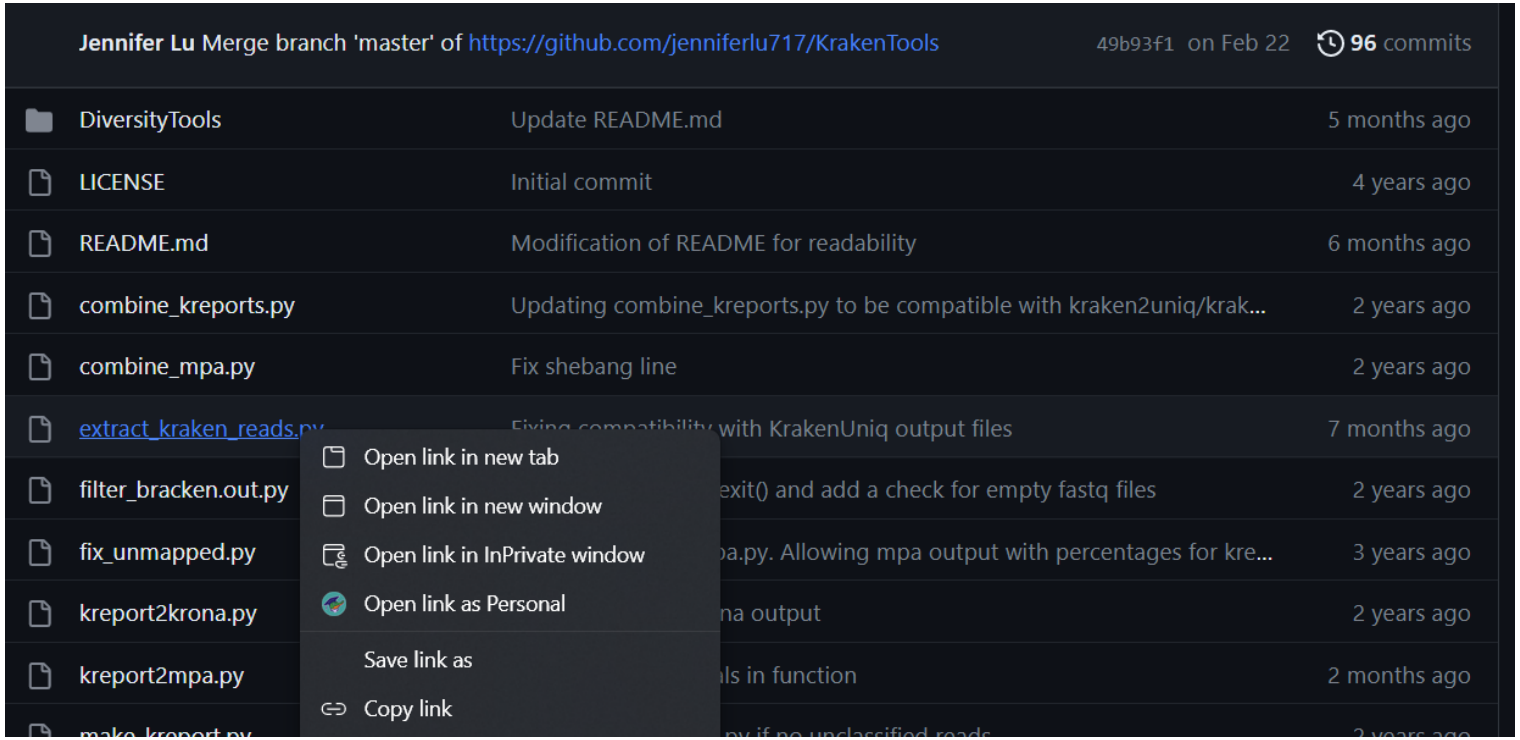
- Instructions: [GitHub - jenniferlu717/KrakenTools: KrakenTools provides individual scripts to analyze Kraken/Kraken2/Bracken/KrakenUniq output files]
- Download the python script, extract_kraken_reads.py:

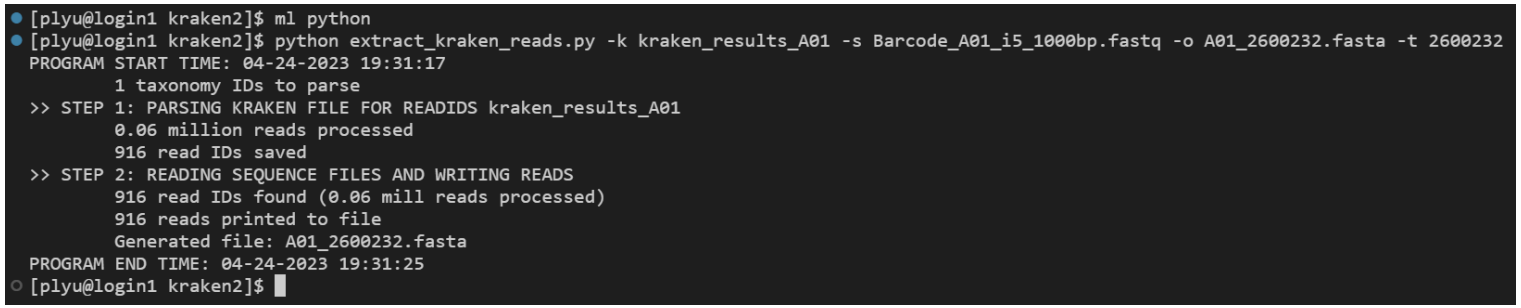
  Right click the file to save link as "extract_kraken_reads.py"

