SNP+: PROGRAMA PARA CALCULAR LA PROBABILIDAD DE *DROPOUT* EN SNPs

Sastre[1], N., Mercadé[1], A., Ramírez[2], O., Sánchez[1,2], A., Francino[1,2], O., Casellas[3], J.
[1]Servei Veterinari de Genètica Molecular, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. [2]Vetgenomics, Parc de Recerca UAB Edifici Eureka, 08193 Bellaterra, Spain. [3]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.
natalia.sastre@uab.cat; joaquim.casellas@uab.cat

INTRODUCCIÓN

Los *Single Nucleotide Polymorphisms*, polimorfismos de base única o SNPs son mutaciones puntuales que implican la substitución de un nucleótido por otro. Son marcadores bialélicos muy abundantes en el genoma pero con una tasa de mutación baja (10^{-9} por generación) (Brumfield et al. 2003; Morin et al. 2004). Son por lo tanto menos polimórficos que los marcadores de tipo microsatélite, siendo la ratio SNP - con una MAF (Frecuencia del Alelo Menor) > 20% - *versus* microsatélite de aproximadamente 4:1. Los SNPs se localizan en regiones no codificantes y codificantes y su uso es muy variado. Por ejemplo, en medicina, la identificación de SNPs suele ir asociada a la búsqueda de alguna enfermedad o susceptibilidad a factores ambientales (Erichsen & Chanock 2004; Amos et al. 2008; Nickels et al. 2013), en agronomía, con la búsqueda de QTLs (*quantitative trait locus*) (Casellas et al. 2008), y en medicina forense con la identificación de individuos a partir de muestras de ADN de baja calidad (Sobrino et al. 2005). La identificación individual o genotipado mediante muestras forenses o no invasivas (pelos, orina, heces, saliva,..) puede resultar infructuosa debido a los errores de genotipado que se producen cuando las muestras tienen menos de 100 copias de gDNA. En estos casos, los efectos estocásticos pueden dominar la reacción y producir *dropout*: la no amplificación de uno de los alelos y por lo tanto, considerar un individuo homocigoto y no heterocigoto para aquel SNP. Existen diferentes soluciones para disminuir el porcentaje de error causado por *dropout* como por ejemplo preamplificar los SNPs de interés y/o realizar réplicas por cada muestra. Aquí presentamos un programa (SNP+) que calcula la probabilidad de *dropout* y el número de réplicas necesario para alcanzar una fiabilidad del 95% al genotipar un individuo. Nuestro programa calcula la probabilidad de error por *dropout,* compara los dos alelos mediante un factor de Bayes, y determina el número de réplicas necesario para cada SNP de un *array*, pudiendo usarse para seleccionar los mejores SNPs sea cual sea la plataforma, desde un único SNP, *arrays* de baja densidad (https://www.thermofisher.com/es/es/home.html) hasta *arrays* de alta densidad con más de 100,000 SNPs (http://www.illumina.com; http://www.affymetrix.com).

MATERIAL Y MÉTODOS

El programa SNP+ analiza cada SNP de forma individual, tomando como punto de partida un vector $\mathbf{y}$ de genotipos ordenado por individuo (m) y réplicas dentro de individuo ($\mathbf{y}' = [\mathbf{y}'_1\ \mathbf{y}'_2\ ...\ \mathbf{y}'_m]$), donde n_1 es el número de réplicas para el primer individuo y $n = n_1 + n_2 + ... + n_m$. Asumiendo que los alelos son A y B, el análisis Bayesiano toma como punto de partida

$$p(f_A,\varepsilon_A,\varepsilon_B|\mathbf{y}) \sim p(\mathbf{y}|f_A,\varepsilon_A,\varepsilon_B)\ p(f_A)\ p(\varepsilon_A)\ p(\varepsilon_B),$$

que se centra en estimar la frecuencia alélica (f_A), así como la probabilidad de que falle el genotipado del alelo A (ε_A) o el B (ε_B). Tomando un genotipado concreto y_i = AA, su verosimilitud Bayesiana se computa como p(AA|AA)p(AA) + p(AA|AB)p(AB), donde p(AA|AB) es la probabilidad de genotipar como homozigoto AA un individuo en realidad heterozigoto, y p(AB) es la probabilidad de que el genotipo real sea heterozigoto. Adicionalmente, el modelo considera que un BB nunca podrá ser genotipado como AA, es decir la posibilidad de falsos alelos es nula. Se consideran distribuciones *a priori* planas para los demás parámetros.

