Identificación de alelos con efecto significativo en la diferenciación de poblaciones

David Garcia[1*],Carlos Carleos[2], Jesus A. Baro[3], Javier Cañon[1]

[1] Departamento de Producción Animal, Universidad Complutense de Madrid
[2] Departamento de Estadística, Universidad de Oviedo
[3] Unidad de Mejora Genética, SERIDA
* Departamento de Producción Animal
Facultad de Veterinaria
Avda. Puerta de Hierro
28040 Madrid
913943758
davidgm@eucmos.sim.ucm.es

RESUMEN

Los sistemas de clasificación racial basados en frecuencias de los alelos de marcadores y genotipos individuales se emplean en programas de conservación, detección de fraudes, análisis de poblaciones híbridas, estudios forenses y pruebas de paternidad. La técnica -asignación de individuos a una raza maximizando la probabilidad de que el individuo pertenezca a ella- es simple, aunque su aplicación rigurosa supone cierta complejidad. Este trabajo revisa uno de los aspectos críticos: la presencia de alelos atípicos. Éstos se definen como los que tienen frecuencias extremas en alguna raza; casos particulares son los alelos específicos de raza y los alelos ausentes en una raza. El enfoque que se ha seguido consistió en examinar para cada locus el comportamiento de una tabla de contingencia con las razas frente al número de alelos. La significación de cada tabla y de cada celda de cada tabla se estimó mediante tres métodos distintos: el habitual, basado en el estadístico ji-cuadrado, y dos aproximaciones por simulación de contrastes exactos: MC y MCMC. En los casos simulados, la significación de la tabla de contingencia se estima como la proporción de simulaciones en que la probabilidad de la tabla observada era superior a la de la simulada. Se estudia la posibilidad de emplear un enfoque análogo para determinar una significación por celda.

PALABRAS CLAVE

marcadores, asignación racial, significación

ABSTRACT

Breed assignment procedures based on marker allele frequencies and individual genotypes can be used for conservation purposes, fraud detection, analysis of hybrid populations, forensic advice, and paternity tests. The technique -allocation of an individual to the breed maximising the probability of that individual belonging to it- is straightforward but rigorously speaking it has a number of weak points. We address one of them in this paper: the presence of outlier alleles. Outlier alleles are those having extreme frecuencies in a certain breed, with breed specific alleles and missing alleles as particular cases. The approach followed to study this topic was to examine for each locus the behaviour of breed vs. number of alleles contingency tables. The significance of each table and of each cell in each table was estimated by three different methods -the usual chi-square statistic and two simulated -MC and MCMC- exact tests. In the simulated cases significance was estimated as the proportion of realisations where the probability of the observed table exceeded that of the simulated one, proceeding analogously for individual cells.

KEYWORDS

markers, breed allocation, significance

INTRODUCCIÓN

Por asignación racial se entiende la clasificación de una muestra de un animal en alguna de una serie de razas a las que el animal pueda o haya podido pertenecer. En los últimos tiempos se han venido desarrollando diferentes técnicas de asignación que tienen en cuenta el genotipo de la muestra y las frecuencias de los alelos de marcadores en las diferentes razas (Banks & Eichert 2000,, Shriver et al. 1997). Algunas de las aplicaciones de estas técnicas son: asignar individuos a razas en programas de conservación étnica, detección de fraudes, análisis de las proporciones de las poblaciones originales en las híbridas, monitorización de la inmigración, estudios forenses y pruebas de paternidad. La información genética empleada (ver la comparación realizada por Blott, 1999) puede ser de marcadores multialélicos como microsatélites, de alta disponibilidad, con métodos estándar de genotipado y muy informativos, o basada en marcadores dialélicos como los SNPs que son mucho más frecuentes y de los que se espera que pronto permitan realizar genotipados masivos a bajo coste.

MÉTODOS

La información sobre microsatélites y otros polimorfismos similares se analiza con métodos estadísticos clásicos de teoría de la decisión para resolver el problema de la asignación racial. El criterio de asignación consiste en asignar un individuo a la raza en la que la aparición de ese individuo sea más probable. Ésta

se toma como la que maximiza la probabilidad de aparición de ese individuo, dado un conjunto de razas y de loci y conocidas las frecuencias de los alelos de dichos loci en cada raza. Se asumirán equilibrios de ligamiento y de Hardy-Weinberg. Nos centraremos en el estudio de la presencia de alelos especialmente influyentes en la clasificación: los alelos atípicos que presentan frecuencias particularmente altas o bajas en alguna población. Casos extremos los constituyen los alelos que aparecen asociados únicamente a una raza, y los alelos que no han sido registrados en alguna raza. La hipótesis nula de partida es la homogeneidad de poblaciones respecto a la distribución de los alelos. Los datos del análisis se presentan en una tabla de contingencia "poblaciones / alelos".

