Máster en Bioinformática

Generación y mantenimiento de datos ómicos

Curso académico 2024-25

Universidad Internacional de Valencia

Dr. Jordi Tronchoni León
jordi.tronchoni@professor.universidadviu.com

15/05/2024

De:
Planeta Formación y Universidades

Tema 6
Métodos de secuenciación

15/05/2024

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

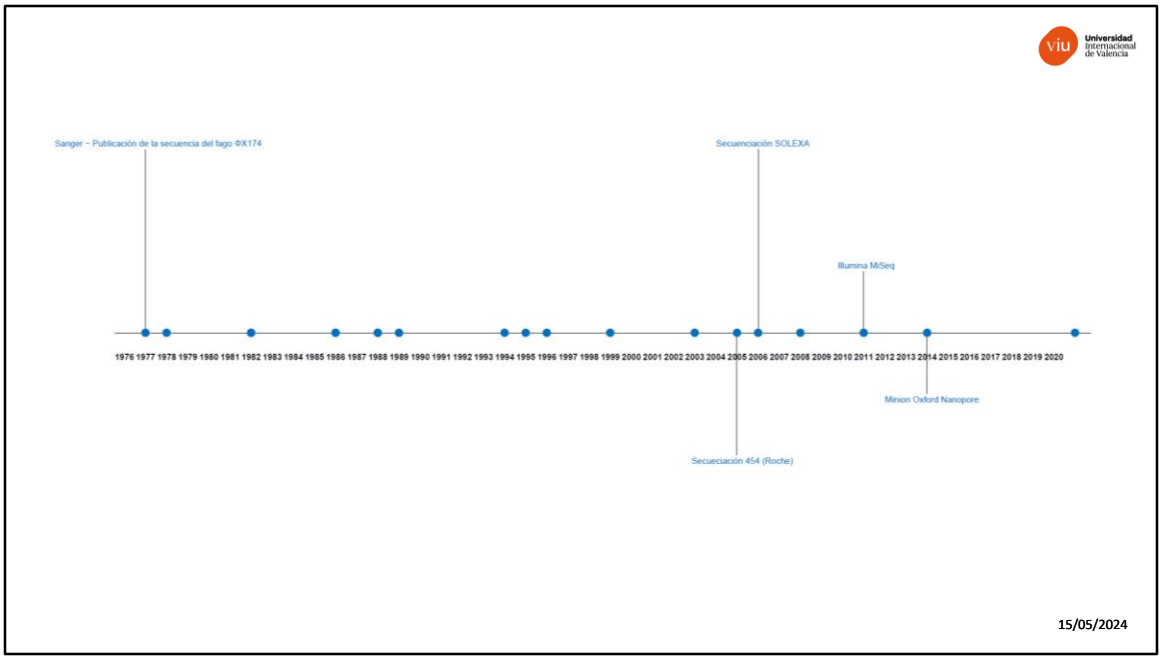
15/05/2024

Tema 6

Métodos de secuenciación

- Primera generación de secuenciadores
  - Sanger
- Segunda generación de secuenciadores
  - Pirosecuenciación 454
  - Secuenciación por síntesis (Illumina)
  - Ion Torrent
- Tercera generación de secuenciadores
  - PacBio (SMRT)
  - Nanopore
  - Comparación de plataformas de secuenciación

15/05/2024

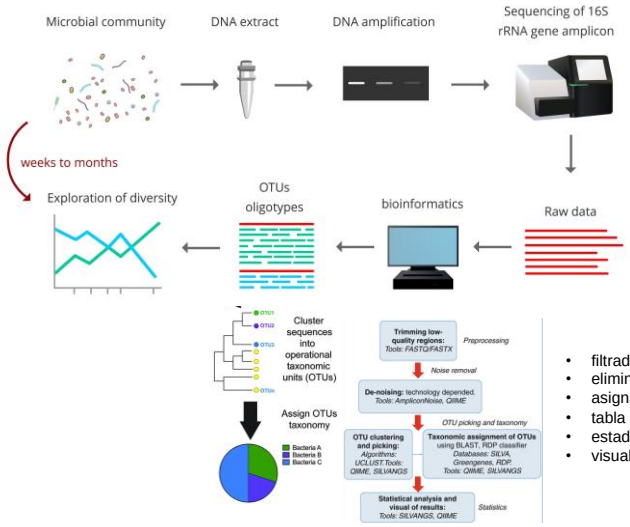


La secuenciación de ADN es el **proceso a través del cual se determina completamente la secuencia de bases de los nucleótidos** (adenina, timina, citosina y guanina) que conforman un fragmento de ADN.

En este tema vamos a conocer cuáles son los sistemas de secuenciación desde un punto de vista histórico y técnico, discutiendo cuáles son lo más importantes y en qué situaciones sería recomendable usar cada uno.

Workflow de bioinformática

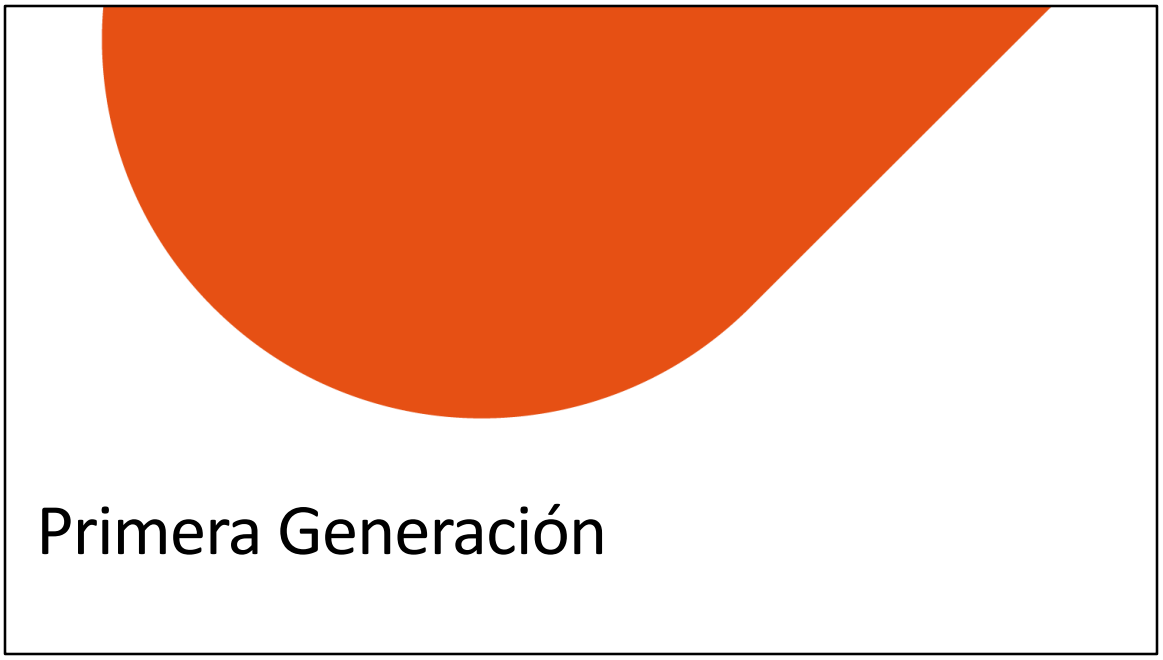
1. **Obtención de la muestra**
2. **Purificación del ADN**
3. **Secuenciación**
4. **Análisis bioinformático**



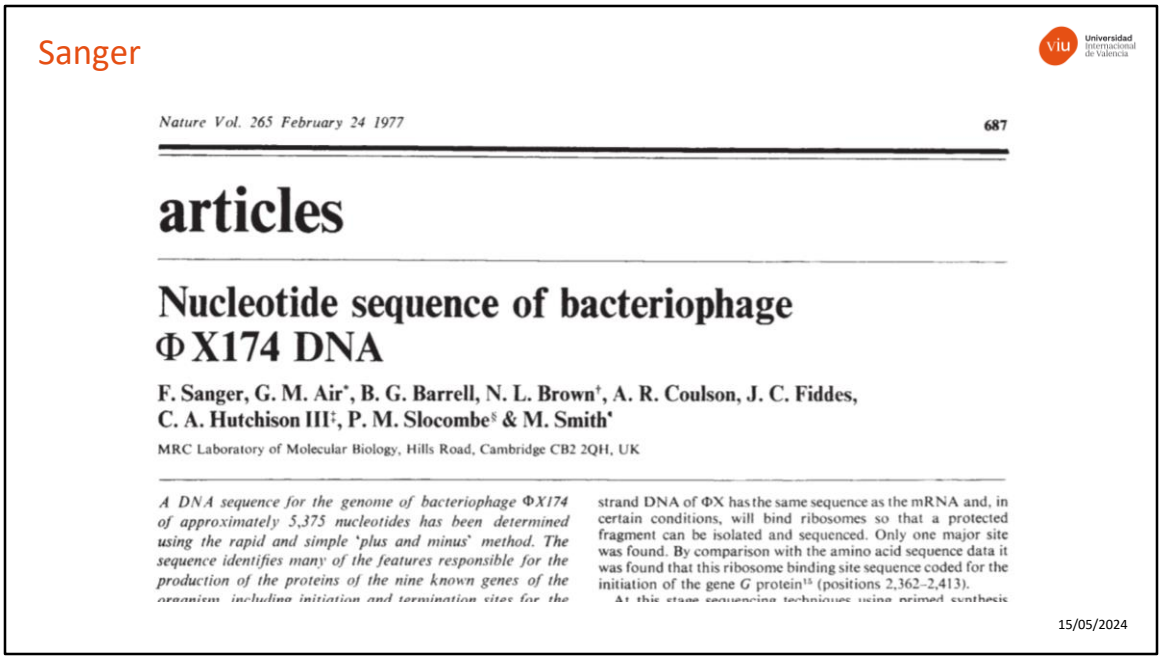
- filtrado y limpieza
- eliminación ruido
- asignación OTUs
- tabla taxonomía
- estadística y
- visualización

15/05/2024

Este sería el pipeline a seguir a la hora de secuenciar y analizar una muestra de ADN. La obtención de la muestra dependerá por tanto del estudio que queramos realizar. Si queremos realizar un estudio de un organismo concreto deberemos, por ejemplo, cultivarlo en placa. Por el contrario, si queremos realizar estudios de comunidades microbianas deberemos obtener las muestras desde entornos naturales, como suelo o agua de mar a distintas profundidades, o muestras humanas, como por ejemplo la mucosa oral. Posteriormente, la purificación del ADN se llevará a cabo utilizando distintos métodos según el organismo con el que estemos trabajando. El primer paso importante es la generación de los fragmentos de ADN previos a la secuenciación. A lo largo del tema destacaremos algunos métodos para ello. Después de la secuenciación utilizando alguno de los métodos que veremos a continuación realizaremos el análisis bioinformático.


Primera Generación

Sanger



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith‖

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[15] (positions 2,362–2,413).

15/05/2024

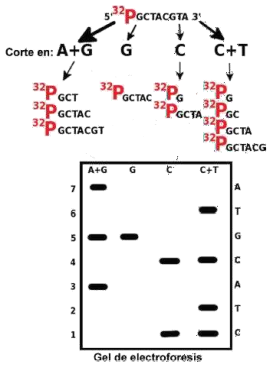
El método de secuenciación por terminación de cadena de Sanger se considera el pionero de las técnicas de secuenciación, abriendo las puertas a la conocida como primera generación de secuenciación. Con la publicación de este artículo en 1977 Sanger demostró un método fiable para lograr la secuencia de ADN del fago PhiX174, utilizando métodos de biología molecular básicos.

Este método permitía secuenciar fragmentos de hasta 900 pares de bases a pesar de ser muy laborioso y costoso.

Maxam-Gilbert

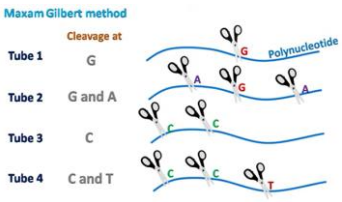
Maxam-Gilbert (secuenciación química)

- Primera técnica de secuenciación de ADN,
- Publicada casi a la vez (Maxam y Gilbert, 1977) que el método por terminación de cadena de Sanger
- Método algo peligroso: uso de productos químicos tóxicos y marcaje radiactivo
- No muy sencillo



Reactivos:

- dimetil sulfóxido o DMSO **(G)**
- ácido fórmico **(A+G)**
- hidracina más sales **(C)**
- hidracina **(C+T)**



15/05/2024

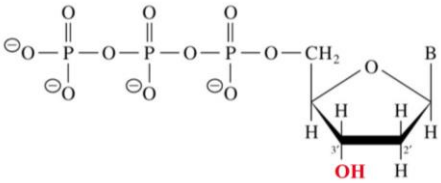
A pesar de considerarse el primero, en realidad ese mismo año Maxam-Gilbert propuso un método de secuenciación química. Por su naturaleza, más peligroso que el de Sanger debido al marcaje radioactivo.

