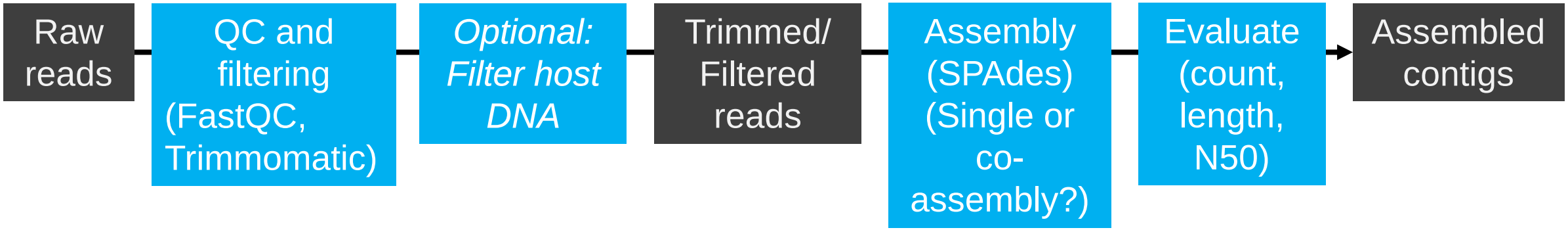
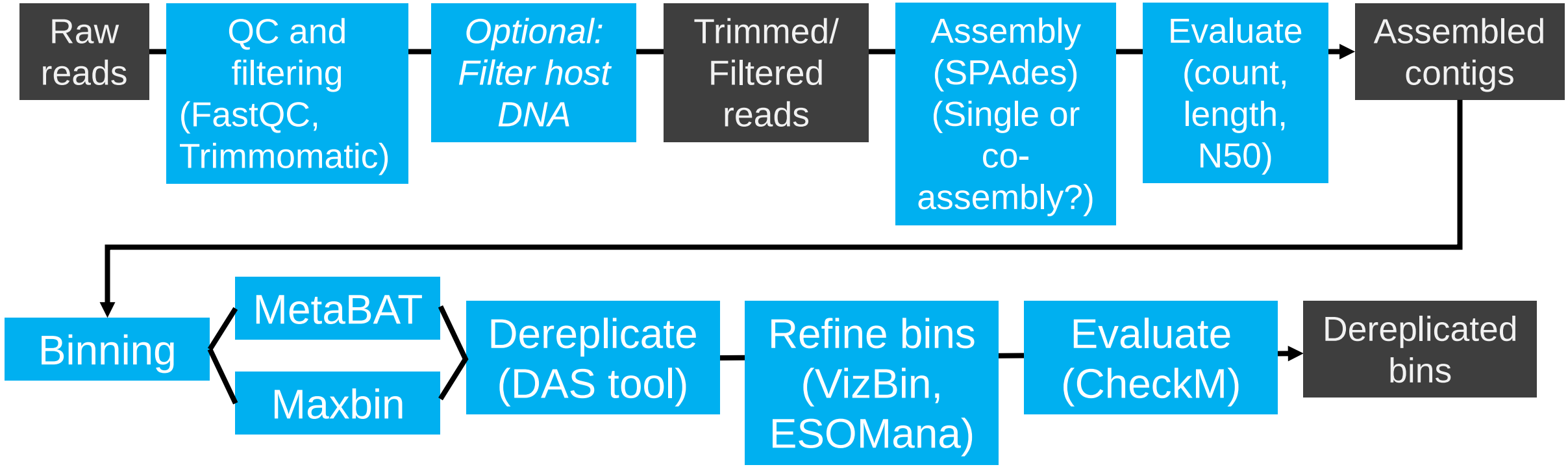
Day 4

