Universidad CAECE

Trabajo de seminario

Análisis de un panel de marcadores de información de ancestría para la caracterización de poblaciones latinoamericanas

Tesista
Juan Manuel Berros
Universidad CAECE

Director
Dr. Hernán Dopazo
Investigador independiente IEGEBA-CONICET
Profesor Adjunto FCEyN-UBA

Lugar de trabajo
Instituto de Ecología, Genética y Evolución de Buenos Aires.
Universidad de Buenos Aires

Agosto 2016

Índice general

Abreviaturas utilizadas en este trabajo

- AFR: ancestría africana.
- AIM: marcador informativo de ancestría.
- CPx1: Control Panel x 1 (438 SNPs).
- CPx10: Control Panel x 10 (4.424 SNPs).
- CPx100: Control Panel x 100 (43.144 SNPs).
- EUR: ancestría europea.
- F_{ST}: Índice de fijación (*fixation index*).
- `GAL_Affy`: Panel producto de la intersección entre los marcadores presentes en `GAL_Completo` y los marcadores que forman parte del *microarray* LAT-1.
- `GAL_Completo`: Panel de AIMs para latinoamericanos diseñado por Galanter *et al.* (2012) [1].
- LAT-1: *Microarray* de SNPs de la serie *Axiom World Arrays*, diseñado por Affymetrix para genotipado de latinoamericanos.
- L: dataset de poblaciones latinoamericanas.
- LE: dataset de poblaciones latinoamericanas y europeas.
- LEA: dataset de poblaciones latinoamericanas, europeas y africanas.
- LEAC: dataset de poblaciones latinoamericanas, europeas, africanas e indias.
- LEACI: dataset de poblaciones latinoamericanas, europeas, africanas, indias y chinas.
- LSBL: *locus-specific branch length*. En un dendograma, la distancia entre nodos debida a la información genética calculada en base al genotipo de un *locus* esecífico.
- MAF: frecuencia del alelo menor.
- NAM: ancestría nativoamericana.
- PCA: análisis de componentes principales.
- SNP: polimorfismo de un solo nucleótido.

1 Introducción

1.1. Proyecto Genoma Humano, HapMap, 1000 Genomas

Durante los últimos 15 años se produjo un gran crecimiento de la cantidad de datos genómicos de poblaciones humanas. Una vez finalizado el Proyecto Genoma Humano y la secuenciación de varios genomas individuales, el esfuerzo internacional se enfocó en la descripción de las variantes presentes en diferentes poblaciones, con el fin de poder describir haplotipos comunes en distintas regiones del mundo.

El International SNP Map Working Group [2], inicialmente, y luego el proyecto HapMap, generaron entre 2001 y 2008 un mapa de haplotipos a nivel global que facilitaron el diseño de estudios de asociación de genomas completos [3, 4, 5, 6]. En sus dos primeras fases, HapMap se encargó de catalogar SNPs comunes, con una frecuencia del alelo menor (MAF) mayor al 5 %. Como una continuación de este proyecto, desde 2010 el Proyecto 1000 Genomas busca catalogar no sólo SNPs sino también otros tipos de variantes, y además incluye variantes de baja frecuencia, con MAF entre 1 % y 5 %, a lo largo de todo el genoma. Adicionalmente, 1000 Genomas reporta también variantes raras, de MAF entre 0,1 % y 1 %, en regiones codificantes. En su fase 3, este proyecto llegó a describir más del 99 % de los SNPs con MAF $> 1\,\%$ [7].

El Proyecto 1000 Genomas, además, amplió a 26 el número de poblaciones incluidas inicialmente por HapMap [8, 9]. Entre ellas, se incluyeron cuatro poblaciones latinoamericanas relevantes para nuestro trabajo: puertorriqueños de Puerto Rico (PUR), colombianos de Medellín (CLM), peruanos de Lima (PEL) e individuos de ancestría mexicana en Los Ángeles (MXL).

En conjunto, estos proyectos de genómica humana generaron un panorama revelador sobre la diversidad de nuestra especie y sus orígenes.

1.2. Migraciones humanas

Los datos genéticos apoyan un modelo de expansión poblacional rápida desde África (estimada en 15 km por generación) a partir de 50 a 60 mil años atrás. La migración se realizó en todas direcciones en Eurasia y llegó en unas decenas de miles de años a todos los continentes, con la excepción de Antártida (**Figura 1.1**). Humanos anatómicamente modernos poblaban África desde mucho antes —hace 130 mil años—, pero no fue sino hasta hace 50 mil años que la llamada "Gran Expansión" comenzaría. La explicación para este "súbito" poblamiento del planeta involucraría una posible acumulación crucial de avances culturales, acaso acompañada de cambios neuroanatómicos no reflejados en el registro fósil [10].

El modelo del fundador serial (**Figura 1.2b**) explica la pérdida de diversidad genética que se observa a medida que aumenta la distancia respecto de África [11]. Este modelo postula que las migraciones ocurren cuando una subpoblación se desprende de la población original para colonizar nuevas regiones. La consecuencia de este patrón migratorio no sólo se ve reflejada en un declive de la diversidad genética, sino también en la pérdida de *diversidad fonémica* en los lenguajes a medida que aumenta la distancia desde África [12].

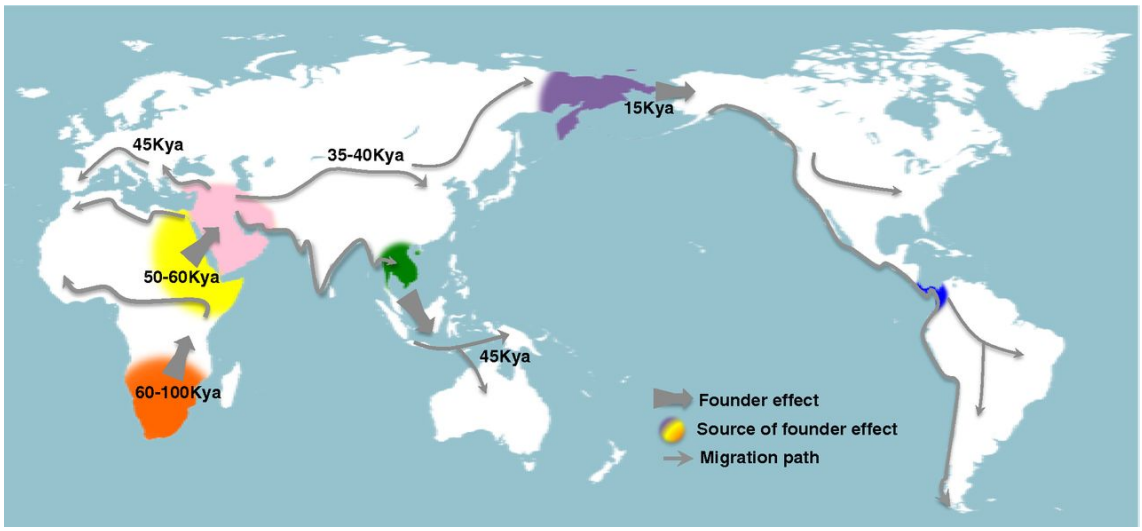
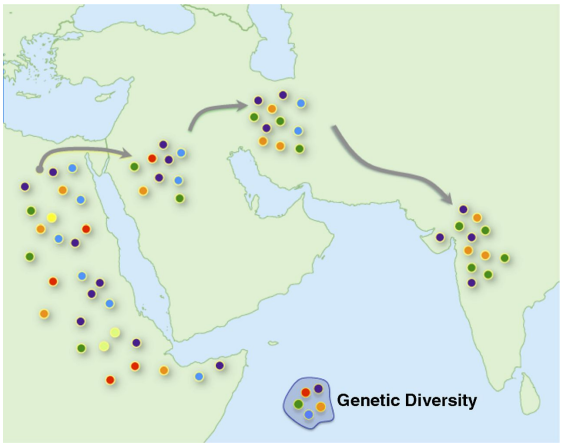
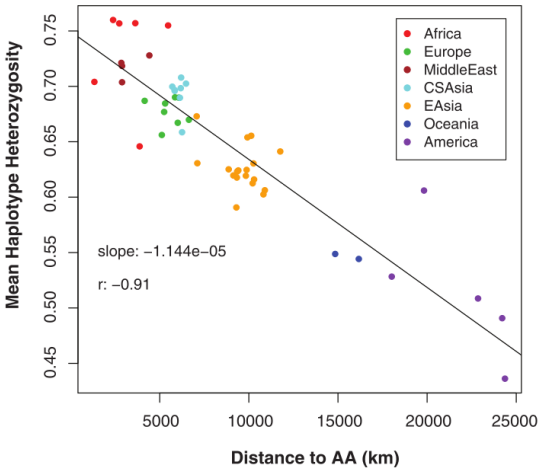


Figura 1.1: La Gran Expansión durante los últimos 100 mil años. Las flechas anchas representan eventos fundador; las flechas delgadas, migraciones; los colores, la fuente de cada evento fundador. Tomado de [10].



(a) Esquema de un modelo de fundador serial. Se ilustra el efecto de cada evento fundador sobre la diversidad genética —la pérdida de alelos. Tomado de [10].



(b) La heterocigosidad de haplotipos de SNPs decrece en función de la distancia desde Addis Ababa (AA), Etiopía. Tomado de [11].

Figura 1.2: Pérdida de diversidad genética al alejarse de África.

1.3. Poblamiento de las Américas

América fue el último continente en ser poblado por la Gran Expansión de los humanos anatómicamente modernos [13]. El descubrimiento de América fue posibilitado por la existencia de Beringia, una masa terrestre que conectó Asia y Norteamérica entre 60 y 10 mil años atrás. El registro fósil sugiere que se trataba de una pradera seca y productiva, habitada por grandes mamíferos, lo que la muestra como una región que podía dar sustento al menos a poblaciones humanas pequeñas [14].

La evidencia genética basada en haplotipos mitocondriales [14] da apoyo a la hipótesis de que hace alrededor de 30 mil años un subconjunto de la población del norte de Asia se estableció en Beringia y se detuvo allí, donde se diferenciaría durante unos 15 mil años.

Esta información es consistente con el hecho de que, si bien Beringia estaba conectada con lo que hoy es Alaska, se encontraba aislada del resto de Norteamérica por masas de hielo que cubrieron la región hasta hace ∼16 mil años, de modo que los proto-americanos de Beringia no pudieron migrar más allá de esa región.

Sólo hacia el final de la última glaciación se abriría un corredor libre de hielo que llevaba hacia el resto del continente (**Figura 1.3**, tercer cuadro). Alrededor de 3 mil años más tarde, el nivel del mar crecería lo suficiente como para reinundar Beringia, creando el estrecho de Bering que hoy separa a Siberia de América con por lo menos 100 kilómetros de agua helada. Estos dos fenómenos crean una ventana temporal definida (entre 15 y 11 mil años atrás) durante la cual los fundadores americanos habrían entrado al continente [15].

Los últimos trabajos genéticos apoyan la hipótesis de que al menos tres oleadas migratorias independientes entraron a América a través de Beringia en ese lapso temporal. La mayoría de los latinoamericanos mestizos actuales deben su ancestría nativa a la primera oleada migratoria, de los llamados "Primeros Americanos". Tras una rápida expansión hacia el sur, las poblaciones derivadas de estos fundadores dieron lugar a la mayor parte de los pueblos nativos del continente. Dos migraciones posteriores desde Beringia darían origen a poblaciones de Chippewan, Saqqaq y Paleo-Eskimos en el norte de Norteamérica, si bien con una gran proporción de mestizaje con los Primeros Americanos [16, 17].

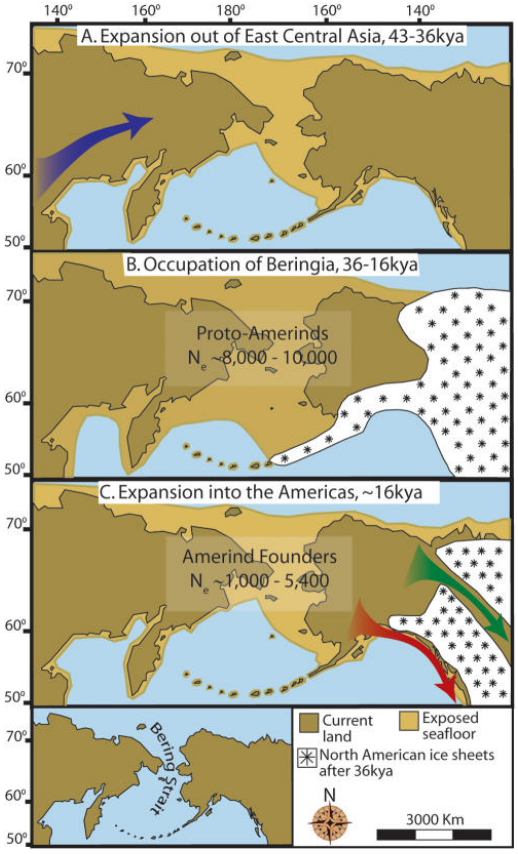


Figura 1.3: Ingreso de los primeros pobladores humanos en América, en tres etapas: expansión a Asia, diferenciación en Beringia y expansión posterior hacia el continente americano. Tomado de [15].

En el presente, se observa como resultado que los nativos americanos son genéticamente más similares a los asiáticos del norte que a otras poblaciones del mundo, menos diversos que las poblaciones de otros continentes y que la heterocigosidad de sus poblaciones decrece hacia el sur [18].

En resumen, el poblamiento de América tuvo su origen en una población de asiáticos que se establecieron previamente en Beringia y se diferenciaron durante 15 mil años. Tras el retraimiento de las capas de hielo en Norteamérica, los "Primeros Americanos" migraron rápidamente desde Beringia por todo el continente, dando lugar a una secuencia de contracciones poblacionales que redujeron progresivamente la diversidad genética. El árbol filogenético de poblaciones nativas americanas es consistente con este relato y los grupos genéticos corresponden de manera general con las familias lingüísticas [16]. Las poblaciones nativas que se formaron a partir del descubrimiento de América conforman el sustrato genético con el que se mezclaron, en los últimos siglos, las oleadas migratorias de la invasión europea y del tráfico de esclavos.

1.4. Latinoamericanos modernos

Los tres acervos genéticos a partir de los cuales se conformó la población latinoamericana contemporánea estaban aislados hasta su encuentro. Por un lado, la población nativa americana y la europea no se encontraron en números importantes hasta la invasión de América iniciada en 1492. A partir de ese momento, se inició un flujo génico constante de europeos —en particular de la Península Ibérica— hacia América, compuesto principalmente por población masculina (como se infiere de la importante diferencia de ancestrías entre los autosomas y los cromosomas X, Y y el ADN mitocondrial [20]).

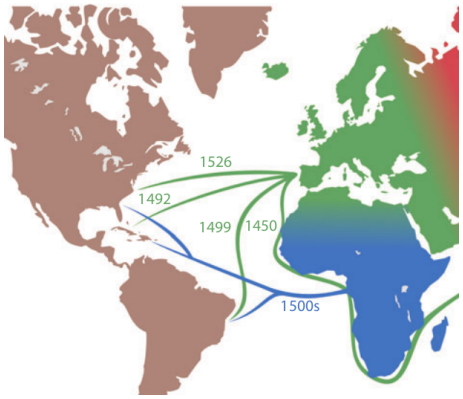


Figura 1.4: Comienzo del flujo migratorio hacia América a partir de la invasión europea. Tomado de [19].

Por otro lado, grandes números de habitantes de África fueron llevados forzosamente a América, en particular en los siglos XVII, XVIII y XIX, como mano de obra esclava para plantaciones de azúcar y café y para minería, entre otras actividades. Se estima en 11,3 millones la cantidad de esclavos transportados en total, con Brasil como principal destino (4 millones) y la América española en segundo lugar (2,5 millones), particularmente las Antillas [21] (**Figura 1.4**).

El flujo migratorio de europeos y africanos no fue uniforme a lo largo del continente. Por ende, en la actualidad hay en Latinoamérica una considerable heterogeneidad genética, ya sea entre distintas regiones de Sudamérica [22], de Centroamérica [23] y en América continental en contraste con las islas del Caribe, así como entre las diferentes islas del Caribe [24]. Esta heterogeneidad depende de numerosos factores, como diferencias históricas en el grado de inmigración desde Europa y de importación de esclavos desde África, la densidad de las poblaciones americanas nativas y la duración del flujo génico.

A pesar de esta complejidad, la misma imagen surge repetidamente: nuestro continente se caracteriza por una gran variación en las proporciones de ancestría americana, europea y africana, con diferencias importantes entre las poblaciones y entre los individuos al interior de cada población.

A continuación resumiremos los principales estudios realizados en los cuatro países relevantes para nuestro trabajo, que fueron recopilados por Salzano y Sans [25]: Perú, Colombia, México y Puerto Rico.

