# EL PODER DE LA GENÓMICA

Ricardo Ramos Ruiz
Responsable Técnico Unidad de Genómica Antonia Martín Gallardo
Parque Científico de Madrid

## INTRODUCCIÓN: EL ORIGEN DE LA GENÓMICA

La especie humana se ha planteado el problema de la herencia genética desde el principio de su historia, en los primeros asentamientos como agricultores y ganaderos. Dominar la naturaleza ha supuesto mejorar las especies silvestres a conveniencia, siendo el único método disponible el cruce de unas variedades con otras. En ese momento el objetivo era conseguir las plantas de recolección en los momentos adecuados, las variedades más adaptadas a cada entorno, los animales de mayor producción de carne o leche, etc. Todas estas actividades demuestran el profundo conocimiento, cuando menos intuitivo, que hemos tendido desde siempre sobre las reglas de la herencia. Sin embargo, no ha sido hasta mediados del siglo XIX cuando los trabajos de Mendel comenzaron a sistematizar las leyes de la Genética que desembocaron casi un siglo más tarde en el descubrimiento de las moléculas que gobiernan la transmisión genética, gracias a los trabajos de Watson, Crick y Franklin. Esta molécula, presente en todas nuestras células, se denomina ácido desoxirribonucleico, **ADN**, la molécula de la herencia.

El ADN es un polímero biológico, lo cual quiere decir que está compuesto por unidades sencillas unidas de forma lineal hasta formar cadenas no ramificadas, en este caso de hasta miles de millones de unidades de longitud. En realidad, no es una sola cadena lineal sino dos cadenas (anti)paralelelas, que adoptan la forma de dos hebras entrelazadas. Solamente hay cuatro constituyentes en la composición del ADN, y son muy semejantes entre sí. Se denominan nucleótidos (desoxi-nucleótidos) y están siempre compuestos por un grupo inorgánico de fosfato, un azúcar de cinco átomos de carbono y un grupo orgánico sencillo llamado base nitrogenada que distingue entre sí a los cuatro nucleótidos posibles (las conocidas bases A, C, G o T). Precisamente la asociación de nucleótidos entre las dos hebras enfrentadas es lo que posibilita que el ADN sea el material genético ya que cada posición en la cadena de un nucleótido define cuál es el nucleótido localizado en la cadena opuesta denominada complementaria por ese motivo. Es decir, una cadena se puede copiar en base a la información de la complementaria, y eso permite que una célula reparta el material genético durante la división celular: cada célula hija será capaz de recomponer todo el material genético partiendo sólo de la mitad.

La enorme simplicidad química de la cadena de ADN es lo que hizo tan difícil pensar que pudiera ser por sí sola la molécula de la herencia, ya que una información de tal envergadura debía contener información muy detallada capaz de generar todas y cada una de las estructuras que nos convierten en seres tan complejos, y una secuencia lineal no parecía suficientemente informativa.

Los mismos nucleótidos que forman parte de nuestro material genético son los que componen el DNA de cualquier otro ser vivo, incluido el de las bacterias más simples. Nuestro material genético sólo se distingue del de otros organismos en el orden en el que se sitúan los nucleótidos en cada cromosoma. Sin duda tal observación sólo se puede explicar en base a que la vida ha aparecido una sola vez en nuestro Planeta y que todos los seres vivos compartimos un único modelo de herencia. Pensemos en una bacteria como *Escherichia coli* que se enfrenta

durante su ciclo vital a la tarea de interpretar a su material genético. Nunca va poder distinguir si su ADN tiene un origen propiamente bacteriano o si procede de una fuente diferente, como por ejemplo humana, ya que son químicamente exactos. La Biotecnología es toda una Ciencia que ha surgido basada en la manipulación del material genético, especialmente dirigida a aprovechar la facilidad y rapidez de crecimiento de los microorganismos en los que se han introducido artificialmente fragmentos definidos de ADN exógeno. En un caso como el descrito, la bacteria utilizará la información genética añadida para producir productos génicos humanos, y lo hará en cantidades grandes y de fácil recuperación. Nuestras bacterias "recombinantes" (así se denominan las bacterias quiméricas constituidas por material genético propio y material exógeno) se convierten en verdaderos micro-reactores biológicos dirigidos a la producción de nuestras proteínas de interés.

## LA INTERPRETACIÓN DE LA INFORMACIÓN GENÉTICA

El material genético no es una molécula pasiva: tiene la capacidad única de promover la generación de un nuevo ser vivo. Es esencial conocer cómo partiendo de ese conjunto de nucleótidos podemos llegar a desempeñar funciones bioquímicas, funciones tan esenciales y diversas como queramos imaginar: comienzan por la construcción de células y tejidos, pero incluyen también el desarrollo de todas las reacciones bioquímicas que desempeñamos: la síntesis de todo tipo de componentes activos como vitaminas, hormonas o moléculas de reserva energética, la capacidad de movimiento, la organización de la lucha contra agentes patógenos…, la lista sería interminable.

Todas las funciones del ADN como portador de información genética se basan en la lectura ordenada de los nucleótidos que lo componen. Por eso, conocer la identidad completa de un ADN supone conocer con detalle el orden exacto en el que se colocan los nucleótidos a lo largo de la cadena.