Sanger

Método de terminación en cadena.
- El ADN se separa por tamaño
- Se utilizan nucleótidos modificados (ddNTPs)

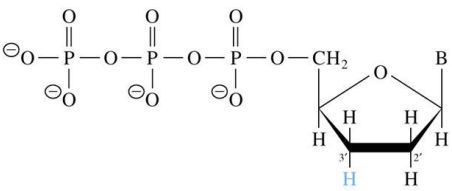
Segundo Premio Nobel en Química en 1980

dNTPs



Deoxinucleótidos

ddNTPs

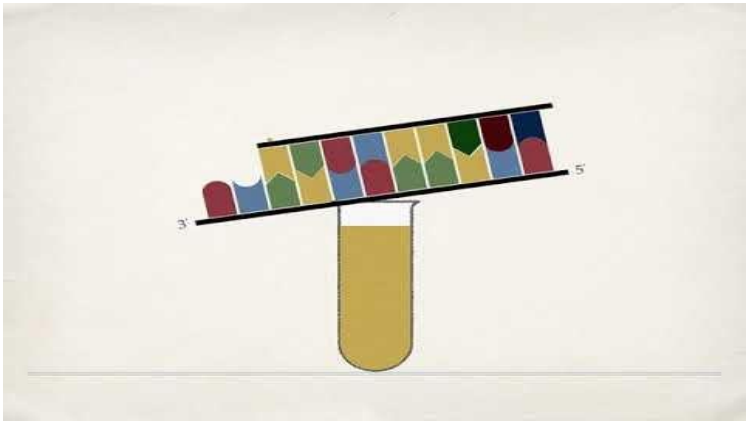


didesoxinucleótidos

15/05/2024

El método de Sanger es un método basado en síntesis, mejor conocido como ”método de terminación de cadena”. En este método se busca amplificar la cadena de ADN utilizando una polimerasa y oligonucleótidos modificados (ddNTPs: didesoxinucleótidos) donde se ha sustituido el grupo OH por un grupo H. Esto deriva en que a la hora de sintetizarse la cadena de ADN con uno de estos nucleótidos se termine el proceso de amplificación debido a que no puede unirse otro nucleótido después.

Sanger



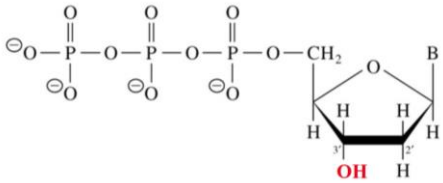
15/05/2024

Se divide la muestra en 4 tubos, introduciendo en cada uno de ellos todos los nucleótidos sin modificar en una proporción, y una de las cuatro bases modificadas en cada uno de los tubos. De esta forma se originará la síntesis de la cadena en cada caso hasta llegar a un oligonucleótido modificado, siendo esta una cadena de un tamaño determinado. Al terminar el proceso, el ADN de cada uno de los tubos de corre en un gel de poliacrilamida donde los fragmentos se separarán por tamaño (a menor tamaño más lejos llegará). Como sabemos qué nucleótido modificado hemos introducido en cada uno de los carriles, sabremos cual es la última base que se ha unido en cada caso. Así generamos un mapa, donde recorriéndolo de abajo a arriba obtendremos la secuencia del ADN.

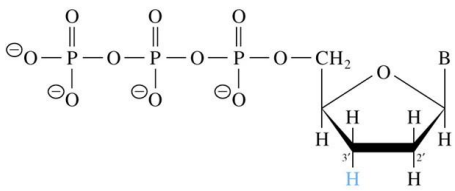
Sanger

Actualmente se sigue utilizando, el principio es el mismo pero con algunas importantes mejoras.

dNTPs



ddNTPs



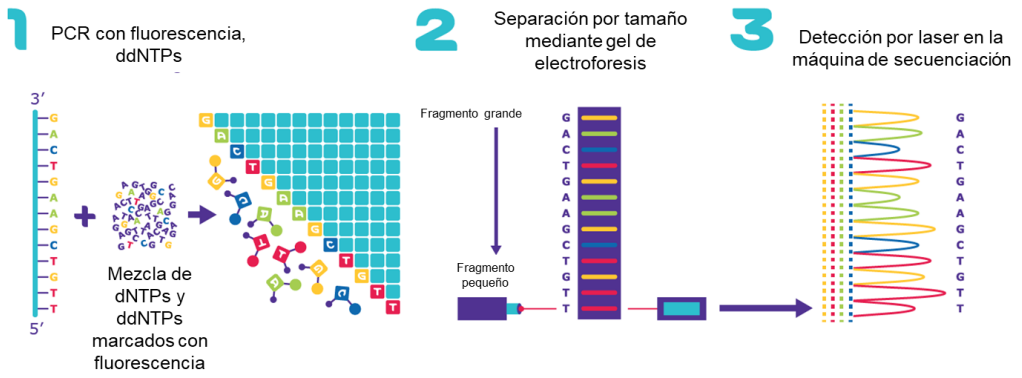
+

fluorocromos

15/05/2024

Actualmente el método Sanger se sigue utilizando siendo el método estándar para secuencias sencillas y relativamente cortas de ADN, pero ha evolucionado significativamente con la inclusión de fluorocromos adheridos en los nucleótidos.

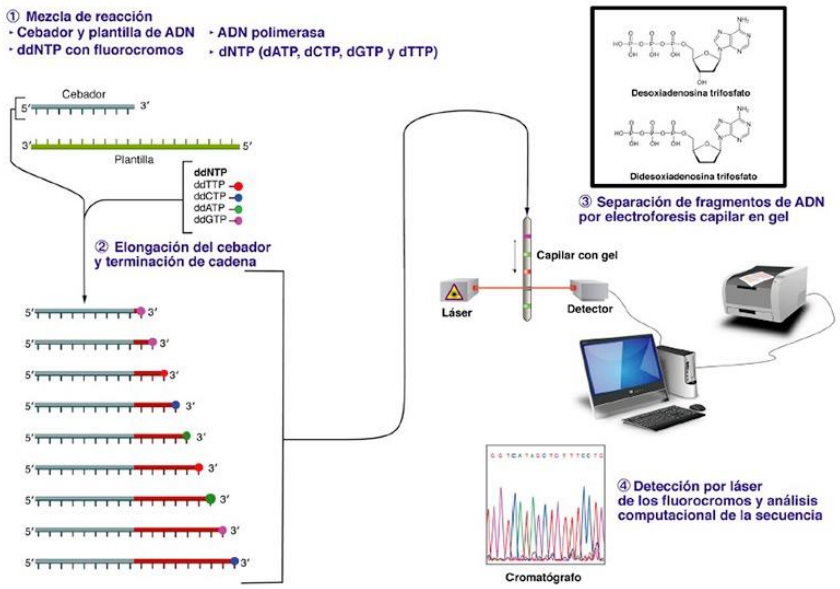
Sanger



15/05/2024

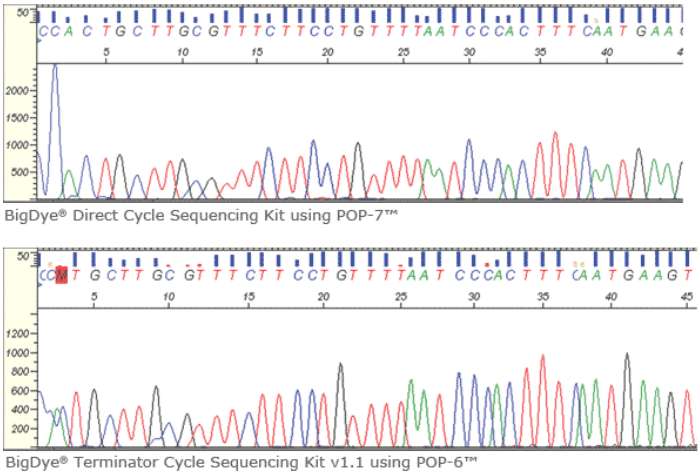
De esta forma podemos realizar la amplificación de la secuencia en un solo tubo y carril de gel. Añadiremos en distintas proporciones nucleótidos sin modificar y nucleótidos modificados a la reacción, junto a la secuencia y la polimerasa. Una vez finalizado correremos la amplificación sobre un gel y en lugar de observar la altura de la banda usaremos un detector de fluorescencia que guardará una señal, la cual es diferente para cada una de las bases, dependiendo de fluorocromo usado. De esta forma formaremos la secuencia del ADN.

Sanger



15/05/2024

Cromatograma



15/05/2024

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. Ambos parámetros 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.

Sanger

ABI 3730xl

Tiempo: 2-3h

Longitud reads: hasta 1000pb

Salida media: 100kpb

Fiabilidad: 99,999%

Características:
Barato
Secuenciación de genoma único

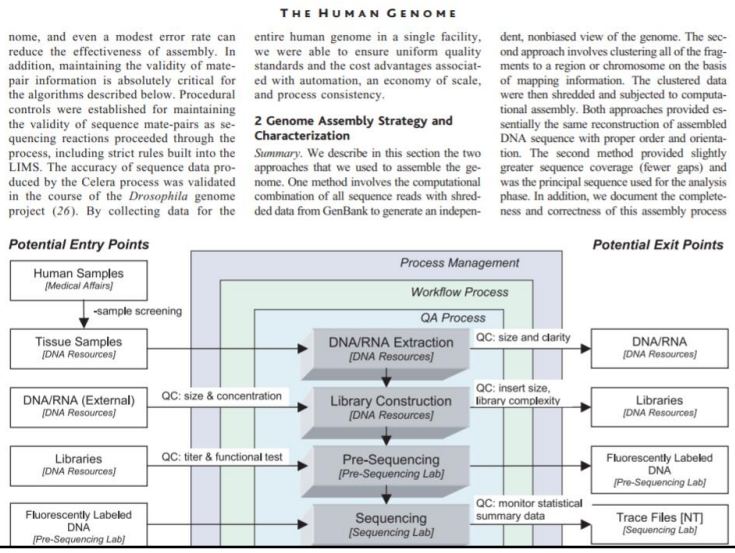
15/05/2024

La secuenciación Sanger se comenzó a desarrollar utilizando metodología de biología molecular sin la ayuda de dispositivos secuenciadores, lo que permitía solo la lectura de secuencias cortas de entre 800 y 1000 pb consumiendo mucho tiempo y recursos. Actualmente, el secuenciador más moderno para secuenciación Sanger es el ABI 3730xl DNA Analyzer. La secuenciación es un proceso rápido, entre 2 y 3 horas y la longitud de las lecturas generadas es de unos 1000 pares de bases, con una salida media de 100.000 pares de bases. De esta forma se pueden secuenciar a día de hoy genomas de un tamaño medio mediante el ensamblaje de las lecturas de 1000pb. Lo que destaca de este sistema de secuenciación es su fiabilidad, de un 99,999% y su bajo coste, abaratado por la evolución de la tecnología a lo largo de los años. Por contra, este sistema de secuenciación solo es útil para realizar la amplificación de genoma único, no permitiendo realizar estudios de comunidades.

Secuenciación del genoma humano

El primer método de secuenciación se denominó Clone by Clone, y estaba basado en el método de Sanger. Se leían secuencias de unos 1000pb de media.

El método Shotgun aplicado a Sanger mejoró el proceso. Aun así el proceso seguía siendo muy lento y caro.



THE HUMAN GENOME

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

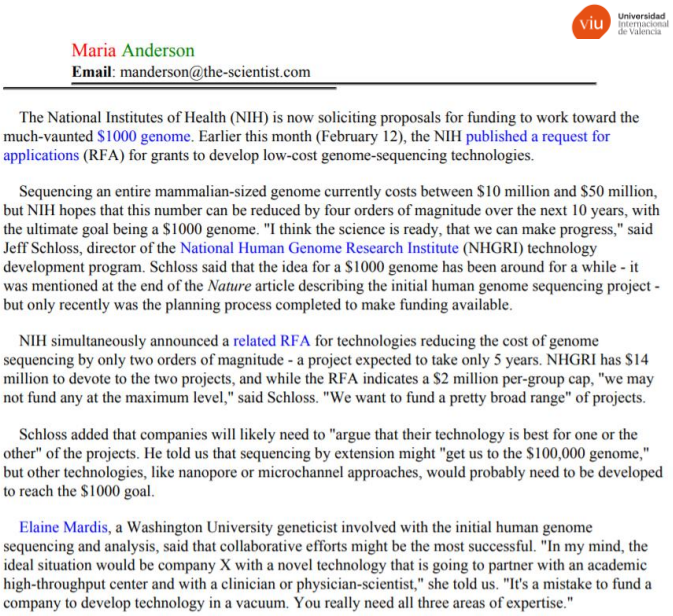
2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

El desarrollo del método Sanger abrió las puertas a unos de los proyectos más ambicioso de la biología molecular, la secuenciación del genoma humano completo. Se desarrolló una metodología basada en clones de fragmentos de ADN en plásmidos denominada ”Clone by Clone”, la cual se utilizó hasta los últimos pasos del proyecto, donde apareció la metodología de secuenciación Shotgun, de la que hablaremos más adelante. Este hito tardó en completarse más de 10 años y millones de dólares contando con la colaboración de más de 3.000 científicos.

the $1000 goal

En 2003, la Fundación para la Ciencia Craig Venter ofreció un premio de 500.000 $ para el que consiguiera secuenciar un genoma por menos de 1000 $



Maria Anderson
Email: manderson@the-scientist.com

The National Institutes of Health (NIH) is now soliciting proposals for funding to work toward the much-vaunted $1000 genome. Earlier this month (February 12), the NIH published a request for applications (RFA) for grants to develop low-cost genome-sequencing technologies.

Sequencing an entire mammalian-sized genome currently costs between $10 million and $50 million, but NIH hopes that this number can be reduced by four orders of magnitude over the next 10 years, with the ultimate goal being a $1000 genome. "I think the science is ready, that we can make progress," said Jeff Schloss, director of the National Human Genome Research Institute (NHGRI) technology development program. Schloss said that the idea for a $1000 genome has been around for a while - it was mentioned at the end of the *Nature* article describing the initial human genome sequencing project - but only recently was the planning process completed to make funding available.

NIH simultaneously announced a related RFA for technologies reducing the cost of genome sequencing by only two orders of magnitude - a project expected to take only 5 years. NHGRI has $14 million to devote to the two projects, and while the RFA indicates a $2 million per-group cap, "we may not fund any at the maximum level," said Schloss. "We want to fund a pretty broad range" of projects.

Schloss added that companies will likely need to "argue that their technology is best for one or the other" of the projects. He told us that sequencing by extension might "get us to the $100,000 genome," but other technologies, like nanopore or microchannel approaches, would probably need to be developed to reach the $1000 goal.

Elaine Mardis, a Washington University geneticist involved with the initial human genome sequencing and analysis, said that collaborative efforts might be the most successful. "In my mind, the ideal situation would be company X with a novel technology that is going to partner with an academic high-throughput center and with a clinician or physician-scientist," she told us. "It's a mistake to fund a company to develop technology in a vacuum. You really need all three areas of expertise."

