UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ

ACTA BIOLÓGICA COLOMBIANA
http://www.revistas.unal.edu.co/index.php/actabiol

ARTÍCULO DE REFLEXIÓN/REFLECTION PAPER

APLICACIONES DE LA BIOINFORMÁTICA EN LA MEDICINA: EL GENOMA HUMANO. ¿CÓMO PODEMOS VER TANTO DETALLE?

Bioinformatics Applications in Medicine: The Human Genome. How Can We See Such Detail?

Clara Isabel BERMUDEZ-SANTANA[1].

[1] Departamento de Biología, Universidad Nacional de Colombia, Sede Bogotá. Cra. 30 n.º 45-03, edificio 421, oficina 207. Bogotá, Colombia.

For correspondence. cibermudezs@unal.edu.co

Received: 13th June 2015, **Returned for revision**: 2nd August 2015, **Accepted**: 1st September 2015.

Associate Editor: María Consuelo Burbano Montenegro.

Citation / Citar este artículo como: Bermudez-Santana C. Aplicaciones de la bioinformática en la medicina: el genoma humano. ¿Cómo podemos ver tanto detalle?. Acta biol. Colomb. 2016;21(1)Supl:S249-258. doi: http://dx.doi.org/10.15446/abc.v21n1sup.51233

RESUMEN

La bioinformática es un campo novedoso que soporta parte de la investigación biológica dirigida a la identificación de variantes génicas que pueden ser descubiertas desde los estudios de genomas completos. Basados en esta motivación se presenta el panorama general de los aportes principales de la bioinformática en el desarrollo del secuenciamiento del primer genoma humano. Adicionalmente se resumen los principales avances en cómputo desarrollados para responder a las demandas requeridas por los métodos de secuenciamiento de última generación para lograr re-secuenciar un genoma humano. Finalmente se introducen algunos de los nuevos retos que deben asumirse para aplicar la genómica personalizada en el desarrollo de la medicina.

Palabras clave: bioinformática, genoma humano, genómica personalizada, secuenciamiento.

ABSTRACT

Bioinformatics is a novel field that supports part of the biological research aimed at identifying gene variants that can be discovered from studies of whole genomes. Based on this motivation the overview of the main contributions of bioinformatics in the development of sequencing the first human genome is presented. Additionally it is summarized the main advances in computing developed to meet the demands to re-sequence a human genome by using the next generation sequencing technologies. Finally some new challenges that should be faced to apply the personalized genomics into the medicine development are introduced.

Keywords: bioinformatics, human genome, personalized genomics, sequencing.

INTRODUCCIÓN

Uno de los principales retos de la medicina reconocido por muchos investigadores desde el secuenciamiento del genoma humano, ha sido identificar a escala genómica la variación génica que puede estar asociada con algunas enfermedades humanas. Líderes mundiales han trabajado en el desarrollo de metodologías experimentales, modelos matemáticos y computacionales para continuar estudiando e identificando las variaciones que pueden detectarse a partir del análisis genómico. Investigadores y líderes del estudio de la genómica, entre ellos, Michael Snyder (Director del Centro de Genómica y Medicina Personalizada de la Universidad de Stanford) nos señala "que la genética y la genómica están experimentando una revolución extraordinaria y nuestra misión es continuar liderando esta revolución para una mejor comprensión de la biología y la salud humana" (Chen y Snyder, 2014).

En el artículo se presenta una reflexión e identificación del esfuerzo del trabajo interdisciplinario liderado por la bioinformática, de los principales desarrollos de programas de cómputo que permitieron llevar a cabo el secuenciamiento del primer genoma humano y de su uso para resolver los problemas derivados de los métodos de secuenciamiento de nueva generación, que como resultado produjeron el secuenciamiento del segundo genoma humano y el secuenciamento de 1000 genomas humanos. Por último se introducen los retos de la era de la genómica personalizada y su posible unión con la medicina.

PRELUDIO BIOINFORMÁTICO

Sin lugar a dudas, los logros de la genética molecular y la biología celular en el pasado han sido acompañados de los avances computacionales necesarios para el procesamiento de la información genética. En palabras de Ouzounis y Valencia (2003) este primer acercamiento a la influencia de la bioinformática sobre la biología molecular, no solo recuerda el continuo avance en la vida moderna gracias al desarrollo de la informática, sino el de su influencia para convertirla en uno de los campos altamente visibles de la ciencia moderna.

En los orígenes de la bioinformática muchos de sus pioneros desarrollaron los principios fundamentales para construir el complejo marco conceptual requerido, desde el punto de vista computacional, para responder a preguntas relacionadas con la variación en las secuencias de los genes, de las proteínas y de los genomas. Estos primeros trabajos no se escaparon de la tarea convencional que se hace en la bioinformática que es en principio trasladar problemas biológicos a problemas computacionales.

Aunque la complejidad de los problemas biológicos no siempre puede ser resuelta computacionalmente, debido a la carencia de algoritmos o modelos matemáticos o por limitantes de equipos de cómputo de alto poder que puedan calcular operaciones para resolverlos, si podemos reconocer en los pioneros de la bioinformática el valor de haber soñado con lo imposible en su época, ya que no se conocía la secuencia del genoma de ningún organismo, y sin embargo hicieron parte de la construcción del andamiaje teórico que le permite hoy en día a miles de investigadores en el mundo realizar el análisis genómico.

Por ejemplo, al inicio de los años 90 ya se habían diseñado e implementado algoritmos para el análisis comparativo de secuencias de proteínas y de genes o para la búsqueda de patrones o repeticiones (Ouzounis y Valencia, 2003), esto cuando aún no se había secuenciado el genoma de un organismo vivo, sólo se habían secuenciado los genomas de los virus ΦX174 (Sanger *et al.*, 1977) y del herpes Epstein-Bar (Baer *et al*., 1984). Años después, en 1995 se publicaron los primeros genomas bacterianos para las especies *Haemophilus influenzae* (Fleischmann *et al*., 1995) y *Mycoplasma genitalium* (Fraser *et al*., 1995).

