ACERCA DEL GENOMA HUMANO

About human genome

Tobías Mojica. Ph.D.
Instituto de Genética
Universidad Nacional de Colombia

Luzardo Estrada. Ph.D.
Instituto de Genética
Universidad Nacional de Colombia

RESUMEN

En este artículo se revisan los eventos, alrededor del secuenciamiento del genoma humano, que han llevado a tanta excitación en los medios noticiosos y académicos en meses recientes. Se explican las estrategias que han llevado a que tengamos dos borradores diferentes pero complementarios, la estrategia llevada a cabo con el dinero de los contribuyentes que consiste en establecer el orden de fragmentos grandes de DNA antes de ser secuenciados y la estrategia llevada a cabo con dineros aportados por la industria privada, con la intención de explotar gananciosamente el conocimiento derivado del genoma humano. El genoma humano a mediados del año 2000 es un borrador incompleto que cubre alrededor del 85% de la secuencia con una precisión de un error en 1000 y el 99% de la secuencia con una precisión menor de 1 en 100 nucleótidos. También se discuten algunas de las posibles avenidas de explotación académica y económica. Nosotros pensamos que la secuencia del genoma ensanchará la brecha entre países avanzados y atrasados.

ABSTRACT

The sequence of the human genome, an undertaking of advanced countries, is nearly complete. In fact The Human Genome Project has around 85% of the genome sequenced 4 times on the average, with an accuracy of roughly 1 in 1000 nucleotides. Celera Genomics, on the other hand, has 99% of the sequence of one person, with an accuracy of slightly less than 1 in 100. The Human Genome project strives to produce a physical map for public consumption following a step by step strategy, in which the researcher sequences short DNA fragments belonging to larger fragments of known relative position. Celera Genomics wants to have very rapidly a physical map which can be quickly used to develop genetic tests and drugs, which can be later sold. We feel that the sequence of the human genome is something, which will widen the gap between advanced and backward countries.

¿QUE ES EL TAL GENOMA HUMANO?

Se trata de los más de 3.100 millones, nadie está seguro aún del número exacto, de bases; A, G, T, C, repetidas muchas veces y en muchos ordenes. Suficientes letras para llenar 250 directorios telefónicos de una ciudad mediana como Bogotá. Las letras representan los compuestos químicos que hacen nuestros genes y cuya secuencia, es decir, su orden lineal, tiene influencia sobre lo que una persona es, las maneras como estudia, camina, come, duerme, pinta, piensa, etc. El conocimiento de tal secuencia es el producto del esfuerzo de muchas personas. Los responsables más notables son dos grupos independientes que se han llamado el aporte público y el privado.

El esfuerzo público, dirigido por Francis Collins, se llama el proyecto del genoma Humano o HGP, y es un consorcio de varios entes internacionales, que incluye el Instituto para el Estudio del Genoma Humano (HGRI) de los Institutos Nacionales de Salud de los Estados Unidos (NIH), 4 centros de secuenciamiento en los Estados Unidos: El departamento de Energía de los Estados Unidos en Walnut Creek, estado de California; La Facultad de Medicina de Washington University en San Luis, estado de Missouri; El Instituto Whitehead para Investigaciones Biomédicas en Cambridge, estado de Massachussets, y la Facultad Baylor de Medicina en Houston, estado de Texas, además del Centro Sanger en Hinxton, cerca de Cambridge, Inglaterra, y laboratorios en Japón, Francia, Alemania y la China.

El trabajo ha involucrado a más de 1.100 científicos a lo largo de una década completa. El enfoque de Collins es lento y tedioso, pedazo a pedazo. Este enfoque empezó con células de la sangre y espermatozoides, luego los científicos dividieron las preparaciones en los 23 cromosomas humanos, y luego tomaron fragmentos largos y los cortaron en fragmentos más cortos que secuenciaron y ordenaron por los extremos superpuestos, y así sucesivamente pasando por genes, luego cromosomas y tarde o temprano todo el genoma. El estilo se puede comparar con la acción de sacar la página de un libro, romperla en pedacitos y volverla a unir utilizando las letras que se sobreponen. En el caso de la secuencia del genoma humano, la «página» es un fragmento de alrededor de 150.000 bases, clonado en un vector conocido como un BAC o cromosoma artificial de bacterias. Este fragmento es cortado en fragmentos más pequeños o subclonos con extremos que se superponen. Estos fragmentos son secuenciados y las secuencias son organizadas por un computador para producir secuencias cada vez más grandes llamadas «contigs». De ahí se pasa a otra página. Cada parte del genoma es secuenciado varias veces con el resultado de que entre más veces se aparece una base en la misma posición, mayor la certi-

