Universidade de Santiago de Compostela
Facultade de Medicina e Odontoloxía
Instituto de Ciencias Forenses "Luís Concheiro"

Aplicaciones forenses de las nuevas tecnologías genómicas

Memoria que presenta para optar al Grado de Doctor,

María del Carmen de la Puente Vila

Santiago de Compostela, Febrero 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

La Doctora María Victoria Lareu Huidobro, Catedrática del Departamento de Ciencias Forenses, Anatomía Patolóxica, Xinecoloxía e Obstetricia e Pediatría de la Universidade de Santiago de Compostela,

CERTIFICA:

Que la presente memoria que lleva por título: “Aplicaciones forenses de las nuevas tecnologías genómicas” de la licenciada en Biología por la Universidade de Santiago de Compostela, María del Carmen de la Puente Vila, ha sido realizada bajo mi dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su presentación ante el Tribunal correspondiente.

Y para que así conste, firmamos la presente en Santiago de Compostela, a 17 de Febrero de 2017.

Prof. Dra. María Victoria Lareu Huidobro

María del Carmen de la Puente Vila

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Este trabajo ha sido parcialmente financiado por una Ayuda de apoyo a la etapa predoctoral del Plan Gallego de Investigación, Innovación y Crecimiento 2011-2015 (Plan I2C) para el año 2012 –Xunta de Galicia–; además del proyecto de investigación Bio2013-42188-R del Ministerio de Economía, Industria y Competitividad, 2013-2016.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AGRADECIMIENTOS

Siempre cuentan que esta es la parte más complicada de escribir pero no te lo crees... Y aquí me tenéis, sonriendo al recordar lo "pipioliña" que era –y, seguramente, aún soy– cuando inicié esta *gymkana* y repasando mentalmente los últimos años, pero sin pulsar ni una sola tecla. En mi imaginaria *gymkana* estáis a mi lado, ayudándome y aplaudiendo cada saltito que doy ante minúsculas vallitas. Habéis sacado tiempo para acompañarme en esta mini-aventura y eso supone muchísimo que agradecer y muchos a quienes agradecer, así que vamos a ello –perdonad cierta informalidad, necesaria para reflejar honestamente mis pensamientos: como sabéis, pensando me disperso un poco–.

Y empiezo por el principio, por quien confió en mí y me dio la oportunidad de enrolarme en su fantástico equipo, aún a riesgo de desmerecerlo. Gracias Maviki por tu cercanía y tus consejos, por animarme a dar hoy un poco más que ayer y por guiarme con destreza durante estos años. Espero haberme aproximado a las expectativas que pusiste en mí.

Chris, gracias por escucharme –con infinita paciencia en los días en los que no logro explicarme– y contar siempre con mi opinión de no experta. Durante este tiempo trabajé sintiéndome valorada y, de ese lujo, eres el principal responsable.

Gracias a Ángel por haber plantado esa semilla que creció con la ayuda de muchos y colocó a Santiago en el mapa de la genética forense. Gracias a Toño por sobreestimar mi capacidad docente.

A las "niñas del cole" quiero agradecerles que supieran ver más allá de mi "toxidez" y que me formaran como investigadora, siempre sin descuidar la parte personal. Gracias a Ana Freire por mantener en secreto mi "patosidad" inicial en el laboratorio, a Fonde por los momentos de sierra y optimista optimización, a Lore por instaurar ese *feeling happy flower* tan necesario en ciertas ocasiones, a Jaco por la complicidad de biblioteca y la resolución de conflictos matemáticos, a Niaz por trasladarme a otros lugares con sus relatos, y a las visitas italianas Laura y Stefania porque aunque fue breve, también fue intenso. Gracias Vane por nuestras comidas "morriña de barrio" modo *on* y tus lecciones de habilidades sociales. Gracias Meluchis por tu excelente trabajo como comisión de festejos: en tu liderazgo y poder de convocatoria recae la responsabilidad de mantener las tradiciones. Gracias Anita por ser mi *follower number one*, debo reconocer que a veces apago el filtro sólo para provocar tu encantadora risa. Gracias Raq por las conversaciones terapéuticas en Thor, por siempre tener hueco para mis "brozas" y por tu cariño –de la bondad todo lo que sale es bueno–. Gracias Carliña por ser "mi dos", por insistir y resistir conmigo frente a las adversidades del análisis de datos y por permanecer ahí desde la no tan lejana lejanía.

A mis padres quiero darles las gracias por enseñarme que todo esfuerzo tiene su recompensa, por dejarme libertad para tomar mis decisiones y apoyarme para lograr mis objetivos: que siempre confiarais en mi criterio es mi principal red de seguridad. Gracias Migui por suponer el punto de referencia atemporal: contigo siempre todo es como siempre, sin que afecten los años o la distancia. Gracias al resto de la familia, por estar siempre pendientes de cómo me va por el mundo de la genética forense.

Y por último y no menos importante, gracias a mis amigos por proporcionarme todos esos "momentos de escape". Como soy realmente afortunada, a partir de estas líneas no solo hay calidad sino también cantidad, así que intentaré ser breve y daros la "chapa" individualmente según vayamos viéndonos.

Gracias a mis "pontinos": Marta –en especial a ti, que regalas tu talento–, Mayte, Patri, Miguel, Prada, Marcos, Jordi, Bita, Rebe, Ledi, Néstor, Iria, Bruno, Guille, Roberto, Ángela y Alma. Como alguien me hizo saber en cierta ocasión especial: los años nos maduran pero no nos separan. Nos quedan muchos éxitos y fracasos que compartir y seguiremos madurando juntos: siempre nos imagino con muchos años más solucionando el mundo durante una interminable partida en cualquier parque del barrio. Gracias a Carmen y Noa, por nuestras ocasionales pero merecidas cenas de actualización.

Gracias a mis coleguillas de "la capi", que fui reclutando para la causa durante la carrera y el máster: Patri, Colo, Laura Vidal, Alba, Elena, Iago, Seo, Laura Martínez, Cynthia, Paula y Jaime. Siempre es buen momento para reírse de uno mismo en buena compañía e intercambiar anécdotas y nuevos conocimientos. Es difícil conectar en poco tiempo, pero con todos vosotros tengo una "banda ancha" que va como un tiro de bien. Gracias, Brais – compartiendo trinchera, la lucha es más llevadera.

"Keep an open mind – but not so open that your brain falls out"

Richard Feynman

Contenido

Abreviaturas

~ – aproximadamente
A – adenina
A-SNPs – SNPs autosómicos
AB – Applied Biosystems
ADN – ácido desoxirribonucleico
ADNmt – ADN mitocondrial
AEC – antes de la era común
AFR – África subsahariana
AIM – *ancestry informative marker*
Al. – alelo
Alt. – alternativo
AMR – América
ARF – *allele read frequency*
ARN – ácido ribonucleico
BGA – *bio-geographical ancestry*
C – citosina
CCD – *charge-coupled device*
CE – *capillary electrophoresis*
CEPH – *Centre d'Etude du Polymorphisme Humain*
CI – código interno
CODIS – *Combined DNA Index System*
Conc. – concentración
Cr. – cromosoma
dbSNP – base de datos de SNPs del NCBI
ddNTP – didesoxinucleótido trifosfato
dNMP – desoxinucleótido monofosfato
dNTP – desoxinucleótido trifosfato
Dp – *discrimination power*
EAS – este de Asia
ESS – *European Standard Set*
et al. – *et alii*
EUR – Europa
EVCs – *externally visible characteristics*
Fig. – figura
F_{ST} – índice de fijación
G – guanina

GWAS – *genome-wide association studies*
HGDP – *Human Genome Diversity Project*
HLA – *human leukocyte antigen*
HV – *hypervariable region*
ID – *identity*
IGV – *integrative genomics viewer*
I_n – *informativeness-for-assignment*
Indel – polimorfismo de inserción/deleción
Infor. – informatividad
ISP – Ion Sphere™ *particle*
Long. – longitud
LR – *likelihood ratio*
MAF – *minor allele frequency*
Mb – megabases
MCMC – *Markov chain Monte Carlo*
ME – Oriente Medio
Mpb – mega pares de bases
MPS – *massively parallel sequencing*
N.º – número
NCBI – *National Center for Biotechnology Information*
NGS – *next generation sequencing*
OCE – Oceanía
Orient. – orientación
p. ej. – por ejemplo
P1000G – Proyecto 1000 Genomas
pb – pares de bases
PC – *principal component*
PCA – *principal component analysis*
PCR – reacción en cadena de la polimerasa
PHR – *peak height ratio*
PPi – pirofosfato
PSD – *population-specific Divergence*
R^2 – coeficiente de determinación
Ref. – referencia
RFUs – *relative fluorescence units*
RMP – *random match probability*
SAS – sur de Asia
SBE – *single base extension*
sd – desviación típica
SNP – *single nucleotide polymorphisms*
SS – *single-source*

STRs – *short tandem repeats*
T – timina
TFS – Thermo Fisher Scientific
TS – *software* Torrent Suite™
v. – versión
vs. – *versus*
Y-SNPs – SNPs de cromosoma Y

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Prefacio

Esta memoria para optar al Grado de Doctor comprende una serie de investigaciones interrelacionadas, propuestas en base a los objetivos iniciales, que permiten expandir las herramientas disponibles para el análisis en genética forense. Así, la memoria se organiza de acuerdo con la siguiente estructura:

En primer lugar, se realiza una breve revisión histórica de la genética forense y se especifican las metodologías y aplicaciones actuales de esta rama del conocimiento. Esta introducción –sección 1– permite identificar las limitaciones actuales y las nuevas metodologías disponibles, profundizando especialmente en aquellas aplicaciones en las que se enmarcan las investigaciones presentadas.

En segundo lugar y atendiendo al contexto actual de la genética forense, se exponen los tres objetivos principales –sección 2– propuestos para esta tesis: por una parte, explorar las posibilidades que ofrece el uso de las nuevas tecnologías en el campo de la genética forense y, por la otra, desarrollar nuevos paneles para la predicción de ancestralidad biogeográfica y el análisis de mezclas de ADN. Para llevar a cabo el desarrollo experimental de los objetivos principales se proponen seis objetivos específicos que representan seis trabajos de investigación.

En tercer lugar se desarrollan individualmente cada uno de los seis trabajos, incluyendo material y métodos, resultados y discusión; y agrupándolos en tres bloques titulados:

- Bloque I: ID-SNPS en MPS –sección 3–
- Bloque II: Ancestralidad biogeográfica –sección 4–
- Bloque III: Mezclas de ADN –sección 5–

En cuarto lugar se realiza una discusión final –sección 6– que engloba los tres bloques y argumenta las limitaciones y las aportaciones de los trabajos expuestos al campo de la genética forense.

Por último se presentan las conclusiones –sección 7–, fruto del proceso experimental y de los resultados obtenidos en cada investigación.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. Introducción

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. Introducción

A modo de introducción de los trabajos presentados en esta tesis: (i) se contextualiza históricamente la genética forense –sección 1.1–; (ii) se presentan los diferentes tipos de marcadores del genoma humano que se utilizan comúnmente en genética forense y los sistemas disponibles para el análisis de los mismos –sección 1.2–; se exponen (iii) las diferentes aplicaciones de la genética forense humana –sección 1.3–; y (iv) la metodología de validación de nuevas técnicas –sección 1.4–.

1.1 LA GENÉTICA FORENSE: CONCEPTO Y EVOLUCIÓN HISTÓRICA

La ciencia forense es un área multidisciplinar que engloba históricamente todas aquellas ramas en las que se realizan labores de peritaje legal –sección 1.1.1–. A partir de finales del siglo XIX, las investigaciones sobre el material hereditario –sección 1.1.2.1– impulsan el nacimiento de la genética forense, que se desarrolla rápidamente a partir de mediados del siglo XX gracias a los avances en las técnicas de análisis de ADN –sección 1.1.2.2–.

1.1.1 La ciencia forense y la genética forense

Etimológicamente, la palabra "forense" deriva del latín *forensis* como perteneciente o relativo al foro (Real Academia Española 2014). En las ciudades del Imperio Romano, el foro era la plaza en la que, entre otras actividades como el comercio o la religión, se llevaban a cabo los juicios públicamente. Así, se aplica el adjetivo forense a aquellas disciplinas que sirven de apoyo a la justicia. En un sentido amplio, la ciencia forense es la aplicación de las técnicas y los principios científicos con el fin de aportar evidencias a investigaciones o determinaciones legales (Tilstone *et al.* 2006). La ciencia forense es un área multidisciplinar que engloba todas aquellas ramas que realizan labores de peritaje legal: psicología, toxicología, antropología, odontología, ingeniería… y, entre ellas, la genética.

Como ejemplos de la importancia de las ciencias forenses en la antigüedad romana han transcendido la autopsia de Julio César –en el año 44 AEC el físico Antistio examinó el cadáver y determinó que tan sólo una de las 23 puñaladas que recibió, la que le atravesó el pecho, causó su muerte– o la exoneración de un hombre ciego acusado de matar a su madre en el siglo I AEC –el jurista y orador Quintiliano presentó como prueba las huellas halladas en el lugar del crimen–. El primer tratado sistemático de medicina forense fue escrito durante el siglo XIII en China, por el jurista Sung Tzhu. El libro, titulado *Hsi Yuan Lu* –traducción aproximada "Cómo evitar los errores"– trata diversos temas como la identificación de armas y heridas o la diferenciación de casos de estrangulación y ahogamiento. A partir del siglo XIX proliferan los ensayos de expertos en diferentes áreas como la toxicología –Mateu Orfila–, la identificación individual mediante las medidas craneales –Alphonse Bertillon–, la

identificación individual mediante las huellas dactilares –Francis Galton–, el análisis de la escritura manual –Albert Osborn– o la balística –Calvin Goddard– (Tilstone *et al.* 2006).

Paralelamente al desarrollo de otras áreas de la ciencia forense, a comienzos del siglo XX se inicia la hemogenética forense –precursora de la genética forense– con el descubrimiento del sistema eritrocitario ABO (Landsteiner 1900). Extendiendo esta línea, se desarrollaron polimorfismos de expresión de otros antígenos (MNS, Rh, Lewis…) y enzimas (glioxilasa, fosfatasa ácida) eritrocitarios; así como proteínas del suero (haptoglobina, transferrina) y antígenos leucocitarios (sistema HLA –*human leukocyte antigen*–). No obstante, las proteínas son inestables en condiciones ambientales, se expresan diferencialmente en los tejidos y la obtención de perfiles altamente discriminatorios se ve muy limitada por la cantidad de material biológico disponible (Goodwin *et al.* 2011). Estos inconvenientes fueron solventados posteriormente por la genética forense.

1.1.2 **La revolución del ADN: la genética forense**

La genética forense es el área de las ciencias forenses que se define como la aplicación de la genética a la resolución de conflictos legales. Entre las posibles aplicaciones se encuentran las pruebas de parentesco, el genotipado de muestras biológicas encontradas en casos de criminalística, la identificación de restos cadavéricos de catástrofes masivas…

La genética se basa en el estudio del material hereditario para el análisis de las variaciones inter e intra-específicas de las poblaciones. Las frecuencias poblacionales de estas variaciones están determinadas por la actuación de las fuerzas evolutivas –mutación y recombinación, selección, deriva genética y migración o flujo genético–. No obstante, las fuerzas evolutivas no actúan de manera uniforme sobre todo el material hereditario: aquellas regiones que influyen en la adaptación de los individuos al medio o a su tasa de reproducción tienden a conservarse, mientras que las regiones no sometidas a selección natural –neutras– acumulan más variaciones (Jobling *et al.* 2004c).

1.1.2.1 La búsqueda del material hereditario

Las teorías evolutivas de mediados del siglo XIX (Darwin y Wallace 1858, Darwin 1859) carecían de la base fundamental de su argumento: no existía un modelo mecanístico de herencia de los caracteres. El desarrollo de estas teorías incentivó la búsqueda del material hereditario: los avances en citología determinaron que se encontraba en el núcleo (Haeckel 1866) y de éste se aisló una substancia muy abundante que fue llamada “nucleína” (Miescher 1871). La fracción no proteica de la “nucleína” se identificó como un ácido compuesto por cuatro bases nitrogenadas, un azúcar –ribosa– y fosfato (Kossel 1886, Levene 1919) que forma una estructura regular (Astbury 1947) en la que dos cadenas antiparalelas se disponen en forma de doble hélice (Franklin y Gosling 1953, Watson y Crick 1953, Wilkins *et al.* 1953): el ADN o ácido desoxirribonucleico.

El ADN se transmite (Griffith 1928, Avery *et al.* 1944, Hershey y Chase 1952) de células madres a hijas a través de la replicación semiconservativa (Meselson y Stahl 1958): las cuatro

posibles bases –A o adenina, C o citosina, G o guanina y T o timina– se enfrentan siempre como C-G y A-T, de manera que una única cadena sirve de molde para la síntesis de la complementaria.

La base de los polimorfismos de expresión que analiza la hemogenética reside también en el ADN, ya que todas las instrucciones necesarias para la síntesis de proteínas se encuentran codificadas en el mismo (Crick 1970). Sin embargo, el análisis de ADN permite acceder a un mayor nivel de polimorfismo –las variaciones en el ADN no codificante son indetectables mediante el análisis de proteínas y la expresión se realiza de acuerdo a un código genético degenerado (Nirenberg y Matthaei 1961) de manera que las variaciones sinónimas son detectables únicamente mediante el análisis de ADN–. Además de aumentar el nivel de polimorfismo, el uso de ADN conlleva las ventajas de mayor estabilidad ambiental y uniformidad en los tejidos.

1.1.2.2 De la ADN *fingerprint* a los perfiles de STRs

La Fig. 1 refleja los principales avances relacionados con la genética forense en el siglo XX, desde los inicios de su precursora, la hemogenética forense. El análisis de los polimorfismos de ADN se inició a partir de los años 1960 gracias al desarrollo de las endonucleasas de restricción, enzimas que reconocen una secuencia característica y cortan el ADN en ese punto concreto, y el *Southern blotting*, una técnica que permite separar fragmentos de ADN en función de su longitud y posteriormente detectarlos mediante la hibridación de una sonda marcada radioactivamente (Southern 1975). También se desarrollan en estos años las primeras metodologías de secuenciación (Maxam y Gilbert 1977, Sanger *et al.* 1977).

La combinación de endonucleasas de restricción y *Southern blotting* permitió el análisis de polimorfismos minisatélites o VNTR –*Variable Number of Tandem Repeats*, repeticiones de tamaño variable de secuencias de entre 9 y 100 pares de bases (pb)– mediante el uso de sondas multi-*locus* (Jeffreys *et al.* 1985). Se obtienen así patrones de bandas que detectan el polimorfismo en varios *loci* simultáneamente (Jeffreys *et al.* 1985). Los patrones de bandas conforman una huella de ADN –*fingerprint*– específica que permite la identificación individual (Gill *et al.* 1985). La ADN *fingerprint* presentaba varias limitaciones: era necesaria una alta cantidad de ADN no degradado para la obtención de la huella y la interpretación de las huellas o la comparación de los resultados entre laboratorios eran complejas. Posteriormente, con la finalidad de permitir interpretaciones más sencillas, las sondas multi-*locus* se sustituyen por sondas uni-*locus* (Wong *et al.* 1987), de manera que se inició la obtención de perfiles –*profiling*– de ADN: se detectan una o dos bandas en individuos homocigotos o heterocigotos, respectivamente, que corresponden inequívocamente a alelos del *locus* a analizar (Decorte 2010).

Sin embargo, el inconveniente de la sensibilidad no fue solventado hasta el descubrimiento de la reacción en cadena de la polimerasa –PCR: *polymerase chain reaction*– (Mullis *et al.* 1986), que permitió reducir drásticamente la cantidad inicial de ADN: un

análisis mediante *Southern blot* requería ~5-10 µg de ADN, mientras las reacciones de PCR requieren ~1000 veces menos (Jobling *et al.* 2004b). Mediante esta técnica, se generalizó el análisis de los polimorfismos STR –*short tandem repeats*– o microsatélites como marcadores de rutina y se establecieron altos niveles de estandarización y calidad de las pruebas (Goodwin *et al.* 2011).

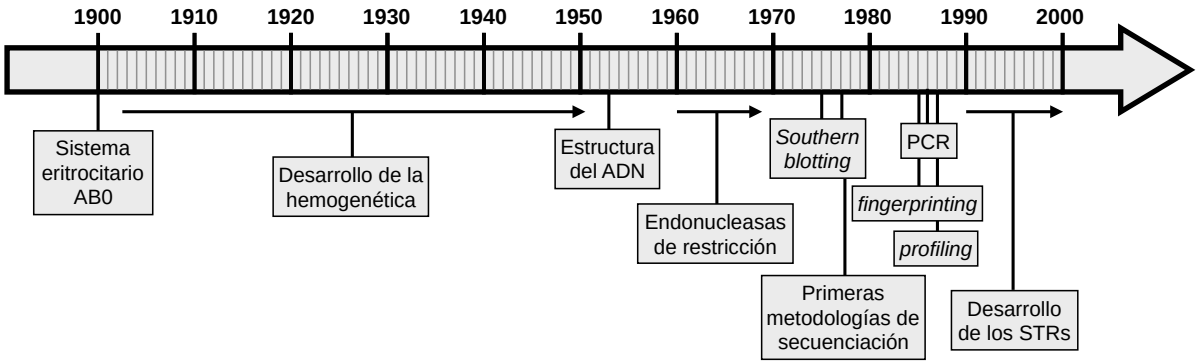


Fig. 1. Principales avances relacionados con la genética forense durante el siglo XX.

1.2 POLIMORFISMOS COMÚNMENTE UTILIZADOS EN GENÉTICA FORENSE

En esta sección se describen los diferentes componentes del genoma humano que se analizan en genética forense –sección 1.2.1–; así como los tipos de polimorfismos más usados –sección 1.2.2– y sus metodologías de análisis –sección 1.2.3–.

1.2.1 Los polimorfismos en el genoma humano

En las células humanas, el ADN se encuentra confinado en dos orgánulos celulares: el núcleo y las mitocondrias. El ADN nuclear consta de ~6400 Mpb (dotación diploide) en cadenas lineales que se empaquetan formando cromosomas: cada núcleo contiene dos copias de cada uno de los 22 autosomas y un par sexual. El par sexual determina el sexo del individuo y puede estar formado por dos cromosomas X –mujer– o un cromosoma X y uno Y –hombre–. Todos los autosomas sufren procesos de recombinación entre pares homólogos durante la profase I de la meiosis. Durante la ovogénesis los dos cromosomas X pueden recombinar a lo largo de toda su longitud; en la espermatogénesis tan solo se produce recombinación entre los cromosomas X e Y en dos pequeños segmentos denominados región pseudoautosómica que representan el ~5% del cromosoma Y. Así, toda la región no recombinante del cromosoma Y –un ~95% del mismo– se hereda como un marcador de linaje patrilineal. A su vez, la tasa de recombinación del cromosoma X es menor, de manera que los bloques de desequilibrio de ligamiento son mas extensos en este cromosoma.

El ADN mitocondrial –ADNmt– está constituido por una única molécula circular de 16569 pb (Anderson *et al.* 1981) de la que existen ~2-10 copias en cada una de las miles de mitocondrias que hay en cada célula. El ADNmt no recombina –no sufre meiosis– y las

mitocondrias de los espermatozoides no pasan a formar parte del material genético del zigoto, de manera que se hereda como marcador de linaje matrilineal. El 90% del ADNmt está constituido por una región codificante muy compactada y conservada. El 10% restante forma la región control, más polimórfica, que contiene dos regiones hipervariables –HVI y HVII (Vigilant *et al.* 1989)– de alto interés en genética forense. Se estima que el ADNmt tiene una tasa de sustitución ~10 veces mayor a la del nuclear (Brown *et al.* 1979), aunque no es uniforme a lo largo de la molécula (Pesole *et al.* 1999). Esta alta tasa de mutación podría ser consecuencia de la ausencia de histonas, de la baja fidelidad de los sistemas de reparación del ADNmt o de la alta presencia de especies reactivas de oxígeno en la mitocondria.

Se estima que el ~99,9% de la dotación genética humana es idéntica entre dos individuos al azar del mismo sexo. Dentro del restante ~0,1% se encuentran las variaciones genotípicas que provocan toda la variabilidad de los individuos, incluida la variabilidad fenotípica. Los *loci* polimórficos son aquellos que presentan múltiples alelos con frecuencias mayores al 1% en una población. La genética forense se sirve polimorfismos de ADN presentes tanto en el ADN mitocondrial como en el nuclear, incluyendo autosomas y el par sexual. En la Tabla 1 se presentan las ventajas e inconvenientes de cada componente del genoma humano, que derivan de sus características y su modelo de herencia.

Tabla 1. Características, ventajas e inconvenientes del ADN nuclear y mitocondrial en genética forense.

	ADN nuclear			ADN mitocondrial
	Autosomas	Cromosomas sexuales		
		Cromosoma X	Cromosoma Y	
Molécula	Cromosomas lineales			Una molécula circular
Nº. de bases	6400 Mpb (dotación diploide)			16569 pb
Copias / célula	2 x 22 cromosomas	– 2 célula femenina – 1 célula masculina	– 1 célula masculina	~2-10 copias / miles de mitocondrias / célula
Regiones	–	–	– Pseudoatosómica: ~5% – No recombinante: ~95%	– Codificante: ~90% – Control: ~10% (HVI y HVII)
Polimorfismos comunes	STRs, SNPs e Indels			SNPs e Indels (principalmente en HVI y HVII)
Recombinación	Sí	Sí (menor que en autosomas)	Región no recombinante= marcador uniparental patrilineal	Marcador uniparental matrilineal
Ventajas de su uso en Genética Forense	– Permite individualizar – Marcadores no sesgados de ancestralidad biogeográfica (AIMs) – Marcadores para la predicción de características fenotípicas	– Puede aportar información en pedigrís complejos	– Puede aportar información por vía paterna en pedigrís complejos – Puede aportar información en mezclas de ADN con componente masculino – Marcador de linaje (origen biogeográfico sesgado)	– Alto número de copias por célula: ventajas en análisis de ADN degradado y ADN *low template* – Puede aportar información por vía paterna en pedigrís complejos – Marcador de linaje (origen biogeográfico sesgado)
Inconvenientes de su uso en Genética Forense	– Bajo número de copias por célula: limitaciones en análisis de ADN degradado y *low template*	– Bajo número de copias por célula: limitaciones en análisis de ADN degradado y *low template* – No individualiza	– Bajo número de copias por célula: limitaciones en análisis de ADN degradado y *low template* – No individualiza	– No individualiza

1.2.2 **Tipos de polimorfismos utilizados comúnmente en genética forense**

Los polimorfismos de ADN de los que actualmente se sirve la genética forense son los más cortos y abundantes del genoma: STRs, SNPs e Indels. Cada una de estas clases de marcadores presenta diversas aplicaciones, ventajas e inconvenientes que derivan de su naturaleza biológica o de su metodología de análisis. Un resumen de las características de cada marcador se presenta en la Tabla 2.

1.2.2.1 STRs

Los STRs –*short tandem repeats*– o microsatélites son reiteraciones en serie de unidades de repetición –núcleos– de entre 2-7 pb (Butler 2012a). Se estima que representan un ~3% del genoma humano y están presentes en todos los cromosomas con una densidad promedio de ~14000 pb/Mpb (Subramanian *et al.* 2003).

Los STRs son polimorfismos de longitud, de manera que los diferentes alelos vienen determinados por el número de repeticiones del núcleo. El alto grado de polimorfismo de estos marcadores deriva de su alta tasa de mutación, que se calcula en torno a 10^{-3}-10^{-4}. El principal mecanismo subyacente a la alta tasa de mutación de los STRs es el *slippage*: durante la replicación del ADN se produce la desalineación de al menos una unidad de repetición entre la cadena molde y la de síntesis, que –si no es reparada por los mecanismos celulares– produce un cambio en el número de repeticiones de la cadena de síntesis respecto a la cadena molde (Fan y Chu 2007). La tasa de mutación de los STRs no es homogénea y depende, entre otros: de la longitud de la unidad de repetición o su secuencia –siendo mayor en los que tienen unidades de repetición más cortas–, de la complejidad de la unidad de repetición –a mayor complejidad menor tasa de mutación– o de la longitud total del alelo –los de mayor longitud tienen una mayor tasa de mutación y tienden a perder repeticiones mientras que los de menor longitud tienen una menor tasa de mutación y tienden a ganar repeticiones– (Brinkmann *et al.* 1998, Ellegren 2000, Ellegren 2004, Eckert y Hile 2009).

Los microsatélites se pueden clasificar en función del número de bases del núcleo como: dinucleótidos, trinucleótidos, tetranucleótidos, pentanucleótidos… –alternativamente: diméricos, triméricos, tetraméricos, pentaméricos…–. Los más comunes en genética forense cuentan con unidades de repetición de entre 4 y 5 pb, debido al balance entre número de *loci* que existe en el genoma, grado de polimorfismo y estabilidad. Además, los STRs se clasifican en función de su grado de variabilidad: los STRs simples están formados por unidades de repetición homogéneas, los STRs compuestos tienen varias unidades de repetición simples adyacentes del mismo tamaño y los STRs complejos varios bloques de unidades de repetición de diferente tamaño variable. En todos los casos, pueden tener secuencias de nucleótidos intercalados que interrumpen las repeticiones o éstas pueden ser incompletas, generándose alelos intermedios.

La principal ventaja de los STRs es su alto grado de polimorfismo. Así, los STRs autosómicos constituyen los marcadores de elección –sección 1.3.1.1– para realizar

identificaciones individuales y análisis de parentesco y se encuentran implementados en las bases de datos de ADN. En mezclas de ADN, el carácter multialélico de los STRs facilita la deconvolución de los componentes y el uso de STRs de cromosoma Y permite obtener información sobre el componente masculino –sección 1.3.1.2.3–. En pruebas de parentesco, el poder de discriminación que ofrecen conjuntos de ~21 STRs es suficiente en la mayoría de los casos; no obstante, en casos de pedigrís complejos puede ser necesario elevar el poder de discriminación –sección 1.3.1.2.4–.

En casos que conllevan el análisis de ADN degradado –sección 1.3.1.2.1– los STRs tienen ciertas limitaciones derivadas de la longitud del polimorfismo. Además, la información que aportan en las investigaciones criminales cuando no existe una coincidencia de perfiles entre un vestigio biológico de la escena del crimen y un sospechoso o las bases de datos de perfiles de ADN –sección 1.3.2– es limitada.

1.2.2.2 SNPs

Los SNPs –*single nucleotide polymorphisms*– son polimorfismos de secuencia, de manera que los diferentes alelos vienen determinados por variaciones de una sola base en un punto particular del genoma (Butler 2012e). Los SNPs son los polimorfismos más abundantes del genoma humano: más de 84 millones de SNPs bialélicos han sido identificados en la Fase III del Proyecto 1000 Genomas (The Genomes Project Consortium 2015).

Las mutaciones que originan los SNPs son producidas principalmente por incorporaciones erróneas de nucleótidos durante la replicación, por modificaciones químicas de las bases –análogos de bases, agentes modificadores de bases o intercalantes– o por daños físicos en el ADN –radiación ionizante– (Jobling *et al.* 2004a). La tasa de mutación se establece en torno a $2{,}3 \times 10^{-8}$ (Nachman y Crowell 2000), aunque no es homogénea a lo largo del genoma. Se debe destacar que la tasa de transiciones[1] es más del doble que la de transversiones[2], pese a existir dos posibles transversiones y una única transición para cada base (Zhang y Gerstein 2003). La baja tasa de mutación confiere a estos marcadores una alta identidad por descendencia –es mucho más probable que dos alelos derivados iguales provengan de uno ancestral mediante un único evento mutacional que mediante dos eventos independientes– y un bajo grado de polimorfismo. De este modo, los SNPs son marcadores típicamente bialélicos, pero existen SNPs multialélicos –trialélicos e incluso tetraalélicos (Phillips *et al.* 2015)– con una menor representación: se han identificado más de 270000 en la Fase III del Proyecto 1000 Genomas (The Genomes Project Consortium 2015).

Los SNPs presentan diferentes ventajas en genética forense ya que, en función de sus características, se pueden escoger conjuntos de marcadores para diferentes aplicaciones (Budowle y van Daal 2008): (i) SNPs de identificación, que permiten individualizar y requieren altos niveles de heterocigosidad y baja heterogeneidad entre poblaciones; (ii) SNPs

[1] Transición: cambio de una base pirimidínica (C, T) a otra pirimidínica; o de una púrica (A, G) a otra púrica.
[2] Transversión: cambio de una base pirimidínica (C, T) a una púrica (A, G), o viceversa.

de características fenotípicas, que permiten establecer probabilidades de que un individuo presente una determinada característica fenotípica; (iii) SNPs de origen biogeográfico, con bajos niveles de heterocigosidad y alta heterogeneidad entre poblaciones; y (iv) SNPs de linaje, que conforman haplotipos de ADNmt y cromosoma Y.

Los SNPs de identificación presentan ventajas en casos de ADN degradado –sección 1.3.1.2.1–, debido al pequeño tamaño de los marcadores. No obstante, presentan limitaciones en las mezclas de ADN debido a su carácter generalmente bialélico –sección 1.3.1.2.3–. Estos SNPs se pueden utilizar en pruebas de parentesco complejas –sección 1.3.1.2.4–, en las que presentan ventajas derivadas de la alta identidad por descendencia de los marcadores, por sí mismos o como complemento de los STRs. Los SNPs de linaje identifican haplotipos de cromosoma Y y ADNmt, pero no permiten individualizar. No obstante, pueden aportar información para identificaciones, principalmente SNPs de ADNmt en casos de ADN degradado –sección 1.3.1.2.2– y de ADN *low template* –sección 1.3.1.2.2–, y para pruebas de parentesco complejas –sección 1.3.1.2.4–. Además, en casos en los que no existe una coincidencia de perfiles de ADN, proporcionan información que puede guiar las investigaciones policiales. Para ello, se utilizan SNPs de linaje y SNPs de predicción de origen biogeográfico –sección 1.3.2.1– y características fenotípicas –sección 1.3.2.2–.

1.2.2.3 Indels

Los Indels –denominados mediante la contracción de los términos inserción/deleción– son polimorfismos de longitud, de manera que los diferentes alelos vienen determinados por el cambio en el número de bases de una secuencia en un punto determinado del genoma (Butler 2012e). Cuando el cambio en el número de bases implica a una única base se les puede denominar SNPs, aunque los mecanismos que generan ambos polimorfismos no son análogos (Jobling *et al.* 2004a). Los Indels más frecuentes son aquellos en los que el cambio de longitud implica unas pocas bases: tan sólo el 4% de los Indels presentan diferencias de más de 16 pb (Weber *et al.* 2002).

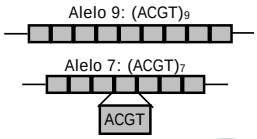
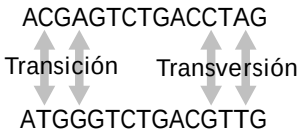
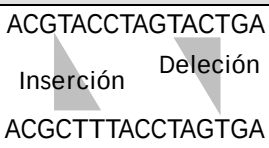
Los Indels constituyen el segundo tipo de polimorfismo más frecuente en el genoma humano, después de los SNPs. En la Fase III del Proyecto Genoma Humano se han identificado más de 3,5 millones de Indels bialélicos y más de 150000 multialélicos (The Genomes Project Consortium 2015).

La tasa de mutación de los Indels se establece en torno a $2{,}3 \times 10^{-9}$ (Nachman y Crowell 2000), muy cercana a la de los SNPs, por lo que comparten con ellos las características de una alta identidad por descendencia y un bajo de grado polimorfismo. La mayoría de las mutaciones de pequeños Indels se producen durante la replicación debido a un alineamiento erróneo de las cadenas, aunque también se producen al replicar las cadenas de ADN que han sido dañadas (Garcia-Diaz y Kunkel 2006).

Los Indels de identificación presentan ventajas y limitaciones similares a los SNPs de identificación, debido a que comparten las características de ser marcadores de pequeño tamaño, generalmente bialélicos y con una alta identidad por descendencia. Así, presentan

ventajas en el análisis de ADN degradado –sección 1.3.1.2.1–; limitaciones en mezclas de ADN (aunque la metodología de análisis presenta ventajas sobre la de los SNPs) –sección 1.3.1.2.3–; y pueden complementar a los STRs en pruebas de parentesco complejas –sección 1.3.1.2.4–. Además, existen paneles de Indels diseñados para la predicción de ancestralidad biogeográfica –sección 1.3.2.1–.

Tabla 2. Características, ventajas y limitaciones de los polimorfismos usados en genética forense.

	STRs	SNPs	Indels
			
Tipo de polimorfismo	Polimorfismo de longitud	Polimorfismo de secuencia	Polimorfismo de longitud
Abundancia en el genoma	~3% del genoma	>84 x 10^6 marcadores	>3,5 x 10^6 marcadores
Tasa de mutación	10^{-3}-10^{-4}	~2,3 x 10^{-8}	~2,3 x 10^{-9}
Grado de polimorfismo	Multialélicos	Generalmente bialélicos	Generalmente bialélicos
Metodología de análisis	PCR con *primers* marcados	Minisecuenciación	PCR con *primers* marcados
Capacidad de *multiplexing*	>15 marcadores	~40 marcadores	>40 marcadores
Tamaño del amplicón	100-500 pb	<100 pb	<100 pb
Información fenotípica	No	SNPs de características fenotípicas	No
Predicción origen biogeográfico	Capacidad limitada	SNPs de origen biogeográfico	Indels de origen biogeográfico
Ventajas de su uso en Genética Forense	- Altamente polimórficos e informativos. - Metodología de análisis sencilla. - Establecidos en las bases de datos de perfiles de ADN. - Deconvolución de mezclas más sencilla debido a su carácter multialélico.	- Tamaño amplicón corto: análisis de ADN degradado. - Baja tasa de mutación: ventajas en pruebas de parentesco complejas. - Predicción de características fenotípicas. - Predicción de origen biogeográfico. -Metodologías que permiten analizar millones de marcadores en ADN de alta calidad y cantidad.	- Metodología de análisis sencilla y combinable con STRs. - Tamaño amplicón corto: análisis de ADN degradado. - Baja tasa de mutación: ventajas en pruebas de parentesco complejas. - Predicción de origen biogeográfico.
Limitaciones de su uso en Genética Forense	- Amplicones largos: limitaciones en el análisis de ADN degradado. - Limitaciones en la predicción de origen biogeográfico. - No informativos para características fenotípicas.	- Bajo poder de discriminación individual de los marcadores. - Metodología de análisis más complicada. - Dificultades en la detección y deconvolución de mezclas de ADN. - No establecidos en bases de datos de perfiles de ADN.	- Bajo poder de discriminación individual de los marcadores. - Dificultades en la detección y deconvolución de mezclas de ADN. - No establecidos en bases de datos de perfiles de ADN.

1.2.3 **Metodologías de análisis de los polimorfismos genéticos**

La estabilidad del ADN a través de los diferentes tejidos permite su análisis a partir de casi cualquier muestra biológica: sangre, saliva, semen, pelo, sudor, hueso, tejidos embebidos en parafina… Los procedimientos a aplicar en la rutina forense se deciden, en cada caso, en función de la información que se desea obtener y del tipo de muestra biológica a analizar. En general, el proceso se inicia con una extracción de ADN adaptada al tipo de vestigio biológico, a partir de la cual se analizan aquellos polimorfismos informativos para cada caso.

1.2.3.1 Metodologías de análisis generales: amplificación y detección

Las diferentes metodologías de análisis de marcadores comparten generalmente un primer paso que consiste en replicar exponencialmente aquellas regiones de interés en cada caso, de manera que se genere una cantidad de las mismas que permita su detección –sección 1.2.3.1.1–. De entre los sistemas de detección disponibles, el más extendido en los laboratorios de genética forense es la electroforesis capilar –sección 1.2.3.1.2– que permite separar fragmentos de ADN de diferente tamaño marcados con fluorocromos.

1.2.3.1.1 Amplificación de la señal mediante PCR

La PCR –*polymerase chain reaction*– o reacción en cadena de la polimerasa es un proceso enzimático en el que una región específica de ADN es replicada varias veces hasta producir un alto número de copias de una secuencia en particular (Mullis *et al.* 1986). Se basa en el uso de la *Taq* polimerasa –una polimerasa altamente termoestable aislada de la bacteria termófila *Thermus aquaticus* (Chien *et al.* 1976)– y oligonucleótidos iniciadores de secuencia específicos –*primers*– que acotan la región de ADN a amplificar (Saiki *et al.* 1988). La PCR requiere cambios cíclicos y precisos en las temperaturas del medio acuoso en el que se produce la reacción, que se controlan mediante un termociclador.

En la Fig. 2 se representan esquemáticamente los componentes, fases y el proceso teórico de amplificación en los primeros ciclos de la reacción. La PCR cuenta con ~30 ciclos en los que se suceden los siguientes pasos: (i) desnaturalización, a ~96°C las cadenas de la doble hélice de ADN se separan entre sí; (ii) *annealing*, los *primers* hibridan con la secuencia molde a una temperatura óptima que depende de la longitud y secuencia de los mismos y que suele estar en torno a los 60°C; y (iii) extensión, la polimerasa copia la cadena molde elongando los *primers*, proceso que habitualmente se realiza a 72°C (temperatura óptima del enzima).

Teóricamente, en cada ciclo se dobla secuencialmente el número de copias iniciales del amplicón –la secuencia comprendida entre los *primers*–. La representación del número de copias del fragmento en cada ciclo frente al número de ciclos sería idealmente una curva exponencial. No obstante, la eficiencia de la reacción nunca es total: no se desnaturaliza todo el ADN, no se produce el *annealing* de los *primers* en todas las cadenas molde existentes o nuevas (sobre todo en los primeros ciclos), la polimerasa pierde eficacia a lo largo de los ciclos… Además, a partir de los 32 ciclos la tendencia exponencial se convierte en un *plateau*

ya que los propios reactivos –los nucleótidos activados (dNTPs), el Mg^{++} (cofactor de la polimerasa), el *buffer* de reacción (que mantiene las condiciones adecuadas de pH del medio), los *primers*…– constituyen factores limitantes.

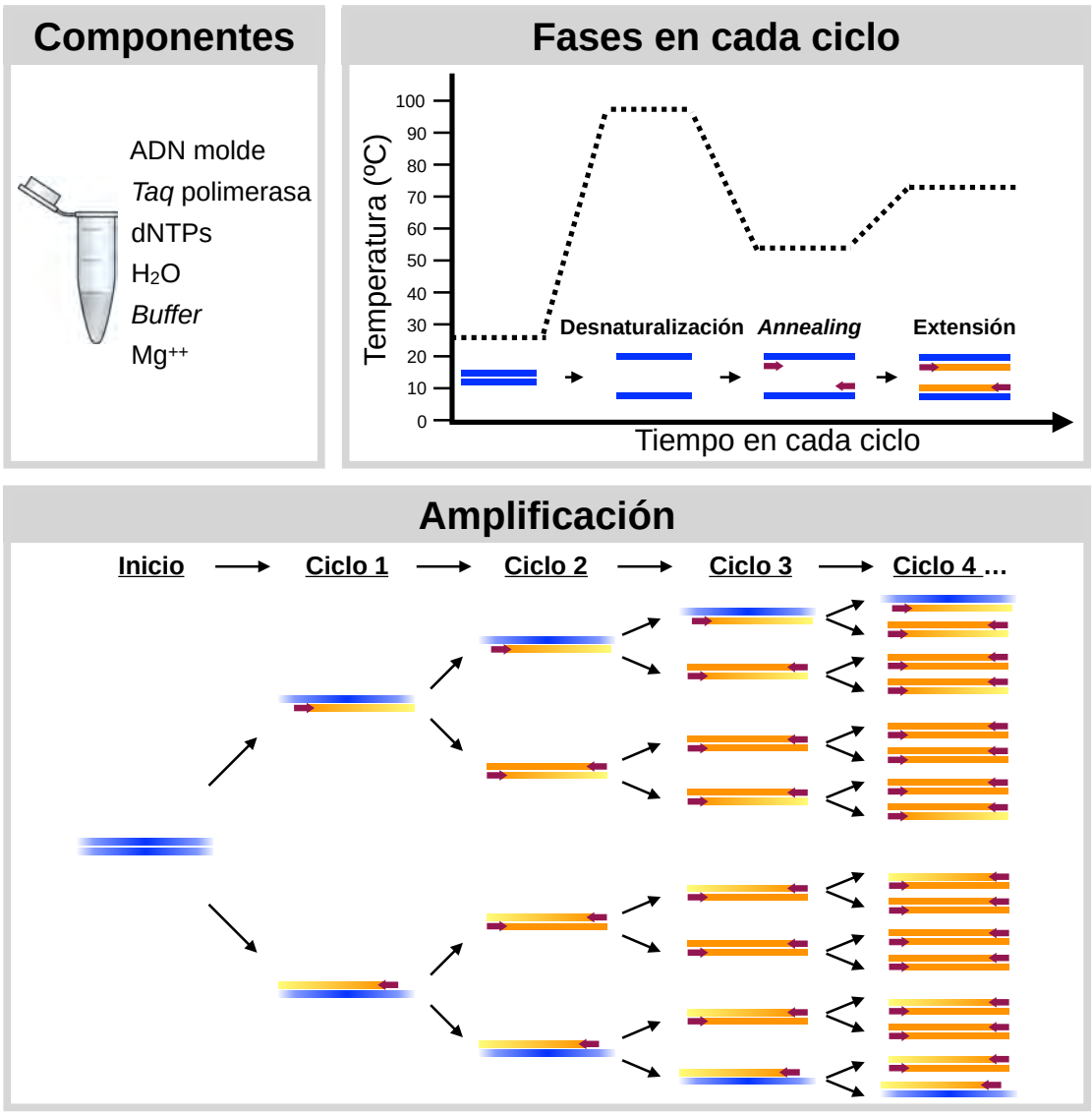


Fig. 2. Esquema de los componentes, las fases de cada ciclo y el proceso de amplificación de la PCR.

1.2.3.1.2 Detección mediante electroforesis capilar

La electroforesis capilar –CE: *capillary electrophoresis*– se lleva a cabo en tubos muy finos de sílice fundida –capilares– que realizan la función análoga a los geles de agarosa o poliacrilamida. Los extremos de los capilares se sumergen en viales que contienen soluciones electrolíticas –*buffers*– y electrodos que se conectan a una fuente de alto voltaje (hasta

~30 kV). Entre ambos extremos de los capilares se establece una diferencia de potencial que provoca la separación de los fragmentos de ADN en función de su relación masa/carga mientras migran a través de la matriz –un polímero viscoso que rellena los capilares– hacia el polo positivo. De esta manera, los fragmentos más cortos alcanzarán antes el extremo positivo del capilar que los más largos. El número de capilares del secuenciador determina el número de muestras que se pueden inyectar simultáneamente –*run*–; además, el polímero, la longitud del capilar y las condiciones de electroforesis pueden ser modificadas para adecuarse a la resolución necesaria en cada caso para la correcta separación de los fragmentos (Butler *et al.* 2004).

En el extremo del capilar más próximo al polo positivo se sitúa un láser que excita los marcadores fluorescentes ligados a los fragmentos de ADN para que emitan fluorescencia. La fluorescencia emitida es captada mediante detectores CCD –*charge-coupled device*–, que permiten registrar múltiples longitudes de onda. Así, fragmentos de ADN que coincidan en tamaño pueden ser analizados simultáneamente mediante el uso de marcadores fluorescentes que emitan en diferentes longitudes de onda (Butler *et al.* 2004).

Además, uno de los fluorocromos se puede reservar para marcar un conjunto de fragmentos de tamaño conocido –*size standard*– que, al ser analizados conjuntamente con la muestra, permiten determinar el tamaño de los fragmentos de la misma.

Finalmente, el secuenciador capilar está acoplado a un ordenador que, mediante un *software* controlador, permite ajustar los parámetros de electroforesis para cada análisis y las longitudes de onda que debe registrar el detector CCD según el conjunto de marcadores fluorescentes utilizados. Además, permite programar las órdenes de inyección de las muestras y recoger los datos en crudo generados (Butler *et al.* 2004). Los datos crudos se pueden procesar mediante el *software* GeneMapper® para generar una representación de los resultados de cada muestra: un electroferograma. En el electroferograma se determinan los tamaños de los fragmentos –referenciándolos al *size standard*– y se representa la intensidad de la señal en unidades relativas de fluorescencia –RFUs: *relative fluorescence units*– para cada uno de los fluorocromos analizados.

1.2.3.2 Metodología de análisis de polimorfismos de longitud

Los polimorfismos de longitud –STRs e Indels– se amplifican mediante una PCR en la que los *primers* son diseñados para hibridar específicamente en las secuencias flanqueantes, de manera que los amplicones recogen las diferencias de tamaño de los alelos. Uno de los *primers* es marcado mediante un fluorocromo en su extremo 5' de manera que, al separarse los amplicones en función de su tamaño en un secuenciador capilar –añadiendo un *size standard* a cada muestra– lo hacen también en función de la longitud del alelo.

Las condiciones de amplificación de cada *locus* se pueden homogeneizar para realizar una PCR de tipo *multiplex* y analizar varios polimorfismos simultáneamente –incluso combinando STRs e Indels– marcando los *primers* con diferentes fluorocromos compatibles. Además, el diseño del electroferograma debe evitar que amplicones que puedan llegar a

solaparse en tamaño se marquen con un mismo fluorocromo. Un esquema de la metodología se muestra en la Fig. 3.

Los kits comerciales incluyen un *ladder* alélico –una colección de fragmentos marcados que contiene la mayoría de los alelos posibles para cada sistema– que se analiza como una muestra más y permite al *software* de análisis identificar los alelos de las muestras de estudio.

En un contexto forense, una de las cualidades más deseables de las metodologías de análisis de marcadores genéticos es una alta capacidad de *multiplexing*, ya que favorece obtener más información empleando una menor cantidad de muestra. Esta técnica permite simultanear el análisis de más de 15 STRs o 40 Indels.

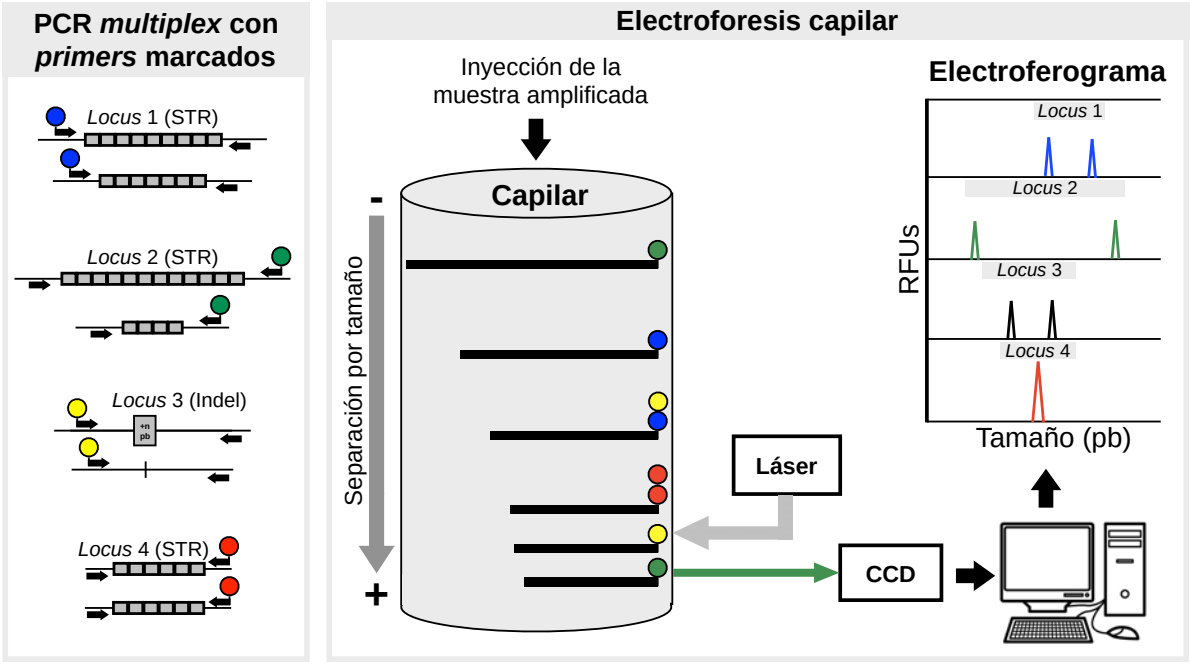


Fig. 3. Análisis de polimorfismos de longitud mediante PCR *multiplex* con *primers* marcados y CE.

Una de las limitaciones de la técnica es la alta experiencia que se requiere para la interpretación de los perfiles –especialmente en casos de mezclas de ADN, ADN *low template* (en baja cantidad) o ADN degradado– ya que el electroferograma se puede ver afectado por una serie de artefactos inherentes a la metodología de detección y a la biología del proceso, (Butler 2005, Butler 2015).

Entre los artefactos relacionados con la biología de los marcadores y la amplificación mediante PCR podemos encontrar: *split-peaks* –la polimerasa añade una adenina al final de las cadenas dobles de ADN, si la adenilación es incompleta aparece un pico de un pb menos–; *drop-outs* –se pierden alelos estocásticamente, ocurre generalmente cuando el ADN molde se encuentra en baja calidad o cantidad–; *drop-ins* –aparecen alelos causados por

amplificaciones inespecíficas–; desbalance de heterocigotos –se produce durante la PCR debido a amplificaciones preferenciales o a efectos estocásticos en ADN *low template*–; alelos nulos –alelos que no son amplificados debido a mutaciones en las secuencias flanqueantes que impiden el *annealing* de los *primers*–; duplicaciones o deleciones cromosómicas que afectan al *locus*… Específicamente en los STRs, debido a las características biológicas de los mismos, aparecen picos *stutter* que se generan mediante *slippage* durante la PCR y que típicamente presentan una unidad de repetición menos que el alelo del que provienen, representando hasta un 15% de su altura.

Entre los artefactos relacionados con la detección podemos encontrar: *pull-ups* –el sistema de detección no filtra completamente la señal en las áreas en las que los espectros de los fluorocromos se solapan–; *split-peaks* –la muestra está sobrecargada y sobrepasa el rango de detección de fluorescencia del equipo–; *dye-blobs* –picos más anchos producidos por fluorocromos escindidos de los *primers*–; burbujas de aire, picos de sales, picos producidos por cambios de voltaje…

1.2.3.3 Metodologías de análisis de polimorfismos de secuencia

En los polimorfismos de secuencia –SNPs– los diferentes alelos tienen la misma longitud, por lo que el uso de *primers* marcados no permitiría diferenciarlos mediante CE. Para distinguir los alelos se pueden seguir dos estrategias fundamentales: (i) determinar el orden de las bases en una región de ADN en la que el polimorfismo quede enmarcado o secuenciar –sección 1.2.3.3.1–; o (ii) interrogar única y específicamente la base del polimorfismo mediante metodologías específicas para SNPs –sección 1.2.3.3.2–.

Los STRs e Indels, pese a ser considerados como polimorfismos de longitud debido a su metodología de análisis más extendida, constituyen a su vez polimorfismos de secuencia y pueden ser analizados como tales mediante secuenciación. En los STRs, debido a la complejidad de la estructura de las repeticiones y a la alta tasa de mutación de estos marcadores –ver sección 1.2.2.1–, la secuenciación cobra una especial importancia ya que permite diferenciar isoalelos –alelos que presentan la misma longitud pero diferente secuencia–. Además, la secuenciación de STRs permite la detección de variantes en las secuencias flanqueantes de los mismos que, conjuntamente con la diferenciación de isoalelos, elevan el poder de discriminación de los marcadores.

1.2.3.3.1 Secuenciación

Secuenciar es determinar el orden preciso de los nucleótidos de una cadena de ADN. Existen dos principales metodologías clásicas de secuenciación: la de Maxam-Gilbert que consiste en inducir modificaciones químicas en el ADN y posteriormente generar roturas específicas en las bases modificadas (Maxam y Gilbert 1977) y la de Sanger que conlleva el uso de terminadores de cadena y la acción de la polimerasa (Sanger *et al.* 1977)–.

Posteriormente, la secuenciación de Sanger adquirió una especial relevancia, ya que fue adaptada para llevar a cabo la detección mediante secuenciadores capilares (Smith *et al.*

1986), elevando su automatización y rendimiento. Gracias a esta adaptación, fue el método elegido para la primera secuenciación del genoma humano (Lander *et al.* 2001, Venter *et al.* 2001). En el ámbito forense, se convirtió en la metodología de secuenciación de rutina, debido a la compatibilidad con el equipamiento básico de los laboratorios.

La metodología de secuenciación de Sanger adaptada al uso de secuenciadores capilares (Sanger *et al.* 1977, Smith *et al.* 1986) se basa en el uso de nucleótidos terminadores o ddNTPs marcados con diferentes fluorocromos para cada una de las posibles bases. Tras una primera amplificación del fragmento de ADN a secuenciar mediante PCR, se produce una segunda reacción de secuenciación en la que el *primer* de secuenciación se extiende con una combinación de ddNTPs marcados y dNTPs no marcados. Cada vez que se incorpora un ddNTP, se finaliza la extensión del *primer* y se genera un fragmento cuya longitud corresponde a la posición en la que se ha añadido el ddNTP marcado. Los fragmentos generados se separan en función de su tamaño en el secuenciador capilar y se recoge la señal fluorescente de los ddNTPs marcados, de manera que la secuencia se puede inferir a partir del electroferograma generado. Un esquema del proceso se muestra en la Fig. 4.

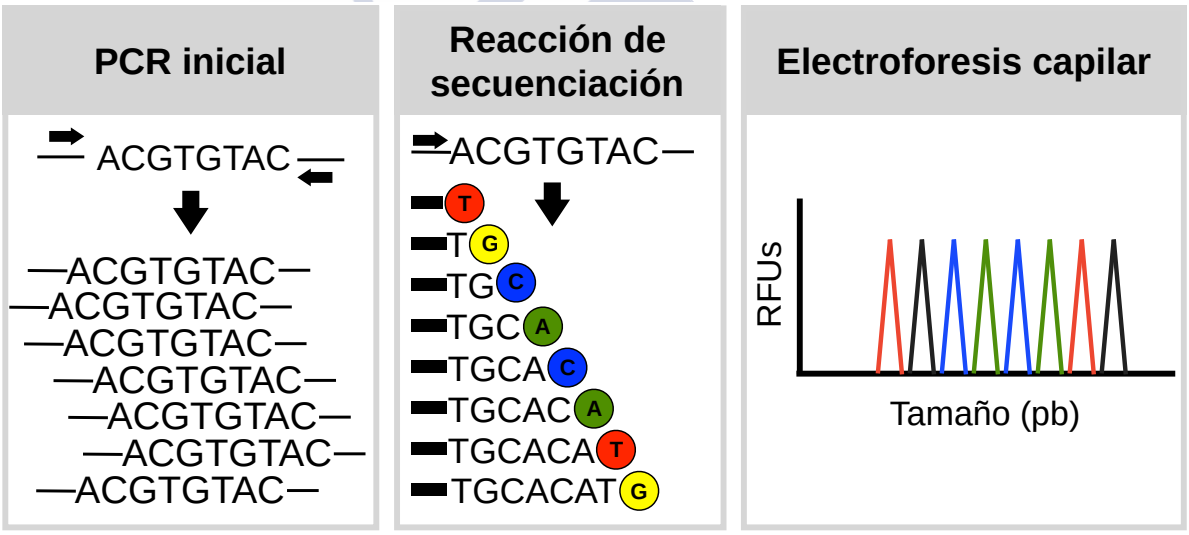


Fig. 4. Secuenciación mediante el método de Sanger con ddNTPs marcados y CE.

A partir del año 2000 se desarrollan nuevas metodologías de secuenciación: las denominadas tecnologías de secuenciación de nueva generación –NGS: *next generation sequencing*– o, más específicamente, de secuenciación masiva en paralelo –MPS: *massively parallel sequencing*–. Existen diferentes aproximaciones de MPS, que comparten la característica de que la secuenciación se lleva a cabo simultáneamente sobre moléculas de ADN clonalmente amplificadas, cuyas señales se detectan individualmente debido a la separación de los procesos en el espacio (Voelkerding *et al.* 2009).

El proceso general de preparación del ADN para secuenciación mediante MPS incluye varios pasos. El primer paso es obtener los fragmentos que se quieren secuenciar: en una secuenciación *de novo* se utiliza el método de *shotgun* –se rompe una alta cantidad de ADN en fragmentos cortos mediante sonicación– mientras que si se pretenden secuenciar posiciones conocidas del genoma se aplican métodos de captura. Los métodos de captura producen un enriquecimiento en aquellas secuencias de interés mediante diferentes metodologías: por un lado, la PCR de captura amplifica aquellas regiones delimitadas por los *primers*, mientras que la captura mediante sondas de hibridación separa los fragmentos de interés de la solución total de ADN. En un segundo paso, los fragmentos a secuenciar se flanquean de adaptadores universales, constituyendo librerías. En un tercer paso, cada uno de los fragmentos de la librería es separado espacialmente y amplificado clonalmente para permitir la secuenciación simultánea (Børsting y Morling 2015).

Existen diferentes metodologías de secuenciación MPS: las tecnologías pioneras son las de (i) pirosecuenciación; (ii) secuenciación mediante ligado; y (iii) uso de terminadores reversibles (Moorthie *et al.* 2011), y posteriormente se incorporó la (iv) secuenciación mediante semiconductores.

i. Pirosecuenciación: se basa en la detección de las incorporaciones de nucleótidos a la cadena de síntesis. Los diferentes nucleótidos se añaden al medio de forma secuencial y cuando se producen incorporaciones en la cadena de síntesis se libera pirofosfato –PPi– en cantidades proporcionales al número de nucleótidos incorporados. El pirofosfato se detecta como luminiscencia a través de una cascada enzimática. El método ya se comercializaba a mediados de los años 90 como reacciones individuales (Ronaghi *et al.* 1996) y fue adaptado posteriormente para MPS (Margulies *et al.* 2005), permitiendo secuenciar un genoma humano por 1,5 millones de dólares americanos en 5 meses (Wheeler *et al.* 2008) frente a los 2700 millones y 13 años que requirió el Proyecto Genoma Humano mediante secuenciación de Sanger.

ii. Secuenciación mediante ligado: se basa en el ligado de sondas marcadas con fluorocromos a un *primer* iniciador. Las sondas complementarias al ADN molde se ligan al *primer* de la cadena creciente y se recoge la señal de fluorescencia de la sonda. Mediante varios ciclos de ligado y escisiones, y tras completar varias cadenas de síntesis iniciadas a diferente altura de la cadena molde, se generan patrones de señales fluorescentes que se traducen en secuencias (Shendure *et al.* 2005, Drmanac *et al.* 2010).

iii. Método de terminadores reversibles –*sequencing-by-synthesis*–: se modifican los nucleótidos marcándolos mediante un fluorocromo diferente para cada base y ligándoles terminadores reversibles que provocan la interrupción de la síntesis de la cadena mediante la polimerasa. Después de ciclo de extensión, se determina la base añadida en función de la fluorescencia emitida y se escinde el terminador, permitiendo que se incorpore un nuevo nucleótido (Bentley *et al.* 2008).

iv. Secuenciación mediante semiconductores: es análoga a la pirosecuenciación, con la diferencia de que en vez de detectar el pirofosfato, se detecta la liberación del ión H^+ que se

produce al incorporar un nucleótido a la cadena de síntesis mediante semiconductores. En caso de que se añadan varios nucleótidos, la cantidad de iones H^+ liberados será proporcional al número de nucleótidos añadidos (Rothberg *et al.* 2011).

Las metodologías de secuenciación de MPS se implementan en diferentes plataformas que se adaptan a las necesidades de los usuarios: las más convenientes en genética forense son las diseñadas para obtener secuencias cortas de varios cientos de pequeños fragmentos –los llamados secuenciadores de sobremesa–. Estos secuenciadores permiten el análisis simultáneo de cientos de los polimorfismos, pudiendo combinar diferentes tipos de marcadores en el mismo análisis.

La comunidad forense está evaluando principalmente las características de dos plataformas: Ion PGM™ de Thermo Fisher Scientific (TFS) –secuenciación mediante semiconductores– y MiSeq de Illumina –secuenciación mediante terminadores reversibles–, que tienen como ventaja la rapidez de los análisis y la flexibilidad del diseño experimental (Børsting y Morling 2015). Las ventajas e inconvenientes de estas dos plataformas no están completamente establecidas, ya que existen pocos estudios comparativos. Además, las casas comerciales realizan continuas actualizaciones de las plataformas y de los *software* de análisis. Por una parte, la secuenciación mediante terminadores reversibles permite una mejor resolución de los tractos homopoliméricos, ya que los nucleótidos se incorporan y detectan de uno en uno; mientras que el uso de semiconductores tiene una mayor tasa de error debido a la pérdida de proporcionalidad entre el número de nucleótidos añadidos y la intensidad de la señal detectada, llegando a producirse alineamientos erróneos en tractos homopoliméricos de sólo dos bases (Loman *et al.* 2012, Ratan *et al.* 2013). Por otra parte, la secuenciación mediante terminadores reversibles presenta errores de lectura que han sido relacionados con la aparición de ciertas características, como repeticiones invertidas o motivos GGC, en la secuencia adyacente (Nakamura *et al.* 2011). Cuando los niveles de lecturas son altos, ambas plataformas presentan unas tasas de genotipado correcto de SNPs muy parecidas (Quail *et al.* 2012). Además, los diferentes *software* de análisis de las secuencias obtenidas y los diferentes parámetros aplicados pueden afectar a los genotipos obtenidos.

A continuación se detalla la metodología de análisis de SNPs mediante la plataforma Ion PGM™, utilizada en varios de los proyectos que integran este trabajo. La plataforma Ion Torrent™ Personal Genome Machine® (Ion PGM™) se basa en la capacidad de los materiales semiconductores para detectar cambios de pH en el medio y convertirlos en cambios de voltaje (Rothberg *et al.* 2011). La actividad ADN polimerasa de elongación 5'→3' conlleva la liberación de un hidrogenión al medio cuando se incorpora un nucleótido a la cadena de síntesis ($[dNMP]_n + dNTP \rightarrow [dNMP]_{n+1} + PPi + H^+$). La cantidad de nucleótidos añadidos es proporcional al cambio de pH que detecta el semiconductor. La adición al medio de los dNTPs se realiza en un orden predeterminado, permitiendo identificar qué nucleótido se ha incorporado.

El flujo de trabajo implica cinco pasos principales: (i) extracción de ADN; (ii) preparación de librerías; (iii) preparación del molde de secuenciación; (iii) secuenciación; y

(v) análisis de los datos generados. Un esquema de los pasos, a partir de la extracción de ADN, se incluye en la Fig. 5.

i. Extracción de ADN

La extracción de ADN puede ser realizada mediante cualquier método de los habituales en genética forense, de manera que se pueden aplicar los protocolos adaptados para cada tipo de evidencia. Esto permite comparar la eficacia de la plataforma frente al método usado rutinariamente en el laboratorio y, además, analizar la misma extracción mediante ambos procedimientos, evitando en lo posible el agotamiento de la muestra. La plataforma permite, por lo tanto, el análisis de muestras de referencia y de evidencias más complejas como tejidos embebidos en parafina o restos óseos con ADN degradado.

ii. Preparación de librerías

En primer lugar, es necesario generar fragmentos que contengan las secuencias que se desea analizar. Generalmente, se realiza una PCR de captura que, a partir del ADN extraído, amplifica simultáneamente cientos de fragmentos. Esta PCR es altamente flexible ya que permite utilizar *primers* de paneles previamente diseñados para diferentes aplicaciones forenses (Daniel *et al.* 2015). Además, se han desarrollo otros métodos de captura especialmente diseñados para ADN degradado, tales como el uso de amplicones redundantes y solapados –*tiling path primers*– o la hibridación de las regiones de ADN de interés con oligonucleótidos complementarios marcados con biotina que se separan del resto de ADN mediante *beads* de estreptavidina –*primer extension capture*– (Briggs *et al.* 2009).

Las librerías para la plataforma Ion PGM™ consisten en colecciones de los fragmentos capturados enmarcados en adaptadores universales que permiten una secuenciación en paralelo. Los *primers* utilizados en la PCR de captura, que forman parte de los amplicones generados pero no de la secuencia objeto de estudio, son parcialmente digeridos a fin de ligar en los extremos de los fragmentos dos adaptadores universales, denominados “A” y “P1”. Aquellos fragmentos que liguen un adaptador diferente en cada uno de los extremos serán válidos para la posterior secuenciación. Así, se puede realizar una PCR previa a la preparación del molde de secuenciación en la que, valiéndose de la universalidad de los adaptadores, se enriquezca la librería en los fragmentos que cumplan esta condición.

Los adaptadores “A” pueden contener *barcodes* (una secuencia corta que permite individualizar cada muestra), en cuyo caso son denominados “X”, que permiten simultanear varias muestras en un mismo análisis. Las librerías se cuantifican para ajustar la concentración a la requerida para la preparación del molde de secuenciación. Si la concentración es insuficiente, se puede realizar una PCR de enriquecimiento. Cuando se analizan simultáneamente varias muestras usando adaptadores “X”, las librerías se combinan en un *pool* equimolar para asegurar una representación proporcional de cada muestra.

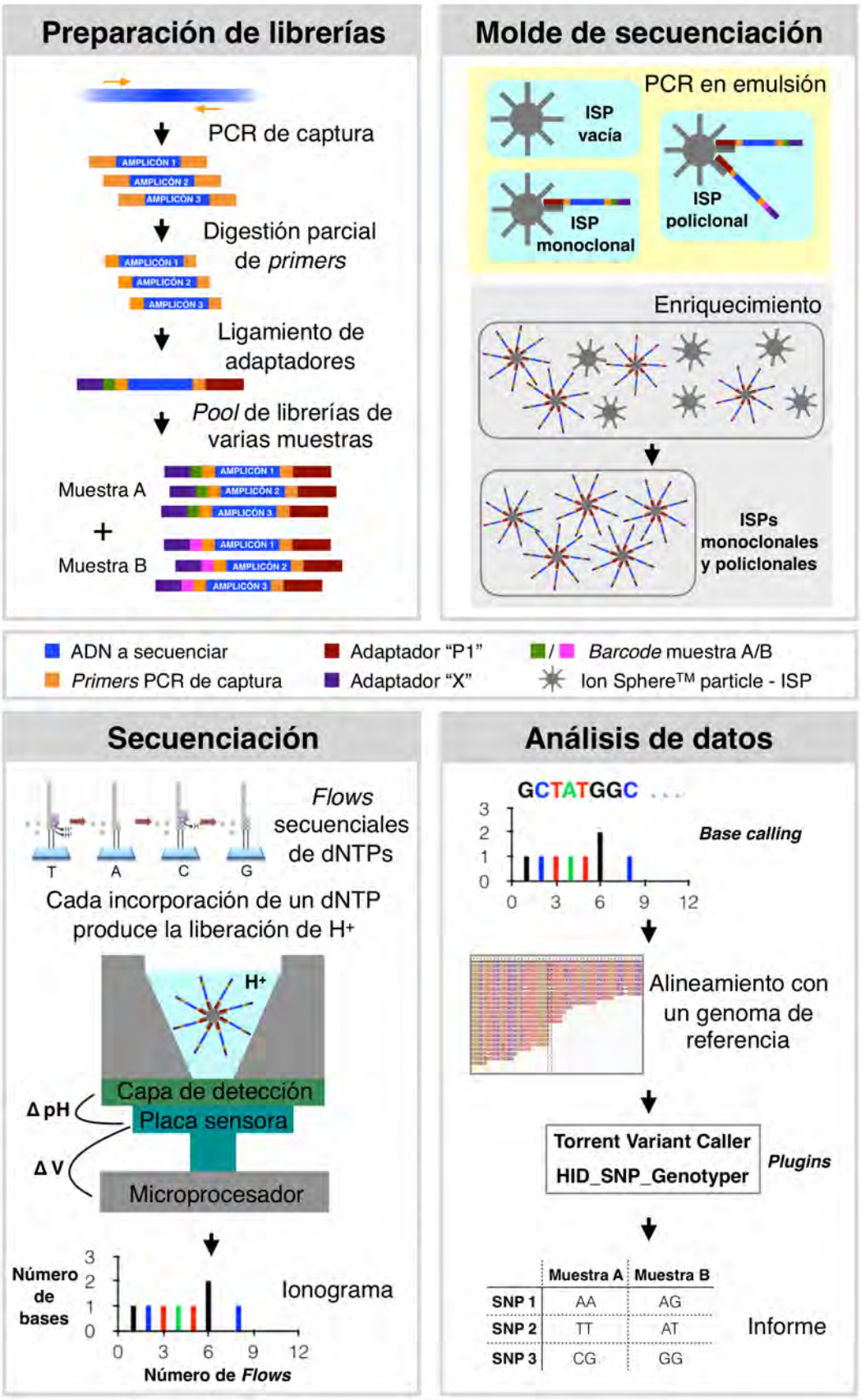


Fig. 5. Esquema del proceso de MPS mediante la plataforma Ion PGM™.

iii. Preparación del molde de secuenciación

Posteriormente, cada uno de los fragmentos de la librería se copia muchas veces sobre un soporte físico, a fin de que la intensidad de la señal generada esté comprendida en el rango del detector. En esta plataforma, el soporte físico consiste en pequeñas esferas llamadas ISPs –Ion Sphere™ *particles*– que, recubiertas de uno de los fragmentos de la librería, forman el molde –*template*– de secuenciación. Para preparar el molde de secuenciación, los adaptadores "P1" de los fragmentos se unen a sus complementarios, que recubren las ISPs, y se realiza una PCR en emulsión. En esta PCR en emulsión se obtienen idealmente moldes monoclonales –ISPs recubiertas de varias copias de un único fragmento de librería–. Un factor crítico en la obtención de ISPs monoclonales es el ajuste de la concentración de la librería o *pool* de librerías, de manera que se maximice la probabilidad de que cada ISP resulte en la misma fase acuosa que un único fragmento de la librería. El proceso posterior de enriquecimiento elimina las ISPs vacías para elevar el rendimiento de secuenciación. No obstante, las ISPs policlonales –con dos o más secuencias diferentes– son secuenciadas y sus resultados se filtran automáticamente durante el análisis de los datos.

iv. Secuenciación

La secuenciación se realiza sobre un chip que contiene millones de pocillos en cada uno de los cuales se encaja una única ISP mediante un proceso de carga. El rendimiento final del análisis está condicionado por la eficacia de este proceso –a mayor número de pocillos vacíos se obtiene un menor número de secuencias–. En el secuenciador Ion PGM™, los pocillos se inundan secuencialmente con los diversos nucleótidos en un orden determinado –*flows*–. En cada uno de los *flows*, los nucleótidos se incorporan a la cadena de síntesis siempre que sean complementarios a la cadena molde. Al incorporarse un nucleótido se libera un ión hidrogenión, siendo el número de iones liberados y, por tanto, el cambio de pH del medio, proporcional al número de nucleótidos añadidos. Cada pocillo tiene en su base una capa de detección sensible a cambios de pH que se conecta a su vez a una placa sensora que los traduce en picos de voltaje. Un microprocesador convierte los picos de voltaje en ionogramas, que constituyen la base del análisis de datos.

Se debe destacar que el flujo de trabajo conlleva muchos pasos manuales, especialmente durante la preparación de la librería. La preparación del molde de secuenciación puede realizarse mediante dos sistemas: Ion OneTouch™ 2 System y Ion Chef™ System. El Ion OneTouch™ 2 System cuenta con dos dispositivos –uno en el que se realiza la PCR en emulsión y otro en el que se lleva a cabo el enriquecimiento en ISPs no vacías– y requiere varios pasos manuales entre ambos dispositivos y la carga manual del chip. El Ion Chef ™ System es un único dispositivo que automatiza todo el proceso desde la preparación del molde hasta la carga del chip. La tendencia hacia la automatización constituye una característica ventajosa para su aplicación en genética forense, ya que permite elevar el grado de estandarización de los procesos. Otra ventaja es la escalabilidad del chip de secuenciación, permitiendo ajustar el número de muestras que se pueden analizar simultáneamente a los requerimientos de cada laboratorio.

v. Análisis de datos

El análisis de datos se realiza mediante el *software* Torrent Suite™. Los ionogramas se traducen en secuencias de bases –*base calling*– que se alinean frente a un genoma de referencia para detectar las posiciones objeto de estudio y generar un informe con los genotipos. Mediante el *software* Torrent Suite™ y sus subprogramas accesorios –*plugins*– se realizan los diferentes análisis requeridos. En el análisis de SNPs, es de especial utilidad la combinación de los *plugins* Torrent Variant Caller y HID_SNP_Genotyper. Ambos *plugins* contienen una serie de parámetros y umbrales de calidad que pueden ser modificados por el usuario en función del grado de rigurosidad deseado o de la finalidad de los análisis. El Torrent Variant Caller identifica todas las variantes de las secuencias obtenidas en relación con el genoma de referencia mientras que HID_SNP_Genotyper simplifica los datos generados por el Torrent Variant Caller y reporta un listado de genotipos de aquellas posiciones de interés para el analista. El *plugin* Genotyper asigna genotipos en base a las probabilidades posteriores calculadas para cada posibilidad, de manera similar a como lo hace el *software* GTAK (McKenna *et al.* 2010). Las probabilidades posteriores se calculan a partir de las probabilidades de cada genotipo, usando los valores de calidad Phred –que se relacionan logarítmicamente con la probabilidad de error del genotipado– y las probabilidades a priori, y teniendo en cuenta los valores de *coverage* y los umbrales para el parámetro minimum_allele_frequency se reportan los valores de calidad del genotipado o QUAL (que pueden variar de entre 0 a varios miles). Los genotipos son asignados si pasan un valor de QUAL específico, además de una serie de parámetros de filtrado que pueden ser modificados por el usuario. Si no se sobrepasan estos umbrales, el genotipo se reporta como NN o N –*no call*–.

