Universidad de Granada
Departamento de Genética

Dinámica evolutiva de los retrotransposones Alu en el genoma humano

MEMORIA PRESENTADA POR
MICHAEL HACKENBERG
PARA OPTAR AL GRADO DE DOCTOR EN BIOLOGÍA

Editor: Editorial de la Universidad de Granada
Autor: Michael Hackenberg
D.L.: Gr. 1079 - 2005
ISBN: 84-338-3483-5

José L. Oliver Jiménez, Profesor Titular del Departamento de Genética de la Universidad de Granada

Informa:

Que el trabajo titulado:

"Dinámica evolutiva
de los retrotransposones Alu
en el genoma humano",

que presenta Michael Hackenberg para la obtención del grado de Doctor en Biología, ha sido realizado bajo su dirección.

Fdo: José L. Oliver Jiménez
El Director de Tesis

Fdo: Michael Hackenberg
El Doctorando

Abreviaturas y Glosario

#/10 kb: Número de elementos por 10 kb.

bp (base pairs): Pares de bases, en general el número de nucleótidos en una cadena de ADN.

Contenido en G+C (%G+C): Fracción molar (porcentaje) de guanina y citosina (G+C).

Clusterización: Fenómeno por el que unos elementos distribuidos en el espacio se aglomeran y forman grumos ("clusters").

kb: Mil pares de bases.

LHGR (Long Homogeneous Genomic Regions): Regiones genómicas largas y composicionalmente homogéneas que predice el programa IsoFinder.

NND (Nearest Neighbour Distance): Distancia al elemento más cercano, por ejemplo entre Alus, genes etc.

IPE: Inserción de un elemento transponible dentro de otro elemento preexistente (**I**nsertion into a **P**reexisting **E**lement).

Isocora: Segmentos largos de ADN (>>300 kb) 'relativamente' homogéneos composicionalmente.

MYA: Million Years Ago (millones de años)

NC (Nivel de clusterización): Desviación estándar de la distribución normalizada de las distancias físicas entre los elementos.

RHD: Recombinación Homóloga Desigual

TE (Transposable Element): Elemento transponible; cualquier secuencia de ADN con capacidad de cambiar de posición en el genoma.

Transición: Mutación de una purina (Adenina, Guanina) a otra purina o de una pirimidina (Timina, Citosina) a otra pirimidina (A←→G; C←→T).

Transversión: Mutación de de una purina a una pirimidina o viceversa.

Resumen

Los retrotransposones Alu son los elementos transponibles más abundantes en el genoma humano. Desde su descubrimiento, se viene proponiendo que pueden haber jugado un papel relevante en la evolución de los primates. Una característica destacada de las Alus es su distribución diferencial a lo largo de los cromosomas en función de su edad evolutiva: las Alus más jóvenes muestran una mayor densidad en regiones ricas en A+T, mientras que las más antiguas son mas densas en regiones ricas en G+C. El objetivo principal de esta tesis es analizar la dinámica evolutiva de las Alus en el genoma humano e identificar los principales factores implicados en su distribución espacial a lo largo de los cromosomas.

Para lograr este objetivo, se han implementado en primer lugar una serie de mejoras metodológicas: 1) desarrollo de una colección de programas en Perl para procesar los datos primarios generados por RepeatMasker e IsoFinder; 2) corrección para sustituciones múltiples, aplicada aquí por primera vez a los alineamientos de Alus, con objeto de estimar con mayor precisión la edad evolutiva; 3) desarrollo de una técnica de desfragmentación de Alus para eliminar la habitual sobreestimación de estos elementos y aproximar mejor su número real en el genoma; y 4) definición precisa, mediante técnicas *in silico*, de los trímeros de Alus y de las Alus solitarias, lo que ha permitido rastrear el genoma en busca de las huellas dejadas por la recombinación.

Se descartan a continuación algunos de los mecanismos propuestos para explicar el cambio de densidad: la interacción Alu/LINE1 (competencia por la retrotransposasa), el ajuste composicional, y la selección positiva (acumulación preferente en el entorno de los genes).

Posteriormente, se presentan resultados que avalan la implicación de la recombinación en el cambio de densidad. Esto ha sido posible mediante la definición *in silico* del trímero resultante de un proceso de recombinación homóloga desigual (RHD) entre Alus, y la simulación del proceso de inserción en el ordenador, lo que ha permitido rastrear el genoma en busca de las huellas dejadas por la recombinación y cuantificar su impacto en el cambio de densidad. Se ha observado una proporción más alta de trímeros en las isocoras L, lo que indica una

mayor actividad RHD, o una mayor supervivencia de sus productos en estas isocoras. Esto sugiere que la RHD podría contribuir al cambio de densidad de las Alus. El análisis de la edad evolutiva de las subfamilias de inserción reciente y la mayor frecuencia de Alus solitarias señalan también a la recombinación como agente causante del cambio de densidad, ya que es el único mecanismo cuya actividad depende de la edad de los elementos.

Se ha estudiado igualmente el proceso de aglomeración (o clusterización) de las Alus a lo largo de los cromosomas. Se empieza proponiendo una nueva medida, independiente de la densidad, para cuantificar la clusterización en la distribución espacial de las Alus. A continuación, mediante simulación del proceso de inserción, se asocian niveles de significación estadística (valores-p) a las frecuencias de clusterización observadas. En tercer lugar, la incorporación en los modelos de inserción de diferentes mecanismos evolutivos, ha permitido evaluar el impacto de cada uno de ellos en la generación y mantenimiento de la clusterización.

Se ha hecho también un análisis de la clusterización de las dianas de inserción, habiéndose comprobado que este proceso no puede explicar la fuerte acumulación preferencial en el entorno Alu que se observa en las isocoras H. Sin embargo, la comparación entre las frecuencias de dianas de inserción y de Alus en los flancos de otras Alus preexistentes ha mostrado que la inserción preferencial en la cola de poli-A de otras Alus preexistentes puede contribuir significativamente a la clusterización marcada de las Alus en las isocoras H. Por último, las inserciones en elementos preexistentes, podrían contribuir igualmente a la generación de la mayor clusterización observada en estas isocoras.

Índice

Capítulo 1

Introducción

1.1 Historia breve de la (retro)transposición

A finales de los años 40, Barbara McClintock identificó dos elementos, causantes de mutaciones en maíz, a los que llamó Ac (Activador) y Ds (Disociador), ambos localizados en el cromosoma 9 (McClintock, 1946 y 1947). Lo impactante de su descubrimiento fue que estos dos elementos ocasionalmente aparecieron en sitios distintos dentro del mismo cromosoma (McClintock, 1948 y 1949). Ella acuñó el término "transposición" para referirse al fenómeno de que al parecer existían unidades genéticas en el genoma del maíz con la capacidad de cambiar de posición. No obstante, los trabajos de Barbara McClintock al principio tenían poca repercusión sobre la forma de entender los genomas, debido a la influencia de la genética mendeliana que postulaba posiciones de genes fijas, y en consecuencia han sido interpretados más bien como una excepción.

Pese a que Barbara McClintock descubrió una segunda pareja de elementos transponibles al principio de los años 50 (sistema supresor-mutador; McClintock, 1951 y 1954), pasaron casi veinte años hasta que en los años 70 los investigadores cobraron conciencia de que estaban ante un fenómeno bastante universal. Las primeras evidencias que avalaban la idea de un genoma dinámico venían de la mano de sistemas genéticos como los virus bacterianos con su capacidad de insertarse en distintos genomas, o de los genes responsables de la

resistencia a antibióticos en bacterias. Además, los trabajos de Susumu Tonegawa en los años 70 sobre el sistema inmunológico mostraron que la gran variedad de los anticuerpos tiene su origen en el reensamblaje de centenares de fragmentos de los genes correspondientes, un procedimiento que permite fabricar centenares de millares de proteínas distintas sin que exista el correspondiente número de genes en el genoma. Es decir, solamente un genoma dinámico puede fabricar esta cantidad de proteínas distintas sin que todo su genoma sea ocupado por genes de anticuerpos.

Desde estos tímidos comienzos, los elementos transponibles han sido detectados en innumerables especies, tanto en organismos procariotas como eucariotas. El contenido en elementos transponibles varía notablemente entre las especies, desde genomas muy compactos como el de *Arabidopsis thaliana* hasta genomas con más del 80% de ADN repetido, como en el caso de *Psilotum nudum* (Lesk, 2005). Esta observación parecía explicar las discrepancias que se observan a menudo entre la complejidad del organismo y el tamaño de su genoma, conocido como paradoja del valor-C. Básicamente, el valor-C es la cantidad total de ADN de un genoma diploide y se esperaba que correlacionara positivamente con la complejidad del organismo. Es decir, cuanto más complejo sea un organismo, más ADN funcional necesitará, lo que debería tener repercusión en el tamaño de su genoma. Sin embargo, hay plantas muy simples, como *Psilotum nudum*, (sin verdaderas hojas, pétalos o frutos) que tienen un genoma 3000 veces más grande que otras aparentemente más complejas, como *Arabidopsis thaliana*. Así que aunque el contenido diferencial en elementos transponibles puede explicar en parte la gran variedad de valores-C, surgen otras cuestiones acerca de las fuerzas evolutivas que regulan el contenido en elementos transponibles. Una de las primeras teorías interpretaba los TEs como "ADN basura" que no desempeña ninguna función en el genoma y que se fija mediante deriva génica (Ohno, 1972). Versiones más recientes de esta teoría hablan de "ADN egoísta", interpretando los TEs como parásitos con el único objetivo de su propia proliferación. Según estas hipótesis la selección purificadora a menudo no es lo suficientemente fuerte para impedir la acumulación de este tipo de ADN (Orgel y Crick, 1980; Doolittle y Sapienza, 1980). Por otro lado, existen las teorías adaptativas que proponen que este tipo de "ADN extra" es tolerado en el genoma debido a sus ocasionales efectos positivos y el potencial de exaptación. En los últimos años, se han encontrado muchos ejemplos que parecen avalar esta hipótesis. Se estima que aproximadamente el 4% de los genes del genoma humano poseen un casete-TE (TE-cassette), lo que indica que a menudo los genomas utilizan los TEs para generar nuevos elementos funcionales (Nekrutenko y Li, 2001; Lorenc y Makalowski, 2003). También se tiene

conocimiento de casos en los que los elementos transponibles asumen funciones en la regulación de los niveles de expresión (Britten, 1996). Otra propiedad de los TEs que se ha relacionado con sus posibles funciones, es su distribución a lo largo de los cromosomas. Con la finalización del proyecto de secuenciación del genoma humano, se abrieron nuevas posibilidades para este tipo de análisis.

1.2 Elementos transponibles en el genoma humano

Un elemento transponible o móvil es una secuencia de nucleótidos con la capacidad de moverse en el genoma, ya sea por sí misma o con la ayuda de secuencias auxiliares. Casi la mitad del genoma humano[1] está compuesto por cuatro clases de elementos transponibles: (1) elementos cortos de ADN disperso (SINE - Short Interspersed Nucleic Element), (2) elementos largos de ADN disperso (LINE – Long Interspersed Nucleic Element), (3) elementos con repeticiones terminales largas (elementos LTR) y (4) transposones de ADN (IHGSC, 2001; Li *et al*., 2001). Los elementos más importantes probablemente son las Alus (SINE) y los LINE1 (LINE), tanto por su frecuencia e impacto evolutivo como porque ambos siguen estando activos en el genoma humano.

1.2.1 El proceso de retrotransposición

Tanto las Alus como los elementos LINE1 son retrotransposones, es decir, su replicación se produce mediante la transcripción inversa de un ARN intermediario (Rogers, 1983). Para transcribirse a un ARN intermediario usan recursos celulares como la polimerasa III (las Alus) y la polimerasa II (los LINE1). El proceso de transcripción inversa a ADN e inserción en el genoma es más complicado. En general se requieren dos enzimas: una transcriptasa inversa y una endonucleasa encargada del corte de la secuencia donde se produce la inserción. Los elementos LINE1 poseen dos marcos abiertos de lectura, ORF1 y ORF2. Se ha podido demostrar que el ORF2 codifica una transcriptasa inversa que además posee un dominio de endonucleasa (Mathias *et al*., 1991; Feng *et al*., 1996; Cost *et al*., 1998). Aunque todavía no se entiende completamente el papel que desempeña la proteína del ORF1, parece que su propensión a fijar expresamente el ARN mensajero del elemento LINE1 es imprescindible para su transposición (Martin y Bushman, 2001; Kolosha y Martin, 2003). Los elementos LINE1 son por lo tanto autónomos, ya que codifican todas las proteínas necesarias para su propia reproducción. Por el contrario, las Alus no codifican

[1] Es virtualmente imposible dar números exactos puesto que probablemente muchos elementos ya no son reconocibles como tales.

ninguna proteína y por lo tanto dependen de enzimas procedentes de otros elementos. Hay varias evidencias que demuestran que las Alus utilizan la maquinaria enzimática de los elementos LINE1 (*in silico*: Jurka, 1997; experimental: Dewannieux *et al*., 2003). En la Figura 1-1 se puede ver un esquema del proceso de retrotransposición.

Dewannieux *et al*. (2003) diseñaron el primer experimento para seguir el proceso de retrotransposición mediante Alus marcadas. En primer lugar, verificaron que la presencia de la proteína del ORF2 es crucial para la replicación de las Alus. Sin embargo, no encontraron evidencia de que lo mismo valiera para la proteína del ORF1. Parece que aunque esta proteína desempeña un papel crucial en la retrotransposición de los propios LINE1, las Alus no requieren su presencia. No obstante, las Alus también necesitan proteínas de funcionalidad similar a la ORF1 (chaperona). Estas proteínas resultaron ser las SRP9/14 del SRP (partícula de reconocimiento de señal). Debido a su origen evolutivo (véase sección 1.2.2.1 El origen evolutivo y la estructura física de las Alus), las Alus muestran una gran afinidad hacia estas proteínas, que utilizan para guiarse hacia los ribosomas (Bovia y Strub, 1996; Bovia *et al*., 1997). Según este modelo, la maquinaria enzimática de los LINE1 no reconoce *per se* a las Alus, sino que éstas, mediante las proteínas SRP9/14, se unen a las proteínas de los LINE1 en los ribosomas (efecto en "*cis*"). Evidencia adicional para este modelo es el hecho de que las Alus con deleciones internas muestran tasas de retrotransposición menores debido a la estructura secundaria alterada por las deleciones.

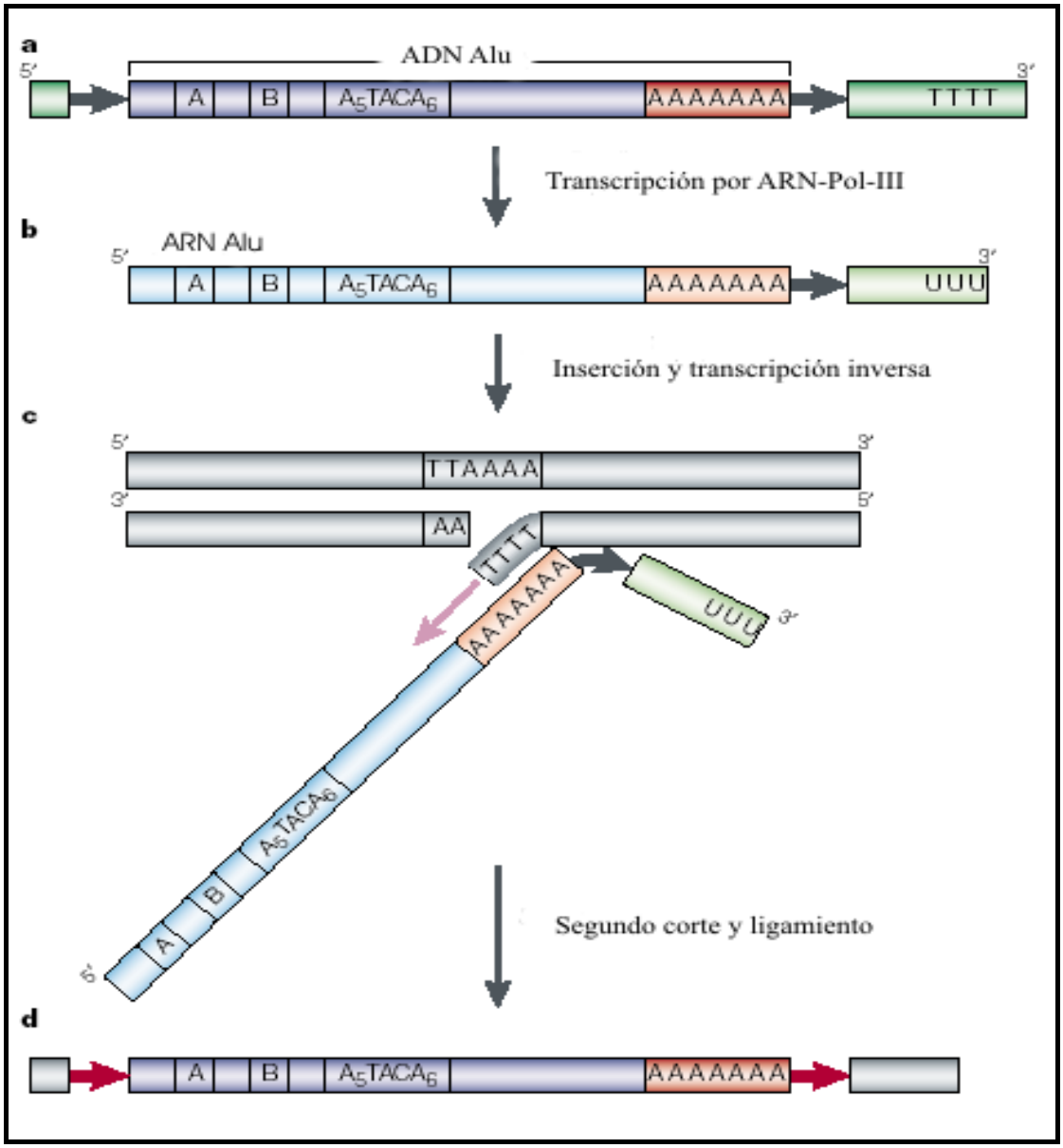

Figura 1-1: Un esquema del proceso de retrotransposición de las Alus. Aparte del primer paso, que consiste en la transcripción de las Alus mediante ARN-pol-III (transcripción mediante ARN-pol-II en el caso de los LINE1), este modelo se puede aplicar también a los elementos LINE1. El segundo paso, muestra la transcripción reversa "*in situ*" (TPRT: Target-Primed Reverse Transcription; Lin y Levin, 1997). La endonucleasa corta la secuencia (preferentemente en 3'-AA/TTTT-5') y el ARNm del elemento se fija con su cola de poli-A en el extremo 3' del corte abierto, el cual debe de ser rico en T. Después de la transcripción reversa, se produce un segundo corte en la hebra contraria y el ligamiento que cierra el ADN (ilustración adaptada de Batzer y Deininger, 2002).

1.2.2 Las Alus

Las Alus surgieron hace aproximadamente 65 millones de años y por lo tanto limitan su presencia a los primates (Deininger y Daniels, 1986). El nombre deriva de un patrón de reconocimiento de una enzima de restricción llamada AluI (Houck *et al*., 1979). Se estima que hay más de 1.300.000 copias Alu, con lo que ocupan un 11% del genoma humano (Li *et al*., 2001). Con una longitud de unos 300 bp, las Alus son elementos relativamente cortos. Además poseen un contenido alto en G+C (en torno al 57%) y son ricas en dinucleótidos CpG. Se estima que un tercio de todos los CpG en el genoma se encuentran en Alus (Stoppa-Lyonnet *et al*., 1990; Hellmann-Blumberg *et al*., 1993).

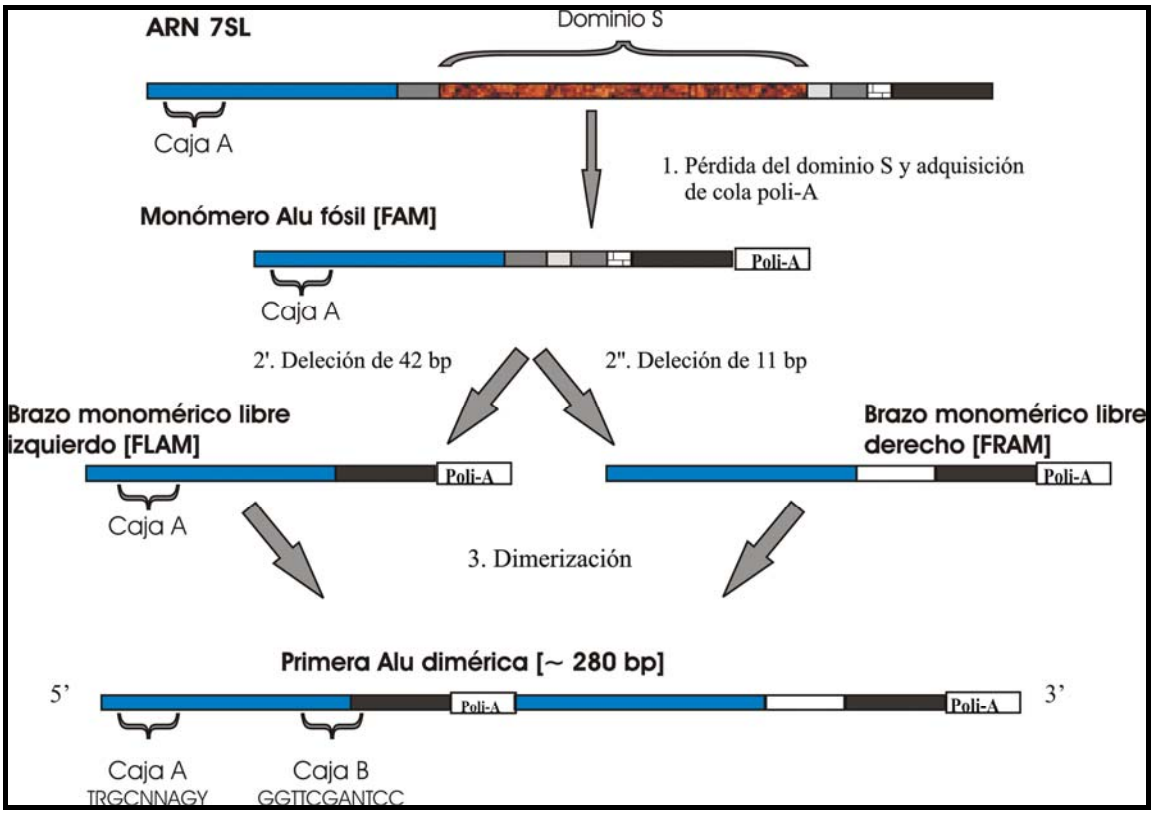

Figura 1-2: Modelo de evolución del gen 7SL hacia la primera Alu dimérica. La evolución ocurrió en tres pasos: 1) pérdida del dominio S del gen 7SL y adquisición de una cola de poli-A, 2) deleciones de 42 bp y 11 bp, respectivamente, que dan lugar a los precursores monoméricos de las Alus, FLAM y FRAM, y 3) la dimerización de FLAM y FRAM. La caja A de las Alus se hereda directamente del gen 7SL, el cual se mantuvo solo en el FLAM y perdió su funcionalidad en el FRAM. El FLAM además adquirió una caja B funcional.

1.2.2.1 El origen evolutivo y la estructura física de las Alus

Una secuencia Alu está compuesta por dos monómeros unidos por un trecho oligo-d(A). Debido a su origen común, los dos monómeros son similares pero no idénticos. Las Alus descienden de un gen 7SL ARN que forma parte del complejo ribosómico (Ullu y Tschudi, 1984). Se supone que una duplicación de este gen en los comienzos de la radiación de los primates y su posterior evolución en varios pasos dio finalmente lugar a las Alus "modernas" (véase Figura 1-2; Quentin, 1992; Batzer y Deininger, 2002; Mighell *et al*., 1997). El monómero izquierdo contiene un promotor de ARN polimerasa III (caja A y B) que no está presente en el monómero derecho (Willis, 1993). La secuencia de las Alus no contiene ningún terminador pol III $d(T)_4$ de ARN, sin embargo este se encuentra a menudo en el flanco 3' de la secuencia genómica debido a la terminación poli-(A) de los transcritos de Alu. Es digno de mencionar que todos los elementos del tipo SINE tienen su origen en genes estructurales de ARN (Deininger y Batzer, 1993; Shedlock y Okada, 2000; Ohshima *et al*., 1996; Ohshima y Okada, 1994; Okada y Hamada, 1997; Okada y Ohshima, 1993).

1.2.2.2 La clasificación de las Alus

La presencia del promotor y la disponibilidad de transcriptasa inversa no son suficientes para que se transcriba o transponga una Alu. Aunque el mecanismo exacto de retrotransposición todavía no se entiende en todos sus detalles, y por lo tanto se desconocen todos los prerrequisitos que deben cumplirse, se sabe que las secuencias flanqueadoras, la estructura secundaria, el estado de metilación y la longitud del trecho de poli-(A) son cruciales para el éxito del proceso de retrotransposición (Batzer y Deininger, 2002; Dewannieux *et al.*, 2003; Roy-Engel *et al*., 2002). En consecuencia, se cree que solamente muy pocas Alus en el genoma tienen la capacidad de transponerse, es decir hacer copias de sí mismas y diseminarlas por el genoma. Este modelo se llama del "gen maestro", y constituye la base para la clasificación de las Alus (Deininger *et al*., 1992). Una mutación en un gen maestro se llama mutación diagnóstica y se encuentra la misma mutación en toda su descendencia. Estas mutaciones permiten la definición de los grupos y (sub)familias de Alus. La Figura 1-3 muestra un alineamiento múltiple de varias secuencias consenso de Alus que sugiere la evolución a partir de un gen maestro. Las edades de las distintas familias se pueden estimar a partir de las distancias evolutivas entre los miembros de cada familia en el genoma (Kapitonov y Jurka, 1995; Batzer *et al*., 1996). A raíz de lo anterior, se divide a las Alus en 3 (sub)grupos: AluJ (más antiguas), AluS (de edad intermedia) y AluY (las más recientes). Los grupos a su vez se componen de varias (sub)familias. En la Figura 1-4 se muestra la filogenia de las diferentes

subfamilias de Alus (véase también el Apéndice C para una estadística básica de las subfamilias analizadas en este trabajo).

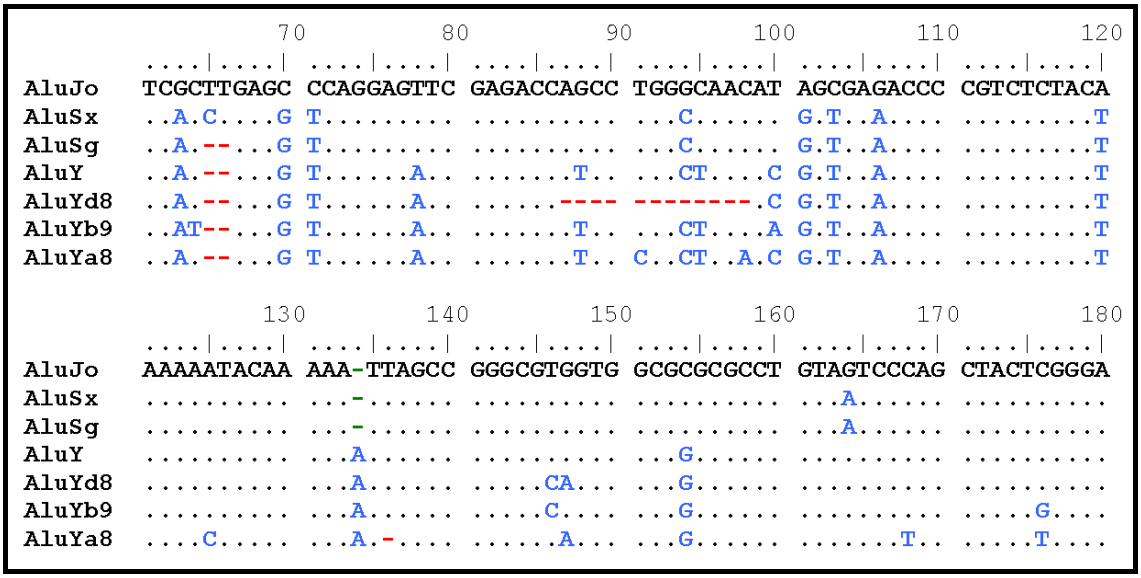

Figura 1-3: Alineamiento múltiple de varias subfamilias de Alus. Se puede observar la evolución de la secuencia AluJo (la más antigua) hacia subfamilias más recientes mediante mutaciones diagnósticas (azul), inserciones (guiones rojos) y deleciones (guiones verdes). Existen mutaciones que se observan en todas las subfamilias descendientes. No obstante, algunas mutaciones se limitan a determinadas subfamilias. La subfamilia AluYd8 es más antigua que Yb9 e Ya8, pero muestra una deleción de 12 bp que no está presente en Yb9 e Ya8. Esto demuestra que la evolución de los genes maestros no es lineal, sino que se forman ramas laterales, indicando así que puede haber más de una subfamilia activa.

Las Alus son todavía activas en el genoma humano, si bien con tasas de replicación muy bajas comparadas con las que se producían hace 40 millones de años, cuando las Alus hicieron su 'despliegue' más pronunciado. Se estima que hoy día ocurre solamente una inserción cada 200 nacimientos, mientras que en el clímax de su actividad replicadora había aproximadamente una inserción por nacimiento (Shen *et al*., 1991; Deininger y Batzer, 1999). Este cambio de dos órdenes de magnitud todavía no tiene una explicación generalmente aceptada. Se estima que varios factores contribuyen a la disminución de la actividad de las Alus jóvenes. Primero, hace 40 millones de años, tres subfamilias LINE1 estaban activas, lo cual probablemente aportara una mayor cantidad de retrotransposasa "libre" (Ohshima *et al*., 2003). Segundo, las Alus mantienen la estructura secundaria del dominio Alu del gen 7SL. Este hecho les permite fijar las proteínas SRP9/14, las cuales supuestamente desempeñan un papel importante en la replicación al dirigir las Alus hacia los ribosomas (Dewannieux *et al*., 2003;

Boeke, 1997). Hay evidencia de que la capacidad de fijar dichas proteínas ha ido disminuyendo durante la evolución de los primates, lo que probablemente ha conducido a una menor actividad de retrotransposición de las Alus jóvenes (Sarrowa *et al*., 1997; Fan *et al*., 1998).

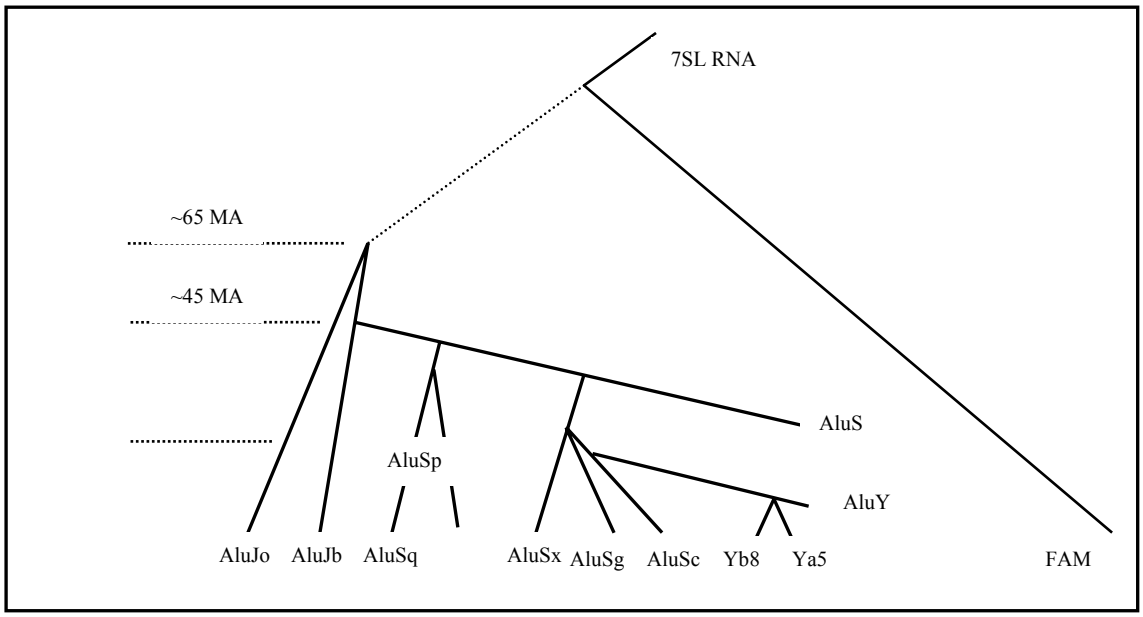

Figura 1-4: Filogenia de las Alus.

1.2.3 Los elementos LINE1

Al contrario que las Alus, los elementos LINE1 no se limitan a los primates, sino que estaban activos ya antes de que se produjera la división entre marsupiales y euterios (Smit, 1995). Poseen una longitud entre 6000 y 8000 bp, sin embargo en el genoma la gran mayoría de los elementos está truncado en su extremo 5', reduciendo así la longitud media a unos 900 bp (véase Tabla 4-2). El número estimado de estos elementos es de aproximadamente 866.600 copias, lo que se traduce en una ocupación en torno al 15.4 % del genoma (IHGSC, 2001, Li *et al*., 2000). Aparte de los dos marcos abiertos de lectura (ORF1 y ORF2), los LINE1 cuentan con regiones no traducidas 5' y 3' y al menos algunos de ellos tienen un promotor interno de polimerasa II (Smit, 1995). La clase de los LINE se incluye en los retrotranspones no-LTR. Suponiendo transferencia vertical, y mediante filogenias basadas en proteínas, se puede dividir esta clase en 11 clados, siendo posiblemente el clado LINE1 el más importante (Malik *et al*., 1999, Smit 1999). El clado de los LINE1 se divide en 47 subfamilias, utilizando para ello las regiones no traducidas 3'-UTR (Smit, 1995). Sólo una de estas subfamilias

(L1Hs) sigue estando activa en el genoma, con un número de elementos entre 30 y 60 (Sassaman *et al*., 1997).

La maquinaria enzimática de transposición de los elementos LINE1 no sólo impulsa la expansión de los propios LINE1, sino que parece estar también involucrada en la aparición de los SINE y los pseudogenes procesados (Esnault *et al*., 2000, Wei *et al*., 2001). Además, se ha observado que los elementos LINE1 promueven el "barajamiento de exones" (exon shuffling, Moran *et al*., 1999) y participan en la reparación de roturas en ambas hebras del ADN (Morrish *et al*., 2002).

1.2.4 Impacto evolutivo de los elementos transponibles

Varios procesos participan en la evolución de los genomas eucariotas, tales como roturas cromosómicas y su translocación posterior a otro cromosoma, duplicaciones de genes y segmentos, reordenación de dominios funcionales como los exones, creación de nuevos genes y conversión génica. En las ultimas dos décadas se ha ido tomando conciencia de que en muchos de estos procesos se ven involucrados elementos transponibles (Kazazian, 2004; Deininger *et al*., 2003; Prak y Kazazian, 2000).

1.2.4.1 TEs – fuente de nuevas estructuras funcionales

Desde hace un tiempo se sabe que en algunas ocasiones los mecanismos celulares 'domestican' a ciertos elementos transponibles, dando lugar así a genes nuevos (Makalowski *et al*., 1994; Britten, 1997; Brosius, 1999; Smit, 1999; Lorenc y Makalowski, 2003; Britten, 2004). Al principio se sospechaba que se trataba de acontecimientos aislados, pero en los últimos años muchos trabajos han demostrado que los TEs pueden promover la variación y diversificación génica de varias maneras distintas. Así, se sabe que más del 5% de los exones internos con splicing alternativo provienen de una secuencia Alu (Sorek *et al*., 2002, Lev-Maor *et al*., 2003). Además, se han identificado 7810 Alus intrónicas con posibilidad de ser exonizadas, lo que constituye un gran potencial para generar genes nuevos (Sorek *et al*., 2004). Otra vía de estudio tiene que ver con la regulación de la expresión génica, la cual se ve afectada a menudo por la presencia de TEs (Brosius, 1999; Hamdi *et al*., 2000; Nigumann *et al*., 2002, Medstrand *et al*., 2001; Landry *et al*., 2001). Además, en un trabajo reciente (Lagemaat *et al*., 2003) se demuestra que hay un exceso de TEs en los transcritos de genes con origen evolutivo reciente, mientras que se ven excluidos de los ARNm de genes con funciones básicas en el desarrollo o el metabolismo. Esto

apoya la suposición de que los TEs desempeñan un papel importante en la diversificación y evolución de los genes en mamíferos.

1.2.4.2 Recombinación – fuerza del "rediseño" genómico

Aparte del impacto directo que ejerce la transposición en el genoma humano, los elementos transponibles también aumentan la tasa de recombinación. Las distintas copias de un elemento transponible suelen compartir una alta similitud de secuencia entre sí, lo cual es un prerrequisito importante para que la recombinación pueda tener lugar (Lambert *et al*., 1999; Babcock *et al*., 2003). La recombinación puede producirse tanto entre dos elementos ubicados en el mismo cromosoma (recombinación intracromosómica) como entre dos secuencias localizadas en cromosomas distintos (intercromosómica). Estos procesos de recombinación pueden provocar duplicaciones, inversiones, deleciones o translocaciones en el genoma (Kazazian y Goodier, 2002; Gilbert *et al*., 2002; Symer *et al*., 2002; Hollies *et al*., 2001; Elliott *et al*., 2005). Muchos de los productos de recombinación se eliminan por selección purificadora; sin embargo, en casos aislados este rediseño se mantiene, brindando variación y posiblemente nueva funcionalidad al genoma.

Las Alus se muestran especialmente proclives a participar en procesos de recombinación. Así, se ha podido demostrar que a menudo una sub-secuencia Alu de 26 bp se encuentra cerca de los puntos de rotura de translocación (Rudiger *et al*., 1995). Además, esta sub-secuencia muestra similitud con la secuencia chi (χ) de *E.coli*, de la cual se sabe que estimula la recombinación. Recientemente, también se ha propuesto la implicación de las Alus en la formación de las duplicaciones segmentales (Bailey *et al*., 2003; Zhou y Mishra, 2005; Abeysinghe *et al*., 2003). Un 27% de éstas terminan en alguno de sus extremos con una secuencia Alu. Debido a la alta frecuencia de recombinación entre Alus, este mecanismo también tiene un impacto notable en la aparición de enfermedades genéticas (Rossetti *et al*., 2004; Mitchell *et al*., 2004). Se estima que un 0.3% tienen su origen en fenómenos de recombinación entre Alus espaciadas (Deininger y Batzer, 1999).

Aunque los TEs pueden propiciar la recombinación, mientras un elemento es polimórfico puede ocurrir justamente lo contrario. Por ejemplo, una inserción reciente no cuenta con una pareja en el otro cromosoma, lo que da lugar a una reducida similitud composicional local entre los dos cromosomas homólogos. Deininger y Batzer (2002) argumentan que este hecho puede influir en las tasas de recombinación meiótica y por lo tanto afectar al desequilibrio de ligamiento

dentro de una población. Según estos autores, esta limitación puede acelerar el proceso de especiación.

1.3 Isocoras en el genoma humano

El genoma humano muestra una alta heterogeneidad espacial respecto a su contenido en G+C, que varía entre aproximadamente el 35 y el 60%. Estas estructuras en mosaico se llaman isocoras, que son trechos largos de ADN (>>300 kb) 'relativamente' homogéneos composicionalmente. Las isocoras se descubrieron originalmente mediante ultracentrifugación analítica de ADN total (Corneo *et al*., 1968; Bernardi *et al*., 1985), y su relevancia se debe al hecho de que tanto la densidad de genes y de elementos transponibles como las tasas de recombinación y mutación varían notablemente en función de la isocora (Bernardi, 2000; Wolfe *et al*., 1989; Oliver *et al*., 2002). Nótese que todos los genomas de mamíferos muestran estas estructuras y que todavía no hay una teoría generalmente aceptada sobre su origen y mantenimiento (Bernardi, 2001; Eyre-Walker y Hurst, 2001; Belle *et al*., 2004).

1.3.1 Segmentación composicional del genoma humano

La disponibilidad de secuencias genómicas completas, permitió el desarrollo de métodos para identificar isocoras. Los primeros intentos utilizaban ventanas móviles para calcular el G+C local de la secuencia. Sin embargo, dada la complejidad de la heterogeneidad composicional del genoma (Román-Roldán *et al*., 1998), esta aproximación puede llevar a resultados inconsistentes (Bernaola-Galván *et al*., 1996; Li, 2001; Oliver *et al*., 2001, 2002).

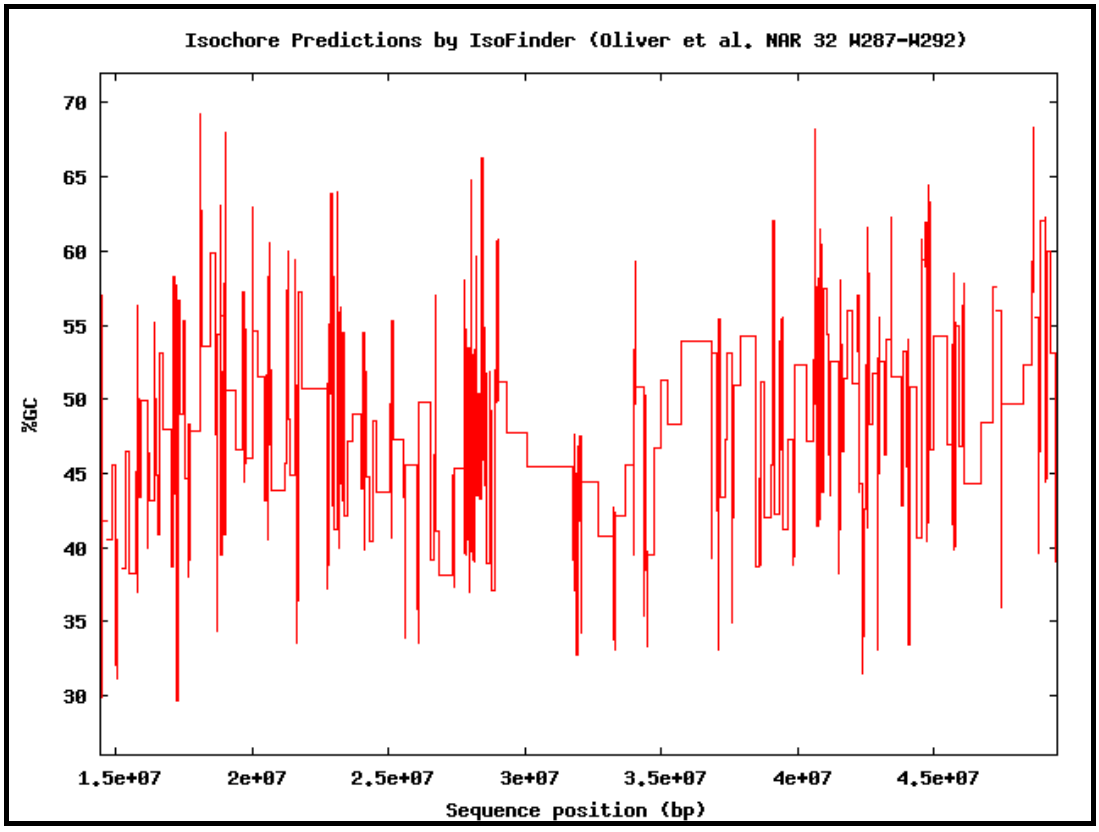

Figura 1-5: Representación grafica de la segmentación del cromosoma 22 generada mediante la interfaz web del programa IsoFinder (Oliver *et al.*, 2004). Se puede apreciar la variación pronunciada del contenido en G+C entre aproximadamente el 35 y el 65%.

En este trabajo se ha utilizado una versión mejorada del algoritmo IsoFinder (Oliver *et al.*, 2004; http://bioinfo2.ugr.es/IsoF/isofinder.html) basada en una versión previamente descrita (Bernaola-Galván *et al.*, 1996; Oliver *et al.*, 2001, 2002). El programa IsoFinder permite predecir regiones largas y homogéneas en el genoma (LHGR - Long Homogeneous Genome Regions), muchas de las cuales se pueden asociar a isocoras de Bernardi (Oliver *et al.*, 2001, 2002). En la Figura 1-5 se muestra una representación grafica de las regiones homogéneas generadas mediante el programa IsoFinder. Se puede apreciar la notable fluctuación del contenido en G+C a lo largo del cromosoma.

1.3.2 Clasificación de las isocoras

Tradicionalmente, según su abundancia, las isocoras se dividen en 4 clases. Las isocoras L, ricas en A+T, que representan el 66.3% del genoma (L por ligero), y las isocoras H (H1, H2 y H3), ricas en G+C, que ocupan el 33.7% restante (H por

heavy o pesado). Es decir, el valor al que se tiene 'acceso' experimentalmente es la abundancia relativa de cada isocora (véase Figura 1-6; Zoubak *et al*., 1996; Bernardi, 2000). Sin embargo, las abundancias se refieren a la cantidad de ADN, y no especifican su localización en el genoma ni facilitan *per se* los rangos de G+C. En consecuencia, para generar una clasificación de las isocoras aplicable a la secuencia del genoma humano se requieren dos pasos. Primero, hay que descomponer o segmentar el genoma en regiones según su contenido en G+C. Segundo, se determinan los rangos de G+C que permiten asignar las distintas regiones a cada clase de isocora.

Por ejemplo, Pavlicek *et al*. (2002) descomponen el genoma mediante ventanas no-solapantes de 10kb. La asignación de las regiones obtenidas de la descomposición a una clase de isocora la llevan a cabo mediante los rangos expuestos en la Tabla 1-1.

Tabla 1-1: Clasificación de isocoras utilizada por Pavlicek *et al*. (2002)

Isocora	**Rangos de G+C**
L1	G+C < 37%
L2	37% <= G+C < 41%
H1	41% <= G+C < 46%
H2	46% <= G+C < 52%
H3	G+C > 52%

En resumen, lo que se entiende como clasificación de las isocoras es la asignación de una etiqueta (L1, L2, H1, H2 o H3) a una región genómica determinada mediante la introducción de rangos de G+C. Nótese que en el ejemplo expuesto arriba los autores elegían rangos fijos de G+C, y por lo tanto las abundancias obtenidas no necesariamente corresponden a las abundancias experimentales.

Por ello, en este trabajo se introduce una clasificación de isocoras generada a raíz de las abundancias experimentales (véase Figura 1-6), que son la única característica accesible de las isocoras (véase 3.2: Clasificación de las isocoras).

1.3.3 El sesgo de las densidades de genes y TEs

En la práctica, ninguna entidad genómica muestra una distribución uniforme en el genoma, sino que varía casi siempre en función del contenido local en G+C (isocoras). El ejemplo más importante es el de los genes, que son muy escasos en

isocoras L (un gen por cada 50-150 kb) y muy abundantes en isocoras H - un gen por cada 5-15 kb (Bernardi, 2000). Así pues, el 88% del genoma (isocoras L y H1) contiene tan sólo el 46% de los genes, mientras que el 12% restante (isocoras H2 y H3) alberga el 54% de los genes (Bernardi, 2000; Mouchiroud *et al.*, 1991; Zoubak *et al.*, 1996). Bernardi llama a este 12% del genoma 'el núcleo génico', al cual atribuye ciertas características destacadas: (i) densidad muy alta de genes; (ii) genes con alto contenido en G+C, intrones cortos y asociados con islas CpG; (iii) replicación temprana; (iv) alta frecuencia de recombinación; y (v) altos niveles de expresión y estructura abierta de la cromatina.

Como hemos mencionado anteriormente, los elementos transponibles también muestran distribuciones sesgadas en función de la isocora. Mientras que los LINE1 se acumulan en las isocoras L1, la máxima densidad de las Alus se encuentra en las isocoras H2 (Meunier-Rotival *et al.*, 1982; Soriano *et al.*, 1983; Bernardi, 2001).

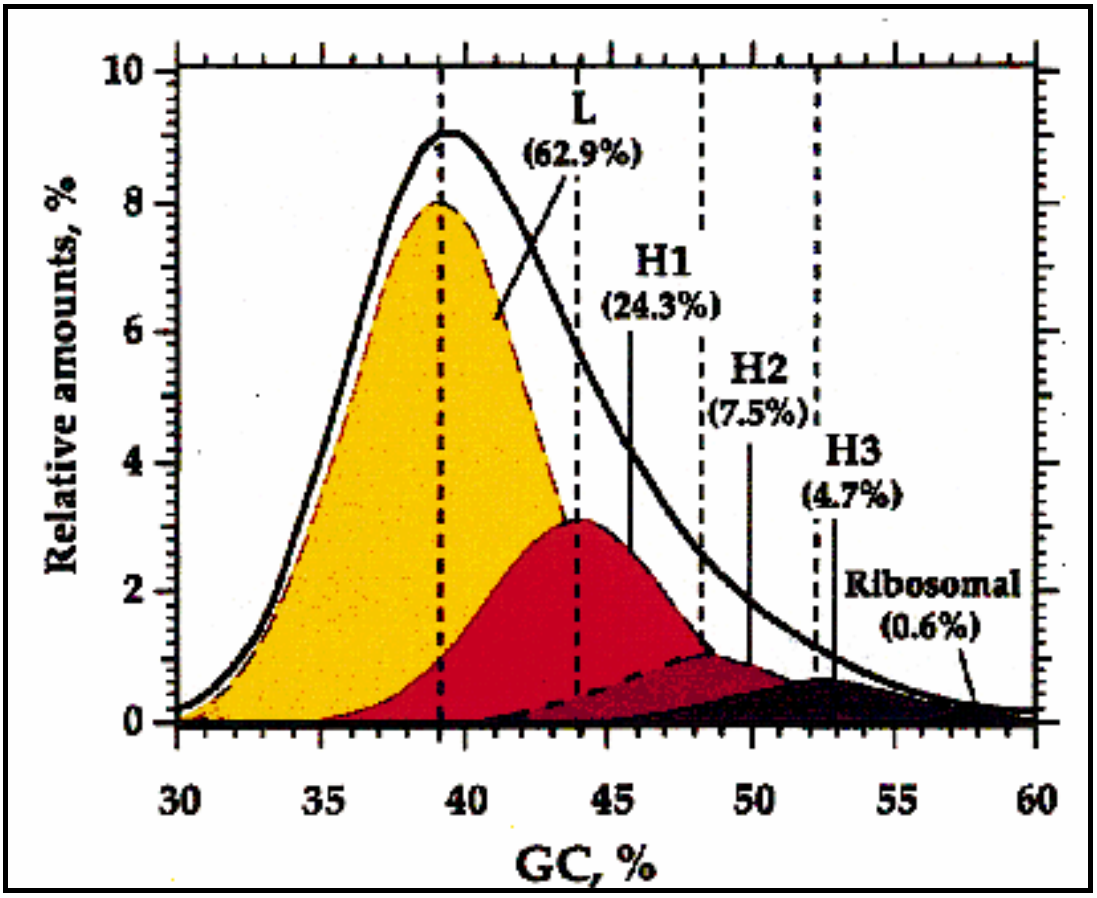

Figura 1-6: Las abundancias experimentales de las isocoras en el genoma humano (ilustración de Bernardi, 2000; Zoubak *et al.*, 1996).

1.3.4 El sesgo de las tasas de mutación

Un lema tradicional de la evolución molecular era que las tasas de mutación puntual eran constantes en el genoma. Por tanto, los cambios observados en las tasas de mutación eran atribuidos a la actuación diferencial de la selección. Sin embargo, Wolfe *et al.* (1989) mostraron que la tasa de mutación silenciosa varía en función del contenido en G+C. Como explicación proponían que la variación en los patrones de mutación se debe a los diferentes tiempos de replicación de las regiones genómicas durante el ciclo celular.

