RNA Seq analysis

De novo assembly

Practical session
The Data

Transcriptome sequencing and comparative expression profiling analysis of *Saccharina japonica* (Kombu).

It is one of the two most consumed species of kelp in China and Japan.

**Overall Design**: mRNA expression of *Saccharina japonica* with 2 different treatment (sample exposed to Dark condition, and sample exposed to blue light respectively) was determined by method of RNA-Seq.

**Citation**: Deng Y, Yao J, Wang X, Guo H, Duan D (2012) Transcriptome Sequencing and Comparative Analysis of *Saccharina japonica* (Laminariales, Phaeophyceae) under Blue Light Induction. PLoS ONE 7(6): e39704. doi:10.1371/journal.pone.0039704
!! WARNING !!

- « TRUE LIFE » data but « re-ingeenered » data
- Reverse analyse
- EdgeR + 1 biological replicate ...

- Selection of 800 genes : 400 NDE– 400 DE vs 70500 unigenes
- 1 millions reads selected vs 24 millions reads sequenced
Data Cleaning
Because...

- Unknown nucleotides
- Bad quality nucleotides
- Adaptors and primers sub-sequences
- Poly A/T tails
- Low complexity sequences
- rRNA sequences
- Contaminant sequences
- Short length sequences

But also:
- Removing singletons
- In-silico normalization
- Sequencing errors correction
- ...

ITMO 2013
NGS data Quality Checking (QC)

• These apply to all NGS data (not just RNAseq).

• Some of these problems can be worked around but others indicate that the lane is bad & must be re-run (or a new library is needed).

• Bias should be corrected in reverse order of their generation
  1. Sequencing biases (bad quality, unknowns)
  2. Library preparation
    a. Adaptors and primers sequences
    b. Poly A/T tails
  3. Biological sample (low complexity, rRNA, contaminants)

• Our favorite NGS QC tools is FastQC.
  http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Data Cleaning

Input (fastq) → Prinseq (Ns, qual) → Cutadapt → Prinseq (polyA, complexity) → riboPicker

FastQC → FastQC → FastQC → FastQC → FastQC → FastQC

Pairs Retrieval

for each condition

Read 1

Read 2

Single 1

Single 2
Normalization step
Normalized assembly step
« Normal » paired data set
« high singleton number » data set
Differential Expression Analysis – N methods

A
Read 1
A
Read 2
Filtered
De-novo
Transcriptome

B
Read 1
B
Read 2

RSEM Align & Estimate

Merge Table

Run DE analysis
DESeq
Filter (on pad)

Run DE analysis
edgeR
Filter (on FDR)

proportional
venn

Compare two
Datasets

Normalized
data matrix +
statistics values

Cut

Filtered
Normalized
data matrix
Pipeline modification

Very few singletons :
• Assembly : (pairedR1 + pairedR2)_norm
• Remapping - filtering : pairedR1 + pairedR2
• Remapping - counting : pairedR1 + pairedR2

Few singletons :
• Assembly : (pairedR1 + pairedR2)_norm
• Remapping - filtering : pairedR1 + singletonR1 + singletonR2
• Remapping - counting : pairedR1 + singletonR1 + singletonR2

Lot of singletons :
• Assembly : (pairedR1 + pairedR2)_norm + (singletonR1 + singletonR2)_norm
• Remapping - filtering : pairedR1 + singletonR1 + singletonR2
• Remapping - counting : pairedR1 + singletonR1 + singletonR2