Relación entre repeticiones de mononucleótidos y categorías de la Ontología de Genes en el Genoma Humano

Relationship between mononucleotide repeats and categories of Gene Ontology in Human Genome

María de los Angeles Zardón Navarro,[I] Miguel Sautié Castellanos,[II] Ivette Camayd Viera,[III] Yulemi González Quesada,[IV] José Luis Hernández Cáceres[V]

Resumen

Los microsatélites son repeticiones en bloques de motivos cortos que abarcan alrededor del 3% del genoma humano. En la actualidad se conocen más de 40 enfermedades neurológicas, neurodegenerativas y musculares, entre otras, asociadas a la inestabilidad mutacional de estas secuencias. La Ontología de Genes puede ser una herramienta muy útil para el estudio del papel funcional de las secuencias repetidas y para profundizar en la comprensión de la etiología molecular de las enfermedades vinculadas a este tipo de secuencia. En este estudio se exploran las distribuciones de frecuencias de repetidos de mononucleótidos en las regiones codificadoras y no codificadoras para la mayor parte de los genes del Genoma Humano. Además se determinan los componentes celulares, las funciones moleculares y los procesos biológicos que aparecen con mayor frecuencia en genes con repetidos de gran tamaño. En general, fueron estadísticamente significativas las categorías de la Ontología de Genes relacionadas con los ácidos nucleicos, fundamentalmente con la transcripción, la regulación del ciclo celular, los procesos musculares, los receptores vinculados a la transducción de señales y a la sinapsis, así como con los implicados en el desarrollo del Sistema Nervioso Central. Se encontraron muy pocas combinaciones significativas de 2 y 3 categorías de Ontología de Genes y éstas estaban relacionadas principalmente con los ácidos nucleicos. Este trabajo contribuirá al establecimiento de patrones de normalidad en términos de frecuencia de repetidos asociados a categorías de Ontología de Genes, información valiosa para la determinación y comprensión de los umbrales de tamaño de repetidos relacionados con el riesgo a padecer determinadas enfermedades.

Palabras clave: Microsatélites, ontología de genes, mononucleótidos, regiones codificadoras o CDS, regiones no codificadoras, intrones.

Abstract

Microsatellites are short tandem repeats that represent the 3 % of the human genome. At present, more than 40 neurological, neurodegenerative, muscular and other diseases associated with the mutational instability of this kind of sequences are known. Gene Ontology may be a very useful tool for studying the functional role of these repetitive sequences and to get an insight into the molecular etiology of these diseases. In this study, the frequency distributions of mononucleotide repeats in coding and non-coding sequences for nearly all genes from the human genome were examined. All the cellular components, molecular functions and biological processes associated with the genes that have the longest repeats were investigated. As a general rule, the statistically significant gene ontology categories investigated corresponded to the nucleic acids, basically those related to the transcription, the cell cycle regulation, the muscular processes, the receptors linked to signals transduction and synapses, as well as those involved in the development of the Central Nervous System. Few significant two and three gene ontology combinations were found and they were predominantly related to nucleic acids. This work will be helpful to establish the normality distributions patterns of repeats associated with the main gene ontology terms, a valuable information for determining and understanding the size threshold of short tandem repeats associated with the risk to contract certain diseases.

Keywords: Microsatellites, gene ontology, mononucleotides, coding regions, CDS, non-coding sequences, introns.

[I] Máster en Ciencias en Bioinformática. Centro Nacional de Genética Médica. mzardon@cngen.sld.cu
[II] Máster en Ciencias en Informática Médica. Centro de Cibernética Aplicada a la Medicina. Instituto de Ciencias Médicas de La Habana.
[III] Licenciada en Bioquímica. Investigador Agregado. Centro Nacional de Genética Médica.
[IV] Máster en Ciencias en Neurociencias. Licenciada en Bioquímica. Investigador Agregado. Centro Nacional de Genética Médica.
[V] Doctor en Ciencias Biológicas. Centro de Cibernética Aplicada a la Medicina. Instituto de Ciencias Médicas de La Habana.
Nota: Los autores I y II tienen el mismo grado de participación en el artículo.

Introducción

Los microsatélites son repeticiones en bloques de motivos cortos que ocupan aproximadamente el 3% del genoma humano.[1] Estos se localizan tanto en regiones codificadoras como no codificadoras.[2] Algunos autores plantean que esta distribución no es azarosa debido a la gran cantidad de funciones en que se encuentran involucrados. Se conoce que tienen efectos directos en la organización de la cromatina, la regulación de la expresión génica,[3]la recombinación, la replicación del ADN, el ciclo celular, el sistema de reparación de errores, entre otros.[4]Los mecanismos que explican su génesis y expansión implican generalmente errores en la replicación del ADN; fundamentalmente a través del mecanismo conocido como deslizamiento de la polimerasa.[5]

Muchos han sido los reportes acerca del vínculo que hay entre mutaciones en microsatélites y ciertos tipos de cánceres humanos.[6,7]Por otro lado, se conocen más de 40 enfermedades neurológicas, neurodegenerativas y musculares asociadas a la inestabilidad mutacional de estas secuencias.[8]Asimismo, un estudio reciente demuestra que en general los microsatélites están sobre-representados en los genes relacionados con enfermedades de herencia mendeliana.[9]

Hasta el momento, la mayor parte de los estudios sobre la distribución de frecuencia de microsatélites se han realizado a nivel genómico y sólo algunos se han centrado en la distribución de estas secuencias repetitivas dentro de la compleja estructura génica. Sin embargo, estos últimos se restringen fundamentalmente a las regiones codificadoras y su carácter es eminentemente descriptivo.[2,10-12]

Por otro lado, algunos trabajos han intentado relacionar la distribución génica de los repetidos de trinucleótidos en bloques con las categorías de la Ontología de genes (GO, del inglés *Gene Ontology*),[12] pero sólo reportan valores de frecuencias y no calculan una medida de asociación estadística entre estas categorías y las distribuciones de frecuencias de repetidos. Además, estos trabajos se han centrado en el análisis de los trinucleótidos, sin extenderse a otros microsatélites.

