EXPLORATORY DATA ANALYSIS IN HUMAN GENOME

Jesús Murga , Edgar Garriga

***Abstract*— En este estudio se va a tratar el genoma humano, realizando un estudio estadistico y crear varias hipotesis sobre dichos datos.**

I. INTRODUCCIÓN E HIPÓTESIS

Se pretende realizar un análisis estadístico del genoma humano desde un punto de vista global de todos los cromosomas. Para ello se disponen de diversas variables que se utilizarán tanto para realizar un análisis descriptivo del genoma como para presentar ciertas hipótesis que se comprobarán su validez estadísticamente
Para el análisis descriptivo se ha trabajado con R y los datos proporcionados han sido modificados entre genes y genes no codificantes para tener unos datos más significativos.Se adjuntan las gráficas correspondientes en el anexo.
En referencia al tamaño (en Mb) podemos ver que la entrada más pequeña hace referencia al cromosoma mitocondrial, para el primer cuartil tendríamos un valor de 80, la mediana son 133Mb y la media 123Mb, el cromosoma de mayor tamaño tiene 248Mb y en este caso hace referencia al cromosoma 1.
Para el caso de la concentración de GC, este dato es utilizado por su posible relación con la cantidad de zonas codificadas. En este caso la concentración mínima se encuentra en el 38.3%, la media es 43.30% y la mayor concentración de GC es de 47.9%.En el caso de las proteínas encontramos la menor concentración se encuentra en el mitocondrio (13) seguido del cromosoma Y con 295. El caso opuesto se encuentra en el cromosoma 1 con 9684 proteínas.
Se han agrupado las regiones no codificantes, por lo que tanto rRNA, tRNA y Other_RNA . Obteniendo el mínimo en 24 regiones no codificantes, teniendo de media 2185 y con un máximo valor de 5045. Finalmente, los genes codificantes, han aportado los siguientes resultados; el mínimo se encuentra en 37 genes, la media son 2030 y el máximo 4778 genes.

II. HIPÓTESIS

A. Genoma codificante

Teniendo en cuenta todos los genes (codificantes y no codificantes) ¿qué proporción corresponde a los codificantes? ¿Mantienen la misma proporción todos los cromosomas? Conociendo la naturaleza de los cromosomas sexuales, ¿hemos de esperar una proporción de regiones diferentes a los autosómicos?

B. Relación tamaño y número de genes

El pensamiento clásico nos dice que el número de genes debe ser mayor en los genomas de mayor longitud. A lo largo de los años se ha demostrado que esta suposición no es cierta, debido a la complejidad de los genomas de mayor tamaño y su alto contenido de regios no codificantes. ¿Cabría esperar a nivel cromosómico que cuanto mayor sean los cromosomas codifiquen para un mayor número de genes? De ser cierto, ¿esperamos mayor densidad génica en los cromosomas de mayor tamaño?

C. Estimación de proteínas

El dogma central de la biología estableció que desde una secuencia de DNA, este se transcribiría a una molécula de mRNA y se traduciría una proteína. Teniendo en cuenta nuestros datos esperaríamos que por cada gen exista una proteína.

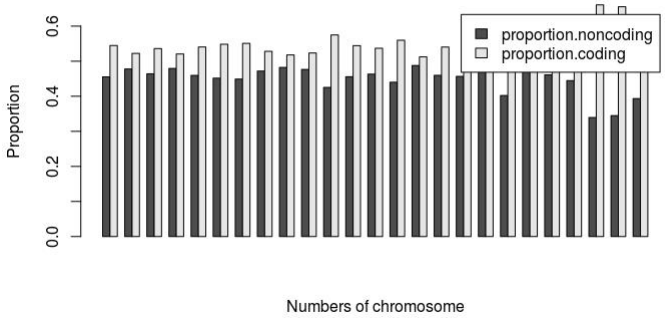
D. Relación %GC y número de genes

Teóricamente conocemos que los genes contienen un mayor %GC que las regiones no codificantes, . ¿Esperamos que los cromosomas con mayor contenido en GC posean un mayor número de genes?

III. RESULTADOSY CONCLUSIONES

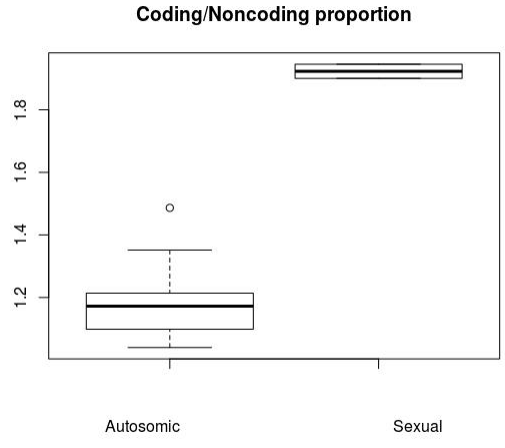
A. Genoma codificante

En primer lugar,para comprobar la proporción de genes codificantes y no codificantes se toman como variables genesy noncoding. A través de sus proporciones respecto al total de genes se realiza un diagrama de barra donde se muestran éstas, observables en el gráfico 1.



