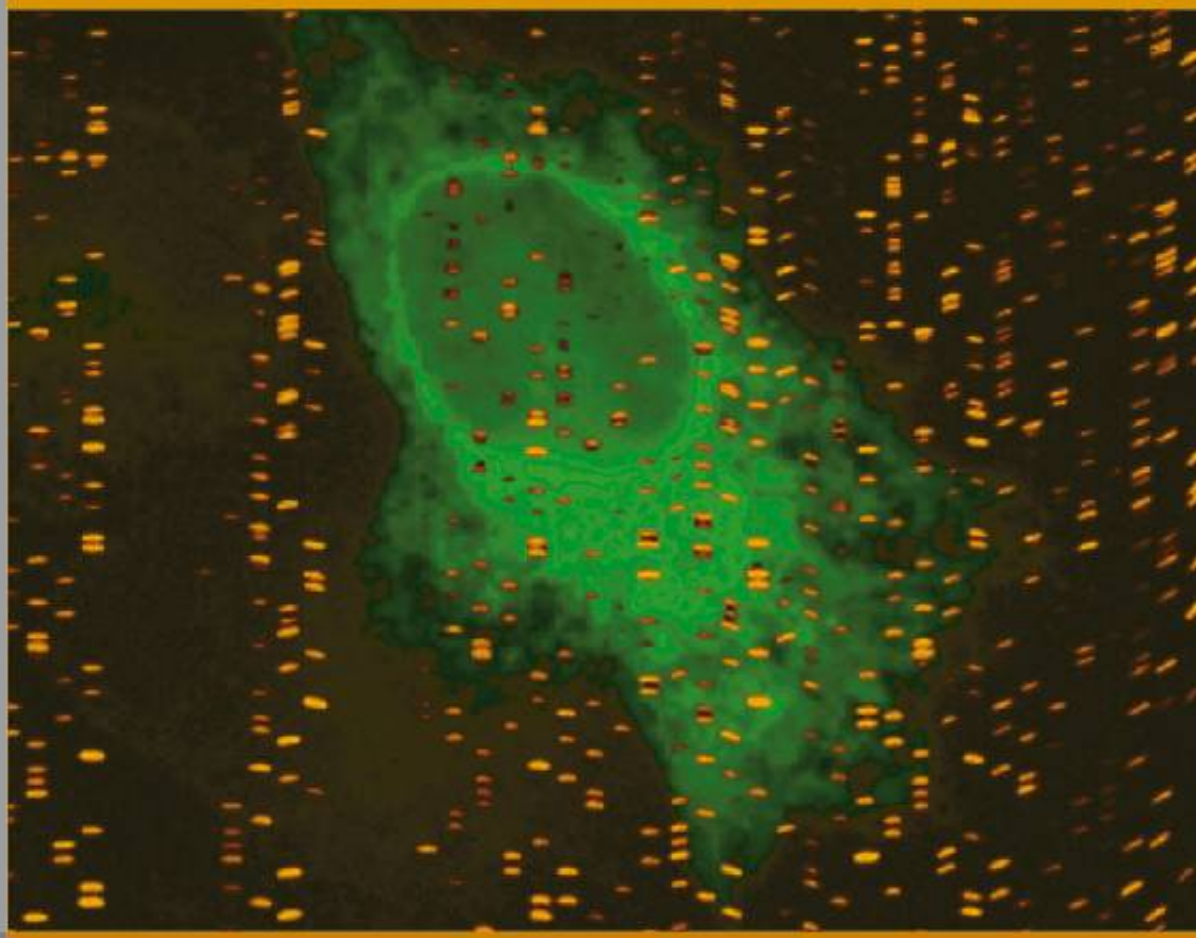
LIBROS DE TEXTO PARA ESTUDIANTES UNIVERSITARIOS

Secuenciado del genoma



2013

UNLPam

Jorge OYHENART

[2013] LIBROS DE TEXTO PARA ESTUDIANTES UNIVERSITARIOS

Secuenciado del genoma

Jorge OYHENART

LIBRO DE TEXTO PARA **ESTUDIANTES UNIVERSITARIOS**

Secuenciado del Genoma
Jorge OYHENART

Diciembre de 2013, Santa Rosa, La Pampa

Diseño y Diagramación: División Diseño-UNLPam

Cumplido con lo que marca la ley 11.723
EdUNLPam - Año 2013
Cnel. Gil 353 PB - CP L6300DUG
SANTA ROSA - La Pampa - Argentina

Índice

Capítulo 1

Las bases de la vida

1

El hombre siempre se ha preguntado cuál es su origen, qué lo hace parecido o diferente de otros organismos, de qué manera podría manipular otros organismos y asegurar la producción de alimentos, dónde radica la diferencia entre la salud y la enfermedad y hasta, si podrá algún día prolongar su vida. Estas, tan bien como otras cuestiones, deben hallar algunas respuestas en la información almacenada en nuestras células.

Hace un siglo se reconocía que la transmisión de la información genética se albergaba en el núcleo de la célula. Debió transcurrir medio siglo para reconocer cómo se almacenaba y transmitía esta información. Su lectura e interpretación comienza hace escasos años. Ningún adelanto científico o tecnológico es un descubrimiento sin precedentes, esto es evidente. Cada hallazgo llega a su tiempo, y cuando ya se han respondido otras preguntas y obtenido las herramientas necesarias.

La lectura de la información contenida en un fragmento de ADN no es, estrictamente hablando un adelanto científico, sino más bien técnico. La lectura y transcripción de la sucesión de cuatro componentes químicos, en un alfabeto compuesto de As, Gs, Ts y Cs.

La lectura de la secuencia del ADN ha sido causa de un incremento notable en nuestros conocimientos y en nuestro bienestar. Impulsando o siendo causa directa del desarrollo, el secuenciado contribuyó muy pronto a caracterizar virus y bacterias, reconocer enfermedades y diferencias entre organismos y hasta, a crear nuevos organismos y producir medicamentos.

A diez años de poder "leer el ADN" el hombre se lanza en una empresa muy atrevida, la lectura del mayor jeroglífico hallado hasta entonces, su propio genoma. La lectura de los 3.000.000.000 de bases que componen el ADN de nuestra especie demandaría quince años de trabajo coordinado entre varios centros de investigación.

El comienzo de este siglo ya ha quedado grabado en nuestra historia como el momento en el cual el hombre logró leer su genoma. Parte de este texto trata de explicar cómo llegó a leerse la secuencia de un genoma, el nuestro, porqué y

para que sirve o servirá.

La lectura de la secuencia del genoma humano persiguió propósitos muy variados y supuso una inversión económica muy grande. El interés por generar nuevos conocimientos -comprender la evolución del hombre y otras especies, avanzar en el estudio de enfermedades-, es balanceado por la necesidad, y el deseo, de generar nuevas fuentes de ingresos –transferir conocimientos, generar tecnología-.

El genoma es una gran fuente de recursos, científicos tan bien como económicos. Transcribir el libro requirió gran cantidad de innovaciones y posibilitó un número asombroso de adelantos tecnológicos. Estos adelantos marcan el inicio de una verdadera revolución, la genómica.

La transcripción de secuencias de genomas enteros en archivos informáticos supone un cambio muy importante en el trabajo por venir. La revolución genómica es acompañada por una informática. Hay que interpretar la información recogida y no hay otro modo de lograrlo que mediante el análisis, y la confirmación.

La lectura del genoma humano, más que sus resultados, significa un punto de inflexión entre lo que era y lo que será. Estudiábamos los genes de a uno, y con gran dificultad para empezar a estudiarlos juntos, de a cientos o miles. Existen herramientas creadas con el sólo propósito de leer grandes cantidades de secuencias. Su desarrollo está en plena expansión y es hoy una fuente inagotable de trabajo y recursos.

El proyecto genoma humano pone de manifiesto que grandes empresas son posibles y que el esfuerzo puede traducirse en el bien común. El avance realizado es realmente vertiginoso, y el análisis genómico ya llega a nuestras vidas. El reconocimiento de individuos y vínculos, el diagnóstico de enfermedades y riesgos, y el uso "racional" de medicamentos están entre sus logros y objetivos inmediatos.

La tecnología debe servir a nuestro bienestar pero también se impone. Detrás de buenas intenciones puede habero otras no tan buenas. La información obtenida se protege y explota comercialmente, se usa para manipular genomas, crear organismos artificiales, o facilitar el clonado de organismos superiores sin propósitos claros, y la medicina se hace personalizada para obtener más beneficios.

Aún no podemos interpretar gran parte de la información almacenada en un genoma, siquiera el de un organismo muy simple. Se requiere aún mucho trabajo, experimental y de análisis. Debemos formar parte de este cambio, al menos manteniéndonos informados, y ser críticos de sus resultados.

HERENCIA PARTICULADA

El siglo XIX transcurrió dejando dos grandes ideas que revolucionarían nuestra manera de interpretar cualquier observación en biología. Charles Darwin hacía pública su visión de la variación de las especies y Gregor Mendel describía su interpretación de la transmisión de la información genética. Debió transcurrir mucho tiempo hasta que ambas ideas pudieran unirse y tomar su forma actual.

El autor de "el origen de las especies" observó que pueden existir ligeras diferencias entre organismos de distintas especies (Figura 1), que hallarían un continuo entre miembros de una misma especie. Darwin sostuvo que toda población consiste de individuos ligeramente distintos unos de otros y que dichas variaciones hacen a la existencia distintas capacidades para adaptarse al medio natural. Estas capacidades deben permitir al organismo reproducirse con más o menos éxito según la naturaleza selecciona los individuos mejor adaptados para sobrevivir y reproducirse. Este proceso se conoce como "selección natural".

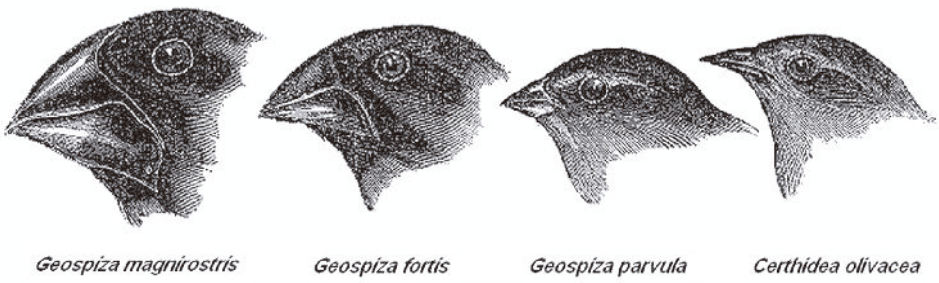


Figura 1. Diferentes especies de pinzones observados por Darwin en las Islas Galápagos. Cada grupo tiene un pico tamaño y forma particular que le permitiría tomar ventaja de las distintas fuentes de alimento.

Transcurrido un gran número de generaciones, los rasgos de los individuos que mejor se adaptaron a las condiciones naturales se vuelven más comunes. Darwin llamó a este proceso "descendencia con modificación". Esta forma de evolución de un grupo biológico se denomina gradualista o gradualismo.

La selección de caracteres más favorables permitiría un cambio continuo en el seno de una especie. Estos cambios constantes se harían evidentes tras períodos muy prolongados, de miles de millones de años y serían el motor para la aparición de nuevas especies.

Durante más de un siglo, la evolución de las especies fue tema de intenso debate y la teoría de Darwin parecía dar un punto final a la discusión. Gran cantidad de pensadores en la época adherirían a sus ideas que pueden reducirse a algunos principios:

- Las especies no tienen una existencia fija ni estática sino que se encuentran en cambio constante.

- La vida se manifiesta como una lucha constante por la existencia y la supervivencia.
- La lucha por la superviviencia provoca que los organismos que menos se adaptan a un medio natural específico desaparezcan y permite que los mejores adaptados se reproduzcan, a este proceso se le llama "selección natural".
- La selección natural, el desarrollo y la evolución requieren de un enorme período de tiempo, tan largo que en una vida humana no se pueden apreciar estos fenómenos.

Gran parte de estos postulados no podrían, y aún no pueden, probarse, explicando porqué la evolución por acción de la selección natural es, aún, una teoría. Sin embargo, la información acumulada durante más de un siglo de investigaciones sostiene esta idea como un hecho.

La teoría de Darwin de la evolución de las especies se sostuvo en un número muy grande de observaciones y tanto su aceptación como el rechazo que pudo provocar en otros, lo llevaron a intentar perfeccionarla con nuevas observaciones y algo de experimentación. Darwin no ahondó en explicaciones sobre el origen de las variaciones genéticas y la transmisión de las mismas. Su postura fue favorable a algunas de las ideas más aceptadas en la época, como la herencia de caracteres adquiridos postulada por J. B. Lamarck, o el modelo de dilución de caracteres explicado más adelante por la influencia de muchos genes en un carácter. Sostener, o tratar de congeniar estos argumentos con su teoría supusieron un gran esfuerzo para Darwin.

La herencia de los caracteres no hallaba en la época muchas explicaciones o más bien, cada ejemplo suponía una explicación propia. La altura de un individuo o una planta, o el color de la piel en el hombre daban en considerar una dilución de factores presentes en cada parental. Producto del cruce de un hombre de color y una mujer blanca resulta, sin excepción, un hijo de piel café con leche, intermedio entre ambos padres. Con esta fusión o dilución de los rasgos de ambos progenitores en su descendiente, cualquier variante ventajosa para un individuo estaba condenada a desaparecer en unas pocas generaciones.

La idea de diluir y eventualmente hacer desaparecer una variante ventajosa parecía contradecir, por otra parte, las formas extremas que se habían obtenido en animales o plantas a través de la selección artificial. La variación podría entonces aparecer por el uso o desaparecer con el desuso, y ser transmisible a la siguiente generación.

Estos razonamientos se adecuaban, como ya se mencionara, a numerosas y variadas, observaciones. Pocas se sostenían en la experimentación. La herencia de los caracteres debía explicarse de alguna manera. Existe tal dilución? Surge nueva variación?

El trabajo de Gregor Mendel probaría de manera clara, que los caracteres heredables no se diluyen. Trabajos posteriores demostrarían a su vez que, con cierta frecuencia, surgen nuevas variantes.

A diferencia de los naturalistas de la época, Gregor Mendel basó su trabajo, en una experimentación sumamente rigurosa y en un análisis detallado de los resultados. Sus experiencias se basaron en una sola especie de plantas, de arveja (*Pisum sativum*), que fueron seleccionadas durante dos años hasta confirmar que sólo transmitían a la progenie las características que se observaran en ellas. Para el análisis se escogieron caracteres discretos que pudieran dar resultados claros y sin transición. Estos detalles no escaparon a quienes examinaron su trabajo y no adhirieron en la época, o lo hicieron de manera ferviente más adelante.

Mendel detalló en la publicación de su trabajo la elección de rasgos que se presentaban como una de dos formas alternativas. La altura de una planta (alta o baja), el color o la textura de una semilla (gris o blanca y lisa o rugosa), el color de los cotiledones (amarillo o verde), la forma o color de un fruto (liso o con constricciones y verde o amarillo) (Figura 2), el color o la posición de una flor en el tallo (blanca o violeta y terminal o axial).

Al cruzar dos líneas puras (homocigotas) con aspecto (fenotipo) diferente, toda la progenie (la generación F1) resultante mostraría la forma característica de uno de los parentales. El rasgo típico del otro parental, permanecería oculto en esta generación F1. Sin embargo, tras la autopolinización o entrecruzamiento entre plantas de dicha generación F1, el carácter oculto reaparecería en la progenie (la segunda generación, o F2). Más notable aún resultaba que la frecuencia con que este rasgo reaparecía resultaba en una proporción constante de 1 a 3 respecto del rasgo manifiesto en la generación anterior (F1).

Mendel concluyó, de manera muy acertada, que un organismo (diploide) contiene elementos responsables de sus caracteres visibles que pueden o no manifestarse en el mismo. Estos elementos se separan durante la formación de las gametas y vuelven a hallar un par en la formación del nuevo zigoto. La expresión de cada elemento en el nuevo organismo dependerá de las propiedades de dicho elemento respecto de su par. Según se manifiesta o no en el híbrido, Mendel llamó a estos elementos, dominante o recesivo respectivamente.

El fenómeno de dominancia de un había sido descrito en el análisis de cruces con plantas de melón por A. Sageret, quien acuñara el término, cincuenta años antes. Sin embargo, el análisis sistemático y la interpretación de Mendel dieron forma clara al fenómeno.

Mendel analizó la herencia simultánea de más de un carácter y dedujo que elementos determinantes de dos caracteres diferentes siguen cursos distintos en la descendencia. En cruces dihíbridos (en los que se consideran dos caracteres diferentes) entre homocigotos para dos caracteres los individuos de la primera generación filial mostrarán las variantes dominantes de cada carácter. A su vez el

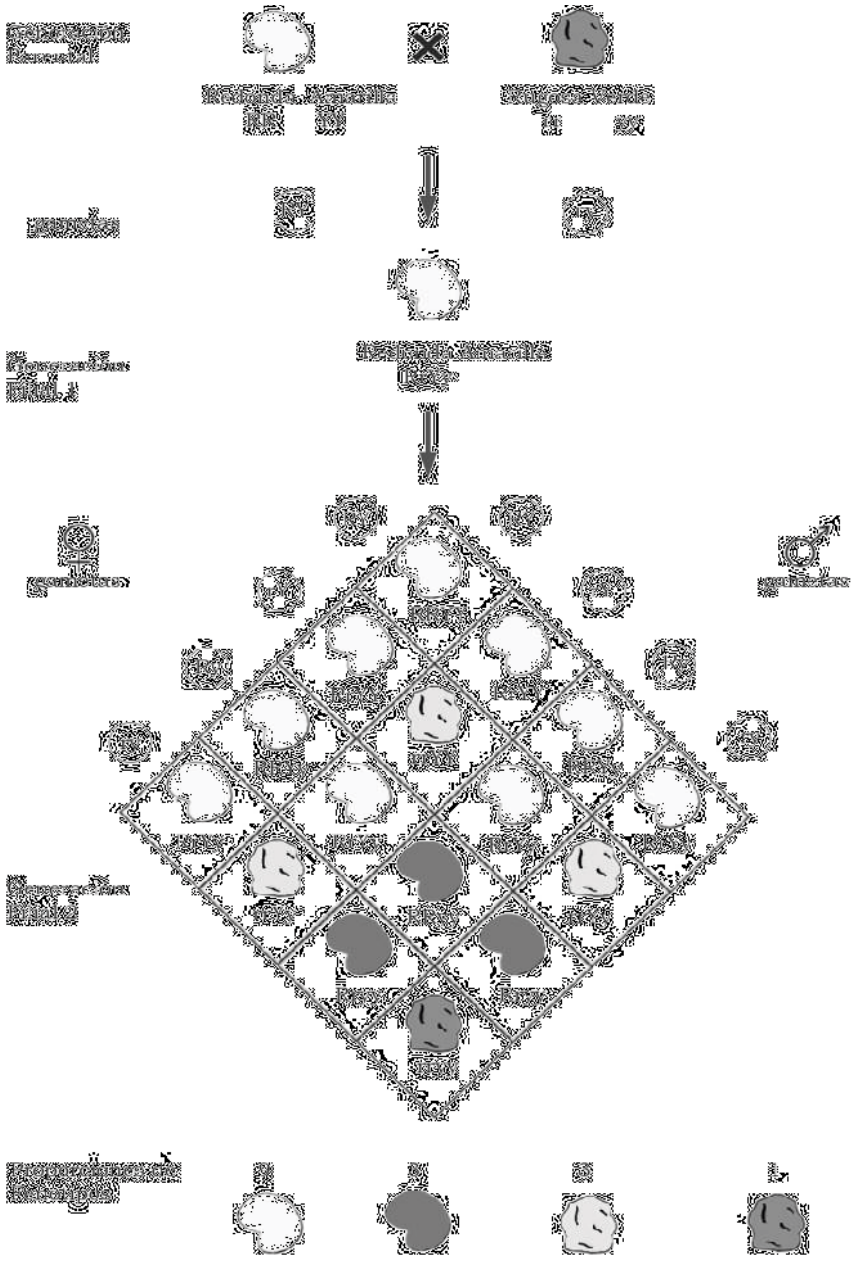


Figura 2. Cruzamiento dihíbrido. Las leyes de Mendel permiten predecir el resultado de un cruzamiento genético. La primera ley establece que los alelos segregan al azar. En la herencia de los alelos "R" y "r" o "Y" e "y" (representando cada letra mayúscula al alelo dominante), en un cruzamiento entre padres heterocigotos, cada miembro de la generación Filial tiene la misma probabilidad de heredar una u otra forma de cualquiera de los padres. La ley de la segregación independiente sostiene que la herencia de un par de alelos no interfiere con la de otro par. Siendo los padres heterocigotos para ambos genes "R" e "Y", cada miembro de la nueva generación tiene la misma probabilidad de heredar cualquiera de las formas alélicas. El tablero de Punnett muestra los genotipos y fenotipos esperados en la generación F2. En el cruce dihíbrido, el gen R la forma (lisa o rugosa) del fruto y el gen Y el color (amarillo o verde) del mismo. Todas las combinaciones posibles de gametas generadas por un organismo heterogigoto en cada lado del cuadrado, permiten estimar fácilmente los genotipos de los descendientes.

cruce de estos individuos o la autopolinización, mostrarán en la siguiente generación proporciones de fenotipos compatibles con la segregación independiente de los dos caracteres. Esto es se hallarán proporciones acordes al producto de dos como la descrita anteriormente (3:1) o 9:3:3:1, según se trate de las dos formas dominantes (9), de una de ellas (3) o de ambas formas recesivas (1).

Aunque el trabajo de Mendel fue publicado en una revista importante y debió ser leído por grandes científicos de la época, este no tuvo mayor repercusión, o aceptación. No es improbable que su propuesta de dos formas alternativas, una aparente y la otra oculta, para un elemento responsable de la herencia de un carácter hubiera sido interpretado como fijista, y por ello opuesto a la flexibilidad que debía permitir la evolución. Entrado el siglo XX los estudios de Mendel comenzarían sin embargo a ser reconocidos y sus premisas serían consideradas como leyes fundamentales de la herencia (Figura 2) .

La experimentación permitiría progresivamente reconocer que el material hereditario es inalterable, o cuando menos, que existen mecanismos para evitar su cambio. La variación preexistente en el seno de una especie permitiría congraciar la idea de evolución por medio de la selección natural con las leyes de la herencia.

Hacia fines del siglo XIX resultaba claro que la variación y la herencia de un carácter podrían adoptar más de una forma. Francis Galton se interesó en la variación y la herencia de numerosos caracteres. Para su estudio desarrolló y empleó herramientas estadísticas que le permitieron –y aún nos permiten- describir la variación continua de organismos de una población, reflejada en la distribución normal.[1] El estudio de estos caracteres acordaba con una propuesta de evolución gradual, o darwiniana aunque calificar y cuantificar precisamente los rasgos dejaba en ocasiones ver patrones de herencia discontinua.

William Bateson describía en 1892 la distribución normal y discontinua de un carácter, y llamaba a este fenómeno dimorfismo. Sin conocer el trabajo de Mendel, Bateson se preguntaba si una variante alternativa, que no desaparece en la progenie, no sería más apta para explicar un cambio o una ventaja selectiva durante la evolución. Bateson había observado que en insectos pueden existir cambios espontáneos que posibilitan el desarrollo de un apéndice incorrecto en lugar de otro (por ejemplo, una pata en lugar de una antena)[2] . Con esta información se preguntaba si una forma de variación discontinua no serviría mejor como sustrato a la selección natural. Basado en estos argumentos sostuvo que la evolución de las especies podría no ser siempre gradual, sino que podría acontecer por saltos.

1 F. Galton logró, entre otras cosas, demostrar la inviabilidad de la teoría de la pangénesis, desarrollar herramientas estadísticas de gran utilidad (como el análisis de regresión y la correlación) en el análisis biométrico, la bioestadística y la genética de poblaciones. También puede reconocerse de su labor el haber reconocido la influencia del ambiente en el desarrollo de un individuo.

2 Bateson describió una mutación "homeótica" conocida como Antennapedia. La palabra homeótico, tan bien como genética y epistasis deben su origen a este autor.

Bateson debió hacer frente a arduas críticas por su ataque a la teoría gradualista de la evolución. Para demostrar su punto de vista emprendió trabajos experimentales en un área por él desconocida, el cruzamiento de plantas. De cruces de plantas pilosas y no pilosas de *Biscutella laevigata* observó que hay características que no se diluyen en la descendencia, sino que persisten ambas variantes. Otros autores llegaban al mismo tiempo a conclusiones semejantes, dos de ellos ya conocían los resultados de Mendel.

> *"It is impossible to be presented with the fact that in Mendelian cases the crossbred produces on an average equal numbers of gametes of each kind, that is to say, a symmetrical result, without suspecting that this fact must correspond with some symmetrical figure of distribution of the gametes in the cell divisions by which they are produced."*
>
> *W. Bateson*

Llegar repetidamente a las conclusiones que alcanzara Mendel dio un gran impulso al estudio de la variación de los seres vivos. Ahora la búsqueda se centraría en los elementos responsables de la transmisión de caracteres, particularmente en la célula germinal.

HERENCIA CROMOSÓMICA

En 1882, con ayuda de un microscopio y de procedimientos de tinción muy elaborados, Walther Flemming había descrito con precisión el comportamiento de los cromosomas durante la división celular. Flemming describió cómo, dentro del núcleo celular en división, toman forma gradualmente estructuras cromófilas discretas. Los cromosomas, llamados por este autor cromatina, se agupan y alinean en el ecuador celular al comienzo de la mitosis. Tras lograr su mayor grado de condensación se separan en mitades idénticas que se dirigen a polos opuestos de la célula y se descondensan dentro dos nuevos núcleos hijos.

Theodor Boveri llevó estas observaciones algo más lejos, trabajando con embriones de Ascaris. Las células de este parásito contienen sólo dos pares de cromosomas. Tras la primera división en el embrión las células somáticas pierden gradualmente sus cromosomas, mediante un proceso aún poco elucidado, y las células germinales conservan los cromosomas que se verán implicados en divisiones reduccionales a medida que se forman las gametas. Boveri pudo trazar la reducción del número de cromosomas en la formación de las células germinales y la restitución del número diploide tras la fertilización. La documentación de la haploidización ocurrida durante la formación de las gametas y la restitución del número cromosómico durante la fertilización dejó ver claramente cómo podría transmitirse la información genética.

La primera explicación de la relación entre los cromosomas y la herencia de los caracteres se debió sin embargo al trabajo de Walter Sutton. Este autor estudió en detalle la meiosis de un ortóptero (*Brachystola magna*) indicando que en sus células se hallan 11 pares de cromosomas. Cada par de cromosomas tiene un tamaño característico por lo que puede seguirse sin ambigüedad su evolución. Existe asimismo un cromosoma individual (responsable de la determinación del sexo) que segrega como cualquier otro aunque no tiene par.

Sutton describió la meiosis y sugirió que cada cromosoma es una entidad estable que se mantiene a través de las generaciones. Del comportamiento de los cromosomas durante la división dedujo que no existe un lado paterno y uno materno durante la división y entonces que su orden en la placa metafásica es aleatoria. Cada célula germinal puede entonces recibir cromosomas maternos y paternos en un orden muy diferente al de otras células germinales. Representando con "n" el número de cromosomas de una especie, el número de probables combinaciones de cromosomas en sus gametas sería de 2n (Sutton 1903).

> *"It has long been admitted that we must look to the organization of the germ cells for the ultimate determination of hereditary phenomena. Mendel fully appreciated this fact and even instituted special experiments to determine the nature of that organization. From them he drew the brilliant conclusion that, while, in the organism, maternal and paternal potentialities are present in the field of each character, the germ cells in respect to each character are pure."*
>
> *W. Sutton. 1903*

Aunque el número de cromosomas de una especie resultara elevado resultaba improbable que cada uno determinara un solo carácter (contuviera un único gen). Experiencias de Bateson y su grupo dejaban ver que de cruzamientos dihíbridos no siempre se hallan cifras acorde con las proporciones esperadas para la segregación independiente. Estos datos se explicaban mejor suponiendo que dos genes se localizan en un mismo cromosoma.

Otros datos sostendrían la existencia de más de un gen en un cromosoma. Especies semejantes pueden tener números de cromosomas muy diferentes y organismos complejos se revelan a veces poseedores de un número escaso de cromosomas (Crow & Crow 2002). Sutton hipotetizó que cada una de estas estructuras lineales debía albergar muchos genes. Pero de ser así, cómo podrían transmitirse independientemente unos de otros? La respuesta a esta contradicción entre la teoría y la observación fue el paso crítico en el desarrollo de técnicas de mapeo genético.

En 1909, Frans A. Janssens describía sus observaciones acerca de un fenómeno curioso que tiene lugar durante la meiosis.[3] De las cuatro cromátidas

3 La théorie de la Chiasmatypie. Nouvelle interprétation des cinèses de maturation.

que se alinean en diploteno de la primera división, dos de ellas se cruzan entre sí mientras que la dos restantes no lo hacen. Janssens dedujo que las cromátidas maternas y paternas que se "entrecruzan", deben romperse y reacomodar segmentos maternos y paternos. Janssens denomina quiasmas a los cruces de cromátidas observados en el microsopio, y propone que de la rotura y unión de segmentos deben surgir nuevas cromátides con material derivado de ambos padres. Las gametas deben entonces tener cromosomas formados por múltiples combinaciones de segmentos maternos y paternos. Esta idea y las experiencias subsecuentes del grupo de Thomas H. Morgan harían posible comprender la segregación de los caracteres.

Morgan adhería a la idea de la evolución de las especies aunque no coincidía con la propuesta del cambio gradual. Este autor sostenía que la teoría de cambios mayores de Hugo de Vries se ajustaba mejor al cambio. De Vries había hallado variaciones en la descendencia de la planta *Oenothera lamarckiana* que atribuía a la ocurrencia de cambios en la información genética. En su "teoría de la mutación" de 1886 postulaba que la evolución, y sobre todo el origen de las especies, podría ocurrir más frecuentemente debido a cambios a gran escala más bien que a través de pequeños cambios graduales propuestos por el darwinismo. Esta forma de evolución se denomina, por contraposición, saltacionismo.

Morgan había establecido en 1908 un laboratorio que haría historia, y para muchos, significa el verdadero nacimiento de la genética. Su propósito sería el de hallar modificaciones susceptibles de justificar la aparición de una nueva especie y su modelo de estudio fue la mosca de la fruta, *Drosophila melanogaster*.

Tras dos años de intentos por inducir cambios en el aspecto de las moscas mediante radiaciones y agentes químicos, Morgan hallaría, entre las moscas de ojos rojos, un macho mutante de ojos blancos. Del cruzamiento de este macho de ojos blancos con una hembra de ojos rojos, surgiría una progenie constituida en su totalidad por moscas de ojos rojos. Sin embargo, en la segunda generación derivada del cruzamiento de estas moscas aparecerían nuevamente machos de ojos blancos. El gen responsable del pigmento de los ojos segregaba, como en los experimentos de Mendel, con una proporción de 3 a 1. La alteración hallada por Morgan era además visible sólo en machos, y por ello estaba ligada al sexo. De esto podía inferirse que el rasgo se hallaba probablemente en uno de los cromosomas sexuales.

Morgan y sus estudiantes fueron muy exitosos en la su búsqueda de moscas mutantes y a medida que se acumulaban otras mutaciones se hacían evidentes patrones de herencia más y más complejos. Tal como la mutación responsable del color blanco en los ojos, una mutación responsable de un ala minúscula (*miniature*) y una responsable del cuerpo completamente amarillo (*yellow*), aparecían también ligadas al sexo. Notablemente, en el análisis de cruzamientos dihíbridos la proporción de combinaciones de genes semejantes a los padres

eran sensiblemente mayores que las combinaciones diferentes, no parentales o recombinantes. Morgan atribuía este efecto a la localización de los tres genes en el mismo cromosoma.

Si un cromosoma aloja distintos genes y éstos segregan, entonces deben poder separarse. Del análisis de la segregación de mutantes *white*, *yellow* y *miniature* resultaba claro que genes alojados en un mismo cromosoma pueden segregar, y además lo hacen en proporciones diferentes. La teoría del quiasmatipo formulada por Janssens proporcionaba a Morgan una explicación para la separación física de estos genes. El quiasma debe ser un fenómeno citológico que refleja el intercambio de material entre cromosomas y este intercambio no tiene otro propósito que el de permitir la separación de los genes y la segregación de caracteres en la descendencia. Luego, suponiendo que el intercambio de fragmentos entre cromátides (el entrecruzamiento) es un evento aleatorio, es decir que existe la misma probabilidad de que se produzca en cualquier posición a lo largo de un par de cromátides alineadas, entonces dos genes distantes se separarán con mayor frecuencia que dos genes que están muy juntos.

El equipo de Morgan acuña los términos ligamiento y recombinación para referirse a la asociación física de los genes y a la posibilidad de separarse. Los genes *white* y *yellow* se veían muy ligados pues mostraban sólo 1.3 % de recombinantes en la descendencia, mientras los genes *white* y *miniature* debían estar más separados pues mostraban 37.2 % de recombinación. Las experiencias en la mosca de la fruta permitían conocer que genes cercanos en un cromosoma se comportan como si estuviesen ligados en tanto que genes en extremos opuestos de un cromosoma pueden segregar más fácilmente, casi como si estuvieran en cromosomas distintos.

Alfred Sturtevant se basó en este principio para inferir la distancia genética y el orden en que los genes se disponen en un mismo cromosoma. A partir de datos de cruces dihíbridos y trihíbridos (trifactoriales) Sturtevant desarrolló un sistema para confeccionar mapas genéticos o mapas de ligamiento (Figura 3). Partiendo del supuesto que el mayor porcentaje de recombinantes que puede obtenerse en un cruzamiento es de 50 %, un valor inferior puede emplearse como estimación de la distancia entre dos genes. Cuando 1 producto entre 100 es recombinante, la frecuencia de recombinación de 1 % y su distancia puede asumirse igual a 1 unidad de mapa (mu, o map unit). Esta distancia se conoce hoy como centimorgan (1 centimorgan o cM).

Así como la probabilidad de un entrecruzamiento entre dos loci aumenta conforme éstos se hallan más separados, también aumentan progresivamente las chances de que ocurra más de un entrecruzamiento. La ocurrencia de ese segundo evento anulará el efecto del primer entrecruzamiento y no permitirá ver recombinantes. Por ésta y otras razones se toman ciertas precauciones en la elaboración de mapas. Por lo general puede deducirse con cierta seguridad la distancia que separa a

dos genes muy cercanos. Sin embargo, la frecuencia de recombinantes puede llevar a subestimar la distancia genética entre genes más separados.

Localizar los elementos responsables de la transmisión de la información genética en los cromosomas y revelar su naturaleza particulada significó un gran avance. La genética adquiriría entonces un gran significado. Sus métodos, y particularmente la posibilidad de prever los resultados experimentales, posibilitarían la creación de nuevas variedades de vegetales y animales útiles al hombre, la obtención de nuevos modelos de estudio, la comprensión de afecciones en el hombre y la posibilidad de comprobar relaciones entre los seres vivos. En los años siguientes el hombre trataría de ir más cerca de la verdadera naturaleza de la información genética.

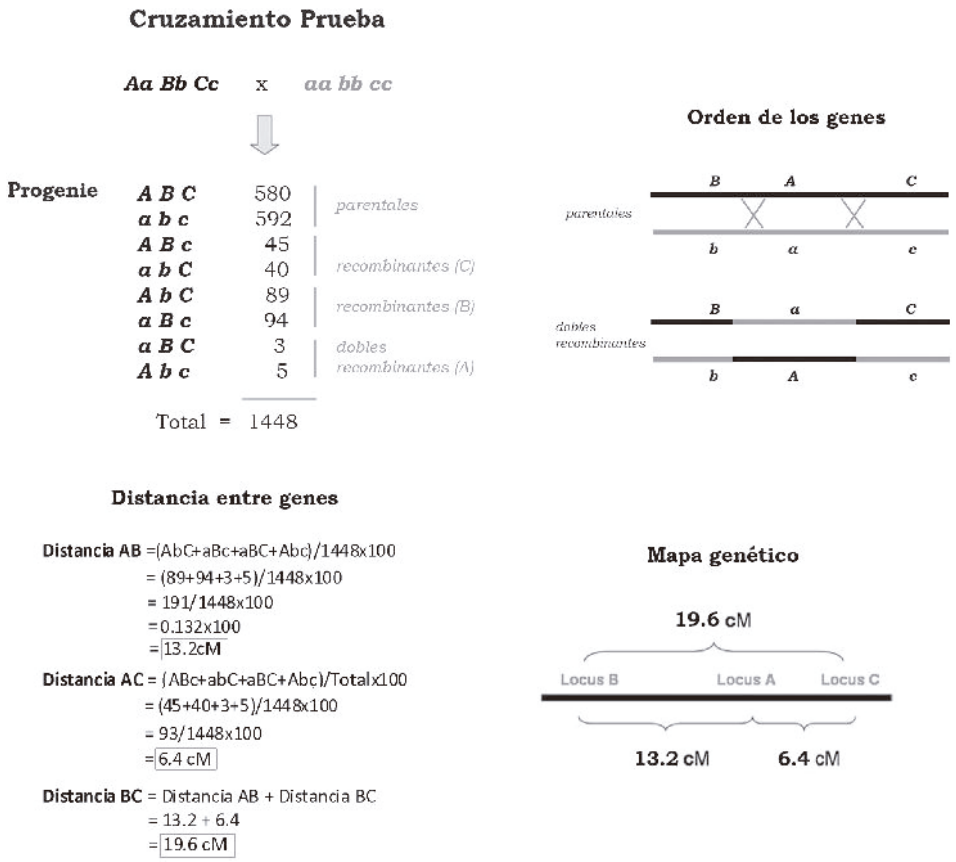


Figura 3. Prueba de tres puntos ideada por A. Sturtevant para determinar el orden y las distancias entre genes. Para el cruce trifactorial se crean organismos heterocigotas para las tres características examinadas y se cruzan con homocigotas recesivos (cruzamiento de prueba). En la progenie se observarán ocho fenotipos diferentes si los tres genes segregan. Si existe ligamiento entre genes, los dos fenotipos más representados corresponden por lo general con los de ambos progentores. Las proporciones entre los fenotipos restantes y el total de individuos obtenidos ayudan entonces a determinar el orden y la distancia entre los tres genes analizados. Los dos fenotipos menos representados corresponden a dobles entrecruzamientos, uno a cada lado del gen que se aloja en medio de los otros (derecha arriba). La distancia entre cada gen extremo y el del centro se estima mediante la suma de los recombinantes de cada clase, más los dobles recombinantes (abajo izquierda). La distancia entre ambos genes situados en los extremos se estima mediante la suma de las dos anteriores. La localización de cada gen y las distancias entre ellos ayudan a trazar una mapa genético (abajo derecha).

LA NUCLEÍNA

El manejo de la información y la divergencia en las técnicas de estudio llevan con frecuencia a la especialización y a la aparición de nuevas áreas de estudio. Hoy es frecuente ver una fisión constante de grandes áreas en dominios de interés cada vez más restringidos. Lo contrario, la fusión o asociación de dos conceptos ocurre más raramente.

El estudio de los genes y de su transmisión a la descendencia fue, desde el comienzo indistinguible de lo que hoy conocemos, o asociamos, con la genética. El estudio del ADN fue sin embargo, parte de un proceso completamente diferente, el estudio de la química de la vida. Debieron transcurrir muchos años antes de poder conectar ambas ideas en una.

Cinco años después de la publicación del trabajo de Mendel un investigador suizo, Johann F. Miescher, daba a conocer su trabajo sobre el aislamiento y la primera caracterización del mayor componente del núcleo celular, el ADN.

El estudio de la célula y sus componentes apenas comenzaba por esa época. Miescher intentaba aislar proteínas a partir de leucocitos cuando dio con un método para separar el núcleo celular. Las células se lavaban con soluciones frescas de ácido clorhídrico diluido durante períodos de varias semanas a baja temperatura (en invierno), y separaban los lípidos mediante agitación en una mezcla de agua tibia y éter. Estabilizada la mezcla, los núcleos celulares se depositaban como un fino granulado en el fondo del recipiente. Al añadir una solución alcalina los núcleos se hinchaban y perdían la capacidad de retener el iodo. La adición de ácido revertía la hinchazón y llevaba a la aparición de un precipitado blanco. Miescher denominaba a esta sustancia precipitable en solución ácida, "nucleína".

> *"DNA is a history book -a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease."*
>
> *Francis Collins*

Para determinar la composición química de la nucleína se requería un mayor grado de pureza. Wilhelm Kühne había descrito un procedimiento para romper las células con el uso de una enzima digestiva, la pepsina, que disuelve el citoplasma sin atacar al núcleo. No existiendo enzimas comerciales, era necesario lavar estómagos de cerdo con ácido clorhídrico diluido y filtrar los contenidos extraídos. El tratamiento de las células con esta solución, dejaba la nucleína libre de proteínas. El análisis de la composición elemental de la nucleína indicó finalmente que estaba formado de carbono, oxígeno, hidrógeno y nitrógeno (elementos que abundan en las proteínas), pero no tenía azufre y sí presentaba grandes

cantidades de fósforo. Este último constituía un hallazgo sorprendente, pues no se conocía ninguna otra molécula orgánica que contuviera fósforo.

Richard Altman mejoraría la técnica para aislar el compuesto libre de proteína, y lo denominaría ácido nucleico. Con esta misma técnica Albretch Kossel y su equipo avanzaría en el conocimiento de su composición. El ácido nucleico se revelaría contituído por bases, dos de ellas ya conocidas, la guanina y la adenina, otras dos ahora caracterizadas, la timina y la citosina[4]. Este mismo grupo reconocería además la existencia de dos tipos de ácidos nucleicos, uno obtenido fácilmente del timo y el otro desde levaduras. Alberto Ascoli reconocería al mismo tiempo la presencia de otra base, el uracilo, en uno de estos componentes.

Con esta información resultaba aún muy difícil inferir qué componente del núcleo podría servir como material hareditario. El grupo de Kossel, como tantos otros, intentaba caracterizar componentes celulares completamente desconocidos hasta entonces. Así describirían las histonas asociadas al ácido nucleico y varios aminoácidos que componen estas proteínas.

Feodor Levene, un investigador ruso sumamente productivo (más de 700 publicaciones entre 1909 y 1930) investigó propiedades de muchos componentes celulares, con particular atención en los azúcares. De sus análisis de los dos ácidos nucleicos Levene concluía la existencia de un azúcar de distinta naturaleza en cada uno de ellos, ribosa en uno y desoxiribosa en el otro. Levene proponía la denominación de ácido ribonucleico (ARN) y ácido desoxirribonucleico (ADN) que llevan desde entonces. Los análisis de la composición química de los ácidos nucleicos indicaban para este autor que estarían constituidos por cantidades equimolares de cada base, resultando de una simple repetición de los cuatro nucleótidos. Levene especulaba que estas moléculas tendrían una función semejante al glucógeno, es decir alguna forma de almacenamiento. Una molécula tan simple no podría ser portadora de la información genética.

El material genético debería adoptar formas diferentes para representar la gran cantidad de funciones que se esperaba de los genes. Las proteínas eran ya conocidas como moléculas poliméricas, muy variables, y estaban formadas por combinaciones diferentes de 20 aminoácidos químicamente diferentes. La fuente de información genética debía necesariamente basarse en proteínas.

El ADN es el material genético

La idea del ADN como portador de la información genética tomó forma gradualmente, impulsada en parte por una experiencia muy original realizada por Frederick Griffith. Su propósito no estuvo relacionado con la naturaleza del gen

4 Guanina y adenina responden a su aislamiento desde el guano (excremento de aves) y el páncreas (griego, aden) respectivamente. Timina y citosina fueron ambas aisladas originalmente del timo.

sino con la comprensión de la inmunidad conferida por la bacteria responsable de una enfermedad entonces mortal, hoy conocida como *Streptococcus pneumoniae*.

Antes del descubrimiento de los antibióticos la neumonía se trataba con antisueros generados inyectando bacterias muertas por calor en un animal. En 1923 Griffith inyectaba ratones con cepas de *S. pneumoniae* y descubría que existen cepas de capaces de generar enfermedad y otras que no lo hacen. La mayor diferencia entre ellas puede percibirse mediante la observación de colonias crecidas en placas de cultivo. Las cepas causantes de enfermedad poseen una apariencia lisa, denominada "S" (del inglés "smooth"), debido a la presencia de una cápsula de polisacáridos, en tanto que las cepas no patogénicas poseen una apariencia rugosa, "R" (del inglés "rough"), debido a la carencia de dicha cápsula.

