GWAS study identifies SNPs variants possibly associated with coronary disease

Miguel Román and Daniel Soto

Contents

1 Abstract **2**

2 Methods **3**

2.1 Packages and tools used 3

2.2 Data description 3

2.3 Quality control 4

3 Results with figures **5**

4 Discussion **8**

5 Bibliography **9**

6 Appendix with supplementary figures **10**

1 Abstract

El objetivo principal de nuestro estudio consiste en identificar variantes de SNPs que esten asociados a enfermedades coronarias. Para ello se realizó un Genome Wide Association Study utilizando sets de datos disponibles mediante el software web-based PLINK.Posteriormente, se realizaron los análisis utilizando R. Los datos fueron tratados para un manejo más eficiente y filtrados para excluir SNPs e individuos que no cumpliesen los requerimientos o pudiesen sesgar nuestros resultados (detallado la siguiente sección). A continuación, se llevó a cabo el GWAS, que se ajustó con la covariable edad, y se representó mediante un Manhatan plot. Aquellas variantes con un nivel de significación superior al threshold empleado se analizaron en su contexto genómico mediante el software LocusZoom. También se llevó a cabo investigación bibliográfica para valorar su idoneidad para futuros estudios moleculares. Considerando todas las variables suggerimos que deberían ampliarse los estudios referentes a la variante rs6089510 y el gen Cdh4.

2 Methods

2.1 Packages and tools used

Para realizar los análisis presentados en nuestro estudio trabajamos con el siguiente conjunto de paquetes para R:

- ggplot2: nos permite generar diversos tipos de plots (histogram, barplot, scatterplot, linearplots, etc) y personalizar la estética, classificación de datos, añadir líneas de significación y guardar los plots en archivos entre otros.

- dplyr: nos facilita poder manipular los datos de forma simple e intuitiva respecto a los comandos base de R. Lo utilizamos para llevar a cabo parte del filtrado de SNPs e individuos o hacer intesección de diversos data.frames.

- ggrepel: se utiliza como complemento a ggplot2 para poder generar etiquetas en los plots sin que las mismas se solapen entre ellas o tapen datos de los plots.

- devtools: lo empleamos para facilitar la instalación de otros paquetes.

- BiocManager: nos permite instalar y trabajar con paquetes del proyecto 'Bioconductor', requerido para realizar análisis estadísticos e interpretar datos genómicos obtenidos por high-throughpout.

- SNPassoc: contiene funciones que nos permiten detectar y filtrar individuos con elevada kinship.

- snpStats: contiene clases y métodos estadísticos aplicados en estudios de associación de SNPs a gran escala.

- SNPRelate: este paquete permite almacenar datos de SNPs en formato binario y se aplica para realizar de forma eficiente análisis IBD (Identity by Descent) a partir d archivos GDS.

Para poder realizar el estudio de asociación utilizamos la función *snp.rhs.tests*. Esta nos permite correr el análisis con variables cuantitativas (bmi, body mass index) usando el argumento *family = 'Gaussian'* y ajustar los datos por la covariable edad.

Posteriormente aplicamos funciones de ggplot2 para crear un manhattan plot que representase nuestros datos. En el mismo aplicamos una significación 1e-05, superior a la que obteníamos al realizar la corrección de Bonferroni era demasiado estricta.

Una vez hecho el GWAS y el manhattan plot, para poder comprobar el contexto genómico de los SNPs significativos, utilizamos el software web-based LocusZoom. Este nos permite introducir los datos del GWAS en un archivo txt i ubicar el SNP (el identificador que le indiquemos) en la versión del assembly del genoma humano que le indiquemos. El sotware nos devuelve un report para cada SNP que incluye los genes en una ventana de 2Mb centrada en la posición del SNP.

2.2 Data description

Los datos que se utilizaron fueron diversos sets descargados mediante el uso del software PLINK. La información corresponde a datos de pacientes afectados de enfermedades coronáreas.

En primer lugar, obtuvimos tres ficheros cada uno con información específica. El bed contiene el genotipo(asignado al data.frame *coronary.genotype*), el bim la anotación de los SNPs (*annotation*) y el fam la información de los individuos (*individuals*).

