



Antropología Biológica

# Evaluación del número mínimo de marcadores para estimar ancestría individual en una muestra de la población argentina

*Evaluation of the minimum number of markers for individual ancestry estimation in an Argentinean population sample*

*María Gabriela Russo\*, Francisco Di Fabio Rocca\*\*, Patricio Doldán\*\*\*, Darío Gonzalo Cardozo\*\*\*\*, Cristina Beatriz Dejean\*\*\*\*\*, Verónica Seldes\*\*\*\*\* y Sergio Alejandro Avena\*\*\*\*\**

\*Equipo de Antropología Biológica, Departamento de Cs. Naturales y Antropológicas, CEBBAD, Fundación de Historia Natural Félix de Azara, Universidad Maimónides. CONICET. Argentina. E-mail: russo.mariagabriela@maimonides.edu; \*\* Equipo de Antropología Biológica, Departamento de Cs. Naturales y Antropológicas, CEBBAD, Fundación de Historia Natural Félix de Azara, Universidad Maimónides. CONICET. Argentina. E-mail: fdifabiorocca@gmail.com; \*\*\* Equipo de Antropología Biológica, Departamento de Cs. Naturales y Antropológicas, CEBBAD, Fundación de Historia Natural Félix de Azara, Universidad Maimónides, Argentina. E-mail: doldan.patricio@gmail.com; \*\*\*\* Equipo de Antropología Biológica, Departamento de Cs. Naturales y Antropológicas, CEBBAD, Fundación de Historia Natural Félix de Azara, Universidad Maimónides. Sección de Antropología Biológica, ICA, Facultad de Filosofía y Letras, Universidad de Buenos Aires. CONICET. Argentina. E-mail: dg.cardozo@gmail.com; \*\*\*\*\* Equipo de Antropología Biológica, Departamento de Cs. Naturales y Antropológicas, CEBBAD, Fundación de Historia Natural Félix de Azara, Universidad Maimónides. Sección de Antropología Biológica, ICA, Facultad de Filosofía y Letras, Universidad de Buenos Aires, Argentina. E-mail: dejeancr@gmail.com; \*\*\*\*\* Sección de Antropología Biológica, ICA, Facultad de Filosofía y Letras, Universidad de Buenos Aires. CONICET. Argentina. E-mail: vseldes@gmail.com; \*\*\*\*\* Equipo de Antropología Biológica, Departamento de Cs. Naturales y Antropológicas, CEBBAD, Fundación de Historia Natural Félix de Azara, Universidad Maimónides. Sección de Antropología Biológica, ICA, Facultad de Filosofía y Letras, Universidad de Buenos Aires. CONICET. Argentina. E-mail: sergioavena@gmail.com

## Resumen

*La estimación de ancestría individual posee gran relevancia en el estudio de la composición poblacional en regiones como Sudamérica, que han atravesado intensos procesos de mestizaje, lo que también tiene implicancia en ciencias de la salud. Debido a esto, es importante conocer los factores que influyen en la confiabilidad de los resultados obtenidos. En este trabajo se evalúa el número mínimo de marcadores informativos de ancestría (AIMs) a partir del cual las estimaciones resultarían aceptables. Se toma como ejemplo el cálculo en individuos provenientes de una muestra poblacional de diferentes regiones de Argentina. Considerando un modelo de tres componentes (nativo americano, euroasiático y subsahariano), se calculó la ancestría de 441 individuos utilizando 10, 20, 30 y 50 AIMs. Los resultados indican que el número de marcadores influye sobre la estimación de ancestría y su precisión aumenta al incrementarse la cantidad de AIMs. Al comparar con las estimaciones obtenidas en un trabajo previo a partir de 99 AIMs, se observó que para el componente minoritario (en este caso subsahariano) se obtiene una buena correlación utilizando al menos 30 marcadores. Se concluye que es necesario considerar en los estudios de ancestría individual el número de marcadores, su capacidad informativa y las características de la población bajo estudio.*

**Palabras clave:** número de AIMs; ancestría individual; población argentina.

## Abstract

*Estimation of individual ancestry has great relevance when studying population composition in regions like South America, where intensive admixture processes have occurred, being also important in biomedical sciences. For that reason, it is important to assess the factors that may affect the reliability of results. In this work, we investigate the minimum number of ancestry informative markers (AIMs) for obtaining acceptable estimations of ancestry. As an example, we take individuals from a population sample of different Argentinean regions. Considering a three component model (Native American, Eurasian and Sub-Saharan), we calculated ancestry of 441 individuals using 10, 20, 30 and 50 AIMs. The results indicate that the number of markers affects ancestry estimation and its accuracy increases with AIMs number. When compared to previous estimations obtained from 99 AIMs, the result shows that at least 30 markers are needed to achieve good correlation values for the minority component (Sub-Saharan in this case). For individual ancestry studies, we suggest to take into account not only the number of markers, but also its informativeness and the background of the studied population.*

**Keywords:** number of AIMs; individual ancestry; Argentinean population.

Recibido 27-10-2015. Recibido con correcciones 26-02-2016. Aceptado 10-05-2016

Revista del Museo de Antropología 9 (1): 49-56, 2016 / ISSN 1852-060X (impreso) / ISSN 1852-4826 (electrónico)

