Pedeciba

Doctorado en Biología

Título: “Estudios composicionales en el genoma humano”

Mag. Victor Sabbia

2011

Orientador: Dr. Héctor Musto

INDICE

AGRADECIMIENTOS

Al Dr. Héctor Musto por su invalorable apoyo, dirección y amistad.

A los miembros del tribunal por sus valiosos aportes y dedicación.

A todos los integrantes del Laboratorio de Organización y Evolución del Genoma que me acompañaron durantes estos años, Yuyo, Rosina, Hugo, Viviana, Andrés.

Al área Biomatemáticas en general.

Al departamento de Bioinformática del Instituto Pasteur de Montevideo.

A mi familia, mis padres, hermanos y especialente a Margot, Ana Paula y Florencia.

A todos los compañeros de Facultad de Ciencias.

A los amigos.

Al PEDECIBA y a la ANII.

INDICE DE ABREVIATURAS

3D	estructura en 3 dimensiones
A	Adenina
AAR	Codón compuesto por Guanina en primera posición, citosina en segunda posicion y cualquier purina (A o G) en la tercera posición
ADN	Ácido desoxiribonucleico
agambi	Anopheles gambiae
AGR	Codón compuesto por Guanina en primera posición, citosina en segunda posicion y cualquier base en la tercera posición
Ala	Alanina
Arg	Arginina
ARN	Ácido ribonucleico
Asn	Asparagina
Asp	Aspartato
AT	Concentración molas de Adenina + Timina
athali	Arabidopsis thaliana
ATY	Codón compuesto por Guanina en primera posición, citosina en segunda posición y cualquier pirimidina (T o C) en la tercera posición
BAMD	3,6-bis(acetato mercurimetil)-1,4-dioxano
Bandas G	Bandas Giemsa
Bandas R	Bandas reversas
BLAST	Herramienta de búsqueda de alineamientos simples
btauru	Bos taurus
C	Citocina
C3	Citocina en tercera posición del codón
CCDS	CDS consenso
CCN	Codón compuesto por Citosina en primera y segunda posición y cualquier base en la tercera posicion
CDS	Secuencias codificantes
celegan	Caenorhabditis elegans
cfamil	Canis familiaris
CoA	Análisis de correspondencia
CodonW	Programa informático desarrollado por John Peden
CpG	dímero de Citocina y Guanina
Cys	Cisteína
dmelan	Drosophila melanogaster
drerio	Danio rerio
ENSEMBL	Sitio web ubicado en www.ensembl.org
EST	Expresión de secuencias marcadas
G	Guanina
G3	Guanina en tercera posición del codón
GC	Concentración molar de Guanina + Citocina
GC1	Concentración molar de Guanina + Citocina en la primera posición del codón

GC2	Concentración molar de Guanina + Citocina en la segunda posición del codón
GC_3	Concentración molar de Guanina + Citocina en la tercera posición del codón
GC_{Chr}	GC medio cromosómico
GC_{exons}	GC de exones
GCG25M	GC medio de los genes más sus correspondientes regiones flanqueantes 25 kb antes y después del ATG inicial y el codón de finalización, enmascarado para secuencias repetidas
GCIG25M	Contenido en GC de las regiones intergénicas sustrayendo un segmento de 25 kb en cada extremo, enmascarado para secuencias repetidas ("negativo" de GCG25M)
$GC_{introns}$	GC de intrones
GCN	Codón compuesto por Guanina en primera posición, citosina en segunda posicion y cualquier base en la tercera posición
GEO	sitio web del Omnibus de Expresión Génica
ggallu	Gallus gallus
GGN	Codon compuesto por Guanina en primera y segunda posición y cualquier base en la tercera posición
gIPT	Implementación GUI de IPT
Gli	Glicina
Glu	Ácido Glutámico
GUI	Interfaz gráfica de usuario
GUI	Interfaz gráfica de usuario
H1	Familia H1 de isocoros
H2	Familia H2 de isocoros
H3	Familia H3 de isocoros
H4	Familia H4 de isocoros
HMI	Índice armónico de masas
hsapie	Homo sapiens
HTM	Mapa del transcriptoma humano
HTML	Lenguaje de marcas de hipertexto
IHGSC	International Human Genome Sequencing Consortium
Ile	Isoleucina
IPT	Programa informático desarrollado en el laboratorio para generar perfiles de isocoros
kb	kilobases
L	Largo del superfragmento para ser considerado homogéneo
L1	Familia L1 de isocoros
L2	Familia L2 de isocoros
Lys	Lisina
Mb	mega pares de bases
mbp	mega pares de bases
mdomes	Monodelphis domestica
Met	Metionina
mmulat	Macaca mulata
mRNA	ARN mensajero

NCBI	Sitio web ubicado en www.ncbi.nlm.nih.gov
olatip	Orizia latipes
P	Valor de tolerancia para añadir fragmentos a superfragmento
Phe	Fenilalanina
PP-plot	gráfica de probabilidad para evaluar cuánto coinciden dos conjuntos de datos
Pro	Prolina
PSI-Pred	Programa informático desarrollado por Jones et al. 1999
ptrogl	Pan troglodites
R^2	Coeficiente de correlación
Ridges	regiones genómicas con aumento en la expresión de genes
rnorv	Rattus norvegicus
RSCU	uso relativo de codones sinónimos
SAGE	Analisis de series de expresiones génicas
scerev	Saccharomices cerevisiae
Ser	Serina
SINE	Elementos génicos cortos omnipresentes
soap	Protocolo de acceso a un objeto simple
SVG	Gráficos vectoriales escalables
T	Timina
Thr	Treonina
tnigro	Tetraodon nigroviridis
Trp	Triptofano
Tyr	Tirosona
UCS	Uso de codones sinónimos
V	Ventana de análisis de nucleótidos
vs	Versus
WEB	Abreviación de World Wide Web En el context significa aplicación que corre sobre internet y se ejecuta en un navegador
wIPT	Implementación web de IPT
WS	Webservices
xml	Lenguaje de marcas extensible

RESUMEN

Hace más de 30 años se postuló que el ADN de los mamíferos está compuesto por mosaicos de isocoros, los cuales se definen como regiones largas (>> 300 kb), composicionalmente homogéneas que pertenecen a un pequeño número de familias, caracterizada cada una de ellas por diferentes niveles de GC. Aunque trabajos posteriores mostraron que el nivel de heterogeneidad puede variar, los isocoros están presentes en todos los vertebrados, plantas y, probablemente, incluso en algunos organismos unicelulares (Bernardi 2004). Una de las características más interesantes de este tipo de organización genómica es la distribución desigual de genes entre las distintas familias. Por ejemplo, en el genoma humano (y, probablemente, en todos los mamíferos) los isocoros ricos en GC contienen más genes que las regiones pobres en GC, lo cual no es "esperable" dado que las regiones ricas en GC representan sólo el 12% del genoma. Además, entre otras funciones, se estableció que estas regiones ricas en GC son puntos importantes de recombinación, replican primero durante el ciclo celular y contienen más islas CpG (revisado en Bernardi, 2004). Aunque las propiedades composicionales del ADN han sido estudiadas y relacionadas con diferentes procesos nucleares, el enfoque utilizado hasta la fecha ha considerado generalmente a los cromosomas como "portadores" de los genes, sin propiedades particulares. En general, lo estudios se han centrado en las características y la distribución de isocoros en cada cromosoma, por lo que aún falta un examen sistemático de las diferencias y similitudes en la composición entre los cromosomas como entidades evolutivas, considerando aspectos estructurales y funcionales.

La determinación de perfiles composicionales genómicos es uno de los grandes desafíos en la era postgenómica. Las estructuras composicionales que han sido caracterizadas durante años a través de análisis de ultracentrifugación, están siendo seriamente cuestionadas tras el análisis in silico de las secuencias completas de genomas. A pesar de que existe software desarrollado capaz de determinar si segmentos largos composicionalmente homogéneos confirman la existencia de los isocoros, en la actualidad no existe una metodología de ventana simple capaz de caracterizar los patrones composicionales existentes en los genomas a partir de su secuencia. Es así que se propuso la generación de esta herramienta y su utilización para determinar el patrón composicional en diferentes genomas. En esta tesis se amplió un algoritmo existente capaz de aportar datos confiables que ayuden a contestar esta importante pregunta. El análisis composicional de genomas a través del algoritmo desarrollado, permitió la determinación de fragmentos largos composicionalmente homogéneos en el genoma de Homo sapiens. A su vez, este conjunto de herramientas permitió la clara identificación de dos espacios con características diferentes, el espacio

homogéneo y el heterogéneo. Se desarrolló además una herramienta nueva para navegar las principales características composicionales genoma, de los genes y sus regiones circundantes. La herramienta está basada en gráficos vectoriales escalables y permite el análisis de características composicionales y de expresión aunque podría ser adaptada fácilmente para mostrar otras características.

Se analizaron las propiedades composicionales del genoma humano, a nivel genómico y cromosómico. Hemos encontrado que cada cromosoma es único y coherente internamente, ya que las regiones cromosómicas génicas (exones, intrones, GC1, GC2 y GC3) y no génicas se correlacionan positivamente con el valor medio de GC del cromosoma y el GC de los isocoros que los contienen, aun cuando la diferencia en la media de GC de los cromosomas es superior al 10% y GC3 solo representa menos del 3% de la longitud cromosómica. Los resultados se discuten a la luz del conocimiento actual sobre los territorios cromosómicos en el núcleo y la presencia de fábricas de transcripción. El papel de la secuencia repetitiva también se considera. La imagen global nos describe un centro del núcleo donde la densidad de genes es más alta, los cromosomas son más cortos y el GC es mayor.

A fin de conocer las principales tendencias en el uso de aminoácidos del genoma humanos se realizó un análisis de correspondencia sobre el uso de los mismos en 14.815 proteínas. Se encontró que existen tres factores principales que influyen en la variabilidad de la composición aminoacídica en estas proteínas, explicando, respectivamente, 20,4%, 14,7% y 9,9% de la variabilidad total. La primera tendencia está fuertemente correlacionada con el contenido de GC de las posiciones primera y segunda del codón y también se correlacionó significativamente con el nivel de GC de las regiones flanqueantes correspondientes y los respectivos intrones. Por lo tanto, la principal fuerza modelando el uso de aminoácidos entre las proteínas humanas son las restricciones composicionales determinadas por el isocoro en el que se encuentra cada gen. La segunda tendencia se correlaciona con la hidropatía de cada proteína y con la frecuencia de hojas-ß. Por último, la tercera tendencia está fuertemente asociada con el uso de Cys y la frecuencia de a-hélices.

Se realizó un análisis de correspondencia sobre uso de codones en los genes humanos que reveló, como era de esperar, que el primer eje está fuertemente correlacionado con la composición de bases en la tercera posición sinónima del codón. Además, en un extremo sobre el segundo eje se localizaron genes con una alta frecuencia de codones NCG y CGN. La gran mayoría de estas secuencias se encontraron en islas CpG, mientras que lo contrario sucede con los genes colocados en el otro extremo. Por lo tanto, las dos principales conclusiones de este estudio son: 1) la influencia de las islas

CpG en el uso de codones, y 2) ya que el segundo eje es ortogonal (y por tanto, independiente) del primero, las islas CpG no necesariamente están relacionadas con genes que se encuentran en isocoros ricos en GC.

ISOCOROS

El genoma de los vertebrados es heterogéneo desde el punto de vista composicional. En otras palabras, es una sucesión de segmentos de ADN de más de 300 kb que internamente se caracterizan por un contenido en GC (concentración molar de Guanina + Citosina) relativamente constante. A estos segmentos se les dio el nombre de isocoros [región igual (Cuny et al., 1981)]. Dado que una región con un determinado contenido en GC está seguida por otra que cambia su composición en trechos relativamente cortos, se dice que el genoma de los vertebrados es un mosaico de isocoros (Macaya et al. 1976; Thiery et al. 1976). Esta característica fue descubierta analizando ADN bovino de alto peso molecular (Filipski et al. 1973) usando centrifugación preparativa en gradientes de densidad en CS_2SO_4 en presencia de iones de plata (Corneo et al. 1968). Posteriormente, el fraccionamiento del ADN usando ligandos más específicos, como el BAMD [3,6-bis(acetato mercurimetil)-1,4-dioxano], llevó al inesperado descubrimiento de la alta heterogeneidad composicional del ADN de alto peso molecular en varios mamíferos (no satélite, no ribosomal). Tanto los isocoros como sus propiedades básicas fueron descubiertas y descritas por el grupo liderado por Giorgio Bernardi (Filipski et al. 1973; Macaya et al. 1976; Thiery et al. 1976; Cuny et al. 1981). Los isocoros son segmentos de ADN >>300 kb, y pueden agruparse en un conjunto discreto de familias caracterizadas, precisamente, por un GC particular (Costantini et al. 2009).

Estas estructuras tienen un tamaño intermedio entre los genes (o grupos de genes) y el de las bandas cromosómicas. En el genoma humano, el rango composicional de los isocoros se ubica entre 30% y 60% de GC, y se han caracterizado 5 familias, dos de las cuales son pobres en GC, y se llaman L1 y L2, y tres familias ricas en estas bases y se las denomina H1, H2 y H3. Desde el punto de vista de la fracción del genoma que corresponde a cada una de ellas, las dos primeras representan en conjunto el 62% del genoma y las últimas tres el 22%, 9% y 3%, respectivamente. El restante 4% del genoma, está constituido por ADN ribosomal y satélite, los cuales podrían también interpretarse como isocoros por su homegeneidad composicional (Bernardi 1989) (Ver figura 1).

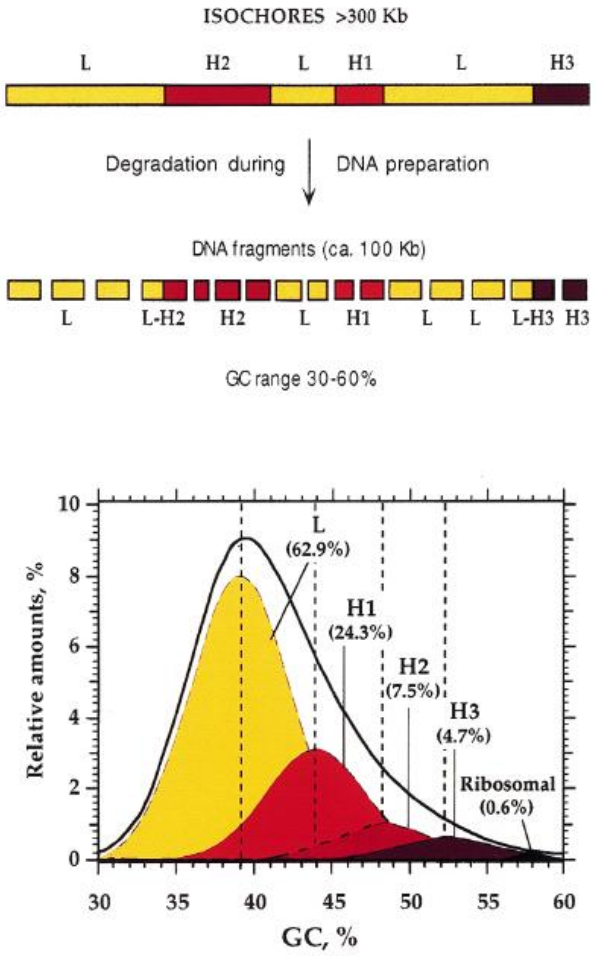
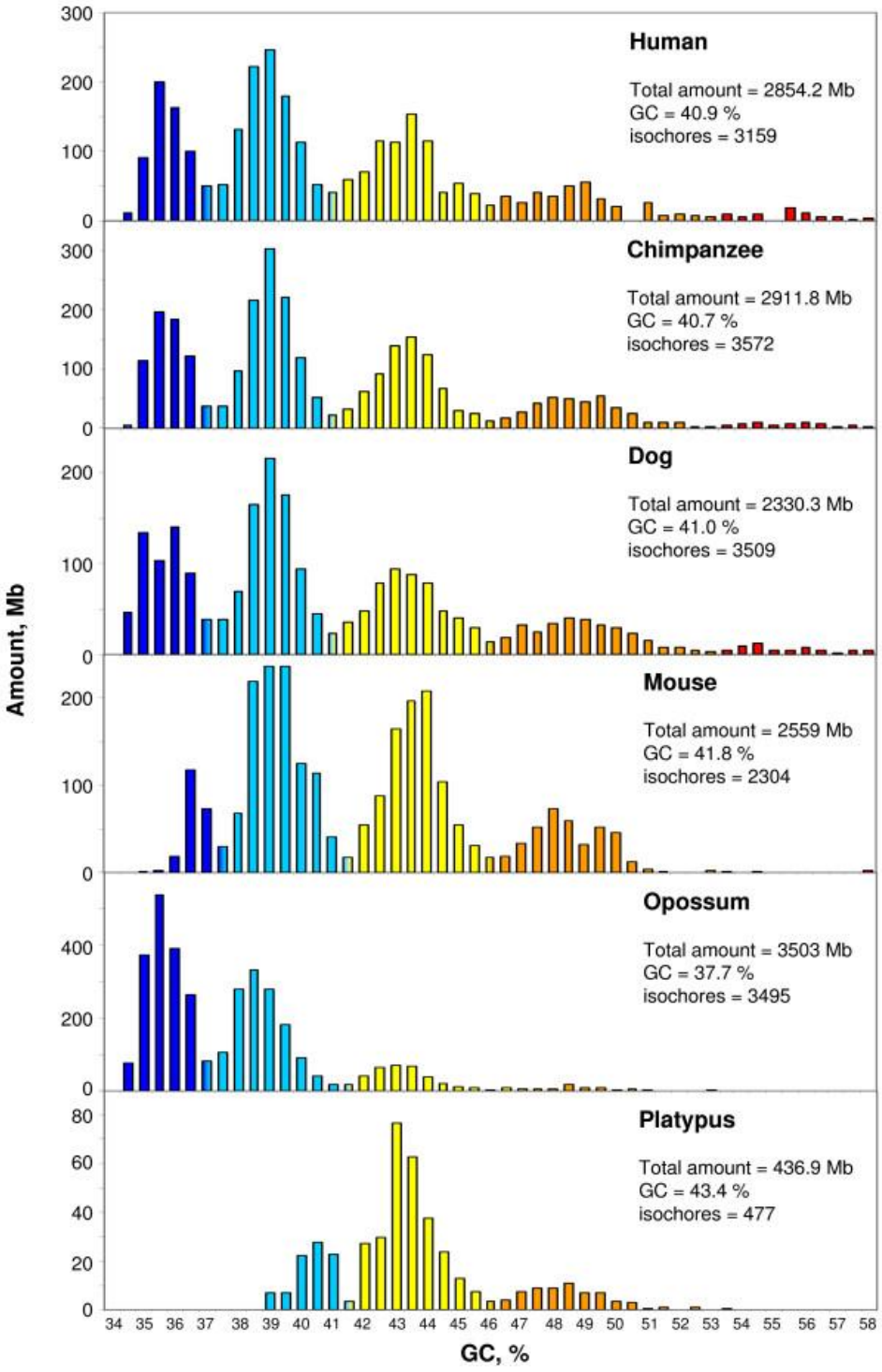


Figura 1. (Tomada de Bernadi, 2000). (Arriba) Esquema de la organización en isocoros del genoma humano. Este genoma, que presenta un perfil común en los vertebrados, es un mosaico de segmentos de ADN largos (>> 300kb en promedio) composicionalmente homogéneos y que puede separarse en familias. Los isocoros se degradan durante las preparaciones normales de ADN a fragmentos de 100kb. El rango de GC de estas estructuras en el genoma humano se encuentra entre 30% y 60% (Bernardi, 1995). (Abajo) El perfil de CsCl del ADN humano se resuelve en sus principales componentes: las familias de fragmentos de L (L1 + L2) y H (H1, H2 y H3).

El patrón general de las familias de isocoros es muy parecido entre los mamíferos y aves y se encuentra también (aunque la heterogeneidad es menor) en los vertebrados de sangre fría (ver figura 2). El contenido relativo de ADN en las diferentes familias define el patrón de isocoros de cada genoma y podría utilizarse como un sello diferencial entre los genomas de los vertebrados, por lo cual los patrones composicionales se han llamado también fenotipos genómicos. Estos difieren (más allá de la similitud general) no sólo entre Clases de vertebrados sino también entre Ordenes de una Clase y entre Familias dentro de un Orden (Bernardi 2000). El análisis de secuencias largas provenientes de bancos de datos (Ikemura y Aota 1988; Ikemura et al. 1990) y el mapeo composicional (Bernardi 1989) de bandas cromosómicas (Gardiner et al. 1990; Krane et al. 1991; Bettecken et al. 1992; Pilia et al. 1993; De Sario et al. 1996; De Sario et al. 1997) y de los cromosomas humanos completos (Costantini et al., 2009) han mostrado que los isocoros varían en su longitud entre 0,2 Mb a varias Mb (De Sario et

al. 1996; De Sario et al. 1997). En la figura siguiente (2), se muestran la distribución de los isocoros en diversos vertebrados, tanto homeotermos como poiquilotermos (Costantini et al. 2009).



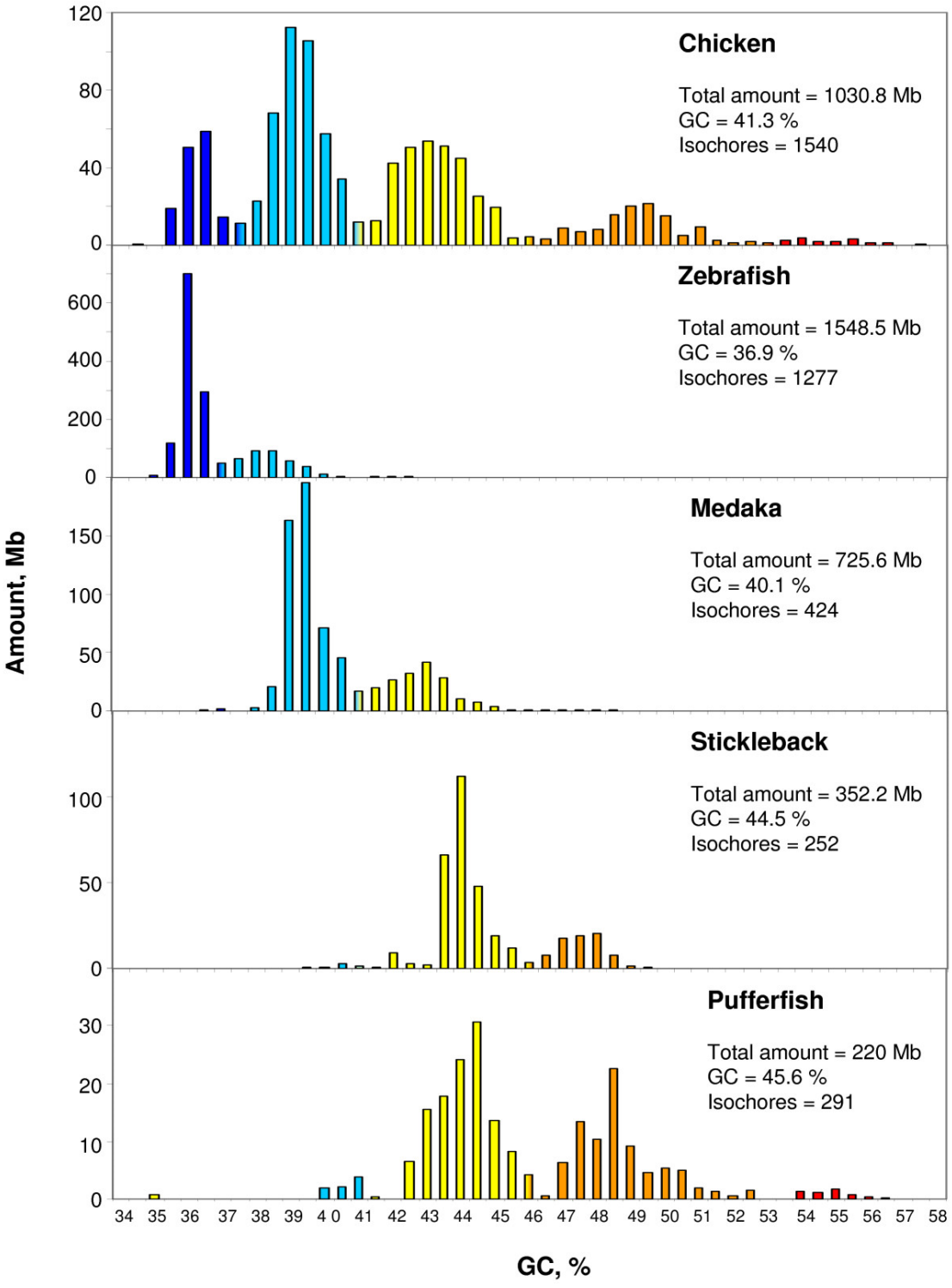


Figura 2. (Tomado de Costantini et al., 2009). Los histogramas muestran la distribución de isocoros agrupados en barras de 0,5% de GC en humano, chimpancé, perro, ratón, zarigüeya, ornitorrinco, gallina y diferentes peces. La cantidad total de secuencias se calcula a partir de la suma de isocoros; los distintos colores representan las cinco familias. Los valores en los mínimos se dividieron entre las dos familias vecinas (barras en el histograma con colores mezclados).

Se han mostrado inclusive, fronteras marcadas entre los isocoros (Fukagawa et al. 1995; Stephens et al. 1999). Estas fronteras serían esperables dado que las moléculas de ADN de diferentes isocoros pueden separarse, lo cual sería imposible si la transición entre estas estructuras fuese menos pronunciada.

A comienzos del siglo XXI, con la secuenciación del genoma humano, el International Human Genome Sequencing Consortium (IHGSC), revisó el concepto de isocoro. Tras el análisis de las secuencias primarias se llegó a la conclusión de que no podía determinarse como homogénea ninguna secuencia de 300 kb (Lander et al., 2001) cuestionándose el mérito del prefijo "iso" de los isocoros y de hecho su existencia. Tal afirmación fue reforzada por Haring y Kypr (2001), quienes no encontraron estructuras de isocoros en los cromosomas 21 y 22. En contraposición, estudios llevados a cabo por otros grupos, encuentran estructuras de este tipo a partir de datos similares (Venter et al. 2001; Pavlicek et al. 2002). Además aparecen fuertes cuestionamientos a los trabajos que niegan la presencia de isocoros en el genoma humano (Bernardi 2001; Clay y Bernardi 2002; Li et al. 2003). Las críticas atacan fundamentalmente la rigurosidad que se aplicó para determinar la homogeneidad interna que debían cumplir los isocoros, la cual no se correspondería con la que se espera encontrar en la naturaleza (Clay et al. 2001), ya que una secuencia que contiene segmentos codificantes necesariamente necesita un grado mayor de libertad para sus codones.

Como nuevas alternativas de análisis de isocoros en el genoma, se crean luego varios programas informáticos a partir de diferentes consideraciones. Así, se utiliza una técnica llamada segmentación composicional para la determinación de isocoros en el genoma humano (Oliver et al. 2001; Oliver et al. 2002) encontrándose con exactitud el límite entre dos isocoros previamente conocidos y generándose una serie de fragmentos largos que podrían asimilarse a los mismos. Este límite también fue identificado con éxito a través de la técnica de análisis de ondeado multirresolucional (Wen y Zhang 2003) asociado a otra técnica derivada del método de curva Z con la cual se identificaron fragmentos largos de ADN composicionalmente homogéneos en los genomas de humano (Zhang y Zhang 2003) y ratón (Zhang y Zhang 2004).

La distribución composicional de fragmentos mostrada en la figura 1 representa el patrón composicional que refleja el modelo de isocoros, el cual podría tomarse como un sello distintivo de cada genoma. Este patrón composicional es muy diferente en vertebrados de sangre fría o caliente. La mayor diferencia es que los primeros son mucho menos heterogéneos que los segundos y solo excepcionalmente llegan a los niveles altos de GC de éstos últimos (Costantini et al. 2009). Existen además, otras diferencias de menor entidad entre ambos tipos de genomas [para una revisión completa, ver (Bernardi 2004)].

Desde hace tiempo se han documentado correlaciones composicionales entre exones (y las distintas posiciones de los codones) y los isocoros en los cuales se encuentran; así como también entre exones e intrones correspondientes al mismo gen (Bernardi et al. 1985; Ikemura 1985; Aota y Ikemura 1986; Bernardi 1986; Bernardi 1989; Aissani et al. 1991; D'Onofrio et al. 1991; Mouchiroud et al. 1991). Estas correlaciones composicionales establecen puentes entre las secuencias codificantes y las no codificantes que las rodean. De hecho, es posible observar una correlación lineal entre los niveles de GC (y los niveles de GC_3) de las secuencias codificantes y el GC de los isocoros en los cuales se ubican estas. Si bien estas correlaciones se conocían desde hace más de dos décadas (Figura 3 a y 3 c), los datos actuales confirman el resultado (figuras 19 a y b, tabla 4).

En la figura 3 puede observarse también una correlación lineal entre el nivel de GC de las regiones codificantes y los intrones de los mismos genes (figura 3 b), confirmada con datos actuales (tabla 4). Las correlaciones de las figuras 3 a y b, son esencialmente similares en el genoma de la gallina (Musto et al. 1999) y también existen en plantas (Carels y Bernardi 2000). De hecho, las correlaciones universales entre las posiciones de los codones se han validado para una gran cantidad de genomas (Bernardi 1990; Bernardi 1991; D'Onofrio y Bernardi 1992) lo que podría indicar que las restricciones composicionales, también llamadas presiones AT o presiones GC por Jukes y Bhushan (1986), operan en la misma dirección, pero no con la misma fuerza en las tres posiciones de los codones y en las secuencias codificantes y no codificantes.

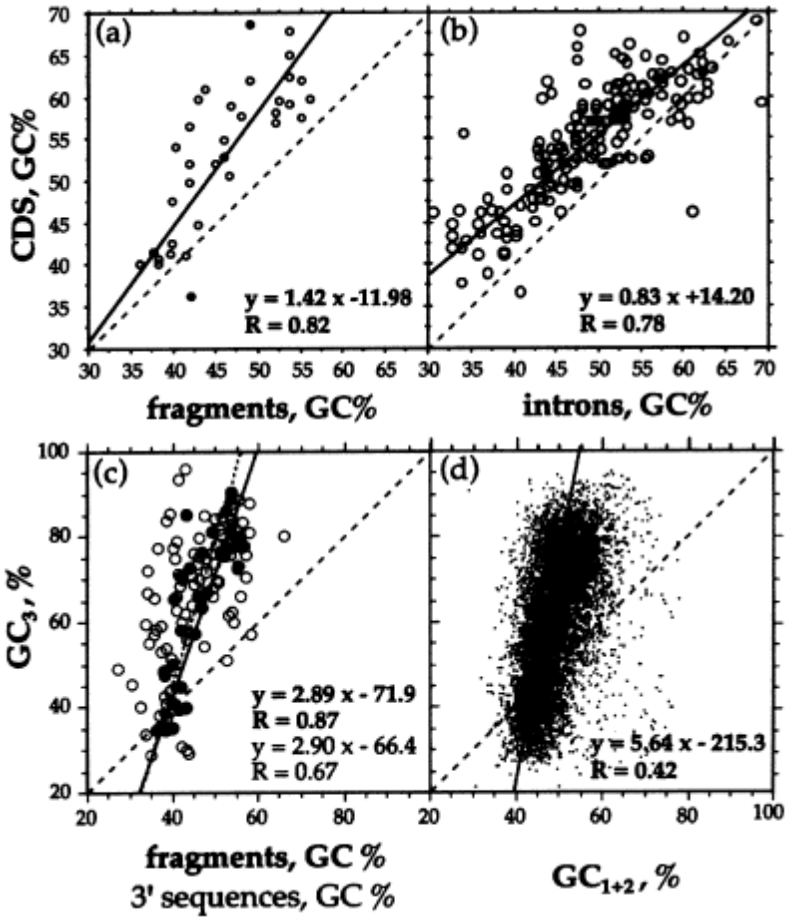


Figura 3. (Tomada de Bernardi, 2000). Correlación entre los niveles de GC de secuencias codificantes humanas y a) el nivel de GC de fragmentos largos de ADN en los cuales se localizan las secuencias, o b) el nivel de GC de los correspondientes intrones. Correlación entre el GC_3 de las secuencias codificantes humanas y c) el nivel de GC de las secuencias donde se encuentran (circulos rellenos) y secuencias flanqueantes más allá de 500 pb desde el codón stop (círculos sin relleno). Las líneas sólida y punteada son las curvas de regresión a través de los dos conjuntos de puntos. d) Correlación entre los valores de $GC_1 + GC_2$ de genes humanos.

LA DISTRIBUCIÓN GÉNICA EN EL GENOMA HUMANO

Las correlaciones composicionales que existen entre los niveles de GC en la tercera posición del codón en los genes humanos (valor promedio para cada gen), con los niveles de GC de las regiones donde los genes se encuentran ubicados, permitieron años atrás determinar y cuantificar su distribución en las diferentes familias de isocoros. Esta aproximación ha permitido encontrar que la distribución no es uniforme, ya que, mientras el 34% de los genes se encuentran en la familia L1 y L2, 38% pertenecen a H1 y H2 y 28% a H3 (Mouchiroud et al. 1991). En la actualidad, con la posibilidad de utilizar los datos del genoma humano completamente secuenciado, esos porcentajes no cambiaron significativamente (Bernardi 2004; Sabbia et al. 2009). Otra manera de ver este resultado, es analizar el número de genes por megabase, lo que se muestra en la figura 4, tomada de Bernardi (2007).

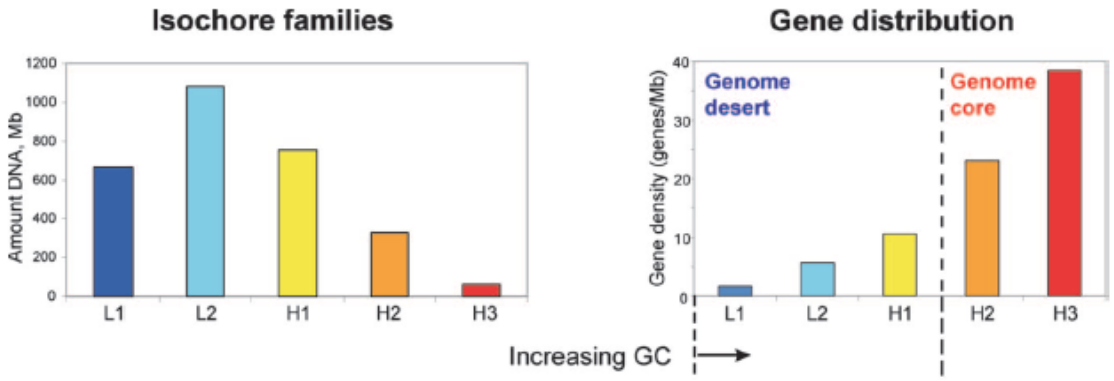


Figura 4. Distribución de ADN y genes en las familias de isocoros del genoma humano. Las principales propiedades estructurales y funcionales asociados con cada espacio génico se listan (en azul para el desierto del genoma y en rojo para el núcleo del genoma). Tomado de Bernardi (2007).

La figura muestra claramente que si se toma en cuenta la cantidad relativa de ADN en las diferentes familias de isocoros, la concentración génica en H3 es significativamente más alta que en L1 y L2, y aproximadamente contiene la misma cantidad de genes que H1 + H2 (ver la figura 4).

Dado que a) como regla general, las familias H3 y H4 son características de los vertebrados de sangre caliente (siendo H4 una familia con GC extremadamente alto y sólo existente en aves), b) que es precisamente en estas familias donde se concentran la mayoría de los genes, y c) tomando en cuenta que el número de genes no varía enormemente entre los vertebrados, se puede concluir que el aumento en la concentración de genes no fue al azar, sino que ocurrió precisamente en las regiones del genoma que se enriquecieron en GC. La línea de razonamiento actual supone que las regiones con GC más alto en vertebrados de sangre fría fueron las que se enriquecieron tanto en la línea que dio origen a mamíferos como a aves (Costantini et al. 2009). Sin embargo, análisis realizados en el presente trabajo con secuencias de genes ortólogos muestran una muy baja correlación entre el GC_3 de mamíferos y aves y aun más baja con vertebrados de sangre fría (figura 37). Como regla general, podríamos decir que la concentración de genes es baja y constante en L1 y L2 y más alta en H3. La existencia de quiebres en aproximadamente 60% de GC_3 y en 46% de GC en los isocoros define dos espacios génicos en el genoma humano (figura 5). En primer lugar tenemos el llamado "genome core" (núcleo genómico) (Bernardi et al. 1985), que estaría formado por las familias H2 y H3, donde la concentración génica es alta (un gen cada 5-15 kb), valor que es comparable con los genomas compactos de eucariotas unicelulares. En segundo lugar tenemos el llamado "empty space" (espacio vacío), formado por las familias de isocoros L y H1 donde la concentración

de genes es muy baja (un gen cada 50 – 150 kb). Se calcula que alrededor del 54% de los genes humanos se encuentran en el pequeño núcleo genómico, mientras que el restante 46% se encuentra en el gran espacio vacío.

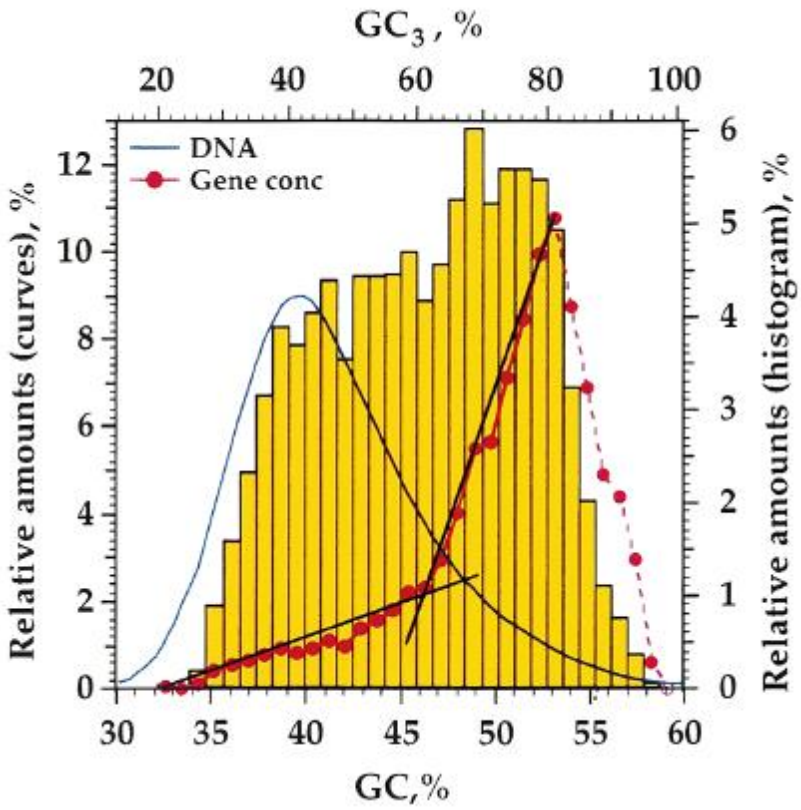


Figura 5. Perfil de concentración de genes (puntos rojos) en el genoma humano, como se obtiene de dividir el valor relativo de genes de cada intervalo de 1,5 % de GC_3 representado por las barras amarillas por la correspondiente cantidad de ADN deducida del perfil de CsCl (línea azul). La posición del histograma de GC_3 relativo al perfil de CsCl es basado en la correlación de GC_3 contra el GC de los isocoros que los contienen (tomado de Bernardi, 2001).

La familia H3 pertenece a una fracción del genoma con propiedades muy importantes, ya que no sólo tiene el nivel más alto de GC y la concentración más alta de genes sino que también presenta la concentración más alta de dímeros CpG (Bernardi 1986), que son los sitios potenciales más importantes de metilación en vertebrados, y también la concentración más alta de islas CpG (Aissani y Bernardi 1991). Estas últimas son secuencias muy ricas en GC (o sea, claramente por encima de la media genómica) caracterizadas por abundantes dímeros CpG no metilados (Bird 1986). Las islas CpG [que se encuentran en general en las regiones 5’ de los genes,

aunque pueden encontrarse también en los primeros exones e intrones (Aissani y Bernardi 1991)] se asocian preferentemente con genes que se expresan en un amplio rango de tejidos (genes "housekeeping") (Gardiner-Garden y Frommer 1987), por lo que este tipo de genes deberían ser más abundantes en H3 que en otras familias de isocoros.

Los genes ubicados en H3 tienen un mayor GC que su entorno genómico, comparados con las secuencias codificantes localizadas en otras familias (Aissani y Bernardi 1991). Además, estos genes y sus islas CpG asociadas están caracterizados por una estructura cromatínica particular, con regiones más laxas, ausencia o pobreza en histona H1, acetilación de las histonas H3 y H4 y un aumento en el espacio nucleosómico (Tazi y Bird 1990). Estas propiedades vuelven a estas regiones cromosómicas más abiertas, lo cual puede comprobarse a través de la sensibilidad al ataque con nucleasa en bandas reversas (Kerem et al. 1984).

La familia H3 tiene presumiblemente el mayor nivel de transcripción debido a su muy alta concentración de genes, especialmente genes "housekeeping". También tendría la tasa más alta de recombinación debido a su estructura caracterizada por la cromatina más abierta. Se caracteriza también por la abundancia de secuencias repetidas como Alu y minisatélites. La tasa tan alta de recombinación en H3 podría además ser responsable en gran parte de la tasa mayor de rearreglos cromosómicos y especiación mostrada por mamíferos en comparación con vertebrados de sangre fría. Existen evidencias, además, que indican que los isocoros H3 podrían actuar como regiones de integración para la mayoría de los retrovirus ricos en GC (Rynditch et al. 1991).

La familia H3 se caracteriza también por un sesgo pronunciado en el uso de codones sinónimos (UCS), y en algunos casos extremos, algunos codones están prácticamente ausentes o se usan con frecuencias extremadamente bajas debido a la muy alta concentración de GC en la tercera posición de los tripletes. Existe también una utilización extrema de aminoácidos, que favorece la presencia de los codificados por codones enriquecidos en G y/o C en las posiciones 1 y 2 del codón (D'Onofrio et al. 1991), como por ejemplo Arg, Ala, Gli y Pro, en lugar de aminoácidos correspondientes a codones con sólo A y/o T en esas posiciones (Phe, Ile, Tyr, Asn, Lys). Por último, las secuencias localizadas en la familia H3 se ubican en las bandas R de los cromosomas metafásicos (Saccone et al. 2002).

Las razones fundamentales de Bernardi para proponer el nombre núcleo del genoma a la familia de mayor contenido en GC son a) la significativa no uniformidad en la distribución de genes descritas para el genoma humano, y b) la existencia de isocoros con concentraciones muy

altas de genes como fenómeno compartido por todos los vertebrados de sangre caliente (Bernardi 2000).

ISOCOROS Y BANDAS CROMOSÓMICAS

Una serie de descubrimientos indican que los isocoros pobres en GC se localizan en las banda G, mientras que los ricos se encuentran en las bandas R en los cromosomas metafásicos humanos (Bernardi 1989). La correlación entre la heterogeneidad composicional dada por la estructura en isocoros y las bandas cromosómicas en vertebrados de sangre caliente, también explican por qué los cromosomas metafásicos de vertebrados de sangre fría, cuyos genomas se caracterizan por un bajo nivel de heterogeneidad composicional, muestran muy poco o ningún bandeado (Cuny et al. 1981).

Al menos para el caso de las bandas R, la correspondencia, sin embargo, no es directa, por la simple razón que los isocoros ricos y pobres en GC existen en una relación 1:2, mientras que las bandas G y R lo hacen en una relación 1:1.