Podemos considerar al ADN como un libro de instrucciones donde están detalladas las instrucciones sobre todas las funciones que somos capaces de desempeñar. La interpretación desde la información escrita en el ADN hasta las moléculas efectoras (en general, las proteínas) se realiza en dos etapas. En la primera, denominada **transcripción**, cada fragmento de ADN se copia a una molécula intermediaria de **ARN**, químicamente muy similar al ADN, también compuesta por una cadena lineal de nucleótidos (ribo-nucleótidos en este caso) pero radicalmente distinta en cuanto a función. Tras la transcripción, la información en "formato ARN" sigue codificada, de una manera muy ingeniosa: cada tres nucleótidos (un triplete) se van a interpretar como un código para un solo componente de las proteínas (un aminoácido). El paso de descodificación del ARN para la síntesis de proteínas se conoce como **traducción** y se realiza en orgánulos especializados denominados ribosomas. Por lo tanto, las proteínas son también polímeros biológicos, pero sus componentes aminoacídicos son más numerosos (en torno a 20) y químicamente mucho más diferentes unos de otros, de forma que la molécula final de proteína tiene toda una gama de residuos químicamente activos capaces de adquirir estructuras tridimensionales muy variadas y complejas capaces de catalizarlas reacciones bioquímicas.

## LOS GENES

Cada uno de los procesos individuales que somos capaces de realizar está gobernado por una entidad genética individual, que denominamos un **gen**. La definición exacta de lo que es un gen cambia según se considere desde un punto de vista genético o bioquímico, pero siempre hace referencia a la estructura básica o mínima de la información genética, la unidad de transmisión de la herencia. En este sentido se esbozó el concepto de << un gen – una proteína >>, que hace referencia a que un gen consiste en una secuencia de tripletes que determinan qué aminoácidos componen una proteína determinada. De forma general, esta proteína será la encargada de realizar la función final, y el gen recibirá el mismo nombre que el de la proteína efectora.

Considerando al material genético como el lugar donde se albergan los genes, nuestro ADN será una simple sucesión de genes localizados unos detrás de los otros. Así es sin duda, pero el ADN no es una yuxtaposición de genes sin solución de continuidad. Por el contrario, la información genética está dispersa a lo largo de la molécula. Esto quiere decir que suele haber una larga separación entre genes contiguos, y al menos en los organismos más evolucionados, los propios genes presentan su información interrumpida, es decir los tripletes que forman parte de cada gen se ensamblan en una secuencia lineal en el ARN, pero suelen estar fragmentados en secuencias parciales separadas a veces por decenas de kilobases (miles de nucleótidos) en el ADN. En la figura 1 se muestra un ejemplo de cómo es la organización de un gen dentro de un cromosoma humano.

En el ser humano, el número total de genes está calculado en torno a unos 20,000. Teniendo en cuenta que el tamaño medio de un ARN que codifica una proteína puede ser de unas pocas kilobases, resulta que la cantidad de material genético encargado de codificar aminoácidos es ciertamente reducida en comparación con los aproximadamente tres mil millones de bases que componen el DNA humano. Por supuesto hay una serie de señales genéticas adicionales en el ADN y el ARN, que marcan a la célula dónde empiezan y terminan los procesos de transcripción y traducción, señales que regulan la eficacia de esos procesos en respuesta al entorno celular, señales que indican a la célula dónde debe comenzar y acabar la replicación del DNA durante la división celular. Tales señales, que también son secuencias lineales de DNA, no explican sin embrago la gran cantidad de material genético que aparentemente no tiene una información precisa y que llevó a acuñar el término de DNA basura en referencia a esta falta de función, como si fuera un DNA sobrante. Actualmente no se cree que el DNA no implicado en traducción, que puede suponer un altísimo porcentaje del material genético total de algunos organismos, sea un DNA inútil. Probablemente participa en funciones más sutiles relacionadas con los mecanismos de recombinación, aportando material genético que permita mantener una variabilidad suficiente a nivel de especie, controlando una excesiva tasa de mutación en las regiones codificantes más sensibles o aportando flexibilidad para la generación de nuevas variantes, entre otras.

Una parte del ADN no codificante sí se procesa a la molécula intermedia de ARN aunque luego no se traduzca a proteína, ya que ejerce su labor directamente así. Se trata especialmente de los ARNs ribosomales y de transferencia que participan como tales en la composición del ribosoma y en la descodificación de los tripletes de nucleótidos en aminoácidos, pero también en toda una larga serie de RNAs de pequeño tamaño que participan en los procesos de maduración de los RNAs codificantes y de los recientemente descubiertos reguladores celulares conocidos como micro-RNAs.

## EXPRESIÓN GÉNICA

No todos los genes están “activos” en todos los momentos y en todas las situaciones. Tal acción causaría que todas las células y tejidos se comportaran exactamente igual, lo cual está absolutamente alejado de la realidad. Todas las células de un organismo sí portan en su núcleo el mismo material genético necesario que codifica todo el catálogo de genes propio de su organismo pero cada célula solamente produce las proteínas que va a utilizar en función de sus necesidades. Algunas de las proteínas se necesitan exclusivamente en un determinado momento del desarrollo embrionario o adulto, otras sólo en respuesta a la falta o sobreabundancia de alimento, otras en presencia de patógenos, otras en respuesta a tratamientos con fármacos, otras se generarán específicamente en unos tipos celulares pero nunca en otros, y así consecutivamente. El proceso de síntesis efectiva de una proteína a partir del gen que la codifica se conoce como **expresión génica**, y el grupo de genes que se expresa en una situación o momento determinados se conoce como **perfil** de expresión génica, y es un marcador de la situación fisiológica de una célula.