<http://revistas.unc.edu.ar/index.php/antropologia/index>

IDACOR-CONICET / Facultad de Filosofía y Humanidades – Universidad Nacional de Córdoba - Argentina

## Introducción

La distribución de la variación genética humana ha sido ampliamente estudiada, siendo tema de discusión si esta variabilidad sigue un patrón clinal o si, por el contrario, se explica mejor por un modelo de "clusters" (Rosenberg et al. 2005). Al margen de este debate, existen ciertos marcadores con frecuencias alélicas diferenciales entre determinados grupos poblacionales que se corresponden con límites geográficos definidos, siendo los más significativos los de nivel continental, los cuales son los elegidos por antonomasia en estudios de ancestría genética.

Las poblaciones cosmopolitas actuales de Argentina, al igual que las del resto de América, reflejan en su acervo génico la presencia de tres grandes poblaciones parentales, cuyos límites se corresponden con divisiones continentales: nativo-americana, euroasiática occidental y africana subsahariana. Estudios previos muestran que existen importantes diferencias regionales, dadas tanto por las cantidades relativas de cada componente como por la intensidad y dirección del proceso de mestizaje (Avena et al. 2012; Corach et al. 2010; García et al. 2011, entre otros).

La miscegenación acaecida se encuentra plasmada diferencialmente en el perfil genético de los individuos y las proporciones de ancestría en cada uno de ellos pueden estimarse usando Marcadores Informativos de Ancestría (AIMs, por su sigla en inglés), cuyas frecuencias alélicas varían entre los grupos caracterizados como parentales. Habitualmente se considera que la diferencia entre una parental y las otras debe ser de al menos un 30% a un 40%. Por ejemplo, la transición a C en el sistema Duffy (determinante del fenotipo NULL) es un excelente marcador, ya que su frecuencia en poblaciones subsaharianas supera el 95%, tiene una prevalencia inferior al 2% en el sur de Europa y está prácticamente ausente en nativo-americanos (Shriver et al. 2003).

Dentro de este contexto, se han desarrollado diversos métodos de estimación de la estructuración poblacional que permiten evaluar la ancestría individual (e.g. Alexander et al. 2009). Uno de los más utilizados es un método bayasiano implementado en el programa *Structure* (Pritchard et al. 2000) y que ha sido empleado tanto para la detección de estructuración poblacional, como para estudios de caso-control y análisis de identificación de individuos.

Por otra parte, el conocimiento de la estructura genética de una población resulta de potencial utilidad en ciencias de la salud pues la prevalencia de diversas patologías con base genética puede variar entre distintos grupos humanos (e.g. Burchard et al. 2003) y puede relacionarse con la ancestría de los individuos que los componen (e.g. Pevrah et al. 2015). Por ejemplo, existen variaciones

interpoblacionales asociadas a genes que predispondrían a determinadas enfermedades, como el cáncer de mama (Bonilla et al. 2015) y de próstata (Pevrah et al. 2015) e hipertensión (Zhu y Cooper 2007), entre muchas otras.

Si bien existen trabajos en los cuales se ha estimado el número de marcadores necesarios para estudiar la estructura poblacional y la ancestría individual (e.g. Ruiz-Linares et al. 2014; Ruiz-Narváez et al. 2010; Tsai et al. 2005; Turakulov y Eastaerl 2003), la gran mayoría utiliza paneles conformados por genotipos basados en grandes cantidades de marcadores en individuos que no provienen de poblaciones argentinas (e.g. HGDP-CEPH –Cann et al. 2002; *HapMap Project –International HapMap Consortium* 2003). Dado que existen diferencias en la estructura poblacional según la región o el continente analizado, resulta necesario obtener estimaciones específicas para cada población (Heinz et al. 2013; Ruiz-Linares et al. 2014) y priorizar un resultado aceptable, considerando también las capacidades metodológicas y su aplicabilidad en distintos ámbitos, tanto en investigación básica como clínica.

Teniendo en cuenta lo expuesto, el objetivo del presente trabajo consiste en estimar el efecto del número de AIMs sobre el cálculo de la ancestría individual en una muestra poblacional de nuestro país y en proponer un número mínimo de marcadores a partir del cual las estimaciones resultan aceptables.

## Materiales y Métodos

### *Estimación de ancestría individual utilizando distinto número de AIMs*

Se emplearon los datos del trabajo realizado por Avena et al. (2012) en el que, utilizando 99 AIMs, se genotipificaron 441 individuos provenientes de distintas regiones de Argentina (Noroeste, Noreste, Centro y Sur), siendo la ancestría promedio 65% euroasiática, 31% nativoamericana y 4% subsahariana. Sobre ese total de marcadores, se realizaron muestreos aleatorios e independientes para obtener conjuntos de 10, 20, 30 y 50 AIMs. Para cada conjunto la selección al azar de los marcadores se realizó cinco veces, con reposición luego de cada muestreo, a fin de obtener réplicas que permitan poner a prueba hipótesis estadísticas (ver más adelante). Con cada muestreo se estimó la ancestría individual utilizando el programa *Structure* 2.3.3 (Pritchard et al. 2000). Asumiendo un modelo tripartito ( $K = 3$ ), se incluyeron en el análisis los genotipos de las poblaciones parentales nativo-americana, euroasiática occidental y subsahariana utilizados en el trabajo original. El período de *burn-in* fue de 100.000 y las probabilidades posteriores se estimaron a partir de 1.000.000 iteraciones bajo el modelo *admixture*, utilizando los parámetros por defecto. Para cada individuo se calculó además el intervalo de probabilidad del 90%.

