Metagenomes in practice

win a prize !

Assemblers: spades, megahit, velvet

```
spades.py --meta \
--pe1-1 ./data/metanom_1.fastq \
--pe1-2 ./data/metanom_2.fastq \
-k 21,33 --only-assembler \
-o spades_assembly --memory 4 --threads 4
```

Metagenomic pipelines

They all use the same tools, the assemblers , the same methods, usually adaptable
Some are easier some harder but they all need cores and memory.

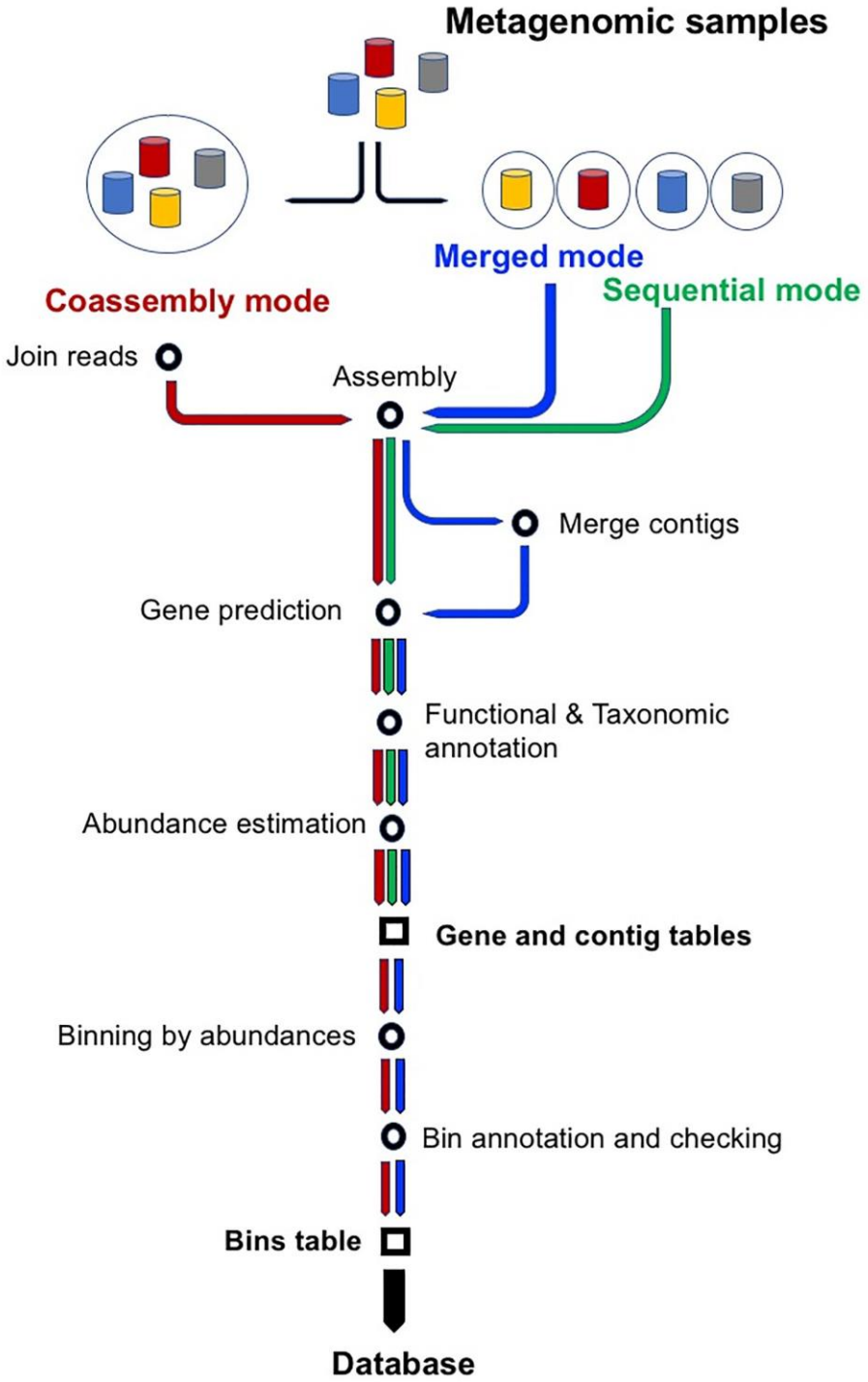
- Quality check trimming
- Assembly
- Normalization
- Annotation
- Binning
- Other functionalities