DOS MODELOS DE EVOLUCIÓN GENÓMICA

Los patrones composicionales constatados permiten conjeturar sobre dos modos de evolución genómica: el modo conservativo y el modo transicional. El modo conservativo estaría caracterizado por la ausencia de cambios composicionales. Este modo se confirma comparando tanto fragmentos largos de ADN de > 100 kb, exones ortólogos, la composición de las tres posiciones de los codones e incluso el GC de intrones de genes ortólogos pertenecientes a genomas diferentes. En el caso de la comparación ratón-rata, la distribución composicional de fragmentos es prácticamente idéntica para las dos especies, y las diferencias en el GC_3 son menores al 1% (Bernardi et al. 1988). Sólo fueron encontradas pequeñas diferencias composicionales cuando se compararon las terceras posiciones de los codones de genes humanos y bovinos (Bernardi 2000) (figura 37). Es de destacar que las conservación composicional se mantiene a pesar de que las terceras posiciones, que cubren un rango entre 30 y 90% de GC, muestran una divergencia (distancia sinónima) promedio del 20% y pertenecen a genomas separados por 60 – 70 millones de años. La explicación más simple para este modo de evolución conservativo podría ser que, para todos los isocoros, los cambios de GC hacia AT son compensados por un número igual de cambios AT hacia GC. Esto no representaría un problema mayor para isocoros con composición cercana al 50% de GC. Sin embargo, la

conservación composicional de isocoros (y sus correspondientes secuencias codificantes) que tienen contenidos de GC extremos es incompatible con un proceso de mutación y fijación al azar, que llevaría el GC hacia ≈ 50%. Mantener esos niveles extremos requiere una fijación de las mutaciones con sesgo en direcciones opuestas para isocoros con alto y bajo GC. Este efecto podría deberse a que el sesgo en las maquinarias de transcripción, replicación y/o reparación se encuentran moduladas por estados locales de la cromatina o por la actividad transcripcional de la misma, lo que podría llevar a niveles diferenciales de síntesis de ADN reparador (Sueoka 1988). Es de notar que no existen evidencias experimentales que avalen este modelo.

Una explicación alternativa es que el modo conservativo de evolución genómica se debe a la selección negativa en contra de las desviaciones composicionales en un rango pequeño de GC (figura 6).

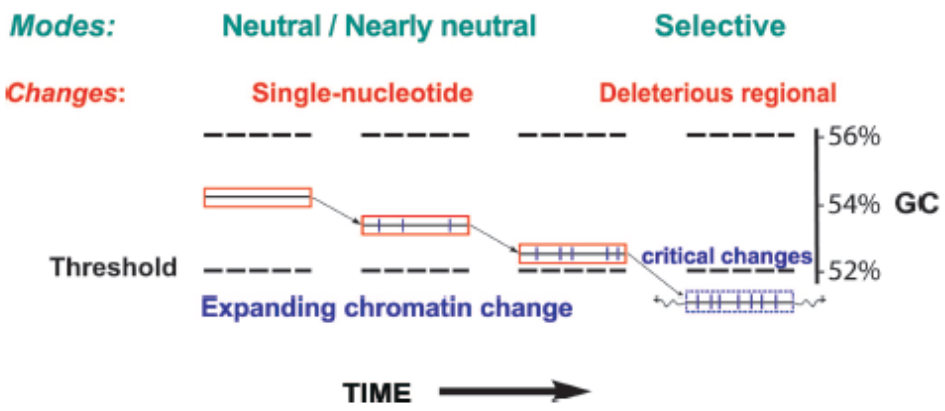


Figura 6. Evolución temporal de cambios en la composición típica de una región rica en GC para un vertebrado de sangre caliente en el modo conservador de la evolución. En una fase temprana, la media de GC de la región, inicialmente estimado como su composición óptima (fijada arbitrariamente aquí en 54%), está disminuyendo debido a la mutación con sesgo hacia AT (las barras azules verticales que cruzan la línea de ADN negro representan el exceso de cambios GC hacia AT), pero se mantiene dentro de un rango de tolerancia (cuyos umbrales arbitrarios son indicados por las lineas horizontales quebradas). En una fase tardía, la media de GC sobrepasa el umbral más bajo (fijado arbitrariamente aquí en 52%), debido a los últimos cambios, los cambios críticos. La cromatina (cajas de color rojo) luego sufre un cambio estructural (caja azul) que es perjudicial para la transcripción y replicación. Hasta entonces, los cambios pueden ser neutrales o casi neutrales. Figura tomada de Bernardi, 2007.

La selección negativa no podría, obviamente, operar a nivel de nucleótidos individuales sino a nivel regional o de isocoro (Bernardi 2004; Bernardi 2007). Algunas divergencias composicionales parecen ser toleradas, pero el nivel de umbral parecería estar muy cerca. De hecho, aún a niveles de tamaño tan pequeños como un gen, la divergencia composicional se mantiene baja en la tercera posición del codón. Quizá se dan cambios cooperativos estructurales (los cuales podrían ser comparables a cambios de fase) a nivel de isocoros más alla del nivel de umbral, y podrían tener consecuencias funcionales deletéreas, como por ejemplo en la transcripción, llevando a una menor adaptación y por lo tanto a la selección

negativa. De acuerdo a esta hipótesis, al menos algunos genes presentes en un isocoro que sufrió una deriva fuera de su nivel óptimo de GC, producirían proteínas deficientes en cantidad y/o calidad. Esta hipótesis no requiere que la selección negativa haga más de lo que normalmente se admite en relación a la mutación deletérea clásica en genes. Resulta interesante mencionar acá cómo secuencias retrovirales integradas en el genoma de mamíferos y localizadas en isocoros de GC similar se expresan y son activas, mientras que secuencias localizadas en isocoros de GC diferente no lo hacen (Rynditch et al. 1991; Zoubak et al. 1992). Esto podría sugerir un efecto del entorno cromosómico en la expresión de los genomas virales integrados.

EL MODELO TRANSICIONAL

El modo transicional o de cambio en el contenido en GC del genoma se encuentra caracterizado por cambios composicionales. Entre los genomas de vertebrados de sangre fría y caliente ocurrieron cambios composicionales mayores; y sucedieron cambios menores entre los genomas de organismos pertenecientes a diferentes Clases. Las transiciones composicionales pueden ser observadas a nivel de fragmentos de ADN, de intrones y de exones, pero se estudian mejor a través de comparaciones composicionales de las diferentes posiciones de los codones en genes ortólogos. Si se hace esta comparación entre vertebrados de sangre fría y caliente (Bernardi 1991), los niveles de GC de las secuencias codificantes de los vertebrados de sangre caliente, tienden a ser iguales o mayores que las de vertebrados de sangre fria (Figura 7), lo que provee una evidencia directa de cambios direccionales en la composición (Perrin y Bernardi 1987; Bernardi et al. 1988). Estas diferencias se espera que sean mayores en la tercera posición respecto a la 1 y la 2 (recordemos que se comparan genes ortólogos, o sea que presentan un nivel de conservación a nivel aminoacídico muy alto). Las transiciones composicionales principales que llevaron a la especiación de los mamíferos y aves, aunque similares, llevó a los genomas de los últimos a adquirir una mayor concentración de GC tanto en las secuencias de ADN no codificantes como en la tercera posición de los codones (Thiery et al. 1976; Bernardi et al. 1988). Si bien estos hallazgos fueron hechos hace más de 20 años, los resultados actuales confirman los mismos.

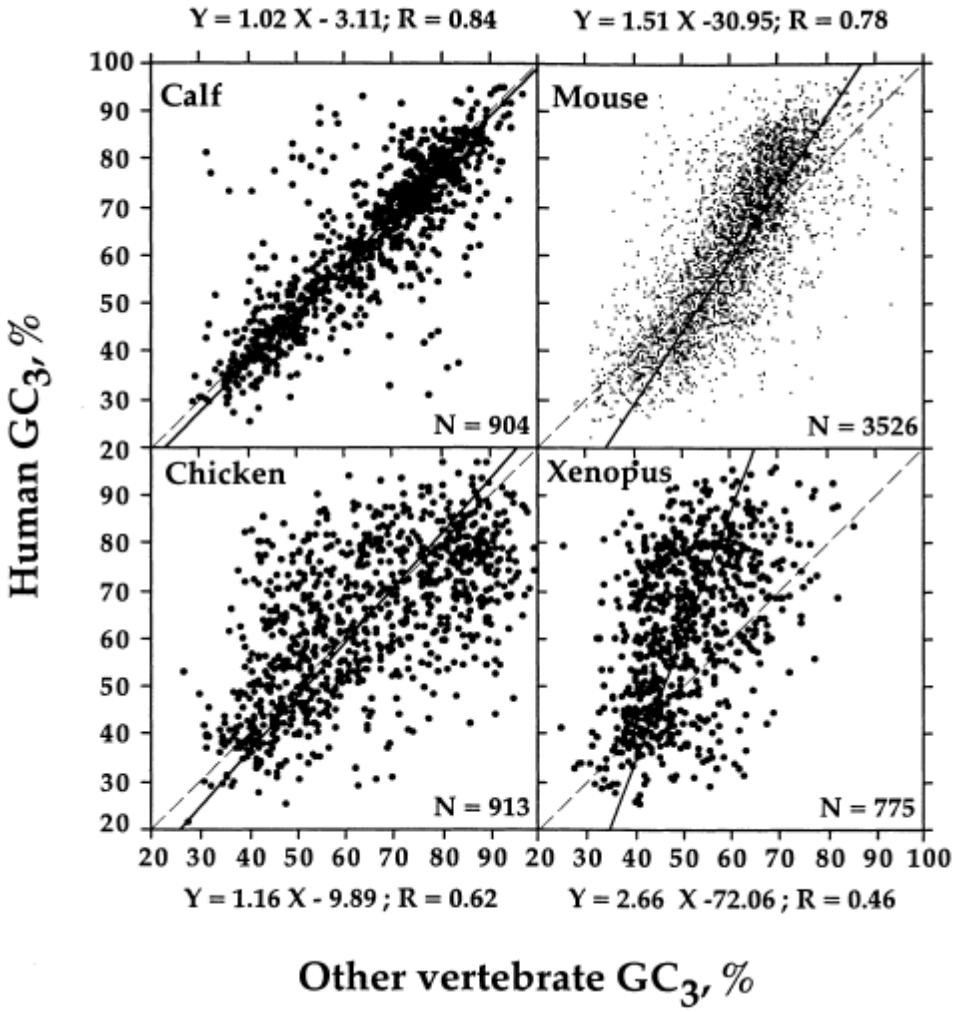


Figura 7. Correlación entre el GC_3 de genes ortólogos de humano contra a) vaca, b) ratón, c) pollo y d) *Xenopus*. N representa el número de genes ortólogos (tomada de Bernardi , 2000).

En términos de composición de bases, el genoma de los vertebrados homeotermos parecería de hecho comprender el **paleogenoma**, caracterizado por isocoros de bajo contenido en GC, *y que no ha cambiado en composición relativa a los isocoros correspondientes en los vertebrados de sangre fría*, y un **neogenoma**, caracterizado por isocoros que se han vuelto ricos en GC (Bernardi 1989). Por revisión, ver (Bernardi 2004). El incremento en GC de los vertebrados de sangre caliente afecta sólo aproximadamente una tercera parte del genoma, que contiene, sin embargo, al menos, dos terceras partes de los genes. De hecho, como ya se ha mencionado, la concentración de GC aumenta en paralelo a la concentración de genes (figura 3).

Las transiciones composicionales menores ocurrieron tanto entre vertebrados de sangre caliente como fría. En el primer caso, las transiciones separaron algunos órdenes de mamíferos y familias (Thiery et al. 1976;

Bernardi et al. 1988). Un caso especial es el de los múridos, cricétidos y espalácidos, que difieren mucho de otros roedores, así como también de la mayoría de los mamíferos investigados. Efectivamente, en esos genomas se observa una distribución más angosta a nivel de los fragmentos de ADN, lo que está acompañado por el mismo fenómeno cuando se observa la distribución de GC_3 (Salinas et al. 1986; Mouchiroud et al. 1988). Dentro de los vertebrados de sangre fría, se han observado varias transiciones composicionales (Bernardi 1990; Bernardi 2004).

TRANSICIONES COMPOSICIONALES GENÓMICAS

Se han postulado dos posibles explicaciones para las transiciones composicionales. La primera, originalmente propuesta para entender las diferentes composiciones de los genomas de procariotas, es que los patrones composicionales (o composiciones genómicas) cambian debido a mutaciones direccionales ocasionadas por un sesgo en las enzimas de replicación y reparación (Sueoka 1961; Freese y Strack 1962; Sueoka 1988; Sueoka 1992).

Esta explicación se apoya en la existencia de una cepa mutante de *Escherichia coli*, la cepa mutT (Cox y Yanofsky 1967), la cual posee una tasa de mutación 1000 veces mayor al nivel espontáneo en cepas salvajes. Además induce solamente transversiones A → C, lo cual se ha reportado que causa un aumento en el GC de 0,3% luego de 1200 – 1600 generaciones. Esta evidencia de cambio instantáneo debería, sin embargo, ser tomada con mucha cautela, dado que la diferencia en el porcentaje reportado se encuentra dentro del error del experimento. Además, las mutaciones direccionales reportadas en cepas mutantes parecen concentrarse en sitios predilectos muy limitados (Yanofsky et al. 1966; Nghiem et al. 1988; Wu et al. 1990) y es, por lo tanto, poco probable que sea debido a ellas que se produce el cambio composicional completo. En cualquier caso, para que esta explicación resultara satisfactoria, debería proveerse evidencia de que los genomas con estos cambios resultan al menos igual de aptos que los no cambiados. Esta evidencia es muy poco probable que se consiga debido al abrumante efecto del alto nivel mutacional.

La ausencia de evidencia directa no excluye *per se* la hipótesis del sesgo mutacional. Sin embargo, existen otros problemas con esta teoría. Algunos de ellos son de carácter natural, otros, se encuentran específicamente asociados a las transiciones composicionales en vertebrados. Para comenzar, la hipótesis del sesgo mutacional implica que los cambios composicionales son irrelevantes (neutrales) en lo que concierne a la organización, función y evolución del genoma y por lo tanto este se

encuentra a merced de la mutación causada por los sistemas de replicación y reparación del ADN. Sin embargo, por lo expuesto más arriba esto no puede ser fácilmente aceptado si se considera la correlación entre la composición de bases y la estructura del ADN, su importancia funcional y que los cambios composicionales en el genoma están acompañados por cambios en el uso de codones, llegando a niveles extremos de sustitución (Osawa et al. 1987) y por cambios en la composición de aminoácidos en las proteínas codificadas. Por estas razones, es difícil de aceptar la interpretación de que las diferencias en la composición de los genomas bacterianos es esencialmente debida al sesgo mutacional como fue originalmente propuesto (Freese y Strack 1962).

Otro argumento general en contra de la hipótesis del sesgo mutacional es que la distribución de los niveles de GC de genomas decrece desde bacterias hacia protistas, y luego baja hacia invertebrados y así hacia vertebrados de sangre fría y luego caliente (Bernardi 1990), indicando que la composición de bases del genoma se encuentra ciertamente no derivando en forma libre entre los seres vivos. De hecho, la composición de bases parece estar más relacionada a la variedad de entornos intra y extrecelulares del organismo en consideración. En vertebrados, cuanto mayor la homeostasis, más angosto el espectro de GC exhibido por los genomas de las diferentes especies.

Existe otro problema importante para la hipótesis del sesgo mutacional en vertebrados. De hecho, como se puede observar en la figura 6, las mutaciones de las maquinarias de replicación y reparación que llevarían a un sesgo mutacional de AT $\rightarrow$ GC, deberían haber ocurrido solamente dos veces en forma independiente en la evolución de los vertebrados, a saber, en los linajes de reptiles que dieron lugar a mamíferos y aves, pero nunca en ninguna de las otras numerosas familias de vertebrados de sangre fría, dado que ninguna otra dio lugar a genomas con las características de los de vertebrados de sangre caliente. Además, las dos series de eventos mutacionales dieron lugar solamente a cambios en los isocoros que habrían de formar el neogenoma en mamíferos y aves y no en los isocoros más abundantes ("paleogenoma") de los vertebrados de sangre caliente. Por último, de esta hipótesis se desprende que una vez alcanzado el nivel de GC de las diferentes familias de isocoros ricos en GC, éstos fueron mantenidos a través de un sesgo mutacional diferente al que les diera origen. Este escenario que involucra procesos que afectan a miles de isocoros físicamente separados dentro de cada genoma, es muy improbable que ocurriera escencialmente a través del sesgo mutacional. De hecho, la idea de que estructuras cromatínicas diferentes en diferentes regiones del genoma podrían llevar a diferentes direcciones en el sesgo mutacional, así como que las actividades transcripcionales diferenciales podrían llevar a varios niveles de síntesis de ADN reparador fue desarrollada por Sueoka (Sueoka 1988).

Sin embargo, esta teoría no explica por qué ocurrió el incremento original en GC en los ancestros de mamíferos y aves. Los problemas recién mencionados también vuelven improbable que cambios en las concentraciones de precursores de nucleótidos en la línea germinal (Wolfe et al., 1989; Eyre-Walker, 1992) pudieran ser responsables de la formación de isocoros ricos en GC en vertebrados de sangre caliente (Bernardi et al. 1988; Wolfe et al. 1989; Eyre-Walker 1992).

LA HIPÓTESIS SELECCIONISTA

La segunda explicación para las transiciones composicionales en los genomas (como la transición mayor que ocurrió en los genomas de los vertebrados) es que las mutaciones direccionales se fijan a través de la selección positiva operando a nivel de los isocoros. De acuerdo a esta explicación, el sesgo mutacional solo provee el mecanismo para el cambio composicional, y luego la selección controla los niveles composicionales de los mismos. En general, las ventajas selectivas asociadas con los patrones composicionales de los genomas pueden ser poco claras, debido a que los patrones son causados por varios factores diferentes cuyos efectos individuales son muy difícil, sino imposibles de aislar. No es posible proveer ninguna explicación para los cambios menores que ocurrieron en los genomas de vertebrados de sangre fría y caliente; aunque el patrón composicional más estrecho que muestran algunos roedores, como la rata, el ratón y el hámster, podrían corresponder a relajaciones parciales de las restricciones composicionales operando sobre los isocoros de alto contenido en GC.

Este problema puede, sin embargo solucionarse en algunas situaciones. Se podría identificar una ventaja selectiva si la misma es dominante y si puede ser evaluada contra un genoma blanco adecuado, una condición que se cumple en el caso de vertebrados. En este caso la división principal en patrones del genoma no ocurrió en un paso principal en la evolución organísmica, como la transición de anamniotes a amniotes o de peces a tetrápodos, o en forma gradual durante la evolución peces → anfibios → reptiles → vertebrados de sangre caliente, sino que ocurrió sola y precisamente en la transición de vertebrados de sangre fría hacia caliente. Esto *sugiere* que el principal factor en el cambio de patrones composicionales del genoma podría estar relacionado con la temperatura corporal. Esta sugerencia no solo provee una explicación al hueco dejado por la hipótesis de sesgo mutacional sino que además puede ser testeada.

El incremento en GC en los genomas de vertebrados de sangre caliente tiene sentido en lo que respecta a las ventajas selectivas porque determina moléculas de ADN, ARN y proteínas termodinámicamente más estables (Bernardi 1986). De hecho, el aumento en GC incrementa la

estabilidad del ADN no solamente en soluciones diluidas sino también en cromosomas como puede observarse con el bandeo R y T, dos técnicas que muestran que el ADN rico y muy rico en GC es más estable frente a la desnaturalización respecto a los ADNs de bandas G. La riqueza en GC también incrementa la estabilidad del ARN, debido a que aumenta la cantidad de estructuras secundarias que vuelven más estable a los transcriptos. Por último, aumenta la estabilidad térmica de las proteínas codificadas, ya que lleva a un incremento en aminoácidos que la acrecientan (como alanina y glicina) y un decremento de los niveles de aminoácidos que la disminuyen (como lisina y serina). Sin embargo, a esta visión clásica del grupo de Bernardi, se le oponen algunos resultados recientes de estudios a nivel de aminoácidos, que comparan genomas completamente secuenciados, tanto a nivel de procariotas como de vertebrados [ver, por ejemplo,(Wang y Lercher 2010)].

La objeción de que algunas bacterias termofílicas tienen genomas ricos en AT, no tiene relevancia para la hipótesis de la temperatura corporal, ya que las comparaciones entre genomas bacterianos termofílicos y mesofílicos no están garantizadas cuando las especies que se consideran están separadas por distancias filogenéticas enormes (por ejemplo cuando se compara Eubacteria con Arquebacteria) y cuando exhiben también enormes diferencias en la fisiología celular (Musto et al. 2005; Musto et al. 2006). Lo mismo ocurre con los genomas de organelos, en los cuales pueden predominar otras ventajas selectivas. Además la estabilidad termodinámica de los genomas podría ser debido no simplemente al aumento en GC sino además a la metilación del ADN o las interacciones ADN – proteínas.

Un argumento independiente en consonancia con esta sugerencia es la similaridad composicional entre genomas de mamíferos y aves, dos clases de vertebrados homeotermos caracterizados por tamaños genómicos distintos y que aparecieron en tiempos geológicos diferentes (> 200 y aproximadamente 150 millones de años atrás, respectivamente). Otro argumento puede ser la fuerte heterogeneidad entre isocoros de plantas originarias de climas áridos (como trigo y maíz) y la pobre heterogeneidad composicional en isocoros de plantas de climas templados (Salinas et al. 1988; Matassi et al. 1989; Montero et al. 1990). En sus patrones composicionales las primeras se parecen a los vertebrados de sangre caliente y las últimas a los de sangre fría. Bajo la hipótesis de sesgo mutacional, todas estas similaridades deben ser vistas como simples coincidencias.

En los últimos 15 años, la disponibilidad de secuencias completas de genomas reposicionó los estudios de composición en el campo de la genómica, y la variable aparentemente anticuada del contenido en GC, ha cobrado un rol protagónico en una gran cantidad de nuevas aplicaciones como por ejemplo la transferencia horizontal de genes. Al mismo tiempo, le dio un nuevo marco a los debates históricos tales como la evolución del contenido en GC [véase, por ejemplo (Hurst et al. 2004; Woodfine et al. 2004; Mijalski et al. 2005; Singer et al. 2005; Sproul et al. 2005; Costantini y Bernardi 2008).

La relación de las propiedades composicionales con procesos genéticos fundamentales como la expresión génica, la replicación y la recombinación ha sido ampliamente estudiada principalmente a nivel génico o de región (Saccone et al. 2002; Bolzer et al. 2005; Goetze et al. 2007). Por ejemplo, la relación entre la expresión de genes y las propiedades composicionales, se examinó utilizando a los primeros como unidad de análisis y volumen de expresión ("houskeeping" frente a genes tejido específicos) y/o niveles de expresión (medido por SAGE, EST, microarrays), como variables funcionales (Gierman et al. 2007). Los resultados obtenidos fueron diferentes, inclusive hasta cierto punto contradictorios, y este hecho obstaculiza una conclusión general con respecto a la relación entre la composición de nucleótidos y la expresión de genes individuales. Sin embargo, el trabajo del grupo del mapa del transcriptoma humano (HTM) identificó regiones con aumento en la expresión de genes (Ridges) en nuestro genoma (Goetze et al. 2007). Estas regiones (o dominios) de genes son densos, ricos en GC, enriquecidos en elementos SINE y también contienen genes con intrones cortos. Por otra parte, las regiones con genes débilmente expresadas, se denominan anti-Ridges, y muestran características contrarias (Saccone et al. 2002; Chakalova et al. 2005). Es importante señalar que los Rigdes han sido descritos también en el genoma del ratón (Bernardi 2004) y se comprobó que estaban relativamente conservados en comparación con el genoma humano (Costantini y Bernardi 2008). No se sabe si la expresión diferencial en Ridges y anti-Ridges se debe a la regulación de genes individuales, o si pueden ejercer un efecto adicional mecanismos en todo el dominio [para revisiones, ver, entre otros (Vinogradov 2001; Semon et al. 2005)]. Desde una perspectiva estructural, los Ridges son en general menos condensados, de forma más irregular y más cercanos al centro nuclear que los anti-Ridges (Bolzer et al. 2005). Todas estas características están de acuerdo con trabajos anteriores de Bernardi y sus colegas que mostraron que en los mamíferos, los isocoros ricos en GC contienen un número significativamente mayor de genes que las regiones más pobres en GC (Bernardi 2004; 2007). Por otro lado, el ciclo de replicación también se asoció con la heterogeneidad composional, ya que las regiones ricas en GC tienden a replicarse antes durante la fase S [para revisiones recientes ver (Costantini y Bernardi 2008;

Hiratani et al. 2009)]. Además, las tasas de recombinación también se encontraron aumentadas en regiones ricas en GC del genoma humano (International HapMap Consortium 2007).

Los cromosomas individuales no se encuentran organizados al azar en el núcleo, sino que ocupan zonas discretas y parecen ser casi inmóviles durante la mayor parte de la interfase. La organización espacial de los cromosomas en territorios cromosómicos constituye una propiedad básica de la arquitectura nuclear (Cremer y Cremer 2010). Las disposiciones de orden superior de la cromatina pueden, en primer lugar, reflejar las limitaciones geométricas, que obviamente, afectan la distribución 3D de territorios cromosómicos mayores y menores comprimidos en el espacio nuclear (Neusser et al. 2007). Ocasionalmente, sin embargo, patrones de proximidad reunidos por azar, pueden haber proporcionado propiedades funcionales ventajosas y en consecuencia se vieron favorecidos por la selección natural. La búsqueda de ensamblajes de cromatina no aleatoria, los mecanismos responsables de su formación y sus implicaciones funcionales son uno de los principales objetivos de investigación de la arquitectura nuclear. Esta búsqueda se encuentra todavía en sus comienzos (Cremer y Cremer 2010). Aunque los cromosomas son relativamente estables, los dominios de la cromatina están sujetos al movimiento browniano y pueden extenderse mucho más allá de los bordes del territorio de su cromosoma. Estos dominios cromosómicos pueden ejercer una influencia general de activación o resultar atenuantes en los genes embebidos en ellos. Los dominios de la cromatina que se replican juntos continúan asociados en todo el ciclo celular. Es razonable suponer la participación de las interacciones entre dominios de la cromatina nuclear y las estructuras subyacentes, como la lámina nuclear, nucleolo y la matriz nuclear (Goetze et al. 2007). Los dominios densos en genes están colocados hacia el interior del núcleo (Saccone et al. 2002; Bolzer et al. 2005; Goetze et al. 2007). Goetze et al. (2007a y b) analizaron las propiedades tridimensionales (3D) específicas de dominios Ridge y anti-Ridge en diferentes tipos de células y observaron que los primeros son menos condensados y están más profundamente situados en el núcleo en comparación con los segundos, independientemente del tipo celular.

De lo dicho anteriormente, se desprende que el contenido en GC, la densidad de genes, la posición nuclear y la actividad transcripcional se correlacionan con el ciclo de replicación. Las RNA polimerasas II (Pol II) no se distribuyen homogéneamente en todo el nucleoplasma, sino que se concentran principalmente en focos altamente enriquecidos, conocidos como fábricas de transcripción (Cseresnyes et al. 2009). La transcripción de genes se produce en impulsos a determinadas frecuencias, y la modulación de la expresión puede producirse por cambios en la probabilidad de que un gen se transcriba. Es interesante observar que los transgenes idénticos integrados en diferentes regiones cromosómicas llegan a adquirir niveles de expresión

que se correlacionan fuertemente con los niveles de expresión de las zonas de la integración (Gierman et al. 2007). Los genes activos tienden a replicarse a comienzos de la fase S mientras que los genes inactivos tienden a hacerlo más tarde (Chakalova et al. 2005). Sin embargo, muchos genes altamente expresados se replican hacia el final de la fase S y muchos genes inactivos se replican en forma temprana. La disrupción y ensamblaje de la cromatina que se produce durante la replicación y transcripción podría proporcionar una ventana de oportunidad para alterar las modificaciones epigenéticas y el estado en la expresión de los genes.

Aunque las propiedades composicionales del ADN han sido estudiadas y relacionadas con diferentes procesos nucleares, el enfoque utilizado hasta la fecha ha considerado generalmente a los cromosomas como simples andamios del ADN sin propiedades particulares. En general, estos análisis muestran las diferencias en GC entre los cromosomas y se han centrado en las características y la distribución de isocoros. Por otra parte, aún falta un examen sistemático de las diferencias y similitudes en la composición entre los cromosomas como entidades evolutivas, considerando aspectos estructurales y funcionales. Es conocido que los cromosomas se diferencian en varias propiedades composicionales ya que poseen diferentes tamaños y formas (en relación con la posición del centrómero), muestran diferentes perfiles de composición, se encuentran en diferentes partes del núcleo, muestran diferentes densidades de genes, niveles de transcripción, ciclo de replicación, tasas de recombinación. Sin embargo no se ha conducido a la fecha un análisis comparativo de las propiedades composicionales y de sus conexiones con aspectos funcionales tomando en cuenta todo el cromosoma.

EL USO DE CODONES

El código genético es, con sólo algunas excepciones, el mismo en todas las especies. Por otra parte, con exclusión de la Met y el Trp, todos los aminoácidos son codificados por más de un codón. Sin embargo, los tripletes sinónimos no son utilizados al azar, incluidos los codones 'stop' (Grantham et al. 1980). Para los organismos unicelulares el patrón general (el uso de codones de cada especie cuando todos los genes se consideran conjuntamente) se acepta como el resultado del sesgo mutacional (que puede ser hacia GC o AT) y la selección natural actuando a nivel de la traducción. Por ejemplo, en *Escherichia coli*, *Bacillus subtilis* y *Saccharomyces cerevisiae*, el uso de codones sinónimos está claramente relacionado con la abundancia relativa de los ARNt isoaceptores, y la correlación es muy fuerte para las secuencias altamente expresadas, que tienden a utilizar con mayor frecuencia los llamados codones mayores. Por el contrario, los genes de baja expresión tienden a mostrar un uso más aleatorio de los codones sinónimos. Además, entre las especies con composición genómica muy sesgada las opciones de uso de codones

sinónimos parecen ser determinadas principalmente (o exclusivamente) por las presiones mutacionales; aunque se han observado excepciones en organismos altamente sesgados en su composición como *Plasmodium falciparum* (Musto et al. 1999; Peixoto et al. 2004) o *Entamoeba histolytica* (Romero et al. 2000). En otras palabras, el uso de codones alternativos puede ser modelado por sesgos en las tasas de mutación entre las bases, por la selección natural en el uso de codones traduccionalmente óptimos para maximizar la tasa de eficiencia de la traducción, o por una combinación de estos dos procesos [para revisiones, ver (Sharp y Matassi 1994; Sharp et al. 1995; Akashi y Eyre-Walker 1998; Akashi 2001; Ermolaeva 2001)].

Entre las especies multicelulares, se encuentran diferentes patrones. Por ejemplo, en *Caenorhabditis elegans* y *Drosophila melanogaster*, que se caracterizan por una amplia variación en el uso de codones, los factores que lo determinan son casi idénticos a los de las especies unicelulares antes mencionadas. De hecho, se ha demostrado que en estos dos organismos existe un subconjunto de codones que son utilizados preferentemente en los genes que muestran un fuerte sesgo (Shields et al. 1988; Stenico et al. 1994). Además, Duret y Mouchiroud (Duret y Mouchiroud 1999) a través del análisis de ESTs demostraron que la frecuencia de los codones preferidos está positivamente correlacionada con el nivel de expresión génica. La selección de sitios silenciosos a nivel de la traducción también se ha reportado como uno de los principales factores que determinan el uso de codones en las plantas *Zea mays* y *Arabidopsis thaliana* (Fennoy y Bailey-Serres 1993; Chiapello et al. 1998; Duret y Mouchiroud 1999).

Entre los vertebrados parecen ser evidentes dos patrones. En primer lugar, en los vertebrados de sangre fría, donde, como se ha dicho, la heterogeneidad intragenómica en el contenido en GC es baja, se ha demostrado que, aunque el principal factor que determina la preferencia de los tripletes es la media para cada gen del contenido en GC de la tercera posición de los codones (GC_3), el segundo factor principal es la selección natural para maximizar la tasa de traducción. Esto se determinó a partir de la observación de que las secuencias altamente expresadas muestran un patrón diferente de uso de codones respecto a los genes de baja expresión. Este efecto ha sido detectado en *Xenopus laevis* (Musto et al. 2001) y en peces de la familia Cyprinidae (Romero et al. 2003). Es interesante observar que en los peces, varios de los supuestos codones óptimos para la traducción son los mismos que en el sapo, a pesar del largo tiempo transcurrido desde la separación de estas especies desde su último ancestro común (Romero et al. 2003). Una situación diferente se encuentra entre los vertebrados de sangre caliente, donde el genoma es composicionalmente muy heterogéneo (Bernardi 2000). Esta compartimentación composicional tiene una gran influencia en el uso de codones, la que es evidenciada por las fuertes correlaciones que existen entre los valores medios del contenido en

GC de las diferentes posiciones de los codones de cada gen y los isocoros en el que los mismos están insertos (Mouchiroud et al. 1991). Esto lleva, como se ha señalado, a que un gen situado en un isocoro pobre en GC mostrará valores de GC_3 (y con menos fuerza, también en GC_2 y GC_1) bajos (y, por consiguiente, un uso de codones sesgado hacia A y T), mientras que lo opuesto es cierto en el caso de un gen ubicado en un isocoro rico en GC. Además, para estas especies se ha sugerido que algunos sitios sinónimos están bajo presión selectiva, probablemente debido a las restricciones que actúan en elementos reguladores incorporados en los exones (Blencowe 2000; Hurst y Pal 2001; Willie y Majewski 2004). Por último, existen algunos indicios que sugieren que la selección en la traducción podría influir en el uso de codones en el genoma humano, aunque su contribución seria, sin duda, muy baja (Urrutia y Hurst 2003; Comeron 2004; Plotkin et al. 2004).

USO DE AMINOÁCIDOS EN EL GENOMA HUMANO

En las diferentes especies los aminoácidos no se utilizan con la misma frecuencia. Esto no se debe solamente a las diferencias físicoquímicas entre ellos, sino también a varios otros factores. Usando análisis multivariados varios artículos han demostrado que cuando diversos procariotas se estudian juntos (análisis intergenómico) el nivel de GC de las especies es la primera tendencia que determina la composición global de las proteínas, mientras que el segundo eje las divide según la temperatura óptima de crecimiento [véase por ejemplo (Singer y Hickey 2000; Kreil y Ouzounis 2001; Tekaia et al. 2002; Lobry y Chessel 2003; Suhre y Claverie 2003; Naya et al. 2004)]. Es importante destacar que la influencia del contenido en GC genómico en la composición media de aminoácidos se ha demostrado previamente en varios procariotas totalmente secuenciados (Lobry 1997). En otras palabras, las especies caracterizadas por niveles bajos de GC exhiben una frecuencia más alta de aminoácidos codificados por los codones ricos en AT en primera y segunda posición (como Lys, Ile, Phe, Asn), mientras que los procariotas que exhiben niveles altos de GC genómico muestran una ocurrencia más alta de residuos codificados por los codones ricos en GC en esas posiciones, como Gly, Ala, Arg y Pro. Además de esta tendencia, las proteínas de hipertermófilos muestran un sesgo fuerte para el uso de los residuos cargados (Asp, Glu, Lys, Arg) a expensas de los residuos polares (Asn, Gln, Ser, Thr). Por otra parte, los análisis intragenómicos han sugerido que los factores como hidrofobicidad, expresividad, aromaticidad (Lobry y Gautier 1994; Rispe et al. 2004), posición respecto a la hebra líder o retrasada de la replicación (Rocha et al. 1999) y costo metabólico (Akashi y Gojobori 2002), juegan un papel en la composición de aminoácidos global de cada especie. A

nivel de eucariotas, se han hecho varios estudios, principalmente en especies unicelulares [véase por ejemplo (Garat y Musto 2000; Chanda et al. 2005)]. Los factores principales que explicaban la variabilidad entre las proteínas eran el sesgo composicional (principalmente en *Plasmodium falciparum*, que tiene un GC genómico de solamente 18% (Pollack et al. 1982), el contenido de Cys (en *Giardia lamblia*) y peso molecular medio, hidropatia, aromaticidad y expresividad de cada gen (en las dos especies). Por lo tanto, parece ser que entre las especies unicelulares, la composición genómica, los factores antedichos, junto con las frecuencias medias de algunos aminoácidos [entre los que Cys parece ser importante, ver también (Zavala et al. 2002)], constituyen tendencias universales que forman la arquitectura de las proteínas; mientras que la temperatura óptima del crecimiento aparece como factor muy importante para los procariotas.

El uso de aminoácidos en el genoma humano (y probablemente de todos los homeotermos) se encuentra limitado por la presencia de isocoros (D'Onofrio et al. 1991). Debido a la relación lineal entre el nivel de GC de las posiciones del codón y el GC de los isocoros donde se encuentran, tanto el uso de codones como la composición de aminoácidos son diferentes para proteínas codificadas por genes localizados en isocoros con niveles de GC distintos. Otra restricción importante para la utilización diferencial de aminoácidos parecería ser la estructura de dominios de la proteína donde la conservación es muy alta (Moses y Durbin 2009). La frecuencia de sustitución de bases, así como inserciones y deleciones es más acentuada en los extremos de los genes permitiendo una evolución más acelerada de las secuencias en los extremos de las proteínas (Yang et al. 2009). Existe además una correlación positiva entre el GC_3 de los genes y la hidropatía de las proteínas codificadas (D'Onofrio et al. 1999). El análisis de proteínas humanas donde se conocía la estructura cristalográfica y la secuencia de bases, reveló que existen diferencias significativas en la composición en la tercera posición del codón en regiones codificando para distintas estructuras secundarias (D'Onofrio et al. 2002). Estas diferencias son más marcadas en genes con alto contenido en GC. Sin embargo, la segunda posición del codón parece ser especialmente importante en la determinación de las diferentes estructuras secundarias (Chiusano et al. 2000).

HIPOTESIS

En base a los antecedentes mencionados nos planteamos las siguientes hipótesis de trabajo:

A la luz de los datos actuales, con varios genomas completamente secuenciados podrían no sustentarse todos los aspectos del modelo de isocoros.

Los cromosomas son unidades de selección y las características composicionales demostradas a nivel de genes y regiones del genoma se mantienen y podrían acentuarse cuando se observan a escala cromosómica.

Los principales factores que modelan el uso de codones y aminoácidos en el genoma humano tienen relación con los aspectos composicionales y estructurales de sus estructuras subyacentes.

OBJETIVOS

OBJETIVOS GENERALES

- Estudiar la composición del genoma humano, y su comparación con otros vertebrados, tanto homeotermos como poiquilotermos.
- Estudiar la vinculación de la composición del genoma humano con el uso de codones sinónimos y uso de aminoácidos.

OBJETIVOS PARTICULARES

- Determinar la presencia de variaciones composicionales regionales en el genoma humano a través del análisis de sus secuencias.
- Determinar la existencia de familias de isocoros
- Desarrollar herramientas de software capaz de realizar estudios sobre la presencia de isocoros y familias de isocoros en secuencias de ADN. Así como herramientas de software para análisis composicional general.
- Desarrollar un portal web asociado a una base de datos sobre isocoros para compartir los estudios con la comunidad científica a través de interfases gráficas y web services con soap y xml.
- Estudiar las principales tendencias en el uso de aminoácidos en el genoma humano
- Estudiar las principales tendencias en el uso de codones en el genoma humano

MATERIALES Y MÉTODOS:

ANALISIS COMPOSICIONAL DE GENES

La base de datos para analizar las correlaciones composicionales de los genes, se obtuvo a partir de la combinación de la colección curada de Refseq (27.000 secuencias) con aproximadamente 30.000 genes putativos mapeados en la versión 35 del genoma humano obtenidos del portal de Ensembl (ftp://ftp.ensembl.org/pub/current_human/data/fasta/cdna/).

Se realizaron blasts recíprocos entre ambas bases de datos, descartando alineamientos que mostraron una identidad menor al 98,5 %. Así, se consiguió depurar una base con 7.800 genes.

HETEROGENEIDAD EN CROMOSOMAS HUMANOS

El genoma humano completo versión 36.1 se descargó del NCBI en ftp://ftp.ncbi.nlm.nih.gov/genomes/. Un conjunto de 20.094 secuencias de genes y referencias de secuencias del genoma humano versión 36,1 se descargaron del sitio ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/ NCBI. El proyecto CCDS es un esfuerzo de colaboración para identificar un conjunto básico de proteínas humanas codificadas en regiones anotadas en forma segura. La localización en el cromosoma de cada gen, así como la estructura exón-intrón se obtuvieron de este conjunto de datos. Las regiones flanqueantes fueron consideradas como las regiones 25 kb antes del codón de inicio más la región 25 kb luego del codón de terminación. Todas las medidas de contenido de GC se realizaron con un software desarrollado internamente.

La localización nuclear de los cromosomas fue tomada de (Bolzer et al. 2005).

DATOS DE EXPRESIÓN

Los datos de expresión se descargaron de "Human Transcriptome Map" para todos los tejidos normales medidos por SAGE (http://bioinfo.amc.uva.nl/HTMseq/controller)

DATOS DE RECOMBINACIÓN

Las frecuencias de recombinación se obtuvieron del HapMap (Consorcio Internacional HapMap, http://hapmap.ncbi.nlm.nih.gov/).

INDICE ARMONICO DE MASA

El índice armónico de masa (HMI) es una medida de similitud de dos distribuciones dadas (Hinloopen 2005) y se puede observar como la superficie total delimitada por la diagonal y la curva en un PP-plot (gráfica de probabilidad para evaluar cuánto coinciden dos conjuntos de datos).

METODO DE CLUSTERING

El análisis de cluster jerárquico se realizó mediante la función "hclust" del lenguaje estadístico "R" (R Team. 2008). La relación completa resultó el método preferido ya que encuentra agrupaciones muy similares, a pesar de que otros métodos tales como "Ward" o "McQuitty" exhiben imágenes conceptualmente similares.

ANALISIS DE USO DE CODONES Y BUSQUEDA DE ISLAS CpG EN EL GENOMA HUMANO

Las secuencias de ADN fueron descargados de la base de datos UniGene (Ftp.ncbi.nlm.nih.gov repositorio / / UniGene). Fueron conservadas un total de 11.657 secuencias completas y no redundantes. El análisis de correspondencia de uso de codones (CoA), la frecuencia de los codones terminados en G y C (GC_3), y el uso relativo de codones sinónimos (RSCU) (Sharp et al. 1986) se calcularon utilizando el programa CodonW, escrito por John Peden obtenido de http://codonw.sourceforge.net/. Los genes que mostraron los valores más extremos de acuerdo con el segundo eje del CoA se localizaron en los cromosomas respectivos utilizando la opción de BLAST en la base de datos ENSEMBL [www.ensembl.org (Hubbard et al. 2005)]. Los resultados con los valores de e más bajos fueron elegidos sólo para los casos $\leq 10^{-3}$ y con una longitud $\geq$ a 150 nucleótidos. RSCU es la frecuencia observada de un codón, dividido por la frecuencia esperada si todos los sinónimos que codifican el aminoácido se utilizan por igual, por lo que valores de RSCU cercanos a 1,0 indican ausencia de sesgos para el uso de ese codón.

ANÁLISIS DE AMINOÁCIDOS EN LAS PROTEÍNAS DEL GENOMA HUMANO

El genoma humano completo versión 35 se descargó de NCBI en ftp://ftp.ncbi.nlm.nih.gov/genomes/. Un conjunto de 14.815 secuencias de genes junto con sus referencias del genoma humano se descargó desde el

CDS consenso (CCDS) en el sitio ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/ de NCBI.

El uso de aminoácidos y el análisis de correspondencia fueron calculados utilizando el programa CodonW (ver más arriba).

El contenido de GC de las regiones circundantes (que nuevamente fue considerada como la región 25 kb antes del codón de inicio más la región 25 kb después del codón de terminación) se calculó para cada gen.

Las frecuencias de las diferentes estructuras secundarias dentro de cada proteína se estimaron usando el programa PSI-Pred (Jones et al. 1999) instalado localmente contra una base de datos de 76.726 proteínas. La longitud mínima (de una α hélice y de una hoja β) se establecieron en seis y cuatro aminoácidos consecutivos respectivamente (Garnier et al. 1978). El nivel de hidropatía de cada proteína se calculó de acuerdo a la escala de Kyte y Doolittle (Kyte y Doolittle 1982). Las proteínas de membrana fueron seleccionadas de acuerdo con la clasificación del sitio de ontología de genes (www.geneontology.org/).

ALGORITMOS Y HERRAMIENTAS DESARROLLADOS

Durante el transcurso de esta tesis se crearon algoritmos y herramientas de interpretación y explotación de resultados. Se desarrollaron herramientas novedosas capaces de mostrar nuevas características del material de estudio con el objetivo de capitalizarlas a través de modelos que expliquen la estructura y dinámica del genoma de los vertebrados.

El primer objetivo perseguido fue el de desarrollar una herramienta de software capaz de realizar estudios sobre la presencia de isocoros y familias de isocoros en secuencias de ADN.

