

Máster en Bioinformática

Generación y mantenimiento de datos ómicos

Curso académico 2024-25



Universidad
Internacional
de Valencia

Dr. Jordi Tronchoni León
Jordi.tronchoni@profesor.universidadviu.es

22/05/2024

De:
 Planeta Formación y Universidades

Tema 7

Pre-procesado y calidad de secuencias

22/05/2024

Tema 1. Introducción a la bioinformática

- 1.1 Historia de la bioinformática
- 1.2 Bioética aplicada al análisis de datos

Tema 2. Principales flujos de trabajo en bioinformática

- 2.1 Genómica
- 2.2 Metagenómica y metataxonómica
- 2.3 Transcriptómica
- 2.4 Proteómica

Tema 3. Gestión de entornos y paquetes

- 3.1 Conda

Tema 4. Bases de datos y herramientas bioinformáticas

- 4.1 Principales bases de datos
- 4.2 Otros recursos online

Tema 5. Alineamiento de secuencias

- 5.1 Introducción al alineamiento de secuencias
- 5.2 Alineamientos Pairwise
- 5.3 Alineamientos Múltiples

Tema 6. Métodos de secuenciación

- 6.1 Primera generación de secuenciadores
- 6.2 Segunda generación de secuenciadores
- 6.3 Tercera generación de secuenciadores
- 6.4 Comparación de plataformas de secuenciación

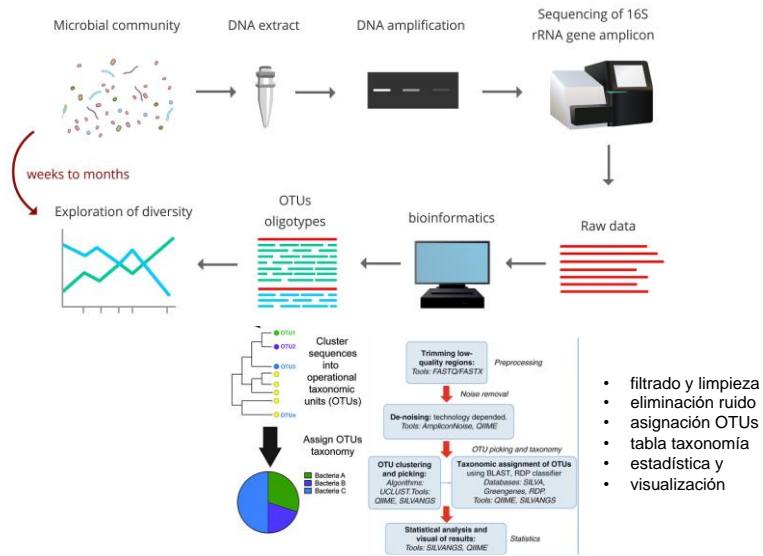
Tema 7. Pre-procesado y calidad de secuencias