Pipelines
SqueezeMeta
Atlas
nf-core
bioBakery4
JAMS
WGSA2

SqueezeMeta

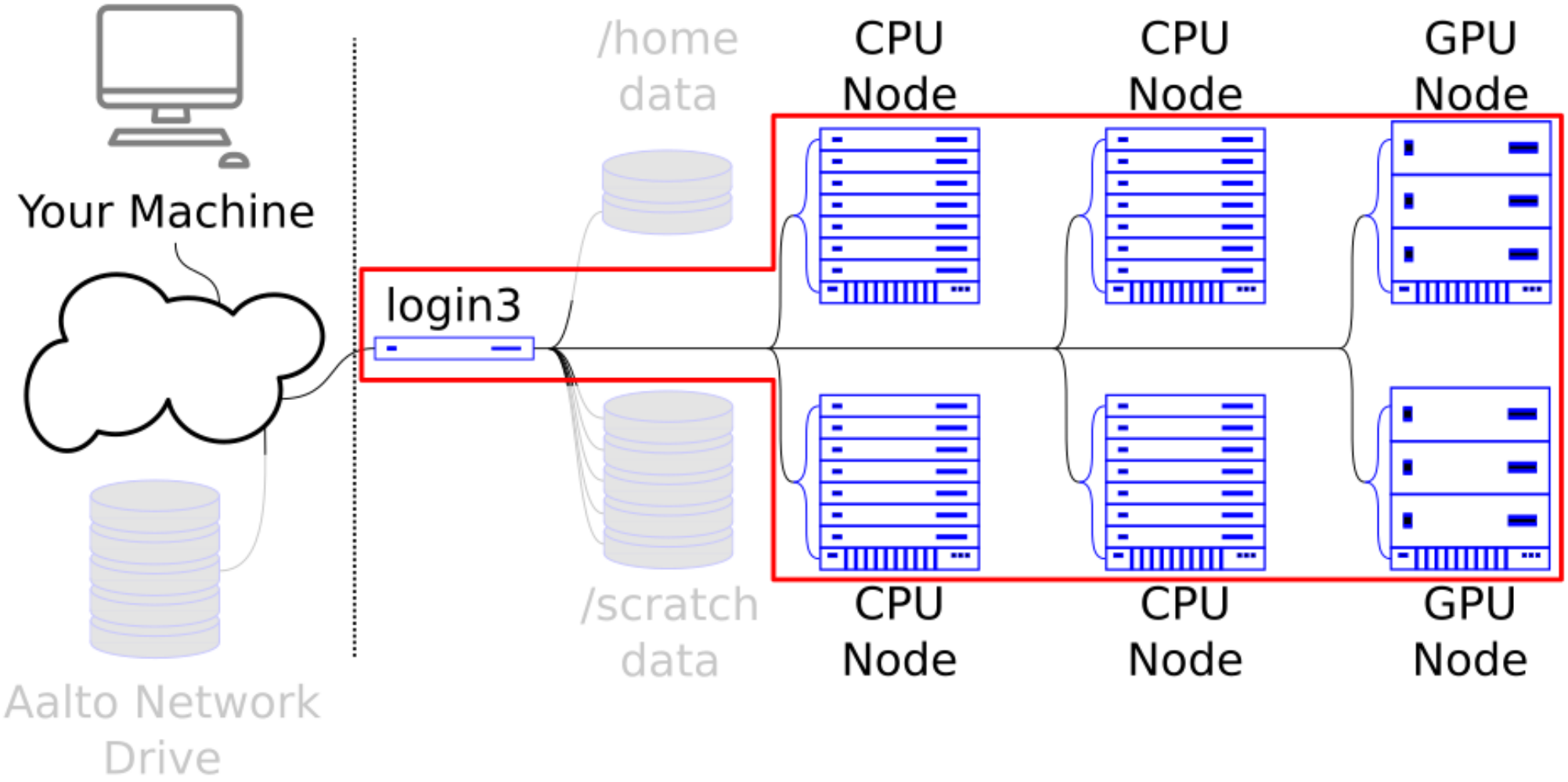
	MG-Rast (Meyer et al., 2008)	Anvio (Eren et al., 2015)	Smash community (Arumugam et al., 2010)	Humann (Abubucker et al., 2012)	fmap (Kim et al., 2016)	MetaWrap (Uritskiy et al., 2018)	Samsa2 (Westreich et al., 2018)	IMP (Narayanasamy et al., 2016)	SqueezeMeta
Assembly	No	No	Yes	No	No	Yes	No	Yes	Yes
Data source	Reads or contigs	Contigs	Contigs	Reads	Reads or contigs	Contigs	Reads (RNA)	Reads	Reads
Gene prediction	Yes	Yes	Yes	No	No	No	No	Yes	Yes
Function assignment	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
RNA assignment	Yes	Yes	No	No	No	No	Yes	Yes	Yes
Taxonomic assignment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gene abundances	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Metagomic comparison	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Co-assembly	No	No	No	No	No	Yes	No	Yes	Yes
Binning	No	Support	No	No	No	Yes	No	Yes	Yes
Bin validation	No	Yes	No	No	No	No	No	No	Yes
Local Installation	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

