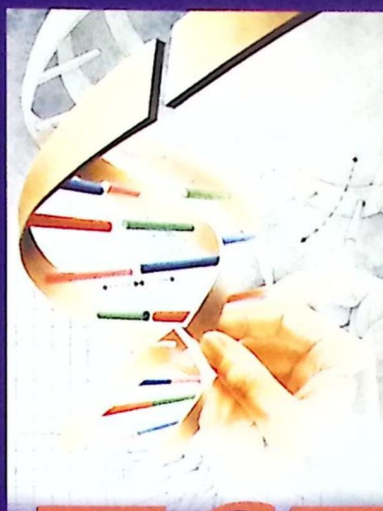




ORESTES

UNA NUEVA
TECNOLOGÍA
PARA EL
DESCUBRIMIENTO
DE GENES

ORESTES

UNA NUEVA TECNOLOGÍA PARA EL DESCUBRIMIENTO DE GENES

EMMANUEL DIAS NETO y ANDREW J. G. SIMPSON

III Premio Alcatel de Tecnología

OCTUBRE 2000

Instituto Ludwig
de Pesquisas
sobre o Câncer

Fundação de Amparo
à Pesquisa do
Estado de São Paulo

Lombardi 0310
610
DS418

LA TECNOLOGÍA ORESTES

BREVE PRESENTACIÓN

El 26 de marzo de 1999 se lanzó, en el auditorio de la *Fundação de Amparo à Pesquisa do Estado de São Paulo-FAPESP*, el Proyecto Genoma Humano del Cáncer, una iniciativa que es el resultado de la asociación entre dicha fundación, una de las más importantes agencias brasileñas de incentivo a la investigación, y el Instituto Ludwig de Investigaciones sobre el Cáncer, una respetada institución científica internacional con sede central en Zúrich, en Suiza, y con una de sus 10 unidades instalada en San Pablo. La meta del proyecto consiste en determinar, en el plazo de dos años, valiéndose de una nueva y revolucionaria tecnología desarrollada por dos investigadores del Instituto Ludwig de San Pablo, y con apoyo de la FAPESP, 700 mil secuencias de genes expresados en tumores malignos de gran incidencia en Brasil. Alcanzada dicha meta, los investigadores involucrados en el proyecto habrán dado, con seguridad, una contribución fundamental para ampliar el conocimiento de la parte del genoma humano que realmente tiene una importancia central, es decir, aproximadamente un 3% del ADN humano que codifica nuestros genes o que son propiamente los genes. Habrán contribuido también para una mayor comprensión acerca de las diferencias genéticas entre células normales y células malignas y, por lo tanto, habrán ayudado a allanar el camino para el desarrollo de nuevos métodos de diagnóstico precoz del cáncer y, más adelante, de nuevas terapias contra la enfermedad.

Palabras vibrantes marcaron la ceremonia de lanzamiento del Proyecto Genoma Humano del Cáncer. Palabras que ni por casualidad eran sinónimo de halagos gratuitos o vacíos. En lugar de ello, reflejaban la aguda consciencia de que se estaba frente a un hecho de indiscutible importancia para la ciencia de Brasil y, de una manera más amplia, para la ciencia contemporánea, con profundas implicaciones sociales y económicas, e incluso con repercusiones sobre la vida humana en su carácter individual.

Todo ello se debe a que los responsables por el proyecto, aquellos que lo formularon y los que decidieron financiarlo, estaban convencidos de que el mismo le permitiría a Brasil formar efectivamente parte del gran esfuerzo internacional, liderado por Estados Unidos e Inglaterra, para develar el genoma humano. Y lo permitiría precisamente porque el proyecto disponía de una carta de triunfo muy especial en su poder, la nueva tecnología de secuenciamiento elaborada en San Pablo y denomi-

nada ORESTES, sigla adoptada para *Open Reading Frames Expressed Sequences Tags*.

Debe subrayarse que el develar el genoma humano, misión del Proyecto Genoma Humano en sus vertientes pública y privada es, sin lugar a dudas, uno de los mayores y más importantes emprendimientos científicos en curso en este pasaje de milenio, en razón de su potencial para alterar radicalmente las bases de la biología y las diversas tecnologías relacionadas a ella, y en función también de sus inmensos efectos posibles sobre la salud e, incluso, sobre la duración normal de la vida humana.



MIGUEL BOYKIN

Fiesta de apertura: la tecnología ORESTES hizo posible el Proyecto Genoma Humano del Cáncer, lanzado en marzo de 1999. En la mesa (de derecha a izquierda), el vicegobernador del estado de San Pablo, Geraldo Alckmin, el presidente de la FAPESP, Carlos Henrique de Brito Cruz, el secretario de Ciencia, Tecnología y Desarrollo Económico del Estado de San Pablo, José Aníbal Pontes, el presidente del Instituto Ludwig, Edward McDermott, y el director presidente de la FAPESP, Francisco Romeu Landi

Estas percepciones estaban todas en escena en el momento en que el Proyecto Genoma Humano del Cáncer, impulsado por la tecnología ORESTES, estaba siendo lanzado (como parte del Programa Genoma de la FAPESP, iniciado por esta Fundación en 1997, bajo la conducción de su director científico, José Fernando Perez). Pasó desde entonces apenas un año y medio, y la tecnología no sólo tiene un indiscutible reconocimiento internacional en medios científicos y en otros más amplios, sino que llevó al Genoma Humano del Cáncer a duplicar sus metas cuantitativas iniciales de secuenciamiento. Y hay más: la excelencia de la tecnología se convirtió en un factor de atracción para poderosos socios interesados en trabajar en conjunto con el grupo de investigadores brasileños responsable por el proyecto, cosa que impulsa la ampliación de los propios objetivos científicos del emprendimiento.

Tanto es así que se formalizará, en una reunión que se realizará aún en octubre en Ginebra, Suiza, una asociación propuesta por el Instituto Nacional del Cáncer (NCI), una de las más respetadas instituciones científicas de Estados Unidos, a la FAPESP y al Instituto Ludwig. A partir de ese hecho, los grupos de investigación de Brasil y EE.UU. comenzarán a secuenciar, con base en las informaciones ya obtenidas por el proyecto, genes enteros, además de fragmentos de genes. Se empieza así a trabajar no sólo con genes expresados en células cancerosas, sino también con cualquier gen del genoma humano. Eso lleva a Robert Strausberg, coordinador de la investigación del NCI, y a Andrew Simpson, coordinador del Proyecto Genoma Humano del Cáncer, a prever que ambos grupos, con sus técnicas diferentes y complementarias, serán los mayores responsables por el secuenciamiento de genes del genoma humano en el mundo.