La ontología de genes es un conjunto de términos o conceptos estandarizados y vinculados entre sí a través de una compleja trama de relaciones mereológicas o de especificación, cuya estructura se puede representar como un grafo acíclico dirigido. Está formada por tres divisiones: los procesos biológicos, las funciones moleculares y los componentes celulares.[13]Su aplicación permite determinar cuáles categorías biológicas tienden a agrupar genes que presentan en sus exones o intrones secuencias repetidas de gran tamaño.

Teniendo en cuenta los criterios anteriores, pretendemos demostrar que existe una asociación estadística entre ciertas categorías del GO y la presencia de secuencias repetitivas de mononucleótidos de gran tamaño tanto en las regiones codificadoras como en las no codificadoras. Este trabajo pretende contribuir al conocimiento del rol funcional de las secuencias repetidas y de las bases biológicas de las enfermedades de expansión. Este estudio de la distribución de los repetidos de mononucleótidos en el interior de la región génica constituye una primera aproximación al estudio de las enfermedades basadas en la expansión de repetidos en bloques localizados en intrones y exones.

Métodos

Los datos fueron extraídos de los archivos gbk del sitio (ftp://ftp.ncbi.nih.gov/genomes/human). Se estudiaron 22 de los cromosomas de *Homo sapiens*. No están incluidos los cromosomas autosómicos 9 y 22. Se seleccionó el ensamblaje alternativo de Celera con la fecha de actualización de los cromosomas del 29 de agosto del 2006.

Se desarrollaron los programas informáticos para la búsqueda y conteo de repetidos utilizando la implementación *Rouge Wave* de la Librería standard de C++, basada en una versión aprobada en marzo de 1988 por la *American National Standards Institute* (ANSI) y la *International Standards Organization* (ISO). Para el procesamiento estadístico de las salidas de estos programas, se prepararon varios *scripts* de Matlab versión 7.0.4.365 que hacen uso de numerosas funciones básicas y de otras pertenecientes sobre todo a la caja de herramientas de estadística versión 5.0.2.

Programas para la búsqueda de repetidos

Para buscar las repeticiones perfectas en bloques de tamaño 1 (mononucleótidos) se consideró al ADN como una secuencia de tipo texto $N_1N_2N_3N_4N_5\ldots N_i\ldots..N_{n-1}N_n$. No se tuvo en cuenta la segunda ley de paridad de Chargaff. Por consiguiente, se asumió la presencia de 4 unidades repetitivas diferentes para tamaño 1 (A, C, G y T). En la región codificadora se tuvieron en cuenta los repetidos de hasta 30 nucleótidos (nt) de tamaño. En cambio, en las regiones no codificadoras se analizaron los repetidos de hasta 70 nt.

Análisis de las distribuciones de las densidades nucleotídica de repetidos

La determinación de la densidad específica de repetidos ($de(x)$) se realizó mediante la ecuación 1

(Ec. 1). Donde *N* es la cantidad de repetidos de tamaño *u* conformada por unidades repetitivas de tamaño $x = 1$ dentro de la región génica de tamaño *t*. La variable indicadora $n_x(i)$ toma valores 0 si no hay repetidos de tamaño *x* o 1 si existen estos en la posición para cada región en estudio.

$$de_{x=1}(u) = \frac{\left(\sum_{i}^{t} n_x(i)\right) * u}{t} \qquad \text{Ec. 1}$$

Todos los cálculos de densidades de repetidos se realizaron en hojas de cálculo de Excel 2003 utilizando las matrices de salida de los programas.

Conteo observado con respecto al conteo esperado por bases

El conteo esperado se calculó asumiendo que la región en estudio se puede describir de acuerdo con un modelo de Bernoulli según el cual las diferentes posiciones del repetido son estadísticamente independientes y las substituciones nucleotídicas son los eventos mutacionales imperantes. Por tanto, la probabilidad de tener un repetido de tamaño *x* equivaldría a la multiplicación de las probabilidades de ocurrencia por posición de cada una de sus unidades repetitivas constituyentes (*p*) para las posiciones internas y para las dos posiciones que definen las fronteras del repetido. De ahí que el conteo esperado sea igual a la multiplicación de esta probabilidad por el tamaño de la región génica estudiada *t* (Ec. 2).

$$Ce_x = \left((1-p)^2 * (p)^u\right) * t \qquad \text{Ec. 2}$$

Dichos análisis se realizaron para cada una de las bases y de manera global utilizando hojas de cálculo de Excel, programas y gráficos implementados en Matlab 7.0.4.

Análisis estadístico de la relación entre las distribuciones de repetidos de mononucleótidos y las categorías de Ontología de genes.

La asociación entre las distribuciones de frecuencia de repetidos de diferentes tamaños por regiones génicas y las categorías del sistema de clasificación de GO se estudió a nivel de las secuencias codificadoras (CDS). Como medida de asociación se tomó la relación entre la proporción total de CDS anotados con determinado término de GO y la proporción de CDS anotados con este mismo término que presentan repetidos de tamaño mayor que un valor dado. De esta manera aquellos términos del GO representados por una cantidad significativamente grande de CDS con repetidos de gran tamaño se consideran asociados de una forma u otra al proceso de expansión de repetidos. Se introdujo la corrección de Bonferroni para comparaciones múltiples. Se estimó la probabilidad de una proporción dada de CDS anotados con determinado GO y con repetidos de tamaño mayor que determinado valor asumiendo que ésta se puede describir mediante una distribución hipergeométrica (Ec. 3) o aproximadamente como una binomial (Ec. 4).[14]Éste análisis se realizó para cada una de las tres divisiones del GO: componentes celulares, función molecular y procesos biológicos y además se llevó cabo a nivel global, por base nucleotídica y en las dos regiones básicas de la estructura génica: exones e intrones.

