
El genoma humano

Pedro García Barreno

Arbor CLXXI, 673 (Enero), 145-179 pp.

El redescubrimiento de las leyes de Mendel sobre la herencia en las semanas que abrieron el siglo XX incitó una búsqueda científica para comprender la naturaleza y el contenido de la información genética que ha impulsado la biología durante los últimos cien años. El progreso científico conseguido se ha fraguado en cuatro fases que se corresponden, aproximadamente, con los cuatro cuartos del siglo XX. El primero estableció las bases celulares de la herencia: los cromosomas. El segundo definió las bases moleculares de la herencia: la doble hélice de ADN. El tercero descifró las bases informativas de la herencia con el descubrimiento de los mecanismos biológicos mediante los que la célula lee la información codificada en los genes; luego, con la invención de la tecnología del ADN recombinante de clonaje y de secuenciación, los científicos pudieron hacer lo mismo. El último cuarto del siglo estuvo marcado por un lento pero constante esfuerzo para descifrar genes primero y, por fin, genomas enteros que han propiciado el desarrollo de la genómica. El día 26 de junio de 2000 se hacía público un «borrador de trabajo» de la secuencia del genoma humano. Las revistas Nature (vol 409, n.º 6822) y Science (vol 291, n.º 5507) dedicaban números especiales a la publicación de la secuencia en el mes de febrero de 2001 (el día 15 Nature y, al día siguiente, Science). «La humanidad ha recibido un gran regalo. La conclusión de la secuencia del genoma humano ofrece una herramienta poderosa para descifrar los secretos de nuestra herencia genética y para precisar nuestro lugar entre otros participantes en la aventura de la vida».

Las bases celulares de la herencia

Pocas dudas pueden albergarse respecto a que los primeros humanos ponderasen las semejanzas entre padres e hijos, y que tales observacio-

nes debieron aplicarlas, en su beneficio, a los organismos que iban domesticando. Sin embargo, la genética moderna, basada en la «teoría genética de la herencia» comenzó con el trabajo de Gregor Johann Mendel.



FIGURA 1. Johann Gregor Mendel (1822-1884)

Mendel (Figura 1) no fue el primero en realizar experimentos de hibridación; pero sí, quién interpretó los resultados en términos de rasgos individuales. Su trabajo seminal —«Experimentos con Plantas Híbridas»—, llevado a cabo en el reducido espacio del jardín de un monasterio y dado a conocer a la Sociedad de Historia Natural de Brunn en 1865, fue el resultado del diseño —de acuerdo con el método científico— de numerosos experimentos, de la comparación de los resultados con modelos matemáticos y de la formulación de una hipótesis para explicar las diferencias observadas. Aunque Mendel concibió un patrón matemático preciso para la transmisión de los rasgos hereditarios, desconoció el mecanismo biológico subyacente. Si bien

cualquier estudiante de nuestros días conoce la figura de Gregor Mendel, su trabajo pasó desapercibido a la comunidad científica de antaño durante 35 años. En 1900, la monografía de Mendel, publicada en 1866, fue «descubierta» por tres botánicos: Hugo de Vries, en Holanda; Carl Correns, en Alemania, y Eric von Tschermak-Seysenegg, en Austria.

«Experimentos con Plantas Híbridas», donde se describe cómo se transmiten los rasgos hereditarios, representa, hoy, una de las referencias más influyentes e imperecederas de la historia de la ciencia. Los experimentos demostraron que los rasgos hereditarios se transmiten mediante pares de factores discretos; los miembros (alelos) de cada par (gen) proceden uno del padre y otro de la madre. Los conceptos más importantes inferidos de los experimentos de Mendel son el principio de segregación —proceso por el que los alelos se separan para producir gametos haploides—, y el principio de combinación independiente de los diferentes pares de alelos. Desde entonces, varios han sido los mojones que han jalonado el camino de la genómica; camino que, en su mayor parte, es el del desarrollo de la biología o genética molecular.

En 1877, Walter Flemming (1843-1905) describió los cromosomas; estructuras que Wilhelm Roux postuló, en 1883, como transportadoras de los factores hereditarios. En 1902, Theodore Boveri y Walter S Sutton confirmaron que un gen es parte de un cromosoma. La palabra gen –última sílaba del término «pangen» utilizado por Darwin– fue utilizada por Wilhelm Johannsen (1857-1927) para referirse a los factores mendelianos, aunque el concepto de «gen» estaba implícito en la visualización de Mendel de un elemento físico o factor (*Anlage*) que actúa como fundamento para el desarrollo de un rasgo. La teoría del gen como una unidad discreta de un cromosoma fue desarrollada por Thomas H Morgan (Figura 2). Por su parte, William Bateson (1861-1926) denominó genética (de la palabra griega «generar») a la incipiente ciencia de la herencia;

el Primer Congreso Internacional de Genética (Londres, 1899), se anunció como «Conferencia Internacional sobre Hibridación».

Por su parte, cuando el trabajo de Mendel era «descubierto» en 1900, Archibald E Garrod (Figura 3) estudiaba enfermedades metabólicas congénitas humanas; una de ellas, la alcaptonuria, está causada por el bloqueo del catabolismo del aminoácido fenilalanina. Garrod propuso que la alcaptonuria está causada por un gen defectuoso que produce una enzima anormal causante del bloqueo metabólico. El concepto de Garrod de «un gen mutado un bloqueo metabólico», fue ignorado –como había pasado con el trabajo de Mendel– durante 30 años. En 1941, George W Beadle y Edward L Tatum (Figura 4) redescubrieron tal concepto; establecieron, utilizando un hongo como material de experimentación, que cada mutación del organismo estudiado se acompañaba de la alteración una determinada vía metabólica. Dado que se conocía que las vías metabólicas estaban gobernadas por enzimas, la

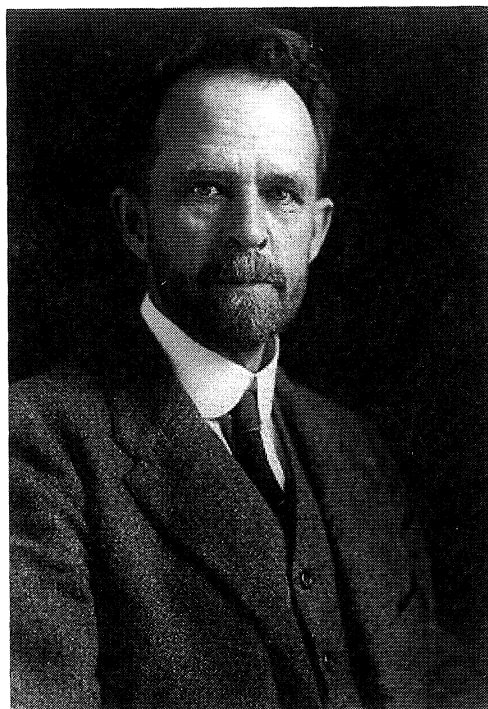


FIGURA 2. Thomas Hunt Morgan (1866-1945). Premio Nobel de Fisiología o Medicina (PNFoM), en 1933, «por sus descubrimientos sobre el papel jugado por los cromosomas en la herencia».

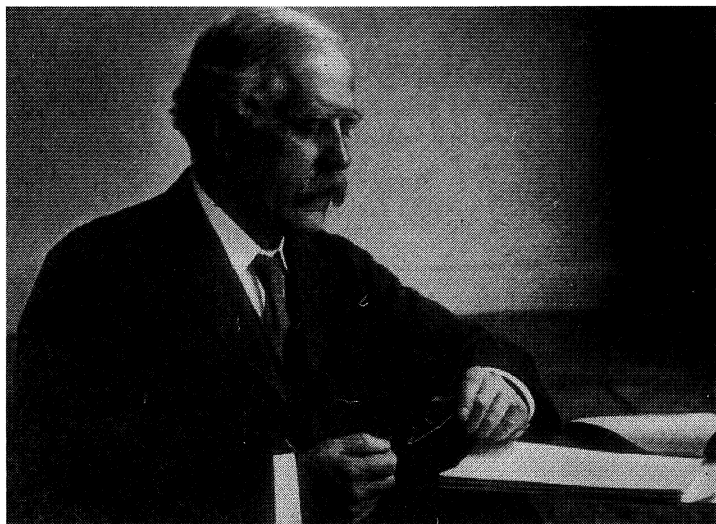


FIGURA 3. Archibald Garrod (1857-1936). *Regius Professor of Medicine* en Oxford

conclusión de tales experimentos fue que cada gen determinaba la estructura de una enzima: hipótesis «un gen → una enzima», luego replanteada como un «un gen → un polipéptido». En 1944, en colaboración con Colin M MacLeod y Maclyn McCarthy, Oswald T Avery (Figura 5)

descubrió que el principio transformante, el gen, es ácido desoxirribonucleico (ADN).

La doble hélice de ADN

La estructura del ADN se definió en 1953; año en que dos jóvenes y desconocidos científicos pusieron en marcha una revolución en las ciencias de la vida cuyas consecuencias persisten. James Watson y Francis Crick determinaron la estructura del ADN: «Deseamos sugerir una estructura para el ADN. Esa estructura tiene hechos originales que son de considerable interés biológico». Su observación ha tenido extraordinarias consecuencias y representa uno de los logros capitales de la ciencia del siglo XX. La biotecnología, la medicina molecular, la terapia génica o el Proyecto genoma humano, son fruto de aquel trabajo (Figura 6).

Una molécula de ADN es una larguísima hebra de cuatro subunidades diferentes denominadas bases: adenina, citosina, guanina y timina (A, C, G y T). Esas bases están unidas entre sí a modo de las cuentas de un collar; ello, mediante un mecanismo de engarce formado por moléculas de azúcar (desoxirribosa) y fosfato. El ADN en el núcleo celular se dispone en dos hebras o bandas paralelas enrolladas entre sí y que forman una superestructura llamada doble hélice y a modo de una escalera de caracol. Las bases de cada una de las bandas interactúan entre sí formando pares de bases que semejan los escalones de la escalera; por su

parte, las moléculas de azúcar y los grupos fosfato de cada banda forman los largueros de esa escalera (Figura 6a). Además, si una determinada posición en una de las bandas está ocupada por la base A, el otro miembro del par, correspondiente a la otra banda, será T, y viceversa. De manera similar, a C corresponderá G, y viceversa. En otras palabras, A es complementaria de T, y C de G. Esta complementariedad se denomina «regla de Chargaff» (Edwin Chargaff, n1905).



FIGURA 4. George Wells Beadle (1903-1989) y Edward Lawrie Tatum (1909-1975) compartieron el PNFoM, en 1958, «por sus descubrimiento de que los genes actúan regulando acontecimientos químicos definidos»

Consecuencia de esta complementariedad estricta entre bases en la doble hélice de ADN es que, si se conoce la secuencia de bases de una de las bandas, puede deducirse automáticamente la secuencia de bases de la banda opuesta. En otras palabras, cada banda de ADN contiene toda la información necesaria para recrear la otra banda. En la célebre publicación sobre la estructura del ADN, Watson y Crick escribieron: «No se nos escapa que el apareamiento específico de bases propuesto sugiere, inmediatamente, un posible mecanismo de copia para el material genético». Ello garantiza un mecanismo de copia fiable cuando la célula se divide; y para formar un organismo complejo, que contiene miles y miles de millones de células formadas a partir de una célula fecundada, se requieren numerosos ciclos de división celular.

