Mapaje físico y caracterización de genes humanos

Roser Gonzàlez Duarte
Departamento de Genética
Universidad de Barcelona.
Diagonal 645. 08071-Barcelona.

1. Introducción

El aislamiento y caracterización de los genes del genoma humano es hoy un objetivo científico prioritario al que se dedican muchos recursos económicos y que motiva a muchos grupos de investigación. A nadie se le escapa la importancia de desvelar la información contenida en los cromosomas humanos ya que, por un lado, permitirá descubrir nuevas funciones y analizar desde nuevas perspectivas los procesos biológicos básicos y, por otro, el conocimiento de los genes implicados en las enfermedades hereditarias es un requisito esencial para el diseño de nuevas estrategias terapéuticas de enfermedades hasta ahora no tratables mediante la terapia convencional.

La Genética Molecular ha avanzado de forma vertiginosa en el último tercio de este siglo. Muchas contribuciones científicas relevantes han hecho posible que hoy podamos manipular, modificar y analizar directamente la molécula de DNA en el tubo de ensayo, tan sólo cuarenta años más tarde que Watson y Crick (1953) propusieran su estructura. Entre los avances metodológicos más sobresalientes hay que destacar el descubrimiento y la caracterización de los enzimas de restricción, las técnicas de clonaje y expresión heteróloga del DNA, la secuenciación, la transferencia de ácidos nucleicos y pro-

teínas a soportes sólidos inertes e hibridación con sondas específicas y, por último, la amplificación de secuencias de DNA (PCR). El conjunto de todas estas tecnologías es lo que ha hecho posible abordar la estructura de genomas complejos, como es el caso del genoma humano.

El elevado tamaño del genoma humano y la abundancia de secuencias repetidas de DNA intercaladas en su estructura son, entre otras, las limitaciones más importantes que hay que sobrepasar para el aislamiento y caracterización de genes. El contenido de DNA del genoma haploide es de $3x10^9$ pares de bases (ó 3.000 Mb) lo que representa un orden de magnitud superior a la del genoma de *Drosophila*, uno de los organismos más estudiados a nivel genético. Las secuencias repetidas de DNA, agrupadas en distintas familias, son un componente constante en las genotecas de DNA, cDNA y en los fragmentos de DNA (sondas) que se utilizan para caracterizar las secuencias codificadoras. Su abundancia es la causa de la frecuente detección de "falsos positivos" en los procesos de hibridación que tanto enlentecen y dificultan el aislamiento y caracterización de genes.

Tan sólo el 20% del genoma humano contiene genes o secuencias relacionadas con su estructura y función. Algunas de las secuencias únicas incluidas en este porcentaje son "reliquias" de genes ancestrales, en la actualidad no codificantes. Otras, son secuencias que proceden de una duplicación reciente de secuencias codificadoras no sometidas a una presión selectiva y por tanto aptas para incorporar modificaciones sucesivas hasta adoptar una nueva función. Finalmente, también hay regiones que forman parte de la familia de secuencias "únicas" necesarias para el control de la expresión de los genes estructurales, aunque la mayor parte de ellas sean todavía desconocidas.

De este 20% se estima que tan sólo un 10% es realmente DNA codificante. El resto, más del 90%, está formado por intrones y regiones no traducidas, pseudogenes y fragmentos de genes originados a partir de duplicaciones incompletas. Por otra parte, el 80% de las 3.000 Mb que constituyen el genoma nuclear está formado por secuencias únicas (70-80% aprox.) y repetidas (20-30% aprox.) y no está asociado a ninguna función codificante (Figura 1). Estos datos ponen de manifiesto por sí mismos que el aislamiento de genes del genoma humano ha de ser una tarea tan costosa como la de encontrar "una aguja en un pajar", y así la han calificado algunos investigadores en sus publicaciones más recientes (1,2).

