Food and Agriculture Organization of the United Nations

Módulo 1

Plataformas de secuenciación y terminología de la secuenciación del genoma completo

Jette S. Kjeldgaard

DTU FOOD

Descargo de responsabilidad

La traducción de esta presentación del idioma inglés al francés fue realizada por la FAO en colaboración con el Instituto Nacional de Alimentos de la Universidad Técnica de Dinamarca (DTU), que actúa como Centro de Referencia de la FAO para la Resistencia a los Antimicrobianos en Dinamarca. Este trabajo se llevó a cabo en el marco del Proyecto de Seguridad Sanitaria Global de la FAO - Vigilancia de Alerta Temprana de Enfermedades Infecciosas, financiado por la United States Agency for International Development (USAID) (Agencia de los Estados Unidos para el Desarrollo Internacional).

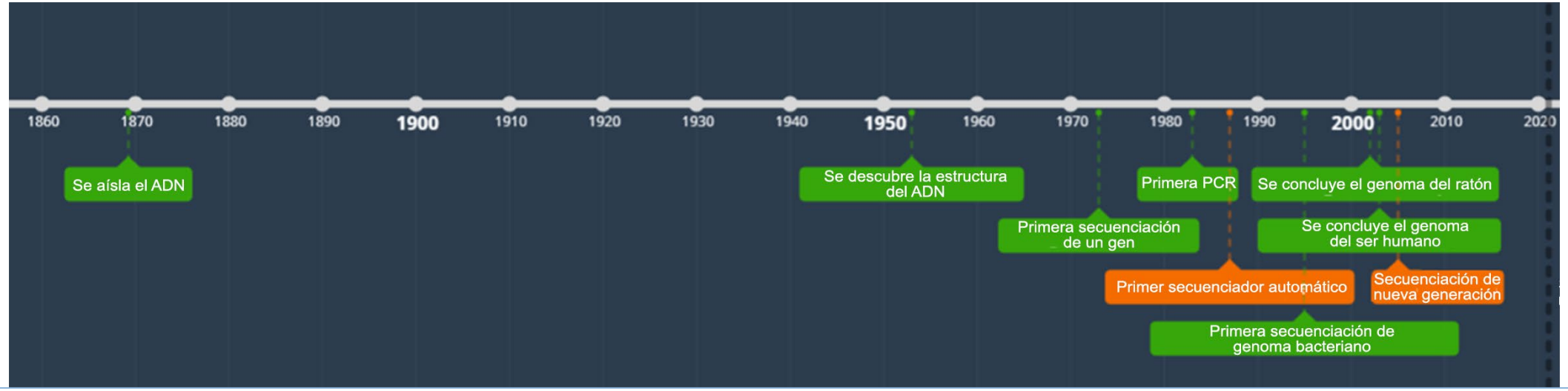
La versión original en inglés de esta presentación fue desarrollada por la DTU como parte de la subvención regional SeqAfrica del Fondo Fleming, financiada por el Departamento de Salud y Asistencia Social del Reino Unido.

Descripción general

- Breve historia de la secuenciación
- Plataformas de secuenciación
  - Lecturas cortas
  - Lecturas largas
- Pros y contras
- Parámetros de secuenciación para el control de calidad

Breve historia de la secuenciación

- En 1869 se aísla por primera vez el ADN.
- En 1953 se descubre la estructura del ADN.
- En 1972 se obtiene la primera secuencia completa de un gen.
- En 1983 se efectúa la primera reacción en cadena de la polimerasa.
- En 1987 se crea el primer secuenciador automático.
- En 1995 se obtiene el primer genoma bacteriano; en 2002, el del ratón; y en 2003, el humano.



Secuenciación de segunda (nueva) generación

- Es posible determinar la secuencia del genoma completo a la vez.