La Torrent Suite™ permite descargar lecturas brutas, tanto alineadas como no alineadas, así como los diferentes archivos de salida de los *plugins* anteriormente expuestos que recogen características de las secuencias a partir de las que se derivan los genotipos. Estos archivos se pueden someter a análisis paralelos o posteriores. En este trabajo se han utilizado hojas de cálculo y el lenguaje R para tratar los datos generados por los *plugins*, el visor de secuencias IGV (Robinson *et al.* 2011) para el escrutinio detallado de las secuencias alineadas y el programa CLC Workbench para la obtención directa de genotipos a partir de secuencias no alineadas.

1.2.3.3.2 Métodos específicos de genotipado de SNPs

Existe una amplia variedad de métodos de genotipado de SNPs aplicables en genética forense, que se pueden clasificar en función de la metodología de diferenciación de los alelos –*allele specific hybridization*, *primer extension*, *oligonucleotide ligation* o *invasive cleavage*–, del formato del ensayo y de la metodología de detección –electroforesis, espectrometría de masas, quimioluminiscencia…– (Syvänen 2001, Sobrino *et al.* 2005). Cuando se tienen muestras de ADN de buena calidad, como en la mayoría de procedimientos de pruebas de parentesco, cualquiera de estos métodos se pueden aplicar para obtener genotipos de hasta miles de SNPs, permitiendo el análisis de parentescos muy lejanos –sección 1.3.1.2.4–. No

obstante, debido a las habituales limitaciones en la cantidad y calidad de las muestras, la metodología de análisis de SNPs más extendida en la comunidad forense es la minisecuenciación (Syvänen 1999) mediante el kit SNaPshot® Multiplex de Thermo Fisher Scientific –SNaPshot–.

El método de SNaPshot conlleva ventajas frente al resto de metodologías en cuanto a sensibilidad y, además, todo el equipamiento necesario –termociclador y secuenciador capilar– forma parte de la instrumentación habitual de un laboratorio de genética forense (Budowle 2004).

El procedimiento consta de una primera amplificación por PCR de la región de interés que comprende el SNP –los amplicones tienen habitualmente una longitud <150 pb– y una segunda reacción de discriminación alélica tipo SBE –*single base extension*– en la que las sondas hibridan con los amplicones hasta la base inmediatamente anterior a la del SNP. La base a cuestionar se extiende con el ddNTP (nucleótido terminador) complementario. Los ddNTPs están marcados con un fluorocromo diferente para cada base, permitiendo la identificación de la base que se ha incorporado al analizar el producto en el secuenciador capilar. Se deben purificar los productos enzimáticamente tanto después de la PCR –para eliminar los *primers* y dNTPs remanentes– como después de la reacción de SBE –para eliminar los ddNTPs fluorescentes que no han sido incorporados–. Un esquema de la metodología SNaPshot se muestra en la Fig. 6.

Las condiciones de PCR inicial e hibridación de las sondas se pueden homogeneizar para analizar simultáneamente varios SNPs. Además, a las sondas se les puede añadir colas no-homólogas de ADN –no intervienen en la hibridación– en el extremo 5' que permitan detectar los SNPs en diferentes rangos de tamaño y distribuirlos en el electroferograma de manera que no se solapen. Así, se pueden alcanzar capacidades *multiplex* de hasta ~40 SNPs.

La interpretación de los perfiles de SNaPshot requiere experiencia por parte del usuario: a los posibles artefactos que afectan a los polimorfismos de longitud, se añaden las particularidades propias de la metodología de SNaPshot. Con esta técnica, los diferentes alelos son reportados mediante diferentes fluorocromos que presentan diferentes intensidades de emisión, de tal manera que no se puede esperar una relación directa entre la altura del pico –en RFUs– y la representación inicial de cada alelo en la muestra. En general, la ratio entre la intensidad de los cuatro fluorocromos es 4:2:1:1 –dR110:dR6G:dTAMRA™:dROX™ o azul:verde:amarillo:rojo– (Sánchez *et al.* 2006). Teniendo en cuenta esta ratio, es relativamente sencillo interpretar perfiles de ADN no mezclado, aunque la deconvolución de mezclas constituye un reto importante.

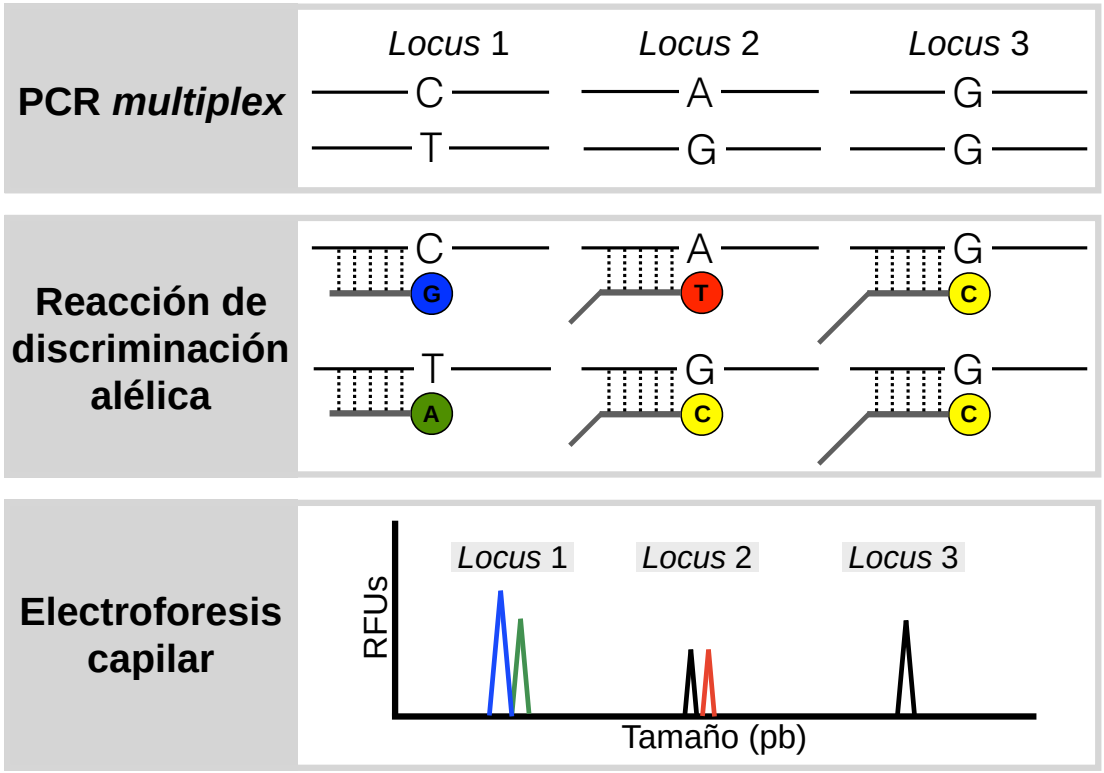


Fig. 6. Análisis de SNPs mediante SNaPshot.

1.3 APLICACIONES DE LA GENÉTICA FORENSE HUMANA

Los diferentes tipos de polimorfismos expuestos en la sección 1.2 pueden ser implementados para diferentes aplicaciones forenses. La aplicaciones más comunes de la genética forense son la identificación individual y las pruebas de parentesco –sección 1.3.1–. Actualmente, se investigan y desarrollan otras aplicaciones basadas en polimorfismos de ADN, dentro de lo que se conoce como ADN *intelligence* –sección 1.3.2–: la predicción de características físicas y de ancestralidad biogeográfica. Además, el uso de otros biomarcadores –sección 1.3.3–, permitirá expandir las aplicaciones de la genética forense humana.

1.3.1 **Identificación individual y parentesco**

Las aplicaciones más comunes de la genética forense, identificación individual y pruebas de parentesco, se basan en establecer la razón de probabilidades entre dos hipótesis a través de la comparación de perfiles de ADN.

En identificaciones individuales se comparan dos o más perfiles de muestras dubitadas o evidencias (un resto biológico de procedencia desconocida) y de muestras indubitadas o de referencia (un resto biológico de procedencia conocida). Las identificaciones se aplican en casos de criminalística y para confirmar la identidad de personas desaparecidas, de los restos

hallados en fosas comunes de conflictos bélicos y de las víctimas de desastres naturales o catástrofes masivas, casos en los que se suelen recurrir a muestras de referencia de familiares. En las pruebas de parentesco se comparan dos o más perfiles de muestras indubitadas o de referencia entre las que se cuestionan relaciones familiares.

Ambas aplicaciones precisan, por tanto, polimorfismos de ADN que permitan individualizar, con alta heterocigosidad y baja heterogeneidad entre poblaciones. Los polimorfismos más comunes y adecuados para llevar a cabo estas aplicaciones son los STRs autosómicos –sección 1.3.1.1–, con los que se resuelven la mayoría de los casos. No obstante, las situaciones en las que el análisis de STRs autosómicos no es lo suficientemente informativo constituye uno de los retos actuales de la genética forense –sección 1.3.1.2–.

1.3.1.1 La prevalencia de los STRs autosómicos

Los STRs autosómicos constituyen los marcadores de elección en casos de identificación individual y parentesco, dado su alto grado de polimorfismo y su elevada heterocigosidad. En el mercado existe una gran variedad de kits (Butler 2012a) que analizan entre 9-21 STRs autosómicos independientes mediante amplicones de ~100-500 pb en reacciones de tipo *multiplex* con *primers* marcados –ver sección 1.2.3.2–. En estos kits, los STRs autosómicos suelen acompañarse de uno o varios marcadores que permiten identificar el sexo de la muestra, como pueden ser la amelogenina y/o Indels y STRs de cromosoma Y. El perfil de STRs autosómicos de cada individuo está constituido por el patrón de alelos que presenta para cada uno de los *loci* analizados.

A pesar de la gran cantidad de STRs autosómicos polimórficos disponibles en el genoma, tan sólo unos pocos se escogen para los análisis de genética forense –los llamados STRs comunes o *core*–. Los STRs *core* tienen núcleos de entre 4 y 5 pb, son *loci* neutros que toleran un mayor grado de variabilidad y no tienen influencia sobre el fenotipo. Además, el uso de STRs *core* permite la comparación de perfiles generados en diferentes laboratorios, la contrapericia y la existencia de bases de datos de ADN en las que se recogen perfiles de interés.

La primera base de datos de perfiles de ADN fue creada en Reino Unido en 1995 (Werrett 1997) y, desde entonces, muchos estados han desarrollado bases de datos atendiendo a sus propias legislaciones. El éxito e impacto de las bases de datos de perfiles de ADN sobre las investigaciones criminales está probado y, a medida que se recogen más perfiles, el número de *cold hits* –coincidencias de perfiles entre la base de datos y una muestra biológica evidencia en un crimen– aumenta. Además, las bases de datos permiten compartir información entre diferentes estados o fuerzas de seguridad del estado, de gran importancia para establecer conexiones entre casos sin resolver (Butler 2012c). No obstante, todavía existen numerosas cuestiones éticas y legales en torno al uso de bases de datos de ADN que deben ser solventadas y reguladas (Guillen *et al.* 2000).

Uno de los requerimientos básicos para el funcionamiento de las bases de datos de perfiles de ADN es la estandarización de los *loci*. En Europa se encuentra extendido el

European Standard Set –ESS–, que consta originalmente de 7 STRs[3] y se expandió posteriormente a 12 *loci*[4] ya que, al compartirse la información a nivel europeo, la probabilidad de coincidencia de perfiles era muy alta. No obstante, los análisis se suelen complementar con otros marcadores como D16S539, D2S1338, D19S433 o SE33. En Estados Unidos se utiliza el *Combined DNA Index System* –CODIS–, que consta de 13 STRs (Hoyle 1998) y se expande hasta 20 a partir de Enero de 2017 (Hares 2015). Los *loci* incluidos en el ESS y el CODIS y sus correspondientes expansiones se recogen en la Tabla 3. Entre ambos sistemas se comparten 12 STRs, lo que favorece las colaboraciones internacionales.

Tabla 3. Marcadores incluidos en el *European Standard Set* (ESS) y el *Combined DNA Index System* (CODIS) y sus correspondientes expansiones.

		ESS (2001)	Expansión ESS (2009)	CODIS (1998)	Expansión CODIS (2017)
STRs autosómicos	D3S1358	✔	✔	✔	✔
	FGA	✔	✔	✔	✔
	D8S1179	✔	✔	✔	✔
	TH01	✔	✔	✔	✔
	VWA	✔	✔	✔	✔
	D18S51	✔	✔	✔	✔
	D21S11	✔	✔	✔	✔
	TPOX			✔	✔
	D5S818			✔	✔
	CSF1PO			✔	✔
	D7S820			✔	✔
	D13S317			✔	✔
	D16S539			✔	✔
	D1S1656		✔		✔
	D2S441		✔		✔
	D10S1248		✔		✔
	D12S391		✔		✔
	D22S1045		✔		✔
	D2S1338				✔
	D19S433				✔
		Número total de marcadores			
		7	12	13	20

[3] Council Resolution of 25 June 2001 on the exchange of DNA analysis results. Official Journal of the European Union 2001/C 187/01.
[4] Council Resolution of 30 November 2009 on the exchange of DNA analysis results. Official Journal of the European Union 2009/C 296/01.

Actualmente, los STRs se están implementando en las tecnologías de MPS (Van Neste *et al.* 2012, van der Gaag *et al.* 2016). Las ventajas que conlleva la MPS frente a la CE incluyen, por una parte, la detección de variantes nucleotídicas tanto en regiones amplificadas que no forman parte del STR como en la secuencia del propio STR –diferenciación de isoalelos– y, por otra parte, la posibilidad de analizar simultáneamente un mayor número de STRs e incluso combinarlos con SNPs. Consecuentemente, el uso de MPS podría elevar el poder de discriminación de los STRs y favorecer la deconvolución de mezclas. No obstante, existen ciertas limitaciones que deben ser solventadas: la sensibilidad de los análisis de MPS es menor y la metodología es más complicada que la CE; los amplicones que se pueden secuenciar mediante MPS son relativamente cortos (<400 bp) por lo que algunos de los STRs más largos ya establecidos no podrían ser implementados; y el alineamiento de las secuencias obtenidas puede resultar complicado. Además, ciertas características son comunes al análisis de STRs mediante CE y MPS, como los *stutter* o el desbalance entre alelos en casos de ADN *low template* (Phillips *et al.* 2014c).

La implementación de las tecnologías de MPS para el análisis de STRs requiere necesariamente un cambio en la nomenclatura de los STRs. La nueva nomenclatura tendría que recoger las variaciones de secuencia de los STRs y de sus secuencias flanqueantes respecto a una secuencia de referencia constante. Además, sería necesario renovar las bases de datos de frecuencias para aprovechar el aumento del poder de discriminación que supone el análisis mediante MPS. Asimismo, se deben aportar soluciones que permitan compatibilizar los resultados de MPS con los de CE, tanto los que se sigan generando como los que ya se recogen en las bases de datos de perfiles de ADN de cada estado (Parson *et al.* 2016).

1.3.1.2 Los retos de la genética forense

La mayoría de los casos que llegan a un laboratorio de genética forense se resuelven mediante el uso de STRs autosómicos, no obstante los STRs presentan ciertas limitaciones. Existen casos en los que se obtienen perfiles nulos o parciales debido a que el ADN se encuentra degradado –sección 1.3.1.2.1– o en muy baja cantidad –sección 1.3.1.2.2–; o perfiles completos cuya interpretación presenta retos como en las mezclas de ADN –sección 1.3.1.2.3– o que no aportan el poder de discriminación exigido por el análisis como en la resolución de pedigríes complejos –sección 1.3.1.2.4–. En estos casos, se pueden complementar los análisis de STRs autosómicos con otros marcadores, o incluso pueden verse desplazados como marcadores de elección para no causar el agotamiento de la muestra.

1.3.1.2.1 ADN degradado

Cuando un organismo comienza a descomponerse, la ausencia de oxígeno y nutrientes desencadena cascadas de señalización que provocan la muerte celular mediante apoptosis o necrosis. Ambas vías provocan la ruptura de las cadenas de ADN mediante la acción de endonucleasas endógenas: en el caso de apoptosis se generan fragmentos de ~200 pb (Arends

et al. 1990) y en el caso de necrosis se producen fragmentos de tamaño variable (Didenko *et al.* 2003, Mizuta *et al.* 2013).

Además de los provocados mediante la apoptosis o necrosis, el ADN sufre daños exógenos, tanto enzimáticos como espontáneos. Estos procesos dependen de las condiciones ambientales: temperatura, humedad, pH, propiedades geoquímicas del suelo, presencia de ácidos fúlvicos y húmicos, radiación… que determinan la tasa de degradación en función del tiempo (Burger *et al.* 1999, Kaiser *et al.* 2008, Alaeddini *et al.* 2010). Por una parte, la ruptura de las membranas celulares de los organismos en descomposición libera al medio fluidos ricos en nutrientes que promueven el crecimiento de microorganismos. Se estima que un ~70% de los microorganismos expresan nucleasas (Antheunisse 1972) que degradan el ADN. Por otra parte, y pese a la alta estabilidad ambiental del ADN frente a otras moléculas orgánicas, se generan daños espontáneos mediante diferentes procesos: hidrólisis –provoca la aparición de sitios abásicos, afectando principalmente a las bases púricas, y deaminaciones de bases nitrogenadas, especialmente de las citosinas–, oxidación –provoca modificaciones de las bases– o, de menor importancia en restos enterrados, radiación ultravioleta –provoca dímeros de pirimidinas– (Lindahl 1993, Höss *et al.* 1996, Goodsell 2001, Gates 2009).

El ADN se conserva en mejores condiciones en restos óseos y dientes (Götherström *et al.* 2002), ya que en tejidos calcificados se encuentra embebido en una matriz de hidroxiapatito y colágeno que lo protege de la degradación (Collins *et al.* 2002). Precisamente, los restos esqueléticos y los dientes constituyen el vestigio biológico de elección para la extracción en casos de ADN degradado. No obstante, tanto en restos óseos y dientes como en tejidos conservados en parafina –donde el ADN sufre degradación durante los procesos de fijación con formaldehído (Greer *et al.* 1991)– los daños en el ADN provocan fallos en la amplificación mediante PCR. La capacidad de *annealing* de los *primers* o la de extensión de la polimerasa pueden verse afectadas por las modificaciones de las bases, los dímeros de timina y otros daños. No obstante, el principal factor que afecta a la amplificación de ADN degradado es la ausencia de una cadena molde íntegra. Así, diversos estudios de STRs autosómicos establecen una correlación negativa, más acusada cuanto más degradado se encuentre el ADN, entre el tamaño del *loci* a amplificar y el éxito de la amplificación del mismo (Whitaker *et al.* 1995, Takahashi *et al.* 1997, Schneider *et al.* 2004).

Teniendo en cuenta el éxito de los *loci* más cortos en los análisis de STRs autosómicos convencionales, una estrategia para el análisis de ADN degradado consiste en disminuir el tamaño de los amplicones: surgen los llamados Mini-STRs (Wiegand y Kleiber 2001). Muchos de los amplicones diseñados para STRs convencionales abarcan regiones más amplias que el propio STR, de manera que son susceptibles de ser analizados con *primers* que hibriden en regiones más próximas al polimorfismo. Los tamaños de los amplicones se reducen de ~100-500 pb en los STRs convencionales a ~150 pb en los Mini-STRs, aunque con diferentes capacidades para cada marcador, debido al propio tamaño del STR y a las limitaciones de las regiones adyacentes.

El uso de Mini-STRs aumenta la probabilidad de éxito de amplificación de los polimorfismos STRs en muestras de ADN degradado (Grubwieser *et al.* 2006, Opel *et al.* 2006, Parsons *et al.* 2007, Welch *et al.* 2011) manteniendo las ventajas de los STRs: alto grado de polimorfismo, compatibilidad con las bases de datos de ADN y metodología de análisis sencilla; por lo que han sido convenientemente integrados en los análisis forenses (Gill *et al.* 2006). La principal desventaja de los Mini-STRs es que se pueden detectar simultáneamente un menor número de marcadores ya que se amplifican en un rango de tamaño ajustado. Otra estrategia consiste en el desarrollo de kits complementarios de manera que los STRs con amplicones cortos en uno de los kits tendrán amplicones largos en el otro kit, y viceversa: se busca la obtención de perfiles completos mediante la combinación de los STRs amplificados con éxito en cada kit.

Las metodologías de análisis de ADN requieren una cadena molde íntegra de al menos el tamaño del polimorfismo más las secuencias adyacentes en las que hibridan los *primers* de la PCR inicial. El tamaño del propio polimorfismo es, por lo tanto, el principal factor limitante en la reducción del tamaño de los amplicones. En este sentido, se han desarrollado paneles de SNPs y pequeños Indels –polimorfismos con un tamaño mucho menor que los STRs– con aplicaciones de identificación individual y parentesco.

Las principales ventajas del uso de SNPs en identificación y análisis de parentesco son la baja tasa de mutación –en comparación con los STRs– y la potencialidad de reducir los tamaños de amplicón para su análisis hasta los ~45-55 pb –favoreciendo así el análisis de muestras de ADN degradado– (Kidd *et al.* 2006, Babol-Pokora y Berent 2008, Fondevila *et al.* 2008, Romanini *et al.* 2012). No obstante, la naturaleza generalmente bialélica de estos marcadores constituye su principal desventaja: un bajo poder de discriminación. Se calcula que conjuntos de ~50 SNPs son suficientes para alcanzar un poder de discriminación comparable a ~16 STRs (Krawczak 1999, Gill 2001, Ayres 2005), escogiendo marcadores con alta heterocigosidad –en SNPs bialélicos, lo más cercano posible a frecuencias 0,5:0,5– y frecuencias distribuidas homogéneamente entre las poblaciones.

Existe una gran cantidad de paneles de SNPs diseñados para aplicaciones de identificación y análisis de parentesco (Inagaki *et al.* 2004, Dixon *et al.* 2005, Lee *et al.* 2005, Kidd *et al.* 2006, Pakstis *et al.* 2007, Pakstis *et al.* 2010, Lou *et al.* 2011, Freire-Aradas *et al.* 2012). El más extendido es el panel creado por el consorcio SNP*for*ID (Sánchez *et al.* 2006), constituido por 52 ID-SNPs –ID: *identity*– polimórficos en tres grandes grupos poblacionales que se amplifican en una única PCR *multiplex* –con amplicones entre 59 y 115 pb– y son genotipados mediante dos reacciones de minisecuenciación. El ensayo se validó en un estudio interlaboratorio en el que se obtuvieron buenos resultados en muestras de ADN degradado y puso de manifiesto la necesidad de incrementar la sensibilidad (Musgrave-Brown *et al.* 2007); además, se comprobó su utilidad en análisis de paternidad (Børsting *et al.* 2008). Finalmente, 49 de los 52 marcadores se incorporaron en un ensayo que presentaba una mayor sensibilidad (Børsting *et al.* 2012).

Una vez probada la utilidad de los paneles de SNPs de identificación y parentesco, es deseable aumentar el poder de discriminación de los mismos. En este sentido, existen dos estrategias fundamentales: aumentar el poder de discriminación individual de cada marcador y/o aumentar el número de los mismos. Para aumentar el poder de discriminación de los marcadores, se debe recurrir a aquellos que presentan mayores niveles de heterocigosidad: los SNPs bialélicos tienen un máximo teórico de 0,5 mientras que los trialélicos de 0,67 y los tetraalélicos de 0,75. El predominio de marcadores no bialélicos con frecuencias lo más balanceadas posible elevaría el poder de discriminación de los paneles de SNPs y, a su vez, solventaría otra de las limitaciones más importantes de los marcadores bialélicos: la detección de mezclas de ADN (Westen *et al.* 2009, Phillips *et al.* 2015). Por otra parte, el desarrollo de la MPS ha permitido ampliar el número de marcadores abordables en un único análisis, solucionando las limitaciones en cuanto a la capacidad de *multiplex* o la cantidad de ADN necesario que conllevaba el uso de otros métodos de genotipado de SNPs. La estrategia principal de las casas comerciales para formar paneles para plataformas de MPS consiste en reunir varios conjuntos de marcadores previamente diseñados y ya establecidos. Así, se han implementado paneles de más de 100 marcadores que alcanzan un alto poder de discriminación y están siendo evaluados por la comunidad forense (Seo *et al.* 2013, Børsting *et al.* 2014, Churchill *et al.* 2016, Grandell *et al.* 2016).

Los Indels comparten las ventajas de los SNPs –mejoran el éxito de amplificación en muestras de ADN degradado y tienen una baja tasa de mutación comparada con los STRs– y, a su vez, conservan la metodología de genotipado sencilla de los STRs –PCR con *primers* marcados y CE–. Por ello, se han desarrollado conjuntos de pequeños Indels con fines de identificación y parentesco, que han demostrado su eficacia como marcadores complementarios en muestras de ADN degradado (Pereira *et al.* 2009, Fondevila *et al.* 2012, LaRue *et al.* 2012, Manta *et al.* 2012, Romanini *et al.* 2012, Bashir y Hassan 2016).

Otra aproximación clásica al análisis de ADN degradado es el uso de ADNmt, principalmente la secuenciación de las regiones hipervariables. La principal ventaja del ADNmt para el análisis de ADN degradado es el alto número de copias que existe en cada célula, que aumenta la probabilidad de que permanezcan cadenas molde íntegras para la PCR inicial. No obstante, se debe tener en cuenta que el ADNmt no permite individualizar y, en caso de tener que recurrir a realizar comparaciones con familiares para las identificaciones, deben serlo por vía materna.

1.3.1.2.2 ADN low template

Los análisis de STRs de ADN *low template*/*low level*/*low copy number* presentan una serie de condiciones definitorias específicas. En primer lugar, la cantidad de ADN disponible para la PCR está muy por debajo del rango óptimo del ensayo, p. ej. cantidades iniciales de ADN de ~100 pg en los kits habituales de STRs. En segundo lugar, se realizan ajustes de los protocolos para aumentar la sensibilidad: habitualmente se incrementa el número de ciclos pero también se puede reducir el volumen de PCR o aumentar el de inyección en CE, realizar

PCRs anidadas o purificar los productos de PCR. En tercer lugar, los perfiles exhiben características específicas provocadas por efectos estocásticos durante la PCR: alto desbalance entre los alelos, *drop-outs* y *drop-ins* y, al incrementase el número de ciclos, mayores niveles de *stutter* (Butler 2012d). Pese a que existen medidas para minimizar los efectos estocásticos –como la generación de perfiles consenso a través de réplicas de PCR– y modelos estadísticos diseñados específicamente para la interpretación de los perfiles generados; la interpretación de los perfiles es muy compleja y presenta cierta controversia (van Oorschot *et al.* 2010).

En el análisis de ADN *low template*, el uso de SNPs e Indels de identificación no comporta ventajas frente al uso de STRs, ya que en el análisis de estos marcadores se pueden dar los mismos efectos estocásticos durante la PCR.

No obstante, una aproximación común para el análisis de ADN *low template* es el uso de ADNmt. A pesar de que no permite individualizar, en vestigios biológicos en los que la cantidad de ADN nuclear es baja, el alto número de copias por célula del ADNmt conlleva una probabilidad mucho mayor de obtener resultados. Además, debido a su localización citoplasmática, el ADNmt permite analizar aquellas células que han perdido el núcleo –p. ej. pelos sin bulbo– (Butler 2012d).

1.3.1.2.3 Mezclas de ADN

En muchos casos de criminalística la evidencia presenta una mezcla de vestigios biológicos de dos o más individuos. Cuando uno de los componentes de la mezcla está representado por células espermáticas, p. ej. en algunos casos de agresiones sexuales, se puede realizar una extracción de ADN diferencial (Gill *et al.* 1985). Mediante la extracción diferencial se producen dos fracciones: una enriquecida en la fracción espermática –los espermatozoides tienen un recubrimiento más resistente y diferente al resto de células– y otra enriquecida en el resto de tipos celulares –habitualmente células epiteliales–. La extracción diferencial permite una interpretación más sencilla de los resultados de los análisis de ADN de las fracciones.

La interpretación de perfiles de STRs de mezclas de ADN requiere una alta experiencia y, actualmente, no existe una aproximación estándar (Budowle *et al.* 2009). En primer lugar, se debe detectar la existencia de una mezcla de ADN en el perfil. Evidencias de que el perfil no procede de un único individuo son la presencia de más de dos alelos en varios *loci* (se han reportado casos en los que muestras procedentes de un único individuo presentan tres alelos para un sistema, pero la probabilidad de que esto ocurra simultáneamente en varios STRs es muy baja) y/o el hecho de que la ratio de altura de los picos de los alelos de un mismo marcador no se corresponda con el rango de valores que presenta dicho marcador habitualmente en heterocigosis.

En segundo lugar, se debe definir qué picos representan alelos reales presentes en la mezcla de ADN. Cuanto más extrema sea la proporción de los contribuyentes de la mezcla, p. ej. ratios de mezclas 1:9, mayor será la probabilidad de que el componente minoritario se

vea afectado por los efectos estocásticos típicos del ADN *low-template*: desbalance entre los alelos de un mismo *locus*, *drop-outs* y *drop-ins* (Butler 2012d); por ello, se generan perfiles consenso a partir de repeticiones desde PCR. Aunque se deben tener en cuenta otros posibles artefactos relacionados con la detección como *pull-ups* o *split-peaks* –ver sección 1.2.3.2–, en el caso de los perfiles de STRs la presencia de picos *stutter* constituye la limitación más importante a la hora de identificar los alelos reales.

Típicamente, los picos *stutter* tienen una repetición menos y pueden llegar a representar hasta el 15% de la altura de los picos reales. No obstante, también se produce una proporción de amplicones con una repetición adicional que representan alrededor del 1-3% en STRs tetraméricos (Gibb *et al.* 2009) y hasta el ~7% en STRs triméricos (Westen *et al.* 2012).

Los niveles de *stutter* son altamente variables entre los diferentes alelos de cada *loci*. Existe una relación directa entre la proporción de *stutter* de cada alelo y el número de repeticiones en STRs simples (Walsh *et al.* 1996, Lazaruk *et al.* 2001). No obstante, en los STRs complejos un mismo alelo puede tener diferentes secuencias internas, y los niveles de *stutter* se correlacionan, en este caso, con la longitud de la porción de repeticiones simples ininterrumpidas más larga del STR (Brookes *et al.* 2012). Esto explica, p. ej., que el alelo 9.3 de TH01, que posee una repetición incompleta que separa las repeticiones simples en dos porciones ininterrumpidas de 5 y 4 repeticiones de longitud –$[TCAT]_5CAT[TCAT]_4$–, tenga tasas de *stutter* más cercanas a las del alelo 5 que al 9 o el 10.

A su vez, los niveles de *stutter* son variables entre los *loci*. Generalmente, la tasa de *stutter* de cada *loci* es inversamente proporcional a la longitud de la unidad de repetición. Por ello, se han desarrollado STRs pentanucleótidos que presentan niveles de *stutter* de entre el 10-20% de los niveles encontrados en STRs tetranucleótidos, de manera que los *stutters* de los alelos pentaméricos representarían entre un 1-3% de la altura de los picos reales (Bacher y Schumm 1998). Algunos STRs pentaméricos se han incorporado en kits comerciales –siendo los más comunes Penta D y Penta E– y otros paneles de STRs (Phillips *et al.* 2013c).

Identificar los picos *stutter* como tales puede ser complicado, especialmente cuando las mezclas incluyen múltiples donantes. Un pico con una repetición menos que uno de los alelos del componente mayoritario puede ser tanto un *stutter* como un alelo real del componente minoritario, o la suma de las señales de ambos. El uso de STRs pentaméricos permitiría descontar de manera más sencilla los *stutter*, facilitando la interpretación de perfiles de mezclas de ADN.

En tercer lugar, se debe estimar el número mínimo de contribuyentes en la mezcla. Aunque el efecto sea menor en sistemas multialélicos, la probabilidad de que existan alelos enmascarados se eleva a medida que aumenta el número de contribuyentes a la mezcla. Por ello, teniendo en cuenta el número de alelos asignados en cada marcador sólo se puede determinar el número mínimo de contribuyentes a la mezcla. Mediante los marcadores incluidos en los kits que permiten identificar el sexo (principalmente amelogenina, pero también Indels y STRs de cromosoma Y) se puede identificar si la mezcla está compuesta únicamente por mujeres o presenta algún componente masculino (Budowle *et al.* 2009). En

este último caso, se puede recurrir al análisis de ADN de cromosoma Y para la detección del componente masculino, teniendo en cuenta la limitación de que no permite individualizar.

En cuarto lugar, la interpretación estadística de las mezclas y la deconvolución de mezclas son campos actualmente en constante debate y desarrollo que requieren colaboraciones internacionales de los expertos de la comunidad forense. En principio, en mezclas en las que existe una diferencia de proporción entre dos contribuyentes, se puede llegar a deconvolucionar los perfiles del componente mayoritario y minoritario, prestando atención a la posibilidad de que los alelos del componente minoritario se encuentren enmascarados (Budowle *et al.* 2009).

Cuando no se obtienen perfiles de STRs de las mezclas, debido principalmente a que el ADN se encuentra degradado, se puede recurrir al análisis de SNPs e Indels de identificación. La mayoría de los SNPs e Indels son bialélicos y, debido a esta condición, presentan importantes limitaciones en la identificación de mezclas de ADN. Una aproximación para detectar mezclas en marcadores bialélicos es el aumento de la heterocigosidad del perfil –se observa un mayor número de alelos, debido a que se combinan los alelos de los componentes individuales–.

Los Indels, debido a su metodología de análisis –ver sección 1.2.3.2–, permiten una interpretación más sencilla de los perfiles de mezclas de ADN, dado que la altura de cada alelo –en RFUs– se correlaciona con la representación inicial del mismo en la muestra original.

En el caso de los SNPs, las limitaciones se ven magnificadas por las características del método de genotipado más usado en forense, la minisecuenciación mediante SNaPshot (Butler *et al.* 2007). En el método de SNaPshot –ver sección 1.2.3.3.2– los diferentes alelos son reportados mediante diferentes fluorocromos, de manera que no existe una relación directa entre la altura del pico y la representación de cada alelo en la muestra original. En los métodos de MPS, en los que se aborda el análisis de un número mucho mayor de SNPs, la detección de las mezclas es más sencilla, ya que existe una buena correlación entre el número de lecturas de cada alelo del SNP y su grado de representación en la mezcla de ADN. En estos casos, podrían llegar a aplicarse métodos de deconvolución de mezclas en principio diseñados para marcadores multialélicos (Gill *et al.* 2015). No obstante, el análisis de SNPs multialélicos –trialélicos e incluso tetraalélicos (Phillips *et al.* 2015)– aumentaría la probabilidad de detección de mezclas de ADN (Phillips *et al.* 2004) en SNaPshot, manteniendo las características beneficiosas de los SNPs en cuanto a identificación y parentesco –principalmente el alto éxito de amplificación en análisis de ADN degradado– (Westen *et al.* 2009).

1.3.1.2.4 Pedigríes complejos

Actualmente, se pueden analizar simultáneamente mediante CE hasta 21 STRs autosómicos, que comportan un poder de discriminación suficiente para aportar altos niveles de probabilidad al contraste de hipótesis en la mayoría de pruebas de parentesco. No obstante,

en ciertos contextos es necesario alcanzar un nivel de discriminación más alto, como pueden ser situaciones en las que se hace la prueba con un familiar en primer grado del real (von Wurmb-Schwark *et al.* 2006), en las que aparecen incompatibilidades mendelianas que pueden ser debidas a la alta tasa de mutación de los STRs o en las que se cuestionan relaciones muy lejanas. Una de las posibles soluciones es añadir más STRs autosómicos. En este sentido, el análisis de STRs mediante MPS, además de aumentar el número de marcadores, permite aumentar el poder de discriminación de cada marcador individualmente al distinguir isoalelos.

Para aumentar el poder de discriminación se pueden añadir otros tipos de marcadores autosómicos –polimorfismos bialélicos– o, siempre que se adecúen a las necesidades del análisis, marcadores no autosómicos –cromosoma X, ADNmt y cromosoma Y–.

El uso del cromosoma X –especialmente de STRs, pero también SNPs (Tomas *et al.* 2010) e Indels (Freitas *et al.* 2010)– puede ser de utilidad en la resolución de casos como pruebas de maternidad, pedigríes nieta-abuela paterna, casos de medias hermanas por parte de padre o pruebas en las que se debe distinguir entre dos posibles padres que a su vez son familiares (Szibor *et al.* 2003, Trindade-Filho *et al.* 2013). Por otra parte, el uso de marcadores uniparentales –ADNmt y cromosoma Y– puede aportar datos en los análisis de pedigríes por las vías materna o paterna, respectivamente.

Para aumentar el poder de discriminación mediante polimorfismos bialélicos –SNPs e Indels– autosómicos, se puede hacer uso de los mismos paneles diseñados para el análisis de muestras de ADN degradado. La baja tasa de mutación de estos marcadores les confiere una alta identidad por descendencia y una baja probabilidad de mutaciones recurrentes. Por ello, son de especial utilidad como marcadores complementarios en la resolución de pedigríes complejos y casos de incompatibilidades mendelianas (Amorim y Pereira 2005, Phillips *et al.* 2008b, Tillmar y Mostad 2014) y pueden llegan a desplazar a los STRs como marcadores de elección en casos de parentescos muy lejanos (Phillips *et al.* 2012b). En ciertos casos, el análisis es inabordable con STRs, pero se puede recurrir a tecnologías que permitan el análisis de cientos de miles de SNPs siempre que se disponga de muestras que cumplan los altos requerimientos de estas plataformas (Lareu *et al.* 2012). Además, los análisis de secuenciación del genoma completo de gemelos monocigóticos pueden identificar mutaciones germinales de SNPs en el esperma del gemelo que es el padre –y no en el otro gemelo– que son transmitidas a la descendencia en más del 80% de los casos (Krawczak *et al.* 2012, Weber-Lehmann *et al.* 2014).

1.3.2 **ADN *Intelligence***

Cuando el perfil de ADN obtenido de una muestra biológica presentada como evidencia en un caso no concuerda con ningún individuo conocido –un sospechoso o un perfil recogido en las bases de datos de perfiles de ADN–, la información que aporta a la investigación es limitada.

Una aproximación que permite aportar información a las investigaciones es la ADN *intelligence*: la inferencia de características del donante de un vestigio biológico a través del ADN contenido en el propio vestigio.

Dos vertientes relacionadas se engloban bajo el término ADN *intelligence*: por un lado, la inferencia de características externas visibles –EVCs: *externally visible characteristics*– denominada ADN *phenotyping* (sección 1.3.2.2); y, por el otro, la predicción de ancestralidad biogeográfica –BGA: *bio-geographical ancestry*– (sección 1.3.2.1).

1.3.2.1 Predicción de ancestralidad biogeográfica

Las poblaciones humanas presentan cierta estructuración genética –sección 1.3.2.1.1–, que se refleja en los marcadores informativos de ancestralidad –sección 1.3.2.1.2–: marcadores de linaje y marcadores informativos de ancestralidad individual –AIMs: *ancestry informative markers*–. La genética forense se sirve prioritariamente de los AIMs para establecer paneles de predicción de ancestralidad biogeográfica –sección 1.3.2.1.4–, para los que se seleccionan los marcadores más informativos –sección 1.3.2.1.3–.

1.3.2.1.1 La estructura genética de las poblaciones humanas

La historia evolutiva de la especies –incluyendo procesos de mutación y recombinación, selección, deriva genética y migración o flujo genético (Jobling *et al.* 2004c)– se recoge en la distribución de la diversidad genética de las poblaciones. La diversidad genética de la especie humana es baja debido a la breve historia evolutiva de la misma; no obstante, los efectos de la historia evolutiva permanecen reflejados en la estructura genética de las poblaciones actuales permitiendo definir la ancestralidad de las mismas.

El primer estudio que abordó la estructuración de las poblaciones, realizado con un número muy limitado de marcadores autosómicos, estimó las diferencias intrapoblacionales e interindividuales en un ~85%; las intergrupales en un 10% y las interpoblacionales e intragrupales en un 5% (Lewontin 1995).

En un estudio posterior se analizaron 377 STRs en muestras del panel HGDP-CEPH (Cann *et al.* 2002) y se estimaron las diferencias intrapoblacionales e interindividuales en un ~93%; las intergrupales en un 4% y las interpoblacionales e intragrupales en un 2,4% (Rosenberg *et al.* 2002). Los análisis mediante STRUCTURE (Pritchard *et al.* 2000) –un programa que genera agrupaciones atendiendo a las similitudes genéticas (ver sección 1.3.2.1.3)– identificaron conjuntos de poblaciones definidos continentalmente cuando se establecían un total de 5 grupos (Eurasia, África subsahariana, este de Asia, América y Oceanía) y 7 grupos (África subsahariana –AFR–, Europa –EUR–, Oriente Medio –ME–, sur de Asia –SAS–, este de Asia –EAS–, América –AMR– y Oceanía –OCE–).

Las mismas muestras analizadas con 650000 SNPs (Li *et al.* 2008) evidenciaron que las diferencias entre los subgrupos de la región de Eurasia eran más débiles, de manera que los 7 grupos generados por el programa STRUCTURE no se correspondían perfectamente con los

subgrupos de Eurasia (Europa, Oriente Medio y centro/sur de Asia). Los resultados del análisis STRUCTURE del trabajo de Li *et al.* (2008) se muestran en la Fig. 7.

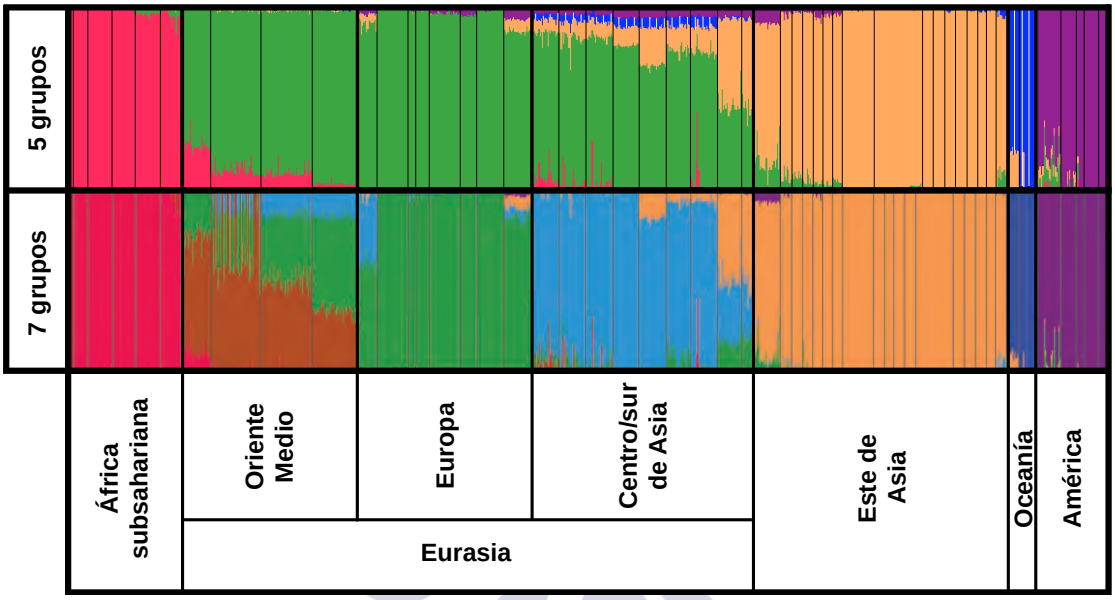


Fig. 7. Análisis STRUCTURE de 650000 SNPs en muestras del panel HGDP-CEPH. Se muestran los resultados para 5 y 7 grupos y su correspondencia con grupos definidos continentalmente. Adaptada de Li *et al.* (2008).

La estructura de la diversidad genética de la especie humana puede ser explicada mediante diversos factores (Phillips 2015). En primer lugar, la población de los diversos continentes durante la expansión *out-of-Africa* conlleva sucesivos eventos de cuello de botella y efecto fundador en los que el tamaño efectivo de la población disminuye y la deriva genética modela las frecuencias génicas (Pickrell y Reich 2014). Así, la mayor diversidad se da en África y disminuye hasta los últimos continentes poblados –América y Oceanía– (Wang *et al.* 2007, Friedlaender *et al.* 2008).

En segundo lugar, las poblaciones no se aparean al azar a nivel mundial. Existen barreras geográficas y geofísicas entre las poblaciones que impiden el flujo de información genética y, a largo plazo, provocan diferencias interpoblacionales (Bamshad *et al.* 2004). Este mismo efecto produce clinas de variabilidad entre poblaciones.

En tercer lugar, estudios de ADN antiguo han identificado posibles hibridaciones entre humanos anatómicamente modernos y formas humanas arcaicas. En estos casos se produciría una introgresión arcaica –un flujo génico desde las formas arcaicas a subconjuntos de poblaciones ancestrales de humanos anatómicamente modernos–. La introgresión arcaica provocaría un aumento en el grado de diferenciación inicial de las poblaciones ancestrales de humanos anatómicamente modernos (Green *et al.* 2010, Reich *et al.* 2010, Gibbons 2011, Sankararaman *et al.* 2012, Pickrell y Reich 2014, Hsieh *et al.* 2016). Eventos recientes de migraciones masivas y posterior hibridación con las poblaciones nativas como la expansión

Austronesia, la expansión Bantú, la expansión del imperio mongólico, el colonialismo europeo, la trata de esclavos…(Pickrell y Reich 2014) se verían reflejados en el grado de *admixture*[5] –generación de una nueva población híbrida a partir de dos o más poblaciones ancestrales previamente aisladas– de las poblaciones actuales. Así, podemos encontrar tanto poblaciones *unadmixed* –en las que los individuos presentan una única ancestralidad biogeográfica– como *admixed* –en las que los individuos presentan diferentes grados de *admixture* entre dos o más ancestralidades biogeográficas–. Al realizar inferencias de ancestralidad biogeográfica sobre individuos, se debe tener en cuenta que un perfil clasificado como *admixed* puede reflejar que el individuo procede de una población *admixed* o que sus ancestros proceden de dos o más poblaciones *unadmixed* diferentes.

En cuarto lugar, los efectos de selección natural varían de acuerdo con los diferentes ambientes en los que se asientan las poblaciones (Coop *et al.* 2009). Así, las presiones selectivas pueden fijar rápidamente mutaciones beneficiosas surgidas al azar, aunque estos procesos son muy poco frecuentes (Hernandez *et al.* 2011). Entre los factores biogeográficos que han provocado estos procesos se encuentran el clima –despigmentación en poblaciones no africanas (Norton *et al.* 2007)–, la presencia de ciertas enfermedades –resistencia a la malaria (Hamblin y Di Rienzo 2000, Tishkoff *et al.* 2001)– o las modificaciones en la dieta derivadas de la práctica de la agricultura –adaptaciones al consumo de leche en adultos (Enattah *et al.* 2008) o de almidón (Perry *et al.* 2007)–.

1.3.2.1.2 Marcadores informativos de ancestralidad biogeográfica

Existen dos tipos de marcadores informativos de ancestralidad biogeográfica: (i) marcadores de linaje y (ii) AIMs.

i. Marcadores de linaje

Los marcadores uniparentales, ADNmt y cromosoma Y –específicamente la región no recombinante del mismo–, se caracterizan por transmitirse como haplotipos o combinaciones de estados alélicos en las líneas materna y paterna, respectivamente. Los diferentes haplotipos se clasifican en haplogrupos atendiendo al estado ancestral o derivado de los polimorfismos. Los haplogrupos constituyen linajes relacionados por descendencia y se pueden representar en árboles filogenéticos tanto de ADNmt (Ingman *et al.* 2000) como de cromosoma Y (Thomson *et al.* 2000). Estos árboles se anclan mediante secuencias de otros primates hominoideos, que representan el estado ancestral de los polimorfismos, y los diferentes clados están determinados por mutaciones filogenéticamente estables (Jobling *et al.* 2004d). Las clasificaciones son jerárquicas y flexibles, permitiendo establecer sublinajes e incorporar nuevos haplotipos.

La idoneidad de los marcadores uniparentales para los estudios filogenéticos poblacionales deriva del hecho de que su tamaño efectivo de población es más bajo que el de

[5] Se mantienen los términos ingleses para evitar confusiones con la palabra mezcla, que se reserva en este trabajo para las mezclas de ADN.

los marcadores autosómicos, de manera que los efectos de la deriva genética son más acusados. Mientras la deriva genética aumenta las diferencias entre las poblaciones, su modelo de herencia implica la ausencia de recombinación a través de las generaciones: los diferentes haplotipos son relativamente estables y la variabilidad se produce únicamente por mutación (Jobling *et al.* 2004a). Como resultado de estos procesos, los haplogrupos de ADNmt y cromosoma Y presentan importantes correlaciones con las diferentes regiones continentales[6], permitiendo trazar la historia de los linajes humanos y sus migraciones a nivel poblacional. A nivel individual, la inferencia del origen biogeográfico se realiza a través de las bases de datos que recogen la frecuencia de los diferentes haplotipos de ADNmt[7] y cromosoma Y[8] a nivel mundial. No obstante, presentan ciertas limitaciones: constituyen un único marcador que no es representativo del genoma total del individuo (Pääbo 2003), se necesitan grandes bases de datos para realizar estimaciones adecuadas de la variabilidad de las poblaciones, pueden llevar a interpretaciones erróneas acerca de la ancestralidad general de un individuo en aquellos casos en los que permanecen linajes distantes atípicos en las poblaciones (King *et al.* 2007) y, aunque se analicen ambos marcadores de linaje, no se recoge la información genética de todos los ancestros del individuo –p. ej., la información genética que proviene del abuelo materno o la abuela paterna no se vería representada–.

ii. AIMs

La ancestralidad biogeográfica de las poblaciones se puede caracterizar mediante el estudio de grandes grupos de marcadores elegidos al azar (Allocco *et al.* 2007, Pardo-Seco *et al.* 2014). No obstante, la elección de los marcadores influye en la capacidad para diferenciar los grupos poblacionales (Enoch *et al.* 2006). Surge por lo tanto una aproximación que consiste en identificar y analizar pequeños conjuntos de marcadores (AIMs) altamente informativos cuyas frecuencias alélicas presenten altas diferencias absolutas entre dos poblaciones ancestrales (Yang *et al.* 2005, Salas *et al.* 2006). Esta aproximación es más adecuada al contexto forense, dado que el número de marcadores que se pueden llegar a analizar se encuentra comúnmente restringido por la cantidad de muestra. Un reto apropiado en este contexto sería la clasificación en 5 grupos continentales –Eurasia, África subsahariana, este de Asia, América y Oceanía (Li *et al.* 2008)– con pequeños conjuntos de AIMs, que representaría un balance entre la cantidad de información que se puede llegar a obtener y el grado de simplificación de la complejidad de la estructura de las poblaciones humanas (Phillips 2015).

Los AIMs son autosómicos, por lo que recogen toda la información acerca de la ancestralidad del individuo evitando los sesgos sexuales que presentan los marcadores de linaje. Además, tienen la ventaja de no requerir un elevado número de muestras para una estimación adecuada de las frecuencias alélicas de la población.

[6] http://www.scs.illinois.edu/~mcdonald/WorldHaplogroupsMaps.pdf
[7] http://empop.online/
[8] https://yhrd.org/

1.3.2.1.3 Selección de AIMs para predicción de ancestralidad biogeográfica

Varios aspectos se deben tener en cuenta a la hora de establecer un ensayo de predicción de ancestralidad biogeográfica con fines forenses: (i) la disponibilidad de datos que nos permitan tanto escoger marcadores como realizar los análisis poblacionales posteriores; (ii) la aplicación de parámetros que nos permitan identificar los mejores AIMs según su grado de informatividad; y (iii) los diferentes sistemas de análisis de datos poblacionales que permiten comprobar la idoneidad de los marcadores escogidos para el ensayo, así como realizar las inferencias de ancestralidad.

i. Bases de datos poblacionales

Actualmente, se puede acceder a través de portales *online* a datos genotípicos recogidos en diferentes bases de datos poblacionales de carácter público. Una de las bases de datos más extensas es la del Proyecto Genoma Humano, que incluye en su Fase III (The Genomes Project Consortium 2015) datos genotípicos de ~81 millones de variantes en 2504 individuos de 26 poblaciones *admixed* y *unadmixed* de África, Europa, este de Asia, sur de Asia y América. Los datos de la Fase III pueden ser descargados para cada variante a través del portal *online* del propio proyecto[9]; mientras que los de la Fase I –que incluye ~28 millones de variantes de 629 individuos de 12 poblaciones (The Genomes Project Consortium 2012)– pueden ser obtenidos simultáneamente para muchas variantes usando el portal SPSmart[10] (Amigo *et al.* 2008). Pese a la importante ampliación en el número de poblaciones entre la Fase I y la Fase III, el Proyecto 1000 Genomas no cubre todas las regiones del mundo.

El panel HGDP-CEPH (Cann *et al.* 2002) ha sido genotipado para más de 650000 SNPs mediante el chip Illumina Human Hap650K (Li *et al.* 2008) y los datos generados son fácilmente accesibles mediante el portal SPSmart. Este panel incluye individuos de poblaciones *unadmixed* de África, Europa, este de Asia, centro y sur de Asia, Oriente Medio, América y Oceanía; de manera que permite cubrir ciertos ámbitos geográficos no incluidos en el Proyecto 1000 Genomas –que no incluye poblaciones *unadmixed* de América ni poblaciones de Oriente Medio u Oceanía–. No obstante, existen ciertas limitaciones en el uso de estos datos para la selección de AIMs. En primer lugar, el número de muestras que se recogen en el panel para algunos de los grandes grupos poblacionales es escaso, de tal manera que los marcadores identificados en base a su grado de divergencia entre grandes grupos poblacionales pueden no ser representativos de toda la variabilidad geográfica, especialmente en los grupos con menor tamaño muestral. En segundo lugar, los SNPs que forman parte del chip Illumina Human Hap650K fueron seleccionados a partir de datos de poblaciones europeas, africanas y americanas por lo que SNPs de otras poblaciones presentes en el panel no se incluyen en la selección. Además, el chip fue diseñado para estudios de asociación, de manera que algunos SNPs próximos a ser fijados en alguna de las poblaciones –y que constituirían buenos AIMs– no fueron incluidos dado su escaso valor en este tipo de estudios

[9] http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice
[10] http://spsmart.cesga.es/

(Phillips 2015). El panel HGDP-CEPH también ha sido genotipado para un gran número de marcadores STRs y los datos poblacionales se pueden obtener en el portal pop.STR[11].

ii. Parámetros que estiman la informatividad de los marcadores

El paso más crucial en el diseño de paneles para la predicción de ancestralidad biogeográfica es la selección de los marcadores informativos para la diferenciación de las ancestralidades biogeográficas que se consideren. La informatividad de cada AIM es más alta cuanto mayores sean las diferencias absolutas entre las frecuencias alélicas de las poblaciones ancestrales (Enoch *et al.* 2006, Salas *et al.* 2006). Así, el parámetro δ mide las diferencias absolutas entre las frecuencias alélicas de dos poblaciones (Shriver *et al.* 1997). Para marcadores bialélicos se define como: $\delta = |p_x - p_y| = |q_x - q_y|$; dónde p_x y p_y representan las frecuencias de uno de los alelos del marcador en las poblaciones X e Y, mientras que q_x y q_y representan las frecuencias del otro alelo. A partir del parámetro δ surgen parámetros más refinados, como el I_n –índice de *informativeness-for-assignment*– (Rosenberg *et al.* 2003) o el índice de Divergencia[12] de Jensen-Shannon (Chen *et al.* 2005). En genética poblacional, predomina el uso del F_{ST} o índice de fijación (Wright 1951), que mide la desviación de la heterocigosidad observada frente a la esperada bajo equilibrio Hardy-Weinberg. Las desviaciones del equilibrio se producen como consecuencia de la acción de las fuerzas evolutivas y, en este caso, se mide el efecto de las subpoblaciones "S" comparado con el total de la población "T", donde influye principalmente la deriva genética. En la práctica, todas estas medidas están relacionadas y, en comparaciones por parejas de poblaciones, tienen valores entre 0, para la mínima divergencia, y 1, para la máxima divergencia (Phillips 2015).

Para realizar estimaciones no sesgadas de proporciones de coancestralidad en individuos *admixed* (Taboada-Echalar *et al.* 2013) es necesario balancear el poder de diferenciación del panel para cada una de las poblaciones. La Divergencia específica de población –PSD: *population-specific Divergence*– para cada *locus*, también denominada LSBL –*locus specific branch lenght*– (Shriver *et al.* 2004), se calcula fácilmente a través del portal Snipper[13] como la Divergencia de un grupo frente al resto de grupos (p. ej., África vs. no África). El acumulativo de los valores de Divergencia específica de población de cada *locus* del panel debe ser equilibrado. Las principales dificultades a la hora de lograr este equilibrio provienen, por una parte, de la distribución desigual de la diversidad humana (p. ej. la divergencia para poblaciones africanas vs. no africanas es mayor que la de poblaciones euroasiáticas vs. no euroasiáticas, de manera que el número de SNPs requeridos para cada comparación puede variar) y por otra, de la falta de SNPs con alelos fijados, que varían los valores acumulados de Divergencia para cada una de las poblaciones (Phillips 2015).

[11] http://spsmart.cesga.es/popstr.php?dataSet=strs_local
[12] La diferenciación entre el fenómeno de divergencia entre poblaciones y el parámetro Divergencia se realiza a lo largo del texto escribiendo la segunda con mayúscula.
[13] http://mathgene.usc.es/snipper/analysispopfile2_new.html

Una vez realizada la selección de marcadores, se deben tener en cuenta otras características como la distancia entre los mismos –para asegurar la independencia de los marcadores– o que las características de las secuencias adyacentes permitan diseñar *primers*. En caso de que algún marcador no cumpla las condiciones, puede ser sustituido por otro que presente una informatividad equivalente.

iii. Sistemas de análisis de datos poblacionales

Existen tres sistemas estadísticos aplicables al análisis de la ancestralidad biogeográfica a partir de datos de variantes: el análisis bayesiano, el análisis de componentes principales y STRUCTURE. Estos sistemas usan datos de poblaciones de referencia y realizan inferencias a partir de los patrones comparativos de variantes detectados (Phillips 2015). Para cada uno de los tres sistemas, se muestran en la Fig. 8 ejemplos de inferencia de ancestralidad biogeográfica para un individuo *unadmixed* y otro *admixed,* usando como referencia tres grandes grupos poblacionales (África subsahariana, Europa y este de Asia) del panel HGDP-CEPH y analizando los 34 AIM-SNPs propuestos por Fondevila *et al.* (2013).

- Análisis bayesiano

El análisis bayesiano se basa en un *training set* –datos genotípicos de un conjunto de muestras de referencia de cada población– para asignar a un nuevo individuo a aquella población en la que se maximice la probabilidad posterior y aportar la razón de verosimilitud –LR: *likelihood ratio*– entre la máxima probabilidad calculada y la siguiente (Phillips 2015). Una forma intuitiva de comprobar la capacidad de predicción de los marcadores incluidos en el análisis es realizar una validación cruzada *–cross-validation–* del *training set*; idóneamente, cada individuo del *traning set* será reclasificado como perteneciente a su población de referencia con altas LR. A la hora de clasificar individuos de ancestralidad desconocida, clasificaciones con altos valores de LR representarían individuos *unadmixed*; mientras que individuos *admixed* presentarían valores de LR bajos, poniendo de manifiesto las escasas diferencias entre las probabilidades calculadas para cada población que intervenga en la coancestralidad. El portal Snipper[14] permite realizar análisis bayesianos en marcadores independientes, tanto asumiendo equilibrio Hardy-Weinberg en las poblaciones como no.

- Análisis de componentes principales

El análisis de componentes principales –PCA: *principal component analysis*– es un análisis multivariante que reduce las dimensiones de los datos manteniendo el mayor grado de información posible. Para ello se calculan un conjunto de variables no correlacionadas, o componentes principales (PC: *principal component*), como combinaciones lineales de las variables originales. Así, la primera PC representa la mayor proporción de varianza y dicha proporción disminuye secuencialmente en la segunda PC, la tercera PC… La combinación de las PCs define el vector propio de cada muestra y habitualmente las tres primeras PCs son suficientes para representar una alta proporción de la variabilidad. Las representaciones de

[14] http://mathgene.usc.es/snipper/analysispopfile_new.html

estas tres variables suelen ser bidimensionales: PC1 vs. PC2, PC1 vs. PC3 y PC2 vs. PC3. En estas representaciones, los individuos de cada población de referencia forman grupos diferenciados mientras que los individuos a clasificar en función de los resultados de las poblaciones de referencia se sitúan o bien cercanos a uno de estos grupos –*unadmixed*– o bien entre dos o más grupos –*admixed*–. El análisis de componentes principales se puede realizar mediante *scripts* del lenguaje R, previa recodificación de los datos mediante el paquete SNPassoc (Gonzalez *et al.* 2007), que permite marcadores bialélicos. Alternativamente, el portal Snipper genera gráficas PC1 vs. PC2 simultáneas al análisis bayesiano, que apoyan la interpretación de las LR obtenidas.

- STRUCTURE

El *software* STRUCTURE (Pritchard *et al.* 2000) se basa en métodos bayesianos para generar agrupaciones atendiendo a las similitudes genéticas. Los individuos son asignados al azar a un número predeterminado de grupos –K–. Después, se estiman las frecuencias de las variantes en cada grupo y se reasigna a los individuos en función de las mismas, y así sucesivamente hasta alcanzar el número de iteraciones determinadas por el usuario. Las primeras iteraciones, denominadas *"burning steps"*, se descartan y se recoge la información de las posteriores, denominadas *"MCMC –Markov chain Monte Carlo– steps"*. Progresivamente, a través de las iteraciones, se llega a estimaciones fiables de las frecuencias de cada grupo y de la proporción en la que cada individuo pertenece a cada grupo. En cada análisis se generan dos matrices que recogen los coeficientes de asignación a los K grupos de cada individuo y cada población. Se debe tener en cuenta que, debido a la metodología, para cada réplica del mismo K se generan matrices diferentes.

Para estimar el K óptimo se realizan múltiples análisis de un rango de Ks y se reúnen los datos generados mediante el portal Structure Harvester (Earl y vonHoldt 2012). Este portal genera gráficas que representan la probabilidad posterior de cada K como media y desviación típica de los múltiples análisis. Para valores de K por debajo del valor óptimo la probabilidad posterior crece, mientras que al llegar a valores próximos o superiores al óptimo tiende a estabilizarse. El punto de inflexión suele representar el valor más apropiado de K, evitando la sobreinterpretación de los datos (Kalinowski 2011), aunque también se debe tener en cuenta la tasa de cambio en las probabilidades entre valores sucesivos de K y la variabilidad entre los múltiples análisis de cada K (Evanno *et al.* 2005). Idóneamente, el K será igual al número de ancestralidades biogeográficas que se pretenden diferenciar y que se recogen en las poblaciones de referencia.

El portal Structure Harvester genera archivos en los que se recogen simultáneamente todas las matrices de coeficientes de asignación a los K grupos de cada individuo y cada población para todos los análisis de un mismo K. A partir de estos archivos, el programa CLUMPP (Jakobsson y Rosenberg 2007) fusiona los múltiples análisis y obtiene matrices únicas para cada K de individuos y poblaciones. A partir de las matrices fusionadas, el programa distruct (Rosenberg 2004) permite generar las clásicas gráficas en las que cada individuo está representado por una barra de K colores. Análogamente, todo el proceso puede

realizarse directamente a partir de los resultados de STRUCTURE mediante el portal CLUMPAK (Kopelman *et al.* 2015).

Idóneamente, aplicando un modelo que permita *admixture*, los individuos de las poblaciones de referencia se presentan en las gráficas como barras de un único y diferente color. Usando un *training set* que cumpla estas condiciones como referencia para el modelo *admixture* POPFLAG –en el que se actualizan las frecuencias alélicas a partir de los individuos marcados como referencia– podemos clasificar a un nuevo individuo. En las gráficas, un individuo *unadmixed* presentará un único color, de manera que su ancestralidad corresponderá con la población de referencia del *training set* que presente el mismo color; un individuo *admixed* presentará dos o más colores cuya representación en la barra atiende a la proporción estimada de coancestralidad de la población de referencia del mismo color.

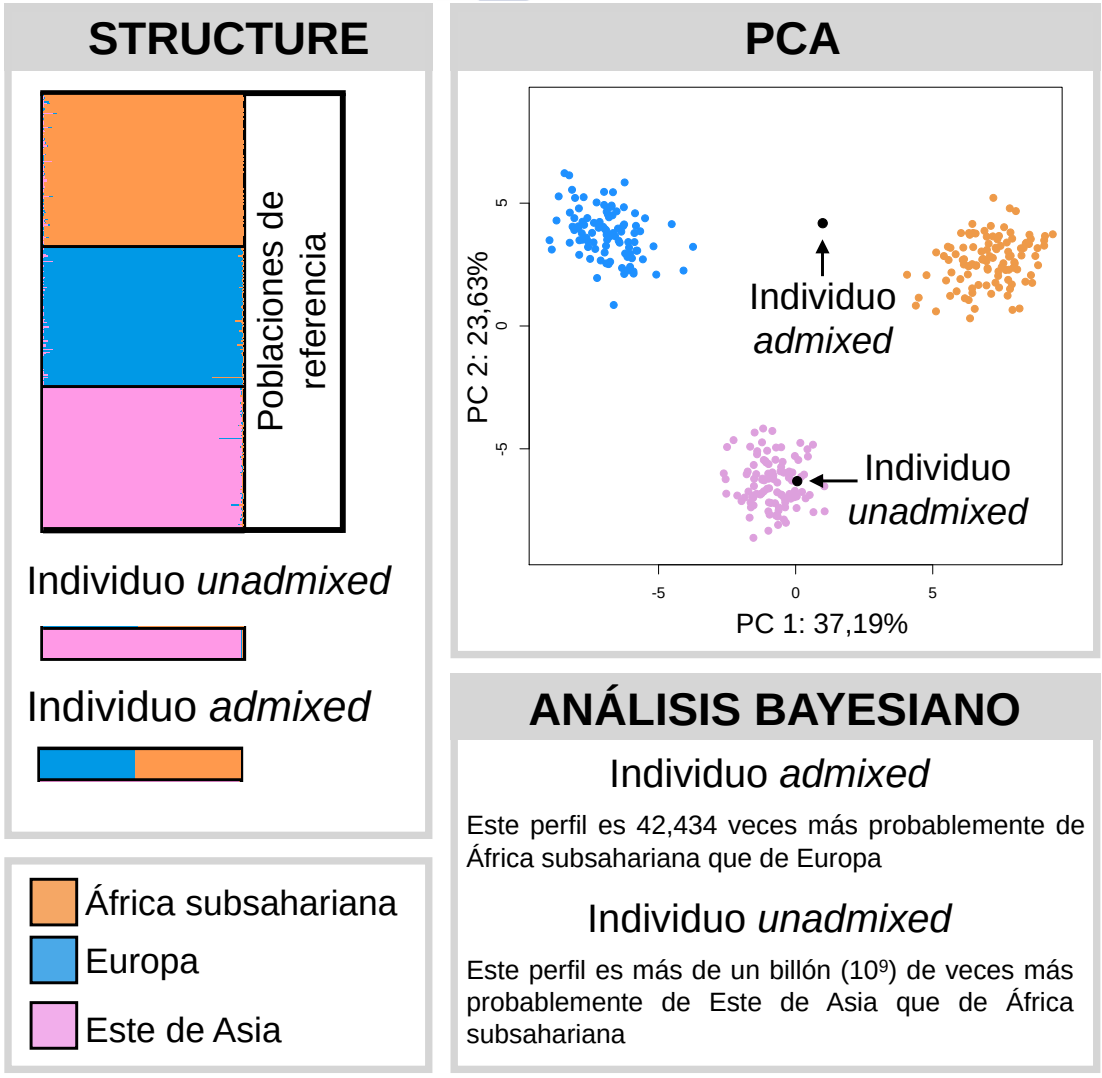


Fig. 8. Ejemplos de clasificaciones mediante análisis bayesiano, PCA y STRUCTURE de dos individuos (*admixed* y *unadmixed*) frente a un *training set* de 3 poblaciones del panel HGDP-CEPH utilizando los 34 AIM-SNPs propuestos por Fondevila *et al.* (2013).

1.3.2.1.4 Paneles de AIMs para predicción de ancestralidad biogeográfica

El uso de los STRs como AIMs está limitado debido a su alta tasa de mutación. Los STRs dinucleótidos están mucho más diferenciados entre las poblaciones que los trinucleótidos o tetranucleótidos (Rosenberg *et al.* 2002, Londin *et al.* 2010). No obstante, los STRs diméricos no forman parte de los STRs *core* comúnmente utilizados en genética forense. Entre los STRs *core* se pueden encontrar alelos específicos de población (Phillips *et al.* 2008a, Phillips *et al.* 2014c) y se puede inferir cierto grado de diferenciación de las poblaciones (Lowe *et al.* 2001, Londin *et al.* 2010, Pereira *et al.* 2011, Phillips *et al.* 2011), que se incrementa al combinarlos con SNPs (Phillips *et al.* 2011) o STRs tetraméricos específicamente diseñados como AIMs (Phillips *et al.* 2013a).

Más adecuados para la predicción de origen biogeográfico son los polimorfismos típicamente binarios –SNPs e Indels–, cuyas bajas tasas de mutación les confieren una alta identidad por descendencia. Entre los paneles de AIM-SNPs, debemos diferenciar aquellos paneles optimizados para ser utilizados en cualquier laboratorio de genética forense –los SNPs se genotipan mediante la tecnología SNaPshot y los ensayos están validados– y paneles teóricos más amplios que reflejan selecciones de marcadores que podrían ser o han sido implementados en otras tecnologías de genotipado de SNPs (como chips de hibridación) o en MPS.

Existen varios paneles diseñados para SNaPshot (Lao *et al.* 2006, Kersbergen *et al.* 2009, Gettings *et al.* 2014, Rogalla *et al.* 2015b), que no logran diferenciar simultáneamente los 5 grupos ancestrales continentales –África subsahariana, Eurasia, este de Asia, América y Oceanía–. El panel 34-plex del consorcio SNP*for*ID (Phillips *et al.* 2007), que posteriormente fue modificado (Fondevila *et al.* 2013), fue diseñado para la diferenciación de Europa, África subsahariana y este de Asia. No obstante, al analizar las muestras del panel HGDP-CEPH se encontraron en los análisis poblacionales patrones que se corresponden con 5 grupos poblacionales –añadiendo América y Oceanía– (Fondevila *et al.* 2013). Este panel se encuentra muy extendido en la comunidad forense (Santos *et al.* 2015) y se han diseñado expansiones que permiten obtener mayores niveles de diferenciación en poblaciones de Eurasia (Phillips *et al.* 2013b), Oceanía (Santos *et al.* 2016) y América.

La mayoría de paneles teóricos se centran en la diferenciación de poblaciones africanas, europeas y nativo americanas (Paschou *et al.* 2007, Kosoy *et al.* 2009, Galanter *et al.* 2012). Actualmente, existen dos paneles diseñados para establecer diferenciaciones de al menos los 5 grupos continentales (Kidd *et al.* 2011, Phillips *et al.* 2014a). El estudio de Kidd *et al.* (2011) comprende el análisis de una gran cantidad de poblaciones y determina que los SNPs de Kosoy *et al.* (2009) permiten diferenciar poblaciones para las que no fueron diseñados. Además, aporta una lista de 55 SNPs complementarios que permiten ajustar el sesgo entre poblaciones y que pueden ser considerados como un panel por sí mismos (Kidd *et al.* 2014). El panel de Phillips *et al.* (2014a), denominado set EUROFORGEN Global AIM-SNP, seleccionó 128 SNPs de diferentes fuentes, así como una lista de posibles sustitutos en caso

de que posteriormente alguno no fuera implementable en MPS, con el objetivo de balancear al máximo la diferenciación de las 5 poblaciones continentales.

Aunque los mejores AIM-Indels no son tan informativos como los mejores AIM-SNPs, la metodología de análisis de Indels presenta la ventaja –frente a SNaPshot– de ser más sencilla y de que los alelos de un mismo marcador son reportados con el mismo fluorocromo, de manera que se simplifica la detección de mezclas de ADN. Así, permiten diferenciar fácilmente un individuo *admixed* de una mezcla de ADN de individuos con dos ancestralidades diferentes (Phillips 2015).

Entre los diferentes ensayos de AIM-Indels (Santos *et al.* 2010, Zaumsegel *et al.* 2013), destaca por su difusión en la comunidad forense (Santos *et al.* 2015) un ensayo que incluye 48 marcadores en una sola reacción (Pereira *et al.* 2012) y permite diferenciar para 4 grupos poblacionales –África, Europa, este de Asia y América–. El poder de diferenciación de los cuatro grupos no está balanceado, de manera que el conjunto de Indels proporciona una mayor divergencia para europeos que para nativo americanos. Así, las estimaciones de proporciones de coancestralidad en poblaciones *admixed* mostraron que el componente nativo americano estaba subestimado frente al europeo cuando se comparaba con un panel más balanceado (Galanter *et al.* 2012) para ambas poblaciones (Taboada-Echalar *et al.* 2013).

1.3.2.2 ADN *phenotyping*

La EVC – *externally visible characteristic*– más simple de predecir es el sexo. Los kits de STRs autosómicos aportan esta información, siendo el marcador más utilizado un Indel que se encuentra en el gen de la amelogenina (Mannucci *et al.* 1994) que tiene una diferencia de 6 pb entre el cromosoma X y el Y. Además, para aumentar la fiabilidad ante una posible deleción del Indel de la amelogenina, se pueden incorporar Indels y STRs de cromosoma Y.

En contraste, la mayoría de los rasgos fenotípicos son caracteres complejos –influyen un número variable de genes con un efecto más o menos fuerte– determinados por la interacción entre el genotipo y el ambiente. Para obtener información sobre el componente genotípico se deben conocer los *loci* funcionales relacionados con cada característica fenotípica. La capacidad para identificar estos *loci* depende principalmente del mayor o menor efecto sobre el fenotipo –penetrancia–, de manera que los rasgos más complejos requieren de amplios estudios de asociación de genomas completos –GWAS: *genome-wide association studies*–, que conllevan importantes inversiones económicas y suelen llevarse a cabo a través de consorcios internacionales (Kayser 2015). Además, dentro de cada *locus* asociado, se deben identificar todos aquellos SNPs que se relacionan con la variabilidad del rasgo.

1.3.2.2.1 Rasgos de pigmentación humana

Los rasgos de pigmentación humana son genéticamente menos complejos que otras EVCs, ya que pueden explicarse mediante un conjunto relativamente pequeño de genes. Por ello, los ensayos de predicción del color de ojos, pelo y piel están más desarrollados. La mayoría de los genes identificados están implicados en la ruta bioquímica de síntesis del

pigmento melanina –melanogénesis– en orgánulos especializados –melanosomas– de células especializadas –melanocitos–. Los melanosomas se transmiten a los queratinocitos en el pelo y la piel, pero permanecen en los melanocitos en los ojos. La melanogénesis puede generar dos tipos de melanina: la eumelanina (formada por pigmentos negros y marrones) y la feomelanina (formada por pigmentos amarillos y rojos). El número de melanocitos presenta una baja variabilidad entre individuos, de manera que el amplio espectro de coloraciones depende del tamaño, cantidad y distribución de los melanosomas; así como de la proporción de eumelanina y feomelanina (Frudakis 2010).

i. Pigmentación de ojos

La pigmentación de ojos (específicamente del iris) es marrón –fenotipo ancestral– en la mayor parte de las poblaciones mundiales; mientras que en poblaciones con ancestralidad europea se presenta una alta variabilidad del rasgo, en un continuo desde el azul al marrón con coloraciones intermedias verde y avellana. El primer sistema de predicción de color de ojos con fines forenses fue denominado IrisPlex (Walsh *et al.* 2011b) e incluye 6 SNPs escogidos entre los más predictivos de un estudio previo (Liu *et al.* 2009). El sistema ha sido validado (Walsh *et al.* 2011a) y expandido en la comunidad forense (Chaitanya *et al.* 2014). Finalmente, el ensayo se ha fusionado con un modelo de predicción de color de pelo, denominado HIrisPlex (Walsh *et al.* 2013, Walsh *et al.* 2014). Los niveles de precisión de las predicciones mediante IrisPlex alcanzan valores altos para los colores marrón y azul, debido principalmente a la extensa base de datos de muestras de referencia. No obstante, la predicción de fenotipos intermedios –verde y avellana– todavía supone un reto. Debido a estas limitaciones se ha propuesto el uso de SNPs adicionales (Mengel-From *et al.* 2010, Ruiz *et al.* 2013, Freire-Aradas *et al.* 2014), que aumentan la precisión de las predicciones en los fenotipos intermedios.

ii. Pigmentación de pelo

El fenotipo ancestral de pigmentación de pelo –negro– se observa en la mayor parte de las poblaciones mundiales; mientras que en poblaciones europeas encontramos una gran variación de fenotipos entre rubio, pelirrojo, castaño y negro. Los primeros estudios se centraron en la predicción del fenotipo pelirrojo, dado el alto grado de correlación de algunas variantes del gen MC1R (Grimes *et al.* 2001, Branicki *et al.* 2007). Uno de los primeros estudios en abordar la clasificación de toda la variabilidad de pigmentaciones constaba de 46 SNPs de 13 genes, con diferente poder de predicción (Branicki *et al.* 2011). Los SNPs más predictivos se incorporaron en un ensayo validado con fines forenses, denominado HIrisplex, que consta de un total de 24 SNPs, de los cuales 6 se incluyen en el modelo de predicción de color de ojos y 22 SNPs se incluyen en el de pelo (Walsh *et al.* 2013, Walsh *et al.* 2014). El modelo de predicción de color de pelo de HIrisPlex es robusto y se apoya en una extensa base de datos de muestras de referencia. No obstante, una de las mayores limitaciones reside en la diferenciación entre castaño y rubio. En este sentido, se espera obtener modelos más exactos gracias al desarrollo de nuevos SNPs asociados al color de pelo (Söchtig *et al.* 2015), a la información que puedan aportar las investigaciones sobre las causas biológicas del cambio de

coloración a lo largo de la vida del individuo –se oscurece de rubio claro en niños a castaño en adultos y se vuelve gris o blanco a partir de cierta edad– y al desarrollo de modelos de predicción de edad.

iii. Pigmentación de piel

La pigmentación de piel presenta una alta heterogeneidad entre las poblaciones mundiales, reflejando la acción de la selección natural en los diferentes ambientes geográficos. El estudio de los *loci* implicados presenta importantes retos dado que los análisis GWAS deben ser aplicados en poblaciones genéticamente homogéneas y el rasgo presenta su mayor variabilidad entre grupos continentales; por ello, es el rasgo de pigmentación menos estudiado. El estudio más completo presentado hasta el momento (Maroñas *et al.* 2014) identifica un conjunto de 10 SNPs, de entre 59 candidatos, que clasifican individuos de diferentes poblaciones en colores de piel negro, intermedio y blanco, con elevadas precisiones de predicción para los fenotipos extremos y menor para los intermedios.