Como ya se ha indicado, el efector final suele ser la proteína, pero una parte esencial del proceso de regulación desde el ADN hasta la proteína suele establecerse en la fase de transcripción. En este sentido, la colección de RNAs presentes en una célula refleja en buena medida el estado en el que se encuentra, su capacidad de respuesta y las funciones activas en cada momento. En el caso de la expresión génica, no sólo es relevante que un ARN se transcriba o no, sino especialmente la cantidad presente de un ARN específico en el interior celular. Mayores niveles de ARN se asocian con mayores niveles de proteínas y un consecuente incremento en la función codificada por dicho gen. En este sentido, el ARN (y por supuesto las proteínas) son un material variable y dinámico, en el mismo sentido que podemos considerar al ADN como un material estático.

Cada una de las técnicas de la Genómica que describiremos a continuación centrará su estudio en uno u otro tipo de material genético según necesite abordar una cuestión de identidad (estática) o definir un estado fisiológico (dinámico). Vayamos a un ejemplo práctico. En el diagnóstico de cáncer de mama y ovario, podemos encontrarnos con casos de enfermos con un claro componente hereditario. El riesgo a contraer esta enfermedad puede provenir de la sustitución de un único nucleótido por otro en la cadena de ADN. Ese cambio modificará el aminoácido que se incorpora en la proteína en una posición determinada afectando la función de la proteína haciéndola más activa, (o impidiendo su actividad, según cada caso), alterando finalmente el ritmo de división celular.

La evaluación de los riesgos genéticos hereditarios adquiridos se basa en el análisis de variantes génicas asociadas a enfermedad, que se analizan en el ADN y por lo tanto están presentes en cualquier célula del paciente (suelen utilizarse por facilidad las células nucleadas de la sangre). Los genes implicados más directamente en la transmisión de cáncer familiar de mama y ovario se conocen como BRCA1 y BRCA2 y el estudio de alteraciones genéticas es puramente cualitativo ya que las únicas opciones son presencia o ausencia de mutación. Sin embargo, en cáncer de mama esporádico (no familiar), el proceso es mucho más complejo ya que se basa en la desregulación de la expresión génica que acaba ocasionando el crecimiento del tumor. El diagnóstico, dirigido en este caso a la clasificación del tipo de tumor y la previsión de su comportamiento y progresión, se basa en los perfiles de expresión de diversos genes (aumento o reducción de su concentración efectiva en la célula), diferentes de BRCA 1 y BRCA2. Los perfiles de expresión génica relacionados con la progresión tumoral son mucho más difíciles de evaluar, ya que se basan en múltiples genes actuando en coordinación cuya expresión se ve

afectada por las células presentes y los patrones de crecimiento celular. Por eso, y a pesar de su enorme potencial, la práctica clínica está aún comenzando a evaluar la utilidad de los perfiles de expresión génica como sistema de diagnóstico. Sin duda se abre un futuro muy prometedor a corto y medio plazo para los actuales sistemas de alta capacidad, que describiremos más adelante.

El análisis de la expresión génica

La naturaleza del ARN como pieza clave del sistema de regulación es en sí mismo un reto para definir con certeza perfiles de expresión. El RNA es una molécula lábil que se degrada con mucha facilidad en disolución y que está sometida al ataque de muchas enzimas celulares que la célula posee justamente para poder someter al ARN a un control eficaz. La forma en que se comenzaron a visualizar perfiles de expresión de ARN de forma sistemática fue mediante la técnica denominada de *Northern blot* (que toma su nombre por analogía al análisis de DNA conocida como *Southetrn blot*). En esta técnica, las moléculas de ARN presentes en un lisado celular se hacen migrar a través de un material semisólido donde se separan en función de su tamaño y se identifican mediante el uso de sondas radiactivas específicas (véase una representación en la figura 2). Como cada proteína posee un número de aminoácidos definido, el número de tripletes de cada ARN codificante y por lo tanto su longitud total será diferente de los demás. Encontramos ARNs de tamaño inferior a una kilobase mientras que otros llegan a longitudes diez veces superiores. La migración de cada especie identifica a cada especie de ARN individual y la intensidad de la señal radiactiva es una medida de la cantidad (nivel de expresión) presente tras un determinado tratamiento. La comparación respecto al estado basal nos dirá si un gen concreto está inducido, reprimido o estable por causa de dicho tratamiento, proporcionando además una medida aproximada de los niveles de modulación, en términos relativos (“número de veces”)

