ESTIMACIÓN DE FRECUENCIAS ALÉLICAS EN *POOLS* DE ADN MEDIANTE EL PROCESAMIENTO DE SEÑALES DE ELECTROFORESIS CAPILAR

M.L. Checa[1], C. Carleos[2], J. A. Baro[3], S. Dunner[1], J. Cañón[1]

[1] Laboratorio de Genética. Dpto. de Producción Animal. Universidad Complutense de Madrid.
[2] Departamento de Estadística e Investigación Operativa. Universidad de Oviedo
[3] SERIDA. Principado de Asturias.

INTRODUCCIÓN

El genotipado a partir de mezclas o *pools* de ADN tiene diversas aplicaciones, entre las que se encuentran la búsqueda de loci implicados en el desarrollo de enfermedades (Kirov et al. 2000; Daniels et al. 1998; Barcellos et al. 1997; Shaw et al. 1998), la búsqueda de loci asociados a caracteres cuantitativos de importancia económica o QTLs (Lipkin et al. 1998).

La amplificación de microsatélites en *pools* de ADN presenta como ventaja más evidente, frente al esfuerzo de genotipados individuales, la reducción del coste de genotipado como consecuencia de una menor necesidad de reactivos y de mano de obra.

A la hora de trabajar con microsatélites amplificados a partir de *pools* de ADN surgen problemas específicos. Por un lado será necesario una cuantificación precisa del ADN de cada uno de los individuos que van a componer el pool, medir la sensibilidad de la PCR para reflejar cuantitativamente las diferencias de amplificación que responden a frecuencias alélicas diferentes, medir la sensibilidad del método de cuantificación del producto de la PCR después de la electroforesis (radioactividad, quimioluminiscencia, fluorescencia…), y la repetibilidad de la técnica de elección. Además, el número de individuos presentes en el pool será una de las decisiones importantes si observamos que tiene una significativa influencia sobre el error de la estimación de las frecuencias alélicas.

Normalmente en los microsatélites no amplifica una única banda por alelo; a la banda principal le acompañan otras secundarias conocidas como “bandas sombra”, que son el resultado de la amplificación simultánea de fragmentos entre uno y tres motivos más cortos que el alelo verdadero, consecuencia del fenómeno conocido como *slippage* (deslizamiento) de la polimerasa. Debido a estas amplificaciones accesorias, en los *pools* de ADN se produce un solapamiento de bandas principales y bandas sombra que dificulta la estimación de las frecuencias alélicas. Además, durante la PCR se producen otros dos artefactos: la adición por parte de la polimerasa de una adenina en el extremo 3’ de la hebra y la amplificación diferencial que favorece a los alelos más cortos, factores éstos que también contribuyen a dificultar la estimación correcta de las frecuencias alélicas.

El objetivo del presente estudio es la formalización de las manipulaciones de los resultados de la electroforesis capilar consecuencia de la amplificación por PCR de los alelos de un microsatélite presentes en un pool de ADN con el fin de estimar las frecuencias alélicas relativas con una precisión suficiente para ser utilizadas en trabajos de localización de QTLs.

MATERIAL Y MÉTODOS

Extracción y cuantificación de ADN

Se extrajo el ADN de cincuenta muestras de sangre de bovino siguiendo un método de extracción fenol-cloroformo, se estimó su concentración mediante fluorímetría (TD-360 Mini-Fluorometer, Turner Designs) y se procedió a la

amplificación individual de los dos microsatélites seleccionados con el fin de determinar el comportamiento de cada uno respecto a las bandas sombra, adición de la adenina y amplificación.

Una vez conocido el genotipo de cada uno de los animales se crearon *pools* de ADN de tres a cincuenta animales atendiendo a sus frecuencias alélicas para estudiar la relación entre el número de animales de un *pool* y el error en la estima de las frecuencias alélicas. Se hicieron dos réplicas de cada pool para estimar el error en la cuantificación de ADN. Una de las réplicas se amplificó dos veces para valorar la variabilidad de resultados debida a la PCR, de modo que para cada *pool* se obtuvieron tres resultados.

