SELECCIÓN Y DETECCIÓN DE INDELS EN EL GENOMA PORCINO A PARTIR DE DATOS DE SECUENCIACIÓN PARALELA MASIVA

Crespo-Piazuelo, D.[1,2,*], Puig-Oliveras, A., Sambache, E., Criado-Mesas, L., Fernández, A.I., Ballester, M., Folch, J.M.

[1]Departamento de Ciencia Animal y de los Alimentos, Facultad de Veterinaria, Universidad Autónoma de Barcelona (UAB), Bellaterra, Barcelona.

[2]Centro de Investigación en Agrigenómica (CRAG), Consorcio CSIC-IRTA-UAB-UB, Edificio CRAG, Campus UAB, Bellaterra, Barcelona.

*daniel.crespo@cragenomica.es

INTRODUCCIÓN

La secuenciación paralela masiva de millones de lecturas o *Next Generation Sequencing* (NGS) ha permitido detectar de una forma más rápida las variaciones en el genoma de los seres vivos. Hasta la fecha, las variaciones que más se han estudiado mediante este método son los polimorfismos de un solo nucleótido (*Single Nucleotide Polymorphism*, SNP) las cuales representan alrededor del 80% sobre el total de las variaciones detectadas (Mullikin *et al.*, 2000; Dawson, 2001; Weber *et al.*, 2002).

Para encontrar el 20% restante de los polimorfismos, el presente trabajo se centra en la detección de las inserciones y deleciones (indels) en el genoma de cerdos Landrace e Ibérico a partir de datos de *Whole Genome Sequencing* (WGS) mediante herramientas específicas.

MATERIAL Y MÉTODOS

Material animal y secuenciación: Los animales secuenciados fueron los fundadores del cruce IBMAP, dos verracos Ibéricos de la estirpe Guadyerbas y cinco cerdas Landrace. Dos razas que difieren en gran medida en la calidad de la carne y la composición de la grasa intramuscular. La secuenciación se realizó en el CNAG (Centro Nacional de Análisis Genómico) mediante un equipo *HiSeq2000* (*Illumina*), con lecturas pareadas de una longitud de 100 pares de bases, las cuales se mapearon usando la herramienta *Burrows-Wheeler Aligner* (*BWA*) (Li y Durbin, 2009) contra el genoma de referencia porcino (Sscrofa10.2). Para cada muestra se obtuvieron cerca de 40 millones de lecturas pareadas con una cobertura media de 15x.

Además, se genotiparon 448 animales de distintos retrocruces (25% Ibérico y 75% Landrace, 25% Ibérico y 75% Duroc y, 25% Ibérico y 75% Pietrain) para una selección de indels, utilizando el sistema *Taqman OpenArrayTM* en un equipo *QuantStudioTM 12K flex Real-Time PCR System* (*ThermoFisher Scientific*).

Programas para la detección de indels: La detección de indels se realizó mediante tres programas distintos siguiendo las recomendaciones de Neuman *et al.* (2013): *Dindel* (Albers *et al.*, 2011), *SAMtools mpileup* (Li *et al.*, 2009) y *Genome Analysis Toolkit* (*GATK*) (McKenna *et al.*, 2010).

Predicción de los efectos de los indels: Se utilizó el programa *Variant Effect Predictor* (*VEP*) (McLaren *et al.*, 2010) de Ensembl para predecir los efectos y consecuencias de los indels. El programa *VEP* también permite predecir la magnitud del impacto que tendrían las variantes sobre el gen en el que estén presentes. Considerándose de gran impacto las

variaciones sobre la pauta de lectura, la ganancia o pérdida de las secuencias consenso de los espliceosomas, y la pérdida del inicio de traducción o del codón de terminación.

RESULTADOS Y DISCUSIÓN

Detección de indels: El programa que más indels detectó fue el *Dindel* con 3.380.221 indels detectados entre todos los individuos. *SAMtools mpileup* y *GATK* detectaron un número similar de indels (2.749.596 y 2.957.377, respectivamente).

Para aumentar la precisión de la detección sólo fueron seleccionados los indels que se encontraron compartidos entre los tres programas (1.928.746 indels, **Figura 1**). Además, debido a que la anotación de los alelos de referencia y alternativos del programa *SAMtools mpileup* es distinta al de los otros dos, fue necesario ajustarla para que coincidiera con ellos. De éstos, se filtraron para cada individuo y se eliminaron aquellos indels que no presentaban el mismo genotipo en al menos dos de los programas, resultando en una lista final de 1.878.218 indels detectados.

Para evitar los falsos positivos, estos indels fueron cribados descartando los que fueran trialélicos y los encontrados en la misma posición pero con un alelo alternativo distinto, debido a la interferencia de los microsatélites. También se eliminaron los indels que estaban fijados en todos los individuos analizados, ya que estos corresponden a variaciones con respecto a la secuencia de referencia, pero no segregan en el cruce IBMAP. En total se obtuvieron 1.631.044 indels, de los cuales un 52,9% fueron deleciones y el resto inserciones (47,1%).

Posteriormente, se utilizó el programa *VEP* para predecir los efectos y consecuencias de los 1.631.044 indels detectados, de los cuales 539.921 indels (un 33,1% del total) se localizaron en regiones génicas. A su vez, *VEP* predijo que 1.289 de estos indels tendrían un gran impacto sobre el gen en el que estaban presentes.