Se han desarrollado otros sistemas de análisis de perfiles de expresión génica además del *Northern blot*, pero la verdadera revolución vino de la mano del desarrollo de la técnica conocida como **PCR** (por sus siglas en inglés *polymerase chain reaction*; la reacción en cadena de la polimerasa). Esta reacción fija una pequeña zona diana (de forma habitual, una porción informativa dentro de un gen) que se copia mediante el mecanismo de replicación de ADN. El producto de cada etapa de amplificación sirve como sustrato para la siguiente reacción de replicación, lo cual ocasiona una reacción exponencial en cadena donde el fragmento escogido se multiplica miles de millones de veces, hasta hacerse visible por técnicas sencillas de biología molecular. Pocos científicos podían sospechar que la investigación en bacterias extremófilas de hace apenas 30 años iba a proporcionar herramientas tan poderosas y de tanta utilidad para el avance de la biología molecular. El uso de polimerasas de replicación procedentes de estos organismos permitió automatizar el proceso de PCR ya que estas enzimas soportan sin pérdida de eficacia los rangos de temperaturas extremas a las que se someten las muestras de ADN a lo largo de los ciclos de PCR. Mediante el uso de la enzima Transcriptasa-Inversa, propia de los retrovirus, que copia cualquier ARN en el ADN molde para la reacción de PCR, esta técnica se puede adaptar para el análisis de la expresión génica tanto de ADN como de ARN, permitiendo que la amplificación por PCR de fragmentos específicos, sea un técnica absolutamente rutinaria en cualquier laboratorio de Genética Molecular.

La reacción de PCR descrita no deja de ser una técnica semi-cuantitativa, que se ha logrado derivar hace algunos años una adaptación conocida como PCR a tiempo real que la transforma en una técnica verdaderamente cuantitativa. La PCR a tiempo real o qPCR permite determinar de forma individual o colectiva qué genes y en qué magnitud exacta se ve afectada la

expresión génica en respuesta a un cambio biológico. Su única limitación, por ser una técnica basada en PCR, es que necesita conocer previamente la secuencia del gen estudiado para diseñar los reactivos con los que conseguir amplificar de forma única el fragmento de interés. En humanos, gran parte del problema quedó resuelto gracias al esfuerzo internacional dirigido a conocer la identidad profunda de nuestro material genético, conocido como Proyecto Genoma Humano. La secuenciación de nuestro genoma completo atravesó un momento muy importante entre los años 2000 y 2003 con la publicación del primer borrador de nuestra identidad genética. La información desvelada gracias a este proyecto constituye una herramienta fundamental con la que poder sondear la presencia o cantidad de cualquier gen, tanto los que ya se conocían en ese momento como todas las nuevas variantes que se han ido describiendo desde entonces.

## ANÁLISIS DE GENOMAS COMPLETOS

El conocimiento sistematizado de nuestro material genético ha permitido, coordinado con los avances tecnológicos, avanzar un paso de gigante en la definición de perfiles genéticos con el desarrollo de la tecnología de *microchips* conocidos como "*microarrays*". Estos dispositivos contienen reactivos específicos para la colección completa de todos los genes de un organismo, dispuestos en una superficie semejante al de una diapositiva o el portaobjetos de un microscopio. El ARN total se incuba con la colección completa de sondas y se obtiene una señal fluorescente, más intensa cuantas más copias hubiera de cada especie de ARN. En un tiempo reducido se puede hacer, a un coste accesible, un barrido completo de todos los genes en busca de cambios asociados al sistema estudiado. Este tipo de análisis es radicalmente distinto de la forma de trabajar que se había venido utilizando. Hasta entonces, los investigadores proponían una hipótesis en la que consideraban si un cambio biológico podía influir sobre el nivel de expresión de un gen o de un cierto número de genes, lo cual no deja de ser el planteamiento de una teoría, siempre limitada por la capacidad de análisis y la exactitud de las premisas asumidas. Al no restringirse el objeto de estudio, se posibilita el hallazgo de asociaciones imprevistas entre enfermedades y perfiles de expresión, de enorme interés para el avance del conocimiento.

Por ejemplo, en el caso mencionado de los perfiles de expresión asociados a cáncer de mama esporádico, algunos de los genes que pueden participar en el diagnóstico están asociados con proceso de división o migración celulares (como cabe esperar en relación a los fenómenos de proliferación o diseminación celular), pero otros son completamente inesperados, y no se hubieran descubierto sin haber empleado sistemas de análisis global.

## LA VARIACIÓN HUMANA

La segunda gran aportación del proyecto Genoma Humano ha sido la determinación de la estructura primaria de todo el material genético no implicado directamente en expresión génica. Como ya se ha indicado, la mayor parte de nuestro ADN tiene una función esencialmente desconocida. Tenemos material genético estructural que forma parte de los centrómeros y telómeros de nuestros cromosomas, tenemos regiones repetitivas largas y regiones repetitivas cortas dispersas a lo largo de todo nuestro ADN, tenemos probablemente material genético residual procedente de antiguas infecciones retrovirales y muchas señales adicionales que participan en los mecanismos de replicación, reparación, transcripción y traducción de los ácidos nucleicos. Todas esas secuencias se hicieron públicas y accesibles gracias al Proyecto Genoma Humano.

Algunas conclusiones que se pudieron alcanzar cuando se completó la secuenciación de nuestro genoma son:

- El tamaño total del Genoma Humano es de $3{,}2 \times 10^9$ bases (3 Gb). No es en absoluto el genoma de mayor tamaño.

- Sólo un 5 % de nuestro genoma codifica proteínas. Existe un 25 % de nuestra ADN muy vacío de genes ("DNA desierto").

- Somos idénticos en un 98 % a animales próximos como los chimpacés; tenemos centenares de genes semejantes a las bacterias.

- Existe un 99,99 % de identidad de persona a persona (lo cual indica también que encontraremos miles de nucleótidos diferentes entre individuos distintos).