- Execute the command:

  $ ml python

  $ python extract_kraken_reads.py -k YOUR_KRAKEN_OUTPUT -s FILTERED_FASTQ -o OUT.fasta -t TAXID





Mash

Required files

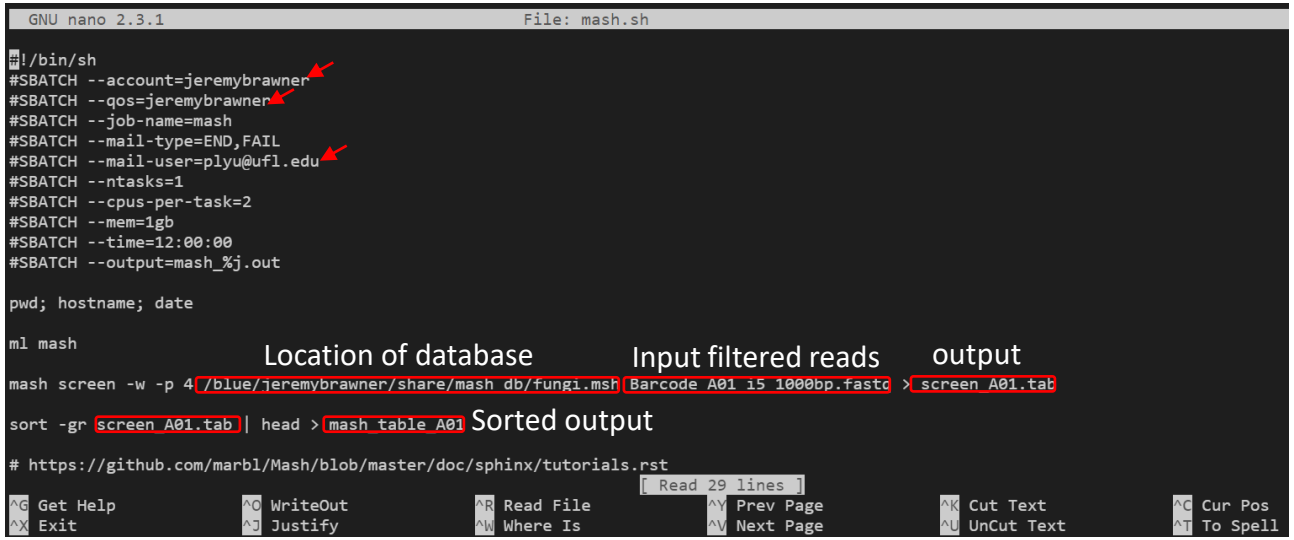
- mash.sh
- Filtered reads (FASTQ)
- Database: please follow the instruction to construct the database (Mash/tutorials.rst at master · marbl/Mash · GitHub)

Edit bash script

- Open “mash.sh” using NANO text editor:

  $ nano mash.sh

- Areas that are pointed by arrows or boxes are need to be changes accordingly.
- Ctrl+X to close/save the text file.



Execute the script and check job status

- ***sbatch*** submits a batch script to Slurm.
  - $ sbatch mash.sh
- ***squeue***: view information about jobs located in the Slurm scheduling queue
  - $ squeue –u plyu

```
[plyu@login1 mash]$ sbatch mash.sh
Submitted batch job 62402769
[plyu@login1 mash]$ squeue -u plyu
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          62402769 hpg-defau     mash     plyu  R       0:02      1 c0702a-s24
[plyu@login1 mash]$
```