https://doi.org/10.3389/fmicb.2018.03349



1. Assembly
2. RNA prediction and classification
3. ORF (CDS) prediction
4. Homology searching against taxonomic and functional databases
5. Hmmer searching against Pfam database
6. Taxonomic assignment of genes
7. Functional assignment of genes
8. Blastx on parts of the contigs with no gene prediction or no hits
9. Taxonomic assignment of contigs, and check for taxonomic disparities
10. Coverage and abundance estimation for genes and contigs
11. Estimation of taxa abundances
12. Estimation of function abundances
13. Merging of previous results to obtain the ORF table
14. Binning with different methods
15. Binning integration with DAS tool
16. Taxonomic assignment of bins, and check for taxonomic disparities
17. Checking of bins with CheckM
18. Merging of previous results to obtain the bin table
19. Merging of previous results to obtain the contig table
20. Prediction of kegg and metacyc patwhays for each bin
21. Final statistics for the run

Running Sqeezemeta in SLURM



Running Sqeezemeta in SLURM

To run the pipeline in basic mode you need to download:

397G ./db

of databases

- #!/bin/bash
- #SBATCH -J squeeze.sh
- #SBATCH --cpus-per-task=256
- #SBATCH --mem=512gb
- #SBATCH --time=05:00:00
- #SBATCH --constraint=cal
- #SBATCH --error=pipe.%J.err
- #SBATCH --output=pipe.%J.out