En estos primeros años se construyó la teoría para la comparación de secuencias de proteínas basada en los trabajos de construcción de las matrices de sustitución y de matrices PAM liderados por Dayhoff (Dayhoff *et al.*, 1978), que posteriormente fue adaptada para el estudio de secuencias de DNA y conceptualmente modificada para el análisis de secuencias más largas. Este primer gran avance es conocido en el lenguaje de la bioinformática como el alineamiento de cadenas y de secuencias utilizado para comparar dos o más secuencias de ADN o ARN o de proteínas y cuantificar su grado de similitud.

Los modelos de alineamiento global para pares de cadenas fueron desarrollados por Needleman y Wunsch (1970), incluyendo restricciones por inserciones o deleciones por Sankoff (1972) y el uso de matrices de mutación por Dayhoff *et al*. (1978) que fueron extendidos para alineamiento locales por Smith y Waterman (1981a;1981b). Posteriormente, Feng y Doolittle (1987) diseñaron los algoritmos que permiten el análisis comparativo múltiple de más de dos cadenas. Por otro lado, se diseñó la familia de algoritmos que basan su búsqueda en bases de datos como FASTA (Wilbur y Lipman, 1983; Lipman y Pearson, 1985), y los basados en perfiles de secuencias por Gribskov *et al*. (1987).

De forma paralela al desarrollo e implementación de algoritmos para el análisis de secuencias, no se puede dejar a un lado la importancia de buscar la manera eficiente de almacenar la información para poder consultarla eficientemente. Ouzounis y Valencia (2003) indican que la creación de las dos primeras fuentes para el almacenamiento de datos de secuencias de genes y proteínas se realizó antes de los 90, conocidas actualmente como el *GenBank* (Bilofsky *et al*., 1986) y el EMBL *Data Library* (Hamm y Cameron, 1986).

Posteriormente en el año 2002, se creó el DDBJ (Banco de Datos del Japón). Además, en estos años se dieron las primeras iniciativas para la construcción de redes de comunidades de la bioinformática, que permitieron, canalizar y difundir los desarrollos en el campo, para permitir una comunicación a nivel mundial entre los investigadores

como lo fueron BIONET (Smith *et al*., 1986; Kristofferson, 1987) y EMBNET (Lesk, 1988).

EL GENOMA HUMANO: INICIOS DE SU CÓMPUTO

El termino bioinformática en sus inicios no era parte del lenguaje utilizado por los biólogos, sino un término común utilizado por matemáticos y científicos de la computación interesados en el tema, situación contrastante con el uso actual de esta disciplina altamente difundida en la biología. Pese a esto, es importante resaltar que a finales de los años 80 ya existían muchos laboratorios de biología molecular que habían iniciado los análisis de comparación de secuencias en micro-computadores en donde podían controlar de forma personal el flujo de análisis (Cannon, 1990). Por esa misma época se incrementó el volumen de datos producto de los resultados experimentales de los trabajos de los biólogos moleculares y se empezó a discutir sobre las necesidades de incremento de capacidad de computo para analizarlos dando inicio a una gran etapa que consistió en evaluar la necesidad de capacidades computacionales para el almacenamiento y procesamiento de información requeridos para el futuro desarrollo del proyecto genoma humano (Kelly, 1989). Entre 1988 y 1989 se había fundado el Centro HUGO (*The Human Genome Organisation*) por parte del Departamento de Energía (DOE) (quienes tuvieron la iniciativa del proyecto en 1986) y el Instituto Nacional de Salud de los Estados Unidos. Posteriormente en 1993, HUGO se transformó en el Instituto Nacional de Investigaciones del Genoma Humano: NHGRI (*National Human Genome Research Institute*), NHGRI, 2016. A finales de 1998 se publicó el primer repositorio de acceso público sobre el genoma humano en el *Genome Database* (GDB) (Letovsky *et al*., 1998), el cual posteriormente trasladó parte de su información al GenBank (Benson *et al.,* 2013) el cual es actualmente manejado por el *"National Center for Biotechnology Information"* (NCBI).

También surgió la iniciativa de publicar los avances del proyecto del genoma humano y se creó la primera serie en 1989 de las hojas informativas oficiales del DOE tituladas: "*Human Genome quarterly*" posteriormente conocida como Human genome news, que hasta el año 2002 publicó las principales noticias asociadas a los avances del proyecto genoma humano. En su primera entrega (Human Genome Quartely, 1989) se publicaron entre otros datos, los objetivos del grupo de trabajo computacional del genoma humano resumidos así: 1. Asesorar sobre los términos técnicos, necesidades, costos y requisitos computacionales. 2. Dar respuesta focalizada para responder a necesidades computacionales requeridas para acompañar los esfuerzos experimentales. 3. Mantener un foro para la discusión detallada sobre los avances e investigación requerida dentro de la comunidad del DOE. 4. Desarrollar protocolos para compartir datos en redes, y 5. Ofrecer un puente de comunicación oficial entre el sector privado y el DOE para negociar permisos y acuerdos de uso comercial derivados del proyecto genoma humano.

Adicionalmente se presentó el resumen general del primer gran taller realizado a finales de 1988 en Santa Fé, Nuevo México, cuyo objetivo fue reforzar la importancia de construir una interface entre ciencia computacional y metodologías de secuenciamiento de ácidos nucleicos. En dicho taller no sólo se identificaron las necesidades de programas de cómputo robustos, también se enfatizó en la necesidad no sólo de construir y extender la base teórica computacional para identificar la funcionalidad de las secuencias de ADN sino en la de desarrollar dispositivos de computo especializados para el análisis de secuencias (Human Genome Quarterly, 1989).

En palabras de DeLissi (2008) se podría resumir que en este taller, sin opacar por su puesto la importancia del taller de Santa Cruz en 1985 (Sinsheimer, 1985), se formalizó la delicada urgencia de evaluar si costos, complejidad técnica experimental y necesidades de métodos computacionales, podían en forma sincrónica e industrializada trabajar en conjunto para balancear la producción de información procesada y la producción masiva producto del secuenciamiento.