ISOCHORE PROFILING TOOLS (IPT)

A partir del objetivo anteriormente mencionado, se comenzó por desarrollar un algoritmo capaz de determinar secuencias largas de ADN composicionalmente homogéneas tomando datos de la secuencia de los genomas. El algoritmo consiste, en forma esquemática, en un motor que recorre el genoma que se está analizando a través de una ventana no solapante de tamaño "V", generando fragmentos caracterizados por su contenido en GC y su posición en el genoma. Esta lista de trozos que se generan, son tomados por el siguiente paso del algoritmo, el cual integra el fragmento a una clase superior que contienen una nueva lista ordenada de los mismos llamada "superfragmento". El nuevo trozo es incluido en la lista si cumple con el siguiente criterio: el GC del fragmento a integrar no debe ser mayor que el GC del elemento con menor porcentaje de GC más un valor "P" de tolerancia, y el GC del fragmento a integrar no es menor que el GC del elemento con mayor porcentaje menos el valor "P" de tolerancia. Así comienza a crecer el superfragmento adicionando los trozos generados por el motor del algoritmo. Una vez que un fragmento en la lista no cumple con el criterio de inclusión, se evalúa el tamaño del superfragmento contra otro parámetro "L" que identifica el largo mínimo del superfragmento para ser considerado un "fragmento largo homogéneo", (considerando el largo del fragmento como la suma de los largos de los fragmentos que constituyen su lista miembro). Si el superfragmento no es lo suficientemente largo se lo considera como un superfragmento perteneciente a la clase heterogénea, si

su largo es mayor o igual a L, comienza una nueva fase del algoritmo, la de retroextensión. Esta consiste en la adición al superfragmento compuesto por elementos [n1…nn] del elemento n-1 perteneciente al superfragmento anterior (siempre y cuando se cumplan 2 condiciones: 1) exista un superfragmento anterior, 2) el fragemento n-1 cumpla con el criterio de inclusión). Si estas 2 condiciones se cumplen, el elemento n-1 pasa a ser el elemento n1 del fragmento actual y comienza un nuevo paso de retroextensión. La retroextensión potencia la capacidad de un superfragmento de captar elementos y así genera superfragmentos homogéneos más largos (figura 8).

El algoritmo se programó en C#, un lenguaje de alto nivel orientado a objetos y se implementó a través de tres interfases, GUI (Grafical User Interfase), WEB y WEBSERVICES, generándose así un sitio web capaz de bridar servicios en el análisis de isocoros en secuencias de ADN.

El conjunto de programas generados extienden las capacidades inicialmente previstas incluyendo funcionalidades para mapeo composicional de genomas como se muestra en la figura 9. Se les dio el nombre de gIPT a la implementación GUI (figura 9) y wIPT a la implementación web. Tanto gIPT como wIPT dependen de un motor de base de datos. Para wIPT se utilizó SQL server 2000 como motor de base de datos, de la forma distribuida gIPT se generaron nueve versiones capaces de conectarse con DB2 6.1 , DB2 7.1, DB2 8.1, SQL Server 7, SQL Server 2000, Oracle 8.1.7 to 8.x.x, Oracle 9.x.x y Access. La salida de los programas es hacia archivos de texto separados por comas o al lenguaje estructurado XML que permite identificar fácilmente la estructura del resultado. El sitio web desarrollado se encontró durante el año 2005 albergado en http://oeg.fcien.edu.uy/ipt/ y tenía 2 áreas principales: la de análisis y la de descarga (figura 10). El área de análisis estaba compuesta por un panel de ingreso de parámetros que invoca a wIPT en forma asincrónica, una vez realizado el análisis wIPT reporta el resultado vía correo electrónico y un panel donde se encuentran una grilla con enlaces para descarga de análisis ya realizados. El área de descarga contenía las nueve versiones de gIPT para descargar en formato comprimido incluyendo un instructivo de instalación, así como algunas herramientas accesorias.

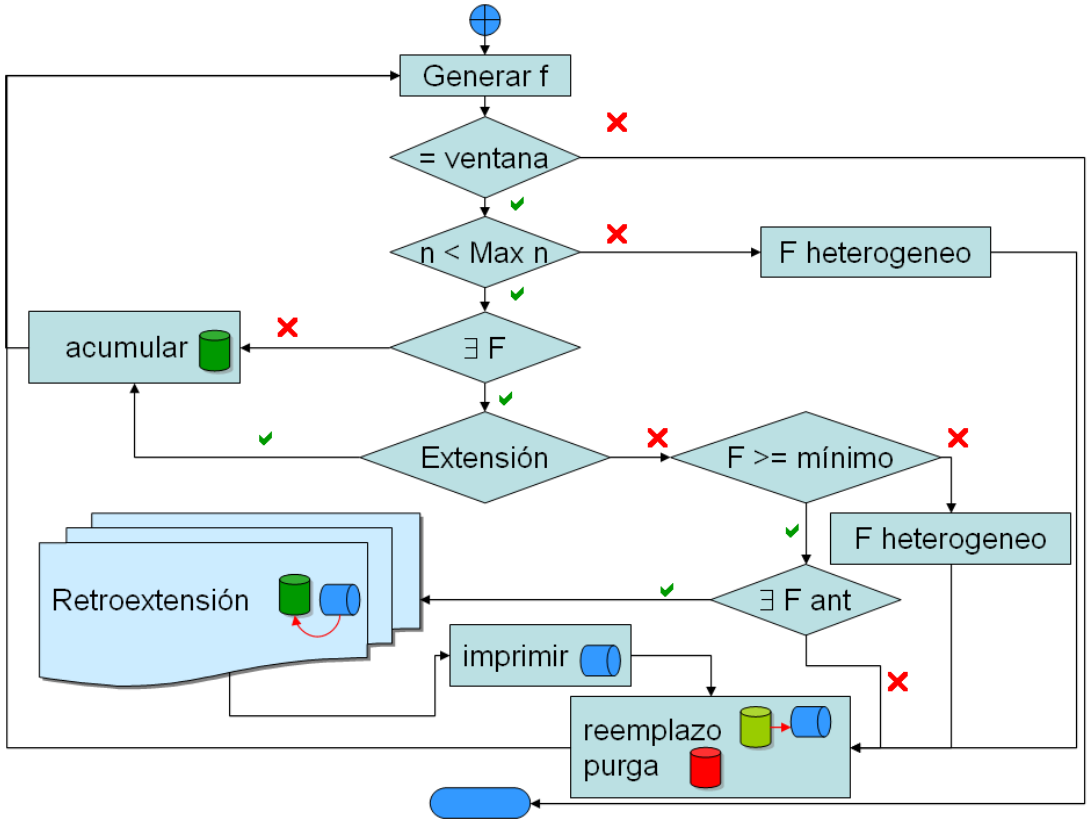


Figura 8. Esquema de flujo del algoritmo.

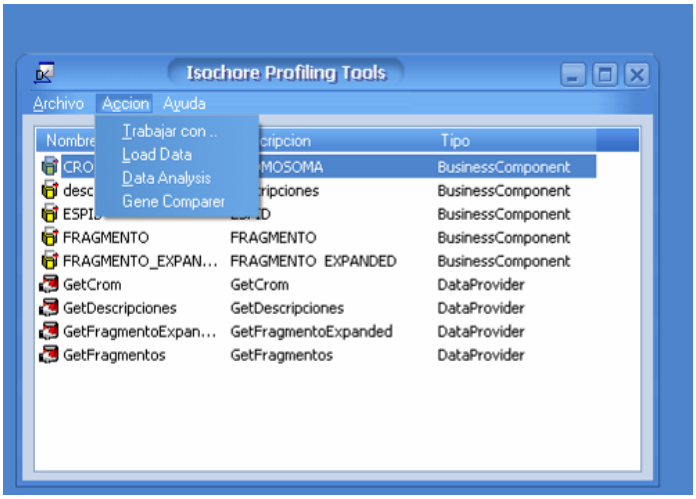


Figura 9. Imagen de captura de pantalla de la aplicación gIPT

Figura 10. Sitio web implementando wIPT

IPT – EXCEL CONVERTER

Entre las herramientas accesorias desarrolladas se encuentra el IPT – Excel converter que toma una salida estructurada XML de wIPT o de gIPT y crea un archivo Excel con algunos análisis gráficos básicos ya predefinidos. Utilizando las funcionalidades de automation de Excel el programa crea una hoja de cálculo con todos los datos extraídos del archivo y luego a través de una macro de Excel los analiza, desplegando al usuario final un análisis básico completo (figura 11).

Figura 11. IPT – Excel converter

Otra herramienta desarrollada es el visualizador de superfragmentos que permite ver el resultado en formato XML, permitiendo personalizar los colores, así como algunas marcas en la salida; por ejemplo variaciones de determinado rango de GC entre los superfragmentos (figura 12). Esta herramienta utiliza el archivo XML de salida de IPT y despliega la información de acuerdo a un esquema de colores configurable. Así, es posible asignar un color a un rango de valores de GC y distinguirlos fácilmente en el gráfico. Además se puede configurar para que el programa marque los sitios donde se encuentra una diferencia de contenido en GC entre 2 fragmentos contiguos, mayor a un valor ingresado. Cuando se hace click sobre cualquier parte del gráfico se despliega una nueva ventana por encima de la ventana principal mostrando los datos del componente seleccionado. Entre los datos visualizables se encuentran la composición del fragmento y los fragmentos anteriores y posteriores, la cantidad de superfragmentos que tienen cada uno de ellos y la diferencia en % de GC entre el fragmento seleccionado y los otros dos. Además se provee una herramienta para avanzar o retroceder hacia el fragmento anterior o posterior.

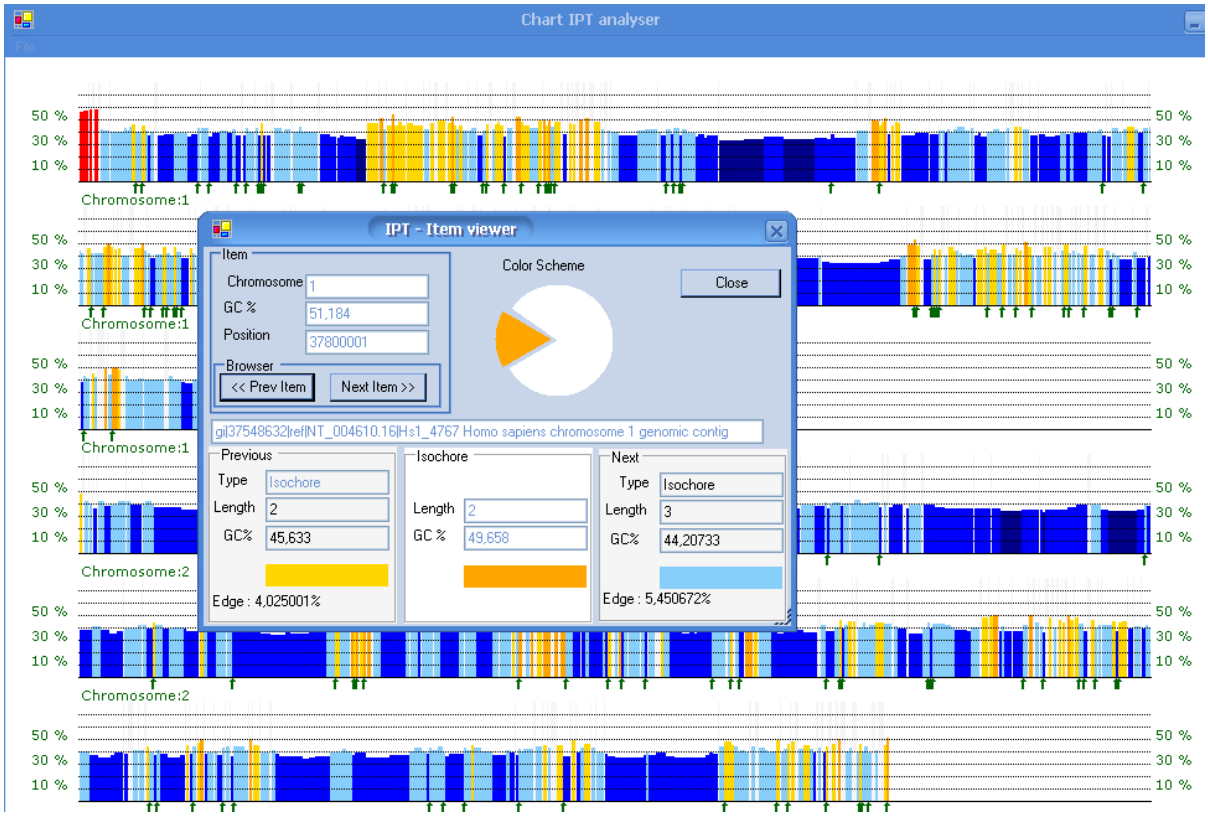
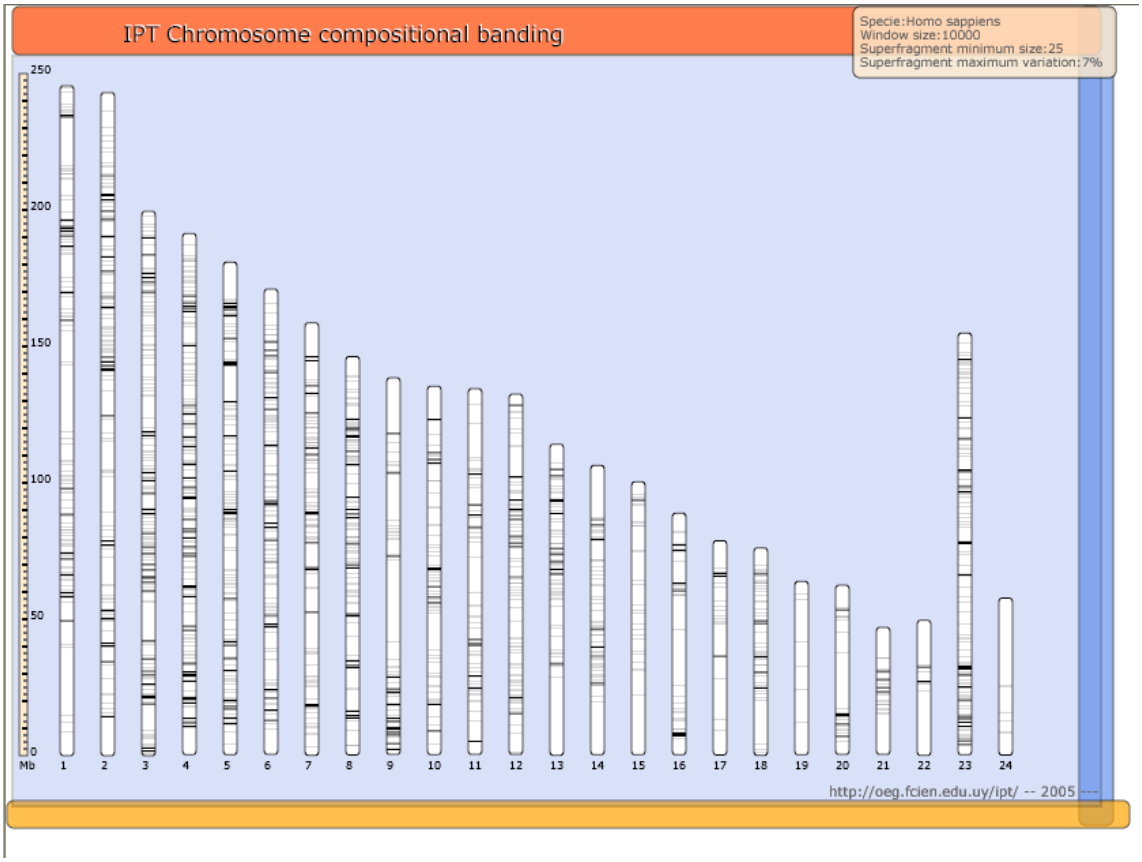
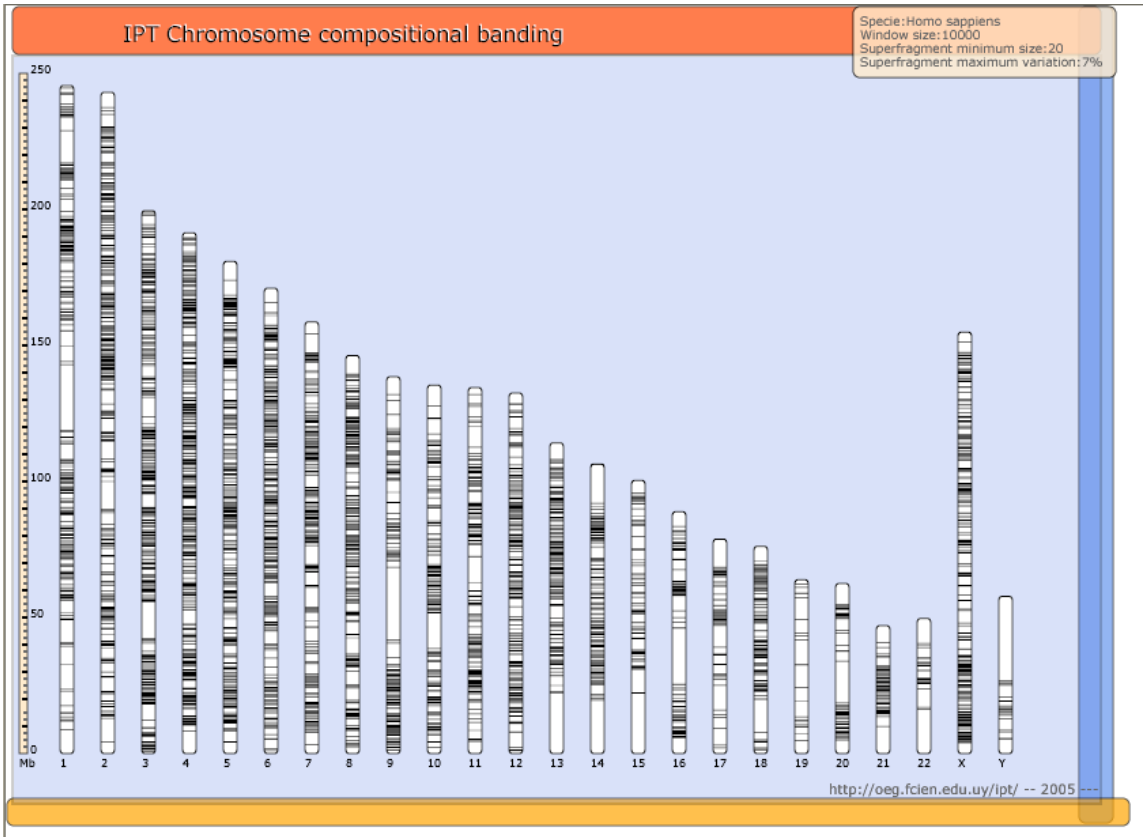
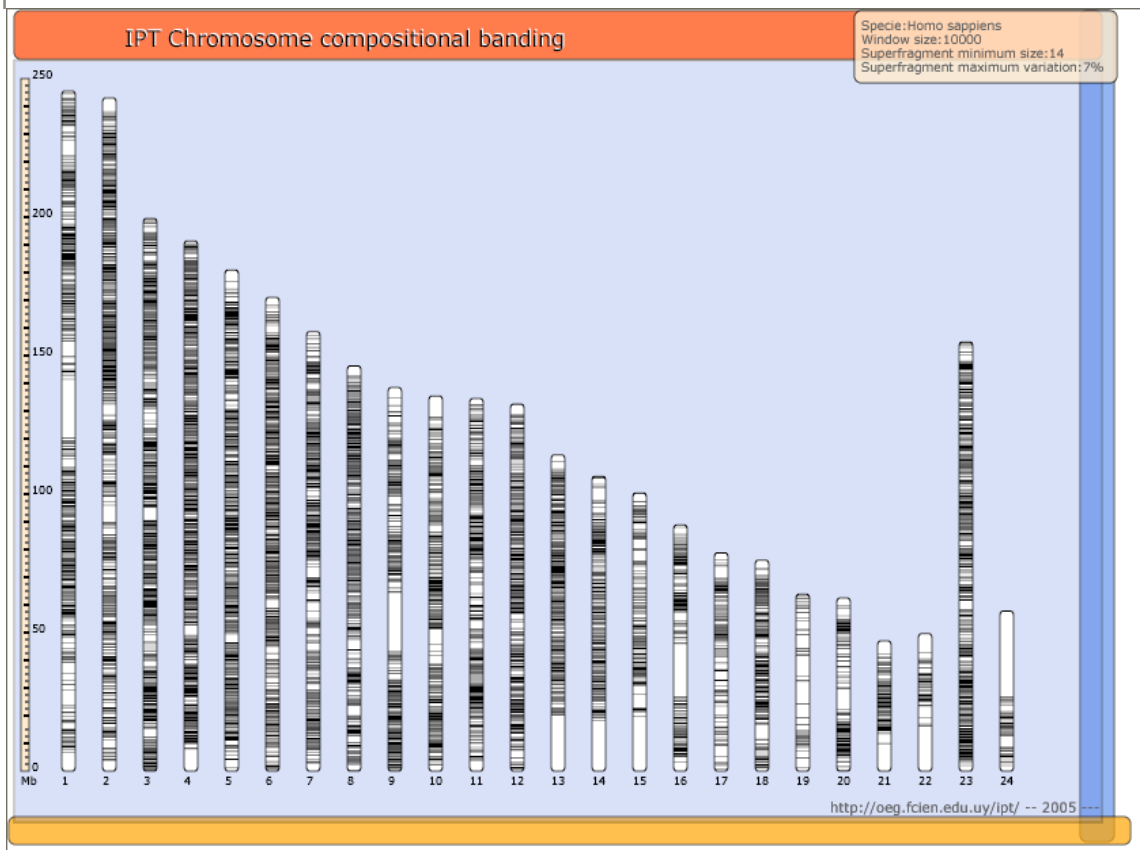


Figura 12. Visualizador de fragmentos marcando diferencias de %GC mayores o iguales a 4%

Otra herramienta desarrollada consiste en el generador de bandeo cromosómico por homogeneidad. Se trata de una salida hacia gráficos vectoriales escalables de los datos producidos tras un análisis IPT. Parte de las ventajas que ofrece son varios niveles de aumento (zoom) en la salida, que permiten la visualización de pequeñas estructuras al detalle, además de brindar una imagen general de las zonas con mayor nivel de homogeneidad, todo lo cual está integrado en una pequeña área de pantalla (figura 13). El bandeo composicional permite distinguir zonas homogéneas (oscuras) de zonas heterogéneas (claras) en los cromosomas. De acuerdo a los parámetros del análisis, que determinarán el grado de exigencia para considerar una zona como homogénea se obtienen resultados con distintas intensidades de oscuro; de forma tal de que cuanto más relajadas las condiciones resultarán más marcadas las bandas homogéneas y eventualmente aparecen nuevas bandas antes imperceptibles. En la figura 13 se puede observar el resultado gráfico de un análisis donde se relajaron progresivamente las condiciones de homogeneidad.


IPT Chromosome compositional banding
Specie:Homo sappiens
Window size:10000
Superfragment minimum size:25
Superfragment maximum variation:7%
250
200
150
100
50
0
Mb
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
http://oeg.fcien.edu.uy/ipt/ -- 2005


IPT Chromosome compositional banding
Specie:Homo sappiens
Window size:10000
Superfragment minimum size:20
Superfragment maximum variation:7%
250
200
150
100
50
0
Mb
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
http://oeg.fcien.edu.uy/ipt/ -- 2005


IPT Chromosome compositional banding
Specie:Homo sappiens
Window size:10000
Superfragment minimum size:14
Superfragment maximum variation:7%
250
200
150
100
50
0
Mb
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
http://oeg.fcien.edu.uy/ipt/ -- 2005

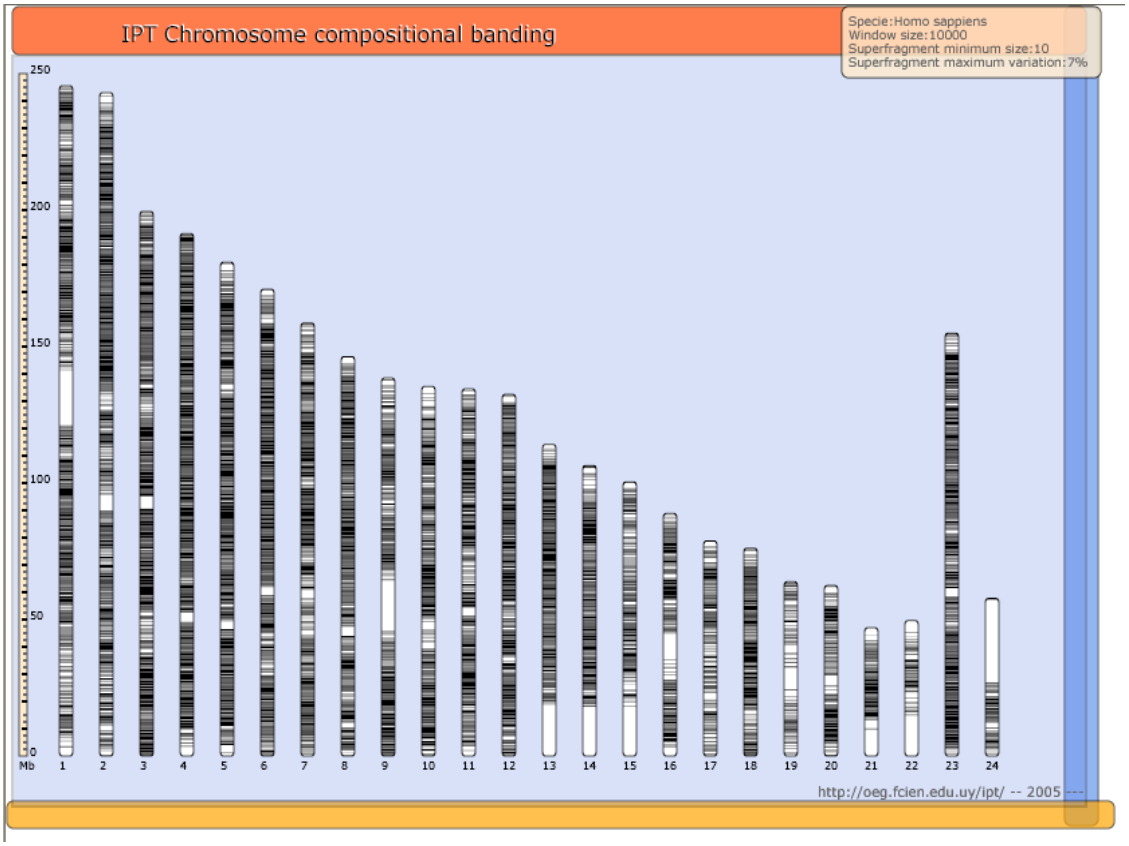


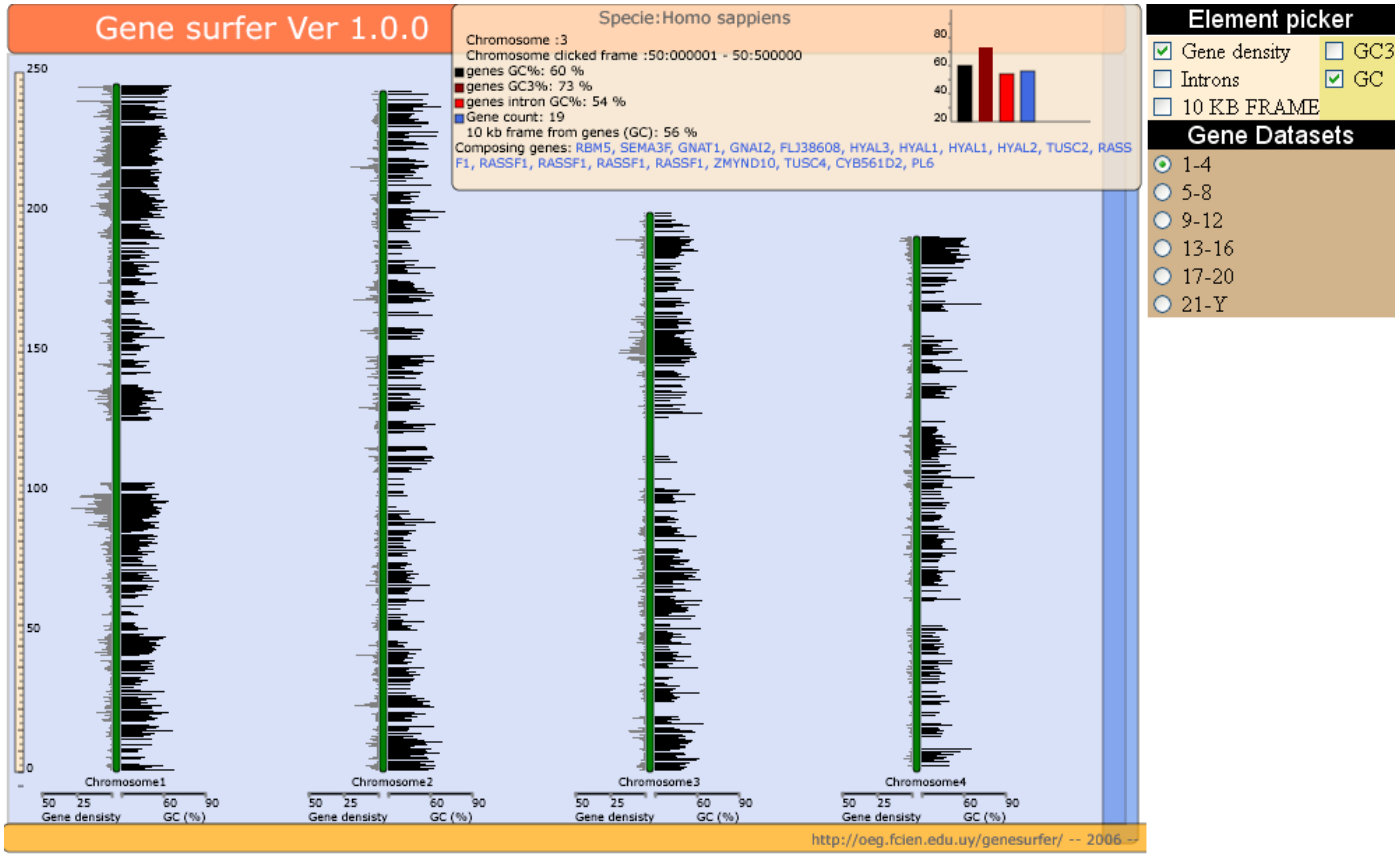
Figura 13. Bandeo composicional cromosómico del genoma humano. Los parámetros fueron: L= 20 P= 7% y a) V = 25 kb, b) V = 20 kb, c) V = 15 kb. d) V = 10 kb Las bandas oscuras representan superfragmentos homogéneos.

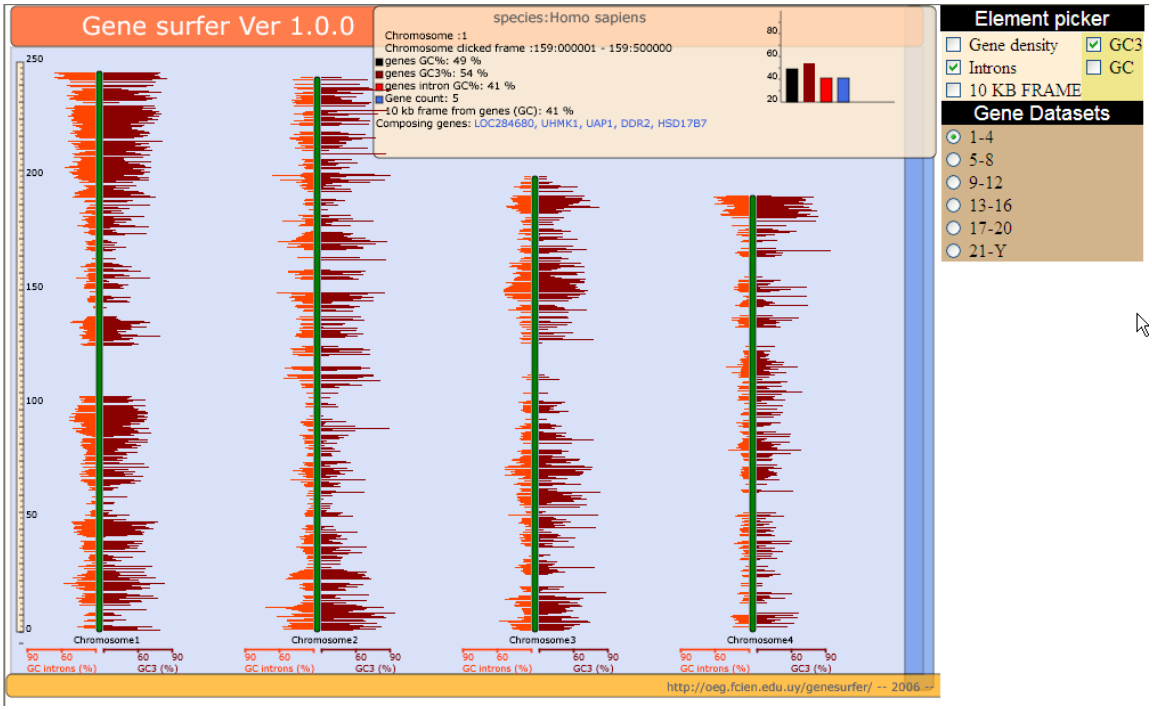
GENESURFER

En la actualidad existen excelentes herramientas para la visualización de la composición del genoma, principalmente en el sitio de ensemble (www.ensemble.org), donde el contenido en GC se puede evaluar en diversas regiones del genoma y a distintos niveles de granularidad. Aunque este sitio es muy útil para evaluar la variación composicional a nivel genómico no es posible analizar las características composicionales de elementos génicos dentro de cada cromosoma, obstaculizando la posibilidad de estudiar la relación entre las características genómicas y génicas. Para lograr mayor detalle en la exploración de los cromosomas completos disponibles y considerando la importancia de localizar las características composicionales observadas de genes a lo largo de ellos, desarrollamos el software "Gensurfer". Esta herramienta es capaz de exhibir varias características composicionales de los genes tales como GC medio, GC_3 medio, GC medio de intrones, GC medio de la región del cromosoma donde

se localiza cada gen (a partir de 10 kb corriente arriba del codón de iniciación hasta 10 kb corriente abajo del codón de terminación). También es capaz de mostrar la densidad génica para una localización dada del cromosoma. Gensurfer funciona sobre HTML, una característica que le permite ser plataforma-independiente. Los datos se almacenan en la base de datos de los gráficos vectoriales (SVG), que es también plataforma independiente con una amplia gama de plug-ins ya desarrollados para los navegadores de internet. Para una lista completa de los visualizadores de SVG y de los kits de desarrollo, se puede consultar el sitio de la fundación de SVG en http://www.svgi.org/. La base de datos del Gensurfer fue construida con la versión 35 de las secuencias consenso de genes humanos (CCDS) que se pueden obtener de http://www.ncbi.nlm.nih.gov/CCDS/, las cuales se encuentran embebidas en el SVG y pueden ser manipuladas fácilmente por el usuario. Este programa tiene un área de resumen para cada punto en el cromosoma que muestra las características composicionales y los nombres de los genes cuando el usuario pasa el indicador del ratón sobre él. Esta es una manera fácil de navegar los cromosomas obteniendo información composicional exacta para los genes (total, discriminando por la posición del codón y los intrones) y también para la región genómica circundante (figura 14).

Resumiendo, Genesurfer es una herramienta gráfica que reúne datos composicionales con datos posicionales de los genes a lo largo del cromosoma. Además podría, si se desea, ampliar la información para otra clase de datos ligados a la localización en el cromosoma.


Gene surfer Ver 1.0.0
Specie:Homo sappiens
Chromosome :3
Chromosome clicked frame :50:000001 - 50:500000
genes GC%: 60 %
genes GC3%: 73 %
genes intron GC%: 54 %
Gene count: 19
10 kb frame from genes (GC): 56 %
Composing genes: RBM5, SEMA3F, GNAT1, GNAI2, FLJ38608, HYAL3, HYAL1, HYAL1, HYAL2, TUSC2, RASSF1, RASSF1, RASSF1, RASSF1, RASSF1, ZMYND10, TUSC4, CYB561D2, PL6
Element picker
Gene density
GC3
Introns
GC
10 KB FRAME
Gene Datasets
1-4
5-8
9-12
13-16
17-20
21-Y
Chromosome1
Chromosome2
Chromosome3
Chromosome4
Gene density
GC (%)
http://oeg.fcien.edu.uy/genesurfer/ -- 2006 --


Gene surfer Ver 1.0.0
species:Homo sapiens
Chromosome :1
Chromosome clicked frame :159:000001 - 159:500000
genes GC%: 49 %
genes GC3%: 54 %
genes intron GC%: 41 %
Gene count: 5
10 kb frame from genes (GC): 41 %
Composing genes: LOC284680, UHMK1, UAP1, DDR2, HSD17B7
Element picker
Gene density
GC3
Introns
GC
10 KB FRAME
Gene Datasets
1-4
5-8
9-12
13-16
17-20
21-Y
Chromosome1
Chromosome2
Chromosome3
Chromosome4
GC introns (%)
GC3 (%)
http://oeg.fcien.edu.uy/genesurfer/ -- 2006 --

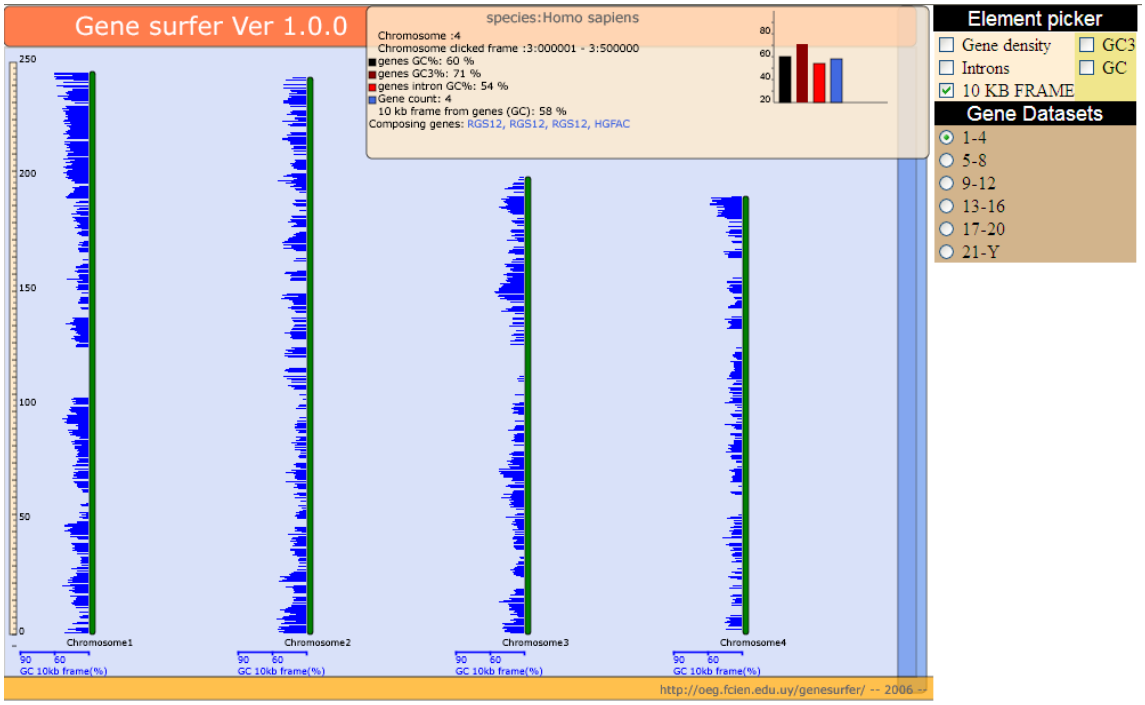


Figura 14. Captura de pantalla de Gensurfer mostrando a) densidad génica y GC y b) cantidad de intrones y % de GC_3 y c) % de GC del marco 10 kb de los genes para los cromosomas 1 a 4 en los tres casos y detalles de la región seleccionada en cada imagen que comienza en la posición 50,000,001 y termina en la posición 50,500,000 del cromosoma 3; y región que comienza en la posición 150,000,001 y termina en la 150,500,000 del cromosoma 1 y región que comienza en la posición 3,000,001 y termina en la 3,500,000 del cromosoma 4, respectivamente.

GENESURFER - EXPRESSION

Utilizando los datos de GEO, (gen expression Omnibus) curados por el grupo de Hugo Naya trabajando en el Instituto Pasteur de Montevideo, se construyó una base de datos relacional y una aplicación web para poder analizar los datos. Esta aplicación web permite la consulta y edición de todos los datos contenidos en la base de datos. Permite además realizar análisis de expresión por tejido y análisis generando un histograma de GC_3 para los genes componentes separándolos en 3 grupos: tejido especificos, intermedios y constitutivos.

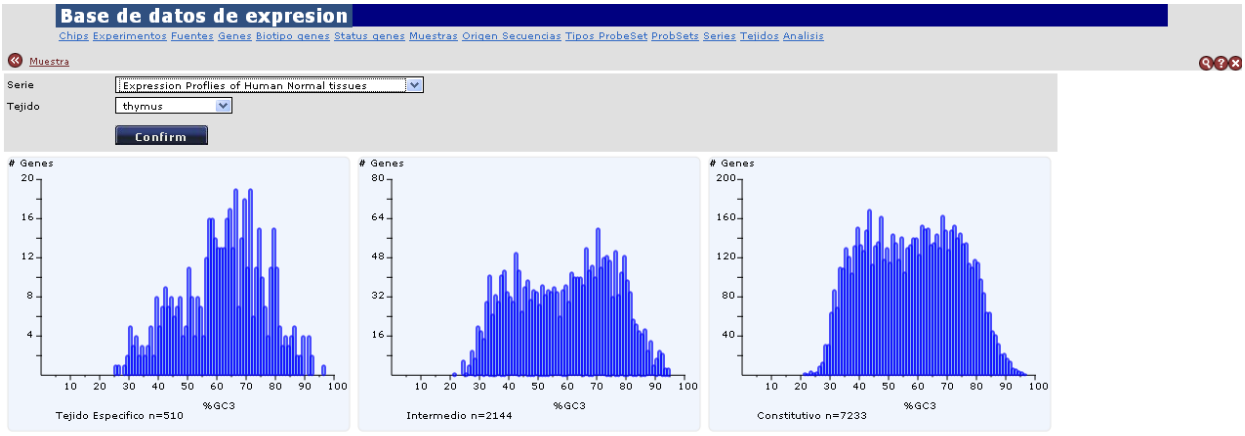
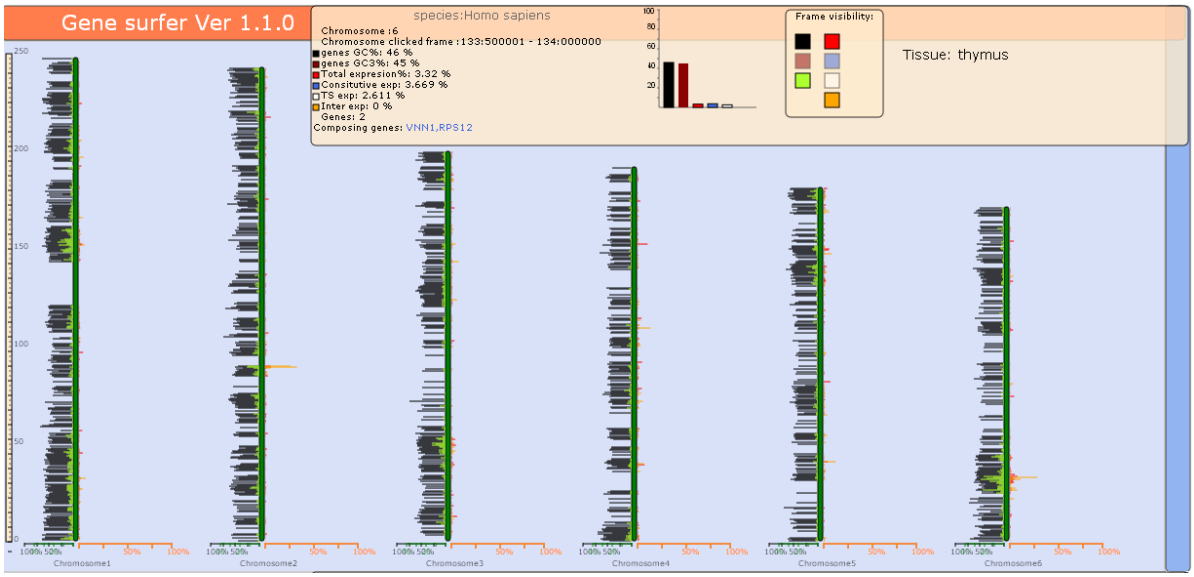


Figura 15. Captura de pantalla de página web, Genesurfer-expression. En este caso se puede ver la clasificación de los genes analizados en una serie para el tejido timo. Para cada grupo de genes se puede ver el perfil de GC_3.