El modelo se aborda mediante un proceso de muestreo de Metropolis-Hastings (Metropolis et al., 1953), y se comparan dos parametrizaciones alternativas ($\varepsilon_A = \varepsilon_B$ vs. $\varepsilon_A \neq \varepsilon_B$) mediante un factor de Bayes. El número de réplicas necesario para alcanzar una fiabilidad del 95% se calcula como log(0,05)/log(ε_A). Todos estos procedimientos se han implementado en el software SNP+, disponible en http://www.casellas.info/software.htlm.

El programa genera los siguientes documentos de salida:
1) Tabla en formato texto de la probabilidad de error, intervalo de confianza, réplicas, factor de Bayes y genotipo de cada individuo para cada SNP.
2) Tabla resumen en formato texto de la probabilidad de error, intervalo de confianza, réplicas y factor de Bayes de todos los SNPs (Figura 1).
3) Tabla en formato csv del genotipo de cada individuo y la probabilidad de error si la hubiera.
4) Tabla en formato texto de la probabilidad de identidad y número de SNPs coincidentes entre dos individuos.

				PROBABILITY (UNGENOTYPED ALLELE)						
	ALLELE			POSTERIOR MEAN		CREDIBILITY INTERVAL (95%)				
SNP	A	B	PARAMETER	Prob.	n(95%)	Probability			n(95%)			BF(2 vs. 1)
1	T	C		* * * not estimable * * *					n.e.			
			P(B=KO)	0.03612	0.90	0.02216 to 0.05305			0.79	to	1.02	
2	T	C	P(A=KO)	0.08160	1.20	0.05073 to 0.11795			1.00	to	1.40	46.608
			P(B=KO)	0.06973	1.12	0.04768 to 0.09576			0.98	to	1.28	
3	C	T	P(X=KO)	0.05120	1.01	0.03585 to 0.06894			0.90	to	1.12	0.274
4	C	T	P(A=KO)	0.02360	0.80	0.00327 to 0.06386			0.52	to	1.09	3.649
			P(B=KO)	0.03501	0.89	0.02162 to 0.05099			0.78	to	1.01	
5	T	C	P(X=KO)	0.02975	0.85	0.01781 to 0.04444			0.74	to	0.96	0.179
6	T	C	P(A=KO)	0.04079	0.94	0.02480 to 0.06088			0.81	to	1.07	2.126
			P(B=KO)	0.05174	1.01	0.03065 to 0.07903			0.86	to	1.18	

Figura 1*. Ejemplo del documento de salida que genera el programa SNP+ (se muestran únicamente los 5 primero SNPs de un array). En el primer caso no se producen dropouts para el alelo A y por lo tanto no es estimable su probabilidad de error ni el factor de Bayes (BF). En el tercer caso, la probabilidad de error de A y B es la misma (BF<1); n: número de réplicas.*

Como ejemplo hemos utilizado SNP+ para calcular las probabilidades de *dropout*, el número de réplicas con una fiabilidad del 95%, y el factor de Bayes para cada SNP en los siguientes cuatro casos reales usando tecnología Open Array: A) 32 SNPs, especie *Homo sapiens* (humano), muestras de saliva (n= 269); B) 192 SNPs, especie *Canis lupus* (lobo), muestras forenses (n=66); 128 SNPs, especie *Ursus arctos* (oso), muestras de C) heces (n=22) y D) pelos (n=114). No se han incluido en el análisis muestras de baja calidad de DNA, es decir, con un *call rate* < 25%.

RESULTADOS Y DISCUSIÓN

La Figura 2 muestra las frecuencias relativas de la probabilidad de *dropout* para los distintos SNPs en los 4 casos estudiados.
La probabilidad de *dropout* fue baja para la mayoría de los casos siendo el valor generalmente < 0,3. En términos de variabilidad y moda de la distribución, el ensayo que obtiene probabilidades de *dropou*t menores es el caso A, probablemente porque las muestras son de saliva y la calidad y/o cantidad del DNA es mayor que en los casos B, C y D donde las muestras son mayoritariamente de pelos y heces. El ensayo con mayor probabilidad de *dropout* es el caso D que corresponde a *Ursus arctos*, concretamente de pelos. Este resultado sorprende si se tiene en cuenta que la calidad y/o cantidad de DNA en pelos es superior que aquel hallado en heces. El hecho de no incluir en el análisis las muestras de heces con un *call rate* < 25% (33%) disminuye la probabilidad de *dropout* y por lo tanto el error de genotipado. El porcentaje de muestras de pelo con un *call rate* < 25%, es decir no incluidas en el análisis, fue claramente inferior (15%).