Un crecimiento sorprendente: en julio de 2000, Mário Covas, gobernador del estado de San Pablo (*en el centro*), anunciaba la duplicación de las metas del proyecto; en la mesa (*de izquierda a derecha*) estaban el director científico y el presidente de la FAPESP, José Fernando Perez y Brito Cruz, el ministro jefe de la Secretaría de Comunicación del gobierno brasileño, Andrea Matarazzo, el secretario José Aníbal Pontes, el director del Instituto del Cáncer de San Pablo, Ricardo Brentani, y el coordinador del Proyecto Genoma Humano del Cáncer, Andrew Simpson

Originalidad y otras características singulares de ORESTES

La tecnología ORESTES, descrita en todos sus detalles en el artículo técnico de sus creadores, Emmanuel Dias Neto y Andrew J. G. Simpson a partir de la página 7, es una metodología para el descubrimiento de genes. Y es original, y muy diferente de la técnica usada, por ejemplo, por el NCI, porque mientras esta última se dedica al secuenciamiento en las extremidades de los genes, la tecnología brasileña permite secuenciar la región central de los mismos, en la cual se concentran las informaciones más relevantes para que los mismos desempeñen sus funciones. Es allí precisamente en donde se sitúa la región codificadora de los genes. Ese dato permite que se tenga una idea inicial de la competencia científica y de la capacidad creativa que la técnica requirió por parte sus autores.

Si a eso se le suma el blanco primordial de aplicación de la tecnología ORESTES, es decir, los tumores malignos, en particular los de alta incidencia en Brasil y en América Latina, en cuyos tejidos genes normalmente raros del genoma humano se expresan con mayor abundancia, se puede tener también una primera idea acerca del impacto positivo que el Proyecto Genoma Humano del Cáncer promete en las batallas que la ciencia libra contra el cáncer. Un impacto que a corto plazo incidirá más sobre el diagnóstico precoz del cáncer, en los estadios en los cuales la afección tiene cura.

Vale la pena resumir algunos aspectos básicos para la apreciación de la tecnología ORESTES en esta breve presentación, aun considerando que los mismos son mucho mejor descritos en el artículo de sus autores.

De este modo, en lo que se refiere al *carácter innovador* de la tecnología, es necesario insistir en el hecho de que ORESTES permite determinar la información más valiosa contenida en los genes, auxiliando sustancialmente en la determina-

ción y caracterización de todos los genes humanos.

Con relación a su *aplicación*, cabe registrar que la obtención de fragmentos de genes expresados (denominados ESTs, del inglés *Expressed Sequences Tags*) es una herramienta de fundamental importancia para la comprensión de los genomas de organismos complejos como el hombre. Después de haberse descifrado el genoma completo, hecho notable recientemente concretado por el Proyecto Genoma Humano, es necesario identificar y determinar la función de todos los

genes activos. La tecnología ORESTES permite alcanzar esos objetivos gracias a sus características singulares. Entre ellas se encuentra la de privilegiar las secuencias derivadas de genes raros del genoma humano, difíciles de obtenerse a través de procesos tradicionales, además de favorecer la obtención de la porción central de los genes expresados, que según dijimos, es la región más rica en informaciones.

Con relación al *valor y al contenido tecnológico de la aplicación*, cabe registrar que ORESTES ofrece una ganancia competitiva, puesto que permite alcanzar a gran escala la porción más informativa de los genes de cualquier organismo, siendo al mismo tiempo una técnica simple, barata y que permite una rápida implementación en los laboratorios de secuenciamiento.

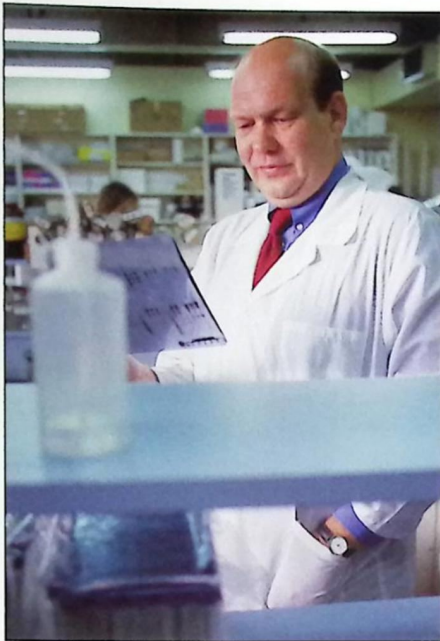
Con referencia a *la escala actual de aplicación de la tecnología y al potencial futuro de aplicación*, observamos que ORESTES es la base del Proyecto Genoma Humano del Cáncer, que involucra a unos 150 investigadores del estado de San Pablo, con una inversión a fondo perdido de 20 millones de dólares. Probablemente el Genoma Humano del Cáncer sea el mayor proyecto brasileño en términos de financiamiento. Su producción de datos avanza a ritmo extremadamente acelerado: ya se han producido más de 700 mil secuencias de ADN, lo cual lo transforma en el segundo proyecto del mundo en importancia en el campo de la investigación genómica. El potencial de aplicación de ORESTES, que empieza a motivar asociaciones internacionales, como la ya mencionada con el NCI, es prácticamente ilimitado, pues la técnica puede emplearse en el genoma de cualquier organismo complejo. La misma podrá usarse como base en proyectos como el del genoma del arroz, que se encuentra en fase de negociación por parte del Ministerio de Agricultura del Brasil y la Empresa Brasileña de Investigación Agropecuaria, (EMBRAPA), los del genoma del hongo *Paracoccidioides brasiliensis* y del parásito causante de la esquistosomo-



MIGUEL BOYAN

Asociación con el NCI: Victor Jongeneel, coordinador de bioinformática del Ludwig en Lausanne (a la izquierda), Robert Strausberg, asistente de dirección del Instituto Nacional del Cáncer-NCI de Estados Unidos y Simpson y Kenneth H. Buetow, de la División de Epidemiología y Genética del Cáncer del NCI, acuerdan los términos de la cooperación entre los grupos de investigación de Brasil y EE.UU.

MIGUEL BOYAN



MIGUEL BOYAN



Creadores exitosos: los autores de la tecnología ORESTES, Andrew Simpson y Emmanuel Dias Neto, trabajando en sus laboratorios en el Ludwig en San Pablo

sis, estos dos últimos con posible financiamiento de la FAPESP.

Con relación a los *impactos de la innovación con miras a la modernización, la eficacia y el aumento de la productividad para el emprendimiento en el cual está siendo aplicada*, la existencia de ORESTES permitió que Brasil entrara en el selecto grupo de países que contribuyen a gran escala para la definición de los genes humanos. El Genoma Humano del Cáncer ya fue capaz de atraer inversiones sustanciales que permitieron equipar de-

cenas de laboratorios de investigación en varias universidades. Siendo un proyecto brasileño, tiene un foco importante en la generación de datos centrados en tumores de incidencia significativa a nivel nacional, que no están siendo estudiados en los países más desarrollados, como los tumores de estómago y de cabeza y cuello. La asimilación de estos hallazgos científicos deberá traducirse en la modernización de los métodos de diagnóstico precoz y tratamiento de neoplasias humanas. Se espera alcanzar también métodos de diagnóstico más eficaces en términos de sensibilidad, carga informativa y especificidad con relación a los diversos tipos de cáncer. A mediano plazo, esos beneficios se extenderán al desarrollo de métodos más adecuados de tratamiento, con menor costo, mayor eficacia y menos efectos colaterales. Los datos generados por el Proyecto Genoma Humano del Cáncer son puestos a disposición, de manera permanente, de toda la comunidad científica internacional, más allá de los lindes de los laboratorios involucrados en el trabajo, del estado de San Pablo y de Brasil.