Por otra parte, descargamos datos fenotípicos a partir de un fichero txt y lo asignamos al data.frame *coronary.phenotype*.

Finalmente usando los dos últimos data.frames mencionados utilizamos la función intersect para generar dos nuevos que contienen los datos con los mismo IDs -> *genotyoe* y *phenotype*.

2.3 Quality control

Antes de llevar a cabo el GWAS realizamos filtrados de los datos descargados con la finalidad de corregir sesgos en los resultados posteriores.

En el caso de los SNPs debemos tener en cuenta para descartar los siguientes casos:

- SNPs con un 'call rate' menor al 95 por ciento, es decir, que el nº de individuos con datos NA sea inferior al 5 por ciento.

- SNPs con una MAF (minor allele frequency) baja, inferior al 5 por ciento.

- SNPs que no superen el test HWE (Hardy-Weinberg Equilibrium)

Los individuos de la muestra también debemos someterlos a filtrado para excluir los siguientes casos:

- Individuos con discrepancias entre el sexo genómico y el reportado en los datos (Figura 6 Annexo).

- Individuos sin genotipo reportado o con valores de heterozigosis que disten de forma significativa de la heterozigosis promedio (Figura 7 Annexo).

- Individuos con una relación de parentesco (kinship) superior a 0,1. Finalmente almacenamos los individuos que han superado los filtrados en las variables *genotype.qc* y *phenotype.qc* para su uso en el GWAS.

3 Results with figures

Los resultados del GWAS pueden observarse en el gráfico inferior. En el mismo se representan los SNPs distribuidos en los 23 cromosomas analizados y con su nivel de significación. Únicamente 4 de ellos se encuentran por encima de nuestro threshold y, por tanto, se consideran significativos (Figura 1). Estos SNPs corresponden a rs6542685,rs1175853, rs1337687 y rs6089510.

Posteriormente localizamos el contexto genómico de cada uno de los 4 SNPs con LocusZoom. De los 4 SNPs significativos solo rs6089510 se solapa con un gen codificante de proteína, CDH4 (Figura 2).El resto se ubican en regiones intergénicas (Figuras 3-5).