### *Análisis del efecto del número de AIMs sobre la estimación de ancestría*

El efecto del número de marcadores sobre la ancestría individual fue estudiado mediante modelos lineales generales y mixtos (GLM, por su sigla en inglés), utilizando el paquete *nlme* (Pinheiro et al. 2015) de R (R Core Team 2014) a través del programa Infostat (Di Rienzo et al. 2013), para cada componente por separado. Este método es preferible frente al Análisis de la Varianza (ANOVA) tradicional, puesto que emplea máxima verosimilitud para estimar los parámetros, permite la elección del mejor modelo que explica la variación observada y posibilita el modelado de la estructura de varianzas (Zuur et al. 2009).

Los supuestos de homogeneidad de varianzas y normalidad se analizaron mediante gráficos de dispersión de residuos. Utilizando los criterios de información de Akaike (1974), o AIC y Bayesiano de Schwarz (1978), o BIC, se evaluaron modelos que diferían en los componentes aleatorios incluidos, así como en la estructura de varianzas. El modelo final elegido consta del efecto del número de marcadores y el componente aleatorio de los individuos, junto con el modelado de varianzas, mediante la función Identidad, con el número de marcadores como criterio de clasificación. La estimación de los parámetros del modelo se realizó por el método de máxima verosimilitud restringida (REML, por su sigla en inglés). Los niveles del factor número de AIMs se compararon mediante el método de Di Rienzo, Guzmán y Casanoves (DGC), recomendado por estos autores para realizar comparaciones múltiples de medias (Di Rienzo et al. 2002).

### *Análisis del efecto del número de AIMs sobre la precisión de la estimación*

Con el objetivo de evaluar la variación en la precisión del cálculo de ancestría en función del número de AIMs, se utilizaron los intervalos de 90% de probabilidad obtenidos para cada estimación individual. La amplitud de estos intervalos fue promediada entre todos los individuos para cada una de las cinco corridas realizadas con cada conjunto de marcadores. El efecto de la cantidad de AIMs fue evaluado mediante GLM para cada componente

por separado. El modelo final fue elegido con el mismo criterio señalado en el apartado anterior, e incluye no sólo el efecto del número de marcadores, sino también la ancestría promedio como covariable. En el caso del componente euroasiático occidental, el término correspondiente a la covariable no fue incorporado debido a que la corrección de la variable respuesta, teniendo en cuenta la ancestría, no mejoraba significativamente el modelo. Por lo tanto, en este caso, el modelo final consta solamente del efecto del número de marcadores. La estimación de los parámetros se realizó nuevamente por el método REML y los niveles del factor número de marcadores se compararon mediante el método DGC.

### *Análisis de correlación entre estimaciones obtenidas con distinto número de AIMs*

Para cada componente por separado, se analizó la correlación entre la ancestría individual obtenida con cada conjunto de AIMs y la calculada a partir de los 99 en el trabajo de Avena et al. (2012). Para cada individuo, se promediaron los valores de ancestría obtenidos en las cinco corridas realizadas con cada conjunto de marcadores. Se estimó el coeficiente de correlación de Pearson con el programa Infostat, a fin de obtener una medida de la aproximación de las estimaciones a la obtenida con 99 AIMs. Debido a las múltiples comparaciones, para evaluar la significación de la correlación se empleó la corrección de Bonferroni (Rice 1989).

## **Resultados**

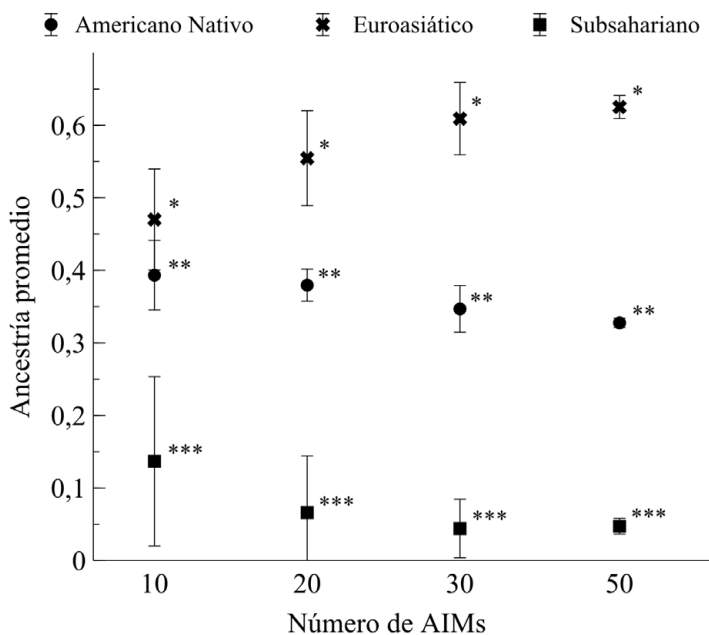
### *Variación de la ancestría individual en función del número de AIMs*