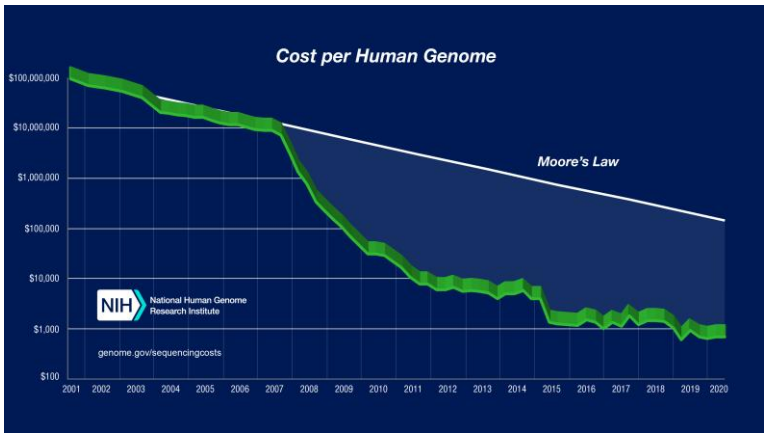
Haber podido desarrollar la secuenciación del genoma humano abrió un debate en el mundo científico: ¿puede abaratarse el coste y reducir el tiempo del proceso de secuenciación? Con esta pregunta la fundación Craig Venter ofreció un premio de 500.000 dólares al que consiguiese lograr este mismo hecho pero con un coste menor a los 1.000 dólares.

Tras varios años de competencia, en 2008, el equipo del Dr. James D. Watson del Baylor College of Medicine fue el ganador del premio. Su método, denominado "Whole Genome Shotgun Sequencing", logró secuenciar un genoma humano por un costo total de 385 dólares, superando con creces el objetivo establecido. El método Shotgun, también conocido como secuenciación aleatoria o secuenciación por escopeta, es una técnica fundamental en la secuenciación del ADN. A diferencia de otros métodos de secuenciación que leen el ADN de forma lineal, el método Shotgun fragmenta el ADN al azar en miles de segmentos más pequeños y luego los secuencía individualmente.

Reto alcanzado

Ley de Moore

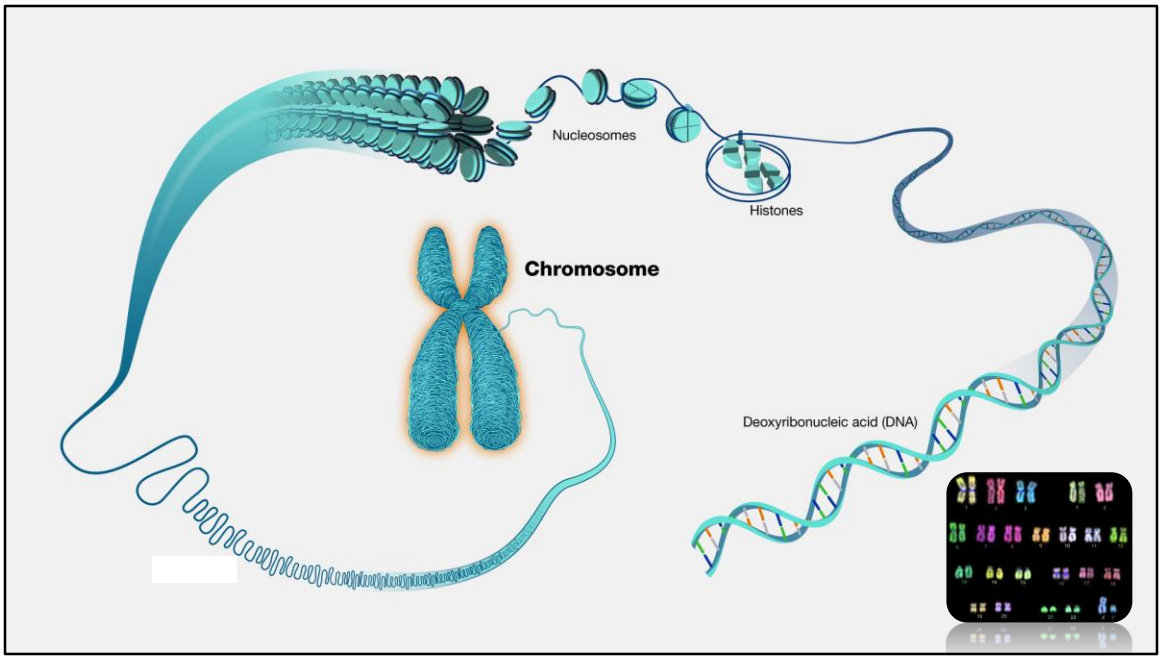
"El número de transistores por unidad de superficie en circuitos integrado se duplicará cada año"



15/05/2024

Esta propuesta fue a principios de siglo 21 y aquí podéis ver cómo ha ido evolucionando a lo largo de los años. Comúnmente, los avances tecnológicos siguen lo que se denomina Ley de Moore, donde se define que la capacidad de computación de un sistema se dobla cada dos años. Extrapolando esta idea a la tecnología de secuenciación, se puede predecir con una función lineal el coste en función del tiempo. Como se puede observar existe un punto de inflexión en 2007 donde el coste de la secuenciación comienza a bajar exponencialmente, hasta llegar a estar a día de hoy en un coste menor a los 1.000 dólares.

Hemos tardado 20 años en pasar de 3 billones de dólares y 13 años de trabajo a menos de 1000 dólares en un par de días de trabajo.



Recordemos que la secuenciación de ADN es el **proceso a través del cual se determina completamente la secuencia de bases de los nucleótidos** (adenina, timina, citosina y guanina) que conforman un fragmento de ADN.

A pesar de escuchar continuamente que el hito de secuenciar el genoma humano fue en el 2001, en realidad lo que obtuvimos fue el primer borrador del genoma humano. Hoy sabemos que aproximadamente ese borrador era del 90 % del genoma. Es decir, si cogemos todas las bases del genoma humano, desde la primera del primer cromosoma hasta la última del último cromosoma (el 100 % del genoma), el borrador tenia “huecos” (gaps) y solo cubría el 90 % del genoma. A esto lo llamamos **cobertura** del genoma, en este caso la *cobertura* del primer borrador del genoma humano era de aproximadamente el 90%.

¿y el 10 % restante? Sencillamente el método de secuenciación no permitía

resolverlo.

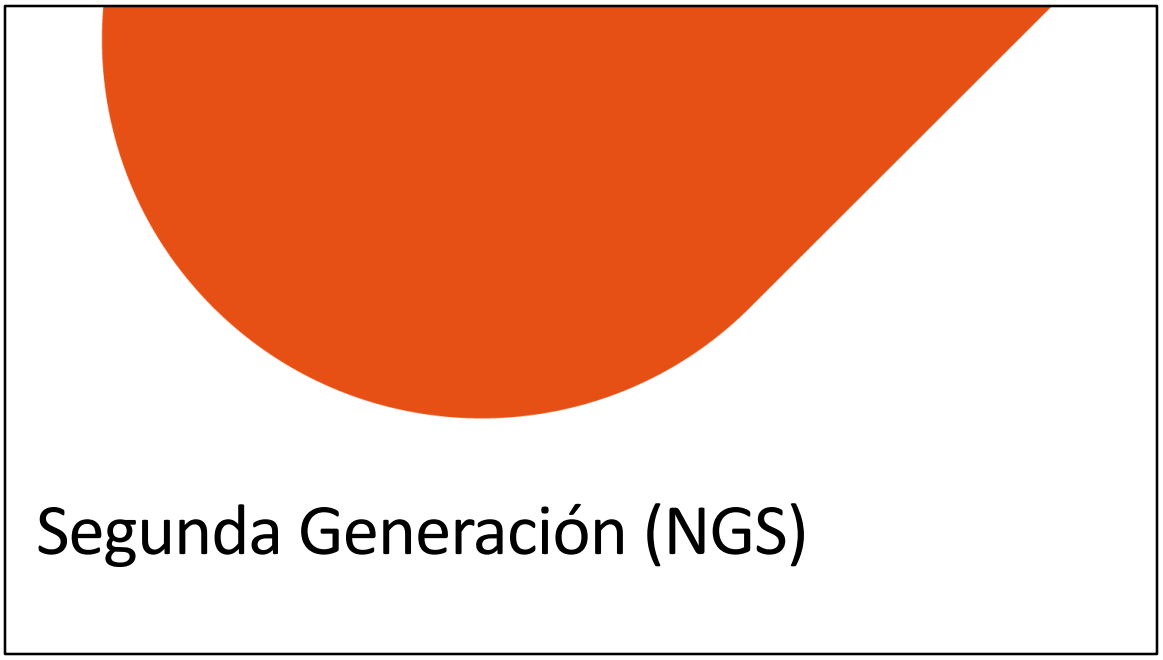
La razón son principalmente los tamaños de lecturas del método, las regiones de alta repetitividad del genoma o las secuencias ricas en GC. Regiones difíciles de resolver con esa tecnología que evitan obtener una secuencia completa.

Por tanto, tras el primer genoma, tenemos dos motivaciones, la principal económica, conseguir reducir el coste para que sea viable, la segunda, rellenar los huecos que nos han quedado (esto no lo conseguiríamos hasta otros 20 años más tarde).

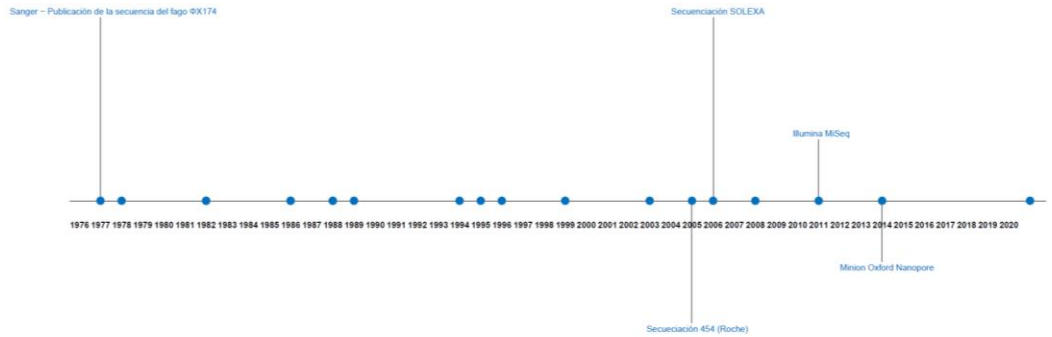
Objetivo de las tecnologías de secuenciación: determina **completamente** la secuencia de un genoma (de forma exacta y barata).

Idealmente lo que buscamos es introducir una muestra genómica y obtener como resultado tantos fragmentos de secuencia como cromosomas existan en esa muestra.

Con estos objetivos en mente, pasamos a la segunda generación de secuenciación, todos los métodos que se desarrollaron para acercarnos a estos objetivos.


Segunda Generación (NGS)

Métodos de secuenciación



15/05/2024

La aparición del equipo de Roche 454 GS System marca el inicio de la secuenciación de nueva generación o NGS.

Inicialmente conocida como Next Generation Sequencing (NGS)

- Disminuye el consumo de recursos respecto a la primera generación.
- Permite secuenciar miles de millones de bases de ADN en un solo experimento (anteriormente decenas de miles de bases), mediante la **secuenciación masiva paralela**
- Procesar múltiples muestras simultáneamente.
- Utiliza la estrategia de secuenciación Shotgun.
- Necesaria la creación de librerías.
- Útil para secuenciar ADN desconocido desde muestras complejas.
- Secuencias más cortas que las obtenidas por método Sanger, pero con mayor número de secuencias o lecturas por nucleótido (***profundidad de secuenciación***).

15/05/2024

Esta bajada en coste y tiempo se debe principalmente al desarrollo de lo que conocemos como secuenciación de segunda generación o next generation sequencing (NGS). La secuenciación de siguiente generación se caracteriza por requerir menos recursos, y por tanto menos costes que la secuenciación Sanger a través del desarrollo de nueva tecnología que principalmente consigue paralelizar los procesos. Utiliza el método Shotgun de secuenciación. Esta segunda generación se caracteriza principalmente por generar lecturas más cortas que las logradas por la secuenciación Sanger, pero con mayor número de secuencias o lecturas por nucleótido (***profundidad de secuenciación***).

A diferencia del anterior método donde un solo fragmento de ADN era secuenciado cada vez, la secuenciación se vuelve masiva y paralela (gran cantidad de fragmentos distintos son secuenciados (leídos) a la vez).

Nuevos pasos en el flujo de trabajo

1. **Obtención de la muestra**
2. **Purificación del ADN (se requiere una cantidad mínima de ADN, 1ug > x > 50ng)**
3. **Fragmentación del ADN**
4. **Creación de una biblioteca/librería**
5. **Secuenciación**
6. **Análisis bioinformático**

15/05/2024

Esta segunda generación incorporan dos nuevos pasos en el flujo de trabajo, la fragmentación del ADN y preparación de librerías. Ambas van ligadas, con el objetivo de preparar una colección de ADN fragmentando que se va a secuenciar. Estos se preparan siguiendo distintos protocolos en un paso previo a la secuenciación, ya sea para facilitar el proceso de secuenciación, marcaje de las secuencias de ADN según la muestra de origen o ambas.

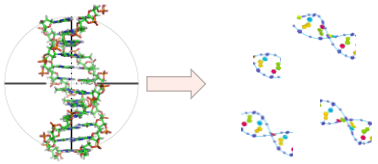
Secuenciación mediante el método *shotgun*

La longitud de los genomas muy larga para los sistemas de secuenciación de primera generación

1- Fragmentación del ADN en partes de tamaño aleatorio
2- Secuenciación
3- Ensamblaje

Tipos

1. Jerárquica (*Hierarchical shotgun sequencing*)
2. De genoma completo (*Whole Genome Shotgun, WGS*)



15/05/2024

Cuando hablamos hoy en día de *shotgun* nos referimos a la secuenciación característica de la segunda generación en la que el ADN es aislado y posteriormente fragmentado siguiendo uno de los métodos mostrados previamente. Esta es la secuenciación *shotgun* del genoma completo (Whole Genome Shotgun, WGS) .