- Se calculó la existencia de no más de 30,000 genes y nos los 100,000 que se suponía hasta entonces. Actualmente ese número parece un poco exagerado y se estima más en torno a 20,000 (aunque ya hemos comentado la dificultad de definir con exactitud qué es un gen). Sin embargo, se piensa que podemos llegar a sintetizar más de 200,000 proteínas distintas, lo cual quiere decir que en promedio cada gen sería capaz de codificar 10 proteínas diferentes, por distintos mecanismos. La versatilidad de los genes en modular la expresión y actividad de los productos génicos es la principal y apasionante tarea que se ha abierto tras el conocimiento de la estructura primaria del Genoma.

El proyecto Genoma Humano se desarrolló utilizando material genético procedente de muy pocos individuos, que lógicamente no puede dar cuenta de toda la variabilidad posible representativa de nuestra especie. Cada individuo tiene unas características externas propias: aspecto exterior, capacidad física, distinta susceptibilidad a enfermedades, etc, que denominamos su **fenotipo**. En muchas ocasiones esas diferencias se pueden basar en cambios localizados con claridad en el material genético (**genotipo**). Las diferencias genéticas entre individuos pueden ser tan sutiles como sustituciones de un nucleótido por otro (los denominados SNPs, por *single nucleotide polymorphisms*), pequeñas sustituciones o inserciones (denominadas *indels*) o cambios de mayor envergadura que suponen reorganizaciones grandes en el ADN. De hecho, los análisis basados en *microarrays* han demostrado que la presencia de repeticiones o duplicaciones génicas de fragmentos relativamente grandes es un proceso mucho más frecuente de lo esperado en el material genético de cada uno de nosotros.

De forma semejante a los estudios en expresión génica, se ha realizado un esfuerzo enorme en describir variantes asociadas a procesos biológicos o que supongan un incremento del riesgo de contraer determinadas enfermedades. Inicialmente estos estudios verificaban si una variante dentro de un gen, siempre previamente definida, se asociaba a un cierto fenotipo. La puesta a punto de la tecnología de *microarrays* para el estudio de variantes en el ADN también ha revolucionado la forma de analizar la variación genética. A día de hoy se pueden correr *microchips* donde se analizan cambios de millones de polimorfismos en paralelo o se pueden correr *microchips* que cubren regiones próximas dispuestas a lo largo de todos los cromosomas, en busca de cambios estructurales grandes en el ADN. De nuevo, un experimento relativamente

sencillo con reactivos universales proporciona una información detallada sobre alteraciones a lo largo de todo el genoma sin que hayamos tenido que reducir nuestras regiones objeto de estudio en función de premisas previas basadas en una información parcial.

## LA NUEVA GENERACIÓN DE LA GENÓMICA

Todos estos estudios dirigidos a conocer cómo funciona y cómo se regulan las funciones del ADN a nivel de genoma completo han llevado al surgimiento de una nueva disciplina denominada **Genómica**, que más allá de las reglas de la Genética, se centra en el estudio de los genomas en su conjunto, su regulación y el control de su actividad.

El progreso de la Genómica se ha basado en el desarrollo de las técnicas de análisis, comenzando por una técnica básica que se desarrolló ya en los años 70 del siglo XX, conocida como **secuenciación**, y que consiste en la identificación de las cadenas de ADN, base a base. En sus inicios la técnica era laboriosa y no se podían alcanzar altas tasas de productividad, pero aun así permitió alcanzar retos tan importantes como identificar genes de forma individual (el primero, ya en el año 1973) y más adelante conocer su estructura, describir familias de genes, encontrar asociaciones filogenéticas, etc. Se descubrieron y purificaron las enzimas que controlan la digestión y reparación del ADN, las encargadas de la replicación, transcripción y traducción, se avanzó en el conocimiento de los mecanismos de expresión y síntesis de proteínas, se aprendió a modificar moléculas de ADN y a introducirlas y extraerlas de sistemas biológicos con facilidad. Dichos logros sentaron la base de la genética moderna y el acceso a las técnicas de Biotecnología basadas en la generación de moléculas de ADN recombinante para múltiples aplicaciones en la industria y en la clínica.

El siglo XXI ha traído lo que se puede considerar la revolución en el campo de la Genómica. Un punto clave ha sido un avance significativo en el conocimiento y acceso a las secuencias de genomas completos, incluido por supuesto el humano, pero no restringido a nuestra especie, incluyendo el de otros mamíferos como rata o ratón, otros sistemas modelo como la mosca de la fruta *Drosophila melonagaster*, la planta *Arabidospsis thaliana*, el nematodo *Caenorabditis elegans*, levaduras, pez cebra, y un creciente número de organismos. En paralelo, la puesta a punto de nuevas tecnologías como la PCR a tiempo real y los *microarrays*, junto con el desarrollo de capacidades de análisis informático más poderosas, han posibilitado abordar todo tipo de proyectos. Algunos avances muy importantes han sido por ejemplo la determinación de las bases genéticas de algunas enfermedades raras o el diagnóstico de enfermedades hereditarias o esporádicas comentadas con anterioridad, pero hay muchas más. Se están determinando las variantes genéticas que suponen un riesgo para muchas enfermedades (incluidas enfermedades tan complejas como la diabetes o los trastornos cardiovasculares), los genes y variedades que mejoran las capacidades de supervivencia o de producción de muchas plantas, los perfiles de expresión asociados a muy diversos estímulos farmacológicos…; las lista es, afortunadamente, interminable.