Considerando la ancestría promedio de la muestra, con todos los conjuntos de AIMs, se obtuvo una predominancia del componente euroasiático occidental, mientras que el subsahariano fue el minoritario. Más allá de esta tendencia general, para cada componente, los conjuntos de 10, 20, 30 y 50 marcadores difieren en la estimación de la ancestría individual ( $p < 0,0001$  en todos los casos). En particular, el valor de la ancestría euroasiática se incrementa a medida que aumenta el número de marcadores, en tanto que para los

Número de AIMs	Americano nativo		Euroasiático		Subsahariano	
	Media	Rango	Media	Rango	Media	Rango
10	0,393	0,332 - 0,444	0,470	0,393 - 0,533	0,137	0,027 - 0,271
20	0,380	0,352 - 0,406	0,555	0,442 - 0,607	0,066	0,030 - 0,206
30	0,347	0,325 - 0,404	0,609	0,552 - 0,649	0,044	0,020 - 0,115
50	0,328	0,319 - 0,337	0,625	0,607 - 0,649	0,047	0,032 - 0,058

**Tabla 1.** Ancestría media y rango (mínimo – máximo) obtenidos con diferentes conjuntos de AIMs para cada componente.

**Table 1.** Average ancestry and range (minimum – maximum) obtained with different set of AIMs for each component.



**Figura 1.** Ancestría media en función del número de marcadores. Se muestra el desvío estándar de las cinco estimaciones promediadas. \*: diferencias significativas al 5% para el componente euroasiático; \*\*: diferencias significativas al 5% para el componente americano nativo; \*\*\*: diferencias significativas al 5% para el componente subsahariano (análisis GLM; método de comparación DGC).

**Figure 1.** Average ancestry vs. number of markers. Standard deviation of five estimates is shown. \*: significant differences at 5% for Eurasian component; \*\*: significant differences at 5% for Native American component; \*\*\*: significant differences at 5% for Sub-Saharan component (GLM analysis; DGC comparison method).

*Correlación entre las estimaciones obtenidas con distinto número de AIMs*

Las estimaciones de ancestría individual realizadas con 10, 20, 30 y 50 AIMs se correlacionan con la obtenida a partir de

componentes americano nativo y subsahariano las estimaciones poseen un comportamiento inverso. Asimismo, tanto el rango como el desvío estándar de la estimación promedio disminuyen notablemente con el incremento en el número de AIMs (Tabla 1 y Figura 1).

99 ( $p < 0,002$ , luego de corrección por Bonferroni), para los tres componentes. Se destaca una tendencia hacia una mayor correlación a medida que aumenta el número de AIMs, evidenciada por una disminución en la dispersión de los puntos (Figura 3).

Por otro lado, para los componentes euroasiático y americano nativo, la variación entre individuos constituye el aporte mayoritario a la varianza total. Sin embargo, es importante destacar que, en los tres casos, el aporte a la variación dado por las diferencias entre las cinco corridas para cada individuo con cada conjunto de marcadores (error aleatorio) es considerable, especialmente en el caso del componente subsahariano, en el cual supera la variación entre individuos (Tabla 2).

Además, existe un efecto del componente estimado puesto que, si bien para el nativoamericano y el euroasiático todos los coeficientes de correlación son mayores a 0,900, para el componente subsahariano se alcanza una correlación de 0,700 sólo cuando se realiza la estimación con 30 AIMs y aumenta a 0,804 para 50 (Tabla 3).

*Variación en la precisión de la estimación en función del número de AIMs*

**Discusión**

Al analizar la amplitud promedio de los intervalos del 90% de probabilidad, se observó que, en líneas generales, mejora la estimación de los tres componentes de la ancestría a medida que aumenta el número de marcadores. Esto se evidencia en la disminución considerable tanto de la amplitud del intervalo ( $p < 0,001$ ) como de su desvío estándar (Figura 2).

Existen diversos factores que influyen en las estimaciones realizadas en estudios de ancestría individual. Los principales refieren a la cantidad de AIMs empleada, su capacidad informativa, la elección de las poblaciones parentales y el número de individuos que integran las mismas (e.g. Tsai et al. 2005). La importancia del número de marcadores utilizados consiste no solamente en su influencia sobre los resultados obtenidos, sino también porque allí radica mayormente el costo que conlleva este tipo de estudios.