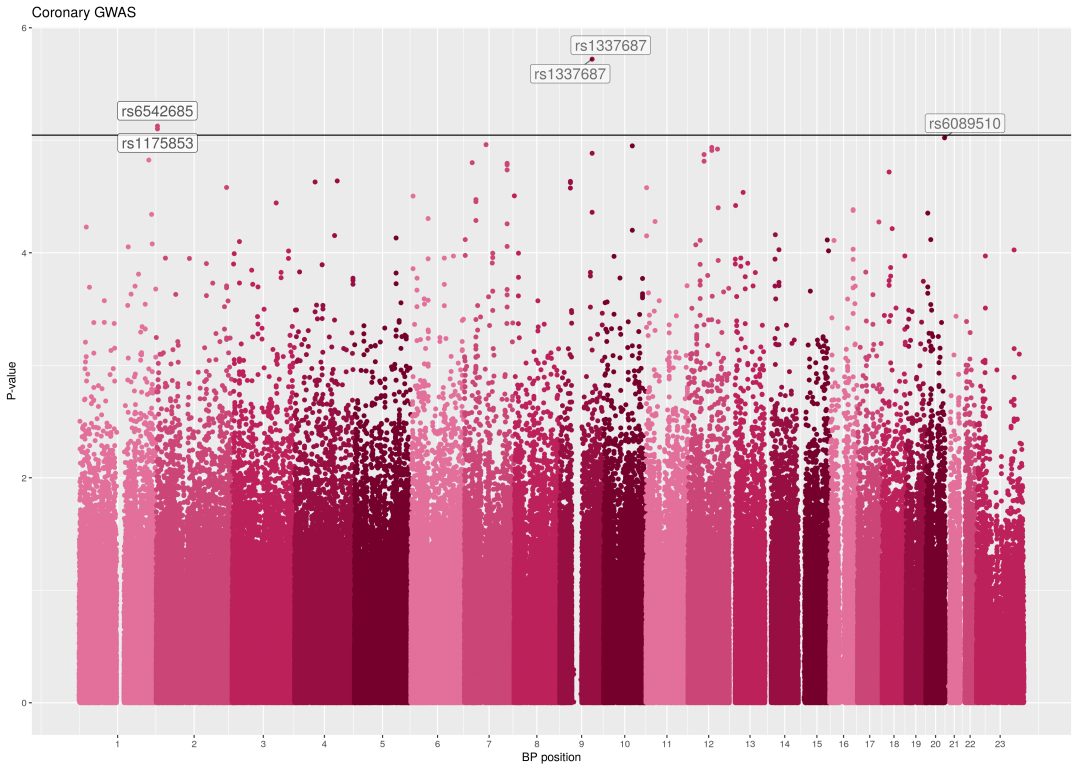

Figure 1 – Manhattan plot for the coronary data. The plot represents the significance of all our SNPs and shows those higher to 1e-05

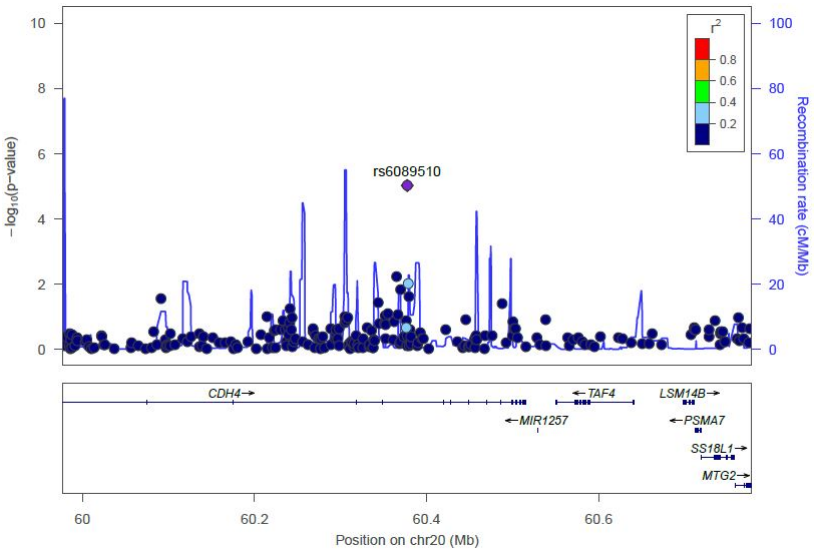

Figure 2 – Report de LocusZoom para rs6089510. El SNP solapa en el gen CDH4 de la cadena forward.

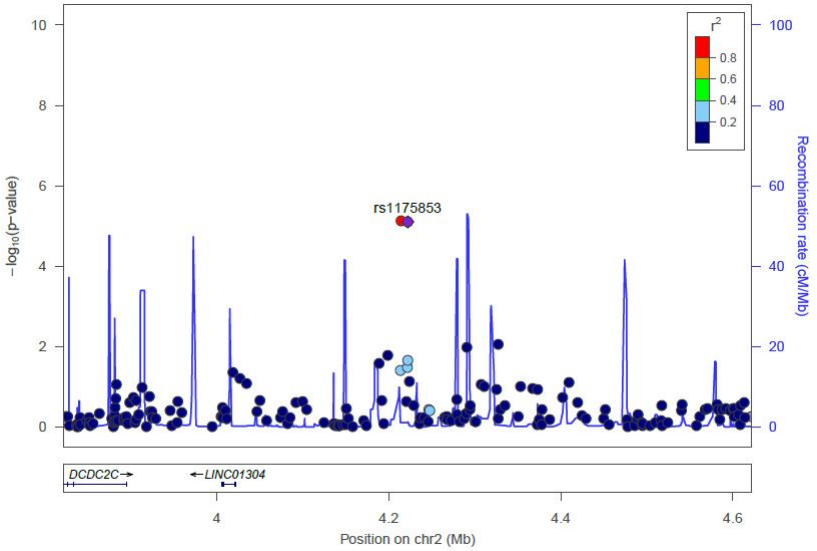

Figure 3 – Report de LocusZoom para rs1175853. El SNP se encuentra en una región intergénica. Gen más próximo: LINC01304, cadena reversa