- 7.1 Calidad de secuencias
- 7.2 Pre-procesado de secuencias

Pre-procesado y calidad de secuencias

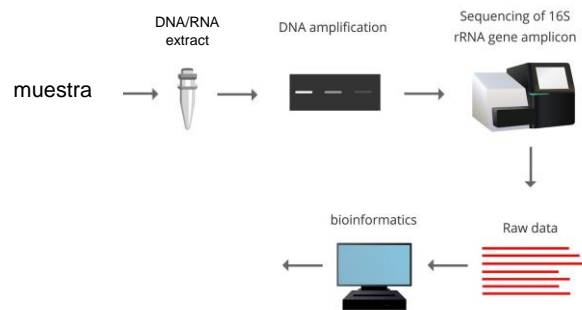
22/05/2024

Flujo de trabajo en metataxonómica

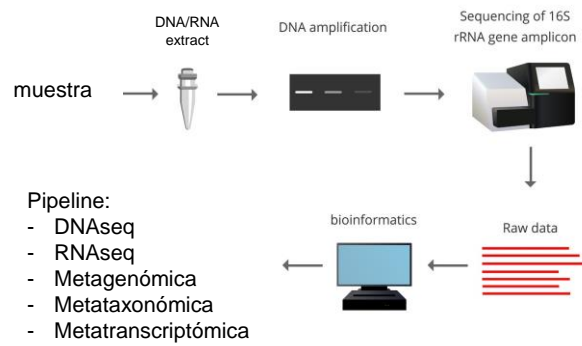


22/05/2024

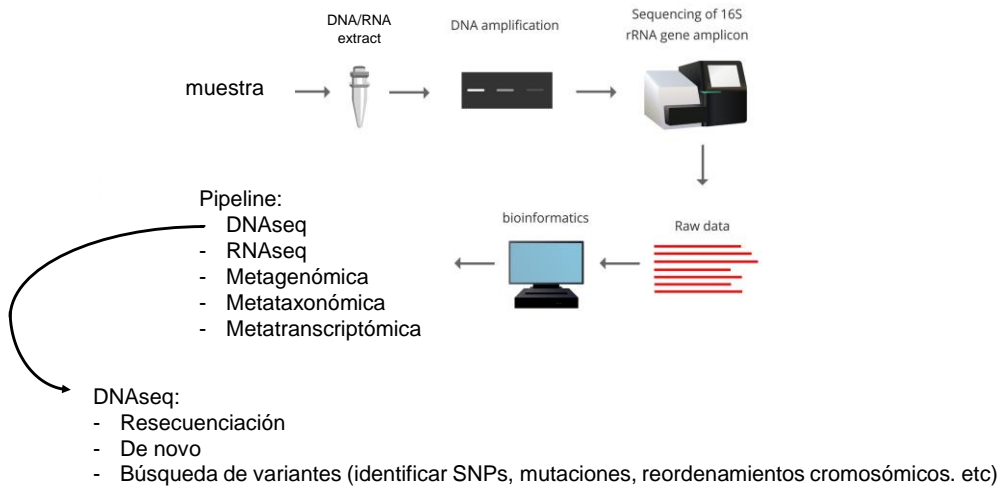
En general



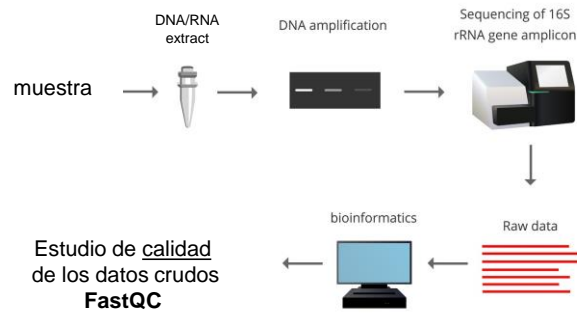
En general



En general



En general



- Limpieza de adaptadores
- Recortado por calidad
- Toma de decisiones

Calidad de secuencias: Sanger

22/05/2024

Sanger



ABI 3730xl

- Tiempo: 2-3h
- Longitud reads: 800 - 1000pb
- Salida media: 100kpb
- Fiabilidad: 99,999%
- Características:
 - Barato
 - Secuenciación de genoma único
- Utilidades:
 - Verificación de la edición del genoma mediada por CRISPR.
 - Control de calidad de la fabricación de ARNm (terapéutico, vacunas...)
 - Confirmación de secuenciación de próxima generación con secuenciación Sanger.
 - Confirmación de SNPs en SARS-CoV-2.
 - En general uso amplio para la confirmación de SNPs y variantes alélicas (cepas de microorganismos, secuenciación masiva etc...)

22/05/2024

Los secuenciadores de Sanger continúan a día de hoy siendo útiles. Algunos ejemplos son:

- Verificación de la edición del genoma mediada por CRISPR.

La secuenciación de Sanger puede utilizarse para determinar la precisión de las técnicas de edición del genoma mediadas por CRISPR en organismos complejos.

- Control de calidad de la fabricación de ARNm con secuenciación de Sanger.

Confirmación de la secuencia durante la fabricación de ARNm terapéutico y de vacunas.

- Confirmación de secuenciación de próxima generación con secuenciación Sanger.

La secuenciación de Sanger puede utilizarse como método para confirmar las variantes identificadas mediante secuenciación de próxima generación (NGS).

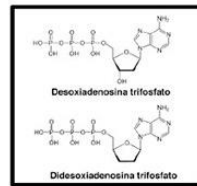
- Confirmación de SNPs en SARS-CoV-2.

En general uso amplio para la confirmación de SNPs y variantes alélicas (cepas de microorganismos, secuenciación masiva etc...)



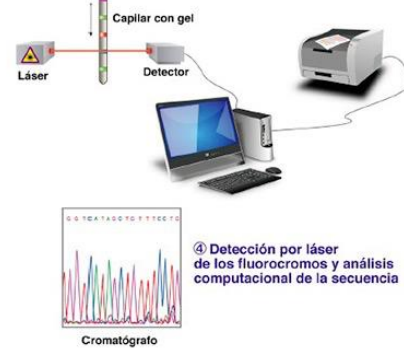
viu
Universidad
Internacional
de Valencia

- Cebador y plantilla de ADN
- ddNTP con fluorocromos
- ADN polimerasa
- dNTP (dATP, dCTP, dGTP y dTTP)



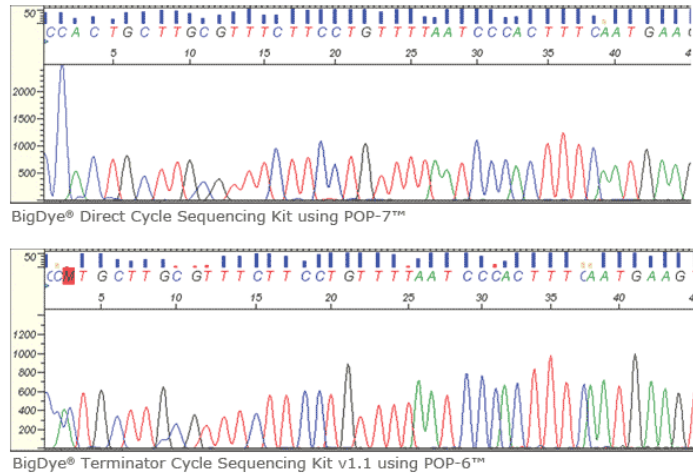
② Elongación del cebador y terminación de cadena

The diagram illustrates the second step of PCR, where the primer (Cebador) binds to the template (Plantilla) and the new strand is extended using dNTPs. The primer is shown as a short blue segment with a 3' end. The template is a long green segment with a 5' end. The extension is shown as a red segment growing from the 3' end of the primer. The dNTPs are labeled as ddNTP (red), dATP (blue), dGTP (green), and dCTP (pink). The extension process is shown in a series of steps, with the red segment growing longer in each step until it reaches the end of the template.