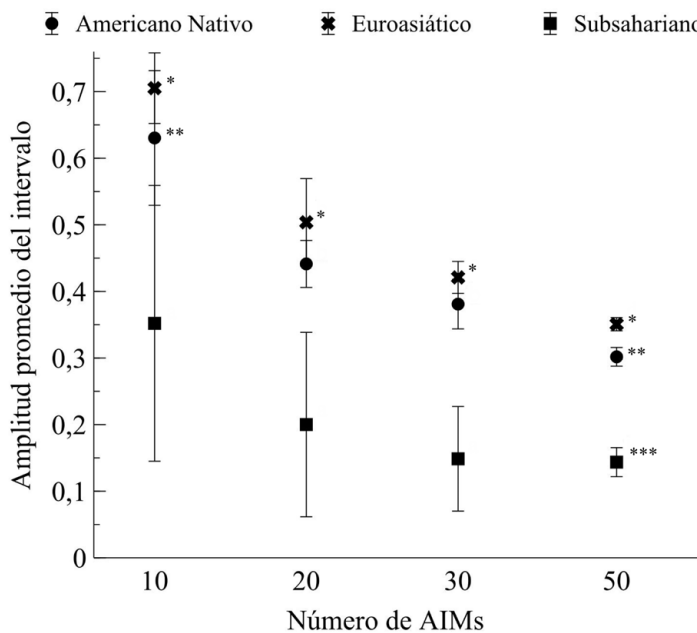
Componentes de varianza (%)	Americano Nativo	Euroasiático	Subsahariano
Individuos o variación "inter"	61,790	59,023	20,161
Error o variación "intra"	38,210	40,977	79,839

**Tabla 2.** Aporte porcentual a la varianza total de los términos aleatorios del modelo para cada componente. El error incluye la variación entre las cinco corridas. (Análisis GLM para el efecto del número de marcadores sobre la ancestría individual.)

**Table 2.** Percentage contribution to total variance of random factors for each component. The error term includes variation between the five estimates. (GLM analysis for the effect of marker number on individual ancestry.)

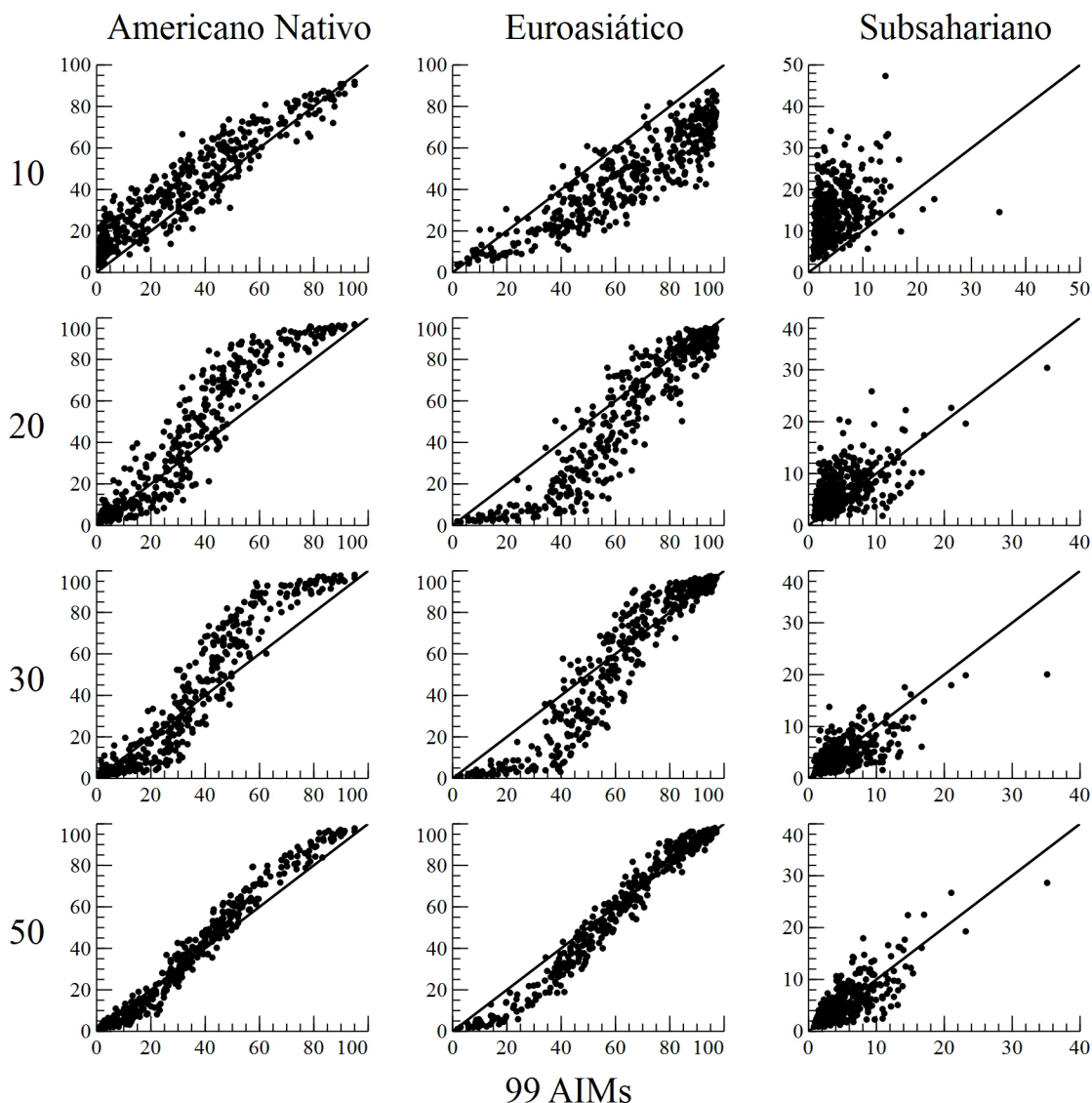
**Figura 2.** Amplitud media del intervalo de probabilidad en función del número de marcadores. Se muestra el desvío estándar de las cinco estimaciones promediadas. \*: diferencias significativas al 5% para el componente euroasiático; \*\*: diferencias significativas al 5% para el componente americano nativo; \*\*\*: diferencias significativas al 5% para el componente subsahariano (análisis GLM; método de comparación DGC).