Entonces, en los inicios de los años 90 se da inicio al auge en la publicación de paquetes informáticos diseñados para incrementar la velocidad de cómputo para la comparación de secuencias de proteínas y de ADN (ya que los algoritmos diseñados por Needleman y Wunsch (1970) y Smith y Waterman (1981a; 1981b) tenían limitaciones de tiempo de ejecución) y aunque previamente existían heurísticas para la comparación de secuencias conocidas como la familia de algoritmos FASTA (Wilbur y Lipman, 1983; Lipman y Pearson, 1985), la técnica sofisticada de filtrado llamada BLAST por su sigla en inglés de *Basic Local Alignment Search Tool* (Altschul *et al.,* 1990), fue elegida para el portal del NCBI por su velocidad y mejor tratamiento estadístico y quizás hoy en día es uno los paquetes de computo ampliamente utilizados para búsqueda de similitud entre secuencias. BLAST sigue siendo actualmente una de las herramientas ampliamente difundida y usada como primera estrategia para identificar homología entre secuencias. Adicionalmente en estos años los primeros recursos sofisticados para la predicción de genes fueron publicados (Ouzounis y Valencia, 2003) por Brunak *et al*., (1990); Mural y Uberbacher, (1991); States y Botstein (1991); Fickett y Tung, (1992) y Guigo *et al*., (1992).

Sin embargo, adicional al problema de comparar y anotar los genes abordado por los pioneros de la bioinformática, es decir la construcción del andamiaje teórico requerido para resolver el problema de asignar sentido a la información secuenciada, un nuevo reto computacional surgió como requisito para poder ensamblar en cadenas largas de ADN, las piezas o lecturas (productos resultantes de las técnicas de secuenciamiento). Hasta el día de hoy las técnicas de secuenciamiento de genomas utilizan métodos que secuencian el ADN no de forma directa sino a partir

de pedazos obtenidos de su fragmentación al azar. No es posible con las plataformas disponibles en el mercado lograr en un único paso o reacción directamente una única lectura que corresponda a la cadena completa de un genoma. El reto computacional que se derivó entonces, fue resolver computacionalmente cómo los fragmentos deberían ser ensamblados o fusionados de forma correcta y consecutiva.

De forma muy general se podría entender hoy en día el proceso de ensamblaje de genomas como un proceso de concatenación de lecturas sobrelapadas de sus extremos iniciales y finales. Henson *et al*. (2012) exponen la idea que es importante buscar las coincidencias de los extremos de las lecturas como máxima para garantizar que los fragmentos ensamblados no sean producto de posibles coincidencias dadas por el azar. Sin embargo, se enfatiza en que los fragmentos ensamblados deben tener sentido en la región del genoma original. Los mismos autores entonces nos conducen a la consideración que si las lecturas concatenadas corresponden a regiones adyacentes en el genoma original entonces, este procedimiento se puede considerar como verdadero, pero si por el contrario si existen dos o más regiones en el genoma donde existiesen sitios potenciales de ubicar las lecturas, no siempre este proceso de concatenación concordaría con las regiones genómicas originales debido a la dificultad que plantearía la ubicación correcta de regiones duplicadas y en caso mas extremo al ensamblaje de regiones repetitivas (Henson *et al*., 2012).

El problema planteado anteriormente está vinculado con la complejidad misma de los genomas y el proceso de su dinámica evolutiva. Se puede pensar que ensamblar correctamente el genoma de una especie depende de su complejidad, la cual es el producto derivado del transcurso del tiempo y que conlleva a una estructura genómica compleja. Estas características se pueden observar cuando se analizan genomas actuales y se encuentran duplicaciones de fragmentos, rearreglos de regiones, entre otras características de alta complejidad. Entonces, la siguiente consideración de Henson *et al*., (2012) nos confirma que el proceso del ensamblaje del genoma es mucho más complicado de lo que podemos suponer, por ejemplo, si en un borrador de un genoma se tuviese que la secuencia X está concatenada como A, X y B y en otra región del mismo genoma ésta está concatenada como C, X y D no se podría descartar por ejemplo que la concatenación A, X y D también existiera debido a que no siempre el sobrelapamiento de los extremos puede resolver el problema de ensamblar las regiones repetitivas. Es por esta razón que los autores sugieren que los ensambles de los genomas deberían finalizar en los extremos de las vecindades donde residen repeticiones y utilizar otras aproximaciones para resolver el ensamblaje de estas regiones complejas.

Los autores de los primeros ensambladores de genomas de gran complejidad como el del humano o el de la mosca de la fruta tuvieron que enfrentarse al problema de diseñar las estrategias de cómputo para manejar la posible complejidad genómica. Los ensambladores genómicos fueron diseñados para lograr atacar al máximo los conflictos originados por múltiples sobrelapamientos entre lecturas. Dentro de los primeros ensambladores se tenían aquellos utilizados para la construcción de mapas físicos por medio de la identificación de huellas derivadas de los fragmentos de restricción en los clones para luego sobreponer aquellos con mayor similitud. Una explicación matemática de las aproximaciones de la época puede consultarse en Lander y Waterman (1988) y su extensión para el ensamblaje de clones utilizando anclas en Arratia *et al*., (1991).

EL PRIMER BORRADOR DEL GENOMA HUMANO

Después de las iniciativas lideradas por el DOE y otros investigadores, desde 1984 hasta 1986 (Sinsheimer, 1985; Palca, 1986), finalmente fue publicado en simultáneo en el 2001 el primer borrador del genoma humano y el primer mapa físico del genoma humano por *The International Human Genome Sequencing Consortium* (2001), (IHGSC, 2001). El consorcio internacional estaba formado por 20 grupos de diferentes países entre ellos Estados Unidos, Inglaterra, Japón, Francia, Alemania y China.

Desde 1997, parte de la estrategia mencionada anteriormente para secuenciar genomas, (es decir fragmentar el genoma humano en pedazos más pequeños para su posterior clonación), ya se había publicado. Este método y su estrategia se conoce en inglés como "*Human Whole-Genome Shotgun Sequencing*" (Weber y Myers, 1997). En esta misma publicación los autores plantearon que se requería gran cantidad de poder de computo para realizar el análisis del secuenciamiento. Adicionalmente indicaron que un *cluster* de computo conformado por estaciones de trabajo que procesaran un millón de instrucciones por segundo, permitiría obtener el genoma ensamblado en 300 días, tiempo que podría disminuirse, si los algoritmos utilizados eran más rápidos y por su puesto si la capacidad de computo se mejoraba.