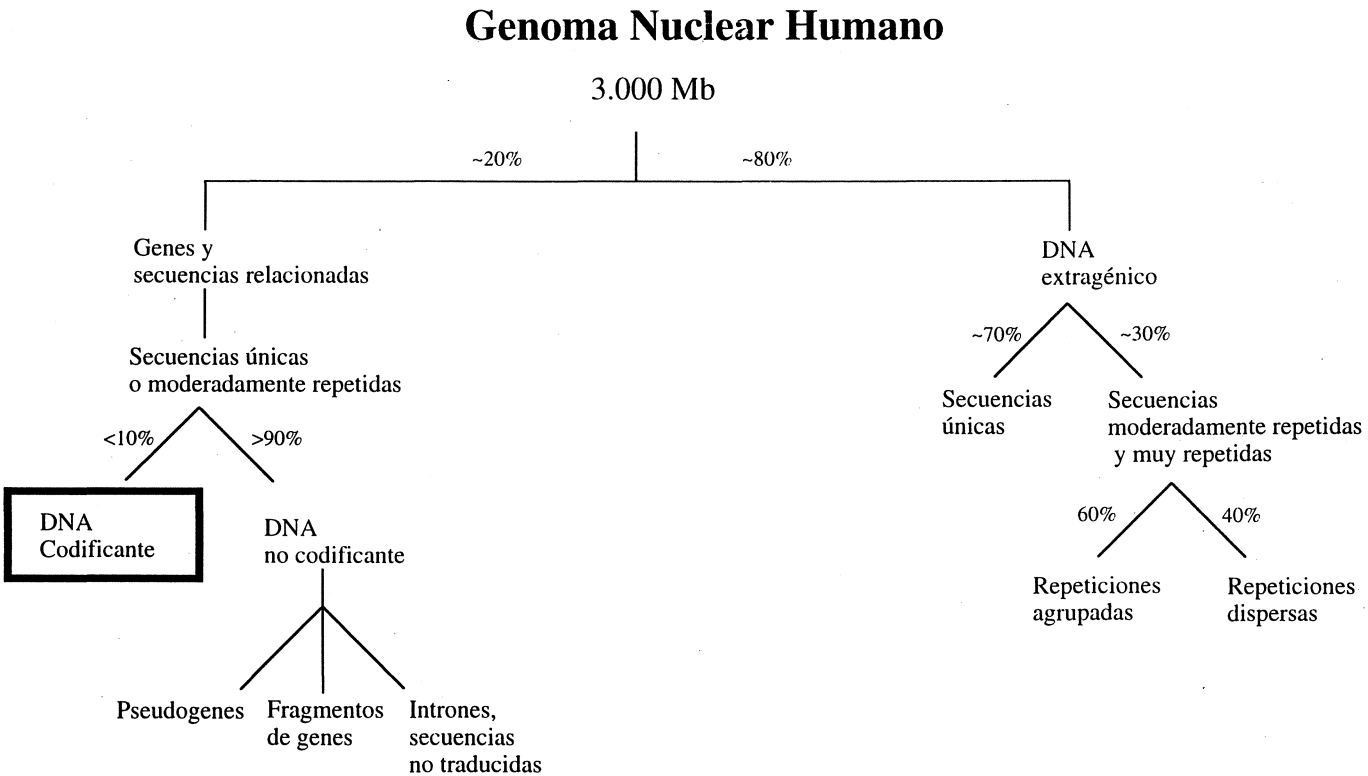


Figura 1. Organización del genoma nuclear humano

2. Marcadores genéticos

Existen muchas secuencias polimórficas en el genoma humano. Se trata en su mayoría de variantes selectivamente neutras localizadas en intrones y regiones flanqueantes de genes, idóneas para el mapaje genético. El estudio de la segregación de estas variantes en una familia afecta de una enfermedad hereditaria nos permite determinar la herencia del cromosoma portador de la deficiencia mediante el análisis de la variante asociada al fenotipo defectivo y así, realizar un diagnóstico genético. Existen varios tipos de secuencias: polimorfismos de dianas de restricción, secuencias minisatélite y microsatélite. Las substituciones de nucleótidos en las primeras, generan presencia o ausencia de diana para un enzima de restricción. En este caso sólo pueden existir dos variantes alélicas en la población y, por tanto, estos polimorfismos presentan un grado de informatividad muy bajo. Las secuencias minisatélite están formadas por un número variable de repeticiones de una secuencia de 5-60 nucleótidos y las microsatélite por una de 1 a 5 nucleótidos. De estas dos últimas, en una misma población pueden coexistir hasta 15 variantes para un mismo *locus*. Su elevada informatividad y abundancia, así como la rápida y fácil detección por PCR de cada variante han sido la causa de que los mini- y microsatélites se hayan convertido en los protagonistas de los tests de ligamiento genético, de los ensayos de paternidad y de los análisis forenses.