- module load squeezemeta/1.6.3

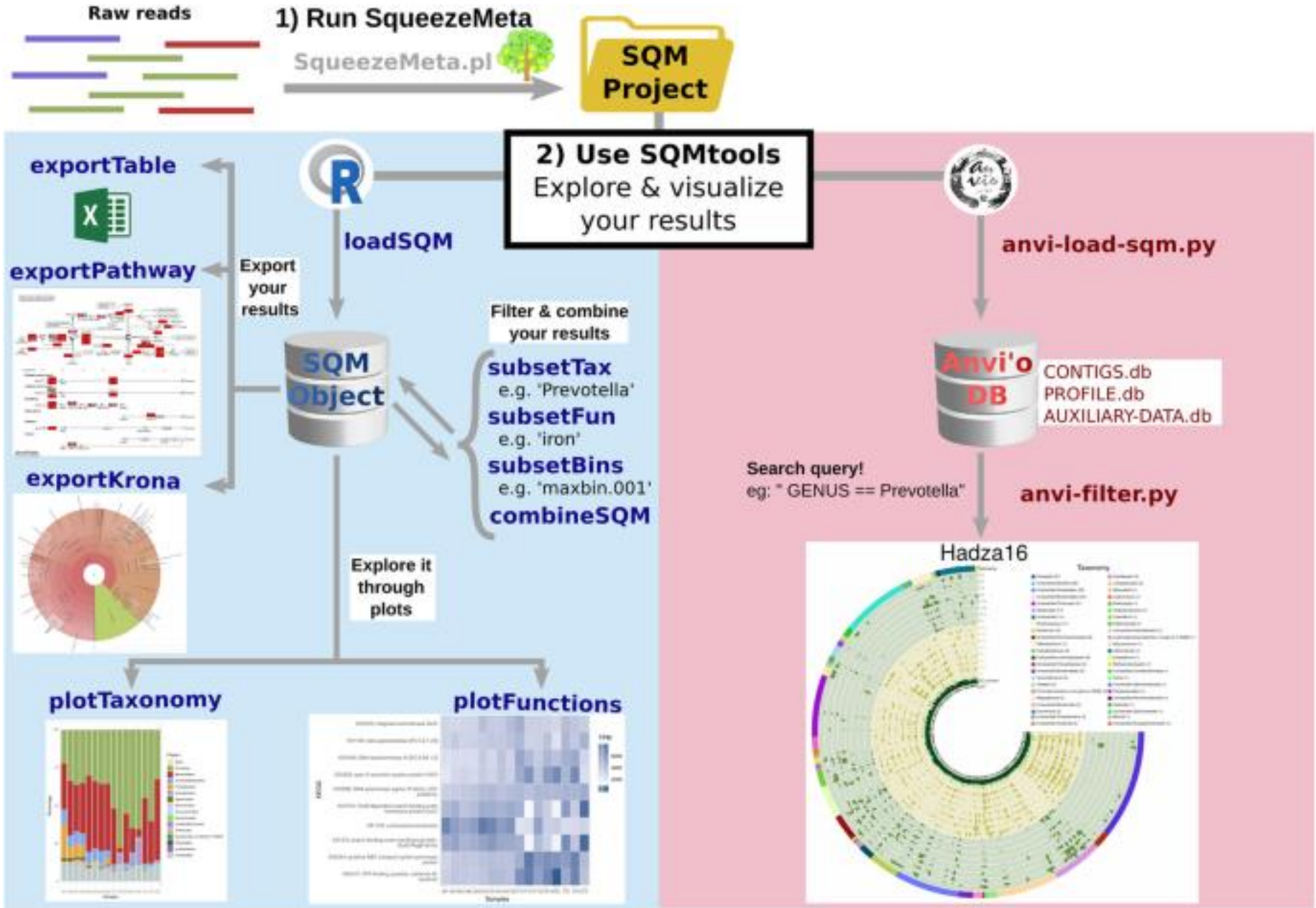
- SqueezeMeta.pl -m coassembly \
- -s samples.txt -f data \
- -t 128 -p metagenome_assembly