Output of Mash

- To view the entire sorted mash output
  $ cat mash_table _A01

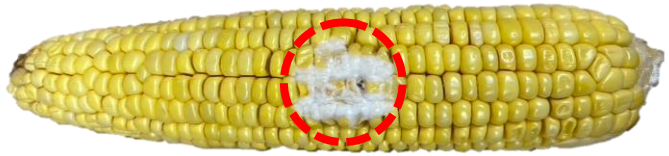
```
[plyu@login1 mash]$ cat mash_table_A01
0.806739        11/1000 2       7.0529e-32      ../../kraken/GCA_022817935.1_ASM2281793v1_genomic.fna   [461 seqs] JAGJIL010000001.1 Trichoderma sp. MLT1J1 Scaffold1, whole genome shotg
un sequence [...]
0.799067        9/1000  4       1.49666e-25     ../../kraken/GCA_019192225.1_ASM1919222v1_genomic.fna   [34 seqs] JAHREO010000001.1 Fusarium sp. TH15 scaffold1, whole genome shotgun seq
uence [...]
0.789561        7/1000  1       2.07059e-19     ../../kraken/GCA_019633535.1_Tulinq235_1.0_genomic.fna  [1742 seqs] WSUE01000001.1 Tulasnella sp. JGI-2019a strain 235 scaffold_1, whole
genome shotgun sequence [...]
0.758338        3/1000  2       6.34812e-08     ../../kraken/GCA_018606675.1_ASM1860667v1_genomic.fna   [15 seqs] JAGYHO010000012.1 MAG: Malasseziomycetes sp. isolate bin8 NODE_1009_len
gth_6427_cov_17.362367, whole genome shotgun sequence [...]
0.758338        3/1000  1       6.34812e-08     ../../kraken/GCA_001990185.1_ASM199018v1_genomic.fna    [331 seqs] MOEQ01000001.1 Salmacisia buchloeana strain OK1 contig001, whole genom
e shotgun sequence [...]
0.758338        3/1000  12      6.34812e-08     ../../kraken/GCA_022577725.1_ASM2257772v1_genomic.fna   [199 seqs] JAKIRV010000100.1 Wickerhamiella dianesei strain PYCC 8330 NODE_100_le
ngth_578_cov_1577.650964, whole genome shotgun sequence [...]
0.743837        2/1000  38      2.62669e-05     ../../kraken/GCA_013186935.1_ASM1318693v1_genomic.fna   [384 seqs] JAAOAX010000001.1 Fusarium thapsinum strain NRRL 22049 NRRL22049_c0000
001, whole genome shotgun sequence [...]
0.743837        2/1000  3       2.62669e-05     ../../kraken/GCA_012980515.1_Ophcf2_genomic.fna [13 seqs] JAACLJ010000001.1 Ophiocordyceps camponoti-floridani strain EC05 scaffold_01, w
hole genome shotgun sequence [...]
0.743837        2/1000  2       2.62669e-05     ../../kraken/GCA_013416785.2_ASM1341678v2_genomic.fna   [400 seqs] SWCQ02000001.1 Fusarium sp. BWC1 scaffold1, whole genome shotgun seque
nce [...]
0.743837        2/1000  2       2.62669e-05     ../../kraken/GCA_003012115.1_Spicellum_roseum_DAOM209012_SROSE_contigs_genomic.fna      [1154 seqs] PXOB01000001.1 Trichothecium sympodia
le strain DAOM 209012 DAOM209012_c0000001, whole genome shotgun sequence [...]
```

Useful resources

- Basic Slurm Commands :: High Performance Computing (nmsu.edu)

Samples

- Index-Host-Pathogen for enriched sequenced:
  - A01-Holly-Unknown (not sure if DNA extracted from plant or pure culture)
  - B01-pure culture-*Tulasnella inquilina* (DNA extracted from fungal pure culture)
  - C01-coconut-Unknown (not sure if DNA extracted from plant or pure culture)
  - D01-corn-*Fusarium verticillioides* (Fc) (DNA extracted from infected tissues)
- Barcode-Host-Pathogen for enriched sequenced:
  - BC13-Holly-Unknown (not sure if DNA extracted from plant or pure culture)
  - BC15-pure culture-*Tulasnella inquilina* (DNA extracted from fungal pure culture)
  - BC17-coconut-Unknown (not sure if DNA extracted from plant or pure culture)
  - BC19-corn-*Fusarium verticillioides* (Fc) (DNA extracted from infected tissues)

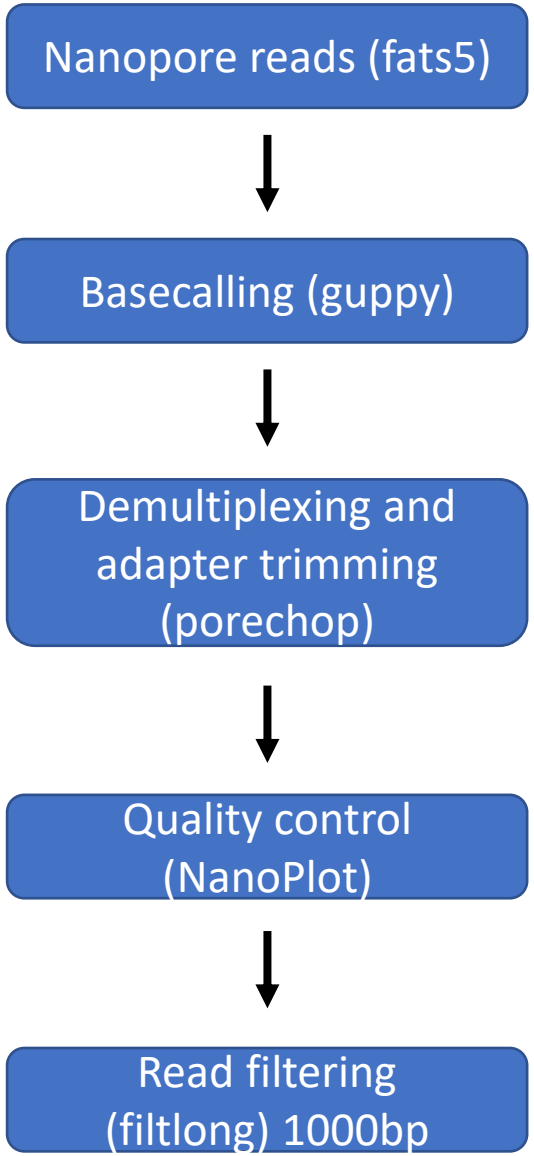


Kernel infected by Fc were collected from the cob. Great number of hyphae was visible.

