REVBIGO: Revista da Facultade de Bioloxía

ARTÍCULO

# Estudio del sesgo mutacional en contenido G+C en posiciones variables (SNPs) de las poblaciones humanas africana, asiática y centroeuropea

Trabajo Fin de Grado
Grado en Biología

Tutor:
Antonio Carvajal

Departamento de Bioquímica, Genética e Inmunología

Pérez Rodríguez, D.

daniel.prz.rodriguez@gmail.com

**Resumen**

Este trabajo tiene como objetivo estudiar el fenómeno de sesgo mutacional en SNPs de tres poblaciones humanas (africana, asiática y centroeuropea). Para ello, se realiza un análisis estadístico de los datos de contenido G+C de todos los SNPs de dichas poblaciones.

## Introducción

El genoma humano está compuesto de cuatro tipos diferentes de bases nitrogenadas que se emparejan mediante enlaces formando la conocida doble hélice de ADN. Estas parejas pueden ser de Adenina (A)-Timina (T) o Citosina (C)-Guanina (G). El tipo de enlace químico que las une es diferente, siendo un doble puente de hidrógeno para el par A-T y uno triple para el G-C (Figura 1). Esta diferencia es importante ya que los enlaces triples son más estables que los dobles y, por tanto, presentan una mayor resistencia a la desnaturalización.

Las bases nitrogenadas parece que no se distribuyen de una manera homogénea en el genoma, sino que hay regiones más ricas en una pareja o en la otra. En general, parece que hay una tendencia hacia la acumulación de A-T en sitios neutros del genoma (aquellos aparentemente no relacionados con ninguna función), y podría existir una acumulación de G-C en zonas codificantes (Yakoychuk *et al*., 2006; Hernández *et al*., 2007). Este fenómeno es conocido como sesgo mutacional.

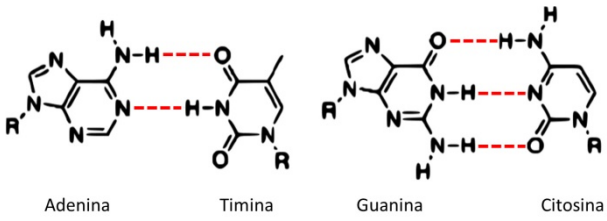


**Figura 1.** Esquema del apareamiento de bases nitrogenadas según el modelo de Watson-Crick.

Esta desigualdad mutacional se cuantifica con el llamado parámetro de sesgo mutacional *k*, que es la ratio de mutación de GC→AT y AT→GC (Sueoka, 1962; Li, 1987; Bulmer, 1991). Los cambios en el contenido G+C genómico (contenido total de citosinas y guaninas del genoma) están fuertemente correlacionados con la variación en *k*.

Asimismo, una mayor presencia G+C puede ser favorecida frente a A+T, debido al efecto de la selección natural junto con fenómenos de sustitución sinónima o conversión sesgada de genes (BGC, por sus siglas en inglés). La BCG ocurre cuando heterocigotos para variantes GC y AT en un nucleótido producen más del 50% de variantes GC en sus gametos como resultado de una reparación sesgada del ADN heterodúplex (Marais, 2003). Este fenómeno produce un cambio en las frecuencias GC frente a las variantes AT.

Por otra parte, el modelo mutacional de Löwding postula que la doble transferencia de protones entre bases (DPT) es la causante de un sesgo mutacional a favor de la conversión GC→AT, ya que se trata de un fenómeno químico espontáneo que ocurre con más frecuencia en los pares GC que en los AT.

Por lo tanto, existen dos fuerzas mutacionales que pueden dar lugar a un sesgo. Por un lado, la BCG que disminuye el valor de k (aumenta GC), y por otro la DPT que actúa de manera opuesta. Debido a que una sustitución AT→GC tiene más probabilidades de ser sinónima (Palidwor *et al*., 2010), se fijarán G y C con más frecuencia en regiones codificantes (donde un cambio en el aminoácido traducido es relevante). Por otra parte, en las regiones neutras (donde no es relevante un cambio de secuencia) suelen fijarse sustituciones GC→AT. Esto sucede así gracias al fenómeno espontáneo DPT (Sueoka, 1962; Li, 1987; Bulmer, 1991).