Analizando la herencia de una enfermedad y los marcadores polimórficos presentes en un conjunto de familias afectas podremos deducir el grado de ligamiento genético entre el *locus* responsable de la enfermedad y el conjunto de marcadores próximos a él. La frecuencia de recombinantes se traducirá en una distancia genética, expresada, generalmente, en centi Morgans (cM). Pero este parámetro genético sólo nos permite inferir una aproximación física (1 cM corresponde aproximadamente a 1 Mb) y de ningún modo nos revela cuál es el gen defectivo ni de qué tipo de secuencia de DNA se trata. Para ello hemos de recurrir a mapas físicos en los que las secuencias de DNA de una región cromosómica están ordenadas y, a partir de un punto de referencia, avanzar fragmento a fragmento hasta alcanzar el gen causante de la enfermedad. La construcción de mapas físicos es extremadamente laboriosa debido a las características mencionadas del genoma humano y siempre es necesario proceder a una clonación de los fragmentos previa a su identificación.

3. Vectores

Los vectores de clonación son moléculas de DNA sintetizadas *in vitro*, diseñadas especialmente para replicar y expresar un DNA exógeno en una

célula huésped. Los primeros vectores de clonación se sintetizaron a partir de fragmentos de plásmidos bacterianos y genomas de bacteriófagos, muy especialmente el del fago lambda (λ). Sin embargo, el tamaño máximo de DNA exógeno que ambos tipos de vectores podían aceptar, 5 y 20 kb respectivamente, constituía una limitación muy importante para la clonación del genoma humano así como para el análisis de subregiones cromosómicas. El desarrollo posterior de vectores cósmidos, híbridos entre plásmidos bacterianos y las regiones cos de los extremos del genoma de l, capaces de aceptar insertos de más de 40 kb, ha facilitado enormemente el clonaje de grandes fragmentos de DNAs eucarióticos. Aún a pesar de sus limitaciones, entre ellas, la obtención de DNAs reordenados artificialmente durante el proceso de clonación, los cósmidos y los cromosomas artificiales de levadura (YACs) son, hoy día, los elementos indispensables para la construcción de mapas físicos del genoma humano. Los cromosomas artificiales son moléculas lineales de DNA que mantienen su autonomía y capacidad de replicación en el huésped gracias a unos elementos genéticos de levadura: secuencias de replicación autónoma (ARS), centrómero (CEN), genes marcadores para la selección de los transformantes y unas secuencias de DNA telomérico de *Tetrahymena*. Además, unas secuencias adicionales derivadas del plásmido bacteriano pBR322 garantizan su propagación en *E. coli*. La principal ventaja de este tipo de vectores, desarrollados a finales de los años 80, es que han permitido incrementar en un orden de magnitud el tamaño del inserto, entre 100 y 1.000 kb (3).

4. Mapas físicos

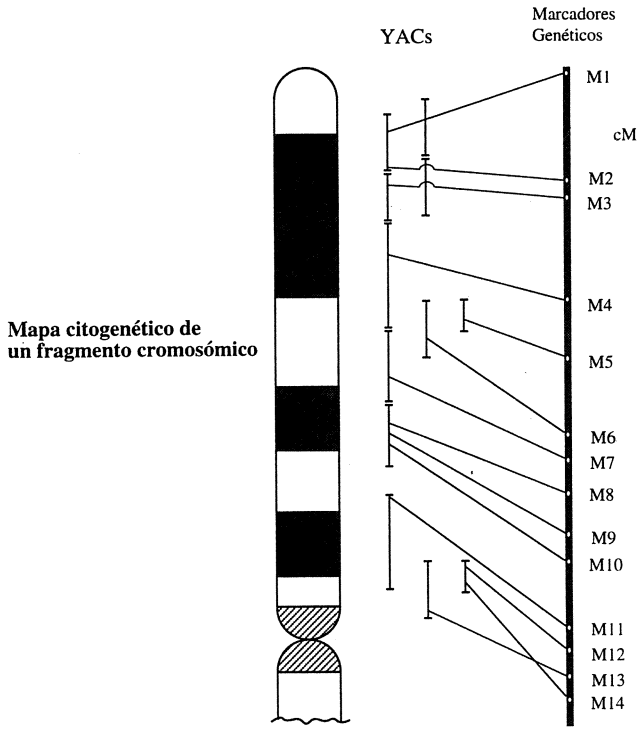
La molécula de DNA es el soporte físico de la información genética. Su estructura determina la síntesis de un RNA que al procesarse dará lugar a una proteína. Para conocer la naturaleza de un gen es necesario conocer su secuencia y a partir de ella, por comparación con las ya existentes, inferir su función. Un gen mutante presentará una secuencia alterada respecto el gen normal. Su presencia asociada a una enfermedad constituye una evidencia importante a favor de que se trata del gen causante de la patología. Generalmente, la ausencia de pistas sobre la naturaleza de los genes responsables de las enfermedades hereditarias descritas dificulta su identificación y de ahí, la necesidad de partir de un marcador genético ligado a la enfermedad, de aberraciones cromosómicas asociadas a ella o de mapajes *in situ* sobre los cromosomas humanos para alcanzar el gen recorriendo una región genómica previamente clonada. Dado que las distancias subcromosómicas que hay que analizar son a menudo muy grandes, del orden de varias megabases (1 Mb = 10^3 kb), es necesario disponer de un conjunto de fragmentos de DNA, con regiones de solapamiento entre ellos, que hayan sido clonados