Evidencia adicional para esto se obtuvo por parte del consorcio de secuenciación de la región MHC en el cromosoma 6 humano (The MHC sequencing consortium, 1999). La región MHC (**M**ajor **H**istocompatibility **C**omplex) alberga dos isocoras experimentalmente verificadas y caracterizadas a nivel de secuencia (Fukagawa *et al*., 1995; Stephens *et al*., 1999; The MHC sequencing consortium, 1999). Las regiones "*clase II"* y "*clase III"* corresponden a isocoras L2 y H3, respectivamente. Una observación importante es que la región "*clase II*" se replica más tardíamente en el ciclo celular que la región "*clase III*" (Tenzen *et al*., 1997). La conexión con los patrones de mutación se debe a la disponibilidad diferencial de los desoxirribonucleósidos trifosfatos durante la replicación del ADN. En un modelo simple, se puede suponer que la replicación empieza con una cierta cantidad de dNTP, que se va agotando durante la replicación. Según este modelo, las isocoras ricas en G+C (isocoras H) se replican más temprano en el ciclo celular que las regiones ricas en A+T (isocoras L). La enzima que lleva a cabo la replicación, ADN polimerasa I, introduce aproximadamente un error cada 10^9 bases (después de reparación). Se puede suponer que la probabilidad con la que se introduce una de las otras tres bases (mutación), viene determinada principalmente por su disponibilidad (es decir, por las concentraciones de dATP, dCTP, dGTP y dTTP). Puesto que las isocoras H se replican al principio del ciclo celular, se gastan antes los dCTP y dGTP, lo que aumenta la probabilidad de mutaciones hacia A o T en las isocoras L.

1.4 La polémica post-insercional

Ya hemos mencionado que las Alus no son transposones autónomos, sino que dependen de una transcriptasa/endonucleasa codificada por los elementos LINE1. La diana consenso de dicha enzima es 3'-AA/TTTT-5', efectuándose el corte en la posición de la barra. Estadísticamente, este patrón ocurre con mayor frecuencia en regiones ricas en A+T y por lo tanto la inserción en isocoras L debe de ser

más probable que en H. El máximo de densidad en isocoras L que muestran los elementos LINE1 confirma esta predicción. Trabajos recientes confirman además que las Alu jóvenes también se insertan preferencialmente en zonas ricas en A+T (Pavlicek *et al*., 2001; IHGSC, 2001). Sin embargo, la polémica surge cuando se toman en consideración todas las Alus, independientemente de su edad. En este caso, las Alus muestran su máximo de densidad en las regiones ricas en G+C.

En los últimos años, se han publicado muchas explicaciones sobre la naturaleza del mecanismo que provoca el cambio de densidad. La mayoría de los mecanismos propuestos incluyen la selección tanto positiva como negativa y/o la recombinación.

1.4.1 Selección positiva

La selección positiva podría actuar si las Alus tuvieran alguna función identificable que beneficiase al organismo. En los últimos años, se han propuesto varios efectos positivos (Chu *et al*., 1998; Schmid, 1998; Deininger y Batzer, 1999). Por ejemplo, se ha observado un aumento de la expresión de las Alus en células sometidas a stress. El ARNm de Alu tiene la capacidad de unirse a una proteína kinasa (PKR) y bloquear su capacidad para inhibir la traducción de proteínas (Chu *et al*., 1998; Schmid, 1998). Por lo tanto, el ARNm de Alu promueve la traducción generalizada de proteínas en condiciones de stress celular. Su localización en cromatina abierta, que muestra una correlación positiva con el G+C de la región genómica, podría facilitar esta tarea (Smit, 1999). Otra posible función de las Alus está relacionada con los niveles de expresión de los genes. Las Alus son ricas en el dinucleótido CpG y al insertarse cerca de un gen, la nueva isla de CpG que aporta al entorno génico puede alterar la expresión del gen (Britten, 1996). Como hemos expuesto anteriormente (sección 1.3.3), los genes muestran una mayor densidad en regiones ricas en G+C, y por lo tanto si las Alus tuvieran una distribución similar, esta podría ser beneficiosa.

A pesar de los indicios a favor de la selección positiva, también existen serias objeciones en contra de esta teoría, puesto que la mayoría de las funciones atribuidas a las Alus no se refieren al conjunto de todas ellas sino a elementos particulares (Deininger y Batzer, 1999). Contradicciones adicionales surgen estudiando Alus polimórficas. El argumento es que la fijación de las Alus debería ser mucho más rápida si fuera por selección positiva (Brookfield, 2001).

1.4.2 Selección negativa

La selección negativa también podría provocar un cambio de densidad, siempre que tuviera la capacidad de excluir o sustraer las Alus con mayor eficacia de las isocoras ricas en A+T que de las regiones ricas en G+C. Se han propuesto varios efectos negativos para el organismo sobre los que la selección podría actuar. Un posible efecto negativo se produciría si la acumulación de Alus (que son ricas en G+C) en regiones ricas en A+T dañase severamente la composición local y/o la estructura de la cromatina en estas regiones del genoma, afectando así a la transcripción de los genes (Rynditch *et al*., 1998).

Bernardi y colaboradores propusieron además una teoría de selección negativa por la cual la mayor parte del ajuste composicional se lograría mediante adquisición y pérdida selectiva de ADN, especialmente de ADN repetitivo (Pavlicek *et al*., 2001). Esta hipótesis se apoya en el análisis comparado de genomas de mamíferos (Pavlicek *et al*., 2002; Paces *et al*., 2004).

Por último, se sabe que la densidad de Alus jóvenes en el cromosoma Y es aproximadamente tres veces más alta que en el cromosoma X y dos veces más alta que en los autosomas (Jurka *et al*., 2002). Estos autores interpretan esta distribución inicial de las densidades como consecuencia de que la replicación de estas Alus se produce primariamente en la línea germinal paterna. Sin embargo, la alta densidad inicial de las Alus en el cromosoma Y disminuye rápidamente. Además, Jurka *et al*. (2004) indican que las Alus jóvenes se hallan en regiones menos densas en Alus que las Alus de subfamilias más antiguas. Concluyen que las inserciones de Alus fuera de 'clusters' de Alus son menos estables en términos evolutivos que las inserciones en los 'clusters' preexistentes. En vista de la distribución de Alus jóvenes tanto en función del cromosoma como en función de la densidad de Alus ('clusters'), Jurka *et al*. (2004) proponen un mecanismo de selección paterna que elimina las Alus de sitios cromosómicos no-neutros.

1.4.3 Recombinación Alu/Alu

La recombinación desigual Alu/Alu puede generar deleciones y duplicaciones. Existen dos posibilidades, ya que este mecanismo puede cambiar el número total de Alus. La primera es que la frecuencia de recombinación entre dos Alus sea más alta en isocoras L (Medstrand *et al*., 2002). La segunda se basa en que la desventaja selectiva impuesta por la recombinación Alu/Alu puede ser más alta en isocoras H (Deininger y Batzer, 1999; Batzer y Deininger, 2002). Ambas propuestas suponen que las deleciones ocurren generalmente con mayor

frecuencia que las duplicaciones. Así pues, este mecanismo daría lugar a una disminución del número total de Alus en isocoras L, y por lo tanto a un aumento relativo de las Alus en isocoras H.

1.4.4 Interacción Alu/LINE1

Un mecanismo que no requiere que actúe ni la selección ni la recombinación es la interacción entre Alus y LINE1s. Se argumenta que las Alus podrían haber cambiado su preferencia de inserción hacia regiones ricas en G+C al fin de evitar la competencia por la retrotransposasa (necesaria para reinsertarse en el genoma) con los elementos LINE1 (Gu *et al*., 2000).

Capítulo 2

Objetivos

La finalización del proyecto de secuenciación del genoma humano brinda nuevas posibilidades para estudiar el papel evolutivo de los elementos transponibles en el genoma completo. Este trabajo se centra en los retrotransposones Alu, que se encuentran solamente en primates. Una característica destacada de estos elementos es su inesperada distribución a lo largo de los cromosomas: mientras que las Alus más jóvenes se acumulan en regiones ricas en A+T, las más antiguas lo hacen en las regiones ricas en G+C. Así pues, debe existir algún mecanismo que provoque este cambio de densidad de las Alus después de su inserción en el genoma. El objetivo principal de este trabajo es analizar los mecanismos que regulan la dinámica evolutiva de las Alus y que pueden haber provocado dicho cambio de densidad. Más concretamente, los objetivos son:

1. Analizar el papel de diferentes mecanismos evolutivos —selección, interacción Alu/LINE1, recombinación, ajuste composicional y localización en el entorno génico— en el cambio de densidad experimentado por las Alus a lo largo de su historia evolutiva.

2. Demostrar la implicación de la recombinación desigual (Alu/Alu) en el cambio de densidad. Esto requiere detectar los productos de recombinación en el genoma mediante técnicas *in silico*. Así mismo, es necesario simular

el proceso de inserción de Alus en el genoma, con objeto de asociar una significación estadística a las frecuencias observadas.

3. Determinar si existe clusterización (aglomeración) en la distribución espacial de Alus a lo largo de los cromosomas. Como se menciona más arriba, la densidad de Alus varía mucho entre diferentes regiones del genoma. Sin embargo, todas las medidas de clusterización empleadas hasta ahora son dependientes de la densidad. En este trabajo trataremos de desarrollar una medida que sea independiente de la densidad y que permita, por tanto, estudiar la variación de la clusterización de Alus en distintas regiones genómicas. Como en el punto anterior, será necesario también simular el proceso de inserción de Alus con objeto de poder asociar una significación estadística a los niveles de clusterización observados. Además, la simulación se hará bajo diferentes supuestos (modelos de inserción) para tratar de identificar los factores implicados en los fenómenos de clusterización.

4. La distribución espacial de las Alus puede venir determinada también por la disponibilidad de las correspondientes dianas de inserción, la inserción preferente y la posible acción diferencial de la selección. Trataremos de discriminar entre estas hipótesis comparando la distribución espacial de las dianas con la distribución de Alus.

5. El estudio riguroso de la dinámica evolutiva de las Alus requiere también una serie de mejoras metodológicas con objeto de estimar con mayor precisión el número de elementos, su edad evolutiva y el contenido local en G+C de la región genómica en que encuentran. En primer lugar, puesto que los TEs muestran un alto grado de fragmentación en el genoma, lo que ha llevado a veces a sobrestimar su número, será necesario implementar un método (desfragmentación) para aglutinar los fragmentos de TEs antes de realizar cualquier cálculo. En segundo lugar, y con objeto de estimar con mayor precisión la edad evolutiva de los elementos, se corregirá la distancia cruda aparente obtenida de los alineamientos mediante la fórmula de Tamura y Nei, que tiene en cuenta los sesgos tanto en las frecuencias de nucleótidos como en la proporción de transiciones y transversiones. En este mismo sentido, se obtendrán estimas independientes de la edad evolutiva según se contemplen o no los sitios ocupados por los dinucleótidos CpG, debido a la alta mutabilidad de estos sitios y a la dinámica diferente que generan. Por último, para estimar el contenido en G+C del entorno Alu se

usarán los segmentos composicionalmente homogéneos obtenidos mediante el algoritmo IsoFinder.

Capítulo 3

Datos y métodos

3.1 Secuencias genómicas y tablas de genes

En este trabajo hemos utilizado la secuencia de referencia (Mayo 2004, versión UCSC hg17) del genoma humano, que se basa en el ensamblaje NCBI 35 generado por el consorcio internacional de secuenciación (IHGSC, 2001). Las secuencias en formato FASTA se han descargado del UCSC Genome Browser (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17).

En distintas partes del trabajo se analizan los TEs en función de su localización respecto a los genes. Desafortunadamente, existen muchas tablas de genes obtenidas en su mayoría con distintos algoritmos de predicción. Una tabla que ha mostrado ser muy fiable es la de RefSeq (Su *et al*., 2004). La información sobre el proceso de obtención y depuración de dicha tabla de genes se puede ver en http://www.ncbi.nih.gov/RefSeq (Pruitt *et al*., 2000; Pruitt y Maglott, 2001; Pruitt *et al*., 2003; Wheeler *et al*., 2005). La tabla de los genes RefSeq se descargó igualmente del UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables).

3.2 Clasificación de las isocoras

Para segmentar el genoma humano en isocoras hemos utilizado el programa IsoFinder (Oliver *et al*., 2004). En la

Figura 3-1 se muestra un archivo de salida de este programa. Nótese que IsoFinder no facilita una clasificación de las isocoras, sino solamente el contenido en G+C de regiones largas y homogéneas (LHGRs, véase Figura 3-1). Es decir, que es el usuario quien debe generar la clasificación a raíz de los contenidos en G+C. Para esto se han utilizado las abundancias de isocoras dadas por Zoubak *et al*. (1996). En concreto, estos autores especifican que el genoma humano está compuesto de la siguiente forma: el 62.9% de isocoras L, el 24.3% de H1, el 7.5% de H2 y el 4.7% de H3[2] (véase también Figura 1-6).

En este trabajo se presentan dos modificaciones de las abundancias dadas por Zoubak y colaboradores. En primer lugar, se divide la isocora L a partes iguales en dos isocoras, L1 y L2. Nótese que de esta forma a L1 y L2 les corresponde la misma abundancia (la mitad de L). En segundo lugar, se fracciona la isocora H3 en dos partes, de manera que la nueva isocora H3 corresponde a dos tercios de la 'antigua' y la nueva isocora H4 está compuesta por el tercio restante. De esta manera, la nueva isocora H4 representa el 1.56% del genoma más rico en G+C (Hackenberg *et al*., 2005).

Con el fin de establecer los rangos de G+C que corresponden a las abundancias dadas por Zoubak y colaboradores, se ha diseñado un algoritmo iterativo. Se empieza con un valor inicial de G+C, que representa la frontera entre las isocoras L1 y L2. Con este valor, se calcula la longitud total de las isocoras que tienen un contenido en G+C menor que el valor de inicio. Con la longitud total se pueden calcular las abundancias. Después, se va aumentando el valor de inicio en pasos de 0.001% de G+C hasta que la abundancia calculada supere la abundancia dada por Zoubak y colaboradores. De esta manera, el valor de G+C fronterizo es excluyente para las isocoras L1 e incluyente para las isocoras L2. De la misma manera se procede para las demás isocoras. Finalmente, mediante los rangos de G+C, se puede asignar a cada LHGR de la tabla de IsoFinder una etiqueta de isocora (L1, L2, H1, H2, H3 o H4). El resultado de este proceso y una estadística básica de las isocoras que hemos usado en este trabajo se pueden ver en 4.1 La nueva isocora H4.

[2] Nótese que el 0.6% restante pertenece a ADN ribosómico. Esta parte se ha distribuido entre las distintas isocoras según su abundancia para obtener el 100%.

Isochore Predictions by IsoFinder

Reference:
J.L. Oliver, P. Carpena, M. Hackenberg and P. Bernaola-Galván
IsoFinder: computational prediction of isochores in genome sequences (NAR 32:W287-W292)

Sequence: out54143/1115113478__Extented_ -- Sig. level = 0.9500 -- Sig. method: Maximum -- Coarse graining: 3000 bp -- SCC: 0.7028E+01

LHGR	From	To	Size	GC%
1	1	284000	284000	43.26
2	284001	705125	421125	36.87
3	705126	765416	60291	41.31
4	765417	1275562	510146	45.95
5	1275563	1330518	54956	39.57
6	1330519	1806902	476384	43.59
7	1806903	2014380	207478	49.20
8	2014381	2142078	127698	46.24
9	2142079	2148190	6112	61.98
10	2148191	2204569	56379	52.47
11	2204570	2210693	6124	45.57
12	2210694	2218401	7708	49.73
13	2218402	2370643	152242	45.87
14	2370644	2442921	72278	52.31
15	2442922	2453122	10201	43.24
16	2453123	2464510	11388	51.70
17	2464511	2713084	248574	43.05
18	2713085	2815535	102451	47.73
19	2815536	3457642	642107	51.87
20	3457643	4028041	570399	40.11
---	---	---	---	---
39	4635600	4647455	11856	52.85

Figura 3-1: Archivo de salida generado por el programa IsoFinder en formato html. El programa facilita las coordenadas, longitudes y contenido en G+C de las regiones genómicas largas y homogéneas o LHGRs (ilustración generada por el servidor web de IsoFinder: http://bioinfo2.ugr.es/IsoF/isofinder.html).

3.3 Preparación de los datos

Para detectar y alinear los elementos transponibles con la correspondiente secuencia consenso de cada familia, hemos empleado el programa RepeatMasker

(Smit y Green, no publicado: http://repeatmasker.genome.washington.edu). El programa se ejecuta con la opción –a para que genere un fichero de salida con los alineamientos. Además, es posible pasar al programa el contenido exacto en G+C de la secuencia genómica para que utilice la matriz de pesos adecuada para el alineamiento. Esta opción es posible utilizarla mediante la aplicación del programa a cada una de las secuencias genómicas de las isocoras. De esta manera, el programa no calcula el contenido en G+C en fragmentos de 50 kb, sino que utiliza el valor de G+C de la isocora. No obstante, los ficheros de salida que genera RepeatMasker no contienen toda la información necesaria, y por tanto hay que procesarlos. A continuación se describen los pasos que hemos dado para ello.

3.3.1 Asignación de genes y TEs a isocoras

La mayoría de los análisis en este trabajo se efectúan en función de la isocora. Por lo tanto, el primer paso consiste en localizar, para cada uno de los genes y TEs, la isocora física en la que residen. Mediante un script en Perl (véase Apéndice A) se asocian a cada gen y TE los valores correspondientes a la isocora: clase de isocora, coordenadas, cromosoma y contenido en G+C. Nótese que a un elemento o gen que solapa con dos isocoras se le asigna la etiqueta "NN", es decir que estos elementos no se utilizan en los análisis en función de la isocora, sino solamente en aquellos análisis que no dependen de la isocora (por ejemplo en un análisis en función del cromosoma).

3.3.2 Análisis de los alineamientos

Muchas características de los elementos transponibles solamente se pueden calcular a partir de los alineamientos entre la secuencia genómica del elemento y su correspondiente secuencia consenso. La Figura 3-2 muestra un alineamiento generado mediante el programa RepeatMasker. A raíz de los alineamientos, se pueden comparar las secuencias genómicas y consenso base a base, con lo que es posible detectar sustituciones, deleciones (guión en la secuencia genómica) e inserciones (guión en la secuencia consenso). Como hemos mencionado en la sección 1.2.2 (Las Alus), las Alus son ricas en dinucleótidos CpG. Una característica destacada de estos sitios es su alta tasa de mutación, que supera en diez veces la tasa normal[3] (Labuda y Striker, 1989; Batzer *et al.*, 1990). Así que a la hora de comparar las tasas de mutación o estimar la edad de los elementos hay que tomar en cuenta estos sitios. La detección fiable de los dinucleótidos CpG en

[3] Las mutaciones en los dinucleótidos CpG se deben sobre todo a la acción de una enzima llamada metilasa, que metila las citosinas, y su rápida mutación posterior hacia TpG.

una secuencia únicamente es posible a partir de los alineamientos, debido a que, ocasionalmente, se puede formar un dinucleótido CpG en la secuencia genómica sin que exista su correspondiente CpG en la secuencia consenso.

```
2408 7.10 0.00 2.63 chr9 50826 51129 (127871) AluY#SINE/Alu 1 296 (15) 1

  chr9_99000296    50826 GGCCGGGTGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA 50875
                         i      i
  AluY#SINE/Alu        1 RGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA 50

  chr9_99000296    50876 GGCGGGCGGATCACGAGGTCAGGAGATCACAACCATCCTGGCTAACATGG 50925
                                                ivi              i
  AluY#SINE/Alu       51 GGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGG 100

  chr9_99000296    50926 TGAAACCCCGTCTCTACTAAAAATACAAAAAAAAAAAAAAATTAGCCAGGT 50975
                                                        --------      i  i
  AluY#SINE/Alu      101 TGAAACCCCGTCTCTACTAAAAATACAAAAAA--------TTAGCCGGGC 142

  chr9_99000296    50976 GTGGTGGCAGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGGAG 51025
                                 i                        v
  AluY#SINE/Alu      143 GTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAG 192

  chr9_99000296    51026 AATGGCATGAACCCAGGAGGTGGAGCTTGCAGTGAGCCAAGACTGTGCCA 51075
                               i       i     i                  i   ii i
  AluY#SINE/Alu      193 AATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCA 242

  chr9_99000296    51076 CTGCACTCCAGCCTGGACAACAGAGCGAGACTCTGTCTCAAAAGAAAGAA 51125
                                     i i              i         i   i
  AluY#SINE/Alu      243 CTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAA 292

  chr9_99000296    51126 AAAA 51129

  AluY#SINE/Alu      293 AAAA 296

Transitions / transversions = 10.00 (20 / 2)
Gap_init rate = 0.00 (1 / 296), avg. gap size = 8.00 (8 / 1)
```

Figura 3-2: Alineamiento entre una AluY y su correspondiente secuencia consenso, generado con el programa RepeatMasker.

A continuación se presentan de forma sinóptica las características de los TEs extraídas a partir de los alineamientos y el fichero "**.out*" generado por RepeatMasker:

- Las coordenadas de la secuencia genómica y del consenso, la longitud del elemento, la divergencia y el número de identificación (extraídos del fichero *"*.out"* de RepeatMasker).

- El contenido en G+C y el número de dinucleótidos CpG tanto en la secuencia genómica como en el consenso.

- Las secuencias genómicas y consenso sin los sitios ocupados por dinucleótidos CpG (necesarias para aplicar la corrección de Tamura y Nei, véase la siguiente sección).
- La matriz de sustitución, con las frecuencias de sustitución entre las 4 bases.
- Las frecuencias de mutación en los dinucleótidos CpG.
- Las frecuencias de las 4 bases en la secuencia consenso (para calcular la probabilidad de mutación).
- El número de deleciones e inserciones en la secuencia genómica para las 4 nucleótidos.

3.3.3 Corrección de Tamura y Nei para sustituciones múltiples

Uno de los objetivos de este trabajo es mejorar algunos de los métodos propuestos en publicaciones de otros autores. El tópico principal de este trabajo es la inesperada diferencia entre las distribuciones de Alus jóvenes y antiguas, y por lo tanto es imprescindible llegar a una estimación fiable de la edad evolutiva[4] de los elementos. En muchas publicaciones anteriores, para estimar la edad de las Alus se utiliza simplemente el número observado de sustituciones por base, es decir el número de nucleótidos diferentes en los alineamientos entre la secuencia genómica y su correspondiente consenso (por ejemplo en Medstrand *et al*., 2002). Este número viene dado por:

$$\hat{p} = \frac{n_d}{n} \qquad 3\text{-}1$$

donde n_d es el número de sustituciones entre la secuencia genómica del elemento y su correspondiente consenso, y n es el número de bases alineadas. Es decir, que estos autores no hacen la corrección para sustituciones múltiples. Además, como hemos señalado en la sección anterior, los dinucleótidos CpG tienen una dinámica evolutiva diferente al resto de los sitios. Esto quiere decir que si se infiere la edad de una Alu sin tomar en cuenta los CpG, se está sobreestimando su edad.

Como es bien conocido, el número observado de sustituciones por base entre dos secuencias proporciona una estimación aceptable de la divergencia evolutiva solo si las dos secuencias son muy similares. Sin embargo, muchos

[4] Tomaremos como "edad evolutiva" o "distancia evolutiva" el número **corregido** de sustituciones por base entre dos secuencias.

elementos transponibles llevan decenas de millones de años en el genoma y por lo tanto muestran una reducida similitud con sus correspondientes consensos. En estos casos, el número de sustituciones por base subestima notablemente la divergencia evolutiva al no tomar en cuenta las mutaciones múltiples. Además, se sabe que las 12 sustituciones posibles entre las bases no tienen todas ellas la misma probabilidad. Hay varios modelos para corregir el número de sustituciones. En este trabajo hemos utilizado la corrección de Tamura-Nei, que toma en cuenta las frecuencias de las bases y las diferencias entre transiciones y transversiones (Tamura y Nei, 1993). La distancia evolutiva (el número corregido de sustituciones por base) se calcula como:

$$d_{TN93} = -2k_1 \ln(1 - \frac{P_1}{k_1} - \frac{Q}{2g_R}) - k_2 \ln(1 - \frac{P_2}{k_2} - \frac{Q}{2g_Y}) - k_3 \ln(1 - \frac{Q}{2g_R g_Y}) \quad (3\text{-}2)$$

donde

$$k_1 = \frac{2g_A g_G}{g_R}$$

$$k_2 = \frac{2g_T g_C}{g_Y}$$

$$k_3 = 2(g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y})$$

donde $g_A, g_C, g_G, g_T, g_R, g_Y$ son las frecuencias de adenina, citosina, guanina, timina, purinas y pirimidinas, P_1 las frecuencias de sustituciones entre A y G, P_2 las frecuencias de sustituciones entre T y C, y Q son las frecuencias de transversiones (sustituciones entre una purina y una pirimidina).

Las frecuencias de las cuatro bases, transversiones y transiciones se calculan a partir de los alineamientos. Nótese que en este modelo no se consideran aparte los dinucleótidos CpG. Sin embargo, en varios análisis es imprescindible hacer esta distinción. Por lo tanto, aplicaremos la corrección tanto a la secuencia entera como a la secuencia sin los sitios ocupados por CpGs.

Como hemos mencionado más arriba, siempre que entre dos secuencias exista una elevada divergencia evolutiva, es importante utilizar la corrección con el fin de no subestimar el número de sustituciones por base. En la Tabla 3-1, se puede ver una comparación entre el número de sustituciones por base y las distancias evolutivas. Los datos en la Tabla 3-1 se han generado mediante las Alus utilizadas en este trabajo. Para varios rangos de sustituciones por base, se ha calculado tanto la media del número de sustituciones como la media de las distancias corregidas. Los datos demuestran claramente que en caso de pocas sustituciones por base, las dos estimaciones dan resultados parecidos. Sin

embargo, según va creciendo la divergencia evolutiva, se ponen de manifiesto diferencias entre las dos estimas. Por ejemplo, en el rango entre 0 y 0.02 sustituciones por base, la diferencia entre los dos modelos es de sólo 0.00089 sustituciones/base. Sin embargo, en el rango entre 0.1 y 0.12, que corresponde a Alus de edad intermedia (AluS), la diferencia es ya de 0.0145.

Tabla 3-1: La tabla muestra las diferencias entre el número de sustituciones por base y las distancias evolutivas (divergencias corregidas mediante el método de Tamura y Nei). Se puede apreciar que según va aumentando el número de sustituciones por base, la corrección se hace más notable. Además se puede observar que las edades evolutivas son notablemente menores si se excluyen los dinucleótidos CpG de la secuencia.

Rango de sustituciones por base	**D**	**$D_{Ta\text{-}Nei}$**	**Diferencia ($D_{Ta\text{-}Nei} - D$)**	**$D_{Ta\text{-}Nei}$ (sin CpG)**
[0,0.02]	0.0114	0.0123	0.0009	0.0100
[0.02,0.04]	0.0312	0.0335	0.0023	0.0187
[0.04,0.06]	0.0516	0.0559	0.0043	0.0267
[0.06,0.08]	0.0706	0.0773	0.0067	0.0353
[0.08,0.1]	0.0906	0.1009	0.0103	0.0477
[0.1,0.12]	0.1095	0.1240	0.0145	0.0633

3.3.4 Desfragmentación de los elementos

Las Alus, al igual que la mayoría de los elementos transponibles, muestran un notable grado de fragmentación en el genoma humano. Hay varios mecanismos que llevan a la fragmentación, como son las inserciones en elementos preexistentes, la recombinación, la inversión 5' y la transducción 3' (en los LINE1; Szak *et al*., 2002). Esto quiere decir que el número de elementos encontrados por el programa RepeatMasker es una sobreestimación del número real debido a la fragmentación. Muchas veces las densidades se calculan en base al número de elementos (véase sección 3.3.5 Cálculo de densidades), y por lo tanto al utilizar los fragmentos se introduce una sobreestimación en el cálculo. Para mejorar el análisis en este aspecto, se ha diseñado un script en Perl (véase Apéndice A) que "aglutina" los fragmentos de la siguiente manera. Primero, RepeatMasker asigna a dos elementos el mismo número de identificación (ID) si la probabilidad de que tengan un origen común es alta. Con este criterio, para todos los TEs se ha determinado si en una ventana de 40 elementos en el flanco 5' existen otros TEs con el mismo número de identificación. Si se detectan

elementos con el mismo ID, el siguiente paso consiste en comprobar la coherencia de las coordenadas de la secuencia consenso. De esta forma, dos elementos se aglutinan si tienen el mismo número de identificación y si la coordenada de la última base en el consenso del elemento 1 y la primera base del consenso del elemento 2 están dentro de un margen de 5 bp (véase Figura 3-3). Los valores básicos (como contenido en G+C o número de mutaciones) del nuevo elemento aglutinado se calculan como la media pesada (por la longitud) de los valores de los fragmentos.

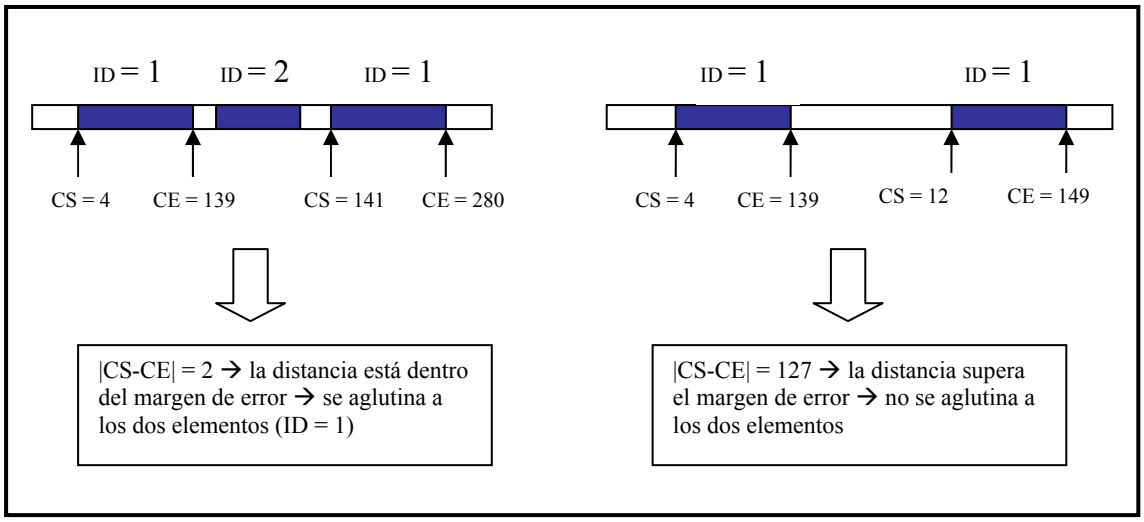

Figura 3-3: Esquema del proceso de "aglutinación" o desfragmentación. Con las flechas se indican las coordenadas de la secuencia consenso (CS y CE son las coordenadas de la primera y última base en el consenso, respectivamente). Se muestran dos casos hipotéticos, en ambos hay dos TEs con el mismo número de identificación, sin embargo solamente el caso de la izquierda está dentro del margen de error (5 bp), y por tanto sería aglutinado.

Tabla 3-2: Grados de fragmentación de las Alus y LINE1 en el genoma humano.

	Alu			**LINE1**		
Isocora	**# elementos**	**# fragmentos**	**Fracción**	**# elementos**	**# fragmentos**	**Fracción**
L1	153570	156503	0.981	195533	297540	0.657
L2	267295	273598	0.977	177037	277937	0.637
H1	335879	347183	0.967	113788	185967	0.612
H2	116275	121411	0.958	25232	42595	0.592
H3	50392	53393	0.944	6980	12017	0.581
H4	10607	11122	0.954	2127	3678	0.578
Total	934018	963210	0.970	520697	819734	0.635

La Tabla 3-2 muestra los grados de fragmentación de las Alus y LINE1 (número de elementos "aglutinados" dividido por el número de fragmentos). Se observa que tanto las Alus como los LINE1 están más fragmentados en las isocoras H que en las L, lo que posiblemente se debe a la correlación positiva que hay entre las tasas de recombinación y el contenido en G+C de la isocora (Fullerton *et al*., 2001). Por otro lado, se observa que el grado de fragmentación difiere profundamente entre los LINE1 y las Alus. El proceso de desfragmentación es más importante en el caso de los LINE1, dado que se reduce el número de elementos aproximadamente un 36%, frente a un 3% en las Alus.

3.3.5 Cálculo de densidades

Existen dos posibilidades para definir y calcular las densidades de los elementos (Alus, LINE1, genes, etc.) dentro de una región dada. La primera medida se basa en el número de unidades, que definimos en este trabajo como:

$$Den_F^R = 10000 \cdot \frac{\sum_{i=1}^{N} d(i)}{Len_R} \tag{3-3}$$

En la ecuación (3-3) se suma sobre todos los elementos (N), siendo *d(i)* la función indicador; ésta vale 1 si el elemento con índice *i* se halla en la región R (Len_R es la longitud de la región R), y 0 en caso contrario. Por lo tanto, Den_F^R representa el número de unidades por 10 kb en la región R. En segundo lugar, también se puede definir la densidad como porcentaje de ocupación, que se calcula como:

$$Den_O^R = 100 \cdot \frac{\sum_{i=1}^{N} Len_i \cdot d(i)}{Len_R} \tag{3-4}$$

En la ecuación (3-4) N es el número de unidades, Len_R la longitud de la región, Len_i es la longitud de la unidad i, y *d(i)* es la función indicador descrita arriba.

3.3.6 Cálculo del exceso de Alus en intrones

Desde hace algún tiempo, se sabe que las Alus muestran densidades más altas en intrones que en los intergenes, es decir que existe un exceso de Alus en intrones (Smit, 1999). Sin embargo, se desconoce si este exceso es una característica general en todo el genoma, o bien se limita a ciertas regiones. Para analizar el exceso en intrones en función de la isocora se utilizan todas las isocoras que cuentan con al menos una región intergénica, es decir dos genes. Además, definimos el entorno génico como la región codificadora más 2 kb en cada flanco. En la tabla de genes RefSeq se anotan también los distintos transcritos

pertenecientes al mismo gen. En estos casos, elegimos siempre el transcrito más largo. Para poder comparar fácilmente las densidades de Alus en intergenes y en intrones, definimos un coeficiente de "exceso en intrones" que viene dado por:

$$R_{Ex} = \frac{\rho_{IV}}{(\rho_{IV} + \rho_{ig})} - 0.5 \qquad \textbf{(3-5)}$$

donde ρ_{IV} y ρ_{ig} son las densidades de Alus en intrones y en intergenes, respectivamente (Hackenberg *et al*., 2005). El coeficiente R_{Ex} nos permite comparar fácilmente las densidades, ya que toma valores negativos si la densidad en intergenes supera a la densidad en intrones, y valores positivos en caso contrario.

3.3.7 La definición *in silico* de los productos de recombinación

Para poder contabilizar la acción de la recombinación homóloga desigual en el genoma humano definimos un trímero (duplicación de una Alu) de la siguiente manera:

- Tres Alus seguidas en la misma hebra y a una distancia menor de 80 bp entre sí.
- No debe haber una cuarta Alu a menos de 200 bp del trímero.
- En el trímero, dos Alus seguidas (primera y segunda o segunda y tercera) deben pertenecer a la misma subfamilia (como consecuencia de la duplicación).
- Las inserciones de Alus dentro de otras Alus (IPEs) ocurren con frecuencia en el genoma humano dando lugar a un trímero de origen no-recombinacional. Para filtrar estos trímeros "falsos", hemos utilizado los números de identificación asignados por RepeatMasker a cada elemento. Si un elemento se divide en dos fragmentos a consecuencia de una inserción, Repeatmasker asigna a ambos fragmentos el mismo número de identificación para subrayar su posible origen común. Pues bien, hemos podido descartar los trímeros no-recombinacionales mediante estos números de identificación.

3.3.8 Detección de IPEs en el genoma

En el genoma humano a menudo se observa que los TEs se insertan dentro de otros elementos transponibles. A una inserción de este tipo la llamaremos IPE (**I**nsertion into a **P**re-existing **E**lement). Para definir una IPE *in silico* utilizamos

los números de identificación que asigna el programa RepeatMasker a cada uno de los elementos que identifica. Así pues, consideramos como IPE una secuencia genómica que cumple los siguientes requisitos:

- Un elemento transponible tiene que estar flanqueado en ambos extremos por el mismo elemento. Consideramos a los elementos flanqueadores como idénticos si pertenecen a la misma subfamilia y si tienen el mismo número de identificación.

- La coordenada de la secuencia consenso del ultimo nucleótido del elemento 5' y la coordenada del primer nucleótido del elemento 3' tienen que hallarse dentro de un margen de 10 bp.

- Los dos elementos de los flancos tienen que hallarse en la misma hebra.

3.3.9 Detección de dianas de inserción

Para detectar las dianas de inserción (5'-TTAAAA-3'), hemos empleado el programa *fuzznuc* del paquete EMBOSS para búsqueda de patrones (Rice *et al*., 2000; véase Apéndice B). Las opciones se eligen de manera que el programa solamente detecte dianas exactas y en ambas hebras. Al igual que hacemos con TEs y genes, el primer paso consiste en localizar las dianas dentro de las isocoras, es decir asignarles una etiqueta. Las dianas las utilizamos para comparar sus frecuencias en ciertas regiones con las frecuencias de Alus. En concreto, las regiones que se analizan en este trabajo son los propios elementos transponibles y sus flancos. El punto clave es la comparación con las frecuencias de Alus, y por tanto hay que tomar en cuenta la resolución del programa RepeatMasker, que es de 10 bp. Este límite tiene como consecuencia que la inserción de una Alu en el extremo de otra Alu preexistente no se detecta como IPE sino como inserción en el flanco (véase Figura 3-4). En consecuencia, consideramos que una diana está dentro de un elemento si se halla dentro de los límites del elemento menos 10 bp en ambos extremos. Y por consiguiente, el flanco abarca los últimos 10 bp del elemento más los primeros 20 bp que siguen (en ambos extremos).

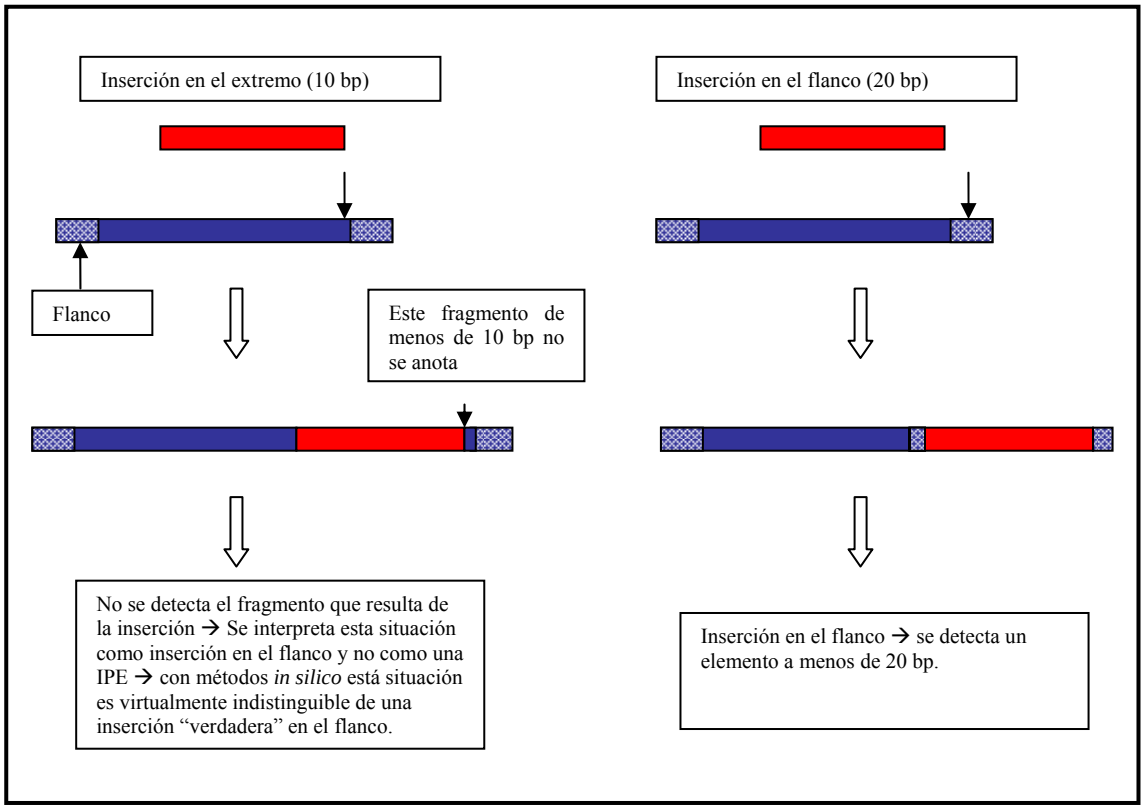

Figura 3-4: Esquema que ilustra el hecho de que es imposible distinguir con métodos *in silico* entre una inserción en el extremo de un elemento preexistente y una inserción en el flanco.

Capítulo 4

Resultados y Discusión

4.1 La nueva isocora H4

En este trabajo se introduce una nueva clasificación de las isocoras. La razón por la que se ha definido una nueva isocora H4 tiene su origen en un hecho conocido: las densidades de las Alus muestran una disminución marcada en las regiones más ricas en G+C (Gu *et al*., 2000). Sin embargo, se sabe que los genes se acumulan en las regiones más ricas en G+C (Mouchiroud *et al*., 1991; Bernardi, 2000). En consecuencia, para dar una representación sólida a estas regiones definimos la nueva isocora H4 como el 1.57% del genoma más rico en G+C (véase el apartado 3.2 Clasificación de las isocoras). La Tabla 4-1 muestra una estadística básica de la clasificación de isocoras utilizada en este trabajo.

En la Tabla 4-2 se muestran algunas características básicas de las Alus y los LINE1 en función de la isocora. En primer lugar, se confirma la distribución opuesta de las densidades. Mientras que los LINE1 se acumulan fuertemente en las isocoras L (máximo de densidad en L1), las Alus muestran el máximo de densidad en las isocoras H3 (Korenberg y Rykowski, 1988; Chen *et al*., 2002). Nótese que Bernardi y colaboradores (Zerial *et al*., 1986; Jabbardi y Bernardi, 1998) encuentran el máximo de densidad de las Alus en las isocoras H2. La diferencia con el resultado presentado aquí se debe a que la isocora H3 de

Bernardi comprende nuestra isocora H4. Esto quiere decir, por tanto, que la isocora H3 de Bernardi abarca regiones que muestran distintos niveles de acumulación de Alus. En concreto, la densidad de Alus en las isocoras H3 de Bernardi corresponde aproximadamente a la suma pesada entre las isocoras H3 y H4 (tal como se definen en esta memoria). Así pues, la aportación de las densidades bajas en las regiones más ricas en G+C hace que Bernardi observe el máximo en H2 y no en H3. Otra propiedad digna de resaltar es la distribución de las longitudes de los LINE1 en función de la isocora. Se observa que la longitud media en las isocoras L dobla aproximadamente la longitud media en las isocoras H3/H4. Por el contrario, las Alus no muestran esta dependencia.

Tabla 4-1: Estadística básica de las isocoras: etiqueta, número, longitud media, longitud total, media del contenido en G+C, porcentaje y rangos de G+C de las 6 isocoras. Los corchetes cerrados en los rangos de G+C indican que el límite está incluido mientras que los corchetes abiertos excluyen el límite.

Isocora	#	Longitud media (bp)	Longitud total (bp)	Media de G+C	% del genoma	Rango de G+C
L1	1353	621586	8.4E+08	36.20	31.65	[0 -37.70[
L2	2249	374712	8.4E+08	39.48	31.65	[37.70 – 41.38[
H1	2371	262255	6.2E+08	43.92	24.45	[41.38 – 47.15[
H2	875	205292	1.8E+08	49.21	7.55	[47.15 – 51.70[
H3	384	186463	7.2E+07	53.39	3.14	[51.70 – 55.65[
H4	167	158348	2.6E+07	58.29	1.56	[55.65 – 100]

Tabla 4-2: Número, densidad (número de unidades por 10 kb) y longitud de Alus y LINE1 en función de la isocora. Se confirma la distribución opuesta entre Alus (máximo de densidad en H3) y LINE1 (máximo en L1). Además, las longitudes de los LINE1 muestran un sesgo marcado en función de la isocora, variando entre aprox. 950 bp en las isocoras L y 460 bp en las isocoras H3/H4.

	Alu			LINE1		
Isocora	#	#/10 kb	Longitud (bp)	#	#/10 kb	Longitud (bp)
L1	153570	1.83	281.9	195533	2.32	978.7
L2	267295	3.17	282.6	177037	2.10	908.7
H1	335879	5.40	281.4	113788	1.83	654.8
H2	116275	6.47	281.8	25232	1.40	514.3
H3	50392	7.04	281.3	6980	0.97	458.8
H4	10607	4.01	278.0	2127	0.80	460.6

En distintos puntos de este trabajo se hace referencia a la distribución de los genes en función de la isocora y por tanto es conveniente tener en cuenta algunas de sus características. La Tabla 4-3 muestra una estadística de los genes RefSeq en función de la isocora. Los datos confirman la correlación positiva entre la densidad y el contenido en G+C de la isocora, y la correlación negativa entre la longitud de los intrones y el G+C (Bernardi, 2001).

Tabla 4-3: Estadística básica de los genes RefSeq (hg17) en función de la isocora. Se puede observar que la longitud de los intrones y la densidad varían drásticamente en función de la isocora, mientras que la longitud de los exones y la media del número de exones por gen no muestran una correlación clara con la isocora.

Isocora	**#**	**#/10 kb**	**% de ocupación**	**Longitud de intrones (bp)**	**Longitud de exones (bp)**	**# exones por gen**
L1	552	0.0244	12.37	13990	426	11.4
L2	1119	0.0383	13.75	9148	420	10.1
H1	1327	0.0611	15.02	6410	446	8.9
H2	612	0.1258	21.28	4733	414	8.9
H3	320	0.2262	32.15	3503	339	10.5
H4	144	0.2761	35.38	2801	330	11.6
Total	4074	0.0685	16.24	7533	419	9.8

Finalmente, la variación de las frecuencias de mutación en función de la isocora se puede estudiar mediante las mutaciones encontradas en las Alus. En la Figura 4-1 se muestran por separado las frecuencias de mutaciones A/T→GC y G/C→A/T. Dos hechos se pueden distinguir claramente. Primero, las mutaciones G/C→A/T ocurren siempre con mayor frecuencia independientemente de la isocora (Eyre-Walker, 1999; Smith y Walker, 2001; Alvarez-Valin *et al*., 2003). Segundo, se puede apreciar un sesgo de las frecuencias de mutación en función de la isocora. Las frecuencias de mutación G/C→A/T son más altas en isocoras L debido a la presión mutacional local. Las mutaciones A/T→G/C muestran probabilidades invertidas, es decir que ocurren con mayor frecuencia en las isocoras H que en L. Esto es justamente lo que se esperaría según el modelo de la presión mutacional local (Wolfe *et al*., 1989). Al final del ciclo celular ya no quedan tantos dCTP y dGTP y por lo tanto la probabilidad de que se produzca una mutación hacia G o C es menor en las isocoras L que en H.

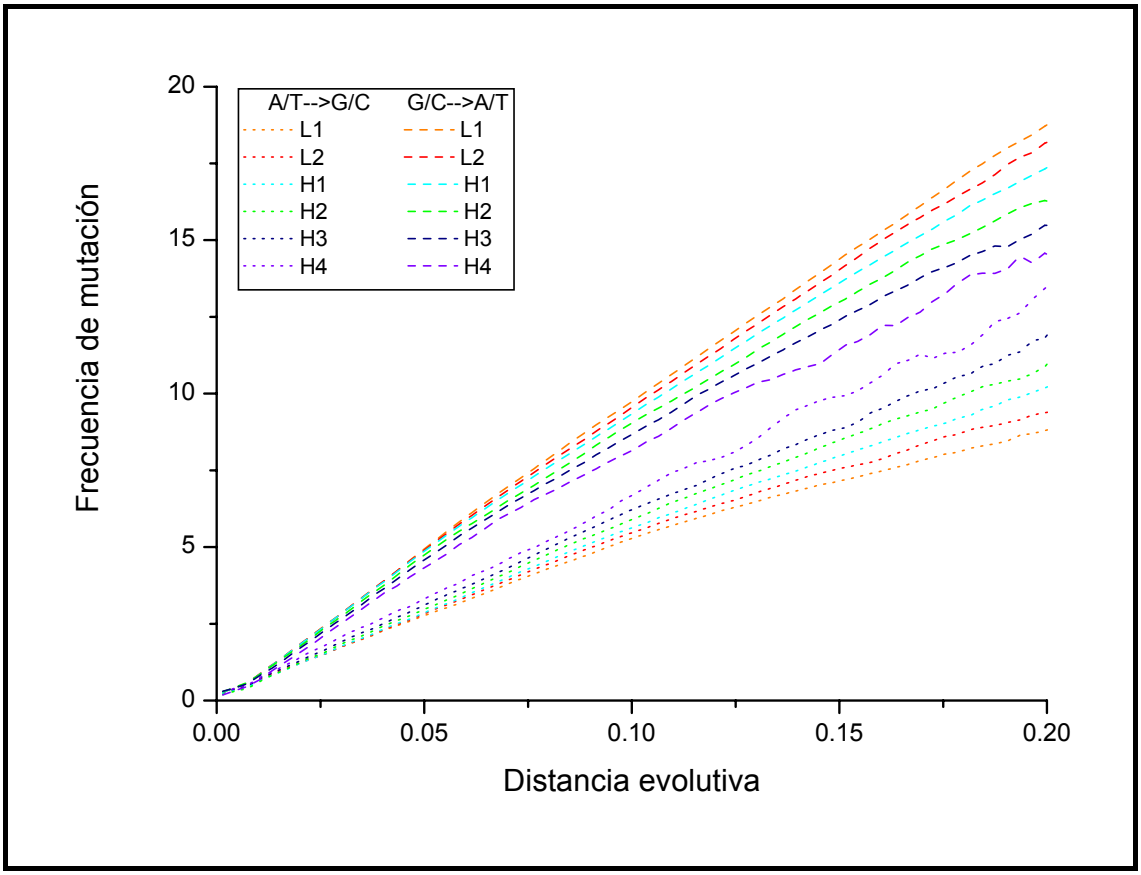

Figura 4-1: Dinámica evolutiva de la presión mutacional G/C ←→ A/T según los datos derivados de los alineamientos entre la secuencia genómica de los elementos y la correspondiente secuencia consenso. Las mutaciones G/C→A/T ocurren con mayor frecuencia en isocoras L que en H, debido a la escasez relativa de G y C al final del ciclo celular. Las mutaciones A/T→G/C se comportan complementariamente, es decir, que son más frecuentes en H que en L.

4.2 El cambio de densidad de las Alus

Existen varias propuestas para explicar el cambio de densidad que muestran las Alus en función de su edad (véase sección 4.1 La nueva isocora H4). En esta parte del trabajo, se analizan varias proposiciones, empleando métodos nuevos con el fin de evaluar el impacto de los distintos mecanismos evolutivos.

4.2.1 Densidades relativas y distancia evolutiva

Con el término "cambio de densidad" nos referimos a que las Alus jóvenes se acumulan en regiones ricas en A+T, mientras que las Alus más antiguas muestran su máxima densidad en regiones ricas en G+C. En publicaciones de otros

autores, el cambio de densidad se ha puesto de manifiesto mediante la comparación entre subfamilias jóvenes del grupo AluY y grupos más antiguos de Alus, como AluS y AluJ (Pavlíček *et al*., 2001; IHGSC, 2001). Aunque el cambio de densidad queda de manifiesto, estos análisis no permiten llegar a conclusiones precisas sobre la dinámica del proceso (velocidad y momento en que se produjo el cambio de densidad). No obstante, estas cuestiones son importantes a la hora de evaluar los mecanismos implicados. Por ello, en este trabajo se analizan las densidades de las Alus en función de su edad evolutiva, con el fin de resolver el cambio de densidad más nítidamente. La edad o distancia evolutiva se calcula a partir de los alineamientos entre los elementos y sus correspondientes secuencias consenso (véase 3.3.3 Corrección de Tamura y Nei), aplicando la corrección para sustituciones múltiples. En consecuencia, a cada Alu se le asigna su edad evolutiva y una etiqueta que indica la clase de isocora en la que reside. A continuación, se agrupan las edades evolutivas en intervalos de 0.005 sustituciones por base y se determina el número de elementos en cada intervalo. Las densidades en función de la isocora y del intervalo vienen dadas por:

$$Den_{Iso}(D_i) = 10000 \cdot \frac{\#_i}{Len_{Iso}} \tag{4-1}$$

donde i es el índice del intervalo y $\#_i$ el número de elementos en ese intervalo.

La mejor forma de poner de manifiesto el cambio de densidad consiste en calcular las densidades relativas en vez de las absolutas. La densidad relativa es la proporción entre las densidades absolutas en las isocoras H y L (H*/L, siendo L la suma pesada de isocoras L1 + L2). De esta manera, las densidades relativas son menores que 1 si el máximo de densidad absoluta se ubica en las isocoras L. La Figura 4-2 (arriba) muestra las cuatro densidades relativas en función de la edad evolutiva.