22/05/2024

Cromatograma



22/05/2024

La lectura lumínica de la secuencia de ADN se representa por picos en un cromatograma. Los colores hacen referencia a la base identificada debido al fluorocromo, la altura a la intensidad lumínica.

Cromatograma



Valor de calidad

QV: 0 - 50
> 40 muy buena calidad
= 20: 1% de probabilidad de error
< 10 indeterminación (N)

En función de la altura, a la intensidad lumínica, la distancia entre los picos.

22/05/2024

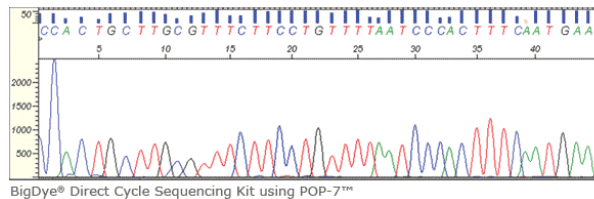
Todos los sistemas de secuenciación estiman la probabilidad de que cada uno de los nucleótidos secuenciados sean erróneos, a este parámetro se le suele llamar calidad. Esta estimación del error es específica de cada tecnología y la calcula el software del equipo. En Sanger, la altura del cromatograma, la intensidad lumínica, junto a la distancia entre los picos se traducen en un valor de **calidad** o como de creíble es el **nucleótido leído**. Siendo generalmente la fiabilidad de la lectura muy alta.

Podemos observar que cada base asignada va acompañada de un valor de calidad (QV: quality value), que está logarítmicamente relacionado con el error de **llamada de base**: $QV = -10 \times \log(\text{probabilidad de error})$. Por ejemplo, una base asignada con QV = 20 tiene una probabilidad de error de 0,01, es decir, un 1% de probabilidades de haber sido leída incorrectamente. El QV depende en última instancia de la forma y la relación señal/ruido del pico, y proporciona la métrica más objetiva para evaluar la confianza de la llamada de base. Por encima de 40 son de alta calidad, por debajo, merece la pena observar el cromatograma. Generalmente, un QV < 10 será traducido en una indeterminación N.

Cromatograma

Basecalling

A partir de los cromatogramas hay que obtener la secuencia de nucleótidos. Este proceso lo hacen automáticamente los programas que leen los cromatogramas, pero conviene que revisarlo manualmente porque en muchas ocasiones se producen fallos al asignar las bases. La secuencia propuesta por el *software* del secuenciador automático casi siempre tiene una gran parte al final que hay que eliminar.



22/05/2024

El proceso mediante el cual el programa asigna las bases en función del resultado del cromatograma se denomina **basecalling**. Generalmente el principio (unas pocas bases) y buena parte del final deberán eliminarse.

Las primeras 20 a 40 bases no suelen estar bien resueltas. Los productos de secuenciación muy cortos no migran de forma predecible durante la electroforesis capilar, y el software de análisis tiene dificultades para asignar bases dentro de esta región, lo que provoca que aparezcan Ns (indeterminaciones) en la secuencia. **Esto se debe tener en cuenta a la hora de diseñar cebadores y que se unan a una distancia de al menos 60 pb, preferiblemente 100 pb, de las bases de interés.**

La mayoría de los protocolos de secuenciación están optimizados para proporcionar la mejor resolución de picos entre 100 y 500 bases aproximadamente. En este intervalo, los picos deben ser nítidos y estar bien espaciados, y la **llamada de bases** es más fiable. **Hacia el final del cromatograma, los picos serán menos definidos y de menor intensidad.** La **llamada de bases** también será menos fiable. Debido a la naturaleza de la polimerización in vitro, los productos de secuenciación más grandes se generan de forma menos eficiente que sus homólogos más cortos. Por lo tanto, **los productos más grandes son menos numerosos y producen una señal más débil.** Además, con cualquier método de electroforesis, resulta cada vez más difícil resolver una diferencia de una sola base a medida que los fragmentos de ADN se hacen más

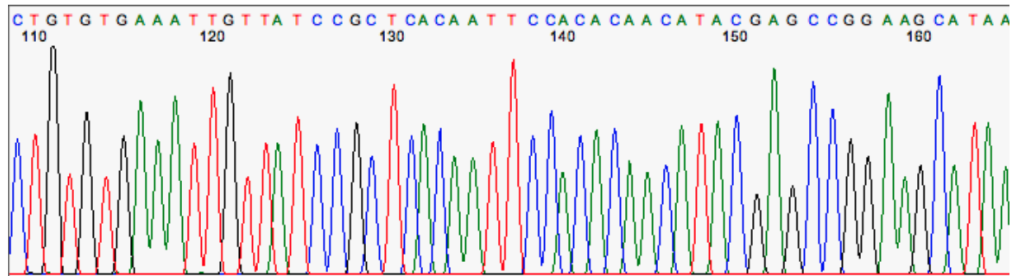
grandes. Por ejemplo, la diferencia de peso molecular entre 100 pb y 101 pb es del 1%, mientras que entre 1.000 pb y 1.001 pb es sólo del 0,1%.

Cromatograma

Ruido

En teoría un cromatograma siempre debería ser perfecto, pero no siempre es así.