Ec. 3. **Función de distribución hipergeométrica**

$$Phip\left(X_{go,t} \middle/ N_t, M, C_{go}\right) = \frac{\binom{C_{go}}{X_{go,t}}\binom{M - C_{go}}{N_t - X_{go,t}}}{\binom{M}{N_t}}$$

Cuyas variables se definen de la siguiente manera:

N_t: Cantidad de CDS con secuencias repetitivas de tamaño mayor que *t*.

$X_{go,t}$: Cantidad de CDS anotados con una instancia dada de la ontología génica (*go*) sea dentro de las categorías de funciones, procesos o componentes y con repetidos de tamaño mayor que *t*.

C_{go}: Cantidad total de CDS anotados con una instancia dada de la ontología génica (*go*).

M: Cantidad total de CDS.

Bajo determinadas condiciones la distribución hipergeométrica se aproxima a la distribución binomial.

Ec. 4. Función de distribución binomial

$$Pbino\left(X_{go,t} \middle/ N_t, p_{go}\right) = \binom{N_t}{X_{go,t}} p_{go}^{X_{go,t}} \; q^{N_t - X_{go,t}}$$

$q = 1 - p_{go}$ Ec. 4

N_t: Cantidad de CDS con secuencias repetitivas de tamaño mayor que *t*.

$X_{go,t}$: Cantidad de CDS anotados con una instancia dada de la ontología génica (*go*) y con repetidos de tamaño mayor que *t*.

p_{go}: Frecuencia total de CDS anotados con una instancia dada de la ontología génica dentro de las categorías de funciones, procesos o componentes.

$$F_{go,t} = \frac{X_{go,t}}{N_t}$$

Se calculó el intervalo de confianza de $F_{go,t}$ para el 100(1-α)= 95% asumiendo que la función que describe su distribución es binomial.

Este análisis se realizó no sólo para los GO anotados independientemente sino también para las combinaciones de dos y de tres GO. En este último caso, se compararon las proporciones generales de CDS anotados con distintas combinaciones de dos y de tres GO, para los CDS que contenían repetidos de tamaño mayor que determinado valor prefijado.

Las tablas donde se muestran los resultados de estos análisis tienen una estructura común, las dos primeras columnas muestran los identificadores y nombres de los términos del GO. Los nombres sólo se refieren en las tablas 1 y 2, pues en las tablas 3 y 4 se repiten las mismas categorías del GO. Seguido de la probabilidad general (*Pt*) de las diferentes instancias del GO estimada a partir de la cantidad de regiones génicas anotadas con estas categorías. Luego viene la cantidad de regiones génicas anotadas con una instancia dada del GO y con repetidos de tamaño mayor que "*x*" (*N*(*x*)) donde *X*=7*nt* para regiones codificadoras y 20 *nt* para regiones no codificadoras.

La siguiente columna muestra la probabilidad asociada a *N*(*x*) suponiendo que esta variable aleatoria tenga una distribución aproximadamente binomial (*pb*(*x*)) o hipergeométrica (*ph*(x)). La sexta columna muestra el valor de la frecuencia observada (*n*>*x*)), calculada a partir de *N*(*x*). En la última columna, se muestran los intervalos de confianza para la distribución binomial para el nivel de significación α=0,05.

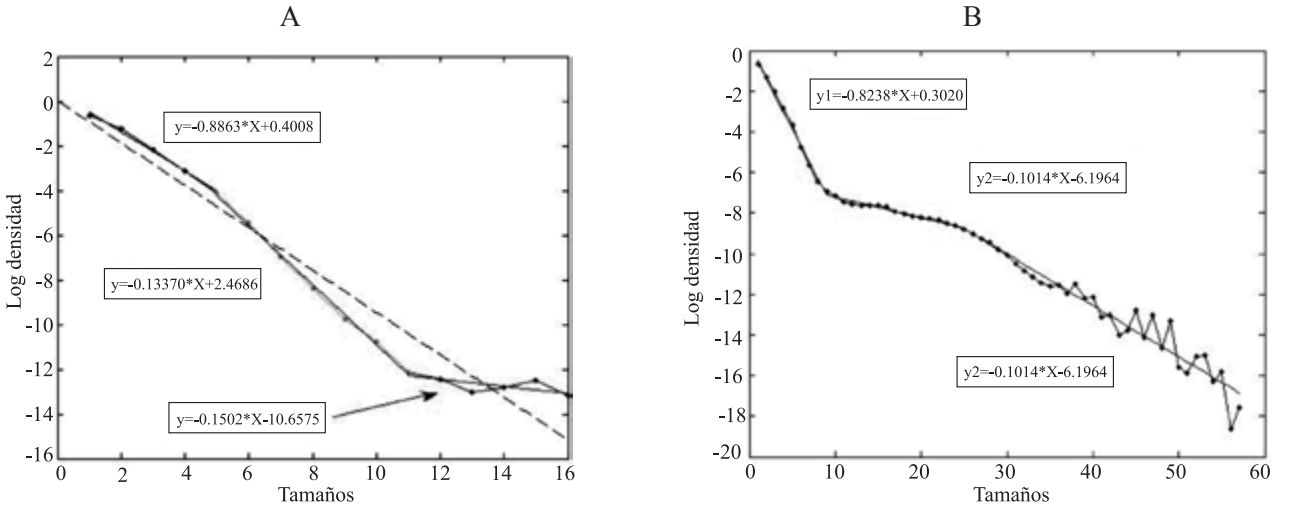
Resultados y Discusión

Cuando se analizó la relación entre la densidad media de repetidos y el tamaño de las repeticiones de mononucleótidos (Figura 1), se evidenció que existían ciertos puntos de inflexión a nivel de los repetidos de tamaño 6 y 12 para las regiones codificadoras y de los repetidos de tamaño 9 y 25 para las no codificadoras. Ello nos sugiere la presencia de mecanismos diferentes implicados en la generación de expansiones.