Se desarrolló una versión de gensurfer capaz de desplegar datos de expresión sobre los diferentes cromosomas del genoma humano. Esta versión de gensurfer, se utiliza como api de dibujo de pantalla en la aplicación web de expresión. Gene surfer expression permite conocer el índice de expresión por sector de cromosoma, disciminando entre los 3 grupos de genes (tejido especificos, intermedios y constitutivos). Permite a su vez comparar el nivel de expresión de estos 3 grupos contra el GC del cromosoma y su GC_3.



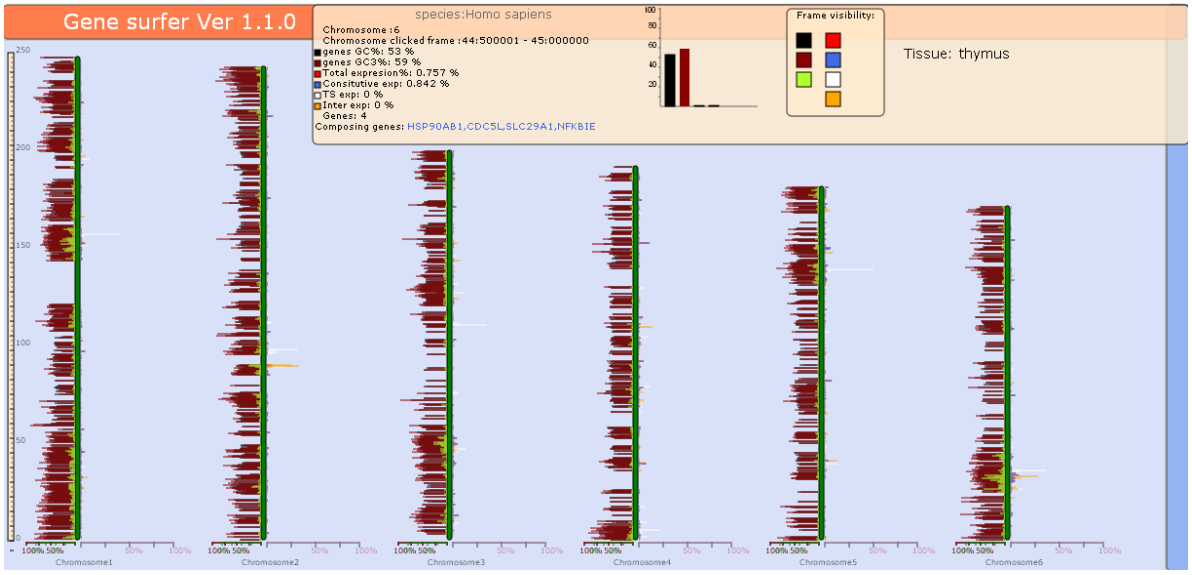


Figura 16. Captura de Pantalla de análisis de Genesurfer expresión proyectando datos de expresión sobre un gráfico de cromosomas.

EXPLORACIÓN DEL ESPACIO PARAMÉTRICO DE IPT

Se exploró el espacio paramétrico evaluándose los resultados obtenidos como porcentaje del genoma retenido como homogéneo vs. heterogéneo. Se fijó el tamaño de ventana en 10 kb y se varió el tamaño mínimo de superfragmento considerado como homogéneo entre L = 20 kb y L = 500 kb tomando 7% como parámetro P. Luego se fijó el tamaño mínimo L en 200 kb y se varió el porcentaje P entre 1% y 20%. El resultado puede verse en los gráficos presentes en la figura 17. En el caso a, cuando se parametriza para una longitud mínima de 20 kb, el 90,2% del genoma es clasificado como homogéneo. En otras palabras, cuando se toma una ventana de 10 kb, en el 90,2% de los casos, es seguido por otro fragmento con una variación menor a 7% de GC. Se aprecia claramente que el porcentaje del genoma que se considera homogéneo disminuye al aumentarse los requisitos. Esto lleva a que solamente el 3,9% del genoma sea homogéneo si se requiere un L de 500 kb. En el caso b, cuando se permite un % de variación interna en el superfragmento de 20%, se clasifica como homogéneo el 90,5% del genoma. Este valor disminuye a medida que disminuye el porcentaje máximo de variación interna aceptado, hasta situarse en 0,01% del total para el caso en el que se permite solamente el 1% de P.

Estos resultados plantean claramente la dificultad existente para definir el criterio de homogeneidad composicional.

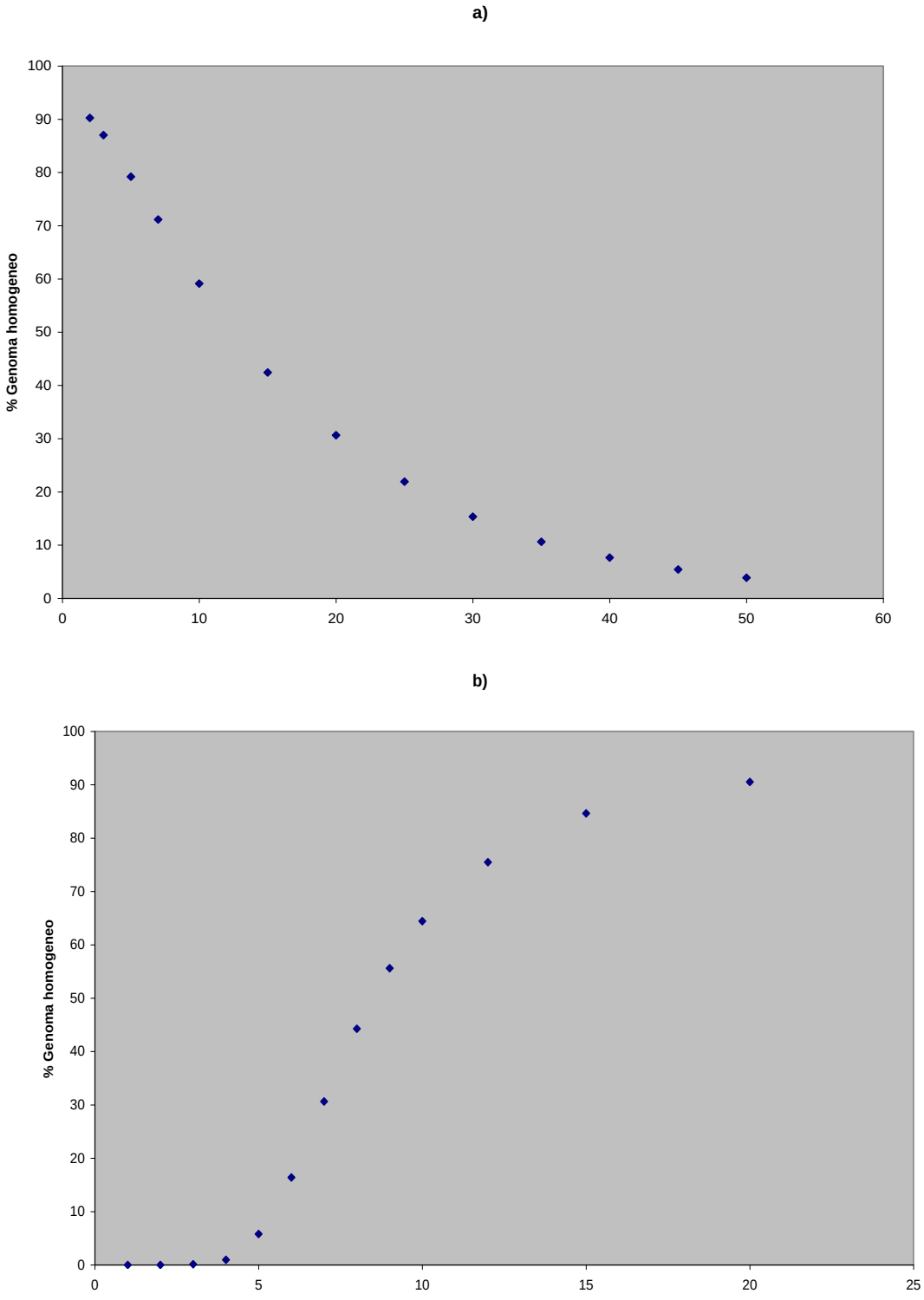


Figura 17. Porcentaje del genoma humano clasificado como homogéneo usando diferentes parámetros de entrada de IPT; utilizando una ventana de 10 kb y a) 7% de P y L entre 20 kb y 500 kb y b) 200 kb de L y P entre 1 y 20 %.

A fin de testar la influencia del orden de los fragmentos en los resultados expuestos arriba, se realizó el siguiente experimento: se cortó el genoma en segmentos no solapantes de 10 kb y se los mezcló al azar 1000 veces, obteniéndose de esa forma 1000 seudogenomas humanos con sus fragmentos desordenados. Se utilizaron estos 1000 genomas como entrada para un análisis con parámetros L = 200 kb y P = 7%. La figura 17 muestra que el resultado depende totalmente del orden natural, ya que ni una sola vez en 1000 se obtiene un 30 % de homogeneidad en el genoma para los parámetros elegidos. Más aún, el resultado más alto es de 0,14 %.

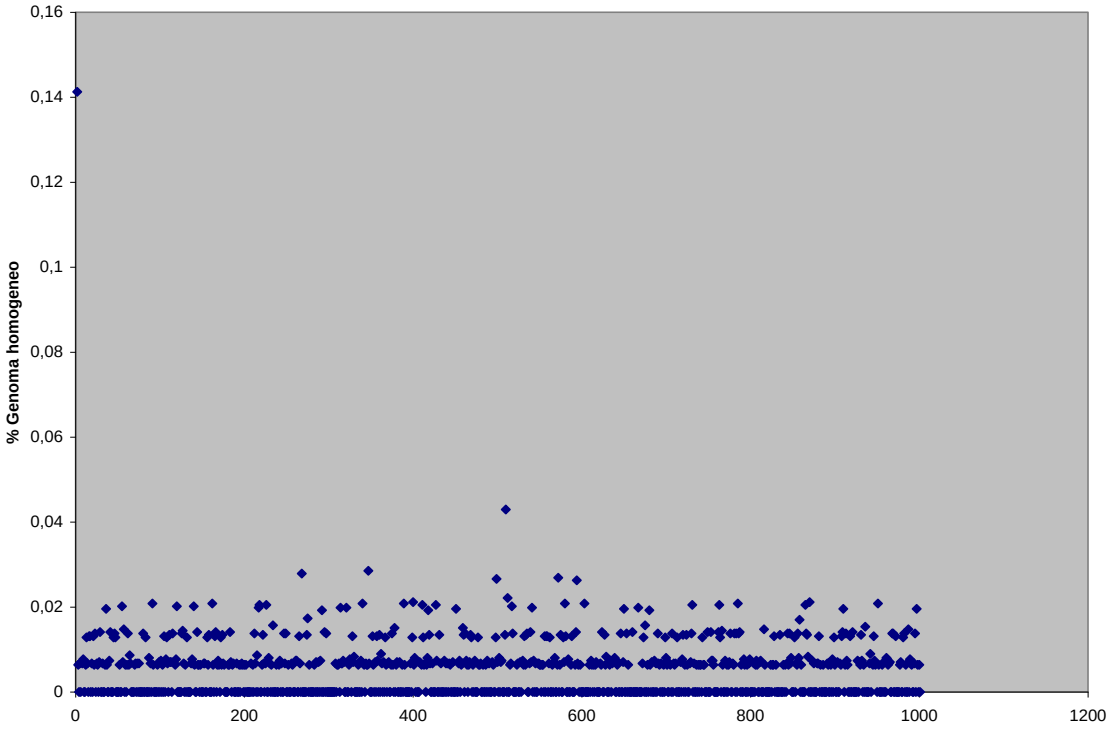


Figura 18. Porcentaje del genoma humano clasificado como homogéneo tras analizar con parámetros de mínimo de tamaño de superfragmento de 200 kb (L) y porcentaje máximo de variación interna de 7% (P) para 1000 réplicas mezcladas al azar del orden natural de fragmentos del genoma humano.

Tomando como parámetros 10 kb de tamaño de ventana (V), 200 kb de largo mínimo de superfragmento (L) y 7% como máximo de variación interna (P), se analizó la distribución de superfragmentos homogéneos y heterogéneos respecto a su GC promedio. La distribución de fragmentos homogéneos tiende a ser más densa a niveles más bajos de GC respecto a la distribución de fragmentos heterogéneos (figura 19a). No se deduce a partir de esta figura la existencia de familias de isocoros. Sin embargo, el gráfico de relación de fragmentos homogéneos respecto a heterogéneos (19b) posee algunos máximos que coinciden con los % de GC reportados por el grupo de Bernardi para las familias de isocoros H1, H2 y H3. Es de notar que el resultado obtenido es altamente dependiente del juego de parámetros del análisis, por lo tanto antes de descartar o asegurar la existencia de familias, es necesario explorar un entorno paramétrico. Los que se escogieron para el análisis surgieron de un trabajo previo de Nekrutenko y Li (2000) quienes utilizan 7% como máximo de variabilidad interna, ya que éste valor es el máximo encontrado en el genoma de la levadura *Saccharomyces cerevisiae* (Nekrutenko y Li 2000). En la tabla 1 se muestran la longitud, la posición y el contenido en GC de los 15 fragmentos más largos que responden al criterio planteado de homogeneidad. De acuerdo a lo esperado, estos superfragmentos son pobres en GC. La distribución desigual entre los diferentes cromosomas resulta llamativa y se estudiará en profundidad más

adelante.

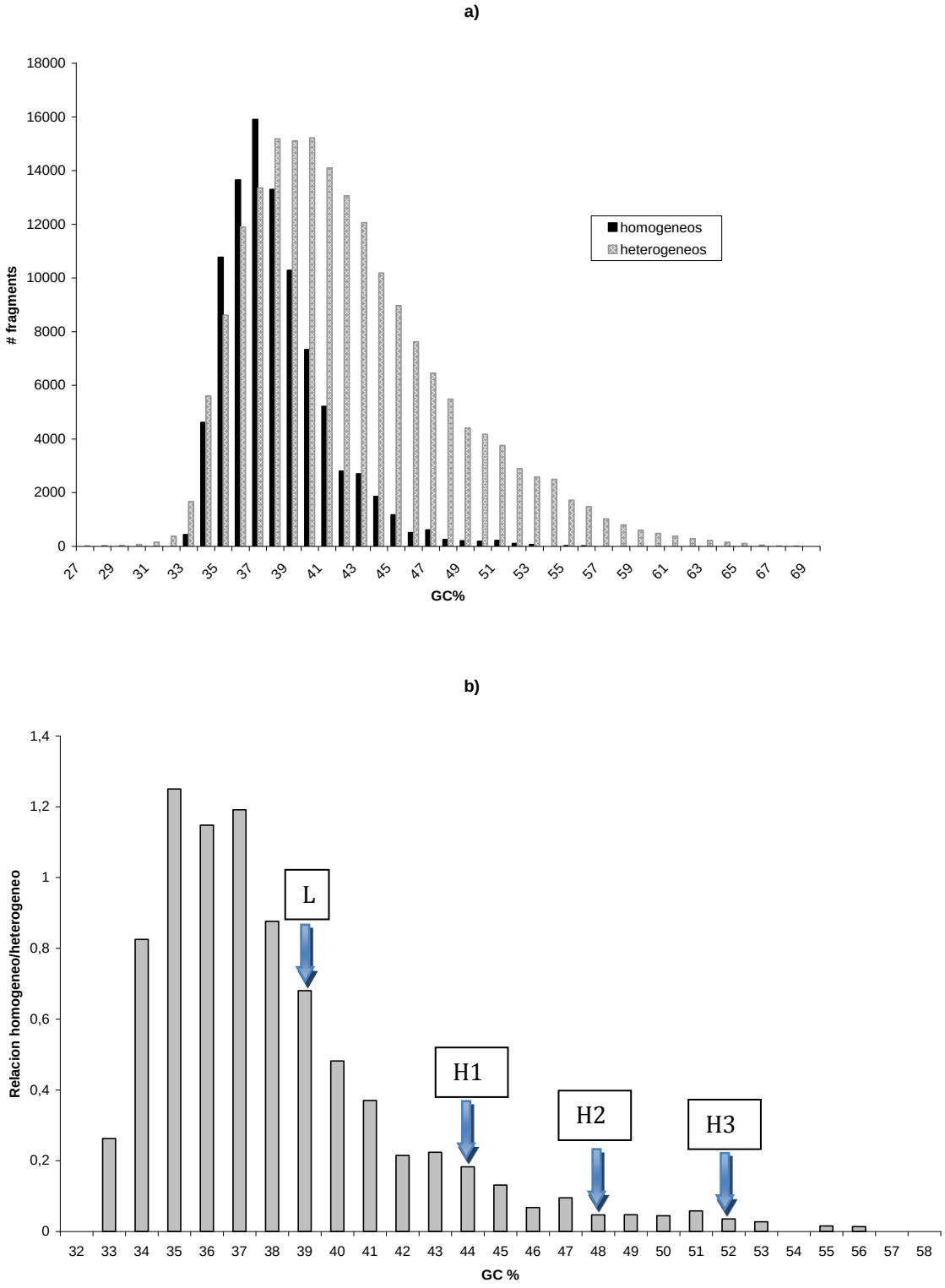


Figura 19. a) Distribución de superfragmentos homogéneos y heterogéneos respecto a su contenido en GC promedio. b) Distribución de la relación homogéneo/heterogéneo de la parte a) se ubicaron en la figura los máximos de las distribuciones de las familias de isocoros según la figura 1.

El análisis de la figura 19 muestra que existen picos en la distribución de la relación hogeneo/heterogéneo en zonas muy cercanas a los picos de las distribuciones encontradas por Bernardi. Si bien no está claro el origen de esta coincidencia, podría estar indicando que la rotura mecánica del ADN que se estima uniforme podría ser menor en las zonas homogéneas y por lo tanto estas zonas estarían sobrerepresentadas formando máximos en la distribución de fragmentos de ADN.

Tabla 1.

Largo de superfragmento (Mb)	Cromosoma	Posición (Mb)	GC %
1110	7	88280	35,75
1050	3	20990	36,46
990	3	28860	36,68
990	21	30240	36,74
980	3	62840	39,00
970	7	17830	36,20
960	X	10040	39,32
940	13	92680	36,57
930	11	82910	37,25
920	21	27270	36,22
900	2	141000	34,51
900	3	173000	39,85
900	9	28180	35,45
890	16	59450	35,78
870	5	164000	35,59

Fragmentos más largos obtenidos después de analizar el genoma humano con parámetros 10 kb V, 200 kb de L y P de 7%.

ANÁLISIS COMPOSICIONAL DE GENES

Se evaluó la frecuencia de genes sobre superfragmentos homogéneos y heterogéneos y por GC de los superfragmentos que los contienen (figura 20). Existe una diferencia clara entre la distribución de los genes en los mismos, dado que la mayoría de los genes se ubican en los fragmentos heterogéneos y además los genes que se ubican en fragmentos homogéneos lo hacen en fragmentos pobres en GC; por otro lado el porcentaje de genes

ubicados en las regiones homogéneas es claramente menor que la frecuencia de estas regiones en el total del genoma.

Por otra parte, se analizó la correlación de GC y GC_3 de cada gen con respecto al GC del superfragmento en el que se ubica (figura 21 y tabla 2). La correlación es mucho más alta tanto para GC como para GC_3 con los fragmentos heterogéneos que con los homogéneos. Por otro lado, se evaluaron las correlaciones de GC y GC_3 del gen con respecto al GC de un marco contenedor de tamaño uniforme que se extiende desde la posición -150.000 del comienzo del gen hasta la posición 150.000 después del codón de finalización (figura 22 y tabla 3). Las diferencias con el análisis anterior son que: 1) en este análisis el marco no contiene al gen, 2) el gen se encuentra equidistante respecto al comienzo y el fin del marco. La similitud entre las correlaciones totales de ambos análisis sugiere que el superfragmento representa correctamente al marco.

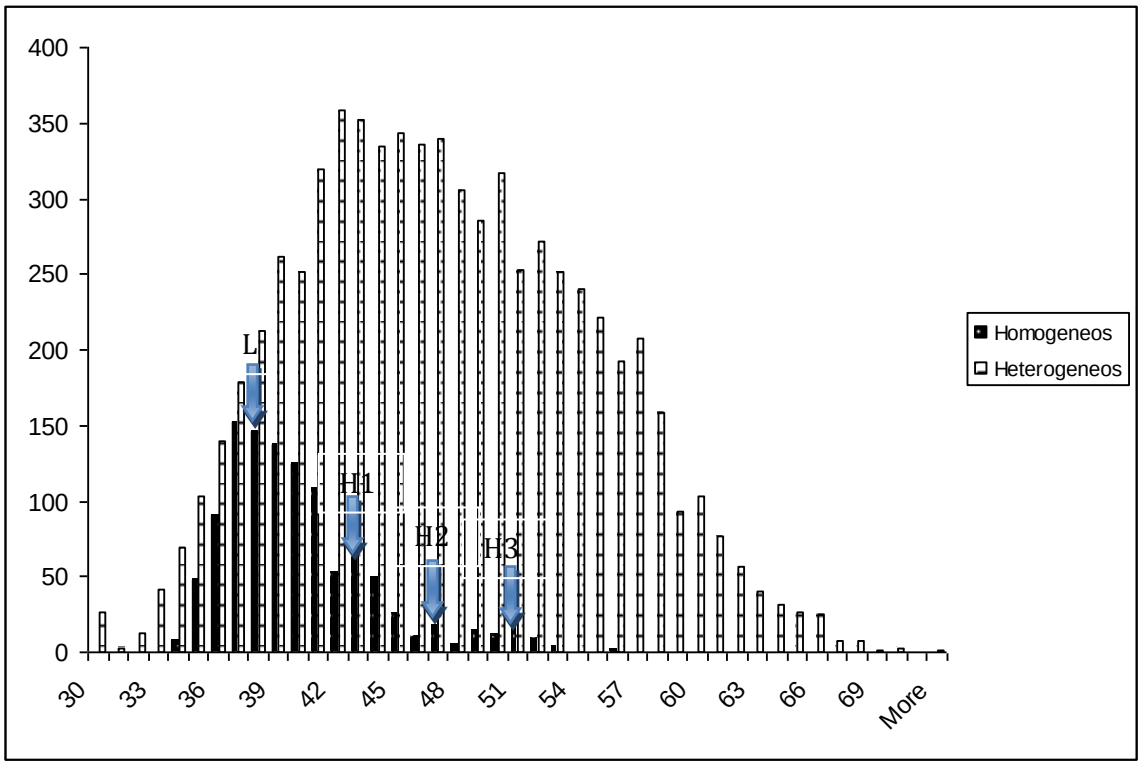


Figura 20. Ubicación de genes en grupos homogéneo y heterogéneo respecto al superfragmento que los contiene, se encuentran indicadas las posiciones de picos de las diferentes familias de isocoros según Bernardi.

Las discontinuidades en la distribución de genes en fragmentos homogéneos parecería coincidir con los máximos de las distintas familias de isocoros. Si bien actualmente no logramos explicar esta particularidad, estudiaremos otras especies completamente secuenciadas para analizar su correlación.

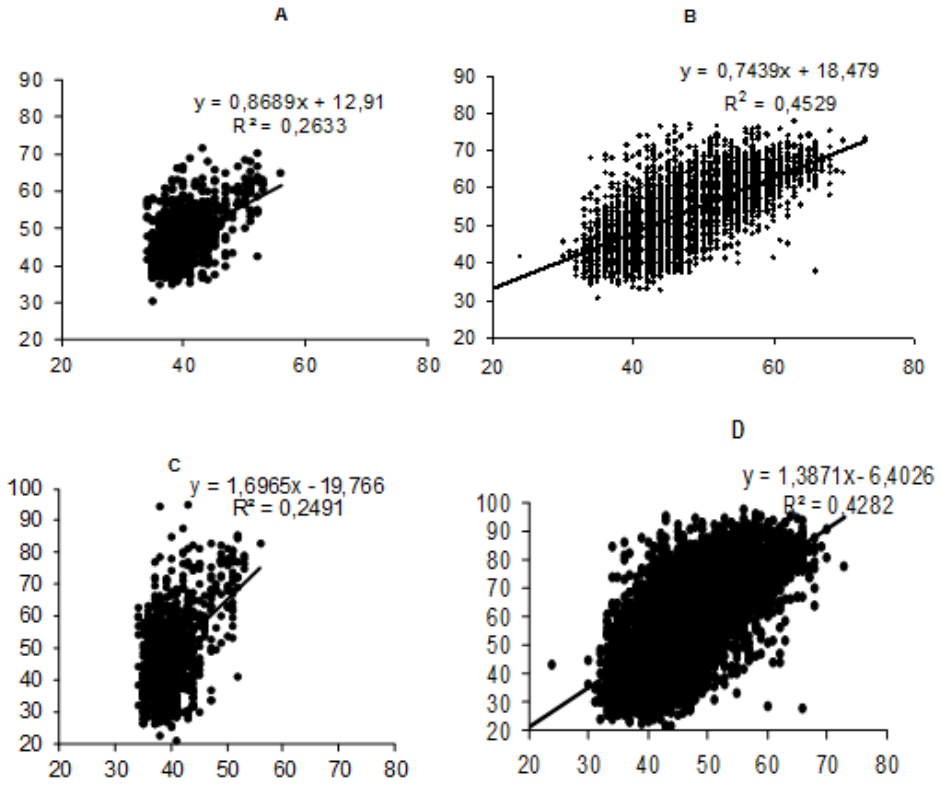


Figura 21. Análisis gráfico de las correlaciones entre los % de GC y GC_3 de los genes y el % de GC del superfragmento que los contiene. A) % de GC de genes vs % de GC para las secuencias ubicadas en superfragmentos homogéneos, B) igual que A, pero considerando fragmentos heterogéneos. C) % de GC_3 de genes vs % de GC de superfragmentos homogéneos, D) igual que C, pero considerando superfragmentos heterogéneos. En las 4 gráficas el eje X es el % de GC de los superfragmentos.

Si bien el genoma ha sido caracterizado en diferentes regiones respecto a su contenido en GC, la presente figura muestra una clasificación en dos espacios con características completamente diferentes, el espacio de regiones homogéneas y el de heterogéneas. Las regiones homogéneas pueden ser tanto ricas en GC como en AT.

Tabla 2.

	SGC Total	SGC Homogéneos	SGC Heterogéneos
GGC	0,48	0,26	0,52
GGC_3	0,45	0,24	0,50

Análisis de correlación entre los superfragmentos y los genes en ellos contenidos expresado como R^2. Se los analiza en forma total y por separado entre superfragmentos homogéneos y heterogéneos. (SGC = % de GC del superfragmento, GGC = % de GC del gene, GGC3 = % de GC_3 del gen.

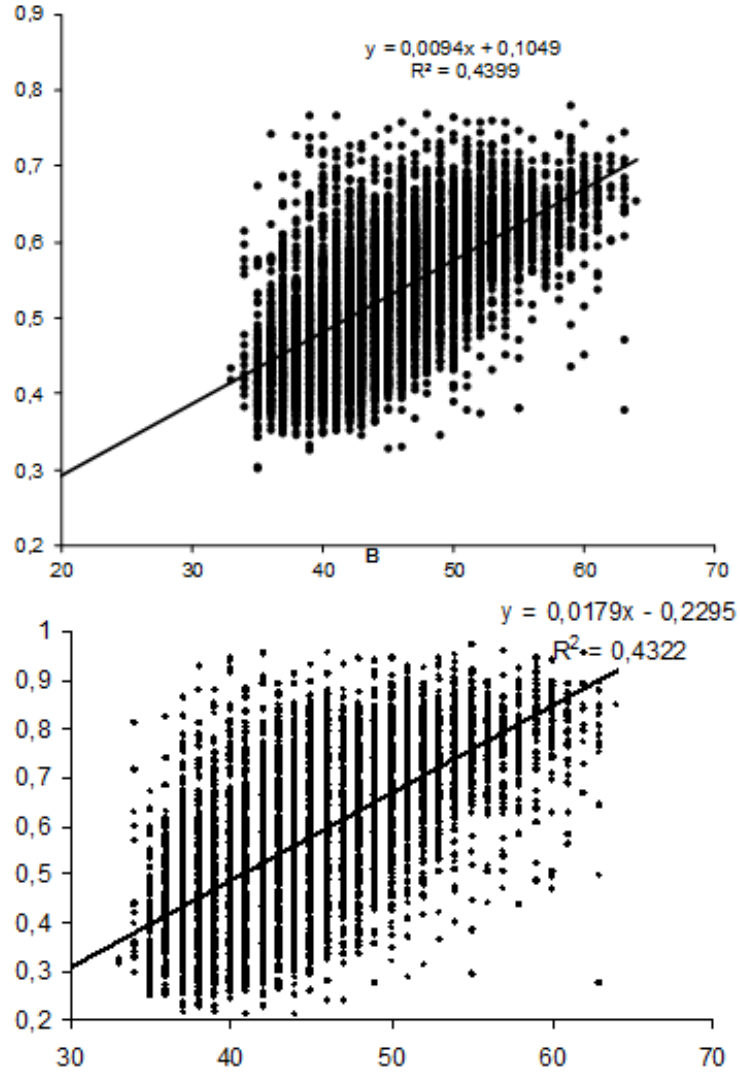


Figura 22. Análisis gráfico de correlación entre el % de GC (A) y GC_3 (B) de los genes y el % de GC del marco de 150 mb que los contiene.

Tabla 3.

	Correlación total
FGC	1
GGC	0,44
GGC3	0,42

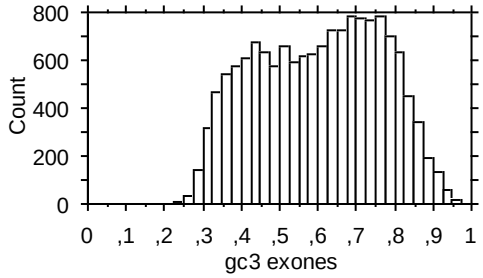
Análisis de correlación entre los % de GC y GC_3 de los genes y el marco de 300 mb que los contiene, para el caso de genes conocidos y nuevos. (FGG = % de GC del marco, GGC = % de GC del gene, GGC3 = % de GC_3 del gen expresado como R^2.

HETEROGENEIDAD EN CROMOSOMAS HUMANOS.

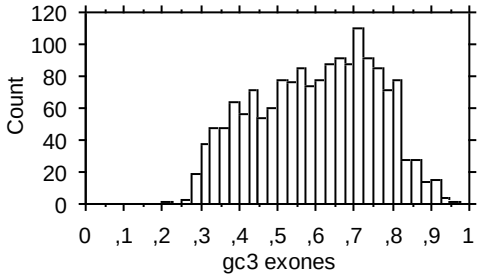
Trabajos anteriores han demostrado ampliamente la omnipresencia de la heterogeneidad en el contenido de GC a lo largo de las regiones del genoma humano en general, y de cada cromosoma en particular. En este sentido, decidimos tomar a los cromosomas completos como unidad de análisis. La figura 23 muestra las fuertes diferencias en la distribución de contenido en GC entre los cromosomas humanos. Estas diferencias se extienden desde el GC_3 (figura 23a) hasta el GC de fragmentos de 20 kb (figura 23b) (prueba de Kruskal-Wallis $p \approx 0$ en los 2 casos). Las diferencias en el contenido global de GC en los cromosomas son cuantitativamente importantes y cubren el 10% (38% - 48%), mientras que el valor medio de GC_3 por cromosoma varía de 49% a 72%. Además, la forma de las distribuciones varía de cromosoma a cromosoma, mostrando sesgos hacia la derecha o izquierda, distribución unimodal o bimodal. Curiosamente, mientras que varios cromosomas mostraron una distribución bimodal de contenido en GC_3 (figura 23a), sólo el cromosoma 19 mostró este comportamiento a nivel de fragmentos. (figura 23b). En la figura 23c, se muestra a todos los cromosomas juntos, marcados con colores diferentes.

a

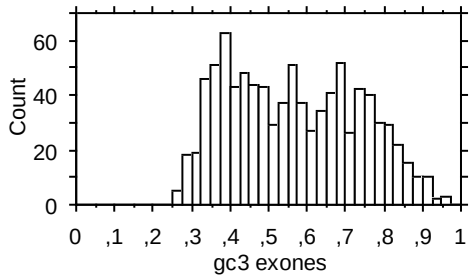
Total N: 14815



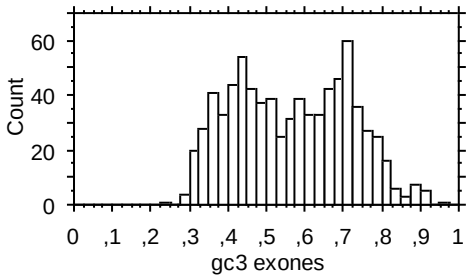
Cromosoma 1 N: 1638



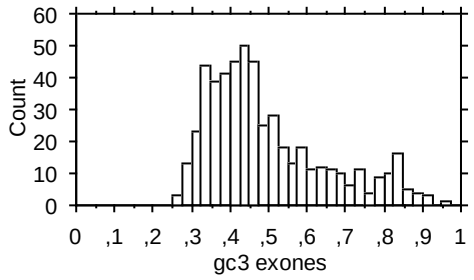
Cromosoma 2 N: 917



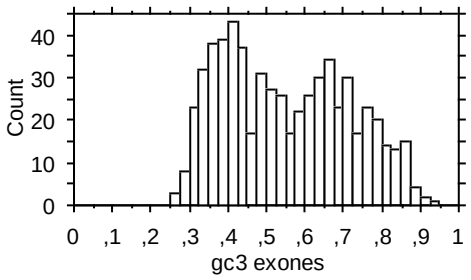
Cromosoma 3 N: 778



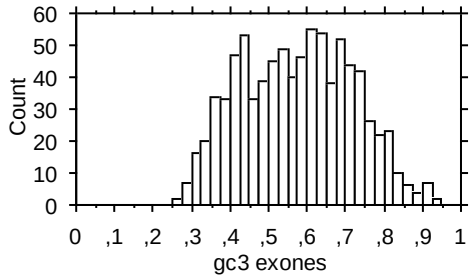
Cromosoma 4 N: 518



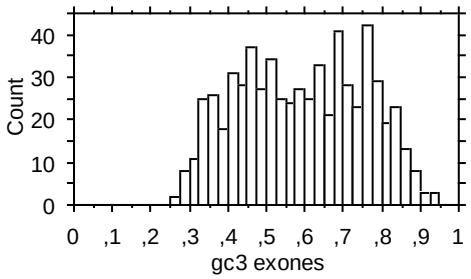
Cromosoma 5 N: 615



Cromosoma 6 N: 849

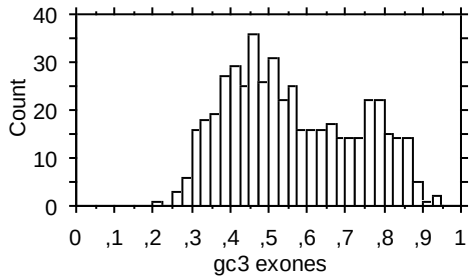


Cromosoma 7 N: 634

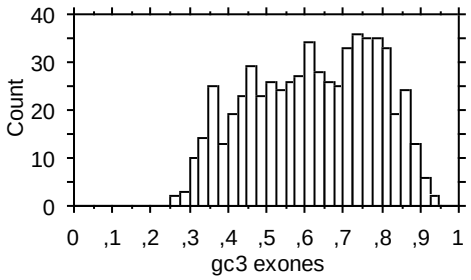


Cromosoma 8 N: 486

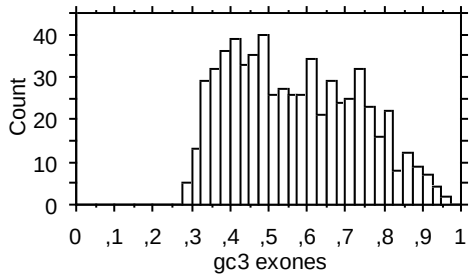
Cromosoma 9 N: 613



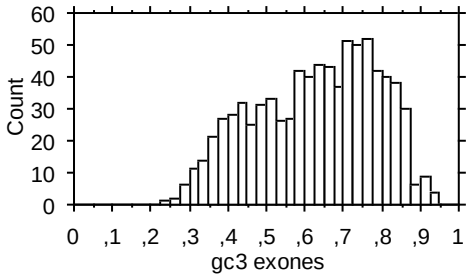
Cromosoma 10 N: 635



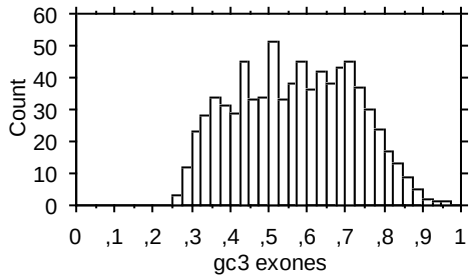
Cromosoma 11 N: 812



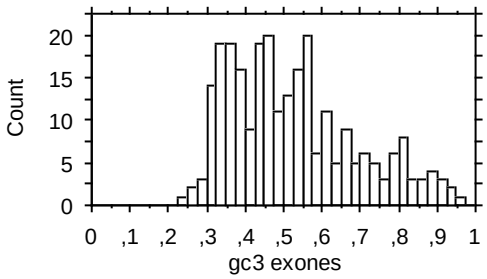
Cromosoma 12 N: 782



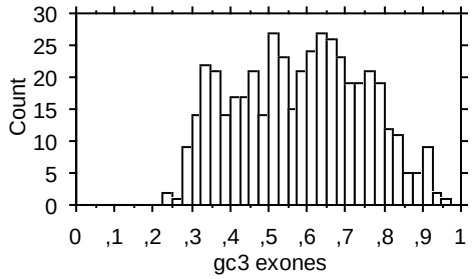
Cromosoma 13 N: 262



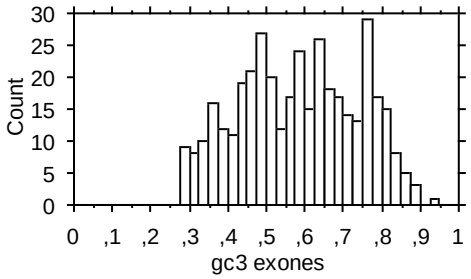
Cromosoma 14 N: 461



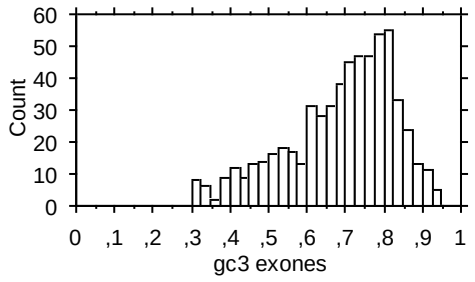
Cromosoma 15 N: 387

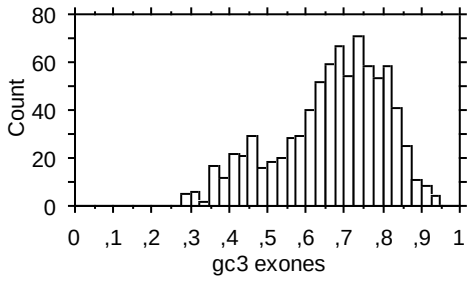


Cromosoma 16 N: 599



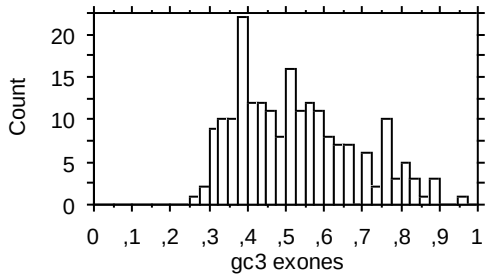
Cromosoma 17 N: 826

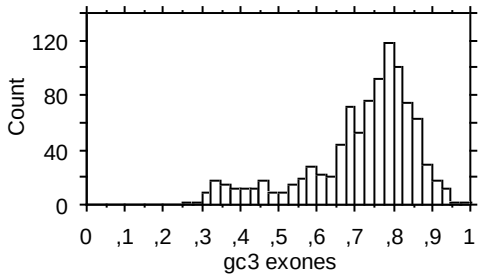




Cromosoma 18 N: 203

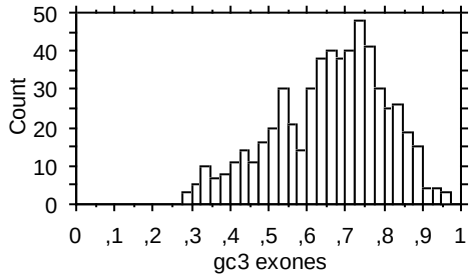
Cromosoma 19 N: 966

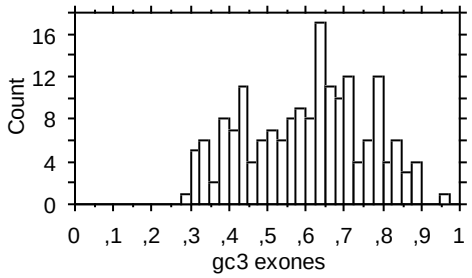




Cromosoma 20 N: 571

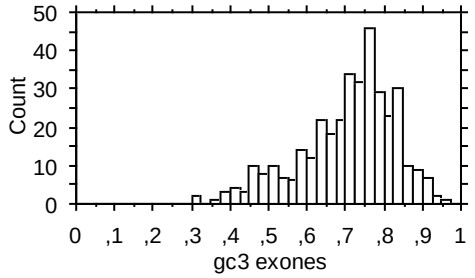
Cromosoma 21 N: 178

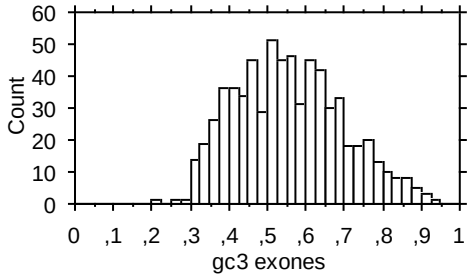




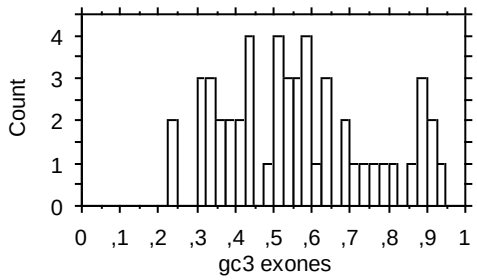
Cromosoma 22 N: 365

Cromosoma X N: 669



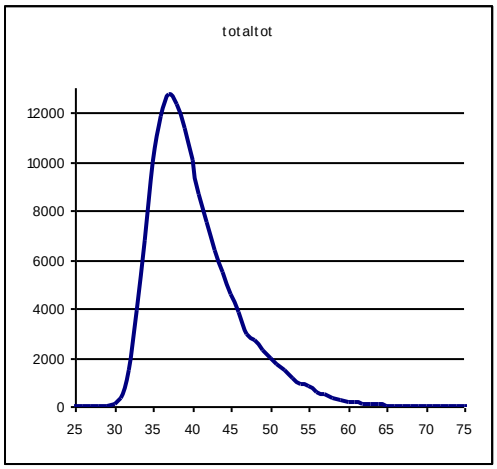


Cromosoma Y N: 51

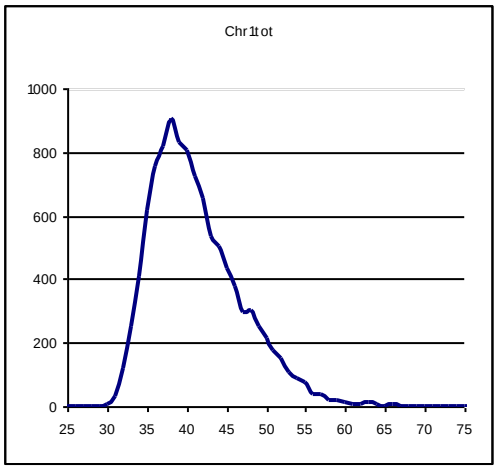


b)