Sin embargo, una vez más estas poderosas técnicas de análisis de alta productividad se han vuelto a ver superadas por una nueva generación de sistemas conocidos como *next-generation sequencers* o Secuenciadores de Genomas. El primer prototipo de estos equipos data del año 2005. pero no ha sido hasta el 2008 cuando esta tecnología se ha comenzado a expandir y generalizarse en la comunidad científica. Los Secuenciadores de Genomas son sistemas de altísima productividad capaces de obtener en una sola carrera tanta información como la

contenida en todo un genoma humano completo. Funcionan directamente a partir del ADN sin necesidad de generar productos intermedios estables, lo cual acelera tremendamente la velocidad de obtención de resultados. Se ha estimado que el plazo de años que se debía dedicar a la preparación y secuenciación de un genoma complejo se ve reducido a un plazo de algunos meses gracias a esta nueva tecnología, y con unas garantías de calidad cuantificables.

Pero quizá la ventaja primordial de esta tecnología es que no necesita información alguna de la secuencia que determina para poder generar resultados. Las técnicas anteriores se han ido construyendo paso a paso: una pequeña secuencia original servía para poder avanzar e ir desentrañando las secuencias contiguas. En el caso de los Secuenciadores Genéticos, no es necesario ningún conocimiento previo sobre el material genético inicial, ni importa si está mezclado ni en qué proporción, con otros ADNs: se secuencia todo el ADN presente y solamente después se ensamblan los fragmentos contiguos hasta recomponer las cadenas originales. La aplicación más directa de esta tecnología es sin duda la secuenciación de genomas desconocidos, por técnicas mucho más productivas que las utilizadas con el Genoma Humano, y se aplica con mucho éxito a la secuenciación de sistemas mixtos, como puede ser la flora microbiana presente en el sistema digestivo de los animales o la población de hongos que crecen en un ecosistema determinado. Por otro lado, también es una técnica de altísima capacidad dedicada a re-secuenciación, que permite determinar y cuantificar todos los cambios genéticos de una persona respecto a la secuencia de referencia con enorme profundidad y eficacia.

Existe una corriente en la Comunidad Científica que discute la idoneidad de las secuencias de referencia de los sistemas modelo, afirmando que están construidas con individuos aislados y por lo tanto representan una variante exclusiva y restringida del genoma de dichos organismos (por no contar con los posibles errores de secuenciación, inconsistencias, imprecisiones, carencias, etc.). En el caso del Genoma Humano, la secuencia publicada se basó en un número muy reducido de personas que no pueden dar cuenta de toda la variabilidad de la especie humana. Se han desarrollado proyectos específicos de análisis de variación en relación a distintos genotipos, como el proyecto HapMap, que ha canalizado el análisis con mucha profundidad de un gran número de genes. Sin embargo, no ha sido hasta la llegada de la nueva generación de secuenciadores cuando se han podido lanzar proyectos de análisis de la variabilidad genética humana a escala global, como son el proyecto de secuenciación de 1,000 genomas que va a analizar muestras de ADN de individuos dispersos por todas las regiones del Planeta, o la secuenciación masiva de las alteraciones genéticas de todos los tipos de cánceres descritos, en busca de perfiles comunes y por supuesto dirigidos a diseñar las mejores estrategias terapéuticas.

No se puede, finalmente, olvidar una crítica que surge como contrapartida a los proyectos de secuenciación masiva. Durante el último tercio del siglo XX la cantidad de secuencias depositadas en las bases de datos públicas fue relativamente pequeña en comparación con la que se está generando en la actualidad, pero fueron secuencias de mucha capacidad informativa. Con esto queremos decir que eran secuencias de ADN que venían acompañadas por un enorme trabajo de fondo que permitían conocer si correspondían o no a genes expresados, en este caso qué variantes de ARN codificaban, qué proteína se podía sintetizar con esa secuencia y en qué tejidos estaban presentes, qué variantes existían en otras especies, qué señales de transcripción o traducción eran operativas y en qué condiciones, con qué otros genes se regulaba su actividad, qué mutaciones producían cambios de actividad significativos, etc. Había todo un trabajo de investigación básica que daba un valor biológico esencial a la mera secuencia. La carrera por secuenciar ha desbordado literalmente las bases de datos, pero estas secuencias están aún en

muchos casos “desnudas”, no tenemos información que nos verifique la relevancia, actividad o interés de muchos de esos fragmentos de ADN. Sin duda en los próximos años los esfuerzos tendrán que dirigirse a desarrollar sistemas de alto rendimiento en los que se pueda evaluar la importancia biológica de las enormes cantidades de secuencias que se están generando en la actualidad.