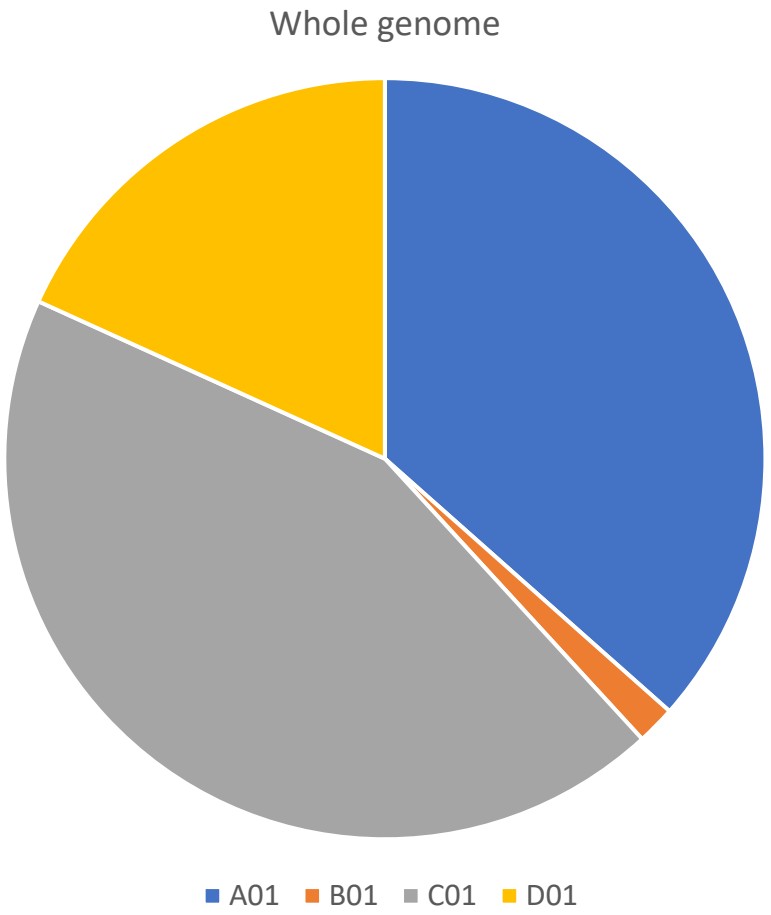
Overview of nanopore reads process

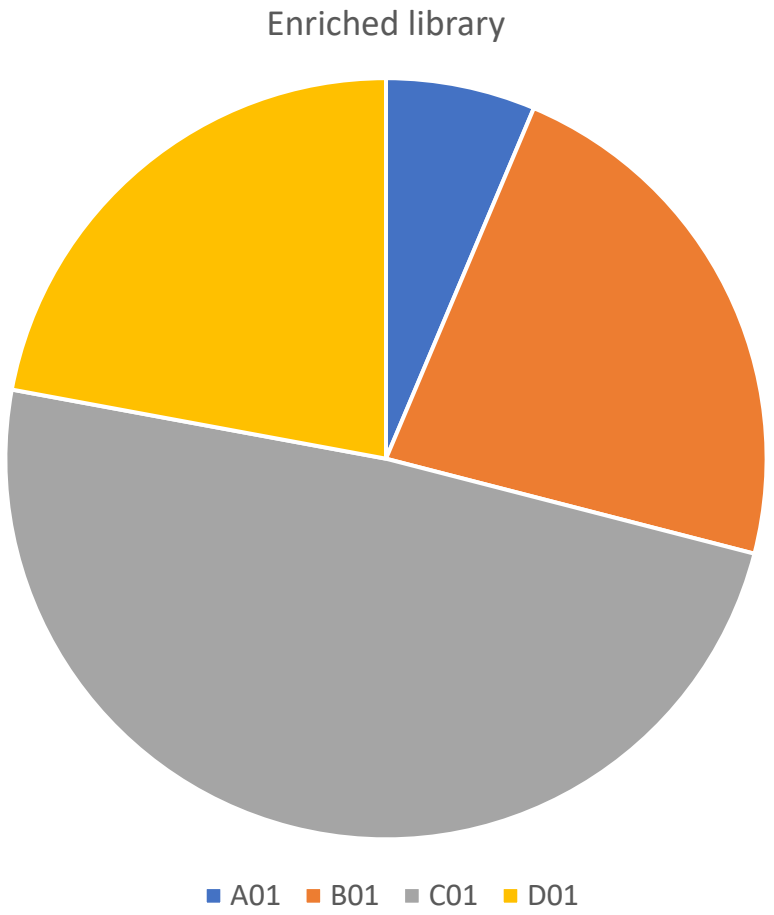
Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



Output reads ratio





DNA pool	Run duration	Reads generated (bases)
Whole genomes	~20 hr	4.15 (8.62 Gb)
Enriched library	~43 hr	15.47 M (18.91Gb)