Para cada cromosoma observamos que generalmente hay mayor proporción de genes codificantes respecto a aquellos noncoding. Si realizamos la diferencia entre dichas proporciones para cada cromosoma, se observa que existe una mayor diferencia entre genes codificantes y noncoding

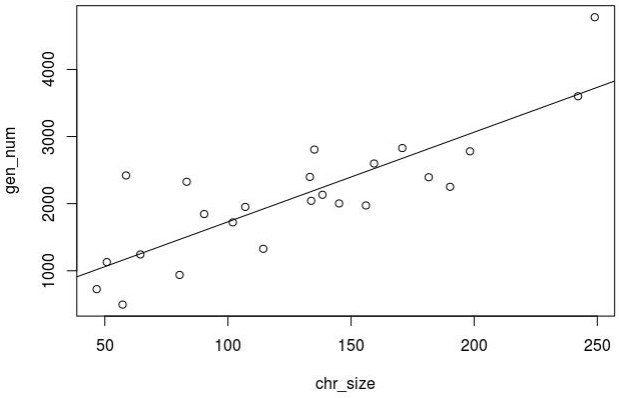
en los cromosomas sexuales y mitocondrial respecto a los autosómicos (Gráfico anexo 1.).
Para comprobar si la proporción de genes codificantes es diferente entre los cromosomas autosómicos y los sexuales, se realiza un test estadístico t-student. Los resultados muestran un p-value= 3.28·1010(¡0,05) y un ICal 95% (0.11, 0.13), lo cual nos indica que existen diferencias entre las proporciones de genes codificantes y el tipo de cromosomas.
Como hemos visto que existen más genes codificantes que noncoding, calcularíamos la relación entre ambos, esperando que la relación sea mayor que 1. Para los cromosomas autosómicos tendríamos una proporción ediade genes codificantes respecto a las no codificantes de 1.17. En cambio,para los cromosomas sexuales tendríamos unarelación media de genes codificantes respecto a los genes noncoding de 1.92.
Al establecer en un diagrama de cajas los datos observamos tres puntos satélites (gráfico anexo 2), que por sus valores podemos saber que hacen referencia al cromosoma 19 y los sexuales.
Diferenciando los cromosomas sexuales y autosómicos en un enun diagrama de cajas, gráfico 2,vemos sigue apareciendo el valor satélite correspondiente al cromosoma 19, el cual posee una proporción de 1.48. Esta elevada proporción podríamos estimar que es debida a que dicho cromosoma es el que posee una mayor densidad génica.Debido a la propia naturaleza de los cromosomas sexuales y por su forma tanto de expresión, como regulación se podría suponer que tuvieran una mayor relación codificante-noncoding.



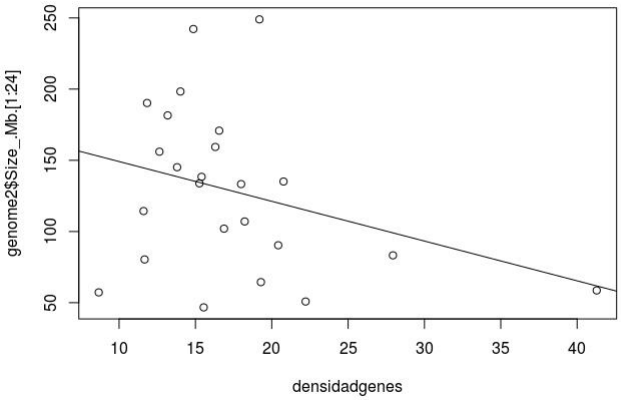
A pesar de que la proporción de codificantes es mayor que la proporción de noncoding, observamos que dicha relación no es tan elevada como cabría esperar, y que gran parte del genoma estaría ocupado por genes noncoding.

B. Relación tamaño y número de genes codificantes

Para solventar la hipótesis descrita previamente hemos decido tomar el número de genes y el tamaño de cada cromosoma. El estudio se realiza utilizando una regresión lineal donde tomamos los valores del tamaño cromosómico y el número de genes tal y como se muestra en el gráfico 3.