Selección de indels relevantes: La selección de los indels más relevantes se realizó eligiendo aquellos que estuvieran localizados en regiones génicas, segregando en la población IBMAP y presentes a frecuencias extremas entre los machos Ibéricos (IB) y las madres Landrace (LD): frecuencia IB = 0 y LD = 0,8-1 o frecuencia IB = 1 y LD = 0-0,2. Así pues, se encontraron 48.395 variantes presentes a estas frecuencias, la mayor parte de las cuales estaban localizadas en regiones intrónicas (68,9%). De las 15.053 variantes restantes que no estaban en intrones (31,1%), 132 variantes se localizaron en regiones codificantes.

Genotipado de los indels detectados y estudios de asociación: Para realizar un estudio más amplio de los indels detectados en nuestra población y analizar su posible efecto sobre caracteres de crecimiento y de composición de ácidos grasos, se realizará el genotipado de una selección de 12 indels sobre el total de los considerados relevantes segregando en la población IBMAP (los 132 situados en regiones codificantes y a frecuencias extremas y los 1.289 con un potencial gran impacto sobre el gen). Por el momento, se ha genotipado en 448 animales pertenecientes a tres fondos genéticos diferentes (Ibérico x Landrace, Ibérico x Duroc e Ibérico x Pietrain) un indel localizado en la región promotora del gen *upstream transcription factor 1* (*USF1*) que codifica para un factor de transcripción relacionado con el metabolismo lipídico.

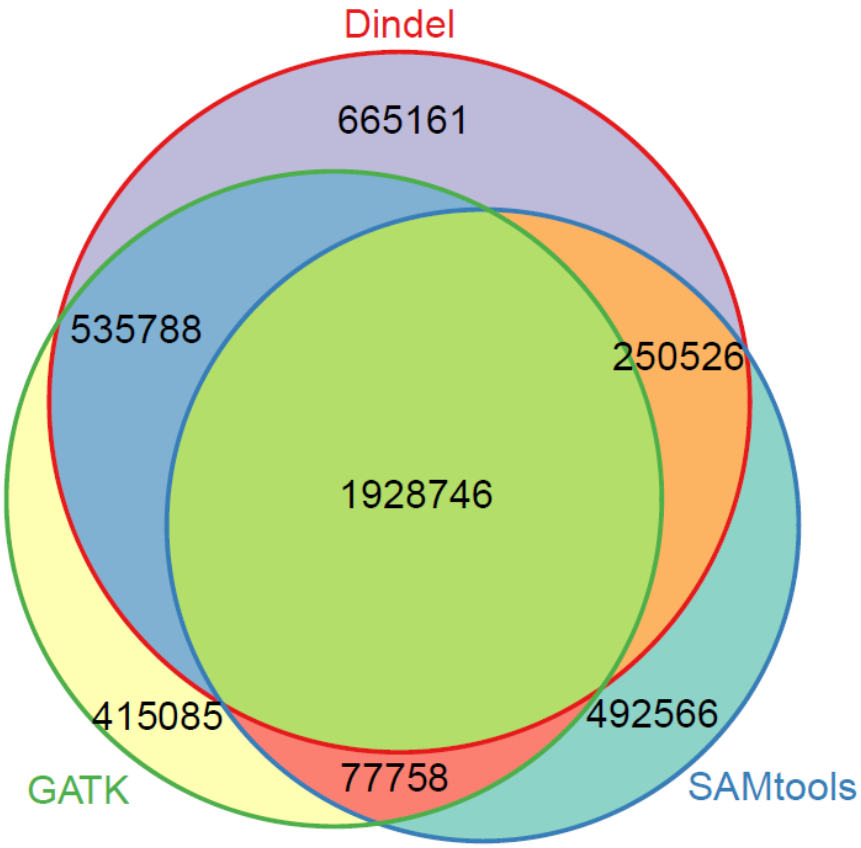

Figura 1. Diagrama de *Venn* de los indels identificados por *Dindel*, *GATK* y *SAMtools mpileup*.

REFERENCIAS BIBLIOGRÁFICAS

• Albers, C.A. *et al.*, 2011. Genome Research 21(6): 961–73. • Dawson, E., 2001. Genome Research 11(1): 170–78. • Li, H. y Durbin, R., 2009. Bioinformatics 25(14): 1754–60. • Li, H. *et al.*, 2009. Bioinformatics 25(16): 2078–79. • McKenna, A. *et al.*, 2010. Genome Research 20(9): 1297–1303. • McLaren, W. *et al.*, 2010. Bioinformatics 26(16): 2069–70. • Mullikin, J.C. *et al.*, 2000. Nature 407(6803): 516–20. • Neuman, J.A. *et al.*, 2013. Briefings in Bioinformatics 14(1): 46–55. • Weber, J.L. *et al.*, 2002. The American Journal of Human Genetics 71(4): 854–62.

Agradecimientos: Este trabajo ha sido financiado por el proyecto AGL2014-56369-C2 (Ministerio de Economía y Competitividad). D. Crespo-Piazuelo ha sido financiado con una beca de Formació i Contractació de Personal Investigador Novell (FI-DGR) de la Generalitat de Catalunya (ECO/1788/2014).