Se observa que el máximo de densidad se localiza en las isocoras L hasta distancias de aproximadamente 0.025, puesto que todas las proporciones son menores que 1. Sin embargo, el cambio de densidad se efectúa rápidamente, y a partir de esta distancia el máximo ya se encuentra en las isocoras H.

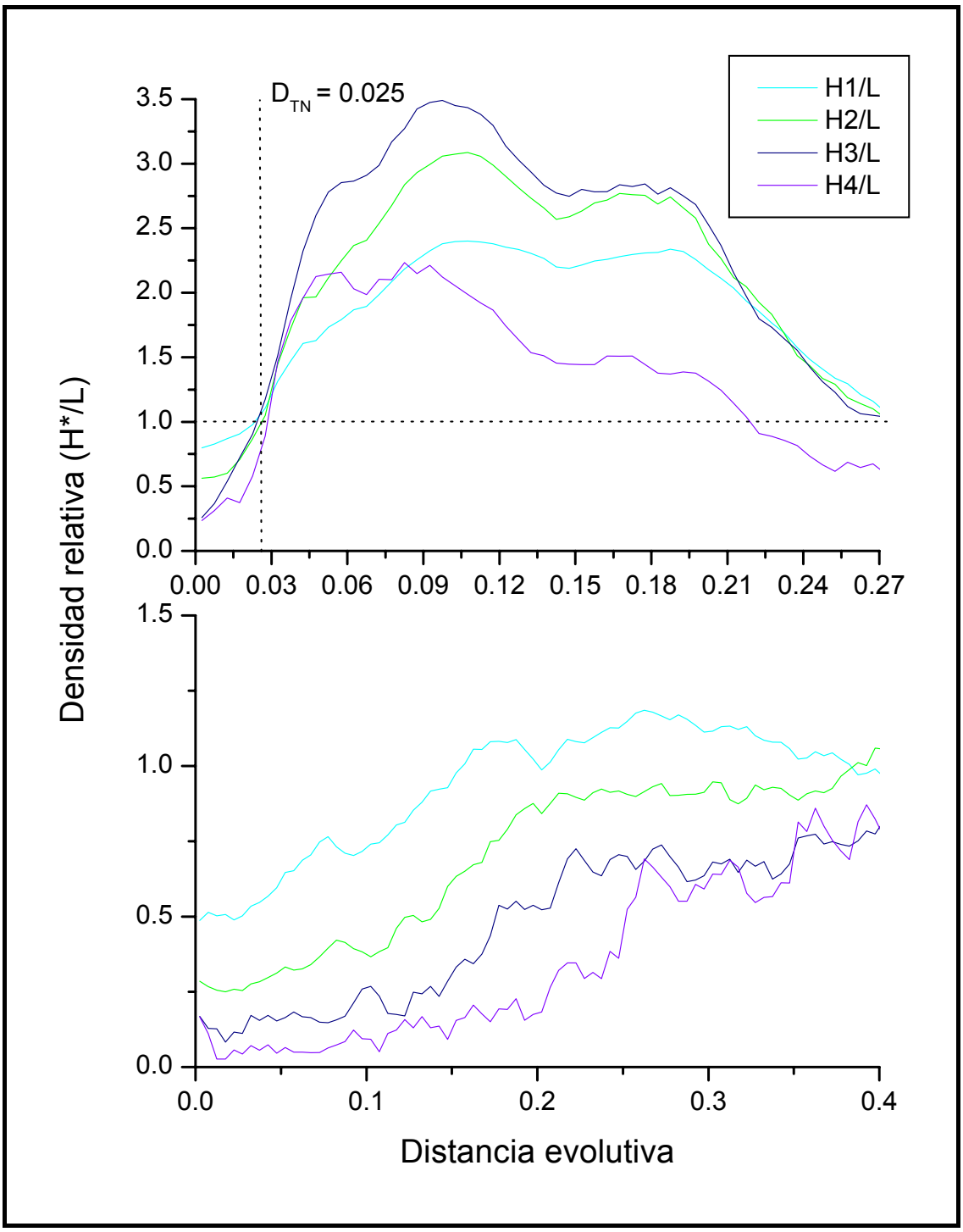

Figura 4-2: Densidades relativas de Alus (arriba) y LINE1 (abajo). El máximo de densidad de las Alus se halla en las isocoras L hasta distancias evolutivas de aproximadamente 0.025, puesto que todas las proporciones son menores que 1. A partir de ahí, se traslada rápidamente hacia las isocoras H3. Los LINE1 (abajo) no muestran ningún cambio de densidad, se insertan preferentemente en las isocoras L donde también se ubica el máximo total.

Como hemos mencionado más arriba, las Alus son ricas en dinucleótidos CpG. Estos sitios evolucionan aproximadamente 10 veces más rápido que los demás (Belle *et al*., 2003; Xing *et al*., 2004; CpG → TpG). En la Figura 4-2, no se han excluido estos dinucleótidos y por tanto la escala de distancias evolutivas no es lineal. En consecuencia, las distancias entre 0.02 y 0.03 (por ejemplo) no representan la misma cantidad de años que las distancias entre 0.08 y 0.09. Sin embargo, el cambio de densidad solo se puede observar de esta forma. A partir de ahora, nos referiremos a este cambio de densidad como el "último", puesto que es de suponer que el cambio de densidad de las Alus ha ido ocurriendo de una manera continua a lo largo de su evolución.

En la Figura 4-2 (abajo) se muestra el mismo análisis para los LINE1. La preferencia de inserción es cualitativamente muy parecida a la de las Alus. Sin embargo, la tendencia hacia regiones ricas en A+T es aún más marcada, puesto que las densidades relativas de los LINE1 jóvenes son menores que las de las Alus jóvenes. Una observación similar la hicieron previamente otros autores (Pavlíček *et al*. 2001). Caben dos posibles explicaciones: o los LINE1 se excluyen de las regiones ricas en G+C, o las Alus se eliminan de las isocoras ricas en A+T.

En resumen, el análisis de la densidad relativa de las Alus en función de la distancia evolutiva pone de manifiesto que debe existir algún mecanismo evolutivo que, actuando tras la inserción, provoque un cambio rápido de su máximo de densidad hacia las isocoras H.

4.2.2 Interacciones Alu/LINE1

Un mecanismo que probablemente influye en la distribución espacial de las Alus es la competencia entre Alus y LINE1 por la retrotransposasa (proceso de reinserción). Gu *et al*. (2000) proponen que las Alu podrían evitar la competencia si se insertasen en regiones ricas en G+C, lo que podría explicar el cambio de densidad.

Para poner a prueba esta hipótesis, hemos comparado las densidades absolutas de las Alus y los LINE1 en función de la isocora y la distancia evolutiva. Sin embargo, hay que tener cuidado a la hora de calcular las densidades de los LINE1. En general, hay dos maneras de calcular las densidades, bien como el porcentaje de ocupación, bien como el número de elementos por región (véase 3.3.5 Cálculo de densidades). Las longitudes de los LINE1 (véase Tabla 4-2) muestran un sesgo muy marcado en función de la isocora. Es decir que en isocoras muy ricas en G+C (H3 y H4) los LINE1 tienen solamente la mitad de longitud que en isocoras ricas en A+T (L1 y L2). Debido a

este sesgo, el porcentaje de ocupación parece ser una medida poco adecuada, ya que llevaría a una subestimación de la densidad de los LINE1 en regiones ricas en G+C. Por otro lado, los LINE1 muestran un alto nivel de fragmentación en el genoma (véase 3.3.4 Desfragmentación de los elementos). Sin embargo, programas como RepeatMasker solamente detectan y anotan los fragmentos. En vista de que el número de fragmentos siempre es más alto que el número real de elementos, la utilización del número de fragmentos provoca una sobreestimación de la densidad.

Para obtener una estima más fiable del número verdadero de elementos, hemos diseñado un script en Perl, que "aglutina" los fragmentos procedentes del mismo elemento (véase 3.3.4 Desfragmentación de los elementos). Además, para poder comparar fiablemente las Alus y los LINE1 es imprescindible utilizar las edades evolutivas calculadas sin los sitios ocupados por los dinucleótidos CpG (debido al gran número de estos sitios en las Alus y su alta mutabilidad).

La Figura 4-3 muestra las densidades calculadas a partir de los elementos aglutinados en función de la distancia evolutiva y para cada isocora. Los LINE1 muestran un mínimo muy marcado en las isocoras L1 y L2, aproximadamente entre distancias evolutivas de 0.03 y 0.15 (Figura 4-3 abajo). Este mínimo coincide en el tiempo con el máximo auge de las Alus (Figura 4-3 arriba), y parece por tanto que hay una especie de interacción entre las Alus y los LINE1. Sin embargo, la inserción de las Alus en las isocoras L no parece haber sido impedida por los LINE1. Más bien al contrario, son las densidades de los LINE1 las que sufren una disminución muy marcada durante el auge de las Alus. Esto quiere decir que son las Alus las que han frenado la expansión de los LINE1 y no al contrario. Podemos concluir, por tanto, que la competencia por la retrotransposasa no es probablemente el mecanismo capaz de explicar el cambio de densidad de las Alus.

Este resultado es coherente con trabajos anteriores sobre el proceso de retrotransposición. Según los modelos propuestos, la competencia por la retrotransposasa tiene lugar cerca de los ribosomas y no dentro del núcleo, lo que descarta igualmente que dicha competencia pueda tener alguna influencia sobre la distribución de las Alus (Boeke, 1997).

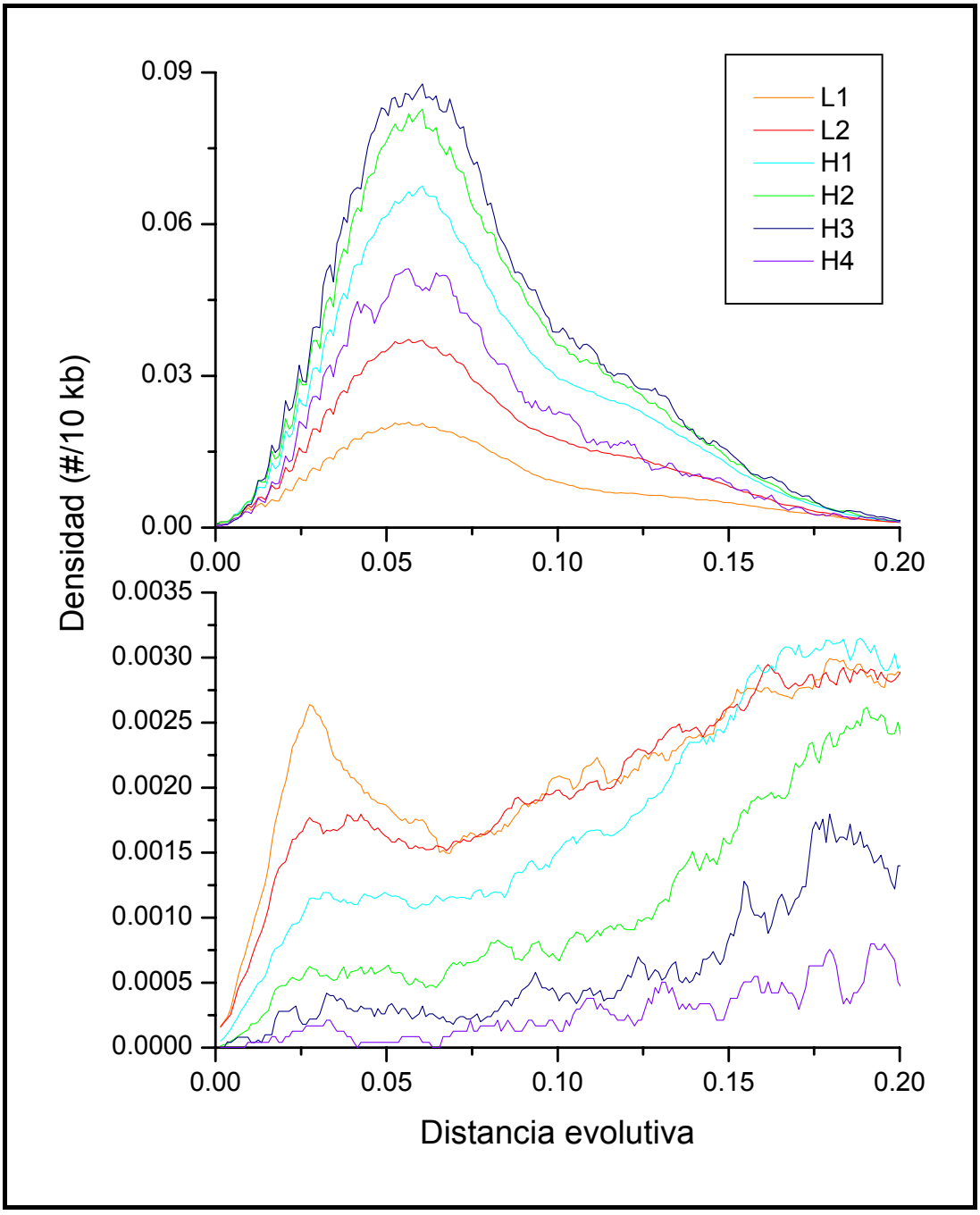

Figura 4-3**:** Las densidades absolutas de las Alus (arriba) y LINE1 (abajo) para todas las isocoras y en función de la distancia evolutiva. Se puede apreciar un "bache" en las densidades de los LINE1, sobre todo en las isocoras L, que coincide en el tiempo con la mayor actividad de transposición de las Alus hace unos 35-40 millones de años.

4.2.3 El ajuste composicional

Las Alus están mejor ajustadas composicionalmente en las isocoras H, debido a su riqueza en G+C. Y lo mismo se puede decir para los LINE1 en las isocoras L. En general, se observa que los elementos transponibles tienden a establecer su máximo de densidad en regiones donde están mejor ajustados composicionalmente (Filipski *et al*., 1989). Por lo tanto, el ajuste composicional podría desempeñar también un papel en el cambio de densidad.

Para probar esta hipótesis, hemos calculado las diferencias en G+C entre las Alus y sus correspondientes secuencias consenso, en función de la distancia evolutiva y para todas las isocoras (véase Figura 4-4). En primer lugar, se aprecia claramente el declive del contenido en G+C de las Alus. Hasta distancias evolutivas entre aproximadamente 0.08 y 1, la reducción del contenido en G+C es más rápida, manifestándose en gradientes mayores, en comparación con la dinámica a partir de estas distancias. La disminución inicialmente más pronunciada del contenido en G+C se puede explicar fácilmente por las tasas más altas de mutación de los dinucleótidos CpG. A este rango de distancias evolutivas se le llama "dominio CpG" (Hackenberg *et al*., 2005). Durante el dominio CpG, no se observa ninguna influencia de la isocora sobre la composición de los elementos. Es decir, el contenido en G+C disminuye en todas las isocoras aproximadamente con la misma intensidad, reflejando así que las mutaciones en los dinucleótidos CpG se deben a la acción de la metilasa, que actúa independientemente de la isocora. Después del dominio CpG, se observa un declive más marcado en isocoras L, poniendo de manifiesto que las Alus están sometidas a mayor presión mutacional en isocoras L que en H. Es decir, en las isocoras L la presión de ajuste composicional al entorno genómico es mayor que en las isocoras H.

La evolución del ajuste composicional se puede observar también mediante los coeficientes de correlación entre el contenido en G+C de la isocora y el de los elementos. La Figura 4-5 muestra los coeficientes de correlación calculados en cada intervalo y en función de la distancia evolutiva. Se observa el mismo comportamiento descrito arriba. Durante el dominio CpG, no hay ningún ajuste composicional y por tanto la correlación no aumenta. Sin embargo, después del dominio CpG, los coeficientes de correlación van creciendo, como consecuencia del ajuste composicional. Nótese que estos resultados son coherentes con la teoría sobre la presión mutacional local (Wolfe *et al*., 1989), ya que esta predice que las tasas de mutación son dependientes del G+C.

Nótese, sin embargo, que el ajuste composicional no se hace notar hasta distancias evolutivas entre 0.08 y 0.1, lo que es claramente posterior al momento

en que se produce el cambio de densidad de las Alus (alrededor de 0.025, véase apartado 4.2.1 Densidades relativas y distancia evolutiva). Por lo tanto, se puede descartar la implicación de este mecanismo en el cambio de la densidad.

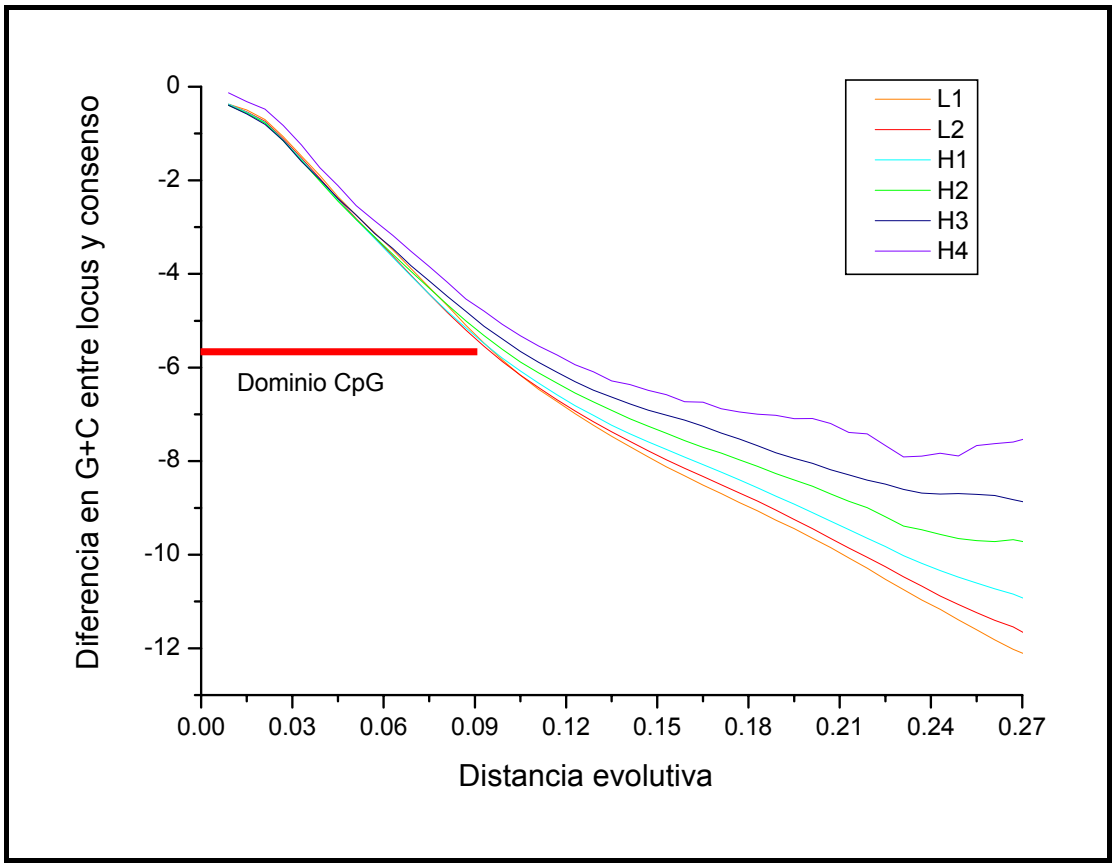

Figura 4-4: Diferencias en G+C entre las secuencias Alu en el genoma y sus correspondientes secuencias consenso. Se pueden distinguir dos fases en la evolución composicional de las Alus. Durante el dominio CpG, el contenido en G+C de las Alus baja rápidamente debido a las mutaciones en los dinucleótidos CpG. Durante esta fase, no se observa ningún ajuste composicional a la isocora. Sin embargo, este es claramente discernible a partir de distancias de aproximadamente 0.1. Se aprecia claramente que las Alus están sometidas a mayor presión composicional en isocoras L, dada la mayor reducción del contenido en G+C en estas isocoras en comparación con las isocoras H.

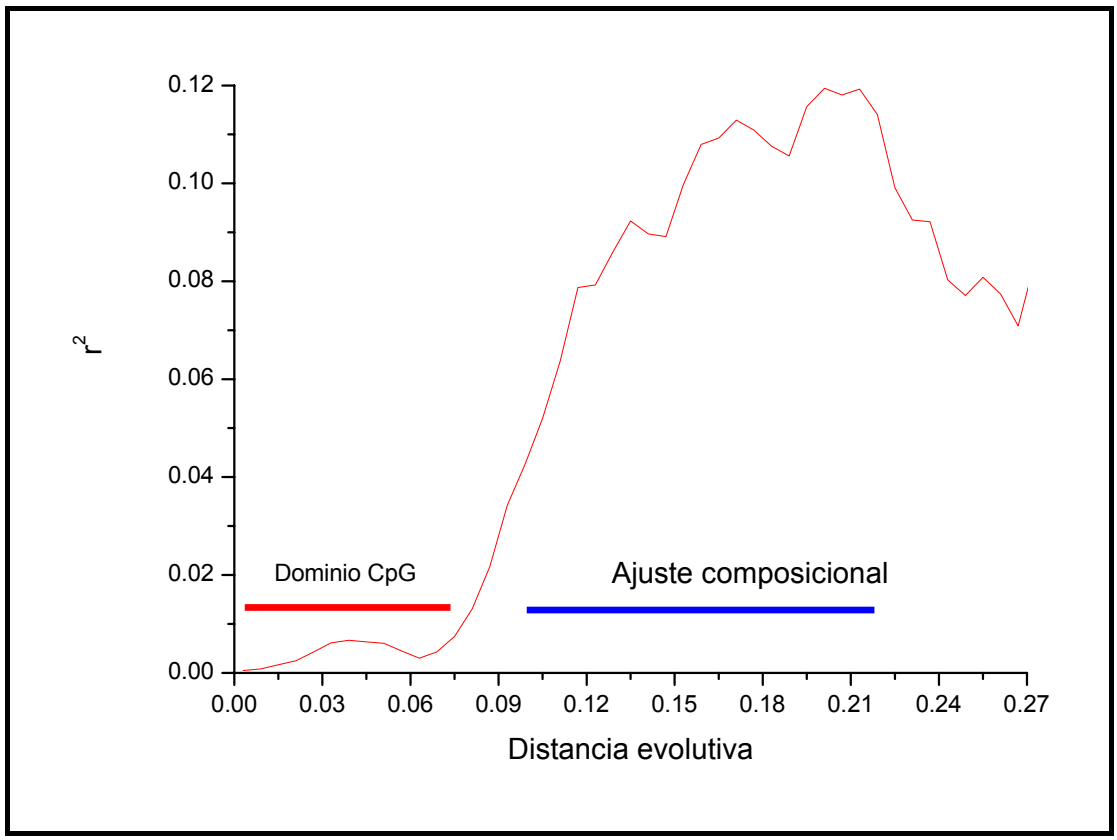

Figura 4-5**:** Coeficientes de correlación entre el contenido en G+C de la isocora y el de los elementos. Se observa que durante el dominio CpG los elementos no se ajustan composicionalmente al G+C de la isocora. Sin embargo, a partir de distancias evolutivas de aproximadamente 0.1, los elementos se van ajustando composicionalmente, ya que los coeficientes de correlación son crecientes.

4.2.4 Alus en el entorno génico

Tanto los genes como las Alus muestran densidades más altas en regiones ricas en G+C (Zoubak *et al*., 1996; Bernardi 2000). En muchas publicaciones se argumenta que este hecho no es casual, sino que las Alus de alguna forma siguen la estela de los genes, acumulándose igualmente en las isocoras H (IHGSC, 2001). Existen varios indicios a favor de esta hipótesis. Por un lado, se sospecha que las Alus podrían estar funcionalmente relacionadas con los genes, por lo que la selección positiva podría fijar las Alus preferentemente en regiones ricas en G+C. Por otro lado, Smit (1999) ha puesto de manifiesto que existe un exceso de Alus en intrones, lo que podría significar que las Alus son mejor aceptadas en el entorno génico que en otras regiones, probablemente debido a su potencial de exonización (Sorek *et al*., 2002; 2004; Lev-Maor *et al*., 2003). Ante este

escenario, la afinidad de las Alus hacia el entorno de los genes podría contribuir también al cambio de densidad de las Alus.

Tabla 4-4: Análisis de las densidades de Alus en función de su localización genómica, tanto para el genoma entero como en función de la isocora. En primer lugar, tomando en cuenta el genoma entero, se confirma el exceso de Alus en intrones. Sin embargo, el análisis en función de la isocora revela que tal exceso se limita a las isocoras L (los valores positivos del coeficiente de exceso en intrones solo se observan en las isocoras L – en color amarillo).

	Intrones		**Intergenes**		
	N	**Densidad**	**N**	**Densidad**	**Exceso intrónico**
Genoma entero	163874	4.640	807434	3.558	0.0660
Isocora					
L1	11297	2.277	18054	2.134	0.0162
L2	33802	4.126	49187	3.964	0.0100
H1	47287	7.022	67336	7.104	-0.0029
H2	19806	7.109	31216	8.011	-0.0298
H3	13063	7.206	18120	8.410	-0.0385
H4	2778	3.996	4300	5.697	-0.0878

Para poner a prueba esta hipótesis, hemos calculado las densidades en intrones y en regiones intergénicas en función de la isocora (para una definición precisa de estas regiones véase 3.3.6 Cálculo del exceso de Alus en intrones). Si la hipótesis fuera correcta, se esperaría un exceso de Alus en intrones en todas las isocoras. Además, el exceso debería ser más pronunciado en isocoras H que en L. Para poder comparar fácilmente las densidades de Alus en intergenes y en intrones, hemos definido un coeficiente de "exceso en intrones" (véase 3.3.6 Cálculo del exceso de Alus en intrones). Recuérdese que el coeficiente R_{Ex} toma valores negativos si la densidad en intergenes supera a la densidad en intrones, y valores positivos en caso contrario. En la Tabla 4-4 se puede ver un resumen del análisis llevado a cabo tanto para el genoma entero como por isocoras. Cuando se analiza el genoma entero, se puede comprobar el exceso de Alus en intrones observado por Smit (1999). Sin embargo, cuando se realiza el análisis en función de la isocora, se pone de manifiesto que tal exceso se limita a las regiones ricas en A+T (isocoras L). Esto descarta la propuesta del IHGSC (2001) de que la acumulación de Alus en regiones ricas en G+C se debe a los efectos positivos

sobre la expresión de los genes. Por el contrario, nuestros resultados parecen confirmar la argumentación de Brookfield (2001), quien propone que los niveles observados de polimorfismo de Alus son inconsistentes con la acción de la selección positiva, puesto que ésta daría lugar a una fijación más rápida de las Alus de la que se observa. Por lo tanto, nuestros resultados son consistentes con las conclusiones derivadas de los estudios de genética de poblaciones de este autor.

4.2.5 Búsqueda de evidencias de recombinación

Las secuencias Alu participan frecuentemente en procesos de recombinación, tanto homóloga como no-homologa (Babcock *et al*., 2003; Deininger *et al*., 2003). La distancia media entre dos Alus en el genoma humano es de unas 3 kb; un emparejamiento incorrecto entre estos elementos, y el subsiguiente entrecruzamiento desigual, es la mayor fuente de deleciones y duplicaciones. La Figura 4-6 muestra un modelo simple de recombinación homóloga desigual (RHD) entre dos Alus. Como se puede ver, este proceso produce la deleción de una Alu en uno de los cromosomas, y su duplicación en el otro (si el quiasma se forma entre las Alus). Por lo tanto, la RHD es capaz de alterar el número absoluto de elementos en el cromosoma; en consecuencia, en varias publicaciones se propone a la RHD como el principal mecanismo evolutivo responsable del cambio de densidad de las Alus (Lobachev *et al*., 2000; Brookfield, 2001; Stenger *et al*., 2001; Batzer y Deininger, 2002; Medstrand *et al*., 2002; Deininger *et al*., 2003; Belle *et al*., 2005).

Si bien es cierto que la RHD altera el número absoluto de elementos en un cromosoma (gameto), su número absoluto en el conjunto de la población se mantiene constante (un cromosoma lleva una deleción y el otro una duplicación). Así que para que haya un efecto neto en términos evolutivos son necesarios más requisitos. Primero, uno de los dos productos de recombinación, deleción o duplicación, debe tener una probabilidad más alta de supervivencia. Por ejemplo, un análisis de reordenaciones cromosómicas en humanos ha revelado que las deleciones ocurren con mayor frecuencia que las duplicaciones (Deininger y Batzer 1999; Kolomietz *et al*., 2002). En la misma línea, existen también datos experimentales sobre la recombinación en células de mamíferos que parecen avalar igualmente frecuencias más altas de deleciones (Lambert *et al*., 1999).

El segundo requisito que debe cumplirse es una probabilidad diferencial de supervivencia o formación de los productos de recombinación en las distintas regiones del genoma. En concreto, los productos de recombinación deben mantenerse o producirse más frecuentemente en isocoras L que en H, lo que

aumentaría la densidad relativa en las isocoras H, provocando así (o participando al menos) en el cambio de densidad. Debido a la alta densidad de genes en isocoras H, parece lógico que la selección actúe con mayor intensidad en contra de los productos de recombinación en isocoras H que en isocoras L.

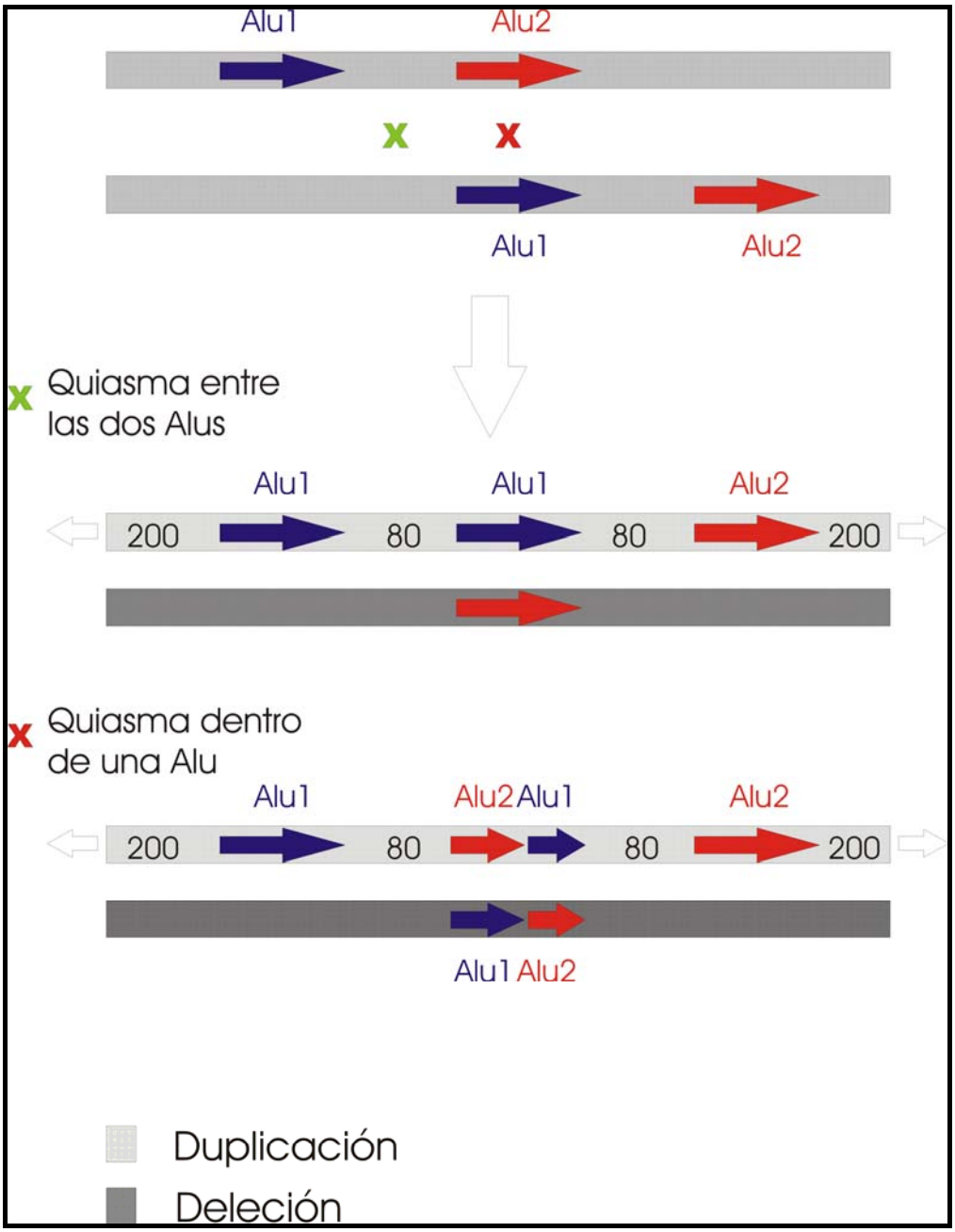

Figura 4-6: Modelo simple de recombinación desigual entre dos Alus. Si el quiasma se forma entre los dos elementos, el resultado es un trímero y un monómero. Si el quiasma se localiza dentro de los elementos, la duplicación la formarán dos Alus enteras y dos fragmentadas (tetrámero), mientras que la deleción estaría formada por dos fragmentos de Alu (dímero).

No obstante, determinar la frecuencia de los dos productos de recombinación directamente con métodos *in silico* es virtualmente imposible, puesto que generalmente las deleciones no parecen dejar ninguna huella en el genoma. En consecuencia, hemos estimado la actividad de la RHD contabilizando solamente las duplicaciones. El caso más sencillo de recombinación es cuando solamente participan dos Alus colocadas en *tandem* (Figura 4-6). Si el quiasma se produce entre las dos Alus, la duplicación estará compuesta por tres elementos colocados en *tandem* (trímero), y por cuatro elementos si el quiasma se produce dentro de las Alus (tetrámero). En nuestro análisis, hemos utilizado solamente los trímeros, ya que el reducido número de tetrámeros encontrados no permite derivar una buena estadística. Para buscar las duplicaciones de Alus en el genoma, hemos diseñado un script en Perl que detecta los trímeros según los criterios expuestos en el apartado 3.3.7 La definición *in silico* de los productos de recombinación. Con los requisitos impuestos a los trímeros, nos aseguramos de que la mayoría de ellos tenga su origen en un proceso de recombinación. No obstante, los trímeros se podrían formar también por azar debido a inserciones casuales que cumplan también con los requisitos impuestos. En consecuencia, para interpretar adecuadamente los datos, habría que comparar el número observado de trímeros con el número esperado por azar. Para estimar las frecuencias esperadas, hemos desarrollado un algoritmo que simula el proceso de inserción de Alus en el genoma. La simulación consta de los siguientes pasos:

1. Se leen todos los elementos en una isocora y se ordenan por edad evolutiva, de más antiguos a más recientes.

2. Se calcula la longitud total de los elementos y se resta de la longitud de la isocora.

3. Se genera una secuencia aleatoria libre de Alus y de la misma longitud que la isocora menos la suma de las Alus.

4. Se inserta el primer elemento (el más antiguo) en la secuencia aleatoria. La posición se elige al azar.

5. Se recalcula la longitud de la secuencia después de la inserción, sumando la longitud del elemento insertado.

6. Se iteran los pasos 4 y 5 para todos los elementos, insertando primero los elementos más antiguos y después los más recientes.

Mediante un programa con este algoritmo, se generan 100 secuencias con Alus insertadas al azar para cada clase de isocora. A continuación, para cada secuencia se determina el número de trímeros, aplicando los criterios descritos más arriba. De esta manera, se calcula el promedio de trímeros que se espera que se formen por azar en cada clase de isocora. Estos promedios los tomaremos como frecuencias esperadas. De la Tabla 4-5 podemos derivar las siguientes conclusiones. Primero, las frecuencias de trímeros observadas en el genoma son mucho más altas que las frecuencias que se esperarían por azar, lo que indica que la recombinación homóloga desigual efectivamente actúa alterando el número total de elementos. Sin embargo, la observación más importante es que las proporciones entre observados y esperados son más altas en isocoras L que en H. Este resultado indica que o bien la frecuencia de recombinación, o bien la probabilidad de supervivencia de los productos de recombinación, son mayores en las isocoras ricas en A+T. Nótese que no hemos medido las frecuencias de las deleciones. Sin embargo, se puede suponer que éstas también ocurren con mayor frecuencia relativa en las isocoras L.

Tabla 4-5: Frecuencias observadas y esperadas de los trímeros de Alus. Las proporciones entre observados y esperados son más altas en isocoras L que en H, lo que significa que los trímeros se forman o mantienen con mayor probabilidad en las isocoras L.

Isocora	# Observados	# Esperados	Obs/Esp
L1	124	1.84	67.39
L2	426	10.43	40.84
H1	804	33.34	24.12
H2	398	15.47	25.73
H3	217	8.40	25.83
H4	64	1.43	44.76
Total	2033	70.91	28.67

Hemos detectado un total de 2033 trímeros en el genoma completo. Este número puede parecer pequeño para explicar el cambio drástico de densidad que muestran las Alus. No obstante, con el modelo empleado en esta sección probablemente solo estamos detectando una parte de los productos de recombinación. Algunos de los factores que pueden dar lugar a esta subestimación son:

- Los requisitos tan restrictivos impuestos a los trímeros, que puede hacer que muchos productos de recombinación pasen “desapercibidos”.

- Los sucesivos procesos de recombinación entre los productos primarios pueden borrar cualquier huella reconocible de nuevas recombinaciones.

- Las Alus también participan en fenómenos de recombinación no-conservadora. Por ejemplo, la recombinación intracromósomica produce una deleción y un trozo de ADN episómico que se pierde antes de la siguiente división celular.

Aunque estos factores pueden llevar a una subestimación del número absoluto de trímeros en el genoma, creemos que no afectan demasiado a las proporciones observados/esperados, que son las que constituyen el punto clave de este análisis.

En resumen, mediante un análisis de las proporciones observados/esperados de los trímeros en función de la isocora, se puede poner de manifiesto que las frecuencias de los productos de recombinación son relativamente más altas en isocoras L que en H (Hackenberg *et al*., 2005). De esta manera, la mayor frecuencia relativa de deleciones en las isocoras L podría conducir al cambio de densidad de las Alus. Nótese, sin embargo, que con estos datos no es posible decidir entre una mayor frecuencia de recombinación (Medstrand *et al*., 2002) o una mayor tasa de supervivencia relativa de los productos de recombinación en las isocoras L (Brookfield, 2001).

4.2.6 Evidencias en favor de la recombinación

Las frecuencias con las que se producen los procesos de recombinación homóloga desigual entre elementos dispersos dependen tanto de la similitud (homología) de secuencia como de la distancia física entre los elementos. Se sabe que una gran similitud y/o una distancia física corta entre las dos secuencias aumentan las tasas de recombinación (Lobachev *et al*., 2000, Medstrand *et al*., 2002). A partir de estas observaciones, hemos obtenido dos indicios experimentales más en favor de la implicación de la RHD en el cambio de densidad de las Alus.

4.2.6.1 Distribución inesperada de las Alus jóvenes

Las Alus se insertan preferentemente en las isocoras L. Las nuevas inserciones son copias exactas de los genes maestros y por tanto estarán libres de mutaciones. En consecuencia, la distancia evolutiva entre el nuevo elemento y su secuencia consenso será nula. Si además de eso, las Alus se excluyeran más frecuentemente de las isocoras L que de las H, se produciría una mayor renovación (turnover) del conjunto de las Alus en las isocoras L. Asumiendo este

escenario, se puede predecir una moderación en la velocidad de evolución composicional de las Alus en las isocoras L. La desaceleración se debería a la entrada continua de nuevas Alus en las isocoras L (con distancia evolutiva nula respecto al consenso) y a la pérdida de otras Alus que probablemente habrían acumulado ya algunas mutaciones. Según este modelo, las Alus estarían sometidas a un proceso de rejuvenecimiento, apareciendo más jóvenes en isocoras L que en H.

Una segunda predicción que se puede hacer es que, si fuese la recombinación el agente que elimina a las Alus, el efecto rejuvenecedor en las isocoras L debería ser más pronunciado en miembros de subfamilias con un origen evolutivo reciente, ya que serían estos los que mostrarían una mayor similitud de secuencia. Esta dependencia de la edad se basa en la correlación positiva que existe entre la frecuencia de recombinación y la similitud de secuencia (Waldman y Liskay 1988; Baker *et al*., 1996).

Tabla 4-6: Distancias evolutivas entre los miembros de 4 subfamilias de Alus y sus correspondientes secuencias consenso. Las dos subfamilias jóvenes (AluYa8, AluYb9) muestran distancias menores en isocoras L que en H, es decir que parecen ser más jóvenes en regiones ricas en A+T. Este comportamiento, que parece contradecir la hipótesis de la presión mutacional local, no se observa en subfamilias más antiguas como AluY y AluJb.

Isocora	**AluYa8**	**AluYb9**	**AluY**	**AluJb**
L1	0.0153 ± 0.021	0.0133 ± 0.018	0.0391 ± 0.023	0.1375 ± 0.040
L2	0.0168 ± 0.030	0.0142 ± 0.023	0.0383 ± 0.026	0.1285 ± 0.041
H1	0.0192 ± 0.036	0.0200 ± 0.031	0.0379 ± 0.027	0.1243 ± 0.042
H2	0.0209 ± 0.027	0.0298 ± 0.039	0.0385 ± 0.026	0.1238 ± 0.049
H3	0.0240 ± 0.041	0.0355 ± 0.045	0.0395 ± 0.037	0.1273 ± 0.060
H4	0.0273 ± 0.039	0.0249 ± 0.029	0.0401 ± 0.043	0.1310 ± 0.069

Combinando ambas predicciones, la hipótesis resultante sería que las Alus pertenecientes a subfamilias con un origen reciente deberían ser más jóvenes en isocoras L que en H. A primera vista, esta hipótesis puede parecer contra-intuitiva, dado que las Alus, que son ricas en G+C, están sometidas a mayor presión mutacional en las isocoras L (véase Figura 4-1). En consecuencia, se esperaría que evolucionasen más rápidamente, y no más lentamente, mostrando por tanto distancias evolutivas más elevadas en las isocoras L que en las H.

Para distinguir entre ambos efectos opuestos, hemos calculado las distancias evolutivas en cuatro subfamilias de distintas edades en función de la

isocora (véase Tabla 4-6) y los coeficientes de regresión entre las distancias evolutivas y el contenido en G+C de la isocora (véase Tabla 4-7).

En las subfamilias jóvenes, AluYa8 y AluYb9, insertadas hace unos 2.5-4 millones de años (Carroll *et al*., 2001; Hedges *et al*., 2004; Otieno *et al*., 2004; Gibbons *et al*., 2004), se observa una correlación positiva entre las distancias evolutivas y el contenido en G+C de la isocora, mientras que no hay correlación en las Alus de edad intermedia (AluY, 20 millones de años). Finalmente, las Alus pertenecientes a una de las subfamilias más antiguas (AluJb, 50-80 millones de años) muestran una correlación negativa. Por tanto, este análisis demuestra que, de acuerdo con nuestra hipótesis, las Alus jóvenes muestran distancias evolutivas menores en isocoras L que en H. Así pues el efecto rejuvenecedor queda limitado a elementos de subfamilias con un origen evolutivo reciente.

Tabla 4-7: Correlación entre la divergencia evolutiva de las Alus y el contenido en G+C de la isocora. Las subfamilias jóvenes muestran una correlación positiva, estadísticamente significativa. Las Alus de una edad intermedia (AluY, aprox. 20 millones de años), no muestran correlación, lo que interpretamos como un estado transitorio entre el "rejuvenecimiento" y el dominio del ajuste composicional. Finalmente, para elementos antiguos (AluJb, aprox. 50-80 millones de años) se observa una correlación negativa, poniendo de manifiesto así la actuación de la presión mutacional a largo plazo.

Familia	Número de isocoras	r	P	Gradiente
AluYa8	490	0.13	0.0030	36.7
AluYb9	551	0.2963	0.0000	77.1
AluY	6934	-0.0017	0.8882	-0.821
AluJb	6822	-0.1796	0.00	-52.1

La dependencia de la edad señala claramente a la recombinación como el agente que elimina las Alus con mayor frecuencia de las isocoras L que de las H. Se sabe que la eficacia de la recombinación está directamente relacionada con la longitud ininterrumpida de la similitud en la secuencia. Cuanto más extenso sea el fragmento similar, más alta será la tasa de recombinación (Waldman y Liskay 1988; Baker *et al*., 1996). Debido a esto, se espera que la recombinación Alu/Alu actúe con mayor intensidad en subfamilias jóvenes (con gran similitud de secuencia entre sí) que en subfamilias antiguas (que habrán sufrido una mayor divergencia). Así pues, al contrario que los demás mecanismos evolutivos, la recombinación actúa especialmente sobre elementos jóvenes y su importancia decae según estos van divergiendo composicionalmente. Este comportamiento de la recombinación explica de una manera concluyente las correlaciones opuestas encontradas en las distintas subfamilias (véase Tabla 4-7). En subfamilias

jóvenes (AluYa8, AluYb9), los elementos tienen una gran similitud de secuencia entre sí, provocando una tasa alta de "rejuvenecimiento". La consecuencia es que los elementos de estas subfamilias aparecen como más jóvenes en isocoras L que en H, dando lugar a las correlaciones positivas que observamos entre la distancia evolutiva y el contenido en G+C de la isocora. Por otro lado, en subfamilias antiguas la recombinación ya no actúa con tanta frecuencia debido al alto número de sustituciones habidas entre los elementos. En consecuencia, la evolución composicional de los elementos viene determinada principalmente por la presión mutacional, lo que se traduce en una correlación negativa. Por último, los elementos de subfamilias con edad intermedia (AluY) no muestran ninguna correlación, lo que interpretamos como un estado transitorio entre el régimen dominado por la recombinación y aquel dominado por la presión mutacional.

En resumen, la distribución inesperada de las distancias evolutivas en subfamilias jóvenes se puede explicar por la acción de la recombinación sobre estos elementos. Una publicación reciente (Callinan *et al*., 2005) parece confirmar este resultado (mediante una comparación entre el genoma humano y el de chimpancé), ya que encuentran un mayor número de deleciones de Alus en regiones ricas en A+T.

4.2.6.2 Las Alus solitarias son más frecuentes en las isocoras L

Aparte de la similitud de secuencia, la recombinación homóloga también requiere cierta cercanía física entre los dos elementos para poder actuar. Las Alus situadas a distancias cortas son más proclives a recombinar que las Alus lejanas. Se sabe que las Alus no están distribuidas uniformemente en el genoma, sino que tienden a agruparse en grumos o "clusters" (Pavlíček *et al*., 2001; Jurka *et al*., 2002, 2004). Sin embargo, existen también muchas Alus fuera de estos "clusters". Mediante un análisis de las Alus "solitarias", se pueden extraer nuevos argumentos en favor de la hipótesis de que la RHD es el principal mecanismo que promueve el cambio de densidad de las Alus.

Para cada Alu en el genoma, hemos determinado su distancia a la Alu más cercana, ya sea en dirección 5' o 3' (distancia NND, nearest neighbour distance). A partir de estas distancias, definimos una Alu solitaria como aquella que muestra una distancia NND mayor de 2 kb. Se toma este valor porque a partir de distancias NND de 2 kb las Alus muestran un máximo de densidad en las isocoras L (véase Figura 4-7).

De esta forma, hemos contabilizado un total de 151.112 Alus solitarias en el genoma humano. Las Alus solitarias, aunque solo sea por su definición, son claramente abundantes en isocoras L. En una publicación reciente de Jurka *et al*.

(2004), se identifican estas Alus solitarias con Alus jóvenes que se han insertado recientemente, y probablemente no estén aún fijadas en el genoma. Sin embargo, solo 27.340 (o 18%) del total de Alus solitarias encontradas en el genoma pertenecen al subgrupo AluY, mientras que el resto son de los subgrupos más antiguos S y J (véase Tabla 4-8).

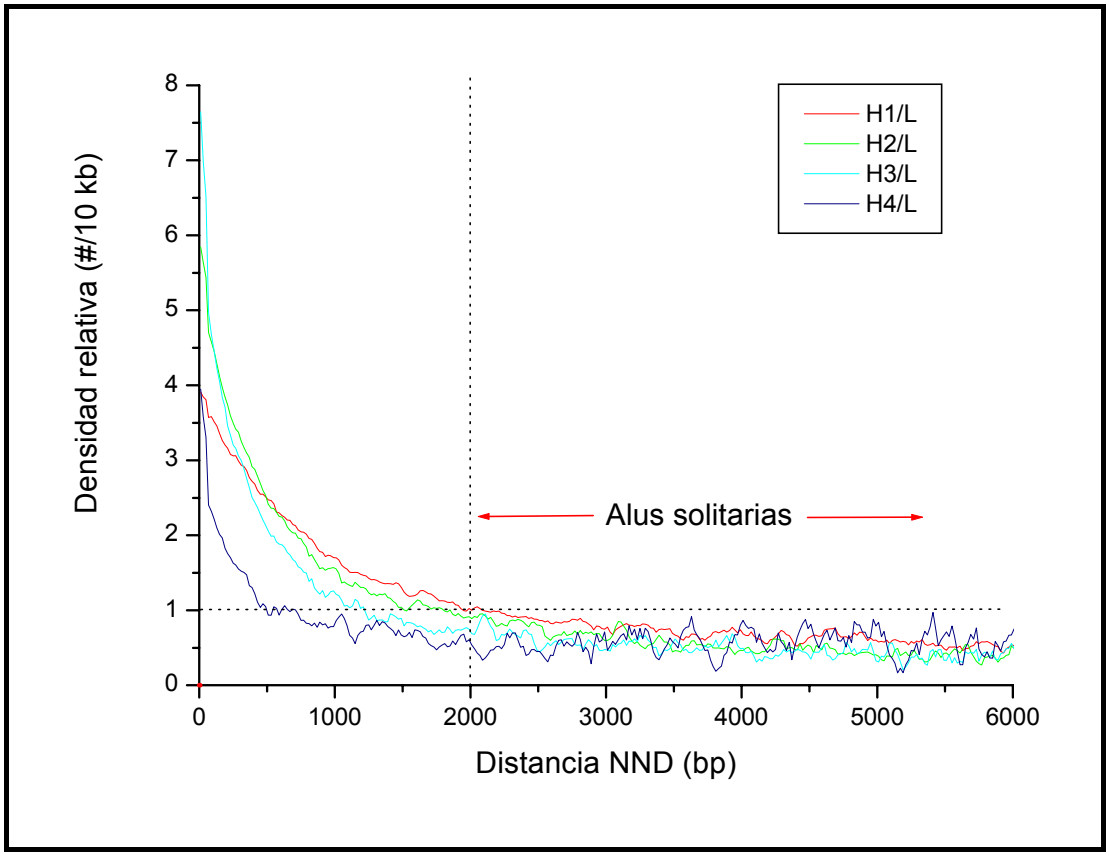

Figura 4-7: Densidad relativa de Alus en función de la distancia a la Alu más próxima (NND). A partir de un valor de NND de aproximadamente 2 kb, el máximo de densidad de las Alus se localiza en las isocoras L (las 4 fracciones son menores de 1).

Este resultado está en contra de que las Alus solitarias pertenezcan mayoritariamente a las subfamilias más jóvenes. En cambio, apoya la acción de la recombinación homologa desigual, ya que ésta tiene lugar preferentemente entre elementos cercanos. Solamente las Alus cercanas entre sí en el cromosoma pueden participar en procesos de recombinación, y en consecuencia, ser eliminadas del genoma a través de este mecanismo. Por el contrario, las Alus solitarias tienen poca probabilidad de participar en procesos de recombinación, y por lo tanto se mantienen durante largos periodos de tiempo en el genoma.

En resumen, el alto número de Alus solitarias pertenecientes a subfamilias antiguas que hemos encontrado en las isocoras L indica probablemente que la recombinación elimina preferentemente las Alus cercanas entre sí, contribuyendo así al cambio de densidad.

