UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Claves y evolución de las tecnologías de Secuenciación Masiva de Segunda Generación

Apellidos, nombre	Vilanova Navarro, Santiago (sanvina@upv.es) Gadea Vacas, José (jgadeav@ibmcp.upv.es)
Departamento	Departamento de Biotecnología
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

La aparición de las tecnologías de secuenciación masiva ha hecho posible que los proyectos de secuenciación de genomas estén al alcance de todos. Hoy en día, cualquier laboratorio con presupuesto ajustado pueden aventurarse a realizar este tipo de proyectos en sus líneas de investigación. Las propuestas son normalmente subcontratadas a empresas de secuenciación, que realizan el trabajo y parte del análisis bioinformático, dejando en manos del investigador la interpretación final de los datos.

A grandes rasgos, el funcionamiento general de un sistema de secuenciación masivo de segunda generación permite la secuenciación simultánea de millones de fragmentos en una única reacción, y por tanto, desbanca a la tecnología de Sanger si el objetivo es precisamente ese: secuenciar muchos fragmentos en un tiempo razonable. Esto, lógicamente, es lo que necesitan los proyectos de genómica. Por eso, desde su descubrimiento, todos los proyectos de escala genómico se secuencian con estas tecnologías.

En este artículo docente vamos a ilustrar las principales claves que han permitido el éxito de estas tecnologías, y también cómo en los últimos años, se han hecho avances en ciertos puntos críticos de la tecnología que han dado lugar a una expansión de tecnologías similares que compiten para un mismo objetivo, de variantes de una misma tecnología en función de la escala del proyecto o de diferentes tecnologías según el objetivo del proyecto. La revolución tecnológica en secuenciación masiva es un campo en continuo desarrollo en los últimos años y continuamente aparecen nuevas aplicaciones y nuevos desarrollos tecnológicos. Para ilustrar esto, usaremos el sistema original de Illumina como punto de partida. En algunos ejemplos, nos quedaremos dentro de Illumina para ver esos avances. En otros, acudiremos a otros sistemas y otras empresas para ver los avances de su competencia.

2 Objetivos

Una vez que el alumno haya leído con detenimiento este documento, será capaz de:

1. Entender las claves que han permitido el éxito de las tecnologías de secuenciación masiva:
    a. La química de los nucleótidos
    b. La paralelización de secuencias
    c. La estrategia de secuenciación

2. Entender que la secuenciación masiva es un campo en continua evolución, cuyos avances permiten acceder a nuevas aplicaciones genómicas a un precio cada vez más bajo.

3 Introducción

La revolución reciente de las estrategias de secuenciación masiva ha permitido que la secuenciación de genomas completos sea una estrategia accesible a un precio razonable. Desde su uso diario en hospitales para la secuenciación de genomas de pacientes hasta los más variados objetivos biotecnológicos pueden ser abordados, algo impensable hace sólo diez años. El éxito de estas tecnologías está basado en avances multidisciplinares en el campo de la química, la biología molecular, la ingeniería y la bioinformática. A continuación, veremos lo aspectos más relevantes que han contribuido al empuje de estas tecnologías.

4 Desarrollo

Para entender el éxito de estas tecnologías, es importante fijarse en avances en tres aspectos diferentes del proceso de secuenciación:

4.1. La química de los nucleótidos.

La síntesis de análogos artificiales de nucleótidos con grupos químicos bloqueando el enlace 3-OH de modo reversible ha sido clave en el éxito de las tecnologías de secuenciación masiva. El sistema de ciclos de secuenciación con adición y retirada de nucleótidos es lo que permite que la paralelización sea una ventaja. Con los dideoxinucleótidos irreversibles de la tecnología de Sanger, la paralelización tampoco hubiera significado un gran avance. Intenta imaginar el proceso de secuenciación de Illumina utilizando los dideoxinucleótidos de la tecnología de Sanger.