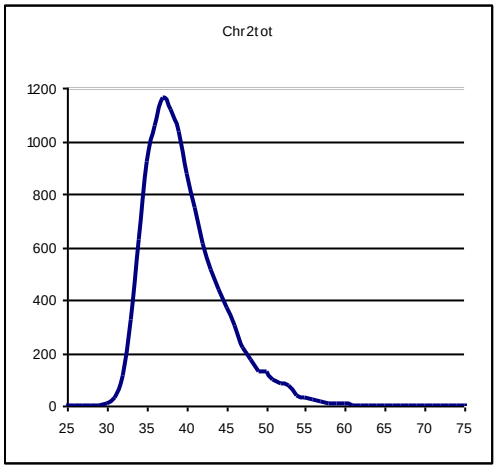
Total



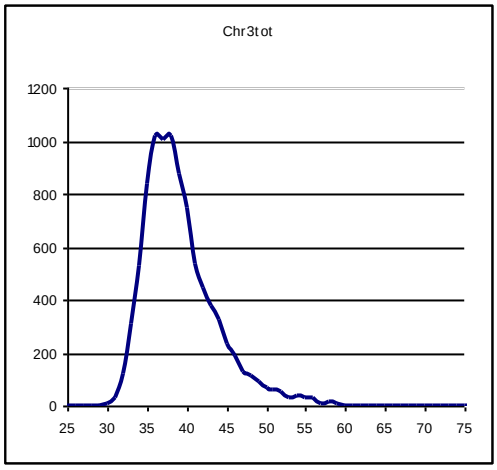
Chr 1



Chr 2

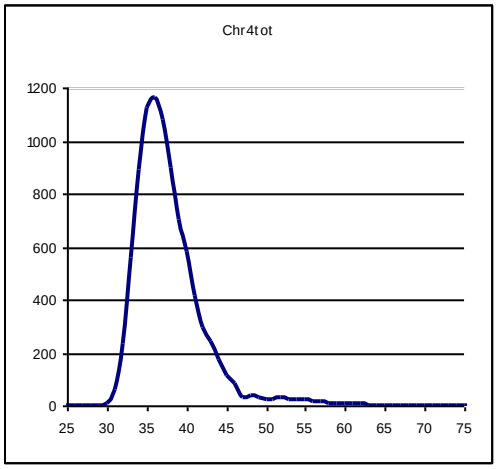


Chr 3

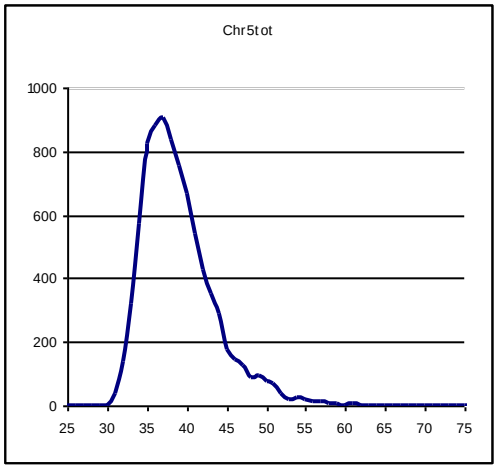


Chr 4

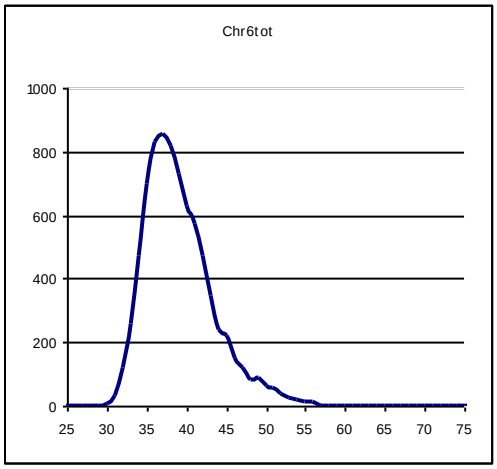
Chr 5



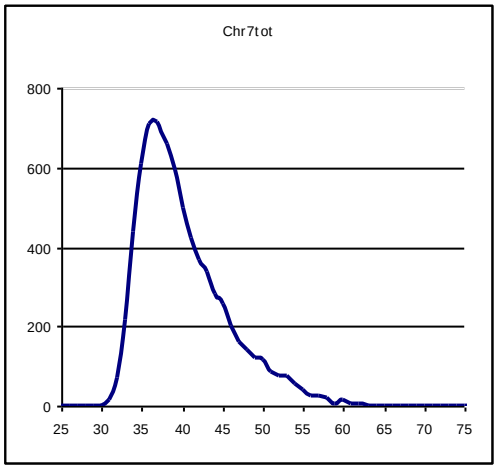
Chr 6



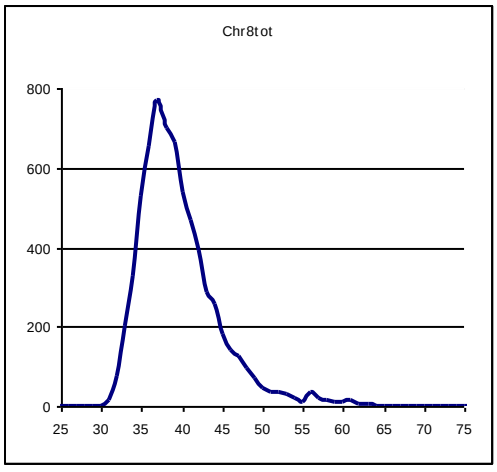
Chr 7



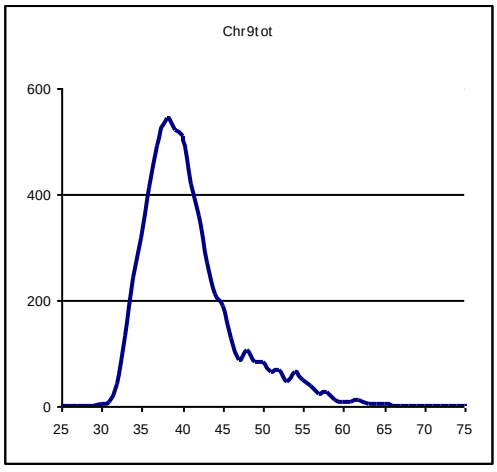
Chr 8



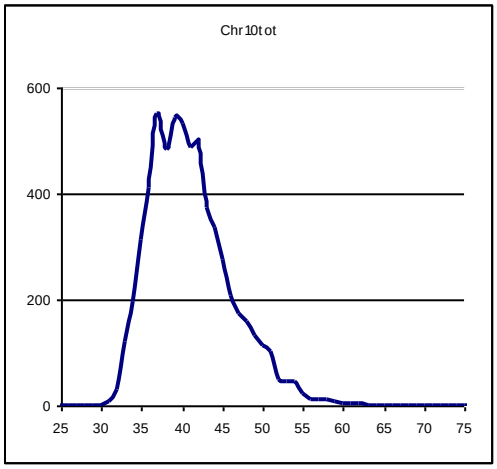
Chr 9



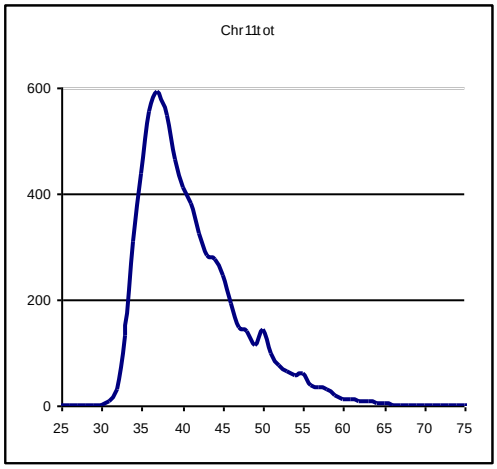
Chr 10



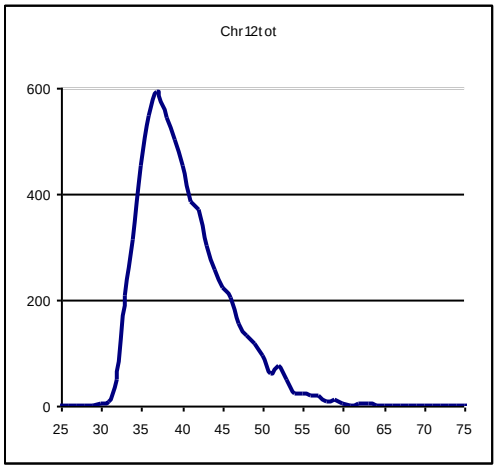
Chr 11



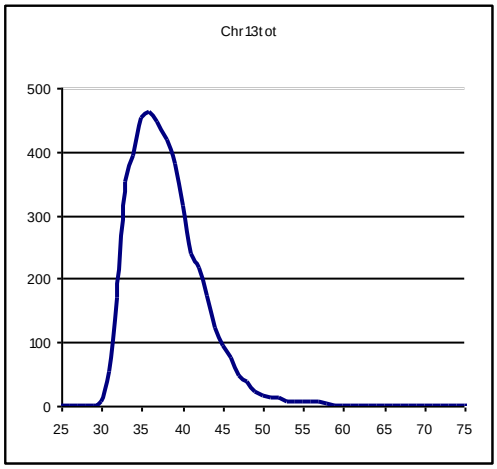
Chr 12



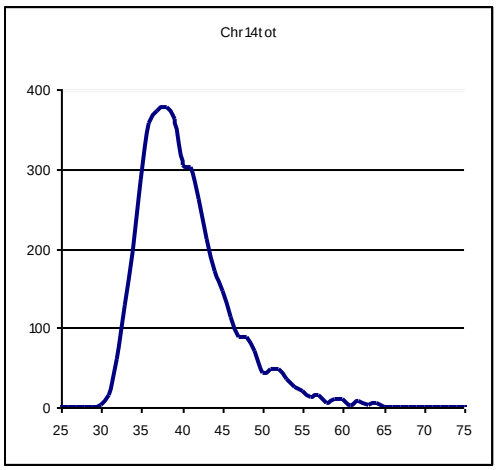
Chr 13



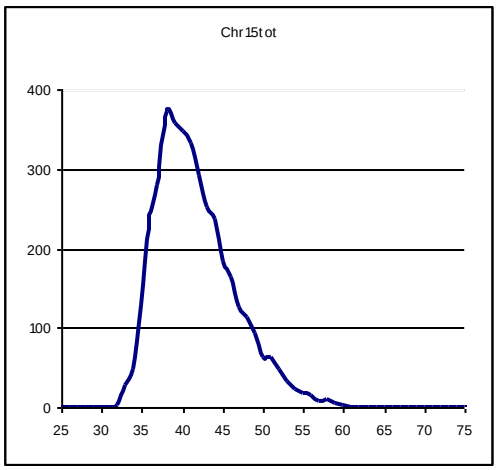
Chr 14



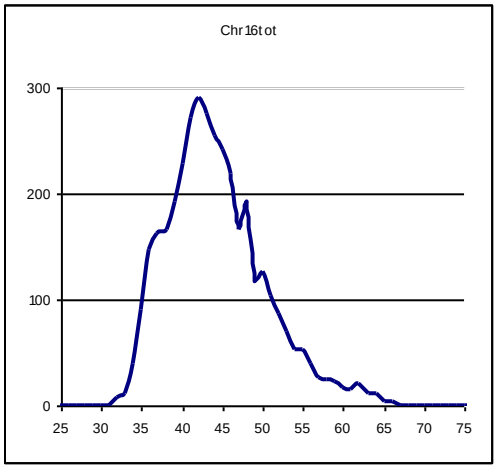
Chr 15

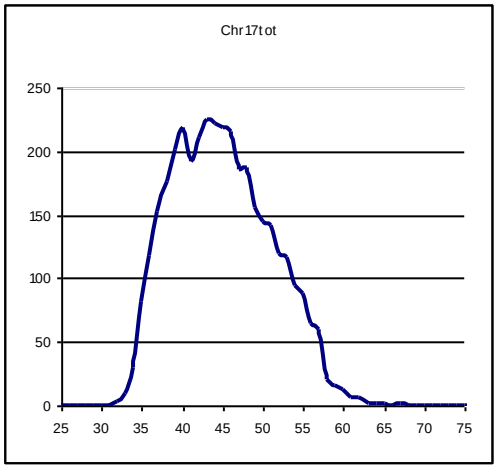


Chr 16



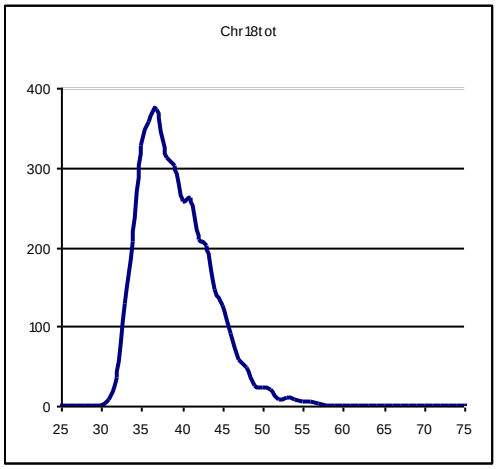
Chr 17

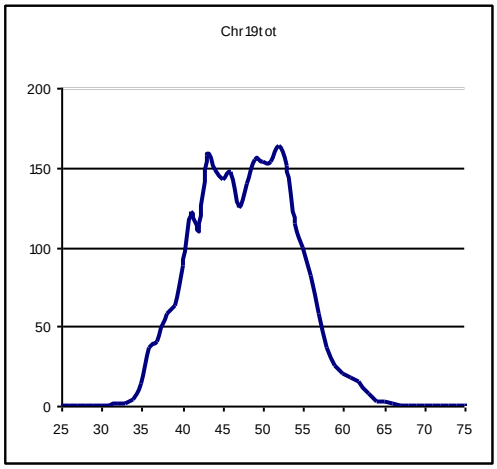




Chr 18

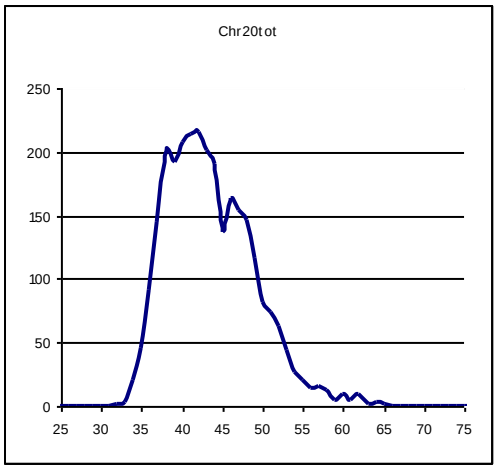
Chr 19

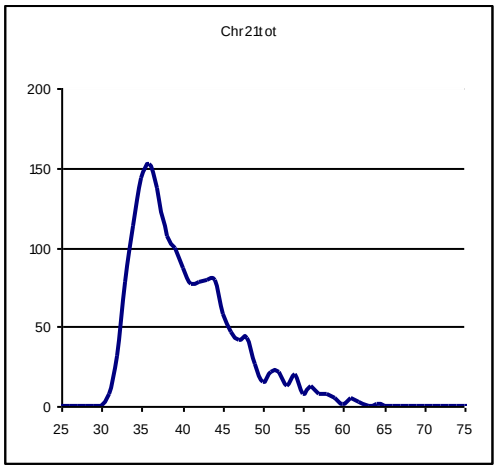




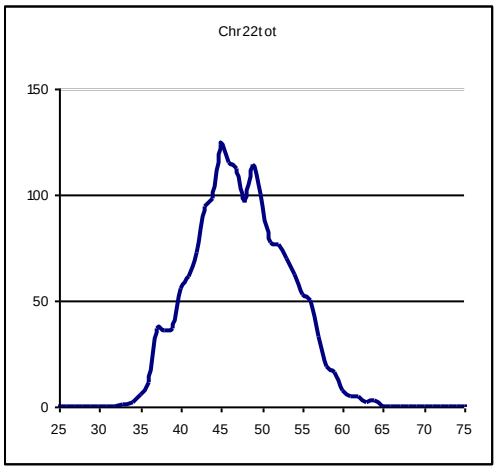
Chr 20

Chr 21

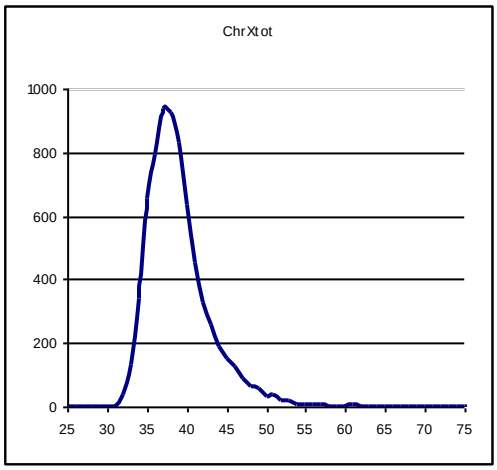




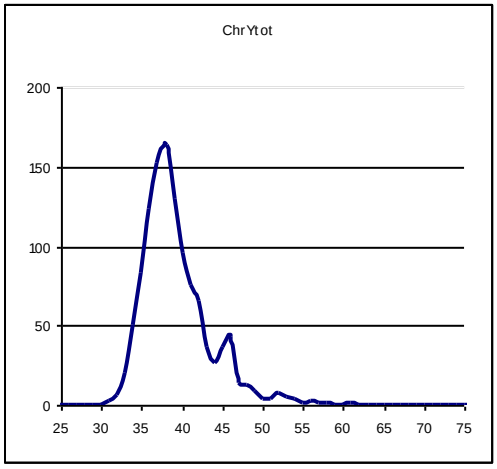
Chr 22



Chr X



Chr Y



c)

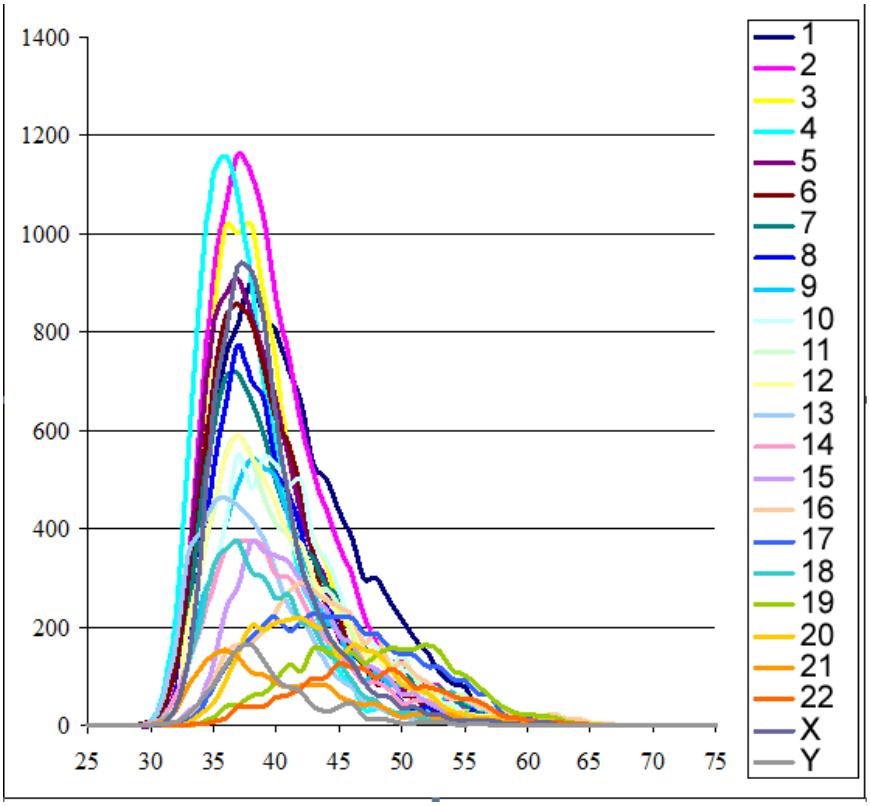


Figura 23 a: Distribución de GC3 en el genoma y en los cromosomas humanos, b: Distribución del contenido de GC en el genoma y los cromosomas humanos, considerando ventanas de 20 kb, c: Distribución del contenido GC de los cromsomas humanos.

Teniendo en cuenta las sorprendentes diferencias entre los cromosomas, éstos se agruparon en base a una medida de disimilitud entre distribuciones. La figura 24 muestra la agrupación de los mismos mediante la distancia HMI (ver métodos) correspondiente a fragmentos de 100 Kb. Dos grandes agrupaciones resultaron del análisis. En el primero se encuentran los cromosomas 1, 10-11, 15-17, 19-20 y 22, mientras que en el otro se encuentran los cromosomas 2-9, 12-14, 18, 21 y X. Si bien los cromosomas pertenecientes a estos grupos difieren fuertemente en el contenido en GC global (44% *vs*. 40%, Kruskal-Wallis $p<1e-4$), la ordinalidad de los miembros (inversamente correlacionada con el tamaño de los cromosomas) aparece como claramente no aleatoria. Un panorama conceptualmente similar se obtiene cuando la agrupación se basa en GCex o GC_3. Otras especies muestran patrones similares, a pesar de algunas idiosincrasias.

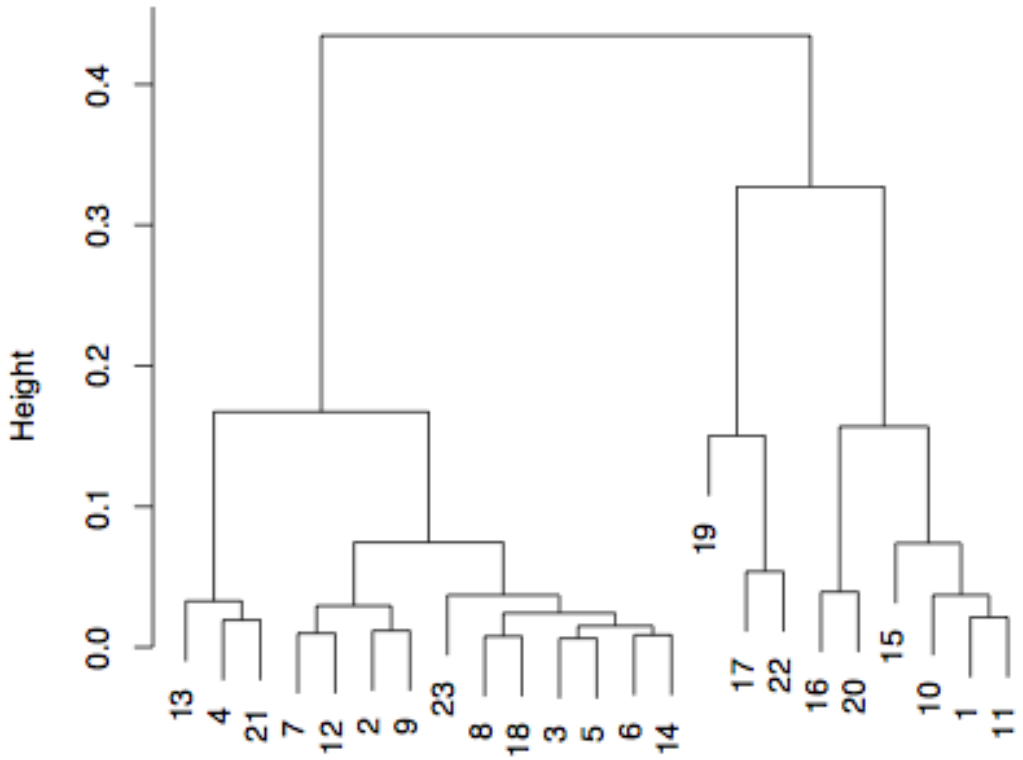


Figura 24: Dendograma de los cromosomas humanos (basado en distancia HMI entre distribución de contenido de GC en fragmentos de 100Kb).

Estas diferencias de composición de los cromosomas humanos no son un resultado menor, ya que sugieren la existencia de posibles aspectos estructurales y funcionales subyacentes. Este resultado, junto con los ya descritos, sugieren claramente que los cromosomas humanos, y sin duda de otros mamíferos, podrían tener historias evolutivas diferentes, las que tienen como consecuencia a) distinto nivel de heterogeneidad vs. homogeneidad composicional, b) genes con distintas propiedades (entre ellas, el uso de codones y de aminoácidos), c) composiciones (GC%) medias y distribuciones de bases diferentes. Estas tres características son particularmente evidentes en mamíferos, pero no en otras especies, ni siquiera de vertebrados. Para obtener más información se llevaron a cabo los siguientes análisis.

Aunque algunas correlaciones composicionales en el genoma humano ya se han descrito, no ha sido reportado hasta la actualidad la muy fuerte correlación entre el contenido de GC medio de los cromosomas y el contenido de GC medio de los genes que se encuentran en ellos. No existen razones obvias para suponer dicha correlación, sobre todo dado que las regiones génicas (incluyendo intrones) aproximadamente representan un 20% del genoma. Sin embargo el contenido en GC de los cromosomas no sólo se correlaciona fuertemente con el contenido en GC de las zonas codificantes (R^2=0,92, p<1-13), sino también con GC_1, GC_2 y GC_3 (figura 25) (R^2=0,81, p<1e-8; R^2=0,54, p<1e-4; R^2=0,85, p<1e-11) de los genes y aun con el contenido en GC de las regiones intrónicas (figura 26) (R^2=0,85, p<1e-11). Es más, cuando cada gen se considera en forma separada, la correlación entre el contenido en GC de los intrones y el de los exones es altamente significativo (R^2= 0,98, p<1e-15). La correlación del contenido en GC de ambas regiones (exónicas e intrónicas) y la que existe entre el GC_3 de las regiones exónicas con las regiones genómicas flanqueantes de cada gen (Tabla 4 y figura 27) resultan también muy significativas, decreciendo lentamente a medida que aumenta la distancia hacia el gen (figura 28). Es significativo que la correlación es más alta con la región flanqueante 5' que con la 3', a pesar de que en la primera existen las islas CpG y la mayoría de los elementos reguladores de la transcripción. Estas correlaciones en su conjunto podrían sugerir un efecto dirigido por los genes en la composición de las regiones cromosómicas, refundando algunos aspectos básicos de la teoría de isocoros planteada por Bernardi y sus colaboradores (Bernardi, 2004). Una correlación significativa y positiva adicional involucra al contenido en GC de los genes (o alguna de sus variantes) con la densidad génica. En la figura 29 se muestra la correlación a nivel de cromosomas (p<1e-6). Además, los cromosomas que pertenecen al agrupamiento de alto nivel de GC (figura 24) son significativamente más densos que su contraparte con GC bajo (Kruskal-Wallis p<1e-3). Esta última correlación se mantiene dentro de los cromosomas cuando consideramos fragmentos de 100Kb. Más aún, el coeficiente de correlación por cromosoma se incrementa con el contenido en GC de estos, lo cual sería un posible subproducto de la disminución del tamaño de los fragmentos sin genes en los cromosomas más cortos ya que el tamaño codificante (suma de tamaño de todos los exones en un cromosoma) está altamente correlacionado con el tamaño del cromosoma (figura 30). Finalmente, el tamaño promedio de los genes por cromosoma decrece claramente con el incremento en el GC promedio de las secuencias codificantes (aproximadamente -4Kb por cada 1% de incremento en GC, R^2=0,68 p<1e-6), también puede verse un efecto de decrecimiento del tamaño del cromosoma a medida que aumenta su composición en GC (figuras 31a y b). Como se ha descrito anteriormente [ver (Bernardi 2004) y

las referencias en ese trabajo], el panorama global muestra regiones codificantes cromosómicas más compactas a medida que aumenta el GC; aquí se muestra que esta característica está relacionada con la arquitectura de cada cromosoma.

En suma, esto indica que los cromosomas son una unidad composicional coherente y los genes que los integran parecen tener un papel importante en determinar el carácter distintivo de cada cromosoma.

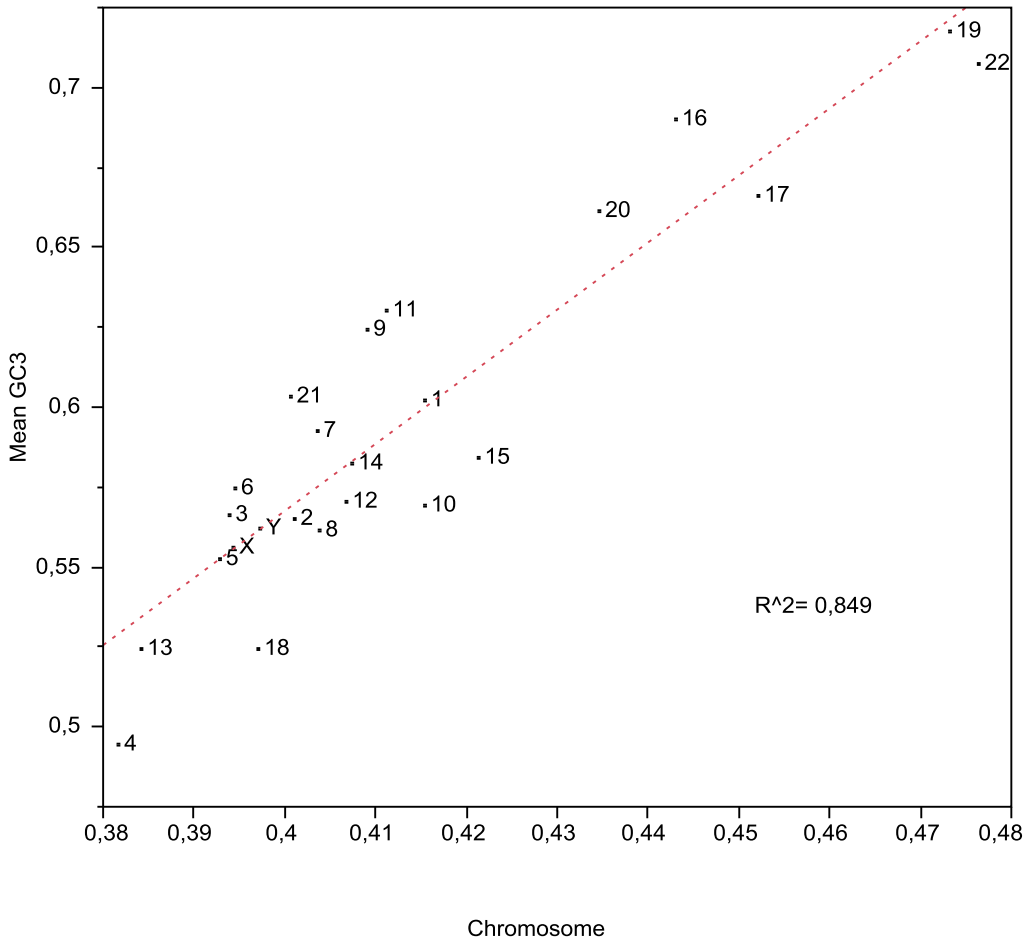


Figura 25: Correlación entre el contenido en GC de los cromosomas y el contenido en GC_3 medio de los genes contenidos en cada uno de ellos.

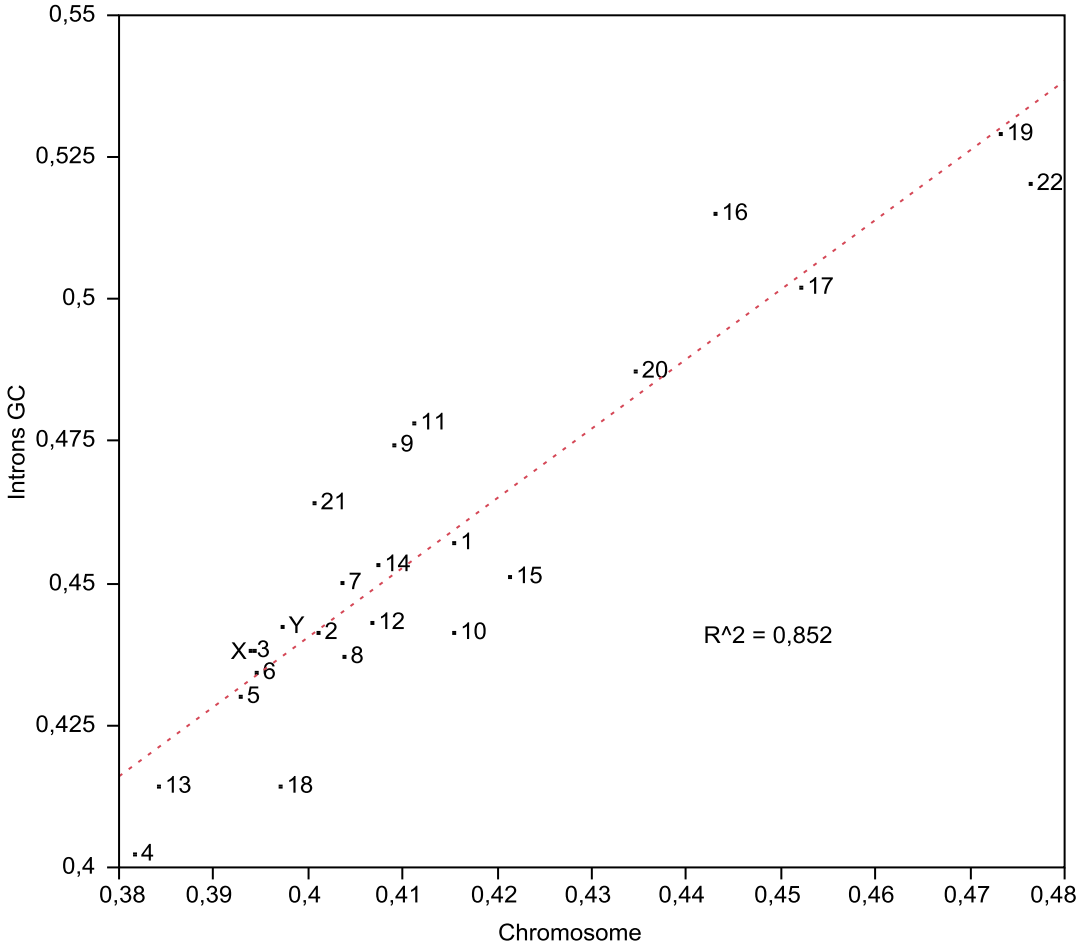


Figura 26: Correlación entre el contenido GC del cromosoma y el contenido medio de GC de los intrones de los genes que contienen.

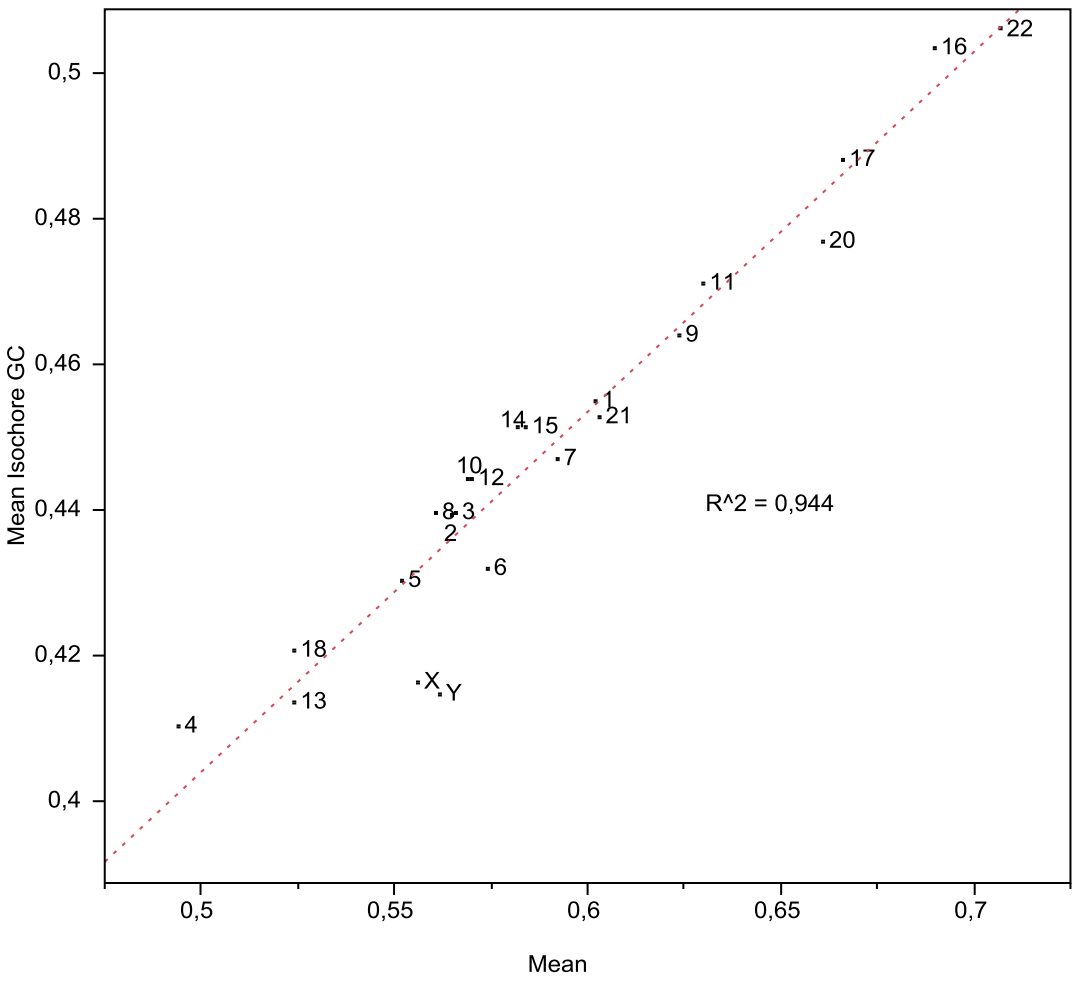


Figura 27: Correlación entre el contenido GC medio de las regiones flanqueantes (25 kb corriente arriba del codón de iniciación + 25 kb corriente abajo del codón de terminación) y el GC_3 medio de los genes contenidos.

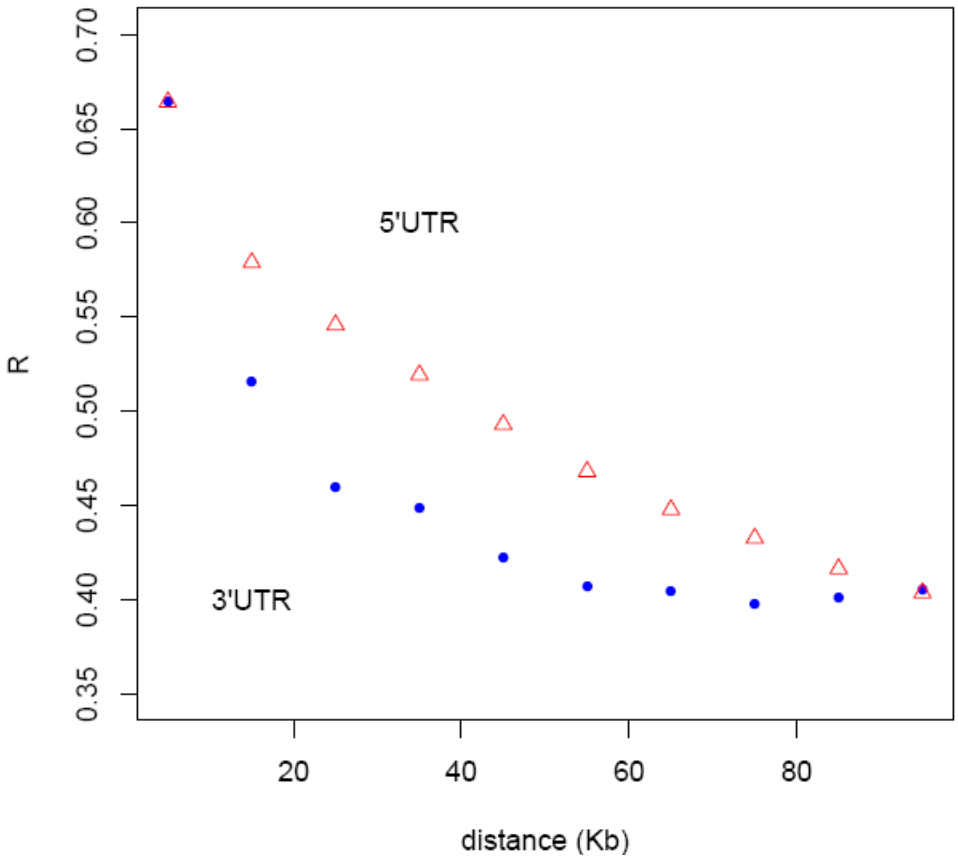


Figura 28: Correlación de GC de zonas flanqueantes (25 kb 5’ utr y 3’ utr) con el GC del gen a distancias incrementales.

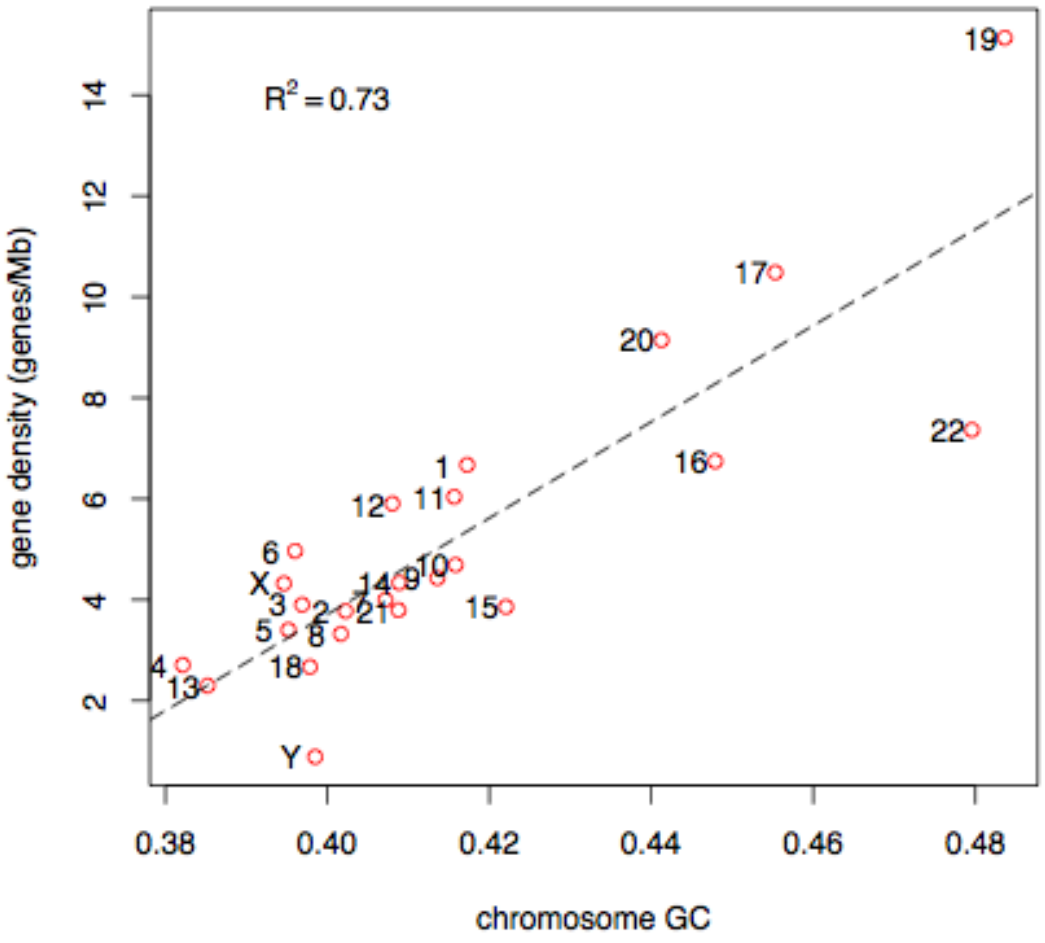


Figura 29: Correlación entre el contenido en GC de los cromosomas y la densidad génica (número de genes por Mb).