Tabla 4-8: Análisis de Alus solitarias en el genoma

Isocora	# Alu	#/10 kb	AluY (%)	AluS (%)	AluJ (%)
L1	60038	0.7139	21.51	57.93	20.14
L2	52424	0.6221	17.46	59.11	23.05
H1	28562	0.4593	14.37	60.78	24.50
H2	6779	0.3774	11.54	62.90	25.31
H3	2320	0.3240	11.25	64.74	23.79
H4	989	0.3740	12.74	60.47	26.49
Todas	151112	0.5850	18.09	59.22	22.30

4.2.6.3 Otros indicios sueltos

Además de los argumentos en favor de la recombinación presentados en las dos secciones anteriores, existen otros, menos rigurosos, que también la avalan. Por ejemplo, en la sección 4.2.4 hemos visto que el exceso de Alus en intrones se limita a las isocoras L. En su momento no discutimos la causa de esta distribución, utilizándola solamente como argumento en contra de que la distribución de los genes influyese sobre la de las Alus. Ahora bien, la distribución diferencial de las Alus en los intrones de isocoras L y H se puede explicar también a través de la acción de la recombinación. En isocoras ricas en G+C, las regiones intergénicas son mucho más cortas que en isocoras L. Por lo tanto, cabe esperar más presión selectiva en contra de las deleciones en regiones intergénicas de las isocoras H que de las L (hay una mayor probabilidad de que la recombinación sea deletérea si la distancia entre dos genes es corta). En consecuencia, la eliminación de elementos debe de ser más frecuente en regiones intergénicas L que en H. Y eso llevaría a un exceso de Alus en los intrones de las isocoras L, pero no en los de las H.

Otro indicio revelador se obtiene a partir de las densidades relativas, mostradas en la Figura 4-2. Las Alus más antiguas, a partir de distancias evolutivas de aproximadamente 0.2 (con CpGs), se acercan a una distribución uniforme en el genoma. Nótese que teóricamente una distribución uniforme vendría dada por valores de 1 en las cuatro densidades relativas (H*/L). La cuestión que surge es: ¿por qué las Alus más antiguas no se acumulan más aún en

las isocoras H? Imaginemos el genoma al principio de la radiación de las Alus. En ese momento había pocas Alus y por tanto las distancias físicas entre ellas debían de ser muy altas. Pero hemos visto que las distancias cortas son un requisito imprescindible para que pueda actuar la recombinación. Puesto que este requisito no se cumplía en los comienzos de la radiación de las Alus, la contribución de la recombinación en ese periodo debió ser baja. En consecuencia, para distancias evolutivas muy grandes se esperaría una distribución más parecida a la distribución inicial, que es justamente lo que se observa en la Figura 4-2

Finalmente, si las deleciones son más frecuentes en regiones ricas en A+T, todos los elementos, y no solo las Alus, deberían acumularse también, aunque lentamente, en regiones ricas en G+C. A primera vista, esa hipótesis parece estar en contra de la fuerte acumulación de los LINE1 en las isocoras L. Sin embargo, la media de la densidad no es un indicador adecuado, y solamente un análisis de las densidades de los LINE1 en función de la distancia evolutiva puede resolver esta cuestión. En la Figura 4-2 (abajo) hemos visto ya las densidades relativas de los LINE1, aunque no hemos discutido este aspecto. Se puede ver claramente que las densidades relativas se aproximan a 1 según va creciendo la distancia evolutiva. Es decir que, aunque el máximo de densidad se mantiene en las isocoras L, los LINE1 se acumulan también, aunque más lentamente, en regiones ricas en G+C. Una comparación entre LINE1 jóvenes y antiguos presentado por Pavlicek *et al*. (2001) ya sugería esta posibilidad.

4.2.7 El dilema del cromosoma Y

Como hemos visto, la recombinación desigual podría ser un mecanismo importante para explicar el cambio de densidad de las Alus (mediante un mayor número relativo de deleciones en las regiones ricas en A+T). El cromosoma Y no tiene homólogo en el genoma y por lo tanto no participa en procesos de recombinación (excepto ciertas regiones cortas llamadas regiones seudo-autosómicas). En consecuencia, la recombinación homóloga desigual no puede eliminar elementos en la mayor parte del cromosoma Y. Por lo tanto, se esperaría una densidad de Alus más alta en el cromosoma Y en comparación con los autosomas. Sin embargo, se han observado densidades más elevadas de Alus en los autosomas (Gu *et al*., 2000; véase también Tabla 4-9), lo que parece estar en contra de la hipótesis de que la RHD pueda participar en el cambio de densidad. No obstante, existen algunas particularidades del cromosoma Y que ofrecen explicaciones alternativas.

La densidad de Alus varía notablemente en función de la isocora (véase Tabla 4-9). Por otro lado, se sabe que el contenido en G+C varía también entre los cromosomas, y por lo tanto se esperarían densidades más bajas en cromosomas ricos en A+T en comparación con los cromosomas ricos en G+C. El cromosoma Y es rico en A+T y, por tanto, la fluctuación de las densidades en función de la isocora podría distorsionar la comparación con los autosomas. Para poner de manifiesto el impacto que tienen los distintos contenidos en G+C de los cromosomas, hemos calculado las densidades tanto para todas las isocoras como solamente para las isocoras L.

Tabla 4-9: Análisis de la densidad de Alus en función del tipo de cromosoma. En primer lugar, se confirma que la densidad de Alus es mayor en los autosomas que en los cromosomas sexuales. Sin embargo, si se limita el análisis a las isocoras L, es decir si se comparan solo regiones con un G+C parecido, esta diferencia desaparece, y los autosomas y el cromosoma Y muestran entonces densidades muy parecidas.

Tipo de cromosoma	**Media de G+C**	**Densidad 10k**	**Densidad 10k (sólo isocoras L)**
Autosomas	42.67	3.669	2.535
X	40.98	2.708	1.952
Y	40.06	2.778	2.626

La Tabla 4-9 muestra la densidad de Alus en función del tipo de cromosoma. En primer lugar, se confirma que las densidades de Alus son menores en el cromosoma Y que en los autosomas. Sin embargo, si comparamos solamente aquellas regiones con un contenido en G+C parecido (isocoras L), esta diferencia desaparece, y las densidades en los autosomas y en el cromosoma Y son entonces muy parecidas. En consecuencia, el G+C del cromosoma es importante a la hora de comparar entre distintos cromosomas.

Por otra parte, aunque la RHD no puede actuar en el cromosoma Y, existen otros tipos de recombinación, como la intracromosómica, que causa también deleciones en el cromosoma Y (Bailey *et al.*, 2003; Prak y Kazazian, 2000). Además, trabajos recientes han mostrado que la conversión génica (en este caso recombinación intracromosómica entre los palíndromos que existen en el cromosoma Y) mantiene la gran similitud de secuencia que se observa entre ciertas regiones dentro del cromosoma Y (Skaletsky *et al.*, 2003; Rozen *et al.*, 2003). En teoría, la conversión génica no altera el número de nucleótidos. No obstante, en caso de un emparejamiento incorrecto, este mecanismo también podría causar deleciones y duplicaciones. De hecho, en estas regiones, a pesar de ser más ricas en G+C, se observan densidades de Alus menores que en otras

regiones del cromosoma Y. En consecuencia, este mecanismo específico del cromosoma Y puede que elimine los elementos de ciertas regiones del cromosoma Y (Jurka *et al*., 2004).

En resumen, lo que intentamos mostrar en esta sección es que las densidades absolutas de Alus en el cromosoma Y no necesariamente contradicen la hipótesis de la recombinación. Primero, una comparación entre regiones con un G+C parecido (isocoras L) muestra que la menor densidad de Alus en el cromosoma Y se debe a la propia riqueza en A+T del cromosoma Y, y la consecuente repercusión sobre las densidades de las Alus. En segundo lugar es probable que existan mecanismos específicos en el Y. Como ejemplo se menciona la posible eliminación de TEs mediante la recombinación intracromosómica.

4.3 La clusterización de las Alus

Como hemos expuesto más arriba, las densidades de las Alus fluctúan notablemente en función de la isocora (véase Tabla 4-2). Esta variación se puede interpretar como una distribución espacial heterogénea de las Alus – más densidad en unas regiones y menos en otras, a lo largo de un cromosoma (véase Figura 4-8).

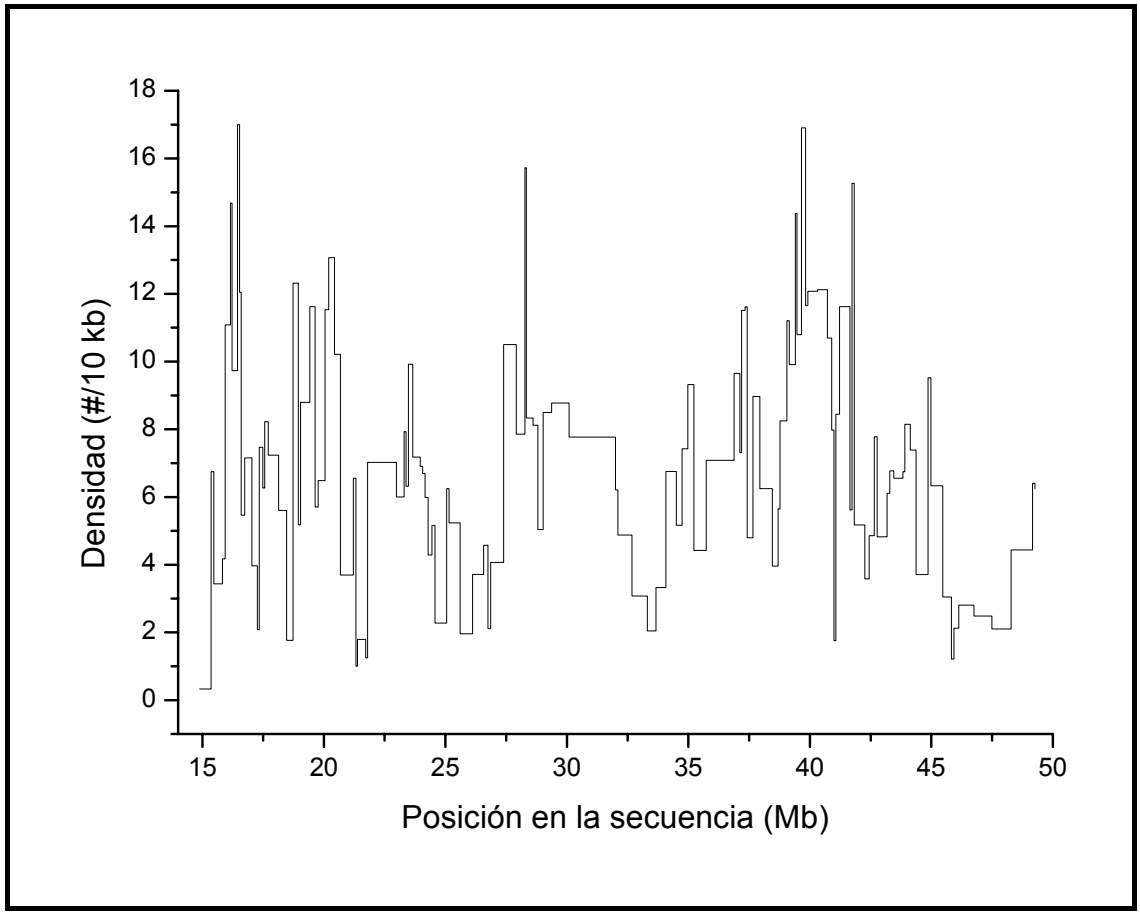

Figura 4-8: Fluctuación de la densidad de las Alus en función de la posición en la secuencia (cromosoma 22). Se observa una marcada heterogeneidad espacial en las diferentes isocoras situadas a lo largo del cromosoma.

La densidad determina la distancia media entre Alus vecinas en una región dada. En las isocoras H3, por ejemplo, la distancia media es de unos 1445 bp, mientras que en las isocoras pobres en Alus (isocoras L) la distancia media alcanza los 3850 bp. Sin embargo, la distancia media no permite por sí misma llegar a ninguna conclusión acerca de la forma de la distribución de distancias. Esta distribución es el punto de conexión con otra medida para cuantificar la distribución: el análisis espacial de patrones de puntos. En muchos campos de investigación donde se emplean estos métodos, los elementos se consideran puntos, es decir sin dimensión. Mediante este tipo de análisis, es posible determinar si los elementos distribuidos en el espacio[5] se atraen, se repelen o si muestran una distribución aleatoria. La Figura 4-9 muestra un ejemplo de la

[5] En este trabajo el espacio es unidimensional (la cadena de ADN), sin embargo en otros campos de investigación se consideran también 2 o 3 dimensiones.

distribución espacial de las Alus. Cada Alu se representa mediante un trazo vertical, como si se tratase de un código de barras. Se muestra la distribución en dos isocoras (L2 y H4), ambas localizadas en el cromosoma 19. La distribución en la isocora L2 parece ser más aleatoria, sin grandes aglomeraciones ("clusters") de Alus. Por el contrario, en la isocora H4 se observan tanto clusters de Alus como varias regiones vacías. Nuestro objetivo en este apartado es cuantificar tanto la distribución espacial como las diferencias que pueda haber entre las distintas clases de isocora, con el fin de identificar los mecanismos evolutivos implicados en la formación de estos patrones espaciales.

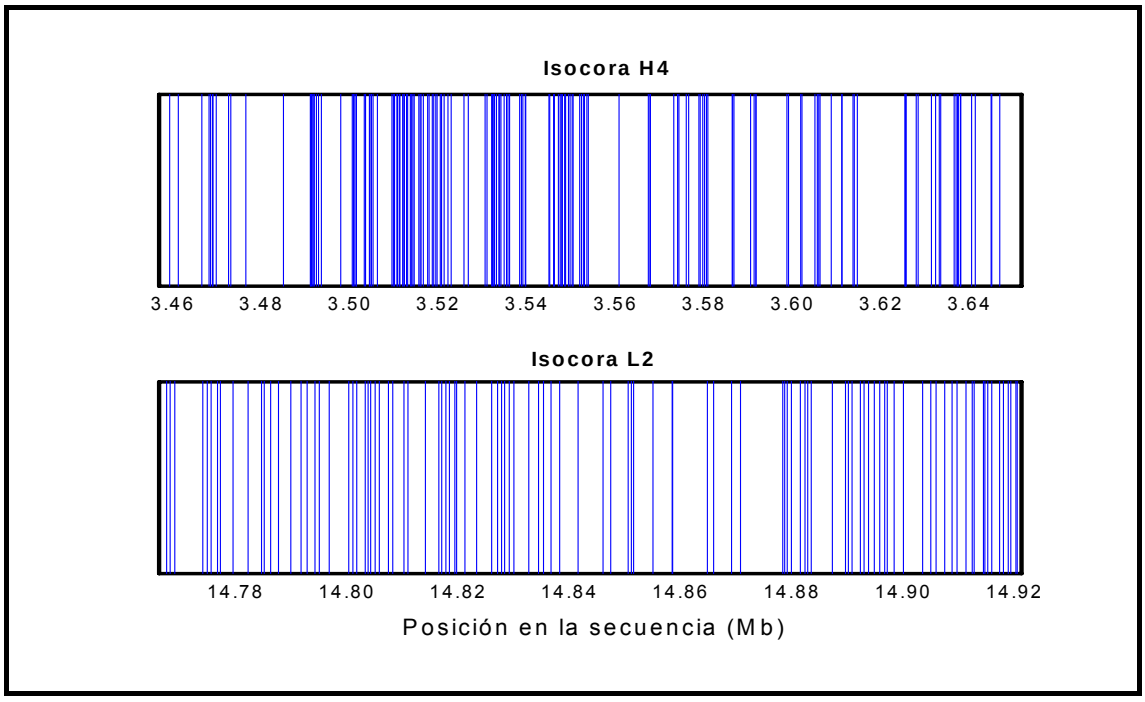

Figura 4-9: Distribución espacial de las Alus en función de la posición en la secuencia. Cada trazo vertical representa una Alu. Arriba se muestra una isocora H4 (163 Alus, 8.37 Alus por 10 kb) y abajo una L2 (105 Alus; 6.78 Alus por 10 kb). Ambas isocoras son del cromosoma 19. Se puede apreciar visualmente una diferencia clara entre las dos isocoras. El objetivo de esta sección consiste en cuantificar estas diferencias.

Si se parte de la suposición razonable de que la inserción de Alus en el genoma es aleatoria, cualquier desviación del patrón aleatorio puede aportar nuevos datos sobre la dinámica evolutiva de las Alus y sobre los mecanismos implicados.

En distintas publicaciones se ha llegado a la conclusión de que las Alus están clusterizadas en el genoma, es decir se atraen y forman grumos o "clusters" (Jurka *et al*., 2004; Pavlicek *et al*., 2001; Slagel *et al*., 1987). Según estos

autores, la clusterización es más pronunciada entre Alus más antiguas y en regiones ricas en G+C. Jurka *et al*. (2004) interpretan la formación de clusters como el resultado de un mecanismo de selección por vía paterna que elimina las Alus más jóvenes de los sitios cromosómicos no-neutros. El mecanismo que lleva a la clusterización sería también el responsable del cambio de densidad, es decir el que provocaría la acumulación rápida de Alus en regiones ricas en G+C.

Sin embargo, Jurka y colaboradores no emplean un método riguroso para medir la clusterización, ya que es dependiente de la densidad. En concreto, lo que hacen es medir la densidad de Alus en una ventana de 25 kb a ambos lados de cada Alu en el genoma. Mediante una comparación entre subfamilias jóvenes y antiguas en función del G+C, derivan conclusiones sobre la clusterización de las Alus. En primer lugar, observan una densidad más alta de Alus alrededor de elementos antiguos, lo que interpretan como tendencia de las Alus a formar clusters. En segundo lugar, observan que la densidad de Alus muestra una correlación positiva con el G+C, y por tanto una "clusterización" más fuerte en isocoras H. Aunque este método se parece bastante a otras medidas locales de patrones espaciales (como la *K* de Ripley, 1981), no comparan las densidades observadas con las esperadas por azar, ni facilitan valores de significación. Además, el uso de ventanas introduce un grado de subjetividad en el análisis.

Aquí proponemos tres mejoras metodológicas para medir la clusterización de las Alus de una forma fiable. Primero, hemos empleado el programa IsoFinder para segmentar el genoma humano en isocoras. Segundo, proponemos un método independiente de la densidad para medir los niveles globales de clusterización en una isocora. Y tercero, establecemos intervalos de confianza mediante la simulación del proceso de inserción. Este último punto permite, por un lado, distinguir entre clusterización efectiva y clusterización debida al azar. Pero además, mediante distintos modelos de inserción, permite analizar también el papel de los diferentes mecanismos evolutivos impicados en la dinámica de las Alus.

4.3.1 Medida de la clusterización de Alus

Para cuantificar la distribución espacial a partir de las distancias entre las Alus, aplicamos un método derivado de la física de sistemas desordenados (Carpena *et al*., en preparación). El método tiene su base en la teoría de matrices aleatorias aplicadas al análisis de distancias entre elementos en un conjunto de datos (Metha, 1991). Carpena y colaboradores han extendido esta teoría desde el dominio cuántico, donde la atracción no actúa (por el principio de exclusión de Pauli) a sistemas microscópicos donde la atracción es una posibilidad. Según este

método, se puede cuantificar la distribución espacial de los elementos (las Alus en nuestro caso) mediante un análisis de la distribución normalizada de las distancias entre Alus. Es decir, se puede determinar si las Alus se atraen, se repelen o se distribuyen aleatoriamente.

Tabla 4-10: Patrones espaciales de las Alus deducidos mediante la desviación estándar de la distribución normalizada de distancias

NC	**Distribución espacial**
1	Aleatoria
< 1	Repulsión
> 1	Atracción/Clusterización

La aplicación del método exige una serie de pasos. Primero, en cada isocora se calcula la distancia de cada Alu a la Alu más próxima en sentido 3'. Segundo, se normaliza la distribución de distancias dividiendo cada distancia observada por la distancia media. Este paso es crucial, ya que elimina la dependencia entre la medida de clusterización y la densidad. Como se menciona más arriba, la densidad de Alus es más alta en isocoras ricas en G+C y por lo tanto la distancia media será más corta. Por lo tanto, si no se normaliza, la distribución tendría una media distinta en cada isocora (según la densidad), y eso llevaría a que no podríamos comparar los resultados. Sin embargo, una vez normalizada, la distribución tiene media 1, independientemente de la densidad de Alus en la isocora, y toda la información sobre la distribución espacial de las Alus está entonces contenida en la forma geométrica que tenga dicha distribución. Así que el tercer paso consiste en determinar dicha forma. Existen principalmente dos posibilidades: 1) ajustando la distribución; y 2) calculando la desviación estándar. En este trabajo utilizamos exclusivamente la desviación estándar: según su valor, se pueden derivar conclusiones acerca de la distribución espacial de las Alus (véase Tabla 4-10). Valores de 1 indican que los elementos se distribuyen aleatoriamente, mientras que valores menores o mayores que 1 indican repulsión o atracción, respectivamente. A la desviación estándar de una distribución normalizada de distancias la llamaremos "nivel de clusterización" (NC). Nótese que cuanto más se apartan de 1 los valores de NC, más marcada será la atracción o repulsión.

El método presentado aquí es una medida "global" de la clusterización en una región. En este trabajo, la región viene dada por la isocora. Es importante destacar que este método es completamente independiente de la densidad, y por lo tanto permite la comparación del nivel de clusterización entre distintas clases de isocora sin que el resultado se vea distorsionado por el sesgo en la densidad.

4.3.2 Clusterización observada

Hemos aplicado el método presentado en la sección anterior a cada isocora del genoma humano que cuente con al menos 30 distancias (31 Alus). En la Tabla 4-11 se muestran los resultados. En primer lugar, se observa una correlación positiva entre el nivel de clusterización y el G+C de la isocora (véase Figura 4-10), con lo que el máximo de clusterización de Alus se localiza en la isocora H4. Este resultado confirma que la clusterización y la densidad son dos medidas independientes, puesto que las Alus presentan su máximo de densidad en las isocoras H3. Así pues, podemos sacar dos conclusiones. En primer lugar, observamos un sesgo de clusterización en función de la isocora, lo que probablemente indica que hay algún mecanismo evolutivo actuando con mayor intensidad sobre las Alus en las isocoras ricas en G+C (véase Tabla 4-11). En segundo lugar, el hecho de que los máximos de clusterización y densidad de Alus se localicen en isocoras distintas, sugiere que los mecanismos implicados en el cambio de densidad y en la formación del sesgo de clusterización son diferentes.

Tabla 4-11: Nivel observado de clusterización por isocora, número de isocoras que entran en el análisis y distancia media entre Alus

Isocora	NC	# isocoras	Distancia media
L1	1.120	1065	4690
L2	1.224	1959	2993
H1	1.335	2084	1807
H2	1.470	737	1494
H3	1.680	305	1445
H4	1.836	79	1909

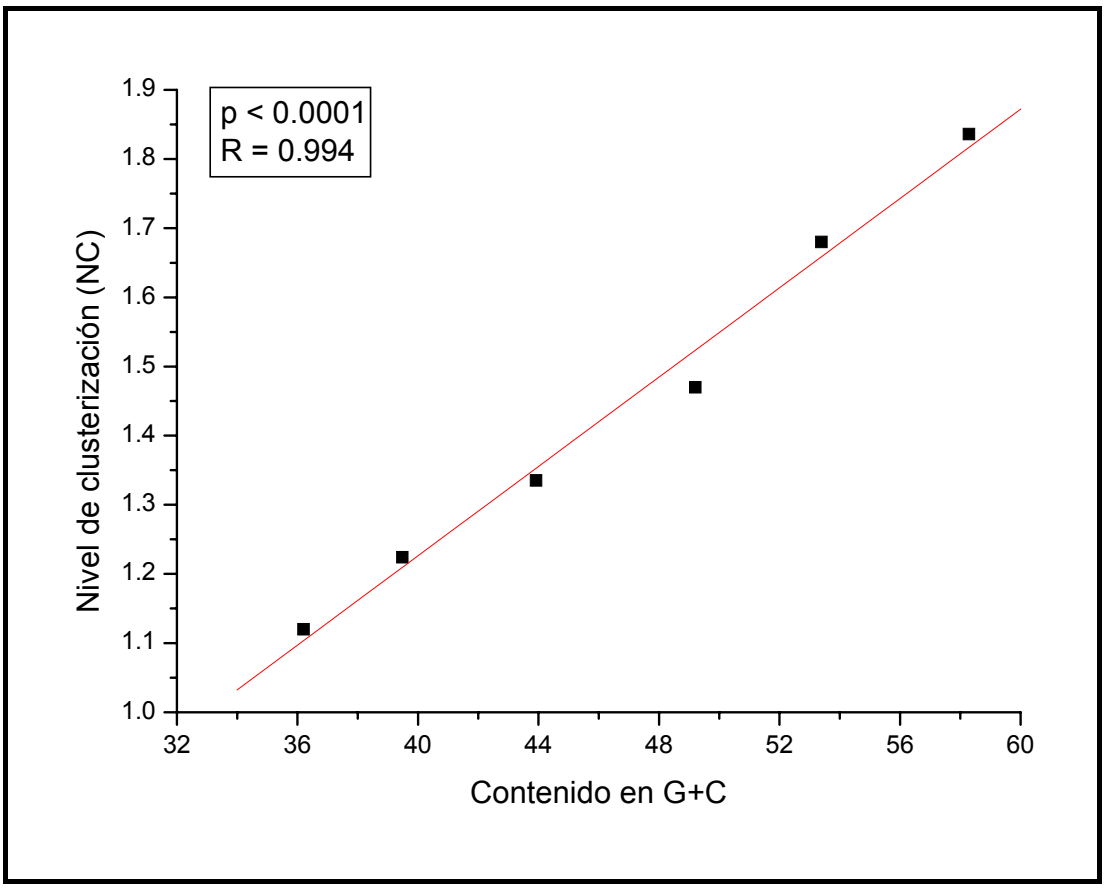

Figura 4-10: Nivel de clusterización de las Alus (NC) en función del contenido en G+C de la isocora.

Nuestro siguiente objetivo será investigar qué mecanismo evolutivo puede estar implicado en la formación del sesgo de clusterización. Como hemos mencionado más arriba, para ello se han implementado distintos modelos de inserción, que nos van a permitir tanto establecer intervalos de confianza como valorar el impacto del mecanismo biológico en que se basa el modelo.

4.3.3 Simulación del proceso de inserción de Alus

En teoría, cualquier nivel de clusterización mayor de 1 indica atracción, mientras que un valor menor de 1 refleja repulsión (véase Tabla 4-10). Ahora bien, esta observación solamente es válida si el número de distancias fuera infinito. Esto quiere decir que si calculamos la desviación estándar de un número finito de distancias, por azar puede resultar un nivel de clusterización ligeramente diferente de 1, aunque el proceso subyacente debería llevar a una distribución de Poisson con una desviación estándar de 1. Así pues, para poder distinguir entre clusterización efectiva y clusterización debida al azar, hay que asociar un nivel

de significación ('P-value') a cada valor de clusterización. Para ello, se han diseñado distintos algoritmos parametrizados que simulan el proceso de inserción de Alus. Comparando el nivel observado y el que obtenemos mediante la simulación, se puede decidir si la clusterización observada se debe o no al azar (véase la próxima sección para una descripción de como se establecen los niveles de significación).

En total hemos diseñado 4 modelos distintos con una estructura jerárquica (véase Figura 4-11). Esto quiere decir que se empieza con un modelo básico y después se van añadiendo nuevas características o restricciones (que corresponden a los diferentes mecanismos evolutivos). El supuesto básico en todos los modelos es que la inserción de Alus en el genoma es aleatoria. Posteriormente, se van añadiendo restricciones. El primer paso consiste siempre en localizar las Alus dentro de la isocora y ordenarlas según su edad evolutiva. El orden de inserción también es el mismo en los distintos modelos: empezamos con la Alu más antigua y después, por orden cronológico, vamos insertando Alus cada vez más jóvenes.

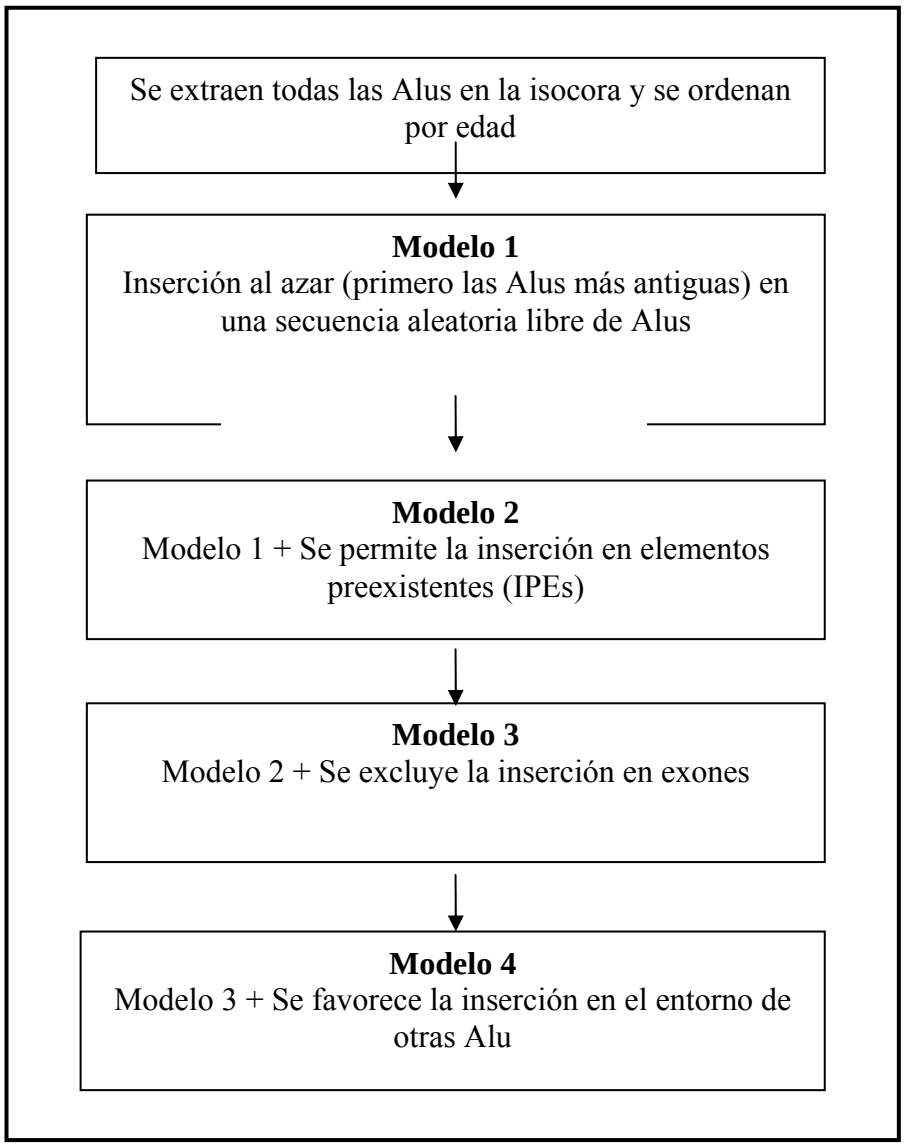

Figura 4-11: Estructura jerárquica de los modelos para simular la inserción de Alus

4.3.3.1 Modelo 1: Inserción aleatoria de puntos

En este modelo tratamos las Alu como objetos sin dimensión. Es decir, no se pueden producir inserciones dentro de otras Alus, como ocurre a menudo en el genoma. Además, insertamos las Alus sin restricción alguna en el área de simulación. Esta área tiene la longitud de la isocora menos la longitud total de las Alus. Este modelo se justifica porque la inserción de Alus en el genoma podría ser aleatoria e independiente del contenido en G+C (Smit, 1999; Arcot *et al.*, 1998).

4.3.3.2 Modelo 2: Inserción en elementos pre-existentes (IPEs)

Este paso trata de conseguir un modelo de inserción más realista. Las Alus se tratan como objetos con dimensión, es decir con longitud, lo que permite que se produzcan inserciones dentro de Alus preexistentes (IPEs). El resultado de un proceso así serían tres Alus muy cercanas (2 fragmentos de Alu en los flancos y una Alu "entera" en medio, véase Figura 4-12).

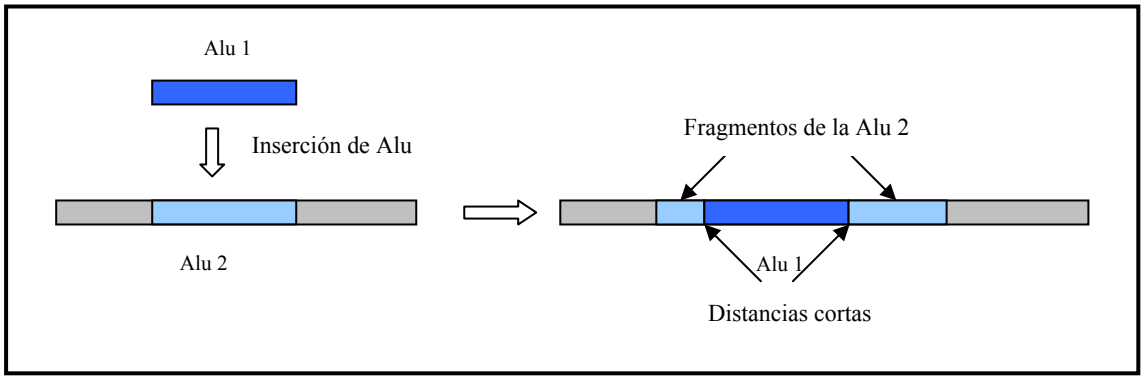

Figura 4-12: Esquema de una inserción de Alus dentro de Alus preexistentes (IPE)

La longitud inicial del área de simulación viene dada nuevamente por la longitud de la isocora menos la longitud total de las Alus. Sin embargo, como las Alus en este modelo tienen longitud, se reajusta el área de simulación después de cada inserción. Es decir, se suma la longitud de la Alu insertada al área de simulación. Si una inserción propuesta cae dentro de un elemento preexistente (es decir, ya insertado anteriormente), se acepta la inserción con una probabilidad P_{IPE}. En este trabajo, a esta probabilidad le asignamos el valor 1 independientemente de la isocora, aceptando pues todas las IPEs.

4.3.3.3 Modelo 3: Exclusión de la inserción en exones

Para acercarnos aún más a un escenario realista de inserción de Alus, en este modelo se toma en cuenta también la estructura génica. En una primera aproximación, podemos suponer que todos los genes existían ya cuando empezó la radiación de las Alus, y que por lo tanto los genes imponían restricciones a la inserción de Alus. Se supone que la gran mayoría de las inserciones de Alus dentro de un exon son eliminadas por selección natural. Se sabe, sin embargo, que existen Alus, o fragmentos de Alus, en algunos ARNm. Pero probablemente la incorporación de un elemento transponible en una región codificadora se debe a la exonización de Alus intrónicas y no a la inserción de Alus dentro de exones funcionales (Sorek *et al*., 2002 y 2004; Lev-Maor *et al*., 2003).

Así pues, los exones siguen siendo regiones genómicas en las que la inserción de Alus no se puede mantener. Además, se sabe que tanto la densidad de genes como la longitud de sus intrones muestran una correlación, positiva y negativa, respectivamente, con el contenido en G+C de la isocora (los intrones son más cortos en regiones ricas en G+C). En consecuencia, tanto la estructura génica[6] como el sesgo de la densidad de genes en función de la isocora pueden tener algún impacto sobre el nivel de clusterización.

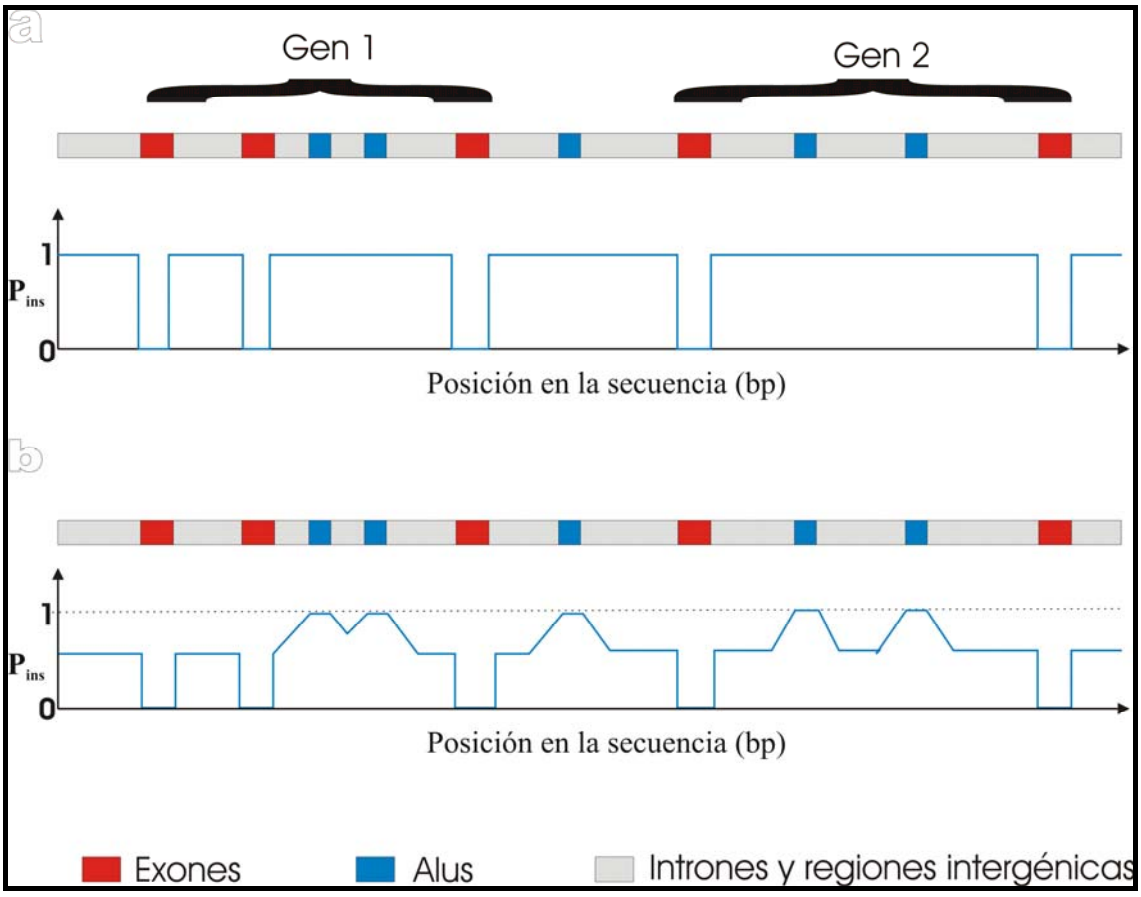

Figura 4-13: La probabilidad de inserción en el modelo 3 (arriba) y modelo 4 (abajo). Nótese que se rechaza cualquier inserción propuesta dentro de un exon (probabilidad 0).

Para definir los loci donde se excluyen las inserciones, hemos utilizado las coordenadas de los genes RefSeq (véase sección 3.1 Secuencias genómicas y tablas de genes). Se inicializa el área de simulación marcando los sitios ocupados por los exones. Las coordenadas de los exones se calculan a partir de sus

[6] La menor longitud de intrones en regiones ricas en G+C podría inducir un mayor agrupamiento de Alus en las isocoras H.

coordenadas genómicas menos la longitud total de las Alus que haya en dirección 5'. De este modo, las nuevas coordenadas representan una aproximación a la situación antes de la inserción de Alus en el genoma.

En este modelo se rechaza cualquier inserción propuesta dentro de un exon y se aceptan todas las demás. En la Figura 4-13 (arriba) se muestra la probabilidad de inserción en función de la posición en la secuencia (isocora).

4.3.3.4 Modelo 4: Acumulación preferente en el entorno de otras Alus

Se sabe que las Alus se colocan a menudo en "*tandem*" en el genoma, con distancias cortas entre sí. Estos "*tandems*" probablemente se originan por la inserción o acumulación preferente cerca de una Alu preexistente (Stenger *et al.*, 2001; Lobachev *et al.*, 2000). En este modelo se toma en cuenta este hecho, favoreciendo la inserción cerca de otra Alu preexistente de forma diferencial en las distintas clases de isocora. Así pues, en primer lugar introducimos una probabilidad en función de la distancia entre la posición propuesta de inserción y la Alu preexistente más cercana (tanto en sentido 5' como en 3'). En concreto, la probabilidad está compuesta por dos términos. El primero describe una probabilidad decreciente linealmente con la distancia:

$$P(d) = 1 - a \cdot d \tag{4-2}$$

donde d es la distancia entre la posición propuesta de inserción y la Alu preexistente más cercana y a es el gradiente; este se elige de manera que $P(d)$ se haga 0 a una distancia de 3 kb (que es aproximadamente la distancia media entre las Alus). Nótese que este término hace que solo se permitan inserciones dentro de una ventana de 3 kb a ambos lados de una Alu preexistente. Además, no depende de la isocora, ya que el gradiente es el mismo para las distintas isocoras.

Con objeto de introducir un término característico para cada clase de isocora y no impedir totalmente cualquier inserción fuera de una ventana de 3 kb, lo que hacemos es solapar el término lineal con una probabilidad constante e individual para cada tipo de isocora. Así pues, la probabilidad a lo largo del área de simulación viene dada finalmente por:

$$P(x) = \max\left[P(d); P_{iso}\right] \tag{4-3}$$

La Figura 4-13 (abajo) muestra la probabilidad de inserción que resulta de la ecuación (4-3). Nótese que P_{iso} regula la intensidad con la que se favorece una inserción cerca de una Alu preexistente. Se puede decir que cuanto más pequeño es el valor de P_{iso}, más se fuerza la acumulación.

Queda ahora por definir la probabilidad P_{iso} para las distintas clases de isocora. Hay dos maneras de hacerlo. Una primera posibilidad sería elegir la probabilidad de manera que descienda linealmente, forzando así una acumulación mayor en las isocoras H que en las L (véase la Tabla 4-12). De esta forma, asignaríamos una probabilidad fija a cada clase de isocora. Otra posibilidad es definir una probabilidad "individual" para cada isocora física:

$$P_i = 1 - Den_i \tag{4-4}$$

donde Den_i es la densidad de genes en la isocora con índice i (la densidad en este caso es la fracción de ocupación de genes en la isocora). Nótese que de esta forma se vincula la intensidad con la que se favorece la inserción en el entorno Alu, a la densidad de genes en la isocora correspondiente. Cuanto más alta sea la densidad de genes, más se fuerza la acumulación en el entorno Alu. Puesto que los genes muestran densidades más altas en las isocoras H, esta manera de definir las probabilidades parece una elección razonable.

Tabla 4-12: Valores de P_{Iso} utilizados en este trabajo

Isocora	P_{Iso} (fija)	P_{Iso} (dependiente de la densidad de genes)
L1	0.7	0.864
L2	0.6	0.846
H1	0.5	0.836
H2	0.4	0.770
H3	0.3	0.669
H4	0.2	0.588

4.3.4 Determinación de los niveles de significación

Para establecer un nivel de significación, hay que formular una hipótesis nula bien definida. La hipótesis nula es la probabilidad con la que se espera encontrar un valor dado. Cuando la distribución de una variable se desconoce, es decir cuando no se deja expresar analíticamente, la única solución consiste en llevar a cabo una simulación numérica de acuerdo con la hipótesis nula. En nuestro caso, la hipótesis nula se especifica en un modelo de inserción, donde se establecen las condiciones con las que se genera la distribución espacial de las Alus. Si se simula la inserción de Alus repetidas veces, se obtiene una distribución empírica de los niveles esperados de clusterización. A partir de esta distribución, se puede derivar el nivel de significación (valor-p):

$$p < 1 - \frac{\sum_{i=1}^{N} d(i)}{N} \tag{4-5}$$

donde *d(i)* es la función indicador; esta vale 1 si el NC observado es mayor que el NC_i (nivel de clusterización obtenido en la simulación i) y 0 en caso contrario. Se toman como estadísticamente significativos los niveles de clusterización con valor-p menor de 0.05. Para cada isocora y modelo de inserción hemos llevado a cabo 1000 simulaciones; así pues, un valor-p menor que 0.05 corresponde a una situación en la que menos de 50 NC_i superan el nivel observado de clusterización.

4.3.5 Exceso de clusterización

Mediante la simulación del proceso de inserción, se puede calcular el número de isocoras con un nivel de clusterización significativo. Nótese que el valor-p que asignamos siempre está vinculado a un modelo de inserción. Así pues, si observamos clusterización significativa (bajo las condiciones de un modelo dado) eso indica que no se puede explicar la clusterización observada mediante los mecanismos evolutivos contemplados en el modelo. Por otro lado, el número de isocoras donde desaparece la clusterización significativa nos indica el grado en que los mecanismos biológicos contemplados en el modelo están implicados en la formación de la clusterización. Para cuantificar este aspecto, definimos el exceso de clusterización de la siguiente manera:

$$Ex_{Clust} = \frac{\#_{Iso}\,(p < 0.05)}{\#_{Iso}} \tag{4-6}$$

donde $\#_{Iso}$ es el número de isocoras (de una determinada clase) y $\#_{Iso}$ ($p < 0.05$) es el número de isocoras que muestran un nivel significativo de clusterización. De esta forma, el exceso de clusterización indica la proporción del número de isocoras que muestran un nivel de clusterización significativo. Este coeficiente puede tomar valores entre 0 y 1. Un valor de 0 indica que la clusterización observada se puede explicar mediante el modelo, mientras que un valor de 1 indicaría que el modelo no es capaz de explicar toda la clusterización observada.

La Figura 4-14 muestra el exceso de clusterización en función de la isocora para los 4 modelos de inserción (hay 5 curvas porque para el modelo 4 consideramos 2 probabilidades P_{Iso} distintas). Con el modelo más básico (1, inserción de puntos), se observa un sesgo entre isocoras L y H. Casi la mitad de las isocoras L no muestran clusterización significativa, mientras que en el 94% de las isocoras H4 la clusterización observada es estadísticamente significativa.

Esto significa que, como era de esperar, debe haber otros mecanismos adicionales implicados en la generación de la clusterización observada. Además, esos mecanismos deben actuar con mayor intensidad en las isocoras H.

Bajo las condiciones del modelo 2, vemos que el número de isocoras con clusterización significativa sigue bajando, lo que sugiere que las IPEs desempeñan algún papel en el origen de la clusterización. Además, se observa que el impacto es más pronunciado en isocoras L que en H, puesto que en L1 la diferencia con el modelo anterior es de un 16%, mientras que en H4 la diferencia se reduce al 7 %.

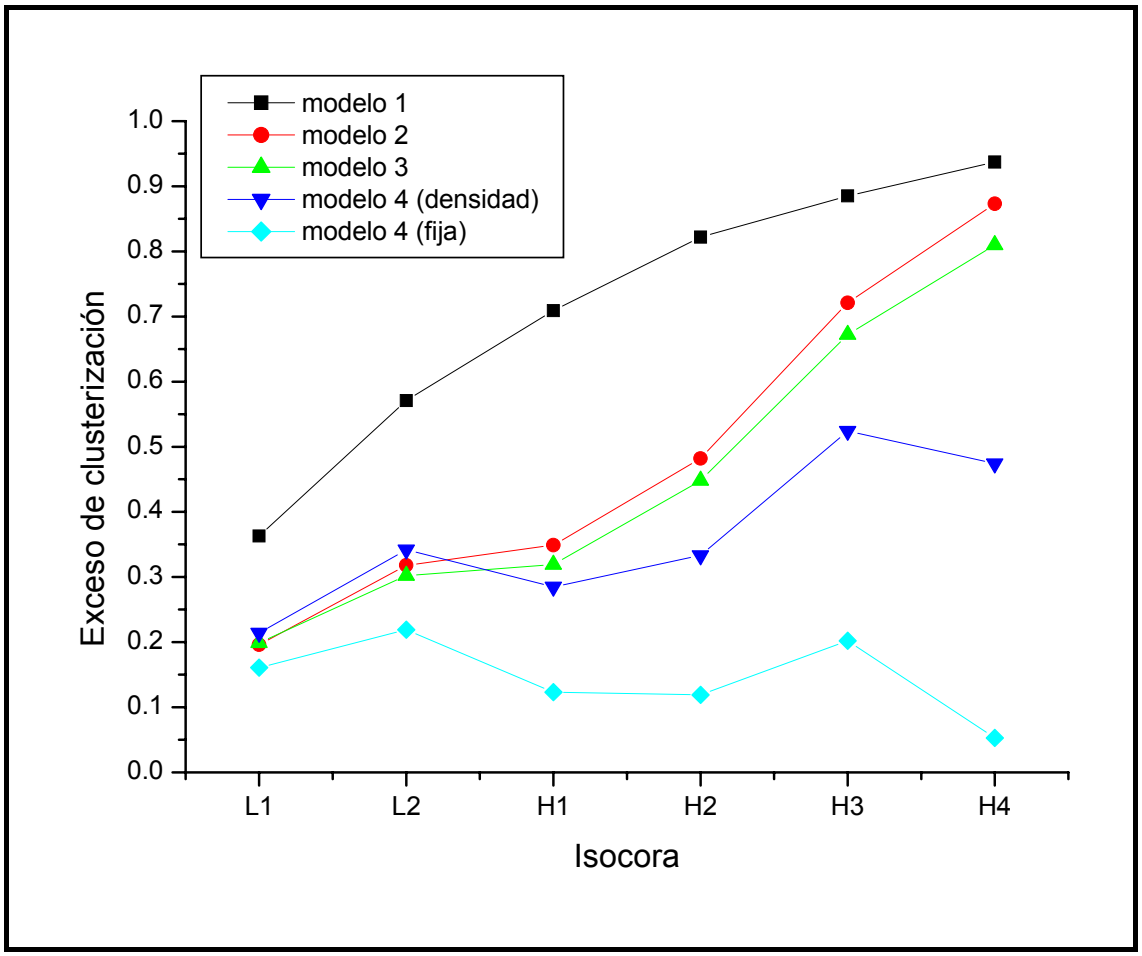

Figura 4-14: Exceso de clusterización en función de la isocora para los 4 modelos de inserción de Alus. Recuérdese que definimos los valores de P_{Iso} de dos maneras distintas: la primera desciende linealmente con el G+C de la isocora (fija), mientras que la segunda depende de la densidad de genes (véase Tabla 4-12 y ecuación (4-4).

Cuando se toma en cuenta la estructura génica (modelo 3), se observa una disminución ligera del exceso de clusterización. El impacto de la exclusión de inserciones de Alus en exones muestra una correlación positiva con el G+C de la isocora. Es decir, mientras que en las isocoras L1 no cambia prácticamente nada, en las isocoras H4 casi un 7% de las isocoras no muestran clusterización significativa en comparación con el modelo 2. Este resultado es coherente ya que existe una correlación positiva entre la densidad de genes y el contenido en G+C de la isocora.

Hasta ahora hemos visto que los tres modelos más básicos pueden explicar en gran parte la formación de clusters en regiones ricas en A+T; sin embargo, en regiones ricas en G+C la clusterización es más marcada, y por lo tanto debe estar involucrado algún otro mecanismo. Este parece ser la acumulación preferente en el entorno Alu. Bajo este modelo, sobre todo para las probabilidades P_{Iso} fijas, el exceso de clusterización baja marcadamente en las isocoras H (véase Figura 4-14). Recuérdese que dentro de este modelo hemos definido dos probabilidades P_{Iso} distintas. Observamos que en el caso de calcular la probabilidad P_{Iso} a partir de la densidad de genes, el exceso de clusterización baja notablemente en las isocoras H, aunque no baja tanto como en caso de las P_{Iso} fijas. Esto demuestra que con el modelo 4 es posible describir los niveles de clusterización, eligiendo adecuadamente la probabilidad P_{Iso} para cada isocora física.

Antes de resumir brevemente esta sección, queremos hacer algunas precisiones acerca de los modelos utilizados. Primero, hay que mencionar que debido al alto coste computacional del modelo 4, no ha sido posible calcular el nivel de clusterización para todas las isocoras del genoma. Así pues, los resultados que presentamos en la Figura 4-14 (para el modelo 4), corresponden solamente a las isocoras de los cromosomas 19, 20, 21, 22, X e Y. Por lo tanto, el exceso de clusterización correspondiente al modelo 4 podría cambiar si se tomase en consideración el genoma entero. Sin embargo, puesto que el análisis incluye un número suficiente de isocoras, la tendencia general observada bajo este modelo probablemente no cambiaría demasiado (véase Apéndice D, Tabla D-6-2). Hay que mencionar también que bajo el modelo 4 (con probabilidades fijas) siguen existiendo isocoras con un nivel de clusterización sin explicar (entre el 12% en H2 y el 22% en L2). Sin embargo, tampoco era de esperar que desapareciera toda la clusterización significativa. Primero, teniendo en cuenta la alta complejidad de la estructura del genoma, los modelos aplicados en este análisis seguramente siguen simplificando el proceso real de inserción. Además, hemos considerado solamente los exones anotados en la tabla de los genes RefSeq. En consecuencia, los exones sin anotar podrían estar provocando una subestimación del nivel real de clusterización. Lo mismo vale también para el

ADN no-codificador conservado entre especies. Se puede suponer que las inserciones dentro de este tipo de ADN serían igualmente seleccionadas en contra.