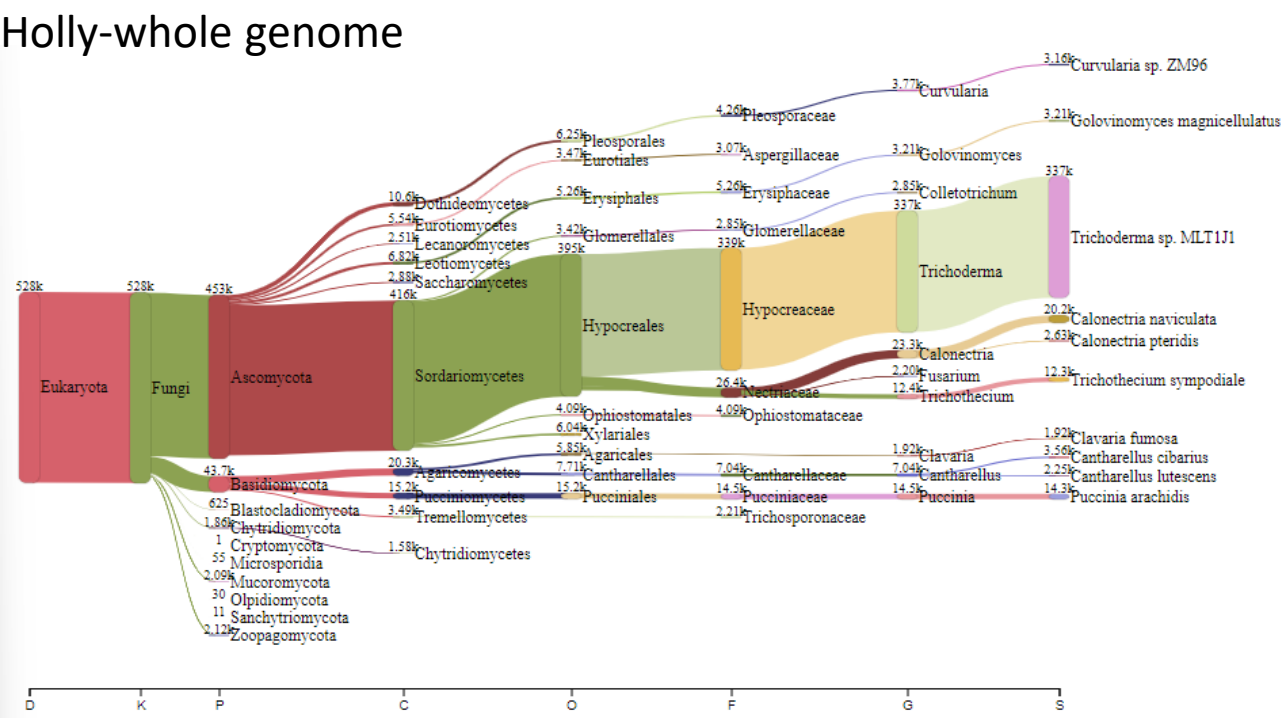
	PDC25540		CIVI-30		coconut/NA		Maize/Fc	
	Whole	Enriched	Whole	Enriched	Whole	Enriched	Whole	Enriched
Mean read length	4,370	1,395	5,153	1,368	3,514	1,374	3,473	1,365
Mean read quality	14	13	14	13	13	13	12	13
Median read length	3,926	1,223	2,662	1,225	3,045	1,229	2,290	1,214
Median read quality	14	12	14	13	13	13	12	13
Number of reads	557,194	56,976	24,935	203,369	664,789	438,507	277,692	198,137
Read length N50	5,666	1,350	9,487	1,313	4,300	1,320	4,955	1,305
Total bases	2,434,661,126	79,500,307	128,486,449	278,129,492	2,335,943,946	602,671,152	964,407,433	270,470,868

	Whole genome sequencing		Enriched sequencing	
	Maize	Fv	Maize	Fv
raw reads post QC	277,692	277,692	198,137	198,137
mapped reads	222,088	34,685	70,840	122,681
mapped reads (%)	80	12	36	62
Increasin of fungal reads (x)				5