Los estudios de predicción de color de ojos, pelo o piel realizan clasificaciones fenotípicas cualitativamente (p. ej. ojos azules, pelo castaño, piel intermedia). La aplicación de escalas cuantitativas reflejaría más fielmente la alta variabilidad de los fenotipos y evitaría posibles subjetividades en las clasificaciones (Kayser 2015). En este sentido, ya se están desarrollando varios trabajos de fenotipado cuantitativo, sobre todo en pigmentación de ojos (Andersen *et al.* 2013).

1.3.2.2.2 Otras características externas visibles

Además de los rasgos de pigmentación, existe un gran número de características externas visibles con alta heredabilidad cuyo estudio se encuentra en fases iniciales: morfología del cabello, patrones de calvicie, altura, morfología facial… Todavía no existen modelos predictivos para la mayoría de estas características, pero se está recopilando información genética sobre los *loci* asociados, que permitirá en el futuro desarrollar modelos predictivos.

i. Morfología del cabello

La morfología del cabello presenta una alta variabilidad –rizo, ondulado o liso– tanto entre las diferentes poblaciones continentales como dentro de ellas, especialmente en el caso de poblaciones europeas. Existen evidencias, a través de estudios de gemelos monocigóticos, de la alta heredabilidad del fenotipo en las poblaciones europeas (Medland *et al.* 2009b) y, en un GWAS del mismo grupo poblacional, se ha identificado como significativamente asociado el gen de la tricohialina (TCHH) –proteína encargada de conferir resistencia mecánica al folículo piloso– además de otros dos genes (WNT10A Y FRAS1) que presentan asociaciones más débiles (Medland *et al.* 2009a). Posteriormente, se confirmó la asociación de 3 SNPs, uno de cada uno de los anteriores genes, estableciéndose un punto de partida para el desarrollo de ensayos más precisos de predicción de morfología del cabello en poblaciones europeas (Pospiech *et al.* 2015).

ii. Patrones de calvicie

La mayoría de las investigaciones están centradas en la alopecia temprana en hombres, mientras que el fenómeno es mucho mas infrecuente y desconocido en mujeres. En hombres se reportan prevalencias del ~20% entre los 20-30 años aumentando un ~10% en cada década de edad. Existe una importante variabilidad en los patrones de calvicie (Hamilton 1951, Norwood 1975) y evidencias de alta heredabilidad –hasta un 80%– (Nyholt *et al.* 2003).

El *locus* más asociado a la calvicie se encuentra en el cromosoma X y contiene los genes AR –gen del receptor de andrógenos– y EDA2R –gen del receptor de la ectodisplasina A2– (Ellis *et al.* 2001, Prodi *et al.* 2008), lo que explica que la calvicie sea un fenómeno predominantemente masculino y supone que el riesgo de calvicie se hereda principalmente del abuelo materno. Además, se ha encontrado asociación con otros *loci* de cromosomas autosómicos (Li *et al.* 2012, Heilmann *et al.* 2013). El primer modelo de predicción de calvicie en hombres se basa en SNPs de los *loci* anteriormente expuestos, y los autores señalan la necesidad de localizar más *loci* informativos para obtener un modelo que alcance un mejor nivel de predicción (Liu *et al.* 2016).

iii. Morfología facial

Existen varios rasgos faciales considerados como monogenéticos y cuya predicción sería relativamente sencilla: hoyuelos en la barbilla o mejillas, pilosidad de las orejas, separación del lóbulo de la oreja, pico de viuda o presencia de pecas (Pulker *et al.* 2007). No obstante, la morfología facial general, en relación a las medidas craneométricas, es un caracter complejo cuyos *loci* asociados no han sido completamente identificados hasta el momento.

Existen dos estudios principales que buscan *loci* funcionales asociados con medidas craneométricas mediante GWAS (Liu *et al.* 2012, Paternoster *et al.* 2012), aunque tan solo concuerdan en uno de los *loci* identificados. Sin embargo, se ha evidenciado la existencia de alta variabilidad en las regiones adyacentes a los cinco *loci* propuestos en total por estos estudios. Los datos concuerdan con una selección negativa dependiente de las frecuencias, viéndose beneficiados los alelos menos frecuentes: la alta variabilidad de los fenotipos habría sido favorecida a raíz de la necesidad de reconocimiento individual y/o mediante selección sexual (Sheehan y Nachman 2014).

En un estudio basado en una aproximación que mitiga los efectos de la ancestralidad –se analizan individuos con ancestralidad mezclada– y el sexo en la morfología facial, se identificaron 24 SNPs de 20 genes como significativamente asociados a ciertas medidas craneométricas (Claes *et al.* 2014a). No obstante, los propios autores recalcan que el efecto de estos 24 SNPs es mínimo y que tanto la ancestralidad como el sexo determinan la mayor parte de la variación fenotípica (Claes *et al.* 2014b). Además, la aproximación estadística y genética llevada a cabo presenta ciertas controversias (Hallgrimsson *et al.* 2014).

iv. Estatura

La heredabilidad de la altura se estima como >80%. Sin embargo, influyen factores ambientales como la alimentación y exposición a ciertas enfermedades, hecho que se pone de manifiesto en el aumento de la estatura media de hombres y mujeres en aquellas poblaciones en las que existe un importante crecimiento económico durante los últimos 150 años (Cole 2003). Los GWAS sobre estatura incluyen decenas de miles de individuos (Lango-Allen *et al.* 2010, Wood *et al.* 2014). Estos estudios han confirmado la alta complejidad del carácter, de manera que conjuntos extensos de 180 y 697 SNPs asociados explican el 10% y 16%, respectivamente, de la variación del fenotipo. Un ensayo de predicción de estatura basado en el conjunto de 180 SNPs ha logrado diferenciar en europeos fenotipos muy altos frente a los de estatura media con un 75% de precisión en la predicción (Liu *et al.* 2014).

v. Edad

La predicción de la edad tendría enormes beneficios para guiar las investigaciones policiales, tanto en sí misma como por la información que podría aportar a la predicción de otras ECVs que se ven modificados por la edad: color de pelo, patrones de calvicie, morfología facial, estatura…

Ciertos cambios en el ADN son dependientes de la edad, como la acumulación de deleciones en ADNmt y la acortación de los telómeros. No obstante, la capacidad predictiva de estos cambios es limitada y existen problemas metodológicos que restringen su uso en genética forense (Meissner y Ritz-Timme 2010). Actualmente, existen otros biomarcadores para la predicción de edad que constituyen alternativas más adecuadas –ver sección 1.3.3–.

1.3.2.3 Consideraciones éticas y legales respecto a la ADN *intelligence*

Ciertos aspectos éticos y legales deben ser considerados en la aplicación del ADN *phenotyping* y la predicción de ancestralidad. En el aspecto ético, se debe encontrar el equilibrio entre posibles riesgos para la privacidad individual y las ventajas que implicaría en las investigaciones criminales. Los marcadores predictivos de ECVs engloban tanto SNPs codificantes como no codificantes –estos marcadores no codificantes se encontrarían en desequilibrio de ligamiento con marcadores codificantes desconocidos o en regiones intrónicas e intergénicas encargadas de la regulación de genes funcionales implicados en el rasgo– mientras que los de BGA son principalmente no codificantes.

En primer lugar, los marcadores predictivos se analizan en muestras anónimas de la escena del crimen, dado que su utilidad conlleva necesariamente la ausencia de un sospechoso. Además, las EVCs no constituyen parte de la privacidad de los individuos ya que no revelan ninguna información que no pueda revelar un testigo ocular (Kayser y Schneider 2009). En este sentido, aunque los genes implicados en la predicción de EVCs pueden estar implicados en ciertas enfermedades, las variantes comunes que producen la variabilidad del rasgo son generalmente diferentes a las mutaciones que determinan la enfermedad (Kayser 2015). Sin embargo, las predicciones de BGA pueden revelar más información de la

visualmente aparente, ya que la apariencia global del individuo no se correlaciona completamente con su ancestralidad biogeográfica (Kayser 2015).

En segundo lugar, el ADN *phenotyping* y la predicción de ancestralidad aportan información que puede reducir el número de sospechosos, análogamente a como lo haría un testigo ocular del crimen, pero aportando evidencia estadística de las predicciones (Kayser y Schneider 2009). En este sentido, la fiabilidad de las predicciones es mayor que la de la información aportada por los testigos oculares, que puede verse influenciada por valoraciones subjetivas o por las circunstancias del crimen (Spinney 2008).

En tercer lugar, las fuerzas de seguridad precisarían la información derivada acerca de las predicciones y no la información genética. De esta manera, la información genética de los marcadores no necesita ser compartida ni almacenada ya que, en caso de encontrar un sospechoso, sería necesario aplicar los marcadores habituales de identificación (Kayser y Schneider 2009).

En último lugar, las predicciones de EVCs o BGA pueden señalar a una parte de la población como sospechosa, provocando reacciones públicas de estigmatización de un colectivo. En este sentido, la información aportada por un testigo ocular, sesgada o no, conllevaría las mimas consecuencias. No obstante, la predicción de EVCs o BGA podría exonerar públicamente a colectivos discriminados. En definitiva, dado que tanto puede estigmatizar como exonerar, el resultado de las predicciones se puede considerar neutro (Kayser y Schneider 2009).

En el marco legal, los diferentes estados deberán crear leyes que regulen el uso de los marcadores con fines de predicción de EVCs o BGA. Hasta el momento, pocas legislaciones recogen el uso de estos marcadores ya que la mayoría no han sido modificadas desde la introducción de los perfiles de ADN. En este sentido, destaca la legislación de Países Bajos, que permite explícitamente el uso de ADN para la inferencia del sexo, BGA y EVCs que no estén relacionadas con enfermedades y sean visibles desde el nacimiento (Kayser y de Knijff 2011).

Muchas legislaciones incluyen cláusulas por las que se permite el uso de otros marcadores, además de los STRs, siempre que sean no codificantes. En este sentido, la predicción de EVCs o BGA a través de marcadores no codificantes no sería ilegal, pero iría en contra del espíritu de las propias legislaciones. Una aproximación más correcta sería regular las diferentes aplicaciones forenses y no los marcadores utilizados para cada fin, que son susceptibles de ser substituidos a medida que aumenta el conocimiento (Kayser 2015).

1.3.3 **Nuevas aplicaciones – nuevos marcadores**

Existe una serie de biomarcadores útiles en genética forense que permiten aplicaciones no factibles mediante los polimorfismos de ADN. Los polimorfismos de ADN, debido a su carácter estable frente a las condiciones ambientales y en los diferentes tipos celulares, representan un nivel estático de información. En un nivel dinámico de información surgen

biomarcadores que implican reordenamientos genéticos, metilaciones de islas CpG –los marcadores epigenéticos más desarrollados debido a su sencillo análisis– o diferencias en la tasa de transcripción.

Las aplicaciones más exploradas actualmente son: (i) predicción de edad e (ii) identificación de tipos celulares. Otras posibles aplicaciones incluyen (Vidaki *et al.* 2013):

– Determinación del origen parental de los alelos a través de los procesos de impronta genética que se producen en algunos *loci* (Nakayashiki *et al.* 2009).

– Autentificación de las muestras de ADN: los métodos de producción de ADN artificial *in vitro* (Lasken y Egholm 2003) no son capaces de imitar los patrones de metilación de islas CpG que se producen *in vivo* (Frumkin *et al.* 2010).

– Diferenciación de gemelos monocigóticos: diferencias en los patrones de metilación de islas CpG pueden permitir la identificación individual de gemelos monocigóticos (Li *et al.*).

– Hora de depósito de la muestra: los ritmos circadianos se traducen a nivel molecular en modificaciones epigenéticas cíclicas que regulan una pequeña cantidad de genes (Albrecht 2006, Feng y Lazar 2012) y que producen diferencias postranscripcionales a lo largo de las 24 horas del día (Pegoraro y Tauber 2008, Lech *et al.* 2016).

i. Predicción de edad

Los episomas de ADN sjTRECs *–signal joint T cell receptor excision circles–* se producen durante los reordenamientos genéticos de las células T y están directamente relacionados con la cantidad de dichas células. La cantidad de células T está, a su vez, inversamente relacionada con la edad ya que desde el nacimiento del individuo se produce progresivamente una involución y pérdida de función del timo. Mediante estos marcadores, se han logrado precisiones de predicción de entre el 88-97% para grupos de edad separados en 20 años y, en edades puntuales, correlaciones >0,8 con desviaciones típicas de 8,9 años (Zubakov *et al.* 2010). Esta correlación ha sido confirmada obteniéndose resultados similares en otro estudio (Cho *et al.* 2014).

Otros marcadores aplicables a la predicción de edad son las metilaciones de islas CpG (Vidaki *et al.* 2013). Diversos estudios de epigenomas han identificado una serie de posiciones CpG (Bocklandt *et al.* 2011, Hannum *et al.* 2013, Horvath 2013) cuyos patrones de metilación permiten predecir la edad con intervalos de desviación de 5 años respecto a la real. Estos marcadores fueron recogidos en ensayos a pequeña escala para diferentes tejidos como sangre (Weidner *et al.* 2014, Freire-Aradas *et al.* 2016), manchas de sangre (Huang *et al.* 2015), semen (Lee *et al.* 2015a) o dientes (Bekaert *et al.* 2015).

ii. Identificación de tipos celulares

La identificación de los diferentes fluidos corporales hallados en la escena del crimen proporciona información de alta utilidad para las investigaciones policiales (Sijen 2015). Aunque existen numerosos ensayos presuntivos aplicables, la necesidad de una mayor

especifidad y de diferenciar un mayor número de tejidos y fluidos celulares ha motivado la búsqueda de nuevos biomarcadores.

Actualmente existen numerosos ensayos (Sijen 2015) que permiten diferenciar fluidos o tejidos basados en ARN mensajero (Fleming y Harbison 2010, Lindenbergh *et al.* 2013, Zubakov *et al.* 2015), microARN (Hanson *et al.* 2009, Park *et al.* 2014b) y patrones de metilaciones de islas CpG (An *et al.* 2013, Park *et al.* 2014a, Lee *et al.* 2015b).

1.4 VALIDACIONES FORENSES

En genética forense, debido a las implicaciones que conlleva la información obtenida a partir de los marcadores, cobra una especial importancia establecer garantías y controles de calidad de los análisis, así como alcanzar un alto nivel de estandarización de los mismos. Desde hace unos años, la acreditación de los laboratorios de genética forense para normas internacionales como las que propone la ISO *–International Organization for Standardization–* constituye un estándar de calidad y, en algunos estados, un requisito para la incorporación de perfiles genéticos en las bases de datos policiales.

Uno de los requerimientos técnicos de la Norma ISO/IEC 17025[15], en la que se establecen los requisitos que deben cumplir los laboratorios de ensayo y calibración, es que los métodos utilizados en los laboratorios de ensayo deben estar validados. Los expertos de cada campo deben establecer, por lo tanto, criterios para llevar a cabo estas validaciones.

El término validación hace referencia al proceso que demuestra que un procedimiento de laboratorio es robusto –los ensayos son exitosos en un alto porcentaje y son necesarias pocas repeticiones–, fiable –los resultados son precisos y reflejan correctamente la muestra que se está analizando– y reproducible –se obtienen los mismos resultados o muy parecidos cada vez que una muestra es analizada– (Butler 2012b).

Existen dos principales niveles de validación: la validación de desarrollo y la validación interna. Las validaciones de desarrollo se realizan generalmente por parte de los fabricantes de equipos instrumentales o reactivos; mientras que las validaciones internas implican la verificación de que los ensayos realizados en el laboratorio funcionan correctamente y, por tanto, los equipos, materiales de laboratorio y reactivos empleados en los mismos (Butler 2012b).

Aunque cada laboratorio es el responsable de establecer su propio plan de validación interna, las comunidades de expertos redactan guías en las que ofrecen ciertas recomendaciones. Se debe tener en cuenta que estas guías no recogen criterios para la validación de ensayos dirigidos a plataformas de MPS debido a la reciente implantación de las mismas. No obstante, se pueden adaptar en la medida de lo posible los criterios para ensayos dirigidos a CE.

[15] ISO/IEC 17025:2005(ES)

En general, para llevar a cabo una validación interna se deben analizar tanto muestras de ADN de concentración conocida y buena calidad –estándares de ADN–, como muestras de casos finalizados o que imiten las condiciones que típicamente se dan en el contexto forense. De esta manera, se pueden establecer de antemano los límites de la metodología a la hora de analizar muestras comprometidas (Butler 2012b).

En la guía de criterios mínimos recomendados para la validación de varios aspectos del proceso de genotipado de ADN[16] de la ENFSI (*European Network of Forensic Science Institutes*) se recogen una serie de parámetros mínimos que deben ser evaluados para la validación interna de un nuevo *multiplex*:

i. Repetitividad: precisión –grado de concordancia entre series de mediciones individuales– y exactitud –grado de conformidad de una medida con su valor real– de los resultados (cuantitativa o cualitativamente) obtenidos por el mismo operador y/o instrumento de detección.

ii. Reproducibilidad: precisión y exactitud de los resultados (cuantitativa o cualitativamente) obtenidos por diferentes operadores y/o instrumentos de detección.

iii. Sensibilidad: rango de cantidades de ADN capaces de producir resultados de genotipado fiables y reproducibles, que debe cubrir el rango de concentraciones de ADN encontrado en las muestras que se deberán analizar.

iv. Análisis de mezclas de ADN: se realiza por replicado sobre una serie de proporciones de mezcla definidas. Se debe evaluar la capacidad de detección de mezclas, así como la capacidad de distinguir los componentes minoritario y mayoritario de las mezclas.

v. Análisis del balance de los marcadores: comprobar el balance de los picos de los alelos de un *locus* heterocigoto y entre los diferentes *loci*. En MPS la altura de los picos de los alelos se relaciona con el número de lecturas de cada alelo y, entre los *loci*, con la media de lecturas de cada marcador frente al resto.

vi. Comprobar que los % de *stutter* se ajustan a los valores indicados por el fabricante. En MPS, los *stutter* de los STRs se calcularían como el % de lecturas del alelo frente a lecturas con una repetición menos.

vii. Estudios de concordancia: para estos estudios se pueden utilizar muestras de ADN control que han sido analizadas mediante otras tecnologías para los mismos marcadores y, por lo tanto, tienen genotipos conocidos.

Otros posibles parámetros incluyen la sensibilidad a inhibidores y ADN degradado o la detección de los componentes masculino y femenino en mezclas cuando se validan kits con marcadores específicos de ADN masculino.

[16] http://www.enfsi.eu/sites/default/files/documents/minimum_validation_guidelines_in_dna_profiling_-_v2010_0.pdf

2. Objetivos

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2. Objetivos

2.1 OBJETIVOS PRINCIPALES

Los objetivos principales de esta tesis son:

1. Explorar las posibilidades que ofrece el uso de las nuevas tecnologías de MPS en el campo de la genética forense, evaluando el rendimiento de diferentes paneles de SNPs de identificación para la plataforma Ion PGM™.

2. Desarrollar nuevos paneles de predicción de ancestralidad biogeográfica, aumentando los niveles de información que se pueden obtener e implementándolos en sistemas tanto de electroforesis capilar como de MPS.

3. Desarrollar paneles de nuevos marcadores, tanto SNPs como STRs, que permitan una detección e interpretación más sencilla de las mezclas de ADN.

2.2 OBJETIVOS ESPECÍFICOS

Con el fin de alcanzar los objetivos principales planteados, se propusieron una serie de objetivos específicos que representan trabajos de investigación independientes. Estos trabajos se distribuyen en tres bloques, en correspondencia con los tres objetivos principales.

- **Bloque I**: ID-SNPs en MPS.
  - Evaluar y validar el panel HID-Ion AmpliSeq™ Identity para la plataforma Ion PGM™ en un estudio interlaboratorio.
  - Evaluar y validar internamente el panel Qiagen SNP-ID en la plataforma Ion PGM™.
- **Bloque II**: Ancestralidad biogeográfica.
  - Diseñar, optimizar y validar un panel de SNaPshot que permita distinguir con igual poder de diferenciación 5 grandes grupos poblacionales continentales.
  - Adaptar el panel EUROFORGEN Global AIM-SNP a la plataforma Ion PGM™; validar el diseño optimizado y analizar nuevas poblaciones.
- **Bloque III:** Mezclas de ADN.
  - Diseñar, optimizar y validar un panel de STRs pentaméricos para electroforesis capilar.
  - Diseñar, optimizar y validar un panel de SNPs de identificación multialélicos mediante SNaPshot.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. Bloque I: ID-SNPs en MPS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. Bloque I: ID-SNPs en MPS

En este bloque se presentan los trabajos de validación interna de dos paneles de SNPs de identificación para la plataforma de MPS Ion PGM™: HID-Ion AmpliSeq Identity v. 2.2 de TFS –sección 3.1 –y Qiagen SNP-ID –sección 3.2 –.

3.1 VALIDACIÓN DEL PANEL HID-ION AMPLISEQ™ IDENTITY V. 2.2

En este trabajo se presenta una validación interlaboratorio del panel HID-Ion AmpliSeq™ Identity v. 2.2 para la plataforma Ion PGM™. Siguiendo un esquema de validación sencillo, se evaluó la calidad de las secuencias obtenidas, la precisión del genotipado, la sensibilidad forense a ADN *low level* y degradado y la capacidad de detección de mezclas de ADN. Los resultados se encuentran publicados en la siguiente referencia:

> Eduardoff M, Santos C, **de la Puente M**, Gross TE, Fondevila M, Strobl C, Sobrino B, Ballard D, Schneider PM, Carracedo Á, Lareu MV, Phillips C (2015). "*Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™*." Forensic Sci Int Genet 17: 110-121.

3.1.1 **Material y métodos**

3.1.1.1 Muestras, extracción de ADN y preparación de mezclas artificiales

Para medir el grado de concordancia de los genotipos y evaluar la calidad de las secuencias obtenidas por los diferentes laboratorios, se pusieron en común una serie de muestras de ADN. Estas muestras incluyen: (i) seis muestras donadas voluntariamente por el personal de los laboratorios (S1-S6), que pudieron ser analizadas repetidamente e intercambiadas entre laboratorios; (ii) controles estándar de ADN 9947A y 007; (iii) controles de ADN Coriell, que permiten contrastar los genotipos obtenidos en los laboratorios con los listados en las bases de datos del Proyecto 1000 Genomas –Fases I y III– y el proyecto Complete Genomics, e incluyen: NA06994, NA07000, NA07029, NA18498, HG00403, NA10540 y NA11200. El uso de los controles de ADN Coriell permite comparar tres sistemas de genotipado diferentes, ya que los genotipos del Proyecto 1000 Genomas han sido obtenidos principalmente mediante la plataforma Illumina HiSeq (The Genomes Project Consortium 2012) y los de Complete Genomics mediante secuenciación por ligado (Drmanac *et al.* 2010).

Para evaluar la sensibilidad de la plataforma Ion PGM™ se prepararon diluciones de los controles de ADN 9947A y 007 y se utilizaron diferentes cantidades iniciales de ADN para la PCR de captura: 10ng, 1 ng, 100 pg, 50 pg y 25 pg. Durante la PCR de captura se aplicaron

un número variable de ciclos –Tabla 4–. Dos *runs*[17] se realizaron con *pools* de librerías a 8 pM, siguiendo las recomendaciones de preparación de librerías de Ion AmpliSeq™; mientras que los 3 *runs* restantes se realizaron con *pools* a 26 pM para determinar si el incremento de la concentración de las librerías permitía obtener mejores resultados en ADN *low level*. Las cantidades iniciales de ADN menores de 1 ng fueron amplificadas con únicamente 25 ciclos o con 5 ciclos adicionales después de la preparación de la librería. Se utilizaron 2 aproximaciones para la reamplificación: (i) reamplificar la mitad de la librería de cada muestra y comparar con la mitad no reamplificada; y (ii) preparar librerías separadas para cada muestra, con y sin reamplificación. Las librerías fueron cuantificadas mediante el Ion Library Quantitation Kit para constituir los *pools* equimolares.

Tabla 4. Diluciones empleadas para el estudio de sensibilidad, analizadas en 5 *runs* diferentes: concentraciones del *pool* de librerías y número de ciclos aplicados en cada caso.
× = misma muestra reamplificada; ∧ = réplicas de librería.

			Runs				
	ADN inicial	Ciclos	8 pM		26 pM		
			Lab 1-A	Lab 1-B	Lab 1-E	Lab 1-F	Lab 1-C
9947A	10 ng	18			•		
	1 ng	21	•		•		
	100 pg	21			×		
	100 pg	25	•				
	50 pg	25	•		×		
	25 pg	25			×		
	100 pg	21+5				×	
	100 pg	25+5	•				
	50 pg	25+5	•			×	
	25 pg	25+5				×	
007	10 ng	18			•		
	1 ng	21	•		•		
	100 pg	21			×		
	100 pg	25	•	∧			∧
	50 pg	25	•	∧	×		∧
	25 pg	25			×		
	100 pg	21+5				×	
	100 pg	25+5	•	∧			∧
	50 pg	25+5	•	∧		×	∧
	25 pg	25+5				×	

Para evaluar la capacidad de detección de mezclas de ADN de la plataforma, se prepararon mezclas de las muestras de ADN S5 (hombre) y S6 (mujer) en ratios de volumen de 1:9, 1:3, 1:1, 3:1 y 9:1. Cada ratio de mezcla fue preparada una única vez y se construyeron

[17] El término *run* se refiere a las muestras analizadas conjuntamente en un mismo chip.

2 librerías. Las dos librerías de cada ratio de mezcla fueron identificadas mediante diferentes *barcodes* para ser secuenciadas en un único chip Ion 316™.

Para evaluar la capacidad de análisis de ADN degradado se utilizó una muestra de ADN antiguo masculina (S7) extraída a partir de restos arqueológicos del siglo XII. Las condiciones de preservación de la muestra en Volders –Austria– se detallan en el trabajo de Bauer *et al.* (2013). La muestra S7 fue analizada en dos réplicas de PCR con la máxima cantidad posible de ADN (un total de 450 pg de acuerdo con Quantifiler® Duo), usando 25 y 25+5 ciclos de reamplificación. Aunque no se dispone de datos de referencia para esta muestra, se evaluó la consistencia de los genotipos entre ambos análisis.

3.1.1.2 Preparación de librerías para Ion PGM™

Las librerías del panel HID-Ion Ampliseq™ Identity Panel v2.2 se construyeron mediante el kit AmpliSeq™ Library 2.0, siguiendo las recomendaciones del fabricante[18]. Previamente, se cuantificaron todas las muestras mediante el kit Qubit® dsDNA HS Assay. La PCR fue realizada, atendiendo a las recomendaciones para 196 pares de *primers*, con 18-21 ciclos. Después de la digestión parcial de los *primers*, se ligaron los adaptadores con *barcode* Ion Xpress™ Barcode Adapters para identificar a cada muestra. El producto fue purificado mediante *beads* magnéticas Agencourt AMPure XP. La calidad de las librerías fue evaluada con los kits Qubit® dsDNA HS Assay, Agilent® High Sensitivity DNA o Ion Library Quantitation y, teniendo en cuenta los resultados, se formaron *pools* equimolares a 100 pM en ≥20 µL para cada *run*, de acuerdo con las recomendaciones del fabricante.

La preparación del molde de secuenciación se realizó mediante el kit Ion OneTouch™ 200 Template v2, siguiendo las recomendaciones del fabricante[19]. Sobre las ISPs que se recuperaron de la PCR de emulsión, se utilizó el kit Ion Sphere™ Quality Control para asegurar un 10-30% de ISPs con *template* antes del enriquecimiento con Ion PGM™ Enrichment Beads. La secuenciación fue llevada a cabo mediante el kit Ion PGM™ Sequencing 200 v2 y chips Ion 314™ o 316™ tipo v1 o v2, siguiendo las recomendaciones del fabricante[20].

3.1.1.3 Análisis de datos

El análisis de datos se realizó mediante el *software* Torrent Suite™ 4.0.2 (TS)[21] y el *plugin* HID_SNP_Genotyper 4.0.1 (Genotyper) con parámetros de baja rigurosidad –*low stringency*–. Se aplicaron los archivos HID_SNP_v2.2.2_hotspots.bed y HID_SNP_v2.2.2_targets.bed, que identifican los marcadores en base al genoma de referencia hg19. Los archivos de salida de Genotyper se analizaron mediante el *software* R v. 3.0.3 (R Core Team 2014).

[18] Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user guide. July (2013).
[19] Thermo Fisher Scientific, Life Technologies: Ion OneTouch™ 200 Template Kit v2 user guide, 2012.
[20] Thermo Fisher Scientific, Life Technologies: Ion PGM™ 200 Sequencing Kit user guide, 2012.
[21] Thermo Fisher Scientific, Life Technologies: Torrent Suite™ software 4.0.2. user guide. November (2013).

3.1.2 **Resultados**

3.1.2.1 *Coverage* obtenido a través de la plataforma Ion PGM™

Uno de los parámetros clave para la obtención de genotipos precisos mediante MPS es la cobertura o *coverage*, que representa el número de veces que cada base ha sido leída en cada *run* de secuenciación. Para aplicaciones genómicas se suele representar como una media del número de lecturas por cada base; no obstante, para el genotipado de SNPs es más adecuado el uso del *coverage* específico de cada SNP, definido como el número de lecturas en la posición del SNP –SNP *Target Reads*–. El valor final de SNP *Target Reads* depende de la plataforma de secuenciación, de los parámetros aplicados para el filtrado de secuencias y de como se procesa la asignación de variantes. En los *runs* de Ion PGM™, el número de pocillos en cada chip define el número máximo de lecturas al determinar el número máximo de ISPs que pueden ser secuenciadas. Los procesos de *pooling* de muestras, preparación del molde de secuenciación (que influye en el número de ISP vacías o policlonales) y la eficiencia de la carga del chip influyen en el número de ISPs monoclonales secuenciadas. Durante el procesamiento de datos en la TS se filtran las lecturas que provienen de ISPs policlonales, las lecturas de baja calidad o las que representan dímeros de adaptadores. Cuando se secuencian múltiples librerías mediante el uso de *barcodes*, el *pooling* equimolar de las mismas pretende lograr una distribución homogénea de las lecturas entre las muestras del *run*.

En este estudio, los 12 *runs* analizados alcanzan los niveles de lecturas esperados de acuerdo con las guías de la casa comercial, evaluados como Mb por *run*, para cada tipo de chip y versión utilizado –ver Fig. 9–. Se debe destacar que en los *runs* que combinan muestras de ADN *low level* y muestras con ADN en cantidad inicial óptima (Lab 1) se filtran más lecturas durante el proceso de *base calling*. Mientras que la cantidad de lecturas filtradas en base a criterios de calidad es similar para todos los *runs*, el porcentaje de lecturas filtradas debido a la presencia de dímeros es ligeramente superior en los *runs* de Lab 1 en los que se combinan muestras con cantidad inicial de ADN óptima y de ADN *low level* ($p=0,029$; $\alpha=0,05$).

Los dímeros podrían haber provocado un sesgo al alza en la cuantificación de las librerías. No obstante, los resultados de cuantificación de librerías no muestran ninguna correlación con la cantidad inicial de ADN, número de ciclos de amplificación, laboratorio en el que se llevaron a cabo o método de cuantificación. En la Fig. 10 se muestran los resultados de la cuantificación de las librerías de los 101 análisis[22] frente a la cantidad inicial de ADN utilizado. Los diferentes *runs* fueron reanalizados para obtener el total de lecturas sin filtrar por calidad y *trimmering* –proceso por el que se eliminan las primeras y últimas bases de las lecturas en función de la calidad o de la secuencia de los adaptadores– individualmente para cada muestra. Se observaron diferencias en el porcentaje de lecturas filtradas para cada muestra, siendo mayor en las muestras de ADN *low level* ($p=2x10^{-6}$; $\alpha=0,05$).

[22] El término análisis se refiere a los resultados de secuenciación de una única muestra en cada *run*.

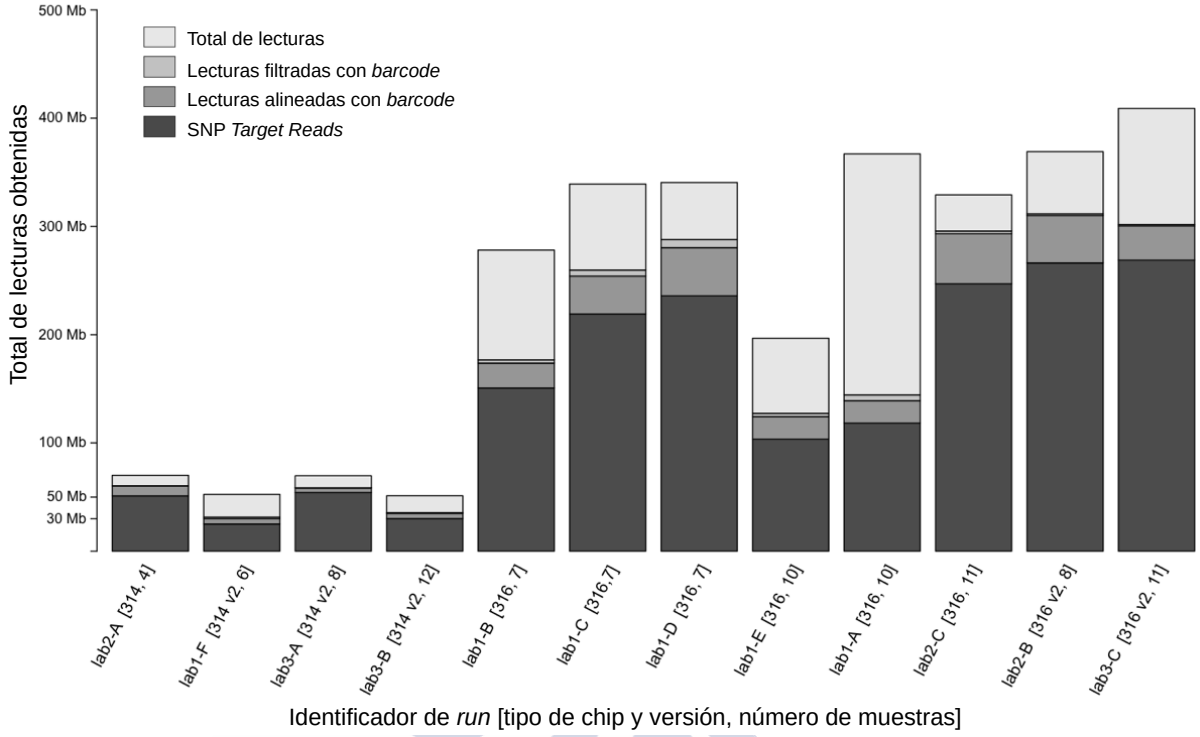


Fig. 9. Proporciones de cuatro tipos de lecturas en los 12 *runs* de Ion PGM™ realizados mediante todo el rango disponible de chips de secuenciación.

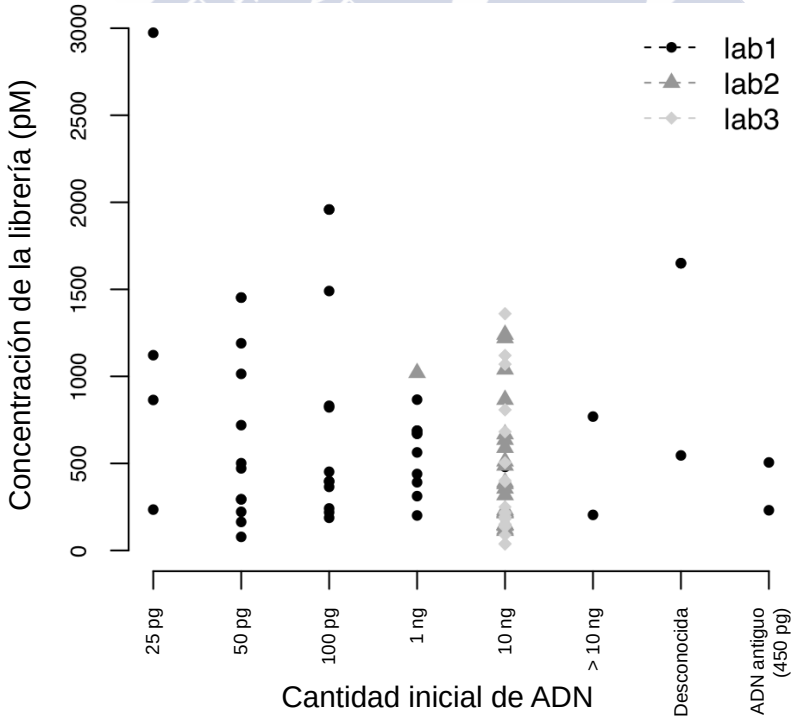


Fig. 10. Concentración de las 101 librerías construidas a partir de diferentes cantidades iniciales de ADN.

La presencia de dímeros afecta más acusadamente a los análisis de controles negativos. Se incluyeron controles negativos en dos *runs* diferentes. En el primer *run* se combinaron 2 librerías de controles de ADN diluidas a 100 pM y preparadas con una cantidad inicial de ADN óptima y 2 librerías de controles negativos no diluidas, el *pool* de librerías se diluyó 2:23 para la preparación del molde de secuenciación y fueron analizadas mediante un chip 314v2. El porcentaje de lecturas policlonales en este *run* ascendió al 51%, un porcentaje más alto que el de los 12 *runs* que presentan un porcentaje de policlonalidad promedio de 30% (desviación típica= 0,8). En el segundo *run*, se combinaron 6 librerías de controles negativos y una librería de control de ADN diluida a 100 pM. El *pool* de librerías no fue diluido antes de la preparación del molde de secuenciación para mantener la concentración entre 1-2 pM. En este *run* se produjeron un 79% de lecturas policlonales.

El 64% (6065/9783) de las lecturas obtenidas para los controles negativos se filtraron debido a la baja calidad de las mismas o a la presencia dímeros de adaptadores. Al igual que en las muestras de ADN *low level*, los controles negativos muestran picos en el histograma de lecturas no filtradas alrededor de 50 pb –ver Fig. 11–, que no aparecen en muestras analizadas con la cantidad inicial de ADN óptima. Entre las lecturas no filtradas, un 76% (2801/3650) se alinearon con el hg19. Cinco de las lecturas se alinearon con el rs1058083, una cantidad comparable a las lecturas de SNPs de cromosoma Y en muestras femeninas –ver sección 3.1.2.2.1–. Al visualizar las lecturas alineadas en IGV, se observa que alrededor de un 28% de las secuencias (1036/3650) alinean en 61 regiones en las que también aparecen, estocásticamente, lecturas en las muestras de ADN *low level*. Estas regiones coinciden con la posición esperada de uno de los *primers* de la PCR de captura, y no se observan en muestras con cantidad inicial de ADN óptima –ver Fig. 12–. El 6% restante de secuencias no filtradas (228/3650) consisten en alineamientos al azar con el resto del genoma.

En la Fig. 13A se muestran las distribuciones de SNP *Target Reads* para los 101 análisis. La distribución de los cuartiles revela variación intra- e inter-*run*. En la Fig. 13C se muestra como en los *runs* de Lab 1, que combinan muestras de ADN *low level* con muestras con cantidad inicial de ADN óptima, la variación entre muestras es mayor. La Fig. 13B muestra la desviación frente a un máximo alcanzable de SNP *Target Reads*, calculado como:

$desviación = \frac{SNP\ Target\ Reads\ observadas - SNP\ Target\ Reads\ esperadas}{Total\ de\ lecturas\ clonales\ en\ el\ chip}$, en donde las SNP *Target Reads* esperadas se corresponden al número total de lecturas clonales en el chip entre el número de muestras incluidas en el mismo y el total de lecturas clonales en el chip se corresponde con el total de lecturas que pasan el filtro de policlonalidad. En comparación con los análisis de ADN *low level*, los análisis de muestras con cantidad inicial de ADN óptima presentan una desviación menor. Además, los análisis de *coverage* de las muestras de ADN *low level* presentan más lecturas *off-target* –fuera del amplicón diseñado para la PCR de captura– ($p=0,00045$; $\alpha=0,05$), que se corresponden principalmente a las lecturas de *primers* de la PCR de captura –ver Fig. 12–. Estas secuencias se alinean correctamente con el genoma de referencia e incrementan el número total de lecturas clonales en el chip.

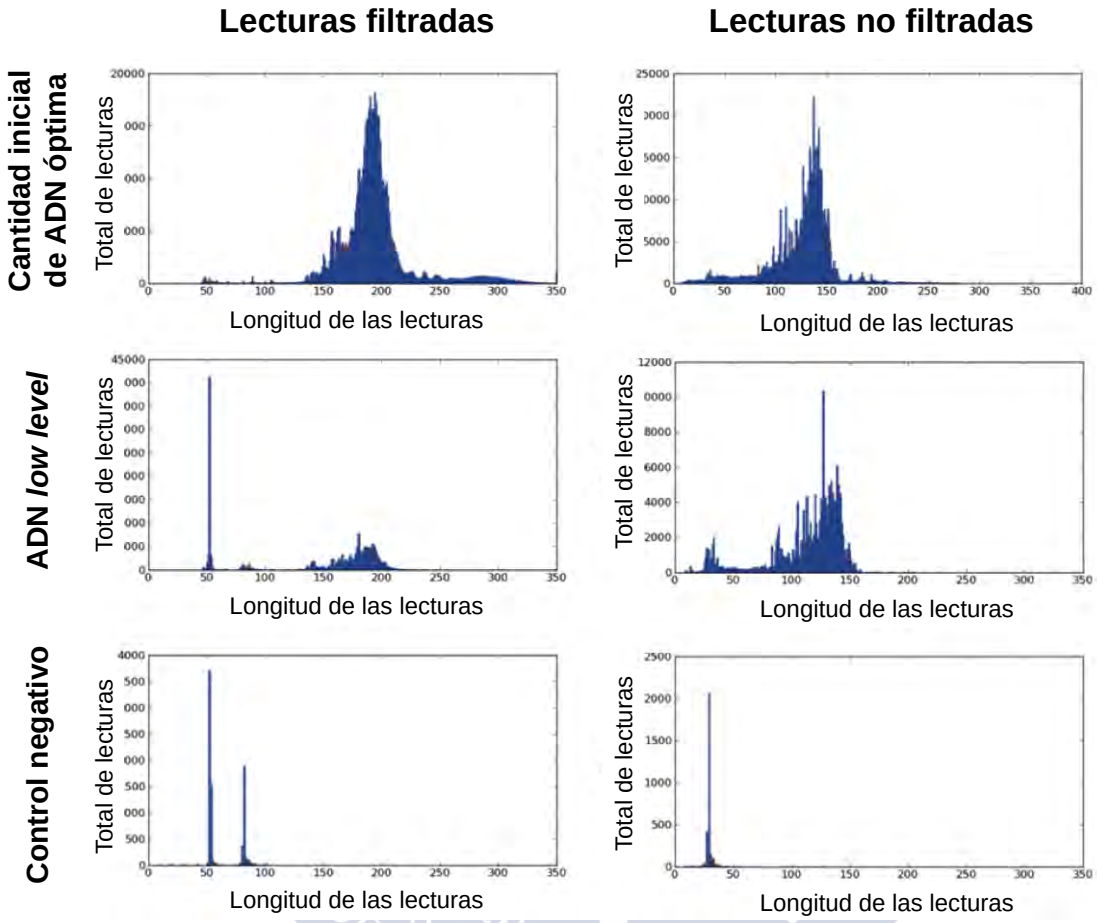


Fig. 11. Histogramas de la longitud de las lecturas de una muestra con cantidad inicial de ADN óptima, una muestra de ADN *low level* y un control negativo, antes (izquierda) y después (derecha) del filtrado.

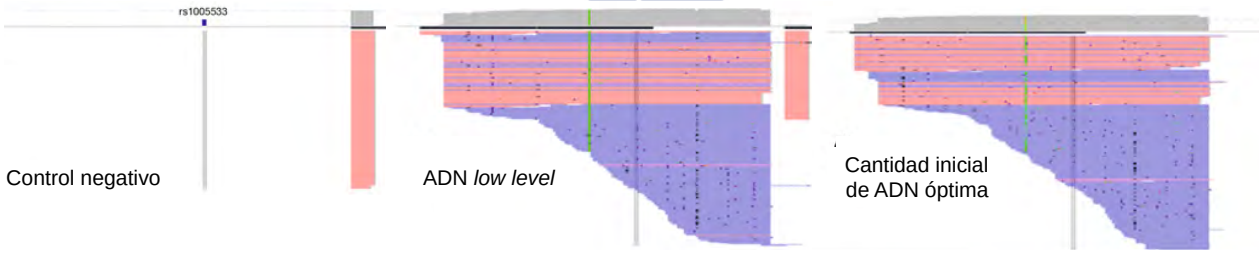


Fig. 12. Comparación de la visualización en IGV de las lecturas obtenidas en la región correspondiente al SNP rs1005533 en una muestra con cantidad inicial de ADN óptima, una muestra de ADN *low level* y un control negativo. Aparecen lecturas en la región en la que se espera que esté uno de los *primers* de la PCR de captura en el control negativo y las muestras de ADN *low level*.

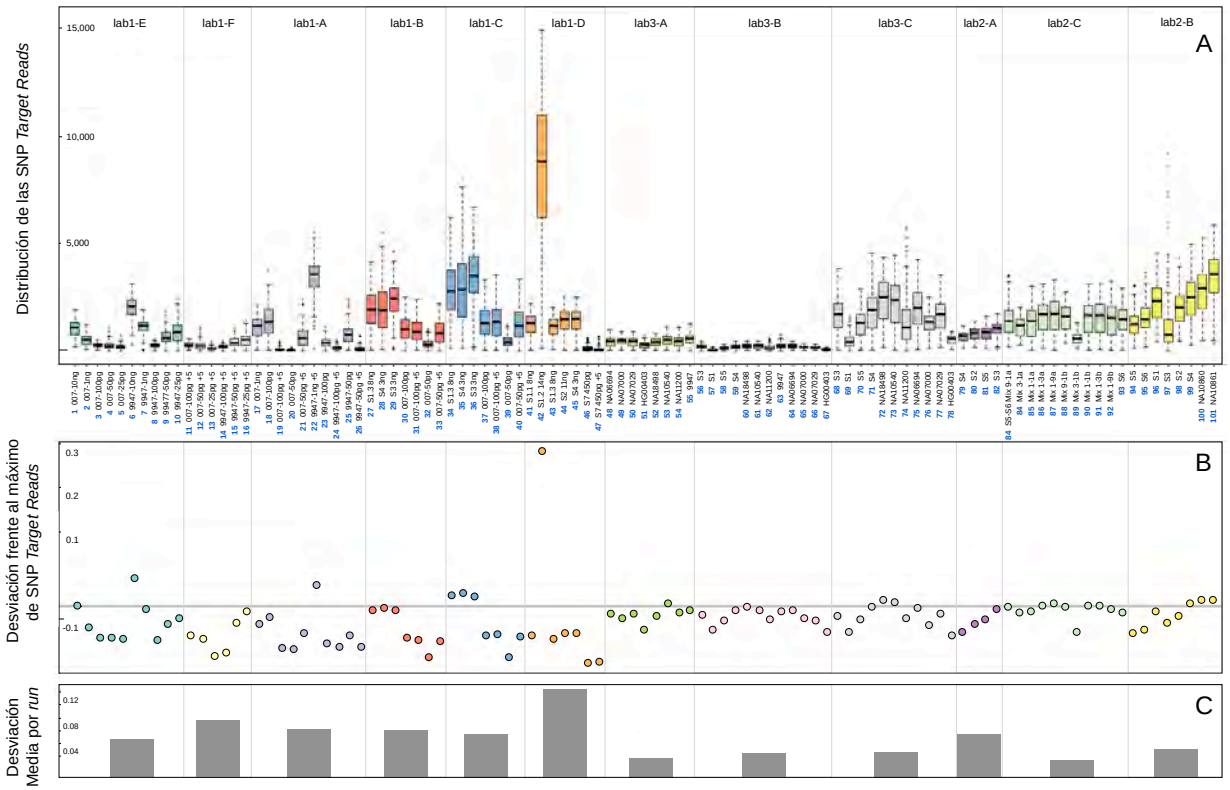


Fig. 13. A) Diagramas de caja que representan el total de SNP *Target Reads* obtenidas en 101 muestras distribuidas en 12 *runs*. B) Desviaciones frente al máximo de SNP *Target Reads*, ajustadas según el tipo de chip de secuenciación. C) Gráficos de barras que representan los valores promedio de las desviaciones en cada *run*.

A la hora de realizar un análisis *multiplex* de SNPs mediante MPS se debe tener en cuenta, por una parte, el umbral mínimo de *coverage* para la obtención de genotipos fiables y, por la otra, el número de muestras que pueden ser secuenciadas en un mismo *run* para llegar al umbral de *coverage* mínimo. En este caso, el umbral de *coverage* propuesto por la casa comercial para la detección de variantes es de 30x para la línea germinal y 500x para somática. El valor de 30x se corresponde adecuadamente con el que proponen otros estudios de secuenciación de genomas completos y de detección de variantes utilizando métodos de captura (Bentley *et al.* 2008, Koboldt *et al.* 2010, Nielsen *et al.* 2011, Elsharawy *et al.* 2012, Quail *et al.* 2012, Sims *et al.* 2014). Sin embargo, los umbrales de *coverage* mínimo dependen generalmente de la aplicación, de los algoritmos utilizados para asignar los genotipos y de los parámetros establecidos en el análisis. Para aplicaciones forenses, un umbral de ~20x puede ser suficiente para detectar variantes en muestras individuales de alta calidad, mientras que las mezclas o el análisis de ADN *low level* requerirán de valores de *coverage* más altos. En este estudio, los valores de *coverage* más bajos para los que se obtuvieron genotipos concordantes en SNP autosómicos (A-SNPs) y SNPs de cromosoma Y (Y-SNPs) son de 13x y 41x, respectivamente. Estos valores concuerdan con el estudio de Daniel *et al.* (2015), que estima un *coverage* mínimo de 20x para la obtención de genotipos fiables. Sin embargo, en mezclas de ADN encontramos que el valor de *coverage* más bajo para el que se obtienen genotipos concordantes con los de la mezcla esperada es de 269x para A-SNPs, mientras que para

Y-SNPs es de 63x en la mezcla de ratio 1:9 (hombre-mujer) y asciende a 274x en la mezcla de ratio 9:1.

Para estimar el número de muestras que se deben cargar por cada *run*, la casa comercial proporciona una serie de guías que permiten calcular el número de muestras que se pueden combinar para alcanzar cierto *coverage* mínimo en el 95% de las bases. En este estudio, las muestras se combinaron en los diferentes tipos de chips disponibles para alcanzar valores de *coverage* mínimo de entre 42-286x en el 95% de las bases –ver Tabla 5–. No obstante, el *coverage* mínimo alcanzado en cada muestra para el 95% de las bases no se incluye entre los archivos generados en los análisis. Los valores de *coverage* mínimo calculados solo fueron alcanzados por el 95% de los SNPs del panel en 8 muestras (31 muestras si se descuentan los SNPs atípicos), todas ellas analizadas con una cantidad inicial de ADN óptima. A partir estos análisis, se infiere que los valores de *coverage* mínimo para el 95% de las bases a los que se debe apuntar son de al menos 62x para obtener valores finales de al menos 13x (*coverage* mínimo de los A-SNPs concordantes). El *run* Lab 3-B fue retirado de los análisis posteriores, ya que ninguna de las muestras con cantidades iniciales de ADN óptimas incluidas alcanzaron el umbral de 13x.

Tabla 5. *Coverage* mínimo esperado para el 95% de las bases para cada uno de los *runs* analizados, en función del tipo de chip utilizado y el número de muestras incluidas.

Run	Tipo de chip	Número de muestras	*Coverage* mínimo para el 95% de las bases
lab1_F	314	6	83
lab2_A	314	4	125
lab3_A	314v2	8	62,5
lab3_B	314v2	12	42
lab1_E	316	10	200
lab1_A	316	10	200
lab1_B	316	7	286
lab1_C	316	7	286
lab1_D	316	7	286
lab2_B	316	11	181,82
lab3_C	316v2	11	181,82
lab2_B	316v2	8	250

En la Fig. 14B se muestra un *heatmap* en el que se ordenan las muestras analizadas en función del *coverage* promedio de cada análisis (creciente de arriba hacia abajo, de manera que la mayoría de las muestras de la parte superior se corresponden a los análisis de ADN *low level*) y de cada SNP (de izquierda a derecha), por separado para A-SNPs y de Y-SNPs. Aunque los patrones de *coverage* de cada SNP se mantienen a través de los diferentes análisis –Fig. 14B– las columnas que corresponden a los SNPs que están más a la izquierda presentan una mayor heterogeneidad que la media. El gráfico de barras de la Fig. 14A indica que el *coverage* promedio de los Y-SNPs es menor, en general, que el de los A-SNPs, tal y como se espera en correspondencia con el número de copias presentes en el genoma.

Como conclusión, las recomendaciones de la casa comercial son útiles para realizar una primera aproximación del número de muestras que se pueden cargar en función del tipo de chip para obtener cierto *coverage* mínimo. No obstante, el número de muestras a analizar en cada chip se debe evaluar individualmente para cada panel y teniendo en cuenta el tipo de muestras (calidad y cantidad de ADN, mezclas de ADN…) que se pretende analizar.

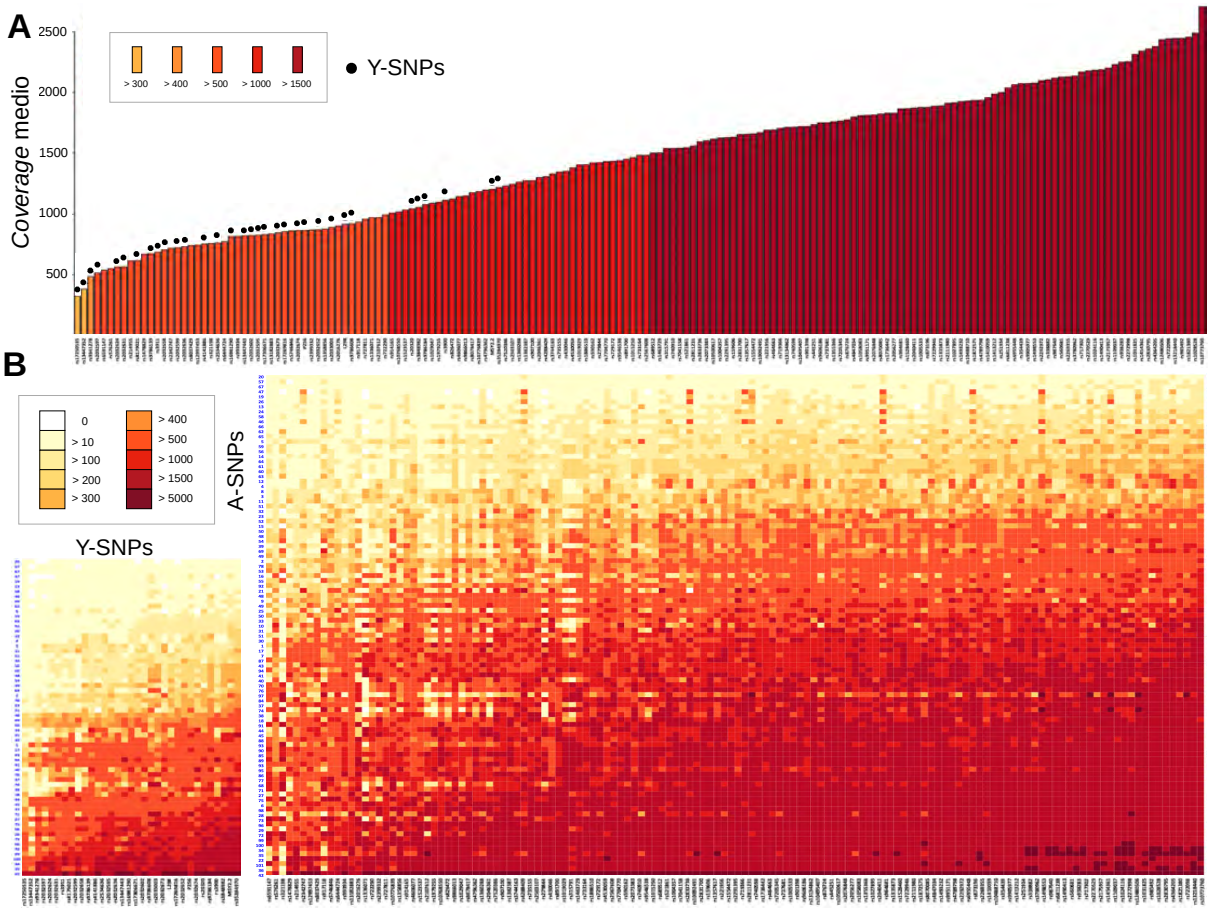


Fig. 14. A) Gráfico de barras en el que se muestra el *coverage* promedio de los SNPs. Los Y-SNPs se marcan con un punto. Los colores corresponden a intervalos de valores promedio de *coverage*. B) *Heatmap* para las muestras en las que se analizan Y-SNPs y A-SNPs: los colores de las celdas corresponden a intervalos de valores de *coverage*. Los análisis se ordenan de arriba abajo en orden creciente de *coverage* promedio. Los SNPs se ordenan de izquierda a derecha en orden creciente de *coverage* promedio.

3.1.2.2 Características de las secuencias que influyen en la obtención de genotipos

Las características de las secuencias se evaluaron mediante la comparación de los diferentes *runs* realizados en los diferentes laboratorios, centrándose en factores que afectan a la fiabilidad de los genotipos obtenidos como el *coverage*, la incorporación errónea de nucleótidos –*misincorporation*–, el balance de las lecturas de los alelos –*allele read frequency* (ARF)– y el sesgo de lectura entre las cadenas *forward* y *reverse* –*strand bias*–. En función de los resultados de estos análisis, se identificaron una serie de SNPs atípicos que deben ser

eliminados del panel o excluidos del análisis de datos en escenarios complejos: ADN *low level*, ADN degradado o detección de mezclas de ADN.

3.1.2.2.1 Tasa de incorporación de bases erróneas

Para evaluar la tasa global de incorporaciones erróneas –*misincorporation*– de bases en el Ion PGM™ (lecturas de bases incorrectas detectadas en la posición del SNP en bajas proporciones), se comparó la incidencia de incorporación de bases no-alélicas del SNP (p. ej. la incorporación de Gs y Ts en un SNP A/C) con la incidencia de incorporación errónea de bases alélicas en homocigotos (p. ej. pequeñas proporciones de lecturas A en un homocigoto CC en un SNP A/C). Si las tasas son similares, se puede establecer una línea base que se corresponde con una tasa de *misincorporation*. Si las tasas son diferentes, se puede conocer el nivel de ADN exógeno que se detecta de manera estocástica (equivalente a una tasa de *drop-in*) en función de la proporción de lecturas alélicas erróneas. En ambos casos, los SNPs atípicos que presentan valores de *misincorporation* por encima de la media pueden ser identificados y tenidos en cuenta a la hora de detectar un componente minoritario <10% en una mezcla de ADN.

La Fig. 15 recoge las frecuencias de *misincorporation* de bases alélicas y no-alélicas, ordenando los SNPs en función del valor promedio las mismas. Se puede observar que las tasas de *misincorporation* de bases alélicas (alelos referencia y alternativo) y no alélicas son similares, y que tan solo superan el 0,2% en 12/169 SNPs. De entre estos 12 SNPs con altas tasas de *misincorporation*, únicamente rs8078417, rs2399332, Y-rs2032597, rs9866013 y rs1523537 alcanzan tasas >1%. Los datos de estos SNPs deben ser descontados de los análisis de mezclas de ADN, ya que presentan patrones de desbalance de homocigotos que pueden ser confundidos con la señal de un alelo del componente minoritario, en particular rs2399332 y rs1523537, que presentan unas tasas desproporcionadas de *misincorporation*.

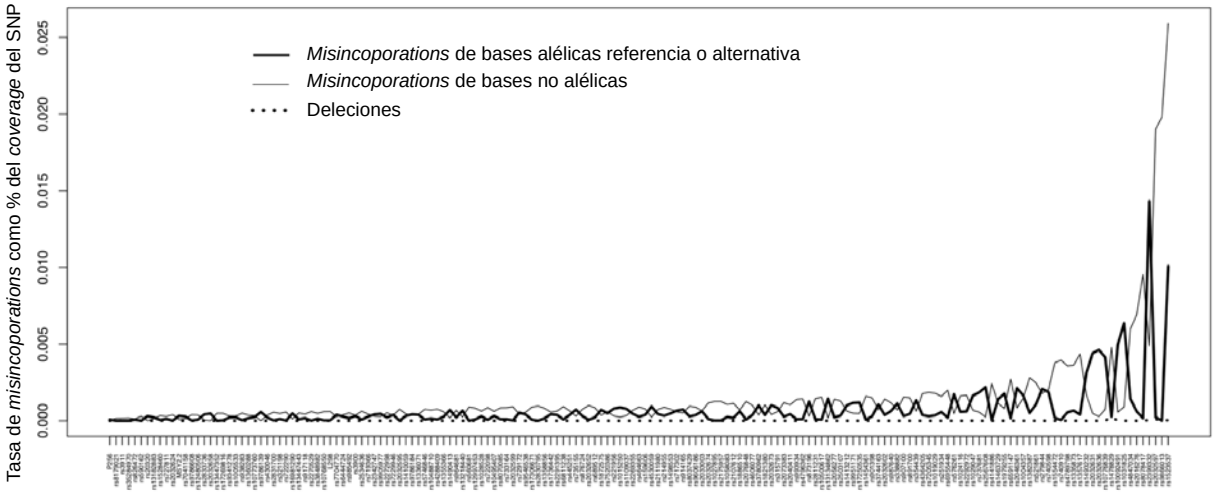


Fig. 15. Tasas de *misincorporation* registradas como presencia de bases alélicas referencia o alternativa erróneas (p. ej. lecturas A en homocigotos GG de un SNP A/G), bases no alélicas o deleciones.

Aunque las tasas de *misincorporation* de bases alélicas y no alélicas son lo suficientemente similares como para descontar una alta incidencia de fenómenos de *drop-in*, se observaron lecturas de amplicones de Y-SNPs en muestras femeninas. La Fig. 16 muestra un total de 34 lecturas de Y-SNP obtenidas en 6 análisis diferentes de las 2 muestras femeninas. No obstante, la obtención de 34 lecturas de ADN específicamente masculino entre más de 2 millones de lecturas indica unos niveles muy bajos de *drop-in*.

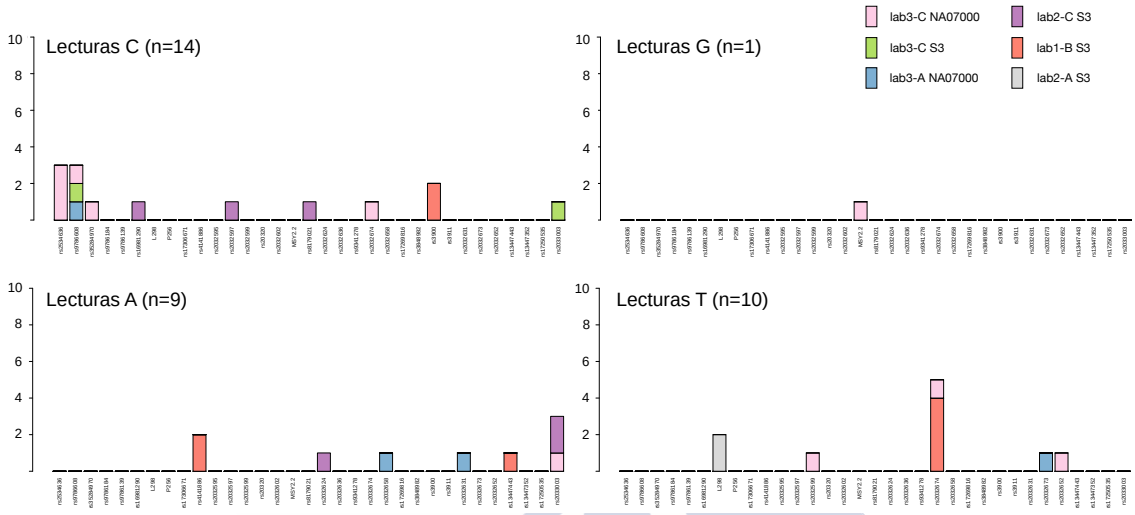


Fig. 16. Lecturas de Y-SNPs detectadas en 6 análisis de las dos muestras femeninas.

3.1.2.2.2 Balance de lecturas de los alelos

Una característica deseable de las metodologías de genotipado aplicadas al campo de la genética forense es la capacidad de diferenciar entre el desbalance de señales de heterocigotos producido estocásticamente durante la PCR y el de la combinación de señales de los alelos en las mezclas de ADN. Esta característica cobra una especial importancia en el análisis de marcadores binarios, como los 136 A-SNPs de este panel, en el que la detección de mezclas de ADN debe realizarse atendiendo al desbalance entre los alelos referencia y alternativo. Además, los Y-SNPs incluidos en el panel, seleccionados para establecer filogenias de cromosoma Y, presentan importantes limitaciones en la detección de mezclas de ADN, ya que la probabilidad de encontrar genotipos múltiples se reduce. Por ello, se evalúa el efecto del parámetro ARF[23] –*Allele Read Frequency*–, definido como número de lecturas de un alelo sobre el total de lecturas del SNP, sobre los genotipos obtenidos.

Se recogieron las frecuencias de las lecturas alélicas de los 169 SNPs para 38 análisis de 28 muestras de ADN masculinas y las de 136 A-SNPs para 10 análisis de muestras de ADN

[23] En el *software* de análisis de Ion PGM™ el término frecuencia alélica se utiliza para describir el porcentaje de lecturas de cada alelo en un marcador. Este término se puede confundir con el utilizado en genética poblacional, por lo que se opta por el término frecuencia de lecturas del alelo –ARF: *allele read frequency*– para diferenciarlo.

femeninas. La Fig. 17 muestra la distribución de las ARF de los alelos de referencia. Las distribuciones de los A-SNPs heterocigotos mostraron un buen nivel de agrupación en torno a los valores de 0,5 que representan el máximo balance teórico. Las distribuciones de Y-SNPs y A-SNPs homocigotos (en torno a los valores de 1 y 0) muestran mayor uniformidad, indicando que las ARF no sobrepasan los umbrales de <0,1 y >0,9.

Para la identificación de SNPs atípicos se valoró aplicar umbrales de ARF en heterocigotos de amplitud equivalente a los de homocigotos (de <0.55 y >0.45), pero una proporción alta de SNPs que presentan genotipos concordantes y fiables mostraron mayores niveles de desbalance. Aplicar umbrales de <0.6 y >0.4, indicados en la Fig. 17 como el área sombreada en gris en los A-SNPs, proporciona un equilibrio adecuado entre la proporción de genotipos fiables obtenidos y el balance de la señal en muestras con una cantidad de ADN inicial óptima. A partir de las medias de los valores de ARF de los alelos de referencia, se pueden identificar la mayoría de SNPs con distribuciones de ARF atípicas; no obstante, una inspección visual de la representación de los valores individuales –Fig. 17– permite identificar los SNPs que presentan valores de ARF sesgados en ambos sentidos, como rs1029047.

Los SNPs rs1029047, rs8037428, rs430046 y rs1523537 fueron identificados en este estudio y, a su vez, en la validación del mismo panel realizada por Børsting *et al.* (2014). Adicionalmente, el SNP rs2107612 fue identificado en este estudio pero no en el de Børsting *et al.* (2014). Además, los SNPs rs10776839, rs4530059 y rs1031825 fueron identificados como poco balanceados por el estudio de Børsting *et al.* (2014) pero muestran valores razonablemente balanceados en este estudio, aunque aparecen valores de ARF atípicos en una pequeña proporción de los análisis para rs4530059 y rs1031825.

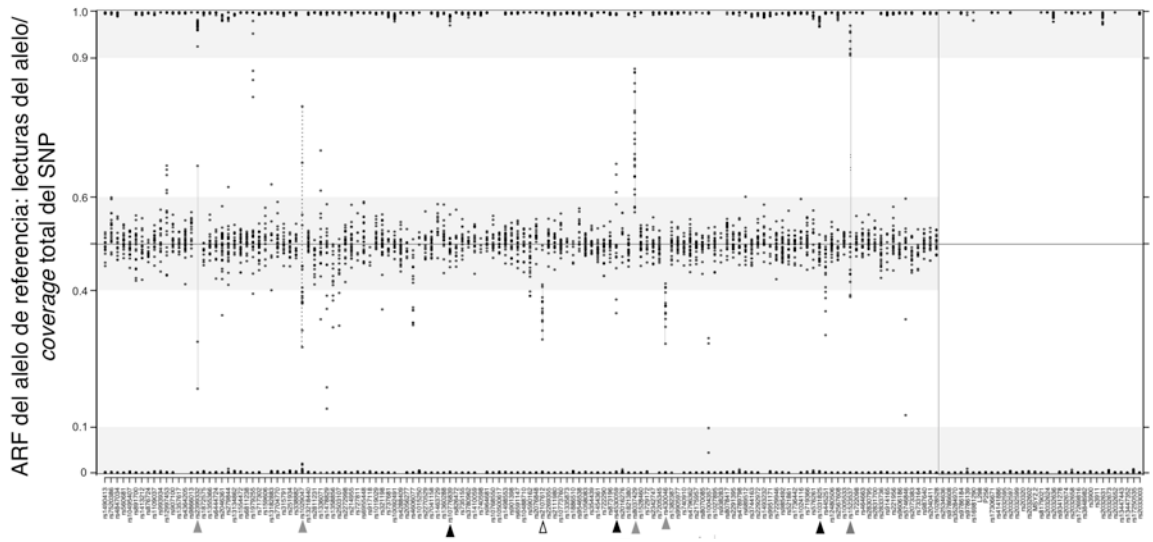


Fig. 17. ARF del alelo de referencia para los 169 SNPs incluidos en el panel (listados en orden cromosómico,Y-SNPs en la derecha). Los umbrales de ARF aplicados para detectar SNPs atípicos se indican mediante un sombreado gris. Se señalan los SNPs identificados a la vez por este estudio y el de Børsting *et al.* (2014) –triángulo gris–, únicamente en el estudio de Børsting *et al.* (2014) –triángulo negro– y únicamente en este estudio –triángulo blanco–.

Las ratios de ARF de homocigotos desviados de los valores ideales de 0 y 1 representan una baja proporción de lecturas que se corresponden con *misincorporations* alélicas o no alélicas. Conforme a los parámetros aplicados por el Genotyper, la proporción de lecturas del segundo alelo debe sobrepasar el 10% para considerar al SNP como heterocigoto, de manera que cuando los valores de ARF de uno de los alelos superan el 90% los genotipos no son asignados erróneamente como heterocigotos.

3.1.2.2.3 Sesgo de lecturas de las cadenas

La plataforma mide el sesgo de lecturas de las cadenas *forward* y *reverse* –*strand bias*– como *coverage* total de las SNP *Target Reads* de la cadena *forward* entre el total de SNP *Target Reads*. En principio, un desbalance de lecturas de las cadenas muy sesgado proporciona genotipos menos fiables.

Los resultados de *strand bias* se muestran en la Fig. 18, en la que los SNPs aparecen ordenados según orden creciente del promedio de *strand bias*. Se observa un amplio rango de valores desde 0,5 (sin sesgo discernible) hasta valores cercanos a 0 y 1 (que corresponden, respectivamente, a lecturas obtenidas únicamente para la cadena *reverse* y *forward*). Para identificar los SNPs atípicos, aplicamos umbrales de *strand bias* de >0,25 y <0,75; que representan diferencias de hasta el triple de lecturas en una cadena respecto a la otra. Un total de 9 SNPs presentaron valores promedio de *strand bias* fuera de estos umbrales –ver Fig. 18–, 3 de los cuales presentan además un alto número de *no-calls* y serán discutidos en la sección 3.1.2.4.

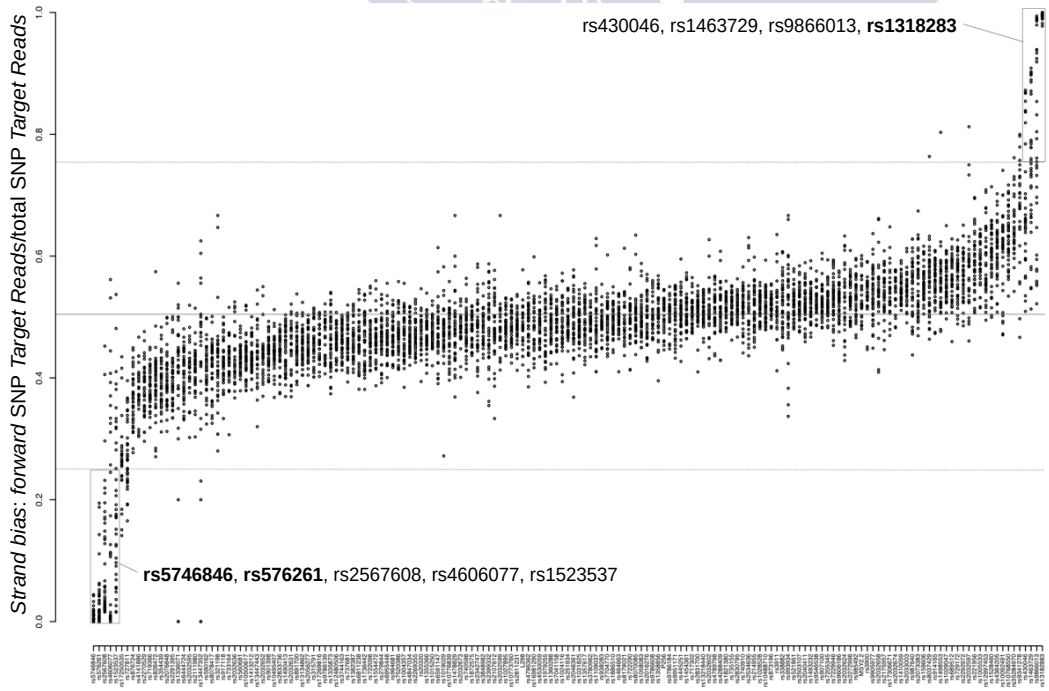


Fig. 18. Distribución de los valores de *strand bias* para 136 A-SNPs. Los SNPs con valores promedio fuera de los umbrales de >0,25 y <0,75. Aparecen destacados en negrita los que además presentan una alta tasa de *no-call*.

3.1.2.3 Concordancia de los genotipos obtenidos

La concordancia de los genotipos se evaluó en 3 niveles diferentes: (i) entre réplicas de la misma muestra en diferentes *runs* del mismo laboratorio –concordancia inter-*run*: 13 muestras en 38 análisis–; (ii) entre muestras analizadas por varios laboratorios –concordancia interlaboratorio: 6 muestras en 24 análisis–; y (iii) comparando los genotipos obtenidos para los controles de ADN Coriell con los listados en bases de datos públicas. Las tasas de concordancia se calculan sobre el total de genotipos obtenidos, para descontar la variación que produce el número de *no-calls* y el número de análisis de cada muestra.

3.1.2.3.1 Concordancia inter-run e interlaboratorio

La tasa de *no-call* de los análisis inter-*run* fue del 1,2% (70/6092), produciéndose *no-calls* en un total de 11 SNPs. El 99,8% de los genotipos asignados fueron concordantes, lo que supone una tasa de discordancia del 0,2% (13/6022). Las discordancias se presentan en los SNPs rs2399332, rs1004357, rs938283, rs1979255 y rs2032597 para 6 muestras diferentes. Las posibles causas de los *no-calls* y las discordancias se detallan en la sección 3.1.2.4.1. Además, se observó una concordancia total, en ausencia de *no-calls*, entre las réplicas de librerías incluidas en los *runs* Lab 1-B y Lab 1-C.