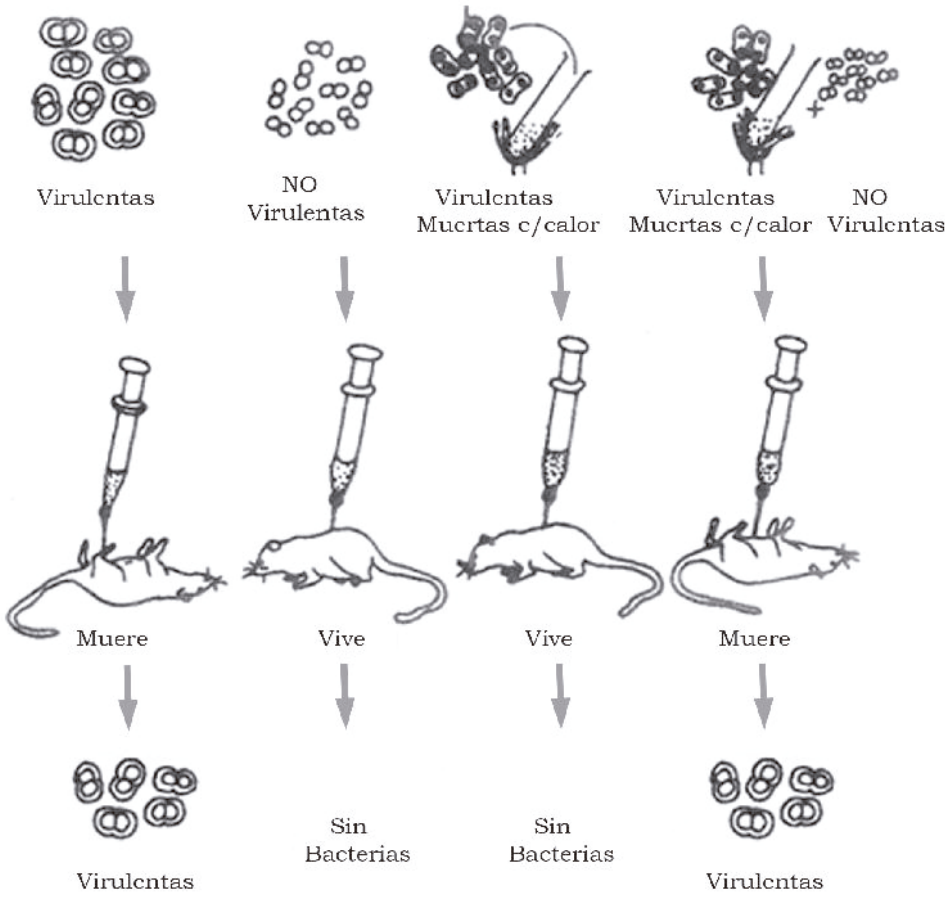


Figura 4. Experimento de F. Griffith con *S. pneumoniae*. La inyección de una cepa virulenta, tipo S, en un ratón es causa de neumonía y muerte del animal. Desde los animales muertos pueden aislarse las mismas bacterias patogénicas. La inyección de una cepa R, o de una cepa S inactivada por calor, no tiene efecto sobre la salud del animal. La inyección de una mezcla de bacterias virulentas muertas y no virulentas vivas lleva a la muerte del animal y desde el mismo pueden aislarse bacterias virulentas. En este caso las bacterias no virulentas debieron volverse virulentas gracias a un principio presente en las bacterias virulentas muertas por calor.

Trabajando con cepas S y R Griffith obtenía en 1928 resultados inesperados. Como en sus experimentos previos, la inyección de una cepa "S" de *S. pneumoniae* en un ratón permitía el desarrollo de neumonía, a la que sobrevenía la muerte del animal. Por el contrario, tras la inyección de una cepa "R" el animal no mostraba signos de enfermedad y permanecía vivo (Figura 4). Lo mismo ocurría cuando inyectaba una cepa "S" muerta por calor. Sorpresivamente, tras la inyección de una mezcla de bacterias virulentas muertas y avirulentas vivas los ratones también morían. Tras la biopsia de estos animales, podía comprobar que, como en la infección por una cepa "S", la muerte sobrevenía tras la invasión de las vías respiratorias y, de manera inesperada, las cepas aisladas de estos animales mostraban una forma "S" en cultivo y el mismo tipo inmunológico que las cepas virulentas.

Griffith dedujo que habría tenido lugar la conversión de una cepa no patógena, en una patógena, gracias a la existencia de un factor proveniente de las células virulentas muertas. Este fenómeno, conocido desde entonces como transformación, es permanente y heredable ya que las bacterias obtenidas de un ratón muerto y reintroducidas en uno vivo también ocasionan su muerte. Griffith, como muchos otros, no reconoció que este principio transformador era el material genético.

> *"The evidence presented supports the belief that a nucleic acid of the desoxyribose type is the fundamental unit of the transforming principle of Pneumococcus Type III."*
>
> *O.T. Avery, C. MacLeod and M. McCarty. 1943*

Recién en 1944 el trabajo descrito adquiriría real importancia. Avery, Macleod y McCarty, retomarían y extenderían los resultados de Griffith con el propósito de descubrir la naturaleza del material transformante (Avery *et al.* 1944). Trataron extractos conocidos por tener el principio transformante con enzimas proteolíticas (quimotripsina y tripsina) y vieron que este tratamiento no alteraba su capacidad de transformar las bacterias inocuas. De manera semejante, el tratamiento con una enzima ribonucleasa (que degrada ARN) tampoco alteraba la capacidad transformante de ese material. Sin embargo, el tratamiento del mismo extracto con una enzima desoxirribonucleasa tendría como consecuencia la pérdida de su actividad. El ADN resultaba así el principio capaz de transmitir propiedades heredables,

Aunque los experimentos de transformación fueron realizados correctamente, muchos científicos dudaban de su veracidad. Purificar enzimas implicaba entonces un enorme trabajo y podía sospecharse la presencia de actividades mezcladas, en particular de proteasas entre las nucleasas, o dudar acerca de la especificidad de cada enzima. Casi diez años más tarde un experimento de diseño muy diferente permitió sostener esta tesis, que el ADN es el material genético.

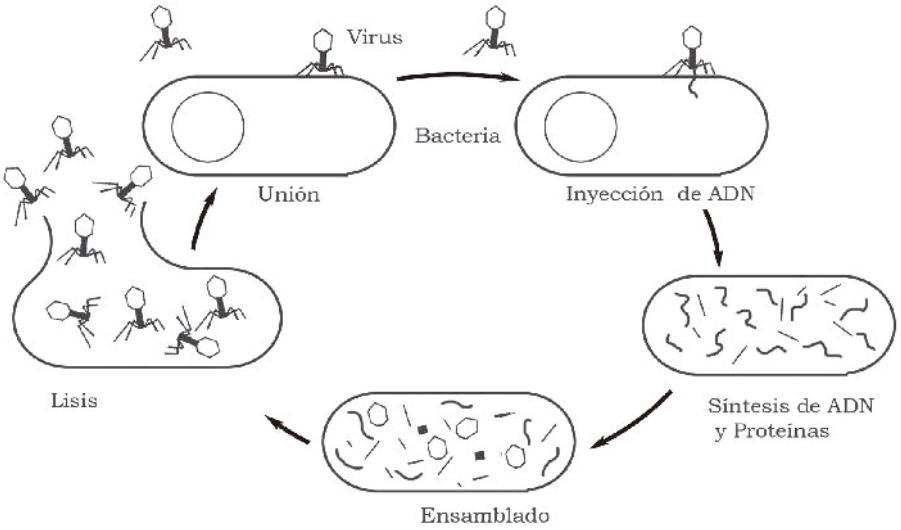


Figura 5. Ciclo lítico de infección de un bacteriófago. Después de la adhesión del fago a la cubierta de la bacteria la partícula viral inyecta su ADN en la célula. La expresión de este material genético fago determina la replicación del mismo y la síntesis de proteínas necesarias para el ensamblado de nuevas cápsides. A la lisis de la bacteria y la liberación de un gran número de partículas infecciosas se suceden nuevas infecciones.

Alfred Hershey y Martha Chase, estudiaban el ciclo de infección de los bacteriófagos. Un fago es una estructura relativamente simple constituídas por un fragmento de ADN envuelto por una cápside proteica. La existencia de un fago se basa en la infección de una célula bacteriana y la producción de nuevas partículas infecciosas (Figura 5). La adhesión de una partícula viral a una bacteria es seguida por la inyección de su ADN. La información contenida en los genes del fago es expresada por la bacteria, y empleada en la síntesis de nuevos fagos. Tras la replicación, nuevos fagos salen de la bacteria y pueden infectar a su vez otras bacterias.

Hershey y Chase diseñaron un experimento para determinar si toda la partícula viral ingresa en la bacteria, o si parte de esta partícula viral queda fuera de la célula. Si ingresaba sólo un componente del fago - ADN o proteína - entonces ese componente debía ser el responsable de la transmisión de la información, es decir el material genético (HERSHEY & CHASE 1952).

Para responder a esto emplearon marcadores radiactivos, recientemente introducidos en la biología experimental. Un isótopo radiactivo del fósforo, el ^{32}P, se empleó para marcar específicamente el ADN, y un isótopo del azufre, el ^{35}S, para marcar las proteínas. Hershey y Chase infectaron bacterias *Escherichia coli* con fagos T2 marcados, y pasado el tiempo necesario para la infección separaron las bacterias de los componentes ligeramente adheridos.[5] Luego, separaron bacterias y virus mediante centrifugación y midieron la cantidad de radioactividad presente en las células.

5 Algunos experimentos notables tienen una parte crítica que requirió mucho esfuerzo. La remoción de las cabezas de los fagos de la cubierta de las bacterias, sólo pudo lograrse mediante el empleo de una licuadora de cocina.

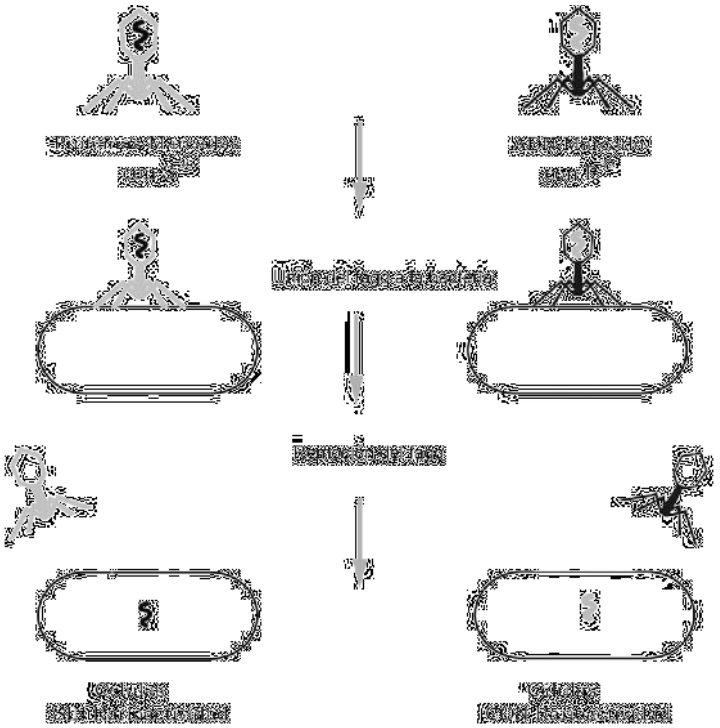


Figura 6. Experimento de Hershey y Chase. Se produjeron partículas virales en presencia de ^{35}S (izquierda) o ^{32}P (derecha) para marcar selectivamente las proteínas o el ADN. Tras la infección de células de E. coli con estas partículas, el cultivo se agitó vigorosamente para desprender las cápsides vacías y fagos no infectantes. Partículas y células se separaron por centrifugación y se midió el contenido de radioactividad en las células. El material responsable de la producción de nuevos fagos debería quedar en la célula.

Bacterias infectadas con fagos marcados con ^{35}S, guardarían sólo pequeñas cantidades de radioactividad, menos del 20% de la empleada. Por el contrario, las bacterias guardaban gran parte (el 70%) del componente marcado con ^{32}P, el ácido nucleico (Figura 6). Hershey y Chase midieron asimismo la radioactividad contenida en los fagos liberados tras la lisis de las bacterias infectadas y pudieron estimar que éstos contenían casi la mitad del ADN de los fagos originales, aunque menos de 1% de la proteína de los mismos.

Estos resultados sugerían que el ADN es el principal componente de un fago que ingresa a la célula bacteriana. Este componente se revelaba como el principal, sino el único, que se transmite a los nuevos fagos.

Asimilar resultados obtenidos con estas formas de vida tan inusuales a organismos celulares podría ser erróneo según algunas opiniones. Para la gran mayoría de los investigadores sin embargo, el experimento proporcionaba pruebas convincentes de que el material genético de un fago es el ADN. Cómo podría el ADN cumplir esta función? La pregunta era para todos, los experimentos decisivos para unos pocos, la interpretación y su comunicación para dos solamente.

La doble hélice

El uso corriente de isótopos radioactivos en los años 50 permitió cuantificar correctamente las cuatro bases y dejar de lado la a teoría de la equimolaridad de bases de Levene. Erwin Chargaff separó los productos de la hidrólisis del ADN mediante cromatografía en papel recientemente desarrollado y cuantificó la abundancia relativa de cada base radiomarcada. La cantidad de guanina apareció igual a la de citosina, y la cantidad de adenina proporcional a la de timina (Chargaff *et al*. 1951, Elson & Chargaff 1952). Aunque Chargaff no sugirió que esto podría indicar el apareamiento entre bases A-T y G-C), la proporcionalidad entre las bases es conocida desde entonces como regla de Chargaff.[6] Esta proporcionalidad se halló constante entre diferentes tejidos de un mismo organismo y distintos organismos de una especie. Sin embargo aunque la proporción (G+C)/(A+T) varía entre una especie a otra. Esta proporción de bases propia de cada especie se conoce como la segunda regla de Chargaff.

Para armar el rompecabezas encastrando bases, azúcares y fosfatos haría falta un molde tridimensional. La cristalografía de rayos X y algunas técnicas concurrentes proporcionarían uno. William Astbury había obtenido en 1938 un patrón de difracción de rayos-x de ADN deshidratado del cual deducía la presencia regular de un espacio de 3.34 Å (Amstrongs) a lo largo del eje de la fibra e interpretaba que este espaciado refleja una sucesión de nucleótidos apilados unos sobre otros. Cada nucleótido debería sobresalir perpendicularmente a lo largo del eje de la molécula y formar una estructura relativamente rígida.

Diez años después J. M. Gulland y D. O. Jordan intentaban titular grupos amino e hidroxilo mediante estudios de viscosidad y birrefringencia, y llegaban a la conclusión es necesario romper puentes de hidrógeno para separar hebras vecinas de ADN. Estudios de viscosidad ponían de manifiesto también que una molécula de ADN podía ser extremadamente larga. Durante su manipulación podía, o debía, evitarse ndo

Rosalind Franklin y Maurice Wilkins obtenían en 1953 patrones de difracción de rayos X que permitían inferir que el ADN tiene, dependiendo de la humedad relativa, dos estructuras secundarias posibles (Wilkins *et al.* 1953, Franklin & Gosling 1953). Una forma A obtenida a baja humedad y un patrón denominado B logrado a humedad elevada. En sus imágenes aparecía un patrón entrecruzado representativo de una estructura helicoidal, con reflexión regular cada 0.34 nm, indicativo del espaciamiento entre pares de bases adyacentes, y otro patrón a 3.4 nm que correspondería a un giro completo de la hélice (cada 10 bases). Franklin y Wilkins deducían que los grupos fosfato deberían estar expuestos al agua en el exterior de la hélice, y las bases se alojarían en el interior de la misma.

6 Algunos genomas de ADN de doble hebra poseen variaciones a esta regla (Szybalski et al. 1966).

No era y no es posible ver el ADN, sin embargo su estructura puede inferirse a través de los datos físicos y químicos acumulados. Proponer un modelo estructural acorde con los resultados experimentales y que explique a la vez cómo esta molécula tan simple puede actuar como reservorio de información era un evento inminente.

En el mismo número de la revista Science en el cual Franklin y Wilkins daban a conocer las imágenes de cristalografía, James Watson y Francis Crick proponen un modelo estructural para el ADN (Watson & Crick 1953b). En base a la información existente propusieron que la molécula tendría una estructura helicoidal de doble cadena. Dedujeron que esta díada es asimétrica, debido a que las cadenas que la componen son equivalentes aunque antiparalelas. Sosteniéndose en la evidencia proporcionada por Chargaf (el contenido de A es igual a T y el de G igual al de C) y en la prueba de la unión de bases mediante puentes de hidrógeno, Watson y Crick proponen el apareamiento entre A y T y entre G y C, cuyos enlaces llevan a la estabilización de la doble hebra. Dos puentes de hidrógeno estabilizan el par A-T y tres puentes de hidrógeno unen G con C.

La característica especial del esquema de apareamiento de bases propuesto consiste en que la geometría relativa de los enlaces que unen las bases a las pentosas es virtualmente idéntico para los pares A-T y G-C. Si una purina siempre se aparea con una pirimidina, entonces una secuencia irregular de bases en una cadena sencilla de ADN podría estar apareada regularmente en el centro de una doble hélice sin perder la simetría. Las reglas de Chargaff aparecieron como una consecuencia obligatoria de una estructura de doble hélice para el ADN.

Sin realizar un solo experimento Watson y Crick daban a conocer un modelo que explicaba muy fácilmente cómo una hebra sencilla del ADN podría actuar de molde para la síntesis de una segunda cadena (Watson & Crick 1953a). Dado que cada base de una cadena tiene una, y sólo una base complementaria, la secuencia de bases de una cadena determina automáticamente la de su pareja, la hebra complementaria. La información contenida en la hebra de ADN puede perpetuar mediante la simple apertura de la doble hélice y la copia de una o ambas hebras.

Tratar de resolver la estructura del ADN tiene un gran mérito y es considerado quizás el avance más importante en la biología durante el siglo 20. El modelo de Watson y Crick de la estructura del ADN fue aceptado rápidamente porque lograba resolver dos cuestiones importantes, reunía toda la evidencia química y física disponible en un solo modelo, y abría el camino para explicar cómo el ADN asumiría su función de portador de la información genética.

Características generales del ADN

El ADN es un polímero lineal constituido por cuatro nucleótidos químicamente distintos que pueden estar unidos entre sí en cualquier orden, conformando cadenas de cientos de miles, o incluso millones, de unidades de longitud.

Cada nucleótido en un polímero de ADN está formado por tres componentes: 1) un azúcar, la 2'-desoxirribosa, derivado de la ribosa en la que el grupo hidroxilo unido al carbono 2' ha sido sustituido por un hidrógeno; 2) una base nitrogenada, pudiendo ser una pirimidina –la citosina o la timina–, o una purina –adenina o guanina–. La base está unida al carbono 1' del azúcar por un enlace β-N-glicosídico desde el nitrógeno número 1 de la pirimidina o el número 9 de la purina. Finalmente 3) un grupo fosfato, que comprende una, dos o tres unidades de fosfato enlazados y unidos al carbono 5 ' del azúcar. Los fosfatos se designan α, β y γ, siendo α el fosfato unido al azúcar (Figura 7).

Una molécula compuesta por el azúcar y la base se denomina nucleósido. La adición de fosfatos convierte a este nucleósido en un nucleótido. Aunque las células contienen nucleótidos con uno, dos o tres grupos fosfato, sólo los nucleótidos trifosfatos actúan como sustratos para la síntesis de ADN. Los nombres químicos completos de los cuatro nucleótidos empleados en la polimerización del ADN son: 2'-desoxiadenosina 5'-trifosfato, 2'-desoxicitidina 5'-trifosfato, 2'-desoxiguanosina 5'-trifosfato y 2'-desoxitimidina 5'-trifosfato. Sus abreviaturas son respectivamente dATP, dCTP, dGTP y dTTP, y cuando estos se representan en una secuencia de ADN se resumen a la expresión de las letras A, C, G y T, respectivamente.

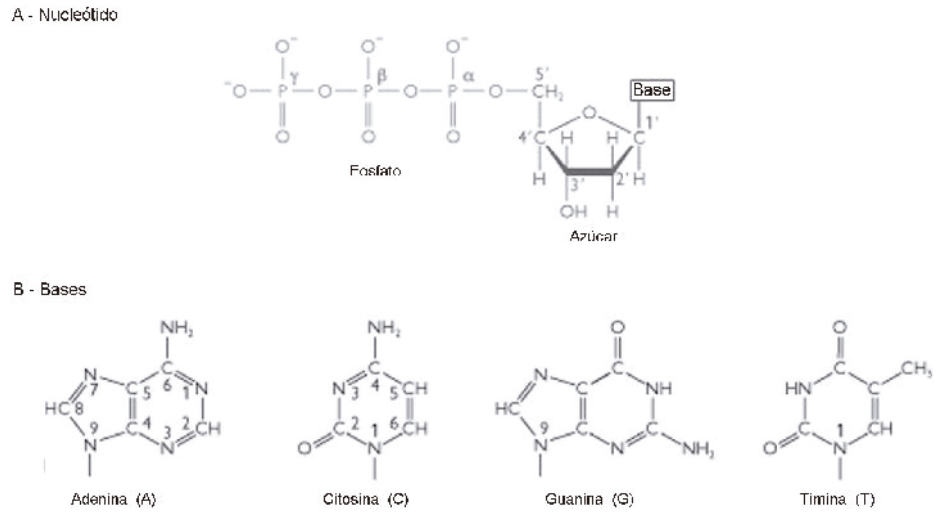


Figura 7. Estructura de un nucleótido. A) La estructura general de un desoxirribonucleótido, compuesto por desoxiribosa, uno a tres grupos fosfato ligados al carbono 5' y una base unida en la posición 1'. (B) Una de las cuatro bases esquematizadas está presente en un desoxirribonucleótido.

En un polinucleótido, nucleótidos individuales están unidos entre sí por enlaces fosfodiéster entre los carbonos 5' y 3'. Desde la estructura de este enlace se puede ver que la reacción de polimerización implica la eliminación de grupos fosfato β y γ del nucleótido y la sustitución del grupo hidroxilo unido al carbono 3' del segundo nucleótido. Los dos extremos del polinucleótido son así químicamente distintos, uno tiene un grupo trifosfato sin reaccionar unido al carbono 5' (5' terminal o 5'-P) y el otro tiene un hidroxilo sin reaccionar unido al carbono 3' (3' terminal o 3'-OH). Esto significa que el polinucleótido tiene un sentido, expresado como 5'→ 3' o 3'→ 5'. Consecuencia importante de esta polaridad del enlace fosfodiéster es que la reacción química necesaria para extender un polímero de DNA en la dirección 5'→ 3' es diferente a la necesaria para hacer una extensión en sentido 3 '→ 5'. Todas las enzimas ADN polimerasa halladas en la naturaleza son capaces únicamente de llevar a cabo la síntesis en la dirección 5'→ 3'. La misma limitación se aplica a la actividad de las enzimas que sintetizan ARN a partir de ADN, las ARN polimerasas.

Características distintivas del ARN

El ARN es un polinucleótido similar al ADN es aunque dos características químicas lo diferencian de este último. En primer lugar, el azúcar en un nucleótido de ARN –o ribonucleótido– es la ribosa. Luego, el ARN contiene uracilo en lugar de timina. Los cuatro nucleótidos sustrato para la síntesis de ARN son entonces: adenosina-5'-trifosfato, citidina 5'-trifosfato, guanosina 5'-trifosfato y uridina 5'-trifosfato. Estos nombres se abrevian a ATP, CTP, GTP y UTP, y en una secuencia se resumen a A, C, G y T, respectivamente.

Tal como ocurre en el ADN, los polinucleótidos del ARN contienen enlaces 3'-5' fosfodiéster. Este enlace es menos estable que en el ADN debido a la presencia del grupo hidroxilo en la posición 2' del azúcar. Por esta razón probablemente las funciones biológicas de ARN no requieren que un polinucleótido exceda unos pocos miles de nucleótidos de longitud.

Estructura de la doble hélice

El apareamiento de bases entre las dos hebras del ADN se basa en la formación de enlaces de hidrógeno entre cada adenina de una hebra y una timina en la otra, o entre una citosina y una guanina en cada hebra respectivamente. Los enlaces de hidrógeno son atracciones electrostáticas débiles (aproximadamente 1.10 kcal mol-1 a 25°C) entre un átomo electronegativo (tal como oxígeno o nitrógeno) y un átomo de hidrógeno unido a un segundo átomo electronegativo. El par guanina-citosina está unido por tres enlaces de hidrógeno en tanto que la

adenina se une a timina mediante dos enlaces del mismo tipo. Debido a la geometría de las bases y las posiciones relativas de los grupos que son capaces de participar en los enlaces de hidrógeno sólo estos pares son permisibles. En cada par pueden participar una purina y una pirimidina puesto que la unión de dos purinas determinaría una hélice de mayor tamaño y un par pirimidina-pirimidina sería demasiado pequeño. La especificidad de las combinaciones de pares de bases, explican las razones de base descubiertas por Chargaff.

El apilamiento de bases añade estabilidad a la doble hélice mediante interacciones hidrófobas entre pares de bases adyacentes. Estas interacciones, ocasionalmente llamadas π-π, surgen como un modo de alejar los grupos hidrofóbicos de los enlaces hidrógeno del agua circundante y confinarlos al interior de la molécula.

Tanto el apareamiento y como el apilamiento de bases tienen importancia en la estabilización de los dos polinucleótidos. El apareamiento sin embargo reviste especial importancia por sus implicaciones biológicas. La limitación del apareamiento de Adenina con Timina y de Guanina con Citosina, sugiere que la replicación del ADN puede resultar en copias perfectas de una molécula original. Para ello sólo basta con separar las hebras pre-existentes y copiar cada una para finalmente obtener dos nuevas cadenas idénticas a la original. Todas las ADN polimerasas celulares siguen este procedimiento de lectura de cada base de una hebra simple y añadido de nucleótidos complementarios para la síntesis de ADN. De manera similar, las polimerasas de ARN siguen este modelo para la síntesis de ARN y la transcripción de la información biológica contenida en las secuencias de la molécula de ADN genómico. Una diferencia entre la síntesis del ADN y la del ARN es que en la segunda la presencia de adenina especifica la presencia de uracilo en lugar de timina.

Variaciones del modelo de la doble hélice

La doble hélice descrita por Watson y Crick recibe el nombre de ADN-B. Los rasgos característicos de esta forma se resumen a un diámetro helicoidal de 2,37 nm, un paso de 0,34 nm por cada par de bases, y una distancia de 3,4 nm por cada revolución o vuelta de hélice, que corresponde a diez pares de bases. La forma B predominaría en las células vivas aunque cada nucleótido en la hélice tiene la flexibilidad necesaria para asumir formas moleculares ligeramente diferentes.

Para que la doble hélice adopte conformaciones diferentes las posiciones relativas de los átomos en el nucleótido deben variar. Entre un gran número de posibilidades los cambios conformacionales más importantes implican la rotación alrededor del enlace β-N-glicosídico, cambiando la orientación relativa de la base al azúcar, y la rotación alrededor del enlace entre los carbonos 3' y 4'. Ambas rotaciones tienen un efecto significativo en la doble hélice: cambiar la

orientación de base influye en el posicionamiento relativo de los dos polinucleótidos, y la rotación alrededor de la unión 3'-4' afecta a la conformación de la cadena principal de azúcar-fosfato.

Desde el año 1950 se han reconocido cambios en las dimensiones de la doble hélice. La llamada forma A de la doble hélice (Figura 8) tiene un diámetro de 2,55 nm, una distancia de 0,29 nm entre pares de bases y un paso de 3,2 nm, correspondiente a 11 pares de bases por vuelta. Variaciones semejantes se conocen como formas B', C, C', C", D, E y T del ADN. Todas estas hélices son dextrógiras como la forma B.

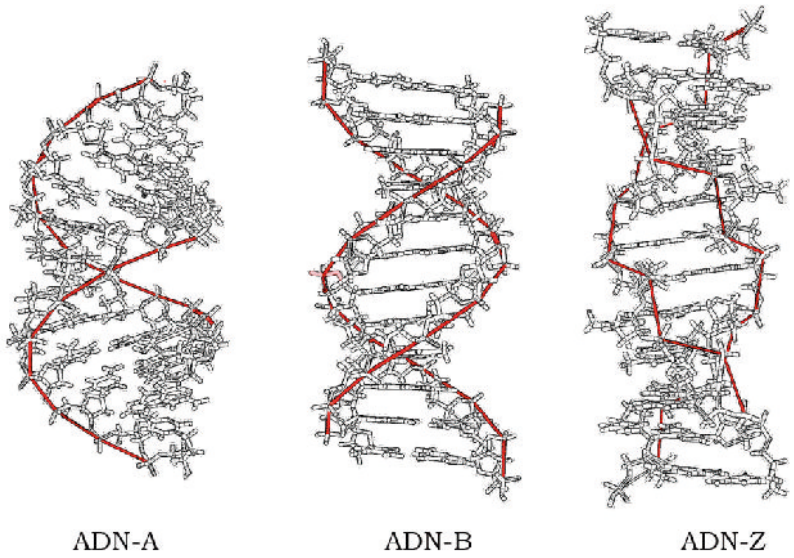


Figura 8. Formas del ADN. La forma dextrógira B del ADN, es la más común dentro de las células. La forma Z, levógira, ocurre ocasionalmente y está asociada con grandes regiones ricas en GC. La forma A puede detectarse en soluciones con escasa cantidad de agua. A diferencia de las anteriores, las bases no son coplanares en esta forma de hélice.

Es conocido que las moléculas de ADN genómico no son del todo uniformes en su estructura y que la sucesión de ciertos nucleótidos se asocia en ocasiones con formas diferentes. Una reorganización más drástica conduce a una forma levógira, conocida como forma Z, que posee asimismo un diámetro inferior, sólo 1,84 nm (Figura 8). Aunque muchos cambios en la estructura se producen en laboratorio cuando las moléculas de ADN se exponen a diferentes humedades relativas, *in vivo* estarían más relacionados con una secuencia particular de bases. La forma Z se asocia en general a secuencias ricas en GC (guanina y citosina).

Las diversas formas de la doble hélice no revelan las diferencias más significativas entre ellas o el sentido que se ha hallado a su existencia. El ADN-B posee dos ranuras en espiral a lo largo de la longitud de la hélice, cada uno de

ellos difiere en ancho y profundidad. Estos surcos se denominan así mayor y menor. El ADN-A posee también dos surcos y su conformación determina que el surco mayor sea más profundo que en la forma B, mientras el surco menor es más amplio y superficial. El ADN-Z es necesariamente diferente con un surco prácticamente inexistente y el otro muy estrecho y profundo.

Las diferencias en el tamaño de los surcos determinan el grado en que las regiones internas de la hélice son accesibles desde la superficie de la estructura. En cada forma del ADN la superficie interna de al menos una de las ranuras está formada por grupos químicos unidos a las bases de nucleótidos. La expresión de la información biológica contenida en el ADN es mediada por proteínas que se unen a la doble hélice y regulan la actividad de los genes contenidos en ella. Para llevar a cabo su función, cada proteína de unión al ADN debe reconocer sitios específicos en un gen o en su proximidad. Muchas enzimas y proteínas de unión al ADN pueden reconocer una estructura o conformación particular en la doble hélice y, sin necesidad de abrir y "leer" cada hebra, aproximarse de una secuencia de nucleótidos específica sin mucha ambigüedad.

Capítulo 2

Secuenciación del ADN

2

El ADN contenido en una célula guarda la información necesaria para la constitución de un organismo. De su lectura deberíamos poder responder a una gran cantidad de cuestiones, qué diferencias existen entre un individuo y otro, o entre dos especies, cómo enfermamos o cómo y cuándo dejamos de ser funcionales.

Las secuencias de ADN reflejan gran parte de las características reconocidas entre los organismos vivos, diferencias marcadas entre distintas especies y escasas variaciones entre individuos de la misma.

Desde comienzos de los años 1960 el secuenciado de ADN fue una meta a la que aspiraron mucho investigadores. El primer ácido nucleico secuenciado fue un ARNt ($tARN^{Ala}$) de levadura (Holley *et al.* 1965). El procedimiento consistía en hidrolizarlo parcialmente con enzimas y fraccionar los productos de reacción en columnas de intercambio iónico. La lectura de 76 nucleótidos requirió siete años de trabajo.

El grupo de Frederick Sanger no lograría esta primera secuencia pero, desde el mismo año, introdujo numerosas técnicas que sirvieron y predominaron en el secuenciado de ARN y ADN. Este equipo integró el marcado del ARN con ^{32}P, una técnica de fraccionamiento más rápida y la visualización de cada base separada mediante autoradiografía (Sanger *et al.* 1965). Para fraccionar los oligonucleótidos recurrieron a una electroforesis bidimensional en acetato de celulosa seguida de otra de intercambio iónico en papel.

La técnica del wandering spot (literalmente, el punto o mancha errante) consistía en digerir parcialmente el ácido nucleico con una exonucleasa y migrar los fragmentos mediante cromatografía bidimensional en papel. Representando cada punto a un fragmento que difiere del siguiente por un nucleótido, la lectura de puntos sucesivos proporciona la secuencia. El sistema se optimizó para distinguir cada base según la posición relativa respecto de sus vecinos (Brownlee *et al.* 1967). El método se emplearía en reacciones de secuenciación complejas, pero no era posible distinguir sin dudas las bases A y G.

Walter Fiers y su grupo lograron en 1972, leer mediante este método la primer secuencia de un gen en la cadena de ARN del bacteriófago MS2 (Jou *et al.* 1972). Dos años más tarde con una combinación de la misma técnica y del método de Maxam y Gilbert (ver más adelante), el mismo grupo lograría revelar los 3,569 nucleótidos contenidos en todo el genoma del mismo fago (Fiers *et al.* 1976).

En 1973 el grupo de F. Sanger publicaba sus resultados con un primer método el secuenciado de ADN sumamente laborioso, que implicaba añadir una a una cada base y verificar su incorporación mediante cromatografía (Sanger *et al.* 1973).

Dos años después el mismo grupo introducía un método de secuenciación basado en la síntesis mediada por la ADN polimerasa I de *E. coli*. Este consistía en reacciones paralelas de síntesis de ADN en las cuales uno de los cuatro nucleótidos se hallaba en concentraciones limitantes para la síntesis. La carencia de un nucleótido determina que la enzima interrumpa la síntesis en sitios en los que debe incorporarlo generando fragmentos de distinta longitud, todos ellos finalizando en el mismo nucleótido. Con este método se introdujo un elemento clave la separación de los fragmentos de ADN sintetizado en un gel de poliacrilamida permitía ver una escalera bandas desde la cual puede deducirse directamente la secuencia (Sanger & Coulson 1975). El procedimiento "más-menos" de secuenciación permitía obtener secuencias de hasta 80 nucleótidos de una sola vez. Aunque era extremadamente laborioso logaron secuenciar con este método la mayor parte de los 5.386 nucleótidos del bacteriófago φX174 (Sanger, Air, *et al.* 1977).

En 1977 dos grupos anuncian dos métodos completamente diferentes para secuenciar el ácido desoxirribonucleico. A. Maxam y W. Gilbert dan detalles de su método de degradación química y el grupo de F. Sanger muestra su versión enzimática mejorada, con nucleótidos terminadores de cadena (Maxam & Gilbert 1977; Sanger, Nicklen, *et al.* 1977).

SECUENCIACIÓN QUÍMICA

El método de degradación química desarrollado por Maxam y Gilbert se basa en el tratamiento de la molécula de ADN con diversos agentes químicos y el clivado de la cadena en posiciones específicas. El procedimiento comienza con el "marcado" del extremo 5' del ADN mediante la adición de un grupo fosfato radioactivo. Para ello se elimina el fosfato terminal mediante una fosfatasa y se transfiere un grupo ^{32}P desde el ATP marcado en posición γ, con ayuda de una enzima quinasa. Tras marcar el ADN, se desnaturalizaba la doble cadena y

se separaban ambas hebras en un gel de poliacrilamida. Habiendo marcado ambas hebras podía procederse al secuenciado de cada una de ellas por separado y confirmar la secuencia de una con la otra.

En el ADN monocatenario se emplea entonces un tratamiento químico que tiene como propósito el de generar la ruptura de la cadena en una proporción pequeña de bases. Con ácido fórmico se eliminan las purinas (A+G), con dimetil sulfóxido se metilan guaninas (y algunas adeninas), con hidrazina se metilan las pirimidinas (C+T), y esta misma reacción se limita a la modificación de las citosinas mediante adición de cloruro de sodio. Las cadenas pueden entonces cortarse en las bases modificadas con ayuda de piperidina. Las cuatro reacciones, con cortes en G, A+G, C y C+T, se migran en paralelo en un gel desnaturalizante de acrilamida. Si los cortes se produjeron con la frecuencia esperada, de aproximadamente uno por molécula, puede verse una serie de fragmentos, cada uno correspondiente al primer sitio de clivaje. Los fragmentos radioactivos pueden visualizarse en una placa sensible a los rayos X, y la secuencia deducida directamente mediante la lectura de las calles (Figura 9).

El método era muy laborioso pero permitía leer la secuencia de fragmentos de 100 bases en un solo experimento.

Figura 9. Método de secuenciación mediante degradación química de Maxam y Gilbert. La molécula de ADN bicatenario se marca con P radioactivo y se trata con agentes químicos que permitirán el clivado de la cadena en posiciones específicas. Las cuatro reacciones se migran en paralelo en un gel desnaturalizante de acrilamida en el que podrá verse una serie de fragmentos, cada uno correspondiente al primer sitio de clivaje. La secuencia se deduce de la lectura de las calles desde el extremo inferior del gel.

SECUENCIACIÓN ENZIMÁTICA

El método "didesoxi" o de "terminación de cadena" ideado por Sanger y Coulson procede mediante la síntesis enzimática de una cadena de polinucleótidos

complementaria. Durante esta síntesis, nucleótidos específicos van "terminando" prematuramente la nueva cadena generada. Esta secuencia puede "leerse" directamente en un gel como en el procedimiento anterior.

La secuenciación por terminación de cadena comienza con la preparación de moléculas de ADN de cadena sencilla como en método descripto previamente. A una de las hebras se hibrida un oligonucleótido desde el cual comenzará la síntesis. El secuenciado comenzará en la misma posición en cada molécula. Este oligonucleótido conocido como cebador (*primer*) es complementario al molde y es un paso crítico de la reacción (Figura 10).

La síntesis de la hebra es catalizada por una enzima ADN polimerasa y requiere la presencia de los cuatro desoxirribonucleótidos trifosfato (dATP, dCTP, dGTP y dTTP) como sustratos. En condiciones óptimas la enzima podrá copiar la hebra polimerizando una nueva cadena de varios miles de nucleótidos. Sin embargo, durante la secuenciación se añade a la reacción una pequeña proporción de un didesoxirribonucleótido (ddATP, o ddCTP, o ddGTP, o ddTTP) junto a los cuatro dNTPs. La ausencia de un grupo 3'-hidroxilo necesario para formar el enlace con el siguiente nucleótido bloquea la síntesis en este nucleótido (Figura 10). Dado que la ADN polimerasa no discrimina entre dNTPs y ddNTPs, un didesoxiribonucleótido puede ser incorporado sin inconvenientes en la cadena en crecimiento.

En una reacción que contiene ddATP, la cadena sintetizada terminará en posiciones opuestas a timidinas en la hebra molde. Así como es necesario controlar la actividad de los productos empleados durante la secueciación de Maxam y Gilbert, en un experimento de terminación de cadena la concentración del ddNTP determinará en gran parte la eficiencia de la lectura. Debido a que hay una gran proporción de dATP la síntesis de la hebra procederá varios cientos de nucleótidos antes que una molécula de ddATP sea incorporada y termine la síntesis. El resultado será entonces un conjunto de nuevas cadenas, de diferentes longitudes, terminadas en ddATP. En reacciones semejantes se sintetizan hebras que terminan con ddCTP, ddGTP y ddTTP. La migración de cada reacción en paralelo en un gel de poliacrilamida permite descifrar la secuencia de ADN desde las posiciones de las bandas en el gel. La banda que se ha movido más lejos representa el fragmento más pequeño de ADN. La sucesión de bases puede inferirse de la lectura de cada banda y su adyacente entre las cuatro calles.

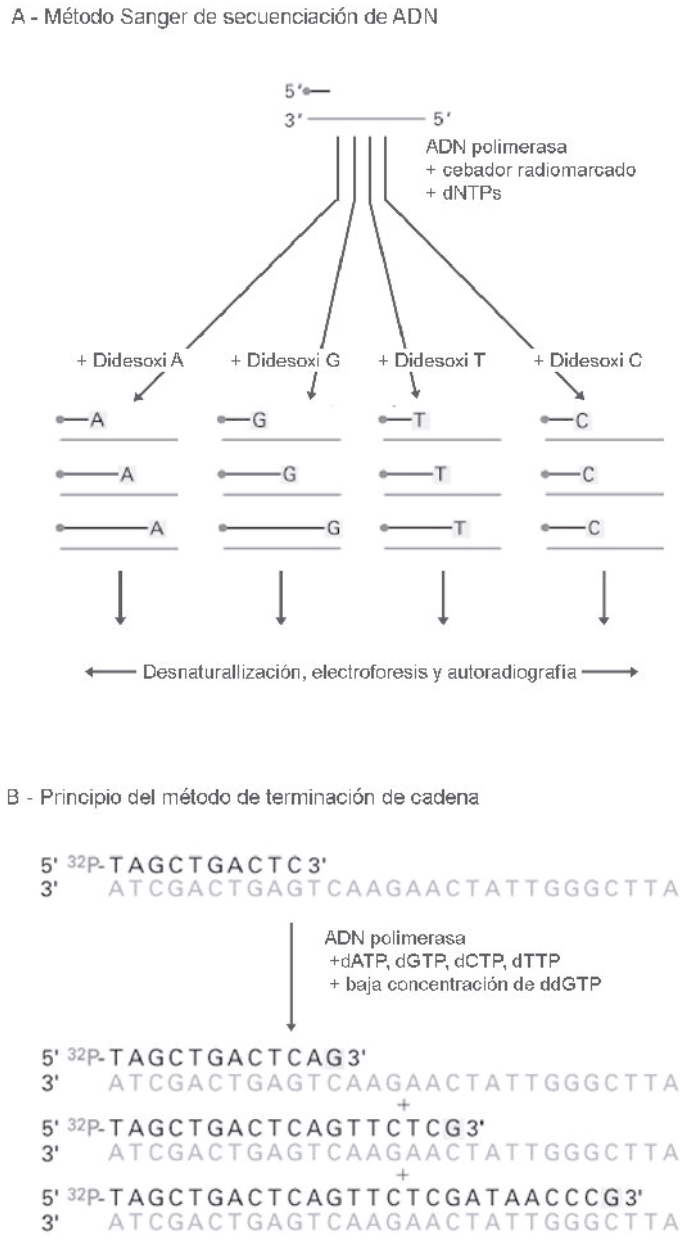


Figura 10. Secuenciación según el método de terminación de cadena de Sanger y Coulson. (A) La secuenciación de un fragmento de ADN consta de cuatro reacciones de síntesis en las cuales se ha incorporado un ddNTP diferente. La reacción conteniendo ddATP terminará la cadena cada vez que deba incorporarse dATP y en su lugar se añada ddATP. De manera semejante las reacciones con ddGTP, ddTTP o ddCTP interrumpirán la reacción de síntesis en cada uno de los sitios en que se hallen C, A o G en la hebra molde. (B) Un oligonucleótido complementario de la hebra molde (cebador o primer) permite iniciar la síntesis con la ADN polimerasa. El marcado del oligonucleótido con el radioisótopo P32 permitirá conocer la altura de la hebra sintetizada mediante autorradiografía del gel de secuenciación. La incorporación de una baja proporción de ddGTP en la mezcla de dNTPs posibilita la interrupción aleatoria de la síntesis y la generación de fragmentos de diferente longitud, todos ellos terminados en G (C en la hebra molde).

EVOLUCIÓN PARALELA

Un trabajo y una técnica en particular son aceptados en mayor o menor medida según el momento en el cual hacen su aparición. El método de secuenciación por medio de la degradación química llegó rápidamente a cada laboratorio. El mismo año en que se diera a conocer este método se obtuvo la secuencia del primer gen humano conocido, el lactógeno placentario (HPL, human placental lactogen) (Seeburg *et al.* 1977).

La degradación química presenta la gran ventaja de no requerir conocimiento alguno de la secuencia a leer. Sin embargo el control preciso de cada reacción siempre requirió gran destreza y el desarrollo del método alternativo llevó al desuso del método químico. El secuenciado de oligonucleótidos tuvo lugar durante mucho tiempo de esta manera y su fundamento está presente en la investigación de patrones de metilación así como en las técnicas para la localización de sitios de unión de proteínas al ADN (footprinting).

El desconocimiento de las secuencias necesarias para diseñar y sintetizar cebadores fue un gran impedimento para el desarrollo inicial del método de terminación de cadena. Sin embargo, simultáneamente con el surgimiento de los métodos de secuenciación se desarrollaban técnicas y herramientas que acompañarían y propulsarían este método en particular.

Manipulación del ADN y enzimas de Restricción

El desarrollo del secuenciado de ADN sobrevino junto al de muchas otras técnicas, hoy corrientes en el laboratorio de biología molecular. El mayor impacto en esta área resultó quizás del descubrimiento de las enzimas de restricción.

Las enzimas de restricción fueron descubiertas durante investigaciones sobre la especificidad de la infección de los virus bacterianos. Estas enzimas forman parte de un mecanismo de defensa de las bacterias. Cuando una cepa de *E. coli* carece de una enzima de restricción y es infectada por un fago, la mayoría de las partículas del virus pueden iniciar una infección exitosa. Por el contrario, cuando la misma cepa contiene una enzima de restricción, la bacteria resiste y la infección fracasa.

La resistencia bacteriana a la invasión viral se basa en el corte e inutilización del ADN inyectado por el fago en la célula receptora. Una enzima de restricción reconoce una secuencia corta, generalmente continua, y cliva al ADN dentro o en un sitio muy próximo de la misma. Las primeras enzimas de restricción purificadas provienen de *E. coli* y *Haemophilus influenzae*, y recibieron así los nombres de EcoRII, EcoRI, HindII y HindIII (Danna & Nathans 1971; H. O. Smith and Welcox 1970; Arber & Linn 1969). EcoRI reconoce una secuencia 5'-GAATTC-3' y corta esta secuencia entre G y A, dejando dos fragmentos con extremos 5'-*G*-3' y 5'-*AATTC*-3' (Figura 11).

