BIOINforMAtics

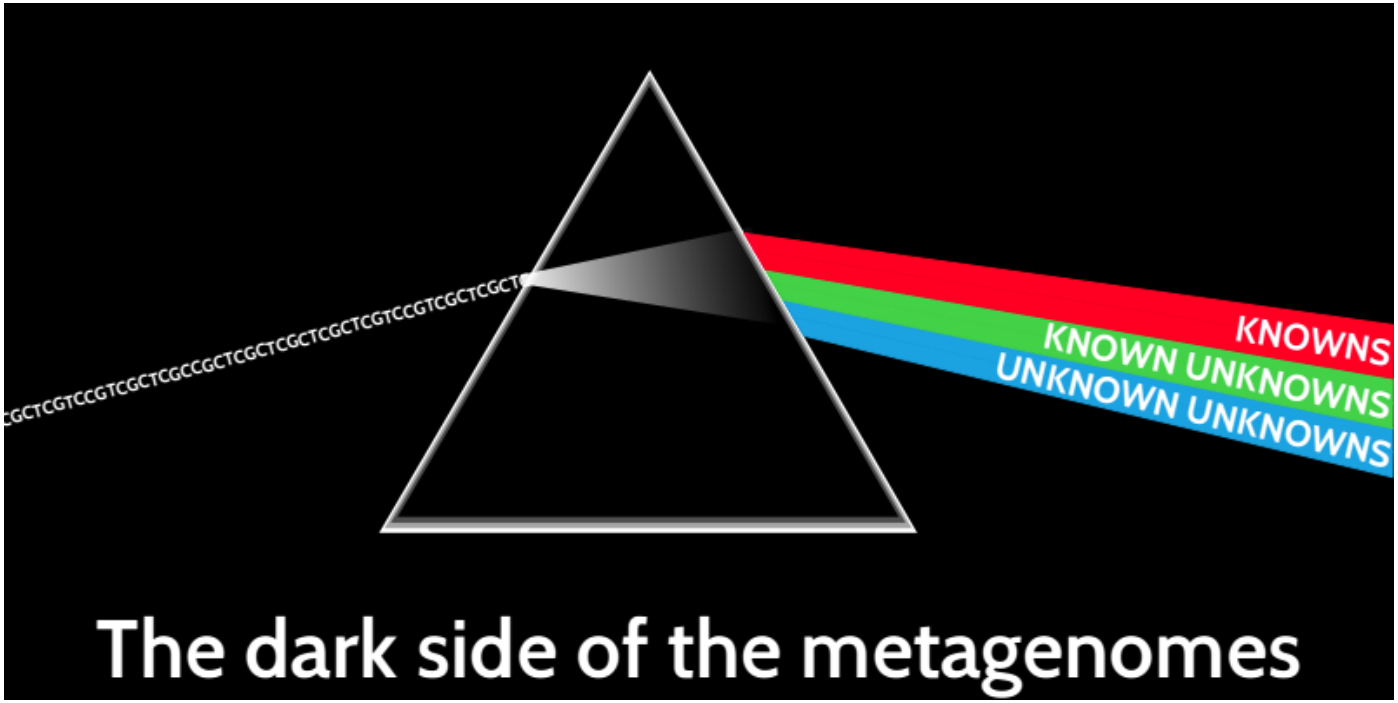
Seminar

Stazione Zoologica Anton Dohrn • Napoli •

Metagenomics: a toolshed to illuminate the dark sides of diversity

Michael Tangherlini, PhD

michael.tangherlini@szn.it

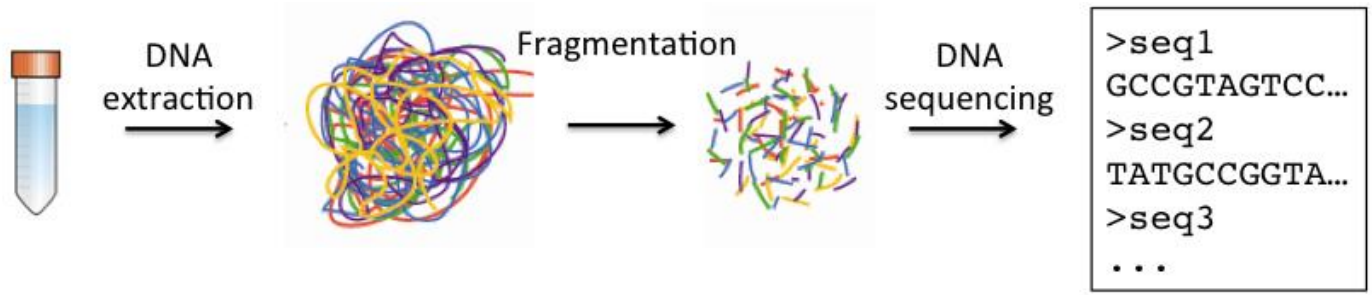


Metagenomics' 3Ws

- What: (shotgun) metagenomics as a strategy for the characterization of natural communities as a whole
- Why: (shotgun) metagenomics as a tool to obtain useful information on the composition of a natural assemblage from a taxonomic and functional point of view
- When: (shotgun) metagenomics as an approach to deal with complex communities without relying on culturing