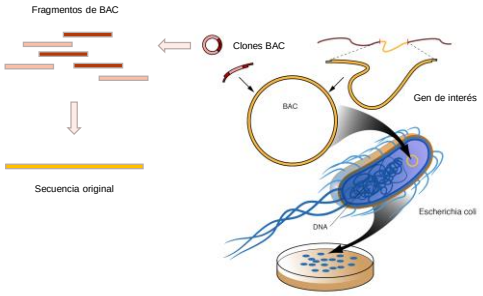
Previamente a esta metodología, la secuenciación *shotgun* era jerárquica (*Hierarchical shotgun sequencing*). Buena parte del proyecto genoma humano se hizo de esta forma. Pero, durante los últimos años de secuenciación del proyecto genoma humano, se propuso un método, que seria computacionalmente mucho más costoso pero más sencillo y menos laborioso (disminuyendo el trabajo humano en el laboratorio) que permitiría hacerlo más rápido y menos costoso. Inicialmente, esta propuesta fue rechazada por Nature y Science que la tildaron de imposible, secuenciación de genoma completo (*Whole Genome Shotgun, WGS*).

La propuesta daría lugar a la segunda generación de secuenciadores.

Jerárquica (*Hierarchical shotgun sequencing*)

- Fragmentos de >100-300kb
- Digestión de los BAC → Fragmentos de distinto tamaño
- Electroforesis en gel → Se ordenan
- Colección aleatoria de secuencias con alta cobertura para cada BAC → ensamblaje
- Cada uno de los contigs generados por cada BAC se superponen y se genera la secuencia original

Pros: Sencillez en el ensamblaje a ser un proceso jerárquico
Contras: Método complejo



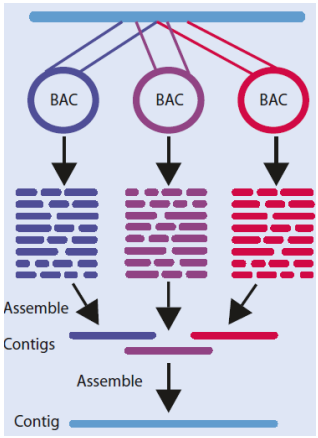
15/05/2024

La secuenciación Shotgun jerárquica (HSS: *Hierarchical shotgun sequencing*) se basa en la clonación de fragmentos de ADN en cromosomas bacterianos artificiales (BAC). Los BAC son vectores de clonación derivados de plásmidos de *Escherichia coli* y tienen la característica de que permiten la inserción de fragmentos de ADN relativamente grandes (>100-300 Kb). Los fragmentos de gran tamaño se clonan dentro de los BAC y se cultivan de forma independiente, originando un gran número de copias de cada uno de los fragmentos. Estos plásmidos se procesan y se realiza una digestión con enzimas específicas para generar fragmentos de distinto tamaños. En un gel de agarosa se corren estos fragmentos de ADN y se ordenan por tamaño. Cada uno de los fragmentos generados se secuencian utilizando alguno de los sistemas de secuenciación que veremos posteriormente, y a partir de la secuencia de estos fragmentos haremos una reconstrucción de la secuencia original en dos pasos.

Jerárquica (*Hierarchical shotgun sequencing*)

- Fragmentos de >100-300kb
- Digestión de los BAC → Fragmentos de distinto tamaño
- Electroforesis en gel → Se ordenan
- Colección aleatoria de secuencias con alta cobertura para cada BAC → ensamblaje
- Cada uno de los contigs generados por cada BAC se superponen y se genera la secuencia original

Pros: Sencillez en el ensamblaje a ser un proceso jerárquico
Contras: Método complejo



15/05/2024

Históricamente, se creía que la secuenciación *shotgun* del genoma completo estaba limitada tanto por el gran tamaño de los genomas grandes como por la complejidad añadida por el alto porcentaje de ADN repetitivo (más del 50% en el caso del genoma humano) presente en los genomas grandes. Por estos motivos, antes de realizar la secuenciación *shotgun*, se tuvieron que utilizar otras estrategias que redujeran la carga computacional del ensamblaje de la secuencia. En la secuenciación jerárquica, también conocida como secuenciación descendente (*top-down sequencing*). Se realiza un mapa genético de distancias de baja

resolución del genoma antes de la secuenciación propiamente dicha (https://www.genome.gov/genetics-glossary/Genetic-Map). A partir de este mapa, se selecciona un número mínimo de (grandes) fragmentos que cubren todo el cromosoma para su secuenciación. De este modo, se requiere la mínima cantidad de secuenciación y ensamblaje de alto rendimiento.

El genoma amplificado se corta primero en trozos más grandes (50-200kb) y se clona en un huésped bacteriano utilizando BACs. Aunque no se conocen las secuencias completas de los **contigs** de BAC, sí se conocen sus orientaciones relativas (gracias al mapa genético físico que previamente se ha creado). Existen varios métodos para deducir este orden y seleccionar los BAC que conforman una ruta que nos dará con el mínimo numero de BACs para secuenciar, el genoma completo. Una vez que se ha encontrado el solapamiento entre los clones y se conoce su orden en relación con el genoma, se realiza una secuenciación shotgun de un subconjunto mínimo de estos contigs que cubre todo el genoma.

Debido a que primero se crea un mapa de baja resolución del genoma, esta secuenciación es más lenta pero depende menos de los algoritmos informáticos que la “secuenciación por escopeta” del genoma completo. Sin embargo, el proceso de creación de bibliotecas BAC y de selección de la rutas de BACs que nos dará el genoma

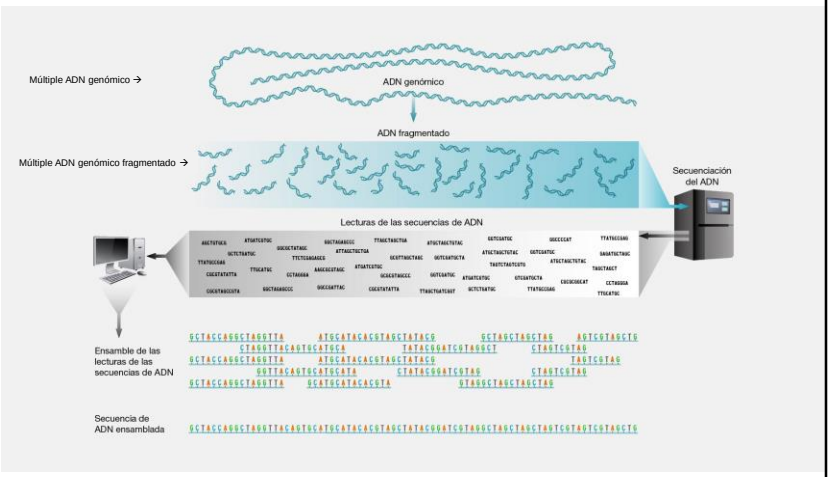
completo, hace que la secuenciación jerárquica sea lenta y laboriosa.

Contig: Un contig es un conjunto de secuencias de ADN contiguas que se han ensamblado a partir de fragmentos de ADN superpuestos. Los contigs se generan típicamente mediante algoritmos computacionales que alinean y unen los fragmentos en función de la superposición entre ellos. Una vez ensamblados, los contigs se pueden utilizar para estudiar la secuencia y organización del ADN y para identificar genes y otros elementos funcionales dentro de la secuencia.

Whole Genome Shotgun, WGS

- Inicialmente para genomas bacterianos (pequeños).
- Craig Venter lo validó para el ensamblaje del genoma humano (2001).
- Sentó las bases de las técnicas de secuenciación seguidas a día de hoy.
- Genómica y metagenómica.

Pros: Metodología sencilla
Contras: El ensamblaje requiere de un nivel de computación exhaustivo para la época



https://www.genome.gov/es/genetics-glossary/Secuenciacion-shotgun

Por otro lado tenemos la secuenciación *Shotgun* de genoma completo, que es a día de hoy la más popular y utilizada y la que ha dado pie al abaratamiento y aumento de eficiencia de las técnicas de secuenciación. En esta secuenciación se parte de la muestra de ADN fragmentando pero no se hace un paso previo de clonación en BAC antes de la secuenciación, si no que se secuencian directamente tras preparar las librerías. Es el método más útil para estudios genómicos y metagenómicos gracias a los avances tecnológicos. Esto se debe a que a pesar de ser una metodología más sencilla que la secuenciación jerárquica, requiere de una mayor capacidad computacional para ensamblar las lecturas en contigs, al no realizarse en dos pasos.

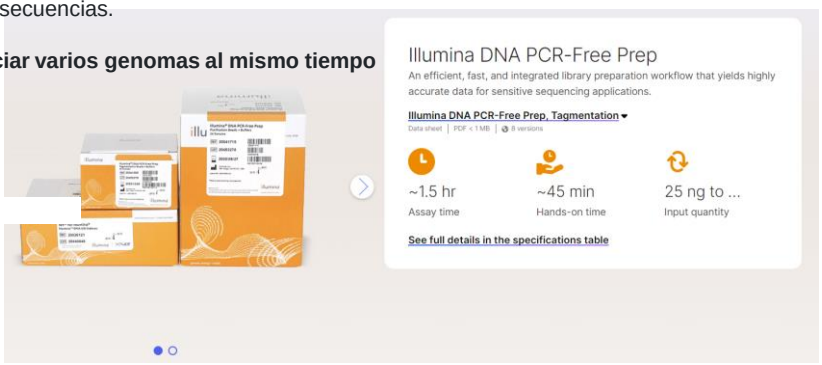
Fragmentación del ADN y preparación de la librería

1. Fragmentación del ADN:
    1. Sonicación
    2. Actividad enzimática
    3. Química

2. Ligación de un adaptador o barcode a la secuencia.

3. Selección por tamaños de las secuencias.

Las librerías permiten secuenciar varios genomas al mismo tiempo



La preparación de las librerías se inicia mediante la fragmentación del ADN a secuenciar en primer lugar según distintos métodos, generando fragmentos de tamaño aleatorio. A estos se les liga un adaptador a los extremos 5’ y 3’ de la secuencia. Estos extremos se añaden por PCR usando primers universales y generan extremos de secuencia conocidos que permitirán la amplificación de los fragmentos en cuestión. Estos fragmentos se seleccionan por tamaño y se secuencian. Destacar que, gracias al uso de la adición de adaptadores a las secuencias amplificadas, es posible secuenciar más de una muestra de forma simultánea.

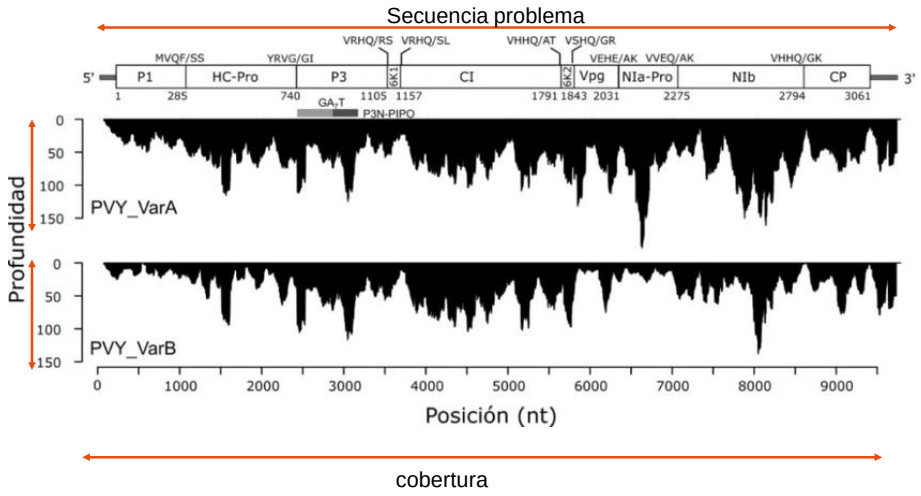
Efectos de la fragmentación del ADN en la NGS

Aunque cada método de fragmentación tiene sus ventajas y desventajas, buscaremos un tamaño en concreto y querremos todos los fragmentos del mismo homogéneo tamaño (no siempre es posible).

Un pico de fragmentos de ADN del tamaño adecuado para la preparación de la librería es ideal para la biblioteca NGS si se planea realizar la secuenciación del genoma completo (WGS) (Panteleeva et. al., 2016).

Además, un estudio que comparaba los métodos físicos y enzimáticos mostró que el rendimiento general era igual para ambos métodos con diferencias menores en la secuenciación NGS. Los autores sugieren que los métodos de fragmentación pueden elegirse únicamente en función de las instalaciones del laboratorio, la viabilidad y el diseño experimental (Kechin et. al., 2021).

Cobertura de la secuenciación vs profundidad de lecturas en la secuenciación



15/05/2024

La cobertura de secuenciación es una medida de la cantidad de nucleótidos que poseen alguna lectura en un experimento de secuenciación, es decir, que han sido analizados. La cobertura se calcula dividiendo el número total de nucleótidos secuenciados (con lecturas) por el número total de nucleótidos en la secuencia completa que se está analizando. Por lo general, se considera que una buena cobertura de secuenciación es aquella que supera el 90%, lo que indica que la mayor parte de la secuencia ha sido analizada.