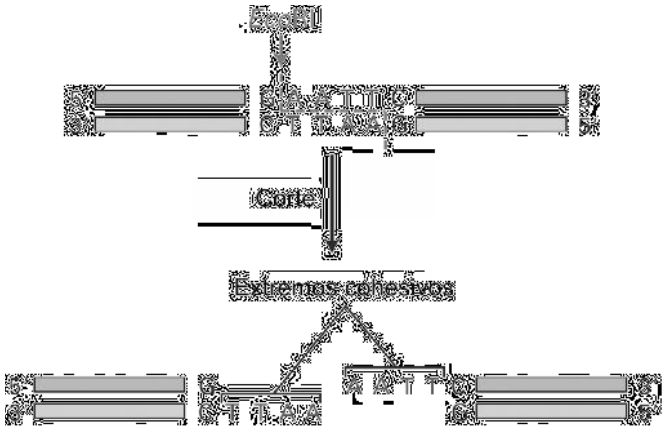


Figura 11. Corte de un segmento de ADN con la enzima EcoRI. Esta enzima reconoce segmentos con la secuencia 5'-GAATTC-3' y corta ambas hebras dejando dos extremos "pegajosos" de cuatro bases.

Para proteger su propio ADN de la escisión por las enzimas de restricción la bacteria posee enzimas que lo modifican, permitiendo reconocerlo del ADN viral. Estas enzimas reconocen la misma secuencia que ciertas enzimas de restricción y anaden un grupo metilo en una base de cada hebra del ADN. Tras la replicación, una hebra metilada sirve a la metilasa de referencia para agregar un metilo en cada posición de la nueva hebra. Las bases metiladas previenen así el reconocimiento de la secuencia por la enzima de restricción. Ambas actividades, restricción-modificación, pueden residir en proteínas separadas que actúan independientemente, o en concierto, como dominios de un complejo multiproteico.

Todos los procariotas estudiados hasta el momento poseen genes que codifican para enzimas de restricción. El secuenciado de genomas completos indica que son comunes a todas las bacterias y arqueobacterias de vida libre.

En la práctica, una enzima de restricción corta un segmento de ADN generando fragmentos de tamaño definido acorde a la existencia de sitios de corte. Estos fragmentos pueden ser modificados, purificados y unidos con otros fragmentos con ayuda de otras enzimas que tienen como sustrato el ADN. Las fosfatasas remueven grupos fosfato del extremo de una hebra de ADN, evitando su unión posterior; las quinasas permiten por el contrari, la adición de un grupo fosfato; y una ligasa permite unir de manera estable dos fragmentos. A fines de los años 70 estas nuevas herramientas llegaban ya a todos los laboratorios de biología molecular más importantes y se volvería cada vez más común el ADN artificial y los organismos genéticamente modificados.

Clonado y ADN recombinante

Dado que el ADN de todos los organismos comparte la misma estructura química, era razonable pensar en tomar un fragmento de un organismo, e introducirlo en otro. Morton Mandel and Akiko Higa, dos científicos de la Universidad de Hawaii, indicaban en 1971 que *E. coli* puede captar ADN exógeno

en presencia de altas concentraciones de calcio iónico en el medio (Mandel & Higa 1970). Este shock osmótico es aún el método habitual de introducir ADN exógeno en una bacteria.

El ADN bacteriano está, en general, contenido en un único fragmento circular, el cromosoma. Muchas bacterias contienen asimismo uno o más fragmentos de ADN extracromosómicos autoreplicantes llamados plásmidos o episomas. Estos poseen genes no asociados a la constitución del microorganismo pero confieren a la bacteria alguna ventaja selectiva, como la resistencia a un antibiótico.

En 1972, estudiando la resistencia a antibióticos, el grupo de Stanley Cohen introdujo plásmidos aislados de distintas fuentes en *E. coli* (Cohen *et al.* 1972). Este procedimiento volvió resistentes a la tetraciclina, kanamycina, neomicina o sulfonamida las bacterias antes sensibles y esta propiedad pudo mantenerse indefinidamente en los nuevos microorganismos.

Los plásmidos representaban una herramienta asombrosa. Estos elementos tienen un tamaño reducido comparado con el cromosoma bacteriano, y pueden replicarse dentro de la bacteria logrando en algunos casos un gran número de copias. Un fragmento de ADN podría ser ligado en un plásmido amplificarse a voluntad dentro de una bacteria (según su capacidad de replicar y la de la bacteria de contenerlo).

Los primeros intentos por reunir dos fragmentos de ADN de fuentes diferentes y crear secuencias que no se encuentran de otro modo en una forma de vida tienen lugar en 1972. El grupo de Paul Berg separó un fragmento del genoma del virus SV40 y lo ligó en el genoma del bacterriófago λ (Jackson *et al.* 1972). Paul Berg advertía en ese mismo momento acerca del peligro de trabajar irracionalmente con estas nuevas estrategias que podrían llevar a la creación, entre otros, de patógenos resistentes a antibióticos o incluso de nuevas formas de vida.

> *"Recombinant DNA molecules constructed in vitro from separate plasmids by the joining of DNA fragments having cohesive termini generated by the Eco*R*1 restriction endonuclease can form biologically functional replicons when introduced in to Escherichia coli by transformation."*
>
> *Boyer and Cohen*

En 1973 Stanley Cohen y Herbert Boyer (quien aislara la enzima Eco RI) cortan y ligan dos fragmentos de ADN de distinto origen y demuestran que ambos son funcionales en *E. coli*. La reunión de dos plásmidos con genes de resistencia a diferentes antibióticos volvía a la bacteria insensible a la presencia de ambos antibióticos en el medio de cultivo (Cohen *et al.* 1973).

Dando un paso más adelante, el mismo grupo probaría que una bacteria puede expresar ADN eucariota. Cortaron un fragmento de ADN ribosomal de la rana *X. laevis,* lo ligaron en un plásmido (pSC101), lo introdujeron en *E.coli* y

demostraron la formación del ARN específico (Morrow *et al.* 1974). La creación de microrganismos recombinantes vería en adelante sólo un límite, la capacidad de la célula receptora para leer las nuevas secuencias. La producción de proteínas de otros organismos, como la insulina humana, en un microorganismo como coli tenía un nombre, biotecnología.

EL SECUENCIADO ENZIMÁTICO REQUIERE UNA HEBRA SIMPLE

La secuenciación de ADN no parecía tomar mucha ventaja del uso de las técnicas del ADN recombinante. El perfeccionamiento de las técnicas de clonado permitiría tener más fragmentos de ADN de un solo tipo gracias a la amplificación del vector plasmídico dentro de la bacteria, pero era necesario aislar ADN monocatenario por desnaturalización con álcali o por ebullición. Luego ambas hebras podían ser separadas si tenían una longitud o un contenido de purinas significativamente diferente, en un gel de poliacrilamida o en un gradiente de cloruro de cesio. No era infrecuente que las hebras pudieran separarse y emplearse en el secuenciado.

Para la producción de hebras de cadena simple para la secuenciación del ADN se diseñaron Los vectores basados en el bacteriófago M13. Este fago tiene un genoma de ADN de cadena sencilla que se convierte en una forma replicativa de doble hebra dentro de *E. coli* (Figura 12). La forma replicativa da origen a más de 100 moléculas dentro de la célula y este número de copias se mantiene por medio de nuevos ciclos de replicación cuando la célula se divide. Las células infectadas liberan continuamente nuevas partículas de fago M13, aproximadamente 1000 por generación, que contienen la versión de cadena simple del genoma.

Los vectores de clonado basados en el fago M13 son moléculas de doble cadena de ADN equivalentes a la forma replicativa del genoma de M13 y pueden ser manipulados de la misma manera que un plásmido. A diferencia de este último las células transformadas con ADN recombinante basado en el fago M13, secretan permanentemente partículas del bacteriófago con ADN de cadena simple. Estos fagos proporcionan un molde muy útil para la secuenciación por terminación de cadena.

Un plásmido puede contener uno o más genes foráneos. Aunque no existe un tamaño límite para el ADN introducido, en la práctica se advierte que más de 10 kb resultan en una disminución considerable del rendimiento de la producción. La partícula viral M13 impone por el contrario, un tamaño máximo de ADN clonado. Su cápside proteica, como la de otros virus, tiene un tamaño definido donde el ADN viral es compactado. Cualquier incorporación de ADN se realiza a expensas de la eliminación de secuencias virales. Tras la eliminación

de algunos segmentos de ADN viral superfluo, la inserción de un fragmento de ADN exógeno se limita a aproximadamente 3 kb.

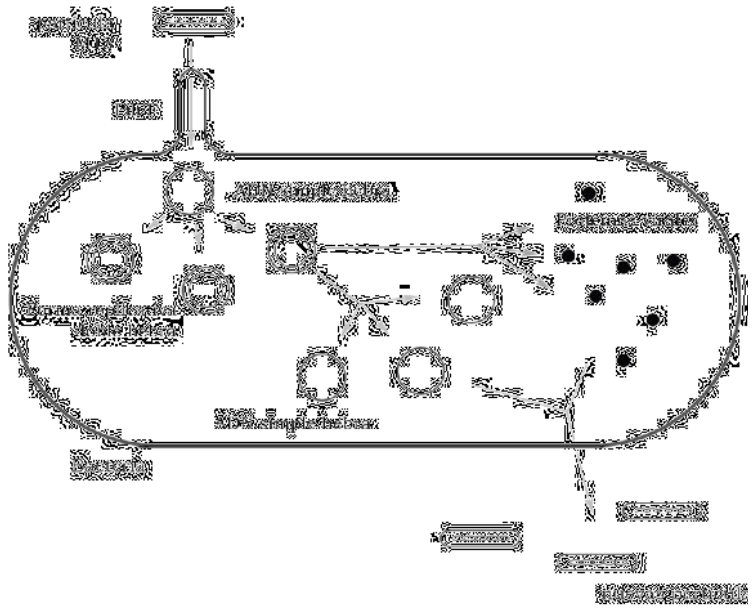


Figura 12. Formas del ADN en el ciclo infeccioso del bacteriófago M13. En partículas de bacteriófago se encuentra ADN de cadena simple que dentro de la bacteria va a adquirir su hebra complementaria. Esta molécula de doble hebra replicativa se conserva dentro de la bacteria y puede ser manipulada de manera semejante a un plásmido bacteriano.

Durante algunos años no se dispuso de muchas enzimas de restricción, y era deseable conocer más genes o regiones contiguas a los mismos. Para poder disponer de secuencias más largas debieron ingeniarse otros métodos basados en vectores basados en el M13 para incluir más cantidad de secuencia. Un fásmido (phasmid o phagemid) reúne la mayor capacidad de almacenamiento de un plásmido y la producción de ADN de cadena sencilla del fago M13 en un vector con fragmentos de ambos. Este plásmido que contiene secuencias del fago (origen de replicación para ADN monocatenario e incorporación en la cápside) se introduce en una bacteria *E. coli* junto a la forma replicativa de un fago auxiliar (contiene genes necesarios para la replicación del ADN y la formación de la cubierta del fago). Así pueden sintetizarse partículas de fagos que contienen ADN monocatenario foráneo proveniente del fásmido. Hasta 10 kb ADN de doble hebra pueden convertirse en ADN monocatenario que sirve de molde para la secuenciación.

Cebadores para el secuenciado enzimático

La secuenciación por el método de terminación de cadena requiere de la presencia de un oligonucleótido cebador, *primer* en inglés, hibridado a la hebra a secuenciar. Este cebador tiene la función indispensable de proporcionar un extremo 3' desde el cual la enzima añadirá un primer nucleótido. Ninguna polimerasa dependiente de ADN puede comenzar la síntesis de ADN en una molécula que es enteramente de hebra única.

En un comienzo, siendo todas las secuencias desconocidas, podían sintetizarse cebadores no específicos que eventualmente hibridaban. Con las enzimas de restricción estos cebadores sintéticos que funcionaban sólo ocasionalmente fueron reemplazados por fragmentos de digestión del mismo ADN a secuenciar. Estos fragmentos hibridaban correctamente aunque su gran tamaño limitaba en ocasiones la resolución de bandas necesaria para la lectura de la secuencia.

El empleo del método de Sanger y Coulson permitía conocer 15-200 bases en un experimento de secuenciación. Para continuar secuenciando de la misma hebra era necesario diseñar un nuevo cebador que hibridara con la región terminal del fragmento recientemente conocido. Este ciclo era muchas veces interrumpido por la dificultad de crear un cebador que hibride correctamente. Un cebador debe hibridar específicamente en la región requerida, y para esto debe reunir ciertas características de tamaño, contenido de GC y secuencia.

Con el uso de vectores y enzimas de restricción fue posible evitar el diseño continuo de cebadores. Se diseñaron plásmidos y vectores virales con sitios cebadores junto a los cuales podría clonarse cualquier fragmento a secuenciar. Estos cebadores universales permiten secuenciar cualquier fragmento desconocido (Heidecker *et al.* 1980; Anderson *et al.* 1980). Si el ADN insertado en el vector era más largo que la secuencia leída por la enzima se obtendría una parte de una secuencia desde ese extremo y otro fragmento de secuencia desde el extremo opuesto. Alternativamente, empleando otras enzimas de restricción pueden superponerse fragmentos. A partir de un gran número de secuencias cortas podía entonces inferirse la secuencia completa.

POLIMERASAS DE ADN

Clave para el secuenciado del ADN por el método de terminación de cadena es la disponibilidad de una ADN polimerasa. Una enzima que copia un ADN existente –o una molécula de ARN– para sintetizar ADN, se llama ADN polimerasa modelo dependiente. Una ADN polimerasa sintetiza un polinucleótido de ADN complementario a una hebra molde o parental, siguiendo las reglas de apareamiento de bases. El nuevo polinucleótido se sintetiza en sentido 5'→3'.

Para iniciar la síntesis de ADN debe haber al menos un corto fragmento de doble hebra que proporciona un extremo 3'. En la célula viva este cebador no es de ADN sino de ARN y es sintetizado por una ARN polimerasa denominada Primasa.

Gran parte de las ADN polimerasas hoy empleadas contienen varias funciones en la misma molécula. Algunas de ellas tienen una función de lectura de prueba (proofreading) o de verificación de la cadena sintetizada. Asociada a ésta suelen tener cierta capacidad de remover nucleótidos de la hebra que acaba de sintetizar, esto es una función exonucleasa 3'→5'. Por otra parte, muchas de ellas poseen una función exonucleasa 5'→3' para remover nucleótidos hibridados delante de la hebra en síntesis. Esta última actividad se limita a veces a la remoción de cebadores de ARN creados por la primasa.

La primer ADN polimerasa empleada en investigación fue aislada a partir de *E. coli* (LEHMAN *et al.* 1958). Esta enzima no asume la replicación del genoma de esta bacteria sino más bien el control y corrección de errores. Tiene actividades 3'→5' y 5'→3' exonucleasa, hecho que limitó su utilidad en la síntesis y el secuenciado de ADN.

El uso de la ADN polimerasa I de coli en la secuenciación de ADN fue posible gracias a la modificación de la enzima de Kornberg. El fragmento de Klenow se obtuvo mediante el corte de la ADN polimerasa I con una proteasa obtenida a partir de *Bacillus subtilis*, la subtilisina (Klenow & Henningsen 1970). El tratamiento da origen a dos fragmentos, uno de ellos retiene las actividades polimerasa y la exonucleasa 3'→5', y carece de la función exonucleasa 5'→3'. El fragmento de Klenow se hace aún desde *E. coli* aunque para evitar la proteólisis, las células contienen un gen carente de la región que determina la actividad exonucleasa 5'→3'. La polimerasa de Klenow no se emplea ya en secuenciación aunque sí es útil en el marcado (radioactivo o fluorescente) de ADN.

Cualquier DNA polimerasa modelo dependiente es capaz de extender un cebador asociado a ADN monocatenario. Sin embargo, no todas las polimerasas pueden resultar útiles para la secuenciación del ADN. Las enzimas elaboradas para la secuenciación deben cumplir algunos criterios mínimos. Deben ser altamente procesivas, es decir deben poder sintetizar polinucleótidos de gran tamaño antes de terminar por alguna razón la síntesis, deben tener una actividad 5'→3' y 3'→5' exonucleasa nula o insignificante. El fragmento de Klenow sintetiza fragmentos cortos de ADN y por ello la lectura no se extiende más de 200-300 bases. La actividad 5'→3' exonucleasa permite a la polimerasa eliminar ADN que ya está unido a la hebra molde pero también nucleótidos de los extremos 5' de las hebras recién sintetizadas. Esta eliminación altera la longitud de los fragmentos y hace imposible leer la secuencia desde un perfil de bandas en un gel de poliacrilamida. La actividad exonucleasa 3'→5' es también indeseable puesto que la polimerasa podría eliminar la terminación de cadena y no permitir la lectura.

Desde la creación del fragmento de Klenow las DNA polimerasas han sido modificadas artificialmente. Por su alta procesividad muchas de las enzimas empleadas en adelante han derivado de ADN polimerasas virales. La ADN polimerasa codificada por el bacteriófago T7 tiene alta procesividad (puede añadir miles de nucleótidos antes de terminar la síntesis) y modificada, no posee actividad exonucleasa (Tabor & Richardson 1987).

La reacción en cadena de la polimerasa

En biología molecular, el desarrollo de cada nueva técnica deriva en gran medida de la comprensión de procesos celulares y del uso de las herramientas que nos proporciona la misma célula. Para experimentar cambios predeterminados en las moléculas de ADN, como cortarla en fragmentos más pequeños, y combinar estos fragmentos para generar formas que no existen en la naturaleza, se emplean enzimas celulares. El método de secuenciación de Sanger emplea una ADN polimerasa para extender un cebador, de manera semejante a como la célula extiende fragmentos de ADN o ARN durante la síntesis.

En 1985 se habían ideado prácticamente todas las estrategias actuales de clonado, pero algunos procedimientos, como cortar un fragmento de ADN desde una célula cualquiera e insertarlo en un vector, podían resultar extremadamente difíciles.

La reacción en cadena de la polimerasa (PCR)[7] aparece en ese momento como un procedimiento sumamente sencillo e ingenioso, mediante el cual se puede amplificar, a voluntad, un fragmento de ADN de doble hebra (Mullis *et al.* 1986). En el trabajo original la doble hebra de ADN, junto a un par de cebadores y desoxirribonucleótidos, se separaba calentando la solución a 100 ºC. Tras la desnaturalización del ADN se bajaba la temperatura para permitir la hibridación de dos oligonucleótidos sintéticos, los cebadores, a la secuencia diana. Se agregaba entonces el fragmento de Klenow de la ADN polimerasa I, y procedía la síntesis de dos nuevas hebras.

> ***Process for amplifying nucleic acid sequences***
> *The present invention is directed to a process for amplifying any desired specific nucleic acid sequence contained in a nucleic acid or mixture thereof. The process comprises treating separate complementary strands of the nucleic acid with a molar excess of two oligonucleotide primers, and extending the primers to form complementary primer extension products which act as templates for synthesizing the desired nucleic acid sequence. The steps of the reaction may be carried out stepwise or simultaneously and can be repeated as often as desired.*
> *Treating the strands with two oligonucleotide primers, for each different specific sequence*

7 La reacción en cadena de la polimerasa (PCR), es uno de los primeros procedimientos registrados en biología molecular. K. Mullis, uno de sus autores, y la corporación Cetus aún multiplican su dinero a un ritmo semejante al que puede amplificarse el ADN.

being amplified, under conditions such that for each different sequence being amplified an extension product of each primer is synthesized which is complementary to each nucleic acid strand, wherein said primers are selected so as to be sufficiently complementary to different strands of each specific sequence to hybridize therewith such that the extension product synthesized from one primer, when it is separated from its complement, can serve as a template for synthesis of the extension product of teshe other primer.

(Mar. 28, 1985)

Aunque la duplicación del ADN era un procedimiento corriente en laboratorios de esa época, la PCR mostraba su originalidad con el empleo de dos cebadores que flanquean la región a amplificar, y mediante la repetición de cada duplicación. Los oligonucleótidos deben unirse específicamente a cada lado de la región que se desea amplificar y deben permitir la síntesis de tal manera que sus productos sean convergentes. El producto de síntesis de un cebador servirá en la ronda siguiente de sustrato al otro cebador y viceversa.

Si la primera ronda de síntesis condujo a la duplicación de cada hebra utilizada como sustrato, repetir el ciclo de desnaturalización, hibridación de primers y síntesis -tras agregar nueva enzima-, permite duplicar el material ya duplicado, y tener $2^2 = 4$ moléculas. Repitiendo esta manipulación 20-27 veces, como se hiciera en el trabajo original, la manipulación resultaba en la nada despreciable cifra de 2^{20} a 2^{27} moléculas por cada una usada como sustrato (Figura 13).

Como otros, este ingenioso procedimiento, sigue un modelo biológico. Durante la replicación del ADN las hebras parentales son separadas por un complejo enzimático –aquí reemplazado por una alta temperatura-, otro grupo de enzimas sintetizará cebadores sobre estas hebras –los cebadores se añaden a la reacción- y una o más enzimas polimerasas sintezan entonces nuevas hebras complementarias. La repetición del ciclo halla su paralelo en la división continua de células hijas.

La reacción en cadena de la polimerasa es entonces la repetición del ciclo de duplicación del ADN, para un fragmento de dicho ADN cuyas secuencias flanqueantes son conocidas. Esta frase encierra a su vez la gran ventaja y el mayor inconveniente de la PCR. Esta técnica no sustituye al clonado de un fragmento de ADN pero facilita enormemente su realización pues permite lograr suficientes copias de un fragmento como para hacerlo "visible" en un gel (con un colorante). La amplificación a voluntad de un fragmento que se desea estudiar resultó entonces, revolucionaria. Por otra parte, la mayor limitación de la PCR redunda de la necesidad de conocer, al menos, las dos secuencias limitantes del fragmento a amplificar.

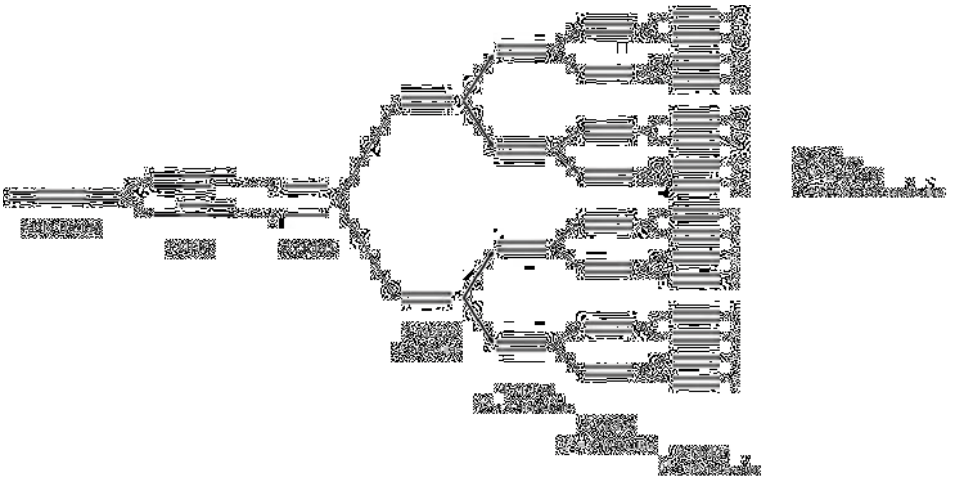


Figura 13. Amplificación del ADN mediante PCR. El esquema representa como puede duplicarse el contenido de ADN en cada roda de amplificación mediada por la polimerasa. Los primeros dos ciclos no representan exactamente el fragmento amplificado pues éste se extiende más allá de los límites impuestos por los cebadores. A partir del tercer ciclo puede contarse un número de moléculas de tamaño definido que se multiplicarán con cada ciclo. En condiciones óptimas el ciclado permitiría obtener 2^{n-2} moléculas.

Automatización de la PCR: Polimerasas Termoestables

Tanto la ADN polimerasa de *E. coli* como las enzimas obtenidas a partir de virus tienen una temperatura óptima de reacción cercana a los 37 °C, la temperatura adecuada para muchos procesos fisiológicos que ocurren en organismos homeotermos, sus comensales o parásitos (muchas bacterias, levaduras y virus). Aumentar la temperatura para desnaturalizar el ADN lleva a la desnaturalización de estas enzimas y por ello, en la PCR inicialmente concebida, era necesario incorporar nueva enzima antes de llevar adelante cada nuevo ciclo de duplicación.

La búsqueda de nuevas enzimas para el trabajo en laboratorio llevó en 1976 a la identificación de una ADN polimerasa de características muy particulares (Chien *et al.* 1976). Esta enzima hallada en la bacteria extremófila *Thermus aquaticus*, habitante de fuentes termales, tiene una temperatura óptima de trabajo de 80 ºC y resiste la temperatura necesaria en un laboratorio para separar las hebras de ADN. El grupo que ideara la PCR incorporó esta enzima en el sistema y demostró el gran potencial de la técnica mejorada (Saiki *et al.* 1988).

El descubrimiento de la polimerasa Taq (por ***T****hermus* ***aq****uaticus*), así como de otras ADN polimerasas termoestables, llevó a la universalización de la PCR. La automatización del procedimiento fue posible con la incorporación de termociclador (la máquina de PCR) (Weier & Gray 1988).

El nuevo método con sus mejoras no permitió solamente el clonado "a voluntad" de fragmentos de ADN. Entre otras mejoras permitió el secuenciado más rápido de fragmentos de mayor tamaño. En un experimento que conjuga el

clonado en un vector M13, el secuenciado con el método dideoxi de Sanger y el diseño de nuevos cebadores al finalizar cada ronda de secuenciación, Strauss *et al* lograron "marchar" a lo largo de un fragmento de 4 Kbp (Strauss *et al.* 1986). Este método será en adelante muy usado para el secuenciado de grandes genomas.

La PCR y las polimerasas termoestables se emplearon pronto en un nuevo método de secuenciación de terminación de cadena. El secuenciado en ciclo térmico tiene la gran ventaja de poder usar la doble hebra como sustrato y puede hacerse a partir de cantidades mínimas de ADN (Sears *et al.* 1992). Esto significa que ya no será necesario clonar un fragmento.

El secuenciado en ciclo térmico sigue los mismos principios que la PCR con la sola excepción de emplear un único cebador. De este modo se secuencia sólo una hebra. Dado que no hay dos cebadores el producto se acumula de manera lineal, y no exponencial como sucede en la PCR. Tal como en el secuenciado didesoxi, la presencia de ddNTPs en la reacción lleva a la terminación de cadena y el resultado se analiza mediante electroforesis.

Automatización del secuenciado: Fluorescencia

La secuenciación de ADN según el método de terminación cadena, tal como el método de degradación química, empleaba deoxirribonucleótidos marcados con un isótopo radioactivo. La migración en un gel de poliacrilamida permitía separar los fragmentos sintetizados y obtener un patrón de bandas, luego de exponer este gel a una placa de radiografía (autorradiografía). Revelar la señal de esta manera insume horas, el manejo de radioactividad es costoso y peligroso, y los deshechos radioactivos deben ser tratados de manera compleja.

Desde 1975 muchos laboratorios intentaban ya localizar genes en el núcleo de una célula mediante la visualización microscópica de fragmentos de ADN o ARN hibridados. Este procedimiento dependía hasta entonces, igualmente, de la utilización de sondas marcadas con radioisótopos. Hacia fines de la misma década algunos grupos incorporaban partículas que contenían moléculas fluorescentes y en 1980 llegó la gran innovación, la incorporación de nucleótidos marcados con un fluorocromo (Bauman *et al.* 1980). Esta mejora no tardó en llegar al secuenciado y favoreció la automatización del proceso.

Un único marcador fluorescente era añadido al cebador y la terminación de secuencia se producía por la incorporación de un didesoxirribonucleótido trifosfato (Ansorge *et al.* 1986; Smith *et al.* 1985). Mediante un detector láser situado en el extremo del gel podía "leerse", el paso de moléculas marcadas. Cuatro reacciones independientes, correspondientes a cada dideoxiribonucleótido terminador de cadena, permitían inferir la secuencia según el paso de los fragmentos de ADN terminados delante del detector.

Un nuevo progreso consistió en asociar cada ddNTP con un marcador fluorescente distinto, las cuatro reacciones podrían así realizarse en un solo tubo (Prober *et al.* 1987). Con derivados de la succinidilfluoresceína se obtenían marcadores con un espectro de emisión propio y muy cercano (505, 512, 519 y 526 nm). Un detector de fluorescencia podría discriminar los distintos espectros de emisión y asociarlos a una base particular, A, C, G o T). La secuenca podía desde entonces leerse directamente a medida que cada banda transitaba delante del detector en una sola calle.

El sistema de detección fluorescente proporciona un aumento importante en el rendimiento de cada experimento y permitió evitar errores que podían surgir de la lectura e interpretación de bandas. Junto a la incorporación de otras técnicas automatizadas, como de robots para la manipulación de muestras y la carga de geles, este sitema permitía ver la generación de datos lo suficientemente rápido como para poder leer todo un genoma en un plazo de tiempo razonable.

Diez años después de la publicación de los dos métodos de secuenciación de ADN, casi todos los esfuerzos se orientaban a la optimización de sólo uno de ellos, el método de Sanger. El desarrollo del método de Maxam y Gilbert vió dos grandes escollos, la imposibilidad de automatización y el uso de marcadores radioactivos. Aunque hasta el momento no se han hallado reactivos capaces de cortar específicamente adeninas y timinas, este parece no haber sido un límite. En 1988 se reportó un protocolo para el secuenciado automático con marcadores fluorescentes por el método de Maxam-Gilbert (Ansorge *et al.* 1988). Utilizando un marcador que no interfiere con la degradación de la molécula de ADN se logró secuenciar 50 oligonucleótidos de 20 bases cada uno en un solo gel. En los años siguientes no hubo muchos mas avances en este método. Esto se debió muy probablemente a razones comerciales.

Capítulo 3

Secuenciado del genoma humano

2

La molécula de ADN contiene la información necesaria para la constitución y funcionamiento de un organismo. Su característica distintiva es una secuencia de nucleótidos. Esta secuencia muestra gran parte de las características reconocidas entre los organismos vivos, diferencias marcadas entre distintas especies y escasas variaciones entre individuos de la misma. El secuenciado de genomas tiene como fin último el de revelar la función cada porción del mismo, y con éste el de comprender la evolución de los organismos y el principio mismo de la vida.

Conocer las 49.000 pb que componen el primer genoma descifrado completamente, el del fago λ, demandó cinco años de trabajo (Sanger *et al*. 1982). Sólo seis años después de este avance, ya se haría pública la intención de conocer las 3.000.000.000 pb del genoma humano. El Proyecto Genoma Humano (PGH) fue concebido con el objetivo de determinar la secuencia de nucleótidos de todo el genoma nuclear humano.

La idea del PGH tiene origen en el Departamento de Energía de Estados Unidos (DOE). Desde la intervención posguerra de Estados Unidos en Japón, este organismo financiaba estudios para medir los efectos de las "nuevas tecnologías" en la salud.[8] El nuevo proyecto del DOE expresaba que "conocer la secuencia del genoma humano contribuiría enormemente a caracterizar diferencias entre padres e hijos".

Dos años después de esta primera divulgación el Instituto Nacional de la Salud (National Institute of Health, NIH) demostraba interés en el proyecto, su propósito sería el de conocer el origen de enfermedades. Este organismo proporcionó el impulso necesario para la internacionalización del PGH. Polos científicos de 17 países (luego 35) se sumaron al PGH. Aunar esfuerzos (y capital)

8 Tras el gran experimento llevado a cabo en las ciudades de Hiroshima y Nagasaki, con sujetos humanos, y sin consentimiento informado, se esperaba medir el efecto de la nueva "tecnología nuclear". Conociendo las alteraciones de las radiaciones en el ADN resultaba interesante medir mutaciones que pudieran transmitirse a la descendencia.

llevó a la conformación del Consorcio Internacional para el Secuenciado del Genoma Humano (International Human Genome Sequencing Consortium, IHGSC) del PGH y de la Organización para el Genoma Humano (Human Genome Organization, HUGO). Este último es el único organismo con sede fuera de los Estados Unidos y debería asegurar la colaboración internacional y regular la asignación de nombres a cada gen identificado. También se crearía la Oficina (o departamento) para la Investigación del Genoma Humano (Office for Human Genome Research, OHGR) bajo la dirección del descubridor de la estructura de la doble hélice, James D. Watson.

Los propósitos del PGH han cambiado a lo largo de los 15 años que ha durado. Sus objetivos iniciales fueron determinar la secuencia de bases que componen el ADN humano e identificar todos genes en el mismo, y perseguía varios ideales: comprender la evolución del hombre, el origen de enfermedades y la definición de la condición humana como una interacción entre el medio ambiente y la herencia. En 2003, terminada la secuenciación, los propósitos aparecen ligeramente diferentes: 1) determinar la secuencia de los 3.000 millones de pares de bases que componen el ADN humano; 2) identificar todos los genes presentes en el ADN humano; 3) almacenar la información obtenida en bases de datos públicas; 4) mejorar las herramientas de análisis; 5) transferir las tecnologías relacionadas al sector privado; 6) tratar los aspectos éticos, legales y sociales que puedan surgir del proyecto.

Dos métodos

La secuenciación del ADN es primordial para el conocimiento de un genoma, pero la longitud de la secuencia impone una gran limitación a su abordaje. La secuencia de una molécula muy larga, como el genoma de un organismo superior, debe ser construida a partir del agrupamiento de secuencias cortas. Para esto no existe otro método que buscar superposiciones que indiquen que un fragmento de una secuencia es también parte de otra. El alineamiento de secuencias contiguas en base a estas superposiciones permite inferir la secuencia principal.

Desde el comienzo de la obtención de secuencias genómicas predominarían dos aproximaciones. Una de ellas, lenta y segura, estimaba necesario avanzar progresivamente y sin perder fragmentos. El procedimiento comienza con la fragmentación del ADN y su inserción en un vector de clonado de gran capacidad, para luego analizar los clones por la existencia de segmentos que solapan. Cada clon se secuencia a medida que se observa superposición. En todo momento se controla la localización (física y genética) de cada segmento en el genoma y se procede a la identificación y secuenciado de los clones vecinos. Este método es denominado del clon contiguo (clon contig) o shotgun jerárquico (Figura 13).

El otro procedimiento, estimaba más apropiado romper un genoma en pequeños fragmentos, luego secuenciar todos los fragmentos obtenidos y ensamblarlos para obtener la secuencia total. Este método recibiría el nombre de secuenciado "shotgun" (literalmente, un "escopetazo", de modo que conservaremos el término inglés).

Cobertura del genoma

Si pudiéramos leer sin interrupciones la secuencia de un fragmento de ADN con el mismo tamaño que el genoma probablemente bastara con una lectura (y eventualmente una segunda para confirmarlo y comprobar la ausencia de errores). El mismo procedimiento de secuenciado hace necesaria la fragmentación, pues un equipo de secuenciación no permite leer certeramente más de 400 bases. Asimismo para el secuenciado de cualquier genoma debemos necesariamente recurrir a su fragmentación para obtener piezas de unos pocos miles de pares de bases, más fáciles de manejar.

A partir de los fragmentos generados se reconstruye la secuencia completa del genoma -o porción de interés en el mismo- en base al solapamiento de unos con otros. Los fragmentos deben superponer para asegurarnos que son parte de fragmentos de ADN de mayor orden.

Si los fragmentos obtenidos deben superponerse para inferir su orden en la gran molécula, el número de fragmentos que deberá secuenciarse debe representar una secuencia más grande que el objetivo. Para obtener suficientes secuencias que solapen y reducir al mínimo la frecuencia de errores, tendremos que llegar a un cierto nivel de redundancia, es decir, producir una cantidad de secuencias aleatorias igual a varias veces la longitud de la secuencia de interés (el genoma). En muchos proyectos de secuenciado de un genoma, se ha leído 10 veces más ADN del que se encuentra en el genoma en estudio.

La cobertura proporcionada por un proyecto de secuenciado debe estar relacionada con el tamaño del genoma. Esta cobertura depende del número de lecturas y del tamaño de dichas lecturas. Como regla general, mayor es el tamaño del genoma, mayor debe ser el número de fragmentos examinados. La cobertura (coverage) o profundidad, [9] o la relación entre la longitud de todas las secuencias obtenidas y la longitud del genoma, indica cuántas veces debe secuenciarse la longitud total del genoma para no perder una base de su secuencia. Por ejemplo,

9 Para algunos autores el término *coverage* es indistinguible de *depth*, o profundidad del secuenciado, que indica cuántas veces ha sido leída cada base de la secuencia diana. Este último tiene el mismo significado cuando se analiza el promedio de lecturas de todo un genoma aunque el análisis detallado de regiones puede dar mayor profundidad a algunas y menor (o nula) a otras debido a la aleatoriedad de la selección de regiones a secuenciar.

si se secuencian 50 millones de bases (Mb) de un genoma de 5 Mb, se tiene una cobertura de diez equivalentes del genoma, o 10X.

Una mayor cobertura no indica que la secuencia final estará completa pues aún quedarán espacios que impedirán la unión de grandes fragmentos tras el ensamblado. Sin embargo, cuanto más grande sea la cobertura mayor será la fracción del genoma conocida, los fragmentos que superpongan serán más numerosos y quedará un mínimo de áreas no secuenciadas. No debe creerse que es posible secuenciar tantas veces un genoma como para que no queden espacios sin secuenciar.

Es posible calcular la cantidad y el tamaño de los "espacios" que quedarán en la secuencia final, e indirectamente, estimar en qué punto debe detenerse un intento de secuenciar un genoma fraccionado aleatoriamente. Aunque la idealización mediante funciones matemáticas puede cambiar entre un genoma y otro, este cálculo es muy útil para el secuenciado de genomas de gran tamaño, tal como el del hombre.

La estimación de la cobertura del genoma deriva es una función descripta inicialmente por Lander y Waterman (Lander & Waterman 1988). Si la longitud promedio de las lecturas (n) es muy pequeña comparado con la longitud (L) de la secuencia diana, y el número de lecturas es muy grande, entonces se puede considerar que la probabilidad de que una base de la secuencia diana esté representada en x lecturas sigue una distribución de Poisson, $C^x e^{-C} / x!$, donde C es la cobertura.

El ensamblado de secuencias del genoma humano según dos valores de cobertura diferentes escogidos arbitrariamente puede variar considerablemente (ver Tabla 1). El cálculo muestra que para un genoma de 3.000.000.000 bases se obtendrá una imagen muy diferente según se invierta en la secuenciación de 9 o de 30 mil millones de bases (3X o 10X).

Con un tamaño de lectura idéntico, e igual a 1.000 bases, una cobertura de 3X resultará en la determinación de bases equivalente a aproximadamente un 95% del genoma. El 5% restante estará representado por aproximadamente 500.000 espacios con un tamaño promedio ligeramente superior a 300 bases.

Alternativamente, si se leyera a 10X, la fracción del genoma conocida podría elevarse al 99.99% y los espacios sin conocer se reducirían a aproximadamente 1.362. Estos espacios tendrían un tamaño promedio de 100 pares de bases.

Aunque el sentido común puede determinar que se va a lograr una mayor cobertura repitiendo la fragmentación y el secuenciado tantas veces como sea necesario, el análisis estadístico y la práctica indican que debe preverse la cobertura y se debe partir de una gran cantidad de ADN si se desea tener un mayor porcentaje del genoma secuenciado.

	3X	**10X**
Siendo ***L*** el tamaño del ADN estudiado y ***N*** el número total de nucleótidos a leer.	L=3.10^9 bases N = 9.10^9 bases	L=3.10^9 bases N = 3.10^{10} bases
C = *N* / *L*	C = (9.10^9) / (3.10^9) = **3 X**	C = (3.10^{10}) / (3.10^9) = **10 X**
Para un tamaño *n* de lectura	n = 1000 bases	n = 1000 bases
ADN leído=1-$e^{-N/L}$	1-e^{-3}=0,95 (aprox. 95 %)	1-e^{-10}=0,9999 (aprox. 99,99 %)
Nº espacios sin secuenciar= $(N/n).e^{-N/L}$	=9.10^6 e^{-3} =448.084 espacios	=3.10^7e^{-10} =1.362 espacios
Tamaño promedio de los espacios = L.n/N	=3.10^9x1000/9.10^9 =333 bases	=3.10^9x1000/3.10^{10} =100 bases

Tabla 1. Cobertura de una gran genoma. Cálculo empleado para estimar el resultado del secuenciado de un genoma según se proceda a analizar una extensión equivalente a tres (3X) o diez (10)X) veces el tamaño del genoma.

SECUENCIADO CLON A CLON

Desde el comienzo de la secuenciación del genoma humano se determinó que, para no secuenciar una región repetidas veces y encarecer innecesariamente el costo del secuenciado, sería necesario avanzar de manera ordenada. Para proceder de esta manera, cualquier secuencia obtenida debería poder relacionarse con secuencias contiguas y con un lugar físico en el genoma. De la contigüidad entre las secuencias obtenidas y buscadas deriva el nombre de este método.

Para secuenciar siguiendo este modelo es necesario fragmentar el ADN e insertar los segmentos obtenidos en un vector de clonado. Una vez logrado el banco o biblioteca de clones, se investigaría el solapamiento entre éstos y se construiría un mapa físico del genoma. A medida que se verificaba la existencia de solapamiento, los clones de interés se fragmentarían y secuenciarían por el método shotgun (Figura 14). Este proceso se repetiría hasta que completar la secuencia de la molécula original de ADN. Siendo el genoma humano muy grande, el secuenciado y la construcción del mapa podían proceder al mismo tiempo que la secuenciación.

El secuenciado por el método del clon contiguo tiene la gran ventaja de asegurar la obtención de la secuencia completa de la molécula original de ADN y, en teoría, no requiere repetir el secuenciado de un mismo clon. Por el contrario, asegurar cada paso imposibilita avanzar rápidamente. El tiempo necesario para conseguir la secuencia completa de genoma humano se estimaba posible en unos 15-20 años.

El PGH esperaba inicialmente poner en evidencia genes o regiones que condicionan la existencia de enfermedades. La estrategia consistiría en secuenciar el total de la información genética de un individuo, y luego hallar variaciones en la secuencia del ADN de personas enfermas. A un costo aproximado de 10 dólares por base se secuencia (sin contar salarios y costos operativos) el proyecto

resultaba inabordable. Para comenzar, se estimó sería razonable estudiar unos pocos individuos y comenzar a conocer la localización de sitios variables en ausencia de enfermedad.

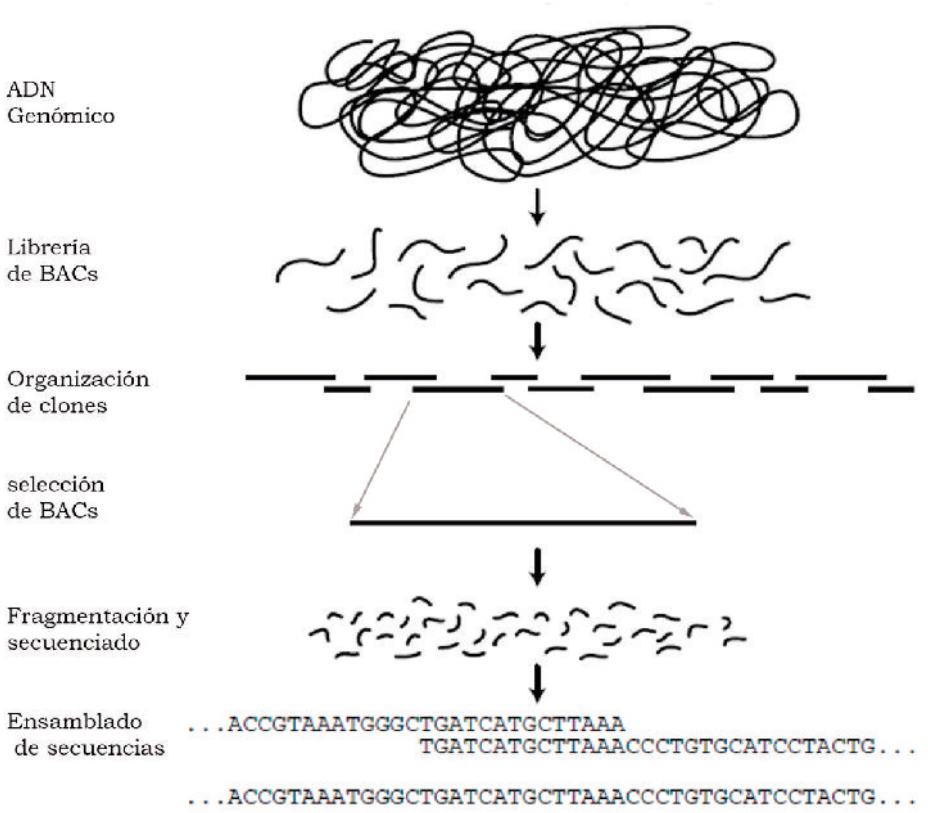


Figura 14. Secuenciado del genoma por el método del clon contiguo. El ADN genómico se fragmenta y clona en vectores de gran capacidad. Estos fragmentos deben entonces ordenarse con ayuda de marcas comunes y referencias con el mapa global. Clones con una localización conocida son seleccionados, fragmentados y secuenciados. El ensamblado de secuencias procede entonces mediante superposiciones.

Grandes vectores

Para comenzar, un genoma debe romperse en fragmentos de tamaño tal que puedan ser manipulados. En este procedimiento los fragmentos se introducen y mantienen en un vector de clonado. Si el vector puede almacenar fragmentos de gran tamaño, un número reducido de ellos permitirá cubrir la totalidad del genoma. Conocido el tamaño de los fragmentos en el vector, el número de clones que deberían analizarse para cubrir el genoma puede estimarse a partir del modelo de Lander y Waterman.