1.4.1. Perú

Perú exhibe valores de ancestría indígena que varían entre 67 % y 98 % entre poblaciones, según un estudio de basado en 40 AIMs [25]. En promedio, la mayoría de las regiones investigadas exhiben una proporción nativa mayor al 90 %. La ciudad de Lima, en particular, tiene un 84 % de componente indígena y un 14 % europeo. En el resto de Perú, la ancestría europea supera el 10 % sólo en algunos casos, mientras que la ancestría africana se mantiene consistentemente baja —entre el 1 % y el 3 % en todas las regiones investigadas.

1.4.2. Colombia

En Colombia, la ancestría indígena oscila ampliamente entre 10 % y 65 % y lo mismo se observa para la ancestría europea. En Medellín en particular, la contribución genética europea está entre el 60 % y el 66 %, por encima de otras partes del país. En general, la variación en el país es demasiado grande como para extraer conclusiones generales, como

sugieren Salzano y Sans [25]. La contribución africana varía entre 1 % y 25 %, pero es mucho mayor en muestras de Antioquía, Cartagena y Quibdó. Los estudios recopilados en el trabajo citado utilizaron principalmente AIMs (entre 11 y 75 marcadores según el trabajo).

En el resto del Sudamérica el paisaje de mestizaje se repite: se encuentra una y otra vez un extenso y muy variable mestizaje de componentes americano y europeo, con una proporción africana en general baja.

1.4.3. México

En México, a partir de 19 reportes listados por Salzano y Sans [25], se observa que el componente indígena es predominante con valores promedio entre 51 % y 56 %, seguido por el componente europeo (40 % a 45 %) y el africano (2 % a 5 %). En algunas regiones particulares, como la ciudad de México, Veracruz y Yucatán, la ancestría nativa americana asciende a valores entre 64 % y 70 % y llega hasta 95 % en un estudio de 156 individuos de Guerrero [26]. En estos trabajos, los marcadores utilizados también fueron mayoritariamente AIMs (desde 24 [26] hasta 1.814 [23]).

Al considerar los patrones de mestizaje en Mesoamérica y Sudamérica, se puede extraer la conclusión señalada por Wang *et al.* [27]: hay una congruencia entre la estructura genética precolombina de cada región y la composición genética actual de los latinoamericanos. Las poblaciones nativas que predominaban en cada territorio antes de la invasión europea dejaron una mayor huella genética en sus propias regiones que en otras partes del continente, lo que tiene especial relevancia en México y Perú, las regiones que antes de la invasión tenían mayor densidad poblacional y más diversidad genética.

1.4.4. Puerto Rico

El componente africano crece notoriamente en los países de las Antillas menores y del Caribe, donde la población taína nativa se redujo drásticamente tras el contacto con los europeos [28], lo que motivó un gran influjo de esclavos africanos como mano de obra. Éste es el caso de Puerto Rico, donde se observa una ancestría africana mayor que en los países hasta ahora mencionados, con valores entre el 21 % y el 32 % (estudio de Via *et al.* utilizando 93 AIMs [28]).

En conjunto, la diversidad genética observada sugiere que el mestizaje entre americanos, europeos y africanos ocurrió en distintas medidas en todo el continente. Este complejo proceso demográfico involucró a numerosas poblaciones indígenas en diferentes grados y a distintas oleadas migratorias tanto de Europa como de África durante cinco siglos. Consecuentemente, los latinoamericanos conforman un continuo de distintos grados de mestizaje con gran variación entre los distintos países y entre las regiones al interior de cada país.

1.5. El concepto de ancestría

Es importante aclarar que lo que denominamos componente ancestral o ancestría implica una simplificación. El componente ancestral africano en latinoamericanos, por ejemplo, tiene una heterogeneidad interna que reducimos al utilizar ese rótulo unificador. Los esclavos de África que llegaron a América provenían de diferentes grupos étnicos de aquel continente: a pulsos migratorios iniciales de Mandenka y Brong le siguieron crecientes contribuciones posteriores de Kongo, Bamoun, Igbo, Fang y Yoruba [24].

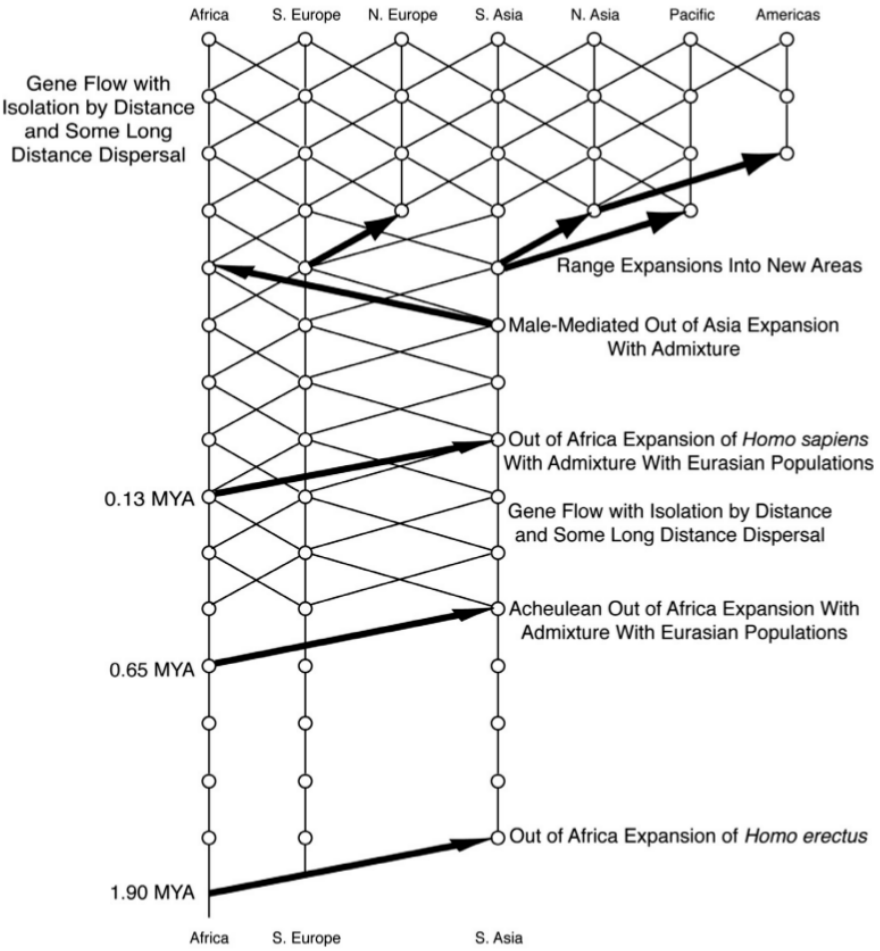


Figura 1.5: Gráfico *trellis* ("enrejado") tomado de [29]. Las flechas verticales indican descendencia genética y las diagonales flujo génico, con las migraciones más importantes en flechas gruesas. Se aprecia que no hay discontinuidades o linajes separados en las poblaciones humanas, conclusión simplificada a la que contribuyen los dendogramas con los que se suele ilustrar la genealogía de nuestra especie.

Lo mismo puede decirse de lo que llamamos componente ancestral europeo. Europa está compuesta por un gradiente de diversidad genética formado por numerosos eventos a lo largo de 40,000 años [30]: el poblamiento inicial de cazadores-recolectores desde el noreste de África, las reducciones poblacionales subsiguientes durante el Último Máximo Glacial en el Pleistoceno, la repoblación del norte del continente, una nueva oleada migratoria durante el inicio del Neolítico (los "agricultores"), las llamadas "invasiones bárbaras" al Imperio Romano y la ocupación musulmana en la Península Ibérica durante siete siglos. No sin razón se ha llamado al resultado genético europeo un palimpsesto de marcas genéticas solapadas unas sobre las otras.

Finalmente, también hay heterogeneidad en el componente ancestral americano, que no sólo varía en proporción entre individuos, sino que también se diferencia internamente en un componente andino (presente en quechuas y aymaras), otro que va desde la Amazonia hasta el norte Argentino (incluyendo a pueblos como los parakana, ticuna, guaraní, wichí) y un componente mesoamericano (que incluye a mayas, mixtecas, tepehuanos) [22]. Estos "subcomponentes" tienen, a la vez, su propia complejidad interna, como ilustra un trabajo de Moreno-Estrada *et al.* sobre México [31].

Así pues, la compleja historia migratoria y demográfica, y la diversidad genética resultante en cada continente, es simplificada al hablar de componentes ancestrales. Es sólo una

cuestión de conveniencia o de interés de la investigación decidir un nivel de diferenciación donde detenerse para el análisis; en este trabajo nos detenemos en el nivel continental. Una investigación reciente, por ejemplo, ha llegado a distinguir genéticamente, en distintas regiones de Europa, entre habitantes de pueblos cercanos cuyas generaciones previas no habían migrado [32]. Esta capacidad de encontrar estructura genética a diferentes niveles, a cada paso exponiendo mayor detalle, ha sido denominada por John Novembre como "la naturaleza fractal de la estructura poblacional humana" [33].

Teniendo en cuenta que deternernos en el nivel continental es una decisión sobre qué queremos investigar y no un reflejo de límites bien definidos entre las poblaciones, no debemos confundir el concepto utilitario de componente ancestral a nivel continental con la idea desacreditada de "raza" como categoría discreta. Las poblaciones ancestrales previas al mestizaje que daremos por sentadas en este trabajo no son tipos "puros"; su heterogeneidad genética se remonta aun más hacia el pasado en la genealogía humana (véase una ilustración en el *trellis* de la **Figura 1.5**). Al hablar de ancestría, damos una idea de trazo grueso sobre el origen genético-geográfico de una parte de los ancestros de un individuo en el "gran pedigree humano", pero esta idea general sobre el origen geográfico no da lugar a un set de categorías discretas para catalogar personas.

Como nota final, cabe recordar que, además, el concepto de "raza" no se puede aplicar a la especie humana en ningún sentido biológico, dado que los humanos conforman un continuo genético sin saltos bruscos entre grupos, tampoco hay linajes evolutivos separados —sino un origen común reciente en África— y las diferencias genéticas que existen no son lo suficientemente marcadas para ameritar una clasificación de ese tipo, como sí ocurre, por ejemplo, en los chimpancés [29].

1.6. Mestizaje y genoma-mosaico

Como consecuencia del continuo genético que caracteriza a nuestra especie, nos encontramos con el problema de que no existen "genotipos diagnósticos" que se encuentren en una población y que estén completamente ausentes en otras. Tanto a nivel global como regional, las poblaciones se diferencian por la acumulación de variaciones pequeñas en las frecuencias alélicas, lo que hace necesario considerar muchos *loci* en simultáneo [34].

A esta complejidad se añade un fenómeno relativamente reciente. La diáspora humana que ocurrió durante los últimos 400 a 600 años resultó en un flujo génico entre poblaciones antes separadas entre sí [19]. Se utiliza el término *admixture* (que traduciremos como "mestizaje") para describir la formación de una población híbrida a partir del encuentro entre poblaciones que habían estado relativamente aisladas hasta ese momento [17]. En ese sentido, la mayoría de los latinoamericanos contemporáneos son mestizos —descendientes, en proporciones que varían ampliamente entre individuos, de ancestros nativos americanos, europeos y africanos, principalmente.

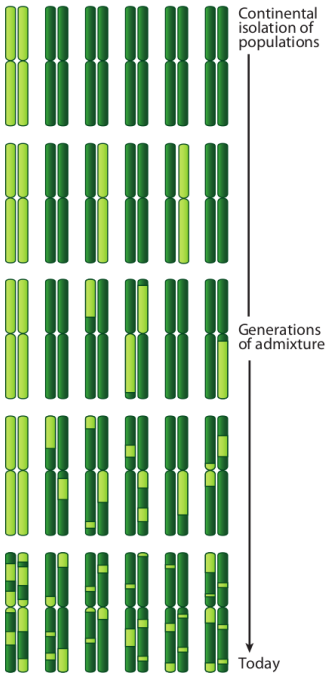


Figura 1.6: Generación de un "genoma-mosaico" en las generaciones que siguen al encuentro entre poblaciones previamente aisladas. Tomado de [19].

Al examinar genomas contemporáneos no se observa directamente la proporción de mestizaje establecida en el primer encuentro entre poblaciones, sino la suma acumulada del flujo génico que existió durante siglos entre las distintas poblaciones. Con el paso de las generaciones desde el encuentro, la recombinación produjo genomas que son un mosaico

de segmentos cromosómicos derivados de una o de otra población ancestral [19] (**Figura 1.6**).

La tarea de inferir la ancestría de un individuo mestizo, entonces, se transforma en una investigación sobre el origen de cada segmento cromosómico, que a la vez depende de la información que pueda extraerse de cada *locus*. El estudio recae en las variantes que tiene el genoma en los *loci* sean polimórficos entre poblaciones, puesto que para esas variantes se podrá inferir un origen con cierta probabilidad.

1.7. Marcadores informativos de ancestría (*AIMs*)

La precisión con la cual se puede inferir la ancestría de un *locus* específico del genoma depende de la diferencia entre frecuencias alélicas para ese *locus* entre las poblaciones ancestrales consideradas. En el caso extremo y raro donde un alelo está fijado en una población y ausente en la otra ($F_{ST} = 1$), al observar ese alelo en un genoma mestizo podremos deducir que fue heredado de la población que lo tenía originalmente. La certeza cae a medida que decrece la diferencia de frecuencias de ese alelo entre las poblaciones ancestrales consideradas. Esto se compensa con la consideración de múltiples alelos y con la selección de los polimorfismos que exhiban mayor diferencia de frecuencias entre poblaciones.

Se denomina AIM (*marcador informativo de ancestría*) a ese tipo de polimorfismo, que exhibe grandes diferencias de frecuencias alélicas entre poblaciones y que, por ende, puede ser utilizado para inferir el origen genético/geográfico de un individuo [1]. Como mencionamos, los alelos con un $F_{ST} = 1$ entre poblaciones proveen la máxima información posible, pero son la excepción. Usualmente, un $F_{ST} > 0{,}5$ es considerado suficiente para que un marcador sirva como AIM.

Por otro lado, un "panel" es un conjunto de marcadores que se genotipan simultáneamente y que fueron seleccionados con un fin determinado, como caracterizar la estructura de una población o estimar el riesgo de tener una enfermedad de causa genética. Si bien en principio cualquier tipo de marcador puede servir, los más utilizados desde el surgimiento de la tecnología de *arrays* son los SNPs, dada su abundancia y su distribución a lo largo del genoma [19]. A partir de los alelos presentes en los *loci* definidos como AIMs, es posible estimar las proporciones de ancestría media en el genoma o por cromosoma.

Dado que las poblaciones nativas de América no eran genéticamente homogéneas al momento de la invasión europea, los alelos que hayan tenido alta frecuencia en una población nativa no necesariamente la tendrán en las otras. Así pues, para detectar ancestría americana en general, podrían elegirse marcadores específicos para *cada una* de las poblaciones nativas, de modo de detectar las ancestrías de diferentes poblaciones americanas. Una estrategia diferente consiste en la elección de marcadores de igual frecuencia en todas las poblaciones nativas y que tengan una frecuencia muy diferente en otros continentes. Esta estrategia es la adoptada por Galanter *et al.* [1] para el diseño del panel de AIMs que analizamos en este trabajo.

La estimación de la ancestría de un individuo tiene diversas aplicaciones. Además del interés personal que puede tener una persona por conocer el origen genético y posiblemente geográfico de sus ancestros, las proporciones de ancestría pueden ser tomadas como determinantes genéticos asociados a características fisiológicas y riesgo de enfermedades. Estimar la ancestría sirve también para controlar si una población está subdividia en grupos con algún grado de diferenciación genética –dato útil para evitar falsas asociaciones en estudios de asociación genética. Otras aplicaciones incluyen el uso forense [35] y la investigación histórica y antropológica, ya que los datos genéticos echan luz sobre la historia migratoria de nuestra especie [17].

1.8. Análisis realizados en este trabajo

1.8.1. Análisis de componentes principales

Alexander, Novembre y Lange [36] distinguen entre programas que estiman la ancestría *global* de un individuo y programas que estiman la ancestría *local* por cada región de sus cromosomas. En el "paradigma de ancestría local", se busca asignar a cada segmento cromosómico un origen poblacional, mientras que en el "paradigma de ancestría global", los programas buscan estimar la proporción promedio con que cada población de origen contribuyó al genoma.

Al estimar la ancestría global de un individuo, hay dos enfoques posibles: la estimación basada en un modelo y la estimación algorítmica [36]. Dentro los efoques de estimación algorítmica se incluye el análisis de componentes principales (PCA), utilizado en este trabajo. El PCA es un método general para obtener estadísticas resumidas en pocas dimensiones a partir de datos multidimensionales, minimizando la pérdida de información [37]. En el caso de la genética de poblaciones, la información genotípica de distintos marcadores conforma la multiplicidad de dimensiones que se busca resumir.