Además, otra clave para la captura de las señales en cada uno de los ciclos es generar un punto de corte químico entre la base y el fluoróforo, que permita la escisión de este último tras cada ciclo de adición de nucleótidos.

Fíjate en el esquema siguiente de los nucleótidos y localiza lo sitios que favorecen la terminación y la unión (reversible por escisión) de los fluoróforos en las tecnologías de Sanger e Illumina:

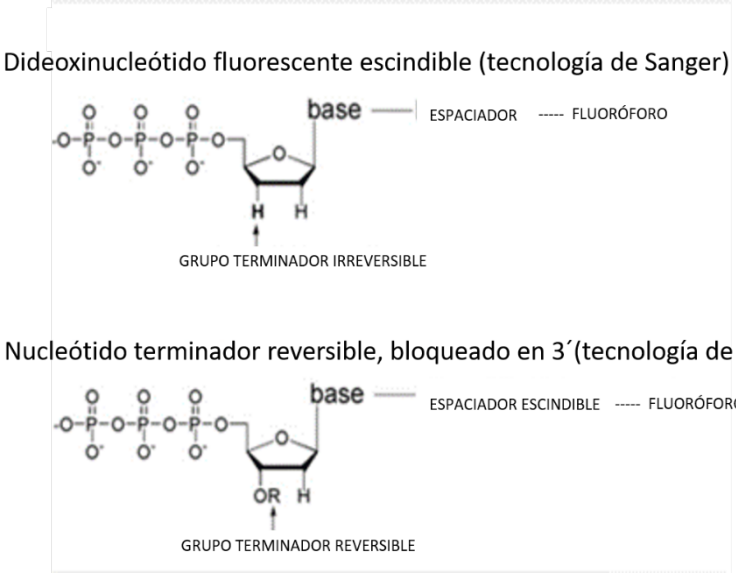


***Figura 1**. Esquema representativo de un nucleótidos irreversibles y reversibles utilizado en secuenciación masiva por las tecnología de Sanger e Illumina, respectivamente. Las flechas señalan el grupo 3-OH necesario para el enlace fosfodiéster.*

Sin embargo, las polimerasas han evolucionado durante millones de años para "preferir" los nucleótidos tal como son en la naturaleza. Esto hace que, en general, las mayoría de las polimerasas conocidas tengan menos afinidad si el nucleótido está modificado artificialmente en el grupo 3-OH, una zona clave para la estructura de una DNA polimerasa, que está ajustada estructuralmente para realizar el enlace fosfodiéster en 3-OH y para poder discernir ribonucleótidos de desoxiribonucleótidos en el 2-, por ejemplo. En resumen, la reacción funciona, es decir, los nucleótidos bloqueados se incorporan a la cadena naciente por la polimerasa, pero lo hacen menos eficientemente que si fueran nucleótidos naturales, sin ninguna modificación en la zona 3-OH.

¿Cómo han avanzado las investigaciones en NGS para conseguir una mayor eficiencia en el proceso de secuenciación con nucleótidos reversibles?

1. Mejorando la DNA polimerasa. Hacían falta nuevas DNA polimerasas con diferentes afinidades para diferentes análogos de nucleótidos. Curiosamente, la rápida expansión de proyectos de secuenciación masiva permitió identificar nuevas secuencias de polimerasas de virus, bacterias y arqueas para ensayar su eficiencia frente a estos análogos de nucleótidos. Los mejores candidatos fueron posteriormente mejorados con ingeniería de proteínas o evolución directa en el laboratorio. Estas líneas de investigación no se limitaban a encontrar la mejor polimerasa para aceptar análogos de nucleótidos, sino la polimerasa más rápida, la más fiel en la síntesis (menor error en la secuenciación) y la más procesiva (aquélla capaz de sintetizar el mayor número de nucleótidos sin "soltarse" del molde). De todos los tipos de DNAs polimerasas, las de las familias A y B de virus (T4, T7, Rb69 o ϕ29), bacterias o arqueas (Vent, Pfu, KOD1) han resultado ser las más adecuadas para secuenciación masiva. La afinidad por los nucleótidos reversibles como los que usa Illumina es alta con una polimerasa de la familia B de Thermococcus (9N). Posteriores modificaciones por mutagénesis dirigida han mejorado aún más las propiedades de las mismas.

