DETECCIÓN DE VARIANTES Y CONCORDANCIA DE GENOTIPOS MEDIANTE SECUENCIACIÓN A COBERTURAS MODERADAS

Ros-Freixedes, R., González-Prendes, R., Gol, S., Solé, E., Pena, R.N. y Estany, J.
Departament de Ciència Animal, Universitat de Lleida - Agrotecnio Center, 25198 Lleida;
rros@ca.udl.cat

INTRODUCCIÓN

La secuenciación de nueva generación ofrece nuevas oportunidades para la identificación de variantes asociadas con la variabilidad genética de caracteres de interés tanto en especies ganaderas como en cultivos (Daetwyler et al., 2014; Nicod et al., 2016; Yano et al., 2016; Schaid et al., 2018). El coste actual de secuenciación hace viable usar tecnología de alto rendimiento para generar grandes cantidades de datos de secuenciación. A diferencia de los chips de genotipado, que permiten genotipar por un conjunto preestablecido de polimorfimos de un solo nucleótido (SNP), las secuencias genómicas contienen las variantes causales de la variación genética de los caracteres, incluyendo variantes con frecuencias bajas y específicas de una población, y otros tipos de variación como inserciones y deleciones cortas (indels) (Das et al., 2015; Gudbjartsson et al., 2015). Estas propiedades hacen que las técnicas de secuenciación se presenten como herramientas muy potentes para desentrañar el control genético de procesos biológicos complejos.

Sin embargo, hasta la fecha la mayoría de variantes descritas en relación a caracteres de interés productivo no han sido debidamente validadas y explican solo una parte modesta de la varianza genética. Nuestro grupo ha descrito varias variantes genéticas asociadas con caracteres de calidad de la carne y metabolismo lipídico, identificadas mediante análisis de asociación del genoma completo basados en chips de genotipado de alta densidad (p. ej., Ros-Freixedes et al., 2016) o mediante estudios de genes candidato genotipados mediante PCR (p. ej., Gol et al., 2018). La secuenciación podría incrementar la eficiencia en identificar variantes asociadas con dichos caracteres.

Para que el uso de la secuenciación sea efectivo y viable económicamente hace falta encontrar estrategias de secuenciación que equilibren la detección de variantes y la precisión del genotipado en estas posiciones. El objetivo de este estudio fue validar el uso de secuenciación de nueva generación a coberturas moderadas (~5x). Por un lado estimamos la capacidad de detección de variantes. Por el otro evaluamos la concordancia de genotipos y alelos entre secuenciación y otras tecnologías de genotipado.

MATERIAL Y MÉTODOS

Para este estudio usamos 40 cerdos de raza Duroc, para los cuales se generaron datos de secuenciación, datos de genotipado con chips de alta densidad, y datos de genotipado mediante PCR-RFLP o PCR-HRM, según el marcador, para un panel de 11 marcadores.

El ADN genómico fue secuenciado con una cobertura de 6,8x (DE=1,2x; mín=4,4x; máx=12,2x). Se prepararon librerías que fueron secuenciadas con la técnica de extremos pareados (*paired-end*) usando un instrumento NovaSeq 6000 (Illumina, San Diego, CA) en el CNAG (Barcelona). Las lecturas generadas fueron alineadas contra el genoma de referencia *Sscrofa11.1* (GenBank: GCA_000003025.6) usando el algoritmo BWA-MEM (Li, 2013) y las variantes fueron detectadas mediante GATK HaplotypeCaller 3.8.0 (DePristo et al., 2011; Poplin et al., 2018). El número de lecturas que contenían el alelo de referencia (nRef) y el alelo alternativo (nAlt) se extrajeron usando una función de apilamiento (*pile-up*) siguiendo las recomendaciones de Ros-Freixedes et al. (2018) para evitar sesgos debidos a la baja cobertura. El genotipo más probable (0, 1, o 2, donde 0 es el homocigoto para el alelo de referencia, 1 el heterocigoto, y 2 el homocigoto para el alelo alternativo) fue calculado usando nRef y nAlt como:

$$P(0)=(1-e)^{nRef}\cdot e^{nAlt},\ P(1)=0{,}5^{nRef}\cdot 0{,}5^{nAlt},\ \text{y}\ P(2)=e^{nRef}\cdot(1-e)^{nAlt},$$

donde ‘e’ es el índice de error (asumido como 0,001), y con las probabilidades ajustadas debidamente para que P(0)+P(1)+P(2)=1.