Con esto en mente, el estudio del contenido G+C utilizando datos de haplotipos de SNPs en poblaciones humanas tiene doble interés; por una parte, podemos comparar el sesgo mutacional genómico ya conocido (asociado a todas las posiciones nucleotídicas), con el asociado exclusivamente a las posiciones variables (SNPs) en haplotipos humanos (este estudio). Por otro lado, podemos desglosar los datos por cromosoma y población, y comparar si existen diferencias para la composición G+C entre los distintos cromosomas, así como entre algunas de las poblaciones humanas que están mejor diferenciadas genéticamente.

Por tanto, este trabajo se centra en el estudio del contenido Guanina + Citosina (G+C) en posiciones variables (SNPs) en tres poblaciones humanas de distinta procedencia geográfica para, posteriormente, estudiar la presencia o ausencia de sesgo mutacional. Los datos provienen del proyecto HapMap (International HapMap Consortium, 2003) y las poblaciones estudiadas provienen de Africa, Asia y Centroeuropa. Se desarrolló un algoritmo eficiente para realizar la extracción y el conteo de bases nitrogenadas a nivel genómico. Posteriormente, se aplicaron pruebas estadísticas (Regresión lineal, ANOVA, Chi cuadrado) para comparar el contenido G+C a nivel poblacional y cromosómico.

## Material y métodos

### Obtención de datos

Los datos empleados en este trabajo se han obtenido del proyecto HapMap (International HapMap Consortium, 2003) y corresponden a los haplotipos de tres poblaciones humanas distintas: centroeuropea (CEU), japonesa-china (JPT-CHB) y africana (YRI). Un haplotipo puede definirse como una combinación de alelos de distintos loci en una región cromosómica. Para referirse a las variaciones nucleotídicas (alelos) dentro de un mismo locus se utiliza el acrónimo SNP (por sus siglas en inglés, Single Nucleotide Polimorphism). Por tanto un haplotipo se puede definir también como un conjunto de SNPs pertenecientes a distintos locus que tienden a transmitirse juntos por estar en una misma región cromosómica. Se emplearon todos los cromosomas exceptuando los sexuales (22 cromosomas en total). El formato de los datos descargados de HapMap consiste en ficheros (uno por cada población y cromosoma) que contienen información sobre cada SNP, su posición dentro del cromosoma, el código del individuo de la muestra y el número de cromosoma. El número de SNPs y el número de individuos muestreados en cada población aparecen indicados en la Tabla 1.

| | | | Individuos | SNPs |
|---|---|---|---|---|
| CEU | Centroeuropea | Residentes en Utah con ancestros del norte y este europeo. | 36 | 116.147 |
| JPT-CHB | China | Residentes en Beijin procedentes de la etnia Han. | 342 | |
| | Japonesa | Japoneses residentes en Tokio. | | |
| YRI | Africana | Tribu yoruba en Ibadan, Nigeria. | 20 | |

**Tabla 1.** Procedencia y tamaño de las muestras empleadas en este trabajo (ftp://ftp.ncbi.nlm.nih.gov/hapmap/).

La lectura y manejo de los datos, formatos, así como los cálculos estadísticos se realizaron mediante el lenguaje de programación y paquete estadístico R (R Development Core Team, 2006) y en algún caso también en Microsoft Excel (2013).

## Métodos estadísticos

### Contraste para el sesgo mutacional (Test $x^2$)

El sesgo mutacional implica la existencia de valores G+C diferentes del esperado si las probabilidades de mutación entre las diferentes bases fueran las mismas. En este caso, esperaríamos que el porcentaje de contenido G+C fuera del 50%. La existencia de sesgo mutacional es ya conocida para la variación neutra en la especie humana (Fu *et al*., 2011). Para contrastar el sesgo mutacional en los datos haplotípicos utilizados en este trabajo, realizamos un test X2 en Excel.