El genoma humano fue inicialmente ensamblado utilizando el algoritmo implementado en el programa GigAssembler diseñado por Kent y Haussler (2001) y el algoritmo diseñado para el ensamblaje de clones implementado por Lander y Waterman (1988) y Arratia *et al*. (1991). Sin embargo, aunque se logró un primer borrador de ensamblaje usando estos métodos, los autores reconocieron las debilidades de ambos algoritmos para resolver problemas asociados al ensamblado de regiones genómicas de alta complejidad producto de duplicaciones y regiones repetitivas teloméricas y subteloméricas (IHGSC, 2001). Parte de los retos que se debían resolver posteriormente para completar en su totalidad la secuencia del genoma humano fue llenar los huecos o *gaps* que se formaron en el primer ensamble del genoma.

Un proceso adicional al problema del ensamblaje de un genoma es poder determinar la cobertura, es decir el

número de veces que una base ha sido secuenciada y que tiene sentido o que puede considerarse como la base o nucleótido que corresponde realmente a esa posición en el genoma original. En este caso, los autores utilizaron la herramienta BLAST para determinar el grado de cobertura del genoma humano. Para ello compararon las lecturas originales usadas para ensamblar el genoma humano y las secuencias disponibles en el GenBank.

Adicionalmente los datos disponibles de cDNAs de la base RefSeq (Pruit y Maglott, 2001) fueron alineadas al borrador del genoma humano encontrando que el 88 % de las bases de los cDNAs podían ser alineadas al genoma con un porcentaje de identidad del 98 % (IHGSC, 2001). De esta forma los autores reportaron que en este primer borrador el 88 % del genoma humano estaba representado, y que con la combinación de datos publicados de secuencias su estimativo incrementaba a un 94 %. Estos valores concordaron con los porcentajes para huecos o *gaps* en los cuales residiría la demás porción del genoma.

Otro de los objetivos de los gestores del proyecto era hacer visible la información obtenida del genoma humano a la humanidad. Una de las herramientas más interesantes diseñadas para visualizar con mucho detalle los genomas (aun disponibles hoy en día y de amplio uso en el campo de la bioinformática) son los navegadores genómicos conocidos en inglés como los *Genome Browsers*. En la misma publicación (IHGSC, 2001) los autores visualizaron la información del genoma humano en este tipo de plataformas que a la fecha han evolucionado a plataformas computacionales que usan sofisticados manejos de bases de datos de tipo relacional para la integración de la información genómica comparativa y evolutiva entre genomas de diferentes especies. Los dos grandes navegadores genómicos utilizados para observar con detalle el genoma humano fueron el UCSC Genome Browser y el EnsEMBL, en constante mantenimiento y curación por la Universidad de California Santa Cruz (Kent *et al*., 2002) y el Instituto de Bioinformática Europeo y el Centro Sanger (Hubbard *et al*., 2002) respectivamente.

Ahora bien, como se ha mencionado anteriormente, el problema de las repeticiones incrementaba la complejidad de ensamblar los genomas, entonces, se diseñó un programa que permite la predicción de elementos repetitivos como prueba en el ensamblaje llamado *Repeatmasker* (Smith *et al*., 1996), el cual se apoya en la base de datos de elementos repetitivos identificados experimentalmente principalmente y curada por el *Genetic Information Research Institute* (GIRI) (Jurka, 1998; Jurka, 2000; Jurka *et al*., 2005).

Además, se construyó el primer índice de genes utilizando los resultados de diferentes aproximaciones de cómputo que alinearon datos de ESTs, mRNAs, cDNAs a secuencias del ADN genómico ensamblado, así como de alineamientos de secuencias previamente reportadas de proteínas. (Birney *et al*., 1996; Gelfand *et al*., 1996; Mott, 1997; Bailey *et al*., 1998; Florea *et al*., 1998). Por último, los autores utilizaron modelos *ab initio* para la predicción de genes basados en los modelos ocultos de Markov (HMM) implementados en programas como GenScan (Burge y Karlin, 1997), Genie (Kulp *et al*., 1996; Reese *et al*., 2000) y Fgenes (Solovyev y Salamov, 1997). Generalmente estos programas utilizaron la información depositada de las bases de datos como EnsEMBL, RefSeq y PFAM, SWISSPROT y TrEMBL como parte del proceso para la construcción de información requerida.

En este primer borrador del genoma humano los autores propusieron construir un índice de genes y un índice de proteínas. El índice de genes construido indicó un total de 24500 genes para esta primera versión del genoma humano (IHGSC, 2001). Así mismo en esta misma época fue presentada la estrategia computacional utilizada para el ensamble del genoma de la mosca de la fruta y conocido como el ensamblador de Celera cuyo diseño algorítmico puede consultarse en Myers *et al*. (2000).

LAS NUEVAS TECNOLOGÍAS DE SECUENCIAMIENTO Y SU IMPACTO EN EL ESTUDIO DEL GENOMA HUMANO A GRAN ESCALA

El proyecto genoma humano ha sido comparado por algunos autores como un proyecto de tanto impacto y de retos para la humanidad como lo fue en su tiempo el proyecto de viajar a la luna. De acuerdo con el NHGRI, 2016 tras una inversión cercana a los 2.700 millones de dolares, los costos requeridos y la necesidad de infraestructura para amplificar a una escala mayor el secuenciamiento de más genomas, conllevó a enfrentar a los investigadores a nuevos retos computacionales y experimentales y a la idea de llevar a cabo un desarrollo automatizado masivo que permitiera identificar variantes genómicas derivadas del análisis genómico con posibles aplicaciones en la medicina personalizada.