The output of Sqeezemeta

- Logs
- Assembly
- Bins
- …Tables

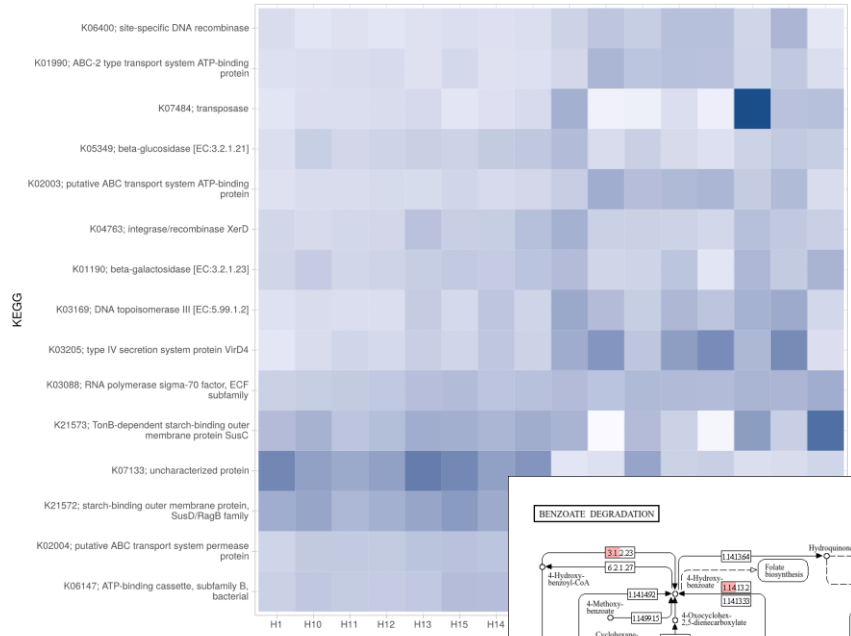
```
drwxr-xr-x  9 adoss  staff  288 Mar  5 17:25 do not open
-rwxr-xr-x@ 1 adoss  staff  168 Mar  5 17:47 spades.sh
adoss@Adams-Laptop meta_exercise % cd ..
adoss@Adams-Laptop meta_exc % ll
total 2220432
drwxr-xr-x  14 adoss  staff         448 Mar  5 17:24 data not used
-rw-r--r--@  1 adoss  staff         253 Mar  5 15:00 meta_excercise.Rproj
drwxr-xr-x   7 adoss  staff         224 Mar  5 17:47 meta_exercise
-rw-r--r--   1 adoss  staff  1136851089 Mar  5 17:29 meta_exercise.zip
drwxr-xr-x@ 14 adoss  staff         448 Mar  5 17:24 quast_megahit
drwxr-xr-x@ 14 adoss  staff         448 Mar  5 17:24 spades_megahit
drwxr-xr-x   8 adoss  staff         256 Mar  6 07:26 sqeezemeta_results
-rw-r--r--   1 adoss  staff         195 Mar  5 17:37 thingstoinstall.txt
adoss@Adams-Laptop meta_exc % cd sqeezemeta_results
adoss@Adams-Laptop sqeezemeta_results % ll
total 64
drwxr-xr-x  15 adoss  staff    480 Mar  6 07:25 metagenome_assembly
-rw-r--r--@  1 adoss  staff    458 Mar  5 12:47 pipe.10654274.err
-rw-r--r--@  1 adoss  staff  17376 Mar  5 12:58 pipe.10654274.out
-rw-r--r--@  1 adoss  staff    126 Mar  5 11:19 samples.txt
-rw-r--r--@  1 adoss  staff    308 Mar  5 11:22 squeeze.sh
adoss@Adams-Laptop sqeezemeta_results % less samples.txt
```