Pese a la simplicidad de estos modelos de inserción, la observación clave es que mediante el modelo 4 se puede explicar el nivel observado de clusterización en la gran mayoría de las isocoras. Parece ser por tanto que la acumulación diferencial de Alus en el entorno Alu desempeña un papel importante en la formación del alto nivel de clusterización en las isocoras H.

Nótese que todavía no hemos hecho ninguna observación acerca del mecanismo biológico que causa esta acumulación preferente en el entorno de otras Alus (modelo 4). Por lo tanto, las siguientes secciones las dedicaremos a tratar de identificar ese mecanismo.

4.4 La clusterización de las dianas

La manera más directa en que se podría generar la clusterización de las Alus, sería una distribución correspondiente de las dianas de inserción. Es decir, una distribución espacial de las dianas que se asemeje a la de las Alus. Para poner a prueba si existe una correlación entre la distribución espacial de Alus y las dianas, hemos calculado el nivel de clusterización de dianas (para la detección de las dianas véase 3.3.9 Detección de dianas de inserción). Sin embargo, determinar el nivel de clusterización de éstas es más complicado que en el caso de las Alus. Eso se debe principalmente al hecho de que los elementos transponibles a menudo insertan consigo mismos nuevas dianas en el genoma (véase Tabla 4-13).

En consecuencia, cabe la posibilidad de que la inserción de nuevas dianas altere la distribución original. Por tanto, hemos de considerar varios escenarios distintos. En primer lugar, determinaremos el nivel de clusterización correspondiente a la distribución de dianas, tal y como se observa hoy día en el genoma. Esto quiere decir que tomaremos en cuenta todas las dianas independientemente de si están dentro o fuera de algún elemento transponible. Por otro lado, haremos tres aproximaciones a distintos estados “ancestrales”, es decir a distintas distribuciones de dianas antes de la inserción de TEs. Para ello distinguiremos entre la distribución de dianas sin los TEs y la situación antes de la radiación de las Alus. Para llegar a esta estimación, se sustraen los TEs, y con ellos las dianas, del genoma en función de la edad evolutiva. Así pues, para reconstruir la situación antes de la radiación de las Alus, se sustraen todos los TEs con una edad evolutiva menor de 0.3. Esta distancia evolutiva marca

aproximadamente el comienzo de la radiación de las Alus. Por otro lado, para estimar la distribución sin los TEs se sustrae cualquier elemento transponible del genoma independientemente de su edad. Nótese que de esta manera se sustraen tanto las dianas que se hallan dentro de un TE como todos los nucleótidos del elemento. En consecuencia, la distancia entre las dianas más cercanas en 5' y 3' se acorta (véase Figura 4-15). Por último, hay que tener en cuenta también que la inserción de una Alu o LINE1 'ocupa' una diana. Así pues, en la tercera aproximación a un estado o una distribución "ancestral" tomamos en cuenta este hecho y sustituimos todas las secuencias Alu y LINE1 por una diana de inserción.

Tabla 4-13: Número y densidad de dianas, tanto en el genoma completo como en las regiones ocupadas por Alus y LINE1. En primer lugar, se confirma que las dianas son más frecuentes en las isocoras ricas en A+T. Por otra parte, se observa que la densidad de dianas dentro de las Alus no depende de la isocora, mientras que la de los LINE1 muestra una correlación negativa con el G+C de la isocora. Esto probablemente quiere decir que las dianas están expuestas a mayor presión mutacional en las isocoras H (es decir que decaen más rápidamente con el tiempo)

	Dianas totales		**Dianas dentro de Alus**		**Dianas dentro de LINE1**	
Isocora	#	10 kb	#	10 kb	#	10 kb
L1	2353385	27.983	24957	5.243	410728	22.123
L2	1910336	22.668	48068	5.620	318098	20.454
H1	1072632	17.250	60775	5.577	155018	21.872
H2	201734	11.230	21678	5.589	26167	21.574
H3	55971	7.817	9552	5.588	5850	19.748
H4	13324	5.039	1909	5.349	1551	17.115
Total	5607382	21.707	166939	5.534	917412	21.432

Las dianas se buscan en ambas hebras y en consecuencia existen dianas solapantes. Por ejemplo, la secuencia 5'-TTTTAAAA-3' ocurre con cierta frecuencia en el genoma. Este patrón contiene la diana 5'-TT/AAAA-3' tanto en 5'→3' como en 3'→5'. Sin embargo, el corte se efectúa siempre en la misma posición en el genoma, independientemente de la orientación de la diana que se utilice para la inserción. Estos patrones se han aglutinado para no distorsionar los resultados.

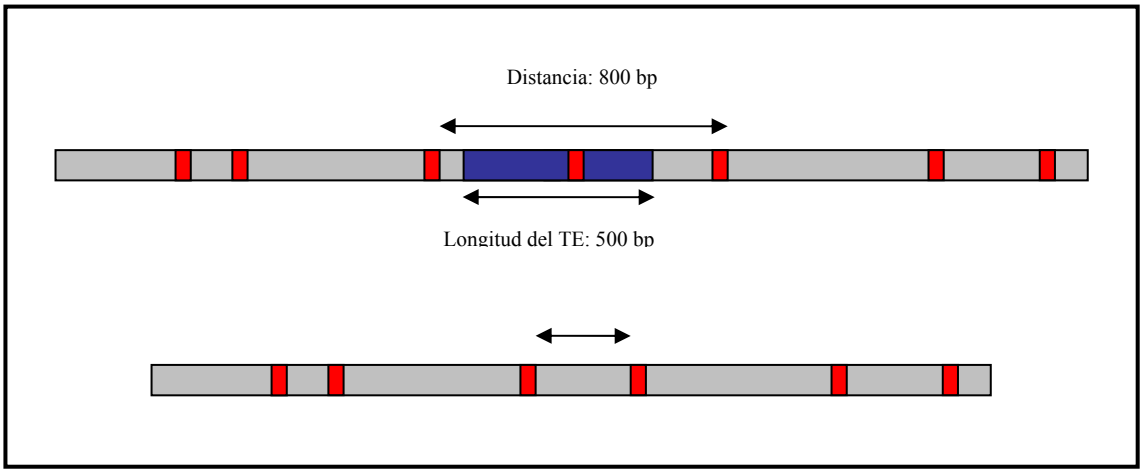

Figura 4-15: Reajuste de las distancias después de la sustracción de un elemento transponible

La Figura 4-16 muestra la comparación entre el nivel de clusterización de Alus y el de dianas de inserción, en función de la isocora. En primer lugar, se observa que no hay diferencias sustanciales entre las cuatro formas de calcular la clusterización de dianas. Sin embargo, se aprecian diferencias importantes entre el nivel de clusterización de Alus y dianas. Ambos tienen aproximadamente la misma magnitud en las isocoras L, mientras que en regiones ricas en G+C las Alus muestran un nivel más alto de clusterización que las dianas. Así pues, en las isocoras L la distribución espacial de las Alus se asemeja a la de las dianas, mientras que en regiones ricas en G+C las dos distribuciones son claramente diferentes. Este resultado sugiere que en isocoras H la distribución espacial de las Alus no está solo determinada por las dianas de inserción, sino que deben estar implicados además otros mecanismos.

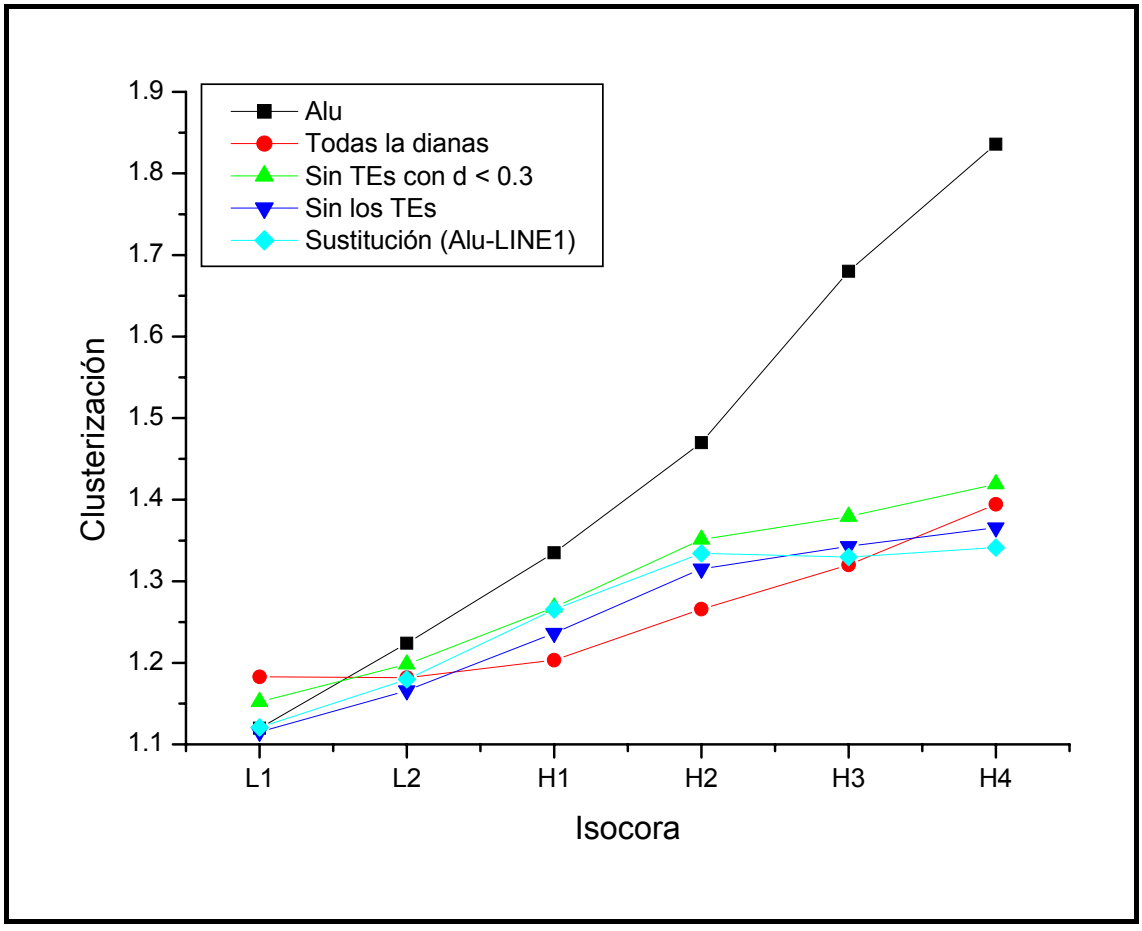

Figura 4-16: Comparación entre la clusterización de dianas y de Alus en función de la isocora. Se observa que las Alus, sobre todo en regiones ricas en G+C, están más clusterizadas que las dianas.

4.5 La base biológica de la acumulación preferente (modelo 4)

En la sección 4.3.5 (Exceso de de clusterización) hemos demostrado que la acumulación preferente en el entorno de otras Alus (modelo 4) permite *describir* la clusterización diferencial en las distintas isocoras. Nótese, sin embargo, que al implementar este modelo no hicimos alusión alguna a los mecanismos biológicos concretos que lo puedan justificar, como en el caso de los modelos 1-3. El objetivo de esta sección es pues tratar de identificar la base biológica de este modelo, añadiéndole así valor *explicativo*. Hay al menos dos posibilidades:

1. Las Alus podrían mostrar alguna preferencia para insertarse en el entorno cercano de otras Alus. Concretamente, se ha propuesto que podrían insertarse preferentemente en la cola de poli-A de otra Alu preexistente

(Stenger *et al*., 2001; El Sawy y Deininger, 2005). Para poner a prueba esta hipótesis, compararemos el número observado de Alus en los flancos de cada Alu con el número de dianas disponibles, analizando luego su variación por isocora.

2. Los clusters de Alus podrían generarse también a través de IPEs, es decir mediante inserciones dentro de Alus previamente insertadas en el genoma. De hecho en el modelo 2 de inserción de Alus ya admitíamos esa posibilidad. Si por alguna razón existiera una mayor propensión a la formación de IPEs en las isocoras H, se podría explicar tambien la acumulación preferente que se observa. Para comprobar esta posibilidad, calcularemos primero las frecuencias esperadas de IPEs y analizaremos luego las proporciones obs/esp de en las diferentes isocoras.

4.5.1 Inserción preferente en el flanco 3' de otras Alus

Si las Alus se insertasen preferentemente en el entorno Alu, eso podría contribuir a la fuerte clusterización de Alus que se observa en las isocoras H, la cual acabamos de ver que no se puede explicar mediante la clusterización de dianas (véase la sección anterior y la Figura 4-16). Nótese que, con el término "inserción preferente", aquí nos referimos a una situación en la que se producirían más inserciones de las esperadas en el entorno Alu. Se sospecha que las Alus podrían tener cierta preferencia para insertarse en la cola de poli-A de otra Alu preexistente (Stenger *et al*., 2002; El Sawy y Deininger, 2005). Para poner a prueba esta hipótesis compararemos el número observado de Alus en los flancos de cada Alu con el número de dianas. La proporción entre las dianas en ambos flancos (#5'/#3') nos da la probabilidad de encontrar Alus en los flancos.

El punto clave de este análisis está en poder determinar con precisión el orden de inserción en un "tandem" de Alus. Es decir, hay que distinguir entre la Alu que estaba antes y la que se insertó después en su flanco. Sin esta distinción no sería posible determinar si una Alu se ha insertado en el flanco 5' o 3' (véase Figura 4-17). Para ello utilizamos la edad evolutiva de cada una de las Alus del tandem. Sin embargo, la edad evolutiva muestra cierta dispersión, y por tanto aunque un elemento tenga una distancia evolutiva ligeramente mayor que otro, podría haberse insertado antes. Por esto, solamente consideramos "tandems" de Alus en los que su edad evolutiva difiera al menos en 0.03 unidades.

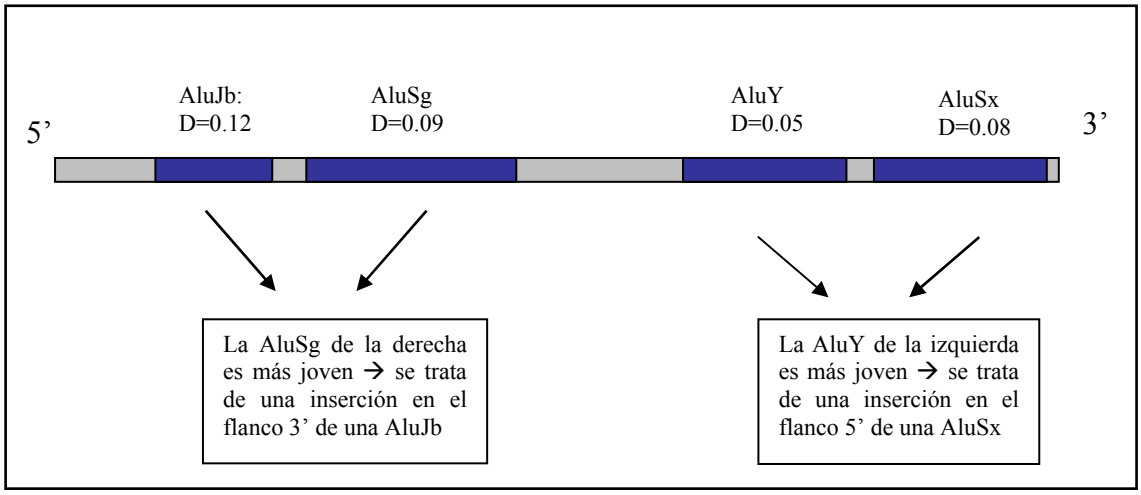

Figura 4-17: Esquema de la determinación del elemento preexistente en tandems de Alus.

En laTabla 4-14 se muestran las frecuencias de dianas y Alus en ambos flancos y las proporciones resultantes. Se confirma que las dianas de inserción son más frecuentes en el flanco 5'. En concreto, observamos aproximadamente 3 veces más dianas en el flanco 5' que en el 3', y por lo tanto se esperarían las proporciones correspondientes de Alus. Sin embargo, se observa todo lo contrario. Las Alus se acumulan más en el flanco 3' pese a que hay menos dianas. Además se observa una correlación positiva con el G+C de la isocora (R=0.93; p=0.0081), que no existe en el caso de las dianas (R=0.78; p=0.067).

Tabla 4-14: Comparación entre el número de dianas y Alus en los flancos de otras Alus

	Dianas de inserción			**Inserciones de Alu**		
Isocora	**# en 5'**	**# en 3'**	**# 5'/# 3'**	**# en 5'**	**# en 3'**	**# 5'/# 3'**
L1	11393	32602	0.349	1018	964	1.056
L2	19545	55874	0.350	3579	3025	1.183
H1	23150	62655	0.369	8073	6278	1.286
H2	7082	19382	0.365	4046	3033	1.334
H3	2788	7509	0.371	2480	1703	1.456
H4	584	1594	0.366	509	364	1.398
Total	64542	179616	0.359	15367	19705	1.282

Este resultado, confirma que existe inserción preferencial de Alus en las colas de poli-A de otras Alus (El Sawy y Deininger, 2005). La inserción preferencial en

este caso probablemente se debe a una mayor accesibilidad de estas regiones (nucleosome phasing, Cost *et al*., 2001). No obstante, el punto importante para nosotros aquí es la observación de que esta preferencia es más fuerte en isocoras H que en L (dada la correlación positiva de las proporciones con el G+C de la isocora). Esto sugiere que la inserción preferente en el flanco 3' de otras Alus contribuye también a la formación de la mayor clusterización de Alus que se observa en las isocoras H.

4.5.2 Sesgo en las proporciones de IPEs

Otro mecanismo que puede generar la acumulación de Alus en el entorno Alu son las inserciones dentro de Alus preexistentes (IPEs). Si la probabilidad de formación de estas IPEs variase con la isocora, tal vez podríamos explicar la acumulación preferente en el entorno Alu. Así pues, en este apartado exploraremos la posibilidad de que haya una mayor probabilidad de IPEs en las isocoras H, lo que explicaría la mayor acumulación preferente que se observa en estas isocoras. La frecuencia observada de IPEs en el genoma se puede medir fácilmente (véase el apartado 3.3.8 Detección de IPEs en el genoma), pero es más problemático estimar el número esperado de IPEs. A continuación, expondremos una aproximación a este problema.

El número esperado de IPEs depende, por un lado, de la densidad del elemento hospedador. Por ejemplo, las Alus ocupan casi el 20% de las isocoras H3 pero sólo el 5% de las L1. En consecuencia, debido al mayor número de elementos hospedadores, se esperan más inserciones Alu/Alu en las isocoras H3. Sin embargo, no es correcto utilizar la densidad que se observa hoy día, ya que el número de muchos elementos hospedadores, sobre todo de aquellos cuya actividad de transposición solapa o solapaba con la de las Alus, no ha sido constante durante la radiación de las Alus. Por lo tanto, en primer lugar hay que corregir tanto para la densidad diferencial del elemento hospedador como para la evolución de la misma.

Para tomar en cuenta la densidad diferencial de los elementos hospedadores en función de la isocora, se calcula la densidad efectiva que tuvo el elemento hospedador cuando se produjo la inserción. La densidad efectiva se define como:

$$D_{eff} = \frac{\sum_{N} (L_N - 20)}{L_{Tot}} \qquad (4\text{-}7)$$

donde N es el número de elementos hospedadores que se hallaban ya en el genoma cuando se produjo la inserción y L_{Tot} es la longitud total de la región que se considera (en este caso, las isocoras). Así pues, la densidad efectiva representa

la proporción de ocupación del elemento hospedador en la región en el momento en que se produjo la inserción. Nótese que, debido al límite de resolución de RepeatMasker, restamos 20 bp de la longitud del elemento hospedador (véase también 3.3.9 Detección de dianas de inserción). La densidad efectiva se puede interpretar como una probabilidad simple de inserción. Por ejemplo, si la densidad del elemento hospedador en un momento dado es de 0.1, es decir un 10% de ocupación, la probabilidad de que una inserción se produzca en un elemento hospedador (preexistente) es igualmente de 0.1. Es decir, si insertamos 100 Alus en una región en la que el 10% del ADN corresponde al elemento hospedador, entonces esperaríamos que un promedio de 10 Alus se insertasen dentro de un elemento hospedador. Así pues, en primer lugar, hemos calculado la probabilidad simple de inserción para cada IPE (densidad efectiva) en función de la isocora (véase Tabla 4-15).

Finalmente, obtenemos el número esperado de IPEs multiplicando la probabilidad de inserción (su media en la isocora) por el número total de elementos en la isocora. Es decir, si analizamos IPEs del tipo Alu/Alu, el número de elementos es el número total de Alus en un tipo de isocora. En la Tabla 4-15 se muestran los resultados.

Cabe destacar dos características importantes. Por un lado, la proporción obs/esp es claramente menor de 1 en todas las isocoras, lo que significa que se detectan menos inserciones Alu/Alu de las esperadas. En segundo lugar, se observa una correlación positiva entre la proporción obs/esp y el G+C de la isocora ($r = 0.92779$; $p < 0.00763$). El valor máximo de esta proporción se encuentra en las isocoras H4.

Tabla 4-15: Análisis de inserciones Alu/Alu: número de Alus insertadas, probabilidad media de inserción, número total de Alus en la isocora, número esperado de IPEs y proporción de observados a esperados (tanto los calculados como los obtenidos por simulación)

Isocora	**# observados**	**Probabilidad de inserción**	**# Alus**	**# esperados**	**obs/esp**	**obs/esp (simulación)**
L1	800	0.0309	153570	4750.7	0.1684	0.2020
L2	2875	0.0545	267295	14571.2	0.1973	0.2407
H1	6623	0.0920	335879	30886.4	0.2144	0.2581
H2	3312	0.1125	116275	13082.2	0.2532	0.3163
H3	2147	0.1236	50392	6227.9	0.3447	0.3941
H4	349	0.0649	10607	688.9	0.5066	0.5688
Total	16106	---	934018	84155.1	---	---

Esta correlación positiva se observa también en los demás tipos de IPEs (véase Tabla 4-16). Para LINE1, TE_{DNA} y LTR como elementos hospedadores, el máximo de la proporción observados/esperados se encuentra también en las isocoras H4 (véase Figura 4-18).

Este resultado confirma que hay una mayor frecuencia de IPEs en las isocoras H, con lo que este mecanismo podría estar contribuyendo tambien a la mayor clusterización observada en estas isocoras. Sin embargo, esta conclusión podría ser prematura: en estos cálculos hemos corregido para la densidad del elemento hospedador y su evolución, pero no hemos tenido en cuenta las dianas de inserción. No obstante, la distribución de las dianas de inserción podría influir decisivamente en la proporción de IPEs, ya que se puede argumentar que no es la densidad del elemento hospedador la que condiciona la probabilidad de inserción, sino el número de dianas. Por lo tanto sería prematuro sacar una conclusión definitiva aún. La incorporación de las dianas a los cálculos es más complicada que la corrección para las densidades debido a múltiples razones. Los problemas y posibles soluciones se exponen en el capitulo 6 Problemas abiertos y perspectivas.

Tabla 4-16: IPEs con LINE1, TE_{DNA} y LTR como elemento hospedador, respectivamente. El elemento insertado siempre es una Alu. Se observa la misma tendencia de la proporción observados/esperados que en el análisis de inserciones Alu/Alu. La región en la que se observan más IPEs con respecto a las esperadas es siempre la isocora H4.

	Hospedador: LINE1		**Hospedador: TE_{DNA}**		**Hospedador: LTR**	
Isocora	**# Alu**	**obs/esp**	**# Alu**	**obs/esp**	**#Alu**	**obs/esp**
L1	20591	0.7197	3904	0.9655	6219	0.4851
L2	31280	0.7816	7608	1.0415	9867	0.5003
H1	30547	0.9401	9292	1.0543	14259	0.5661
H2	8503	1.1445	2923	1.2133	4312	0.6870
H3	2539	1.2205	1030	1.3881	1305	0.7696
H4	726	1.9637	190	1.8996	353	1.1739

Otro problema pendiente sería determinar la razón de la mayor propensión a formar IPEs en las isocoras H. Una posibilidad para esto es que la abundancia de elementos funcionales (exones, promotores, etc.) en las isocoras H determine la eliminación a través de selección purificadora de la mayoría de las inserciones de Alus producidas en estos sitios. Sin embargo, las Alus preexistentes podrían ser 'sitios seguros' para nuevas inserciones, ya que la probabilidad de eliminación

sería menor. Por lo tanto, la acumulación de Alus en estos sitios seguros llevaría a un mayor incremento de la clusterización en los entornos Alu de las isocoras más ricas en G+C.

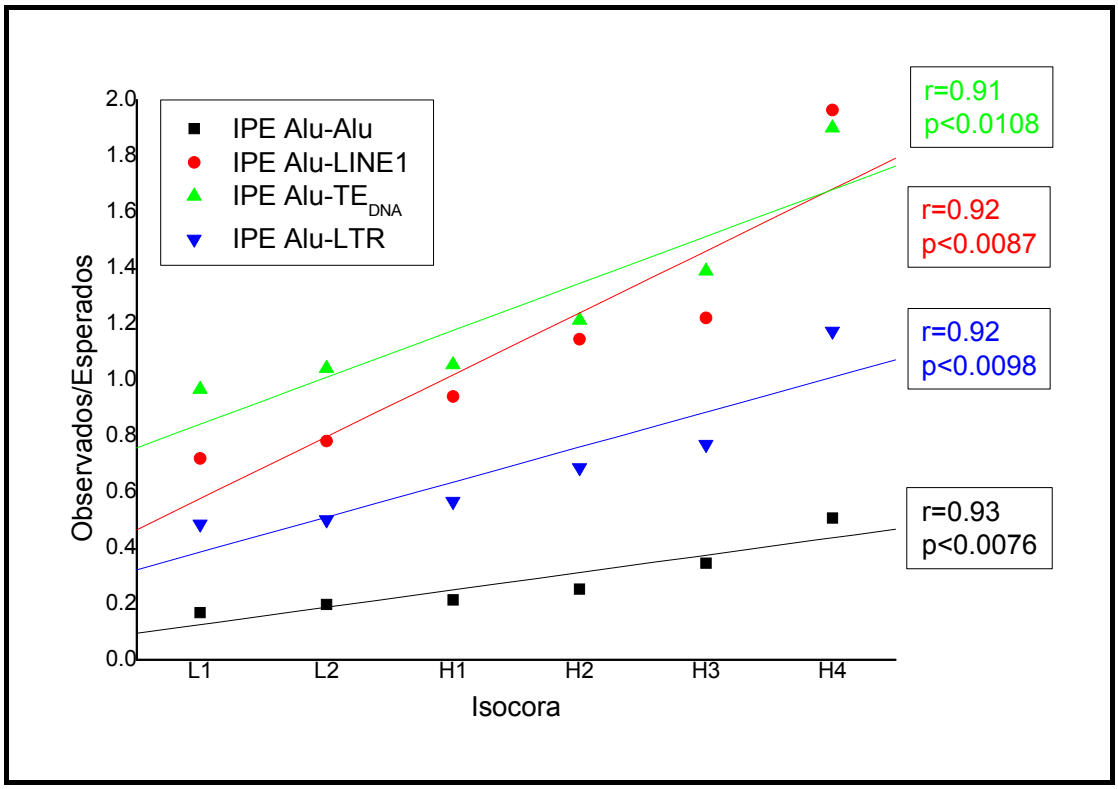

Figura 4-18: Proporción de IPEs (obs/esp). Se observa una correlación positiva con el G+C de la isocora independientemente de cuales sean los elementos hospedadores.

Capítulo 5

Conclusiones

El análisis llevado a cabo sobre los cambios de densidad de los retrotransposones Alu en diferentes regiones del genoma humano permite obtener las siguientes conclusiones:

1. La comparación de las densidades absolutas de Alus y LINE1 en diferentes isocoras y en función de su edad evolutiva, descarta la intervención de la competencia por la retrotransposasa en el cambio de densidad de las Alus.

2. Las diferencias en G+C entre las Alus y sus correspondientes secuencias consenso en función de la distancia evolutiva, indican que el ajuste composicional no comienza hasta que se alcanzan distancias evolutivas de 0.08-0.1. Puesto que el cambio de densidad ocurre mucho antes, alrededor de distancias de 0.025, esto demuestra que el ajuste composicional no interviene en el cambio de densidad de las Alus.

3. La comparación de las densidades de Alus en intrones y en regiones intergénicas en el genoma completo, confirma un exceso de Alus en los intrones. Sin embargo, se observa que este exceso se limita a las isocoras L, lo que descarta cualquier influencia del entorno génico en el cambio de densidad de las Alus.

4. La definición *in silico* del trímero resultante de un proceso de recombinación homóloga desigual (RHD) entre Alus, y la simulación del proceso de inserción, han permitido rastrear el genoma en busca de las huellas dejadas por la recombinación y cuantificar su impacto en el cambio de densidad. Se observa una proporción más alta de trímeros en las isocoras L, lo que indica una mayor actividad RHD, o una mayor supervivencia de sus productos en estas isocoras. Esto sugiere que la RHD podría contribuir al cambio de densidad de las Alus.

5. La edad evolutiva de las subfamilias de inserción reciente aporta datos adicionales en favor de la recombinación. Las Alus jóvenes muestran distancias evolutivas menores en isocoras L que en H, lo que señala a la recombinación como agente causante del cambio de densidad, ya que es el único mecanismo cuya actividad depende de la edad de los elementos.

6. El rastreo del genoma humano en busca de 'Alus solitarias' (aquellas que muestran una distancia a la Alu más próxima superior a 2 kb), ha revelado que este tipo de Alus son más frecuentes en las isocoras L. Este resultado apunta de nuevo a la RHD, ya que ésta actúa eliminando sobre todo a las Alus próximas, contribuyendo así al cambio de densidad.

7. Se propone una nueva medida para cuantificar la aglomeración (o clusterización) en la distribución espacial de las Alus, basada en las distancias normalizadas entre los elementos. La independencia de esta medida con respecto a la densidad permite comparar distintas regiones genómicas sin que el resultado se vea distorsionado por el sesgo en la densidad.

8. Para asociar un nivel de significación estadística (valor-p) a las frecuencias de clusterización observadas, se han diseñado distintos modelos para simular la inserción de Alus en el genoma, pudiendo así distinguir entre clusterización efectiva y clusterización aleatoria. Además, la incorporación en estos modelos de diferentes mecanismos evolutivos, ha permitido evaluar el impacto de cada uno de ellos en la generación de la clusterización.

9. Se observa que la mayor clusterización de Alus se produce en las isocoras H4. Esto confirma que la densidad y la clusterización son dos medidas diferentes de la distribución espacial de Alus, puesto que la densidad muestra su máximo en las isocoras H3. Por lo tanto, y al contrario de lo

propuesto por otros autores, los mecanismos implicados en el cambio de densidad y en la clusterización diferencial de las Alus parecen ser distintos.

10. La clusterización observada en la mayoría de las isocoras L se puede explicar en base al azar y a las inserciones en elementos preexistentes (IPEs). Por el contrario, el mayor grado de clusterización observado en las isocoras H no se puede explicar por azar, por las IPEs ni por la exclusión de la inserción en exones. Solo cuando se contempla la acumulación preferente en el entorno Alu de manera diferenciada para cada isocora, es posible explicar los niveles de clusterización observados en las isocoras más ricas en G+C.

11. La clusterización de las dianas de inserción muestra cualitativamente la misma correlación positiva con el G+C de la isocora que las Alus. Sin embargo, en las isocoras H la clusterización de Alus es mucho más fuerte que la de dianas. Así pues, la clusterización de dianas no puede explicar la fuerte acumulación preferencial en el entorno Alu que se observa en las isocoras H.

12. Una comparación entre las frecuencias de dianas de inserción y de Alus en los flancos de otras Alus preexistentes muestra que existen aproximadamente 3 veces más inserciones de Alus en la cola de poli-A de lo esperado. Además, esta preferencia de inserción es más fuerte en isocoras H que en L (correlación positiva: $R=0.93$; $p=0.0081$). De este modo, la inserción preferente en la cola de poli-A de otras Alus preexistentes contribuye significativamente a la acumulación observada de Alus en el entorno Alu.

13. Tomando en cuenta la densidad efectiva de los diferentes elementos hospedadores en el momento en que se produjo la inserción, se pueden calcular las frecuencias esperadas de las inserciones dentro de elementos preexistentes (IPEs). Hemos encontrado una correlación positiva entre la proporción obs/esp y el G+C de la isocora ($r = 0.92779$; $p < 0.00763$), lo que sugiere la participación de este mecanismo en la generación de la mayor clusterización observada en las isocoras H.

14. Por último, el estudio presentado aquí no hubiera sido posible sin una serie de mejoras metodológicas que nos han permitido seguir la dinámica evolutiva de las Alus en el genoma completo: 1) el desarrollo de una serie de programas en Perl ha permitido procesar los datos primarios, integrando así la voluminosa información generada por RepeatMasker o IsoFinder

cuando se aplican a genoma completo; 2) la corrección para sustituciones múltiples, aplicada aquí por primera vez a los alineamientos de Alus, ha permitido estimar con mayor precisión la edad evolutiva de los elementos; 3) la técnica de desfragmentación de Alus desarrollada en este trabajo ha permitido eliminar la habitual sobreestimación de estos elementos y aproximar con mayor precisión su número real en el genoma; y 4) la definición precisa, mediante técnicas *in silico*, de los trímeros de Alus o de las Alus solitarias, ha permitido rastrear el genoma en busca de las huellas dejadas por la recombinación.

Capítulo 6

Problemas abiertos y perspectivas

Como último punto de esta memoria, queremos dar aquí una perspectiva breve del trabajo. Por un lado, apuntaremos algunas posibles soluciones al problema de la estimación fiable del número esperado de IPEs. En segundo lugar, daremos algunas perspectivas de futuro acerca de la aplicación de los métodos presentados aquí a otros problemas.

Dinámica de las dianas de inserción

Como hemos mencionado en el apartado 4.5.2 (Sesgo en las proporciones de IPEs), para poder estimar el número esperado de IPEs hay que tomar en cuenta la dinámica de las dianas de inserción. Algunos problemas son:

- La probabilidad de que se produzca, por ejemplo, una inserción Alu/Alu viene dada en una primera aproximación por la proporción de dianas dentro y fuera de las Alus. Muchos TEs insertan consigo mismos nuevas dianas, lo que hace que se esté alterando continuamente la proporción entre dianas dentro y fuera de las Alus.

- Algunos elementos, como las Alus, no contienen ninguna diana de inserción si lo que se contempla es su secuencia consenso; sin embargo,

éstas se detectan a menudo en la secuencia genómica de los elementos. Así pues, se pueden generar nuevas dianas de inserción a través de mutación, lo que dificulta aún más la reconstrucción de su dinámica evolutiva.

Por lo tanto, la distribución de dianas depende tanto de la dinámica de los elementos transponibles como de la presión mutacional. En el futuro, intentaremos tomar en cuenta estos dos factores mediante la simulación de un modelo de inserción y mutación de los elementos transponibles.

Aplicación de la medida de clusterización a otros problemas

La medida presentada en 4.3.1 (Medida de la clusterización de Alus) se puede aplicar igualmente a cualquier elemento genómico, como genes, islas CpG, etc. Pensamos que la medida puede ser especialmente útil para abordar la correlación entre la distribución espacial de genes y sus niveles de expresión (Hurst *et al*., 2004; Altuvia *et al*., 2005; Ma *et al*., 2005).

Referencias

Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat. 22(3):229-44.

Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H. (2005) Clustering and conservation patterns of human microRNAs. Nucleic Acids Res. 33(8):2697-706.

Arcot SS, Adamson AW, Risch GW, LaFleur J, Robichaux MB, Lamerdin JE, Carrano AV, Batzer MA. (1998) High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. J Mol Biol. 281(5):843-56.

Alvarez-Valin F, Lamolle G, Bernardi G. (2003) Isochores, GC3 and mutation biases in the human genome. Gene 300(1-2):161-8.

Babcock M, Pavlíček A, Spiteri E, Kashork CD, Isohikhes I, Shaffer LG, Jurka J, Morrow BE. (2003) Shuffling of Genes Within Low-Copy Repeats on 22q11 (LCR22) by Alu-Mediated Recombination Events During Evolution. Genome Res 13:2519-2532.

Bailey JA, Liu G, Eichler EE. (2003) An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 73(4):823-34.

Baker MD, Read LR, Beatty BG, NG P. (1996) Requirements for Ectopic Homologous Recombination in Mammalian Somatic Cells. Mol Cell Biol. 16:7122-7132.

Batzer MA, Kilroy GE, Richard PE, Shaikh TH, Desselle TD, Hoppens CL, Deininger PL. (1990) Structure and variability of recently inserted Alu family members. Nucleic Acids Res. 18(23):6793-8.

Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E. (1996) Standardized nomenclature for Alu repeats. J Mol Evol. 42(1):3-6.

Batzer MA, Deininger PL. (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3:1-10.

Belle EMS, Duret L, Galtier N, Eyre-Walker A. (2004) The decline of isochores in mammals: An assessment of the GC content variation along the mammalian phylogeny. J Mol Evol. 58(6):653-660.

Belle EM, Webster MT, Eyre-Walker A. (2005) Why are young and old repetitive elements distributed differently in the human genome? J Mol Evol. 60(3):290-6.

Bernaola-Galván P, Román-Roldán R, Oliver JL. (1996) Compositional segmentation and long-range fractal correlation in DNA sequences. Phys Rev E 53:5181-5189.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953-958.

Bernardi G. (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3-17.

Bernardi G. (2001) Misunderstandings about isochores. Part 1. Gene 276:3-13.

Boeke JD. (1997) LINEs and Alus — the polyA connection. Nat Genet. 16:6–7.

Bovia F, Strub K. (1996) The signal recognition particle and related small cytoplasmic ribonucleoprotein particles. J Cell Sci. 109:2601-8.

Bovia F, Wolff N, Ryser S, Strub K. (1997) The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. Nucleic Acids Res. 25(2):318-26.

Britten RJ. (1996) DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A. 93(18):9374-7.

Britten RJ. (1997) Mobile elements inserted in the distant past have taken on important functions. Gene 205:177-182.

Britten RJ. (2004) Coding sequences of functioning human genes derived entirely from mobile element sequences. Proc Natl Acad Sci U S A. 101(48):16825-30.

Brookfield JF. (2001) Selection on Alu sequences? Curr Biol. 11:900–901.

Brosius J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115-134.

Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA. (2005) Alu retrotransposition-mediated deletion. J Mol Biol. 348(4):791-800.

Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA. (2001) Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol. 311(1):17-40.

Chen C, Gentles AJ, Jurka J, Karlin S. (2002) Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. Proc Natl Acad Sci USA. 99:2930-2935.

Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW. (1998) Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. Mol Cell Biol. 18:58-68.

Corneo G, Ginelli E, Soave C, Bernardi G. (1968) Isolation and characterization of mouse and guinea pig satellite DNAs. Biochemistry 7:4373-4379.

Cost GJ, Boeke JD. (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 37:18081-18093.

Cost GJ, Golding A, Schlissel MS, Boeke JD. (2001) Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res. 29(2):573-7.

Deininger PL, Daniels GR. (1986) The recent evolution of mammalian repetitive DNA elements. Trends Genet. 2:76-80.

Deininger PL, Batzer MA, Hutchison CA, Edgell MH. (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet. 8:307-311.

Deininger PL, Batzer MA. (1993) Evolution of retroposons. Evolutionary Biology 27:157-196. Plenum Press. New York.

Deininger PL, Batzer MA. (1999) Alu Repeats and Human Disease. Mol Genet Metab. 67:183-193.

Deininger PL, Batzer MA (2002) Mammalian Retroelements. Genome Research 12: 1455-1465.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 13:651–658.

Dewannieux M, Esnault C, Heidmann T. (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 35:41-8.

Doolittle WF, Sapienza C. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601-603.

El-Sawy M, Deininger P. (2005) Tandem insertions of Alu elements. Cytogenet Geneome Res. 108:58-62.

Elliott B, Richardson C, Jasin M. (2005) Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. Mol Cell. 17(6):885-94.

Esnault C, Casella JF, Heidmann JF. (2001) Tetrahymena thermophila ribozyme-based indicator gene to detect transposition of marked retroelements in mammalian cells. Nucleic Acids Res. 30(11):e49.

Eyre-Walker A. (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochors and junk DNA. Genetics 152:675-683.

Eyre-Walker A, Hurst LD. (2001) The evolution of isochores. Nature Reviews 2:549-555.

Fan H, Goodier JL, Chamberlain JR, Engelke DR, Maraia RJ. (1998) 5' processing of tRNA precursors can be modulated by the human La antigen phosphoprotein. Mol Cell Biol. 18(6):3201-11.

Feng G, Moran JV, Kazazian HH, Boeke JD. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Nat Genet. 22:130.

Filipski J, Salinas J, Rodier F. (1989) Chromosome localization-dependent compositional bias of point mutations in Alu repetitive sequences. J Mol Biol. 206:563-6.

Fukagawa T, Sugaya K, Matsumoto K, Okumura K, Ando A, Inoko H, Ikemura T. (1995) A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. Genomics 25:184–191.

Fullerton SM, Bernardo Carvalho A, Clark AG. (2001) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 18(6):1139-42.

Gibbons R, Dugaiczyk LJ, Girke T, Duistermars B, Zielinski R, Dugaiczyk A. (2004) Distinguishing humans from great apes with AluYb8 repeats. J Mol Biol. 339(4):721-9.

Gilbert N, Lutz-Prigge S, Moran JV. (2002) Genomic deletions created upon LINE-1 retrotransposition. Cell 110(3):315-25.

Gu Z, Wang H, Nekrutenko A, Li WL. (2000) Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. Gene 259:81-88.

Hackenberg M, Barnaola-Galván P, Carpena P, Oliver L. (2005) The biased distribution of Alus in human isochores might be driven by recombination. J Mol Evol. 60:365-377.

Hamidi S, Salo T, Kainulainen T, Epstein J, Lerner K, Larjava H. (2000) Expression of alpha(v)beta6 integrin in oral leukoplakia. Br J Cancer. 82(8):1433-40.

Hasegawa M, Kishino H, Yano T. (1989) Estimation of branching dates among primates by molecular clocks of nucleic DNA which slowed down in Hominoidea. J Human Evol. 18: 461-476.

Hedges DJ, Callinan PA, Cordaux R, Xing J, Barnes E, Batzer MA. (2004) Differential alu mobilization and polymorphism among the human and chimpanzee lineages. Genome Res. 14(6):1068-75.

Hellmann-Blumberg U, Hintz MF, Gatewood JM, Schmid CW. (1993) Developmental differences in methylation of human Alu repeats. Mol Cell Biol. 13(8):4523-4530.

Hollies CR, Monckton DG, Jeffreys AJ. (2001) Attempts to detect retrotransposition and de novo deletion of Alus and other dispersed repeats at specific loci in the human genome. Eur J Hum Genet. 9(2):143-6.

Houck CM, Rinehart FP, Schmid CW. (1979) A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol. 132:289-306.

Hurst LD, Pal C, Lercher MJ. (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 5(4):299-310.

IHGSC (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Jabbari K, Bernardi G. (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 224:123-128.

Jurka J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872-1877.

Jurka J. (2000) Repbase Update, a database and an electronic journal of repetitive elements. Trends Genet. 16:418-419.

Jurka J, Krnjajic M, Kapitonov VV, Stenger JE, Kokhanyy O. (2002) Active Alu elements are passed primarily through paternal germlines. Theor Popul Biol. 61(4):519-30.

Jurka J, Kohany O, Pavlíček A, Kapitonov VV, Jurka MV. (2004) Duplication, coclustering, and selection of human Alu retrotransposons. Proc Nat Acad Sci USA 101:1268-1272.

Kapitonov V, Jurka J. (1995) The Age of Alu Subfamilies. J Mol Evol. 42:59-65.

Kazazian HH Jr, Goodier JL. (2002) LINE drive. retrotransposition and genome instability. Cell. 110(3):277-80.

Kazazian HH Jr. (2004) Mobile elements: drivers of genome evolution. Science 303(5664):1626-32.

Kolomietz E, Meyn MS, Pandita A, Squire JA. (2002) The role of Alu Repeat Clusters as Mediators of Recurrent Chromosomal Aberrations in Tumors. Genes Chromosomes Cancer 35:97-112.

Kolosha VO, Martin SL. (2003) High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). J Biol Chem. 278:8112-8117.

Korenberg JR, Rykowski MC. (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell 53:391-400.

Labuda D, Striker G. (1989) Sequence conservation in Alu evolution. Nucleic Acids Res. 17:2477-2491.

Lagemaat van de LN, Landry JR, Mager DL, Medstrand P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 19(10):530-6.

Lambert S, Saintigny Y, Delacote F, Amiot F, Chaput B, Lecomte M, Huck S, Bertrand P, Lopez BS. (1999) Analysis of intrachromosomal homologous recombination in mammalian cell, using tandem repeat sequences. Mutat Res. 433:159-168.

Landry JR, Medstrand P, Mager DL. (2001) Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. Genomics 76(1-3):110-6.

Lesk A. (2005) Introduction to Bioinformatics. Oxford University Press.

Lev-Maor G, Sorek R, Shomron N, Ast G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. Science 300(5623):1288-91.

Li W. (2001) Delineating relative homogeneous G+C domains in DNA sequences. Gene 276:57-72.

Li WH, Gu Z, Wang H, Nekrutenko A. (2001) Evolutionary analyses of the human genome. Nature 409:847-849.

Lin JH, Levin JL. (1997) Self-primed reverse transcription is a mechanism shared by several LTR-containing retrotransposons. RNA 3:952-953.

Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA. (2000) Related Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J. 19:3822-30

Lorenc A, Makalowski W. (2003) Transposable elements and vertebrate protein diversity. Genetica 118:183-191.

Ma PC, Chan KC, Chiu DK. (2005) Clustering and re-clustering for pattern discovery in gene expression data. J Bioinform Comput Biol. 3(2):281-301.

Makalowski W, Mitchell GA, Labuda D. (1994) Alu sequences in the coding regions of mRNA: A source of protein variability. Trends Genet. 10:188-193.

Malik HS, Burke WD, Eickbush TH. (1999) The Age and Evolution of Non-LTR Retrotransposable Elements. Mol Biol Evol. 16(6):793–805.

Martin SL, Bushman FD. (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol. 21:467-475.

Mathias SL, Scott AF, Kazazian HH, Boeke Jr. JD, Gabriel A. (1991) Reverse transcriptase encoded by a human transposable element. Science 254:1808-1810.

McClintock B. (1946) Year Book Carnegie Inst. Washington 45, 176–186.

McClintock B. (1947) Year Book Carnegie Inst. Washington 46, 146–152.

McClintock B. (1948) Year Book Carnegie Inst. Washington 47, 155–169.

McClintock B. (1949) Year Book Carnegie Inst. Washington 48, 142–154.

McClintock B. (1951) Year Book Carnegie Inst. Washington 50, 174–181.

McClintock B. (1954) Year Book Carnegie Inst. Washington 53, 254–260.

Medstrand P, Landry JR, Mager DL. (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem. 276(3):1896-903.

Medstrand P, van de Lagemaat LN, Mager DL. (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 12:1483-95.

Metha ML. (1991). Random Matrices. Academic Press, Boston.

Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G. (1982) Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc Natl Acad Sci USA 79:355–359.

Mighell AJ, Markham AF, Robinson PA. (1997) Alu sequences. FEBS Letters 417:1-5.

Mitchell M, Dai L, Savidge G, Alhaq A. (2004) An Alu-mediated 31.5-kb deletion as the cause of factor XI deficiency in 2 unrelated patients. Blood. 104(8):2394-6.

Moran JV, DeBerardinis RJ, Kazazian HH Jr. (1999) Exon shuffling by L1 retrotransposition. Science 283:1530-1534.

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 31(2):159-65.

Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G. (1991) The distribution of genes in the human genome. Gene 100:181-7.

Nekrutenko A, Li W-H. (2001) Transposable elements are found in a large number of human protein-coding genes. Trends in Genetics 17(11):619-621.

Nigumann P, Redik K, Matlik K, Speek M. (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79(5):628-34.

Ohno S. (1972) So much “junk” in out genomes. Evolution of Genetic Systems (Smith, H.H. ed.) 366-370.

Ohshima K, Okada N. (1994) Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. J Mol Biol. 243:25–37.

Ohshima K, Hamada M, Terai Y, Okada N. (1996) The 3′ ends of tRNA-derived short interspersed repetitive elements are derived from the 3′ ends of long interspersed repetitive elements. Mol Cell Biol. 16:3756–3764.

Ohshima K, Hattori M, Yada Y, Gojobori T, Sakaki Y, Okada N. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biology 4:R74.

Okada N, Hamada M. (1997) The 3′ ends of tRNA-derived SINEs originated from the 3′ ends of LINEs: a new example from the bovine genome. J Mol Evol. 44:S52–S56.

Okada N, Ohshima K. (1993) A model for the mechanism of initial generation of short interspersed elements (SINEs). J Mol Evol. 37:167–170.

Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R. (2001) Isochore chromosome maps of eukaryotic genomes. Gene 276:47-56.

Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, Hackenberg M, Bernaola-Galván P (2002) Isochore chromosome maps of the human genome. Gene 300:117-127.

Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. (2004) IsoFinder: computational prediction of isochores in genome sequences. Nucl Acids Res. 32 (Web Server Issue): W287–W292.

Orgel LE, Crick FHC. (1980) Selfish DNA: the ultimate parasite. Nature 284:604-607.

Otieno AC, Carter AB, Hedges DJ, Walker JA, Ray DA, Garber RK, Anders BA, Stoilova N, Laborde ME, Fowlkes JD, Huang CH, Perodeau B, Batzer MA. (2004) Analysis of the human Alu Ya-lineage. J Mol Biol. 342(1):109-18.

Paces J, Zika R, Paces V, Pavlicek A, Clay O, Bernardi G. (2004) Representing GC variation along eukaryotic chromosomes. Gene 333:135-41.

Pavlíček A, Jabbari K, Paces J, Paces V, Henjar J, Bernardi G. (2001) Similar integration but different stability of Alus and LINES in the human genome. Gene 276:39-45.

Prak ET, Kazazian HH Jr. (2000) Mobile elements and the human genome. Nat Rev Genet. Nov;1(2):134-44.

Pruitt KD, Katz KS, Sicotte H, Maglott DR. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 16(1):44-7.

Pruitt KD, Maglott DR. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucl Acids Res. 29(1):137-40.

Pruitt KD, Tatusova T, Maglott DR. (2003) NCBI Reference Sequence project: update and current status. Nucl Acids Res. 31(1):34-7.

Quentin Y. (1992) Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. Nucl Acids Res. 20:487-493.

Rice P, Longden I, Bleasby A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16(6):276—277.

Ripley BD. (1981) Spatial statistics. John Wiley. Chichester, UK.

Rogers J. (1983) Retroposons defined. Nature 301:460.

Román-Roldán R, Bernaola-Galván P, Oliver JL. (1998) Sequence compositional complexity of DNA through an entropic segmentation method. Physical Review Letters 80(6): 1344-1347.

Rossetti LC, Goodeve A, Larripa IB, De Brasi CD. (2004) Homologous recombination between AluSx-sequences as a cause of hemophilia. Hum Mutat. 24(5):440.

Roy-Engel AM, Salem AH, Oyeniran OO, Deininger L, Hedges DJ, Kilroy GE, Batzer MA, Deininger PL (2002) Active Alu Element “A-Tails”: Size Does Matter. Genome Research 12:1333-1344.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature. 19;423(6942):873-6.

Rudiger NC, Gregersen N, Brandt-Kielland MC. (1995) One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. Nucleic Acids Res. 23:256-260.

Rynditch A, Zoubak S, Tsyba L, Tryapitsina-Guley N, Bernardi G. (1998) The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222:1-16.

Sarrowa J, Chang DY, Maraia RJ. (1997) The decline in human Alu retroposition was accompanied by an asymmetric decrease in SRP9/14 binding to dimeric Alu RNA and increased expression of small cytoplasmic Alu RNA. Mol Cell Biol. 17(3):1144-51.

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr. (1997) Many human L1 elements are capable of retrotransposition. Nature Genet. 16:37-43.

Schmid CW. (1998) Does SINE evolution preclude Alu function? Nucl Acids Res. 26:4541-4550.