Needed inputs

- Sequence read files (e.g. FASTA and FASTQ)
- Produced with several technologies, both in-house @SZN (i.e. Ion Proton) and outside (e.g. Illumina)



Big Data, Bog Data

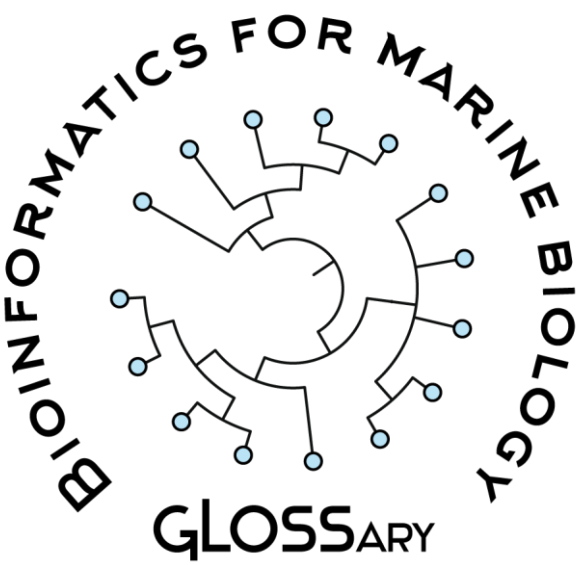
- First problem: huge amount of sequences forbids thorough analysis
- B*i*g Data -> B*o*g Data

Big Data, Bog Data

- The massive amounts of data provided by NGS sequencing would require an enormous computational power…
- …which we currently lack.

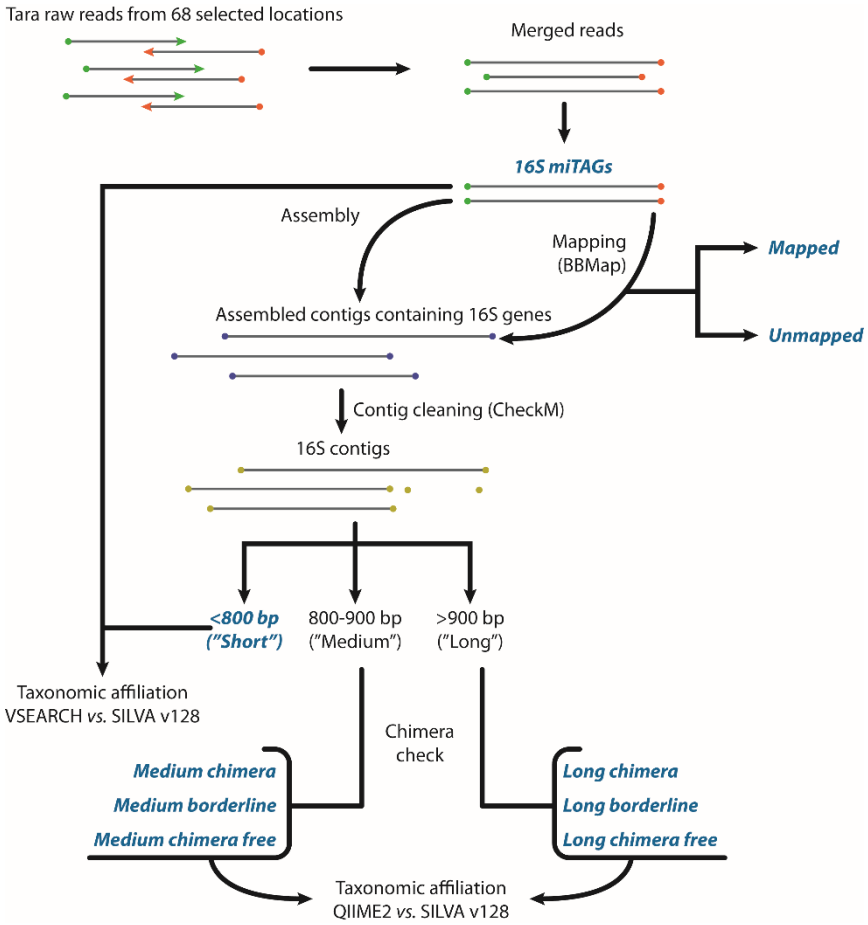
- But strategies can be employed to circumvent such issue
- E.g. subsampling, normalization and focusing on specific fractions of a bigger sample