Genome annotation (continued)
Report findings
Presentation of data

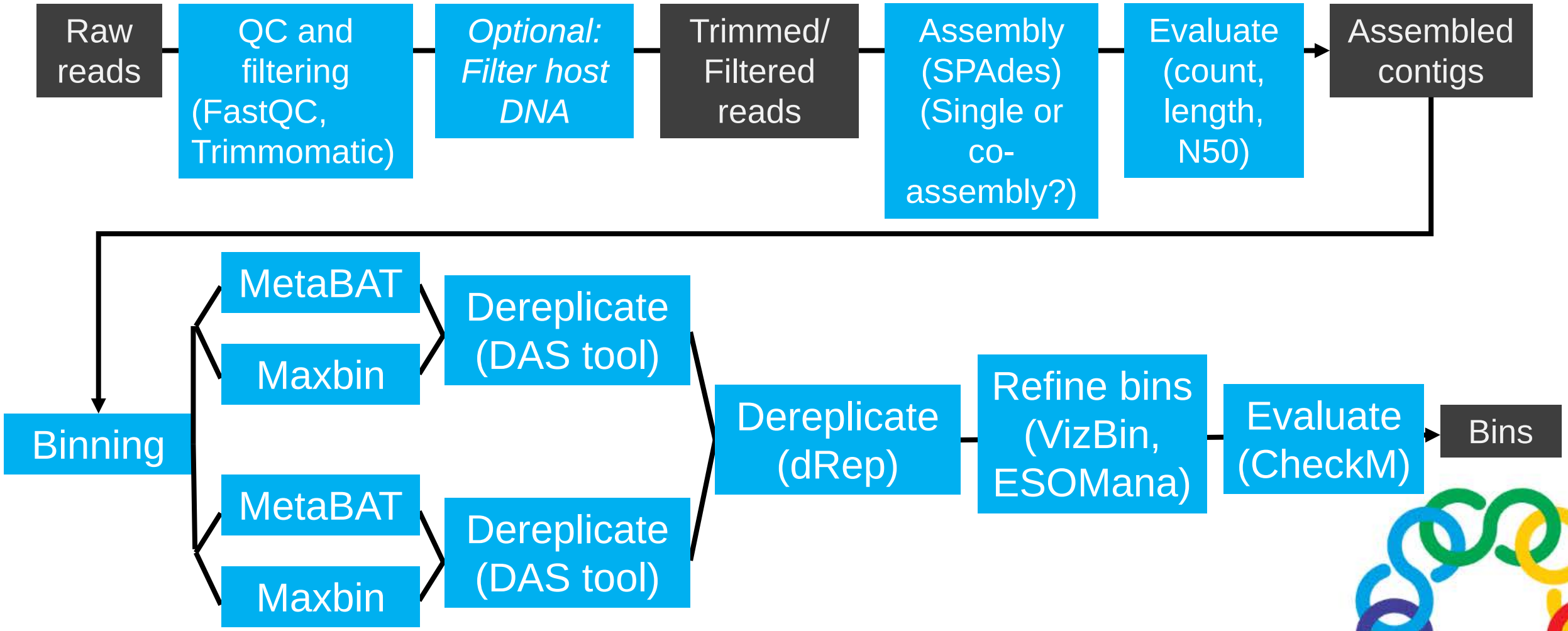
Workshop overview



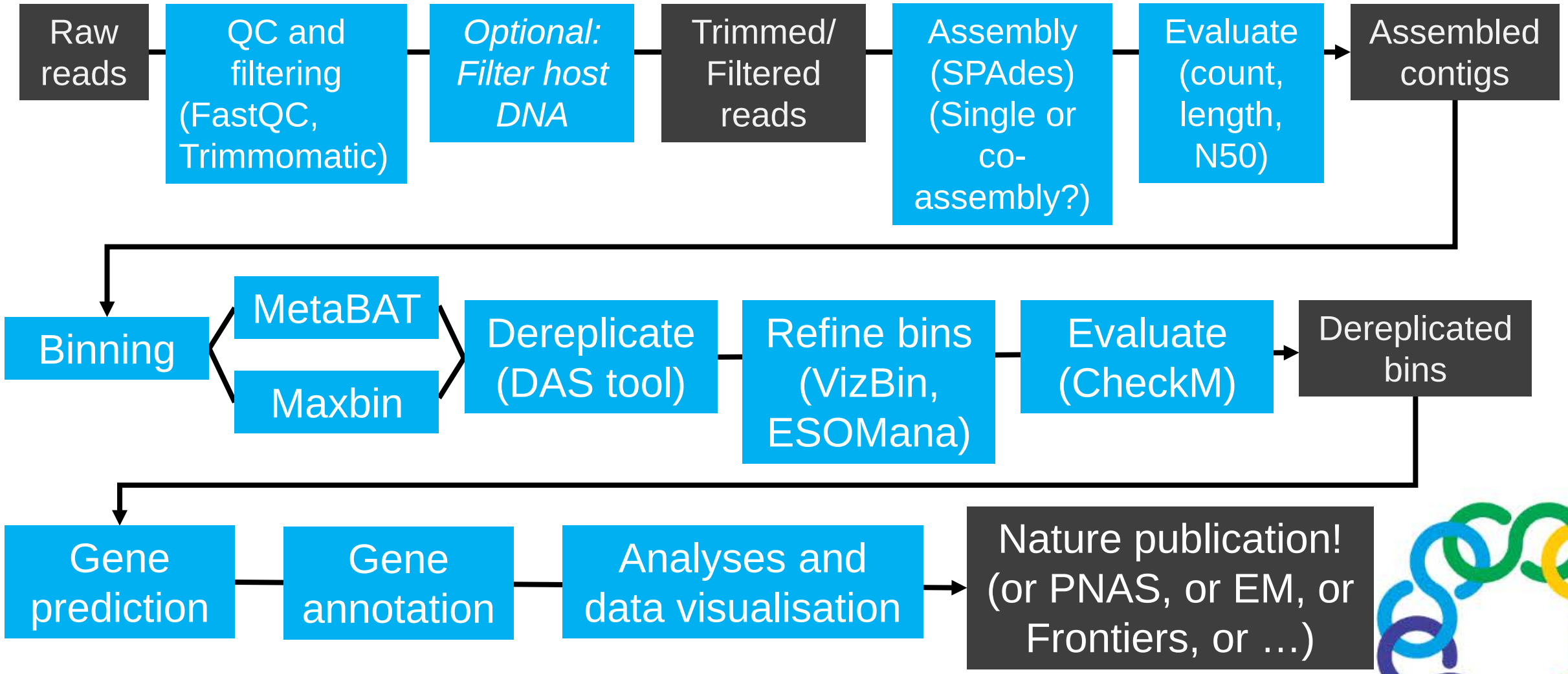
Workshop overview



Workshop overview

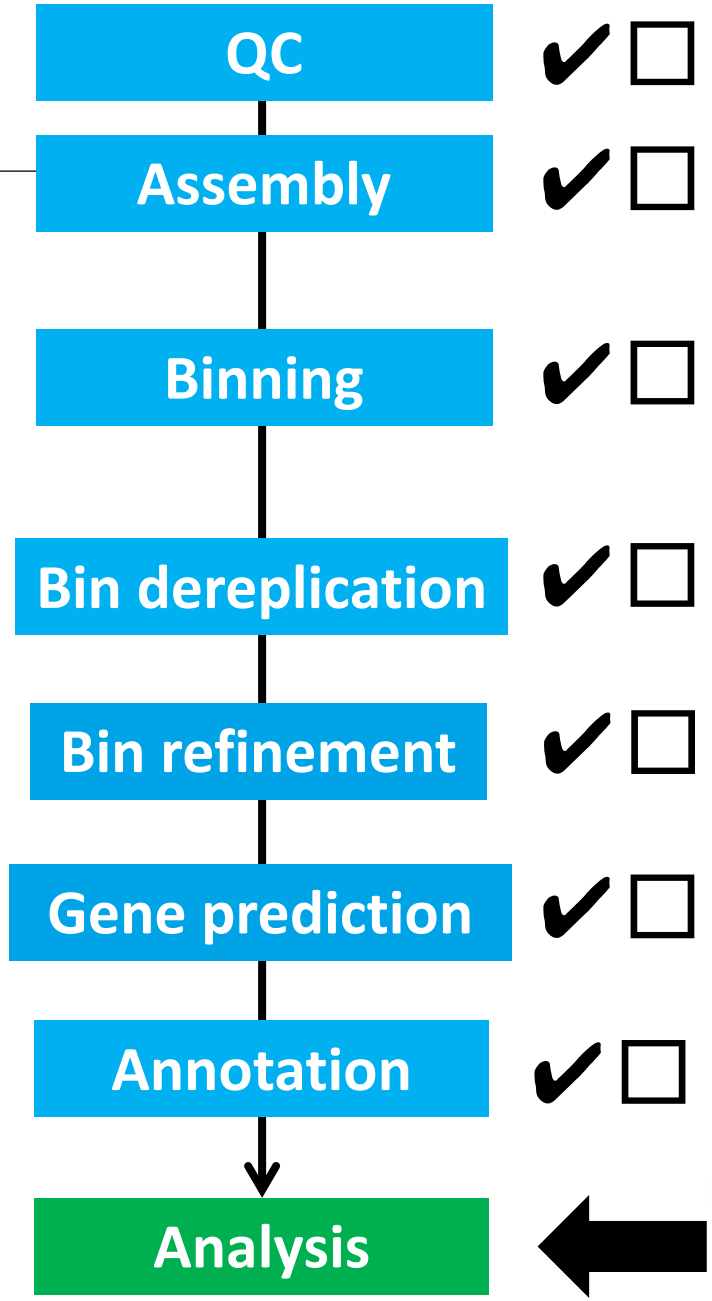


Workshop overview



Day overview

- **Goals:**
  - **Gene annotations (continued)**
  - **Visualizing metagenomic data**
  - **Group task**



DRAM

DRAM annotation

Distilling and Refining Annotations of Metabolism (DRAM; Shaffer et al. 2020. Nucleic Acids Research 48(16)**)**

- Tool for gene prediction and gene annotation of MAGs (DRAM-v for viruses)
  - Functional annotation:
    - BLAST-style searches:
      - KEGG (if provided),
      - UniRef 90 (if desired)
      - MEROPS
    - HMM searches
      - Kofam, Pfam, dbCAN2 (CAZy)
      - VOGDB
    - tRNAs and rRNAs also detected

DRAM annotation

Distilling and Refining Annotations of Metabolism (DRAM; Shaffer et al. 2020. Nucleic Acids Research 48(16)**)**

- Genome annotations to metabolic functions in three levels:

1. RAW
Each gene nucleotide and amino acid sequence with annotations

	fasta	scaffold	gene_positio	start_positio	end_position	strandednes	rank	kegg_id	kegg_hit	uniref_id	uniref_hit	uniref_taxon	uniref_RBH
bin_0_1f935	bin_0	1f9359e86e6	1	205	1371	1	B	K02338	DNA polyme	Q7V9E7_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	2	1375	2151	1	B			Q7V9E6_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	3	2191	4593	1	B	K23269	phosphoribos	PURL_PROM	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	4	4653	6110	1	B	K00764	amidophospl	Q7TV87_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	5	6146	8635	-1	B			A0A163AH7(	UniRef90_A(	Cyanobacteri	TRUE
bin_0_1f935	bin_0	1f9359e86e6	6	8713	9606	-1	B			Q7V9E3_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	7	9616	10590	-1	B	K18979	epoxyqueuos	Q7V9E2_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	8	10677	11291	1	B			Q7V9E1_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	9	11363	12112	1	B			Q7V9E0_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	10	12142	12777	1	B	K03625	transcription	A0A162EFM(	UniRef90_A(	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	11	12777	14231	1	B	K03110	fused signal	Q7V9D8_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	12	14355	15698	1	B	K07315	phosphoserir	A0A163R2M	UniRef90_A(	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	13	15728	17140	1	B	K01755	argininosucc	ARLY_PROM	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	14	17264	17872	1	C			Q7V9D6_PR(	UniRef90_Q	Cyanobacteri	FALSE
bin_0_1f935	bin_0	1f9359e86e6	15	17882	18886	-1	B	K05539	tRNA-dihydr	A0A163N6K6	UniRef90_A(	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	16	18956	19462	1	C	K07305	peptide-met	A0A163N6J4	UniRef90_A(	Prochlorococ	FALSE
bin_0_1f935	bin_0	1f9359e86e6	17	19434	20711	1	B			Q7V9D3_PR(	UniRef90_Q	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	18	20686	21966	-1	B	K02653	type IV pilus	Q7V9D2_PR(	UniRef90_Q	Cyanobacteri	TRUE
bin_0_1f935	bin_0	1f9359e86e6	19	21983	23059	-1	B	K02669	twitching mo	A0A163N6E5	UniRef90_A(	Prochlorococ	TRUE
bin_0_1f935	bin_0	1f9359e86e6	20	23070	24887	-1	B			Q7V9D0_PR(	UniRef90_Q	Prochlorococ	TRUE