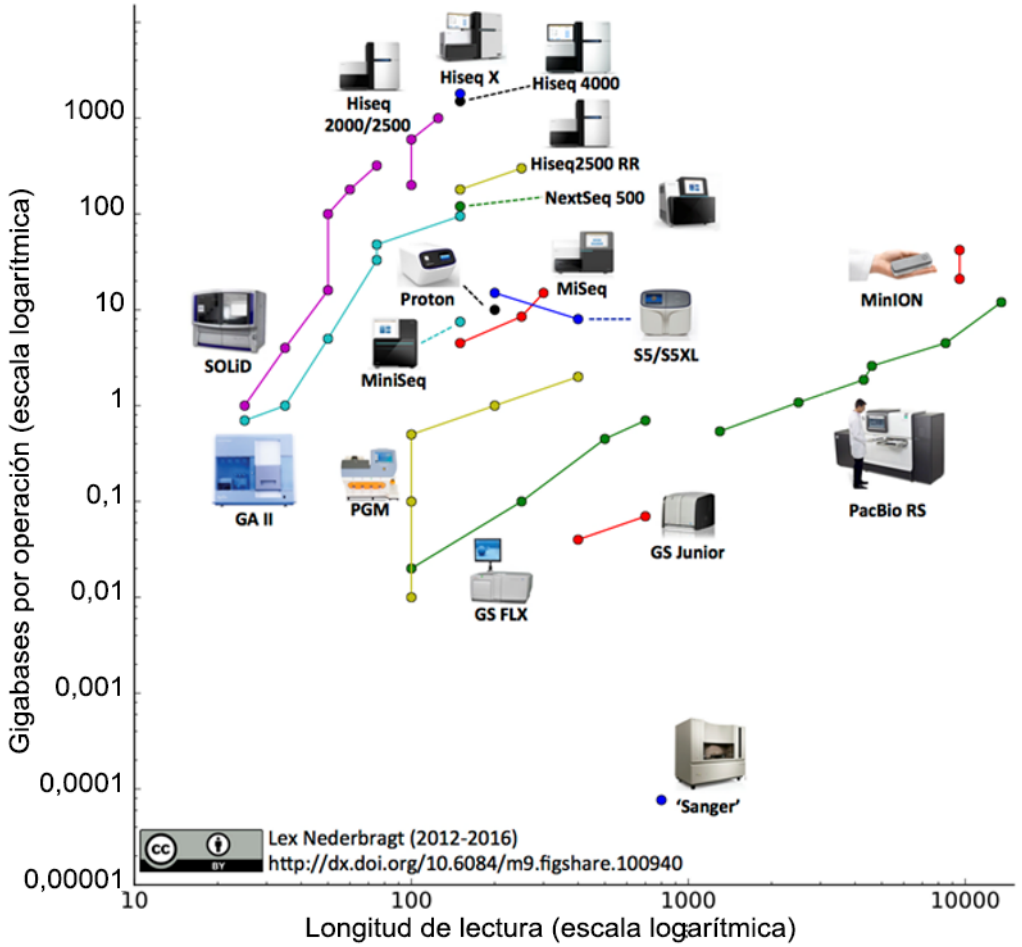
- El proceso general consiste en:
  - Fragmentar el genoma en numerosas partes más pequeñas.
  - Tomar aleatoriamente un fragmento de muestra.
  - Determinar la secuencia del fragmento mediante una de las numerosas tecnologías posibles.

Plataformas de secuenciación

- **Tecnologías de lectura corta** (50 - 300 bases)
  - Illumina (MiSeq, HiSeq, etc.)
  - 454
  - Ion Torrent

- **Tecnologías de lectura larga**
  - Pacific Biosciences (PacBio) (~20 kb)
  - Oxford Nanopore Technologies (MinION) (hasta 200 kb)