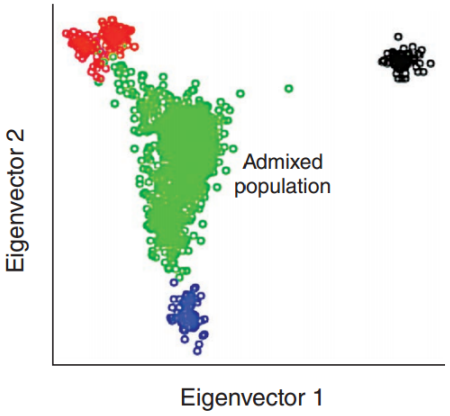


Figura 1.7: PCA esquemático donde los individuos de la población verde son mestizos con mezcla de las poblaciones roja, negra y azul. Se llama *Eigenvectors* a los componentes principales. Tomado de [17].

El PCA ubica a los individuos a lo largo de ejes de variación, lo que lo hace flexible para describir poblaciones con variación continua [33]. En el espacio de los primeros componentes, los individuos más similares genéticamente tienden a agruparse y los individuos mestizos se distribuyen entre las poblaciones de las que provienen según su grado de ancestría en cada una (véase la **Figura 1.7**). Las similitudes y diferencias genéticas son inferidas a partir de las distancias euclidianas a lo largo de esos ejes.

Con todo, debe abordarse esta técnica con prudencia. Novembre y Stephens [37] mostraron que el PCA produce ciertos patrones distintivos al analizar poblaciones en un modelo de aislamiento por distancia, aun en ausencia de movimientos migratorios. Novembre relata la queja de un colega: "interpretar los primeros componentes principales de un dataset complejo se parece a leer la borra del café" [17]. Si no se acompaña el análisis de los componentes principales y sus *clusters* con conocimientos complementarios de otras disciplinas —e.g. historia, arqueología, lingüística— se corre el riesgo de intentar *adivinar* la historia demográfica en la distribución de las muestras a lo largo de los ejes. En el caso de los latinoamericanos, está ampliamente documentado que la población del continente es el resultado del mestizaje entre tres poblaciones previamente aisladas, de modo que el PCA probará ser una herramienta útil para dilucidar las composiciones ancestrales europea, africana y americana, pero más allá de este punto hemos de proceder con cuidado.

1.8.2. Estimación de ancestrías con el programa *Admixture*

Un método de estimación global de ancestría fue implementado por el programa *admixture*, que calcula coeficientes o proporciones de ancestría de cada individuo como parámetros de un modelo estadístico [36]. El modelo se construye postulando K "poblaciones ancestrales" o *clusters*, cada una de las cuales será caracterizada por un conjunto de frecuencias alélicas para los *loci* presentes en los genotipos utilizados. El método estima, para cada

individuo, qué proporción de su genotipo proviene de cada población ancestral, mientras que en simultáneo calibra las frecuencias alélicas que definen a esas poblaciones.

Es importante distinguir el *input* de este método de su *output*. El número de poblaciones o *clusters* es determinado de antemano al correr el programa —el valor K— y se utiliza como *input*, junto a los genotipos observados. Como *output*, se obtienen las proporciones de ancestría de cada genotipo. Véase como ejeplo la **Figura 1.8**, donde los investigadores analizaron muestras de todos los continentes con un valor de $K = 7$.

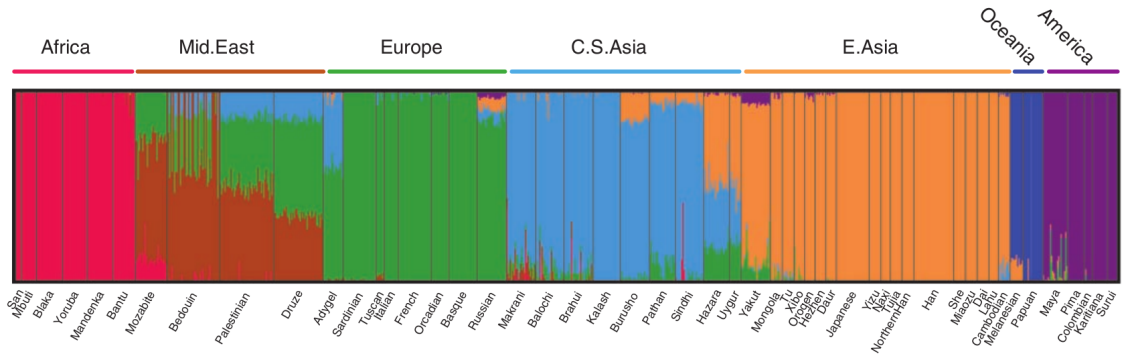


Figura 1.8: Ejemplo de corrida de *frappe* (programa similar a *admixture*) con $K = 7$. Los investigadores utilizaron muestras de todos los continentes. Cada línea vertical representa un individuo y los diferentes colores, diferentes ancestrías. Tomado de [11]

Cabe aclarar que los *clusters* no necesariamente corresponden con poblaciones ancestrales. Es trabajo del investigador elegir un valor K adecuado para correr el programa. Como sugieren los desarrolladores de *admixture*, la elección debe estar guiada por el conocimiento de la historia de una población. En este trabajo, nos basamos en el conocimiento previo sobre el origen de los latinoamericanos modernos y priorizamos el análisis con $K = 3$, salvo en los datasets que incluyen poblaciones adicionales más allá de América, Europa y África.

Un segundo modo de decidir un valor de K viene dado por *admixture* mismo, que incluye el error de validación cruzada para cada K elegido. Este valor se obtiene "ocultando" deliberadamente un subconjunto de los genotipos, prediciendo sus valores con el modelo y luego calculando el error de esa predicción [36]. Con esto se pueden comparar corridas del programa con diferentes valores de K y elegir la que arroje el menor error de validación cruzada.

Precauciones al interpretar de los *clusters* de *admixture*

Teniendo en cuenta cómo funciona *admixture*, no puede decirse que el programa descubra o permita inferir que los latinoamericanos tienen cierto número de poblaciones ancestrales. Al contrario, ese dato es una hipótesis inicial de la que el programa depende. El mismo algoritmo puede revelar subestructura poblacional indefinidamente, corriendo para el valor de K que se le provea, con indiferencia respecto de la historia demográfica real de las muestras. En cierto sentido, se puede afirmar que el *clustering* con cualquier valor de K podría, en principio, tener una interpretación razonable asociada a los niveles continental, subcontinental, regional, etc.

Es por esto que, en consonancia con lo antes argumentado acerca del concepto de "raza", no debe introducirse la idea de "pureza ancestral" en un análisis como el de *admixture*. Los porcentajes reportados dependen, por un lado, del valor de K elegido por el investigador, y por el otro, de qué poblaciones actuales se utilicen como *proxies* de las poblaciones ancestrales.

1.9. SNP *arrays* y panel de AIMs de Galanter *et al.* (2012)

Durante la última década, los costos de genotipado se redujeron en varios órdenes de magnitud [38]. Parte de esta reducción de costos se debe al desarrollo de la tecnología de SNP *arrays*, basada en *microarrays* de oligonucleótidos, cuyos formatos actuales pueden analizar simultáneamente una o varias muestras de ADN en busca de las variantes de más de 1 millón de SNPs, con precisión y reproducibilidad mayores al 99 % [39].

Los *arrays* (también llamados *chips*) consisten en una placa de vidrio de alrededor de $1,5cm^2$, dividida en una "grilla" donde cada sección (del orden algunos micrómetros de lado, según el modelo) lleva unidas millones de copias de la hebra complementaria al alelo de un SNP particular. Estas sondas se unen por complementariedad al ADN de una muestra en caso de que esa variante esté presente.

Una de las compañías dedicadas al diseño y comercialización de *arrays* de SNPs es Affymetrix, cuyos paneles de la serie *Axiom World Arrays* fueron optimizados para genotipar diferentes poblaciones del mundo. Dentro de esta serie, el diseño del *array* LAT 1, realizado en 2011, tomó en cuenta las ancestrías principales que componen a los latinoamericanos [40]. El número final de SNPs incluidos en el *array* LAT 1 es de 817.810.

Al año siguiente, Galanter *et al.* [1] publicaron el diseño de un panel de 445 AIMs, optimizado para la caracterización de poblaciones latinoamericanas. La selección de AIMs estuvo orientada a la detección de los tres componentes ancestrales predominantes en nuestro continente –americano nativo, europeo y africano– y se buscó que el panel fuera portable entre distintas poblaciones latinoamericanas, es decir, que sirviera para su utilización en distintos países del continente. Como el diseño del *array* LAT 1 es previo a esta publicación, no todos los AIMs de Galanter *et al.* (2012) fueron incluidos.

1.10. Objetivos del trabajo

Se prevé la utilización de LAT 1 en el proyecto del Consorcio y Biobanco de Población Argentina: PoblAr [41]. Por este motivo, un primer objetivo de este trabajo consiste en determinar cuáles de los AIMs elegidos por Galanter *et al.* (2012) están incluidos en el *array*, y establecer si es posible una estimación de ancestría de igual precisión sin los marcadores que falten.

Para examinar esto, definimos el panel reducido `GAL_Affy` como la intersección entre los SNPs presentes en el *array* LAT 1 de Affymetrix y los AIMs del panel de Galanter *et al.* (2012) (`GAL_Completo` de aquí en adelante). El objetivo del trabajo consiste, precisamente, en comprobar si el panel reducido `GAL_Affy` permite una estimación de ancestrías igual de precisa que la que genera el panel `GAL_Completo`.

Un segundo objetivo del trabajo, relacionado al punto anterior, es evaluar la posibilidad de realizar un diseño del *array* de menor cantidad de marcadores y, por ende, de menor costo, que sirva para estimar la ancestría en muestras de la población argentina y que no pierda sensibilidad o precisión. Con este fin, diseñamos una serie de subpaneles reducidos basados en `GAL_Affy` y evaluamos la estimación de ancestría que permiten, en comparación con el panel completo.

Finalmente, como análisis complementario, evaluamos cuántos SNPs elegidos al azar en el *array* LAT 1 son necesarios para lograr una determinación de ancestrías similar a la que posibilitan los AIMs optimizados del panel `GAL_Completo`. Se busca con esto justificar la elección de AIMs y demostrar su poder de estimación de ancestría en contraste con marcadores aleatorios del genoma. A la vez, se evalúa si hay alguna ventaja en un diseño de marcadores al azar.

En todos los casos, la comparación se realiza de dos modos. En primer lugar, analizamos si la distribución de los individuos en un análisis de componentes principales es similar

al utilizar los marcadores de distintos paneles. En segundo lugar, examinamos si las proporciones de ancestría arrojadas por el programa *admixture* son similares para el mismo conjunto de muestras, basándose en los genotipos de los diferentes paneles.

2 Materiales y Métodos

2.1. *Software* utilizado

Utilizamos el programa *plink* [42] para la extracción de los genotipos definidos por cada panel. La conversión a diferentes formatos fue también realizada con *plink* así como la generación de los paneles de SNPs aleatorios, con la ayuda de scripts en Bash y Python [43].

La mayor parte de la manipulación de los datos de genotipos y de información sobre la población de origen de cada muestra fue realizada en Python, con la librería *pandas* [44]. Para el análisis de componentes principales utilizamos la *suite* de programas EIGENSOFT [45, 46] y para la generación de gráficos las librerías *matplotlib* [47] y *seaborn* [48]. La mayor parte del código fue ejecutado en el ambiente interactivo provisto por *Jupyter* [49].

Para la determinación de proporciones de ancestría a partir de los genotipos generados utilizamos el programa *admixture* [36].

2.2. Descarga de paneles y genotipos

Descargamos la información de los 817.170 SNPs del Axiom World Arrays modelo LAT 1 (versión 35) de Affymetrix del sitio web de la compañía [50]. La lista de los 445 AIMs del panel definido por Galanter *et al.* [1] se encuentra publicada en el sitio web de PLoS Genetics.

Los genotipos utilizados corresponden al *release* 20130502 del Proyecto 1000 Genomas fase 3 y fueron descargados desde los servidores ftp del proyecto (*build* GRCh37.p13 del genoma de referencia). Para extraer los SNPs definidos por cada panel a partir de los archivos de variantes por cromomosoma (`.vcf`), utilizamos *plink* con el siguiente loop en Bash:

```
for chromosome in {1..22}; do

    # Extracts the panel SNPs list to a new file, per chromosome
    plink --vcf ALL.chr${chromosome}.phase3.20130502.genotypes.vcf.gz
          --extract GAL_Completo.snps
          --make-bed --out GAL_Completo_chr_${chromosome}

done

# Merge the chromosome files in a single .bed
ls *.bed | sed "s/.bed//" > files_to_merge
plink --merge-list files_to_merge --make-bed --out GAL_Completo
```

La extracción del panel `GAL_Affy` fue realizada también con *plink*. De los 445 SNPs definidos en el panel original, dejamos afuera tres SNPs que no son bialélicos para facilitar el análisis: rs12065716, rs2510719, rs2242865.

2.3. Creación de paneles control con marcadores aleatorios

Tras la extracción de los genotipos definidos por `GAL_Completo` y `GAL_Affy`, creamos tres paneles adicionales de marcadores elegidos al azar a partir de los 817 mil disponibles en el

array LAT 1. El primero, denominado `CPx1` (por *control panel* x 1), tiene aproximadamente la misma cantidad de SNPs en cada cromosoma que la encontrada en `GAL_Completo`. Los paneles `CPx10` y `CPx100`, por otro lado, tienen una cantidad de SNPs diez y cien veces mayor, respectivamente. El detalle se muestra en la **Sección 3.1.1**, **Tabla 3.1**.

La generación de los paneles aleatorios tuvo por únicas condiciones que los SNPs estuvieran disponibles en el array LAT 1 y que no estuvieran ligados. Por cada cromosoma, decidimos tomar el número de AIMs definidos en `GAL_Completo` multiplicado por el factor utilizado en cada panel aleatorio (1, 10 ó 100). Definida así la cantidad de SNPs por cromosoma, buscamos maximizar la distancia entre ellos y finalmente eliminamos un SNP de cada par que estuviera relacionado con un $r^2 > 0{,}4$ en ventanas de 100 kbp. Los principales comandos utilizados para esto fueron:

```
for n in 1 10 100; do
    for chromosome in {1..22}; do
        # Extract the control panel SNPs list to a new file, per chromosome
        plink --vcf ALL.chr${chromosome}.phase3.20130502.genotypes.vcf.gz
              --extract CPx${n}.${chromosome}.snps
              --make-bed --out CPx${n}.${chromosome}

        # Prune the panel SNPs to remove markers in linkage disequilibrium
        plink --bfile CPx${n}.${chromosome}
              --indep-pairwise 100 25 0.4

        plink --bfile CPx${n}.${chromosome}
              --extract CPx${n}.${chromosome}.prune.in
              --make-bed --out CPx${n}.${chromosome}.pruned
    done

    # Merge the chromosome files in a single .bed
    ls CPx${n}.*.pruned.bed | sed "s/.bed//" > files_to_merge
    plink --merge-list files_to_merge --make-bed --out CPx${n}
done
```

2.4. Creación de subpaneles de `GAL_Affy`

Para analizar cuál es el mínimo de AIMs con el que se mantiene el *clustering* de las muestras por población, creamos una serie de subpaneles de `GAL_Affy` de tamaños decrecientes entre 175 y 5 AIMs, a intervalos de 5 marcadores. Establecimos como requisito para los subpaneles que el LSBL acumulado para las tres ancestrías tuviera un valor similar. El LSBL es el *locus-specific branch length*, una medida de cuánto permite un marcador diferenciar entre poblaciones; el LSBL acumulado de un panel es simplemente la suma del LSBL todos sus marcadores, que en nuestro caso se basa en el F_{ST}. Este valor fue tomado del material suplementario de Galanter *et al.* [1].

El proceso de creación de subpaneles es parecido al detallado por Galanter *et al.* para la selección de sus AIMs [1]. Comenzamos con la elección del AIM de mayor LSBL de cada ancestría. Luego añadimos un nuevo AIM (el de mayor LSBL disponible) para la ancestría que tenga una menor suma total de LSBL en el subpanel parcial. Repetimos este paso, agregando AIMs de distintas ancestrías de manera alternada, hasta llegar al número total de AIMs deseado. El resultado es un subpanel donde el LSBL acumulado para cada ancestría (americana, europea y africana) tiene un valor similar.

2.5. Creación de los datasets

Los datos del Proyecto 1000 Genomas incluyen 2.504 muestras de 26 poblaciones. En este trabajo utilizamos las cuatro poblaciones latinas disponibles, dos poblaciones europeas, dos poblaciones africanas y tres poblaciones asiáticas.

Los esclavos forzados a migrar a América provenían mayoritariamente de las costas del oeste de África [24], razón por la cual incluimos en el análisis a los yorubas de Ibadán, Nigeria (YRI). No obstante, incluimos también a una población del este africano, los luhyas de Webuye, Kenya (LWK), dado que Galanter *et al.* (2012) utilizaron muestras de luhyas para evaluar la calidad de su panel. La razón es que los luhyas, en tanto hablantes de lenguas bantúes, están relacionados genéticamente a otros grupos que hablan lenguas de la misma familia en el oeste.

