

Introducción a las técnicas de secuenciación masiva – Aplicaciones translacionales (Gonzalo Gómez)

Introducción

Tecnologías

NGS y enfermedad

Retos y conclusiones

Introducción

En los 1970's Sanger propone un método de secuenciación y electroforesis de nucleótidos dideoxi- que se convirtió en el primer método de secuenciación que permitiría obtener secuencias realmente relevante. Se inicia en 1991 el proyecto genoma humano (HGP). En el año 2001 se presenta el primer borrador del genoma. El proyecto genoma humano termina en el 2003 al entregar la secuencia definitiva. Desde el punto de vista de la tecnología, se ha visto un incremento espectacular en el *throughput* de secuencias iniciando con el método Sanger que fue el utilizado en HGP. La segunda generación, denominada con el confuso nombre de **Next Generation Sequencing o NGS**, arrancó en el 2004-2005 con máquinas automatizadas de compañías con Illumina, Roche, Applied Biosystems. Hay una tercera generación, que apenas está incursionando actualmente que se basa en métodos de detección no óptica basada en semiconductores, que fue iniciada en 2010 y que promete una reducción sustancial en el tiempo y el costo. Esta nueva generación la llaman la era de la **secuencia de mil dólares**. La primera compañía en proponer esta tecnología es Life Technologies con el Ion Torrent, disponible para la secuenciación de exomas desde el 2012.

El costo por megabase a lo largo de los años. En el 2006 se produce una caída exponencial del costo por megabase, que supera incluso la caída del coste de Mbyte de memoria computacional (El genoma secuenciado en el HGP costó unos US \$ 1-3 billones, en 13 años. En la década actual se espera el genoma de los mil dólares, en 2 semanas). Ahora el reto no es la velocidad de generación de datos, sino la capacidad de manejo de datos – por su físico VOLUMEN. El costo de secuenciación decrece más rápidamente que la ley de Kryder para el coste del almacenamiento de datos. El reto bioinformático es tan grande como el de secuenciación pura. Las potencias mundiales en High-Thoroughput- Sequencers son USA y China. También hay en Europa, Australia, México, Brasil y Argentina. En el Beijing Genomics Institute (BGI) tienen 137 Illumina HSeq 2000, 27 ABI SOLID 4 y Roche GS FLX Titanium. Eso es una capacidad mayor de 10.000 Gbp/day (más de 100 genomas humanos a 30x / día). Aproximadamente 40% de los empleados del BGI son bioinformáticos.

Tecnologías – máquinas de secenciación automática

Las primera generación son las que permitieron el HGP. La segunda generación son NGS, deep sequencing. La tercera generación, o de genoma de los US \$1000, que estamos esperando.

NGS

Las principales máquinas que se usan son la plataforma Illumina. El pipeline típico de trabajo: 1. DNA genómico, 2. Fragmenta, 3. Eliminar DNA que no hibrida con primers de exones. 4. Se introduce en el secuenciación.

Illumina: Fija cada fragmento a una fase sólida con polimerasa que va secuenciando el DNA complementario. Cada nucleótido tiene un color diferente. Se repite 30-60 veces, y cada foto se va superponiendo mejorando la resolución de la detección.

Roche 454 utiliza el método de pirosecuenciación en donde la señal lumínica identifica el nucleótido que se observa en el resultado donde la secuencia se establece según la ubicación relativa.

Applied BioSystems – SOLID secuencia dinucleótidos, con códigos de colores. La tasa de error es menor que las otras tecnologías.

Todas estas requieren de detección ÓPTICA que las hace complejas desde el punto de vista tecnológico.

Secuencia de \$1000 dólares

Ion Torrent (<http://www.iontorrent.com/>) – es una máquina pequeña, de no tan alto throughput. Se basa en membranas semiconductoras en donde se colocan las secuencias en una serie de pozos y la polimerasa al producir un nucleótido emite un protón, que baja el pH, que es leído y se genera la secuencia. No es basado en tecnología óptica.

MinION – es un dispositivo como un USB que permite tomar sangre, utilizando un software que le permite secuenciar DNA y proteínas específicas por electricidad.

Paso de la bioinformática

Una vez obtenido toda la secuencia, por cualquiera de las tecnologías ya mencionadas, se procede a compararla contra secuencias de referencias. Posteriormente, se procesa al paso de Alineación para mapear las lecturas a la referencias, identificar variantes, con visualización. Finalmente, se procede a la interpretación: analiza las variantes, y predice. Todo esto requiere de software sofisticados que continúan en evolución.

Aplicaciones

Toneladas. Ejemplo: El proyecto de los 1000 genomas – mapa genético de las variantes genéticas asociadas a 20 poblaciones humanas. Ha mostrado más de 38 millones de SNPs, entre otras cosas.

NGS y enfermedad

Variantes genómicas y enfermedades

Enfermedades monogénicas - Mendelianas

Nature Genetics – Exome sequencing makes medical genomics a reality (Leslie G Biesecker, 2010) que identificó el gen causal del síndrome de Miller utilizando NGS en

4 pacientes. En algunas enfermedades metabólicas neurológicas ha permitido hacer translación al tratamiento. También se ha utilizado para encontrar fusiones crípticas que causan enfermedad (Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene – APL especial (JAMA 2011)).

Proyectos de investigación

1. The 1000 Mendelian Project
2. Cancer Genome Project: cada país se encarga de un cáncer. Es cartografiar el cáncer. España con CLL. Se almacenan en el ICGC Data accession Site.

Conclusiones y retos

Qué es enfermedad, qué es riesgo y qué es variación. Cómo manejar la información. Cómo se integra con otras estrategias como epigenoma. Es algo así como se debió haber sentido Tycho Brahe antes de su Copérnico.

Versión 1.0 – 21.04.2013