DRAM annotation

Distilling and Refining Annotations of Metabolism (DRAM; Shaffer et al. 2020. Nucleic Acids Research 48(16)**)**

- Genome annotations to metabolic functions in three levels:

2. DISTILLATE
Taxonomy (GTDB-tk), quality statistics (checkM), and key metabolisms summarized by genome

9	K02303	uroporphy	Siroheme b	Antibiotic Resistance		1	1	3	1	1	1	0	1	1	1
0	K02304	precorrin-2	Siroheme b	Antibiotic Resistance		0	1	1	1	1	0	0	1	1	1
1	K02492	glutamyl-tf	Siroheme b	Antibiotic Resistance		1	1	1	1	1	0	1	1	1	1
2	K02496	uroporphy	Siroheme b	Antibiotic Resistance		0	0	1	1	1	0	0	0	0	0
3	K03794	sirohydroc	Siroheme b	Antibiotic Resistance		0	0	0	0	0	0	0	0	0	0
4	K13542	uroporphy	Siroheme b	Antibiotic Resistance		0	0	0	0	0	0	0	0	1	1
5	K13543	uroporphy	Siroheme b	Antibiotic Resistance		0	0	0	0	0	0	0	0	0	0
6	K14163	glutamyl-tf	Siroheme b	Antibiotic Resistance		0	0	0	0	0	0	0	0	0	0
7	K07464	cas4; CRISP	Subtype I-A	CRISPR	Type I CRISI	0	0	0	0	0	1	0	0	1	2
8	K07725	csa3; CRISP	Subtype I-A	CRISPR	Type I CRISI	0	0	0	0	0	0	0	0	0	0
9	K19074	csa2; CRISP	Subtype I-A	CRISPR	Type I CRISI	0	0	0	0	0	0	0	0	0	0
0	K19075	cst2, cas7;	Subtype I-A	CRISPR	Type I CRISI	0	0	0	0	0	0	0	1	0	1

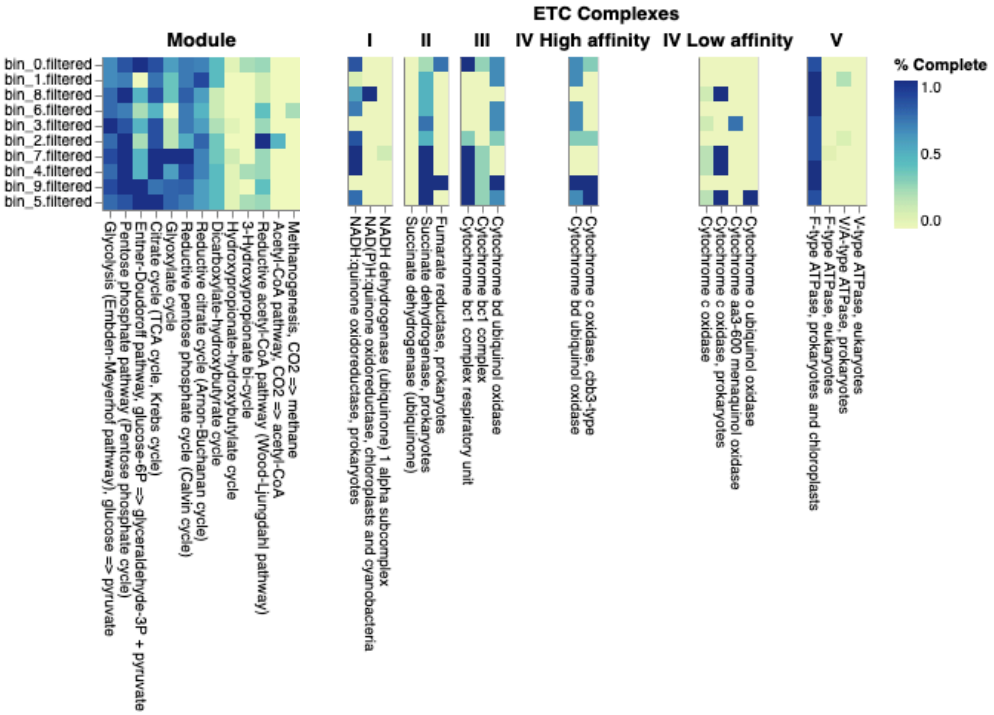
MISC | carbon utilization | Transporters | Energy | Organic Nitrogen | rRNA | tRNA | +

DRAM annotation

Distilling and Refining Annotations of Metabolism (DRAM; Shaffer et al. 2020. Nucleic Acids Research 48(16)**)**

- Genome annotations to metabolic functions in three levels:

3. PRODUCT
Interactive heatmap of key metabolic functions by genome

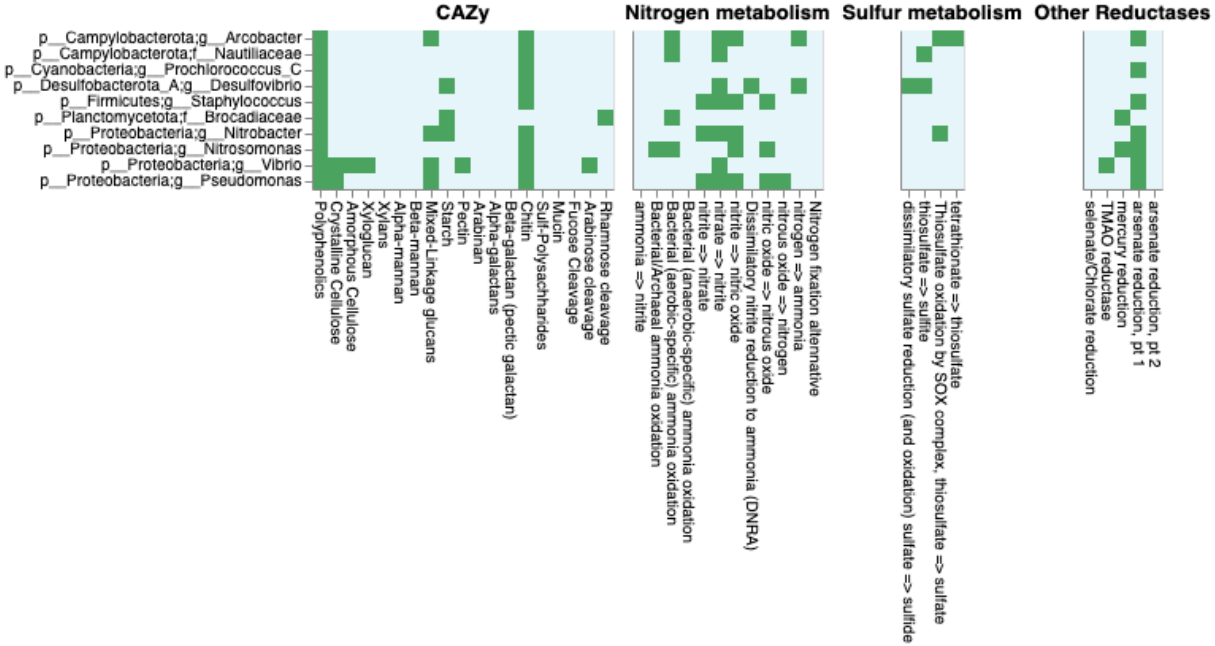


DRAM annotation

Distilling and Refining Annotations of Metabolism (DRAM; Shaffer et al. 2020. Nucleic Acids Research 48(16)**)**