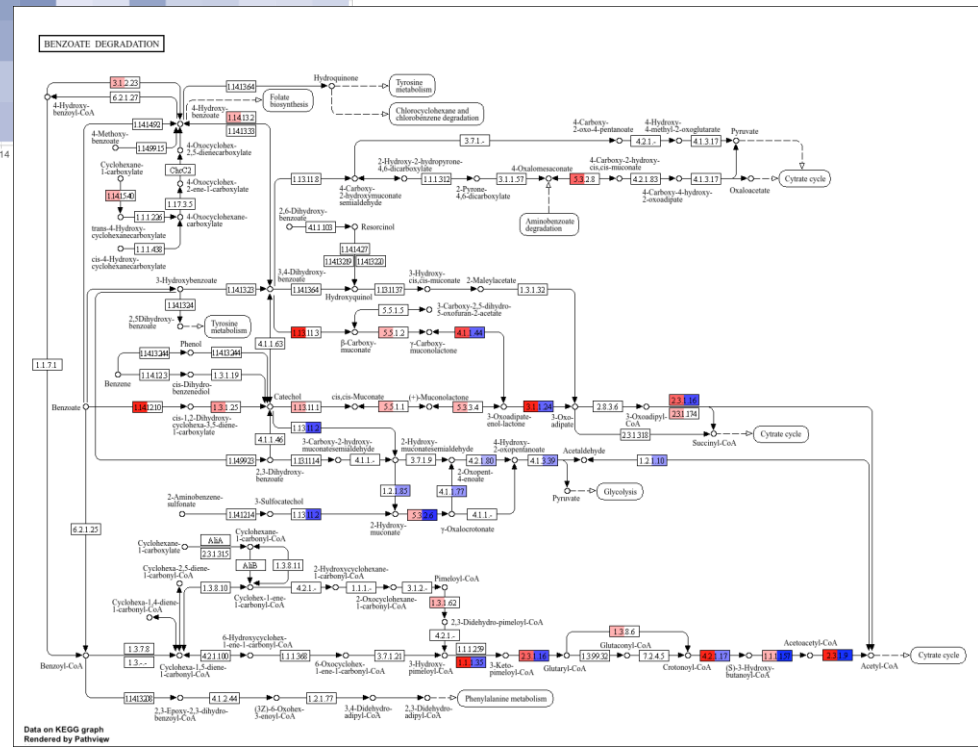
SQMtools


Raw reads
1) Run SqueezeMeta
SqueezeMeta.pl
SQM Project
2) Use SQMtools
Explore & visualize your results
exportTable
exportPathway
exportKrona
Export your results
loadSQM
SQM Object
Filter & combine your results
subsetTax
e.g. 'Prevotella'
subsetFun
e.g. 'iron'
subsetBins
e.g. 'maxbin.001'
combineSQM
Explore it through plots
plotTaxonomy
plotFunctions
anvi-load-sqm.py
Anvi'o DB
CONTIGS.db
PROFILE.db
AUXILIARY-DATA.db
Search query!
eg: " GENUS == Prevotella"
anvi-filter.py
Hadza16