En términos generales, el mecanismo de replicación del ADN es sencillo; primero, las dos bandas complementarias de ADN se desenrollan y se separan la una de la otra. Luego, se copian dos nuevas hebras a partir de cada una de las dos bandas por separado y usando la regla A:T/C:G. Dado que cada banda independiente de ADN dirige la síntesis de su complementaria, el resultado de la replicación es la síntesis de dos moléculas de doble banda de ADN idénticas. La pieza principal de la máquina copiadora, responsable de la replicación del ADN, es una molécula denominada ADN polimerasa.

¿Cómo accede la célula a la información génica almacenada en su ADN?. El programa genético completo de un organismo se denomina genoma. El genoma puede imaginarse como una secuencia de pares de bases (pb = A:T, C:G, G:C o T:A); de 100 millones a 3000 millones de pares



FIGURA 5. Oswald Theodore Avery (1877-1955)

según el organismo. Si una base se representa por una letra impresa en páginas similares a las que acogen el presente artículo, el genoma humano corresponde a una enciclopedia de un millón de páginas. El ADN en el núcleo de una célula en reposo (interfase) forma, enrollado sobre un eje de proteínas (histonas), una larga hebra continua de cromatina. Cuando se inicia la división celular –mitosis–, la cromatina se condensa y dispone en estructuras discretas o cromosomas. El número de cromosomas varía entre las diferentes especies, desde un par en una lombriz (*Ascaris lumbricoides*) hasta más de cien en algunas mariposas y crustáceos; la especie humana tiene 23 pares de cromosomas.

Hay varias clases de ADN en el genoma humano. En general,

las secuencias codificantes representan menos del 5% del genoma; son secuencias representadas una sola vez por genoma haploide y ubicadas en una cromatina laxamente empaquetada denominada eucromatina. Por su parte, secuencias iterativas –denominadas ADN basura– representan el 60% del genoma y se localizan en una cromatina densamente condensada denominada heterocromatina.

Un gen típico puede subdividirse en dos componentes independientes pero funcionalmente interrelacionados; uno de ellos, región codificante, contiene la información necesaria para sintetizar una proteína que ejecutará la función génica correspondiente. Sin embargo, la mayoría de los genes contienen, en su región codificante, cortas secuencias que se expresan (exones que codifican, aproximadamente, 50 codones) interrumpidas por otras largas secuencias (10 kb) silentes (intrones, «basura»): genes discontinuos. Ello significa que los genes transcriben exones e intrones en un pre-ARN mensajero (pre-mARN), que será procesado –en términos generales excluyendo los intrones– en un ARN mensajero (mARN) que traducirá, en la maquinaria ribosómica cito-plasmática, la correspondiente pro-

teína. Richard J Roberts y Phillip A Sharp (Figura 7) recibieron el PNFoM de 1993 «por su descubrimiento de genes discontinuos». Tanto la transcripción como el procesamiento del transcrito (pre-mARN) son tareas flexibles que permiten lecturas y procesamientos alternativos.

La otra parte del gen, denominada región reguladora o de control –que no es codificante– es un interruptor de «encendido-apagado» que determina si la región codificante ha de expresarse o no. Junto a genes codificantes de proteínas, existen otros muchos genes que transcriben ARNs no codificantes (ncARNs) como producto final. Tal ncARN incluye: ARNs de transferencia que, adaptándose a los tripletes del ARN mensajero, colocan uno tras otro los aminoácidos que construyen una cadena peptídica; ARN ribosómicos (rARNs) involucrados en la estructura de los ribosomas; pequeños ARNs nucleolares requeridos en el procesamiento de los rARNs, y pequeños ARNs nucleares que son componentes críticos de los espliceosomas –macrocomplejos ribonucleoproteicos– encargados de procesar en el núcleo los pre-mARNs en mARNs que traducen proteínas. Existen otros ARNs sin función bien conocida.

Una primera característica del genoma es la existencia de regiones ricas y pobres en el dinucleotido GC; regiones que pudieran tener funciones biológicas diferentes: densidad génica, composición de secuencias iterativas o correspondencia con el patrón de bandas cromosómicas y con la tasa de recombinación. Otro hecho distintivo es la distribución de las llamadas islas CpG; dinucleotido singular debido a su escasez en el genoma humano. Por su parte, se desconoce el papel que el ADN iterativo juega en la célula, aunque representa una extraordinaria fuente de información, todavía de difícil acceso, sobre los procesos biológicos. Las repeticiones constituyen un rico registro paleontológico en el que se esconden claves evolutivas. Como marcadores pasivos, proporcionan apuntes

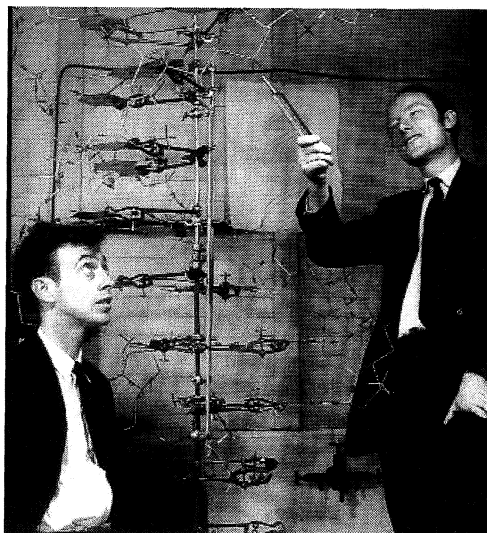


FIGURA 6. James Dewey Watson (n1928) y Francis Harry Compton Crick (n1916) compartieron con Maurice Hugh Frederick Wilkins (n1916) el PNFoM de 1962 «por sus descubrimientos sobre la estructura molecular de los ácidos nucleicos y su importancia en la transferencia de información en la materia viva».

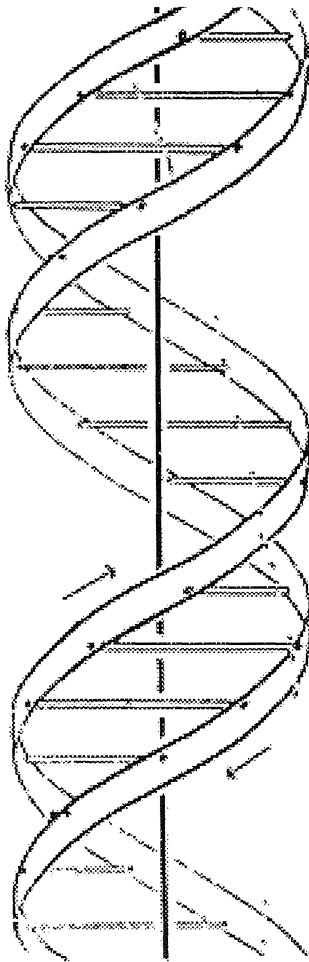


FIGURA 6a. La figura es un mero diagrama. Las dos bandas enrolladas representan las dos cadenas azúcar-fosfato y las curvas horizontales los pares de bases que estabilizan la estructura. La línea vertical señala el eje de la doble hélice (Nature, 25 abril, 1953).

para estudiar procesos de mutación y selección; es posible reconocer cohortes de iteraciones nacidas al unísono y seguir sus destinos en diferentes regiones del genoma o en especies diferentes. Como marcadores o agentes activos las iteraciones han reestructurado el genoma al provocar reordenamientos ectópicos, crear por entero nuevos genes y remodelar genes existentes; y también arrojan luz sobre la estructura y dinámica cromosómica, y proporcionan herramientas útiles para estudios de genética médica y de poblaciones.

El ADN iterativo puede concentrarse en aglomerados (*cluster*) –satélites– o dispersarse –transposones–. Conglomerados de secuencias repetidas o iterativas representan el 10-15% del genoma y comprenden despliegues de cortas repeticiones dispuestas en un orden cabeza-cola. Tales repeticiones en tándem, sin relación entre ellas, se denominan, colectivamente, satélites de ADN y cuyas longitudes varían desde unos pocos nucleótidos hasta varios millones. Su localización también es muy heterogénea; algunas repeticiones se encuentran, sólo, en las heterocromatinas pericentromérica o telomérica, y algunas solo se encuentran en un determinado cromosoma. Las secuencias simples iterativas (*simple sequence repeats* – SSRs) son estructuras repetidas muy frecuentes en el genoma humano; son repeticiones en tándem de una unidad particular. Las SSRs de una unidad repetida corta ($n = 1-13$ bases) se denominan microsátélites; aquellas con

unidades repetidas mayores ($n = 14-500$ bases) se denominan minisátélites. Las SSRs comprenden, aproximadamente, el 3% del genoma, registrándose una SSR cada 2 kb y siendo el componente principal repeti-

ciones dinucleotídicas. Las SSRs desempeñan un importante papel en estudios de genética humana por mostrar un extraordinario grado de polimorfismo de longitud en las poblaciones humanas.

Otra clase de secuencias iterativas de pares de bases incluye familias de secuencias dispersas por todo el genoma, a veces incluso dentro de los genes y que representan el 45% del ADN total. Estas secuencias tienen las propiedades de los elementos transponibles. La mayor parte de la genética clásica se centró en la localización de los genes en los cromosomas; cada gen o más precisamente cada alelo, ocupa un lugar fijo (locus) en un determinado cromosoma, con lo que la estructura de un determinado mapa genético es prácticamente invariable. Sin embargo, a comienzos de los 1940s los investigadores encontraron que algunas secuencias de ADN pueden cambiar de posición. Tales secuencias móviles se denominan elementos génicos transponibles o, simplemente, transposones. Los transposones fueron descubiertos por Barbara McClintock a raíz de sus estudios sobre la inestabilidad genética del maíz.

Los transposones son elementos móviles que pueden mudarse, por ellos mismos, desde una posición a otra en una molécula de ADN. La mayoría de las secuencias repetidas en el genoma humano derivan de elementos transponibles. El 45% del genoma humano pertenece a esta clase de secuencias; la mayoría del remanente del genoma, constituido por secuencias únicas de ADN, debe haber derivado, también, de transposones ancestrales aunque, en la actualidad, irreconocibles; en los mamíferos, la casi totalidad de los transposones pertenecen a cuatro tipos. Tres de ellos se mudan de una localización a otra a través de un ARN intermediario utilizando su autocapacidad de transcriptasa inversa (retrotransposones o elementos móviles homólogos a la forma integrada de retrovirus): SINEs (secuencias cortas dispersas, *short interspersed sequences*), LINEs (secuencias largas dispersa, *long interspersed sequences*) y LTRs (retrotransposones flanqueados por largas secuencias repetidas terminales homólogas a las que flanquean los retrovirus, *long terminal repeats*). El cuarto tipo lo representan varias familias de transposones ADN, similares a los transposones bacterianos.