Sin embargo, los costos invertidos en el secuenciamiento del primer borrador del genoma humano, plantearon la necesidad de utilizar métodos de secuenciamiento mucho más económicos y eficientes. Esto condujo al desarrollo de lo que se conoce hoy en día como tecnologías de secuenciamiento de nueva generación. Para una revisión más detalla de su desarrollo en comparación con procedimientos computacionales y otras aplicaciones se puede consultar a Zhang *et al*. (2011) y Henson *et al*. (2012), y Bermudez-Santana (2011) respectivamente. Los nuevos métodos de secuenciamiento incrementaron la cobertura, basaron sus métodos de amplificación en otras estrategias sin usar amplificación por clones biológicos y disminuyeron dramáticamente los costos de secuenciamiento.

Pero por otro lado, los tamaños de las lecturas producto de este nuevo secuenciamiento disminuyeron en comparación con las obtenidas por los métodos de secuenciamiento de Sanger (1975) y Sanger *et al*. (1977), es decir de una longitud cercana a los 1000 pares de bases máximas obtenidas por el secuenciamiento de Sanger, se

pasó a una longitud que variaba entre los 35 y 500 pares de bases dependiendo de la tecnología. Entonces ahora, los problemas de ensamblaje de lecturas se centraron en la búsqueda de sobrelapamiento de lecturas de menor tamaño y en la búsqueda de estrategias de almacenamiento y de procesamiento de información producto del alto volumen de datos de las nuevas metodologías de secuenciamiento que puede superar las gigas en cantidad de información para un experimento.

Dentro de las primeras tecnologías, la primera técnica de segunda generación introducida en el mercado en 2005 se conoció como 454 (Gilles *et al*., 2011) cuyo tamaño de lecturas era cercano a 600 pares de bases (actualmente esta tecnología se encuentra fuera de uso comercial). Con esta tecnología fue posible el secuenciamiento del segundo genoma humano perteneciente a James Watson (Wheeler *et al*., 2008 en Zhang *et al*., 2011). Posteriormente en el 2007 se lanzó al mercado Illumina con productos de tamaño de lecturas hoy en día cercano a los 100 pares de bases y cuya metodología fue utilizada para el resecuenciamiento del genoma humano en el 2008 (Bentley *et al*., 2008). Otro método comercializado por *Life Technologies* (CA, USA), SOLiD de *Life Technologies: Applied Biosys*tems ha sido también utilizado dentro de este marco de secuenciadores de segunda generación y por primera vez usado en el estudio del posicionamiento de nucleosomas (Valouev *et al*., 2008). Para una revisión del tema de secuenciadores de segunda generación consulte a (Metzker, 2010)**.**

Posteriormente surgieron los métodos de tercera generación que se basan en el secuenciamiento directo de ADN o conocido como secuenciamiento en tiempo real de cadena única (en inglés single-molecule real-time (SMRT) sequencing) (Benjamin *et al*., 2010) comercializado por *Pacific Biosciences*–y los métodos comercializados por *Life Technologies: Ion Torrent* (Rusk, 2011). Estas tecnologías pueden producir un tamaño final de lecturas en promedio cercano a 14000 pares de bases y 200 pares de bases, respectivamente. Los métodos de cuarta generación en proceso de desarrollo y comercialización se basan en la tecnología del Nanoporo que promete tener aplicaciones en la medicina personalizada, por ejemplo en Mikheyev (2014) puede consultarse una de sus aplicaciones en otras áreas de la ciencia.

Aunque las nuevas tecnologías incrementaron la eficiencia y disminuyeron los costos, por ejemplo del costo de secuenciamiento del primer genoma humano de 2.700 millones de dólares americanos se pasó a un costo cercano a los 5.000 dólares, los nuevos tamaños de lecturas no permitían continuar con las aproximaciones utilizadas de ensamblaje tradicionales. Henson *et al*. (2012) enfatizan que la extensión de concatenación de lecturas y el posible sobrelapamiento múltiple con más regiones se podría incrementar conllevando a una mayor cantidad de gaps en los genomas ensamblados. Problemas similares fueron resaltados por (Wold y Myers, 2008).

Para el caso particular de genoma humano, incluso extensible a genomas ya ensamblados, el proceso general reportado para el ensamblaje utilizando productos de secuenciamiento de última generación es en primer lugar el mapeo o alineamiento de lecturas al genoma humano de referencia (Zhang *et al*., 2011) o proceso conocido como mapeo para genomas re-sequenciados. Este tipo de pasos para ensamblar genomas utilizando genomas previamente secuenciados difiere en procedimiento de los pasos utilizados en el ensamblaje *de novo,* procedimiento que se fundamentó orginalmente en los principios de los grafos de Bruijn (Pevzner *et al*. 2011).

Para una mejor documentación sobre pasos a tener en cuenta en este tipo de ensamblaje se puede consultar a Baker (2012). A diferencia del ensamble de *de novo*, en el ensamble con genoma de referencia se debe mapear con alta confiabilidad las pequeñas lecturas al genoma de referencia, millones de lecturas cortas deben ser mapeadas al genoma de referencia que para el caso del humano, correspondería a las versiones secuenciadas del genoma humano.

Sin embargo para lograr con gran efectividad el mapeo, el genoma humano es comprimido en un conjunto de índices y se usan estructuras de datos conocidas como árboles de sufijos que facilitan no solo la compresión de cadenas de larga magnitud que corresponden a cada cromosoma sino el proceso de búsqueda eficiente de coincidencias de los millones de lecturas obtenidas de un experimento de secuenciamiento de última generación en el genoma de referencia, una explicación matemática detallada puede consultarse en Grossi y Vitter (2005).

Dentro de las diferentes aplicaciones de mapeo más utilizadas se encuentran MAQ (*Mapping and Assembly with Quality*) desarrollada por Li *et al*. (2008, 2009), Bowtie por Langmead *et al*. (2009), Segemehl por Hoffman *et al*. (2009) y SOAP por Li *et al*. (2008); entre otras aplicaciones disponibles que pueden ser consultadas en Zhang *et al*. (2011).