Además, realizaremos también un test X2 con el objetivo de comprobar si existe alguna diferencia significativa entre el contenido G+C de los haplotipos de SNPs y el G+C cromosómico (todas las posiciones del cromosoma sean fijas o polimórficas). Para ello, primeramente se calculó el contenido G+C promedio de todos los individuos para cada población (valor observado) y se comparó con el valor esperado (50% para contrastar sesgo mutacional, y el valor de referencia para los cromosomas humanos para realizar la última comparación mencionada).

### Modelo lineal de dos factores: ANOVA multivariante en R

El objetivo principal de este análisis es averiguar si hay diferencia significativa en el contenido de G+C entre las poblaciones y/o los cromosomas estudiados. Para ello, se llevó a cabo un ANOVA multivariante.

### ANOVA Multivariante

El análisis ANOVA (de sus siglas en inglés ANalysis Of VAriance) es una colección de modelos estadísticos que se emplean para determinar cuándo un conjunto de datos pertenece a una misma muestra (y todas sus diferencias con respecto a la media pueden ser explicadas debido a errores aleatorios) o pertenece a muestras diferentes (existe un factor población) (Martínez, 2008). En nuestro caso, realizamos un modelo lineal de dos factores con interacción, en el que todo valor observado puede ser descompuesto en media (μ), efecto factor 1 (efecto poblacional, α), efecto factor 2 (efecto cromosoma, β), efecto de la interacción entre los factores población y cromosoma (αβ) y efecto aleatorio (ε).

$$Y_{i,j} = \mu + \alpha_i + \beta_j + (\alpha_i\beta_j) + \varepsilon_{i,j}$$

Para realizar el ANOVA multivariante en R se usó la función aov, en la que se fijó la variable dependiente e independiente. Esto con el propósito de averiguar si las variables están relacionadas entre sí. La sintaxis fue la siguiente:

ANOVA=aov (GC_Content~Population*Chromosome, tablaANOVA)

Primero se escribió la variable dependiente; contenido de G+C ("GC_Content"), y después se especificaron las independientes; "Population" (población) y "Chromosome" (cromosoma). Finalmente, se indica que los datos para llevar a cabo el ANOVA se encuentran en la estructura de datos (dataframe) llamada tablaANOVA.

| Población | Individuo | Cromosoma | Contenido G+C |
|---|---|---|---|
| 1 | 1 | 1 | 52,6 |
| 1 | 2 | 1 | 52,5 |
| 1 | 3 | 1 | 52,6 |
| 1 | 4 | 1 | 52,7 |
| 1 | 5 | 1 | 52,6 |
| 1 | 6 | 1 | 52,9 |
| 1 | 7 | 1 | 52,9 |

**Tabla 2.** Fragmento de los datos tabulados listos para ser analizados por la función aov de R. Los factores aparecen agrupados en columnas y las poblaciones codificadas en números.

En el ANOVA aplicado se estudió la relación existente entre pertenecer a una población y poseer una cantidad determinada de G+C y si esta cantidad en cada cromosoma dependía del factor población. Para ello, fue necesario adaptar las tablas de datos a un formato especial requerido por la función ANOVA. Este formato requería volcar todos los datos de todas las poblaciones y de cada individuo en una sola tabla, separando los factores en columnas (población, individuo, cromosoma, contenido GC). Además, las poblaciones tuvieron que ser codificadas en forma de números (1=CEU, 2=YRI, 3=JPT-CHB). Como ejemplo, ver la Tabla 2.
Una vez obtenida, se aplicará la función aov de R. De esta manera, se obtendrá una tabla resultado donde se observen los niveles de significación para los emparejamientos establecidos.