SQM

lvl1	lvl2	lvl3	type	rows/names	columns	data
$orfs	$table		dataframe	orfs	misc. data	misc. data
	$abund		numeric matrix	orfs	samples	abundances (reads)
	$bases		numeric matrix	orfs	samples	abundances (bases)
	$cov		numeric matrix	orfs	samples	coverages
	$cpm		numeric matrix	orfs	samples	coverages per million of reads
	$tpm		numeric matrix	orfs	samples	tpm
	$seqs		character vector	orfs	(n/a)	sequences
	$tax		character matrix	orfs	tax. ranks	taxonomy
$contigs	$table		dataframe	contigs	misc. data	misc. data
	$abund		numeric matrix	contigs	samples	abundances (reads)
	$bases		numeric matrix	contigs	samples	abundances (bases)
	$cov		numeric matrix	contigs	samples	coverages
	$cpm		numeric matrix	contigs	samples	coverages per million of reads
	$tpm		numeric matrix	contigs	samples	tpm
	$seqs		character vector	contigs	(n/a)	sequences
	$tax		character matrix	contigs	tax. ranks	taxonomies
	$bins		character matrix	contigs	bin. methods	bins
$bins	$table		dataframe	bins	misc. data	misc. data
	$length		numeric vector	bins	(n/a)	length
	$abund		numeric matrix	bins	samples	abundances (reads)
	$percent		numeric matrix	bins	samples	percentages (reads)
	$bases		numeric matrix	bins	samples	abundances (bases)
	$cov		numeric matrix	bins	samples	coverages
	$cpm		numeric matrix	bins	samples	coverages per million of reads
	$tax		character matrix	bins	tax. ranks	taxonomy
$taxa	$superkingdom	$abund	numeric matrix	superkingdoms	samples	abundances (reads)
		$percent	numeric matrix	superkingdoms	samples	percentages (reads)
	$phylum	$abund	numeric matrix	phyla	samples	abundances (reads)
		$percent	numeric matrix	phyla	samples	percentages (reads)
	$class	$abund	numeric matrix	classes	samples	abundances (reads)
		$percent	numeric matrix	classes	samples	percentages (reads)
	$order	$abund	numeric matrix	orders	samples	abundances (reads)
		$percent	numeric matrix	orders	samples	percentages (reads)
	$family	$abund	numeric matrix	families	samples	abundances (reads)
		$percent	numeric matrix	families	samples	percentages (reads)
	$genus	$abund	numeric matrix	genera	samples	abundances (reads)
		$percent	numeric matrix	genera	samples	percentages (reads)
	$species	$abund	numeric matrix	species	samples	abundances (reads)
		$percent	numeric matrix	species	samples	percentages (reads)
$functions	$KEGG	$abund	numeric matrix	KEGG ids	samples	abundances (reads)
		$bases	numeric matrix	KEGG ids	samples	abundances (bases)
		$cov	numeric matrix	KEGG ids	samples	coverages
		$cpm	numeric matrix	KEGG ids	samples	coverages per million of reads
		$tpm	numeric matrix	KEGG ids	samples	tpm
		$copy_number	numeric matrix	KEGG ids	samples	avg. copies
	$COG	$abund	numeric matrix	COG ids	samples	abundances (reads)
		$bases	numeric matrix	COG ids	samples	abundances (bases)
		$cov	numeric matrix	COG ids	samples	coverages
		$cpm	numeric matrix	COG ids	samples	coverages per million of reads
		$tpm	numeric matrix	COG ids	samples	tpm
		$copy_number	numeric matrix	COG ids	samples	avg. copies
	$PFAM	$abund	numeric matrix	PFAM ids	samples	abundances (reads)
		$bases	numeric matrix	PFAM ids	samples	abundances (bases)
		$cov	numeric matrix	PFAM ids	samples	coverages
		$cpm	numeric matrix	PFAM ids	samples	coverages per million of reads
		$tpm	numeric matrix	PFAM ids	samples	tpm
		$copy_number	numeric matrix	PFAM ids	samples	avg. copies
$total_reads			numeric vector	samples	(n/a)	total reads
$misc	$project_name		character vector	(empty)	(n/a)	project name
	$samples		character vector	(empty)	(n/a)	samples
	$tax_names_long	$superkingdom	character vector	short taxa names	(n/a)	long taxa names
		$phylum	character vector	short taxa names	(n/a)	long taxa names
		$class	character vector	short taxa names	(n/a)	long taxa names
		$order	character vector	short taxa names	(n/a)	long taxa names
		$family	character vector	short taxa names	(n/a)	long taxa names
		$genus	character vector	short taxa names	(n/a)	long taxa names
		$species	character vector	short taxa names	(n/a)	long taxa names
	$tax_names_short		character vector	long taxa names	(n/a)	short taxa names
	$KEGG_names		character vector	KEGG ids	(n/a)	KEGG names
	$KEGG_paths		character vector	KEGG ids	(n/a)	KEGG hierarchy
	$COG_names		character vector	COG ids	(n/a)	COG names
	$COG_paths		character vector	COG ids	(n/a)	COG hierarchy
	$ext_annot_sources		character vector	(empty)	(n/a)	External databases





https://github.com/jtamames/SqueezeMeta/wiki/Using-R-to-analyze-your-SQM-results

Preparation:

- Download the data
- Install SQMtools library in R
- Install “pathview” library in R (NOT “pathviewr”)

- `BiocManager::install("pathview")`

- Install assembler - spades and quast in one conda environment

- `conda create -n genome_assembly -c bioconda \ spades quast`
- `conda activate genome_assembly`

Tasks:

- Assemble the data using spades
- Check the quality of the assembly using quast

- Backup: Check which assembly is better using quast

- Read the results of SqueezeMeta pipeline in R and :
  - Check the bins and decide which are good and which not
  - Choose your 3 favorite function, plot taxonomy related to them and pathways