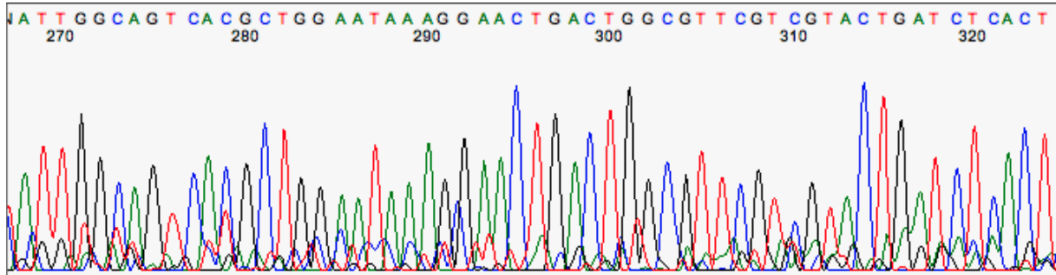
Un buen cromatograma:



22/05/2024

Cromatograma

Cromatograma con mucho ruido:



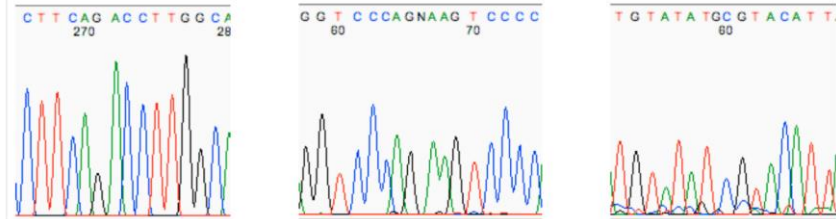
Por dímeros del cebador u otras amplificaciones no deseadas, cebadores mal diseñados, burbujas, capilares viejos etc...

La razón habitual de ruido de fondo es la presencia de algún dímero de cebador u otra amplificación no deseada presente en la muestra. Otras posibilidades son capilares de secuenciación viejos, burbujas en el capilar, cebadores mal diseñados o viejos (caducados). El método utilizado para eliminar los cebadores de PCR también puede ser una causa.

Cromatograma

Errores en el *basecalling*

Algunas veces el secuenciador automático no es capaz de colocar los picos correspondientes a las distintas bases a una distancia equidistante. Por ejemplo, esto sucede con frecuencia en el dinucleótido GA.



- Izquierda: mal espaciado bien interpretado.
- Centro: mal espaciado que introduce una N extra.
- Derecha: mal espaciado y fondo que introduce una base extra.

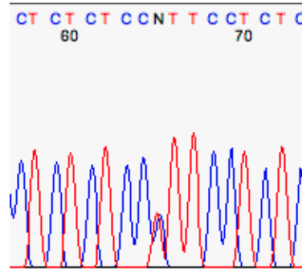
22/05/2024

Cromatograma

Heterocigotos

Los programas de *basecalling* suelen interpretar los heterocigotos como N.

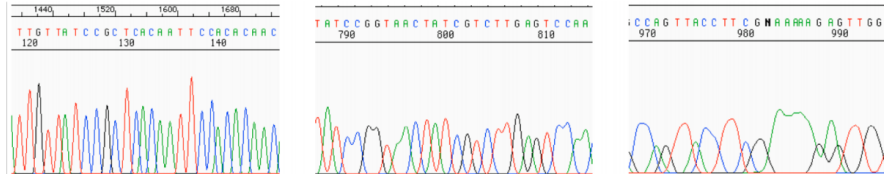
Hay programas especializados en detectar estos picos dobles y en etiquetarlos adecuadamente.



Cromatograma

Pérdida de resolución

Incluso las buenas secuencias pierden resolución al avanzar la secuencia, debido a la cromatografía. Este es uno de los motivos que hacen que las lecturas no tengan más de 700-800 pb.





Calidad de secuencias: fastq

Texto plano

Formatos simples

Secuencia en texto plano:

```
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTCTCCGGG
GCCACGGCCACCGCTGCCCTGGAGGGTACGCCCCACCGG
CCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCA
GCCTCCTGACTTTCCTCGCTTGGTAGTGGACCTCCAGGCCAGTGCC
GGGCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGG
AAGGCGCACCCCATCCGCGCGCCGGGACAGAATGCCCTGCAGGAA
CTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
```

22/05/2024

Recordemos.

Secuencia en formato FASTA

En el formato FASTA la secuencia comienza con un nombre y una descripción. Esta línea se distingue porque siempre comienza con el signo '>'. A continuación, sigue la secuencia propiamente dicha con el formato en texto plano.

```
>nombre_secuencia1 descripción  
ACAAGATGCCATTGTCCCCGGCTCCTGCTGCTGCTCTCCGGG  
GCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCC  
GAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC  
CTCCTGACTTTCTCGCTTGGTGGTTTGAGTGGACCTCCAGGCCAGT  
GCCGGGCCCCCATAGGAGAGGAAGTCGGGAGGTGGCCAGGCGGC  
AGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGACAGAATGCC  
CTGCAGGAACCTTCTTGGAAGACCTTCTCCTCGCAAATAAACCTC  
ACCCATGAATGCTCACGC
```