Los vectores más comúnmente utilizados, plásmidos y cósmidos, permiten clonar fragmentos de tamaño inferior a 45 kilobases. Vectores basados en un fago (λ es el más empleado) pueden albergar unos 15 kb de ADN foráneo, y en un cósmido pueden insertarse hasta 40 kb de secuencia. El clonado de fragmentos de ADN para el secuenciado de genomas de organismos superiores no se ha basado sin embargo en el empleo de éstos vectores, sino en otros de mayor capacidad. Los cromosomas artificiales de bacteria y levadura se desarrollaron para este fin y tienen 10-1.000 veces más capacidad que un plásmido, (Burke *et al.* 1987).

El cromosoma artificial de levadura (Yeast artificial Chromosome, o YAC) es, desde el comienzo del PGH, el vector de mayor capacidad conocido. Como su nombre lo indica, está diseñado con secuencias de, y para ser empleado en, una levadura. Un YAC contiene las secuencias mínimas para la replicación de manera estable en la levadura *Saccharomyces cereviciae* (centrómero, telómeros y centros de replicación) y permite la inserción de fragmentos de 100-3.000 kb (Murray & Szostak 1983).

Una biblioteca con insertos del genoma humano basada en YACs, podría cubrirlo de manera aceptable con menos de 20.000 clones. El PGH basó sus primeros esfuerzos en la inserción de fragmentos de ADN humano en YACs. Esta estrategia debió sin embargo ser abandonada cuando se descubrió que algunos clones YAC podían sufrir reordenamientos y contener fragmentos no contiguos del ADN transportado.

En lugar del cromosoma artificial levadura debió recurrirse a un vector que albergara fragmentos de ADN de menor tamaño. El Cromosoma Artificial de Bacteria (Bacterial Artificial Chromosome, o BAC) deriva de un plásmido F de conjugación de *E. coli* (O'Connor *et al.* 1989). Un BAC puede alojar fragmentos 100-200 veces mayores que un plásmido convencional, 150 kilobases en promedio, y no presenta el defecto observado en los YACs. Algunos laboratorios han desarrollado y empleado vectores semejantes basados en secuencias del fago P1, denominado PAC (Phage Artifitial Chromosome) o cromosoma artificial de fago.

En el PGH el ADN humano se cortó en fragmentos de aproximadamente 150-350 kpb y éstos se ligaron en BACs para obtener una biblioteca con aproximadamente 300.000 clones. Con esto se lograría cubrir 2.841.366.484 pares de bases con un mínimo de 26.614 fragmentos. Cada uno de ellos debería ser relacionados con una posición física ("mapeado") en el genoma. Habiendo relacionado un clon con una posición particular en el genoma, cada BAC "listo para secuenciar", sería secuenciado por el método shotgun (Figura 14).

Balizado del genoma

En el laboratorio, no es difícil hallar un fragmento de ADN vecino de otro. Se marca con radioactividad el primer segmento, se lo hibrida con un banco de fragmentos representando el genoma de origen y se recuperan los clones que dan una señal específica. Relacionar un fragmento de ADN con una posición particular en un cromosoma y en el genoma, cuando el inserto es 20.000 veces más pequeño que ese genoma, no es posible sin el empleo de técnicas mucho más laboriosas.

Para relacionar un fragmento con un sitio particular es necesario un mapa. Genes y otras características de interés hallan relaciones entre sí y dentro de un genoma mediante el mapeo, o el posicionamiento de unos respecto de otros.

En el estudio de genes y genomas existen dos tipos de mapas, los mapas físicos y los mapas genéticos. Un mapa físico se obtiene relacionando rasgos o características con secuencias, y en particular, con un lugar en un cromosoma. El mapeo físico depende hoy mayormente del examen de las secuencias de ADN.

Cuando se emplean técnicas genéticas para posicionar una referencia se obtiene un mapa genético. Su obtención depende de experimentos de cruzamiento (como el descrito anteriormente desarrollado por Sturtevant) en organismos con cortos tiempos entre generaciones. En el hombre u otros organismos con largos tiempos generacionales el mapeo genético se realiza en base al examen de la historia familiar (análisis de pedigrees o genealogías).

La dificultad del análisis genético en el hombre ha llevado a la investigación de la separación de fragmentos de ADN mediante un procedimiento artificial. El mapeo se logra tras la separación de marcadores físicos o genéticos lograda mediante la ruptura del ADN.

Mapeo físico

La construcción de un mapa físico tiene por objeto facilitar el ensamblado final del genoma. En la estrategia del clon contiguo, se secuencia cada uno de los fragmentos principales ya ordenados en el genoma entero. Esto reduce la dificultad de montaje de fragmentos con un tamaño máximo de 300.000 pares de bases, en lugar de los 3.000.000.000 que totaliza el genoma humano.

Esta estrategia permite también focalizar el trabajo de terminación, dejando a voluntad un fragmento en el que se está trabajando para completar la secuencia de otro. Con este método resulta asimismo más fácil distribuir el trabajo entre laboratorios colaboradores con un mínimo de coordinación, verificar la validez de una secuencia ensamblada y, en parte, evitar problemas asociados al polimorfismo de secuencias. En la secuenciación shotgun (ver más adelante),

incluso cuando se secuencia ADN de un solo individuo, se ensamblan secuencias que provienen de dos cromosomas diferentes.

El primer paso fue obtener clones de gran tamaño, el siguiente, ordenar estos clones entre sí y a lo largo de los cromosomas humanos. El posicionamiento de un clon respecto de otro implica el uso de diferentes técnicas. Las técnicas más empleadas para hallar áreas comunes entre los diferentes clones son 1) el uso de perfiles de restricción, 2) la hibridación de clones y fragmentos de los mismos, y 3) el uso de STS.

El primer procedimiento implica digerir los distintos clones con enzimas de restricción, elaborar una lista de fragmentos correspondientes obtenidos tras la digestión y luego buscar si diferentes clones presentan fragmentos del mismo tamaño (Figura 15 y Figura 17). Cuando las enzimas de restricción cortan los mismos lugares, y generan fragmentos del mismo tamaño, hay una gran probabilidad de que estos clones correspondan a una región genómica común. Este procedimiento es de gran utilidad cuando la redundancia de clones es grande y debe ser complementado con otras técnicas.

La hibridación de clones y fragmentos de los mismos permite establecer si existen secuencias comunes. El procedimiento comprende la preparación de una sonda de hibridación con un segmento de restricción de un clon o un fragmento amplificado mediante PCR. Luego este fragmento marcado se hibrida con toda una biblioteca genómica para recuperar los clones que pueden hibridar. La verificación de la identidad o continuidad se realiza entonces mediante cartas de restricción y/o secuenciado de cada uno.

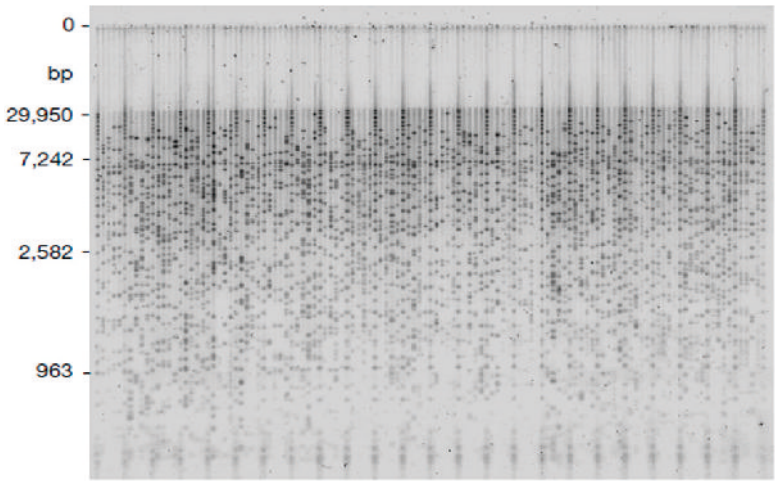


Figura 15. Fingerprinting de alto rendimiento. Cada línea vertical representa una calle del gel de agrosa en el cual se ha sembrado el inserto de un BAC digerido con la enzima HindIII . Cada cinco calles puede verse un marcador que sirve de referencia para estimar el tamaño de los fragmentos. Los números a la izquierda indican el tamaño en pares de bases de bandas representativas.

El paseo cromosómico o la marcha sobre el cromosoma (*cromosome walking*) es un método que consiste en tomar un clon de una biblioteca de ADN creada de manera aleatoria, o con al menos dos enzimas de restricción diferentes, y utilizar el inserto como sonda de hibridación contra todos los otros clones en la biblioteca. Los clones cuyos insertos dan señales positivas de hibridación y se solapan con la sonda, se utilizan a su vez como nuevas sondas de hibridación. Este método permite identificar vecinos próximos y avanzar progresivamente a lo largo de un fragmento de ADN de mayor tamaño, como un cromosoma.

Este método se empleó originalmente para avanzar distancias cortas mediante la superposición de clones preparados con vectores λ o cósmidos y tiene la gran ventaja de avanzar a lo largo de un cromosoma de manera segura (Garber *et al.* 1983). Sin embargo, el tiempo necesario para cubrir largas distancias puede medirse en meses y hasta años. El avance puede asimismo ver interrupciones cuando una sonda contiene una secuencia repetida. La hibridación puede llevar a la superposición de clones pero también a identificación de clones con secuencias repetidas semejantes dispersas en todo el genoma.

La hibridación no específica puede reducirse mediante el bloqueo de las secuencias repetidas por prehibridación con el ADN genómico no marcado pero esto no soluciona totalmente el problema. Cuando el paseo se lleva a cabo con clones de gran capacidad alojados en vectores BAC o YAC la especificidad de las hibridaciones es más problemática debido a la presencia de más secuencias repetidas y a su abundancia en genomas de gran tamaño. En este caso suele emplearse sólo un extremo del inserto como sonda de hibridación o, tras el secuenciado, como sustrato de PCR. El extremo secuenciado se emplea para el diseño de cebadores y éstos se utilizan en reacciones de PCR con todos los otros clones en la biblioteca. Un clon que da un producto de PCR del tamaño correcto debe contener un inserto que superpone. Para acelerar procesos de secuenciado a gran escala, en lugar de realizar una PCR con cada clon individual, se mezclan clones y se reduce la cantidad de reacciones de PCR.

Los clones también pueden ser relacionados entre sí y con una posición física en el genoma mediante secuencias conocidas como STS (Sequence Tagged Sites), sitios de secuencia marcada o de secuencia única. Estas son secuencias cuya ubicación en el genoma es conocida, son relativamente cortas (200 a 500 pb) y pueden amplificarse fácilmente mediante PCR.

La secuencia de ADN de un STS puede contener elementos repetitivos, que aparecen en otras partes del genoma, pero las secuencias en ambos extremos del sitio son únicas y conservadas. En la mayoría de los casos los marcadores STS son co-dominantes, es decir que permiten se distinguir hetorocigotos de homocigotos.

A diferencia de otros marcadores que permiten asignar una secuencia a un sitio particular en el genoma, los STS y el procedimiento para distinguirlos pueden ser descritos detalladamente y cualquier laboratorio puede simplemente

buscar su secuencia, sintetizar los cebadores específicos, y ejecutar la PCR para amplificar un STS desde ADN genómico. Esta PCR produce un patrón simple y reproducible en gel de agarosa o de poliacrilamida.

La IRE-PCR es una PCR en la que se emplean cebadores que se unen a secuencias de elementos repetitivos dispersos en el genoma (IRE, Interspersed Repeated Element). Algunas repeticiones como las de elementos Alu son muy frecuentes en el genoma con una distancia promedio de uno cada 4 kb. La amplificación de fragmentos desde un BAC de 150 Kb podría dar, en teoría, 38 productos de PCR de diversos tamaños, que resultan en una huella digital del clon. Este tipo de PCR se ha empleado excepcionalmente cuando un clon resultaba difícil de identificar.

Los grupos de clones solapantes obtenidos por medio de estos métodos pueden posicionarse a lo largo de los cromosomas gracias a las referencias proporcionadas por mapas de híbridos de radiación o de ligamiento. Cuando todos los clones se han ordenado y posicionado, y se determinan sus distancias reales (en pares de bases) se dispone de un mapa físico del genoma.

A continuación, puede seleccionarse un conjunto mínimo de grandes clones solapantes y emprender el secuenciado. En la práctica, la elección de los clones para ser secuenciados puede lograrse a medida que progresa la secuenciación, permitiendo en cierta medida minimizar las regiones de solapamiento entre los clones más importantes (Figura 16).

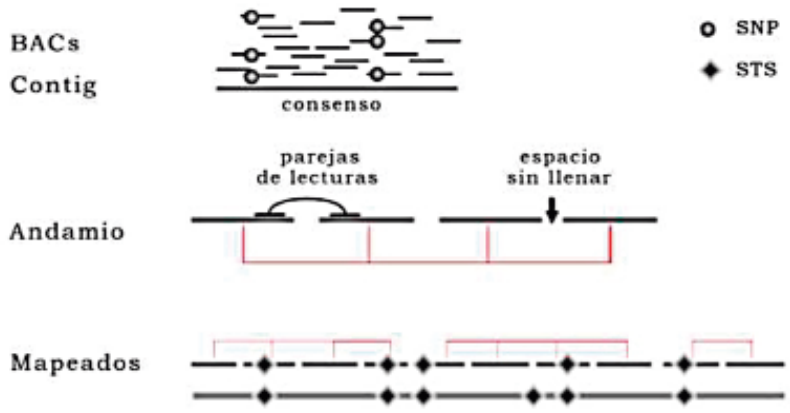


Figura 16. Un mapa de STSs permite reducir el número de fragmentos secuenciados. Consensos de secuencias logradas a partir de BACs u otros vectores, pueden posicionarse en un mapa global logrado gracias a la identificación de marcadores STS. Espacios libres entre secuencias posicionadas podrán llenarse subsecuentemente desde sus extremos dado que, con cierta seguridad, ya se conocerán las secuencias inmediatas adyacentes.

Mapeo genético

Un mapa de ligamiento permite disponer de marcadores, o puntos de referencia (mojones o balizas), ordenados a lo largo de los cromosomas, mediante la medición de su enlace. En 1922 se podían reconocer más de 50 genes en *D. melanogaster*. Todos ellos correpondían a rasgos visibles como el color de los ojos (nueve genes diferentes), forma o color del cuerpo, presencia y número de pelos o características de las alas, y podrían pronto ser relacionados con una posición en los cromosomas del insecto.

Los marcadores bioquímicos han sido ampliamente reconocidos y utilizados con microorganismos (básicamente su capacidad de crecer en ausencia de ciertos componentes). Algunos genes responsables de características trazables entre generaciones se han hallado y empleado asimismo en el hombre y otros mamíferos: los grupos sanguíneos, variantes de proteínas de suero y de las proteínas inmunológicas tales como los antígenos leucocitarios humanos (HLA).

Aunque se conocen muchos genes que pueden ser empleados como marcadores, no es posible trazar un mapa con suficiente detalle para posicionar cientos de contigs en el genoma humano. El tamaño de los genes es asimismo muy variable y con frecuencia son muy grandes o están muy esparcidos. Finalmente, sólo una fracción del número total de genes posee distintas formas alélicas que pueden distinguirse convenientemente.

Pocos años antes de iniciado el PGH se descubrían marcadores que podrían ser empleados en el trazado de mapas genéticos en organismos superiores. Los marcadores de ADN son secuencias polimórficas, es decir que tienen dos o más formas alternativas, o no polimórficas. Estas secuencias no guardan relación con la presencia de genes y se hallan distribuídas de manera más o menos regular en un genoma. El estudio de su segregación permite la elaboración de mapas genéticos.

El mapeo genético humano fue posible tras el descubrimiento de los polimorfismos de longitud de fragmentos de restricción. El análisis de RFLPs se basa en alteraciones en los sitios de corte de enzimas de restricción (Figura 17). Así como un gen puede tener dos o más formas, la presencia o ausencia de un sitio de restricción puede ser vista como dos formas alternativas del mismo sitio. La presencia o ausencia de un sitio de corte da la posibilidad de observar diferentes longitudes en los fragmentos generados por cada enzima. Cada organismo o individuo tiene así un perfil de restricción característico que lo distingue.

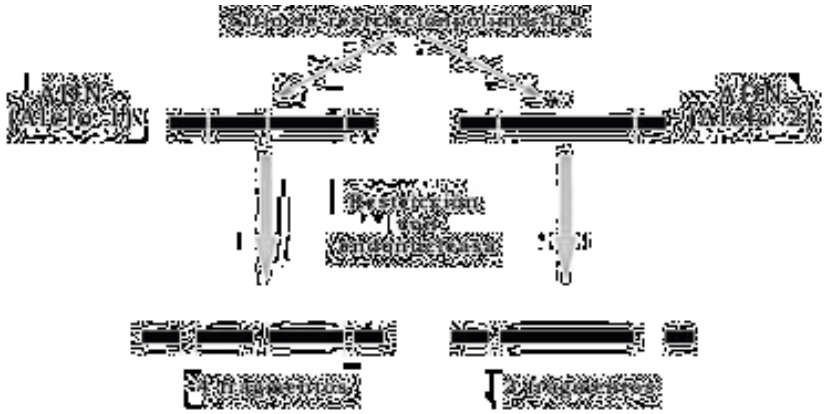


Figura 17. Principio del análisis de polimorfismos de restricción. La existencia o no de un sitio de corte para una endonucleasa de restricción determina la obtención de un número diferente de fragmentos de ADN. La ausencia de un sitio determina asimismo la existencia de un fragmento cuyo tamaño equivale a la suma del tamaño de dos fragmentos menores en el ADN que contiene el sitio de corte.

El valor de los RFLP en el mapa genético humano está limitado a la presencia/ausencia de un sitio de restricción. Para hallar regiones variables es necesario analizar numerosas familias. El primer mapa de ligamiento humano incluyó 393 RFLPs y 403 sitios polimórficos deducidos a partir del estudio de tres generaciones entre 21 familias (Donis-Keller *et al.* 1987). Este mapa, tenía una densidad media de un marcador por cada 10 Mb. Aunque este avance era significativo, el valor no alcanzaría para relacionar cada clon con una posición física.

El consorcio para el estudio del genoma humano se fijó entonces el objetivo de aumentar la densidad de marcadores en el mapa genético a uno por cada 1 Mb (aproximadamente 4000 marcadores esparcidos en el genoma).

Esto sería posible gracias a la inclusión de SSLPs (Single Sequence Lenght Polymorphism) y SNPs (Single Nucleotide Polimorphism). Los SSLPs son secuencias repetidas de longitud variable. A diferencia de los RFLPs los SSLPs contienen diferentes números de unidades de repetición y por ello pueden ser multialélicos.

Los SSLPs más conocidos y empleados guardan dos formas características conocidas como minisatélites o microsatélites. Los minisatélites, o repeticiones en tándem de número variable (VNTR, Variable Number Tamdem Repeats), son secuencias en las que la unidad de repetición alcanza hasta 25 pb. Los microsatélites, o repeticiones cortas en tándem (STRs), poseen repeticiones más cortas de 2 a 4 nucleótidos (Figura 18). Los primeros pueden alcanzar kilobases de longitud en tanto que los segundos raramente alcanzan los 300 pares de bases.

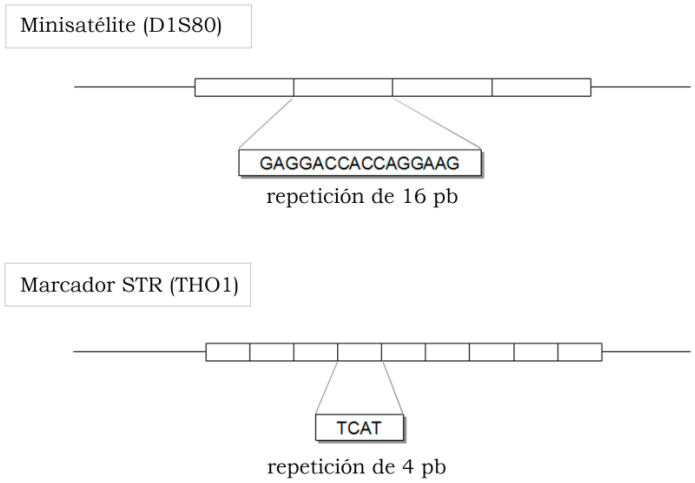


Figura 18. Repeticiones cortas. Ejemplos de dos repeticiones cortas, un minisatélite, con repeticiones de 16 pares de bases, y un microsatélite, de tan sólo cuatro pares de bases en cada unidad.

Los microsatélites han alcanzado más popularidad que los minisatélites como marcadores de ADN porque pueden ser observados rápidamente mediante PCR y porque se distribuyen uniformemente en todo el genoma. El análisis de filiación en la actualidad se basa en el empleo de este tipo de marcadores (Figura 19).

En 1994 el objetivo del consorcio de una mayor densidad de marcadores se había superado. Con el sostén del Centro de Estudios de Polimorfismos Humanos (CEPH, Centre d'Etude du Polymorphisme Humain) el mapa del genoma contaba con 5800 marcadores SSLPs y la densidad alcanzada era de un marcador cada 0.7 Mb (Murray *et al.* 1994). Este mapa permitía localizar muchos genes asociados con enfermedades genéticas en el hombre (como el gen que codifica la distrofina, responsable de la distrofia muscular de Duchenne, y cftr, cuyas mutaciones dan origen a la fibrosis quística). Dos años después el mapa del genoma humano incluía 1250 SSLPs adicionales. Esta cifra se apenas acerca al 10% de los microsatélites existentes en el genoma humano (aproximadamente $6{,}5 \times 10^5$).

El otro tipo de marcador polimórfico empleado para medir la variación entre individuos, que ha resultado muy útil en el trazado de mapas, se conoce como SNP, o polimorfismo de nucleótido único. Cuando se comparan secuencias de ADN de varios individuos se hallan sitios en los cuales algunos tienen una base mientras que otros poseen una distinta. Curiosamente, pudiendo presentarse hasta cuatro variantes (una por cada base), la mayoría de los SNPs hallados en el genoma humano existe bajo una de dos formas. Este hecho representa una limitación en el mapeo genético humano pues existe una alta posibilidad de que un SNP no muestre variación entre muchos individuos de una misma familia.

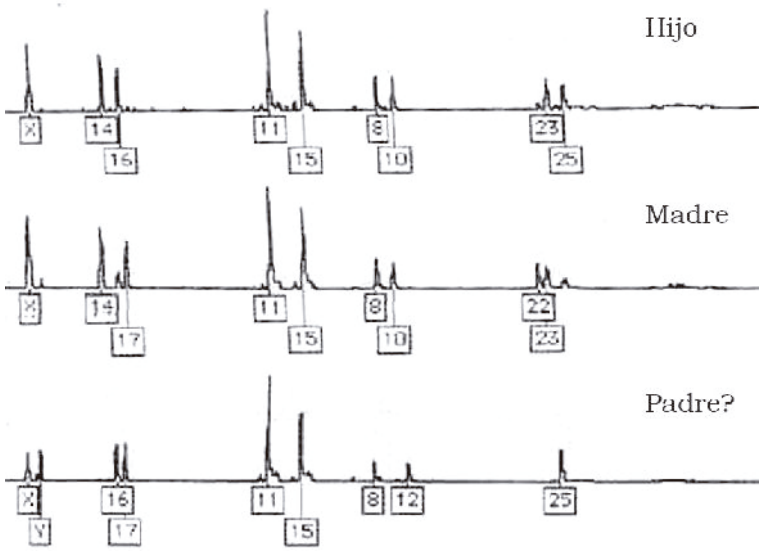


Figura 19. Marcadores STR empleados en la identificación de personas y vínculos. Cinco marcadores empleados en la identificación de personas. El primero permite la identificación inequívoca del sexo de la persona, un pico a la izquierda indica la presencia de un cromosoma X y uno más a la derecha el de un cromosoma Y. De arriba hacia abajo, las dos primeras personas son de sexo femenino y la tercera de sexo masculino. Los cuatro marcadores que siguen de izquierda a derecha alojan en autosomas y son variables entre personas. Estos permiten confirmar relaciones biológicas entre individuos. En el hijo, arriba, se hallan dos marcadores diferentes, de 14 y 16 repeticiones. Si se busca cada uno de éstos entre las dos personas debajo, una es la madre y la restante el posible padre, se encuentra el alelo de 14 repeticiones en el perfil materno y el de 16 en el probable padre. En los tres marcadores que siguen ocurre algo semejante.

La ventaja del uso de estos marcadores consiste en que se presentan en número muy elevado en el genoma y pueden identificarse rápidamente mediante métodos automatizables, como la hibridación de oligonucleótidos en soporte sólido o en solución.

Los marcadores moleculares empleados en los mapas de ligamiento son una herramienta indispensable en la validación y relación de un mapa físico con un sitio particular de un cromosoma.

Sin embargo las comparaciones entre los mapas genéticos y las posiciones reales de los genes en las moléculas de ADN han demostrado que existe una gran diferencia entre uno y otro. Algunas regiones de los cromosomas son más propensas a la existencia de entrecruzamiento que otras y así la distancia en el mapa genético no se relaciona exactamente con la distancia física entre dos marcadores.

Secuencias expresadas

Los genes eucariotas son demasiado largos y variables para ser empleados como los marcadores citados hasta el momento. Sin embargo, la secuencia de cada uno de ellos tendría gran utilidad en el ensamblado del genoma humano.

Recuperar o identificar la secuencia de un gen puede ser un procedimiento largo que involucra en general muchas técnicas. Sin embargo, es posible recuperar sin demasiado esfuerzo el resultado de la expresión de un gen, en particular el ARN que dará origen a una proteína. El procedimiento consiste en aislar el ARN de la célula, especialmente el ARN mensajero (libre de ARN ribosomal y ARN de transferencia), luego este ARN se retrotranscribe para obtener un ADN complementario (ADNc). Este procedimiento es posible gracias al empleo de una ADN polimerasa dependiente de ARN hallada en retrovirus. Las transcriptasas reversas más empleadas provienen del virus de la leucemia de Moloney y el virus de la mieloblastosis aviar (Spiegelman *et al.* 1971; Buell *et al.* 1978).

El híbrido ARN-ADNc se separa entonces, simplemente calentando la solución, y se agregan entonces cebadores específicos –si se desea recuperar una secuencia particular- o cebadores muy cortos que hibridarán de manera aleatoria en todos los ADNc y en varias regiones de cada uno. Tras éstos se añade una polimerasa de ADN que permitirá reconstruir una doble hebra de ADN. Este podrá entonces servir para investigar la presencia de un transcripto particular o a la creación de una librería de ADNc.

Tras el aumento del rendimiento del secuenciado a comienzos de los 90 pudieron hacerse librerías de expresión, o librerías que representan todos los ARNm de una célula, tejido u organismo. Estas secuencias de ADNc son conocidas como ESTs (expressed sequence tags, traducidos comúnmente como sitios de secuencia rotulada o sitios de secuencia etiquetada, haciendo referencia al sitio en que aloja cada gen como marcador físico). Los ESTs tienen gran uso en el descubrimiento de genes, en el mapeo y la identificación de regiones codificantes en el genoma (Adams *et al.* 1991; McCombie *et al.* 1992; Martin-Gallardo *et al.* 1992). En el mapeo en particular, los ESTs serían esenciales para anotar y validar los genes estimados mediante análisis informático. Además de contribuir a la predicción de genes transcriptos los ESTs contribuyen directamente a empalmar contigs y cubrir espacios de secuencia gracias al gran espacio existente entre exones.

A pesar de sus muchos usos, el número de ESTs humanos en las bases de datos públicas sumarían menos de 40.000 a fines de 1994 y dos años después este número aumentaba sin mucha certitud a 44.000 (Hillier *et al.* 1996). Este valor ya daba una idea que la complejidad del hombre como especie no es reflejo de un gran número de genes.

Análisis genealógico

En el hombre no es posible cartografiar genes mediante la selección de los genotipos de los padres y el diseño de cruzamientos. Las frecuencias de recombinación y la distancia entre genes y marcadores deben ser determinadas a partir

del examen de genotipos de miembros de distintas generaciones de una o más familias en las cuales se halla un rasgo particular. Los datos disponibles son así limitados y la interpretación es a menudo difícil.

El material de análisis consiste en general de familias con un gran número de hijos, y en las cuales pueden analizarse los miembros de al menos tres generaciones. El Centre d'Études du polymorphisme Humain (CEPH) en París alberga colecciones de familias que reúnen estas condiciones, incluyen cuatro abuelos y al menos ocho niños de la segunda generación, permitiendo estudios genealógicos (Grausz 1993). La colección está disponible en forma de líneas celulares que pueden ser solicitadas por cualquier investigador para el estudio de familias y asignación de marcadores de ADN con el único compromiso de presentar los resultados obtenidos a la base de datos del CEPH.

El análisis genealógico tiene como propósito principal el estudio de enfermedades genéticas. El marcador genético que desea localizarse es un gen en el cual un alelo determina la enfermedad y otro el estado de salud. El estudio genealógico se realiza como se ejemplifica en la Figura 20 para una enfermedad genética transmitida desde la abuela materna y presente en cuatro de seis niños. Del estudio de ligamiento con un microsatélite A, para el cual se ha demostrado la presencia de cinco alelos (A1, A2, A3, A4 y A5) se deduce que sólo un niño es recombinante y que el alelo de la enfermedad está en el mismo cromosoma que A1.

En la especie humana, para detectar ligamiento y estimar la frecuencia de recombinación más probable entre genes se emplea un estimador conocido como LOD score, o puntuación LOD (Morton 1955). Un valor Z se calcula como el logaritmo de la razón entre la probabilidad de la descendencia dada la existencia de ligamiento con una frecuencia de recombinación R respecto del caso sin ligamiento. Una ventaja de este método es que es pueden compararse e incluso sumarse valores obtenidos para diferentes familias (asumiendo un mismo modelo genético) aumentando la confianza en el resultado. Por el contrario, el análisis LOD requiere un modelo genético preciso respecto al modo de herencia, penetrancia y frecuencias génicas, por lo cual es difícilmente aplicable a enfermedades complejas.

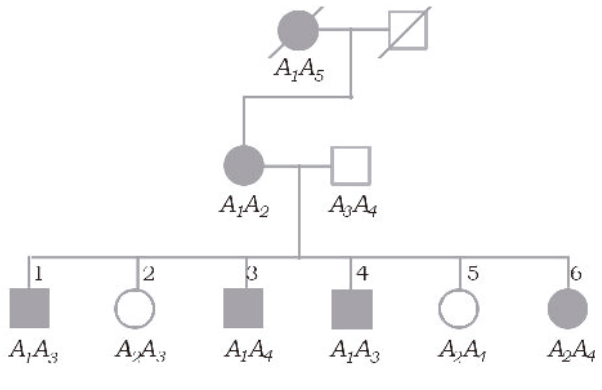


Figura 20. Ejemplo de análisis de pedigree en humanos. El árbol genealógico muestra la herencia de una enfermedad genética en una familia de dos padres vivos y seis niños, con información acerca de los abuelos maternos. La enfermedad viene acompañada del marcador A1 en las madres y los niños 1, 3 y 4. Los niños 2 y 5 tienen el alelo sano acompañado del marcador A2, proveniente de la madre y ausente en la abuela. La niña 6 tiene el alelo de la enfermedad y el alelo microsatélite A2. 1, 2, 3, 4 y 5 corresponden a genotipos parentales y 6 es la única recombinante. El gen de la enfermedad y el microsatélite están ligados.

Híbridos de Radiación

El mapeo mediante híbridos de radiación es un enfoque genético llevado a cabo con células somáticas y presenta grandes ventajas para la construcción de mapas de contigs de alta resolución y de largo alcance del genoma humano

En una primera versión del procedimiento, se hicieron híbridos de células somáticas de hámster que contienen todos los cromosomas de esta especie junto a un único cromosoma humano. Estas células se irradian letalmente con el propósito de romper cada cromosoma en fragmentos cuyo tamaño depende de la dosis de radiación. Tras la ruptura del ADN, fragmentos del cromosoma humano se fusionarán espontáneamente en distintos sitios de los cromosomas de hámster. La célula irradiada se rescata entonces mediante fusión con una célula de hámster no irradiado en condiciones en que sólo híbridos de células somáticas irradiadas y no irradiadas pueden formar colonias viables. En estos "híbridos de radiación", pueden hallarse fragmentos de cromosomas humanos dispersos en los cromosomas de hámster, que segregan de manera estable (Benham *et al.* 1989).

ADN aislado a partir de clones híbridos de radiación independientes proporciona luego el sustrato para ordenar y determinar las distancias que separan distintos marcadores en el genoma humano. La frecuencia de ruptura entre dos marcadores inducida por la radiación se utiliza como una medida de la distancia y el orden de los marcadores se determina de manera análoga al análisis de ligamiento (Cox *et al.* 1990). Al igual que en la cartografía meiótica, la confianza en el orden de los marcadores puede evaluarse mediante métodos estadísticos.

El mapeo mediante híbridos de radiación puede, a diferencia de la cartografía genética, integrar marcadores polimórficos y no polimórficos. Una ventaja adicional de esta cartografía es que los paneles para la construcción de mapas pueden ser generados en muy diferentes niveles de resolución mediante la manipulación experimental de la dosis de radiación.

El mapeo físico del genoma humano ha avanzado progresivamente mediante la ubicación de marcadores STS (Hudson *et al.* 1995) genes (Deloukas *et al.* 1998; Schuler *et al.* 1996) y hasta su integración posicionando cada BAC a ser secuenciado (McPherson *et al.* 2001). La densidad de marcadores STS en ese entonces estaba próxima a uno por cada 100 kb fijados como objetivo para el mapeo físico en el inicio del Proyecto Genoma Humano. Lejos de detenerse con el inicio del secuenciado el mapeo del genoma ha continuado y ha incrementado la densidad de marcadores con el propósito de reconocer la diversidad entre genomas de distintos individuos (Abecasis *et al.* 2012).

El secuenciado final

El consorcio para el PGH tenía una idea clara desde el comienzo, el secuenciado procedería una vez asignada una posición a cada clon. Fragmentos de gran tamaño se dividirían en otros de menor tamaño (de aproximadamente 2 kb) cuyos extremos se secuenciarían individualmente. Estas secuencias se ensamblarían entonces hasta proporcionar una secuencia lo más completa posible del de mayor tamaño. El montaje de secuencias de contigs debía así, paso a paso, dar la secuencia del genoma humano completo.

En un comienzo el secuenciado no representaba la mayor parte del trabajo y se llevaba a cabo manualmente hasta la introducción de un instrumento creado y vendido por la compañía Applied Biosystems. En 1987 Applied lanzaría al mercado el primer secuenciador automático de ADN, el ABI 310, que permite la lectura de fragmentos que transitan por un gel de poliacrilamida con un detector láser en el extremo. Con él serían posibles lecturas de 400 pares de bases y casi 2000 bases de secuencia por hora. El secuenciado de dos genes de receptores cardíacos de rata aseguraba el primer éxito del instrumento (Gocayne *et al.* 1987).

Aunque la capacidad de secuenciado de algunos centros asociados con el PGH es realmente grande, su mayor mérito fue asociar cada fragmento de ADN con una localización física en el genoma. El consorcio internacional concluyó a principios del año 2000, un mapa físico del genoma humano que cubre el 97% del mismo (McPherson *et al.* 2001). La secuenciación, a una cobertura de 7,5X, había permitido montar un esbozo de secuencia que cubría el 87% del genoma. El 28% de esa secuencia estaba entonces en la fase de "terminado".

Con el paso de 7,5X a 10X y el acabado de huecos y regiones de baja calidad, el consorcio comunicaría en abril de 2003 una secuencia completa del genoma humano. Esta secuencia cubre el 99% del genoma humano con una precisión de 99,99%.

Parte de la inversión y todas las ideas derivadas del PGH terminan siendo públicas. Parte de estas ideas y algo del capital terminaron sin embargo sirviendo al desarrollo de empresas con grandes ambiciones. Una de ellas emprendería el proyecto alternativo.

SECUENCIADO ALEATORIO

El secuenciado aleatorio tiene su origen, como otros métodos ya descriptos, en el laboratorio de F. Sanger (Sanger *et al.* 1982) (Sanger *et al.* 1980). En su forma original el procedimiento consistió en fragmentar el genoma, o una región de interés, por medio de cortes parciales con una enzima de restricción con diana en un sitio frecuente, o por medio de sonicación. Los fragmentos se clonaban entonces en vectores apropiados y se secuenciaban tantos fragmentos como

fuera posible. La superposición de las secuencias obtenidas permitía entonces el ensamblado de una de mayor orden.

Requisito indispensable para el secuenciado con el método de Sanger es el conocimiento de una pequeña porción de secuencia para poder diseñar y sintetizar primers. Sin embargo, un segmento de ADN desconocido puede introducirse en un vector y secuenciarse a partir de los extremos conocidos de éste último (ya se describió el empleo del fago M13 con este propósito). Este método tiene la gran ventaja de permitir el secuenciado de cualquier fragmento de ADN que pueda ser clonado en un vector con las ventajas que provee el mismo clonado (conservar el fragmento, disponer de grandes cantidades del mismo y eventualmente, estudiar su expresión).

Con el método shotgun no es necesario trazar un mapa genético o físico (Figura 21). En principio, secuenciar mucho permitiría alinear todos los fragmentos de un genoma pequeño. Aunque el costo podría superar al del secuenciado progresivo de fragmentos vecinos, el tiempo necesario para su consecución se demostraba significativamente menor. La mayor desventaja del procedimiento se halló en la ausencia de ciertas regiones del genoma en el ensamblado final.

Sanger y su grupo descifraron mediante este procedimiento las 2771 pares de bases del genoma mitocondrial humano (Sanger *et al.* 1980) y poco tiempo después las 48.502 bases que componen el genoma del bacterófago λ (Sanger *et al.* 1982).

Durante la década de 1990 hubo un extenso debate respecto del secuenciado de un genoma de mayor tamaño con el enfoque shotgun. La mayor limitante era entonces la capacidad de los sistemas informáticos existentes para alinear los fragmentos obtenidos. Estas dudas se resolvieron recién en 1995 cuando se publicó la secuencia del genoma de la bacteria *Haemophilus influenzae* (Fleischmann *et al.* 1995).

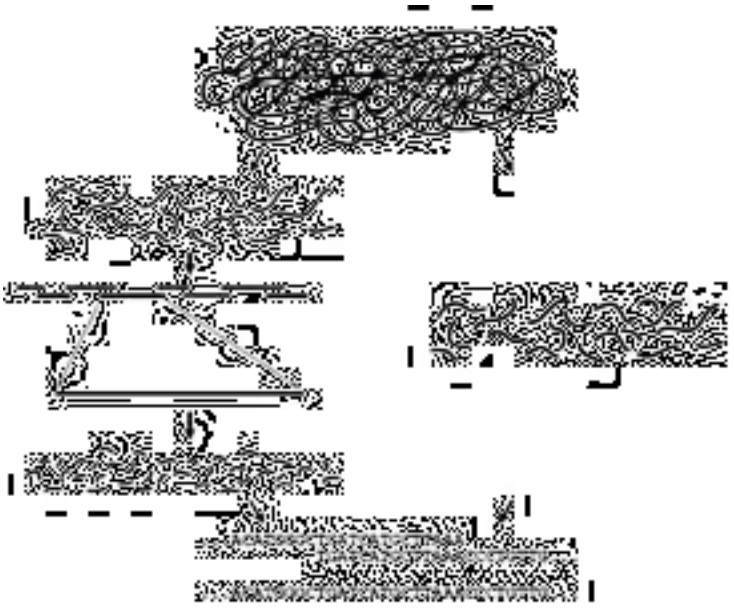


Figura 21.
Procedimiento general del secuenciado shotgun. El nombre de este procedimiento define la intención de romper el ADN y secuenciarlo directamente (derecha), ahorrando pasos tendientes a organizar la información (izquierda).

El grupo privado dirigido por C. Venter en el Instituto para la Investigación Genómica (The Institute for Genomic Research, MD, USA) empleó el método shotgun para conocer los 1.830 kpb del genoma de *H. influenzae*. Para conocer la secuencia completa se fragmentó el ADN de manera aleatora mediante ultrasonido. Los fragmentos se separaron por electroforesis en gel de agarosa, se seleccionaron aquellos con un tamaño de 1,6-2,0 kb y se ligaron en un vector plasmídico. De la biblioteca resultante se aislaron 19.687 clones y obtuvieron 24.304 secuencias de tamaño mayor a 400 bases. Para la cobertura completa se estimaba necesario disponer de un mínimo de 5.3 veces el genoma de la bacteria y las secuencias obtenidas sumaron seis veces la longitud del genoma (11.631.485 bases). El ensamblado de estas secuencias insumiría 30 horas de procesamiento en una computadora con 512 Mb de RAM, y daría lugar a 140 secuencias largas (contigs) no alineadas. El método es muy rápido pero se obtinene fragmentos del genoma de la bacteria.

Para unir los fragmentos resultantes debió recurrirse a otras estrategias. Mediante hibridación se buscaron vectores que pudieran contener las secuencias terminales de dos contigs diferentes, del mismo modo que en el procedimiento del clon contig, y se hallaron 99 clones que cerrarían un número semejante de lagunas del secuenciado.

Suponiendo que los fragmentos de ADN necesarios para llenar los 42 espacios aún vacíos no estarían presentes en la biblioteca y representarían secuencias que resultan inestables en el vector plasmídico, se clonaron fragmentos de 15-20 kb en un vector derivado del fago λ. Esta nueva biblioteca se hibridó con 84 oligonucleótidos que representaban secuencias idénticas a los extremos de los contigs aislados. Así pudieron cerrarse 23 brechas adicionales. Para completar los espacios aún presentes debieron emplearse otras técnicas.

Conocer la secuencia de este primer genoma bacteriano demandó casi tres años de trabajo, sin contar el análisis. Un genoma pequeño podía entonces ser secuenciado de forma relativamente rápida mediante el sistema shotgun. Esto se vería reflejado en la puesta en marcha de numerosos proyectos genoma para otros microorganismos en ese mismo instituto y en otros en el mundo. La organización y estandarización del proceso resultó, por ejemplo, en la secuenciación del genoma de Mycoplasma genitalium, de 580.070 pb, en tan sólo ocho semanas (Fraser *et al.* 1995) 070 base pairs.

El secuenciado aleatorio global resulta hoy el método más abordable, rápido y económico, cuando el objetivo es el secuenciado de un genoma pequeño, tal como el de una bacteria. Sin embargo con genomas de mayor tamaño y más complejos la estrategia presenta grandes limitaciones. Tras el ensamblado de secuencias de *H. influenzae* se estimaba que quedaría un 0.67% del genoma sin cubrir, distribuido en unos 130 espacios con una longitud promedio de 100 pares de bases. Estos valores se vieron reflejados con gran exactitud en el resultado. Para obtener una

cobertura semejante del genoma humano sería necesario disponer, de algo más de 15 millones de secuencias, y esto dejaría aún muchos baches.

Otro método para un mismo propósito

En 1998, la corporación que aportaba la mayor cantidad de insumos y sistemas de análisis al PGH, PerkinElmer, crea su área de biosistemas (PE Biosystems), agrupando empresas destinadas a la elaboración de secuenciadores, termocicladores, reactivos de PCR, de secuenciación y programas informáticos, entre otros. A esta gran reorganización sigue el anuncio que el grupo estaba en condiciones de descifrar el genoma humano. Más aún, PE Biosystems anunciaba que podrían decodificar todo el ADN humano para el año 2001, cuatro años antes que lo previsto por el consorcio público, y por un costo diez veces inferior (sólo 200 millones de U$S). La declaración tuvo gran impacto en el trabajo del consorcio académico que acortaba el plazo esperado para completar la lectura hasta 2003.

El grupo PE creaba en pocos meses Celera Genomics, una empresa destinada exclusivamente a la secuenciación genómica, bajo la dirección de C. Venter. Este proyecto genoma privado comenzaría inmediatamente.

El ADN para el estudio del genoma humano provino de cinco individuos, dos hombres y tres mujeres de distintos grupos raciales. Tras clivar y fraccionar el ADN por tamaño se obtuvieron bibliotecas de fragmentos de aproximadamente 2 Kpb, 10 Kpb y 50 Kpb. El empleo de vectores de distinta capacidad y naturaleza tiene como propósito mejorar la cobertura global del genoma. Tres razones justifican este diseño: 1) minimizar problemas de incompatibilidad de algunos fragmentos de ADN con el vector o el huésped, 2) evitar la pérdida de fragmentos, y 3) saltar grandes regiones de repeticiones mediante el empleo de vectores de gran tamaño.

El mismo grupo trató de llegar con el secuenciado aleatorio al genoma completo de *Drosophila* y en éste se vió que la mayoría de secuencias repetidas tienen una longitud de 8 kb o menos. Empleando vectores con fragmentos de 10 Kb se podría "saltar"[10] sobre grandes secuencias repetidas (Figura 22).

Cada vector se leería entonces desde ambos extremos y las secuencias se controlarían para excluir aquellas muy cortas o conteniendo fragmentos de ADN bacteriano o del vector. Con la incorporación de robots en casi toda la manipulación y de 75 secuenciadores del nuevo modelo ABI PRISM 3700, íntegramente automatizados, este grupo podría obtener 175.000 lecturas de 500-750 pb por día. En nueve meses se obtenían 27.271.853 de lecturas totalizando una cobertura de 5.11 X.

10 Este procedimiento recibe el nombre de "chromosome jumping"(jump= saltar), siguiendo el modelo del anteriormente citado "chromosome walking".