Con respecto a las muestras europeas, la elección de las muestras ibéricas de España (`IBS`) se explican porque esa región constituyó el principal origen de la inmigración hacia América durante la invasión. Por otro lado, la elección de las muestras británicas de Inglaterra y Escocia (`GBR`) tuvo como objetivo contrastar las muestras del sur de Europa con muestras del norte del mismo continente.

Para realizar análisis adicionales, sumamos además poblaciones del este y del sur de Asia. El detalle de todas las poblaciones, el número de genotipos y el código con el que se nombran sus muestras se reúnen en la **Tabla 2.1** y se puede observar en la **Figura 2.1**.

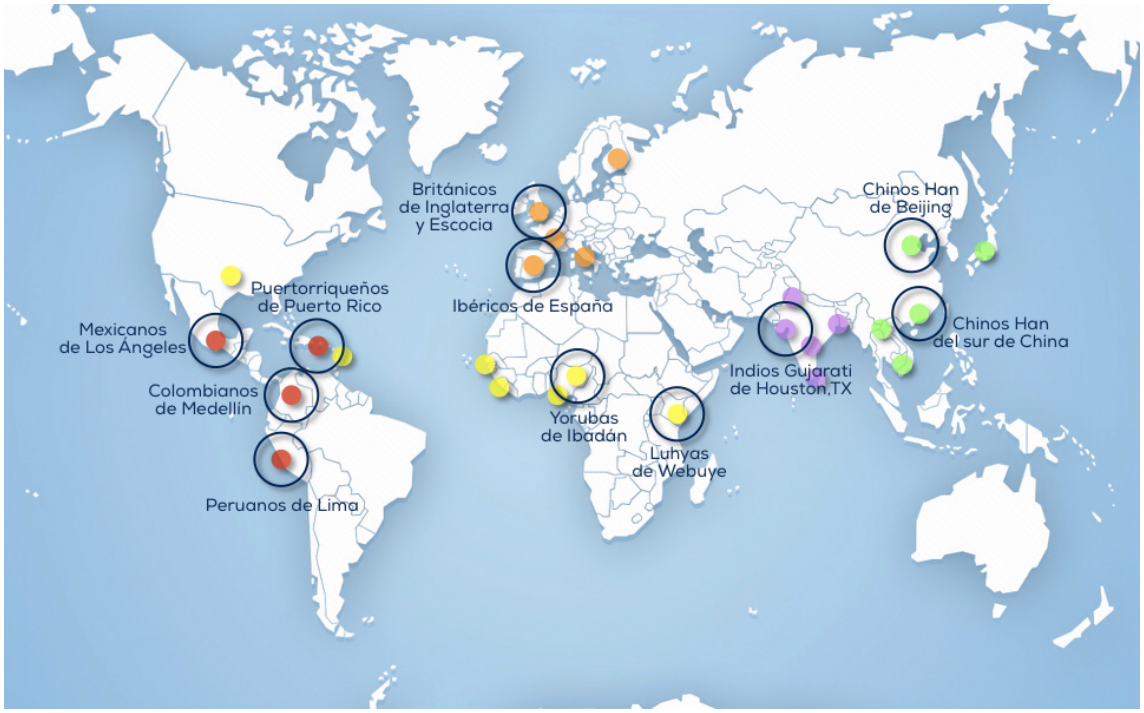


Figura 2.1: Se detallan en el mapa todas las poblaciones que forman parte del Proyecto 1000 Genomas, con un círculo alrededor de las utilizadas en este trabajo.

Con estas poblaciones disponibles definimos cinco datasets. En cada dataset unimos los genotipos de las poblaciones latinas con distintas combinaciones de poblaciones de los otros continentes; el detalle se da en la **Tabla 2.2**. Si bien los análisis fueron realizados sobre todos los datasets, el de mayor relevancia para nuestros objetivos es el dataset que denominamos `LEA`, compuesto por latinoamericanos, europeos y africanos.

Tabla 2.1: Poblaciones de 1000 Genomas utilizadas en este trabajo.

Región	Población	Descripción	Muestras
AFR	LWK	Luhyas de Webuye, Kenya	99
	YRI	Yorubas de Ibadán, Nigeria	108
AMR	CLM	Colombianos de Medellín, Colombia	94
	MXL	Mexicanos de Los Angeles, EEUU	64
	PEL	Peruanos de Lima, Perú	85
	PUR	Puertorriqueños de Puerto Rico	104
EAS	CHB	Chinos Han de Bejing, China	103
	CHS	Chinos Han del sur de China	105
EUR	GBR	Británicos de Inglaterra y Escocia	91
	IBS	Ibéricos de España	107
SAS	GIH	Indios Gujarati de Houston, Texas	103

Tabla 2.2: Datasets definidos para los análisis.

Código	Poblaciones	Muestras
L	Latinos: PUR, CLM, MXL, PEL	347
LE	Latinos, Europeos: PUR, CLM, MXL, PEL, GBR, IBS	545
LEA	Latinos, Europeos, Africanos: PUR, CLM, MXL, PEL, GBR, IBS, LWK, YRI	752
LEAC	Latinos, Europeos, Africanos, Chinos: PUR, CLM, MXL, PEL, GBR, IBS, LWK, YRI, CHS, CHB	960
LEACI	Latinos, Europeos, Africanos, Chinos, Indios: PUR, CLM, MXL, PEL, GBR, IBS, LWK, YRI, CHS, CHB, GIH	1063

2.6. Análisis de componentes principales

El análisis de componentes principales fue realizado con *smartpca*, parte de la *suite* de programas de EIGENSOFT [45, 46] y con ayuda de *scripts* en Python para automatizar el proceso. Para cada dataset se realizaron cinco PCAs: dos de ellos con los genotipos de los paneles `GAL_Completo` y `GAL_Affy` y los tres restantes con los genotipos de los paneles aleatorios, `CPx1`, `CPx10` y `CPx100`. Además, realizamos PCAs utilizando el dataset `LEA` para evaluar todos los subpaneles de `GAL_Affy` generados.

Los datos de entrada del PCA fueron las matrices de genotipos en formato de "dosis alélica", donde 0 equivale al homocigoto del alelo de referencia, 1 al heterocigoto y 2 al homocigoto del alelo alternativo —considerando siempre marcadores bialélicos (véase como ejemplo de este tipo de matriz la **Tabla 2.3**). Esta manera de expresar los genotipos permite obtener una matriz numérica con la cual realizar un PCA.

Tabla 2.3: Matriz de genotipos con muestras por fila y SNPs por columna.

	rs6685064	rs12085319	rs2745285	rs4920310	rs6684063	...
HG01112	0	1	1	0	0	...
HG01113	0	1	1	1	1	...
HG01119	0	1	1	1	1	...
HG01121	1	0	2	0	1	...
HG01122	2	1	1	0	0	...
...						

2.7. Estimación de la proporción de ancestrías con *admixture*

Utilizamos el programa *admixture* 1.3.0 [36] para estimar la proporción de ancestrías en cada muestra. Generamos archivos `.bed` con *plink* para cada combinación de dataset y panel (5 datasets por 5 paneles) y en cada caso realizamos el análisis con valores de K entre 2 y 9. El formato de entrada para *plink*, el archivo `.bed`, contiene la información de los genotipos por muestra y marcador tal como se observa en la **Tabla 2.3**.

Una versión simplificada del script utilizado para el cálculo de ancestrías se muestra a continuación:

```
for panel in GAL_Completo GAL_Affy CPx1 CPx10 CPx100; do
    for dataset in L LE LEA LEAC LEACI; do

        # For each panel / dataset combination, make a new bedfile
        plink --bfile ${panel}
              --keep-fam ${dataset}.samples
              --make-bed --out ${panel}.${dataset}

        # Use the new bedfile to estimate ancestries with Ks 2 through 9
        for K in {2..9}; do
            admixture --cv ${panel}.${dataset}.bed ${K}
        done

    done
done
```

Reunimos luego los valores de error de validación cruzada en un único archivo, con el siguiente comando:

```
rubycode='puts $_.match("(\\d).: (.*)").to_a[1..-1].join(",")'
grep CV */*.log | ruby -lane $rubycode > CV_error_summary
```

Seguimos el mismo procedimiento para correr *admixture* con los subpaneles de AIMs, esta vez limitándonos al dataset de poblaciones `LEA`.

3 Resultados

3.1. Comparación entre los paneles

3.1.1. Cantidad de SNPs y LSBL acumulado

De los 445 AIMs definidos en `GAL_Completo`, 179 se encuentran en el array LAT 1 de Affymetrix y conforman el panel que definimos como `GAL_Affy`. Por otro lado, definimos subpaneles de `GAL_Affy` con tamaños decrecientes hasta llegar a 15 AIMs, manteniendo en cada reducción los marcadores más informativos de ancestría. Finalmente, los paneles `CPx1`, `CPx100` y `CPx100` corresponden a una selección al azar de 438, 4,4 mil y 43,1 mil SNPs que forman parte de LAT 1. El objetivo de estos paneles fue contrastar el desempeño de los paneles `GAL` con un panel aleatorio de igual tamaño y con otros dos 10 y 100 veces más grandes. El detalle se da en la **Tabla 3.1** y la relación entre paneles se representa en la **Figura 3.1**.

Tabla 3.1: Cantidad de SNPs por panel, sin incluir los subpaneles.

Panel	SNPs	Fracción de `GAL_Completo`
`GAL_Completo`	445	1,00
`GAL_Affy`	179	0,40
`CPx1`	438	0,98
`CPx10`	4.424	9,94
`CPx100`	43.144	96,95

Cada AIM del panel `GAL_Completo` fue seleccionado por Galanter *et al.* para diferenciar a una población ancestral en particular, con el criterio de que su frecuencia alélica se distinga de la frecuencia en las otras poblaciones. Por ejemplo, en la posición 1.201.155 del cromosoma 1 (rs6685064), la presencia del alelo "C" permite inferir ancestría europea con cierta probabilidad, dado que la frecuencia de esa variante en europeos contemporáneos

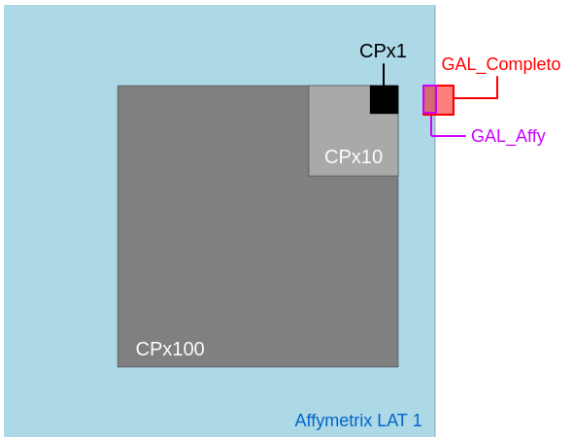


Figura 3.1: Diagrama de Venn simplificado de los paneles.

es de 0,93, mientras que sólo llega a 0,24 en nativos americanos y 0,42 en africanos. Como marcador complementario, en rs7598069 el alelo alternativo "A" aporta evidencia de ancestría americana, dado que el alelo de referencia sólo se encuentra en una frecuencia de 0,011 en americanos nativos, mientras que tiene frecuencias altas en Europa y África (0,880 y 0,920). Véase el ejemplo en la **Tabla 3.2**.

Tabla 3.2: AIMs que permiten inferir ancestría europea y americana. La presencia del alelo subrayado permitiría inferir la ancestría mencionada en la última columna, gracias a que su frecuencia varía entre continentes.

	Crom.	Posición	A1	A2	Frec. A1 AMR	Frec. A1 EUR	Frec. A1 AFR	Ancestría a inferir
rs6685064	1	1.201.155	C	T	0,244	0,93	0,422	EUR
rs7598069	2	98.127.823	G	A	0,011	0,61	0,971	AMR

Como puede apreciarse, la información se encuentra en la *diferencia* de frecuencias alélicas entre una población continental y las demás. Con esta idea de que cada marcador aporta información sobre una de las ancestrías en particular, podemos contar cuántos AIMs en cada panel 'pertenecen' a (i.e. permiten inferir información sobre) cada componente ancestral (**Tabla 3.3**).

Tabla 3.3: Proporción de AIMs para inferir cada componente ancestral

	`GAL_Completo`		`GAL_Affy`	
Ancestría a inferir	AIMs	Fracción del panel	AIMs	Fracción del panel
AFR	115	26 %	62	35 %
EUR	202	45 %	71	40 %
NAM	129	29 %	47	26 %

Con la reducción del panel `GAL_Completo`, observamos que `GAL_Affy` sufre un aumento de la proporción de AIMs africanos (de 26 % a 35 %), mientras que la proporción de AIMs americanos y europeos disminuye. Dado que no todos los AIMs aportan de igual manera a la distinción de ancestrías, recurrimos al LSBL (*locus specific branch length*) acumulado de los AIMs de cada grupo y comparamos ese valor entre paneles (**Tabla 3.4**). El LSBL fue calculado por Galanter *et al.* por cada AIM individual, en base al F_{ST} específico del *locus*.

Tabla 3.4: LSBL acumulado por panel y por componente ancestral

Componente	`GAL_Completo` LSBL acumulado	`GAL_Affy` LSBL acumulado
AFR	73.0	39.6
EUR	77.9	28.3
NAM	74.5	27.0

En consonancia con la mayor cantidad de AIMs africanos, también se observa que el LSBL acumulado para esa ancestría es mayor respecto de las otras ancestrías en `GAL_Affy`. Dado que por diseño se buscó en `GAL_Completo` un valor similar de LSBL acumulado para los tres componentes, la reducción implica que el panel `GAL_Affy` podría sobreestimar el componente africano. Sin embargo, veremos en los análisis posteriores que la ancestría

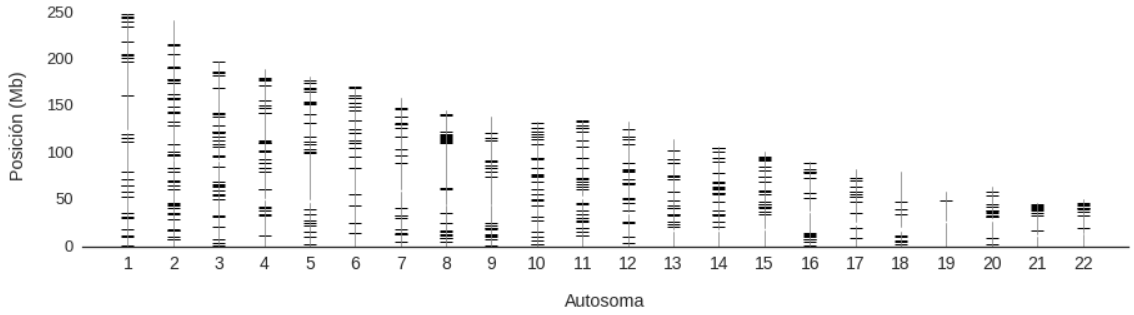
africana no parece ser sobreestimada por este panel reducido en contraste con el panel completo.

3.1.2. Distribución de los AIMs en el genoma

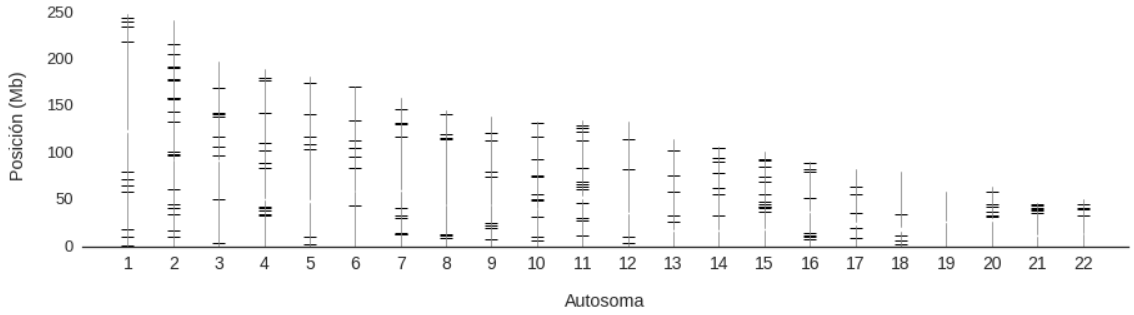
Una comparación visual rápida de la distribución de marcadores por cromosoma en cada panel nos muestra que la eliminación de 266 alelos deja en `GAL_Affy` algunas regiones cromosómicas sin genotipar, como por ejemplo una gran parte del cromosoma 1 (**Figura 3.2a** y **3.2b**). Sin embargo, esto puede decirse en gran medida también del panel original `GAL_Completo`, en particular cuando se lo compara con un panel de SNPs distribuidos regularmente en el genoma —por ejemplo, nuestro `CPx1` (**Figura 3.2c**).