Shedlock AM, Okada N. (2000) SINE insertions: powerful tools for molecular systematics. Bioessays 22:148–160.

Shen MR, Batzer MA, Deininger PL. (1991) Evolution of the master Alu gene(s). J Mol Evol. 33:311-320.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 423(6942):825-37.

Slagel V, Flemington E, Traina-Dorge V, Bradshaw H, Deininger P. (1987) Clustering and subfamily relationships of the Alu family in the human genome. Mol Biol Evol. 4:19–29.

Smit AFA, Tóth G, Riggs AD, Jurka J. (1995) Ancestral, Mammalian-wide Subfamilies of LINE-1 Repetitive Sequences. J Mol Biol. 246:401–417.

Smit AFA. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657-663

Smith NGC, Eyre-Walker A. (2001) Synonymous codon bias is not caused by mutation bias in G+C rich genes in humans. Mol Biol Evol. 18:982-986.

Sorek R, Ast G, Graur D. (2002) Alu-containing exons are alternatively spliced. Genome Res. 12:1060-1067.

Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D, Ast G. (2004) Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. Mol Cell. 14(2):221-31.

Soriano P, Meunier-Rotival M, Bernardi G. (1983) The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. Proc Natl Acad Sci USA 80:1816–1820.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka ED, Wilkinson M, Birney E. (2002) The Bioperl Toolkit: Perl modules for the life sciences. Genome Research. 12(10):1611-8.

Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. (2001) Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res. 11:12-27.

Stephens R, Horton R, Humphray S, Rowen L, Trowsdale J, Beck S. (1999) Gene Organisation, Sequence Variation and Isochore Structure at the Centromeric Boundary of the Human MHC. J Mol Biol. 291:789-799.

Stoppa-Lyonnet D, Carter PE, Meo T, Tosi M. (1990) Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. PNAS 87(4):1551-1555.

Su A, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. PNAS 101(16):6062–6067.

Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. (2002) Human L1 retrotransposition is associated with genetic instability in vivo. Cell 110(3):327-38.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. (2002) Molecular archeology of L1 insertions in the human genome. Genome Biology 3(10):1-18.

Tamura K, Nei M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10:512-526.

The MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. Nature 401(6756):921-3.

Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T. (1997) Precise switching of DNA replication timing in the GC content transition area in the human MHC. Mol Cell Biol. 17:4043-4050.

Ullu E, Tschudi C. (1984) Alu sequences are processed 7SL RNA genes. Nature 312:171-172.

Waldman AS, Liskay RM. (1988) Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol Cell Biol. 8:5350-7.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33(Database issue):D39-45.

Willis IM. (1993) RNA polymerase III. Genes, factors and transcriptional specificity. Eur J Biochem. 212(1):1-11.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol. 21(4):1429-39.

Wolfe K, Sharp P, Li W-H. (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283-285.

Xing J, Hedges DJ, Han K, Wang H, Cordaux R, Batzer MA. (2004) Alu element mutation spectra: molecular clocks and the effect of DNA methylation. J Mol Biol. 344(3):675-82.

Zerial M, Salinas J, Filipski J, Bernardi G. (1986) Gene Distribution and nucleotide sequence organization in the human genome. Eur. J. Biochem. 160:479-485.

Zhou Y, Mishra B. (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci U S A. 102(11):4051-6.

Zoubak S, Clay O, Bernardi G. (1996) The gene distribution of the human genome. Gene 174: 95–102.

Apéndices

A. Listado de programas en Perl

A continuación se da una descripción breve de los scripts en Perl que se han escrito para procesar los datos primarios (IsoFinder y RepeatMasker) y llevar a cabo los distintos análisis.

Procesamiento de la salida de IsoFinder

Script: make_classification.pl	
Descripción:	
Implementa un algoritmo iterativo para establecer la clasificación de las isocoras según las abundancias de Zoubak (véase 3.2 Clasificación de las isocoras).	
Entrada:	• Las abundancias de Zoubak • La salida de IsoFinder
Salida:	• Fichero con los rangos de G+C de las isocoras (la clasificación)

Script: convert_IsoF-out.pl	
Descripción:	
Añade una etiqueta de isocora a cada línea del fichero de salida del programa IsoFinder, según el contenido en G+C de la LHGR correspondiente.	
Entrada:	• Salida de IsoFinder • Salida del script make_classification.pl (los rangos de G+C)
Salida:	• Como la salida de IsoFinder & etiqueta de isocora

Detección de los elementos transponibles

Script: Extract_Iso.pl	
Descripción:	
Mediante el programa "extractseq" del paquete EMBOSS se extrae la secuencia genómica de cada isocora	
Entrada:	• Tabla con las coordenadas de isocoras
Salida:	• Fichero en formato FASTA de la secuencia genómica de cada isocora

Script: Run_RM.pl	
Descripción:	
Para cada isocora se lanza el programa RepeatMasker	
Entrada:	• Tabla con las coordenadas de isocoras • Directorio con los ficheros FASTA
Salida:	• Salida del programa RepeatMasker

Procesamiento de la salida de RepeatMasker

Script: GetTE-values.pl	
Descripción:	
El script cruza la salida de RepeatMasker con la tabla de isocoras y calcula las características de los TEs como se describe en la sección 3.3 Preparación de los datos	
Entrada:	• Tabla de isocoras • Los dos ficheros de salida de RepeatMasker
Salida:	• Tabla con todos los valores de los TEs (un fichero por isocora)

Script: Glue_TEs.pl	
Descripción:	
Aglutina los TEs fragmentados según los criterios expuestos en 3.3.4 Desfragmentación de los elementos	
Entrada:	• Tabla de isocoras • Directorio con los ficheros de salida del script GetTE-values.pl
Salida:	• Tabla con todos los valores de los TEs (un fichero por isocora)

Búsqueda in silico de productos de recombinación

Script: SearchRHD_prod.pl	
Descripción:	
Este script implementa la búsqueda de productos de recombinación según los criterios expuestos en 3.3.7 La definición in silico de los productos de recombinación	
Entrada:	• Tabla generada con el script Glue_TEs.pl (véase más arriba)
Salida:	• Tabla que mantiene el formato del fichero de entrada & añade un número de identificación para cada trímero

Cálculo del exceso en intrones

Script: Alu_IntronEx.pl	
Descripción:	
El programa lee sucesivamente para cada isocora los datos de las Alus y genes, calculando entonces la densidad de Alus en intrones e intergenes (para la definición de intergenes véase 3.3.6 Cálculo del exceso de Alus en intrones)	
Entrada:	• Tabla de Alus (una por isocora) • Tabla de genes • Tabla de isocoras
Salida:	• Tabla con las densidades en intrones e intergenes (una línea por isocora)

Cálculo de la clusterización

Script: Calc_Clustering.pl	
Descripción:	
El programa lee sucesivamente para cada isocora los datos de los elementos y los genes, calculando entonces la clusterización observada. Después lleva a cabo la simulación del proceso de inserción bajo un modelo dado (elegido por el usuario)	
Parámetros (obligatorios):	• Modelo de inserción • Elementos a usar (subfamilias, grupos, etc.)
Parámetros (opcionales):	• La especie (por defecto: *Homo sapiens*) • El rango de cromosomas (por defecto: todos) • Número de simulaciones (por defecto: 500) • Número de procesadores (por defecto: 2)
Entrada:	• Tabla de elementos (una por isocora) • Tabla de genes (por cromosoma – el script asigna los genes a la isocora) • Tabla de isocoras • Fichero de configuración de los modelos de inserción
Salida:	• Tabla con los niveles de clusterización y sus valores-P asociados (una línea por isocora)

Detección de las dianas

Script: TE_InsPat.pl	
Descripción:	
El programa lee sucesivamente para cada isocora los datos de los elementos y las dianas de inserción. Después se cruzan las coordenadas y se detecta la localización de la diana (dentro de un TE, en el flanco, o fuera). Si la diana reside dentro de un TE o en su flanco se determina adicionalmente la orientación de la diana respecto al elemento.	
Parámetros (opcionales):	• Definición del flanco (por defecto: últimos 10 bp del TE y los 20 bp que siguen) • Límite de distancia evolutiva (por defecto: 0.3)
Entrada:	• Tabla de elementos (una por isocora) • Tabla de isocoras • Tabla de las dianas de inserción
Salida:	• Tabla con la localización y orientación de las dianas

B. Programas

A continuación se listan los programas más importantes utilizados en este trabajo.

IsoFinder (Oliver *et al.*, 2004):

El programa IsoFinder realiza la segmentación composicional de una secuencia de ADN generando regiones genómicas largas de composición homogénea. El programa está disponible a través de una interfaz web: http://bioinfo2.ugr.es/IsoF/isofinder.html.

EMBOSS (The European Molecular Biology Open Software Suite; Rice *et al.*, 2000):

EMBOSS es un conjunto de programas para el análisis de secuencias. EMBOSS se distribuye libremente bajo las licencias GPL y LGPL y se puede descargar en http://emboss.sourceforge.net/. En este trabajo se han empleado sobre todo los programas "extractseq" (para extraer un subsecuencia de ADN) y "fuzznuc" (para detectar patrones de ADN).

BioPerl (http://www.bioperl.org/; Stajich *et al.*, 2002)

BioPerl es una colección amplia de módulos en Perl que facilitan tanto el manejo y análisis de secuencias como el uso de aplicaciones externas.

RepeatMasker (Smit y Green, no publicado)

El programa RepeatMasker permite la detección de elementos repetitivos en un gran número de especies. El programa y las librerías necesarias se pueden solicitar en la siguiente página Web: http://repeatmasker.genome.washington.edu.

C. Las subfamilias de Alus

La Tabla C-1 muestra una estadística básica de las subfamilias de Alus más frecuentes en el genoma humano.

Tabla C-1: Las subfamilias de Alus más frecuentes en el genoma humano.

Familia	#	Longitud media (bp)	Edad evolutiva	Tiempo de inserción (MYA)
AluSx	291680	297.5	0.0755	37
AluY	121827	291.5	0.0389	19
AluJo	115467	276.3	0.1378	81
AluJb	108754	271.2	0.1293	81
AluSq	81605	293.2	0.0705	44
AluSg	71565	294.3	0.0601	31
AluSp	44532	291.2	0.0514	37
AluSc	43533	283.4	0.0561	35
AluSg1	5277	288.9	0.0559	---
AluYb9	2808	302.4	0.0173	3-4
AluYa8	2603	297.9	0.0172	3-4
AluYd8	873	276.9	0.0434	8
AluYg6	582	260.5	0.0289	---

D. Datos suplementarios sobre la clusterización

En las tablas de esta sección se facilita más información acerca de las simulaciones de los modelos de inserción.

Tabla D-1: Número de isocoras en el análisis, número de isocoras con clusterización significativa, exceso de clusterización y nivel de clusterización en las isocoras significativas para los modelos 1, 2 y 3.

	Modelo 1: Inserción de puntos (p < 0.05)			
Isocora	**# isocoras**	**# isocoras significativas**	**Exceso de clusterización**	**NC en isocoras significativas**
L1	1065	387	0.363	1.256
L2	1959	1118	0.571	1.357
H1	2084	1478	0.709	1.450
H2	737	606	0.822	1.558
H3	305	270	0.885	1.757
H4	79	74	0.937	1.882
	Modelo 2: Inserción en elementos preexistentes (p < 0.05)			
Isocora	**# isocoras**	**# isocoras significativas**	**Exceso de clusterización**	**NC en isocoras significativas**
L1	1065	209	0.196	1.316
L2	1959	622	0.318	1.447
H1	2084	728	0.349	1.598
H2	737	355	0.482	1.722
H3	305	220	0.721	1.848
H4	79	69	0.873	1.922
	Modelo 3: Exclusión en exones (p < 0.05)			
Isocora	**# isocoras**	**# isocoras significativas**	**Exceso de clusterización**	**NC en isocoras significativas**
L1	1065	212	0.199	1.313
L2	1959	592	0.302	1.452
H1	2084	665	0.319	1.605
H2	737	330	0.448	1.738
H3	305	205	0.672	1.865
H4	79	64	0.810	1.959

Tabla D-6-2: Modelo 4: Número de isocoras en el análisis, número de isocoras con clusterización significativa y exceso de clusterización tanto para las probabilidades de inserción fijas como para las probabilidades variables obtenidas a partir de la densidad de genes.

	Modelo 4: Inserción preferente en el entorno Alu (p < 0.05)				
		Probabilidad (fija)		Probabilidad (variable)	
Isocora	# isocoras	# isocoras significativas	Exceso de clusterización	# isocoras significativas	Exceso de clusterización
L1	56	9	0.161	12	0.214
L2	196	43	0.219	67	0.342
H1	260	32	0.123	74	0.285
H2	126	15	0.119	42	0.333
H3	84	17	0.202	44	0.524
H4	19	1	0.053	9	0.474

E. Publicaciones relacionadas con la memoria

A continuación se relacionan los trabajos publicados, señalando a que parte de la memoria corresponden.

Oliver *et al.* (2002)

Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, **Hackenberg M**, Bernaola-Galván P. (2002) Isochore chromosome maps of the human genome. GENE 300:117-127 http://bioinfo2.ugr.es/publi/Oliver_etal_2002.pdf

Este trabajo describe la estructura de isocoras del genoma humano generada con el algoritmo IsoFinder. Por otro lado, contiene también un análisis descriptivo de la distribución de genes y elementos transponibles en función de la isocora. A raíz de este primer análisis descriptivo se comenzó la investigación que se resume en esta memoria acerca de los mecanismos biológicos que generan la distribución espacial de los retrotransposones Alu.

ELSEVIER

Gene 300 (2002) 117–127

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Isochore chromosome maps of the human genome

José L. Oliver[a,*], Pedro Carpena[b], Ramón Román-Roldán[c], Trinidad Mata-Balaguer[a], Andrés Mejías-Romero[a], Michael Hackenberg[a], Pedro Bernaola-Galván[b]

[a]*Departamento de Genética, Instituto de Biotecnología, Universidad de Granada, Granada, Spain*
[b]*Departamento de Física Aplicada II, Universidad de Málaga, Málaga, Spain*
[c]*Departamento de Física Aplicada, Universidad de Granada, Málaga, Spain*

Received 21 December 2001; received in revised form 19 August 2002; accepted 18 September 2002

Abstract

The human genome is a mosaic of isochores, which are long DNA segments ($\gg$ 300 kbp) relatively homogeneous in G + C. Human isochores were first identified by density-gradient ultracentrifugation of bulk DNA, and differ in important features, e.g. genes are found predominantly in the GC-richest isochores. Here, we use a reliable segmentation method to partition the longest contigs in the human genome draft sequence into long homogeneous genome regions (LHGRs), thereby revealing the isochore structure of the human genome. The advantages of the isochore maps presented here are: (1) sequence heterogeneities at different scales are shown in the same plot; (2) pair-wise compositional differences between adjacent regions are all statistically significant; (3) isochore boundaries are accurately defined to single base pair resolution; and (4) both gradual and abrupt isochore boundaries are simultaneously revealed. Taking advantage of the wide sample of genome sequence analyzed, we investigate the correspondence between LHGRs and true human isochores revealed through DNA centrifugation. LHGRs show many of the typical isochore features, mainly size distribution, G + C range, and proportions of the isochore classes. The relative density of genes, Alu and long interspersed nuclear element repeats and the different types of single nucleotide polymorphisms on LHGRs also coincide with expectations in true isochores. Potential applications of isochore maps range from the improvement of gene-finding algorithms to the prediction of linkage disequilibrium levels in association studies between marker genes and complex traits. The coordinates for the LHGRs identified in all the contigs longer than 2 Mb in the human genome sequence are available at the online resource on isochore mapping: http://bioinfo2.ugr.es/isochores.

Keywords: Isochore maps; Compositional segmentation; Chromosome domains; Comparative genomics; Alus; Long interspersed nuclear elements; Single nucleotide polymorphisms

1. Introduction

The availability of the human genome draft sequence offers an unprecedented opportunity to bring sequence patterns into line with the chromosome structures revealed by modern molecular cytogenetics, such as chromosome domains or high-resolution chromosome bands. Isochores – long DNA segments ($\gg$ 300 kbp) fairly homogeneous in G + C, revealed by analytical ultracentrifugation of bulk DNA (Macaya et al., 1976; Bernardi et al., 1985; Bernardi, 1995, 2000) – may be the structures linking both organization levels. In fact, isochores have been successfully related to chromosome bands (Saccone et al., 1993).

One conventional way to visualize sequence heterogeneity is the moving-window approach. This simple technique consists of sliding a window of arbitrary length along the sequence, and then computing the GC content of each window. This procedure dates from the earliest times of sequence analysis when only short, and often homogeneous, sequences were available. However, with the discovery that eukaryotic genomes are multi-scale complex systems made up of fairly homogeneous isochores of different composition (Macaya et al., 1976; Bernardi et al., 1985; Bernardi, 2000) and with the subsequent finding of long-range correlations in eukaryotic DNA sequences (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992; Bernaola-Galván et al., 2002a), this

Abbreviations: LHGR, long homogeneous genome region; bp, base pair; kbp, kilobase pair; G + C, guanine plus cytosine content; SNP, single nucleotide polymorphism; MY, millions of years; SINE, short interspersed nuclear element; LINE, long interspersed nuclear element.

* Corresponding author. Departamento de Genetica, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain. Fax: +34-958-244073.
E-mail address: oliver@ugr.es (J.L. Oliver).

0141-933/02/$ - see front matter
PII: S0378-1119(02)01034-X

practice becomes untenable. Sliding a window of arbitrary length and step over long, heterogeneous and correlated sequences may lead to misleading results (see Li, 2001, for a recent review). However, GC-plots routinely accompany the publication of every new genome sequence, the long-range patterns being identified only by eye. This happens, for example, with the 'isochores' tentatively identified on human chromosomes 21 (Hattori et al., 2000) and 22 (Dunham et al., 1999).

Two other more recent techniques (Nekrutenko and Li, 2000; Häring and Kypr, 2001), also based on moving windows, use the random (uncorrelated) model to test sequence homogeneity. The pitfalls in such an approach have already been noted (Bernardi, 2001; see also Li et al., 2002). A key problem is that moving windows do not enable the accurate location of isochore boundaries before carrying out the homogeneity test. Therefore, it is not surprising that one of these techniques (Nekrutenko and Li, 2000) failed to detect the only isochore boundary experimentally characterized to date (Fukagawa et al., 1995, 1996; Stephens et al., 1999), while the other (Häring and Kypr, 2001) was unable to detect any isochores in the human chromosomes 21 and 22.

An alternative tool to analyze genome heterogeneity is compositional segmentation (Bernaola-Galván et al., 1996, 2001; Li et al., 1998; Román-Roldán et al., 1998; Oliver et al., 1999, 2001; Li, 2001). Domains of all sizes can be simultaneously detected by this method, and isochore boundaries can be accurately determined to single base pair resolution.

A recently derived hierarchical segmentation method (Oliver et al., 2001; Román-Roldán et al., 2002) is used here to divide the longest contig of each human chromosome into non-overlapping, relatively homogeneous genome regions, called long homogeneous genome regions (LHGRs). To investigate to what extent these regions may correspond to the true isochores identified by the Bernardi group through DNA centrifugation, we analyze here several LHGR features, such as size distribution, G + C range, and proportions of the different compositional classes in a wide sample of human genome sequence. We also analyzed the relative densities of genes, Alu and long interspersed nuclear element (LINE) repeats and the different types of single nucleotide polymorphisms (SNPs) in these regions.

2. Materials and methods

Different freezes, from October 2001 to February 2002, of the public human genome draft sequence available at NCBI (Lander et al., 2001; ftp://ncbi.nlm.nih.gov/genomes/H_sapiens) were used to compile information for different parts of this work. All the contigs longer than 2 Mb in the human genome were segmented using our hierarchical algorithm (for a complete list see the online resource on isochore mapping: http://bioinfo2.ugr.es/isochores). The

Table 1
Longest human contigs by chromosome analyzed in this study (NCBI October 2001 freeze[a])

Chromosome	Accession	Contig version	Contig length (bp)
1	NT_004424	6	6,311,978
2	NT_005375	6	4,746,219
3	NT_005927	6	19,259,936
4	NT_006204	6	5,458,445
5	NT_006907	6	4,272,479
6	NT_007592	6	19,443,354
7	NT_007819	6	12,615,535
8	NT_008271	6	3,868,249
9	NT_008413	6	8,724,786
10	NT_008609	6	8,702,417
11	NT_009151	6	24,188,643
12	NT_009714	6	5,170,685
13	NT_024524	6	10,245,455
14	NT_025892	5	16,139,217
15	NT_010194	6	10,898,583
16	NT_010604	6	4,049,516
17	NT_010718	6	8,843,538
18	NT_010895	6	4,073,989
19	NT_026483	4	4,069,655
20	NT_011362	6	26,179,448
21	NT_011512	4	28,511,026
22	NT_011520	8	23,083,944
X	NT_011687	6	6,615,739
Y	NT_011875	7	9,946,786
			Total: 275,419,622 (8.6% of the genome)

A complete list of the contigs analyzed can be found at the online resource on isochore mapping: http://bioinfo2.ugr.es/isochores.

[a] ftp://ncbi.nlm.nih.gov/genomes/H_sapiens.

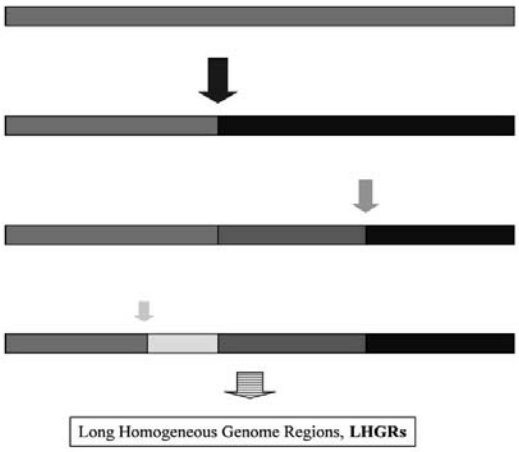

Fig. 1. Schematic representation of the segmentation algorithm used to locate LHGRs on sequence contigs. The successive cuts are given in a hierarchical way: at each scale, the cut maximizing the overall compositional complexity of the sequence is chosen, a procedure equivalent to maximizing the statistical significance of each cut (see Oliver et al., 2001, for details).

LHGRs identified on the longest contigs of each chromosome (Table 1) were used for most of the statistical comparisons described in this paper.

The segmentation algorithm used here (Oliver et al., 2001, 2002; Román-Roldán et al., 2002) is based on the original method developed by our group (Bernaola-Galván et al., 1996, 1999, 2000; Román-Roldán et al., 1998; Oliver et al., 1999; Grosse et al., 2002), but with several improvements (all aimed at addressing the specific isochore mapping problem). A schematic representation of the improved method is shown in Fig. 1. Two features should be emphasized:

(a) The cuts on the sequence are made one by one, in a hierarchical way, which may be more appropriate in searching for homogeneous segments in the long-range correlated, fractal landscape of eukaryotic DNA. In such a multi-scale landscape, the statistical significance of isochore boundaries may depend on the scale being considered. The hierarchical procedure guarantees the choice of the most significant cut at each scale.
(b) Short-scale sequence heterogeneity below 3 kbp is filtered out; this coarse graining of the sequence is a requirement imposed by the experimental characterization of isochores through DNA centrifugation (Bettecken et al., 1992; Bernardi, 2000). Filtering out heterogeneities below 3 kbp is also justified by the analysis of correlations in human chromosomes 20 and 21 (Bernaola-Galván et al., 2002a), which show a clear shift between two heterogeneity regimes just at this scale.

Therefore, our procedure tries to identify isochores by closely following Bernardi's early definition, i.e. 'fairly homogeneous regions', which implies accepting a certain level of internal heterogeneity. The 'strict isochores', unsuccessfully searched for by other authors (Lander et al., 2001; Häring and Kypr, 2001), simply cannot exist in natural DNA (Bernardi, 2001).

Parameter settings were as in the previous work (Oliver et al., 2001), i.e. a tract length of 3 kbp was used for coarse graining of GC content, and a 0.05 threshold was set for the P value in t-tests. These settings provide a high stability in detecting isochores in the human MHC region: the same isochore structure was obtained with coarse graining ranging from 2 to 30 kbp (see Fig. 2 in Oliver et al., 2001).

This segmentation method has been used to accurately predict the boundary between classes II and III of the human MHC region (Oliver et al., 2001), the only isochore boundary experimentally determined to date (Fukagawa et al., 1995; Stephens et al., 1999; The MHC Sequencing Consortium, 1999). The method has also been used to uncover isochore-like regions in other eukaryotic genomes (Oliver et al., 2001). More recently, we are also using this method to explore sequence heterogeneity in prokaryotic genomes (Bernaola-Galván et al., 2002b).

Gene ('CDS' line) and SNP ('variation' line) coordinates were taken from contig annotations. Chromosome contigs were scanned for Alu and LINE repeats using the program RepeatMasker (http://repeatmasker.genome.washington.edu), which identifies full-length and partial members of all the known repeat families represented in RepBase (Jurka, 2000; http://www.girinst.org/~server/repbase.html).

3. Results and discussion

3.1. Isochore chromosome maps

We applied our segmentation algorithm to the longest contig of each human chromosome available at the NCBI web server. As an example, the isochore chromosome maps for the longest contigs of chromosomes 21 and 22 are shown in Figs. 2 and 3. A more compact representation is used in Fig. 4 to show the isochore maps of the 24 longest contigs in the human chromosome complement. Isochore chromosome maps of every long human contig are regularly updated at the online resource on isochore mapping: http://bioinfo2.ugr.es/isochores. These maps graphically display the mosaic organization of the human genome (Bernardi et al., 1985; Bernardi, 2001; Pavlícek et al., 2001), composed by many regions of fairly homogeneous GC contents (see Li et al., 2002 for a recent reassessment of isochore homogeneity). The advantages of these maps over previous approaches based on moving windows are: (1) heterogeneities at very different scales are shown in the same plot; (2) pair-wise differences in GC content between adjacent regions are all statistically significant; (3) the

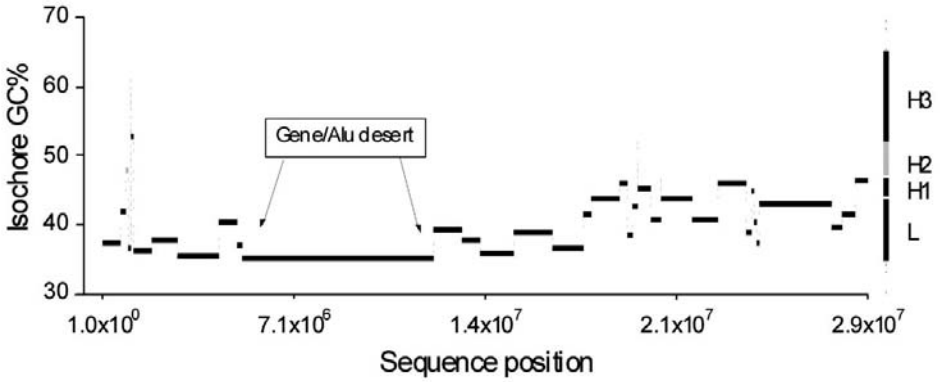

Fig. 2. Isochore chromosome map of the longest contig of human chromosome 21. The gene and Alu desert of 7.1 Mb is indicated by the arrows.

isochore boundaries are accurately defined to single base pair resolution; and (4) both gradual and abrupt isochore boundaries can be simultaneously revealed. As recently stressed by Bernardi (2001), this generalized mosaic structure along all the human chromosomes contradicts the suggestion (Eyre-Walker and Hurst, 2001) that the isochore structure accounts for 'only some parts' of the genome.

3.2. The relative amounts of DNA in the different compositional families

The LHGRs we found were classified into compositional families on the basis of their respective GC content, according to the GC values of Zoubak et al. (1996). The relative amounts of DNA in L, H1, H2 and H3 LHGRs in the longest contig of each human chromosome (Fig. 5) were fairly similar to the proportions experimentally found in the entire human genome by DNA centrifugation (e.g. Zoubak et al., 1996).

3.3. Statistical features of LHGRs

The size distribution of LHGRs, the distribution of GC contents and the GC differences between adjacent LHGRs are shown in Fig. 6. The LHGR size distribution was strongly skewed, with the highest value being 7.1 Mb (corresponding to the gene desert of chromosome 21) and an average size of 463 kbp. Many of the smaller LHGRs may correspond to GC-skewed repeats (Alus, LINEs), CpG islands or attached scaffold regions. LHGR GC content ranges from 30.8% to 64.3%, thus being consistent with the known GC range in isochores (Bernardi, 2001). The GC differences between adjacent isochores range from 1.5% to 24.5%. The smaller compositional shifts were noted between L-L adjacent LHGRs, while the greater ones occurred between L-H LHGRs. We observed, therefore, both gradual and abrupt GC shifts between adjacent isochores, depending on the neighbors considered. This observation contrasts with all previous reports in which only abrupt isochore boundaries were found (see, for example,

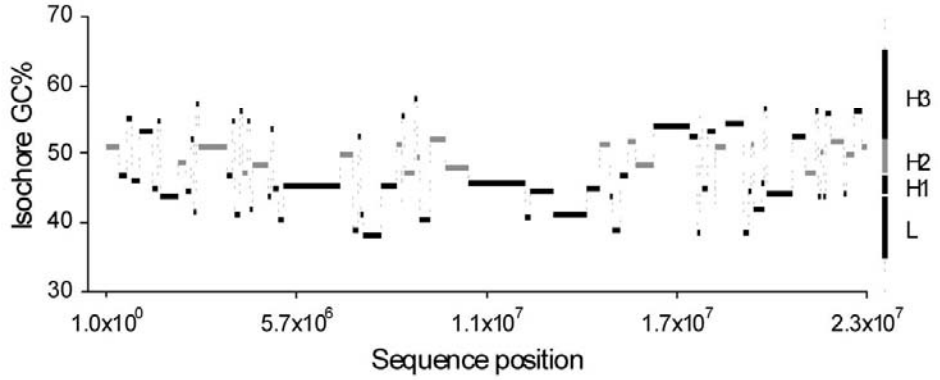

Fig. 3. Isochore chromosome map of the longest contig of human chromosome 22.

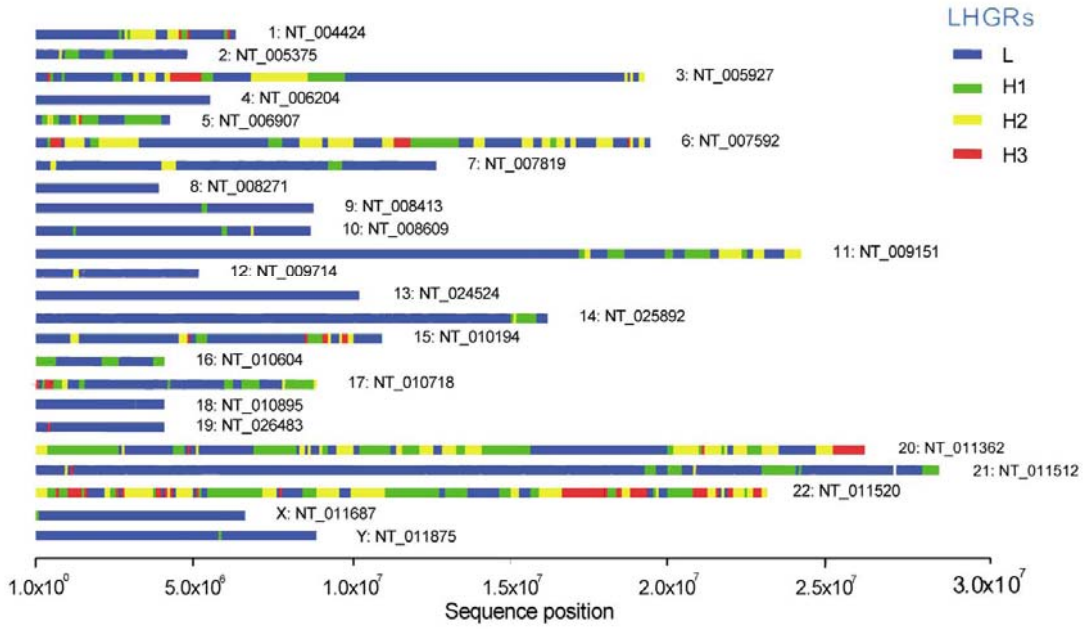

Fig. 4. Isochore chromosome maps of the longest contig in the human chromosome complement. The October 2001 freeze of NCBI contigs was used.

Fukagawa et al., 1995; Stephens et al., 1999). Note that the moving-window plot used by most authors only allows for the detection of abrupt transitions, while our segmentation method can reveal both gradual and abrupt isochore boundaries.

3.4. LHGR size variation with GC content

The different LHGR families show a strong variation in size, depending on the GC content, GC-poor LHGRs being significantly larger than GC-rich ones (Table 2). This

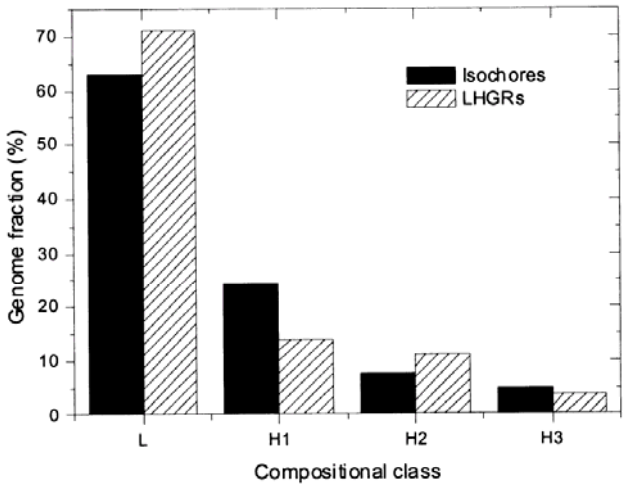

Fig. 5. The relative amounts of DNA in the different compositional LHGR families. The LHGRs in the longest contig of each chromosome (NCBI, October 2001 freeze), amounting to a total of 275.4 Mb (8.6% of the genome), were compared to the isochores detected by DNA centrifugation in the entire genome (Zoubak et al., 1996). LHGR G + C ranges (taken from Zoubak's paper) were: L1-L2 (GC% < 44), H1 (44 ≤ GC% < 47), H2 (47 ≤ GC% < 52) and H3 (GC% ≥ 52).

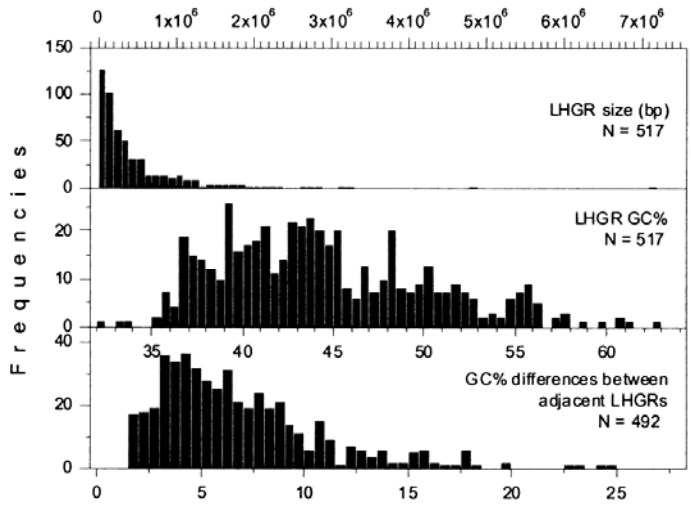

Fig. 6. Size distribution (above), GC content (middle) and GC differences between adjacent LHGRs in the longest contig of each human chromosome. A total of 517 LHGRs were considered. Contigs were taken from the NCBI October 2001 freeze, amounting to a total of 275 Mb (8.6% of the genome).

relationship was previously noted for the isochores detected by DNA centrifugation (Bettecken et al., 1992; Pilia et al., 1993; De Sario et al., 1996, 1997).

3.5. Variation of gene density in human LHGRs

In isochores detected by DNA centrifugation, Bernardi and coworkers (Bernardi et al., 1985; Mouchiroud et al., 1991; Zoubak et al., 1996; Bernardi, 2000) observed that gene density increases from a very low average in L isochores to a 20-fold higher level in H3 isochores. The recent release of the human genome draft sequence (Lander et al., 2001; Venter et al., 2001) propitiated a reexamination of this relation; while the first of the analyses, using 20 kbp windows along the assembled sequence, confirms the original observation, the second one, using 50 kbp windows, questioned the relative strength of the correlation. Thus, Venter et al. (2001) found that the correlation between GC content and gene density was not as skewed as observed by Bernardi's group, a higher proportion of genes being located in the GC-poor regions than had been previously observed in isochores. We therefore check this relation by using the human isochore boundaries accurately determined through our segmentation algorithm. Fig. 7 illustrates the close relationship we found between LHGR G + C and gene density (number of genes per kilobase). These results were remarkably similar to those of Bernardi's group (Mouchiroud et al., 1991; Zoubak et al., 1996; Bernardi, 2000, 2001), with our gene density values also falling on two straight lines crossing each other at about 46% GC. The less skewed distribution observed by Venter et al. (2001) may be due to (1) the specific values chosen for the window length and/or step, or (2) a wrong definition of the GC ranges assigned to each isochore family.

Table 2
Sizes of LHGRs (in kb) belonging to different families

LHGR	*N*	Mean	SE	Minimum	Maximum
L	276	615	49	9	7105
H1	84	399	49	16	2293
H2	97	281	29	3	1794
H3	60	144	28	6	1121

An analysis of the variance shows that size differences were statistically significant ($P < 10^{-5}$). The NCBI October 2001 freeze of contigs was used to compile this table.

3.6. Variation in the densities of Alu and LINE repeats

The density of Alu and LINE repeats is known to vary with isochore GC content (Soriano et al., 1983; Smit, 1999; Lander et al., 2001). To investigate if this relation is also true for LHGRs, we analyzed in detail the variations in Alu density along the LHGRs detected by our segmentation algorithm in 131 contigs longer than 3.5 Mb in the human genome (NCBI February 2002 freeze). We found a relationship between LHGR GC content and Alu density. However, the strength of such a relationship depends on the genetic age of the Alu family considered. Fig. 8 shows the average densities of two Alu families of different ages; the genetic ages of Alu families were taken from Kapitonov and Jurka (1996). While the density of the old Alu S family is strongly dependent on the isochore GC content, no

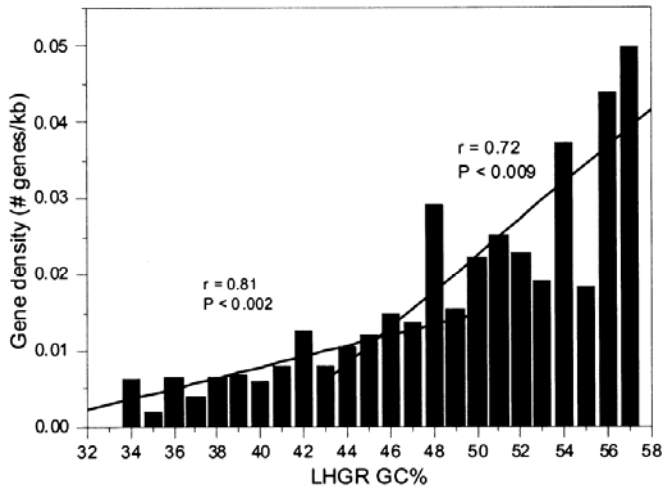

Fig. 7. Gene density vs. LHGR GC content. A total of 1096 genes located on 495 LHGRs from the longest contigs of each human chromosome were used for the comparison.

relationship was found for the youngest Alu Y family. LINE L1 density shows just the opposite pattern to that of the old Alus, being more frequent in L isochores and practically absent in the H3 isochores. Therefore, the density of Alu and LINE repeats in LHGRs follows the patterns previously found for isochores (Soriano et al., 1983; Smit, 1999; Lander et al., 2001).

3.7. Compositional correlations between gene GC content and LHGR G + C

Working with the isochores detected by DNA centrifugation, it has been convincingly shown that the GC content of genes matches the G + C of the isochores harboring them (Bernardi et al., 1985). Figs. 9 and 10 and Table 3 show that

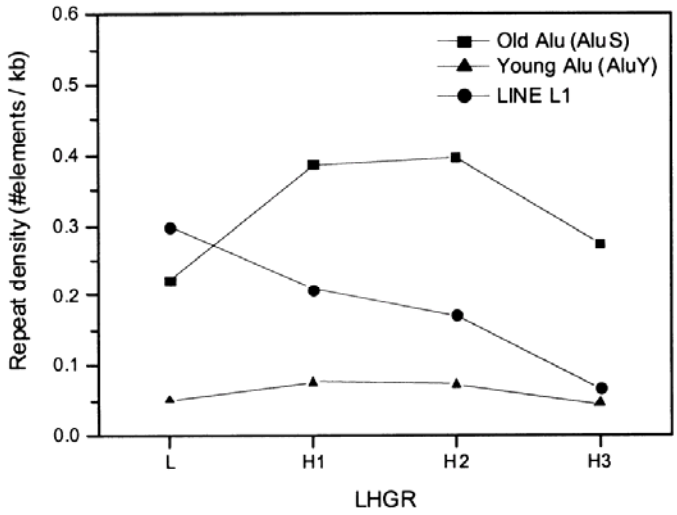

Fig. 8. Alu and LINE densities in the different LHGR families from 131 contigs longer than 3.5 Mb in the human genome (NCBI February 2002 freeze). Around 330,000 Alus and 345,000 LINEs in 2048 LHGRs were used to compile this figure. Old (S) and young (Y) Alus and the older LINE L1 elements were included in the comparison.

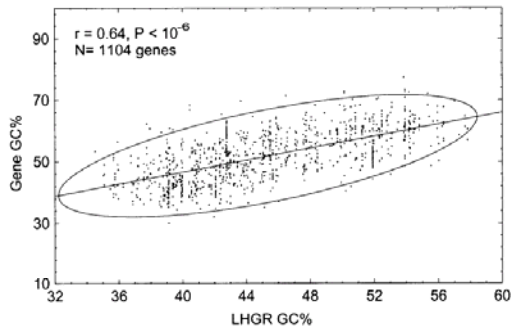

Fig. 9. Gene GC content vs. LHGR G + C ($r = 0.64$, $P < 10^{-6}$). A total of 1104 genes from the longest contig of each chromosome were included in the comparison. The ellipse shows 95% confidence intervals.

this rule also holds for the LHGRs detected by segmenting human contigs, thus being consistent with previous results (see, for example, Eyre-Walker and Hurst, 2001).

3.8. The compositional adjustment of Alus and LINES to LHGR GC content

The study of sequence repeats, such as Alus and LINEs, can also reveal compositional correlations. An advantage of using repeats instead of genes for this purpose is that the genetic ages of the different repeat families are known (Kapitonov and Jurka, 1996; Mighell et al., 1997), and therefore the evolution towards the compositional matching with the host sequence (a process known as 'compositional adjustment'; Oliver et al., 1990; Martínez Zapater et al., 1993) can be analyzed in more detail. Figs. 11–13 show the correlation plots between the GC content of repeats and the G + C of the LHGR harboring them for AluJo (81 MY old), AluSp (37 MY old) and AluYa5 (4 MY old) families, respectively. The strongest correlation, and therefore the best compositional adjustment, was for the oldest AluJo repeats, while no correlation was found for the youngest

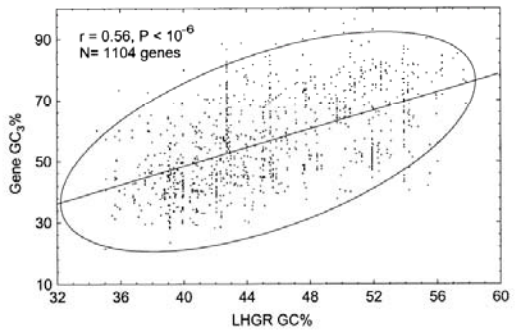

Fig. 10. Gene GC_3 content vs. LHGR G + C ($r = 0.56$, $P < 10^{-6}$). A total of 1104 genes from the longest contig of each chromosome were included in the comparison. The ellipse shows 95% confidence intervals.

Table 3
Gene GC content (%GC) and G + C at third codon positions (%GC_3) in LHGR families

LHGR	*N*	%GC ± SE	%GC_3 ± SE
L	526	47.05 ± 0.29	49.10 ± 0.55
H1	172	52.32 ± 0.48	57.08 ± 0.94
H2	263	55.86 ± 0.36	63.62 ± 0.77
H3	143	60.78 ± 0.45	69.41 ± 0.98

Data from the longest contig of each chromosome.

AluYa5 elements. The trend to increase the compositional adjustment with time is confirmed by the fact that the very old LINE L2 repeats (>120 MY old) show the strongest correlation ($r = 0.73$, Fig. 14). Therefore, the compositional adjustment of Alu and LINE repeats to the isochores harboring them is a time-dependent process, thus fitting well within the framework of the neutral theory of molecular evolution (Kimura, 1983). In this way, Alus and LINEs, taken all together, appear to be neither beneficial nor harmful to the host; adaptive or maladaptive functions, if any, need to be demonstrated on an individual basis.

3.9. SNP density in different LHGR families

The SNP density (SNPs/kilobase) varies considerably among and within human chromosomes, the observed distribution of SNPs in 100 kbp fragments of the draft genome sequence showing far more pronounced variance than expected by chance (Venter et al., 2001).

By using our segmentation algorithm on the longest contig of each human chromosome, and by collecting all the annotated SNPs save those at CpG sites, we also found such a relationship for LHGRs (Fig. 15 and Table 4). A trend for SNP density to increase with LHGR GC level can be appreciated, the differences being significant between L and H3 LHGRs ($P < 10^{-3}$).

Fig. 15 also shows the variation in the densities of the six

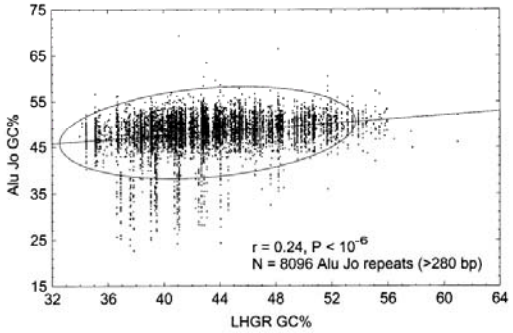

Fig. 11. AluJo GC content vs. LHGR G + C ($r = 0.24$, $P < 10^{-6}$). A total of 8096 AluJo repeats larger than 280 bp located on the longest contig of each chromosome (NCBI October 2001 freeze) were used in the comparison. The ellipse shows 95% confidence intervals.

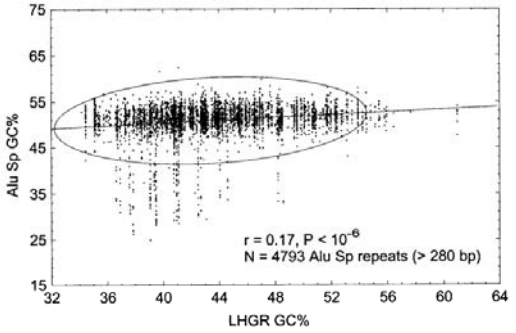

Fig. 12. AluSp GC content vs. LHGR G + C ($r = 0.17$, $P < 10^{-6}$). A total of 4793 AluSp repeats larger than 280 bp located on the longest contig of each chromosome (NCBI October 2001 freeze) were used in the comparison. The ellipse shows 95% confidence intervals.

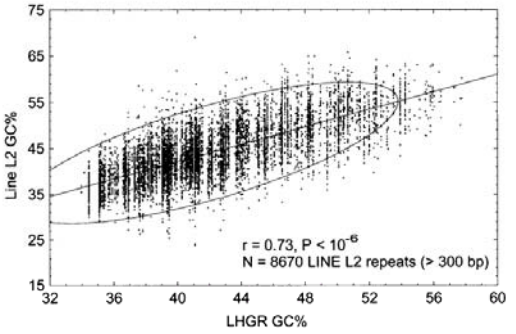

Fig. 14. LINE L2 GC content vs. LHGR G + C ($r = 0.74$, $P < 10^{-6}$). A total of 8670 LINE L2 repeats larger than 300 bp located on the longest contig of each chromosome (NCBI October 2001 freeze) were used in the comparison. The ellipse shows 95% confidence intervals.

possible base changes at the SNPs mapping on the different LHGR families. Considerable variations were observed among the different types of base changes, but, for a given base change, only slight variations among LHGR families were detected.

Lastly, we analyzed the ratio of transition and transversion substitutions in LHGR families (Table 4), and found variations with roughly the typical 2:1 ratio for mammalian genomes (Graur and Li, 2000). A slight trend for the transition/transversion ratio to increase with the LHGR G + C content appeared, although statistical significance was not reached.

3.10. Conclusions

By means of a hierarchical segmentation algorithm, specifically designed to determine the most statistically significant partition of a DNA sequence at each scale, here we have drawn isochore maps defined to single base pair resolution for the longest contig of each human chromosome. The fairly homogeneous regions found (LHGRs) displayed many of the features (G + C range, proportion of isochore classes, size distribution and relationship with gene and Alu densities) of the isochores identified through the centrifugation of vertebrate DNA fragments. The known correlations between different biological features (gene, repeat and SNP densities) and the isochore G + C content were also observed in LHGRs.

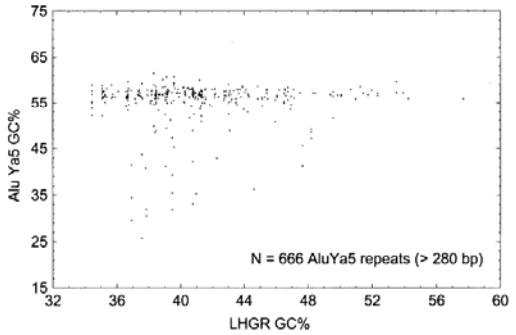

Fig. 13. AluYa5 GC content vs. LHGR G + C. A total of 666 AluYa5 repeats larger than 280 bp located on the longest contig of each chromosome (NCBI October 2001 freeze) were used in the comparison.

The isochore chromosome maps of the human genome presented here show several advantages over previous approaches based on moving windows: (1) sequence heterogeneities at different scales are simultaneously shown; (2) pair-wise differences in GC content between adjacent regions are all statistically significant; (3) isochore boundaries are defined to single base pair resolution; and (4) both gradual and abrupt isochore boundaries are simultaneously revealed.

The computational prescreening of isochore boundaries may have many applications in genomics: (1) the changes in replication timing known to occur at isochore boundaries (Tenzen et al., 1997) can now be exhaustively searched for at the predicted LHGR boundaries; (2) the genomic sequences can now be scanned for gene-rich regions, as we found that gene density depends heavily on the GC content of the LHGRs; (3) improvements in computational gene identification are also expected, as the specific compositional parameters of the corresponding isochores

Table 4
SNP density and ratio of transition to transversion substitutions in LHGR families

LHGR	SNP density (# SNP/kb) ± SE	Transition/transversion rate ± SE
L	0.47 ± 0.02	2.00 ± 0.05
H1	0.51 ± 0.06	2.09 ± 0.08
H2	0.57 ± 0.04	2.17 ± 0.09
H3	0.68 ± 0.07	2.10 ± 0.15

All the annotated SNPs (112,826) in the longest contig of each human chromosome, save those at CpG sites, were analyzed.

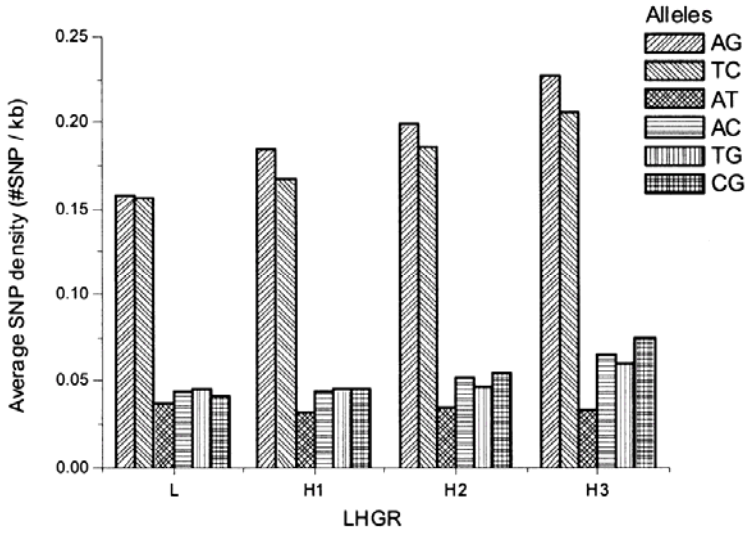

Fig. 15. Densities of different SNPs in LHGR compositional families. All the annotated SNPs in the longest contig of each human chromosome, save those at CpG sites, were analyzed. The densities of the six possible base changes are shown.

can now be taken into account as input for gene-finding programs (Burge and Karlin, 1997); in fact, we have recently shown (Carpena et al., 2002) that the prediction of the coding proportion in a sequence is better when LHGRs, instead of moving windows, are used; (4) in the same way, other programs making use of local compositional parameters to predict sequence patterns, as RepeatMasker (http://repeatmasker.genome.washington.edu), could be improved by considering LHGRs instead of moving windows; (5) the transitions from long-range to short-range linkage disequilibrium can coincide with switches in the isochore pattern (Eisenbarth et al., 2000, 2001); if so, the precise delimitation of isochore boundaries can help to predict the levels of linkage disequilibrium, thereby facilitating association studies, the most powerful current tool for the identification of genes underlying complex traits; and (6) the analysis of isochore chromosome maps in different genomes may allow new insights in the field of comparative genomics.