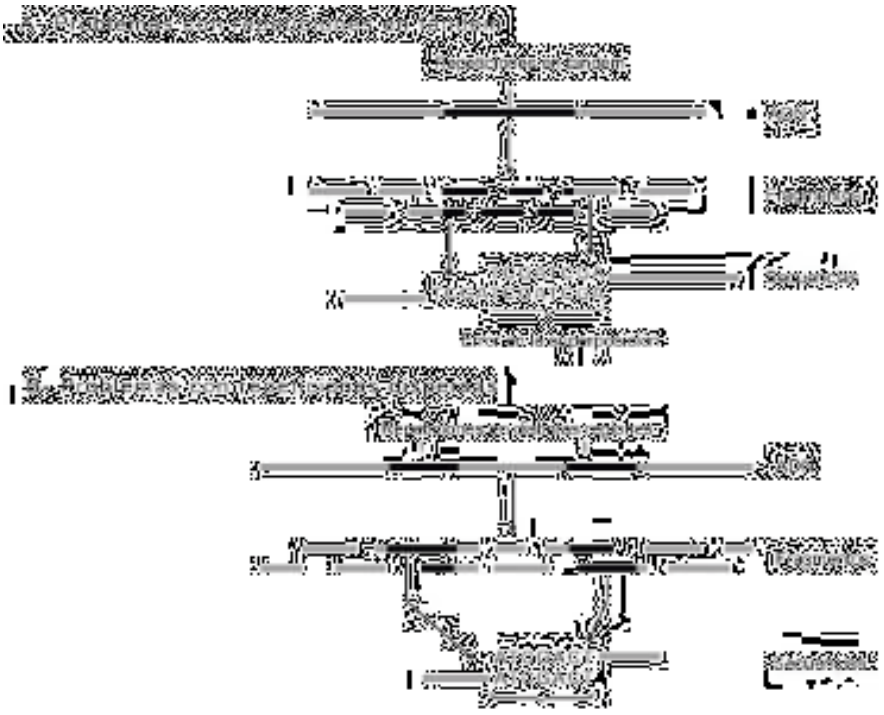


Figura 22. Secuencias repetidas en el genoma. Las secuencias repetidas pueden hallarse en grupos alojados en un solo sitio (A) o dispersas en el genoma (B). Tanto unas como otras representan un gran problema al momento de ensamblar fragmentos. La longitud de las primeras pueden estimarse erróneamente por la superposición incorrecta y las segundas pueden llevar al ensamblado de secuencias distantes, incluso de distintos cromosomas, así como a la pérdida de fragmentos.

Las secuencias obtenidas por Celera dejarían por supuesto grandes lagunas. Para reducir algunos de estos espacios el grupo tomó las secuencias obtenidas hasta entonces por el consorcio público (de bases de datos accesibles para todo el mundo), cada secuencia se fraccionó en segmentos de 550 pb y se mezclaron con sus propios datos para agregar 2.9 X de cobertura adicional.

La experiencia con el secuenciado aleatorio demostraría que una cobertura próxima a 8 veces la longitud del genoma dejaría leer más de 98 % del genoma. El trabajo de Celera (con secuencias públicas) produjo 2.586 Gbp de secuencia, es decir menos del 90 % del genoma humano. Esta secuencia está fragmentada por 93.857 espacios sin secuencia o sin secuencia segura. La falta de cobertura en ciertas regiones presenta grandes problemas cuando se trata de examinar genes. Los estudios del genoma humano se basan casi íntegramente en los resultados del consorcio público pero los estudios de la secuencia del genoma de Drosophila realizado por Celera han sugerido que hasta el 6500 de los 13 600 genes podrían contener errores significativos de la secuencia (Karlin *et al.* 2001).

El hecho más remarcable del proyecto de Celera no fue quizás la enorme inversión en robótica y la capacidad de organización y trabajo de un solo

laboratorio pues esta inversión retornaría con creces tras atraer clientes. El gran desarrollo se basó en los sistemas de análisis. Cinco programas ordenados en línea se diseñaron para trabajar y ordenar las secuencias.

FIN DE LA CARRERA

La primera secuencia de un cromosoma humano completo (el número 22) se publicó en diciembre de 1999 (Dunham *et al.* 1999) y la secuencia del cromosoma 21 apareció unos meses más tarde (Hattori *et al.* 2000). El 26 de junio del año 2000, Francis Collins (director del PGH) y Craig Venter (director de Celera Genomics), anunciaban conjuntamente la inminente finalización del proyecto.

En el mes de enero del año 2001 la revista Science daba a conocer la secuencia del genoma humano obtenida por el grupo privado y en febrero del mismo año la revista Nature publicaba el primer borrador del genoma humano cuyo descifrado completo se estimaba posible para fines del año 2003 (McPherson *et al.* 2001; Venter *et al.* 2001; Lander *et al.* 2001).

El secuenciado del genoma humano quedará, quizás, en la memoria colectiva como uno de los más grandes hitos en la historia de las ciencias. Para muchos tendrá otro significado. Celera usó los mapas y las secuencias obtenidos por el consorcio público. Sin embargo rechazó la propuesta de hacer públicos sus datos a través del GenBank o mediante acceso a sus propias bases de datos. Peor aún, Celera habría patentado muchas de las secuencias obtenidas (ver más adelante).

Las secuencias del genoma anunciadas en 2001 son borradores, no secuencias finales completas. La versión obtenida por el método clon a clon, más completa que la de Celera, cubre apenas el 90% del genoma. Las aproximadamente 320 Gpb que restan secuenciar se consisten mayormente en duplicaciones muy difíciles de ensamblar y heterocromatina constitutiva, es decir regiones de los cromosomas en las que el ADN está muy empaquetado y se cree contienen pocos o ningún gen. En el porcentaje cubierto del genoma, cada parte ha sido secuenciada por lo menos cuatro veces, proporcionando un nivel "aceptable" de precisión. Sin embargo sólo el 25% se había secuenciado las 8-10 veces necesaria para considerar el trabajo "terminado". La secuencia obtenida consta de aproximadamente 50.000 andamios con un tamaño promedio de 54,2 kb.

EL RESULTADO

El PGH es probablemente la mayor empresa que se haya visto hasta el momento agrupando múltiples laboratorios colaborando con un solo propósito. Ciertamente el mayor porcentaje de los laboratorios intervinientes están en los

Estados Unidos de América y esto refleja exactamente la enorme inversión en ciencia que realiza ese país. El aporte del resto de países intervinientes en el consorcio también está ligada a esta inversión.

El mayor logro de este proyecto es sin duda la lectura del genoma del hombre y las puertas que abre este trabajo, y el conocimiento adquirido, al desarrollo de nuevos proyectos. Disponer de la secuencia del genoma permite hoy un análisis diferente del procedimiento y de la secuencia. El análisis de esta última en particular nos ha llevado a cierta reflexión respecto de nuestra posición respecto de otros organismos. En el aspecto de mayor interés, su relación con la salud humana, no era de esperar obtener mayores resultados inmediatos, aunque poco a poco se perciben algunos avances.

El proyecto y la enorme inversión monetaria realizada dieron un impulso inusitado al desarrollo de nuevas tecnologías. Siendo estas transferidas a, o directamente desarrolladas por, compañías privadas, se percibe ya un gran interés por la vulgarización de su uso y el consumo global de instrumentos y reactivos. Todo ello deriva por supuesto en algunas fricciones ante el ocasional –y ya demostrado- interés por transgredir algunas barreras éticas.

Límites metodológicos

Ninguno de los proyectos para el estudio del genoma humano logró su secuencia completa. Debemos continuar o debemos conformarnos por el momento con examinar sólo una porción, aquella alcanzable.

En teoría, cuantas más secuencias se lean se estará más cerca del ensamblado completo y de la lectura sin interrupciones de un genoma. Cada genoma presenta sin embargo regiones particularmente inaccesibles dando como saldo, una lectura final incompleta.

El genoma humano posee en total 3,08 Gb. Hoy, habiendo transcurrido más de diez años desde el anuncio del secuenciado del genoma humano, no podemos vanagloriarnos de conocerlo todo pues la porción "leída" de manera confiable apenas supera las 2,85 Gb. Desde entonces es cada vez más frecuente hablar de la parte conocida como de la "porción secuenciable" del genoma.

Las técnicas e instrumentos modernos mejor diseñados permiten raramente una lectura que se aproxime al 97% del total del genoma humano en una lectura. En ésta, abundan errores de lectura y quedan muchos fragmentos libres. Lecturas posteriores y coberturas de más de 30X raramente han podido hallar sin ambigüedad la secuencia de más del 90 % del genoma humano. Esto resalta el mérito del secuenciado paciente, clon a clon, y un resultado global comprobado que hoy sirve de referencia.

Del principio que la cobertura no es uniforme a lo largo del genoma, y sigue una distribución de Poisson, se desprende que el aumento en el número de

lecturas debe ser considerable para permitir leer más que en la primera ronda. Este aumento permite sin embargo confirmar secuencias ya leídas, y con esto asegurar la asignación de una base a una posición particular.

Es necesario aclarar que el objetivo de un proyecto de secuenciación genómica es conocer la secuencia íntegra. Sin embargo es más importante que esta se lea correctamente. El grado de cobertura en un proyecto de secuenciado no nos da una idea clara de la confiabilidad de la secuencia final obtenida si no va acompañado de otros valores. Citar el grado de cobertura del genoma, junto al tamaño promedio de las lecturas nos da una idea más acertada aunque aún no es completamente confiable. Para quienes esperan un valor que indique la confiabilidad de la lectura realizada, como quienes estudian las variaciones de secuencias entre individuos, un valor más apropiado es el porcentaje del genoma con asignaciones de bases correctas.

El porcentaje del genoma (humano) correctamente asignado es, con las mejores estrategias hoy logradas, del 93%. La imposibilidad de conocer el resto del genoma reposa en la dificultad para leer regiones ricas en GC y en la asignación de fragmentos a regiones para las cuales no existen marcadores físicos o genéticos. Mediante la secuenciación solamente, no es posible aún estimar el tamaño y la importancia de regiones compuestas casi exclusivamente de secuencias repetidas, abundantes en telómeros, centrómeros y brazos cortos de algunos cromosomas.

Aunque los mayores límites parecen surgir del mismo proceso de secuenciado, las mayores dificultades se hallan en la preparación y la pérdida del material de partida. Cuando se generan fragmentos cientos de miles de veces más pequeños que el genoma de interés, hay una probabilidad de que algunos estén menos representados. Cada paso en la preparación del ADN representa entonces una posibilidad de perder ciertas secuencias.

Ahora estamos en condiciones de asumir que durante el proceso de secuenciado de genomas completos se perderán fragmentos. Cuánto y cómo se perderá depende de la manipulación del proceso y la intención determinará cuánto queremos ver y qué tan cuidadosos debemos ser.

El interés hace al método

Tanto el secuenciado aleatorio como el de clon contiguo se revelaron como estandartes de proyectos que no siguieron estrictamente estos procedimientos. El consorcio público emplearía un también el secuenciado aleatorio y el grupo privado recurriría a los mapas y secuencias de su concurrente. Proyectos tan grandes, es hoy claro, no podrían estar ligados a un solo método. Desde entonces es posible ver que la terminología se ajusta mejor a cada proyecto de secuenciado, definiendo el método con más exactitud.

El secuenciado "basado en mapas", "basado en BACs" o "clon a clon", como el procedimiento empleado por el consorcio internacional recibe el nombre de secuenciado Shotgun Jerárquico. Este involucra la obtención y organización de un gran número de clones con insertos (típicamente de 100–200 kb) que cubren el genoma seguido del secuenciado aleatorio de cada clon de interés. Para algunos autores, este procedimiento describe un proceso más simple que aquel que evolucionara lo largo de 15 años. El procedimiento de ordenar los clones BACs podría reducirse al uso del fingerprints pero la localización de marcadores físicos y genéticos como ESTs o STS se hace en general indispensable.

La técnica, empleada para secuenciar el genoma de la rata (Gibbs *et al.* 2004) consiste en un secuenciado aleatorio del genoma completo combinado con el clonado en vectores BAC a baja cobertura (*x* 1). Los clones BAC se emplearon como esqueleto sobre el cual irían montándose los contigs logrados con el secuenciado aleatorio. Este método se conoce como *Secuenciado mixto* (mixed strategy sequencing)

En el *Secuenciado selectivo* (RRS o Reduced Representation Sequencing) se seleccionan fragmentos del genoma para evitar secuenciar grandes regiones sin, o de escaso, interés. Para hacer esto se prepara una biblioteca de fragmentos de restricción seleccionados por tamaño, se clonan y se secuencian todos los fragmentos de manera aleatoria. La selección de fragmentos de restricción de un tamaño determinado reduce la porción del genoma efectivamente secuenciado (Altshuler *et al.* 2000). Este método no es un secuenciado de genoma completo aunque en genomas muy grandes, como el del maíz (*Zea mays*) se ha empleado como aproximación para reconocer genes codificantes (Barbazuk *et al.* 2005). La selección puede basarse en otros métodos como la captura de secuencias metiladas por filtración, la selección de secuencias repetitivas basada en la tasa relativa de reasociación (Springer *et al.* 2004).

Gran parte de las plataformas de secuenciado a gran escala emplean actualmente una selección de secuencias. Algunos ejemplos de ellos son la separación por tamaño (mediante electroforesis en gel) para tener una mayor representación de ciertos RNAs (Glazov *et al.* 2008), el secuenciado con bisulfito para el análisis del genoma metilado (Meissner *et al.* 2008), la separación de cromosomas para el mapeo de sitios de ruptura en cromosomas translocados, el enriquecimiento de sitios de unión de factores de transcripción tras la precipitación de la cromatina (CHiP-Seq) (Cuddapah *et al.* 2009).

El secuenciado de ESTs es asimismo un secuenciado selectivo en el que intenta conocerse la secuencia de todos los transcriptos, y por ende los genes expresados, de un determinado tejido (Adams *et al.* 1991). En este se obtiene una única secuencia desde cada extremo de cada ADNc clonado, por lo que resulta relativamente económico.

TRATAMIENTO DE LA INFORMACIÓN

La lectura y la comprensión de un genoma no serían posibles sin una computadora y programas adecuados para el manejo de secuencias. Las razones para esto son obvias, no muchas personas pueden leer palabras tan largas escritas con cuatro letras.

Sabemos interpretar algunas señales que inmersas en esta hebra. Un gen que codifica una proteína tiene cierta estructura, un comienzo y un final (o varios de cada uno), que puede contener regiones no codificantes y pueden existir regiones regulatorias dentro y fuera de cada uno de esos genes. El código en el que se escribe la secuencia de una proteína es conocido desde hace largo tiempo pero no es posible saber aún como se lee todo el resto. Y esto sólo suponiendo que ya hubiésemos llegado al momento en que pudiéramos interpretar la información recogida.

La secuencia de un genoma, hemos visto, no puede (por el momento) recogerse íntegramente en un solo intento. Fragmentar el genoma ha sido una necesidad y fragmentarlo más o menos está ligado al método experimental empleado. Para obtener la secuencia completa de un genoma es necesario montar esos fragmentos. Un ordenador es la herramienta indispensable para lograrlo.

La computadora ha estado presente en el laboratorio de biología molecular desde su origen. En 1964 ya se desarrollaban sistemas de análisis y procesamiento de la información. Estos primeros intentos se orientaron a la comprensión de una secuencia proteica y muy pronto a las del ARN. En poco tiempo se intentaría comparar unas secuencias con otras, buscar regiones con características particulares, y predecir estructuras y señales particulares. A mediados de la misma década era posible analizar secuencias de ARN, predecir la existencia de puentes entre bases y la formación estructuras secundarias. Después de 1965 era posible conocer la secuencia de una proteína desde la secuencia de un ARNm o inferir la secuencia del mensaje desde la proteína (Upham 1970).

Con la revolución genómica y la lectura del código guardado en los organismos biológicos, la creación y mantenimiento de bases de datos se volvió habitual. Desde el secuenciado del genoma del bacteriófago ΦX174, universidades y laboratorios de Estados Unidos disponían de sistemas propios para el almacenamiento de información. En 1981 la Fundación Nacional para las Ciencias (NSF, National Science Foundation) dispuso la creación de una red para promover la conectividad. En 1986 este mismo organismo disponía la existencia de supercomputadoras en todo el país y aseguraría la conexión de cada laboratorio e institución educativa a internet con una velocidad de 56 kbit/s.

Las bases de datos y bibliografía más visitadas surgen como consecuencia del proyecto genoma humano. En 1988 nace el Centro Nacional para la Información Biotecnológica (NCBI), organismo que administrará desde

entonces, entre otros, el Banco de datos Genéticos (GenBank) y el index de revistas científicas relacionadas con la biología y la medicina (PubMed). En el primero, así como en dos sitios pares de Europa (European Molecular Biology Laboratory, EMBL) y Japón (DNA Data Bank of Japan, DDBJ), se registran y almacenan todas la secuencias de ADN, ARN y proteínas de todos los organismos conocidos. Las tres bases son de acceso público y proveen un gran número de herramientas para el análisis de secuencias.

La incorporación de nuevas estrategias para el secuenciado de ADN permite hoy adquirir miles de millones de bases al año. Este increíble adelanto en la capacidad de obtener información requiere un incremento consecuente en la capacidad de almacenar y procesar la información recogida.

El Consorcio Internacional de Secuenciación del Genoma Humano y Celera Genomics produjeron 23 y 27 mil millones de bases de secuencias crudas respectivamente. Desde entonces esta cantidad ha crecido exponencialmente debido a la incorporación de nuevos instrumentos de análisis (ver más abajo). Durante el año 2012 el NCBI anunció que no podría hacer frente a la demanda creciente de almacenamiento de secuencias y el organismo público ha aceptado aportes privados para su sostén.

Las secuencias obtenidas durante el secuenciado no son, en general, volcadas directamente en estas bases de datos. Estos fragmentos de tamaño reducido deben ser ensamblados en otros de mayor orden y su secuencia debe ser verificada (curada). Sólo entonces se ponen a disposición del público y pueden servin de referencia a estudios posteriores.

Ensamblado de Secuencias de ADN

El tamaño y la complejidad de un genoma condicionan la elección de un método de secuenciación. De manera semejante, el tamaño del genoma, su complejidad, y el tamaño de las secuencias obtenidas determinarán en gran medida qué estrategia de montaje debe, o puede, emplearse para ensamblar de manera exitosa la secuencia final.

La estrategia empleada para el ensamblado depende de varios factores, entre ellos la calidad de los datos de secuencia, el tamaño de los insertos, el carácter uni o bidireccional de las lecturas, la construcción de una biblioteca genómica y el vector de clonación, la selección de clones para el secuenciado, y fundamentalmente, la disponibilidad de información adicional (mapas de genes, secuencias consenso, ESTs, genes conocidos confirmados, etc) (Figura 23).

La mayoría de los programas de montaje siguen un esquema básico de superposición-disposición-consenso (Scheibye-Alsing *et al.* 2009) que procede según:

1. Examen de la calidad de las secuencias y limpieza de las lecturas.
2. Detección de coincidencias entre lecturas. Falsas superposiciones, lecturas duplicadas, quiméricas o con autocomplementariedad (incluyendo secuencias repetitivas) deben ser identificadas y dejadas de lado para estudio posterior.
3. Agrupamiento de secuencias para obtener una distribución de contigs de secuencia terminada.
4. Alineamiento múltiple para obtener secuencias consenso de cada grupo de contigs (con valores de calidad para cada base asignada).
5. Identificación de sitios mal ensamblados mediante inspección manual y validación de datos de calidad.

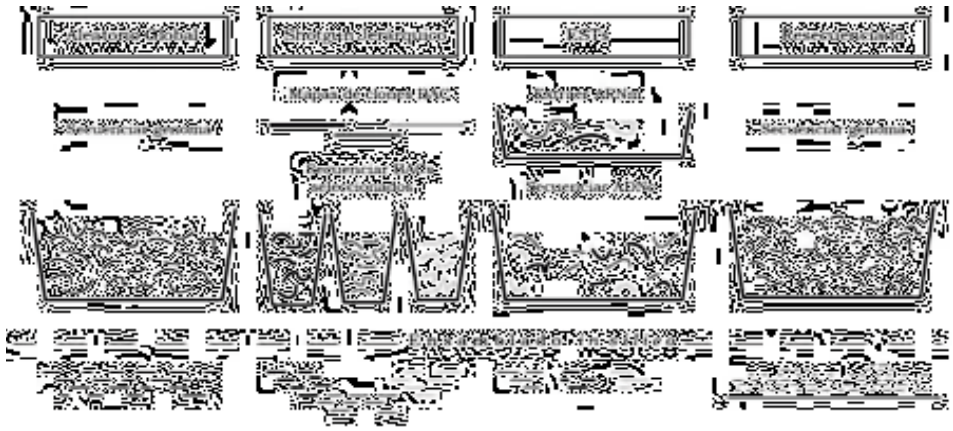


Figura 23. Ensamblado de secuencias. El método de secuenciado y la disponibilidad de información complementaria determinan en gran medida el método de ensamblado más apropiado.

Limpieza de secuencias

Tal como las técnicas de laboratorio, los programas para el ensamblado se desarrollaron y fueron adquiriendo complejidad conforme se detectaban problemas. Los ensambladores de primera generación requerían de la concurrencia de otros programas que permitieran "limpiar" las secuencias. Cada lectura de secuencia debe ser examinada para verificar la presencia de secuencias contaminantes, ADN del vector de clonado o de *Escherichia coli*. También es común buscar y eliminar regiones con secuencia de baja calidad, lecturas muy cortas (menos de 100 bp en secuenciado Sanger), adaptadores y colas o señales de poliadenilación.

Una característica esencial en la secuenciación es la asignación de valores de calidad a las secuencias crudas. Los valores de calidad indican la probabilidad de que cada base asignada (base call) durante la lectura sea correcta. Para obtener estos datos se emplean programas llamados basecallers, como PHRED (Ewing & Green 1998). En la fase de montaje los valores de calidad ayudarán en la distinción entre errores de secuencia y polimorfismos, y las secuencias finales de baja

calidad podrán corregirse o no emplearse en el ensamblado. La lectura de calidad se asocia en general con la lectura del instrumento, la intensidad de cada señal, el espaciado o la regularidad con la que se registra cada señal.

Con las secuencias corregidas puede comenzar el ensamblado. Para esto el programa debe tomar una porción de una secuencia y buscar entre el resto de los fragmentos la existencia de una porción idéntica. El tamaño de la región así como la existencia de una proporción de no-identidad pueden -o deben- definirse con antelación. Hallada la similitud entre dos fragmentos, o parte de ellos, el programa debe tomarlos como si fueran uno y continuar hacia uno u otro extremo, así prolongando un contig.

El alineamiento de secuencias se hace corriente con programas que se pueden emplear en un laboratorio pequeño con ordenadores corrientes de escasa capacidad. Programas muy difundidos de uso público como CLUSTAL (Higgins & Sharp 1988) y BLAST (Altschul *et al.* 1990) permiten alinear y ordenar fragmentos de una pequeña base de datos y con secuencias almacenadas en servidores distantes. Los primeros ensambladores desarrollados ya escapan sin embargo a la capacidad de trabajo para un ordenador corriente. Estos podían manejar genomas de varios millones de bases como el de una bacteria. El ensamblador del TIGR (Sutton *et al.* 1995) ha sido uno de los más empleados y ya se desarrollaban otros con fines comerciales.

El trabajo de Celera deja ver que la dificultad de disponer de un único programa capaz de manejar toda la secuencia obtenida a partir de fragmentos de un gran genoma. El ensamblado de la secuencia del genoma humano requirió de cinco programas sucesivos. Screener permitía hallar y marcar microsatélites (con módulos de menos de 6 pb) y reconocer todos los elementos repetitivos conocidos (Alu, Line, y ADN ribosomal). Overlapper se empleaba entonces para comparar cada lectura con todas las otras en busca de superposiciones completas de un mínimo de 40 pb. Para este cómputo solamente se emplearon unidades multiprocesador con cuatro procesadores y 4 gigabytes de RAM. Cuarenta máquinas semejantes trabajando en paralelo completarían la comparación en aproximadamente 5 días. Un tercer programa, Unitigger, se empleó entonces para discriminar contigs reales de otros obtenidos debido a la presencia de repeticiones. Los contigs así separados o al menos limpios de un exceso de elementos repetitivos se ensamblaron en supercontigs o andamios de grado superior con otro programa, Scaffolder. Este se diseño para encontrar secuencias distantes presentes en los extremos de BACs con secuenciado terminal.

Identificación de superposiciones y alineamiento

Actualmente existen más de veinte sistemas para el ensamblado de secuencias (muy pocos de ellos públicos). Aunque el trabajo de ensamblado de

fragmentos es esencialmente el mismo, se perciben grandes diferencias en su manera de abordar el proceso. Estas diferencias son muy visibles en el tratamiento de las superposiciones, el requerimiento de información adicional y la lectura de secuencias muy cortas.

Presentando al programa millones de secuencias el programa, éste va a utilizar algoritmos que permiten la detección de identidad (superposición) entre palabras, o pequeños fragmentos de secuencias (words, k-words o k-mers) (Gotoh 1982; Smith & Waterman 1981). Cada superposición se emplea a su vez para hallar otras secuencias con la misma, y deducir si representan una superposición real o un artefacto. El tamaño de la palabra empleada varía entre distintos ensambladores, y puede ser fija, predeterminada, o variable.

El alineamiento emplea luego el mismo algoritmo para todas las estrategias (Smith & Waterman 1981). Este se logra mediante la construcción de matrices con puntuaciones (scores) asignadas a cada superposición

Aunque alinear dos secuencias es simple, lograr alinear más de dos y obtener una secuencia que represente a todas, un consenso, no lo es. A medida que se agregan palabras a la comparación el número de cálculos crece exponencialmente. Esto se ha resuelto mediante la implementación de algoritmos de tipo voraz (heuristic Greedy algorithm). Estos algoritmos a añaden una secuencia a la vez de manera reiterada y eligen una opción en cada paso en vez de probar todas las múltiples combinaciones posibles.

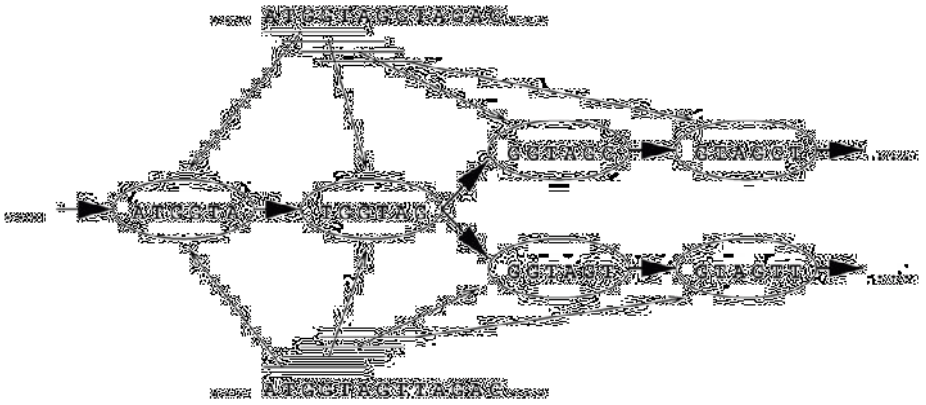


Figura 24. Construcción de un Gráfico. Dos lecturas construidas a partir de k-mers k = 6. Los límites de los nodos vienen determinados por el tamaño de las lecturas. Un nucleótido diferente lleva a la división del gráfico y a una ambigüedad en la secuencia. La frecuencia de lecturas de uno y otro tipo ayudará a determinar si es un error de secuencia o un polimorfismo.

Tras el alineamiento múltiple puede llegarse a una secuencia consenso, en la cual se representa cada posición con la base más comúnmente hallada. En algunos casos el programa asigna un consenso según palabras de una longitud definida (12-mer en geneDistiller).

La ruptura aleatoria de un genoma diploide suponía un gran problema para resolver la existencia de variantes entre cromosomas homólogos, desde un polimorfismo de un nucleótido hasta grandes deleciones. Sin embargo el uso de métodos gráficos ha permitido evitar el uso de alineamientos múltiples, particularmente cuando se obtiene un gran número de secuencias cortas. Un procedimiento de este tipo se ejemplifica en la Figura 24.

Un problema mayor en el gran genoma de un mamífero es la presencia de repeticiones. Las secuencias repetidas pueden representar más del 50 % del total del genoma y son de naturaleza muy variada: secuencias derivadas de transposones, duplicaciones en tándem, palíndromos, o secuencias dispersas como genes ribosomales, centrómeros, heterocromatina y retroposones.

Dependiendo de la profundidad del secuenciado los elementos repetitivos pueden hallarse por comparación con otros genomas o directamente verificando la existencia de excesos de lecturas en ciertas regiones (Figura 25). En ambos casos las secuencias repetidas se dejan de lado para el ensamblado final.

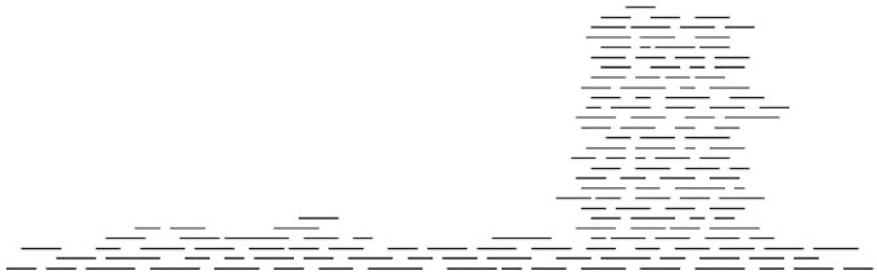


Figura 25. Detección de repeticiones mediante observación de la profundidad de secuenciado. Regiones repetidas se ponen en evidencia con excesiva representación de algunas lecturas.

Existen programas y bases de datos para hallar y borrar, u ocultar al ensamblador, secuencias repetidas (Wheeler *et al.* 2013).

Los ESTs son sumamente útiles al ensamblado cuando los genes se hallan interrumpidos por intrones dado que permiten saber que dos porciones distantes se hallan sobre una misma hebra del cromosoma. Sin embargo su uso presenta algunas dificultades. El estudio de secuencias expresadas, particularmente cuando los ADNc representan fragmentos y no mensajeros completos, lleva con frecuencia a su agrupamiento de secuencias semejantes. Esto ha llevado a menudo a tomar fragmentos de diferentes genes por uno solo. Algo semejante ocurre con genes superpuestos, en el mismo sentido o en sentidos opuestos.

ESTs muy representados debido a su elevado nivel de expresión pueden en ocasiones ocultar genes menos representados. Asimismo, grupos de genes expresados suelen tener cientos de miles de secuencias que hacen difícil obtener un consenso. Para reducir este efecto se suelen eliminar transcriptos de genes housekeeping. These procedures can alleviate some of the problems, however

Armado de cromosomas

Tras el ensamblado de lecturas individuales en contigs, éstos se posicionan en el genoma para constituir andamios (scaffolds o super-contigs). Los contigs se ordenan y orientan hasta que puede armarse un andamio que representa un cromosoma. Espacios entre contigs deberían llenarse entonces mediante una estrategia apropiada o resecuenciando.

La información para el ensamblado de super contings no viene necesaria o únicamente de las lecturas. Para cubrir sitios distantes se recurre a las posiciones de marcadores físicos (particularmente STSs), a secuencias ESTs y, si se dispone de ellas, a secuencias de extremos de clones largos.

Terminación

Concluido el ensamblado la secuencia aún requiere un examen detallado. Existirán regiones de baja calidad o baja (o nula) cobertura, y regiones probablemente mal ensambladas. La inspección es, sin dudas, manual. Por esta razón un programa que construye gráficos es de suma utilidad.

Existen programas específicamente diseñados para esto. Consed (Gordon et al., 1998), Autofinish (Gordon *et al.*, 2001), BACcardi (Bartels *et al.*, 2005), y GAP4 (Bonfield *et al.*, 1995), permiten, recorrer lecturas, contigs y andamios en busca de errores, marcar regiones sospechosas y reordenarlas o volver sobre ellas más tarde.

Para la verificación de superposición entre clones el consorcio público se impuso un límite de superposición de secuencias terminales mínimo de 2 kb con una identidad mayor al 99.6 % (Consortium 2004). En el ensamblado de 26.720 contigs, se aceptaron 308 con un porcentaje de identidad menor al valor exigido. El error estimado de este ensamblado de toda la secuencia es de 1 por cada 100.000 bases.

Cuando el análisis informático con secuencias del mismo genoma humano no fue suficiente para el llenado de brechas se recurrió a la búsqueda de fragmentos de ESTs y a secuencias homólogas en el genoma del ratón. Cuando estas dos estrategias no alcanzaron el consorcio internacional volvió a la hibridación mediante sondas basadas en secuencias del ratón y al paseo cromosómico en el laboratorio húmedo.

La secuencia del genoma humano lograda por el PGH consta de 2.851.330.913 nucleótidos, presentes casi todo ellos en la porción eucromática. La secuencia está interrumpida por 341 brechas, entre las cuales 33 están en regiones heterocromáticas y totalizan 198 Mb, las restantes cubren aproximadamente 28 Mb. Aunque el grado de cobertura empleado fue de 6 a 10 X para la

mayoría de los clones secuenciados, algunas brechas no pudieron cubrirse con una cobertura de 30 X. Se estima que el genoma completo posee 3.08 Gb.

Resecuenciado

El número creciente de secuencias de genomas en las bases de datos hace posible el secuenciado de un gran número de fragmentos y el análisis de éstos sobre una secuencia de referencia (Scheibye-Alsing *et al.* 2009). Este procedimiento, el resecuenciado, no hace posible cubrir regiones refractarias o la asignación correcta en regiones de baja cobertura pero el análisis es más simple, y el costo final del proceso se reduce significativamente.

El resencuenciado se ve como una estrategia útil para poner en evidencia la variación humana con distintos propósitos. Para reducir aún más los costos y dar una utilidad rápida a esta tecnología aparecen intentos por seleccionar y limitar las lecturas a ciertas regiones de mayor interés. Un método permite por ejemplo la selección de fragmentos correspondientes a aproximadamente exones humanos mediante hibridación en microchips y el secuenciado posterior de los fragmentos seleccionados (Porreca *et al.* 2007).

El avance de la tecnología del secuenciado (ver más adelante) y el desarrollo continuo de instrumentos de análisis de secuencia han hecho posible reducir enormemente el tiempo de lectura e interpretación. Los quince años necesarios para obtener el primer genoma han podido en 2007 reducirse a tan sólo dos meses para tener la secuencia del genoma de James D. Watson (Wheeler *et al.* 2008).

Se obtuvieron 106,5 millones de lecturas que representan 24,5 billones de bases (una cobertura de 7,4 X) que se filtraron analizando el alineamiento de cada lectura con un genoma de referencia mediante BLAT (Kent 2002). Cada lectura se incluyó en el análisis subsecuente si estaba representada en más del 90%, demostraba un único sitio de homología en el genoma, contenía menos de 5 diferencias de secuencia y menos de 5 indels. Los 93 millones de lecturas resultantes de la selección se alinearon con Cross-match. Las diferencias resultantes se examinaron manualmente (y se desarrolló un algoritmo) para juzgar la presencia de variantes y errores. Las secuencias de baja calidad o escasa similitud se recortaron para eliminar las bases finales, donde acumulan los mayores errores de lectura, y se ensamblaron con ATLAS (Havlak *et al.* 2004).

Capítulo 4

¿Qué hay en el genoma?

4

Desde la primera asociación de una secuencia de nucleóticos con un polipéptido (JACOB & MONOD 1961) la interpretación de fenómenos moleculares ha dependido del reconocimiento de particularidades en una secuencia. La identificación de secuencias nos ha permitido desde entonces inferir la existencia de otras regiones con actividad semejante. Determinar la existencia de un gen, la presencia de intrones, de regiones regulatorias, de modificaciones en el ADN, el ARN o las proteínas, y hasta la estructura, función y localización de una proteína son posibles gracias a inferencias (Stamatoyannopoulos 2012).

Gran parte de las medidas recogidas a partir de la secuencia del genoma completo son inferencias. Conocemos las particularidades de algunas secuencias, hemos verificado que algunas características pueden extenderse a otras, y ahora tratamos de extrapolarlas a las secuencias de todo el genoma. Tratamos de descubrir algo oculto y por el momento vamos a tientas. Sabemos que la información que vamos a obtener puede cambiar o ser confirmada a medida que lo permitan nuevas observaciones.

Muchos aspectos del genoma eran conocidos antes de su secuenciado completo. Más interesados en las secuencias codificantes, aquellas que dan un producto proteico, hemos podido acceder con cierta facilidad a los transcriptos y muchas de ellas nos resultan conocidas. Una gran proporción del genoma consiste de secuencias repetidas y secuencias no codificantes cuya función permanece largamente desconocida.

GENES

Completa la secuencia del genoma humano, el análisis debía responder a una pregunta formulada desde el comienzo de la genética contemporánea: cuántos genes tiene el hombre? Por ese entonces, los cambios inducidos en el ADN (mutaciones) mediante radiaciones permitían ver variaciones groseras en rasgos visibles. Conforme avanzamos en el conocimiento del gen, de las proteínas

codificadas por algunos de ellos y de las múltiples formas que éstas pueden adoptar, algunos investigadores sugieren que el genoma humano debe albergar al menos 100.000 genes.[11] La era genómica guardaría sin embargo algunas sorpresas.

Estimaciones basadas en el examen de ESTs y de la secuencia de los primeros genomas (Venter *et al.* 2001) indicaron que el genoma humano contiene entre 25.000-35.000 genes y sus secuencias codificantes ocupan menos del 2% del total del ADN (Figura 26).

El catálogo de referencia provisto por el PGH, Ensembl 34d, indica que, la fracción de secuencia leída tiene 22.287 loci (Consortium 2004). La longitud total del genoma cubierta por estos genes es de 34 Mb, equivalente al 1,2 % del genoma conocido. Estimaciones más recientes dan un total de 20.687 genes codificantes (The ENCODE Project Consortium 2012).

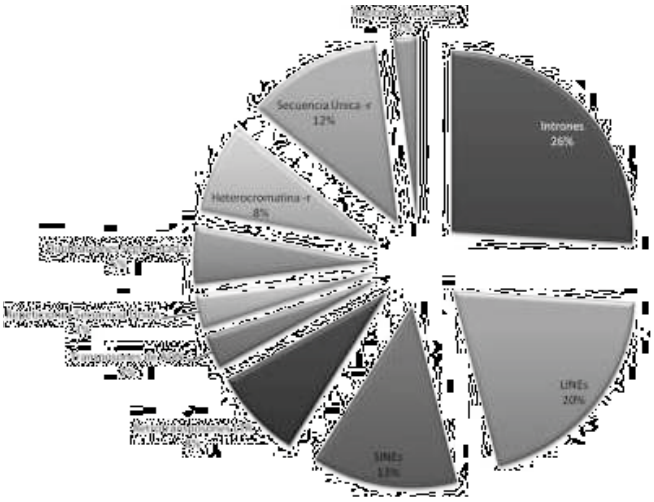


Figura 26. Componentes del Genoma Humano. Aproximadamente un 2% del genoma humano correspondería a regiones que pueden dar origen a proteínas. Intronesy regiones repetidas, incluyendo elementos móviles cubren la mayor parte de las secuencias estudiadas. Regiones escasamente cubiertas hasta el momento se agrupan como secuencia única y heterocromatina restante (r). Por detalles de cada una ver texto a continuación.

El mayor problema en estas estimaciones es la predicción correcta de secuencias que pueden dar origen a proteínas de tamaño menor a 100 aminoácidos. Presuponiendo que las regiones sin secuencia definida puedan albergar algunos genes el valor final no debería superar lo 25.000 genes.

El número de genes debía explicar para muchos las grandes diferencias existentes entre el hombre y el chimpancé, el ratón o la mosca, pero el número obtenido no sugiere mayores diferencias. Si la complejidad cada organismo se basa en la cantidad de genes, entonces el hombre no es tres veces más complejo que una mosca de la fruta (*D. melanogaster*) y menos de dos veces más que el gusano nemátodo *Caenorhabditis elegans*. El proyecto genoma del chimpancé y del mono Rhesus permitieron confirmar en cierta medida estos hallazgos pues

11 Müller había estimado en 1930, que la mosca debería tener alrededor de 20.000 genes y el hombre no más de 30.000.

ambos comparten un número similar de genes con el hombre (Hubbard *et al.* 2005). Las diferencias deben entonces ser cualitativas.

La mayoría de los genes humanos da origen a más de un transcripto, se estima que de cada locus deben surgir 1,5-1,7 de ellos. El uso diferencial de sitios de inicio de la transcripción y de terminadores prematuros puede llevar a la existencia de secuencias más largas o más cortas en 5' o en 3. De manera semejante el mensaje puede ser interpretado de manera diferente por los ribosomas y llevar a obviar o añadir regiones para traducir la información de manera diferente (Figura 27).

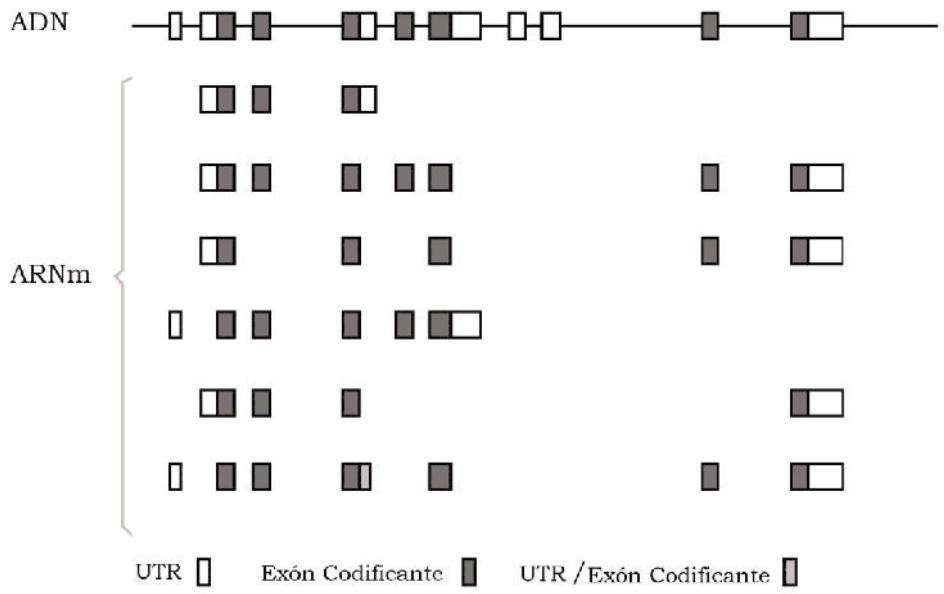


Figura 27. Estructura de un gen humano. El gen esquematizado arriba según aparece o puede estimarse a partir de la secuencia de ADN puede dar origen a transcriptos en los que se halla una secuencia 5' diferente (presente en el 4to y 6to ARNm) o de una región alternativa (en el último transcripto), posible gracias a una lectura alternativa. El procesamiento del ARN heterogéneo nuclear puede a su vez dar ARNm en los que se hallan ciertos exones y otros en los que se encuentran variantes que dan como resultado proteínas con características diferentes.

Cada transcripto puede a su vez ser procesado de manera diferente. Los genes eucariotas son típicamente discontinuos, compuestos por regiones codificantes (exones) separadas por largas regiones no codificantes (intrones). Durante la expresión génica, el ARN se sintetiza copiando la secuencia del gen entero, incluyendo exones e intrones. Los intrones presentes en el pre-ARN mensajero se eliminan durante el splicing para dar un ARNm que saldrá del núcleo y dirigirá la síntesis de una proteína. Con frecuencia un pre-ARNm sufre splicings alternativos o diferenciales, desde los cuales surge una serie de diferentes combinaciones de exones (Figura 22). Como en las demás estimaciones, las cifras

varían. Los genes catalogados contendrían 231.667 exones, un promedio de 10,4 exones por locus y 9,1 exones por transcripto según el estudio finalizado por el PGH (Consortium 2004). Estudios más recientes indican que un gen codificante puede dictar, en promedio, la existencia de 6.3 transcriptos diferentes (Howald *et al.* 2012). El mismo estudio afirma que más del 90 % de los genes codificantes dan origen a más de un transcripto.

Hasta el momento sólo ha podido compararse globalmente la información originada en la transcripción (ESTs). Con estos datos llegamos a menos de 200.000 proteínas encargadas de todas las funciones celulares. Aún es necesario combinar esta información con datos de estudios de expresión de proteínas totales (proteómica) para tener una idea más certera de lo que realmente ocurre.

No existe un gen modelo. El genoma cuenta con genes de estructura muy variada, algunos carentes de intrones y en el extremo opuesto uno con 363 exones, uno de ellos de 17 kpb, esparcidos en una región de más de 280 kpb. Este gen codifica una proteína muscular conocida como Titina, cuyo tamaño varía entre 27.000 y 33.000 aminoácidos.

Un gen con una actividad conocida no reside generalmente en un sitio vecino a otro con una función semejante, o implicado en una misma ruta metabólica. Así es común hallar un gen que codifica para un factor de transcripción junto a otro implicado en la síntesis de oligosacáridos o uno que dará origen a una proteína de secreción. Aunque esto era ya conocido, o se preveía, el estudio del genoma permitió confirmarlo.

Asimismo los genes se distribuyen de manera muy desigual a lo largo del genoma. Algunos cromosomas, como el 1 y el 19, contienen grupos muy densos de genes en tanto que pueden hallarse grandes regiones aparentemente libres de secuencias codificantes. A pesar de la escasa proporción del genoma cubierta por exones, la superficie del genoma cubierta por genes, desde el primer codón de inicio hasta el último, alcanzaría el 30 % del genoma.

Pseudogenes

El genoma humano, como el de otros organismos, contiene muchas secuencias sin función reconocida que se pueden ver como vestigios evolutivos. Pseudogenes y genes truncados son dos formas frecuentes en este tipo de secuencias.

Un pseudogen es una copia *no funcional* de un gen. Existen dos tipos de pseudogenes, los seudogenes convencionales y los seudogenes procesados. Los primeros surgen de duplicaciones de un gen original o de fragmentos de mayor tamaño y suelen residir en sitios cercanos al ocupado por el primero. Si la función del nuevo gen es superflua un evento de mutación, y su inactivación, pasará desapercibido. La acumulación de más mutaciones marcará su divergencia del

original. El tiempo desde la duplicación y el ritmo de acumulación de mutaciones determinan que en algún momento ya no sea posible reconocerlo como derivado de otro.