**Realización de controles para el ANOVA**

Debido al procesamiento que tienen los datos desde la fuente original hasta que finalmente se obtiene el resultado del ANOVA, es necesario verificar que no han sufrido modificaciones y que todos siguen correctamente emparejados (datos de población con sus correspondientes datos de contenido G+C, contenido de G+C con el cromosoma correspondiente, etc…); también es necesario asegurarse de que no hay contenido duplicado o ausente.
Para ello, se llevan a cabo cuatro controles diferentes en los que se modifican los datos de entrada al programa (de tal manera que se puede predecir el resultado). Estos datos son procesados automáticamente por el script que los agrupará y transformará en una tabla adecuada para posteriormente aplicar el ANOVA y arrojar un resultado que será el que se evalúe. Si no coincide con el resultado esperado, habría que revisar el código.
Los cuatro controles son los que figuran en la Tabla 3.

| | |
|---|---|
| CONTROL 1 | Consiste en una tabla con los datos de la misma población repetidos 3 veces; por tanto, se elimina cualquier efecto posible de población. En este caso, se repitió la población CEU. De esta manera podremos asegurarnos de que, si el código es correcto, esperamos que no haya efecto en el factor poblacional. |
| CONTROL 2 | Consiste en una tabla con los datos del cromosoma 1 de cada población repetidos 22 veces. Por tanto, eliminamos cualquier posible efecto de cromosoma. En este caso, esperamos que no haya efecto de cromosoma. |
| CONTROL 3 | Se va a eliminar cualquier efecto de población y cromosoma. Para ello se usa el cromosoma 1 de CEU para reemplazar el resto de los cromosomas en todas las poblaciones (CEU, YRI y JPT-CHB). De esta manera, esperamos que no haya efecto ni de población ni de cromosoma. |
| CONTROL 4 | Se va a contrastar la media y varianza del contenido G+C de la tabla que entra en el ANOVA con la media y varianza de las tablas G+C obtenidas tras realizar el conteo desde los datos originales. Este control tiene el objetivo de detectar cualquier error relacionado con la transferencia de datos de las tablas originales a las tablas del ANOVA. Bien porque no se copien todos o porque haya datos que se repiten. De esta manera, la media y la varianza entre las dos tablas tiene que ser la misma. |

**Tabla 3.** Controles empleados para verificar la validez de los datos usados para calcular el ANOVA.

**Función para el modelo lineal de dos factores**

Una vez obtenidos los resultados del ANOVA y comprobado que existe una relación entre las variables estudiadas, se aplicó la función de R lm. Esta función corresponde al modelo lineal de dos factores ya aplicado en el ANOVA y ofrece una salida más detallada de las interacciones entre factores que la función previa aov (Adler, 2009). Se empleó el mismo dataframe y sintaxis que para aov:

linearmod=lm(GC_Content~Chromosome*Population, tablaANOVA)

Finalmente se obtuvo un resumen del análisis mediante *summary (linearmod)* donde aparecían desglosadas las relaciones por cromosoma, población y sus interacciones.).

## Resultados y discusión

**Contenido G+C**

Tras haber realizado el conteo de bases nitrogenadas en los haplotipos de las poblaciones, se representaron en una gráfica (Figura 2) el contenido en % G+C de los haplotipos de los cromosomas de cada población junto con el del genoma humano de referencia.

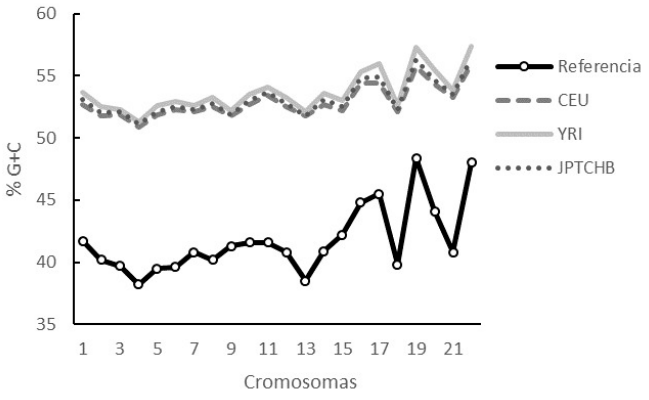


**Figura 2.** Comparativa del contenido de G+C de los haplotipos por cromosomas de las poblaciones CEU, YRI y JPT-CHB frente al genoma humano de referencia.