Acknowledgements

We would like to thank Giorgio Bernardi and Oliver K. Clay for encouragement and helpful discussions. Their warm hospitality to J.L.O. and P.B. at the Laboratory of Molecular Evolution of the Stazione Zoologica of Naples during the summer of 2001 is also acknowledged. We are also grateful to Dr Arian Smit for providing the RepeatMasker computer program. This work was supported by grant BIO99-0651-CO2-01 from the Spanish Government. The help of David Nesbitt with the English version of the manuscript is appreciated.

References

Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys. Rev. E 53, 5181–5189.

Bernaola-Galván, P., Oliver, J.L., Román-Roldán, R., 1999. Decomposition of DNA sequence complexity. Phys. Rev. Lett. 83, 3336–3339.

Bernaola-Galván, P., Grosse, I., Carpena, P., Oliver, J.L., Román-Roldán, R., Stanley, H.E., 2000. Finding borders between coding and non-coding regions by an entropic segmentation method. Phys. Rev. Lett. 85, 1342–1345.

Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L., 2001. Mapping isochores by entropic segmentation of long genome sequences. In: Sankoff, D., Lengauer, T. (Eds.), RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology, Montreal, Canada, ACM Press, New York, pp. 217–218.

Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L., 2002a. Study of statistical correlations in DNA. Gene 300, 105–115.

Bernaola-Galván, P., Oliver, J.L., Carpena, P., Clay, O., Bernardi, G., 2002b. Quantifying intragenomic heterogeneity in prokaryotic genomes, in preparation.

Bernardi, G., 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. 29, 445–476.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bernardi, G., 2001. Misunderstandings about isochores. Part 1. Gene 276, 3–13.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Bettecken, T., Aissani, B., Müller, C.R., Bernardi, G., 1992. Compositional mapping of the human dystrophin-encoding gene. Gene 122, 329–335.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 8–94.
Carpena, P., Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 2002. A simple and species-independent coding measure. Gene 300, 95–104.
De Sario, A., Geigl, E.-M., Palmieri, G., D'Urso, M., Bernardi, G., 1996. A compositional map of human chromosome band Xq28. Proc. Natl. Acad. Sci. USA 93, 1298–1302.
De Sario, A., Roizes, G., Allegre, N., Bernardi, G., 1997. A compositional map of the cen-q21 region of human chromosome 21. Gene 194, 107–113.
Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., et al., 1999. The DNA sequence of human chromosome 22. Nature 402, 489–495.
Eisenbarth, I., Vogel, G., Krone, W., Vogel, W., Assum, G., 2000. An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. Am. J. Hum. Genet. 67, 873–880.
Eisenbarth, I., Striebel, A.M., Moschgath, E., Vogel, W., Assum, G., 2001. Long-range sequence composition mirrors linkage disequilibrium pattern in a 1.13 Mb region of human chromosome 22. Hum. Mol. Genet. 10, 2833–2839.
Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.
Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H., Ikemura, T., 1995. A boundary of long-range G + C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. Genomics 25, 184–191.
Fukagawa, T., Nakamura, Y., Okumura, K., Nogami, M., Ando, A., Inoko, H., Saitou, N., Ikemura, T., 1996. Human pseudoautosomal boundary-like sequences: expression and involvement in evolutionary formation of the present-day pseudoautosomal boundary of human sex chromosomes. Hum. Mol. Genet. 5, 123–132.
Graur, D., Li, W.-H., 2000. Fundamentals of Molecular Evolution, 2nd Edition, Sinauer Associates, Sunderland, MA.
Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L., Stanley, H.E., 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence measure. Phys. Rev. E 65, 041905-1–041905-16.
Häring, D., Kypr, J., 2001. No isochores in the human chromosomes 21 and 22? Biochem. Biophys. Res. Commun. 280, 567–573.
Hattori, M., et al., 2000. The DNA sequence of human chromosome 21. Nature 405, 311–319.
Jurka, J., 2000. RepBase update: a database and an electronic journal of repetitive elements. Trends Genet. 16, 418–420.
Kapitonov, V., Jurka, J., 1996. The age of Alu subfamilies. J. Mol. Evol. 42, 59–65.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge.
Lander, E.S., Waterston, R.H., Sulston, J., Collins, F.S., et al., 2001. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Li, W., 2001. Delineating relative homogeneous G + C domains in DNA sequences. Gene 276, 57–72.
Li, W., Kaneko, K., 1992. Long range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 655.
Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J.L., 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Res. 8, 916–928.
Li, W., Bernaola-Galván, P., Carpena, P., Oliver, J.L., 2002. Isochores merit the prefix 'Iso', submitted for publication.
Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol. 108, 237–254.
Martínez Zapater, J.M., Marín, A., Oliver, J.L., 1993. Evolution of base composition in T-DNA genes from *Agrobacterium*. Mol. Biol. Evol. 10 (2), 437–448.
Mighell, A.J., Markham, A.F., Robinson, P.A., 1997. Alu sequences. FEBS Lett. 417, 1–5.
Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., Bernardi, G., 1991. The distribution of genes in the human genome. Gene 100, 181–187.
Nekrutenko, A., Li, W.H., 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. 10, 1986–1995.
Oliver, J.L., Marín, A., Martínez Zapater, J.M., 1990. Chloroplast genes transferred to the nuclear plant genome have adjusted to nuclear base composition and codon usage. Nucleic Acids Res. 18 (1), 65–73.
Oliver, J.L., Román-Roldán, R., Pérez, J., Bernaola-Galván, P., 1999. SEGMENT: identifying compositional domains in DNA sequences. Bioinformatics 15, 974–979.
Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.
Oliver, J.L., et al., 2002. IsoFinder: finding isochore boundaries on large sequence contigs, in preparation.
Pavlícek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.
Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.
Pilia, G., Little, R.D., Aïssani, B., Bernardi, G., Schlessinger, D., 1993. Isochores and CpG islands in YAC contigs in human X26.1-qter. Genomics 17, 456–462.
Román-Roldán, R., Bernaola-Galván, P., Oliver, J.L., 1998. Sequence compositional complexity of DNA through an entropic segmentation method. Phys. Rev. Lett. 80, 1344–1347.
Román-Roldán, R., et al., 2002. Information-theoretic symbolic sequence segmentation by maximum discrepancy ordering, in preparation.
Saccone, S., De Sario, A., Wiegant, J., Rap, A.K., Della Valle, G., Bernardi, G., 1993. Correlations between isochores and chromosomal bands in the human genome. Proc. Natl. Acad. Sci. USA 90, 11929–11933.
Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.
Soriano, P., Meunier-Rotival, M., Bernardi, G., 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. USA 80, 1816–1820.
Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J., Beck, S., 1999. Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. J. Mol. Biol. 291, 789–799.
Tenzen, T., Yamagata, T., Fukagawa, T., Sugaya, K., Ando, A., Inoko, H., Gojobori, T., Fujiyama, A., Okumura, K., Ikemura, T., 1997. Precise switching of DNA replication timing in the GC content transition area in the human MHC. Mol. Cell. Biol. 17, 4043–4050.
The MHC Sequencing Consortium, 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401, 921–923.
Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.
Voss, R., 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.
Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.

Oliver *et al*. (2004)

Oliver JL, Carpena P, **Hackenberg M**, Bernaola-Galván P. (2004) IsoFinder: computational prediction of isochores in genome sequences. Nucl Acids Res 32 (Web Server Issue): W287–W292 http://bioinfo2.ugr.es/publi/IsoFinder_NAR_04.pdf

En este trabajo se introduce un algoritmo mejorado de IsoFinder y se publica la interface web de este programa. A raíz de este trabajo, debido a los distintos niveles de acumulación de Alus observados en las isocoras H3, se decidió introducir una nueva isocora H4.

Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W287–W292***
DOI: 10.1093/nar/gkh399

IsoFinder: computational prediction of isochores in genome sequences

José L. Oliver*, Pedro Carpena[1], Michael Hackenberg and Pedro Bernaola-Galván[1]

Departamento de Genética, Instituto de Biotecnología, Facultad de Ciencias, Universidad de Granada and [1]Departamento de Física Aplicada II, Universidad de Málaga, Spain

Received January 22, 2004; Revised March 4, 2004; Accepted March 25, 2004

ABSTRACT

Isochores are long genome segments homogeneous in G+C. Here, we describe an algorithm (IsoFinder) running on the web (http://bioinfo2.ugr.es/IsoF/isofinder.html) able to predict isochores at the sequence level. We move a sliding pointer from left to right along the DNA sequence. At each position of the pointer, we compute the mean G+C values to the left and to the right of the pointer. We then determine the position of the pointer for which the difference between left and right mean values (as measured by the *t*-statistic) reaches its maximum. Next, we determine the statistical significance of this potential cutting point, after filtering out short-scale heterogeneities below 3 kb by applying a coarse-graining technique. Finally, the program checks whether this significance exceeds a probability threshold. If so, the sequence is cut at this point into two subsequences; otherwise, the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. This leads to the decomposition of a chromosome sequence into long homogeneous genome regions (LHGRs) with well-defined mean G+C contents, each significantly different from the G+C contents of the adjacent LHGRs. Most LHGRs can be identified with Bernardi's isochores, given their correlation with biological features such as gene density, SINE and LINE (short, long interspersed repetitive elements) densities, recombination rate or single nucleotide polymorphism variability. The resulting isochore maps are available at our web site (http://bioinfo2.ugr.es/isochores/), and also at the UCSC Genome Browser (http://genome.cse.ucsc.edu/).

INTRODUCTION

Mammalian genomes are made up of isochores, long DNA segments ($\gg$300 kb) fairly homogeneous in G+C which were first revealed by analytical ultracentrifugation of bulk DNA (1–3). The relevance of the isochore model for genome biology is based on observations of gene and SINE (short interspersed repetitive elements) densities, as well as recombination frequency, which are all higher in (G+C)-rich isochores, whereas LINEs (long interspersed repetitive elements) are denser in (G+C)-poor isochores (3). Besides compositional differences, genome segments separated by isochore boundaries also differ in replication timing, as in the isochores of the human major histocompatibility complex (MHC) locus (4), or in recombination rates, as in the human neurofibromatosis NF1 region (5). Isochores are found in a large variety of taxa, including plants and cold-blooded vertebrates, although they are more conspicuous in the genome of warm-blooded vertebrates [see (3) and references therein]. The isochore concept has increased our appreciation of the complexity and compositional variability of eukaryotic genomes (6), having been recently summarized as 'a fundamental level of genome organization' (7). The evolutionary origin and maintenance of isochores in today's genomes is currently the object of active debate (7–11).

The recent availability of the draft human genome sequence allowed for a direct test of the isochore model. A first analysis denied the existence of isochores in the human chromosomes 21 and 22 (12), while a second one ruled out a strict notion of isochores as compositionally homogeneous, concluding that isochores do not appear to deserve the prefix 'iso' (13). Both approaches, however, present serious drawbacks (14–17). First, denying the existence of isochores means denying the existence of compositional discontinuities in the human genome and going back to a genome organization characterized by a continuous compositional variation, a view shown to be wrong in the early 1970s (18). Second, the methodological problem common to the approaches denying isochores is that they take as a reference the random, uncorrelated model

*To whom correspondence should be addressed. Tel: +34 958243261; Fax: +34 958244073; Email: oliver@ugr.es

W288 Nucleic Acids Research, 2004, Vol. 32, Web Server issue

(in which every nucleotide is free to change) to test sequence homogeneity. This would lead to the absurd conclusion that only highly repetitive DNA sequences would be homogeneous. However, since 1981, when the heterogeneity within isochore families was quantified (19), we have known that the homogeneity of isochores is only relative—hence their definition as *fairly* homogeneous regions (3). Indeed, we have recently shown that when the appropriate scale and statistical test are chosen, isochores still merit the prefix 'iso' (20).

A third analysis (6), also based on the random, uncorrelated sequence model, reports a high degree of compositional heterogeneity in the human genome, but fails to detect the most conspicuous isochore boundary experimentally characterized to date, the one separating Class II and Class III regions of the human MHC locus (21–23).

In the past few years, we have been developing an algorithm, based on compositional segmentation, able to predict isochore boundaries at the sequence level (17,24–26). Most of the long homogeneous genome regions (LHGRs) predicted by this algorithm can be identified with Bernardi's isochores, given their correlation with biological features such as gene density, SINE and LINE densities, recombination rate and SNP (single nucleotide polymorphism) variability (17,26). Its efficacy as compared with other methods has also been proven (26,27). Now, after optimizing the algorithm by making several key improvements, we describe its implementation in a computer program (IsoFinder) running on the web (http://bioinfo2.ugr.es/IsoF/isofinder.html).

METHODS

The partition of a DNA sequence into fairly homogeneous, isochore-like segments is performed by means of a modified version of the entropic compositional segmentation. This algorithm was proposed in 1996 (24) and has proven useful in finding homogeneous regions (28,29) and measuring the compositional complexity or heterogeneity of DNA sequences (25).

While the original algorithm was designed to maximize the global difference in composition between adjacent segments and was intended for finding segments at all scales, the modified version presented here maximizes the G+C contrast between adjacent segments and looks for large-scale isochore-like segments.

The algorithm works as follows. Consider a DNA sequence of length L, along which we move a sliding pointer from left to right. At each position of the pointer, we compute the mean G+C values to the left and to the right of the pointer. To measure the difference between left and right mean values, we use the t-statistic. We next determine the position of the pointer for which t reaches its maximum value, t_{max}. Let us assume that this point divides the sequence into two subsequences of lengths L_{left} and L_{right}.

Next, we determine the statistical significance of the candidate to be a cutting point. As we wish to detect only isochore-like segments, we need to avoid the influence of short-scale heterogeneities on the statistical significance. Thus, we filter out the heterogeneity below a given minimum length (ℓ_0); i.e. we divide both subsequences into non-overlapping windows of length ℓ_0 and compute the G+C content in each window. In this way, we convert the subsequence of length L_{left} (L_{right}) into an array of L_{left}/ℓ_0 (L_{right}/ℓ_0) real numbers corresponding to the G+C content of each window of size ℓ_0. The online version of IsoFinder allows the user to choose among three different values for ℓ_0 (1, 2 and 3 kb) to perform the filtering procedure. However, the results presented in previous works (17,26) were obtained with $\ell_0 = 3$ kb, which corresponds to a homogeneity criterion for mammalian isochores, derived from ultracentrifugation of DNA at different molecular weights (2).

To determine the validity of the candidate to be a cutting point, we have to verify that the mean value of G+C of the left-hand-side windows is significantly different from the mean value of G+C in the right-hand-side windows. In doing so, we again use the t-Student statistics, but now computed on the two arrays of real numbers obtained after the filtering procedure, thus obtaining t_{filt}. In this way, the estimator of the difference in composition between left and right subsequences is not affected by short-scale heterogeneities (below ℓ_0).

Since the candidate to be a cutting point was found as the one maximizing the difference in composition, the statistical significance of t_{filt} cannot be measured by the standard Student's distribution. Thus, here we use the distribution of t_{filt} values [$P(\tau = t_{filt})$] obtained by means of Monte Carlo simulations (30). The significance level $P(\tau)$ of a possible cutting point with $t_{filt} = \tau$ is defined as the probability of obtaining the value τ or lower values within a random sequence. Thus, a series of N random numbers of fixed mean would remain unsegmented with probability $P(\tau)$. Finally, we check whether this significance exceeds a selected threshold P_0, usually taken to be 95%. If so, the sequence is cut at this point into two subsequences; otherwise, the sequence remains undivided. If the sequence is cut, the procedure continues recursively for each of the two resulting subsequences created by each cut. All resulting segments have a statistically significant difference in their means. The process stops when none of the possible cutting points has a significance exceeding P_0, and we say that the sequence has been segmented at the 'significance level P_0'. Our method leads to the partitioning of a DNA sequence into LHGRs with well-defined mean G+C levels, each significantly different from the mean G+C level of the adjacent LHGRs. Other algorithms proposed later (27,31) do not provide the statistical significance of the resulting partition of the sequence.

IMPLEMENTATION AND WEB INTERFACE

IsoFinder core routines were developed using the Lahey/Fujitsu Fortran 95 compiler under Debian Linux. A graphical gnuplot routine (http://www.gnuplot.info/) was used to generate the isochore maps. Lastly, a Perl CGI script for the Apache web server was used to integrate data input/output. Text fields to choose the significance level and the coarse-graining tract length to compute the statistical significance of cutting points are provided in the launch page (http://bioinfo2.ugr.es/IsoF/isofinder.html). The sequence to be segmented can be uploaded from a file on the local machine or pasted into the appropriate text field. Raw sequences and some of the standard sequence formats (EMBL, GenBank or FASTA) are accepted.

The output web page provides links to the results of sequence segmentation in three formats: (i) coordinates, sizes and G+C contents of the predicted isochores as an HTML table (Figure 1); (ii) the same but in plain text; and (iii) the isochore map of the sequence in PNG format (Figure 2). These results

Nucleic Acids Research, 2004, Vol. 32, Web Server issue **W289**

Isochore Predictions by IsoFinder

Sequence: out585990/1070968482__MHC -- Sig. level = 0.9500 -- Sig. method: Maximum --
Coarse graining: 3000 bp -- SCC: 0.6214E+01

LHGR	From	To	Size	GC%
1	1	299270	299270	46.08
2	299271	354226	54956	39.57
3	354227	364029	9803	53.95
4	364030	833239	469210	43.43
5	833240	1040717	207478	49.20
6	1040718	1168415	127698	46.24
7	1168416	1174527	6112	61.98
8	1174528	1230906	56379	52.47
9	1230907	1237030	6124	45.57
10	1237031	1244738	7708	49.73
11	1244739	1396980	152242	45.87
12	1396981	1469257	72277	52.32
13	1469258	1479458	10201	43.24
14	1479459	1490846	11388	51.70
15	1490847	1739420	248574	43.05
16	1739421	1841871	102451	47.73
17	1841872	2483966	642095	51.87
18	2483967	3054365	570399	40.19
19	3054366	3065923	11558	45.49
20	3065924	3074335	8412	51.84
21	3074336	3080554	6219	47.39
22	3080555	3088089	7535	51.36
23	3088090	3159420	71331	38.31
24	3159421	3384907	225487	42.95
25	3384908	3444780	59873	55.84
...	...	...	...	...
37	3661923	3673778	11856	52.85

Figure 1. Coordinates, sizes and GC contents of the predicted isochores as an HTML table.

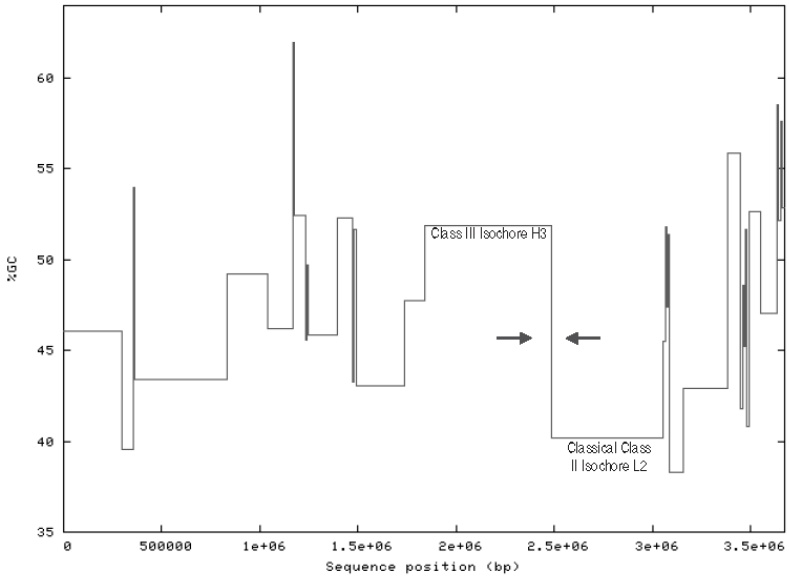

Figure 2. Isochore map of the human MHC region. The H3 and L2 isochores in this region are indicated. The arrows point to the pronounced isochore boundary between these two isochores.

remain available on the server for 24 h, and the user can access them through the hyperlinks provided in the output page.

EXAMPLES

MHC isochores

Class II (in an L2 isochore) and Class III (in an H3 isochore) regions of the human MHC are the only fully characterized isochores determined at the sequence level to date (21–23). Figure 2 shows the segmentation we made of the consensus sequence for the human MHC produced by the Human Chromosome 6 Sequencing Group at the Sanger Centre (http://www.sanger.ac.uk/HGP/Chr6/published_consensus.fasta). The IsoFinder algorithm enables precise location of the MHC isochore boundaries. The first three predicted boundaries were at positions 2483966, 3054365 and 1841872, corresponding to the sequence junction separating L2 and H3 isochores, the centromeric end of the isochore L2, and the telomeric end of the isochore H3, respectively. The remaining cuts on this sequence all fell outside of these two isochores, thus defining other homogeneous regions within the MHC sequence.

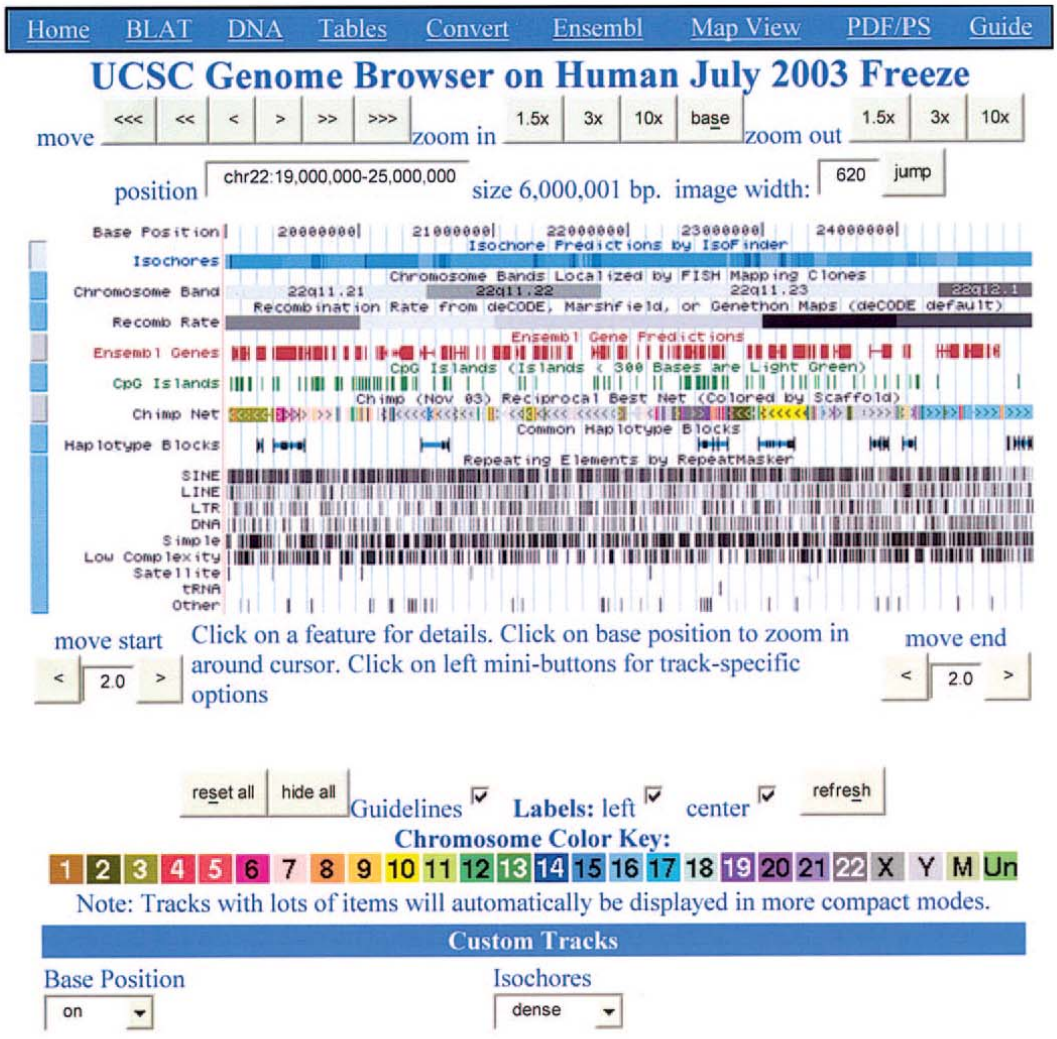

Figure 3. Isochore structure of a region of human chromosome 22 spanning 6 Mb between coordinates 19 000 000 and 25 000 000. Chromosome bands, recombination rate, gene density, CpG islands, best alignments with the chimpanzee genome, haplotype blocks and repeat densities are jointly shown with the isochore structure (top track) of this region.

Nucleic Acids Research, 2004, Vol. 32, Web Server issue **W291**

Isochore visualization with the Genome Browser at UCSC

The Genome Browser at UCSC (http://genome.ucsc.edu/) stacks annotation tracks beneath genome coordinate positions, allowing rapid visual correlation of different types of information. The user can look at a whole chromosome to get a feel for gene density, open a specific cytogenetic band to see a positionally mapped disease-gene candidate, or zoom in to a particular gene to view its spliced ESTs and possible alternative splicing. Isochore predictions by IsoFinder have also been integrated with all this comprehensive genome information (http://bioinfo2.ugr.es/isochores/), being available also at the UCSC genome browser as a custom track (http://genome.cse.ucsc.edu/goldenPath/customTracks/custTracks.html). As an example, the top track in Figure 3 shows the isochore structure of a region of human chromosome 22 integrated with information about chromosome bands, recombination rate, gene density, CpG islands, best alignments with the chimpanzee genome, haplotype blocks and repeat densities.

CONCLUSIONS AND FUTURE DEVELOPMENTS

The IsoFinder web server allows accurate and reliable isochore predictions in genome sequences. Four key advantages of the isochores predicted by IsoFinder are that (i) the isochore heterogeneity at different genome scales is shown in the same plot; (ii) pair-wise compositional differences between adjacent isochores are all statistically significant; (iii) isochore boundaries are accurately defined to single base pair resolution and (iv) both gradual and abrupt isochore boundaries are simultaneously revealed.

Apart from the MHC isochores (see above), no other experimentally confirmed isochore dataset exists. Therefore, the algorithm has not been validated against experimental data, using, e.g. parameters such as specificity or sensitivity. Instead, we have used numerical simulations to estimate the error in isochore-boundary determination [see section 2.3 and figure 1 in (26)]. We found that for typical isochore sizes (≈300 kb), the relative error ranges from 0.15 to 0.05%.

IsoFinder has also been compared with other available methods. In particular, an extensive comparison of IsoFinder against the wavelet multiresolution method (27) found a very good agreement between both algorithms, the differences being always lower than 1%. The remaining available methods either deny the existence of isochores (12,13) or fail in detecting even the clearer isochores experimentally identified to date (6).

Our algorithm has also been successfully used to relate isochore chromosome structure to gene density, SINE and LINE densities and SNP variability (17,26). We are now using IsoFinder to analyse the evolutionary factors driving the biased distribution of Alu retrotransposons in human isochores (M. Hackenberg and J. L. Oliver, submitted for publication). There are several other potential applications of the algorithm described here. First, one could scan the predicted LHGR boundaries searching for changes in replication timing known to occur at isochore boundaries (4). Second, anonymous, recently obtained genomic sequences can now be quickly and accurately scanned for gene-rich regions, as we found that gene density depends heavily on the G+C content of the LHGRs. Third, we have recently shown (32) that the prediction of the coding proportion in a sequence is better when LHGRs, instead of moving windows, are used. Therefore, improvements in computational gene identification are also expected, as the specific compositional parameters of the corresponding isochores can now be taken into account as input for gene-finding programs. Fourth, in the same way, other programs making use of local compositional parameters to predict sequence patterns, such as RepeatMasker, could be improved by considering LHGRs instead of moving windows. Finally, studies of comparative genomics could benefit from the detailed isochore chromosome maps provided by IsoFinder.

ACKNOWLEDGEMENTS

The help of David Nesbitt with the English version of the manuscript is appreciated. This work was supported by the Spanish Government (Grants Nos. BIO2002-04014-C03-01 and 02) and Plan Andaluz de Investigación (CVI-162). M.H. acknowledges a predoctoral grant from the University of Granada (Spain).

REFERENCES

1. Macaya,G., Thiery,J.P. and Bernardi,G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.*, **108**, 237–254.
2. Bernardi,G., Olofson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.
3. Bernardi,G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
4. Tenzen,T., Yamagata,T., Fukagawa,T., Sugaya,K., Ando,A., Inoko,H., Gojobori,T., Fujiyama,A., Okumura,K. and Ikemura,T. (1997) Precise switching of DNA replication timing in the GC content transition area in the human MHC. *Mol. Cell. Biol.*, **17**, 4043–4050.
5. Eisenbarth,I., Vogel,G., Krone,W., Vogel,W. and Assum,G. (2000) An isochore transition in the Nf1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am. J. Hum. Genet.*, **67**, 873–880.
6. Nekrutenko,A. and Li,W.H. (2000) Assesment of compositional heterogeneity within and between eukariotic genomes. *Genome Res.*, **10**, 1986–1995.
7. Eyre-Walker,A. and Hurst,W. (2001) The evolution of isochores. *Nat. Rev. Genet.*, **2**, 549–555.
8. Eyre-Walker,A. (1992) Evidence that both G+C rich and G+C poor isochores are replicated early and late in the cell cycle. *Nucleic Acids Res.*, **20**, 1497–1501.
9. Francino,M.P. and Ochman,H. (1999) Isochores result from mutation not selection. *Nature*, **400**, 31–32.
10. Fryxell,J.J. and Zuckerkandl,E. (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.*, **17**, 1371–1383.
11. Piganeau,G.I., Mouchiroud,D., Duret,L. and Gautier,Ch. (2002) Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *J. Mol. Evol.*, **54**, 129–133.
12. Häring,D. and Kypr,J. (2001) No isochores in the human chromosomes 21 and 22? *Biochem. Biophys. Res. Commun.*, **280**, 567–573.
13. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (International human genome sequencing consortium) (2001) Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.
14. Bernardi,G. (2001) Misunderstandings about isochores. Part 1. *Gene*, **276**, 3–13.
15. Clay,O. and Bernardi,G. (2002) Isochores: dream or reality? *Trends in Biotech.*, **20**, 237.

16. Li,W. (2001) Delineating relative homogeneous G+C domains in DNA sequences. *Gene*, **276**, 57–72.
17. Oliver,J.L., Carpena,P., Román-Roldán,R., Mata-Balaguer,T., Mejías-Romero,A., Hackenberg,M. and Bernaola-Galván,P. (2002) Isochore chromosome maps of the human genome. *Gene*, **300**, 117–127.
18. Filipski,J., Thiery,J.P. and Bernardi,G. (1973) An analysis of the bovine genome by Cs_2SO_4–Ag density gradient centrifugation. *J. Mol. Biol.*, **80**, 177–197.
19. Cuny,G., Soriano,P., Macaya,G. and Bernardi,G. (1981) The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. *Eur. J. Biochem.*, **115**, 227–233.
20. Li,W., Bernaola-Galvan,P., Carpena,P. and Oliver,J.L. (2003) Isochores merit the prefix 'Iso'. *Comp. Biol. Chem.*, **27**, 5–10.
21. Fukagawa,T., Sugaya,K., Matsumoto,K., Okumura,K., Ando,A., Inoko,H. and Ikemura,T. (1995) A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics*, **25**, 184–191.
22. Stephens,R., Horton,R., Humphray,S., Rowen,L., Trowsdale,J. and Beck,S. (1999) Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.*, **291**, 789–799.
23. The MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature*, **401**, 921–923.
24. Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (1996) Compositional segmentation and long-range fractal correlation in DNA sequences. *Phys. Rev. E.*, **53**, 5181–5189.
25. Román-Roldán,R., Bernaola-Galván,P. and Oliver,J.L. (1998) Sequence compositional complexity of DNA through an entropic segmentation method. *Phys. Rev. Lett.*, **80**, 1344–1347.
26. Oliver,J.L., Bernaola-Galván,P., Carpena,P. and Román-Roldán,R. (2001) Isochore chromosome maps of eukaryotic genomes. *Gene*, **276**, 47–56.
27. Wen,S.Y. and Zhang,C.T. (2003) Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochem. Biophys. Res. Comm.*, **311**, 215–222.
28. Azad,R.K., Bernaola-Galván,P., Ramaswamy,R. and Rao,J.S. (2002) Segmentation of genomic DNA through entropic divergence: power laws and scaling. *Phys. Rev. E*, **65**, 0519091–0519096.
29. Grosse,I., Bernaola-Galván,P., Carpena,P., Román-Roldán,R., Oliver,J.L. and Stanley,H.E. (2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E*, **65**, 0419051–04190516.
30. Bernaola–Galván,P., Ivanov,P.Ch., Amaral,L.A.N. and Stanley,H.E. (2001) Scale invariance in the nonstationarities of human heart rate. *Phys. Rev. Lett.*, **87**, 168105: 1–4.
31. Zhang,C.T. and Zhang,R. (2004) Isochore structures in the mouse genome. *Genomics*, **83**, 384–394.
32. Carpena,P., Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (2002) Simple and species-independent coding measure. *Gene*, **300**, 97–104.

Hackenberg y Oliver (2003)

Hackenberg M, Oliver JL. Alu Evolution and Mutation Pressure. RECOMB 2003 (Berlin, Germany) http://bioinfo2.ugr.es/recomb03_abstract.pdf

Esta aportación en forma de Poster al congreso RECOMB, contiene los análisis preliminares acerca de algunos mecanismos que podrían contribuir al cambio de densidad. Además, se presenta un análisis descriptivo de la evolución composicional de las Alus en función de la isocora.

Alu Evolution and Mutation Pressure

Michael Hackenberg[1] and José L. Oliver[1]

Keywords: Repetitive Elements, Alus, Isochores, Mutation Pressure, Genetic Distances

1 Introduction.

Since the finding that the Alu repetitive element family is much more abundant in GC-rich isochores (maximum in H2) than in AT-rich ones, an ongoing discussion exists on which mechanism leads to this biased distribution ([2], [4], [5]). It is known ([4]) that Alus prefer to insert in GC-poor genomic regions and therefore the question rises which mechanism shifts the density maximum from GC-poor to GC-rich regions. An extended opinion is that the Alus shift towards GC-rich regions to lessen the mutation pressure upon them (see [1], [4]). Alus are GC-rich and by installing in GC-rich isochores it is thought that the increased compositional match increases likewise their stability in the GC-rich parts of the genome.
However, taking into consideration all Alus irrespective of their age, it can hardly be determined if this mechanism is the most important or just plays a minor role. To trace the evolution of the Alus in the human genome, we determined their densities and GC-contents as a function of their age. We first align each Alu in the genome to its family consensus using the RepeatMasker algorithm. The Tamura-Nei distance was then used to estimate the Alu ages. The partition of the contigs into isochores was performed using the IsoFinder segmentation algorithm of Oliver, et al (see [3] and references therein).

2 Compositional adjustment of repeats to the isochores

By analysing the Alus as a function of their age, it is possible to monitor the changes suffered by the Alus over time. In Figure 1 we show a) the relative density ratios H*/L (densities in the H isochores divided by the densities in the L isochores) as a function of their age and b) the course of the Alu GC-contents in the different isochores also as a function of their age. It can be seen (graph b) that up to genetic distances of around 0.08 no correlation between the GC of the isochore and the element becomes established. Just from these genetic distances on, the elements seem to note the influence of the isochore. However, as can be inferred from graph a, at these genetic distances the maximum has already shifted to the H2 isochores. This means that long before the elements seem to note the compositional pressure of the isochore, the density maximum has changed. By calculating the GC-content without the CpG dinucleotides (not shown here), it can be observed that the change in GC suffered by young elements is predominantly due to mutations of the CpGs, which do not seem to depend on the isochore. The %GC without the CpGs hardly vary with the Alu age and the change stays within 1%. If mutation pressure was the main mechanism, one might expect a greater influence of the isochore on the element's base composition.
As a second quantitative measurement of mutation pressure strength we have calculated the number of deletions and insertions suffered by the elements, which is only possible using aligned sequences. Although the results confirm the general validity of the mutation pressure theory (more GC insertions in GC-rich than in AT-rich parts and more AT insertions in AT-rich than in GC-rich regions), we found again that the differences of this mutation pattern between the different isochores is just weakly pronounced. We found that just one out of ten elements show one G or C additionally inserted in H3 compared with the L isochores. The AT insertion mechanism acts

[1] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E-18071-Granada, Spain. E-mail: genmol@ugr.es

slightly stronger but also just one out of five repeats has one A or T additionally inserted in an L isochore compared with H3.

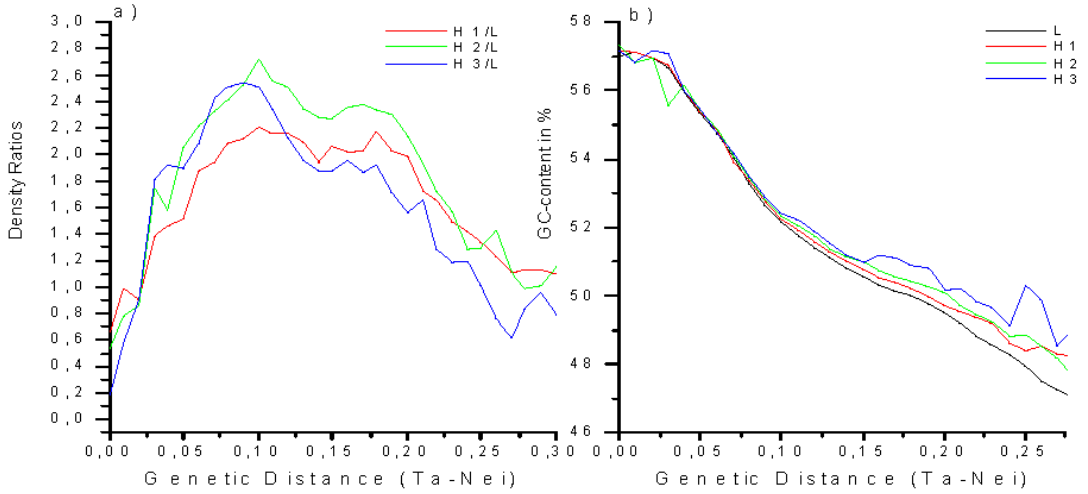

Figure 1: Density ratios between H and L isochores (a). It can be seen that the densities shift very fast from the L isochores towards the H isochores. The GC-content of the Alus in the different isochores are depicted in (b). The GC-content do not show a dependence of the isochores below genetic distances of around 0.08.

3 Summary.

An analysis of the Alus as a function of their age is presented. The main change in GC is caused by CpG mutations, which do not depend on the isochore. A correlation between the GC-content of the isochores and the elements is not established until genetic distances of around 0.08. As the density shift takes place yet around genetic distances of 0.03 we infer that, although mutation pressure has definitely an influence on the base composition it may not be sufficient to explain the density shift and the biased distribution of the Alus in the human genome. The weak influence of mutation pressure on the distribution of the Alu elements is also confirmed by the analysis of AT and GC insertions.

4 References and bibliography.

[1] Giorgio Bernardi. Misunderstandings about isochores. Part 1. Gene 276 (2001) p. 3-13.

[2] Zhenglong Gu, Haidong Wang, Anton Nekrutenko, Wen-Hsiung Li. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. Gene 259 (2000) p. 81-88.

[3] José L. Oliver, Pedro Bernaola-Galván, Pedro Carpena and Ramón Román-Roldán. Isochore chromosome maps of eukaryotic genomes. GENE 276: 47-56.

[4] A. Pavlicek, K. Jabbari, J. Paces, Vaclav Paces, Jiri Henjar, Giorgio Bernardi. Similar integration but different stability of Alus and LINES in the human genome. Gene 276 (2001) p. 39-45.

[5] Arian FA Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current Opinion in Genetics & Development, 9 (1999) p. 657-663.

Hackenberg *et al*. (2004)

Hackenberg M, Bernaola-Galván P, Carpena P, Oliver JL. Alu clustering in the human Genome: Origins and Consequences. RECOMB 2004 (San Diego, USA). http://bioinfo2.ugr.es/publi/Recomb04_Abstract.pdf

Se presentan algunos análisis preliminares acerca de la clusterización de Alus en el genoma humano.

H10. Alu Clustering in the Human Genome: Origins and Consequences

Michael Hackenberg[1] and Jose L. Oliver[1]

Keywords: Alu, Isochores, Recombination, Evolution, Negative Selection, Alu Clustering

1 Introduction

Approximately 10% of the human genome is made up of Alu elements. Although the Alu repeats have been the focus of investigation over the last few years, many aspects of their genomic role still remain unsolved. For example, they appeared in the genome some 50-80 MYA and had their major spread approx. 30 MYA, which is about two orders of magnitude higher than the present amplification rate [1]. Other topics which are still controversial deal with the Alu distribution over the human genome. In general, Alu density is positively correlated with the GC content of the genomic region (isochore) in which they reside [2]. However, recently inserted Alus present a different pattern. As they depend on the LINE-1 transposition machinery, they show an initial density maximum in GC-poor regions. In [3], we analysed the Alu distribution as a function of evolutionary age and isochore membership, as well as the densities of possible recombination outcomes, inferring that recombination is probably the most important mechanism driving the density shift from GC-poor (L) to GC-rich (H) isochores. Here, we present an analysis of the Alus as a function of isochore membership and physical distance to the next Alu repeat (Distance to the Nearest Neighbour - DNN), which reinforces our argument and sheds light on this controversy from a different side.

We analysed the human reference sequence (April 2003 freeze, UCSC version hg15), based on NCBI Build 34 and produced by the International Human Genome Sequencing Consortium (IHGSC), downloaded from (ftp://genome.ucsc.edu/goldenpath/10april2003/bigZips/chromFa.zip). The partition of the chromosomes into isochores was performed using the IsoFinder segmentation algorithm (see [4] and references therein). The Alu densities indicated have been calculated as the number of elements per 10 kb.

2 Results and Discussion

As a measure for Alu clustering, we used the distance to the next Alu (DNN). The Alu density ratios (densities in H isochores divided by densities in L isochores) in the different isochores as a function of DNN are shown in Figure 1. For short distances, the densities are up to 8 times higher in H than in L isochores, which corresponds to an extremely higher clustering in the H isochores. For example, in the GC-richest isochore H4, almost 40% of all Alus are closer than 20 bp to each other, whereas in the GC-poorest isochore L1 this fraction declines to 5%. With growing distances to the closest Alu, however, the density maximum shifts gradually towards the L isochores; thus, Alus with distances greater than 2000 bp to the next Alu show a density maximum in L isochores. This means that interactions with other Alus are crucial for the density shift, as single Alus with long distances to the next Alu remain in L isochores. There are two possible explanations: Alu/Alu recombination and preferential insertion in or near preexisting Alus, both leading to pronounced Alu clustering. It is known that the poly-A tail of an Alu may constitute a good insertion target for

[1] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain. E-mail: genmol@ugr.es, oliver@ugr.es

new insertions [5]. Analysing the densities of Alu/Alu insertions (i.e., insertions in the AT-rich linker, unpublished), we found that the insertion probability is positively correlated with the isochore GC content. We believe, nevertheless, that the preferential targeting on preexisting Alus in H isochores can only partially contribute to form the pronounced overall Alu density maximum in H2/H3 isochores, but that the density shift and the formation of the biased distribution over isochores is driven mainly by recombination [3].

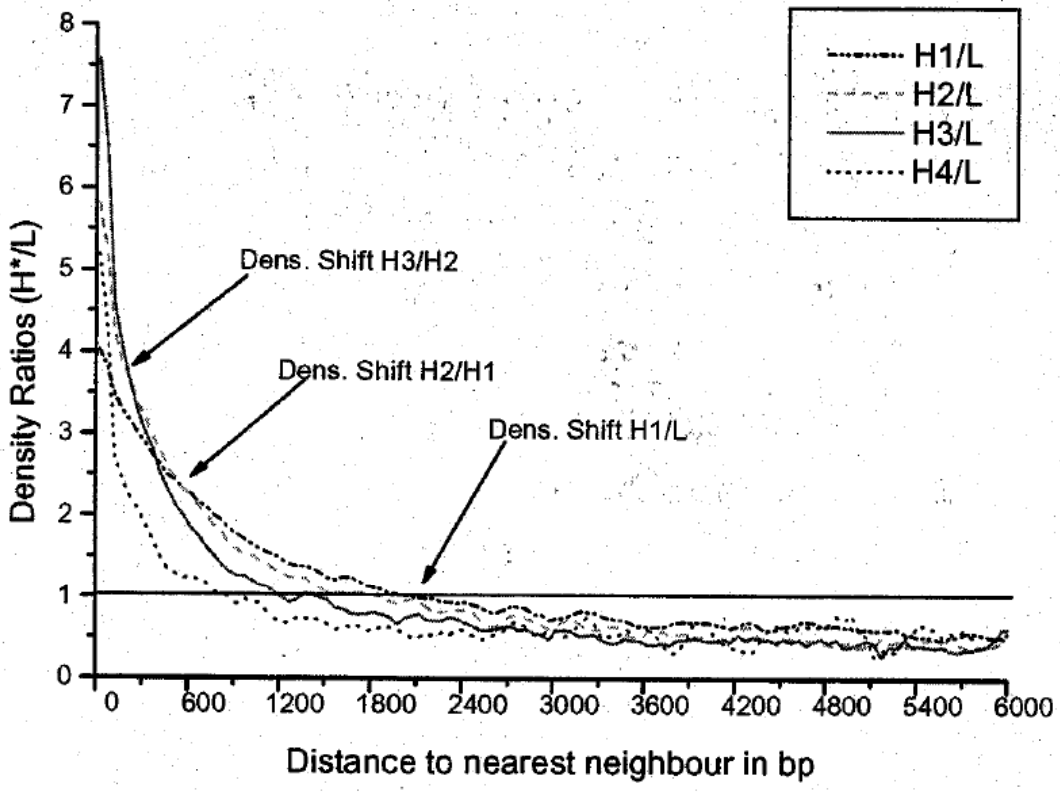

Figure 1: Alu density ratios as a function of DNN (Distance to Nearest Neighbour) and isochore membership are shown. It can be seen that very closely spaced Alus have a pronounced density maximum in the H3 isochore. With growing distance, the density maximum shifts gradually towards L isochores.

3 References and bibliography

[1] Batzer M.A., Deininger P.L. 2002. Alu repeats and human genomic diversity. Nature Reviews Genetics 3:1-10.

[2] Bernardi G. 2001. Misunderstandings about isochores. Part 1. Gene 276:3-13.

[3] Hackenberg M, and Oliver J.L. 2004. .The biased distribution of Alus in the human isochores might be driven by recombination. Submitted.

[4] Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. IsoFinder: computational prediction of isochores in genome sequences, in preparation

[5] Stenger, J.E. et al. 2001. Biased Distribution of Inverted and Direct *Alus* in the Human Genome: Implications for Insertion, Exclusion, and Genome Stability. Genome Research 11:12-27.

Hackenberg *et al*. (2005)

Hackenberg M, Bernaola-Galván P, Carpena P, Oliver JL. (2005) The biased distribution of Alus in human isochores might be driven by recombination. J Mol Evol 60:365-377 (http://bioinfo2.ugr.es/publi/jme05.pdf)

Este trabajo comprende en gran medida las secciones 4.1 y 4.2 de esta memoria.

J Mol Evol (2005) 60:365–377
DOI: 10.1007/s00239-004-0197-2

JOURNAL OF MOLECULAR EVOLUTION

The Biased Distribution of Alus in Human Isochores Might Be Driven by Recombination

Michael Hackenberg,[1] Pedro Bernaola-Galván,[2] Pedro Carpena,[2] José L. Oliver[1]

[1] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain
[2] Departamento de Fisica Aplicada II, Universidad de Málaga, Málaga, Spain

Received: 28 June 2004 / Accepted: 1 October 2004 [*Reviewing Editor:* Dr. Jerzy Jurka]

Abstract. Alu retrotransposons do not show a homogeneous distribution over the human genome but have a higher density in GC-rich (H) than in AT-rich (L) isochores. However, since they preferentially insert into the L isochores, the question arises: What is the evolutionary mechanism that shifts the Alu density maximum from L to H isochores? To disclose the role played by each of the potential mechanisms involved in such biased distribution, we carried out a genome-wide analysis of the density of the Alus as a function of their evolutionary age, isochore membership, and intron vs. intergene location. Since Alus depend on the retrotransposase encoded by the LINE1 elements, we also studied the distribution of LINE1 to provide a complete evolutionary scenario. We consecutively check, and discard, the contributions of the Alu/LINE1 competition for retrotransposase, compositional matching pressure, and Alu overrepresentation in introns. In analyzing the role played by unequal recombination, we scan the genome for Alu trimers, a direct product of Alu Alu recombination. Through computer simulations, we show that such trimers are much more frequent than expected, the observed/expected ratio being higher in L than in H isochores. This result, together with the known higher selective disadvantage of recombination products in H isochores, points to Alu Alu recombination as the main agent provoking the density shift of Alus toward the GC-rich parts of the genome. Two independent pieces of evidence the lower evolutionary divergence shown by recently inserted Alu subfamilies and the higher frequency of old stand-alone Alus in L isochores support such a conclusion. Other evolutionary factors, such as population bottlenecks during primate speciation, may have accelerated the fast accumulation of Alus in GC-rich isochores.

Key words: Alu LINE1 Retrotransposons Alu Alu recombination Selection Isochores Human genome

Introduction

A large part of the human genome (about 45%) is made up of mobile, repeated elements (Jurka 1995; Batzer and Deininger 2002; Deininger et al. 2003). The two most frequent interspersed repeats are Alu and LINE1 retrotransposons, with around 1,100,000 and 700,000 copies in the genome, respectively. Both mobilize (i.e., retrotranspose) via an RNA intermediate. Alus, relatively short elements (≈300 bp), are rich in GC and CpG dinucleotides, containing roughly one-third of all CpGs in the human genome. LINE1s are rather long, GC-poor elements (≈6500 bp), although in the genome they are often 5′ truncated and their average length may reach less than 500 bp in GC-rich isochores (see Table 1). The LINE1 element is an autonomous retrotransposon that has two open reading frames, one of which codes for a retrotranscriptase/endonuclease (retrotranspos-

Correspondence to: José L. Oliver; *email:* oliver@ugr.es

366

Table 1. Numbers, average lengths, GC ranges and mean %GC of the different isochore classes

	Isochores				Alus			LINE1		
Isochore class	*N*	Average length (bp)	%GC range	Mean %GC	*N*	%GC	Average length (bp)	*N*	%GC	Average length (bp)
L1	1,020	797,025	<37.77	36.33	150,157	51.03	282	178,476	33.95	974
L2	1,634	495,487	37.77–41.32	39.51	258,670	51.16	283	163,087	34.51	887
H1	1,843	335,261	41.32–46.91	43.80	333,435	51.41	281	110,117	35.07	658
H2	724	260,731	46.91–51.23	48.86	120,123	51.64	281	26,574	36.58	529
H3	315	235,565	51.23–54.75	52.72	52,553	51.93	281	7,777	38.46	483
H4	200	163,698	≥54.75	57.32	16,322	52.53	279	2,542	41.41	435
Total	5,736				931,260			488,633		

Note. Only isochores longer than 50 kb were included. The numbers, lengths, and mean %GC contents of the Alus and LINE1s harbored by the different isochore classes are also shown. Note that in the definition of the isochores (GC ranges) the lower boundary is included and the upper one excluded.

ase). It is has been recently demonstrated that LINE1 and Alus are amplified by the same enzymatic apparatus of active LINE1 elements (Dewannieux et al. 2003). Therefore, Alus should have the same insertion preference for AT-rich regions as LINE1. The binding site of the endonuclease is TT/AAAA, the slash indicating the cutting point. Indeed, the preference for AT-rich regions shown by young Alus has recently been demonstrated (Pavlíček et al. 2001; IHGSC 2001).