independientemente. Este conjunto de clones o "contig" es el material físico que se deberá analizar.

De lo expuesto anteriormente se desprende la necesidad de relacionar (o integrar) los mapas físicos con los genéticos y citogenéticos. La rápida caracterización de los genes del genoma humano depende en gran medida de la asociación entre este tipo de datos que hoy tiende a ser completado con el mapaje de los productos de transcripción (4). Esta tarea, conceptualmente sencilla, es muy costosa y ardua desde el punto de vista experimental. (Figura 2)

El trabajo realizado por el grupo de Weissenbach en el cromosoma 7 en colaboración con otros grupos es un buen ejemplo de ello (5).

Figura 2. Integración del mapa citogenético, físico y genético de un fragmento de un cromosoma humano. Se indican los YACs (solapados) que cubren la región y los marcadores genéticos localizados sobre los YACs. Mediante experimentos de hibridación puede además determinarse el conjunto de cósmidos que están contenidos en un YAC (cosmid pocket). Las distancias genéticas entre los marcadores genéticos se expresan en centi Morgans (cM). El mapaje de los YACs sobre una extensión de cromosomas humanos por hibridación in situ (FISH) permite relacionar el mapa físico con el citogenético.



5. Estrategias para el aislamiento de genes

Se han descrito metodologías muy diversas para, partiendo de YACs, identificar las secuencias que corresponden a genes estructurales humanos. podríamos distinguir tres grandes tipos de estrategias: 1) "Exon Trapping", 2) Selección de cDNAs y 3) Caracterización Directa de cDNAs.

a) "Exon Trapping" (6)

Se basa en la selección *in vivo* de los sitios de procesamiento del RNA ("splicing") que flanquean a los exones presentes en el DNA genómico. Para ello, se ha diseñado un vector con un fragmento del gen tat del virus HIV-1 que contiene un intron flanqueado por los exones correspondientes. El sitio de introducción del inserto está en la región intrónica. Si el DNA humano que se analiza contiene un exon completo, con sus propias señales de "splicing", esta secuencia aparecerá en el mRNA cuando se exprese el fragmento de tat y se obtendrá un cDNA amplificable por PCR, cuando se utilicen como cebadores oligonucleótidos específicos para los exones flanqueantes del gen viral. Si, en cambio, no hay exones en el inserto que se ensaya se recompondrá una diana de restricción y, previa digestión, este DNA no podrá ser amplificado.

Si bien esta estrategia es conceptualmente atractiva y técnicamente factible, no todos los experimentos realizados han permitido alcanzar resultados positivos. Una de las limitaciones mas importantes es la abundancia de "falsos positivos", mayoritariamente debida a regiones crípticas de "splicing" presentes en el inserto genómico. Otra es la obtención de productos de PCR redundantes que, al presentar un tamaño ligeramente distinto entre ellos, serán inicialmente contabilizados como si correspondieran a distintos exones.

b) Selección de cDNAs (7)

A partir de una genoteca de DNA o de una región subcromosómica clonada en cósmidos o YACs se obtiene un DNA puro que se inmoviliza sobre un soporte inerte y se selecciona mediante hibridación con cDNAs procedentes de una genoteca. Estos cDNAs se han obtenido a partir de los mRNAs de un tejido o de líneas celulares específicas y por tanto corresponden a genes que se transcriben. Existen diversas variantes de esta metodología, tampoco exenta de limitaciones importantes, entre ellas, la presencia de secuencias repetidas de DNA en el 90% de las regiones 3' no codificantes de los mRNAs y diversos contaminantes frecuentes en las gentotecas de cDNA. Este método permite analizar muestras de clones de cósmidos previamente ordenados, de forma que la localización del híbrido sobre el filtro permite identificar inmediatamente el clon al cual pertenece.

