Datos de RNA-Seq

Guillermo Ayala Gallego

2024-04-02

Table of contents

Introducción **1**
Flujo de trabajo . . . 1
Objetivos . . . 2
Repositorios . . . 2
Formato FASTA . . . 2
Formato FASTQ . . . 3
Phred . . . 3
Phred . . . 3
Ejemplos . . . 4
`SummarizedExperiment` . . . 4

Introducción

Flujo de trabajo

1. Diseño experimental.
2. Protocolos de extracción del RNA.
3. Preparación de las librerías. Se convierte el RNA en cDNA y se añaden los adaptadores para la secuenciación.
4. Se secuencian las lecturas cDNA utilizando una plataforma de secuenciación.
5. **Alineamiento de las lecturas secuenciadas a un genoma de referencia.**
6. **Resumen del número de lecturas alineadas a una región.**
7. **Normalización de las muestras para eliminar diferencias técnicas en la preparación.**
8. **Estudio estadístico de la expresión diferencial incluyendo en lo posible un modelo.**
9. Interpretación de los resultados desde el punto de vista biológico.

Objetivos

1. ¿Dónde podemos obtener datos de secuenciación? De otro modo: ¿qué repositorios podemos utilizar?
2. ¿Cómo podemos realizar búsquedas en los metadatos con objeto de obtener experimentos que tenga que ver con lo que me interesa?
3. ¿Cómo bajarlos?
4. ¿Cómo contar las lecturas alineadas sobre regiones genómicas (genes, exones o ...)?

Repositorios

1. NCBI SRA
2. EBI ENA
3. DDBJ

Formato FASTA

- El formato **fasta** está basado en texto.
- Se utiliza representar secuencias bien de nucleótidos bien de aminoácidos.
- Tanto unos como otros son representados por una sola letra.
- También tiene símbolos para representar un hueco (gap) o parada en la traducción o bien que no se sabe el nucleótido o aminoácido.
- Tiene una línea que comienza con el símbolo > al que sigue una descripción de la secuencia.
- En la siguiente línea empieza la secuencia de bases o aminoácidos. Se recomienda que no tener más de 80 columnas y se pueden tener todas las filas que se precisen.
- Un ejemplo

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken     MADQLTEEQIAEFKEAFSLFDKD
PEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADI*
```

Formato FASTQ

- El formato **fastq** es el más popular para datos de secuencias.
- Consiste de cuatro líneas por lectura.
- La primera que comienza con el carácter `@` y contiene el **nombre de la secuencia** con alguna descripción opcional de la misma.
- La segunda línea contiene la secuencia con las letras que correspondan dependiendo del tipo (nucleótidos, aminoácidos).
- La tercera línea que comienza con + contiene información opcional sobre la secuencia.
- La cuarta línea cuantifica la confianza o calidad en la determinación de cada base recogida en la segunda línea (**Phred**).
- Un ejemplo

```
@SRR1293399.1 ILLUMINA-545855_0026_FC629BG:6:1:1022:5049 length=50
ACAGGGACGCCATCGAATCCGGATCNTNNNNNNNNNNNNANNNNNNNNNN
+SRR1293399.1 ILLUMINA-545855_0026_FC629BG:6:1:1022:5049 length=50
dee\edYcdcbbYS]bb_]]Ua^BBBBBBBBBBBBBBBBBBBBBBBBBB
```

Phred

- ¿Cómo se cuantifica la confianza o precisión?

1. Phred asigna los picos de fluorescencia a una de las cuatro bases: **base call**.
2. P: probabilidad para una base dada de ser mal asignada o clasificada
3. Se muestra:
$$Q = -10 \log_{10} P.$$
4. Una probabilidad P muy pequeña de clasificación incorrecta se traduce en un valor grande de Q.
5. Se muestra el caracter ASCII que ocupa la posición $33 + Q$.

Phred

- Simulamos las probabilidades de clasificación incorrecta y generamos al azar las bases.