- Genome annotations to metabolic functions in three levels:

3. PRODUCT
Interactive heatmap of key metabolic functions by genome



Presentation of data

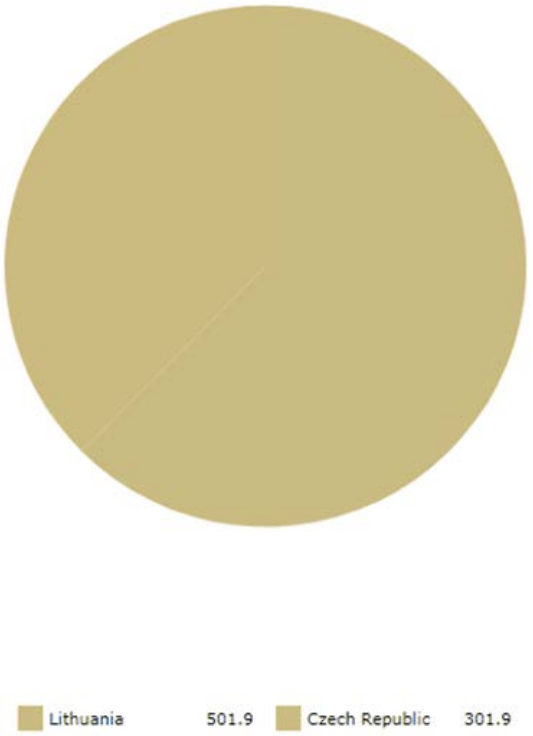
Presentation of data

How do we report/present our data?

1. **Heatmaps of bin and viral contig coverage across samples**
2. **Ordinations to investigate relatedness of samples**
3. **KEGG pathway maps**
4. **Gene synteny analysis**
5. **Heatmaps of genomic features**
6. **Inference of gene trees**
7. **Creating metabolic schematics**