Sobre la base de experimentos iniciados a finales de los 1950s, Howard Temin propuso,



FIGURA 7. Richard J Roberts (n1943) y Phillip A Sharp (n1944).



FIGURA 8. Martin Howard Temin (1934-1994) y David Baltimore (n1938), descubridores de la transcriptasa inversa, compartieron con Renato Dilbecco (n1914) el PNFoM de 1975 «por sus descubrimientos sobre la interacción entre virus tumorales y el material genético celular».

en 1964, la hipótesis del provirus de ADN, que establecía que el genoma de los virus tumorales ARN se copiaba en ADN en la célula infectada; una propuesta que iba directamente en contra del dogma central de la biología universalmente aceptado. En 1970, Temin y David Baltimore (Figura 8), independientemente, descubrieron una enzima viral –transcriptasa inversa– que sintetiza ADN a partir de un molde ARN. Temin concluyó su publicación de 1970 con la aseveración de que sus resultados «constituyen

una prueba relevante de que la hipótesis del provirus de ADN es correcta, y de que los virus tumorales ARN tienen un genoma ADN cuando están en las células y un genoma ARN cuando están como viriones (partículas virales libres). Estos resultados tendrán implicaciones importantes en las teorías de carcinogénesis viral y, posiblemente, en las teorías de transferencia de información en otros sistemas biológicos». Como predijo Temin, el descubrimiento de la síntesis de ADN dirigida por ARN –transcripción inversa– condujo avances importantes en la comprensión del cáncer, de los retrovirus humanos y de los reordenamientos génicos. La transcriptasa inversa ha sido una herramienta crítica en el clonaje de ADN y, con ello, ha impactado en todas las áreas de la biología celular y molecular contemporáneas. La transcriptasa inversa desbancó el que se había denominado, hasta entonces, «dogma central de la biología molecular»: la unidireccionalidad del flujo de información. La ruta: [ADN → ARN → proteína], quedó establecida como: [ADN ↔ ARN → proteína].

Si la replicación del ADN garantiza la continuidad de la información génica de padres a hijos, la transcripción regula la expresión de esa información en cada célula; la transcripción dicta lo que cada célula es. La ARN polimerasa transcribe porciones determinadas –genes– de una de las bandas de la doble hélice de ADN en un polímero lineal similar de bases, el ácido ribonucleico (ARN, que contiene el azúcar ribosa en vez de desoxirribosa, y las bases A, C, G, y U –uridina– en vez de T). El PNFoM de 1959 recayó en Severo Ochoa y Arthur Kornberg por su descubrimiento de los mecanismos de biosíntesis del ARN y del ADN (Figura 9).

La transcripción sigue un principio de complementariedad similar al indicado para la replicación del ADN. En el caso de la transcripción, la complementariedad es similar a excepción de que a una A en la banda de lectura del ADN le corresponderá una U en el ARN complementario; pero a una T en el ADN corresponderá una A en el ARN. Esto es, a la secuencia de bases ATCG en el ADN le corresponde la secuencia de bases UAGC en el ARN. La banda de ADN que es transcrita se denomina banda codificante. Una vez que ha sido completada la síntesis del ARN complementario a un gen determinado, se libera como una molécula lineal sencilla: pre ARN mensajero. Este ARN será procesado y convertido en ARN mensajero que dirigirá la síntesis de la proteína codificada en el correspondiente gen; ello, en un proceso denominado traducción.

La traducción refleja el hecho de que un lenguaje molecular se traduce a otro. El primer lenguaje, el del ADN y ARN, se escribe mediante el orden de las cuatro bases y obedece, esencialmente, a las mismas reglas de complementariedad; ADN y ARN son dialectos de un mismo lenguaje. El lenguaje de proteína es muy diferente al del ADN/ARN y requiere una traducción compleja que convierte la secuencia de bases, en los ácidos nucleicos, en una secuencia de aminoácidos, en las proteínas. Los aminoácidos son completamente diferentes de las bases. Frente a las cuatro bases, veinte son los aminoácidos que forman las proteínas. Dada la gran diferencia estructural entre aminoácidos y bases no debe extrañar que se requiera una compleja maquinaria de traducción para convertir la secuencia de bases de un gen, en otra gramaticalmente correcta de aminoácidos en la correspondiente proteína.

¿Cómo una secuencia de solo cuatro bases en un gen dirige la síntesis de una proteína formada por una secuencia de 20 aminoácidos? La solución a este problema de codificación es que cada aminoácido está especificado por una combinación de tres bases contiguas. El código que relaciona la secuencia de bases en el ARN con la secuencia de aminoácidos en una proteína se denomina código genético. El código genético asigna cada uno de los 20 aminoácidos a un determinado grupo contiguo de tres

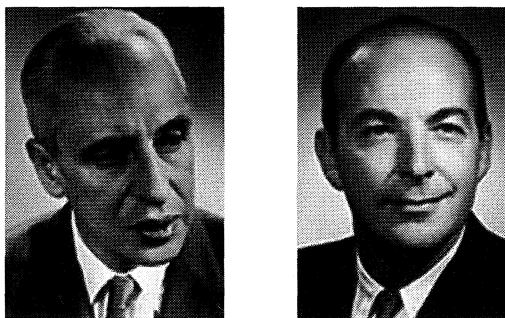


FIGURA 9. Severo Ochoa (1905-1993) y Arthur Kornberg (n.1918) compartieron el Premio Nobel de Medicina y Fisiología, en 1958, «por su descubrimiento de los mecanismos involucrados en la síntesis biológica del ácido ribonucleico y del ácido desoxirribonucleico».



FIGURA 10. Robert W Holley (1922-1993), Har Gobind Khorana (n1922) y Marshall W Nirenberg (n1927) compartieron el PNFoM, en 1968.

bases, tripletes o codones en el ARN. Existen 64 combinaciones posibles de cuatro bases agrupadas en tripletes (4^3). La correspondencia se asegura, primero porque algunos aminoácidos están especificados por más de un triple (degeneración o redundancia del código) y, segundo, porque algunos triples no codifican aminoácidos sino que se utilizan como señales de terminación del proceso de traducción. El PNFoM de 1968 se concedió a Robert W Holley, Har G Khorana y Marshall Nirenberg «por su interpretación del código genético y su función en la síntesis de proteínas» (Figura 10).

Finalmente, debe apuntarse un hecho importante respecto a las diferencias estructurales entre el ADN y las proteínas. Mientras que la secuencia de bases apenas influye en la estructura en doble hélice del ADN, la secuencia de aminoácidos tiene enormes consecuencias sobre la estructura de la proteína. Siguiendo el dictado de la secuencia lineal de aminoácidos (estructura primaria de la proteína) cada proteína se pliega (estructura secundaria) y adopta, finalmente, una estructura tridimensional (estructura terciaria), compleja y distintiva, y esta gran diversidad estructural permite a las diferentes proteínas desarrollar funciones celulares específicas. Ello significa que mientras que posibles cambios de una base por otra (polimorfismo o mutación) no alteran la estabilidad del ADN, la sustitución de un aminoácido por otro puede tener graves consecuencias: enfermedad molecular.

Una mutación es una alteración en la secuencia de bases de un gen. En general, el cambio de base sucede en la región codificante del gen, aunque existen mutaciones que alteran la región reguladora; así, algunas formas de cáncer resultan de mutaciones reguladoras que activan genes involucrados en promover crecimiento celular de manera incontrolada. Las mutaciones codificantes alteran la secuencia de bases en un gen de varias for-

mas; el tipo más simple de mutación consiste en el cambio de una base por otra, mutación puntual que puede tener graves repercusiones en la estructura tridimensional funcional del producto de su expresión. Otras mutaciones afectan mayor número de bases tanto por delección como por inserción. Las mutaciones de uno u otro tipo pueden ser espontáneas y consecuencia de errores durante la replicación del ADN, o inducidas por la acción de diversos mecanismos mutagénicos. Los agentes mutágenos pueden ser físicos como, por ejemplo, la radiación ultravioleta o químicos como diferentes sustancias mutagénicas; en ambos casos, los mutágenos reaccionan con las bases del ADN cambiando una base en otra, o reducen la fidelidad de la replicación del ADN. Una característica común a todos los agentes mutagénicos es que son altamente carcinogénicos.

Pero no todas las mutaciones son deletéreas. El código genético es universal para todas las formas de vida conocidas. Una de las principales implicaciones de esta universalidad es que toda la vida existente en la Tierra deriva de una común forma de vida ancestral, en la que se desarrolló el código genético que hoy utilizan todos sus descendientes. Esta evidencia incontrovertible de un ancestro común de todas las formas de vida confirma una de las predicciones más destacadas de Charles R. Darwin (Figura 11). En *Sobre el Origen de las Especies por Medio de la Selección Natural*, Darwin concluyó que «.. probablemente todos los seres orgánicos que han vivido sobre la tierra descienden de alguna forma primordial en la que se originó la vida». Dado que, en tiempos de Darwin, el conocimiento de las bases genéticas o moleculares de la vida eran escasos, tal hipótesis intuitiva representa uno de los logros, en biología, de mayor alcance intelectual. La evolución es una de las ideas unificadoras en biología.

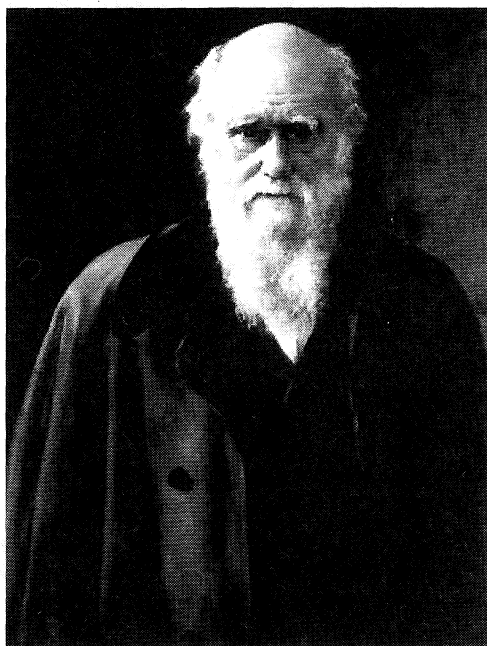


FIGURA 11. Charles Robert Darwin (1809-1882).

Para explicar esa evolución a partir de un ancestro común, Darwin invocó un mecanismo que denominó selección natural. De acuerdo con ello, no todos los organismos tienen las mismas posibilidades de sobrevivir y

reproducirse. Los mejor adaptados tendrán mayor descendencia y sus características se perpetuarán y expandirán en la población. Sin embargo, la teoría de Darwin no explica satisfactoriamente la herencia. Aunque Darwin y Mendel fueron contemporáneos no tuvieron conocimiento el uno del otro; fue necesario que los seguidores de Darwin incorporaran los descubrimientos de Mendel a la teoría de la selección natural. La teoría sintética de la evolución explica la transformación de una especie mediante selección natural (especiación). La evolución es un proceso de cambio; al nivel molecular este proceso implica la inserción, delección o sustitución de bases (mutaciones) en el ADN. Si esas mutaciones ofrecen alguna ventaja se irán acumulando hasta dar lugar a una serie de caracteres bastante diferentes a los originales que, a lo largo de millones de años, irán desembocando en especies diferentes. En ello, los transposones han jugado un papel decisivo.