dumbre de que la identificación que hace el computador es correcta. Todo el genoma es secuenciado varias veces, 4 veces sirven para tener una tasa de error menor de 1 en 100 y alrededor de 11 veces para llegar a una tasa de error de 1 en 100000 bases. A mitad del año 2000 el mapa consiste de BACS que cubren el 85% de las regiones del genoma que contienen genes. Cada BAC difiere de los otros en el grado de terminación de la secuencia y del ordenamiento de ésta. Sólo alrededor del 24% de la secuencia parece terminada en una forma precisa, otro 22% en forma casi acabada y el 38% está en borrador. Todo esto suma un 84%, el 16% restante está siendo secuenciado, excepto un 3% que no se deja clonar. Este genoma es un mosaico de 6-10 individuos.

El esfuerzo privado, es decir con capital privado y con el propósito de explotar comercialmente los resultados, se debe a unas pocas compañías con Celera Genomics de Rockville, Estado de Maryland a la cabeza. El cerebro detrás de Celera es J. Craig Venter. Celera Genomics tiene el enorme capital de 900 millones de dólares sólo para el año 2000. Celera tiene 300 secuenciadores hechos por PE Biosystems (la compañía dueña de Celera) y uno de los supercomputadores más poderosos del mundo. La compañía Celera Genomics utiliza un enfoque diferente del enfoque utilizado por Francis Collins, llamado de escopetazo, que le había dado a su presidente buenos resultados en el secuenciamiento de genomas más pequeños como los de bacterias. En lugar de arrancar una página cada vez, Celera rompe en millones de pedacitos todo el libro. Los pedacitos se superponen y luego se ensamblan por los extremos que se superponen, con la ayuda del supercomputador. El genoma humano se rompe en fragmentos de 2.000 bases, luego de 100.00 bases y luego de 50.000 bases. Se dice que Celera tiene alrededor del 99% del genoma humano secuenciado, pero la fuente de DNA es de un solo individuo y con cada base secuenciada tres o menos veces, es decir la confiabilidad está por debajo de uno en cien. Además de Celera Genomics, otras compañías notables involucradas en el secuenciamiento y explotación del genoma humano son las siguientes: Human Genome Sciences (*www.hgsi.com*) de Rockville, Estado de Maryland, bajo la dirección de William A. Haseltine, tiene 525 millones de dólares de capital para desarrollar y expandir la producción y mercadeo de drogas basadas en el genoma humano. Ya tiene varias drogas en investigación clínica. Incyte Genomics (*www.incyte.com*) de Palo Alto, Estado de California, dirigida por Roy A. Whitfield, suministra acceso a bases de datos genómicos y vende acceso a clonos de DNA representados en las bases de datos. Tiene 622 millones de dólares de capital para convertir la información genómica en un buen negocio. Millenium Pharmaceuticals (*www.mlnm.com*) de la ciudad de Cambridge, Estado de Massachussets, dirigida por Mark J. Levine, tiene como función el desarrollo de pruebas y terapias personalizadas. Tiene un capital de 700 millones de dólares para traducir la información genómica a productos patentables, incluyendo drogas y pruebas diagnósticas.

EL GENOMA HUMANO TIENE UNA HISTORIA RECIENTE

La historia del secuenciamiento del genoma humano es como sigue:

Marzo, 1986. El Departamento de Energía de los Estados Unidos, DoE, efectúa una reunión en Santa Fe para discutir los planes para secuenciar el genoma humano.

Abril, 1987. Empieza en DoE un programa del genoma a pequeña escala, DoE sugiere gastar 1.000 millones de dólares en 7 años.