La existencia de umbrales o valores críticos de tamaños de repetidos en el mecanismo de deslizamiento de la polimerasa ha sido ampliamente estudiada. Ejemplo de ello lo constituyen estudios *in silico* en *Saccharomyces cerevisiae*,[15] en los que se plantea que es necesario un umbral de alrededor de ocho bases nucleotídicas para que ocurran expansiones por deslizamiento de la polimerasa. Otros estudios han reportado[16] un umbral de nueve mononucleótidos a partir del cual tiene lugar la inestabilidad mutacional. Estos resultados coinciden con lo encontrado para los intrones, es decir, el punto de inflexión en 9 (Figura 1 B), pero no así con el observado en los repetidos de la región codificadora (Figura 1 A). Es de señalar que cuando Lai y Sun, en el 2003, estudian el genoma en general, no lo separan por regiones, a ello se suma que solo el 2% del genoma codifica para proteínas, por lo que el aporte de las mismas puede ser despreciable en un estudio a escala genómica. Es por ello que nuestros resultados no son contradictorios.

Además, se observa una mayor restricción sobre los tamaños de los repetidos en las regiones codifica-

Figura 1. Logaritmo de la densidad media de repetidos contra el tamaño de las repeticiones de mononucleótidos. Se observan las ecuaciones que describen las rectas de mejor ajuste estimadas según el método de los mínimos cuadrados para cada una de las tres regiones. A: Región Codificadora. B: Región No Codificadora.



doras, hecho también reportado con anterioridad.[10,11] Ello indica que en las regiones codificadoras la función biológica, y la presión evolutiva vinculada a esta, contribuyen a que el equilibrio entre la contracción y la expansión se desplace más hacia la primera si se compara con el equilibrio entre estas dos fuerzas opuestas en la región no codificadora. Es muy probable que los mecanismos involucrados en el aumento de estas repeticiones puedan no ser los mismos en estas regiones. Conforme como Calabrese y Durrett (2003),[17] señalan, sería demasiado engorroso describir el proceso de deslizamiento de la polimerasa usando funciones simples. Nuestros resultados apoyan la tesis que plantea que existen diferentes mecanismos para el incremento de las expansiones, y que estos dependen no sólo del tamaño del repetido, sino también de las regiones del ADN, así como de la composición nucleotídica. (Figura 1 y 2).

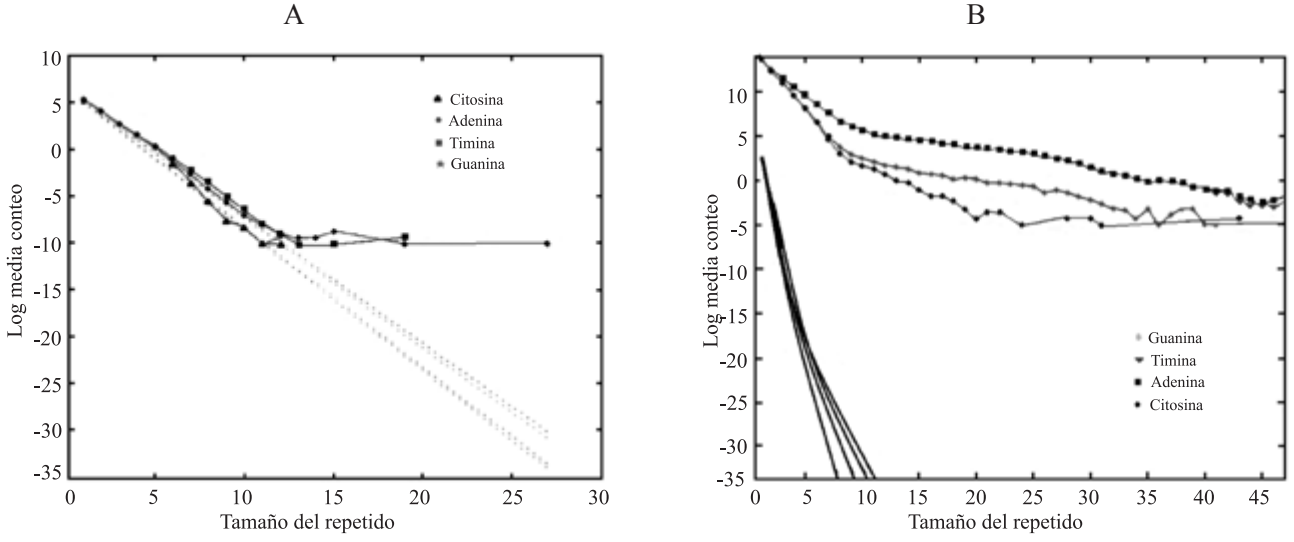
Comparación entre los repetidos de bases diferentes

Cuando se estudió la distribución de repetidos para cada una de las bases, se obtuvo que las cantidades de repetidos de nucleótidos de A y T eran superiores en todas las regiones a las de C y G (Figura 2). Estudios previos revelan que las repeticiones más comunes en el genoma humano son las secuencias de poli(A) y poli(T),[10,11]llegando a ser estas, incluso, 300 veces más frecuentes que las de poli(G) o poli(C).[11]Es posible que la abundancia de repeticiones de estas bases esté influida por su efecto en la estructura del ADN. Si los repetidos de poli(A) y poli(T) son más fáciles de escindir desde el punto de vista energético que los de poli(G) y poli(C), entonces existirían fuerzas que tienden a fijarlos evolutivamente con mayor probabilidad; estos mismos resultados fueron observados en bacterias.[18]

Los resultados expuestos en la figura 2 A muestran tres hechos interesantes: 1) La dinámica de los repetidos hasta el tamaño de cinco unidades para cada una de las cuatro bases se explica según un modelo de Bernoulli, 2) Los repetidos de C y G alcanzan menor longitud (como máximo hasta 12 nt) que los repetidos de A (como máximo hasta 27 nt) y T (como máximo hasta 20 nt) y 3) Las frecuencias observadas de los repetidos de más de cinco unidades de tamaño, para las bases C y G son menores que las esperadas de acuerdo con el modelo de cadena aleatoria de Bernoulli. La figura 2 B por el contrario demuestra que las frecuencias observadas de repetidos en las regiones no codificadoras son mucho mayores que las esperadas según este modelo, siendo esta tendencia más marcada en los repetidos de A y T.