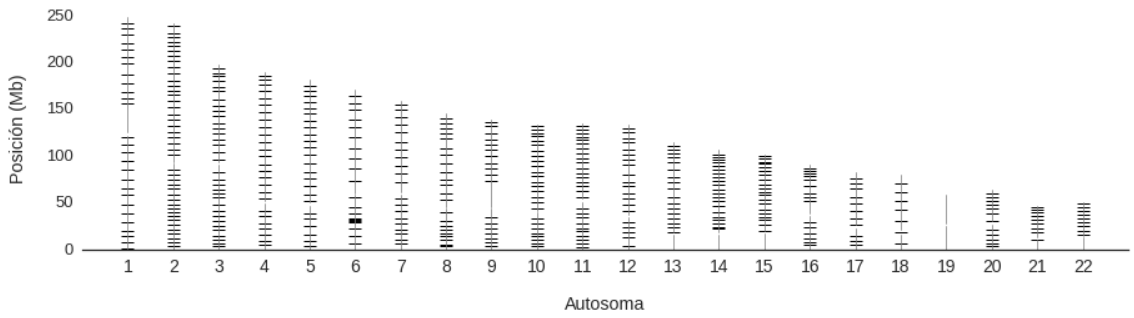
Dado que buscamos una inferencia de ancestría a nivel global —es decir, para todo el genoma y no para cada fragmento cromosómico—, la distribución uniforme y la alta densidad de los marcadores no son requisitos. Para nuestros fines, alcanza con elegir SNPs muy informativos de ancestría en regiones específicas.

(a) `GAL_Completo`. 445 AIMs



(b) `GAL_Affy`. 179 AIMs





(c) `CPx1`. 438 SNPs repartidos uniformemente.

Figura 3.2: Distribución de marcadores a lo largo del genoma, contraste entre paneles de AIMs y un panel de SNPs.

Así pues, en principio `GAL_Affy` y sus subpaneles podrían predecir ancestrías sin problemas, en tanto la distribución pareja de los marcadores no es necesaria y tampoco se cumple en el panel original `GAL_Completo`.

3.1.3. Frecuencias alélicas por continente

En línea con lo expuesto en la **Sección 3.1.1**, podemos contrastar a `GAL_Completo` y `GAL_Affy` con los paneles de control `CPx1`, `CPx10` y `CPx100` por la informatividad de sus marcadores, en función de la diferencia de frecuencias que presentan los SNPs entre las diferentes poblaciones ancestrales.

Así pues, obsérvese en la **Figura 3.3** que si un AIM dado de `GAL_Completo` tiene alta frecuencia en un continente, mostrará una baja frecuencia en los otros continentes (y viceversa). En los paneles `CP`, en cambio, los marcadores no necesariamente exhiben grandes diferencias de frecuencia en las distintas poblaciones ancestrales, de modo que el origen del alelo en un genoma contemporáneo puede ser incierto. Nótese que en el panel `CPx10` (**Figura 3.4**), a cada nivel de frecuencias en África se corresponde una nube indistinta de frecuencias en América y Europa. Peor aún, las frecuencias en los tres continentes parecen relacionadas, lo que dificulta más la tarea de inferir ancestría a partir de la presencia de un alelo determinado.

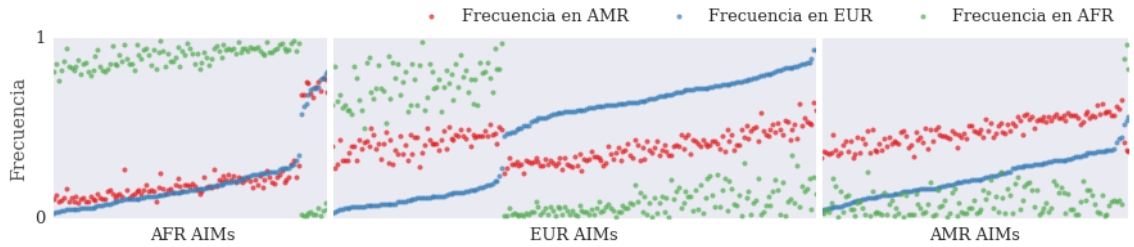


Figura 3.3: Frecuencia alélica de cada AIM en los tres continentes. Panel `GAL_Completo`.

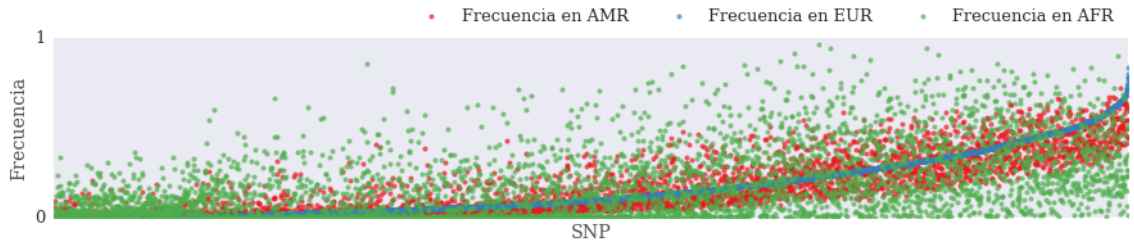


Figura 3.4: Frecuencia alélica de cada SNP en los tres continentes. Panel `CPx10`.

Lo que se aprecia en el contraste es la diferencia entre paneles *de AIMs*, por un lado, y paneles *de SNPs*, por el otro. En los primeros, las diferentes frecuencias alélicas en cada población ancestral sirven para inferir el origen probable del alelo. En los últimos, muchos alelos están presentes en frecuencias parecidas en los distintos continentes, de modo que no se puede inferir su origen al encontrarlos en un genoma.

Los paneles `CP`, sin embargo, sirven también para estimar ancestrías porque, en esa mezcla de marcadores con diferentes frecuencias alélicas ancestrales, *algunos marcadores* al menos exhibirán la diferencia amplia que caracteriza a los AIMs. Como demostración, filtramos el panel `CPx10` quedándonos únicamente con los SNPs que tienen una diferencia de frecuencias mayor a 0.30 entre África y Europa y ordenamos la serie según si la diferencia es positiva o negativa. El resultado se aprecia en la **Figura 3.5**; los SNPs elegidos podrían conformar un panel que determine las ancestrías africana y europea y sería, de hecho, un

panel de AIMs. En el primer grupo, el alelo de referencia sirve para predecir ancestría africana; en el segundo grupo, para predecir ancestría europea.

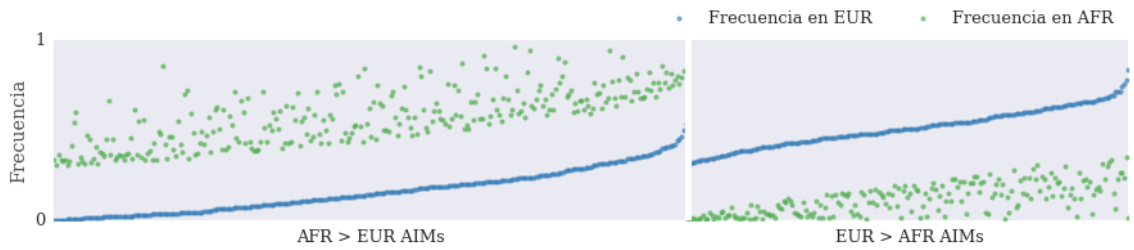


Figura 3.5: Frecuencia alélica de cada SNP en Europa y África. Panel `CPx10` filtrado: sólo se muestran los marcadores con una diferencia de frecuencias mayor a 0.3 entre esos dos continentes, separados en dos categorías según el signo de la diferencia.

Así pues, *a priori* es de esperar una buena determinación de ancestrías también con los paneles aleatorios, porque, por así decirlo, contienen un panel de AIMs en su interior. Por otro lado, como los SNPs útiles se encuentran mezclados con muchos otros que no permiten la inferencia de ancestría, es de esperar que los paneles aleatorios sólo repliquen la estimación de ancestría de los paneles de AIMs con una cantidad mucho mayor de marcadores totales.

3.2. Análisis de componentes principales

3.2.1. Paneles de AIMs y paneles control

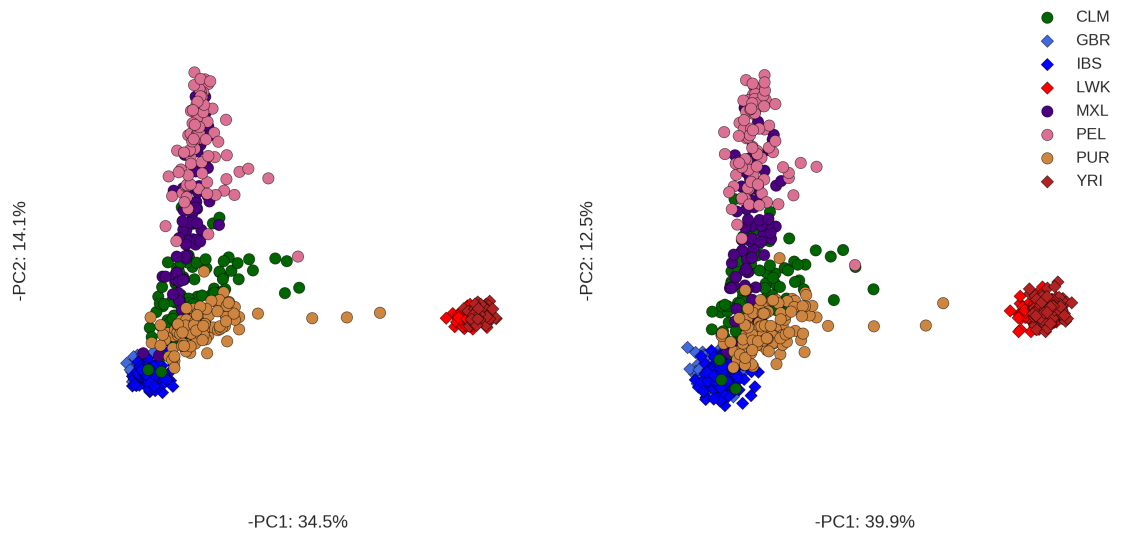


Figura 3.6: PCA usando los paneles `GAL_Completo` (izquierda) y `GAL_Affy` (derecha) con las muestras de europeos, africanos y latinoamericanos del dataset `LEA`. Se observa que la distribución en los dos primeros ejes de variación es similar: clusters europeo y africano y un gradiente de mestizaje en los latinoamericanos.

En la **Figura 3.6** graficamos la distribución de individuos de poblaciones latinoamericanas, europeas y africanas (dataset `LEA`) a lo largo de los dos primeros ejes de variación en dos análisis de componentes principales: uno basado en los marcadores de `GAL_Completo` y otro en los marcadores de `GAL_Affy`.

Comprobamos que ambos paneles generan una distribución similar de las muestras en el espacio de los dos primeros componentes principales. A lo largo del PC 1 (eje x), los individuos se reparten entre un extremo de ancestría no-africana y un extremo de ancestría africana. Este eje separa, por ende, a los individuos de poblaciones `YRI` y `LWK`, que forman un cluster africano bien definido (color rojo en la figura), de los individuos del resto del mundo, entre los cuales se distingue un cluster europeo de `GBR` e `IBS` (azul en la figura) y prácticamente todos los latinoamericanos. Se aprecia, sin embargo, que algunos individuos `PUR` y `CLM` exhiben un alto grado de ancestría africana (se acercan al cluster rojo), aunque en la mayoría de las muestras latinoamericanas el componente africano es relativamente bajo.

A lo largo del PC 2 (eje y) se observa un gradiente de ancestría europea, donde las muestras `PUR` y `CLM` exhiben un componente europeo predominante, mientras que las muestras `MXL` y `PEL` se alejan del cluster europeo. Esta distribución coincide con los estudios citados en la **Sección 1.4**: las poblaciones mexicana y peruana tienen, en promedio, un componente nativo americano mayor. Por ende, podemos sugerir que el PC 2 reparte a los individuos entre un extremo de ancestría europea y un extremo de ancestría americana.

En general, se observa que el resultado es muy similar entre ambos paneles, aunque los *clusters* están mejor definidos con los 445 marcadores de `GAL_Completo` que con el subconjunto de 179 AIMs de `GAL_Affy`.

La misma distribución se observa, a grandes rasgos, en un análisis de componentes principales realizado con los marcadores al azar del panel `CPx1` en la **Figura 3.7** (izquierda). Sin embargo, puede notarse que los *clusters* están en gran medida desagregados, lo que muestra que se pierde precisión al estimar ancestrías con marcadores al azar en lugar de AIMs. Hace falta un panel de 4.424 marcadores al azar, como el panel `CPx10` (**Figura 3.7**, derecha), para lograr una precisión similar a la que posibilitan los AIMs.

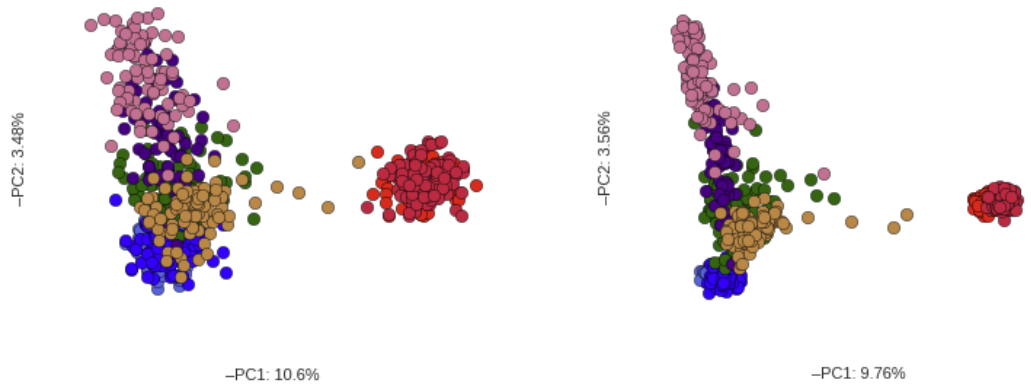


Figura 3.7: PCA con los marcadores al azar del panel `CPx1` (izquierda) y del panel `CPx10` (derecha). Puede observarse la incipiente desagregación de los *clusters* al usar el panel `CPx1`, debida a la menor precisión en la estimación de ancestrías cuando los marcadores utilizados fueron elegidos al azar. Recién con un número 10 veces mayor de SNPs al azar, como en el panel `CPx10` a la derecha, puede igualarse la precisión de los paneles de AIMs observada en la **Figura 3.6**.

3.2.2. Paneles de AIMs progresivamente reducidos

Al utilizar los subpaneles de `GAL_Affy`, en los que se restan AIMs hasta dejar sólo 5, comprobamos que remover marcadores genera *clusters* progresivamente más dispersos, lo

que equivale a reducir la precisión con la que se estima la ancestría..

Como puede observarse en la **Figura 3.8**, los *clusters* africano y europeo se hacen menos compactos a medida que se utilizan menos marcadores, a la vez que las muestras de distintas poblaciones latinoamericanas se entremezclan hasta confundirse. El umbral de desagregación aceptable dependerá de cuán precisa sea la estimación buscada, pero a grandes rasgos podemos señalar que una selección de los mejores 50 AIMs de `GAL_Affy` revelan la misma distribución general de muestras que el panel original (véase el PCA de `GAL_Affy_SubPanel_50` en la figura mencionada).

Es interesante señalar también que la precisión lograda al utilizar los 438 SNPs al azar de `CPx1` puede igualarse con tan sólo 50 AIMs seleccionados, lo que evidencia la practicidad de utilizar este tipo de marcadores a la hora de reducir el tamaño de un panel.

Podemos concluir que la tarea fundamental para la que el panel `GAL_Completo` fue diseñado (distinguir tres ancestrías en distinta proporción en latinoamericanos) es lograda también por `GAL_Affy` y por sus versiones reducidas hasta el subpanel de 50 AIMs, pero que por debajo de este número de marcadores, la distribución de las muestras se vuelve progresivamente confusa.

Limitaciones de los paneles basados en Galanter *et al.* (2012)

El diseño del panel `GAL_Completo`, cuyo fin es distinguir tres componentes ancestrales, se pone de manifiesto *como limitación* al explorar las dimensiones más altas del espacio de los componentes principales. Los paneles de AIMs evaluados permiten ordenar las muestras latinoamericanas según tres ancestrías, pero no distinguen otros tipos de diferencias genéticas, ya sea ancestrías de otras partes del mundo (como del sur o del este de Asia) o ancestrías a nivel subcontinental (como de Europa ibérica en contraste con Europa del norte).

Esta limitación es especialmente clara cuando se contrasta a el PCA realizado con los AIMs de `GAL_Completo` y un PCA realizado con los 43.144 marcadores de `CPx100`. En la **Figura 3.9** se observa que el panel `GAL_Completo`, más allá del PC 2, confunde a todas las muestras en una nube indiferenciada, mientras que en la **Figura 3.10** se aprecia que el panel `CPx100` da información nueva hasta los componentes principales 8 y 9.

Como ejemplo de la información que `GAL_Completo` no brinda, el panel `CPx100` permite distinguir como *clusters* bien definidos —y separados de las muestras latinoamericanas— a las muestras de China, por un lado, y de indios gujarati, por el otro (*clusters* turquesa y magenta). Más aun, en los PC 5 y PC 6 se diferencian como *clusters* independientes las dos poblaciones africanas del dataset: `YRI` y `LWK`. Otros componentes principales distinguen a algunos individuos como *outliers* respecto sus poblaciones, información que también se pierde con el panel `GAL_Completo` o cualquiera de sus versiones reducidas.