**Figure 2.** Average width of probability interval vs. number of markers. Standard deviation of five estimates is shown. \*: significant differences at 5% for Eurasian component; \*\*: significant differences at 5% for Native American component; \*\*\*: significant differences at 5% for Sub-Saharan component (GLM analysis; DGC comparison method).



**Figura 3.** Ancestría individual (%) obtenida con cada conjunto de marcadores en función de la calculada a partir de 99 AIMs para los tres componentes.

**Figure 3.** Individual ancestry (%) obtained with each marker set vs. the panel of 99 AIMs for the three components.



Número de AIMs	Americano Nativo	Euroasiático	Subsahariano
10	0,937	0,901	0,387
20	0,948	0,937	0,603
30	0,956	0,950	0,700
50	0,988	0,985	0,804

**Tabla 3.** Coeficientes de Pearson para la correlación entre la ancestría individual calculada con cada conjunto de marcadores y la obtenida con los 99 AIMs. En todos los casos  $p < 0,002$  luego de corrección por Bonferroni.

**Table 3.** Pearson's coefficient for pairwise correlations between individual ancestry obtained with each marker set and for the panel of 99 AIMs.  $p < 0.002$  in every case after Bonferroni correction.

En este trabajo se observó una dependencia de las estimaciones de ancestría respecto al número de marcadores utilizados (Figura 1). El incremento del número de AIMs aumenta la precisión de las estimaciones, puesto que se reduce el desvío estándar, hecho que se observó en las cinco corridas independientes (Figura 1), y disminuye la amplitud promedio de los intervalos de probabilidad para la ancestría individual (Figura 2).

Si bien todas las estimaciones de ancestría promedio varían a medida que aumenta el número de marcadores, al utilizar 30 AIMs se obtiene una aceptable correlación con la estimación realizada a partir de 99 marcadores para los tres componentes (Figura 3 y Tabla 3). Por otro lado, las estimaciones realizadas con 50 AIMs deben tomarse con precaución debido a que, al realizar cinco muestreos sobre un total de 99 marcadores, existe una alta probabilidad de que gran parte de los marcadores se repitan en cada conjunto. En efecto, la frecuencia promedio de repetición de un marcador en las corridas ha sido de 1,85 para 30 AIMs y 2,69 para 50. De esta forma, es esperable que disminuya notablemente la variación entre diferentes corridas realizadas con 50 AIMs (Figuras 1 y 2) y, por lo tanto, la diferencia significativa entre las estimaciones con 30 y 50 marcadores podría deberse en parte a esa disminución en la variabilidad y no a una diferencia real en la ancestría promedio en este caso.

Es importante destacar que tanto el componente euroasiático como el americano nativo presentaron un comportamiento diferencial con respecto al subsahariano, principalmente en los análisis de correlación. Esta diferencia radica en que solamente para este último se evidenció una variación notable en las correlaciones entre la estimación con 99 AIMs y las realizadas con 10, 20, 30 y 50 (Tabla 3). Es decir, mientras que para las estimaciones de ancestría euroasiática y nativoamericana ya se obtienen elevadas correlaciones con 10 AIMs; la estimación del componente subsahariano sólo resulta aceptable a partir de 30. Una posible explicación a este fenómeno se halla en el método empleado, el cual podría estar estimando el componente minoritario con un mayor error, especialmente cuando se emplea una menor cantidad de marcadores. Esto se condice con el hecho de que el único caso para el cual se observó que la

varianza entre las cinco corridas supera a la variación entre individuos corresponde al componente subsahariano (Tabla 2).

Tsai et al. (2005) proponen el uso de al menos 100 marcadores para una estimación correcta. Sin embargo, los autores también indican que el número necesario disminuye a medida que aumenta la capacidad informativa, puesto que al usar AIMs con una diferencia de más de 0,5 entre las frecuencias alélicas de las poblaciones parentales, se requerirían menos de 50 marcadores para mejorar la exactitud de las estimaciones. Turakulov y Easteal (2003), si bien recomiendan el uso de por lo menos 65 marcadores para obtener resultados confiables, observaron que existen ciertas combinaciones de 30, para los cuales se obtienen buenas estimaciones. Estos trabajos refuerzan la idea de la alta dependencia que existe con la capacidad informativa de los AIMs seleccionados. Por ejemplo, considerando un modelo tripartito, si las frecuencias alélicas son opuestas sólo entre dos de las poblaciones parentales y la tercera posee valores intermedios, una muestra de individuos con influencia de esta última población únicamente resultaría indistinguible de otra con influencia de las dos poblaciones restantes en igual proporción. Idealmente, para este tipo de estudios se deberían seleccionar aquellos marcadores cuyas frecuencias alélicas posean una diferencia mayor al 40% entre una población parental respecto a las otras dos, siendo, a su vez, la diferencia entre estas dos restantes menor al 5%. Es importante destacar que, como consecuencia del muestreo aleatorio, en el presente trabajo los marcadores con mayor capacidad informativa se encuentran distribuidos homogéneamente entre los distintos conjuntos de AIMs empleados. Con lo cual, en este caso, su capacidad informativa no explicaría la mayor parte de las diferencias encontradas entre las distintas corridas realizadas sino que, como ya se ha mencionado, el efecto principal se debe a la cantidad de marcadores empleados.