Multi-fasta

Secuencia en formato FASTA

En el formato FASTA la secuencia comienza con un nombre y una descripción. Esta línea se distingue porque siempre comienza con el signo '>'. A continuación, sigue la secuencia propiamente dicha con el formato en texto plano. Pueden incluirse varias secuencias en un mismo fichero. Es un formato muy utilizado.

```
>nombre_secuencia1 descripción
ACAAGATGCCATTGTCCCCGGCTCCTGCTGCTGCTCTCCGGG
GCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCC
GAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCTGACTTTCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGT
GCCGGGCCCCATAGGAGAGGAAGTCGGGAGGTGGCCAGGCGGC
AGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACCTTCTCTGGAAGACCTTCTCTCTGCAAATAAACCTC
ACCCATGAATGCTCAGC

>nombre_secuencia2 descripción
ACAAGATGCCATTGTCCCCGGCTCCTGCTGCTGCTCTCCGGG
GCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCC
GAGACAGCGAGCATATGCAGGAAGCGGCAGGA
```

22/05/2024

¿Qué se genera en la secuenciación?

Las secuencias se generan dando lugar a archivos fastq, archivos fasta que poseen parámetros de calidad para las secuencias.

```
@SRR6043417.1 1 length=125
CCGGGGGGCGGAGACGGGGAGGAGGAGGACGGACGGACGGACGGGGCCCCCGAGCCACCTTCCCGCGGGCCTTCCAGCCGTCGCGGCGCACCGCCGCGGTGAA
+SRR6043417.1 1 length=125
/;<B8FFBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF7B/FF
@SRR6043417.2 2 length=125
GGACCTTGAGAGCTTGTGGAGTTCTAGCAGGGAGCGCAGCTACTCGTATACCTTGACGAAGACCGTCTCTCTATCGGGGATGGTCGCTCTTCGACCGAGCGCAGCTTCGGGA
+SRR6043417.2 2 length=125
/;<B8B8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFF/
```

Encabezado
/ID

Secuencia

```
>CDX67397 gene=BnaA07g14370D
ATGGGATCTCCACCATCTGTATTATGTCTAATCTTATTACTCTTGTTCATCATATTCTTGTGTACAA
GCCAGCGAGCACCACGTGGGTTTTGTGATACCCTAATCCACTGTGGCAACATCTCAGTTGGCTTCCCTTT
CTGGGGAGAGAATCGTCATCAAGATGCGGTCATCCTTCTCTGAAACTTAAGTCAACAAACACTCAAAC
ACAACCACTCTTTTCATCTCAGGCTACAACACAGTCTCTCCATATAGACAACACGACCAACATCATTC
GACTTTTCAGACAAGATTTCTCAACTTCTTCTGCTCCGCTTCATTCTCTCCGACCTCTGCCTTCTGT
```

22/05/2024

Las secuencias obtenidas tendrán el siguiente formato, generalmente **.ab1** en el caso de las secuencias **Sanger** y **.fastq** para las **tecnologías de segunda generación**. Las tecnologías de tercera generación producen un tipo de archivo que deberemos convertir a **.fastq** o el mismo software del secuenciador traducirá. Desde estos archivos generaremos archivos **.fasta** que son los que trabajaremos en los subsiguientes análisis. Los archivos **.fastq** son secuencias con calidad, que nos permitirá evaluar y preprocesar nuestras secuencias antes de realizar cualquier análisis.

Evaluación de la calidad y preprocesado

1. La primera línea siempre comienza con @ y contiene información sobre la descripción de la secuencia y una identificación única.
2. La segunda línea contiene la secuencia sin procesar.
3. Por lo general, la tercera línea solo contiene el signo «+» o el signo «+» y una repetición de la identificación de la secuencia.
4. La cuarta línea contiene los valores de calidad y debe contener el mismo número de caracteres que la línea de secuencia. Se codifican código ASCII dependiendo de la tecnología de secuenciación

```
@SRR6043417.1 1 length=125
CCGGGGGGCGGAGACGGGGAGGAGGACGGACGGACGGACGGGGCCCCCGAGCCACCTCCCCCGGGCCTCCAGCCGTCGGAGCCGGTCGGCGCACCCGCCGCGTGGA
+SRR6043417.1 1 length=125
/ <BBFFBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF7B/FF
@SRR6043417.2 2 length=125
GGACCTTGAGAGCTTGTGGAGGTTCTAGCAGGGAGCGCAGCTACTCGTATACCTTGACCAAGACCGTCTCTCTATCGGGATGGTCGTCCTTCGACCGAGCGCGCAGCTTCGGGA
+SRR6043417.2 2 length=125
/ <BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFF/
```

22/05/2024

La estructura que observamos en los archivos .fastq es la que se observa en la imagen. Para cada una de las lecturas tenemos:

- una cabecera o ID única
- seguida de la lectura de la secuencia
- en la tercera línea tenemos una línea que contiene el símbolo "+" o el símbolo y una repetición del identificador
- por último la cuarta línea contiene la calidad, en formato de caracteres ASCII de forma individual para cada una de las bases leídas. Es importante tener en cuenta que esta codificación de la calidad es dependiente del método de secuenciación.