En resumen, los análisis de componentes principales nos muestran que el panel `GAL_Affy` estima ancestrías con una precisión equiparable a la de `GAL_Completo`, mientras que subpaneles de hasta 50 AIMs podrían ser utilizados con el mismo fin, aunque la información se vuelva un poco más confusa. Por otro lado, también observamos que los paneles de marcadores aleatorios tomados del *array* LAT 1 producen resultados similares en los primeros componentes principales, pero para ello necesitan un número de marcadores varias veces mayor. Como ventaja, los paneles aleatorios del orden de decenas de miles de SNPs (i.e. `CPx100`) se revelan como poderosas herramientas y proveen información no presente en los paneles especializados basados en el diseño de Galanter *et al.*.

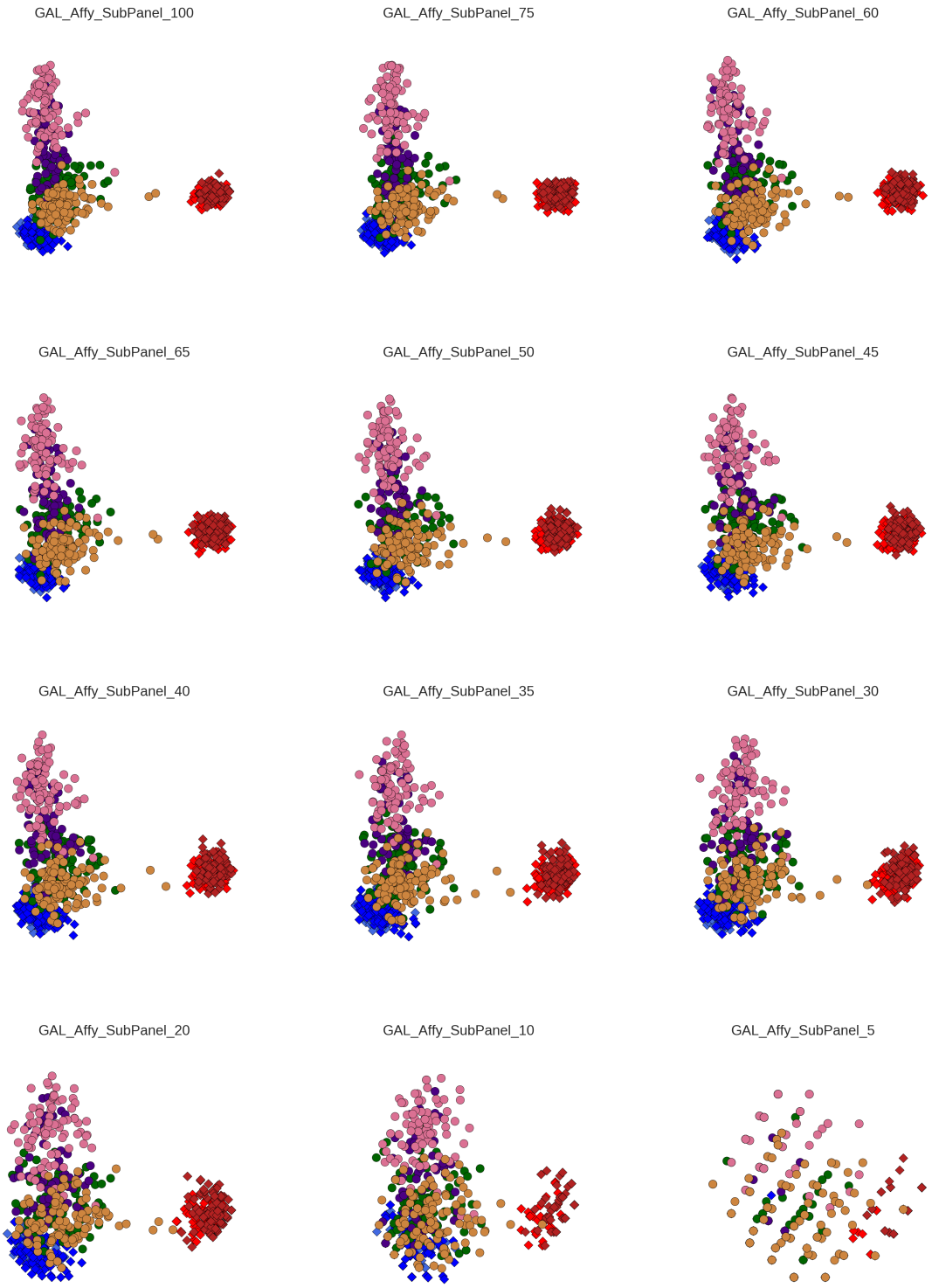


Figura 3.8: PCAs usando subpaneles de `GAL_Affy` de diferente cantidad de AIMs, con poblaciones del dataset `LEA`. Se observa la progresiva desagregación de los clusters africano (rojo) y europeo (azul), mientras que las poblaciones latinoamericanas también se dispersan y confunden a medida que se remueven marcadores.

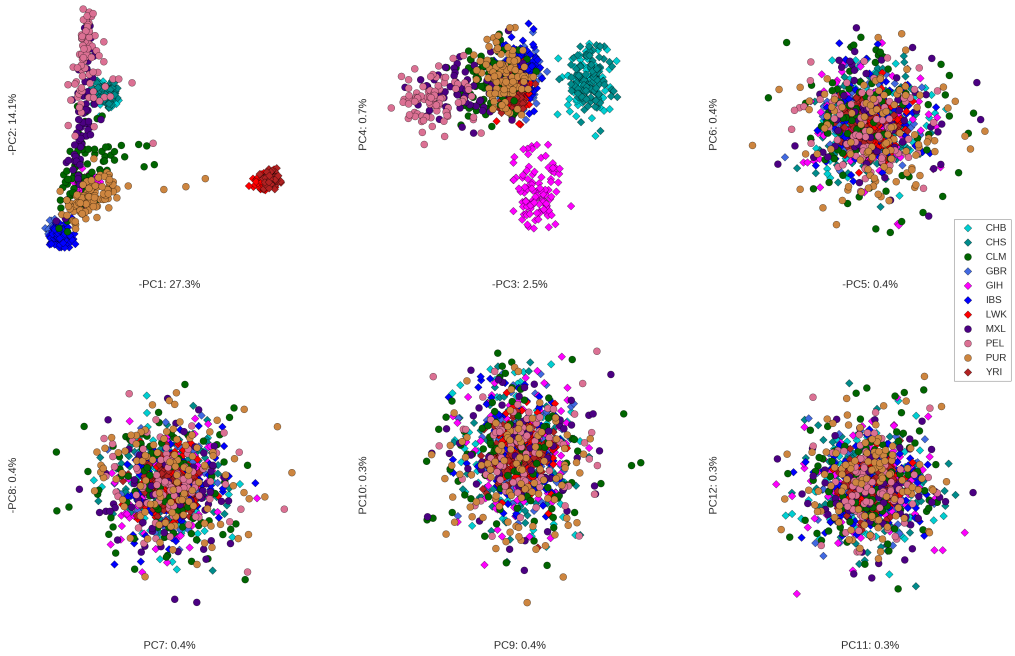


Figura 3.9: Componentes principales 1 a 12 de un PCA basado en `GAL_Completo`. El dataset `LEACI` incluye poblaciones de India y de China, que se distinguen en los PC 3 y 4. Los otros componentes principales no revelan diferencias entre las muestras.

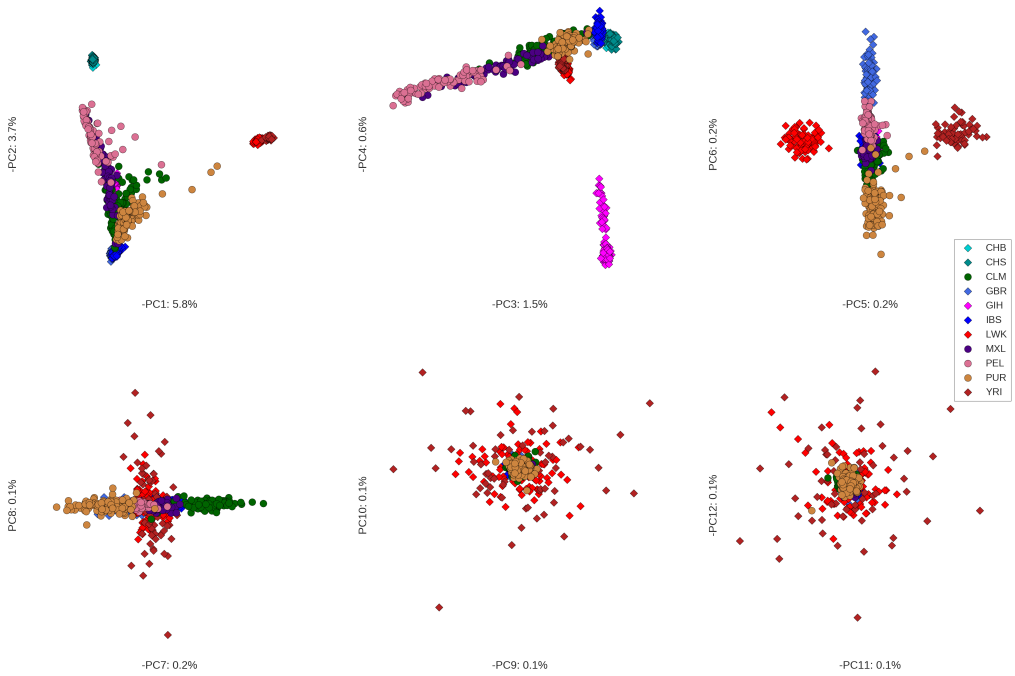


Figura 3.10: Componentes principales 1 a 12 de un PCA basado en `CPx100`. La población china se distingue como un cluster celeste en el PC 2, mientras que la población india se distingue como un cluster de color magenta en el PC 4. Adicionalmente, las dos poblaciones africanas se diferencian a lo largo del PC 5.

3.3. *Admixture*

Corrimos el programa *admixture* para valores de K entre 2 y 9, utilizando los genotipos generados con todas las combinaciones de paneles y datasets. Los valores de K que mostraron menor error de validación cruzada (*cross-validation error*) fue en la mayoría de los casos consistentes con la cantidad de poblaciones continentales utilizadas (**Figura 3.12** y **Figura 3.13**).

Por ejemplo, el dataset `LEAC` reúne poblaciones de 4 regiones —África, Europa, América, este de Asia. Consistentemente, el menor error de validación cruzada fue encontrado con un $K = 4$. Como excepción a este comportamiento, los datasets con menos de 3 poblaciones continentales (i.e. datasets `L` y `LE`) arrojaron un error menor para valores de $K = 3$. Consideramos este resultado consistente con el diseño del panel de Galanter *et al.* (2012), que busca encontrar en las muestras los tres componentes ancestrales relevantes para latinoamericanos, incluso si no se proveen poblaciones africanas o europeas modernas como *proxies* de las ancestrales.

En la **Figura 3.11** graficamos los resultados de *admixture* para el dataset `LEA` (latinoamericanos, europeos, africanos) con $K = 3$, utilizando los paneles `GAL_Completo`, `GAL_Affy` y `CPx1`. Recordemos que el *output* de *admixture* es una proporción de pertenencia de cada individuo a uno de los K *clusters*, que pueden interpretarse como poblaciones de origen. Se aprecia en la figura que la estimación de ancestrías es similar para los paneles comparados, lo que anima la idea de que `GAL_Affy` no pierde poder de estimación de ancestría frente a `GAL_Completo`. En este caso, el panel de marcadores al azar `CPx1` muestra resultados similares, aunque predice mayores proporciones de ancestría americana en europeos y africanos, lo que puede considerarse un mayor error o imprecisión.

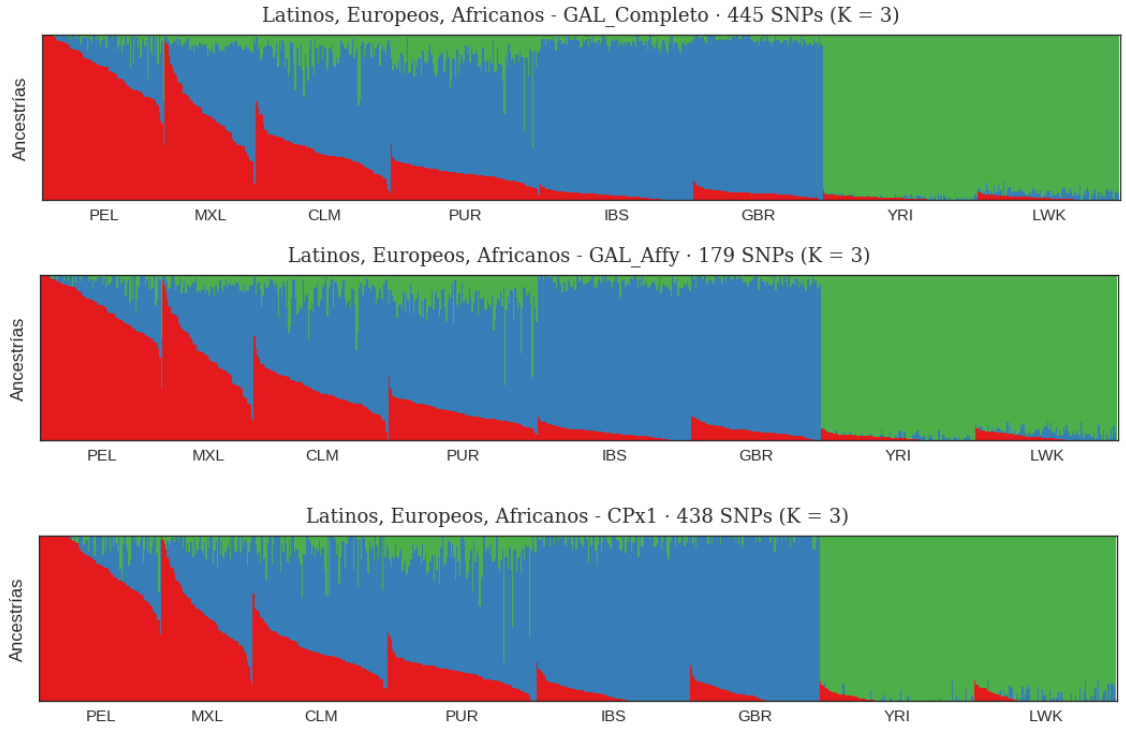


Figura 3.11: Proporciones de ancestría estimadas por *admixture* con un $K = 3$ utilizando los paneles `GAL_Affy`, `GAL_Completo` y `CPx1`. Cada línea vertical representa un individuo y los tres colores indican la proporción de pertenencia a cada uno de los *clusters* o poblaciones postuladas. Puede apreciarse que la estimación es muy similar con los distintos paneles.

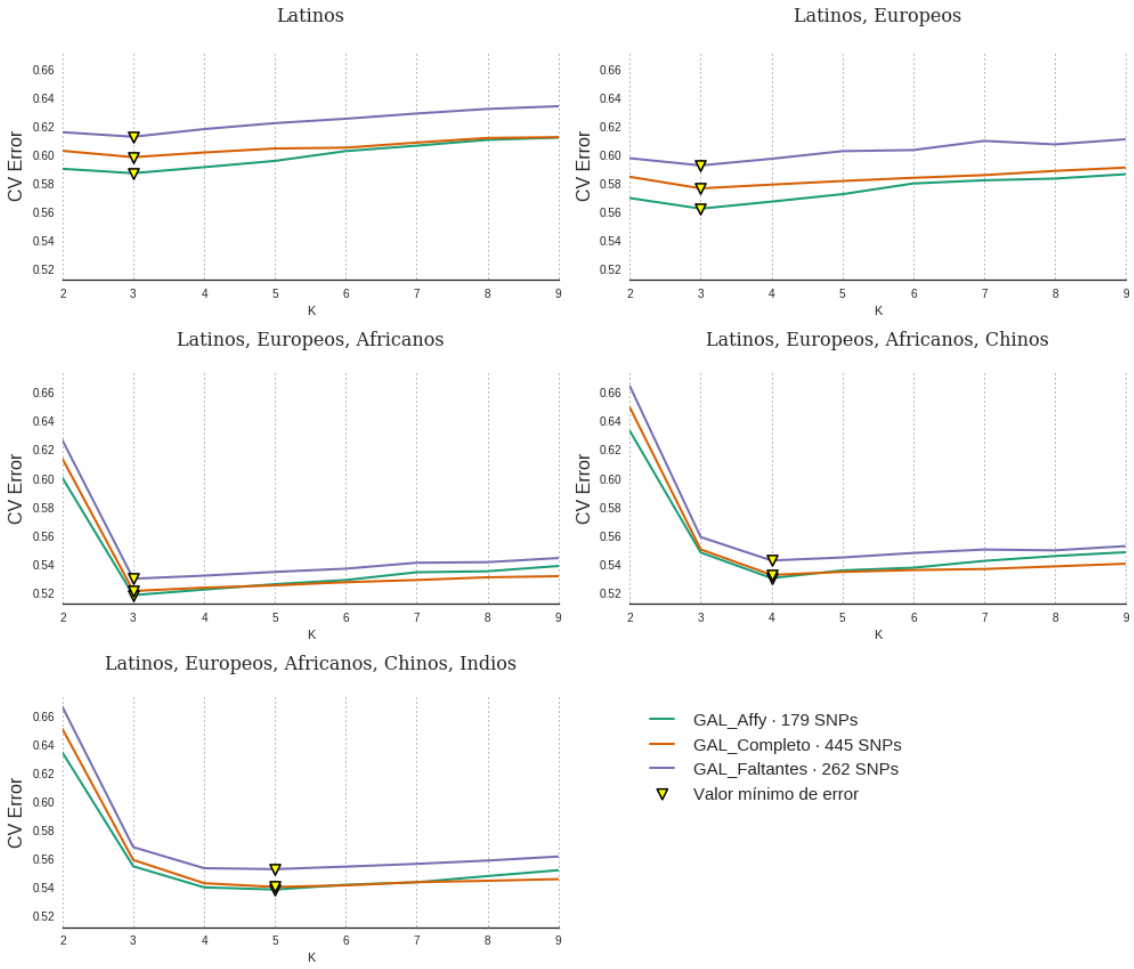


Figura 3.12: Error de validación cruzada al correr *admixture* para los paneles `GAL`, en todos los datasets y con distintos valores de K.