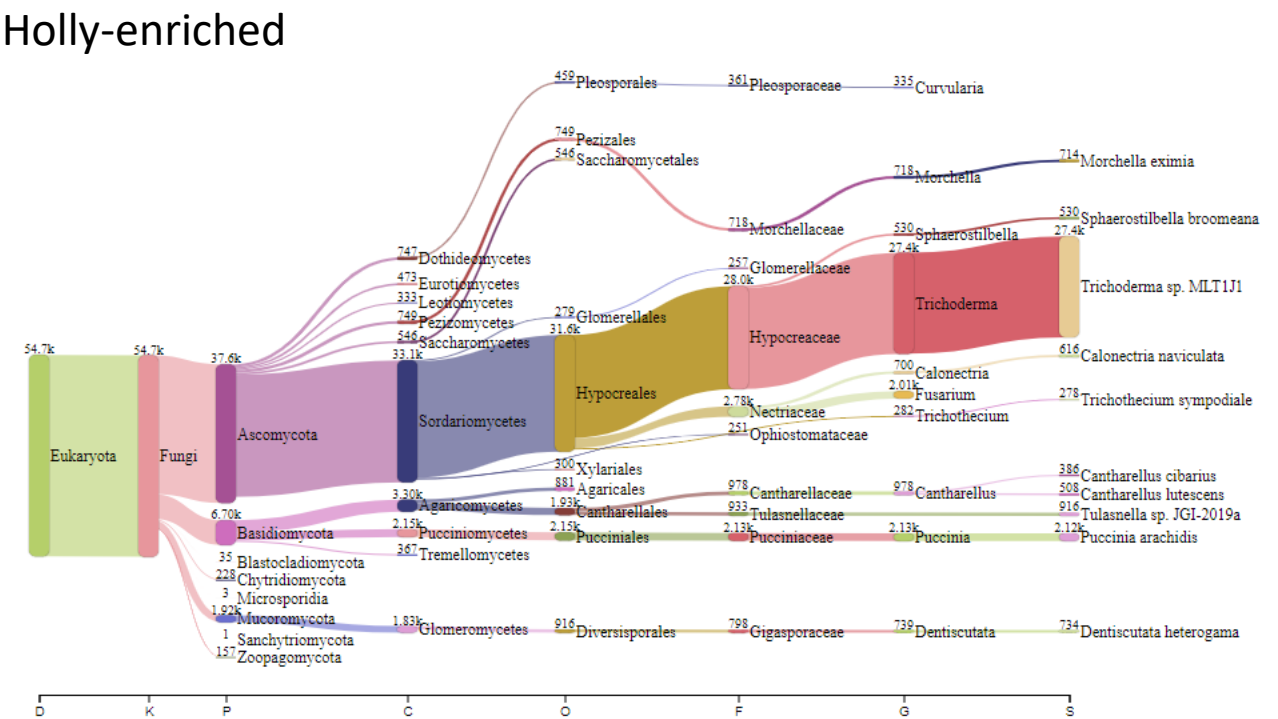
Kernel infected by Fc were collected from the cob. Great number of hyphae was visible.