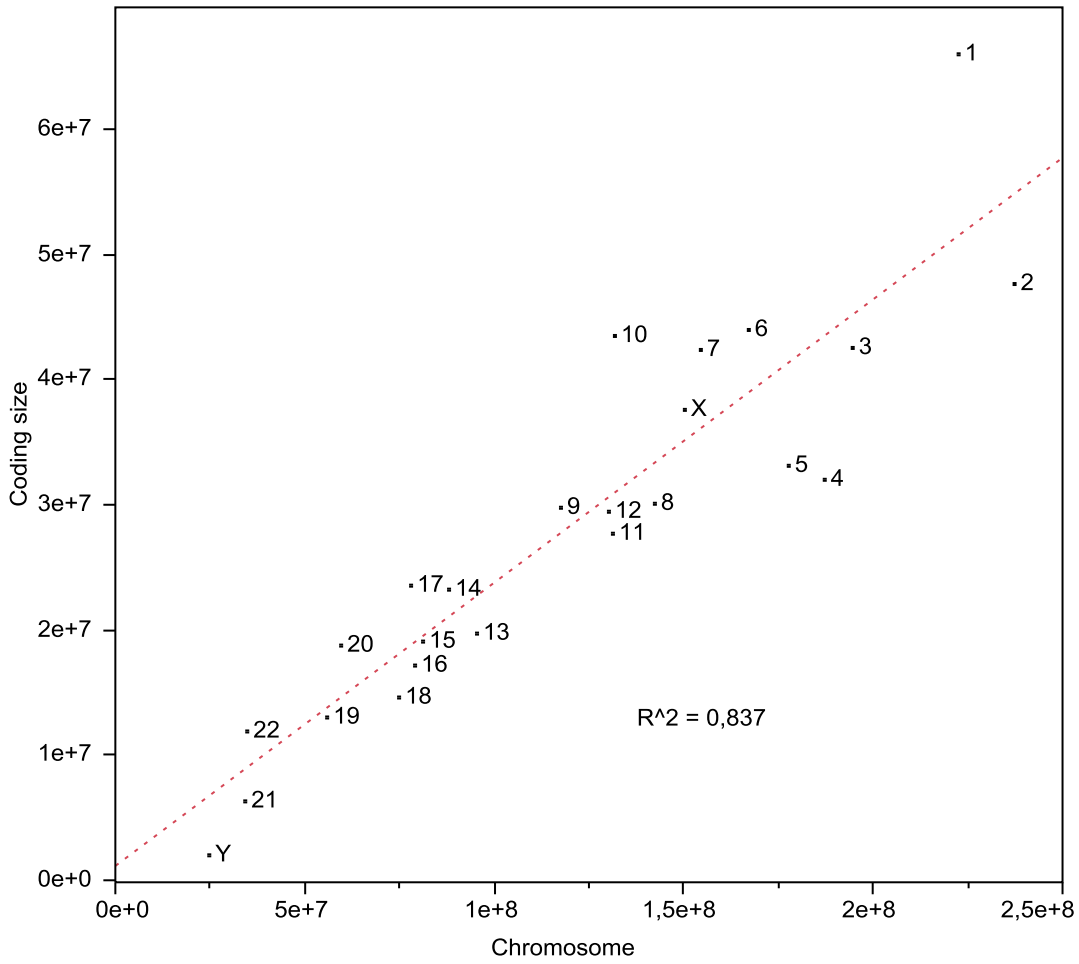
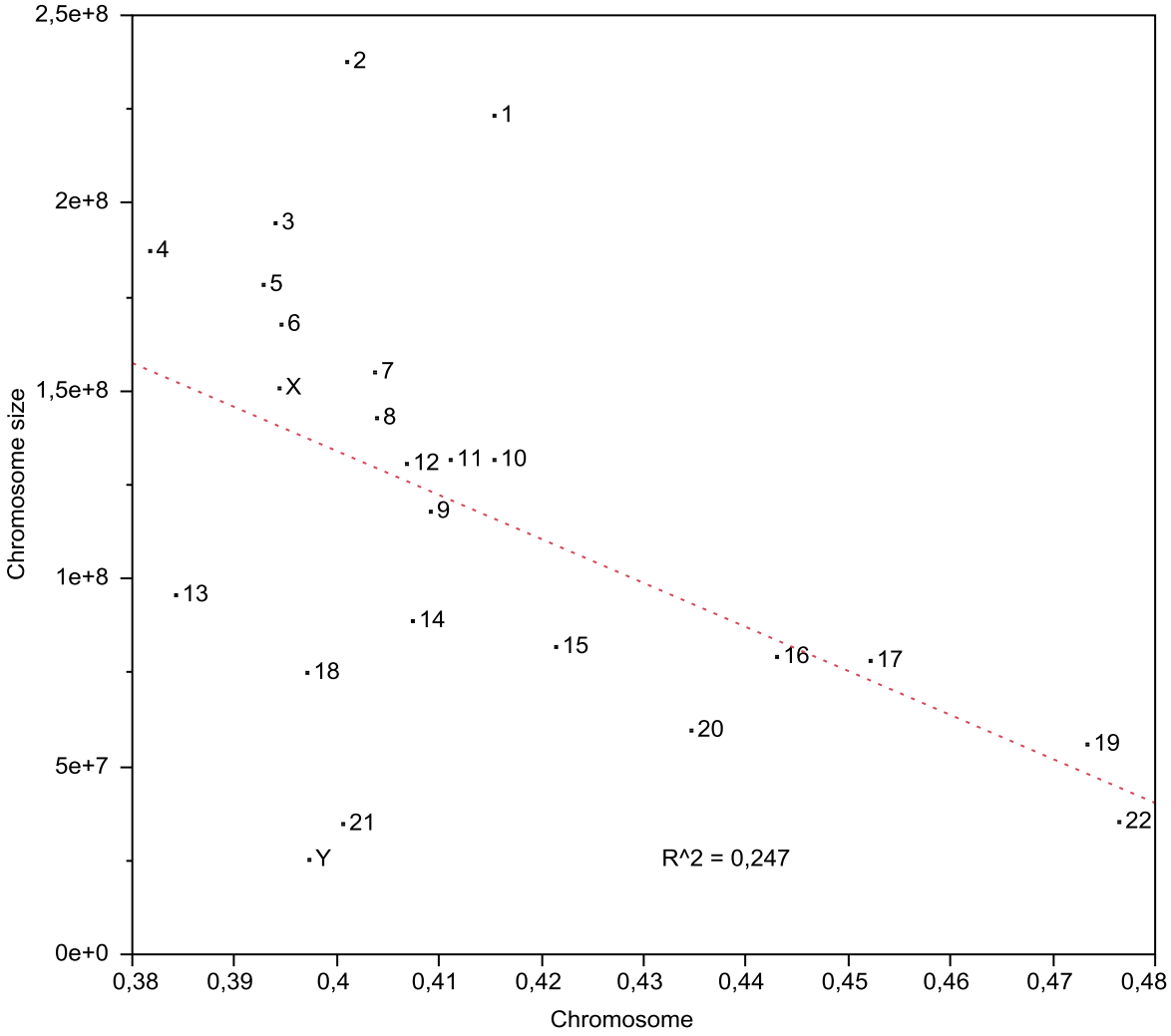


Figura 30: Correlación entre el tamaño de los cromosomas y el tamaño codificante de cada uno (medido como la suma de las bases en los exones en los respectivos cromosomas).


2,5e+8
2e+8
1,5e+8
1e+8
5e+7
0e+0
Chromosome size
0,38
0,39
0,4
0,41
0,42
0,43
0,44
0,45
0,46
0,47
0,48
Chromosome
2
1
3
4
5
6
7
X
8
12
11
10
9
13
14
15
16
17
18
20
19
21
22
Y
R^2 = 0,247

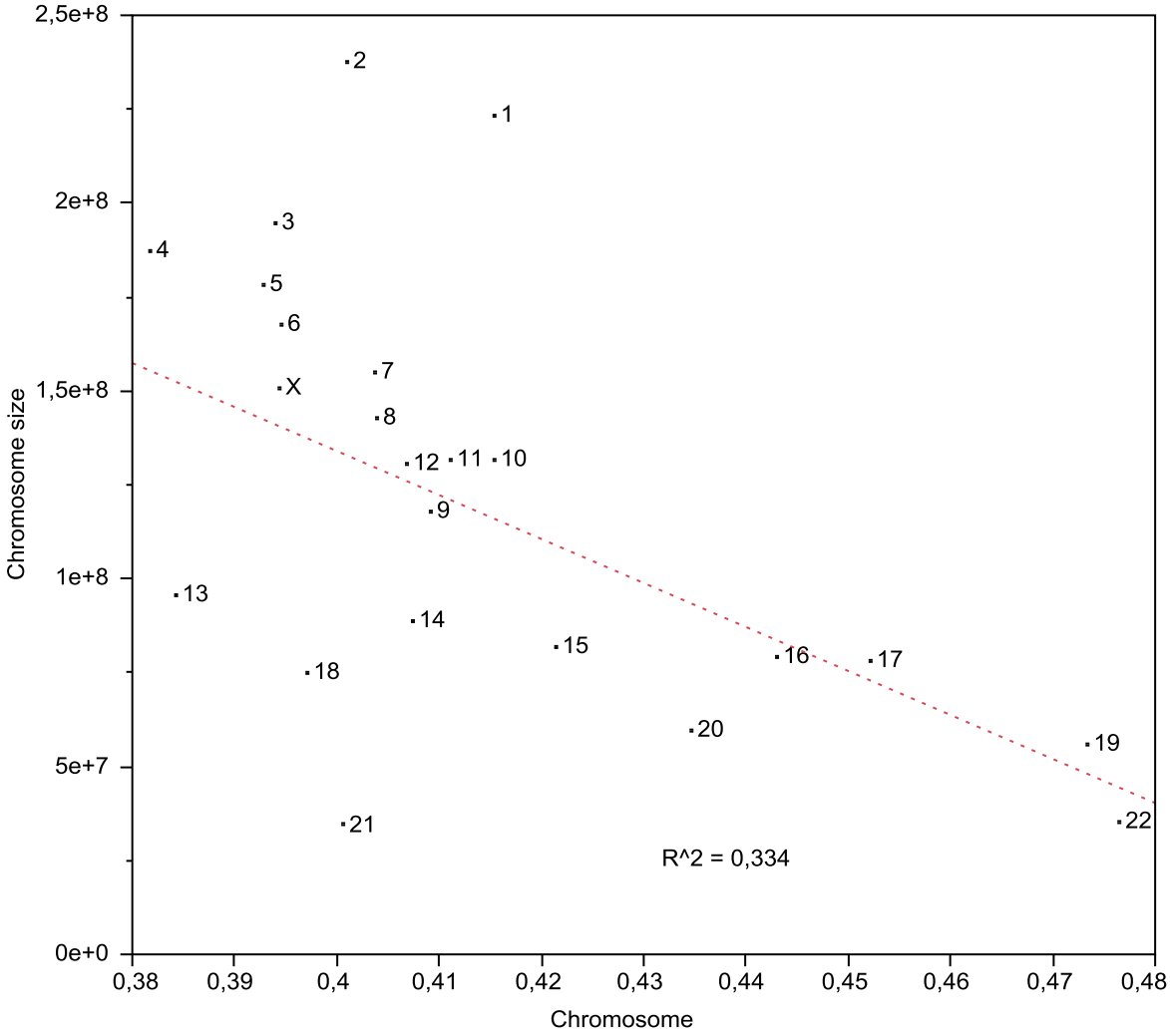


Figura 31: Correlación entre el contenido en GC de los cromosomas y sus respectivos tamaños. a: considerando el cromosoma Y y b: sin considerar al cromosoma Y.

Tabla 4.

	GC exones	GC intrones
marco 5' 100Kb	0,43	0,59
marco 5' 75Kb	0,44	0,61
marco 5' 50Kb	0,46	0,62
marco 5' 25Kb	0,48	0,64
marco 5' 10Kb	0,48	0,62
GC_3	0,85	0,65
marco 3' 10Kb	0,48	0,62
marco 3' 25Kb	0,42	0,57
marco 3' 50Kb	0,39	0,53
marco 3' 75Kb	0,37	0,52
marco 3' 100Kb	0,36	0,51

Correlaciones composicionales a nivel génico.

SOBRE LOS ISOCOROS Y SUS FAMILIAS

Desde la publicación de la secuencia del genoma humano, se ha dado una gran controversia sobre la existencia de isocoros y sus familias. Se han desarrollado una batería importante de programas que inspeccionan la secuencia en busca de patrones composicionales a gran escala basados en diferentes principios físicos y matemáticos (Oliver et al. 2001; Oliver et al. 2002; Wen y Zhang 2003; Zhang y Zhang 2003; Zhang y Zhang 2004). Si tomamos en cuenta las regiones génicas y correlacionamos el contenido en GC de los genes con los genes siguientes en el cromosoma, el índice de correlación es muy grande. Cuando se analiza la autocorrelación entre los genes para los distintos cromosomas desfasándolos en forma creciente, se pueden observar zonas de autocorrelación muy grande en algunos cromosomas (figura 32) Estas distribuciones nos retrotraen a la imagen original de Bernardi del genoma como un mosaico de zonas de contenido de GC constante y podrian estar dándonos un indicio sobre la estructura de los isocoros. Además, si consideramos los distintos niveles de GC, se pueden observar varias agrupaciones de genes con perfiles distintivos de GC que podrían representar las familias de isocoros visibles a través de la centrifugación analítica (Bernardi 2000).

Inspeccionando dentro de las características génicas antedichas, el contenido en GC cromosómico se correlaciona fuertemente con el GC de los exones y de los intrones (figura 25 y 26). Además, cuando cada gen se considera por separado, la correlación entre el GC de ambas regiones es muy significativa. A su vez, las correlaciones entre ambas con las regiones génicas anteriores y posteriores al gen también son muy importantes, disminuyendo lentamente a medida que la distancia de la secuencia circundante al gen aumenta. Estos resultados apuntan hacia un efecto dirigido por los genes en la composición de las regiones cromosómicas. En otras palabras, la estructura en isocoros del genoma de los mamíferos podría ser el resultado de efectos todavía no claros actuando sobre la composición de los genes, disminuyendo a medida que la distancia desde las unidades de transcripción aumenta.

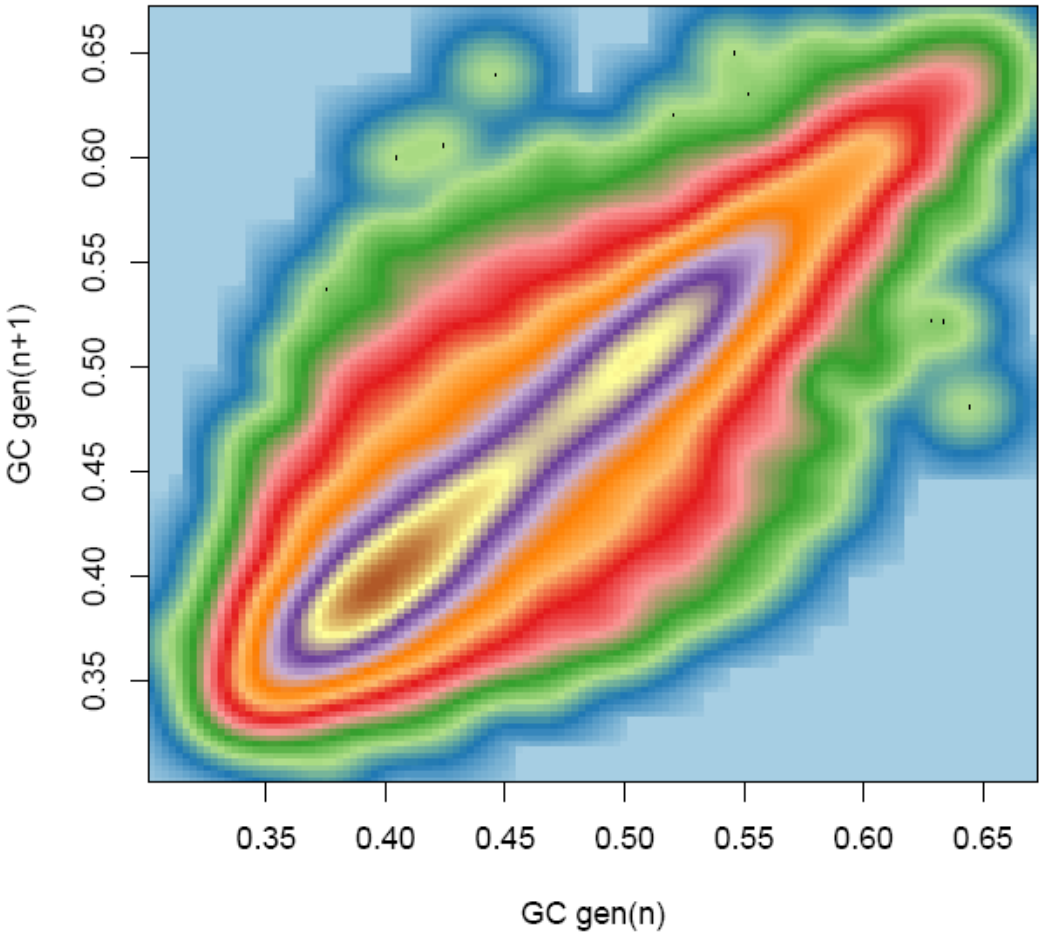


Figura 32: Correlacion entre el contenido GC de los genes y los genes siguientes en el cromosoma, los diferentes colores explican niveles de densidad difereneciales en los puntos desde el nivel más bajo de densidad (celeste) hasta el nivel más alto (marrón).

A nivel genómico, la variación en el uso de aminoácidos está fundamentalmente asociada con el contenido en GC y el uso de cisteína (D'Onofrio et al. 1991), Sabbia et al., 2007; ver más abajo). Llamativamente, los mismos factores modulan las diferencias entre los cromosomas. Así, el primer eje del análisis de correspondencia en el uso de aminoácidos por cromosoma se correlaciona fuertemente con el promedio del contenido en GC de las regiones codificantes (R^2 = 0,87, p < 1e-10) pero además con el contenido en GC de los cromosomas (R^2 = 0,81, p< 1e-8). El segundo eje se correlaciona con la proporción de cisteína en las secuencias que codifica cada cromosoma (R^2 = 0,45, p <5e-4). El cambio en el contenido en GC lleva a un cambio pequeño pero estadísticamente significativo en la frecuencia de aminoácidos codificados totalmente por GC o AT en la primera y segunda posición de los codones. Además, los cromosomas tienen una pequeña pero significativa diferencia en la fracción de residuos hidrofóbicos (Kruskal-Wallis p < 5e-4), que podrían indicar una posible diferencia en la clase de genes que son codificados por los diferentes cromosomas.

Las diferencias de frecuencias para los distintos dinucleótidos están fundamentalmente asociadas con el contenido en GC de los cromosomas. En un análisis de componentes principales, el primer eje explica más del 99% de la varianza y además se correlaciona muy significativamente con el contenido en GC (R^2 = 0,99, p < 1e-15). Llamativamente, cuando se agrupan los cromosomas basados en estas frecuencias, los cromosomas pequeños y ricos en GC (16, 17, 19, 20 y 22) se agrupan consistentemente sin importar el método de agrupamiento (“complete”, “ward” y “average”).

DISTANCIAS DE LOS CROMOSOMAS HACIA EL CENTRO DEL NÚCLEO

En las últimas décadas, se han logrado notables avances en el conocimiento de la sub-estructura nuclear y su organización, dando lugar a una nueva visión de la relación estructura - función del núcleo, en particular con los procesos genéticos fundamentales (Bolzer et al. 2005). Recientemente, estos mismos autores y otros grupos reportaron una correlación significativa (R^2 = 0,81, p < 1e-8) entre una medida resumen de la distancia al centro del núcleo y el tamaño de los cromosomas para el genoma humano (Bolzer et al. 2005; Goetze et al. 2007) (figura 35). En estos, se ha encontrado una correlación negativa. Como era de esperar de

los resultados mencionados, una correlación significativa marginal existe entre la distancia al centro del núcleo y el contenido en GC (R^2 = -0,18, p < 0,03). Sin embargo, como se muestra en la figura 33, si el cromosoma Y no se considera esta correlación aumenta sustancialmente (R^2 = -0,33, p < 0,003). Cuando los cromosomas se agrupan en base en su contenido en GC (figura 24) la mediana de las distancias al centro del núcleo en los dos principales grupos son ligeramente diferentes (Kruskal-Wallis p < 0,05).

Se ha reportado que regiones ricas en GC y en genes se agrupan cerca del centro del núcleo, mientras que las más pobres en genes y GC se localizan en la periferia. Esta triple asociación de tamaño cromosómico, distancia al centro del núcleo y contenido en GC es indicativo de un vínculo entre las propiedades composicionales, características estructurales de los cromosomas enteros y su organización en el núcleo. En la figura 33 se puede observar la correlación entre la distancia de los cromosomas al centro del núcleo y el tamaño del cromosoma (R^2 = 0,81, p < 1e-8) También la distancia al centro del núcleo se encuentra altamente correlacionada con la densidad génica (figura 35). Cuando los cromosomas son agrupados en base a su contenido en GC (figura 24) la mediana de las distancias al centro del núcleo en los dos grupos principales es mínimamente diferente (Kruskal-Wallis p < 0,05).

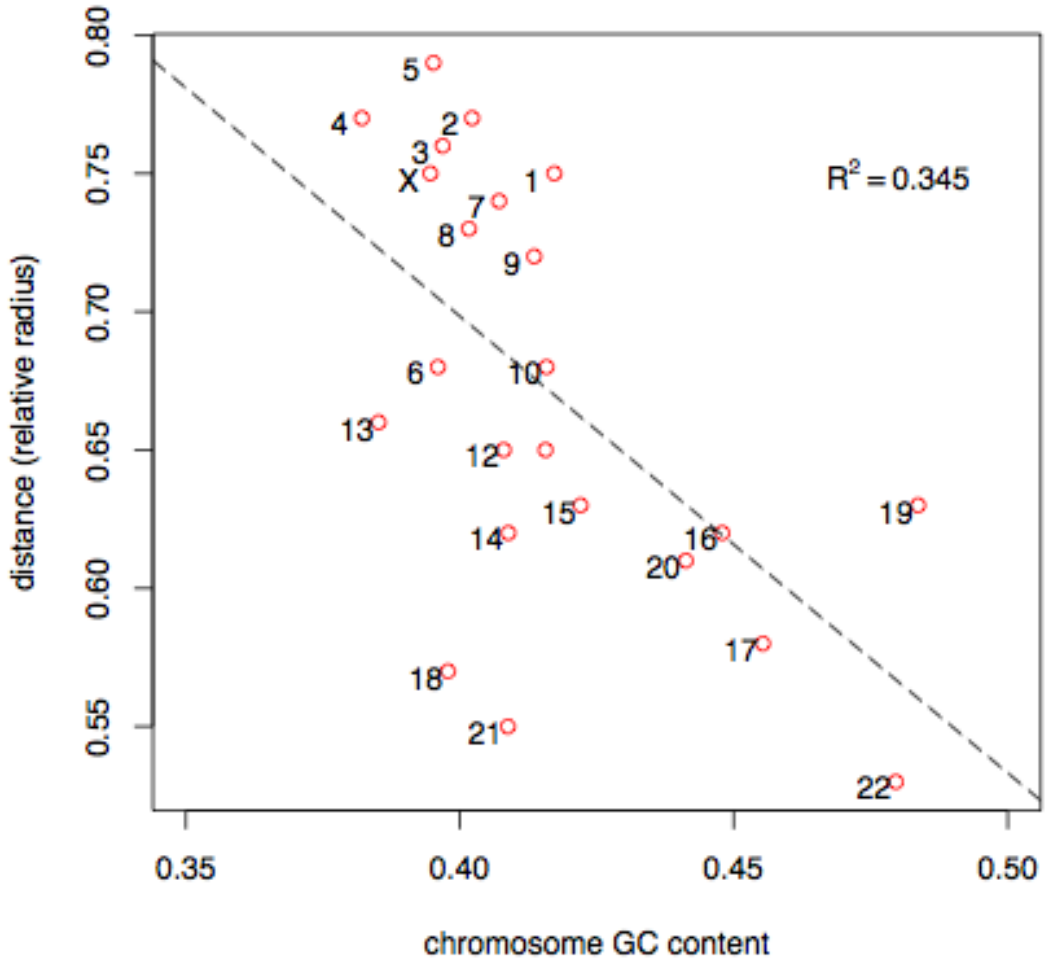


Figura 33: Correlación entre el contenido en GC de los cromosomas y su distancia al centro del núcleo, excluyendo el cromosoma Y.

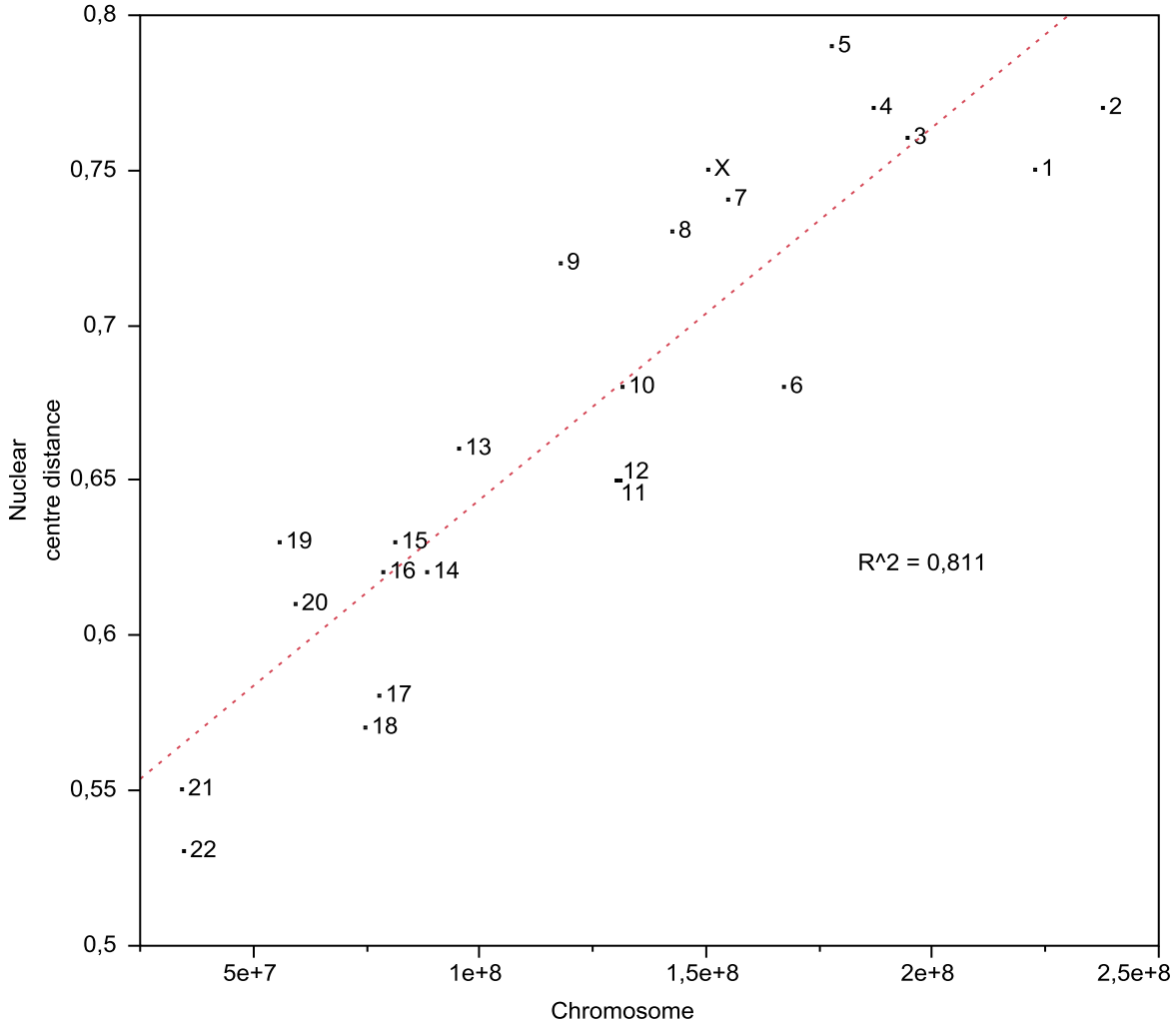


Figura 34: Correlación entre la distancia del cromosoma al centro del núcleo y el tamaño de cada cromosoma.

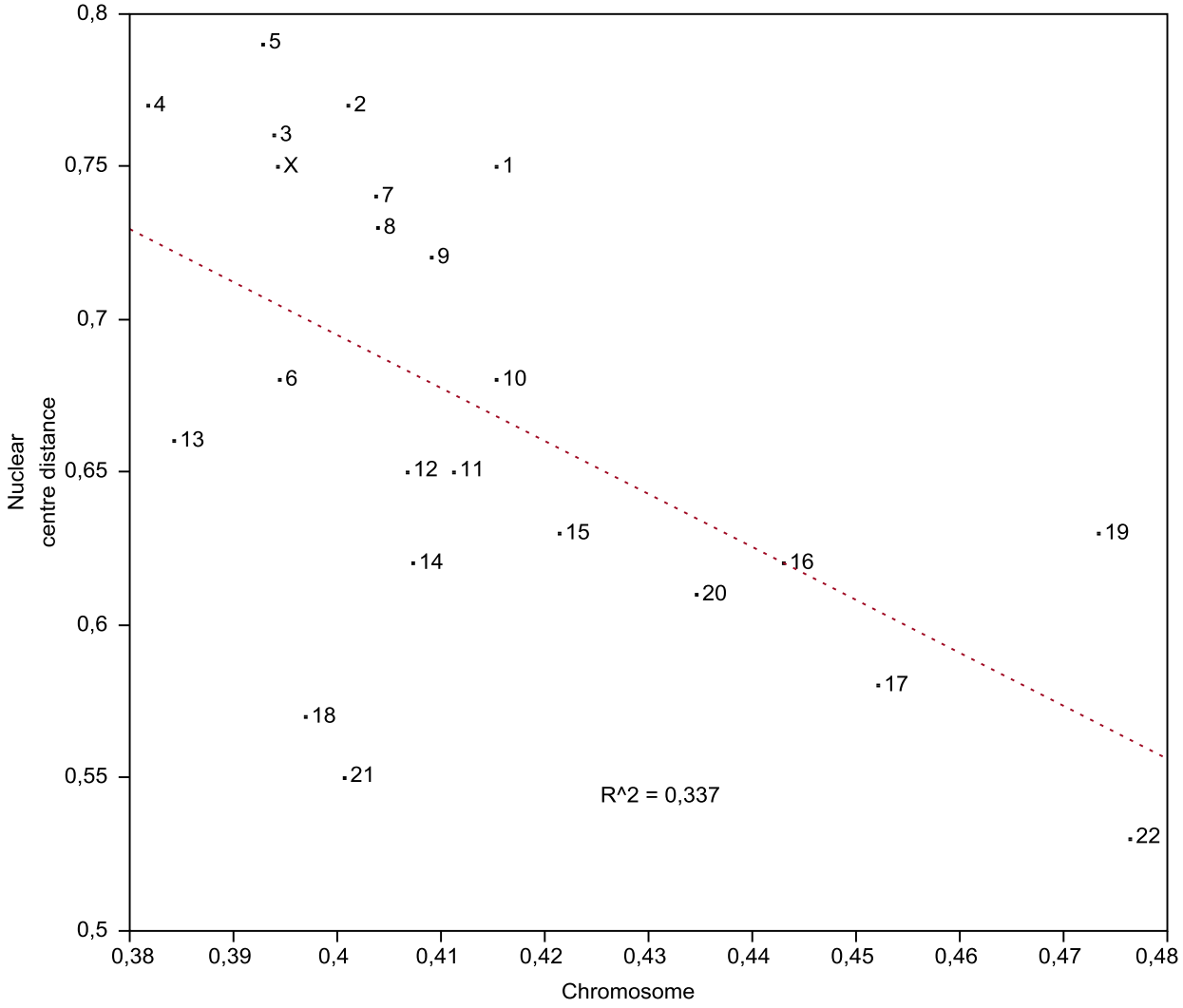


Figura 35: Correlación entre la distancia del cromosoma al centro del núcleo y la respectiva densidad génica (sin considerar el cromosoma Y).

ESTRUCTURA Y FUNCION

Se disecaron los cromosomas en varios segmentos estructurales y funcionales y se analizaron las correlaciones composicionales entre ellos. Los cromosomas se dividieron en grupos de: exones (las tres posiciones de los codones y todas juntas), intrones, regiones génicas e intergénicas, considerando y sin considerar las secuencias repetitivas (los coeficientes de Pearson para las correlaciones se muestran en la Tabla 5). Las correlaciones composicionales observadas entre cualquiera de los diferentes conjuntos de estructura y función de cada una de las regiones cromosómicas indica claramente que la variabilidad en el contenido en GC cromosómico no se produce por la diferencia de la contribución de un determinado tipo de estas

regiones, sino por la contribución, con diferencias en la fuerza, de todas ellas. En otras palabras, cualquier región de una determinada clase estructural o funcional dentro de un cromosoma rico en GC tiende a ser más rico que un segmento perteneciente a la misma clase pero en un cromosoma pobre en GC.

Tabla 5.

	GC_{Chr}	GC_{G25M}	GC_{IG25M}	$GC_{introns}$	GC_{exons}	GC_1	GC_2	GC_3
GC_{Chr}		0,90	0,98	0,88	0,88	0,81	0,50	0,88
GC_{G25M}			0,83	0,98	0,98	0,86	0,61	0,98
GC_{IG25M}				0,81	0,79	0,77	0,40	0,81
$GC_{introns}$					0,98	0,86	0,62	0,98
GC_{exons}						0,88	0,67	0,98
GC_1							0,38	0,88
GC_2								0,62

Correlaciones composicionales entre diferentes partes estructurales y/o funcionales del los cromosomas humanos. **GC_{Chr}** : GC medio cromosómico; **GC_{G25M}**: GC medio de los genes más sus correspondientes regiones flanqueantes 25 kb antes y después del ATG inicial y el codón de finalización, enmascarado para secuencias repetidas; **GC_{IG25M}**: Contenido en GC de las regiones intergénicas sustrayendo un segmento de 25 kb en cada extremo, enmascarado para secuencias repetidas ("negativo" de **GC_{G25M}**). Todos los coeficientes de correlación son altamente significativos ($p<0,001$) y expresados como R^2.

Los Ridges se han definido como regiones densas en genes, ricas en GC, enriquecidas en repetidos SINE, y que contienen genes con intrones cortos (Caron et al. 2001). Sobre esta base, investigamos si la asociación entre la riqueza en GC y el alto nivel de expresión era sólo local o, a su vez, podría extenderse al nivel de los cromosomas en su conjunto. Por lo tanto, hemos comparado el nivel global de expresión por Mb de cada cromosoma (obtenidos a partir de HTM, ver métodos) con el contenido en GC cromosómico; los resultados se muestran en la figura 36. Como puede verse, existe una fuerte correlación entre ambas variables. Aunque se trata de una medida aproximada de expresión global, y pueden existir muchas excepciones (como expresión tejido específica o patrones temporales), creemos que es una buena aproximación para el nivel medio de expresión espacio-temporal de cada cromosoma. Aunque a nivel de gen no puede obtenerse una clara asociación entre el contenido en GC y la expresión (Vinogradov 2001; Semon et al. 2005), al examinar los Ridges esta asociación surge sin problemas (Caron et al. 2001; Versteeg et al. 2003). Aquí mostramos que la correlación sigue existiendo en la escala cromosómica.

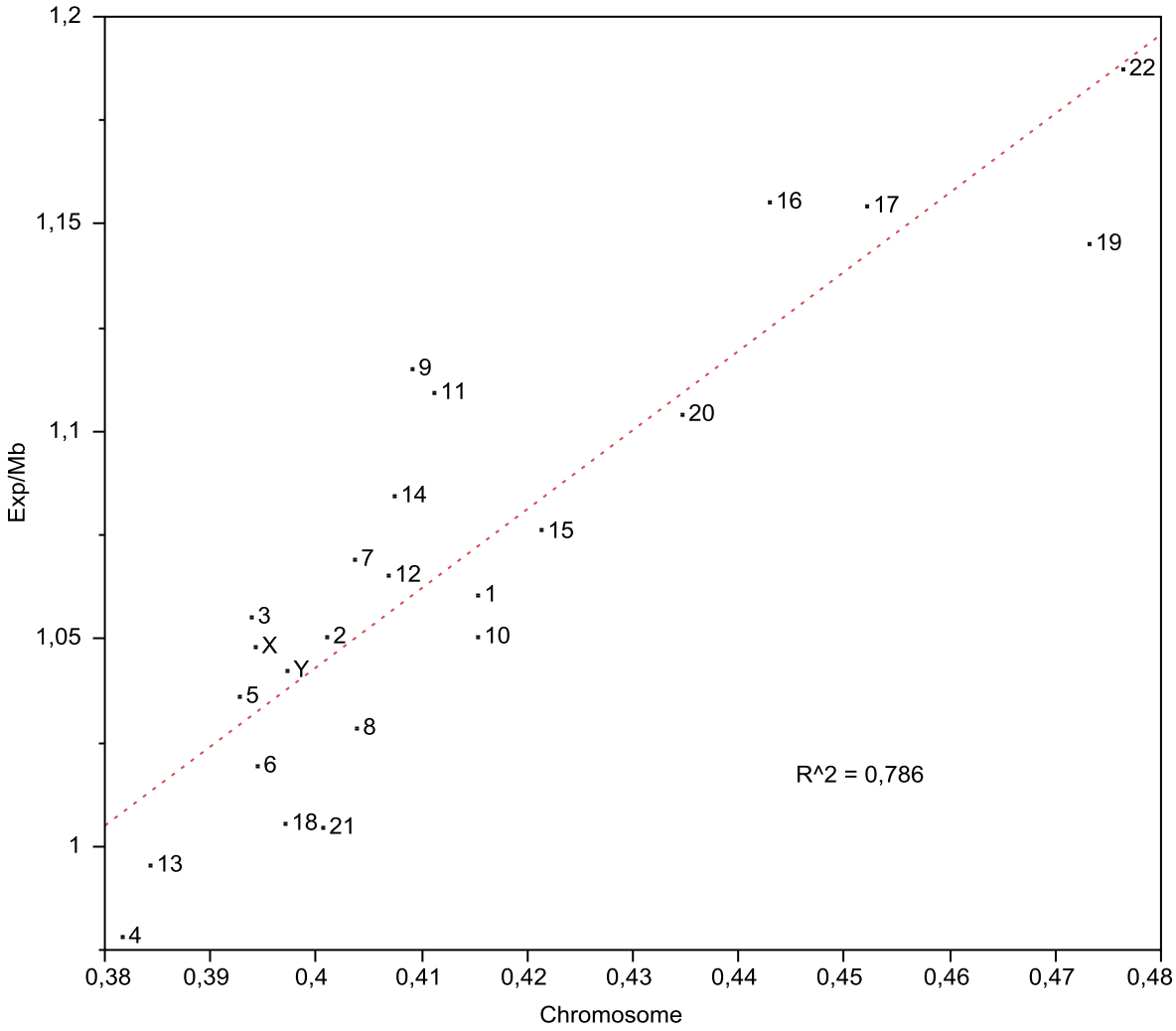


Figura 36. Correlación entre el GC de los cromosomas y su expresión media.

Woodfine et al. (2004) proporcionaron un mapa detallado de los tiempos de replicación de todo el genoma humano con una escala de resolución de 1 Mb, utilizando *tiling arrays* genómicos. Encontraron varias características que se correlacionan con el ciclo replicación: las regiones densas en genes, ricas en GC, ricas en ALU y otras SINES tienden a replicarse antes durante la fase S. Estas características también se correlacionan entre sí. La frecuencia de expresión también se encontró correlacionada con el ciclo de replicación, lo que podía esperarse teniendo en cuenta que los Ridges comparten muchas de estas propiedades. En este sentido, una fuerte correlación de $R^2 = 0,92$ entre el ciclo de replicación de los cromosomas y el contenido en GC se reporta en ese artículo. Una vez más, las propiedades composicionales están presentes cuando se considera el proceso de replicación al nivel de organización del cromosoma.

Numerosos artículos han vinculado las regiones de alto contenido en GC con mayores frecuencias de recombinación. En línea con nuestro

razonamiento anterior, hemos examinado la posibilidad de ampliar esta asociación a la escala cromosómica. La figura 37 muestra la tasa de recombinación por Mbp (ver Métodos) de cada cromosoma contra su contenido en GC, indicando que los cromosomas ricos en GC tienen una mayor tasa de recombinación general por Mbp que los cromosomas pobres en estas bases.

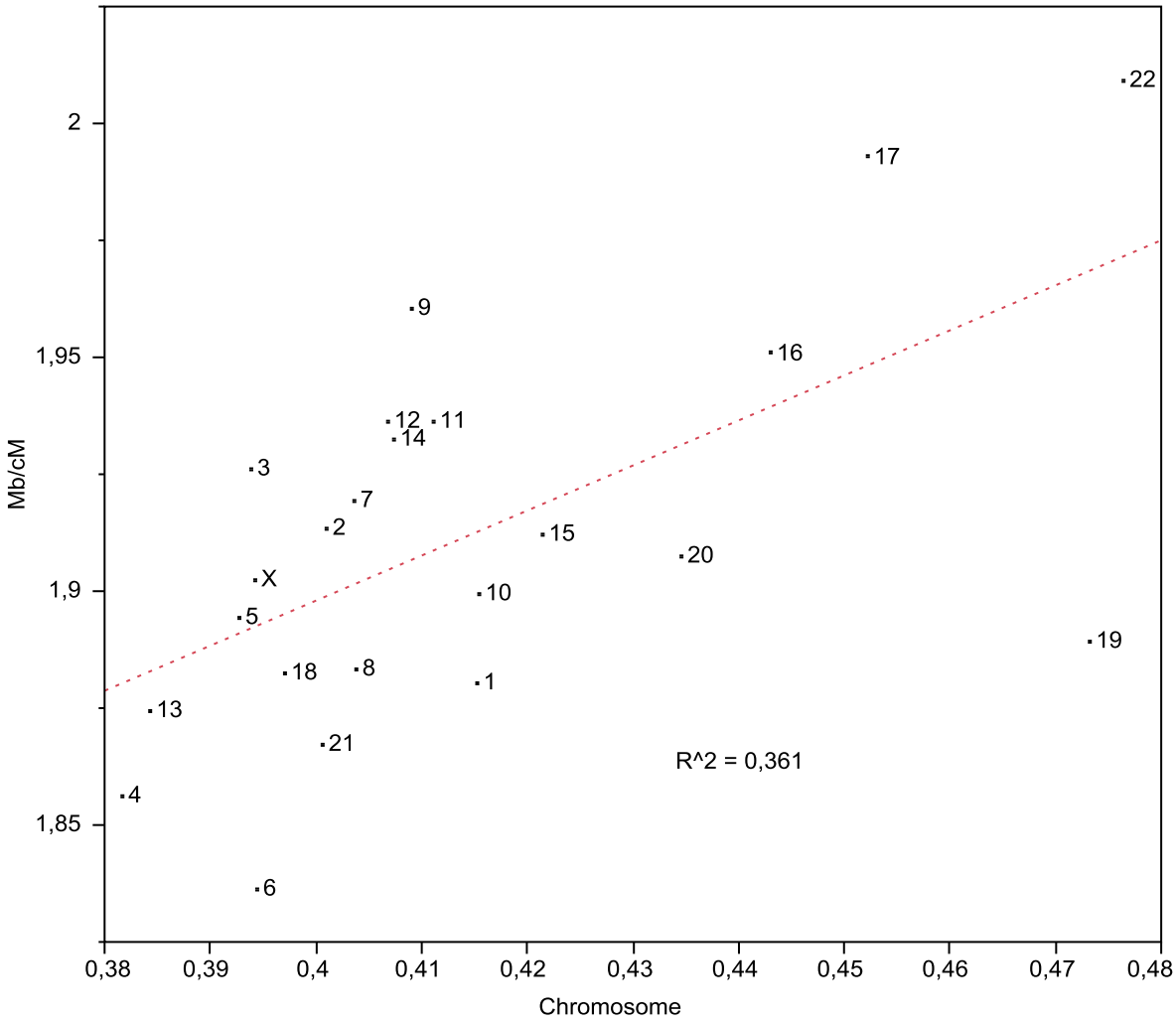


Figura 37. Correlación entre el GC de los cromosomas y la respectiva tasa de recombinación por Mbp.

Como ya hemos mencionado, la asociación del contenido en GC con tres procesos genéticos fundamentales, como son la expresión génica, la replicación y la recombinación, ha sido ampliamente reportado (Semon et al. 2005; Myers et al. 2006; Graffelman et al. 2007; Huvet et al. 2007). Sin embargo, estos análisis previos se han realizado principalmente a la escala local de los genes o regiones cromosómicas, y hasta el momento la escala cromosómica no fue examinada. Como hemos mostrado, todos estos procesos están relacionados con el contenido en GC a nivel de organización de cada cromosoma.