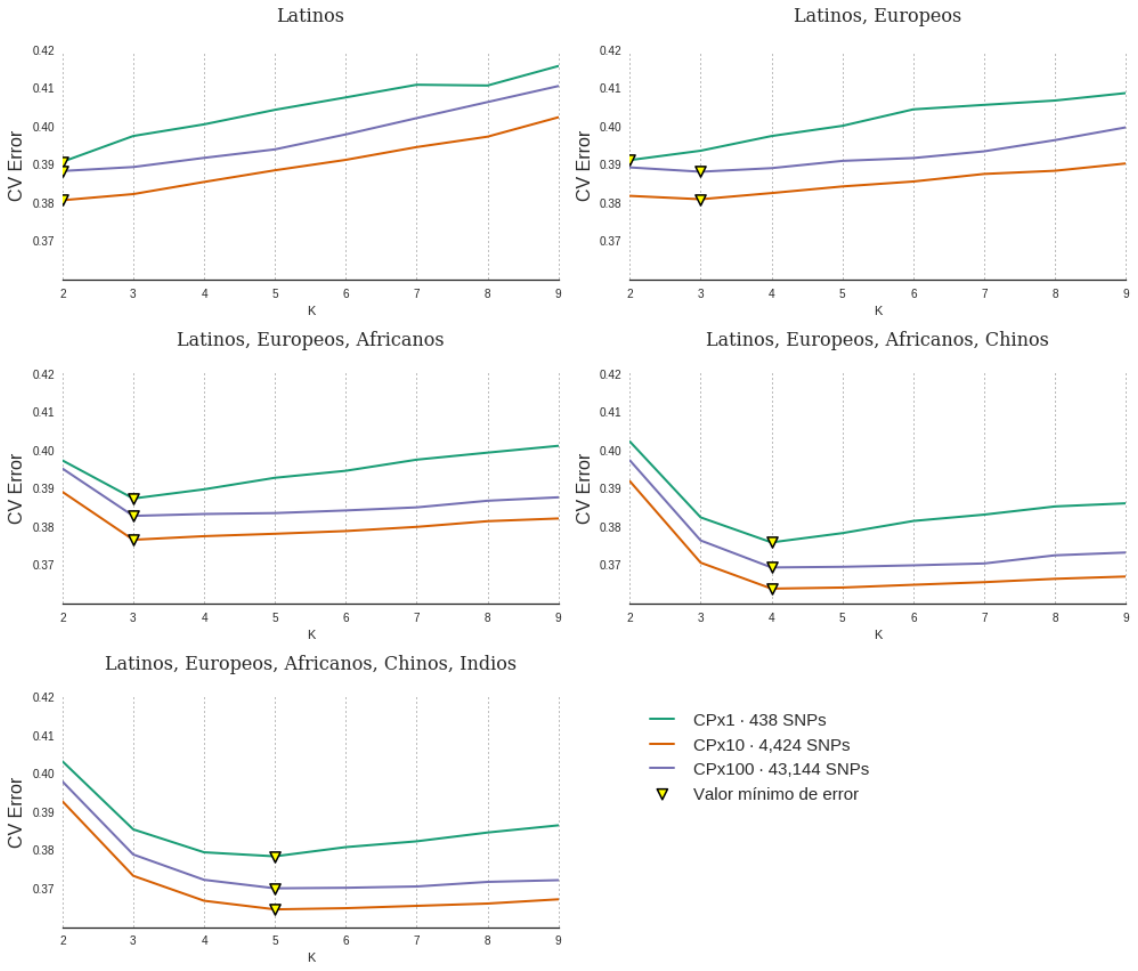


Figura 3.13: Error de validación cruzada al correr *admixture* para los paneles `CP`, en todos los datasets y con distintos valores de K.

Estos resultados replican, con otro método, la distribución entre tres ancestrías que observamos en las dos primeras dimensiones del PCA y coincide con los conocimientos previos sobre cada una de las poblaciones latinoamericanas analizadas. A diferencia del PCA, *admixture* ofrece como resultado proporciones de pertenencia de cada individuo a las poblaciones postuladas. Estos valores permiten una comparación precisa de la estimación de ancestría con los distintos paneles.

En la **Figura 3.14** se observa que al correr *admixture* con los subpaneles se sobreestima ligeramente la ancestría europea y se subestima la ancestría americana. La diferencia en media es pequeña, de 1 punto porcentual, de modo que no representa un problema para la utilización de los subpaneles. Un test t de Student con datos pareados confirma que esa diferencia es significativa. Esto implica que un individuo a quien se le adscriba 75 % de componente americano nativo con un subpanel podría tener, en verdad, 74 % o 76 %. Por otro lado, la ancestría africana estimada con los subpaneles no se diferencia en media de la ancestría africana estimada con `GAL_Completo`, salvo en los paneles más reducidos.

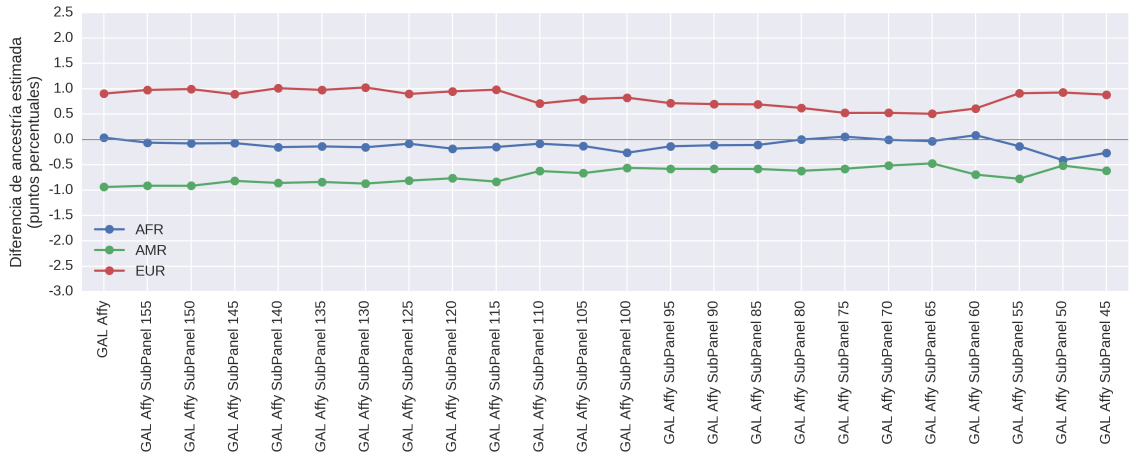


Figura 3.14: Diferencia media de las ancestrías estimadas por cada panel con respecto a la ancestría estimada por el panel base, `GAL_Completo`. Se observa que hay una consistente sobreestimación de la ancestría europea (serie EUR en rojo), relacionada a una consistente subestimación de la ancestría americana (serie AMR en verde). Sin embargo, los valores de esas diferencias con `GAL_Completo` son pequeños, de alrededor de 1 punto porcentual o menos. Véase, no obstante, la dispersión de las diferencias en la **Figura 3.15**.

La **Figura 3.15** detalla la dispersión de las diferencias de estimación de ancestría y ofrece una guía práctica de utilización de los subpaneles. Estableciendo un umbral pragmático de ± 10 puntos porcentuales de diferencia en la estimación de ancestría, observamos que en los paneles de más de 95 AIMs todas las muestras tienen diferencias menores a ese valor. Paneles de menor cantidad de AIMs disminuyen su precisión con diferencias que superan ese límite, al menos para algunas muestras. Esto permite tener una idea de qué magnitud de error de estimación puede conllevar la remoción de AIMs en los subpaneles. Con todo, el problema afecta sólo a las muestras en los extremos de la distribución, mientras que más del 50 % de las muestras se mantiene dentro de los límites definidos, lo que permite en gran medida confiar en los paneles de poca cantidad de AIMs.

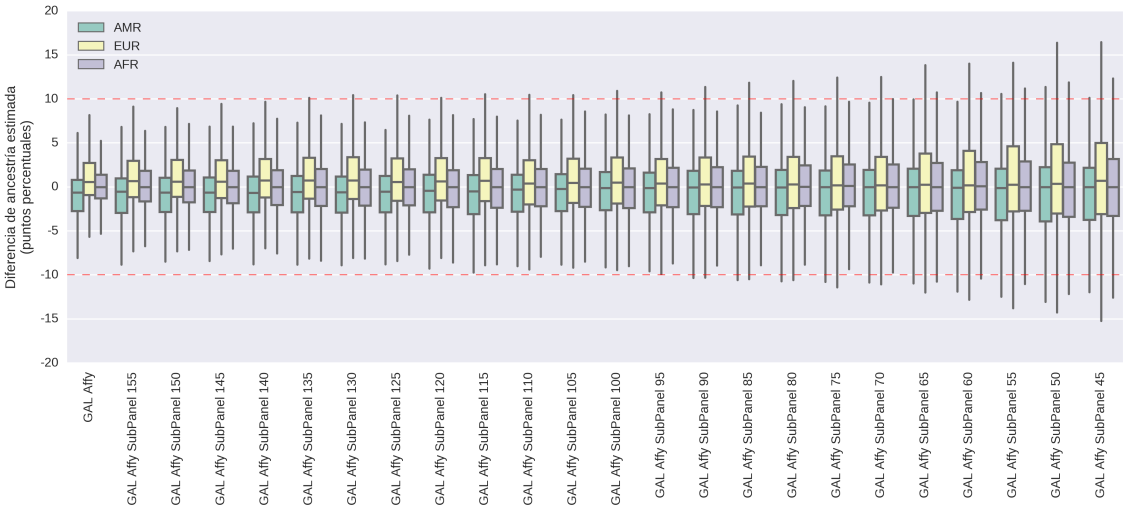


Figura 3.15: Diferencia en las ancestrías estimadas por cada subpanel con respecto a la ancestría estimada por el panel `GAL_Completo`. Al reducir la cantidad de AIMs, la estimación se vuelve más errática y los valores se diferencian cada vez más respecto del panel completo. La dispersión de las diferencias ubica a los subpaneles de más de 95 AIMs en una "zona confiable", definida arbitrariamente para aquellos subpaneles que para todas su muestras estiman ancestrías dentro de un rango de ±10 puntos porcentuales respecto del valor original de `GAL_Completo` (líneas rojas punteadas).

Limitaciones de los paneles `GAL`

Como en la sección anterior, aquí también podemos mostrar la limitación de los paneles `GAL` al comparar sus resultados con los paneles `CP` más allá de tres componentes ancestrales (i.e. con valores de $K > 3$). Ejemplo excelente de esto es el contraste entre la **Figura 3.16** y la **Figura 3.17**. Al correr *admixture* con $K = 8$, el dataset `LEACI` y el panel `CPx100`, descubrimos componentes bien definidos a nivel subcontinental que no se obtienen al utilizar los paneles `GAL`. Las poblaciones `LWK` y `YRI` son asignadas a *clusters* diferentes, replicando la distribución a lo largo del PC 5 en el PCA con el mismo panel. Consecuentemente, la ancestría africana en latinoamericanos también se divide entre ambos componentes (naranja y verde en la figura).

Por otro lado, puede observarse que el *cluster* europeo se divide y aparece un componente común a españoles y latinoamericanos, no compartido por las muestras británicas. Este fenómeno coincide con el hecho de que la ancestría europea predominante en nuestro continente sea española, en particular en los países de origen de las muestras utilizadas. En particular, los puertorriqueños exhiben una mayor proporción de este "componente europeo-ibérico" (marrón) y baja proporción del componente europeo "general" (azul). Esto coincide en cierta medida con lo que Moreno-Estrada *et al.* señalaron como el "componente europeo específico de latinos" [24], producto del cuello de botella que siguió a la colonización española y de la posterior deriva génica.

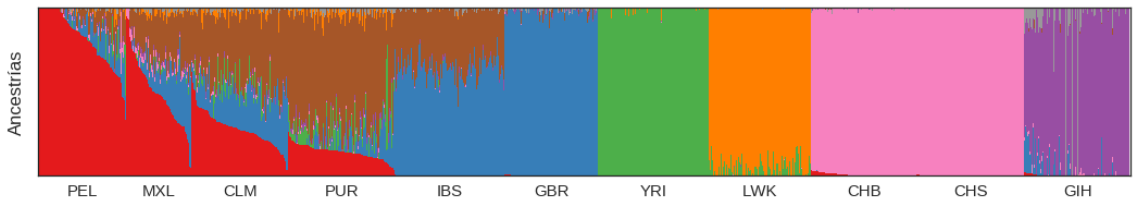


Figura 3.16: Resultados de *admixture* para el dataset `LEACI`, utilizando el panel `CPx100` ($K = 8$). Cada línea vertical representa una muestra y cada color un componente ancestral o cluster. Las poblaciones africanas `YRI` y `LWK` se asignan a *clusters* diferentes (naranja y verde), así como la población india `GIH` (violeta y gris en algunas muestras), mientras que ambas poblaciones chinas `CHS` y `CHB` son asociadas en el mismo *cluster* (rosa). Aparece un componente ibérico-latinoamericano (marrón), ausente en los europeos del norte `GBR`.

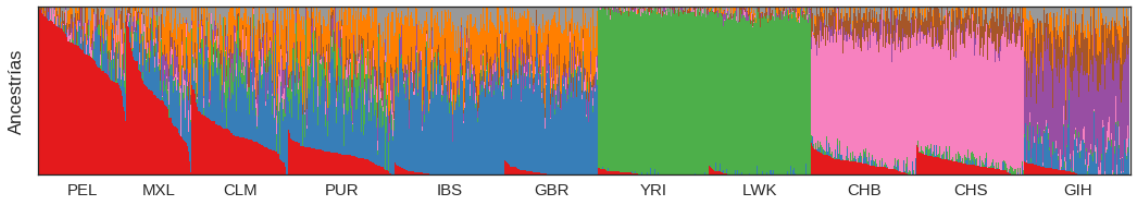


Figura 3.17: Resultados de *admixture* para el dataset `LEACI`, utilizando el panel `GAL_Completo` ($K = 8$). En contraste con la **Figura 3.16**, las poblaciones de Asia son asignadas en parte al *cluster* 'americano' (rojo), dado que el panel confunde en parte la ancestría nativa americana y la asiática. Las dos poblaciones africanas, además, no son diferenciadas y aparecen componentes dispersos (naranja, marrón, gris) sin una lectura demográfica clara.

4 Conclusiones

Nuestro trabajo comenzó motivado por una limitación: del diseño de 445 AIMs para latinoamericanos elegidos por Galanter *et al.* (2012), no todos están disponibles en el *array* comercial LAT 1 de Affymetrix [50]. Al realizar el cruce entre ambos conjuntos de marcadores, encontramos que sólo 179 de los AIMs del panel se encuentran en el *array* LAT 1. Un primer objetivo, entonces, consistió en determinar si este subconjunto de marcadores permite una estimación de ancestrías de igual precisión que la que posibilita el panel completo.

La primera conclusión a la que llegamos es que el panel reducido de 179 AIMs, que denominamos `GAL_Affy`, permite estimar ancestrías con precisión y cumple el mismo objetivo que el panel completo de 445 AIMs. Comprobamos que la precisión se mantiene con dos métodos distintos: un análisis de componentes principales y una corrida de *admixture* con $K = 3$. En ambos casos, las muestras utilizadas fueron de 347 individuos latinoamericanos del proyecto 1000 Genomas, junto a muestras de África y Europa del mismo proyexto, utilizadas como *proxies* de poblaciones ancestrales.

