IDENTIFICACIÓN Y VALIDACIÓN DE CNVs EN PORCINO

Revilla[1,2], M., Puig-Oliveras, A., Crespo-Piazuelo,D., Fernández, A.I., Ballester, M. y Folch, J.M.
[1]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Barcelona. [2]Centre de Recerca en Agrigenòmica (Consorci CSIC-IRTA-UAB-UB), Edifici CRAG, Campus UAB, 08193 Bellaterra, Barcelona. manuel.revilla@cragenomica.es

INTRODUCCIÓN

Los CNVs (*Copy Number Variants*) son un tipo de polimorfismo genético estructural que consiste en la variación en el número de copias de un fragmento de DNA. Los CNVs pueden alterar la expresión génica y afectar a la variación de caracteres de importancia económica en animales domésticos.

En los últimos años, la secuenciación paralela masiva de millones de lecturas o *Next Generation Sequencing* (NGS) ha permitido el análisis global del genoma y se han desarrollado estrategias para la identificación de CNVs (Metzker, 2010). Entre las ventajas que el enfoque basado en NGS proporciona, se encuentran: una mayor resolución, una estimación más precisa del número de copias y una mayor capacidad para identificar nuevos CNVs (Alkan *et al*., 2011).

Se han descrito varias estrategias para la detección de CNVs utilizando datos de NGS: *paired-end mapping* (PEM), *split read* (SR), *read depth* (RD), *assembly de novo* de un genoma (AS), así como una combinación de los enfoques anteriores (CB) (Zhao *et al*., 2013). Debido a la alta cobertura generalmente proporcionada por los datos de NGS, los enfoques basados en RD se han convertido en un buen estimador para definir el número de copias. Ello es debido a que la cobertura de una región genómica está relacionada con el número de copias de la región (Teo *et al*., 2012).

En un trabajo anterior, se identificaron 49 CNVs segregando en un cruce entre el cerdo Ibérico y Landrace (IBMAP) utilizando los genotipos del panel de 60 K SNPs (Ramayo-Caldas *et al*., 2010). Fernández *et al.* (2014) identificaron 65 regiones de CNVs segregando en animales ibéricos utilizando datos de NGS y genotipos del panel de 60 K SNPs.

El objetivo del presente trabajo es identificar CNVs segregando en la población IBMAP a partir de datos de NGS del genoma de animales fundadores y estudiar el efecto de los CNVs en la determinación de caracteres de crecimiento y calidad de la carne en porcino.

MATERIAL Y MÉTODOS

Material animal y secuenciación: El material animal utilizado procede del cruce IBMAP, generado por el cruce inicial de verracos Ibéricos Guadyerbas y cerdas Landrace (Pérez-Enciso *et al*., 2000). Dos machos Ibéricos y cinco hembras Landrace se secuenciaron en un equipo HiSeq2000 (Illumina), con lecturas pareadas de 75 pb de longitud, generándose una media por animal de más de 307 millones de lecturas. Estas lecturas fueron mapeadas contra el genoma de referencia porcino (*Sscrofa10.2*), con una cobertura media de 13,3 X.

Detección de CNVs: La detección de CNVs se realizó utilizando el programa *Control-FREEC* (Boeva *et al*., 2012), basado en el enfoque RD. La detección de los CNVs se realizó analizando cada muestra contra el genoma de referencia en cerdos (*Sscrofa10.2*). Con el objetivo de refinar el análisis de los datos de secuenciación de alto rendimiento y evitar falsos positivos se utilizó el fichero de *mappability* creado con *gem-indexer* (http://gemlibrary.sourceforge.net). El valor de significación de cada CNV se calculó mediante el test no paramétrico *Wilcoxon Rank Sum Test*.

Solapamiento de CNVRs: La identificación de CNVRs (agrupación de CNVs próximos en el genoma con al menos un par de bases de solapamiento) se realizó con la herramienta *MultiIntersect* de *BEDTools* (Aaron *et al*., 2010).

Actualización de Assembly: Para comparar las regiones anotadas en el *Assembly 9* porcino (*Sscrofa 9*) por Ramayo-Caldas *et al*. (2010, 2012), se actualizaron las posiciones al *Assembly* 10 (*Sscrofa10.2*) mediante la herramienta *Assembly Converter* de *Ensembl*.

Identificación de genes y función biológica: La anotación de los genes ubicados dentro de los CNVRs se realizó utilizando *BioMart* de *Ensembl* (ensembl.org/biomart) (Genes 78 Database, *Sscrofa10.2* Dataset). La anotación de la función biológica se realizó utilizando la base de datos de *Mouse Genome Informatics* (MGI, http://www.informatics.jax.org/batch).