En los análisis de concordancia interlaboratorio se obtuvo un porcentaje de genotipos asignados concordantes del 99,7% (3751/3763), con una tasa de *no-call* del 0,8% (29/3792). Se observaron discordancias en cinco muestras, en los mismos SNPs que presentaron discordancias en los análisis inter-*run*, con una tasa del 0,3% (12/3751). No se encontraron discordancias entre los análisis de 9947A.

3.1.2.3.2 Concordancia entre Ion PGM™ y bases de datos online

Existen genotipos disponibles para 4 de los 7 controles de ADN Coriell (NA06994, NA07000, HG00403 y NA18498) en la base de datos del Proyecto 1000 Genomas Fase I. Los datos de los Y-SNPs no se encuentran compilados y 4 A-SNPs no están listados. Por lo tanto, la concordancia entre la plataforma y la base de datos del Proyecto 1000 Genomas se evaluó sobre 1056 genotipos de 132 SNPs, con una tasa de *no-call* de 2,4% (25/1056) producida en los SNPs rs1029047, rs13182883, rs13447352, rs2399332 y rs5746846. La tasa de concordancia fue del 99,5% (1026/1031) con un 0,5% de discordancias debidas a los SNPs rs8078417, rs10768550 y rs2399332 –ver Tabla 6–. No obstante, 2 discordancias se resuelven al comparar con los datos de la Fase III del Proyecto 1000 Genomas, de manera que rs2399332 permanece como el único SNP discordante y tasa de concordancia corregida se eleva al 99,9%.

La base de datos de Complete Genomics incluye 5 de los controles de ADN Coriell (los anteriores y NA07029) y datos para todos los SNPs del panel, permitiendo la comparación de 1624 genotipos. Las comparaciones se hicieron en relación a los datos de Complete Genomics basados en el *software* de ensamblaje v. 2.2.0.26, excepto para los genotipos de NA06994

para los que se utilizó la v. 2.2.0.19. Además de 30 *no-calls* obtenidos en Ion PGM™, un total de 8 genotipos presentan resultados ambiguos en Complete Genomics, resultando en un tasa global de *no-calls* de 2,3% (38/1624). No obstante el 99,7% de los genotipos obtenidos (1583/1586) resultaron concordantes. Un total de 3 genotipos discordantes se producen en los SNPs rs2032597 y rs2399332 –Tabla 6–. El SNP rs2399332 presenta discordancias entre los diferentes análisis de la misma muestra en Ion PGM™, mientras que los genotipos del Proyecto 1000 Genomas y Complete Genomics son concordantes.

En general, las comparaciones entre los genotipos asignados mediante la plataforma Ion PGM™ y los listados en las bases de datos indican una tasa alta de concordancia, que alcanza el 99,8%.

Tabla 6. Detalles del análisis de concordancia entre los genotipos obtenidos mediante Ion PGM™ para los controles de ADN Coriell y los datos recogidos de las bases de datos de Complete Genomics y el Proyecto 1000 Genomas (P1000G).

SNP	Control de ADN Coriell	Genotipo Ion PGM™	Genotipo Complete Genomics	Genotipo P1000G Fase I	Genotipo P1000G Fase III	Comentarios
Y-rs2032597	NA06994	T	C	–	–	Ver sección 3.1.2.4.1
Y-rs2032597	NA07029	T	C	–	–	Ver sección 3.1.2.4.1
rs2399332	NA18498	TT	GT	GT	GT	Ver sección 3.1.2.4.1
rs2342747	NA07000	AG	NN	AG	AG	*No-call* en Complete Genomics
rs4288409	NA18498	AC	NN	AC	AC	*No-call* en Complete Genomics
rs4847034	NA07000	GG	GN	GG	GG	*No-call* en Complete Genomics
rs4847034	NA07029	GG	GN	GG	GG	*No-call* en Complete Genomics
rs8078417	HG00403	TT	TT	CT	TT	Error en la Fase I del P1000G
rs10768550	NA18498	CT	CT	CC	CT	Error en la Fase I del P1000G

3.1.2.4 SNPs atípicos

Los SNPs atípicos fueron identificados atendiendo a los resultados de los análisis de concordancia y a los parámetros de calidad de las secuencias. Los SNPs fueron incluidos en las siguientes categorías en función del riesgo que supone el genotipado incorrecto de los mismos: (i) SNPs que presentan genotipos discordantes; (ii) SNPs que presentan *no-calls*; (iii) SNPs con parámetros de calidad de las secuencias desviados de los umbrales definidos; y (iv) SNPs que no pertenecen al resto de categorías –SNPs con buen rendimiento–.

La Fig. 19 representa los SNPs incluidos en cada una de las categorías, de manera que el 85,2% de los SNPs se enmarcan en la categoría de SNPs con buen rendimiento. Un total de 5 SNPs presentan discordancias, 9 presentan *no-calls* y 11 presentan parámetros de calidad de las secuencias desviados de los umbrales definidos, aunque produjeron genotipos totalmente concordantes. Estos SNPs fueron analizados más detalladamente, examinando en profundidad los archivos .vcf y las secuencias alineadas en el *software* IGV (Robinson *et al.* 2011).

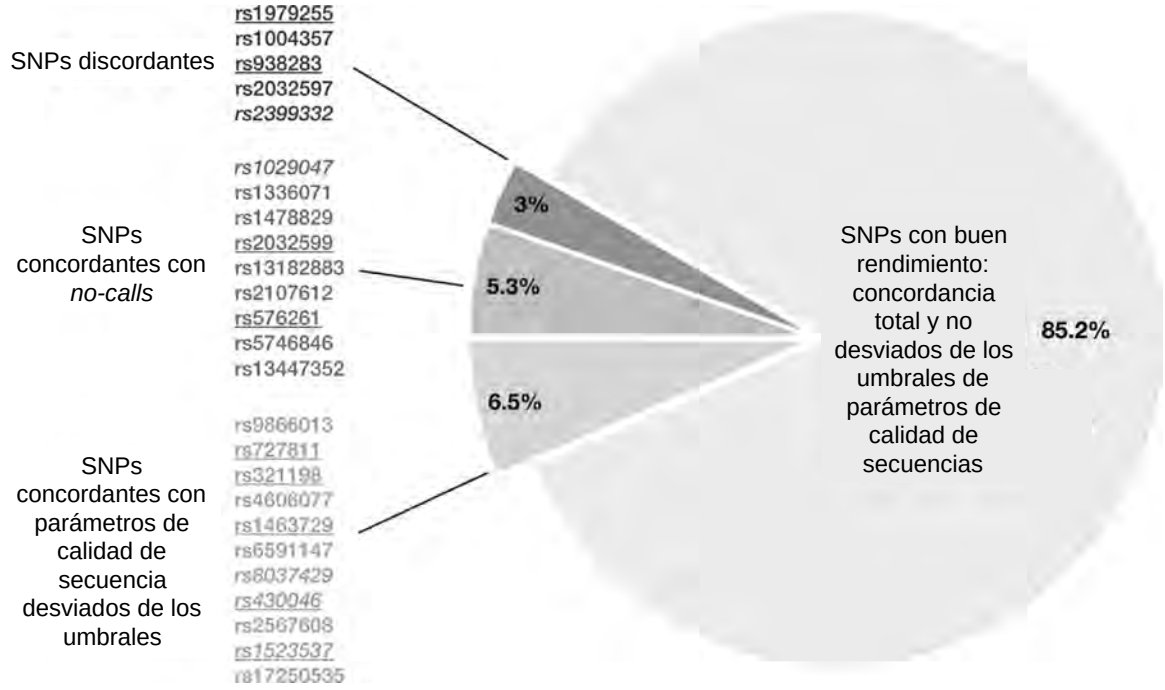


Fig. 19. Representación esquemática de las proporciones de SNPs con buen rendimiento y atípicos. Los SNPs subrayados están incluidos en versiones posteriores del panel. Los SNPs en cursiva deben ser excluidos del panel según el estudio de Børsting *et al.* (2014), conjuntamente con rs10776839, rs4530059 y rs1031825 (que no presentaron ninguna característica problemática en este estudio).

3.1.2.4.1 SNPs discordantes

Un total de 5 SNPs presentan resultados discordantes. Los SNPs rs2032597 y rs2399332 muestran diferencias genotípicas entre réplicas en más de una muestra y comparten la característica de presentar tractos homopoliméricos en las regiones adyacentes a la posición del SNP.

El Y-SNP A/C rs2032597 fue genotipado como T (no alélico) en un 20% de los análisis de muestras masculinas. Tal y como se muestra en la visualización en IGV –Fig. 20–, la base inmediatamente *upstream* de la posición del SNP es C (la base ancla –*anchor*– para el alineamiento). Cuando el genotipo del SNP presenta el alelo C, se produce un alineamiento erróneo de una alta proporción de las lecturas en ambas direcciones de secuenciación, generándose una falsa inserción C y convirtiéndose la C alélica del SNP en la base ancla. Esto produce que el tracto poli-T situado *downstream* se desplace una base hasta la posición del SNP. Como el SNP es hemicigótico, cuando el número de lecturas T supera el umbral de ARF mínimo de Genotyper (>10%) el genotipo es asignado como T en lugar de C.

El A-SNP G/T rs2399332 está situado entre un tracto poli-T. El examen de las secuencias en IGV reveló que varias lecturas G tenían una base T extra en el tracto poli-T *downstream* a la posición del SNP. Esta base T extra causa un alineamiento erróneo de las secuencias, ya que la base alélica G se considera como una inserción y la T se coloca en la posición del SNP. Como esto ocurre habitualmente a una frecuencia <10%, Genotyper asigna el genotipo GG

correctamente en la mayoría de las muestras. No obstante, se reportan genotipos GT discordantes cuando la frecuencia de las lecturas T en la posición del SNP excede el umbral de >10%. Este fenómeno explica la alta tasa de *misincorporation* del SNP –Fig. 15– y las discordancias con las bases de datos –Tabla 6–. Además, este SNP presenta unos valores de ARF en muestras heterocigotas claramente desviados de los esperados –ver Fig. 17–. Estas muestras alcanzan valores de desbalance de ARF de hasta 0,2:0,8 (20% de secuencias G) que, en la visualización de IGV, no se corresponden con alineamientos erróneos del tracto homopolimérico. Por este motivo, se revisó la secuencia contexto del SNP para buscar polimorfismos en el lugar de unión de los *primers*, que puedan afectar a la amplificación de las cadenas que portan el alelo G. Se encontraron varios SNPs en la región que abarca 30 bases *downstream* y *upstream* de los límites del amplicón, donde probablemente se encuentre la región de unión de los *primers*. En concreto, es muy probable que el SNP rs3299333 se encuentre en la región de unión del *primer forward*, ya que está localizado a 10 pb del extremo 5' del amplicón. Además, si el *primer reverse* es lo suficientemente largo el SNP rs9866331, situado a 25 pb del extremo 3' del amplicón, podría estar interfiriendo en el balance de la PCR. Dependiendo de la eficiencia de la PCR y del grado en el que los SNPs próximos afectan a la unión de los *primers*, las lecturas del alelo G de rs2399332 en heterocigotos podrían no superar el umbral de ARF del 10% del Genotyper, de manera que se reportarían como homocigotos TT, tal y como sucede en las réplicas discordantes de la muestra S5 –Tabla 6–.

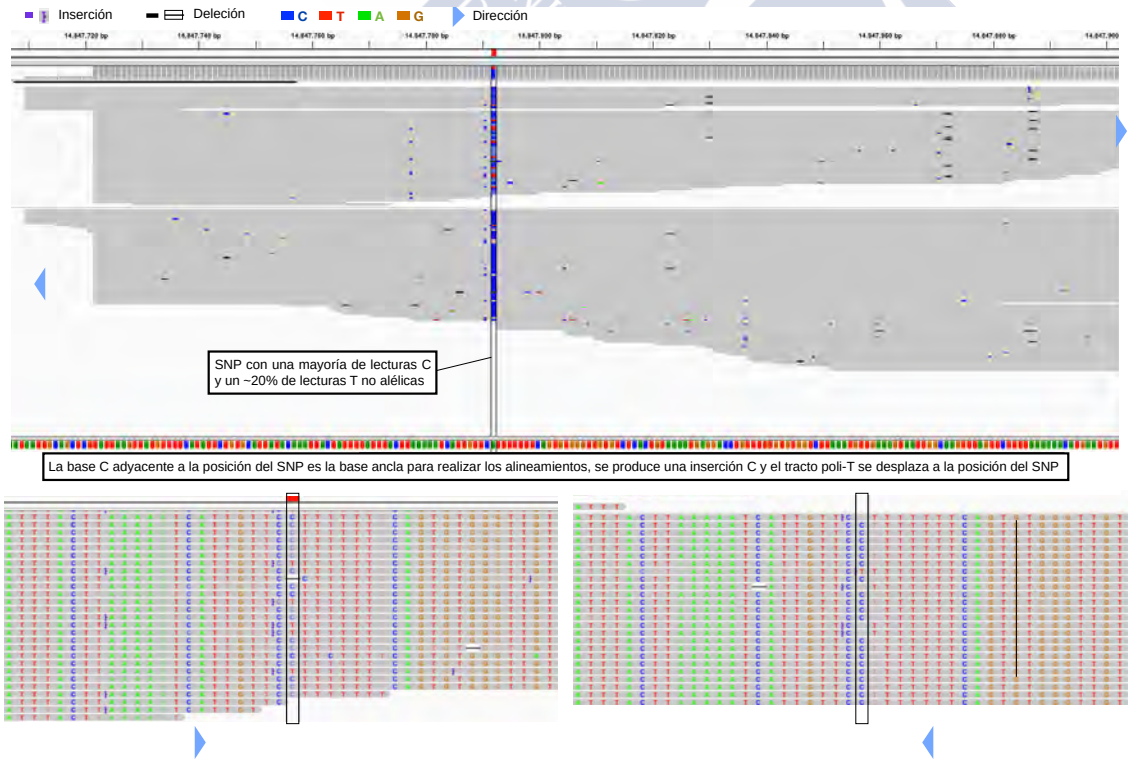


Fig. 20. Visualización del Y-SNP rs2032597 en IGV. Se muestra el alineamiento erróneo de una proporción de cadenas debido a que la base del SNP coincide con la base ancla.

Los 3 SNPs restantes presentaron genotipos discordantes en 1-2 análisis de una única muestra. En los SNPs rs1979255 y rs1004357, los heterocigotos presentan valores balanceados de ARF salvo en la muestra discordante. En el SNP rs938283 presenta valores de ARF balanceados incluso en la muestra discordante y los análisis de IGV no revelaron ninguna característica que pueda causar alineamientos o genotipos erróneos en las muestras analizadas.

3.1.2.4.2 SNPs con no-calls

En la Fig. 21 se resumen el total de SNPs con *no-calls* o *drop-outs* para 74 análisis (se excluyen las mezclas y los análisis del *run* Lab 3-B). En los análisis de concordancia, se encontraron *no-calls* en un total de 9 SNPs. En primer lugar, los SNPs rs5746846, rs576261 y rs13182883 presentan un *coverage* insuficiente en una de las cadenas –ver Fig. 18–. En estos SNPs la secuenciación se inicia en ambas cadenas pero una de ellas no llega a la posición del SNP, tal y como se muestra para rs13182883, con un 99,4% de *strand bias*, en la Fig. 22. Este fenómeno se produce en la mayoría de los SNPs que se encuentran en los extremos de la distribución que se muestra en la Fig. 18 y permanece inexplicable.

En segundo lugar, los SNPs rs13447352 y rs1336071 presentan consistentemente un número de lecturas bajo, de manera que no alcanzan ni el umbral de *coverage* mínimo total, ni el de cada una de las cadenas. En los SNPs rs2032599, rs2107612 y rs1478829 se observa el mismo efecto, pero únicamente en un análisis (rs1478829 presentó 0 lecturas).

En último lugar, otros parámetros que produjeron ocasionalmente *no-calls* al no alcanzarse los valores umbral de Genotyper fueron la calidad mínima de la variante (min_variant_score = 10) y el máximo *common signal shift* (filter_unusual_predictions = 0.3); afectando este último principalmente a rs1029047. La visualización en IGV de este SNP produjo cierta incertidumbre en cuanto a los genotipos. Tal y como se muestra en la Fig. 23, este SNP A/T se encuentra entre dos tractos poli-T y poli-A y abundantes artefactos tipo Indel, de manera que se producen errores sistemáticos en el alineamiento. Este SNP ha sido identificado como atípico por Børsting *et al.* (2014) y como discordante por Seo *et al.* (2013).

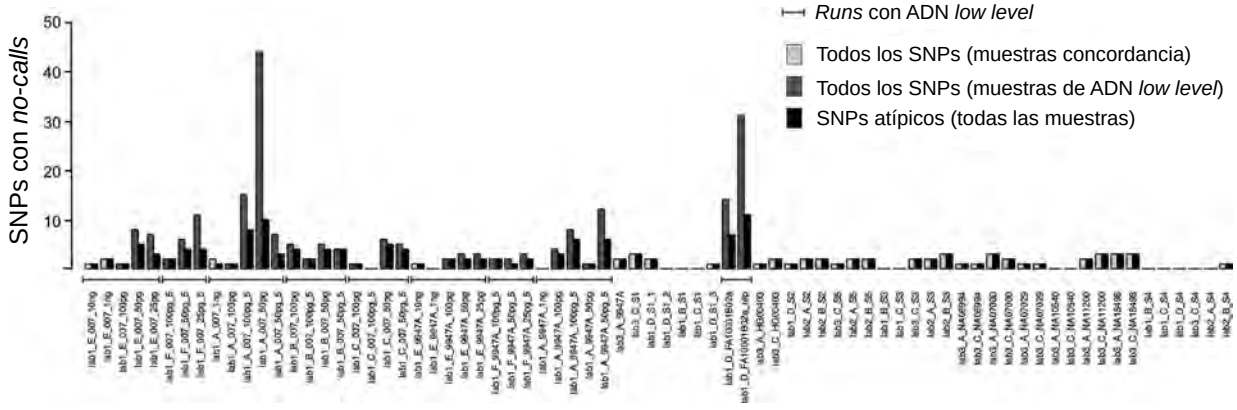


Fig. 21. Número total de SNPS que muestran *no-calls* en los de muestras de concordancia o ADN *low level*.

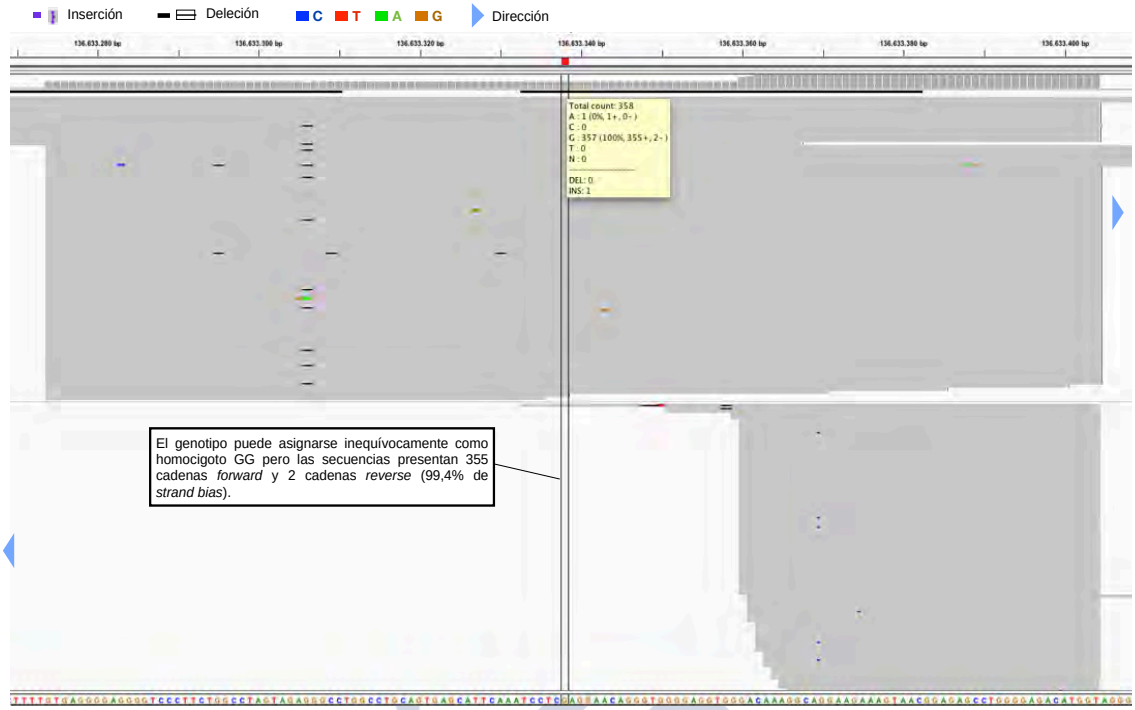


Fig. 22. Visualización en IGV de las lecturas del SNP rs13182883, que presenta un acusado *strand bias*. Las cadenas *reverse* son iniciadas pero se detienen después de ~40 pb.

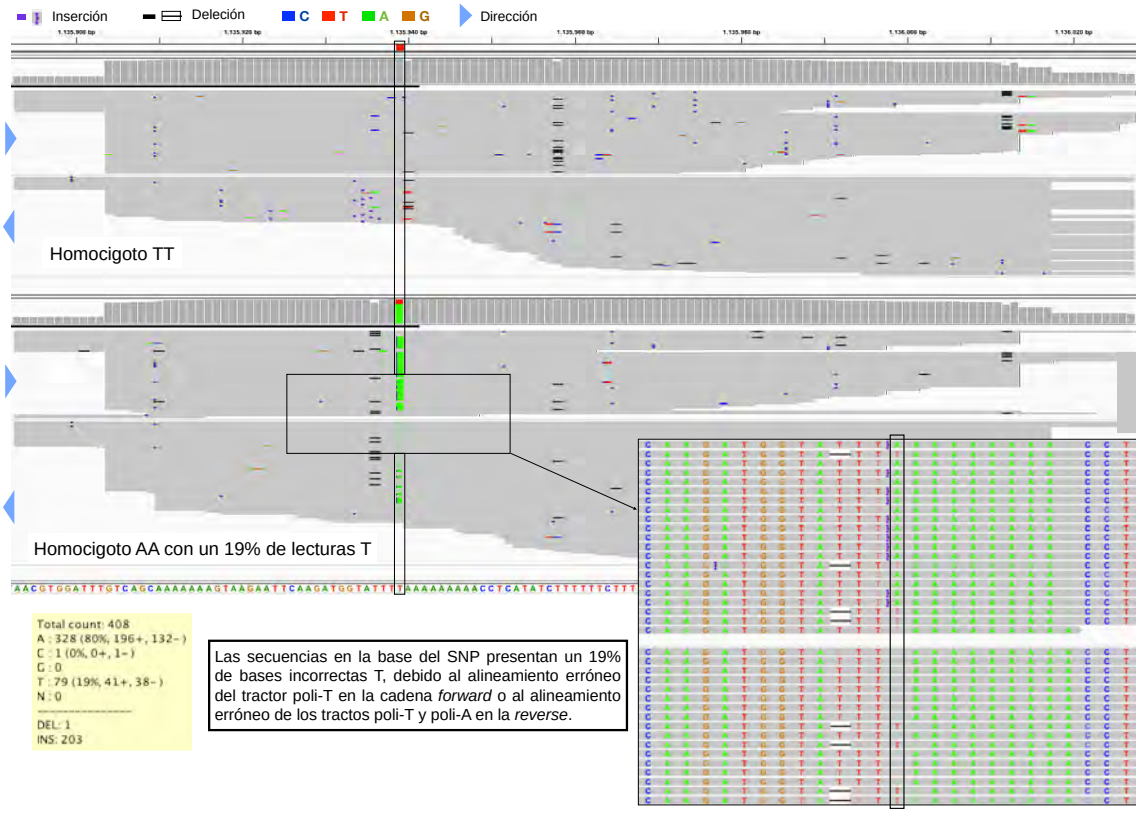


Fig. 23. Visualización en IGV de las lecturas del SNP A/T rs1029047, situado entre tractos poli-T y poli-A.

3.1.2.4.3 SNPs con parámetros desviados de los umbrales definidos

Un total de 11 SNPs fueron incluidos en la categoría de SNPs con parámetros de calidad de las secuencias desviados de los umbrales definidos –Fig. 19– a pesar de no presentar discordancias o *no-calls* en el genotipado. No obstante, los valores promedio para los diferentes parámetros evaluados indican un comportamiento atípico consistente, especialmente en cuanto a *coverage* y *strand bias*. Las secuencias de estos SNPs fueron analizadas en IGV, pero no se observó ninguna característica que evidenciara algún problema concreto de secuenciación.

Se muestra como ejemplo el SNP rs430046 –Fig. 24– que, a pesar de presentar un *strand bias* acusado y un considerable número de deleciones en la posición del SNP, produjo genotipos concordantes en todas las réplicas. En principio, no existe ninguna razón para dudar de los genotipos asignados principalmente a partir de la información de una única cadena, a pesar del incremento de *no-calls* que se observa generalmente en estos marcadores cuando no llegan al umbral de *coverage* mínimo por cadena de Genotyper.

Fig. 24. Visualización en IGV de las lecturas del SNP rs430046, que presenta un acusado *strand bias*. Las lecturas abarcan a otros 3 SNPs y presentan un artefacto tipo Indel.

3.1.2.5 Evaluación de la sensibilidad de Ion PGM™

El nivel de datos obtenidos a partir de las muestras de ADN *low level* se puede deducir de la Fig. 21, en la que se indica el número de *no-calls* de cada uno de los análisis. Para cantidades iniciales de ADN de 100, 50 y 25 pg, se aprecia un incremento en el número de *no-calls*, en comparación a los análisis con cantidad de ADN inicial óptima, aunque los *runs* Lab 1-E y Lab 1-F mantienen una buena tasa de obtención de genotipos. El SNP rs2016276 presenta un número de *no-calls* desproporcionadamente elevado en las diluciones de 100, 50 y 25 pg.

En los análisis de concordancia de muestras con cantidad inicial de ADN óptimo no se obtienen genotipos para entre 1-3 SNPs, mientras que en muestras de ADN *low level* no se obtienen genotipos para entre 8-12 SNPs, indicando un alto grado de sensibilidad del ensayo. Los análisis con 5 ciclos adicionales de amplificación de las librerías no produjeron mejores resultados en cuanto a sensibilidad.

La pérdida de la información de 8-12 SNPs produce un efecto mínimo sobre los valores de probabilidad de coincidencia al azar –RMP: *random match probability*–. En la Fig. 25 se muestra como, incluso con la pérdida de la información del 40-50% de los SNPs (incluyendo los SNPs atípicos) se obtienen valores de RPM similares a los del panel de STRs GlobalFiler.

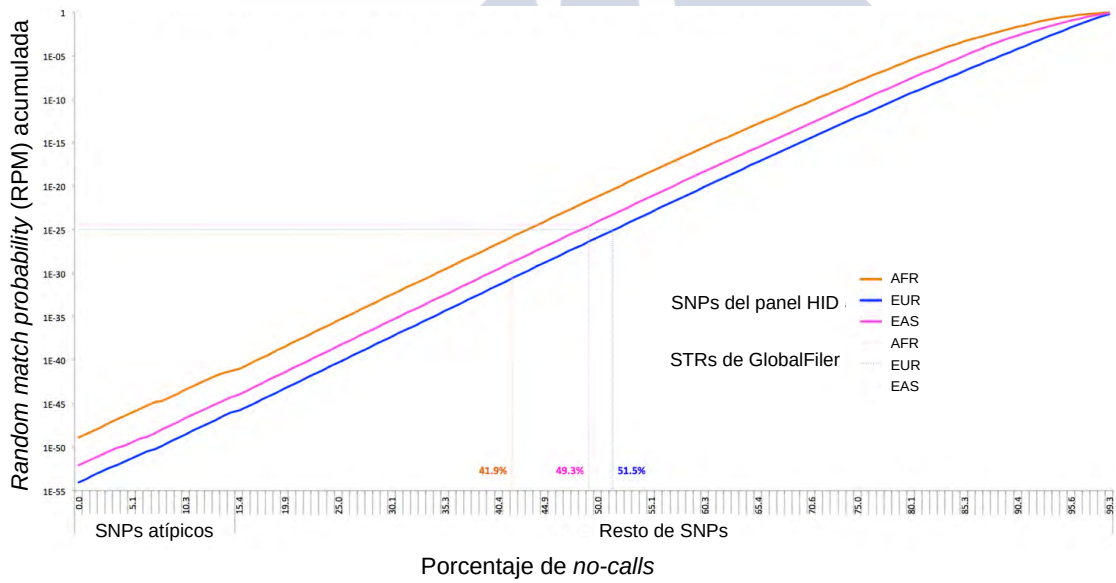


Fig. 25. Reducción de la RMP acumulada para los SNPs del panel HID en poblaciones de AFR, EUR y EAS en función del número de *no-calls*. Se muestran los puntos de intersección en los que se obtiene una RMP comparable a la de GlobalFiler para las tres poblaciones.

Los dos análisis de la muestra de ADN antiguo S7 presentaron un nivel de SNP *Target Reads* y una media de longitud de las secuencias menores que el resto de análisis, aunque se obtuvieron genotipos con un nivel relativamente alto de concordancia entre réplicas. En total,

un 75,7% (128/169) de los genotipos fueron asignados idénticamente en ambas réplicas, y uno de los análisis produjo genotipos para 23 SNPs más (un total de 89,3% de genotipos obtenidos). El mayor número de *no-calls* se produjo en el análisis en el que se realizó una reamplificación de la librería de 5 ciclos (25+5 vs. 25); además, el análisis sin reamplificación produjo más secuencias (128 vs. 72) y de mayor calidad (QUAL= 422,7 vs. 285,5).

La diferencia entre la tasa de genotipos asignados idénticamente (75,7%) y la tasa de resultados idénticos entre ambos análisis teniendo en cuenta los *no-calls* comunes (81,7%), sugiere que la pérdida de información para ciertos SNPs puede ser sistemática frente a estocástica. No obstante, se deben realizar otros ensayos con muestras de ADN degradado para comprobar esta hipótesis. A pesar de que no se dispone de genotipos de referencia para esta muestra, la heterocigosidad fue del 51% que, en comparación con el 46% esperado para la misma población, sugiere una baja tasa de *drop-out*.

La muestra degradada produjo una tasa de SNPs no genotipados mayor que la de la mayoría de los análisis de ADN *low level*. A pesar de que los datos son muy limitados, los resultados parecen indicar una buena sensibilidad de la plataforma para el análisis de muestras de ADN degradado o inhibido. La alta sensibilidad de los sistemas MPS a ADN *low level* ya ha sido señalada por otros estudios (Seo *et al.* 2013, Børsting *et al.* 2014), pero los efectos específicos del análisis de muestras de ADN altamente degradado necesitan ser evaluados en mayor profundidad, de manera que se caracterice la efectividad de las plataformas en casos de identificación de personas desaparecidas que requieran el análisis de restos óseos.

3.1.2.6 Análisis de mezclas de ADN

La detección de mezclas de ADN en los análisis de SNPs bialélicos mediante SNaPshot presenta retos, debido principalmente al desbalance entre las señales de los diferentes alelos, que son reportados mediante diferentes fluorocromos. En MPS, los SNPs presentan genotipos heterocigotos bien balanceados, favoreciendo el análisis de mezclas de ADN. La detección de una mezcla de ADN como tal es de gran importancia, ya que evita conclusiones erróneas en las investigaciones forenses. Además, se espera que las mejoras en los sistemas de análisis estadístico de mezclas permitan realizar inferencias cuando el perfil de uno de los componentes sea conocido (p. ej., la víctima de una agresión sexual).

3.1.2.6.1 Variación de las frecuencias de lectura de alelos en mezclas de ADN

La primera aproximación considerada para la detección de mezclas es la distribución de las ARF. En la Fig. 26 se muestran los valores de ARF del alelo de referencia en las cinco ratios de mezclas de ADN S5:S6 y en las muestras individuales S5 y S6. Las ARF de las muestras individuales S5 y S6 presentan distribuciones acordes con las esperadas –ver sección 3.1.2.2.2–, entre los rangos de 0-10% y 90-100% para homocigotos y 40-60% para heterocigotos, con 2-3 A-SNPs presentando valores atípicos. En contraste, en las réplicas de la mezcla 1:1 existe una dispersión muy evidente: la mayoría de las ARF se sitúan en los

rangos 10-40% y 60-90%. Así, la ausencia de balance de las ARF de los SNPs produce un efecto similar al desbalance de los picos de los STRs.

Aunque las muestras S5 y S6 tienen distribuciones de ARF similares, sus genotipos son diferentes para la mayoría de los A-SNPs. Estas diferencias afectan a las ARFs y, como consecuencia, a los genotipos reportados en las diferentes ratios de mezclas. Así, si S5 es el componente minoritario de las mezclas de ratios 1:3 y 1:9 y heterocigoto para un SNP homocigoto en S6, la frecuencia del alelo minoritario es de 1:7 (12,5%) y 1:19 (5%), respectivamente. De esta manera, en las ratios de mezcla más extremas puede no detectarse el alelo minoritario dado que no supera el umbral de >10% necesario para que Genotyper asigne un genotipo heterocigoto.

Atendiendo a las distribuciones de ARF presentadas en la Fig. 26, se puede observar como las ARF de la mezcla 9:1 en particular son muy parecidas a las de las muestras individuales; y como las ARF de la mezcla de ratio opuesto 9:1 presentan un mayor desbalance. El contraste entre las mezclas 1:9 y 9:1 explica como la detección del componente minoritario depende de los genotipos, es decir, de la combinación de homocigotos y heterocigotos y del grado de contraste de los componentes de la mezcla.

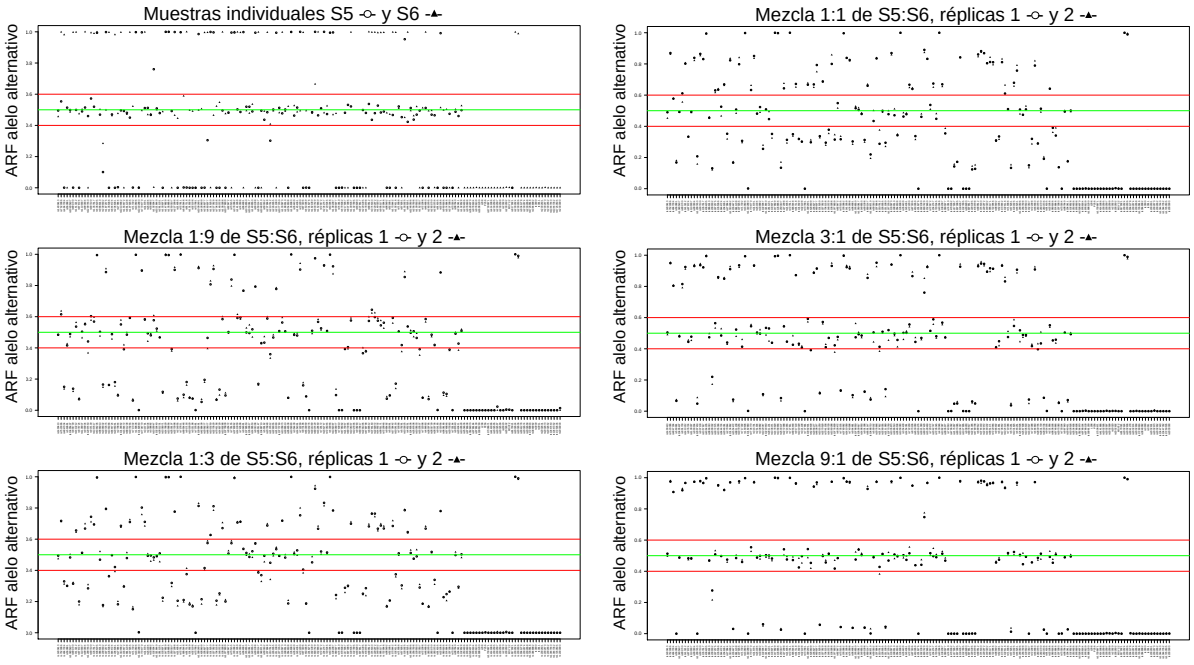


Fig. 26. ARF del alelo alternativo de los SNPs del panel (según el orden de salida de Genotyper, los Y-SNPs en la derecha de las gráficas) para las muestras individuales S5 y S6 y las dos réplicas de las diferentes ratios mezclas de ADN.

3.1.2.6.2 Cambios en los niveles observados de heterocigosidad

En una segunda aproximación para evaluar la capacidad de detección de mezclas del panel, se contaron el número de genotipos de A-SNPs reportados como heterocigotos por Genotyper. Las muestras individuales S5 y S6 presentan alrededor de un 50% de heterocigosidad que se eleva al 86,8% para la mezcla esperada calculada a través de la combinación de los genotipos obtenidos para S5 y S6 –ver Tabla 7–.

Tabla 7. Número y porcentaje de homocigotos, heterocigotos y *no-calls* para 136 A-SNPs en las muestras S5 y S6 y la mezcla esperada calculada a través de la combinación de genotipos.

	Muestras individuales				Mezcla esperada	
	S5		S6			
	N.º	%	N.º	%	N.º	%
Homocigotos	70	51,47	62	45,59	17	12,5
Heterocigotos	64	47,06	71	52,21	118	86,77
No-calls	2	1,47	3	2,21	1	0,74

Dependiendo de los genotipos de los donantes y de la ratio de mezcla, los SNPs heterocigotos pueden dividirse en balanceados (proporciones equivalentes de homocigotos opuestos o ambos componentes heterocigotos) y desbalanceados (resto de combinaciones que presentan desbalance entre los alelos). En la Tabla 8 se observa como, pese a que la heterocigosidad se eleva respecto a las muestras individuales para todas las ratios de mezclas, la proporción de heterocigotos desbalanceados se eleva desde el 60% en la mezcla de ratio 1:1 hasta el 73% en el resto de ratios más extremos.

Tabla 8. Balance de los SNPs heterocigotos de las mezclas esperadas de ratio 1:1 frente al resto de ratios. Los SNPs se dividen en balanceados, desbalanceados o indeterminados (no se puede calcular porque no se obtuvieron genotipos para alguna de las muestras individuales).

	Ratios de mezcla			
	1:1		Otros	
	N.º	%	N.º	%
Balanceados	45	38,14	31	26,27
Desbalanceados	70	59,32	84	71,19
Indeterminados	3	2,54	3	2,54

3.1.2.6.3 Efectos de los parámetros de análisis sobre la detección de mezclas

Un tercer aspecto evaluado es el efecto de los parámetros de Genotyper sobre la obtención de genotipos precisos en las mezclas de ADN. La versión de Genotyper utilizada en este trabajo permite determinar el número de lecturas analizadas para inferir los genotipos, estableciendo un valor de *downsampling*. La comparación de los análisis de las muestras de concordancia con valores de *downsampling* de 400 (valor por defecto) y 10000 o 20000

(incrementando el valor hasta un punto en el que no se produce una disminución del número de lecturas) reveló que los cambios en este parámetro no tienen efecto sobre los genotipos. Esto se debe a que la reducción de las lecturas se produce al azar, originando pequeños cambios en la proporción de lecturas de los alelos. En las mezclas de ADN, estos pequeños cambios pueden significar la diferencia entre alcanzar o no alcanzar el umbral necesario para asignar un genotipo heterocigoto, por lo que los genotipos se verían afectados. Versiones posteriores del *plugin* Genotyper presentan valores de *downsampling* por defecto de 1000000, de manera que este parámetro ya no precisa ser considerado.

Las réplicas de diferentes ratios de las mezclas de ADN (incluidas en el *run* Lab 2-C) fueron analizadas con los parámetros de bajo rigor para línea germinal *–Germline low stringency–*, incluyendo los valores por defecto *downsampling*. Además, se analizaron todas las réplicas con las mismas condiciones y el parámetro de *downsampling* modificado para incluir todas las lecturas (downsample_to_coverage=10000). De entre los 1360 genotipos posibles para todas las ratios y réplicas, el 4,4% de los genotipos asignados presentaron diferencias entre las opciones de *downsampling*, de las que el 40% se deben a *no-calls*.

Con un *downsampling* 10000 se asignan más genotipos; no obstante, no se resuelven las limitaciones a la hora de detectar alelos minoritarios que no alcanzan la proporción del 10% de minimum_allele_frequency necesaria para reportar los SNPs como heterocigotos. Comparando los genotipos reportados en los análisis *Germline low stringency* con *downsampling* en 10000 con los genotipos esperados para la mezcla (la combinación de los genotipos individuales de los componentes) se producen discordancias para al menos una de las réplicas de las diferentes ratios de mezcla en 87/136 A-SNPs (64%). En total, un 17,53% de los genotipos fueron discordantes con la mezcla esperada mientras que un 1,76% produjeron *no-calls* (una precisión de genotipado del ~80%).

Las aplicaciones clínicas de la plataforma Ion PGM™ incluyen la detección de mutaciones somáticas (p. ej. genética de cáncer) en las que la base mutada está presente en una proporción muy baja en comparación con la base de referencia. El desbalance de las ARF de las mezclas de ADN imita los patrones de las mutaciones somáticas, por lo que los parámetros optimizados para detectar estas variantes a baja frecuencia son adecuados para el análisis de mezclas. Así, se reanalizaron las mezclas aplicando parámetros *Somatic*, que establecen umbrales más bajos de minimum_allele_frequency (2% frente al 10%), calidad de las secuencias y *coverage* de cada cadena. En este tipo de parámetros, el *downsampling* por defecto es 5 veces más alto (2000). Este parámetro por defecto fue comparado con el de 10000 pero, al contrario que en los análisis de *Germline*, no se produjeron diferencias en los genotipos asignados para los A-SNPs. Además, comparando los genotipos obtenidos con los de la mezcla esperada tan solo un 1,32% (de 1360 genotipos posibles) presenta diferencias (un 1,03% *no-calls*) de manera que la precisión de genotipado se eleva al 97,65%.

3.1.2.6.4 Y-SNPs en las mezclas de ADN

El cuarto aspecto considerado del análisis de mezclas de ADN son los Y-SNPs. Para estos SNPs no se espera un segundo alelo, dado que las mezclas están compuestas por un componente femenino y otro masculino. Se debe tener en cuenta que la selección de Y-SNPs afecta a la probabilidad de encontrar un segundo alelo en mezclas con múltiples componentes masculinos y este efecto debe ser explorado.

En muestras individuales masculinas, los Y-SNPs presentan alrededor de la mitad de *coverage* que los A-SNPs. Así, cuando en una mezcla simple de dos componentes uno de ellos es masculino, el *coverage* promedio relativo de los Y-SNPs frente a los A-SNPs se correlaciona aproximadamente con la ratio de la mezcla de ADN. En la Fig. 27 se muestra como el *coverage* promedio observado de los Y-SNPs representa un 55% del *coverage* promedio de los A-SNPs en la muestra individual de S5 y decae progresivamente a través de las diferentes ratios de mezclas hasta alcanzar el 9% en la mezcla de ratio 1:9 (S5:S6), la de menor proporción de componente masculino. Así, la presencia de bajos niveles relativos de *coverage* en los Y-SNPs puede ser indicativa de la presencia de un componente minoritario masculino en una muestra forense de origen desconocido.

En cuanto a la precisión del genotipado de los Y-SNPs en las mezclas, no se observaron diferencias entre los distintos umbrales de *downsampling*. Sin embargo, la tasa de *no-call* es mayor cuando se aplican parámetros *Somatic*, especialmente en la mezcla de ratio 1:9. La reducción del umbral de minimum_allele_frequency conjuntamente con el bajo *coverage* (especialmente si el componente minoritario es el masculino) produce una disminución de los valores de calidad Phred y genera *no-calls*. No obstante, los genotipos asignados bajo parámetros *Germline* y *Somatic* son concordantes con los de la muestra individual S5.

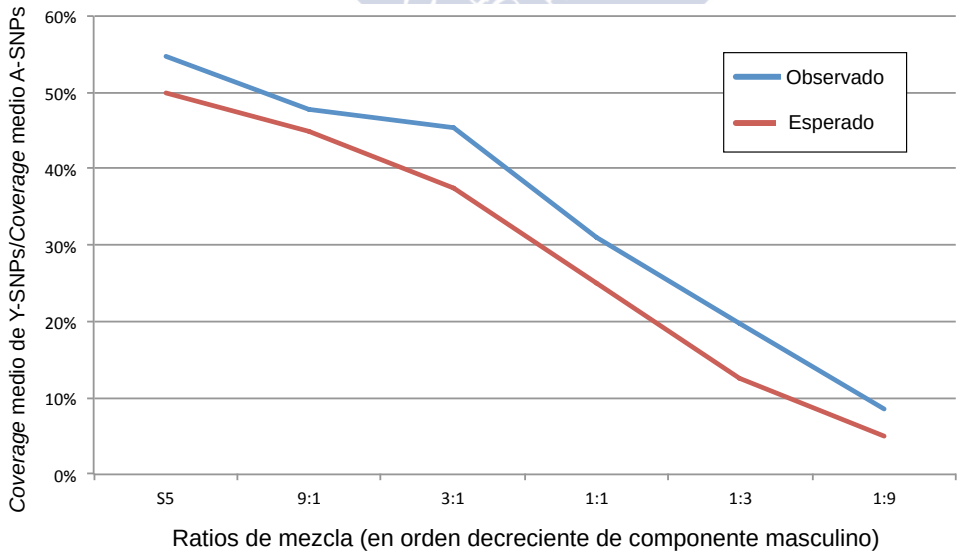


Fig. 27. Valores observados y esperados de *coverage* promedio de Y-SNP vs. *coverage* promedio de A-SNPs para la muestra individual de S5 y las diferentes ratios de mezclas de ADN.

3.1.2.6.5 Consideraciones para el análisis de mezclas de ADN

Como conclusión, las mezclas presentan patrones de ARF –Fig. 26– discernibles de los de las muestras individuales, con un alto número de SNPs heterocigotos fuera de los rangos de 40-60%, al menos para las ratios de mezcla analizadas. Conjuntamente con el incremento del nivel de heterocigotos y la reducción de la proporción del *coverage* de los Y-SNPs, se pudo indicar la presencia de una mezcla de ADN en todos los casos.

Los análisis iniciales de mezclas de dos componentes (femenino y masculino) presentados indican que los parámetros de análisis *Germline* deben ser utilizados para el análisis inicial de muestras forenses de origen desconocido. Si se presentan indicadores de la presencia de mezcla de ADN como los descritos, los datos deben ser reanalizados con parámetros *Somatic* para obtener genotipos más precisos. Aún así, se debe tener en cuenta la probabilidad de que los alelos minoritarios no sean detectados en casos en los que las mezclas de ADN presenten ratios más extremas. Los Y-SNPs deben ser analizados independientemente con parámetros *Germline* que garantizan una mayor proporción de genotipos asignados y mejores valores de calidad Phred.

3.1.2.7 Visualización detallada de las secuencias contexto con IGV

La evaluación de las regiones contexto de cada marcador en IGV permite comprobar características de las secuencias que pueden influir en los alineamientos, como Indels y tractos homopoliméricos, y localizar otros polimorfismos no incluidos en el panel que pueden ser de utilidad. Como ejemplo, en el amplicón del SNP rs430046 –Fig. 24– se incluyen 3 SNPs bien caracterizados que podrían elevar el poder de discriminación de los análisis. No obstante, en los extremos de los amplicones o cerca de tractos homopoliméricos –ver Fig. 28– se pueden producir artefactos tipo SNP provocados por alineamientos erróneos de las secuencias.

En contraste con los buenos resultados que se presentan para los SNPs, el genotipado y descubrimiento de Indels a través de las secuencias obtenidas por la plataforma presenta importantes limitaciones. Pequeños errores de secuenciación en tractos homopoliméricos tienden a producir artefactos tipo Indels que en la mayoría de los casos se corresponden con deleciones. Estos artefactos tipo Indel son fácilmente identificables ya que suelen ocurrir exclusivamente en una de las cadenas –ver Fig. 24 y Fig. 28–. Se debe tener en cuenta que futuras mejoras del *software* de alineamiento podrían evitar este efecto y permitir el correcto genotipado de polimorfismos Indel.

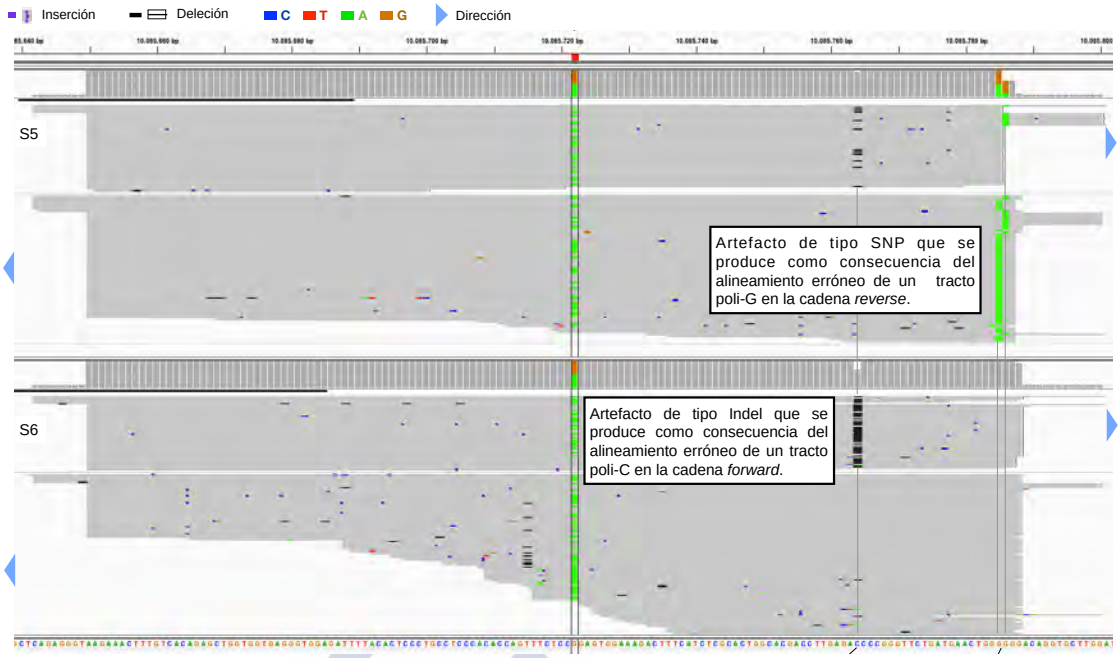


Fig. 28. Visualización en IGV de las lecturas del SNP rs1109037, en cuyo amplicón se producen artefactos de tipo SNP e Indel.

3.1.3 **Discusión**

Las evaluaciones de sensibilidad y precisión del genotipado realizadas en este trabajo proporcionan argumentos favorables a la aplicación de las tecnologías MPS en el análisis forense. Los datos obtenidos por los tres laboratorios presentan buenos valores de *coverage* y produjeron genotipos fiables para la mayoría de los SNPs del panel. No obstante, se han encontrado un total de 5 SNPs que presentan discordancias y deberían ser excluidos del panel. De entre estos 5 SNPs, rs1004357 y rs2032597 han sido excluidos de la versión revisada del panel y rs2399332 ha sido identificado como problemático en el estudio de Børsting *et al.* (2014). Los restantes dos SNPs, rs1979255 y rs938283, permanecen en el panel y deben ser evaluados en mayor profundidad. Además, el SNP rs2107612 presenta valores de ARF desbalanceados en heterocigotos y debería ser excluido del panel, conjuntamente con los 8 SNPs identificados en el estudio de Børsting *et al.* (2014). En último lugar, el SNP rs1029047 produjo genotipos inconsistentes en todos los estudios de MPS que lo incluyen (Seo *et al.* 2013, Børsting *et al.* 2014). Este SNP presenta características en la secuencia contexto –Fig. 23– que afectan al alineamiento de las secuencias y, por consiguiente, a la fiabilidad de los genotipos; aunque los análisis mediante SNaPshot (Sánchez *et al.* 2006) no se ven afectados. Así, la implementación de los SNPs en MPS requiere un escrutinio detallado de las secuencias contexto de los SNPs.

La estimación del número óptimo de muestras que se pueden analizar simultáneamente en cada tipo de chip y sus correspondientes versiones supuso un importante reto, tanto a la hora de armonizar los *runs* de los diferentes laboratorios como a la hora de asegurar un

coverage mínimo que permitiera evaluar la sensibilidad forense de la plataforma. Dado que los análisis de ADN *low level* acentúan las diferencias de *coverage* de los SNPs del panel, la optimización del número de muestras requerirá un especial cuidado si se adapta esta tecnología a la rutina forense. No obstante, los chips de secuenciación alcanzan niveles muy altos de lecturas y los usuarios pueden optar por una metodología conservadora, cargando un número de muestras menor al que sugieren las guías. Además, existe cierto consenso en que umbrales de *coverage* mínimo de 15-20x son suficientes para asegurar la fiabilidad de los genotipos obtenidos mediante MPS, al menos para muestras de referencia (Bentley *et al.* 2008, Quail *et al.* 2012, Daniel *et al.* 2015).

Aunque el *software* Torrent Suite™ proporciona parámetros de calidad de las secuencias en los archivos de salida, permite poco margen para establecer umbrales para dichos parámetros. La aplicación de umbrales más rigurosos puede evitar el genotipado incorrecto de SNPs atípicos. Dado que la detección de mezclas con marcadores bialélicos presenta limitaciones en comparación con el uso de marcadores multialélicos, es importante que se pueda valorar correctamente el desbalance de los patrones de ARF. En este sentido, y de acuerdo con las conclusiones del estudio de Børsting *et al.* (2014), el *software* todavía necesita importantes mejoras para su aplicación en la rutina forense, aunque se están produciendo continuas actualizaciones y revisiones. En particular, aunque los parámetros recomendados para aplicaciones forenses son los de *Germline*, es necesario aplicar parámetros *Somatic* para un análisis adecuado de las mezclas de ADN,

3.2 VALIDACIÓN DEL PANEL QIAGEN SNP-ID

En este trabajo se realizó una validación interlaboratorio del panel Qiagen SNP-ID en la plataforma Ion PGM™. Siguiendo un esquema de validación sencillo, se evaluó la calidad de las secuencias obtenidas, la precisión del genotipado, la sensibilidad forense a ADN *low level* y degradado y la capacidad de detección de mezclas de ADN. Los resultados se encuentran publicados en la siguiente referencia:

> **de la Puente M**, Phillips C, Santos C, Fondevila M, Carracedo Á, Lareu MV (2017). *"Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing."* Forensic Sci Int Genet 28: 35-43.

3.2.1 **Material y métodos**

3.2.1.1 El panel Qiagen SNP-ID

El diseño de *primers* del panel Qiagen SNP-ID utiliza una aproximación de PCR solapante –*tiling*– Grandell *et al.* (2016) mediante la que se pretende evitar el efecto de los polimorfismos de las regiones flanqueantes sobre el *annealing* de los *primers*. Se diseñaron parejas de *primers* dobles para cada uno de los SNPs, de manera que existen 4 posibles amplicones para cada marcador, siendo el tamaño promedio de los amplicones más cortos de 134 pb. Para el diseño la PCR de captura *multiplex*, se balancearon las temperaturas de *melting* de los *primers* y se seleccionaron aquellos menos estables en el extremo 3' (secuencias ricas en AT), a fin de minimizar posibles amplificaciones inespecíficas.

La PCR de captura combina 140 SNPs autosómicos de dos paneles de identificación previamente publicados: los 52 SNPs del panel SNP*for*ID (Sánchez *et al.* 2006) y los 92 de *Kiddlab* (Pakstis *et al.* 2010), con 4 SNPs en común entre los dos paneles.

3.2.1.2 Muestras de ADN

La validación del panel se realizó siguiendo una estructura lo más sencilla posible. Se diseñó un único *run* en el que se incluyeron 16 librerías: (i) 5 controles ADN Coriell para evaluar la concordancia de los genotipos obtenidos; (ii) 3 diluciones seriadas de uno de los controles de ADN Coriell; (iii) dos réplicas de mezclas de ADN de controles Coriell de ratios 1:1, 1:3 y 1:9; (iv) dos réplicas de ADN extraído de un hueso degradado.

Para evaluar la concordancia de los genotipos obtenidos se utilizaron 5 controles de ADN Coriell diluidos a 1 ng/µL: NA10540, NA18498, NA06994, NA11200 y HG00403. Los genotipos se compararon con los obtenidos para todos los controles mediante SNaPshot para los 52 marcadores de SNP*for*ID (Sánchez *et al.* 2006) y con los listados en la base de datos del Proyecto 1000 Genomas (The Genomes Project Consortium 2015), generados mediante la tecnología Illumina HiSeq, para NA18498, NA06994 y HG00403.

Para evaluar la sensibilidad del panel se prepararon diluciones seriadas de NA11200 a 0.5 ng/µL, 0.25 ng/µL y 0.125 ng/µL.

Para evaluar la capacidad de detección de mezclas de ADN se prepararon una única vez mezclas de volumen de los controles NA18498 y HG00403 a 1 ng/µL en las siguientes proporciones: 1 a 1; 1 a 3; y 1 a 9. Cada mezcla se analizó por duplicado.

Para evaluar la capacidad de análisis de muestras de ADN degradado, se analizó por duplicado una única muestra de ADN degradado extraída de restos esqueléticos (fémur). Los análisis previos con Quantifiler® Duo no indicaron indicios de inhibición y los resultados de cuantificación fueron de 0,017 ng/µL de ADN. El análisis de los SNPs del panel SNP*for*ID mediante SNaPshot produjo un ~90% del perfil.

3.2.1.3 Preparación de las librerías y secuenciación

Las librerías de ADN fueron preparadas mediante el kit GeneRead™ DNAseq Targeted Panels v2 (Qiagen, Hilden, Alemania) y el kit de *primers* Qiagen SNP-ID multiplex PCR. La PCR de captura fue realizada siguiendo las indicaciones del fabricante[24], a excepción de la cantidad inicial de ADN que se redujo 20 veces. Por tanto, las reacciones de PCR utilizaron 1 µL de las muestras descritas en la sección 3.2.1.2; a excepción de la muestra de ADN degradado y el control negativo, para las que se incluyó el volumen inicial máximo de 8 µL. Las condiciones de PCR incluyeron 20 ciclos de amplificación de 4 min a 60°C.

Después de la purificación de los productos de PCR, se realizó una evaluación simple de la eficiencia de amplificación mediante el sistema Agilent High Sensitivity D1000 ScreenTape (Agilent Technologies, Santa Clara, EEUU). En este punto se comprobó que el control negativo estaba libre de ADN por lo que no fue incorporado en análisis posteriores. Para maximizar el rendimiento del chip de secuenciación, durante la preparación de las librerías se incluyeron *barcodes* para individualizar cada muestra combinando los kits Qiagen GeneRead Adapter L Set 12-plex (*barcodes* 1-12) y TFS Ion Xpress™ Barcode Adapters (*barcodes* 13-16). Después de purificar las librerías, se utilizó el sistema Agilent High Sensitivity D1000 ScreenTape para la cuantificarlas. Cuando fue necesario, las librerías se diluyeron para combinarlas en un *pool* equimolar a 25 pM. La secuenciación se realizó en un chip Ion 316™ v2, preparado mediante el kit Ion PGM™ Hi-Q™ Chef.

3.2.1.4 Análisis de datos

Las secuencias alineadas obtenidas en la plataforma Ion PGM™ (archivos .bam y .bai) se analizaron de 3 maneras diferentes: (i) usando el *software* Torrent Suite™ v. 5.0.2 y el *plugin* HID_SNP_Genotyper versión 4.2 –Genotyper–; (ii) usando el *software* Biomedical Genomics Workbench v. 2.5.1 (CLC Bio, Qiagen) y aplicando un *workflow* personalizado previamente descrito (Grandell *et al.* 2016) –Workbench–; y (iii) mediante la visualización detallada de las secuencias alineadas en *software* IGV v. 2.3.40 (Robinson *et al.* 2011).

[24] Qiagen: GeneRead™ DNAseq Targeted Panels V2 Handbook. June (2014).

Los parámetros por defecto de Workbench (aplicados a las muestras de concordancia) incluyen una frecuencia mínima del alelo de 0,2 para la asignación de un genotipo heterocigoto y ningún umbral de *coverage* mínimo. Los parámetros por defecto de Genotyper (aplicados a todas las muestras) incluyen una frecuencia mínima del alelo de 0,1 para la asignación de un genotipo heterocigoto y un *coverage* mínimo de 6x en la posición del SNP. Para el reanálisis de las mezclas de ADN con Genotyper, se mantienen todos los parámetros por defecto menos la frecuencia mínima del alelo, que se establece en 0,02.

Las secuencias brutas fueron alineadas con el genoma de referencia humano GRCh37/hg19 y los SNPs en las regiones de interés fueron identificados de acuerdo con dbSNP build 144, usando UCSC Genome Browser (Kent *et al.* 2002).

A partir de los archivos de Genotyper se calcularon una serie de parámetros de calidad de las secuencias utilizando hojas de cálculo tipo Excel, que permitieron identificar SNPs atípicos. El sesgo de lectura de las cadenas –*strand bias*– se calculó como porcentaje de *coverage forward*/*coverage* total. El sesgo de lectura de cadenas para cada alelo –*strand bias per allele*– se calculó como porcentaje de lecturas *forward* del alelo/ (lecturas *forward* del alelo + lecturas *reverse* del alelo). La frecuencia de lecturas de los alelos –ARF: *allele read frequency*– se calculó como porcentaje de lecturas del alelo/*coverage* total del SNP. La tasa de incorporación errónea de nucleótidos –*misincorporation*– fue calculada como porcentaje de lecturas no alélicas/*coverage* total del SNP. Se debe tener en cuenta que los umbrales de *strand bias*, *strand bias per allele*, ARF y *misincorporation* aplicados para la identificación de SNPs atípicos no se corresponden con los umbrales aplicados por Genotyper o Workbench para la asignación de genotipos.

Para detectar haplotipos en las siguientes parejas de SNPs: rs10768550-rs10500617 (coordenadas GRCh37 11:5098714-5099393) y rs9606186-rs5746846 (coordenadas 22:19920359-19920646) se obtuvieron datos del Proyecto 1000 Genomas Fase III mediante la herramienta *online* Data Slicer[25] y se procesaron en Excel.

3.2.2 **Resultados**

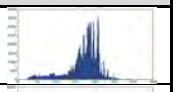
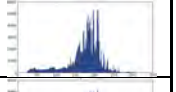
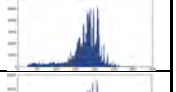
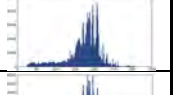
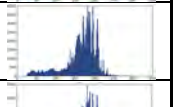
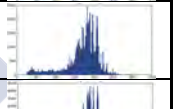
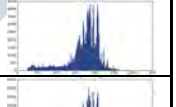
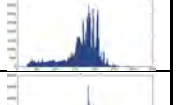
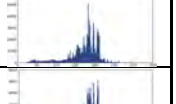
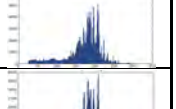
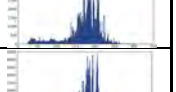
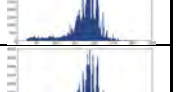
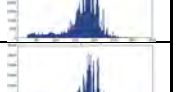
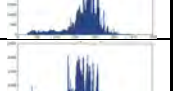
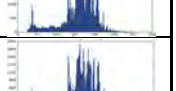
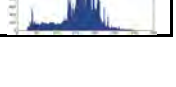
3.2.2.1 Rendimiento de la preparación de la librería y secuenciación

En la Tabla 9 se muestran los resultados de cuantificación de las PCR de amplificación y de las librerías mediante el sistema Agilent High Sensitivity D1000 ScreenTape. Los valores de cuantificación de PCR son menores en las diluciones seriadas y las dos réplicas de ADN degrado, correspondiéndose con una menor cantidad inicial de ADN. Los valores de cuantificación de librerías muestran más variabilidad que los de PCR, posiblemente debido a la acumulación de diferencias en la PCR inicial, la amplificación de la librería y los múltiples pasos de purificación.

[25] http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice

No obstante, los resultados de cuantificación de las librerías permitieron formar un *pool* equimolar a 25 pM. La réplica A de la mezcla de ADN 1:9 fue la única que mostró un pico extra de 77 pb en el electroferograma obtenido, indicando la presencia de dímeros de adaptadores –ver Fig. 29–.

Tabla 9. Para cada muestra se recogen los resultados de cuantificación del producto de PCR y de las librerías después de la purificación, el *coverage* promedio de los SNPs, la longitud promedio de las lecturas y un histograma de longitud de las mismas (número de lecturas vs. longitud).
*esta muestra presenta un pico en aproximadamente 77 pb que no se corresponde con la librería.

Set de muestras	Muestra	Cuantificación PCR (pg/µL)	Cuantificación librería (pg/µL)	*Coverage* promedio de los SNPs	Longitud promedio de las lecturas	Histograma de la longitud de las lecturas
Muestras de concordancia (1 ng)	NA10540	152	235	944,81	168	
	NA18498	145,4	134	1424,46	171	
	NA06994	117	182	1398,49	171	
	NA11200	177	239	1215,57	171	
	HG00403	117	117	1118,13	163	
Diluciones seriadas de NA11200	0,5 ng	86,7	66,9	650,7	168	
	0,25 ng	68,7	53,9	1193,69	169	
	0,125 ng	25,3	22,5	966,19	163	
Mezclas de ADN de NA18498 y HG00403 (1 ng)	1:1 - réplica A	168	31,2	822,43	167	
	1:1 - réplica B	109,1	248	1316,03	170	
	1:3 - réplica A	114	90,8*	929,99	169	
	1:3 - réplica B	147,6	192	1109,71	173	
	1:9 - réplica A	184	175	1035,87	170	
	1:9 - réplica B	164,6	184	805,44	170	
ADN extraído de fémur (0,136 ng)	Réplica 1	20,8	15,7	704,29	158	
	Réplica 2	19,8	15,8	524,54	158	

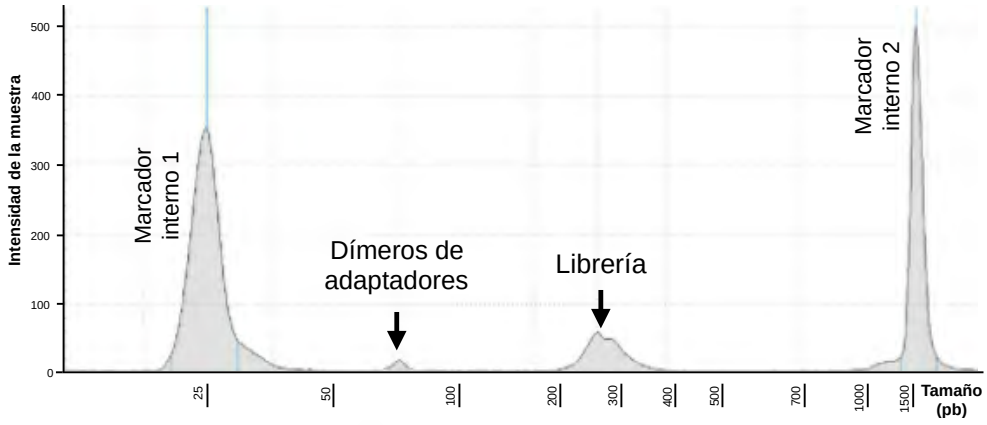


Fig. 29. Electroferograma de la librería purificada de la réplica A de la mezcla de ADN de ratio 1:9, analizada mediante el sistema Agilent High Sensitivity D1000 ScreenTape.

El informe del *run* indica una densidad media de carga del chip del 69% – ver Fig. 30A–, con un 100% de enriquecimiento en las Ion Sphere™ Particles (ISPs) –Fig. 30B–. Los datos de filtrado de lecturas indican un nivel relativamente alto de policlonalidad (secuencias heterogéneas en las ISPs) del 27% –Fig. 30B–, un valor que se encuentra dentro de los límites del rango de valores habitual de 10-30%, pero cercano al límite superior. El histograma de lecturas del *run* –ver Fig. 30C– muestra un mayor número de lecturas para las longitudes que se corresponden con los rangos de los amplicones de PCR, empezando en ~160 pb y extendiéndose por encima de los 200 pb. No obstante, los histogramas individuales de cada librería, mostrados en la Tabla 9, indican diferencias entre las muestras. En concreto, las dos réplicas de ADN degradado tienen un menor número de lecturas y éstas aparecen distribuidas en un rango de longitudes más amplio y sesgado hacia longitudes más cortas.

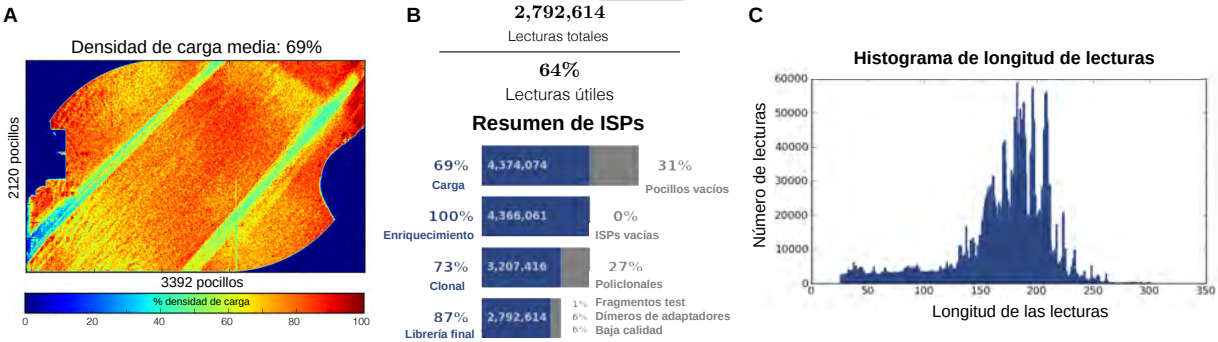


Fig. 30. A) Rendimiento de carga del chip. B) Resumen de las características de las ISPs. C) Histograma de longitud de las lecturas del *run*.

3.2.2.2 Visualización en IGV de las lecturas y potenciales problemas de genotipado

Las secuencias generadas en el análisis de las muestras de concordancia se visualizaron en IGV para detectar características como: *strand bias*, estructuras inusuales de las lecturas que puedan afectar a la obtención de genotipos, artefactos tipo Indel derivados del alineamiento erróneo de las cadenas e incorporaciones erróneas en la posición de los SNPs del panel que no alcanzan la frecuencia mínima necesaria para ser asignadas como genotipo.

Aunque ciertas características de las secuencias no afectan a la fiabilidad de genotipado del SNP incluido en el panel, pueden influir en otros polimorfismos informativos que se encuentran en los amplicones, si estos se incorporaran en posteriores análisis. Centrándose en los SNPs incluidos en el panel, un 25% de los marcadores presentan al menos una característica que puede afectar al genotipado. Estas características se pueden dividir en dos categorías: (i) estructuras de secuencias inusuales (lecturas que no cubren uniformemente el amplicón del SNP), principalmente casos en los que las lecturas son incompletas o una de las cadenas presenta un número de lecturas desproporcionadamente bajo con respecto a la otra; y (ii) alineamientos erróneos próximos a la posición del SNP.

En la primera categoría se encuentran un total de 25 SNPs que mostraron un *strand bias* acusado o *coverage* bajo: rs891700; rs9866013; rs13182883; rs7704770; rs2272998; rs727811; rs321198; rs737681; rs4288409; rs4606077; rs1015250; rs2270529; rs1360288; rs430046; rs8070085; rs8078417; rs1024116; rs576261; rs12480506; rs2567608; rs1005533; rs1523537; rs722098; rs2831700 y rs5746846.

A excepción de rs4606077, rs2270529 y rs1523537 (que se encuentran incluidos a su vez en la segunda categoría), los SNPs de esta categoría no revelaron ninguna otra característica que pueda influir en la fiabilidad de los genotipos obtenidos. Por tanto, la mayoría de ellos presentan simplemente un riesgo más alto de *no-calls* (p. ej. en casos en los que se analiza ADN *low level*) derivado de la reducción del número de lecturas; pero no un riesgo de imprecisión del genotipado.

Un ejemplo de SNP con acusado *strand bias* es rs430046 –ver Fig. 31– que, a pesar de presentar un buen balance de lecturas de los alelos C y T en heterocigotos, muestra un número de lecturas muy bajo en la cadena *reverse*. La visualización de las secuencias obtenidas en IGV indica que el SNP próximo rs381840 A/G, situado en un tracto complejo de nucleótidos G y A, parece afectar a la correcta extensión de las cadenas *reverse*. En general, se ha observado que las cadenas incompletas que presentan los SNPs de esta categoría tienden a truncarse en tractos homopoliméricos o regiones repetitivas.

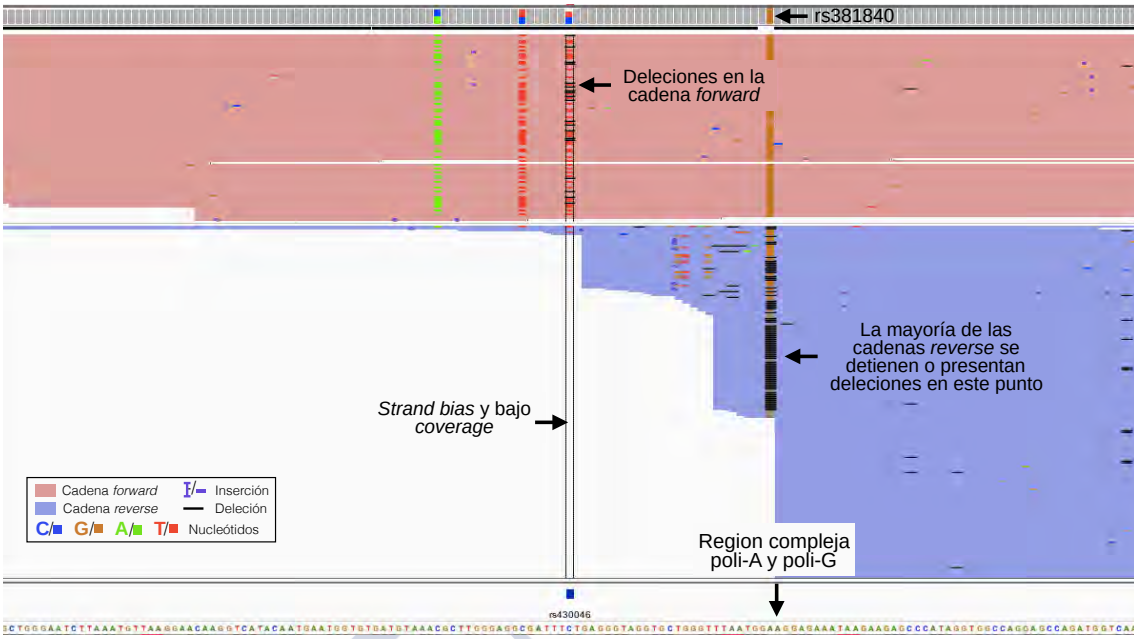


Fig. 31. Visualización de una muestra heterocigota para rs430046 (C/T).

La segunda categoría incluye SNPs que presentan alineamientos erróneos, incorporaciones erróneas de nucleótidos o artefactos tipo Indel en la posición del SNP, y *strand bias per allele*. En total, 12 SNPs del panel presentan un riesgo alto de producir genotipos discordantes.

Los SNPs rs4847034, rs1554472, rs4796362, rs1004357, rs733164 y rs1821380 presentan incorporaciones de nucleótidos no alélicos en la posición del SNP en una de las cadenas –Fig. 32, Fig. 33, Fig. 34, Fig. 35, Fig. 36 y Fig. 37, respectivamente–. Los primeros 4 SNPs presentan incorporaciones erróneas no alélicas T causadas por el desplazamiento de un tracto poli-T adyacente en el que se lee una T extra. Cuando esto ocurre, el nucleótido del SNP es alineado como una inserción. En rs1821380 y rs733164, la causa de la aparición de lecturas no alélicas T y C, respectivamente, no se pudo dilucidar mediante la visualización de las secuencias en IGV. El SNP rs2270529 es un SNP T/C embebido en la estructura TT[T/C]GTT –ver Fig. 38–, varias lecturas G aparecen en la posición del SNP en las cadenas *reverse*, seguidas de una inserción T que desplaza el tracto poli-T. También en sentido *reverse*, una proporción de cadenas muestran una inserción C situada antes de la posición del SNP cuando este presenta una lectura T. Estas inserciones C ocurren a consecuencia de la lectura de una T extra en el tracto TT próximo al SNP, aunque este fenómeno ocurre, al igual que las lecturas G en la posición del SNP, a muy baja frecuencia.

Para todos estos SNPs, la proporción de incorporaciones erróneas de nucleótidos no alélicos puede llegar a alcanzar los valores umbrales del *software* para ser asignados como genotipos. En este caso, la corrección manual de los genotipos es sencilla, ya que se pueden identificar fácilmente al no ser alelos reconocidos del SNP.