Como se espera, dada la relación evolutiva entre las especies las diferencias en las propiedades composicionales de los cromosomas no es distintivo del genoma humano. A pesar de esto, aun en los eucariotas no es un fenómeno universalmente extendido. En la figura 38 se puede observar la varianza en el contenido de GC de dos fuentes principales, en y entre cromosomas de varias especies que cubren los principales grupos eucariotas. Mientras que en eucariotas unicelulares como la levadura *S. cereviseae* o aun en plantas la varianza total explicada por los cromosomas (coeficiente de determinación ajustado por el numero de cromosomas) es alrededor o menor a 1%, en mamíferos, aves y algunos grupos de insectos este valor es claramente superior y puede llegar a más de 10%.

En la figura 38 se ilustran las distancias entre las distribuciones del GC génico en los cromosomas de algunas especies consideradas. Si bien es cierto que el abordaje es completamente diferente con respecto a la metodología de la tabla 6 puede observarse la misma tendencia general. Sin embargo, en este caso la heterogeneidad en insectos voladores parece ser muy grande. Este efecto es esencialmente debido al pequeño número de cromosomas en ambas especies consideradas (3 y 4 cromosomas). De hecho, en las dos especies solo un cromosoma difiere sustancialmente de los demás (5% de GC en las medianas de *Anopheles gambiae* y 10% de GC en *Drosophila melanogaster*, lo que lleva a que una mayor fracción de distancias se vuelva muy grande.

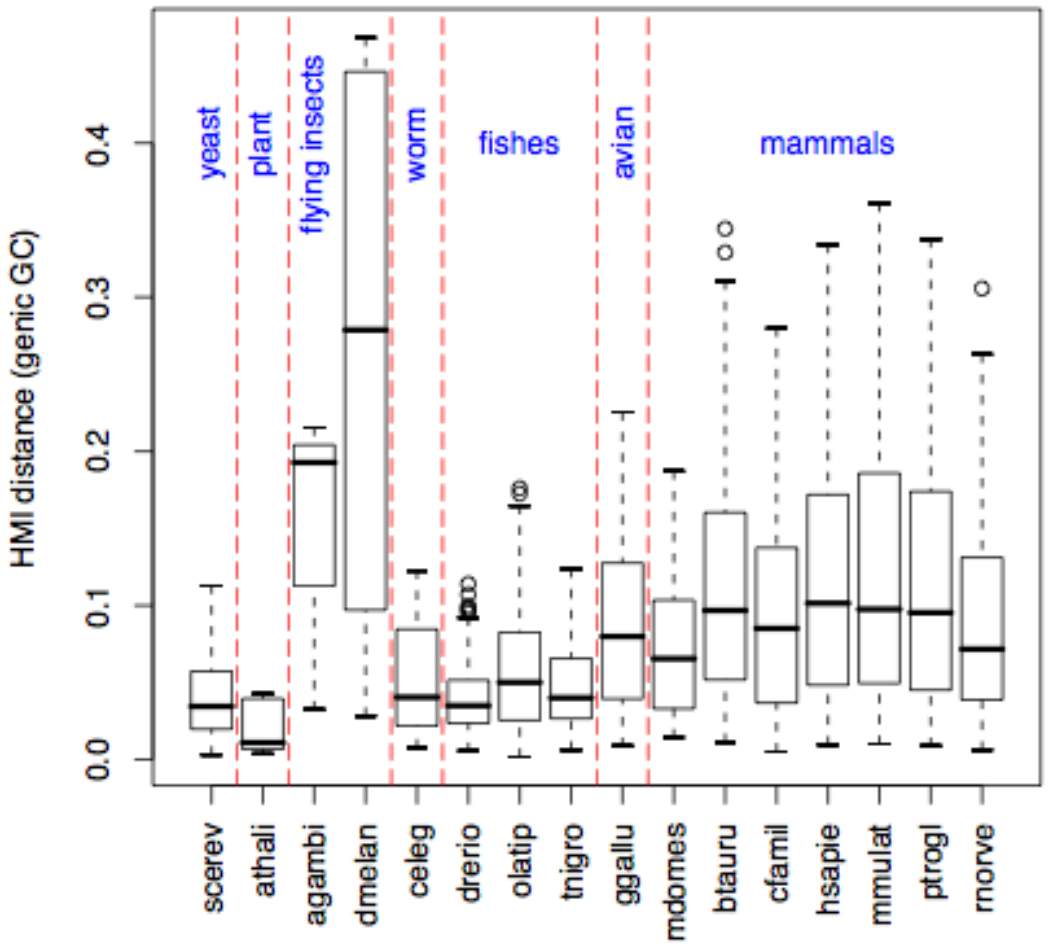


Figura 38. Distancias entre distribuciones cromosómicas de GC génico en diferentes especies. Abreviaturas: scerev, *Saccharomyces cerevisiae;* athali, *Arabidopsis thaliana;* agambi, *Anopheles gambiae;* dmelan; *Drosophila melanogaster;* celegan, *Caenorhabditis elegans;* drerio *Danio rerio;* olatip, *Orizia latipes*; tnigro, *Tetraodon nigroviridis;* ggallu, *Gallus gallus;* mdomes, *Monodelphis domestica; btauru, Bos taurus*, cfamil, *Canis familiaris*, hsapie, *Homo sapiens*, mmulat, *Macaca mulata*; ptrogl, *Pan troglodites*; rnorv, *Rattus norvegicus*.

Tabla 6.

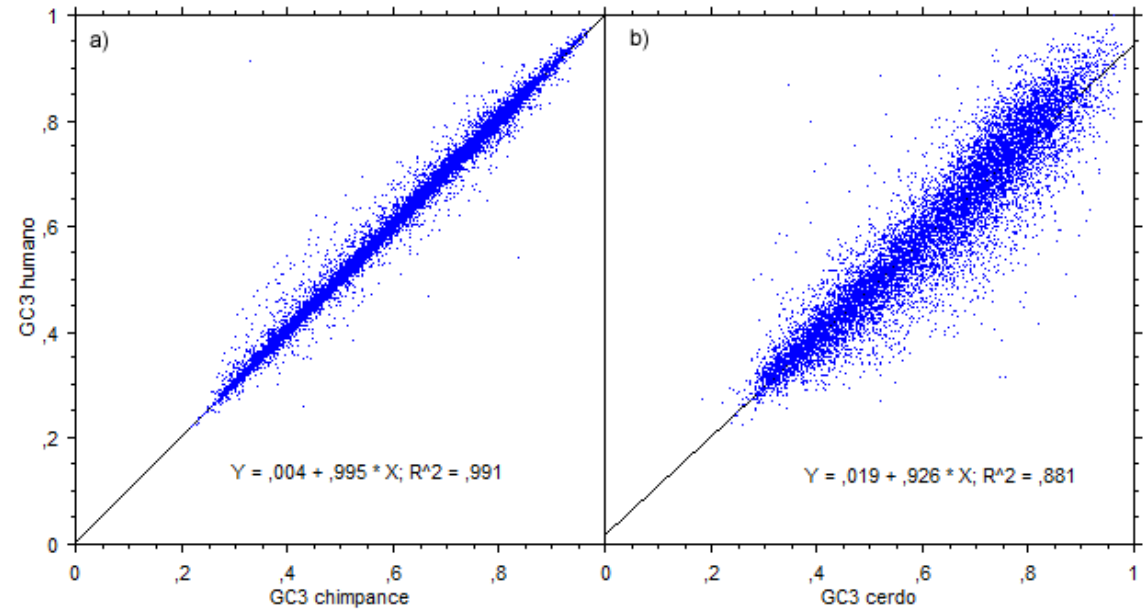
Especie	crom	GC gen	R2% (GCgen)	GC3	R2% (GC3)
Scerev	16	0,402	1.0	0,394	1,6
Athali	6	0,445	0,0	0,427	0,1
Agambi	3	0,551	3.9	0,682	3,0
Dmelan	4	0,538	8.7	0,649	10,2
Celeg	6	0.434	2.6	0.414	1,2
Drerio	25	0,501	1,1	0,559	1,2
Olatip	23	0,530	2,4	0,629	2,8
Tnigro	18	0,555	0,8	0,654	0,6
Ggallu	17	0,488	4,4	0,525	5,2
Mdomes	9	0,486	2,1	0,518	1,6
Btauru	28	0,540	10,7	0,630	10,9
Cfamil	28	0,528	7,4	0,603	7,9
Hsapie	23	0,531	11,0	0,602	11,3
Mmulat	21	0,519	11,0	0,579	11,3
Ptrogl	24	0,521	10,2	0,580	10,8
Rnorve	21	0,518	5,5	0,588	6,1

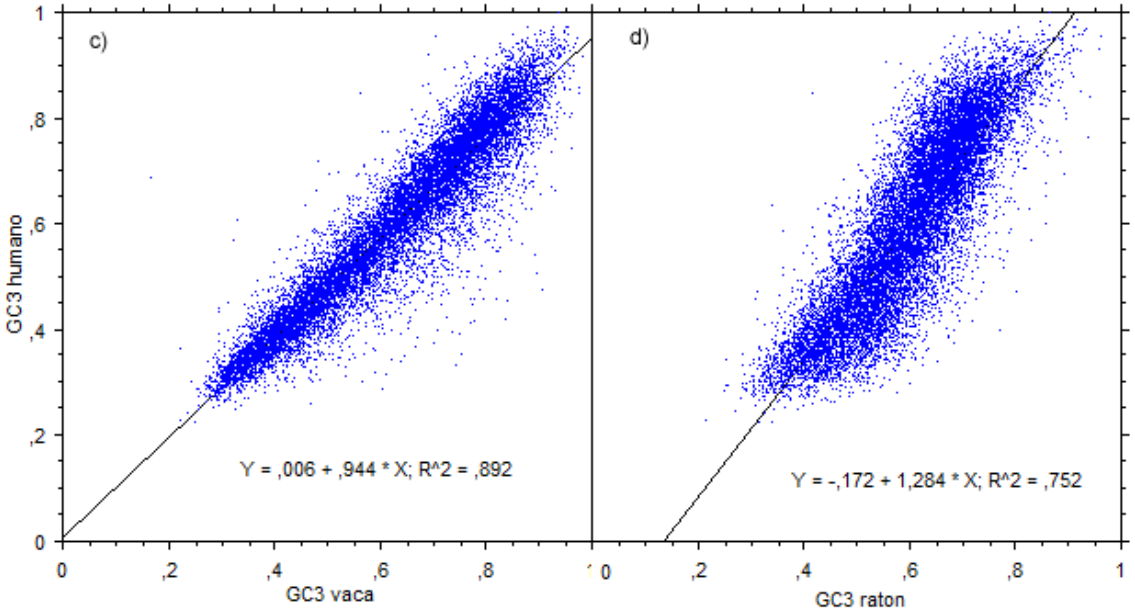
Coeficiente de determinación (%) por el efecto cromosómico en el contenido en GC y GC_3 génicos. Abreviaturas: scerev, *Saccharomices cerevisiae;* athali, *Arabidopsis thaliana;* agambi, *Anopheles gambiae;* dmelan; *Drosophila melanogaster;* celegan, *Caenorhabditis elegans;* drerio *Danio rerio;* olatip, *Orizia latipes*; tnigro, *Tetraodon nigroviridis;* ggallu, *Gallus gallus;* mdomes, *Monodelphis domestica; btauru, Bos taurus*, cfamil, *Canis familiaris*, hsapie, *Homo sapiens*, mmulat, *Macaca mulata*; ptrogl, *Pan troglodites*; rnorv, *Rattus norvegicus*.

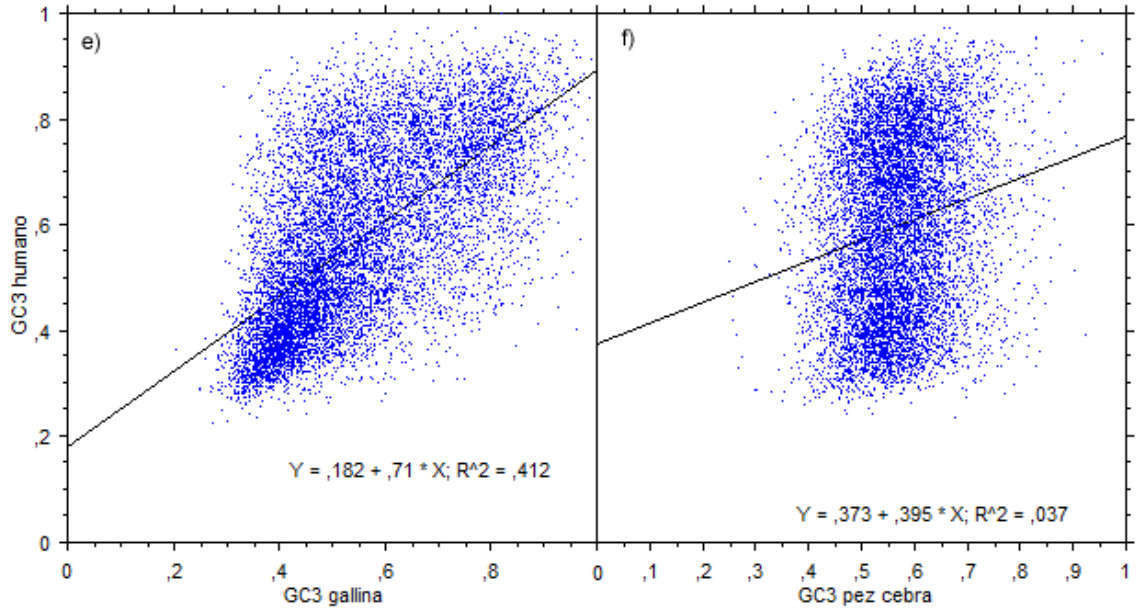
LA PERSISTENCIA DE LOS ISOCOROS ENTRE ESPECIES

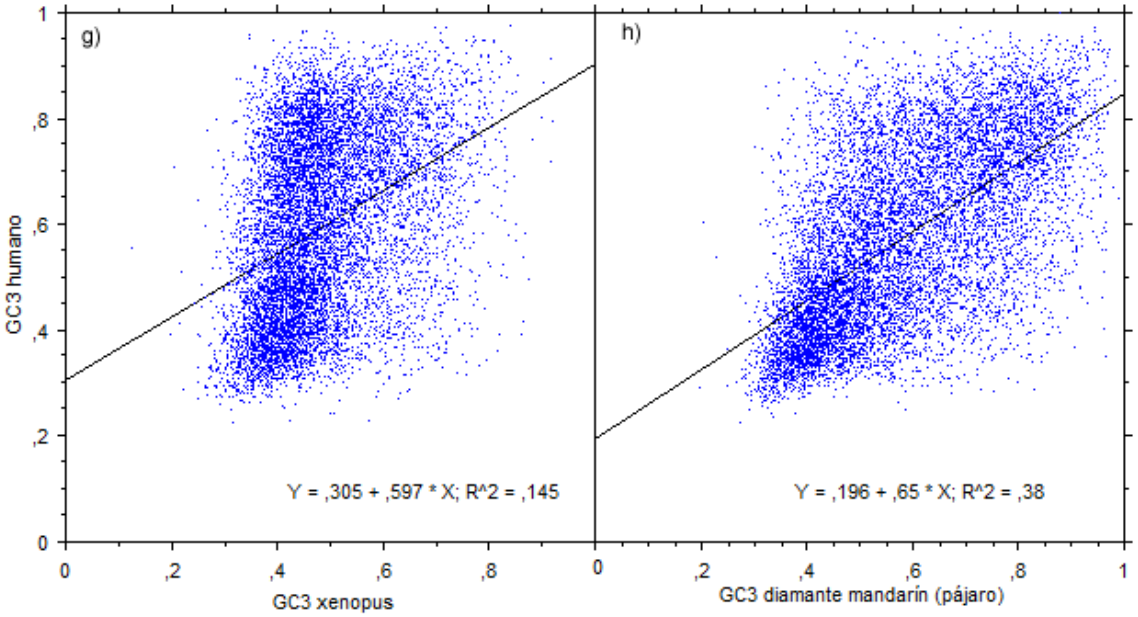
Los análisis realizados sobre genes ortólogos hace 20 años permitieron realizar generalizaciones sobre la composición de isocoros y de familias de isocoros, tanto en los mamíferos como en las aves. De esta manera, a través del análisis de las correlaciones entre el GC_3 de genes ortólogos de

hombre y otras especies como la vaca, ratón y gallina (figura 7), se desarrolló la teoría de que, a pesar de que la transición hacia isocoros ricos en GC se había dado dos veces en la evolución en forma independiente, habían sido los mismos genes (y los mismos isocoros) los que se habrían enriquecido; postulándose asimismo que el incremento en GC se habría dado en las regiones de los poiquilotermos que presentaban niveles relativamente más elevados de GC. Si bien las correlaciones encontradas apoyaban esta teoría, en el momento que se realizaron se disponía de un nivel de información muy inferior al actual. Es así que para esta tesis se realizaron nuevamente las correlaciones entre genes ortólogos pero esta vez en lugar de utilizar < 1000 genes por par de especies comparadas, se utilizaron en promedio 10.000 (figura 39). La correlación es muy alta *entre* especies de mamíferos (figura 39 a-d) o aves (figura 39 i), sin embargo los valores caen mucho cuando se correlacionan los valores de mamíferos *vs.* aves (figura 39 e y h), ya que existen genes con GC_3 extremadamente bajos en humano que en gallina u otras aves son muy altos y viceversa. En conjunto la tendencia es a formar en lugar de una recta, un gran cuadrado para valores de GC_3 en humano > 60%, donde los puntos se reparten sin un órden determinado. La correlacion entre genes ortólogos humanos y de *Xenopus* así como entre humanos y pez cebra también es muy baja en contra de la teoría de que los genes que se enriquecieron en GC eran aquellos que ya se encontraban más enriquecidos (figura 39 f y g). Resulta interesante ver que prácticamente no existe correlación entre 2 especies de peces (figura 39 j) o entre peces y *Xenopus*. Entre mamíferos las diferencias más notorias se encuentran con el ratón ya que presenta el "minor shift" composicional que hace que la correlación se mueva en forma de sigmoide en valores extremos de GC_3 (figura 39 d). La comparación de genes ortólogos de mamíferos y aves con *Xenopus* (figura 39 g y 39 k) y entre sí (figura 39 e), nos permite tener una idea primaria de cómo se ha comportado la evolución sobre las diferentes regiones comoposicionales. Desde el punto de vista de conservación composicional, encontramos que es mucho más intensa en niveles bajos de GC_3. Sobre las regiones del "paleogenoma" de bajo contenido en GC opera selección negativa. No sucede lo mismo en el "neogenoma" ya que el enriquecimiento en GC ha sido diferente en aves y mamíferos y a su vez también en diferentes genes que los ya enriquecidos en GC de *Xenopus*. Si comparamos los genes de de esta especie con los genes de peces encontramos que ni siquiera puede hablarse de conservación del "paleogenoma" (figura 39 l).


a)
b)
GC3 humano
GC3 chimpance
GC3 cerdo
Y = ,004 + ,995 * X; R^2 = ,991
Y = ,019 + ,926 * X; R^2 = ,881
0
,2
,4
,6
,8
1


c)
d)
GC3 humano
GC3 vaca
GC3 raton
Y = ,006 + ,944 * X; R^2 = ,892
Y = -,172 + 1,284 * X; R^2 = ,752
0
,2
,4
,6
,8
1


e)
f)
GC3 humano
GC3 gallina
GC3 pez cebra
Y = ,182 + ,71 * X; R^2 = ,412
Y = ,373 + ,395 * X; R^2 = ,037
0
,2
,4
,6
,8
1
,1
,3
,5
,7
,9


g)
h)
GC3 humano
GC3 xenopus
GC3 diamante mandarín (pájaro)
Y = ,305 + ,597 * X; R^2 = ,145
Y = ,196 + ,65 * X; R^2 = ,38
0
,2
,4
,6
,8
1

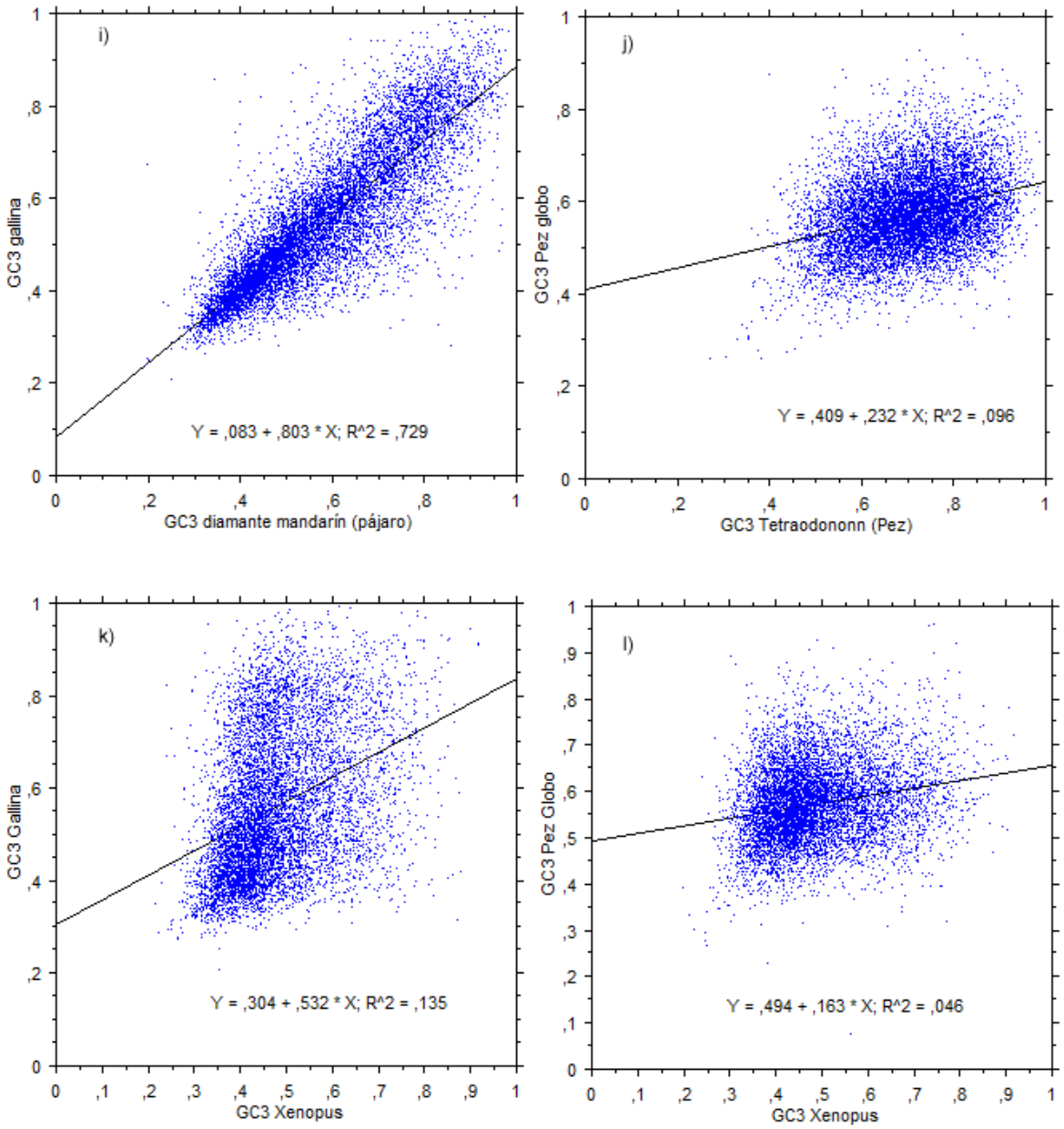


Figura 39. Correlación de GC3 de genes ortólogos entre diferentes especies de vertebrados. Las especies comparadas están mostradas en los ejes de las figuras.

Se realizó un análisis de correspondencia (CoA) sobre la matriz de frecuencias de aminoácidos en 14.815 proteínas humanas, que representan un porcentaje muy alto del total, que se piensa está entre 20.000 y 25.000 (Consortium, 2004) para entender cuáles son las tendencias principales que modelan el uso global de aminoácidos en nuestro proteoma. Con este objetivo, se aprovecharon tres características, que son a) la disponibilidad del genoma humano completo, b) una colección "limpia" de genes no-redundantes, que representan más del 50% de la población supuesta de genes en los seres humanos, y c) la localización exacta de cada gen y de la región circundante correspondiente, que ayuda a entender la influencia de la estructura de isocoros de nuestro genoma en el uso de aminoácidos. Los primeros tres ejes representaron 20,4%, 14,7% y 9,9% de la variabilidad total, respectivamente. La distribución de los residuos para los primeros dos ejes se muestra en figura 40.

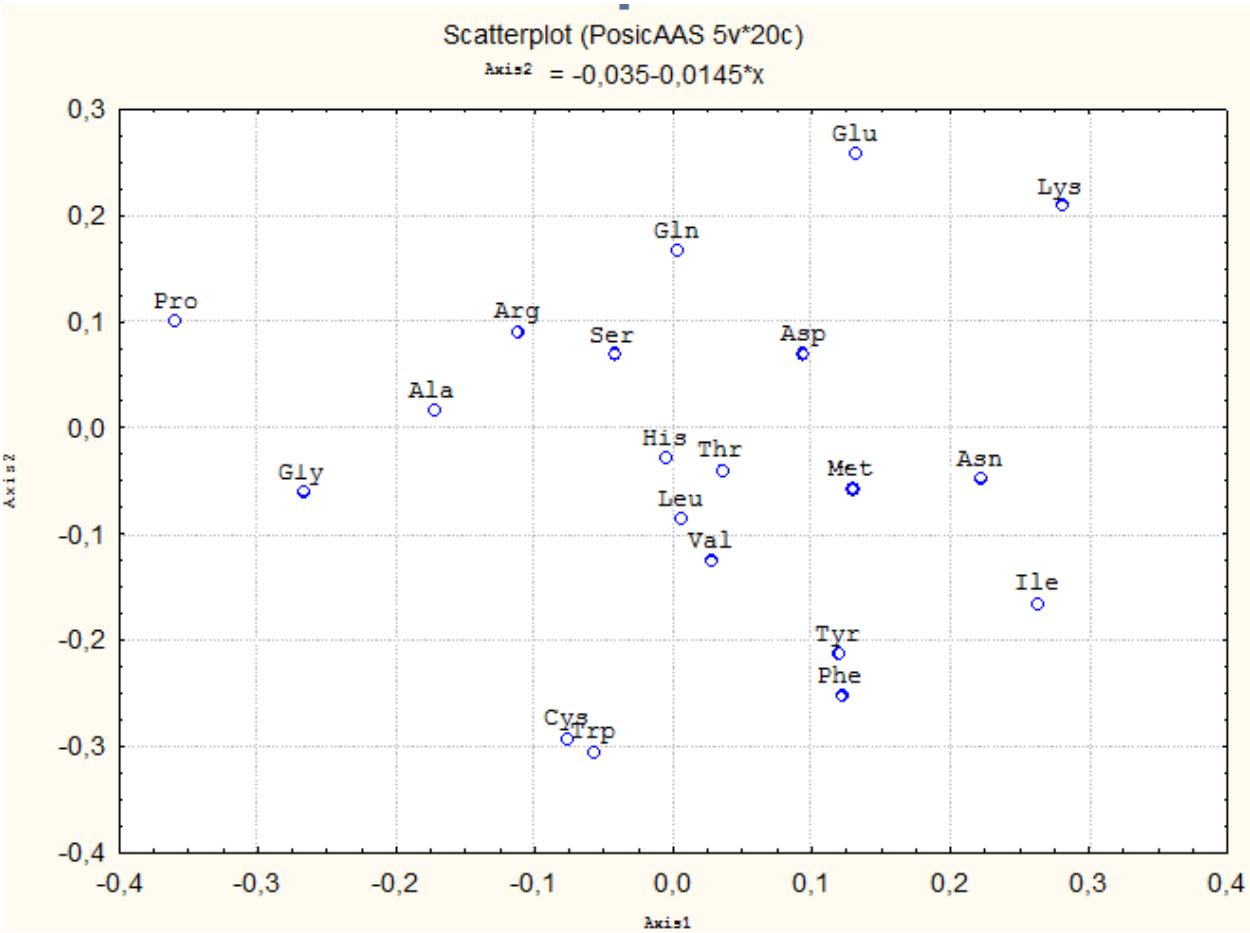


Figura 40. Representación de la posición de los aminoácidos en el plano definido por los dos primeros ejes del CoA.

Como puede verse en esta figura, el primer componente (eje horizontal) discrimina los aminoácidos según el contenido de GC en la primera y segunda posiciones del codón: en el lado izquierdo del plano (valores negativos) se encuentran residuos como Pro (CCN), Gly (GGN), Ala (GCN) y Arg (CGN + AGR), mientras que Lys (AAR), Ile (ATA + ATY) y Asn (AAY) ocupan el extremo opuesto. Sin embargo, una inspección más cercana de los aminoácidos clasificados según la posición respecto a este eje, sugiere que el contenido en GC en la segunda posición del codón tendría una importancia mayor que en el primero. De hecho, los primeros siete residuos en el lado izquierdo de la distribución muestran G o C en esta posición mientras que una A o un T se encuentran en la misma posición de los ocho aminoácidos con los valores positivos mayores. Estos resultados sugieren fuertemente que la principal fuerza para el uso de aminoácidos entre las proteínas humanas son las restricciones composicionales que actúan en cada secuencia, como ha sido postulado previamente por D'Onofrio et al. (D'Onofrio et al. 1991) estudiando solamente el 10% de las secuencias analizadas aquí. Para confirmar esta hipótesis, se correlacionaron las posiciones de los genes sobre el eje 1 contra GC_1 (figura 41a) y contra GC_2 (figura 41b).

Se encontró que, aunque las dos correlaciones son altamente significativas, los valores de R^2 fueron más altos para GC_2 que para GC_1. Es interesante observar que la correlación con GC_3, aunque significativa, es mucho más baja puesto que el valor de R^2 es 0,22, comparado con 0,48 para GC_1 y 0,85 para GC_2 (véase la figura 41). El hecho de que GC_3 represente una posición más relajada del codón respecto a las restricciones composicionales de las proteínas hace que explique mucho menos variabilidad sobre el eje 1.

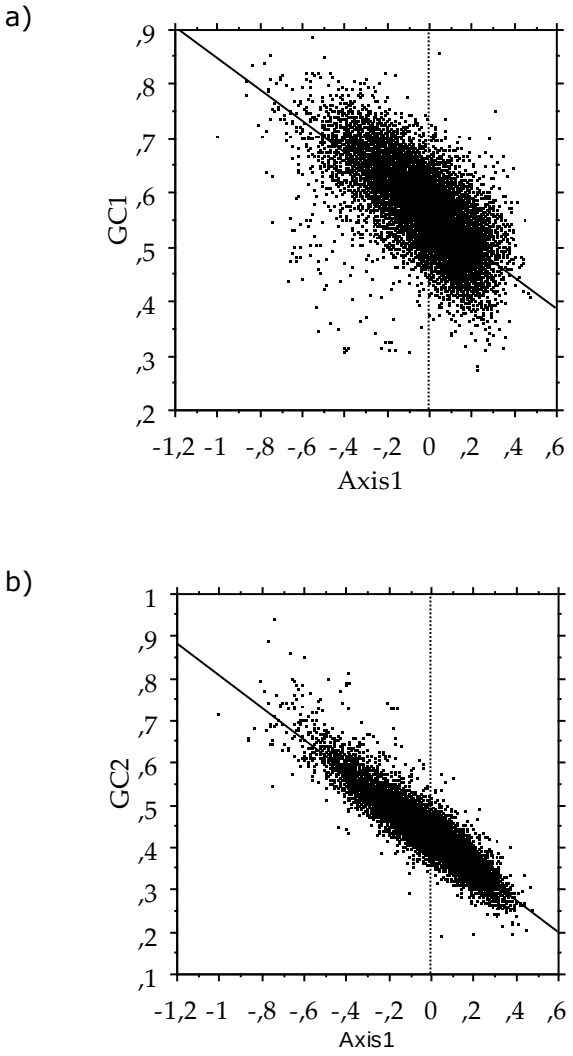


Figura 41. Correlaciones entre la posición de cada gen a lo largo del primer eje del análisis de correspondencia *vs*. GC_1 (a) y GC_2 (b) de cada secuencia. Los valores de R^2 son -0,48 y -0,85, respectivamente, y en los dos casos, $p < 0,0001$.

Para confirmar los resultados publicados por el grupo de Bernardi (D'Onofrio et al. 1991), es decir la importancia de la región (isocoro) donde cada secuencia se encuentra para el uso de los aminoácidos, correlacionamos la posición de cada gen en el primer eje contra el contenido en GC de los intrones correspondientes y de los isocoros que los contienen. Las dos correlaciones fueron altamente significativas ($P < 0,0001$), con valores de R^2 de -0,26 y -0,17, respectivamente. Además, puesto que el GC de todas las posiciones de los codones se correlacionan con las regiones circundantes y los intrones (véase la tabla 7), podemos concluir que la principal fuerza en la arquitectura de las proteínas humanas es la localización física de cada gen, reforzando la interpretación anterior de D'Onofrio et al. (D'Onofrio et al. 1991). Este hallazgo es muy importante si

tenemos en cuenta los avances en genómica humana de los 20 años pasados. De hecho, la cantidad enorme de datos generados a través de técnicas del alto rendimiento de procesamiento demostró patrones complejos en la expresión de genes en diversos tejidos humanos y etapas de desarrollo. La idea de que la localización del gen en el cromosoma y el contenido en GC del isocoro sean la principal fuerza en la composición de aminoácidos, a pesar de la posible existencia de otros factoes como la función del gen, los requisitos espacio-temporales exactos de la expresión, la cantidad e las interacciones de cada proteína, etc. determina nuevamente diversos aspectos de la discusión seleccionista-neutralista.

Tabla 7**.**

Posición codón	Isocoros	Intrones
GC_1	0,31	0,38
GC_2	0,10	0,18
GC_3	0,45	0,66

Correlaciones composicionales entre las tres posiciones de los codones y el contenido en GC de los isocoros respectivos y los intrones. En cada caso, el valor de p es < 0,0001 y se expresa como R^2.

En la figura 42 se muestra la correlación existente entre la posición de cada secuencia respecto al eje 2 (que explica el 14,7% de la variabilidad total) contra el nivel de hidropatía de cada proteína (calculado de acuerdo a la escala de Kyte-Doolittle). Como puede observarse, ambas variables están fuertemente relacionadas. Los valores negativos en este eje corresponden a las proteínas hidrofóbicas. Este resultado demuestra que, al igual que en varios procariotas y especies eucariotas unicelulares, sucede que este factor es crucial para la arquitectura de las proteínas humanas. Además, la posición de las secuencias respecto a este eje se correlaciona también con la frecuencia de hojas β en cada proteína (R^2 = -0,22, P < 0,0001). Este resultado resulta esperable, ya que los residuos hidrofóbicos tienden a ser más frecuentes en estas estructuras secundarias.

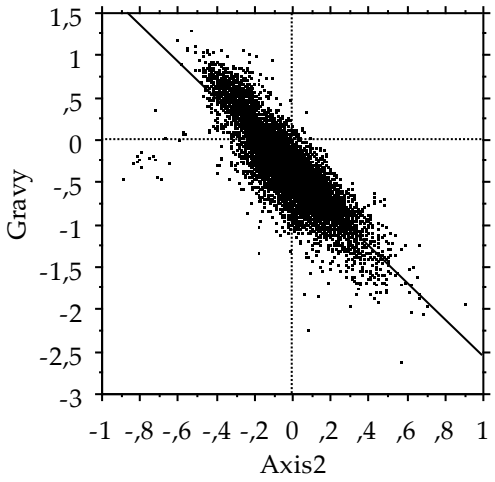


Figura 42. Correlación entre la posición de cada gen a lo largo del segundo eje del análisis de correspondencia y el nivel de hidropatía de cada proteína (Gravy score). El valor de R^2 es -0,71 y $p < 0,0001$.

A nivel individual de aminoácidos, en la figura 40 puede observarse que los más hidrofóbicos están abajo en el plano (valores negativos para el eje 2) mientras que la localización opuesta es ocupada por los residuos hidrofílicos. En lo que respecta a las estructuras secundarias, se analizó la distribución de α-hélices, hojas β y coiled-coil (α-hélices dobles) para el conjunto total de proteínas (figura 43). Encontramos que las frecuencias de residuos en regiones de α-hélices y "coiled-coil" en cada proteína, se correlacionan ($R^2 = -0,17$, $p < 0,001$; $R^2 = 0,23$, $p < 0,001$) con la posición de las secuencias a lo largo del tercer eje, que explica el 9,9% de la variabilidad total.

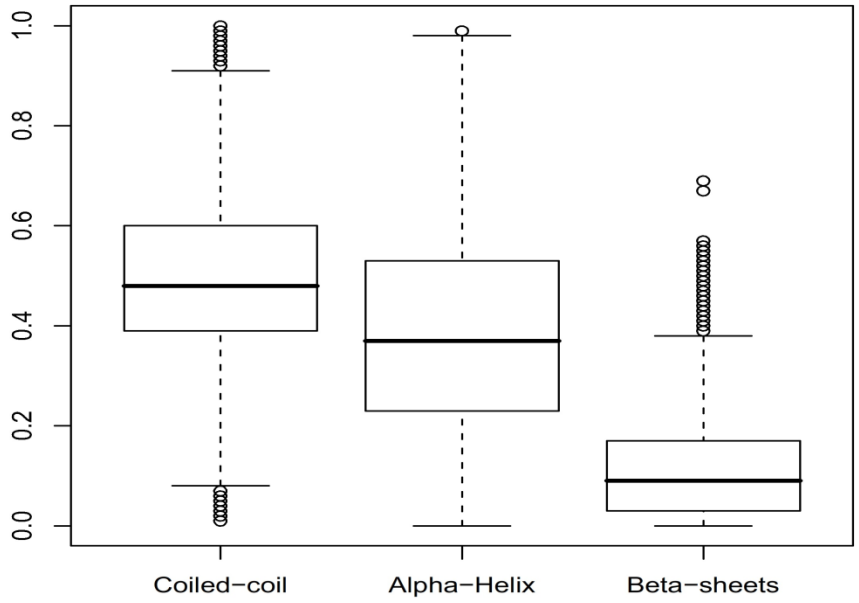


Figura 43. Distribución de las fracciones de residuos en cada proteína humana clasificadas por tipo de estructura secundaria. La línea central corresponde a la media, las "cajas" corresponden al primer y tercer cuartiles mientras que las patillas están a 1,5 intervalo intercuartílico.

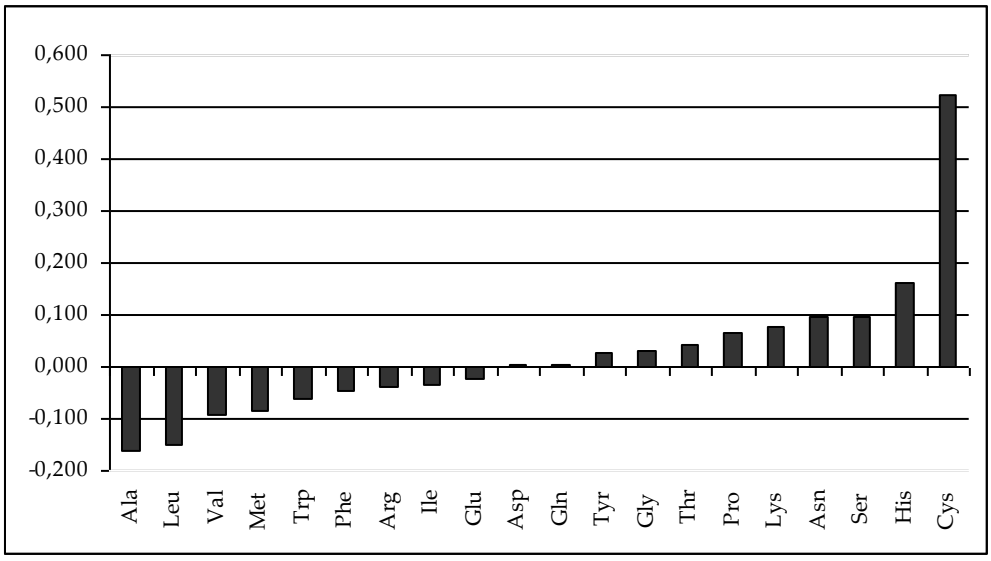


Figura 44. Posición relativa de cada aminoácido respecto al eje 3 del análisis de correspondencia.

Cuando la posición de cada aminoácido se grafica contra la posición respectiva respecto al eje 3 (figura 44) puede observarse que la mayor parte de la variabilidad es debida al valor extremo de la Cys. Este residuo se utiliza generalmente en niveles bajos, y cuando algunas proteínas dentro de un genoma incrementan su concentración relativa, puede ser detectado

generalmente como una de las fuentes de variabilidad (ejes) más prominentes del análisis multivariado [véase por ejemplo (Garat y Musto 2000; Zavala et al. 2002)]. Las proteínas humanas no son una excepción, ya que la Cys representa solamente el 2,27% de los residuos totales, ranqueando por encima solamente de Pro (2,18%) y Trp (1,25%). Sin embargo, mientras que estos últimos aminoácidos casi siempre se utilizan en niveles muy bajos, hay un grupo de proteínas que muestran niveles extremadamente altos de Cys (incluyendo un grupo de 35 donde representa más del 20% de todos los aminoácidos). Entre este grupo, hay proteínas asociadas a queratina, metalotioneínas, etc. La correlación entre el uso de Cys y la posición de cada proteína a lo largo del eje 3 se puede observar en la figura 45.

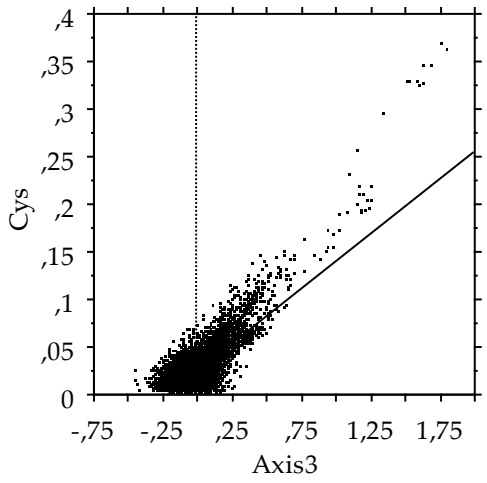


Figura 45. Correlación entre la posición de cada gen a lo largo del tercer eje del análisis de correspondencia y la frecuencia de Cys en cada proteína. El valor de R^2 es 0,61 y $p < 0,0001$.

En resumen, hemos mostrado que la principal fuerza que modela el uso de aminoácidos en las proteínas humanas está dada por el contenido en GC de los isocoros en que se encuentra ubicado cada gen. En segundo lugar está la influencia del nivel de hidropatía de cada proteína y la frecuencia de hojas β. Finalmente, el tercer factor se relaciona con el uso de Cys y la frecuencia de α-hélices. En conjunto, estos tres primeros factores explican el 45% de la variabilidad total encontrada en las proteínas humanas. Dado que cada uno de los ejes consecutivos explican una proporción más baja de la variabilidad total (por ejemplo, el cuarto corresponde al 8,1%) la imagen que hemos mostrado, aunque incompleta, es una buena representación de los factores globales que modelan la arquitectura de nuestras proteínas y muy probablemente, la de otros mamíferos.

Como se sabe desde hace mucho tiempo, hay una gran variación en el contenido en GC_3 entre los genes humanos (Sharp et al. 1988; Aissani et al. 1991; Sharp et al. 1995; Bernardi 2000). En el caso de las secuencias analizada en este trabajo, esa variabilidad se muestra en la figura 46. Además, como era de esperar, se puede observar una fuerte correlación entre el contenido en GC_3 de cada gen y el isocoro donde cada secuencia se encuentra (Bernardi et al. 1985) (ver figura 47). Para esta muestra de genes, el $R^2 = 0{,}51$, $p < 0{,}0001$. Por lo tanto, en el genoma humano, como era esperable, la altísima variabilidad en GC_3 es el factor principal modulando el uso de codones.

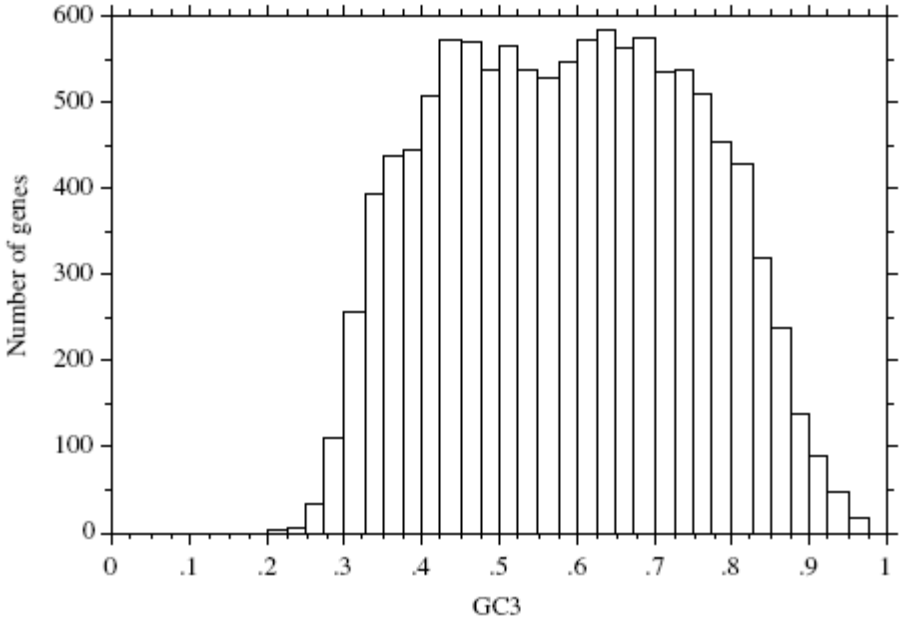


Figura 46. Número de genes para diferentes niveles de GC_3.

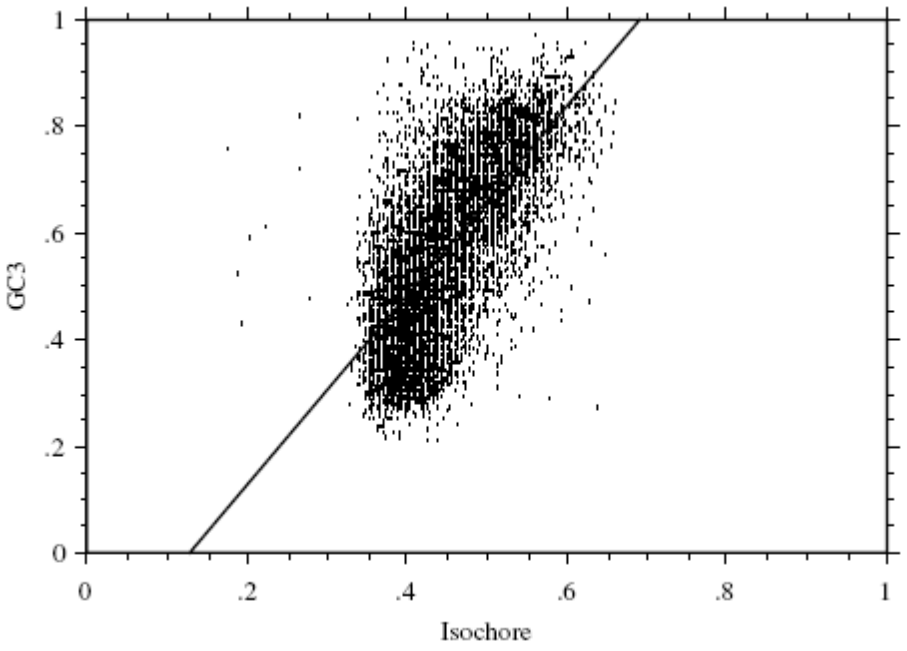


Figura 47. Correlación entre GC_3 y el GC del isocoro donde se encuentran los genes de la muestra.

Esto se ve claramente cuando el conjunto de datos se estudia mediante análisis de correspondencia (CoA). Por lo general, cuando los dos primeros ejes se trazan en un plano, es posible directamente a través de la inspección visual detectar las tendencias biológicas, que son a menudo difícil (o imposibles) de entender en los espacios de mayores dimensiones. En este trabajo, el análisis se realizó sobre el RSCU de datos (excluyendo Met, Trp y codones de terminacion) para reducir al mínimo los efectos de la composición en aminoácidos. La figura 48a muestra la posición de los genes en el plano definido por el primer eje (horizontal) y el segundo (vertical), que representan el 38% y el 4,3% de la variación total, respectivamente.

Como dijimos más arriba, la única fuente cuantitativamente importante de variación está fuertemente correlacionada con el nivel de GC_3 de cada gen (figura 48B; $R^2 = 0,97$, $p < 0,0001$). Tomados individualmente, C_3 y G_3 no contribuyen por igual a la correlación. De hecho, para C_3, $R^2 = 0,86$, $p < 0,0001$, mientras que para G_3, $R^2 = 0,70$, $p < 0,0001$). Esta diferencia podría deberse a dos factores: en primer lugar, hay más tripletes terminados en C que en G con al menos un codón sinónimo (16 *vs*. 13) y porque el aumento de tripletes terminados en G es menor en valores altos de GC_3 que los terminados en C. Como se ha mencionado anteriormente estos resultados confirman, con un conjunto de datos mayor, resultados anteriores, entre otros, para el genoma humano (Sharp et al. 1988) y vertebrados de sangre fría (Musto et al. 2001; Romero et al. 2003).

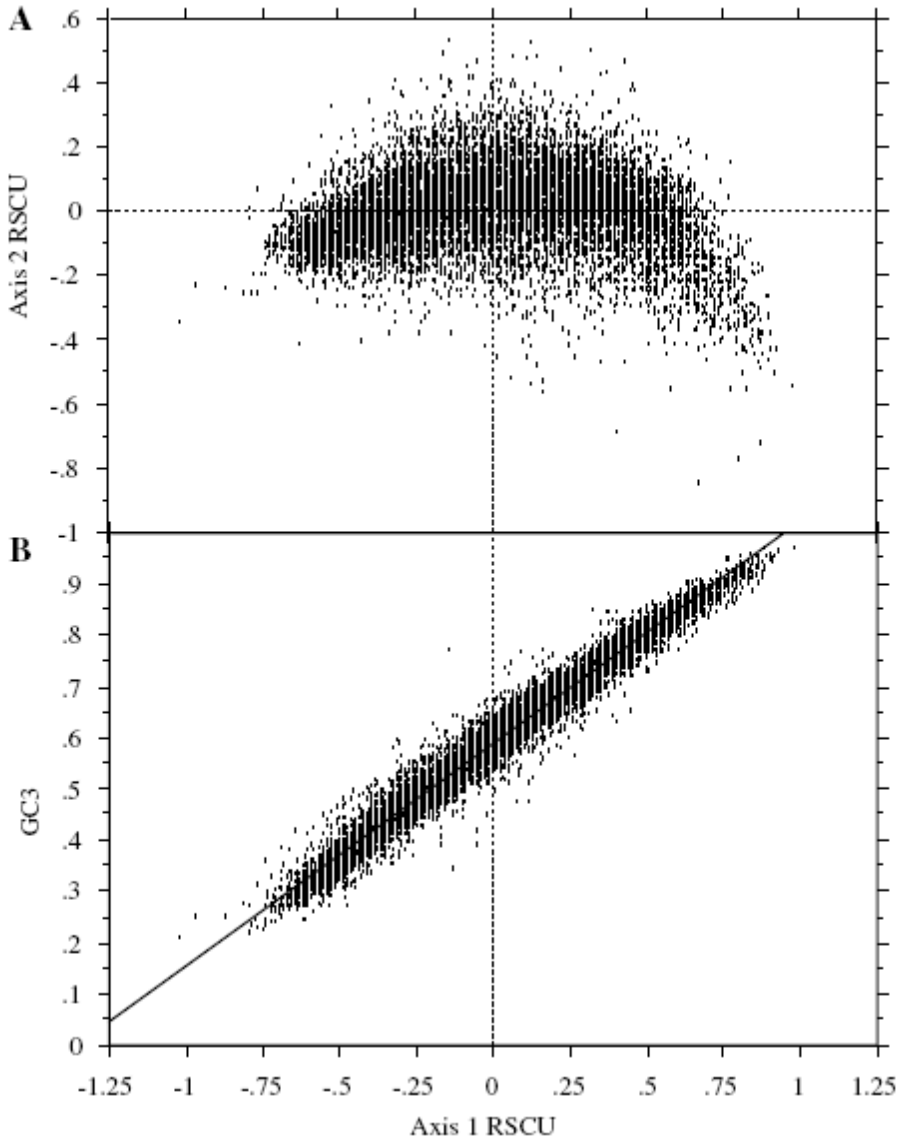


Figura 48. A: Distribución de los genes en el plano definido por los dos ejes principales del CoA. B: Correlación entre la posición de cada gen en el primer eje y el contenido en GC_3 de las diferentes secuencias. Para la última figura, $R^2 = 0,97$, $p < 0,0001$.

El resultado más interesante surge del análisis de los genes de acuerdo con el segundo eje generado por el CoA. En la figura 49 se muestra la posición de los codones en el plano definido por los dos ejes principales. Como era de esperar, ya que el primer eje está fuertemente correlacionado con GC_3, los tripletes terminados en G o C se ubican en un extremo del eje 1, mientras que los que terminan en A o T se colocan en el lado opuesto de este eje.

En relación con el eje 2, se puede ver que en un extremo se agrupan varios codones que contiene el dinucleótido CpG, siendo aquellos que muestran los valores extremos TCG (Ser), GCG (Ala), CCG (Pro), CGC (Arg) y ACG (Thr), mientras que los otros tres tripletes que contienen el

mencionado dinucleótido (CGA, CGG y CGT, codificando para Arg) tienden a colocarse en el mismo extremo. Es interesante observar que el codón colocado en el lado opuesto de este grupo es AGG, que también codifica para Arg. A primera vista, esto podría sugerir que el uso de tripletes codificando Arg es la tendencia biológica en relación con la segunda fuente de variación en el uso de codones en los genes humanos. Sin embargo, este no es el caso por dos razones: en primer lugar, como se mencionó anteriormente, de los cinco tripletes más extremos sólo uno codifica para este residuo y, en segundo lugar, AGA (el otro miembro de esta familia con seis miembros sinónimos) no se encuentra "cerca" de AGG a lo largo de este eje (véase la figura 49).

La segunda posibilidad es que esta fuente de variación esté relacionada con la existencia de islas CpG. De hecho, en el genoma de mamíferos hay regiones que fueron tempranamete llamdas HTF (fragmentos diminutos HpaII), que suman aproximadamente ~ 1-2% del genoma, y se caracterizan por un a) aumento del contenido en GC y b) la presencia de agrupaciones de dinucleótidos CpG no metilados. Estas regiones están dispersas por todo el genoma en islas de ~2.000 pb, y existen aproximadamente 30.000 islas en el genoma humano haploide.

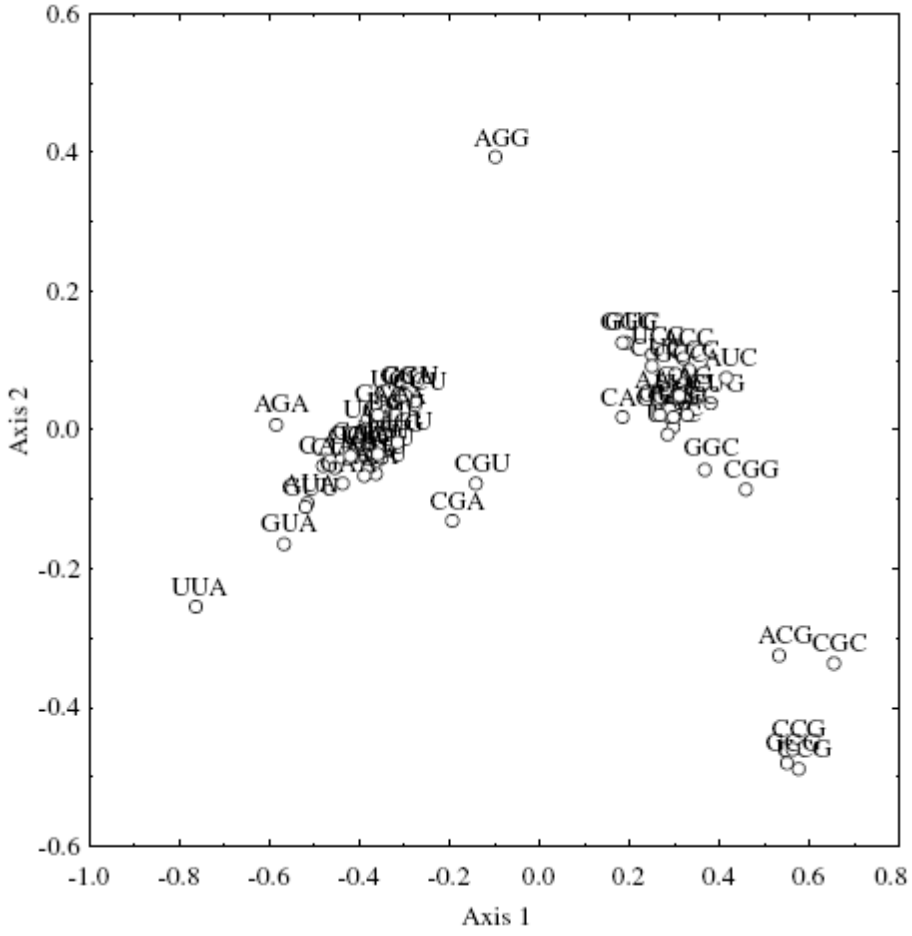


Figura 49. Distribución de los codones de los genes humanos sobre el plano definido por los dos ejes principales del CoA.

Se ha sugerido que aproximadamente el 60% de todos los genes humanos están asociados a una isla típicamente en el extremo 5', y cerca del 85% de las islas se ha determinado que están dentro de la zona comprendida entre -500 pb a +1.500 pb del sitio de inicio de la transcripción (Brown y Bird 1986; Antequera y Bird 1993; Cross y Bird 1995). Esta asociación con regiones promotoras sugieren que factores actuando en trans se unen a estos sitios y evitan la metilación y posterior degradación de la citosina (Antequera 2003). En consonancia con esta función, se ha demostrado que la metilación de novo de las islas puede resultar en represión transcripcional (Bird 2002).

Para probar si la asociación con estas importantes regiones influye en el uso de codones en el genoma humano, se ordenaron los 11.657 genes utilizados en este estudio de acuerdo con su posición sobre el segundo eje y, a continuación, las secuencias que mostraban los valores extremos de acuerdo con el segundo eje del CoA (110 de cada extremo) fueron localizados en el cromosoma respectivo utilizando la opción BLAST en la base de datos de Ensembl. De acuerdo con la posición de los codones (ver figura 49) se esperaba que los genes que muestran los valores más negativos estuvieran representados por secuencias localizadas en islas CpG. Con los criterios descritos en Materiales y Métodos, se pudieron localizar 102 genes, y de estos, 75 (73,5%), se encontraban efectivamente en islas CpG. Por otra parte, de los 110 genes exhibiendo los valores más positivos en el eje 2, 99 pudieron ser localizados con precisión, y de éstas 98 (99%) no se asociaron con islas CpG. Estos resultados apoyan firmemente que el estar o no asociado a una isla CpG es el segundo factor principal modelando el uso de codones en el genoma humano (y posiblemente en otros mamíferos).

A fin de confirmar esta hipótesis, se correlacionaron la posición de cada secuencia en el eje 2 con las respectivas frecuencias del dinucleótido CpG, discriminando entre posiciones del codón 1-2, 2-3 y 3-1. Como puede observarse en la figura 50, la frecuencia de CpG aumenta significativamente en las tres posiciones consideradas en los genes que muestran los valores más negativos en el segundo eje. Es interesante observar que, si bien algunos otros dinucleótidos se correlacionan con la posición de los genes en ese eje, los valores de R^2 son sólo marginalmente significativos. Por lo tanto, llegamos a la conclusión de que la presencia de una secuencia dentro de una isla CpG es, en efecto, el segundo factor principal modelando el uso de codones a nivel global en los genes humanos.

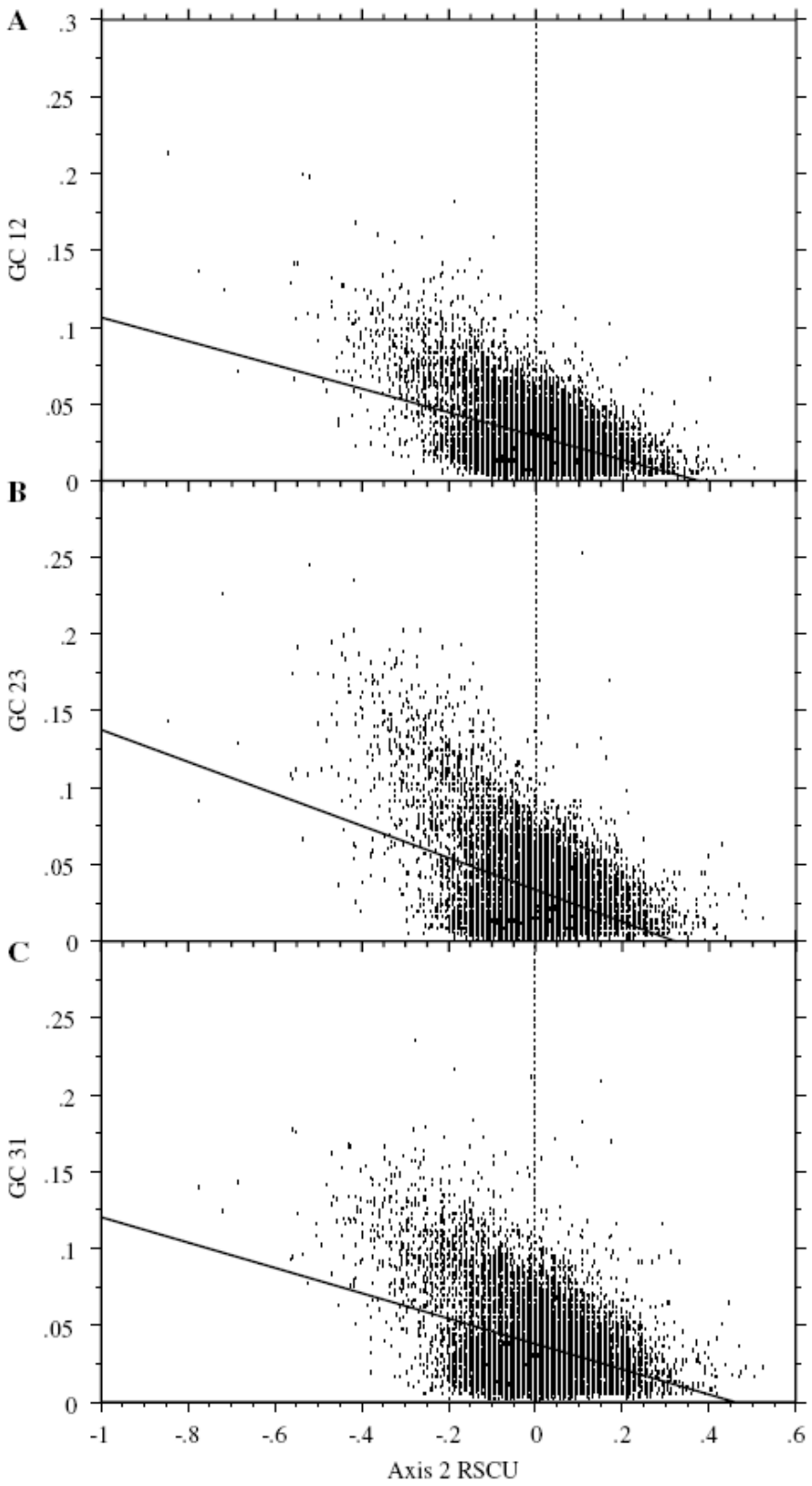


Figura 50. Frecuencias de CpG en las posiciones de los codones 1-2 (a), 2-3 (b), y 3-1 (c) de los genes contra su respectiva posición sobre el segundo eje del CoA. Los valores son a: R^2= 0,2 p < 0,0001; b: R^2 = 0,19 p < 0,0001; c: R^2 = 0,13 p < 0,0001.

Dado que el CoA define una serie de ejes ortogonales, se puede concluir que no necesariamente los genes que muestran altos niveles de GC en la tercera posición se encuentran en islas CpG. Esto es importante porque por lo general se ha asumido que estas estructuras están asociadas con genes situados en isocoros ricos en GC y, a su vez, estas secuencias se asume que son siempre ricas en GC_3, dada la fuerte correlación composicional positiva que existe entre ambas variables [véase, por ejemplo, (Bernardi 2000)]. En otras palabras, nuestros resultados sugieren que una fracción de genes incorporados en las islas CpG no se encuentran en isocoros ricos en GC.

En el transcurso de la tesis se desarrollaron una serie de herramientas que se utilizaron para entender la estructura génica y cromosómica desde el punto de vista composicional; es de destacar que parte de estos aplicativos de software se colocaron en internet para su uso público.

Estas herramientas nos permitieron caracterizar 2 espacios en el genoma humano, el espacio homogéneo y el heterogéneo cada uno con características diferentes en cuanto a su perfil de GC y presencia de genes. Además constatamos la importancia del órden en los fragmentos de ADN en el genoma humano.

Hemos realizado un análisis exhaustivo de las propiedades composicionales de cada cromosoma humano, y mostrado que estas entidades son unidades composicionales coherentes. La ya conocida relación entre las propiedades composicionales de las regiones de ADN y la expresión a nivel génico, ciclo de replicación y frecuencias de recombinación se volvió a examinar y ampliar a todo el cromosoma. Además, los aspectos composicionales y funcionales de los cromosomas se relacionan con la posición nuclear y el tamaño del cromosoma. Demostramos la existencia de un fuerte vínculo entre la composición y las características estructurales/funcionales a nivel cromosómico, junto con el hecho de que los cromosomas presentan un cierto nivel de coherencia de composición, y se comportarían como unidades de selección. En este sentido, creemos que nuestro trabajo contribuye a la tendencia general de los últimos años que re-enmarca los aspectos de la organización y función cromosómica desde una nueva perspectiva, más allá del papel histórico como unidades de segregación y recombinación.

Encontramos que la principal fuerza que modela el uso de aminoácidos en las proteínas humanas está dada por el contenido en GC de marco de ADN donde se encuentra ubicado cada gen. En segundo lugar está la influencia del nivel de hidropatía de cada proteína y la frecuencia de hojas β. Finalmente, el tercer factor se relaciona con el uso de Cys y la frecuencia de α-hélices. En conjunto, estos tres primeros factores explican el 45% de la variabilidad total encontrada en las proteínas humanas. Dado que cada uno de los ejes consecutivos explican una proporción más baja de la variabilidad total (por ejemplo, el cuarto corresponde al 8,1%) la imagen que hemos mostrado, aunque incompleta, es una buena representación de los factores globales que modelan la arquitectura de nuestras proteínas y muy probablemente, la de otros mamíferos.

En lo que refiere al uso de codones confirmamos que el principal factor modelando el uso de codones en el genoma humano es el contenido en GC de la tercera posición de los codones. A su vez, y esto es un aporte original, encontramos que la segunda fuerza es la asociación de los genes con las islas CpG. Dado que este factor es independiente del factor antes mencionado, nuestros resultados sugieren que una fracción de genes incorporados en las islas CpG no se encuentran en isocoros ricos en GC. Es necesario señalar que estos resultados no son contradictorios con resultados que muestran que algunos codones se encuentran bajo presión selectiva. Por ejemplo, se ha sugerido que ciertas mutaciones 'sinónimas' afectan el splicing y/o la estabilidad del mRNA, y que hay un contenido en GC específico de genes que se expresan únicamente en un determinado tejido [para una revisión, véase (Chamary et al. 2006)]. Lo que se muestra aquí se enmarca en tendencias principales que actúan globalmente en el genoma humano (es decir, afectan a la inmensa mayoría de los genes) y, por lo tanto, deben entenderse como complementarias con otros factores que, por cierto, siguen en vigor para secuencias individuales o grupos de genes. Naturalmente, esto sería nuevamente aplicable también a otros mamíferos.

Aissani, B. y Bernardi, G. (1991). "CpG islands, genes and isochores in the genomes of vertebrates." Gene **106**(2): 185-195.
Aissani, B., et al. (1991). "The compositional properties of human genes." J Mol Evol **32**(6): 493-503.
Akashi, H. (2001). "Gene expression and molecular evolution." Curr Opin Genet Dev **11**(6): 660-666.
Akashi, H. y Eyre-Walker, A. (1998). "Translational selection and molecular evolution." Curr Opin Genet Dev **8**(6): 688-693.
Akashi, H. y Gojobori, T. (2002). "Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis." Proc Natl Acad Sci U S A **99**(6): 3695-3700.
Antequera, F. (2003). "Structure, function and evolution of CpG island promoters." Cell Mol Life Sci **60**(8): 1647-1658.
Antequera, F. y Bird, A. (1993). "Number of CpG islands and genes in human and mouse." Proc Natl Acad Sci U S A **90**(24): 11995-11999.
Aota, S. y Ikemura, T. (1986). "Diversity in G + C content at the third position of codons in vertebrate genes and its cause." Nucleic Acids Res **14**(16): 6345-6355.
Bernardi, G. (1986). "Compositional constraints and genome evolution." J Mol Evol **24**(1-2): 1-11.
Bernardi, G. (1986). "The human genome and its evolutionary context." Cold Spring Harb Symp Quant Biol **51 Pt 1**: 479-487.
Bernardi, G. (1989). "The isochore organization of the human genome." Annu Rev Genet **23**: 637-661.
Bernardi, G. (1990). "Compositional patterns in the nuclear genome of cold-blooded vertebrates." J Mol Evol **31**(4): 265-281.
Bernardi, G. (1990). "Compositional transitions in the nuclear genomes of cold-blooded vertebrates." J Mol Evol **31**(4): 282-293.
Bernardi, G. (1991). "The human genome: a view from above." Boll Soc Ital Biol Sper **67**(5): 459-474.
Bernardi, G. (2000). "Isochores and the evolutionary genomics of vertebrates." Gene **241**(1): 3-17.
Bernardi, G. (2001). "Misunderstandings about isochores. Part 1." Gene **276**(1-2): 3-13.
Bernardi, G. (2004). Structural and evolutionary genomics. Natural selection in genome evolution. Amsterdam, Elsevier.
Bernardi, G. (2007). "The neoselectionist theory of genome evolution." Proc Natl Acad Sci U S A **104**(20): 8385-8390.
Bernardi, G., et al. (1988). "Compositional patterns in vertebrate genomes: conservation and change in evolution." J Mol Evol **28**(1-2): 7-18.
Bernardi, G., et al. (1985). "The mosaic genome of warm-blooded vertebrates." Science **228**(4702): 953-958.
Bettecken, T., et al. (1992). "Compositional mapping of the human dystrophin-encoding gene." Gene **122**(2): 329-335.
Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.
Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature **321**(6067): 209-213.
Blencowe, B. J. (2000). "Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases." Trends Biochem Sci **25**(3): 106-110.
Bolzer, A., et al. (2005). "Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes." PLoS Biol **3**(5): e157.
Brown, W. R. y Bird, A. P. (1986). "Long-range restriction site mapping of mammalian genomic DNA." Nature **322**(6078): 477-481.

Carels, N. y Bernardi, G. (2000). "The compositional organization and the expression of the Arabidopsis genome." FEBS Lett **472**(2-3): 302-306.
Caron, H., et al. (2001). "The human transcriptome map: clustering of highly expressed genes in chromosomal domains." Science **291**(5507): 1289-1292.
Clay, O. y Bernardi, G. (2002). "Isochores: dream or reality?" Trends Biotechnol **20**(6): 237.
Clay, O., et al. (2001). "Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses." Gene **276**(1-2): 15-24.
Comeron, J. M. (2004). "Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence." Genetics **167**(3): 1293-1304.
Corneo, G., et al. (1968). "Isolation of the complementary strands of a human satellite DNA." J Mol Biol **33**(1): 331-335.
Costantini, M. y Bernardi, G. (2008). "Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes." Gene **410**(2): 241-248.
Costantini, M. y Bernardi, G. (2008). "Replication timing, chromosomal bands, and isochores." Proc Natl Acad Sci U S A **105**(9): 3433-3437.
Costantini, M., et al. (2009). "The evolution of isochore patterns in vertebrate genomes." BMC Genomics **10**: 146.
Cox, E. C. y Yanofsky, C. (1967). "Altered base ratios in the DNA of an Escherichia coli mutator strain." Proc Natl Acad Sci U S A **58**(5): 1895-1902.
Cremer, T. y Cremer, M. (2010). "Chromosome territories." Cold Spring Harb Perspect Biol **2**(3): a003889.
Cross, S. H. y Bird, A. P. (1995). "CpG islands and genes." Curr Opin Genet Dev **5**(3): 309-314.
Cseresnyes, Z., et al. (2009). "Analysis of replication factories in human cells by super-resolution light microscopy." BMC Cell Biol **10**: 88.
Cuny, G., et al. (1981). "The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity." Eur J Biochem **115**(2): 227-233.
Chakalova, L., et al. (2005). "Replication and transcription: shaping the landscape of the genome." Nat Rev Genet **6**(9): 669-677.
Chamary, J. V., et al. (2006). "Hearing silence: non-neutral evolution at synonymous sites in mammals." Nat Rev Genet **7**(2): 98-108.
Chanda, I., et al. (2005). "Proteome composition in Plasmodium falciparum: higher usage of GC-rich nonsynonymous codons in highly expressed genes." J Mol Evol **61**(4): 513-523.
Chiapello, H., et al. (1998). "Codon usage and gene function are related in sequences of Arabidopsis thaliana." Gene **209**(1-2): GC1-GC38.
Chiusano, M. L., et al. (2000). "Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code." Gene **261**(1): 63-69.
D'Onofrio, G. y Bernardi, G. (1992). "A universal compositional correlation among codon positions." Gene **110**(1): 81-88.
D'Onofrio, G., et al. (2002). "The base composition of the genes is correlated with the secondary structures of the encoded proteins." Gene **300**(1-2): 179-187.
D'Onofrio, G., et al. (1999). "The correlation of protein hydropathy with the base composition of coding sequences." Gene **238**(1): 3-14.
D'Onofrio, G., et al. (1991). "Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins." J Mol Evol **32**(6): 504-510.
De Sario, A., et al. (1996). "A compositional map of human chromosome band Xq28." Proc Natl Acad Sci U S A **93**(3): 1298-1302.
De Sario, A., et al. (1997). "A compositional map of the cen-q21 region of human chromosome 21." Gene **194**(1): 107-113.

Duret, L. y Mouchiroud, D. (1999). "Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis." Proc Natl Acad Sci U S A **96**(8): 4482-4487.
Ermolaeva, M. D. (2001). "Synonymous codon usage in bacteria." Curr Issues Mol Biol **3**(4): 91-97.
Eyre-Walker, A. (1992). "The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals." Genet Res **60**(1): 61-67.
Fennoy, S. L. y Bailey-Serres, J. (1993). "Synonymous codon usage in Zea mays L. nuclear genes is varied by levels of C and G-ending codons." Nucleic Acids Res **21**(23): 5294-5300.
Filipski, J., et al. (1973). "An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation." J Mol Biol **80**(1): 177-197.
Freese, E. y Strack, H. B. (1962). "Induction of mutations in transforming DNA by hydroxylamine." Proc Natl Acad Sci U S A **48**: 1796-1803.
Fukagawa, T., et al. (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**(1): 184-191.
Garat, B. y Musto, H. (2000). "Trends of amino acid usage in the proteins from the unicellular parasite Giardia lamblia." Biochem Biophys Res Commun **279**(3): 996-1000.
Gardiner-Garden, M. y Frommer, M. (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-282.
Gardiner, K., et al. (1990). "A compositional map of human chromosome 21." EMBO J **9**(6): 1853-1858.
Garnier, J., et al. (1978). "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins." J Mol Biol **120**(1): 97-120.
Gierman, H. J., et al. (2007). "Domain-wide regulation of gene expression in the human genome." Genome Res **17**(9): 1286-1295.
Goetze, S., et al. (2007). "The three-dimensional structure of human interphase chromosomes is related to the transcriptome map." Mol Cell Biol **27**(12): 4475-4487.
Goetze, S., et al. (2007). "Three-dimensional genome organization in interphase and its relation to genome function." Semin Cell Dev Biol **18**(5): 707-714.
Graffelman, J., et al. (2007). "Variation in estimated recombination rates across human populations." Hum Genet **122**(3-4): 301-310.
Grantham, R., et al. (1980). "Codon catalog usage and the genome hypothesis." Nucleic Acids Res **8**(1): r49-r62.
Hinloopen, J. a. V. M., Charles. (2005). "Comparing Distributions: The Harmonic Mass Index." Tinbergen Institute Discussion Paper No. 05-122/1

Available at SSRN: http://ssrn.com/abstract=873831.
Hiratani, I., et al. (2009). "Replication timing and transcriptional control: beyond cause and effect--part II." Curr Opin Genet Dev **19**(2): 142-149.
Hubbard, T., et al. (2005). "Ensembl 2005." Nucleic Acids Res **33**(Database issue): D447-453.
Hurst, L. D. y Pal, C. (2001). "Evidence for purifying selection acting on silent sites in BRCA1." Trends Genet **17**(2): 62-65.
Hurst, L. D., et al. (2004). "The evolutionary dynamics of eukaryotic gene order." Nat Rev Genet **5**(4): 299-310.
Huvet, M., et al. (2007). "Human gene organization driven by the coordination of replication and transcription." Genome Res **17**(9): 1278-1285.
Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms." Mol Biol Evol **2**(1): 13-34.
Ikemura, T. y Aota, S. (1988). "Global variation in G+C content along vertebrate genome DNA. Possible correlation with chromosome band structures." J Mol Biol **203**(1): 1-13.

Ikemura, T., et al. (1990). "Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions." Genomics **8**(2): 207-216.

International HapMap Consortium, F. K., Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. (2007). "A second generation human haplotype map of over 3.1 million SNPs. ." Nature **449**: 851 -861.

Jones, S., et al. (1999). "Protein-DNA interactions: A structural analysis." J Mol Biol **287**(5): 877-896.

Kerem, B. S., et al. (1984). "Mapping of DNAase I sensitive regions on mitotic chromosomes." Cell **38**(2): 493-499.

Krane, D. E., et al. (1991). "Rapid determination of nucleotide content and its application to the study of genome structure." Nucleic Acids Res **19**(19): 5181-5185.

Kreil, D. P. y Ouzounis, C. A. (2001). "Identification of thermophilic species by the amino acid compositions deduced from their genomes." Nucleic Acids Res **29**(7): 1608-1615.

Kyte, J. y Doolittle, R. F. (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol **157**(1): 105-132.

Li, W., et al. (2003). "Isochores merit the prefix 'iso'." Comput Biol Chem **27**(1): 5-10.

Lobry, J. R. (1997). "Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species." Gene **205**(1-2): 309-316.

Lobry, J. R. y Chessel, D. (2003). "Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria." J Appl Genet **44**(2): 235-261.

Lobry, J. R. y Gautier, C. (1994). "Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes." Nucleic Acids Res **22**(15): 3174-3180.

Macaya, G., et al. (1976). "An approach to the organization of eukaryotic genomes at a macromolecular level." J Mol Biol **108**(1): 237-254.
Matassi, G., et al. (1989). "The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants." Nucleic Acids Res **17**(13): 5273-5290.
Mijalski, T., et al. (2005). "Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues." Proc Natl Acad Sci U S A **102**(24): 8621-8626.
Montero, L. M., et al. (1990). "Gene distribution and isochore organization in the nuclear genome of plants." Nucleic Acids Res **18**(7): 1859-1867.
Moses, A. M. y Durbin, R. (2009). "Inferring selection on amino acid preference in protein domains." Mol Biol Evol **26**(3): 527-536.
Mouchiroud, D., et al. (1991). "The distribution of genes in the human genome." Gene **100**: 181-187.
Mouchiroud, D., et al. (1988). "The compositional distribution of coding sequences and DNA molecules in humans and murids." J Mol Evol **27**(4): 311-320.
Musto, H., et al. (2001). "Translational selection on codon usage in Xenopus laevis." Mol Biol Evol **18**(9): 1703-1707.
Musto, H., et al. (2005). "The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor." Biochem Biophys Res Commun **330**(2): 357-360.
Musto, H., et al. (2006). "Genomic GC level, optimal growth temperature, and genome size in prokaryotes." Biochem Biophys Res Commun **347**(1): 1-3.
Musto, H., et al. (1999). "Compositional correlations in the chicken genome." J Mol Evol **49**(3): 325-329.
Musto, H., et al. (1999). "Synonymous codon choices in the extremely GC-poor genome of Plasmodium falciparum: compositional constraints and translational selection." J Mol Evol **49**(1): 27-35.
Myers, S., et al. (2006). "The distribution and causes of meiotic recombination in the human genome." Biochem Soc Trans **34**(Pt 4): 526-530.
Naya, H., et al. (2004). "Correspondence analysis of amino acid usage within the family Bacillaceae." Biochem Biophys Res Commun **325**(4): 1252-1257.
Nekrutenko, A. y Li, W. H. (2000). "Assessment of compositional heterogeneity within and between eukaryotic genomes." Genome Res **10**(12): 1986-1995.
Neusser, M., et al. (2007). "Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates." Chromosoma **116**(3): 307-320.
Nghiem, Y., et al. (1988). "The mutY gene: a mutator locus in Escherichia coli that generates G.C----T.A transversions." Proc Natl Acad Sci U S A **85**(8): 2709-2713.
Oliver, J. L., et al. (2001). "Isochore chromosome maps of eukaryotic genomes." Gene **276**(1-2): 47-56.
Oliver, J. L., et al. (2002). "Isochore chromosome maps of the human genome." Gene **300**(1-2): 117-127.
Osawa, S., et al. (1987). "Role of directional mutation pressure in the evolution of the eubacterial genetic code." Cold Spring Harb Symp Quant Biol **52**: 777-789.
Pavlicek, A., et al. (2002). "A compact view of isochores in the draft human genome sequence." FEBS Lett **511**(1-3): 165-169.
Peixoto, L., et al. (2004). "The effect of expression levels on codon usage in Plasmodium falciparum." Parasitology **128**(Pt 3): 245-251.
Perrin, P. y Bernardi, G. (1987). "Directional fixation of mutations in vertebrate evolution." J Mol Evol **26**(4): 301-310.
Pilia, G., et al. (1993). "Isochores and CpG islands in YAC contigs in human Xq26.1-qter." Genomics **17**(2): 456-462.

Plotkin, J. B., et al. (2004). "Tissue-specific codon usage and the expression of human genes." Proc Natl Acad Sci U S A **101**(34): 12588-12591.
Pollack, Y., et al. (1982). "The genome of Plasmodium falciparum. I: DNA base composition." Nucleic Acids Res **10**(2): 539-546.
Rispe, C., et al. (2004). "Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids." Genome Res **14**(1): 44-53.
Rocha, E. P., et al. (1999). "Universal replication biases in bacteria." Mol Microbiol **32**(1): 11-16.
Romero, H., et al. (2000). "Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote Entamoeba histolytica." Gene **242**(1-2): 307-311.
Romero, H., et al. (2003). "The influence of translational selection on codon usage in fishes from the family Cyprinidae." Gene **317**(1-2): 141-147.
Rynditch, A., et al. (1991). "The isopycnic, compartmentalized integration of Rous sarcoma virus sequences." Gene **106**(2): 165-172.
Sabbia, V., et al. (2009). "Composition profile of the human genome at the chromosome level." J Biomol Struct Dyn **27**(3): 361-370.
Saccone, S., et al. (2002). "Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds." Gene **300**(1-2): 169-178.
Salinas, J., et al. (1988). "Compositional compartmentalization and compositional patterns in the nuclear genomes of plants." Nucleic Acids Res **16**(10): 4269-4285.
Salinas, J., et al. (1986). "Gene distribution and nucleotide sequence organization in the mouse genome." Eur J Biochem **160**(3): 469-478.
Semon, M., et al. (2005). "Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance." Hum Mol Genet **14**(3): 421-427.
Sharp, P. M., et al. (1995). "DNA sequence evolution: the sounds of silence." Philos Trans R Soc Lond B Biol Sci **349**(1329): 241-247.
Sharp, P. M., et al. (1988). "Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity." Nucleic Acids Res **16**(17): 8207-8211.
Sharp, P. M. y Matassi, G. (1994). "Codon usage and genome evolution." Curr Opin Genet Dev **4**(6): 851-860.
Sharp, P. M., et al. (1986). "Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes." Nucleic Acids Res **14**(13): 5125-5143.
Shields, D. C., et al. (1988). ""Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons." Mol Biol Evol **5**(6): 704-716.
Singer, G. A. y Hickey, D. A. (2000). "Nucleotide bias causes a genomewide bias in the amino acid composition of proteins." Mol Biol Evol **17**(11): 1581-1588.
Singer, G. A., et al. (2005). "Clusters of co-expressed genes in mammalian genomes are conserved by natural selection." Mol Biol Evol **22**(3): 767-775.
Sproul, D., et al. (2005). "The role of chromatin structure in regulating the expression of clustered genes." Nat Rev Genet **6**(10): 775-781.
Stenico, M., et al. (1994). "Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases." Nucleic Acids Res **22**(13): 2437-2446.
Stephens, R., et al. (1999). "Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC." J Mol Biol **291**(4): 789-799.
Sueoka, N. (1961). "Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein." Proc Natl Acad Sci U S A **47**(8): 1141-1149.
Sueoka, N. (1988). "Directional mutation pressure and neutral molecular evolution." Proc Natl Acad Sci U S A **85**(8): 2653-2657.
Sueoka, N. (1992). "Directional mutation pressure, selective constraints, and genetic equilibria." J Mol Evol **34**(2): 95-114.

Suhre, K. y Claverie, J. M. (2003). "Genomic correlates of hyperthermostability, an update." J Biol Chem **278**(19): 17198-17202.
Tazi, J. y Bird, A. (1990). "Alternative chromatin structure at CpG islands." Cell **60**(6): 909-920.
Team, R. D. C. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Vienna.
Tekaia, F., et al. (2002). "Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis." Gene **297**(1-2): 51-60.
Thiery, J. P., et al. (1976). "An analysis of eukaryotic genomes by density gradient centrifugation." J Mol Biol **108**(1): 219-235.
Urrutia, A. O. y Hurst, L. D. (2003). "The signature of selection mediated by expression on human genes." Genome Res **13**(10): 2260-2264.
Venter, J. C., et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.
Versteeg, R., et al. (2003). "The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes." Genome Res **13**(9): 1998-2004.
Vinogradov, A. E. (2001). "Intron length and codon usage." J Mol Evol **52**(1): 2-5.
Wang, G. Z. y Lercher, M. J. (2010). "Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes." BMC Evol Biol **10**(1): 263.
Wen, S. Y. y Zhang, C. T. (2003). "Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis." Biochem Biophys Res Commun **311**(1): 215-222.
Willie, E. y Majewski, J. (2004). "Evidence for codon bias selection at the pre-mRNA level in eukaryotes." Trends Genet **20**(11): 534-538.
Wolfe, K. H., et al. (1989). "Mutation rates differ among regions of the mammalian genome." Nature **337**(6204): 283-285.
Woodfine, K., et al. (2004). "Replication timing of the human genome." Hum Mol Genet **13**(2): 191-202.
Wu, E. D., et al. (1990). "Identification of the mutations in the prfB gene of Escherichia coli K12, which confer UGA suppressor activity." Jpn J Genet **65**(3): 115-119.
Yang, H., et al. (2009). "Evolutionary pattern of protein architecture in mammal and fruit fly genomes." Genomics **93**(1): 90-97.
Yanofsky, C., et al. (1966). "Amino acid replacements and the genetic code." Cold Spring Harb Symp Quant Biol **31**: 151-162.
Zavala, A., et al. (2002). "Trends in codon and amino acid usage in Thermotoga maritima." J Mol Evol **54**(5): 563-568.
Zhang, C. T. y Zhang, R. (2003). "An isochore map of the human genome based on the Z curve method." Gene **317**(1-2): 127-135.
Zhang, C. T. y Zhang, R. (2004). "Isochore structures in the mouse genome." Genomics **83**(3): 384-394.
Zoubak, S., et al. (1992). "Compositional bimodality and evolution of retroviral genomes." Gene **119**(2): 207-213.