**Contraste para el sesgo mutacional (Test $x^2$)**

Como se puede apreciar en la Tabla 4, para las comparaciones entre las posiciones variables (SNPs) de las 3 poblaciones y el valor de referencia, los valores p de la prueba Chi cuadrado ($\chi^2$) son menores que el nivel de significación α=0,05. Por tanto, el contenido en G+C es significativamente distinto para el genoma completo comparado con el asociado a las posiciones variables de cada población o de su promedio.

| Población | Valor p (α=0,05) |
|---|---|
| CEU vs referencia | 7,1 E-07 |
| YRI vs referencia | 2,5 E-08 |
| JPT-CHB vs referencia | 5,8 E-07 |
| Media vs referencia | 2,3 E-07 |
| Media vs sin sesgo mutacional | 0,99979 |

**Tabla 4.** Resultado de la prueba Chi cuadrado en el supuesto de homogeneidad en el contenido G+C.

Se compara cada población con el GC del cromosoma de referencia y el promedio de las tres poblaciones con el cromosoma de referencia. En la última fila se indica el valor de la prueba para un supuesto de no sesgo.

El contenido G+C del genoma presenta sesgo mutacional pero no hay sesgo mutacional en el contenido G+C asociado a las posiciones variables.

Por tanto, parece interesante constatar que el sesgo mutacional conocido en cromosomas humanos hacia T+A, es decir % de G+C menor del 50 %, (Fu *et al*., 2011) no se manifiesta cuando estudiamos únicamente las posiciones variables. Las posibles causas de esto podrían estar relacionadas con el estatus evolutivo de las posiciones variables identificadas en el HapMap, pero en cualquier caso se requiere un estudio más profundo para poder entender el resultado.

**Realización de controles del proceso de transformación de datos previo al ANOVA**

CONTROL 1

| Factor | GL | Suma Cuadrados | F | P |
|---|---|---|---|---|
| Población | 2 | 0 | 0 | 1 |
| Cromosoma | 21 | 3723 | 2755 | <2e-16 (***) |
| Interacción | 42 | 0 | 0 | 1 |
| Códigos de signif. | 0 (***) 0,001 (**) 0.01 (*) 0,05 (.) 0,1 ( ) 1 | | | |

**Tabla 5.** Salida de la consola de R tras haber aplicado un ANOVA a una tabla con los valores de la población CEU sustituyendo los de YRI y JPTCHB. De esta manera se elimina el efecto de población.

El primer control (misma población) implicaba que en ningún caso el factor población podía ser significativo (el cromosoma podría serlo o no). Como se puede observar en la Tabla 5, el valor p para el factor población es 1, por lo que se acepta la hipótesis nula (las medias son iguales) y, por tanto, el resultado del primer control es satisfactorio.

CONTROL 2

El segundo control (mismo cromosoma) implicaba que en ningún caso el factor cromosoma podía ser significativo (la población podría serlo o no). Como se puede observar en la Tabla 6, desaparece el efecto de cromosoma por lo que el resultado del segundo control es satisfactorio.

| Factor | GL | Suma Cuadrados | F | P |
|---|---|---|---|---|
| Población | 2 | 199 | 58,534 | <2e-16 (***) |
| Cromosoma | 21 | 10 | 0,270 | 1,000 |
| Interacción | 42 | 39 | 0,548 | 0,992 |
| Códigos de signif. | 0 (***) 0,001 (**) 0.01 (*) 0,05 (.) 0,1 ( ) 1 | | | |

**Tabla 6.** Salida de la consola de R tras haber aplicado un ANOVA a una tabla con el primer cromosoma de cada población repetido 22 veces (todos los cromosomas se sustituyeron por el número 1). De esta manera se elimina el efecto de cromosoma.