El hecho de que la expansión de los repetidos está mucho menos restringida en los intrones se podría explicar por dos razones: por la menor densidad en estas zonas de secuencias con requerimientos funcionales estrictos, o por la función moduladora que tienen los homopolímeros de procesos como la transcripción y el *splicing*.[3,19]Por otro lado, se sabe que en ciertos mecanismos de expansión como el deslizamiento de la ADN polimerasa, la tasa de crecimiento del repetido aumenta con la cantidad de unidades repetitivas.[5]Por

Figura 2. Logaritmo del conteo medio observado de repetidos (Líneas con marcadores), logaritmo del conteo medio esperado según el modelo de Bernoulli (líneas sin marcadores) contra el tamaño de los repetidos para las cuatro bases, A, G, T y C, A: Región Codificadora B: Región no Codificadora.



ello al incrementarse el tamaño del microsatélite aumenta también su inestabilidad y con ello su propensión a contraerse, de ahí que la probabilidad de que se fijen tamaños mayores de repetidos en las regiones codificadores sea mucho menor que si se tuviera en cuenta sólo el proceso de expansión/contracción del tamaño del repetido impuesta por la estructura/función de la proteína a nivel evolutivo.

Por otra parte, se conoce que las alteraciones en el tamaño de homopolímeros de A pueden generar varios tipos de cáncer. Por ejemplo, una de las causas fundamentales del cáncer de colon no polipósico es el aumento en el número de repetidos de A, el cual provoca corrimientos en el marco de lectura de los genes que codifican para las proteínas del Sistema Reparador de Errores.[6,7]

Análisis estadístico de la relación entre las distribuciones de repetidos de mononucleótidos y las categorías de ontología de genes.

Para determinar el tamaño umbral de los repetidos por región, se tuvo en cuenta el cálculo previo de los puntos de inflexión (Figura 1), de manera que se seleccionaron los GO representados por una cantidad significativamente grande de CDS con repetidos de tamaño por encima de este valor umbral. Estos análisis se realizaron en las dos regiones de la estructura génica, para todos los polímeros de mononucleótidos simultáneamente y para las cuatro bases por separado.

En términos generales, para las dos regiones estudiadas, y en las tres divisiones principales de GO se cumple que existen muchos repetidos de gran tamaño en CDS asociados a proteínas del núcleo y de una forma u otra vinculados al ADN, es decir, regiones que codifican para proteínas unidoras a ácidos nucleicos, a metales y para proteínas que participan en los procesos básicos de la transcripción o replicación como las que interactúan con otras proteínas formando parte de complejos, actuando como factores de transcripción y receptores de membrana (Tablas 1 y 2). En este sentido se destacan también, aunque en menor medida, los CDS vinculados a: transducción de señales, sistema nervioso central, procesos musculares, ciclo celular y traducción. Reportes previos confirman parte de estos resultados,[4,12] por otro lado se conoce que muchas enfermedades neuro-degenerativas y musculares[8] tienen como causa fundamental la expansión del tamaño de repetidos en CDS asociados a estas mismas categorías GO. Sin embargo, lo sorprendente radica en que muchas de estas enfermedades están basadas en la expansión de tripletes, no de mononucleótidos. Es decir, las mismas categorías de GO relacionadas con estas enfermedades contienen también proporciones significativas de CDS con repetidos grandes de los nucleótidos A, T, C y G no sólo en las regiones codificadoras sino también en las no codificadoras.[3, 6, 20]

Tabla 1. Selección de instancias de las distintas categorías representativas a nivel de las regiones codificadoras con repetidos de tamaño superior al umbral de 7 nucleótidos.

No GO	GOP	Pt	N(x)	pb (x) /ph (x)	frecuencia (n>x)	ic binomial(0.05)
0006355	regulación de la transcripción	0,001	536	3,43 E-11/1,31 E-13	0,004	[0,003 0,007]
0006350	transcripción	0,005	373	1,66 E-7/ 3,46 E-8	0,008	[0,006 0,011]
0007268	transmisión sináptica	0,015	74	5,40 E-4/ 7,79 E-5	0,022	[0,018 0,026]
			GOC			
0005634	núcleo	0,161	1159	4,20 E-23/8,84 E-29	0,211	[0,201 0,222]
0005622	intracelular	0,027	189	1,51 E-3/ 3,98 E-4	0,034	[0,029 0,039]
			GOF			
0005515	proteínas de unión	0,144	976	4,96 E-13/ 7,69 E-16	0,178	[0,168 0,188]
0046872	unión a iones metálicos	0,085	597	1,59 E-9/ 1,02 E-12	0,109	[0,100 0,117]
0008270	unión al Zinc	0,083	640	1,49 E-18/ 3,50 E-23	0,116	[0,108 0,125]
0000166	unión a nucleótidos	0,075	549	1,24 E-11 / 3,56 E-13	0,100	[0,092 0,108]
0003677	unión al ADN	0,045	401	1,72 E-20/4,14 E-26	0,073	[0,066 0,080]
0003676	unión a ácidos nucleicos	0,033	270	2,75 E-10 /1,30 E-11	0,049	[0,043 0,055]
0005524	unión al ATP	0,061	491	1,99E-17//7,60 E-22	0,089	[0,082 0,097]

GOP: Ontología génica de procesos biológicos, GOF: Ontología génica de funciones moleculares y GOC: Ontología génica de componentes celulares.