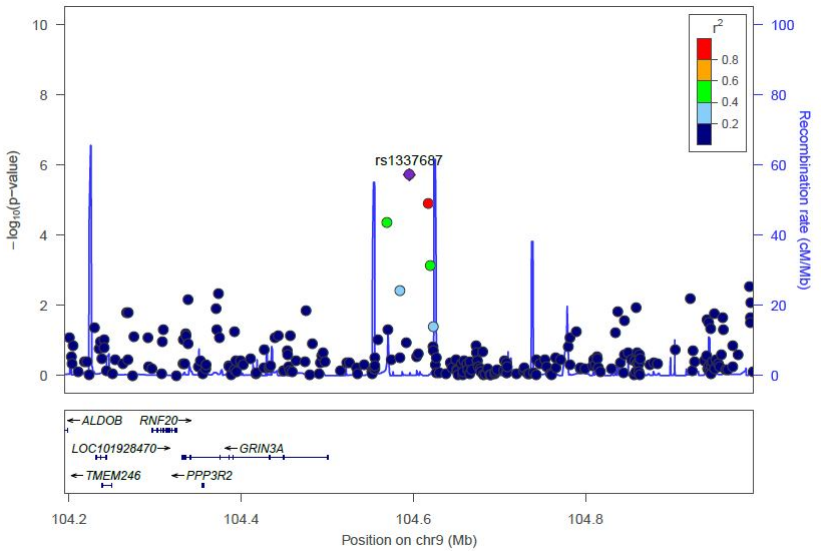

Figure 4 – Report de LocusZoom para rs1337687. El SNP se encuentra en una región intergénica. Gen más próximo: GRIN3A, cadena reversa

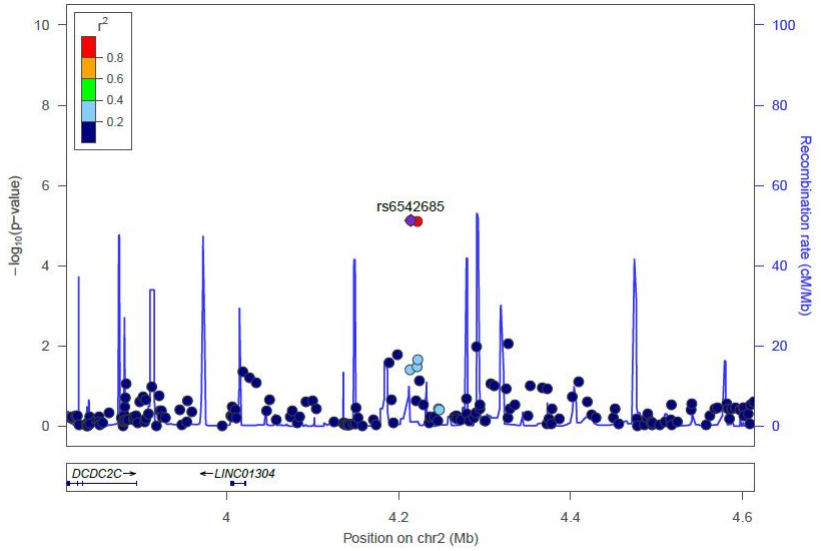

Figure 5 – Report de LocusZoom para rs6542685. El SNP se encuentra en una región intergénica. Gen más próximo: LINC01304, cadena reversa

4 Discussion

Los resultados de nuestro análisis nos han aportado SNPs candidatos possiblemente asociados a las enfermedades coronáreas y su ubicación en el genoma. Sin embargo, antes de llevar a cabo cualquier estudio molecular tenemos que investigar más sobre los genes y/o mecanismos de regulación génica que podrían verse modificados con cada alelo. Así pues, préviamente debemos conocer si el gen CDH4 se ha descrito como implicado en rutas metabólicas moleculares o celulares asociadas a la enfermedad. De la misma forma, también sería de interés identificar que funciones moleculares presentan los genes próximos a los otros SNPs y estudiar posibles efectos en la regulación (enhancers, inhibidores, etc).