ADN recombinante

Establecidas las bases de la biología o genética molecular, hubo que esperar algunos años hasta que se dispusieron de herramientas eficaces para manipular la estructura genética; ello, aunque las técnicas de la genética clásica se habían demostrado eficaces en el análisis de los genes y de los cromosomas. El avance más significativo fue el desarrollo, en los 1970s, del ADN recombinante; una técnica mediante la que un fragmento de ADN puede ser seccionado y separado de un genoma donante, e insertado, transplantado o recombinado en otro genoma receptor. Werner Arber, Paul Berg, Daniel Nathans y Hamilton Smith fueron los protagonistas (Figura 12).



FIGURA 12. Werner Arber (n1929), Daniel Nathans (n1928) y Hamilton O Smith (n1931) compartieron el PNFoM, en 1978, «por el descubrimiento de las enzimas de restricción y su aplicación a problemas de genética molecular».

El primer paso en el desarrollo de la tecnología del ADN recombinante fue la caracterización de las endonucleasas de restricción: enzimas que seccionan el ADN en secuencias específicas de bases. Tales enzimas se identificaron en bacterias, donde proporcionan un mecanismo de defensa contra la entrada de ADN extraño (por ejemplo, virus infectivos o bacteriófagos) en la bacteria. Las bacterias disponen de una amplia variedad de endonucleasas de restricción que escinden el ADN en más de cien sitios distintos de reconocimiento. Lugares que constan de secuencias específicas de cuatro a ocho pares de bases dispuestas simétricamente; característica por la que tales secuencias se denominan palíndromos. Palíndromo –proviene de la griega «recurrente»– define una palabra o expresión que resulta lo mismo leída en un sentido que en otro, como «anilina»; también un número capicúa es un palíndromo. En genética se refiere a un fragmento de ADN en el que una secuencia de pares de bases se lee de igual manera en uno u otro sentido a partir de un eje de simetría. Tales secuencias representan el sustrato de las enzimas de restricción que rompen la molécula en el entorno del eje de simetría de las secuencias y en las dos cadenas.

Como las endonucleasas de restricción seccionan el ADN sobre secuencias específicas, tales enzimas pueden utilizarse para romper una molécula de ADN en trozos distintivos. Por ejemplo, una determinada enzima de restricción (*EcoRI*, aislada de la bacteria *Escherichia coli*, un comensal normal de nuestro intestino grueso) reconoce una secuencia característica de seis pares de bases (GAATTC). Esta secuencia se repite cinco veces en el genoma de un cierto virus bacteriano (bacteriófago λ , cuya célula diana es *E. coli*); de este modo, la enzima de restricción rompe el ADN del bacteriófago en seis fragmentos de diferente tamaño. Estos fragmentos pueden separarse según su tamaño utilizando una técnica denominada electroforesis en gel, que separa diferentes moléculas sobre la base de su velocidad de migración en un campo eléctrico y en el que el gel actúa como un tamiz retardando, selectivamente, el movimiento de las moléculas de mayor tamaño. La localización de los sitios de reconocimiento de múltiples y diferentes endonucleasas de restricción se utiliza para generar detallados mapas de restricción de moléculas de ADN, tal como genomas virales. Tras la separación electroforética de los distintos trozos de ADN de restricción, estos pueden recogerse para estudios posteriores, en especial la determinación de las secuencias de bases (secuenciación del ADN) que forman los genomas virales.

Sin embargo, la digestión por endonucleasas de restricción no proporciona la suficiente resolución para analizar moléculas de ADN de mayor tamaño, tal como el genoma humano. Por ejemplo, la enzima *EcoRI* seña-

lada rompe la molécula de ADN con una frecuencia estadística determinada ($1/4^6$ pares de bases). Ello significa que una molécula de ADN algo mayor que el genoma del bacteriófago señalado será rota en diez fragmentos; segmentos de ADN que podrán ser separados e identificados satisfactoriamente con la técnica de electroforesis en gel. Pero la misma enzima originaría más de 500,000 fragmentos al actuar sobre una molécula de ADN humano; tal cantidad de fragmentos es imposible de separar con nitidez mediante la técnica apuntada, que proporcionaría una mancha continua mas que un patrón discreto de fragmentos de ADN reconocibles. En este caso, la obtención de fragmentos puros de ADN se consigue mediante la técnica del clonaje molecular.

La estrategia básica en el clonaje molecular es insertar un fragmento de ADN de interés, en una molécula ADN receptora (vector) capaz de replicarse de manera independiente en una célula hospedadora. El resultado será una molécula recombinante (ADN recombinante) que contendrá el ADN insertado junto a las secuencias del vector. Ello permite obtener grandes cantidades del ADN donante insertado. Por ejemplo, puede clonarse un determinado fragmento de ADN humano en un vector bacteriófago λ . Tal molécula recombinante puede introducirse en la bacteria *Escherichia coli*, hospedador habitual del bacteriófago λ , donde se replica eficazmente originando una progenie de millones de fagos que contienen el inserto de ADN humano. Ese ADN recombinante puede digerirse con la misma endonucleasa de restricción utilizada para obtener el inserto inicial y recuperar los millones de copias producidas de ese inserto (ADN donante, utilizando terminología de trasplante). Ello proporciona la cantidad suficiente del fragmento de ADN humano puro para su posterior análisis y manipulación. También pueden clonarse secuencias ARN. Aquí, el primer paso es sintetizar una copia ADN a partir del molde original ARN; ello, utilizando transcriptasa inversa. El producto, denominado ADN complementario (ADNc, porque es complementario al ARN utilizado como molde), puede ser insertado en un vector ADN siguiendo el esquema descrito en el párrafo anterior. Dada la complejidad de los genes eucarióticos, la posibilidad de clonar ADNc ha sido crítica para comprender la estructura y función de tales genes.

Proyecto Genoma Humano

Su objetivo es, una vez conseguida la secuencia completa del ADN, identificar todos los genes y su función. Ello ha de representar un broche de oro al «conócete a ti mismo» y la llave de la Medicina Molecular. El

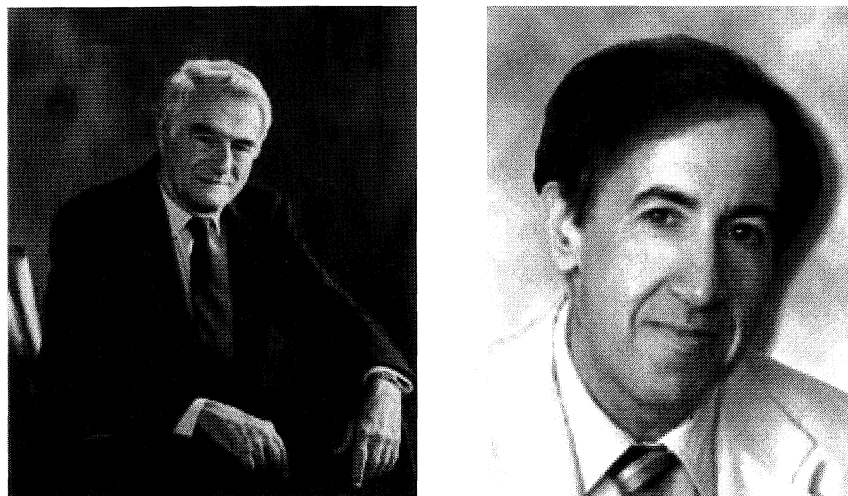


FIGURA 13. Robert Sinsheimer y Charles DeLisi.

Proyecto genoma humano tiene su origen en las iniciativas de Robert L. Sinsheimer y de Charles DeLisi (Figura 13), a mediados de los 1980s. DeLisi –físico y jefe de biomatemáticas de los Institutos Nacionales de Salud de EE.UU– fue el director de la Oficina de Salud y Medio Ambiente del Departamento de Energía, en Washington DC. El departamento –sus raíces alcanzan el Proyecto Manhattan y había financiado investigación sobre los efectos biológicos de la radiación, en especial los referentes a mutaciones génicas– mantenía una División de Ciencias de la Vida en el Laboratorio Nacional de Los Álamos, Nuevo Méjico; aquí, en 1983, estableció una base de datos –«Genbank»– para la informatización de las secuencias de ADN que iban analizándose. DeLisi estaba interesado en aprovechar esos datos para comprender las bases genéticas de las enfermedades humanas. Sinsheimer –biólogo molecular– había declarado, en 1969, que la biología molecular abría nuevas e ilimitadas posibilidades para la humanidad. «Es la primera vez –decía– que una criatura viva conoce su origen y puede diseñar su futuro». En 1977 fue nombrado rector del campus de Santa Cruz de la Universidad de California y, en 1984, estableció un proyecto, en dicha universidad, encaminado a determinar los detalles del genoma humano.

En junio de 1985 en Santa Cruz (California) y en marzo de 1986 en Santa Fe (Nuevo Méjico), hubo sendas reuniones, independientes, lideradas por Sinsheimer y por DeLisi, respectivamente, para discutir los aspectos técnicos de un proyecto para descifrar el genoma humano. Días después de la reunión convocada por DeLisi, Renato Dulbecco, presiden-

te del Instituto Salk (California), declaró (Editorial, *Science* 7 marzo 1986) que la ciencia había alcanzado un punto crucial en la investigación del cáncer y que cualquier avance en tal sentido pasaba por la secuenciación completa del ADN del genoma humano. La relación entre el mapeo génico humano y la biología del cáncer había sido apuntada por M Siniscalco en 1979. Dulbecco, que no había participado en las conferencias de Santa Cruz y Santa Fe, reclamó que los EE.UU. deberían afrontar la secuencia del genoma humano como una empresa comparable, en esfuerzo y en espíritu, al programa que había conducido a la conquista del espacio.

Los acontecimientos se sucedieron a una velocidad vertiginosa. En junio de 1986, Walter Gilbert y Paul Berg copresidieron una sesión sobre el genoma humano en un simposio organizado en Cold Spring Harbor sobre la Biología Molecular del *Homo sapiens*; uno de los participantes, Gilbert, denominó la secuencia completa del genoma humano el «santo grial» de la biología. A esta reunión siguieron otras importantes: en septiembre de 1986, una del Consejo Nacional de Investigación de los EE.UU. (*National Research Council, NRC*), que nominó un Comité para el Mapeo y la Secuenciación del Genoma Humano (*Committee on Mapping and Sequencing the Human Genome, CMSHM*); una segunda, convocada por la Fundación «Instituto de Medicina Howard Hughes», tuvo lugar en agosto de 1987, y, en octubre de ese mismo año, los Institutos Nacionales de Salud de EE.UU. se planteaban su papel en el futuro Proyecto. Tras varias reuniones a lo largo del año 1987, el CMSHN-NCR emitió, en febrero de 1988, un informe: «Mapeo y Secuenciación del Genoma Humano» (*Mapping and Sequencing the Human Genome*), de cuyo «*Executive summary*» puede entresacarse:

«Los humanos llevan tiempo intrigados por las fuerzas que dan forma a ellos y a otros organismos. ¿Qué código dicta el color de los ojos, del pelo o la forma de una flor?. Hace más de 100 años Gregor Mendel descubrió que tales rasgos hereditarios están controlados por unidades celulares que, posteriormente, se conocieron como genes. En años recientes, la comprensión de esos genes ha crecido considerablemente tras el conocimiento de la biología molecular del ADN, la molécula gigante de la que están formados los genes. Ahora es posible obtener la descripción última de los genes y del ADN; ello, desde el reciente desarrollo de las técnicas que permiten mapear (localizar) los genes en el ADN de cualquier organismo y luego secuenciar (ordenar) cada una de las unidades de ADN, conocidas como nucleótidos, que constituyen los genes.