El segundo objetivo del trabajo consitió en encontrar una cantidad mínima de AIMs con los cuales pudiera realizarse la misma estimación de ancestría, sin perder precisión. Como vimos, con 95 marcadores seleccionados puede lograrse una precisión cercana a la del panel completo, dentro de 10 puntos porcentuales de diferencia en los casos de mayor error.

No obstante, paneles de menor cantidad de AIMs también permiten distinguir la ancestría de individuos latinoamericanos, si se admite una pérdida de precisión mayor en un porcentaje pequeño de las muestras, mientras que la mayor parte de los individuos será analizado correctamente. La cantidad mínima de SNPs a utilizar dependerá, entonces, de cuánta precisión sea necesario obtener y de cuán costosos o riesgosos sean los errores de estimación.

Así pues, la reducción del costo con un diseño reducido del panel de AIMs se ofrece como una alternativa viable. Esta posibilidad es de particular interés en escenarios donde por limitaciones de presupuesto la única opción factible es incluir unos pocos marcadores de ancestría.

Finalmente, los análisis realizados con el panel `GAL_Completo` y los paneles reducidos fueron también aplicados a paneles de SNPs elegidos al azar a lo largo del genoma. Diseñamos tres paneles de control, con 438, 4.424 y 43.144 marcadores, con la idea de contrastar la estimación basada en esos *loci* con la estimación basada en los AIMs.

Observamos que, a igual número de marcadores, la precisión se reduce drásticamente al utilizar SNPs al azar. En este sentido, los 438 marcadores del panel `CPx1`, en comparación con los 445 marcadores de `GAL_Completo`, producen una estimación de ancestrías mucho más imprecisa. Sólo se iguala el poder informativo de los AIMs al incrementar la cantidad de marcadores al azar un orden de magnitud o más, con paneles como nuestro `CPx10` y `CPx100`.

Una ventaja de los paneles de marcadores al azar es que permiten descubrir *clusters* o ancestrías para las que el panel de Galanter *et al.* (2012) no fue diseñado. Pudimos comprobar esto al graficar más de dos componentes principales y también al correr *admixture* con valores de K mayores a 3, incluyendo poblaciones del sur y del este de Asia. No obstante, esta ventaja se obtiene a un precio elevado: mantener el número de marcadores y lograr menor precisión, o multiplicar el tamaño del panel por 10 o por 100. Puede suponerse, además, que esta capacidad de discriminar otras ancestrías podría ser compensada en un

panel de AIMs con la elección de marcadores específicos de ancestría asiática o de otras regiones.

Como continuación de la investigación desarrollada en este trabajo, es de interés realizar una inferencia de *haplotipos* en las muestras y estimar ancestrías por región en cada cromosoma, en lugar de un valor global para todo el genoma. Un análisis de estas características requeriría aumentar el número de marcadores y controlar su distribución en el genoma, de modo que se ubiquen AIMs a cada lado de los *hotspots* de recombinación. Un desarrollo de este tipo permitiría predecir con mayor certeza si un marcador asociado a enfermedad en un estudio previo estará o no asociado a ese fenotipo en una muestra analizada, según la ancestría del segmento cromosómico que lo contiene.

Bibliografía

[1] Joshua Mark Galanter *et al.* «Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas». En: *PLoS Genetics* 8.3 (8 de mar. de 2012). Ed. por Greg Gibson, e1002554. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002554. URL: http://dx.plos.org/10.1371/journal.pgen.1002554 (visitado 03-03-2016).

[2] Ravi Sachidanandam *et al.* «A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms». En: *Nature* 409.6822 (15 de feb. de 2001), págs. 928-933. ISSN: 0028-0836. DOI: 10.1038/35057149. URL: http://www.nature.com/doifinder/10.1038/35057149 (visitado 01-06-2016).

[3] The International HapMap Consortium. «The International HapMap Project». En: *Nature* 426.6968 (18 de dic. de 2003), págs. 789-796. ISSN: 0028-0836.

[4] The International HapMap Consortium. «A haplotype map of the human genome». En: *Nature* 437.7063 (27 de oct. de 2005), págs. 1299-1320. ISSN: 0028-0836, 1476-4679. DOI: 10.1038/nature04226. URL: http://www.nature.com/doifinder/10.1038/nature04226 (visitado 07-03-2016).

[5] Kelly A. Frazer *et al.* «A second generation human haplotype map of over 3.1 million SNPs». En: *Nature* 449.7164 (18 de oct. de 2007), págs. 851-861. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature06258. URL: http://www.nature.com/doifinder/10.1038/nature06258 (visitado 07-03-2016).

[6] David M. Altshuler *et al.* «Integrating common and rare genetic variation in diverse human populations». En: *Nature* 467.7311 (2 de sep. de 2010), págs. 52-58. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09298. URL: http://www.nature.com/doifinder/10.1038/nature09298 (visitado 07-03-2016).

[7] Adam Auton *et al.* «A global reference for human genetic variation». En: *Nature* 526.7571 (30 de sep. de 2015), págs. 68-74. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature15393. URL: http://www.nature.com/doifinder/10.1038/nature15393 (visitado 07-03-2016).

[8] Richard M. Durbin *et al.* «A map of human genome variation from population-scale sequencing». En: *Nature* 467.7319 (28 de oct. de 2010), págs. 1061-1073. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09534. URL: http://www.nature.com/doifinder/10.1038/nature09534 (visitado 07-03-2016).

[9] Gil A. McVean *et al.* «An integrated map of genetic variation from 1,092 human genomes». En: *Nature* 491.7422 (31 de oct. de 2012), págs. 56-65. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11632. URL: http://www.nature.com/doifinder/10.1038/nature11632 (visitado 03-03-2016).

[10] B. M. Henn, L. L. Cavalli-Sforza y M. W. Feldman. «The great human expansion». En: *Proceedings of the National Academy of Sciences* 109.44 (30 de oct. de 2012), págs. 17758-17764. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1212380109. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1212380109 (visitado 07-03-2016).

[11] J. Z. Li *et al.* «Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation». En: *Science* 319.5866 (22 de feb. de 2008), págs. 1100-1104. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1153717. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1153717 (visitado 02-04-2016).

[12] Q. D. Atkinson. «Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa». En: *Science* 332.6027 (15 de abr. de 2011), págs. 346-349. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1199295. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1199295 (visitado 10-03-2016).

[13] L. L. Cavalli-Sforza, Paolo Menozzi y Alberto Piazza. *The history and geography of human genes*. Princeton, N.J: Princeton University Press, 1994. 541 págs. ISBN: 978-0-691-08750-4.

[14] Erika Tamm *et al.* «Beringian Standstill and Spread of Native American Founders». En: *PLoS ONE* 2.9 (5 de sep. de 2007). Ed. por Dee Carter, e829. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000829. URL: http://dx.plos.org/10.1371/journal.pone.0000829 (visitado 06-03-2016).

[15] Andrew Kitchen, Michael M. Miyamoto y Connie J. Mulligan. «A Three-Stage Colonization Model for the Peopling of the Americas». En: *PLoS ONE* 3.2 (13 de feb. de 2008). Ed. por Henry Harpending, e1596. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0001596. URL: http://dx.plos.org/10.1371/journal.pone.0001596 (visitado 03-03-2016).

[16] David Reich *et al.* «Reconstructing Native American population history». En: *Nature* 488.7411 (11 de jul. de 2012), págs. 370-374. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11258. URL: http://www.nature.com/doifinder/10.1038/nature11258 (visitado 03-03-2016).

[17] Mark A. Jobling *et al.*, eds. *Human evolutionary genetics*. 2nd ed. New York: Garland Science, 2013. 670 págs. ISBN: 978-0-8153-4148-2.

[18] Andrés Ruiz-Linares. «How Genes Have Illuminated the History of Early Americans and Latino Americans». En: *Cold Spring Harbor Perspectives in Biology* 7.6 (jun. de 2015), a008557. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a008557. URL: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a008557 (visitado 01-04-2016).

[19] Cheryl A. Winkler, George W. Nelson y Michael W. Smith. «Admixture Mapping Comes of Age *». En: *Annual Review of Genomics and Human Genetics* 11.1 (sep. de 2010), págs. 65-89. ISSN: 1527-8204, 1545-293X. DOI: 10.1146/annurev-genom-082509-141523. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-082509-141523 (visitado 03-03-2016).

[20] K. Bryc *et al.* «Genome-wide patterns of population structure and admixture in West Africans and African Americans». En: *Proceedings of the National Academy of Sciences* 107.2 (12 de ene. de 2010), págs. 786-791. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0909559107. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0909559107 (visitado 03-03-2016).

[21] Neil Frankel. *The Atlantic Slave Trade And Slavery in America*. URL: http://www.slaverysite.com/ (visitado 02-04-2016).

[22] Julian R. Homburger *et al.* «Genomic Insights into the Ancestry and Demographic History of South America». En: *PLOS Genetics* 11.12 (4 de dic. de 2015). Ed. por Eduardo Tarazona-Santos, e1005602. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005602. URL: http://dx.plos.org/10.1371/journal.pgen.1005602 (visitado 06-03-2016).

[23] I. Silva-Zolezzi *et al.* «Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico». En: *Proceedings of the National Academy of Sciences* 106.21 (26 de mayo de 2009), págs. 8611-8616. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903045106. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0903045106 (visitado 03-03-2016).

[24] Andrés Moreno-Estrada *et al.* «Reconstructing the Population Genetic History of the Caribbean». En: *PLoS Genetics* 9.11 (14 de nov. de 2013). Ed. por Eduardo Tarazona-Santos, e1003925. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003925. URL: http://dx.plos.org/10.1371/journal.pgen.1003925 (visitado 03-03-2016).

[25] Francisco Mauro Salzano y Mónica Sans. «Interethnic admixture and the evolution of Latin American populations». En: *Genetics and Molecular Biology* 37.1 (2014), págs. 151-170. ISSN: 1415-4757. DOI: 10.1590/S1415-47572014000200003. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572014000200003&lng=en&nrm=iso&tlng=en (visitado 06-03-2016).

[26] Carolina Bonilla *et al.* «Admixture analysis of a rural population of the state of Guerrero, Mexico». En: *American Journal of Physical Anthropology* 128.4 (dic. de 2005), págs. 861-869. ISSN: 0002-9483, 1096-8644. DOI: 10.1002/ajpa.20227. URL: http://doi.wiley.com/10.1002/ajpa.20227 (visitado 01-06-2016).

[27] Sijia Wang *et al.* «Geographic Patterns of Genome Admixture in Latin American Mestizos». En: *PLoS Genetics* 4.3 (21 de mar. de 2008). Ed. por Gil McVean, e1000037. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000037. URL: http://dx.plos.org/10.1371/journal.pgen.1000037 (visitado 03-03-2016).

[28] Marc Via *et al.* «History Shaped the Geographic Distribution of Genomic Admixture on the Island of Puerto Rico». En: *PLoS ONE* 6.1 (31 de ene. de 2011). Ed. por John Relethford, e16513. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0016513. URL: http://dx.plos.org/10.1371/journal.pone.0016513 (visitado 07-03-2016).

[29] Alan R. Templeton. «Biological races in humans». En: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44.3 (sep. de 2013), págs. 262-271. ISSN: 13698486. DOI: 10.1016/j.shpsc.2013.04.010. URL: http://linkinghub.elsevier.com/retrieve/pii/S1369848613000460 (visitado 07-03-2016).

[30] K. R. Veeramah y J. Novembre. «Demographic Events and Evolutionary Forces Shaping European Genetic Diversity». En: *Cold Spring Harbor Perspectives in Biology* 6.9 (1 de sep. de 2014), a008516-a008516. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a008516. URL: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a008516 (visitado 03-03-2016).

[31] A. Moreno-Estrada *et al.* «The genetics of Mexico recapitulates Native American substructure and affects biomedical traits». En: *Science* 344.6189 (13 de jun. de 2014), págs. 1280-1285. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1251688. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1251688 (visitado 07-03-2016).

[32] Colm O'Dushlaine *et al.* «Genes predict village of origin in rural Europe». En: *European Journal of Human Genetics* 18.11 (nov. de 2010), págs. 1269-1270. ISSN: 1018-4813, 1476-5438. DOI: 10.1038/ejhg.2010.92. URL: http://www.nature.com/doifinder/10.1038/ejhg.2010.92 (visitado 10-03-2016).

[33] John Novembre y Sohini Ramachandran. «Perspectives on Human Population Structure at the Cusp of the Sequencing Era». En: *Annual Review of Genomics and Human Genetics* 12.1 (22 de sep. de 2011), págs. 245-274. ISSN: 1527-8204, 1545-293X. DOI: 10.1146/annurev-genom-090810-183123. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-090810-183123 (visitado 01-04-2016).

[34] N. A. Rosenberg. «Genetic Structure of Human Populations». En: *Science* 298.5602 (20 de dic. de 2002), págs. 2381-2385. ISSN: 00368075, 10959203. DOI: 10.1126/science.1078311. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1078311 (visitado 01-04-2016).

[35] Christopher Phillips *et al.* «Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation». En: *PLoS ONE* 4.8 (11 de ago. de 2009). Ed. por Lluis Quintana-Murci, e6583. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006583. URL: http://dx.plos.org/10.1371/journal.pone.0006583 (visitado 03-04-2016).

[36] D. H. Alexander, J. Novembre y K. Lange. «Fast model-based estimation of ancestry in unrelated individuals». En: *Genome Research* 19.9 (1 de sep. de 2009), págs. 1655-1664. ISSN: 1088-9051. DOI: 10.1101/gr.094052.109. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.094052.109 (visitado 03-03-2016).

[37] John Novembre y Matthew Stephens. «Interpreting principal component analyses of spatial population genetic variation». En: *Nature Genetics* 40.5 (mayo de 2008), págs. 646-649. ISSN: 1061-4036. DOI: 10.1038/ng.139. URL: http://www.nature.com/doifinder/10.1038/ng.139 (visitado 06-03-2016).

[38] NIH. *The Cost of Sequencing a Human Genome.* URL: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/ (visitado 24-06-2016).

[39] GeneChip Microarrays: Student Manual. *Manufacturing of GeneChip Microarrays and Building Models.* URL: http://www.affymetrix.com/estore/about_affymetrix/outreach/educator/microarray_curricula.affx#1_3.

[40] Thomas J. Hoffmann *et al.* «Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm». En: *Genomics* 98.6 (dic. de 2011), págs. 422-430. ISSN: 08887543. DOI: 10.1016/j.ygeno.2011.08.007. URL: http://linkinghub.elsevier.com/retrieve/pii/S0888754311001960 (visitado 05-03-2016).

[41] Portal web del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). *Avanza la creación de un biobanco genómico para la Argentina.* URL: http://www.conicet.gov.ar/2015/10/22/avanza-la-creacion-de-un-biobanco-genomico-para-la-argentina/ (visitado 22-10-2015).

[42] Shaun Purcell *et al.* «PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses». En: *The American Journal of Human Genetics* 81.3 (sep. de 2007), págs. 559-575. ISSN: 00029297. DOI: 10.1086/519795. URL: http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524 (visitado 03-03-2016).

[43] Python Software Foundation. *Python Language Reference, version 3.5.* URL: http://www.python.org.

[44] Wes McKinney. «Data Structures for Statistical Computing in Python». En: *Proceedings of the 9th Python in Science Conference*. Ed. por Stéfan van der Walt y Jarrod Millman. 2010, págs. 51 -56.

[45] Alkes L Price *et al.* «Principal components analysis corrects for stratification in genome-wide association studies». En: *Nature Genetics* 38.8 (ago. de 2006), págs. 904-909. ISSN: 1061-4036. DOI: 10.1038/ng1847. URL: http://www.nature.com/doifinder/10.1038/ng1847 (visitado 25-04-2016).

[46] Nick Patterson, Alkes L. Price y David Reich. «Population Structure and Eigenanalysis». En: *PLoS Genetics* 2.12 (2006), e190. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.0020190. URL: http://dx.plos.org/10.1371/journal.pgen.0020190 (visitado 06-03-2016).

[47] John D. Hunter. «Matplotlib: A 2D Graphics Environment». En: *Computing in Science & Engineering* 9.3 (2007), págs. 90-95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160265 (visitado 03-03-2016).

[48] Michael Waskom *et al.* «seaborn: v0.7.0 (January 2016)». En: (2016). DOI: 10.5281/zenodo.45133. URL: http://dx.doi.org/10.5281/zenodo.45133 (visitado 04-04-2016).

[49] Fernando Perez y Brian E. Granger. «IPython: A System for Interactive Scientific Computing». En: *Computing in Science & Engineering* 9.3 (2007), págs. 21-29. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.53. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160251 (visitado 03-03-2016).

[50] Affymetrix. *Axiom World Arrays*. URL: http://www.affymetrix.com/catalog/prod640001/AFFY/Axiom%26%23174%3B+World+Arrays#1_1.