Febrero, 1988. El consejo Nacional de Investigaciones de los Estados Unidos anuncia apoyo para el Proyecto del Genoma Humano (HGP).

Marzo, 1988. Primeros esfuerzos de secuenciamiento en los Institutos Nacionales de Salud de los Estados Unidos (NIH).

Abril, 1988. La oficina de Evaluación de Tecnología, del congreso de los Estados Unidos apoya el HGP.

Septiembre, 1988. Se establece la oficina del genoma humano en NIH, el profesor James Dewey Watson, codescubridor de la estructura de doble hélice del DNA en 1953, es nombrado director.

Octubre, 1989. Los NIH establecen el Centro Nacional para el Genoma Humano (NCHGR), con Watson a la cabeza.

Abril, 1990. Se publican los planes de los NIH y del DoE para secuenciamiento.

Julio, 1991. Craig Venter, revela que NIH ha solicitado patentes en fragmentos aislados de DNA extraídos del cerebro.

Abril, 1992. Watson se retira y se nombra a Francis Collins en su lugar.

Junio, 1992. Venter abandona NIH para fundar el Institute for Genomic Research, (TIGR), en Rockville, Estado de Maryland. La compañía farmacéutica SmithKline-Beecham financia la empresa con 125 millones de dólares, con la intención de desarrollar los descubrimientos comercialmente, a través de una compañía llamada Human Genome Sciences.

Julio, 1992. Wellcome Trust, de Inglaterra entra en la arena de los genomas con 50 millones de libras esterlinas, para apoyar, entre otros, el secuenciamiento del genoma del nemátodo *Caenorhaditis elegans*. El proyecto del Genoma Humano (HGP) es ya una tarea multinacional, de los países avanzados.

Octubre, 1993. NIH y DoE publican revisiones de sus planes de secuenciamiento. Se abre el Centro Sanger en Hixton, Inglaterra, financiado y operado por Wellcome Trust y el Consejo de Investigaciones Médicas del Reino Unido.

Septiembre, 1994. Investigadores Franceses y Norteamericanos publican un mapa de ligamiento genético completo.

Diciembre, 1995. Investigadores franceses y norteamericanos publican un mapa físico con 15000 marcadores moleculares.

Febrero, 1996. El acuerdo de Bermuda. Los diferentes socios se ponen de acuerdo para hacer públicas las secuencias en las siguientes 24 horas después de determinadas.

Abril, 1996. Se publica la secuencia completa del genoma de la levadura *Saccharomyces cerevisiae*.

Enero, 1997. NCHGR se convierte en Instituto y ahora se llama NIHGR, Instituto Nacional para la Investigación del Genoma Humano.

Junio, 1997. TIGR se separa de la empresa del genoma humano.

Mayo, 1998. Se forma la compañía Celera para secuenciar el genoma humano en «tres años», sin seguir el acuerdo de Bermuda. En respuesta Wellcome Trust duplica su apoyo al HGP.

Octubre, 1998. NIH y DoE prometen, para finales del 2003, un tercio del genoma secuenciado completamente y un borrador del resto del genoma. La secuencia completa para 2005.

Diciembre 1998. Se publica la secuencia completa del genoma de *C. Elegans*, un organismo multicelular.

Septiembre, 1999. Se inicia el secuenciamiento del genoma de ratón, por parte de los NIH.

Diciembre, 1999. Se publica la secuencia (casi) completa del primer cromosoma humano, el cromosoma 22.

Enero, 2000. Celera anuncia la compilación de secuencias que cubren el 90% del genoma humano.

Marzo 2000. Celera y otros investigadores académicos publican la secuencia «substancialmente» completa del genoma de la *Drosophila melanogaster* o mosca de la fruta.

Abril 2000. Celera anuncia la terminación de un borrador crudo de la secuencia del genoma de un individuo.

Mayo, 2000. Se publica la secuencia del cromosoma humano 21.

Junio, 2000. Celera y HGP anuncian mancomunadamente la terminación de un borrador de la secuencia del genoma humano.

El presidente Clinton saludó el borrador del genoma humano con palabras muy importantes, diciendo que el genoma humano es:

«The most wondrous map ever produces by humankind»

(El mapa más maravilloso producido por la especie humana).