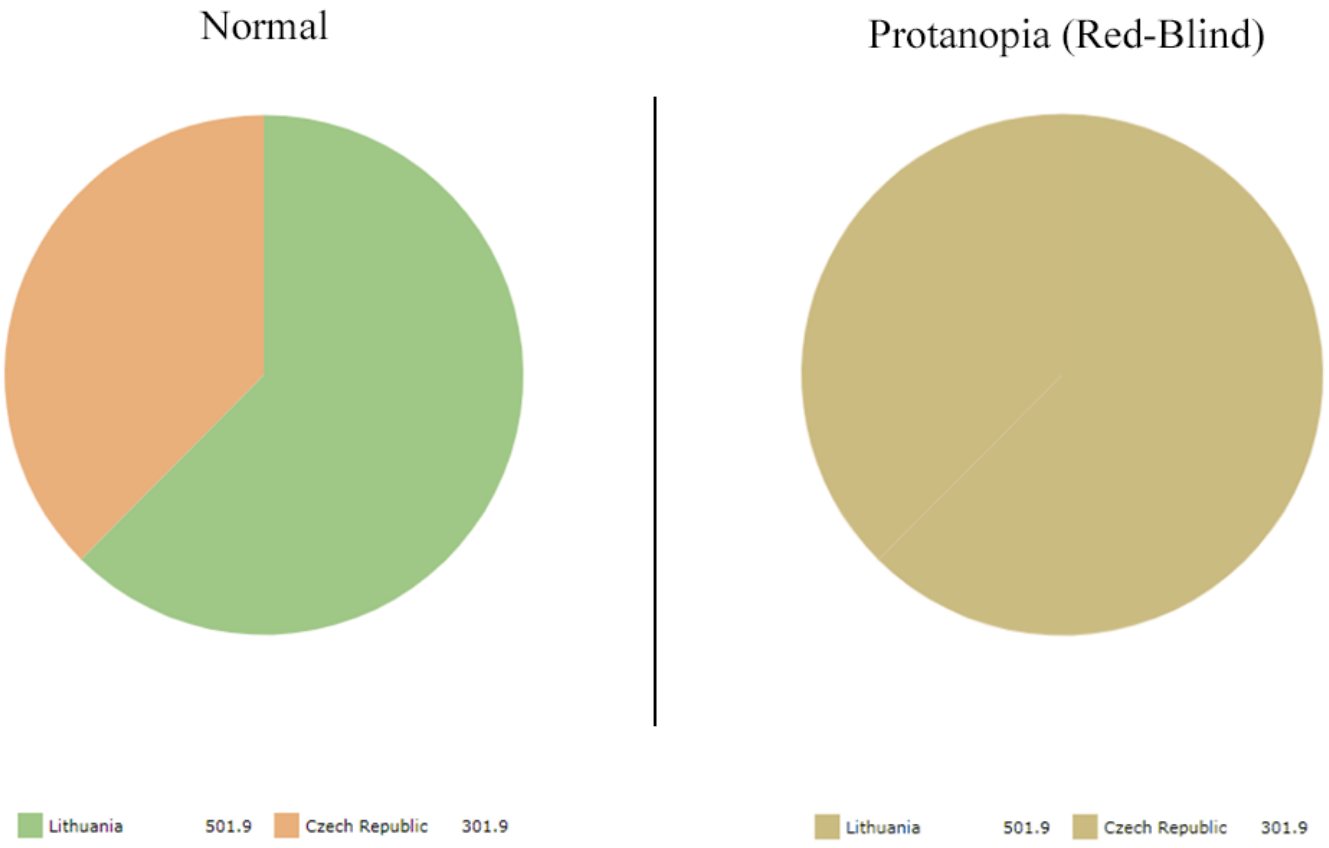
Data visualisation and accessibility

The fundamental point of data visualisation is *communication*

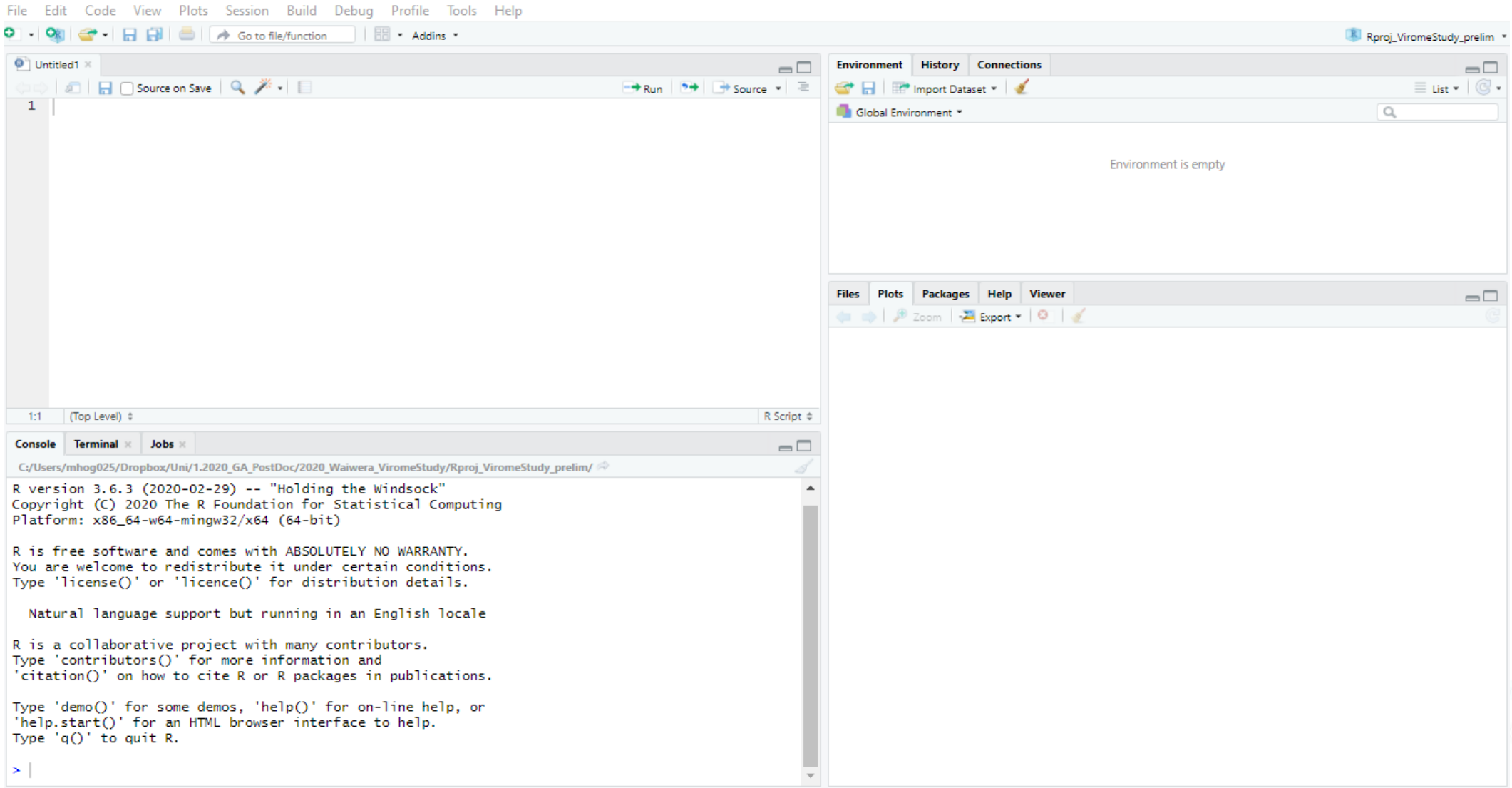


Data visualisation and accessibility

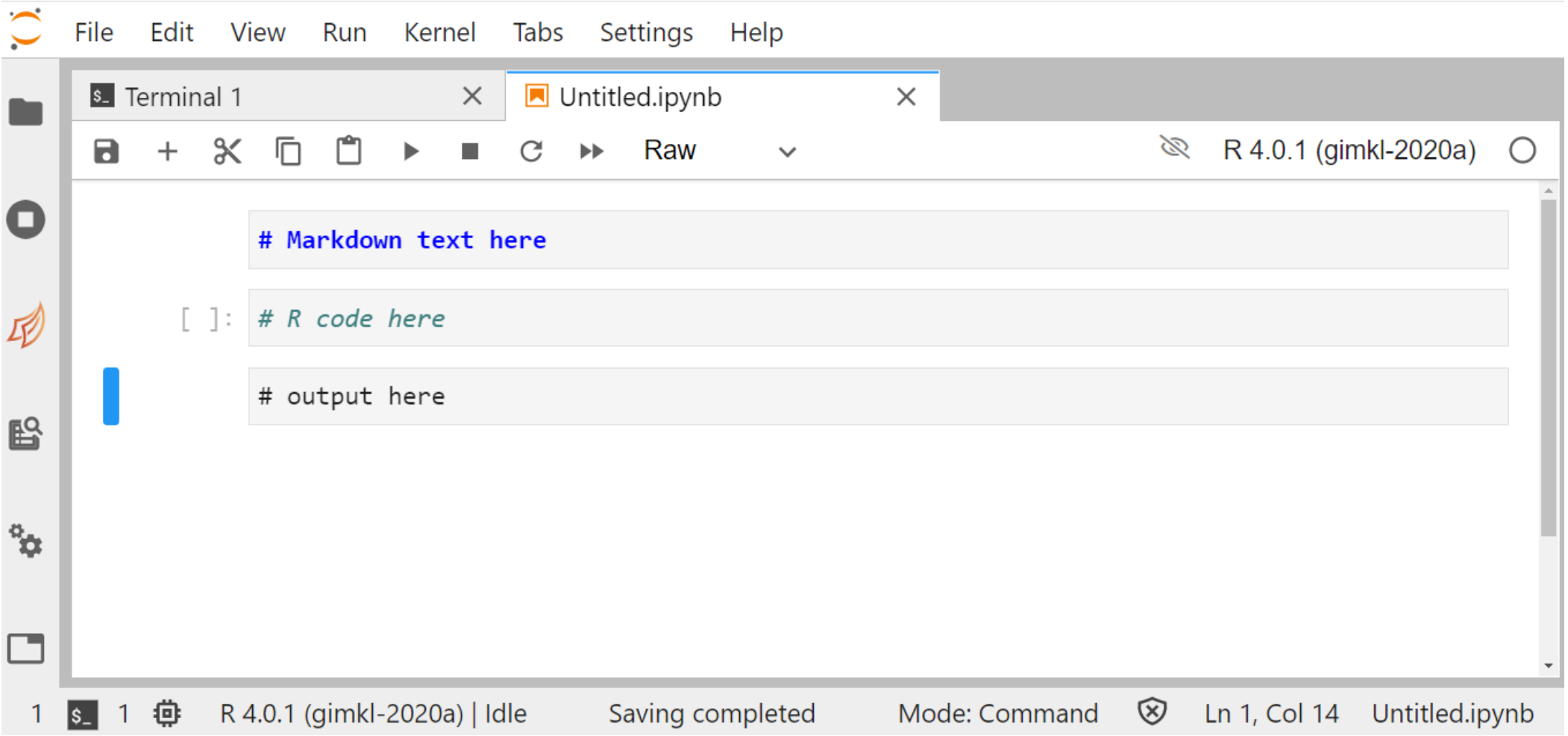
The fundamental point of data visualisation is *communication*