Por otro lado, Magi *et al*. (2015) indican que una vez el proceso de mapeado finaliza, una de las estrategias siguientes es identificar el conjunto de lecturas que representan estadísticamente variaciones con relación al genoma de referencia. Con estas estrategias es posible identificar lo que se conoce como variantes de nucleótido sencillas. Para ello se utilizan herramientas flexibles que almacenan de forma genérica los alineamientos resultantes de mapeo, estas herramientas pueden almacenar resultados de mapeo realizado por diferentes estrategias. Un ejemplo de éstas es SAM-Tools desarrollado por Li *et al*. (2009) que ha sido utilizada en el proyecto de secuenciamiento de los 1000 genomas humanos (The 1000 Genomes Project Consortium, 2010). Otra herramienta disponible es GATK desarrollada por McKenna *et al*. (2010). Para una revisión de métodos utilizados para la identificación de variantes se puede consultar a Medvedev *et al*. (2009) y Pirooznia *et al*. (2014).

FUTURO Y PERSPECTIVAS DE LA GENÓMICA PERSONALIZADA

Una de las preguntas derivadas del éxito del secuenciamiento del genoma humano podría plantearse en relación con su posible uso para el diagnóstico de enfermedades, es decir, ¿Tendrá la genómica un impacto en la práctica médica y por consiguiente, también en la salud humana? Con los cambios en los precios del secuenciamiento del genoma humano, que han disminuido dramáticamente en los últimos años (se ha llegado a pronosticar un valor de 1000 dólares americanos por genoma) la idea de poder vernos más en detalle no es tan lejana y cada vez está al alcance del soporte al diagnóstico médico. Pero a la vez, la posibilidad de conocer la secuencia del genoma de un paciente genera problemática desde el punto de vista ético y logístico, especialmente cuando se desee integrar la información genómica y su posible uso para el desarrollo de un fármaco o cuando se requiera planificar un esquema de combinación de fármacos dirigido a pacientes interesados en enfoques de la medicina personalizada. Pero por otro lado, aun no se ha evaluado las implicaciones que en salud pública tengan los efectos de preexistencias detectadas por el análisis genómico para alcanzar una trabajo o una afiliación a los sistemas de salud.

Offit (2011) enfatiza que en el caso de la medicina personalizada por ejemplo, la genómica personalizada construye los principios para la integración de la genética dentro de la práctica médica. Sin embargo, Offit (2011) también recuerda que mecanismos epigenéticos deberán ser incorporados al modelo genómico de estudio de la enfermedad humana, para lograr entender la enfermedad como un modelo genético multifactorial que involucra el ambiente asi como otros modificadores genéticos. Desde el punto de vista de la farmacología, la posible optimización de la terapia podría lograrse por el uso de información obtenida de los estudios genómicos, sin dejar atrás la importancia de evaluar la complejidad del actuar de muchos medicamentos.

Autores como Sadee y Dai (2005) mencionan que el éxito del descubrimiento de nuevos fármacos derivados de los estudios genómicos o de estudios farmacogenómicos deberían tener en cuenta los múltiples procesos que involucran la respuesta de los pacientes a fármacos y sus combinaciones, si es así, ahora la bioinformática tendrá nuevos retos para el manejo masivo de datos de las historias clínicas nutridas con la información genómica y requerirá el uso de la informática aplicada para conectar los posibles resultados obtenidos de estudios de polimorfismo de genes, del análisis cuantitativo de factores genéticos así como de la evaluación de los fenómenos epigenéticos. Todos estos en conjunto requerirán por supuesto estar conectados con las respuestas a nivel proteómico y metabolómico. Por tanto, el descubrimiento de nuevos fármacos asi como su posible uso comercial dependerá de la construcción de protocolos para la extrapolación de sus posibles beneficios asi como de la necesidad de construir un estructura jurídica que proteja la confidencialidad de los pacientes.

Guttmacher *et al*. (2010) enfatizan que para utilizar de forma apropiada y efectiva la información genómica derivada de un individuo se debe poseer de una infraestructura científica, logística y ética. La pregunta actual es si estamos preparados o si estamos construyendo una infraestructura para ello. En este sentido seis reglas principales se deben seguir para construir una agenda en salud pública que involucre la genómica como apoyo al diagnóstico médico que pueden ser consultadas en Burke *et al*. (2012). Sin embargo debido a la complejidad bio-sico-social de muchas enfermedades, las políticas de salud pública se verán enfrentadas con retos de alta complejidad como lo indica McBride *et al*. (2008).

Aunque se mantienen igualmente muchos debates éticos alrededor del tema de la genómica personalizada y su posible impacto en la medicina personalizada muchos investigadores, entre ellos Harol Elliot Varmus (premio Nobel de medicina y fisiología en 1989), indican que la genómica es tan sólo un modo de hacer ciencia y no medicina. Entonces de sus palabras podríamos pensar que a los protocolos existentes en medicina tan sólo se les debe incorporar la información genómica como una información complementaría similar a los resultados de laboratorio convencionales. Al respecto el debate continuará en los próximos años.

AGRADECIMIENTOS

El autor agradece a los organizadores y colaboradores de la Catedra José Celestino Mútis Semestre I -2015 por su arduo trabajo en el desarrollo exitoso de la Cátedra. Igualmente agradece al Profesor Peter F. Stadler de la Universidad de Leipzig por sus discusiones sobre el desarrollo de la bioinformática y su influencia en el surgimiento de la genómica. Al Profesor Humberto Arboleda del Instituto de Genética de la Universidad Nacional de Colombia, en el marco del proyecto Hermes 28240, por sus comentarios sobre la genómica personalizada en medicina y a los evaluadores anónimos del manuscrito quienes colaboraron en mejorar su contenido y forma. Al laboratorio de Biologia Computacional del Departamento de Biologia de la Universidad Nacional de Colombia dotado con equipos donados por el programa de equipamiento del DAAD a la Facultad de Ciencias.