Plataformas de secuenciación II

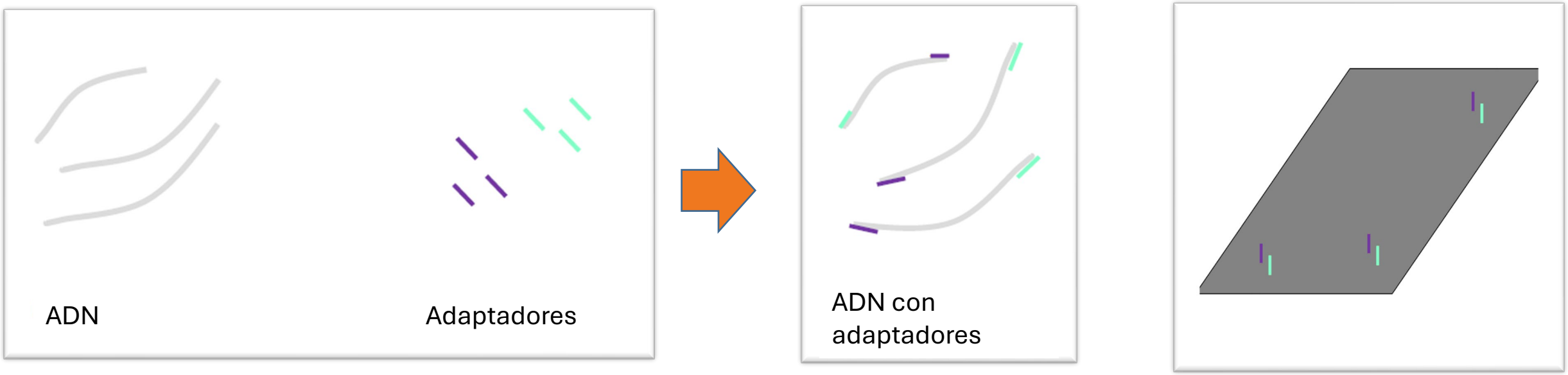


Características que se han de considerar:

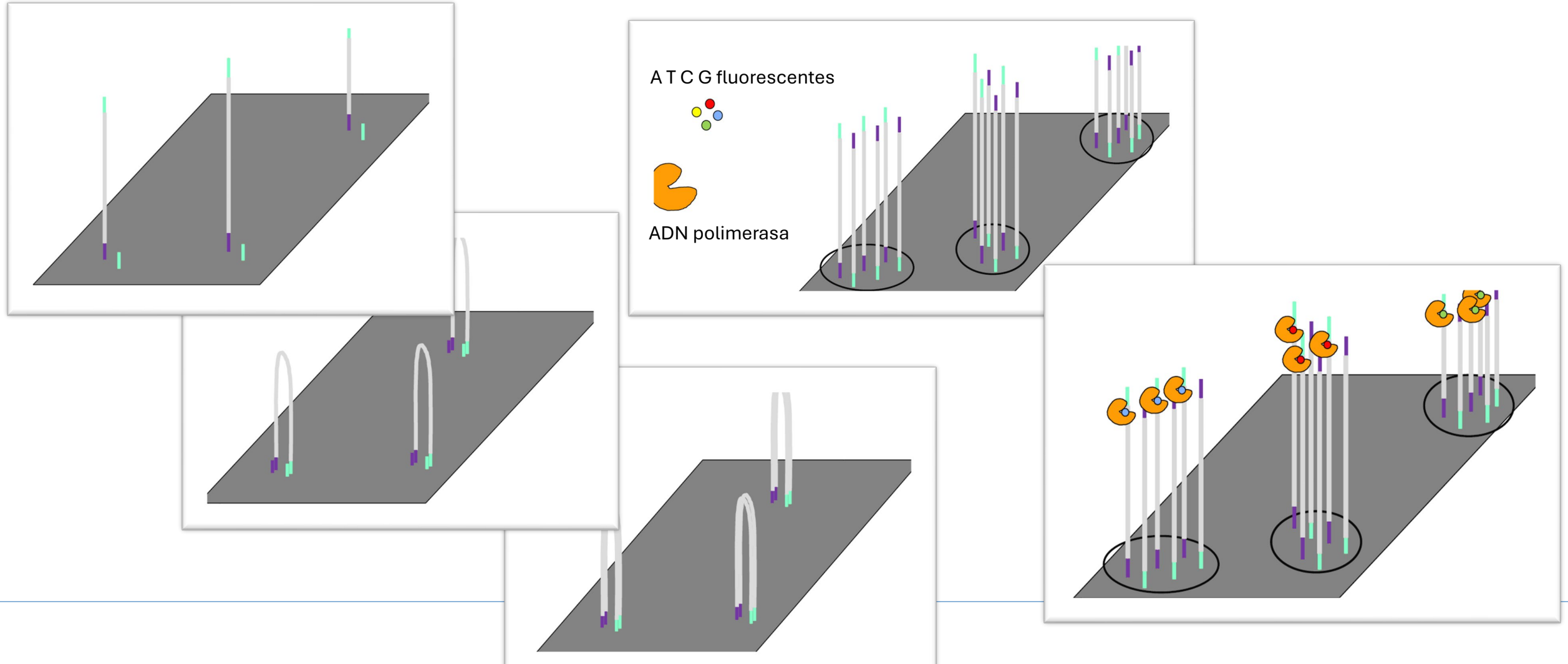
- Rendimiento/capacidad
- Tiempo de ejecución
- Costos de instrumentación y servicio
- Costo y disponibilidad de reactivos
- Infraestructura
- Capacitación del personal