c) Caracterización Directa de cDNAs (8)

El concepto que subyace a esta estrategia es que puede abordarse la búsqueda directa de genes a partir de un fragmento genómico grande, por ejemplo un YAC que contenga un inserto de 550 kb del genoma humano, utilizando como sonda una mezcla de cDNAs no procedentes de las genotecas convencionales. La diferencias principales respecto el método anterior es que tanto el fragmento genómico inicial como los cDNAs "sonda" se encuentran en solución, con el fin de minimizar la irreproducibilidad asociada a los procesos de hibridación sobre soportes físicos. La segunda característica diferencial es el proceso cuidadoso de obtención de los cDNAs (no se parte de genotecas de cDNA comerciales) y el bloqueo de las secuencias repetidas, lo que permite obtener una muestra más representativa y pura de los productos de transcripción presentes en un tejido o un tipo celular y incrementar en más de 1000 veces (enriquecer) la concentración de un cDNA específico en sólo un proceso de selección.

Las tres estrategias mencionadas son complementarias y pueden ser utilizadas conjuntamente. De hecho lo son por muchos autores. A su vez, se describen continuamente nuevas variantes que mejoran considerablemente el rendimiento de cada una de ellas. Nuevas técnicas para la caracterización de regiones 5' de un gen, como las versiones actualizadas de la detección de islas ricas en secuencias CpG, "island rescue PCR", o para la identificación rápida por PCR de los productos de transcripción de un DNA contenido en un YAC o la de obtención de nuevos marcadores para el mapaje físico son, entre otras, aportaciones muy valiosas (9,10,11,12).

Cuando recordamos que, una vez redescubiertas las leyes de Mendel, las bases de la ciencia que hoy denominamos Genética fueron establecidas hacia 1900, (W. Sutton, T. Boveri, T.H. Morgan, C. Bridges, entre otros) es impresionante y alentador al mismo tiempo considerar el gran reto que cierra este mismo siglo y abre las puertas de una nueva medicina en el siglo XXI "El conocimiento del genoma humano".

6. Referencias

1. Bird A.P. (1995) Gene number, noise reduction and biological complexity. TIG 11:94-100.
2. Brennan M.B. & Hochgeschwender U. (1995) So many needles, so much hay. Hum.Mol.Genet. 4:153-156
3. Monaco A.P. & Larin Z. (1994) YACs, BACs, PACs and MACs: artificial chromosomes as research tools. Tibtech 12:280-286.

4. Gardiner K. & Mural R.J. (1995) Getting the message: identifying transcribed sequences. TIG 11:77-79
5. Green E.D., Idol J.R., Mohr-Tidwell R.M., Braden V.V., Peluso D.C., Fulton R.S., Massa H.F., Magness C.L., Wilson A.M., Kimura J., Weissenbach J.& Trask B. (1994) Integration of physical, genetic and cytogenetic maps of human chromosome 7:isolation and analysis of yeast artificial chromosome clones for 117 mapped genetic markers. Hum. Mol. Genet.3:489-501
6. Church D.M., Stotler C.J., Rutter J.L., Murrell J.R., Trofatter J.A.& Buckler A. J. (1994) Isolation of genes from complex sources of mammalian genomic DNA using exon amplification.Nat.Genet.6:98-105
7. Hochgeschwender U.& Brennan M.B. (1991) Identifying genes within the genome:new ways for finding the needle in a haystack.BioEssays 13:139-144
8. Lovett M. (1994) Fishing for complements:finding genes by direct selection.TIG 10:352-357
9. Patel K., Cox R., Shipley J., Kiely F., Frazer K., Cox D.R., Lehrach H.& Sheer D. (1991) A novel and rapid method for isolating sequences adjacent to rare cutting sites and their use in physical mapping.Nucleic Acids Res.19:4371-4375
10. Hochgeschwender U. (1992) Toward a transcriptional map of the human genome.TIG 8:41-44
11. Valdes J.M., Tagle D.A.& Collins F.S. (1994) Island rescue PCR:A rapid and efficient method for isolating transcribed sequences from yeast artificial chromosomes and cosmids.Proc.Natl.Acad.Sci.USA 91:5377-5381
12. John M.D., Robbins C.A.& Myers M. (1994) Identification of genes within CpG-enriched DNA from human chromosome 4p16.3. Hum. Mol. Genet. 3:1611-1616

Agradecimientos: **A Xavier Lacasta por su valiosa colaboración en la elaboración de las figuras**