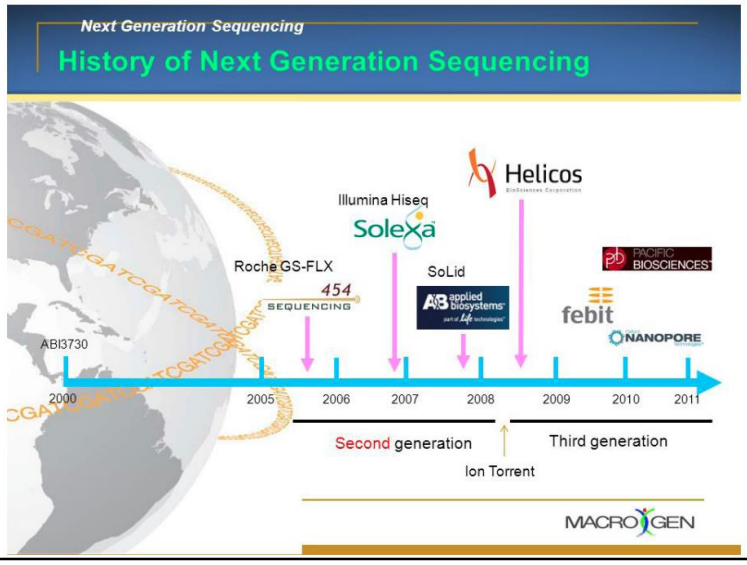
En cambio, la profundidad de lectura, se refiere a la cantidad de lecturas que existen para una base determinada de la secuencia problema. El número de veces que esa base ha sido leída. Esa es la profundidad de lectura para esa base. La profundidad de lectura media para un genoma que ha sido analizado, será la media de la profundidad de lectura por cada base de ese genoma.
También podemos referirnos a profundidad de lectura, hablando de un secuenciador, no de un genoma problema o una secuencia problema, como la cantidad de lecturas que es capaz de proporcionar por carrera o run de secuenciación.

Estos dos parámetros son muy importantes en bioinformática, ya que permite evaluar la calidad y fiabilidad de los datos obtenidos en un experimento de

secuenciación.

En la figura superior tenemos la misma secuencia secuenciada dos veces (A y B). En el eje de las Y, el histograma negro invertido hace referencia a la cantidad de lecturas que hay por posición nucleotídica de la secuencia, es por tanto una representación ***real*** de la profundidad de lectura. El eje de las X, en cambio, representa de la lectura superior cuantos nucleótidos están leídos, es la cobertura de la secuencia.

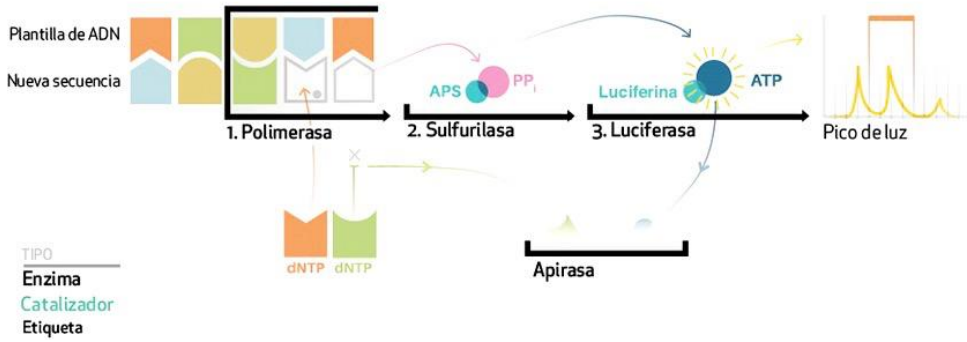
Secuenciadores de segunda generación



15/05/2024

Método de Pirosecuenciación

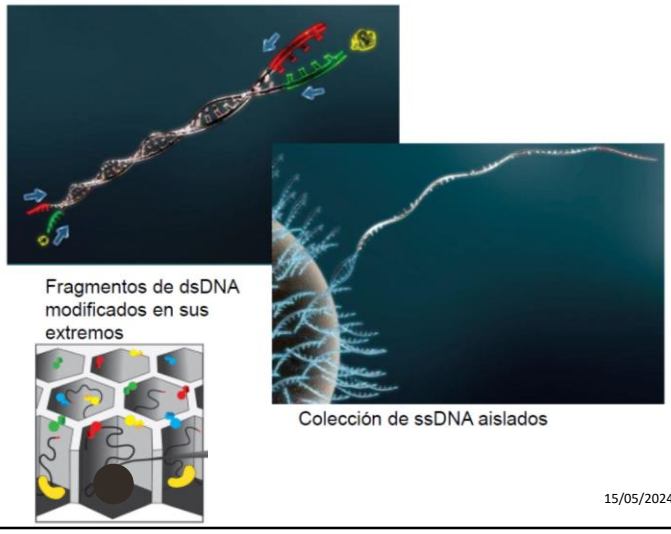
- Publicada en 1988 (Hyman, 1988)
- Más rápida y barata pero menos precisa que el método Sanger
- No requiere gel de electroforesis ni marcadores fluorescentes o ddNTP



15/05/2024

Pirosecuenciación 454 (Roche) - 2005

- La metodología deriva de la pirosecuenciación.
- Los fragmentos se generan mediante nebulización, generando fragmentos entre 400-800pb
- Se crean librerías añadiendo adaptadores biotinilados a los externos del ADN
- Las librerías (ADN molde) se preparan sobre microesferas con avidina que van de forma independiente sobre pocillos de una placa de análisis

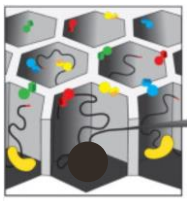


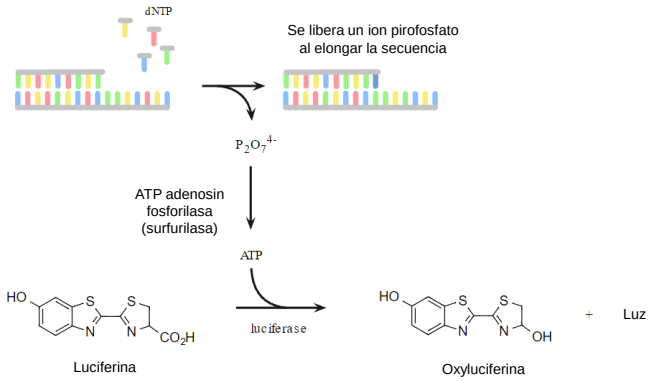
15/05/2024

La primera tecnología de segunda generación fue desarrollada por Roche, denominada pirosecuenciación 454. Las características de esta secuenciación es que la fragmentación se da a través un paso de nebulización, generando fragmentos de entre 400 y 800pb, en un paso previo a la preparación de librerías mediante la adición de adaptadores biotinilados. Estos adaptadores son necesarios para el proceso de secuenciación, ya que van a permitir a las librerías adherirse a unas microesferas con avidina donde se va a producir todo el proceso.

454 (Roche)

- A las microplacas con microesferas se les añade sulfurilasa, luciferasa y luciferina.
- Se sintetiza un nucleótido y se produce luz por cada nucleótido unido.
- La secuencia se va leyendo poco a poco. No es necesario correr el resultado en un gel.





15/05/2024

La pirosecuenciación se basa en la reacción química producida por la luciferina, la cual produce luz. El proceso es el siguiente: a las placas con microesferas se les añade sulfurilasa y luciferasa además de los nucleótidos y la polimerasa. Al unirse un nucleótido se libera un ion pirofosfato que aprovechará la surfurilasa para generar un ATP y producir la luciferasa luz a partir de la luciferina. De esta manera tenemos una señal lumínica que nos evita tener que correr los resultados en un gel, **siendo una secuenciación en tiempo real**. Para diferenciar qué nucleótido es el que se acaba de unir, el proceso es secuencial y se va añadiendo un nucleótido cada vez. Si se produce la señal se efectúa un lavado y se añade otro nucleótido, proceso que se seguirá hasta completar la secuencia. Si tenemos el caso donde se adicionen más de un mismo nucleótido, la intensidad de la luz aumenta, siendo detectado por el sistema. Este último punto es también su debilidad al no ser capaz de diferenciar siempre correctamente la señal cuando más de un nucleótido igual se incorpora.

454 (Roche)

GS FLX Titanium XL+

Tiempo: 24h

Longitud reads: Hasta 1000pb (media de 700pb)

Capacidad de secuenciación por carrera: 1 millón de secuencias

Salida media: 700Mpb

Fiabilidad: 99,75%

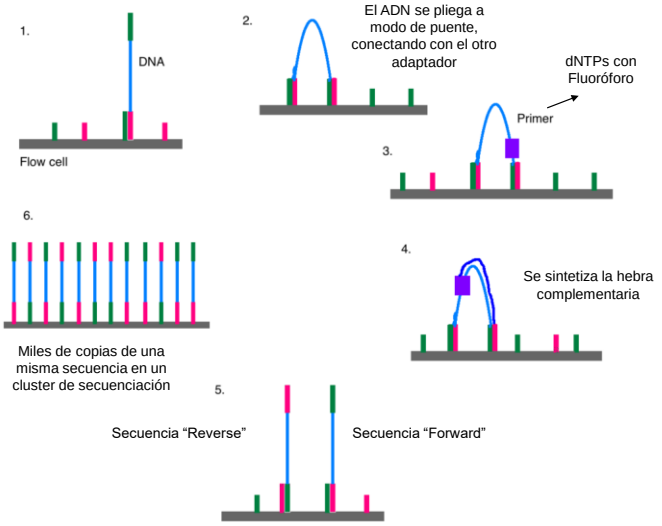
Características:
- Capacidad de paralelizar el trabajo más eficientemente que Sanger
- Más barato que Sanger
- Problemas al medir homopolímeros

15/05/2024

La pirosecuenciación 454 de Roche se caracteriza por presentar lecturas de un tamaño similar a las obtenidas mediante Sanger, pero con una profundidad mucho mayor, de hasta 1 millón de secuencias, dando una salida total de 700 millones de pares de bases nitrogenadas. La capacidad de paralelizar los procesos y el abaratamiento de los costes son las principales ventajas respecto a la primera generación de secuenciación. Como punto negativo, el proceso es más lento que Sanger principalmente por la metodología secuencial que sigue, y su fiabilidad es más baja que la de Sanger, pero manteniéndose en unos niveles del 99.75%. Por último, destacan los problemas a la hora de medir homopolímeros, y es que a pesar de que la adición de más de una base nitrogenadas simultáneamente sea cuantificable, no es un criterio preciso dando lugar a errores de secuenciación.

Illumina (Solexa) - 2007

- Tecnología de terminador reversible, secuenciación por síntesis o “Bridge-PCR”.
- Preparación de la librería – Adaptadores P5 y P7.
- Es la plataforma de secuenciación más usada – DNAseq, transcriptómica y metagenómica.
- Opción de “Pair-end reads” – Secuencias complementarias sintetizadas a la vez.



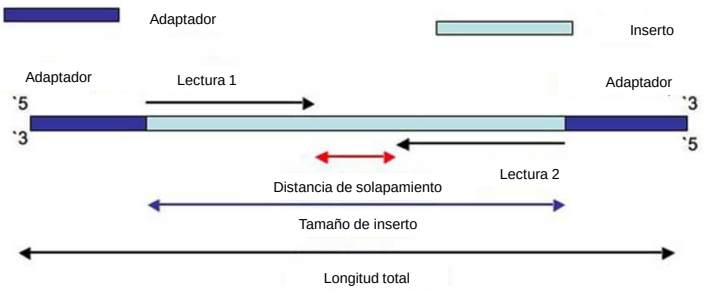
15/05/2024

Uno años más tarde a la aparición de la secuenciación de Roche dio a luz la secuenciación de Illumina (Solexa). Es un sistema de secuenciación por síntesis denominada ”de terminador reversible” o ”Bridge PCR”, que mejora en muchos aspecto a la tecnología de pirosecuenciación. El primer paso, como en el resto de tecnologías de segunda generación, es la preparación de las librerías a partir de nuestro ADN fragmentados, lo que comúnmente se lleva a cabo utilizando los adaptadores P5 y P7 (por su secuencia). Estos adaptadores se unen a secuencias complementarias existentes en una ”flow cell”, que es donde se dará la reacción. Luego, el ADN que posee el otro adaptador en su otro extremo se dobla formando un puente y se adhiere a otro adaptador complementario. En este momento comienza la síntesis de la hembra complementaria y se van uniendo nucleótidos a los que se les ha adherido un fluoróforo, lo que permite determinar qué nucleótido se ha unido secuencialmente. Esto se da hasta formar la hebra complementaria al completo. Una vez sintetizada, se separan las hebras, dando lugar a dos secuencias complementarias, denominadas ”forward” y ”reverse”. El sistema entonces libera una de las dos hembras y se repite el proceso. De esta forma se comienza la síntesis de secuencias ”pair-ends”.

Illumina (Solexa) - 2007

"Pair-end reads"

Se genera una secuencia y su secuencia complementaria, dejando entre ambas lo que se conoce como inserto.

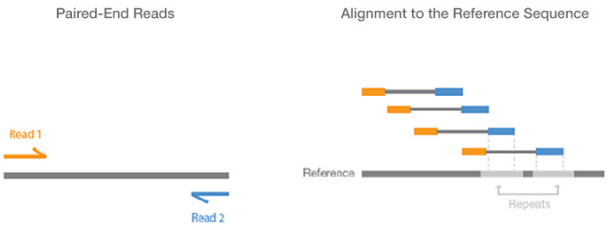


15/05/2024

Las secuenciación "pair-ends" se lleva a cabo mediante la síntesis y amplificación de las dos hebras complementarias de un fragmentos de ADN. Esta tecnología presenta una gran ventaja respecto a la secuenciación de una cadena simple, ya que al tener la cadena complementaria podemos trabajar con el concepto "longitud de inserto" para ayudarnos durante el posterior paso de ensamblaje. El tamaño de inserto se define como la cantidad de nucleótidos (en valor absoluto) total del fragmento amplificado. Por ejemplo, si hemos amplificado un fragmento de 1.000pb y la secuenciación solo da lugar a fragmentos de 300pb, el tamaño de inserto será de 1.000pb pero dejando una distancia de solapamiento de 400pb.