Tabla de ejemplo. La siguiente tabla recoge las frecuencias alélicas observadas para el locus CSSM en 50 individuos de las razas bovinas "asturiana de los valles", "asturiana de la montaña" y "pirenaica":

ASM	0	3	0	10	13	5	31	2	10	2	18	6
ASV	1	1	3	17	10	4	24	6	16	1	8	9
PIR	0	6	1	14	40	0	16	0	10	3	4	6

Existen varios enfoques para analizar la significación de las celdas (casillas) bajo la hipótesis de homogeneidad. En primer lugar, comentaremos tres estrategias de contraste empleadas con la tabla en conjunto.

El problema radica en contrastar la hipótesis de distribución homogénea de los alelos en todas las poblaciones, lo que equivale a contrastar las hipótesis de independencia entre filas y columnas. El estadístico tradicionalmente empleado es el χ^2:

$$X = \sum_{celdas} \frac{(\text{obs.} - \text{esp.})^2}{esp.}$$

Como regla general, no se recomienda su uso cuando hay celdas con menos de 5 observaciones. En tal caso, en una tabla 2×2 es factible emplear el método exacto de Fisher, pero con tablas mayores, como es el caso de los microsatélites, que son muy polimórficos, el número de combinaciones lo hace computacionalmente impracticable. Se ha de recurrir a procedimientos de Montecarlo. Se describirán dos de ellos, que denominaremos MC (Montecarlo clásico) y MCMC (Montecarlo con cadenas de Markov).

Montecarlo clásico: El algoritmo para el cálculo de la significación de una tabla frente a la hipótesis nula de independencia u homogeneidad es como sigue:

1. Calcular la probabilidad de la tabla observada T:

$$\Pr[T] = \frac{\prod_{i=1}^{nfilas}(T_{i\cdot}!)\prod_{j=1}^{ncols}(T_{\cdot j}!)}{T_{\cdot\cdot}!\prod_{i=1}^{nfilas}\prod_{j=1}^{ncols}(T_{ij}!)}$$

2. Generar aleatoriamente N tablas bajo la hipótesis de independencia y con los mismos recuentos marginales que T: $T_1, \ldots, T_N$.

3. Calcular las probabilidades de estas tablas.

4. Compararlas con Pr[T] y estimar la significación (p-valor) de T como:

$$\hat{P}_v = \frac{\#\{T_x | \Pr[T_x] \leq \Pr[T], x \in \{1, \ldots, N\}\}}{N}$$

MCMC: Cuando se emplea simulación de Montecarlo con cadenas de Markov, la generación de nuevas tablas bajo la hipótesis nula es mucho más rápida. Con la tabla de ejemplo, la generación de 100000 tablas mediante Montecarlo clásico llevó 3 min 52 s, en tanto que 10000000 tablas MCMC consumieron solamente 49 s. El algoritmo propuesto por Guo y Thompson (1992) cambia un alelo en cada iteración y por tanto sólo se modifican los valores de cuatro celdas respecto a la tabla generada en la iteración previa. La tabla de partida es la observada. Dados $1 \leq i, i' \leq nfilas$ y $1 \leq j, j' \leq ncols$, el cambio afecta únicamente a las casillas (i,j), (i',j), (i,j') y (i',j').

Aunque es mucho más rápido que el MC, se requiere la generación de muchas más tablas para obtener la misma precisión en las estimaciones, al no ser éstas independientes. Hastings (1970) demostró que en estas circunstancias puede obtenerse un estimador del p-valor $\hat{P}_v$ asintóticamente normal y que converge en probabilidad a su valor real.

De la tabla a la celda. Se plantean ahora estrategias centradas en la significación de las celdas y no de las tablas. Se pretende determinar cuáles de las casillas de cierta tabla de contingencia son las responsables de que la tabla se aleje de la hipótesis nula de independencia.

Estadístico χ^2. La propia forma del estadístico sugiere asignar como "índice de rareza o atipicidad" el sumando correspondiente a cada celda. El estadístico χ^2 es la suma, celda a celda, de la diferencia cuadrática "tipificada" entre el valor observado de la celda y el valor esperado (el que tendría la celda, por término medio, bajo la hipótesis nula). Este valor ofrece una medida comparativa de la contribución de cada celda al alejamiento de la hipótesis nula.

Métodos de Montecarlo. Por cada tabla generada en los métodos MC y MCMC, se registran los valores observados (k) en cada celda (i,j) y se incrementan unos contadores que acumulan el número de veces que la celda (i,j) adopta el valor k. Así se obtiene una estimación de la distribución de probabilidades de los diferentes valores que puede adoptar cada celda bajo la hipótesis nula. La distribución de los valores de una celda ofrece toda la información estadística sobre su significación. Para resumir la relación entre la distribución y el valor observado, existen varias opciones:

1. "tipificación": si E es la media muestral, y D la desviación típica muestral, dar el valor $\frac{k-E}{D}$ o su cuadrado como índice de atipicidad; este valor indica la rareza de los valores observados en la celda: cuanto mayores son, más atípico es el valor observado, por lo que permite la comparación entre celdas. Sin embargo, no tiene una significación propia nítidamente definida, salvo que se suponga que la distribución por celda es más o menos normal (una corrección por continuidad sería adecuada en este caso).