Los pseudogenes procesados derivan de la reinserción en el genoma de copias de ARNm. Los pseudogenes procesados se reconocen fácilmente por la ausencia de intrones y secuencias reguladoras de la transcripción. Su origen podría deberse a la copia del ARN a ADN (mediante transcripción reversa) y la inserción de este último en el genoma. La actividad de enzimas retrovirales es muy probablemente responsable de estos accidentes. La ausencia de señales necesarias para la transcripción sugiere que estas copias nunca fueron funcionales y pudieron comenzar a acumular mutaciones desde su inserción.

Cuando la similitud de una secuencia se limita a un extremo de un gen, no se refiere a éste como pseudogen sino más bien como gen truncado. Estos fragmentos de genes pueden también ser observados en todo el genoma.

A escala global, distinguir genes verdaderos de genes procesados no es posible sin cierto margen de error. El grado de cobertura es importante para confirmar la existencia de polimorfismos, variantes de splicing, codones de parada. Un criterio básico de búsqueda se limita a comparar genes con intrones con genes sin intrones pero la detección de secuencias codificantes o similares en ausencia de ESTs debe necesariamente incluir la divergencia de la secuencia codificante, la presencia de promotores y sitios reguladores adyacentes.

El número total de elementos arrojado por el primer análisis global supone la existencia de más de 20.000 pseudogenes (Consortium 2004). Trabajos más recientes han podido buscar pseudogenes mediante la comparación del genoma de muchos individuos, la comparación con secuencias de otros primates y el resecuenciado. Estos trabajos predice la existencia de 11.224 elementos (Pei *et al.* 2012).

De manera inesperada aparear genes y sudogenes con la gran cantidad de secuencias disponibles desde el año 2000 ha llevado a identificar parálogos de genes que, se presume, vuelven a ser funcionales tras un período de inactividad (Venter *et al.* 2001). Experimentos de RT-PCR-secuenciado llevaron a concluir que 863 elementos presuntamente inactivos, producen transcriptos (Pei *et al.* 2012). Estos elementos se hallan en regiones de cromatina activa y asociados a promotores. Hallar que algunos pseudogenes se volvieron genes verdaderos da cuenta de algunos cambios que nos deparan los estudios globales del genoma.

Genes nuevos y genes viejos

La variación entre individuos es gran parte un reflejo de la existente dentro del genoma. El cambio proporciona nuevo sustrato sobre el que actúa la selección. Desde el hace mucho tiempo se hipotetiza que la presencia de nuevo material genético, originado con la inserción de secuencias endógenas o exógenas,

podría proveer este sustrato (McClintock 1953). La duplicación de un gen proporcionaría una forma que puede evolucionar independientemente de la primera y la adquisición de una nueva función. Uno de los aspectos más relevantes de los proyectos de secuenciado de genomas completos surge de la posibilidad de medir el origen de secuencias a gran escala.

Para la búsqueda de secuencias casi idénticas deben establecerse parámetros de referencia tales como una tasa de sustitución de bases entre sitios sinónimos (KS). Esta tasa de sustitución expresa indirectamente los cambios acontecidos en un período evolutivo expresado en millones de años. Un KS de 0,15 equivale aproximadamente al tiempo distante de la separación de homínidos y roedores. En la búsqueda de pares de genes humanos con un KS igual a 0,30 (un KS de 0.15 para cada gen), se hallaron 1.183 genes (Consortium 2004). Este valor representa groseramente las diferencias que existen entre las secuencias codificantes de ambos grupos.

De manera semejante, un KS igual o menor que 0.015 permitiría evaluar cambios ocurridos en un período geológico más cercano, particularmente duplicaciones ocurridas hace menos de 3–4 milones de años. En esta búsqueda se han puesto en evidencia algo más de 300 genes El surgimiento de los primates podría así estar asociado a una explosión de duplicaciones. Alternativamente, estas similitudes podrían reflejar la homogeneización producto de un gran número de eventos de conversión génica que enmascara la divergencia. Buscar estos cambios en el genoma de otros mamíferos puede poner en evidencia los acontecimientos que llevaron a la separación de unos de otros.

Las secuencias de genes funcionales y de genes que han perdido su función no deben evolucionar al mismo ritmo. En el genoma completo es posible buscar pseudogenes no procesados con escasas diferencias respecto de la copia aún activa en otros órdenes. Nuestra evolución en este caso acontece por pérdida de funciones.

En el análisis de genes ortólogos presentes en el ratón común, la rata y el hombre, en los cuales la copia del último ha perdido su función en el hombre, se hallaron 33 secuencias. Esta cuenta incluye pseudogenes humanos en los cuales puede trazarse sin dudas el origen y guardan un elevado grado de similitud. De éstos, 19 pseudogenes humanos con dos codones de parada por secuencia se encuentran en una forma semejante en el genoma del chimpancé. Entre los 14 pseudogenes restantes, con un único codón de parada interrumpiendo la secuencia activa hallada en roedores, 8 tienen la misma interrupción en chimpancé, 5 aún serían funcionales en el mono y 1 parece ser un polimorfismo responsable de una proteína truncada. De los 32 pseudogenes fijados en la población humana 10 corresponden a receptores olfatorios. El número de receptores olfatorios funcionales habría decrecido en los primates respecto de los roedores, y sería menor en el hombre que en otros primates.

ARN NO CODIFICANTE

El resultado más sorprendente de los estudios del genoma completo es, probablemente, reconocer que la mayoría de las secuencias transriptas, funcionales, no llevan a la producción de proteínas. El término ARN no codificante se emplea para todo ARN que no es traducido para dar una proteína. Aunque muchos de ellos son conocidos, como los ARNs de transferencia, ribosomales, ARNsn (small nuclear RNAs) y ARNsno (small nucleolar RNAs), otros sólo comienzan a percibirse como importantes en el funcionamiento de la célula (Mattick & Makunin 2006).

ARNr

El ribosoma está constituido por dos subunidades formadas por cuatro ARNs y proteínas. La subunidad mayor posee ARNr de 28S 5.8S rRNAs (large subunit, LSU, rRNA) y un ARNr de5S. La subunidad menor contiene ARNr de 18S (small subunit, SSU rRNA). Los genes LSU y SSU se agrupan en un segmento de 44 Kb que se estimaba repetido 150-200 veces en el genoma, principalmente en los brazos cortos de los cromosomas 13, 14, 15, 21 y 22 (Long & Dawid 1980). Ninguno de los proyectos genoma realizado hasta el momento ha podido ensamblar todas las copias de estas repeticiones y en general carecen de segmentos completos del módulo mencionado (Venter *et al.* 2001; Lander *et al.* 2001; Consortium 2004).

Algo semejante ocurre con los ADNr 5S, agrupados asimismo en tándem. Habría 200-300 repeticiones como estas en el cromosoma 1, pero tampoco se han podido ensamblar por presentar sitios para enzimas empleadas en el clonado y por haberse excluido deliberadamente.

ARNsno

Los snoARNs dirigen, entre otras actividades, el splicing del ARNhn. En el grupo pueden reconocerse dos familias: C/Dbox, relacionados con la metilación de 2'-O-ribosa, y H/ACA empleados en pseudouridilaciones (Weinstein & Steitz 1999). Tras La búsqueda de estos elementos se reconocieron 97 secuencias. Casi todos ellos estarían en una sola copia (Lander *et al.* 2001). No existe por el momento suficiente información acerca de sus secuencias para emprender una búsqueda más exhaustiva.

ARNsn

ARNs que forman parte del spliceosoma también estarían subrepresentados en estos primeros estudios debido a que estos presentan rasgos semejantes

a los ARNr (debido a la presencia de sitios de restricción y a que fueron dejados de lado por presentar regiones de baja complejidad).

Se hallaron copias de casi todos ellos en la porción eucromática secuenciada (21 de 22 conocidos). Estos están dispersos en todo el genoma y algunos en múltiples copias, como los ARNsn para las partícula U6 (44 genes) y U1 (16 genes) (Lander *et al.* 2001; Consortium 2004).

ARNs pequeños

Los ARNnc incluyen ARNs de pequeño tamaño, o procesados para dar productos de tamaño reducido, como miARNs (micro), snARNs y snoARNs, y grandes transcriptos, entre los que se hallan elementos que entrelazan y superponen con transcriptos sintetizados en el mismo o diferente sentido. Los intrones también pueden ser incluídos entre los ARNnc. La mayoría de estos transcriptos tienen funciones desconocidas. Sus roles implican, en general, el reconocimiento de ácidos nucleicos. Algunos llevan adelante reacciones catalíticas relativamente simples, como ciertos intrones o las partículas que integran el complejo de splicing. Muchos parecen controlar señales relacionadas con la expresión variable de genes durante el desarrollo, la arquitectura de la cromatina, la memoria epigenética, la transcripción, el splicing, la edición, la traducción y el metabolismo.

Estudios comparativos entre secuencias genómicas y ARNs de tamaño mayor a 200 nucleótidos en líneas celulares sugieren que hasta un 60% del genoma sería transcripto (Thurman *et al.* 2012). El 30 % de estas secuencias serían intergénicas en tanto que el resto aloja en intrones o superpone con secuencias codificantes.

Los miARN han adquirido cierta relevancia en los últimos años por haberse reconocido su rol en procesos relacionados con la regulación del desarrollo, de la expresión génica en el cáncer y hasta el silenciamiento de transposones durante la maduración de las células germinales . El genoma humano codifica unas 100 familias conservadas de estos pequeños RNAs que actúan en general uniéndose a un mensajero y determinando su degradación. El análisis genómico ha permitido determinar que un miARN hallaría secuencias complementarias en unos 200 ARN mensajeros diferentes. Entre elementos altamente conservados se han podido detectar 39 UTRs con secuencias de 7 bases complementarias con las bases 2–8 de miARNs (Lewis *et al.* 2003).

REGIONES REGULADORAS

Regiones 5'-UTR, 3'-UTR (untranslated regions) y reguladores distantes representan una buena proporción del genoma. Las regiones UTRs abarcarían 21 Mb, aproximadamente 0.7 % del genoma conocido.

El examen de sitios de unión para más de cien factores de transcripción y componentes de ARN polimerasas llevó a estimar que más del 6 % del genoma estaría cubierto por sitios regulatorios (Whitfield *et al.* 2012).

Sitios hipersensibles a la DNAsa I y a la nucleasa micrococal se han interpretado desde hace tiempo como sitios accesibles a factores de transcripción y por ende, como sitios activos del ADN. Extrapolaciones que incluyen esta clase de secuencias arriesgan que las regiones regulatorias podrían alcanzar proporciones muy superiores al 20 % del genoma (Hesselberth *et al.* 2009).

Comparando el genoma del hombre con el del ratón parece evidente que el 6 % de la secuencia no codificante está muy conservada, y entonces habría sido seleccionada durante los últimos 100 millones de años (Chinwalla *et al.* 2002).

Entre los genomas del hombre, ratón y rata existen unos 500 elementos muy bien conservados (200 bases o más idénticas), que no se hallan en, o asociados a, secuencias codificantes. Variando la longitud de secuencias se hallan decenas a cientos de miles de secuencias muy conservadas entre estos genomas. En algunos casos estos elementos pueden trazarse hasta un elemento común que diera origen al humano y el de un pez. Muchos elementos no codificantes muy bien conservados residen en áreas carentes de genes aunque flanqueadas por genes con patrones de expresión específico de ciertos tejidos y/o muy importantes en el desarrollo embrionario (Figura 28) (Gibbs *et al.* 2004).

El secuenciado actual de casi 30 genomas de mamíferos permitió identificar millones de secuencias conservadas no codificantes y estimar que representan más de 2/3 de todas las secuencias bien conservadas. Estos estudios evolutivos empiezan a poner en evidencia la existencia de elementos funcionales no sólo presentes en promotores, UTRs, enhancers e aisladores (insulators) sino también familias de estructuras secundarias en los ARNs.

Aunque muchos de estos elementos se hallan conservados, todos parecen evolucionar más rápido que las secuencias codificantes. Aproximadamente el 20 % se secuencias conservadas no codificantes no son reconocibles en el genoma de un marsupial en tanto que sólo el 1 % de secuencias codificantes no hallan equivalente (Mikkelsen *et al.* 2007). Estos elementos habrían surgido en algún momento hace 90-180 millones de años (la separación de los marsupiales y el origen de los placentados) o se habrían perdido en el linaje de marsupiales. La proporción de elementos de este tipo conservados entre mamíferos y aves (310 millones de años) es significativamente menor, sólo el 30 % de los hallados en mamíferos, y en peces (450 millones de años) la similitud se aproxima a cero. El cambio, o recambio, en las secuencias de estas estructuras no codificantes conservadas sugiere que la evolución de las especies podría sostenerse en la innovación de secuencias no codificantes más bien que en las secuencias codificantes (King & Wilson 1975).

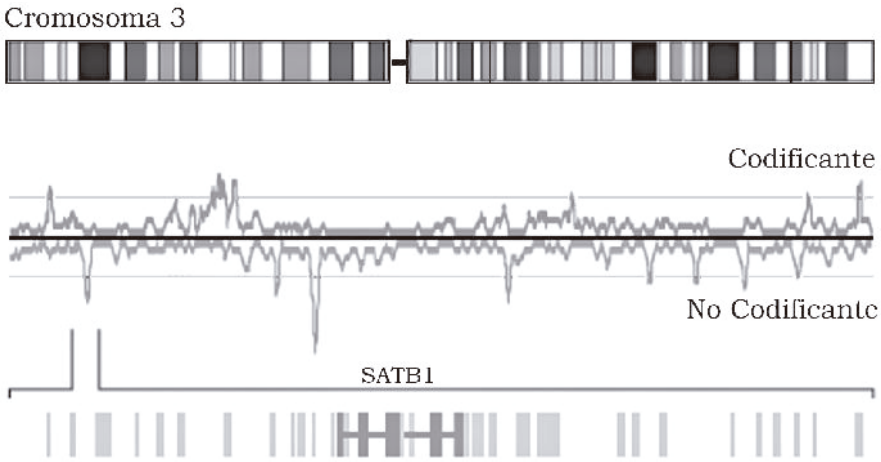


Figura 28. Distribución de regiones codificantes y non codificantes conservadas entre el hombre, el perro y el ratón. Debajo del ideograma del cromosoma 3 humano puede verse la distribución de transcriptos bien conservados destinados a la síntesis de proteínas (codificante) y segmentos no transcriptos con alto grado de conservación. El detalle de una región del brazo corto muestra en gris claro secuencias conservadas con transcriptas y en gris oscuro un gen implicado en el desarrollo embrionario (SATB1).

SECUENCIAS REPETIDAS

La mayor dificultad para el ensamblado del genoma humano deriva de la abundancia de regiones repetidas. Regiones intergénicas e intrones se hallan amplificados por una gran cantidad de elementos que se repiten en todo el genoma. Durante largo tiempo, estas secuencias se creyeron restos de ADN que se incorporaron, copiaron y diseminaron por el genoma sin representar alguna utilidad para el huésped, en ocasiones siendo perjudiciales. Este ADN, en apariencia inútil para el huésped recibió el nombre de "ADN basura".

Más del 50% del ADN se halla en forma de secuencias repetidas dos o más veces en el genoma. De manera simplificada se agrupa a las secuencias repetidas en cinco tipos diferentes según su estructura, origen y contenido: repeticiones simples, duplicaciones segmentadas, pseudogenes, grupos de repeticiones en tándem (como centrómeros, telómeros, brazos cortos de cromosomas acrocéntricos y grupos de genes ribosomales) y, repeticiones intercaladas o derivadas de transposones.

Repeticiones simples

La repeticiones de secuencia simple (Simple sequence repeats, SSRs), se describieron por su utilidad como marcadores en la confección de mapas del genoma. Repeticiones perfectas o ligeramente imperfectas de un módulo (k-mer)

de 1-13 bases tal como (A)n, (CA)n o (CGG)n se denominan microsatélites y con módulos de 14-500 bases se denominan minisatélites.

Las SSRs surgirían de apareamientos desiguales durante la recombinación de fragmentos en la meiosis. Con algunas excepciones, como las repeticiones del hexanucleótido (TTAGGG)n de los telómeros, las repeticiones en tándem son variables entre individuos, por lo que se estima que no están bajo una fuerte presión selectiva.

La sumatoria de bases de todas las SSRs reconocidas alcanzaría para cubrir el 3 % del genoma humano, el 0.5% cubierto por repeticiones de un dinucleótido. Entre estas últimas las SSRs con los dinucleótidos AC y AT representan 50% y 35% respectivamente en tanto que repeticiones del dímero GC están apenas representadas (0.1%). Estos valores indican que en promedio existiría una repetición cada 2 kb. El número de repeticiones no superpuestas asciende a más de 400 por Mb (Lander *et al.* 2001).

Duplicaciones segmentadas

Se denomina así a copias de bloques de 1-200-kb presentes en dos o más regiones diferentes del genoma. El origen de estas duplicaciones no es claro, no pueden suponerse originadas en eventos de apareamiento desigual, no se encuentran restos de secuencias flanqueantes asimilables a sistemas de inserción y su similitud es muy elevada, explicando probablemente un origen reciente.

Las duplicaciones segmentadas se han agrupado en dos categorías según las copias sean intracromosómicas o se encuentren en cromosomas no homólogos. Cada una de ellas presenta particularidades de gran interés.

Las duplicaciones intracromosómicas que conservan un elevado grado de similitud pueden llevar a recombinación entre fragmentos parálogos[12] y pérdida de segmentos cromosómicos. En el cromosoma 17 por ejemplo hay tres copias de una repetición de 200 kb separadas por 5 Mb y dos copias de un fragmento de 24 kb separadas por 1.5 Mb. Estas copias son 99% idénticas y la pérdida de segmentos cromosómicos tras la recombinación da lugar en un caso al síndrome de Smith-Magenis y en el otro al síndrome de Charcot-Marie-Tooth (Lander *et al.* 2001). Un ejemplo semejante aparece con duplicaciones en la región AZF-C del cromosoma Y.

12 Siguiendo la denominación homólogos, atribuida a los dos cromosomas o las dos copias de un mismo gen presentes en un organismo diploide (derivadas de ambos progenitores), se denomina genes (o fragmentos) parálogos a dos copias semejantes presentes en un mismo organismo, y que puede suponerse tienen un origen común. Dos genes codificantes parálogos dan por lo general origen a dos proteínas con un alto grado de semejanza, que asumen funciones muy parecidas, o la misma función en tejidos diferentes. En el mismo sentido dos genes son ortólogos cuando puede trazarse su origen a través de diversos grupos, como en la comparación de genes entre la mosca y el hombre.

Las duplicaciones en cromosomas no homólogos se hallan por lo general en regiones teloméricas y pericentroméricas. Copias de un segmento de 9.5 kb del locus de la adrenoleucodistrofia situado en Xq28 se hallan en regiones pericentroméricas de los cromosomas 2, 10, 16 y 22 (Horvath *et al.* 2000).

Hasta completar la secuenciación del genoma humano la identificación de duplicaciones se basó en el hallazgo fortuito de sondas de ADN que hibridan en distintos sitios del mismo, o distintos cromosomas y la asociación de estos datos con sitios de ruptura de alguno de esos cromosomas. Disponer de la secuencia de todo el genoma permite explorarlo y desarrollar métodos más rápidos y/o certeros de búsqueda de causas de enfermedad aunque debería ponerse mucho cuidado en confirmar que estas suposiciones no provienen de errores en el ensamblado. El gran tamaño y similitud de las duplicaciones han hecho de estas repeticiones la mayor dificultad para ensamblar clones y obtener la secuencia completa del genoma humano.

El análisis de los primeros dos cromosomas completamente secuenciados, el 21 y el 22, muestra que la mayor parte de las regiones próximas de los centrómeros consisten casi exclusivamente de duplicaciones intercromosómicas. Un fragmento de 1.5 Mb adyacente del centrómero 22 contiene 90% de duplicaciones. En el mismo cromosoma, el 52% de las duplicaciones con origen en otro cromosoma se alojan cerca del centrómero, equivalente a tan sólo el 5 % del cromosoma. El resto de los cromosomas parece presentar rasgos semejantes (Lander *et al.* 2001).

La búsqueda de segmentos de más de 1 kb de extensión y más de 90 % de identidad en la secuencia terminada por el consorcio internacional reveló que un 5.3% del genoma está cubierto por duplicaciones segmentadas, el 45% de ellas corresponde a duplicaciones intercromosómicas (Consortium 2004). Las duplicaciones consisten en término medio de segmentos de 10-50 kb con una homología superior al 96 % (Lander et al. 2001).

La distribución de duplicaciones segmentadas y la proporción de duplicaciones intra e intercromosómicas varía a lo largo del genoma. El cromosoma Y representa un caso extremo con duplicaciones que cubren un 25 % de su secuencia, incluyendo un bloque de 1,45 Mb con 99,97 % de identidad (Skaletsky *et al.* 2003).

La mayoría de las duplicaciones en el hombre tiene un tamaño mayor a 10 kb. Este rasgo resulta típico del genoma humano cuando se lo compara con el de la mosca o el nematodo. La proporción de duplicaciones idénticas en el genoma humano sería asimismo mayor que en el del ratón o la rata. Estos genomas han sido secuenciados por el método shotgun y no han sido completados por lo cual estas ideas deben ser confirmadas.

Elementos transponibles (ETs)

Las repeticiones intercaladas, o ETs, constituyen casi el 45% del genoma humano (Lander *et al.* 2001; Consortium 2004; Pei *et al.* 2012). Se estima que gran parte del ADN restante, de secuencia única, podría tener origen en estos elementos.

Los ETs en los genomas de mamíferos aparecen bajo cuatro formas. Tres de ellos transponen a través de intermediarios de ARN (transposones de clase I) y son 1) los elementos nucleares intercalados largos o LINEs (long interspersed nuclear elements), 2) los elementos nucleares intercalados cortos, o SINEs (short interspersed nuclear elements), y 3) las repeticiones terminales largas, LTR (long terminal repeat), o retroposones. El otro tipo de elemento transpone mediante copia directa de su ADN (transposones de clase II), son 4) los transposones de ADN. Los cuatro tipos de ETs se distinguen por las características de sus secuencias (Figura 29).

Las secuencias LINE tienen aproximadamente 6 kb de extensión, tienen un promotor de polimerasa II y dos secuencias codificantes (ORFs, open Reading frame). Traducido el mensajero, éste, junto a las proteínas recientemente sintetizadas son transportados al núcleo. La endonucleasa traducida desde en ORF2 corta allí una hebra del ADN receptor. En el sitio clivado se yuxtapone la cola poli-A del mensajero y la transcriptasa reversa emplea el fragmento complementario como cebador para sintetizar una copia de ADN del ET desde el extremo 3'. La transcripción reversa no resulta en general eficaz por lo que se reinsertan en el ADN huésped muchos fragmentos cortos, con un promedio de 900 pb en las LINE1. Los nuevos sitios de inserción están flanqueados por duplicaciones de 7-20 pb del sitio blanco. Las enzimas provenientes de transcriptos LINE serían las mayores responsables de la transcripción reversa en el genoma, incluyendo la de fragmentos SINE, no autónomos.

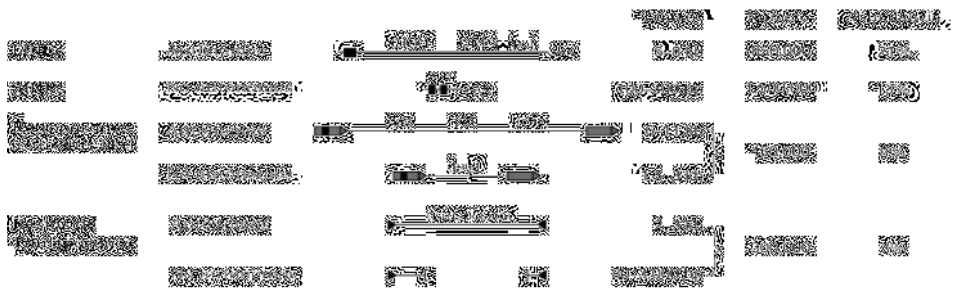


Figura 29. Elementos transponibles hallados en el genoma humano. En el hombre, como en otros mamíferos, se hallan 6 clases de elementos transponibles, tres de ellos autónomos. De su estructura puede deducirse que derivan de tres grupos básicos de elementos, dos derivan de LINE y Retroposones, y pertenecen a la clase I de ETs, el restante comprende los transposones de ADN, miembros de la clase II. SINEs, retroposones no autónomos y restos de transposones no autónomos representan a su vez vestigios de los elementos replicantes autónomos. Los genes hallados en estos elementos codifican secuencias necesarias para transcripción reversa y la integración del elemento. Los genes entre paréntesis representan vestigios de los mismos.

En el genoma humano se encuentran 900.000 copias de LINE, equivalente a un 20% de secuencia. Entre ellas se encuentran tres tipos de secuencias: LINE1, LINE2 y LINE3, de las cuales sólo los LINE1 están activos (Chinwalla *et al.* 2002).

Los ET de tipo SINE son fragmentos de 100-400 pb, con un promotor de polimerasa III que no contienen genes codificantes. Su transposición no sólo depende de las enzimas codificadas por elementos LINE, sino que comparten, en general, su extremo 3' con un LINE (Okada *et al.* 1997). El promotor de casi todos estos elementos deriva de secuencias de ARNt. En el hombre existen tres familias de elementos SINE: Alu, MIR y Ther2/MIR3. Las repeticiones Alu son los únicos SINEs activos. Estos no comparten su extremo 3' con LINEs y tienen un promotor derivado del componente 7SL de la partícula de reconocimiento de señal. Su presencia en el genoma humano asciende a más de 1.100.000 copias, el 70% del total de SINEs. La extensión de secuencia ocupada por las SINES representaría aproximadamente un 13 % del genoma.

Los retrotransposones, retroposones o retrovirus endógenos tienen largas repeticiones terminales (LTR, long terminal repeats) directas en las cuales se hallan reguladores de la transcripción. Existen elementos autónomos con secuencias codificantes gag y pol, que permiten la síntesis de una proteasa, una transcriptasa reversa, una ARNasa H (que degrada el ARN en el híbrido ARN-ADN) y una integrasa (Malik *et al.* 2000). Los retrovirus parecen tener origen en estos elementos móviles tras adquirir genes que dirigen la síntesis de una cápside o cubierta (env, por envelope) (Malik *et al.* 2000). A diferencia del mecanismo de retrotranscripción de LINEs, el ADN de retroposones es, como el de retrovirus exógenos, sintetizado en el citoplasma a partir de un ARNt cebador. En el genoma humano se han contabilizado más de 440.000 copias (8% del genoma) de retroposones entre los que se reconocen 4 tipos. Los ERVs o retrovirus endógenos específicos de vertebrados (clases I-III) y un pequeño grupo específico de mamíferos, MaLR (Mammalian-Apparent Long-Terminal Repeat Retrotransposon). El 85 % elementos LTR carece de una repetición terminal .

Los transposones de ADN hallados en mamíferos semejan a los transposones bacterianos. Estos tienen repeticiones terminales invertidas y codifican para una transposasa que se une al ADN en sitios próximos a estas repeticiones para cortar el elemento y pegarlo en otro sitio en el genoma. En el hombre hay al menos 300.000 (3% del total secuenciado) de estos elementos móviles entre los cuales se reconocen 7 tipos diferentes (RepBase, http://www.girinst.org/server/repbase.html).

ETs en la evolución del genoma

Los elementos transponibles son tan frecuentes y sus actividades están tan bien caracterizadas que es posible deducir el origen de uno a partir de otro, e inferir cambios que representan el paso de cientos de millones de años dentro de nuestros genomas (Smit 1996). La comparación entre elementos de un mismo, o distintos genomas será cada vez más factible gracias al secuenciado de decenas de otros organismos.

Las familias LINE1 y Alu representan el 60% de ETs en el genoma humano pero este efecto no se observa en los genomas de *D. melanogaster*, de *C. elegans* ni en el de *Arabidopsis thaliana* (Chinwalla *et al.* 2002). El genoma de estos tres organismos no tiene de hecho un grupo de ETs dominante y todos ellos tienen gran cantidad de familias de transposones. Esta diferencia podría explicarse por el mecanismo de transposición de los elementos frecuentes en el hombre y los otros organismos. LINEs y SINEs aseguran la transposición del elemento activo, dado que el mensajero y las proteínas codificadas por el mismo vuelven al núcleo donde se realiza la inserción. En los transposones por el contrario se sintetiza una proteína que debe volver al núcleo y transponer el elemento pero este mecanismo no garantiza la reinserción de un elemento activo. En tanto que el mecanismo de LINEs y SINEs garantiza su transmisión vertical, los transposones de ADN deben contar con mecanismos de transferencia horizontal para garantizar su existencia. El hombre, como otros mamíferos, contaría con mecanismos más desarrollados de defensa contra vectores, como virus, y estaría menos expuesto a eventos de transferencia horizontal.

Las secuencias de ETs del hombre y el ratón surgieren que el genoma del segundo cambia a un ritmo más acelerado, con un nivel de sustitución 1,7 veces mayor. Las secuencias LINE2 y MIR habrían llegado al genoma común antes de la radiación pero en el genoma del ratón son escasos. Asimismo, se estima que muchos transposones y retrovirus endógenos estarían aún activos en el genoma murino.

El contraste entre el contenido de elementos móviles en los genomas humano y murino sugiere estos podrían tener relación con las grandes diferencias que existen entre ambos (Casavant *et al.* 2000).

El estudio de la distribución de ETs podría dar una idea del carácter deletéreo o la ventaja que puede representar la inserción de los mismos en el genoma. Algunas regiones del genoma tienen una densidad de ETs muy grande. La mayor detectada es una región de 525 kb ene Xp11 con 89 % de la superficie cubierta por secuencias de ETs. Contrariamente algunas regiones están virtualmente desprovistas de estas secuencias, probablemente indicando las consecuencias nefastas de su inserción. Los grupos de genes homeóticos HOXA, HOXB, HOXC y HOXD (Figura 30) ejemplifican este hallazgo.

Figura 30. Densidad de ETs en un segmento del cromosoma 22 humano. Las barras encima de la línea horizontal representan genes codificantes activos en el hombre. Las barras debajo de la línea horizontal representan repeticiones intercaladas. Note la densidad de estos elementos a lo largo del cromosoma y la ausencia de estas secuencias "parásitas" en el segmento ocupado por el grupo de genes HOXD.

Recombinación entre cromosomas homólogos

Poder combinar la información proporcionada por los marcadores físicos y genéticos permite ver la cantidad de recombinación de la información genética que puede existir en el hombre (Yu *et al.* 2001)

La integración de mapas físico y genético del genoma humano incluyó 5.282 loci polimórficos cuyas posiciones se conocen en megabases (físico) y centimorgans (genético). En la Figura 26 se detalla la combinación de ambas medidas a lo largo del cromosoma 12 (Consortium 2004). La pendiente entre puntos deja ver la distancia genética en cM aproximada para cada distancia recorrida desde el centrómero hacia ambos telómeros. Resulta llamativo que no existe recombinación a nivel del centrómero. Luego, aunque existen sitios adyacentes con escasos eventos de intercambio, el porcentaje de recombinación aumenta consecuentemento con la distancia desde el centrómero. Los valores obtenidos indican que en promedio el brazo corto sufre más eventos de recombinación que el brazo largo, aproximadamente 2cM por Mb para el primero y 1 cM por Mb para el siguiente.

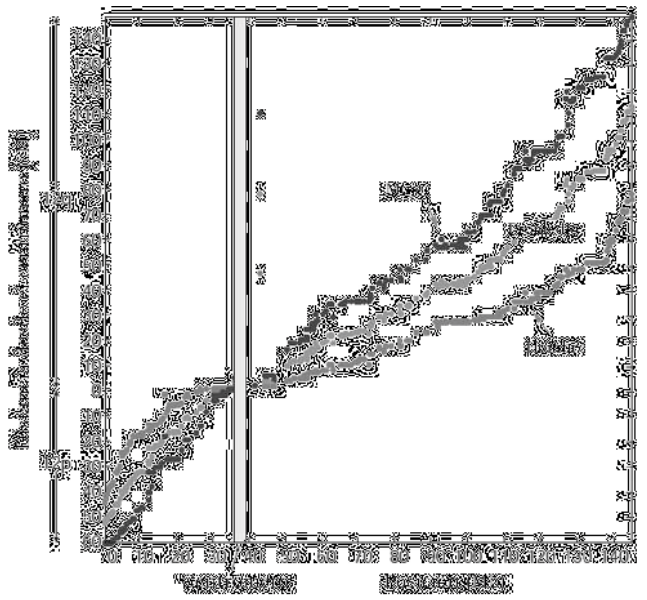


Figura 31. Relación entre distancias físicas (Mb) y genéticas (cM) a lo largo del cromosoma 12. Las frecuencias de recombinación son sensiblemente superiores en los extremos delcromosoma. En la meiosis femenina el intercambio de información es más pronunciado que en la masculinas.

En la Figura 31 se detallan valores de recombinación medidos en hombres y mujeres y resulta claro que existe un mayor porcentaje de intercambio entre cromosomas femeninos. Los resultados obtenidos en este cromosoma son representativos del resto del complemento.

Capítulo 5

Herencia del proyecto genoma

Emprender el secuenciado del genoma humano fue una decisión política sumamente importante. Como otras decisiones de esta clase, tuvo la intención de promover la mejora, impulsar el desarrollo, e inevitablemente, debió alimentar la crítica. El resultado del proyecto genoma humano no se limita al uso de la información obtenida y la tecnología desarrollada en ciencia, este influye muchos aspectos de la sociedad. Como en todo gran proyecto pueden verse lados buenos y otros no tanto.

El PGH ha cumplido gran parte de sus objetivos y marcado el comienzo de una verdadera *revolución genómica*. Los resultados científicos y las posibilidades a venir son sin duda invalorables. En otros aspectos aún no puede apreciarse el verdadero valor del emprendimiento, aunque el capital destinado estos últimos años deja suponer grandes avances.

MEJORAS Y MÁS MEJORAS

Uno de los más grandes logros del PGH es el impulso que dio a la obtención de nuevas tecnologías. Aunque muchos han cuestionado el enorme gasto que implicó el proyecto Genoma Humano, ese mismo gasto se habría recuperado –al menos para los Estados Unidos- y multiplicado 10 veces en 2010 con la sola creación de nuevos métodos de adquisición de secuencias.

Cada uno de los centros en los que se llevó adelante la secuenciación dentro del proyecto genoma humano, luchó durante una década con geles de poliacrilamida. Gran parte del trabajo de confección de geles pasaría muy pronto a manos de industrias privadas para optimizar el tiempo de trabajo y asegurar la calidad. Estas mejoras harían crecer la capacidad de trabajo pero la adquisición de información no podría superar el límite de 400 bases por corrida.

La terminación de cadena con ddNTPs marcados con fluorescencia (Dye-terminators) (Smith *et al.* 1986) que posibilitara el PGH, es aún la base del

método más empleado aunque los procedimientos generales han variado progresivamente. La mejora del procedimiento de secuenciado fue y sigue siendo objeto de interés.

El secuenciado capilar

La electroforesis capilar se desarrolló hace algo más de 30 años como método alternativo a la electroforesis en placa (Jorgenson & Lukacs 1981). El polímero que sirve a la separación de fragmentos de ADN está encerrado en un capilar de vidrio de apenas 50-100 μm de diámetro y la migración tiene lugar gracias a un potencial de 300-400 V/cm. Estas condiciones proporcionaban un gran incremento en la capacidad de separación.

En 1989 aparece el primer instrumento comercial basado en una columna de características semejantes a la descrita, dotado de un detector UV/Visible, de un sistema de inyección automática y de un sistema computarizado para el análisis de datos. El capilar de un metro permitía realizar una corrida en un tiempo 5 veces inferior a un gel en placa y se estimaba podía alcanzar una resolución de 6.000 bases (Swerdlow & Gesteland 1990). Tal resolución no se alcanzó con frecuencia. La columna debía ser reemplazada tras cada corrida y no se podía manejar un gran número de muestras.

Se ensayarían nuevos materiales para la confección de los capilares (dióxido de silicio) y se revestirían los mismos, la poliacrilamida daría su lugar a otros polímeros derivados de la celulosa u óxidos polietileno y se ajustarían las condiciones de migración (Karger & Guttman 2009). Ligando nuevos colorantes a los nucleótidos y empleando nuevos algoritmos que permitían una mejor asignación de bases, la lectura de más de 1.000 bases estaba garantizada.

Muy pronto se adaptó el sistema para la excitación-detección y el monitoreo simultáneo de múltiples capilares (Huang *et al.* 1992). Una cámara adaptada podría recoger información del paso de >1 base/segundo/capilar. En 1998 un equipo comercial estaría preparado para inyectar nuevo polímero en el capilar, cargar la reacción a secuenciar, correr la muestra bajo una corriente constante y detectar el paso de moléculas fluorescentes. La señal podría entonces interpretarse como una secuencia de 500–1000 bases de gran calidad.

El aparato empleado por Celera durante el secuenciado del genoma humano puede recoger y analizar 96 muestras menos de 4 horas. Otros sistemas desarrollados incrementan la generación de datos aún más, pues permiten obtener 1024 secuencias al mismo tiempo.

El desarrollo del secuenciado capilar de alto rendimiento dependería del trabajo de laboratorios públicos y privados, quienes hacen grandes esfuerzos por

ocupar el mercado.[13] Pero el secuenciado de alto rendimiento ya no estaría ligado tan estrechamente al método tradicional de Sanger.

Secuenciado masivo

El nombre de Next Generation sonaba hace algunos años como ficción. Clasificar un producto como primera, segunda o enésima generación indica que éste incorpora mejoras progresivamente. Correcciones de errores respecto de una versión anterior, incorporación de tecnología que lo hace más eficaz –raramente más económico- con el uso de reactivos, e incluso menos contaminante.

La fórmula *next generation sequencing*, o "el secuenciado que viene" (o que se viene) comenzó a emplearse de manera corriente para designar productos en fase experimental.[14] Las empresas alimentan en ocasiones nuestra imaginación anunciando el inminente lanzamiento de algún producto desde su etapa embrionaria, y preparar así el mercado. En este caso particular, hablar del secuenciado que vendrá, tenía este objetivo, y el de atraer inversiones para finalizar el desarrollo de productos de costo muy elevado.

Durante los 15 años que duró el proyecto genoma llevado adelante por el consorcio público hubo muchos grupos interesados en desarrollar tecnología capaz de aliviar el trabajo, tanto en los Estados Unidos como en otros países. Como podrá verse no hubo un nuevo método sino varios. Estos tienen una característica fundamental, pueden obtener la secuencia de un número muy elevado de fragmentos simultáneamente, es decir que son capaces del secuenciado masivo en paralelo (Massive Parallel Sequencing o MPS).

Los métodos de secuenciado masivo que llegan actualmente al público concuerdan en algunos aspectos. El procedimiento general implica fijar fragmentos de ácido nucleico a un sustrato, al cual se añade y extrae de manera cíclica los diferentes sustratos necesarios para la reacción. Durante un ciclo se recoge una imagen que indicará en qué moléculas hubo una reacción. El secuenciado de todas las moléculas inmobilizadas procede en paralelo.

Pirosecuenciación

Cada uno de los secuenciadores de segunda generación incorpora cambios que pueden trazarse por el examen de la literatura como la integración de muchos pequeños progresos. Uno de ellos sin embargo, el primero en llegar al

13 Las empresas Applied BioSystems, Molecular Dynamics, SpectraMedix, Beckman y LiCor fabrican secuenciadores capilares de distinto porte.

14 Algunos autores siguen usando el término asociándolo a la segunda generación de máquinas y procesos de secuenciado (masivo) de ADN.

mercado, incorpora una técnica sumamente original conocida como pirosecuenciación que cambia sensiblemente varios aspectos del secuenciado convencional.

La técnica que sirve de base al método de pirosecuenciado fue ideada por un laboratorio sueco y se denomina ELIDA (enzymatic luminometric inorganic pyrophosphate (PPi) detection assay) (Nyrén 1987; Nyrén *et al.* 1993; Ronaghi *et al.* 1998). Se basa en la detección del pirofosfato (PPi) liberado durante una reacción y la producción de luz por medio de una reacción enzimática. Para adaptarla al secuenciado de ADN se unieron tres procesos mediados por enzimas. Durante la secuenciación la ADN polimerasa libera pirofosfato (PPi) cada vez que añade un nucleótido a la hebra de ADNP. (1987. Este difosfato es empleado por una ATP-sulfurilasa para la generación de ATP. La producción de ATP sirve a su vez a la producción de luz visible por la enzima luciferasa (Figura 32).[15]

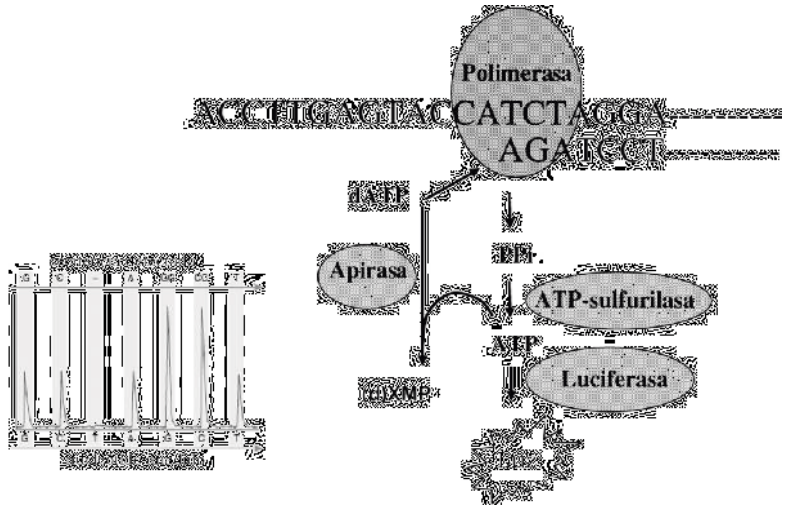


Figura 32. Pirosecuenciación. El ADN de cadena simple se hibrida con un cebador y se incuba con las enzimas ADN polimerasa, ATP sulfurilasa, luciferasa y apirasa, así como con los sustratos adenosina 5 'fosfosulfato (APS) y luciferina. La adición de uno de los cuatro dNTPs (el dATP es reemplazado por desoxiadenosina tiotrifosfato, dATPαS, para evitar que éste sea empleado por la luciferasa) permite la síntesis. Si el dNTP complementario está en la solución la ADN polimerasa puede incorporarlo y libera pirofosfato (PPi) estequiométricamente. La ATP sulfurilasa emplea este PPi para obtener ATP en presencia de adenosina 5'fosfosulfato. El ATP es a su vez empleado por la luciferasa para la conversión de luciferina en oxyluciferina que genera luz visible en cantidades proporcionales a la cantidad de ATP presente. Los nucleótidos no incorporados y el ATP no utilizado son degradados por la enzima apirasa, y la reacción puede reiniciar con el agregado de otro nucleótido. La cantidad de luz generada es proporcional a la cantidad de nucleótidos incorporados como puede verse a la izquierda.

En este procedimiento se advierten rápidamente varios cambios respecto del secuenciado convencional. No es necesario un gel para separar fragmentos ya que la lectura de secuencia procede simultáneamente con la síntesis (en directo o *real time*). De esta particularidad deriva otro nombre que se da al proceso, el de secuenciación con la síntesis (*sequencing by synthesis*). La pirosecuenciación

15 La pirosecuenciación es propiedad original de Pyrosequencing AB, empresa adquirida por Qiagen en 2008. La tecnología se licencia al grupo 454 Life Sciences, ahora propiedad de Roche Diagnostics.

resulta por esta razón más rápida que la secuenciación mediante terminadores de cadena.

La producción de luz es semejante con cada nucleótido de modo que éstos sólo pueden diferenciarse mediante el agregado de cada uno de manera independiente. Se inyecta entonces una solución conteniendo todos los reactivos necesarios y uno de los nucleótidos. Un sistema óptico capta entonces qué moléculas emiten luz. El nucleótido agregado debe ser entonces eliminado para poder agregar el siguiente. Aunque puede parecer que estas manipulaciones complican la tarea, el proceso de secuenciado requiere que una serie repetitiva de adiciones, un procedimiento fácilmente automatizable que permite la realización de muchos experimentos en paralelo.

La lectura de varias bases idénticas consecutivas se argumentó como un límite de este método. Sin embargo la intensidad de luz emitida es proporcional a la cantidad de nucleótidos incorporados y el sistema es capaz de asignar varias "letras" idénticas sucesivas.

La empresa 454 Life Sciences adaptó el sistema descripto al secuenciado masivo en paralelo. Fragmentos de una librería de ADN se inmovilizan en micropartículas (de aproximadamente 20 μm de diámetro) que se incorporan en una gota independiente de una emulsión. Cada pequeña gota en esta solución representará un microreactor en el que se amplifica un solo fragmento de ADN mediante PCR (emPCR, *emulsion PCR*). Las partículas cargadas de clones de cada fragmento se disponen entonces en placas que contienen cientos de miles de pocillos, cada uno de ellos con capacidad para albergar una partícula.[16] A la placa cargada se añaden micropartículas (de menor tamaño) a las cuales están ligadas las enzimas necesarias para la reacción. Para que tenga lugar la reacción de secuenciación el instrumento hace fluir secuencialmente sobre la placa los cuatro nucleótidos. La reacción de la luciferasa se registra entonces en una imagen desde la cual se deducirá la incorporación de un nucleótido complementario de cada ADN molde. Tras cada registro tiene lugar un lavado para eliminar el nucleótido utilizado, y entonces se carga el siguiente (Figura 33).