2. Mejorando la estructura química de los nucleótidos. Otro punto de mejora en los últimos años ha sido actuar sobre los nucleótidos. En los últimos años se han desarrollado nucleótidos terminadores reversibles NO bloqueados en el grupo -OH. Como hemos dicho, tener el grupo 3-OH libre facilita la incorporación por las polimerasas. Esta nueva generación de nucleótidos, por tanto, son más eficientes en la incorporación. La paralización de la cadena naciente (que necesita seguir ocurriendo en una secuenciación NGS) se consigue mediante la adición del mismo grupo fluróforo en la zona de la base nitrogenada, que dificulta estructuralmente la adición del siguiente nucleótido

Fíjate en las estructuras siguientes y observa cómo cambia de posición el sitio responsable de la terminación de la reacción, del 3-OH en el primero, a la base en el segundo:

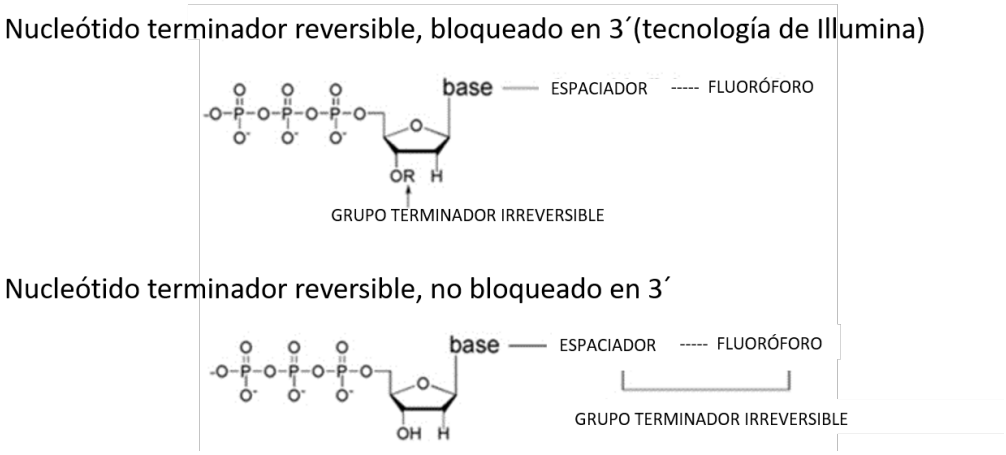


***Figura 2**. Esquema representativo de un nucleótido reversibles con y sin bloqueo en el grupo 3-OH. Las flechas señalan el grupo 3-OH necesario para el enlace fosfodiéster.*

En la siguiente figura puedes observar la diferencia estructural entre un nucleótido terminador reversible bloqueado en 3-OH (concretamente el 3´-O-azidometil, desarrollado por Illumina), y un nucleótido terminador reversible no bloqueado (desarrollado por la empresa Helicos, cuyo sistema de secuenciación no está ya disponible). Fijaros como el prominente grupo químico (en rojo) está cercano al grupo 3-OH, (que en este caso no está bloqueado). Esta estructura facilita la incorporación del nucleótido por la polimerasa, pero no la adición de un siguiente nucleótido, haciendo por la tanto el mismo papel que si el grupo 3-OH estuviera bloqueado. El posterior corte en la zona de la flecha libera en un único paso el fluoróforo y este grupo interferente:

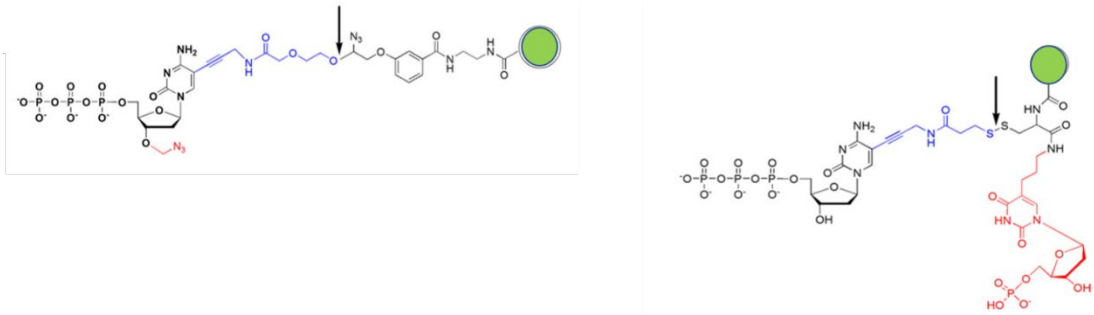


Figura 3. *Estructura química de un nucleótido terminador reversible (izquierda) con el grupo 3-OH bloqueado por un grupo químico (rojo) y otro (derecha) con el grupo bloqueante en la zona de la base nitrogenada (rojo). La flecha indica el punto de escisión para la incorporación del siguiente nucleótido. En verde, el grupo fluoróforo.*

4.2. La paralelización.

Ya hemos visto como la capacidad de poder paralelizar millones de reacciones de secuenciación es la clave para el éxito de las secuenciación masiva, que facilita así disponer de una gran cantidad de información de secuencia en un tiempo reducido. La creación de soportes sólidos donde están inmovilizados los fragmentos a secuenciar es un aspecto común a todas estas tecnologías, aunque el funcionamiento de cada una de ellas sea distinto en muchos detalles. Vamos a ilustrar dos aspectos de las tecnologías de secuenciación masiva en este apartado:

a) la capacidad de paralelización de estos soportes,

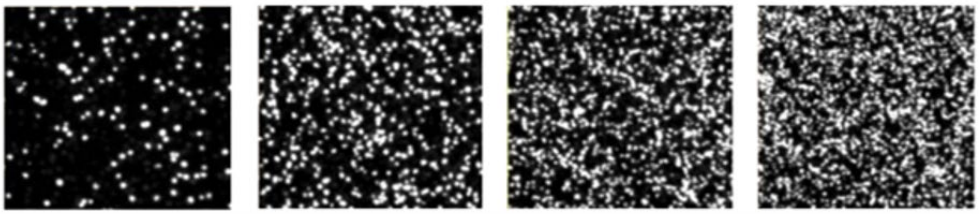
b) la versatilidad actual de estos soportes para acoplarse a proyectos genoma de diferente magnitud.

La base de estas tecnologías (usemos el sistema de illumina como ejemplo) es la célula de flujo (flowcell en inglés), un dispositivo miniaturizado de cristal con varios canales individualizados a través de los cuales pueden bombearse polimerasas, nucleótidos y tampones para realizar la reacción de secuenciación. Estas células de flujo tendrán (tras el proceso de construcción de las librerías) los fragmentos preparados para su secuenciación.

Vamos a ver la capacidad de paralelización que tienen las células de flujo de Illumina. Estas células de flujo han ido mejorándose en los últimos diez años. Los avances se han centrado en aumentar la paralelización (para acoplarse a proyectos genoma cada vez más ambiciosos en cuanto a capacidad de secuenciación, como la secuenciación de miles de genomas humanos, por ejemplo) o bien disminuirla!, (para acoplarse a proyectos donde una excesiva paralelización , es decir, cuando la cantidad de fragmentos que puede aceptar una célula de flujo es mayor que el que se desea secuenciar, ya que la tecnología estaría infrautilizándose, con el consecuente perjuicio a nivel económico; un ejemplo sería la secuenciación de pequeños genomas o grupos (paneles) de genes.

