ANÁLISIS INTEGRATIVO DE DATOS GENÓMICOS RELACIONADOS CON LA GRASA INTRAMUSCULAR DEL CERDO

Gonzalez-Prendes[1], R., Ramayo-Caldas[2], Y., Ros-Freixedes[1], R., Solé[1], E., Estany[1], J. y Pena[1], RN.

[1] Departament de Ciència Animal, Universitat de Lleida–Agrotecnio Center, Lleida. [2]Animal Breeding and Genetics Program, IRTA, Torre Marimon, Caldes de Montbui.

rayner.prendes@ca.udl.cat

INTRODUCCIÓN

Los avances en la genómica han permitido el análisis de diferentes fuentes de información biológica en paralelo, incluyendo información genómica, transcriptómica, epigenómica, proteómica y metabolómica (Hasin et al., 2017). En cerdos el estudio de cada uno por separado ha permitido obtener importantes avances en el conocimiento de las bases moleculares de los caracteres como la grasa intramuscular (GIM; Wang et al., 2017) y su composición (Estany et al., 2017). Los análisis integrativos prometen brindar información más completa sobre los sistemas biológicos en estudio (Cesar et al., 2018). En este trabajo se ha implementado un análisis integrativo de información genotípica y transcriptómica con fenotipos de GIM y su composición medidos en cerdos Duroc con el fin de identificar SNPs, factores de transcripción y genes candidatos asociados al metabolismo de los lípidos en esta población.

MATERIAL Y MÉTODOS

Muestras biológicas y diseño experimental: Se partió de una población de 256 cerdos Duroc criados en condiciones comerciales. Se realizó la extracción de los lípidos totales de las muestras de *gluteus medius* de toda la población y el contenido de GIM y su composición fueron cuantificados por cromatografía de gases (Bosch et al., 2009). Se seleccionaron 40 animales distribuidos equilibradamente entre dos tratamientos de vitamina A (con y sin suplemento) en la dieta para secuenciar el transcriptoma del músculo semimembranoso. **Datos de genotipado y control de calidad:** Toda la población fue genotipada con el Porcine SNP 70K BeadChip (GeneSeek, Illumina). En el control de calidad se eliminaron los SNPs localizados en más de dos posiciones, los que tenían una frecuencia alélica inferior al 5 % y los que mapearon en los cromosomas sexuales. Finalmente, se obtuvieron un total de 41.135 SNPs informativos con una tasa de genotipado superior al 95%. **Medición de la expresión génica mediate RNA-seq:** Se aisló el ARN de las 40 muestras de músculo semimembranoso mediante TRI Reagent (Invitrogen) y el kit *Direct-zol™ RNA Miniprep Plus* (Zymo Research). Después de comprobar la integridad, las muestras fueron secuenciadas en el Centre Nacional d'Anàlisi Genòmica en un equipo HiSeq 2500XL (Illumina, Inc.). Las muestras fueron mapeadas con el programa STAR (Dobin et al., 2013) de acuerdo al genoma de referencia porcino Sscrofa11.1 y los transcritos se cuantificaron con el programa Feature Counts (Liao et al., 2014). **Análisis de los datos:** Se realizó un análisis de asociación GWAS y se seleccionaron los genes y SNPs localizados en las regiones asociadas a GIM y su composición. Mediante un análisis factorial entre grupos con el programa *Multi-Omics Factor Analysis* (Argelaguet et al., 2018), detectamos los factores que mejor capturan la variación en las tres fuentes de información: fenotipos, SNPs y RNAseq en el subgrupo de 40 cerdos seleccionados. De esta forma, se obtuvieron tanto los ejes que mejor describen/capturan las variables latentes de heterogeneidad específicas de cada uno, así como las compartidas por las tres fuentes de datos y que por consecuencia sugieren fuentes de interacción. Por último, para detectar factores de transcripción (TF) relacionados con los genes candidatos obtenidos con MOFA, se calculó el factor de impacto de regulación génica (RIF) (Reverter et al., 2010). Con este valor RIF identificamos los TF co-expresados con genes diferencialmente expresados según la suplementación con vitamina en el pienso de los cerdos.

RESULTADOS Y DISCUSIÓN

En el estudio GWAS con los 256 animales se identificaron 30 regiones génicas asociadas a 8 caracteres de GIM y su composición, distribuidas en los cromosomas SSC1, SSC2, SSC3, SSC4, SSC5, SSC7, SSC8, SSC9, SSC12 y SSC14. En total, se seleccionaron 290 genes y 555 SNPs en estas regiones con los que se realizó el análisis multifactorial para identificar genes y SNPs candidatos relacionados con el metabolismo de los ácidos grasos.