En el caso de rs6089510,el SNP es descrito en RefSNP (NCBI) como una variante intrónica que no se ha asociado a ninguna condición clínica o publicación. Sin embargo, dada su posición podría tener efectos sobre el splicing alternativo entre otras opciones. El gen Cdh4 codifica una cadherina, una proteína de adhesión intercelular, que se ha asociado principalmente al desarrollo neuronal, muscular y nefrítico. En la sección "Phenotypes" de la entrada del gen encontramos un estudio GWAS previo (*Aouizerat, B. E. et al., 2011*). Basándonos en estos hechos suggeriríamos iniciar estudios moleculares que contrasten los efectos fenotípicos del SNP.

En los casos de rs1175853 y rs6542685 no existe referencia alguna sobre sus efectos en las regiones intergénicas en RefSNP. Su gen más cercano, LINC01304, corresponde a un lncRNA cuya expresión está restringida a los testículos. Por tanto, consideramos que no serían objetivos de estudios moleculares. Dado que en este caso existe muy poca información disponible no creemos recomendable llevar a cabo estudios moleculares.

Por último, rs1337687 se describe en RefSNP como una variante intrónica de un gen sin caracterizar de NCBI (LOC105376187), un ncRNA. Su otro gen más próximo en LocuZoom, GRIN3A, codifica una subunidad receptora de glutamato asociada principalmente a procesos fisiológicos y patólogicos en sistema nervioso. Por otra parte, hay dos papers que han asociado el gen con patologias de la arteria coronaria (*(Wojczynski, M. K. et al., 2013; Lin, Y. J. et al., 2013)*. Así pues, la decisión en este caso particular sería un poco compleja puesto que hay una gran falta de información sobre el posible gen con el SNP pero no debe descartar-se una afectación de secuencias reguladoras de la expresión de GRIN3A.

A pesar de todos los resultados y consideraciones obtenidas debemos tener en cuenta que nuestro análisis GWAS contaba con un threshold más permisivo del que obtendríamos mediante la corrección de Bonferroni. Por ello sería preferible realizar nuevamente el estudio aplicando diferentes métodos de corrección de la significación para corroborar nuestros resultados.

5 Bibliography

Aouizerat, B. E., Vittinghoff, E., Musone, S. L., Pawlikowska, L., Kwok, P. Y., Olgin, J. E. and Tseng, Z. H. (2011). GWAS for discovery and replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. BMC cardiovascular disorders, 11, 29. doi:10.1186/1471-2261-11-29

Wojczynski, M. K., Li, M., Bielak, L. F., Kerr, K. F., Reiner, A. P., Wong, N. D., … Reilly, M. P. (2013). Genetics of coronary artery calcification among African Americans, a meta-analysis. *BMC medical genetics*, 14, 75. doi:10.1186/1471-2350-14-75

Lin, Y. J., Chang, J. S., Liu, X., Hung, C. H., Lin, T. H., Huang, S. M., … Tsai, F. J. (2013). Association between GRIN3A gene polymorphism in Kawasaki disease and coronary artery aneurysms in Taiwanese children. *PloS one*, 8(11), e81384. doi:10.1371/journal.pone.0081384