Partamos de los secuenciadores HiSeq de Illumina (esta serie de secuenciadores desbancó en 2010 a la serie anterior de Illumina, los GAs, o Genome Analizers, al aumentar por cinco la velocidad de adquisición de datos al mejorar sus sistemas ópticos de adquisición de imagen). Este cambio de sistema supuso también un cambio en el diseño de la célula de flujo. Esta mantuvo un sistema de 8 canales, pero pronto incorporó un patrón ordenado en la deposición de fragmentos en el soporte, que hasta ese momento se hacía de modo aleatorio.

Aquí puedes ver el sistema aleatorio de generación de los clusters de fragmentos según el método original de Illumina, que requería una optimización en los protocolos de generación de librerías para obtener una densidad óptima de clusters en cada célula de flujo. Fíjate en la poca eficiencia en cuanto a número de fragmentos en la figura de la derecha, que mantiene grandes "espacios" de la célula de flujo sin fragmentos, y como un diseño aleatorio, aunque más optimizado, consigue "rellenar" más la célula de flujo de fragmentos sin que interfieran unos con otros. Sin embargo, el diseño es aún ineficiente, ya que siguen quedando "espacios" sin aprovechar:



Bajo número de clusters *Excesivo número de clusters*

Figura 4. *Imagen representativa de una célula de flujo de la tecnología Illumina con patrón de fragmentos aleatorio. A la izquierda, el número de fragmentos es muy bajo, y la eficiencia de la paralelización es baja. Sin embargo, una alta densidad de fragmentos tampoco es óptima (derecha), ya que las señales interfieren entre los diferentes fragmentos.*

La serie HiSeq incorporó pronto un sistema ordenado de generación de clusters de fragmentos. Esto permitió aumentar la paralelización de la tecnología Illumina respecto a su diseño original. Fijate en la siguiente figura cómo con este diseño, el "espacio" de la célula de flujo está optimizado, permitiendo la máxima paralelización posible de cada una.

En otro ámbito, las aplicaciones de las NGSs son cada vez más diversas. Para grandes proyectos de secuenciación, ya no se quiere secuenciar un genoma de referencia, sino resecuenciar miles de genomas individuales en un tiempo corto, requiriendo secuenciadores con MÁS capacidad de secuenciación. Para aplicaciones de diagnóstico o genomas pequeños, por el contrario, tanta potencia de secuenciación no es efectiva a nivel de coste, por lo que precisa de secuenciadores con MENOS capacidad de secuenciación.

La serie NovaSeq de Illumina mejora de nuevo el sistema, mejorando la óptica y las celulas flujo (aumentando la miniaturización del sistema ordenado de clusters de los HiSeq), y generando además un sistema escalable en función de la necesidad de secuenciación. Para ello, centrándonos en las células flujo, aparecen ya cuatro tipos, con diferente versatilidad dependiendo del sistema de secuenciación. Fijaros en la cantidad de lecturas (fragmentos) que es capaz de secuenciar cada una (M: millones; B: billones)

	NovaSeq 6000			
Tipo de célula de flujo	SP	S1	S2	S4
Nº de lecturas sencillas	650–800 M	1.3–1.6 B	3.3 B–4.1 B	8-10 B

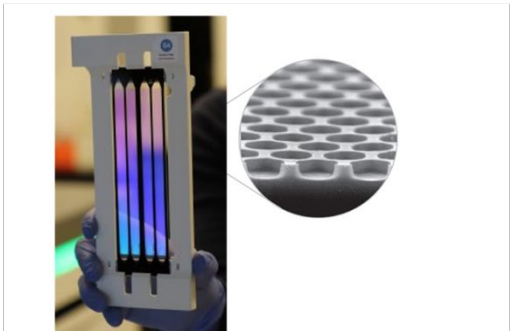


Figura 5. *Células de flujo de Illumina Novaseq con patrón ordenado para el espaciamiento de los fragmentos a secuenciar. Los diferentes tipos de células se adaptan al objetivo deseado con diferentes capacidades (número de lecturas sencillas).*

Finalmente, puedes echar un vistazo a los diferentes secuenciadores de Illumina en su página web, diseñados para proyectos genómicos que requieren diferente capacidad de secuenciación. La seria Nova Seq, con sus cuatro tipos de células de flujo versátiles en función de la capacidad de secuenciación requerida en cada proyecto, es actualmente el sistema más avanzado. En el gráfico siguiente puedes ver también la capacidad de secuenciación de cada célula de flujo NovaSeq, en comparación con las capacidades que ofrecían los secuenciadores HiSeq.

