
DTU

DTU Health Technology
Bioinformatics

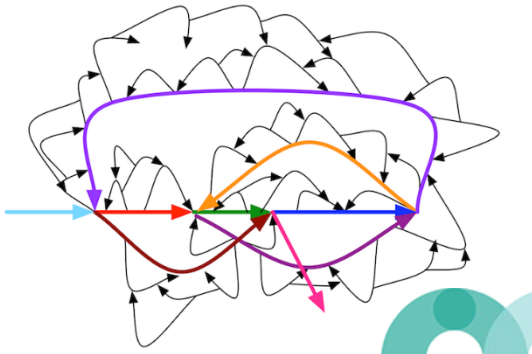
Week 8 Recap

Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk

Last Week…last year…

Two weeks ago…

- Metagenomics *de novo* assembly
- Nonpareil exercises



clinical microbiomics

Nonpareil exercise questions

Q1: What is nonpareil and what is it used for?

- Nonpareil is used to estimate the coverage of metagenome datasets.
- It uses the redundancy of the reads in metagenomic datasets to estimate the average coverage and predict the amount of sequences that will be required to achieve “nearly complete coverage”.

Nonpareil exercise questions

Q2: Why do we use the R1 reads to ensure higher quality?

- We generally see that the R2 reads will be of lower quality, due to imprecisions on the sequencing machines for R2.

Nonpareil exercise questions

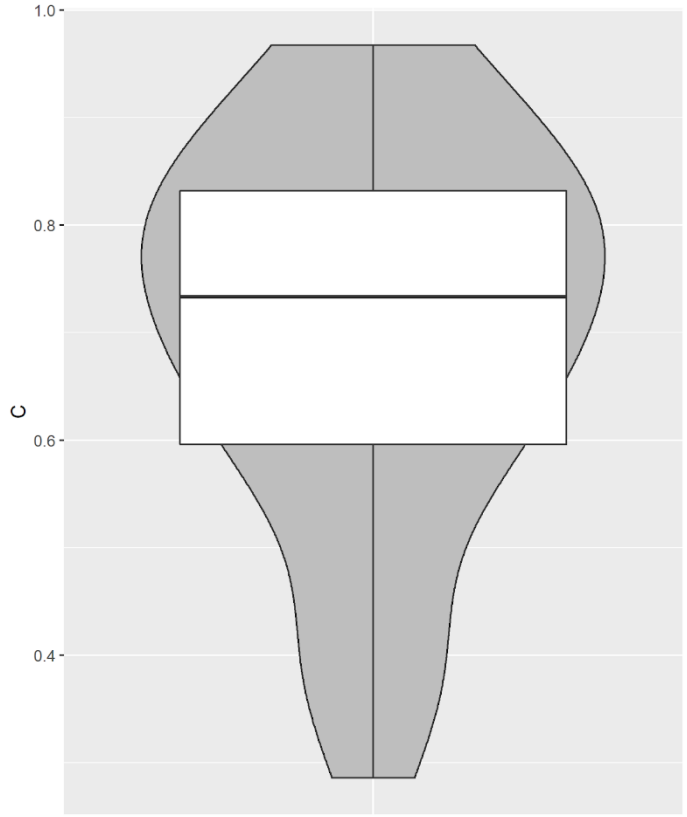
Q3: Briefly explain what the Rscript is used for

- It collects the nonpareil results for all samples and creates a summary.
- Returns a matrix with the following values for the dataset:
- kappa: "Redundancy" value of the entire dataset.
- C: Average coverage of the entire dataset.
- LRstar: Estimated sequencing effort required to reach the objective average coverage (star, 95)
- LR: Actual sequencing effort of the dataset.
- modelR: Pearson's R coefficient betweeen the rarefied data and the projected model.
- diversity: Nonpareil sequence-diversity index (Nd). This value's units are the natural logarithm of the units of sequencing effort (log-bp), and indicates the inflection point of the fitted model for the Nonpareil curve. If the fit doesn't converge, or the model is not estimated, the value is zero (0).

Nonpareil exercise questions

Q4: What can you tell about the coverage?