Illumina (Solexa) - 2007

Pair-end files
Es común obtener dos lecturas de una sola molécula. Ejemplos de estas técnicas son las lecturas parejas de Illumina (Pair-end reads). En estos casos para cada lectura hay otra lectura pareada. Una forma común de almacenar esas lecturas emparejadas es crear archivos fastq, uno para la primera lectura de los pares y otro para la segunda. En este caso, los archivos deben contener las lecturas exactamente en el mismo orden.



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Illumina System	Supported insert size
NextSeq	550 bp (with NextSeq v2 reagents) 350 bp (with NextSeq v1 reagents)
HiSeq X	350 bp (with HiSeq X Reagent Kit v2) 450 bp (for TruSeq Nano DNA and TruSeq DNA PCR-Free libraries)

15/05/2024

La elección del tamaño de inserción adecuado es una parte importante de la planificación del experimento de secuenciación. La selección del tamaño suele realizarse tras la fragmentación del ADN de entrada y la adición del adaptador, mediante electroforesis en gel o perlas. La elección de los posibles tamaños de inserto está limitada por factores técnicos del proceso de secuenciación, en Illumina, especialmente por el paso de amplificación en puente.

Illumina (Solexa) - 2007

MiSeq

Tiempo: 4-55h

Longitud reads: 2 x 300pb

Capacidad de secuenciación: 25 millones de secuencias

Salida media: 15Gpb

Fiabilidad: 99,25%

Características:
- Mejora en todos los aspectos a 454
- Mismo precio que 454 a mayor cantidad de información

15/05/2024

Debido a su gran popularidad, Illumina gestiona una amplia colección de secuenciadores que utilizan su tecnología, dividiéndose en baja y gran escala. Los de baja escala se caracterizan por su tamaño reducido y precio económico, con la posibilidad de existir dentro de un laboratorio de tamaño medio. Dependiendo de la profundidad de secuenciación cada carrera de secuenciación, puede oscilar entre las 4 y 55 horas, proceso en el que se dan lecturas de hasta 2 x 300 pb con una profundidad total de 25 millones de lecturas. Por otro lado, su fiabilidad sigue siendo bastante alta, similar a la de la secuenciación 454. Al comparar ambas estrategias lo que resulta más llamativo es la mayor profundidad de secuenciación lograda, algo significativo al tener en cuenta que los precios de ambas técnicas es similar. Por contra, las lecturas son más cortas, pero se soluciona parcialmente al tener la información en lecturas ”pair-ends”.

Illumina (Solexa) - 2007

Tiempo: 13-44 horas

Longitud reads: 2x150pb

Capacidad:
5 billones de secuencias

Salida media: 750Gpb

HiSeq

Tiempo: 13-44 horas

Longitud reads: 2x250pb

Capacidad: 20 billones de secuencias

Salida media: 6000Gpb

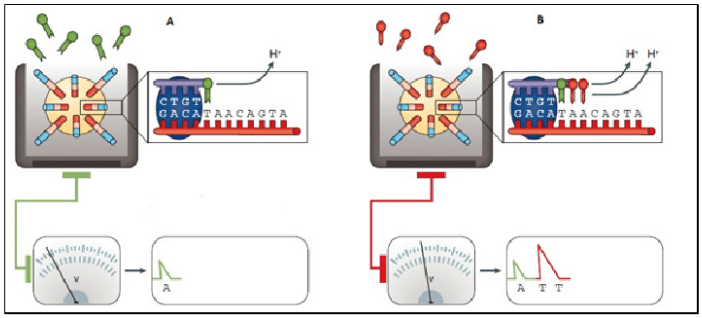
NovaSeq

15/05/2024

Por otro lado tenemos los secuenciadores a gran escala, donde destacan HiSeq y su evolución actual denominado NovaSeq. De estos secuenciadores nos debemos quedar con los datos de longitud de lectura, más cortas que su hermano menor MiSeq, pero con una profundidad de secuenciación que los convierte actualmente en los sistemas que más información generan por carrera de secuenciación. En este caso, NovaSeq al utilizar las últimas tecnologías es capaz de dar lugar a un total de 20 billones de secuencias por carrera.

Ion Torrent (Life Technologies) - 2010

- Similar a la secuenciación con 454 (síntesis), pero aquí se mide la concentración de hidrógeno en el medio.
- Medición mediante un ISFET (Ion Sensitive Fiel Effect Transistor)
- Librerías construidas sobre microesferas



15/05/2024

Otro de los sistemas que se han desarrollado dentro de la denominada secuenciación de segunda generación es la tecnología ”Ion Torrent” de Life Technologies. Este sistema es único ya que es una evolución de la tecnología 454 de Roche con la diferencia de que al sintetizarse un nucleótido no se genera luz, si no que se consigue variar la concentración de hidrógeno del medio. Al liberarse iones hidrógeno cambia el pH del medio, pudiéndose medir la cantidad de los mismos que se han sintetizado. Al sistema se añaden los nucleótidos uno a uno como en el paso de pirosecuenciación, por lo que sabemos qué nucleótidos son los que se están uniendo.

Ion Torrent (Life Technologies) - 2010

Ion Proton

Tiempo: 2-4h

Longitud reads: 200pb

Capacidad: 330 millones de secuencias

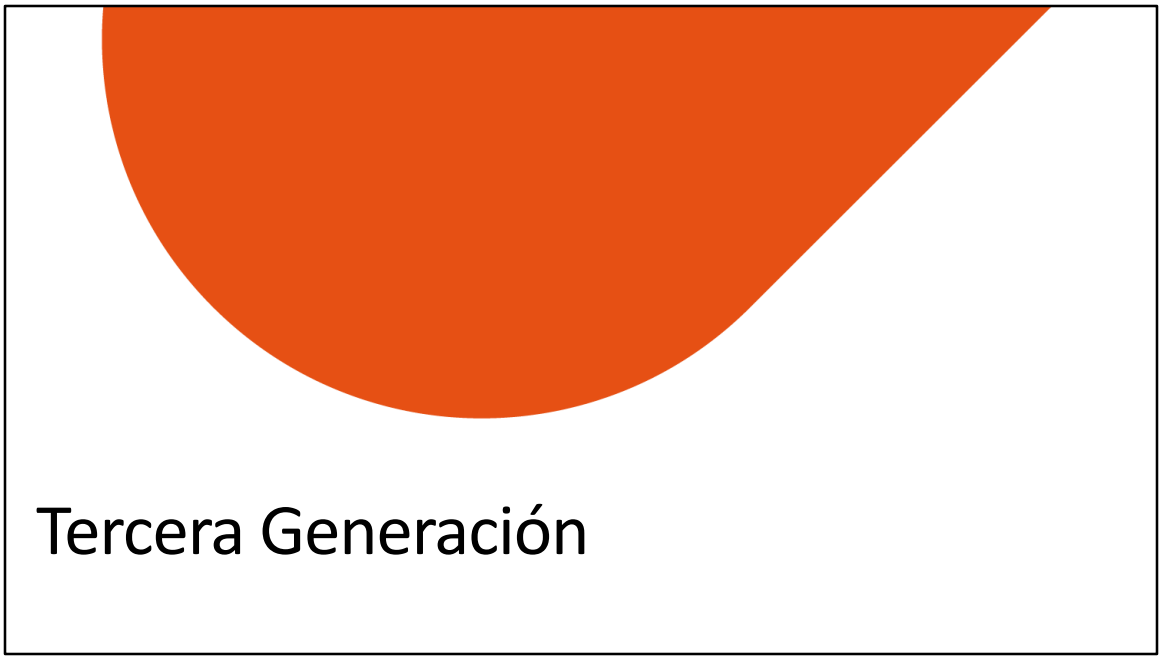
Salida media: 66.6Gpb

Fiabilidad: 98%

Características:
- Es el método más veloz
- Su alto porcentaje de errores

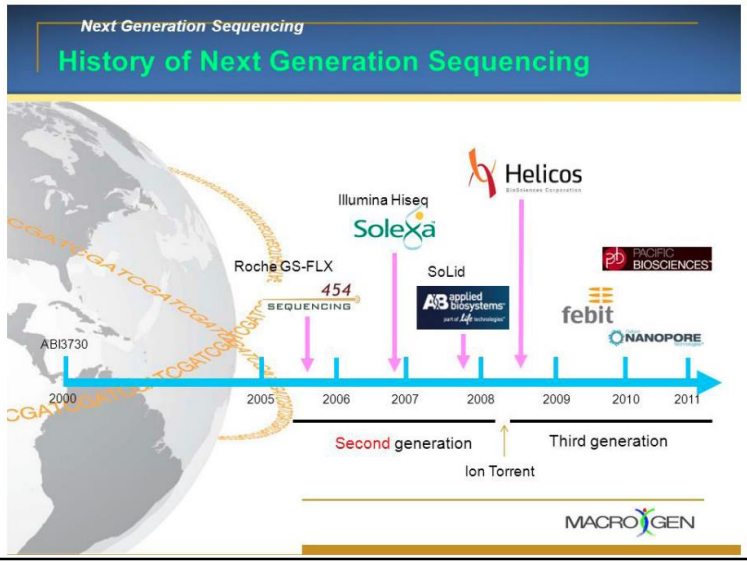
15/05/2024

A pesar de ser una evolución de la tecnología 454 de Roche, Ion Torrent no ha triunfado por quedar muy por detrás de la tecnología Illumina. En este caso tenemos lecturas de unos 200pb de longitud aunque a una gran profundidad, pero la fiabilidad en este proceso cae por problemas existentes en el sistema de detección de homopolímeros.



Tercera Generación

Secuenciadores de tercera generación



15/05/2024

La segunda generación de secuenciadores acercó la secuenciación a un gran número de laboratorios y proyectos consiguiendo su primer objetivo, el abaratamiento de los costes. La disminución en la mano de obra humana por la implementación del *Whole Genome Sequencing Shotgun* contribuyó a este abaratamiento. La tecnología que mejor funcionó fue la que proporcionaba mucha cantidad de secuencias al menor coste, Illumina. Su principal problema, la longitud de las secuencias, no permitía resolver las regiones difíciles del genoma y requería de una gran potencia de calculo debido al tamaño de las “piezas del puzle”.

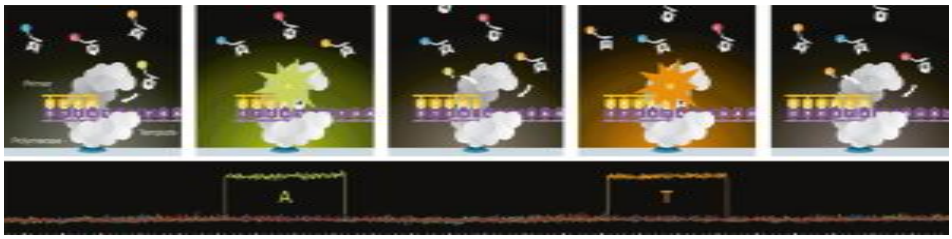
La tercera generación de secuenciadores (NGS3) representa un avance significativo en la tecnología de secuenciación de ADN, introduciendo nuevas capacidades y superando algunas de las limitaciones de las generaciones anteriores.

Características principales:
Lectura larga de secuencias:
Secuenciación de ADN nativo: La NGS3 no requiere amplificación previa del ADN, lo que minimiza el riesgo de introducir errores y sesgos durante el proceso de secuenciación.
Detección de modificaciones epigenéticas: Algunas plataformas permiten la

detección de modificaciones epigenéticas, como la metilación del ADN, proporcionando información adicional sobre la regulación de la expresión génica. Mayor velocidad y rendimiento: algunas plataformas que ofrecen velocidades y rendimientos cada vez mayores, permitiendo secuenciar genomas completos en pocas horas o incluso minutos.

PacBio - Pacific Biosciences 2011

- Revolución tecnológica, lecturas de más de 1000pb.
- SMRT: Simple Molecule Real Time sequencing
- La secuenciación se lleva a cabo en una estructura nanofotónica rodeada de aluminio (ZMW - zero-mode waveguide). Dentro de esta estructura se da la síntesis con dNTPs marcados con fluoróforos.
- Se añade una molécula de ADN polimerasa de phi29 a la superficie.
- Se requiere un ADN de alta calidad (libre de contaminantes y alto grado de integridad)



15/05/2024

Hasta este momento los sistemas de secuenciación sólo han podido generar secuencias cortas las cuales deben ser ensambladas posteriormente para su análisis, lo que supone en muchas ocasiones un gran reto a nivel computacional y metodológico. La tercera generación de secuenciación se caracteriza por intentar paliar este problema, con el coste de dar lugar a una menor profundidad de secuencias. Bajo esta premisa se implantó la tecnología SMRT de PacBio. Mediante esta metodología podemos obtener secuencias de más de 1.000pb en un proceso de secuenciación que se da en una estructura nanofotónica llamada ZMW. Aquí se da una síntesis de la cadena con nucleótidos con fluoróforos, similar a los de los métodos de segunda generación.

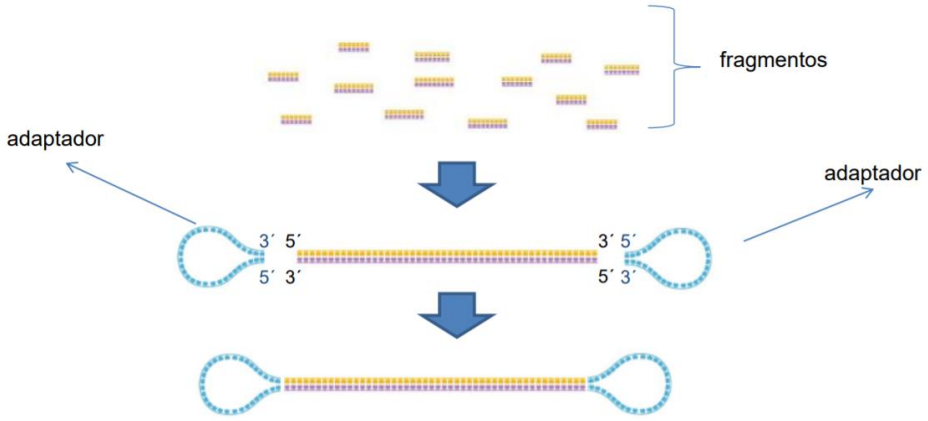
PacBio - Pacific Biosciences 2011

ZMW

15/05/2024

Esta es la ”flow cell” utilizada para la secuenciación con PacBio, donde encontramos las regiones ZMW.