Tabla 2. Selección de instancias de las distintas categorías representativas a nivel de las regiones no codificadoras con repetidos superiores al umbral de 20 nucleótidos.

No GO	GOP	Pt	N(x)	pb (x)/ph (x)	frecuencia (n>x)	ic binomial(0.05)
0006350	Transcripción	0,055	336	1,26 E-3 / 2,04 E-5	0,064	[0,058 0,069]
0007399	desarrollo del sistema nervioso	0,013	57	1,51 E-3/ 1,70 E-3	0,018	[0,015 0,023]
0007268	transmisión sináptica	0,009	40	1,42 E-4/ 7,72 E-4	0,017	[0,013 0,023]
0007517	inhibición del desarrollo muscular	0,005	50	2,17 E-3 / 5,24 E-4	0,008	[0,006 0,010]
0006936	inhibición de la contracción muscular	0,004	40	1,12 E-3 / 1,33 E-3	0,007	[0,006 0,010]
			GOC			
0005634	Núcleo	0,172	3282	2,27 E-3/ 1,27 E-7	0,179	[0,173 0,185]
0016020	membrana del fotoreceptor	0,145	3050	2,82E-5/1,41 E-19	0,153	[0,148 0,158]
0005737	Citoplasma	0,055	1102	3,70 E-4/ 1,07 E-5	0,153	[0,148 0,159]
0005886	membrana plasmática	0,028	109	4,58 E-5/ 4,77	0,029	[0,026 0,031]
0016021	integral de membrana	0,124	2565	NS/ 8,15 E-9	0,129	[0,124 0,134]
0005887	Complejo ATPasa	0,052	1161	NS/ 6,30 E-12	0,056	[0,052 0,059]
			GOF			
0005515	proteínas de unión	0,158	3368	9,48 E-15/ 2,06 E-30	0,169	[0,164 0,175]
0046872	unión a iones metálicos	0,092	1843	1,16 E-8 / 2,05 E-9	0,098	[0,094 0,102]
0008270	unión al Zinc	0,089	1902	3,56 E-12/3,30 E-20	0,095	[0,091 0,100]
0000166	unión a nucleótidos	0,083	1842	5,12 E-23/3,51 E-36	0,092	[0,088 0,096]
0003677	unión al ADN	0,046	801	2,19 E-3/ 7,61 E-5	0,051	[0,048 0,055]
0003700	factores de transcripción	0,044	758	2,15 E-3/ 5,26 E-8	0,039	[0,037 0,042]

GOP: Ontología génica de procesos biológicos, GOF: Ontología génica de funciones moleculares y GOC: Ontología génica de componentes celulares.

Hasta el momento el efecto de los microsatélites en la regulación génica está ampliamente documentado, ya sea a nivel de la transcripción, la replicación, la traducción o en la unión a proteínas.[3]Ello justifica en parte los resultados anteriormente descritos, pues permite explicar la tolerancia a los repetidos de gran tamaño en estas regiones. Además se sabe que variaciones en el número de repeticiones podrían modular la capacidad de unión de las proteínas que los contienen. Incluso, pueden formar estructuras cooperativas e interactuar con factores de transcripción.[2,3]

Combinaciones de dos y de tres categorías de Ontología Génica

Se encontraron varias combinaciones interesantes entre las categorías de la Ontología Génica de Funciones moleculares (GOF). Es decir, proporciones significativas de CDS con repetidos grandes asociados simultáneamente a actividades de unión a proteínas (a complejos proteicos o a otras macromoléculas) y a las de unión a metales. Los genes que tienen repetidos de tamaño por encima del umbral y combinaciones de tres GOF significativas, fueron aquellos que se distinguen por la presencia de: 1) dominios de unión al Zinc o de unión a metales, 2) a nucleótidos y 3) a otra proteína, complejos proteicos o macromoléculas.

Dentro de las combinaciones internas de la Ontología Génica de Procesos Biológicos (GOP) fueron significativas las que vinculan los procesos de regulación de la transcripción con los de la transcripción propiamente dicha y con los procesos que utilizan la unión a receptores para señalizar cambios celulares. Dentro de las combinaciones de tres GOP sobresalieron las que incluyen conjuntamente los procesos de transcripción, los de regulación de los mismos y aquellos procesos basados en la unión a receptores (datos no mostrados).

El análisis de las proporciones significativas de CDS con repetidos grandes y anotados con combi-

naciones entre las instancias de distintas categorías de GO, llamadas combinaciones externas, mostró que existen varias combinaciones de dos y de tres GO interesantes desde el punto de vista biológico (Tablas 3 y 4). Dentro de las que se destacan, en ambas regiones, las combinaciones entre las ontologías de procesos biológicos (GOP) y las de componentes celulares (GOC) vinculadas al proceso de modulación de la frecuencia, la tasa y la extensión de la transcripción y la localización en el núcleo. Cuando se examinaron las combinaciones externas entre los GOP y los GOF hallamos significativas las que relacionan el proceso que modula la frecuencia, la tasa y la extensión de la transcripción con las funciones representadas por proteínas de unión a otras proteínas, a ácidos nucleicos y a metales. El análisis de las combinaciones externas entre categorías de las tres ontologías GO mostró como más significativas las que vinculaban simultáneamente el proceso de modulación de la frecuencia, la tasa y la extensión de la transcripción con la localización nuclear y con las siguientes funciones: unión a proteínas, a complejos proteicos y a macromoléculas; las de unión a metales y, asimismo, las de unión a ácidos nucleicos (Tablas 3 y 4).