GLOSSary



- GLOSSary: an explorable database of 16S genes reconstructed from the Tara Oceans project
- Prokaryotic 16S rDNA-related sequences were extracted from the prokaryotic fraction of the Tara Oceans samples – miTAGs
- We downloaded and assembled these raw sequences to produce longer (near full-length to full-length) gene sequences
- These were stored in a database together with sample and sequence metadata

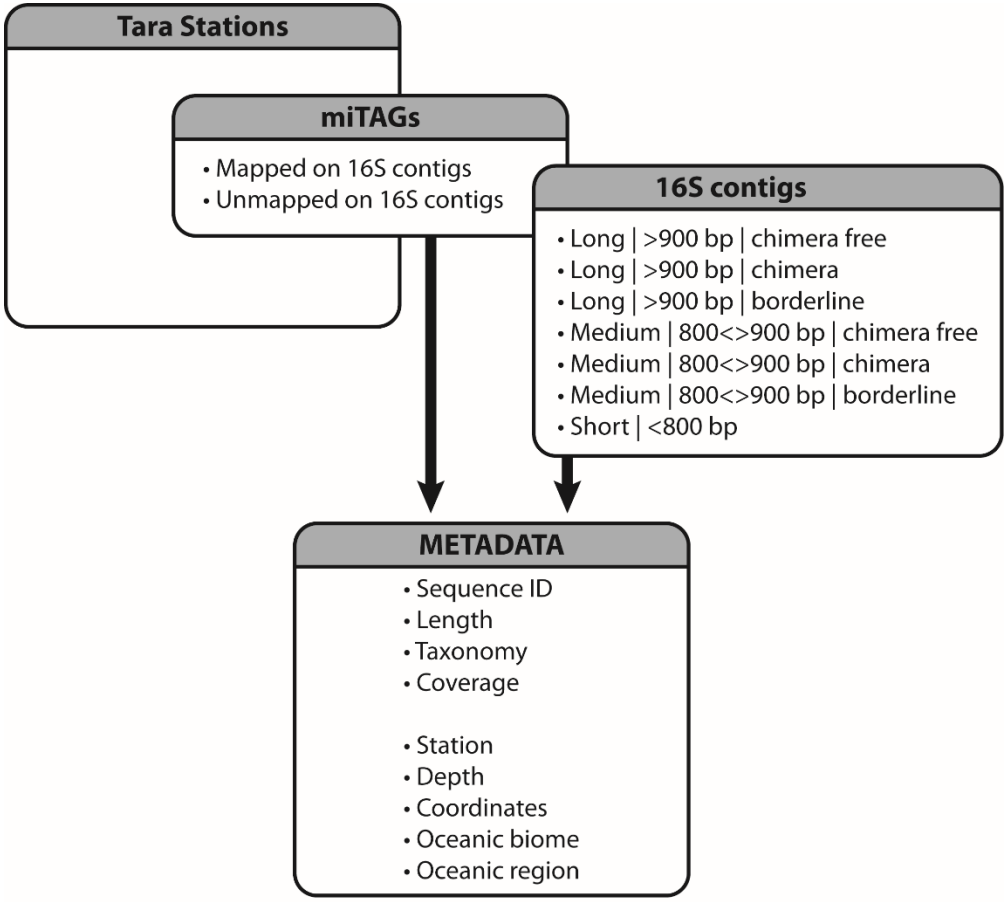
GLOSSary



- Pipeline used:
- MiTAG sequence download
- Sequence assembly and read mapping
- Length filtering (>900 bp, 800-900 bp, <800 bp)
- Chimera check and filter
- Taxonomic affiliation