Cabe subrayar también a la *utilización de tecnologías de la información* impulsada por la tecnología ORESTES. A decir verdad, el intercambio de información tiene un papel fundamental en todo el trabajo de investigación generado por la técnica. El método ORESTES permite la obtención de secuencias de ADN extremadamente ricas en información. Todos los datos generados por los grupos de secuenciamiento, distribuidos en los 33 laboratorios de investigación en las ciudades del estado de San Pablo, son transmitidos vía INTERNET a la coordinación de bioinformática. Esos datos son comparados con los de los bancos de datos internacionales para la caracterización de las secuencias. El correcto y cuidadoso análisis de dichos datos permite que los mismos se transformen en información sumamente valiosa, que permite su aplicación a corto plazo. Las tecnologías de la información facilitan la interconexión de todos los grupos asociados y, lo más importante, permiten que cualquier investigador, de cualquier país del mundo, tenga libre acceso a las informaciones generadas en el seno del Proyecto Ge-

noma Humano del Cáncer. La libertad de intercambio de información garantiza una rápida utilización del conocimiento, lo que hace factible el retorno de los beneficios a la comunidad de la manera más rápida posible.

En lo que se refiere a su *carácter ejemplar y a los efectos emuladores e inductores de la aplicación de la técnica ORESTES*, su implementación ha surtido efectos que superan todas las expectativas de sus autores. El Proyecto Genoma Humano del Cáncer le permitió a Brasil producir hasta ahora el 13,2% de todas las ESTs humanas disponibles en bancos de datos públicos. Como fue dicho, el mismo es el segundo proyecto mundial en importancia, sólo atrás en número de datos del CGAP (*Cancer Genome Anatomy Project*) iniciado como mínimo tres años antes. Cuando los tumores humanos son analizados en detalle, se verifica que el proyecto generó 64 mil datos derivados de tumores de mama (el doble que las informaciones generadas por el resto del mundo junto). También para tumores de estómago se generó el doble de datos que los que todo el mundo generó, y en relación con los tumores de cabeza y cuello, el proyecto brasileño generó 4 veces más datos. La disponibilidad de una tecnología alternativa atrajo una inversión de 20 millones de dólares al Genoma del Cáncer, convirtiéndolo en -más allá del ser el mayor proyecto de investigación nacional- uno de los proyectos que disponen de mayor cantidad de recursos provenientes de la iniciativa privada.

Con relación a los *equipamientos utilizados en el proceso de creación y aplicación de la tecnología ORESTES y al grado de cooperación de instituciones y agentes externos*, es necesario destacar -en la creación- el uso termocicladores, sistemas de electroforesis y sistemas de análisis de datos del tipo WorkStations SUN e IBM,

además de sistemas de comunicación por líneas dedicadas y de alta velocidad. En la aplicación de tecnología, se utilizaron secuenciadores de ADN de última generación, basados en la tecnología capilar. Al respecto de la cooperación externa, se obtuvo un apoyo financiero de 10 millones de dólares del Instituto Ludwig de Investigaciones sobre el Cáncer. Y en el ámbito científico específicamente, el proyecto contó con la cooperación externa bajo la forma de la valiosa participación de un comité de evaluación externo, integrado por las siguientes personalidades: Dr. Phillip Sharp (premio Nobel de Medicina por el descubrimiento de los intrones), Dr. John Sgouros (ex coordinador de Bioinformática del Proyecto Genoma de Levadura, Inglaterra), Dr. Marcelo Bento Soares (Universidad de Iowa y Proyecto Genoma Humano, EE.UU.), Dr. Victor Jongeneel (Office for Information and Technology, Suiza).

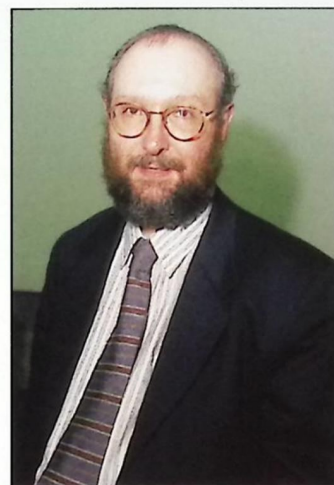
Con referencia a la consolidación de la tecnología Orestes a través de la *patente de innovación y su registro referente a la propiedad industrial*, se ha formulado un pedido de patente internacional ante la *World Intellectual Property Organization* #WO 00/31299 (02.06.2000). International Application Number: PCT/US99/27430



MIGUEL BOYAN



EDUARDO CESAR



EDUARDO CESAR

Evaluación externa: los científicos (*en sentido horario*) Phillip Sharp, premio Nobel de Medicina por el descubrimiento de los intrones, Marcelo Bento Soares, de la Universidad de Iowa y miembro del Proyecto Genoma Humano de Estados Unidos, y John Sgouros, ex coordinador de bioinformática del Proyecto Genoma de la Levadura en Inglaterra, examinaron y le dieron su aval a la tecnología Orestes y al Proyecto Genoma Humano do Cáncer

ORESTES

UNA TECNOLOGÍA BRASILEÑA PARA EL DESCUBRIMIENTO DE GENES

EMMANUEL DIAS NETO y ANDREW J. G. SIMPSON

ADN, diversidad y evolución

Toda la rica diversidad de los 6 mil millones de habitantes de nuestro planeta, con su carga de contrastes y toda la diversidad de cuerpos y mentes, tiene como base las informaciones contenidas en el núcleo de cada una de nuestras células. La molécula clave de toda esta información es el ADN. En última instancia, toda la sutileza de nuestra especie, nuestro arte, nuestras conquistas científicas, nuestras enfermedades y un gran número de factores que influyen en nuestro comportamiento son determinados por un grupo relativamente pequeño de instrucciones contenidas en nuestro ADN, en nuestros cerca de 80 mil genes. Con seguridad no es una exageración decir que el descubrimiento de los genes humanos echará luz sobre muchos de los misterios de nuestra vida y de nuestra biología.