Microsatélites amplificados

Los marcadores estudiados fueron los siguientes dinucleótidos: BM143 (Bishop et al., 1994) y RM188 (Barendse et al., 1994). En ambos casos el cebador correspondiente al extremo 5' de la secuencia se marcó con el fluorocromo TET.

Electroforesis capilar de los productos PCR en ABI 310.

El secuenciador automático ABI 310 usa electroforesis capilar con un polímero desnaturalizante (POP4, Perkin-Elmer Cetus) bajo condiciones de temperatura, voltaje, etc., recomendadas por el fabricante para el análisis de microsatélites. A partir del electroferograma de cada una de las amplificaciones individuales se extrajeron los datos de altura de pico y área de pico del alelo principal y de las bandas sombra accesorias, mediante el programa GENESCAN.

Correción de los artefactos de la PCR (bandas sombra, adición de la adenina en el extremo 3' y amplificación diferencial) en los *pools*.

Cada uno de los picos observados en un *pool* es el resultado de la superposición de la banda principal y un número de picos accesorios generados por las bandas sombra de los alelos contiguos. Todos los cálculos se hicieron para altura de pico y área de pico.

La corrección de las bandas sombra se lleva a cabo de la siguiente manera:

Llamamos *altura relativa o área relativa ($AR_{n,i}$) de una banda sombra* al cociente $A_{n,i}/A_{n,}$ donde *n* es el alelo, *i* es el orden de la banda sombra expresado como la diferencia en número de bases con la banda principal.

Los valores de $AR_{n,i}$ fueron medidos partiendo de la información que proporcionan tanto los homocigotos como los heterocigotos en alelos con diferencia en suficiente número de bases para que no exista ningún solapamiento de picos.

Una vez calculados los $AR_{n,i}$ de todas las bandas sombra para todos los alelos se ajustó una recta de regresión (Lipkin et al., 1998) para predecir el comportamiento de las bandas sombra de aquellos alelos no analizados. Posteriormente se corrigieron los picos de los *pools* aplicando la siguiente ecuación:

$$AC_n = A_n - \sum_{i=2}^{r} (AC_{n-i} AR_{n-i,i})$$

donde AC_n se define como la altura o área corregida para las bandas sombra y AC_{n-i} es la altura o área corregida de A_{n-i}. Por tanto para un *pool* con *k* bandas habrá *k* ecuaciones con *k* incógnitas AC_n.

En el caso de la banda resultante de la adición de la adenina en el extremo 3', al estar situada una base por encima del alelo principal no interfiere con la banda principal de ningún alelo en microsatélites dinucleótidos. Es posible reducir al máximo el tamaño de la banda alargando el tiempo de elongación principal de la PCR.

Queda por corregir la amplificación diferencial ya que en la PCR se favorece la amplificación de la banda principal de los alelos más cortos, al contrario que en las bandas sombra, donde los alelos más largos tienen bandas sombra más altas. Teniendo en cuenta que ambos tipos de bandas son resultado de la amplificación de los alelos que existían en el pool, las bandas sombra se producen en detrimento de las principales. La corrección practicada fue la siguiente:

$$AV_n = \sum_{i=-2}^{0} AC_n AR_{n,i}$$

siendo AV_n el valor corregido de altura o área para bandas sombra y amplificación diferencial.

La correción expuesta es un caso particular del tratamiento de la PCR como proceso "amplificador". Mediante la teoría del procesamiento electrónico de señales (Eykhoff 1974, Papoulis 1977) las respuestas reproducibles de un amplificador pueden ser modelizadas con precisión. Perlin (1995) considera que existe una relación lineal entre la *señal de entrada* (las frecuencias de los diferentes alelos que integran el pool) y la de *salida* (la lectura del secuenciador en forma de alturas o áreas de la curva). Sea $\boldsymbol{x}$ el vector de frecuencias alélicas reales (estimado arriba por los AV_n), y $\boldsymbol{A}$ la matriz formada por los elementos $AR_{n,i}/\sum_j AR_{n,j}$. Según su modelo, el vector $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$ predice las concentraciones relativas de ADN presentes en el producto de la PCR. Puesto que $\boldsymbol{y}$ es el dato observado en el experimento, se puede estimar $\boldsymbol{x}$ mediante el producto matricial $\boldsymbol{A}^{-}\boldsymbol{y}$, operación que resume los pasos detallados arriba. La matriz $\boldsymbol{A}^{-}$ es una inversa generalizada de $\boldsymbol{A}$, la cual puede obtenerse por diferentes métodos (p.ej. optimización minimo-cuadrática). Los valores ajustados negativos se truncan a cero.