GLOSSary

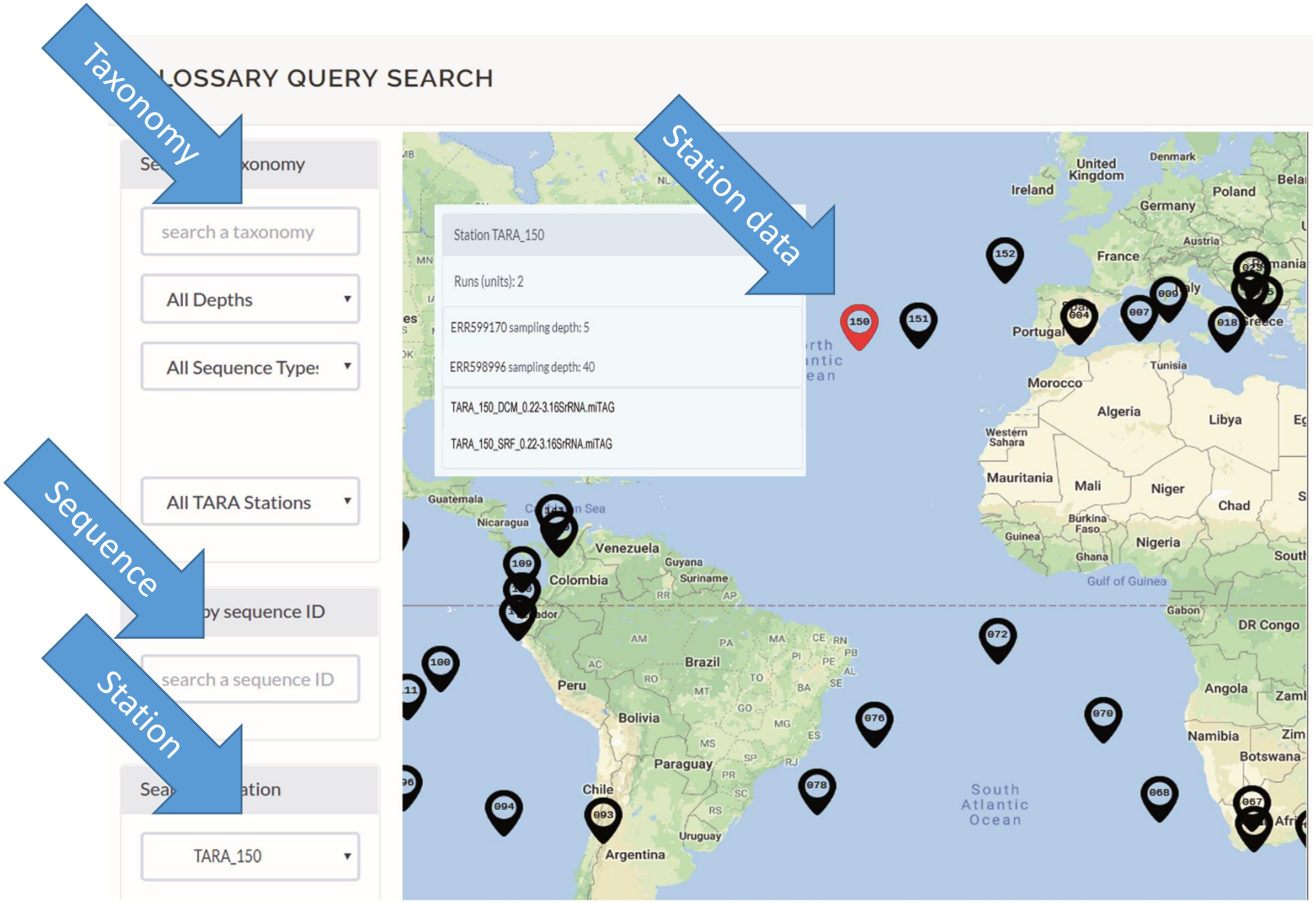
- Metadata within GLOSSary:
  - Sequences:
    - Mapped/unmapped
    - Sequence length
    - Chimeric/non-chimeric
    - Taxonomy
    - Coverage
  - Stations:
    - Sequence Ids
    - Depth
    - Coordinates
    - Biomes
    - Regions