Los mismos 40 cerdos fueron genotipados con el chip GGP-Porcine HD BeadChip (~70k SNP; GeneSeek, Lincoln, NE). Un total de 47.222 SNP mostraron segregación en la muestra. Finalmente, los cerdos también fueron genotipados para un panel de 11 marcadores, escogidos por su asociación con caracteres productivos, mediante protocolos de PCR-RFLP (A1 a A4) o PCR-HRM (B5 a B11).

Se analizó la capacidad de detección de las variantes del chip y del panel de marcadores mediante secuenciación, así como la concordancia de genotipos y alelos obtenidos por secuenciación con el chip y los genotipados por PCR, asumiendo, en principio, que estos últimos representan los verdaderos genotipos.

RESULTADOS Y DISCUSIÓN

Un total de 20.381.114 variantes fueron detectadas en los 40 cerdos secuenciados, de las cuales 15.941.676 eran SNP bialélicos y 3.774.997 eran indels bialélicos.

El 96% de los SNP del chip que segregaban en estos 40 cerdos fueron identificados con éxito con los datos de secuenciación (45.365 de 47.222). La Figura 1 muestra la concordancia de genotipos entre secuenciación y el chip de genotipado para estos SNP. La concordancia entre los genotipos asignados con el chip y a partir de los datos de secuenciación fue muy alta, con un promedio de 0,96 (DE=0,05). En general, el 98% de los SNP tenían una concordancia de al menos 0,85 (44.259), el 95% de al menos 0,90 (43.253) y el 81% de al menos 0,95 (36.901). Una fracción pequeña de SNP presentó concordancias muy inferiores a estos valores, lo cual sugiere posibles errores de mapeado para estos SNP o incluso regiones genómicas de alta complejidad que dificultan el alineamiento de lecturas.

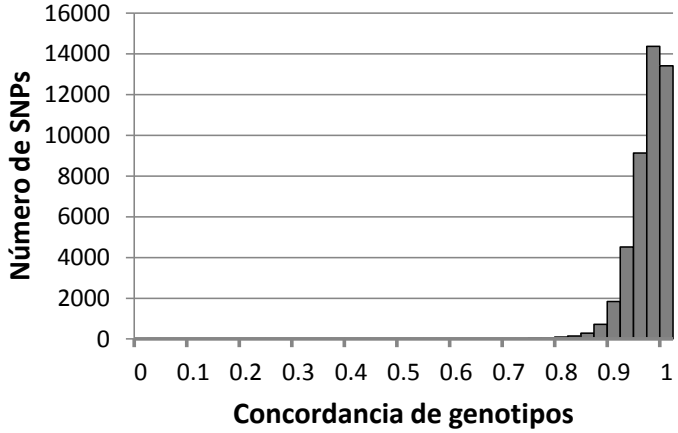

***Figura 1**. Concordancia de genotipos entre secuenciación y chip de genotipado.*

Con el panel de marcadores de interés se obtuvieron resultados similares. La Tabla 1 muestra la cobertura de secuenciación obtenida para cada uno de los marcadores considerados, si fue detectado como variante y las concordancias de genotipo y alelo obtenidas. De los 11 marcadores que conformaban el panel, 9 fueron identificados con éxito. En los 2 marcadores que no fueron identificados (A3 y B9), la cobertura conseguida (7,4x o solo 2,0x) fue menor que en los otros 9 (8,7x; mín=7,8x; máx=9,8x), aparentemente debido a un sesgo en el alineamiento en contra del alelo alternativo. Esto sugiere que los marcadores A3 y B9 podrían estar en regiones muy polimórficas o con estructuras complejas.

En consonancia, la concordancia de alelos fue también muy alta. La mayor parte de disconcordancias entre genotipos fueron confusiones entre un homocigoto y el heterocigoto, y muy raramente entre homocigotos opuestos. A efectos de estudios de asociaciones marcador-carácter, los errores entre homocigoto y heterocigoto son mucho más leves que entre homocigotos opuestos. Aun así, la menor concordancia para el marcador B7 sugiere un posible error en el genotipado o en el mapeado.

Es importante notar que, dejando de lado las anomalías de los marcadores A3, B7 y B9, las concordancias fueron consistentemente menores en los marcadores genotipados mediante PCR-RFLP que mediante PCR-HRM. Esto indica que la menor concordancia para los marcadores A1, A2 y A4 no se debe a errores de la secuenciación sino del genotipado por PCR-RFLP, que es un método menos sensible que PCR-HRM.