Cuanto más de nuestros genes estén mapeados y sus ADNs secuenciados, dispondremos de una fuente cada vez más útil – una base de datos esen-

cial que facilitará la investigación en bioquímica, fisiología, biología celular y medicina. Esta base de datos tendrá su mayor impacto sobre el cuidado de la salud y la prevención de la enfermedad así como sobre nuestro conocimiento de las células y los organismos. La concepción de organizar un gran proyecto para mapear y secuenciar el ADN en los genes y en las regiones intergénicas que los conectan (el ADN completo o genoma humano) recibe cada vez más atención en el mundo. Varios países han expresado su interés en apoyar tal proyecto. Para evaluar lo que los Estados Unidos de Norte América deberían hacer en esta área, el *NRC* estableció el *CMSHM*, cuyas conclusiones se recogen en este documento, y en donde el Comité explora cómo, cuando y por qué debemos mapear y secuenciar el ADN en el genoma humano. Para llevar a cabo esos objetivos, el comité alcanzó las siguientes conclusiones:

- Conseguir un mapa, una secuencia y un conocimiento cada vez mayores del genoma humano exige una acción especial que deberá organizarse y establecerse para ese propósito. Tal esfuerzo especial en las próximas dos décadas potenciará significativamente el progreso en biología y medicina humanas.
- Los problemas técnicos asociados con el mapeo y la secuenciación de los genomas humano y de otros organismos son lo suficientemente grandes para que un programa científico como este requiera un esfuerzo diversificado y sostenido para mejorar nuestras posibilidades para analizar las complejas moléculas de ADN. Aunque las capacidades necesarias no existen todavía, la manera en que deben desarrollarse parece evidente. Las tecnologías avanzadas requeridas emergerán de un esfuerzo común que recalque proyectos piloto y desarrollo tecnológico. Una vez establecidas, estas tecnologías no solo harán factible completar el análisis del genoma humano y otros, sino que también harán contribuciones importantes a muchas otras áreas de la biología básica y la biotecnología.
- Los objetivos iniciales más importantes se esforzarán en adquirir un mapa de alta resolución de ligamiento génico del genoma humano, una colección de clones ordenados de ADN y una serie de mapas físicos complementarios de resolución progresivamente creciente. El objetivo último será obtener la secuencia nucleotídica completa del genoma humano, comenzando a partir del material de la colección de clones ordenados de ADN. La consecución de este objetivo requerirá mayores (pero alcanzables) avances en el manejo y en la tecnología de secuenciación del ADN.
- Un enfoque genético comparativo es esencial para interpretar la información en el genoma humano. Más aún, deberán llevarse a cabo en paralelo estudios intensivos de aquellos organismos que proporcionen modelos particularmente útiles para comprender la estructura, función y evolución génica humana.
- El esfuerzo para mapear y secuenciar comenzará como una serie de programas encaminados a mejorar el desarrollo tecnológico, competitivos y

revisados por pares. Los fondos deben incluir tanto ayudas individuales como a grupos de mediano tamaño interdisciplinares de científicos e ingenieros.

- El proyecto genoma humano deberá diferir de la marcha de la investigación actual en cuanto que los subproyectos deben incrementar en un factor de 5 a 10 sus eficacias de mapeo, secuenciación, análisis o interpretación de la información.
- Los progresos hacia los objetivos antes expuestos requerirán el establecimiento de instalaciones centralizadas bien dotadas, incluyendo un banco central para los fragmentos de ADN clonados generados en el esfuerzo de mapeo y secuenciación y un centro de datos para la colección computarizada y la distribución de grandes cantidades de información de las secuencias de ADN. El comité sugiere que los grupos que soporten tales servicios sean seleccionados mediante concurso abierto.

Sobre la base de esas conclusiones y en vista de la importancia y magnitud de la tarea, el comité recomienda unos fondos adicionales de \$200 millones anuales, que no deberán retrotraerse del presupuesto federal actual para investigación en ciencias biomédicas. Por su parte, la mayoría del comité recomienda que una agencia federal única deberá liderar el proyecto. Esta agencia deberá recibir y administrar los fondos para el proyecto y deberá ser responsable de las operaciones del banco central y del centro de datos, así como administrar el sistema de revisión por pares utilizado para determinar los beneficiarios de los fondos. Deberá trabajar en estrecha colaboración con un Consejo Asesor Científico (*Scientific Advisory Board, SAB*) para desarrollar e implementar un alto estándar en la revisión por pares. El *SAB*, compuesto en principio por expertos científicos de renombre en campos relevantes, deberá asesorar no solo en cuanto a la revisión por pares sino sobre el control de calidad, cooperación internacional, coordinación de esfuerzos entre los laboratorios y las operaciones del banco central y los centros de datos...».

El «*Executive summary*» concluye: «El Comité cree firmemente que debe emprenderse un proyecto para mapear y secuenciar el genoma humano. Son evidentes las implicaciones éticas, sociales y legales de tal esfuerzo, pero está convencido de que pueden afrontarse adecuadamente. El proyecto incrementaría enormemente nuestro conocimiento y comprensión de la biología humana y permitiría un rápido progreso para incidir en el diagnóstico y, finalmente, en el control de muchas enfermedades humanas. Puede vislumbrarse que el proyecto conduciría, también, al desarrollo de nuevas tecnologías y a la producción de mapas y de secuencias de diversos organismos, lo que proporcionaría información relevante de la mayor importancia para mejorar nuestro conocimiento de toda la Biología».

En octubre de 1988 se creaba la Oficina para la Investigación del Genoma Humano de los Institutos Nacionales de Salud, bajo la presidencia de James Watson y con el mandato de llevar a cabo el Proyecto Genoma Humano. Watson, *Mr DNA*, movilizó vocaciones y recursos en EE.UU., Europa y Japón logrando el establecimiento de un Consorcio Internacional para el desarrollo del proyecto. «Cómo primer proyecto de *Gran Ciencia* en biología –comentaba DeLisi– el mapeo y desciframiento de la secuencia completa del ADN humano estimulará la investigación en muy diversos campos, desde la tecnología de las computadoras a la química teórica». Escudriñar el genoma humano –en el caso del Consorcio Internacional un mosaico de material genético correspondiente a linfocitos y espermatozoides de seis a diez individuos anónimos– representa una colosal empresa que, sobre la información proporcionada por los cariotipos de las diferentes cromosopatías, aberraciones cromosómicas o anatomía mórbida cromosómica se ha abordado combinando tres estrategias sucesivas y a la vez solapantes: mapas genéticos, mapas físicos y secuenciación.

Los cromosomas pueden teñirse y examinarse microscópicamente revelando patrones distintivos de bandas claras y oscuras (mapa cromosómico). Por ejemplo, la tinción de cromosomas metafásicos con el colorante Giemsa (bandeo G) muestra bandas claras que contienen genes constitutivos y específicos de tejidos, SINEs y regiones de ADN ricas en GC, y bandas más oscuras que contienen menos genes, más LINEs y menor contenido en GC. Las aberraciones cromosómicas contribuyen de manera significativa a malformaciones congénitas, siendo responsables de >50% de todos los abortos espontáneos. Además, el 0.7% de los neonatos –y el 2% de los nacidos vivos en mujeres >35 años– presentan anomalías cromosómicas significativas. También numerosos cánceres resultan de este tipo de anomalías. El análisis citogenético de los cromosomas en linfocitos periféricos ha facilitado la identificación de numerosas anomalías cromosómicas. En tales análisis los cromosomas son detenidos en metafase y teñidos con Giemsa. El bandeo G proporciona 350-550 bandas por genoma haploide; 1 banda corresponde a 5-10 millones bp y acoge entre 1 y varios genes. La numeración de las bandas de cada cromosoma procede del centrómero al telómero de cada uno de los brazos. Los cromosomas se ordenan –cariotipo– por pares y de acuerdo con su tamaño, a la vez que pueden identificarse por su patrón de bandeo. Los cariotipos pueden revelar deleciones, duplicaciones y reorganizaciones cromosómicas, y pueden ayudar a localizar la región de ADN responsable de una enfermedad génica. A parte de tales aberraciones, en ocasiones pueden detectarse, a través del microscopio, una serie de variaciones cromosó-

micas; sirvan de ejemplo el aspecto desenrollado del brazo largo del cromosoma 1, los cromosomas en anillo, las piezas cromosómicas extra o los mosaicismos.

Las preparaciones de cromosomas intactos pueden utilizarse, también, para el diagnóstico genético molecular, el análisis de deleciones y duplicaciones génicas y el mapeo de genes en el genoma. La hibridación *in situ* fluorescente (FISH) utiliza sondas de cADN marcadas con sustancias fluorescentes, con lo que bajo el microscopio de fluorescencia puede determinarse la localización relativa de una secuencia particular de ADN; ello permite la localización de la sonda cADN en regiones subcromosómicas que la técnica de bandeado cromosómico no puede determinar. El análisis estándar por FISH proporciona resoluciones de 5-10 Mb y sirve de puente entre el análisis citogenético estándar y el estudio molecular detallado utilizando ADN purificado. Utilizando sondas multicolor y cromosomas interfásicos, en los que el ADN es menos compacto, puede obtenerse una panorámica con una resolución de 100 kb.

Por otro lado, en ocasiones puede cuantificarse la cantidad o dosis de una enzima codificada por un determinado gen (dosis génica). Es posible mostrar, ante la deleción de una porción cromosómica, que la cuantía de una determinada enzima es, solo, la mitad de la cuantificada en controles, o ante una duplicación cromosómica que los niveles son el 150% de los controles. De esta manera se localizaron los genes de la fosfatasa ácida (c2), adenilato quinasa (c9) o superóxido dismutasa (c21). Sin embargo, uno de los descubrimientos más sorprendentes ha sido que células somáticas de especies diferentes, cuando crecen el mismo medio de cultivo y en presencia de ciertos promotores, funden produciendo células híbridas. Sobre la base de que las células híbridas pierden progresivamente y durante las sucesivas divisiones celulares cromosomas de una de las especies, es posible adjudicar diferentes actividades enzimáticas a cromosomas determinados.