GLOSSary

- Using GLOSSary:
  - Map exploration
    - Click on a station for further data
  - Keyword search
    - Type a taxonomic keyword (e.g. «Gammaproteobacteria»)
    - Retrieve information on distribution of associated sequences
  - BLAST search
    - Paste 16S rDNA sequence
    - BLAST against GLOSSary database + SILVA v128

GLOSSary


Taxonomy
Station data
Sequence
Station
LOSSARY QUERY SEARCH
xonomy
search a taxonomy
All Depths
All Sequence Type
All TARA Stations
by sequence ID
search a sequence ID
ation
TARA_150
Station TARA_150
Runs (units): 2
ERR599170 sampling depth: 5
ERR598996 sampling depth: 40
TARA_150_DCM_0.22-3.16SrRNA.miTAG
TARA_150_SRF_0.22-3.16SrRNA.miTAG
United Kingdom
Ireland
Denmark
Poland
Germany
France
Austria
Portugal
Tunisia
Morocco
Algeria
Libya
Western Sahara
Mauritania
Mali
Niger
Chad
Burkina Faso
Guinea
Nigeria
Ghana
Gulf of Guinea
Gabon
DR Congo
Angola
Namibia
Botswana
Guatemala
Nicaragua
Venezuela
Guyana
Suriname
Colombia
Peru
Brazil
Bolivia
Paraguay
Chile
Argentina
Uruguay
South Atlantic Ocean

GLOSSary

GLOSSARY QUERY SEARCH

Search by taxonomy **A**

alcaligenaceae

All Depths

All Sequence Types

All TARA Stations

Search by sequence ID **B**

TARA-004-DCM_215

Search by station

All TARA Stations

Tags	Hits	Download
contig | long | 900bp | chimeraFree	16	Get fasta
contig | long | 900bp | borderline	4	Get fasta
contig | long | 900bp | chimera	4	Get fasta
contig | medium | 800bp | chimeraFree	6	Get fasta
contig | short	204	Get fasta
mitag | raw | mapped	6945	Get fasta
mitag | raw | unmapped	2674	Get fasta

TARA-004-DCM_215

Station: TARA_004

Sequence Length: 231

Sequence:
CATTGCAGCCCGCGTGTGGCCCCAGAGTTTCGGGGCATACTGACCTGCCGTGGCCCCTTC
CTTCCTCCGCATTAACTGCGGCGGTCCCCCTAATTCGCCCTACTACACCAGGGTGTAATA
GTGGCAACTAGAGGCAGGGATCTCGCTCGTTACCTGACTTAACAGGACATCTCACGGCAC
GAGCTGGCGACGGCCATGCACCACCTCTCAGCTTGTCTGGTAAAGTCTTCA

Sampling Depth: DCM

Taxonomic Affiliation: archaea, thaumarchaeota, marine group i, unknown order, unknown family, candidatus nitrosopelagicus, marine metagenome