CONTROL 3

El tercer control (misma población y cromosoma) implicaba que ningún factor podía ser significativo. Como se puede observar en la Tabla 7, no hay efecto de cromosoma, población, ni interacción, por lo que el resultado del tercer control es satisfactorio.

| Factor | GL | Suma Cuadrados | F | P |
|---|---|---|---|---|
| Población | 2 | 0 | 0,000 | 1,000 |
| Cromosoma | 21 | 8 | 0,255 | 1,000 |
| Interacción | 42 | 31 | 0,508 | 0,997 |
| Códigos de signif. | 0 (***) 0,001 (**) 0.01 (*) 0,05 (.) 0,1 ( ) 1 | | | |

**Tabla 7.** Salida de la consola de R tras haber aplicado un ANOVA a una tabla con los valores del cromosoma 1 de la población CEU sustituyendo todos los de YRI y JPT-CHB. De esta manera se elimina cualquier relación entre factores.

CONTROL 4

Los valores de ambas salidas,; media y varianza de la tabla original (52,976 y 1,871), y de la tabla con formato ANOVA (52,976 y 1,871), coincidieron exactamente por lo que el resultado del cuarto control es satisfactorio.

**ANOVA multivariante**

Como ya se indicó en el apartado de Métodos, en el ANOVA se fijaron como variables independientes los factores población y cromosoma y como variable respuesta el contenido G+C.

| Factor | G. Libertad | Sum. Cuadrados | Media | F valor | P |
|---|---|---|---|---|---|
| Población | 2 | 203 | 101.6 | 1678.6 | <2e-16 |
| Cromosoma | 21 | 15366 | 731.7 | 12093.2 | <2e-16 |
| Interacción | 42 | 50 | 1.2 | 19.6 | <2e-16 |
| Residuos | 8558 | 518 | 0.1 | | |

**Tabla 8.** Salida del ANOVA tras aplicar la función de R aov.

El resultado de la prueba ANOVA (Tabla 8) es significativo para los factores población, cromosoma e interacción (valor *p* <2e-16). Esto quiere decir que existen diferencias relevantes entre el contenido G+C de los factores y que esta diferencia también está presente cuando se tienen en cuenta posibles interacciones entre ambos elementos (población y cromosoma).

**Modelo lineal de dos factores**

El análisis del modelo lineal para desglosar la información de cada población y cromosoma utiliza uno de los valores como referencia. En el caso de las poblaciones la referencia es la población 1 que se compara con la 2 y la 3, en el caso de los cromosomas se emplea el 1 y en el análisis de la interacción se compara la población 1 (CEU) y el cromosoma 1 con el resto. Este es el motivo por el que no aparecen en los gráficos ni el cromosoma 1, ni la población CEU.

- Factor cromosoma:

Como se aprecia en la Figura 3, el factor cromosoma es significativo para todos menos para el 10 y 14; es decir, estos cromosomas dentro de cada población, no tendrían contenido G+C diferente con respecto al cromosoma 1.

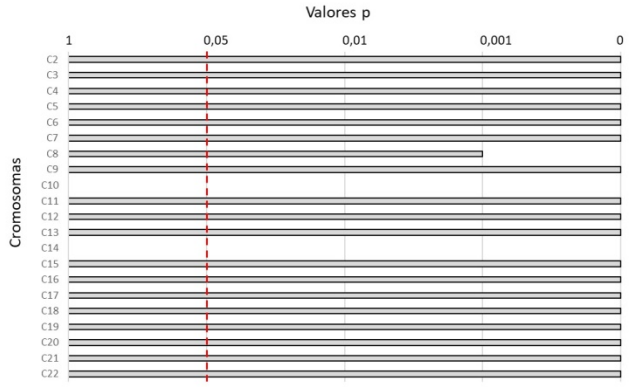


**Figura 3.** Resultado del contraste del factor cromosoma frente a contenido G+C. Aparecen representados los 21 cromosomas en el eje Y (el cromosoma 1 se emplea como intercepto).

- Factor población:

Como se aprecia en la Figura 4, el factor población es significativo en las dos comparaciones (CEU vs YRI y CEU vs JPT-CHB).