Validación funcional de CNVRs: Se utilizaron muestras de DNA de 117 animales del BC1_LD (25% Ibérico, 75% Landrace) y 22 animales de distintas razas (6 Large White, 5 Landrace, 5 Duroc y 6 Ibéricos). Los *primers* se diseñaron con el software *Primer Express*® (Applied Biosystems), a partir de la secuencia porcina de referencia *Sscrofa 10.2*. Las reacciones de PCR se llevaron a cabo por triplicado utilizando el kit *SYBR*® *Select Master Mix* (Life Technologies) en un ABI PRISM 7900HT Sequence Detection System (Applied Biosystems). Para cuantificar y normalizar los datos de expresión se empleó el método $2^{-\Delta\Delta CT}$.

RESULTADOS Y DISCUSIÓN

Se detectaron un total de 1.423 CNVs incluyendo duplicaciones y deleciones en los 7 animales utilizados en el estudio. El número de CNVs detectados en cada una de las muestras varió entre 264 (Ibérico) y 121 (Landrace), con una media de 203 CNVs por muestra (Tabla 1). Con el objetivo de identificar CNVRs que difirieran entre las dos razas, se agruparon los CNVs en función de su frecuencia en la población. Utilizando como criterio una frecuencia en la raza Ibérica de 1 y de 0-0,4 en la raza Landrace se detectaron un total de 178 CNVRs; mientras que con una frecuencia en la raza Ibérica de 0 y de 1-0,4 en la raza Landrace se detectaron 304 CNVRs.

Para seleccionar los CNVRs con posible efecto sobre el fenotipo de los animales estudiados se anotaron los genes presentes en estos CNVRs y su posible función, identificándose un total de 326 genes. Asimismo, se llevó a cabo una selección de CNVRs teniendo en cuenta trabajos anteriores de nuestro grupo y/o identificados por otros autores mediante búsqueda bibliográfica.

De los 178 CNVRs identificados en Ibérico, 12 CNVRs fueron identificados previamente en el trabajo de Ramayo-Caldas *et al*. (2010) y de los 304 CNVRs identificados en Landrace, 16 coincidían. El número de CNVRs que concuerdan con los descritos por Fernández *et al*. (2014) fue menor, 7 CNVRs en Ibérico y 4 en Landrace.

También, se identificaron un total de 7 CNVRs en Ibérico y 5 CNVRs en Landrace que solapaban con regiones genómicas asociadas con la composición de ácidos grasos en músculo (Ramayo-Caldas *et al*., 2012).

Por último, se compararon los genes localizados en los CNVRs con los genes diferencialmente expresados en el transcriptoma de hígado (Ramayo-Caldas *et al*., 2012b), grasa dorsal (Corominas *et al*., 2013) y músculo (Puig-Oliveras *et al*. 2014) de animales con fenotipos extremos para la composición de ácidos grasos en músculo, identificándose un total de 7 genes en Ibérico y 4 genes en Landrace.

De los 178 CNVRs identificados en Ibérico, 38 CNVRs cumplían como mínimo con uno de los criterios anteriormente detallados, conteniendo un total de 21 genes. Mientras que para los 304 CNVRs identificados en Landrace, 45 CNVRs cumplían alguno de los criterios, identificándose 38 genes en estos CNVRs. Entre estos CNVRs se seleccionaron 5 para realizar su validación funcional mediante PCR cuantitativa en tiempo real. Estos CNVRs se encuentran distribuidos en diferentes regiones del genoma porcino. Los CNVRs seleccionados contienen genes funcionales como son *CYP4X1*, *KIT, MOGAT2* y *PRKG1*, y un *novel protein* cuyo ortólogo humano es *CLCA4*. A excepción del CNVR, localizado en un intrón del gen *CLCA4*, los demás CNVRs incluyen exones codificantes. Los genes *CLCA4* y *KIT* fueron previamente identificados en el trabajo realizado por Ramayo-Caldas *et al*. (2010). El uso de datos de NGS ha permitido obtener una mayor resolución en los límites de los CNVs en comparación con el uso de datos genotípicos de SNPs, como es el caso del CNVR del gen *KIT* en el SSC8, pasando de un tamaño de 413 Kb (Ramayo-Caldas *et al.*, 2010) a 283 Kb.