## EL PARQUE CIENTÍFICO DE MADRID

El Parque Científico de Madrid es una Fundación sin ánimo de lucro fundada por las Universidades Complutense y Autónoma de Madrid cuyo principal cometido es el apoyo a la actividad de I+D+I a los centros públicos y privados tanto de la Comunidad de Madrid como fuera de ella. El Parque Científica cuenta con departamentos de apoyo a la iniciativa empresarial junto con laboratorios de apoyo a la Investigación, como es la Unidad de Genómica. Esta Unidad está repartida en dos laboratorios, uno ubicado en el campus de Moncloa y otro en Cantoblanco. En total cuenta con un conjunto de más de 15 técnicos y está apoyado científicamente por profesores y catedráticos de ambas Universidades. Las Unidades de Genómica están preparadas para intentar responder con criterios de eficacia y calidad a cualquier problema que un científico pueda plantear dentro del ámbito de la Genómica. Si recorremos los principales servicios que las Unidades de Genómica del Parque Científico de Madrid ofrecen, nos encontraremos las siguientes actividades:

Secuenciación de ADN. Esta es la técnica básica de la Genómica. Se emplea de forma mayoritaria para conocer con el máximo detalle (secuencia primaria) un fragmento de ADN. Se puede utilizar para determinar la posible presencia de una mutación en un gen de susceptibilidad, y también para conocer si se ha preparado correctamente una molécula recombinante o qué mutación se ha incorporado en un organismo procariota tras un proceso de mutagénesis dirigida. En nuestro laboratorio utilizamos también la técnica de secuenciación para identificar taxonómicamente bacterias u hongos en base a la secuenciación de una región variable que se puede comparar con las bases de datos en busca de homologías, y para determinar la presencia de mutaciones recurrentes en pacientes con cáncer.

Análisis de fragmentos. Esta técnica se utiliza para determinar variantes de ADN basadas en el tamaño diferencial de fragmentos de ADN definidos, que recibe el nombre de Genotipado. Un caso típico es el análisis de microsatélites de ADN, que son regiones repetitivas de ADN que se presentan en un número distinto de veces en los cromosomas de cada persona. La secuenciación convencional de marcadores genéticos es normalmente capaz de discriminar entre géneros y especies diferentes, pero no resuelve diferencias por debajo de ese nivel. La identificación basada en marcadores de microsatélites puede ser capaz de discriminar cepas y subcepas, lo cual tiene un enorme impacto por ejemplo en la identificación de patógenos alimentarios. La técnica de análisis de fragmentos en humanos es la base de la huella genética, en la que se analiza una serie de aproximadamente 15 marcadores distintos, muy polimórficos (es decir, con muchas variantes posibles en la población). La combinación de esos 15 marcadores es propia de cada individuo y prácticamente imposible que se repita por azar en cualquier otra persona. Solamente habrá homologías con los parientes más relacionados, ya que heredamos una copia de cada marcador de cada uno de nuestros padres y presentaremos marcadores comunes con nuestros hermanos, que transmitiremos a nuestros hijos. La presencia y sobre todo la ausencia de homología en estos marcadores tienen una utilidad forense evidente sirviendo como un auténtico DNI molecular que permite nuestra identificación inequívoca allí donde se necesite.

PCR a tiempo real. Es el tipo de ensayo de cuantificación por excelencia. Se utiliza para medir los niveles de expresión génica tras cualquier tipo de tratamiento al que se ve sometido una célula u organismo. Aprovechando la enorme capacidad de amplificación de la PCR, se aplica también para detectar y cuantificar DNA específico presente en una preparación: por ejemplo la cantidad de DNA circulante de un patógeno en sangre, el nivel de contaminación con alimentos transgénicos de un producto, la presencia de un determinado microorganismo en una muestra medioambiental, siempre en términos cuantitativos (grado de inducción en la expresión, porcentaje de contaminación, número de microorganismos por mililitro). La PCR a tiempo real es un técnica tan poderosa que también se puede adaptar a ensayos no cuantitativos como el genotipado de polimorfismos de tipo SNP o la determinación de presencia/ausencia de patógenos contaminantes o transgenes en ensayos denominados a tiempo final.

*Microarrays*. Se dirigen especialmente a la determinación de perfiles de expresión génica a escala global. Se utilizan para cualquier sistema en el que se quiera descubrir qué cambios pueden ser los más relevantes tras un cierto tratamiento o en una situación biológica determinada cuando no se puede o no se desea restringir el estudio a genes de respuesta definidos. Cuando se aplican sobre ADN, los *microarrays* son capaces de establecer sitios de reorganización cromosómica, determinar regiones duplicadas o de pérdida de heterozigosidad, o de analizar millones de SNPs para genotipar polimorfismos de interés biológico descritos hasta el momento.

*Secuenciación masiva NGS*. Es una técnica recién implementada en nuestros laboratorios. Los primeros ensayos se están dirigiendo a proyectos tan diversos y de tanta relevancia como la determinación de la identidad de los virus presentes ciertos medios naturales, la localización de variantes genéticas de plantas en relación con su supervivencia, la identificación de ARNs con capacidad reguladora y al desarrollo de sistemas de alta productividad dirigidos a realizar de forma rápida y eficaz ensayos de diagnóstico genético para las principales enfermedades hereditarias, de interés en la práctica asistencial en los hospitales.

En su conjunto, se llega a dar respuesta a más de un centenar de grupos de investigación, departamentos hospitalarios y empresas de biotecnología situadas tanto dentro de la Comunidad de Madrid como en el resto de España, en Europa y América, siempre bajo criterios de calidad y con la máxima agilidad y rigor científico posibles. La razón de ser de nuestro laboratorio no es otra que hacer disponible a la comunidad científica unos equipamientos altamente especializados junto con una experiencia y un conocimiento profundo a nivel tecnológico y científico que permitan a los grupos de investigación obtener los resultados deseados para el avance de su investigación y del conocimiento científico en general.