«Epic-making triumph of science and reason»

(Un triunfo de la ciencia y la razón de proporciones épicas)

«A revolution in medical sciences whose implications will far surpass even the discovery of antibiotics»

(Una revolución en ciencias biomédicas cuyas implicaciones sobrepasarán aún el descubrimiento de los antibióticos)

Aún así, y al contrario de lo que mucha gente ha pensado, no tendremos un entendimiento acerca de nosotros mismos. Tal entendimiento está a décadas de nosotros. Pero los entendimientos que se empiezan a tener son importantes. Por ejemplo, compañías farmacéuticas intentan fabricar medicinas para entenderse con genes específicos en un conjunto de esfuerzos técnicos llamados colectivamente Farmacogenómica. Otras compañías diseñan pruebas de laboratorio que revelan mutaciones particulares, lo cual permitirá predecir si una persona va a tener, por ejemplo, Corea de Huntington, y muchos científicos y buena parte del público, vislumbran esperanzas de una terapia génica realista, la adición de genes buenos al cuerpo de un paciente. Los retos para las nuevas industrias son, sin embargo, formidables. Unos son técnicos, por ejemplo, una cosa es conocer la secuencia de un gene y otra es saber como funciona. Otros retos son legales; desde que nivel de conocimiento de un gene se puede patentar? Otros retos son de naturaleza social; por ejemplo el enfrentarse a la posibilidad de que a una persona se le diagnostique una enfermedad incurable.

Desde el punto de vista de la investigación, el siguiente paso, el paso realmente importante, es el establecimiento del significado de las secuencias, es decir la anotación del genoma humano. ¿Qué quiere decir cada nucleótido del genoma humano?. La gran cantidad de datos disponibles, no sólo del genoma humano sino de genomas de otros organismos, ha llevado a que ya se vea que la carrera tiene un blanco nuevo, diferente de las secuencias, el desarrollo de algoritmos mejores y mejores y de paquetes más y más fáciles de usar para el análisis del genoma humano y de los otros organismos involucrados en la carrera del secuenciamiento. Estos desarrollos representan la bioinformática. Otras compañías privadas, además de Celera, tales como Double Twist, Incyte, Compugen, etc, parecen tener más recursos económicos y más deseos de desarrollar software y hardware más y más avanzados.

Los cromosomas (casi) complemente secuenciados nos dicen muchas cosas

El cromosoma 21 es el autosoma más pequeño. Una copia extra produce el famoso síndrome de Down, la causa más frecuente del retardo mental humano. En éste cromosoma se han mapeado también algunos 20 loci anónimos que especifican desordenes monogénicos y predisposiciones a desordenes complejos. Además, la pérdida de heterozigocidad se ha informado asociada a la producción de algunos tumores sólidos. Un grupo de 63 autores (del Japón, de Alemania, Francia y Suiza) llamado el consorcio para mapear y secuenciar el cromosoma 21, informa la secuencia y el catálogo de genes del brazo largo del cromosoma 21. La secuencia presentada es de 33.546.361 pares de bases. El contig más largo es de 25.491.867 pares de bases. Sólo necesitan ser secuenciadas unas 100 kilobases del brazo largo comprendidas en 3 discontinuidades en los clonos y siete discontinuidades en la secuencia, lo cual da una cobertura del 99.7% del brazo largo. Adicionalmente se presenta la secuencia de 281.116 pares de bases del brazo corto. La secuencia permite identificar duplicaciones, probablemente involucradas en la generación de anomalías cromosómicas, y estructuras repetidas en las regiones pericentroméricas y teloméricas. La anotación analítica del cromosoma 21 reveló 127 genes conocidos, 98 que se predicen por motivos de secuencia y 59 pseudogenes.