Evaluación de la calidad y preprocesado

- Las calidades de las secuencias se almacenan en el archivos RAW (archivo de salida o archivo en formato crudo) en formato .fastq
- Cada una de las técnicas de secuenciación **siguen su propio criterio de puntuación Phred**

```
@SRR6043417.1 1 length=125
CCGGGGGGCGGAGACGGGGAGGAGGAGGACGGACGGACGGACGGGGCCCCCGAGCCACTTCCCCCGGGCCTTCCAGCCGTCCCGGAGCCGGTCGGGGCGCACCGCCCGGTGAA
+SRR6043417.1 1 length=125
/;<BBFFBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFF7B/FF
@SRR6043417.2 2 length=125
GGACCTTGAGAGCTTGTTGGAGGTTCTAGCAGGGGAGCGCACTACTCGTATACCTTGACCGAAGACCGGTCCTCTCTATCGGGGATGGTCGTCTTCGACCGAGCGCAGCTTCGGGA
+SRR6043417.2 2 length=125
/;<BBBFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFF/
```

22/05/2024

Con esto hay que tener claro dos conceptos: Las calidad de las secuencias se almacenan en archivos RAW, en formato .fastq (con calidades) y que el sistema de calidad es dependiente de cada sistema de secuenciación. Para facilitar el análisis y la interpretación de los resultados, estos valores se suelen cambiar a una escala normalizada por todas las tecnologías de secuenciación, la **escala Phred**. Phred era originalmente un programa de *base calling* pero ahora es sobre todo conocido como la **escala del valor de calidad**. Esta se define como **Phred Score = $-10 \times \log$ (probabilidad de error)**.

Phred Score

Las secuencias generadas por cada método de secuenciación pueden contener errores.

En el año 1990 se calculó por primera vez la probabilidad de estos errores por posición dentro de la secuencia. Se denominó Phred.

$$Q = -10 \log_{10} P \qquad P = 10^{\frac{-Q}{10}}$$

Puntuación de calidad Phred	Probabilidad de error por base	Fiabilidad por base
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

22/05/2024

Debido a que las secuencias leídas pueden tener errores se estableció un código de puntuación llamado "Phred" que cuantifica la calidad de la secuencia base a base. El criterio de puntuación es el que se indica en la tabla. Por ejemplo, si la calidad de la secuencia es de media 20, encontraremos un error cada 100 bases leídas, encontrándonos en una fiabilidad media de secuencia del 99%.

Phred Score



22/05/2024

Cada una de las tecnologías de secuenciación sigue su propio criterio de Phred-score. Todos los métodos van desde 0 a 40 (o 41) puntos de calidad a excepción de la tecnología PacBio, que se mueve en una escala que llega a 93 puntos totales. **Es importante tener en cuenta que los caracteres para cada tecnología no son los mismos, y esto es crítico a la hora de preprocesar las muestras.** Si vamos a preprocesar muestras de Sanger no podemos utilizar el criterio de Illumina 1.5, ya que para Illumina las bases con mayor valor de Sanger, son las más bajas.

Phred Score

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

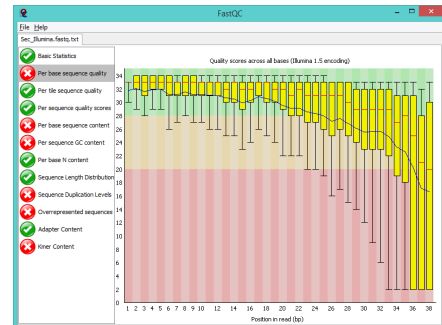
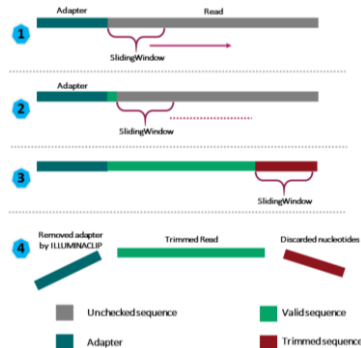
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

22/05/2024

Paso 1: Control de calidad de secuencias



Filtrado de secuencias por calidad, búsqueda y limpieza de adaptadores.

22/05/2024

Al igual que ocurría con la secuenciación por Sanger pero por causas distintas, una lectura o *read* tendrá distintas calidades a lo largo de la secuencia. Es por ello que el **primer paso** en nuestro análisis bioinformático es el de **control de calidad de las secuencias**. En este paso, se evalúa la calidad y **confiabilidad** de las lecturas **obtenidas**. Este proceso garantiza la integridad de los datos y ayuda a identificar posibles problemas que podrían afectar los resultados posteriores del análisis.

Normalmente conocido como **pre-procesado de las muestras**, en este paso se realiza un análisis que evalúa la calidad de las muestras mediante distintas métricas como son la **calidad media a lo largo de las lecturas**, el **número de lecturas**, la **presencia de contaminantes**, o la **presencia de restos de adaptadores e índices**. Posteriormente se deciden los pasos a seguir, generalmente involucrado el filtrado por calidad, la eliminación de adaptadores o el recorte de las lecturas en las zonas de baja calidad. Por lo general, las primeras bases no serán óptimas y a lo largo de la lectura la calidad ira decayendo.

Una baja calidad, es una menor fiabilidad de que los nucleótidos han sido correctamente leídos. Por tanto, no podemos confiar en que la lectura haya sido correcta, si la calidad no esta por encima de un determinado umbral. Cuanta menor sea la calidad de nuestra lectura, mayor será la probabilidad de arrastrar errores de lectura que en última instancia dificultarán el procesado de los datos.

Es por tanto una buena praxis “limpiar” las lecturas para quedarnos únicamente con la bases que tienen poca probabilidad de ser incorrectas. Para ello, necesitamos primero evaluar la calidad de nuestras lecturas.

