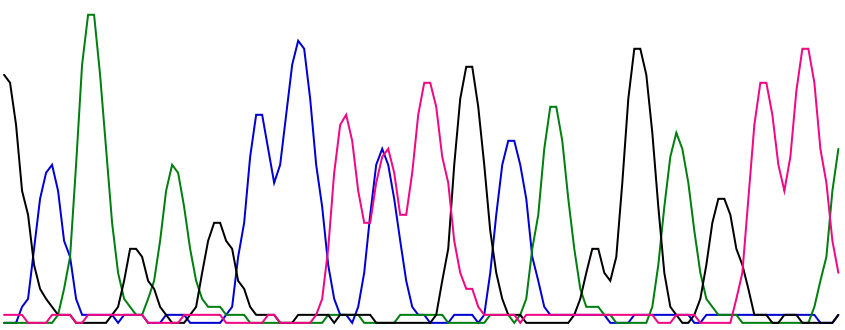
VARIACIÓN HAPLOIDE EN SECUENCIAS NUCLEARES HUMANAS: EL PSEUDOGÉN GBA



ROSA MARTÍNEZ ARIAS

2001

Dipòsit legal: B.7786-2004
ISBN: 84-688-5717-3

Departamento de Ciencias Experimentales y de la Salud
UNIVERSITAT POMPEU FABRA

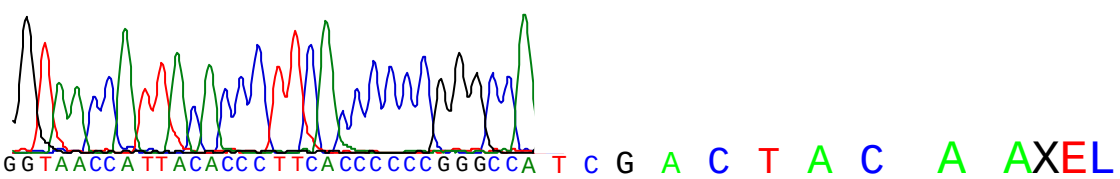
VARIACIÓN HAPLOIDE EN SECUENCIAS NUCLEARES HUMANAS: EL PSEUDOGÉN GBA

Memoria presentada por Rosa Martínez Arias para optar al grado de Doctor por la Universitat Pompeu Fabra.

Este trabajo se ha realizado bajo la dirección de Jaume Bertranpetit i Busquets, en la Unidad de Antropología del Departamento de Biología Animal de la Facultad de Biología, Universidad de Barcelona, y en la Unidad de Biología Evolutiva del Departamento de Ciencias Experimentales y de la Salud, Universitat Pompeu Fabra. Se ha englobado dentro de los Programas de Doctorado: Biología Animal II, Antropología Biológica, Universidad de Barcelona (bienio 1997-1999) y Ciencias de la Salud y de la Vida, Universitat Pompeu Fabra (bienio 1998-2000)

Jaume Bertranpetit i Busquets
Director de la Tesis

Rosa Martínez Arias


GGTAACCATTACACCCTTCACCCCCCGGGCCA T C G A C T A C A AXEL

AGRADECIMIENTOS

A lo largo de cuatro años mucha gente ha colaborado de una manera u otra en la realización de este trabajo. Quiero expresarles ahora mi agradecimiento:

A Jaume Bertranpetit, por haberme aceptado en su grupo de investigación. Le agradezco muchísimo su optimismo y su entusiasmo, el tiempo que ha dedicado a este trabajo y ese arte tan especial de convertir en buenos los puntos malos.

A los que me han acompañado en el día a día del laboratorio. A Aida Andrés por los repetitivos "*gene cleans*" y por su paciencia al ayudarme en el laboratorio. A Elena Bosch por su ayuda con los cultivos celulares (y los múltiples traslados del material y las células) y por la información sobre las últimas "formalidades" necesarias en la tesis. A Francesc Calafell, por contestar a mil preguntas, por estar siempre dispuesto a discutir sobre el efecto del efecto del efecto... por todas las explicaciones que no se ha cansado de darme, por las numerosísimas lecturas y correcciones con las que tanto me ha ayudado. A Jordi Clarimón por sus explicaciones sobre estadística y la burocracia que hizo por mí. A David Comas, por las secuencias que ha repasado (y aquellos heterozigotos escondidos...), por las lecturas tediosas a los manuscritos y, como no, por los consejos de buen hermano mayor (mayor pero muy joven!). A Oscar Lao, por la ayuda con la bibliografía y su querida fórmula de Ewens. A Eva Mateu, por tantas buenas horas, y tanto trabajo (y *papers*, y terapia...) compartidos. A Anna Pérez, por los consejos sobre el Mac (y por dejarme ocupar un buen espacio en él!), y por las cuatro últimas secuencias de este trabajo, que pasaron por el *servei de seqüenciació*. A Stéphanie Plaza por las clases de francés y por cuidar tan bien del 377 durante el tiempo en que lo "compartimos". A Marta Soldevila, por el muérdago de la suerte. A Mónica Vallès por mil pequeñas cosas que han hecho que el trabajo en el laboratorio fuera más fácil. Y a todos ellos y también al Dr. Josep Pons, a Bárbara Arias, Lourdes Fañanás, Blanca Gutiérrez, Cristina Junyent, Carles Lalueza, Araceli Rosa y Neus Valveny, por las anécdotas del día a día, por los consejos de laboratorio, por los ánimos de trabajo y también en las crisis "post-aeropuertarias", por hacerme reír tanto, por los cafés compartidos en las inolvidables y divertidísimas sobremesas, y por todos los ratos entrañables que me han hecho pasar.

A David De Lorenzo, por las explicaciones sobre selección, por los largos e-mails con más y más y más preciadísimas explicaciones sobre el maravilloso mundo de los tests, y por contestar siempre con buen humor a tantas preguntas tontas...

A Julio Rozas, por las explicaciones sobre diversidad y evolución, y por haber creado un programa tan fantástico como el DnaSP.

A Montse Aguadé y Carme Segarra, por las tan útiles clases de evolución.

A Luís Pérez-Jurado y Mónica Bayés, por ayudarnos a entender "cosas de genética" como la conversión génica y la formación de heterodúplex.

A la gente del Departamento de Genética de la UB que nos ha prestado siempre ayuda en momentos de crisis. A Marc Valls (que me enseñó a clonar!), a Gemma Marfany, por los consejos y la *speedvac*, a Agustí Munté por los consejos sobre las "columnitas". A Montse Papaceit, de la Unidad de Evolución, por permitirme dar los primeros pasos en un laboratorio, y a Elvira Juan y Francesc Mestres, por su ayuda mientras estuve allí.

A Ramón, Anna y Amaya, de los servicios científico-técnicos de la UB.

A David Moreso y Alex García por su valiosísima ayuda en los pequeños y grandes problemas que he tenido con los ordenadores (y sus alrededores).

A Elena, Jose, Carles, Oscar, Magano, Raquel, y a todos los "nuevos" compañeros del CEXS por animar el departamento (y algunos fines de semana en el lab). A Anna Magre, Josep Gibert y Marina Losada por su ayuda a conseguir tantos artículos. A Tàmara Monràs, Carmen Trigueros, Carme Abelló, Montserrat Saladrigas, Lluïsa Tarragó, Núria Margalef, Xavier Ardite, Constantí Serrano, Cristòfol Moreno, Pietat Estany, Octavi Izquierdo, Lydia García, y a todo el personal administrativo y de soporte de la UPF, por hacer la vida diaria (y burocrática!) más fácil.

A Jordi Camí, Miguel López-Botet, y Luís Pérez-Jurado por ser tan buen tribunal del DEA.

Las estancias en otros laboratorios han sido importantísimas para el desarrollo de este trabajo, y muy enriquecedoras desde el punto de vista personal. Mis más sinceras gracias a todos los que me han aceptado en sus grupos de trabajo:

De la Unidad de Hematología Molecular en el Instituto de Medicina Molecular del John Radcliffe Hospital en Oxford, a John B. Clegg, Yan Tat-Liu, Nicole, Marc y Julie.

Del Departamento de Antropología de la Universidad de Penn State en Pensilvania, a Kenneth M. Weiss por su gran afabilidad y su buen humor, por las discusiones sobre el trabajo, y por mantenerme al día de los *papers*. A Anne Buchanan (mi Hada Madrina) por su inestimable ayuda con todos los problemas que tuvimos en el laboratorio, por el tiempo que me dedicó, y su inagotable simpatía. A Frances Hayashida, por su enorme generosidad, por acogerme tan bien y hacerme sentir como en casa. A Malia Fullerton, por los consejos y las dudas resueltas. A Esteban Parra, por su alegría contagiosa y sus explicaciones sobre bases de datos. A Marc Shriver, por escuchar y darme consejos. A David, Carolina, Carry, Blanca, Amie, Ellen, a la gente del laboratorio, y a todos los que hicieron que mi estancia allí fuera inolvidable.

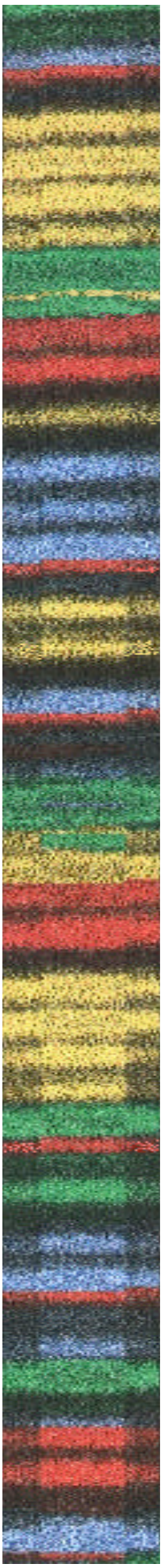
Del Centro de Investigación sobre el Cáncer Charles Bruneau, en el Hospital Ste-Cathérine en Montreal, a Damian Labuda y Ewa Zietkiewicz, por su gran simpatía, por las pacientes explicaciones y por las horas que pasaron tratando de “resolver” la filogenia de haplotipos. Muchísimas gracias a Andrew, a Vania y a Mme. Lavergne por enseñarme Montreal y hacerme sentir tan a gusto.

Gracias también a Eva, Miguel Ángel, José, Sión, a la familia Indorf, a la familia Martínez-Arias y a la familia Schechert, por darme apoyo, cariño y ánimos. Mil gracias a Montse y a David, por tantas pequeñas cosas tan importantes y por estar siempre ahí!

Gracias a Axel, por estar siempre cerca durante tanto tiempo y a pesar de tanta distancia, por su gran paciencia, por su apoyo total y constante. (That’s what I’ve been doing here these years!...)

Finalmente, quiero expresar mi agradecimiento a toda la gente que ha permitido que su DNA se utilice para que estudios como este sean posibles.

Durante el tiempo que ha durado este trabajo he disfrutado de una beca de colaboración para estudiantes de tercer ciclo de la Universidad de Barcelona, y de una beca FPI del Ministerio de Educación y Cultura (MEC), AP96. Tengo que agradecer también al MEC las becas para las visitas a otros laboratorios que me ha concedido durante tres años. Este trabajo ha contado también con el apoyo de los siguientes organismos y proyectos: Dirección General de Investigación Científica y Técnica, proyecto PB98-1064; Generalitat de Catalunya: Grup de Recerca Consolidat 1998SGR00009; Acció integrada amb Quebec DGR-CIRIT ABM/ACS/HI 1998-17.



ÍNDICE

NOTA: Las Figuras y Tablas en los capítulos preparados como artículos independientes (apartado 2 de Material y Métodos, y capítulos I, II, III y IV de Resultados) siguen una numeración independiente a la del resto del trabajo.

PRESENTACIÓN

Variation is not an incidental aspect but has alwa ys been intrinsic to the human genome. Evolution works by selecting from available biological variation; this is the process by which genes, DNA sequences, and chromosomes have come to be in their current configuration, and there has never been a time when any species had an invariant genome.

Kenneth M. Weiss

Este estudio es una contribución al análisis de la diversidad en el DNA humano. Hemos trabajado con una zona no codificante autosómica, el pseudogén homólogo al gen de la glucocerebrosidasa o gen GBA, de la que hemos analizado la variabilidad en una muestra de representación mundial. El análisis de la diversidad en el pseudogén GBA (psGBA) se ha realizado desde la perspectiva de la genética de poblaciones, desde la que hemos tratado de hacer inferencias aplicables a la historia de las poblaciones humanas. Se ha tomado también un punto de vista de la evolución molecular más genómico para establecer la dinámica de la región con la que hemos trabajado y para entender las causas del espectro de variación obtenido. Para ello, hemos analizado el papel de fenómenos como la mutación, la recombinación, la conversión génica y, especialmente, la selección. Hemos detectado que la selección, a través de su efecto en el contexto genómico cercano a psGBA, ha tenido un papel crucial en la formación de la variabilidad en este *locus*. Por otra parte, mutaciones en el homólogo funcional de psGBA son la causa principal de la enfermedad de Gaucher. Se sabe que psGBA es importante en la producción de alelos complejos GBA-psGBA, que llevan a los tipos más severos de este trastorno. Hemos intentado ver cómo el conocimiento de la variabilidad en psGBA ayuda al reconocimiento de estos alelos complejos, y por tanto al entendimiento de las bases moleculares de la enfermedad de Gaucher. Finalmente, el disponer de datos de variabilidad de dos regiones parálogas, un gen y un pseudogén, situadas en la misma región cromosómica (GBA, de variación ya conocida, y psGBA, analizado en este trabajo), nos ha permitido comparar los patrones de mutación que presenta una secuencia bajo diferentes presiones selectivas.

INTRODUCCIÓN

I know nothing about genetics, but genetics knows about me.

Günter Grass

1. ANÁLISIS DE DIVERSIDAD EN EL DNA

A pesar de que la secuenciación completa del genoma humano se halla en las fases finales, su entendimiento apenas empieza. Continúa la investigación para comprender e interpretar la secuencia del genoma. Entre los aspectos abiertos destaca la comprensión de la diversidad existente en el genoma, y el entendimiento de la distribución de esta diversidad.

El entendimiento de la variabilidad de una región genómica ayuda a la comprensión de los mecanismos de producción, mantenimiento y eliminación de variabilidad, como mutación, recombinación y conversión génica, desde el punto de vista genómico, y deriva genética, migración y selección desde el punto de vista poblacional. Por otra parte, los análisis genéticos también suponen una importante contribución al conocimiento de la historia de las poblaciones. El análisis de la diversidad en un *locus* concreto puede llevar a un mayor conocimiento sobre cuestiones históricas específicas de las poblaciones en las que se ha analizado este *locus*. La estructura de la variación genética puede mostrar patrones demográficos, tales como expansiones antiguas, reducciones drásticas de la población, o migraciones. Estos estudios también permitirán la reconstrucción de la filogenia de la especie si la muestra es de representación global y contamos con datos de variación en otras especies. Los dos aspectos mencionados serían los dos objetivos principales de los análisis genéticos: la comprensión de la región genómica en sí, y la de las poblaciones en las que se ha analizado esta región. Expuesto de manera inversa, el patrón de variabilidad de una región genómica se ve influenciado por factores genéticos y por factores poblacionales. Fenómenos de ambos tipos pueden resultar en patrones similares de variabilidad genética (Owens y King, 1999; Bertranpetit, 2000).

La variabilidad que los seres vivos portan en su material genético, el DNA, empezó a ser estudiada de manera rigurosa a principios de la década de los años 60. Los primeros estudios se realizaron con proteínas, comparando patrones de movilidad electroforética entre individuos de poblaciones y especies distintas. Una vez el análisis pudo alcanzar a la molécula de DNA, el número de estudios sobre patrones de variación creció rápidamente. Las mejoras en las técnicas de secuenciación del DNA y el desarrollo de programas informáticos capaces de manejar estos datos de secuencia han seguido desarrollándose y mejorando desde entonces. Este importante soporte biotecnológico, junto al desarrollo de modelos evolutivos para el tratamiento de los datos genéticos, han

estimulado en gran medida que los análisis de diversidad engloben cada vez a más especies, más poblaciones, y más regiones genómicas. Estos estudios contribuyen a que mejoremos nuestra comprensión de las relaciones evolutivas entre los seres vivos, históricas en la especie humana, y amplían nuestro conocimiento sobre los procesos genómicos.

Los primeros trabajos sobre variabilidad en la molécula de DNA humano se iniciaron con el trabajo de Cann, Stoneking y Wilson en 1987, y se centraron en el estudio de la diversidad en el DNA mitocondrial (mtDNA). El mtDNA ha sido desde entonces y durante la última década ampliamente utilizado en el estudio genético de las poblaciones humanas, debido a las propiedades que lo hacen único y que facilitan su estudio: herencia materna, ausencia de recombinación, alta tasa de sustitución, y alto número de copias. En especial, el análisis de la región de control ha demostrado ser particularmente informativo (Vigilant *et al*., 1991). Más tarde se añadieron los estudios con microsatélites y la porción no recombinante del cromosoma Y, que sería el equivalente al mtDNA, pero con transmisión paterna. Gran número de SNPs (*Single Nucleotide Polymorphism*) y microsatélites informativos están siendo identificados en el cromosoma Y humano en los últimos años (Underhill *et al*., 2000).

El campo de la genética de poblaciones humanas se ha ido ampliando cada vez más hacia el DNA nuclear. En los estudios sobre la variabilidad de secuencias de DNA, los cromosomas X e Y han sido los primeros objetivos de análisis. La amplificación de un solo alelo en hombres evita el problema de asignar la fase entre posiciones heterozigotas que presentan los autosomas. Las regiones que se han estudiado a nivel de secuencia en el cromosoma X humano hasta el momento son: el gen de la distrofina, en una región de 7,6 kilobases (kb) (Zietkiewicz *et al*., 1997; Zietkiewicz *et al*., 1998); el gen PDHA1, en una región de 1,8 y más tarde 4,2 kb (Hey, 1997; Harris y Hey, 1999); regiones intrónicas de 7 genes, con un total de 11,4 kb analizadas (Nachman *et al*., 1998); 10,2 kb de una región no codificante (Kaessman *et al*., 1999); y el gen ZFX, en una región de 1,1 kb (Jaruzelska *et al*., 1999). Los estudios que se han centrado en el cromosoma Y para el análisis de secuencias son los siguientes: se han analizado 18,3 kb del gen SRY (Whitfield *et al*., 1995), 2,6 kb de una zona flanqueante a un polimorfismo YAP (*Y Alu polymorphism*) (Hammer, 1995), el gen ZFY en una región de 0,7 kb (Dorit *et al*., 1995; Jaruzelska *et al*., 1999), y 81 kb correspondientes a cuatro genes: SMCY,UTY1, DBY, y DFFY (Shen *et al*., 2000).

El primer estudio que aplicó datos de variación de secuencias de DNA haploides autosómicas a la reconstrucción de la historia de las poblaciones humanas se basó en

una región de 3 kb del cromosoma 11 humano que incluye al gen de la β-globina, para la que se secuenciaron 48 cromosomas de individuos melanésicos, y más tarde 349 cromosomas de poblaciones de todo el mundo (Fullerton *et al*., 1994; Harding *et al*., 1997). El segundo *locus* analizado fue un fragmento del gen para la lipoproteína lipasa: se secuenciaron 9,7 kb en 142 cromosomas pertenecientes a tres poblaciones (Nickerson *et al*., 1998; Clark *et al*., 1998). Se ha estudiado también el gen ACE (*angiotensin converting enzyme*), aunque la muestra analizada es pequeña (cinco individuos afro-americanos y seis individuos euro-americanos) (Rieder *et al*., 1999), y 0,9 kb pertenecientes al *locus* del receptor de la melanocortina 1, que se analizaron en una muestra de representación mundial (Rana *et al*., 1999). El último estudio hasta la fecha se ha centrado en el gen de la apolipoproteina E, del que se han analizado 5,5 kb en cuatro poblaciones (Fullerton *et al*., 2000).

Si bien la dificultad de la obtención de haplotipos supone una importante desventaja en el estudio de *loci* autosómicos, su análisis (aparte del conocimiento *per se* de la diversidad del DNA en zonas nucleares) aporta una serie de ventajas a la genética de poblaciones. Las relaciones filogenéticas entre secuencias de mtDNA resultan en un árbol génico matrilineal, que no ofrece información del flujo génico a través de la línea paterna. El cromosoma Y presenta el rasgo equivalente, pero con las líneas paternas. Esta característica es una ventaja en el análisis de la evolución de mujeres y hombres por separado, pero una clara desventaja si el objetivo del análisis es la evolución del conjunto de una población, una especie, o de un *locus*. El estudio de secuencias autosómicas y del cromosoma X nos aportan a la vez información sobre la filogenia de ambas líneas, paterna y materna. Otro inconveniente de los estudios sobre mtDNA y cromosoma Y es que se trabaja con un tamaño de población efectiva (aquella parte de la población que puede pasar sus genes a la descendencia) menor al del DNA nuclear, con lo que el tiempo de coalescencia se acorta, reduciendo la profundidad temporal que alcanzan sus análisis. El tiempo medio de vida de la variabilidad nuclear autosómica es cuatro veces mayor (tres veces en el caso del cromosoma X) que el del mtDNA y cromosoma Y, por lo que para obtener una visión que alcance un pasado más lejano, en principio es más conveniente el estudio de *loci* autosómicos.

2. MECANISMOS DE GENERACIÓN Y MODIFICACIÓN DE LA DIVERSIDAD GENÓMICA

Bajo el modelo neutralista, la mayoría de los cambios moleculares en la evolución se deben a mutaciones neutras; la diversidad en la secuencia de DNA se introduce en la población por mutación, y su destino final (fijación o pérdida) será determinado por la deriva genética. La mayor parte de la variabilidad molecular intra-específica sería selectivamente neutra, y se mantendría por deriva genética. Los polimorfismos de DNA representarían una situación transitoria, en la que un alelo escogido al azar, por fenómenos de deriva, va camino de su fijación.

Sin embargo, frecuentemente la mutación y la deriva no son las únicas fuerzas que actúan para modelar los niveles de diversidad de un *locus*, y la consideración de otros factores puede llevar a modelos más cercanos a la realidad. La interacción entre mecanismos moleculares como mutación, recombinación, y conversión génica a nivel genómico, y deriva genética, migración, y selección a nivel poblacional, determinarán en gran manera la evolución de una secuencia de DNA.

Antes de describir la región genómica en la que hemos trabajado, describiremos brevemente algunos procesos que han podido modelar su evolución.

2.1. MECANISMOS MOLECULARES

MUTACIÓN

Las mutaciones, o cambios en la molécula de DNA, suponen la fuente primaria de creación de diversidad en el DNA. Las deleciones e inserciones hacen referencia a la eliminación o inserción de uno o más nucleótidos en el DNA. Las sustituciones suponen el reemplazo de una base nucleotídica por otra, y se separan en transiciones (si la sustitución es de una pirimidina por otra o de una purina por otra) y transversiones (si la sustitución se produce de una pirimidina a una purina o a la inversa). Determinados

aspectos relacionados con fenómenos mutacionales serán discutidos en una sección específica más adelante.

RECOMBINACIÓN

La recombinación homóloga supone la recombinación (*crossover*) en meiosis (o raramente en mitosis) entre secuencias de DNA homólogas, idénticas o altamente similares, y normalmente supone la rotura de cromátidas no hermanas de un par de homólogos, y la posterior unión de los fragmentos para generar nuevas cadenas recombinantes. El intercambio de DNA que se produce es equivalente, ya que la rotura y unión de las cromátidas se da en la misma posición en cada cromátida. Los intercambios por tanto se dan entre las mismas posiciones nucleotídicas de los alelos de un *locus* (Strachan y Read, 1996) (Figura 1a).

En la recombinación no homóloga (entrecruzamiento desigual, *unequal crossover*) la recombinación se da entre secuencias no alélicas, o entre cromátidas no hermanas de un par de homólogos. Las secuencias entre las que se da el entrecruzamiento comparten una alta homología, que posibilita el mal apareamiento. El intercambio resulta en una deleción en una de las cromátidas y una inserción en la otra, y por tanto no es equivalente pero sí recíproco. Este fenómeno ocurre principalmente en lugares donde hay repeticiones en tándem, de forma que la homología entre las distintas repeticiones facilita el apareamiento de repeticiones no alélicas (Strachan y Read, 1996) (Figura 1b).

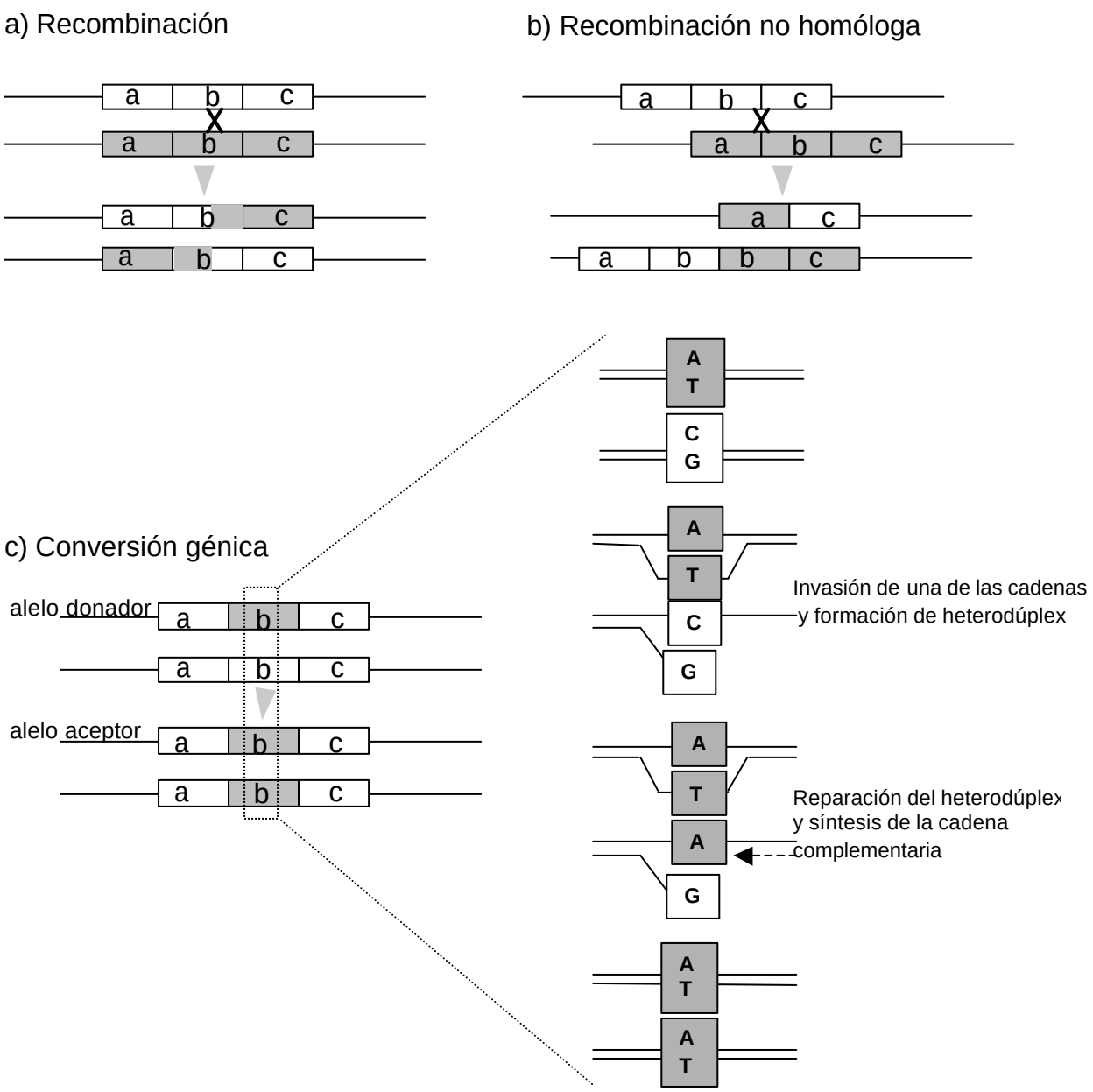
CONVERSIÓN GÉNICA

La conversión génica supone la modificación de un alelo (aceptor de información) determinada por otro que no resulta alterado (el donador de secuencia o de información), y puede abarcar fragmentos desde pocos a varios centenares de pares de bases. Un posible modelo molecular para explicar la conversión se inicia con la formación de heterodúplex, en mitosis o meiosis, entre una cadena del gen donador y la cadena complementaria del gen aceptor. La conversión se produciría entonces mediante la reparación de desapareamientos por los enzimas correctores de DNA, que harían a la secuencia aceptora perfectamente complementaria a la secuencia donadora en la zona convertida. Por último, la secuencia aceptora se replica, con lo que la conversión se completa (Figura 1c). La conversión puede ser interalélica (entre los alelos de un mismo *locus*), o bien *interlocus* (entre alelos de *loci* distintos entre los que existe homología) (Ohta, 1983; Strachan y Read, 1996). La conversión génica puede tener una dirección

preferente hacia la conversión de una de las secuencias homólogas en particular, lo que se denomina conversión génica sesgada (Li, 1997).

En organismos superiores (en los que los cuatro productos de la meiosis no pueden ser recuperados y estudiados) la conversión génica no puede demostrarse sin ambigüedad, ya que no es distinguible de un doble entrecruzamiento. No obstante, la probabilidad de un fenómeno de doble entrecruzamiento dentro de un segmento de pocas kilobases es extremadamente pequeña (Broman y Weber, 2000). Por otra parte, existen diferencias entre el entrecruzamiento desigual y la conversión génica. A diferencia del entrecruzamiento desigual, la conversión no causa cambio en el número de secuencias génicas. Así, la conversión puede suponer un mecanismo corrector u homogeneizador de secuencias, mientras que el entrecruzamiento sólo lleva a la deleción o inserción de fragmentos génicos (Li, 1997). En este sentido se ha sugerido que esta transferencia no recíproca de información intracromosómica puede ser un importante mecanismo que mantenga la homogeneidad de secuencias entre genes repetidos (Nagylaki, 1982).

Figura 1: Representación esquemática de algunos mecanismos moleculares para la generación de diversidad (**a**, **b** y **c** representan secuencias con alta homología).



2.2. MECANISMOS POBLACIONALES

DERIVA GENÉTICA

La deriva genética es una consecuencia de la transmisión aleatoria de alelos de una generación a la siguiente, con lo que las frecuencias alélicas pueden modificarse simplemente debido a procesos estocásticos. Fluctuaciones aleatorias en generaciones sucesivas pueden producir que a lo largo del tiempo las frecuencias alélicas se alejen cada vez más del estado inicial. La deriva tiene una gran influencia sobre poblaciones pequeñas, ya que el azar en la transmisión de alelos hará variar su composición genética más fácilmente que en poblaciones de gran tamaño. Cuanto menor sea la población, mayores serán los cambios en las frecuencias alélicas producidos por deriva. En consecuencia, este fenómeno tiene una especial importancia al considerar tiempos en la historia de la humanidad en los que los tamaños poblacionales eran reducidos, ya que el posterior crecimiento de la población hace que se mantenga en el tiempo el paisaje genético producido por deriva cuando el tamaño poblacional era pequeño.

MIGRACIÓN

El desplazamiento de individuos de un lugar a otro también puede producir cambios en la composición del acervo (o *pool*) genético. Si dos poblaciones diferentes genéticamente intercambian individuos a lo largo del tiempo, las características genéticas de la población de los individuos que migran se difundirán con ellos en la nueva población. Como resultado, las diferencias genéticas entre ambas poblaciones disminuirán. El efecto de la migración dependerá de la diferencia genética inicial que mantenían las poblaciones, y de la proporción de individuos que se desplace.

2.3. SELECCIÓN

La selección se sitúa en la frontera entre los mecanismos moleculares y los mecanismos poblacionales: depende de factores externos, pero sus efectos se manifiestan a nivel molecular en regiones más o menos extensas del genoma, pero no en el genoma en su conjunto. Esto es lo que nos ha llevado a considerar a la selección en una sección separada.

La selección natural puede ejercer diferentes grados de presión sobre una variante alélica. Esta presión es un término relativo ya que las probabilidades de vivir y/o reproducirse se asocian a un alelo o genotipo en relación a otro alelo o genotipo. La presión selectiva conferirá distintas probabilidades de sobrevivir y reproducirse (las cuales pueden variar también en ambientes distintos) al individuo portador de un determinado alelo, y por tanto hará incrementar o disminuir la frecuencia del alelo en la población en generaciones sucesivas. A un alelo que no está sometido a ningún tipo de presión selectiva se le denomina neutro.

A continuación definiremos brevemente algunos tipos generales de fenómenos selectivos.

SELECCIÓN POSITIVA

La selección positiva (o direccional) es la selección que se da sobre una variante alélica favorecida en relación a otras, en un *locus* concreto. Esta variante aumentará su frecuencia en la población hasta fijarse. El tiempo de fijación dependerá del tamaño efectivo de la población.

Una posible consecuencia de la selección positiva es el ARRASTRE GENÉTICO (*hitchhiking*, *genetic sweep*, o barrido selectivo), que se produce cuando una mutación es selectivamente favorable, y "arrastra" consigo a las variantes neutras ligadas a ella. En un primer momento, la variante seleccionada positivamente aumenta su frecuencia en la población hasta llegar a fijarse, eliminando la variación neutra ligada. A continuación se produce una fase de recuperación de la variación en zonas ligadas, a través de procesos de mutación y deriva, que va acompañada de un exceso transitorio de variantes alélicas únicas en la zona bajo arrastre (Kaplan *et al*., 1989; Charlesworth *et al*., 1993; De Lorenzo, 1998) (Figura 2a).

SELECCIÓN PURIFICADORA

Consiste en la eliminación de alelos deletéreos de la población. Es el caso bien conocido de eliminación de variantes genéticas que son letales o producen enfermedad (y que por tanto tienen una reducida eficacia biológica).

Un efecto paralelo al arrastre pero donde la selección determinante es purificadora en vez de positiva, es la SELECCIÓN DE FONDO (*background selection*), la eliminación de variantes deletéreas en un *locus* concreto de una población, que produce a su vez una reducción de la variabilidad neutra ligada a dichas variantes. La selección

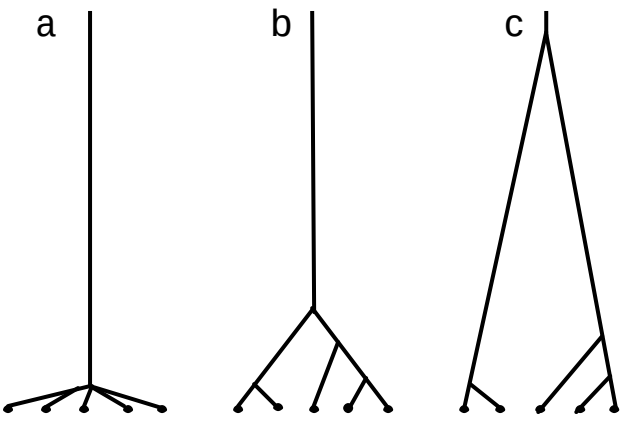
de fondo no varía la topología de la genealogía de los *loci* a los que afecta (como consecuencia de que no fija ninguna variante), sino que sólo reduce su tamaño (Charlesworth *et al*., 1993).

Tanto el arrastre genético como la selección de fondo representan casos de selección en los que el destino final de una variante nucleotídica depende del entorno genético en el que se encuentra, y no solamente del efecto que esta variante cause en la eficacia biológica del individuo. La influencia de ambos fenómenos depende de la intensidad de la selección en las variantes seleccionadas, y de la tasa de recombinación en la región cromosómica afectada, aumentando su efecto si la recombinación es baja, y viceversa.

SELECCIÓN BALANCEADORA

La selección balanceadora (*balancing selection*, selección equilibradora) se da cuando el *locus* que arrastra la variación ligada no es una mutación seleccionada positivamente, sino un polimorfismo mantenido por la selección. Es decir, los heterozigotos son favorecidos sobre ambos homozigotos. En la región ligada al *locus* seleccionado, la selección balanceadora mantiene durante largos períodos de tiempo polimorfismos que de otro modo serían eliminados de la población por deriva genética. De esta manera se retarda el tiempo de fijación de nuevas mutaciones y se incrementa el número de alelos. Como consecuencia, en los *loci* ligados las nuevas mutaciones tenderán a acumularse, por lo que existirá un exceso de posiciones segregantes mantenidas a frecuencias intermedias (Figura 2c).

Figura 2. Posibles genealogías génicas de una muestra aleatoria de cinco alelos, bajo arrastre genético reciente (a), bajo deriva genética (b), y bajo el efecto de un polimorfismo equilibrado (c) (modificado de Aguadé y Langley, 1994).



3. LA DUPLICACIÓN GÉNICA

La región genómica en la que se ha centrado este trabajo es fruto de una duplicación génica en tándem. Entender el fenómeno de duplicación es por tanto importante para la mejor comprensión de la zona de interés.

La duplicación génica hace referencia principalmente a mecanismos de recombinación no homóloga entre elementos genómicos. Transposición y retrotransposición de secuencias también originan elementos génicos duplicados, pero implican mecanismos genéticos distintos. La transposición supone que la copia de un elemento transponible se inserta en una región genómica distinta de la original, produciendo una copia. Si el elemento transponible ha incorporado a un segmento genómico, éste también resultará duplicado. La retrotransposición implica a una molécula de RNA mensajero (mRNA) como intermediaria en el proceso de duplicación. De ella se tratará más adelante.

Estudios en eucariotas han apuntado a que las tasas de duplicación génica son relativamente altas (del orden de entre 10^{-4} y 10^{-6} por *locus* y generación), si bien los valores varían mucho dependiendo del *locus* y la especie estudiados. Se ha sugerido que la tasa de duplicación espontánea depende de múltiples factores, como el *locus* de que se trate, la región genómica en que se sitúe (el *background* genético), la disponibilidad de mecanismos de recombinación y la especie (Shapira y Finnerty, 1986).

Para generar y mantener una variabilidad genética útil en los organismos, los mecanismos básicos parecen ser una estructura multigénica, la transferencia de dominios génicos y el uso combinatorial de la información genética. La duplicación génica es el mecanismo más importante para la creación de una estructura multigénica, posibilitando la incorporación de nuevos genes en el genoma y por tanto el desarrollo de nuevas funciones génicas. En relación a esto, el alto contenido en DNA de los organismos superiores parece haberse producido principalmente por duplicación génica. En general, el incremento de DNA en el genoma se relaciona con un incremento en la complejidad del organismo, si bien no es suficiente con un alto contenido en DNA para desarrollar un organismo complejo, sino que su genoma debe contar con un gran número de genes con funciones distintas (Nei, 1987). A su vez, a mayor contenido en DNA y mayor redundancia génica, crece la posibilidad de incrementar el número de genes por recombinación entre dominios básicos (Gilbert, 1978).

La duplicación génica ha desempeñado un importante papel en la creación de complejidad ya desde los primeros estadios de la evolución de los organismos. Se ha hipotetizado que la mayoría de genes eucariotas fueron producidos por duplicación y elongación de genes primordiales o repeticiones oligoméricas que existieron en un estado temprano de la evolución (Ohno, 1984). En este sentido, la alta homología entre algunos exones dentro de un mismo gen ha hecho pensar que los intrones serían remanentes de las regiones intergénicas originales que fueron duplicadas simultáneamente con las regiones codificantes (Nei, 1987).

En definitiva, la duplicación en tándem (de un gen o de un segmento génico) es importante para incrementar el número de genes: genes redundantes, que se adquieren cuando se necesita mucha cantidad de un producto génico y, fundamentalmente, genes con nuevas funciones.

3.1. LAS FAMÍLIAS GÉNICAS

Como consecuencia de la duplicación génica, gran número de genes en organismos superiores no se hallan como una sola copia en el genoma, sino agrupados en lo que se denomina familias génicas (o *clusters* génicos), cuyos miembros pueden estar dispersos o agrupados en un mismo cromosoma. Las familias multigénicas se definen como grupos de genes con homología de secuencia y funciones relacionadas, y las llamadas familias supergénicas son grupos mayores de genes que contienen uno o más dominios o módulos de origen común. Un dominio hace referencia a una unidad funcional de una proteína, y un módulo sería una unidad estructural ligeramente menor que un dominio.

Para desarrollar genes con nuevas funciones, la familia génica debe acumular diversidad en las copias duplicadas por selección natural positiva. La fijación en la población de una duplicación puede producirse puramente por deriva genética, pero la tasa de fijación se acelerará si la duplicación es ventajosa.

La familia de las β-globinas en mamíferos es quizá el *cluster* génico mejor estudiado. Se ha utilizado con frecuencia como modelo y ejemplo para entender y explicar distintos mecanismos genómicos. Su organización es diferente entre especies de mamíferos, y fenómenos de duplicación génica, deleción, conversión y recombinación parecen haberse dado frecuentemente durante su evolución. En la familia de las β-

globinas existen cinco genes funcionales producidos por duplicación en tándem, con alta homología entre ellos (ε, funcional sólo en estado embrionario, γ^G y γ^A funcionales en estado fetal, δ y β activos en estado adulto). A su vez esta familia génica se une a la de las α-globinas y la mioglobina para formar la superfamilia de genes globina (Nei, 1987). Las funciones de algunos pares de homólogos son considerablemente diferentes (como la hemoglobina y la mioglobina), mientras que otros pares mantienen esencialmente la misma función (hemoglobinas β y α).

El hecho de que los miembros de una familia génica lleven a cabo funciones relacionadas parece claro, ya que los nuevos genes provienen de un gen ancestral común. En general, dos proteínas homólogas cuyas secuencias aminoacídicas difieran en más del 50% tendrán funciones distintas, mientras que no existe diferenciación funcional entre proteínas cuya divergencia en las secuencias sea menor del 15% (Nei, 1987). Debido a la similitud de funciones, la evolución de los genes de un *cluster* puede ir ligada, de manera que cambios en la función o en la regulación de uno de ellos afecte al resto. La evolución conjunta también se manifiesta en el grado de diversidad dentro de la familia génica. Los *clusters* evolucionan bajo mecanismos que diversifican y mecanismos que homogeneizan su información genética, como mutación, entrecruzamiento desigual y conversión génica. El ritmo de diversificación u homogeneización será ajustado por la evolución para cada familia, que funcionará como una unidad evolutiva (Ohta, 1983; Nagylaki, 1982; Ohta, 1990). Incluso las familias génicas que se consideran más estables, como la del RNA ribosomal o las histonas, sufren procesos de diversificación y homogeneización. A medida que los miembros de un *cluster* diverjan entre sí, la homogeneización se hará más rara.

Dependiendo del modo en que evolucione el *cluster*, genes homólogos (miembros de una familia génica) de distintas especies (ortólogos) pueden ser más semejantes entre sí que genes homólogos dentro de la misma especie (parálogos).

4. LAS COPIAS NO FUNCIONALES: LOS PSEUDOGENES

Si se producen dos duplicados de un gen, uno de ellos puede mutar y adquirir una nueva función, o bien puede convertirse en no funcional, siempre y cuando la otra copia mantenga la función original. En el primer caso obtendremos un nuevo gen, en el segundo caso un pseudogén. Los pseudogenes constituyen una pequeña parte de la gran porción del genoma eucariota formada por DNA aparentemente no funcional. Estas secuencias han sido calificadas como *dead genes* y se definen como aquellos segmentos genómicos homólogos a un gen funcional, derivados del mismo y no alélicos, que contienen cambios nucleotídicos que impiden su correcta expresión, y que por tanto no son capaces de producir un producto proteico funcional. El primer pseudogén eucariota fue descrito por Jacq y colaboradores en 1977, en el *cluster* génico del gen 5S del oocito de *Xenopus laevis*. El gen 5S codifica para el componente del ribosoma RNA5S, y se encuentra repetido en múltiples copias ordenadas en tándem; en cada copia, el gen va seguido de un pseudogén 5S altamente conservado.

Dentro de un *cluster* génico, la creación de nuevos genes que se conviertan en duplicados funcionales, en duplicados que varíen para resultar en genes con nuevas funciones, o en pseudogenes, es una dinámica continua que modula la evolución de la familia. Cabría considerar a las familias génicas existentes como el resultado de múltiples pruebas de ensayo y error para la creación de nuevos genes. El número de genes varía ampliamente de una familia génica a la otra, pero la presencia de pseudogenes en todas ellas parece ser una constante.

Walsh (1995) desarrolló un modelo para estimar la probabilidad de formación de un nuevo gen, en comparación con la probabilidad de que la nueva copia derive a no funcional. Según este modelo, una vez se ha producido la duplicación de un gen, es más probable que la copia duplicada evolucione hacia un pseudogén (una mutación deletérea se fija) que el que se convierta en un gen con una función nueva (una mutación ventajosa se fija y es protegida por la selección de la acumulación subsecuente de mutaciones deletéreas), simplemente por que las mutaciones deletéreas ocurren con mayor frecuencia que las ventajosas.

La frecuencia de los pseudogenes es variable dependiendo del gen en concreto y del organismo de que se trate. Normalmente son una minoría respecto al número total de

genes, si bien se han descrito casos en los que los pseudogenes forman la mayor parte de la familia génica (ver revisión de Cooper, 1999). Hay que considerar que existe una tendencia a infravalorar el número de pseudogenes existentes, porque la secuencia de pseudogenes antiguos puede haber divergido del gen parental hasta tal punto que ya no sea reconocida por pruebas moleculares derivadas del homólogo funcional. La acumulación de mutaciones hace que la identidad del pseudogén desaparezca gradualmente, hasta que pasa a existir simplemente como DNA no funcional. También puede que existieran pseudogenes que han sido eliminados de las poblaciones actuales por procesos de deleción de secuencias.

Li (1997) demostró que la tasa de pérdida o de fijación de un pseudogén en una población depende principalmente del tamaño efectivo de la población (N_e) y de la tasa de mutación (μ). Esto es así porque el proceso se da principalmente por deriva genética. En una población muy grande, el efecto de la deriva es negligible, de modo que los duplicados no funcionales puede que no se fijen nunca en la población. En poblaciones finitas, la fijación de pseudogenes sí puede darse (por ejemplo cuando N_e es 10^6, y μ es 10^{-5} por generación y por secuencia, el tiempo de fijación de un pseudogén es de unas 10^7 generaciones).

Desde el momento en que un pseudogén se establece como una nueva secuencia en el genoma, empiezan a actuar dos procesos evolutivos. En primer lugar, se acumulan rápidamente mutaciones puntuales que diferenciarán gradualmente las secuencias del pseudogén y del gen correspondiente. La composición nucleotídica del pseudogén se irá asemejando progresivamente a la de sus nuevos alrededores, proceso que se denomina asimilación de la composición (Li, 1997). El segundo proceso evolutivo es que el pseudogén reduce su tamaño con relación al gen; este “acortamiento de longitud” (*length abridgment*) está causado por un exceso de deleciones sobre inserciones. Se ha estimado que un pseudogén procesado (de los que trataremos a continuación) pierde la mitad de su DNA en unos 400 millones de años (Li, 1997). Este proceso es más lento que la tasa de producción de pseudogenes, por lo que el acortamiento de longitud no es un factor determinante del tamaño total del genoma.

4.1. MECANISMOS DE PRODUCCIÓN Y TIPOS

Si bien la definición de pseudogén es clara, el término abarca un amplio abanico de regiones genómicas. Los pseudogenes pueden ser truncados (si sólo contienen un fragmento del gen fuente) o enteros, pueden contener intrones o no y, aunque la mayoría de pseudogenes son silenciosos en cuanto a transcripción y traducción, pueden transcribirse, e incluso traducirse a péptido aberrante. Su localización es frecuentemente cercana al gen funcional correspondiente, pero pueden hallarse también en cromosomas distintos.

Las características mencionadas dependen en gran medida del mecanismo que originó el pseudogén, y de hecho su clasificación se basa en estos mecanismos. Los mecanismos básicos son dos: duplicación, de la que ya hemos tratado y mencionaremos sólo brevemente aquí, y retrotransposición, de la que trataremos más adelante en este apartado. No obstante, para el origen de algunos pseudogenes ambos mecanismos se combinan, mientras que otras veces ninguno de los dos interviene, caso de los pseudogenes de copia única originados por inactivación de un gen funcional. Se han descrito también otros casos complejos. Por ejemplo, se han detectado pseudogenes formados a partir de pseudogenes preexistentes, como los pseudogenes β^x y β^z en la familia multigénica de las β-globinas en cabra, que se formaron a partir de la duplicación de un pseudogén ancestral (Cleary *et al*., 1981). Es interesante observar que la duplicación génica actúa en todo tipo de secuencias, ya sean funcionales o no. Existen otros casos inusuales, como el del pseudogén híbrido β-δ-globina hallado en lemures, mezcla de fragmentos del pseudogén de la β-globina y el gen de la δ-globina por entrecruzamiento desigual (Jeffreys *et al*., 1982). También se han identificado pseudogenes originados por duplicación a partir de otro pseudogén más transposición: los pseudogenes relacionados con el gen de la neurofibromatosis tipo 1 (NF1) parecen haberse generado por duplicaciones tanto a partir del gen NF1 original como a partir de duplicaciones del mismo, que se duplicaron a su vez y se transpusieron a otros cromosomas (Luijten *et al*., 2000). El mismo mecanismo de duplicación y transposición de copias no funcionales parece haberse dado en los pseudogenes de la familia de receptores olfatorios (Rouquier *et al*., 1998).

Los pseudogenes originados por duplicación génica se denominan no procesados, es decir, retienen la estructura exón-intrón de los genes funcionales fuente.

Se asume que estas secuencias se originan por simple duplicación de secuencias génicas funcionales, pero se inactivan dado que su redundancia impide que la selección mantenga su potencial de ser expresadas. Se diferencian dos tipos, que podrían reflejar mecanismos distintos en la generación de pseudogenes: copias completas, normalmente ligadas al gen, y copias parciales, que se encuentran algunas veces no ligadas al gen de origen. Muchas duplicaciones incluyen a la región promotora, por lo que algunos duplicados son transcritos, en algún caso incluso con mayor eficiencia que el gen original (Cooper, 1999). No obstante, las mutaciones acumuladas impiden la traducción a péptido funcional.

La retrotransposición origina pseudogenes no ligados al gen funcional correspondiente, es decir, en posición genómica distante, y frecuentemente en cromosomas distintos. Este tipo de pseudogenes se denomina pseudogenes procesados o retropseudogenes. Para su origen se supone un proceso de transcripción reversa del mRNA procesado y posterior inserción aleatoria en el genoma del DNA complementario (cDNA) de doble cadena resultante con la ayuda de retrovirus (la retrotranscriptasa inversa es un enzima característico de retrovirus). La situación distal del gen fuente se debe a que la retrotransposición requiere de un intermediario de mRNA que es móvil. El origen a partir de una molécula de mRNA se hace evidente por la ausencia de intrones y la frecuente presencia de segmentos poli-A en los extremos 3' del pseudogén. Normalmente se encuentran *direct repeats* de 9 a 17 pares de bases (pb) flanqueando al pseudogén, adquiridos probablemente durante la integración en el genoma. Al igual que los pseudogenes producidos por duplicación, estas secuencias acumulan lesiones que evitan su expresión. Los pseudogenes procesados suelen truncarse durante la retrotransposición, y en particular la rotura del extremo 5' es frecuente. Además, el proceso de transcripción reversa es propenso a acumular diferencias entre el RNA molde y el cDNA. En todo caso, aunque no adquirieran ninguna mutación, su expresión se vería limitada por la ausencia de elementos promotores activos. A menos que el gen derive de un gen transcrito por la RNA polimerasa de tipo III (con promotor interno), la copia será insertada en una región genómica que muy probablemente no dispondrá de las secuencias reguladoras que residen en regiones no transcritas, y permanecerá inactivo. Existen algunos casos excepcionales, como el pseudogén de la glutamina sintetasa humana, que se ha ubicado a 3' de un promotor funcional que permite su transcripción, aunque la secuencia no se expresa debido a las mutaciones que ha acumulado (Cooper, 1999).

Para ser heredados, los pseudogenes procesados deben originarse en la línea germinal, y por tanto sus genes de origen deben ser expresados en dicha línea. En concordancia con esto, muchos pseudogenes procesados corresponden a genes de los que se sabe que se expresan en las células germinales, o a genes *housekeeping*. Los genes *housekeeping* son aquellos que codifican para proteínas necesarias en todos los tipos celulares para el crecimiento, el metabolismo o la replicación, y que por tanto se expresan en todos los tejidos. No obstante, algunos pseudogenes procesados derivan de genes específicos de tejido, lo que requiere que el RNA cruce barreras celulares. Estos pseudogenes pueden ser consecuencia de "tránscritos filtrados" (fenómeno de *leaky transcription*), que parecen darse para cada gen y en cada célula (Cooper, 1999). También se ha sugerido que la molécula de RNA podría ser encapsulada en el virión de un retrovirus y transportada a la línea germinal, donde sería retrotranscrita (Li, 1997).

A pesar de que los pseudogenes procesados son abundantes en mamíferos, son relativamente raros en otros vertebrados y no se han encontrado en invertebrados. En mamíferos, algunos genes tienen una sola copia de un pseudogén procesado, mientras que otros poseen centenares (como los genes snRNA -*small nuclear* RNA-). Esta proporción puede reflejar el nivel de la transcripción del gen activo en la línea germinal, aunque también es probable que esté determinada por características particulares del gen fuente, que determinarán la eficiencia de la transcripción reversa y de la integración de los tránscritos (Cooper, 1999).

Se han descrito también algunos casos de pseudogenes semi-procesados, en los que el mRNA es retrotranscrito e integrado en el genoma antes de que las secuencias intrónicas hayan sido completamente eliminadas (Cooper, 1999).

Un último mecanismo de producción de pseudogenes viene dado por la inactivación de genes funcionales que no son eliminados del genoma. A este tipo de secuencias se las ha denominado pseudogenes unitarios, ya que el pseudogén así originado será de copia única y no tendrá un duplicado funcional (Graur y Li, 2000). Dado que el organismo puede permitirse la pérdida de estas secuencias, los genes inactivados no serán funcionalmente esenciales, ya que de otro modo serían eliminados rápidamente de la población. Se han detectado algunos de estos pseudogenes que perdieron funcionalidad durante la evolución de primates, y cuyos ortólogos correspondientes pueden encontrarse activos en genomas de vertebrados inferiores (Cooper, 1999).

4.2. MUTACIONES EN PSEUDOGENES

Dado que carecen de función biológica y por tanto no están sujetos a restricciones funcionales, en los pseudogenes no existe ninguna limitación para la acumulación de todo tipo de mutaciones que pueden impedir cualquier etapa de la expresión génica: codones de paro prematuro de la transcripción (mutaciones sin sentido o *nonsense*), inserciones y deleciones que desplacen la pauta de lectura (mutaciones *frameshift*), sustituciones (mutaciones erróneas o *missense*), o reordenamientos génicos más complejos; las mutaciones pueden localizarse en regiones reguladoras o codificantes, en zonas que afecten al proceso de *splicing*, al inicio de la transcripción, etc. Una sola de estas mutaciones puede ser suficiente para impedir la expresión. Las múltiples mutaciones que normalmente contienen los pseudogenes hacen difícil identificar aquella que fue la causa primera de la pérdida de función.

El cambio en la evolución de los pseudogenes ha sido de particular interés porque pueden ser utilizados para testar y verificar el modelo neutralista. Según la teoría neutralista de la evolución (Kimura, 1983), se espera que la tasa de sustituciones nucleotídicas sea más alta para genes funcionalmente menos importantes que para genes funcionalmente más importantes, ya que estos últimos estarían sujetos a una selección purificadora más intensa. Por tanto, sería de esperar que los pseudogenes evolucionaran a tasas más altas que las secuencias funcionales, ya que las mutaciones se acumularían sin ser sujeto de selección.

Se ha observado que la proporción de deleciones en pseudogenes es de una por cada 40 sustituciones, mientras que se da una inserción cada 100 sustituciones (Ophir y Graur, 1997). La tasa de sustitución se ha estimado en $4{,}7 \times 10^{-9}$ por nucleótido y por año a partir de pseudogenes de globinas, mientras que para los homólogos funcionales correspondientes las tasas estimadas son $0{,}85 \times 10^{-9}$, $0{,}70 \times 10^{-9}$ y $2{,}34 \times 10^{-9}$, para la primera, segunda y tercera posición dentro de un codón en zona codificante (Li *et al*., 1981; Nei, 1987). A partir de diversos genes eucariotas se estimó la tasa de sustitución en $4{,}6 \times 10^{-9}$ y $0{,}88 \times 10^{-9}$ por nucleótido y año, para posiciones sinónimas y no sinónimas respectivamente, dentro de zona codificante, y como $4{,}85 \times 10^{-9}$ por nucleótido y por año en pseudogenes (Li *et al*., 1985; Nei, 1987). Esta tasa de mutación superior en pseudogenes, aunque sólo ligeramente mayor que la tasa en posiciones sinónimas de zonas codificantes, se ha puesto de relevancia en otros estudios (Li, 1997). La tasa de mutación más alta observada en pseudogenes respecto a regiones codificantes, y más

alta en zonas sinónimas que no sinónimas dentro de la región codificante, es una predicción de la teoría neutralista y por tanto este dato se ha utilizado para apoyarla. En principio, todas las mutaciones que se dan en pseudogenes son selectivamente neutras y tienen la misma probabilidad de ser fijadas en la población (Li *et al*., 1984).

En relación al punto anterior, se ha sugerido que los pseudogenes mostrarían un patrón de mutación equivalente al patrón de mutación espontánea en el DNA, ya que estas secuencias darían una imagen menos sesgada de la ocurrencia de fenómenos mutacionales que las secuencias funcionales bajo selección. Por ello, los pseudogenes se han utilizado también como modelo para estimar las tasas de mutación relativas entre los cuatro nucleótidos (Li *et al*., 1981). Estudios de variación en pseudogenes demostraron que la dirección de las mutaciones no es al azar entre los cuatro nucleótidos. Se sabe que las transiciones y en particular C a T y G a A se dan más frecuentemente que las transversiones (Li *et al*., 1984; Li *et al*., 1987; Nei, 1987). A pesar de que bajo mutaciones al azar la proporción esperada de transiciones es del 33%, en la práctica, en DNA nuclear, se observa una proporción cercana al 60% (incluso si se excluyen los dinucleótidos hipermutables CpG) (Gojobori *et al*., 1982; Li *et al*., 1984; Cooper y Krawczak, 1990). Algunos nucleótidos son también más mutables que otros; las frecuencias relativas con que mutan los cuatro nucleótidos son: adeninas 19%, timidinas 16%, citidinas 32%, y guaninas 33% (Gojobori *et al*., 1982; Li *et al*., 1984; Li, 1997). Un 64,5% de todas las mutaciones resultan en adeninas o timidinas (el valor esperado si las sustituciones fueran al azar sería del 50%), nucleótidos que a su vez son poco mutables. Por tanto, esperaríamos que los pseudogenes, así como otras regiones no sujetas a restricciones funcionales, se enriquecieran en A y T con el paso del tiempo (Li, 1997), aunque esto depende en gran manera de la composición de la región cromosómica en que se hallen (Saccone *et al*., 1999; Bernardi *et al*., 2000).

Estudios sobre pseudogenes han apuntado a que la alta tasa de sustitución nucleotídica observada en estas regiones, en comparación con las zonas codificantes, depende de manera importante de mutaciones en los dinucleótidos CpG. Sólo un 3% de las citosinas en el DNA humano están metiladas, pero de estas, un 90% se encuentra en dinucleótidos CpG. Los residuos de citosina que se encuentran en dinucleótidos CpG son dianas para la metilación en el átomo de carbono 5. El residuo de 5-metil-citosina puede deaminarse espontáneamente, dando lugar a un residuo timidina. Como la timidina se encuentra ahora mal apareada con una guanina, el desapareamiento es reconocido por una DNA glicosilasa específica, que reemplaza la T por una C. No obstante, este proceso de reparación es altamente ineficiente y, como resultado, los dinucleótidos CpG son

lentamente reemplazados por TpG (y CpA en la cadena complementaria). Después de que un gen se convierte en pseudogén, los cambios en los dinucleótidos CpG ya no están sujetos a ninguna restricción funcional, y pueden por tanto contribuir sustancialmente a las transiciones C a T y G a A si la frecuencia de CpG (metilados) era relativamente alta después de que se produjera la silenciación del gen (Li *et al*., 1984; Strachan y Read, 1996).

Como el ritmo de sustituciones es normalmente mayor en un pseudogén que en su homólogo funcional, las diferencias en la secuencia de nucleótidos entre ambos *loci* se atribuyen principalmente a mutaciones que han ocurrido en el pseudogén, a menos que el tiempo de divergencia entre ambos sea muy largo (Nei, 1987). A partir de esta divergencia se ha intentado estimar la edad de los pseudogenes. La "edad" de un pseudogén hace referencia al tiempo que ha pasado desde la duplicación y/o la inactivación, ya que los pseudogenes pueden inactivarse desde el momento de la duplicación o ser activos durante un tiempo. Estas estimas se basan en el grado de homología del par gen/pseudogén y en la proporción relativa de cambios sinónimos y no sinónimos entre el gen y el pseudogén. En este último caso se asume que después de la inactivación el pseudogén acumulará cambios a la misma velocidad que los cambios silenciosos se acumulan en el gen activo. No obstante, sí existe selección contra los cambios silenciosos en los genes activos. Además, si se han producido fenómenos de conversión génica del gen al pseudogén, el reloj molecular de las zonas convertidas será reinicializado, lo que llevará a una subestima de la edad del pseudogén. Se ha observado que la edad de un pseudogén no está siempre linealmente relacionada con el número de mutaciones presentes, y se ha sugerido que esto puede ser en parte debido a la conversión génica (Cooper, 1999). Otra aproximación a la edad de un pseudogén se realiza por la presencia o ausencia de secuencias ortólogas en especies de filogenia conocida, lo que puede ser un indicador de la antigüedad de la duplicación génica.

Una vez una copia génica se inactiva y pasa a ser un pseudogén, la probabilidad de que se vuelva a activar es muy baja. En primer lugar, porque si bien existen muchas formas diferentes de que un gen se inactive, una mutación que active a un pseudogén debe producirse exactamente en el mismo lugar en que se produjo la primera y revertir exactamente al nucleótido o nucleótidos iniciales. La probabilidad de que esto ocurra es prácticamente nula. Además, una vez un gen se hace no funcional, muchas mutaciones destructivas ocurrirán en este *locus* (por la falta de presión selectiva purificadora), con lo

que la probabilidad de reactivación decrece con el tiempo. Existe una excepción a esto, y es cuando la zona de un pseudogén con las mutaciones responsables de su inactividad es reemplazada por la secuencia nucleotídica funcional a través de recombinación desigual o conversión génica (Nei, 1987). Esto ocurriría durante el tiempo en que el grado de divergencia entre gen y pseudogén no fuera demasiado grande, y llevaría a un duplicado redundante, disminuyendo por tanto la probabilidad de crear un gen de nueva función. Esta reactivación por conversión génica es especialmente difícil para los pseudogenes procesados, ya que normalmente se sitúan a gran distancia cromosómica del gen fuente (Li *et al*., 1997).

Marshall y colaboradores (1994) estimaron en 6 millones de años el tiempo durante el que era probabilísticamente aceptable que un pseudogén volviera a la actividad. Después de 10 millones de años, la acumulación de múltiples mutaciones inactivadoras lo haría improbable, a menos que algún tipo de presión selectiva preservase intacto el pseudogén.

4.3. SIGNIFICADO EVOLUTIVO

Por lo que se ha expuesto anteriormente, los pseudogenes serían intentos fallidos de la evolución, intentos de producir nuevos genes que no tuvieron éxito. No obstante, también se ha propuesto que estas secuencias podrían desempeñar algunas funciones y tener cierta importancia en la evolución del genoma.

Ya en 1933, Haldane sugirió que la mutación terminaría por inactivar las duplicaciones génicas, generando pseudogenes. También él remarcó que incluso si una duplicación no lleva a un nuevo gen, aún puede tener una importante función evolutiva, ya que hace variar el orden génico y varía las relaciones de ligamiento entre genes activos (Walsh, 1995). Se sabe que los pseudogenes pueden influir también en la evolución de otras secuencias funcionales al actuar como secuencias donadoras en la conversión génica o participando en fenómenos de recombinación (Cooper, 1999).

Un ejemplo es la familia génica de receptores olfatorios (RO), en la que de 90 secuencias RO distribuidas por todo el genoma un 70% son pseudogenes. Se ha sugerido que la alta proporción de pseudogenes puede ser debida a su importancia en el mantenimiento y la generación de diversidad en los genes RO activos, a través de

mecanismos de recombinación y, principalmente, conversión génica (Rouquier *et al*., 1998; Glusman *et al*., 2000).

Parece que los pseudogenes de la familia de RO presentan una especial plasticidad funcional. Se han descrito dos pseudogenes RO que han desarrollado nuevas funciones no codificantes. Uno de ellos se ha convertido en una zona reguladora, *enhancer* (Buettner *et al*., 1998), y el otro ha evolucionado hacia una zona estructural de unión a la matriz nuclear (*matrix-attachment region,* MAR) (Gimelbrant y Mclintock, 1997). Se cree que la organización de la cromatina respecto a estas zonas MAR afecta a los dominios cromosómicos y contribuye a posicionar a los genes, en este caso a otros genes RO, para facilitar su transcripción (Gimelbrant y Mclintock, 1997; Cooper, 1999). Otro de los pseudogenes RO parece haber evolucionado a isla CpG, ya que presenta un 65.4% en contenido C+G y un 9.3% de su secuencia corresponde a dinucleótidos CpG (Glusman *et al*., 2000).

Otras funciones que se han propuesto para los pseudogenes son la restricción de la expresión génica, al ocupar la maquinaria de transcripción en tránscritos que no serán útiles, y la regulación de la expresión del gen funcional, en caso de que el pseudogén fuera transcrito en sentido opuesto al gen de manera que se formaran heterodúplex no funcionales de RNA gen-pseudogén (Vanin *et al*., 1980).

También se ha sugerido que los pseudogenes pueden tener un papel regulador, actuando como espaciadores intergénicos. En la familia de las globinas de ratón, conejo, y humanos, los genes de β-globina están separados de otros genes globina por pseudogenes, que podrían desempeñar funciones reguladoras en *cis* durante el desarrollo, marcando el orden de expresión de los genes (Cleary *et al*., 1981). Otro ejemplo de pseudogenes como espaciadores sería el del DNA 5S de *Xenopus laevis* que, como ya se ha mencionado, se constituye por repeticiones en tándem de unidades de gen más pseudogén 5S. La presencia de un pseudogén en cada una de las copias y su alta conservación se intentaron explicar sugiriendo que el pseudogén actuaría como un espaciador, que sería transcrito y formaría parte del tránscrito primario del RNA 5S (Jacq *et al*., 1977).

¿Porqué se mantienen los pseudogenes en el genoma si son “finales” evolutivos? Se ha intentado encontrar alguna función biológica que explique la presencia de más secuencias en el genoma de las estrictamente necesarias para producir proteínas. Como hemos visto, quizá sí que alguna vez los pseudogenes evolucionen hacia nuevas funciones (no deberíamos llamarlos entonces ya pseudogenes), y parece que una función importante, como secuencias propiamente no funcionales, es la de actuar como reserva

de diversidad genética, que sería transferida al gen. No obstante, en la mayoría de casos puede que sean simplemente vestigios evolutivos, resultado de la duplicación génica y la posterior divergencia de las secuencias. Algunos de estos fragmentos de DNA no funcional pueden ser eliminados, pero muchos de ellos parecen permanecer en el genoma simplemente porque no interfieren con la fisiología celular, lo que daría una respuesta simple a la pregunta anterior (Nei, 1987).

5. EL GEN GBA Y EL PSEUDOGÉN GBA

El presente estudio se ha centrado en el pseudogén de la glucocerebrosidasa (psGBA), el duplicado no funcional del gen que codifica para la proteína glucocerebrosidasa (gen GBA). Ambos genes se sitúan en la región cromosómica 1q21 humana (Figura 3), separados por 16 kb, y comparten un 96% de homología (Horowitz *et al*., 1989; Zimran *et al*., 1990; Winfield *et al*., 1997).

Un clon de cDNA del gen GBA se secuenció por primera vez a partir de una librería de cDNA de fibroblasto (Sorge *et al*., 1985). El pseudogén fue clonado por Reiner y colaboradores en 1988, al aislar dos clones de cDNA GBA específicos, uno de los cuales se comprobó que no se trataba de un gen homólogo a GBA como se sugirió en un principio, sino de un pseudogén. Horowitz y colaboradores obtuvieron en 1989 la secuencia completa del fragmento de DNA genómico correspondiente a GBA y psGBA.

El gen GBA tiene una longitud de 7,6 kb, que se reparten en 11 exones y 10 intrones. psGBA presenta el mismo número de exones e intrones, pero su longitud es de unas 5,7 kb. La diferencia en longitud se debe a una deleción de 55 pb en el exón 9 de psGBA (flanqueada en el gen por un *short inverted repeat* que puede haber desempeñado un papel en su deleción), y a la presencia de distintas combinaciones de secuencias Alu en GBA: 313 pb en el intrón 2, correspondientes a una Alu invertida; 626 pb en el intrón 4, que corresponden a dos secuencias Alu invertidas; 320 pb en el intrón 6, correspondientes a una secuencia Alu entera más el fragmento 5' de una Alu invertida; y 277 pb en el intrón 7, que abarcan dos repeticiones en tándem de una secuencia Alu, más el fragmento 3' truncado de otra Alu. En psGBA hay sólo una secuencia Alu, en el intrón 7 (Horowitz *et al*., 1989). El mecanismo de inserción de las secuencias Alu resulta en la duplicación de un fragmento de pocos pares de bases en el que la secuencia Alu se inserta.

Figura 3: Cromosoma 1 humano.
La flecha indica la localización de GBA y psGBA.

Como consecuencia, se crean repeticiones nucleotídicas que se orientan en la misma dirección a lo largo del DNA y que flanquearán a la secuencia Alu en su nueva localización. Estos *direct repeats* que se encuentran en GBA están presentes como copia única en psGBA. Dado que las secuencias Alu son elementos móviles, su presencia en GBA pero no en psGBA hace suponer que el gen ancestral no las contenía (excepto quizá la secuencia Alu que se encuentra en el intrón 7 de psGBA y GBA) y que, en este aspecto, la secuencia de psGBA es más similar a la del gen original que la secuencia de GBA. Después de la duplicación, la copia que hoy es el gen activo habría sido modificada por la inserción de secuencias Alu, mientras que la otra copia acumuló mutaciones que la hicieron no funcional.

Los promotores de genes eucariotas se separan en dos grandes grupos. Por una parte se consideran los que contienen cajas TATAA y CAAT. El motivo TATAA supone un lugar de reconocimiento para el factor de transcripción TFIID que dirige el inicio de transcripción unos 30 pares de bases a 3'. Un segundo tipo de promotores, los promotores sin TATAA, frecuentemente contienen varios motivos GGCGGG (puntos potenciales para la unión del factor de transcripción SP1), y generalmente dirigen el inicio de transcripción desde varios lugares. Las proteínas *housekeeping* suelen tener lugares de reconocimiento para SP1 en sus promotores, mientras que genes altamente regulados tienen cajas TATAA y CAAT (Dynan, 1986). El gen GBA presenta a la vez características de gen *houskeeping* y de gen específico de tejido. Por una parte, la transcripción se inicia en múltiples lugares (Doll *et al*., 1995; Beutler y Grabowski, 1995), y su expresión parece ser generalizada, si bien con acusadas diferencias en los niveles de mRNA y de actividad enzimática en distintos tejidos (Reiner y Horowitz, 1988; Doll *et al*., 1995). Por otra parte, el motivo de unión a SP1 no se encuentra en el promotor de GBA. El gen GBA contiene dos motivos TATAA y dos CAAT en la zona promotora, que se encuentran conservados (excepto por una sustitución A a G en la segunda caja CAAT) en el pseudogén (Horowitz *et al*., 1989). En conjunto, estos datos apuntan a que el enzima glucocerebrosidasa no es un enzima *housekeping* típico.

Doll y colaboradores (1995) demostraron que elementos reguladores a 5' del TATAA *box* son dispensables para la expresión en algunos tipos celulares (podrían controlar la expresión diferencial de GBA en distintos tejidos), mientras que elementos en el exón 1 son esenciales para la expresión. Los elementos del exón 1 actúan como *enhancers* transcripcionales y pertenecen a la región 3' no traducida (3'UTR, *untranslated region*), por lo que no afectan a la estabilidad de la proteína, sino a los niveles de

transcripción del mRNA. Moran y colaboradores (1997) identificaron cuatro elementos reguladores en el promotor GBA y sugirieron que la disponibilidad de los factores de transcripción que se unen a estos motivos controlaría los niveles de transcripción de GBA. En psGBA se mantienen estos motivos reguladores, a excepción de dos sustituciones nucleotídicas en uno de ellos.

El gen GBA se transcribe desde dos codones ATG activos. El situado más hacia 5' (en el final del exón 1) inicia una proteína que contiene un péptido señal de 39 residuos, con la mitad amino terminal hidrofílica, mientras que el situado a 3' (en medio del exón 2) inicia un péptido señal hidrofóbico de 19 residuos (Sorge *et al*., 1987). El codón ATG iniciador más *upstream* se traduce preferencialmente, y produce de dos a tres veces más proteína que el ATG situado más hacia 3'. Se ha sugerido que las dos formas podrían generarse por mecanismos de *splicing* diferencial, aunque no se han encontrado los mRNA correspondientes (Beutler y Grabowski, 1995). Sí se han detectado dos mRNA de GBA de distinto tamaño (de 2,2 y 2,6 kb), pero no debidos a los dos ATG de inicio, sino a dos señales alternativas de poliadenilación, separadas por unos 470 pb (Reiner *et al*., 1988). El enzima glucocerebrosidasa necesita el péptido señal para el transporte a través de la membrana del retículo endoplasmático rugoso, desde donde pasará al aparato de Golgi y finalmente al lisosoma. Durante el procesamiento post-traduccional se produce una escisión proteolítica, y el péptido señal se libera. La proteína madura tiene 497 aminoácidos y un peso molecular de 55,6 kDa (aunque varía dependiendo de la presencia o no del péptido señal y del estado de glicosilación de la proteína).

Los dos codones de inicio de la traducción y las dos señales de poliadenilación se encuentran conservados en psGBA.

psGBA presenta un rasgo poco frecuente entre los pseudogenes como es la transcripción, aunque la actividad del promotor de psGBA no alcanza los niveles del promotor de GBA (Reiner *et al*., 1988). Los lugares donadores y aceptores de *splicing* todavía se encuentran en el pseudogén, a excepción de los lugares donadores (5') de los intrones 2 y 4, en los que el GpT que debe iniciar cada intrón ha variado a ApT. Los lugares ApG aceptores de *splicing* con los que debe finalizar cada intrón, se encuentran todos conservados. Se han descrito dos tránscritos procesados para psGBA, que apuntan a un mecanismo de *splicing* alternativo en el pseudogén. Uno de los tránscritos no contiene secuencias intrónicas, ni tampoco el exón 2, la mayor parte del exón 3, y el exón 4, de manera que los exones a 5' de los lugares donadores de *splicing* erróneos han sido delecionados (Sorge *et al*., 1990). El segundo tránscrito no contiene los exones 2 y 3 ni la mayor parte del exón 4 y presenta 75 pb del intrón 4 (Imai *et al*., 1993).

A través del estudio de polimorfismos que no afectan al producto génico de GBA, se han descrito dos haplotipos GBA mayoritarios, definidos por 12 polimorfismos en desequilibrio de ligamiento (Beutler *et al*., 1992). Estos dos haplotipos se denominan Pv1.1+ y Pv1.1- (o simplemente + y -) según la ausencia o la presencia de una diana PvuII en el intrón 6, respectivamente. En población caucasoide el haplotipo - se encuentra con una frecuencia de aproximadamente el 70% y el haplotipo + representa el 30% restante, mientras que en población asiática y africana estas frecuencias se invierten (Beutler *et al*., 1992; Glenn *et al*., 1994; E. Mateu, comunicación personal). Se han encontrado otros dos haplotipos, si bien en muy baja frecuencia: uno de ellos es un cambio en uno de los polimorfismos en un fondo de haplotipo +, y el otro, encontrado sólo en individuos africanos, es un cambio también de un polimorfismo sobre un fondo - (Beutler *et al*., 1992).

5.1. LA ENFERMEDAD DE GAUCHER

5.1.1. DESCRIPCIÓN

La enfermedad de Gaucher es el trastorno de acumulación lisosómica de lípidos más prevalente en humanos, con unos 80.000 afectados en todo el mundo. Se transmite como rasgo autosómico recesivo y forma parte de una serie de trastornos metabólicos debidos a la deficiencia de alguno de los enzimas lisosómicos implicados en el catabolismo celular. Como resultado, estas deficiencias producen la acumulación de sustrato no metabolizado en el interior de los lisosomas.

En el caso de la enfermedad de Gaucher, el trastorno resulta principalmente de mutaciones en el gen GBA, que anulan la síntesis de proteína o bien llevan a la síntesis de un producto proteico, la glucocerebrosidasa (EC 3.2.1.45), con nula o baja estabilidad y actividad catalítica. La proteína glucocerebrosidasa (también denominada β-glucosidasa ácida, glucosilceramidasa o D-glucosil-N-acilesfingosina glucohidrolasa) es una hidrolasa lisosómica que hidroliza el glucoesfingolípido glucocerebrósido (un metabolito intermediario en la síntesis y degradación de glicoesfingolípidos complejos, y un componente de las membranas celulares) a glucosa y ceramida. La deficiencia en glucocerebrosidasa lleva a la acumulación de glucocerebrósido en el interior de los lisosomas de macrófagos, desencadenando una patología multisistémica. En los

macrófagos se produce la etapa más importante en el catabolismo de los lípidos, y dado que estas células son muy abundantes en hígado, bazo y huesos, éstos son los órganos más afectados. Los casos más graves presentan también una degeneración del sistema nervioso central.

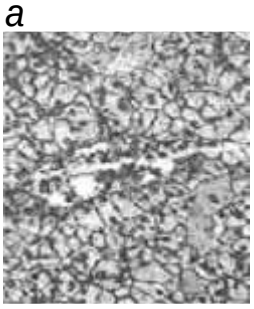
El almacenamiento intralisosómico de glucocerebrósido produce que los macrófagos aumenten de tamaño. A estas células se las llama entonces células Gaucher (Figura 4). La acumulación de estas células provoca la hipertrofia del hígado y el bazo. La causa de las anormalidades esqueléticas no se ha determinado con exactitud, aunque se ha sugerido que también podrían ser debidas a la presencia de células Gaucher en la médula ósea. En el cerebro, el rasgo más característico, a parte de la presencia de células Gaucher, es la pérdida neuronal. Secundariamente, se produce un proceso de fibrosis (gliosis en el cerebro) por la formación de fibras reticulares alrededor de las células Gaucher.

Se han descrito algunos casos, muy poco frecuentes, en los que la enfermedad de Gaucher se debe a una deficiencia en saposina C, un péptido activador necesario para la correcta función de la glucocerebrosidasa. También una mala función de las proteínas responsables del transporte de glucocerebrosidasa a los lisosomas (LAMP, *lysosome-associated membrane proteins*) puede ser responsable de la enfermedad (Beutler y Grabowski, 1995; Zimmer *et al*., 1999).

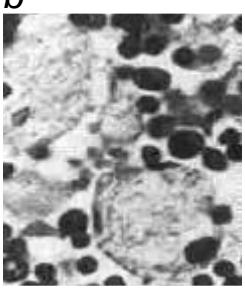
El fenotipo clínico de la enfermedad de Gaucher es muy heterogéneo. El espectro de síntomas es amplio, abarcando desde enfermos que mueren a las pocas semanas de vida hasta individuos que llegan asintomáticos a edades muy avanzadas. Dependiendo de la ausencia o presencia y severidad de las manifestaciones neurológicas, y de la edad de aparición de los primeros síntomas, se han establecido tres tipos clínicos principales de la enfermedad. El tipo 1 (crónico o no neuronopático, OMIM 230800) presenta un amplio rango de manifestaciones viscerales, la edad de inicio varía entre la infancia y la edad adulta, y es el único tipo que no presenta degeneración del sistema nervioso. El tipo 2 (neuronopático agudo, OMIM 230900) es el más grave: se inicia a una edad muy temprana, presenta una seria degeneración del sistema nervioso que puede llevar a retraso mental y produce la muerte en la infancia. El tipo 3 (neuronopático crónico o subagudo, OMIM 231000) también presenta degeneración del sistema nervioso, pero su curso es menos grave que el del tipo 2, y se manifiesta más tarde.

Figura 4. Células Gaucher en hígado (*a*), bazo (*b*), y médula ósea (*c* y *d*).

a

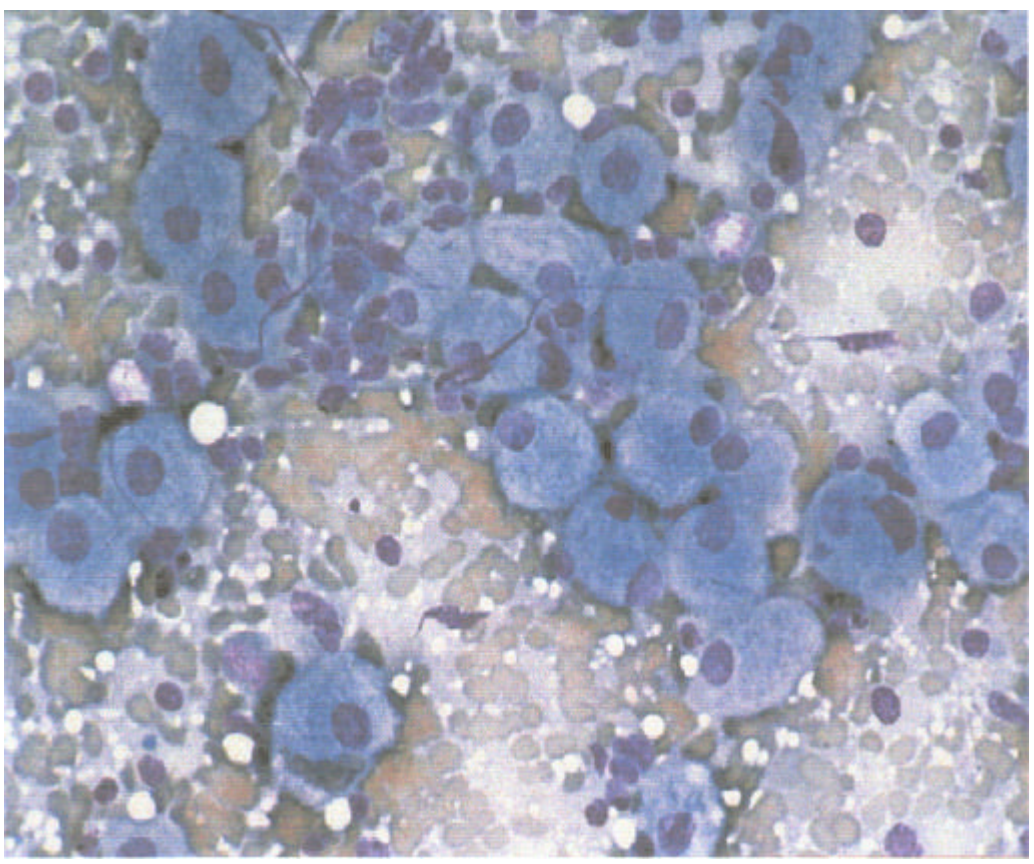


b



c

d



Los tres tipos de enfermedad se han descrito en casi todas las poblaciones. Las prevalencias oscilan entre 1/60.000 a 1/360.000 para el tipo 1, 1/500.000 para el tipo 2, y 1/100.000 para el tipo 3 (Grabowski, 1997). Estas frecuencias se han estimado a partir de pacientes sintomáticos (tipo 1), y a partir de nacimientos vivos (tipos 2 y 3) con lo que la frecuencia real de la enfermedad probablemente sea mayor. El tipo 1 es el más frecuente y se manifiesta más comúnmente en la población de judíos ashkenazitas. En esta población, la enfermedad de Gaucher representa el trastorno genético más común, con una incidencia de 1/855 de nacimientos con riesgo de enfermedad (Beutler y Gelbart, 1998), y una frecuencia de heterozigotos portadores de alelos Gaucher en la población del 5% (Motulsky, 1995). El tipo 3 es especialmente común en la población de Norrbotten, un grupo genéticamente aislado del norte de Suecia (Beutler y Grabowski, 1995), en el que la causa de la alta frecuencia de la enfermedad se ha atribuido a un efecto fundador seguido de deriva genética (Dahl *et al*., 1993). En el caso de la población judía ashkenazita se han propuesto dos hipótesis para explicar la alta frecuencia de la enfermedad: (a) una ventaja selectiva de los heterozigotos en algún momento de la historia de la población (Motulsky, 1995; Boas, 2000); (b) un efecto fundador con posterior deriva genética, que estaría asociada a la expansión que esta población experimentó en Europa desde el siglo XVI al XIX (Peleg *et al.* 1998).

Se ha intentado desarrollar una terapia para la enfermedad de Gaucher desde distintos frentes. Por una parte, la terapia sintomática supone una medida paliativa de los síntomas que mejore la calidad de vida del enfermo. La esplenectomía total o parcial y la sustitución de articulaciones son ejemplos de estas medidas, que no corrigen no obstante la causa primaria de la enfermedad. Otra alternativa es el transplante de médula ósea. Dado que los macrófagos derivan de células madre hematopoyéticas, el transplante ha revertido totalmente la enfermedad en algunos casos. Aun así, los riesgos de transplante son considerables incluso en los casos en que el donante es compatible. También en los casos en que el transplante se ha producido de manera favorable se han descrito efectos secundarios a largo plazo en el crecimiento. La restitución enzimática es el tercer campo en que se desarrolla una terapia. La administración reiterada de enzima glucocerebrosidasa modificada puede detener el curso de la enfermedad y revertir la acumulación de glucocerebrósido y con ello las manifestaciones clínicas. Las desventajas son la dependencia de las inyecciones enzimáticas de por vida, y el elevado coste del tratamiento, que hace que no sea una posibilidad real para la mayoría de pacientes. El último frente en el que se trabaja para obtener una terapia es la transferencia génica. La

estrategia sería transformar vía retrovírica células madre hematopoyéticas del paciente *in vitro*, y reintroducirlas posteriormente en la médula ósea del mismo paciente. Esto iría unido a la eliminación de las células propias, que no producen glucocerebrosidasa, por radiación o quimioterapia (Beutler y Grabowski, 1995).

5.1.2. MUTACIONES EN GBA. psGBA COMO CAUSA DE ENFERMEDAD

Hasta la fecha se han descrito 147 mutaciones causantes de la enfermedad de Gaucher. La gran mayoría son sustituciones que provocan cambio de aminoácido (115 entre *missense* y *nonsense*), aunque también se han descrito cinco sustituciones que afectan al *splicing*, seis pequeñas deleciones, cinco pequeñas inserciones, una inserción-deleción, dos grandes deleciones y trece alelos complejos (HGMD 119262, diciembre 2000; *GeneDis Database*, diciembre 2000; y para los alelos complejos que no se citan en estas bases de datos: Eyal *et al*., 1990; Latham *et al*., 1990; Hatton *et al*., 1997; Sinclair *et al*., 1998; Sarria *et al*., 1999; Filocamo *et al*., 2000).

Las relaciones fenotipo-genotipo en la enfermedad de Gaucher no se hallan completamente establecidas. Existe cierta relación entre el fenotipo clínico y determinadas mutaciones (ciertos alelos tienden a determinar enfermedad leve, moderada o severa), pero la variación fenotípica es muy amplia y no se relaciona un determinado fenotipo a un único genotipo. Probablemente, los síntomas de la enfermedad se deban a la influencia simultánea de diversos factores, entre los que el más determinante sea la presencia de mutaciones en el gen GBA, pero también otros, como la influencia de psGBA, que se sitúa muy cercano al gen y es importante en la formación de alelos complejos no funcionales con GBA, factores ambientales como infecciones, variabilidad genética ligada en los promotores de los diferentes alelos, o variabilidad genética no ligada, que puede incluir a muchos otros genes, como el de la saposina C o los genes codificantes de las LAMP (Beutler y Grabowski, 1995; Winfield *et al*., 1997; Zimmer *et al*., 1999; Koprivica *et al*., 2000).

Los alelos complejos resultan de reordenaciones entre fragmentos de GBA y de psGBA, ya sea por mecanismos de entrecruzamiento desigual o por fenómenos de conversión génica en los que psGBA actúa como donador de información. En algún

trabajo se ha utilizado el término alelo complejo para hacer referencia a alelos GBA con más de una mutación puntual no provenientes de psGBA (Eyal *et al*., 1991; Grace *et al*., 1999). En este trabajo nos referiremos a alelos complejos sólo para denominar reordenaciones genómicas GBA-psGBA.

En la Tabla 1 se resumen los alelos complejos que se han descrito hasta el momento. Los alelos complejos más frecuentes son RecNciI y RecTL. RecNciI contiene tres mutaciones puntuales en el exón 10 del gen GBA: L444P, A456P y V460V. RecTL contiene las tres mutaciones de RecNciI más D409H, en el exón 9 de GBA (Eyal *et al*., 1990; Latham *et al*., 1990; Zimran *et al*., 1990; Hong *et al*., 1990). Los dos alelos se han descrito como el posible resultado de fenómenos de conversión génica, o de un doble entrecruzamiento desigual. Hong y colaboradores (1990) sugirieron que la conversión génica era la causa más probable, aunque un fenómeno de doble entrecruzamiento se diagnosticó en un paciente con el alelo RecNciI (Zimran *et al*., 1990).

La presencia de alelos complejos agrava la severidad de la enfermedad; como excepción, también se ha descrito un paciente con el alelo RecNciI con Gaucher tipo 1 (Latham *et al*., 1990, Hong *et al*., 1990; Beutler y Gelbart, 1998).

Algunas de las mutaciones causantes de enfermedad detectadas en GBA corresponden a la secuencia de psGBA (Tabla 2). Se ha sugerido que estas mutaciones pueden producirse por dos mecanismos: mutación puntual y fenómenos de recombinación o conversión génica en los que psGBA transfiera información al gen. Esto es así por que la mayoría de estas mutaciones se han encontrado en dos estados: aisladas en un alelo GBA y formando parte de un alelo complejo. Ni la mutación que afecta al *splicing* en el intrón 2, ni las mutaciones R48W, N188S y G325R se han relacionado por el momento con ningún alelo complejo. No obstante, este alelo complejo podría haber estado presente y no haber sido detectado debido al proceso de *screening* de mutaciones concretas en GBA o, dada la alta homología que comparten psGBA y GBA, por la presencia de variantes alélicas de psGBA no conocidas, que se han confundido con variantes de GBA.

Tabla 1: Resumen de los alelos complejos psGBA-GBA. La nomenclatura de los alelos corresponde a la del artículo en que se describieron por primera vez.
[a] Bajo el término "mutaciones" se resumen las mutaciones puntuales y los fragmentos de DNA correspondientes a la secuencia de psGBA encontrados en alelos GBA.

Alelo	Mutaciones [a]	Autores
RecNciI	L444P, A456P, V460V	Eyal *et al*., 1990; Latham *et al*., 1990; Zimran *et al*., 1990; Hong *et al*., 1990
RecTL	D409H, L444P, A456P, V460V	Eyal *et al*., 1990; Latham *et al*., 1990;
Alelo complejo	R120W, N188K, V191G, S196P, G202R, F213I	Latham *et al*., 1991
RecA456P	L444P, A456P	Hatton *et al*., 1997
c1263del+RecTL	c1263del, D409H, L444P, A456P, V460V	Hatton *et al*., 1997
Alelo de fusión	secuencia psGBA desde el intrón 3 al exón 11	Reissner *et al*., 1998
Alelo complejo	secuencia psGBA en el intrón 9 y exón 10	Sinclair *et al*., 1998
Rec(1263del55; 1342G>C)	1263del55, D409H	Sarria *et al*., 1999
L444P+V460V	L444P, V460V	Hodanova *et al*., 1999
Rec(g4889-6506)	secuencia psGBA desde el exón 8 al exón 11	Hodanova *et al*., 1999
Alelo recombinante	secuencia psGBA en la zona 3' no traducida	Stone *et al*., 2000
RecFS	L444P, V460V	*GeneDis Database* (comunicación personal de A. Rolfs)
Alelo recombinante	Secuencia psGBA desde el exón 6 al exón 11	Filocamo *et al*., 2000

Tabla 2: Mutaciones causantes de la enfermedad de Gaucher descritas en GBA cuyo estado mutado corresponde a la secuencia de psGBA (según Beutler y Gelbart, 1998). La secuencia de psGBA corresponde a la descrita por Horowitz y colaboradores (1989) (Genbank J03060). Las mutaciones que no se han relacionado con ningún alelo complejo se marcan con un asterisco.
[a] Se indica la numeración correspondiente al cDNA.

GBA® psGBA	Posición	Sustitución de aminoácido		psGBA
G→A	intrón 2	*splicing*		A
C→T	exón 3	*48Arg → Trp	(R48W)	T
C→T	exón 5	120Arg → Trp	(R120W)	T
A→G	exón 6	*188Asn → Ser	(N188S)	G
T→G		188Asn → Lys	(N188K)	G
T→G		191Val → Gly	(V191G)	G
T→ C		196Ser → Pro	(S196P)	C
G→A		202Gly → Arg	(G202R)	A
T→A		213Phe → Ile	(F213I)	A
G→A	exón 8	*325Gly → Arg	(G325R)	A
1263-1317[a] del	exón 9	*frameshift*		-
G→C		409Asp → His	(D409H)	C
T→ C	exón 10	444Leu → Pro	(L444P)	C
G→ C		456Ala → Pro	(A456P)	C
G→ C		460Val → Val	(V460V)	C

6. CONTEXTO GENÓMICO: LA REGIÓN CROMOSÓMICA 1q21

6.1. GENES CIRCUNDANTES

GBA y psGBA se encuentran separados por sólo 16 kb en la región cromosómica 1q21 humana, una zona muy rica en genes funcionales (Zimran *et al*., 1990; Ligtenberg *et al*., 1990; Vos *et al*., 1992; Winfield *et al*., 1997). A continuación detallamos brevemente los genes flanqueantes a GBA descritos hasta el momento, distribuidos a lo largo de 75 kb en 1q21 (de 5' a 3'). Por su posición cercana a GBA se ha sugerido que la variabilidad en estos genes podría contribuir a la heterogeneidad fenotípica de la enfermedad de Gaucher (Winfield *et al*., 1997). No obstante, hasta ahora la participación de estos genes en la patogénesis de la enfermedad se desconoce, y de hecho se ignoran los detalles sobre sus funciones. Todos ellos, excepto MTX1, se transcriben en la misma dirección.

El gen CLK2 se localizó entre 21,5 y 32 kb *upstream* de GBA. Contiene 12 exones y codifica para una proteína kinasa. Propin1 se halló entre 15 y 21 kb *upstream* de GBA, se extiende 6,5 kb y contiene 9 exones; es homólogo a una proteína transportadora de membrana en ratón (SCAMP37), pero dada la relativamente baja homología (65%) y que SCAMP37 pertenece a una familia de proteínas, se supone que el gen humano puede ser ortólogo a otro miembro del *cluster*. Cote1 se sitúa entre 6 y 14 kb a 5' de GBA y se cree que contiene 12 exones, si bien el cDNA entero no ha sido aislado; no mantiene homología con ningún gen descrito en Genbank hasta la fecha y su función se desconoce (Winfield *et al*., 1997). Siguiendo de 5' a 3', después de Cote1 se sitúa el gen GBA, seguido por psMTX1, el pseudogén del gen MTX1. Unas 10 kb *downstream* empieza psGBA e inmediatamente a 3' de psGBA se sitúa el gen MTX1, seguido de THBS3 (Vos *et al*., 1992).

El gen MTX1 consta de ocho exones que se distribuyen a lo largo de 6 kb y codifica para la metaxina 1, una proteína de la membrana mitocondrial externa implicada en la importación de sustancias. Se sabe que MTX1 es esencial para la supervivencia celular y para el desarrollo temprano del embrión de ratón (Bornstein *et al*., 1995). Recientemente, se ha identificado una proteína, denominada metaxina 2, que se une a la

metaxina 1 formando un complejo proteico, aunque la función precisa de MTX1 en el mecanismo de importación no se ha descrito (Armstrong *et al*., 1999). MTX1 es el primer componente del aparato de translocación proteica mitocondrial en mamíferos para el que se ha descrito una mutación heredada (Armstrong *et al*., 1997).

Se creyó durante un tiempo que MTX1 compartía un promotor bidireccional con THBS3. Aunque la posición cercana (1,3 kb separan sus inicios de traducción) y la orientación convergente de ambos genes sugirieron esta posibilidad, varias líneas de evidencia apuntan a una regulación independiente (Collins *et al*., 1996; Armstrong *et al*., 1997; Collins *et al*., 1998). El gen THBS3 consta de 23 exones que se distribuyen a lo largo de 12 kb, sólo tres de las cuales son codificantes. Este gen codifica para una proteína (trombospondina 3) secretada a la matriz extracelular y de función hasta el momento desconocida, aunque su pertenencia a la familia de las trombospondinas sugiere que tiene un papel en el control de la función de la matriz extracelular. El gen THBS3 en ratón se regula por un elemento *enhancer* situado en el intrón 6 del gen MTX1. Este elemento es específico de promotor, ya que estimula al promotor basal de THBS3 a una distancia de 5,5 kb, pero no tiene efecto sobre el promotor, más próximo, de MTX1 (Collins *et al*., 1998). La región a unas 3 kb a 3' de THBS3, corresponde al gen MUC1. Este gen codifica para una glicoproteína tipo mucina (episialina) presente en la parte luminal de la mayoría de las células epiteliales (Ligtenberg *et al*., 1990).

Se conoce otro gen que también se localiza en la región 1q21, el gen de la piruvato kinasa tipo L (PKLR). Mutaciones en este gen provocan una enfermedad anémica hemolítica. La distancia entre los finales 5' de GBA y PKLR es de 71 kb y la dirección de transcripción es convergente. El desequilibrio de ligamiento descrito para los polimorfismos de GBA se extiende y abarca al *locus* PKLR. Se han descrito dos polimorfismos en PKLR, que también se organizan en dos grupos haplotípicos mayoritarios y que están en fuerte desequilibrio de ligamiento con los dos haplotipos más frecuentes de GBA (Glenn *et al*., 1994). La recombinación entre ambos *loci* se ha estimado en el 0,071% (Glenn *et al*., 1994; Demina *et al*., 1998; Boas, 2000).

6.2. LA DUPLICACIÓN GBA-psGBA

La duplicación cromosómica que dio lugar a psGBA afectó también al gen MTX1, originando dos genes funcionales, GBA y MTX1, y dos pseudogenes, psGBA y psMTX1. A diferencia de psGBA, psMTX1 no se transcribe (Long *et al*., 1996). Winfield y colaboradores (1997) acotaron los extremos de la zona duplicada por la alineación de las zonas flanqueantes de GBA y psGBA. La duplicación se extiende desde unas 5,5 kb a 5' de GBA hasta el exón 2 de MTX1. La alineación de más de 4 kb a 3' de GBA y psGBA (correspondiente a la mayoría de MTX1 y psMTX1) mostró más de un 97% de homología. A pesar de que psMTX1 no contiene el exón 1 ni parte del 2, el resto de sus secuencias exónicas presenta un 99% de homología con MTX1. La alineación de la secuencia de la región a 5' de psGBA (zona intergénica entre GBA y psGBA) con la región a 5' de GBA mostró un 88% de homología al eliminar de la alineación varias inserciones Alu y un segmento de 6,1 kb a 5' de psGBA. El final 5' de esta región de homología muestra también el final 5' de la duplicación.

En el estudio de Winfield también se dató esta duplicación cromosómica, entre 25 y 40 millones de años. La duplicación tuvo que producirse antes de la divergencia entre grandes simios y monos del viejo mundo, hace 25 millones de años, ya que la duplicación se halla también en monos rhesus. La presencia de un tipo específico de secuencia Alu (perteneciente a la familia Sx) en las dos copias de la duplicación, y de la que se conoce la antigüedad (40 millones de años), acotó el periodo máximo.

OBJETIVOS

Todo lo que una persona puede imaginar, otros pueden hacerlo realidad.

Julio Verne

Nuestros objetivos para el estudio del patrón de diversidad humana en psGBA han sido:

a) Aportar información sobre la cantidad de variación que existe en la secuencia de DNA en el genoma nuclear humano, así como caracterizar de qué tipo de variación se trata, cómo se organiza y cuáles son los mecanismos que la han originado.

b) Estudiar una región genómica nuclear autosómica que proporcionara mayor alcance temporal a los estudios de genética de poblaciones que el DNA mitocondrial, el cromosoma Y o el cromosoma X. El patrón de diversidad genética en psGBA podría ser otra pieza en la reconstrucción de la historia de las poblaciones humanas, que contribuiría al entendimiento de la diversidad genética humana y de la evolución del propio genoma humano.

c) A pesar de su importancia en las bases moleculares de la enfermedad de Gaucher, hasta el momento la única secuencia genómica completa disponible del pseudogén para la glucocerebrosidasa era la que se describió en el artículo de Horowitz y colaboradores (1989) (Genbank J03060). El conocimiento de la variabilidad en psGBA podría ayudar a la identificación de alelos GBA complejos y por tanto a entender mejor las complejas relaciones fenotipo-genotipo de este trastorno.

d) Al estudiar una región no funcional, un pseudogén en este caso, en principio estamos considerando una región no sometida a presión selectiva, en la que podríamos observar los fenómenos mutacionales en su estado de más completa aceptación. Existen datos sobre la variabilidad de su homólogo funcional (GBA), y ambos *loci* se sitúan en el mismo contexto genético. Por tanto, podríamos comparar el espectro mutacional de psGBA con el espectro mutacional de GBA (bajo selección purificadora) y ver cuál es la diferencia entre mutación producida y mutación aceptada.

MATERIAL Y MÉTODOS

El análisis destruye los conjuntos. Algunas cosas, las cosas mágicas, han sido hechas para permanecer enteras. Si uno las observa por partes, desaparecen.

Robert James Waller

1. MUESTRAS

Con la intención de obtener una representación global de la variabilidad humana en psGBA, constituimos la muestra del estudio con diez poblaciones representantes de las principales áreas geográficas mundiales (Figura 5). Del África subsahariana hemos incluido a pigmeos biaka, del sudoeste de la República Central Africana, y a una muestra de tanzanos del distrito Kilombero en la región de Morogoro del sudeste de Tanzania. Del África del norte incluimos a la población saharaui, del Sahara occidental. Como representante del sudoeste asiático incluimos a una representación de drusos, del norte de Israel. Para representar a la variabilidad en Europa incluimos a la población catalana, de la provincia de Gerona, y a una muestra de población vasca proveniente de la provincia de Guipúzcoa. Del continente asiático analizamos una muestra de la población yakut, del noreste de Siberia, y de chinos Han, la etnia mayoritaria en China. De América incluimos una muestra de población maya, del estado de Campeche en el Yucatán. Para representar el área del Pacífico analizamos una muestra de nasioi, de la isla de Bougainville en Melanesia.

De cada una de las poblaciones se analizó la región psGBA en diez cromosomas pertenecientes a cinco individuos no emparentados entre sí.

Se incluyeron en la muestra dos chimpancés (*Pan troglodytes*) no emparentados entre sí, y dos gorilas (*Gorilla gorilla*) no emparentados entre sí, en los que también se analizaron los haplotipos de psGBA. Además, el gen funcional GBA se secuenció en un chimpancé.

1.1. OBTENCIÓN Y CUANTIFICACIÓN DE DNA

El DNA de las muestras de biaka, maya, yakut, chinos, drusos y nasioi se obtuvo de líneas celulares de linfoblastos mantenidas en el laboratorio de K.K. Kidd y J.R. Kidd en la Universidad de Yale.

Las muestras de primates no humanos se obtuvieron del Zoológico de Barcelona.

Las muestras de tanzanos fueron proporcionadas por la Dra. Clara Menéndez de la Unidad de Epidemiología y Bioestadística del Hospital Clínico de Barcelona.

Las muestras de saharauis, vascos y catalanes provienen de distintas campañas de recogida de muestras efectuadas por miembros de nuestro laboratorio.

El DNA de las muestras de saharauis, vascos, catalanes, chimpancés y gorilas se extrajo de muestras de sangre, a partir de *vacutainers* de 5 ml con citrato sódico. El DNA de muestras de tanzanos se extrajo a partir de 1 ml de sangre. A partir de las muestras de sangre de algunos individuos saharauis fue posible establecer líneas celulares de linfoblastos inmortalizadas por infección con el virus Epstein-Barr, lo que supone una fuente indefinida de DNA .

Para la extracción de DNA seguimos un procedimiento estándar basado en la lisis celular seguida por digestión con proteinasa K, una serie de extracciones con fenol-cloroformo para eliminar las proteínas de la muestra, y finalmente precipitación del DNA extraído con etanol (Sambrook *et al*., 1989). La cuantificación de DNA en las muestras se realizó midiendo la absorbancia en un espectrofotómetro a 260 nm (una unidad de absorbancia a 260 nm equivale a 50 ng/µl de DNA de doble cadena). Se calculó la relación de absorbancias a 260 nm y 280 nm ($D.O._{260}/D.O._{280}$) para determinar la relación de ácidos nucleicos y proteínas en cada muestra: un valor de 1.8 resulta óptimo, mientras que un valor inferior indica exceso de proteínas o restos de fenol en la muestra, y un valor superior indica presencia en exceso de RNA. Las muestras se llevaron a una concentración de trabajo de 100 ng/µl.

Figura 5. Situación geográfica de las diez poblaciones analizadas en este estudio.

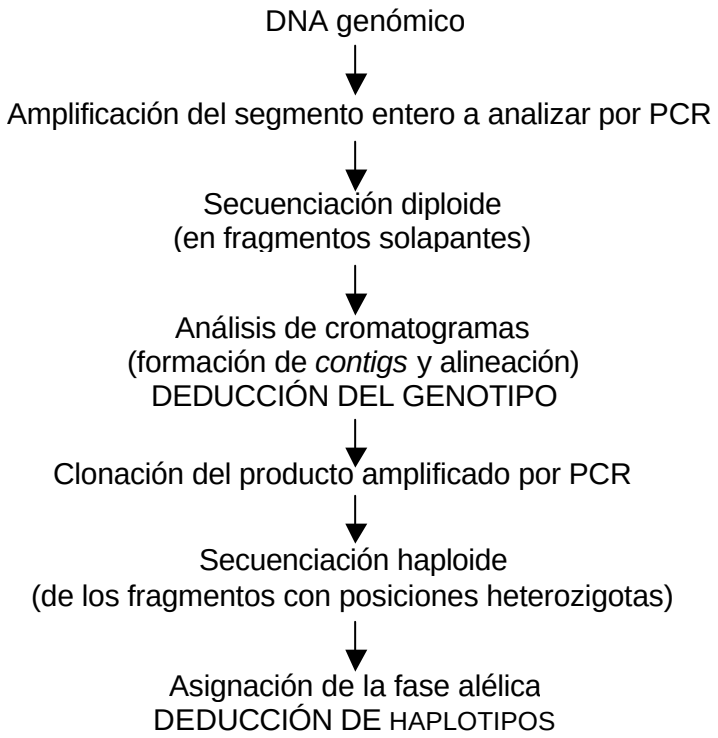
2. OBTENCIÓN DE DATOS

Para la reconstrucción de filogenias con secuencias de DNA necesitamos trabajar con haplotipos. De la amplificación por PCR o de la secuencia de DNA directa a partir de un *locus* autosómico de un individuo obtenemos el genotipo. El genotipo no especifica la fase en que se distribuyen los lugares heterozigotos. Los haplotipos expresan la relación de fase de todos los lugares heterozigotos en un *locus* determinado y para un individuo concreto. Para convertir datos de genotipos autosómicos en haplotipos, normalmente se requiere el análisis de genealogías, o un esfuerzo experimental adicional, como la amplificación específica de alelo. También se han desarrollado aproximaciones informáticas para la inferencia de haplotipos, que sustraen haplotipos conocidos de genotipos con múltiples heterozigotos, iniciando una cascada de resolución de haplotipos. No obstante, los métodos de inferencia no suelen resolver todos los haplotipos de la muestra, especialmente si las posiciones segregantes son numerosas (Clark, 1990).

El objetivo principal de la obtención de datos del presente trabajo es conseguir haplotipos a partir de un *locus* autosómico, es decir, queremos conocer la secuencia exacta de los dos alelos que porta cada individuo. Para ello, desarrollamos una nueva aproximación metodológica, a partir de técnicas ya conocidas, que hemos preparado en forma de manuscrito y que se presenta a continuación. El trabajo realizado en el laboratorio se presenta de manera esquemática en la Figura 6. La metodología seguida se puede resumir en las siguientes etapas: amplificación diploide del *locus* psGBA, secuenciación diploide de DNA y obtención del genotipo, clonaje de los alelos, secuenciación haploide de DNA de los fragmentos en que se habían detectado posiciones heterozigotas y, finalmente, asignación de la fase y deducción de alelos.

En total se secuenciaron alrededor de 600 kb de DNA, y se obtuvo el haplotipo psGBA de 108 cromosomas (100 cromosomas humanos, cuatro de chimpancés y cuatro de gorilas).

Figura 6: Aproximación metodológica para la obtención de haplotipos a partir de organismos diploides.



DETERMINATION OF LONG-RANGE HAPLOID DNA SEQUENCES IN HUMANS: APPLICATION TO THE GLUCOCEREBROSIDASE PSEUDOGENE

Rosa Martínez-Arias, David Comas, Jaume Bertranpetit

Publicado en DNA Sequence:

Martinez-Arias R, Bertranpetit J, Comas D "Determination of haploid ***DNA*** sequences in humans: application to the glucocerebrosidase pseudogene", ***DNA Sequence***, 2002 Feb; 13 (1), pp. 9-13

DETERMINATION OF LONG-RANGE HAPLOID DNA SEQUENCES IN HUMANS: APPLICATION TO THE GLUCOCEREBROSIDASE PSEUDOGENE

Rosa Martínez-Arias, David Comas, Jaume Bertranpetit.

Unitat de Biologia Evolutiva
Facultat de Ciències de la Salut i de la Vida
Universitat Pompeu Fabra
Doctor Aiguader 80
08003 Barcelona, Spain

Running title: Determination of haploid DNA sequences in humans

Address correspondence to:
Jaume Bertranpetit
Unitat de Biologia Evolutiva
Facultat de Ciències de la Salut i de la Vida
Universitat Pompeu Fabra
Doctor Aiguader 80
08003 Barcelona, Spain
E-mail: jaume.bertranpetit@cexs.upf.es
Telephone: 34+93+5422840
Fax: 34+93+5422802

Key words: haplotype, diploid organisms, sequencing analysis, genotype, glucocerebrosidase pseudogene.

ABSTRACT

Objective: Variation analysis in the human genome at the sequence level, specially human genetic population analysis, are hampered by the difficulty to ascertain haplotypes on autosomal regions. We have designed a new methodological approach to obtain autosomal haploid sequences from diploid organisms.

Methods: First, genotypes are unambiguously determined through long-range PCR and diploid DNA sequencing. Second, the allelic phase is discerned by cloning the whole PCR-amplified segment, and sequencing a single clone for those fragments which presented a heterozygous position. The second allele is deduced from the genotype, and the phase reconfirmed by sequencing a second clone.

Results: A hundred human chromosomes were analysed for a 5.4 kilobases encompassing the glucocerebrosidase pseudogene on human chromosome 1. Haplotypes were unambiguously ascertained for all samples.

Conclusion: The manner to combine the used techniques makes this approach a novelty. Our approach allows to obtain haploid sequences from diploid organisms in a less time consuming and more accurate manner than in other used procedures.

INTRODUCTION

The accurate description of a genomic region relies on the DNA sequence. For diploid organisms this means not only to recognise the variant sites, but also to determine the phase among them in the two chromosomes of an individual, that is, to reconstruct the haploid sequences or haplotypes. There is discrepancy between which method may give more accurate results most efficiently and, in fact, studies on haploid sequences are scarce in humans. The reports on X-linked and Y-linked loci were the first aim of the analysis, since single alleles can be directly amplified from males. For autosomal regions the analysis is much more difficult and elusive. On regions with previously unknown variability different strategies have been applied: (i) Heterozygote detection and amplification of only one allele by allele-specific PCR [1]. This is a very laborious approach, since allele-specific PCR conditions need to be settled for each heterozygous site. (ii) Diploid DNA sequencing together with software for haplotypes inference, and sequencing of multiple clones [2]. (iii) Haplotype inference by a maximum-parsimony approach, plus haplotypes confirmation by cloning and sequencing representative PCR-amplified segments [3]. (iv) Diploid DNA sequencing together with software for haplotypes inference, and allele-specific PCR [4]. The protocol described here was designed in order to avoid indirect inference of alleles, which may be problematic in certain cases [5]. Also, we tried to maximise accuracy while optimising time and efforts; this procedure is easier to perform and less time consuming than those used in other approaches. None of the methodologies applied are new; the whole approach is what represents a novelty.

MATERIALS AND METHODS

Both chromosomes from fifty unrelated individuals were analysed. DNA was extracted from fresh blood using a standard phenol-chloroform method after digestion with proteinase K. Some samples were obtained from lymphoblastoid cell lines maintained in K.K. Kidd and J. R. Kidd laboratory at Yale University.

In order to amplify 5.7 kb encompassing psGBA, long-range PCR was performed using psGBA specific primers (Table 1). The amplification conditions were as follows: 0.24mM of dNTPs, $6x10^{-5}$ mM of each primer, from 200 to 400 ng of genomic DNA, 1.5 mM $MgCl_2$ buffer, and 3.5 units of Expand™ High Fidelity PCR System kit (Roche), in 50 µl of final volume. The PCR profile starts with a denaturation step of 2 min at 94 ºC, followed by 10 cycles of 15 sec at 94 ºC, 30 sec at 60 ºC, 4 min at 68 ºC, 20 cycles with

the same conditions but with 20 additional elongation seconds per cycle, and a final elongation step at 72 ºC for 8 min. The PCR products were run in a 1% agarose gel to check for amplification and specificity.

PCR products were purified with QIAquick PCR purification kit (Qiagen), and directly sequenced with the ABI PRISM dRhodamine Terminator Cycle Sequencing kit (PE Biosystems) according to supplier's instructions. A battery of primers was designed to obtain the complete sequence of the region (Table 1). The fragments read with each of these primers overlap between successive reactions. The products of the sequence reaction were run in an ABI PRISM 377 automated sequencer (PE Biosystems). The ABI sequence software (version 3.0, PE Biosystems) was used for electrophoresis lanes tracking and for a first basecalling. From the 5.7 kb amplified, around 150 base pairs (bp) from the initial and final segments were not considered, since not all the positions could be clearly confirmed for all the samples. A stretch of 5,420 bp was considered with full confidence.

The chromatograms were imported into the Seqman II software (Lasergene package, DNASTAR Inc.), assembled and aligned. DNA contigs were created for each sample along the whole psGBA segment. This allowed us to check the coincidence between overlapping segments, and to assure the necessary coverage for all positions. The bases that the ABI basecaller software had miscalled as well as the heterozygous positions assigned as homozygous were corrected. The detection of heterozygous sites is the most critical aspect for obtaining correct genotypes. Thus, all chromatograms were also visually screened. The ABI basecaller software does not call heterozygotes, and when two peaks are appreciably different in size assigns the value of the taller of the two. As a result, if we align the chromatograms and only check the nucleotide sequence, as "tract of letters", and not the trace itself (two simultaneous options with the Seqman II software and most of the alignment programmes), we may miss many of these heterozygous sites. The comparison among individuals is essential when doubtful, or lower than expected, peaks appear. The peaks found in heterozygotes have a notably smaller height than the peaks for the same position in a homozygous individual. As an example, Figure 2 (a) shows a fragment of a diploid DNA sequence of a sample heterozygous for a particular position where C and G are present. The individuals homozygous for nucleotides C (b) and G (c) have higher peaks than the heterozygote G/C for this particular position, whereas the remaining peaks of the sequence do not vary significantly from one individual to the other.

The complementary DNA strand was sequenced in a few doubtful cases, where a peak was suspected to be very low, and could be confused with background noise. That for, reverse primers were designed (Table 1). When this strategy did not solve the position, the fragment was re-sequenced from a new PCR product from genomic DNA. In any case, once the heterozygous sites were detected, they were all verified by re-sequencing the region which included the variant site using a different PCR product from genomic DNA in order to reject PCR artifacts. The probability that a PCR error is produced twice in the same position, and towards the same nucleotide, is negligible. Moreover, the possibility of PCR artifacts is considerably reduced with the use of the High Fidelity enzyme mix (with an error rate of 8.5×10^{-6}) instead of Taq polymerase alone (with an error rate of 2.6×10^{-5}) (protocols for Expand™ High Fidelity PCR System kit, Roche).

In the samples with more than one heterozygous position, the 5.7 kb psGBA segment was cloned using the TOPO-XL PCR cloning kit (Invitrogen, Groningen, The Netherlands) following supplier´s instructions. Only those colonies incorporating an insert in the polycloning site and thus interrupting the *ccd*B lethal gene (recombinants) survive. This facilitates the screening for positive recombinants (colonies with plasmids containing the correct insert). Minipreps (Wizard Plus SV Minipreps DNA Purification System kit, Promega) were done from several colonies per sample. The products were digested with the restriction enzyme Eco RI, in order to free the insert from the vector, and to detect positive recombinants. On average, one every three colonies was a positive recombinant. Miniprep products were used as templates for the haploid DNA sequencing.

Those psGBA fragments in which heterozygous positions had been detected were sequenced in one clone per sample. The sequencing conditions were the same as specified above. The chromatograms from haploid DNA sequencing were analysed with the Seqman II software, assembled, and compared with the previously obtained diploid contigs. The diploid sequence obtained initially was used as a reference frame to discard PCR errors in the haploid sequences. Once the heterozygous sites from one allele were established in phase, the sequence of the other allele could be directly deduced from the genotype. For each sample, a second clone was sequenced in order to reconfirm the phase. No chimeric allele was found.

RESULTS AND DISCUSSION

We analysed the glucocerebrosidase pseudogene (psGBA) in 100 human chromosomes (GeneBank AF267177). A tract of 5,7 kilobases (kb) encompassing psGBA was PCR-amplified, and the genotypes were obtained through automated sequencing and chromatograms analysis. Automated sequencing is still problematic for the unambiguous detection of heterozygous sites. Specific software exist (PolyPhred; [6]) but chromatograms with insertions or deletions cannot be read, and its accuracy is not yet complete. Eye checking of the chromatograms for the detection of heterozygotes is still important, because the direct software alignment of the chromatograms can lead to miss some information. The need of this careful visual screening has been previously stated [7]. When the chromatograms are clean the visual detection of heterozygotes is easy, because two neatly depicted peaks in a single position are seen. Length differences, and thus lag between both allele sequences, are detected as shown in Figure 1, where a clear trace abruptly breaks in a tract of double peaks. The lag can be delimited by visually separating the two overlapping sequences (Figure 1b), and confirmed by sequencing the complementary DNA strand. Besides, this superposition of peaks show us a glimpse about how heterozygous sites would look with a given sequencing chemistry.

The haplotypes on those samples without or with only one heterozygous site are unambiguous. For the other samples, the PCR-amplified segment was cloned. In order to discern the phase, we sequenced a clone from each cloned sample, only for those fragments where a heterozygous site had been detected in the diploid DNA sequence. To avoid the possibility of phase disruption due to cloning of heteroduplexes plus bacterial repairing, and thus to reconfirm the allelic phase, a second clone was sequenced. The phase given for the second clone always matched the previously deduced phase, suggesting that the frequency of chimeric alleles might be lower than previously reported [8], at least under our specific protocol. It should be considered that the probability of heteroduplex formation of long DNA molecules is very low, because of the handicap of the long fragments against the primers on the reannealing cycles. Also, the longer reannealing times used in the study of Jansen and Ledley [8] favour the formation of heteroduplexes.

It is essential to clone the whole PCR amplified segment, and to sequence always the same clone for a given sample in order to work on the same allele. The haploid DNA sequencing always confirmed the previously detected heterozygous sites, since the

relative size of the peak in the haploid sequence became higher than in the diploid sequence (Figure 2, samples (a) and (d)).

Once all sequencing was done, PCR errors were summarised. In a total of 550 kb of diploid DNA sequence, four putative heterozygous sites were not present when checking through a new PCR from genomic DNA and re-sequencing. Thus, these cases are likely to be PCR errors. On the other hand, in some sites the haploid DNA sequence did not match the diploid sequence in a particular homozygous position. These sites were considered as PCR errors; they would have been spread during the amplification in a few molecules in the last PCR cycles, and thus they are in a frequency low enough not to be detected on the diploid sequencing, but high enough to be picked in the cloning process. Along 33 kb of haploid DNA sequencing, 35 errors were detected. When comparing between both types, PCR error detected on haploid sequences (seen on sequences of clones) was significantly higher than the PCR error detected on diploid sequences (χ^2=265.4, 1.d.f., p<0.0001). This fact might be relevant when haplotypes are discerned relying only on cloning, without considering the diploid DNA sequence.

The unambiguous resolution of the diploid sequence (the genotype) is the essential step of this approach: it avoids to sequence many clones per sample, and allows to sequence only those fragments with a heterozygous site. Given that the probability of heteroduplexes formation is extremely low for long DNA fragments, sequencing a single clone would have probably been sufficient. This makes our protocol especially useful for the processing of large amounts of samples.

ACKNOWLEDGEMENTS

We received valuable help on the setting of the methods, as well as helpful comments on an earlier version of the manuscript, from Kenneth M. Weiss and Anne Buchanan. We are indebted to Luís Pérez-Jurado for explanations about heteroduplexes. We are grateful to Kenneth K. Kidd and Judith R. Kidd for sharing DNA samples. This research was supported by Dirección General de Investigación Científica y Técnica grant PB98-1064, and by Generalitat de Catalunya, Grup de Recerca Consolidat 1999SGR00009. R. M-A. received a fellowship from the Spanish Ministry of Education and Culture (AP96).

REFERENCES

1. Harding R , Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB: Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 1997; 60:772-789.
2. Clark, GA , Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF: Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 1998; 63:595-612.
3. Jin L, Undrehill P, Doctor V, Davis RW, Shen P, Cavalli-Sforza LL, Oefner PJ: Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proc Natl Acad Sci USA 1999; 96:3796-3800.
4. Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor ST, Stengård JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, and Sing CF: Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 2000; 67 (4): 881-900.
5. Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 1990; 7(2):111-122.
6. Nickerson DA, Tobe VO, Taylor SL: Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 1997; 25: 2745-2751.
7. Shimmin LC , Miller J, Tran HN, Li WH: Contrasting levels of DNA polymorphism at the autosomal and X-linked visual color pigment loci in humans and squirrel monkeys. Mol Biol Evol 1998; 15 (4):449-455.
8. Jansen R, Ledley FD: Disruption of phase during PCR amplification and cloning of heterozygous target sequences. Nucleic Acids Res 1990; 18 (17):5153-5156.

FIGURE LEGENDS

Figure 1: Effect of a triplet deletion in one of the alleles in a diploid DNA sequence. (a) homozygote individual without the deletion; the CTC deletion in (b) is underlined in red. (b) heterozygous individual, with the deletion in one of the alleles. Both alleles can be ascertained if the two nucleotides are read in every double peak. The resulting sequence of a segment of the overlapped alleles is written in (b), and the missing CTC is indicated in red.

Figure 2: Comparison of the peaks found in diploid (in heterozygotes and homozygotes) and haploid sequences. (a) C/G heterozygote, miscalled C by the ABI software; (b) C/C homozygote; (c) G/G homozygote; (d) haploid sequence for the individual (a) showing a G in the heterozygous site. The height of the nucleotides in homozygotes, (b) and (c), is clearly higher than the height of the peaks in heterozygotes (a).

Table 1: Location and sequence of the PCR and sequencing primers.

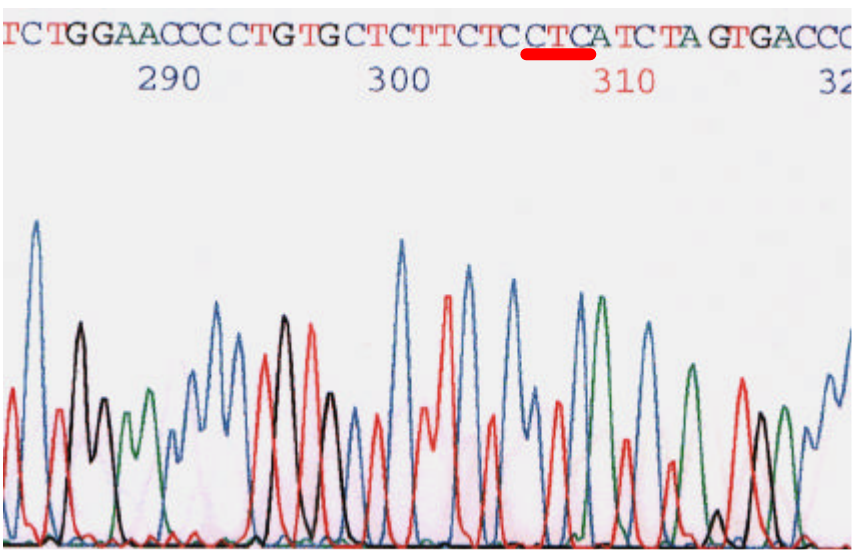
PRIMER	POSITION [a]	SEQUENCE, 5' to 3'
PCR primers		
F [b]	-175 to -148	acatcacggtagcctcagcatgttgtg
R	5462 to 5491	ccccaagactggtttttctactctcatgac
Sequencing primers		
F4[c]	-26 to -5	ggaatctttacccgattctcca
F18	313 to 335	ctagtgaccctgaggtgatggag
F5	440 to 464	tgaatccagggccatcatggctctt
R5	546 to 567	cccttctgatgaaaactctctg
R12	810 to 830	cagatgagtgagtcaaggcag
F7	930 to 952	tgagtgactgagaccaactttgg
F8	1296 to 1318	ggtgtcagtgatcaccatggagt
F17	1866 to 1889	agctgtgacttctccatccgcacc
R3	2089 to 2114	agtgcatccggttcagccattagccc
F23	2327 to 2348	ctgggccagatacattgtgaag
F21	2578 to 2601	gccaggttcctggatgcctatgct
R14	2698 to 2719	gaacatcagcgagacttcattg
F11	3354 to 3378	gaggaactagaagttccagaagcct
R8	3568 to 3588	cactctgctcccagaacttgg
F22	3784 to 3806	gtcttgcctttccttcacaggtc
F13	4039 to 4061	tagaacctcctgtaccatgtggt
R11	4222 to 4241	ctacaccacgagggagcagg
R13	4301 to 4322	caggaagtgactaggtagcaac
F14	4410 to 4431	gcagagccttgcaggagttat
F15	4782 to 4804	ggagacaatctcacctggctact
R10	5422 to 5444	ctcctcgttgttggtacagccgg

[a] Numeration is according to GeneBank AF267177.

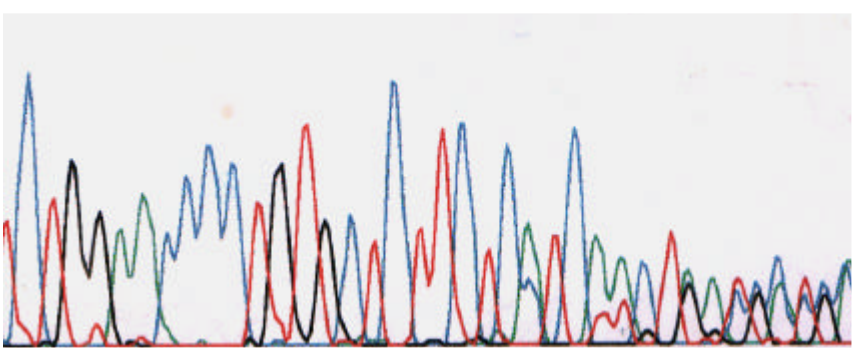
[b] F stands for forward and R for reverse.

[c] In bold, primers needed to obtain the whole contig.

a)



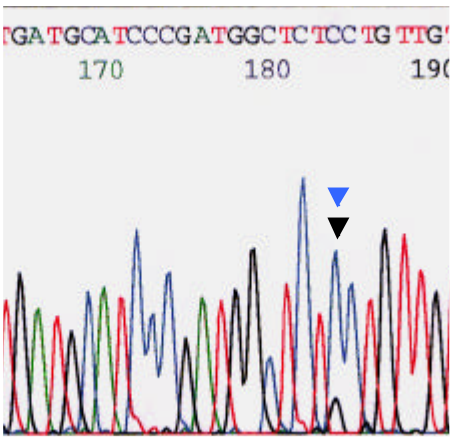
b)



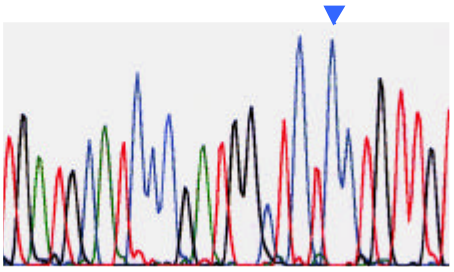
CTCCTCATCTAGTGA
CTCATCTAGTGACCC

DIPLOID

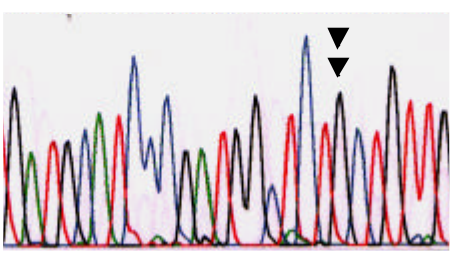
a)



b)

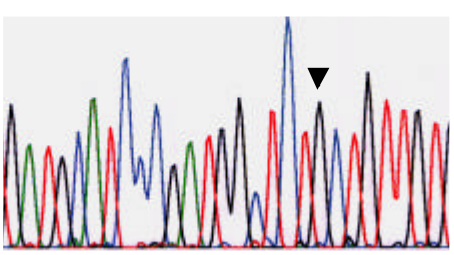


c)



HAPLOID

d)



3. ANÁLISIS DE DATOS

Los análisis que hemos utilizado para tratar los datos haplotípicos y de diversidad obtenidos se detallan en cada uno de los capítulos del apartado de Resultados. A continuación listamos solamente los programas informáticos que hemos utilizado, y los análisis que hemos realizado con cada uno de ellos.

DnaSP (*DNA Sequence Polymorphism*, versión 3.14; Rozas y Rozas, 1999), para el cálculo de los siguientes parámetros: H o diversidad haplotípica; los estimadores de la variabilidad nucleotídica, $\hat{q}_p$, la heterozigosidad media observada por nucleótido (Nei y Li, 1979) y $\hat{q}_W$, la diversidad nucleotídica esperada a partir del número de sitios segregantes (Watterson, 1975); los estadísticos D de Tajima (Tajima, 1989) y D, F, F* y D* de Fu y Li (Fu y Li, 1993), que se calcularon para testar posibles desviaciones del modelo neutro; la divergencia Dxy entre psGBA en humanos y chimpancés; la comparación del ratio polimorfismo/divergencia mediante el test HKA (Hudson *et al*., 1987); la representación en *sliding windows* de distintos parámetros de diversidad, para observar como varían estos parámetros a lo largo de la zona de interés; el estimador de la tasa de recombinación $\hat{C}$ (Hudson, 1987).

Network 2.0b (Bandelt *et al*., 1995) se utilizó para establecer conexiones en red (*median joining networks*) entre los haplotipos de nuestra muestra. Este método ha demostrado ser muy útil para visualizar relaciones filogenéticas entre distintos haplotipos, porque muestra gráficamente toda la información contenida en los datos, es decir, establece todas las conexiones filogenéticas posibles entre los haplotipos. También predice haplotipos que no están presentes en nuestra muestra pero que pueden existir en la población o haber existido en el pasado, ayuda a identificar homoplasia y posibles recombinaciones, que aparecerán como reticulaciones en la red, y es útil para visualizar haplogrupos (Bandelt *et al*., 1995).

El paquete Phylip 3.5c (Felsenstein, 1989) se utilizó para calcular matrices de distancias calculada con el programa DNADIST, y para construir árboles *neighbor-joining* (Saitou y Nei, 1987) a partir de estas matrices. El algoritmo de *neighbor-joining* permite representar en forma de árbol una matriz de distancias genéticas (que contiene el valor

obtenido para cada par de comparaciones). Como características, genera árboles sin raíz, tiende a minimizar la longitud total de las ramas del árbol, no presupone una tasa constante de evolución, y por tanto no fuerza una misma longitud para todas las ramas del árbol. Esta última característica implica que la correlación cofenética (correlación entre las distancias originales y las distancias representadas en el árbol) es mejor con éste que con otros algoritmos de representación de árboles filogenéticos. A fin de estimar el error estadístico asociado a las distancias genéticas, y para encontrar una medida de la robustez estadística de los árboles genéticos, utilizamos métodos de *bootstrap*, basados en el remuestreo: los alelos que intervienen en el cálculo de las distancias genéticas se remuestrean de manera que en cada remuestra hay alelos representados más de una vez y alelos ausentes (pero siempre conservando el número total de alelos). A partir de cada remuestra se calcula una nueva matriz de distancias y un nuevo árbol. Cada vez que un grupo concreto aparece en los árboles *bootstrap* se cuenta y el número total se da como porcentaje en el árbol original. Estos porcentajes serán los indicadores de la robustez estadística de cada grupo o *cluster*.

El programa Sites (Hey y Wakeley, 1997) se utilizó para calcular el parámetro de recombinación γ.

El programa Genetree (Griffiths y Tavaré, 1994; Griffiths y Tavaré, 1998a; Griffiths y Tavaré, 1998b) se utilizó para estimar tiempos de coalescencia entre alelos y la edad de las mutaciones. La teoría de la coalescencia consiste en la reconstrucción de la genealogía de una muestra de alelos, tomando como base que los alelos siguen el modelo neutralista, el modelo mutacional de infinitos lugares (*infinite-sites model*; asume que cada nueva mutación ocurre en un nucleótido que no había mutado anteriormente), y que el tiempo hasta el ancestro común de dos alelos depende del tamaño efectivo de la población. Es decir, la probabilidad de que dos linajes se fusionen en el tiempo es función del tamaño poblacional. Esta teoría describe en términos probabilísticos las relaciones ancestrales que esperaríamos encontrar en una muestra de secuencias, antes de que estas secuencias aparezcan. Las ramas del árbol generado convergen en un nodo cuando las secuencias comparten por primera vez un ancestro, de manera que en la raíz del árbol se localiza el ancestro común más reciente de la totalidad de la muestra (MRCA, *most recent comon ancestor*) (Tavaré *et al*., 1997). Cada MRCA es aleatorio en el sentido de que diferentes muestras producirán distintos ancestros; cada uno de los árboles "parciales" permitiría visualizar diferentes partes del árbol total que uniría a todos los

individuos que han existido a lo largo de la historia de la población (Griffiths y Tavaré, 1994).

Arlequin 2.000 (Schneider *et al*., 2000), se utilizó para realizar los siguientes análisis:

- El cálculo del estadístico Fst, que indica la cantidad de la variabilidad total que se encuentra entre las poblaciones. Se puede calcular como la variancia genética (heterozigosidad) entre poblaciones, dividida por la variancia genética de la población total. Para dos poblaciones se puede utilizar como medida de la distancia genética. Así por ejemplo, si dos poblaciones divergen y permanecen genéticamente aisladas, se espera que Fst crezca con el tiempo.

- La estima de frecuencias haplotípicas a partir de datos genotípicos de psGBA y GBA, así como para estimar la probabilidad de desequilibrio de ligamiento entre los haplotipos estimados de ambos *loci*.

- La matriz de distancias genéticas *intermatch/mismatch* entre poblaciones. Entre dos poblaciones *i* y *j* esta distancia se calcula como la media de diferencias nucleotídicas entre todos los pares de secuencias posibles que contengan una secuencia de la población *i* y una secuencia de la población *j*; a esta media se le sustrae la diversidad media dentro de cada población, medida como la media de diferencias nucleotídicas entre pares de secuencias pertenecientes a la misma población. De la comparación de todos los cromosomas de una población con todos los cromosomas de una segunda población se obtiene una medida de la distancia genética entre las dos poblaciones. Además de considerar las diferencias en frecuencias haplotípicas, esta medida de distancia genética considera también la diferencia molecular (y previsiblemente filogenética) entre haplotipos distintos. Al comparar distintas poblaciones entre ellas obtendremos una matriz de distancias. La información contenida en esta matriz se puede representar de distintos modos, por ejemplo en forma de árbol filogenético, o bien en un plano mediante el análisis de coordenadas principales, de manera que las distancias entre las poblaciones en el plano sean lo más proporcionales posible a las distancias de la matriz.

NTSYS (versión 1.70; Applied Biostatistics, Inc.) se utilizó para realizar el análisis de coordenadas principales, a partir de la matriz de distancias genéticas entre poblaciones calculada con el programa Arlequin 2.000. Las coordenadas principales fueron representadas gráficamente mediante Microsoft Excel 97.

El paquete SPSS 9.0.1 se utilizó para el análisis de componentes principales. Este análisis permite construir pocas variables que sintetizan la información contenida en un número alto de variables observadas. La ventaja del análisis es que posibilita la representación gráfica de datos multidimensionales en un número reducido de dimensiones. Estas nuevas variables se ordenan según la cantidad de variabilidad inicial que recogen, y son independientes unas de otras. Procesos poblacionales más antiguos, que operaron sobre tamaños poblacionales reducidos, y por tanto tendrían más impacto en la diferenciación genética posterior de las poblaciones, darán componentes principales más altos, es decir, que explicarán mayor parte de la variabilidad. El análisis se realiza a partir de variables medidas en una serie de poblaciones (en nuestro caso las frecuencias haplotípicas). Al identificar grupos de variables correlacionadas entre ellas, obtendremos una representación de la relación entre las poblaciones, y de cuáles son las variables más correlacionadas entre sí. La diferencia de este análisis con el de coordenadas principales es que este último se hace a partir de una matriz de distancias genéticas, mientras que el análisis de componentes principales no presupone un modelo genético, sino que es un tratamiento puramente estadístico.

El exceso de *singletons* (variantes nucleotídicas que aparecen en la muestra una sola vez), se testó con el estadístico Fs de Fu, (Fu, 1997), calculado a través del programa disponible en la página www.hgc.sph.uth.tmc.edu/fu/.

RESULTADOS

Com a organismes evolucionats en la naturalesa, som summament sensibles a les seves característiques de conjunt; com a organismes racionals tenim tendència a una dissecció i anàlisi de la naturalesa.

Ramon Margalef

Capítulo I:

SEQUENCE VARIABILITY OF A HUMAN PSEUDOGENE

Rosa Martínez-Arias, Francesc Calafell, Eva Mateu, David Comas, Aida Andrés, Jaume Bertranpetit

Consultable en:

Martinez-Arias R, Calafell F, Mateu E, Comas D, Andres A, Bertranpetit J. ”Sequence variability of a human pseudogene”. ***Genome Research***. 2001 Jun;11(6):1071-85

Capítulo II:

SELECTION SHAPING VARIABILITY ON A HUMAN PSEUDOGENE

Rosa Martínez-Arias, David De Lorenzo, Eva Mateu, Francesc Calafell, Jaume Bertranpetit

(Sometido a consideración editorial; *Genetics*)

SELECTION SHAPING VARIABILITY ON A HUMAN PSEUDOGENE

Rosa Martínez-Arias,*[,1] David De Lorenzo, H[,1] Eva Mateu,* Francesc Calafell, *Jaume Bertranpetit*

*Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, 08003 Barcelona, Spain

HDepartment of Evolutionary Biology, Institute of Zoology, Ludwig-Maximilians University, 80333 Munich, Germany

1, These authors have contributed equally to this work.

RUNNING HEAD: selection on a human pseudogene

KEY WORDS: selection, pseudogene, hitchhiking, glucocerebrosidase

CORRESPONDING AUTHOR:
Jaume Bertranpetit
Unitat de Biologia Evolutiva
Universitat Pompeu Fabra
Dr Aiguader, num. 80
08003, Barcelona
Spain

Telephone number: 34-93-542 28 40
FAX number: 34-93-542 28 02

e-mail: jaume.bertranpetit@cexs.upf.es

ABSTRACT

Pseudogenes are considered to be free from selective pressure and therefore to accumulate more variation than functional genes do. However, reduced levels of nucleotide diversity have been observed on a human autosomal glucocerebrosidase pseudogene (*psGBA*), the non-functional duplicate of the *GBA* gene. A high number of singletons and unique haplotypes, and two haplotypes found at high frequency are the most striking features on the *psGBA* variability pattern. This pattern has been analysed through theoretical evolutionary models. The observed frequency spectrum of segregating sites and the allelic partition do not fit the pattern expected under the neutral model. In addition, Fu´s *Fs* and HKA tests are statistically significant. Therefore, the amount of variation in *psGBA* cannot be understood under a neutral model. The possible role of population events, purifying selection, background selection, and balancing selection are discussed. Genetic hitchhiking seems to be the likeliest cause for the variability pattern observed, either as two successive selective sweeps, or as a single hitchhiking event plus recombination. Our findings point to genome context as a key element to understand variation in the human genome.

INTRODUCTION

The glucocerebrosidase pseudogene (*psGBA*), is the non-functional duplicate of the glucocerebrosidase gene (*GBA*). Mutations on *GBA* produce Gaucher disease in humans, the most prevalent sphingolipid accumulative disorder (OMIM 230800, 230900 and 231000, for Gaucher type 1, type 2, and type 3, respectively). *psGBA* and its functional counterpart are 16 kilobases (kb) apart at human chromosome 1q21, a centromeric region very rich in functional genes (Winfield *et al.*, 1997). The duplication that originated *psGBA* also affected the *MTX1* gene, producing two functional genes, *GBA* and *MTX1*, and two pseudogenes, *psGBA* and *psMTX1*. Both pseudogenes have maintained a high degree of structural and sequence homology with their functional counterparts (96% for *psGBA* and higher for most of *psMTX1*). All four loci are located in a region of less than 40 kb (Long *et al.*, 1996). *psGBA* is transcribed, but the aberrant splicing and the accumulated changes scattered along the sequence break the reading frame and prevent the translation to functional product (Sorge *et al.*, 1985; Imai *et al.*, 1993).

By definition, pseudogenes do not have any biological function and therefore are considered free from any selective pressure. Since purifying selection would not act on them, it is expected that all types of mutations will accumulate at a faster pace than for their respective functional genes, and that mutation will not be confused with mutation acceptance due to selective processes. Studies on pseudogenes have indeed shown that they evolve at a faster rate than their functional counterparts. In fact, substitution rate on pseudogenes has been suggested as a global estimate of the spontaneous point mutation rate (Li *et al*., 1981; Li, 1987). However, the variability observed in *psGBA* is lower than expected. Along 100 human chromosomes and a region surveyed of 5,420 nucleotides encompassing the whole *psGBA* fragment, 18 segregating sites were found, comprising 17 substitutions, and a single three base pairs (bp) deletion (R. Martínez-Arias, F. Calafell, E. Mateu, D. Comas, A. Andrés, J. Bertranpetit, unpublished results). Nucleotide diversity at *psGBA* (p=0.00044) is lower than for other human nuclear loci, and interestingly lower than for coding regions, studied at the sequence level: p=0.0011 for 49 autosomal loci (Li and Sadler, 1991); p=0.0018 for the *b-globin* gene (Harding *et al.*, 1997); p=0.002 for the Melanocortin 1 Receptor locus, *MC1R* (Rana *et al*. 1999); p=0.0008 as an average for 75 genes (Halushka *et al.*, 1999), p=0.0005 for the Apolipoprotein E gene (Fullerton *et al*., 2000). This first appreciation will be tested assuming a neutral model for the pseudogene, and the possible causing factors of the reduction in variability will be analysed.

Many factors may have acted in the past to account for the variability pattern found on a given genomic region. In the simplest, neutral model, only mutation and random genetic drift affect the variability level of a locus. The consideration of other genome-related and population-related mechanisms, may lead to a more realistic view in most cases. These processes will leave a particular footprint on the structure and extent of the variation on the genome. One of the challenges of present day population genetics and molecular evolution is to unravel these footprints, that is, to ascertain which of these factors have lead to the observed patterns of genomic variation on a given locus.

In our previous manuscript we have described the diversity pattern on *psGBA*, and the role of mutation, recombination and gene conversion on its shaping. The purpose of the present paper is to infer the importance of selection on *psGBA*. The analysis of the patterns of variation through theoretical models may allow to understand which forces have shaped the variability pattern of *psGBA*, and may help to rule out alternative possibilities. The influence of the genetic context will be shown to be extremely important. The general idea of pseudogenes as duplicated genomic regions free for mutational

events has to incorporate the genetic context concept, which may be of great importance for understanding variation in the human genome.

MATERIALS AND METHODS

Ten chromosomes from each of ten worldwide populations were analysed: Biaka Pygmies (from the Central African Republic), Tanzanians (from the region of Morogoro in the South East of Tanzania), Saharawi (from Western Sahara), Druze (from Northern Israel), Basques and Catalans (from the Iberian Peninsula), Yakut (from Siberia), Han Chinese, Mayan (from Yucatan), and Nasioi (from Melanesia). Four chimpanzee (*Pan troglodytes*) chromosomes were also included in the sample.

A 5.7 kb segment encompassing the *psGBA* region was PCR-amplified. The PCR product was directly sequenced on an automated sequencer (ABI PRISM 377, PE Biosystems), using the ABI PRISM dRhodamine Terminator Cycle Sequencing Ready Reaction kit with Ampli Taq Polymerase (PE Biosystems). The same primer pair and PCR conditions were used to amplify and sequence human and chimpanzee samples. Details on the PCR and sequencing conditions and primers are available under request.

The resulting chromatograms were assembled and analysed with the help of the Seqman II software (Lasergene package, DNASTAR Inc.). Heterozygous sites were detected and the genotypes for all the individuals were obtained. A 5,420 nucleotides stretch encompassing *psGBA* was unambiguously ascertained for all the samples. All heterozygous sites were confirmed through a second PCR amplification and re-sequencing. The whole segment amplified was cloned for those samples that presented more than one heterozygous site, and therefore whose haplotype ascertainment was not obvious. Tracts with heterozygous sites were re-sequenced in one allele from each cloned sample, in order to discern the phase among them. The sequence of the other allele was inferred, and the phase was reconfirmed by sequencing a different clone.

DnaSP software (DNA Sequence Polymorphism version 3.14; Rozas and Rozas, 1999) was used to estimate diversity and divergence parameters, and to perform tests such as the Hudson, Kreitman and Aguadé (HKA) (Hudson *et al.*, 1987), Tajima's *D* (Tajima, 1989), and Fu and Li *D* and *F* (Fu and Li, 1993). The neutral diversity parameter $q = 4N_e\mu$, where N_e is the effective population size, and μ is the mutation rate per nucleotide per generation, was estimated as $\hat{q}_p$, the average heterozygosity per nucleotide site obtained from the mean number of pairwise differences among the sample

sequences (Nei and Li, 1979) and $\hat{q}_W$, estimated through the expected nucleotide diversity from the number of segregating sites (Watterson, 1975). $\hat{q}_W$ and divergence (*Dxy*) along sliding windows through *psGBA* was also performed with DnaSP software.

Network 2.0b analysis software was used to establish median joining networks among haplotypes. This method can infer haplotypes that could have existed in the past or are not present in our sample, and show possible recombinations and homoplasy, which will be visualised as reticulations in the network (Bandelt *et al.*, 1995).

Computer simulations were performed by generating neutral samples (using coalescence theory) and calculating in each sample the parameter or parameters under study. Once obtained the probability distribution, the probability associated with the observed value of the parameter under neutral conditions can be calculated.

Excess of singletons was tested with Fu's *Fs* statistic (Fu, 1997), calculated using the program available at www.hgc.sph.uth.tmc.edu/fu/.

Linkage disequilibrium analysis between *psGBA* and neighbouring regions were calculated with the Arlequin 2.000 package and its significance calculated using Fisher's exact test. A normal approach was used for establishing the 95% confidence interval of the number of segregating sites, calculated as $E(S) \pm 1.96\ Var(S)$, being $E(S)=q\sum_{i=1}^{n}\frac{1}{i}$ and $Var(S)=E(S)+q^2\sum_{i=1}^{n}\frac{1}{i^2}$ (Watterson, 1975).

For the expected number of alleles, Ewens' sampling formula was used (Ewens, 1972) $E(k)=\frac{q}{q}+\frac{q}{q+1}+\cdots+\frac{q}{q+2n-1}$ and

$$Var(k)=E(k)-\left[\frac{q^2}{q^2}+\frac{q^2}{(q+1)^2}+\cdots+\frac{q^2}{(q+2n-1)^2}\right]$$

RESULTS

Diversity pattern on *psGBA*: Haplotypes for a 5,420 nucleotide stretch encompassing the *psGBA* locus were ascertained for 100 human chromosomes (GenBank AF267177). Twenty-five different haplotypes were found, defined by 18 segregating sites, 17 substitutions and a 3-bp deletion. A polymorphic poly-adenine tract was observed, but was not included in the analysis due to its specific evolutionary pattern. Two different haplotypes were found among four chimpanzee chromosomes (GenBank 272642). In

humans, *psGBA* presented a nucleotide diversity value of p=0.00044. The $\hat{q}_p$ estimator was 2.39, and $\hat{q}_W$ was 3.28. Nucleotide diversity is low in comparison to other autosomal data (Li and Sadler, 1991; Harding *et al.*, 1997; Clark *et al.*, 1998; Rieder *et al.*, 1999; Halushka *et al.*, 1999; Rana *et al*., 1999) and especially if considering that the analysed tract is a pseudogene, believed to accumulate high levels of variation. A median network (Bandelt *et al.*, 1995) representing the haplotype phylogeny was divided in two main clusters, each of them centered on one of the two most frequent haplotypes of the sample, i.e., haplotypes A (29% of the chromosomes) and B (23% of the chromosomes). Seven different haplotypes radiate from haplotype A, creating a clear star-like phylogeny, which was not present around haplotype B (Figure 1).

Using the formula $\mathrm{E}(S)=q\,a_n$ (Watterson, 1975) we have calculated the expected number of segregating sites under a non-recombinant neutral model. This value depends on the neutral diversity parameter $q=4N_e\mu$. An estimate of q (0.002 per site) obtained from a previously analysed autosomal region with 90% of non coding nucleotides for which selection has not been detected (the Lipoprotein Lipase locus, *LPL*; Nickerson *et al.*, 1998; Clark *et al.*, 1998) was used as an external reference. Using this value, the expected number of segregating sites is $\mathrm{E}(S)$=56 and its variance *Var (S)*=247. As for long sequences and large sample sizes, the distribution of the number of segregating sites approaches a normal distribution, we can therefore estimate a 95% confidence interval as $\mathrm{E}(S) \pm 1.96\sqrt{Var(S)}$, which gives the interval (25, 88). This means that if we sample repeatedly 100 chromosomes and sequence 5,420 nucleotides in a population with $4N_e\mu$=0.002, 95% of the samples will present a number of segregating sites ranging from 25 to 88. If we consider this interval as an estimation of the level of nucleotide variation in neutral human autosomal loci, it is clear that the observed value of 17 segregating sites at *psGBA*, off its low range, is lower than the value expected under neutrality. To be more precise, the probability of finding 17 segregating sites or less in our sample is 0.0012. This observation is in agreement with the lower density of segregating sites of *psGBA*, 1 in 301 bp, when comparing with other autosomal loci: 1/76 bp for the *b-globin* locus (Harding *et al*., 1997), 1/110 for the *LPL* locus (Clark *et al.*, 1998), 1/159 bp for the *MC1R* locus (Rana *et al*., 1999).

Besides the low number of segregating sites, there is a very irregular pattern in the distribution of their frequencies, which presents an excess of variants little polymorphic (singletons and doubletons) and an excess of variants highly polymorphic (appearing in more than 30 alleles) (Figure 2). This distribution differs significantly (χ^2_{3df}=13.06,

p=0.0045) from the expected neutral frequency spectrum at equilibrium (obtained following Watterson, 1975).

In a model with no recombination (Ewens, 1972), the expected number of alleles and its variance are 11.9 and 8.1 respectively. Therefore, the probability of obtaining 24 alleles or more (the number really found, not considering the haplotype produced by the 3 pb deletion) in a 100-sequences sample with 17 segregating sites using a non-recombination model is $p<0.0001$. This may be caused by forces like recombination or recurrent mutation. To compare the observed allelic partition with the expectations under a neutral model, we have generated a set of samples (conditioned not only to the number of segregating sites and alleles, but also to the recombination rate) and then obtained the distribution of allele frequencies in that sample set (Figure 3). The observed allele distribution does show some deviations from the neutral expectations: i) an excess of alleles at frequency one (due to the excess of singletons, as observed in the polymorphisms distribution). ii) two alleles found at high proportion (which distribute the observed sequences in two sets as seen in the network).

In order to test if the excess of rare alleles (in relation to the expected under a mutation-drift model) is significant, Fu's *Fs* was applied, which was –13.451 ($p<0.050$). This value indicates a clear deviation from the expected allele frequencies spectrum in the sense of an excess of rare alleles.

Nevertheless, the excess of alleles observed does not imply that levels of variation in *psGBA* locus are high. In fact, nucleotide variation is markedly reduced in this region. In spite of the low number of segregating sites, and due to the excess of variants at low frequency, the number of alleles is higher than expected.

Mutation, recombination and linkage disequilibrium: One first possible explanation for the reduction of variation at *psGBA* would be a reduced mutation rate in this region. Substitution rate has been estimated as $1.23\pm0.22\times10^{-9}$ per nucleotide and per year from the nucleotide differences between human and chimpanzee, human and gorilla, and chimpanzee and gorilla *psGBA* sequences, and considering a divergence time of 5, 7, and 7 million years, respectively (R. Martínez-Arias, F. Calafell, E. Mateu, D. Comas, A. Andrés, J. Bertranpetit, unpublished results). This value fits well within the range of values previously estimated for human nuclear genes (Li *et al.*, 1987: 1×10^{-9} as average synonymous rate for several genes; Harding *et al.*, 1997: 1.34×10^{-9}, for the *b-globin* gene; 1.3×10^{-9} for the *LPL* gene, calculated from the data in Clark *et al.*, 1998) and pseudogenes (Li *et al.*, 1981: 4.6×10^{-9}, from several globin pseudogenes; Nachman and

Crowell: 1.25×10^{-9}, from 18 processed pseudogenes). Thus, low mutation rate can be discarded as the cause for the low variability.

Recombination is another essential mechanism on the variability shaping. We computed the level of recombination in *psGBA* as $C=4N_ec$, N_e being the effective population size and c the recombination rate per nucleotide and generation. An estimator of $4N_ec$, the gamma recombination parameter (γ; Hey and Wakeley, 1997), was calculated as 1.271. This estimator is less biased for the number of polymorphisms, small sample sizes, and short sequences than Hudson's $\hat{C}$ estimator (Hudson, 1987; Hey and Wakeley, 1997). From this estimate, a ratio of recombination events per mutation (calculated as $4N_ec/4N_em$, where $4N_ec$ is estimated from γ, and $4N_em$ is estimated from $\hat{q}_W$) of 0.39 was obtained, which points to mutation as a more important factor than recombination on shaping variation at *psGBA*.

The importance of recombination can be further assessed by measuring linkage disequilibrium on this region. When considering linkage disequilibrium within *psGBA* through each pair of segregating positions, linkage was significant in 8 out of 45 pairwise comparisons using Fisher's exact test. This number is not as high as would be expected in a region of reduced recombination, but this may be due to the low frequency of most of the polymorphisms found. With low frequencies, even Fisher's exact test does not have enough power to obtain significant results (Slatkin, 1994). For this reason, we have extended the linkage disequilibrium analysis to a haplotype level including neighbouring regions.

As for *psGBA*, also for the *GBA* gene, located 16 kb upstream of *psGBA*, two haplotypes are found at high frequencies (Beutler *et al.*, 1992; Glenn *et al.*, 1994). The *GBA* haplotypes, named - and +, are defined by 12 polymorphisms in tight linkage. It is also well established that *GBA* haplotypes are completely linked to polymorphisms in the *PKLR* gene, located 70 kilobases upstream of *GBA* (Glenn *et al.*, 1994; Demina *et al.*, 1998). In our previous paper, independence between *psGBA* A/B haplotypes and –/+ *GBA* haplotypes was tested with a chi-square test. Data for the *GBA* haplotypes were extracted from two *GBA* and one *PKLR* polymorphisms analysed by E. Mateu (E. Mateu, F. Calafell, R. Martínez-Arias, A. Pérez-Lezaun, A. Andrés, J. Bertranpetit, unpublished results) for almost 900 individuals that include 94 chromosomes of the present work. Only those individuals for which the phase could be unambiguously ascertained (54 out of the 94 chromosomes) were included in the test. A chi-square value of 8.21 with p=0.004 was obtained, pointing to the fact that both groups of haplotypes are linked (-/B and +/A).

However, this method can be biased towards the inclusion of homozygous individuals in the subsample set, for whom the ascertainment is more likely. Therefore, haplotype frequencies were estimated from genotypic *GBA* and *psGBA* data using a likelihood ratio test (Slatkin and Excoffier, 1996). This test gave two most frequent haplotypes, -/B (23.4 %) and +/A (21.3 %). In order to determine whether the resolution of haplotypes are significantly non-random, a likelihood ratio test between pairs of loci, whose empirical distribution is obtained by a permutation procedure, was calculated. The test was significant (p<0.005), and thus, a strong linkage disequilibrium is present in a wide region encompassing *psGBA*.

Neutrality tests: In order to verify if neutrality could be discarded for *psGBA*, some neutrality tests were applied. We have already stated the significant result of the Fu's *Fs* test. Tajima's *D* (Tajima, 1989) tests the difference between $\hat{q}_p$ and $\hat{q}_W$, both diversity estimators of the neutral parameter $4N_e m$. If the population sample fits the infinite sites model, $\hat{q}_p$ and $\hat{q}_W$ have the same expected value, and D=0. Under non-neutral evolution, since $\hat{q}_W$ is calculated using the number of segregating sites only, it is more affected by low-frequency variants than $\hat{q}_p$, which is calculated using the frequency of each variant. Therefore, an excess of singletons (as expected in cases of recent selective sweeps or a population expansion) will produce negative values of Tajima's *D* (Braverman *et al.*, 1995). The same idea is followed by the *D* and *F* tests (Fu and Li, 1993). The *D* test compares the mutations of the external and internal branches of the gene genealogy. The *F* test compares the $\hat{q}_p$ parameter and mutations at the external branches. We considered an outgroup (the chimpanzee in our case) to calculate the number of external variants. Also in this case, negative values of *D* and *F* indicate an excess of recent mutations (in the external branches of the genealogy).

The calculated Tajima's *D* statistic was –0.764 (not significant, p=0.180 under a zero-recombination model, p=0.185 under a model with recombination, with 1000 coalescence simulations). Fu and Li's *D* statistic was -1.42 (not significant, 0.05<p<0.10; critical point calculated with 1000 simulated samples of size 100: -1.78. The critical point is the value from which higher values have a probability lower than the 5% to be detected), and the *F* statistic was -1.40 (not significant, 0.05<p<0.10; critical point calculated with 1000 simulated samples of size 100: -1.74).

If there is a specific constraint that reduces the levels of intraspecific polymorphism on *psGBA*, the levels of interspecific divergence should be affected as well. To test that, the divergence between human and chimpanzee along the *psGBA* sequence was computed (as *Dxy*) (Figure 4). There is not any region with a marked decrease in divergence that could explain the observed reduced levels of polymorphism. The nucleotide divergence was 0.012, a value consistent with other studies (Whitfield *et al.*, 1995; Harding *et al.*, 1997). We applied the HKA test (Hudson *et al.*, 1987), which is based on the prediction of the Neutral Theory of Molecular Evolution (Kimura, 1983) that regions of the genome that evolve at high rates will present high levels of polymorphism within and between species. When comparing diversity between chimpanzee and human sequences within *psGBA* (one half of is compared against the other half of the locus), no significant differences on the ratio polymorphism/divergence were found between both halves of *psGBA* (χ^2_{1df}=3.74, p= 0.053). Therefore we can consider that *psGBA* evolves as a single unit.

We also used the HKA test to compare *psGBA* with an external reference, the *LPL* locus, as variation at individual nucleotide sites in *LPL* is consistent with a neutral gene balance between mutation and random genetic drift (Clark *et al.*, 1998). The comparison between both loci showed a significant result (χ^2_{1df} =3.99, p=0.046), due to a reduced level of intraspecific polymorphism at the *psGBA* locus. The polymorphism/divergence proportions are 0.83 for the *LPL* region and 0.25 for *psGBA* (more than a three-fold difference between both loci). Under the neutral model, this proportion is expected to be constant within species. Even in the case of regions with reduced mutation rate, both polymorphism and divergence would be reduced, keeping the proportion between these two quantities the same. For the analysed regions, the significant difference between the ratio polymorphism/divergence shows a significant reduction in the levels of polymorphism at *psGBA*. The significant value of the HKA test is an indicator that selection shaped the evolution of *psGBA*.

DISCUSSION

psGBA presents low nucleotidic diversity in comparison to what would be expected for a pseudogene, and in general for a neutral genome region. An excess of rare alleles has been observed, and the variability is reduced given the present number of segregating sites. Substitution rate was shown to be in the range of the values observed for other loci. Recombination seems to be of lesser impact, though present. Linkage

disequilibrium data also points to a low recombination rate on *psGBA* and neighbouring regions. We have observed a "two-pattern" model in *psGBA* (with two haplotypes at high frequency) which also applies for a much wider region encompassing *psGBA* at human 1q21 .

Locus specific phenomena: The low diversity on *psGBA* could be locus specific or due to a population event affecting the whole genome. A bottleneck or a population expansion could shape the spectrum of polymorphism into a pattern similar to that observed in *psGBA*. Once the population has gone through a reduction in the number of chromosomes, or has started expanding quickly, the mutations that are newly generated will not have time to increase in frequency and therefore, many variants at low frequency will be found, that is, $\hat{q}_p$ recovers with a slower rate than the number of segregating sites. However, the possibility of a extensive population event can be discarded when we compare *psGBA* with other nuclear loci, also studied for human worldwide samples. Population events would produce a reduced variability on all loci, while a locus-specific selective event would have only a local effect. As mentioned above, no reduction of variability has been observed in any of the nuclear loci studied up to date: *b-globin* gene (studied for a sample of 349 chromosomes from nine populations representing five continents; Harding *et al.*, 1997), a mostly intronic region of the *LPL* gene (studied for a sample of 142 chromosomes from three continents; Clark *et al.*, 1998), also a mostly intronic region of the X-linked Dys 44 gene (studied for a sample of 260 chromosomes from thirteen populations representing four continents; Zietkiewicz *et al.*, 1997; Zietkiewicz *et al.*, 1998) and an intron of the X-linked ZFX gene (studied in 336 chromosomes from fifteen populations representing four continents; Jaruzelska *et al.*, 1999). Therefore, the low variability at *psGBA* is a locus-specific phenomenon.

Fst indicates that 12.8% of the variation could be explained due to differences among populations (significantly different from zero, $p<0.001$), and suggests that population subdivision could not have had a strong influence on the *psGBA* variability pattern (R. Martínez-Arias, F. Calafell, E. Mateu, D. Comas, A. Andrés, J. Bertranpetit, unpublished results). Nevertheless, even if geographic structure would be present on the variability of *psGBA*, this would not explain the high number of alleles observed (tested through computer simulations considering a given population substructure. Results not shown).

Non-neutrality at *psGBA*, and possible genetic hitchhiking: In our sample there is a higher than expected proportion of highly polymorphic sites (Figure 2), and there are two consequences for this fact. First, the value of Tajima's *D* and Fu and Li's *D* and *F* are affected by the frequencies of the segregating sites, in the sense that the observed excess of polymorphisms at low frequencies are compensated by the excess of polymorphisms at medium frequencies, therefore making the observed statistics closer to 0 and reducing the power of the tests. Second, the fact that we observe this kind of polymorphism pattern is good evidence that natural selection is playing an important role in the evolution of *psGBA*. Only selection would yield the excess of highly polymorphic sites we observe.

Neither Tajima's *D* nor Fu and Li statistics *D* and *F* were significant. This does not allow to reject neutrality, but does not indicate absence of selection either. These tests were indeed very close to the significance level. There are several factors of the *psGBA* diversity pattern which reduce their power. First, as noted, they are affected by the reduced number of polymorphisms at *psGBA*. Second, these tests assume no recombination, but there is evidence of it on *psGBA* (even if low). And third, they may not be powerful enough to detect selection on *psGBA* with this given sample size (Simonsen *et al.*, 1995). The failure of these tests to detect selection on loci under selective pressure has been stated before (Jaruzelska *et al.*, 1999; Hamblin and Di Rienzo, 2000; Fullerton *et al*., 2000; Gilad *et al.* 2000). Nevertheless, the HKA test (Hudson *et al.*, 1987) is statistically significant when comparing *psGBA* with the *LPL* locus, indicating that both loci have evolved with a different relation of polymorphism versus divergence and, therefore, under different selective constraints. Since *LPL* presumably evolved neutrally, this is a strong indication that *psGBA* did not.

Even if Tajima and Fu and Li tests are not significant, they still give some information. It is noticeable that all these statistics are negative, which indicates an excess of variants at low frequency. Since population events have been rejected, this fact may be a consequence of the lost of variation by a selective sweep and the consequent generation of new variants at low frequency by mutation. Moreover, Fu's *Fs* was statistically significant, which is an evidence for an excess of rare alleles produced by recent mutations. While Fu and Li's *D* is more powerful than Tajima's *D* and *Fs* statistics to detect background selection, the contrary is true for detecting hitchhiking events (Fu, 1997). Also, *Fs* is not that sensitive to the low number of segregating positions as Tajima's and Fu and Li's tests. The *Fs* statistic may be affected by recombination or gene conversion events (Fu, 1997), but none of these phenomena does not seem to be very

important on the current variability (R. Martínez-Arias, F. Calafell, E. Mateu, D. Comas, A. Andrés, J. Bertranpetit, unpublished results). Since the test is highly significant, the consideration of these forces would not alter its significance. Actually, this significant result is very probably due to the high number of polymorphisms at reduced frequencies, which increases the number of observed alleles.

In summary, the possibility that *psGBA* is evolving under a neutral mutation-genetic drift model can be discarded. The results of the tests and the excess of singletons point to genetic hitchhiking on *psGBA*. Under genetic hitchhiking (or selective sweep), positive selection of an advantageous allele leads to the increase and eventual fixation of this variant together with neutral alleles on linked loci. If the sweep is complete, the neutral variation linked to the fixed allele will be entirely produced after the sweep. On the whole, polymorphism levels near the selected locus will be reduced, and new mutations on the hitchhiked tract will account for the excess of singletons observed.

Selection on linked loci and possible balancing selection: The high proportion of variants at low frequency is not the only striking feature on *psGBA*. Under genetic hitchhiking one haplotype would be selected, but two *psGBA* haplotypes have been found to be very frequent. Linkage disequilibrium results show that these two haplotypes and the two most common haplotypes for *GBA* and *PKLR* genes (16 and 86 kb upstream from *psGBA*, respectively) are linked. Thus, selection would maintain two long haplotypic groups, and a strong linkage disequilibrium would have preserved them over time, mixed at a very low extent by recombination. This fact would suggest that the selective pressure could be due to more than one different selective events or to balancing selection. In this case, the sweeping locus would be a balanced polymorphism maintained by selection, which also maintains polymorphism at linked sites.

GBA and *PKLR* are the only genes of the 1q21 region for which variability data is currently available. From these data it is not possible to know why two main haplotypes exist, since they do not have any functional meaning. Therefore, selection is not acting directly on them, but must influence this region through the effect on other linked sites.

Assuming that mutations are equally produced in expressed and non-expressed regions, and that purifying selection determines their acceptance degree, the fact that coding regions have an observed lower mutation rate than non-coding regions has to be due to a stronger purifying pressure. Since the variability in *psGBA* is lower than expected, could it be due to purifying selection in the pseudogene itself? This seems extremely unlikely, since the two transcripts that have been reported for *psGBA* are

aberrant and not translated (Sorge *et al.*, 1985; Imai *et al.*, 1993), and no regulatory mechanism has been described for *psGBA* either. The *Ka/Ks* ratio (assuming the same frameshift for the pseudogene as for the gene) in this region is less than 1 (0.626). Usually, a ratio up to 1 is considered to be evidence of neutral evolution, and a ratio higher than 1 is considered as evidence for positive selection, that in our case can be rejected.

Balancing selection is suggested by the low recombination rate and the haplotypic structure (two most frequent haplotypes present in the sample). During the time that a balancing polymorphism is maintained by selection, new mutations will tend to accumulate in the region closely linked to the selected sites. That would cause an excess of linked segregating positions, probably kept at intermediate frequencies, and a deficit (and not an excess as in our case) of singletons. This pattern would make the value of Tajima's *D* be positive, and not negative as in our case, and thus, balancing selection could be discarded. Moreover, under balancing selection, a *Ka/Ks* ratio significantly higher than 1 is expected (Nei, 1987). Under specific circumstances, such as when only specific positions are allowed to change while others have stronger constrictions, the *Ka/Ks* ratio could be close to 1 even in the case of balancing selection (Nielsen and Yang, 1998). Nevertheless, in the case of a pseudogene, a value of 0.626 makes it extremely unlikely that balancing selection is acting directly on it, although it does not give any information about its effect on neighbouring loci.

Possibility of other selective events: The effect of natural selection on positions closely linked to *psGBA* can affect the levels of nucleotide variation observed at that locus. The elimination of deleterious mutations (background selection, Charlesworth *et al.*, 1993) or the selection for advantageous mutations (hitchhiking effect, Kaplan *et al.*, 1989) affect in unique and different ways variation at linked sites. We have already briefly stated the characteristics of genetic hitchhiking.

Under background selection, neutral mutations associated with deleterious alleles in a loci will be eliminated from the population, and thus the variability in the linked area will be reduced. In some cases, background selection can produce, in the same way as hitchhiking does, an excess of variants at low frequency (Charlesworth *et al.*, 1993). In populations with low effective population size, the average time of loss of a variant does not change much under the background selection model. But the fixation time (which, if long, provides populations with variants at medium and high frequencies) is significantly reduced. In this case, the number of segregating sites at low frequency may be relatively

increased. However, an excess of medium frequency variants, as it is observed, can only be explained by positive selection on loci linked to *psGBA*, and thus, background selection can be discarded as the main force shaping variation on this locus.

In the same way, change in population size and population subdivision could as well, in certain conditions, explain an excess in low frequency variants, but they cannot account for an excess of high or medium frequent variants. Moreover, these would be phenomena observed at other loci, which is not the case.

Hitchhiking and the effect of recombination: Taking into account all the previous considerations, a hitchhiking model can best explain the observed data on *psGBA*. To explain the presence of two basic haplotypes, we can postulate that two consecutive selective sweeps have extended through a wide chromosomal area, that is, before one has had time to become fixed in the linked background, the other starts. This hypothesis, named traffic hypothesis, has been suggested for the white locus in Drosophila (Kirbi and Stephan, 1996). The two successive selective sweeps hypothesis is supported by the considerable long and gene-rich region where linkage disequilibrium has been observed, where many genes are susceptible of generating variants being selectively swept (Winfield *et al.*, 1997; Vos *et al.*, 1992; Ligtenberg *et al.*, 1990). However, if recombination is present at the *psGBA* locus, even at very reduced levels, the hitchhiking effect would be incomplete, and not all variation would be removed. If there is a recombination event previous to the end of the selective sweep, the regions contiguous to the selected locus will present two or more alleles at medium frequencies (Figure 5). This last hypothesis would explain the observed variability pattern without the requirement for a second selective event.

The haplotypes network is also in concordance with these two hypothesis, with one main haplotype from which all the unique haplotypes are descendants (haplotype A; the one probably linked from the beginning to the selected variant), and a second most frequent haplotype without a bunch of haplotypes radiating from it (haplotype B; secondarily swept either by being recombined with the selected variant, or because being linked to a new selected variant). It is interesting to notice that this last haplotype is surrounded by reticulations in the network, which are indicative of recombination or homoplasy. Nonetheless, even when this seems to be the likeliest hypothesis to explain the variability pattern on *psGBA*, we cannot fully prove it.

Last remarks: With the present data about *psGBA* and the neighbouring loci we cannot make any inference about the meaning of these two long haplotypes, preserved in more than 90 kb in human chromosome 1, or about what is being selected. Nevertheless, our survey of variation at *psGBA* points out to the effect of selection on this area. Positive selection in a nearby locus, and its effect on the surrounding loci (including *psGBA*) seems to be the most plausible hypothesis.

Fst gives additional information about the selection acting on *psGBA*. The *Fst* value found for *psGBA* (12.8%) is very similar to other values described (Fst= 15.5% for microsatellite loci, Fst= 15.6% for RFLP loci, Barbujani *et al*., 1997; Fst= 14.7% for a X-linked locus, Zietkiewicz *et al*., 1997; Fst= 19.7% for the LPL locus, Clark *et al*., 1998; Fst from 11% to 18% for autosomal markers, Jorde *et al*., 2000), and suggests that the selective pressure acting on *psGBA* is previous to the populations subdivision existing in humans: if the selection was recent and homogeneous worldwide, the value of *Fst* would be lower than the values for other loci, while if the selection was recent but geographically heterogeneous, *Fst* would be higher than for other loci.

In our previous paper we estimated a coalescence time of 200,000 years for *psGBA* haplotypes (R. Martínez-Arias, F. Calafell, E. Mateu, D. Comas, A. Andrés, J. Bertranpetit, unpublished results). For the observed alleles this time is correct. But selection has made that other alleles where lost, and thus that the coalescence time of *psGBA* was shortened. How much shortened, is something we cannot infer. On the other hand, we know that selection affects all human populations, and thus must be previous to their spread. This clearly points to that 200,000 years ago the differentiation of the ancestors of present human populations had not started.

Certainly, the understanding of the genome dynamics, which often focuses in specific genes or regions for which clear functions are known, be it either directly or through their disease-causing mutations, needs the comprehension of several additional factors, including the genetic environment (i.e., surrounding genes) and the levels of recombination in the region under study. The present study stresses the importance of the dynamics for larger regions, containing several genes or other non-coding elements, where the understanding of each of them is only possible under a general framework. A genomic approach might be needed for the understanding of the diversity of pseudogenes and other regions where selection, in general terms, is erroneously assumed to be absent.

ACKNOWLEDGEMENTS

We thank Damian Labuda and Ewa Zietkiewicz for valuable help on the phylogenetic analysis and helpful suggestions on the work. We are grateful to Kenneth M. Weiss and Anne Buchanan for technical assistance and helpful comments on an earlier version of the manuscript. We thank Kenneth K. Kidd, Judith R. Kidd, and B. Bonné-Tamir for sharing DNA samples This research was supported by Dirección General de Investigación Científica y Técnica (Spanish Government) grant PB98-1064, by Generalitat de Catalunya, Grup de Recerca Consolidat 1999SGR00009, and by Acció integrada amb Quebec DGR-CIRIT ABM/ACS/HI 1998-17. R. M-A. received a fellowship from the Spanish Ministry of Education and Culture (AP96).

REFERENCES

BANDELT, H., P. FORSTER, B.C. SYKES, and M.B. RICHARDS, 1995 Mitochondrial portraits of human populations using median networks. Genetics **141:** 743-753.

BEUTLER, E. C., C. WEST, and T. GELBART, 1992 Polymorphisms in the human glucocerebrosidase gene. Genomics **12:** 795-800.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. Genetics **140:** 783-796.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CLARK, G. A., K. M. WEISS, D. A. NICKERSON, S. L. TAYLOR, A. BUCHANAN *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595-612.

DEMINA, A., E. BOAS and E. BEUTLER, 1998 Structure and linkage relationships of the region containing the human L-type pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hematopathol. Mol. Hematol. **11:** 63-71.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87-112.

FU, Y. X., 1997 Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics **147:** 915-925.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

FULLERTON, S.M., A. G. CLARK, K. M. WEISS, D. A. NICKERSON, S. T. TAYLOR, J. H. STENGÅRD, V. SALOMAA, E. VARTIAINEN, M. PEROLA, E. BOERWINKLE, and C. F. SING, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67** (4): 881-900.

GILAD, Y., D. SEGRÉ, K. SKORECKI, M. W. NACHMAN, D. LANCET, D. SHANON, 2000 Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat. Genet. **26:** 221-224.

GLENN, D., T. GELBART and E. BEUTLER, 1994 Tight linkage of pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hum. Genet. **93:** 635-638.

HALUSHKA, M.K., J.-B. FAN, K. BENTLEY, L. HSIE, N. SHEN, *et al.*, 1999

Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239-247.

HAMBLIN, M.T., A. DI RIENZO, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66** (5): 1669-1679.

HARDING, R., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX *et al.*, 1997 Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772-789.

HEY, J., and J. WAKELEY,1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833-846.

HUDSON, R.R.,1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245-250.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

IMAI, K., M. NAKAMURA, M. YAMADA, A. ASANO, S. YOKOYAMA *et al.*, 1993A novel transcript from a pseudogene for human glucocerebrosidase in non-Gaucher disease cells. Gene **136:** 365-368.

JARUZELSKA, J., E. ZIETKIEWICZ and D. LABUDA, 1999 Is selection responsible for the low level of variation in the last intron of the ZFY locus? Mol. Biol. Evol. **16:** 1633-1640.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887-899.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KIRBI, D. A., and W. STEPHAN, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144:** 635-645.

LI, W. H., T. GOJOBORI and M. NEI, 1981 Pseudogenes as a paradigm of neutral evolution. Nature **292:** 237-239.

LI, W.H., and L. A. SADLER. 1991 Low nucleotide diversity in man. Genetics **129:** 513-523.LIGTENBERG, M. J. L., H. L. VOS, M. C. GENNISSEN and J. HILKENS, 1990Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini. J. Biol. Chem. **265:** 5573-5578.

LONG, G. L., S. WINFIELD, K. W. ADOLPH, E. I. GINNS and P. BORNSTEIN, 1996 Structure and organization of the human metaxin gene (MTX) and pseudogene. Genomics **33:** 177-184.

NACHMAN, M. W., and CROWELL S. L., 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156**: 297-304.

NEI, M.,and LI W.-H., 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Genetics **76**(10): 5369-5273.

NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

NIELSEN, R., and Z. YANG, 1998 Likelihood for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929-936.

RANA, B.K., D. HEWETT-EMMETT, L. JIN, B. H. CHANG, N. SAMBUUGHIN, *et al*.
1999 High polymorphism at the human melanocortin 1 receptor locus. Genetics **151:** 1547-1557.

RIEDER, M. J., S. L. TAYLOR, A. G. CLARK, N. A. NICKERSON, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59-62.

ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.

SCHNEIDER, S., J. M. KUEFFER, D. ROESSLI and L. EXCOFFIER, 2000 *Arlequin (ver. 2000): a software environment for the analysis of population genetics data.* Genetics and Biometry Lab., Geneva.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413-429.

SLATKIN, M.,1994 Linkage disequilibrium in growing and stable populations. Genetics **137**: 331-336.

SLATKIN, M., and L. EXCOFFIER, 1996 Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity **76:** 377-383.

SORGE, J., E. GROSS, C. WEST and E. BEUTLER, 1990 High level of transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gauchier disease. J. Clin. Invest. **86:** 1137-1141.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

VOS, H. L., S. DEVARAYALU, Y. DE VRIES and P. BORNSTEIN, 1992 Thrombospondin 3 (Thbs3), a new member of the thrombospondin gene family. J. Biol. Chem. **267:** 12192-12196.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

WHITFIELD, L.S., J. E. SULSTON, P. N. GOODFELLOW, 1995 Sequence variation of the human Y chromosome. Nature **378:**379-380.

WINFIELD, S. L., N. TAYEBI, B. M. MARTIN, E. I. GINNS and E. SIDRANSKY, 1997 Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. Genome Res. **7:** 1020-1026.

ZIETKIEWICZ, E., V. YOTOVA, M. JARNIK, M. KORAB-LASKOWSKA, K. K. KIDD *et al.*, 1997 Nuclear DNA diversity in worlwide distributed human populations. Gene **205:** 161-171.

ZIETKIEWICZ, E., V. YOTOVA, M. JARNIK, M. KORAB-LASKOWSKA, K. K. KIDD *et al.*, 1998 Genetic structure of the ancestral population of modern humans. J. Mol. Evol. **47:** 146-155.

FIGURE LEGENDS

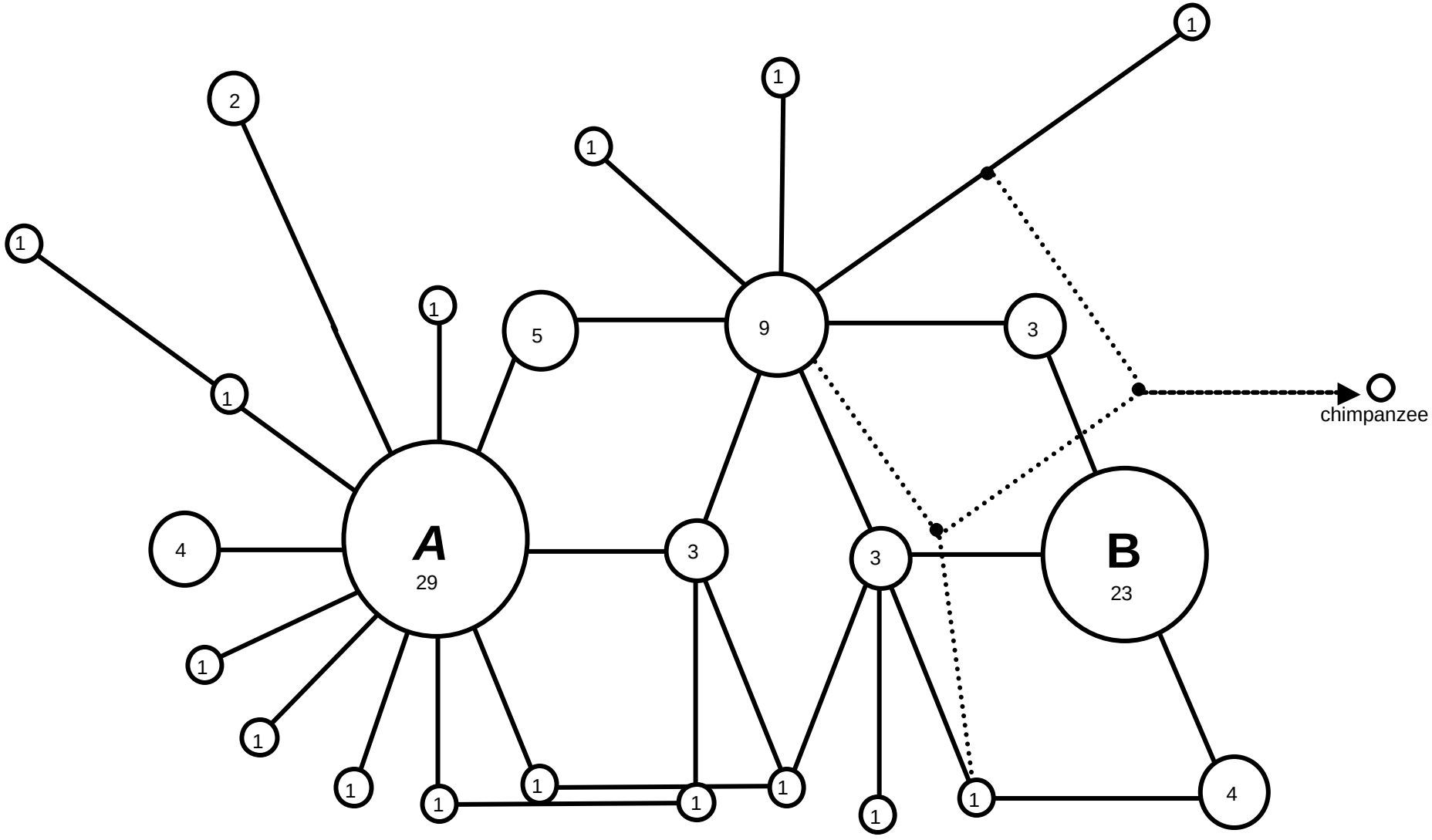
Figure 1: Median network relating the 25 *psGBA* haplotypes and the chimpanzee state. Haplotype frequency is shown inside the circles. Most frequent haplotypes, A and B, are indicated. Circle areas are proportional to the frequency of the haplotypes, and branch lengths to the circle centre are proportional to the number of mutational events they represent (one or two), except for the links with the chimpanzee haplotype. Black circles indicate haplotypes not present in the sample and necessary to link the human with the chimpanzee haplotype.

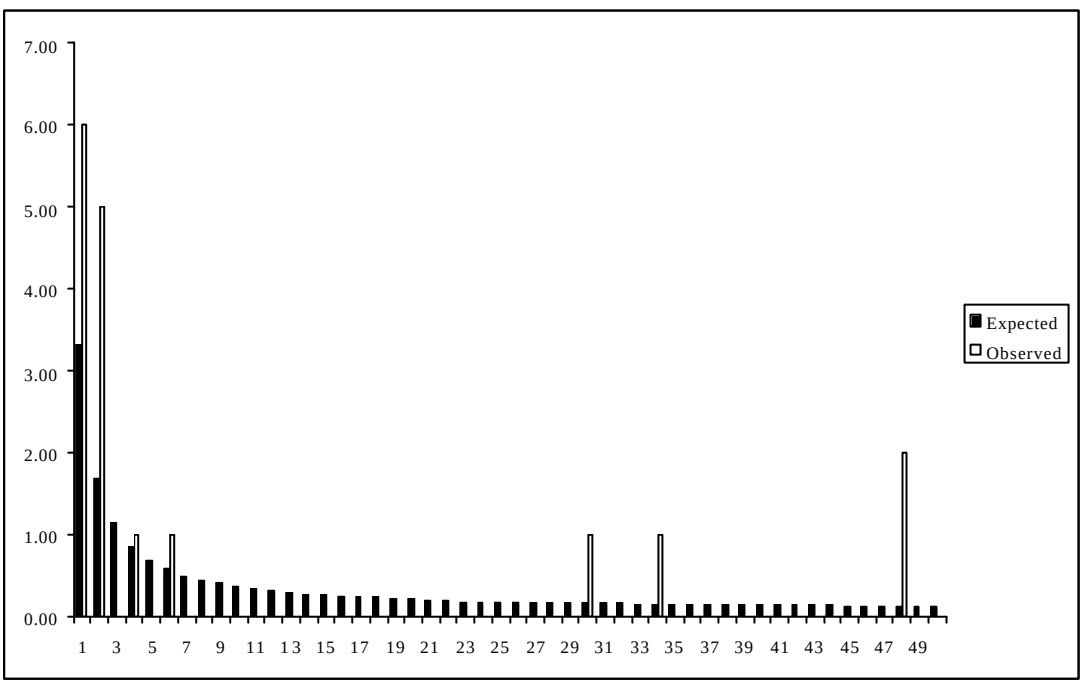
Figure 2: Frequency spectrum of polymorphisms on *psGBA*. Observed (solid bars) and expected under the neutral model (open bars) nucleotide variant frequencies are shown. Only substitutions have been taken into account. A chi-square test indicated significant differences when comparing both distributions (χ^2_{3df}=13.06, p<0.05; frequencies were grouped in singletons, doubletons, medium frequency –appearing from 4 to 30 times in the sample- and high frequency –appearing more than 30 times- variants). The less frequent variant in each site has been represented.

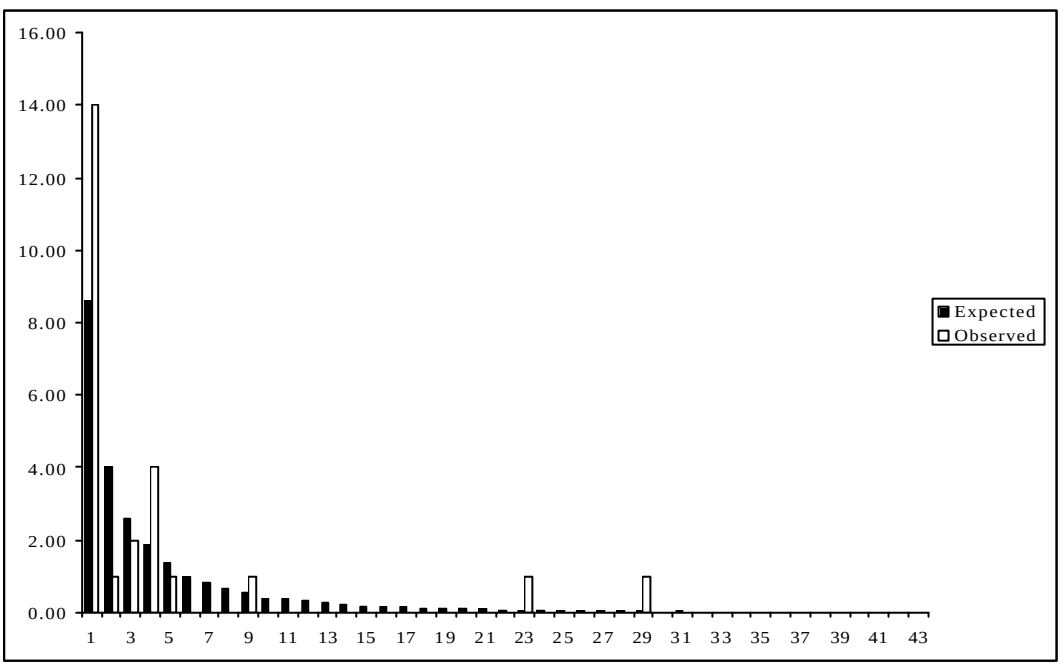
Figure 3: Frequency spectrum of alleles on *psGBA*. Observed (solid bars) and expected under the neutral model (open bars) allele frequencies are shown. The expected frequency spectrum was estimated using computer simulations, generating samples with 17 polymorphisms and a recombination rate of 16.9 (parameter $C=4N_ec$, as Hudson estimator, is required for the developed simulations). Thus, this expected distribution of allele frequencies is conditioned to 17 segregating sites and 24 alleles, with a recombination rate $C=4N_ec=16.9$.

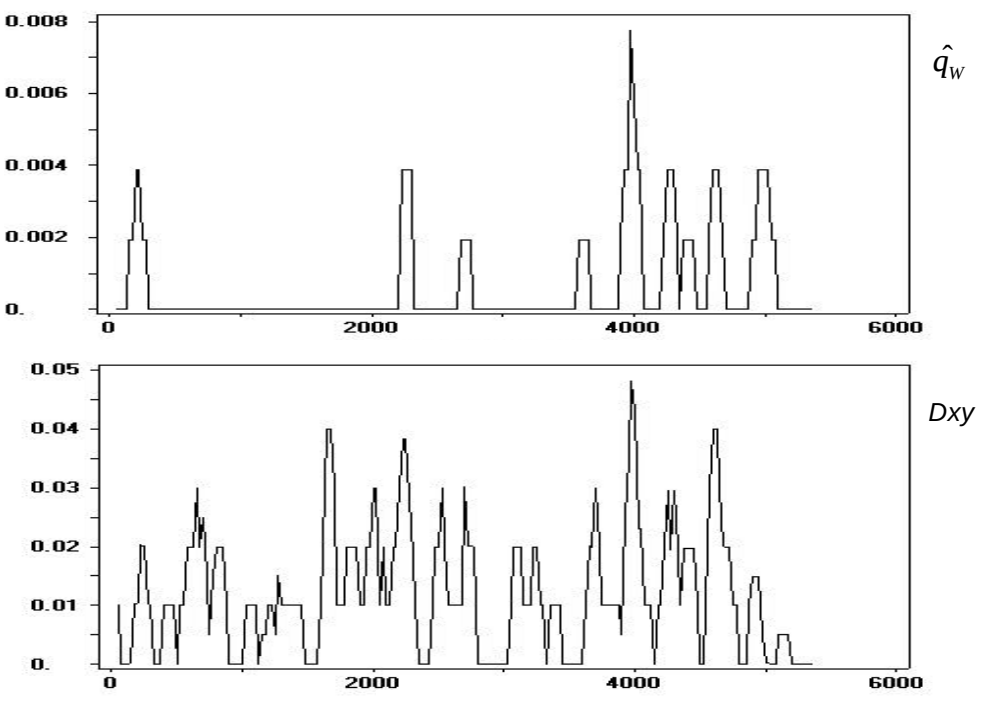
Figure 4: Nucleotide diversity as $\hat{q}_W$, and divergence (*Dxy*) along *psGBA*. A window length of 100 nucleotides moved in steps of 25 nucleotides along the sequence was used. Divergence (*Dxy*) is calculated from $Dxy=\sum_{ij} X_i\ Y_j\ d_{ij}$, where d_{ij} is the nucleotide substitutions between the *ith* haplotype from species X and the *jth* haplotype from species Y (Nei, 1987; equation 10.20).

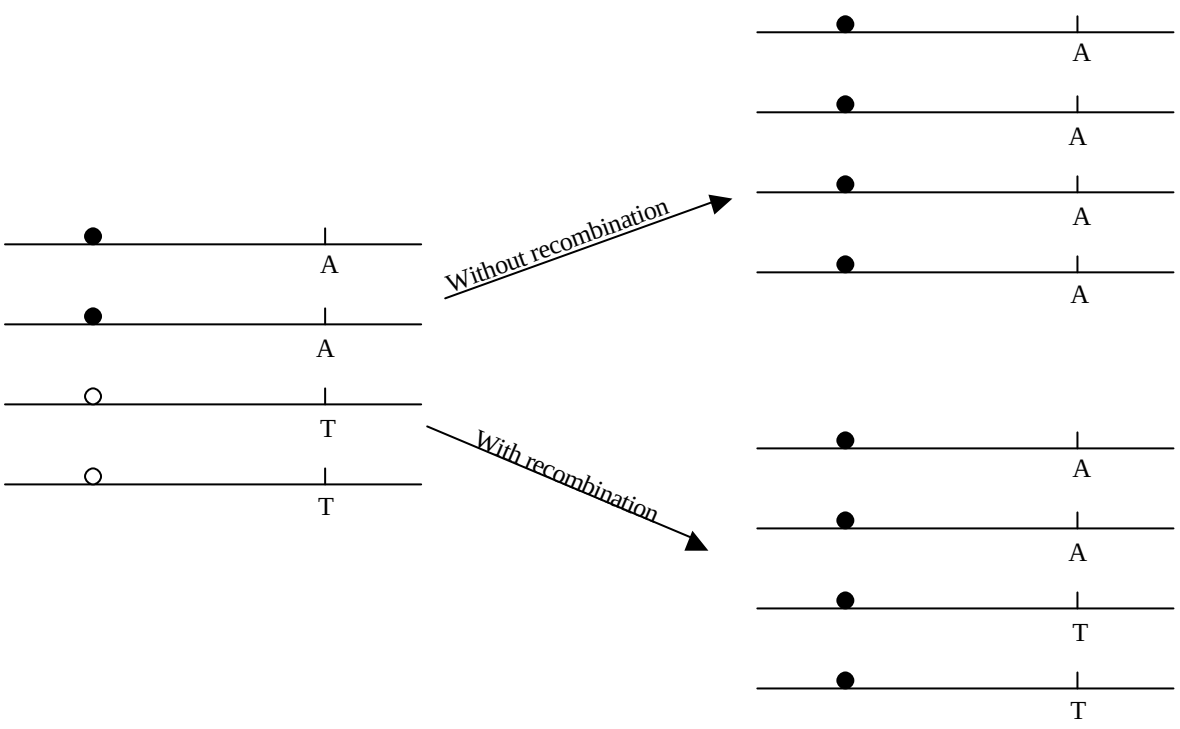
Figure 5: Effect of a recombination event previous to the end of the selective sweep. If the recombination is produced in an early stage, the regions neighbouring to the selected locus will present two or more alleles at medium frequencies.


1
2
1
1
1
5
9
3
1
chimpanzee
4
A
29
3
3
B
23
1
1
1
1
1
1
1
1
1
1
4


7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Expected
Observed


16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43
Expected
Observed


0.008
0.006
0.004
0.002
0.
0
2000
4000
6000
$\hat{q}_w$
0.05
0.04
0.03
0.02
0.01
0.
0
2000
4000
6000
Dxy
Nucleotide position
1
2
3
4
5
6
7
8
9
10
11


A
A
T
T
Without recombination
With recombination
A
A
A
A
A
A
T
T

Capítulo III:

GLUCOCEREBROSIDASE PSEUDOGENE VARIATION AND GAUCHER DISEASE: RECOGNISING PSEUDOGENE TRACTS IN GBA ALLELES

Rosa Martínez-Arias, David Comas, Eva Mateu, Jaume Bertranpetit

Consultable en:

Martinez-Arias R, Comas D, Mateu E, Bertranpetit J. “Glucocerebrosidase pseudogene variation and Gaucher disease: Recognizing pseudogene tracts in GBA alleles”. ***Human Mutation***. 2001 Mar;17(3):191-8

Capítulo IV:

PROFILES OF ACCEPTED MUTATION: FROM NEUTRALITY IN A PSEUDOGENE TO DISEASE-CAUSING MUTATION ON ITS HOMOLOGOUS GENE

Rosa Martínez-Arias, Eva Mateu, Jaume Bertranpetit, Francesc Calafell

Publicado en:

Martinez-Arias R, Mateu E, Bertranpetit J, Calafell F. “Profiles of accepted mutation: from neutrality in a pseudogene to disease-causing mutation on its homologous gene”. ***Human Genetics*** . 2001 Jul;109(1):7-10, © 2001 by Springer Verlag

PROFILES OF ACCEPTED MUTATION: FROM NEUTRALITY IN A PSEUDOGENE TO DISEASE-CAUSING MUTATION ON ITS HOMOLOGOUS GENE

Rosa Martínez-Arias, Eva Mateu, Jaume Bertranpetit, Francesc Calafell

Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, 08003 Barcelona, Spain

KEY WORDS: mutation profiles, neutral mutation, accepted mutation, glucocerebrosidase gene, glucocerebrosidase pseudogene.

CORRESPONDING AUTHOR:
Francesc Calafell
Unitat de Biologia Evolutiva
Universitat Pompeu Fabra
Dr Aiguader, num. 80
08003, Barcelona
Spain
Telephone number: 34-93-542 28 41
FAX number: 34-93-542 28 02
e-mail: francesc.calafell@cexs.upf.es

ABSTRACT

We have compared the mutational pattern of the glucocerebrosidase gene (GBA) and the glucocerebrosidase pseudogene (psGBA), two highly homologous regions under different selective pressures and within the same genomic background. Mutations on GBA may lead to Gaucher disease, an inborn metabolic disorder. Disease-causing mutations and neutral variation in the gene have been compared to neutral variation in the pseudogene. This comparison offers a unique opportunity to better understand the action of purifying selection, since the differences between mutational patterns can be attributed to different selective pressures. A similar frequency of CpG dinucleotides was observed in GBA than in psGBA, and CpG pairs were mutated with the same high frequency in both regions. However, nucleotides not in CpG pairs were more likely to contribute to disease-causing mutation than to accepted polymorphisms. This pattern, which resulted in a higher transition to transversion ratio in the pseudogene, may be due to CpG avoidance on critical gene regions.

INTRODUCTION

Mutational processes are basic to understand the genetic variation and the evolution of the genome. Mutation is not a pure random change on DNA. Many factors influence such a "blind" process, such as DNA context, especially the flanking bases (as it is the case in the hypermutable CpG dinucleotides), and a complex substitution probability matrix (Gojobori et al., 1982; Li et al., 1984). Selection-related constraints as, for instance, changes inducing aminoacid substitutions, are also a determinant point. As for functional regions, there is a discrepancy between on one hand, the chemical changes that result in mutations on DNA, and on the other hand, what can actually be detected, since only "accepted mutations", those that have not been wiped out by purifying selection, can be observed as fixed substitutions or at polymorphic frequencies. Thus, the mutational process is modulated by a large number of effects beyond the mere nucleotide change.

Paralogous genomic regions under different selective pressures may help to understand the mutational process not only in its production stage, but also on its acceptance and detectability. An interesting way to examine this is to analyse the pattern of substitutions on a gene and a non-functional copy, both located within the same genomic background. In this report, we focus on the different mutational processes that have been acting on the glucocerebrosidase gene (GBA; GenBank J03059) and the glucocerebrosidase pseudogene

(psGBA; GenBank AF267177), a 96% homologous non-functional duplicate. GBA consists on 11 exons which spread along 8.8 kilobases (kb) on human 1q21 chromosome. psGBA is 5.7 kb long and it is located 16 kb downstream from GBA (Ginns et al., 1985; Zimran et al., 1990; Winfield et al., 1997). The difference in length is due to several Alu insertions on four GBA introns (Horowitz et al., 1989). GBA codes for the glucocerebrosidase enzyme (EC.3.2.1.45), a misfunction of which leads to Gaucher disease, the most prevalent lipid storage disorder in humans (OMIM 230800, OMIM 230900, OMIM 231000). We examine the proportion of each type of transition and transversion, the nucleotide composition, and the presence of CpG mutational hotspots on both functional and non-functional loci. Eventually, we expect to characterise the patterns of accepted (polymorphisms on GBA), selectively wiped out (disease-causing mutations on GBA), and total (neutral variation on psGBA) mutational events.

MATERIALS AND METHODS

GBA Gaucher-causing mutations were extracted from the Human Gene Mutation Database (HGMD 119262, December 2000). New mutations in GBA that produce Gaucher disorder are continuously being described. At present, 138 mutations have been reported in the HGMD: 120 substitutions (115 causing aminoacid replacement, and 5 altering splicing), 12 small indels, 2 gross deletions, and 4 complex alleles produced by recombination or gene conversion events with psGBA. We have analysed the pattern of disease-causing nucleotide substitutions by considering as ancestral the normal allele.

Polymorphisms at the GBA gene: A total of 16 polymorphisms have been reported in the literature for the GBA gene; all but one of those are nucleotide substitutions. Beutler et al. (1992) described 12 of these polymorphisms: an insertion-deletion in intron 7 and 11 nucleotide substitutions (one of these with three different nucleotide states described, and which we have treated as two polymorphisms with two different substitution events). Eight of these polymorphisms are located in introns and three in the 5' GBA flanking region. All these polymorphisms are in marked linkage disequilibrium with each other, and define two common haplotypes, designed as Pv1.1- and Pv1.1+, with frequencies of around 70% and 30% in Caucasoid population, respectively, and with approximately inverted frequencies in East Asian and African populations (Beutler et al., 1992; Glenn et al., 1994; Mateu et al., unpublished data). Additional polymorphisms were found at position 1033 of exon 2 (Horowitz

et al. 1993), and at position 4586 of intron 7 (Amaral et al., 1997; Lau et al., 1999). We also included in the analysis a polymorphism found in position 6416 of GBA exon 10 (E. Beutler, personal communication).

For 12 of the 15 polymorphic single nucleotide variants, the direction of the nucleotide changes has been determined assuming as ancestral the state in the chimpanzee GBA. In those cases in which the chimpanzee GBA sequence was not determined (i.e., polymorphisms at positions -802, -725, -614) we assumed as ancestral the nucleotide state which is most common in African populations (namely, haplotype Pv1.1+) (Mateu et al., unpublished data). A number of studies have shown that African populations have an ancient origin and tend to preserve ancestral human states (Vigilant et al., 1991; Harding et al.,1997; Hey, 1997; Underhill et al., 1997; Zietkiewicz et al., 1997; Clark et al., 1998; Shen et al., 2000).

psGBA polymorphisms: psGBA variability data on humans and chimpanzee was extracted from Martínez-Arias et al. (submitted) (GenBank AF267177; GenBank AF272642). Among 100 human chromosomes, 19 variable sites were found on the psGBA locus (Martínez-Arias et al., submitted; GenBank AF267177): a three-nucleotide deletion, a polymorphic polyA tract, and 17 single nucleotide substitutions. To determine the direction of the substitutions, the ancestral state for human psGBA was assumed to be the state at each polymorphic position that would lead to the most parsimonious psGBA phylogeny. Ancestral states at the polymorphisms match those in the chimpanzee psGBA (GenBank 272642), except for four positions (Martínez- Arias et al., submitted).

RESULTS AND DISCUSSION

Variability distribution

The nucleotide substitutions leading to 120 Gaucher-causing mutations, to 15 single nucleotide polymorphisms in the GBA gene and to 17 single nucleotide polymorphisms in the GBA pseudogene are summarised in Table 1. The higher number of disease-causing mutations can be attributed to ascertainment bias, since mutations were described from the screening of Gaucher patients. Moreover, comparison to chimpanzee GBA and psGBA sequences has yielded estimated substitution rates of $0.87 \pm 0.11 \times 10^{-9}$ and $1.23 \pm 0.22 \times 10^{-9}$ per nucleotide and year for GBA and psGBA, respectively. Obviously, the higher number of

Gaucher-causing mutations at GBA is a product of ascertainment bias and not of a faster mutation rate.

Out of 120 disease-causing nucleotide substitutions in GBA, 115 resulted in an aminoacid change. Divided by their position in the codon, 48% fell in the first position, 46% in the second position, and 6% in the third. Those frequencies do not different significantly (χ^2_{2df} = 4.61, p=0.1) from those expected if non-synonymous changes were caused by mutations happening at random and independently of nucleotide position (i.e., 42.6%, 44.8% and 12.6%; Nei, 1987).

CpG hypermutability and suppression

CpG dinucleotides have long been recognised as mutation hotspots, changing approximately ten times faster than any other dinucleotide in the nuclear genome (Cooper and Krawczak, 1990). A marked underrepresation of CpG dinucleotides relative to the expected frequencies based on nucleotide frequencies has been detected in vertebrate genomes. On coding regions, CpG pairs are found at an average 37% of the expected frequency, while on non-coding regions, CpG dinucleotides reach 20-25% of their expected frequency (Cooper and Gerber-Huber, 1985; Cooper and Krawczak 1989; Cooper and Krawczak, 1990; Schorderet and Gartler, 1992; Deinard et al., 1999). The GBA gene contains 151 CpG pairs in 8850 bp (52 in exons and 99 in introns), while its pseudogene harbours 80 CpG pairs in 5420 bp. Corrected by sequence length, the CpG contents are not different between gene and pseudogene (χ^2_{1df} = 1.12, p=0.291). The expected frequencies for CpG dinucleotides based on C and G frequencies are 615 for GBA and 388 for psGBA; thus, the actual CpG frequencies are respectively 24.5% and 20.6% of the expected values, a highly significant decrease in both cases (χ^2_{2df} = 376.1 and χ^2_{1df} = 211.8 respectively, both with p≅0). The level of CpG suppression in GBA exons is stronger than the coding-region genome average (29.2% in GBA vs. average 37%, χ^2_{1df} = 4.63, p=0.03), while CpG suppression in psGBA and in GBA introns (22.7%) falls within the observed range for non-coding regions.

Regarding the mutability of CpG dinucleotides, 14 out of the 151 (9.3%) CpG pairs in GBA are found to have mutated, either to disease-causing changes (10 out of 14) or to polymorphisms (4 of 14). The proportion of variable CpG pairs is similar in psGBA: 7 out of 80 (8.8%) CpG pairs have resulted in a segregating position. Nucleotides not in a CpG pair are clearly less likely to be substituted: 106 out of 8548 nucleotides (1.24%) in GBA not in a CpG

are found to be variable, while this fraction is 10 out of 5260 nucleotides (0.19%) in psGBA. Given that CpG content is similar between GBA and psGBA, the contribution of mutations in or out of a CpG context in both regions can give a measure of the acceptability of both types of mutations. Ten out of 120 (8.3%) disease-causing mutations are in a CpG context, while that is the case for 7 out of 17 (41.2%) psGBA polymorphisms. The difference is highly significant (χ^2_{1df}= 14.8, p≅0.0001), and may imply that CpG dinucleotides are selectively avoided in the critical gene region.

Transitions and transversions

Table 1 shows nucleotide substitutions in Gaucher-causing disease in GBA, GBA polymorphisms and psGBA polymorphisms, subdivided by nucleotide change. If mutation were a merely random change, 33% of all nucleotide substitutions would be transitions. The average observed value across the genome is 60% (Gojobori et al., 1982; Li et al., 1984; Cooper and Krawczak, 1990). 54.2% of the nucleotide substitutions in GBA that cause disease are transitions; this proportion is 73.3% for GBA polymorphisms and 88.2% for psGBA polymorphisms. There are significantly more transitions in psGBA than in GBA disease-causing mutations (χ^2_{1df} = 7.11, p=0.008), but the other two pairwise comparisons (i.e., GBA disease mutations vs. GBA polymorphisms and GBA polymorphisms vs. psGBA polymorphisms) were not statistically significant.

Transitions have been described to be biased towards C to T and G to A changes (Gojobori et al., 1982; Li et al., 1984; Deinard et al., 1999), partly because mutation in hypermutable CpG pairs is usually either C to T or G to A. In our dataset, these two types of changes account for 73.8% of the disease-causing transitions, for 81.8% of the GBA transition polymorphisms, and for 80% of the psGBA transitions. Mutations within CpG pairs explain, respectively, 20.8%, 44.4% and 58.3% of these two types of changes. Again, the difference is significantly different only for GBA disease-causing mutations vs. psGBA polymorphisms (χ^2_{1df} = 6.65, p=0.01).

Overall, the pattern of transitions and transversions in psGBA, and of the types of transitions, seem to be due to a relative preponderance of mutations in CpG pairs. However, as discussed above, both the frequency of CpG pairs and the fraction of those that have been found to be variable is similar among GBA and psGBA. Thus, the difference in mutation

patterns seems to be due to an overrepresentation of non-CpG mutations in GBA rather than to a lack of mutated CpG.

Nucleotide composition

Studies on non-coding regions have shown that spontaneous mutations result more often in A or T than on C or G, partly because of CpG mutation, and leading to a A+T average proportion of 57% (Gojobori et al., 1982; Li et al., 1984), as compared to 50% on coding regions, which, given their functional constraints, are not free to accept all spontaneous mutations. Contrary to those expectations, GBA is slightly richer in A+T than psGBA (47.2% in GBA and 46.5% for psGBA). Two factors can explain this observation: i) GBA contains six Alu insertions not present in psGBA; when those are excluded from the count, the A+T content of GBA drops to 46.2%; and ii) given their physical proximity and high homology, up to 13% of the psGBA sequence bears the traces of gene conversion events from GBA (Martínez-Arias et al., submitted). Gene conversion from GBA to the non-coding psGBA is not selectively disadvantageous and homogenises psGBA with GBA. Moreover, the maintenance of a high C+G content even on non-functional tracts of this region (1q21) may be correlated with its richness in functional genes (Saccone et al., 1999; Bernardi et al., 2000).

In summary, we have compared the pattern of nucleotide substitutions that cause disease (and that are, thus, sieved out by selection and on their way to extinction or mutation-selection equilibrium) with accepted, polymorphic nucleotide substitutions in the same gene and in an adjacent non-functional duplicate. We have found that non-CpG mutations are overrepresented among non-accepted mutations, as if selection had already excluded many (but not all) CpG pairs from those parts of the gene sequence were their mutation would be more functionally disruptive.

ACKNOWLEDGEMENTS

We thank Kenneth K. Kidd, Judith R. Kidd, and B. Bonné-Tamir for sharing DNA samples. This research was supported by Dirección General de Investigación Científica y Técnica (Spanish Government) grant PB98-1064, and by Generalitat de Catalunya, Grup de Recerca Consolidat 1998SGR00009. R. M-A. received a fellowship from the Spanish Ministry of Education and Culture (AP96).

REFERENCES

Amaral O, Marcao A, Pinto E, Zimran A Miranda SMC (1997) Distinct haplotype in Non-Ashkenazi Gaucher patients with N370S mutation. Blood Cells Mol Dis 23 (22): 415-416

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

Beutler E, West C, Gelbart T (1992) Polymorphisms in the human glucocerebrosidase gene. Genomics 12: 795-800

Clark GA, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, and Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595- 612

Cooper DN, Gerber-Huber S (1985) DNA methylation and CpG suppression. Cell Differentiation 17: 199-205

Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83: 181-188

Cooper DN, Krawczak M (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: paterns and predictions. Hum Genet 85: 55-74

Deinard A, Dorit R, Castiglione C, Jiang Z, Becker D, Ruddle F, Schugart K, Kidd K (1999) Evolution of the HOXB6 intergenic region: motif conservation at the lateral plate mesoderm (LPM) enhancer element. J Exp Zool 285: 170-176

Ginns E, Choudary PV, Tsuji S, Martin B, Stubblefield B, Sawyer J, Hozier J, Barranger J (1985) Gene mapping and leader polypeptide sequence of human glucocerebrosidase : implications for Gaucher disease. Proc Natl Acad Sci Usa 82: 7101-7105

Glenn D, Gelbart T, Beutler E (1994) Tight linkage of pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hum Genet 93: 635-638

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18: 360-369

Harding, RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, and Clegg JB (1997) Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans. Am J Hum Genet 60: 772- 789

Hey J (1997) Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol Biol Evol 14(2): 166-172

Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, Beutler E (1989) The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 4: 87-96.

Karlin S, Mrázek J. 1996. What drives codon choices in human genes?. J Mol Biol 262: 459-472

Horowitz M, Tzuri G, Eyal N, Berebi A, Kolodny EH, Brady RO, Barton NW, Abrahamov A, Zimran A (1993) Prevalence of nine mutations among Jewish and non-Jewish Gaucher disease patients. Am J Hum Genet 53: 921-930

Lau EK, Tayebi N, Ingraham LJ, Winfield SL, Koprivica V, Stone DL, Zimran A, Ginns EI, Sidransky E (1999) Two novel polymorphic sequences in the glucocerebrosidase gene region enhance mutational screening and founder effect studies of patients with Gaucher disease. Hum Genet 104: 293-300

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21: 58-71

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York:

Saccone S, Federico C, Solovei I, Croquette M-F, Della Valle G, Bernardi G (1999) Identification of the gene-richest bands in human prometaphase chromosomes. Chromosome Res. 7: 379-386

Schorderet D, Gartler SM (1992) Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA 89: 957-961

Shen P, Wang F, Underhill PA, Franco C, Yang W-H, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, David RW, Cavalli-Sforza LL, and Oefner PJ (2000) Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA 97: 7354- 7359

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, and Oefner PJ (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7: 996- 1005

Vigilant L, Stoneking M, Harpending H, Hawkes K, and Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503- 1507

Winfield SL, Tayebi N, Martin BM, Ginns EI, Sidransky E (1997) Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. Genome Res 7: 1020-1026

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M, and Labuda D (1997) Nuclear DNA diversity in worldwide distributed human populations. Gene 205: 161- 171

Zimran A, Sorge J, Gross E, Kubitz M, West C, Beutler E (1990) A glucocerebrosidase fusion gene in Gaucher disease. J Clin Invest 85: 219-222

Table 1: Summary of the single nucleotide substitutions found in GBA gene and pseudogene. Substitutions due to metilated CpG dinucleotide (mCpG) are indicated in the last row. The GBA thriallelic polymorphism (G-C, G-A) in position 2834 has been counted as two different changes, G to C and G to A.

Substitution	GBA disease-causing mutations		GBA polymorphisms		psGBA polymorphisms	
Transitions						
C→T	23	19.2%	3	20%	8	47.0%
T→C	9	7.5%	-	-	2	11.7%
A→G	8	6.6%	2	13.3%	1	5.8%
G→A	25	20.8%	6	40%	4	23.5%
Total transitions	65	54.2%	11	73.3%	15	88.2%
Transversions						
T→G	11	9.2%			1	5.8%
G→T	10	8.3%				
T→A	4	3.3%				
A→T	4	3.3%	1	6.6%		
C→G	5	4.2%	1	6.6%	1	5.8%
G→C	11	9.2%	1	6.6%		
C→A	5	4.2%				
A→C	5	4.2%	1	6.6%		
Total transversions	55	45.8%	4	26.6%	2	11.7%
Total substitutions	120	100%	15	100%	17	100%
mCpG	10	8.3%	4	26.6%	7	41.2%

Capítulo V:
ANÁLISIS POBLACIONALES

El propósito de este capítulo es caracterizar la variabilidad de psGBA dentro de cada una de las poblaciones del estudio, así como determinar la estructura poblacional de dicha variabilidad.

Algunos parámetros que nos informan sobre determinadas características relacionadas con los distintos grupos poblacionales ya se han presentado en los capítulos anteriores. Por ejemplo, el estadístico Fst apunta a que la mayor parte de la diversidad en psGBA se debe a diferencias dentro de las poblaciones y no a diferencias entre ellas. Mediante los análisis de coalescencia hemos visto que el haplotipo humano más antiguo es particular de la población biaka, mientras que no parece haber una estructuración geográfica entre el resto de haplotipos, al menos entre aquellos que siguen el modelo de infinitos lugares (que son los que hemos utilizado para aplicar el modelo de coalescencia). También hemos demostrado que el patrón de baja diversidad encontrado en psGBA no se debe a un fenómeno poblacional, sino que es específico de *locus*.

Los análisis que presentamos en este apartado aportan información más detallada sobre los parámetros de diversidad y sobre la distribución de los haplotipos psGBA en cada población en particular. La comparación entre todos los cromosomas de una población con todos los cromosomas del resto de poblaciones se hará a través del cálculo de la matriz de distancias, que visualizaremos mediante el análisis de coordenadas principales. La comparación de las frecuencias haplotípicas en cada población se llevará a cabo mediante el análisis de componentes principales. Ambos análisis nos mostrarán relaciones interpoblacionales y por tanto pondrán de manifiesto, si es que existen, patrones geográficos en la variabilidad de psGBA.

La distribución de haplotipos en las poblaciones, a la que haremos referencia en los siguientes análisis, se halla en la Tabla 2 del capítulo I de Resultados.

A continuación mostramos algunos parámetros de diversidad en cada población (Tabla 3), el espectro de frecuencias haplotípicas por población (Figura 7), y la distribución de alelos por población dentro del *median joining network* entre haplotipos (Figura 8).

Tabla 3: Parámetros de diversidad en las distintas poblaciones de este estudio. Se han analizado 10 cromosomas en cada una de las poblaciones. Hap: número de haplotipos distintos; S: número de posiciones segregantes (sólo se han considerado sustituciones nucleotídicas); H: diversidad haplotípica; pi: diversidad nucleotídica por nucleótido; k: número medio de diferencias nucleotídicas; $\hat{q}_w$: estimador de theta según el número de sitios segregantes (Watterson, 1975); estadísticos D* y F* (Fu y Li, 1993); estadístico D (Tajima, 1989). Ninguno de los tres últimos estadísticos es significativamente distinto de 0 en ningún grupo. Como africanos se agrupan tanzanos y biaka, como europeos se agrupan catalanes y vascos, y como asiáticos se agrupan chinos y yakut.

	biaka	tanzanos	saharauis	vascos	catalanes	drusos	yakut	chinos	maya	nasioi	africanos	europeos	asiáticos	Total
Hap	5	6	5	6	5	4	4	5	4	5	9	9	7	25
S	5	6	5	7	5	7	7	6	4	4	9	8	8	17
H	0,800	0,844	0,822	0,844	0,822	0,800	0,644	0,844	0,733	0,844	0,874	0,837	0,753	0,853
Pi	0,00028	0,00043	0,00035	0,00050	0,00037	0,00055	0,00045	0,00043	0,00023	0,00025	0,00037	0,00044	0,00042	0.00044
k	1,511	2,356	1,911	2,756	2,022	3,000	2,467	2,356	1,298	1,400	2,000	2,395	2,321	2,396
$\hat{q}_w$	1,767	2,121	1,767	2,474	1,767	2,474	2,474	2,121	1,414	1,414	2,537	2,255	2,225	3,284
D*	-0,686	0,775	0,638	-0,126	0,638	0,376	-0,126	0,204	-0,338	-0,338	-0,686	-0,919	-0,354	-1,336
F*	-0,740	0,782	0,629	0,026	0,698	0,567	-0,110	0,298	-0,379	-0,296	-0,808	-0,692	-0,260	-1,343
D	-0,581	0,457	0,326	0,481	0,578	0,899	-0,013	0,457	-0,339	-0,037	-0,727	0,208	0,098	-0,764

Figura 7. Espectro de frecuencias haplotípicas para las diez poblaciones del estudio. k: para cada población se ha representado el número de haplotipos encontrados k veces (por ejemplo: en vascos se ha encontrado un haplotipo único cuatro veces, un haplotipo presente dos veces, y un haplotipo presente cuatro veces).

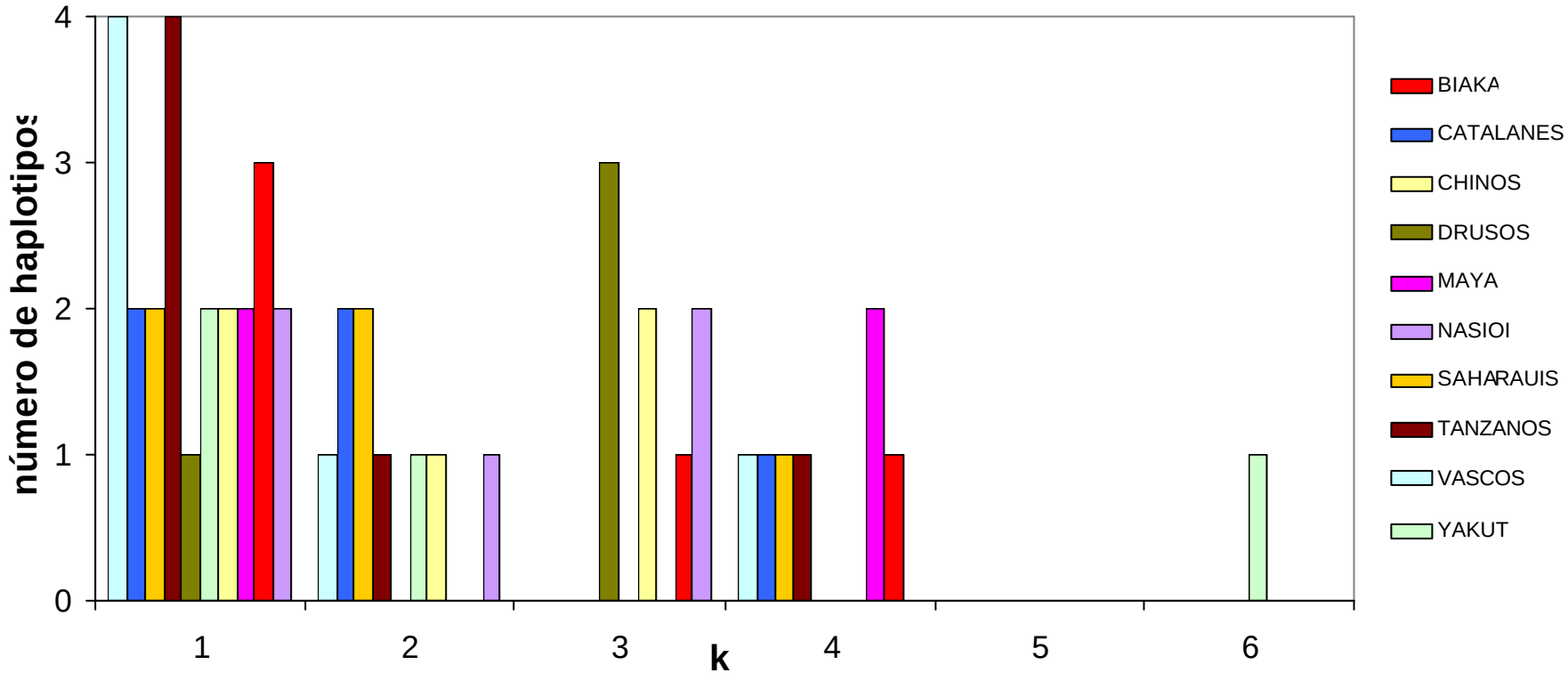
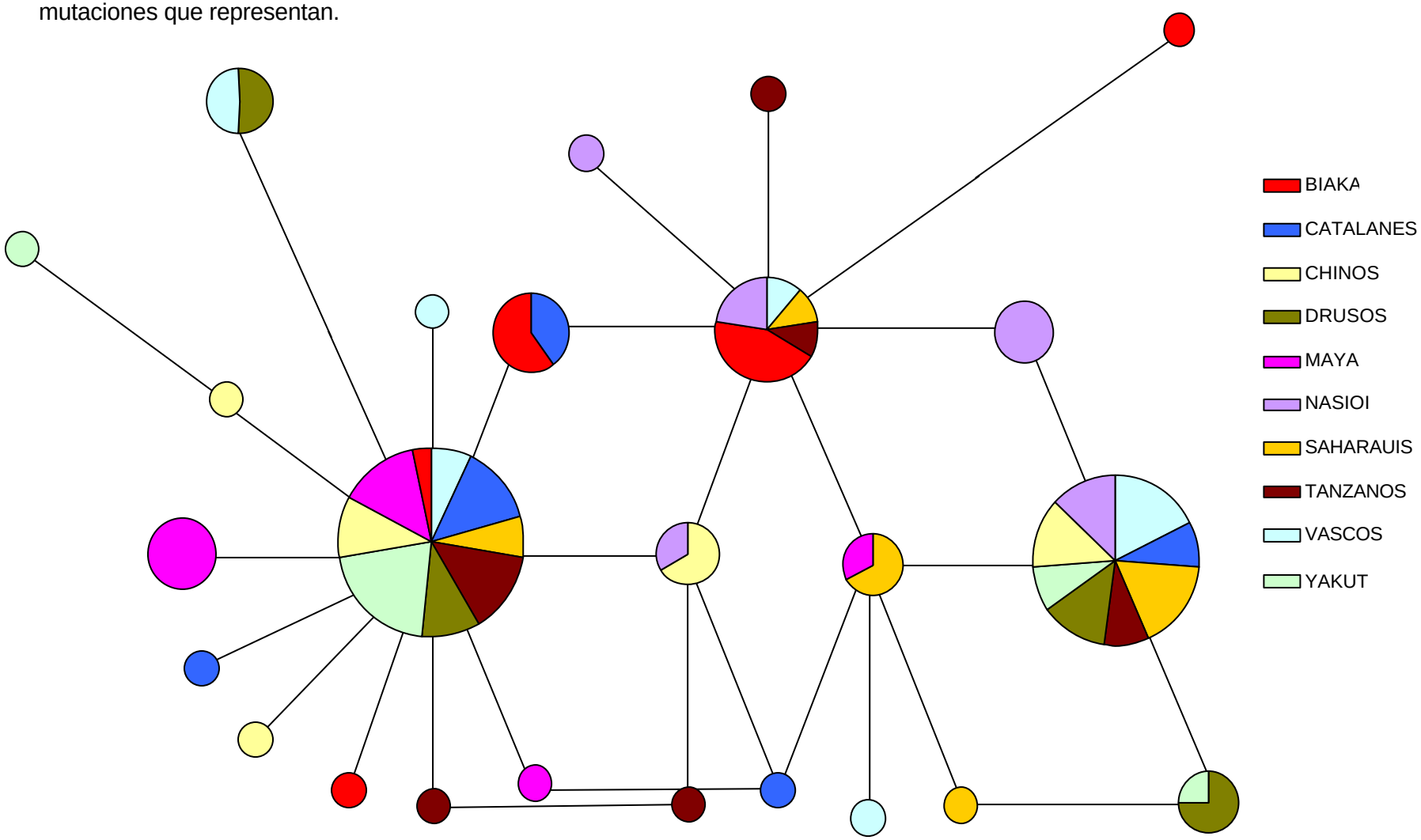


Figura 8. *Median joining network* (MJN) entre los 25 haplotipos humanos, con la distribución haplotípica en las distintas poblaciones del estudio. La representación del MJN sigue las mismas características que el representado en el capítulo I de Resultados. El área del círculo es proporcional a la frecuencia del haplotipo y la longitud de las ramas que unen dos haplotipos es proporcional al número de mutaciones que representan.



Análisis de coordenadas principales

De la representación en el plano de la matriz de distancias genéticas entre poblaciones (Tabla 4), obtenemos el gráfico de coordenadas principales (Figura 9). Este gráfico representa en el plano las distancias entre poblaciones del modo más similar posible a las distancias genéticas reales, facilitando así la interpretación de los datos de la matriz.

La primera coordenada principal explica el 62,5% de la variancia total, y la segunda coordenada principal explica el 21,4% de la variancia total. Las poblaciones más destacadas por su separación en el gráfico son las poblaciones nasioi y maya. Esta separación puede ser debida a que son las dos únicas poblaciones con haplotipos privados presentes más de una vez. La población maya presenta como haplotipos privados el haplotipo 4 (en cuatro cromosomas) y el haplotipo 5 (en un cromosoma). La población nasioi presenta como haplotipos privados el haplotipo 26 (en tres cromosomas) y el haplotipo 27 (en un cromosoma). A la separación de nasioi y maya puede también contribuir el hecho de que son las dos poblaciones con menor diversidad nucleotídica interna (Tabla 3), ya que en la matriz de distancias genéticas las distancias entre poblaciones se corrigen por este factor.

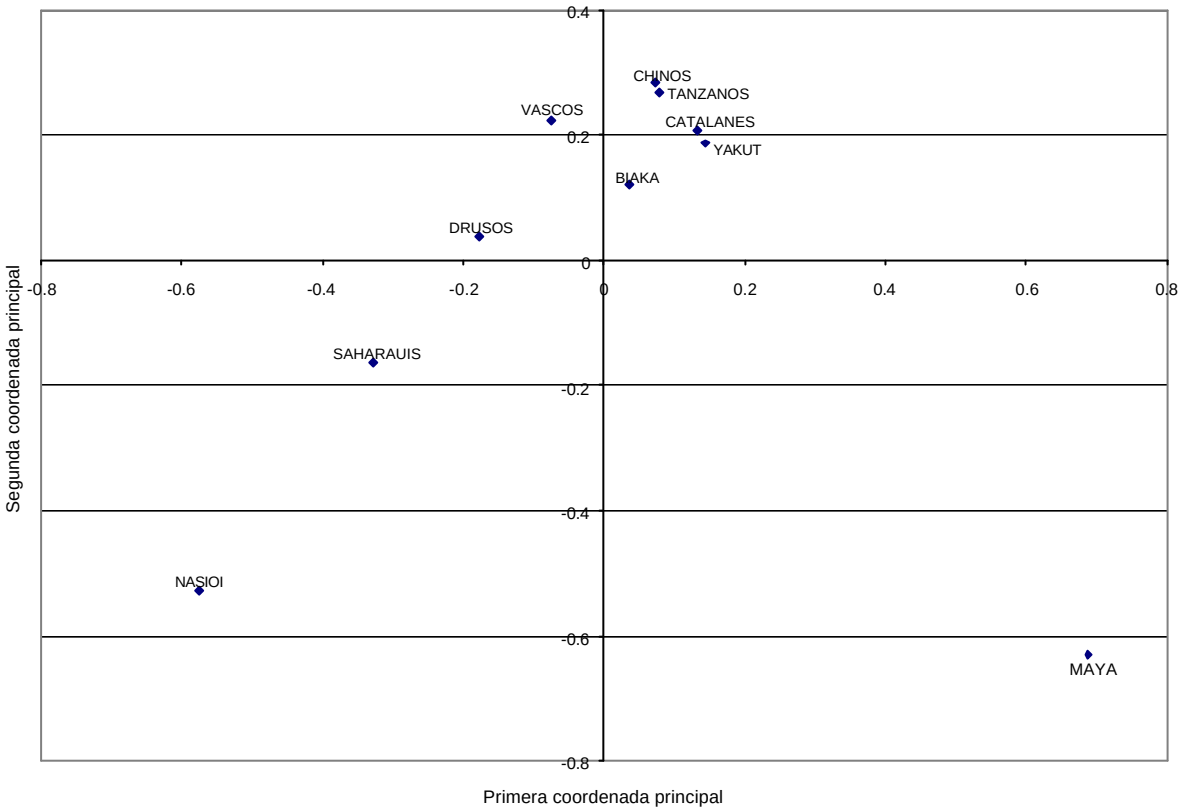
Con una distancia menor, se separan del resto de poblaciones los grupos de drusos y saharauis.

El resto de las poblaciones se agrupan sin especial estructuración.

Tabla 4. Matriz de distancias *intermatch*/*mismatch* entre poblaciones. Sobre la diagonal: número medio de diferencias de *pairwise* entre poblaciones. En la diagonal: número medio de diferencias de *pairwise* dentro de cada población. Por debajo de la diagonal: media corregida de las diferencias de *pairwise*, es decir, distancia *intermatch*/*mismatch*.

	vascos	catalanes	saharauis	tanzanos	drusos	yakut	chinos	maya	biaka	nasioi
vascos	**3.355**	*2.700*	*2.560*	*2.860*	*3.040*	*2.940*	*2.800*	*3.060*	*2.740*	*2.540*
catalanes	0.011	**2.022**	*2.380*	*2.080*	*2.700*	*2.060*	*2.060*	*1.840*	*1.960*	*2.460*
Saharauis	-0.073	0.413	**1.911**	*2.520*	*2.400*	*2.660*	*2.460*	*3.000*	*2.320*	*1.760*
tanzanos	0.004	-0.108	0.386	**2.355**	*2.940*	*2.320*	*2.220*	*2.140*	*2.060*	*2.420*
drusos	-0.137	0.188	-0.055	0.262	**3.000**	*2.900*	*2.820*	*3.220*	*2.940*	*2.440*
yakut	0.028	-0.184	0.471	-0.091	0.166	**2.466**	*2.240*	*2.060*	*2.300*	*2.720*
chinos	-0.055	-0.128	0.326	-0.135	0.142	-0.171	**2.355**	*2.120*	*2.220*	*2.420*
maya	0.737	0.184	1.400	0.317	1.075	0.182	0.297	**1.288**	*2.140*	*3.200*
biaka	0.306	0.193	0.608	0.126	0.684	0.311	0.286	0.740	**1.511**	*2.060*
nasioi	0.162	0.748	0.104	0.542	0.240	0.786	0.542	1.855	0.604	**1.400**

Figura 9. Análisis de coordenadas principales.



Análisis de componentes principales (ACP)

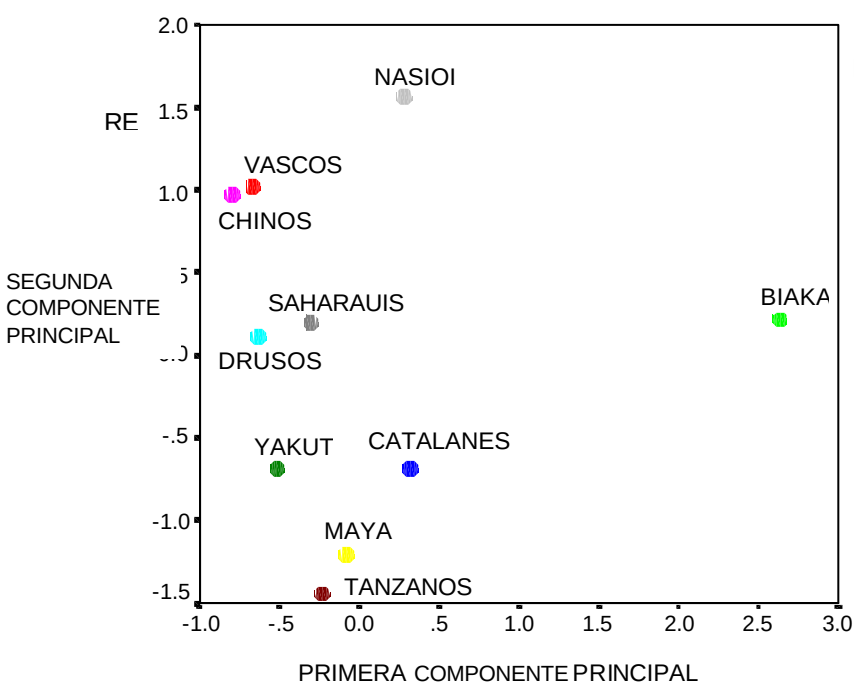
Utilizamos las frecuencias de los 25 haplotipos humanos para inferir afinidades genéticas entre las poblaciones del estudio. Consideramos a las diez poblaciones separadamente (Tabla 5 y Figura 10) y también agrupándolas por continentes (Tabla 6 y Figura 11).

En el primer caso, grupos de poblaciones pertenecientes al mismo continente no quedan agrupados. La primera CP (que explica el 17,7% de la variancia total) separa a la población biaka del resto. Esto se debe principalmente a los haplotipos 24 y 30, que son privados de esta población, y son los que presentan la correlación más alta en la primera CP (r=0,924 en ambos). En segundo lugar, la separación de la población biaka puede explicarse porque es la población con mayor cantidad de cromosomas con los haplotipos 10 (cuatro cromosomas) y 8 (tres cromosomas), que son los siguientes en presentar un mayor coeficiente de correlación. La segunda CP (15,3% de la variancia total) distingue a las poblaciones de nasioi y de tanzanos, separadas en ambos extremos del eje. La correlación más alta de esta segunda CP la presenta el haplotipo 3, con una correlación negativa (r=–0,754), que apunta a que la población nasioi se diferencia del resto por la ausencia del haplotipo 3, presente en todas las poblaciones excepto en ella.

Tabla 5. Coeficientes de correlación entre las frecuencias haplotípicas y la primera y segunda componentes principales (CP). De entre las 25 variables (frecuencias haplotípicas) consideradas en el análisis, se indican aquellas que presentan una correlación (en valor absoluto) mayor de 0,7. Un mayor coeficiente de correlación, indica una mayor contribución de la variable en la construcción de la componente principal. Se indica también el porcentaje de la variancia explicado por cada componente, y el porcentaje de la variancia acumulado.

Variable	Primera CP	Segunda CP
Haplotipo 3	-	-0,754
Haplotipo 10	0,854	-
Haplotipo 8	0,877	-
Haplotipo 24	0,924	-
Haplotipo 30	0,924	-
% de la variancia	17,7	15,3
% acumulado	17,7	33,0

Figura 10. Análisis de componentes principales utilizando las frecuencias de los 25 haplotipos humanos hallados. Se han considerado las diez poblaciones del estudio por separado.



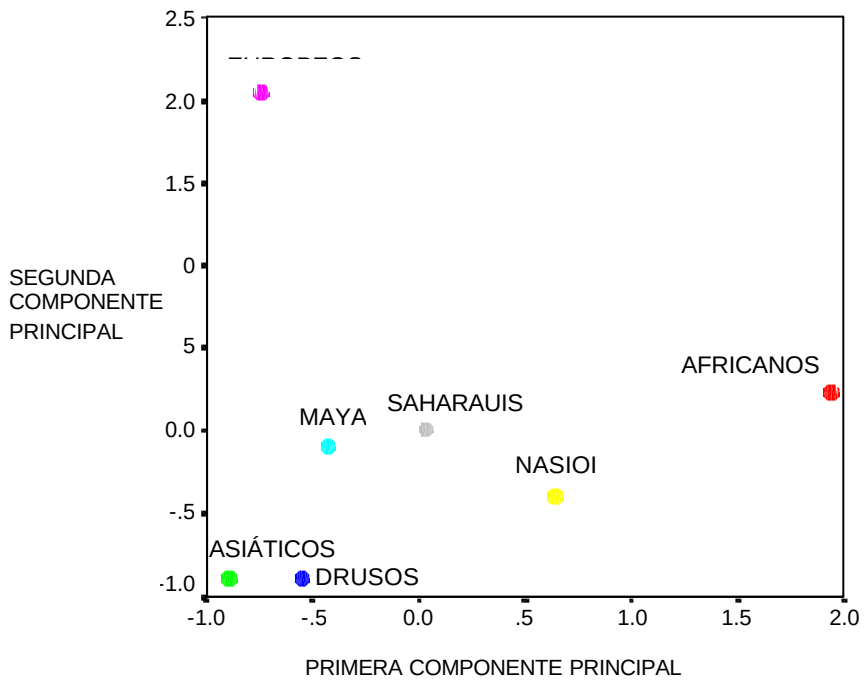
También realizamos el análisis de CP agrupando las poblaciones por continentes del siguiente modo: africanas (biaka y tanzanos), europeas (vascos y catalanes), asiáticas (chinos y yakut). Las cuatro poblaciones restantes también se incluyeron en el análisis (Tabla 6 y Figura 11).

La primera CP (31,9% de la variancia total), separa a las poblaciones africanas del resto. Esta distinción se debe principalmente al haplotipo 10 (con una correlación de 0,951), que está presente mayoritariamente en poblaciones africanas: vascos (un cromosoma), saharauis (un cromosoma), tanzanos (un cromosoma), biaka (cuatro cromosomas) y nasioi (dos cromosomas). También en menor medida, la separación de las poblaciones africanas se debe a los haplotipos 6 y 11 (presentes sólo en tanzanos) y al haplotipo 30 (presente sólo en biaka). La segunda CP (25,6% de la variancia total) separa a los europeos de los asiáticos, y se debe fundamentalmente a los haplotipos 16, 21 y 22, presentes sólo en poblaciones europeas (un cromosoma de cada haplotipo).

Tabla 6. Coeficientes de correlación entre las frecuencias haplotípicas y la primera y segunda componentes principales (CP). Se indican aquellas variables que presentan una correlación (en valor absoluto) mayor de 0,7. Se indica también el porcentaje de la variancia explicado por cada componente, y el porcentaje de la variancia acumulado.

Variable	Primera CP	Segunda CP
Haplotipo 10	0,951	-
Haplotipo 6	0,881	-
Haplotipo 11	0,855	-
Haplotipo 16	-	0,902
Haplotipo 21	-	0,902
Haplotipo 22	-	0,902
Haplotipo 30	0,855	-
% de la variancia	31,9	25,6
% acumulado	31,9	57,5

Figura 11. Análisis de componentes principales utilizando las frecuencias de los 25 haplotipos humanos hallados. Se han considerado las diez poblaciones del estudio agrupadas por continentes (ver texto).



DISCUSIÓN

Las ideas no duran mucho. Hay que hacer algo con ellas.

Santiago Ramón y Cajal.

En este trabajo hemos analizado la variabilidad de una zona no codificante autosómica humana, el pseudogén de la glucocerebrosidasa (psGBA). Los principales resultados del estudio se han preparado en forma de artículos para su publicación independiente, cada uno centrado sobre un aspecto distinto de los patrones, las causas, o las implicaciones de la diversidad que hemos encontrado en psGBA. Hemos presentado estos manuscritos como distintos capítulos y en cada uno de ellos se han discutido ya los resultados correspondientes.

A continuación discutiremos los aspectos más relevantes de este trabajo, e intentaremos integrar los principales resultados obtenidos.

OBTENCIÓN DE HAPLOTIPOS

Los aspectos metodológicos, el modo de llegar a obtener haplotipos del *locus* psGBA, fueron una parte dificultosa del trabajo. No existe ningún método estándar para este proceso, y los procedimientos con que nos encontramos al iniciar el estudio no resultaron aplicables o convincentes para los resultados que queríamos obtener. Los métodos que se habían utilizado hasta entonces eran: (i) ARMS-PCR, que no nos pareció fácilmente aplicable por la dificultad de poner a punto multitud de reacciones de amplificación específicas que abarcaran regiones relativamente largas (Harding *et al*., 1997). (ii) La inferencia de alelos, que preferíamos evitar por la posibilidad de inferir alelos que no fueran reales y de encontrarnos con muestras de las que no se podría deducir la fase, y que por tanto tendrían que ser excluidas del estudio (implementado en el programa Hapinfer; Clark, 1990). (iii) La clonación del producto de PCR y la secuenciación de múltiples clones para deducir la secuencia consenso de ambos alelos, técnica que no nos pareció asequible por el tiempo (y el coste económico) que consume conseguir la secuencia completa de muchos clones de secuencia larga (Clark *et al*., 1998).

Como metodología alternativa a las anteriores, pusimos a punto el protocolo que hemos presentado en el apartado de Material y Métodos. Las ventajas de esta metodología se han discutido ya. Dado que en nuestro caso la formación de alelos quimera es altamente improbable, el análisis de un único clon hace ventajosa esta metodología para análisis haploides de zonas autosómicas.

La inferencia de alelos ha sido la técnica más ampliamente aplicada en los estudios de genética de poblaciones humanas. El tipo de inferencia al que hemos hecho referencia (programa Hapinfer) asigna dos haplotipos concretos a cada individuo. Se han desarrollado otros programas que, como alternativa, no asignan haplotipos a individuos concretos, sino que realizan estimas de las frecuencias haplotípicas en la población (esto evita desestimar muestras del estudio). Este es el caso de los programas Haplo (Hawley y Kidd, 1995) y Arlequin (Excoffier y Slatkin, 1995), que utilizan el algoritmo de "maximización de la esperanza" (*expectation-maximization*) para la inferencia de haplotipos. Se ha demostrado experimentalmente que este método resuelve correctamente aquellos haplotipos que son frecuentes en la población, pero no es capaz de estimar con precisión la frecuencia de haplotipos raros (Tishkoff *et al.*, 2000). En nuestro caso, conocer la presencia y el número de haplotipos únicos y a baja frecuencia en la muestra ha sido determinante para deducir el fenómeno de arrastre genético que ha actuado en psGBA.

A modo de comparación y una vez obtenidos los haplotipos reales, utilizamos el programa Hapinfer (Clark, 1990) para inferir los haplotipos de nuestra muestra. La deducción de haplotipos fue correcta en 44 de los 50 individuos. A partir de los seis individuos en los que la fase no se infirió correctamente, Hapinfer creó tres haplotipos que no se encuentran presentes en nuestra muestra, a la vez que no detectó cinco de los haplotipos reales. Estos cinco haplotipos son todos únicos en la muestra y se hallan en individuos con múltiples lugares heterozigotos. Hapinfer no puede asignar haplotipos únicos a menos que se presenten en individuos que no sean heterozigotos múltiples. Los resultados ponen de manifiesto la necesidad de recurrir a métodos moleculares para la determinación de haplotipos, al menos para los propósitos de este trabajo.

ANÁLISIS DE DIVERSIDAD

Como se ha visto, el número de posiciones segregantes y la diversidad nucleotídica hallados en psGBA son menores de lo que se esperaría para una secuencia no expresada y por tanto presumiblemente neutra respecto a la selección natural. Una alta proporción de haplotipos únicos y dos haplotipos a alta frecuencia son los otros rasgos destacados del patrón de diversidad de psGBA. El valor de la diversidad nucleotídica en psGBA en humanos (0,00044 por nucleótido) indica que, en promedio,

dos secuencias escogidas al azar en nuestra muestra sólo diferirán en dos nucleótidos. El valor de la diversidad nucleotídica promedio en *loci* autosómicos es de 0,0009 (calculado a partir de los estudios indicados en el capítulo I de Resultados: Li y Sadler, 1991; Fullerton *et al*., 1994; Harding *et al*., 1997; Clark *et al*., 1998; Grimsley *et al*., 1998; Rieder *et al*., 1999; Halushka *et al*., 1999; Rana *et al*., 1999; Fullerton *et al*., 2000). La diferencia entre ambos valores resulta más acusada si tenemos en cuenta que la mayoría de *loci* a partir de los que se ha calculado este promedio son genes funcionales, sometidos en mayor o menor grado a presión selectiva purificadora.

Para el análisis de la diversidad genética, el estudio se ha dividido en dos grandes bloques: en un primer momento nos centramos en la comprensión de la variabilidad en psGBA a través del análisis de distintos procesos moleculares de generación y modificación de diversidad. También hemos analizado las relaciones filogenéticas entre los haplotipos psGBA en humanos y en las otras dos especies incluidas en el estudio. En segundo lugar, profundizamos en el análisis del mecanismo que probablemente haya sido más determinante en la generación del patrón de diversidad de psGBA: la selección.

MUTACIÓN, RECOMBINACIÓN Y CONVERSIÓN GÉNICA

Se han detectado 19 posiciones variables en psGBA: 17 sustituciones, una deleción y un tracto de poli-A. En las sustituciones destaca la alta proporción (41,2%) debida a dinucleótidos CpG. En la única deleción observada (308-310delCTC), los tres nucleótidos delecionados forman parte de un *direct repeat* CTTCTCctcATC (el *direct repeat* está subrayado y los nucleótidos delecionados se representan en minúscula). Secuencias *direct repeat*, especialmente de motivos de tres pares de bases, han sido descritas como *hotspot* de deleciones (Efstratiadis *et al*., 1980; Krawczak *et al*., 1991).

La tasa de sustitución que se ha encontrado en psGBA entra en el rango de valores descritos para otros *loci* humanos, tanto genes como pseudogenes.

Se ha sugerido que la tasa de sustitución en pseudogenes es una buena estima de la tasa de mutación espontánea en el DNA. Quizá cabría considerar que estimas previas de tasas de mutación en pseudogenes, alrededor de $4{,}7\times10^{-9}$ por nucleótido y por año (/nt a), se han realizado en *loci* en los que no se ha probado que no exista algún tipo

de presión selectiva (Li *et al*., 1981; Li *et al*., 1985; Li *et al*., 1997). Por otra parte, en un reciente estudio sobre 18 pseudogenes procesados, la tasa de mutación se ha estimado en $1{,}25\times10^{-9}$/nt a (Nachman y Crowell, 2000), valor muy próximo al estimado para psGBA ($1{,}23\times10^{-9}$/nt a).

No obstante, en general y a partir de los datos conocidos actualmente, la tasa de sustitución estimada para psGBA se acerca más a las tasas previamente estimadas para genes funcionales (1×10^{-9}/nt a, Li *et al.*, 1987; $1{,}34\times10^{-9}$/nt a, Harding *et al.*, 1997; $1{,}3\times10^{-9}$/nt a, Clark *et al.*, 1998) que a las tasas estimadas en otros pseudogenes (exceptuando el trabajo de Nachman y Crowell, 2000). Hemos estimado la tasa de sustitución a partir de datos de divergencia entre especies; podríamos pensar que la selección ha sesgado la estima haciendo disminuir la variabilidad en psGBA de modo que la divergencia aparente entre especies también disminuyera. Sin embargo, esto no resulta demasiado verosímil, ya que la presión de selección debería de ser muy baja para empezar a actuar antes de los 5 millones de años que separan a humanos de chimpancés, o de los 7 millones de años que separan a humanos y chimpancés de gorilas, teniendo en cuenta que hemos estimado que el ancestro común más reciente de los alelos psGBA tiene una edad de 200.000 años. Parece muy improbable que los datos de divergencia entre especies estén subestimados por el efecto de la selección sobre psGBA, y por tanto no afectarían a la estima de la tasa de mutación sobre este *locus*.

La recombinación parece haber actuado en la generación de haplotipos en psGBA, si bien no hemos podido cuantificar su efecto con precisión. Hemos calculado los dos estimadores del parámetro de recombinación $C=4N_ec$ más utilizados: $\hat{C}$ (Hudson, 1987) y γ (Hey y Wakeley, 1997), pero ambos sobrestiman la tasa de recombinación real en psGBA, ya que no pueden discernir entre fenómenos de recombinación, mutación recurrente o conversión génica. La sobrestima de $\hat{C}$ cuando se compara con el resultado de γ ($\hat{C}$=16.9 frente a γ=1.27) parece clara. A partir del parámetro γ estimamos una proporción de una recombinación producida cada 2,56 mutaciones, de lo que se extrae que la recombinación (sin olvidar que aquí el término “recombinación” incluye también a otros fenómenos de incremento de diversidad), si bien presente, ha sido menos importante que la mutación en la producción de variabilidad en psGBA.

Como dato complementario que pondría de manifiesto la presencia de recombinación en psGBA y la región circundante, la explicación más plausible para el

patrón de diversidad haplotípica observado sería la de un barrido selectivo acompañado por recombinación, tal y como hemos visto en el capítulo II de Resultados.

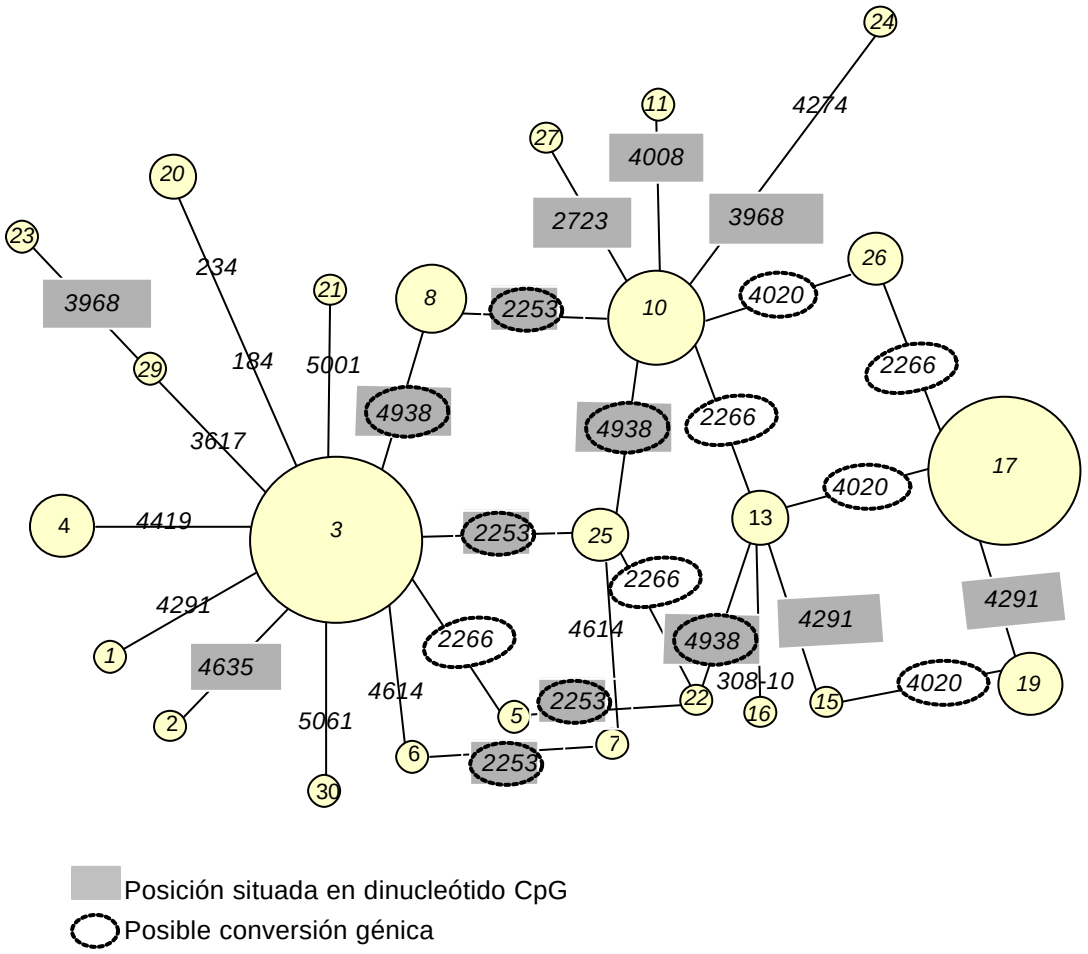
La representación de todos los árboles de máxima parsimonia entre los haplotipos, que se consigue mediante un *median joining network* (MJN), permite investigar los mecanismos productores de variabilidad. Bajo un modelo de infinitos lugares (*infinite-sites model*), el camino que lleva de un haplotipo a otro sería siempre único. De lo contrario, es decir, si se observan reticulaciones en el MJN, se obtiene evidencia de la actuación de mecanismos que incumplen el modelo de infinitos lugares, como recombinación, mutación recurrente y conversión génica. Podemos intentar establecer el origen de las reticulaciones en el MJN, y por tanto de los haplotipos que se pueden haber producido por uno de estos fenómenos, por distintos mecanismos. La composición nucleotídica, y en concreto la presencia de presuntos *hotspots* mutacionales nos informará sobre la posibilidad de mutación recurrente. No se debe descartar una mutación recurrente fuera de un *hotspot*, pero estos *hotspots* incrementan la probabilidad de que se produzca una mutación recurrente. La comparación entre las secuencias psGBA y GBA nos informará sobre la conversión génica. La construcción de haplotipos extendidos psGBA-GBA nos informará sobre procesos de recombinación.

De las seis reticulaciones presentes en el MJN, en principio cinco serían explicables por fenómenos de mutación recurrente ya que, al menos en dos de las ramas de cada reticulación, las mutaciones que conectan un haplotipo con otro se encuentran en dinucleótidos CpG, considerados *hotspots* mutacionales. Aparte de las mutaciones en dinucleótidos CpG, sólo hemos detectado una posición variable situada en otro posible *hotspot* mutacional, en este caso en un tracto de mononucleótidos. La posición 5061 es la tercera C en un tracto de cinco citosinas, pero esta posición no se halla dentro de ninguna reticulación.

Al situar en el MJN las posiciones que podrían haber estado afectadas por conversión génica (2253, 2266, 4020 y 4938), éstas se localizan en las ramas que crean las reticulaciones del MJN (Figura 12). En consecuencia, en la mayoría de casos no es posible discernir si las reticulaciones están producidas por uno de los tres fenómenos (mutación recurrente, conversión génica o recombinación) o por una combinación entre ellos. Lo que sí podemos afirmar es que una estima de la recombinación o de la homoplasia en un *locus* sólo a partir de las reticulaciones del MJN (como se hace normalmente), puede llevar a una sobrestima de estos fenómenos en dicho *locus*.

Un primer análisis de la tabla en la que se muestran las posiciones variables de los haplotipos humanos (capítulo I de Resultados) sugiere que los haplotipos 1, 5, 7, 8, 22 y 25 son haplotipos recombinantes. No obstante, la construcción de haplotipos extendidos psGBA-GBA muestra que es muy poco probable que los haplotipos 7 y 25 sean recombinantes (capítulo I de Resultados). Del resto de los haplotipos, en el 1, 8 y 22 la posibilidad de recombinación viene dada por posiciones que se sitúan en dinucleótidos hipermutables CpG, con lo que no es posible discernir entre recombinación y mutación recurrente debida a un CpG. El único haplotipo recombinante parece ser el haplotipo 5. Podemos también sospechar que este haplotipo se ha producido por conversión génica desde GBA y no por recombinación: (i) observando el MJN, las dos posiciones que tienen que mutar para producir el haplotipo 5 podrían haber estado afectadas por conversión génica desde GBA (Figura 12); (ii) por recombinación, el haplotipo 5 tendría que haberse producido por una combinación de los haplotipos 3 y 22. La localización geográfica de los haplotipos 22 (encontrado sólo en población catalana) y 5 (privado de la población maya), hace improbable esta combinación y aumenta la posibilidad de una conversión génica. No obstante, también hay que considerar que el alcance poblacional de este estudio es limitado y que no hemos detectado todos los haplotipos presentes en cada población. No podemos asegurar que este haplotipo se haya producido realmente por recombinación propiamente dicha, o por conversión génica.

Figura 12. *Median joining network* de haplotipos; corresponde al que hemos representado en el capítulo I de Resultados. El nombre de cada haplotipo se presenta dentro de los círculos, y las mutaciones en las ramas que conectan los haplotipos. El área de los círculos es proporcional a la frecuencia del haplotipo, y la longitud de las ramas es proporcional al número de mutaciones que representan. Se indican las posiciones que pueden haber estado afectadas por mutación recurrente en dinucleótidos CpG y por conversión génica.



La conversión génica de GBA a psGBA ha sido importante en la evolución del pseudogén. La alineación de las secuencias de GBA y psGBA en humanos y la comparación con sus homólogos en chimpancé, apunta a que al menos un 13% de la secuencia de psGBA en humanos proviene de GBA por fenómenos de conversión génica. Estos fragmentos habrían homogeneizado parcialmente (todavía los pseudogenes mantienen más homología entre ellos que la que hay entre pseudogenes y genes dentro de la misma especie) la región psGBA-GBA, y explicarían, al menos en

parte, la alta homología entre ambos *loci*. Si la tasa de conversión génica fuera muy alta, disminuiría el polimorfismo en psGBA, ya que eliminaría toda la variabilidad que se fuera produciendo. Este no parece ser el caso de psGBA, ya que a pesar de la alta homología que comparte con GBA, existen múltiples diferencias puntuales particulares de cada uno de los *loci*. Probablemente, también entre MTX1 y psMTX1 se den mecanismos de conversión génica que sean en parte responsables del 99% de homología que presentan ambos *loci* entre las regiones duplicadas.

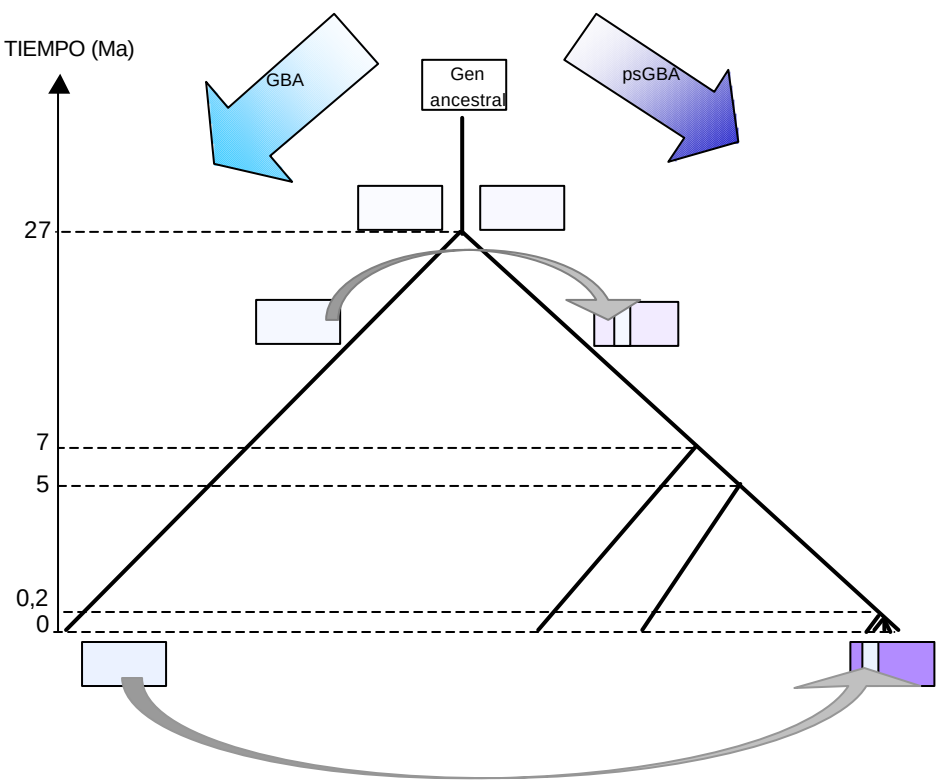
Los fenómenos de conversión génica que hemos detectado son antiguos. No podemos datarlos con exactitud pero sí sabemos que son anteriores al tiempo de coalescencia de los haplotipos psGBA actuales, porque los segmentos en los que hemos detectado conversiones se encuentran fijados en todos los alelos psGBA analizados, es decir, la conversión se había producido ya en el alelo psGBA a partir del cual se ha originado toda la diversidad actual (Figura 13).

El efecto de estas conversiones génicas en la comparación entre psGBA y GBA es el de reinicializar el reloj molecular de fragmentos de psGBA, eliminando la variabilidad que hubieran podido acumular hasta entonces. Como hemos visto, conversiones en sentido inverso, de psGBA a GBA, se eliminan de la población si afectan a la funcionalidad del gen. Los posibles fragmentos de conversión que hemos detectado en regiones intrónicas y de los que no podemos determinar cuál ha sido el *locus* aceptor y cuál el donador, serían los únicos en los que se ha podido dar una transferencia de psGBA a GBA que se habría fijado con el tiempo.

Conversiones hacia psGBA que se produjeran después del TMRCA (*time to the most recent common ancestor*) incrementarían el polimorfismo del *locus*, ya que tendrían el efecto de introducir variantes nuevas. Estas conversiones aumentarían la estima de la tasa de mutación, ya que no podemos diferenciar si la posición variable está causada por un fenómeno o por otro. Este es el caso de cuatro posiciones segregantes en psGBA (2253, 2266, 4020 y 4938). Dos de ellas se sitúan en dinucleótidos hipervariables CpG (2253 y 4938), lo que quizá haría más probable que se hubieran originado por mutación recurrente en dinucleótidos CpG que por conversión génica. Sobre las otras dos posiciones no es posible distinguir entre ambos mecanismos. Merece ser destacado el hecho de que la conversión génica producida antes del TMRCA, que afecta a un mínimo de 709 nucleótidos, ha tenido mucha más importancia en la formación del pseudogén actual que la conversión génica que haya podido producirse después del TMRCA, que quizá afecte a 4 posiciones.

Los fragmentos de conversión génica detectados no afectan a la estima de fenómenos posteriores a la divergencia de haplotipos psGBA (excepto el efecto de sobrestimar la mutación aparente que ya se ha comentado). Habría que tener en cuenta el efecto de la conversión génica en las estimas de fenómenos que se dieran antes de la coalescencia de los alelos psGBA actuales (como en la estima del tiempo de duplicación GBA-psGBA), pero no en fenómenos posteriores (estimas de los tiempos de coalescencia y de la edad de las mutaciones).

Figura 13. Conversión génica (flechas curvas) en psGBA a lo largo del tiempo. La flecha superior representa los fenómenos de conversión génica que hemos detectado. La flecha inferior representa fenómenos de conversión que podrían haberse dado. Los tiempos que se indican corresponden a: duplicación génica GBA-psGBA (27 Ma); divergencia humanos-gorilas (7 Ma); divergencia humanos-chimpancés (5 Ma); ancestro común más reciente para las moléculas de psGBA actuales (200.000 años). El eje temporal no se representa a escala. Ma: millones de años.



El efecto homogeneizador de conversiones génicas anteriores al TMRCA es el que hemos tenido en cuenta en la estima de la edad de duplicación de psGBA y GBA. El primer dato calculado (23 millones de años) se referiría al 87% de psGBA no afectado por conversión génica. Una vez corregido por el efecto de la conversión, la edad de la duplicación se sitúa en 27 millones de años, poco antes de la divergencia de los monos del viejo mundo. Más que a la duplicación génica, este tiempo apuntaría a la edad de la inactivación de psGBA, ya que éste podría haber sido funcional durante un tiempo después de la duplicación.

La conversión génica se ha destacado como un mecanismo importante en la evolución de muchos pares gen-pseudogén (Cooper, 1999). Quizá el ejemplo más sorprendente sea el del gen de la 21-hidroxilasa esteroide (CYP21), ya que un 75% de los alelos CYP21 mutados son producto de conversiones génicas en las que el pseudogén psCYP21 transfiere fragmentos de DNA al gen funcional (Tusié-Luna y White, 1995). La recombinación se ha considerado un mecanismo esencial para incrementar la diversidad genética entre individuos. Quizá cabría considerar a la conversión génica como un mecanismo con el mismo efecto.

RELACIONES FILOGENÉTICAS

Hemos representado la filogenia de los haplotipos humanos mediante dos métodos: *median joining networks* (MJN) (Bandelt *et al*., 1995) y mediante el algoritmo de *neighbor-joining* (Saitou y Nei, 1987).

Los árboles de *neighbor-joining* han resultado visualmente muy informativos cuando comparamos diferentes especies entre sí (humanos, chimpancés y gorilas), o los distintos *loci* (psGBA y GBA), pero no cuando representamos sólo las relaciones entre los haplotipos psGBA en humanos (nos informa en todo caso de la falta de estructura geográfica entre los haplotipos).

El árbol de *neighbor-joining* que muestra la relación entre las secuencias de psGBA en humanos, chimpancés y gorilas (capítulo I de Resultados), refleja la baja diversidad entre haplotipos humanos. Como dato a destacar, entre los tres haplotipos hallados en gorila la distancia es mayor que entre los 25 haplotipos hallados en humanos. Nuestros datos, por tanto, reflejan el patrón de baja diversidad en el DNA nuclear humano que ya se ha descrito en otros estudios (Li y Sadler, 1991; Huang *et al*., 1998). Por otra parte, a pesar de que sólo hemos analizado cuatro cromosomas de gorilas, la diversidad

observada (mayor que en humanos) podría indicar que la selección (en forma de barrido selectivo) no ha actuado sobre psGBA en esta especie, aunque quizá sólo apunte a que se trata de una especie más antigua, y por tanto con más tiempo para acumular variación, que la humana.

En el MJN de haplotipos humanos, los alelos se dividen en dos grandes grupos, que podríamos definir como los haplotipos que rodean al haplotipo 3, y el resto, con el haplotipo 17 como mayoritario entre ellos. A pesar de que la distribución de diferencias de *pairwise* muestra diferencias entre estos dos haplogrupos, la división entre ellos no es estricta. El escaso número de posiciones segregantes (sólo cuatro posiciones separan el haplotipo 3 del 17) y la probable ocurrencia de mutación recurrente, conversión génica y recombinación, hacen que la división entre haplogrupos no sea inequívoca y clara.

La estructura de haplotipos estrellada alrededor del haplotipo 3 (presente en el 29% de los cromosomas) no se presenta alrededor del haplotipo 17 (presente en el 23% de los cromosomas). Aunque normalmente una filogenia estrellada apunta a una creación reciente de haplotipos (ha habido tiempo sólo para la formación de ramas cortas), en el caso concreto de psGBA la estructura del MJN podría indicar que el haplotipo 3 es más antiguo que el haplotipo 17. A partir del haplotipo 17 no habría habido tiempo de que se produjeran mutaciones que llevaran a la radiación de haplotipos alrededor del mismo. Al extender los haplotipos psGBA hacia el gen GBA, el par de haplotipos GBA - /psGBA 17 y el par GBA + /psGBA 3 están ligados, tal y como ya se ha comentado a lo largo del trabajo. No obstante, sí hemos detectado algunos individuos con la combinación de haplotipos GBA - /psGBA 3, mientras que no hemos detectado ningún par GBA + /psGBA 17. Esto sería compatible con que el haplotipo 17 es de formación reciente, de manera que las recombinaciones GBA + /psGBA 17 no habrían tenido tiempo de producirse. Las edades estimadas por el programa Genetree para los haplotipos 3 y 17 concordarían con la segunda parte del razonamiento: el haplotipo 3 tiene una edad estimada de 74.400 años y el haplotipo 17 de 38.000 años.

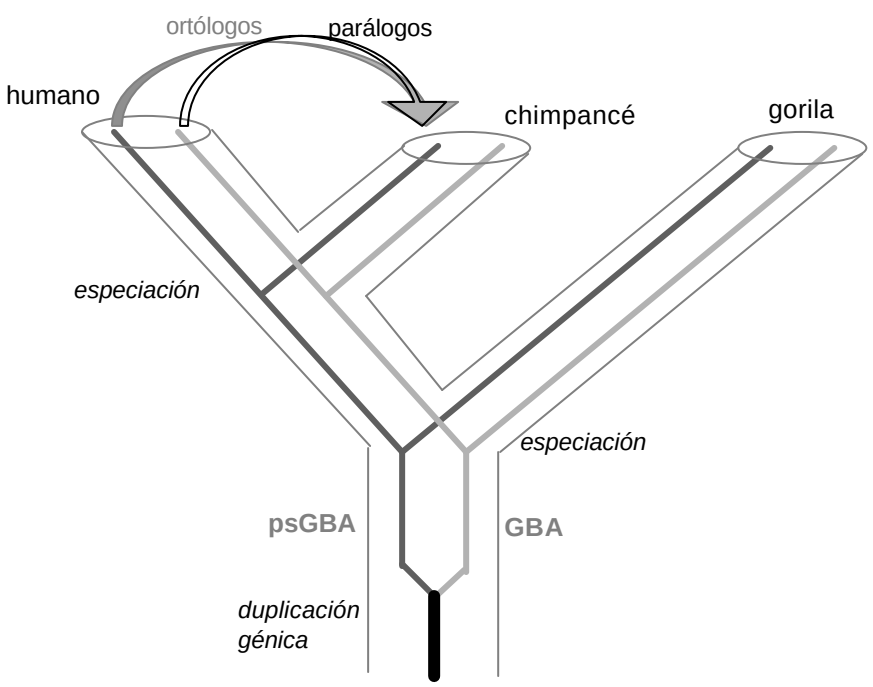
Una hipótesis alternativa sería que se hubieran producido dos barridos selectivos consecutivos en psGBA, cada uno de ellos arrastrando a uno de los haplotipos psGBA mayoritarios, 3 y 17, y que el barrido que afecta al haplotipo 17 fuera el más reciente, de manera que no habría habido tiempo de que se produjeran mutaciones a partir de este haplotipo.

Entre especies, la homología (calculada como número de nucleótidos que no han experimentado sustituciones, por secuencia) entre las secuencias es muy alta para todos los pares de comparaciones que hemos llevado a cabo: la homología entre psGBA en humanos y psGBA en chimpancés es del 98,6% (78 sustituciones y seis inserciones/deleciones en 5420 pb), al igual que entre psGBA en chimpancés y psGBA en gorilas (77 sustituciones y ocho inserciones/deleciones en 5420 pb). Entre psGBA en humanos y psGBA en gorilas la homología es del 98% (99 sustituciones y ocho inserciones /deleciones en 5420 pb).

Entre genes hemos encontrado la homología más alta: 99,1% entre GBA en humanos, y GBA en chimpancés (62 sustituciones y cinco inserciones/deleciones en 7.156 pb).

La homología más baja se da entre los pares gen-pseudogén: 95,5% de homología entre GBA y psGBA en chimpancé (259 sustituciones y 26 inserciones/deleciones), y 95,3% de homología entre GBA y psGBA en humanos (258 sustituciones y 29 inserciones/deleciones), valor ligeramente inferior a lo que se había descrito previamente (96%; Horowitz *et al*., 1989)

La estructura de los árboles de *neighbor-joining* concuerda con el hecho de que la duplicación psGBA-GBA es anterior a la separación de especies humana, de chimpancés y de gorilas. La relación filogenética gen-especie habría evolucionado del siguiente modo:



ASPECTOS POBLACIONALES

Como se ha mostrado en el capítulo V de Resultados, ni el análisis de componentes principales, ni el de coordenadas principales han mostrado un patrón geográfico claro entre las poblaciones. La ausencia de estructura geográfica también se puede apreciar en el árbol de *neighbor-joining* para los haplotipos humanos, en el que los valores de *bootstrap* son muy bajos, y en la distribución de haplotipos en las poblaciones, que no presenta ningún tipo de orden geográfico determinado.

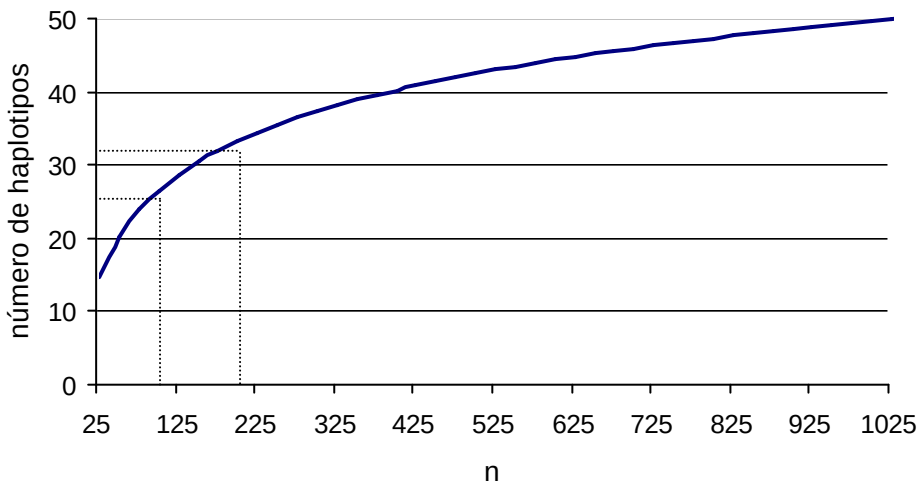
Los únicos puntos a destacar serían: (i) la mayor separación de las poblaciones maya y nasioi del resto, debido principalmente a que son las que presentan mayor cantidad de cromosomas con haplotipos privados; (ii) la separación de la población biaka, debido a su contenido en haplotipos 10 y 8, y a los haplotipos privados 24 y 30. La población biaka también queda destacada de forma indirecta en el análisis de coalescencia por ser la única población que presenta el haplotipo más antiguo (haplotipo 24).

Cabe considerar que el alcance poblacional de este estudio es limitado. El hecho de que no observemos estructura poblacional con los datos de variación de psGBA puede tener diversos motivos. En primer lugar, la variabilidad genética es baja, por lo que no hay muchas posiciones segregantes que permitan diferenciar grupos. Por otra parte, según las estimas del programa Genetree, la mayoría de las mutaciones son anteriores al tiempo de diferenciación de las poblaciones humanas (entre 40.000 y 50.000 años, a partir de datos arqueológicos y paleontológicos; Klein, 1995). Es decir, la variabilidad no se estructura entre las diferentes poblaciones del análisis. Por último, quizá con una muestra mayor hubiéramos observado algún tipo de estructura geográfica; cien cromosomas son suficientes para obtener una visión global de la variabilidad genética en un *locus*, pero quizá no para caracterizar poblaciones a un nivel más local.

No obstante, y respecto al tamaño muestral, dada la baja diversidad observada en psGBA y visto el efecto de la selección en este locus, no parece probable que aumentando el número de cromosomas aumentemos considerablemente el número de haplotipos. A partir de la fórmula de Ewens (Ewens, 1972) que expresa la relación entre el tamaño muestral y el número de alelos bajo condiciones de neutralidad, hemos calculado el número de haplotipos esperados para distintos tamaños muestrales (Figura 14). Para realizar este cálculo se utilizó un pequeño *script* escrito *ad hoc* en Quick Basic (SAMPLEK, francesc.calafell@cexs.upf.es). Se observa que doblando el tamaño muestral

llegamos a 32 haplotipos, mientras que para obtener el doble de haplotipos el tamaño muestral tendría que multiplicarse por diez. El barrido selectivo que condiciona a psGBA haría que el número de haplotipos aún fuera menor de lo calculado.

Figura 14. Número esperado de haplotipos para psGBA dados distintos tamaños muestrales (n), y a partir de la fórmula de Ewens (Ewens, 1972). El tamaño muestral se ha incrementado en saltos de 25 cromosomas. Los haplotipos obtenidos (25 haplotipos a partir de 100 cromosomas), y el número de haplotipos que esperaríamos obtener a partir de 200 cromosomas (32 haplotipos) se indican mediante líneas discontinuas.



Si bien no podemos profundizar en un análisis poblacional entre grupos humanos distintos, sí podemos realizar un análisis más general que, dada la representación global de la muestra, abarcaría a toda la especie. Se han podido extraer algunas inferencias sobre la historia de las poblaciones humanas, que se han derivado en gran parte de la aplicación de la teoría de la coalescencia.

Utilizamos el método de coalescencia para estimar la edad de las mutaciones y el tiempo del ancestro común más reciente (TMRCA). Bajo un modelo de población constante, el tiempo de coalescencia de los haplotipos psGBA es de 200.000 años, es decir, la diversificación de la variabilidad de psGBA es posterior a 200.000 años. Bajo un modelo de población en expansión, la edad estimada es de 90.000 años; si la población crece, se aceleran los fenómenos de coalescencia y se llega antes al TMRCA. La

verosimilitud de los árboles de coalescencia bajo ambos modelos poblacionales no resultó significativamente distinta.

El tiempo de coalescencia de 200.000 años es el más reciente encontrado para la especie humana hasta el momento a partir de datos autosómicos. Un tiempo reciente hasta el MRCA, y el hecho de que el haplotipo más próximo al estado ancestral se encuentre sólo en una población subsahariana, concuerdan con la hipótesis del "*Out of Africa*". Según esta hipótesis, los humanos anatómicamente modernos emergieron en África hace entre 100.000 y 200.000 años, y subsiguientemente se extendieron al resto del mundo reemplazando a las formas humanas arcaicas. Como alternativa, la hipótesis multirregional propone que hubo una única salida, de *Homo erectus*, del continente africano hace unos 2 millones de años, y que la transformación de humanos anatómicamente modernos ocurrió en distintas partes del mundo de forma paralela. Dado que la coalescencia de las moléculas debe de ser anterior al tiempo de coalescencia de la población, una edad de 200.000 años concordaría con la hipótesis *Out of Africa*. Es decir, la coalescencia de los genes sería anterior a una migración reciente fuera del continente africano. La diversidad genética actual de psGBA se habría generado dentro de los humanos anatómicamente modernos, no en el *pool* genético de *Homo erectus*.

Si un grupo migrador hubiera salido del continente africano y a continuación hubiera experimentado una expansión de población, la migración podría datarse. En el caso de psGBA esta migración se situaría en 200.000 años. Podemos acotar un poco más la edad de la migración fuera de África, por la posición en la genealogía de psGBA de la mutación 3968. Esta mutación, con una edad estimada de 164.000 años, se encuentra presente tanto en poblaciones africanas como no africanas, lo que la sitúa con anterioridad a cualquier posible cuello de botella poblacional y a cualquier expansión de población y, en consecuencia, de haplotipos. La mutación más antigua hallada sólo fuera de África es la mutación 184, con una edad estimada de 19.000 años. La edad de esta mutación, posterior a la posible expansión de población fuera de África, situaría el límite inferior de la edad de la migración.

SELECCIÓN EN psGBA

La baja variabilidad en psGBA y el inesperado patrón de la partición alélica, nos hicieron indagar en el posible papel de fenómenos selectivos en la formación de la diversidad actual de psGBA.

En primer lugar, la separación entre fenómenos genéticos y fenómenos poblacionales es importante, porque fenómenos poblacionales como una expansión, llevarían al mismo patrón de partición alélica, y a iguales resultados de los tests de selección. A fin de separar y diferenciar entre ambos tipos de fenómenos, es necesario comparar más de un *locus* para las mismas poblaciones, de modo que obtengamos comparaciones con el mismo marco histórico de referencia. Factores poblacionales tienden a actuar en todo el genoma a la vez, mientras que factores genómicos tendrán un impacto mucho más restringido. La comparación de distintos *loci* en muestras de representación global nos ha llevado a ver que la reducción de variabilidad en psGBA no se debe a un efecto general sobre todo el genoma, y por tanto no es un fenómeno poblacional.

La descripción de la variabilidad en una región genómica se lleva a cabo normalmente con el cálculo de parámetros que asumen ciertos presupuestos: neutralidad y el modelo de infinitos lugares en el proceso de ocurrencia de mutaciones (supone ausencia de fenómenos de recombinación, mutación recurrente y conversión génica), apareamientos al azar y una población constante en el tiempo y sin subestructuración. Es sabido que estos modelos suelen ser una simplificación de la realidad. Para verificar si las observaciones se corresponden con estos supuestos, se han desarrollado una serie de tests de contraste de la hipótesis neutralista, como el test HKA, basado en la proporción de variación esperada en dos *loci* en dos especies distintas (Hudson *et al*., 1987), y los tests basados en distintos estimadores del parámetro de diversidad $\theta=4N_e\mu$, como el estadístico D de Tajima (Tajima, 1989) o los tests F* y D* de Fu y Li (Fu y Li, 1993), y los tests basados en la proporción de alelos esperados, como el test Fs de Fu (Fu, 1997; De Lorenzo, 1998).

No detectamos desviaciones del modelo neutro en psGBA con los tests de Tajima y Fu y Li. Como ya se ha comentado en el capítulo II de Resultados, esto se ha descrito también para otros *loci* humanos, para los que existen evidencias externas e

independientes de que están sometidos a presión selectiva. Sin embargo, otras observaciones indican que la selección ha tenido un efecto en psGBA (y probablemente sobre *loci* cercanos): el bajo número de posiciones segregantes en psGBA respecto a lo que sería de esperar en un pseudogén, la significación estadística de los tests HKA y Fs de Fu, y la comparación del patrón de partición alélica observado con el esperado bajo neutralismo. El conjunto de resultados apunta a que psGBA ha estado bajo el efecto de un barrido selectivo, que debe de haber afectado a cierta región cromosómica en 1q21.

Otro aspecto relacionado con la selección es la neutralidad de la duplicación génica psGBA-GBA. Podemos considerar que la duplicación en sí es neutra: si hubiese sido favorable, es decir, si tener dos copias funcionales del gen GBA confiriera una ventaja sobre tener una, entonces las mutaciones de pérdida de función que originaron psGBA no se habrían fijado, ya que habrían estado en desventaja respecto a los cromosomas con las dos copias del gen activo; si la duplicación fuese deletérea entonces no se habría llegado a fijar nunca. Dado que la duplicación es neutra, no hay ninguna fuerza que retenga activo a psGBA, o que acelere su inactivación.

Aparte del arrastre genético, hemos indagado en otro tipo de selección, la selección purificadora sobre el gen GBA, que nos ha permitido observar las diferencias entre mutación producida, aceptada y eliminada (capítulo IV de Resultados). Contamos con dos secuencias altamente homólogas y en la misma región genómica (GBA y psGBA), lo que nos ha posibilitado comparar sus espectros de mutación. Nuestro objetivo era dilucidar si las diferencias entre los espectros mutacionales podían ser entendidas por la acción de la selección purificadora sobre GBA. Observaríamos la mutación total producida (en psGBA), la proporción de estas mutaciones que se eliminan (mutaciones causantes de enfermedad en GBA) y la mutación que finalmente es aceptada (polimorfismos en GBA). Hemos observado que el contenido de dinucleótidos CpG es similar entre el gen GBA y psGBA, pero la contribución de pares CpG al polimorfismo en psGBA es mayor que su aporte a las mutaciones en GBA causantes de la enfermedad de Gaucher.

El arrastre genético que se extiende en 1q21 no influenciaría a la selección purificadora propia del gen GBA, con la excepción de que podría costar más eliminar mutaciones que, sin llevar a enfermedad, fueran ligeramente desventajosas, cuando estas mutaciones se produjeran en el haplotipo que está siendo seleccionado.

EFECTO DE LA SELECCIÓN EN LAS MEDIDAS POBLACIONALES Y DE DIVERSIDAD

Para extraer conclusiones que fueran aplicables a la historia de las poblaciones, como la estima del tiempo de coalescencia de los haplotipos de un *locus*, el *locus* en cuestión debería ser neutro. El análisis de coalescencia sólo considera, por ahora al menos, condiciones de neutralidad. Teniendo en cuenta que existe selección sobre psGBA, y que un barrido selectivo afecta al espectro de variabilidad, creando un exceso de mutaciones y haplotipos únicos, ¿cuál es entonces la validez de las estimas temporales que hemos obtenido? Un punto a tener en cuenta es que el tiempo de coalescencia estimado, de 200.000 años, es correcto para las moléculas de psGBA presentes en la muestra. Además, aunque sólo hemos utilizado en el análisis de coalescencia los haplotipos que siguen el modelo de infinitos alelos, hemos observado que este subconjunto no presenta unos resultados de variabilidad significativamente distintos a los del conjunto entero de haplotipos. Por tanto, las estimas temporales son aplicables a toda la muestra de haplotipos psGBA.

El efecto de la selección sobre el tiempo de coalescencia viene dado de manera indirecta por la eliminación de alelos psGBA de la población. Sin selección observaríamos más moléculas de psGBA (que se distribuirían en un patrón de partición alélica distinto), y por tanto encontraríamos tiempos más antiguos, seguramente más similares a los encontrados en otros *loci* autosómicos (750.000 años, Harding *et al*., 1997; 1.200.000 años, Clark *et al*., 1998). Si dispusiéramos de más datos de coalescencia de haplotipos de otros *loci* autosómicos neutros, podríamos especular que el tiempo de coalescencia de los haplotipos psGBA, sin la acción de la selección, se situaría dentro del rango dado por esos valores. En todo caso, con los datos disponibles, sabemos que el tiempo de coalescencia sería mayor sin selección, pero no podemos estimar cuánto. Para ello necesitaríamos saber cuántas y qué moléculas han sido eliminadas de la población, y esto resulta imposible de predecir. Las edades de las mutaciones y del TMRCA dependerían de la fuerza de la presión selectiva y del tiempo en que empezó a actuar.

Sin embargo, hay otro punto a tener en cuenta que apunta a que aun sin selección el tiempo de coalescencia de psGBA seguiría concordando con la teoría *Out of Africa*. Un barrido selectivo que afecte a toda la variabilidad en un *locus* (como se observa para psGBA) es más probable que se dé cuando el tamaño poblacional y el espacio que ocupa la población son reducidos, y estas características son asunciones de la hipótesis *Out of Africa*.

La presión selectiva puede variar con el ambiente, y por tanto podría haber tenido efectos distintos sobre poblaciones localizadas en ambientes diferentes, modificando las diferencias genéticas entre las poblaciones. Este punto no parece ser de gran importancia en psGBA dado el grado menor de separación geográfica observado y, teniendo en cuenta que el valor obtenido de Fst es muy semejante al que se da en otros *loci* autosómicos (Capítulo II de Resultados; Barbujani *et al*., 1997; Zietkiewicz *et al*., 1997; Clark *et al*., 1998; Jorde *et al*., 2000).

El hecho de que la selección sobre psGBA haya eliminado a lo largo del tiempo parte de la genealogía de haplotipos repercute en la cantidad de diversidad nucleotídica que observamos, que es menor de lo que sería sin selección. No obstante, la selección no afecta a la tasa de sustitución que hemos calculado, ya que ésta se estima a partir de datos de divergencia, y como ya se ha comentado antes, es muy poco probable que el arrastre genético empezara a actuar hace más de cinco millones de años de evolución independiente de humanos y chimpancés, o antes de siete millones de años en el caso de la separación de gorilas.

EL PAPEL DE psGBA EN LA ENFERMEDAD DE GAUCHER

Dilucidar la importancia de la variabilidad en psGBA en una enfermedad humana ha sido otro aspecto de este trabajo. Conocer la secuencia y la estructura de psGBA fue muy importante para entender algunos fenómenos mutacionales en el gen GBA, como la formación de alelos complejos, y también para diseñar estrategias adecuadas para buscar mutaciones en GBA (Hong *et al*., 1990; Tayebi *et al*., 1996; Finckh *et al*., 1998). No obstante, hasta el momento se conocía la secuencia de un único alelo psGBA (Horowitz *et al*., 1989).

El método de identificación de alelos GBA en pacientes Gaucher es a veces problemático para la correcta clasificación del alelo mutado entre alelo recombinante y alelo con una mutación *de novo*. Un *screening* mutacional restringido como alternativa a secuenciar el segmento GBA entero, y también el hecho de que las mutaciones provenientes de psGBA pueden encontrarse en fragmentos de DNA completamente homólogos a GBA, dificultan la correcta identificación de alelos. El conocimiento de la

variabilidad de psGBA posibilita una mejor identificación de las secuencias de ambos *loci* y ayuda a la clasificación de alelos GBA causantes de enfermedad.

Los alelos complejos GBA-psGBA llevarán probablemente a las formas de enfermedad más graves. Esto implica que el conocimiento de los alelos que porta el paciente sea importante para predecir el curso de la enfermedad. Conocer de antemano el pronóstico de la enfermedad es un punto clave a la hora de aplicar terapias preventivas que eviten, por ejemplo, una degeneración nerviosa irreversible.

Respecto a las mutaciones causantes de enfermedad en GBA que se han descrito como provenientes de psGBA, y teniendo en cuenta las variantes que hemos encontrado en psGBA, la Tabla 2 del apartado de Introducción (página 53) quedaría modificada en las posiciones siguientes (las nuevas variantes nucleotídicas se indican en negrita):

Sustitución nucleotídica GBA→psGBA	Posición	Sustitución de aminoácido	Nucleótido en psGBA
A→G	exón 6	188Asn → Ser	**A**/G
G→C	exón 10	456Ala → Pro	**G**/C

Hemos comentado ya en el capítulo III de Resultados cómo el conocimiento de estas variables afecta y simplifica la interpretación de algunos alelos complejos descritos en la literatura.

OBSERVACIONES FINALES

Los procesos genómicos de creación, mantenimiento y eliminación de variabilidad están demostrando ser más versátiles y dinámicos de lo que quizá se creyó en un principio. Los pseudogenes son un ejemplo interesante para observar parte de la riqueza de mecanismos de que dispone el genoma.

Hemos descrito el papel de distintos fenómenos genómicos en un *locus* autosómico, que han apuntado a la importancia del contexto genómico para la interpretación de los datos de una región o de un *locus* concreto. Sin considerar las regiones circundantes y los fenómenos que pueden darse en ellas (selección positiva en este caso), no es posible entender la variabilidad en psGBA.

Podríamos considerar al genoma como un continuo donde determinados puntos o zonas experimentan máximos de determinados fenómenos (como por ejemplo, selección positiva en un *locus* ligado a psGBA). La influencia de estos fenómenos se extendería en la región circundante en mayor o menor grado (en este caso dependiendo por ejemplo de la tasa de recombinación en la región), y los efectos de distintos mecanismos confluirían (por ejemplo en GBA se unen el efecto del arrastre genético más la propia selección purificadora, y en psGBA se unen el efecto del arrastre más la neutralidad de las mutaciones, etc.), haciendo de cada región del genoma el resultado de una combinación compleja y única.

Se han hecho algunas inferencias sobre la historia de las poblaciones humanas, que concuerdan con la teoría de un origen africano reciente para los humanos anatómicamente modernos. Los datos que hemos obtenido en psGBA apuntan a que la diversificación de los humanos modernos es posterior a la diversificación del pseudogén, y por tanto a 200.000 años. No obstante, las posibles inferencias a partir de datos genómicos a la historia humana quizá deberían hacerse con mucha cautela, al menos en el caso de psGBA. La selección ha jugado muy probablemente un papel importante en la formación del espectro de variabilidad de psGBA, y por tanto la genealogía de psGBA, su corto tiempo de coalescencia, quizá se deba más al efecto de fenómenos genómicos que poblacionales. Este trabajo refleja la importancia de entender la dinámica de la región genómica de estudio antes de inferir conclusiones poblacionales.

Lo que sí nos indican los datos de psGBA, es que es posible hallar unos tiempos de coalescencia tan recientes para *loci* autosómicos, que en principio esperaríamos que nos aportaran mayor profundidad temporal en los análisis que los datos provenientes de cromosomas X, Y y mtDNA.

Quizá la función más importante de los pseudogenes sea la de actuar como reserva de variabilidad genética para el correspondiente homólogo funcional. Esta reserva tendría un efecto beneficioso; por ejemplo, procurar adaptación frente a cambios ambientales, o crear genes de función mejorada (Rouquier *et al*., 1998; Glusman *et al*., 2000). El contrapeso evolutivo de esta ventaja sería que, así como se puede transferir "variabilidad" ventajosa, también se transfieren secuencias que resultarán en alelos del homólogo funcional defectivos y llevarán a enfermedad, como en el caso del par GBA/psGBA.

Resulta difícil extraer conclusiones que puedan generalizarse a otros pseudogenes a partir de los datos obtenidos en psGBA. Primero, porque los fenómenos selectivos a los que está sometida esta región la hacen única. Una excepción a esto sea posiblemente el *locus* psMTX1, en el que con probabilidad se ha dado el mismo fenómeno de arrastre genético que ha sufrido psGBA. Segundo, por que una de las conclusiones que hemos podido extraer de este trabajo es que, si bien con la idea del contexto genómico siempre presente, la historia que nos cuenta cada región genómica es distinta, y que la historia completa de un genoma y de las poblaciones y especies que lo portan no se hallará replicada en cada *locus*, sino que se deberá resolver haciendo encajar las piezas de un complicado puzzle.

Las obras no se acaban, se abandonan.
Paul Valéry

Muchas veces, si no siempre, la investigación podría ramificarse sin un final preciso. Al acabar una tesis doctoral, algunas preguntas, determinadas por los objetivos iniciales de la tesis, quedan contestadas, pero también se empiezan a formular muchas otras, o surgen nuevos aspectos que sería interesante analizar. En el caso concreto de este trabajo, hay algunas de estas "ramificaciones" que podrían marcar puntos de inicio de análisis futuros.

Sería interesante ver, por ejemplo, cuál es la diversidad de psGBA en poblaciones en las que la enfermedad de Gaucher es especialmente prevalente, como judíos ashkenazitas, suecos de la región de Norrbotten y en población japonesa, donde se han detectado fenotipos especialmente severos para mutaciones que en población occidental producen fenotipos más leves; también, desde otro punto de vista, en cromosomas con una mutación concreta causante de la enfermedad de Gaucher.

Quizá el estudio de los polimorfismos encontrados en psGBA en muestras mayores podría mostrar subestructuras poblacionales humanas que con el presente tamaño muestral no se han podido detectar.

Podríamos analizar la variabilidad de psGBA en poblaciones más amplias de chimpancés y de gorilas, para averiguar qué cantidad de variabilidad presentan y si la selección ha tenido algún efecto en estas especies. Estos análisis, junto a análisis de coalescencia en estas dos especies ayudarían a acotar más el tiempo en que empezó a producirse el arrastre genético en psGBA.

Dado que hemos obtenido la secuencia de GBA y psGBA en chimpancés, podríamos analizar los patrones de conversión génica en esta especie (alineando las secuencias de psGBA y GBA en chimpancé y comparándolas con psGBA y GBA en humanos), para compararlos con el patrón de conversión génica detectado en humanos.

Y quizá el punto más intrigante que queda abierto es la identificación del gen (o genes) y las variantes alélicas que están siendo seleccionadas positivamente, y que arrastran consigo la diversidad circundante, así como entender el motivo de esta selección y ver hasta dónde alcanza el efecto del arrastre.

BIBLIOGRAFÍA

A

Aguadé M, Langley CH. 1994. Polymorphism and divergence in regions of low recombination in *Drosophila*, pp 67-76. En*: Non-neutral Evolution: Theories and Molecular Data*, Golding GB ed. Chapman & Hall, New York.

Amaral O, Marcao A, Pinto E, Zimran A, Miranda SMC. 1997. Distinct haplotype in Non-Ashkenazi Gaucher patients with N370S mutation. Blood Cells Mol Dis 23 (22): 415-416.

Armstrong LC, Komiya T, Bergman EB, Mihara K. 1997. Metaxin is a component of a preprotein import complex in the outer membrane of the mammalian mitochondrion. J Biol Chem 272: 6510-6518.

Armstrong LC, Saenz AJ, Bornstein P. 1999. Metaxin 1 interacts with metaxin 2, a novel related protein associated with the mammalian mitochondrial outer membrane. J Cell Biochem 74: 11-22.

B

Bandelt H, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141: 743-753.

Barbujani G, Magagni A, Minch E, Cavalli-Sforza L. 1997. An appointment of human DNA diversity. Proc Natl Acad Sci USA 94: 4516-4519.

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

Bertranpetit J, Calafell F. 1996. Genetic and geographical variability in cystic fibrosis: evolutionary considerations. En: *Variation in the human genome*, pp 97-118. Wiley, Chichester (Ciba Foundation Symposium 197).

Bertranpetit J. 2000. Genome, diversity, and origins: the Y chromosome as a storyteller. Proc Natl Acad Sci USA 97 (13): 6927-6929.

Beutler E, West C, Gelbart T. 1992. Polymorphisms in the human glucocerebrosidase gene. Genomics 12: 795-800.

Beutler E, Nguyen NJ, Henneberger MW, Smolec MJ, McPherson RA, West C, Gelbart T. 1993a. Gaucher disease: gene frequencies in the Ashkenazi Jewish population. Am J Hum Genet 52: 85-88.

Beutler E, Gelbart T, West C. 1993b. Identification of six new Gaucher disease mutations. Genomics 15: 203-205.

Beutler E, Grabowski GA. 1995. Gaucher disease. En: *The metabolic and molecular bases of inherited disease*, 7ª edición, pp 2641-2670. Scriver CR y Beaudet A ed. Mc Graw-Hill, New York.

Beutler E, Gelbart T. 1998. Hematologically important mutations: Gaucher disease. Blood Cells Mol Dis 24 (1): 2-8.

Boas FE. 2000. Linkage to Gaucher mutations in the Ashkenazi population: effect of drift on decay of linkage disequilibrium and evidence for heterozygote selection. Blood Cells Mol Dis 26(4): 348-359.

Bornstein P, McKinney CE, LaMarca ME, Winfield S, Shingu T, Devarayalu S, Vos HL, Ginns EI. 1995. Metaxin, a gene contiguous to both thrombospondin 3 and glucocerebrosidase, is required for embryonic development in the mouse: implications for Gaucher disease. Proc Natl Acad Sci USA 92: 4547-4551.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency-spectrum of DNA Polymorphisms. Genetics 140: 783-796.

Broman KW, Weber JL. 2000. Characterization of human crossover interference. Am J Hum Genet 66: 1911-1926.

Buettner JA, Glusman G, Ben-Arie N, Ramos P, Lancet D, Evans GA. 1998. Organization of olfactory receptor genes on human chromosome 11. Genomics: 53: 56-68.

C

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325: (6099): 31- 6.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7(2): 111-122.

Clark GA, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595-612.

Cleary ML, Schon EA, Lingrel JB. 1981. Two related pseudogenes are the result of a gene duplication in the goat β-globin locus. Cell 26: 181-190.

Collins M, Bornstein P. 1996. SP1-binding elements, within the common metaxin-thrombospondin 3 intergenic region, participate in the regulation of the metaxin gene. Nucleic Acids Res 24: 3661- 3669.

Collins M, Rojnuckarin P, Zhu Y-H, Bornstein P. 1998. A far upstream, cell type-specific enhancer of the mouse thrombospondin 3 gene is located within intron 6 of the adjacent metaxin gene. J Biol Chem 273: 21816- 21824.

Cooper DN, Gerber-Huber S. 1985. DNA methylation and CpG suppression. Cell Differ 17: 199-205.

Cooper DN, Krawczak M. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83: 181-188.

Cooper DN, Krawczak M. 1990. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Hum Genet 85: 55-74.

Cooper DN. 1999. Pseudogenes and their formation, pp 265-296. En: *Human Gene Evolution*, Cooper DN ed. BIOS Scientifics Publishers Limited, Oxford.

D

Dahl N, Hillborg PO, Olofsson A. 1993. Gaucher disease (Norrbottnian type III): probable founders identified by genealogical and molecular studies. Hum Genet 92: 513-515.

De Lorenzo D. 1998. Variabilidad nucleotídica y recombinación: región *l(2)gl* en *Drosophila*. Tesis Doctoral, Universidad de Barcelona.

Deinard A, Dorit R, Castiglione C, Jiang Z, Becker D, Ruddle F, Schugart K, Kidd K. 1999. Evolution of the HOXB6 intergenic region: motif conservartion at the lateral plate mesoderm (LPM) enhancer element. J Exp Zool 285: 170-176.

Demina A, Boas E, Beutler E. 1998. Structure and linkage relationships of the region containing the human L-type pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hematopatol Mol Hematol 11 (2): 63-71.

Devine EA, Smith M, Arredongo-Vega FX, Shafit-Zagardo B, Desnick RJ. 1982. Regional assignment of the structural gene for human acid β-glucosidase to q42→qter on chromosome 1. Cytogenet Cell Genet 33: 340-344.

Dipple KM, McCabe ERB. 2000. Phenotypes of patients with "simple" mendelian disorders are complex traits: thresholds, modifiers and systems dynamics. Am J Hum Genet 66: 1729-1735.

Doll RF, Bruce A, Smith F. 1995. Regulation of the human acid β-glucosidase promoter in multiple cell types. Biochim Biophys Acta 1261: 57-67.

Dorit RL, Akashi H, Gilbert W. 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268: 1183-1185.

Dynan WS. 1986. Promoters for housekeeping genes. TIG: 196-197.

E

Efstratiadis A, Posakoni JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ. 1980. The structure and evolution of the human β-globin gene family. Cell 21: 653-668.

Eto Y, Ida H. 1999. Clinical and molecular characteristics of Japanese Gaucher disease. Neurochem Res 24 (2): 207-11.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. Theor Popul Biol 3: 87-112.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.

Eyal N, Firon N, Wilder S, Kolodny EH, Horowitz M. 1991. Three unique base pair changes in a family with Gaucher disease. Hum Genet 87: 328-332.

Eyal N, Wilder S, Horowitz M. 1990. Prevalent and rare mutations among Gaucher patients. Gene 96: 277-283.

F

Felsenstein J. 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics 5: 164-166.

Filocamo M, Bonuccelli G, Mazzotti R, Giona F, Gatti R. 2000. Identification of a novel recombinant allele in three unrelated Italian Gaucher patients: implications for prognosis and genetic counseling. Blood Cells Mol Dis 26 (4): 307-311.

Finckh U, Seeman P, Von Widdern OC, Rolfs A. 1998. Simple PCR amplification of the entire glucocerebrosidase gene (GBA) coding region for diagnostic sequence analysis. DNA Seq 8 (6): 349- 56.

Fu YX. 1997. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics 147: 915-925.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Fullerton SM, Harding RM, Boyce AJ, Clegg JB. 1994. Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc Natl Acad Sci USA 91: 1805- 1809.

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor ST, Stengård JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67 (4): 881-900.

G

Gilad Y, Segré D, Skorecki K, Nachman MW, Lancet D, Shanon D. 2000. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat Genet 26: 221-224.

Gilbert W. 1978. Why genes in pieces?. Nature 271: 501.

Gimelbrant AA, McClintock TS. 1997. A nuclear matrix attachment region is highly homologous to a conserved domain of olfactory receptors. J Mol Neurosci 9: 61-63.

Ginns E, Choudary PV, Tsuji S, Martin B, Stubblefield B, Sawyer J, Hozier J, Barranger J. 1985. Gene mapping and leader polypeptide sequence of human glucocerebrosidase: implications for Gaucher disease. Proc Natl Acad Sci USA 82: 7101-7105.

Glenn D, Gelbart T, Beutler E. 1994. Tight linkage of pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hum Genet 93: 635-638.

Glew RH, Venkatakrishnan G, Hubbell CA, Beutler E, Geil JD, Lee RE. 1991. A case of nonneurologic Gaucher's disease that biochemically resembles the neurologic types. J Neuropathol Exp Neurol 50 (2): 108-117.

Glusman G, Sosinsky A, Ben-Asher E, Avidan N, Sonkin D, Bahar A, Rosenthal A, Clifton S, Roe B, Ferraz C, Demaille J, Lancet D. 2000. Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. Genomics 63: 227-245.

Gojobori T, Li WH, Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18: 360-369.

Grabowski GA. 1997. Gaucher disease: gene frequencies and genotype/phenotype correlations. Genet Test 1 (1): 5-12.

Grace ME, Ashton-Prolla P, Pastores GM, Soni A, Desnick RJ. 1999. Non-pseudogene-derived complex acid β-glucosidase mutations causing mild type 1 and severe type 2 Gaucher disease. J Clin Invest 103 (6): 817-823.

Graur D, Li WH. 2000. Fundamentals of molecular evolution, 2ª edición. Sinauer Associates Inc., Sunderland, Massachussetts.

Griffiths RC, Tavaré S. 1994. Ancestral inference in population genetics. Stat Sci 9 (3): 307-319.

Griffiths RC, Tavaré S. 1998a. The age of a mutation in a general coalescent tree. Commun Stat 14 (1&2): 273-295.

Griffiths RC, Tavaré S. 1998b. The ages of mutations in gene trees. Ann Appl Prob 9: 567-590.

Grimsley C, Mather KA, Ober C. 1998. HLA-H: a pseudogene with increased variation due to balancing selection at neighboring loci. Mol Biol Evol 15(12): 1581-1588.

H

Haldane JBS. 1933. The part played by recurrent mutation in evolution. Am Nat 67: 5-19.

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22: 239-247.

Hamblin MT, DI Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66 (5): 1669-1679.

Hammer MF. 1995. A recent common ancestry for human Y chromosomes. Nature 378: 376-378.

Harding R , Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60: 772-789.

Harris EE, Hey J. 1999. X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA 96: 3320-3324.

Hatton CE, Cooper A, Whitehouse C, Wraith JE. 1997. Mutation analysis in 46 British and Irish patients with Gaucher's disease. Arch Dis Child 77: 17-22.

Hawley ME, Kidd KK. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86: 409-411.

Hey J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol Biol Evol 14(2): 166-172.

Hey J, Wakeley J. 1997. A coalescent estimator of the population recombination rate. Genetics 145: 833-846.

Hodanová K, Hrebícek M, Cervenková M, Mrázová L, Vepreková L, Zeman J. 1999. Analysis of the β-glucocerebrosidase gene in Czech and Slovak Gaucher patients: mutation profile and description of six novel mutant alleles. Bood Cells Mol Dis 25 (18) 30: 287-298.

Hong CM, Ohashi T, Yu XY, Weiler S, Barranger JA. 1990. Sequence of two alleles responsible for Gaucher disease. DNA Cell Biol 9 (4): 233-241.

Horowitz M, Zimran A. 1994. Mutations causing Gaucher disease. Hum Mutat 3: 1-11.

Horowitz M, Tzuri G, Eyal N, Berebi A, Kolodny EH, Brady RO, Barton NW, Abrahamov A, Zimran A. 1993. Prevalence of nine mutations among Jewish and non-Jewish Gaucher disease patients. Am J Hum Genet 53: 921-930.

Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, Beutler E. 1989. The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 4: 87-96.

Huang W, Fu YX, Chang BHJ, Gu X, Jorde LB, Li WH. 1998. Sequence variation in ZFX introns in human populations. Mol Biol Evol 15 (2): 138-142.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147-164.

Hudson RR. 1987. Estimating the recombination parameter of a finite population model without selection. Genet Res 50: 245-250.

I

Ida H, Rennert OM, Iwasawa K, Kobayashi M, Eto Y. 1999. Clinical and genetic studies of Japanese homozygotes for the Gaucher disease L444P mutation. Hum Genet 105: 120-126.

Imai K, Nakamura M, Yamada M, Asano A, Yokoyama S, Tsuji S, Ginns E. 1993. A novel transcript from a pseudogene for human glucocerebrosidase in non-Gaucher disease cells. Gene 136: 365-368.

J

Jacq C, Miller JR, Brownlee GG. 1977. A pseudogene structure in 5S DNA of Xenopus laevis. Cell 12: 109-120.

Jansen R, Ledley FD. 1990. Disruption of phase during PCR amplification and cloning of heterozygous target sequences. Nucleic Acids Res 18 (17): 5153-5156.

Jaruzelska J, Zietkiewicz E, Labuda D. 1999. Is selection responsible for the low level of variation in the last intron of the ZFY locus?. Mol. Biol. Evol. 16: 1633-1640.

Jaruzelska J, Zietkiewicz E, Batzer M, Cole DEC, Moisan JP, Scozzari R, Tavaré S, Labuda D. 1999. Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. Genetics 152: 1091-1101.

Jeffreys AJ, Barrie PA, Haris S, Fawcett DH, Nugent ZJ. 1982. Isolation and sequence analysis of a hybrid δ-globin pseudogene from the brown lemur. J Mol Biol 156: 487-503.

Jin L, Undrehill P, Doctor V, Davis RW, Shen P, Cavalli-Sforza LL, Oefner PJ. 1999. Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proc Natl Acad Sci USA 96: 3796-3800.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66: 979-988.

K

Kaessman H, Heiβig F, Haeseler A, Pääbo S. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22: 78-81.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics 123: 887-899.

Kim JW, Liou BB, Lai MY, Ponce E, Grabowski GA. 1996. Gaucher disease: identification of three new mutations in the Korean and Chinese (Taiwanese) populations. Hum Mutat 7: 214- 218.

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kirbi DA, Stephan W. 1996. Multi-locus selection and the structure of variation at the white gene of *Drosophila melanogaster*. Genetics 144: 635-645.

Klein RG. 1995. Anatomy, behavior, and modern human origins. Journal of World Prehistory 9: (2): 167-198.

Koprivica V, Stone DL, Park JK, Callahan M, Frisch A, Cohen IJ, Tayebi N, Sidransky E. 2000. Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease. Am J Hum Genet 66: 1777-1786.

Krawczac M, Cooper DN. 1991. Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum Genet 86: 425-441.

L

Latham TE, Grabowski GA, Theophilus BDM, Smith FI. 1990. Complex alleles of the acid β-glucosidase gene in Gaucher disease. Am J Hum Genet 47: 79-86.

Latham TE, Theophilus BDM, Grabowski GA, Smith FI. 1991. Heterogeneity of mutations in the acid β-glucosidase gene of Gaucher disease patients. DNA Cell Biol 10 (1): 15-21.

Lau EK, Tayebi N, Ingraham LJ, Winfield SL, Koprivica V, Stone DL, Zimran A, Ginns EI, Sidransky E. 1999. Two novel polymorphic sequences in the glucocerebrosidase gene region enhance mutational screening and founder effect studies of patients with Gaucher disease. Hum Genet 104: 293-300.

Li WH, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292: 237-239.

Li WH, Wu CI, Luo CC. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21: 58-71.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonimous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2(2): 150-174.

Li WH, Tanimura M, Sharp PM. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25: 330-342.

Li WH, Sadler LA. 1991. Low nucleotide diversity in man. Genetics 129: 513-523.

Li WH. 1997. Molecular Evolution. Sinauer Associates Inc., Sunderland, Massachusetts.

Ligtenberg MJL, Vos HL, Gennissen MC, Hilkens J. 1990. Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini. J Biol Chem 265: 5573-5578.

Long GL, Winfield S, Adolph KW, Ginns EI, Bornstein P. 1996. Structure and organization of the human metaxin gene (MTX) and pseudogene. Genomics 33: 177-184

Luijten M, Wang YP, Smith BT, Westerveld A, Smink LJ, Dunham I, Roe BA, Hulsebos TJM. 2000. Mechanism of spreading of the highly related neurofibromatosis type 1 (NF1) pseudogenes on chromosomes 2, 14 and 22. Eur J Hum Genet 8: 209-214.

M

Marshall CR, Raff EC, Raff RA. 1994. Dollo's law and the death and resurrection of genes. Proc Natl Acad Sci USA. 91: 12283- 12287.

Masuno M, Tomatsu S, Sukegawa K, Orii T. 1990. Non-existence of a tight association between a ^{444}leucine to proline mutation and phenotypes of Gaucher disease: high frequency of a Nci I polymorphism in the non-neuronopathic form. Hum Genet 84: 203-206.

Moran D, Galperin E, Horowitz M. 1997. Identification of factors regulating the expression of the human glucocerebrosidase gene. Gene 194: 201-213.

Motulsky AG. 1995. Jewish diseases and origins. Nat Genet 9: 99-101.

N

Nachman MW, Bauer VL, Crowell SL, Charles FA. 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics 150: 1133-1141.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.

Nagylaki T. 1982. Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics 100: 315-337.

Nei M.1987. Molecular Evolutionary Genetics. Columbia University Press, Inc., New York.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Genetics 76(10): 5369-5273.

Nickerson DA, Tobe VO, Taylor SL. 1997. Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25: 2745-2751.

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19: 233-240.

Nielsen R, Yang Z. 1998. Likelihood for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.

O

Ohno S. 1984. Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. J Mol Evol 20: 313- 321.

Ohta T. 1990. How gene families evolve. Theor Popul Biol 37: 213- 219.

Ohta T. 1983. On the evolution of multigene families. Theor Popul Biol. 23: 216- 240.

Ophir R, Graur D. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene 205: 191-202.

Owens K, King MC. 1999. Genomic views of human history. Science 286: 451-453.

P

Peleg L, Frisch A, Goldman B, Karpaty M, Narinsky R, Bronstein S, Frydman M. 1998. Lower frequency of Gaucher disease carriers among Tay-Sachs disease carriers. Eur J Hum Genet. 6: 185-186.

R

Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambuughin N, Lin M, Watkins S, Bamshad M, Jorde LB, Ramsay M, Jenkins T, Li WH. 1999. High polymorphism at the human melanocortin 1 receptor locus. Genetics 151: 1547-1557.

Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, Bandelt HJ. 1998. Mitochondrial DNA analysis of Northwest African populations reveals genetic exchange with European, Near-Easter, and sub-Saharan populations. Ann Hum Genet 62: 531-550.

Reiner O, Horowitz M. 1988. Differential expression of the human glucocerebrosidase-coding gene. Gene 17: 469-478.

Reiner O, Wigderson M, Horowitz M. 1988. Structural analysis of the human glucocerebrosidase genes. DNA 7(2): 107-116

Reissner K, Tayebi N, Stubblefield BK, Koprivica V, Blitzer M, Holleran W, Cowan T, Almashanu S, Maddalena A, Karson EM, Sidransky E. 1998. Type 2 Gaucher disease with Hydrops Fetalis in an Ashkenazi Jewish family resulting from a novel recombinant allele and a rare splice junction mutation in the glucocerebrosidase locus. Mol Genet Metab 63: 281-288.

Rieder MJ, Taylor SL, Clark AG, Nickerson NA. 1999. Sequence variation in the human angiotensin converting enzyme. Nat Genet 22: 59-62.

Rouquier S, Taviaux S, Trask BJ, Brand-Arpom V, Engh G, Demaille J, Giorgi D. 1998. Distribution of olfactory receptor genes in the human genome. Nat Genet 18: 243-250.

Rozas J, Rozas R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15(2): 174-175.

S

Saccone S, Federico C, Solovei I, Croquette M-F, Della Valle G, Bernardi G. 1999. Identification of the gene-richest bands in human prometaphase chromosomes. Chromosome Res 7: 379-386.

Saitou N, Nei M. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular Cloning, 2ª edición. Cold Spring Harbor Laboratory Press, New York.

Sarria AJ, Giraldo P, Perez-Calvo JI, Pocoví M. 1999. Detection of three rare (G377S, T134P and1451delAC), and two novel mutations (G195W and Rec 1263del55;1342G>C) in Spanish Gaucher disease patients. Hum Mutat 14 (1): 88.

Schneider S, Kueffer JM, Roessli D, Excoffier L. 2000. Arlequin ver.2000: a software environment for the analysis of population genetics data. Geneva, Genetics and Biometry Laboratory. University of Geneva, Switzerland.

Schorderet D, Gartler SM. 1992. Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA 89: 957-961.

Shafit-Zagardo B, Devine EA, Smith M, Arredongo-Vega F, Desnick RJ. 1981. Assignment of the Gene for acid β-Glucosidase to human chromosome 1. Am J Hum Genet 33: 564-575.

Shapira SK, Finnerty VG. 1986. The use of genetic complementation in the study of eukaryotic macromolecular evolution: rate of spontaneous gene duplication at two loci of *Drosophila melanogaster*. J Mol Evol 23: 159-167.

Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, David RW, Cavalli-Sforza LL, Oefner PJ. 2000. Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA 97: 7354-7359.

Shimmin LC , Miller J, Tran HN, Li WH. 1998. Contrasting levels of DNA polymorphism at the autosomal and X-linked visual color pigment loci in humans and squirrel monkeys. Mol Biol Evol 15 (4): 449-455.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

Sinclair G, Choy FYM, Humphries L. 1998. A novel complex allele and two new point mutations in type 2 (acute neuronopathic) Gaucher disease. Blood Cells Mol Dis 24 (20): 420-427.

Slatkin M, Excoffier L. 1996. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 76: 377-383.

Slatkin M.1994. Linkage disequilibrium in growing and stable populations. Genetics 137: 331-336.

Sorge J, Gross E, West C, Beutler B. 1990. High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. J Clin Invest 86: 1137-1141.

Sorge J, West C, Kuhl W, Treger L, Beutler E. 1987. The human glucocerebrosidase gene has two functional ATG initiator codons. Am J Hum Genet 41: 1016-1024.

Sorge J, West C, Westwood B, Beutler E. 1985. Molecular cloning and nucleotide sequence of human glucocerebrosidase cDNA. Proc Natl Acad Sci USA 82: 7289-7293.

Stone DL, Tayebi N, Orvisky E, Stubblefield B, Madike V, Sidransky E. 2000. Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. Hum Mutat 15: 181-188.

Strachan T, Read AP. 1996. Human Molecular Genetics. Bios Scientific Publishers Ltd., Oxford.

Strasberg PM, Skomorowski MA, Warren IB, Hilson WL, Callahan JW, Clarkes JTR. 1994. Homozygous presence of the crossover (fusion gene) mutation identified in a Type II Gaucher disease fetus: is this analogous to the Gaucher knock-out mouse model?. Biochem Med Metab Biol 53: 16-21.

T

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis. Genetics 123: 585- 595.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. Genetics 145: 505-518.

Tayebi N, Cushner S, Sidransky E. 1996. Differentiation of the glucocerebrosidase gene from pseudogene by long-template PCR: implications for Gaucher disease. Am J Hum Genet 59: 740-741.

Tayebi N, Reissner K, Lau E, Stubblefield B, Klineburgess A, Martin B, Sidransky E. 1998. Genotypic heterogeneity and phenotypic variation anomg patients with Type 2 Gaucher's disease. Pediatr Res 43 (5): 571-578.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. 2000. Recent common ancestry of human Y chromosome: evidence from DNA sequence data. Proc Natl Acad Sci USA 97: 7360-7365.

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. 2000. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67: 518- 522.

Tusié-Luna MT, White PC. 1995. Gene conversions and unequal crossovers betwen CYP21 (steroid 21-hydroxylase gene) and CYP21P involve different mechanisms. Proc Natl Acad Sci USA 92: 10796-10800.

U

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7: 996-1005.

Underhill PA, Shen P, Lin A, Passarino G, Yang WH, Kauffman E, Bonné-Tamir B. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet 26: 358-361.

V

Vanin EF, Goldberg GI, Tucker PW, Smithies O. 1980. A mouse α-globin-related pseudogene lacking intervening sequences. Nature 286: 222-226.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. 1991. African populations and the evolution of human mitochondrial DNA. Science 253: 1503-1507.

Vos HL, Devarayalu S, de Vries Y, Bornstein P.1992. Thrombospondin 3 (Thbs3), a new member of the thrombospondin gene family. J Biol Chem 267: 12192-12196.

W

Walsh JB. 1995. Selection and biased gene conversion in a multigene family: consequences of interallelic bias and threshold selection. Genetics 112: 699- 716.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276.

Whitfield LS, Sulston JE, Goodfellow PN. 1995. Sequence variation of the human Y chromosome. Nature 378: 379-380.

Winfield SL, Tayebi N, Martin BM, Ginns EI, Sidransky E. 1997. Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. Genome Res 7: 1020-1026.

Z

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M, Labuda D. 1997. Nuclear DNA diversity in worldwide distributed human populations. Gene 205: 161-171.

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M, Labuda D. 1998. Genetic structure of the ancestral population of modern humans. J Mol Evol 47: 146-155.

Zimmer KP, Le Coutre P, Aerts HMFG, Harzer K, Fukuda M, O'Brien JS, Naim HY. 1999. Intracellular transport of acid β-glucosidase and lysosome-associated membrane proteins is affected in Gaucher's disease (G202R mutation). J Pathol 188: 407-414.

Zimran A, Sorge J, Gross E, Kubitz M, West C, Beutler E. 1990. A glucocerebrosidase fusion gene in Gaucher disease. J Clin Invest 85: 219-222.

Zimran A, Gelbart T, Westwood B, Grabowski GA, Beutler E. 1991. High frequency of the Gaucher disease mutation at nucleotide 1226 among Ashkenazi Jews. Am J Hum Genet 49: 855-859.

DIRECCIONES ELECTRÓNICAS DE INTERÉS

Las direcciones electrónicas que hemos utilizado en este trabajo se listan a continuación:

BASES DE DATOS:

Genbank, http://www.ncbi.nlm.nih.gov/Genbank
GBA humano (Genbank J03059)
GBA en chimpancé (Genbank AF285236)
psGBA humano (Genbank J03060), (Genbank AF267177)
psGBA en chimpancé (Genbank AF272642)
psGBA en gorila (Genbank AF272641)

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim
Enfermedad de Gaucher tipo 1 (OMIM 230800), tipo 2 (OMIM 230900) y tipo 3 (OMIM 231000)

The Human Gene Mutation Database (HGMD), http://www.uwcm.ac.uk/
GBA humano (HGMD 119262)

GeneDis, http://www.tau.ac.il/~racheli/genedis/gaucher/gaucher.html
Enfermedad de Gaucher

Genzyme, http://www.genzyme.com
Células Gaucher

Expasy, http://www.expasy.ch/enzyme/
Enzima glucocerebrosidasa (EC 3.2.1.45)

PROGRAMAS INFORMÁTICOS:

Arlequin ver.2000, http://lgb.unige.ch/arlequin

DnaSP, http://www.bio.ub.es/~julio/DnaSP.html

Genetree, http://www.maths.monash.edu.au/~mbahlo/mpg/gtree.html

Networks, http//www.fluxus-engineering.com/fluxe02.htm

Sites, http://heylab.rutgers.edu/

Test Fs de Fu, http://www.hgc.sph.uth.tmc.edu/fu/

APÉNDICE

Durante el tiempo que ha durado la preparación del trabajo de tesis doctoral propiamente dicho, de forma paralela y totalmente independiente llevamos a cabo otro estudio, que presentamos a continuación. El objetivo era obtener la secuencia de DNA de los exones del gen TYR en gorilas, y tratar de entender las causas del albinismo en "Copito de Nieve", el único gorila albino conocido hasta el momento. TYR codifica para la proteína tirosinasa, el enzima clave en la producción de melanina.

A continuación presentamos la publicación resultante de este estudio.

TYROSINASE GENE IN GORILLA AND THE ALBINISM OF "SNOWFLAKE"

Rosa Martínez-Arias, David Comas, Aida Andrés, Maria-Teresa Abelló, Xavier Domingo-Roura, Jaume Bertranpetit

Publicado en Pigment Cell Research:

Martinez-Arias R, Comas D, Andres A, Abello MT, Domingo-Roura X, Bertranpetit J. "The tyrosinase gene in gorillas and the albinism of 'Snowflake'". ***Pigment Cell Research***. 2000 Dec;13(6):467-70

TYROSINASE GENE IN GORILLA AND THE ALBINISM OF "SNOWFLAKE"

Rosa Martínez-Arias[1], David Comas[1], Aida Andrés[1], Maria-Teresa Abelló[2], Xavier Domingo-Roura[1, 3], Jaume Bertranpetit[1].

(1) Unitat de Biologia Evolutiva
Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra
Doctor Aiguader 80
08003 Barcelona, Spain
(2) Parc Zoològic de Barcelona
Parc de la Ciutadella
08003 Barcelona, Spain
(3) Wildlife Conservation Research Unit
Department of Zoology, University of Oxford
South Park Road
OX13PS Oxford, United Kingdom

RUNNING TITLE: Tyrosinase gene and albinism in the gorilla

Address for correspondence:
Jaume Bertranpetit (1)
Phone: 34-93-5422840; Fax number:34-93-5422802
E-mail: jaume.bertranpetit@cexs.upf.es

ABSTRACT

The sequence of the tyrosinase (*Tyr*) gene coding tracts has been obtained for the gorilla (*Gorilla gorilla gorilla*). The five exons of the gene were sequenced in three gorillas and a non-albino human. Tyrosinase gene has been found to be a very conservative locus with a very low substitution rate. Some nucleotide and aminoacid differences were found between the gorilla and human tyrosinase coding sequences. One of the gorillas included in the study is the only known case of albinism in a gorilla (Snowflake). Mutations on the *TYR* gene lead to Oculocutaneous 1 albinism, the most common type of albinism in humans (OMIM accession number 203100). *TYR* gene encodes for the tyrosinase enzyme (E.C. 1.14.18.1), whose activity was found to be completely inhibited in Snowflake, indicating that a mutation in the *Tyr* gene is the likely cause of his albinism. Nonetheless, no nucleotide changes were detected that could account for the lack of *Tyr* product or tyrosinase activity in Snowflake. Explanations of these findings are discussed.

KEY WORDS: TYR gene, gorilla, albinism, Snowflake.

Among the genes involved in melanin synthesis that have been isolated, the tyrosinase gene family has been widely studied. The product of the *TYR* gene is the tyrosinase protein (E.C. 1.14.18.1), a melanocyte specific enzyme. Tyrosinase catalyzes three distinct reactions in the melanin byosynthetic pathway and is the key enzyme on it. The *TYR* locus was mapped at human chromosome 11q14-21 using *in situ* hybridization (1). Both the human and mouse tyrosinase genes are estimated to span more than 70 kilobases (kb) (2) and both contain five exons ranging in size from 148 to 819 base pairs (bp). A second tyrosinase locus, *TYRL* or tyrosinase-like locus, has been mapped at 11p11.2-cen in humans, and consists of a truncated tyrosinase gene having only exons 4 and 5, highly homologous to the *TYR* exons (3). No transcription product of the *TYRL* locus is detectable in human melanocytes, and it is considered to be a pseudogene. Analysis of non-human primates has shown that gorilla genome contains the *TYRL* locus while, unexpectedly, it is absent in the chimpanzee (4).

The aim of the present study is to obtain the transcribed DNA sequence of the tyrosinase gene in the gorilla. Interestingly, in the present study we have included the only known albino gorilla, called *Floquet de Neu* (Snowflake) (Fig.1). Snowflake was captured in Equatorial Guinea in October 1966, when tooth development indicated an age of approximately two years. His hair was white, his skin pink without pigmented nevi or freckles, and his eyes blue. In November 1966 he arrived at the Zoologic Park of Barcelona, where a complete dermatological examination and skin biopsies were performed. Nor visible pigment neither generalised genetic abnormalities were found. Melanocytes in the lower epidermis and hair bulbs showed only stage I and II premelanosomes, a characteristic trait of tyrosinase-negative oculocutaneous albinism (OCA) (5).

Hair bulbs from Snowflake, two normally pigmented unrelated gorillas (Ndengue-G1; and Xebo-G4), and one normally pigmented daughter (Kena-G3) were analysed. The DOPA oxidase assay, that measures the conversion of DOPA to DOPAchrome catalysed by tyrosinase, was performed according to (6). The specific activity of the enzyme, measured as A475/mg protein, was zero (lack of tyrosinase activity) for Snowflake, and 7.1, 6.45 and 2.02 for G1, G3, and G4, respectively. The lack of enzymatic function in Snowflake points to an alteration in the gene coding for the tyrosinase protein itself, the *Tyr* locus, as it is the case in most affected humans.

Subsequently, DNA was extracted from fresh blood of Ndengue-G1, a daughter of Snowflake (Machinda-G2), and Snowflake using a standard phenol-chloroform extraction method. DNA from a non-albino human of European descent without family history of albinism was also included in the study. Each of the five exons (including the 3'UTR of the exon 5) was amplified completely and independently from 100 ng of genomic DNA with oligonucleotides previously reported and designed from human sequences (7). The cycling conditions were as follows: a denaturation step of 94 ºC for 5 min; 30 cycles of 94 ºC for 45 s, 50 ºC for 45 s, 72 ºC for 45 s and a final elongation step of 72 ºC for 5 min. The amplification products were purified

using Gene Clean (BIO 101), and the sequence reactions were performed with the same primers mentioned above on each strand, by means of the ABI PRISM dRhodamine Terminator Cycle Sequencing Ready Reaction Kit with Ampli Taq polymerase (PE Biosystems). The human known sequence (GenBank accession number M27160) was used as a reference.

The complete sequence of the tyrosinase exons comprising a total of 1,878 base pairs was obtained for all the individuals (Table 1). All the differences found between the human sequence obtained, in relation to the GenBank *TYR* sequence, are synonymous except for a single aminoacid change that has been detected in position 174. This change has been already described as a potentially harmless substitution (8), and is not associated to any pigmentation disorder. Nor insertions or deletions have been found, and thus the codon numbering is the same for humans and gorillas.

Molecular evolutionary studies have shown that gorilla and human nuclear DNA differ by an average of 2.1% at the nucleotide level (9,10). In the present work the tyrosinase locus shows a high homology degree between gorillas and humans, with a 0.21% difference (four fixed substitutions in 1,878 bp, see Table 1), pointing to a high conservation of the tyrosinase protein. The observed number of substitutions is significantly less than that expected for the average human-gorilla divergence ($\chi_1^2 = 32.5$, $p \approx 0$). According to Li (1997) (11), we can estimate the substitution rate (r_{Tyr}) of this region as:

$r\,Tyr = K Tyr / 2\,T_S = 0.0021/(2 \times 7 \times 10^6) = 0.15 \times 10^{-9}$ substitutions per year, where K_{Tyr} is the number of substitutions per site between the two species in the tyrosinase locus, and T_S is the time of divergence between humans and gorillas (assumed to be 7 million years). This estimate is one order of magnitude lower than that obtained for other autosomal loci (12, 13, 14).

From the four fixed differences found between humans and gorillas, three are synonymous and only one, in nucleotide position 893, produces an aminoacid change (arginine in humans and glutamine in gorillas). Interestingly, this position has also a different aminoacid (leucine) in mouse, suggesting that this substitution is not crucial to the enzymatic activity. Another not fixed aminoacid difference is found between gorillas and humans: nucleotide 1306 (exon 4, codon 418) a heterozygote position results in two codons codifiying for different aminoacids (glycine in humans, and glycine and serine in gorillas), and thus two different types of enzymes are produced in a single individual. This fact is not unusual: in humans about 24,000 to 40,000 non-synonymous substitutions have been estimated to be found in heterozygosity (15). The fact that two non-related individuals are heterozygotes makes it likely that both alleles are frequent in gorillas. Apart from the above mentioned changes, all the other nucleotidic differences between gorillas and humans are transitions in the third position of a codon.

In exons 4 and 5 a high number of heterozygote positions were observed in the samples. The likeliest explanation is that the exons of the *Tyrl* locus, and the exons 4 and 5 of the *Tyr* locus itself, are both amplified with the same primer set. Therefore the sequence obtained for

exons 4 and 5 could be the result of the mixture of both PCR products (*Tyr* and *Tyrl*). That makes it not possible to differentiate, in exons 4 and 5, between real *Tyr* locus heterozygotes and artifacts produced by coamplification of the *Tyr* and *Tyrl* loci.

The only particular difference of Snowflake *Tyr* locus in relation to both pigmented gorillas (G1 and G2) is located in nucleotide 627 of the first exon, a transition in heterozygosity that does not produce aminoacid change. Considering the possibility that his albinism could be the result of the effect of a compound heterozygote (as it is the case in many human individuals), each of the heterozygote positions observed in Snowflake has been carefully checked, but all of them are either also present in the other gorillas or affecting the third position of a codon, and therefore the hypothesis of finding a compound heterozygote can be rejected, as for coding tracts are concerned. Also the putative heterozygote sites found in exons 4 and 5 (due to a possible coamplification of *Tyr* and *Tyrl*) were checked. None of these substitutions, if actually present in the *Tyr* locus, would be responsible for the albinism of Snowflake, since they do not imply any alteration in the tyrosinase protein. Nevertheless, it seems clear that Snowflake is a compound heterozygote, with two different mutations in other than coding tracts, given the fact that several heterozygote positions are found along the *Tyr* exons, and therefore a mutation in homozygosity is unlikely.

None of the changes observed in Snowflake´s *Tyr* exons sequence, with regard to humans or the non-albino gorillas, is responsible for his albino condition. As a possible genetic alteration, we could also reject a non-homologous recombination between the *Tyr* and *Tyrl* loci, which would lead to a truncated gene. This alteration could have not been detected by PCR, since we would always obtain amplification for one of the alleles for those exons missing in the truncated gene (three first exons). However, this can be excluded with the sequence obtained, since no heterozygous sites would have been found for exons 1, 2 and 3. Hence, the attention of further research must be focused on some alternative explanations: one of the several *cis* regulatory elements located upstream of the translational start site could be altered, or a splicing site in any of the introns could be affected, skipping exon sequences or retaining intron fragments. In this sense, splicing defects have been estimated to make up between 8 and 15 % of all single base pair substitutions causing human genetic disease, and therefore its importance should not be overlooked (16). As another possible explanation, a mutation could be placed in the exonic sequence complementary to the amplification primers, as the primers used for the amplification of each of the five TYR exons are located at the beginning of each exonic tract (7).

We have shown that TYR is a very conserved locus, with a low substitution rate. On the other hand, the cause of the albinism in Snowflake, the only known albino gorilla, still remains an open question. The product of the *TYR* locus, the tyrosinase enzyme, is essential in the melanin biosynthetic pathway, and therefore its inactivation can be the direct responsible for the albinism in the gorilla. We have ascertained that there is no mutation responsible of the albinism in the

exons of the *Tyr* locus. Molecular analysis of the introns and 3' untranslated region, as well as mRNA transcription and expression analysis, could shed further light to the comprehension of albinism in the gorilla.

REFERENCES

1. Barton DE, Kwon BS, Francke U. Human Tyrosinase gene, mapped to chromosome 11 (q14-q21), defines second region of homology with mouse chromosome 7. Genomics 1988; 3: 17-24.
2. Shibahara S. Mutations of the tyrosinase gene in oculocutaneous albinism. Pigment Cell Research 1992; 5: 279-283.
3. Giebel LB, Strunk KM, Spritz RA. Organization and nucleotide structure of the human Tyrosinase gene and a truncated tyrosinase-related segment. Genomics 1991; 9: 435-445.
4. Oetting WS, Stine OC, Townsend D, King RA. Evolution of the Tyrosinase related gene (TYRL) in primates. Pigment Cell Res 1993; 6: 171-177.
5. Ferrer L, Fernández-Figueras MT, Castells A, Fernández J. Albinism in a gorilla (*Gorilla gorilla gorilla*). Proc Am Acad Vet Derm abstracts 1997.
6. Petrescu SM, Petrescu AJ, Titu HN, Dwek RA, Platt FM. Inhibition of *N*-glycan processing in B16 melanoma cells results in inactivation of tyrosinase but does not prevent its transport to the melanosome. J Biol Chem 1997; 272: 15796-15803.
7. Giebel LB, Strunk KM, King RA, Hanifin JM, Spritz RA. A frequent tyrosinase gene mutation in classic, tyrosinase-negative (type IA) oculocutaneous albinism. Proc Natl Acad Sci USA 1990; 87: 3255-3258.
8. Chintamaneni CD, Halaban R, Kobayashi Y, Witkop CJ, Kwon BS. A single base insertion in the putative transmembrane domain of the tyrosinase gene as a cause for tyrosinase-negative oculocutaneous albinism. Proc Natl Acad Sci USA 1991; 88: 5272-5276.
9. Ruvolo, M. A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. Mol Phylogenet Evol 1996; 5: 202-219.
10. Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 1998; 9: 585-598.
11. Li W-H. Molecular Evolution. Sunderland Massachusetts: Sinauer Associates, Inc., Publishers; 1997.
12. Li WH, Tanimura M, Sharp PM, 1987 An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25: 330-342.
13. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans. Am J Hum Genet 1997; 60: 772-789.
14. Clark, G.A., K.M. Weiss, D.A. Nickerson, S.L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, C.F. Sing. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595-612.

15. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterisation of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999; 22: 231-238.

16. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 1992; 90: 41-54.

ACKNOWLEDGEMENTS

This research has been possible thanks to the cooperativeness of the authorities of the Barcelona Zoologic Park, E. Tomàs and F. Costa. We also would like to thank J. Fernández, J. Xampeny, and J. Fàbregues, for their help in obtaining gorilla blood and hair samples and the photograph of the albino gorilla. We are grateful to F. Calafell for helpful suggestions and comments, and to M. Vallés for her technical assistance. We thank S.M. Petrescu, A.J. Petrescu, and F.M. Platt for performing the DOPA oxidase assay. R.M-A. received a fellowship from the Spanish Ministry of Education and Culture (AP96). This research was supported by Dirección General de Investigación Científica y Técnica in Spain (PM98-018 and PB98-1064), and by Direcció General de Recerca, Generalitat de Catalunya (1998SGR0009).

FIGURE LEGENDS

Figure 1: Snowflake, the albino gorilla, in the Zoologic Park of Barcelona. In the insert, Snowflake two years old. Photographs courtesy of the Zoologic Park of Barcelona.

Table 1: Aminoacid and nucleotide changes found in the TYR coding sequences [1]

EXON	POSITION	G1		G2		SNOWFLAKE		HUMAN		GenBank	
1[2]	nt: 447 Aa: 131	TA**C**	Tyr	TA**C**	Tyr	TA**C**	Tyr	TAT	Tyr	TAT	Tyr
(819 bp)	nt: 575 Aa: 174	TCT	Ser	TCT	Ser	TCT	Ser	T**A**T	**Tyr**	TCT	Ser
	nt: 627 Aa: 191	CCT	Pro	CCT	Pro	CC**T/C**	Pro	CCT	Pro	CCT	Pro
2	nt: 891 Aa: 279	TT**A/G**	Leu	TTA	Leu	TT**A/G**	Leu	TTA	Leu	TTA	Leu
(217 bp)	nt: 893 Aa: 280	C**A**G	**Gln**	C**A**G	**Gln**	C**A**G	**Gln**	CGG	Arg	CGG	Arg
3	nt: 1065 Aa: 337	GC**A**	Ala	GC**A**	Ala	GC**A**	Ala	GCG	Ala	GCG	Ala
(148 bp)	nt: 1152 Aa: 366	CC**C**	Pro	CC**C**	Pro	CC**C**	Pro	CCT	Pro	CCT	Pro
4	nt: 1236 Aa: 394	CC**A/G**	Pro	CC**A/G**	Pro	CC**A/G**	Pro	CC**A/G**	Pro	CCA	Pro
(182 bp)	nt: 1305 Aa: 417	AA**C/T**	Asn	AA**C/T**	Asn	AA**C/T**	Asn	AA**C/T**	Asn	AAT	Asn
	nt: 1306 Aa: 418	**G/A**GT	Gly/**Ser**	**G/A**GT	Gly/**Ser**	**G/A**GT	Gly/**Ser**	GGT	Gly	GGT	Gly
5 [3]	nt: 1413 Aa: 453	GC**A**	Ala	GC**A**	Ala	GC**A**	Ala	GC**A/G**	Ala	GCG	Ala
(512 bp)	nt: 1446 Aa: 464	GC**G/C**	Ala	GCG	Ala	GC**G/C**	Ala	GC**G/C**	Ala	GCG	Ala
	nt: 1569 Aa: 505	AG**T/C**	Ser	AG**T/C**	Ser	AG**T/C**	Ser	AG**T**	Ser	AGC	Ser

(1) The differences are with regard to the human GenBank sequence (accession number M27160), and are shown in bold. Abbreviations: nt: nucleotide; Aa: aminoacid; bp: base pairs. The nucleotide sequences reported in this paper have been submitted to GenBank and have been assigned the accession numbers: 321050, 321054, 321068, 321069, 321071, 321081, 321083, 321092, 321095, 321093, 321103, 321108, 321117, 321122, 321128, 321127, 321134, 321133, 321138, and 321630. Nucleotides are numbered from the first nucleotide of the ATG initiation codon, and aminoacids from the amino terminal aminoacid of the tyrosinase mature protein, according to (8). Nucleotide substitutions found in the 3' untranslated region of exon 5 have not been included in the table, since they are not expected to play any role in the enzymatic activity

(2) The length of each exon is indicated in brackets.