Finalmente, para estimar las frecuencias alélicas (FA_n) se utilizó la siguiente ecuación:

$$FA_n = \frac{AV_n}{\sum_m AV_m} \times 2N$$

siendo N el número de animales que componen el pool.

Como indicador de la precisión de las frecuencias alélicas se estimó la varianza de la distribución del error técnico V_T, término utilizado por otros investigadores en trabajos anteriores (Lipkin et al., 1998; Darvasi & Soller, 1994), como media muestral de las diferencias cuadráticas entre la frecuencia real (genotipado individual) y las frecuencias estimadas.

Otra manera de medir la precisión de las estimaciones de las frecuencias alélicas es ver en que medida nos alejamos de la recta deseable $x=y$ mediante la regresión de las frecuencias estimadas con los *pools* (x) sobre las frecuencias reales (y).

RESULTADOS Y DISCUSIÓN

Intensidad relativa de las bandas sombra ($AR_{n,i}$).

En la figura 1 se muestran las observaciones y la recta de regresión de las intensidades relativas de cada una de las bandas sombra, considerando bien el valor de la altura del pico, bien el área que encierra, sobre el tamaño del alelo. En ambos casos los valores de altura proporcionan un mejor ajuste que los valores de las áreas.

Como puede observarse, las rectas de regresión para ambos microsatélites, tanto cuando se utiliza la altura como el área, tienen una pendiente similar para valores de $AR_{n,i}$ iguales, siendo la correspondiente a i=-2 la que tiene una pendiente mayor. En el trabajo de Lipkin et al. (1998), se utilizaba como variable regresora el número de repeticiones de los alelos en lugar del tamaño final del mismo.

Estos resultados demuestran que se podría simplificar enormemente el trabajo de puesta a punto inicial que exige trabajar con microsatélites en *pools*, reduciendo aún más los costes de reactivos y mano de obra.

Estimación de las frecuencias alélicas en los *pools* de ADN.

La varianza debido a repeticiones no resultó ser diferente de 0, tanto entre réplicas de PCR ($6{,}88.10^{-6}$) como entre réplicas de preparación del pool ($7{,}18.10^{-5}$), por lo que siempre se utilizó como valor estimado final de las frecuencias alélicas a partir de los *pools* la media de todas las repeticiones.

En la Tabla 1 aparecen los valores de V_T para los dos microsatélites cuando se utilizan los valores de altura o los de área. Los valores de V_T más bajos en ambos microsatélites (BM143, 0,0009; RM188, 0,0014) se obtienen cuando se utilizan las medidas de altura de pico, lo que representa el mismo resultado, como era previsible, a partir del mejor ajuste de la regresión de las bandas sombra para altura. Teniendo en cuenta un intervalo de confianza del 95 % la frecuencia de cualquier alelo se encontraría dentro de un rango de $x \pm 6{,}3\%$. Basándonos en este error medio de muestreo (V_T = 0,001), en el caso de familias de 500 medios hermanos y un gen con efecto aditivo de 0,25 desviaciones típicas fenotípicas, el resultado obtenido se correspondería con una pérdida de potencia del 15% aproximadamente, lo que se traduciría en la necesidad de incrementar el tamaño de familia en un 26 % (130 animales más) para obtener la misma potencia de detección de asociación entre el locus y el carácter cuantitativo.

En cuanto a la evolución del error técnico en función del número de individuos dentro del *pool* los resultados obtenidos demuestran que no parece que exista ningún incremento, al menos hasta un tamaño final de 50 animales que es el *pool* de mayor tamaño utilizado (Fig. 2).