Sin embargo, no se trata de una cuestión meramente cuantitativa, sino también cualitativa, siendo recomendable la selección de marcadores apropiados para el estudio de cada población en particular. Por ejemplo, debe considerarse si el proceso de mestizaje tuvo

influencia de dos o tres poblaciones parentales e incluso utilizar parentales más específicas acorde a cada historia demográfica (en el caso argentino, por ejemplo, lo ideal serían americanos nativos del Cono Sur, europeos del sur y subsaharianos del centro-oeste de África y Mozambique).

En este trabajo se plantea el uso de 30 AIMs como un mínimo aceptable, lo cual concuerda con lo observado por Ruiz-Narváez et al. (2010). Estos autores, tomando diferentes cantidades (30, 60, 100, 150 y 200) de un total de 1.757 AIMs, encontraron una alta correlación de estimación (0,89) utilizando 30 marcadores. Adicionalmente, un trabajo reciente (Ruiz-Linares et al. 2014), evaluó la ancestría individual en poblaciones de Colombia con un panel de 50.000 marcadores, hallando una correlación de 0,7 con 30 AIMs para los tres componentes, y valores idénticos para las estimaciones poblacionales. En este último caso, el componente minoritario (subsahariano) posee una frecuencia mucho mayor que en la muestra analizada en el presente trabajo, lo cual podría explicar que las correlaciones para los tres componentes se hayan visto afectadas de manera similar. De esto se desprende que el número mínimo de AIMs con los que se obtiene una estimación aceptable de ancestría individual también depende de la frecuencia de todos los componentes en la población mestizada, en particular del menos representado.

Finalmente, se destaca la importancia de reflexionar en torno a aspectos metodológicos. Mientras que en disciplinas tales como morfometría geométrica esta temática ha sido tradicionalmente abordada (González et al. 2004; González-José et al. 2011, entre muchos otros), en antropología molecular la discusión metodológica se ha incrementado sensiblemente a partir del desarrollo de los estudios de ancestría individual (e.g. Ruiz-Narváez et al. 2010; Tsai et al. 2005; Turakulov y Easteal 2003), siendo su tratamiento crucial puesto que las estimaciones de dicha ancestría son altamente dependientes de la metodología empleada.

## Conclusiones

En este trabajo pudo observarse que las estimaciones de ancestría individual tienen una confiabilidad aceptable utilizando, como mínimo, 30 marcadores seleccionados sobre la base de su poder discriminatorio; siendo preferible el empleo de aquellos que presenten mayores diferencias en las frecuencias alélicas entre las poblaciones parentales.

El hecho de que las estimaciones de ancestría individual dependan en gran medida del número de AIMs utilizados remarca la importancia de conocer la cantidad de marcadores necesarios antes de emprender estudios de este tipo. Adicionalmente, las características particulares de cada población analizada en relación a su estructura genética, y especialmente la presencia relativa de

los componentes a nivel continental, deberían ser consideradas en la elección de los marcadores apropiados.

Ciudad Autónoma de Buenos Aires,  
29 de febrero de 2016

## Agradecimientos

Los autores deseamos agradecer a la Lic. María Bárbara Postillone por los comentarios realizados a una versión previa del manuscrito y a los revisores, cuyos comentarios contribuyeron a mejorar sustancialmente este trabajo. Agradecemos el apoyo financiero del Subsidio intramuros de la Fundación Científica Felipe Fiorellino, de la Fundación de Historia Natural Félix de Azara y de CONICET.