El genoma humano está representado por la totalidad del material genético contenido en una célula. La fuente de nuestra identidad, nuestro genoma completo, se preserva en cada una de nuestros billones de células. En un nivel más básico, nuestro genoma es nuestro ADN, el ácido desoxirribonucleico, un polímero natural compuesto por tan sólo cuatro tipos de nucleótidos o bases, repetidos miles de millones de veces. En la famosa doble hélice de ADN, estas bases forman pares de una manera fija y complementaria. Los pares formados son A/T, T/A, C/G y G/C. La secuencia y ordenamiento de estas bases forma la estructura primaria del ADN. En los cromosomas, las hélices de ADN se enroscan y empaquetan formando estructuras más complejas con un alto nivel de organización.

Esta estructura organizacional se mantiene con pequeñas alteraciones en todas las especies vivas, y nuestro genoma actual es fruto de millones de años de evolución. Durante esos millones de años, el material genético fue experimentando continuas modificaciones seleccionadas por el ambiente, que permitieron el surgimiento de todas las especies, incluyendo al homo sapiens. Esas alteraciones fueron siendo incorporadas a nuestro ADN a través de diversos mecanismos. Tramos de DNA se insertaron a través de virus, se produjeron mutaciones después del contacto con sustancias químicas o rayos ultravioleta, se incorporaron errores acaecidos durante procesos de división celular, algunos genes se desligaron o se perdieron y todas las alteraciones que redundaron en alguna ganancia evolutiva para un individuo se fijaron en su genoma. Como resultado de todo esto, nuestra especie

tiene hoy un genoma de 3.120 millones de bases. En ese vasto contenido del genoma están desparramados entre 50 y 80 mil genes humanos. Los genes se encuentran distribuidos a lo largo de todo el genoma y su contenido informacional se encuentra dispersos de un modo discontinuo como si fueran palabras separadas en sílabas. En la biología molecular, cada "sílabas" es denominada exón, separada de otra "sílabas" por una región llamada intrón. Esta organización está representada en la figura 1. Si sumamos cada una de las "sílabas" de todos los genes totalizamos tan sólo cerca de un 3% del contenido del genoma. Solamente este pequeño porcentaje está directamente comprometido con el funcionamiento de la maquinaria celular para la construcción y el mantenimiento funcional de un ser humano. Las demás regiones del genoma (el 97%) se constituyen de ADN no informativo.

El Proyecto Genoma Humano

Para que podamos comprender amplia y totalmente la estructura completa de funcionamiento de nuestro organismo, con todas sus características fisiológicas, estructurales y funcionales, es absolutamente fundamental que tengamos un total conocimiento de las informaciones contenidas en nuestros genes. Este conocimiento se adquiere a través de la lectura de las bases nucleotídicas que componen los genes (secuenciamiento) y la determinación de las regiones cromosómicas en las cuales dichos genes se encuentran (mapeamiento). El secuenciamiento y el mapeamiento del ADN constituyen la más importante actividad del Proyecto Genoma Humano. El desciframiento de toda la información genética de nuestro genoma fue un desafío que comenzó a planearse aún du-

rante en los años 80. Las limitaciones técnicas fueron siendo gradualmente superadas y la entrada de empresas privadas en el sector llevó a un aumento de la competitividad y a una impresionante reducción del período inicialmente estimado para el término de esta etapa de descodificación.

De este modo, el día 26 de junio de 2000 la empresa privada Celera Genomics y la iniciativa pública del Proyecto Genoma Humano (financiado fundamentalmente por el Department of Energy, el National Institutes of Health y el The Wellcome Trust) anunciaron conjuntamente el término del primer montaje del genoma humano, hecho que fue anunciado por el presidente de Estados Unidos, Bill Clinton, y rápidamente difundido en toda la prensa mundial. Ciertamente, nuestra generación será mencionada por las generaciones venideras como la responsable por la decodificación del genoma humano. Una conquista única que influirá en todos los aspectos de la vida humana en las próximas generaciones.



Figura 1

Estructura básica de la organización de un gen de organismos complejos. Las regiones espaciadoras e intrones son representadas en azul, mientras que los exons o regiones transcritas de un gen son representados en rojo. La yunción de estos exones, transformados en moléculas de mARN, forma los genes propiamente dichos, que equivalen a tan sólo el 3% de nuestro genoma. La porción interna de estos mARNs contiene la información central para la síntesis proteica. En los proyectos genoma, los mARNs son secuenciados parcialmente, generando las ESTs. Se muestra la distribución posicional de estas ESTs generadas con ORESTES y con técnicas tradicionales.

Los desafíos de la era post-genoma

Tras el primer montaje del genoma humano, el proyecto avanza hacia su segunda fase: la interpretación del genoma codificado. El montaje y la lectura de toda la secuencia de bases del genoma humano equivalen a la realización de una gran sopa de letras con 3.120 millones de letras. En la próxima fase, el primer desafío será el de encontrar todos los genes humanos en esta inmensa sopa. El desafío no es sencillo, pues los genes normalmente se encuentran divididos en pedazos (exones) equivalentes a las sílabas de una palabra. Los análisis por computadora se mostraron limitados en el proceso de identificación de estas "sílabas" siendo de ese modo incapaces de identificar correctamente a todos los genes. Actualmente, tan sólo poco más de 10 mil genes humanos son completamente conocidos y las estimaciones del número total de genes son extremadamente variadas, oscilando entre 35 mil y 120 mil (Liang *et al.*, 2000; Ewing & Green, 2000; Crollius *et al.*, 2000).

De esta manera, para ser capaces de determinar cuáles son e identificar correctamente cada uno de los genes humanos, debemos procurar alternativas que ofrezcan una posibilidad real y segura. Para ello, nada mejor que valernos de los procesos celulares que naturalmente reflejan los genes bajo cada una de sus formas alternativas. El 3% de nuestro genoma, que contiene todos nuestros genes expresados, es responsable por la síntesis de proteínas y, durante este proceso, se sintetizan las moléculas intermediarias denominadas mARNs. De esta forma, la obtención y caracterización de la secuencia de estas mARNs representa un atajo seguro para la determinación de la secuencia de los genes, permitiendo economizar un 97% del trabajo y de las inversiones. Un barrido rápido de los genes expresados, usando el secuenciamiento parcial de estas moléculas de mARNs, fue descrito por Adams *et al.*, en 1991, y utilizado con éxito por diversos grupos mundiales trabajando con diversos organismos. La técnica, denominada "EST sequencing" (secuenciamiento de "etiquetas de genes expresados"), permitió un aumento dramático en el número de genes humanos conocidos (aunque parcialmente) y permitió identificar una gran serie de genes importantes en diversas patologías y procesos celulares humanos. La técnica consiste en obtener una secuencia parcial, normalmente derivada de una de las extremidades de un gen. Esta secuencia, aunque parcial, es suficiente para determinar si un gen es nuevo o no, trazando así el perfil de genes activos en un tejido. Actualmente, las técnicas ESTs se usan como base en una serie de proyectos en el mundo, incluso por parte de grandes grupos tales como el CGAP (Cancer Genome Anatomy Project – www.ncbi.nlm.nih.gov/CGAP/).