Una forma visual de evaluar la calidad de los archivos .fastq es generando lo que se conoce como gráficos “fastqc”. Aquí podemos observar la calidad media por base de un conjunto de secuencias desde un archivo .fastq. Aquí podemos ver como a lo largo de las 38 bases de las secuencias del archivo la calidad media varía, estando en el extremo 5’ en torno a los 32 y cayendo al final por debajo de 30. Este perfil es el que comúnmente encontraremos, ya que **durante los procesos de secuenciación existe un error sistemático en los extremos 3’ (tres prima) donde la calidad siempre es más baja**. La bajada de calidad en el extremo 3’ de las lecturas que provienen del secuenciador Illumina se debe a una serie de factores relacionados con la química y el proceso de secuenciación. **Alguno de estos factores se debe a la degradación del cebador durante el proceso utilizado para iniciar la secuenciación, lo que provoca una disminución de la señal y, por lo tanto, una menor calidad de la lectura a medida que se avanza en la secuencia**. La probabilidad de errores de incorporación de bases aumenta en el extremo 3’ de la lectura debido a la menor eficiencia de la enzima polimerasa y a la mayor acumulación de productos de reacción secundarios. De igual manera, **los secuenciadores Illumina tienen un número limitado de ciclos de secuenciación**, lo que significa que las lecturas más largas, que incluyen el extremo 3’, tienen menos ciclos disponibles para su secuenciación completa, lo que puede conducir a una menor calidad en esa región. En menor medida, las estructuras secundarias del ADN, como los bucles en horquilla, pueden dificultar la extensión del cebador y la incorporación de bases en el extremo 3’, lo que resulta en una menor calidad de la lectura.

De igual manera, **durante el análisis de pre-procesado de las muestras, nos fijaremos en la presencia de adaptadores e índices que hayan quedado en nuestro archivo fastq**.

Finalizada la primera ronda de limpieza de lecturas realizaremos un nuevo análisis de calidad para ver como ha mejorado nuestro dataset de lecturas. Es común realizar

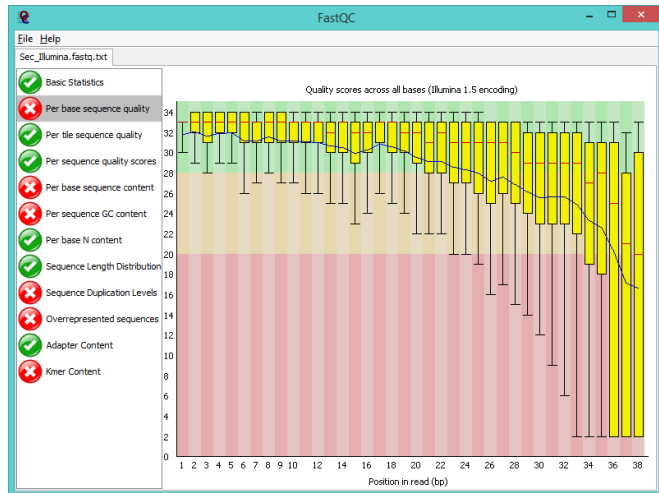
varias rondas de evaluación y limpieza hasta que estemos satisfechos con el resultado.

Paso 1: El Control de calidad de las secuencias

Las secuencias en bruto (*raw data*) pueden tener algunas regiones que podrían ser problemáticas, por ejemplo, secuencias vectoriales o adaptadoras y que sería aconsejable eliminar para evitar problemas con los análisis posteriores.

Algunos de estos problemas son:

- Adaptadores
- Baja calidad
- Baja complejidad
- Contaminantes
- Duplicados
- Error de corrección



Adaptadores

La principal diferencia práctica, en el contexto del análisis de secuencia, entre los vectores y los adaptadores es la longitud de la secuencia, los adaptadores son secuencias cortas y son comunes en las lecturas de NGS. Se añaden durante la creación de librerías. Se pueden usar para separar experimentos o muestras dentro del secuenciador. Para los vectores largos podríamos usar *Blast*, pero buscar los adaptadores, que son cortos, con el algoritmo *Blast* estándar no es el mejor enfoque. Es mejor usar el algoritmo *Blast-short* también implementado por el software *NCBI Blast*.

Cuando los adaptadores tienen menos de 15 pares de bases, los algoritmos utilizados por los alineadores pueden fallar. Una alternativa en este caso es buscar coincidencias exactas o usar el software *cutadapt*.

22/05/2024

Cuando las lecturas de secuenciación salen del secuenciador, suelen contener información adicional en forma de adaptadores, índices y otros elementos. Estos son componentes clave en el proceso de secuenciación de nueva generación (NGS, por sus siglas en inglés).

Adaptadores:

Los adaptadores son secuencias cortas de ADN que se añaden a los extremos de las moléculas de ADN durante la preparación de la biblioteca.

Estos adaptadores suelen contener secuencias que son reconocidas por las plataformas de secuenciación y permiten la unión de las moléculas de ADN al soporte sólido (como una lámina de flujo en la tecnología Illumina).

Índices:

Los índices son secuencias cortas de ADN únicas que se incorporan a cada muestra durante la preparación de la biblioteca.

Estos índices actúan como códigos de barras únicos para identificar cada muestra de manera única.

La presencia de índices permite la multiplexación, es decir, la secuenciación simultánea de múltiples muestras en una misma carrera de secuenciación.