En la figura 3 aparecen enfrentadas la estimación de las frecuencias alélicas por genotipado individual y la estimación desde los pools. La correlación fue alta, r=0,93 para ambos microsatélites frente a valores obtenidos por Breen et al. (1999) entre r=0,96 y r=0,79. La nube de puntos alrededor de la recta se hace más dispersa a medida que las frecuencias alélicas crecen, lo que indica que la precisión en la estima es mejor para alelos cuya frecuencia de aparición dentro del pool es baja.

CONCLUSIONES

Los resultados presentados proporcionan unas excelentes expectativas para justificar la utilización de *pools* de ADN en estudios de búsqueda de genes asociados a caracteres productivos o enfermedades, especialmente en las etapas iniciales de búsqueda cuando se pretende localizar regiones candidatas a lo largo de todo el genoma de una especie mediante genotipado selectivo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la CICYT y Fondos FEDER a través de los proyectos: **1FD97-0042** y **2FD97-1191.**

BIBLIOGRAFÍA

Bishop M.D., Kappes S.M., Keele J.W., Stone R.T., Sunden S.L.F.,RA Hawkins G.A., Solinas-Toldo S., Fries R., Grosz M.D., Yoo J., Beattie C.W. 1994. A Genetic Linkage Map for Cattle. *Genetics* **136**:619-639.

Barcellos, L.F.; Klitz, W.; Leigh Field, L.; Tobias, R.; Bowcock, A.M.; Wilson, R.; Nelson, M.P.; Nagatomi, J. and Thomson, G. 1997. Association mapping of disease loci, by use os a pooled DNA genomic screen. *Am. J. Hum. Genet*.**61**:734-747.

Barendse W. et al. 1994. A genetic linkage map of the bovine genome. *Nat. Genet.* **6**: 227-235.

Daniels, J.; Holmans, P.; Williams, N.; Turic, D.; McGuffin, P.; Plomin, R. And Owen, M.J. 1998. A simple method for analyzing microsatellites allele image patterns generated from DNA pools and its application to allelic association studies. *Am. J. Hum. Genet*. **62**:1189-1197.

Darvasi A., Soller M. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics*.**138**:1365-1373.

Eykhoff P. 1974. *System identification: parameter and state estimation*. John Wiley & Sons. Chichester.

Kirov, G.; Williams, N.; Sham, P.; Craddock, N. And Owen, M.J. 2000. Pooled genotyping of microsatellite markers in parent-pffspring trios. *Genome Research.* **10**:105-115.

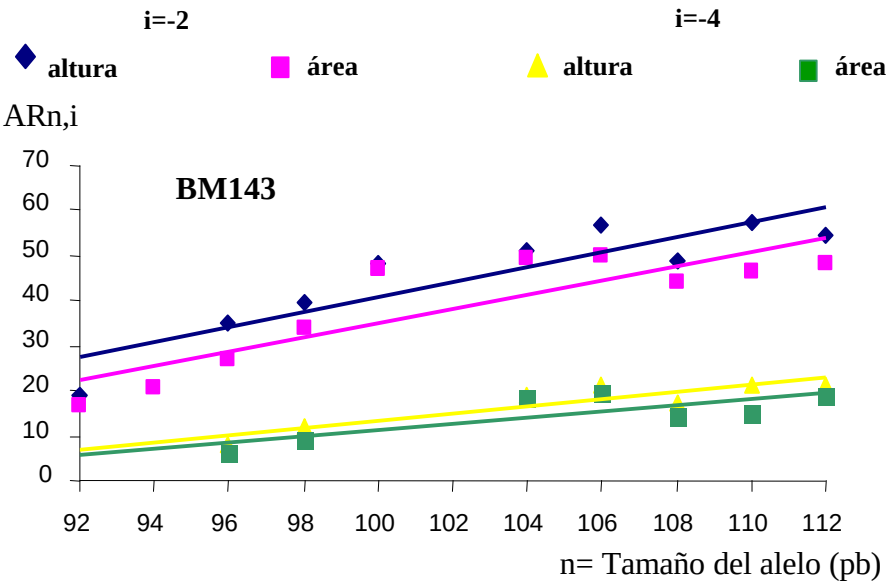
Lipkin, E.; Mosig, M.O.; Darvasi, A.; Ezra, E.; Salom, A.; Frideman, A. And Soller, M. 1998. Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. *Genetics.* **149**: 1557-1567.