Fast taxonomic classifications of metagenomic sequence data

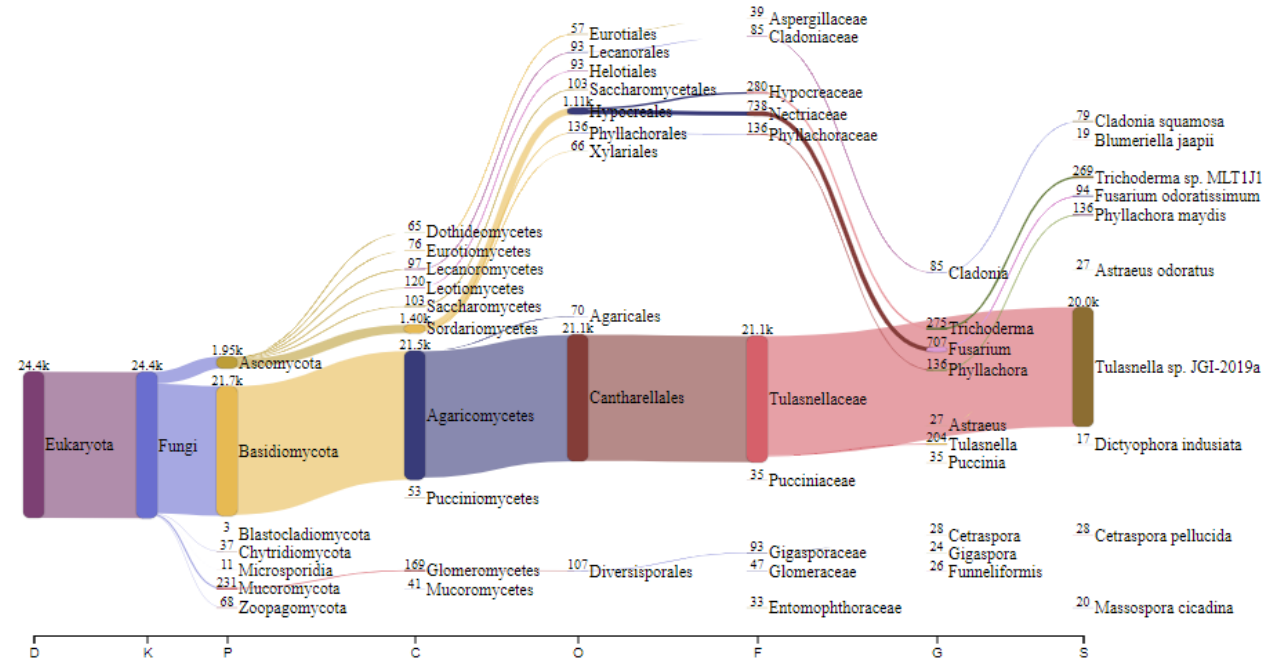
Holly-whole genome



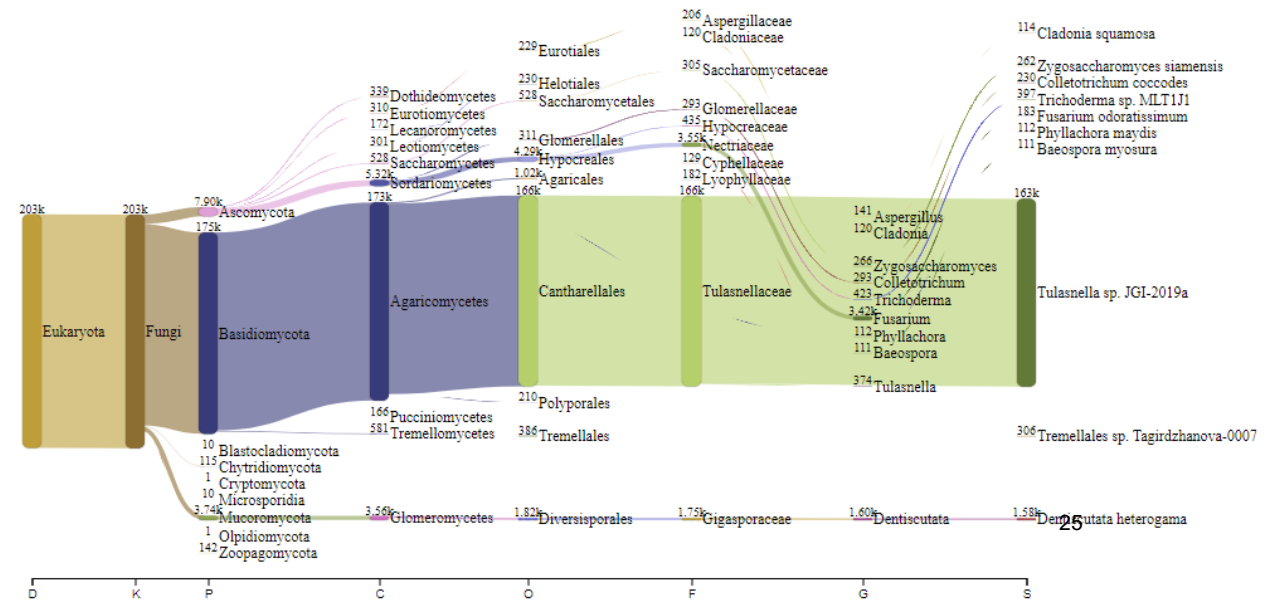
Holly-enriched



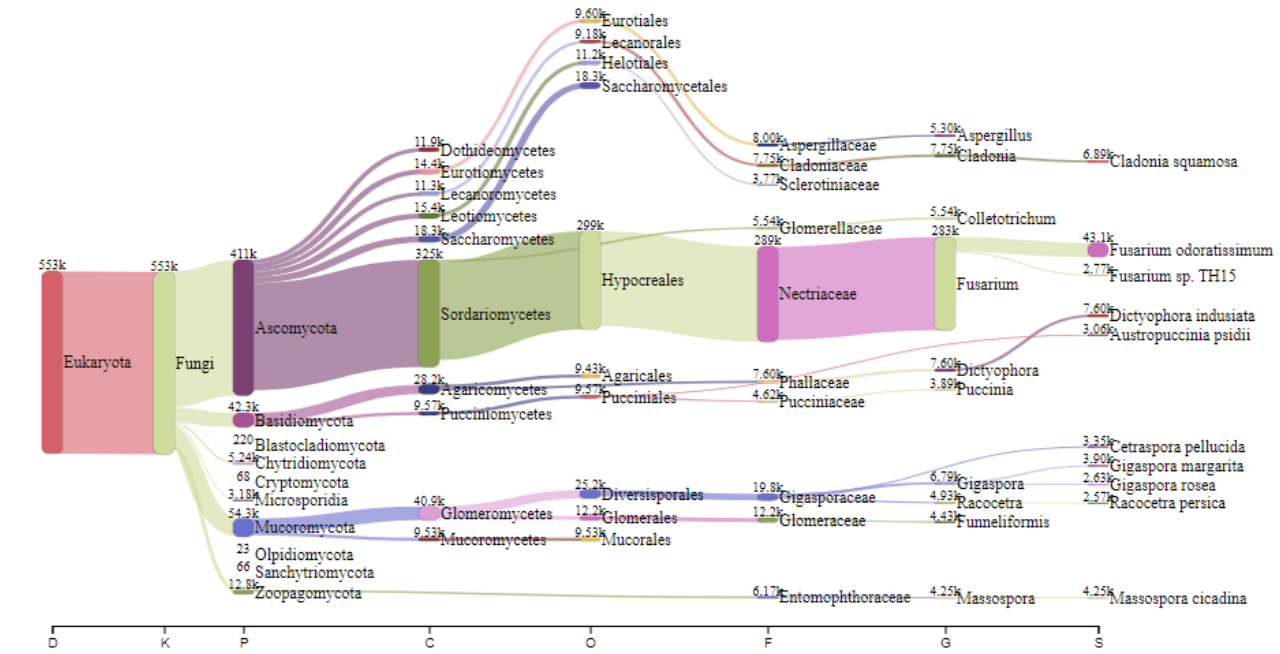