While the density maximum of the LINE1 elements remains in AT-rich regions, the maximum for Alus is found in the H2 isochores (Bernardi 2001). Thus, if the insertion pattern is the same for Alus and LINE1s, the question arises: What evolutionary mechanism shifts the Alu density maximum from L to H isochores? This question has raised controversy in recent years, and various proposals have been advanced, most of which involve either positive or negative/purifying selection. Positive selection could act only if Alus had some identifiable function that would favor the organism. In recent years, various positive effects of Alus have been proposed (Chu et al. 1998; Schmid 1998; Deininger and Batzer 1999). For example, in many species, Alus are transcribed under conditions of stress, and the resulting RNAs specifically bind a particular protein kinase (PKR), blocking its ability to inhibit protein translation (Chu et al. 1998; Schmid 1998). Alu RNAs would thus promote protein translation under stress. Their location in open chromatin (which tends to correlate with GC content of the genomic region) appears to facilitate this task (Smit 1999). Alus can also insert into mature messenger RNAs via a splicing-mediated process termed exonization (Lev-Maor et al. 2003). Through exonization, intronic Alus can be converted into new coding exons. Indeed, about 5% of alternatively spliced internal exons in the human genome originate in Alu sequences (Sorek et al. 2002). Another possible function of Alus relates to gene expression levels. Alus are CpG rich, and by inserting near a gene this newly introduced CpG island may alter the expression pattern of the gene (Britten 1996). Since the genes are denser in GC-rich than in GC-poor regions, an equivalent distribution of the Alus may be positively selected for. However, generalized positive selection for Alus is problematic (Deininger and Batzer 1999) that is, most of the attributed Alu functions refer to individual or few Alu repeats and not to the full set of repeats.

Consequently, most approaches to this problem involve negative selection. This mechanism could also provoke the Alu density shift if it were able to exclude or remove Alus more effectively from AT-rich than from GC-rich regions. Different negative influences of the Alus in AT-rich regions have been proposed. For example, the accumulation of Alus in GC-poor regions is counter-selected because it would severely change the local composition and/or the chromatin structure of these genomic regions, possibly affecting gene transcription (Rynditch et al. 1998). Bernardi's group also proposed a negative selection theory in which the major part of compositional matching is achieved by selective gain and loss of DNA, particularly repetitive DNA (Pavliček et al. 2001). This hypothesis is supported by comparative analyses of mammalian genomes (Pavliček et al. 2002; Paces et al. 2004).

A related mechanism is the compositional matching of repeats to the isochores harboring them. Compositional matching (or adjustment) to different genomic GC contents has been demonstrated in homologous genes and noncoding sequences of microorganisms and mitochondrial genomes (Jukes and Bushan 1986), as well as in genes moved between genomes (Oliver et al. 1990; Martínez-Zapater et al. 1993). Alu sequences can undergo compositional matching as well (Filipski et al. 1989) base substi-

tutions decrease the average GC content of Alus located in AT-rich regions, whereas the Alus located in GC-rich genomic regions do not change their already high GC content. In this way, compositionally nonmatching, recently inserted Alus could undergo stronger mutational pressure than the repeats adjusted to their host isochores. Since most of the elements found in the genome have adjusted their density maximum to regions where they have a highest compositional match, compositional matching can be considered responsible for the observed density distribution.

Another mechanism which appears to influence the distribution of these retrotransposons, and which does not require selection to act, involves Alu/LINE1 interactions. It has been argued (Gu et al. 2000) that Alus may switch their insertion preferences toward the GC-richer parts, thereby avoiding the competition for retrotransposase with the LINE1 elements in the L isochores.

A last mechanism that might be involved in the Alu density shift is the unequal homologous recombination among Alus repeats (Alu Alu recombination), which is thought to be a recurrent process. Some authors (Deininger and Batzer 1999; Batzer and Deininger 2002) have argued that the selective disadvantage posed by Alu Alu recombination is higher in H than in L isochores, given the positive correlation between gene density and GC content. The deletion rate of Alu elements should then be higher in L isochores. The evolutionary outcome of such a process would be a progressive decline in the relative density of Alus in L isochores, thus leading to the density shift.

However, by considering only average values or combining old with young elements, it is difficult to identify which of these is the main mechanism involved. Here, we present a new approach to this subject by analyzing Alu and LINE1 elements as a function of their evolutionary age, isochore membership, and intron vs. intergene location. This enables us to trace Alu evolution in the genome and in different genomic compartments, thereby disclosing the roles of the different mechanisms potentially involved in the Alu density shift.

Data and Methods

Chromosome Sequences and Gene Data

We used the human reference sequence (April 2003 freeze; UCSC version hg15), based on NCBI Build 34 and produced by the International Human Genome Sequencing Consortium (IHGSC). The 24 chromosome sequences, in FASTA format, were downloaded from ftp://genome.ucsc.edu/goldenpath/10april2003/bigZips/chromFa.zip. Gene data were derived from the GeneID database downloaded from the UCSC Genome Browser (http://genome.cse.ucsc.edu/).

Finding Isochores

In analyzing the relation of repeats with GC content, it is necessary to avoid the subjectivity implicit in choosing a window size to compute the surrounding GC for repeats, which would lead to unpredictable results (Bernaola-Galván et al. 1996; Li 2001; Oliver et al. 2001, 2002). Therefore, we first located isochores, thus obtaining GC values truly representative of the genomic environment in which the repeats are inserted.

Chromosome sequences were partitioned into fairly homogeneous genome regions (isochores) by using IsoFinder (Oliver et al. 2004; http://bioinfo2.ugr.es/IsoF/isofinder.html), an improved version of the segmentation algorithm described earlier (Bernaola-Galván et al. 1996; Oliver et al. 2001, 2002). Briefly, we move a sliding pointer from left to right along the DNA sequence. At each position of the pointer, we compute the mean G+C values to the left and to the right of the pointer. We then determine the position of the pointer for which the difference between left and right mean values (as measured by the t statistic) reaches its maximum. Next, we determine the statistical significance of this potential cutting point, after filtering out short-scale heterogeneities below 3 kb by applying a coarse-graining technique. Finally, the program checks whether this significance exceeds a probability threshold. If so, the sequence is cut at this point into two subsequences; otherwise, the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. This leads to partitioning of a DNA sequence into long homogeneous genome regions (LHGRs) with a well-defined mean GC level, each significantly different (at the 95% confidence) from the mean GC level of the adjacent regions. LHGRs may be assimilated into Bernardi's isochores (Oliver et al. 2001, 2002). The coordinates, sizes, and GC contents for all isochores identified in each human chromosome are available at our Web site (Online Resource on Isochore Mapping: http://bioinfo2.ugr.es/isochores) and also at the UCSC Genome Browser (http://genome.cse.ucsc.edu/).

We classified isochores into discrete compositional classes (Table 1) using the isochore abundances reported by Bernardi's group (Bernardi et al. 1985; Bernardi 2000). However, the classification given here differs slightly from that given by Bernardi, as we introduced a new human isochore (H4) by splitting the old H3 isochore into two new ones. The new H3 isochore corresponds to two-thirds of the old H3, while the new H4 is made up of the remaining one-third richest in GC of the old H3 (1.6% of the total DNA). We have chosen this classification because Alu density markedly decreases in this newly introduced H4 isochore, whereas gene density reaches its maximum. In this way, Alu and gene densities are positively correlated in GC-poor isochores but negatively correlated in the GC-richest isochore.

Scanning the Human Genome Sequence for Alus and LINE1s

To mask the repeats in the genome and align them with their respective family consensus, we used the RepeatMasker algorithm by Arian Smit (A.F.A. Smit and P. Green, unpublished data: http://repeatmasker.genome.washington.edu) and Repbase Update (Jurka 2000). We obtained the alignments between each element found in the genome (locus) and their family consensus by using the –a option in RepeatMasker. The proportion of nucleotide sites at which the two aligned sequences differ (p) was then transformed into evolutionary distance using the method of Tamura and Nei (1993). This method takes into account both the transition/transversion and the GC-content biases (Nei and Kumar 2000). Although we masked the whole genome, only repeats located in isochores longer than 50 kb were used here (see Table 1 for iso-

chore statistics). The total amount of DNA in the present analysis was thus limited to 2.58 Gb.

The repeats are often found fragmented in the genome. Based on the identification number (ID) provided by RepeatMasker, we joined together those fragments. As this ID number is assigned also to repeats that have potentially the same origin, we take into account the consensus coordinates to assemble only *true* fragments, and we used exclusively those *reassembled* repeats.

A total of 931,260 (≈85%) Alus and 488,633 (≈70%) LINE1 retrotransposons in 5736 human isochores longer than 50 kb were analyzed. Table 1 shows the number, average lengths, GC range, and mean %GC of the different isochore classes analyzed in this study. The numbers, lengths, and mean %GC contents of the Alus and LINE1s harbored by different isochore classes are also shown.

Perl Scripts

Perl scripts were developed to parse RepeatMasker output and compute the repeat GC content and evolutionary distance to the family consensus, as well as to perform all the subsequent data analyses. These scripts are available from the authors upon request.

Simulation of the Alu Insertion Process

To check the observed recombination products against the ones found in a random insertion process, we carried out a simulation of Alu insertion in the following way. (1) We used RepeatMasker to generate an exhaustive list of the Alus present in a particular genomic sequence we wished to simulate. (2) The Alus were ranked by age (evolutionary distance). (3) We first inserted the oldest Alu of the ranked list into a random position of a clean (Alu-free) random sequence, (4) We iterated step 3 for all the remaining Alus in the ranked list. By inserting the Alus in this ordered way, we generated the best approximation to the true insertion history. We also took into consideration the possibility of insertions into preexisting Alus and the "blow-up" of the sequence produced by the Alu insertions. In this way, at each step of the simulation process the sequence generated increases its length, finally reaching the same length as the original genome sequence we wish to simulate. In computing the expected values of Alu trimers for a given genome sequence, we generated a hundred simulated sequences and then averaged the observed numbers of Alu trimers in each sequence.

Results and Discussion

The Density Shift: Alus Are Now Preferentially Located in H Isochores

To analyze the evolution of retrotransposon densities in the genome, we computed the *density ratios* i.e., the densities in the H isochores divided by the densities in the L isochores (H*/L, where L is the weighted sum of L1 + L2 isochores). This means, for example, that if the ratios are lower than 1, the absolute-density maximum is located in the L isochores.

Figure 1 shows the Alu and LINE1 density ratios as a function of age, estimated by evolutionary distance (see Data and Methods). Below evolutionary distances of around 0.025, the absolute Alu density maximum is located in the L isochores, given that all the ratios are smaller than one. However, the density maximum shifts very fast toward the H isochores, which from this distance on register the maximum for Alus.

Note that the scale in Fig. 1 is not linear because CpG dinucleotides were included in the computations. When the fast CpG mutations were excluded, the density shift could not be resolved so clearly. We refer to this observable density shift at 0.025 as the "last" density shift, as it can be assumed that Alus during their evolution continually shifted their density.

Figure 1 (bottom) shows that the insertion pattern of the LINE1s is similar to that of Alus, but the AT-rich preference of LINE1 elements is more pronounced. This same observation was previously made by Pavliček et al. (2001), who offered two possible explanations: either the LINE1 elements are excluded from the GC-rich isochores or the Alus are removed from the AT-rich isochores. In this way, all happens as if some mechanism, acting after the insertion of these retrotransposons into the genome, were able to remove or exclude the Alus more effectively from the L than from the H isochores. What, then, is the nature of this mechanism(s)?

Factors Mediating the Initial Insertion Process: Alu/LINE1 Competition for Retrotransposase

The first mechanism that may change the Alu genomic distribution is competition for retrotransposase mediating the initial insertion process. This enzyme is encoded only by LINE1 but is used by both LINE1 and Alu repeats to retrotranscribe and reinsert into the genome. Gu et al. (2000) proposed that, since Alus use the retrotransposase encoded by LINE1, and since LINE1 prefers to insert into AT-rich regions, Alus could avoid competition for enzyme with LINE1 elements if they insert into GC-rich regions, which are replicated at different times from AT regions in the cell cycle. In this way, Alus could have shifted their density maximum to H isochores.

To test this hypothesis, we plotted the absolute densities of Alus and LINE1s as a function of isochore membership and evolutionary distance (Fig. 2). The LINE1s show a pronounced density minimum in the L1 and L2 isochores between evolutionary distances of 0.04 and 0.15 (Fig. 2, bottom), which appears concurrently with the maximum spread of the Alus (Fig. 2, top). Therefore, although the competition between the two elements is evident, the Alu insertion into L isochores does not appear to have been hampered. Instead, the massive Alu radiation has apparently somewhat impeded the LINE1 propagation, particularly in these isochores. Therefore, it appears that Alu/LINE1 competition for retrotrans-

369

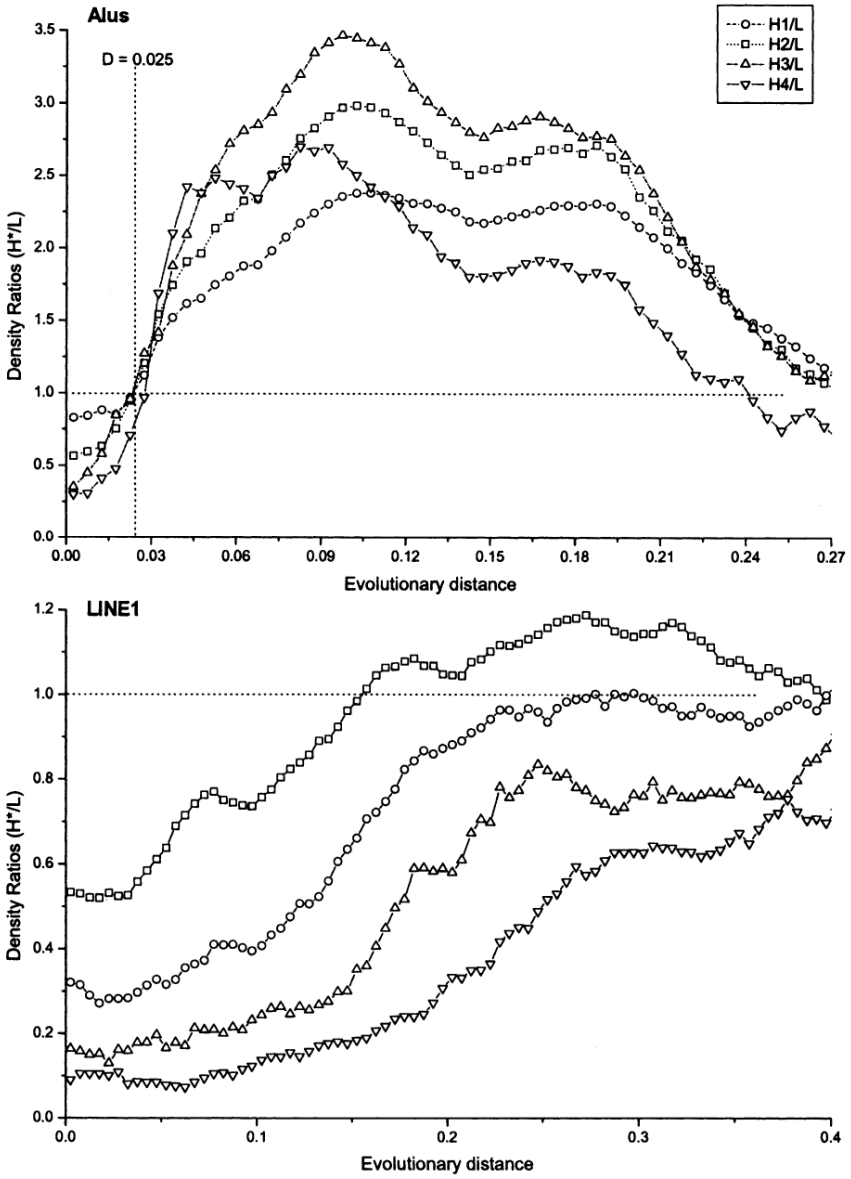

Fig. 1. Alu (**top**) and LINE1 (**bottom**) density ratios (H*/L) as a function of age, where L is the weighted density sum of L1 and L2 isochores. The CpG mutations are included to compute the genetic distance, and therefore the time scale (abscissa) is not linear.

posase does not hinder Alu insertion and spread in the L isochores.

In their paper, Gu et al. (2000) assume that the competition between LINE1 and Alu RNAs for LINE1 ORF2 proteins takes place in the nucleus. This assumption is probably wrong. Current evidence (e.g., the LINE1 cis-preference) suggests that the interaction takes place on ribosomes (see, e.g., Boeke 1997), and consequently no competition occurs near the nuclear DNA.

The Density Shift of the Alus Occurred Long Before They Began to Match the Host Isochore Composition

Being GC rich, Alus have a higher compositional match in the H isochores, while the GC-poor LINE1s match better to the L isochores. In general, most elements tend to fit their density maximum in regions where they compositionally match (Filipski et al. 1989). If so, compositional matching might be responsible for the density shift of the Alus.

To test this hypothesis, we analyzed the differences in GC content between the Alus and their family consensus as a function of age and isochore membership (Fig. 3). The GC-content decay of Alus is clearly appreciated in this plot. Up to evolutionary distances between 0.008 and 0.1, the slope is steeper, which corresponds to a higher rate of CpG mutations. Alus are rich in CpG dinucleotides that tend to mutate faster than the remaining sites. We call this range of evolutionary distances "the CpG domain." During the CpG domain, no influence of the isochore on the composition of the Alus can be appreciated,

370

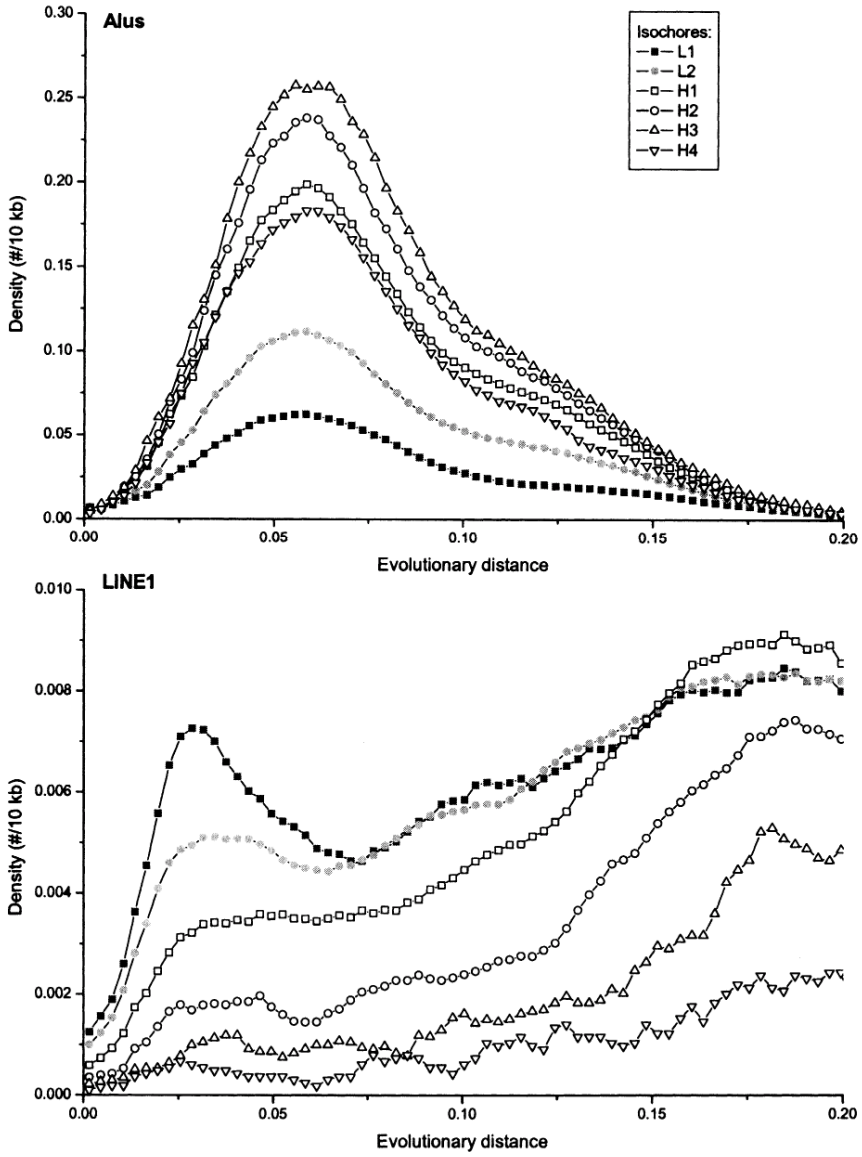

Fig. 2. Alu (**top**) and LINE1 (**bottom**) absolute densities as a function of isochore membership and evolutionary age. The LINE1 show a very pronounced density minimum in the L isochores, which happens concurrently with the massive spread of the Alus.

since the decay in Alu GC is dominated at this stage by fast mutations at the CpGs. After the CpG domain, the Alu GC decay is stronger in L than in H isochores, thus revealing a higher compositional matching pressure on Alus in L isochores.

By plotting the correlation between the isochore and the Alu GC content in different "age bins" (Fig. 4), the two stages in Alu evolution the CpG domain and the compositional matching to the isochore can be more clearly distinguished. During the CpG domain, the correlation remains low. Afterward, the correlation begins to increase and the effects of the compositional matching become appreciable. The constant increase in r^2 during this second stage indicates a time dependence of Alu compositional matching, a result that agrees with the regional mutation-pressure hypothesis (Wolfe et al. 1989), which postulates that the mutation rate varies with chromosome regions. CpG levels of Alus have been previously shown to correlate with the GC levels of the long sequences in which they are located (Jabbari and Bernardi 1998). Our results also agree with previous observations that Alus accumulate point mutations at the rate expected for unselected DNA sequences, except for the special case of CpG dinucleotides (Jurka 1995; Schmid 1998).

As mentioned above, the density shift already takes place at evolutionary distances of 0.025. This signifies that long before Alus are influenced by the isochore on their base composition, Alu density shifts toward the H isochores. This observation rules out the matching mutational pressure as the agent of the density shift.

371

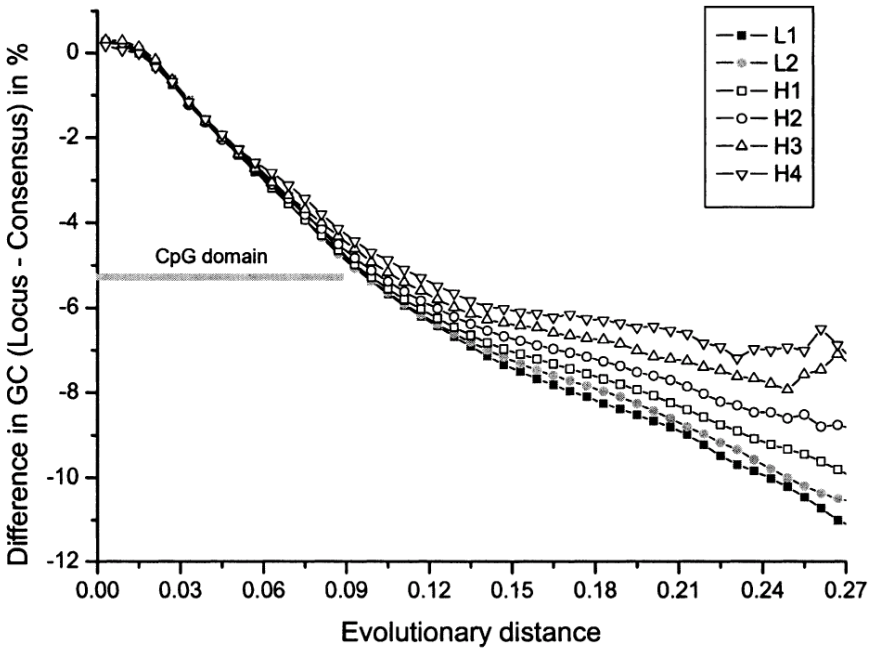

Fig. 3. Differences in GC content between the Alus and their family consensus. Two different slopes can be appreciated, corresponding to the CpG domain and the subsequent compositional matching of the Alus to the isochore.

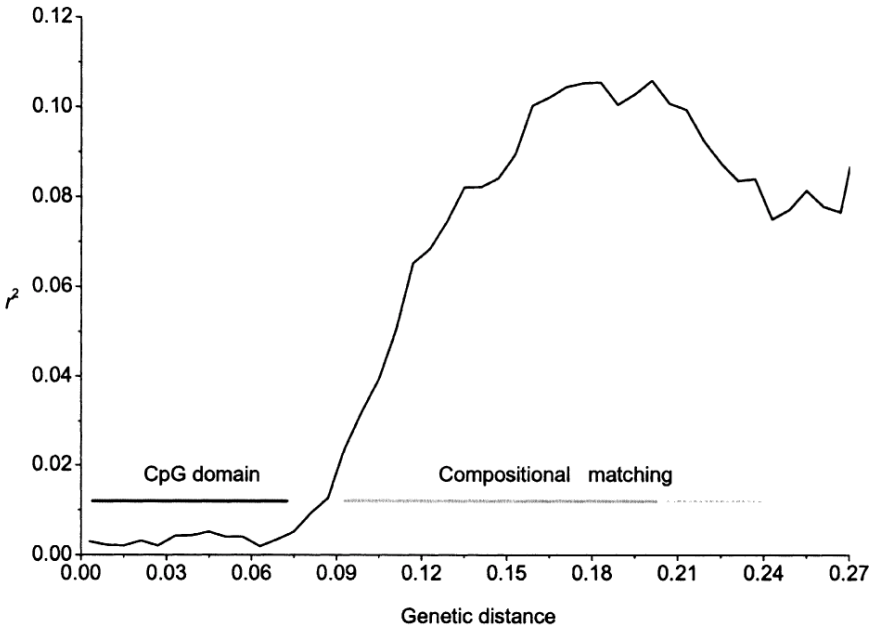

Fig. 4. Correlation between the isochore and the Alu GC content in different "age bins." The r^2 coefficient is plotted against the genetic distance.

Alu Overrepresentation in Introns Is Limited to L Isochores

Like Alus, the genes are also denser in GC-rich isochores (Zoubak et al. 1996; Bernardi 2000). By assuming a causal effect in this association, another hypothesis to explain the Alu density shift can be formulated: Alus would have followed the genes by concentrating in the GC-rich isochores, either through selective forces similar to those acting on genes (IHGSC 2001) or, more probably, through a hitchhiking effect. In fact, Alus are overrepresented in introns and scarcer in intergenic sequences (Smit 1999). The higher gene (intron) density in H isochores could then explain the higher Alu density in GC-rich isochores.

To check this hypothesis we need to compare intron vs. intergenic *Densities* in the different isochores. The Alu excess in introns was measured as $R_{Ex} = [\rho_{IV}/(\rho_{IV} + \rho_{ig})] - 0.5$, where ρ_{IV} and ρ_{ig} are intron and intergenic Alu densities, respectively. The coefficient R_{Ex} is negative if intergenic densities are higher than intron densities and positive in the opposite case.

372

Table 2. Alu densities in introns and intergenic regions of the entire genome and the different isochores

	Introns		Intergenic regions		
	N	Density	*N*	Density	Intron excess
Entire genome	550,908	4.376	431968	3.12	0.084
Isochore					
L1	30,400	2.088	66164	1.752	0.044
L2	97,372	3.405	87347	2.917	0.039
H1	141,086	5.168	101627	5.897	−0.033
H2	57,537	6.257	32569	7.561	−0.047
H3	28,055	7.389	13568	8.977	−0.049
H4	8,641	5.289	4190	7.328	−0.081

Note. Intergenic regions were taken as sequence segments beginning 2 kb downward (3′) from the stop codon of a gene and ending 2 kb upward (5′) from the initiation codon of the next gene.

Table 2 shows the Alu excess in introns we found for the entire genome and for the different isochores. When the entire genome is considered, Alus are in fact overrepresented in introns, thus confirming the observation by Smit (1999). However, when R_{Ex} is computed per isochore, we got positive values in L, but negative values in H, isochores. Thus, Alu overrepresentation is limited to the introns of L isochores. Hence, the density shift to H isochores cannot be explained by the Alu concentration in the introns of these regions. When F5′ and F3′ flanking regions of genes, instead of introns, were used in the comparison with intergenic regions, the same conclusion was reached (not shown). The observation that in genes of GC-rich isochores the Alus are underrepresented compared to intergenic Alu copies rules out the hypothesis by IHGSC (2001) that Alu accumulate in genes and thus in GC-rich regions because they have some benefit for gene expression. Brookfield (2001) already suggested that the hypothesis of IHGSC (2001) is inconsistent with our knowledge of human population genetics.

Measuring Alu Alu Recombination Activity in the Genome

Alu sequences can serve as substrates for either homologous or nonhomologous recombination events (Babcock et al. 2003; Deininger et al. 2003). The Alu repeat occurs approximately once every 3 kb in the human genome, and mispairing between such repeats, and a subsequent unequal crossing-over event between the mispaired repeats, may be a frequent cause of deletions and duplications (Fig. 5). In fact, Alu Alu recombination has been proposed to be a possible mechanism driving the spatial, postinsertional Alu dynamics in the human genome (see, e.g., Lobachev et al. 2000; Brookfield 2001; Stenger 2001; Batzer and Deininger 2002; Medstrand et al. 2002; Deininger et al. 2003). However, an additional requisite for Alu Alu recombination to provoke a net

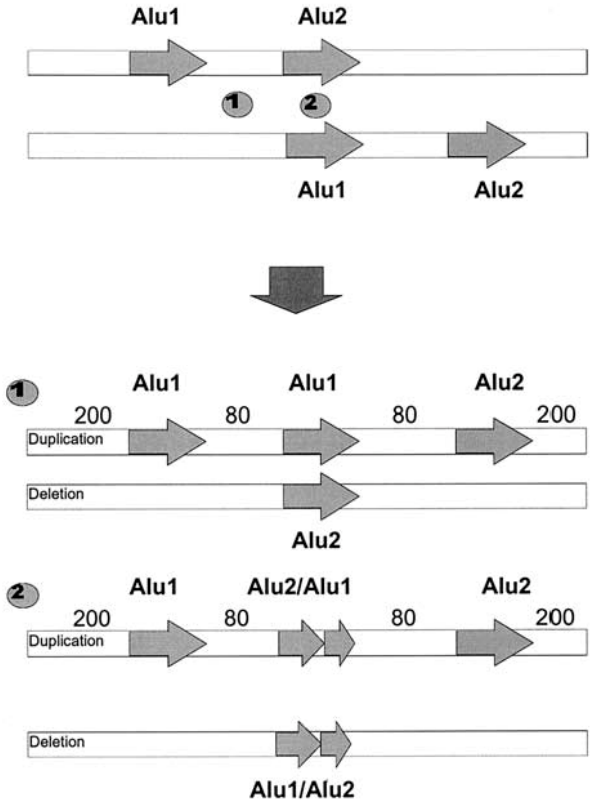

Fig. 5. Duplication products coming from the unequal homologous recombination between two directly oriented Alus. Two classes of Alu duplications are produced, depending on the specific site where the crossing-over occurs (1 or 2). Inner distances of 80 bp between the Alus involved and distances of 200 bp to the flanking Alus were used to scan the genome in search of these recombination products.

effect on Alu densities is that recombination products (deletion/duplication) must have a differential chance to survive in the distinct genomic regions.

Direct measurement of Alu Alu recombination products in the genome sequence is difficult, as deletions usually do not leave identifiable footprints. Therefore, we estimated unequal recombination activity by counting only the duplication products. In

Table 3. Frequencies of Alu trimers (derived from Alu/Alu recombination) in the different isochore classes

Isochore	No. observed	No. expected	Obs/exp ratio
L1	124	1.84	67.39
L2	426	10.43	40.84
H1	804	33.34	24.12
H2	398	15.47	25.73
H3	217	8.40	25.83
H4	64	1.43	44.76
Total	2033	70.91	28.67

Note. The expected frequencies were derived from simulation experiments (see text). The comparisons between observed and expected frequencies were all statistically significant (chi-square test, $p < 0.001$).

the simplest case, if the crossing-over takes place between the two Alus, the duplication consists of three elements arranged in tandem (trimer), while if the recombination event occurs within the Alus, the product will be a tandem of four Alus (tetramer). We focus here on the frequencies of Alu trimers, as the tetramer numbers we found in the different isochores were too small (and noisy) to allow statistical tests (see below).

To track Alu duplications in the genome, we designed several Perl scripts to parse the RepeatMasker output. To locate these recombination products, we searched for three Alus in tandem separated by less of 80 bp and with no other Alu closer than 200 bp. Additionally, we imposed the condition that two consecutive Alus in a trimer must be of the same subfamily, due to the duplication. To filter out trimers originated by insertion of Alus into other preexisting Alus, we used the ID number provided by RepeatMasker, afterward discarding trimers in which the two flanking Alus were actually the same element. In this way, we excluded most "nonrecombinogenic" trimers that did not originate through unequal Alu Alu recombination.

However, another alternative source for Alu trimers is simply chance formation, in the "normal" Alu-insertion process over evolution. To control for this possibility, we performed computer simulations by randomly inserting Alus into artificial sequences (see Data and Methods). After generating 100 simulated sequences for each isochore class, we scanned them for Alu trimers, using the same setup as for the genome scan (see above). The frequencies of Alu trimers were then averaged for the 100 simulation runs for each isochore. We consider these values to be the expected chance frequencies.

Table 3 shows that the observed frequencies of Alu trimers in the genome were far higher than those expected to arise merely by chance. Furthermore, the observed/expected ratios were higher in L isochores, thus pointing to a higher rate of unequal recombination, or a higher survival rate of recombination products, in these isochores. This suggests that the other recombination product (Alu deletion, which we cannot detect directly) should also be more frequent in L isochores.

The total number of trimers detected (2033) may appear too low to explain the dramatic shift in the Alu distribution. It should be taken into account, however, that we are detecting probably only one part of the recombination events. There may be a number of reasons for this underestimation. First, we imposed restrictive conditions on our genome scan (see above) to exclude most of the nonrecombinogenic Alu trimers. Second, the products of successive recombination events involving trimers cannot be detected with the setup we are using. Lastly, other recombination mechanisms, not quantified in this paper, may be also operating, in particular, those nonconservative ones preferentially producing deletions. For example, the intrachromatid single-strand annealing mechanism produces one genomic deletion and one episomal DNA that is lost before or during the next cell division.

Even so, the relatively higher number of recombination products we detected in L isochores is not enough to change the Alu densities in human isochores. It is necessary, furthermore, for each recombination product to have a differential survival probability. It seems that this would be the case. The analysis of human genetic disorders caused by Alu Alu recombination has shown that deletions are by far more abundant than duplications, despite that even in-frame duplications/deletions can cause disease (Deininger and Batzer 1999; Kolomietz et al. 2002). Along the same line, experimental evidence for recombination in mammalian cells strengthens the idea of the increased frequency of deletions compared to duplications (Lambert et al. 1999). If this were the general rule, the higher level of unequal Alu Alu recombination, together with the higher selective tolerance toward deletions in L isochores, could have promoted the preferential removal of Alus from GC-poor isochores and, therefore, their relative increase in the GC-rich genome regions (density shift).

If deletions by unequal recombination are more frequent in AT-rich DNA, all other repeats should also (though slowly) accumulate in GC-rich DNA. However, the opposite is true; with the exception of Alu and MIR SINEs, all other repeats are more dense in AT-rich DNA. Nevertheless, when the density ratios were plotted against the evolutionary age of LINE1 elements (Fig. 1b), an accumulation trend of the older LINE1 elements in the GC-rich genome regions can be appreciated. A comparison between young and old LINE1s in Pavliček (2001) also suggested this trend.

374

Table 4. Average genetic distances (±SD), determined with the Tamura–Nei method, for the AluYa5, AluYb8, and AluY subfamilies and the AluJ subgroup in the different isochores

Isochore	AluYa5	AluYb8	AluY	AluJ
L1	0.010 ± 0.012	0.016 ± 0.023	0.039 ± 0.023	0.143 ± 0.042
L2	0.010 ± 0.014	0.019 ± 0.032	0.038 ± 0.025	0.133 ± 0.040
H1	0.015 ± 0.030	0.029 ± 0.051	0.038 ± 0.023	0.129 ± 0.042
H2	0.018 ± 0.034	0.031 ± 0.046	0.038 ± 0.026	0.127 ± 0.046
H3	0.015 ± 0.015	0.041 ± 0.049	0.039 ± 0.032	0.129 ± 0.053
H4	0.019 ± 0.033	0.032 ± 0.032	0.039 ± 0.032	0.129 ± 0.055

A Role for Alu Alu Recombination: Two Independent Pieces of Evidence

The frequency of unequal homologous recombination between repeated Alu elements depends on both sequence similarity and physical distance on the chromosome between the two repeats involved. High sequence similarity and/or short physical distance should enhance recombination rates (Lobachev et al. 2000; Medstrand et al. 2002). On the contrary, a low level of sequence similarity or a large physical distance should lower the recombination rate. We took advantage of these two relationships and collected two sets of observations that support the implication of unequal recombination in the preferential Alu removal from L isochores. Below, these observations are described in detail.

Recently Inserted Alu Subfamilies Are Younger in L Than in H Isochores. New Alus are preferentially inserted in L isochores. If, likewise, the Alus were more frequently removed from these regions, a high turnover rate of Alu elements should be expected in these isochores. New inserted Alus are probably mutation-free copies of the master genes (Deininger et al. 1992). Therefore, due to the constant inflow of new, mutation-free Alus and the removal of other Alus, many of which have probably accumulated mutations, the compositional evolution of Alus in L isochores should become scaled down compared to those in H isochores. If so, we can formulate a first testable prediction: Alus should appear "younger" in L than in H isochores.

In addition, if the mechanism for Alu removal were Alu Alu recombination, the rate of which directly depends on sequence similarity, we can add that the rejuvenation effect in L isochores should be stronger for members of the recently inserted subfamilies (more similar) than for the older ones (less similar). Summarizing both premises, we can formulate our hypothesis in the following way: Members of recently inserted Alu subfamilies will appear "younger" in L than in H isochores. This hypothesis may at first seem counterintuitive, since, given the Alu GC-richness, a faster rate of molecular evolution (compositional matching), and therefore faster aging, can be expected for Alus in L isochores, compared to those located on H isochores.

Table 5. Correlation between genetic distance (Tamura–Nei method) and isochore GC content in the AluYa5, AluYb8, and AluY subfamilies and the AluJ subgroup

	N	*r*	*p*	Slope
AluYa5	1836	0.14	$<10^{-6}$	35.32
AluYb8	1648	0.21	$<10^{-6}$	27.81
AluY	5610	−0.01	0.31	−7.46
AluJ	5689	−0.22	$<10^{-6}$	−64.29

To discriminate between these two opposing effects, we computed the average evolutionary distances in Alu subfamilies of different ages (Table 4). We also plotted the evolutionary distances for each Alu against the isochore GC content, then computed the regression coefficients (Table 5). We observed a positive correlation between evolutionary distance and isochore GC content in young, recently inserted Alu subfamilies (AluYa5 and AluYb8, inserted 3 4 MYA [see Mighell et al. I997]), but no correlation (AluY, inserted 20 MYA) or a negative correlation (AluJb, inserted 50 80 MYA) was found in the older ones. In this way, as predicted by our hypothesis, the rejuvenation effect in L isochores is limited to only recently inserted Alu subfamilies.

This age dependence of the rejuvenation process might point to recombination as the mechanism provoking the Alu removal from L isochores. It is known that the efficiency of recombination is directly related to the length of uninterrupted regions of nucleotide identity, with higher rates of recombination being associated with longer stretches of nucleotide identity (Waldman and Liskay 1988; Baker et al. 1996). In this way, Alu Alu recombination should depend on the pairwise similarity between the sequences involved, with older Alu elements being less similar and less prone to recombine than younger, more similar Alu insertions (Batzer and Deininger 2002). Therefore, unlike any other evolutionary mechanism, recombination should act especially upon young, more similar elements, and its effects should decline sharply with growing sequence divergence or age of the elements.

In this way, the opposite correlations we observed between evolutionary distance and isochore GC content for Alu members of different ages (Tables 3 and 4) can be readily explained through the action of recombination. In recently inserted Alu subfamilies, such as AluYa5 or AluYb8, the average identity between repeats, and therefore the rate of Alu Alu recombination, should be high, and a strong turnover of repeats should be expected. The consequence is that these Alus should appear younger in L than in H isochores, and a positive correlation between evolutionary distance and isochore GC content is then expected. Conversely, in the older Alu subfamilies, as AluJb, the average identity between repeats, and therefore the rate of Alu Alu recombination should be low, and no turnover should take place. Therefore, Alu evolution would be driven only by compositional matching pressure, and a negative correlation between evolutionary distance and isochore GC content should be expected. In subfamilies of intermediate age, such as AluY, the lack of correlation we observed may be due to equilibrium between a moderate turnover rate and the increasing pressure of compositional matching.

In summary, the rejuvenation we found in recently inserted Alu subfamilies from L isochores may be due to the preferential Alu removal by Alu Alu recombination. Therefore, unequal recombination could have contributed significantly to the Alu density shift in human isochores.

Old Stand-Alone Alus are More Frequent in L Isochores. Besides sequence similarity, unequal recombination requires a physical neighborhood on the chromosome between the two repeats involved. Closely spaced elements are more prone to provoke unequal recombination than the more distant ones. It is known that Alu elements are not uniformly distributed and tend to cluster with each other (Pavliček et al. 2001; Jurka et al. 2002, 2004). However, other retroposed elements exist outside the clusters. By measuring the densities of these "stand-alone" Alus, we found significant variations among different isochores, which may again suggest a role for unequal recombination in the preferential removal of Alus from L isochores.

For each Alu, we define a "nearest-neighbor distance" (NND) as the length (bp) separating this particular Alu element from the nearest one (in either the 5′ or the 3′ direction in the two DNA strands). We then considered "stand-alone" Alus as those showing NND > 2 kb (from this distance on, the Alus show a density maximum in L isochores). We found a total of 127,511 stand-alone Alus. Their densities in the different isochores are shown in Fig. 6. Stand-alone Alus are clearly abundant in L isochores and scarce in H isochores, a result consistent with the higher clustering found by Jurka et al. (2004) in GC-rich regions.

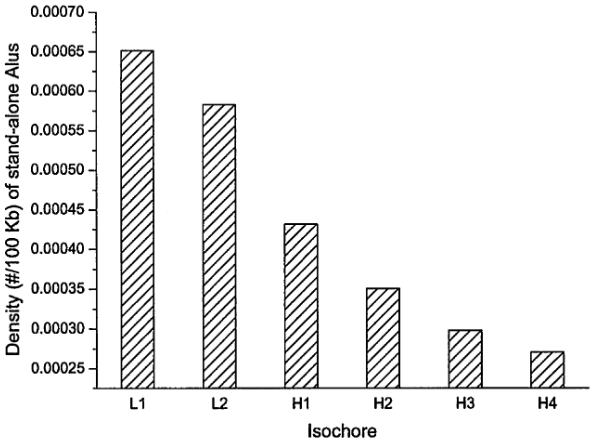

Fig. 6. Densities (number/100 kb) of stand-alone Alus (NND > 2 kb) in the different isochores. Most (81%) of these stand-alone Alu copies are from older AluS and AluJ families.

Jurka et al. (2004) observed, furthermore, that recently retroposed elements are likely to be inserted outside the existing clusters, which should lead to an overrepresentation of young members among the stand-alone class. We confirmed such overrepresentation, as 19% of stand-alone Alus are members of the youngest group AluY (this group represents 10% of all Alu copies in the genome). However, the proposal of these authors that stand-alone elements appear to be rapidly eliminated from the genome was not supported by our genome scan. We observed that 81% of stand-alone Alu copies are from older AluS and AluJ families, thus indicating that many old elements are able to survive outside the clusters.

We interpret the high frequencies of stand-alone Alus in L isochores as a consequence of the preferential elimination of closely spaced Alus by unequal recombination. Only the Alus that are neighbors on the chromosome seem to take part in unequal recombination events, the isolated, stand-alone ones remaining untouched for long periods in the genome. In this way, the higher frequencies of old stand-alone Alus we detected in L isochores may reflect the higher activity of Alu Alu recombination eliminating closely spaced Alus from these GC-poor isochores, thereby contributing to the density shift of Alus toward GC-rich isochores.

Conclusions

The genome-wide analysis of Alu densities we performed allowed us to rule out Alu/LINE1 competition for retrotransposase mediating the initial insertion process, compositional matching pressure,

and Alu overrepresentation in introns as mechanisms driving the Alu density shift in the human genome. However, the higher level of unequal recombination we detected in L isochores points to Alu Alu recombination, coupled with the differential selective disadvantage of their recombination products, as the main agent provoking the Alu density shift in human isochores. Two independent pieces of evidence the lower evolutionary divergence shown by recently inserted Alu subfamilies and the higher frequency of old stand-alone Alus in L isochores support such a conclusion.

However, other recombination mechanisms as well as other evolutionary factors are probably involved. In particular, the contractions in population size or bottlenecks during primate speciation events (Hedges et al. 2004) may have been an important helper mechanism for Alu expansions. For example, the evolutionary distance of 0.025 at which the last density shift occurred (see Fig. 1) corresponds to 6.4 MY, which is also the date of the putative bottleneck during the human/chimpanzee split. It may be, therefore, that the Alu expansion around the speciation event was accelerated by such a population bottleneck, thus contributing to the fast accumulation of Alus in GC-rich isochores.

Acknowledgments. Helpful comments from A. Marín, J.P. Martínez-Camacho, M. Ruiz-Rejón, and two anonymous reviewers are greatly appreciated. We are also grateful to A. Smit for providing the RepeatMasker computer program. This work was supported by the Spanish Government (Grants BIO2002-04014-C03-01/02 to J.L.O. and P.B. and BFM2002-00183 to P.C. and P.B.) and Plan Andaluz de Investigacion (CVI-162). M.H. acknowledges a predoctoral grant from the University of Granada (Spain). The help of David Nesbitt and Christopher Previti with the English version of the manuscript is also appreciated.

References

Babcock M, Pavliček A, Spiteri E, Kashork CD, Isohikhes I, Shaffer LG, Jurka J, Morrow BE (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by *Alu*-mediated recombination events during evolution. Genome Res 13:2519–2532

Baker MD, Read LR, Beatty BG, NG P (1996) Requirements for ectopic homologous recombination in mammalian somatic Cells. Mol Cell Biol 16:7122–7132

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3:1–10

Bernaola-Galván P, Román-Roldán R, Oliver JL (1996) Compositional segmentation and long-range fractal correlation in DNA sequences. Phys Rev E 53:5181–5189

Bernardi G, Olofson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17

Bernardi G (2001) Misunderstandings about isochores. Part 1. Gene 276:3–13

Boeke JD (1997) LINEs and Alus—The polyA connection. Nat Genet 16:6–7

Britten RJ (1996) DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci USA 93:9374–9377

Brookfield JF (2001) Selection on Alu sequences? Curr Biol 11:900–901

Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW (1998) Potential Alu function: Regulation of the activity of double-stranded RNA-activated kinase PKR. Mol Cell Biol 18:58–68

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67:183–193

Deininger PL, Batzer MA, Hutchison CA, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet 8:307–311

Deininger PL, Moran TV, Batzer MA, Kazazian HH Jr (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13:651–658

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35:41–48

Filipski J, Salinas J, Rodier F (1989) Chromosome localization-dependent compositional bias of point mutations in Alu repetitive sequences. J Mol Biol 206:563–566

Gu Z, Wang H, Nekrutenko A, Li WL (2000) Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. Gene 259:81–88

Hedges DJ, Callinan PA, Cordaux R, Xing J, Barnes E, Batzer MA (2004) Differential Alu mobilization and polymorphism among the human and chimpanzee lineages. Genome Res 14:1068–1075

IHGSC (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Jabbari K, Bernardi G (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 224:123–127

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44

Jurka J (1995) Origin and evolution of Alu repetitive elements. In: Maraia RJ (ed) Impact of short interspersed elements (SINEs) on the host genome. Landes, Austin, TX, pp 25–41

Jurka J (2000) Repbase Update, a database and an electronic journal of repetitive elements. Trends Genet 16:418–419

Jurka J, Krnjajic M, Kapitonov VV, Stenger JE, Kohkanyy O (2002) Active Alu elements are passed primarily through paternal germlines. Theor Popul Biol 61:519–530

Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV (2004) Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci USA 101:1268–1272

Kolomietz E, Meyn MS, Pandita A, Squire JA (2002) The role of *Alu* repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes Chromosomes Cancer 35:97–112

Lambert S, Saintigny Y, Delacote F, Amiot F, Chaput B, Lecomte M, Huck S, Bertrand P, Lopez BS (1999) Analysis of intrachromosomal homologous recombination in mammalian cell, using tandem repeat sequences. Mutat Res 433:159–168

Lev-Maor G, Sorek R, Shomron N, Ast G (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science 300:1288–1291

Li W (2001) Delineating relative homogeneous G + C domains in DNA sequences. Gene 276:57–72

Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA (2000) Related inverted Alu repeats unstable in

yeast are excluded from the human genome. EMBO J 19:3822–3833

Martínez Zapater JM, Marín A, Oliver JL (1993) Evolution of base composition in T-DNA genes from Agrobacterium. Mol Biol Evol 10:437–448

Medstrand P, van de Lagemaat LN, Mager DL (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 12:1483–1495

Mighell AJ, Markham AF, Robinson PA (1997) Alu sequences. FEES Lett 417:1–5

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Oliver JL, Marín A, Martínez Zapater JM (1990) Chloroplast genes transferred to the nuclear plant genome have adjusted to nuclear base composition and codon usage. Nucleic Acids Res 18:65–73

Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R (2001) Isochore chromosome maps of eukaryotic genomes. Gene 276:47–56

Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, Hackenberg M, Bernaola-Galván P (2002) Isochore chromosome maps of the human genome. Gene 300:117–127

Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P (2004) IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res 32(Web Server Issue):W287–W292

Paces J, Zika R, Paces V, Pavliček A, Clay O, Bernardi G (2004) Representing GC variation along eukaryotic chromosomes. Gene 333:135–141

Pavliček A, Jabbari K, Paces J, Paces V, Henjar J, Bernardi G (2001) Similar integration but different stability of Alus and LINES in the human genome. Gene 276:39–45

Pavliček A, Clay O, Bernardi G (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? FEES Lett 523:252–253

Rynditch A, Zoubak S, Tsyba L, Tryapitsina-Guley N, Bernardi G (1998) The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222:1–16

Schmid CW (1998) Does SINE evolution preclude Alu function? Nucleic Acids Res 26:4541–4550

Smit AFA (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657–663

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12:1060–1067

Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA (2001) Biased distribution of inverted and direct Alus in the human genome: Implications for insertion, exclusion, and genome stability. Genome Res 11:12–27

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Waldman AS, Liskay RM (1988) Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol Cell Biol 8:5350–5357

Wolfe K, Sharp P, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. Gene 174:95–102

F. Otras publicaciones

Hackenberg M, Carpena P, Bernaola-Galván P, Oliver JL. (en preparación) Alu clustering in the human genome.

Hackenberg M, Metzner C, Hoffmann M, Döhler GH. (2000) Subband selective disorder in a quasi-2D system and its effects on the intersubband spectrum. Physica E 7:216-219.

Metzner C, Steen C, Winkler R, **Hackenberg M**, Döhler GH. (2000) Intersubband transitions in coupled wells with disorder, Physica E 6:606-610.

Metzner C, Steen C, Hoffmann M, **Hackenberg M**, Döhler GH. (2000) Interplay of disorder and tunneling in coupled quantum well structures - Tuning the intersubband lineshape by an electric field. Physica E 7:722-725.

Metzner C, **Hackenberg M**, Doehler GH. Collective many-body effects in the intersubband absorption of localized electrons, DPG spring meeting en Münster (Germany), March 1999.

Hackenberg M, Metzner C, Hoffmann M, Doehler GH. Subband selective disorder in a quasi-2D system and its effect on the intersubband spectrum. ITQW-99 (5th International Conference on Intersubband Transitions in Quantum Wells) in Bad Ischl (Austria), September 1999.

Hackenberg M, Metzner C, Doehler GH. Effect of interface roughness on optical intraband transitions, DPG spring meeting en Regensburg (Germany), March 2000.