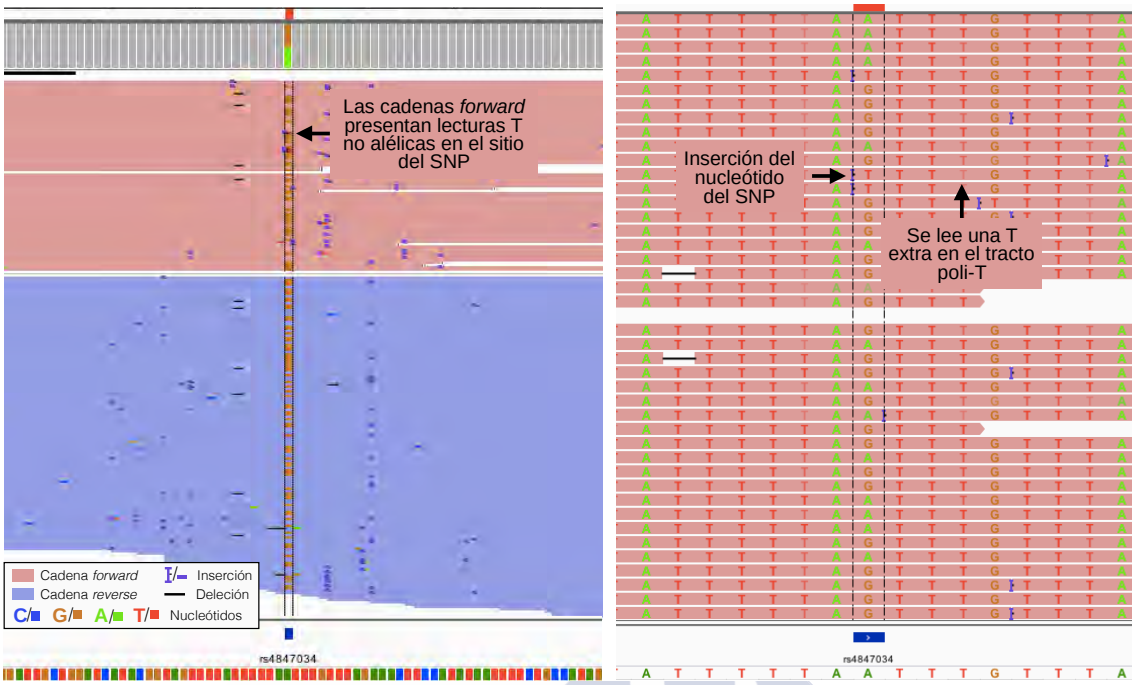


Fig. 32. Visualización en IGV de una muestra heterocigota para el SNP A/G rs4847034.

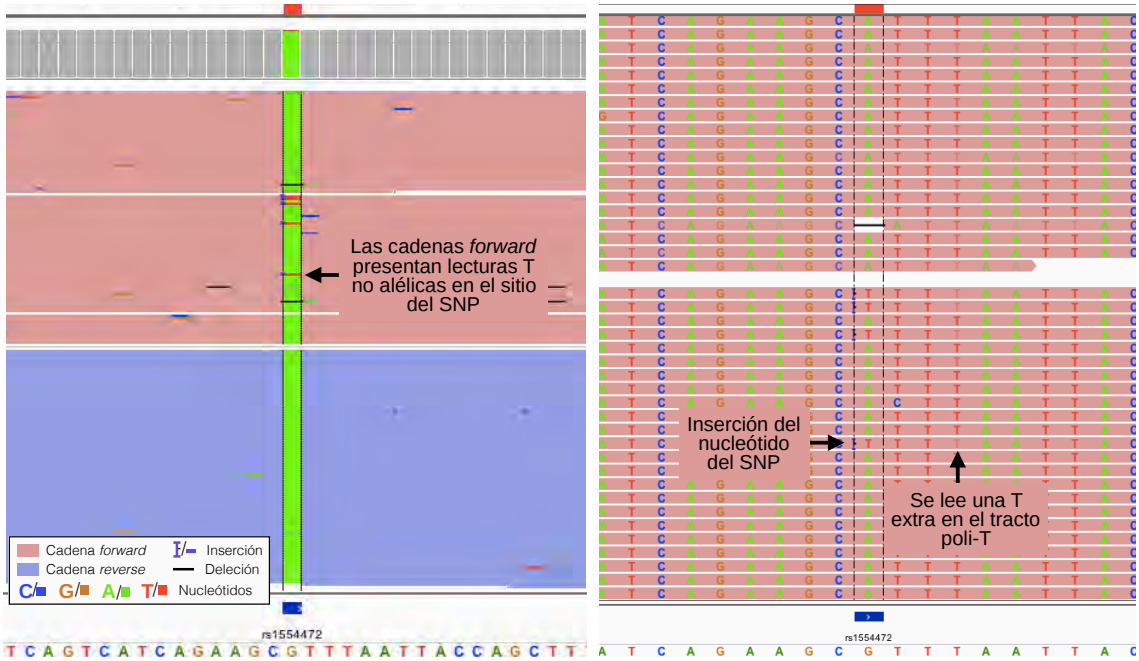


Fig. 33. Visualización en IGV de una muestra homocigota AA para el SNP G/A rs1554472.

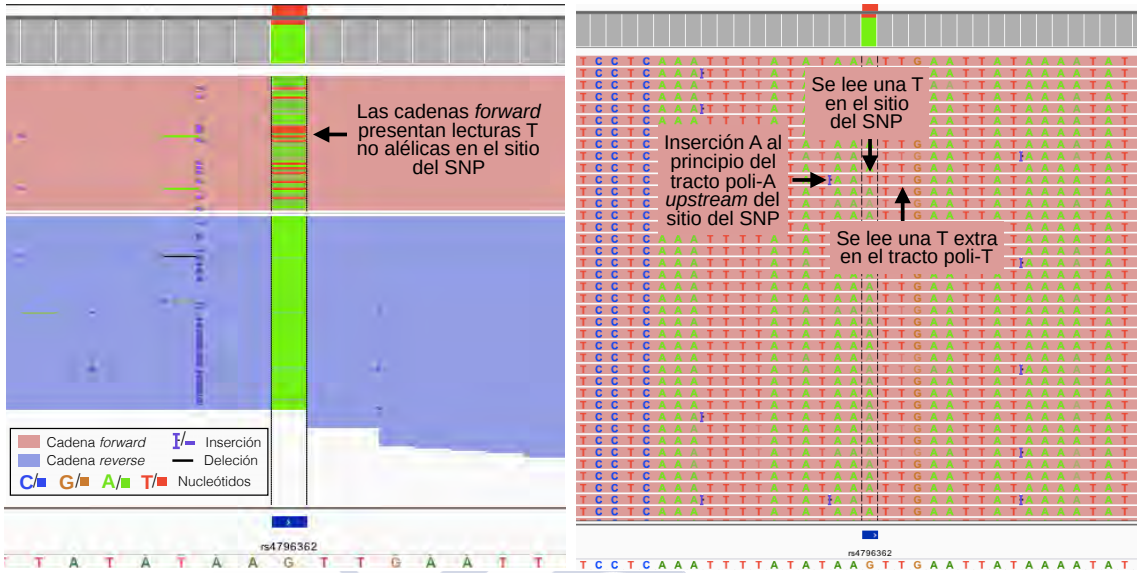


Fig. 34. Visualización en IGV de una muestra homocigota AA para el SNP G/A rs4796362.

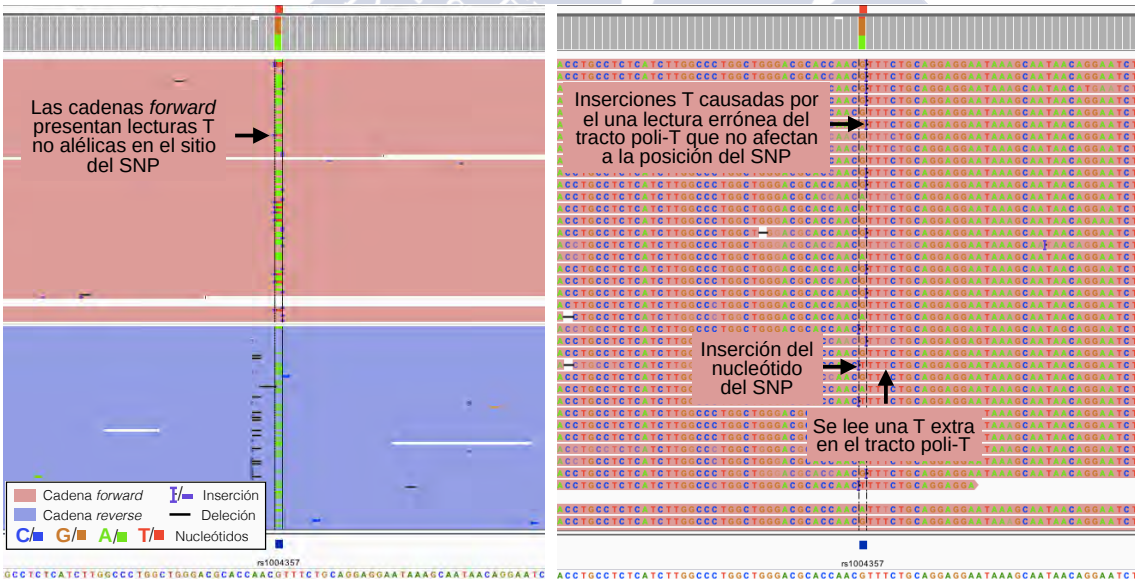


Fig. 35. Visualización en IGV de una muestra heterocigota para el SNP G/A rs1004357.

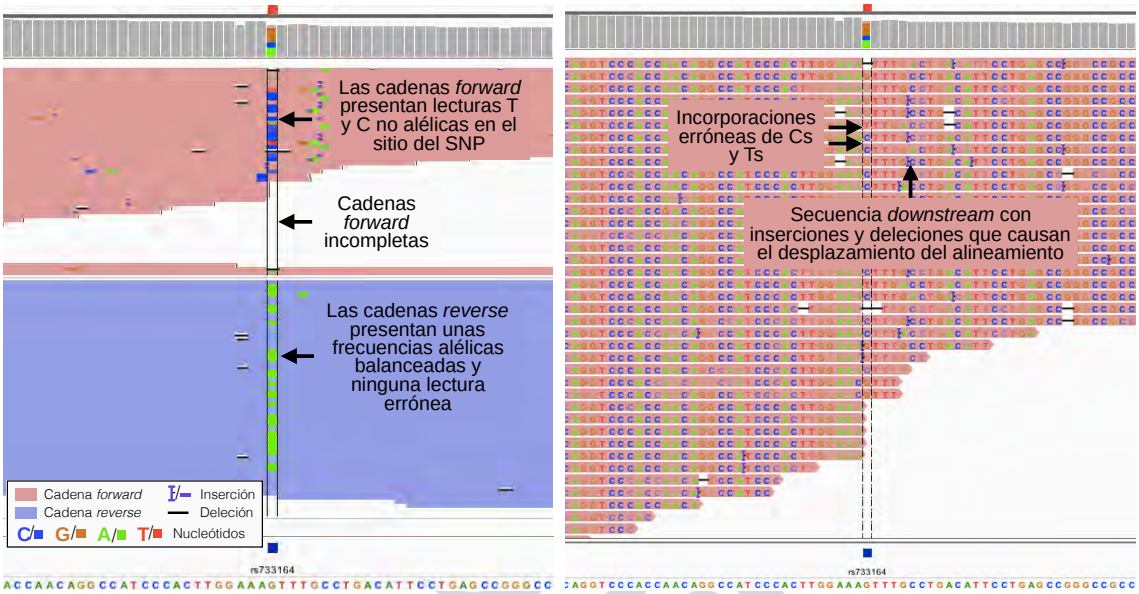


Fig. 36. Visualización en IGV de una muestra heterocigota para el SNP G/A rs4733164.

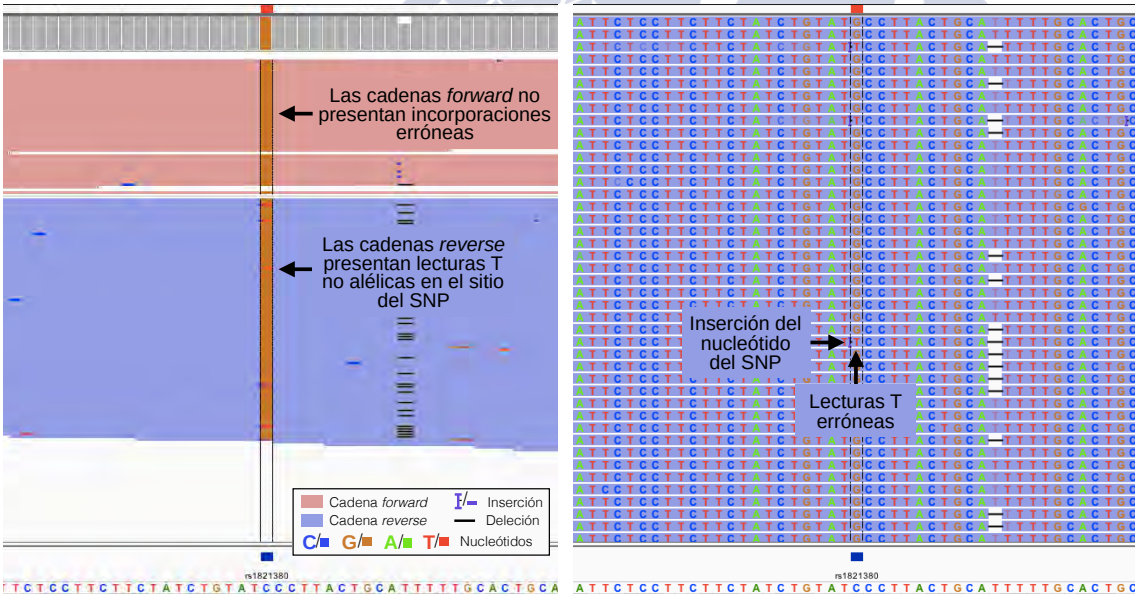


Fig. 37. Visualización en IGV de una muestra homocigota CC para el SNP C/G rs1821380.

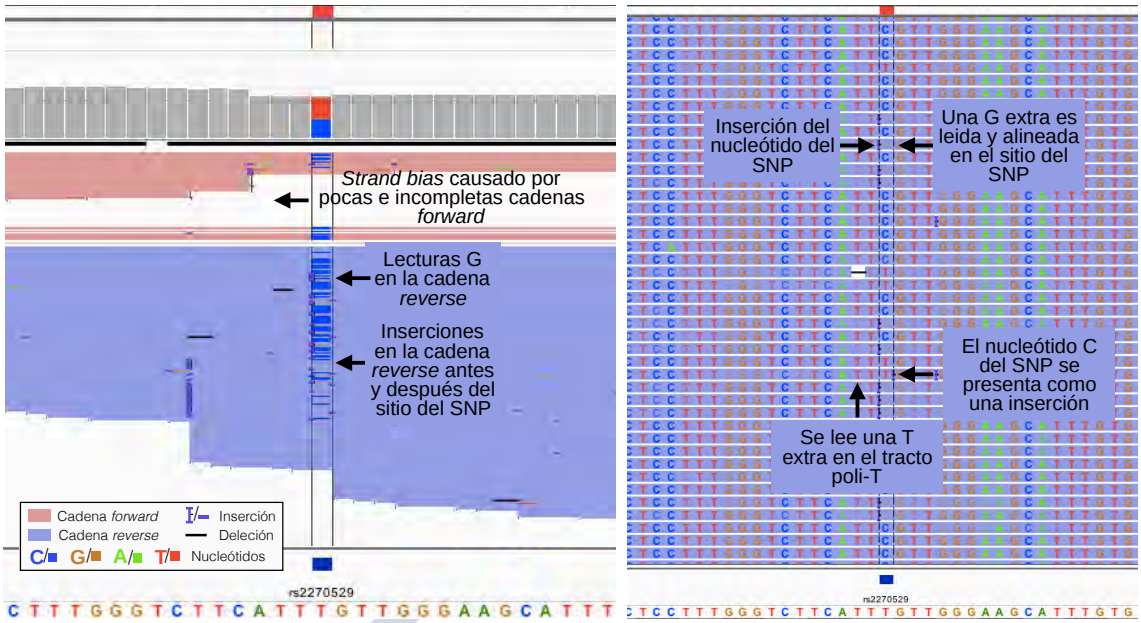


Fig. 38. Visualización en IGV de una muestra heterocigota para el SNP T/C rs2270529.

Además, se identificaron una serie de SNPs con tendencia a alineamientos erróneos que producen la incorporación errónea de bases alélicas en la posición del SNP y/o que presentan un desbalance de las lecturas alélicas en heterocigotos.

En primer lugar, el SNP T/A rs1029047 –Fig. 39– presenta tractos poli-T y poli-A adyacentes, con la estructura (TATTT[T/A]AAAAAAAAA). Los tractos homopoliméricos producen falsas inserciones y deleciones en las posiciones +1 y -2 respecto al SNP, provocando el desbalance de los genotipos heterocigotos e impidiendo un genotipado correcto en caso de mezclas de ADN, tal y como indican otros estudios (Seo *et al.* 2013, Børsting *et al.* 2014, Grandell *et al.* 2016).

En segundo lugar, el SNP T/G rs2399332 –Fig. 40– se encuentra embebido en un tracto poli-T. A raíz del conteo erróneo de Ts del tracto homopolimérico se generan falsas lecturas T y las lecturas alélicas G se alinean como inserciones (principalmente en las cadenas *forward*).

En tercer lugar, el SNP T/C rs4606077 –Fig. 41– se encuentra embebido en un tracto poli-C corto. La mayoría de las lecturas *forward* acaban en el tracto poli-C en la posición -10, y las restantes presentan un claro desbalance alélico debido a la baja presencia de lecturas C. Este efecto ocurre también para los SNPs flanqueantes de las posiciones +10 (rs58774517) y +11 (rs1869434). Se puede asumir así que la mayoría de las cadenas *forward* truncadas en el tracto poli-C corresponden al alelo C, generando el desbalance alélico y *strand bias*. No obstante, se podrían corregir los genotipos infiriéndolos a partir de las cadenas *reverse*.

En cuarto lugar, el SNP G/C rs445251 –Fig. 42– se encuentra embebido en una región repetitiva GGTT[G/C]GTG. Este SNP presenta deleciones en las cadenas *reverse*, en su mayoría de alelos C, de manera que los heterocigotos están desbalanceados y los homocigotos

CC presentan bajo *coverage*. Además, se presentan incorporaciones erróneas A en las cadenas *reverse*, seguidas de incorporaciones erróneas en otras posiciones del amplicón. Estas características indican problemas de alineamiento en el área que rodea al SNP que provocan incorporaciones erróneas, *strand bias*, *strand bias per allele* y un *coverage* bajo en la posición del SNP. No obstante, las lecturas de la cadena *forward* parecen fiables y las incorporaciones erróneas A son fácilmente detectables.

Por último, el SNP T/C rs1523537 –Fig. 43– presenta una estructura de lecturas inusual. La mayoría de las cadenas *forward* no alcanzan la posición del SNP y, en las restantes, una proporción presenta incorporaciones erróneas G debidas a errores en el alineamiento. En las cadenas *forward* predominan los alelos T, produciéndose *strand bias per allele*; mientras que en muestras homocigotas CC alrededor del ~10% de las lecturas son Ts. Sin embargo, las cadenas *reverse* parece fiables y se podrían corregir los genotipos, al menos para muestras de referencia.

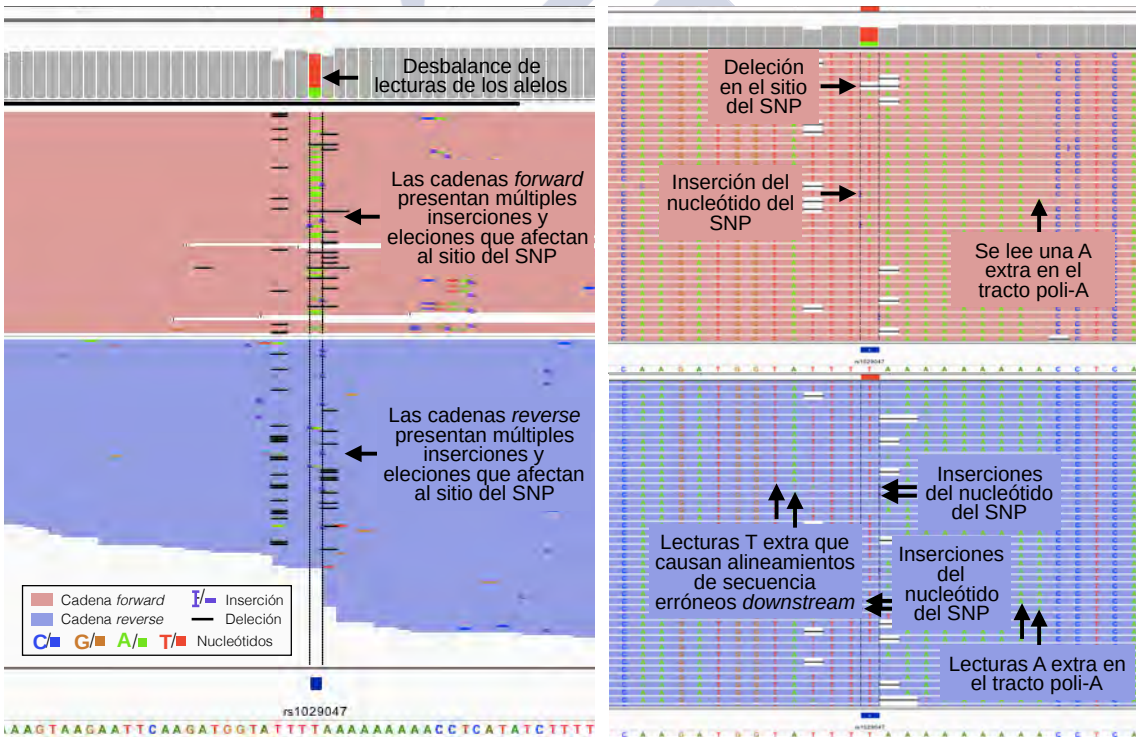


Fig. 39. Visualización en IGV del SNP T/A rs1029047.

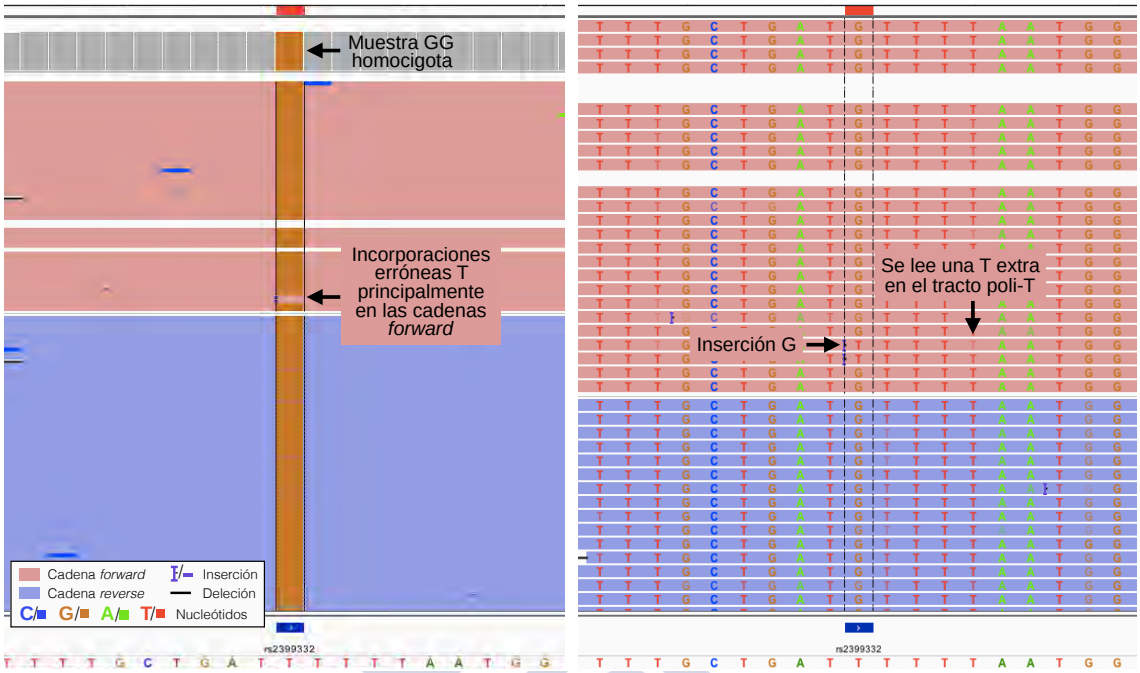


Fig. 40. Visualización en IGV de una muestra homocigota GG para el SNP T/G rs2399332.

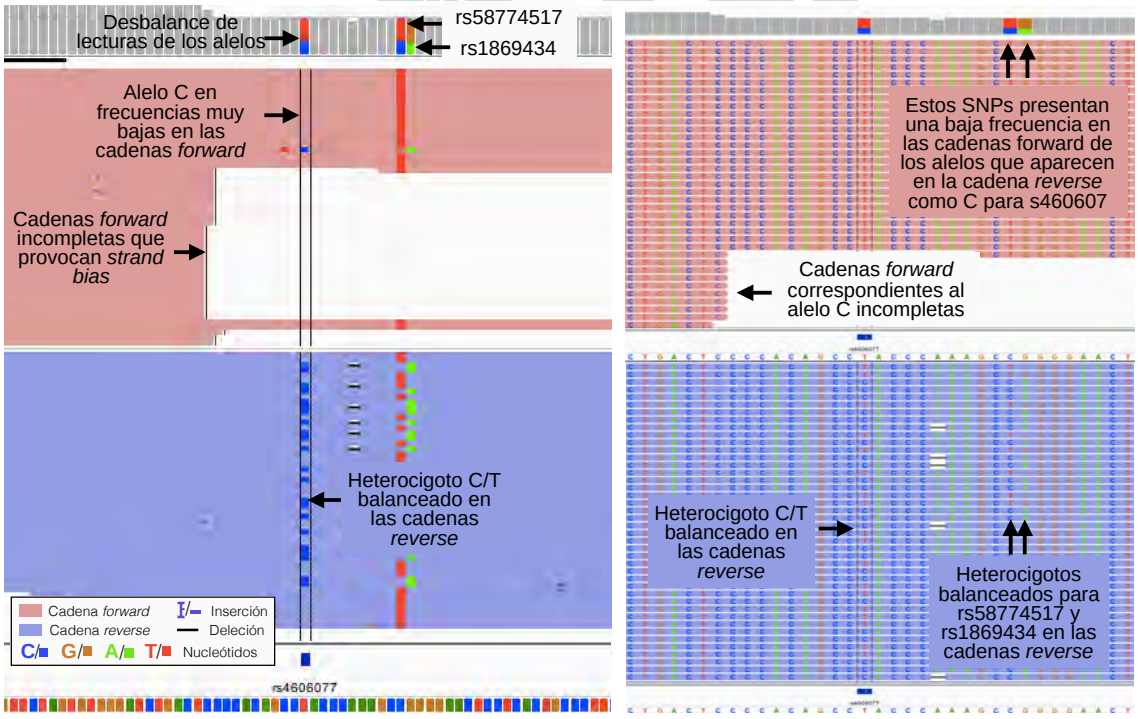


Fig. 41. Visualización en IGV de una muestra heterocigota para el SNP T/C rs4606077.

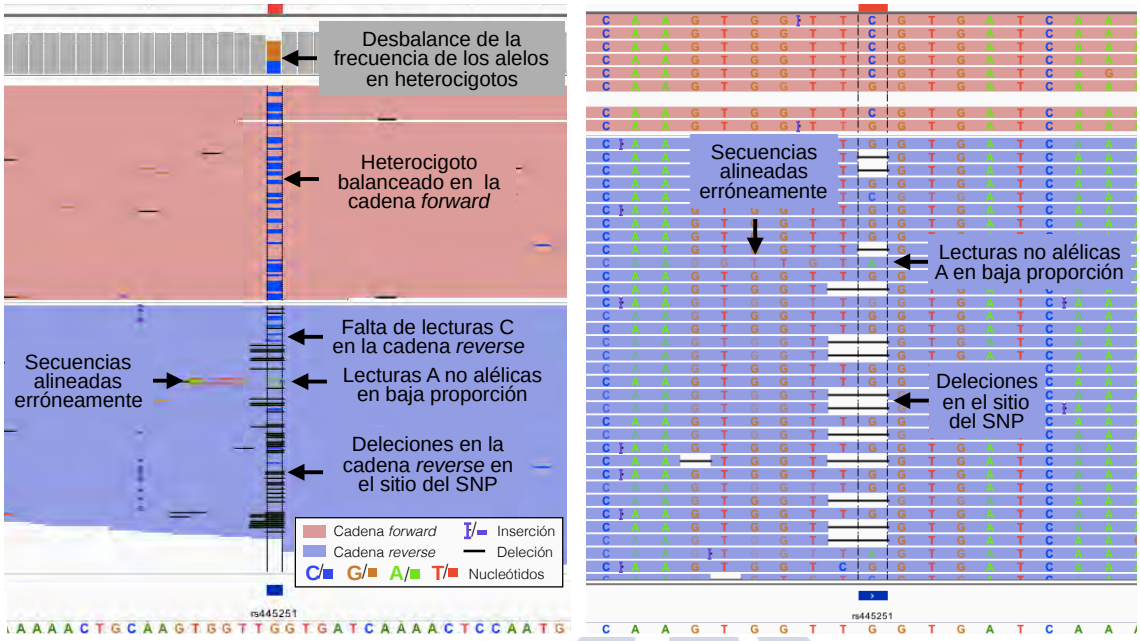


Fig. 42. Visualización en IGV de una muestra heterocigota para el SNP G/C rs445251.

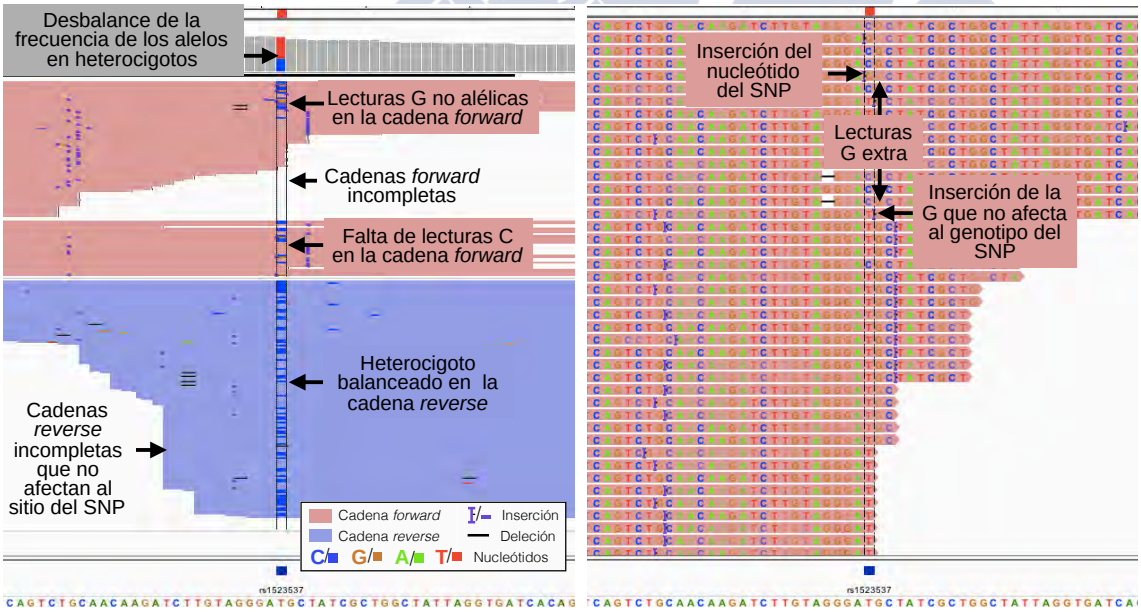


Fig. 43. Visualización en IGV de una muestra heterocigota para el SNP T/C rs1523537.

3.2.2.3 Concordancia del genotipado

La concordancia de los genotipos obtenidos para los 5 controles de ADN Coriell se evaluó comparándolos, para 3 de las 5 muestras, con los de la base de datos del Proyecto 1000 Genomas para 139 de los 140 SNPs del panel (todos los SNPs del panel excepto rs938283, que no se encuentra listado) y, para los 5 controles, con los obtenidos mediante SNaPshot para 52 de los 140 SNPs. Los genotipos fueron inferidos independientemente mediante los *software* Genotyper y Workbench a partir de las secuencias.

La concordancia entre los genotipos obtenidos mediante el panel Qiagen SNP-ID y los listados por el Proyecto 1000 Genomas alcanzó una tasa de 99,52%. No obstante, las 2 discordancias encontradas difieren entre los diferentes métodos de análisis y se centran en el SNP rs445251 en Workbench y en los SNPs rs1004357 y rs5746846 en Genotyper. En la Tabla 10 se muestran las discordancias y las causas de las mismas, que se dan en SNPs previamente identificados en la sección 3.2.2.2 como problemáticos. Se debe tener en cuenta que las diferencias encontradas para el rs445252 en Workbench no son estrictamente discordancias, sino genotipos parciales que destacan la deleción de las secuencias en una de las cadenas –Fig. 42–. Si se corrigen los genotipos de rs1004357 y se infiere el genotipo de rs5746846, que presenta un *coverage* muy bajo, se obtiene un 100% de concordancia con el Proyecto 1000 Genomas.

La concordancia entre los genotipos del panel Qiagen SNP-ID y SNaPshot alcanzó el 98,64% –Tabla 10–. Las 7 discordancias encontradas se deben a las dificultades de interpretación de los perfiles de SNaPshot (Daniel *et al.* 2015). Los genotipos discordantes fueron resueltos al analizar los SNPs en *singleplex* en SNaPshot, de manera que se alcanza finalmente un 100% de concordancia.

Hasta donde se puede apreciar con esta evaluación limitada, el *software* Workbench funciona de manera equivalente a Genotyper. Una opción clave que no se incluye en estos *software* es la posibilidad de que el usuario pueda modificar los parámetros de análisis de cada SNP individualmente. Los problemas de incorporaciones erróneas de nucleótidos a bajos niveles, alineamientos erróneos debidos a la presencia de tractos homopoliméricos, *strand bias* y artefactos tipo Indel descritos en la sección 3.2.2.2 avalan la idea de que los parámetros deben adaptarse a las características de cada SNP.

Tabla 10. Discordancias encontradas en las comparaciones entre los genotipos obtenidos mediante el panel Qiagen SNP-ID (usando los *software* Genotyper y Workbench) y los del Proyecto 1000 Genomas (P1000 Genomas) o los obtenidos mediante SNaPshot.

Controles de ADN Coriell: concordancia con el Proyecto 1000 Genomas	Marcador	Muestra	Genotyper	Workbench	P1000 Genomas		Causa de la discordancia
	rs445251	NA18498	CC	C	CC		Deleción en la posición del SNP
		NA06994	CC	C	CC		Deleción en la posición del SNP
	rs1004357	NA18498	AT	AA	AA		Incorporaciones erróneas T
	rs5746846	HG00403	NN	GG	GG		*Coverage forward* bajo
Controles de ADN Coriell: concordancia con SNaPshot	Marcador	Muestra	Genotyper	Workbench	52-plex	*singleplex*	Causa de la discordancia
	rs251934 (A43)	NA10540	CC	CC	NN	CC	Problemas en SNaPshot
		NA06994	CC	CC	NN	CC	Problemas en SNaPshot
		NA11200	CT	CT	NN	CT	Problemas en SNaPshot
	rs729172 (A16)	NA18498	AA	AA	AG	AA	Problemas en SNaPshot
		NA06994	AA	AA	AG	AA	Problemas en SNaPshot
		HG00403	AA	AA	AG	AA	Problemas en SNaPshot
	rs917118 (A07)	NA11200	GT	GT	GN	GT	Problemas en SNaPshot
ADN extraído de fémur: concordancia con SNaPshot	Marcador	Réplica 1		Réplica 2		52-plex	Causa de la discordancia
	rs938283 (A33)	TT		TT		CT	Tipado erróneo en SNaPshot
	rs917118 (A07)	AC		AC		CN	Tipado erróneo en SNaPshot

3.2.2.4 Parámetros de calidad de las secuencias

Los datos de las secuencias de los 5 controles de ADN Coriell se utilizaron para definir los parámetros de calidad de las secuencias: *coverage* promedio del marcador, *strand bias*, *strand bias per allele* (alelos referencia y alternativo), tasa de *misincorporation* y valores de ARF.

El promedio de *coverage* de los SNPs del panel –SNP *Target Reads*– para todos los análisis fue de 1220,3 lecturas; con un rango entre 22,2x y 2239,2x. No obstante, el valor de *coverage* promedio más bajo de las muestras de concordancia es de 211,2x. Estos niveles están muy por encima del umbral de *coverage* mínimo de 15-20x señalado por varios estudios de MPS (Bentley *et al.* 2008, Quail *et al.* 2012, Daniel *et al.* 2015). En la Fig. 44 se ordenan los SNPs en función creciente de los niveles de *coverage* promedio (basados en las muestras de concordancia), indicando unos niveles razonables de homogeneidad del número de secuencias obtenidas para cada SNP del panel. Aunque la mezcla de ADN 1:9 réplica A presenta algunos problemas inesperados de *coverage*, todos los valores de *coverage* de las diluciones seriadas y el ADN extraído de fémur se encuentran por encima del umbral mínimo de 20x y tan solo un 5-6% de SNPs presentan un *coverage* <200x. Cuando se comparan los resultados de este estudio con los de Grandell *et al.* (2016), que evalúa el mismo panel mediante una plataforma de secuenciación diferente (Illumina MiSEq), las principales diferencias en *coverage* entre los marcadores que componen el panel pueden ser explicadas en base a la PCR inicial de captura. Como esta PCR es común a ambos análisis, se esperan similitudes en la distribución de los *coverage* relativos de los marcadores. Dos SNPs fueron identificados como infrarrepresentados por Grandell *et al.* (2016): rs1360288 y rs105883; que se encuentran entre los marcadores con menor *coverage* promedio en este estudio (SNPs en la izquierda en la Fig. 44), con valores de *coverage* promedio por debajo de 400x. De entre los 5 marcadores con menor proporción de *coverage* normalizado en el estudio de Grandell *et al.* (2016), los 2 anteriores más rs5746846, rs873196 y rs2567608 se encuentran listados entre los

7 SNPs con menor *coverage* de este estudio. Las diferencias de *coverage* en rs9951171 y rs1005533, identificados en este estudio como SNPs con bajo *coverage*, puede deberse a las diferencias en la cantidad de ADN inicial (20 veces menor en este estudio) o a un efecto estocástico causado por limitado número de muestras analizado (5 vs. 49).

Los valores de *strand bias* indicaron que 11 SNPs (rs891700; rs9866013; rs13182883; rs727811; rs321198; rs430046; rs576261; rs2567608; rs1005533; rs722098 y rs5746846). presentan lecturas sesgadas fuera de un umbrales del 25-75%, como se muestra en la Fig. 45. Estos 11 SNPs también fueron identificados mediante la visualización de las lecturas en IGV –sección 3.2.2.2– como marcadores con riesgo de *no-call* si se aplican valores umbrales rigurosos para los parámetros de *coverage* mínimo. Además, el SNP rs4606077 presentó un importante *strand bias per allele*, con lecturas del alelo alternativo muy sesgadas a la detección en la cadena *reverse* –tal y como se detalla en la sección 3.2.2.2–.

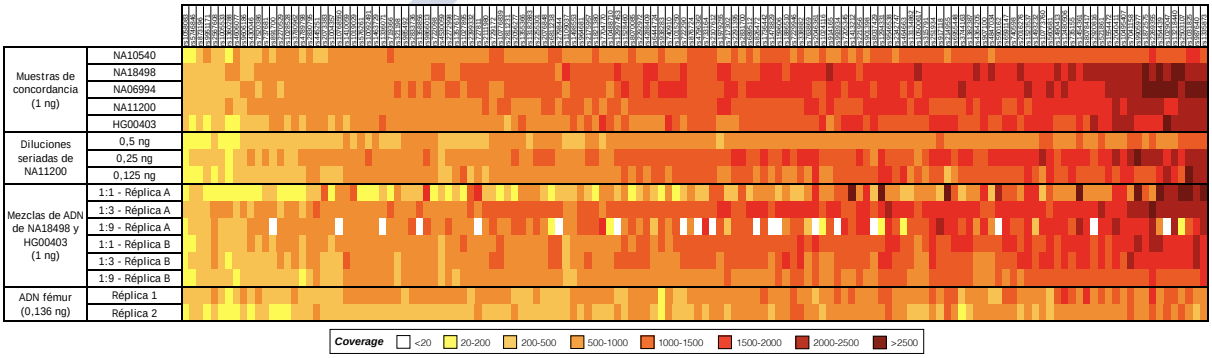


Fig. 44. *Heatmap* en el que se presentan los valores de *coverage* de los 140 SNPs incluidos en el panel para cada muestra analizada. Marcadores ordenados por *coverage* promedio creciente en muestras de concordancia.

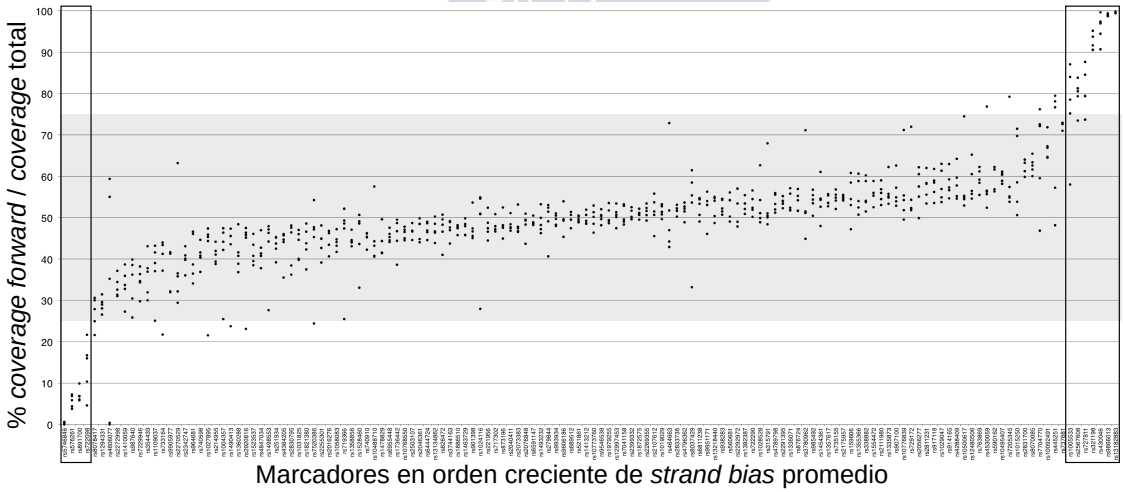


Fig. 45. Valores de *strand bias* (como % de *coverage forward* / *coverage* total) observados en las muestras de concordancia para los 140 marcadores incluidos en el panel. Los marcadores con lecturas muy sesgadas (<25% o >75%) se enmarcan en negro. Los marcadores se ordenan por *strand bias* promedio creciente.

Un total de 9 SNPs mostraron tasas altas de *misincorporation* (lecturas no alélicas > 1%): rs4847034, rs1554472, rs2270529, rs1821380, rs4796362, rs1004357, rs445251, rs1523537 y rs733164; tal y como se describe en la sección 3.2.2.2. Se debe destacar que los altos valores que presentan los SNPs rs1004357 (10,32%) y rs733164 (6,12%) indican que la frecuencia de lecturas de nucleótidos no alélicos puede alcanzar el umbral del 10% necesario para asignar genotipos heterocigoto en Genotyper, causando discordancias como la recogida en la Tabla 10 para rs1004357.

Los valores de ARF que se muestran en la Fig. 46 indican patrones que se corresponden adecuadamente con los rangos esperados para muestras heterocigotas (40-60%) y homocigotas (hasta el 5% y por encima del 95%). Los 10 SNPs que se indican en la Fig. 46 se desvían de estos rangos en algunas muestras: rs7520386, rs2046361, rs2056277, rs2833736, rs1029047, rs1478829, rs4606077, rs430046, rs3744163 y rs1523537. Los 4 primeros SNPs de esta lista no habían sido identificados como marcadores problemáticos durante los análisis de IGV presentados en la sección 3.2.2.2 y presentan una única observación fuera de los valores definidos de ARF.

En el estudio de Grandell *et al.* (2016) se identifican 3 SNPs con valores de ARF desviados de los umbrales: : rs2399332, rs4530059 y rs1029047. Los autores explican este efecto para rs2399332 y rs4530059 mediante la presencia de polimorfismos en las regiones de unión de los *primers*, pero estos SNPs no presentan valores de ARF desviados en este estudio. No obstante, el SNP rs2399332 presentó incorporaciones erróneas debidas a la presencia de un tracto homopolimérico –ver Fig. 40–. Se debe destacar que este SNP fue identificado utilizando un set de *primers* diferente en el trabajo presentado en la sección 3.1 como SNP con incorporaciones erróneas, y por Børsting *et al.* (2014), como SNP con desbalance alélico. El SNP rs4530059 se identificó como SNP con desbalance alélico por Børsting *et al.* (2014), pero no en el trabajo presentado en la sección 3.1.

Además, tan solo rs430046 y rs1523537 coinciden con los 4 SNPs identificados en el trabajo presentado en la sección 3.1 como SNPs con valores de ARF desviados, donde se identificaron además rs8037429 y rs803749 (en este trabajo se encuentran dentro de los valores umbral establecidos). Por último, rs1029047 presenta desviaciones de las ARF debido a la presencia de un tracto homopolimérico –ver Fig. 39– y fue identificado por Grandell *et al.* (2016) y, mediante un set de *primers* diferente, por Seo *et al.* (2013), Børsting *et al.* (2014) y en el trabajo presentado en la sección 3.1. A la vista de estos resultados, establecer un set de valores de referencia que identifique los marcadores más balanceados y fiables a la hora de detectar mezclas de ADN es un proceso que debe ser realizado por cada laboratorio para cada panel y sistema de secuenciación.

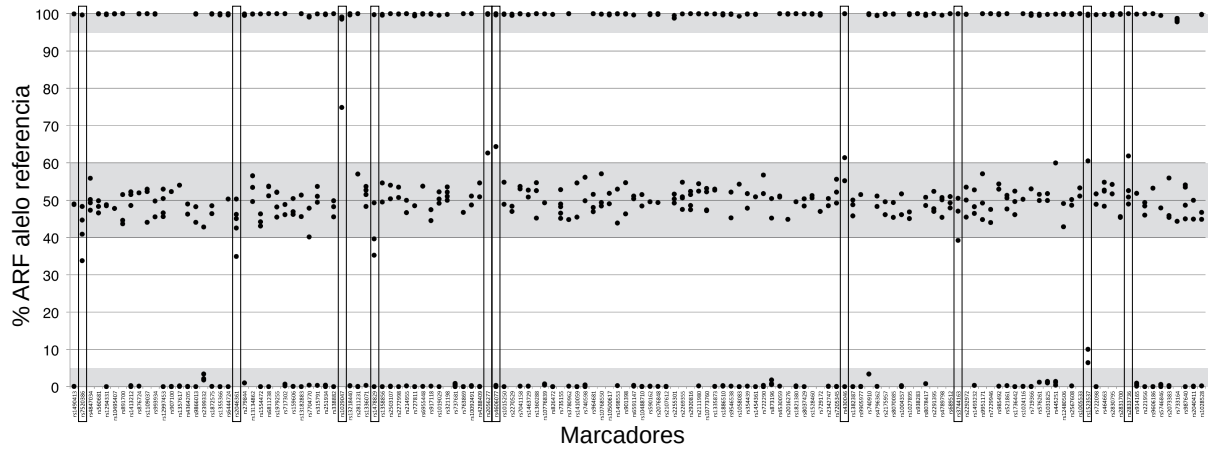


Fig. 46. Valores de ARF como porcentaje de lecturas del alelo de referencia / *coverage* total para las muestras de concordancia. Los marcadores se muestran en orden cromosómico. Las áreas remarcadas en gris representan los valores esperados para heterocigotos (entre 40-60%) y homocigotos (<5% y >95%). Los marcadores que se desvían de estos umbrales están enmarcados en negro.

3.2.2.5 Evaluación de la sensibilidad forense del panel

Los análisis de diluciones seriadas indicaron un 100% de genotipos asignados y concordantes para cantidades iniciales de ADN de 0,5 ng, 0,25 ng y 0,125 ng. Los niveles de *coverage* fueron comparables a los de análisis con cantidad inicial de ADN de 1 ng, con valores promedio de SNP *Target Reads* de: 650,7x para 0,5 ng; 1193,7x para 0,25 ng y 966,2x para 0,125 ng. Aunque el *coverage* final se ve afectado por la ecualización de las librerías antes de la preparación del molde de secuenciación, se observa una correlación entre la cantidad inicial de ADN y los resultados de cuantificación de la PCR inicial y de las librerías –presentados en la Tabla 9–. Se debe destacar también que los valores de ARF de las diluciones seriadas tienden a desviarse de los típicos para muestras heterocigotas –ver Fig. 47A en comparación con Fig. 46–, un efecto señalado por Grandell *et al.* (2016) utilizando el mismo panel en la plataforma MiSeq. De hecho, la mayor desviación frente a las ARF balanceadas habituales de heterocigotos se observaron en los análisis de ADN degradado extraído de un fémur –ver Fig. 47B–.

La muestra de ADN degradado (0,136 ng de ADN inicial) produjo, tras la PCR inicial, más de 0,4 ng de ADN para la preparación de las librerías. Los valores de cuantificación de las librerías de las dos réplicas fueron ligeramente inferiores a los de la dilución seriada de 0,125 ng –ver Tabla 9–. La media de la longitud de las secuencias de ambas réplicas fue ligeramente menor que la media del resto de muestras (158 vs. 168,74), tal y como se muestra en la Tabla 9. Para las réplicas 1 y 2 se obtuvieron valores promedio de *coverage* de 704,28x y 524,54x; respectivamente –Tabla 9–. Estos valores son inferiores a los de otras muestras analizadas, pero ningún SNP presentó valores de *coverage* por debajo del umbral de 20x –Fig. 44–. Se obtuvieron genotipos para todos los SNPs del panel, concordantes entre ambas réplicas. La comparación de los genotipos asignados por Genotyper con los 46 detectados

mediante SNaPshot indicó una concordancia del 96,73%. La causa de las discordancias observadas en dos SNPs (rs938283 y rs1031825, ver Tabla 10) no pudo ser resuelta, pero estos SNPs presentaron un 100% de concordancia en los controles de ADN y valores de *coverage* por encima de la media, sugiriendo que los genotipos de SNaPshot podrían haber sido identificados erróneamente.

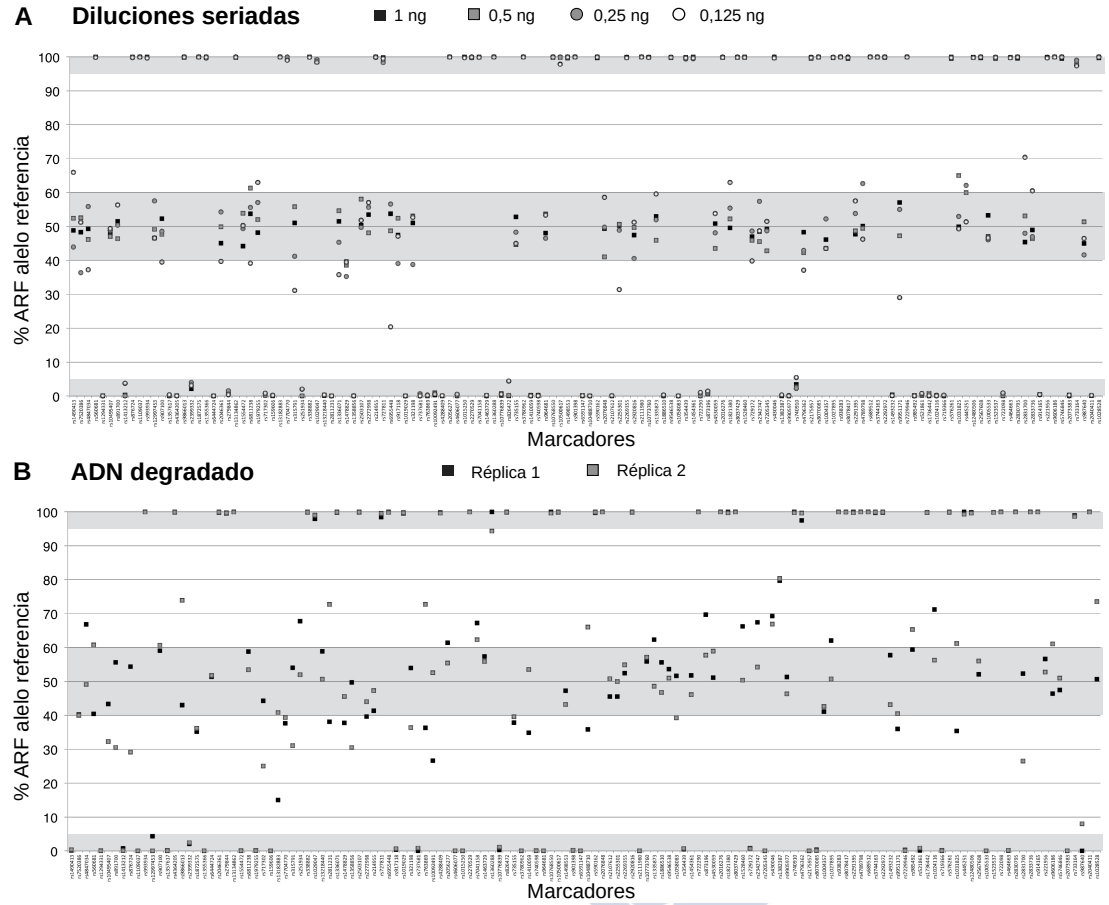


Fig. 47. Valores de ARF como porcentaje de lecturas del alelo de referencia / *coverage* total para las muestras de concordancia. Los marcadores se muestran en orden cromosómico. Las áreas remarcadas en gris representan los valores esperados para heterocigotos (entre 40-60%) y homocigotos (<5% y >95%). A) Diluciones seriadas de NA11200, la leyenda indica el símbolo de cada cantidad inicial de ADN. B) Muestra degradada extraída de un fémur, la leyenda indica el símbolo de cada réplica.

3.2.2.6 Detección de mezclas de ADN

El balance de lecturas de los alelos de cada SNP en muestras individuales permite establecer umbrales de ARF para homocigotos y heterocigotos –Fig. 46–. No obstante, las lecturas de mezclas de ADN presentan patrones desbalanceados, derivados de la adición de copias extra de uno de los alelos –ver Fig. 48–. Así, las mezclas de ADN presentan distribuciones de ARF diferenciadas de las muestras individuales, con un alto número de

heterocigotos fuera del umbral 40-60%. Estos patrones fueron distinguibles para todas las ratios de mezcla analizadas, especialmente en 1:1 y 1:3. Además, la proporción de heterocigotos se eleva desde el ~48% en las muestras individuales que componen las mezclas (NA18498 y HG00493), hasta el ~80% en la mezcla de ADN de ratio 1:1.

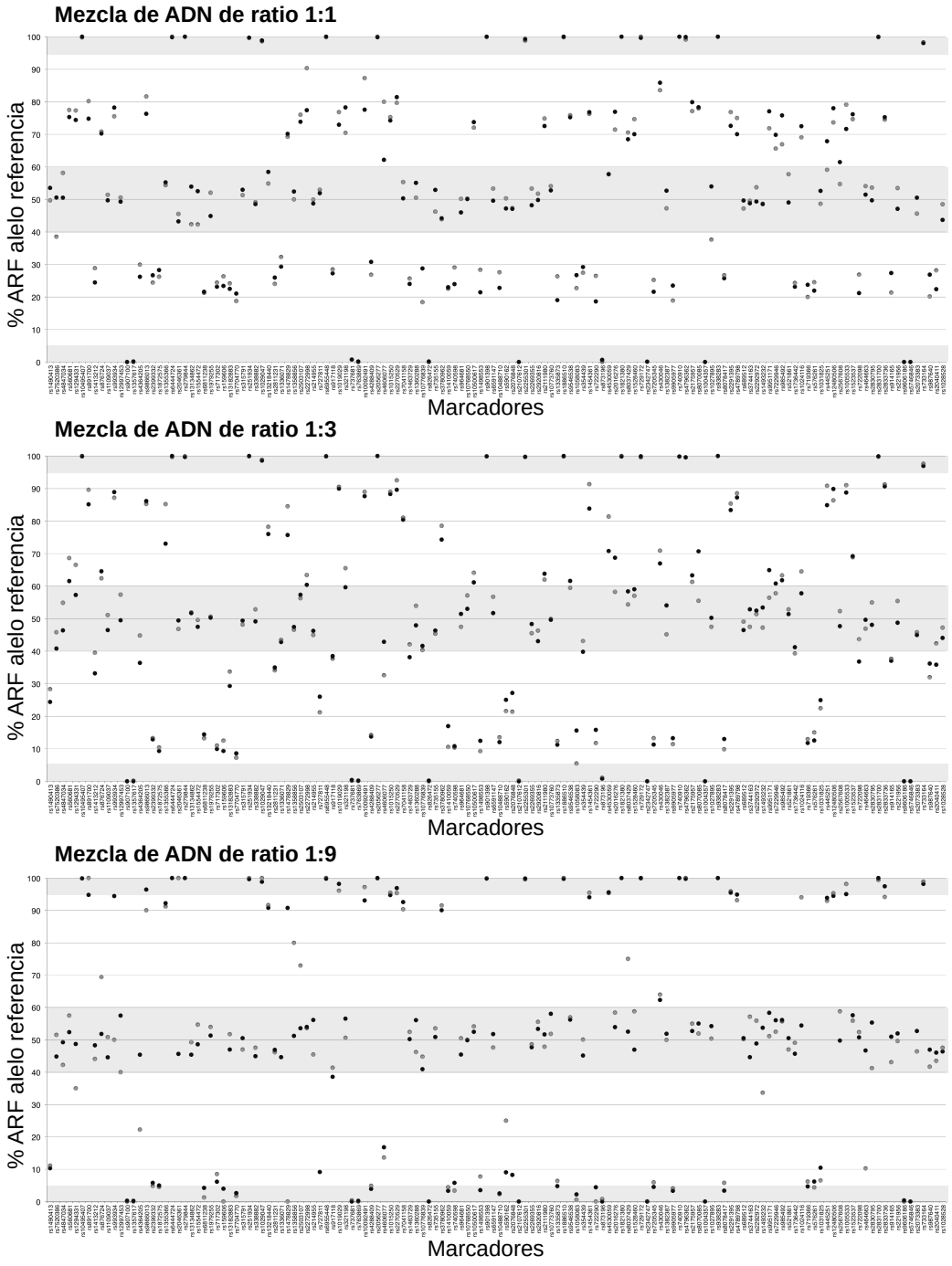


Fig. 48. Valores de ARF del alelo de referencia como lecturas del alelo/*coverage* total para ambas réplicas de las tres ratios de mezcla de ADN (1:1, 1:3 y 1:9). Los marcadores se muestran en orden cromosómico. Las réplicas A y B se representan como círculos rellenos o vacíos, respectivamente. Las áreas en gris representan los patrones de muestras individuales para heterocigotos (entre 40-60%) y homocigotos (<5% y >95%).

En el estudio de mezclas de ADN del panel HID-Ion AmpliSeq Identity v. 2.2 –ver sección 3.1.2.6.3– se compararon los resultados de dos configuraciones de parámetros diferentes para Genotyper: *Germline* (que incluye el parámetro minimum_allele_frequency=0,1) y *Somatic* (que incluye minimum_allele_frequency=0,02). Los resultados indicaron que la configuración *Somatic* es la más efectiva a la hora de detectar variantes a baja frecuencia. Dado que la versión actual de Genotyper no permite seleccionar estas configuraciones, la configuración por defecto (minimum_allele_frequency=0,1) se reservó para *Germline* y el parámetro minimum_allele_frequency se modificó a 0,02 para *Somatic*, dejando el resto de parámetros en su configuración inicial. Al reducir el grado de *stringency* de los parámetros, se pretende reducir el número de discordancias entre los genotipos asignados y los de la mezcla esperada, detectando los alelos minoritarios que se encuentran en frecuencias entre el 2-10%. El uso del umbral de 0,02 redujo la tasa de *drop-out* del 11,4% al 0,8% (7 *drop-outs* en 6 SNPs en la mezcla de ratio 1:9). Inexplicablemente, el SNP rs10488710 presentó 2 *drop-outs* en ambas réplicas de la mezcla de ratio 1:9 a pesar de que las ARF del alelo minoritario presentan valores de 0,023 y 0,025, respectivamente. Las diferentes configuraciones de parámetros no afectaron significativamente a la tasa de *no-call* (4,3% con parámetros por defecto y 3,5% con minimum_allele_frequency=0,02), que se debe a un bajo *coverage* en todos los casos.

Cuando se aplica el *software* Genotyper al análisis de mezclas de ADN, se debe tener en cuenta que bajar el nivel de *stringency* del parámetro minimum_allele_frequency eleva el riesgo de asignar incorrectamente un genotipo heterocigoto a un SNP homocigoto (*drop-in*), ya que existen SNPs homocigotos que muestran valores de ARF desviados de los valores ideales 0 y 1 en muestras individuales –ver Fig. 46–. En este estudio, no se encontraron *drop-ins* cuando se aplicaron los umbrales de 0,1 o 0,02. Además, se debe tener un especial cuidado al analizar los SNPs atípicos. En este sentido, se tuvieron que realizar dos intervenciones. En primer lugar, en el componente HG00403 el SNP rs5746846 tuvo que ser corregido tal y como se describe en la sección 3.2.2.2, ya que presenta un importante *strand bias* que provoca que la secuencia sea leída únicamente en la dirección *reverse* y se obtengan *no-calls*. En segundo lugar, todos los genotipos del SNP rs1004357 se corrigieron a homocigotos AA después de la visualización de las muestras mediante IGV. En la sección 3.2.2.2 se describe este SNP como problemático debido a un tracto poli-T adyacente a la posición del SNP que produce lecturas T no alélicas.

En resumen, el análisis de un número limitado de mezclas de ADN indica que la reducción del parámetro minimum_allele_frequency en Genotyper es un paso crucial para la obtención de genotipos de mezclas de ADN. El ajuste de este parámetro debe realizarse retrospectivamente en aquellos casos en los que la muestra sugiera la presencia de una mezcla de ADN, observándose una mayor heterocigosidad y patrones de ARF desviados.

3.2.2.7 Estimación de frecuencias haplotípicas a partir del Proyecto 1000 Genomas

En la Tabla 11 se resumen las frecuencias haplotípicas estimadas a partir de los datos del Proyecto 1000 Genomas para los pares de SNPs rs10768550-rs10500617 (679 nucleótidos de separación) y rs9606186-rs5746846 (287 nucleótidos de separación) para 4 grupos poblacionales de referencia (AFR, EUR, SAS y EAS) y poblaciones *admixed* de AMR.

En el par de SNPs rs10768550-rs10500617, el haplotipo CA solo se observa 2 veces y el TT es muy infrecuente en poblaciones no-AFR (la presencia de este haplotipo en poblaciones *admixed* de AMR se debe probablemente a *admixture* con poblaciones AFR). En el par de SNPs rs9606186-rs5746846 el haplotipo GC ocurre a alta frecuencia, pero no se observaron haplotipos CG, sugiriendo una tasa de recombinación muy baja entre ambos SNPs. Así, la inclusión de estos dos pares de SNPs en el panel reduce ligeramente la informatividad general del panel y requiere del uso de estimaciones de frecuencias haplotípicas en lugar de estimaciones de frecuencias alélicas independientes.

Tabla 11. Estimaciones de las frecuencias haplotípicas a partir de datos del Proyecto 1000 Genomas para los dos pares de SNPs más próximos del panel Qiagen SNP-ID.
Al. Ref.: alelo de referencia. Al. Alt.: alelo alternativo. *excluye las poblaciones *admixed* ASW y ACB.

	Par de SNPs					
		rs10768550	rs10500617		rs9606186	rs5746846
	Al. Ref.	C	T	Al. Ref.	C	C
	Al. Alt.	T	A	Al. Alt.	G	G
	Haplotipo	Número de haplotipos observados	Frecuencia haplotípica	Haplotipo	Número de haplotipos observados	Frecuencia haplotípica
AFR*	CT	631	0.6260	GG	468	0.4643
	TA	227	0.2252	CC	388	0.3849
	CA	1	0.0010	CG	0	0.0000
	TT	149	0.1478	GC	152	0.1508
		1008			1008	
EUR	CT	740	0.7356	GG	545	0.5417
	TA	265	0.2634	CC	416	0.4135
	CA	0	0.0000	CG	0	0.0000
	TT	1	0.0010	GC	45	0.0447
		1006			1006	
SAS	CT	763	0.7802	GG	528	0.5399
	TA	215	0.2198	CC	340	0.3476
	CA	0	0.0000	CG	0	0.0000
	TT	0	0.0000	GC	110	0.1125
		978			978	
EAS	CT	611	0.6062	GG	716	0.7103
	TA	396	0.3929	CC	279	0.2768
	CA	1	0.0010	CG	0	0.0000
	TT	0	0.0000	GC	13	0.0129
		1008			1008	
Admixed AMR	CT	482	0.6945	GG	413	0.5951
	TA	198	0.2853	CC	239	0.3444
	CA	0	0.0000	CG	0	0.0000
	TT	14	0.0202	GC	42	0.0605
		694			694	

3.2.3 **Discusión**

En este estudio se evaluó un nuevo *multiplex* forense para MPS que incluye 140 SNPs. Los datos obtenidos indican un rendimiento adecuado del panel. Aunque los resultados se basan en un número limitado de muestras, las mismas muestras fueron genotipadas en el estudio presentado en la sección 3.1 para la mayoría de los marcadores con el mismo sistema de detección MPS pero diferente PCR de captura y metodología de preparación de librería, obteniéndose niveles de *coverage* comparables entre ambos kits. El análisis detallado de las secuencias obtenidas en IGV y de los genotipos asignados revela que los problemas de genotipado de ciertos SNPs no son causados por la metodología de construcción de librerías de Qiagen, sino que se relacionan con las características de la secuencia contexto de dichos SNPs. Por ello, muchos de los SNPs identificados como problemáticos han sido eliminados de la versión final del panel de ID-SNPs desarrollado por TFS para el sistema Ion PGM™ –el panel Precision ID Identity–. Así, el uso del panel Qiagen SNP-ID requiere un análisis cuidadoso de ciertos SNPs y la comprobación de los genotipos asignados para los mismos mediante la visualización de las secuencias en IGV.

Se aplicaron 2 *software* de genotipado diferentes (Genotyper y Workbench) a los mismos datos de secuencias. Ambos sistemas proporcionaron un rendimiento casi idéntico a la hora de obtener genotipos consistentes en muestras control. En Genotyper, un único SNP en una de las muestras presenta un genotipo discordante con el Proyecto 1000 Genomas, que se debe a que las incorporaciones no alélicas sobrepasan el umbral de frecuencia del 10% (puede ser corregido manualmente) y se produce un *no-call* por *coverage* bajo (aunque el genotipo puede ser inferido a través de los archivos generados por Genotyper). En Workbench se obtiene una deleción, debido a un artefacto tipo Indel en una de las cadenas, en el SNP rs445251 (detectable mediante la comprobación de las secuencias obtenidas en IGV). Así, tras la corrección de los genotipos de 1-2 SNPs fácilmente identificables, ambos sistemas de análisis produjeron un 100% de concordancia con los datos del Proyecto 1000 Genomas, indicando una alta fiabilidad.

Los SNPs que requieren una inspección manual cuando se analiza ADN *low level* o degradado pueden ser fácilmente reemplazados o eliminados del panel. Además, los dos pares de SNPs próximos no son aplicables como marcadores independientes y presentan haplotipos con una informatividad similar al uso de un único SNP de cada par (con la excepción del par rs9606186-rs5746846 en las poblaciones AFR y SAS). No obstante, la aplicación de paneles de SNPs de amplicones cortos como herramienta para la identificación forense en casos que presentan ADN degradado requiere la comprobación minuciosa de los genotipos obtenidos, tanto cuando se aplican los sistemas de MPS como cuando se aplican paneles más pequeños de SNaPshot.

4. Bloque II: Ancestralidad Biogeográfica

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. Bloque II: Ancestralidad biogeográfica

En este bloque se presentan dos trabajos en los que se adapta el panel teórico EUROFORGEN Global AIM-SNP (Phillips *et al.* 2014a) a dos metodologías diferentes de análisis de SNPs: electroforesis capilar mediante SNaPshot –sección 4.1– y MPS mediante la plataforma Ion PGM^{TM} –sección 4.2–.

4.1 PANEL G-AIMS NANO

En este trabajo se presenta un nuevo panel de predicción de ancestralidad biogeográfica, capaz de diferenciar de forma balanceada 5 poblaciones continentales. El panel compila los marcadores más informativos del set EUROFORGEN Global AIM-SNP (Phillips *et al.* 2014a) en un ensayo SNaPshot optimizado para su uso en genética forense. Los resultados se encuentran publicados en la siguiente referencia:

> **de la Puente M**, Santos C, Fondevila M, Manzo L, Carracedo Á, Lareu MV y Phillips C (2016). "*The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs*." Forensic Sci Int Genet 22: 81-88.

4.1.1 Material y métodos

4.1.1.1 Genotipos de SNPs de las poblaciones de referencia y muestras de ADN

Los datos genotípicos de individuos de diferentes poblaciones se obtuvieron a partir del Proyecto 1000 Genomas Fase III (The Genomes Project Consortium 2015) y de los resultados del análisis del panel HGDP-CEPH (Li *et al.* 2008) a través del portal SPSmart (Amigo *et al.* 2008). Se escogieron como poblaciones de referencia aquellas que presentaron los niveles más bajos de variación intrapoblacional entre las 5 ancestralidades biogeográficas consideradas: África subsahariana (AFR), Europa (EUR), este de Asia (EAS), Oceanía (OCE) y nativa de América (AMR). Asimismo, se recogieron datos de otras poblaciones de estudio del Proyecto 1000 Genomas, tanto de poblaciones *unadmixed* (a modo de set test) como de poblaciones *admixed*.

Las poblaciones de referencia se compilaron del Proyecto 1000 Genomas: AFR 108 YRI (*Yoruba in Ibadan, Nigeria*[26]); EUR 99 CEU (*Utah Residents with North and Western European ancestry*) y EAS 103 CHB (*Han Chinese in Beijing, China*) y de los datos obtenidos para el panel HGDP-CEPH: OCE 28 individuos (17 *Papuan from New Guinea* y 11 *Melanesian from Bougainville*) y AMR 64 individuos (14 *Karitiana from Brazil*, 8 *Surui from Brazil*, 21 *Maya from Mexico*, 14 *Pima from Mexico*, y 7 *Piapoco from Colombia*).

[26] Se mantienen las descripciones originales de las poblaciones recogidas en las bases de datos.

Las poblaciones *unadmixed* comprenden: 99 AFR LWK (*Luhya in Webuye, Kenya*), 113 AFR GWD (*Gambian in Western Divisions in the Gambia*), 85 AFR MSL (*Mende in Sierra Leone*), 99 AFR ESN (*Esan in Nigeria*), 107 EUR TSI (*Toscani in Italia*), 99 EUR FIN (*Finnish in Finland*), 91 EUR GBR (*British in England and Scotland*), 107 EUR IBS (*Iberian Population in Spain*), 104 EAS JPT (*Japanese in Tokyo, Japan*); 105 EAS CHS (*Southern Han Chinese*), 99 EAS KHV (*Kinh in Ho Chi Minh City, Vietnam*) y 93 EAS CDX (*Chinese Dai in Xishuangbanna, China*).

Las poblaciones *admixed* comprenden: 61 ASW (*Americans of African Ancestry in SW USA*), 96 ACB (*African Caribbeans in Barbados*), 104 PUR (*Puerto Ricans from Puerto Rico*), 94 CLM (*Colombians from Medellin, Colombia*), 64 MXL (*individuals with Mexican Ancestry from Los Angeles USA*) y 85 PEL (*Peruvians from Lima, Peru*).

Con el fin de evaluar la sensibilidad forense del ensayo, se analizaron muestras comprometidas procedentes de casos de rutina y ADNs control, incluyendo: (i) 5 muestras de ADN de cada uno de los grupos poblacionales, previamente analizadas en un ejercicio colaborativo de predicción de ancestralidad (Santos *et al.* 2015); (ii) ADN extraído a partir de restos esqueléticos degradados; (iii) diluciones seriadas del control de ADN 9947A de 1 ng/µL; 0,5 ng/µL; 0,25 ng/µL; 0,125 ng/µL; 0,064 ng/µL; 0,032 ng/µL y 0,016 ng/µL.

4.1.1.2 Selección de AIM-SNPs y diseño del ensayo SNaPshot

A partir del panel EUROFORGEN Global AIM-SNP, se seleccionó un subconjunto de AIM-SNPs atendiendo a los siguientes criterios: (i) diferenciación de los 5 grupos poblacionales, de manera que los valores de Divergencia específica de población (PSD: *population-specific Divergence*) resulten lo más balanceados posible; (ii) inclusión de SNPs trialélicos, para favorecer la detección de mezclas de ADN y (iii) conservación de una distancia mínima de 1 Mb entre marcadores sinténicos, para asegurar la independencia de los mismos. Los detalles de los SNPs seleccionados y sus frecuencias alélicas en cada una de las poblaciones de referencia se muestran en la Tabla 12.

Los 31 SNPs seleccionados –panel G-AIMs Nano– fueron incluidos en un ensayo de SNaPshot, que se diseñó y optimizó siguiendo recomendaciones previamente publicadas (Sánchez y Endicott 2006).

Las reacciones de PCR, ajustadas a un volumen final de 10 µL, constan de:

- 1 µL de Buffer II (100 mM Tris-HCl; pH 8,3; 500 mM KCl)
- 1,8 µL de $MgCl_2$ a 25 mM
- 0,1 µL AmpliTaq Gold® DNA Polymerase (a 5 U/µL)
- 0,4 µL GeneAmp® 10 mM dNTP Mix with dTTP (Applied Biosystems, AB)
- 1 µL de seroalbúmina bovina a 3.2 mg/ml
- 1,5 µL de *mix* de *primers*
- 1 ng de ADN

Tabla 12. Características de los AIM-SNPs seleccionados y frecuencias del alelo de referencia (AR) en cada una de las poblaciones de referencia.
Infor: Informatividad. Trial.: marcadores trialélicos. CI: código interno. Cr.: cromosoma.

Detalles del SNP						Frecuencia del alelo de referencia				
Infor.	SNP	CI	Cr.	Posición	AR	AFR	EUR	EAS	OCE	AMR
AFR	rs2814778	Y7	1	159174683	A	0.005	1.000	1.000	1.000	0.992
	rs1871534	S3	8	145639681	C	0.981	0.000	0.000	0.000	0.000
	rs2789823	Y2	9	136769888	G	0.935	0.000	0.000	0.000	0.000
EUR	rs1426654	Y6	15	48426484	A	0.014	1.000	0.029	0.000	0.039
	rs16891982	S1	5	33951693	C	1.000	0.020	0.985	1.000	0.984
	rs12142199	R4	1	1249187	G	0.977	0.177	0.971	1.000	1.000
	rs8072587	S2	17	19211073	C	0.986	0.182	1.000	1.000	0.817
	rs9522149	Y3	13	111827167	T	0.972	0.237	0.995	1.000	0.977
	rs4749305	R5	10	28391596	A	0.389	0.909	0.078	0.036	0.008
EAS	rs17822931	R3	16	48258198	C	1.000	0.869	0.029	0.875	0.650
	rs1229984	Y4	4	100239319	A	0.000	0.015	0.709	0.071	0.000
	rs3827760	R6	2	109513601	T	1.000	1.000	0.063	0.946	0.109
	rs6437783	Y1	3	108172817	C	0.259	0.146	0.995	0.589	0.891
	rs12594144	K1	15	64161351	C	1.000	0.889	0.121	0.607	0.177
	rs4657449	R7	1	165465281	G	0.912	0.909	0.102	0.000	0.117
OCE	rs9908046	R9	17	53563782	C	0.958	0.929	0.883	0.018	0.992
	rs3751050	R1	11	9091244	A	0.972	0.924	0.966	0.089	0.961
	rs2139931	Y10	1	84590527	A	0.898	0.753	0.879	0.018	0.898
	rs715605	R10	22	30640308	T	0.866	0.914	0.985	0.089	1.000
	rs6054465	Y11	20	6673018	T	0.972	0.859	0.743	0.036	0.859
	rs9809818	M1	3	71480566	C	0.019	0.116	0.869	0.982	0.820
AMR	rs12498138	Y5	3	121459589	G	1.000	0.949	0.922	0.911	0.094
	rs10483251	K2	14	21671277	G	0.921	0.798	0.898	0.712	0.024
	rs2080161	M2	7	13331150	T	0.981	0.758	0.689	0.820	0.000
	rs8137373	Y9	22	41729216	G	0.833	0.707	0.927	0.982	0.023
	rs1557553	Y8	22	44760984	C	0.949	0.904	0.714	0.786	0.094
	rs12402499	R2	1	101528954	G	1.000	0.919	1.000	1.000	0.258
	rs4792928	R8	17	42105174	T	1.000	0.960	0.345	0.804	0.195
Trial.	rs2069945	B1	20	33761837	CG	0.153 / 0.796	0.480 / 0.420	0.680 / 0.277	0.036 / 0.536	0.766 / 0.234
	rs4540055	H1	4	38803255	AC	0.069 / 0.514	0.793 / 0.010	0.301 / 0.068	0.536 / 0.250	0.605 / 0.000
	rs5030240	V1	11	32424389	CA	0.278 / 0.389	0.712 /0.055	0.228 / 0.039	0.093 / 0.278	0.258 / 0.023

En la Tabla 13 se recogen los diseños de los *primers* de PCR y la concentración de los mismos en el *mix* de *primers*. Las reacciones se llevaron a cabo en un termociclador GeneAmp® PCR System 9700 o 2700 (AB) bajo las siguientes condiciones: 10 min a 95°C; 32 ciclos de 30 s a 95°C, 40 s a 62°C y 1 min a 72°C; con una extensión final de 20 min a 72°C.