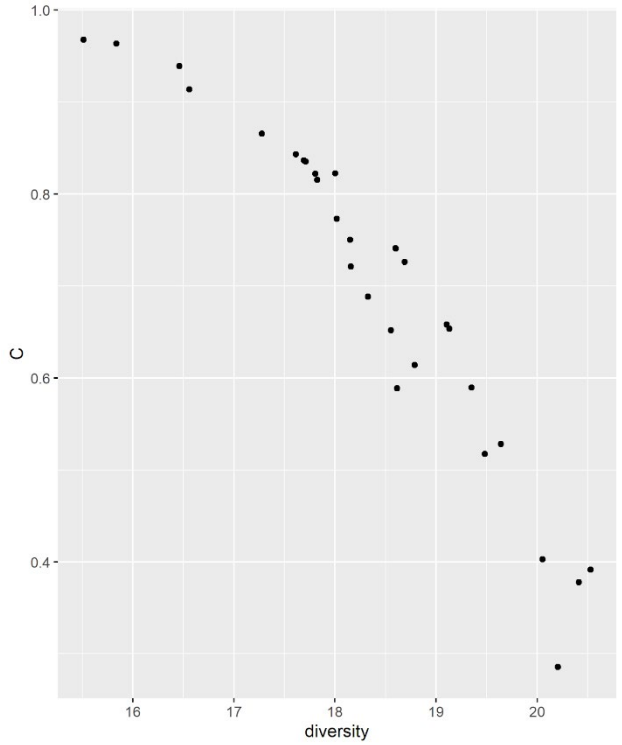
- So in general we have pretty good (more than 50% of the metagenome) coverage of most samples



Nonpareil exercise questions

Q5: Do the coverage correlate with the diversity?

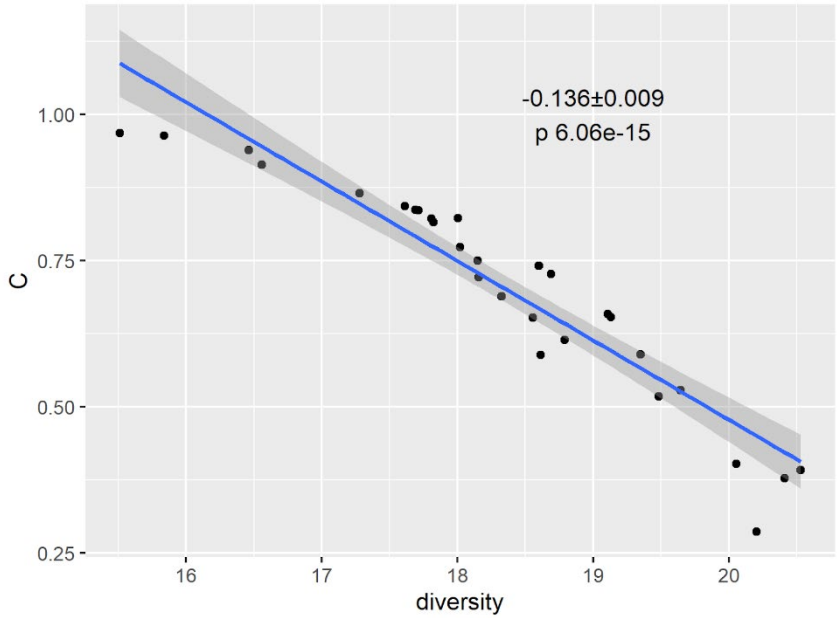
- Not surprisingly there is an obvious negative correlation between diversity of metagenome and coverage since all samples were sequenced with similar number of reads.



Nonpareil exercise questions

Q6: Do the model support your answer from Q5?

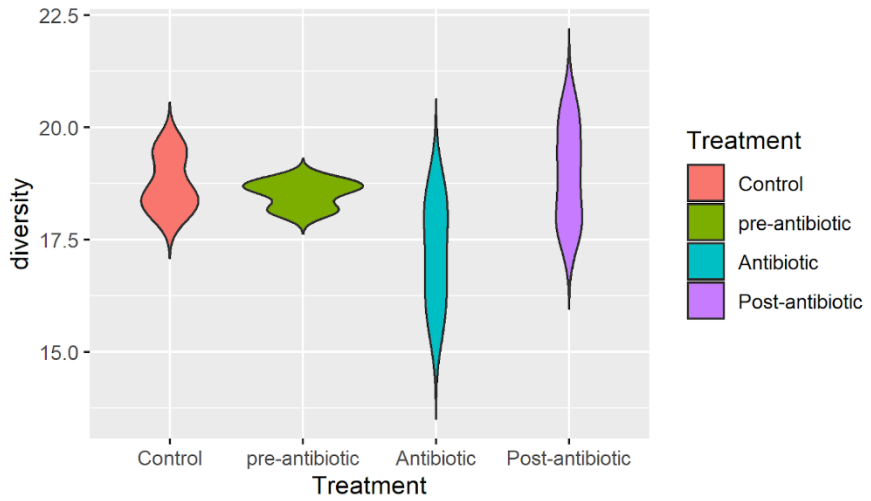
- Yes, we see that the coverage is negatively correlated with the diversity.