Se encontraron un total de 5 factores de variación latentes (LF1 a 5) que explican al menos un 2 % de la varianza en uno de los grupos de datos. De forma acumulativa, los 5 factores identificados explicaron el 23% de la variabilidad de la expresión de los genes, el 60 % de los fenotipos estudiados y el 68 % de la variación de los SNPs (**Figura 1**). De ellos, 4 fueron específicos de uno de los grupos de datos (LF2 – fenotipos, LF3 – SNPs y LF4/5 - expresión). En cambio, LF1 fue el que mayor porcentaje de la varianza explicó en las 3 matrices (SNPs, RNAseq y fenotipos), indicando una co-variación conjunta (interacción) de las tres fuentes de información. Como estudio preliminar, nos centramos en el factor LF1 y seleccionamos los 40 genes y 40 SNPs con valores más extremos de ***loading factor*** (cargas de las variables en el factor latente).

Entre los genes seleccionados se encuentran, entre otros, el gen "*stearoyl-CoA desaturase*" (*SCD*) que interviene en la formación de ácidos grasos monoinsaturados, ampliamente estudiado en nuestra población (Ros-Freixedes et al., 2016; Estany et al., 2014), el "*cytochrome c oxidase homologo 15*" (*COX15*), que participa en la cadena respiratoria mitocondrial (Petruzzella et al, 1998) y el "*ATP binding cassette subfamily C member 2*" (*ABCC2*) cuya función está relacionada con el transporte de moléculas a través de las membranas celulares. De los 40 genes y 40 SNPs, el 30 % co-localizaron en la misma región genómica SSC14 (108-114 Mb; **Tabla 1**) mientras que el resto mapearon en regiones diferentes. La sobrerrepresentación de la región del SSC14 probablemente responda a un efecto de arrastre del gen *SCD* sobre la expresión de los genes vecinos, un efecto conocido como "***expression piggybacking"***, descrito en varias especies y que representa una fuente más de variabilidad en la expresión de los genes (Ghanbarian y Hurst, 2015).

Tabla 1*. Colocalización de genes y SNPs con valores extremos en el Factor de variación latente 1 (LF1) en la región SSC14:108-114 Mb.*

Región	Gen (acrónimo)	*loading factor*	SNPs en la región	*loading factor*
SSC14(108,9-108,9 Mb)	*HOGA1*	0,02	INRA0046585	0,81
SSC14(109,0-109,1 Mb)	*ZFYVE27*	0,03	ASGA0066024	0,81
SSC14(110,9-111,0 Mb)	*ABCC2*	-0,05	ASGA0066098	1,13
SSC14(111,5-111,6 Mb)	*SEC31B*	0,07	MARC0050687	1,00
SSC14(112,3-112,5 Mb)	*BTRC*	-0,03	WU_10,2_14_122239321	-1,85
			INRA0046731	-1,85
			MARC0083967	-1,85
SSC14(112,6-112,7 Mb)	*FBXW4*	-0,07	ASGA0066158	-1,85
SSC14(113,8-113,8 Mb)	*AS3MT*	-0,02	INRA0046767	0,98
SSC14(113,4-113,4 Mb)	*CUEDC2*	-0,03	DIAS0004744	-1,72
SSC14(114,4-114,5 Mb)	*NEURL1*	-0,02	H3GA0042130	0,98
			ASGA0066225	0,98

Con el objetivo de comprender mejor la regulación de los genes que más discriminan según el LF1 y conocer si existe un efecto de la vitamina A sobre su regulación se calculó el factor RIF. Con este enfoque, seleccionamos los 60 TF más extremos según el valor RIF, donde destacan el receptor del ácido retinoico beta (RARB), cuya función se ha demostrado que promueve la oxidación de los ácidos grasos y contribuye el control de la homeostasis (Li et al., 201; Levi et al., 2015), y algunas de sus parejas con las que forman heterodímeros (LXRB), o con los que compiten (YY1); dos miembros de la familia homebox (HOXA13 y

HOXD8), cuya expresión es afectada directamente por el ácido retinoico (Szatmari et al., 2010; Marshall et al., 1996); o el regulador de la actividad mitocondrial PPRC1.
En conclusión, aunque los resultados son preliminares, la integración de datos fenotípicos, genotípicos y transcriptómicos nos ha permitido identificar fuentes de variación entre los genes expresados y los SNPs, así como posibles fenómenos de arrastre de expresión.