Papoulis A. 1977. *Signal Analisis*. McGraw-Hill. New York.

Perlin M.W., Lancia G., Ng S.K. 1995. Towards fully automated genotyping: genotyping microsatellite markers by deconvolution. *Am. J. Hum. Genet.* **57**:1199-1210.

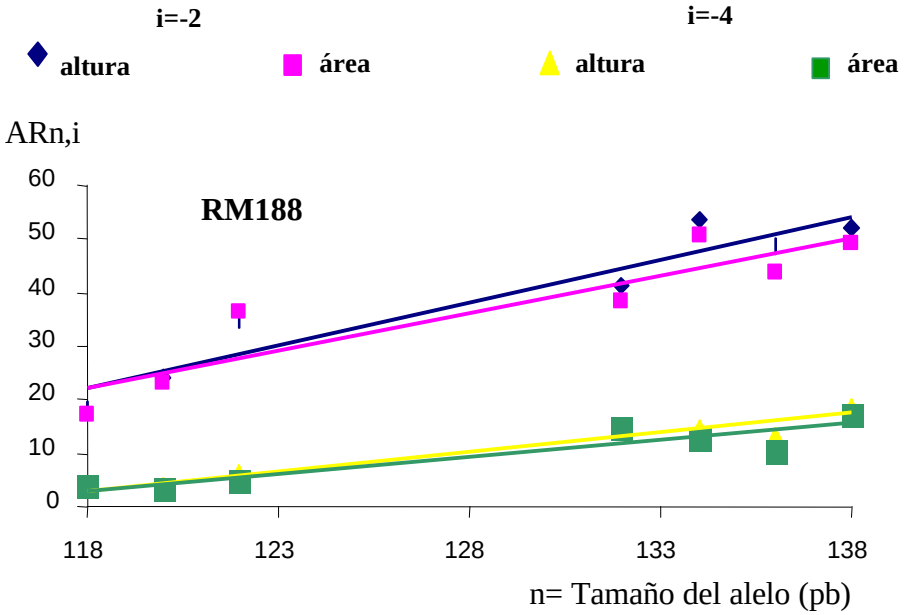
Shaw, S.H.; Carrasquillo, M.M.; Kashuk, C.; Puffenberger, E.G. and Chakravarty, A. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Research.* **8**:111-123.

Figura 1:
Regresión de las bandas sombra relativas (*AR*) sobre el tamaño del alelo (*n*). *r* es el coeficiente de correlación.
Figure 1:
Regression of the relative shadow bands (AR) on the length of allele (n). r is the correlation coefficient.



BM143 (altura):
$AR_{(n,-2)}$ = 1,64n-123,5 r=0,89
$AR_{(n,-4)}$ = 0,78n-64,62 r=0,90

BM143 (área):
$AR_{(n,-2)}$ = 1,58n-123,1 r=0,86
$AR_{(n,-4)}$ = 0,68n-56,54 r=0,79



RM188 (altura):
$AR_{(n,-2)}$ = 1,59n-165,15 r=0,95
$AR_{(n,-4)}$ = 0,72n-64,62 r=0,97

RM188 (área):
$AR_{(n,-2)}$ = 1,41n-143,75 r=0,91
$AR_{(n,-4)}$ = 0,63n-71,54 r=0,92

Figura 2:
Evolución del error técnico (V_T) a medida que aumenta el número de alelos dentro del pool.
Figure 2:
Evolution of the technical error (V_T) with the increasing of the number of alleles included in the pool