Todos estos resultados sugieren incuestionablemente que las proteínas que tienden a tener repetidos de gran tamaño, están vinculadas a procesos, componentes y funciones relacionadas fundamentalmente con los ácidos nucleicos, dentro de las que se destacan las vinculadas a la regulación de la transcripción. Como se mencionó previamente, dentro de estas últimas se destacan los factores de trascripción, además de las proteínas de unión al ADN; lo que concuerda con reportes previos.[12]

En la actualidad numerosos grupos estudian la utilidad inherente de estas secuencias repetitivas.[3,4] Otros se enfocan en los microsatélites vinculados a enfermedades, profundizando en los mecanismos que generan su inestabilidad.[8,20] Es por ello que el conocimiento sobre la asociación entre las distribuciones de repetidos nucleotídicos y las categorías de GO puede aportar mucho al estudio del rol biológico de las secuencias repetitivas que se encuentran en cualquiera de las regiones génicas, con tamaños superiores a lo esperado según modelos de cadenas aleatorias de tipo Bernoulliano o Markoviano. De esta manera se podrían aportar nuevas ideas para la predicción de genes con propensión a desarrollar enfermedades a partir de

Tabla 3. Selección de las combinaciones externas de dos y de tres categorías de Ontología Génica, en genes con repetidos de tamaño superior al umbral de 7 nucleótidos en las regiones codificadoras.

Comb. De GO	No GO	Pt	N(x)	pb (x)/ph (x)	frecuencia (n>x)	ic binomial(0.05)
GOC-GOF	0005634y0005515	0,044	383	1,83 E-17/4,71E-22	0,069	[0,063 0,077]
	0016020y0005515	0,028	161	3,09 E-2/3,44 E-2	0,029	[0,025 0,034]
	0016021y0005515	0,0+17	94	3,85 E-2/4,22 E-2	0,022	[0,018 0,026]
	0005887y0005515	0,011	56	2,81 E-2/2,65 E-2	0,017	[0,013 0,020]
	0005886y0005515	0,008	27	7,31 E-4/2,94 E-4	0,011	[0,009 0,014]
GOC-GOP	0005634y0006355	0,068	483	5,00 E-9/8,11 E-11	0,088	[0,080 0,096]
	0016020y0006355	0,001	12	9,92 E-3/5,28 E-3	0,006	[0,004 0,009]
GOF-GOP	0008270y0006355	0,031	265	8,02 E-8/1,13 E-13	0,041	[0,036 0,046]
	0046872y0006355	0,027	226	6,77 E-8/3,79 E-12	0,048	[0,042 0,054]
	0005515y0006355	0,017	190	1,73 E-17/9,34 E-23	0,036	[0,031 0,041]
	0005524y0006355	0,003	37	3,50 E-4/7,49 E-5	0,014	[0,011 0,018]
	0000166y0006355	0,003	38	4,29 E-5/3,98 E-6	0,034	[0,029 0,039]
GOF-GOP-GOC	0008270y0006355 y0005634	0,029	239	2,36 E-7/2,09 E-11	0,038	[0,033 0,043]
	0046872y0006355 y0005634	0,025	208	2,21 E-6/3,77 E-10	0,043	[0,038 0,049]
	0005524y0006355 y0005634	0,003	33	4,72 E-4/1,08 E-4	0,014	[0,011 0,017]
	0000166y0006355 y0005634	0,002	34	4,11 E-5/3,44 E-6	0,032	[0,027 0,037]

Nota: Los nombres de los GOP, GOC y GOF se pueden encontrar en las tablas 1 y 2.

Tabla 4. Muestra una selección de las combinaciones externas de dos y de tres categorías de Ontología Génica, en genes con repetidos de tamaño superior al umbral 20 nucleótidos en las regiones no codificadoras.

Comb. De GO	GO RNC	Pt	N(x)	pb (x)/ph (x)	Frecuencia (n>x)	ic binomial(0.05)
GOC-GOF	0005634y0005515	0,049	1068	1,25 E-4/1,04 E-17	0.053	[0.050 0.057]
	0016020y0005515	0,022	664	2,04 E-3/1,05 E-8	0.033	[0.031 0.036]
	0016021y0005515	0,019	400	1,52 E-2/8,57 E-3	0.021	[0.019 0.023]
	0005737y0005515	0,012	265	1,30 E-2/7,02 E-4	0.013	[0.011 0.015]
	0005886y0005515	0,009	211	1,23 E-2/1,80 E-4	0.010	[0.009 0.012]
GOC-GOP	0016021y0006355	0,001	28	4,46 E-2/4,65 E-3	0.004	[0.003 0.005]
	0016020y0006355	0,001	24	4,43 E-2/3,18 E-3	0.006	[0.005 0.007]
GOF-GOP	0046872y0006355	0,029	623	3,79 E-3/1,16 E-6	0.036	[0.034 0.039]
	0005515y0006355	0,019	430	2,26 E-3/2,17 E-7	0.032	[0.030 0.035]
	0000166y0006355	0,003	86	2,00 E-2/1,87 E-4	0.026	[0.023 0.028]
GOF-GOP-GOC	0003700y0006355 y0005634	0,035	636	2,73 E-4/6,91 E-9	0.004	[0.003 0.005]
	0046872y0006355 y0005634	0,027	586	5,08 E-3/9,07 E-6	0.034	[0.031 0.036]
	0005515y0006355 y0005634	0,017	389	2,42 E-3/2,05 E-9	0.032	[0.029 0.034]
	0000166y0006355 y0005634	0,003	73	2,58 E-2/1,06 E-3	0.024	[0.022 0.026]

Nota: Los nombres de los GOP, GOC y GOF se pueden encontrar en las tablas 1 y 2.

expansiones deletéreas de repetidos en bloques o por el contrario de genes que son prácticamente invulnerables para este tipo de mutación por el hecho de estar fuertemente protegidos por fuerzas evolutivas básicas que favorecen el acortamiento o la estabilización de tales secuencias, mecanismos que en este trabajo se describen sobre la base de un espacio ontológico tridimensional: las funciones moleculares, los componentes celulares y los procesos biológicos.