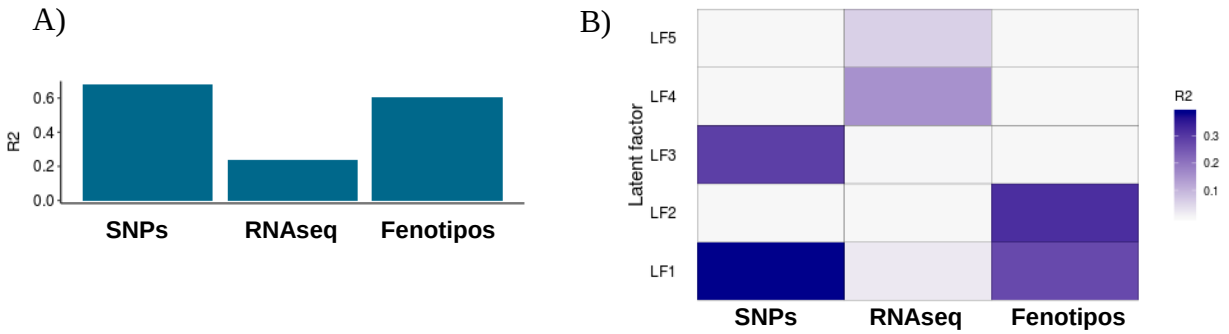

Figura 1*. Varianza explicada por los factores latentes (LF). A) Total de la varianza explicada (R^2) por los 5 factores en cada grupo de datos. B) Varianza capturada por cada factor en los tres grupos de datos.*

REFERENCIAS BIBLIOGRÁFICAS

• Argelaguet, R. et al. 2018. Mol. Syst. Biol. 14: e8124. • Bosch, L. et al. 2009. Meat Sci. 82: 432-437. • Cesar, A.S.M. et al. 2018. BMC Genomics. 19: 499. • Dobin, A. et al. 2013. Bioinformatics 29:15-21. • Estany, J. et al. 2017. J Anim. Sci. 95: 2261-2271. • Ghanbarian, A.T. & Hurst, L.D. 2015. Mol. Biol. Evol. 32: 1748-66. • Hasin, Y. et al. 2017. Genome Biol. 18:83. • Levi, L., et al. 2015. Nat. Commun. 6: 8794. • Li, Y. et al. 2013. J. Biol. Chem. 288: 10490-504. • Liao, Y. et al. 2014. Bioinformatics 30: 923-30. • Marshall, H. et al. 1996. FASEB J. 10:969-78. • Petruzzella, V. et al. 1998. Genomics 54:494-504 • Reverter, A. et al. 2010. Bioinformatics 26:896-904. • Ros-Freixedes, R. et al. 2016. PLoS One 11: e0152496. • Szatmari, I. et al. 2010. Sytem Cells 28:1518–29. • Wang, Y. et al. 2017. BMC Genomics 18:780.

Agradecimientos: Experimento financiado MINECO y fondos FEDER (AGL2015-65846-R). E. Solé disfruta de una beca predoctoral de la Universidad de Lleida.

INTEGRATIVE ANALYSIS OF GENOMIC DATA RELATED WITH PIG INTRAMUSCULAR FAT

ABSTRACT: In the present study an integrative approach with DNA variants, mRNA and lipid related traits was used to identify transcription factors and candidate genes related to lipid metabolism. With this objective, a population of 256 Duroc pig was genotyped with the 70k porcine BeadChip, and 22 intramuscular and fat composition traits determined in the *gluteus medius* muscle. From a subgroup of 40 pigs, the transcriptome of the semimembranosus muscle was sequenced with the HiSeq 2500XL (Illumina, Inc.). A total of 290 genes and 555 SNPs from the genome wide association study (GWAs) candidate regions were selected for the multi-omics factor analysis. With this approach, we identify 5 hidden factors which together explained 23% of variation in gene expression, 60% of SNPs and 68% of lipid traits data. Latent Factor 1 (LF1) was active in all datasets (SNPs, RNAseq and traits) whereas factors 2, 3, 4 and 5 were active in a single data modality. The 30% most extreme genes co-localized with SNPs in the SSC14 (108-114 Mb) genome region indicating a gene *expression piggybacking* effect. In addition we provide a list of genes and transcription factors that contribute to broaden knowledge of the genetic regulation of fatty acids traits.

Keywords*:* fatty acids, pigs, integrative analysis