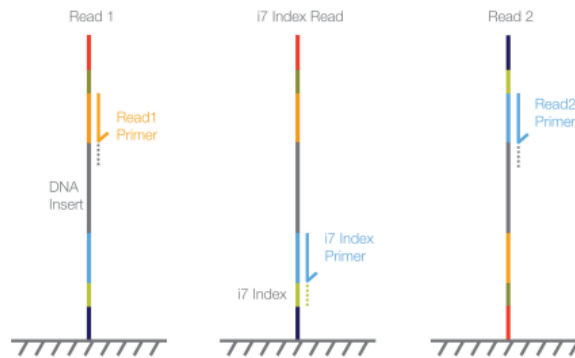
Secuencias de control de calidad (QC):

En algunas lecturas, se pueden encontrar secuencias específicas que se utilizan para evaluar la calidad de la lectura en términos de precisión y confiabilidad.

Estas secuencias a menudo contienen patrones conocidos que se utilizan para calibrar y corregir posibles errores de secuenciación.

Estos elementos son esenciales para el procesamiento y análisis de los datos de secuenciación. Los adaptadores e índices son particularmente útiles para la asignación de lecturas a muestras específicas en experimentos de alto rendimiento, como la secuenciación de genomas completos o experimentos de ARN-seq.

Figure 1 Single-Indexed Sequencing



22/05/2024

Existen múltiples manuales donde se especifican las posibilidades de índices y adaptadores para distintas máquinas:
https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/indexed-sequencing-overview-guide-15057455-08.pdf

Limpieza o filtrado por calidad

Para algunos análisis podría ser aconsejable eliminar las regiones de mala calidad. Algunas personas desaconsejan el recorte de baja calidad de las secuencias, ya que incluso las regiones de baja calidad tienen información en ellas. Pero es común que el *software* que se utilizará más adelante en el pipeline no trate particularmente bien las regiones de baja calidad de las lecturas, en este caso es importante eliminar estas regiones. El enfoque habitual para deshacerse de las regiones de baja calidad es hacer un “análisis de ventana” estableciendo un umbral para la calidad. Prinseq (Schmieder y Edwards, 2011) o seq_crums pueden hacer esta limpieza. Alternativamente, Lucy (Chou y Holmes, 2001) puede limpiar las lecturas largas.

Si las lecturas no tienen calidad podemos estimar qué regiones tienen pobre calidad observando la densidad de Ns que se encuentran en la secuencia.

Duplicados

Idealmente, dos lecturas duplicadas deberían tener la misma secuencia y podríamos buscarlas simplemente buscando secuencias idénticas, pero debido a los errores de secuenciación podrían no ser idénticas, sino simplemente muy similares. Si tenemos un genoma de referencia, un método habitual para eliminar estos duplicados es eliminarlos una vez que los hayamos alineado con la referencia. Si no tenemos una referencia, podríamos buscar al menos lecturas que sean idénticas. El *software* PRINSEQ tiene un módulo para filtrar lecturas duplicadas idénticas.

Regiones de baja complejidad

Las lecturas de baja complejidad pueden afectar a varios análisis posteriores. Estas lecturas poco complejas pueden ser una carga especial para los ensambladores, por lo que, en algunos casos, podría ser recomendable filtrarlas. La distribución *NCBI Blast* incluye Polvo, un programa para enmascarar regiones de baja complejidad. *Ngs_crums* también tiene un ejecutable de filtrado de baja complejidad.

Contaminantes

En las muestras a analizar pueden existir diferentes tipos de contaminantes:

- Debido a la preparación de la muestra, por ejemplo: *E. coli*
- Mitocondrial y de cloroplastos en muestras genómicas
- *rRNA* en transcriptomas
- patógenos en muestras infectadas

Toma de decisiones

El uso o no de este tipo de programas dependerá por tanto de diferentes factores y no únicamente de la calidad de las secuencias. Será importante su posterior uso por otros programas o el tipo de trabajo que estamos realizando, secuenciación *de novo*, RNAseq, búsqueda de SNPs, etc.



22/05/2024

Análisis de calidad de secuencias mediante FastQC



22/05/2024

FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) tiene como objetivo proporcionar una forma sencilla de realizar algunas comprobaciones de control de calidad en los datos de secuencias en bruto procedentes de pipelines de secuenciación de alto rendimiento. Proporciona un conjunto modular de análisis que se puede utilizar para hacerse una idea rápida de si los datos presentan algún problema que se deba conocer antes de realizar cualquier otro análisis.

Las principales funciones de FastQC son:

- Importación de datos de archivos BAM, SAM o fastq (cualquier variante)
- Proporcionar una visión general rápida para indicar en qué áreas puede haber problemas
- Gráficos y tablas de resumen para evaluar rápidamente los datos
- Exportación de resultados a un informe permanente basado en HTML
- Funcionamiento sin conexión para permitir la generación automática de informes sin ejecutar la aplicación interactiva

FastQC – ejemplos de informes

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

[Good Illumina Data](#)

[Bad Illumina Data](#)

[Adapter dimer contaminated run](#)

[Small RNA with read-through adapter](#)

[Reduced Representation BS-Seq](#)

[PacBio](#)

[454](#)

¡Gracias!



Universidad
Internacional
de Valencia

universidadviu.com

De:
🌐 Planeta Formación y Universidades