Secuenciadores de lectura corta (Illumina)

- ¿Cómo funciona la secuenciación de Illumina?
  - Secuenciación mediante síntesis



Secuenciadores de lectura corta (Illumina)



Diferentes de Illumina

- HiSeq, NextSeq, MiSeq…
- El proceso químico en general es el mismo
- HiSeq ofrece más lecturas, requiere más tiempo y cuesta más
- MiSeq es más rápida y barata, pero ofrece menos lecturas
- NextSeq utiliza un sistema de dos colorantes (ciclos más rápidos y plataforma más económica [que HiSeq])

Comparación de tecnologías de lectura corta

Illumina

- Gran precisión
- Tasa de error ~0,1%
- Cierta subrepresentación en regiones ricas en AT y GC
- Alto rendimiento

Ion Torrent

- Tiempo de ejecución y flujo de trabajo más rápidos
- Requiere más tiempo del personal
- Tasa de error ~1%
- Problemas con homopolímeros

454

- Lecturas más largas (hasta 700 pares de bases)
- Tamaño de inserto más largo
- Costo elevado por megabase
- Tasa de error ~1%
- Problemas con homopolímeros
- Discontinuada

Secuenciación de tercera generación

- Lecturas largas
- Una sola molécula
- En tiempo real
- Plataformas:
  - Pacific Biosciences (PacBio)
  - Oxford Nanopore Technologies (MinIon)

Comparación

- Illumina
  - Lecturas cortas
  - Alta precisión

Costo:

- MiSeq: costo bajo de instrumento, costo mayor por gigabyte de datos
- NextSeq: costo medio-alto de instrumento, costo menor por gigabyte de datos
- HiSeq: costo alto de instrumento, costo bajo por gigabyte de datos

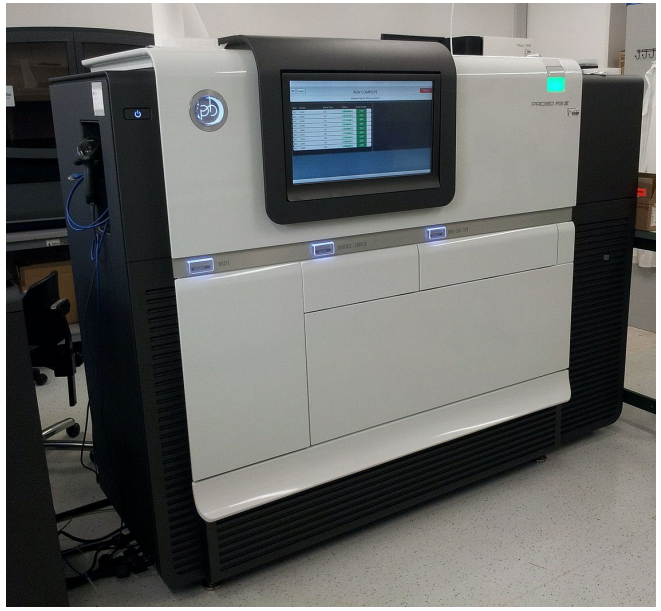
MiSeq Series

NextSeq Series

HiSeq Series

Comparación (II)