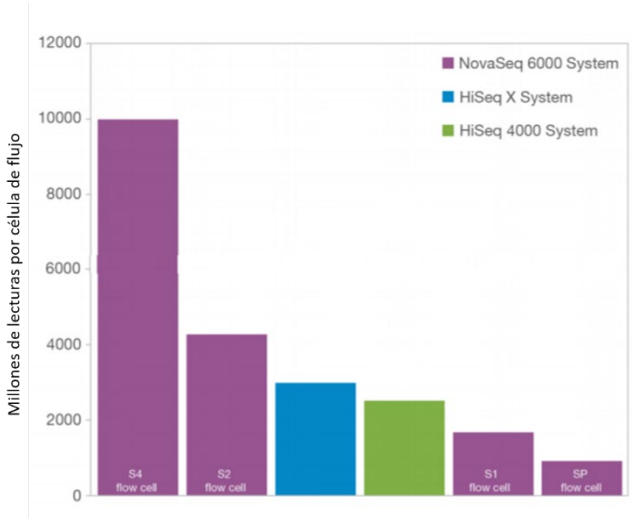


Figura 6. *Comparación de la capacidad de secuenciación (en número de lecturas), en función del sistema (Hiseq, Novaseq) y la célula de flujo (flowcell) utilizada.(adaptado de www.illumina.com)*

4.3. La estrategia de secuenciación.

Aunque la base de las tecnologías NGS está principalmente basada en los dos puntos anteriores (una gran paralelización de fragmentos, acoplada a un eficiente sistema de secuenciación), en los últimos años las diferentes empresas de secuenciación masiva han mejorado diferentes puntos del proceso de secuenciación, con el fin de hacer un proceso global más eficiente. Vamos a ver dos ejemplos de estas mejoras:

4.3.1 Reducción de costosos mecanismos ópticos.

Aunque la propia tecnología de secuenciación de Illumina basada en ciclos con nucleótidos reversibles marcados, tal como la hemos contado anteriormente, supuso una revolución en el proceso de secuenciación, ésta ha ido mejorándose en los últimos años en diferentes aspectos. Uno de ellos tiene que ver con la reducción del coste del equipamiento, fabricando equipos que contiene únicamente uno o dos láseres, en lugar de los cuatro (uno por fluoróforo) que requiere el sistema original (un sistema cíclico de secuenciación con 4 nucleótidos marcados cada uno de ellos con un fluoróforo diferente). Pero, para utilizar un equipo con dos o un láseres, el sistema de marcaje de nucleótidos debía cambiar. Así es como lo solucionaron en Illumina:

En el sistema de cuatro canales, las bases se identifican usando cuatro diferentes fluoróforos, uno para cada base, y cuatro imágenes por ciclo de secuenciación (fíjate en la figura siguiente, panel izquierdo. Este es el sistema original de Illumina). El ciclo de secuencia comienza cuando las cuatro bases marcadas diferencialmente se agregan a la célula de flujo. Después de la incorporación de un nucleótido por fragmento, se capturan cuatro imágenes distintas utilizando cuatro diferentes bandas de longitud de onda. Hacen falta, por tanto, cuatro láseres en el secuenciador. Las imágenes se procesan para determinar qué nucleótidos se incorporaron en cada posición del grupo a través de la celda de flujo. Por lo tanto, con cuatro canales, cada ciclo de secuenciación requiere cuatro fluoróforos y cuatro imágenes para determinar la secuencia de ADN. MiSeq y HiSeq utilizan cuatro canales.