Tulasnella inquilina-whole genome



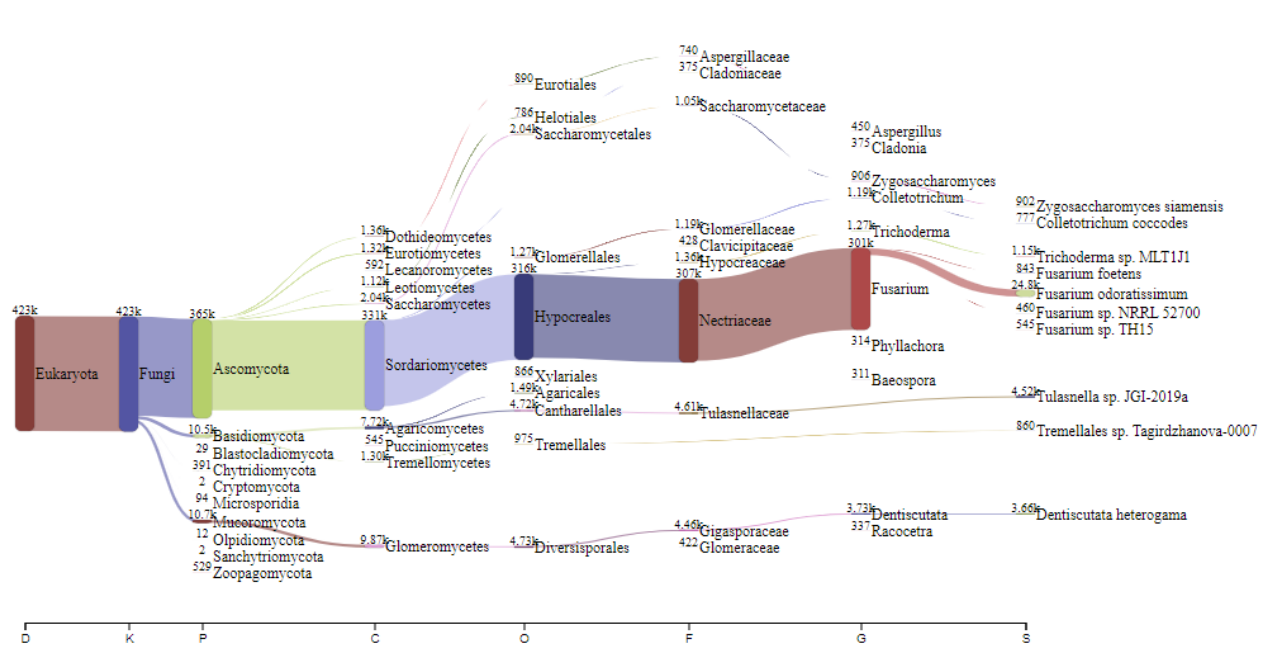
Tulasnella inquilina-enriched



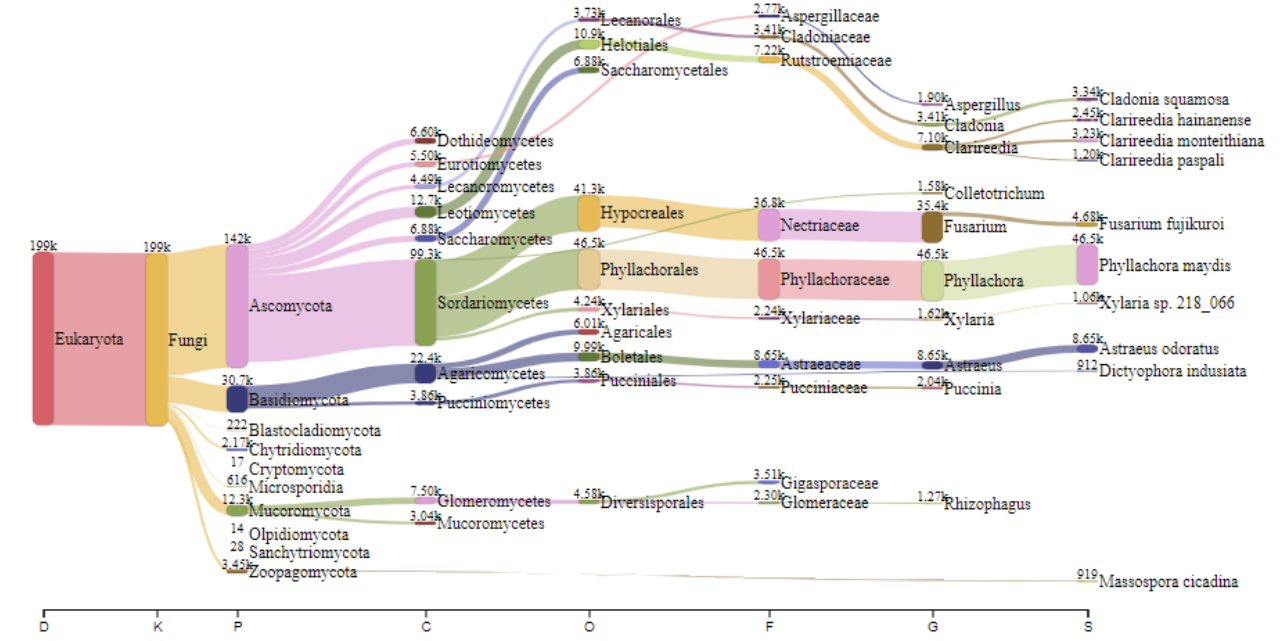
Coconut-whole genome



Coconut-enriched



Corn-whole genome



Corn-enriched