6 Appendix with supplementary figures

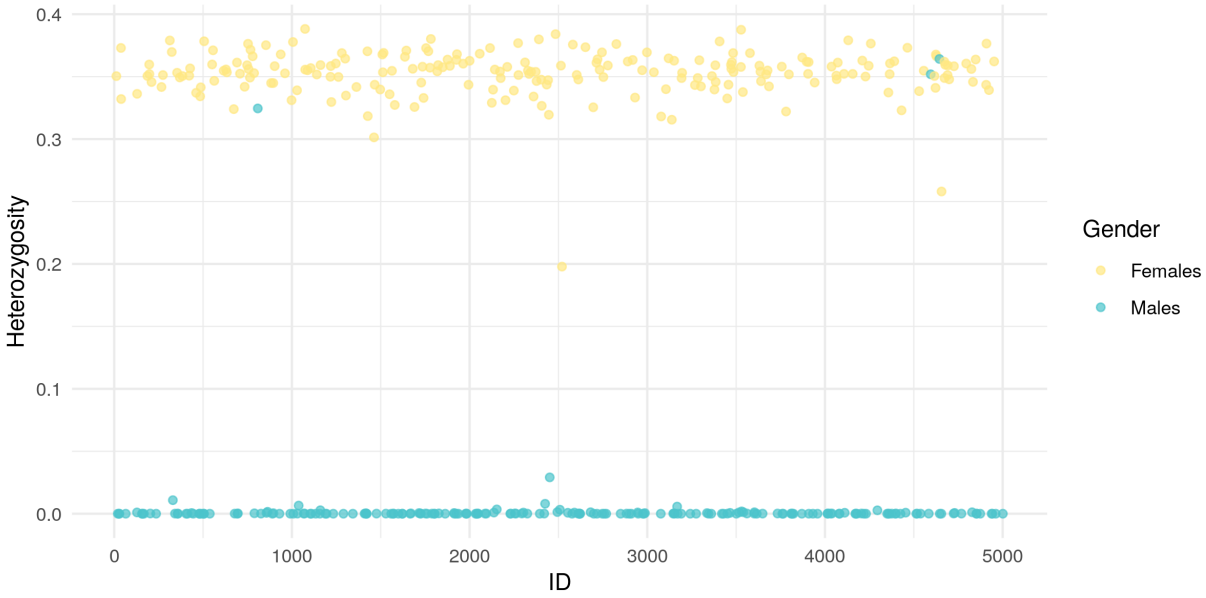

Figure 6 – Sex discrepancies in our sample

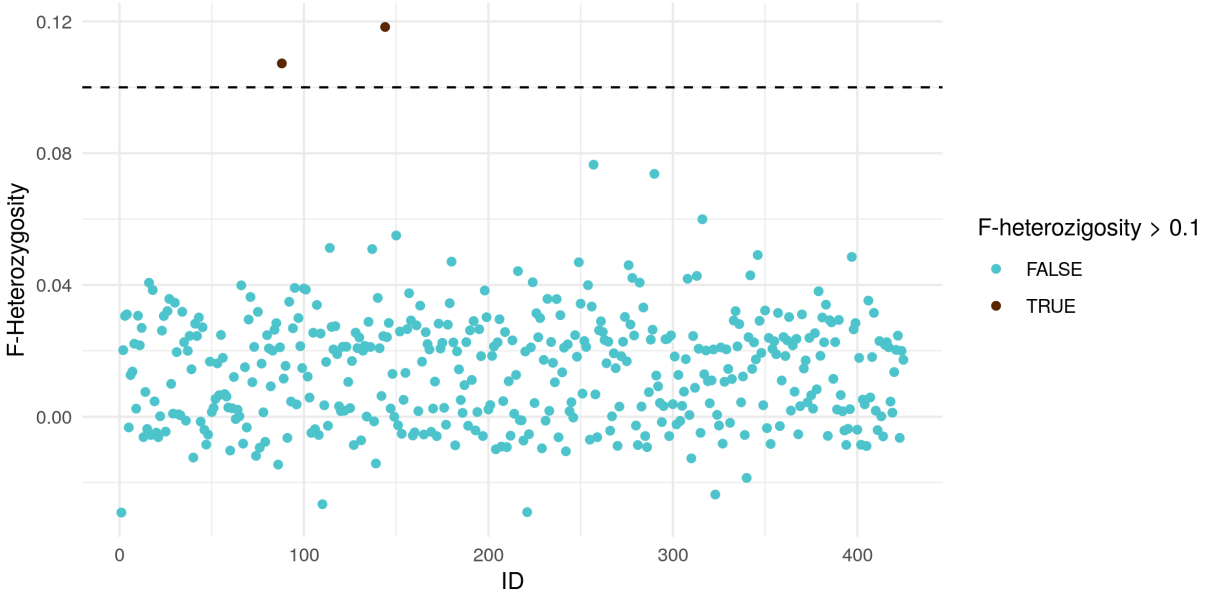

Figure 7 – Outlying heterozygozis rate individuals