```
x = sample(Biostrings::DNA_ALPHABET[1:4],10,replace=TRUE)
y = runif(10,min=0,max=.001)
names(y) = x
y
```

```
           A            A            A            G            C            T
3.275159e-04 6.759368e-04 3.409596e-04 4.405459e-04 3.872925e-04 3.753785e-05
           A            G            T            T
7.955901e-04 2.272534e-04 5.364215e-04 2.730879e-04
```

- Calculamos los `Q` valores.

```
(Q = round(-10 * log10(y)))
```

```
 A  A  A  G  C  T  A  G  T  T
35 32 35 34 34 44 31 36 33 36
```

- Codificación Sanger

```
intToUtf8(Q+33)
```

```
[1] "DADCCM@EBE"
```

Ejemplos

- `tamidata::PRJNA266927`
- `tamidata::PRJNA297664`
- `tamidata::PRJNA297798`

`SummarizedExperiment`

- Cargamos el paquete.

```
library(SummarizedExperiment)
```

- Leemos unos datos de `tamidata`

```
data(PRJNA297664,package = "tamidata")
```

- ¿Qué clase es?

```
class(PRJNA297664)
```

```
[1] "RangedSummarizedExperiment"
attr(,"package")
[1] "SummarizedExperiment"
```

- ¿Cuántos genes y muestras tenemos?

```
dim(PRJNA297664)
```

```
[1] 7126    6
```

- La matriz de conteos es

```
head(assay(PRJNA297664))
```

- ¿Qué devuelve `assay`?

```
class(assay(PRJNA297664))
```

```
[1] "matrix" "array"
```

```
head(which(apply(assay(PRJNA297664),1,sum) > 10))
```

```
  ICR1   LSR1   NME1 RDN5-1 RDN5-6 RNA170
     4      5      6     42     47     50
```

- Variables fenotípicas

```
colData(PRJNA297664)
```

```
DataFrame with 6 rows and 4 columns
   SampleName         Run      treatment replication
  <character> <character>       <factor>   <numeric>
1  GSM1900735  SRR2549634 Wild                     1
2  GSM1900737  SRR2549636 Wild                     3
3  GSM1900739  SRR2549638 SEC66 deletion           2
4  GSM1900736  SRR2549635 Wild                     2
5  GSM1900738  SRR2549637 SEC66 deletion           1
6  GSM1900740  SRR2549639 SEC66 deletion           3
```

- ¿Y es?

```
class(colData(PRJNA297664))
```

```
[1] "DFrame"
attr(,"package")
[1] "S4Vectors"
```

- Nombres de las variables fenotípicas.

```
names(colData(PRJNA297664))
```

```
[1] "SampleName"  "Run"         "treatment"   "replication"
```

- Y accedemos a los valores de la variable `treatment` con

```
colData(PRJNA297664)[,"treatment"]
```

```
[1] Wild           Wild           SEC66 deletion Wild           SEC66 deletion
[6] SEC66 deletion
Levels: Wild SEC66 deletion
```

- Datos relativos a las filas o características o genes.

```
head(rownames(PRJNA297664))
```

```
[1] "15S_rRNA" "21S_rRNA" "HRA1"     "ICR1"     "LSR1"     "NME1"
```

- ¿Hay más? Pues sí.

```
head(rowData(PRJNA297664))
```

```
DataFrame with 6 rows and 4 columns
                ORF         SGD    ENTREZID     ENSEMBL
        <character> <character> <character> <character>
15S_rRNA    15S_rRNA          NA          NA          NA
21S_rRNA    21S_rRNA          NA          NA          NA
HRA1            HRA1  S000119380     9164866          NA
ICR1            ICR1  S000132612     9164906          NA
LSR1            LSR1  S000006478     9164871          NA
NME1            NME1  S000007436     9164967          NA
```

- ¿Algo más? De hecho, mucho más.

```
rowRanges(PRJNA297664)[3]
```

```
GRangesList object of length 1:
$HRA1
GRanges object with 1 range and 2 metadata columns:
      seqnames      ranges strand |   exon_id   exon_name
         <Rle>   <IRanges>  <Rle> | <integer> <character>
  [1]        I 99305-99868      + |        30      HRA1.1
  -------
  seqinfo: 17 sequences (1 circular) from an unspecified genome; no seqlengths
```