Sampling Fraction: 0.22-3

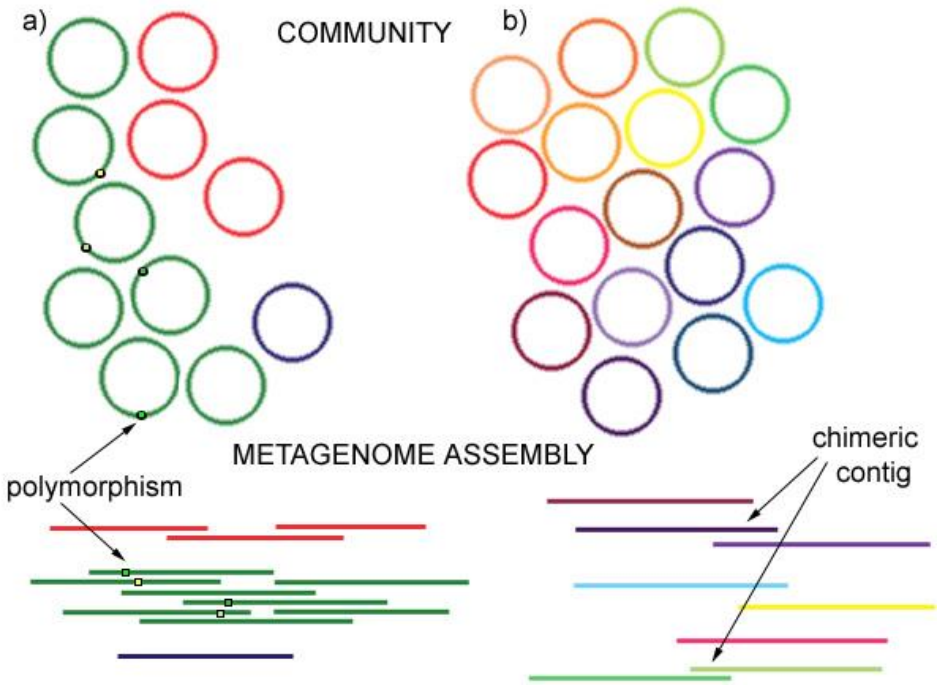
Show miTAGs | Get miTAGs fasta

Assembly

- Assembly is the first step for a correct representation of the sequence assemblage
- Through assembly we can perform two separate tasks at once:
  - Reducing the dataset complexity: fewer reads from the huge amounts usually produced by recent technologies
  - Increasing efficiency of taxonomic assignment and functional identification: by assembling we produce longer sequences (contigs and scaffolds) which can provide more information about the organisms we are characterizing

Assembly

- Metagenomic assembly is more complicated than genome assembly
- Requires specific pipelines for the correction of putative mis-assemblies