Presentation of data: R

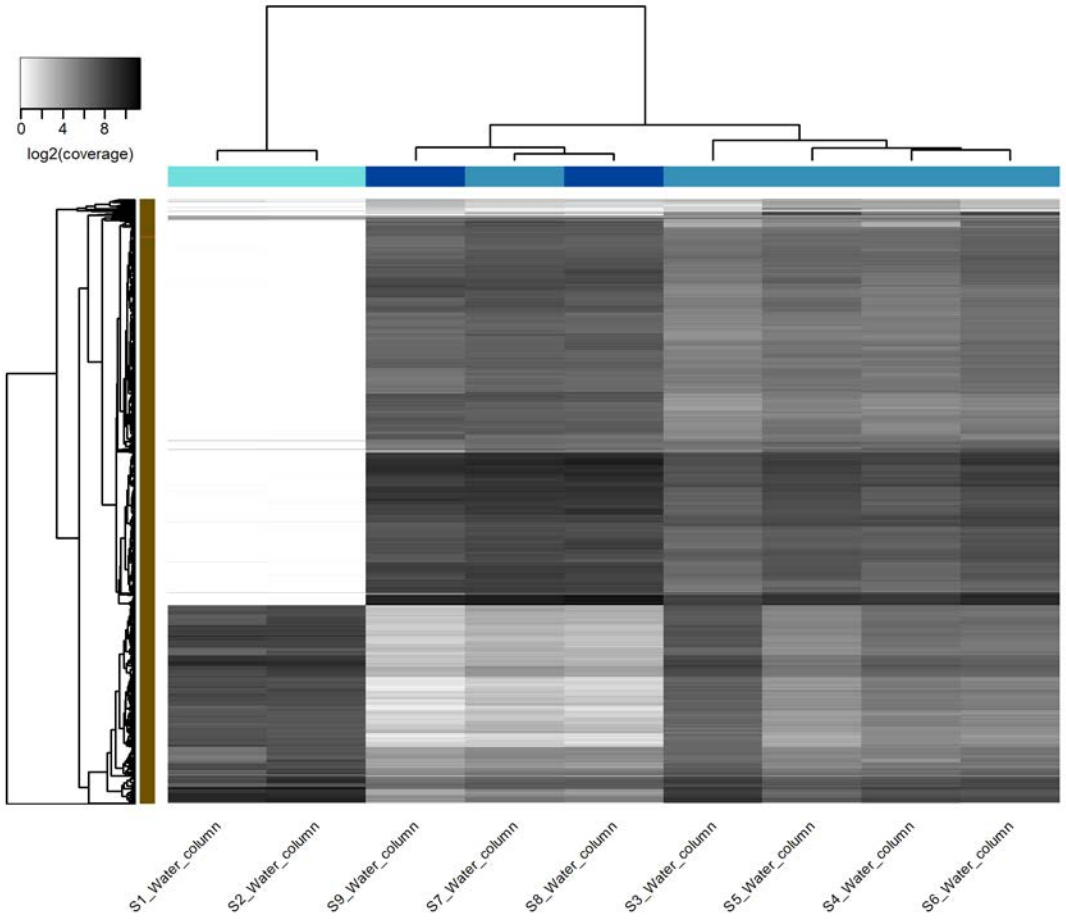


Presentation of data: R in Jupyter Lab



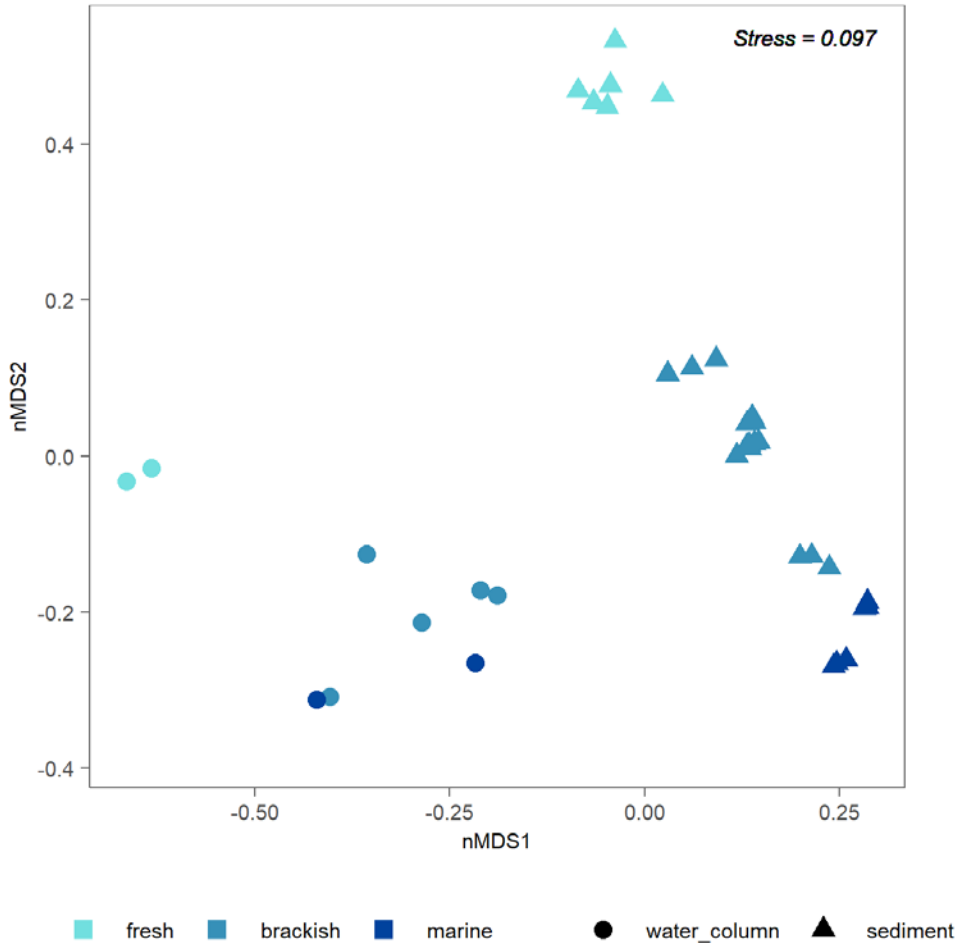
Presentation of data

1. **Heatmaps of bin and viral contig coverage across samples**



Presentation of data

2. *Optional:* Ordinations to investigate relatedness of samples



Task: Presentation of data

- **Build per-sample coverage heatmaps**
- ***Optional: Build nMDS ordination plots***

Presentation of data

3. KEGG pathway maps

Task: Presentation of data

- Build KEGG pathway maps for nitrogen metabolism

Presentation of data

4. Gene synteny analysis

A more informative view of gene content

- Shows local gene context
- More detailed than reporting gene table
- Sometimes absence of genes from operon is biologically informative

Presentation of data

4. Gene synteny analysis

Gene coordinates

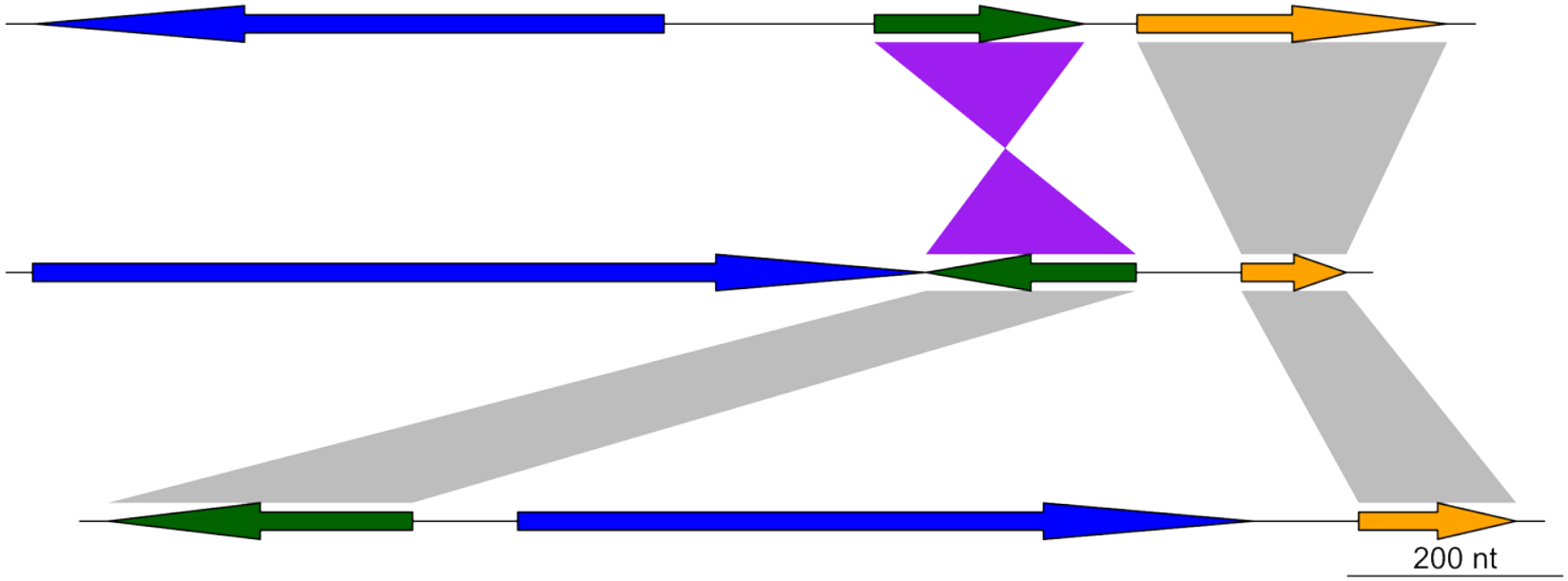
- Prodigal
- BLAST

Wrangling

- R (tidyr)
- R (dplyr)

Presentation

- R (genePlotR)



Task: Presentation of data

- Build gene synteny plots for sulfur assimilation

Presentation of data

5. Heatmaps of genomic features

Simple figure to display complex data tables

- *M* genomes x *N* features in one place
- Presence/absence or relative abundance (multi-copy)
- Fixed layout, or clustering by patterns