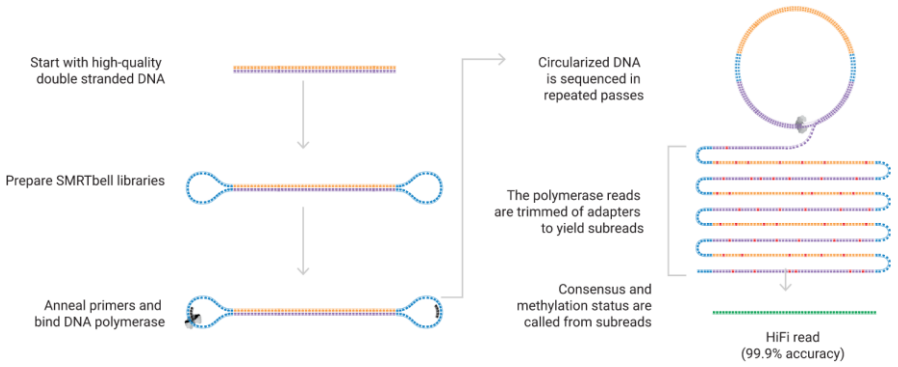
PacBio - Pacific Biosciences 2011



15/05/2024

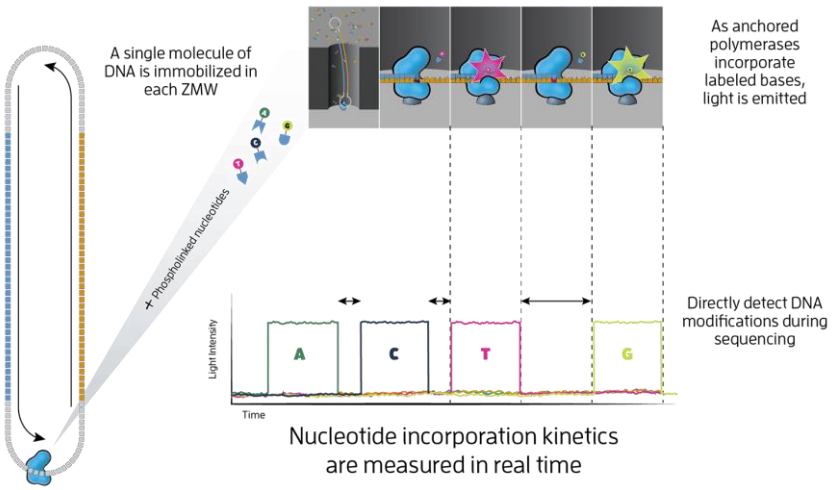
La principal diferencia en este nuevo método es que se originan estructuras circulares en un sistema llamado “hairpin”. Tras fragmentar el ADN, los adaptadores, con forma de horquilla, se adhieren a moléculas de ADN de doble cadena. Esto permite a la polimerasa del fago phi29 (de alta procesividad) generar copias de ambas cadenas y formar fragmentos de ADN de gran longitud. Este proceso emula a la replicación por “circulo rodante” que muchos virus de ADN circular llevan a cabo.

PacBio - Pacific Biosciences 2011



15/05/2024

PacBio - Pacific Biosciences 2011



15/05/2024

PacBio - Pacific Biosciences 2011

PacBio RS II

Tiempo: 4h

Longitud reads: 15-50Kpb

Capacidad: 1 millón de secuencias

Salida media: 2.8Gpb

Fiabilidad: 85% en un solo pase, se requieren múltiples pases

Características:
- Secuencias de gran longitud
- La tasa de error es muy alta

15/05/2024

Las principales características de este sistema de secuenciación es su rapidez y principalmente la longitud de las lecturas, de entre **15 y 50 kpb**. Esto permite, por ejemplo, la secuenciación de algunos virus al completo. El principal problema con los sistema de tercera generación es su baja fiabilidad, siendo en este caso 85% de media, aunque los últimos avances están mejorándola por encima del 90%.

PacBio - Pacific Biosciences

REVIO SYSTEM
Long-read sequencing

Industry-leading accuracy with HiFi sequencing for complete views of genomes, epigenomes, and transcriptomes

Throughput to run up to 1,300 human HiFi genomes per year

Smart consumables for simple handling and less plastic waste

Powerful compute with Google Health DeepConsensus onboard

ONSO SYSTEM
Short-read sequencing

Greater level of sensitivity to detect rare variants

Reduced requirement for sequencing coverage depth versus SBS sequencers

Contiguous reads through homopolymer and difficult to sequence regions

Low duplication rate, no index hopping

Rapid conversion of existing P5/P7 libraries for sequencing on the Onso system

SEQUEL IIe SYSTEM
Long-read sequencing

Long, accurate HiFi reads with DNA methylation direct from the instrument in every run

Supports a wide range of applications – targeted sequencing, RNA sequencing, and whole-genome sequencing

Throughput match for microbial genome sequencing, AAV vector sequencing, Iso-Seq, and more

Featured in hundreds of peer-reviewed publications

05/2024

PacBio - Pacific Biosciences

15/05/2024

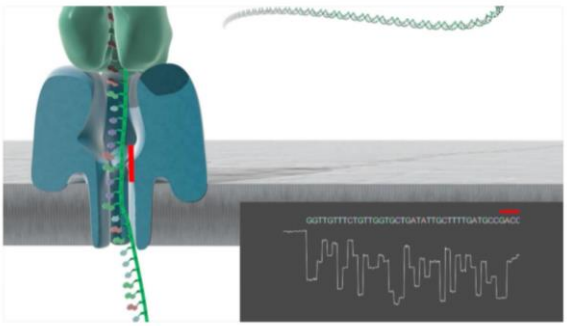
Oxford Nanopore - 2015

15/05/2024

Aquí vemos cómo se secuencia la cadena de ADN en su transcurso a través de los poros.

Oxford Nanopore - 2015

- Secuenciación de bases al pasar a través de nanoporos con una corriente iónica.
- Basado en el mismo sistema de *hairpin*.
- Los nanoporos son dispuestos en membranas conductoras de la electricidad, de modo tal de que cuando son atravesados por moléculas de ssDNA, en función de sus nucleótidos, ocurren cambios de potencial que pueden ser detectados.
- Debido a la alta velocidad del proceso se guardan fragmentos consecutivos de, generalmente, 5 nucleótidos.
- Alta tasa de error, mejorando con el tiempo.



15/05/2024

El último sistema de secuenciación que a día de hoy está revolucionando el mundo del DNASeq y RNASeq está desarrollado por Oxford Nanopore. Este sistema está basado en el mismo sistema de ”hairpin” que PacBio con la diferencia de que aquí la cadena de ADN cruza una serie de poros arrastrada por una proteína motora. En este poro existe una corriente de iones que se desestabilizan con el paso de las bases nitrogenadas, generando un perfil diferente para cada una de las bases. Debido a la velocidad del proceso no se leen las bases de una en una, si no generalmente de cinco en cinco, generando perfiles únicos en el detector. Principalmente debido a esta velocidad, este sistema está ligado a una baja fiabilidad de secuenciación, la cual ha mejorado con el tiempo. Durante el año 2015 la fiabilidad era menor del 60%.

Oxford Nanopore - 2015

MinION

Contiene 512 canales, cuatro poros cada uno.

El primer dispositivo que se comercializó, por un precio de unos 1.000 dólares, fue el MinION. Este dispositivo se caracteriza por tener el tamaño de un smartphone y requerir solo una conexión USB a un ordenador para funcionar.

Oxford Nanopore - 2015

MinION

Tiempo: 16h

Longitud reads: 15-50Kpb – 4 Mpb

Capacidad: 1 millón de secuencias

Salida media: 50Gpb

Fiabilidad: Porcentaje de error alto, mejorando con nuevas versiones

Características:
- Secuencias de gran longitud
- La tasa de error es muy alta

GridION

Capacidad: 5 millón de secuencias

Salida media: 250Gpb

15/05/2024

El MinION se caracteriza por presentar unas especificaciones similar a las de PacBio, con la principal diferencia de su tamaño. Con el paso de los años los dispositivos de secuenciación han evolucionado, y a día de hoy se comercializa el GridION, cinco secuenciadores MinION colocados en serie para aumentar la profundidad de secuenciación.

Oxford Nanopore - 2015

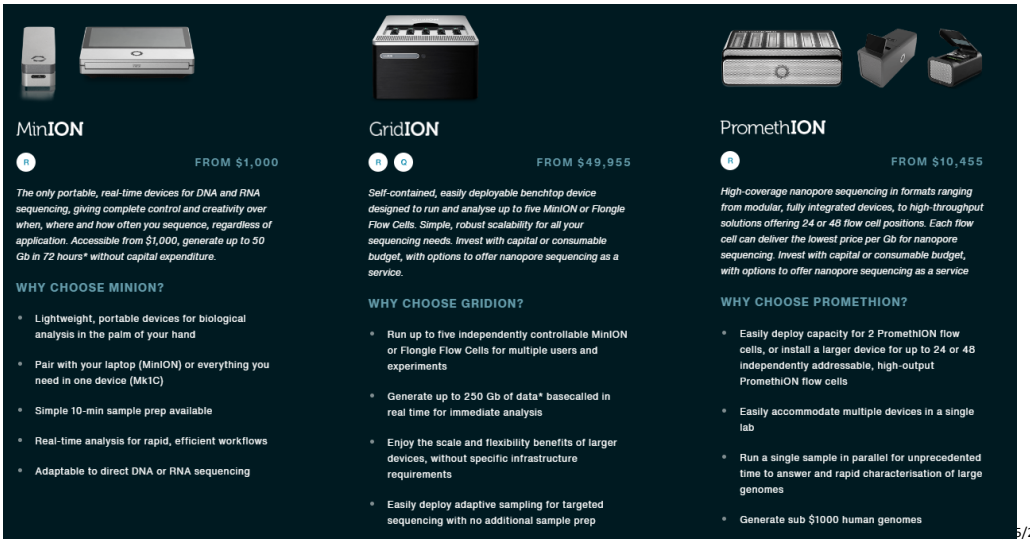
We have developed a new generation of sensing technology that uses nanopores - nano-scale holes - embedded in high-tech electronics, to perform comprehensive molecular analyses.

Our first products sequence DNA and RNA. We offer the only sequencing technology to combine scalability from portable to ultra-high throughput formats with real-time data delivery and the ability to elucidate **accurate**, rich biological data through the analysis of short to ultra-long fragments of native DNA or RNA. The sensing platform has the potential to be adapted for the analysis of other types of molecules, for example proteins.

View products >

15/05/2024

Oxford Nanopore - 2015



MinION

R — FROM $1,000

The only portable, real-time devices for DNA and RNA sequencing, giving complete control and creativity over when, where and how often you sequence, regardless of application. Accessible from $1,000, generate up to 50 Gb in 72 hours without capital expenditure.*

WHY CHOOSE MINION?

- Lightweight, portable devices for biological analysis in the palm of your hand
- Pair with your laptop (MinION) or everything you need in one device (Mk1C)
- Simple 10-min sample prep available
- Real-time analysis for rapid, efficient workflows
- Adaptable to direct DNA or RNA sequencing

GridION

R O — FROM $49,955

Self-contained, easily deployable benchtop device designed to run and analyse up to five MinION or Flongle Flow Cells. Simple, robust scalability for all your sequencing needs. Invest with capital or consumable budget, with options to offer nanopore sequencing as a service.

WHY CHOOSE GRIDION?

- Run up to five independently controllable MinION or Flongle Flow Cells for multiple users and experiments
- Generate up to 250 Gb of data* basecalled in real time for immediate analysis
- Enjoy the scale and flexibility benefits of larger devices, without specific infrastructure requirements
- Easily deploy adaptive sampling for targeted sequencing with no additional sample prep

PromethION

R — FROM $10,455

High-coverage nanopore sequencing in formats ranging from modular, fully integrated devices, to high-throughput solutions offering 24 or 48 flow cell positions. Each flow cell can deliver the lowest price per Gb for nanopore sequencing. Invest with capital or consumable budget, with options to offer nanopore sequencing as a service

WHY CHOOSE PROMETHION?