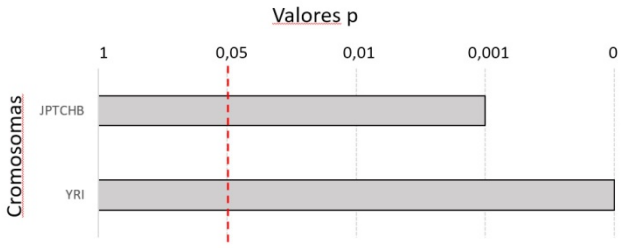


**Figura 4.** Resultado del contraste de los factores población frente a contenido G+C. Aparecen representados las 2 poblaciones en el eje Y (la 1 se emplea como intercepto) frente a su valor p.

**Interacción cromosoma y población:**

Una vez que sabemos que hay un efecto diferenciador de contenido G+C según se pertenezca a una población u otra y según se estudie un cromosoma u otro, nos interesa ahora saber si determinadas poblaciones interaccionan con determinados cromosomas a la hora de diferenciarse en su contenido G+C. En la Figura 5 se muestra el resultado del contraste de los factores cromosoma y población YRI frente al contenido G+C. Puede observarse como todos los cromosomas excepto el 16 y el 20 presentan interacción significativa con la población. Esto significa que, por ejemplo, si bien el efecto del cromosoma 2 es diferente al del resto de cromosomas, esta diferencia también se ve afectada según estemos evaluando este mismo cromosoma en la población europea (CEU) o la africana (YRI).

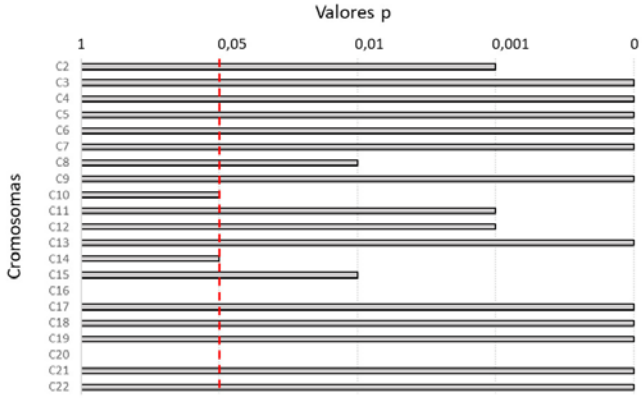


**Figura 5.** Resultado del contraste de los factores cromosoma y población 2 (YRI) frente a contenido GC. Aparecen representados los 21 cromosomas (el cromosoma1 se emplea como intercepto) en el eje Y frente a su p-valor.

**Factores cromosoma y población JPT-CHB**

En la Figura 6 se muestra el resultado del contraste de los factores cromosoma y población JPT-CHB frente al contenido G+C. Puede observarse como los cromosomas 2, 5, 8, 9, 11, 13, 16, 17 y 18 presentan interacción significativa con la población.

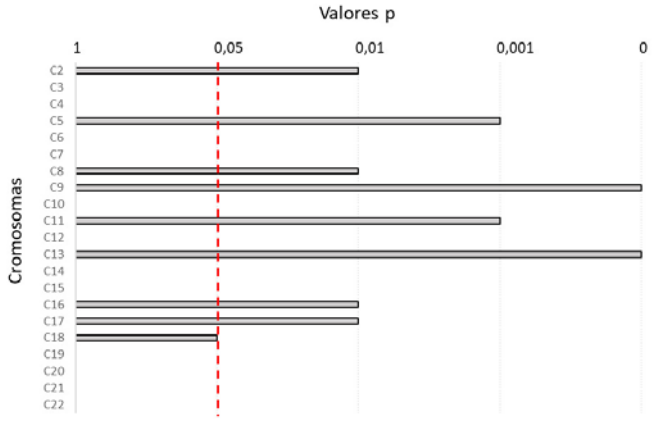


**Figura 6.** Resultado del contraste de los factores cromosoma y población 2 (YRI) frente a contenido GC. Aparecen representados los 21 cromosomas (el cromosoma1 se emplea como intercepto) en el eje Y frente a su p-valor.