Presentation of data

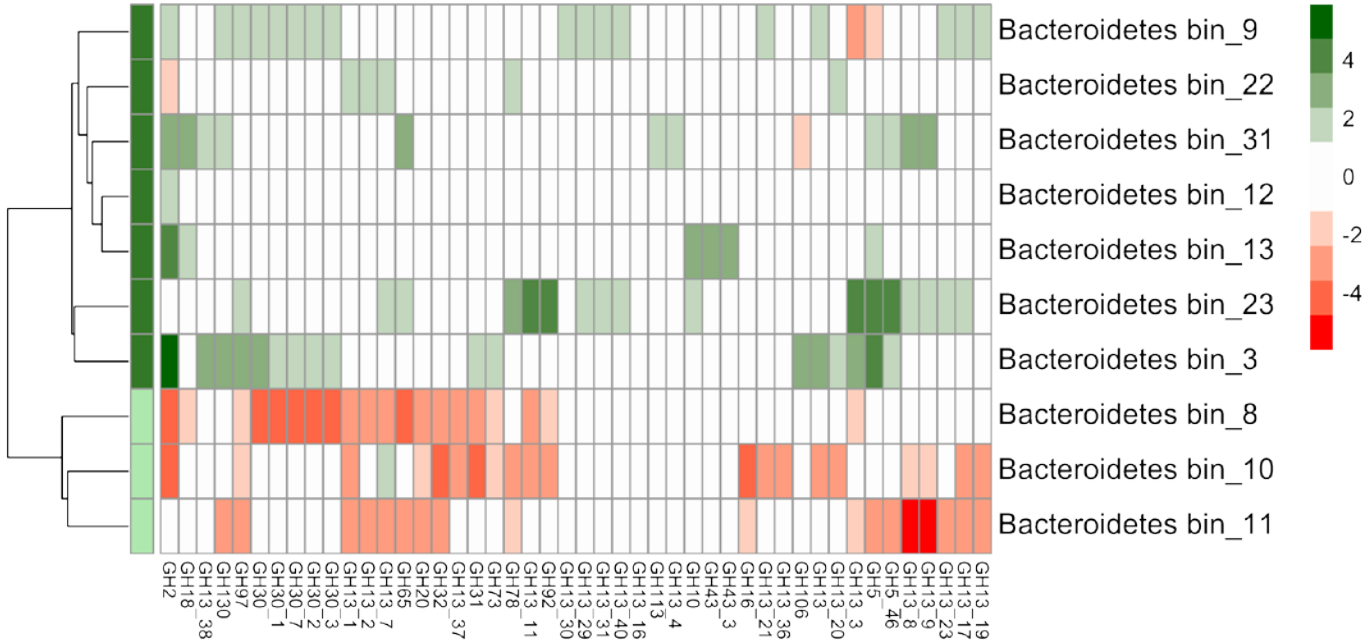
5. Heatmaps of genomic features

Feature table
- R (dplyr)

Preprocessing
- R (vegan)

Wrangling
- R (tidyr)
- R (tibble)

Presentation
- R (pheatmap)



Task: Presentation of data

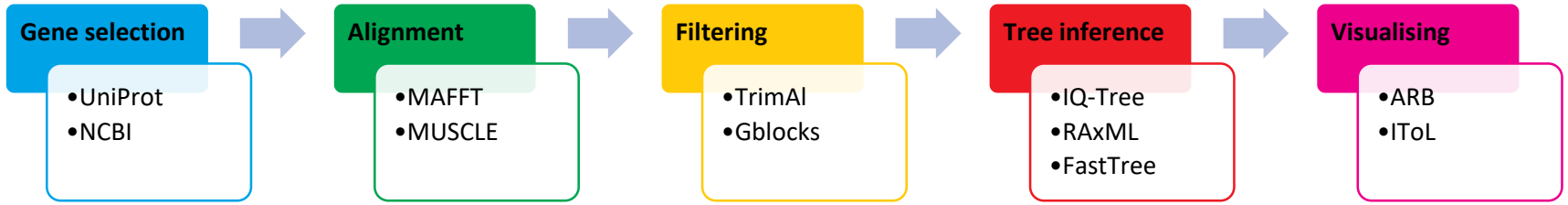
- ***Optional: Build a heatmap of CAZy annotations***

Presentation of data

6. Inference of gene trees

Gene trees are a great way to present

- Confirmation of annotation
- Novelty of detection / horizontal gene transfer
- Rate of evolution in the feature



Presentation of data

6. Inference of gene trees

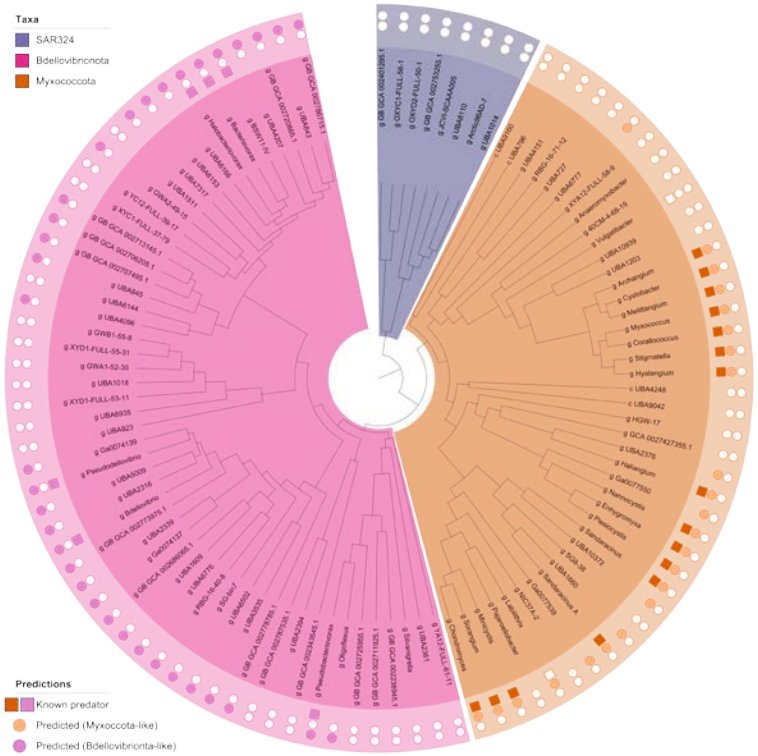
Gene selection
- UniProt
- NCBI

Alignment
- MAFFT
- MUSCLE

Filtering
- TrimAl
- Gblocks

Tree inference
- IQ-Tree
- RAxML
- FastTree

Visualising
- ARB
- IToL



Presentation of data

7. Creating metabolic schematics

Summarise the entire core metabolism of an organism into a single figure

Handley et al. (2014) doi:10.1111/1462-2920.12453

Presentation of data

7. Creating metabolic schematics

Summarise the entire core metabolism of an organism into a single figure

What's the magic tool for producing these?

- Illustrator, Inkscape, GIMP (and a lot of time)
- Use tools for picking colour schemes
  - ColorBrewer2 (http://colorbrewer2.org)
  - IWantHue (https://medialab.github.io/iwanthue/)

Task: Prep for group presentations

- **Use the white board for illustrations**
- **Things to include:**
  - The attribute you found
  - Details about the attribute
  - The organism(s) you found it in
  - A brief explanation of biological relevance
  - The tools and annotations you used
  - Anything else?

