Sobre los Errores de Secuenciación en la Estimación de ROHs

R. Peiró-Pastor[1], W. M. Rauw[1], O. Torres[1], M. García-Gil[1], L. Gómez-Raya[1]*

[1] Departamento de Mejora Genética Animal, INIA-CSIC, Ctra de la Coruña km 7.5, 28040, Madrid
** Corresponding autor: ramon.peiro@inia.csic.es*

Resumen

Un error de secuenciación ocurre cuando al ensamblar una secuencia se añade una base no real en vez de la base real. Los errores de secuenciación se estiman entre el 1 y el 0.1% del total, lo cual significa entre 1.2 y 12 millones de errores de secuenciación en el genoma de la gallina con 1.2 Gb. El gran tamaño de los genomas implica que los errores de secuenciación pueden tener un gran impacto en la estimación de la consanguinidad usando ROHs (*Runs Of Homozygosity*), particularmente en los de gran tamaño. Algunos programas como PLINK, permiten añadir un número de heterocigotos fijo cuando se buscan e identifican ROHs (Purcell *et al*., 2007). Sin embargo, el número esperado de heterocigotos depende del tamaño del ROH. Un ROH de tamaño grande tendrá un mayor número de errores de secuenciación identificando heterocigotos en vez de homocigotos. En este estudio modelamos la ocurrencia y detección de errores de secuenciación utilizando un método secuencial en el que se asume que los errores son independientes y que se distribuyen en el ROH como una binomial con probabilidad de heterocigoto = error asumido de secuenciación. Cada posición del SNP se añade a las anteriores y se calculan heterocigotos en el fragmento. A continuación, se realiza un test similar al LOD *score* en el análisis de ligamiento. El test es el logaritmo en base 10 de la razón de probabilidades de obtener el número esperado de heterocigotos dependiendo del tamaño del fragmento y del número de heterocigotos observado en el fragmento en cuestión. Para fragmentos de más de 100 SNPs, se utilizan las aproximaciones de Poisson y de Stirling. El algoritmo termina cuando el LOD > 3, es decir, es 1000 veces más probable obtener el número de heterocigotos esperados (a ese tamaño de ROH) que el del observado en la realidad. El procedimiento se inicia de nuevo para identificar el siguiente ROH en el mismo cromosoma. Este algoritmo se ha aplicado a las secuencias del cromosoma 1 de 20 gallinas de la raza Castellana Negra. La Figura 1 muestra como el número de heterocigotos cambia con el tamaño del ROH llegando a más de 120 heterocigotos permitidos en ROH de tamaño grande después de asumir un error de secuenciación de 0.1%. La Tabla 1 muestra como asumiendo errores de secuenciación de 0.01 los coeficientes de consanguinidad se incrementan en 3 a 4 unidades porcentuales. Este método permite investigar como es la distribución de errores de secuenciación (de homocigoto a heterocigoto) en ROHs de tamaño grande en el genoma.

Keywords: Error de Secuenciación; Next Generation Sequencing, Runs of Homozygosity, Gallinas, Castellana Negra

References

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D.,et al. (2007). PLINK: a tool set for whole-genome association and population based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi: 10.1086/519795

Figure 1. Relación entre el número de heterocigotos permitido y el tamaño del ROH cuando se asume un error de secuenciación de 0.1%.

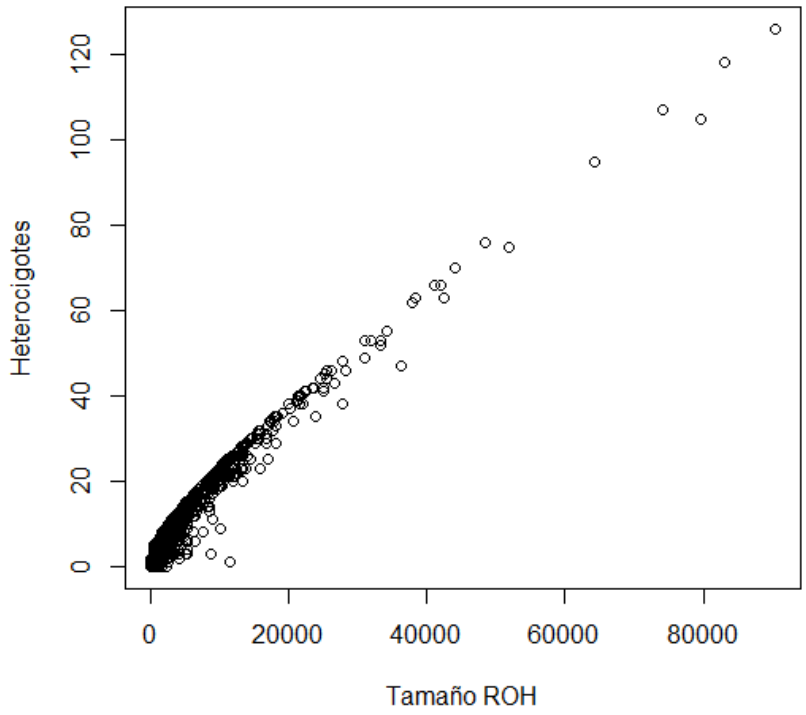

Tabla 1. Coeficiente de consanguinidad para el cromosoma 1 en 20 individuos asumiendo errores de secuenciación de 1 y 0.1%.

	Error de Secuenciación			Error de Secuenciación	
Ind	0.01	0.001	Ind	0.01	0.001
1	0.44	0.41	16	0.48	0.45
4	0.44	0.41	17	0.33	0.30
5	0.36	0.32	18	0.38	0.34
7	0.46	0.42	19	0.43	0.39
9	0.41	0.38	20	0.35	0.31
10	0.38	0.35	21	0.39	0.36
12	0.42	0.38	22	0.45	0.41
13	0.37	0.33	23	0.45	0.42
14	0.35	0.32	24	0.47	0.44
15	0.54	0.51	25	0.47	0.44