Tabla 13. Diseño de *primers* de PCR de los marcadores incluidos en el ensayo y concentraciones de los mismos en el *mix* de *primers*.
CI: código interno. Conc.: concentración.

SNP	CI	*Primer forward*	*Primer reverse*	Conc. (µM)
rs10483251	K2	GCACGTTCTTAACCTTGGCTAT	TTCTGAATATCCCACCCACAA	0.40
rs12142199	R4	AGGCCTTGATGTGCTTGAAC	CGAGAAGGCCAACCACTACT	0.30
rs1229984	Y4	ATTCTGTAGATGGTGGCTGTAGGA	CTGCCTCATGGCCTAAAATCA	0.75
rs12402499	R2	TGAAGGGTATTACTAGTGGC	TTGACAGACTTCTGCTTTTG	1.35
rs12498138	Y5	TCTTCTTCAGGGAATCCTGT	GAGTTACATAGGATTTGCGAG	1.75
rs12594144	K1	CCTACAAGACCACCCACCAG	GGACCCATGGTCATTCCATA	0.40
rs1426654	Y6	AATTCAGGAGCTGAACTGCC	TGTTCAGCCCTTGGATTGTC	1.50
rs1557553	Y8	TAATACAAGAGCCGCCTGGA	CTTGCAAGGAACTGCAGCTAT	0.65
rs16891982	S1	GAATAAAGTGAGGAAAACACGGAGT	GTTTCTCATCTACGAAAGAGGAGTC	0.50
rs17822931	R3	CCTAGAGTCCCCCAAACCTC	CACTTCTGGGCATCTGCTTC	0.50
rs1871534	S3	ACATCCTGCAGACCTTCCTG	CAGACCTTGGGCGTCAGAT	0.65
rs2069945	B1	GCAAACCTTGGCTCTGCTAC	CCTTTCCCCAGTGGCTTAAT	0.60
rs2080161	M2	GAGTATGATATAATTTTGTTCCTGCTG	TGGACTTTATGGGTTGTTGTTTT	0.50
rs2139931	Y10	AGTCTTGGCTAGGGCGTTAGTA	CTCCTAGTCATGGTTGATGTGG	1.50
rs2789823	Y2	AGAGGGCTTCTGTTCACACC	ATGCACCACTACTGTCCAAG	0.25
rs2814778	Y7	AACCTGATGGCCCTCATTAGT	ATGGCACCGTTTGGTTCAG	0.40
rs3751050	R1	GAAGGCTCCCAACTCGTTAG	GTCATTAAAGTCAACCTAGGC	1.35
rs3827760	R6	TGCTCAGCTCCACGTACAAC	CTCTTCAGGCCGAAGCTCT	0.30
rs4540055	H1	TGTGCCTCTGATCACTTTTGAATAC	CCTAGCCAACTCCAGAGTTCAT	0.40
rs4657449	R7	CCCCTCGGGAGAAAACATAG	TTCTAGAGTTGAATGAGGGTCAGA	0.85
rs4749305	R5	CAGCCCAACCTACTCCTCTG	TCCCTACAAAGTGGCAAACC	1.25
rs4792928	R8	TCTCTCAGGATATCCCTTTGG	AAAATCTTGATTCTGTATCGCAGTC	3.00
rs5030240	V1	CCAAAGTGCCAGGATCACAG	TCCCTAGAAATCCTTCAGCC	0.85
rs6054465	Y11	TATGGCCTCAGGTTCTCCAC	CACATGATCTCACCGTTTCCT	2.00
rs6437783	Y1	GCAATGAGATTAGTTGCACTGG	ATTATATGCCCACCCTGCTC	0.30
rs715605	R10	CCCAGCTAGGGCTAGACACC	TCAAAGACTGAGCCATGCAC	0.40
rs8072587	S2	TGGCAACCTCACATGGTAGA	CCAGGGGAGGTAGAAAGAGG	2.00
rs8137373	Y9	CCAGAGCTTTGCAGCACTTT	CAAGGACGCAGCTCTCTCA	2.00
rs9522149	Y3	AGAAAGGAGAGGAAACACCG	TCAGCAACTTCTAGTCCTCG	0.30
rs9809818	M1	TGTGTGGTTTTCTCAGCGAC	AGCATGGTATGAGCACTGAG	5.00
rs9908046	R9	CCTTGGCATGTTCCTCTCTC	TCAGAGGAATTAGAAAGGCCTAAA	0.30

Para la purificación post-PCR se combinaron 2,5 µL de producto de PCR con 1 µL de 1:3 Illustra™ ExoStar™ 1-Step (GE Healthcare). Se incubó a 37°C durante 45 min y seguidamente se inactivó el enzima a 85°C durante 15 min.

La reacción de SBE, con un volumen final de 3 µL, consta de:

- 1,25 µL de SNaPshot® Multiplex Ready Reaction Mix (1:2)
- 0,75 µL de *mix* de sondas de SBE
- 1 µL de producto purificado de PCR

En la Tabla 14, se recogen los diseños de sondas para SBE y su concentración en el *mix* de sondas. Las condiciones de la reacción de SBE comprenden 33 ciclos de 10 s a 96°C, 5 s a 59°C y 30 s a 60°C.

Tabla 14. Diseño de sondas de SBE para los marcadores incluidos en el ensayo, tamaño y concentración de las mismas en el *mix* de sondas. Las bases en minúscula corresponden a secuencias utilizadas como modificadores de la movilidad. Conc.: concentración.

SNP	CI	Sonda SBE	Tamaño	Conc. (μM)
rs10483251	K2	gtgccacgtcgtgaaagtctgacaaGGAAAAAGTTATGTGACCAGATT	48	0.45
rs12142199	R4	aacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTCAAACATGTTCCTCTGCAC	63	0.35
rs1229984	Y4	caactgactaaactaggtgccacgtcgtgaaagtctgacaaACCACGTGGTCATCTGTG	59	1.20
rs12402499	R2	aggtgccacgtcgtgaaagtctgacaaCTGCTTTTGATTTCAAGTATCAGT	51	1.70
rs12498138	Y5	caactgactaaactaggtgccacgtcgtgaaagtctgacaaGGGAATCCTGTTATTCACATTA	63	4.00
rs12594144	K1	cacgtcgtgaaagtctgacaaGCTCCCACCCTGAAAAAGA	40	1.00
rs1426654	Y6	ctctctctctctctctctctctctctctctctctctctctctctctctttCGCTGCCATGAAAGTTG	67	3.00
rs1557553	Y8	taaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCCAAAGCCCCTGGAAAAA	93	0.80
rs16891982	S1	cacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGGTTGGATGTTGGGGCTT	80	1.00
rs17822931	R3	aactgactaaactaggtgccacgtcgtgaaagtctgacaaCTGCATTGCCAGTGTACTC	59	0.80
rs1871534	S3	ctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCTGGCAGTGGGTGCA	105	1.80
rs2069945	B1	aaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTATTATTAACCCCAGTCTACATG	76	1.10
rs2080161	M2	cacgtcgtgaaagtctgacaaTTTTTTGTTTTTTTTTTGCACTCATCA	48	0.80
rs2139931	Y10	acaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTTCGTGTTGATGAGAAAATTTCA	109	1.10
rs2789823	Y2	aaactaggtgccacgtcgtgaaagtctgacaaGGAGGTGAGCTTCACGGGG	51	1.10
rs2814778	Y7	agtctgacaactaggtgccacgtcgtgaaagtctgacaactaggtgccacgtcgtgaaagtctgacatCTCATTAGTCCTTGGCTCTTA	89	0.80
rs3751050	R1	ctctctctctctctctctctcttgtTTAGGAGAGTTGAGACATC	44	1.40
rs3827760	R6	actaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTGGCGCCACGTTTTCACA	89	1.00
rs4540055	H1	ctctctctctctctctctctctctctctctctctctctctctctctctctctctcGAAGCAGTGATCAGCAC	72	0.50
rs4657449	R7	aaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGAGCTAAGGAAAGATACGTG	93	0.50
rs4749305	R5	caacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCAGTAAATAGTAACTCCATCTTC	67	1.30
rs4792928	R8	caactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCGCAGTCTACTAGTTGTCC	101	1.80
rs5030240	V1	ctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctcCACAGGAGTGAGCCACTGC	84	1.30
rs6054465	Y11	tgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCACATGCAAAATCAGGATAATAATG	113	1.60
rs6437783	Y1	tgccacgtcgtgaaagtctgacaaCTGGTTGAGGCACACTATTA	44	0.90
rs715605	R10	gaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCACCCTAAGGGGACAGA	113	0.90
rs8072587	S2	aactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCAGTCTCCTGCCCGGC	97	2.00
rs8137373	Y9	acaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGAGTGTTTTGTGGGCCTC	101	2.30
rs9522149	Y3	ctctctctctctctctctctctctctctctgacaatctGAGGTCCTTGCAGCTCC	55	1.00
rs9809818	M1	tctctctctctctctctctctccTCCTAATAAGAGCTGGC	40	3.00
rs9908046	R9	agtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGGAGGTAGGAGCACCCA	109	0.60
			$(NH_4)_2SO_4$:	4.00 mM

El producto total de la reacción de SBE se combinó con 1 µL de 1:2 Illustra™ Shrimp Alkaline Phosphatase (GE Healthcare) para la purificación, se incubó a 37°C durante 80 min y se inactivó la enzima a 85°C durante 15 min.

Los productos de SBE purificados se prepararon para CE añadiendo 1 µL del producto a 9,5 µL de Hi-Di™ Formamide (AB) y 0,25 µL de GeneScan™ 120 LIZ® Size Standard (AB). La electroforesis se llevó a cabo en un secuenciador ABI Prism 3130*xl* Genetic Analyser, con capilares de 36 cm y polímero POP-4™ bajo condiciones estándar. Los electroferogramas se visualizaron usando AB GeneMapper® ID Software v. 3.2.1.

4.1.1.3 Análisis de la variación poblacional en los SNPs seleccionados

Los valores de PSD y Divergencia entre pares de poblaciones se calcularon mediante el portal Snipper, usando la opción de *cross-validation*[27] y marcando los perfiles como p. ej. AFR vs. no-AFR o comparando cada par de poblaciones. Los valores obtenidos en Snipper representan Divergencia de Shannon y se convirtieron a valores de I_n (Rosenberg *et al.* 2003) multiplicándolos por 0,693. Asimismo, mediante el portal Snipper se realizó la *cross-validation* de las poblaciones de referencia y el cálculo de las LRs de las clasificaciones de perfiles frente a un *training set* que incluye los datos de las poblaciones de referencia o eligiendo entre las clasificaciones disponibles en el portal.

Los análisis poblacionales en STRUCTURE v. 2.3.4 (Pritchard *et al.* 2000) se realizaron siguiendo recomendaciones publicadas previamente (Porras-Hurtado *et al.* 2013). Los parámetros escogidos consistieron en cinco iteraciones (desde K=1 a K=9) con 100000 *burnin steps* y 100000 *MCMC steps*, frecuencias alélicas correlacionadas y modelo *admixture* (sin POPFLAG al analizar únicamente las poblaciones de referencia; con POPFLAG=1 para los individuos de las poblaciones de referencia y POPFLAG=0 para los individuos de las poblaciones del set test y *admixed* cuando se analizan conjuntamente). El valor óptimo de K se estimó computando los valores mediante el portal Structure Harvester (Earl y vonHoldt 2012) y siguiendo recomendaciones previamente publicadas (Evanno *et al.* 2005). Las gráficas que representan los coeficientes de ancestralidad de cada individuo para cada K se representaron mediante CLUMPAK v. 1.1 (Kopelman *et al.* 2015) o la combinación de CLUMPP v. 1.1.2 (Jakobsson y Rosenberg 2007) y distruct v. 1.1 (Rosenberg 2004). Los análisis de PCA se realizaron mediante el *software* R v. 3.1.2 (R Core Team 2014) y un *script* personalizado basado en el paquete SNPassoc (Gonzalez *et al.* 2007). Los cálculos de F_{ST} y las gráficas en los que se representan los valores se realizaron mediante el programa Arlequin v. 3.5 (Excoffier y Lischer 2010).

Para evaluar la capacidad de inferencia de ancestralidad del panel G-AIMs Nano se realizaron comparaciones con otros dos paneles de AIMs que comprenden 46 Indels (Pereira *et al.* 2012) y 34 SNPs (Fondevila *et al.* 2013). Se compilaron los datos de estos marcadores para las poblaciones de referencia y se aplicaron los análisis de STRUCTURE y PCA equivalentes a los descritos. Para el panel de 46 Indels, se usaron únicamente datos de los 44 marcadores listados por la Fase III del Proyecto 1000 Genomas.

[27] http://mathgene.usc.es/snipper/analysispopfile2_new.html

4.1.2 **Resultados**

4.1.2.1 Características y balance del panel

Los SNPs seleccionados muestran distribuciones de frecuencias muy contrastadas entre los 5 grupos poblacionales continentales. Cada uno de los 28 SNPs bialélicos presenta un alelo muy próximo a ser fijado (frecuencia entre 0,9 y 1) en al menos uno de los grupos poblacionales, tal y como se representa en el *raster plot* de la Fig. 49.

En la Fig. 50 se resumen las frecuencias alélicas de los marcadores en los 5 grupos poblacionales de referencia en forma de gráficos circulares. Los SNPs seleccionados se encuentran bien distribuidos en el genoma, con una distancia entre marcadores sinténicos suficiente para que no se encuentren ligados o en desequilibrio de ligamiento y puedan considerarse como estadísticamente independientes, tal y como se muestra en la Fig. 51.

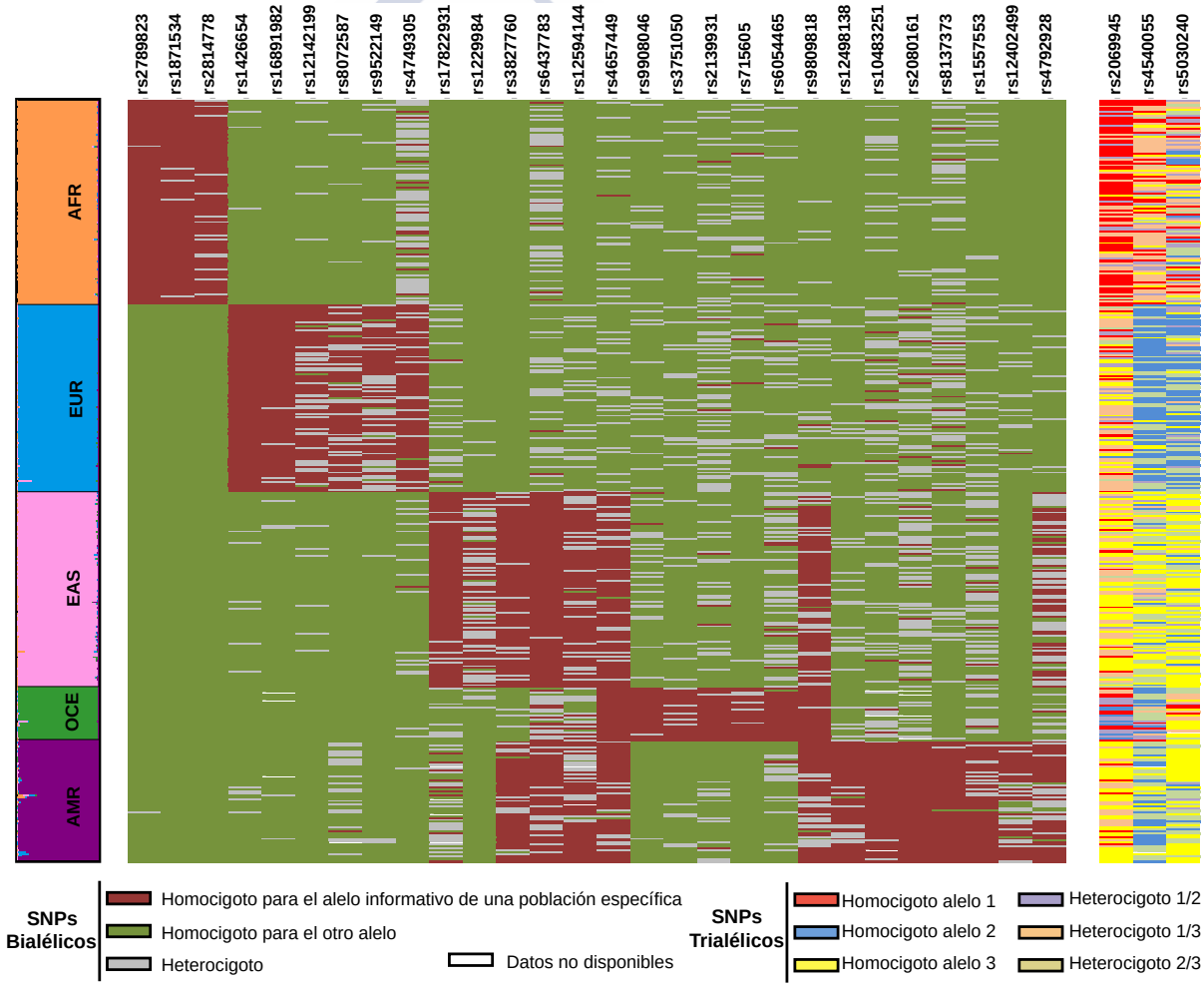


Fig. 49. *Raster plot* de los marcadores incluidos en el ensayo para los individuos de las poblaciones de referencia. Para evidenciar la ancestralidad biogeográfica (AFR, EUR, EAS, OCE, AMR) de cada individuo del set de referencia se incluyen los resultados del análisis de STRUCTURE para K=5.

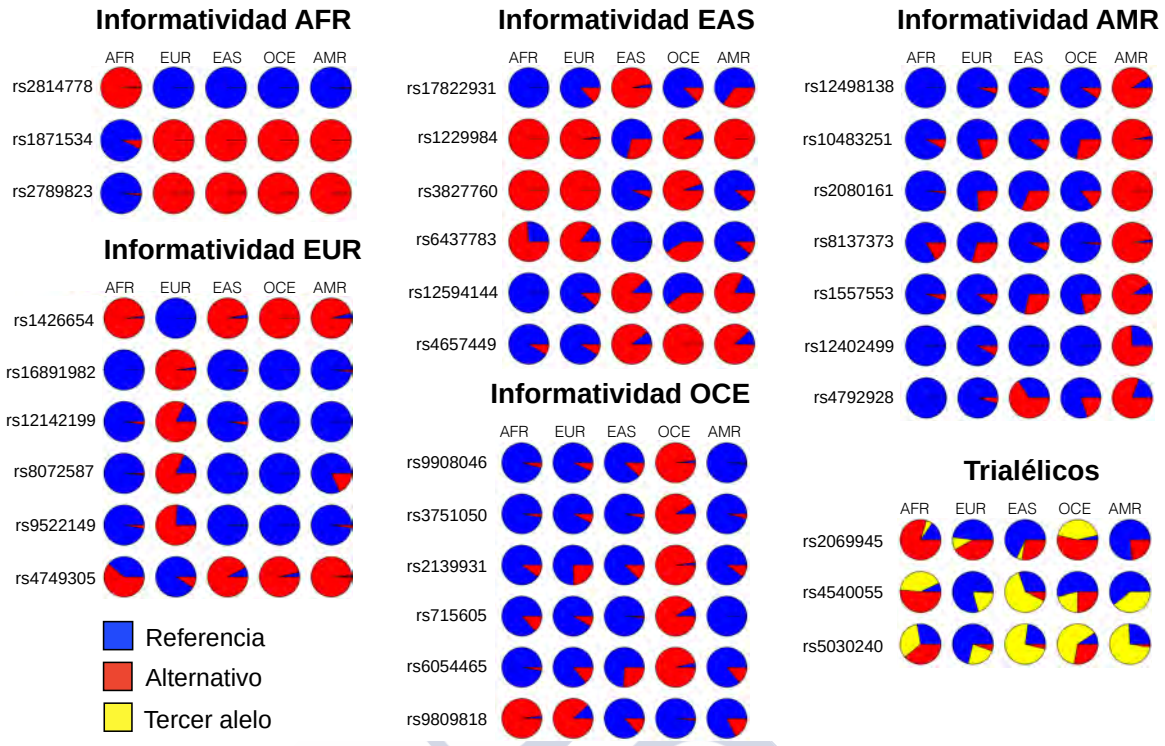


Fig. 50. Representación de las frecuencias alélicas de los SNPs en los 5 grupos poblacionales de referencia.

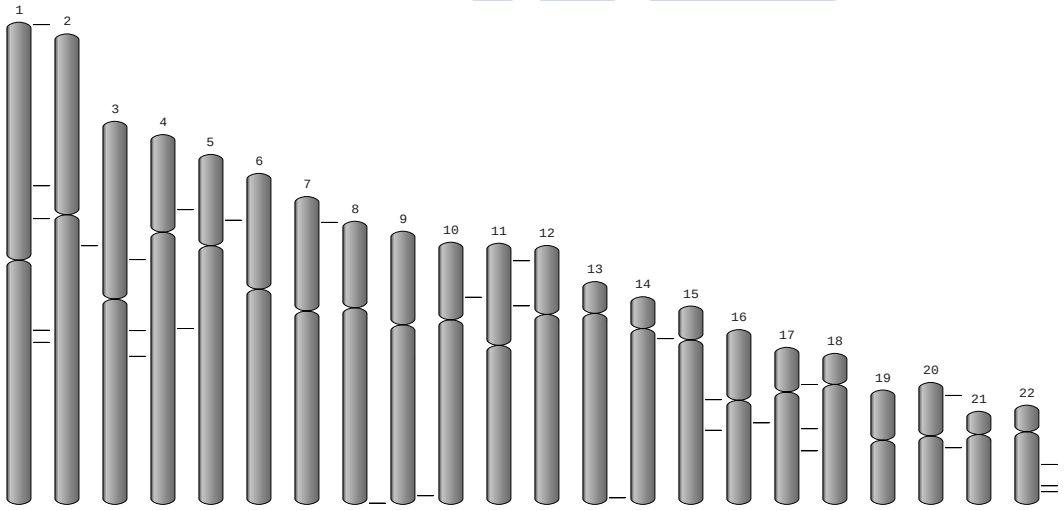


Fig. 51. Representación de los cromosomas autosómicos y la posición de los SNPs incluidos en el ensayo.

Para asegurar que los valores acumulados de PSD permanecen balanceados en los diferentes grupos poblacionales tras reducir el número de SNPs del ensayo desde los 128 marcadores del panel original hasta los 31 del ensayo para CE, se realizaron cálculos de los valores de PSD y Divergencia entre pares de poblaciones para cada marcador, y de sus valores acumulados para el conjunto de marcadores del panel. Estos resultados se recogen en la Tabla 15 y se representan en la Fig. 52. Los resultados indican unos valores de PSD acumulada en EAS de 3,39; relativamente bajos en comparación con la media de los grupos,

que alcanza el valor de 4,18. El valor acumulado de PSD de AFR presenta los niveles más elevados (4,74), que se mantienen comparables al resto de grupos poblacionales. La reducida diferenciación entre EAS vs. AMR y EAS vs. OCE se debe a las similitudes de estas poblaciones en la distribución de las frecuencias alélicas de ciertos SNPs como rs4657449 y rs9809818, que se pueden apreciar en la Fig. 49. El efecto es más acusado en la comparación EAS vs. AMR, poblaciones muy próximas, con poco grado de divergencia entre las mismas y que presentan distribuciones muy similares de los SNPs rs3827760, rs6437783 y rs12594144. Futuros ajustes del panel podrían abordar el desbalance en la PSD de EAS agregando marcadores específicamente informativos para esta población.

En la Fig. 53 se presentan los valores de F_{ST} y la media de las diferencias genotípicas entre pares de poblaciones, calculadas con Arlequin. Los resultados revelan valores altos de F_{ST} entre poblaciones de diferentes grupos y bajos entre poblaciones del mismo grupo, siguiendo la misma distribución que la media de las diferencias genotípicas interpoblacionales. Las poblaciones *admixed* del Proyecto 1000 Genomas presentan valores altos de media de diferencias genotípicas intrapoblacionales, tal y como se espera en base a los patrones complejos de variación que caracterizan al *admixture* entre poblaciones.

Tabla 15. Valores I_n de PSD y Divergencia entre pares de poblaciones de los marcadores incluidos en el ensayo para los grupos poblacionales de referencia.
Los SNPs se ordenan en función de su PSD individual dentro de cada grupo de informatividad.

SNPs		Valores PSD					Divergencia entre pares de poblaciones									
Infor.	SNP	AFR	EUR	EAS	OCE	AMR	AFR-EUR	AFR-EAS	AFR-OCE	AFR - AMR	EUR-EAS	EUR-OCE	EUR-AMR	EAS - OCE	EAS-AMR	OCE-AMR
AFR	rs2814778	0.672	0.131	0.134	0.083	0.108	0.663	0.663	0.634	0.656	0.000	0.672	0.000	0.000	0.000	0.001
	rs1871534	0.641	0.128	0.131	0.081	0.107	0.631	0.632	0.603	0.624	0.000	0.641	0.000	0.000	0.000	0.000
	rs2789823	0.565	0.121	0.123	0.076	0.100	0.556	0.557	0.527	0.548	0.000	0.565	0.000	0.000	0.000	0.000
EUR	rs1426654	0.117	0.622	0.093	0.081	0.069	0.641	0.001	0.000	0.003	0.611	0.117	0.595	0.001	0.000	0.002
	rs16891982	0.126	0.620	0.104	0.073	0.087	0.629	0.001	0.000	0.002	0.606	0.126	0.603	0.000	0.000	0.000
	rs12142199	0.076	0.402	0.068	0.061	0.083	0.393	0.000	0.000	0.002	0.383	0.076	0.423	0.001	0.003	0.000
	rs8072587	0.099	0.358	0.113	0.069	0.003	0.405	0.001	0.000	0.047	0.425	0.099	0.218	0.000	0.058	0.043
	rs9522149	0.063	0.353	0.092	0.055	0.056	0.334	0.004	0.001	0.000	0.377	0.063	0.341	0.002	0.003	0.000
	rs4749305	0.001	0.309	0.095	0.100	0.155	0.162	0.072	0.105	0.141	0.404	0.001	0.514	0.004	0.017	0.005
EAS	rs17822931	0.190	0.053	0.432	0.039	0.000	0.039	0.613	0.037	0.129	0.428	0.190	0.034	0.434	0.251	0.036
	rs1229984	0.089	0.070	0.320	0.018	0.071	0.002	0.336	0.018	0.000	0.313	0.089	0.001	0.238	0.328	0.015
	rs3827760	0.219	0.209	0.319	0.098	0.203	0.000	0.559	0.012	0.500	0.558	0.219	0.499	0.471	0.003	0.415
	rs6437783	0.078	0.153	0.268	0.001	0.104	0.010	0.359	0.057	0.223	0.459	0.078	0.311	0.157	0.031	0.062
	rs12594144	0.237	0.097	0.222	0.000	0.130	0.032	0.487	0.149	0.430	0.334	0.237	0.283	0.136	0.003	0.101
	rs4657449	0.176	0.164	0.177	0.210	0.130	0.000	0.380	0.497	0.364	0.376	0.176	0.360	0.017	0.000	0.022
OCE	rs9908046	0.020	0.008	0.000	0.528	0.042	0.002	0.010	0.562	0.006	0.003	0.020	0.015	0.463	0.030	0.626
	rs3751050	0.020	0.002	0.016	0.451	0.011	0.006	0.000	0.477	0.000	0.004	0.020	0.003	0.467	0.000	0.459
	rs2139931	0.017	0.002	0.011	0.433	0.014	0.019	0.000	0.480	0.000	0.013	0.017	0.019	0.458	0.000	0.480
	rs715605	0.000	0.003	0.039	0.422	0.041	0.003	0.030	0.345	0.036	0.015	0.000	0.020	0.502	0.001	0.516
	rs6054465	0.062	0.005	0.005	0.408	0.004	0.022	0.061	0.553	0.022	0.011	0.062	0.000	0.306	0.011	0.408
	rs9809818	0.245	0.120	0.172	0.227	0.102	0.021	0.446	0.603	0.399	0.319	0.245	0.276	0.026	0.002	0.042
AMR	rs12498138	0.085	0.034	0.020	0.011	0.443	0.011	0.020	0.024	0.519	0.002	0.085	0.437	0.000	0.401	0.387
	rs10483251	0.055	0.006	0.040	0.000	0.429	0.016	0.001	0.038	0.497	0.010	0.055	0.370	0.028	0.469	0.301
	rs2080161	0.144	0.007	0.001	0.036	0.424	0.064	0.091	0.044	0.624	0.003	0.144	0.366	0.033	0.314	0.462
	rs8137373	0.020	0.000	0.068	0.093	0.406	0.011	0.011	0.037	0.402	0.043	0.020	0.298	0.009	0.506	0.593
	rs1557553	0.076	0.042	0.000	0.002	0.325	0.004	0.053	0.031	0.436	0.030	0.076	0.380	0.003	0.219	0.270
	rs12402499	0.061	0.007	0.059	0.032	0.324	0.021	0.000	0.000	0.361	0.021	0.061	0.252	0.000	0.361	0.335
	rs4792928	0.171	0.111	0.108	0.011	0.178	0.008	0.297	0.064	0.413	0.239	0.171	0.350	0.113	0.014	0.199
Trial.	rs2069945	0.114	0.002	0.045	0.201	0.087	0.078	0.156	0.118	0.210	0.022	0.114	0.065	0.289	0.017	0.384
	rs4540055	0.217	0.148	0.055	0.024	0.084	0.355	0.147	0.142	0.300	0.130	0.217	0.027	0.098	0.063	0.101
	rs5030240	0.083	0.117	0.062	0.062	0.054	0.125	0.123	0.052	0.135	0.135	0.083	0.124	0.067	0.001	0.085
Valor acumulado		4.739	4.404	3.392	3.986	4.374	5.263	6.111	6.210	8.029	6.274	4.739	7.184	4.323	3.106	6.350

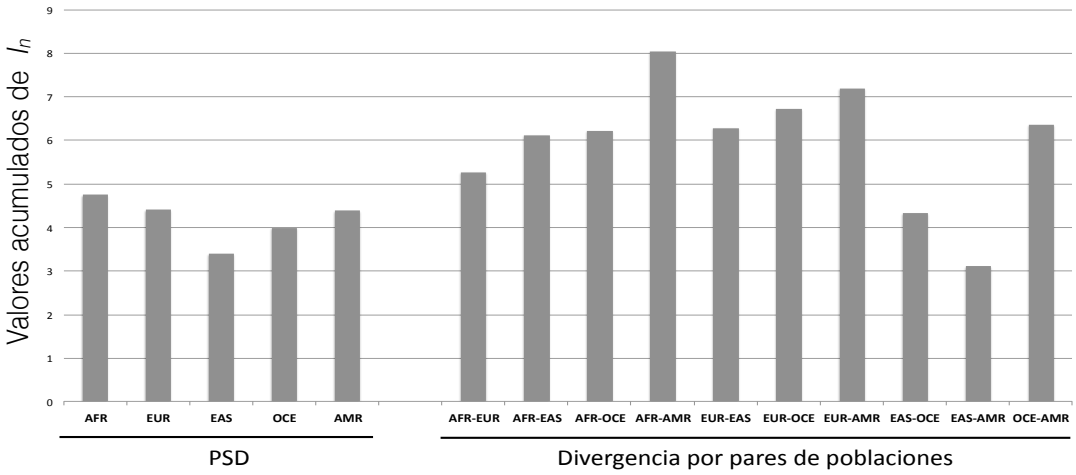


Fig. 52. Representación de los valores acumulados de I_n para PSD y Divergencia por pares de poblaciones.

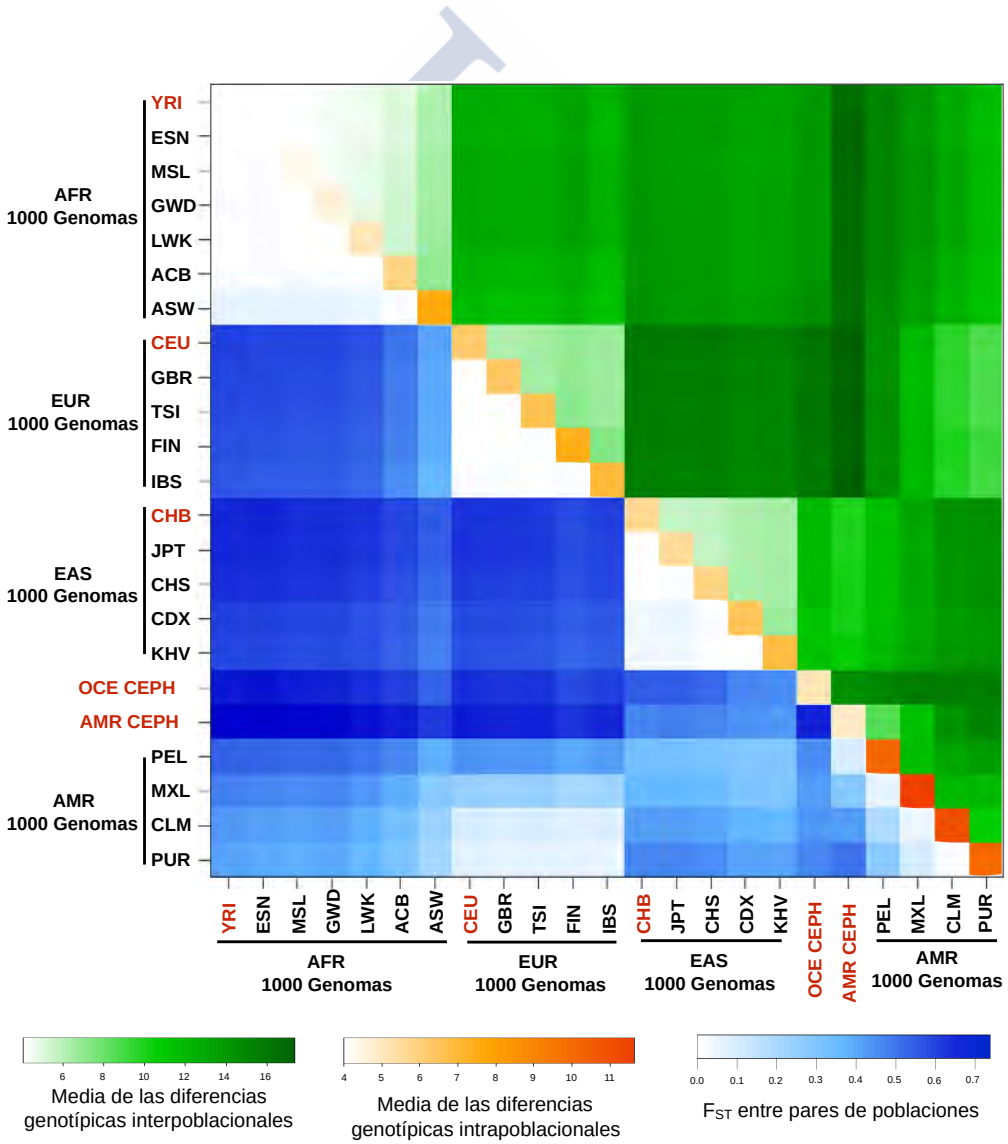


Fig. 53. Media de las diferencias genotípicas interpoblacionales e intrapoblacionales y valores de F_{ST} entre pares de poblaciones. Las poblaciones marcadas en rojo representan el set de referencia. Los identificadores de cada población se encuentran listados en la sección 4.1.1.1.

4.1.2.2 Capacidad del panel para la inferencia de ancestralidad

En la Tabla 16 se muestran los resultados de *cross-validation* de las poblaciones de referencia, con un 100% de éxito de asignación de ancestralidad para cada población. Además, el éxito de asignación permanece en el 100% en todos los grupos cuando se excluyen los 14 marcadores más informativos del panel (aquellos que presentan los niveles de divergencia más altos entre los 5 grupos de referencia), lo que indica que el panel es capaz de proporcionar altos niveles de informatividad incluso cuando cierta proporción de SNPs no puedan ser genotipados (p. ej. ADN *low template* o degradado).

El análisis de STRUCTURE del set de poblaciones de referencia (sin POPFLAG) produjo un patrón de 5 grupos que se corresponde adecuadamente con las 5 poblaciones de referencia. Los otros dos paneles de AIMs con los que se compara este set de marcadores, uno de 34 SNPs y otro de 46 Indels (datos para 44 Indels) permiten diferenciar 5 grupos, pero en ambos casos el número óptimo de grupos (K) estimados es menor de 5 –Fig. 54–.

Tabla 16. Resultados de *cross-validation* de las poblaciones de referencia, incluyendo los datos genotípicos de los marcadores del panel. Los valores resaltados en gris corresponden al porcentaje de individuos del set de referencia correctamente clasificados.

	AFR	EUR	EAS	OCE	AMR
Población de origen AFR	100.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Población de origen EUR	0.00 %	100.00 %	0.00 %	0.00 %	0.00 %
Población de origen EAS	0.00 %	0.00 %	100.00 %	0.00 %	0.00 %
Población de origen OCE	0.00 %	0.00 %	0.00 %	100.00 %	0.00 %
Población de origen AMR	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %

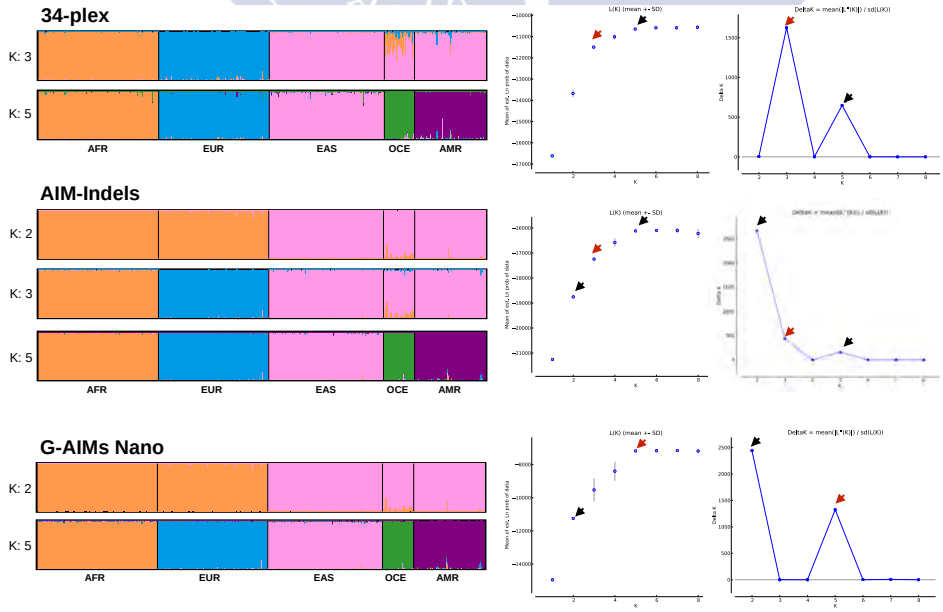


Fig. 54. Análisis STRUCTURE de tres paneles de AIMs para el set de poblaciones de referencia. Las gráficas de la derecha representan el Ln de la media de las estimaciones de la probabilidad de los datos y los valores de Delta K para cada uno de los posibles número de grupos –K– según el método propuesto por Evanno *et al.* (2005). Flechas rojas: K más probable, se muestran los resultados a la izquierda. Flechas negras: otros K altamente probables, de los que también se muestran sus resultados a la izquierda.

Los PCA revelan que el panel diseñado, en comparación con los otros ensayos, incrementa la separación entre las poblaciones de referencia y reduce la dispersión de los individuos de las mismas, tal y como se muestra en la Fig. 55. La separación es especialmente evidente en la representación PC2 vs. PC3, en la que las nubes de puntos que representan individuos de cada población no se solapan y se encuentran aproximadamente equidistantes.

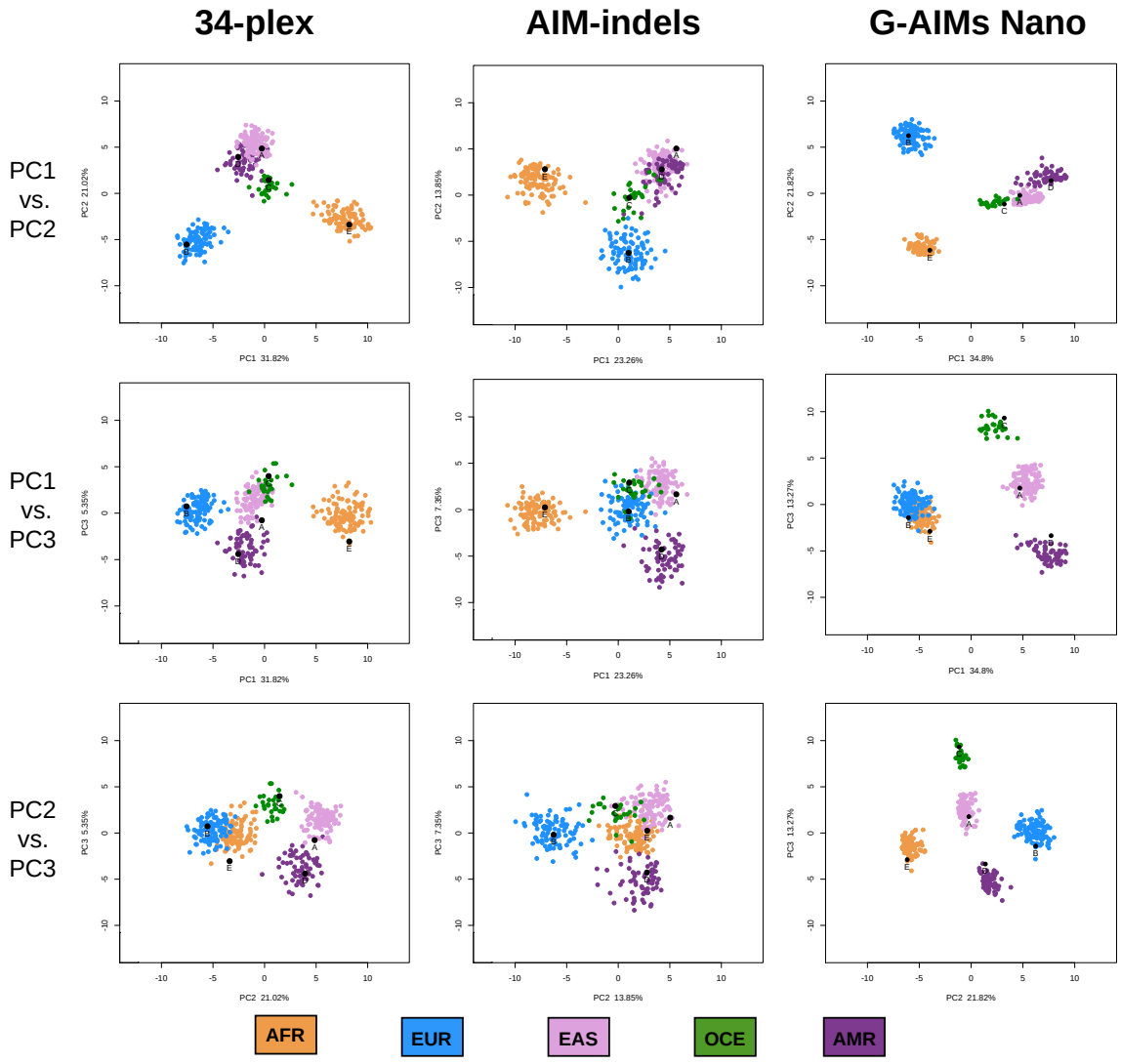


Fig. 55. Comparación de los análisis de PCA de los 3 paneles de marcadores para las poblaciones de referencia (se excluyen los marcadores trialélicos, de manera que el número de marcadores analizados es de 28 para G-AIMs Nano, 32 para 34-plex y 44 para AIM-Indels). Utilizando el PCA de las poblaciones de referencia se clasifican los controles del ejercicio de ancestralidad realizado por Santos *et al.* (2015).

El análisis STRUCTURE de los datos de las poblaciones del Proyecto 1000 Genomas escogidas como set test y *admixed* (marcadas como poblaciones de estudio: POPFLAG=0) produjeron patrones consistentes con los análisis realizados con panel EUROFORGEN

Global AIM-SNP (Phillips *et al.* 2014a, Phillips 2015). Así, las gráficas generadas mediante STRUCTURE usando los 31 marcadores del ensayo, que se muestran en la Fig. 56, se corresponden a las generadas usando los 128 SNPs del panel original, indicando que la población PEL presenta las mayores proporciones de coancestralidad nativo americana y la población PUR presenta predominantemente coancestralidad europea. Asimismo, en los PCA de las poblaciones *admixed*, presentados en la Fig. 56, los individuos se distribuyen entre las nubes de puntos que forman las poblaciones de referencia que contribuyen en la *admixture*, siendo más cercanos al grupo al que corresponde la mayor proporción de coancestralidad.

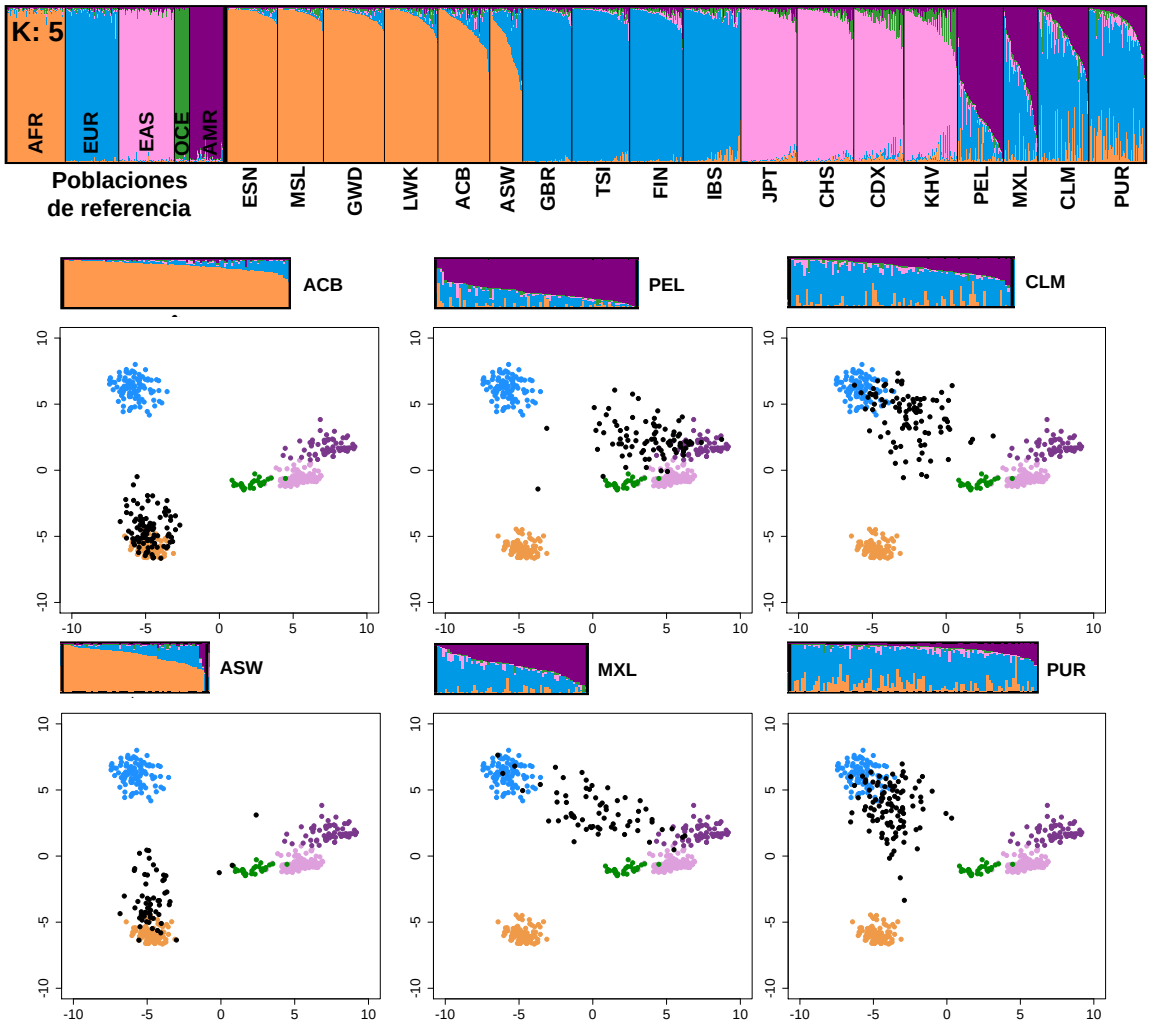


Fig. 56. Análisis STRUCTURE de las poblaciones *unadmixed* (set test) y *admixed* del Proyecto 1000 Genomas, marcadas como poblaciones de estudio frente al set de referencia. Las poblaciones *admixed* se acompañan de PCAs (PC1 vs. PC2) en los que se representan los individuos en negro frente a los de las poblaciones de referencia, que se representan en su color conforme al análisis STRUCTURE.

El ensayo de 31 SNPs se utilizó para estimar la ancestralidad de los controles de ADN utilizados en un ejercicio de colaboración de ancestralidad (Santos *et al.* 2015) que evalúa dos paneles de 34 SNPs (Fondevila *et al.* 2013) y 46 Indels (Pereira *et al.* 2012). Se analizaron

mediante PCA y Snipper los 5 controles (muestras A-E) de ancestralidad conocida correspondiente a cada uno de los cinco grupos poblacionales continentales. En los análisis PCA, presentados en la Fig. 55, cada uno de los controles se posiciona en el medio del grupo poblacional correspondiente a su ancestralidad conocida. Las LRs –*Likelihood ratios*– obtenidas mediante Snipper (análisis bayesiano) se muestran en la Tabla 17 y los resultados indican que se alcanzan unos valores altos para los 5 controles, en comparación con los otros 2 paneles. En la Fig. 57 se confirma, para los 5 controles del ejercicio de ancestralidad y el ADN control 9947A, que mediante este panel se alcanzan LRs altas con perfiles parciales, aunque no se obtengan los genotipos de los 14 marcadores más informativos del panel.

Tabla 17. LRs de las estimaciones de ancestralidad calculadas en Snipper para los controles de ancestralidad conocida A-E, en comparaciones de cinco grupos. *Todas las asignaciones fueron correctas excepto en la muestra A, clasificada erróneamente como AMR mediante 34-plex.

		Resultados de LR			
Muestra	Ancestralidad	34-plex	AIM-Indels	80 marcadores	G-AIMs Nano
A	EAS	6.8E+00*	5.5E+06	9.7E+06	2.9E+16
B	EUR	4.4E+16	1.7E+11	1.0E+28	1.3E+29
C	OCE	1.0E+07	4.0E+07	1.5E+14	1.5E+16
D	AMR	1.0E+05	1.2E+09	1.1E+14	4.2E+08
E	AFR	6.1E+19	2.8E+21	3.6E+40	2.0E+29

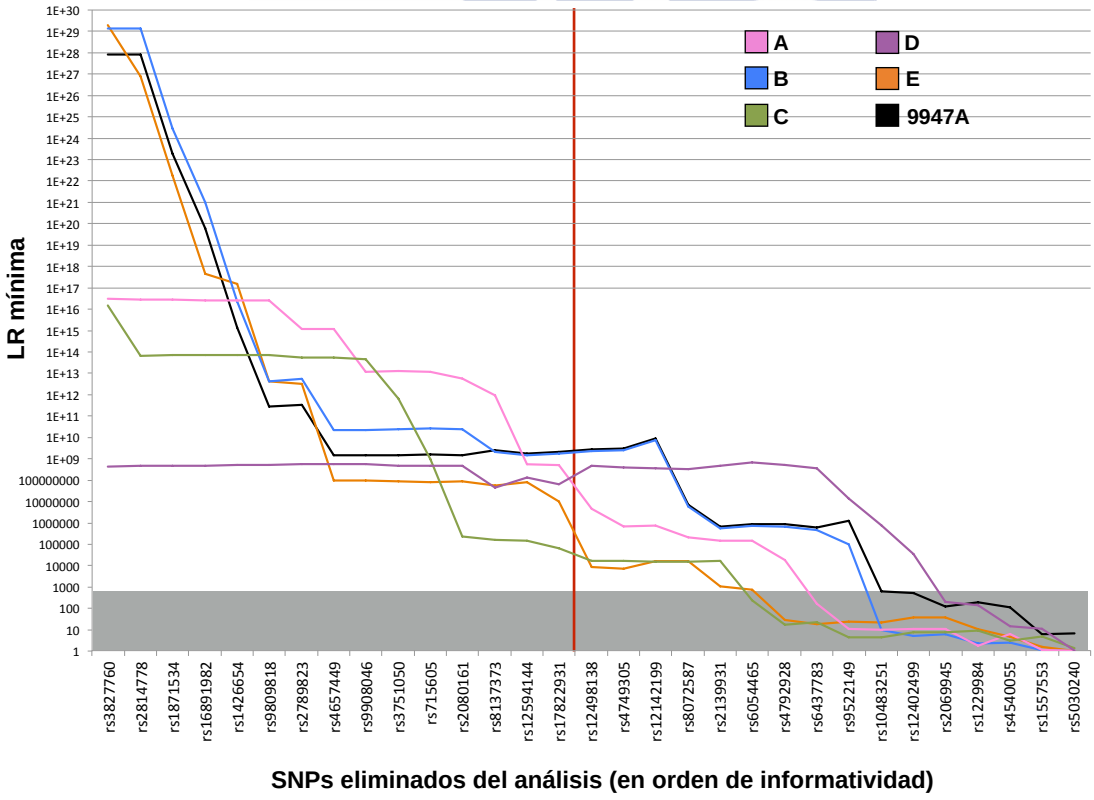


Fig. 57. LRs de las estimaciones de ancestralidad de 5 controles (A-E) y del control de ADN 9947A, eliminando progresivamente los SNPs más informativos del panel. El sombreado gris representa un valor umbral de LR de 1000, la línea roja representa el punto en el que la mitad de los marcadores son eliminados.

4.1.2.3 Evaluación forense del panel de SNaPshot

En la Fig. 58 se muestra un perfil de SNaPshot del control de ADN 9947A, con 1 ng de ADN inicial. Las diluciones seriadas de 9947A presentaron perfiles completos con 0,5 ng; 0,25 ng; 0,125 ng y 0,064 ng de ADN inicial. Los análisis de 0,032 ng y 0,016 ng de ADN presentaron fenómenos de *drop-out* de *locus* y alélico; aunque se obtuvieron perfiles con más del 80% de los marcadores.

La sensibilidad forense del ensayo fue evaluada mediante extractos de ADN de biopsias, huesos (cráneo, fémur y tibia) y dientes identificados como degradados o inhibidos, que mostraban perfiles de STRs en los que se amplificaban entre el ~35-95% de los STRs. En los perfiles de SNaPshot se analizaron con éxito entre el ~20-70% de los marcadores del panel, produciendo LRs de asignación >60000 en todos los casos.

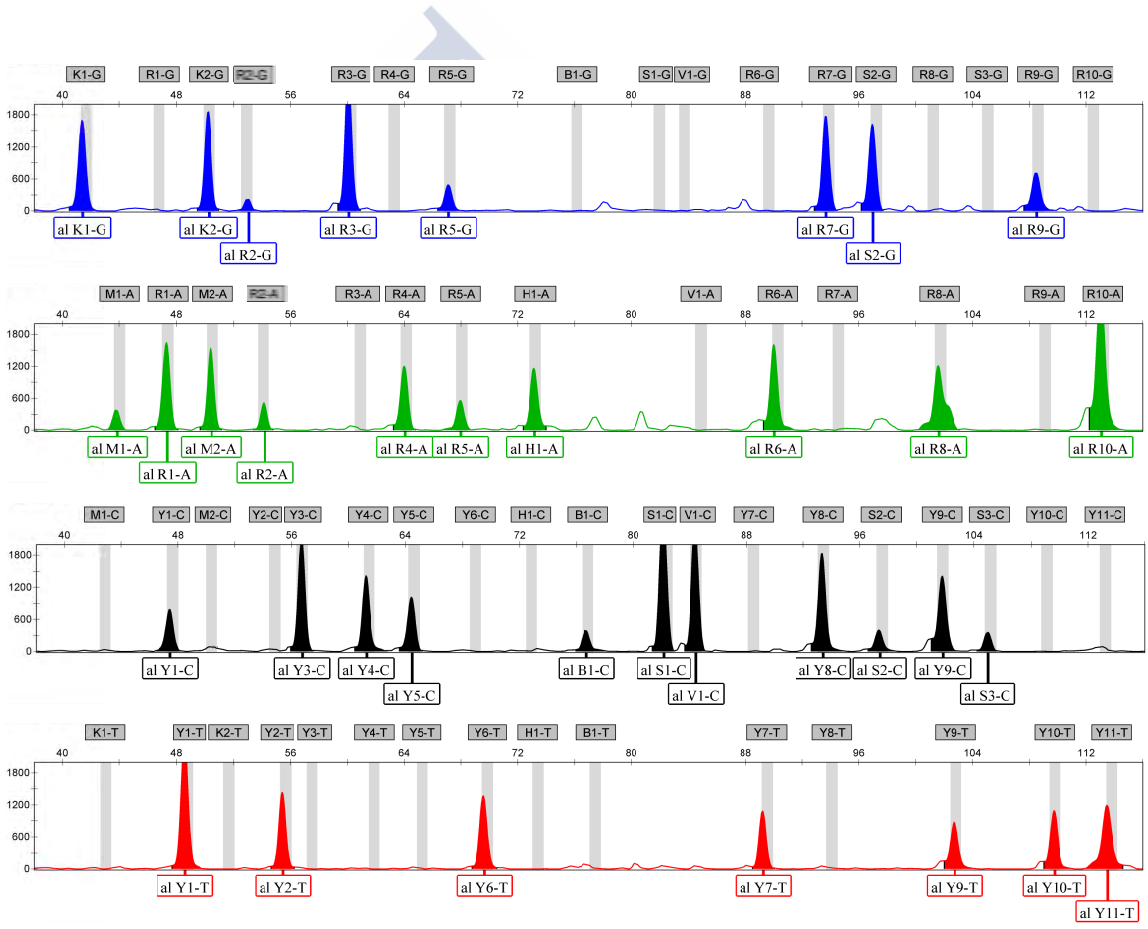


Fig. 58. Perfil de SNaPshot del control de ADN 9947A (1 ng) para los 31 SNPs incluidos en el panel.

4.1.3 **Discusión**

Durante la construcción del panel EUROFORGEN Global AIM-SNP (Phillips *et al.* 2014a) se abordó la posibilidad de diseñar pequeños subconjuntos de marcadores para su análisis en SNaPshot. Este estudio presenta una recopilación de 31 SNPs, que representan

principalmente los SNPs más informativos del panel EUROFORGEN Global AIM-SNP. El subconjunto de 31 SNPs mantiene la capacidad de diferenciar los 5 grupos poblacionales definidos continentalmente. De esta manera, el panel G-AIMs Nano extiende las comparaciones de tres grupos ya abordadas en otros ensayos SNaPshot (Lao *et al.* 2010, Fondevila *et al.* 2013, Rogalla *et al.* 2015a, Wei *et al.* 2015), adicionando dos grupos poblacionales: nativo americanos y oceánicos. Cada uno de estos grupos poblacionales contribuye a los patrones de *admixture* de las extensas regiones que ocupan. Por este motivo, una de las prioridades fue mantener el balance de las PSD acumuladas de las poblaciones, aunque el objetivo no es sencillo teniendo en cuenta la reducción del número de marcadores en un 75%. La PSD acumulada de EAS disminuyó desproporcionadamente en comparación con la del resto de grupos. La adición de 1 o 2 AIM-SNPs que eleven selectivamente la PSD de EAS podría solventar este inconveniente.

La selección de SNPs informativos para AMR y OCE conlleva el uso de datos procedentes del panel HGDP-CEPH, con tamaños poblacionales pequeños en comparación con los datos del Proyecto 1000 Genomas. En este sentido, puede haberse producido un sesgo al escoger los 31 marcadores de este análisis, de manera que el poder para diferenciar poblaciones todavía no caracterizadas de AMR y OCE puede verse reducido. No obstante, es improbable que este sesgo haya impedido recoger SNPs más divergentes que los 22 AMR y los 28 OCE presentes en el panel EUROFORGEN Global AIM-SNP. La inclusión de los 5 SNPs más informativos del panel EUROFORGEN Global AIM-SNP para AMR y OCE, con alelos muy próximos a ser fijados, asegura que el panel sea casi igual de informativo para los 5 grupos poblacionales.

Además, los valores de LR obtenidos para asignaciones en base a 5 grupos poblacionales superan a los obtenidos con otros paneles de AIMs diseñados para CE –ver Tabla 17 –. Cuando se analizan las mismas muestras control de ancestralidad conocida con los 31 SNPs de G-AIMS Nano, 34 SNPs o 46 Indels; se obtienen LR más altas con el conjunto de 31 SNPs que con la combinación de los otros 80 marcadores para 3 de las 5 poblaciones, y todas las LR obtenidas superan con un margen considerable (entre 3 y 16 órdenes de magnitud) a las obtenidas con el conjunto de 34 SNPs. Así, el ensayo G-AIMs Nano se presenta como la mejor opción para la predicción de ancestralidad en sistemas de CE para aquellos laboratorios que no dispongan de plataformas de MPS. La combinación de este panel con el de 46 Indels, que permite la detección de mezclas de ADN (Pereira *et al.* 2012), constituye la aproximación más completa disponible con sistemas de CE.

Como conclusión, este panel ha recopilado un set de AIMs altamente informativos y con valores de PSD acumulados bien balanceados para 5 grupos poblacionales definidos continentalmente. Esta característica minimiza el sesgo en las estimaciones de proporciones de coancestralidad cuando se analizan individuos *admixed*. Además, el ensayo SNaPshot presenta altos niveles de sensibilidad (se obtienen perfiles completos con tan solo 64 pg de ADN) y permite el análisis de muestras de ADN degradado, convirtiéndolo en un ensayo de alta utilidad en rutina forense.

4.2 ADAPTACIÓN A ION PGM™ Y VALIDACIÓN DEL PANEL GLOBAL AIM-SNP

En este trabajo se adapta el panel teórico EUROFORGEN Global AIM-SNP (Phillips *et al.* 2014a) a la plataforma de MPS Ion PGM™. El rendimiento del panel final se evaluó en 5 laboratorios diferentes siguiendo un esquema de validación sencillo. Además, se realizaron análisis de ancestralidad incluyendo todas las nuevas poblaciones de la Fase III del Proyecto 1000 Genomas y un total de 551 individuos de 14 nuevas poblaciones de estudio. Los resultados se encuentran publicados en la referencia:

> Eduardoff M, Gross TE, Santos C, **de la Puente M**, Ballard D, Strobl C, Børsting C, Morling N, Fusco L, Hussing C, Egyed B, Souto L, Uacyisrael J, Syndercombe-Court D, Carracedo Á, Lareu MV, Schneider PM, Parson W, Phillips C (2016). *"Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™."* Forensic Sci Int Genet 23: 178-189.

4.2.1 **Material y métodos**

4.2.1.1 Muestras de ADN y datos poblacionales

Para los estudios de concordancia de genotipos se seleccionaron siete controles de ADN Coriell de ancestralidad conocida, representando cada uno de los cinco grupos poblacionales diferenciados por el panel EUROFORGEN Global AIM-SNP (Phillips *et al.* 2014a). Los controles de ADN Coriell incluyen: el trío NA06994, NA07000 y NA07029 (EUR); NA18498 (AFR); HG00403 (EAS); NA10540 (OCE) y NA11200 (AMR). El uso de estos controles permite la comparación de los genotipos obtenidos mediante tres tecnologías de MPS diferentes, dado que sus genotipos se encuentran en las bases de datos del Proyecto 1000 Genomas –obtenidos mediante la plataforma Illumina HiSeq (The Genomes Project Consortium 2012)– y Complete Genomics –obtenidos mediante secuenciación por ligado (Drmanac *et al.* 2010)–. El control de ADN 9947A, disponible en la mayoría de laboratorios de genética forense, se incluyó como control universal.

Los datos genotípicos poblacionales de los SNPs que componen el panel Global AIM-SNP se obtuvieron a partir de tres fuentes: (i) datos de la Fase III del Proyecto 1000 Genomas (The Genomes Project Consortium 2015); (ii) datos de los análisis del panel HGDP-CEPH de Li *et al.* (2008) obtenidos a través del portal SPSmart (Amigo *et al.* 2008) y (iii) genotipos generados en este estudio para muestras de poblaciones estudio de interés. Las descripciones de las poblaciones se exponen en la Tabla 18.

Como poblaciones de referencia –poblaciones 1-6 en la Tabla 18– se utilizaron poblaciones del Proyecto 1000 Genomas con bajos niveles de *admixture* y variación intrapoblacional: ESN para AFR, GBR para EUR y JPT para EAS; además de dos conjuntos de poblaciones del panel HGDP-CEPH: un conjunto de dos poblaciones para OCE y otro de cinco poblaciones para AMR. Para los análisis que incluyen las poblaciones de SAS (Sur de Asia) se eligió como referencia la población GIH. Las poblaciones ESN, GBR, JPT y GIH se

escogieron como referencia por presentar los niveles más bajos de media de las diferencias genotípicas intrapoblacionales en los análisis que se presentan en la Fig. 66. Mediante esta estrategia de selección de poblaciones de referencia se pretende compensar el contraste entre los tamaños poblacionales de los datos del Proyecto 1000 Genomas y los del panel HGDP-CEPH (de los que se compilan los grupos poblacionales de OCE y AMR) de manera que los análisis de STRUCTURE no se vean afectados (Onogi *et al.* 2011).

El resto de poblaciones *admixed* y *unadmixed* incluidas en el Proyecto 1000 Genomas se utilizaron como set test –poblaciones de 7-28 en la Tabla 18–. Las poblaciones *unadmixed* de SAS se incluyeron en aquellos análisis que evalúan la capacidad del set Global AIM-SNP para diferenciar EUR y SAS, grupos poblacionales menos divergentes entre sí que los cinco grupos poblacionales para los que el panel fue diseñado originalmente.

Las 14 poblaciones de estudio –poblaciones 29-42 de la Tabla 18– comprenden un total de 551 muestras recogidas bajo consentimiento informado por escrito, seleccionadas para expandir el ámbito geográfico de las bases de datos –ver Fig. 59–. Para todas las poblaciones, se obtuvo la aprobación de los comités de ética de las instituciones pertinentes, que se sometió a una evaluación posterior por parte de la Comisión Europea. La extracción de ADN de las muestras se realizó mediante QIAamp® DNA Mini Kit (Qiagen), EZ1 DNA Investigator Kit (Qiagen) y otros métodos descritos previamente (Egyed *et al.* 2007).

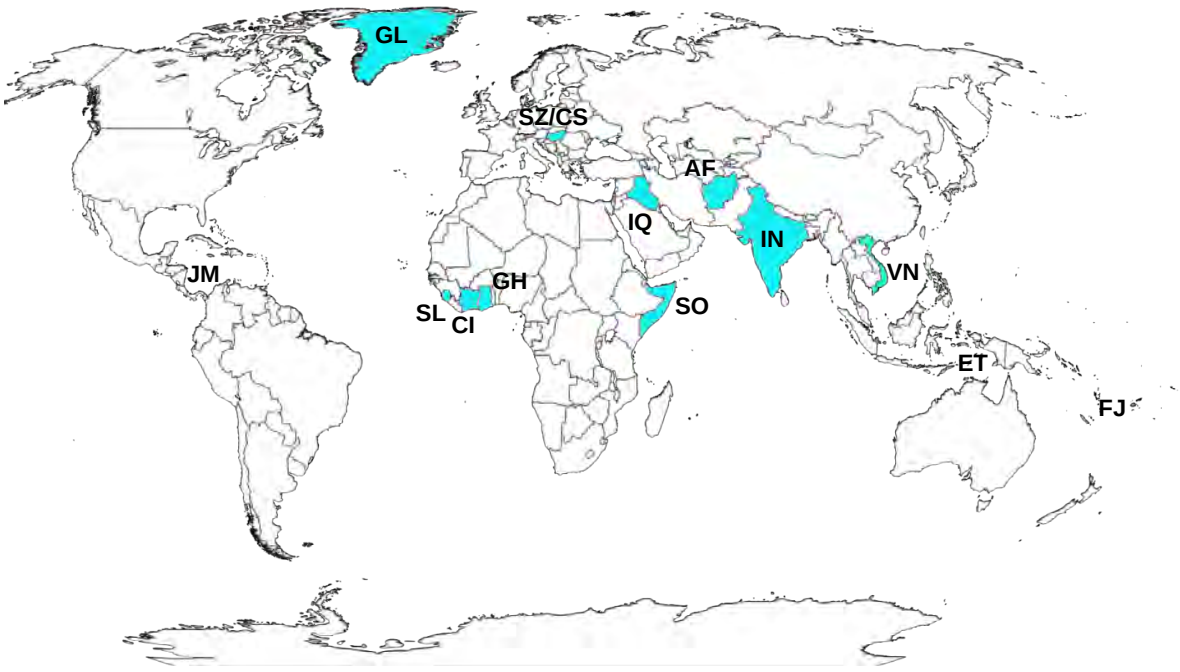


Fig. 59. Situación geográfica de las 14 poblaciones de estudio.

Tabla 18. Poblaciones incluidas en los análisis de ancestralidad. Pob: código de población. N.º: número de población. *N*: número de individuos. ME: Oriente Medio. P1000G: Proyecto 1000 Genomas. *Adm.*: *Admixed*.

Set	N.º	Pob	Grupo	*N*	Datos	Descripción
Referencia	1	ESN	AFR	99	P1000G	*Esan in Nigeria*
	2	GBR	EUR	91	P1000G	*British in England and Scotland*
	3	JPT	EAS	104	P1000G	*Japanese in Tokyo, Japan*
	4	OCE	OCE	28	HGDP-CEPH	*17 Papuan (New Guinea); 11 Melanesian (Bougainville)*
	5	AMR	AMR	64	HGDP-CEPH	*14 Karitiana/8 Surui (Brazil); 21 Maya/14 Pima (Mexico); 7 Piapoco (Colombia)*
	6	GIH	SAS	103	P1000G	*Gujarati Indian from Houston, Texas*
Test	7	YRI	AFR	108	P1000G	*Yoruba in Ibadan, Nigeria*
	8	MSL	AFR	85	P1000G	*Mende in Sierra Leone*
	9	GWD	AFR	113	P1000G	*Gambian in Western Divisions in the Gambia*
	10	LWK	AFR	99	P1000G	*Luhya in Webuye, Kenya*
	11	CEU	EUR	99	P1000G	*Utah Residents with North and Western European ancestry*
	12	TSI	EUR	107	P1000G	*Toscani in Italia*
	13	IBS	EUR	107	P1000G	*Iberian Population in Spain*
	14	FIN	EUR	99	P1000G	*Finnish in Finland*
	15	CHS	EAS	105	P1000G	*Southern Han Chinese*
	16	CHB	EAS	103	P1000G	*Han Chinese in Beijing, China*
	17	CDX	EAS	93	P1000G	*Chinese Dai in Xishuangbanna, China*
	18	KHV	EAS	99	P1000G	*Kinh in Ho Chi Minh City, Vietnam*
	19	PJL	SAS	96	P1000G	*Punjabi from Lahore, Pakistan*
	20	BEB	SAS	86	P1000G	*Bengali from Bangladesh*
	21	STU	SAS	102	P1000G	*Sri Lankan Tamil from the UK*
	22	ITU	SAS	102	P1000G	*Indian Telugu from the UK*
	23	ACB	*Adm.*	96	P1000G	*African Caribbeans in Barbados*
	24	ASW	*Adm.*	61	P1000G	*Americans of African Ancestry in SW USA*
	25	PEL	*Adm.*	85	P1000G	*Peruvians from Lima, Peru*
	26	MXL	*Adm.*	64	P1000G	*Individuals with Mexican Ancestry from Los Angeles USA*
	27	CLM	*Adm.*	94	P1000G	*Colombians from Medellin, Colombia*
	28	PUR	*Adm.*	104	P1000G	*Puerto Ricans from Puerto Rico*
Estudio	29	GH	AFR	35	Este trabajo	*Ghana*
	30	SL	AFR	45	Este trabajo	*Sierra Leone*
	31	CI	AFR	33	Este trabajo	*Ivory Coast*
	32	JM	*Adm.*	45	Este trabajo	*Jamaica*
	33	SO	AFR	46	Este trabajo	*Somalia*
	34	CS	EUR	49	Este trabajo	*Csango from Hungary*
	35	SZ	EUR	50	Este trabajo	*Szeklers from Hungary*
	36	IN	SAS	32	Este trabajo	*India*
	37	AF	SAS	35	Este trabajo	*Afghanistan*
	38	IQ	ME	34	Este trabajo	*Kurdish from Iraq*
	39	VN	EAS	32	Este trabajo	*Vietnam*
	40	ET	OCE	50	Este trabajo	*East Timor*
	41	FJ	OCE	12	Este trabajo	*Fiji*
	42	GL	AMR	53	Este trabajo	*Greenlanders*

4.2.1.2 Preparación de las muestras de ADN para MPS

Para la preparación de las librerías se utilizó el kit Ion AmpliSeq™ Library 2.0 siguiendo las recomendaciones del fabricante[28]. Todas las librerías fueron realizadas con cantidades iniciales de ADN entre 1-10 ng, menos las que corresponden a las de evaluación de la sensibilidad forense del ensayo (diluciones seriadas, ADN degradado y mezclas de ADN). La mayoría de las muestras se amplificaron mediante el protocolo de volumen completo –*full volume*– propuesto por el fabricante, mientras que una proporción de las mismas se amplificó con la mitad de volumen en todas las reacciones –*half volume*–. Independientemente del volumen de reacción, la PCR de captura de los fragmentos de interés constó de 18 ciclos para 10 ng de ADN inicial o 21 ciclos para <10 ng de ADN inicial. Las librerías fueron cuantificadas, siguiendo las recomendaciones del fabricante, con los kits Ion Library TaqMan® Quantitation o Qubit® ds DNA HS Assay. El término "+5" hace referencia a los 5 ciclos de amplificación de la librería que se aplican antes de la cuantificación mediante Qubit® según el protocolo y que no se requieren si se realiza la cuantificación mediante TaqMan®. Las librerías de los controles de ADN y de un total de 82 muestras de las poblaciones de estudio fueron amplificadas mediante los protocolos de *full volume* y *half volume*, permitiendo realizar comparaciones entre ambos protocolos.

Siguiendo las recomendaciones del fabricante, se purificaron las librerías con *beads* magnéticas AMPure XP y se incluyeron *barcodes* Ion Xpress™ Barcode Adapters para individualizar cada muestra del *run*. Las librerías se agruparon en un *pool* equimolar de entre 8-12 pM para la preparación del molde de secuenciación mediante el kit Ion OneTouch™200 Template v2. Siguiendo las recomendaciones del fabricante[29] los resultados de la preparación del molde de secuenciación se evaluaron con el kit Ion Sphere™ Quality Control. La secuenciación se realizó utilizando el kit Ion PGM™ Sequencing 200 v2 y loschips Ion 316™ o 318™ v2, siguiendo las recomendaciones del fabricante[30].

La sensibilidad del panel fue evaluada utilizando diluciones del control de ADN Coriell NA07000, de manera que se construyeron librerías a partir de las siguientes cantidades iniciales de ADN: 10 ng (18 ciclos), 1 ng (21 ciclos), 500 pg (21 y 25 ciclos), 250 pg (21 y 25 ciclos), 100 pg (21 y 25 ciclos), 50 pg (25 y 25+5 ciclos) y 10 pg (25+5 ciclos). Además, se analizaron dos muestras de ADN extraído de restos esqueléticos –Hueso-1 y Hueso-2– previamente evaluadas para indicadores de degradación e inhibición. Los ensayos de Quantifiler® Duo sugieren que Hueso-1, que presenta valores de C_T>31 para el control interno de PCR –IPC: *internal PCR control*–, se encuentra inhibido. Los resultados de Quantifiler® Trio presentaron ratios de concentración fragmento corto/largo de 2,45 para Hueso-1 y 5,49 para Hueso-2, sugiriendo degradación en ambos casos, aunque los resultados del IPC no indicaron inhibición. Ambas muestras fueron amplificadas con 21+5 ciclos.

[28] Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user guide. April (2014).
[29] Thermo Fisher Scientific, Life Technologies: Ion OneTouch™ 200 Template Kit v2 user guide (2014).
[30] Thermo Fisher Scientific, Life Technologies: Ion PGM™ 200 Sequencing Kit user guide (2014).

En cada caso, se siguieron las recomendaciones del fabricante para el análisis de muestras limitadas para calcular el número de ciclos de la PCR de captura y decidir si se aplicaban 5 ciclos adicionales de amplificación a las librerías con bajo rendimiento (<100 pM).

Para la evaluación de mezclas de ADN, se prepararon mezclas artificiales de dos controles de ADN de ancestralidad conocida: NA07000 (SS1 –EUR– 8,30 ng/µL) y NA18498 (SS2 –AFR– 7,76 ng/µL), cuantificadas con el kit Qubit® ds DNA HS Assay, en ratios de volumen de 1:9, 1:3, 1:1; 3:1 y 9:1. Cada librería de las mezclas de ADN fue secuenciada dos veces en *runs* diferentes (réplicas A y B).