Obstáculos de las técnicas usuales de generación de ESTs

La utilización de ESTs en el descubrimiento de los genes mostró un gran valor. Sin embargo, existen dos grandes obstáculos técnicos que hicieron que varios proyectos fueran abandonados. El primero de ellos se refiere a un factor ligado a la distribución de los diversos genes dentro de las células. Mientras algunos pocos genes son muy activos, con miles de mARNs dentro de una célula, la gran mayoría de ellos tiene un bajo nivel de actividad, es decir, tenemos pocas unidades funcionales representadas en cada célula (Gráfico 1).

Para generar las ESTs se arman las denominadas bibliotecas de cADN, que reflejan la población de los genes activos en una célula. Siendo así, los genes raros son poco representados, y su descubrimiento es caro, pues se requiere la generación de millares de ESTs para que los genes raros tengan probabilidades estadísticas de

estar representados. Por otra parte, la alta representatividad de los genes abundantes lleva a un secuenciamiento redundante de los mismos, aumentando los gastos de un proyecto e impidiendo la obtención de datos informativos referentes a nuevas secuencias.

Un segundo problema hallado en este abordaje se refiere a la limitación técnica de secuenciamiento de ADN existente aún en los días actuales. Los genes humanos tienen en promedio cerca de 4 mil bases para ser secuenciadas. Los métodos actuales permiten, en un proyecto a gran escala, secuenciar sólo unas 500 bases. Siendo así, tan sólo las extremidades de los genes son secuenciadas cuando se utiliza esta estrategia. La porción central normalmente no tiene representación. Con ello, no se alcanza la secuencia completa de los genes, y la porción interna de los mismos, aquella que contiene la información necesaria para generar las proteínas, no es determinada. Ese hecho puede ser fácilmente detectado por análisis en los bancos de datos, en donde es notable la pequeña frecuencia de ESTs derivadas de la porción central de los genes (Figura 2b).

Siendo así, para que los frutos del Proyecto Genoma Humano sean totalmente aprovechados, se hace necesario encontrar alternativas tecnológicas que permitan:

- 1) El acceso directo a las moléculas correspondientes sólo a los genes, evitando el 97% no informativo del genoma.
- 2) El acceso a la población de genes raros, permitiendo un avance en la dirección de descubrir los genes poco expresivos, evitando los genes abundantes ya ampliamente secuenciados.
- 3) La generación de datos referentes a la porción central y más informativa de estos genes, de manera de facilitar los análisis comparativos que permitan inferir la función biológica, además de montar la secuencia completa de todos los genes.

ORESTES: Una alternativa tecnológica para el descubrimiento de la porción central de genes humanos, con énfasis en los genes raros

De manera general, podemos decir que una biblioteca rica en información debe contener un gran número de diferentes títulos. Esto también puede pensarse para nuestros genes. La diversidad de informaciones es proporcional al número de genes. Si analizamos con cuidado el gráfico 1, podemos observar que la mayor diversidad de informaciones se encuentra en la categoría correspondiente a los genes raros, en donde está más del 90% de los genes humanos. De esta manera, para favorecer a los genes raros, una alternativa sería disponer de una tecnología que favoreciera la diversidad y riqueza de informaciones y no la abundancia de secuencias de bajo contenido informacional. Esta es exactamente una de las ideas centrales que sustentan a ORESTES.

ORESTES tiene en cuenta dos factores:

- 1) El descubrimiento de genes raros implica el descubrimiento de una población de genes bastante compleja en términos de diversidad de secuencia de ADN. Esta población con alta diversidad o gran cantidad de información puede determinarse preferencialmente si este factor es tenido en cuenta en el diseño experimental. Siendo así, el factor nivel de expresión (importante para los genes abundantes) no debe ser determinante, y la complejidad informacional de las moléculas de este grupo debería ser favorecida.

GRÁFICO 1

Clase	Número de diferentes genes/célula	Representatividad (copias de cada gen por célula)
Abundantes	<10	12.000
Intermedios	500	300
Raros	11.000	10

Fuente: Huang et al., 1999

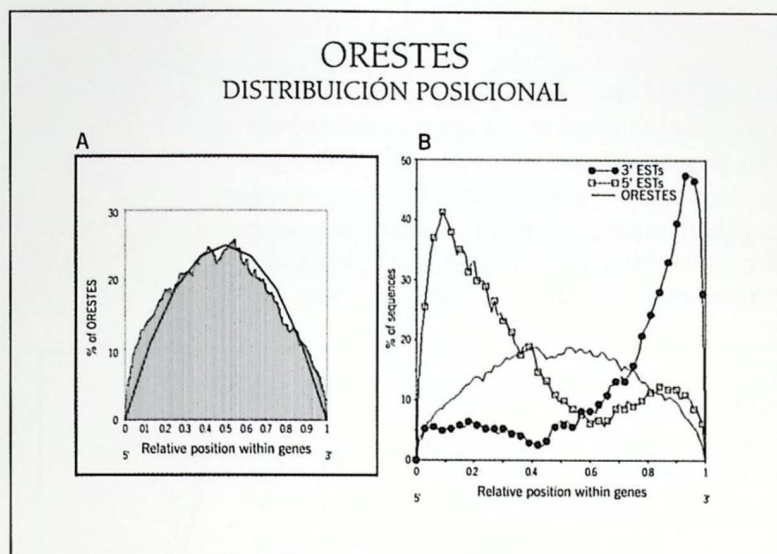


Figura 2

Distribución posicional de ORESTES. 2a - La línea representa la distribución teórica de ORESTES prevista de acuerdo con la fórmula discutida en el texto. Las barras representan la distribución real observada después de 10 mil ORESTES derivadas de tumores de mama. 2b - Los datos observados en ORESTES a partir de la distribución posicional son comparados con los datos observados para las secuencias derivadas de otros proyectos genoma a partir de cada una de las dos extremidades. Podemos ver la complementariedad de los dos abordajes.

2) Para alcanzar la parte central de las moléculas, no podríamos iniciar la reacción, como en los demás proyectos mundiales, utilizando moléculas iniciadoras ancladas en las extremidades de los genes. Deberíamos utilizar una gran serie de moléculas diferentes que se vinculen en porciones más internas de los diversos genes.

Siendo así, toda la idea que soporta la técnica ORESTES se basa en la complementariedad de moléculas. Una complementariedad que permite la efectivización de las reacciones de una manera alternativa. Esta complementariedad entre moléculas se da siempre en donde la diversidad de informaciones es mayor. Así las cosas, el elevado nivel de diversidad en la población de los genes más raros hace que éstos se vean favorecidos y ellos serían, teóricamente, los blancos principales de esta técnica. El hecho de que no usemos

moléculas complementarias a la extremidad de los genes, lo que resultaría en secuencias de las extremidades, hace que tengamos ESTs altamente informativas provenientes de la porción central y generalmente poco representada de los genes. En la jerga de la biología molecular, esta región es denominada Moldura Abierta de Lectura, o Open Reading Frame. Surgió entonces la técnica ORESTES (*Open Reading Frame ESTs*).

