Biotecnología de Plantas
Posgrado

Cinvestav

Transcriptómica

Clase 2 - Formato y calidad de RNA-Seq

Bioinformática y Bioestadística 2021

Selene L. Fernández-Valverde
regRNAlab.github.io
@SelFdz

Objetivos de aprendizaje

En esta clase aprenderemos:

- Que tipo de datos obtenemos de un experimento de secuenciación masiva
- Que es el formato FastQ
- Como sabemos si los datos obtenidos tienen la calidad suficiente para ser analizados

Fuentes de error

- Existen dos fuentes principales de error:
  - **Error humano:** mezcla de muestras (en el laboratorio o cuando se recibieron los archivos), errores en el protocolo
  - **Error técnico:** Errores inherentes a la plataforma (e.g. secuencias de mononucleotidos en pyrosecuenciacion) - Todas las plataformas tienen cierto de nivel de error que se debe tomar en cuenta cuando se esta diseñando el experimento.

Errores en preparación de la muestra

- Error del usuario (e.g. etiquetar equivocadamente una muestra)
- Degradación de ADN/ARN por métodos de preservación
- Contaminación con secuencias externas
- Baja cantidad de ADN de inicio

Errores en preparación de la librería

- Error del usuario (e.g. contaminar una muestra con otra, contaminar con reacciones previas, errores en el protocolo)
- Errores de amplificación por PCR
- Sesgo por (cebadores) primers (sesgo de unión, sesgo por metilación, dímeros de cebadores [primer dimers])
- Sesgo por captura (Poly-A, Ribozero)
- Errores de máquina (configuración errónea, interrupción de la reacción)
- Quimeras
- Errores de índice, adaptador (contaminación de adaptadores, falta de diversidad de índices, códigos (barcodes) incompatibles, sobrecargo)

Errores de secuenciación e imagen

- Error del usuario (e.g. sobrecarga de la celda)
- Desfase (e.g. extensión incompleta, adición de múltiples nucleótidos)
- Fluoróforos muertos, nucleótidos dañados y señales superpuestas
- Contexto de la secuencia (e.g. alto contenido de GC, secuencias homologas y de baja complejidad, homopolímeros).
- Errores de máquina (e.g. laser, disco duro, programas)
- Sesgos de cadena

El reto - diferenciar señales biológicas de ruido/errores

- Controles negativos y positivos - ¿Qué espero?
- Réplicas técnicas y biológicas - ayudan a determinar la tasa de ruido
- Conocer los tipos de errores comunes en determinada plataforma

El formato FastQ (.fastq)

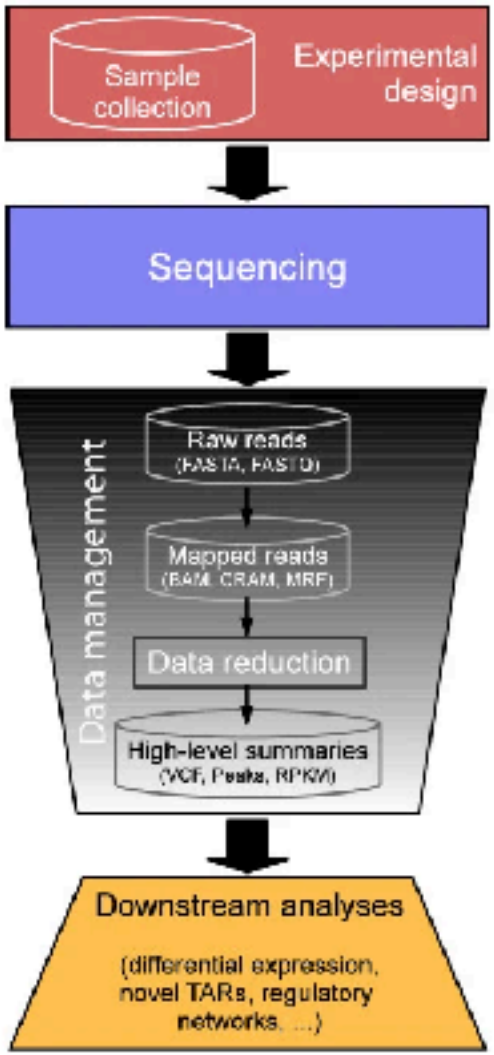
```
@SRR2467151.1 SALLY:472:C6NCDACXX:8:1101:1366:1867 length=101
NTCTCTTGCTCCAAGTCCTTGTTCTTCTCTTTCAACTCCTCGTACTTCTCCTGCCATGATTCGGCTTCCA
+SRR2467151.1 SALLY:472:C6NCDACXX:8:1101:1366:1867 length=101
#4=DDFFFFHFHDGHJIJHIIHICFH<C<FHIAEFHHGHHGIIJIIJGGFGHIIJCFH?FDFEGHGGEEH
@SRR2467151.2 SALLY:472:C6NCDACXX:8:1101:1372:2000 length=101
CAATAACATACCAATAGTCGCTGTCAGTAGCCATGGCAATAATACTAATTTGTGAACGGTCACCTAGCTC
+SRR2467151.2 SALLY:472:C6NCDACXX:8:1101:1372:2000 length=101
CCCFFFFFHHHHHJJJJIJJJJJIIJJEHIJJJJJJJIJJJJJJJJJJJJJJIIIJJJJIJJJJIJJJJI
@SRR2467151.3 SALLY:472:C6NCDACXX:8:1101:1599:1867 length=101
NAGTAATAGACAAATAGTCTCTAAGGATTGTTTACCACGGTCGACATAATCACCCAAAAATAGATAATTA
```