La longitud de una lectura promedio con este método ronda los 300-500 nucleótidos. Un solo instrumento puede, en condiciones óptimas, generar información equivalente a 100.000.000 de pares de bases. La capacidad del sistema se demostró en 2005 con el secuenciando del genoma de *M. genitalium* en una sola corrida (Margulies *et al.* 2005). Este sistema es hoy propiedad de la empresa Roche.

16 La reacción en pocillos opacos tiene como propósito limitar la difusión de luz.

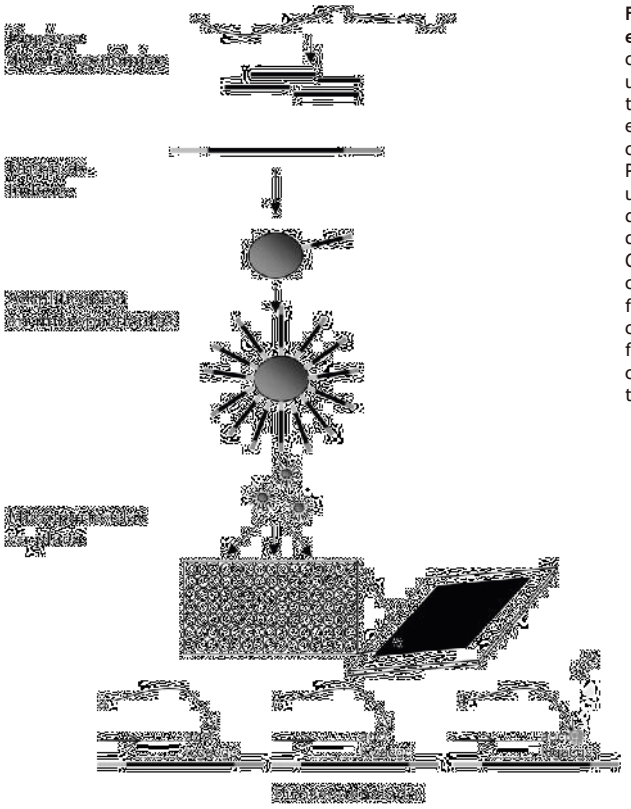


Figura 33. Secuenciado masivo con el sistema 454. El proceso comienza con la ruptura del ADN genómico y la unión covalente de fragmentos cortos de ADN que permiten la fijación en micropartículas. A este paso de clonado sigue una amplificación por PCR de cada partícula inmersa en una micela. Las partículas son luego depositadas en microplacas en las que tendrá lugar el secuenciado. Cada picopocillo contiene una micropartícula con clones de un único fragmento de ADN. Sobre las partículas inmovilizadas el sistema hará fluir alternativamente soluciones con un único dNTP y captará instantáneamente la producción de luz.

Grupos reversibles

Otro procedimiento, que tiene origen en el instituto Sanger de Secuenciación Genómica (Cambridge, Inglaterra), llegaría al mercado con un sistema sensiblemente diferente del anterior (Bentley *et al.* 2008). El secuenciado procede escencialmente mediante la terminación de cadena del método de Sanger, aunque los grupos fluorescentes acoplados a los desoxirribonucleótidos puedes removerse y dejar un extremo 3' libre para continuar la síntesis.

Los cuatro nucleótidos, marcados cada uno con un fluoróforo diferente, se encuentran entonces en la misma solución. El equipo inyecta esta solución y lee cada base incorporada mediante un sistema óptico capaz de asociar la emisión de fluorescencia con un sitio preciso en el soporte sólido (la placa). Luego lava la solución presente y remueve los fluoróforos para la incorporación de un nuevo nucleótido marcado en la cadena creciente. La lectura instantánea del nuevo nucleótido agregado permite dar a esta tecnología, como a la anterior, el nombre de *sequence by synthesis*.

En este método cada señal proviene de grupos de moléculas amplificadas *in situ* a partir de fragmentos individuales unidos de manera aleatoria a una placa.

Para lograr esto se fragmenta el ADN a secuenciar y se ligan a cada fragmento adaptadores que permitirán su unión al soporte y amplificación. Esta última tiene lugar en condiciones isotérmicas y mediante la formación de puentes (*bridges*), posibles gracias a la presencia de dos adaptadores/cebadores diferentes ligados al soporte a gran concentración (Figura 34).

El instrumento comercializado por Illumina (Solexa) logra lecturas de un tamaño promedio de 75 bases, bastante menor que el sistema 454. Los nucleótidos con terminadores reversibles y la enzima adaptada para su incorporación no permitirían lecturas más largas. La densidad de marcadores en el soporte permite sin embargo producir más de 1.000.000.000 de bases en una sola corrida.

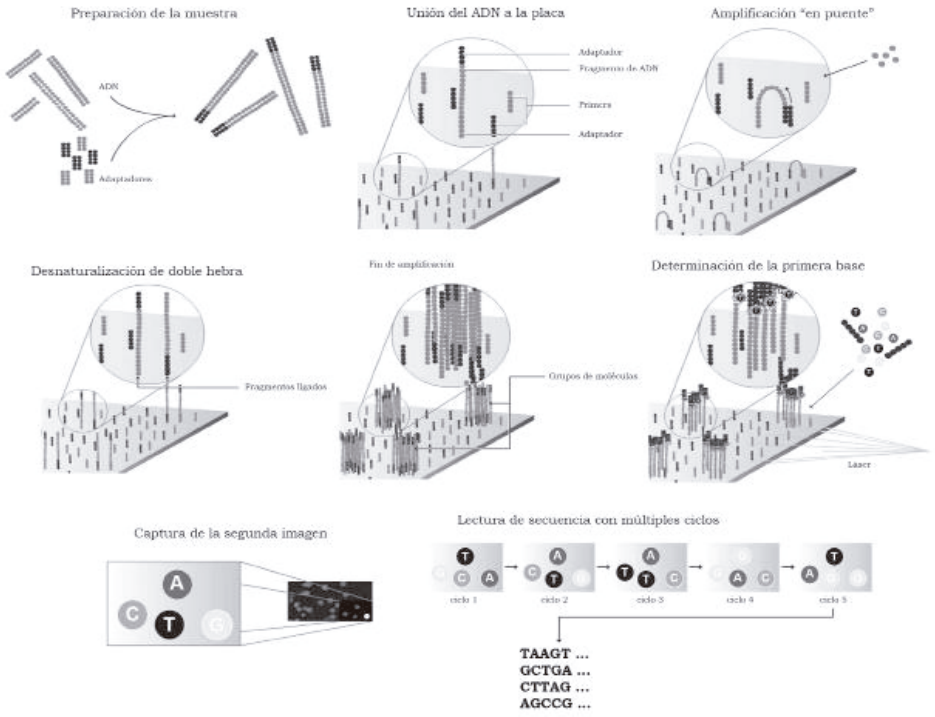


Figura 34. Amplificación en puente y detección de grupos reversibles. Fragmentos de ADN se ligan a dos adaptadores diferentes. El ADN obtenido se desnaturaliza y se hibrida con uno de los oligonucleótidos presente en el soporte sólido. Hibridados, se copian "en puente" desde el extremo 3' del oligonucleótido unido a la superficie gracias a la existencia de adaptadores complementarios. Múltiples ciclos de copiado y desnaturalización permiten poblar la placa de micropuntos en los que se agrupan clones de cada molécula original. El secuenciado puede proceder mediante la adición de las cuatro bases en simultáneo, la captura del color y la remoción del grupo coloreado. Este ciclo puede repetirse a voluntad.

Secuenciado mediado por ligasa

Un tipo muy particular de secuenciado consiste prescindir de la ADN polimerasa y en lugar de esta emplear una ADN ligasa. Siendo esta enzima

extremadamente sensible al correcto apareamiento de dos bases puede agregarse a la hebra molde un oligonucleótido que podrá ligarse con otro, si y sólo si, éste está correctamente apareado.

Este principio fue desarrollado en 1990 en Inglaterra, para el secuenciado masivo de fragmentos expresados (ADNc) por el grupo liderado por Sydney Brenner (premio Nobel de fisiología). El MPSS (massive parallel sequencing signature) desarrollado por este grupo se basa en la captura del ADN mediante micropartículas y emplea un sistema complejo de ligado de adaptadores seguido de la decodificación de la secuencia mediante la adición alternativa de fragmentos complementarios de cuatro bases (Figura 35). Brenner fundó en 1992 la compañía Lynx Therapeutics para comercializar esta tecnología que no logró salir al mercado. El principio del secuenciado masivo se empleó sin embargo en el producto (descripto más arriba) de Solexa, que fusionó con la original, y es hoy parte de Illumina.

Otro laboratorio de Harvard sí desarrolló el concepto de secuenciado mediado por una ligasa y llevó al mercado un sistema para el secuenciado masivo en paralelo. Polony sequencing conjugó la preparación de una librería mediante la adición de adaptadores, la emPCR, el secuenciado mediado por una ligasa y la lectura mediante un microscopio de fluorescencia automatizado para secuenciar el genoma de *E. coli* (Shendure *et al.* 2005). Esta tecnología se licenció a Agencourt Biosciences y luego a Applied Biosystems. Hoy es parte del sistema SOLiD (Supported Oligonucleotide Ligation and Detection system), que se lanzó al mercado en 2008.

La preparación de la muestra en el sistema SOLiD es semejante al empleado por Roche con la preparación de una librería de fragmentos y la amplificación clonal mediante emPCR sobre micropartículas. Este grupo no emplea un soporte con celdas y, dado que el tamaño de las partículas es inferior (aproximadamente 1 μm) al de Roche, puede alcanzar una densidad cientos de veces superior. Con SOLiD se podrían producir alrededor de 20.000.000.000 de bases de secuencias cortas (20-50 b) por corrida.

El procedimiento de secuenciado de este sistema consiste en exponer la simple hebra a un grupo de sondas de 8 bases con un hidroxilo libre en el extremo 3', un colorante fluorescente en 5' y un sitio de corte entre el 5to y 6to nucleótido. Las primeras dos bases del extremo 3' se investigan por su compolementariedad con la hebra a secuenciar. Las bases 3, 4 y 5 son degeneradas y así se espera que hibriden con cualquier secuencia en el molde. Las bases 6 a 8 también son variables aunque se espera sean removidas con el colorante tras la lectura. El corte tras esta lectura deja un grupo fosfato 5' para ligar el fragmento siguiente. De este modo los sitios n+1 y n+2 hibridan correctamente, como n+6 y n+7, y pueden ser asignados. La secuencia de las bases n+3, n+4 y n+5 quedará indeterminada durante esta ronda aunque se podrá leer en las siguientes.

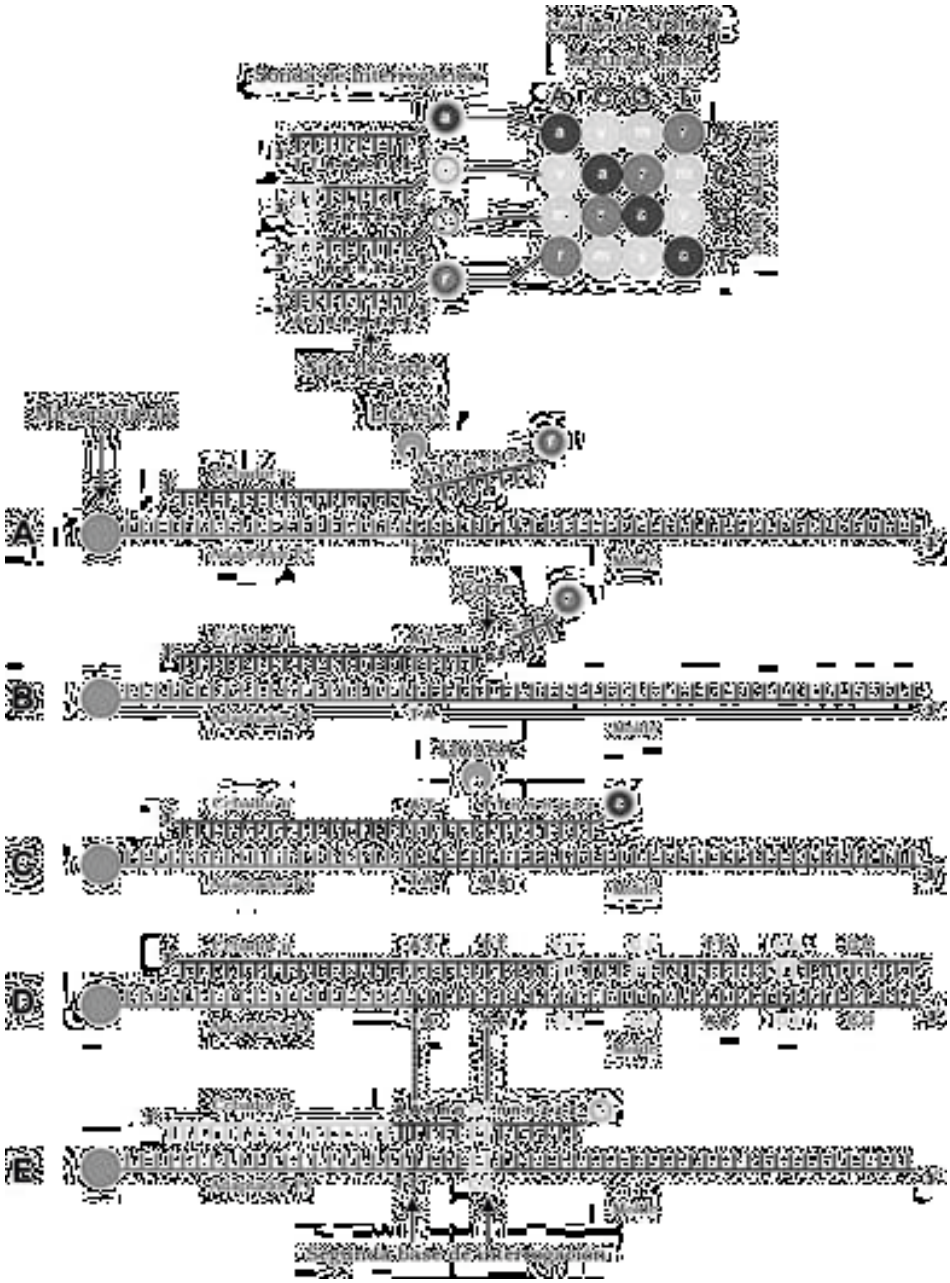


Figura 35. Secuenciado mediado por ligasa. El procedimiento consiste en interrogar un fragmento de ADN con oligonucleótidos que contienen dos bases de secuencia conocida y las restantes degeneradas (generadas por un proceso aleatorio de síntesis). Las dos primeras bases determinan el color al cual la sonda está asociada (v: verde, r: rojo, m: amarillo, a: azul). El apareamiento correcto de estas bases es vital para que la misma sea unida al fragmento en extensión por una ligasa. Las naturaleza de las bases siguientes sólo podrá revelarse con la repetición del ciclo desde un cebador más corto. Cada ciclo permite la incorporación de 6 o 7 sondas en la secuencia interrogada y el desplazamiento de bases permite completar la interrogación, y confirmar cada base hibridada y ligada.

El secuenciado completo consiste en cinco rondas, cada una comprendida por 5-7 cycles (Anon 2009). Cada ronda comienza con la adición de un cebador complementario del adaptador situado en el extremo del fragmento sustrato de secuenciación. El cebador de la primera ronda tiene un tamaño igual al adaptador, el siguiente tendrá una base menos y el siguiente una menos que el anterior. Al cabo de cinco rondas podrá conocerse, y estar seguro de ello, la secuencia de las cinco bases interrogadas (las dos conocidas en el primer ciclo y las tres desconocidas). Al cabo cinco rondas de el resultado será una secuencia de 25 bases confirmada por varias lecturas. Durante un ciclo se incorpora un octámero (8-mer) y se liga de acuerdo a sus bases 1 y 2. La sonda no hibridada se lava y puede medirse la fluorescencia de las sondas que hibridaron. Estas son entonces clivadas entre los nucleótidos 5 y 6 y puede comenzar un nuevo ciclo conteniendo sondas marcadas.

Cada sonda posee uno de cuatro colores empleados en el secuenciado convencional con terminadores. Sin embargo, siendo 16 las posibilidades de combinar las cuatro bases en dímeros la interpretación final depende de la incorporación de todo ellos. La base incorporada en la última es la única base conocida, la última posición del adaptador. Entonces conociendo el color de ésta, se podrá determinar cuál es la adyacente (con su propio color) y así con las bases subsecuentes. La lectura repetida permite minimizar los errores.

Secuenciado con semiconductores

Durante la incorporación de un desoxirribonucleótido en la cadena naciente de ADN se libera pirofosfato e hidrógeno. El método de secuenciado desarrollado por Ion Torrent Systems[17] (licenciado por DNAelectronics), se basa en un sistema de detección extremadamente sensible, capaz de captar la liberación de un ión hidrógeno. (Pennisi 2010; Rusk 2011).

Tal como ocurre en los métodos de secuenciado masivo ya descriptos, la síntesis avanza paso a paso, mediante la adición cíclica de soluciones con sólo uno de los cuatro desoxirribonucleótidos al soporte sobre el cual se encuentran adheridas las moléculas blanco. A diferencia de estos sistemas, el de Ion Torrent no requiere que el dNTP sea un terminador de cadena o que esté marcado con fluorescencia. Si el nucleótido es complementario, la polimerasa lo incorpora a la hebra y se produce un ión hidrógeno. La liberación de este ión hidrógeno es captada por un transistor extremadamente sensible (ISFET, ion-sensitive field-effect transistor) y transmitida al circuito, tal como en un aparato empleado para medir el pH de una solución. Secuenciado y lectura son, con más razón que en otros ya descriptos, simultáneos. La detección mediada por un semiconductor permite

17 Tras del lanzamiento del secuenciador, Ion Torrent Systems fue adquirido por Life Technologies.

prescindir de un complejo (y caro) sistema óptico, así como de un sistema de emisión-detección e interpretación de señales (Figura 36).

De manera semejante a la señal de luz obtenida durante la pirosecuenciación, el aparato presenta cierta limitación en la asignación correcta del número de bases cuando se halla frente a un homopolímero. Si se incorpora repetidas veces el mismo dNTP la señal debe representar la liberación de una cantidad equivalente de iones hidrógeno. Es así que resultaría difícil discernir claramente más allá de 5 a 6 bases iguales apiladas.

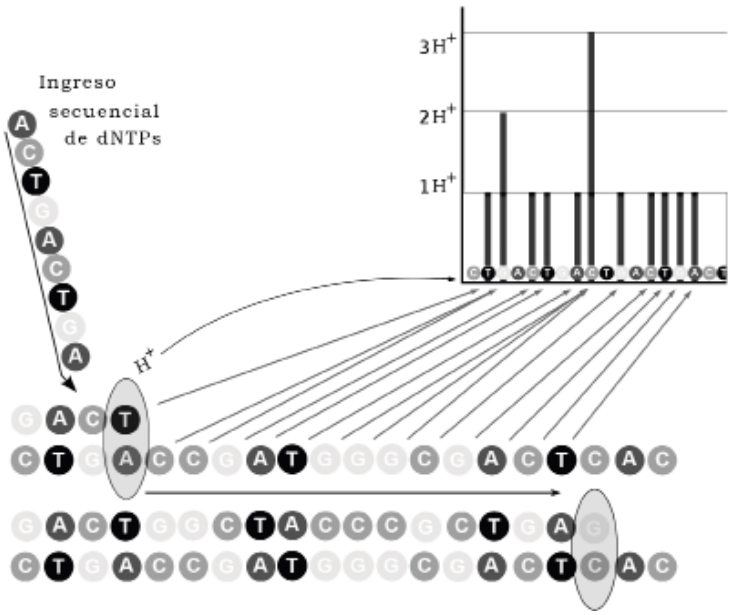


Figura 36. Principio del secuenciado mediante detección de iones liberados durante la síntesis. En este método cada dNTP es añadido independientemente según un ciclo predefinido y la liberación de un protón indica su incorporación. La presencia de iones hidrógeno por un soporte altamente sensible es asociada a la utilización de una o más bases consecutivas (arriba derecha).

Otra diferencia radical con los sistemas anteriores consiste en la preparación de la muestra para secuenciar dado que no requiere una amplificación de la molécula blanco. Cada lectura representa la síntesis de la hebra complementaria de una única molécula. Esto es posible gracias a la gran sensibilidad del sistema de detección, que permite la identificación de tan sólo un ión hidrógeno. El soporte del secuenciado en el cual se fija el ADN, contiene pocillos nanoscópicos —aproximadamente 400 en el ancho de un cabello humano. Estos pocillos asientan sobre una capa sensible situada encima de otra que transmite la corriente eléctrica.

Las lecturas logradas con este sistema comercializado desde 2011 son cortas y el rendimiento total significativamente menor que en los descriptos. Las lecturas oscilan entre 50 y 100 bases y pueden leerse aproximadamente 100 Mb de secuencia por corrida.

Secuenciado directo sin amplificación

Agregar una a una las bases que formarán parte de la cadena naciente de ADN impide a la enzima polimerasa leer de manera ininterrumpida la hebra que debe secuenciar. Por más rápido que sea el instrumento este proceso no puede acelerarse más allá del límite del mecanismo de inyección y lavado. Algunos grupos se plantearon la posibilidad de crear un instrumento capaz de permitir la lectura continua, es decir detectar cada adición de nucleótido al mismo tiempo que procede la enzima.

El sistema de secuenciado desarrollado por Pacific Biosciences bajo el nombre de SMRT (single-molecule, real-time) permite leer la síntesis de una única molécula de ADN en tiempo real y sin interrupciones (Eid *et al.* 2009). Cada reacción tiene lugar en celda "nanofotónica"conocida como ZMW (zero-mode waveguide) –de tan sólo 70 nm de diámetro y 100 de profundidad–, diseñada para reducir el volumen de observación y evitar la interferencia por las ondas de luz. Un microscopio confocal adaptado para esta medición puede captar la actividad de una única ADN polimerasa fijada a la pared de la celda sin interferencia desde el resto de las moléculas marcadas libres en solución.

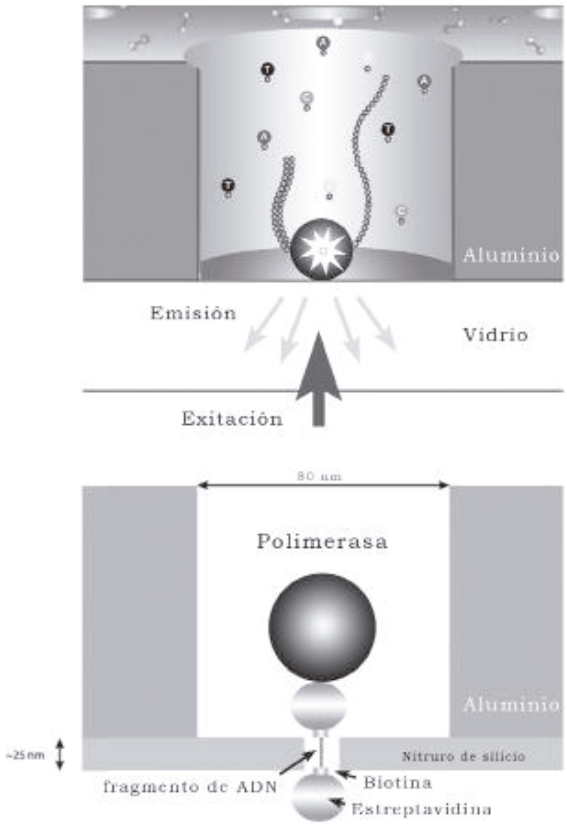


Figura 37. Secuenciado en celda nanofotónica desarrollado por Pacific Biosciences. Una única enzima polimerasa unida al fondo de un micropocillo de aproximadamente 80nm de diámetro realiza la síntesis de una única hebra de ADN. El sistema permite que la emisión y lectura de fluorescencia tenga lugar desde el fondo de cada pocillo La unión de una única enzima se ha logrado mediante un ancla de dos proteínas de muy alta afinidad (estreptavidina-biotina) y un fragmento que media su unión covalente (linker de ADN). La lectura individual y la flexibilidad del anclaje permiten medir la actividad de otras proteínas y complejos de síntesis, como el ribosoma.

De manera semejante a otros sistemas cada dNTP está ligado a un fluoróforo con un espectro de emisión característico. A diferencia de otros sistemas, el colorante fluorescente está ligado al dNTP por el fosfato terminal. Esto permite liberar el grupo exitable dejando ADN sin modificaciones (sin fluorescencia). El pulso fluorescente detectado por el instrumento se produce cuando la ADN polimerasa (modificación de la polimerasa del bacteriófago ϕ29) retiene el dNTP complementario en la región de detección de la celda ZMW y finaliza con el clivado del grupo colorante-pirofosfato (Figura 37).

El sistema elaborado por Pacific Bio es capaz de leer 3000 celdas simultáneamente, muy poco comparado con la capacidad de los secuenciadores más populares. Sin embargo, y para sorpresa de muchos cada celda puede registrar lecturas de más de 1.000 bases, un avance significativo respecto de los otros.

Numerosos artículos científicos sugieren un enorme potencial para este instrumento, no sólo en el secuenciado de ácidos nucleicos, sino de modificaciones y alteraciones del ADN y hasta síntesis de proteínas (Loomis *et al.* 2012). El sistema está presente en el mercado desde 2011 pero no se encuentran trabajos que confirmen estas mejoras.

Secuenciado a través de nanoporos

Desde el inicio del PGH y durante 15 años Applied Biosystems tuvo el monopolio del secuenciado de ácidos nucleicos. El proyecto genoma parece haber sido un gran motor para el desarrollo de métodos alternativos pues hoy, existen más de 5 instrumentos comercializados destinados al secuenciado masivo y ya no pueden contarse con ambas manos los proyectos en marcha. Muchos de ellos orientan y otros reorientan sus objetivos a medida que se dan a conocer resultados del funcionamiento de otros. El secuenciado masivo, la obtención de secuencias de mayor longitud y hasta la construcción de productos de menor costo o que prescindan del uso de reactivos costosos.

Existen aún muchos proyectos en curso, algunos de los cuales se desarrollan de manera secreta. Una estrategia sin embargo ha creado grandes expectativas, tantas que se ha llevado el 60 % de los fondos para investigación de nuevas tecnologías de secuenciado. La idea de estos métodos tiene origen en investigaciones realizadas 1996 cuando un grupo demostraba que es posible hacer pasar una hebra de ADN o ARN a través de un poro generado en una membrana mediante un campo eléctrico (Kasianowicz *et al.* 1996). Gracias a la alteración del flujo de iones ocasionado por el pasaje de la macromolécula este grupo lograba medir la longitud de los fragmentos que atraviesan el poro y sugería muy claramente que desarrollar esta estrategia podría permitir conocer la secuencia y/o características del ácido nucleico. Para esto debía lograrse un poro y un sistema de medición en el cual cada purina o pirimidina que pasa a través,

bloquea la corriente reflejando exactamente su tamaño y naturaleza química. El grupo sugería que para secuenciar de esta manera deberían reunirse otras condiciones: la apertura del canal debe reflejar la presencia de un solo nucleótido, la resolución de la medición debe ser mayor que el movimiento del nucleótido, el movimiento retrógrado debe ser mínimo y, el canal y la membrana deben soportar temperaturas o condiciones necesarias para reducir la estructura secundaria del polinucleótido.

Al menos cuatro grupos intentaron desarrollar este principio. En 2012 la empresa Oxford Nanopore Technologies anunció el lanzamiento de un sistema capaz de secuenciar ADN durante su tránsito a través de un poro. Para sorpresa de muchos el instrumento no sería costoso, espacioso ni pesado. El secuenciador de ADN MiniON es una celda de apenas 150 que se vendería por menos de 900 dólares, y tiene un adaptador USB que debe conectarse a una computadora personal. Para leer una secuencia de ADN se deposita una muestra de ADN digerido en el chip, en el cual dobles hebras de longitud variable se asocian a cada poro de una matriz. Una enzima permeasa asociada a cada poro comienza entonces a disociar las hebras de la doble cadena y a pasar una de ellas a través del poro (Figura 38). La alteración de la corriente de iones que fluye a través del mismo poro se registra a través de semiconductores con capacidad para enviar 33.000 impulsos por segundo. Habiendo disociado toda la doble cadena y enviado una hebra a través del poro, la enzima podría hacer pasar la complementaria en sentido inverso, por lo que puede suponerse existe un sistema de corrección de errores en la medición.

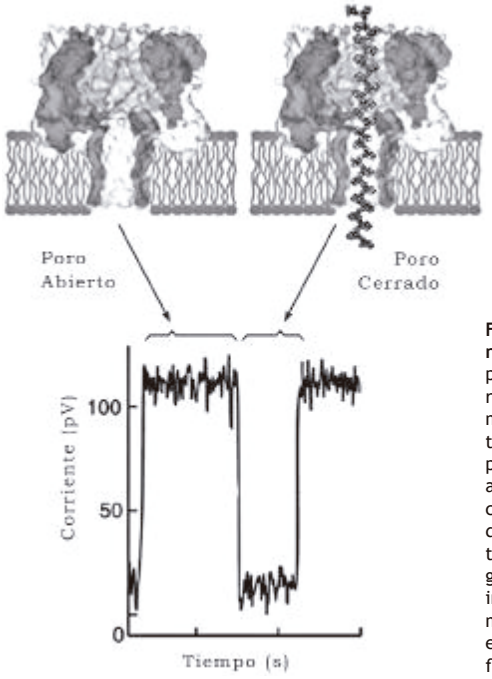


Figura 38. Modelo de secuenciado a través de nanoporos. El pasaje de un metabolito a través de un poro en una membrana puede ser medido desde la alteración del flujo de iones y corriente a través de la misma superficie. En el modelo arriba el poro está estructurado en base a una proteína inserta en una bicapa lipídica que permite el paso de distintos compuestos. La alteración de la corriente se muestra con un resultado claro entre dos estado, abierto o cerrado. Los procesos desarrollados hasta ahora se basan en el uso de distintas proteínas, membranas y metabolitos, habiendo algunos que prescinden del uso de proteínas. El sistema intenta adaptarse ala detección específica de distintos metabolitos, modificados o no, y a reconocer cada uno en base a una señal característica en el detector del flujo de corriente.

Este secuenciador de bolsillo no tiene capacidad para secuenciar un genoma humano completo, aunque podría lograrse una baja cobertura con cuatro o más de ellos. La intención de esta empresa, como la de muchas otras, no parece ya el genoma humano completo sino la secuenciación rápida (en cuestión de minutos) y segura de virus o una bacteria pequeña a partir de cantidades ínfimas de material. El mercado del diagnóstico estaría así más rápidamente al alcance de la mano.

SECUENCIADO GENÓMICO: GASTO O INVERSIÓN?

El secuenciado del genoma pudo verse, como otros proyectos científicos, egoísta. Un profesional que desea llevar adelante un proyecto personal, con un resultado incierto, a costa de un gran influjo de recursos económicos que podrían ser destinados a otros propósitos. Contrariar la justificación de tal gasto resulta en general inaudito, el científico sabe porqué y cómo debe hacerse, y en general puede estimar un beneficio para esa investigación.

El proyecto genoma nace ciertamente en el laboratorio de investigación pero la decisión de llevarlo adelante surge de un organismo (el Departamento de Energía) y se transforma en una política de estado. Resultaba difícil ver en un comienzo el gran potencial de esta empresa, pero algunos visionarios parecen no haberse equivocado. A los doce centros de investigación financiados por el NIH y el DOE en Estados Unidos, se sumaron universidades e institutos de investigación de Inglaterra, Francia, Alemania, Japón y China.

El esfuerzo requirió una suma considerable de dinero, 3.800.000.000 U$S sólo en los Estados Unidos. Esta gran inversión ha sido muy cuestionada en ese país, aunque representaría menos del 1,5 % del presupuesto del NIH en ese mismo período de tiempo. La inversión realizada ha sido mucho menos criticada en los restantes países participantes en el proyecto. Esto podría deberse a la menor inversión realizada por esos países, a la difusión limitada de información concerniente y/o a una mejor aceptación del posible retorno.

Un informe difundido en 2011 (Battelle HGP Report) intentó demostrar que la inversión en el PGH habría sido recuperada y ampliamente superada.[18] En este informe se estima que cada dólar invertido en el PGH desde 1988 se transformó en 140 dólares hasta mayo 2011. Los 796 mil millones recuperados se verían relacionados con la creación de empleos, de industrias y tecnología. Hay quienes cuestionan el procedimiento para llegar a estas estimaciones aunque el beneficio económico brindado por el proyecto es hoy indudable (Drake 2011).

18 El presidente de los Estados Unidos se sirvió de datos de este informe para lanzar un proyecto de mapeo del cerebro humano en febrero del 2013.

El interés de capitales privados da, por lo general, una buena estimación del retorno que podría tener un proyecto y este interés está directamente relacionado con la confianza de hallar un mercado. En el año 2000 una empresa monopolizaba la industria relacionada con la producción de secuenciadores y productos consumibles. Tras seis años del anuncio de la finalización inminente del proyecto habrían surgido al menos 10 empresas con proyectos originales que cambiarían el concepto de la secuenciación.

Finalizado el proyecto genoma muchos laboratorios vieron un estímulo a la innovación, la Fundación X Prize para el desarrollo de las ciencias daría 10 millones de dólares a quien demostrara ser capaz de secuenciar mucho, bien, rápido y barato. El límite impuesto fue claro: secuenciar 100 genomas diploides, en 10 días, que las lecturas obtenidas cubran más del 98 % y con una seguridad superior a 1 error cada 100.000 bases, y por un costo inferior a 10.000 dólares por genoma.

El capital necesario para el desarrollo de pequeñas empresas surgiría en gran parte de fondos públicos. Investigadores y laboratorios de este medio son incitados continuamente a levar a la práctica sus desarrollos. Los fondos llegan luego de manos privadas, pequeños fondos de inversión, financistas, bancos y grandes grupos farmacéuticos.

Secuenciar una base en el 1980 tenía un costo aproximado de 10 dólares. Diez años después, con la llegada del primer secuenciador automático, este valor bajaba a sólo 1 dólar. La caída en los valores del secuenciado siguió unos años el curso esperado por el uso cada vez más frecuente de la tecnología. Un laboratorio como Celera podía en 2001 secuenciar un genoma humano por aproximadamente 100 millones de dólares. Este costo llegaría a 1 décimo en 2008. Con la introducción de los métodos de secuenciado masivo el costo decreció a una velocidad aún mayor, pudiendo hoy hacerse por apenas 5.000 U$S. A este ritmo antes de 2015 sería posible alcanzar el objetivo fijado por la fundación Prize, un genoma por menos de 1.000 U$S.

No es difícil ver qué pasaba por la cabeza de los directores de proyectos relacionados con el secueciado del genoma humano así como de las empresas que se lanzaron en programas asociados: "porqué no pensar que cualquiera debería poder secuenciar su genoma". Los costos del proceso disminuyeron, ahora lo hacen el de los instrumentos. Para justificar todo el esfuerzo es necesario masificar su uso.

Desde 2006 gran parte de las pequeñas empresas que llevan a término sus desarrollos son "rápidamente fagocitadas" por grandes grupos económicos. Roche adquiere gran parte de 454 por 140 millones de U$S, Life Technologies hace lo propio con el desarrollo de Ion Torrent por 375 millones de U$S, e Illumina con Solexa al comienzo y hoy con Oxford Nannopore Technologies.

Estas inversiones no dan cuenta por el momento más que de una prospección, apresurarse para ganar un mercado por el momento en desarrollo. Life Technologies habría destinado 2/3 de sus fondos para investigación y desarrollo a las nuevas tecnologías para el secuenciado de ADN. Hoy, con una buena posición en el mercado y una gran deuda por detrás habría sido adquirida por Thermo Fisher Scientific (por la módica suma de 15.800 millones).

El secuenciado del genoma ha hallado muchos nichos. Prueba de esto son las intenciones de Ion Torrent de ofrecer un producto de bajo costo que pueda encontrar un espacio en cualquier laboratorio de diagnóstico. Secuenciar un virus o una bacteria pequeña en tan sólo 1 hora con un sistema económico, tomar muestras de tierra, agua u otra fuente y caracterizar, a campo, la flora medioambiental, disponer de un secuenciador del tamaño de un teléfono celular, están entre las múltiples promesas que ya garantizan el consumo masivo.

USO Y ABUSO DE LA INFORMACIÓN

El libre acceso a la información obtenida durante, y desde, el PGH ha sido, y es aún, sujeto de algunos debates. Aunque la reflexión comienza antes de los años 90, el anuncio de la lectura de la secuencia completa del genoma humano puso de relieve algunas preguntas que exigen respuestas prontas: ¿Puede alguien apropiarse de una secuencia de ADN? ¿A quién pertenecen las secuencias obtenidas a partir de seres humanos? ¿Quién tiene la capacidad para obtener la secuencia de un fragmento de ADN, puede protegerla y explotar comercialmente su uso?

Es comprensible que una empresa que basa su actividad en el desarrollo de productos biológicos y derivados, como medicamentos, enzimas recombinantes, reactivos diagnósticos basados en alteraciones en las secuencias de ADN, o compañías que buscan la mejora de semillas mediante la incorporación de genes de resistencia, deseen y puedan proteger una secuencia que sirve de base a su desarrollo tecnológico. En todo caso, sería válido, protegerla si otras compañías pueden explotarle con el mismo propósito.

Estados Unidos tiene una larga tradición en la protección de inventos o descubrimientos, entre los que se hallan productos biológicos. La composición de un material, un método para su obtención o el uso de uno o más productos combinando ambas cosas, puede ser protegido mediante una patente. El poseedor, o detentor de esta patente tiene el derecho de excluir a otros de su producción, uso, venta o importación, durante un período de 20 años desde la solicitud. Sin considerar el posible uso o proceso de producción, se pueden patentar también productos naturales "aislados" de su estado natural. La adrenalina (ya en 1906), la insulina y la vitamina B12 estuvieron protegidas, se comercializaron, y han favorecido el origen de grandes empresas. Siguiendo esta definición

se patentaron bacterias recombinantes, plantas y animales transgénicos y hasta células no modificadas aisladas a partir de un organismo (células madre, células cancerosas obtenidas de pacientes).

La industria biotecnológica debe la mayor parte de su éxito a la protección de productos y procedimientos patentados. En 1982 una patente otorgada a la empresa Lilly, protegerá la secuencia de un ADNc humano que codifica para la hormona de crecimiento. La concesión da origen a un gran litigio con otra empresa, Genentech, que también intentó protegerla. Las universidades inglesas y norteamericanas han incitado durante mucho tiempo a investigadores a proteger sus hallazgos y a crear empresas sobre esa base. Patentar y producir estos descubrimientos equivale a obtener grandes beneficios en ausencia de competencia. Ambos, universidad e investigadores obtienen grandes cantidades de dinero gracias a esto.

El uso del gen humano BRCA1, y modificaciones del mismo que llevan al desarrollo de un cáncer de mama, está limitado desde 1994 por una patente que beneficia a la universidad de Utah y a la empresa Myriad Genetics. Esta empresa ha logrado cientos de patentes que protegen secuencias aisladas de ADN, métodos para diagnosticar el cáncer, basados en modificaciones existentes en secuencias de ADN, y métodos para identificar drogas empleando ADN aislado.

Tras el secuenciado del genoma humano la empresa Celera, la misma que empleara los mapas y secuencias obtenidas por el consorcio público para ensamblar las secuencias que obtuviera mediante secuenciado aleatorio, reveló abiertamente que no pondría a disposición del público las lecturas de secuencias obtenidas. El debate abierto comenzó entonces. Celera presentó documentación para patentar más de 6.000 genes y habría obtenido al menos 500 de éstas. Gran parte de los genes y polimorfismos protegidos suponen una utilidad diagnóstica (celera.com/celera/intellectual_property).

La lucha por proteger la información no concierne sólo a la empresas, sino a cada hombre. Hoy, la protección no lleva sólo al desarrollo en "exclusividad" de una bacteria, un animal, una planta o un test diagnóstico sino al simple resguardo de la idea hasta que alguien, pagando, pueda desarrollar un producto necesario para nuestro bienestar (y su beneficio económico). Si ninguna persona o empresa manifiesta interés por esas secuencias, el detentor de la patente podrá, en veinte años, olvidarla o modificarla para que aún tenga vigencia. La protección hace posible que nadie llegue a producir un producto benéfico si no es posible o no se desea pagar a la empresa detentora. Por el contrario, la protección impone que el producto producido tenga a menudo un precio caprichoso, por la protección y la falta de competencia. Aunque la investigación biotecnológica tiene un alto costo, las empresas suelen tener márgenes de ganancias muy amplios.

Pacientes, donantes y particulares han demostrado interés en marchar contra grandes compañías en busca de beneficios por haberse explotado material

biológico obtenido a partir de ellos. Son escasos los que han obtenido lo que esperaban. Asociaciones de pacientes que buscan estrategias terapéuticas ligadas a la existencia de secuencias patentadas suelen asimismo emprender acciones en busca de su explotación libre o a bajo costo. Salvo raras excepciones, a comienzos del año 2013 es aún posible patentar genes.

Celera, como algunas otras empresas, respondieron a reclamos y pusieron a disposición del público parte de los datos recogidos durante el secuenciado del genoma humano. El acceso a sus bases de datos se limita sin embargo a un tiempo máximo de uso y de secuencias visualizadas desde un mismo ordenador o institución. Los problemas relacionados con el uso público de secuencias del genoma humano (y de otros organismos) no se limitan desde entonces a la existencia de patentes por secuencias obtenidas.

El archivo de secuencias más grande hoy existente es el Archivo de Lecturas de Secuencias (Sequence Read Archive, SRA). Este sistema, puesto en marcha por el NIH como parte de la Asociación Internacional de Archivos (INSDC) en el NCBI, el Instituto Europeo de Bioinformática y la Base de Datos de ADN de Japón, tiene la función de guardar datos de secuenciación de alto rendimiento. Aunque hoy hay secuencias alineadas con patrones de referencia diferencia, en ésta base se almacenan secuencias crudas, tal como salen de un secuenciador.

La capacidad de producir datos de los sistemas modernos de secuenciación se refleja claramente en el crecimiento constante del SRA (Figura 31). Creada a fines del 2008 hoy alberga 1.5 x 10^{15}, equivalentes a aproximadamente 600.000 genomas humanos completos. En octubre del año 2011 el NIH anunció que este archivo no podría hacer frente a la constante demanda de incorporación de datos. Desde entonces, compañías especializadas en el almacenamiento y gestión de la información "ayudan" a cubrir la demanda. DNAnexus, financiada por Google Cloud Storage, es la primera de estas compañías, y probablemente ya se sumen otras.

Capitales privados ya han tomado parte del control de la gestión de la información y también de la producción. En la Figura 39, se muestra el total de bases almacenadas en la SRA y las secuencias de acceso público, que pueden ser buscadas y analizadas por cualquier persona. A apenas dos años de la intervención de capitales privados en el almacenamiento y gestión de datos en 2011, algo más de la mitad de las secuencias pertenecen a empresas privadas. Esto no indica necesariamente que están compañías estén protegiendo con patentes cada base que obtienen. Esto sugiere que avanzan a un ritmo muy acelerado en la búsqueda de información, y hasta no confirmar que una secuencia no tiene utilidad comercial no la harán pública.

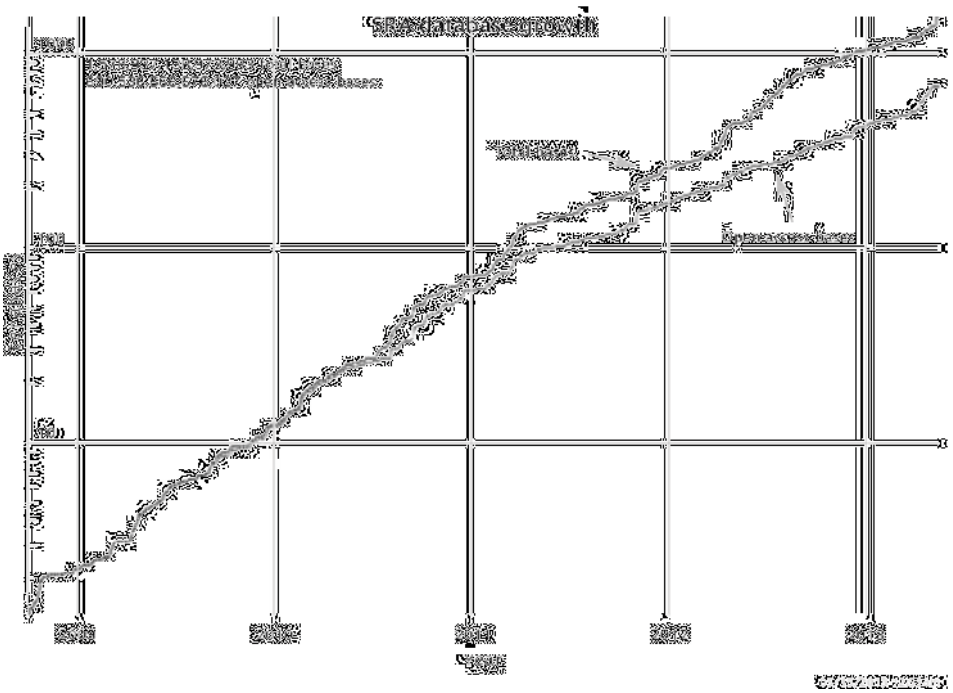


Figura 39. Crecimiento del Archivo de Lecturas de Secuencias desde el origen de la base de datos. Note el ritmo de crecimiento de los datos totales y de libre acceso (open Access). El número de bases incorporadas está expresado en terabases.

Es deseo de –casi– todos, el libre acceso a la información, pero esta libertad podría tener consecuencias no deseadas para algunas personas. Alguien que donara material genético para estudios podría ser localizado o el acceso fácil a la tecnología que se acerca podría permitir a cualquiera hacer un examen de ADN, a sí mismo o a otros. Un empleador, una compañía de seguros o una prestataria de salud, podrían acceder a la información genética y seleccionar sus empleados o clientes en base a la existencia de defectos genéticos.