Para dar una idea de la complejidad de la tarea se describe aquí la geografía del cromosoma 21. El cromosoma fue convertido en una sucesión continua de fragmentos grandes que incluyen 182 BAC (Cromosomas Artificiales de Bacterias) 111 PAC (Cromosomas Artificiales del Fago P1), 101 P1, 81 cósmidos, 33 fórmidos y 5 productos de PCR. Los autores utilizaron dos estrategias para generar la secuencia:

Se aislaron clonos a partir de las bibliotecas genómicas organizadas, por medio de hibridización no-isotópica y a gran escala; de este conjunto de experimentos se ensamblaron contigs y la localización de la mayoría de los clonos fue verificada por FISH. Las discontinuidades se llenaron por un método llamado multipoint clone walking.

Clones iniciales fueron escogidos usando marcadores STS seleccionados que fueron secuenciados o por lo menos parcialmente secuenciados. Luego los clonos iniciales fueron extendidos en ambas direcciones con nuevos clonos genómicos identificados o por PCR o por búsqueda de secuencias (*www.tigr.org*).

El mapa final consiste de 518 clonos bacterianos, ya descritos, formando 4 contigs grandes, con las discontinuidades ya mencionadas.

De los 98 genes putativos que se predicen por genómica, 13 son parecidos a genes de proteínas conocidas, 17 son marcos abiertos de lectura anónimos y 68 son unidades anónimas de transcripción.

El cromosoma 22 fue secuenciado por 115 autores del Centro Sanger, quienes informan la secuencia de la región eucromática del cromosoma 22. Esta secuencia consiste de 12 segmentos contiguos que ocupan 33 megabases que contienen por lo menos 545 genes y 134 pseudogenes. El cromosoma 22 es el segundo más pequeño de los cromosomas autosómicos (después del cromosoma 21) y comprende del 1.6% al 1.8% del DNA Genómico; es uno de los 5 cromosomas humanos acrocéntricos, cada uno de los cuales comparte una similaridad substancial en secuencia en el brazo corto, el cual contiene los genes repetidos del RNA ribosómico y una serie de otra secuencia repetida en tandem. No hay evidencia de secuencias codificadoras de proteínas (genes) en el brazo corto del cromosoma 22, mientras que evidencia directa e indirecta sugieren que el brazo largo es inusualmente rico en regiones que especifican proteínas. La alteración de la dosis génica de partes del brazo largo del cromosoma 22 es la etiología responsable de varias anomalías congénitas incluyendo el síndrome de ojo de gato (CES) y el síndrome velocardiofacial/DiGeorge (VCFS, DGS). Otras regiones asociadas con la enfermedad humana son el locus asociado a la esquizofrenia y las secuencias involucradas en ataxia espinocerebelar (SCA10).

Para identificar los clonos genómicos del cromosoma 22 se construyeron mapas de clonos utilizando cosmidos, fosmidos,

BAC, y PAC. Los clonos pertenecientes al cromosoma 22 se identificaron analizando bibliotecas en BAC y PAC, que representan más de 20 equivalentes genómicos, usando marcadores STS conocidos como pertenecientes al cromosoma 22, o utilizando bibliotecas en cosmidos o fosmidos derivadas del cromosoma 22 por flow sorting. Los contigs se ensamblaron basándose en mapas de restricción y STS, y se ordenaron unos en relación con los otros utilizando el mapa marco estándar. Los contigs resultantes se extendieron y se unieron por medio de ciclos repetidos de chromosome walking, utilizando secuencias de ambos extremos de los contigs. La secuencia completa cubre 33.4 Megabases de 22q con 11 discontinuidades, con una precisión calculada de menos de un error por cada 50000 bases.

La organización genómica de varias especies de mamíferos está conservada. La comparación de mapas, genéticos y físicos, entre especies, puede ayudar a predecir la localización de los genes en otras especies. Las relaciones mejor estudiadas son las que existen entre los hombres y los ratones. De los 160 genes identificados en el cromosoma 22 con ortólogos en el ratón, 113 son ortólogos murinos con localizaciones cromosómicas conocidas con mucha conservación en cuanto a la localización.

¿CUANTOS GENES EN EL GENOMA HUMANO?