4.2.1.3 Análisis de datos

Los resultados brutos de secuenciación fueron analizados mediante el *software* Torrent Suite™ 4.2 (TS) y el *plugin* HID_SNP_Genotyper v. 4.2 (Genotyper), aplicando los parámetros de línea germinal –*Germline*– de baja rigurosidad[32] –*low stringency*–. Se construyeron dos archivos de extensión .bed para los SNPs incluidos en el set Global AIM-SNP: *target* –identifica los amplicones– y *hotspot* –identifica los SNPs–, basados en el genoma de referencia hg19 (GRCh37). Los dos archivos generados por Genotyper (.csv y .vcf) para cada análisis se procesaron con R v. 3.0.3 (R Core Team 2014) y/o Excel.

4.2.1.4 Criterios de exclusión de marcadores o muestras y corrección de genotipos

Los análisis de concordancia de los controles de ADN y las comprobaciones preliminares de los genotipos de las muestras de las poblaciones de estudio indicaron ciertos problemas de genotipado para 4 de los SNPs incluidos en el panel y algunas de poblaciones de estudio, tales como variantes específicas de población que provocan una alta proporción de genotipos NN o *no-call* y alineamientos erróneos causados por tractos homopoliméricos. El análisis detallado de las secuencias brutas –visualización de los archivos .bam y . bai en IGV (Robinson *et al.* 2011)– y de los archivos .vcf permitió la corrección manual de los genotipos de los SNPs rs595961, rs6875659 y rs12402499. No obstante, el SNP rs2080161 fue excluido de los análisis –ver sección 4.2.2.5.2–.

Los genotipos de las poblaciones de estudio se revisaron para prevenir sesgos en los análisis poblacionales derivados de SNPs o muestras con bajo rendimiento. Los SNPs con bajo rendimiento se definieron como aquellos que presentan tasas de *no-call* más altas de lo habitual, a causa de una baja calidad de la secuenciación o al bajo número de secuencias obtenidas –*coverage*–. En las muestras de estudio de buena calidad se observan pocos *no-call*, pero las de baja calidad presentan tasa altas de *no-call*. Por este motivo, se eliminaron de los análisis poblacionales las muestras con <95% de genotipos asignados –muestras con bajo rendimiento–, aplicando un grado de *stringency* mayor que el umbral de 90% empleado en un estudio similar (Nassir *et al.* 2009). Para la mayoría de las muestras de bajo rendimiento, el

[31] Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user guide. April (2014).
[32] Thermo Fisher Scientific, Life Technologies: Torrent Suite™ software 4.2 user guide (2015).

alto número de *no-calls* se debe a un bajo *coverage* promedio de la muestra. Por ello, se inspeccionaron los archivos .vcf de las muestras con un *coverage* promedio <200x para asegurar la fiabilidad de los genotipos asignados. Además, todos los genotipos obtenidos con un *coverage* menor de 30x fueron confirmados o rechazados mediante la revisión de los archivos .vcf. Los umbrales de *coverage* se establecieron en un mínimo de 20x para heterocigotos y de 10x para homocigotos, con valores mínimos por alelo y cadena –*strand*– de secuenciación (*forward* o *reverse*) de 10x o 5x, respectivamente. Los genotipos se corrigieron manualmente a NN –*no-call*– cuando los valores de *coverage* no alcanzaron los umbrales, las frecuencias alélicas no estaban balanceadas (40-60% en heterocigotos, >90% para homocigotos) o las lecturas de cada *strand* no estaban balanceadas (entre el 25-75%). Así, p. ej., un genotipo homocigoto con 15x de *coverage* total, 3x de *forward strand* y 12x de *reverse strand* se corrige a NN; mientras que un genotipo homocigoto con 15x de *coverage* total, 7x de *forward strand* y 8x de *reverse strand* se mantiene.

4.2.1.5 Análisis de ancestralidad poblacional

Mediante el portal Snipper[33], se calcularon los valores de Divergencia de Shannon para cada SNP, en comparaciones por pares y para cada grupo poblacional frente al resto. Los valores de Divergencia de Shannon se convirtieron en el parámetro I_n multiplicándolos por 0,693 y los valores obtenidos para cada SNP se acumularon para obtener los valores globales de PSD y Divergencia entre pares de poblaciones. Asimismo, el portal Snipper se utilizó para obtener las LRs de las clasificaciones, aportando un *training set* con los datos de los individuos de las poblaciones de referencia.

Los análisis de STRUCTURE v. 2.3.4 (Pritchard *et al.* 2000) se realizaron siguiendo recomendaciones previas (Porras-Hurtado *et al.* 2013). Se asumieron entre 1-9 poblaciones (K=1 a K=9) y se realizaron 5 réplicas para cada valor de K. Los análisis se llevaron a cabo considerando el modelo de ancestralidad *admixture* con frecuencias alélicas correlacionadas. Cada *run* de análisis constó de 100,000 *burnin steps* y 100,000 *MCMC steps* para alcanzar unas estimaciones precisas de las probabilidades posteriores. Los valores óptimos de K se estimaron computando los resultados con Structure Harvester (Earl y vonHoldt 2012) y siguiendo guías publicadas previamente (Evanno *et al.* 2005). Las gráficas de proporciones de ancestralidad se construyeron combinando los *software* CLUMPP v. 1.1.2 (Jakobsson y Rosenberg 2007) y distruct v. 1.1 (Rosenberg 2004). Los análisis PCA se realizaron con el *software* R v. 3.1.2 y un *script* personalizado.

Las estimaciones de las frecuencias alélicas poblacionales, la media de las diferencias genotípicas intrapoblacionales e interpoblacionales, los valores de F_{ST} y el test exacto para el equilibrio Hardy-Weinberg se realizaron en Arlequin v. 3.5 (Excoffier y Lischer 2010).

[33] http://mathgene.usc.es/snipper/analysispopfile2_new.html.

4.2.2 **Resultados**

4.2.2.1 Diseño del ensayo para Ion PGM™ y tasa de conversión a MPS

El set EUROFORGEN Global AIM-SNP constituye uno de los primeros paneles forenses personalizados adaptado para MPS por TFS a través del diseño de una PCR *multiplex* de captura con *primers* AmpliSeq™. Se debe señalar que diferentes paneles de SNPs de identificación previamente establecidos han sido adaptados a MPS con éxito por parte de las casas comerciales de las plataformas; no obstante, en este caso se han realizado ciertos ajustes en los marcadores finalmente incluidos en el ensayo. La tasa de conversión del set EUROFORGEN Global AIM-SNP puede ser tomada como un indicador de la capacidad de adaptación a MPS de sets teóricos de marcadores escogidos con diferentes fines. Durante el diseño original del panel se llevó a cabo un escrutinio detallado de las secuencias contexto de los 128 marcadores seleccionados. Pese a ello, tres SNPs (rs5757362, rs2282107 y rs7246968) no pudieron ser incluidos en la PCR *multiplex* de captura dado que se sitúan en regiones repetitivas que conllevan problemas de unión inespecífica de los *primers* y, consecuentemente, podrían elevar la cantidad de lecturas *off-target*. Además, rs2282107 y rs7246968 se sitúan próximos a largos tractos homopoliméricos, lo que impide una secuenciación eficaz. Dos de los tres SNPs sustitutos que se propusieron: rs2837352 (para rs5757362) y rs16946159 (para rs2282107) presentan propiedades parecidas en cuanto al poder de diferenciación de las poblaciones, pero se sitúan en regiones diferentes, por lo que pudieron ser incorporados con éxito al ensayo. El SNP sustituto rs7250345 (para rs7246968) constituía un buen candidato, ya que se encuentra en la misma región y posee frecuencias alélicas casi idénticas; no obstante, los motivos repetitivos de la secuencia contexto comprenden elementos SINE muy largos, por lo que toda la región se veía afectada y finalmente se incorporó el sustituto rs11048128.

En la Tabla 25 se presentan los valores actualizados de PSD acumulada del panel, tras la sustitución de los SNPs durante el proceso de adaptación a MPS y la exclusión del SNP atípico rs2080161. Al sustituirse tan solo 3 SNPs de los 128 del ensayo inicial, los cambios en los valores acumulados de PSD son mínimos, pese a que se aprecia una bajada del valor de EAS. En general, una tasa de conversión del 97,6% parece bastante adecuada; no obstante, análisis posteriores revelaron problemas de alineamientos en los SNPs rs595961, rs6875659 y rs2080161, causados por la presencia de tractos homopoliméricos. Este tipo de problemas de alineamiento provocaron la exclusión del SNP rs2080161 y podrían haber sido detectados durante la fase de diseño de *primers*.

4.2.2.2 Concordancia del genotipado

La concordancia del genotipado se evaluó en tres niveles: (i) comparando los genotipos de muestras control de ADN idénticas, preparadas y analizadas en cinco laboratorios diferentes –concordancia interlaboratorio: 37 análisis–; (ii) comparando los genotipos obtenidos para los controles de ADN Coriell a través de la plataforma Ion PGM™ con los de

las bases de datos públicas del Proyecto 1000 Genomas y Complete Genomics; y (iii) comparando análisis de las mismas muestras con protocolos *full volume* y *half volume* –8 análisis de controles de ADN y 82 de muestras de poblaciones de estudio–. Para que no interfieran las diferencias en el número de análisis y *no-calls* de cada muestra, las tasas de concordancia se calculan sobre el número de genotipos asignados.

4.2.2.2.1 Concordancia interlaboratorio

La concordancia entre los genotipos asignados alcanzó el 99,81% (4707/4716), con una tasa de *no-call* del 0,42% (20/4736). Se observaron discordancias en 3 muestras diferentes en los SNPs rs6875659, rs2080161, rs9934011 y rs9908046 –ver Tabla 19–, de manera que la tasa de discordancia es del 0,19% (9/4716).

4.2.2.2.2 Concordancia entre genotipos de Ion PGM™ y bases de datos online

Para cuatro de los controles de ADN Coriell: NA06994, NA07000, HG00403 y NA18498 –19 análisis–, la base de datos del Proyecto 1000 Genomas lista los 128 SNPs del panel Global AIM-SNP. La tasa de concordancia de los genotipos asignados resultó en un 99,84% (2414/2418), con una tasa de *no-call* del 0,58% (14/2432) causada por 8 SNPs diferentes –ver sección 4.2.2.5.3 y Tabla 24–. Se encontraron cuatro genotipos discordantes (tasa de discordancia del 0,16%) para el SNP rs6875659, en cuatro análisis diferentes del mismo control de ADN –ver Tabla 19–.

Cinco de los controles de ADN Coriell: NA06994, NA07000, HG00403, NA18498 y NA07029 se encuentran listados en Complete Genomics, de manera que las tasas de concordancia se basan en 2944 genotipos de 23 análisis. La tasa de *no-call* es del 0,47% (14/2944); y el 99,86% de los genotipos asignados (2926/2930) son concordantes. El mismo SNP que causa las discordancias con el Proyecto 1000 Genomas causa las discordancias con Complete Genomics –Tabla 19– resultando en una tasa del 0,14%.

4.2.2.2.3 Concordancia entre protocolos full volume y half volume

Si se comparan los protocolos *full volume* y *half volume* para los 8 controles de ADN se obtiene una tasa de concordancia del 99,57% (1/2038 genotipo discordante en el SNP rs2080161) y una tasa de *no-call* de 0,49% (10/2048).

Además, se compararon los resultados de ambos protocolos para 82 muestras de siete poblaciones de estudio. Un total de 41 muestras presentan diferencias entre ambos protocolos (tanto *no-calls* como discordancias). Treinta y dos muestras (71,19%) presentan <3 diferencias en los SNPs rs595961 y rs2080161 –ver sección 4.2.2.5.2– o en SNPs con altas tasas de *no-call* (>4 *no-calls* en 551 genotipos): rs4979274, rs499827, rs310644 y rs1366220 –ver sección 4.2.2.5.3–. Entre estos 6 SNPs, tan solo rs595961 produjo genotipos discordantes (11), mientras que el resto produjo un *no-call* en uno de los análisis. En general, se observaron 31 diferencias debidas a *no-call*, que en 14 (45%) de los casos se producen en el protocolo *full volume*. En 7 muestras, se encontraron >5 diferencias entre ambos protocolos,

principalmente *no-calls* debidos a un bajo *coverage* en el protocolo de *half volume*. Una muestra AFR produjo un genotipo discordante para rs2814778 en el protocolo *half volume*, mostrando un 11% de lecturas T en *forward strands* en contraste con el 100% de lecturas T en ambas *strands* obtenidas mediante el protocolo *full volume*. La visualización de las secuencias obtenidas para esta muestra en IGV sugiere incorporaciones erróneas durante la PCR de emulsión o un alineamiento erróneo de las secuencias. En resumen, el protocolo *full volume* produjo más genotipos fiables de SNPs para 14 muestras; mientras que el protocolo *half volume* para 11 (44%). Así, el genotipado con el protocolo *half volume* constituye una estrategia factible que reduce costes, asumiendo un bajo riesgo de pérdida de calidad de los datos.

Tabla 19. Análisis de concordancia entre los genotipos asignados mediante Ion PGM™ en 5 laboratorios para los controles de ADN Coriell y las bases de datos online.

SNP	Control de ADN Coriell	N.º de análisis discordantes	Genotipo discordante	Genotipo concordante entre laboratorios	Complete Genomics	Proyecto 1000 Genomas
rs6875659	CTR_NA18498	4/5	AG	AA	AA	AA
rs2080161	CTR_NA11200	3/5	AC	CC	–	–
rs9934011	CTR_NA11200_lab3	1/5	CT	CC	–	–
rs9908046	CTR_NA10540_lab3	1/5	CT	TT	–	–

4.2.2.2.4 Corrección manual de los genotipos

Los análisis de concordancia indican problemas de genotipado para cuatro SNPs. La causa de los problemas fue investigada para determinar las medidas adecuadas para la corrección manual de los genotipos o excluir el SNP si no se puede garantizar la obtención de genotipos fiables. Los genotipos de los SNPs rs2080161, rs595961 y rs6875659 están afectados por tractos homopoliméricos cercanos a la posición del SNP, que causan alineamientos erróneos de las secuencias y/o el truncamiento de las mismas antes de alcanzar la posición del SNP.

El amplicón del SNP rs2080161 –Fig. 60– comprende varios tractos poli-T en ambas direcciones de secuenciación. Como consecuencia de una serie de alineamientos erróneos de los tractos poli-T se generan genotipos no fiables, lo que provocó la exclusión de este SNP de los análisis posteriores e indica que debe ser excluido del panel.

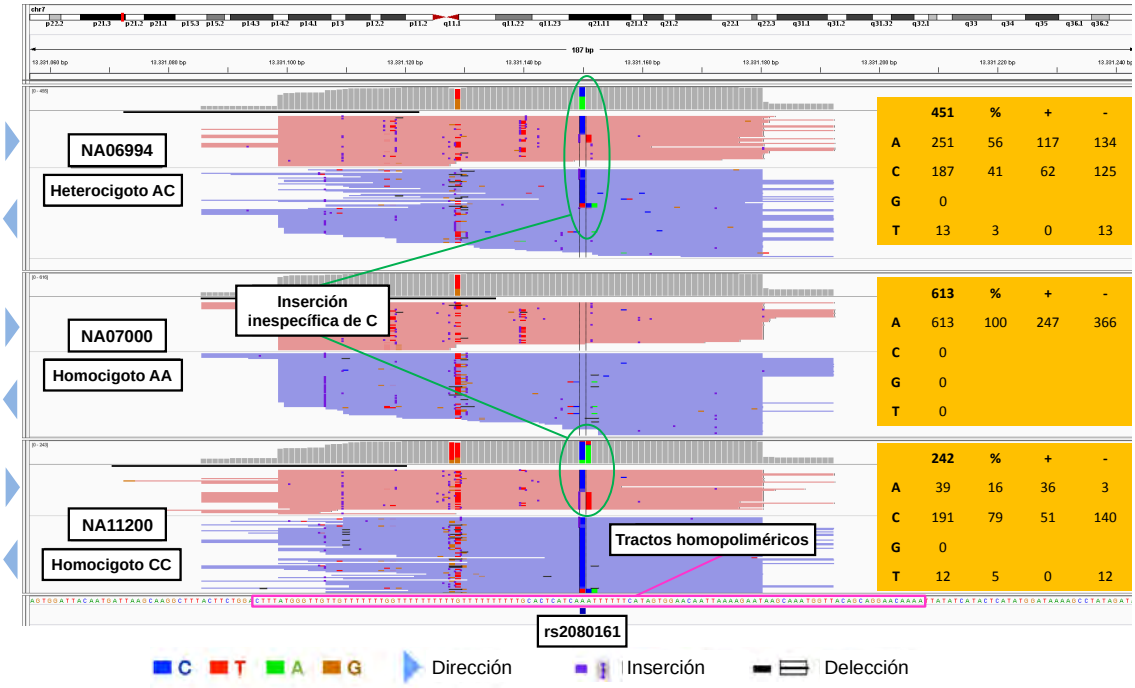


Fig. 60. Visualización en IGV del SNP A/C rs2080161 en diferentes muestras. Las tablas recogen información sobre las lecturas en la posición del SNP. Este SNP se excluye del panel.

Los genotipos de los SNPs rs6875659 y rs595961 –Fig. 61 y Fig. 62, respectivamente– se ven afectados en una de las cadenas, mientras que la complementaria presenta un *coverage* balanceado de los alelos en muestras heterocigotas. Así, los genotipos de rs6875659 y rs595961 fueron corregidos manualmente basándose en las lecturas desglosadas en los archivos .vcf e infiriendo los genotipos a partir de las lecturas *reverse* y *forward*, respectivamente. En el SNP rs6875659, las muestras de ancestralidad AFR (donde el alelo más frecuente es A) son más propensas a presentar genotipos incorrectos que el resto de muestras de las poblaciones de estudio (en las que el alelo más frecuente es G). El alelo A alarga el tracto poli-A adyacente, situándose como la última de las 5 As de la cadena *forward* y causando las discordancias observadas en 4/5 análisis del control de ADN Coriell NA18598 (AFR) –ver Tabla 19–.

Los genotipos corregidos de los controles de ADN Coriell para los SNPs rs6875659 y rs595961 se corresponden con los de las bases de datos, avalando la aplicabilidad de las correcciones manuales. Esta corrección manual de genotipos no es directa, sencilla o deseable; no obstante, se espera que futuras mejoras del *software* de análisis solventen la mayoría de los problemas de alineamiento o incluyan nuevas opciones como inferir genotipos a partir de una única cadena.

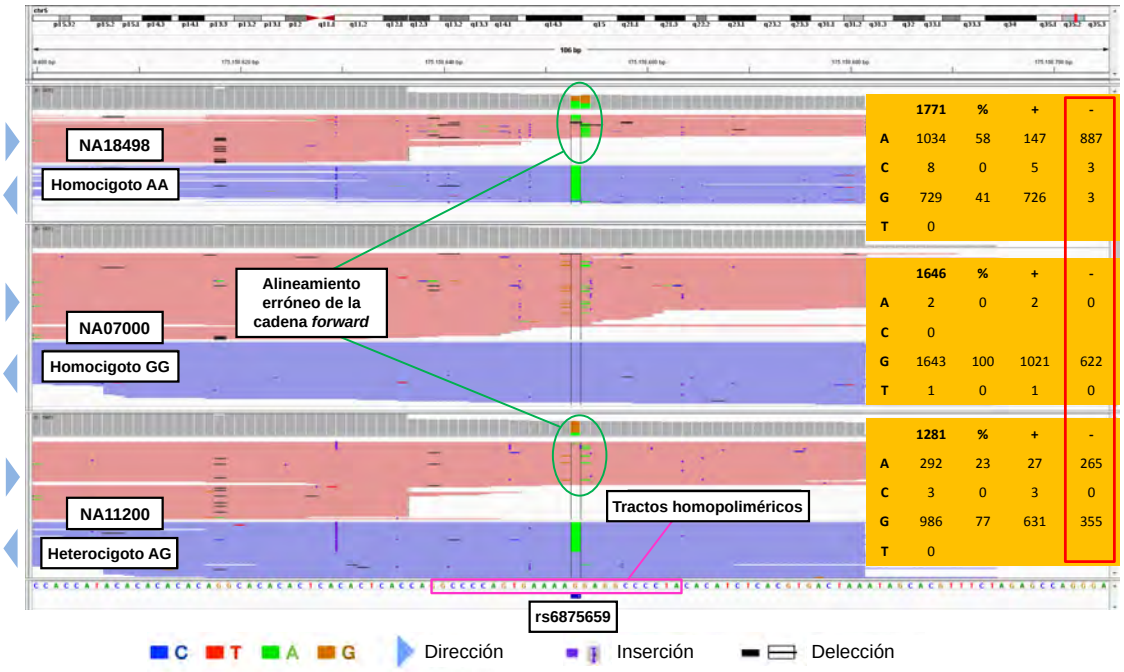


Fig. 61. Visualización en IGV del SNP A/G rs6875659 en diferentes muestras. Las tablas recogen información sobre las lecturas en la posición del SNP. Los genotipos se corrigieron en base a la cadena *reverse*.

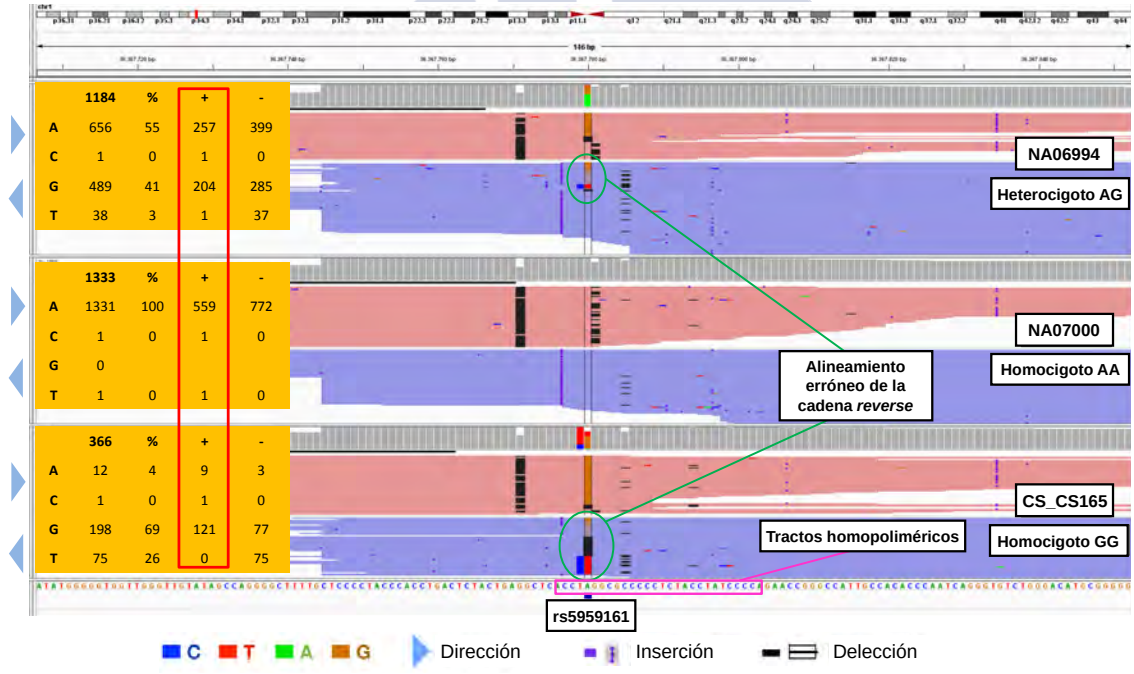


Fig. 62. Visualización en IGV del SNP A/G rs5959161 en diferentes muestras. Las tablas recogen información sobre las lecturas en la posición del SNP. Los genotipos se corrigieron en base a la cadena *forward.*

Aunque no se observaron discordancias para el SNP rs12402499 –Fig. 63–, las muestras de poblaciones de estudio AFR presentaron un alto número de *no-calls* (22%, 44/204). La secuencia contexto de este SNP presenta un Indel específico de población (rs146348214, TTGA/–) en la posición adyacente al SNP, con una frecuencia del alelo deleción del 15% en poblaciones AFR. Tanto el SNP como el Indel son secuenciados correctamente, pero son identificados como una única variante por Genotyper, de manera que las muestras que presentan la deleción son genotipadas como NN. Los genotipos de las muestras de poblaciones AFR con *no-call* para rs12402499 se corrigieron e infirieron manualmente revisando los archivos .vcf.

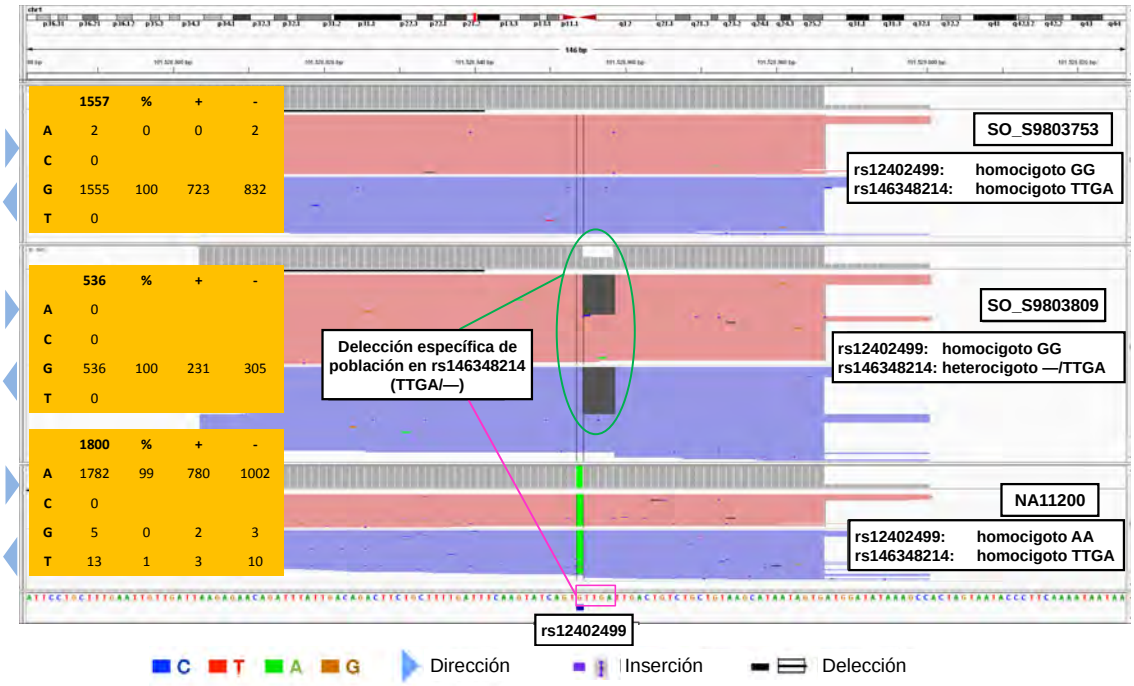


Fig. 63. Visualización en IGV del SNP A/G rs12402449 en diferentes muestras. Las tablas recogen la información sobre las lecturas de la posición del SNP.

4.2.2.3 Sensibilidad del ensayo Global AIM-SNP y análisis de ADN degradado

Las diluciones seriadas del control de ADN Coriell NA07000 presentaron una concordancia total para cantidades de ADN inicial entre 10 ng y 100 pg, usando 21 y 25 ciclos de amplificación. La única excepción fue el SNP rs715605, que obtuvo *no-calls* debidos a bajo *coverage* con cantidades de ADN inicial ≤100 pg. A su vez, el SNP rs187153 presentó *no-calls* con cantidades de ADN inicial ≤50 pg. Por debajo de 100 pg se observan *no-calls*, *drop-ins* y *drop-outs* –Tabla 20–. La reamplificación de las librerías con 5 ciclos adicionales (+5) no conlleva una mayor sensibilidad, sino que produce un incremento de genotipos parciales y *no-calls*. A pesar de ello, la muestra de cantidad inicial de ADN 10 pg (25+5c) produjo un 48% (62/128) de genotipos concordantes.

La muestra Hueso-1 no produjo resultados para ninguno de los marcadores analizados. La muestra Hueso-2 (cantidad inicial de ADN= 726 pg) produjo 4 *no-calls* y un *coverage* promedio de 430x –ver Tabla 20–. La tasa de *no-calls* de Hueso-2 es mayor que la de las diluciones con una cantidad de ADN inicial equivalente.

Tabla 20. *Coverage* promedio, *no-calls*, *drop-ins* y *drop-outs* (por alelo y marcador) de las librerías de las diluciones seriadas, con cantidades iniciales de ADN de 50, 25 y 10 pg y de la muestra Hueso-2. c: ciclos.

	50 pg 25c	50 pg 25+5c	25 pg 25c	25 pg 25+5c	10 pg 25+5c	Hueso-2
Coverage promedio	265	183	267	211	48	430
No-calls	3	3	4	3	9	4
Drop-in de alelo	1	0	2	1	1	0
Drop-out de alelo	1	2	5	10	7	0
Drop-out del *locus*	0	0	1	15	49	0
Total	5	5	12	29	66	4

4.2.2.4 Análisis y detección de mezclas de ADN

Los AIM-SNPs se escogen de manera que se maximizan las diferencias de frecuencias entre las poblaciones, presentando alelos casi fijados en alguna población. Por ello, son marcadores con una capacidad limitada a la hora de detectar mezclas de ADN de individuos que comparten ancestralidad, en comparación con los SNPs de identificación que presentan un bajo contraste de frecuencias alélicas entre las poblaciones ancestrales. No obstante, se espera que un panel de AIM-SNPs presente una mayor heterocigosidad (% de genotipos heterocigotos en el perfil) en mezclas de ADN de componentes con ancestralidades diferentes. Al comparar los genotipos de las muestras individuales SS1 y SS2 –*single source* 1 y 2– con los de la mezcla esperada (basados en la combinación de los genotipos de las muestras individuales), se observa un aumento del 40% en el nivel de heterocigosidad (Fig. 64A). Al comparar los niveles de heterocigosidad observados para las diferentes ratios de mezcla con los de la mezcla esperada se observa que son muy próximos para la mezcla 1:1, mientras el nivel de heterocigosidad desciende a medida que las ratios de mezcla son más asimétricas –Fig. 64B– al no detectarse los alelos del componente minoritario de la mezcla (*drop-out*).

La distribución asimétrica de los genotipos *no-call* y *drop-outs* en las diferentes ratios de mezcla (1:9 vs. 9:1 y 1:3 vs. 3:1) se debe, probablemente, a que el componente SS1 presenta una concentración ligeramente más alta. El análisis detallado de concordancia entre réplicas indica que los *drop-outs* no se deben a que el alelo minoritario no amplifique durante la PCR, sino a que la frecuencia de lecturas del mismo no alcanza el umbral mínimo de 0,10 para el parámetro minimum_allele_frequency, necesario para asignar un genotipo heterocigoto en Genotyper. Por ello, los datos fueron reanalizados ajustando este parámetro a 0,02. Los resultados de la Fig. 64C indican como Genotyper detecta una proporción más alta de alelos minoritarios al cambiar el parámetro. De hecho, el 90% de los *drop-outs* que se producen con

el umbral de 0,1 se solventan aplicando el umbral de 0,02. La mayoría de los *no-call* de los análisis de mezclas de ADN se concentran en un pequeño número de SNPs atípicos: rs4979274, rs310644, rs11048128 y rs7151991 –ver sección 4.2.2.5.3 y Tabla 24–. Además, el SNP rs12402499 presentó *no-calls* en todas las ratios en las que el componente de la mezcla SS2 (AFR) era mayoritario, causado por la deleción específica de población –ver sección 4.2.2.2.4–.

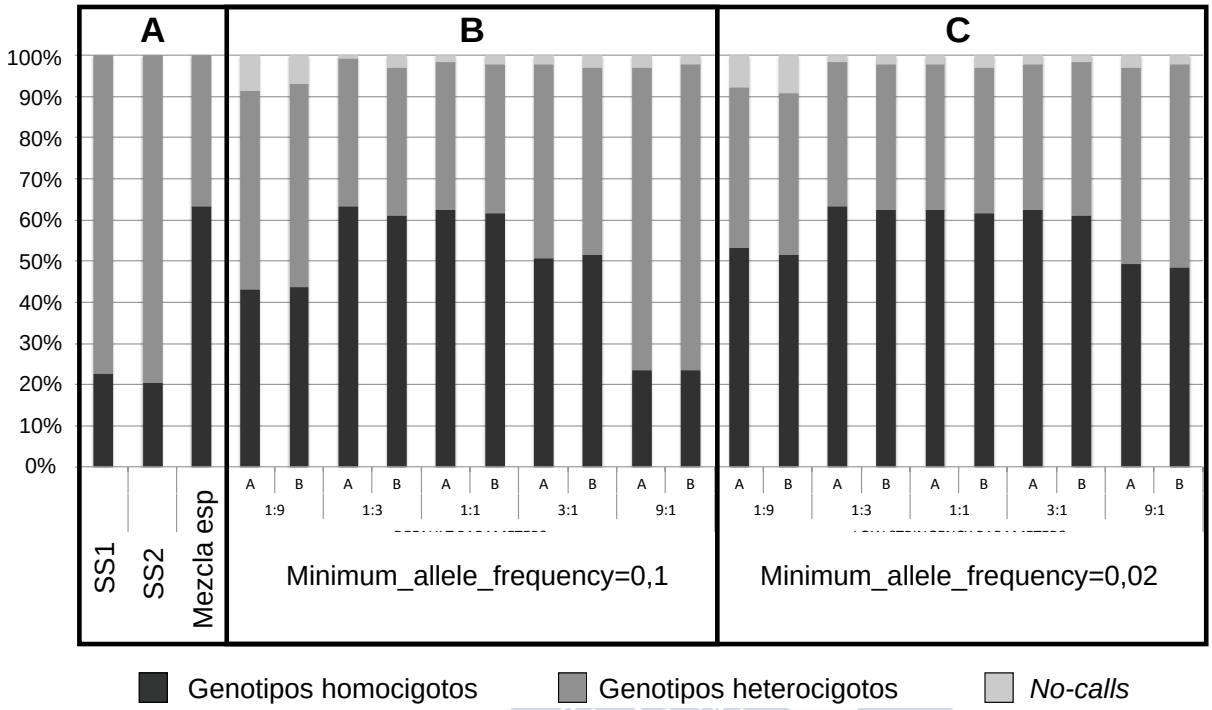
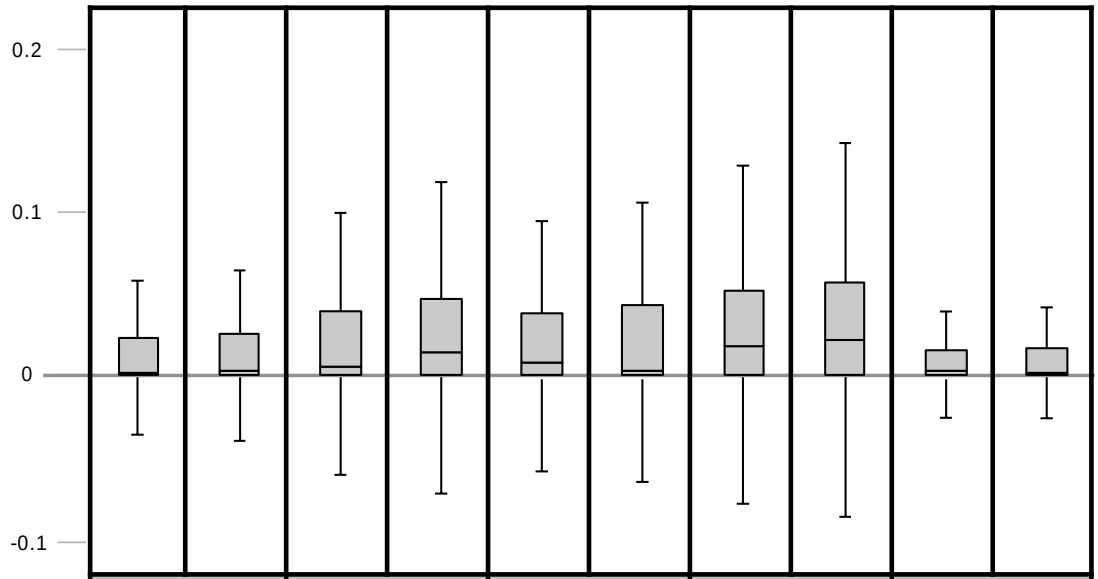


Fig. 64. Porcentaje de *loci* heterocigotos, homocigotos y *no-call* para los 127 SNPs del panel (se excluye rs20180161). De derecha a izquierda: SS1; SS2; Mezcla esp= mezcla esperada (genotipos calculados teóricamente, combinando los de SS1 y SS2); y las diferentes ratios de mezcla (1:9, 1:3, 1:1, 3:1, 9:1) en dos réplicas (A y B) con diferentes parámetros de análisis (minimum_allele_frequency= 0,1 y 0,02).

En la Tabla 21 se muestran las desviaciones entre los valores observados y esperados de frecuencia de lecturas de un alelo presente en la muestra SS1 para las diferentes ratios de mezclas en 123 SNPs (se excluye rs2080161 y cuatro SNPs trialélicos que presentan tres alelos en los genotipos esperados de la mezcla). En general, la correlación entre los valores observados y esperados es alta, mayor del 95% en todos los casos (valores de R^2). No obstante, se observa una tendencia hacia mayores desviaciones en las ratios de mezcla más balanceadas.

Tabla 21. Los diagramas de caja representan, para 123 SNPs del panel, las diferencias entre los valores observados y esperados de frecuencia de lecturas de un alelo presente en la muestra SS1 para las réplicas de las diferentes proporciones de mezclas de ADN. Para cada réplica, se recogen: (i) los valores correspondientes al R^2 de un modelo de regresión lineal entre las frecuencias observadas y esperadas, (ii) la media de las diferencias, y (iii) la desviación típica (sd: *standard deviation*) de las diferencias.



	1:9		1:3		1:1		3:1		9:1	
	A	B	A	B	A	B	A	B	A	B
media	0.012	0.013	0.022	0.023	0.020	0.020	0.025	0.027	0.005	0.006
sd	0.029	0.031	0.043	0.042	0.041	0.043	0.041	0.046	0.026	0.030
R^2	0.998	0.998	0.994	0.994	0.989	0.988	0.982	0.978	0.992	0.989

La representación gráfica de las ARF –*Allele Read Frequency*– de cada SNP es un método útil para la identificación de mezclas de ADN, ya que la mayoría de las ARF aparecerán desplazadas frente a los patrones típicos de muestras individuales. En la Fig. 65 aparecen representadas las ARF de las mezclas de ADN de diferentes ratios conjuntamente con las individuales de las muestras SS1 y SS2. Se observa que el desplazamiento es más acusado en la ratio 1:1 y que las réplicas de cada ratio presentan patrones muy similares. Además, al representarse específicamente la ARF de un alelo del componente SS1 se observa como la frecuencia de los mismos decrece a través de las diferentes ratios de mezcla, alcanzando mínimos de 0,05% en la mezcla 1:9 (la mezcla con menor proporción de SS1).

Es importante explorar hasta qué punto se diferencian las mezclas de ADN con componentes de ancestralidades diferentes de los individuos *admixed*, dado que en ambos casos se produce un aumento de los niveles de heterocigosidad. En la Tabla 22 se recogen los niveles de heterocigosidad media observados en poblaciones *unadmixed* (poblaciones de referencia AFR y EUR) y *admixed* (poblaciones que presentan coancestralidad AFR y/o EUR). En los datos se aprecia como en las poblaciones *admixed* se eleva la heterocigosidad entre un 20-80% respecto a las *unadmixed*. Sin embargo, gracias a la alta correlación entre la

proporción de los alelos en el ADN inicial (equivalente a las ARF esperados) y los ARF observados es relativamente sencillo distinguir ambas situaciones: las ARF de los individuos *admixed* mostrarán patrones similares a los de las muestras individuales (*single-source* SS1 y SS2 en la Fig. 64). Además, se espera que esta información esté disponible de antemano, tras los análisis iniciales con STRs de rutina.

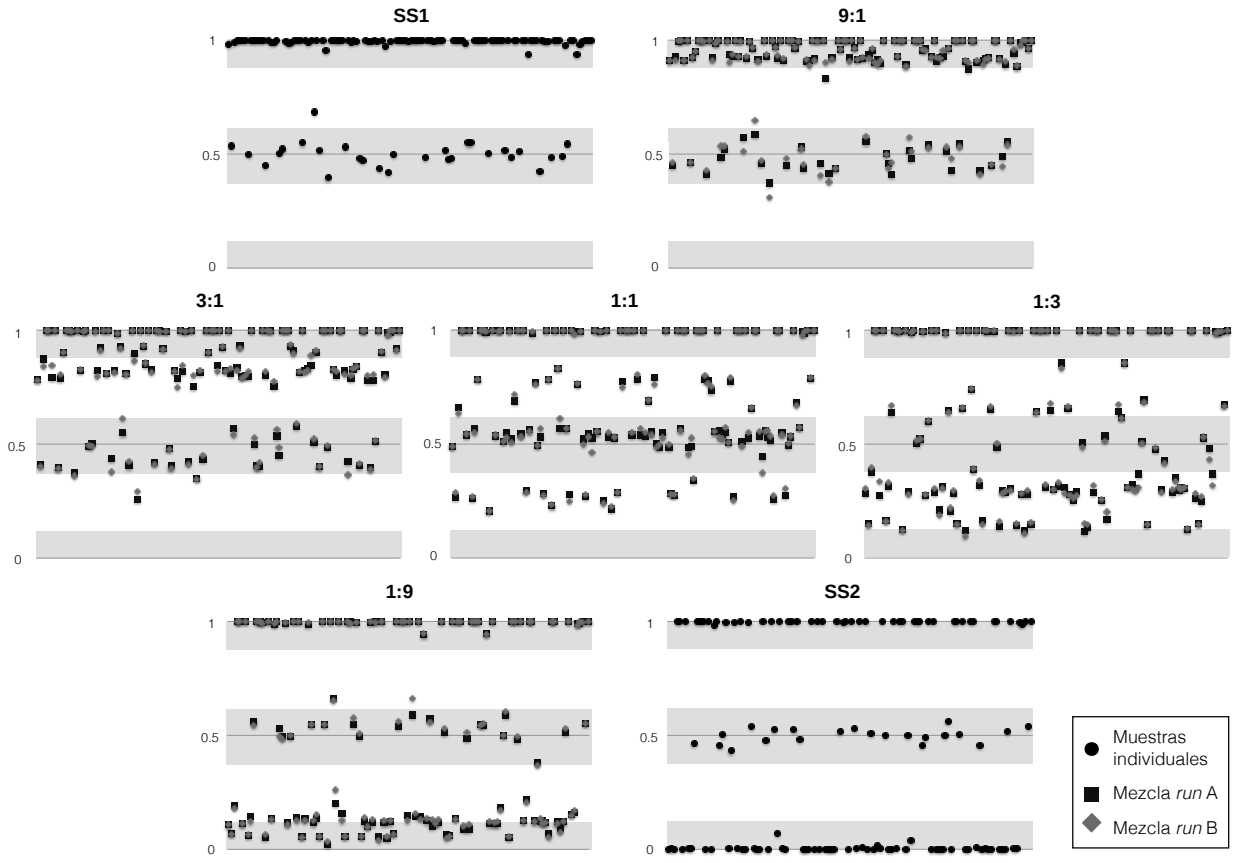


Fig. 65. Representación del porcentaje de lecturas de un alelo presente en la muestra EUR (SS1) sobre el *coverage* total. De arriba hacia abajo y de derecha a izquierda: SS1 (ADN *single-source* EUR), mezclas de ADN de ratios 9:1, 3:1, 1:1, 1:3, 1:9 (en ambos *runs*) y SS2 (ADN *single-source* AFR).

Tabla 22. Heterocigosidad media observada en poblaciones del *unadmixed* del Proyecto 1000 Genomas englobadas en las superpoblaciones AFR y EUR y en todas las poblaciones *admixed* del mismo. PEL, MXL, CLM y PUR presentan un componente europeo; ACB y ASW un componente africano.

Tipo de población	Población	Heterocigosidad media
Unadmixed	AFR	0,165
	EUR	0,245
Admixed	ASW	0,270
	ACB	0,218
	PEL	0,329
	MXL	0,372
	CLM	0,252
	PUR	0,343

Los 6 SNPs trialélicos incluidos en el panel EUROFORGEN Global AIM-SNP proporcionan un método adicional para identificar mezclas. No obstante, Genotyper no permite la detección automática de un tercer alelo. En la Tabla 23 se muestra como la ARF esperada del alelo minoritario en SNPs trialélicos es igual que en bialélicos, por lo que las ARF observadas en SNPs bialélicos sirven de guía para establecer el rango de ARFs del tercer alelo en SNPs trialélicos. No obstante, se debe tener un especial cuidado para no confundir el alelo minoritario con una proporción de nucleótidos incorporados erróneamente, en especial cuanto más extrema sea la ratio de la mezcla. En 4/6 SNPs trialélicos, los genotipos de la mezcla esperada presentaban tres alelos (rs2184030, rs4540055, rs433342 y rs17287498). El tercer alelo pudo ser detectado para todas las ratios de mezcla, examinando detalladamente los valores de ARF. En el caso de la mezcla de ADN de ratio 9:1, se pudo detectar el alelo minoritario con una ARF del 3%, un valor muy próximo al 5% esperado.

Tabla 23. Frecuencias de lecturas de los alelos esperadas para las diferentes ratios de mezclas de ADN evaluados. Se muestran todas las combinaciones posibles de genotipos de dos muestras individuales (SS1 y SS2). Nótese que los marcadores trialélicos se comportan como bialélicos siempre que los genotipos de las muestras individuales comprendan tan solo 2 de los 3 alelos posibles.

		Genotipos		ARF esperada para las diferentes ratios de mezclas de ADN															
				1:9			1:3			1:1			3:1			9:1			
		SS 1	SS 2	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Tipo de SNP	Bialélicos	AA	AA	1	0	-	1	0	-	1	0	-	1	0	-	1	0	-
		AB	AA	0,95	0,05	-	0,875	0,125	-	0,75	0,25	-	0,625	0,375	-	0,55	0,45	-
		BB	AA	0,9	0,1	-	0,75	0,25	-	0,5	0,5	-	0,25	0,75	-	0,1	0,9	-
		AA	AB	0,55	0,45	-	0,625	0,375	-	0,75	0,25	-	0,875	0,125	-	0,95	0,05	-
		AB	AB	0,5	0,5	-	0,5	0,5	-	0,5	0,5	-	0,5	0,5	-	0,5	0,5	-
		BB	AB	0,45	0,55	-	0,375	0,625	-	0,25	0,75	-	0,125	0,875	-	0,05	0,95	-
		AA	BB	0,1	0,9	-	0,25	0,75	-	0,5	0,5	-	0,75	0,25	-	0,9	0,1	-
		AB	BB	0,05	0,95	-	0,125	0,875	-	0,25	0,75	-	0,375	0,625		0,45	0,55	-
		BB	BB	0	1	-	0	1	-	0	1	-	0	1	-	0	1	-
	Trialélicos	AA	BC	0,1	0,45	0,45	0,25	0,375	0,375	0,5	0,25	0,25	0,75	0,125	0,125	0,9	0,05	0,05
		AB	AC	0,5	0,05	0,45	0,5	0,125	0,375	0,5	0,25	0,25	0,5	0,375	0,125	0,5	0,45	0,05
		AB	BC	0,05	0,5	0,45	0,125	0,5	0,375	0,25	0,5	0,25	0,375	0,5	0,125	0,45	0,5	0,05
		AB	CC	0,05	0,05	0,9	0,125	0,125	0,75	0,25	0,25	0,5	0,375	0,375	0,25	0,45	0,45	0,1
		AC	AB	0,5	0,45	0,05	0,5	0,375	0,125	0,5	0,25	0,25	0,5	0,125	0,375	0,5	0,05	0,45
		AC	BB	0,05	0,9	0,05	0,125	0,75	0,125	0,25	0,5	0,25	0,375	0,25	0,375	0,45	0,1	0,45
		AC	BC	0,05	0,45	0,5	0,125	0,375	0,5	0,25	0,25	0,5	0,375	0,125	0,5	0,45	0,05	0,5
		BB	AC	0,45	0,1	0,45	0,375	0,25	0,375	0,25	0,5	0,25	0,125	0,75	0,125	0,05	0,9	0,05
		BC	AA	0,9	0,05	0,05	0,75	0,125	0,125	0,5	0,25	0,25	0,25	0,375	0,375	0,1	0,45	0,45
		BC	AB	0,45	0,5	0,05	0,375	0,5	0,125	0,25	0,5	0,25	0,125	0,5	0,375	0,05	0,5	0,45
		BC	AC	0,45	0,05	0,5	0,45	0,05	0,5	0,25	0,25	0,5	0,125	0,375	0,5	0,05	0,45	0,5
		CC	AB	0,45	0,45	0,1	0,375	0,375	0,25	0,25	0,25	0,5	0,125	0,125	0,75	0,05	0,05	0,9

4.2.2.5 Evaluación del rendimiento de los SNPs incluidos en el ensayo

La identificación de los SNPs atípicos se basa en: (i) los resultados de concordancia, sensibilidad y análisis de mezclas de controles de ADN y la revisión de los genotipos de las poblaciones de estudio, y (ii) la evaluación de cuatro parámetros de calidad de las secuencias en base a los valores promedio de las muestras de las poblaciones de estudio: *coverage*, balance de la frecuencia de lectura de los alelos –ARF–, tasas de incorporación errónea de nucleótidos –*misincorporation*– y sesgo de cadena de cada alelo –*strand bias per allele*–. En función de los resultados, los SNPs fueron clasificados en los siguientes grupos –ver Tabla 24–: (i) SNPs con genotipos discordantes; (ii) SNPs con *no-calls*; y (iii) SNPs con un buen rendimiento y alta fiabilidad de los genotipos.

Tabla 24. Clasificación de los SNPs del ensayo en función de los análisis de calidad aplicados.

SNPs discordantes	SNPs con *no-calls*			SNPs con buen rendimiento y alta fiabilidad de los genotipos
	En controles de ADN	En poblaciones de estudio	En ambos sets	
rs595961 rs6875659 rs2080161 rs9934011 rs9908046	rs12498138 rs4979274 rs310644	rs1366220 rs16946159 rs203150 rs715605	rs11048128 rs7151991 rs499827	112 SNPs restantes

4.2.2.5.1 Parámetros clave de calidad de las secuencias

El principal factor limitante en los análisis de MPS es el *coverage*. Entre las muestras de las poblaciones de estudio se obtuvieron valores promedio de *coverage* por SNP entre 106-1647x. Esta variación en los valores de *coverage* por SNP se debe a diferencias en la eficiencia de amplificación en la PCR *multiplex* inicial que incluye los 128 marcadores, y ya ha sido descrita en paneles de SNPs para Ion PGM™ de tamaños similares (Seo *et al.* 2013, Børsting *et al.* 2014). No obstante, el 95% de los SNPs del panel (122/128) mostró un *coverage* promedio >300x. En los 6 SNPs restantes (rs2080161, rs12498138, rs4979274, rs310644, rs1366220 y rs203150) se observó una tasa elevada de *no-calls* (>1% en 551 muestras) o genotipos discordantes.

Otro factor clave es el balance de los alelos, una variable crítica a la hora de asignar genotipos heterocigotos fiables e identificar mezclas de ADN. En este sentido, el parámetro ARF equivale a las ratios de la señal en marcadores heterocigotos detectados mediante CE. Un total de 123 SNPs mostraron rangos de valores de ARF comprendidos entre los umbrales de >90% para homocigotos y entre 40-60% para heterocigotos. Cuatro de los cinco SNPs restantes mostraron valores ligeramente desviados de los umbrales (entre 61-65%). Un único SNP (rs310644) presentó desviaciones muy marcadas, y fue identificado a su vez como un marcador con bajo *coverage* y alto número de *no-calls* en análisis de concordancia y mezclas de ADN.

Además del balance de los alelos, la tasa de incorporación errónea de nucleótidos o *misincorporation* (porcentaje de lecturas no alélicas entre todas las lecturas del SNP) constituye un factor importante a la hora de genotipar los SNPs e identificar los alelos minoritarios de las mezclas de ADN. La tasa de *misincorporation* es <1% para todos los SNPs, salvo rs595961 con un 2,9% y rs2789823 con un 1,8%. La secuencia contexto de estos SNPs indica que la aparente incorporación errónea de nucleótidos es causada por el alineamiento erróneo de las secuencias debido a la presencia de tractos homopoliméricos cercanos al SNP.

Por último, el parámetro *strand bias per allele*, que mide la ratio de secuencias de cada alelo en cada cadena, puede afectar a la calidad de las lecturas y a los genotipos obtenidos. De entre los 3 SNPs que presentaron lecturas de una de las cadenas afectadas por las características de la secuencia contexto, tan solo rs6875659 tiene un valor promedio no comprendido en el rango de 25-75% –ver sección 4.2.2.2.4 y Fig. 61–.

4.2.2.5.2 SNPs con genotipos discordantes y exclusión de rs2080161

Cuatro SNPs mostraron genotipos discordantes en los análisis de controles de ADN: rs9934011, rs9908046, rs6875659 y rs2080161. El SNP rs595961 presentó discordancias en varias muestras de poblaciones de estudio. Todas las discordancias se producen como consecuencia de alineamientos erróneos causados por tractos homopoliméricos en las secuencias contexto. Mientras que estos alineamientos erróneos ocurrieron una única vez y en diferentes controles de ADN para los SNPs rs9934011 y rs9908046 (probablemente debido a efectos estocásticos); los SNPs rs2080161, rs595961 y rs6875659 presentaron discordancias en análisis diferentes del mismo control de ADN o en varias muestras de las poblaciones de estudio, indicando un error sistemático. Los genotipos de estos SNPs presentan un sesgo en una de las cadenas debido a la presencia de tractos homopoliméricos –ver sección 4.2.2.2.4–. El SNP rs595961 tiene dos tractos poli-C de 5 y 4 Cs consecutivas entre los 25 nucleótidos adyacentes al SNP, causando que una alta proporción de lecturas de la cadena *reverse* sean poco fiables (Fig. 62). Asimismo, el SNP rs6875659 presenta las lecturas de la cadena *forward* afectadas por tractos poli-C y poli-A (Fig. 61). Los genotipos de estos dos SNPs pueden ser corregidos manualmente infiriendo el genotipo correcto a partir de la cadena no afectada. No obstante, el SNP rs2080161 se encuentra rodeado por tractos poli-T de más de 5 nucleótidos en ambas direcciones (Fig. 60), por lo que no es posible realizar la corrección manual y el SNP debe ser excluido del panel.

4.2.2.5.3 SNPs con no-calls

Diez SNPs presentaron más *no-calls* que la media en controles de ADN y muestras de poblaciones de estudio (>1 en 551 muestras). La mayoría de *no-calls* se deben a bajo *coverage* (<5-30x) o baja calidad de las secuencias en ciertas muestras. Estos datos concuerdan con el umbral de *coverage* mínimo establecido en otros estudios de MPS para la

obtención de genotipos fiables de SNPs (Bentley *et al.* 2008, Nielsen *et al.* 2011, Quail *et al.* 2012, Sims *et al.* 2014).

En general, una revisión manual extensiva de las secuencias y los archivos de salida de Genotyper y el establecimiento de umbrales para parámetros clave de calidad de las secuencias aseguran que los genotipos utilizados para los análisis poblacionales sean fiables. No obstante, los SNPs que requirieron correcciones manuales podrían ser reemplazados, en revisiones futuras del panel, por *loci* con patrones de informatividad similares y secuencias contexto menos problemáticas. Además, se debe priorizar el desarrollo y la mejora del *software* de análisis para MPS, de manera que se resuelvan los problemas de alineamiento que se producen en los tractos homopoliméricos.

4.2.2.6 Análisis poblacionales

En la Fig. 66, Fig. 67 y Fig. 70 se muestran los resultados de diferentes análisis de ancestralidad para las 14 poblaciones de estudio y las poblaciones test del Proyecto 1000 Genomas, utilizando 5 grupos poblacionales de referencia.

Las distribuciones de la media de las diferencias genotípicas interpoblacionales e intrapoblacionales y de los valores de F_{ST} en las poblaciones test y de estudio –Fig. 66– se corresponden adecuadamente con los patrones que se esperan en función de su situación geográfica y sus niveles de *admixture*. Ejemplos de ello son: (i) la población SO presenta un menor grado de diferenciación con las población EUR en comparación con otras poblaciones AFR y al igual que las poblaciones ASW y ACB, que presentan *admixture* con componente EUR; (ii) la población GL aparece menos diferenciada de las poblaciones *admixed* de AMR, revelando cierto grado de coancestralidad entre EUR y AMR; (iii) las poblaciones FJ y ET aparecen poco diferenciadas con OCE; y (iv) las poblaciones *admixed* de AMR muestran los mayores niveles de variación intrapoblacional, con patrones similares a IN, AF y GL.

El gráfico de STRUCTURE (Fig. 67) indica patrones de agrupación que concuerdan con las distribuciones geográficas y los niveles de *admixture* de cada población. Los resultados que se presentan son para el K óptimo de 5, infiriendo los grupos a partir de 5 poblaciones de referencia. Este análisis se realiza excluyendo las poblaciones SAS, ya que la diferenciación de esta población de la EUR no fue un factor a tener en cuenta en la selección original de los SNPs del panel. Esta selección de marcadores se refleja en los resultados de los valores de PSD acumulados para comparaciones de 5 poblaciones frente a los de 6 poblaciones, en las que SAS alcanza valores de ~3 mientras que el resto de poblaciones alcanza valores de entre 11-15 (ver Tabla 25).

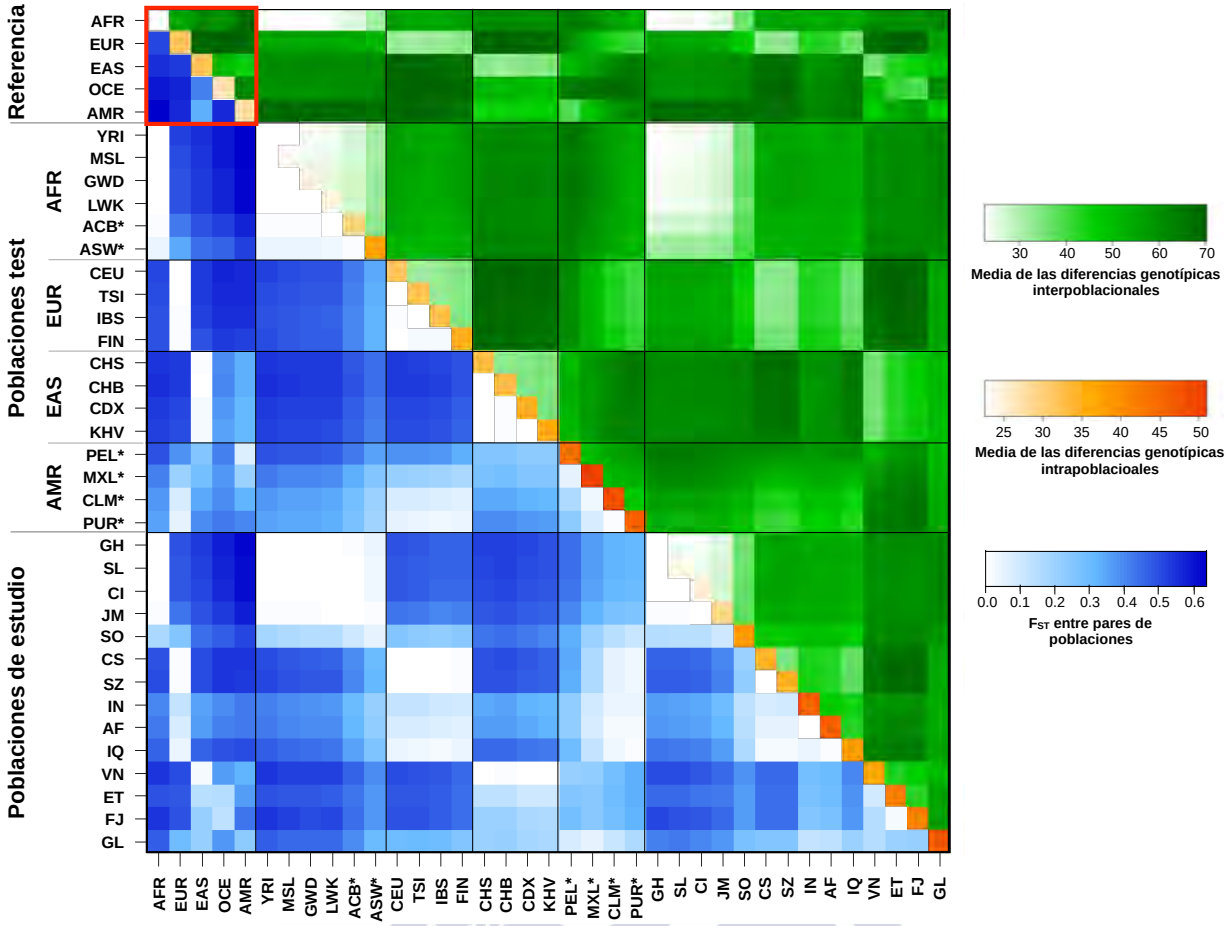


Fig. 66. Valores d F_{ST} y media de las diferencias genotípicas intrapoblacionales e interpoblacionales. Las poblaciones de referencia se indicadan en un recuadro rojo. Las poblaciones test *admixed* se indican con un asterisco.

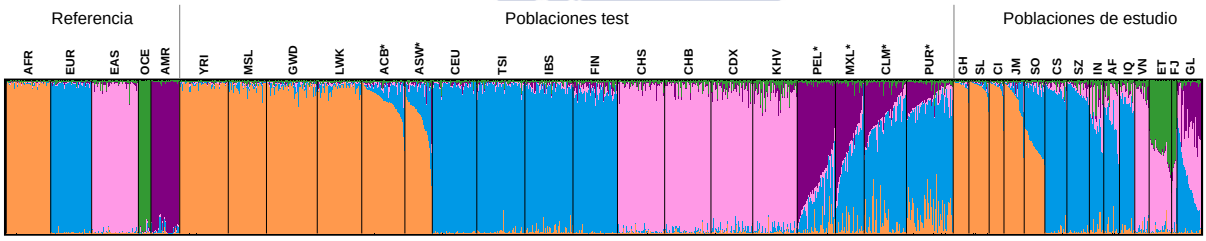


Fig. 67. Resultados de STRUCTURE de las poblaciones test y estudio en comparación con 5 poblaciones de referencia. Valor óptimo de K=5. Los individuos de las poblaciones test *admixed* (marcadas con un asterisco) y de estudio se ordenan en orden decreciente del componente mayoritario.

Tabla 25. Valores de PSD acumulados de los127 SNPs del panel para comparaciones de 5 y 6 grupos.

	AFR	EUR	EAS	OCE	AMR	SAS
5 grupos	16.448	15.132	11.080	14.105	14.547	–
6 grupos	15.225	11.238	10.727	14.482	14.060	3.060

Para alcanzar niveles adecuados de diferenciación de SAS se deben incorporar marcadores informativos para esta población, como los incluidos en Eurasiaplex (Phillips *et al.* 2013b), que permitan balancear la PSD acumulada y aseguren un análisis no sesgado. No obstante, se incluye un análisis STRUCTURE con las 6 poblaciones de referencia –Fig. 68– y otro con la clasificación de poblaciones test SAS y poblaciones de estudio con posible componente SAS frente al set de 6 poblaciones de referencia –Fig. 69–. A pesar de la reducida divergencia de SAS, los resultados de STRUCTURE indican un K óptimo de 6 para los análisis que incluyen las 6 poblaciones de referencia –Fig. 68–.

En los análisis de K=5 (Fig. 67) destacan los patrones de agrupamiento de SO y GL. En primer lugar, la población SO se encuentra posicionada en el extremo oriental del continente africano, produciéndose *admixture* con poblaciones de ME y SAS. Las muestras de la población SO manifiestan patrones de coancestralidad de proporciones casi equivalentes entre AFR y EUR, considerándose Eurasia como un grupo poblacional que incluye EUR, ME y SAS. No obstante, cuando se añade una población SAS como referencia (Fig. 69) la población SO sigue presentando coancestralidad EUR en la mayoría de las muestras. Estos resultados para K=6 subrayan la falta de divergencia entre SAS y EUR –ver Tabla 26–, así como la necesidad de reconfigurar o suplementar el panel para adecuarlo a la diferenciación de 6 poblaciones. En segundo lugar, las muestras de la población GL muestran patrones complejos de agrupación. Estos patrones reflejan su origen particular a partir de migraciones de Siberia y el noreste de Asia (Colonna *et al.* 2011), diferente al resto de poblaciones AMR *admixed* analizadas que presentan *admixture* con un componente EUR reciente.

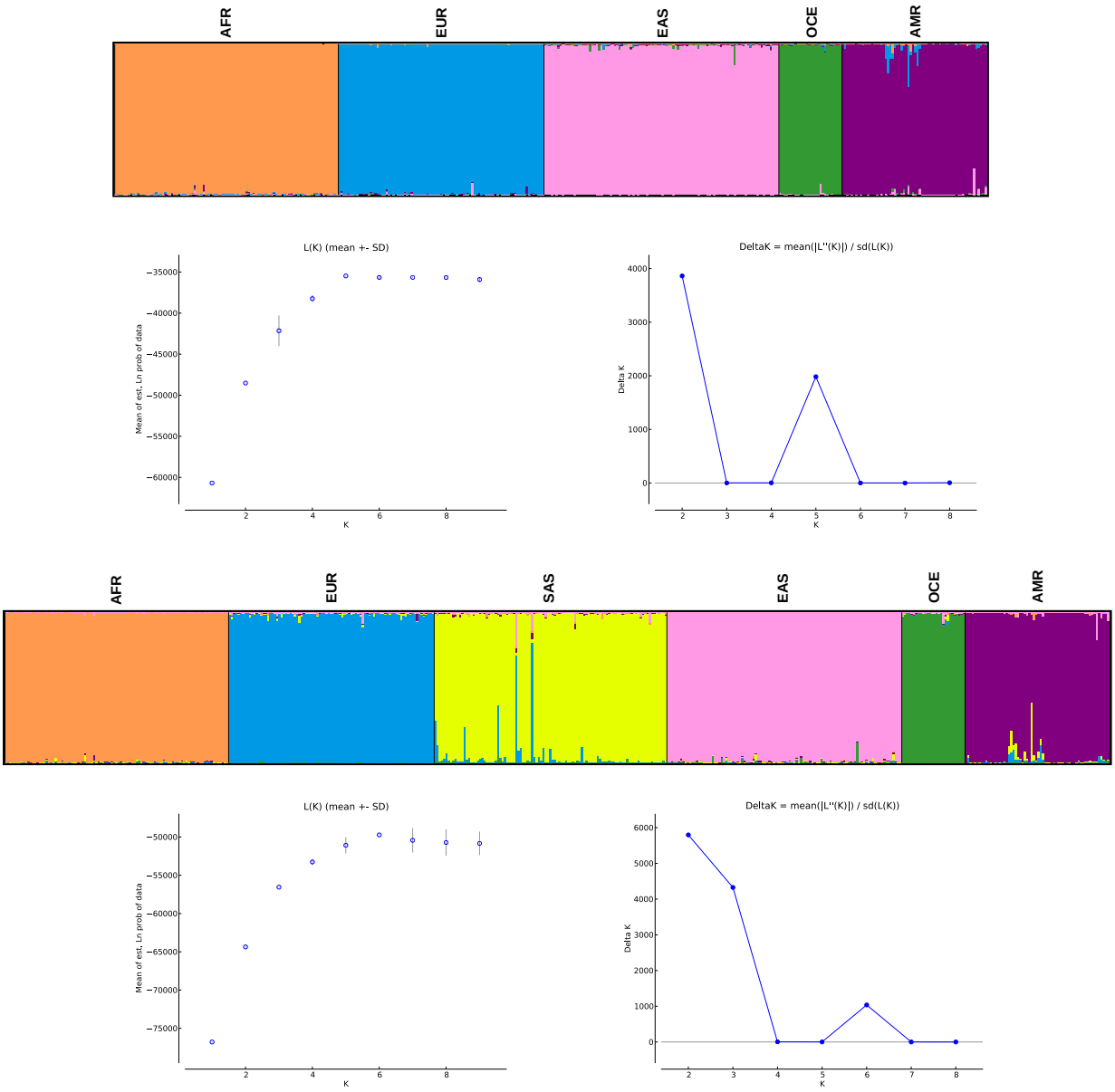


Fig. 68. Comparación de los análisis de ancestralidad para 5 (AFR, EUR, EAS, OCE y AMR) y 6 (+ SAS) grupos de referencia. Los análisis de STRUCTURE se realizaron con 5 réplicas de K=1 a K=9 bajo el modelo de *admixture* y frecuencias alélicas correlacionadas (100000 *burning steps* y 100000 *MCMC steps*). Para estimar el K óptimo se generaron, mediante Structure Harvester, gráficos del Ln de la media de las estimaciones de probabilidad de los datos y los valores de Delta K. Se obtuvieron valores de K óptimos de 5 y 6 para 5 y 6 poblaciones de referencia, respectivamente.

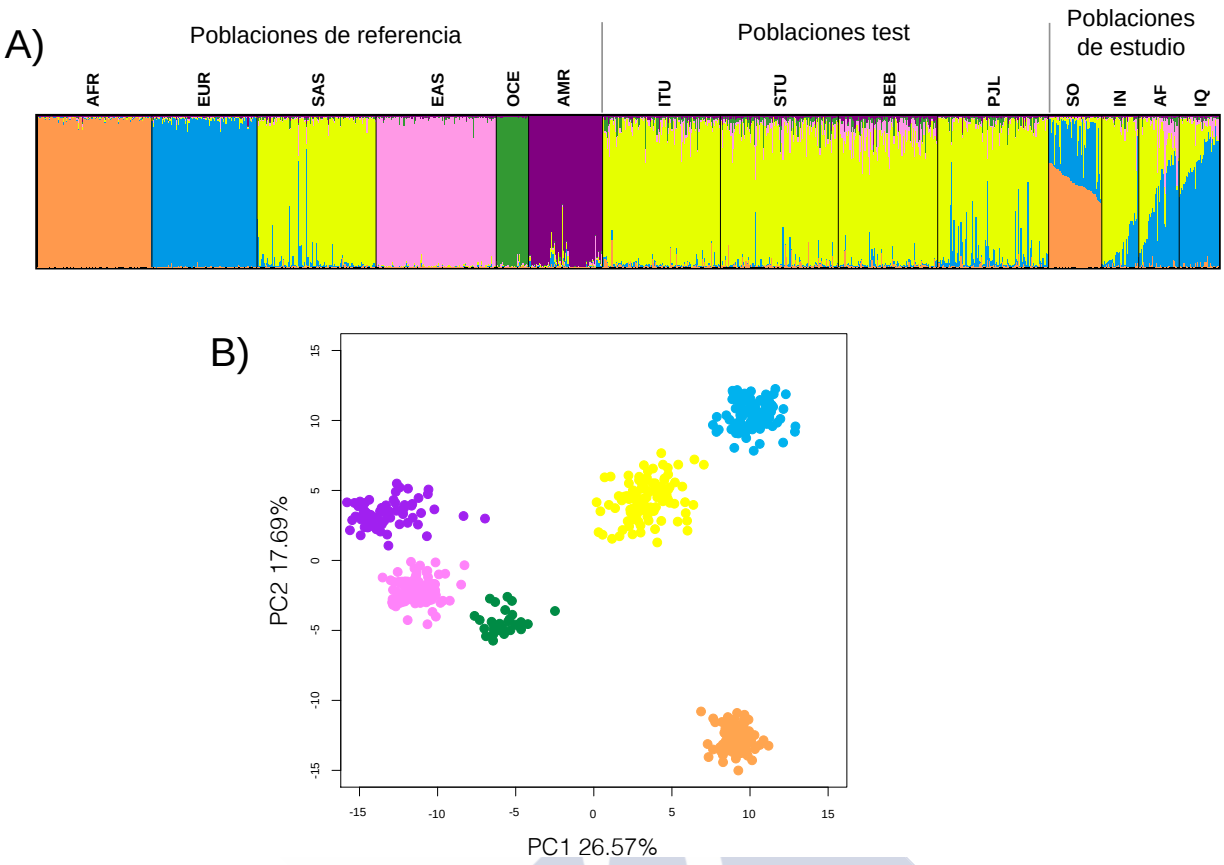


Fig. 69. Análisis de ancestralidad con 6 grupos de referencia para poblaciones test SAS y poblaciones de estudio con posible componente SAS. A) Resultados del análisis STRUCTURE. Se realizaron 5 réplicas de K=1 a K=9 bajo el modelo de *admixture* POPFLAG y frecuencias alélicas correlacionadas (100000 *burning steps* y 100000 *MCMC steps*). El K óptimo se estimó en 6. B) PCA de las 6 poblaciones de referencia (PC1 vs. PC2).

Tabla 26. Valores acumulados de divergencia entre pares de poblaciones de los 127 SNPs del panel.

	AFR	EUR	EAS	OCE	AMR	SAS
AFR	–					
EUR	18.673	–				
EAS	21.781	22.325	–			
OCE	23.386	26.084	15.136	–		
AMR	28.520	23.937	10.567	22.167	–	
SAS	13.530	4.586	11.467	16.198	15.610	–

La Fig. 70 muestra cuatro gráficas de PCA que indican las posiciones de las poblaciones de referencia (1 y 2) y, frente a éstas, de las poblaciones test del Proyecto 1000 Genomas *unadmixed* (3) y *admixed* (4). Los grupos de puntos de cada población se distribuyen en correspondencia con los resultados de la media de las diferencias genotípicas (Fig. 66) y los análisis STRUCTURE (Fig. 67).

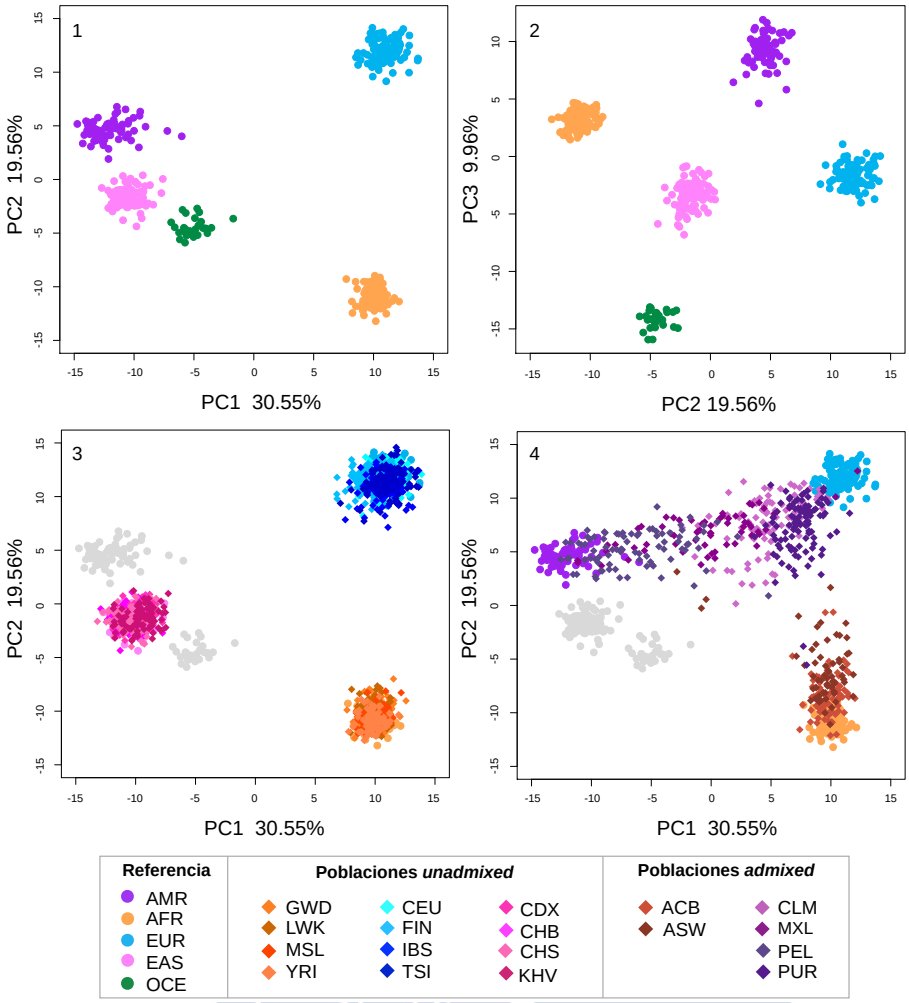


Fig. 70. Gráficos de PCA para poblaciones test comparadas con 5 poblaciones de referencia. Se muestran los gráficos de PC1 vs. PC2 (1) y de PC2 vs. PC3 (2) de las poblaciones de referencia, así como los gráficos PC1 vs. PC2 de las poblaciones test *unadmixed* (3) y *admixed* (4) frente a las referencias (en gris).

En la Fig. 71 se muestran análisis detallados de las poblaciones de estudio en base a 5 grupos de referencia. Para cada población se incluyen, individualmente, los análisis STRUCTURE, PCA y las gráficas que representan las LR ordenadas obtenidas a través de la opción de *cross-validation* en Snipper. Aunque las LR no son generalmente informativas para individuos *admixed*, ordenar las LR obtenidas en una población en una gráfica $\log_{10}$LR puede resultar ilustrativo. Para los individuos de las poblaciones de estudio, las LRs de las clasificaciones se encuentran por encima de la línea de LR=1 y tan solo 2 individuos de GL se encuentran por debajo de esta línea, al ser asignados EUR en vez de AMR. En estos casos y todos los que tienen LRs bajas, los individuos se identifican claramente en el PCA y el gráfico de STRUCTURE, ya que presentan proporciones de coancestralidad superiores a la media de la población. En la Fig. 72, se muestran análisis detallados de las poblaciones de estudio con componente SAS en base a 6 grupos de referencia. Las poblaciones IN, AF y IQ muestran inferencias de ancestralidad SAS, con tan solo 5 muestras con probabilidades más altas de EUR, aunque con valores bajos.