Dado el tamaño y la complejidad del genoma humano, la identificación y localización de los genes no es trivial. Los mapas genómicos intentan facilitar la localización de un gen de interés. Dos tipos de mapas (genéticos o físicos) describen el orden de los marcadores y las distancias entre ellos en los cromosomas. Uno de los objetivos del Proyecto genoma humano es incrementar la resolución de esos mapas hasta que se conozca la localización exacta de cada uno de los genes. Un mapa genético determina la posición relativa de un gen o de un locus sobre la base de las frecuencias de recombinación relativas a otros loci en el mismo cromosoma. Cualquier secuencia polimórfica cuyo patrón de herencia pueda seguirse, es útil para mapeo genético. Los marcadores polimórfi-

cos más utilizados son las repeticiones de secuencias simples (satélites) y los cambios mononucleotídicos. El mapa genético se construye asignando la frecuencia con que dos marcadores se heredan conjuntamente mediante estudios de ligamiento genético. La localización cromosómica de un gen puede determinarse rastreando la herencia de marcadores polimórficos.

La frecuencia observada de recombinación entre dos loci es una función de su distancia y se expresa en centimorgan (cM); ello, en recuerdo de Thomas H Morgan, un genetista que estudió el ligamiento génico y estableció el concepto de que la frecuencia de recombinación varía como una función de la distancia entre dos locus génicos. Si la frecuencia de recombinación entre dos loci es del 1%, ambos loci distan 1 cM. Por tanto, el mapa genético se construye calculando, en estudios de ligamiento, la frecuencia con que dos marcadores se heredan juntos. Ligamiento genético es la tendencia de los genes a heredarse conjuntamente como resultado de su localización sobre el mismo cromosoma. Dado que la frecuencia de recombinación incrementa como una función de la distancia genética, cuanto más próximos se encuentren dos loci mayor posibilidad habrá de que no recombinen y que, por el contrario, se hereden juntos (ligamiento génico). Un conjunto de marcadores lo suficientemente próximos para que se hereden conjuntamente definen un haplotipo. A efectos de identificar un locus cromosómico que segregue con una enfermedad es necesario determinar el genotipo de muestras de ADN de varios miembros de varios pedigrís. Luego, puede determinarse si ciertos alelos marcadores cosegregan con la enfermedad.

Los marcadores utilizados en la confección de los mapas genéticos corresponden a diferentes polimorfismos de secuencias de ADN que se distribuyen por todo el genoma. Un polimorfismo ADN es una alteración en una determinada secuencia ADN que se observa con una frecuencia >1% en una población determinada. A diferencia de las mutaciones, estas alteraciones en la secuencia no condicionan efectos adversos en la función de los genes y pueden considerarse variaciones neutras en la secuencia ADN. Los polimorfismos son importantes porque permiten rastrear un gen –refiriéndolo a una región polimórfica cromosómica determinada– en pedigrís. Los polimorfismos más importantes son los debidos a variaciones en la longitud de secuencias simples iterativas, y a variaciones de un nucleotido en una posición dada (polimorfismos mononucleotídicos, *single nucleotide polymorphisms* - SNPs). El *International SNP Map Working Group* anunció, en noviembre de 2000, un mapa de más de un millón de SNPs distribuidos a través del genoma humano, que proporciona una densidad de secuencias de un SNP cada 1.0 kb.

Una clase especial de polimorfismo es el polimorfismo de longitud del fragmento de restricción (*restriction fragment-length polymorphism*, RFLP). Si el cambio de una base ocurre en un sitio de reconocimiento para una endonucleasa de restricción, puede destruir la secuencia de reconocimiento de esta enzima y abolir dicho sitio de reconocimiento; alternativamente, el polimorfismo puede crear un sitio nuevo de restricción. Polimorfismos de este tipo modifican el tamaño de los fragmentos que resultan de una digestión con esa enzima. El tamaño alterado de los segmentos de restricción puede detectarse mediante electroforesis en gel. Si un determinado sitio de restricción ha sido destruido, la banda polimórfica correspondiente al fragmento de restricción estudiado será mayor, y si se ha generado uno nuevo, la banda será de menor tamaño. Tales polimorfismos proporcionan «huellas dactilares» del ADN.

Por su parte, el mapa físico es la distancia de hecho (expresada en pares de bases –pb– de secuencia ADN) entre genes. En términos físicos, 1 cM equivale, aproximadamente, a un millón de pares de bases (1 Mb). El mapa físico refleja la ordenación y distancias entre genes; un mapa que tiene varios niveles de resolución. Un mapa físico de baja resolución indica qué cromosoma ubica un gen particular. El uso de técnicas como la hibridación *in situ* (FISH) –antes citada– permite determinar la localización (cartografía o mapeo) de un determinado gen en un cromosoma. Utilizando diferentes sondas marcadas con diferentes distintivos fluorescentes es posible «pintar» los cromosomas y demostrar las localizaciones relativas de los diferentes genes, aunque ello sigue significando un bajo nivel de resolución. Mayor resolución se consigue clonando trozos de ADN en diferentes vectores de clonaje y estimando las distancias mediante el solapamiento de los fragmentos. El acortamiento de los fragmentos incrementa la resolución. En una primera fase los grandes fragmentos de restricción se clonan en cromosomas artificiales de levadura (YACs); luego se fragmentan en otros más pequeños que se clonan en cromosomas artificiales de bacterias (BACs), cósmidos, fagos y, por último, en plásmidos; ello, de tal manera, que un segmento de ADN queda cubierto por una serie de clones contiguos (*contigs*). La utilización de sitios etiquetados por la secuencia (*sequence tagged sites*, STSs) como unidad estándar para la confección del mapa físico, ha representado una gran ventaja. Los STSs sirven como contraseñas que permiten solapar fragmentos clonados y disponerlos en el mismo orden que ocupan en el genoma. Los STSs constan de 200-500 pb.

El objetivo final del Proyecto genoma humano es conocer la secuencia completa de los pares de bases que componen el ADN nuclear humano: la secuenciación del genoma. Sin embargo, la tecnología actual no permite la secuenciación directa del ADN completo de un cromosoma; es necesario

manejar trozos lo suficientemente pequeños representados por los BACs. Sobre la base de este esquema general, se dispone de dos estrategias de secuenciación: «abordaje jerárquico» (*hierarchical shotgun*) utilizado por el Consorcio Internacional, y «abordaje genómico global» (*whole genome shotgun*) aplicado por *Celera Genomics*. En el primer caso se escinde el ADN de cada cromosoma de manera ordenada y en fragmentos cada vez más pequeños (YACS BACs cósmidos) lográndose el reensamblaje de las secuencias de acuerdo con el orden relativo –conocido sobre la base de los diferentes marcadores genéticos localizados en el mapa genético, en especial STSs– de cada uno de los fragmentos. La segunda estrategia escinde aleatoriamente el ADN en pequeños segmentos que, tras su clonaje, se preparan para su secuenciación; la reconstrucción se logra mediante el solapamiento de las secuencias de los extremos de los fragmentos.

En 1975, Frederick Sanger –que había recibido el PNQ de 1958 por su trabajo sobre la estructura de las proteínas, especialmente la insulina– anunciaba que había desarrollado un método para determinar eficazmente el orden de los pares de bases en un genoma; por su parte, Alan Maxam y Walter Gilbert desarrollaron, de manera completamente independiente y el mismo año, un método de secuenciación diferente (Figura 14). Pocos años después, numerosos grupos habían logrado automatizar el proceso. El primer prototipo práctico fue desarrollado por un equipo en el Instituto Tecnológico de California, en 1986. Este prototipo fue rápidamente convertido en un instrumento comercial y puesto en el mercado en 1987. El crecimiento de la capacidad de secuenciación ha sido explosivo. En 1976 un investigador era capaz de secuenciar 5 kb a lo largo de un año; la capacidad de secuenciación conseguida por *Celera Genomics* es de 100,000 kb / día, lo que ha hecho que la máquina PRISMA 3700 (Perkin-Elmer) sea una pieza clave de su proyecto.



FIGURA 14. Paul Berg (n1926), Walter Gilbert (n1932) y Frederick Sanger (n1918) compartieron el Premio Nobel de Química, en 1980, «por sus contribuciones referentes a la determinación de las secuencias de bases en los ácidos nucleicos».

Si los secuenciadores son las herramientas claves, la bioinformática es el cerebro del Proyecto, principalmente en el caso de *Celera Genomics*. Las 230 máquinas PRISMA 3700 –completamente robotizadas en sus funciones– que dispuso Celera vomitan un flujo continuo de datos –señales que representan las bases adenina, citosina, guanina y timina– que, a través de una red de fibra óptica, han estado alimentando un computador desarrollado por Compaq expresamente para Celera: un sistema en paralelo con una capacidad de computación de 1.3 teraflop (tera = 10^{12}) o 1.3 billones de operaciones por segundo, que hacen de este artificio la computadora civil más potente del mundo. Solo tiene un rival: ASCI Red, construida por Intel para el gobierno USA y con el objetivo de modelar explosiones nucleares. Tan fantástico apoyo computacional se debe a la exigencia de la estrategia seguida por Celera –opuesta a la del consorcio internacional– de secuenciar segmentos aleatorios de ADN que deben ser correctamente ordenados.

El 26 de junio de 2000, en una Conferencia de prensa pactada tras duras negociaciones entre el Consorcio Internacional para la Secuenciación del Genoma Humano y *Celera Genomics* se presentó, en la Casa Blanca, un «borrador de trabajo» (*working draft*) del Genoma Humano. El acto lo encabezó el presidente William J Clinton; fueron coprotagonistas J Craig Venter, presidente de Celera y Francis Collins, director del Consorcio Internacional (Figura 15), y fueron testigos representantes de los países copartícipes en el Proyecto: Tony Blair, Primer Ministro del Reino Unido por videoconferencia y los distintos embajadores en presencia física:

«El consorcio público Proyecto Genoma Humano da a conocer, hoy, la conclusión de un *borrador de trabajo* de la secuencia del genoma humano – la huella genética del ser humano–. Este borrador incluye dos tareas: colocar largos fragmentos de ADN en orden correcto para completar todos los cromosomas, y determinar la secuencia de ADN de esos fragmentos. El ensamblaje que hoy se da a conocer consta de fragmentos solapantes que cubren el 97% del genoma humano, representando un 85% del genoma la secuencia hasta ahora ensamblada.

La secuenciación se disparó durante el último año; el 60% de la secuencia se ha conseguido en los últimos seis meses. Durante este tiempo el consorcio ha proporcionado mil bases de secuencia bruta por segundo –7 días a la semana, 24 horas cada día–. La calidad media del «borrador de trabajo» supera con mucho las expectativas originales del consorcio para este producto intermedio. Los centros del consorcio han proporcionado mayor cantidad de secuencia que la esperada.