La validación funcional realizada para el CNVR del gen *KIT* en razas puras nos ha permitido testar la validez de la técnica utilizada para determinar el número de copias, observándose que los individuos Ibéricos y Duroc presentaban sólo una copia en cada cromosoma, mientras que otras razas como Large White y Landrace presentan gran variación (Tabla 2).

Se analizaran los demás CNVRs seleccionados así como la asociación de estos CNVRs con caracteres de crecimiento y composición de ácidos grasos. Los CNVRs detectados pueden ser útiles en futuras evaluaciones genómicas como marcador de caracteres de importancia económica.

REFERENCIAS BIBLIOGRÁFICAS

• Aaron *et al*., 2010. Bioinformatics. 26, 841-2. • Alkan *et al*., 2011. Nat. Rev. Genet. 12, 363-76. • Boeva *et al*., 2012. Bioinformatics. 28, 423-5. • Corominas *et al*., 2013. BMC Genomics. 14, 843. • Fernández *et al*., 2014. Anim. Genet. 45, 357-66. • Metzker, 2010. Nat. Rev. Genet. 11, 31-46. • Pérez-Enciso *et al*., 2000. J. Anim. Sci. 78, 2525-31. • Puig-Oliveras *et al*., 2014. PLoS One. 9, e99720. • Ramayo-Caldas *et al*., 2010. BMC Genomics. 11, 593. • Ramayo-Caldas *et al*., 2012. J. Anim. Sci. 90, 2883-93. • Ramayo-Caldas *et al*., 2012b. BMC Genomics. 13, 547. • Teo *et al*., 2012. Bioinformatics 28, 2711-8. • Zhao *et al*., 2013. BMC Bioinformatics. 11, S1.

Agradecimientos: Este trabajo ha sido financiado por el proyecto AGL2011-29821-C02 (Ministerio de Economía y Competitividad). M. Revilla ha sido financiado con una beca de Formació i Contractació de Personal Investigador Novell (FI-DGR) de la Generalitat de Catalunya (ECO/1639/2013). A. Puig-Oliveras ha sido financiada con una beca de la Universitat Autònoma de Barcelona (PIF, 458-01-1/2011). Agradecemos a J.L. Noguera (IRTA) su contribución en la obtención del material animal.

Tabla 1. *Estadísticos de los CNVs para las 7 muestras analizadas (Sscrofa10.2)*

Muestra	Raza	Número de CNVs			Longitud total media (kb)		
		Total	#duplicaciones	#deleciones	CNVs	Duplicaciones	Deleciones
I10	Ibérico	264	124	140	545,01	27,21	1003,64
I20	Ibérico	210	105	105	693,98	48,64	1339,32
IN9	Landrace	121	88	33	40,98	23,99	86,29
IN28	Landrace	225	131	94	25,02	23,28	27,44
IN39	Landrace	196	122	74	32,43	31,50	33,95
IN47	Landrace	222	132	90	22,00	17,88	28,05
IN51	Landrace	185	114	71	31,03	25,50	39,93
Media		203	117	87	198,64	28,29	365,52

Tabla 2. *Cuantificación relativa por qPCR del gen KIT y su CNV estimado tomando como referencia el Ibérico (raza que menos variación presenta entre las distintas réplicas técnicas*)

Gen	Nº de Animales				
	# 2 copias	# 3 copias	# 4 copias	# 5 copias	# 6 copias
Large White	-	-	1	2	3
Landrace*	-	1	2	1	1
Duroc	5	-	-	-	-
Ibérico*	6	-	-	-	-

*Ninguno de los individuos Ibérico o Landrace analizados pertenecía al proyecto IBMAP.

CNVs IDENTIFICATION AND VALIDATION IN SWINE

ABSTRACT: The objective of this study was to identify copy number variants (CNVs) associated to growth traits and fatty acid composition in swine. Based on *Control-FREEC* software, we detected CNVs using the WGS of 7 individuals (2 Iberian boars and 5 Landrace sows). A total of 1,423 CNVs were identified with an average size of 198.64 kb in all individuals. Among them, 178 CNVRs were identified using a frequency of 1 for Iberian samples and 0-0.4 for Landrace. Using a frequency of 1-0.4 for Landrace and 0 for Iberian, 304 CNVRs were evidenced. 38 and 45 CNVRs respectively, overlapped with at least one of the selection criteria used. 59 genes located in these regions were identified after gene annotation. Interestingly, some of these genes (*CLCA4, CYP4X1, KIT, MOGAT2* and *PRKG1*) have been previously associated with lipid metabolism. This study provided new insights into genomic structural variations that may be affecting economically relevant traits in pigs.

Keywords: CNV, Swine, *KIT*.