Los resultados expuestos parecen confirmar que hay diferencias en contenido G+C asociado a las posiciones variables en los distintos cromosomas, algo que ya era conocido para el genoma humano (Sabbia *et al*., 2009). Sin embargo, raramente se ha estudiado de modo comparativo la composición G+C entre distintas poblaciones humanas (International HapMap Consortium,. 2007; Dutta *et al*., 2018) y hasta donde sabemos no se han estudiado posibles diferencias asociadas a determinados cromosomas y poblaciones (interacciones). Nuestros resultados parecen indicar que podría haber diferencias en la composición G+C dependientes de determinados cromosomas en poblaciones específicas. Sería interesante saber si las causas de estas diferencias podrían explicarse debido a diferencias históricas, selectivas, etc, asociadas a la composición genética específica de esos cromosomas en las poblaciones. Por ejemplo, cabría especular si pudieran estar relacionadas con alguna región del genoma donde se encuentran genes importantes tales como los del metabolismo o el desarrollo embrionario que pudieran haberse visto afectados de manera diferente a lo largo de la historia evolutiva en las distintas etnias humanas. Sin embargo, esto requiere un análisis mucho más profundo y completo que queda más allá de los objetivos de este trabajo.

## Conclusiones

- Los resultados obtenidos en este trabajo muestran que el sesgo mutacional a favor de A+T (G+C < 50%), cuya existencia es conocida para el genoma humano, no parece cumplirse cuando nos fijamos en las posiciones variables (SNPs), al menos en aquellas presentes para los tamaños de muestra y las poblaciones estudiadas del proyecto HapMap.

- Se hallaron diferencias significativas en el contenido G+C entre las poblaciones y los cromosomas estudiados, así como interacciones significativas entre algunos cromosomas y poblaciones.

- Sería interesante indagar en las posibles causas de las diferencias en el contenido G+C encontradas.

## Bibliografía


Adler, J. (2009). *R in a nutshell, a desktop quick reference*. San Francisco, USA: O'Reilly.

Bulmer, M. G. (1991). The selection–mutation–drift theory of synonymous codon usage. *Genetics*. 129: 897-907.

Cockell, S. (2012). [on line]. GCc-Content across different human chromosomes. *Biostars*. <https://www.biostars.org/p/16169/> [Consultado 09/05/2018].

Dutta, R., Saha-Mandal, A., Cheng, X., Qiu, S., Serpen, J., Fedorova, L., Fedorov, A. (2018). 1000 human genomes carry widespread signatures of GC biased gene conversion. *BMC genomics*. 19(1): 256. doi:10.1186/s12864-018-4593-1.

Fu, L. Y., Wang, G. Z., Ma, B. G., Zhang, H. Y. (2011). Exploring the common molecular basis for the universal DNA mutation bias: Revival of Löwdin mutation model. *Biochem Biophys Res Commun*, 409(3): 367-371.

Hernández, R. D., Williamson, S. H., Zhu, L., Bustamante, C. D. (2007). Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Mol Biol Evol*. 24(10): 2196-2202.

International HapMap Consortium. (2003). The international HapMap project. *Nature*. 426: 789-796.

International HapMap Consortium. (2007). A second-generation human haplotype map of over 3.1 million SNPs. *Nature*. 449(7164): 851.

Li, W. H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol*. 24(4): 337-345.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends Genet*. 19: 330-338.

Martínez, R. (2008). *El análisis multivariante en la investigación científica*. Madrid, España: La Muralla.

Palidwor, G. A., Perkins, T. J., Xia, X. (2010). A general model of codon bias due to GC mutational bias. *PLoS One* 5(10): e13431. https://doi.org/10.1371/journal.pone.0013431.

Sabbia, V., Romero, H., Musto, H., Naya, H. (2009). Composition profile of the human genome at the chromosome level. *J Biomol Struct Dyn*. 27(3): 361-369.

Sokal, R.R.,; Rohlf, F.J. (1995). *Biometry*, New York, USA: W.H. Freeman and Company.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci*. USA. 48: 582-592.

Yakovchuk, P., Protozanova, E., Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*. 34(2): 564-574. doi:10.1093/nar/gkj454.