Referencias bibliográficas

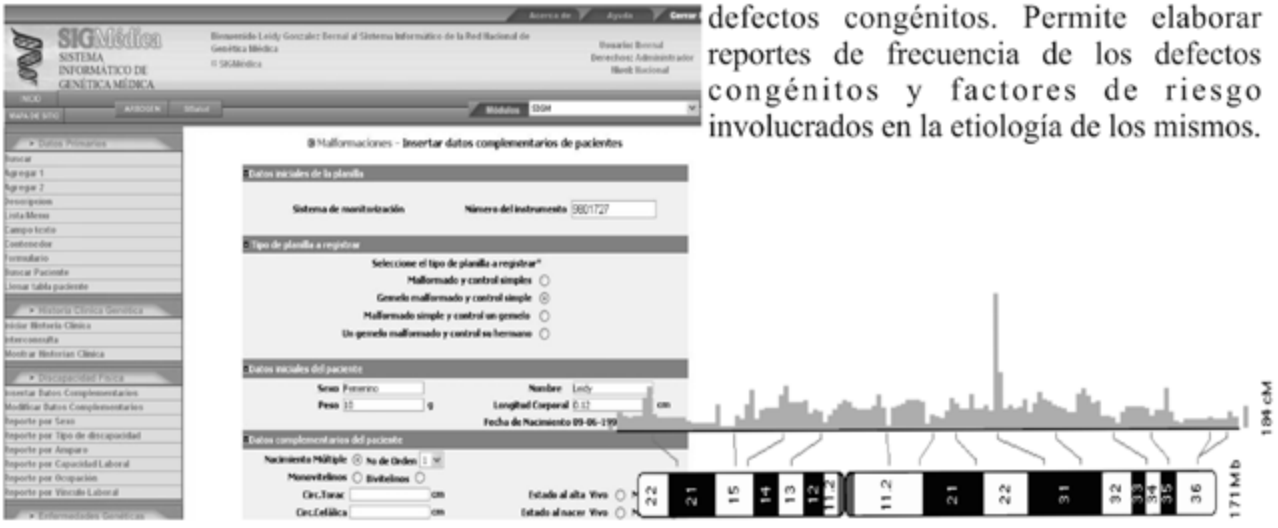
1. Venter, J.C., and coll. The sequence of the human genome. Science. 2001; 291(5507):1304-51.
2. Subirana, J.A. and Messeguer X. Structural families of genomic microsatellites. Gene. 2008; 408(1-2):124-32.
3. Li, Y. C., Korol, A. B. Fahima T., Beiles A., and Nevo E. Microsatellites Within Genes: Structure, Funtion, and Evolution. A review. Mol Ecol. 2004;11:2453-65.
4. Fondon, J.W., Hammock E.A.D., Hannan A. J. and King D. G. Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci. 2008; 31(7):328-34.
5. Kruglyak, S., Durrett T., Schug D. and Aquadro C.F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA 1998; (95):10774-78.
6. Aquilina, G. and Bignami M. Mismatch repair in correction of replication errors and processing of DNA damage. Journal of Cell Physiology. 2001;187:145-54.
7. Hussein, M.R. and Wood G.S. Building bridges in cancer: mismatch repair and microsatellite instability. American Journal of Dermatopathology. 2002; 24:76-81.
8. Pearson, C.E., Nichol Edamura K., and Cleary J.D., Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet. 2005; 6(10):729-42.
9. Madsen, B.E., Villesen P., and Wiuf C., Short bloques repeats in human exons: a target for disease mutations. BMC Genomics. 2008; 9: p. 410.
10. Subramanian, S., Madgula V. M., George R., Mishra R. K., Pandit M., Kumar C. S., Singh S and coll. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. Bioinformatics. 2003; 19(5):549-52.

11. Subramanian, S., Mishra R.K., and Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003; 4(2):R13.
12. Alba, M.M. and Guigo R. Comparative analysis of amino acid repeats in rodents and humans. Genome Res. 2004;14(4):549-54.
13. Rhee, S.Y., Wood V., Dolinski K., and Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008;9(7): 509-15.
14. Drãghici S., Khatri P., Martins R., Ostermeier GC. and Krawetz SA. Global functional profiling of gene expression. Genomics. 2003, 81:98-104.
15. Rose, O., and Falush D.: A threshold size for microsatellite expansion. Mol. Biol. Evol. 1998;15:613-615.
16. Lai, J. and Fengzhu S. The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. Mol. Biol. Evol. 2003;20(12):2123-31.
17. Calabrese, P. and Durrett R. Dinucleotides repeats in the Drosophila and Human genome have complex, length dependent mutation processes. Mol. Biol. Evol. 2003; 20:715-725.
18. Coenye, T. and Vandamme P. Characterization of Mononucleotide Repeats in Sequenced Prokaryotic. Genomes. 2005;85:105-107.
19. Bell, G. I. and Jurka J. The Length Distribution of Perfect Dimer Repetitive DNA Is Consistent with Its Evolution by an Unbiased Single-Step Mutation Process. J. Mol. Evol. 1997;44:414-21.
20. Parniewski P. and Staczec P. Molecular mechanisms of TNR instability. Adv. Exp. Med. Biol. 2002; 516:1-25.

Sistema Informático para la Red Nacional de Genética Médica: Registro Cubano de Malformaciones Congénitas

SIGMédica

Herramienta para gestionar la información relacionada con los recién nacidos con defectos congénitos. Permite elaborar reportes de frecuencia de los defectos congénitos y factores de riesgo involucrados en la etiología de los mismos.



Esta herramienta permite a los especialistas de cada provincia del país generar la información que antes viajaba en planillas sobre una plataforma informática, conocer y manejar los datos de los recién nacidos con defectos congénitos, los productos de embarazos múltiples así como las interrupciones de causa genética desde su origen y actuar de una manera rápida como un sistema de vigilancia epidemiológica de los defectos congénitos.