## BIBLIOGRAFÍA RECOMENDADA

M. W. Pfaffl (2004): *Quantification strategies in real-time PCR*, in:"A-Z of quantitative PCR" (Ed. S.A. Bustin). Chaper 3 pages 87 – 112. International University Line (IUL)

A. Jeffreys, V. Wilson y S. Thein (1985). *Individual-specific 'fingerprints' of human DNA*. Nature, 316, 6023, pp. 76-79.

A. Glas et al. (2006). *Converting a breast cancer microarray signature into a high-throughput diagnostic test.* BMC Genomics Vol 7, 278-287.

*Michael L. Metzker* (2010). *Sequencing technologies —the next generation*. Nature, Vol. 11, 31-46.

Página del *National Center for Biotechnology Information* (NCBI): www.ncbi.nih.gov.

Páginas del consorcio Laboratorio Europeo de Biología Molecular (EMBL)-Instituto Europeo de Bioinformática (EBI)-Insituto Sanger www.ensembl.org; www.ebi.ac.uk.

Página de acceso a la información del Proyecto Genoma Humano: www.genome.gov.

Figura 1

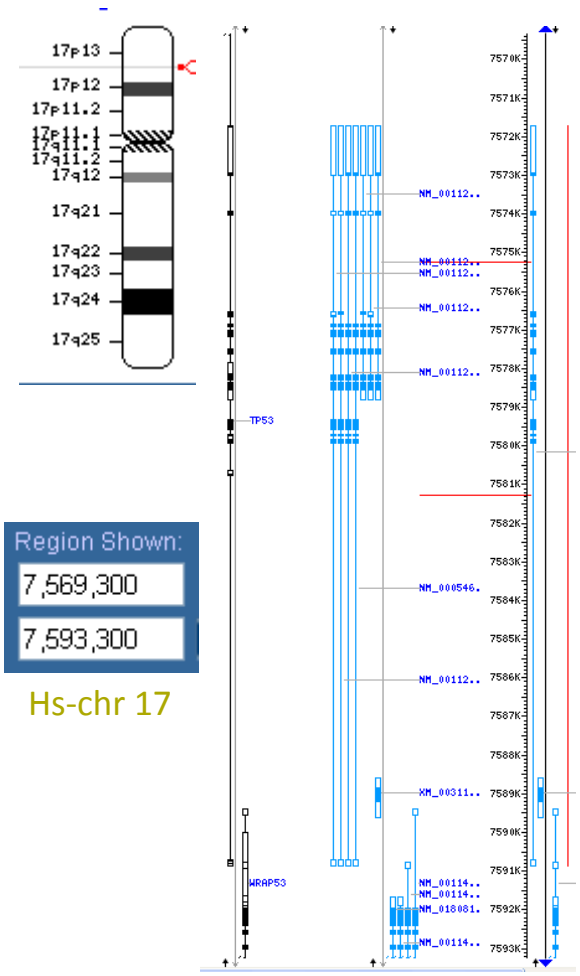


Figura 1.
Esquema de la organización de una región del cromosoma 17 del genoma humano.

Se representa (arriba a la izquierda) la estructura del cromosoma, y se amplía (derecha) una pequeña región del brazo corto, donde reside el gen TP53 implicado en progresión tumoral y un regulador de TP53 conocido como WRAP53, ambos representados por distintas variantes de ARN. (Tomada de www.ncbi.nlm.nih.gov)

Figura 2

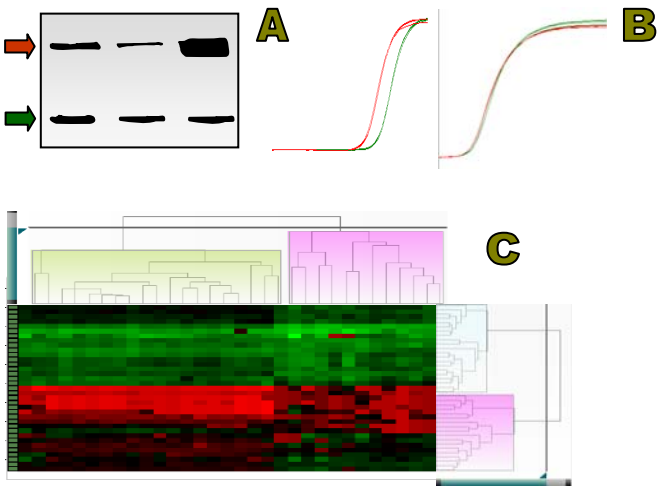


Figura 2.
Algunos ejemplos de sistemas de análisis de expresión génica.

A: localización de bandas de ARN correspondientes a dos genes diferentes (señalados con flechas de distinto color) mediante la técnica de *Northern blot*. El gen localizado en la región superior muestra una reducción y una inducción progresiva en las situaciones representadas.
B: Medida de la expresión génica mediante PCR a tiempo real. Los diagramas representan curvas de amplificación analizadas en dos genes diferentes, uno regulado (izquierda) y otro estable (derecha).
C: Generación de dendogramas que clasifican muestras (panel superior) y genes (panel lateral) en base a perfiles de expresión multigénica obtenidos mediante PCR a tiempo real o hibridación en *microarrays*.