Para obtener la representación de un determinado punto dentro de un gen, es necesario que tengamos moléculas complementarias que se vinculen de los dos lados de este punto. La chance de que estos dos eventos de complementariedad se produzcan es proporcional al largo de la secuencia de cada uno de los lados de este punto. Siendo así, si el tamaño de un gen es 1 y la distancia desde el punto determinado hasta una de las extremidades es tenida como S , la probabilidad de que tengamos una interacción para cada uno será S y $1-S$. Como resultado de ello, la probabilidad de representación de este punto será $S(1-S)$. En términos prácticos, la posibilidad de que una ORESTES contenga exactamente la porción central de un determinado gen es igual a $0,5 \times 0,5 = 0,25$, mientras que la probabilidad de que represente un punto relativo al 10% final o inicial es igual a $0,1 \times 0,9 = 0,09$, etc. Al plotear los valores de esta fórmula simple para las diversas regiones de un gen, tendremos la línea sólida observada en la figura 2a. Las barras en esta misma figura representan los datos reales derivados de las primeras 10 mil ORESTES derivadas de tumores de mama. La semejanza de las dos representaciones permite concluir que ORESTES funciona como estaba previsto, siendo la única técnica disponible para la generación de datos a gran escala referentes a la porción central de los genes.

ORESTES

Esta técnica comenzó a ser desarrollada en el año 1993, con genes derivados del parásito humano *Schistosoma mansoni*, el agente causante de la esquistosomosis. Se generaron y analizaron cerca de mil ESTs, que permitieron evaluar algunas características de ORESTES. Los genes abundantes, ya conocidos en este parásito, no tuvieron representación y se obtuvo un alto porcentaje de secuencias inéditas (Dias Neto, *et al.*, 1997).

Posteriormente pasamos a evaluar con mayor detalle las características de ORESTES. Para ello era necesaria su evaluación en un modelo

ya bastante estudiado, donde pudiéramos tener gran número de genes conocidos –verificando la posición exacta de nuestras ESTs– y pudiéramos evaluar en mayor escala la capacidad de obtener genes raros. Comenzamos entonces a generar un número similar de datos derivados de un tumor colon-rectal humano, tejido bastante estudiado y que permitía una real evaluación de ORESTES. Para nuestra satisfacción, aun trabajando con el genoma más estudiado, continuamos con un alto porcentaje de datos nuevos referentes a secuencias humanas jamás descubiertas por otros grupos mundiales. Frente a estos datos, el Instituto

Ludwig decidió financiar con 250 mil dólares un proyecto piloto para generar 10 mil ORESTES derivadas de tumores de mama. Para generar esa cantidad de datos en un corto lapso de tiempo, fuimos auxiliados por algunos grupos del estado de San Pablo que efectuaron el secuenciamiento de los clones que generamos. Se montó también un laboratorio de bioinformática, capaz de vérselas con la gran cantidad de datos generados, permitiendo así un análisis minucioso. Una vez más los datos se mostraron sorprendentes. Las 10 mil ORESTES confirmaron las predicciones de distribución central y mostraron una capacidad de reflejar genes raros por lo menos equivalente a las mejores del mundo. Estos datos fueron reunidos bajo la forma de un artículo científico publicado por la revista de la Academia Nacional de Ciencias de Estados Unidos (PNAS), tras ser recomendado por el Premio Nobel, Dr. Phillip Sharp (Massachusetts Institute of Technology – MIT). Las figuras de este trabajo (presentadas también en las figuras 2a y 2b)) demuestran que las ORESTES son realmente derivadas de la porción codificadora de los genes. Los análisis también demuestran con claridad que los genes raros son favorecidos y más fácilmente representados.

La alta calidad de los datos hizo que una vez más recibiéramos un apoyo inestimable por parte del Instituto Ludwig, que esta vez se sumó al apoyo de la FAPESP. El total financiado a “fondo perdido” fue de 10 millones de dólares, repartido en partes iguales entre el Ludwig y la FAPESP, surgiendo entonces así el Proyecto Genoma del Cáncer Humano Ludwig/FAPESP (*Human Cancer Genome Project - HCGP*). El día 26 de marzo de 1999 el HCGP fue lanzado por las direcciones del

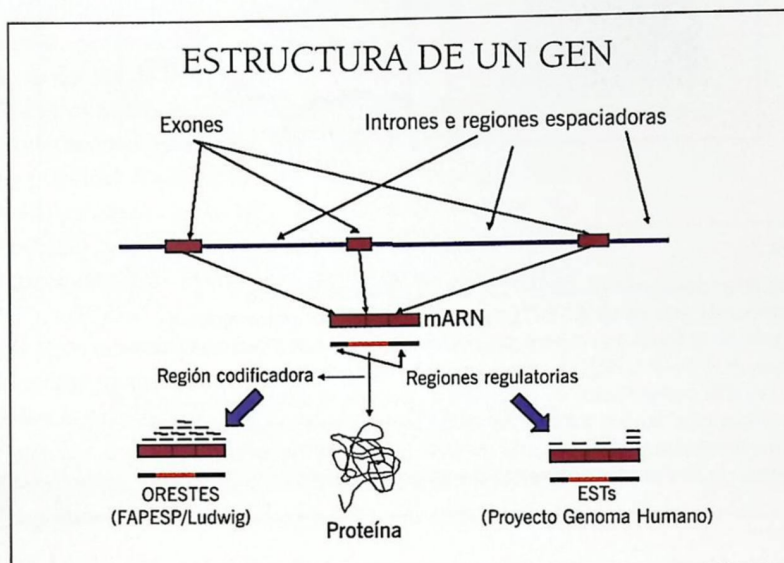


Figura 3

Ilustración que representa las fases iniciales de organización del Proyecto Genoma del Cáncer Humano – Ludwig/FAPESP.

Instituto Ludwig y de la FAPESP. Se formó una red de 33 laboratorios, involucrando a cerca de 200 investigadores. El objetivo entonces delimitado fue extremadamente audaz: descubrir los genes involucrados en la formación de tumores de gran importancia en Brasil. La organización del proyecto se sustenta en una coordinación general, a cargo de Andrew JG Simpson. Los tejidos que se utilizan son generalmente microdisecados por el Dr. Fernando Soares, jefe del Departamento de Patología del Hospital del Cáncer. Este material es tratado por las coordinaciones específicas, a saber:

Coordinador de ARN: Luis Fernando Lima Reis

Coodinador de Bibliotecas de cADN: Emmanuel Dias Neto

Coordinadores de secuenciamento: Fernando Costa (Universidad Estadual de Campinas, UNICAMP, San Pablo), Marco Antônio Zago (Ribeirão Preto, San Pablo), Maria A. Nagai (Fac. de Medicina de la Universidad de San Pablo, USP), Marcelo Briones (Universidad Federal de San Pablo, UNIFESP) y Sérgio Verjovski (Inst. de Química, USP).

