Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico “Dead Zone”

Thrash et. al. 2017

Duplicating the results

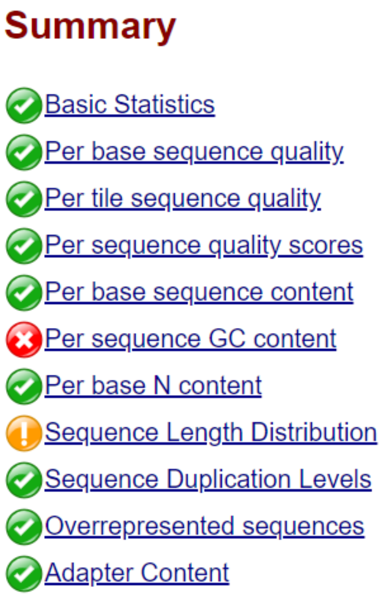
Patrick Hennig - 2021-05-28

Sequencing Data

- Biological samples from 2 sites
- Illumina paired end reads
  - One run per site for DNA
  - One run per site for mRNA
- Equal to 25% of the paper’s reads

Read Quality check

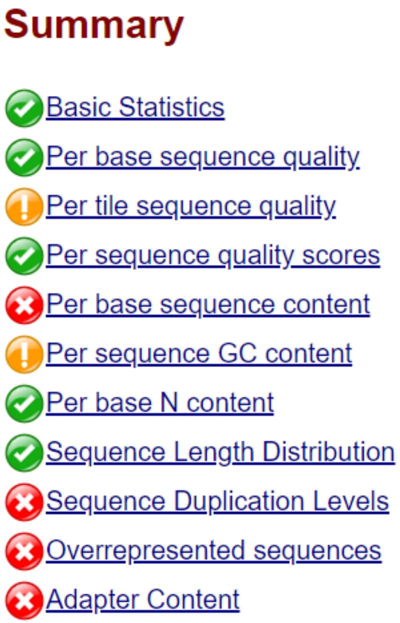
FastQC



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

DNA



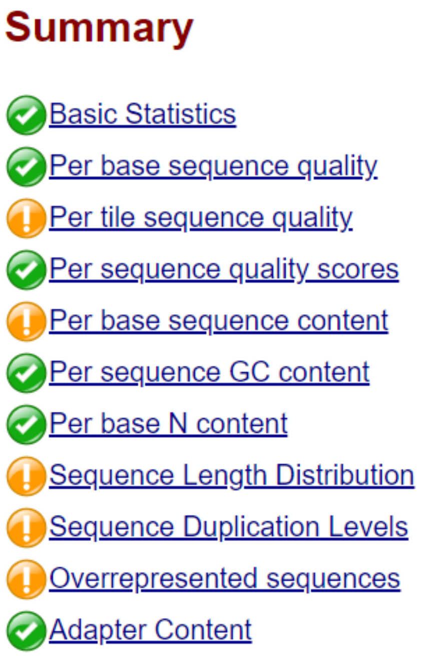
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Adapter Content

mRNA

Trimming - mRNA

- Leading/Trailing low quality sequences
- Adapter sequences
- Sliding window
- Trimmomatic



Summary

✓ Basic Statistics
✓ Per base sequence quality
! Per tile sequence quality
✓ Per sequence quality scores
! Per base sequence content
✓ Per sequence GC content
✓ Per base N content
! Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✓ Adapter Content

FastQC

Metagenome assembly and QC

- Megahit assembler
- 231Mbp total genome length
  - Paper: 217Mbp
- N50 980
- 346472 contigs
- Quast QC

DNA binning

- Metabat
- Tetranucleotide freqencies + coverage
- 26 bins estimated
- Not all contigs used

DNA binning QC

- CheckM
- Marker genes → estimate completeness, contamination
- Rough taxonomic ID assignment

DNA binning QC

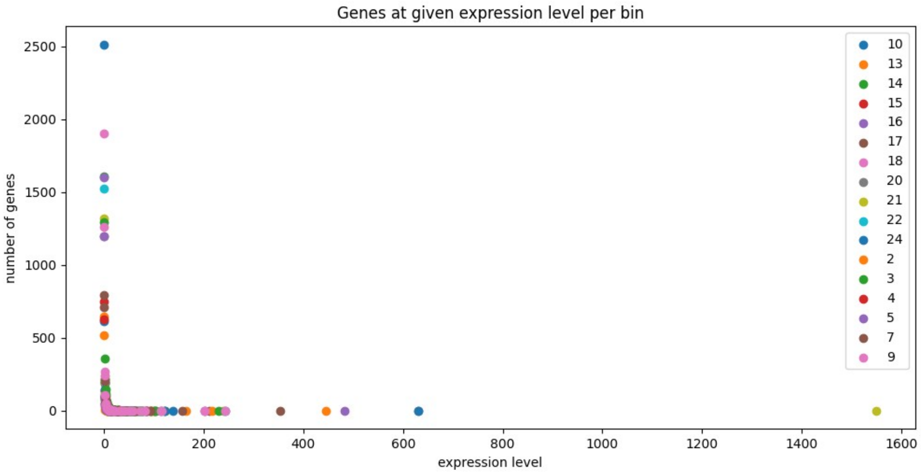
- Continue with bins:
  - >10% completeness and
  - <10% contamination
- 17/26 bins left
- Rough taxonomic assignment:
  - Mostly Kingdom only

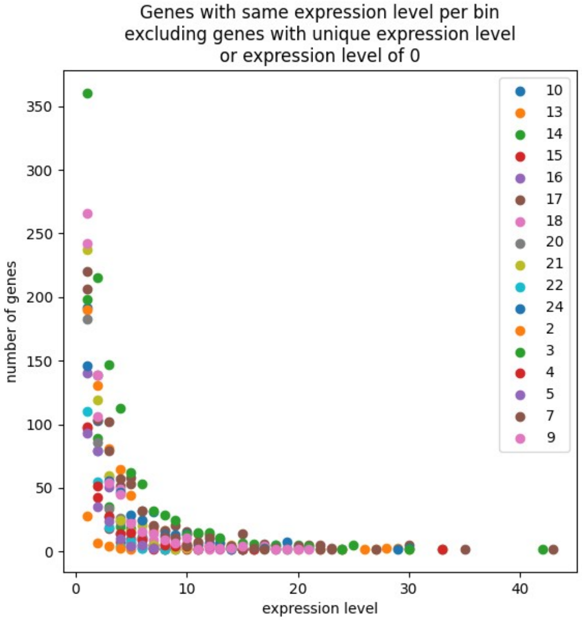
Functional annotation

- Annotation using Prokka
- Different algorithms for e.g.:
  - Genes
  - CDS
- ca. 28k genes & CDS across metagenome
- Paper: 140k genes

Analysis of bin activity

- Mapping using bwa
- Mapped reads counted using htseq





Extra analysis

- Taxonomic refinement
  - Classes determined for 10 bins
- Comparison of expression levels between bins
- (Ortholog gene clustering)