Los resultados obtenidos indican que una cobertura moderada (~5x) provee un buen balance entre detección de variantes y el grado de certidumbre de los genotipos para las variantes detectadas. Coberturas menores permiten secuenciar un mayor número de individuos, lo cual aumenta la capacidad de detectar variantes con frecuencias bajas (Le y Durbin, 2011). Sin embargo, una cobertura demasiado baja reduce la capacidad de detectar variantes a nivel de

individuo, aunque no necesariamente a nivel de población, y aumenta la incertidumbre de genotipado en dichas posiciones. Una cobertura de 5x parece proveer buenas concordancias entre los genotipos obtenidos por secuenciación o por otras tecnologías, no muy distintas de las obtenidas con 6-10x (Ros-Freixedes et al., 2018).
En conclusión, este estudio demuestra la elevada capacidad de detección de variantes de secuenciación masiva a ~5x y la buena concordancia entre tres métodos de asignación de alelos (secuenciación masiva, chips de SNP y PCR-RFLP/PCR-HRM). La secuenciación resulta una herramiento prometedora para la identificación de variantes asociadas con caracteres de interés.

***Tabla 1**. Cobertura de secuenciación y concordancia de genotipos y alelos para los once marcadores de interés por su asociación con caracteres productivos*

Marcador[a]	nRef[b]	nAlt[b]	Cobertura media	Detectado	Concordancia de genotipos	Concordancia de alelos
A1	222	122	8,6x	Sí	0,89	0,95
A2	268	78	8,7x	Sí	0,89	0,95
A3	292	5	7,4x	No	0,50	0,58
A4	304	47	8,8x	Sí	0,88	0,94
B5	100	233	8,3x	Sí	0,98	0,99
B6	97	215	7,8x	Sí	0,95	0,98
B7	160	205	9,1x	Sí	0,58	0,78
B8	274	92	9,2x	Sí	0,95	0,98
B9	79	0	2,0x	No	0,34	0,63
B10	176	165	8,5x	Sí	0,98	0,99
B11	257	134	9,8x	Sí	1,00	1,00

[a]A: marcador genotipado mediante protocolo de PCR-RFLP; B: de PCR-HRM.
[b]nRef: número de lecturas con el alelo de referencia; nAlt: con el alelo alternativo.

REFERENCIAS BIBLIOGRÁFICAS

• Daetwyler, H.D. et al. 2014. Nat. Genet. 46: 858-865. • Das, A. et al. 2015. BMC Genomics 16: 1043. • DePristo, M.A. et al. 2011. Nat. Genet. 43: 491-498. • Gol, S. et al. 2018. Sci. Rep-UK 8: 14336. • Gudbjartsson, D.F. et al. 2015. Nat. Genet. 47: 435-444. • Le, S.Q. & Durbin, R. 2011. Genome Res. 21: 952-960. • Li, H. 2013. arXiv: 1303.3997. • Nicod, J. et al. 2016. Nat. Genet. 48: 912-918. • Poplin, R. 2018. bioRxiv: 10.1101/201178. • Ros-Freixedes, R. et al. 2016. PLoS One 11: e0152496. • Ros-Freixedes, R. et al. 2018. Genet. Sel. Evol. 50: 64. • Schaid, D.J. et al. 2018. Nat. Rev. Genet. 19: 491-504. • Yano, K. et al. 2016. Nat. Genet. 48: 927-934.

Agradecimientos: Proyecto financiado por el MINECO y fondos FEDER (AGL2015-65846-R). E. Solé es beneficiaria de una beca doctoral de la Universitat de Lleida.

VARIANT DISCOVERY RATE AND GENOTYPE CONCORDANCE OF SEQUENCING AT A MODERATE COVERAGE

ABSTRACT: Next-generation sequencing is a promising powerful tool for the discovery of variants associated with traits of interest in livestock and crops. For that purpose, generally sequencing strategies that balance high discovery rate with high genotyping accuracy are required, which could be achieved with moderate sequencing coverages. The objective of this study was to evaluate variant discovery rate and genotype and allele concordance between next-generation sequencing at ~5x, genotyping marker arrays, and genotyping by PCR-RFLP or PCR-HRM. Around 96% of the variants from a marker array were successfully discovered, with very high genotype and allele concordances. Genotype and allele concordances with markers genotyped by PCR were also high and indicated that sequencing was more accurate than PCR-RFLP. In conclusion, sequencing at a moderate coverage (~5x) provided a suitable balance between variant discovery and genotyping accuracy.

Keywords: sequencing, variants, discovery rate, genotype concordance