En lugar de usar un fluoróforo diferente para cada base, el sistema de dos canales simplifica la detección de nucleótidos mediante el uso de únicamente dos fluoróforos y dos imágenes para determinar las cuatro bases (Figura siguiente, panel central). Hacen falta equipos con únicamente dos láseres. Las timinas están marcadas con un fluoróforo verde, las citosinas están marcadas con un fluoróforo rojo y las adeninas están marcadas con fluoróforos rojos y verdes. Las guaninas no se marcan, por lo que no darán señal.

Finalmente, el sistema iSeq 100 utiliza un único fluoróforo, dos pasos de procesos químicos y dos pasos de captura de imagen por ciclo de secuenciación. En el sistema de un canal, la adenina tiene una fluoróforo escindible y su señal se captura en una primera imagen. La citosina tiene un grupo enlazador que puede unir un fluoróforo, y se captura en la segunda imagen. La timina tiene un fluoróforo permanente, y, por lo tanto, se captura en ambas imágenes y la guanina no contiene fluoróforo. Los nucleótidos se identifican por el

análisis de los diferentes patrones de emisión para cada base en las dos imágenes. Como veis, precisa de un equipo con un solo láser, adecuado para secuenciadores muy pequeños. (Figura siguiente, panel derecho).

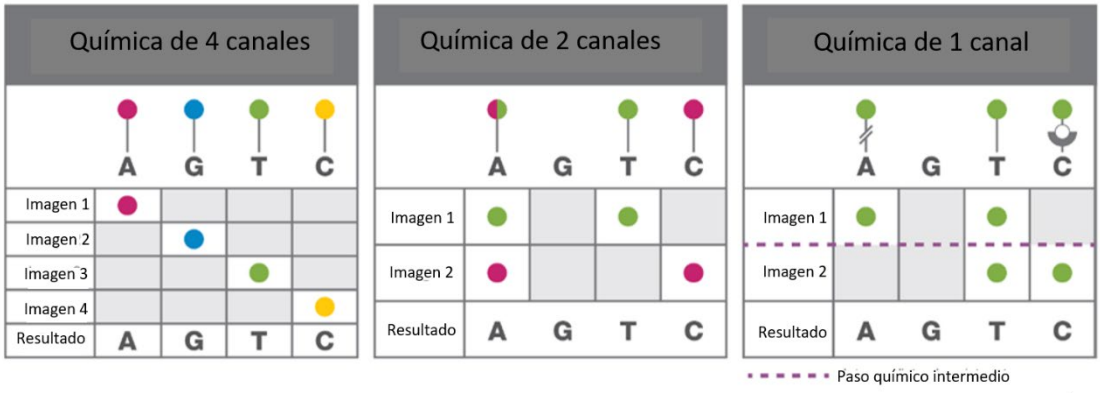


Figura 7. *Comparación de las estrategias de cuatro, dos y un láser para la adquisición de señal proveniente de los nucleótidos incorporados. (adaptado de www.illumina.com)*

4.3.2 Secuenciación codificada.

Presentamos aquí otra innovación dentro de las tecnologías de secuenciación masiva que permite aumentar la eficiencia de la secuenciación y por tanto reducir el coste. Como hemos visto, aunque un proyecto se secuenciación grande como la secuenciación de genomas eucariotas necesite mucha paralelización, en muchos casos el proyecto deseado es la secuenciación de genomas pequeños o subgenomas (transcriptomas, exomas, paneles, regiones concretas del genoma) cuyos requerimientos de paralelización están por debajo de la capacidad de carga de una célula de flujo. (Imagina pagar la gasolina de un viaje por Blablacar a Madrid entre 5 personas, o tu sola/o, el viaje por persona te costará menos si el coche va lleno!).