2. estimar el "p-valor"; según la "definición" empleada, puede ser:

(a) calcular la suma de las probabilidades estimadas más bajas que la probabilidad estimada del valor k: $\sum_{\Pr[l]\leq\Pr[k]}\Pr[l]$.

(b) calcular el doble del mínimo de: la suma de las probabilidades estimadas para los valores menores o igules que k, y la suma de las probabilidades estimadas para los valores mayores o iguales que k (Manly & al. 1998): $2\times\min\{\sum_{l\leq k}\Pr[l],\sum_{l\geq k}\Pr[l]\}$.

Los valores calculados en 2 pretenden estimar directamente una significación, pero como estimadores resultan inconsistentes en el caso de que exista al menos un valor con la misma probabilidad teórica que el observado: Sea k el valor observado en cierta celda. Sea P_v el p-valor asociado y $\hat{P}_v$ el estimador 2a. Sean k y k' dos valores que puede adoptar la celda (i,j), y supóngase que $\Pr[k|H_0]=Pr[k'|H_0]$. Sea p_k la estimación empírica de $\Pr[k|H_0]$ y $p_{k'}$ la de $Pr[k'|H_0]$. Cuando el número de iteraciones es muy grande, $\Pr[p_k>p_{k'}]\approx 0'5\approx\Pr[p_k<p_{k'}]$, con lo que existe $\epsilon=\frac{p_k}{2}>0$ tal que $\Pr[|\hat{P}_v-P_v|>\epsilon]\approx 0'5$ no tiende a cero y $\hat{P}_v$ no es débilmente consistente.

RESULTADOS

A partir de la tabla de ejemplo, obtendremos cada uno de los estadísticos antes descritos. En cuanto a la significación global de la tabla frente a la hipótesis de homogeneidad, el estadístico χ^2 toma un valor de 67'79 (P_v=1'47E-6). Las estimaciones del p-valor mediante Montecarlo son las dos prácticamente nulas (MC: 1'0E-5, MCMC: 2'0E-7).

Los valores del estadístico χ^2 celda a celda fueron:

0'33	0'03	1'33	0'98	3'04	1'33	2'27	0'17	0'33	0'00	6'40	0'14
1'33	1'63	2'08	0'81	5'76	0'33	0'00	4'17	1'33	0'50	0'40	0'57
0'33	2'13	0'08	0'01	17'19	3'00	2'48	2'67	0'33	0'50	3'60	0'14

Los valores de Montecarlo (MC y MCMC) tipificados fueron:

0'50	0'05	2'02	1'69	5'71	2'04	4'48	0'26	0'57	0'00	10'68	0'24
2'01	2'54	3'14	1'40	11'23	0'52	0'01	6'42	2'27	0'76	0'65	0'93
0'50	3'29	0'12	0'01	32'52	4'62	4'84	4'09	0'57	0'76	6'01	0'23

Los p-valores estimados, fueron:

1'00	1'00	0'31	0'21	0'02	0'17	0'04	0'72	0'57	1'00	0'00	0'81
0'33	0'17	0'11	0'29	0'00	0'49	1'00	0'02	'014	0'67	0'54	0'34
1'00	0'89	1'00	0'86	0'00	0'03	0'03	0'06	0'57	0'40	0'01	0'81

CONCLUSIONES

Hemos querido obtener medidas objetivas para la identificación de las celdas "causantes" del alejamiento de la hipótesis de homogeneidad en tablas de contingencia. El estimador del nivel crítico (p-valor) propuesto no es consistente cuando existe un valor de la celda que bajo hipótesis nula tiene la misma probabilidad que el observado. Aunque este hecho es presumiblemente infrecuente, conviene aportar adicionalmente algún valor de diagnóstico más estable, como los "índices tipificados".

REFERENCIAS

Banks, M.A.; Eichert, W.(2000) WHICHRUN (version 3.2): A computer program for population assignment of individuals based on multilocus genotype data. Journal of Heredity, 91:87-89

Blott, S.C.; Williams, J.L.; Haley, C.S.(1999) Discriminating among cattle breeds using genetic markers. Heredity, 82:613-619

Guo, S.W.; Thompson, E.A.(1992) Performing the exact test of Hardy-Weinberg proportions for multiple alleles. Biometrics, 48:361-372

Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97-109.

Manly, B.F.J. Randomization, bootstrap and Monte Carlo methods on biology. Chapman & Hall, 1998.

Shriver, M.D.; Smith, M.W.; Jin, L.; Marcini, A.; Akey, J.M.; Deka, R.; Ferrell, R.E.(1997) Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics, 60:957-964