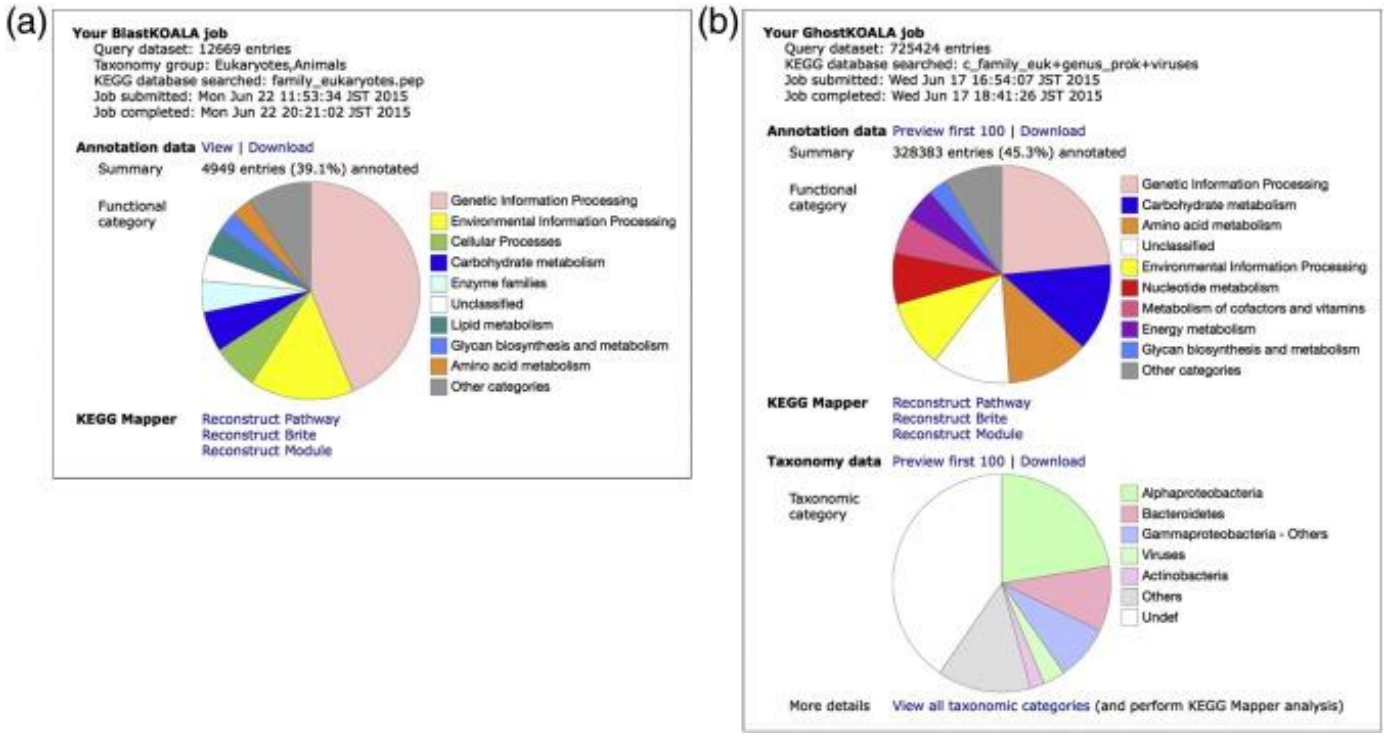
Assembly

- Steps that can be automated:
  - Read quality check
  - Read merging
  - Read assembly
- Steps that require supervision:
  - Assembly quality check
  - Assembly strategy comparison

- Outputs:
  - Statistics of quality check
  - Assembly statistics
  - Set of contigs/scaffolds ready for downstream analyses

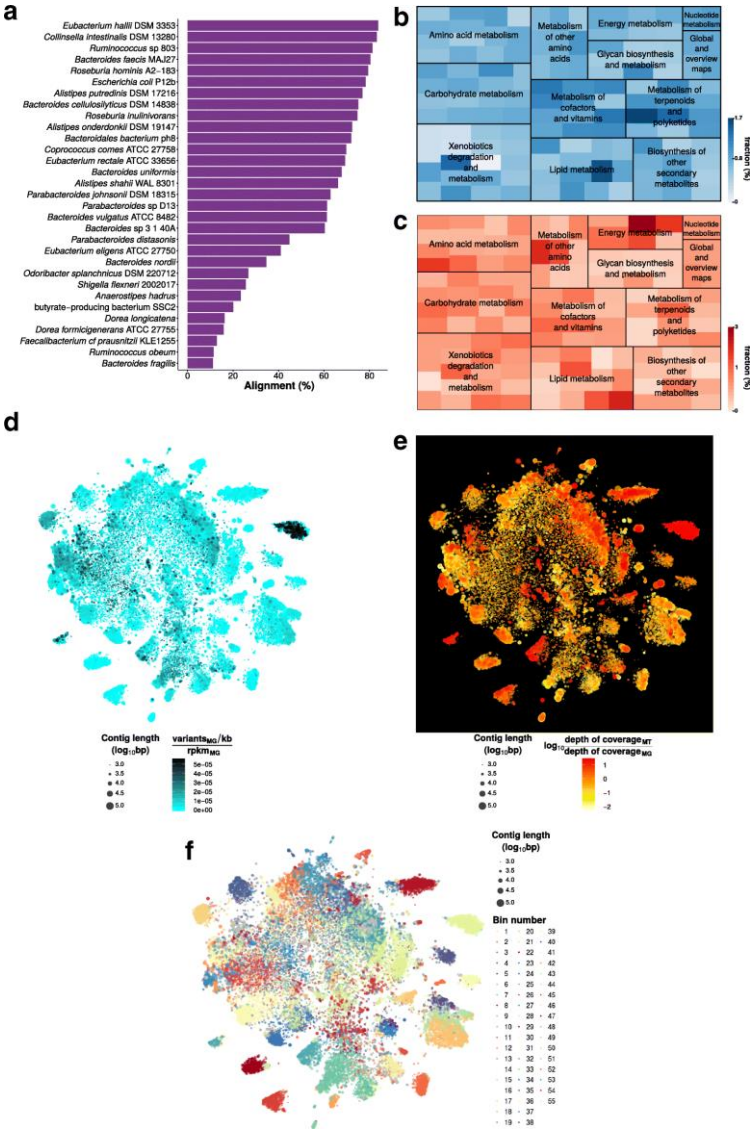
Annotation

- After assembly, contigs and scaffolds can be annotated, e.g. information can be added to each contig for taxonomic and functional analyses