- PacBio
- Lecturas largas
- Tasas de error altas
- Precio:
  - Costo medio alto de instrumento
  - Costo muy alto por gigabyte de datos



Comparación (III)

- Nanopore
  - Lecturas muy largas
  - Tasas de error altas
  - Portátil

- Precio:
  - Costo muy bajo de instrumento
  - Costo alto por gigabyte de datos

MinIon

GridIon

Comparación de tecnologías de lectura larga

Pacific Biosciences

- Lecturas largas (Máx.: 50 kb. Promedio: 10-15 kb)
- Tasa de error ~15% (pase único)
- La precisión se puede mejorar al 99,999%
- Bajo rendimiento
- Caro (costo de adquisición)

Oxford Nanopore

- Lecturas muy largas (hasta 900 kb)
- Tasas de error elevadas (3% a 8%)
- Portátil

Sigue habiendo ciertas limitaciones en el procesamiento de

Producto de la secuenciación

Datos brutos – archivos de fastq

1 lectura: 4 líneas

¿Qué significa fastq? Fasta **(rápido) + puntuaciones de q**uality **(calidad)**

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG

 BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihiihggeeccdddcccacWTT^acc[ab `]`[ b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG

bb eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[ X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA

bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

Producto de la secuenciación

Datos brutos – archivos de fastq

1 lectura: 4 líneas

¿Qué significa fastq? Fasta (rápido) + puntuaciones de quality (calidad)

Encabezamiento/ID

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG

_BP`ccceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG

bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA

bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

Producto de la secuenciación

¿Qué significa fastq? Fasta (rápido) + puntuaciones de quality (calidad)

Datos brutos – archivos de fastq

1 lectura: 4 líneas

Secuencia de ADN

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG

_BP`cceggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG

bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA

bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

Producto de la secuenciación

¿Qué significa fastq? Fasta (rápido) + puntuaciones de quality (calidad)

Datos brutos – archivos de fastq

1 lectura: 4 líneas

Campo de nombre (opcional)

```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG

_BP`cccggcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab_`]`[_b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG

bb_eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[_X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA

bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh_fhggdgegeaceeacbdcbcc\^aa]``_^bb]bcccccbac_a^b
```