Alrededor de estos centros tenemos los laboratorios de secuenciamento distribuidos en las ciudades anteriormente mencionadas y también en Botucatu, Araraquara y São José dos Campos (todas en el interior de San Pablo). La organización, desde la obtención de los tejidos hasta la generación de ORESTES puede verse en la figura 3.

La principal institución mundial de investigaciones sobre el cáncer es el NCI (National Cancer Institute) uno de los principales Institutos del NIH (National Health Institute, EE.UU.). A pedido nuestro, su director, el Dr. Richard Klausner, emitió su parecer sobre nuestro proyecto. Dicho dictamen constituyó una importante fuente de reconocimiento a nuestro trabajo, sirviendo como estímulo para continuar el mismo. A continuación, las palabras del Dr. Klausner:

"I am impressed and pleased with the enormous effort of the Brazilian cancer genome project. The production of this amount of valuable information and the generous and open sharing with the international scientific community is a model of the spirit of the international fight against cancer. This fight can only be waged with the full participation of the talents of people throughout the world and the free and open exchange of ideas and information".

Proyecto Genoma del Cáncer Humano Ludwig/FAPESP - Invertir en lugar de esperar

En el momento del lanzamiento del HCGP, existía solamente un gran proyecto mundial centrado en la generación de ESTs derivadas de tumores: el CGAP (Cancer Genome Anatomy Project). Dicho proyecto, lanzado en 1997 con financiamiento del Instituto Nacional del Cáncer de Estados Unidos, tenía como objetivos principales la determinación de los genes involucrados en los tumores de mayor importancia epidemiológica en Estados Unidos, tales como los de mama, pulmón, próstata y colon. Cuestiones políticas llevaron a que el Instituto Nacional de Salud de EE.UU. diera prioridad a la generación de datos relativos a los tumores más importantes en dicho país.

No obstante, sabemos que los tipos, la preponderancia y la mortalidad por cáncer son diferentes entre los países del primero y del tercer mundo. Cuestiones raciales, geográficas, culturales y socioeconómicas actúan de manera importante generando índices bastante diversificados entre diferentes países. Datos recientes sobre la incidencia y la mortalidad por cáncer en todo el mundo fueron publica-

dos por Pisani *et al.*, 1999 y por Parkin *et al.*, 1999). De modo general, vemos que la incidencia y la mortalidad de la mayoría de los tipos de cáncer viene creciendo de una manera impresionante en los últimos años. Se estima para el año 2010 un aumento del 52,4% en las muertes por cáncer, totalizando cerca de 7.900.000 muertes en ese año. El panorama es desalentador en todo el mundo, y aún peor en los países en desarrollo, donde el crecimiento porcentual de muertes por cáncer tiende a duplicarse con relación al crecimiento en los países desarrollados. Los datos pueden visualizarse mejor en el gráfico 2:

GRÁFICO 2				
	Año 2000	A	Año 2010	B
Mundo	6.366.000	22,8%	7.899.000	52,4%
Países desarrollados	2.669.000	14,5%	3.019.000	29,5%
Países en desarrollo	3.697.000	29,6%	4.880.000	71,1%

A: crecimiento porcentual en el año 2000 con relación a 1990.

B: crecimiento porcentual en el año 2010 con relación a 1990.

Fuente: Pisani *et al.*, 1999. Estimates of the worldwide mortality from 25 cancers in 1990. International Journal of Cancer 83:18-29.

En el trabajo de Parkin *et al.*, 1999, vemos que los tumores de cuello de útero aún tienen una alta incidencia en países del tercer mundo, especialmente en América Central (31,2 casos/100.000 habitantes) y América del Sur (27 casos/100.000 habitantes). En los países desarrollados esta tasa es inferior a 14/100.000. Los índices son aún más preocupantes para tumores nasofaríngeos (agrupados en nuestros estudios como tumores de cabeza y cuello) en los cuales la incidencia comparada, normalizada para 100.000 habitantes de sexo masculino entre países desarrollados y en desarrollo es de 35.7 frente a 4.1 respectivamente. El cuadro no mejora cuando observamos la mortalidad causada por el cáncer (Pisani *et al.*, 1999). Si consideramos todo el mundo, tenemos anualmente 197 mil muertes tan sólo por cáncer de la cavidad oral. El 74% de la mismas se produce en países en desarrollo. De las 190 mil muertes anuales por cáncer de cuello de útero, el 78% ocurre en países en desarrollo. De esta manera, para descubrir los genes vinculados a tumores importantes en nuestro país, que quedaron fuera de las prioridades del programa americano, teníamos dos opciones: esperar o invertir.

El HCGP fue el fruto del coraje y la osadía de invertir. Invertir en el descubrimiento de genes importantes para tumores que afectan a países del tercer mundo, con el coraje de apostar a una tecnología alternativa totalmente desarrollada en Brasil y con la osadía de participar, sin haber sido invitados, en selecto club de países desarrollados ligados al secuenciamiento de genes humanos.

No tenemos dudas de que el gran éxito de este proyecto reside en el compromiso del grupo de cerca de 200 investigadores brasileños de altísima capacidad, distribuidos en 33 laboratorios del estado de San Pablo. Se suma a ello el espíritu humanitario de los pacientes que permitió el estudio del material tumoral y la colaboración de diversos grupos de cirugía, que fueron de un valor inestimable.

Resultados actuales del HCGP

Tenemos la satisfacción de haber alcanzado, cerca de 8 meses antes del plazo, la meta de medio millón de secuencias humanas, todas producidas por el método ORESTES. De ese total, cerca del 30% representa fragmentos nuevos, jamás obtenidos en ningún otro proyecto. Además de este 30%, otras ORESTES permitieron el descubrimiento de nuevas formas de expresión de genes humanos ya conocidos, o determinaron por primera vez tejidos en donde algunos genes son activos. Nuestros datos permitieron descubrir genes incluso en fragmentos del genoma humano exhaustivamente estudiados por otros grupos mundiales. Un ejemplo de ello puede graficarse a través del análisis del cromosoma 22 humano, secuenciado por el Sanger Centre en

GRÁFICO 3

TUMOR	América del Sur (clima templado)		América del Sur (clima tropical)		ESTs generadas (x1000)	
	Incidencia	Mortalidad*	Incidencia	Mortalidad*	CGAP	HCGP1
Boca/Faringe	8,4	1,1 (11)	10,9	7,9 (8)	3,5	147,6
Estómago	19,6	6,2 (2)	23,3	30,8 (1)	22,0	38,8
Mama	76,1	5,6 (4)	28,6	15,1 (4)	15,5	70,9
Próstata	22,9	3,5 (6)	16,8	11,9 (6)	14,7	19,2
Colon	26,8	5,9 (3)	9,6	14,6 (5)	53,5	62,5
Pulmón	53,3	10,2 (1)	15,1	22,4 (2)	61,8	5,4
Cuello de útero	28,2	2,6 (8)	24,5	15,6 (3)	0,9	21,4
TOTAL	235,3	35,1	128,8	105,8	171,9	365,8

La tasa de incidencia es presentada para grupos de 100 mil individuos. Mortalidad: número total de muertes anuales causadas por determinado tipo de cáncer (en miles). (Fuente: Pisani et al., 1999. Estimates of the worldwide incidence from 25 cancers in 1990. International Journal of Cancer 83:18-29; Parkin et al., 1999. Estimates of the worldwide mortality from 25 cancers in 1990. International Journal of Cancer 80:827-841.