Presentation of findings

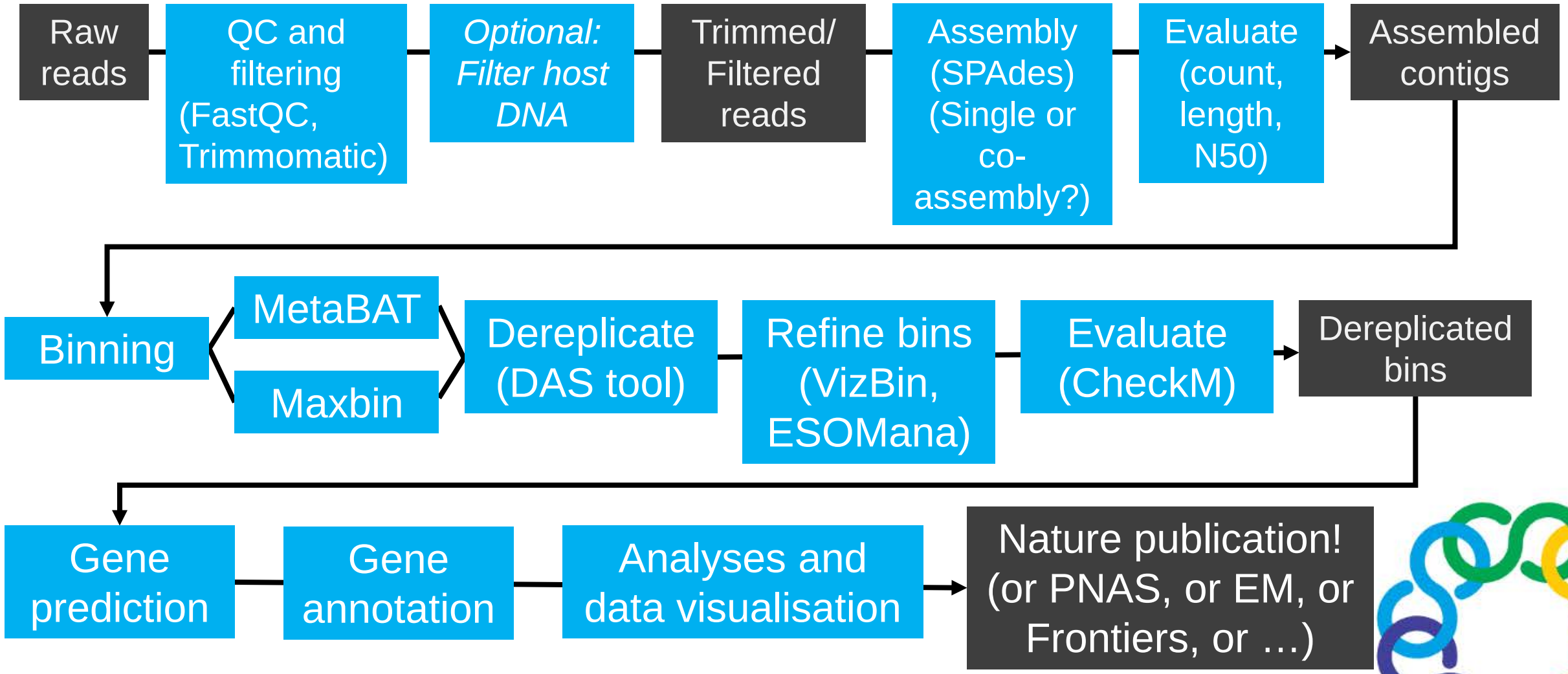
Task: Report findings

Report your findings with regards to your objective

1. What did you find?
2. How did you make the discoveries?
   - Functional prediction only
   - Functional prediction and taxonomic context
3. Each group present for 5 mins each (max!)

Wrap up and Q/A

Workshop overview



Genomics Aotearoa - Resources

Genomics Aotearoa – GitHub repositories

https://github.com/GenomicsAotearoa/

- Metagenomics Summer School material
- RNA seq workshop
- Environmental metagenomics
  - Metagenomic annotation and binning
- Methods and musings
  - Bin cluster refinement
  - Genome assembly ont
  - Metagenomic ont

Genomics Aotearoa - Resources

genomics aotearoa

Search

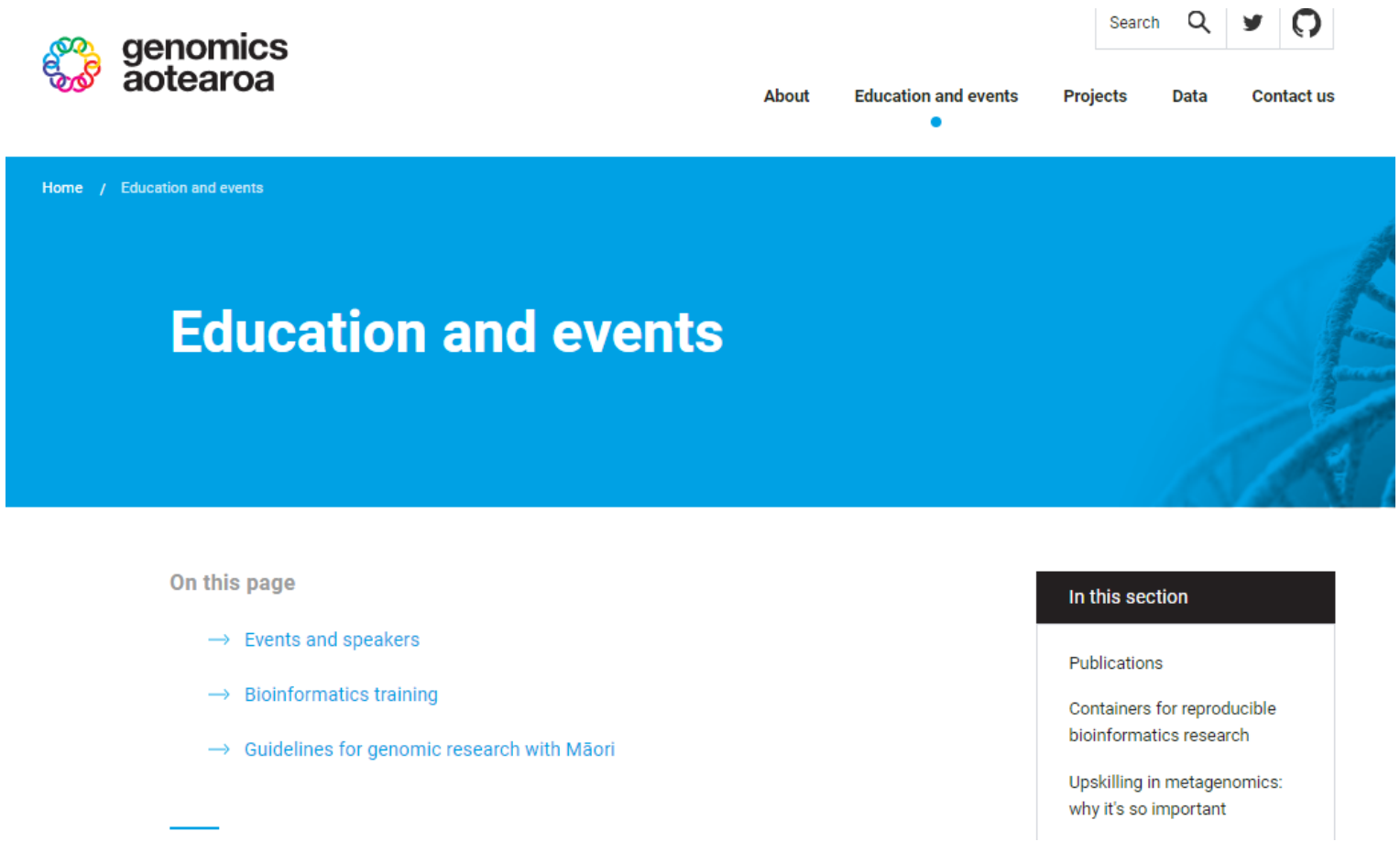
About | Education and events | Projects | Data | Contact us

Genomics Aotearoa seminar series

We are holding a free presentation series via zoom, with plenty of opportunities for questions and discussion. To attend an upcoming seminar, or get materials from previous seminars, email genomics.aotearoa@otago.ac.nz.

Other workshops

https://www.genomics-aotearoa.org.nz/education-events



Wrap up and Q/A