Producto de la secuenciación

¿Qué significa fastq? Fasta **(rápido) + puntuaciones de** q**uality (calidad)**

Datos brutos – archivos de fastq

1 lectura: 4 líneas

Puntuaciones de calidad

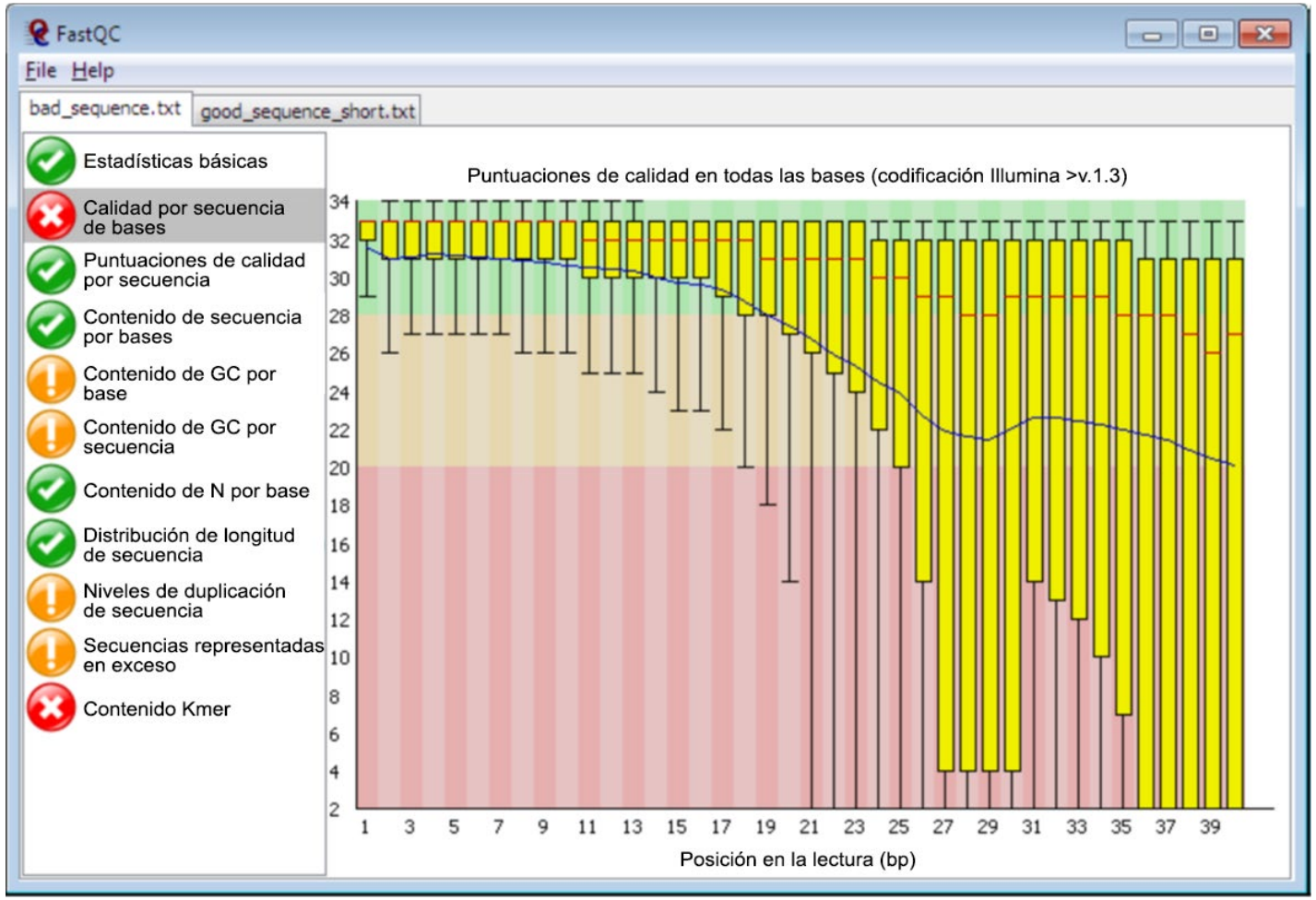
```
@FCC0CD5ACXX:1:1101:1103:2048#ACCGT/1
ACNGTGTTTTTAGTTATTGTTTTGTTAAGTTGGGTTTTTTGTACCCAATAGCCAACAAGCCGCCTTTATGGCGGTTTTTTTGTGCCTGAAAAGTGGGCG

 BP`cccegcegihiiighiifhihfddgfhi^efgfhhhhhegiiiiiiiihiihihggeeccdddcccacWTT^acc[ab `]`[ b`^BBBBBBB
@FCC0CD5ACXX:1:1101:1165:2058#ACGTT/1
ACGTTAGCAGAATCGCTTTCTGTTCGTTTTCCACCTGCGACAGACGCACCGGACCACGGTTGGCGAGATCGTCGCGCAGAATATCGGCGGCACGCTGCG

bb eeceefeggehhdagfghhiihfghighhffhifhhcghfdhiihafgdceba`a\aaccc^V]^baccaccXaaX^bbcccaac[ X]]a[aacX
@FCC0CD5ACXX:1:1101:1135:2082#AGCGT/1
AGCGTGACAAACATTTTATTGCGCCCGGTTTTATCCAGCTTGAATGCCTGACGAAAGAAGATGATGGTGACGACGATGGAGAGAACAATCAGCACCAGA

bbbeeeeefggfgiihgiigiiiiiiiffgifgeghiiihhfefffhhhfgh fhggdgegeaceeacbdcbcc\^aa]`` ^bb]bcccccbac a^b
```