De acuerdo a Pitágoras, los números limitan lo ilimitado y constituyen la verdadera naturaleza de las cosas; y rigen formas e ideas, y son la causa de dioses y demonios. Lo cual nos trae al enorme y antiguo problema acerca del número total de genes en el genoma humano. Nunca se ha sabido cuantos genes son especificados por el genoma humano y durante años se ha pensado que el número es de alrededor de 100000, hoy día muchos científicos se inclinan a pensar que el número debe estar más cerca de los 50000 que a los 100000. De todas maneras la primera tarea de anotación es realmente titánica. Cristóbal Colón nunca hubiese descubierto este continente americano si su tarea hubiese sido tan difícil. Con tanto del genoma humano ya secuenciado es posible hacer nuevos cálculos acerca del número de genes en nuestro genoma. Nuevos cálculos sugieren un límite inferior de 30000 y un límite superior de 120000 genes, la enorme dicotomía depende del tipo de análisis. Por ejemplo, Brent Ewing y Philip Green derivan cálculos de 34,700 y 33,630 multiplicando una muestra representativa de genes (como por ejemplo el número de genes identificados en el cromosoma 22) por el número de veces que otra muestra (como por ejemplo una colección de secuencias EST) puede ser dividida por el número de genes comunes entre las dos muestras. Jean Weissenbach y sus colaboradores también calculan un número bajo de 30000 comparando el genoma del pufferfish con lo que hay de la secuencia del genoma humano. John Quackenbush y sus colaboradores por otra parte parecen encontrar una verdadera mina de genes humanos y calcular que hay alrededor de 120000 genes, y lo hacen eliminando artefactos percibidos de las secuencias EST, como por ejemplo, EST solitarios y sin la cola de poliA, luego ensamblando las demás secuencias en contigs y comparándolos con colecciones anotadas de genes que especifican proteínas.

Los análisis de Green y de Quackenbush se basan en una comparación entre los genes anotados y el transcriptoma, el cual se asume que representa cabalmente al genoma. La identidad de los genes se establece por medio de una combinación de predicción ab initio y búsquedas de similaridad contra los contenidos de las bases de datos de DNA y de proteínas. Ambos medios de identificación tienen límites desconocidos y los mismo ocurre con la suposición popular de que EST representan genes transcritos, aún cuando están agrupados en una secuencia consenso. La transcripción ilegitima, como la que puede ocurrir en secuencias tales como Alu, pueden generar contigs consenso que son espurios. Muchos cálculos del número de genes humanos se basan en la extrapolación de un número pequeño sacado de un segmento pequeño del genoma, tomando diferentes cálculos de la densidad de los genes a lo largo del genoma y del tamaño real de éste. Que debe ser contado como un gene? Todo parece indicar que los cálculos del número de genes humanos son, la mayoría, muy grandes y que las bases de datos de EST pueden contener sólo el 40% de la fracción que codifica proteínas. Métodos anteriores estaban basados en enfoques indirectos como por ejemplo la medida de la complejidad celular por cinética de reasociación del RNA, determinación de islas CpG, reglas evolutivas o la suposición de que secuencias de DNA complementario representan genes.

Vemos una enorme operación y una operación llevada a cabo por los países avanzados, una operación multimillonaria con muchos resultados esperables y quizás buenos, pero seguramente un resultado inevitable será el aumento grande de la distancia entre los países de diferentes clases. Vemos una operación en la cual los países atrasados, por necesidad absoluta, son apenas espectadores, no podemos ser otra cosa, Vemos una operación en la cual han entrado los capitales grandes privados, indudablemente mucha gente piensa que se puede ganar mucho dinero de este conocimiento.

LITERATURA CITADA

ANONIMO. Rival genome sequencers celebrate a Milestone Together. Science 288: 2294

ANONIMO. World leaders heap praise on human genoma landmark. Nature 405: 983-985

BROWN, K. The human genome bussiness today Scientific American, p. 40-45. July 2000

HOWARD, H. The bioinformatics gold rush Scientific Am. p. 46-51. July 2000.

HATTORI, M. *et al.* (63 autores) The DNA sequence of human chromosome 21 Nature 405: 311-319. 2000

DUNHAM, I. *et al.* (115 autores) «The DNA sequence of human chromosome 22» Nature 402: 489-495. December 2, 1999.

SAMUEL, A. J. y R. APARICIO. How to count...human genes. Nature genetics 25:129-130. July 2000.