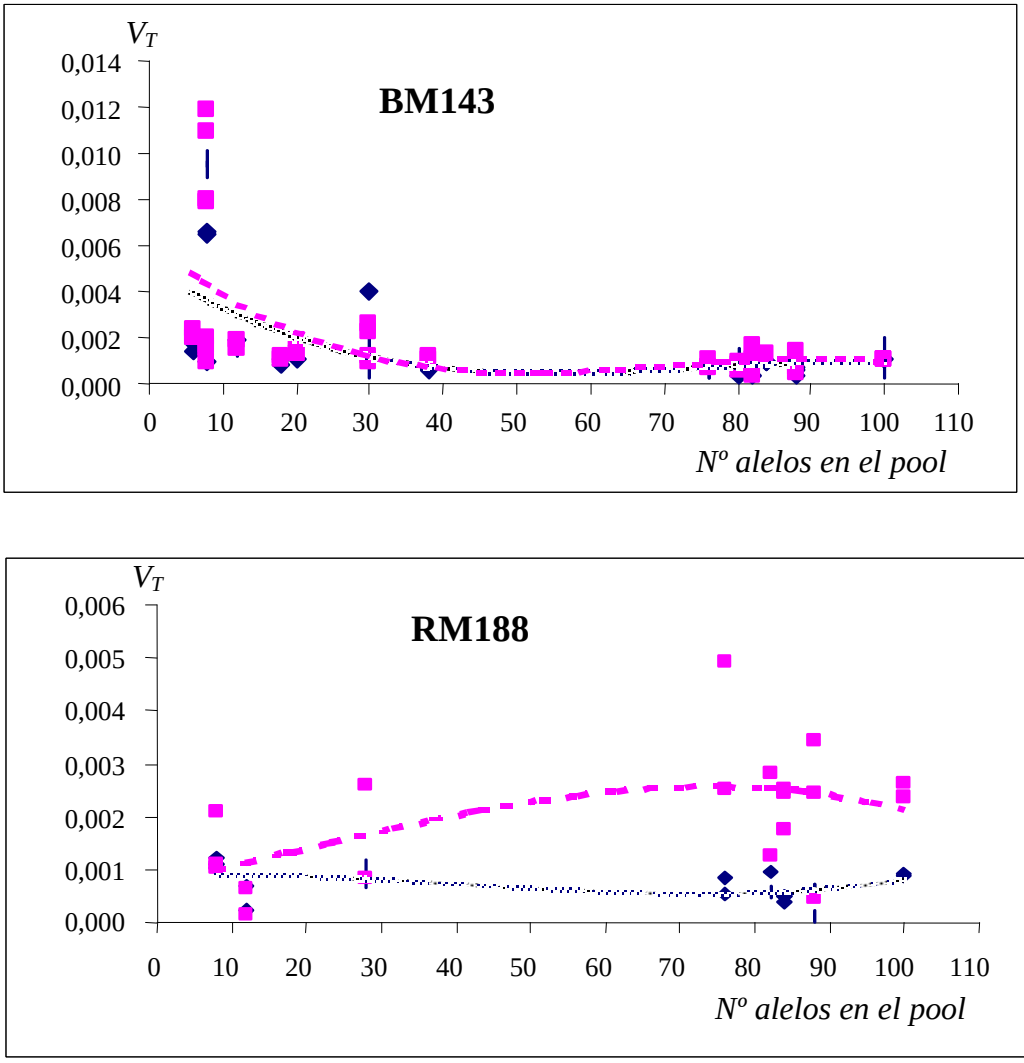
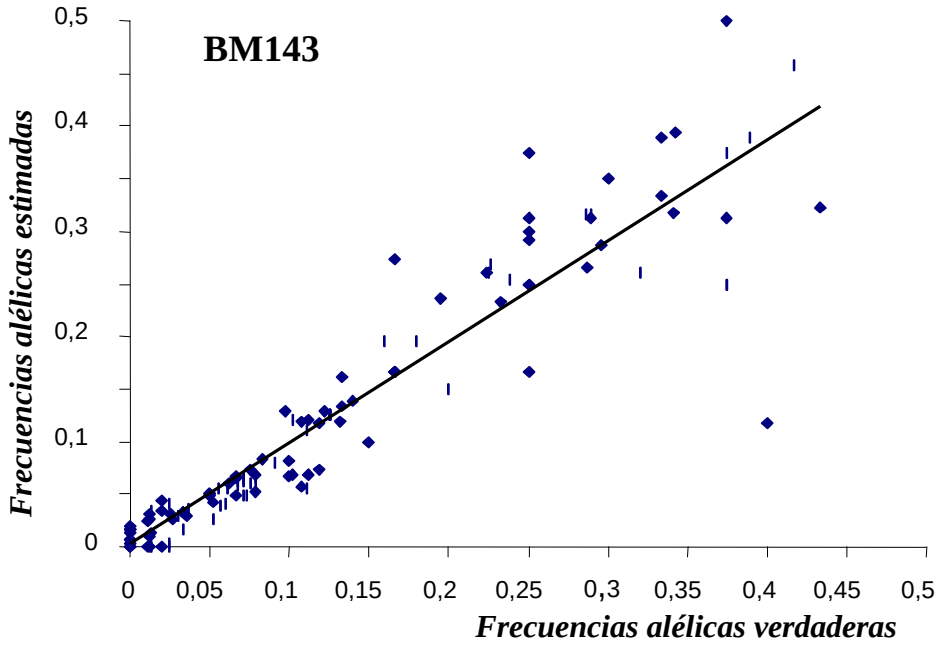