Pasos posteriores a la secuenciación

- Control de la calidad
  - Ajuste de los adaptadores y las lecturas de baja calidad
  - Corrección de error
- Ensamblado
- Validación
- Análisis de datos

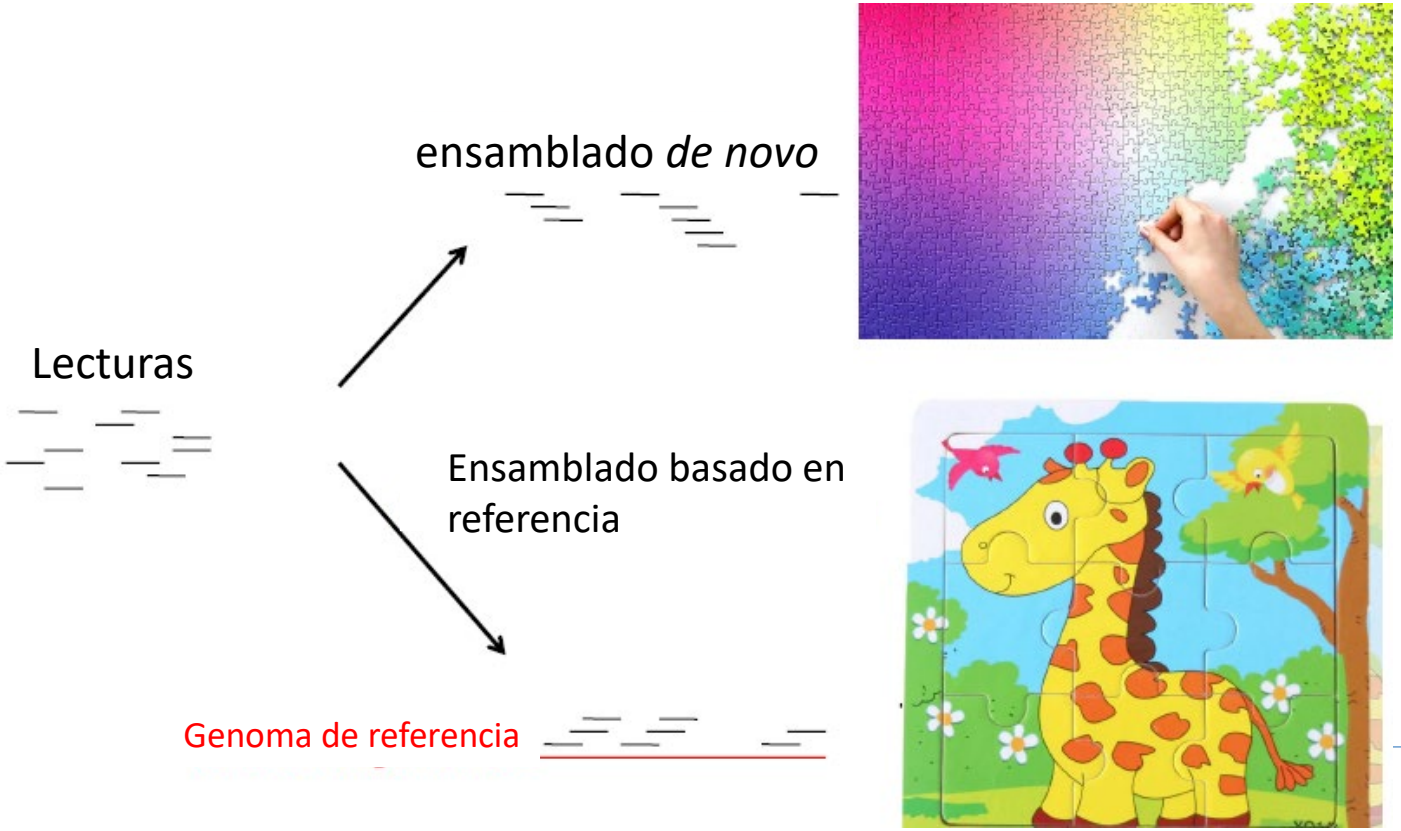
Control de la calidad

- Diferentes herramientas disponibles
- Control de calidad de las lecturas brutas
  - Porcentaje de bases de calidad
  - Número de lecturas
- Control de calidad del ensamblado
  - Parámetros del control de calidad
    - Tamaño del genoma
    - Número de cóntigos
    - N50
    - Cobertura



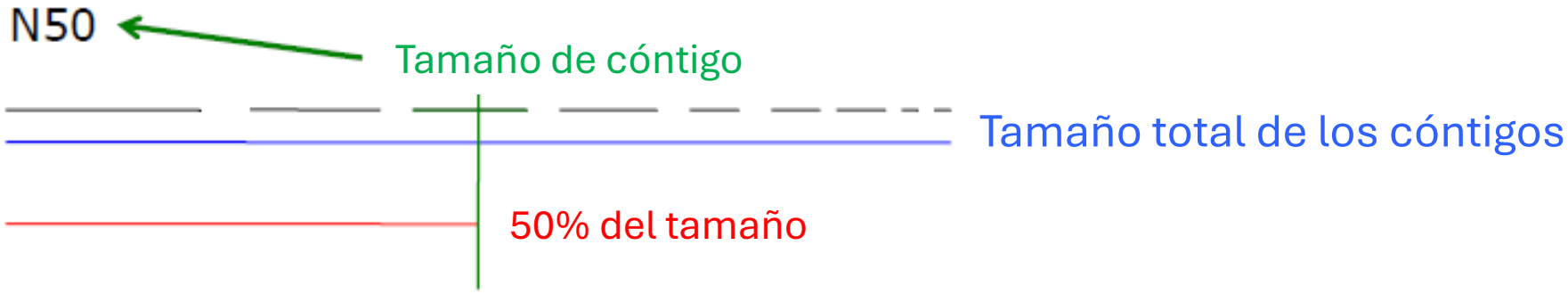
¿En que consiste el ensamblado?

- Ensamblar las lecturas brutas en fragmentos más largos de ADN denominados cóntigos
  - Cartografía con relación a referencia
  - *Ensamblado de novo*



Parámetros del control de calidad: directrices

- Número de cóntigos
  - <500 cóntigos (cuantos menos, mejor)
- Tamaño del genoma ensamblado (tamaño total de los cóntigos)
  - Cerca del tamaño previsto (+/- 20%)
- N50; más alto es mejor (>30.000)



- Más largo es mejor

Parámetros del control de calidad: directrices

- Cobertura
  - El número de veces que los datos de lectura cubren el genoma
  - Preferiblemente más de 20 veces

$$C = N \cdot \frac{L}{G}$$

- N: Número de lectura
- L: Longitud de lectura
- G: Tamaño de genoma (meta o ensamblado)

Ejemplo:
N = 5 millones
L = 100 bp
G = 5 Mbp
C = 5*100/5 = 100X
Como media, 100 lecturas cubren cada posición en el genoma.

- Profundidad
  - El número de lecturas que cubren un nucleótido particular en cada posición en el genoma

$$\frac{\text{reads}}{\text{site}} = \text{depth}$$

El equipo de secuenciación dependerá de las necesidades

- Gran variedad de equipos
  - Rendimiento
  - Disponibilidad de reactivos
  - Conveniencia del análisis de secuencias

¡El control de la calidad es importante!

- Garantizar la validez de las secuencias.
- Asegurar que la calidad es adecuada para realizar análisis ulteriores del genoma.
- Garantizar la comparabilidad de los resultados.

→ Control de calidad
Análisis del genoma

¡Gracias!