Consecuencia de todo ello es que el «borrador de trabajo» está mucho más cerca de la versión final de lo que el consorcio había previsto para este mo-

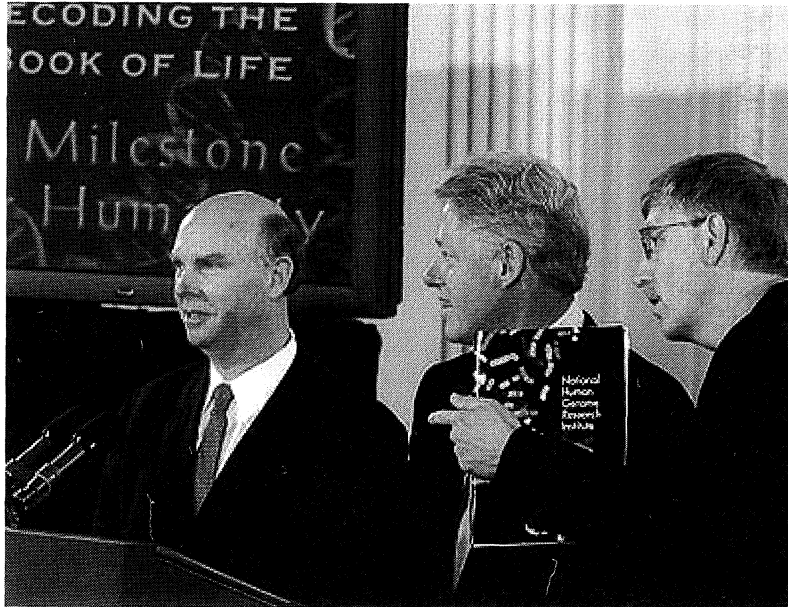


FIGURA 15. J Craig Venter y Francis Collins, «apadrinados» por William Jefferson. Clinton (n1946), anunciaron, el 26 de junio de 2000, un borrador de trabajo del genoma humano

mento. Aproximadamente el 50% de la secuencia del genoma se encuentra en forma casi definitiva, y el 24% lo está definitivamente. La secuencia genómica está organizada en segmentos contiguos de, aproximadamente, 200,000 bases. La fiabilidad media de la secuencia de ADN en este ensamblaje es del 99.9 %. La información de la secuencia por parte del proyecto público ha sido continua, inmediata y de libre disposición, sin restricciones para su utilización o redistribución. La información es consultada a diario por científicos de la academia y de la industria.

Hasta este momento ya han sido identificadas algunas decenas de miles de genes a partir de la secuencia del genoma. El análisis de la secuencia disponible muestra los 38.000 genes presupuestos y confirmados por evidencia experimental. Hay muchos miles de predicciones génicas adicionales que han de ser comprobados experimentalmente. Docenas de genes involucrados en enfermedades han sido identificados accediendo al borrador de trabajo.

El objetivo del consorcio para la primavera de 2000 fue producir una versión tipo borrador de trabajo de la secuencia humana, un ensamblaje que contiene fragmentos solapantes que cubren aproximadamente el 90% del genoma y que se secuencia en forma de borrador de trabajo; ello es, con algunas lagunas y ambigüedades. El objetivo último del consorcio es producir una secuencia completamente finalizada; en otras palabras, sin lagunas y con el 99.9% de fiabilidad. La fecha para este último objetivo había

sido el año 2003, pero a la vista de los resultados actuales es completamente seguro que la obtención de la secuencia definitiva se acortará significativamente.

En un anuncio paralelo, *Celera Genomics* anunció hoy que ha completado su propio ensamblaje de la secuencia del ADN humano. Los proyectos público y privado utilizan automatismos y tecnología de secuenciación similares, pero emplean distintos enfoques para secuenciar el genoma humano. El proyecto público usa una estrategia denominada «abordaje jerárquico» (*hierarchical shotgun*); el proyecto de Celera utiliza un enfoque denominado «abordaje genómico global» (*whole genome shotgun*). El abordaje jerárquico tiene la ventaja de que la localización global de cada secuencia individual se conoce con certeza, pero requiere construir un mapa de largos fragmentos que cubra el genoma. El abordaje global no requiere este paso, pero tiene otros requerimientos en la fase de ensamblaje. Ambas estrategias alinean la secuencia a lo largo de los cromosomas humanos utilizando mojones localizados en el mapa físico producido en el Proyecto Genoma Humano.

La producción de secuencias más allá de lo esperado ha ido pareja de una sorprendente cosecha de variaciones génicas humanas –los llamados polimorfismos nucleotídicos simples o SNPs–. El Proyecto Genoma Humano tenía el objetivo de descubrir 100,000 SNPs antes del año 2003. Hasta ahora, con las secuencias ensambladas y otros datos acumulados por el Consorcio SNP, los científicos ya han identificado más de 300,000 SNPs y seguramente dispondrán hasta un millón de SNPs a finales del año 2000. Los SNPs proporcionan una poderosa herramienta para estudiar las enfermedades y la historia humanas. La secuenciación –determinar el orden exacto de las cuatro bases químicas del ADN denominadas A, T, C y G– se ha conseguido en el Proyecto Genoma Humano merced a los avances tecnológicos en el desciframiento del ADN y a la naturaleza colaborativa del esfuerzo, que ha incluido 1,000 científicos de casi todo el mundo que han trabajado juntos con eficacia.

El Consorcio internacional para la secuenciación del genoma humano incluye científicos de 16 instituciones en Francia, Alemania, Japón, China, Gran Bretaña y Estados Unidos. Los cinco centros principales están localizados en: *Baylor College of Medicine*, Houston, Texas; *Joint Genome Institute* en Walnut Creek, CA; *Sanger Centre* cerca de Cambridge, Inglaterra; *Washington University School of Medicine*, St. Louis; y *Whitehead Institute*, Cambridge, Massachusetts. Juntos, estos cinco centros han generado el 82% de la secuencia.

El proyecto ha sido tan estrechamente coordinado que ninguna región del genoma ha sido desatendida a la vez que se han minimizado las duplicaciones. Los participantes en el consorcio internacional se han adherido a los estándares de calidad del proyecto y a la política de publicidad de los datos. El proyecto se financia mediante ayudas de las agencias gubernamentales.

mentales y fundaciones privadas en varios países, que incluyen el Instituto Nacional para la Investigación del Genoma Humano de los Institutos Nacionales de Salud y el Departamento de la Energía, de EE UU, y el *Wellcome Trust* en Inglaterra.

El coste global de la secuenciación del borrador de trabajo ha sido, aproximadamente, de 300 millones de dólares USA; de los que, aproximadamente, 150 millones han sido financiados por los Institutos Nacionales de Salud de EE UU. El coste de la secuenciación del genoma humano se refiere frecuentemente a tres mil millones de dólares. Sin embargo, esa cifra se refiere a la estimación inicial del Proyecto Genoma Humano para un periodo de 15 años (1990-2005) y para una amplia gama de actividades científicas relacionadas con la genómica, que incluyen estudios de enfermedades humanas, organismos experimentales (como bacterias, levaduras, lombrices, moscas y ratón), desarrollos de nuevas tecnologías para investigación biomédica, métodos computacionales para el análisis de genomas y aspectos éticos, legales y sociales relacionados con la genética».

El Consorcio Internacional y Celera representan dos maneras de hacer ciencia: burocracia y ortodoxia de planteamientos frente a iniciativa e innovación. El Consorcio inició su marcha en 1990, Celera entró en la competición en 1998 con una inversión de \$300 millones. Celera anunció que lograría la secuencia completa del genoma humano en tres años, adelantándose en cuatro a los objetivos del Consorcio; ello, sobre la base del análisis de los datos que el Consorcio facilitaba «día a día» en bancos de datos de libre acceso, de una nueva estrategia de secuenciación que obvia una parte de la cadena metodológica simplificándola y abaratándola y que había sido testada en la secuenciación de genomas más simples (Celera publicó el genoma de *Drosophila* en marzo de 2000), y de la utilización de máquinas ultrarrápidas –secuenciadores automatizados y computadoras–, de novísima generación.

El Proyecto Genoma Humano ha sacado a la superficie un serio dilema político sobre la base de una serie de conflictos de intereses. Temeroso de que Celera patentara la secuencia, el Consorcio reaccionó. El *Wellcome Trust*, inglés, incrementó de inmediato su aporte al proyecto comprometiéndose a que el *Sanger Centre* completara un tercio de la secuencia; por su parte, EE. UU. consolidó su esfuerzo y creó cuatro supercentros de secuenciación. El Consorcio, además, replanteó sus objetivos que desplazó desde la consecución de una secuencia definitiva a producir un borrador del 90% del genoma para la primavera de 2001, la fecha anunciada por Celera. Un año después, el Consorcio revisó sus plazos a la baja señalando la primavera de 2000 como tope; ello, a efectos de adelantarse a los intereses de Celera. Llegando juntos a la meta de esta prime-

ra etapa, el Consorcio y Celera han pactado el reconocimiento del borrador como patrimonio de la humanidad. A partir de ese momento comenzó una segunda etapa, más dura, cuya meta es la transferencia de secuencias concretas al mercado del diagnóstico médico: los chips de ADN. En resumen, el genoma humano es para el Consorcio Internacional un bien público o un bien común, para *Celera Genomics* es una fuente de beneficios.

El Proyecto Genoma Humano se centra, ahora, en convertir el «borrador de trabajo» en una «obra definitiva». Ello se conseguirá rellenando las lagunas existentes en la secuencia presentada e incrementado la fiabilidad de la secuencia global hasta el 99.99%. Aunque la versión denominada «borrador de trabajo» es útil para la mayor parte de la investigación médica, una secuencia tan perfecta como sea posible es crítica para obtener toda la información contenida en los datos de la secuencia humana. Ello ya se ha conseguido para los cromosomas 21 y 22, así como para el 24% del total del genoma. La secuenciación del c21, el más pequeño de los cromosomas humanos, ha deparado un hallazgo inesperado: la pobreza de genes en el cromosoma. Si el número total esperado de genes para el genoma humano era, aproximadamente, 100000, la predicción para el c21 era 800 a 1000 genes. Sin embargo, el consorcio internacional para c21 ha encontrado, solamente, 225 genes. Los genes c21 más los correspondientes a c22 –el primer cromosoma humano secuenciado– son 770; si tal es la tendencia en el resto de los cromosomas, el número total de genes del genoma humano sería, aproximadamente, 35000; una cifra muy lejana de los 80000-100000 esperados.

La cifra de partida se debe a Walter Gilbert, uno de los pioneros de la genómica quién, en los 1980s, estimó en, aproximadamente, 100000 los genes humanos; ello sobre la base de las bases por el secuenciadas en segmentos de ADN. A finales de los 1990s, Craig Venter barajó 50000-80000 genes, e *Incyte Pharmaceuticals* y *Human Genome Sciences*, otras dos firmas norteamericanas que entraron en la «arena» genómica, subieron la cifra a 120000-150000. Por otro lado, el número de genes aceptado a la baja ha deparado otra «noticia»: que la diferencia cuantitativa génica no es tan dispar entre las distintas especies. Esta estimación a la baja –penuria genómica– no es tan sorprendente. La dotación genómica (14000 genes) de la *Drosophila melanogaster* es inferior a la dotación (19000 genes) del *Caenorhabditis elegans*, aunque la mosca es más compleja que el gusano. La contestación debe estar en relación con el proteoma. Es cada vez más evidente el papel que el procesamiento alternativo del ARN juega en expandir la diversidad proteómica, lo que puede explicar la discrepancia aparente entre el número de genes y la complejidad del organismo. El procesamiento (*splicing*) alternativo puede generar más

transcriptos que el correspondiente al número de genes en un genoma dado. Además, una secuencia codificante puede comenzar a leerse en distintas bases; tal transcripción o lectura alternativa incrementa las posibilidades de expresión de una secuencia dada. El desarrollo de un catálogo completo de los transcriptos alternativos (transcriptoma) de un genoma y determinar su función, será el principal reto de la era post-genómica. Otra vez, un dogma biológico ha sido cuestionado; ahora se plantea la revisión del esquema estándar: [un gen → una proteína].