REFERENCIAS

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-410. Doi:10.1016/S0022-2836(05)80360-2

Arratia R, Lander ES, Tavaré S, Waterman MS. Genomic mapping by anchoring random clones: A mathematical analysis. Genomics. 1991;11(4):806-827. Doi:10.1016/0888-7543(91)90004-X

Baer R, Bankier A, Biggin M, Deininger P, Farrell P, Gibson T, *et al*. DNA sequence and expression of the B95-8 Epstein—Barr virus genome. Nature. 1984;310:207-211. Doi:10.1038/310207a0

Bailey LCJr, Searls DB, Overton, GC. Analysis of EST-driven gene annotation in human genomic sequence. Genome Res. 1998;8:362-376. Doi:10.1101/gr.8.4.362

Baker M. De novo genome assembly: what every biologist should know. NatureMethods. 2012;9:333–337. Doi:10.1038/nmeth.1935

Benjamin AF, Dale W, Jessa L, Kevin T, Eric O, Tyson AC, *et al*., Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010; 7(6): 461–465.doi:10.1038/nmeth.1459

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218): 53–59. Doi:10.1038/nature07517

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36-42. Doi: 10.1093/nar/gks1195

Bermudez-Santana C. Buscando agujas en un pajar: viajes de RNAs pequeños *in silico* e *in vitro*. Acta biol Colomb. 2011;16(3):103-114.

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitte FI, Rindone W, *et al*. The GenBank genetic sequence data bank. Nucleic Acids Res. 1986;14(1):1-4. Doi:10.1093/nar/14.1.1

Birney E, Thompson JD, Gibson TJ. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res. 1996;24:2730-2739. Doi:10.1093/nar/24.14.2730

Brunak S, Engelbrecht J, Knudsen S. Neural network detects errors in the assignment of mRNA splice sites. Nucleic Acids Res. 1990;18:4797-4801. Doi:10.1093/nar/18.16.4797

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268:78-94: Doi:10.1006/jmbi.1997.0951

Burke W, Burton H, Hall AE, Karmali M, Khoury MJ, Knoppers B, *et al*. Extending the reach of public health genomics: What should be the agenda for public health in an era of genome-based and "personalized" medicine? Genet Med. 2010;12(12):785-791. Doi:10.1097/GIM.0b013e3182011222

Cannon G. Nucleic acid sequence analysis software for microcomputers. Anal Biochem. 1990;190(2):147-153. Doi:10.1016/0003-2697(90)90172-6

Chen R. Snyder M. Promise of personalized omics to precision medicine. Wiley Interdiscip Rev Syst Biol Med 2013;5(1):73-82. Doi: 10.1002/wsbm.1198.

Dayhoff MO, Schwartz RM, Orcutt, BC. A model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure. Vol. 5. suppl. 3. Dayhoff MO, editor. Washington, DC: Biomed Res Found; 1978. p. 345-352.

DeLissi C. Santa Fe 1986: Human genome baby-steps. Nature. 2008;455(16):876-878. Doi:10.1038/455876a

Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25:351-360. Doi:10.1007/BF02603120

Fickett JW, Tung CS. Assessment of protein coding measures. Nuc Acids Res. 1992;20: 6441-6450. Doi:10.1093/nar/20.24.6441

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995; 269(5223):496-512. Doi:10.1126/science.7542800

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. 1998;8(9):967-974.

Fraser C, Gocayne J, White O, Adams M, Clayton R, Fleischmann R, *et al*. The Minimal Gene Complement of Mycoplasma genitalium. Science. 1995;270(5235):397-404. Doi:10.1126/science.270.5235.397

Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced sequence alignment. Proc Natl Acad Sci USA. 1996;93:9061-9066. Doi:10.1073/pnas.93.17.9061

Gilles A, Meglécz E, Pech N. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011;12:245. Doi:10.1186/1471-2164-12-245

Gribskov M, McLachlan M, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA. 1987;84:4355–5358. Doi:10.1073/pnas.84.13.4355

Grossi R, Vitter JS. Compressed Suffix Arrays and Suffix Trees, with Applications to Text Indexing and String Matching. SIAM J Sci Comput. 2005;35(2):378-407. Doi:10.1137/S0097539702402354

Guigo R, Knudsen S, Drake N, Smith TF. Prediction of gene structure. J Mol Biol. 1992;226:141-157. Doi:10.1016/0022-2836(92)90130-C

Guttmacher AE, McGuire AL, Ponder B, Stefánsson K. Personalized genomic information: preparing for the future of genetic medicine. Nat Rev Genet. 2010;11:161-165. Doi:10.1038/nrg2735

Kelly MJ. Computers: the best friends a human genome ever had. Genome. 1989;31(2):1027-1033. Doi:10.1139/g89-177

Kent, WJ, Haussler D. GigAssembler: an algorithm for the initial assembly of the human working draft. Technical Report UCSC-CRL-00-17. Santa Cruz, California: University of California at Santa Cruz; 2001. p. 1-11.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, *et al*. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006. Doi:10.1101/gr.229102.

Kristofferson D. The BIONET electronic network. Nature. 1987;325:555-556. Doi:10.1038/325555a0

Kulp D, Haussler D, Reese MG, Eeckman FH. A generalized hidden Markov model for the recognition of human genes in DNA. ISMB. 1996;4:134-142.

Hamm GH, Cameron GN. The EMBL Data Library. Nucleic Acids Res. 1986;14:5–9. Doi:10.1093/nar/14.1.5

Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. Pharmacogenomics. 2012;13(8):901–915. Doi:10.2217/pgs.12.72

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, *et al*. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009:5(9):E1000502. Doi:10.1371/journal.pcbi.1000502

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L. *et al*. The Ensembl genome database project. Nucl Acids Res. 2002;30(1):38-41. Doi:10.1093/nar/30.1.38

Human Genome Quarterly. Oak Ridge National Laboratory Health and Safety Research Division Information Research and Analysis Section United States Department of Energy Office of Health and Environmental Research. ISSN: 1044-0828 1989.

IHGSC. The International Human Genome Sequencing Consortium, *et al*. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. Doi:10.1038/35057062

Jurka J. Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol. 1998;8:333-337. Doi:10.1016/S0959-440X(98)80067-5

Jurka J. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. 2000;9:418-420. Doi:10.1016/S0168-9525(00)02093-X

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogent Genome Res. 2005;110:462-467. Doi:10.1159/000084979

Lander ES, Waterman MS. Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis. Genomics. 1988;2(3):231–239. Doi:10.1016/0888-7542(88)90007-9

Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;10:R25. Doi:10.1186/gb-2009-10-3-r25

Lesk AM. The EMBL data library. In: Lesk AM, editor. Computational Molecular Biology. Sources and Methods for Sequence Analysis. Oxford: Oxford University Press; 1988. p. 55–65.