SNP+ puede ser utilizado para diseñar ensayos o paneles evitando aquellos SNPs que requieran un gran número de réplicas porque conducen a error o bien, una vez diseñado el panel, realizar un número x de réplicas para dar un correcto genotipo cuando se trabaje con muestras no invasivas.

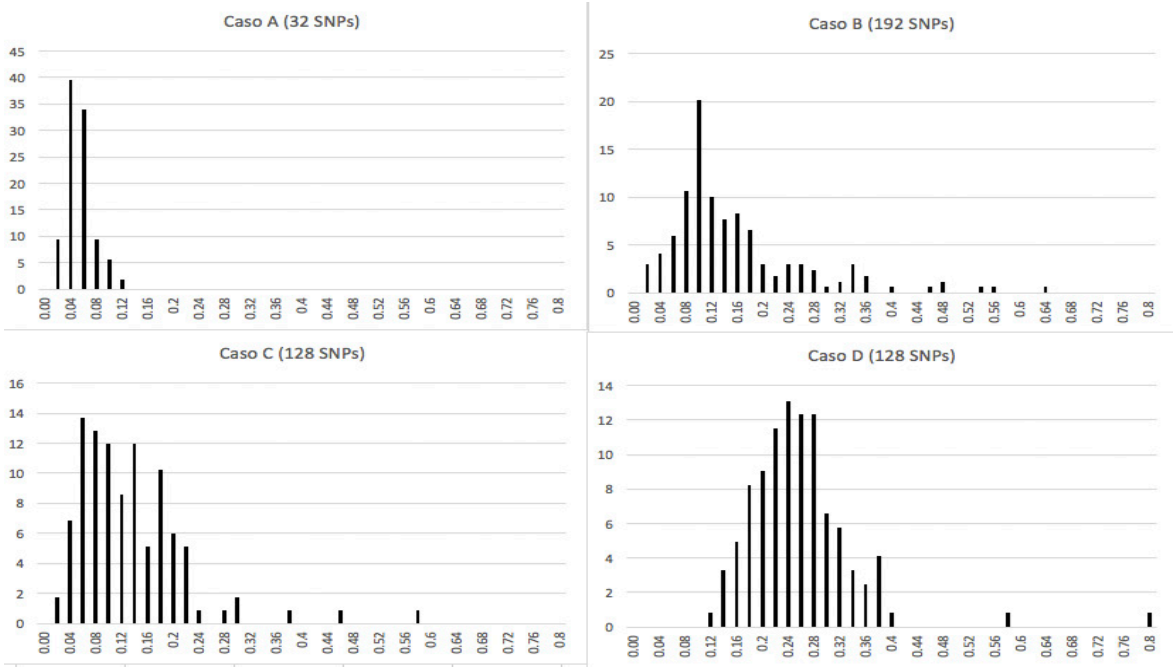

***Figura 2**. Histogramas que muestran las frecuencias relativas (%) de la probabilidad de dropout en los 4 casos analizados con tecnología Open Array: A) 32 SNPs, especie Homo sapiens, muestras de saliva; B)192 SNPs, especie Canis lupus, muestras forenses; 128 SNPs, especie Ursus arctos, muestras de C) heces y D) pelos.*

REFERENCIAS BIBLIOGRÁFICAS

Brumfield, R.T. et al. 2003 Ecology and Evolution, 18: 249-256 • Morin P.A. et al. 2004 Trends in Ecology and Evolution, 19(4): 208-216 • Erichsen H.C. & Chanock S.J. 2004 Br J Cancer 90:747–751 • Amos C.I. et al. 2008 Nat. Genet. 40:616–622 • Nickels S. et al. 2013 PLOS Genetics 9:e1003284 • Casellas J. et al. 2008 Animal 2: 177-183 • Sobrino B. et al. 2005 Forensic Sci Int 154:181–194 • Metropolis, N. et al. 1953. J. Chem. Phys. 21:1087-1092.

SNP+: SOFTWARE TO CALCULATE ALLELIC DROPOUT LIKELIHOOD IN SNPs

ABSTRACT: Genotyping individuals using forensic or non-invasive samples such as hair, faecal and saliva samples increases the risk of allelic amplification failure (dropout) due to the low quality and quantity of DNA. One way to decrease the rate of allelic dropout is to increase the number of replicates per sample. Here, we have developed a software (SNP+) to calculate dropout likelihood, number of replicates and Bayes factor per SNP in order to decrease genotyping errors. Our software can be used to select SNPs from a single SNP array to high density arrays with more than 100,000 SNPs.

Keywords*:* SNP, software, dropout, forensic samples