## Bibliografía

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. doi:10.1109/TAC.1974.1100705
- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655–1664. doi:10.1101/gr.094052.109
- Avena, S., Via, M., Ziv, E., Pérez-Stable, E. J., Gignoux, C. R., Dejean, C., Huntsman, S., Torres-Mejía, G., Dutil, J., Matta, J. L., Beckman, K., Burchard, E. G., Parolin, M. L., Goicoechea, A., Acreche, N., Boquet, M., Ríos Part, M. C., Fernández, V., Rey, J., Stern, M. C., Carnese, R. F. & Fejerman, L. (2012). Heterogeneity in genetic admixture across different regions of Argentina. *PLoS ONE*, 7(4), e34695. doi:10.1371/journal.pone.0034695
- Bonilla, C., Bertoni, B., Hidalgo, P. C., Artagaveytia, N., Ackermann, E., Barreto, I., Cancela, P., Cappetta, M., Egaña, A., Figueiro, G., Heinzen, S., Hooker, S., Román, E., Sans, M. & Kittles, R. A. (2015). Breast cancer risk and genetic ancestry: a case-control study in Uruguay. *BMC Women's Health*, 15(1), 1–10. doi:10.1186/s12905-015-0171-8
- Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., ... Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *The New England Journal of Medicine*, 348(12), 1170–1175. doi:10.1056/NEJMs025007
- Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J. & Cavalli-Sforza, L. L. (2002). A Human Genome Diversity Cell Line Panel. *Science*, 296(5566), 261–262.
- Corach, D., Lao, O., Bobillo, C., van Der Gaag, K., Zuniga, S., Vermeulen, M., van Duijn, K., Goedbloed, M., Vallone, P. M., Parson, W., de Knijff, P. & Kayser, M. (2010). Inferring Continental Ancestry of Argentineans from Autosomal, Y-Chromosomal and Mitochondrial DNA. *Annals of Human Genetics*, 74(1), 65–76. doi:10.1111/j.1469-1809.2009.00556.x

- Di Rienzo, J. A., Guzman, A. W., & Casanoves, F. (2002). A multiple-comparisons method based on the distribution of the root node distance of a binary tree. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(2), 129–142. doi:10.1198/10857110260141193
- Di Rienzo, J. A., Casanoves, F., Balzarini, M. G., Gonzalez, L., Tablada, M., & Robledo, C. W. (2013). InfoStat versión 2013. Córdoba: Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. <http://www.infostat.com.ar>
- García, A., Tovo-Rodrigues, L., Pauro, M., Callegari-Jacques, S. M., Salzano, F. M., Hutz, M. H., & Demarchi, D. A. (2011). Caracterización del mestizaje en poblaciones del centro de Argentina a partir de marcadores moleculares informativos de ancestralidad (AIM). En M. F. Cesani (Ed.), *Libro de Resúmenes de las Décimas Jornadas Nacionales de Antropología Biológica* (p. 136). City Bell: Asociación de Antropología Biológica Argentina.
- González, P., Bernal, V., Perez, S., Del Papa, M., Gordon, F., & Ghidini, G. (2010). El error de observación y su influencia en los análisis morfológicos de restos óseos humanos. Datos de variación discreta. *Revista Argentina de Antropología Biológica*, 6(1), 35-46.
- González-José, R., Escapa, I., Neves, W. A., Cúneo, R., & Pucciarelli, H. M. (2011). Morphometric variables can be analyzed using cladistic methods: A reply to Adams et al. *Journal of Human Evolution*, 60(2), 244–245.
- Heinz, T., Alvarez-Iglesias, V., Pardo-Seco, J., Taboada-Echalar, P., Gómez-Carballa, A., Torres-Balanza, A., Rocabado, O., Carracedo, A., Vullo, C. & Salas, A. (2013). Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. *Forensic Science International: Genetics*, 7(5), 537–542. doi:10.1016/j.fsigen.2013.05.012
- Peprah, E., Xu, H., Tekola-Ayele, F., & Royal, C. D. (2015). Genome-Wide Association Studies in Africans and African Americans: Expanding the Framework of the Genomics of Human Traits and Disease. *Public Health Genomics*, 18(1), 40–51.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2015). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3. <http://cran.r-project.org/package=nlme>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.r-project.org/>
- Rice, W. R. (1989). Analyzing Tables of Statistical Tests. *Evolution*, 43(1), 223–225.
- Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1(6), e70.
- Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Burley, M. W., Konca, E., de Oliveira, M. Z., Veronez, M. R., Rubio-Codina, M., Attanasio, O., Gibbon, S., Ray, N., Gallo, C., Poletti, G., Rosique, J., Schuler-Faccini, L., Salzano, F. M., Bortolini, M. C., Canizales-Quinteros, S., Rothhammer, F., Bedoya, G., Balding, D. & Gonzalez-José, R. (2014). Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genetics*, 10(9):e1004572.
- Ruiz-Narváez, E. A., Rosenberg, L., Wise, L. A., Reich, D., & Palmer, J. R. (2011). Validation of a Small Set of Ancestral Informative Markers for Control of Population Admixture in African Americans. *American Journal of Epidemiology*, 173(5), 587–592. doi:10.1093/aje/kwq401
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461–464.
- Shriver, M.D., Parra, E. J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N., Baron, A., Jackson, T., Argyropoulos, G., Jin, L., Hoggart, C. J., McKeigue, P. M. & Kittles, R. A. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. *Human Genetics*, 112(4), 387–399.
- Tsai, H.-J., Choudhry, S., Naqvi, M., Rodriguez-Cintron, W., Burchard, E. G., & Ziv, E. (2005). Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. *Human Genetics*, 118(3), 424–433.
- Turakulov, R., & Easteal, S. (2003). Number of SNPs loci needed to detect population structure. *Human Heredity*, 55(1), 37–45.
- Zhu, X., & Cooper, R. S. (2007). Admixture Mapping Provides Evidence of Association of the VNN1 Gene with Hypertension. *PLoS ONE*, 2(11), e1244.
- Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer New York.