Figura 3

Regresión de las frecuencias alélicas estimadas utilizando pools sobre las frecuencias verdaderas en el caso de los microsatélites BM143 y RM188

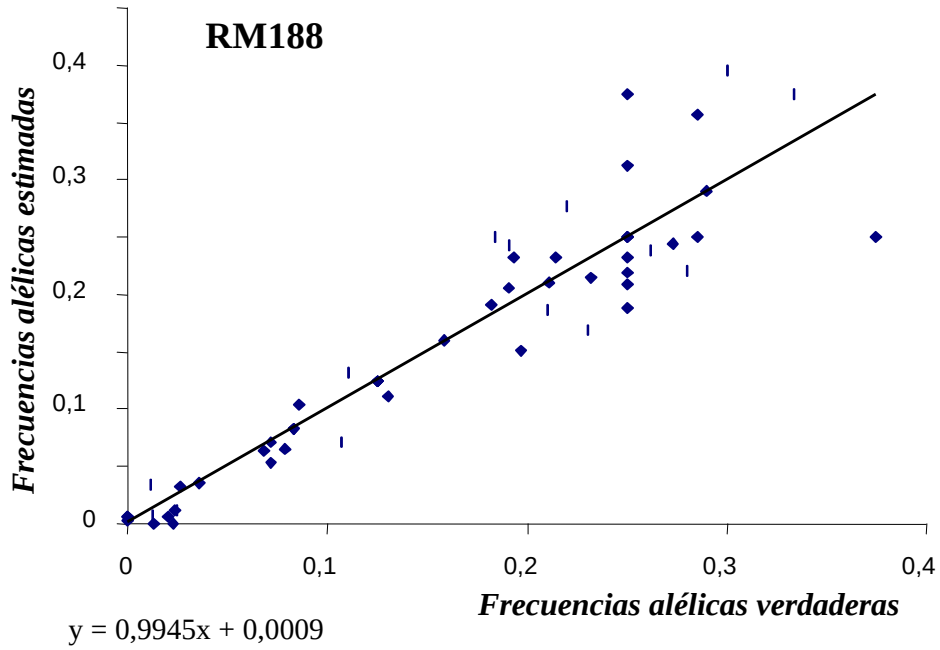
Figure 3

Regression of allelic frequencies estimates using pools on the true values for BM143 and RM188 microsatellites



$y = 0,958x + 0,0026$

$R^2 = 0,8676$



$y = 0,9945x + 0,0009$

$R^2 = 0,8675$

Tabla 1

Valores de V_T para los dos microsatélites cuando se utilizan la altura o el área de los picos

Table 1

V_T values for both microsatellites when high or area of the peaks are used

	Vt (altura)	Vt (área)
BM143	0,0018	0,0032
RM188	0,0008	0,0026