Annotation

- Genes can be identified (easier for prokaryotes and viruses, more complicated for eukaryotes) on contigs and analyzed to infer their taxonomic identity and function
- Reads can be back-mapped to contigs (or genes) to infer their relative abundances in the sample and to create abundance tables for subsequent ecologic analyses



Annotation

- Needed input:
  - Contig/scaffold file
- Output:
  - Contig/scaffold – gene – annotation table

Contig name	Gene name	Annotation	Annotation source
Contig_1	Gene_1 (start-end)	SoxA	UniProt
Contig_1	Gene_2 (start-end)	SoxC	UniProt
Contig_2	Gene_3 (start-end)	16S rDNA	SILVA

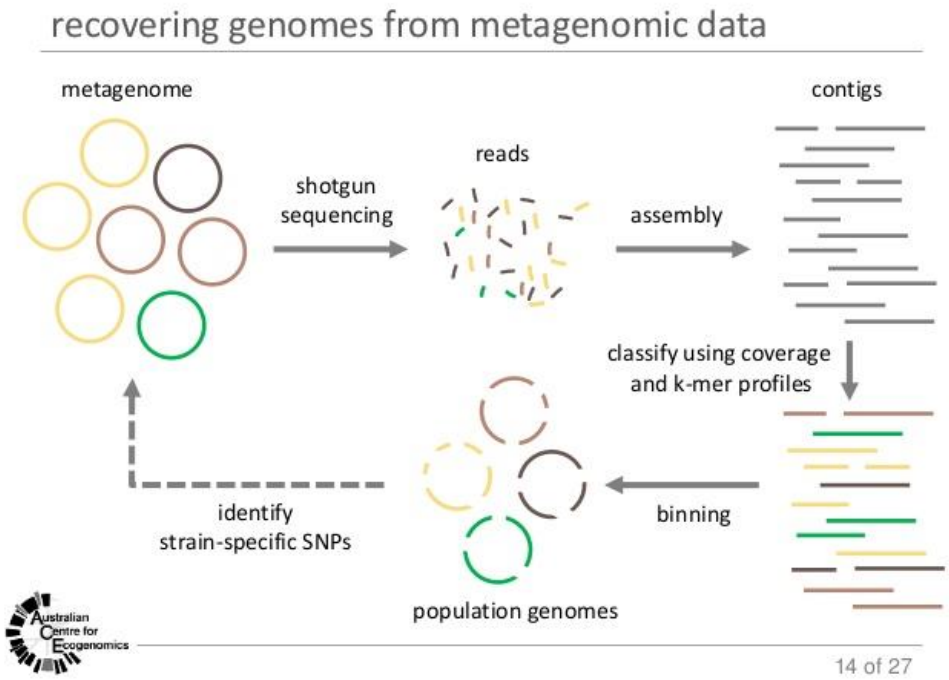
Annotation

- Steps that can be automated:
  - Gene finding
  - Gene annotation
  - Gene -> protein translation

- Steps that require supervision:
  - Annotation quality check
  - Contig table construction

Binning

- After metagenome assembly, single (prokaryotic and viral) genomes can be extracted and isolated from the metagenome
- This can be done exploiting different features of all contigs (e.g. taxonomy, presence/absence of marker genes, differential abundance of reads, tetranucleotide frequencies)



Binning

- Single putative genomes can be further analyzed for:
  - Taxonomic identification
  - Functional assignments
  - Identify mis-assemblies
  - Check for potential contamination

Binning

- Needed input:
  - Contig/scaffold file (for binning)
  - Original sequence data (for mapping)

- Output:
  - Genome bin sequence file
  - Genome bin quality file

Binning

- Steps that can be automated:
  - Genome binning
  - Bin quality check
  - Bin reassembly
  - Bin annotation

- Steps that require supervision:
  - Assessment of quality check results
  - Finding the most suitable binning strategy
  - Identifying reliable bins


Thanks for your patience