- Easily deploy capacity for 2 PromethION flow cells, or install a larger device for up to 24 or 48 independently addressable, high-output PromethION flow cells
- Easily accommodate multiple devices in a single lab
- Run a single sample in parallel for unprecedented time to answer and rapid characterisation of large genomes
- Generate sub $1000 human genomes

Oxford Nanopore - 2015

In development

Learn about nanopore technology >

MinION Mk1D

- Combines established MinION technology with leading tablet manufacturers
- Nanopore accessory with sequencer and other peripherals
- New nanopore software will run and analyse sequencing data on-tablet, whilst utilising e.g. Bluetooth, Wi-Fi and 5G

SmidgION

- Designed to be our smallest sequencing device so far
- Same nanopore sensing technology as MinION and PromethION
- Designed for use with a smartphone in any location

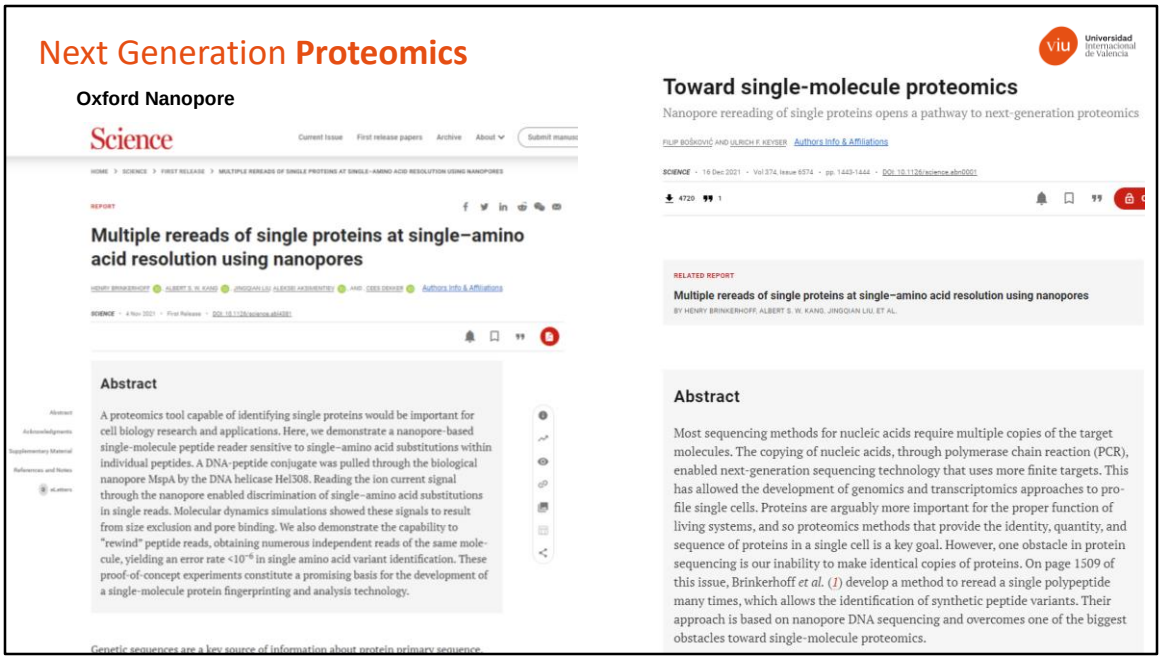
Learn more >

Plongle

- High-throughput, real-time, long-read sequencing for smaller tests or experiments
- Cost-efficient, 96-well plate format
- Compatible with high-throughput automation

Learn more >

15/05/2024



Uno de los últimos avances que se están produciendo mediante el uso de secuenciadores nanopore es el que de describe en este artículo publicado en Science (noviembre y diciembre, 2021). En el mismo se explica la metodología para secuenciar directamente proteínas, generando la cadena de aminoácidos en lugar de una secuencia de ADN.

Alfaro, J.A., Bohländer, P., Dai, M. *et al.* The emerging landscape of single-molecule protein sequencing technologies. *Nat Methods* **18**, 604–617 (2021). https://doi.org/10.1038/s41592-021-01143-1

https://www.nature.com/articles/s41592-021-01143-1

Ying, YL., Hu, ZL., Zhang, S. *et al.* Nanopore-based technologies beyond DNA sequencing. *Nat. Nanotechnol.* **17**, 1136–1146 (2022). https://doi.org/10.1038/s41565-022-01193-2

https://www.nature.com/articles/s41565-022-01193-2#citeas

Métodos de Secuenciación de ADN

Retos de la secuenciación de nueva generación → hacia una secuencia por cromosoma sin errores.

- Regiones de difícil secuenciación
- Errores de secuencia
- Tamaño de las lecturas (*reads*)
- Disminuir requerimientos computacionales

15/05/2024


Comparativa de secuenciadores

¿Qué secuenciador debo usar?

- El output más grande lo generan las máquinas de Illumina (NovaSeq), y a menor escala MiSeq.
- La longitud de lectura más grande la genera PacBio, con su alternativa más económica MinION.

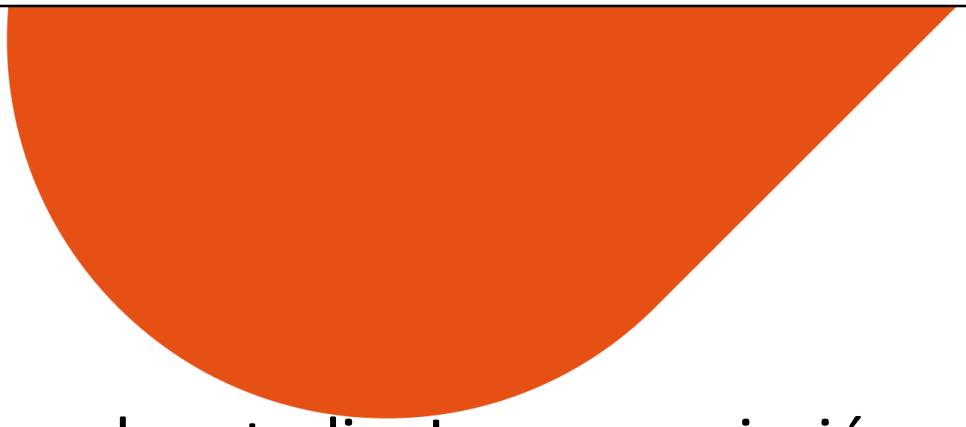
La combinación de métodos (secuenciación híbrida) es/era? la mejor solución para estudiar genomas eucarióticos, de gran longitud y complejidad

(hi-fi PacBio puede acabar con esta necesidad)

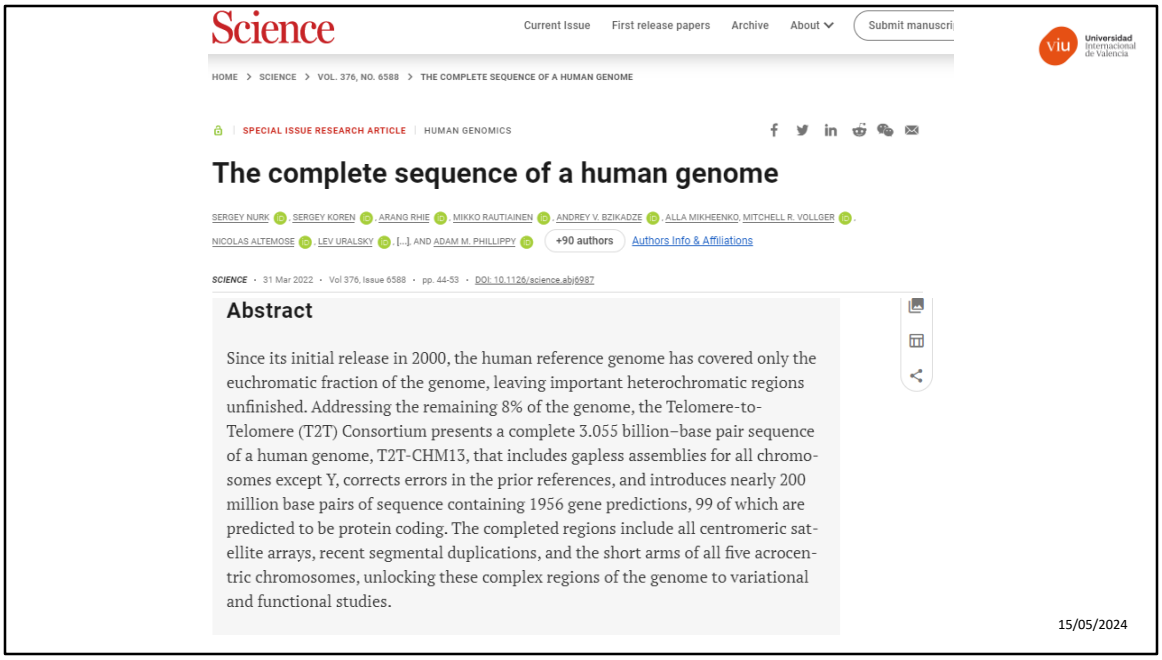
15/05/2024

A la hora de elegir el mejor método de secuenciación dependerá del tipo de estudio que vayamos a realizar. Si secuenciamos un solo genoma y queremos estudiar variabilidad a nivel de nucleótido único la mejor alternativa puede ser Sanger o Illumina, si queremos estudiar un metagenóma de un entorno complejo o amplicones 16S usaremos Illumina MiSeq. Por el contrario, si queremos secuenciar un genoma de alta concentración de regiones repetitivas, que presentarán dificultades en su ensamble, lo mejor será usar secuenciadores de tercera generación.

Todas estas decisiones dependerán finalmente de la diversidad de genomas que vayamos a secuenciar, de su longitud y de si dependemos de un paso posterior de ensamblaje. Una de las opciones que se han tomado durante los años en que la secuenciación de tercera generación tenía un alto porcentaje de errores ha sido realizar una secuenciación híbrida, utilizando secuenciadores de segunda y tercera generación. Las lecturas producidas por los métodos de tercera generación servirán como guía para las lecturas cortas de segunda generación, usando la información de esta última como solución a la secuencia original.



Caso de estudio: La secuenciación completa del genoma humano



https://www.science.org/doi/10.1126/science.abj6987

Cell line and sequencing

As with many prior reference genome improvement efforts (*1*, *8*, *17*–*20*), including the T2T assemblies of human chromosomes X (*14*) and 8 (*21*), we targeted a complete hydatidiform mole (CHM) for sequencing. Most CHM genomes arise from the loss of the maternal complement and duplication of the paternal complement postfertilization and are, therefore, homozygous with a 46,XX karyotype (*22*). Sequencing of CHM13 confirmed nearly uniform homozygosity, with the exception of a few thousand heterozygous variants and a megabase-scale heterozygous deletion within the rDNA array on chromosome 15 (*23*) (figs. S1 and S2). Local ancestry analysis shows that most of the CHM13 genome is of European origin, including regions of Neanderthal introgression, with some predicted admixture (*23*) (Fig. 1A). Compared with diverse samples from the 1000 Genomes Project (1KGP) (*24*), CHM13 possesses no apparent excess of singleton alleles or loss-of-function variants (*25*).

We extensively sequenced CHM13 with multiple technologies (*23*), including 30× PacBio circular consensus sequencing (HiFi) (*16*, *20*), 120× Oxford Nanopore ultralong-read sequencing (ONT) (*14*, *21*), 100× Illumina PCR-Free sequencing (ILMN) (*1*), 70× Illumina Arima Genomics Hi-C (Hi-C) (*14*), BioNano optical maps (*14*), and single-cell DNA template strand sequencing (Strand-seq) (*20*) (table S1). To enable assembly of the highly repetitive centromeric satellite arrays and closely related segmental duplications, we developed methods for assembly, polishing, and validation that better utilize these available datasets.

15/05/2024

Para conseguir el objetivo propuesto hace falta una estrategia de secuenciación que usa distintas tecnologías pero también un diseño de *que secuenciar*.

Cell line and sequencing

As with many prior reference genome improvement efforts (*1*, *9*, *13*, *24*, *28*, *29*), including the T2T assemblies of human chromosomes X (*11*) and 8 (*12*), we utilized a complete hydatidiform mole for sequencing. CHM genomes arise from the loss of the maternal complement and duplication of the paternal complement postfertilization and are, therefore, homozygous for one set of alleles. This simplifies the genome assembly problem by removing the confounding effect of heterozygous variation. We selected CHM13 for its stable 46,XX karyotype compared to other CHMs (*11*), but later found that CHM13 does possess a low level of heterozygosity, notably including a megabase-scale heterozygous deletion within the rDNA array on Chromosome 15, which was revealed by both FISH and nanopore sequencing (Figs. S1-2, Note S1). This and other identified heterozygous variants appear fixed in CHM13 and may have arisen during growth of the mole or passaging of the cell line. Local ancestry analysis shows the majority of the CHM13 genome is of European origin, including regions of Neanderthal introgression, with some predicted admixture from other populations (*30*) (**Fig. 1A**, Note S2).

Over the past 6 years, we have extensively sequenced CHM13 with multiple technologies (Note S3), including 30× PacBio circular consensus sequencing (HiFi) (*29*), 120× Oxford Nanopore ultra-long read sequencing (ONT) (*11*, *12*), 100× Illumina PCR-Free sequencing (ILMN) (*1*), 70× Illumina / Arima Genomics Hi-C (Hi-C) (*11*), BioNano optical maps (*11*), and Strand-seq (*29*). Here we developed new methods for assembly, polishing, and validation that better utilize these datasets. In contrast to the first T2T assembly of Chromosome X (*11*)—which relied on ONT sequencing to create a backbone that was then polished with other technologies—we shifted to a new strategy that leverages the combined accuracy and length of HIFI reads to enable assembly of highly repetitive centromeric satellite arrays and closely related segmental duplications (*12*, *22*, *29*).

15/05/2024

Para conseguir el objetivo propuesto hace falta una estrategia de secuenciación que usa distintas tecnologías pero también un diseño de *que secuenciar*.

Recent advances in sequencing technologies have greatly increased the accuracy and length of sequencing reads. Pacific Biosciences' high-fidelity (HiFi) reads can achieve accuracies of over 99.9% with read lengths of 18–25 kilobases (kb) and Oxford Nanopore Technologies (ONT) reads routinely reach median lengths of 50–150 kb with accuracies around 95%. Recently, ONT has demonstrated the ability to generate relatively shorter reads (median 25–35 kb) at 99.9% accuracy. For convenience when not referring to a specific technology, we will refer to

"long, accurate reads" (LA reads) as those with lengths greater than 10 kb and accuracy greater than 99.9%, and

"ultra-long reads" (UL reads) as those with lengths over 100 kb and accuracies over 90%.

Verkko: telomere-to-telomere assembly of diploid chromosomes
Mikko Rautiainen, Sergey Nurk, Brian P. Walenz, Glennis A. Logsdon, David Porubsky, Arang Rhie, Evan E. Eichler, Adam M. Phillippy, Sergey Koren
bioRxiv 2022.06.24.497523; doi: https://doi.org/10.1101/2022.06.24.497523

15/05/2024

https://www.biorxiv.org/content/10.1101/2022.06.24.497523v1.full

¡Gracias!


viu
Universidad
Internacional
de Valencia
universidadviu.com
De:
Planeta Formación y Universidades