* Los números que están entre paréntesis indican la posición de un tumor específico en términos de mortalidad;

(1) Actualizado al 20 de septiembre de 2000.

Inglaterra. Cuando los datos de ORESTES fueron usados en el mapeamiento de genes en este cromosoma, se descubrieron 144 nuevos genes (Souza et al., en proceso de publicación).

Los números obtenidos hablan por sí solos. El HCGP ya es el segundo proyecto de generación de ESTs humanas en importancia, tan sólo atrás del CGAP/NIH. Para diversos tumores humanos ya somos el primer grupo mundial en términos de secuencias de genes. Cuando evaluamos nuestra contribución al estudio de tumores específicos, constatamos que nuestros

grupo lidera en número de ESTs generadas a partir de tumores de mama, cabeza y cuello, estómago, próstata, colon intestinal y testículos. En total, tenemos hoy (al 2 de octubre de 2000) 701.471 ORESTES generadas. Entre ellas, cerca de 30% representan secuencias humanas inéditas.

La productividad de nuestro proyecto con relación a algunos de los principales tipos de tumores es presentada junto a la incidencia y mortalidad de los mismos en diferentes regiones sudamericanas.

El impacto de estos tumores es dramático, causando tan sólo en América del Sur cerca de 141 mil de las 764,7 mil muertes anuales por cáncer.

Las ORESTES producidas, después de pasar por nuestro control de calidad, son depositadas en bancos de datos públicos, de acceso gratuito, permitiendo su uso por parte de toda la comunidad científica internacional (www.ncbi.nlm.nih.gov).

Este gran número de datos de alta calidad, proporcionado por el uso de ORESTES, hizo que el financiamiento inicial se duplicara de 10 a 20 millones de dólares totalmente voltados a la investigación. Con ello dispondremos de un parque instalado de 11 secuenciadores automáticos de tecnología capilar, lo que permitirá una significativa aceleración del proyecto. Las metas fueron reevaluadas y tenemos el compromiso de generar un millón de ORESTES hasta el final de 2000.

Perspectivas finales

Cuando se concluya, el Proyecto Genoma Humano generará un inmenso y detallado atlas de la biología humana. El HCGP, a través de la técnica de ORESTES permitirá el mapeamiento de un alto número de genes humanos dentro de dicho atlas. El atlas completo del genoma humano suministrará informaciones a nivel de las secuencias de ADN, lo que permitirá diagnosticar, estudiar, tratar, curar y prevenir la mayoría de las enfermedades humanas. La investigación biomédica pasará por un antes y un después de lo que fue y será hecho por el PGH. Podrán crearse nuevas drogas, para que las mismas actúen tan sólo en vías específicas de agentes infecciosos, podrán diseñarse otras drogas para suplir las funciones de genes defectuosos, pudiendo

éstas usarse en el tratamiento y prevención de enfermedades como el cáncer, la diabetes y las afecciones cardíacas. No sólo los médicos deberán estar listos para lidiar con esta nueva tecnología, sino también toda la sociedad, que deberá estar preparada para vérselas con esta nueva perspectiva.

El PGH viene repleto de expectativas, pero ciertamente con una carga aún mayor de implicancias sociales. Los datos generados corren el riesgo de ser mal utilizados, y la posibilidad de invasión de la privacidad individual debe ser considerada seriamente. El PGH no promete curar el cáncer o corregir la esquizofrenia o el Mal de Alzheimer; no promete tampoco definir los cambios moleculares precisos que se producen en la transformación de un embrión en adulto, ni promete explicar los comportamientos agresivos antisociales. Sin embargo, tenemos la seguridad de que si algún día logramos descubrir los misterios de la carcinogénesis, si algún día comprendemos la bioquímica de los disturbios neurológicos y si realmente tenemos la pretensión de comprender los procesos de crecimiento y desarrollo, es necesario que tengamos antes que nada un cuadro completo del genoma humano. En ese momento, el "código de los códigos" estará realmente descifrado. El gran desafío será el utilizar esas informaciones a conciencia y con responsabilidad.

Bibliografía:

- Crollius, HR, Jaillon, O, Bernot, J, DaSilva, C, Bouneau, L, Fischer, C, Fizames, C, Wincker, P, Brottier, P, Quetier, F, Saurin, W & Weissenbach, J. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics* 25: 235-238.
- Dias Neto, E, Harrop, R, Corrêa-Oliveira, R, Wilson, RA, Pena, SDJ & Simpson AJG. 1997. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the efficient generation of ESTs from nanogram quantities of mRNA. *Gene* 186: 135-142.
- Dias-Neto, E, Corrêa, RG, Verjovski-Almeida, S, Briones, MRS, Nagai, MA, Silva Jr, W, Zago, MA, Bordin, S, Costa, FF, Goldman, GH, Carvalho, AF, Matsukuma, A, Baia, GS, Simpson, DH, Brunstein, A, Oliveira, PSL, Bucher, P, Jongeneel, CV, O'Hare, M, Soares, F, Brentani, RR, Reis, LFL, Souza, SJ & Simpson, AJG. 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings of the National Academies of Sciences (USA)* 97: 3491-3496.
- Ewing, B & Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics* 25: 232-234.
- Huang, GM, Ng, W, Farkas, J, He, L, Liang, HA, Gordon, D, Yu, J & Hood, L. Prostate Cancer expression profiling by cDNA sequencing analysis. 1999. *Genomics* 59: 178-186.
- Liang, F, Holt, I, Pertea, G, Karamycheva, S, Salzberg, SL & Quackenbush, J. 2000. Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics* 25: 239-240.
- Parkin, DM, Pisani, P & Ferlay, J. 1999. Estimates of the worldwide mortality from 25 cancers in 1990. *International Journal of Cancer* 80:827-841.
- Pisani, P, Parkin, DM, Bray, F & Ferlay, J. 1999.
- Souza, SJ *et al*; 2000. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proceedings of the National Academies of Sciences* (en proceso de publicación).