Con todo, cabe recordar las impresiones de Ralph Waldo Emerson –recogidas en *Science* de 22 de diciembre de 2000– tras visitar, en 1833, el Gabinete de Historia Natural del Jardín Botánico de París: «Los límites de lo posible se han ampliado... y lo real se nos hace más extraño que lo imaginario». Para *Science*, «la explosión en la secuenciación genómica» hace actual la expresión de Emerson y «justifica su elección como Acontecimiento Científico del año (*Breakthrough of the Year 2000*)». Cualquiera puede apreciar su inmenso poder y potencialidad; también su asombrosa complejidad y su imperfección. Poco después de la publicación formal de la *secuencia dorada*, Tony Blair declaró: «Hoy somos testigos de una revolución en la ciencia médica cuyo impacto puede superar al que supuso el descubrimiento de los antibióticos». Justo cinco días antes, el temor a la creación de un «gueto genético» estremeció Londres cuando una de las compañías aseguradoras más importantes de Inglaterra reveló que había utilizado ilegalmente datos de análisis genéticos experimentales para evaluar algunas peticiones de seguros. Sólo de manera gradual, mediante la investigación meticulosa y concienzuda, podrá clarificarse el verdadero impacto social y médico de la ciencia genética.

Historia del Proyecto Genoma Humano

1865

Mendel expone su trabajo ante la Sociedad de Historia Natural de Brünn.

1900

«Descubrimiento» del trabajo de Mendel por de Vries, Correns y von Tschermak-S.

1902

Boveri y Sutton demuestran la presencia de pares de cromosomas en especies diploides.

1905

Bateson denomina «genética» a la ciencia de la herencia.

1909

Johannsen introduce la palabra «gen».

- 1911**
Morgan postula la base cromosómica del ligamiento genético.
- 1928**
Griffith descubre un «principio transformador» en *Diplococcus pneumoniae*.
- 1941**
Beadle y Tatum establecen el «dogma» «un gen - una enzima».
- 1944**
Avery, MacLeod y McCarthy demuestran que el «principio transformador» en pneumoco es ADN.
- 1950**
McClintock propone los «elementos transportables».
- 1952**
Hershey y Chase demuestran que el material genético del bacteriófago T2 es ADN.
- 1953**
Watson y Crick publican la estructura de doble hélice de ADN sobre la base de los datos de difracción de rayos X de Wilkins y de la regla de correspondencia de bases de Chargaff.
- 1956**
Tjio y Levan establecen que la dotación cromosómica normal diploide de la especie humana es 46.
- 1958**
Kornberg aísla la ADN polimerasa I.
- 1959**
Ochoa aísla la RNA polimerasa.
- 1964**
Se establece la colinearidad entre genes y productos polipeptídicos sobre la base de los trabajos de Yanofski *et al* y de Brenner *et al*.
Temin propone la hipótesis la formación de provirus ADN a partir de virus tumorales ARN.
- 1966**
Nirenberg y Khorana completan el código genético.
- 1970**
Nathans y Smith aíslan la primera endonucleasa de restricción.
Baltimore identifica la transcriptasa inversa en virus tumorales ARN.
- 1972**
El laboratorio de Berg consigue la primera molécula de ADN recombinante *in vitro*.
- 1977**
Breathnach, Mandel y Chambon y Jeffreys y Flavell, demuestran intrones en genes eucarióticos.
Maxam y Gilbert y Sanger, Nicklen y Coulson publican técnicas de secuenciación de ADN.
Sanger *et al* publican la secuencia completa (5387 nucleótidos) del fago FX174.
- 1978**
Se descubre procesamiento (*splicing*) de ARNs derivados de adenovirus en tres laboratorios diferentes.
- 1980**
Botstein y Davis y Skolnick y White proponen un método para mapear la totalidad del genoma humano sobre la base de RFLPs.

1982

Wada propone la utilización de robots y la secuenciación automática.

1984

Científicos del MRC (UK) descifran la secuencia completa del virus Epstein-Barr (170 kb).

1985

Sinsheimer, Rector de la Univ. California en Santa Cruz, organiza una reunión para discutir la viabilidad de secuenciar el genoma humano.

Mullis *et al* desarrollan la PCR.

Sanger *et al* publican la secuencia completa (48502 pares de bases) del fago lambda.

1986

Brenner urge a la Unión Europea a elaborar un programa colaborativo para mapear y secuenciar el genoma humano.

DeLisi, del Departamento de Energía (DOE) de EE.UU., organiza una reunión para discutir la estrategia para secuenciar el genoma humano.

Dulbecco apoya la secuenciación del genoma humano.

La secuenciación del genoma humano se debate en una reunión sobre «La Biología Molecular del *Homo sapiens*» en *Cold Spring Harbor Laboratory*.

Hood y Smith anuncian el primer prototipo de secuenciadora automática de ADN.

DeLisi inicia el estudio del genoma humano consiguiendo, para el año, 1987 \$5.3 millones a tal efecto.

1987

Gilbert crea *Genome Corp.*

Burke, Olson y Carle construyen cromosomas artificiales de levadura (YACs) como herramientas de clonaje, incrementando x10 el tamaño del inserto.

Donis-Keller *et al* publican el primer mapa genético con 403 marcadores.

1988

La *National Science Foundation* (EE.UU.) endorsa el Proyecto Genoma Humano al que dota con \$200 millones.

Wyngaarden, director de los Institutos Nacionales de la Salud (NIH, EE.UU.), reclama el protagonismo de ellos en el Proyecto.

Se celebra la 1.ª Reunión anual sobre el Genoma Humano.

Watson acepta dirigir la Oficina para la Investigación del Genoma Humano de los NIH. Establece que el 3% de los fondos se destinarán al estudio de aspectos éticos y sociales.

Los NIH y el DOE firman un acuerdo de colaboración sobre el Proyecto Genoma Humano.

1989

El Comité Asesor para el ADN recombinante de los Institutos Nacionales de la Salud de EE.UU recomienda la aprobación del primer trasplante génico en humanos (terapia génica).

Tsui, Collins *et al* clonan el gen de la fibrosis quística.

La oficina dirigida por Watson es elevada de categoría: Centro Nacional para la Investigación del Genoma Humano (NCHGR).

1990

Los NIH y el DOE publican un plan de cinco años. Los objetivos incluyen una mapa genético completo, un mapa físico con marcadores cada 100 kb y la secuencia de 20 Mb de ADN en organismos modelo, para el año 2005.

Los NIH comienzan un ambicioso programa a gran escala a efectos de secuenciar los genomas de *Mycoplasma caporicolum*, *Escherichia coli*, *Caenorhabditis elegans* y *Saccharomyces cerevisiae*.

Los NIH y el DOE señalan el uno de octubre como fecha del inicio del Proyecto.

Lipman y Myers publican el algoritmo BLAST para alinear secuencias.

1991

Craig Venter anuncia una estrategia alternativa de secuenciación.

Japón inicia la secuenciación del genoma del arroz.

1992

Tras una discusión con Healey, directora de los NIH, con motivo de la patente de genes, Watson dimite como director del NCHGR.

El *Wellcome Trust*, de Inglaterra, entra en el Proyecto aportando \$ 95 millones.

Simon *et al* desarrollan cromosomas artificiales bacterianos (BACs) para clonaje.

Cohen *et al*, del Centro para el Estudio de los Polimorfismos Humanos (CEPH, París) y *Généthon* mapean el cromosoma 21.

Equipos de EE.UU y de Francia completan mapas genéticos murino (marcadores espaciados 4.3 cM) y humano (marcadores espaciados 5 cM).

1993

Francis Collins es nombrado director del NCHGR.

Los NIH y el DOE revisan los objetivos para el periodo 1993-98, que incluyen secuenciar 80 Mb de ADN para finales de 1998 y completar el genoma para 2005.

El *Wellcome Trust* y el Consejo de Investigación del Reino Unido (NRC) abren el Centro Sanger, el mayor laboratorio de secuenciación del consorcio internacional.

1994

Murria, Cohen *et al* publican un mapa completo de ligamiento génico del genoma humano, con marcadores espaciados 0.7 cM.

1995

Venter y Fraser y Smith publican la primera secuencia de un organismo vivo: *Haemophilus influenzae*, 1.8 Mb.

El gobierno japonés aporta \$15.9 millones para garantizar el trabajo de varios grupos de secuenciación durante los 5 años siguientes.

Investigadores del Instituto Whitehead y de *Généthon* publican un mapa físico del genoma humano conteniendo 15000 marcadores.

1996

Affymetrix comercializa chips de ADN.

Un consorcio internacional publica el genoma completo del *S. cerevisiae*.

1997

El NCHGR es promovido a Instituto Nacional para la Investigación del Genoma Humano.

Blattner, Plunkett *et al* publican la secuencia completa del ADN de *E. coli*: 5 Mb.

1998

Japón, EE.UU., UE, China y Corea del Sur establecen una colaboración internacional para secuenciar el genoma del arroz.

Green, Swing *et al* publican un programa (PHRED) para interpretar automáticamente datos de secuenciación.

Venter anuncia la creación de una compañía, Celera Genomics, para secuenciar el genoma humano en tres años y con un capital de \$300 millones.

El *Wellcome Trust* (UK) y los NIH y el DOE (EE.UU) reaccionan ante el anuncio de Venter incrementando los fondos y redefiniendo los objetivos: conseguir un «borrador

de trabajo» de la totalidad del genoma humano para el año 2001, y completarlo en 2002-03.

Sulston, Waaterson *et al* completan la secuencia genómica de *C. elegans*.

1999

Los NIH adelantan la conclusión de los datos al año 2000, y lanzan un proyecto para secuenciar el genoma del ratón en 3 años y con un presupuesto de \$130 millones.

Investigadores ingleses, japoneses y norteamericanos completan la secuencia del cromosoma 22.

2000

Celera e investigadores académicos secuencian el genoma (180 Mb) de la mosca *Drosophila melanogaster*.

El Consorcio Internacional para el genoma humano publica la secuencia completa del cromosoma 21.

En una ceremonia en la Casa Blanca, el Consorcio Internacional y Celera anuncian la consecución de un «borrador de trabajo» del genoma humano.

Un consorcio internacional publica la secuencia completa de un vegetal: *Arabidopsis thaliana*.

En diciembre fracasan los planes para una publicación conjunta del «borrador de trabajo» por parte del Consorcio Internacional y de Celera.

2001

El Consorcio Internacional publica su trabajo en *Nature* (15 de febrero) y Celera en *Science* (16 de febrero).