La solución a esto sería compartir gastos. Para eso, estas tecnologías pronto implementaron el uso de adaptadores codificados para la construcción de librerias (llamados *barcodes* o *index* en inglés, según la tecnología). El uso de un adaptador diferente para cada librería permite "juntar" fragmentos provenientes de librerías diferentes en una misma célula de flujo, de forma que esta se llene a la máxima capacidad. Todos los fragmentos de todas las librerías se secuencian simultáneamente. Pero se secuencia el fragmento, y también el codificador. que lleva dicho fragmento. Tras la adquisición de las señales y la obtención de las secuencias, cada fragmento es asignado bioinformáticamente al proyecto de secuenciación del que provenía. Puedes ver un resumen de esta idea en la siguiente figura. Esta idea, junto a la fabricación de secuenciadores de diferente capacidad, como hemos visto, ha expandido enormemente las posibilidades de secuenciación de NGS en toda la escala de proyectos de secuenciación.

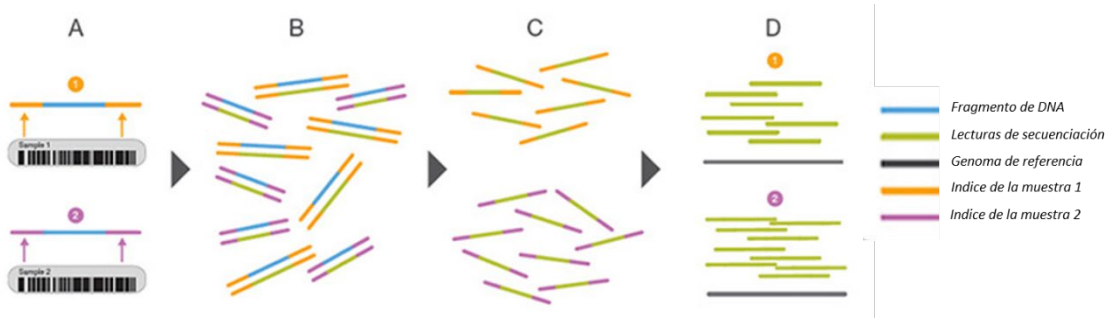


Figura 8. *Codificación de fragmentos mediante índices de ADN. (A) Dos muestras diferentes reciben dos adaptadores que se diferencian en una secuencia índice que actúa de código de barras. (B). Las librerías se realizan para las dos muestras independientemente, pero (C) las lecturas se separan posteriormente a la secuenciación gracias al código de barras de cada muestra, de modo que (D) cada lectura puede mapearse con el genoma de referencia del que provenga cada proyecto. (adaptado de www.illumina.com)*

5 Cierre

En este artículo docente hemos visto las claves del éxito de la secuenciación masiva para proyectos genómicos. También hemos visto como no se trata de una tecnología estática, sino que diferentes mejoras en diversos aspectos de la tecnología han permitido, a partir de un esquema básico de secuenciación masiva propuesto inicialmente, aumentar la rapidez del sistema, reducir el precio por reacción, abaratar equipamiento, así como expandir el abanico de aplicaciones, atendiendo a la diferente necesidad de secuenciación. Estas mejoras han incluido aspectos de la propia química de la secuenciación, búsqueda de nuevas enzimas, diferentes aspectos relacionados con la robótica, la ingeniería y la mejora del equipamiento óptico, así como ingeniosas ideas con base de biología molecular y bioinformática para optimizar al máximo el proceso, como la aparición de sistemas con codificadores.

6 Bibliografía

6.1 Artículos de revisión

Chen F, Dong M, Ge M, Zhu L, Ren L, Liu G, Mu R. The history and advances of reversible terminators used in new generations of sequencing technology. Genomics Proteomics Bioinformatics. 2013 Feb;11(1):34-40. doi: 10.1016/j.gpb.2013.01.003.

Chen, Cheng-Yao. “DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present.” *Frontiers in microbiology* vol. 5 305. 24 Jun. 2014, doi:10.3389/fmicb.2014.00305

Illumina “CMOS Chip and One-Channel SBS Chemistry” 2014. https://www.illumina.com/content/dam/illuminamarketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf