FACULTAD DE CIENCIAS EXACTAS Y NATURALES
UNIVERSIDAD DE BUENOS AIRES
Departamento de Ecología, Genética y Evolución

REPETICIONES EXACTAS EN EL GENOMA HUMANO: UNA INTERPRETACIÓN BIOLÓGICA BASADA EN SU RELACIÓN CON LAS DUPLICACIONES SEGMENTALES

Autor: **LEANDRO MIQUET**

Director: **Prof. Dr. ESTEBAN HASSON**
Codirectora: **Prof. Dra. VERÓNICA BECHER**
Director Asistente: **Dr. JULIÁN MENSCH**

Tesis para optar por el título de Licenciado en Ciencias Biológicas

OCTUBRE DE 2011

Lugares de trabajo: Laboratorio de Evolución (EGE) y Grupo Kapow (DC)

ÍNDICE

COLABORADORES

Como autor de esta Tesis quiero destinar este espacio para agradecer a quienes colaboraron con el trabajo en alguna de sus múltiples instancias, dando crédito a sus aportes.

A mis directores, los Dres. Esteban Hasson y Julián Mensch (Laboratorio de Evolución) y la Dra. Verónica Becher (grupo Kapow), por haberme ofrecido la posibilidad de desarrollar esta Tesis y guiarme a lo largo del trabajo.

Al Lic. Pablo Barenbaum, (grupo Kapow) por haber llevado a cabo los cálculos de cobertura de bases de datos, sin los cuales no podría haberse concebido el Capítulo 3 de esta Tesis.

Al Dr. Hernán Dopazo (Centro de Investigaciones Príncipe Felipe, Valencia) por sus aportes a la discusión de nuestros resultados.

Al Dr. Guillermo Folguera por su asesoramiento epistemológico.

A Gabriela Russo por sus aportes a la revisión del manuscrito.

A la Dra. Paula Cramer (LFBM) por su somero pero acertado asesoramiento en etapas tempranas de la investigación.

RESUMEN

El grupo de investigación Kapow, del Departamento de Computación de esta Facultad, desarrolló un algoritmo capaz de hallar todas las repeticiones exactas (en adelante patterns) existentes cualquier secuencia de letras. Al aplicar este algoritmo al genoma humano se observó que existe una gran cantidad de patterns llamativamente largos, que cubren gran parte de su extensión. Muchas de estas repeticiones idénticas no corresponden a ningún elemento genómico reportado previamente. La motivación de este trabajo fue elaborar una interpretación biológica de estas secuencias repetidas en el genoma humano. Para abordar este problema se trabajó sobre un modelo que pretende explicar el origen de los patterns sin necesidad de conocer su posible función biológica. Nuestra hipótesis postula que los patterns deben su origen a la duplicación de bloques genómicos más grandes, que los contienen. Estos bloques genómicos caen dentro de lo que se ha definido como duplicaciones segmentales (en adelante DS). Se implementaron dos estrategias metodológicas complementarias. La primera de ellas consistió en tomar un conjunto de patterns escogido *a priori* y evaluar si su comportamiento se ajustaba a las predicciones de nuestro modelo. Se observó que el patrón de ocurrencia en el genoma del conjunto de patterns estudiado es consistente con que el origen de una parte de ellos se deba a la generación de DS. El análisis filogenético de los homólogos de esas DS dentro del Orden Pimates mostró que presentan un patrón evolutivo que las valida como verdaderos elementos duplicados a lo largo de la historia evolutiva del grupo. La segunda estrategia metodológica consistió en tomar la totalidad de los patterns computados por el grupo Kapow y evaluar qué proporción de ellos cae dentro de las bases de datos de DS existentes. Se observó que la proporción de patterns contenidos en DS aumenta al incrementarse su longitud: cuanto más grande es un pattern, más probable es que esté contenido en una DS.

El primer abordaje permitió establecer que el modelo de duplicación-divergencia de grandes bloques genómicos es capaz de explicar el origen de al menos una parte de los patterns del genoma humano. Mediante el segundo abordaje se pudo evaluar cualitativamente el poder explicativo del modelo a escala genómica, pudiéndose concluir que es directamente proporcional a la longitud de los patterns.

El modelo de duplicación-divergencia de grandes bloques genómicos puede dar cuenta del origen de buena parte de los patterns del genoma humano, aunque sin establecer un mecanismo unívoco para explicar la conservación de los mismos a lo largo del tiempo.

1- INTRODUCCIÓN

1.1- Antecedentes

1.1.1- Marco teórico: secuencias repetidas en el genoma humano

La existencia de ADN repetitivo en genomas eucariotas fue postulada a partir de estudios de reasociación varios años antes de que se desarrollaran las técnicas de secuenciación, y varias décadas antes de que se completara el primer proyecto de un genoma eucariótico [5]. El concepto de ADN basura (*junk DNA*, concebido casi veinte años antes de que se iniciara el Proyecto Genoma Humano [25]) aglutinaba a estas secuencias repetitivas y al denominado ADN "egoísta" o "parásito", todos ellos elementos portadores de información presuntamente inútil [26].

En el año 2001 se publicó el primer borrador de la secuencia del genoma humano, cuyo análisis preliminar tuvo como uno de sus resultados más notorios la alta cantidad de secuencias repetidas presentes [15]. La clasificación dentro de la que caen estos elementos es amplia: repeticiones originadas por distintos tipos de elementos transponibles (de alto número de copias, que cubrirían cerca de la mitad de la extensión del genoma), repeticiones de secuencia simple (satélites) y duplicaciones segmentales (en adelante DS). El Consorcio Internacional de Secuenciación del Genoma Humano incluye en este último grupo a bloques de secuencia genómica cuyo tamaño varía entre 1 y 200 kb que aparecen duplicados en una o más ubicaciones en el genoma. Según una primera estimación de abundancia en el genoma, confirmada posteriormente mediante metodologías de detección específicas [1; 6], las DS cubrirían un 3,6 % de la secuencia genómica completa. Actualmente, las DS están definidas como bloques genómicos de más de 1 kb de longitud, que aparecen al menos dos veces en el genoma con una similitud mayor o igual al 90 % [22].

La duplicación de segmentos genómicos se ha propuesto como un mecanismo importante en la evolución de los genomas para explicar la aparición de familias multigénicas

a partir de genes ancestrales [24]. Desde esta perspectiva, el esfuerzo por estudiar secuencias parálogas estaba enfocado en las regiones codificantes. Por su parte, el interés por estudiar la variación poblacional estaba centrado en polimorfismos nucleotídicos. En un informe de 1998, los grupos de planeamiento del programa estadounidense de investigación del genoma humano mencionaban el desarrollo de tecnologías para la detección de polimorfismos de nucleótido simple (SNP), la elaboración de mapas de SNP, la creación de bases de datos de cDNA y la evaluación de SNP en regiones codificantes entre los principales objetivos para el último tramo del proyecto. La principal motivación tenía que ver con la posibilidad de asociar estas variantes a fenotipos patológicos, ya sea con miras a una aplicación clínica (diagnóstico basado en secuencias de ADN) o básica (estudio de las bases genéticas de patologías). Aunque en ese mismo documento se remarca que "otros tipos de variación, como en el número de copias, inserciones, deleciones, duplicaciones y rearreglos también existen, pero en baja frecuencia y [de] distribución pobremente comprendida" [8], los estudios de DS ya habían permitido la acumulación de una buena cantidad de información, en un principio exclusivamente mediante técnicas citogenéticas, hibridización o utilizando bibliotecas construidas en distintos vectores, etc. [10, 35], y más adelante a través del análisis de secuencias genómicas publicadas en bases de datos [11, 32]. Se describieron muchos casos de DS asociadas a diversos desórdenes genómicos [2, 9, 16, 22], e incluso se logró identificar las bases genéticas de ciertas patologías a partir de mapas de DS [30].

Además de sus implicancias en cuestiones relacionadas con la salud, las DS han despertado interés desde el punto de vista evolutivo. La generación de duplicaciones se postuló como mecanismo capaz de responder a preguntas pendientes sobre la evolución de los genomas, siendo la falta de correlación entre la divergencia fenotípica y a nivel de secuencia entre humano y chimpancé uno de los ejemplos más conocidos [18]. Para el grupo de los grandes simios se han identificado familias de DS exclusivas de cada linaje, observándose una aceleración de rearreglos genómicos en el humano, respecto a otro tipo de mutaciones (puntuales y actividad de transposición) [2, 6, 21, 22]. Se cree que los efectos de la dinámica de rearreglos podrían explicar parte de las diferencias en niveles de expresión reportadas entre chimpancé y humano [4, 12].

Actualmente existen dos metodologías *in silico* para la detección de DS [22]:

i) Comparación a partir del genoma completo ensamblado (*whole-genome assembly comparison*): se parte del genoma completamente ensamblado, se lo divide en fragmentos de 400 kb, se eliminan las repeticiones de alto número de copias y se comparan los fragmentos entre sí mediante alineamientos de a pares [1].
ii) Detección a partir de datos de secuenciación escopeta (*whole-genome shotgun detection*): se parte de todos los fragmentos (*reads*) obtenidos por la secuenciación del genoma mediante estrategias tipo escopeta (*whole-genome shotgun*) y se los alinea de a pares contra una secuencia de referencia, que puede no estar completamente ensamblada [2].

Cada uno de estos métodos tiene sus ventajas y desventajas, pero ambos tienen en común que utilizan alineamientos de secuencias, en general locales utilizando el algoritmo BLAST o derivados. La aplicación de estas metodologías a genomas de homínidos ha permitido: construir mapas de DS de cada genoma; identificar DS propias de cada linaje y datar los eventos de duplicación; estudiar la distribución de DS a lo largo del genoma; establecer correlaciones entre DS y densidad de genes; determinar el tipo de genes en los que están enriquecidas las DS del linaje humano y establecer si existe alguna correlación entre famlias de DS y tasas de evolución adaptativa [2, 6, 21, 31].

1.1.2- Problema biológico

Recientemente, el grupo de investigación de algoritmos sobre palabras/secuencias (Kapow) del Departamento de Computación de esta Facultad, dirigido por la Profesora Doctora Verónica Becher, desarrolló un algoritmo capaz de hallar repeticiones exactas en cualquier secuencia finita de caracteres. Este algoritmo (en adelante nos referiremos a él como *Findpat* o *el algoritmo del grupo Kapow*) se basa en el concepto de patrón (en adelante *pattern*), que está definido como “secuencia finita de letras que ocurre más de una vez de

manera idéntica, y cuyas extensiones ocurren menos veces". Por ejemplo, la secuencia *abcdeabcdfbcdebcd* contiene a los patterns *abcd*, *bcde* y *bcd*, que tienen 2, 2 y 3 ocurrencias, respectivamente. Aunque *bc* aparece repetido 3 veces, no es un pattern porque su extensión *bcd* ocurre la misma cantidad de veces. La aplicación de Findpat a un genoma permite hallar absolutamente todos los patterns de cualquier longitud presentes en él [3]. Cuando hablemos de patterns en un genoma nos referiremos a una secuencia explícita sin denotar su ubicación, y llamaremos *ocurrencia* a cada una de las ubicaciones en el genoma donde aparezca dicha secuencia.

Al aplicar Findpat a una de las hebras del genoma humano en búsqueda de patterns de al menos 40 letras se obtuvieron resultados inesperados, entre los cuales se destacan los siguientes:

- Gran parte del genoma está cubierto por secuencias de longitud mayor a 40 que aparecen repetidas de manera exacta al menos dos veces, y existen regiones cromosómicas extensas cuya cobertura es cercana al 100% (Figura 1)
- Cada cromosoma comparte el 10% de su secuencia en patterns de 40 o más letras con el resto de los cromosomas
- El pattern más largo es una secuencia de 67632 nucleótidos, que aparece dos veces en el cromosoma 1
- Existen patterns cuyas ocurrencias no se solapan con ningún gen ni repetición reportados en bases de datos (nos referiremos a ellos como *novel patterns*)

El presente trabajo de Tesis de Licenciatura debe su origen al interés de los miembros del grupo Kapow en hallar una interpretación biológica de los resultados de la aplicación de sus algoritmos al genoma humano. La motivación principal es comprender el significado de estas secuencias repetidas en forma idéntica a lo largo del genoma. ¿A qué se debe esta alta similitud? ¿Están altamente conservadas por cumplir una función biológica importante, o su existencia es una consecuencia de otros procesos relacionados con la dinámica evolutiva del genoma?

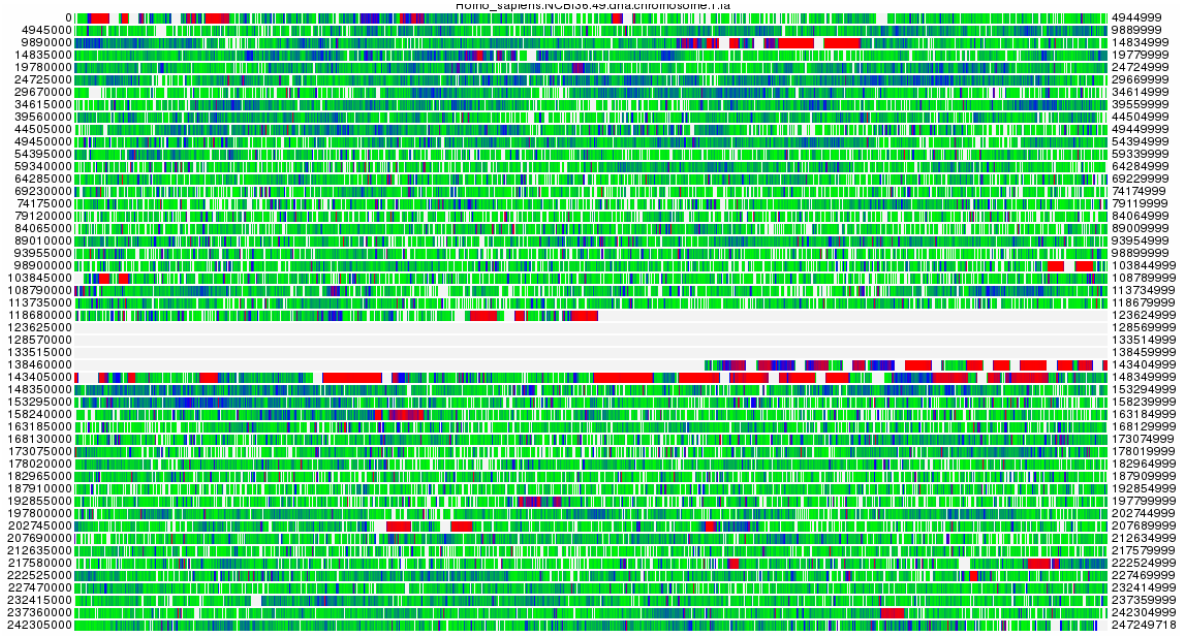

Figura 1. Cobertura del cromosoma 1 por patterns de longitud mayor a 39 nucleótidos. Los números denotan las posiciones en el cromosoma. Cada píxel representa 5000 posiciones en el genoma. Rojo: 100% de cobertura. Gris: 0% de cobertura. Colores intermedios: porcentaje de cobertura intermedios. Tomado de Becher V., *et al.* 2009. Efficient computation of all perfect repeats in genomic sequences of up to half a Gigabyte, with a case study on the Human genome. *Bioinformatics*25: 1746-53.

Estos interrogantes despiertan especial interés en el caso de los novel patterns, que cubren el 0,85% del cromosoma 1 y pueden tener longitudes superiores a 500 nucleótidos, ya que resultan buenos candidatos para la búsqueda de funciones biológicas no descritas.

Este trabajo se propone atacar el problema de los patterns respondiendo algunas de estas preguntas.

1.1.3- Definiciones

A continuación se recapitulan las definiciones de aquellos elementos mencionados entre nuestros antecedentes que se utilizarán a lo largo del trabajo:

Duplicación segmental (DS): bloque genómico de más de 1 kb de longitud, que aparece al menos dos veces en el genoma con una similitud mayor o igual al 90%.

Pattern: secuencia que ocurre más de una vez de manera idéntica, y cuyas extensiones ocurren menos veces.

Ocurrencias de un pattern: ubicaciones en el genoma de un pattern.

Novel pattern: pattern que cumple con que ninguna de sus ocurrencias se solapa con algún elemento de las bases de datos de genes y repeticiones de Ensembl (compiladas por el Dr. Javier Herrero, EMBL-EBI).

1.2- Objetivos e hipótesis

El objetivo general de este trabajo es elaborar una interpretación biológica de las repeticiones exactas en el genoma humano, identificadas por el grupo Kapow mediante su algoritmo de búsqueda de patterns.

La primera interpretación que surgió para este fenómeno fue que muchas de estas secuencias (al menos las más largas) estuvieran repetidas de manera idéntica debido a una fuerte presión selectiva sobre su función biológica. Sin embargo, se decidió abordar el problema haciendo uso de un modelo teórico que fuera independiente de la posible funcionalidad de las secuencias estudiadas.

La hipótesis de trabajo es la siguiente: “los patterns existen en el genoma como consecuencia de la duplicación y divergencia de grandes bloques genómicos, que los contienen”

Los objetivos particulares son:

Objetivo 1: determinar si la hipótesis permite explicar el origen de al menos una parte de los patterns detectados en el genoma.

Objetivo 2: realizar una validación evolutiva de los resultados obtenidos en la puesta a prueba de la hipótesis. Si los patterns deben su existencia a eventos de duplicación, entonces debería poder inferirse la historia evolutiva de tales eventos, tanto dentro del linaje humano como en ramas más profundas de la filogenia.

Objetivo 3: determinar el poder explicativo de la hipótesis a escala genómica (*i.e.* qué proporción del total de los patterns son explicados por la hipótesis)

Para la puesta a prueba del modelo postulado en la hipótesis se propusieron dos abordajes distintos y complementarios:

A partir de casos particulares: utilizando conjuntos de patterns escogidos según un criterio establecido *a priori*. Mediante este abordaje se pretende atender los Objetivos 1 y 2.

A partir de bases de datos completas: utilizando todos los patterns detectados por el grupo Kapow y las bases de datos de DS de acceso público. Mediante este abordaje se pretende atender el Objetivo 3.

2- PUESTA A PRUEBA DE LA HIPÓTESIS A PARTIR DE CASOS PARTICULARES

Mediante este primer abordaje se pretende determinar si existen patterns cuyo origen pueda ser explicado por la hipótesis, respondiendo así al Objetivo 1. En esta instancia no se busca estudiar el poder explicativo del modelo, sino hallar únicamente un conjunto de patterns que se ajuste a sus predicciones.

2.1- Predicciones

Para esta primera puesta a prueba de la hipótesis se hace uso de cuatro de las predicciones que pueden desprenderse de ella (Figura 2):

Pr1) Deben encontrarse conjuntos de patterns dependientes entre sí: que aparezcan asociados en un orden y a una distancia determinados, formando un *patrón de patterns* (en adelante PP)

Pr2) Un pattern que forma parte de un PP no debe ocurrir nunca sin estar asociado a él

Pr3) Las secuencias entre patterns (dentro de cada PP) deben tener alta similitud entre sí

Pr4) La secuencia de un dado PP no debe ocurrir nunca en el genoma sin contener a todos los patterns que lo componen

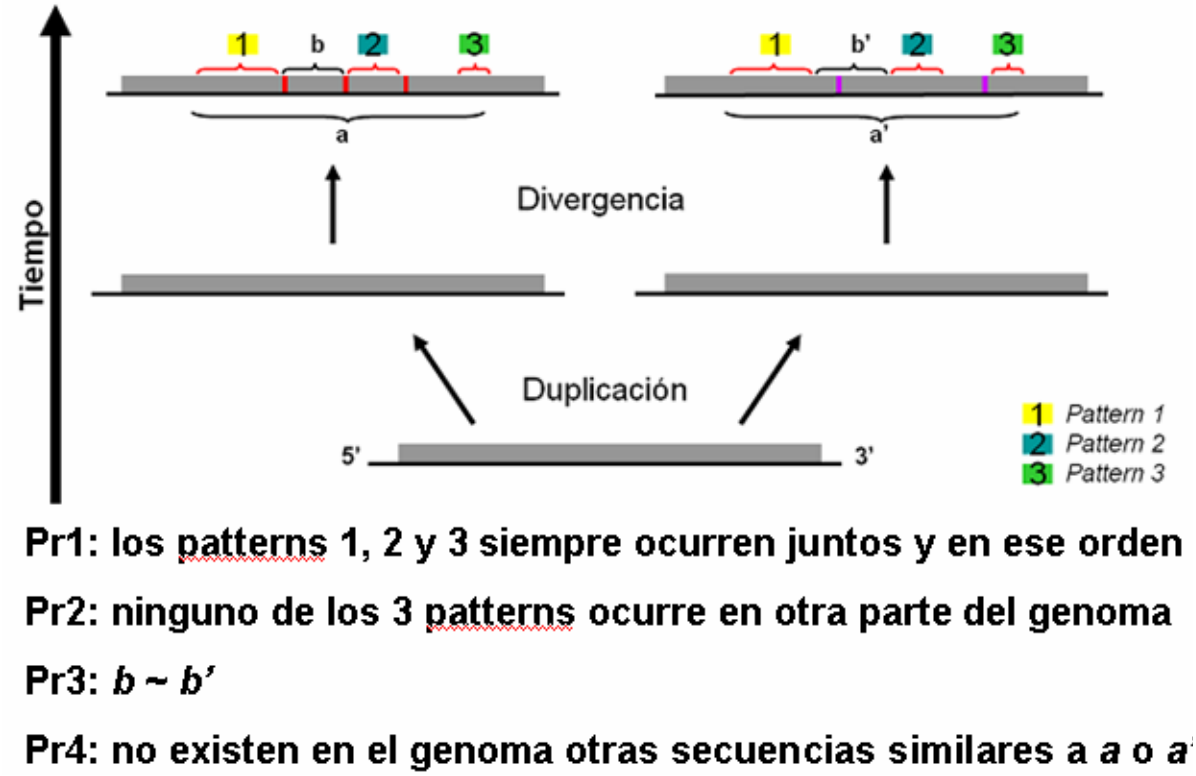

Figura 2. Ejemplificación de las 4 predicciones que se desprenden de la hipótesis. El diagrama esquematiza un evento de duplicación-divergencia de un bloque genómico (rectángulo gris). En un momento dado se generan dos copias de este bloque, inicialmente idénticas. Con el correr del tiempo cada una de ellas comienza a adquirir mutaciones propias (líneas rojas y violeta, respectivamente). Como resultado de la divergencia de estos bloques duplicados se originan los patterns 1, 2 y 3.

2.2- Selección del conjunto de patterns a estudiar

El criterio para la selección de los patterns a analizar ha sido arbitrario. Responde a nuestro sesgo personal por intentar explicar los patterns que nos resultaron más interesantes: los más largos de entre los que no pertenecieran a ningún elemento ya reportado. Este criterio permite testear la hipótesis.

De la base de datos de patterns generada por el grupo Kapow se tomaron aquellos que cumplieran con los siguientes requisitos: i) ser novel patterns (*i.e.* que sus posiciones en el genoma no se solapen con la de ningún elemento reportado previamente), ii) ocurrir al menos una vez en el cromosoma 1 y iii) tener una longitud de al menos 700 nucleótidos. Se obtuvieron así 24 patterns, que fueron identificados con números escogidos arbitrariamente (Tabla 1).

Tabla 1. Set de patterns escogidos para la puesta a prueba de la hipótesis. A cada uno se le asignó un número que lo identifica. La cantidad de posiciones iniciales informadas responde a la cantidad de veces que el pattern ocurre en el cromosoma 1.

Número de pattern	Longitud (posiciones)	Posiciones iniciales en el cromosoma 1
1	1576	144324447
		144504699
2	1313	148617657
		148779685
3	1143	577406
4	1056	146127300
		147669132
5	1048	340440
6	1027	419732
7	936	144339153
		147930351
8	928	120807460
		206106976
9	916	144327499
		144507750
10	903	24933
11	886	629062
12	825	144048774
		206511480
13	881	143737002
		147924524
(Continúa al lado)		

Número de pattern	Longitud (posiciones)	Posiciones iniciales en el cromosoma 1
(Continuación)		
14	771	121977
15	780	144323666
		144503918
16	768	648110
17	762	223745280
18	760	144333739
		144513988
19	717	121097675
		206547876
20	712	120788828
		206088299
21	705	104205829
		104299967
22	705	148620075
		148782104
23	703	340778
24	701	228751657
		228767297
		228780706

2.3- Materiales y Métodos

2.3.1- Abordaje metodológico general

Conciliación entre los patterns como definición matemática y entidad biológica

En nuestro abordaje nos propusimos considerar a cada pattern como un elemento genómico que (funcional o no) aparece repetido de manera exacta más de una vez por alguna causa. Sin embargo, hay que tener en cuenta que la aplicación de la definición matemática de pattern no tiene por qué coincidir exactamente con el resultado de los procesos biológicos que moldean los genomas. Piénsese por ejemplo en un caso como el que se esquematiza en la Figura 3, en el que un elemento sufre dos eventos de duplicación. Así, el genoma cuenta con tres versiones de este elemento original, una de las cuales no es susceptible de ser reconocida por un algoritmo de búsqueda de patterns por haber divergido de las otras dos.

Para poder trabajar en una interpretación basada en mecanismos biológicos fue necesario entonces asegurarse de que el conjunto de datos incluyera no sólo a los patterns sino también a todos los elementos biológicamente relevantes asociados a ellos. Con este fin se buscaron en el genoma todos los parálogos de cada uno de los 24 patterns.

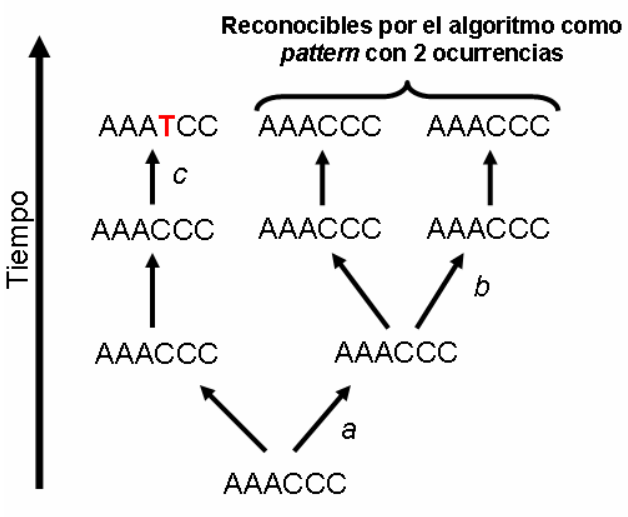

Figura 3. Esquema de la evolución de un elemento genómico. La secuencia AAACCC inicial sufre un primer evento de duplicación (*a*) generándose 2 copias. Posteriormente, una de esas copias se duplica (*b*) dando lugar a un total de tres copias, una de las cuales sufre luego un cambio en una posición (*c*). El resultado de este proceso es que el genoma contiene tres elementos con un origen evolutivo común, uno de los cuales no sería reconocible como tal por un algoritmo de búsqueda de patterns.

A partir de los resultados de estas búsquedas se definieron *familias*, cada una de ellas consistente en todas las ocurrencias de un pattern y todos los parálogos hallados a partir de él:

fi-j-k(R) es el miembro número *k* de la familia número *i*, y se ubica en el cromosoma *j*. R indica que el miembro está ubicado en sentido inverso respecto a fi-1-1. (*i.e.* es reverso complementario a él)

El número de familia corresponde al del pattern utilizado para la búsqueda de parálogos (Tabla 1). La numeración de los miembros de cada familia respeta el orden físico de los mismos a lo largo del cromosoma, según orden creciente de posición.

Con los resultados de la búsqueda de parálogos quedaron definidas 24 familias (Figura 4a).

Puesta a prueba del modelo

Para la contrastación de la hipótesis a partir de las 4 predicciones se tomaron todos los miembros de las 24 familias, se los rotuló con el número que identificara la familia y se los ordenó por posición (Figura 4b). Con todas las ocurrencias de los patterns y sus parálogos así ordenados se analizó si las 24 familias se comportaban de manera independiente o si, por el contrario, ocurrían respetando un orden de algún tipo. Para ello se investigaron los patrones de ocurrencia de los miembros de las distintas familias. Si las familias fueran independientes, sus órdenes de ocurrencia (localización en el genoma y ordenamiento relativo) serían aleatorios. En cambio si no fueran independientes, verificaríamos la existencia de regiones genómicas distintas conteniendo la misma secuencia de miembros de dichas familias. En adelante diremos que cada conjunto de familias dependientes entre sí es un *patrón de patrones* (PP), y nos referiremos a las distintas regiones genómicas donde aparece un PP como *ocurrencias* del mismo. Existirán PP distintos (que ocurrirán una determinada cantidad de veces en el genoma) según qué familias estén formando parte de ellos:

Pi-j-k(R) es la ocurrencia número *k* del PP número *i*. Esta ocurrencia está en el cromosoma *j*.
R indica que la ocurrencia está ubicada en sentido inverso respecto a Pi-1-1.

La disposición de los miembros de las 24 familias en PP se utilizó para poner a prueba el modelo a partir de las predicciones Pr1 y Pr2 (Figura 4c).

Una vez definidos los PP formados por las 24 familias se procedió a su análisis estructural. Utilizando la posición inicial, la orientación relativa y la longitud total de los elementos contenidos en un PP (*i.e.* los miembros de todas las familias involucradas en él), se calculó la distancia entre ellos para cada ocurrencia del mismo. También se calculó la longitud total de cada ocurrencia del PP, definida como la cantidad de nucleótidos entre la posición inicial del primer elemento hasta la posición final del último. Los valores de estos parámetros se compararon entre todas las ocurrencias de cada PP, para la puesta a prueba a partir de la predicción Pr1.

Para la puesta a prueba a partir de la predicción Pr3 se hizo, para cada PP, una comparación de las secuencias de todas sus ocurrencias entre sí mediante alineamientos globales y locales. La bondad de los alineamientos obtenidos se evaluó utilizando como criterios el porcentaje de similitud y su longitud (Figura 4c).

Para la puesta a prueba a partir de la predicción Pr4 se realizó un análisis exhaustivo de posibles parálogos de la secuencia de cada PP, mediante búsquedas contra todo el genoma. Una vez determinados todos los parálogos, se analizó si se había hallado alguno distinto a los definidos a partir de los miembros de las 24 familias (*i.e.* si existe alguna ocurrencia que no contiene a los patterns o a sus parálogos) (Figura 4c).

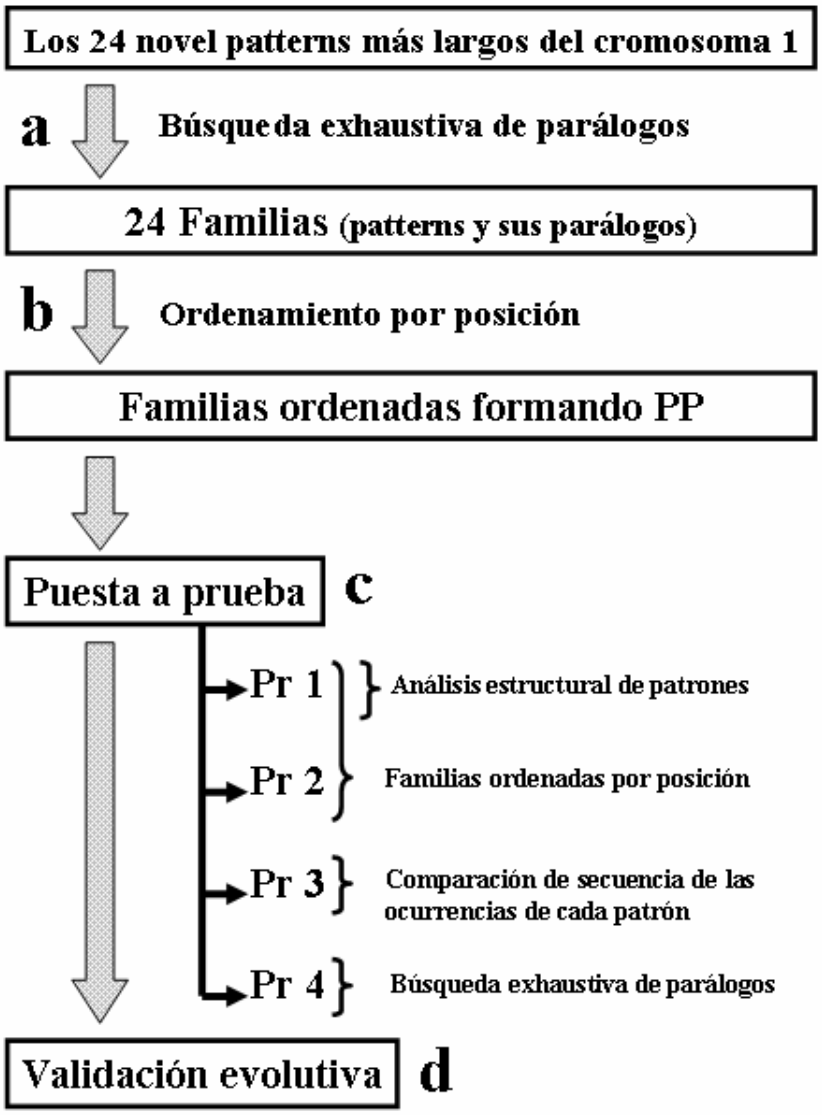

Figura 4. Esquema de trabajo para la puesta a prueba del modelo a partir de los 24 novel patterns más largos del cromosoma 1.

Validación evolutiva

Una consecuencia de la puesta a prueba de la hipótesis a partir de las 24 familias es la conclusión de que existen ciertos patterns asociados a (*i.e.* contenidos en) regiones genómicas presuntamente duplicadas (DS). Si es cierto que estas regiones deben su origen a la duplicación de elementos genómicos en algún momento de la historia, entonces dichos eventos de duplicación deberían poder ser estudiados (Figra 4d).

A pesar de que la estructura lógica de este razonamiento es similar a la del método hipotético deductivo, no hablamos de "contrastación evolutiva" sino de "validación evolutiva". Esto se debe a que la predicción en cuestión ("dichos eventos de duplicación deberían poder ser estudiados") no se desprende de la hipótesis, e incluso es independiente de la existencia de patterns en el genoma. Al estudiar los eventos de duplicación mencionados arriba lo que se está validando son las DS en cuestión como entidades biológicas.

La validación evolutiva se llevó a cabo solo para P1, que fue tomado como caso de estudio. Para inferir su historia en el linaje humano se determinó la estructura de las DS que contienen a cada una de sus ocurrencias en el genoma humano, y se compararon entre sí (a nivel de secuencia y también en cuanto a presencia de rearreglos genómicos). Para la inferencia de la historia evolutiva de estas DS a lo largo de la filogenia se identificaron las secuencias ortólogas a cada ocurrencia de P1 en todos los genomas incluidos en el análisis. Se reconstruyó la historia de las DS a partir del análisis de presencia de ortólogos en las distintas especies y mediante análisis filogenéticos a partir de las secuencias nucleotídicas de los mismos.

Comparación entre las DS halladas y bases de datos de DS

Se analizó si las DS halladas a partir de las 24 familias estaban incluidas dentro de las bases de datos de DS de acceso público o si, por el contrario, no habían sido reportadas como

tales. Se utilizaron las posiciones iniciales y finales de las DS halladas para determinar la existencia de solapamiento con la ubicación de las DS reportadas previamente.

2.3.2- Materiales y métodos detallados

Especies incluidas en el análisis

La elección de las especies a incluir en el análisis tuvo que ver con la disponibilidad de genomas publicados. Se decidió trabajar con todos los genomas disponibles de primates e incluir los genomas de perro (*Canis familiaris*), ratón (*Mus musculus*) y pollo *(Gallus gallus*).

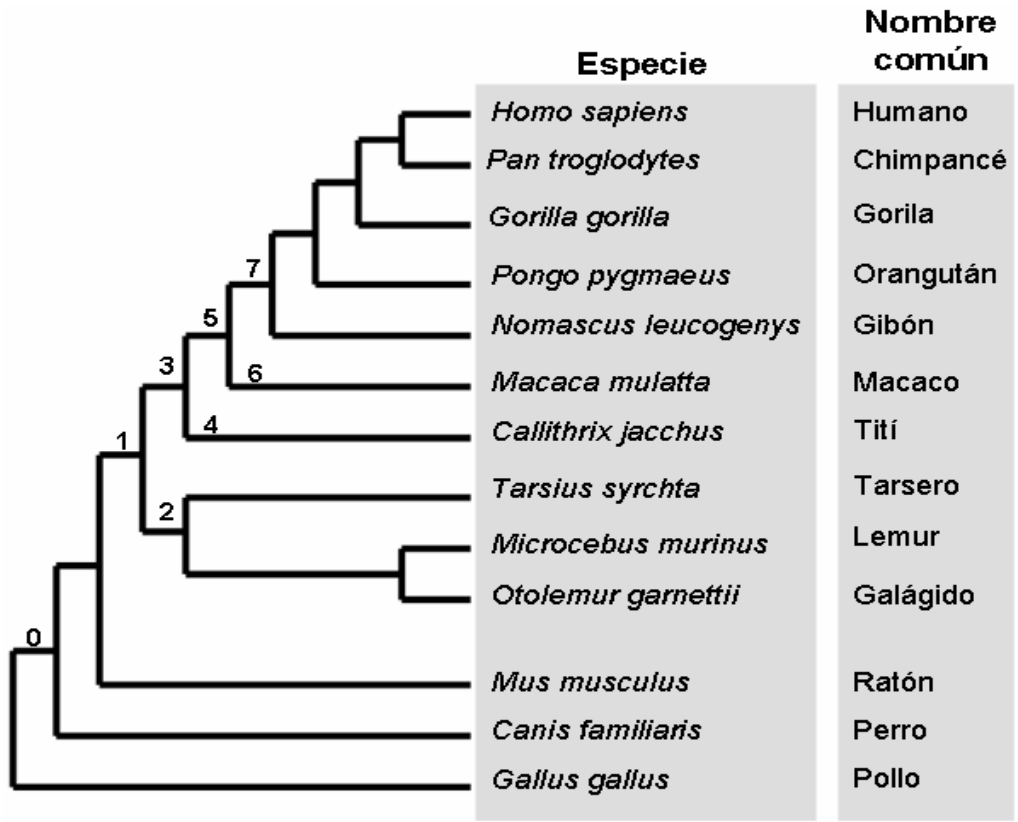

Figura 5. Relaciones filogenéticas entre las especies incluidas en el análisis. Los números denotan el grupo definido por cada nodo. *0*: Boroeutheria. *1*: Orden Primates. *2*: Suborden Prosimii. *3*: Suborden Anthropoidea. *4*: Infraorden Platyrrhini. *5*: Infraorden Catarrhini. *6*: Superfamilia Cercopithecoidea. *7*: Superfamilia Hominoidea. Modificado de Fleagle J. G. 1998. Primate adaptation and evolution, second edition. Academic Press, Nueva York.

Las relaciones filogenéticas entre todas las especies con las que se trabajó se muestran en la Figura 5.

Bases de datos genómicas

La versión del genoma humano con la que se trabajó es la más reciente: GRCh37/hg19, publicada en febrero de 2009 [15]. Para el resto de las especies también se utilizaron las últimas versiones de los correspondientes proyectos genoma:

- Chimpancé (*Pan troglodytes*): Pan_troglodytes-2.1, marzo de 2006 [34]
- Gorila (*Gorilla gorilla gorilla*): gorGor3, diciembre 2009
- Orangután (*Pongo pygmaeus*): PPYG2, septiembre de 2007 [20]
- Macaco (*Macaca mulatta*): MMUL 1.0, febrero de 2006 [28]
- Gibón (*Nomascus leucogenys*): Nleu 1.0, febrero 2007 [29]
- Tití (*Callithrix jacchus*): C_jacchus3.2.1, enero de 2010
- Tarsero (*Tarsius syrichta*): tarSyr1, julio de 2008
- Lemur (*Microcebus murinus*): micMur1, junio de 2007
- Galágido (*Otolemur garnettii*): otoGar1, mayo de 2006
- Ratón (*Mus musculus*): NCBIM37, abril de 2007 [23]
- Perro (*Canis familiaris*): CanFam 2.0, mayo de 2006 [19]
- Pollo (*Gallus gallus*): WASHUC2, mayo de 2006 [14]

Base de datos de DS

La comparación entre las posiciones de las DS halladas y la ubicación de las DS previamente reportadas se llevó a cabo a través del navegador de genomas de la Universidad de California, Santa Cruz (UCSC genome browser) [17]. La base de datos de DS con la que cuenta dicho navegador fue generada por el grupo de investigación del Dr. Evan Eichler (Eichler Lab, Departamento de Ciencias Genómicas de la Escuela de Medicina, Universidad de Washington).

Extracción de secuencias genómicas

Todas las secuencias genómicas con que se trabajó fueron extraídas en formato FASTA del navegador de genomas de Ensembl (www.ensembl.org).

Búsquedas en genomas

Todas las búsquedas en genomas a partir de una secuencia problema (en adelante secuencia *query*) se llevaron a cabo utilizando el algoritmo BLASTN a través de la interfase de Ensembl (www.ensembl.org/Multi/blastview). En todos los casos se aplicaron los filtros *dust* y *RepeatMasker* sobre las secuencias query. Se utilizaron los parámetros que se muestran en la Tabla 2.

Tabla 2. Valores de los parámetros utilizados para las búsquedas en genomas mediante BLASTN. *E*: máximo valor E reportado (*maximum E-value for reported alignments*). *W*: tamaño de palabra para ensemillado (*word size for seeding alignments*). *M*: puntuación por identidad (*match score*). *N*: puntuación por no identidad (*mismatch score*). *Q*: costo del primer "carácter hueco" (en adelante *gap*) (*cost of first gap character*). *R*: costo de los gaps restantes (*cost of second and remaining gap characters*).

Búsqueda contra genoma de	**E**	**W**	**M**	**N**	**Q**	**R**
Humano, Chimpancé, Gorila, Orangután, Macaco, Titi, Tarsero, Lemur, Galágido, Perro	10	15	1	-3	3	3
Gibón	10	9	1	-1	2	1
Ratón	10	8	1	-1	2	1
Pollo	10	4	1	-1	2	1

Criterios de homología

Para determinar si el resultado de una búsqueda correspondía a una secuencia homóloga de la secuencia query se evaluaron dos parámetros: i) porcentaje de similitud del alineamiento y ii) diferencia de longitud entre la secuencia hallada y la secuencia query, relativa a la longitud de esta última. Para que una secuencia fuera considerada homóloga a la

secuencia query estos dos parámetros debían superar un cierto valor límite. Los valores límite variaron para cada tipo de búsqueda.

Cuando se trabajó con secuencias query largas (hasta 16 kb para el caso de los PP), las búsquedas dieron como resultado una serie de alineamientos contiguos cubriendo la totalidad de cada uno de los homólogos hallados (en lugar de un único alineamiento por cada homólogo).

Búsqueda de parálogos de los patterns

La secuencia de cada uno de los patterns fue utilizada como secuencia query en una búsqueda contra el genoma humano utilizando BLASTN. Para considerar a un resultado de la búsqueda como parálogo del pattern en cuestión se adoptó como criterio que su similitud fuera mayor al 96 % y que la diferencia de longitud (entre el resultado y la secuencia query) fuera menor al 10 % de la longitud total del pattern.

Búsqueda de parálogos de los PP

Se tomó la secuencia completa de una de las ocurrencias de cada PP estudiado y se la utilizó como secuencia query en una búsqueda mediante BLASTN contra el genoma humano.

Búsqueda de ortólogos de los PP

Para buscar los ortólogos de las DS que contienen a las ocurrencias de un PP se tomó la secuencia de una de ellas y se la utilizó como secuencia query en una búsqueda mediante BLASTN contra el genoma completo de la especie correspondiente. Como la elección de una única ocurrencia a utilizar como secuencia query habría sido arbitraria, se realizó una búsqueda de ortólogos a partir de cada una de ellas.

Para cada uno de los homólogos hallados hubo que decidir de cuál de las ocurrencias del PP era ortólogo. Para ello se adoptaron los siguientes criterios: cromosoma en que los presuntos ortólogos están ubicados en cada especie, orientación relativa y ubicación relativa a

marcadores homólogos entre las especies. La información sobre la sintenia de las regiones genómicas involucradas se obtuvo del navegador de Ensembl.

Para la búsqueda de ortólogos podría haberse optado por utilizar la secuencia completa de las DS como secuencia query, en lugar de las ocurrencias de los PP. La decisión de no trabajar con la secuencia completa de las DS responde a la existencia de variación estructural entre las ocurrencias (*i. e.* inversiones, inserciones y deleciones de longitud considerable entre las distintas ocurrencias de una DS). Las ocurrencias de los PP no presentan entre ellas variaciones de este tipo, con lo cual resultaron más apropiadas para ser utilizadas como secuencia query en la búsqueda.

Comparación de pares de secuencias

Para la comparación de pares de secuencias nucleotídicas se utilizaron dos herramientas:

Alineamientos locales: se utilizó el algoritmo BLASTN a través de la interfase de NCBI (blast.ncbi.nlm.nih.gov/Blast.cgi, herramienta *bl2seq* para alinear dos o más secuencias). Se filtraron regiones de baja complejidad de las secuencias y los parámetros utilizados fueron: *E*=10; *W*=11; *M*=2; *N*=-3; *Q*=5; *R*=2.

Alineamientos globales: se utilizó el algoritmo ClustalW implementado mediante el programa BioEdit [13].

Para la comparación de pares de secuencias aminoacídicas se utilizó el algoritmo BLASTP a través de la interfase de NCBI (blast.ncbi.nlm.nih.gov/Blast.cgi, herramienta *bl2seq* para alinear dos o más secuencias). Se utilizó la matriz BLOSUM62 y los parámetros utilizados fueron: *E*=10; *W*=3; *Q*=11; *R*=1.

Determinación de la estructura de las DS que contienen a las ocurrencias de P1

Para definir los límites de las DS que contienen a P1 se trabajó con sus secuencias flanqueantes. Se extrajeron las secuencias 5' y 3' flanqueantes a cada una de sus ocurrencias. Se decidió extraer un máximo de 200 kb flanqueantes, tanto hacia 5' como hacia 3'. Sin embargo, la cantidad de posiciones extraídas dependió de la disponibilidad de datos, ya que varias de las ocurrencias se ubican en zonas del genoma pobladas de regiones de secuencia desconocida. En estos casos se extrajeron las secuencias flanqueantes hasta la aparición de la primera tanda de posiciones indefinidas. La secuencia más corta obtenida fue la 3' flanqueante a P1-1-1, de 27238 nucleótidos de longitud. Se compararon todas las secuencias flanqueantes entre sí (5' y 3' respectivamente) mediante alineamientos locales. Se decidió no hacer alineamientos globales por dos motivos: i) el costo computacional asociado a la gran longitud de las secuencias y ii) la posible existencia de grandes inversiones, inserciones y deleciones. Para todas las comparaciones se analizó cada uno de los alineamientos obtenidos, comenzando por los de mayor valor E. Utilizando las posiciones de los alineamientos y su porcentaje de similitud se determinaron los límites a partir de los cuales deja de haber similitud entre las ocurrencias del PP (*i.e.* los límites de la DS) y se detectaron posibles inserciones o deleciones. La orientación relativa de los alineamientos permitió determinar la existencia de inversiones en algunas ocurrencias respecto a las demás.

Análisis filogenéticos

Para la construcción de las matrices básicas de datos se partió de alineamientos globales múltiples. Como las DS estudiadas presentaron una cantidad considerable de inserciones y deleciones especie-específicas, no se utilizó la totalidad de las mismas para estos alineamientos. Antes del alineamiento global se realizaron alineamientos locales de cada uno de los homólogos contra la secuencia de P1-1-1. A partir de ellos se escogió la región de la secuencia para la que estuvieran representadas la mayor parte de las especies. La matriz básica de datos más inclusiva que se consiguió contempla a todas las especies del Suborden

Anthropoidea (Figura 5). Las tres secuencias de Gorila del cromosoma 1 quedaron excluidas de esta matriz de datos porque eran considerablemente más cortas al resto y no solapantes entre sí (*i.e.* todas presentan alta similitud con P1-1-1, pero en zonas distintas).

Se trabajó con 2 matrices básicas de datos:

i) La más inclusiva posible: 16 terminales, 11905 posiciones

ii) Las ocurrencias de P1 del linaje humano: 7 terminales, 17632 posiciones

Para las reconstrucciones filogenéticas se utilizó el programa PAUP* versión 4.0b10 [33]. Se emplearon 2 metodologías de reconstrucción:

Máxima parsimonia

Se asignó el mismo peso a todos los caracteres utilizados. Para las búsquedas se implementó el algoritmo Branch-and-Bound. El soporte de los nodos se estudió mediante bootstrap de 10000 réplicas.

Máxima verosimilitud

El modelo de evolución molecular se escogió de entre 56 posibles mediante un análisis jerárquico de relación de verosimilitud (hLRT) utilizando el programa Modeltest 3.7 [27]. Se escogió un modelo de evolución molecular para cada una de las matrices básicas de datos con que se trabajó. Se realizaron búsquedas heurísticas de 10 réplicas, con secuencia aleatoria de adición de taxones. El soporte de los nodos se estudió mediante bootstrap de 1000 réplicas.

2.4- Resultados

2.4.1- Conciliación entre los patterns como definición matemática y como entidad biológica

Búsqueda de parálogos y definición de las familias

La búsqueda exhaustiva de parálogos de los 24 patterns mostró que salvo dos de ellos (12 y 17) todos presentan al menos uno que no había sido detectado por el algoritmo de Kapow. Muchos de los parálogos hallados son incluso idénticos en secuencia al pattern al que corresponden. En principio esto puede resultar contradictorio, ya que el algoritmo de Kapow encuentra patterns (que por definición son secuencias idénticas) de manera exhaustiva. Lo que ocurre es que todos los parálogos idénticos hallados se encuentran sobre la hebra complementaria a la de su pattern. Como el algoritmo de Kapow se corrió originalmente sobre una de las dos hebras, todas las repeticiones exactas ubicadas en la otra hebra pasaron desapercibidas hasta que se realizó la búsqueda mediante BLAST a partir de la secuencia de los patterns.

Se hallaron 230 parálogos en total: 119 en el cromosoma 1 y el resto en los cromosomas 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 15, 16, 19, 20 e Y. Las familias más numerosas son f5, f14, f16 y f24, con 26 miembros cada una. f12 y f17 son las menos numerosas, con solo dos miembros.

2.4.2- Puesta a prueba del modelo

Análisis de dependencia de las familias entre sí

Al ordenar los miembros de las 24 familias por posición en el cromosoma 1 se puso en evidencia la dependencia de algunas de las familias entre sí (Tabla 3). Inspeccionando las filas de esta tabla puede apreciarse a simple vista que existen patrones de aparición para los miembros de determinadas familias.

El caso más claro es el de los miembros de las familias f15, f1, f9, f13, f18 y f7, que aparecen siempre asociados en ese orden formando el PP P1 (amarillo). P1 tiene 5 ocurrencias en el cromosoma 1, denominadas P1-1-1, P1-1-2, P1-1-3, P1-1-4 y P1-1-5-R. P1-1-5-R está en sentido reverso respecto a las otras cuatro ocurrencias. Como consecuencia de ello, el orden de las familias está invertido (*i.e.* f7, f18, f13, f9, f1, f15) y todos los elementos que lo componen son también reversos respecto a sus respectivos parálogos en las otras ocurrencias.

Lo mismo ocurre con las familias f16, f5, f23 y f14, que aparecen siempre asociadas entre sí formando P2 (magenta). P2-1-2-R y P2-1-5-R están en sentido reverso respecto a las otras 4 ocurrencias, y como consecuencia de ello el orden de las familias está invertido (*i.e.* f14, f23, f5, f16). Estas dos ocurrencias presentan la particularidad de ser las únicas (en el cromosoma 1) que solo son reconocidas como tales cuando se las ordena por posición final (véase *Análisis de P2*, más adelante).

Las familias f2 y f22 aparecen siempre asociadas formando P3 (verde), que tiene 6 ocurrencias en el cromosoma 1, cuatro de ellas en sentido inverso a las otras dos.

Los miembros de f24 aparecen repetidos en tándem formando P4 (anaranjado), que está formado por una única familia y consta de una única ocurrencia.

Lo mismo ocurre con f21, cuyos miembros se repiten en tándem formando P5 (celeste).

Existen otros conjuntos de familias que parecen presentar una cierta dependencia entre sí, pero que no cumplen con los requisitos para ser considerados PP:

- f6: sus miembros ocurren siempre asociados a algún miembro de f3
- f8: sus miembros ocurren siempre asociados a algún miembro de f20
- f11: sus miembros ocurren siempre entre un miembro de f3 y uno de f16
- f12: sus miembros ocurren siempre asociados a algún miembro de f19

Tabla 3. Todos los miembros de las 24 familias definidas para el cromosoma 1, ordenados por posición. Cada color denota uno de los 5 PP formados.

n° de familia	Miembro de la familia	Posición de inicio en el cromosoma	Posición final en el cromosoma	Patrón formado
10	f10-1-1	24933	25835	
3	f3-1-1	95835	96979	
16	f16-1-1	121771	122538	P2-1-1
5	f5-1-1-R	121837	122884	
23	f23-1-1-R	121844	122546	
14	f14-1-1	121977	122747	
20	f20-1-1	163064	163759	
3	f3-1-2	242170	243314	
14	f14-1-2-R	340577	341347	P2-1-2-R
23	f23-1-2	340778	341480	
5	f5-1-2	340440	341487	
16	f16-1-2-R	340786	341553	
11	f11-1-1-R	359741	360626	
3	f3-1-3-R	411919	413061	
6	f6-1-1	419732	414088	
3	f3-1-4	577406	578548	
11	f11-1-2	629062	629947	
16	f16-1-3	648110	648877	P2-1-3
5	f5-1-3-R	648176	649221	
23	f23-1-3-R	648183	648883	
14	f14-1-3	648316	649084	
20	f20-1-2	698904	699599	
20	f20-1-9-R	38706629	38707308	
4	f4-1-1-R	87914814	87915881	
21	f21-1-1	104120661	104121368	P5-1-1
21	f21-1-2	104167039	104167754	
21	f21-1-3	104205829	104206533	
21	f21-1-4-R	104230680	104231384	
21	f21-1-5	104261164	104261879	
21	f21-1-6	104299967	104300671	
20	f20-1-3	120788828	120789539	
8	f8-1-1	120807460	120808387	
19	f19-1-1	121097675	121098391	
15	f15-1-1	143727341	143728122	P1-1-1
1	f1-1-1	143728124	143729699	
9	f9-1-1	143731171	143732087	
13	f13-1-1	143737002	143737882	
18	f18-1-1	143737416	143738175	
7	f7-1-1	143742829	143743763	
8	f8-1-2-R	143943721	143944648	
20	f20-1-4-R	143962568	143963279	
12	f12-1-1	144048774	144049598	
19	f19-1-2	144084865	144085581	
15	f15-1-2	144323666	144324445	P1-1-2
1	f1-1-2	144324447	144326022	
9	f9-1-2	144327499	144328415	
13	f13-1-2	144333328	144334205	
18	f18-1-2	144333739	144334498	
7	f7-1-2	144339153	144340088	
22	f22-1-1-R	144387954	144388656	P3-1-1
2	f2-1-1-R	144389754	144391065	
15	f15-1-3	144503918	144504697	P1-1-3
1	f1-1-3	144504699	144506274	
9	f9-1-3	144507750	144508666	
13	f13-1-3	144513577	144514454	
18	f18-1-3	144513988	144514747	
7	f7-1-3	144519402	144520336	
4	f4-1-2	146127300	146128355	
4	f4-1-3	147669132	147670187	
15	f15-1-4	147914849	147915628	P1-1-4
1	f1-1-4	147915630	147917200	
9	f9-1-4	147918675	147919591	
13	f13-1-4	147924524	147925404	
18	f18-1-4	147924938	147925697	
7	f7-1-4	147930351	147931286	

(Continúa al lado)

n° de familia	Miembro de la familia	Posición de inicio en el cromosoma	Posición final en el cromosoma	Patrón formado
(Continuación)				
22	f22-1-2-R	147979117	147979819	P3-1-2
2	f2-1-2-R	147980917	147982228	
22	f22-1-3-R	148225757	148226459	P3-1-3
2	f2-1-3-R	148227561	148228873	
2	f2-1-4	148617657	148618969	P3-1-4-R
22	f22-1-4	148620075	148620780	
2	f2-1-5	148779685	148780997	P3-1-5-R
22	f22-1-5	148782104	148782809	
2	f2-1-6	149526152	149527463	P3-1-6-R
22	f22-1-6	149528562	149529264	
7	f7-1-5-R	149577138	149578073	P1-1-5-R
18	f18-1-5-R	149582730	149583484	
13	f13-1-5-R	149583018	149583895	
9	f9-1-5-R	149588853	149589769	
1	f1-1-5-R	149591240	149592815	
15	f15-1-5-R	149592817	149593596	
20	f20-1-5	206088299	206089010	
8	f8-1-3	206106976	206107903	
12	f12-1-2	206511480	206512304	
19	f19-1-3	206547876	206548592	
20	f20-1-6	222682323	222683017	
17	f17-1-1	223745280	223746041	
6	f6-1-2-R	224091839	224092870	
3	f3-1-5	224099544	224100688	
16	f16-1-4	224125955	224126722	P2-1-4
5	f5-1-4-R	224126021	224127064	
23	f23-1-4-R	224126028	224126730	
14	f14-1-4	224126161	224126927	
20	f20-1-7	224169262	224169956	
14	f14-1-5-R	228151429	228152195	P2-1-5-R
23	f23-1-5	228151630	228152328	
5	f5-1-5	228151292	228152335	
16	f16-1-5-R	228151638	228152401	
24	f24-1-1	228744934	228746635	P4-1-1
24	f24-1-2	228747175	228747876	
24	f24-1-3	228749416	228750117	
24	f24-1-4	228751657	228752358	
24	f24-1-5	228753898	228754600	
24	f24-1-6	228756129	228756813	
24	f24-1-7	228758332	228759034	
24	f24-1-8	228760574	228761275	
24	f24-1-9	228762814	228763516	
24	f24-1-10	228765055	228765756	
24	f24-1-11	228767297	228767998	
24	f24-1-12	228769537	228770238	
24	f24-1-13	228771762	228772463	
24	f24-1-14	228774003	228774704	
24	f24-1-15	228776234	228776935	
24	f24-1-16	228778475	228779176	
24	f24-1-17	228780706	228781407	
24	f24-1-18	228782911	228783415	
6	f6-1-3-R	243170214	243171245	
3	f3-1-6	243177908	243179051	
16	f16-1-6	243203745	243204512	P2-1-6
5	f5-1-6-R	243203811	243204859	
23	f23-1-6-R	243203818	243204518	
14	f14-1-6	243203951	243204722	
20	f20-1-8	243249814	243250508	

P1
P2
P3
P4
P5

Es importante destacar que al extender el análisis de dependencia de familias al resto del genoma se observó que la dependencia entre familias se mantiene en todos los cromosomas. Al ordenar todos los miembros por posición (en cada cromosoma) se

recuperaron los mismos PP que habían sido definidos para el cromosoma 1 (*i.e.* las mismas familias y en el mismo orden). Más aún, no existe ningún miembro de las familias involucradas en PP que ocurra alguna vez en el genoma sin estar formando parte del PP correspondiente. En la Tabla 4 se muestra el resultado para el cromosoma 5, a modo de ejemplo. Los resultados correspondientes al resto del genoma se muestran en el Material Suplementario 1, Tabla S1.

Tabla 4. Todos los miembros de las 24 familias del cromosoma 5, ordenados por posición.

n° de familia	Miembro de la familia	Posición de inicio en el cromosoma	Posición final en el cromosoma	Patrón formado
2	f2-5-1	49900427	49901718	P3-5-1-R
22	f22-5-1	49902824	49905236	
7	f7-1-5-R	149577138	149578073	P1-5-1-R
18	f18-1-5-R	149582730	149583484	
13	f13-1-5-R	149583018	149583895	
9	f9-1-5-R	149588853	149589769	
1	f1-1-5-R	149591240	149592815	
15	f15-1-5-R	149592817	149593596	
20	f20-5-1	180723078	180723773	
14	f14-5-1	180767192	180767962	P2-5-1-R
23	f23-5-1	180767393	180768095	
5	f5-5-1	180767055	180768102	
16	f16-5-1	180767401	180768168	
11	f11-5-1	180786611	180786599	
3	f3-5-1	180838469	180839611	
6	f6-5-1	180846289	180847315	

P1 y P3 presentaron una única ocurrencia fuera del cromosoma 1, en el cromosoma 5. P2 presentó 15 ocurrencias fuera del cromosoma 1, en los cromosomas 2, 3, 4, 5, 6, 8, 10, 11, 16, 19, 20 e Y (Material Suplementario 1, Tabla S1). Las familias que componen P4 y P5 no tienen miembros fuera del cromosoma 1. Las familias que mostraban dependencia con otras sin formar PP también conservaron en alguna medida la asociación mencionada arriba: f6 siempre ocurre asociado a f3; f11 ocurre muchas veces asociado a f3 y f16 (aunque también ocurre solo en algunos cromosomas); f8 y f12 no ocurren fuera del cromosoma 1. También se observó que f20 suele aparecer asociado a ocurrencias de P2.

Análisis estructural de P1

La longitud promedio de las 6 ocurrencias de P1 es de 16463 nucleótidos. La diferencia de longitud entre ellas es pequeña: P1-1-5-R, la más larga, supera a la más corta (P1-1-3) por 195 nucleótidos, que equivalen al 1,18 % de la longitud promedio (Tabla 5). Las otras cinco ocurrencias difieren en su longitud en entre 0,018 % y 0,2 % de la longitud promedio.

La distancia entre los miembros contiguos de las 6 familias (calculada por diferencia entre la posición inicial de uno y la final del anterior) también se mantiene entre las ocurrencias de P1. Todos los miembros de f1 ocurren 2 nucleótidos río abajo de la última posición de un miembro de f15. Lo propio ocurre con las familias restantes, que se suceden a distancias que no varían más del 2,1 % de la distancia promedio (Tabla 5).

La poca variabilidad en todos estos parámetros puede apreciarse también mediante los valores bajos de CV que presenta cada uno.

Tabla 5. Análisis estructural de las ocurrencias de P1. Para cada ocurrencia se informa su longitud total, calculada por diferencia entre la posición final del último miembro y la inicial del primero, y la distancia entre los miembros consecutivos de las familias que lo componen, calculada por diferencia entre la posición inicial de un miembro y la final del miembro inmediatamente anterior. Para todas estas variables se muestran los parámetros estadísticos que se consideraron más relevantes: promedio, desvío estándar (Desv. std.), Coeficiente de variación (CV) y diferencia máxima resultante de hacer todas las comparaciones posibles de a pares, tanto absoluta (Máx. dif.) como porcentual relativa al promedio (Máx. dif. %).

		Distancia entre miembros de las familias				
Ocurrencia	**Longitud**	**15-1**	**1-9**	**9-13**	**13-18**	**18-7**
P1-1-1	16424	2	1472	4915	-466	4654
P1-1-2	16423	2	1477	4913	-466	4655
P1-1-3	16420	2	1476	4911	-466	4655
P1-1-4	16438	2	1475	4933	-466	4654
P1-1-5-R	16615	2	1471	4958	-466	4657
P1-5-R	16458	2	1502	4949	-476	4639
Promedio	16463	2	1478,8	4929,8	-467,7	4652,3
Desv. std.	75,79	0	11,58	20,14	4,08	6,62
CV	0,005	0	0,008	0,004	0,009	0,001
Máx. dif.	195	0	31	47	10	18
Máx. dif. %	1,18	0,02	2,1	1	2,1	0,4

La distancia entre miembros de f13 y f18 tiene signo negativo. Esto significa que los patterns 13 y 18 están solapados.

La comparación de la secuencia nucleotídica de las 6 ocurrencias de P1, realizada mediante alineamientos globales, mostró que todas ellas son altamente similares (Tabla 6). Las ocurrencias del cromosoma 1 son más similares entre sí (no menos del 98,6 %) que lo que es cualquiera de ellas a la ocurrencia del cromosoma 5. Aún así, las ocurrencias más disímiles (P1-5-1-R respecto a P1-1-1) presentan una similitud del 93,8 %.

Tabla 6. Comparación a nivel de secuencia de las 6 ocurrencias de P1. Los valores que se informan son el porcentaje de similitud entre las secuencias de las ocurrencias correspondientes, calculados a partir del alineamiento global. En el cálculo se tuvo en cuenta la cantidad de gaps. Los valores máximo y mínimo se muestran resaltados.

	P1-1-1	**P1-1-2**	**P1-1-3**	**P1-1-4**	**P1-1-5-R**
P1-1-1					
P1-1-2	99				
P1-1-3	99	**99,7**			
P1-1-4	99,4	99	98,9		
P1-1-5-R	98,6	99	99	98,7	
P1-5-1-R	**93,8**	93,9	93,9	93,9	94

Las 6 ocurrencias de P1 analizadas hasta aquí fueron halladas a partir de los patterns y sus parálogos. Sin embargo, no se puede asegurar que no existan en el genoma otros parálogos de P1, no detectables mediante la metodología de búsqueda de patrones de dependencia entre familias. Para hacer un sondeo exhaustivo de ocurrencias de P1 se utilizó la secuencia de P1-1-1 como secuencia query en una búsqueda mediante BLASTN contra el genoma humano completo. Se hallaron 6 parálogos en total, que coinciden exactamente con las 6 ocurrencias de P1 definidas a partir de las 24 familias. Esta búsqueda de parálogos se repitió utilizando como secuencia query a las otras 5 ocurrencias de P1, obteniéndose el mismo resultado en todos los casos.

Análisis estructural de P2

De las 21 ocurrencias de P2 hay 10 que se dan en sentido reverso, 10 en sentido derecho y una (P2-6-2-R) que presenta problemas para su clasificación.

Las 10 ocurrencias en sentido reverso tienen la particularidad de que solo son reconocidas como tales (*i.e.* los miembros de las familias que los componen respetan el orden correspondiente a P2) cuando el ordenamiento se hace por posición final en vez de inicial (Tabla 3 y Material Suplementario 1, Tabla S1). Esto no ocurría con P1, cuyas ocurrencias en sentido reverso pueden ser detectadas independientemente de si se las ordena por posición inicial o final. La causa de esta diferencia tiene que ver con las posiciones relativas de los miembros de las familias. En el caso de P1 los miembros de las familias no se solapan entre sí: f15 está 2 nucleótidos río arriba de f1, ésta 1478 nucleótidos (en promedio) río arriba de f9, etc. La única excepción es la de f13 y f18, que tienen un solapamiento de 466 nucleótidos (Tabla 5). En el caso de P2, en cambio, los miembros de las 4 familias que lo componen están siempre solapados entre sí (Tabla 7). Una consecuencia de esto es que si se toma una ocurrencia de P2 y se la ubica en sentido reverso, el ordenamiento por posición inicial varía respecto al ordenamiento por posición final (Material Suplementario 2, Figuras S1a y S1b).

El caso de P2-6-2-R es problemático porque los miembros de las familias no respetan el orden esperado para P2 así se las ordene por posición inicial o final. Sin embargo, esta ocurrencia tiene la particularidad de que sus miembros de f5 y f14 son notablemente más cortos que los del resto (33 % y 26 % más cortos, respectivamente; véase Material Suplementario 1, Tabla S1). Teniendo esto en cuenta se puede ver que si la longitud de los miembros de f5 y f14 fuera la del resto de las ocurrencias, el orden de las familias por posición final sería el esperado (Material Suplementario 2, Figura S1c). Podemos decir entonces que se trata de una ocurrencia reversa de P2.

La longitud promedio de las 21 ocurrencias de P2 es de 959 nucleótidos. Puede verse que hay tres ocurrencias que son notablemente más cortas que las demás. Una de ellas es P2-6-2-R, cuya diferencia de longitud se atribuye a f5 y f14. Las otras dos son P2-3-1-R y P2-4-2. En estos casos, se observa que la diferencia de longitud es atribuible a sus miembros de f16, que son 63 y 61 nucleótidos más cortos que el pattern original, respectivamente (Material Suplementario 1, Tabla S1). Si se excluyen estas 3 ocurrencias del análisis la longitud promedio es de 974,9 nucleótidos, con un CV de 0,002. La diferencia entre la ocurrencia más

larga (P2-1-6) y la más corta (P2-1-4) es de 5 nucleótidos, que corresponde al 0,5 % de la longitud promedio (Tabla 7).

Tabla 7. Análisis estructural de las ocurrencias de P2. Para cada una de las 21 ocurrencias se informan los mismos parámetros que en la Tabla 5. (#): parámetro estimado excluyendo a las ocurrencias P2-3-1-R, P2-4-2 y P2-6-2-R.

	Distancia entre miembros de familias			
Ocurrencia	**Longitud**	**16-5**	**5-23**	**23-14**
P2-1-1	976	-701	-1040	-569
P2-1-2-R	976	-701	-1040	-569
P2-1-3	974	-701	-1038	-567
P2-1-4	972	-701	-1036	-569
P2-1-5-R	972	-697	-1036	-565
P2-1-6	977	-701	-1041	-567
P2-2-1-R	973	-698	-1037	-566
P2-3-1-R	910	-701	-1040	-569
P2-4-1-R	973	-698	-1032	-566
P2-4-2	907	-703	-1043	-570
P2-5-1-R	976	-701	-1040	-569
P2-6-1	976	-701	-1040	-569
P2-6-2-R	775	-694	-702	-569
P2-8-1	975	-700	-1039	-568
P2-10-1-R	976	-701	-1040	-569
P2-11-1	972	-701	-1036	-569
P2-16-1-R	976	-701	-1040	-569
P2-19-1	976	-701	-1040	-569
P2-20-1-R	976	-701	-1040	-569
P3-Y-1-R	976	-700	-1039	-568
P2-Y-2	976	-700	-1039	-568
Promedio	959	-700,1	-1022,8	-568,2
Promedio (#)	974,9	-700,3	-1038,5	-568,1
Desv. std	46,7	1,9	73,5	1,3
Desv. std (#)	1,71	1,27	2,28	1,3
CV	0,002	0,002	0,002	0,002
CV (#)	0,002	0,002	0,002	0,002
Max dif	202	9	341	5
Max dif (#)	5	4	11	4
Max dif %	21,1	1,3	33,3	0,9
Max dif % (#)	0,5	0,6	1,1	0,7

Todas las distancias entre los miembros contiguos de las familias que forman P2 son negativas. Esto implica que las secuencias de dichos miembros están solapadas en el genoma (un miembro comienza antes de que el anterior haya terminado; Tabla 7 y Material Suplementario 2, Figura S1). Estas distancias (o grado de solapamiento, en este caso) varían

muy poco entre las ocurrencias de P2, aún incluyendo en el análisis a las 3 ocurrencias más cortas. La única excepción es P2-6-2-R, que muestra respecto al resto de las ocurrencias un menor solapamiento entre f5 y f23, atribuible a la diferencia de longitud de f5 mencionada arriba.

Al someter a las 21 ocurrencias a un alineamiento global se constató la diferencia de longitud de las tres ocurrencias más cortas: P2-3-1-R y P2-4-2 presentan 65 y 68 gaps iniciales, respectivamente, que corresponderían al acortamiento de f16; P2-6-2-R presenta 200 gaps finales, que corresponderían al acortamiento de f5 (Material Suplementario 2, Figura S1).

Todas las ocurrencias son altamente similares entre sí. Las más disímiles (P2-4-2 respecto a las 2 ocurrencias del cromosoma Y) presentan una similitud del 95,3 %. Las 2 ocurrencias del cromosoma Y son idénticas entre sí. Lo mismo ocurre con P2-1-2-R, P2-5-1-R y P2-19-1 (Material Suplementario 2, Tabla S2).

El análisis exhaustivo de parálogos permitió hallar 21 en total, que coinciden exactamente con los definidos a partir del análisis de dependencia de familias. Este resultado fue independiente de la ocurrencia que se utilizara como secuencia query.

Análisis estructural de P3

De las 7 ocurrencias de P3 hay 4 que se dan en sentido reverso. La longitud promedio de las 6 ocurrencias del cromosoma 1 es de 3116,2 nucleótidos. La diferencia de longitud entre ellas es pequeña: la más larga (P3-1-5-R) supera a la más corta (P3-1-1) por 13 nucleótidos, que equivalen al 0,42 % de la longitud promedio. P3-5-1-R, la única ocurrencia del cromosoma 5, tiene una longitud notablemente mayor a la longitud promedio de las ocurrencias del cromosoma 1. Esta diferencia es atribuible a que su miembro de f22 es 1710 nucleótidos más largo que el resto de las ocurrencias (Tabla 8 y Material Suplementario 1, Tabla S1).

La distancia entre los miembros de f2 y f22 varía muy poco entre las 7 ocurrencias, tomando un valor máximo equivalente al 0,82 % de la distancia promedio (Tabla 8).

La comparación de la secuencia nucleotídica de las 6 ocurrencias del cromosoma 1 mostró que todas ellas son altamente similares. Las más disímiles entre sí (P3-1-5-R con P3-1-6-R) presentan una similitud del 98,8 % (Material Suplementario 3, Tabla S3). Dada la diferencia de tamaño entre P3-5-1-R y las ocurrencias del cromosoma 1, se decidió someter su secuencia a un alineamiento local mediante BLASTN para 2 secuencias (*E* = 10; *W* = 28; *M* = 2; *N* = -3; *Q* = 5; *R* = 2) contra la secuencia de P3-1-1. A partir de este alineamiento se detectó una inserción de 1700 nucleótidos. Utilizando la posición exacta de esta inserción se removieron las posiciones correspondientes de la secuencia de P3-5-1-R, y se la incluyó en el alineamiento global junto con las otras 6 ocurrencias. Las ocurrencias del cromosoma 1 son más similares entre sí que a P3-5-1-R. Aún así, la similitud con P3-5-1-R es siempre mayor al 94 % (Material Suplementario 3, Tabla S3).

Tabla 8. Análisis estructural de las ocurrencias de P3. Para cada una de las 7 ocurrencias se informan los mismos parámetros que en la Tabla 5. (#): parámetro estimado excluyendo a P3-5-1-R.

	Distancia entre elementos	
Ocurrencia	**Longitud**	**22-2**
P3-1-1	3111	1098
P3-1-2	3111	1098
P3-1-3	3116	1102
P3-1-4-R	3123	1106
P3-1-5-R	3124	1107
P3-1-6-R	3112	1099
P3-5-1-R	4809	1106
Promedio	3358	1102,29
Promedio (#)	3116,17	1101,67
Desvío std	639,85	4,03
Desvío std (#)	5,98	4,03
CV	0,191	0,004
CV (#)	0,002	0,004
Max dif	1698	9
Max dif (#)	13	9
Max dif %	50,57	0,82
Max dif % (#)	0,42	0,82

Análisis estructural de P4

Este PP tiene una única ocurrencia en todo el genoma, de 36473 nucleótidos de longitud. Las repeticiones en tándem de los miembros de f24 que componen esta ocurrencia están separadas por una distancia de 1535,6 nucleótidos en promedio. Esta distancia varía muy poco: los dos miembros contiguos más alejados presentan una distancia 22 nucleótidos mayor a la del par más cercano, que equivalen al 1,4 % de la distancia promedio (Tabla 9).

Todas las secuencias que separan a cada par de miembros de f24 son altamente similares entre sí. Las más disímiles son las secuencias entre 6-7 y 12-13, respectivamente, y presentan una similitud del 97,4 % (Material Suplementario 4, Tabla S4).

Tabla 9. Análisis estructural de la única ocurrencia de P4. Se muestra la distancia entre cada uno de los miembros y el inmediatamente anterior. Se informan los mismos parámetros estadísticos que en la Tabla 7.

Miembro	Longitud	Distancia al anterior
1	701	--
2	701	1540
3	701	1540
4	701	1540
5	702	1540
6	684	1529
7	702	1519
8	701	1540
9	702	1539
10	701	1539
11	701	1541
12	701	1539
13	701	1524
14	701	1540
15	701	1530
16	701	1540
17	701	1530
1-17	36473	--
Promedio	700,176	1535,625
Desvío std	4,187	6,908
CV	0,006	0,004
Máx. dif	18	22
Máx dif %	0,026	1,433

Tabla 10. Análisis estructural de la única ocurrencia de P5. Se muestra la distancia entre cada uno de los miembros y el inmediatamente anterior. Se informan los mismos parámetros estadísticos que en la Tabla 7.

Miembro	Longitud	Distancia al anterior
1	707	--
2	715	45671
3	704	38075
4	704	24147
5	715	29780
6	704	38088
Promedio	708,167	35152,200
Desvío std	5,419	8333,394
CV	0,008	0,237
Máx. dif	18	21524
Máx dif %	2,542	61,231

Análisis estructural de P5

Este PP tiene una única ocurrencia en todo el genoma, de 180010 nucleótidos de longitud. La distancia que separa a cada uno de los miembros de f21, de 35152,2 nucleótidos en promedio, es altamente variable. La distancia máxima es 21524 nucleótidos mayor a la mínima (lo cual representa un 61 % de la distancia promedio), y el valor de CV es 2 órdenes de magnitud superior al observado en P4 (Tabla 10). Esta falta de constancia en la distancia entre miembros hace que P5 no pueda ser considerado un verdadero PP.

Resultado de la puesta a prueba del modelo

Las observaciones realizadas para los 24 novel patterns escogidos son consistentes con las 4 predicciones con las que se trabajó. Parte de los patterns resultaron ser dependientes entre sí, formando PP como los que se postulan en la predicción Pr1. El comportamiento de P1, P2, P3 y P4 se ajusta a las predicciones Pr2, Pr3 y Pr4. Esto nos permite decir que el modelo postulado en la hipótesis de trabajo lograría explicar exitosamente el origen de los patterns involucrados en estos 4 PP.

2.4.3- Validación evolutiva

Los resultados del apartado anterior permitieron concluir que el modelo de duplicación-divergencia propuesto en la hipótesis es capaz de explicar el origen de un conjunto de patterns, pudiéndose identificar las regiones genómicas duplicadas involucradas. En este apartado se pretende evaluar desde una perspectiva evolutiva si dichas regiones genómicas se comportan como verdaderas duplicaciones.

Determinación de la estructura de las DS que contienen a las ocurrencias de P1

Las ocurrencias de P1 resultaron estar ubicadas en una zona del genoma rica en regiones de posiciones indefinidas. En 8 de las 12 secuencias flanqueantes extraídas se

obtuvieron menos de 200 kb, siendo la menor de ellas de 27238 nucleótidos de longitud (Material Suplementario 5, Tabla S5).

El resultado de hacer todas las comparaciones posibles entre las secuencias 5' flanqueantes agrupó a las 6 ocurrencias de P1 en 3 grupos:

i) P1-1-1 y P1-1-4. Presentan entre sí una similitud del 99% hasta 82816 y 82835 nucleótidos río arriba, respectivamente.

ii) P1-1-2, P1-1-3 y P1-1-5-R. Presentan entre sí una similitud del 99% hasta 49158, 52173 y 52212 nucleótidos río arriba, respectivamente.

iii) P1-5-1-R. Presenta similitud del 95% con el resto de las ocurrencias, pero solo hasta 1400 y 4000 nucleótidos (aproximadamente) río arriba con i) y ii), respectivamente.

Al comparar la secuencia 5' flanqueante de cualquiera de los miembros de i) con la de cualquiera de los miembros de ii) se observó una similitud del 99% solo a lo largo de los primeros 2300 nucleótidos aproximadamente. Río arriba de esa primera zona se observaron alineamientos largos de 97% de similitud, pero orientados al reverso. Esta región de similitud con orientación reversa cubre aproximadamente los 40 kb río arriba de los miembros de ii), pero con una interrupción de aproximadamente 6 kb.

De manera similar al caso anterior, el resultado de la comparación de todas las secuencias 3' flanqueantes agrupó a las 6 ocurrencias en 3 grupos:

i) P1-1-1, P1-1-2, P1-1-4 y P1-1-5-R. Presentan similitud mayor al 98% hasta donde existe disponibilidad de datos (excepto en el caso de P1-1-4, donde la similitud se interrumpe 67493 nucleótidos río abajo existiendo aún más posiciones disponibles).

ii) P1-5-1-R. Presenta similitud de alrededor del 95% con los miembros de i). Esta similitud se interrumpe en la posición 2242 río abajo para retomarse 14102 nucleótidos más adelante.

iii) P1-1-3. Presenta similitud mayor al 98% con los miembros de i) y cercana al 95% con el miembro de ii), pero sólo a lo largo de los primeros 6500 nucleótidos (aproximadamente) río abajo.

En el caso de las secuencias 3' flanqueantes no se observaron grandes regiones de alineamientos de orientación reversa.

Teniendo en cuenta todos los alineamientos obtenidos de las comparaciones mencionadas arriba se dedujo una arquitectura para las DS que contienen a las ocurrencias de P1. La estructura postulada se esquematiza en la Figura 6. Pueden identificarse características estructurales que diferencian y agrupan a las 6 DS. La inversión de una gran región río arriba de la ocurrencia de P1 (flecha azul) es característica de las DS 2, 3 y 5, al igual que la inserción *a*. Las inserciones *c* y *b*, en cambio, están presentes solo en la DS 6. Estas particularidades podrían usarse como caracteres para deducir las relaciones de parentesco entre las DS (véase el apartado *Reconstrucción de la historia evolutiva de la DS que contiene a las ocurrencias de P1*, más adelante).

Tabla 11. Ubicación de las 6 ocurrencias de la DS que contiene a P1 en el genoma. Se informa además la longitud de cada una, calculada por diferencia entre la posición final y la inicial.

n° de DS	Ocurrencia de P1	Cromosoma	Posición inicial	Posición final	Longitud
1	P1-1-1	1	143644525	143771003	126478
2	P1-1-2	1	144274483	144401745	127262
3	P1-1-3	1	144451746	144526797	75051
4	P1-1-4	1	147832033	147998780	166747
5	P1-1-5-R	1	149521690	149661090	139400
6	P1-5-1-R	5	49958509	50058284	99775

De la estructura de estas 6 porciones genómicas duplicadas se desprenden las posiciones iniciales y finales de cada una en el genoma humano, a partir de las cuales puede calcularse su longitud (Tabla 11).

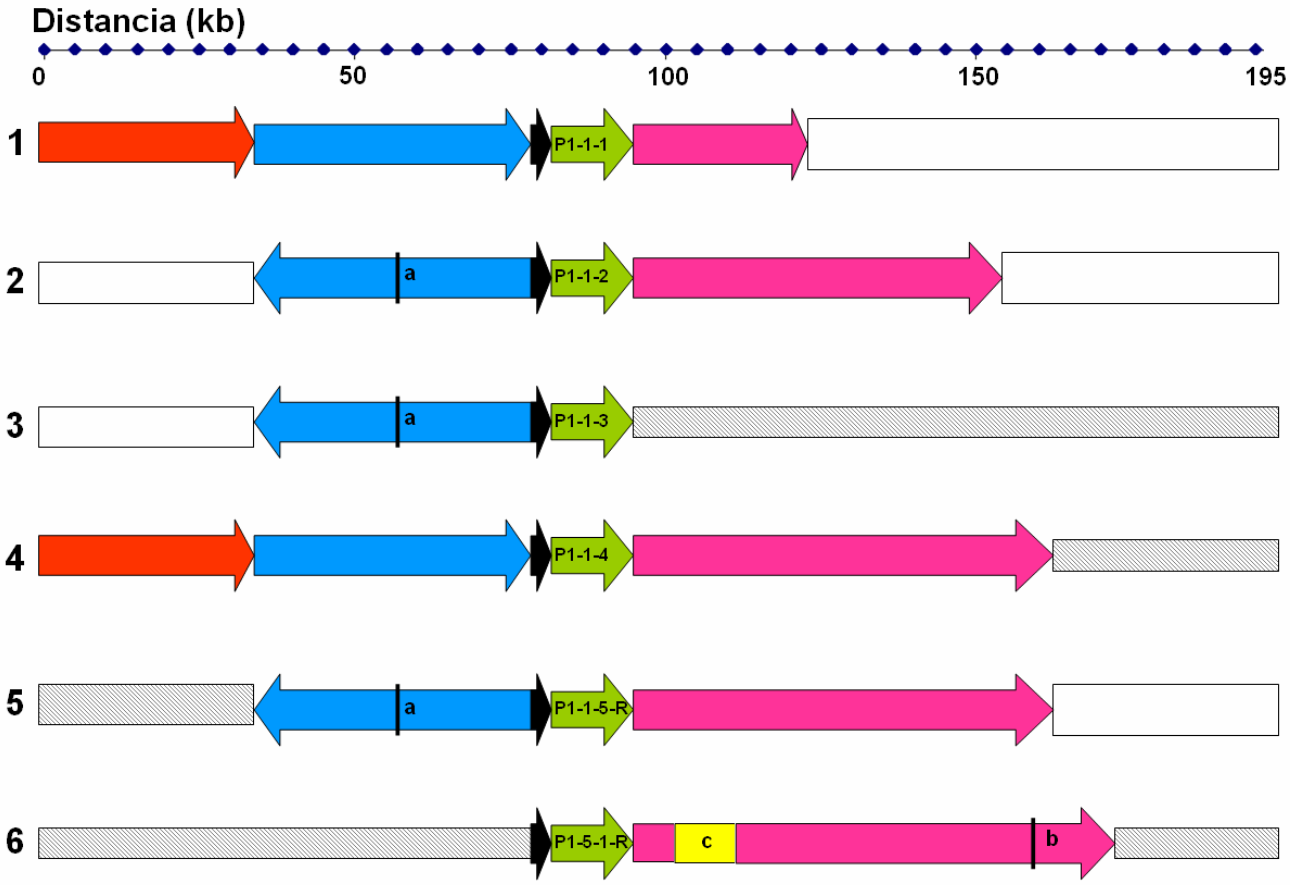

Figura 6. Estructura postulada para las 6 DS que contienen a las ocurrencias de P1. Las flechas en verde corresponden a las ocurrencias de P1. Las flechas de color denotan regiones altamente similares, y su orientación indica el sentido (derecho o reverso) de la similitud. Los bloques en gris denotan ausencia de alta similitud con el resto de las regiones. Los bloques blancos corresponden a regiones del genoma en las que no hay disponibilidad de datos. *a*: inserción de aproximadamente 6 kb. *b:* inserción de aproximadamente 6 kb. *c:* inserción de aproximadamente 14 kb.

Una consecuencia de la gran extensión de las DS que contienen a las ocurrencias de P1 es que P3 también estaría contenido en ella: la distancia entre una ocurrencia de P1 y la inmediatamente contigua de P3 (aproximadamete 47800 nucleótidos, Tabla 3) es siempre menor a la cantidad de posiciones que la DS posee río abajo de ella (flechas magenta, Figura 6). Cabe preguntarse entonces por qué inicialmente no pudo definirse un único PP que contuviera P1 y P3. El hecho de que en la Tabla 3 no se observen miembros de f2 y f22 río abajo de P1-1-1 tendría que ver con que en esa zona solo se cuenta con 27238 posiciones disponibles (Material Suplementario 5, Figura S5 y Figura 6). Todas las otras ocurrencias de P1 presentan una ocurrencia de P3 río abajo. Podríamos predecir entonces que los datos no disponibles río abajo de P1-1-1 seguramente contienen algún parálogo de los patterns 2 y 22. Restaría explicar por qué P1-1-4 y P1-1-5-R presentan río abajo más de una ocurrencia de P3.

Una posible explicación sería la duplicación en tándem de regiones que contendrían a P3-1-2 y P3-1-6-R (respectivamente), posterior a los eventos de duplicación que habrían generado las DS que se postulan en la Figura 6.

Todas las DS que contienen a P1 se ubican relativamente cerca del centrómero. Las del cromosoma 1 se encuentran en una región acotada del brazo *q* (regiones q21.1 y q21.2), mientras que la del cromosoma 5 se ubica en la región pericentromérica (Figura 7).

Todas las DS del cromosoma 1 contienen elementos anotados, la mayoría de los cuales han sido validados experimentalmente (Material Suplementario 7, Tabla S6). La DS del cromosoma 5 contiene un único elemento anotado. Se trata de un gen reportado como codificante bajo el nombre de PARP8 (Entrez ID: 79668; Ensembl: ENSG00000151883), cuyo producto proteico está validado experimentalmente. Este gen pertenece a la familia PARP, compuesta por enzimas poli (ADP-ribosa) polimerasas. Todos los grupos de eucariotas poseen algún miembro de esta familia génica, cuya función ancestral habría estado relacionada con la reparación de ADN dañado. En animales existen varios genes PARP que cumplen esta función ancestral, pero se han descrito otros cuya actividad no está relacionada con ella. PARP8 pertenece a este segundo grupo, aunque su rol biológico no ha sido estudiado en profundidad [7].

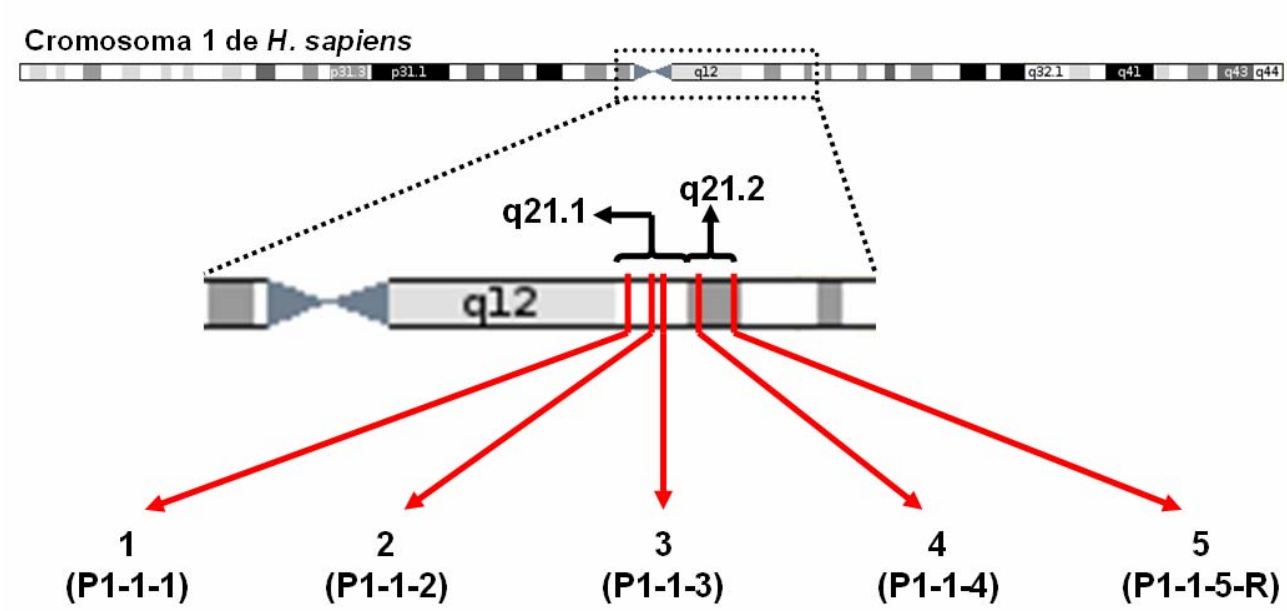

Figura 7. Ubicación de las 6 ocurrencias de P1 en el cromosoma 1.

Búsqueda de ortólogos de las ocurrencias de P1

El resultado de la búsqueda de homólogos de P1 en otras especies fue independiente de cuál de sus ocurrencias se utilizara como secuencia query.

La Tabla 12 resume la información de todos los homólogos hallados. Las 12 especies de mamíferos analizadas presentaron un elemento homólogo a la secuencia de alguna ocurrencia de P1 en una región sinténica al cromosoma 5 del genoma humano. Los miembros de la superfamilia Hominoidea (*i.e.* Humano, Chimpancé, Gorila, Orangután y Gibón) son los únicos que presentaron otros homólogos además de éste, siempre en el cromosoma 1.

En todas las especies se observa que el homólogo ubicado en la región sinténica al cromosoma 5 del genoma humano es parte de un elemento reportado como codificante bajo el nombre de PARP8. Considerando esto y teniendo en cuenta las relaciones sinténicas de las regiones involucradas, postulamos que los homólogos nº 6, 9, 13, 14, 17, 19, 20, 21, 22 y 24 serían ortólogos entre sí (Tabla 12).

Nuestra metodología de búsqueda de homólogos no arrojó ningún resultado para el genoma del pollo. Sin embargo, este genoma contiene un gen reportado como homólogo de PARP8, bajo el mismo nombre, ubicado en el cromosoma sexual Z (Entrez ID: 427198; Ensembl: ENSGALG00000014880). Esta discrepancia se debe a la alta divergencia entre los genes PARP8 de humano y pollo, atribuible a las regiones intrónicas. Como P1-5-1-R contiene tanto a los intrones como a los exones de PARP8, el genoma del pollo no contiene entonces ninguna secuencia altamente similar a esta ocurrencia de P1, a pesar de que este gen sí tiene su ortólogo en el genoma del pollo. De hecho, al comparar la secuencia aminoacídica de los genes PARP8 de ambas especies se observó una similitud del 93 %.

En el caso del genoma de Gorila, los tres homólogos del cromosoma 1 presentan la particularidad de ser notablemente más cortos que el resto. En principio esto podría sugerir que en realidad no se trata de homólogos legítimos, sino de elementos que son altamente similares a parte de la secuencia de las ocurrencias de P1 pero no por haberse generado a partir de la misma secuencia ancestral. Sin embargo, al analizar el entorno genómico de estas secuencias se puede ver que la menor longitud de los alineamientos no se debe a que la

elevada similitud con las ocurrencias de P1 caiga, sino a que la secuencia se ve interrumpida por regiones de posiciones indefinidas. Teniendo en cuenta esto, resulta razonable pensar que se trata de homólogos *bona fide* de los cuales no se cuenta con la secuencia completa.

Tabla 12. Secuencias homólogas a las de las ocurrencias de P1 halladas en los 13 genomas incluidos en el estudio. Para cada homólogo se informa su ubicación en el genoma en cuestión y el cromosoma humano al que dicha región es sinténica. También se muestran las 6 ocurrencias del genoma humano, ya descriptas en la Tabla 5.

n° de homólogo	Especie	Cromosoma	Sinténico al humano	Posición inicial	Longitud
1	Humano	1	--	143727341	16424
2	Humano	1	--	144323666	16423
3	Humano	1	--	144503918	16420
4	Humano	1	--	147914849	16438
5	Humano	1	--	149577138	16615
6	Humano	5	--	49963447	16458
7	Chimpancé	1	1	114658964	14176
8	Chimpancé	1_random	1	4873075	16471
9	Chimpancé	5	5	65325058	16549
10	Gorila	1	1	124188559	2779
11	Gorila	1	1	124579578	2413
12	Gorila	1	1	127646283	6044
13	Gorila	17	5	46715142	14805
14	Orangután	5	5	48839681	16427
15	Orangután	1	1	102161959	16492
16	Gibón	GL397354.1	1	655294	14832
17	Gibón	GL397396.1	5	289274	16489
18	Macaco	6	1	48211509	17094
19	Titi	2	5	156980469	16868
20	Tarsero	GeneScaffold_4957	5	1401	23298
21	Lemur	GeneScaffold_2628	5	97465	17976
22	Galágido	GeneScaffold_3103	5	42552	16172
23	Ratón	13	5	117794924	18395
24	Perro	4	5	67460081	19275
	Pollo	--	--	--	--

Reconstrucción de la historia evolutiva de la DS que contiene a las ocurrencias de P1

El análisis de la presencia de homólogos en las distintas especies permitió postular la reconstrucción que se esquematiza en la Figura 8. El estado ancestral del grupo de mamíferos estudiados (representantes del grupo de placentarios Boroeutheria) sería la presencia de una única copia de lo que en *H. sapiens* hemos llamado P1-5-1-R, formando parte del gen PARP8.

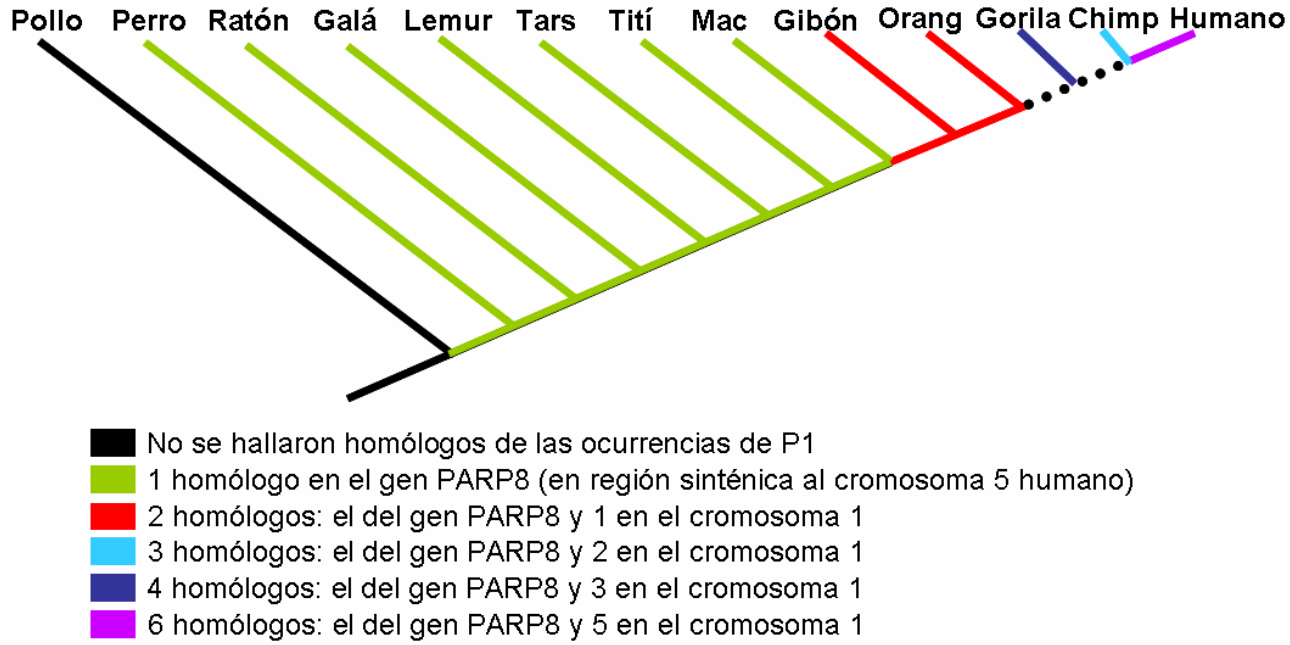

Figura 8. Reconstrucción de la historia de la DS que en humano contiene a P1. El caracter que se mapea en el árbol filogenético es "cantidad de homólogos de la secuencia de P1 humano". Los estados del caracter se denotan mediante los distintos colores de las ramas. La línea punteada indica que para la rama en cuestión existe más de una alternativa de mapeo. *Galá*: galágido; *Tars*: tarsero; *Mac*: macaco; *Orang*: orangután; *Chimp*: chimpancé.

Un primer evento de duplicación podría mapearse en la rama posterior a la divergencia de Macaco. Se trataría de una duplicación con salto de cromosoma (del sinténico al 5 de humano al sinténico al 1 de humano). Como consecuencia de ello los genomas de Gibón y Orangután presentan, además de la contenida en PARP8, una segunda copia en su cromosoma 1. La segunda tanda de eventos de duplicación se postula luego de la divergencia de Orangután. Las nuevas copias se ubican en el cromosoma 1, lo cual sugiere que podrían haberse generado a partir de la copia preexistente en ese mismo cromosoma (se asume que la duplicación es menos probable si implica salto de cromosoma [9, 37]). Entre este nodo y la

divergencia del linaje humano hay dos alternativas igualmente parsimoniosas (rama punteada):

i) Luego de la divergencia de Orangután se generaron 2 nuevas copias en el cromosoma 1, y el linaje del Chimpancé perdió una de ellas.

ii) Luego de la divergencia de Orangután se genera 1 nueva copia en el cromosoma 1, y en el linaje de Gorila se genera una 1 copia más en el mismo cromosoma

Una última serie de eventos de duplicación tendría lugar en el linaje humano luego de la divergencia con el chimpancé. Al igual que antes, todas las copias nuevas que aparecen en este linaje se encuentran en el cromosoma 1, sugiriendo que deben su origen a algunas de las preexistentes en ese cromosoma.

Los resultados del análisis filogenético son consistentes con esta hipótesis evolutiva. Las relaciones de parentesco entre los homólogos presentes en los genomas del Infraorden Catarrhini se muestran en la Figura 9 y Material Suplementario 6, Figura S2. En ambas reconstrucciones se observa que los homólogos presentes en el cromosoma sinténico al 5 de humano (*i.e.* las asociadas al gen PARP8) comparten un ancestro común más cercano, y el orden de su divergencia coincide con el de las especies (Figura 5). Lo mismo ocurre con las copias del cromosoma sinténico al 1 de humano, que además serían de divergencia más tardía que las anteriores, lo cual es consistente con el hecho de que la versión ancestral de la DS sea la asociada a PARP8. Esto también apoya la idea de que la proliferación de parálogos en el cromosoma 1 luego de la divergencia de Orangután se debe a eventos de duplicación a partir del parálogo preexistente en ese cromosoma. Siendo así, en la historia de esta DS habría ocurrido un único evento de salto de cromosoma, luego de la divergencia de Macaco.

Cabe aclarar que en la reconstrucción por Máxima Parsimonia (Material Suplementario 6, Figura S2) la relación entre las copias del genoma de Chimpancé entra en conflicto con el orden de divergencia de las especies. Hay que tener en cuenta que además de ser el único caso conflictivo, el nodo en cuestión es el único que presenta un bajo valor de soporte.

Para la reconstrucción de la secuencia de eventos de duplicación que originó las 6 ocurrencias de P1 en el genoma humano se puede hacer uso de dos fuentes de datos ya mencionadas: la secuencia nucleotídica de las ocurrencias y los rearreglos genómicos presentes en las DS que las contienen.

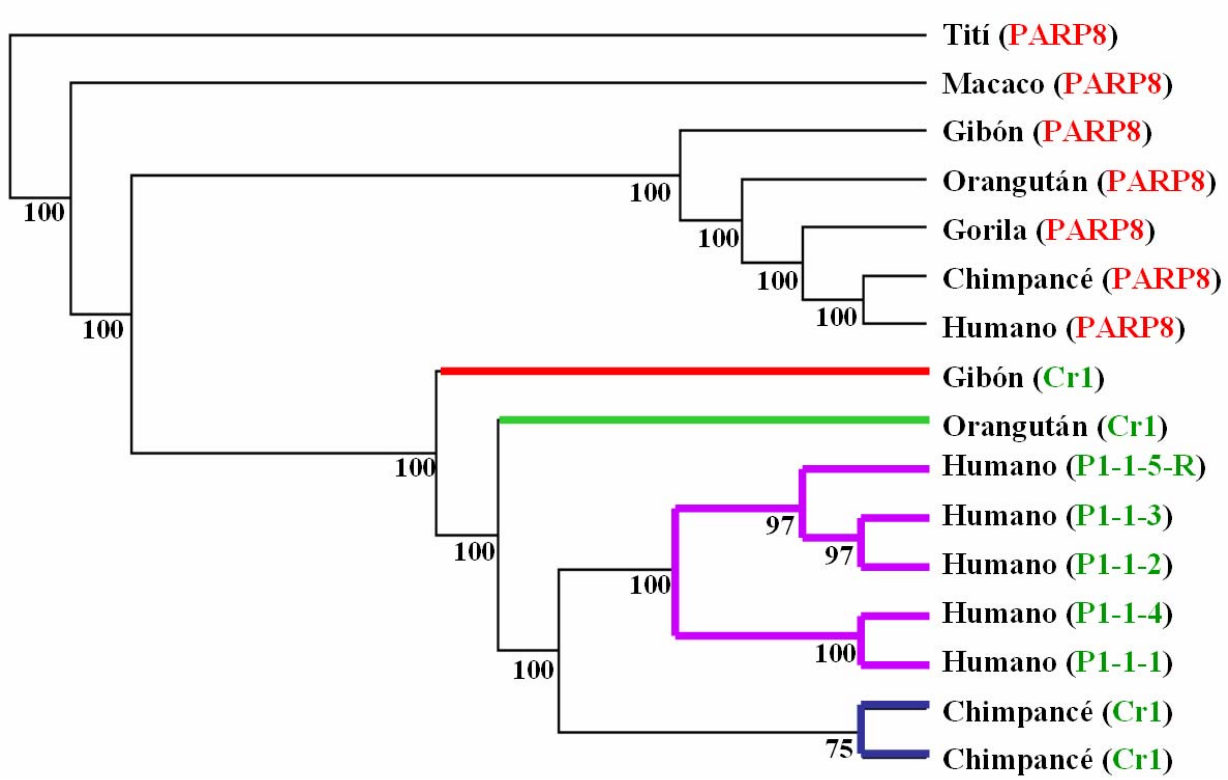

Figura 9. Reconstrucción de la filogenia de los homólogos de P1 del Infraorden Catarrhini mediante el método de Máxima Verosimilitud. Se utilizó el homólogo del genoma de Titi como grupo externo. *PARP8*: homólogo asociado al gen PARP8, en el cromosoma sinénico al 5 humano. *Cr1*: homólogo ubicado en el cromosoma sinténico al 1 humano. Las líneas de color indican las ramas correspondientes a cada especie. Los números junto a los nodos indican el valor de soporte sobre 1000 réplicas mediante bootstrap. Se utilizó el modelo TrN+G con los siguientes parámetros: Nst=6; Rmat=(1.0000 3.8872 1.0000 1.0000 4.9813); Gamma=1.0231; Pinv=0.

La estructura de las DS postulada en la Figura 6 permite definir 2 caracteres discretos para agrupar las 6 ocurrencias de P1: i) presencia de una región invertida río arriba de la ocurrencia de P1 (flechas azules, Figura 6), en adelante *inversión 5'* y ii) presencia de la inserción *a*. Estos caracteres agrupan a P1-1-2 con P1-1-4, al resto de las ocurrencias del cromosoma 1 entre sí, y a P1-5-1-R aparte. El análisis a nivel de la secuencia es consistente con estos agrupamientos. Al reconstruir la filogenia de estas ocurrencias mediante el método de Máxima Parsimonia (Figura 10) se observa que P1-1-5-R habría sido la primera copia en aparecer en el cromosoma 1, y que las otras 4 se habrían generado más recientemente a partir

de ella. No puede establecerse un orden unívoco para los eventos de duplicación, pero sí puede verse que la inversión 5' y la desaparición de la inserción *a* (que a la luz de la historia evolutiva inferida debería ser considerada una deleción) tienen lugar en la rama que da origen a P1-1-3 y P1-1-2. Una reconstrucción mediante el método de Máxima Verosimilitud arrojó un árbol de idéntica topología, que no se muestra.

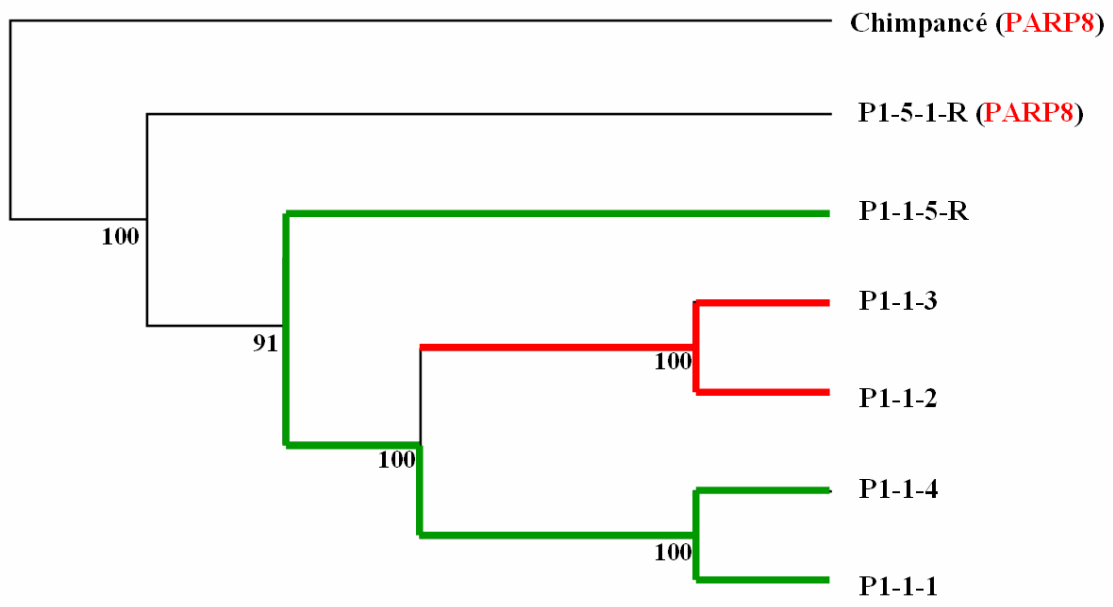

Figura 10. Reconstrucción de la filogenia de las secuencias de las 6 ocurrencias de P1 del genoma humano, mediante el método de Máxima Parsimonia. Se utilizó el homólogo del cromosoma 5 de Chimpancé como grupo externo. *PARP8*: homólogo asociado al gen PARP8, en el cromosoma sinénico al 5 humano. Las líneas rojas indican presencia de la inversión 5' y ausencia de la inserción *a*. Las líneas verdes indican ausencia de la inversión 5' y presencia de la inserción *a* (Figura 6). Los números junto a los nodos indican el valor de soporte sobre 10000 réplicas mediante bootstrap.

2.4.4- Comparación de los resultados con bases de datos de DS

En los dos apartados anteriores se lograron definir y validar 6 DS a partir de un subconjunto de los 24 patterns incluidos en el análisis. En este apartado se pretende evaluar si estas DS ya habían sido reportadas mediante los métodos de detección de DS existentes o si, por el contrario, eran desconocidas hasta el momento. Nos referiremos a las 6 DS utilizando la numeración de la Figura 6 y la Tabla 11.

Cada una de las dos primeras DS coincide en sus posiciones inicial y final con una DS reportada en la base de datos. Esta coincidencia es casi exacta: las posiciones varían en una única unidad.

Las DS nº 3, 4 y 5 coinciden en su posición inicial con alguna DS reportada, pero la base de datos reporta además otras DS inmediatamente contiguas que se extienden más allá de su posición final. En la Figura 11a se muestra una toma del navegador de genomas para el caso de la DS nº 3, a modo de ejemplo.

En el caso de la DS nº 6 se observa que su posición inicial coincide con las de DS reportadas, pero que se extiende mucho más allá de la posición donde éstas terminan (Figura 10b).

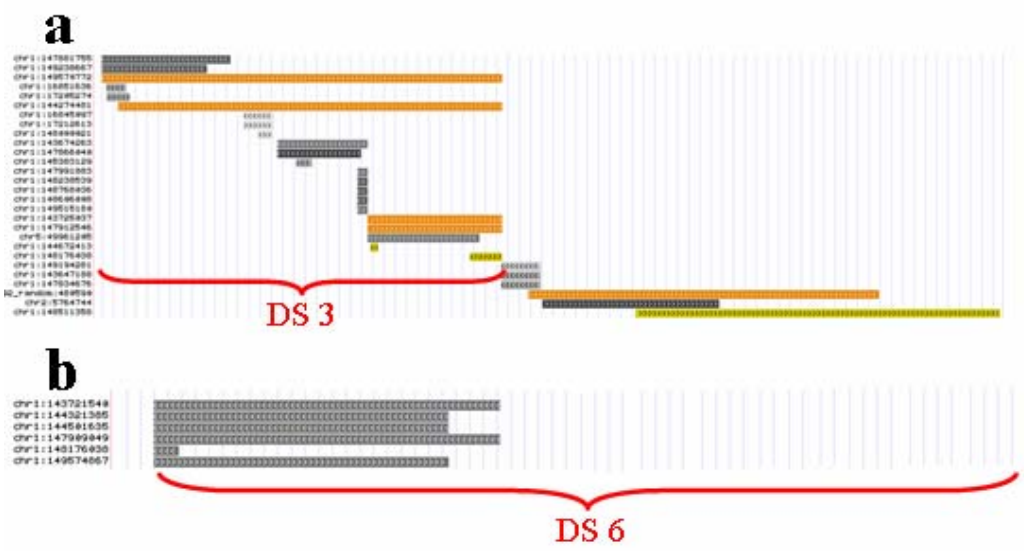

Figura 11. Capturas de la pantalla del navegador de genomas de la UCSC para dos regiones del cromosoma 1. Cada rectángulo de color corresponde a una DS de la base de datos. Para cada una de ellas figura, a la izquierda, la ubicación de su pareja (el formato de la base de datos de DS se describe en Material Suplementario 8, Figura S3 y en el apartado 3.1). En rojo se marca la extensión de dos de las DS definidas por nosotros a partir de los 24 novel patterns más largos del cromosoma 1. *a*: la DS 3 coincide en su extensión solo con una parte de la región duplicada. *b*: la DS 6 se extiende más allá del límite de las DS de la base de datos.

Estos resultados respaldan nuestro hallazgo de DS a partir de los 24 patterns estudiados, ya que en los 6 casos se hallaron DS ya reportadas dentro de las posiciones correspondientes. Sin embargo, las discrepancias en la longitud de las mismas indican que existe una cierta complementariedad entre las dos metodologías de detección de DS. Una de las DS que hallamos es más extensa de lo que se reporta, y hallamos otras 3 cuya extensión reportada sería mayor a la que determinamos. Para una discusión en profundidad véase el apartado 4.4, en *Discusión*.

2.5- Conclusiones

El conjunto de patterns escogido para el análisis se comporta de manera consistente con las 4 predicciones con las que se trabajó. Se observó que 14 de los 24 patterns estudiados están formando parte de PP que resultaron ser DS. Esto no implica que los 10 restantes no formen parte de DS, sino que la metodología empleada no fue capaz de determinarlo.

Las DS halladas mostraron un patrón evolutivo que se ajusta a lo esperado para elementos genómicos que sufren eventos de duplicación sucesivos. Esto apoya la idea de que las DS que contienen a los patterns estudiados son realmente el resultado de procesos de duplicación-divergencia.

Adicionalmente, todas las DS halladas resultaron coincidir total o parcialmente con DS reportadas previamente mediante metodologías completamente distintas.

En conjunto, estos resultados nos permiten decir que el modelo de duplicación-divergencia propuesto en nuestra hipótesis es capaz de explicar el origen de al menos una parte de los patterns detectados por el grupo Kapow.

3- PUESTA A PRUEBA DE LA HIPÓTESIS A PARTIR DE BASES DE DATOS COMPLETAS

La primera puesta a prueba de la hipótesis nos permitió decir que el modelo de duplicación-divergencia de grandes bloques genómicos es capaz de dar cuenta del origen de una parte de los patterns existentes (Capítulo 2). A través de este segundo abordaje se pretende determinar el poder explicativo de nuestra hipótesis a escala genómica. Para ello nos proponemos responder a la siguiente pregunta: ¿qué proporción de todos los patterns existentes son explicados por el modelo propuesto?

Como la definición de pattern no tiene implícita una longitud mínima, sabemos de antemano que el modelo nunca podría ser capaz de explicar absolutamente todos los patterns. Piénsese por ejemplo en el pattern *AT*, de longitud 2: se conocen muchas secuencias no pertenecientes a DS que contienen este dinucleótido. Es por este motivo que para esta puesta a prueba del modelo elegimos utilizar una predicción que tuviera en cuenta la longitud de los patterns: “la cantidad de ocurrencias de patterns explicadas por el modelo debe aumentar al incrementarse la longitud de los patterns analizados”.

3.1- Materiales y métodos

Para responder esta pregunta se evaluó cuántos de los patterns presentes en todo el genoma estaban asociados a DS. Computacionalmente fue necesario contar la cantidad de ocurrencias de patterns de todo el genoma que están contenidas dentro de DS previamente reportadas.

Base de datos de DS

Se utilizó la base de datos generada por el grupo de investigación del Dr. Evan Eichler (Eichler Lab, Departamento de Ciencias Genómicas de la Escuela de Medicina, Universidad de Washington), gentilmente suministrada por él y por el Lic. John Huddleston.

Esta base de datos contiene todas las DS detectadas mediante la aplicación de las metodologías desarrolladas por el grupo [1, 2], y son las mismas que aparecen reportadas en el navegador de genomas de la Universidad de California, Santa Cruz (UCSC genome browser) [17]. Estas metodologías de detección de DS encuentran pares de secuencias de más de 1000 kb que presentan una similitud de al menos 90%. Como consecuencia de ello, las bases de datos generadas presentan a las DS siempre de a pares, aunque se trate de familias de duplicaciones con más de dos miembros (Material Suplementario 8, Figura S3).

En la base de datos utilizada cada fila corresponde a una DS y las columnas informan distintos parámetros calculados, de los cuales solamente utilizamos los 6 que se muestran en la Tabla 13.

Tabla 13. Formato de la sección que se utilizó de la base de datos de DS

Nº de cromosoma del 1º miembro del par de DS	**Posición inicial del 1º miembro del par de DS**	**Posición final del 1º miembro del par de DS**	**Nº de cromosoma del 2º miembro del par de DS**	**Posición inicial del 2º miembro del par de DS**	**Posición final del 2º miembro del par de DS**

Esta base de datos se transformó cortando las 3 columnas correspondientes al 2º miembro del par de DS y pegándolas a continuación de la última fila de las 3 primeras columnas. Se obtuvo así un listado de 62273 intervalos (que ya no contiene información sobre pares de DS) no ordenados por posición, con el formato que se muestra en la Tabla 14.

Tabla 14. Formato del listado de intervalos generado por transformación de la base de datos de DS

Nº de cromosoma	Posición inicial	Posición final

Base de datos de patterns

Los patterns computados por el grupo Kapow están almacenados en una base de datos en la que cada pattern está especificado en dos líneas: una que describe explícitamente su secuencia, el número total de ocurrencias y su longitud, y otra que lista la posición inicial de cada una de las ocurrencias. Por ejemplo:

TGGATAACTTTTT #2 (13)
>25103413 >16670527

Para cada cromosoma existe un archivo que lista a los patterns y sus ocurrencias por longitud creciente de pattern, pero sin un orden dentro las ocurrencias listadas. Se trabajó con las bases de datos de patterns de longitud mayor o igual a 40.

Cálculo de cobertura de bases de datos (llevado a cabo por el Lic. Pablo Barenbaum, del grupo Kapow)

Para calcular la cantidad total de ocurrencias de patterns que caen dentro de la base de datos de DS se llevó a cabo una intersección de intervalos (Figura 12): se recorrieron secuencialmente todas las ocurrencias de cada uno de los patterns, evaluando para cada una si estaba contenida dentro de alguna de las DS de la base de datos. Consideramos a una ocurrencia de un pattern como contenida dentro de una DS si se verificaba en simultáneo que: i) su posición final fuera mayor que la posición inicial de la DS, y ii) su posición inicial fuera menor que la posición final de la DS (Figura 12b).

Como resultado de este procedimiento se obtuvo un archivo con el formato que se muestra en la Tabla 15. El archivo de salida informa además la cantidad total de DS, patterns, ocurrencias y ocurrencias en DS.

Tabla 15. Formato del archivo de salida de la intersección de intervalos llevada a cabo para calcular la cobertura de DS por patterns.

Longitud de pattern	Cantidad de ocurrencias en DS	Cantidad de ocurrencias fuera de DS	Cantidad total de ocurrencias

Este procedimiento se repitió para los 24 archivos de patterns provistos por el grupo Kapow (1 por cada cromosoma). Se compilaron los datos de los 24 archivos de salida en uno solo, a partir del cual se calculó la proporción de ocurrencias de patterns contenidas en DS para cada longitud de pattern.

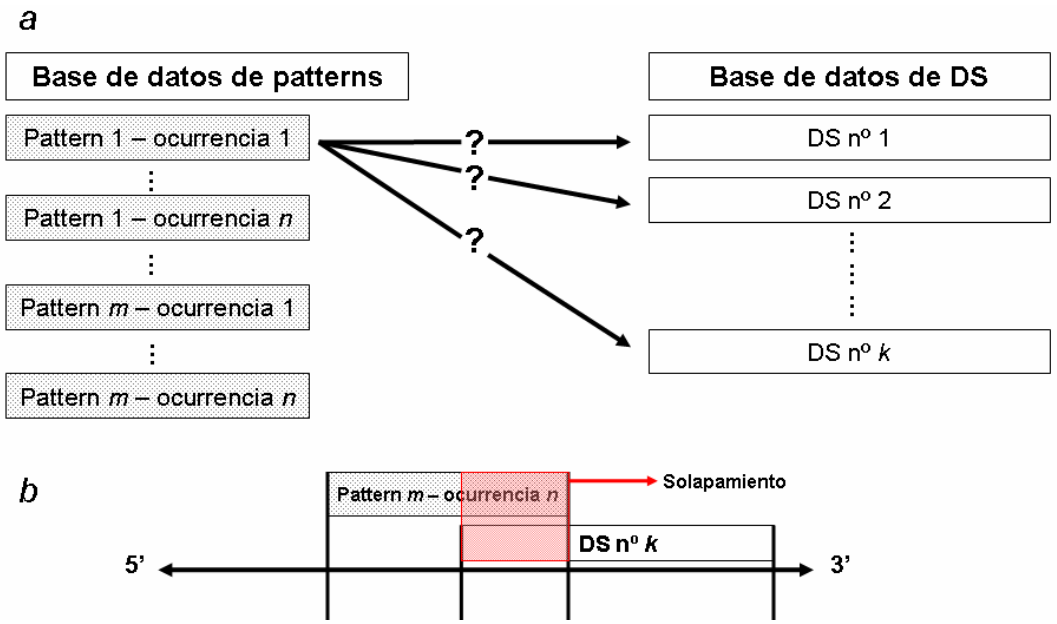

Figura 12. Esquema de la intersección de intervalos llevada a cabo para calcular la cobertura de DS por patterns. *a*: para cada ocurrencia de pattern se determinó si presentaba solapamiento con alguna de las DS de la base de datos. *b*: esquema gráfico de las condiciones que debía verificar una ocurrencia *n* para ser considerada solapante a una DS *k*.

3.2- Resultados

Existe una cantidad considerable de valores de longitud de pattern que verifican una alta proporción de ocurrencias contenidas dentro de DS. Muchos de los patterns (entre ellos los correspondientes a las 7 longitudes mayores) tienen todas sus ocurrencias dentro de DS, aunque también hay una cantidad considerable de longitudes con valores de proporción intermedios, e incluso con ninguna de sus ocurrencias solapadas con alguna DS (Figura 13).

El gráfico de la Figura 13 no permite poner a prueba el modelo a partir de nuestra predicción porque sólo da información sobre la proporción de cobertura de DS por ocurrencias patterns para cada valor de longitud por separado. Por ese motivo se utilizaron los mismos datos para calcular la proporción acumulada y graficarla en función de la longitud de pattern, como se muestra en la Figura 14. En este caso, los valores del eje de las ordenadas corresponden a la proporción de ocurrencias contenidas en DS, para patterns de longitud igual o mayor al valor del eje de las abscisas.

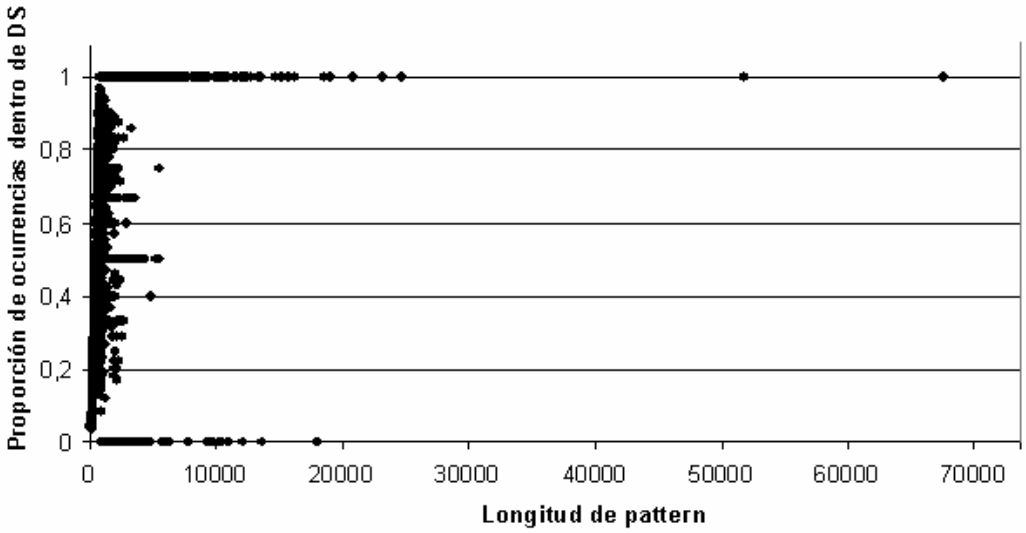

Figura 13. Proporción de ocurrencias de patterns contenidas en DS en función de la longitud de pattern.

La cantidad de ocurrencias de patterns que se solapan con la base de datos de DS aumenta a medida que se incrementa la longitud de pattern para longitudes menores a 1103. Para longitudes entre 1103 y 18517 se observa una oscilación entre 0,79 y 0,91, y para longitudes mayores o iguales a 18517 la proporción alcanza un valor de 1 (Figura 12a). Cualquier valor de longitud mayor o igual a 451 verifica que más de la mitad de las ocurrencias de patterns están contenidas en alguna DS (Figura 12b).

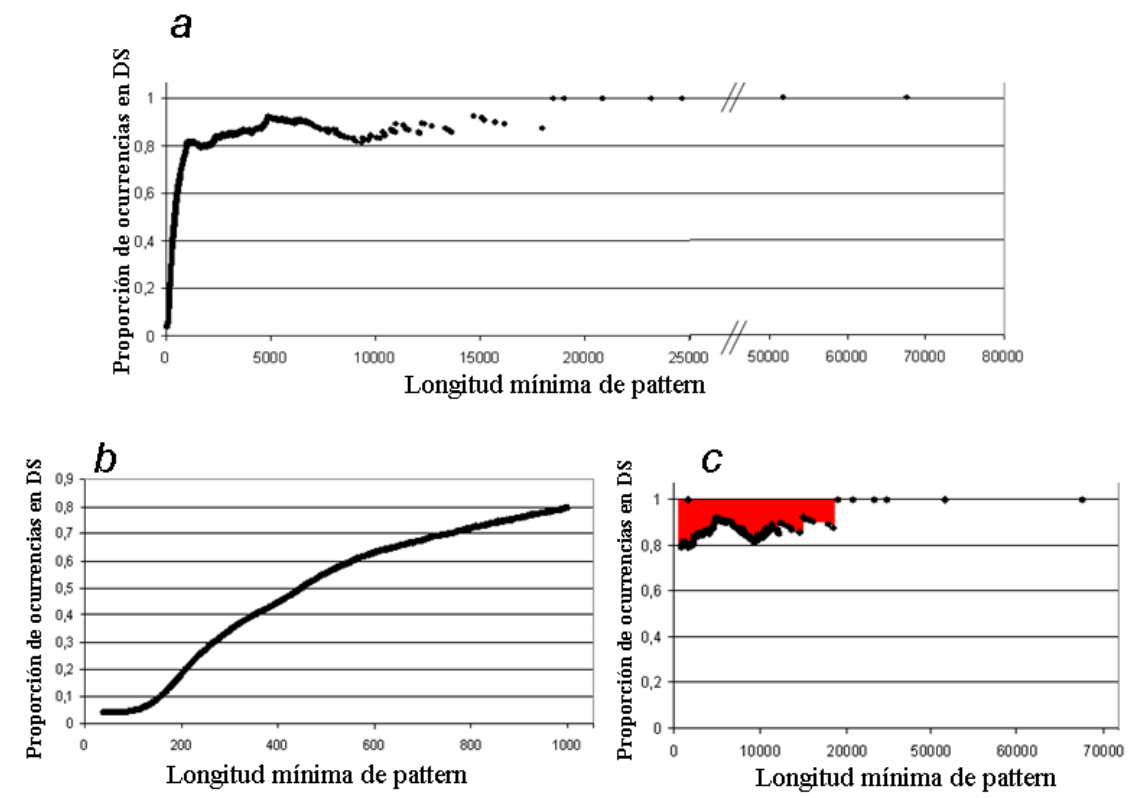

Figura 14. Proporción de ocurrencias de patterns contenidas en DS en función de la longitud mínima de pattern. *a*: para todas las longitudes de patterns incluidas en el análisis. *b*: para longitudes menores a 1000. *c*: para longitudes mayores a 1000. El área marcada en rojo señala la proporción de ocurrencias de patterns de longitud mayor a 1000 que no están contenidas en DS.

3.3- Conclusiones

Los resultados del análisis de proporción de ocurrencias de patterns contenidas en DS son consistentes con nuestra predicción. Si consideramos a dicha proporción como un estimador del poder explicativo de nuestra hipótesis a escala genómica, podemos decir que el modelo que proponemos explicará mejor el origen de un pattern cuanto mayor sea su longitud. Si se toma un pattern de una longitud dada, nuestros resultados permiten estimar la probabilidad de que sus ocurrencias deban su origen a un evento de generación de DS.

Hay que tener en cuenta que esta manera de estimar el poder explicativo del modelo tiene sentido únicamente para patterns de longitudes menores a 1000: cualquier pattern de longitud mayor o igual a ese valor será por definición una DS, cuando son justamente las DS las utilizadas como estimadores de la cantidad de secuencias duplicadas. Así, para longitudes mayores a 999 la interpretación de la proporción de patterns contenidos en DS como un estimador de los contenidos en bloques duplicados llevaría a un razonamiento circular (sin embargo, véase el apartado 4.4 más adelante para el análisis de los patterns de longitudes mayores).

Las metodologías de detección de DS actuales, a partir de las cuales se generaron las bases de datos, se basan en una definición precisa de DS. De esta definición se desprende que toda secuencia de más de 1000 posiciones que aparezca al menos 2 veces de manera idéntica será considerada una DS. En consecuencia, cualquier pattern con longitud igual o mayor a 1000 será por definición una DS. Sin embargo, nuestros resultados muestran que existen patterns de longitud mayor a 1000 que no se solapan con la base de datos de DS (Figura 12c). Esto implica que existen DS que no han sido detectadas por las metodologías actuales, y que se pusieron en evidencia a partir del cómputo de patterns mediante el algoritmo del grupo Kapow.

4- DISCUSIÓN

La aplicación al genoma humano del algoritmo de búsqueda de patterns desarrollado por el grupo Kapow dio resultados inesperados. La cantidad y longitud de secuencias repetidas de manera idéntica asombró a los miembros del grupo y despertó el interés por darles una explicación biológica. Fue ese interés el que dio inicio a este trabajo, que se llevó a cabo con la participación de miembros de Kapow y del Laboratorio de Evolución.

La motivación inicial fue lograr explicar estas secuencias repetidas en términos de alguna función biológica responsable de su conservación. Finalmente se optó por buscar un modelo sencillo que pudiera dar cuenta del origen de las repeticiones sin necesidad de contar con información sobre su función.

4.1- Un modelo capaz de explicar el origen de patterns

El modelo postulado se basa en la acción secuencial de dos mecanismos biológicos conocidos: i) la duplicación de regiones genómicas y ii) la divergencia de las duplicaciones generadas por adquisición de mutaciones propias de cada una. Cuando un evento de duplicación da origen a dos secuencias, inicialmente idénticas, éstas comienzan a divergir. En cada punto del proceso de divergencia existirán regiones de las dos secuencias hermanas en las que aún no habrán ocurrido mutaciones, conservando entonces la identidad. Según el modelo propuesto, los patterns hallados por el grupo Kapow son esas regiones que aún no han divergido.

Este modelo logra explicar el origen de al menos una parte de los patterns observados. El análisis en profundidad de un conjunto de 24 patterns mostró que el comportamiento de muchos de ellos se ajusta a lo esperado para el escenario de duplicación y divergencia de una secuencia más larga que los contiene. Más aún, las secuencias largas que contienen a estos patterns (nos hemos referido a ellas como DS, pero sin embargo véase el apartado 4.4, más adelante) presentan ellas mismas un patrón evolutivo propio (cuya profundidad en la

dimensión filogenética va más allá del genoma humano) que respalda la idea de que deban su existencia a eventos de duplicación seguidos de divergencia.

Una segunda estrategia metodológica permitió la puesta a prueba del modelo a escala genómica. Las bases de datos generadas por el grupo Kapow permiten tener acceso a la totalidad de los patterns. Al utilizar la base de datos de DS disponible actualmente como estimador de la cantidad total de regiones duplicadas en el genoma, la proporción de patterns contenidos en esas DS se utilizó como estimador del poder explicativo del modelo a escala genómica. El resultado de este análisis mostró que el poder explicativo del modelo aumenta al incrementarse el largo de pattern. No es sorprendente que el modelo falle al intentar explicar los patterns de menor longitud, ya que cualquier elemento genómico (duplicado o no) tendrá una probabilidad relativamente alta de contener alguna secuencia corta que aparezca repetida de manera idéntica en otras ubicaciones del genoma. Sin embargo, se observa que la proporción de patterns explicados por el modelo aumenta abruptamente al incrementarse su longitud. Esto permitiría estimar la probabilidad de que un pattern cualquiera deba su origen a un evento de duplicación de una secuencia más larga.

4.2- Alcances del modelo y escenarios complementarios o alternativos

La decisión de trabajar sobre un modelo independiente de la función biológica de los patterns implicó que nos limitáramos a explicar únicamente su origen, sacrificando nuestras aspiraciones a responder preguntas relacionadas con las causas de su permanencia en el genoma. A la hora de explicar el porqué de la identidad de secuencia entre las ocurrencias de un pattern existe un abanico de escenarios posibles, y el modelo que postulamos para su origen es compatible con más de uno de ellos. Los dos grandes escenarios que consideramos particularmente relevantes son:

a) Las ocurrencias de los patterns bajo estudio no cumplen ninguna función biológica. Cada uno de los bloques genómicos inicialmente duplicados sufre mutaciones en su secuencia, todas ellas neutras, que al fijarse generan la divergencia. Cuanto mayor sea el tiempo transcurrido desde el evento de duplicación, mayor será el grado de divergencia entre las duplicaciones. En consecuencia, la extensión de las regiones de las duplicaciones que aún no han divergido será también función del tiempo de divergencia. De esta manera, las ocurrencias de patterns largos podrían ser explicadas por la duplicación reciente de secuencias sin función biológica. Dentro de este escenario se esperaría que el largo de los patterns no funcionales hallados disminuyera a lo largo de la evolución del genoma.
b) Los patterns bajo estudio cumplen una función biológica importante. Dentro de los bloques genómicos inicialmente duplicados existen posiciones asociadas a algún proceso biológico, que estarían entonces sometidas a presiones selectivas. Según este escenario, la longitud de los patterns sería una consecuencia de las restricciones funcionales propias de las duplicaciones.

Existen evidencias de restricciones funcionales en DS [21, 22], lo cual respaldaría la idea de una distribución relativamente amplia de escenarios similares a b). El escenario a), por su parte, tiene como virtud su sencillez en términos de mecanismos biológicos involucrados, lo cual permitiría su utilización como modelo nulo (véase el apartado 4.6, más adelante).

Es necesario mencionar alguna explicación alternativa a nuestro modelo para el origen de los patterns. Una posibilidad es que los patterns se hayan duplicado como tales, no formando parte de secuencias más largas que luego divergieron. Nuestros resultados no se ajustan a este modelo, pero no podemos descartar que sea capaz de explicar alguna parte de los patterns.

4.3- Los patterns como definición matemática y entidad biológica

El concepto de pattern surge como definición matemática y es aplicable a cualquier secuencia de letras. Cuando la secuencia a la que se la aplica es un genoma, deberían surgir inmediatamente preguntas sobre la naturaleza de los patterns hallados. ¿Son entidades biológicas reales, o artefactos de la aplicación de una definición matemática?

Dentro del genoma de una especie, un pattern “llamativamente largo” (como los estudiados en el Capítulo 2) hace pensar que sus ocurrencias no pueden tener orígenes independientes. Patterns de estas características corresponderían a lo que biológicamente llamamos secuencias parálogas. En el Capítulo 2 de este trabajo hemos mostrado que existen patterns cuyas ocurrencias cumplen con los requisitos para ser consideradas secuencias parálogas. Sin embargo, las familias de parálogos a las que pertenecen estos patterns constan de otros miembros, que no fueron detectados por el algoritmo de Kapow por no ajustarse a la definición matemática de pattern. En nuestro caso de estudio, esta aparente contradicción entre definición matemática y entidad biológica en realidad no es tal: aunque el concepto de pattern no sea suficiente para hallar todos los miembros de la familia de parálogos, sí se cumple que todas las ocurrencias de estos patterns sean parálogas entre sí. Si pensamos en los patterns en general, más allá de nuestro caso de estudio, aparece un punto de conflicto: ¿cuán largo debe ser un pattern para considerar que sus ocurrencias son secuencias parálogas? Volviendo al ejemplo extremo del pattern *AT*, de longitud 2, la hipótesis de que todas las secuencias *AT* del genoma tienen un origen común no podría sostenerse. Operativamente podrían hallarse soluciones sencillas basadas en cálculos de probabilidades y adopción de líneas de corte, pero eso no quita que exista al menos un sentido en el que la definición de pattern entre en conflicto con las entidades biológicas conocidas.

El aspecto potencialmente más conflictivo de la aplicación de la definición matemática de pattern a los genomas es su relación con el concepto biológico de homología. Para la Biología Evolutiva el estatus de homología implica la existencia de ancestralidad común, y es aplicable a cualquier característica observable en los organismos vivos, desde estructuras

morfológicas o distribuciones geográficas hasta nucleótidos individuales en el genoma [36]. Resulta entonces necesario preguntarse por el comportamiento de los patterns en términos de ancestralidad. Piénsese en una situación como la que se esquematiza en la Figura 15. Un genoma ancestral sufre la duplicación de una secuencia, originándose un par de duplicaciones. Si hiciéramos el cómputo de patterns sobre la secuencia de este genoma ancestral inmediatamente después de la duplicación, hallaríamos al pattern *A*, con dos ocurrencias. Si luego de un tiempo en el que las duplicaciones divergen una respecto de la otra (fijando mutaciones independientes) se volviera a realizar el cómputo, se hallarían los patterns *B* y *C*, ambos con una ocurrencia en cada duplicación. *B* y *C* estarían ausentes en la secuencia ancestral con lo cual deberían ser mapeados como adquisiciones independientes, a pesar de estar contenidos en secuencias que sí son homólogas. En consecuencia, los patterns no se comportan como entidades biológicas si se los trata como caracteres discretos cuyos 2 estados sean "presencia del pattern" y "ausencia del pattern".

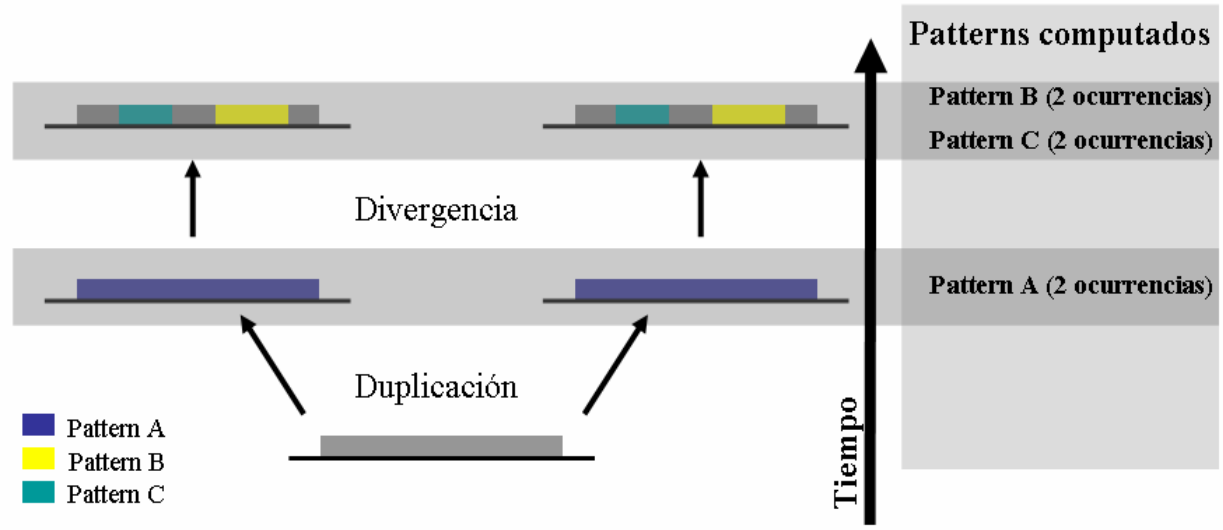

Figura 15. Esquema de la evolución de un conjunto de patterns en un genoma. Un bloque genómico (rectángulo gris) sufre un evento de duplicación generando dos copias inicialmente idénticas. Si se buscaran patterns en el genoma en ese momento se obtendría *A*, cuyas 2 ocurrencias coinciden exactamente el bloque duplicado. Si se repitiera la búsqueda de patterns luego de que las duplicaciones diverjan se hallarían *B* y *C*, ambos con una ocurrencia en cada duplicación. Como *B* y *C* están ausentes en la secuencia ancestral, deberían ser mapeados como adquisiciones independientes a pesar de pertenecer a secuencias con un origen común.

El resultado de la aplicación de la definición matemática de pattern no siempre coincide con el resultado de los procesos biológicos que actúan sobre el genoma. Esto no le quita valor al pattern como herramienta para el estudio de los genomas, ya que resulta útil para el cómputo exhaustivo de secuencias potencialmente homólogas y altamente

conservadas. Los inconvenientes surgirán en la medida en que se saquen conclusiones con implicancias evolutivas a partir de la comparación de parámetros que no contemplen la historia evolutiva de los patterns (tales como cantidad, diversidad, longitud promedio, etc.) sin considerar el impacto de la posible falta de ancestralidad común. Para conocer la naturaleza y magnitud de dicho impacto hacen falta estudios detallados sobre el comportamiento de los patterns, desde las dos perspectivas posibles: experimentalmente (mediante el análisis de la evolución de los patterns a lo largo de la filogenia) y teóricamente (incorporando parámetros que reflejen procesos biológicos –como la especiación, duplicación y divergencia– al desarrollo formal de las teorías de patterns).

4.4- Duplicaciones segmentales: consideraciones ontológicas y metodológicas

El concepto de DS es uno de los que surgen como consecuencia de la observación de regiones repetidas en la secuencia obtenida por el Proyecto Genoma Humano [8]. El estudio de las DS mediante herramientas bioinformáticas generó la necesidad de una definición tan precisa como arbitraria: longitud mayor a 1000 posiciones y similitud mayor o igual al 90% [22]. La gran ventaja de esta definición operacional fue la posibilidad de desarrollar metodologías capaces de generar las bases de datos de DS disponibles actualmente [1, 2, 22]. La desventaja tiene que ver con lo artificial que puede resultar el agrupamiento de los distintos elementos genómicos susceptibles de caer dentro de la categoría de DS. Un análisis ontológico riguroso de las DS excedería por mucho los objetivos de este trabajo. Desde nuestra posición personal, los grupos de elementos biológicos vinculados entre sí mediante procesos que impliquen ancestralidad común gozan de un estatus ontológico superior al de los que agrupen elementos por características compartidas. Las bases de datos de DS pertenecerían al segundo tipo de grupos, ya que solo se basan en los aspectos fenomenológicos provistos por la definición, dejando de lado los mecanísticos. Esto no implica que todas las DS estén agrupadas arbitrariamente (de hecho, hemos dedicado parte del

Capítulo 2 a validar un conjunto de DS como verdaderas duplicaciones), pero la categoría así definida no permite decir a qué procesos deben su existencia los elementos que contiene.

Desde el punto de vista metodológico, el formato de las bases de datos de DS presenta un obstáculo extra para la dilucidación de los mecanismos biológicos subyacentes. Todas las DS aparecen reportadas de a pares, aún en los casos en que se trate de una familia numerosa (Material Suplementario 8, Figura S3). Esto implica que una misma región duplicada puede estar solapada con más de una de las DS reportadas en la base de datos (Figura 11). El hecho de que las bases de datos tengan estas características es consecuencia de la naturaleza de las metodologías de detección actuales, que se basan en alineamientos pareados de poblaciones de porciones genómicas, seguidos de la aplicación de líneas de corte definidas estadísticamente [1, 2]. Otro aspecto importante de estas metodologías es que las búsquedas de DS que llevan a cabo no son exhaustivas. Esto se puso en evidencia experimentalmente en el Capítulo 3 de este trabajo, cuando se observó que existen patterns de longitud mayor a 1000 que a pesar de ser DS por definición no se solapan con ninguna DS reportada en la base de datos (Figura 14c).

Desde nuestro punto de vista, las bases de datos de DS están a la espera de mejoras en dos aspectos complementarios: su enriquecimiento cuantitativo (por perfeccionamiento de las metodologías de detección) y el avance hacia una definición de DS cada vez menos arbitraria.

4.5- Detección de duplicaciones segmentales: una propuesta metodológica alternativa

En el Capítulo 2 se logró identificar una familia de 6 DS a partir de un conjunto de patterns. Inspirándonos en esos resultados hemos concebido una nueva metodología para la detección de DS en cualquier genoma ensamblado. La estructura de esta metodología (que en

esencia es similar a lo descrito en la primera parte del apartado 2.4.2) puede sintetizarse de la siguiente manera:

1- Cómputo exhaustivo de patterns de un rango de longitud dado en el genoma de interés (mediante el algoritmo Findpat del grupo Kapow)
2- Asignación de un número a cada pattern hallado
3- Generación de una base de datos que liste todas las ocurrencias identificadas con el número correspondiente a su pattern y su posición en el genoma
4- Ordenamiento de todas las ocurrencias de los patterns por posición en el genoma
5- Cómputo exhaustivo de patterns de cualquier longitud en la secuencia de números de pattern generada en el ordenamiento del paso anterior (nótese que los patterns hallados en esta segunda instancia de cómputo no serán secuencias nucleotídicas sino numéricas)

Cada uno de los patterns numéricos hallados en el último paso denotará un conjunto de patterns de secuencia nucleotídica que ocurren siempre asociados, de manera no independiente. La ubicación de las ocurrencias de estos conjuntos de patterns no independientes corresponderá a una región candidata a contener una DS. Las DS candidatas deberán ser luego validadas. El único parámetro a fijar por el usuario sería la longitud de los patterns computados en el paso 1.

A partir de este esquema básico podrían desarrollarse 2 modificaciones que mejorarían el poder de detección del método:

i) Agregar una instancia de búsqueda exhaustiva de parálogos no idénticos de los patterns mediante alineamientos antes del paso 2, y tratarlos como ocurrencias de los mismos. Esto permitiría detectar DS que de otra forma quedarían ocultas por contener parálogos no idénticos de los patterns.

ii) Reemplazar el paso 5 por un cómputo exhaustivo de patrones no exactos. Esto permitiría considerar a conjuntos de patterns no independientes que, sin dejar de formar parte de DS, no fueran absolutamente dependientes entre sí.

La aplicación de esta metodología a escala genómica requeriría de un trabajo de programación que excede los objetivos de este proyecto. Resulta difícil predecir la eficiencia computacional que tendría un algoritmo inspirado en esta propuesta. Sin embargo, podemos anticipar algunas ventajas y desventajas respecto a las metodologías existentes para la detección de DS.

Una primera ventaja es el carácter exhaustivo de las búsquedas de patterns mediante el algoritmo Findpat. La contraparte de esta ventaja es el hecho de que la búsqueda de patterns solo halla secuencias idénticas, con lo cual si no se implementa la modificación i) nuestra metodología perdería poder de detección al aumentar la divergencia entre las DS. Sin embargo, ese inconveniente podría solucionarse reduciendo la longitud de los patterns computados en el paso 1.

La ventaja más importante tiene que ver con la forma de detección de las familias de duplicaciones. Si una secuencia ancestral se duplica secuencialmente generando 3 o más secuencias hermanas, cuando el algoritmo Findpat busque los patterns contenidos en ellas hallará una ocurrencia por cada una, y para cada pattern. En el archivo de salida figurarán los patterns con la ubicación de todas sus ocurrencias, cada una de las cuales corresponderá a uno de los miembros de la familia de duplicaciones. Esto permitiría que las DS no se hallen únicamente de a pares, como mediante las metodologías actuales, sino respetando un agrupamiento consistente con el patrón de ancestralidad común.

La principal desventaja radicaría en el hecho de que nuestra metodología requiere que el genoma esté completamente ensamblado, y además será sensible a la bondad del mismo. Una de las metodologías existentes no tiene este requisito ya que utiliza los fragmentos genómicos producidos durante la etapa de secuenciación de los proyectos genoma, antes del ensamblado final. Este punto es crítico teniendo en cuenta que buena parte de los genomas no están ensamblados.

Si tomamos a las DS halladas en el Capítulo 2 como un simulacro de la aplicación de la metodología que proponemos y las comparamos con los resultados de las metodologías existentes, podemos ver que existe cierta complementariedad entre ambas. Algunas de ellas coinciden exactamente, mientras que en las restantes se verifica que una u otra metodología falla al determinar su extensión.

4.6- Otras posibles aplicaciones de la asociación entre patterns y duplicaciones segmentales

En el apartado 4.2 se postuló la ausencia de función biológica de los patterns como uno de los escenarios posibles para explicar su permanencia en el genoma, luego de su origen por duplicación de la secuencia más larga que los contiene. La sencillez de este escenario permitiría su utilización como modelo nulo para la puesta a prueba de una hipótesis de neutralidad selectiva. Piénsese en un par de duplicaciones de largo l con un valor m de divergencia entre ellas (estimable a partir de su porcentaje de similitud), cada una de las cuales contiene una ocurrencia de un pattern de largo n. En principio se podría calcular la probabilidad de hallar patterns de largo n en una secuencia de largo l que acumula una cantidad $m/2$ de mutaciones distribuidas al azar. Si se conociera la distribución de esta probabilidad condicional podría diseñarse una prueba estadística que permita decidir si la longitud de los patterns hallados es significativamente mayor a la esperada por azar, en cuyo caso podría pensarse en un escenario alternativo a la neutralidad.

A pesar de que el genoma humano se encuentre ensamblado, existen aún regiones de posiciones indefinidas. La detección de regiones duplicadas podría servir como herramienta para la elucidación de las secuencias faltantes. Considérese un caso similar al que se ve en la Figura 6, en el cual existen algunas DS cuyo límite está determinado por la aparición de regiones de posiciones indefinidas. El hecho de que otras de las DS altamente similares se extiendan más allá de dicho límite podría utilizarse para predecir en cierta medida la

secuencia de las regiones faltantes (en el ejemplo de la Figura, las DS 2, 4 y 5 podrían aportar información sobre la secuencia río debajo de la DS 1).

CONSIDERACIONES FINALES

En este trabajo se atacó el problema de la interpretación biológica de los patterns del genoma humano mediante dos estrategias metodológicas diferentes, basadas en el análisis detallado de un conjunto de patterns y en la superposición de bases de datos completas, respectivamente. Estas estrategias resultaron complementarias, ya que la primera permitió validar los resultados desde un marco evolutivo –sacrificando el poder de extrapolación de las conclusiones– mientras que la segunda permitió el estudio del fenómeno a escala genómica, a costas de estar sujeto a una definición arbitraria de DS.

Se logró poner a prueba un modelo capaz de explicar el origen de parte de los patterns en el genoma, y se estimó su poder explicativo. A partir de los escenarios y modelos alternativos planteados, estamos en condiciones de especular sobre la naturaleza de los patterns que, según nuestros resultados, no serían explicados por el modelo que proponemos (Figura 14b). Una alternativa es preguntarse si caen dentro de alguna otra categoría genómica ya descripta (un buen candidato sería la que comprende a los elementos transponibles). Preguntas de este tipo podrían responderse mediante cálculos de cobertura como los realizados en el Capítulo 3. Hay que tener en cuenta que también puede tratarse de patterns contenidos en DS aún no reportadas como tales debido a la imperfección de los métodos de detección existentes. Otra alternativa sería tomar los patterns no contenidos en DS como candidatos a ser explicados por modelos de origen alternativos al propuesto en este trabajo (*i.e.* duplicación independiente de grandes bloques genómicos) o vestigios ultra conservados de DS altamente divergentes.

Como primera aproximación del grupo Kapow y el Laboratorio de Evolución al problema de la interpretación biológica de patterns, este trabajo se acotó al estudio de su

origen. Queda pendiente el desarrollo de modelos capaces de explicar la permanencia de los mismos a lo largo del tiempo y, como objetivo último, la dilucidación de sus posibles funciones biológicas.

Por fuera de esta línea de investigación, destacamos la necesidad de que el desarrollo de los marcos teóricos de patterns y DS ocurra con miras a lograr una relación más cercana y clara con los mecanismos biológicos conocidos.

Por último, nos interesa destacar la importancia que tuvo el carácter interdisciplinario de nuestro trabajo en el desarrollo de esta tesina. Los aspectos más ricos de la investigación surgieron del aporte conjunto de las dos disciplinas, que no habría sido posible sin la colaboración de miembros de los dos grupos de investigación involucrados. Desde nuestra experiencia percibimos este hecho como una prueba más del potencial que ofrece el trabajo interdisciplinario y entre grupos independientes, en el ámbito científico en general y en particular dentro de nuestra Facultad.

5- REFERENCIAS BIBLIOGRÁFICAS

1. Bailey J. A., *et al*. 2001. Segmental Duplications: organization and impact within the current human genome project assembly. *Genome Research* 11: 1005-1017
2. Bailey J. A., *et al*. 2002. Recent segmental duplications in the human genome. *Science* 297: 1003-1007
3. Becher V., *et al*. 2009. Efficient computation of all perfect repeats in genomic sequences of up to half a Gigabyte, with a case study on the Human genome. *Bioinformatics* 25: 1746-1753
4. Blekhman R., *et al*. 2009. Segmental duplications contribute to gene expression differences between humans and chimpanzees. *Genetics* 182: 627–630
5. Britten R. J. y Kohne D. E. 1968. Repeated sequences in DNA. *Science* 161: 529-40
6. Cheng Z., *et al*. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437: 88-93
7. Citarelli, M, *et al*. 2010. Evolutionary history of the poly(ADP-ribose)polymerase gene family in eukaryotes. *BMC Evolutionary Biology*, 10:308
8. Collins F. S., *et al*. 1998. New Goals for the U.S. Human Genome Project: 1998-2003. *Science* 282: 682-689
9. Eichler E. E. 1998. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Research* 8: 758-762
10. Eichler E. E., *et al*. 1996. Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Human Molecular Genetics* 7(5): 899–912
11. Eichler E. E., *et al*. 2001. Divergent origins and concerted expansion of two segmental duplications on chromosome 16. *The American Genetic Association* 92:462–468
12. Enrad W., *et al*. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 190-193
13. Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.

14. International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018):695-716
15. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921
16. Ji, Y. *et al*. 2000. Structure of chromosomal duplications and their role in mediating human genomic disorders. *Genome Research* 10: 597-610
17. Kent W.J., *et al*. 2002. The human genome browser at UCSC. *Genome Research* 12(6):996-1006
18. King M. C. y Wilson A. C. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–106.
19. Lindblad-Toh K., *et al*. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438: 803-819.
20. Locke D.P., *et al*. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331):529-33
21. Marques-Bonet T., *et al*. 2009. A burst of segmental duplications in the african great ape ancestor. *Nature* 457: 877–881
22. Marques-Bonet T., *et al*. 2009. The origins and impact of primate segmental duplications. *Trends in Genetics* 25: 443–454
23. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520-62
24. Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, Nueva York
25. Ohno S. 1972. So much "junk" DNA in our genome. Evolution of Genetic Systems. H. H. Smith. ed. Gordon and Breach, New York. pp. 366–370
26. Orgel L. E. y Crick F. H. 1980. Selfish DNA: the ultimate parasite. *Nature* 284: 604–607
27. Posada D. y Crandall K.A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14 (9): 817-818
28. Rhesus Macaque Genome Sequencing and Analysis Consortium, *et al*. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822):222-234

29. Roberto R., *et al*. 2007. Molecular refinement of gibbon genome rearrangements. Genome *Research* 17(2):249-257
30. Sharp A. J., *et al*. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics* 38:1038-1042
31. She, X., *et al*. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 431: 927-930
32. Stefan Kirsch S., *et al*. 2005. Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Research* 15: 195-204
33. Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) version 4b10. Sinauer Associates, Sunderland, Massachusetts.
34. The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69-87
35. Trask B. J., *et al*. 1998. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Human Molecular Genetics* 13(7): 2007-2020
36. Wagner G. P. 2007. The Developmental genetics of homology. *Nature Genetics* 8:473-479
37. Zhang L., *et al*. 2005. Patterns of Segmental Duplication in the Human Genome. *Molecular Biology and Evolution* 22(1):135–141

6- MATERIAL SUPLEMENTARIO

Material Suplementario 1: análisis de dependencia de familias a todo el genoma

Tabla S1. Todos los miembros de las 24 familias en todo el genoma (excepto los cromosomas 1 y 5), ordenados por posición en cada cromosoma. En magenta se indican las ocurrencias de P2, que es el único PP observado fuera de los cromosomas 1 y 5. Se resaltan los elementos cuyas longitudes respecto a las del resto de los miembros de su familia generaron conflicto en la definición de los PP, y se indica su longitud.

Cromosoma	n° de familia	Miembro de la familia	Posición de inicio en el cromosoma	Posición final en el cromosoma	Patrón formado	Miembros más cortos
2	10	f10-2-1	114345183	114346085		
	14	f14-2-2	243070986	243071753	P2-2-1-R	
	23	f23-2-1	243071187	243071886		
	5	f5-2-1	243070849	243071893		
	16	f16-2-1	243071195	243071959		
	3	f3-2-1	243096753	243097897		
	3	f3-2-3-R	243180359	243181503		
3	20	f20-3-1	197917748	197918443		
	14	f14-3-1	197961520	197962290	P2-3-1-R	
	23	f23-3-1	197961721	197962423		
	5	f5-3-1	197961383	197962430		
	16	**f16-3-1**	**197961729**	**197962430**		**701**
4	3	f3-4-1	119340655	119341800		
	6	f6-4-1	119348474	119349500		
	20	f20-4-1	119521642	119522336		
	14	f14-4-1	119568274	119569041	P2-4-1-R	
	23	f23-4-1	119568475	119569174		
	5	f5-4-1	119568142	119569181		
	16	f16-4-1	119568483	119569247		
	20	f20-4-2-R	120360504	120361196		
	16	**f16-4-2-R**	**165196422**	**165197125**	P2-4-2	**703**
	5	f5-4-2-R	165196422	165197469		
	23	f23-4-2-R	165196426	165197133		
	14	f14-4-2-R	165196563	165197329		
6	11	f11-6-1	113884	114769		
	16	f16-6-1	132932	133699	P2-6-1	
	5	f5-6-1	132998	134045		
	23	f23-6-1	133005	133707		
	14	f14-6-1	133138	133908		
	5	**f5-6-2-R**	**170921793**	**170922495**	P2-6-2-R	**702**
	14	**f14-6-2-R**	**170921793**	**170922362**		**569**
	23	f23-6-2R	170921793	170922495		
	16	f16-6-2-R	170921801	170922568		
	11	f11-6-2-R	170941020	170941008		
	3	f3-6-1	170996115	170997257		
	6	f6-6-1	171003928	171004954		
7	20	f20-7-1	39801904	39802597		
	20	f20-7-2	45823283	45823938		
	20	f20-7-3-R	56475182	56475877		
	20	f20-7-4-R	65942811	65943506		
	20	f20-7-5	66329153	66329839		
	20	f20-7-6-R	75783147	75783838		
8	6	f6-8-1	62316	63342		
	3	f3-8-1	70016	71158		
	11	f11-8-1	124055	124951		
	16	f16-8-1	143114	143880	P2-8-1	
	5	f5-8-1	143180	144226		
	23	f23-8-1	143187	143888		
	14	f14-8-1	143320	144089		

(Continúa en la siguiente hoja)

Cromosoma	n° de familia	Miembro de la familia	Posición de inicio en el cromosoma	Posición final en el cromosoma	Patrón formado
(Continuación)					
9	10	f10-9-1	25046	25948	
	17	f17-9-1	138944776	138945537	
	20	f-20-9-1	141144171	141144866	
10	20	f20-10-1	38706629	38707308	
	14	f14-10-1	38753082	38753852	P2-10-1-R
	23	f23-10-1	38753283	38753985	
	5	f5-10-1	38752945	38753992	
	16	f16-10-1	38753291	38754058	
11	11	f11-11-1	94934	95819	
	16	f16-11-1	113991	114758	P2-11-1
	5	f5-11-1	114057	115100	
	23	f23-11-1	114064	114766	
	14	f14-11-1	114197	114963	
	20	f20-11-1	158270	158965	
12	10	f10-12-1	77842	78744	
15	11	f11-15-1	102408491	102408479	
	10	f10-15-1	102505328	102506222	
16	20	f20-16-1	90205128	90205823	
	14	f14-16-1	90250597	90251367	P2-16-1-R
	23	f23-16-1	90250798	90251500	
	5	f5-16-1	90250460	90251507	
	16	f16-16-1	90250806	90251573	
	3	f3-16-1	90276355	90277500	
19	10	f10-19-1	66541	67443	
	11	f11-19-1	164557	165442	
	16	f16-19-1	183925	184692	P2-19-1
	5	f5-19-1	183991	185038	
	23	f23-19-1	183998	184700	
	14	f14-19-1	184131	184901	
	20	f20-19-1	228437	229132	
20	14	f14-20-1	62932716	62933486	P2-20-1-R
	23	f23-20-1	62932917	62933619	
	5	f5-20-1	62932579	62933626	
	16	f16-20-1	62932925	62933692	
Y	14	f14-Y-2-R	26435699	26436468	P2-1-R
	23	f23-Y-1	26435900	26436601	
	5	f5-Y-1	26435562	26436608	
	16	f16-Y-2-R	26435908	26436675	
	3	f3-Y-1	26467450	26468595	
	6	f6-Y-1	26475258	26476280	
	6	f6-Y-2-R	27486164	27487186	
	3	f3-Y-2-R	27493849	27494994	
	16	f16-Y-1	27525765	27526532	P2-Y-2
	5	f5-Y-2-R	27525832	27526878	
	23	f23-Y-2R	27525839	27526540	
	14	f14-Y-1	27525972	27526741	

Material Suplementario 2: análisis estructural de P2

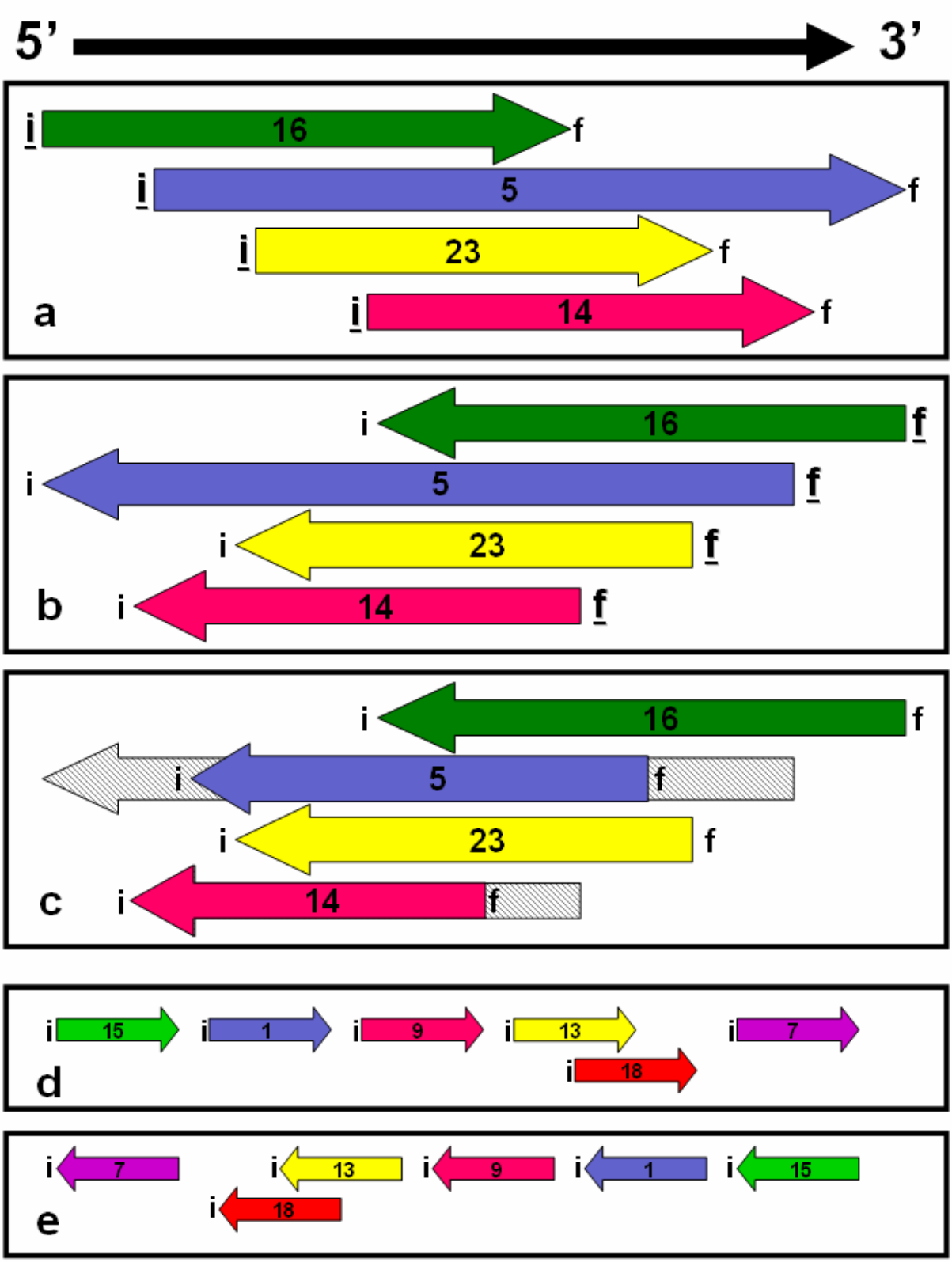

Figura S1. Efectos de la disposición de las familias en el ordenamiento de las mismas por posición inicial y final. A cada familia le corresponde un color. El sentido de las flechas indica la orientación del elemento. *i*: posición inicial. *f*: posición final. Se muestra la disposición de las familias en: una ocurrencia no reversa P2 (*a*), una ocurrencia reversa de P2 (*b*), P2-6-2-R (*c*), una ocurrencia no reversa de P1 (*d*) y una ocurrencia reversa de P1 (*e*). En el caso de P2, el ordenamiento por posición inicial difiere del ordenamiento por posición final (*a* y *b*). Esto no ocurre en el caso de P1 (*d* y *e*).

Tabla S2. Comparación a nivel de secuencia de las 21 ocurrencias de P2. Los valores que se informan son el porcentaje de similitud entre las secuencias de las ocurrencias correspondientes, calculados a partir del alineamiento global. En el cálculo se tuvo en cuenta la cantidad de gaps. Los valores máximo y mínimo se muestran resaltados.

	P2-1-1	P2-1-2-R	P2-1-3	P2-1-4	P2-1-5-R	P2-1-6	P2-2-1-R	P2-3-1-R	P2-4-1-R	P2-4-2	P2-5-1-R	P2-6-1	P2-6-2-R	P2-8-1	P2-10-1-R	P2-11-1	P2-16-1-R	P2-19-1	P2-20-1-R	P2-Y-1-R	P2-Y-2
P2-1-1		0,998	0,996	0,987	0,985	0,992	0,988	0,995	0,99	0,966	0,998	0,995	0,995	0,993	0,995	0,994	0,996	0,998	0,995	0,978	0,978
P2-1-2-R			0,997	0,988	0,986	0,993	0,989	0,997	0,991	0,967	**1**	0,996	0,997	0,994	0,996	0,995	0,997	**1**	0,996	0,979	0,979
P2-1-3				0,986	0,984	0,991	0,987	0,994	0,989	0,964	0,997	0,994	0,994	0,992	0,994	0,993	0,995	0,997	0,994	0,977	0,977
P2-1-4					0,981	0,986	0,982	0,988	0,984	0,962	0,988	0,985	0,988	0,987	0,989	0,992	0,986	0,988	0,989	0,972	0,972
P2-1-5-R						0,984	0,98	0,983	0,982	0,959	0,986	0,983	0,983	0,985	0,987	0,982	0,984	0,986	0,987	0,972	0,972
P2-1-6							0,987	0,991	0,989	0,964	0,993	0,99	0,991	0,992	0,994	0,989	0,991	0,993	0,994	0,977	0,977
P2-2-1-R								0,985	0,991	0,959	0,989	0,986	0,985	0,988	0,99	0,985	0,987	0,989	0,99	0,973	0,973
P2-3-1-R									0,987	0,965	0,997	0,994	0,994	0,991	0,994	0,997	0,994	0,997	0,994	0,977	0,977
P2-4-1-R										0,96	0,991	0,988	0,987	0,99	0,992	0,987	0,989	0,991	0,992	0,977	0,977
P2-4-2											0,967	0,964	0,964	0,964	0,967	0,967	0,964	0,967	0,967	**0,953**	**0,953**
P2-5-1-R												0,996	0,997	0,994	0,996	0,995	0,997	**1**	0,996	0,979	0,979
P2-6-1													0,994	0,991	0,993	0,992	0,994	0,996	0,993	0,976	0,976
P2-6-2-R														0,991	0,994	0,997	0,994	0,997	0,994	0,977	0,977
P2-8-1															0,995	0,99	0,992	0,994	0,995	0,979	0,979
P2-10-1-R																0,992	0,994	0,996	0,997	0,98	0,98
P2-11-1																	0,993	0,995	0,992	0,975	0,975
P2-16-1-R																		0,997	0,994	0,977	0,977
P2-19-1																			0,996	0,979	0,979
P2-20-1-R																				0,98	0,98
P2-Y-1-R																					**1**
P2-Y-2																					

Material Suplementario 3: análisis estructural de P3

Tabla S3. Comparación a nivel de secuencia de las 7 ocurrencias de P3. Los valores que se informan son el porcentaje de similitud entre las secuencias de las ocurrencias correspondientes, calculados a partir del alineamiento global. En el cálculo se tuvo en cuenta la cantidad de gaps.

	P3-1-1	P3-1-2	P3-1-3	P3-1-4-R	P3-1-5-R	P3-1-6-R	P3-5-1-R
P3-1-1		0,997	0,991	0,989	0,989	0,996	0,944
P3-1-2			0,992	0,99	0,989	0,998	0,944
P3-1-3				0,995	0,995	0,991	0,948
P3-1-4-R					0,998	0,989	0,948
P3-1-5-R						0,988	0,947
P3-1-6-R							0,944
P3-5-1-R							

Material Suplementario 4: análisis estructural de P4

Tabla S4. Comparación a nivel de secuencia de las posiciones ubicadas entre los 17 miembros de f24. Los valores que se informan son el porcentaje de similitud entre las secuencias de las ocurrencias correspondientes, calculados a partir del alineamiento global. En el cálculo se tuvo en cuenta la cantidad de gaps. Los valores máximo y mínimo se muestran resaltados.

	1--2	2--3	3--4	4--5	5--6	6--7	7--8	8--9	9--10	10--11	11--12	12--13	13--14	14--15	15--16	16--17
1--2		**0,999**	**0,999**	0,996	0,991	0,982	0,998	0,995	0,998	0,998	0,996	0,985	0,997	0,988	0,997	0,989
2--3			0,998	0,997	0,99	0,983	0,998	0,996	0,997	**0,999**	0,996	0,986	0,998	0,989	0,996	0,99
3--4				0,996	0,991	0,981	**0,999**	0,994	0,998	0,998	0,996	0,986	0,998	0,989	0,998	0,99
4--5					0,988	0,981	0,996	0,997	0,994	0,996	0,995	0,985	0,995	0,987	0,994	0,987
5--6						0,977	0,99	0,987	0,99	0,99	0,988	0,992	0,989	0,996	0,989	0,987
6--7							0,982	0,983	0,981	0,983	0,981	**0,974**	0,982	0,977	0,979	0,977
7--8								0,994	0,997	0,997	0,996	0,985	0,997	0,988	0,997	0,989
8--9									0,993	0,996	0,994	0,983	0,994	0,985	0,992	0,986
9--10										0,996	0,994	0,984	0,997	0,988	0,996	0,988
10--11											0,996	0,985	0,997	0,988	0,996	0,989
11--12												0,988	0,996	0,988	0,994	0,988
12--13													0,985	0,991	0,984	0,992
13--14														0,99	0,996	0,989
14--15															0,987	0,996
15--16																0,99
16--17																

Material Suplementario 5: determinación de los límites de la DS que contiene a las ocurrencias de P1

Tabla S5. Cantidad de posiciones flanqueantes extraídas para cada ocurrencia, tanto hacia 5' como hacia 3' para cada una de las ocurrencias de P1

	Posiciones extraídas	
	Hacia 5'	Hacia 3'
P1-1-1	82816	27238
P1-1-2	49185	61656
P1-1-3	52173	102076
P1-1-4	200 kb	94752
P1-1-5-R	200 kb	67493
P1-5-1-R	200 kb	200 kb

Material Suplementario 6: reconstrucción de la historia evolutiva de las DS que contienen a las ocurrencias de P1

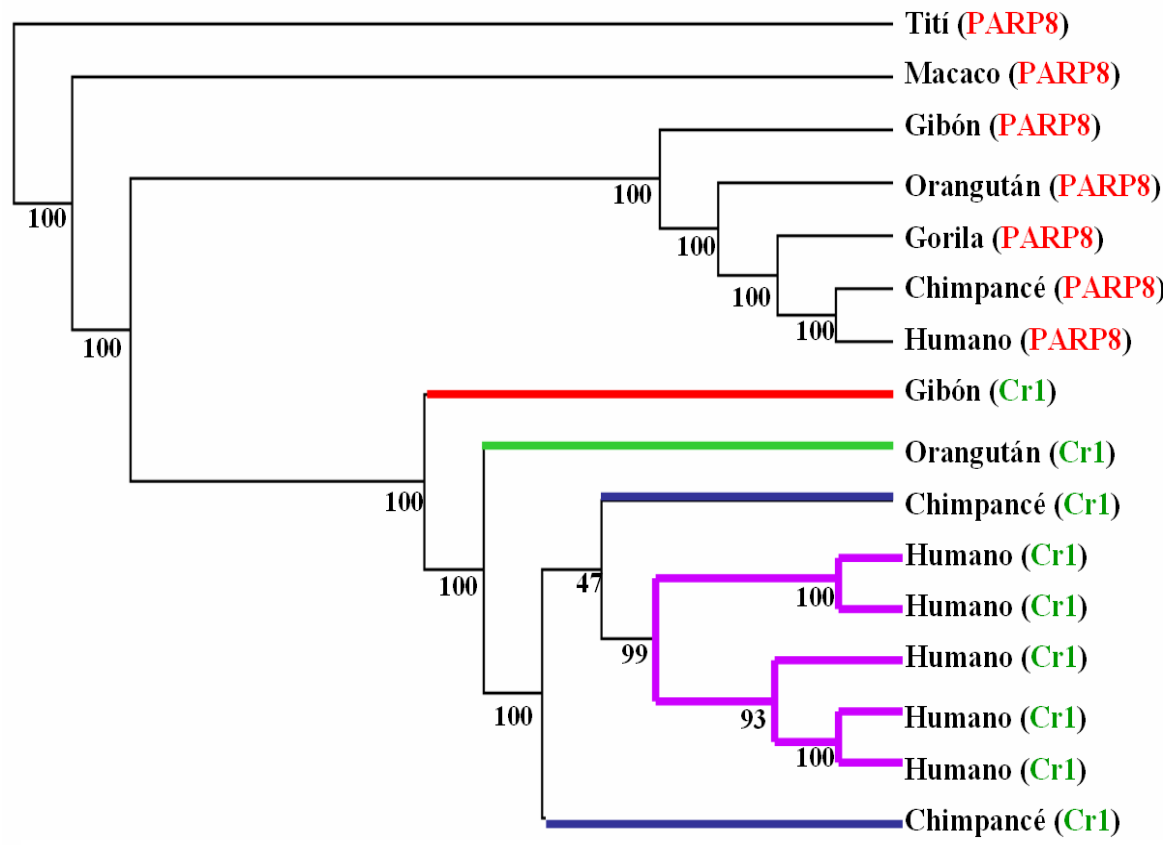

Figura S2. Reconstrucción de la filogenia de los homólogos de P1 del Infraorden Catarrhini mediante el método de Máxima Parsimonia. Se utilizó el homólogo del genoma de Titi como grupo externo. *PARP8*: homólogo asociado al gen PARP8, en el cromosoma sinénico al 5 humano. *Cr1*: homólogo ubicado en el cromosoma sinténico al 1 humano. Las líneas de color indican las ramas correspondientes a cada especie. Los números junto a los nodos indican el valor de soporte sobre 10000 réplicas mediante bootstrap.

Material Suplementario 7: determinación de la estructura de las DS que contienen a las ocurrencias de P1

Tabla S6. Algunos de los elementos reportados dentro de cada una de las 6 DS definidas a partir de las 24 familias. Se informa el nombre con el que se reporta cada elemento y la categoría a la que pertenece.

n° de DS	Ocurrencia de P1	Elementos reportados	Categoría
1	P1-1-1	PPIAL4G	Codificante
		BX284650.1	RNA no codificante
2	P1-1-2	PPIAL4B	Codificante
		BX248398.1	RNA no codificante
		AL592284.1	Codificante
3	P1-1-3	Al592284.1	RNA procesado
		AL592284.1	Codificante
4	P1-1-4	PPIAL4A	Codificante
		RP11495p10,2	RNA no codificante
5	P1-1-5-R	PPIAL4C	Codificante
		RP11-277L2.2	RNA no codificante
6	P1-5-1-R	PARP8	Codificante

Material Suplementario 8: puesta a prueba de la hipótesis a partir de bases de datos completas

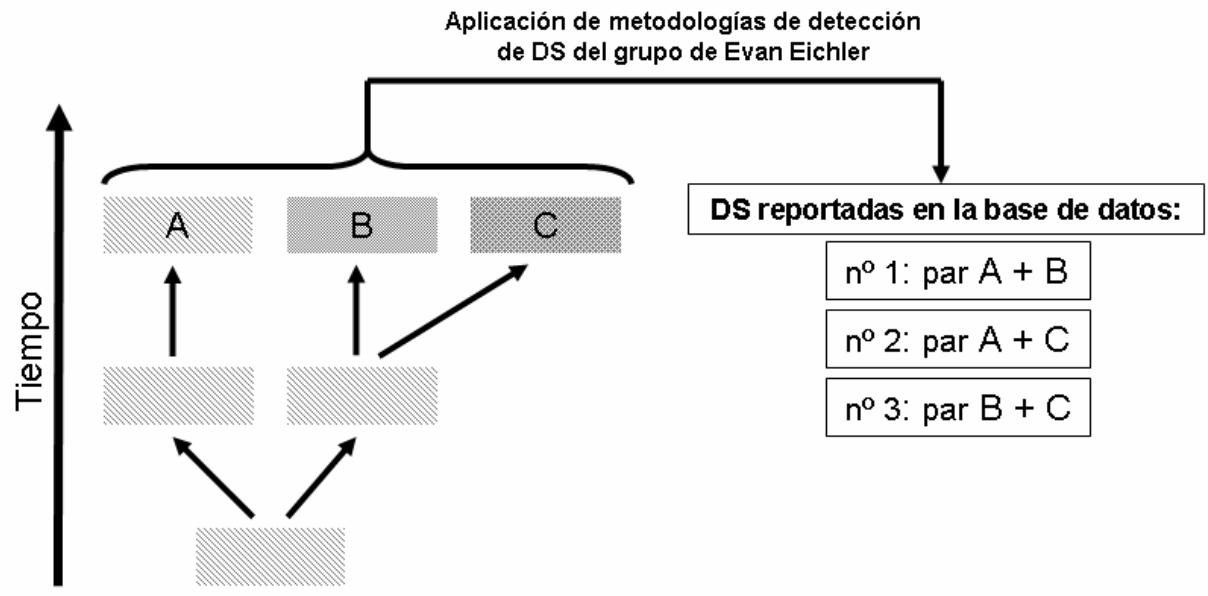

Figura S3. Relación entre el parentesco de un grupo de duplicaciones y la forma en que son reportadas por en la base de datos de DS. Se esquematiza la historia evolutiva de 3 duplicaciones emparentadas, *A*, *B* y *C*, muy similares entre sí. Las metodologías de detección de DS actuales reconocen esa similitud mediante comparaciones de a pares, pero al no asociar a las 3 duplicaciones no dan información sobre su parentesco.