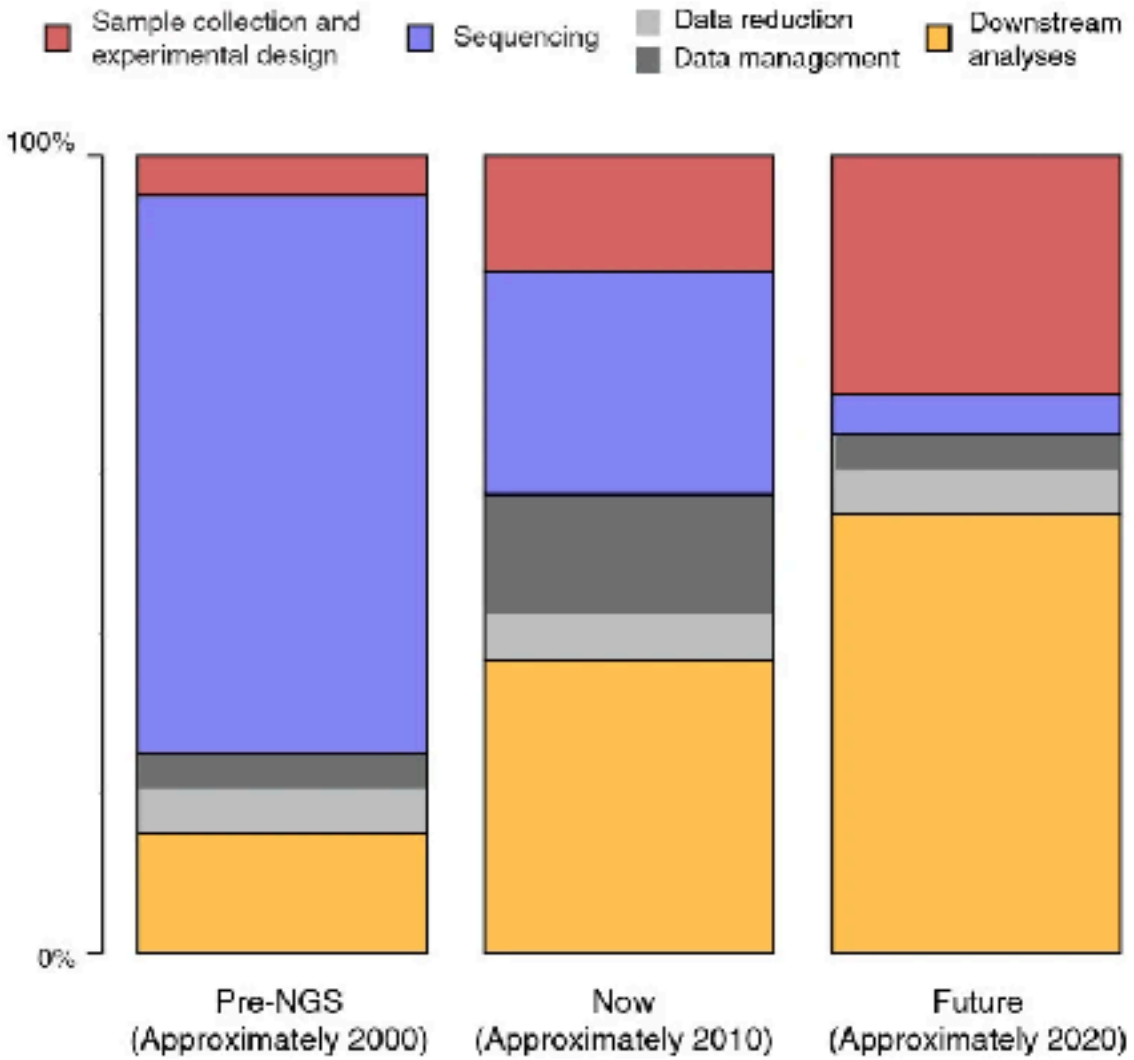
Software de análisis de y control de calidad

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- RseQC (http://rseqc.sourceforge.net/)
- RNA-SeQC (https://www.broadinstitute.org/cancer/cga/rna-seqc)
- Picard (http://broadinstitute.github.io/picard/)

Otras medidas de control de calidad

- ¿Cuál es el gen mas expresado en mi muestra?
- ¿Tienen mis observaciones sentido con respecto a lo que se de mi sistema?
- ¿Algunas de las secuencias sobre representadas sugieren errores humanos o errores en el protocolo?
- ¿Qué tan similares/diferentes son mis réplicas?





Sboner. A. et al. *Genome Biology*, *2011*

Práctica - análisis de calidad usando FastQC

https://liz-fernandez.github.io/PBP_transcriptomics_2020/