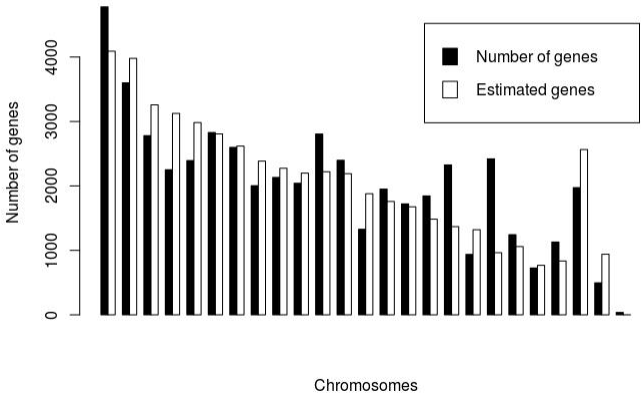


Observando el gráfico podemos discernir que cuanto mayor número de genes, mayor tamaño cromosómico. Así pues,para determinar numéricamente la correlación se utiliza untest de correlaciones. El valor obtenido de la correlación de ambas proporciones es de 0.83. Sin embargo,no lo podemos considerar estadísticamente significativo debido a un p-value= 4.91·107 y un ICal 95% (0.64, 0.92).
Con estos datos no se puede afirma rque exista una relación directa entre el tamaño y el número de genes en los cromosomas. Debemos tener en cuenta que esta correlación hace referencia al número de genes en cada cromosoma, para realizar un estudio más completo tomamos en cuenta la densidad genética de cada cromosoma y observamos si existe alguna correlación entre ésta y el tamaño.
Observando el gráfico 4, y el estudio de la correlación de ambos parámetros observamos que hay una correlación negativa de -0.41, un p-value= 0.131 (¿0,05) y un ICal 95% (-0.63, 0.09). A pesar de no tener un elevado índice de correlación, tenemos un p-value significativo y podríamos esperar que la densidad de genes fuera inversamente proporcional al tamaño del cromosoma.



Si estimamos el número de genes en base al tamaño de los cromosomas, observaríamos una distribución de genes en la cual cuanto mayor sea el cromosoma mayor número de genes poseería. Si comparamos dicha estimación con el

número real de genes por cromosoma, gráfico 5, observamos que no existe una correspondencia entre ambos datos, y que para un gran número de casos los genes estimados superan a los genes reales. También podemos observar que en los cromosomas más pequeños el número de reales es mayor que el estimado.

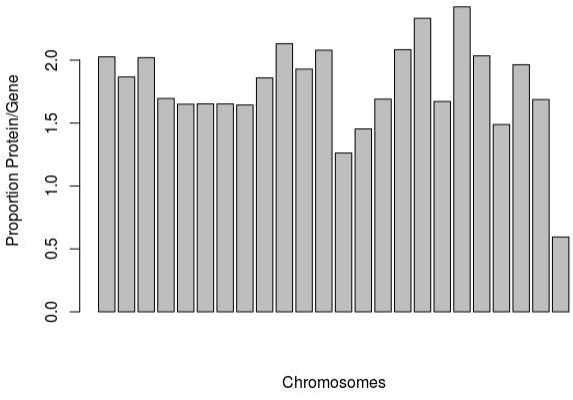


Con todos estos resultados podemos esperar que los genes no se repartan en base al tamaño y que se distribuyen de forma aleatoria, aunque si podría existir una relación entre la densidad de genes y el tamaño cromosómico

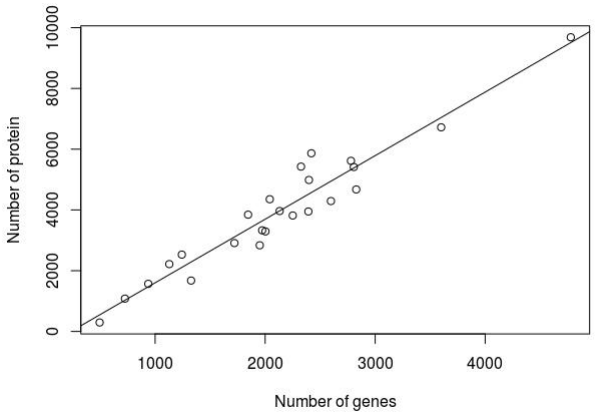
C. Estimación de proteínas

Teniendo en cuenta que por cada gen debería existir una proteína habría el mismo número de genes que de proteínas. Mediante una proporción, gráfico 6, entre el número de proteínas y el número de genes observamos que la relación se acerca a 2, lo que indica que hay más proteínas que genes.

$$\frac{TotalProtein}{TotalGenes} = 1.86 \tag{1}$$



Como ya sabemos la suposición del dogma central de la biología ha sido matizada, y conocemos que cada gen codifica más de una proteína mediante procesos de splicing alternativo y modificaciones post-traduccionales que hacen incrementar el número de proteínas por gen.



D. Relación %GC y número de genes

Para dicha hipótesis se va a trabajar con el tamaño del cromosoma, la concentración de GC por cromosoma y el número de genes por cada gen.
Se ha calculado que de media hay 41.30% de GC en el genoma. También se ha demostrado que de media existen 2112 genes.
El valor de la correlación es negativo, por lo que la relación es inversa. Mientras un valor sube el otro baja. Aun así, al obtener un resultado de -0.26 no se puede decir que existe relación entre concentraciónde GC y tamaño. Cosa que era esperada.