Letovsky SI, Cottingham RW, Porter CJ, Li PW. GDB: the Human Genome Database. http://www.gdb.org. Nucleic Acids Res. 1998;26(1):94-99. Doi:10.1093/nar/26.1.94

Li H, Ruan J. Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851-1858. Doi: 10.1101/gr.078212.108

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25 (14):1754-1760. Doi:10.1093/bioinformatics/btp32

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Doi:10.1093/bioinformatics/btp352

Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24:713-714. Doi:10.1093/bioinformatics/btn025

Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227:1435–1441. Doi:10.1126/science.2983426

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. 2010;20(9):1297-303. Doi:10.1101/gr.107524.110

Magi A, D'Aurizio R, Palombo F, Cifola I, Tattini L, Semeraro R, *et al*. Characterization and identification of hidden rare variants in the human genome. BMC Genomics. 2015;16(1):340. Doi:10.1186/s12864-015-1481-9

McBride C, Alford S, Reid R, Larson E, Baxevanis A, Brody L. Putting science over supposition in the arena of personalized genomics. Nat Genet. 2008;40(8):939-942. Doi:10.1038/ng0808-939

Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature methods. 2009;6(11):S13-S20. Doi:10.1038/nmeth.1374

Metzker ML. Sequencing technologies — the next generation. Nat Rev Gen. 2010;11:31-46. Doi:10.1038/nrg2626

Mott R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci. 1997;13:477-478. Doi:10.1093/bioinformatics/13.4.477

Mural RJ, Uberbacher EC. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA. 1991;88:11261–11265. Doi:10.1073/pnas.88.24.11261

Mikheyev AS, Mandy MYT. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Res. 2014;14(6):10971102. doi:10.1111/1755-0998.12324

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ. A Whole-Genome Assembly of *Drosophila*. Science. 2000;287:2196-2204. Doi:10.1126/science.287.5461.2196

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence

of two proteins. J Mol Biol. 1970;48:443-453. Doi:10.1016/0022-2836(70)90057-4

NHGRI. *National Human Genome Research Institute. [Sitio oficial]*. Última actualización 26 de Enero de 2016. [cited 5 February 2016]. Available on: http://www.genome.gov/

Offit K. Personalized medicine: new genomics, old lessons. Hum Genet. 2011;130:3–14. Doi:10.1007/s00439-011-1028-3

Ouzounis C, Valencia A. Early bioinformatics: the birth of a discipline a personal view. Bioinformatics. 2003;19(17):2176–2190. Doi:10.1093/bioinformatics/btg309

Palca J. Human genome-Department of Energy on the map. Nature. 1986;321:371. Doi:10.1038/321371a0

Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci. 2001;98:9748–9753. Doi:10.1073/pnas.171285098

Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, *et al*. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 2014;8:1-14. Doi:10.1186/1479-7364-8-14

Pruit KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 2001;29:137–140. Doi:10.1093/nar/29.1.137

Reese MG, Kulp D, Tammana H, Haussler D. Genie—gene finding in *Drosophila melanogaster*. Genome Res. 2000;10:529-538. Doi:10.1101/gr.10.4.529

Rusk N. Torrents of sequence. Nat Meth. 2011;8(1): 44-44. doi:10.1038/nmeth.f.330.

Sade W, Zunyan Dai Z. Pharmacogenetics/genomics and personalized medicine. Hum Mol Genet. 2005;14(2):R207-R214. Doi:10.1093/hmg/ddi261

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–448. Doi:10.1016/0022-2836(75)90213-2

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74(12):5463–5467. Doi:10.1073/pnas.74.12.5463

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, *et al*. Nucleotide sequence of bacteriophage ΦX174 DNA. Nature. 1977;265(5596):687-695. Doi:10.1038/265687a0

Sankoff D. Matching sequences under deletion-insertion constraints. Proc Natl Acad Sci USA. 1972;69(1):4-6. Doi:10.1073/pnas.69.1.4

Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, David N, *et al*. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014;133:1–9. Doi:10.1007/s00439-013-1358-4

Sinsheimer RL. The Santa Cruz Workshop—May 1985. Genomics. 1985;5:954–956. Doi:10.1016/0888-7543(89)90142-0

Smith TF, Waterman MS. Comparison of biosequences. Adv Appl Math. 1981a;2:482-489. Doi:10.1016/0196-8858(81)90046-4

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981b;147:195–197. Doi:10.1016/0022-2836(81)90087-5

Smith DH, Brutlag DL, Friedland P, Kedes LH. BIONET: a national computer resource for molecular biology. Nucleic Acids Res. 1986;14:17–20. Doi:10.1093/nar/14.1.17

Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996-2010. Available from: http://www.repeatmasker.org/

Solovyev V, Salamov A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc Int Conf Intell Syst Mol Biol.1997;5:294–302.

States DJ, Botstein D. Molecular sequence accuracy and the analysis of protein coding regions. Proc Natl Acad Sci USA. 1991;88(13):5518-5522. Doi:10.1073/pnas.88.13.5518

Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, *et al*. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. 2002;30(1):27-30. Doi:10.1093/nar/30.1.27

The International Human Genome Mapping Consortium. A physical map of the human genome. Nature. 2001;409:934-941. Doi:10.1093/nar/30.1.27

The 1000 Genomes Project Consortium. A map of human genome variation from population–scale sequencing. Nature. 2010;467(7319):1061-1073. Doi:10.1038/nature09534

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H *et al.,* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Research. 2008;18:1051–1063. Doi:10.1101/gr.076463.108

Weber JL, Myers EW. Human Whole-Genome Shotgun Sequencing. Genome Res. 1997;7:401-409. Doi:10.1101/gr.7.5.401

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, *et al*. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452:872–876. Doi:10.1038/nature06884

Wilbur WJ, Lipman DJ. Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci. USA. 1983;80:726-730. Doi:10.1073/pnas.80.3.726

Wold B, Myers RM. Sequence census methods for functional genomics. Nat Methods. 2008; 5:19-21. Doi:10.1038/nmeth1157

Zhang J, Chiodinic R, Badra A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011;38(3):95-109. Doi:10.1016/j.jgg.2011.02.003