Existen en varios países (no en el nuestro) leyes que tienen como propósito limitar, y hasta penar, la discriminación de una persona a causa de la información contenida en sus secuencias de ADN. En Francia y algunos países de la comunidad europea estas normas comienzan a tomar forma ya en 1991. En los Estados Unidos la protección del individuo toma algo más de tiempo. Solo en 2008 se llegará a una ley (GINA, Genetic Information Nondiscrimination Act) que prohibe a sistemas de seguro de salud y empleadores negar a una persona el acceso a una cobertura de salud o al empleo basándose "únicamente" en el uso de información genética. Como puede verse aún hay algunos detalles para corregir en ese texto.

AVANCES CIENTÍFICOS Y EN MEDICINA

Twenty-five years ago, biologists debated the value of sequencing the human genome. Today, young scientists struggle to imagine the nature of research in the antediluvian era, before the flood of genomic data.

Eric S. Lander, 2012.

La investigación genómica apenas comienza, pero avanza tan rápidamente que no es posible manejar toda la información disponible. El análisis de la literatura puede dar buenos indicios de esta explosión denominada revolución genómica. El uso de la palabra genoma ha crecido de manera lineal desde el desarrollo de las técnicas de secuenciado hasta ser empleada en 47.000 artículos científicos durante el año 2012. La frase "genoma humano" por el contrario tiene un crecimiento exponencial desde el lanzamiento de PGH y se acerca en 2012 a los 8.000 trabajos publicados. La frase "whole genome sequencing" ha sido citada menos de 200 veces por año hasta el 2007, pero desde el año siguiente, con la aparición de los secuenciadores de tercera generación, las citas se han duplicado cada año (Figura 40).

Los cambios introducidos por el secuenciado y análisis del genoma se advierten en distintas áreas. Como ocurriera con la revolución biotecnológica de los años 90' ya se percibe algún impacto en la economía, al menos en la de un país. En las ciencias como en la medicina ya se detectan algunos aunque lo mejor estaría por llegar. Aquí se destacan solo unos pocos que reciben gran atención actualmente.

En el año 2000 el número de genomas secuenciados era escaso: 38 bacterias, un hongo (*Saccharomyces cerevisiae*), dos invertebrados (*Drosophila melanogaster* y *Caenorhabditis elegans*) y una planta (*Arabidopsis thaliana*), todos con genomas relativamente simples y de tamaño reducido. Diez años más tarde esta cifra se multiplica a tal punto que disponemos de las secuencias de genomas casi completos de más de 250 eucariotas, y 4.000 bacterias y virus.

A pesar del avance notable en el número de proyectos con otros organismos, el análisis del genoma humano sigue siendo el que atrae más atención, y fondos. La genómica ha hallado nichos en muchos estudios no vinculados con ése, desde la mejora de animales y plantas hasta el estudio de comunidades de microorganismos, pero el mayor interés, y beneficio, sigue estando en el estudio del hombre.

Llegar a comprender el origen de la enfermedad e instalar herramientas diagnósticas adecuadas requiere sin embargo conocernos más. Para determinar el estado de enfermedad es necesario definir de antemano que es normal. Los estudios de variación en el hombre tienen ese objetivo.

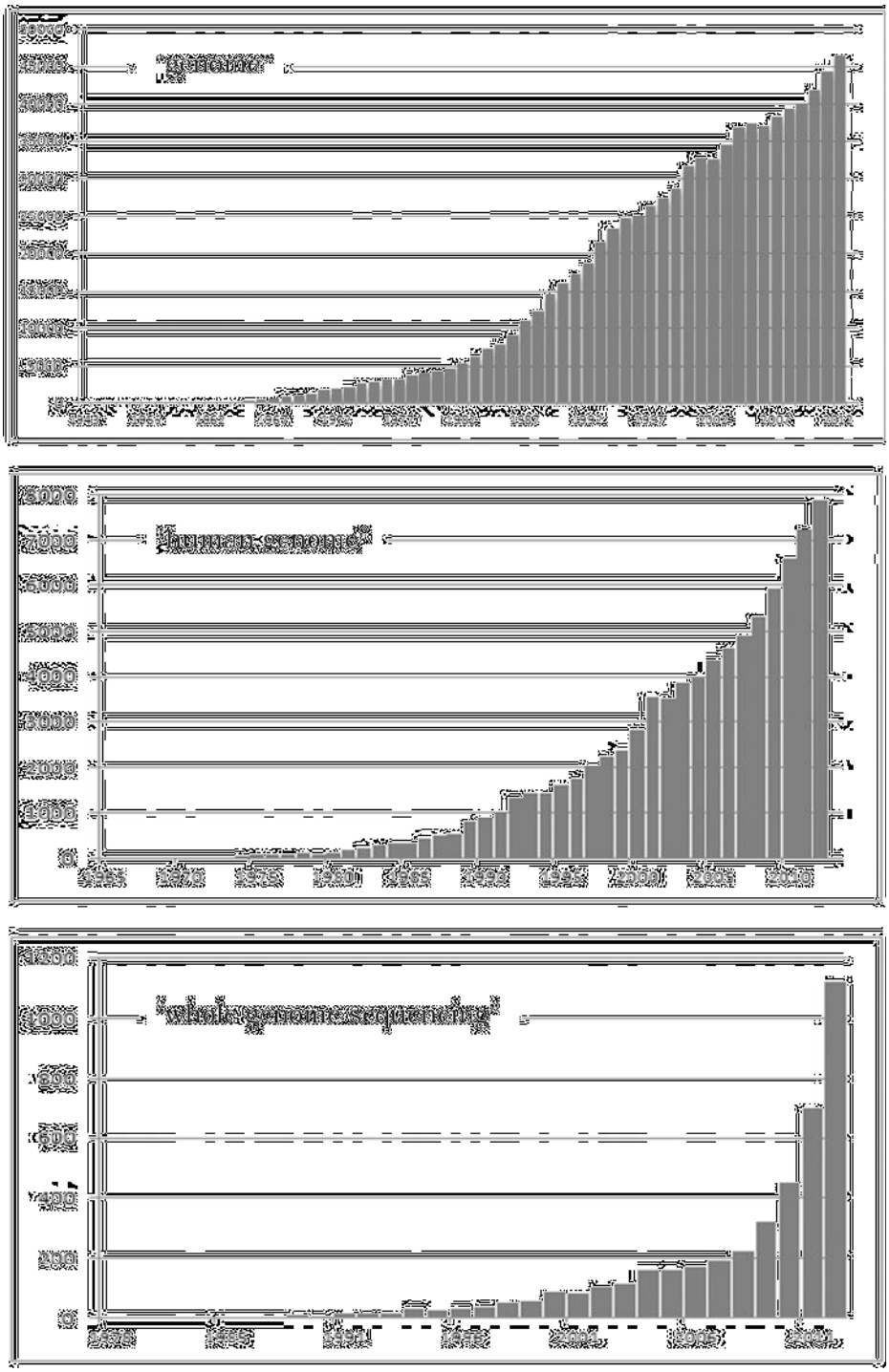


Figura 32. Uso de palabras clave en la literatura científica. El empleo de las palabras “genoma”, “genoma humano” y “secuenciado de todo el genoma” reflejan la explosión de trabajos científicos relacionados.

Variación humana

El secuenciado masivo permite resecuenciar con cierta facilidad y trabajar sobre un genoma de referencia, confirmar hallazgos obtenidos hasta el momento y buscar variaciones entre individuos. Analizar la variación entre individuos de una especie debe proporcionarnos datos muy importantes acerca del potencial de cambio dentro de la especie y principalmente, sugerirnos dónde están los límites de lo "normal".

El fin del PGH y el lanzamiento de los primeros equipos de secuenciado masivo dieron comienzo a nuevos proyectos como el 1000 Genomes Project. Este comenzó con un objetivo, secuenciar y analizar 1.000 genomas humanos de distintos puntos del globo con el propósito de determinar la variación del genoma a nivel de secuencias expresadas (exoma) y del genoma global (whole genome sequencing). Concluida una fase I de trabajo se hizo pública la comparación del genoma de 1.092 individuos (Abecasis *et al.* 2012).

El catálogo de 1.420.000 marcadores SNP logrado por el PGH se estimaba que permitiría haber descubierto todas las variantes con una frecuencia mayor al 5 %. Con el 1000Genomes se ha confirmado esta sospecha y extendido el límite a 38.000.000 de SNPs, 1.400.000 de indels bi-alélicos y 14.000 grandes deleciones. Del número de muestras, la cobertura alcanzada con estos sistemas de análisis y la comparación con datos ya disponibles, se estima que es posible detectar un SNP presente en una frecuencia de 1% con un 99,3% de seguridad en todo el genoma, y uno que esté a 0,1% de frecuencia con un 70% de certeza.

Agrupando poblaciones según el componente predominante de ascendencia: cinco comunidades de Europa, tres grupos de África, tres de Asia del Este y tres de América (ninguno de América del Sur) se encuentra que variantes presentes con frecuencias >10 % se encuentran en casi todas las todas las poblaciones estudiadas. Por el contrario, el 17% de las variantes de baja frecuencia en el rango de 0,5-5% se observan en un solo grupo ascendencia, y el 53% de las variantes raras, menos de 0,5%, se observan en una misma población.

Antes del año 2000 se pensaba que un único evento de migración ocurrido hace 50.000-100.000 años desde África, llevaba a la sustitución progresiva de formas humanas arcaicas en Europa y Asia. Este reemplazo, y el origen de las poblaciones modernas, habría ocurrido sin mezcla evidente en la actualidad. El análisis genómico ha cambiado radicalmente nuestra comprensión de estos fenómenos, dando una imagen más rica de la mezcla de la población y la selección natural.

En la actualidad existe evidencia que sugiere que las poblaciones del sur de Asia son el producto de repetidas mezclas entre antiguas poblaciones europeas y vecinos y sus genes reflejan la geografía más bien que caminos de migraciones humanas o familias lingüísticas. El análisis del genoma mostró europeos y

asiáticos, y no así africanos, han heredado 1-4 % de su genoma de neandertales, indicando el flujo de genes desde África hacia Oriente Medio (Green *et al.* 2010).

Los proyectos que intentan analizar la diversidad humana llevan con frecuencia a ver los aspectos más polémicos del análisis del hombre como especie. Aunque algunos científicos creen que los datos de diferentes poblaciones no revelan más que la unidad de la raza humana, datos actuales ya demuestran que los patrones de variabilidad genética reflejan la existencia de grupos geográficos. Es de esperar que estos resultados no estimulen el debate en círculos no científicos.

Es difícil determinar el efecto de la selección en el hombre, aunque el gran número de secuencias disponibles ya habría hecho posible estimar algún grado de selección positiva. Desconociendo qué rasgo está bajo selección, es posible determinar la existencia de cierta selección positiva por la existencia de una alta frecuencia de un haplotipo de largo alcance. Los haplotipos son variantes genéticas estrechamente correlacionadas, que se reflejan en un desequilibrio de ligamiento y están separadas por puntos calientes (más frecuentes que lo esperado) de recombinación. Cuando un gen del grupo es seleccionado favorablemente, y no existe recombinación entre ellos, sus vecinos son también seleccionados. El análisis de datos de otro proyecto destinado a ver la variación humana, el HapMap Project, ha revelado al menos 300 regiones del genoma que han estado bajo selección positiva durante los últimos 5.000-30.000 años. En algunos casos las regiones han podido reducirse un único gen (Grossman *et al.* 2010).

Enfermedades raras

Siendo financiado en su mayor parte por el Instituto Nacional de la Salud, el PGH debía casi todos sus logros a las áreas médicas. Cuando se puso en marcha el PGH se conocían menos de 100 genes relacionados con enfermedades raras heredables. Con la creación de mapas genéticos y físicos completos desde las primeras etapas del PGH la lista de genes relacionados con enfermedades heraditarias comenzó a crecer y se estima hoy se conocen más de 2850 defectos transmisibles según el modelo mendeliano (http://www.omim.org/).

Asignar una enfermedad monogénica a una región cromosómica y luego a un gen causante implicaba hasta ahora analizar muchas familias afectadas. Cada vez es más frecuente que se trate de eliminar la elaboración de mapas genéticos y con el secuenciado masivo en paralelo, debería poder interrogarse cada región deseada resecuenciando todo el exoma y/o el genoma.

Hasta el momento esta tarea no es del todo clara. Aunque pudieran filtrarse todas las variantes comunes, cualquier persona seguirá teniendo, como mínimo, 150 variantes de secuencia codificante raras, que afectan el 1% de sus genes, y aproximadamente 100 veces más variantes no codificantes (Abecasis *et al.* 2012). La única excepción susceptible de dar resultados claros en esta búsqueda

es el caso de enfermedades recesivas causadas por mutaciones en una secuencia codificante.

Esta estrategia ha funcionado en algunos casos cuando se busca respuesta a un fenómeno particular, se dispone de un número de pacientes y se lleva a cabo el secuenciado. Las secuencias deben ser filtradas y el secuenciado limitado al exoma demostró. Desde el 2009 se han demostrado algunos ejemplos de este enfoque para identificar genes que subyacen ciertas condiciones, trastornos ligados al cromosoma X, rasgos autosómicos recesivos o dominantes y mosaicismo somático (Choi *et al.* 2009; Ng *et al.* 2009; Lindhurst *et al.* 2011).

La genómica se convierte progresivamente en una herramienta diagnóstica. Algunos laboratorios de citogenética ya emplean microarrays de ADN que facilitan la detección de desarreglos cromosómicos clínicamente significativos, avanzando la evaluación de niños con retrasos idiopáticos del desarrollo, discapacidad intelectual, autismo y defectos en el parto.

En otras áreas específicas se ven grandes esfuerzos por emplear la tecnología, en ocasiones injustificados. Falta información y es necesario obtenerla, ordenarla y comprenderla. Un nuevo esfuerzo por determinar el origen de enfermedades con herencia mendeliana ha surgido recientemente. El mismo NIH financia tres centros para la Genómica Mendeliana (Centers for Mendelian Genomics, CMGs) que proveerán un servicio gratuito de secuenciado completo de exoma y genoma (ES/WGS) así como análisis y ayuda a investigadores de todo el mundo, interesados en buscar genes causales de fenotipos mendelianos desconocidos (http://mendelian.org).

Enfermedades comunes

Las enfermedades comunes tienen, no raras veces, un componente genético. Una variante de un gen "predispone" al desarrollo del estado mórbido y en muchas ocasiones el componente genético implica mucho más de uno, es poligénica. Antes del año 2000, sólo se conocía una decena de enfermedades de esta clase, gran parte de ellas relacionadas con componentes del sistema HLA (Human Leukocyte Antigens). Una década después la lista se extiende a más de 1.100 loci implicados en más de 160 enfermedades, casi todos ellos desde 2007.

Las enfermedades mendelianas raras son, por lo general, causados por mutaciones raras, porque la selección actúa en contra de los alelos causales. Por el contrario, las enfermedades comunes pueden estar asociadas a variantes de genes comunes en la población Una hipótesis formulada en 2001 postula de hecho, que las enfermedades comunes se asocian con frecuencia con variantes alélicas comunes (CD/CV, common disease- common variant), o polimorfismos con frecuencias mayores al 1 % (Reich & Lander 2001).

La hipótesis de CD/CV se basa en la premisa: puesto que la mayoría (el 99%) de la variación genética en la población humana proviene de variantes comunes, los alelos de susceptibilidad para una enfermedad común incluirán muchas variantes comunes excepto si los alelos han tenido un gran efecto deletéreo en la aptitud reproductiva durante largos períodos. Para las enfermedades comunes, muchos alelos de susceptibilidad pueden haber sido sólo ligeramente perjudiciales, neutros o incluso ventajosos. Ejemplos de esto pueden incluir enfermedades de aparición tardía, enfermedades resultantes de cambios en las condiciones de vida, como la diabetes y enfermedades del corazón, los rasgos morfológicos, y los alelos con efectos pleiotrópicos que dan como resultado en el equilibrio de selección.

Con la explosión de datos representando variantes comunes, haplotipos y genotipos, pudo comenzar a contrastarse esta hipótesis. Los resultados hasta ahora son alentadores. De estos estudios surgen tres resultados principales: (1) la mayoría de las enfermedades comunes, como otros rasgos, pueden ser influenciados por un gran número de loci; (2) la gran mayoría de las variantes comunes en estos loci tienen un efecto moderado, aumentando el riesgo de 10-50% (similares a los efectos de muchos factores de riesgo ambientales), y (3) los loci implicados incluyen la mayoría de los genes que se encuentran por análisis de ligamiento, pero se hallan muchos más genes no conocidos no previamente. Algunos científicos y médicos argumentaron que tales descubrimientos no son útiles para la comprensión de la enfermedad y no podían tener consecuencias terapéuticas. Algunos resultados sugieren sin embargo lo contrario.

Estudios de Asociación con Genomas Completos (Genome Wide Association Studies), revelan hasta el momento factores asociados con la degeneración de la mácula en el adulto, la enfermedad de Crohn, enfermedades autoinmunes, la diabetes tipo 2, algunos trastornos psiquiátricos (Lander 2011). Rasgos poligénicos que atraen nuestra atención desde el comienzo de la genética, como la altura, también son sujeto de estudio a través de estos métodos. Para concluir, aunque la lista se extiende, la eficacia o el efecto adverso de ciertas drogas, un área conocida desde hace un tiempo como farmacogenética.

Cáncer

El cáncer es una enfermedad genética, esto es claro desde 1980 (Dulbecco 1986). Mutaciones somáticas, a menudo asociadas con factores predisponentes, contribuyen a su desarrollo. En el año 2000 se reconocían 80 genes implicados en el cáncer, particularmente asociados con la producción de tumores sólidos. Muchos de ellos se reconocían por su asociación con oncogenes virales y con ensayos de transformación y unos pocos por mapeo genético.

Con el descubrimiento de cada gen asociado a una neoplasia, las empresas farmacéuticas han corrido para desarrollar inhibidores. Con blanco en proteínas implicadas mayormente en la señalización, se han reemplazado progresivamente los inhibidores del crecimiento, sumamente tóxicos para el paciente. Pero una terapia adecuada está aún muy ligada a la detección de mutaciones en cada paciente.

Un estudio de más de 3.000 muestras obtenidas de 26 tipos de cáncer, identificó recientemente más de 150 alteraciones recurrentes en el número de copias, sólo una cuarta parte de ellos contiene genes reconocidos por estar implicados en el cáncer (Beroukhim *et al.* 2010). Esto indica claramente que quedan muchos más por descubrir. Estudios semejantes revelaron una clase completamente nueva de genes asociados que consiste en factores de transcripción linaje específico (MITF en el melanoma, NKX2 en el cáncer de pulmón y Sox2 en el cáncer de esófago), translocaciones relacionadas tumores en la próstata y otras halladas con mucha frecuencia en algunos cánceres de pulmón. En este último ya habría ensayos clínicos para probar una nueva droga (Lander 2011).

Como en otros aspectos, el Cancer Genome Atlas es un proyecto destinado a secuenciar tantos genomas de tipos de cáncer diferente como sea posible mediante técnicas de secuenciado masivo. Este enfoque ya da algunos frutos.

Conocer el genoma se vuelve una necesidad

Hallar una explicación a un fenómeno desconocido siempre fue motivo para apresurarse. El conocimiento es en ocasiones sinónimo de gloria y/o una gran fuente de recursos. Desde el primer impulso proporcionado por el proyecto genoma humano la secuenciación de genomas enteros ha hallado muchos nichos.

Se ha desarrollado tecnología capaz de generar gran cantidad de información y se dispone de un número creciente de sistemas de análisis. Esta tecnología ha ganado espacio en numerosas áreas de investigación científica y médica, lo hace ahora en el diagnóstico y muy pronto llegará a la identificación de personas.

Muy pocos pueden ocultar la intriga que genera disponer de esta tecnología y el deseo de emplearla. Algunos investigadores pueden darle uso, y de manera semejante muchos laboratorios privados desearían hacerlo. Incluso muchos particulares desearían conocer su genoma saben que, disponiendo de recursos, podrían hacerlo. Pero, ¿se justifica realmente el uso o para qué serviría esta información?.

El genoma humano ha generado muchas expectativas durante la última década y se transforma en el centro de atención de los años por venir. Gran parte de la información obtenida recientemente permite anunciar resultados grandiosos. La gran capacidad de comunicación que alcanzamos hace posible asimismo que deseemos informarnos y se nos informe con distintos propósitos.

La publicidad de sistemas en desarrollo se ha vuelto corriente en un intento por atrapas clientes y no es raro, no saber para qué se compra un producto o se contrata un servicio.

Por el momento sólo un laboratorio de investigación público o de una gran empresa puede justificar "plenamente" el uso de la nueva tecnología, y para hacerlo debe contar con muchos recursos económicos. Como en el origen del automóvil o de la computadora hallar los medios y justificar su empleo son dos problemas mayores que dejarán de existir con el correr de los años.

No estamos en posición de dominar o de marcar un rumbo en las decisiones próximas pero podemos ser críticos, y decidir si es válido, el uso de cada nuevo producto o servicio que llega o se pone a nuestra disposición. Conocer las distintas facetas de la revolución genómica es una opción, pero se transforma poco a poco en una necesidad.

Llegar primero no es siempre sinónimo de estar seguro de la veracidad de los resultados obtenidos. Es necesario confirmar y validar cada hallazgo realizado. Pacientes con enfermedades raras puede recurrir a sistemas públicos deseosos de investigar el origen de su afección. No existen hasta el momento soluciones milagrosas.

Mientras el catálogo de secuencias de ADN y variantes crece, la atención se centra cada vez más en el transcriptoma, el proteoma y otras áreas en conocer todo lo que sea posible dentro de la célula. Conjugar y comprender toda esta información será muy difícil. Por el momento, la búsqueda de las claves para comprender cómo emplea la célula la información contenida en el genoma, avanza por esta vía.

Guía de estudio

CAPÍTULO 1- Las bases de la vida

- ADN y genes fueron descubiertos durante la década de 1860. ¿Porqué cree que no se halló una conexión entre ambos hasta 80 años más tarde?.

- ¿Porqué no sería posible congraciar las ideas de Darwin y Mendel antes de 1900?

- ¿Qué pruebas condujeron a pensar que el material genético se encuentra en el núcleo de las células?

- Dibuje, en detalle, un polinucleótido corto de ADN que contiene los cuatro nucleótidos. Represente el ARN transcripto a partir de ese fragmento.

- ¿Porqué se creía que las proteínas almacenarían la información genética?

- Describa algún experimento que haya indicado que los genes están hechos de ADN.

- Describa el ciclo lítico de infección de un bacteriófago.

- ¿Fueron los experimentos de Avery y Hershey-Chase inmediatamente aceptados por la comunidad científica? Porqué?

- ¿Qué diferencia hay entre apareamiento y apilamiento de bases? ¿Qué influencia tienen estas interacciones en la estructura de la doble hélice?

- Enumere las pruebas que llevaron a Watson y Crick a deducir que una molécula de ADN celular es una doble hélice.

CAPÍTULO 2 - Secuenciación del ADN

- ¿Cómo contribuyó el clonado de ADN al secuenciado? ¿Y la PCR?
- Enumere algunas enzimas empleadas en la investigación con ADN recombinante.
- ¿Cómo puede emplearse un bacteriófago para clonar ADN?
- Diagrame un experimento de secuenciación mediante terminadores de cadena.
- ¿Cómo se procede al secuenciado mediante degradación química?
- ¿Qué ventajas presenta un vector basado en el fago M13 para el secuenciado enzimático?
- ¿En qué difiere el secuenciado automático del procedimiento clásico del secuenciado con terminadores de cadena?

CAPÍTULO 3- Secuenciación del genoma humano

- ¿Porqué no pueden emplearse YACs en la construcción de librerías genómicas?
- ¿Porqué es útil un mapa en el secuenciado genómico?
- Evalúe la importancia del mapeo físico y el mapeo genético en el secuenciado del genoma humano.
- ¿Qué marcadores se emplean en el mapeo físico?
- ¿En qué contribuyó el trabajo de Mendel al mapeo genético?
- Describa de manera sucinta el secuenciado global aleatorio con un organismo de genoma pequeño como *Haemophilus influenzae.*
- Evalúe críticamente el método clon a clon como medio de secuenciado de un gran genoma eucariota.

- Describa ventajas y desventajas del método shotgun de secuenciado. ¿Qué precauciones se toman para asegurar la calidad y cobertura de la secuencia obtenida por este método?

- ¿Qué diferencia existe entre una brecha (gap) física y una de secuencia? ¿Cómo pueden llenarse ambas?

- ¿Qué método/s se emplea/n para obtener clones superpuestos en el secuenciado clon a clon?

- Justifique las siguientes afirmaciones:
  1. En años futuros será posible utilizar la bioinformática para obtener una descripción completa de las ubicaciones y las funciones de los genes en una secuencia del genoma.
  2. En los próximos años la bioinformática será obsoleta debido al desarrollo de métodos de experimentación rápida y eficaz para la localización y asignación de funciones a los genes en una secuencia del genoma.

- ¿Porqué el método clon a clon daría más seguridad que el secuenciado shotgun en la determinación de la secuencia de un gran genoma?

- ¿Cómo se reconoce la existencia de repeticiones durante el ensamblado de secuencias?

- ¿Cómo pueden llenarse brechas físicas en una secuencia obtenida mediante secuenciado aleatorio?

- ¿Cómo se distingue un polimorfismo de un error de secuencia durante el ensamblado de secuencias?

CAPÍTULO 4- ¿Qué hay en el genoma?

- Dibuje y anote la estructura de un gen humano "promedio".

- Defina elementos repetitivos intercalados y elementos repetitivos en tándem.

- Experimentos de inactivación de genes indican que algunos genes son redundantes en el genoma del ratón. ¿Ha confirmado este hallazgo el análisis del genoma? ¿Qué efecto puede esto tener a escala evolutiva?

- ¿Qué características distinguen las dos grandes clases de elementos transponibles? ¿Y los pseudogenes?

- Enumere diferencias entre LINEs y SINEs.

- Esquematice un segmento del cromosoma 2 del hombre incluyendo tantos elementos como pueda.

- ¿Qué diferencias y similitudes esperaría encontrar si se comparara su genoma con el sus padres?

- En el trabajo "la secuencia del genoma humano", de Venter y colaboradores, se caracteriza el genoma funcional como compuesto de: 23.2% de genes relacionados con la expresión, replicación y mantenimiento del genoma, 21.1% relacionado con la transmisión de señales, 17.5% con funciones bioquímicas generales en la célula y 38.2% con actividades varias. ¿Cree que habría una categorización más informativa de los genes codificantes?

- ¿Qué uso corriente se da a los microsatélites?

CAPÍTULO 5- Herencia del proyecto Genoma

- ¿Qué procedimientos no convencionales se emplean para el secuenciado de ADN?

- ¿Puede hallar alguna similitud entre un método de secuenciado masivo y el método de degradación química de Maxam y Gilbert?

- Suponiendo que usted tenga los medios necesarios para secuenciar su genoma completo, ¿lo haría? Porqué?

- Esquematice cómo funciona el secuenciado a través de nanoporos.

- Usted ha aislado una nueva especie de bacteria cuyo genoma es una molécula de ADN de aproximadamente 2,6 MB. Escriba un plan

detallado del proyecto que le permita obtener la secuencia del genoma de esa bacteria.

- ¿Cómo puede la genómica comparativa contribuir al estudio de genes de enfermedades humanas?

- Adopte y defienda una de las siguientes posiciones:
  1. Estoy a favor del secuenciado del genoma humano,
  2. Estoy en contra del secuenciado del genoma humano

- Defina ligamiento parcial.

- Defina haplotipo.

- ¿Podría emplearse un esquema de secuenciado masivo para el estudio de comunidades de microorganismos?

- Describa algunos caminos que sigue la investigación en el genoma humano.

- Si usted no tuviera impedimento alguno para hacerlo, ¿secuenciaría o haría secuenciar su genoma? ¿Porqué?

Referencias

Abecasis, G.R. et al., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), pp.56–65.

Adams, M.D. et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science (New York, N.Y.)*, 252(5013), pp.1651–1656.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Altshuler, D. et al., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803), pp.513–516.

Anderson, S. et al., 1980. A short primer for sequencing DNA cloned in the single-stranded phage vector M13mp2. *Nucleic acids research*, 8(8), pp.1731–1743.

Anon, 1965. CENTENARY OF MENDEL'S PAPER. *British Medical Journal*, 1(5431), pp.368–374.

Anon, 2009. *SOLiD DNA Sequencing*, Available at: http://www.youtube.com/watch?v=nlvyF8bFDwM&feature=youtube_gdata_player [Accessed May 12, 2013].

Ansorge, W. et al., 1986. A non-radioactive automated method for DNA sequence determination. *Journal of biochemical and biophysical methods*, 13(6), pp.315–323.

Ansorge, W. et al., 1988. Non-radioactive automated sequencing of oligonucleotides by chemical degradation. *Nucleic acids research*, 16(5), pp.2203–2206.

Arber, W. & Linn, S., 1969. DNA Modification and Restriction. *Annual Review of Biochemistry*, 38(1), pp.467–500.

Avery, O.T., Macleod, C.M. & McCarty, M., 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of experimental medicine*, 79(2), pp.137–158.

Barbazuk, W.B., Bedell, J.A. & Rabinowicz, P.D., 2005. Reduced representation sequencing: a success in maize and a promise for other plant genomes.

BioEssays: news and reviews in molecular, cellular and developmental biology, 27(8), pp.839–848.

Bauman, J.G. et al., 1980. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-melabelled RNA. *Experimental cell research*, 128(2), pp.485–490.

Bauman, J.G. et al., 1981. Rapid and high resolution detection of in situ hybridisation to polytene chromosomes using fluorochrome-labeled RNA. *Chromosoma*, 84(1), pp.1–18.

Bejerano, G. et al., 2004. Ultraconserved elements in the human genome. *Science (New York, N.Y.)*, 304(5675), pp.1321–1325.

Benham, F. et al., 1989. A method for generating hybrids containing non-selected fragments of human chromosomes. *Genomics*, 4(4), pp.509–517.

Bentley, D.R. et al., 2008. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature*, 456(7218), pp.53–59.

Beroukhim, R. et al., 2010. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283), pp.899–905.

Brownlee, G.G., Sanger, F. & Barrell, B.G., 1967. Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli. *Nature*, 215(5102), pp.735–736.

Buell, G.N. et al., 1978. Synthesis of full length cDNAs from four partially purified oviduct mRNAs. *The Journal of biological chemistry*, 253(7), pp.2471–2482.

Burke, D.T., Carle, G.F. & Olson, M.V., 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science (New York, N.Y.)*, 236(4803), pp.806–812.

Casavant, N.C. et al., 2000. The end of the LINE?: lack of recent L1 activity in a group of South American rodents. *Genetics*, 154(4), pp.1809–1817.

Chargaff, E. et al., 1951. The Composition of the Desoxyribonucleic Acid of Salmon Sperm. *Journal of Biological Chemistry*, 192(1), pp.223–230.

Chien, A., Edgar, D.B. & Trela, J.M., 1976. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *Journal of bacteriology*, 127(3), pp.1550–1557.

Chinwalla, A.T. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–562.

Choi, M. et al., 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, 106(45), pp.19096–19101.

Cohen, S.N. et al., 1973. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 70(11), pp.3240–3244.

Cohen, S.N., Chang, A.C.Y. & Hsu, L., 1972. Nonchromosomal Antibiotic Resistance in Bacteria: Genetic Transformation of Escherichia coli by R-Factor

DNA*. *Proceedings of the National Academy of Sciences of the United States of America*, 69(8), pp.2110–2114.

Consortium, I.H.G.S., 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.

Cox, D.R. et al., 1990. Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science (New York, N.Y.)*, 250(4978), pp.245–250.

Crow, E.W. & Crow, J.F., 2002. 100 years ago: Walter Sutton and the chromosome theory of heredity. *Genetics*, 160(1), pp.1–4.

Cuddapah, S. et al., 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research*, 19(1), pp.24–32.

Danna, K. & Nathans, D., 1971. Specific Cleavage of Simian Virus 40 DNA by Restriction Endonuclease of Hemophilus Influenzae*. *Proceedings of the National Academy of Sciences of the United States of America*, 68(12), pp.2913–2917.

Deloukas, P. et al., 1998. A physical map of 30,000 human genes. *Science (New York, N.Y.)*, 282(5389), pp.744–746.

Donis-Keller, H. et al., 1987. A genetic linkage map of the human genome. *Cell*, 51(2), pp.319–337.

Drake, N., 2011. What is the human genome worth? *Nature News*. Available at: http://www.nature.com/news/2011/110511/full/news.2011.281.html [Accessed May 21, 2013].

Dulbecco, R., 1986. A turning point in cancer research: sequencing the human genome. *Science (New York, N.Y.)*, 231(4742), pp.1055–1056.

Dunham, I. et al., 1999. The DNA sequence of human chromosome 22. *Nature*, 402(6761), pp.489–495.

Eid, J. et al., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910), pp.133–138.

ELSON, D. & CHARGAFF, E., 1952. On the desoxyribonucleic acid content of sea urchin gametes. *Experientia*, 8(4), pp.143–145.

Ewing, B. & Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3), pp.186–194.

Fiers, W. et al., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551), pp.500–507.

Fleischmann, R.D. et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223), pp.496–512.

FRANKLIN, R.E. & GOSLING, R.G., 1953. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356), pp.740–741.

Fraser, C.M. et al., 1995. The minimal gene complement of Mycoplasma genitalium. *Science (New York, N.Y.)*, 270(5235), pp.397–403.

Garber, R.L., Kuroiwa, A. & Gehring, W.J., 1983. Genomic and cDNA clones of the homeotic locus Antennapedia in Drosophila. *The EMBO journal*, 2(11), pp.2027–2036.

Gibbs, R.A. et al., 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982), pp.493–521.

Glazov, E.A. et al., 2008. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. *Genome research*, 18(6), pp.957–964.

Gocayne, J. et al., 1987. Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family. *Proceedings of the National Academy of Sciences of the United States of America*, 84(23), pp.8296–8300.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3), pp.705–708.

Grausz, J.D., 1993. Gene mapping of the mammalian genome: the CEPH and Genethon initiative. *Current opinion in biotechnology*, 4(6), pp.665–671.

Green, R.E. et al., 2010. A Draft Sequence of the Neandertal Genome. *Science*, 328(5979), pp.710–722.

Grossman, S.R. et al., 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, 327(5967), pp.883–886.

Hattori, M. et al., 2000. The DNA sequence of human chromosome 21. *Nature*, 405(6784), pp.311–319.

Havlak, P. et al., 2004. The Atlas genome assembly system. *Genome research*, 14(4), pp.721–732.

Heidecker, G., Messing, J. & Gronenborn, B., 1980. A versatile primer for DNA sequencing in the M13mp2 cloning system. *Gene*, 10(1), pp.69–73.

HERSHEY, A.D. & CHASE, M., 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1), pp.39–56.

Hesselberth, J.R. et al., 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods*, 6(4), pp.283–289.

Higgins, D.G. & Sharp, P.M., 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1), pp.237–244.

Hillier, L.D. et al., 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome research*, 6(9), pp.807–828.

HOLLEY, R.W. et al., 1965. NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID. *The Journal of biological chemistry*, 240, pp.2122–2128.

Horvath, J.E., Schwartz, S. & Eichler, E.E., 2000. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome research*, 10(6), pp.839–852.

Howald, C. et al., 2012. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome research*, 22(9), pp.1698–1710.

Huang, X.C., Quesada, M.A. & Mathies, R.A., 1992. DNA sequencing using capillary array electrophoresis. *Analytical chemistry*, 64(18), pp.2149–2154.

Hubbard, T. et al., 2005. Ensembl 2005. *Nucleic acids research*, 33(Database issue), pp.D447–453.

Hudson, T.J. et al., 1995. An STS-based map of the human genome. *Science (New York, N.Y.)*, 270(5244), pp.1945–1954.

Jackson, D.A., Symons, R.H. & Berg, P., 1972. Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 69(10), pp.2904–2909.

JACOB, F. & MONOD, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3, pp.318–356.

Jorgenson, J.W. & Lukacs, K.D., 1981. Free-zone electrophoresis in glass capillaries. *Clinical chemistry*, 27(9), pp.1551–1553.

Jou, W.M. et al., 1972. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature*, 237(5350), pp.82–88.

Karger, B.L. & Guttman, A., 2009. DNA Sequencing by Capillary Electrophoresis. *Electrophoresis*, 30(Suppl 1), pp.S196–S202.

Karlin, S., Bergman, A. & Gentles, A.J., 2001. Genomics: Annotation of the Drosophila genome. *Nature*, 411(6835), pp.259–260.

Kasianowicz, J.J. et al., 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24), pp.13770–13773.

Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. *Genome research*, 12(4), pp.656–664.

King, M.C. & Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, 188(4184), pp.107–116.

Klenow, H. & Henningsen, I., 1970. Selective Elimination of the Exonuclease Activity of the Deoxyribonucleic Acid Polymerase from Escherichia coli B by Limited Proteolysis*. *Proceedings of the National Academy of Sciences of the United States of America*, 65(1), pp.168–175.

Lander, E.S., 2011. Initial impact of the sequencing of the human genome. *Nature*, 470(7333), pp.187–197.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Lander, E.S. & Waterman, M.S., 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3), pp.231–239.

LEHMAN, I.R. et al., 1958. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. *The Journal of biological chemistry*, 233(1), pp.163–170.

Levy, S. et al., 2007. The Diploid Genome Sequence of an Individual Human. *PLoS Biology*, 5(10).

Lewis, B.P. et al., 2003. Prediction of mammalian microRNA targets. *Cell*, 115(7), pp.787–798.

Ley, T.J. et al., 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218), pp.66–72.

Lindhurst, M.J. et al., 2011. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *The New England journal of medicine*, 365(7), pp.611–619.

Long, E.O. & Dawid, I.B., 1980. Repeated genes in eukaryotes. *Annual review of biochemistry*, 49, pp.727–764.

Loomis, E.W. et al., 2012. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research*. Available at: http://genome.cshlp.org/content/early/2012/10/11/gr.141705.112 [Accessed May 13, 2013].

Malik, H.S., Henikoff, S. & Eickbush, T.H., 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome research*, 10(9), pp.1307–1318.

Mandel, M. & Higa, A., 1970. Calcium-dependent bacteriophage DNA infection. *Journal of molecular biology*, 53(1), pp.159–162.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–380.

Martin-Gallardo, A. et al., 1992. Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genetics*, 1(1), pp.34–39.

Mattick, J.S. & Makunin, I.V., 2006. Non-coding RNA. *Human molecular genetics*, 15 Spec No 1, pp.R17–29.

Maxam, A.M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), pp.560–564.

McClintock, B., 1953. Induction of Instability at Selected Loci in Maize. *Genetics*, 38(6), pp.579–599.

McCombie, W.R. et al., 1992. Expressed genes, Alu repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3. *Nature Genetics*, 1(5), pp.348–353.

McPherson, J.D. et al., 2001. A physical map of the human genome. *Nature*, 409(6822), pp.934–941.

Meissner, A. et al., 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205), pp.766–770.

Mikkelsen, T.S. et al., 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141), pp.167–177.

Morrow, J.F. et al., 1974. Replication and transcription of eukaryotic DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 71(5), pp.1743–1747.

Mullis, K. et al., 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1, pp.263–273.

Murray, A.W. & Szostak, J.W., 1983. Construction of artificial chromosomes in yeast. *Nature*, 305(5931), pp.189–193.

Murray, J.C. et al., 1994. A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science (New York, N.Y.)*, 265(5181), pp.2049–2054.

Ng, S.B. et al., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), pp.272–276.

Nyrén, P., 1987. Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical biochemistry*, 167(2), pp.235–238.

Nyrén, P., Pettersson, B. & Uhlén, M., 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical biochemistry*, 208(1), pp.171–175.

O'Connor, M., Peifer, M. & Bender, W., 1989. Construction of large DNA segments in Escherichia coli. *Science (New York, N.Y.)*, 244(4910), pp.1307–1312.

Okada, N. et al., 1997. SINEs and LINEs share common 3' sequences: a review. *Gene*, 205(1-2), pp.229–243.

Pei, B. et al., 2012. The GENCODE pseudogene resource. *Genome biology*, 13(9), p.R51.

Pennisi, E., 2010. Semiconductors Inspire New Sequencing Technologies. *Science*, 327(5970), pp.1190–1190.

Porreca, G.J. et al., 2007. Multiplex amplification of large sets of human exons. *Nature methods*, 4(11), pp.931–936.

Prober, J.M. et al., 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (New York, N.Y.)*, 238(4825), pp.336–341.

Reich, D.E. & Lander, E.S., 2001. On the allelic spectrum of human disease. *Trends in genetics: TIG*, 17(9), pp.502–510.

Ronaghi, M. et al., 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, 242(1), pp.84–89.

Ronaghi, M., Uhlén, M. & Nyrén, P., 1998. A sequencing method based on real-time pyrophosphate. *Science (New York, N.Y.)*, 281(5375), pp.363, 365.

Rusk, N., 2011. Torrents of sequence. *Nature Methods*, 8(1), pp.44–44.

Saiki, R.K. et al., 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)*, 239(4839), pp.487–491.

Sanger, F. et al., 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of Molecular Biology*, 143(2), pp.161–178.

Sanger, F. et al., 1982. Nucleotide sequence of bacteriophage lambda DNA. *Journal of molecular biology*, 162(4), pp.729–773.

Sanger, F., Air, G.M., et al., 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596), pp.687–695.

Sanger, F. et al., 1973. Use of DNA Polymerase I Primed by a Synthetic Oligonucleotide to Determine a Nucleotide Sequence in Phage f1 DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 70(4), pp.1209–1213.

Sanger, F., Brownlee, G.G. & Barrell, B.G., 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of molecular biology*, 13(2), pp.373–398.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), pp.441–448.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Scheibye-Alsing, K. et al., 2009. Sequence assembly. *Computational biology and chemistry*, 33(2), pp.121–136.

Schuler, G.D. et al., 1996. A gene map of the human genome. *Science (New York, N.Y.)*, 274(5287), pp.540–546.

Sears, L.E. et al., 1992. CircumVent thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo-) DNA polymerase. *BioTechniques*, 13(4), pp.626–633.

Seeburg, P.H. et al., 1977. Nucleotide sequence of a human gene coding for a polypeptide hormone. *Transactions of the Association of American Physicians*, 90, pp.109–116.

Shendure, J. et al., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, 309(5741), pp.1728–1732.

Skaletsky, H. et al., 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942), pp.825–837.

Smit, A.F., 1996. The origin of interspersed repeats in the human genome. *Current opinion in genetics & development*, 6(6), pp.743–748.

Smith, H.O. & Welcox, K.W., 1970. A Restriction enzyme from Hemophilus influenzae: I. Purification and general properties. *Journal of Molecular Biology*, 51(2), pp.379–391.

Smith, L.M. et al., 1986. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071), pp.674–679.

Smith, L.M. et al., 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic acids research*, 13(7), pp.2399–2412.

Smith, T.F. & Waterman, M.S., 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), pp.195–197.

Spiegelman, S., Watson, K.F. & Kacian, D.L., 1971. Synthesis of DNA complements of natural RNAs: a general approach. *Proceedings of the National Academy of Sciences of the United States of America*, 68(11), pp.2843–2845.

Springer, N.M., Xu, X. & Barbazuk, W.B., 2004. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant physiology*, 136(2), pp.3023–3033.

Stamatoyannopoulos, J.A., 2012. What does our genome encode? *Genome Research*, 22(9), pp.1602–1611.

Strauss, E.C. et al., 1986. Specific-primer-directed DNA sequencing. *Analytical Biochemistry*, 154(1), pp.353–360.

Sutton, G. et al., 1995. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *Genome Science and Technology*, 1(1), pp.9–19.

SUTTON, W.S., 1903. THE CHROMOSOMES IN HEREDITY. *Biological Bulletin*, 4, pp.231–251.

Swerdlow, H. & Gesteland, R., 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, 18(6), pp.1415–1419.

Szybalski, W., Kubinski, H. & Sheldrick, P., 1966. Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. *Cold Spring Harbor symposia on quantitative biology*, 31, pp.123–127.

Tabor, S. & Richardson, C.C., 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, 84(14), pp.4767–4771.

The ENCODE Project Consortium, 2012. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, 489(7414), pp.57–74.

Thurman, R.E. et al., 2012. The accessible chromatin landscape of the human genome. *Nature*, 489(7414), pp.75–82.

Upham, R.A., 1970. The computer-aided determination of peptide sequences. *The Biochemical journal*, 117(2), p.1P–2P.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–1351.

WATSON, J.D. & CRICK, F.H., 1953a. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361), pp.964–967.

WATSON, J.D. & CRICK, F.H., 1953b. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), pp.737–738.

Weier, H.U. & Gray, J.W., 1988. A programmable system to perform the polymerase chain reaction. *DNA (Mary Ann Liebert, Inc.)*, 7(6), pp.441–447.

Weinstein, L.B. & Steitz, J.A., 1999. Guided tours: from precursor snoRNA to functional snoRNP. *Current opinion in cell biology*, 11(3), pp.378–384.

Wheeler, D.A. et al., 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189), pp.872–876.

Wheeler, T.J. et al., 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic acids research*, 41(Database issue), pp.D70–82.

Whitfield, T.W. et al., 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome biology*, 13(9), p.R50.

WILKINS, M.H.F., STOKES, A.R. & WILSON, H.R., 1953. Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356), pp.738–740.

Yu, A. et al., 2001. Comparison of human genetic and sequence-based physical maps. *Nature*, 409(6822), pp.951–953.

Santa Rosa, LP, Diciembre de 2013.