Nonpareil exercise questions

Q7: What do we see on the violin plot? How do the treatment affect the diversity?

- So metagenome diversity seems lowest during antibiotic-treatment and significantly lower compared to post-antibiotic treatment samples. I am fairly certain that with more samples, we would have significantly lower diversity in the fish being treated with antibiotics compared with all three other conditions.

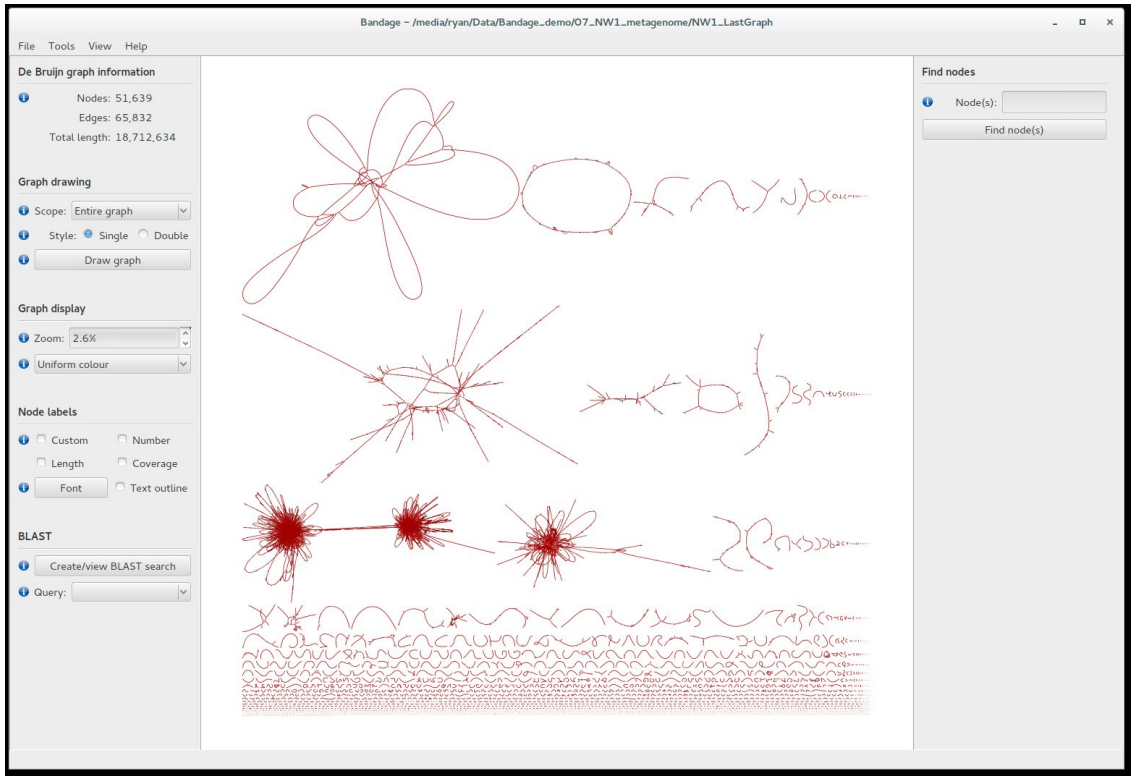


De novo assembly

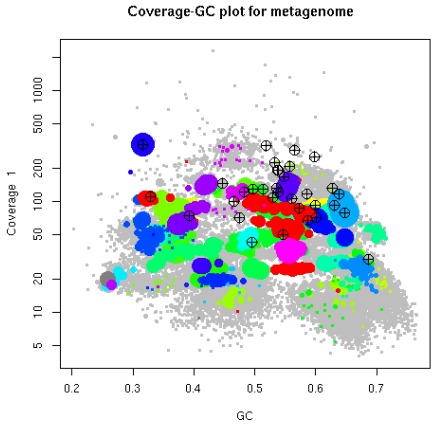
- Metagenomics *de novo* assembly is even harder
- Currently we use de Bruijn graphs to assemble

De Bruijn graphs

- Really struggles with repetitive regions
  - Horisontally transferred regions
  - Closely related strains
  - Etc.



Today



- Kaiju exercise
- Metagenomic binning
- Start the project work!