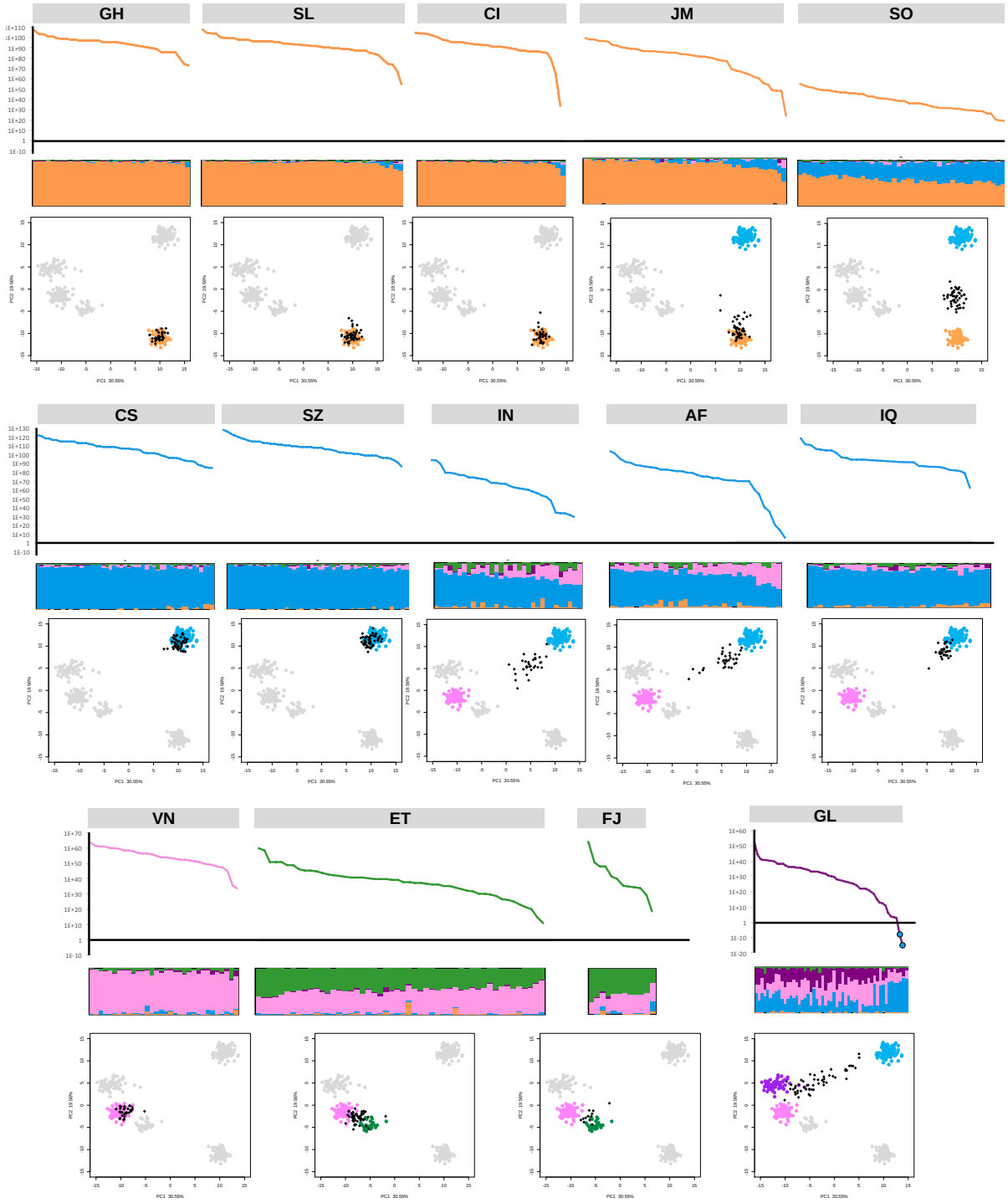


Fig. 71. Análisis de ancestralidad detallado para las 14 poblaciones de estudio en comparación con 5 grupos de referencia. Para cada población, se muestran las gráficas de $\log_{10}$ de las LRs (eje Y) obtenidas mediante la opción de *cross-validation* de Snipper para cada individuo (eje X), en orden decreciente. Los individuos clasificados erróneamente se representan mediante círculos azules (indicando la asignación como EUR). En los gráficos de STRUCTURE para K=5 los individuos están ordenados de manera que se corresponden con la gráfica de LR. Los gráficos de PCA muestran PC1 vs. PC2 para las poblaciones de referencia (en gris o en su color correspondiente). Las coordenadas de los individuos de las poblaciones de estudio (en negro) se calculan en función de los grupos de referencia.

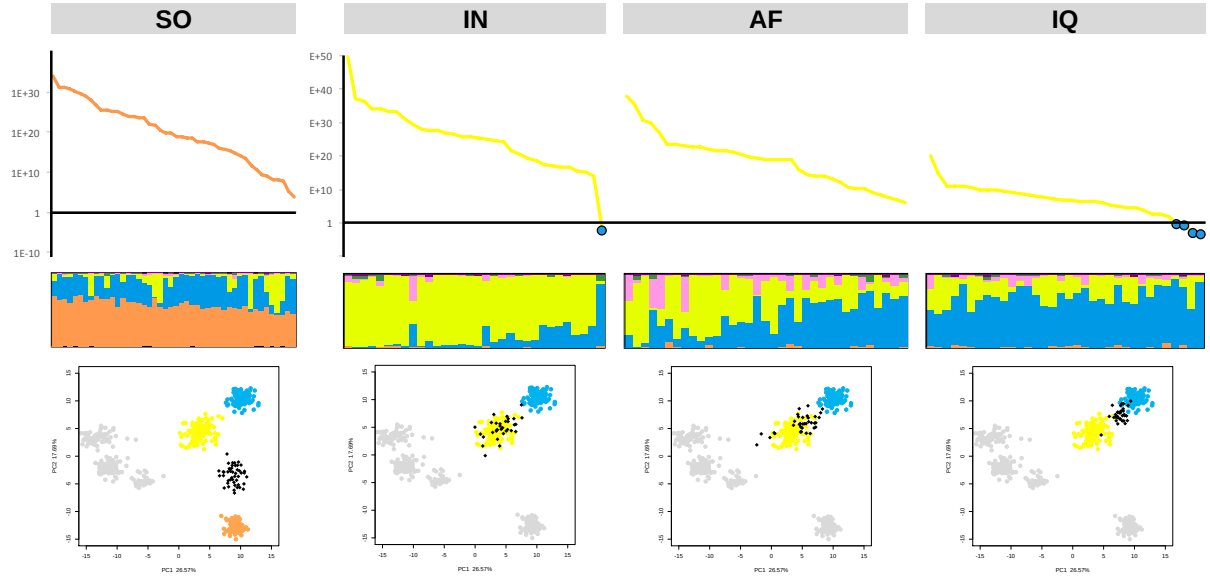


Fig. 72. Análisis de ancestralidad detallado para las poblaciones de estudio con posible componente SAS en comparación con 6 grupos de referencia. Para cada población se muestran las gráficas de $\log_{10}$ de las LR (eje Y) obtenidas mediante la opción de *cross-validation* de Snipper para cada individuo (eje X), en orden decreciente. Los individuos clasificados erróneamente se representan mediante círculos azules (indicando la asignación como EUR). En los gráficos de STRUCTURE para K=6 los individuos están ordenados de manera que se corresponden con la gráfica de LR. Los gráficos de PCA muestran PC1 vs. PC2 para las poblaciones de referencia (en gris o en su color correspondiente). Las coordenadas de los individuos de las poblaciones de estudio (en negro) se calculan en función de los grupos de referencia.

4.2.3 **Discusión**

En este trabajo, se han incorporado con éxito un total de 125 de los 128 marcadores originales del panel EUROFORGEN Global AIM-SNP y 3 marcadores sustitutos en una PCR *multiplex* para Ion PGM™. Se espera que se puedan adaptar adecuadamente paneles de SNPs más extensos, diseñados para análisis de ancestralidad más detallados (como los que probablemente sean necesarios para diferenciar las poblaciones ME y SAS de EUR) o para predecir características fenotípicas complejas. Además, la mayoría de los SNPs de predicción de características fenotípicas se sitúan en regiones codificantes, por lo que es menos probable que se encuentren afectados por una baja complejidad de la secuencia contexto. El hecho de que el SNP rs2080161 tuviera que ser finalmente excluido del panel indica que se debe mejorar el análisis detallado de la secuencia contexto de cada SNP durante el proceso de diseño de los paneles personalizados de Ion PGM™. Aún así, la pérdida de este AIM-SNP y la sustitución de otros 3 produjo un efecto mínimo sobre el balance original del panel de las PSD acumuladas para 5 poblaciones de referencia. Este balance entre las poblaciones permitió un análisis no sesgado en base a 5 grupos de referencia de varias poblaciones de estudio con patrones complejos de *admixture*.

En los análisis de ancestralidad en base a 6 grupos de referencia, existe una importante reducción de la PSD acumulada de la población SAS, indicando que las inferencias de proporciones de coancestralidad en poblaciones de estudio *admixed* (como SO) pueden

presentar un sesgo de infraestima de la proporción del componente SAS frente al resto de grupos. A pesar de esto, las diferentes aproximaciones utilizadas para el análisis poblacional lograron diferenciar, en la mayoría de los casos, entre las poblaciones SAS y EUR.

En una futura revisión del panel se podrían añadir AIM-SNPs que permitan diferenciar las subpoblaciones de Eurasia, de manera que se cubran las necesidades de aquellos laboratorios forenses que, debido a la demografía de la región en la que operan, necesiten diferenciar las poblaciones EUR, ME y SAS. Además, los SNPs que requieren una corrección manual de los genotipos deben ser reemplazados por otros marcadores, manteniendo el ajuste de las PSD acumuladas del panel. En este sentido, algunos SNPs sustitutos son fáciles de encontrar: p. ej. para rs2080161 se puede considerar rs2080162, con frecuencias alélicas idénticas pero sin tractos homopoliméricos en las regiones flanqueantes.

En este estudio se evaluaron detalladamente los resultados de un conjunto de mezclas de ADN de diferentes ratios, formadas por dos componentes. Los AIMs tienen la potencialidad de añadir información útil en la deconvolución de mezclas de ADN, ya que se puede inferir la ancestralidad de los componentes en mezclas simples. La MPS proporciona la posibilidad de detectar mezclas de ADN, ya que las ARF obtenidas constituyen un indicador fiable de la proporción de cada alelo en la muestra inicial. Las gráficas de ARF presentados en la Fig. 65 muestran como, para las diferentes ratios de mezclas, los patrones se diferencian de los presentados en las gráficas individuales.

5. Bloque III: Mezclas de ADN

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. Bloque III: Mezclas de ADN

En este bloque se presentan dos trabajos que desarrollan paneles de marcadores de utilidad para el análisis de mezclas de ADN, optimizados para electroforesis capilar: STRs pentaméricos –sección 5.1– y ID-SNPs multialélicos –sección 5.2–.

5.1 STRs PENTAMÉRICOS

En este trabajo se presenta un panel de 9 STRs pentaméricos con bajas tasas de *stutter*, suplementados con 2 marcadores específicos de cromosoma Y. El panel está optimizado y validado para su uso en genética forense. Los resultados del estudio están siendo recopilados y preparados para su publicación.

5.1.1 Material y métodos

5.1.1.1 STRs pentaméricos candidatos

Se identificaron un total de 10 STRs con repeticiones pentaméricas y características apropiadas de entre los 783 *loci*, que incluyen 643 microsatélites, del set Marshfield (Pemberton *et al.* 2009, Pemberton *et al.* 2013). La selección de STRs se llevó a cabo atendiendo a los siguientes criterios: (i) que fueran autosómicos, (ii) que tuvieran el mayor número de alelos posible de manera que presenten un nivel de polimorfismo útil, (iii) que no tuvieran secuencias de baja complejidad en las regiones flanqueantes y (iv) que presentaran motivos de repetición de 5 pb regulares y uniformes. Los datos genotípicos de los 783 marcadores para el panel HGDP-CEPH (Cann *et al.* 2002) están disponibles públicamente en los portales de Rosenberg lab[34] y Marshfield[35].

Los detalles genómicos de los 10 STRs, que se detallan en la Tabla 27, revelan que el motivo de repetición "AAAA-C/T" es una característica común: un total de 6 *loci* presentan este patrón. Los 4 *loci* restantes presentan un segundo nivel de uniformidad con motivos de repetición "ATA-A/C-A/C". Ninguno de los STRs pentaméricos identificados coincide con STRs pentaméricos previamente establecidos o desarrollados, que presentan los siguientes identificadores y posiciones cromosómicas en base al ensamblaje GRCh37 del genoma humano: Penta B=7:134201476; Penta C=9:37920273; Penta D=21:45056086; Penta E=15:97374245; D10S2325=10:12797054. Dos de los STRs presentan motivos de repetición compuestos: TTTAT02=$[TTTAT]_a[TTCAT]_b[TTTAT]_c$ y TAAAA06=$[TAAAA]_a[ATAAA]_b$. En total, 4 de los STRs pentaméricos presentados en este estudio están situados a menos de 12 Mb de STRs comunes en genética forense (Phillips *et al.* 2012a), este ligamiento físico se

[34] http://www.stanford.edu/group/rosenberglab/data/rosenbergEtAl2005/combinedmicrosats-1048.stru
[35] http://research.marshfieldclinic.org/genetics/genotypingData_Statistics/humanDiversityPanel.asp

debe tener en cuenta a la hora de realizar pruebas de parentesco, pero no cuando se calculan probabilidades de coincidencia de perfiles con fines de identificación combinando las estimaciones de frecuencias de los alelos (Budowle *et al.* 2011, O'Connor *et al.* 2011, Gill *et al.* 2012, Tillmar y Phillips 2017).

Tabla 27. Detalles genómicos de los 10 STRs pentaméricos seleccionados para el desarrollo de una PCR *multiplex*. La línea final en gris muestra el STR que no pudo ser incorporado en el *multiplex* optimizado. ID: identificador. *$[TTTAT]_a[TTCAT]_b[TTTAT]_c$ **$[TAAAA]_a[ATAAA]_b$.

ID Rosenberg	Unidad de repetición	Coordenadas amplicón GRCh37 (hg19)	Nº de repeticiones GRCh37	Tamaño secuencia GRCh37	STR *core* más próximo	Distancia entre STRs (Mb)
GTTTT02	GTTTT	1:768048-768183	9	136	D1S1656	230.14
TTTAT02	TTTAT*	4:8194051-8194321	12	271	FGA	147.31
TAAAA06	TAAAA**	4:31759939-31760203	13	265	FGA	123.75
AAAAC01	AAAAC	9:4177035-4177221	8	187	-	-
TTTTA02	TTTTA	9:138420913-138421045	13	133	-	-
ATAAC02	ATAAC	12:10764618-10764920	11	303	vWA / D12S391	4.67 / 1.69
AACAT01	AACAT	13:77891970-77892198	13	229	D13S317	4.83
ATACC01	ATACC	16:82249557-82249860	11	304	D16S539	4.14
SCA10	ATTCT	22:46191141-46191337	14	197	D22S1045	8.65
AAAAT02	AAAAT	12:133547766-133547881	7	116	vWA / D12S392	127.45 / 121.45

5.1.1.2 Muestras de ADN, datos poblacionales y análisis de datos

Se compilaron genotipos para los 10 STRs candidatos a partir de los disponibles en las bases de datos para el panel HGDP-CEPH y, dado el pequeño tamaño muestral de las poblaciones, se agruparon en 3 superpoblaciones –AFR, EUR y EAS– para obtener estimaciones de las frecuencias alélicas más representativas.

Un total de 103 individuos AFR incluyen: 21 *Central African Republic-Biaka Pygmies*, 13 *Congo-Mbuti Pygmies*, 11 *Kenya-Bantu N.E.*, 6 *Namibia-San*, 22 *Nigeria-Yoruba*, 22 *Senegal-Mandenka* y 8 *South Africa-Bantu*.

Un total de 157 individuos EUR incluyen: 24 *France-Basque*, 28 *France-French*, 12 *Italy (Bergamo)-North Italian*, 28 *Italy-Sardinian*, 8 *Italy-Tuscan*, 15 *Orkney Islands-Orcadian*, 25 *Russia-Russian* y 17 *Russia Caucasus-Adygei*.

Un total de 227 individuos EAS incluyen: 10 *Cambodia-Cambodian*, 10 *China-Dai*, 10 *China-Daur*, 43 *China-Han*, 8 *China-Hezhen*, 8 *China-Lahu*, 10 *China-Miaozu*, 10 *China-Mongola*, 7 *China-Naxi*, 9 *China-Oroqen*, 10 *China-She*, 10 *China-Tu*, 10 *China-Tujia*, 9 *China-Xibo*, 10 *China-Yizu*, 28 *Japan-Japanese* y 25 *Siberia-Yakut*.

Además, se recogieron un total de 94 muestras de ADN de donantes voluntarios y no relacionados de la población gallega –NO España–, a fin de estimar las frecuencias

poblacionales y compararlas con las del grupo EUR y evaluar el rendimiento del *multiplex* optimizado.

Para calcular las frecuencias alélicas, los tamaños de los alelos reportados por Rosenberg fueron convertidos en genotipos atendiendo al número de repeticiones del STR. El cálculo de las frecuencias alélicas y de los diferentes parámetros de informatividad forense se realizó mediante el *software* Promega Powerstats y los valores acumulados se obtuvieron mediante hojas de cálculo.

5.1.1.3 Construcción y optimización del *multiplex*

Los diseños de *primers* incluidos en la Tabla 28 permiten amplificar simultáneamente los 10 STRs candidatos, marcados con los fluorocromos 6FAM (azul), JOE (verde) y TMR (amarillo/negro). Se añadieron colas de ADN no humano en los extremos 5' de los *primers* como modificadores de la movilidad para asegurar la separación electroforética de 4 de los STRs, manteniendo fragmentos de amplificación cortos para aumentar el éxito en la amplificación de ADN degradado. De los 10 STRs candidatos, 9 fueron implementados en el *multiplex* final; el STR AAAAT02 afectó al balance del *multiplex* hasta el punto de no poder ser corregido mediante el ajuste de las concentraciones de los *primers* y fue eliminado de los ajustes posteriores del panel.

Los 9 STRs pentaméricos fueron suplementados con 2 marcadores específicos de cromosoma Y: el STR DYS391 (de Knijff *et al.* 1997) y el Indel rs2032678; ambos incluidos en el kit GlobalFiler™ [36]. Para DYS391, se adaptaron los Set 1 y Set 2 de *primers* listados en STRbase[37], reduciendo el tamaño de los amplicones en 181 pb y 61 pb, respectivamente.

Las secuencias finales de los *primers* diseñados para DYS391 son: CTATTCATTCAATCATACACCCATA y 6FAM-ATAGGTAGGCAGGCAGATAG.

Las secuencias de los *primers* diseñados para el Y-Indel rs2032678 son: CCCAAATCAACTCAACTCCAG y TMR-GATACCTTTGTTTCTGTTCATTCTT. Estos *primers* generan fragmentos de 94 pb para el alelo inserción y de 89 pb para la deleción.

[36] http://tools.lifetechnologies.com/content/sfs/manuals/4477604.pdf
[37] http://www.cstl.nist.gov/strbase/str_y391.htm

Tabla 28. Diseño de *primers* para los 10 STRs pentaméricos. Se indican los genotipos del control de ADN 9947A. Las bases en minúsculas del extremo 5' de los *primers* corresponden a modificadores de movilidad. ID: identificador. Conc.: concentración en el *mix* de *primers*.

ID Rosenberg	Genotipo 9947A	*Primer forward*	*Primer reverse*	Conc. (µM)
GTTTT02	13	GCGACAAAGCAAGACTCCAT	TMR-GCGTAAGCAGGTTTGATGGT	0,15
TTTAT02	9,1	cattttgctgccggttataaCGGAATCACCAGGAACTCTC	6FAM-TTGAACCCAGTAGGCAGAGG	0,05
TAAAA06	13	GGTTTAAGGAGATGAGATACATTA	TMR-CCTCTTTCTGCAACCCTGAA	0,3
AAAAC01	3,5	cattttgctgccggtagttggaCCCCGAAATTTACACCATCATT	TMR-CCTTTGGAGAAAACTGAACTCG	0,2
TTTTA02	12	JOE-AGCCTAGGTGGCAGAGTGA	TGCAGAAGGAAAAGAACTGCAG	0,075
ATAAC02	10,11	JOE-TGGGGACGGTAAGGTTAAATAC	AGTGCAAGTGATCCCTCACC	0,075
AACAT01	12,13	GCTGAGGTAGGAGAACTGCT	JOE-GCTTGCAGGATGTTTGATCC	0,1
ATACC01	14,15	cattttgctgccggtctcgtATGTTTCATCCCGAAGGACATG	6FAM-TTGAACTGAAGAGGCAAGCAG	0,1
SCA10	12,15	cattttgctgccggtAGAAAACAGATGGCAGAATGATAA	6FAM-GCCTGGGCAACATAGAGAGA	0,05
AAAAT02	-	CAAGATCGCGCCACTGCA	6FAM-GGCCAAACTCCCTCCATTAA	-

Los *ladders* alélicos de referencia se construyeron combinando alelos amplificados en *singleplex*, principalmente a partir de muestras heterocigotas. Los *ladders* prototipo, agrupados en función de su fluorescencia, fueron sometidos a electroforesis y se registraron las alturas relativas de los picos para reajustar los componentes y mejorar paulatinamente el balance. Posteriormente, cada *ladder* prototipo fue diluido entre $1x10^{-7}$-$1x10^{-9}$ y reamplificado mediante una PCR de 35 ciclos, siguiendo las condiciones de la *multiplex* optimizada. Los *ladders* prototipo fueron combinados para la CE y rebalanceados para compensar las diferencias de emisión/detección entre los fluorocromos 6FAM, JOE y TMR.

Las amplificaciones *multiplex* se realizaron en un volumen final de 10 µL e incluyen:

- 5 µL de 2x Qiagen Multiplex PCR kit
- 1 µL de *mix* de *primers*
- 1-3 µL de ADN (cantidad óptima ~1 ng)
- 1-3 µL de H_2O

Las concentraciones de cada STR pentamérico en el *mix* de *primers* se listan en la Tabla 28. Para evitar la formación de dímeros de *primers* durante la PCR, se añadió 1 µL de sulfato amónico $(NH_4)_2SO_4$ 200 mM por cada 50 µL de *mix* de *primers*. Los *primers* de DYS391 se añadieron a una concentración final de 1,5 µM y los del Y-Indel a 0,8 µM.

Las reacciones se realizaron en un termociclador GeneAmp® 9700 (AB) bajo las siguientes condiciones: 15 min a 95°C; 28 ciclos de 30 s a 94°C, 60 s a 63°C y 60 s a 65°C; y 65 min a 72°C. Los productos de PCR se prepararon para electroforesis capilar combinando 1 µL del producto amplificado con 9,5 µL de una mezcla 30:1 de Hi-Di™ Formamide (AB) con Internal Lane Standard 500 (Promega), marcado con el fluorocromo CC5. La electroforesis se llevó a cabo en un secuenciador ABI Prism 3130xl Genetic Analyzer usando un capilar de 36 cm, matriz G5 y polímero POP-4™ (AB). Los electroferogramas se visualizaron en el *software* Genemapper v. 4.0 (AB).

5.1.1.4 Medida de las ratios de *stutter* en STRs pentaméricos vs. tetraméricos

Las muestras de ADN de individuos de la población gallega fueron genotipadas con el ensayo optimizado de STRs pentaméricos y, en paralelo, con el kit Promega Powerplex® ESX-17. Las ratios de *stutter* fueron recogidas para los STRs de cada ensayo en función del tipo de fluorocromo y longitud. Ambos sets presentan rangos de tamaños comparables de entre 90 y 300-320 pb, aunque los STRs más largos en ESX-17 tienen tamaños más cortos.

5.1.1.5 Evaluación del rendimiento forense del *multiplex* de STRs pentaméricos

Para evaluar la capacidad del *mutiplex* optimizado para el análisis de ADN degradado, se utilizaron 8 muestras de ADN extraídas a partir de restos esqueléticos: fémur, diente y peroné. En todos los casos, se obtuvieron consentimientos informados de los familiares vivos más cercanos. Los resultados de Powerplex® ESX-17 para dichas muestras habían sido obtenidos previamente.

Las 94 muestras de la población gallega descritas en la sección 5.1.1.2 se utilizaron para evaluar el rendimiento del panel y medir el balance intra- e inter-*locus* del *multiplex* final.

Para evaluar el comportamiento del *multiplex* en mezclas de ADN, se prepararon cuatro mezclas artificiales simples, usando controles de ADN Coriell y combinando pares de muestras en ratios 1:3. Las mezclas de ADN comprenden ADN de dos donantes de diferentes orígenes biogeográficos, a fin de maximizar el número potencial de alelos observados para cada STR. Las mezclas fueron genotipadas con el *multiplex* de STRs pentaméricos y Powerplex® ESX-17, y se aplicó un umbral analítico de 50 RFUs (establecido en el laboratorio para los detectores 3130xl).

Finalmente, para evaluar la sensibilidad del ensayo, se prepararon diluciones seriadas de las muestras de ADN Coriell a concentraciones de 1 ng/µL; 0,5 ng/µL; 0,25 ng/µL; 0,125 ng/µL y 0,063 ng/µL.

5.1.2 **Resultados**

5.1.2.1 Características de la calidad de los perfiles *multiplex*

En la Fig. 73A se muestran los *ladders* alélicos desarrollados para cada uno de los STRs pentaméricos que recoge el ensayo. Los *ladders* de los STRs TTTAT02 y TTTTA02 muestran efectos de adenilación incompleta, produciendo picos de -1 pb y +1 pb, respectivamente. Este problema se podría resolver ajustando las condiciones de PCR (p. ej. una extensión final más larga) o cambiando el marcaje de un *primer* al opuesto. En la Fig. 73B se muestran los perfiles obtenidos para el control de ADN femenino 9947A y el control de ADN Coriell masculino HG00403, alineados con el *ladder* alélico. Los perfiles expandidos se muestran en la Fig. 74 y Fig. 75, respectivamente. Aunque los diseños de *primers* iniciales presentaban artefactos derivados de interacciones entre los mismos, un rediseño selectivo eliminó todos estos artefactos, menos un pico no específico entre las posiciones de los alelos de AAAAC01

que migra consistentemente a 208,3 pb. Este pico se marca en el perfil de 9947A de la Fig. 73B y es más evidente en el perfil femenino, aunque puede ser fácilmente descontado debido a que no se corresponde con ninguna posición alélica.

El genotipado de las 94 muestras de la población gallega produjo entre 57 y 73 heterocigotos por *locus*, indicando un balance alélico intra-*locus* promedio, medido como ratio de altura de los picos –PHR: *peak height ratio*– en el rango 1-1,3. No obstante, estos valores promedio se ven afectados desproporcionalmente por el STR TAAAA06. El PHR promedio inter-*locus* para cada marcaje fluorescente fue de 1,02 para 6FAM; 1,12 para JOE y 1,16 para TMR (con una desviación estándar de ~0,1 en cada caso).

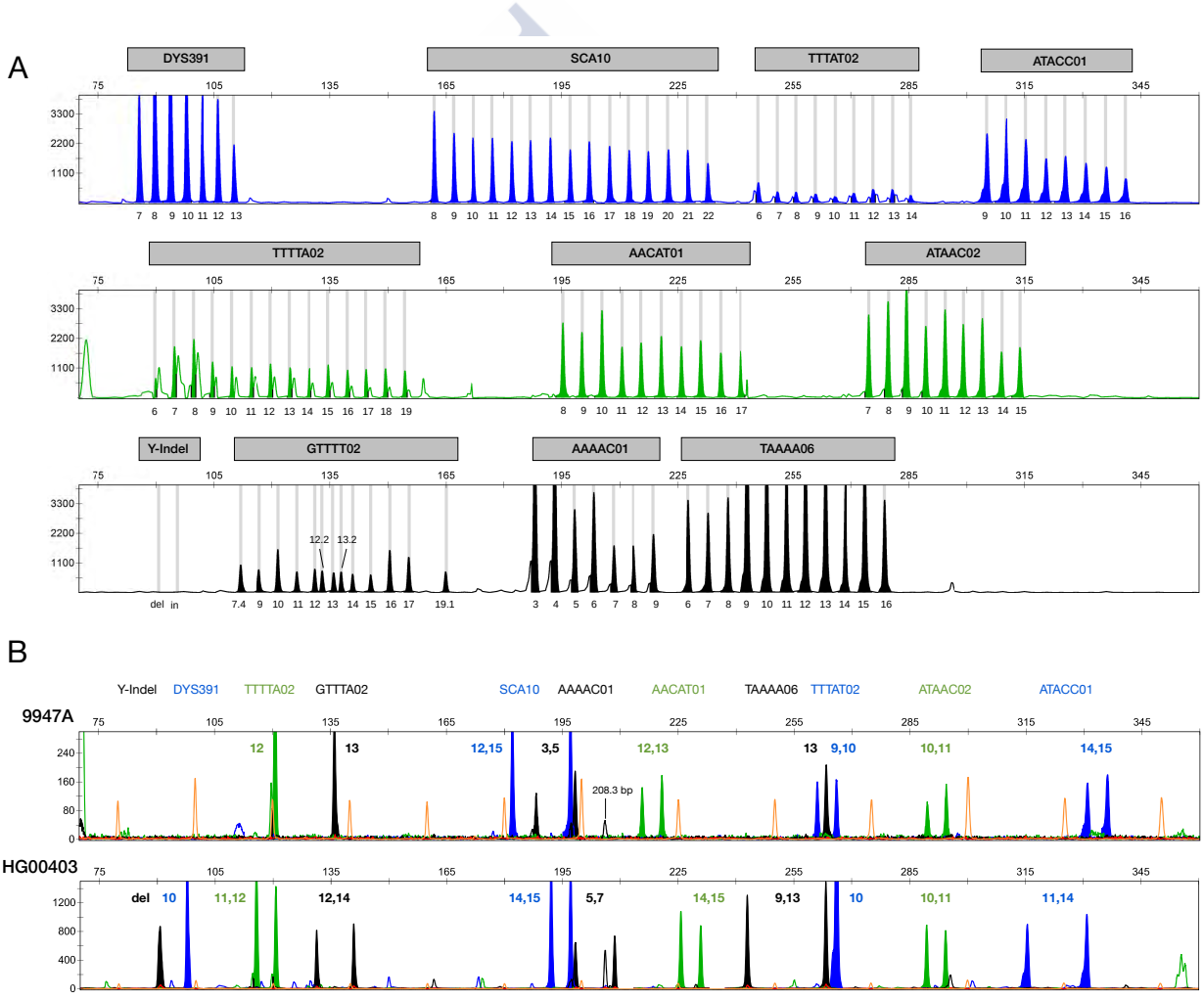


Fig. 73. A) *Ladder* alélico construido para los 9 STRs pentaméricos. El número de repeticiones se muestra debajo de cada pico. B) Perfiles típicos del ensayo 11-plex, análisis realizados con 1 ng de ADN inicial del control de ADN 9947A (femenino) y el control de ADN Coriell HG00403 (masculino). Los perfiles expandidos se muestran en la Fig. 74 y Fig. 75.

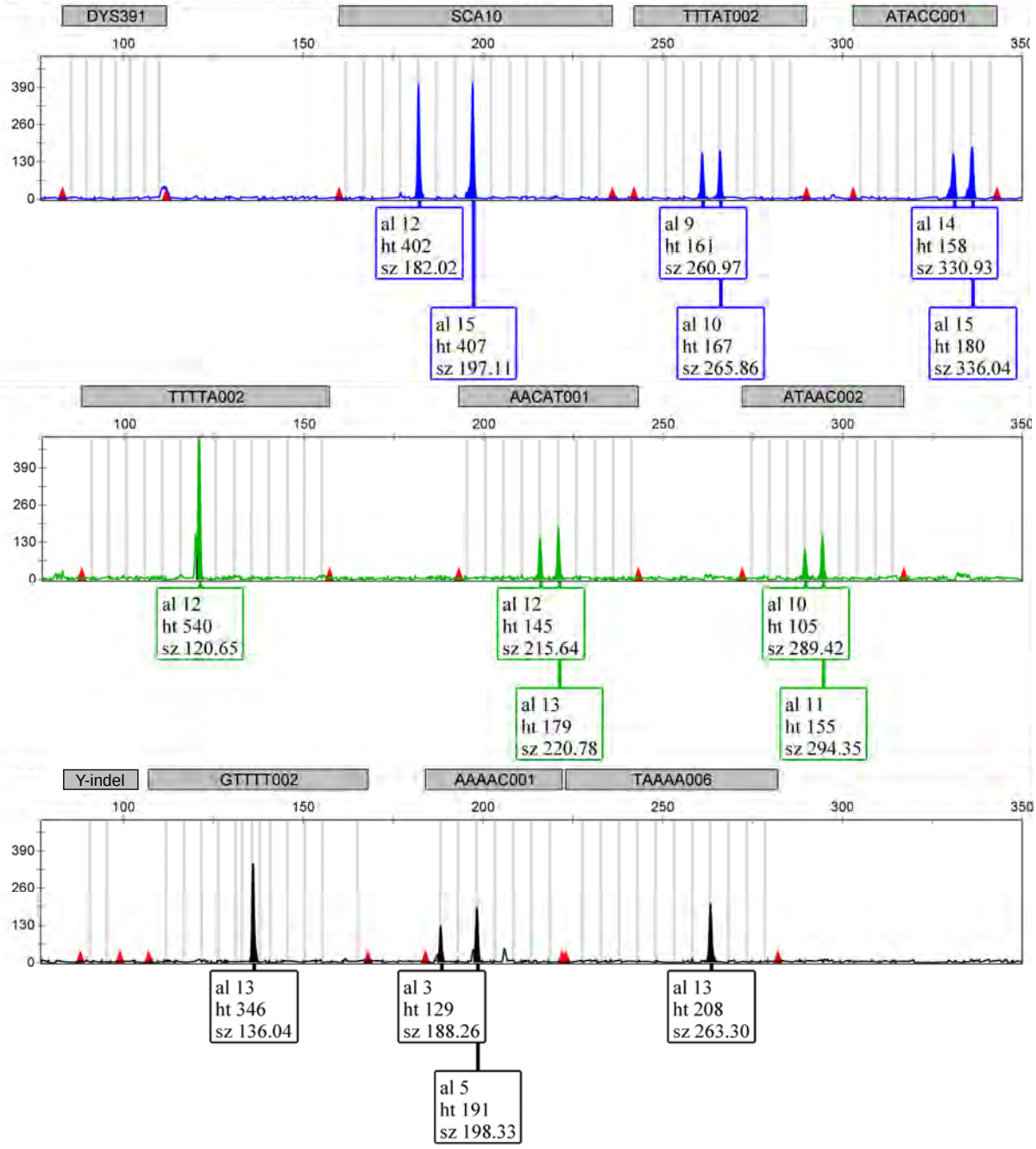


Fig. 74. Perfil expandido de 11-plex del control de ADN 9947A (1 ng).

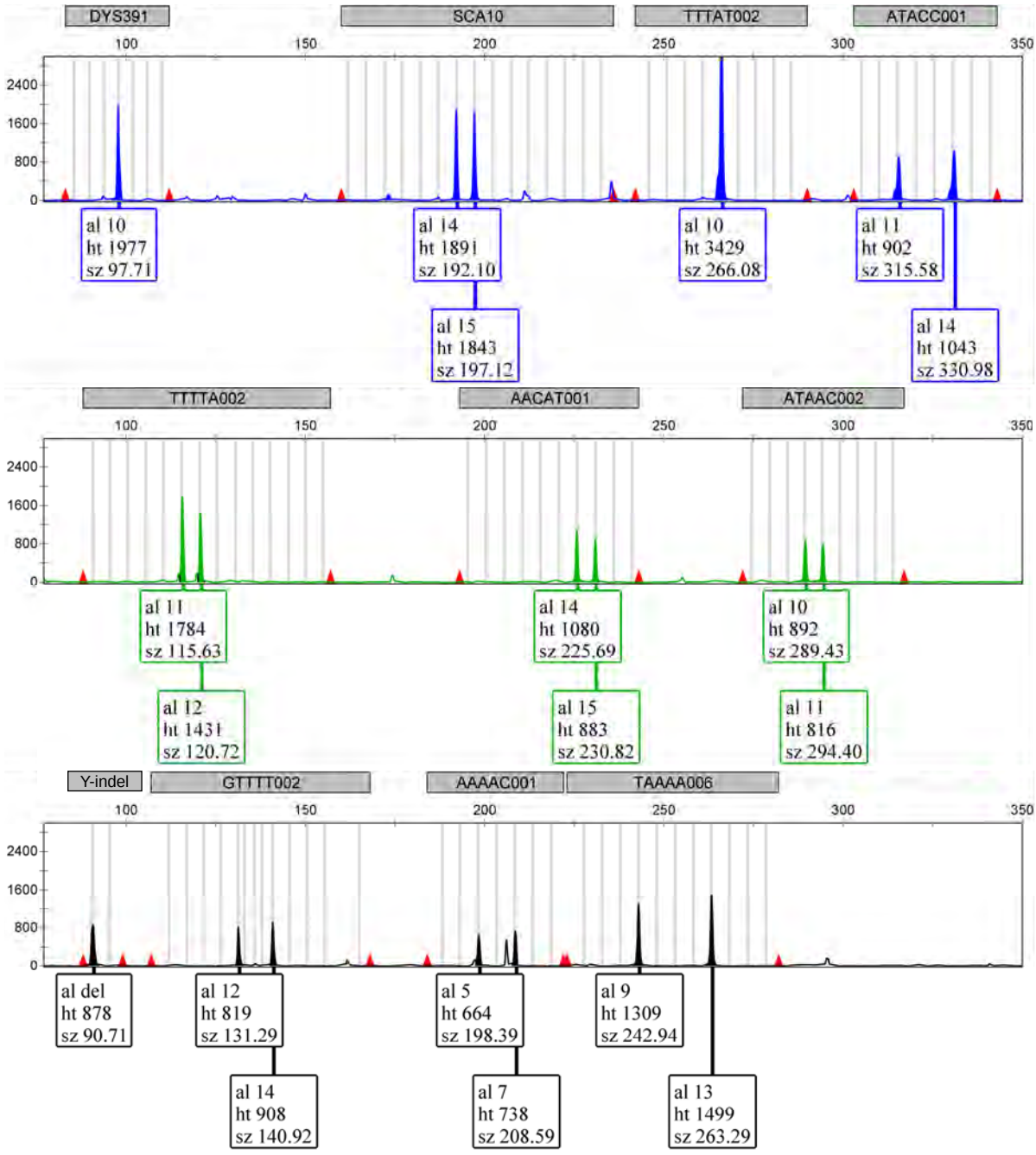


Fig. 75. Perfil expandido de 11-plex del control de ADN Coriell HG00403 (1 ng).

5.1.2.2 Patrones de variación poblacional e informatividad forense

Las distribuciones de las frecuencias alélicas de los 9 STRs pentaméricos incorporados con éxito en el *multiplex* para los conjuntos de poblaciones AFR, EUR y EAS del panel HGDP-CEPH se representan en la Fig. 76 y se listan en la Tabla 29.

Los STRs pentaméricos presentan diversos niveles de polimorfismos con SCA10, ATAAC02, TTTTA02, TAAAA06 y AACAT01 indicando los niveles más altos de informatividad global (aunque TAAAA06 presenta un nivel de polimorfismo más bajo en la población EAS, con las frecuencias agrupadas en los alelos 9 y 13). Los 5 STRs pentaméricos más informativos alcanzan niveles de poder de discriminación –Dp: *discrimination power*– promedio ≥90% (93,8%; 91,9%; 91,8%; 89,9% y 89,9%; respectivamente). Las RMPs acumuladas de los 9 STRs pentaméricos se resumen en la Fig. 77A, en la que se comparan con los valores de los 15 STRs incluidos en el kit Identifiler™ y los 16 de NGM SElect™ (AB). Los 9 STRs pentaméricos presentan valores acumulados de RMP entre 2-3 órdenes de magnitud menores que los 9 mejores marcadores de Identifiler™ NGM SElect™.

A excepción de SCA10, el nivel de informatividad forense de los *loci* pentaméricos está muy por debajo del de los *loci* comúnmente utilizados en genética forense. En la Fig. 77B se muestra una comparación entre los 9 STRs pentaméricos, 18 STRs del CODIS-A (Hares 2015), STRs extra de GlobalFiler™ y los Pentas B, C, D y E. Así, se comparan los valores promedio de Dp de los STRs para los 3 grupos poblacionales del HGDP-CEPH con el valor promedio de Dp de 0,926 de los STRs del CODIS-A, utilizando datos de otros estudios (Phillips *et al.* 2011, Phillips *et al.* 2013c, Phillips *et al.* 2014b). De entre los nuevos STRs pentaméricos evaluados, únicamente SCA10 sobrepasa el valor promedio de Dp del CODIS-A, aunque ATAAC01 y GTTTT02 se acercan a dicho valor. La mayoría del resto de los STRs pentaméricos se encuentran entre los de menor Dp de entre los evaluados; de hecho, AAAAC01 es menos informativo que TPOX, el marcador con menor Dp del CODIS-A.

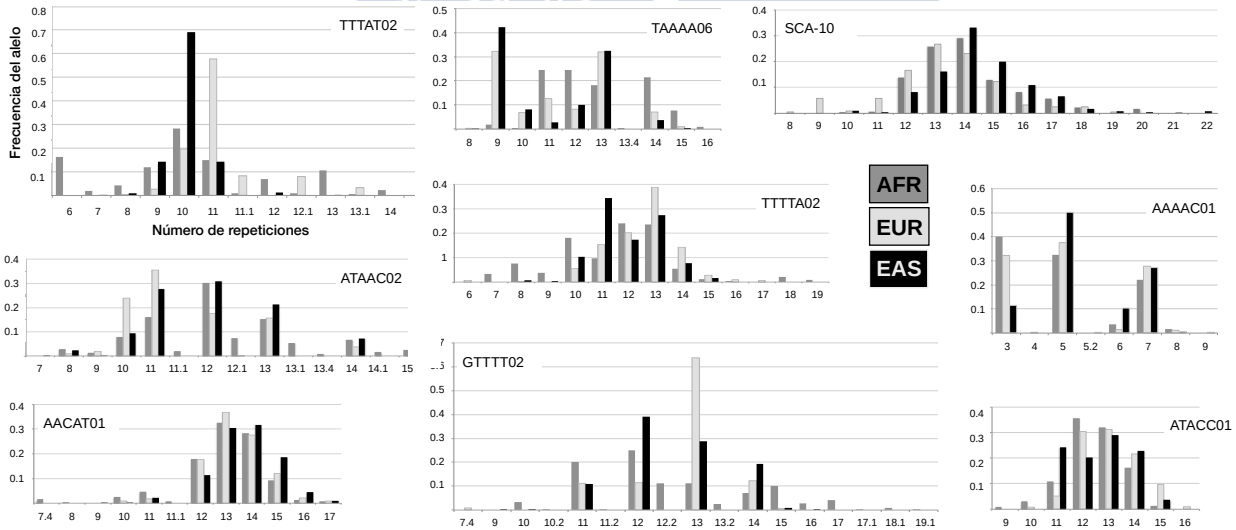


Fig. 76. Gráficos de barras que representan las estimaciones de frecuencias presentadas en la Tabla 29.

Tabla 29. Parámetros de informatividad forense y estimaciones de las frecuencias alélicas de 9 STRs pentaméricos a partir de los datos del panel HGDP-CEPH, combinando las poblaciones como se describe en la sección 5.1.1.2. RMP= *random match probability*. Dp=*discrimination power*. PE= probabilidad de exclusión. Het= heterocigosidad.

TTTTA02	AFR	EUR	EAS
RMP	0,043	0,252	0,123
Dp	0,958	0,748	0,877
PE	0,599	0,205	0,265
% Het	80	51,9	57,8
Alelos	AFR	EUR	EAS
6		0,007	
7	0,0336		
8	0,0756	0,0035	0,0072
9	0,0378		0,0048
10	0,1807	0,0559	0,1039
11	0,0966	0,1538	0,343
12	0,2395	0,2028	0,1739
13	0,2353	0,3881	0,2729
14	0,0546	0,1434	0,0773
15	0,0126	0,028	0,0169
16	0,0042	0,0105	
17		0,007	
18	0,021		
19	0,0084		

TAAAA06	AFR	EUR	EAS
RMP	0,0761	0,0987	0,1271
Dp	0,9239	0,9013	0,8729
PE	0,4366	0,5521	0,372
% Het	70,54	77,42	66,2
Alelos	AFR	EUR	EAS
8		0,0032	0,0023
9	0,0179	0,3175	0,4236
10	0,0045	0,0635	0,081
11	0,2455	0,1302	0,0278
12	0,2455	0,0889	0,0995
13	0,183	0,3111	0,3241
13,4	0,0045		
14	0,2143	0,0698	0,037
15	0,0759	0,0159	0,0046
16	0,0089		

TTTAT02	AFR	EUR	EAS
RMP	0,1217	0,1074	0,1015
Dp	0,8783	0,8926	0,8985
PE	0,3913	0,5611	0,5639
% Het	67,54	77,92	78,08
Alelos	AFR	EUR	EAS
6	0,1637		
7	0,0177	0,0033	0,0023
8	0,0442	0,0267	0,0093
9	0,1195	0,1967	0,1419
10	0,2832	0,5767	0,6907
11	0,1504	0,0833	0,1419
11,1	0,0088		
12	0,0708	0,08	0,0116
12,1	0,0088		
13	0,1062	0,0333	0,0023
13,1	0,0044		
14	0,0221		

ATAAC02	AFR	EUR	EAS
RMP	0,0486	0,0992	0,0945
Dp	0,9514	0,9008	0,9055
PE	0,6811	0,5432	0,5414
% Het	84,3	76,92	76,82
Alelos	AFR	EUR	EAS
7			0,0023
8	0,0289	0,0096	0,0227
9	0,0124	0,0192	0,0023
10	0,0785	0,2404	0,0932
11	0,1612	0,3558	0,2773
11,1	0,0207		
12	0,3017	0,1763	0,3091
12,1	0,0744	0,0032	
13	0,1529	0,1571	0,2136
13,1	0,0537		
13,4	0,0083		
14	0,0661	0,0385	0,0727
14,1	0,0165		
15	0,0248		0,0068

SCA10	AFR	EUR	EAS
RMP	0,0524	0,0663	0,0686
Dp	0,9476	0,9337	0,9314
PE	0,5781	0,6084	0,5906
% Het	78,86	80,5	79,55
Alelos	AFR	EUR	EAS
8	0,0041		
9	0,0569		
10	0,0081	0,0031	0,0091
11	0,0569	0,0063	0,0046
12	0,1667	0,1384	0,0818
13	0,2683	0,2579	0,1614
14	0,2317	0,2893	0,3318
15	0,122	0,1289	0,2
16	0,0325	0,0818	0,1091
17	0,0244	0,0566	0,0659
18	0,0244	0,0220	0,0159
19	0,0041		0,0068
20		0,0157	0,0045
21			0,0023

ATACC01	AFR	EUR	EAS
RMP	0,1217	0,1074	0,1015
Dp	0,8783	0,8926	0,8985
PE	0,3913	0,5611	0,5639
% Het	67,54	77,92	78,08
Alelos	AFR	EUR	EAS
9	0,0088		
10	0,0307	0,0065	
11	0,1096	0,0519	0,242
12	0,3553	0,3052	0,2032
13	0,3202	0,3117	0,29
14	0,1623	0,2175	0,2283
15	0,0132	0,0974	0,0365
16		0,0097	

AAAAC01	AFR	EUR	EAS
RMP	0,173	0,178	0,17
Dp	0,827	0,822	0,83
PE	0,398	0,419	0,276
% Het	68	69,4	58,7
Alelos	AFR	EUR	EAS
3	0,402	0,322	0,115
4		0,003	
5	0,324	0,376	0,502
5,2			0,002
6	0,037	0,013	0,103
7	0,221	0,277	0,271
8	0,016	0,01	0,005
9			0,002

AACAT01	AFR	EUR	EAS
RMP	0,0921	0,1073	0,1033
Dp	0,9079	0,8927	0,8967
PE	0,5098	0,4224	0,5594
% Het	75	69,62	77,83
Alelos	AFR	EUR	EAS
7,4	0,0167		
8	0,0042		
9			0,0024
10	0,025	0,01	0,0024
11	0,0458	0,019	0,0212
11,1	0,0083		
12	0,1792	0,1772	0,1132
13	0,325	0,3671	0,3042
14	0,2833	0,2753	0,316
15	0,0917	0,1203	0,1863
16	0,0125	0,0222	0,0448
17	0,0083	0,0095	0,0094

GTTTT02	AFR	EUR	EAS
RMP	0,0526	0,0925	0,1012
Dp	0,9474	0,9075	0,8988
PE	0,5353	0,4832	0,6299
% Het	70,54	77,42	66,2
Alelos	AFR	EUR	EAS
7,4		0,009	
9			0,002
10	0,033		0,002
10,2	0,004		
11	0,2	0,111	0,11
11,2	0,004		
12	0,25	0,114	0,392
12,2	0,113		
13	0,113	0,636	0,289
13,2	0,025		
14	0,071	0,123	0,1930
15	0,1	0,006	0,009
16	0,029		0,0020
17	0,042		
17,1	0,004		
18,1	0,008		
19,1	0,004		

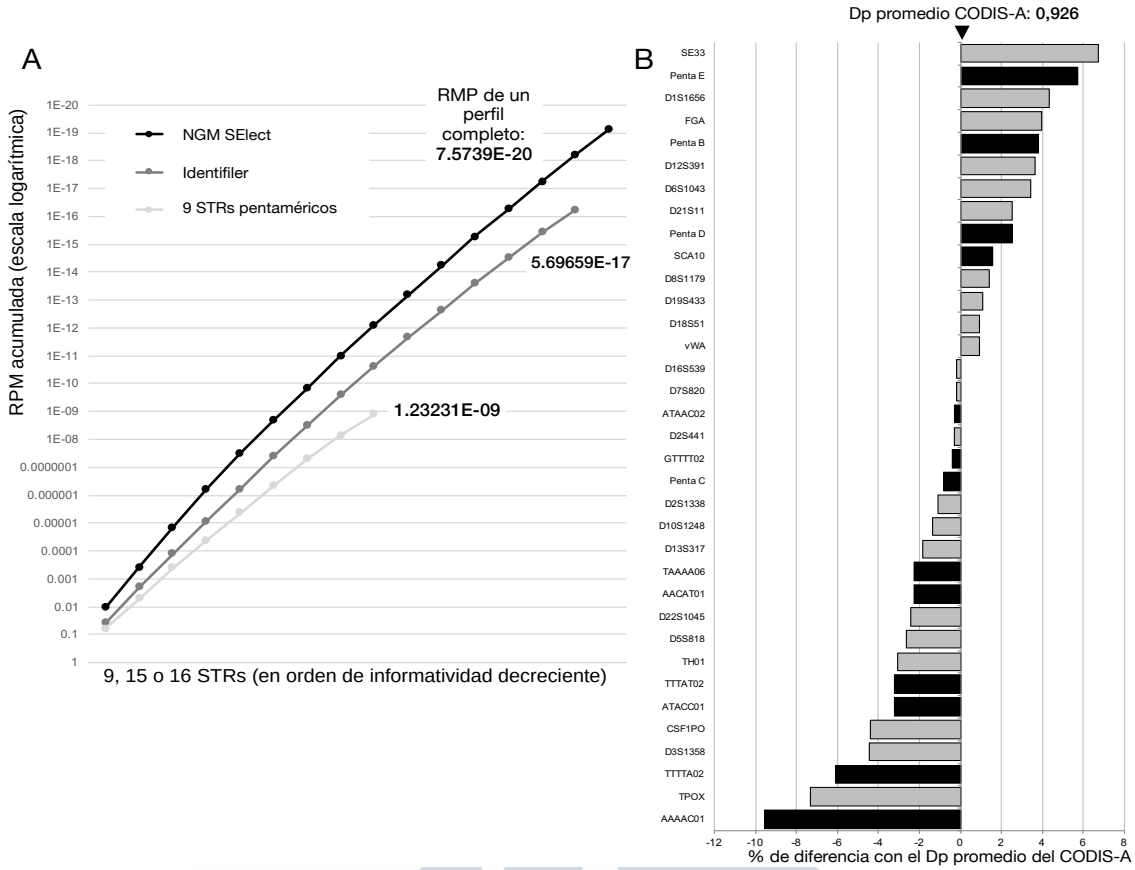


Fig. 77. A) RMP –*random match probability*– acumulada de los 9 STRs pentaméricos en comparación con los 15 STRs de Identifiler™ y los 16 de NGM SElect™. Los valores de RPM se calculan como promedio de los de AFR, EUR y EAS –Tabla 29–. Los STRs se ordenan de más a menos informativos y se muestran los valores de RMP que se espera obtener de perfiles completos. B) Comparación del poder de discriminación (Dp) de 22 STRs comúnmente utilizados (en gris) y 13 STRs pentaméricos (en negro) en relación al promedio de Dp del CODIS-A (0,926). Los valores de Dp se calcularon como promedio de los de AFR, EUR y EAS –Tabla 29–.

En la Tabla 30 se muestran las estimaciones de las frecuencias alélicas de los 9 STRs pentaméricos para la población gallega. Los test exactos de equilibrio Hardy-Weinberg indican que no existe una desviación detectable del equilibrio (p=0,00556; corrección de Bonferroni). Las estimaciones se corresponden adecuadamente con las obtenidas a partir de datos del panel HGDP-CEPH para la población EUR, con una correlación de R^2=0,927.

Finalmente, el alelo deleción del Y-Indel rs2032678 no se detectó en ninguna de las muestras de la población gallega. No obstante, se descubrió la deleción en el control de ADN Coriell HG00403 –ver Fig. 73 y Fig. 75–, de origen EAS (*Southern Han Chinese*). Este Indel no se encuentra listado en la base de datos del Proyecto 1000 Genomas, pero en la colección de cromosoma-Y Stanford Oefner se observaron 63 individuos (de origen desconocido) con la inserción TTCTC y 9 con la deleción (identificadas como EAS), correspondiéndose a frecuencias de 0,875 y 0,125; respectivamente. Por tanto, se necesitan estudios adicionales que determinen la frecuencia del Indel rs2032678 en los diferentes grupos poblacionales, ya que probablemente pueda aportar información adicional sobre la población de origen además de cumplir su papel como marcador específico de ADN masculino.

Tabla 30. Estimaciones de las frecuencias alélicas de los 9 STRs pentaméricos en la población gallega. Al.:alelo. Frec: frecuencia.

SCA10	
Al.	Frec
9	0,0053
10	0,0000
11	0,0319
12	0,1702
13	0,2766
14	0,3032
15	0,0904
16	0,0691
17	0,0372
18	0,0053
20	0,0053

ATACC01	
Al.	Frec
10	0,0053
11	0,0798
12	0,3138
13	0,3085
14	0,2181
15	0,0691
16	0,0053

GTTTT02	
Al.	Frec
7,4	0,0000
11	0,1543
12	0,1436
12.2	0,0053
13	0,5691
14	0,1117
15	0,0106
17	0,0053

TTTAT02	
Al.	Frec
7	0,0000
8	0,0000
9	0,2713
10	0,5213
11	0,0851
12	0,0745
13	0,0372
14	0,0106

AAAAC01	
Al.	Frec
3	0,3723
4	0,0000
5	0,3617
6	0,0106
7	0,2394
8	0,0160

TTTTA02	
Al.	Frec
6	0,0213
8	0,0053
9	0,0053
10	0,0160
11	0,2128
12	0,1915
13	0,3883
14	0,1117
15	0,0266
16	0,0160
17	0,0000
18	0,0053

ATAAC02	
Al.	Frec
8	0,0000
9	0,0106
10	0,2340
11	0,3404
12	0,2181
12.1	0,0053
13	0,1277
14	0,0532
15	0,0106

AACAT01	
Al.	Frec
10	0,0160
11	0,0372
12	0,2181
13	0,2926
14	0,3138
15	0,0957
16	0,0266
17	0,0000

TAAAA06	
Al.	Frec
8	0,0000
9	0,3245
10	0,0851
11	0,1383
12	0,0851
13	0,3298
14	0,0372
15	0,0000

5.1.2.3 Comparación de las ratios de *stutter*

Los valores de las ratios de *stutter* en los 9 STRs pentaméricos de este panel y 11 STRs tetraméricos y uno trimérico del kit Powerplex® ESX-17 se recogen en la Fig. 78. La comparación de las ratios de *stutter* de los 9 STRs pentaméricos y 12 STRs comunes de tamaño equivalente indican una reducción de los picos *stutter* en los STRs pentaméricos. Únicamente los STRs pentaméricos ATAAC02 y SCA10 exceden marginalmente una tasa de *stutter* del 3%; y tan solo el STR tetramérico TH01 presenta valores comparables a la tasa media de *stutter* de los STRs pentaméricos analizados, con una ratio de *stutter* del ~2% (debida en gran parte al efecto del alelo 9.3). El resto de los *loci* de Powerplex® ESX-17 analizados presentan una actividad *stutter* notablemente más alta. En conjunto, los STRs pentaméricos tienen tasas de *stutter* más de tres veces menores (3,4) que los 12 *loci* de Powerplex® ESX-17 analizados, con tasas promedio del 2,15% y 7,32%, respectivamente.

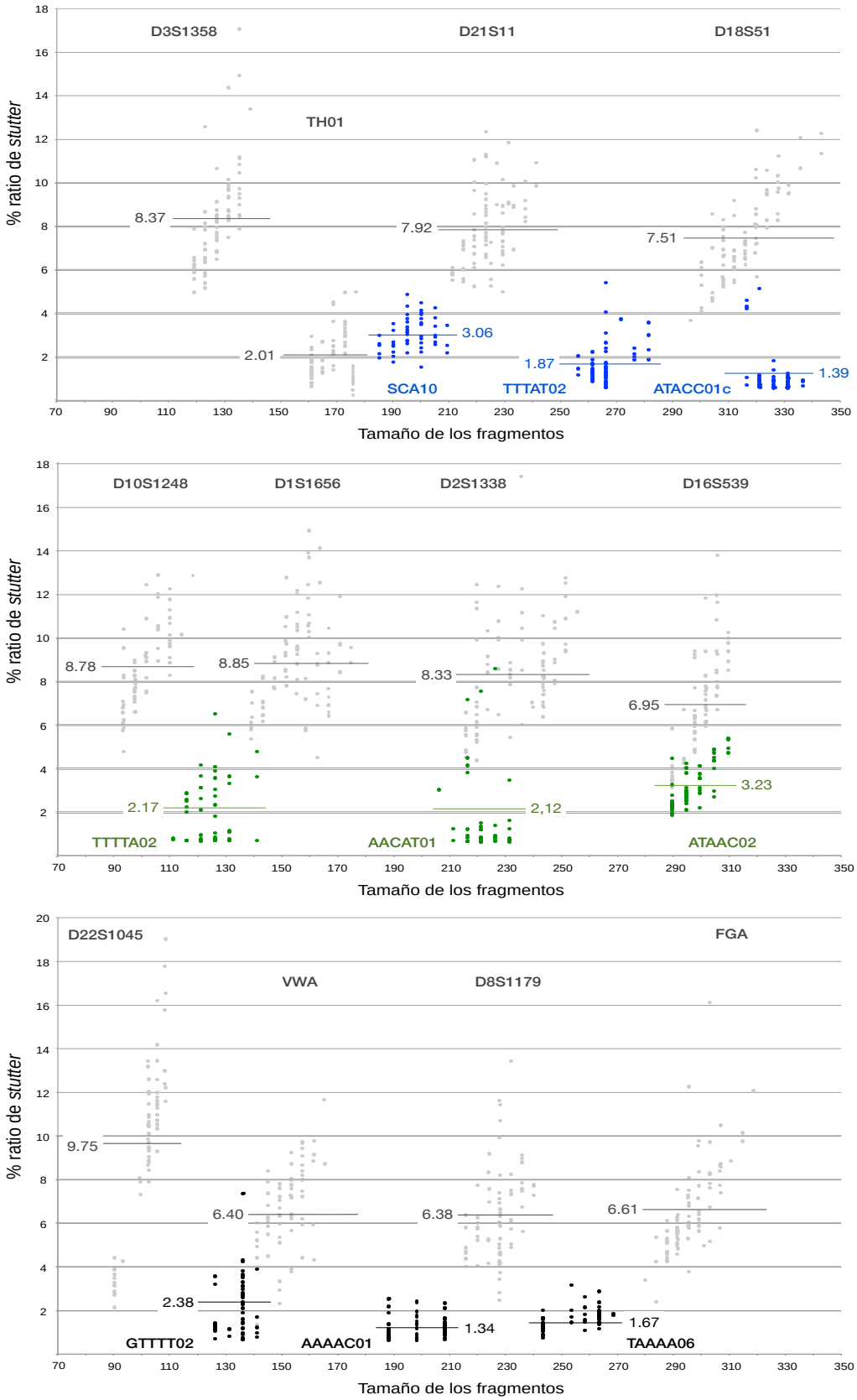


Fig. 78. Ratios de *stutter* (altura del *stutter*/altura del alelo) obtenidos del análisis de 94 muestras con 9 STRs pentaméricos y 12 de los 16 STRs de Promega Powerplex® ESX-17. Los 3 fluorocromos se muestran por separado. Los valores promedio para cada STR se indican dentro de cada grupo.

5.1.2.4 Evaluación del rendimiento forense del panel

Las muestras de ADN extraídas a partir de material esquelético produjeron resultados similares a los obtenidos originalmente con Powerplex® ESX-17 con el panel de STRs pentaméricos. Una muestra altamente inhibida presentó genotipos para únicamente 2 de los 9 STRs pentaméricos, dos muestras mostraron perfiles completos y las restantes 5 muestras perfiles casi completos, con *drop-out* consistente en los marcadores ATACC01 y ATAAC02 (con tamaños de amplicón >300 pb). Los perfiles de Powerplex® ESX-17 fueron ligeramente más completos en todas las muestras de ADN degradado analizadas, aunque las muestras más comprometidas produjeron perfiles incompletos en ambos paneles. Las formulaciones mejoradas del *buffer* de PCR de los kits comerciales, diseñadas para controlar los efectos de inhibición en muestras como las estudiadas, podrían influir en la tasa de éxito de amplificación cuando se comparan kits comerciales con paneles personalizados. Por ello, los reactivos de PCR suplementarios diseñados para controlar la inhibición, como 5X AmpSolution™ Reagent de Promega, pueden suponer importantes mejoras en la optimización de *multiplexes* personalizados.

Las diluciones seriadas del control de ADN indican que el ensayo tiene una sensibilidad adecuada para su uso en genética forense. Se obtuvieron perfiles completos con cantidades iniciales de ADN de hasta 0,063 ng. No obstante, la cantidad inicial de ADN óptima es de 1 ng de ADN.

5.1.2.5 Evaluación de mezclas de ADN artificiales

Se evaluaron una serie mezclas de ADN de ratio 1:3 para realizar una comparación inicial exploratoria entre los STRs pentaméricos y el kit Powerplex® ESX-17. Los resultados de la comparación entre los dos paneles refuerzan el concepto de que la actividad *stutter* reducida de los STRs pentaméricos facilita la interpretación de las mezclas de ADN, cuando se utilizan estos marcadores en combinación con los STRs tetraméricos establecidos.

La Tabla 31 muestra los datos típicos de altura de picos de una de las mezclas de ADN para el kit Powerplex® ESX-17 y los STRs pentaméricos. Ambos paneles presentan indicadores consistentes que permiten la detección de mezclas de ADN. Sin embargo, los perfiles de Powerplex® ESX-17 presentan 5-6 picos por encima de los 50 RFUs en la mayoría de los STRs, lo que conlleva el riesgo de sobreestimar el número mínimo de contribuyentes si no se descuentan adecuadamente las señales no alélicas de las posiciones de los picos *stutter*. Los análisis de las mismas muestras con el panel 11-plex presentado en este trabajo presentan un número máximo de 4 picos por *locus*, de manera que la deconvolución de la mezcla resulta más sencilla. De hecho, los perfiles de 11-plex permitieron asignar los alelos observados a ambos contribuyentes de la mezcla en todos los casos.

En cada perfil de las mezclas de ADN, al menos un 75% de los STRs pentaméricos mostraron más de 2 picos, a pesar de la variabilidad reducida de los mismos. Se debe destacar que en varias ocasiones los picos alélicos del contribuyente minoritario coincidieron con

posiciones *stutter* en ambos paneles, pero los STRs pentaméricos fueron más sencillos de detectar debido a que las RFUs de estos picos eran demasiado altas en comparación con las RFUs del alelo del que provendría el *stutter* -1.

Tabla 31. Resultados del análisis de una de las mezclas para Powerplex® ESX-17 y los STRs pentaméricos. Se reportan todas las señales alélicas por encima de 50 RFUs (Al.1, Al.2, Al.3…) y las alturas de las mismas en RFUs (H.1, H.2, H.3…)

		Señal alélica >50 RFUs								Altura de los picos (RFUs)							
		Al.1	Al.2	Al.3	Al.4	Al.5	Al.6	Al.7	Al.8	H.1	H.2	H.3	H.4	H.5	H.6	H.7	H.8
Powerplex® ESX-17	AMEL	X	Y	–	–	–	–	–	–	7980	2693	–	–	–	–	–	–
	D3S1358	13	14	15	16	17	–	–	–	685	8001	6574	328	3484	–	–	–
	TH01	5	6	9	9.3	–	–	–	–	160	8155	3285	3531	–	–	–	–
	D21S11	28	29	–	–	–	–	–	–	3319	7822	–	–	–	–	–	–
	D18S51	11	12	13	15	16	17	–	–	308	3549	5763	440	5765	2695	–	–
	D10S1248	12	13	14	15	–	–	–	–	437	4278	6408	3096	–	–	–	–
	D1S1656	11	12	13	14	17	17.3	–	–	530	7292	693	6170	2607	2360	–	–
	D2S1338	15	16	17	18	19	20	24	25	295	6814	519	181	2681	3664	731	7645
	D16S539	8	9	10	11	12	–	–	–	139	4351	273	5106	5142	–	–	–
	D22S1045	11	15	16	17	–	–	–	–	1448	3750	358	2186	–	–	–	–
	vWA	14	16	17	18	19	–	–	–	1511	290	3955	4566	2664	–	–	–
	D8S1179	9	10	12	13	14	19	–	–	129	3035	315	5731	2011	248	–	–
	FGA	17	18	20	21	22.1	23.1	24.1	–	113	3721	178	3898	2017	157	1323	–
	D2S441	10	11	13	14	–	–	–	–	361	6878	284	6749	–	–	–	–
	D12S391	17.3	18.3	20	21	22	–	–	–	690	8504	4837	7335	5361	–	–	–
	D19S433	11	12	13	14	–	–	–	–	331	6043	771	8289	–	–	–	–
	SE33	16	17	18	26.2	27.2	28.2	–	–	306	4420	2180	121	1434	1196	–	–
STRs pentaméricos	DYS391	9	10	–	–	–	–	–	–	116	1483	–	–	–	–	–	–
	SCA10	12	13	14	17	–	–	–	–	109	2194	855	1954	–	–	–	–
	TTTAT002z	1	2	–	–	–	–	–	–	1012	999	–	–	–	–	–	–
	ATACC001	11	12	15	–	–	–	–	–	210	993	633	–	–	–	–	–
	TTTTA002	10	13	14	15	–	–	–	–	702	1167	1056	729	–	–	–	–
	AACAT001	12	13	–	–	–	–	–	–	1068	1672	–	–	–	–	–	–
	ATAAC002	10	11	12	13	–	–	–	–	170	531	613	217			–	–
	rs2032678	In	–	–	–	–	–	–	–	309	–	–	–	–	–	–	–
	GTTTT002	11	12	13	–	–	–	–	–	359	300	1533	–	–	–	–	–
	AAAAC001	3	5	7	–	–	–	–	–	749	255	796	–	–	–	–	–
	TAAAA006	9	12	13	–	–	–	–	–	1450	945	594	–	–	–	–	–

5.1.3 **Discusión**

Los 9 STRs pentaméricos que han sido incorporados con éxito en el nuevo *multiplex* presentan una actividad *stutter* marcadamente más baja; más de tres veces menor que los STRs tetraméricos de tamaño equivalente. El ensayo está adecuadamente optimizado para poder obtener perfiles de buena calidad a partir de muestras de ADN *low level* o extraídas de material esquelético, típicas de los casos de rutina forense.

Las observaciones para estos nuevos *loci* son similares a las primeras descripciones de los STRs pentaméricos desarrollados para aplicaciones forenses, que reportaron niveles de

tasas de *stutter* por debajo del 2%, mientras que las tasas de *stutter* de los STRs tetraméricos se sitúan entre el 2-10%, con ciertos alelos excediendo el 15%[38]. Todos los STRs pentaméricos desarrollados con fines forenses comparten la característica de la baja tasa de *stutter*, que potencialmente simplifica el análisis de las mezclas de ADN (como se comprueba en el análisis exploratorio presentado en este trabajo), especialmente en casos complejos que impliquen perfiles de múltiples donantes. En la mayoría de casos, el número de picos identificados en los STRs pentaméricos proporciona una inferencia correcta del número mínimo de contribuyentes a la mezcla. La dificultad para diferenciar los alelos del componente minoritario de picos *stutter* en los análisis de STRs se incrementa a medida que la ratio de mezcla es más extrema, hasta el punto en que la altura de ambos es semejante: la reducción de la actividad *stutter* permite diferenciarlos más fácilmente.

A pesar de la reducción del poder de discriminación que presentan la mayoría de los STRs pentaméricos a causa de su menor variabilidad, los 9 nuevos STRs y los dos *loci* específicos de cromosoma Y permiten expandir los datos disponibles para los análisis. Por lo tanto, el ensayo que constituye un avance en la búsqueda de *loci* adicionales que puedan ser aplicados a cuando los kits de STRs de elección no proporcionen información suficiente.

Dado que actualmente se pueden genotipar más de 20 STRs autosómicos en un único *multiplex*, como con los kits GlobalFiler™ o Promega Powerplex® Fusion, puede parecer innecesario el desarrollo de nuevos STRs. No obstante, debido a la habitual presencia de mezclas de ADN en los casos de rutina, es necesario seguir revisando y perfeccionando la metodología de obtención de perfiles a partir de este tipo de muestras. Una identificación eficiente de los picos en posiciones *stutter* como artefactos y no como alelos de un componente minoritario puede proporcionar una interpretación más segura de los perfiles más complejos.

Se debe destacar también que, dado que los *stutter* son un fenómeno derivado de la PCR, seguirán constituyendo un problema en la interpretación de mezclas aunque los secuenciadores de CE sean reemplazados por sistemas de MPS. Una de las ventajas de los sistemas MPS en relación al genotipado de STRs es el potencial para analizar un mayor número de marcadores mediante una única reacción de amplificación; de manera que aumenta el interés en el desarrollo de nuevos STRs suplementarios, más aún si no solo incrementan el poder de discriminación sino que poseen características de amplificación especiales que permiten mejorar, p. ej., el análisis de mezclas de ADN.

[38] https://www.promega.com/-/media/files/resources/conference-proceedings/ishi-09/oral-presentations/08.pdf?la=en

5.2 ID-SNPS MULTIALÉLICOS

En este trabajo se presenta un panel SNaPshot de ID-SNPs multialélicos. Las características de estos marcadores permiten, por una parte, elevar el poder de discriminación individual de los SNPs y, por la otra, favorecer la detección de mezclas de ADN. Los resultados del estudio están siendo recopilados y preparados para su publicación.

5.2.1 **Material y métodos**

5.2.1.1 Datos poblacionales y muestras de ADN

Los datos genotípicos de individuos de las 5 poblaciones definidas continentalmente (AFR, EUR, EAS, OCE, AMR) para los SNPs seleccionados se obtuvieron a través de dos fuentes: la base de datos del Proyecto 1000 Genomas Fase III (The Genomes Project Consortium 2015) y el genotipado del panel de HGDP-CEPH (Cann *et al.* 2002) con el *multiplex* optimizado. Los cálculos de las frecuencias poblacionales y los diferentes parámetros de informatividad de los SNPs seleccionados se realizaron en hojas de cálculo.

Para los grupos AFR, EUR y EAS se combinaron los datos de todas las poblaciones recogidas en la Fase III del Proyecto 1000 Genomas, exceptuando las poblaciones *admixed* ACB y ASW. Teniendo en cuenta el pequeño tamaño muestral del panel HGDP-CEPH se combinaron las poblaciones para obtener un total de 28 individuos OCE (17 *Papuan from New Guinea* y 11 *Melanesian from Bougainville*) y 64 individuos AMR (14 *Karitiana from Brazil*, 8 *Surui from Brazil*, 21 *Maya from Mexico*, 14 *Pima from Mexico*, y 7 *Piapoco from Colombia*).

Para evaluar la capacidad de detección de mezclas del panel, se prepararon 4 mezclas artificiales de ratio 1:3 a partir de muestras de ADN diluidas a 1 ng/µL. Las muestras de ADN se recogieron de donantes voluntarios bajo consentimiento informado utilizando hisopos bucales. El ADN fue extraído mediante el kit QIAamp DNA Micro (Qiagen) y cuantificadas posteriormente con el kit Quantifiler® Duo DNA Quantification (AB), siguiendo en ambos casos las recomendaciones del fabricante.

5.2.1.2 Selección de ID-SNPs multialélicos y diseño del panel SNaPshot

A partir de una recopilación de SNPs multialélicos recogidos de la base de datos del Proyecto 1000 Genomas Fase III (The Genomes Project Consortium 2015) se seleccionó un conjunto de SNPs candidatos atendiendo a los siguientes criterios: (i) altos niveles de heterocigosidad en los grupos poblacionales de AFR, EUR y EAS (priorizando EUR); (ii) separación mínima de 1 Mb entre marcadores sinténicos para asegurar su independencia; y (iii) secuencias contexto adecuadas para su análisis mediante tecnologías de genotipado de SNaPshot y MPS: sin tractos homopoliméricos ni regiones de baja complejidad.

Un total de 29 SNPs se implementaron en un ensayo SNaPshot, que se diseñó y optimizó siguiendo recomendaciones previamente publicadas (Sánchez y Endicott 2006).

Las reacciones de PCR, ajustadas a un volumen final de 10 µL, constan de:

- 1 µL de Buffer II (100 mM Tris-HCl; pH 8,3; 500 mM KCl)
- 1,8 µL de $MgCl_2$ a 25 mM
- 0,1 µL AmpliTaq Gold® DNA Polymerase (a 5 U/µL)
- 0,4 µL GeneAmp® 10 mM dNTP Mix with dTTP (Applied Biosystems, AB)
- 1 µL de seroalbúmina bovina a 3.2 mg/ml
- 1,5 µL de *mix* de *primers*
- 1 ng de ADN

En la Tabla 32 se recogen los diseños de los *primers* de PCR y la concentración de los mismos en el *mix* de *primers*. Las reacciones se llevaron a cabo en un termociclador GeneAmp® PCR System 9700 o 2700 (AB) bajo las siguientes condiciones: 10 min a 95°C; 35 ciclos de 30 s a 95°C, 40 s a 62°C y 1 min a 72°C; con una extensión final de 20 min a 72°C.

Tabla 32. Diseño de *primers* para los 29 SNPs recogidos en el ensayo.
CI: código interno. Tam.: Tamaño del amplicón. Conc.: concentración.

CI	SNP	*Primer Forward*	*Primer Reverse*	Tam. (pb)	Conc. (µM)
Tri1	rs10063649	CCTTATTTCTTCCAGGAGTTTTGTAGT	AAAGCAAAACAAACAAAACAGCA	110	5,00
Tri2	rs10023685	TCTTCTCTTCCTCCTCATCTCCT	GGCTTAGTGATAAAATAGTGCTTGG	99	4,00
Tri3	rs6512916	TGGGTTGTGTTTTGGGTTCCTA	CAACGTCTCATTCCACTATGTGT	81	4,00
Tri4	rs2894257	TCCAATCTCCACATAGTAGCTAGA	TTTATTTCTCAAGCCGGCCGAT	120	2,30
Tri5	rs470767	ACTGAAAATAGAGCAATCCAAAGAC	TGATGCTTATCTGGTGAATTTGG	108	1,00
Tri6	rs1903613	AGCAAGGACACAAATTGGTAGAAAA	TTTCCCAGTCACACAGGCCATC	89	1,00
Tri7	rs1241304	ACATTTCCTCGTTACTGGTTACCA	TGCTTTGAAGAGGGACACACTT	81	0,50
Tri8	rs2328264	CTGCATGGCGGGTCAGGAT	ACTCTCACCAAAGTCATGCCTATAA	77	1,50
Tri9	rs2407301	CTCCCTGCTCCAGTTGTCC	GGTTTTATTCATTTTAGGGAGACATGG	96	0,35
Tri10	rs23595	CCAGACATTCCCATCCAGAGAA	CTTTCCTCCCATCTCCTCAGGA	120	0,60
Tri11	rs2780786	GGCCTCCTGTTAACTCACATAA	GTCTAGCACTGTCCTTGGCAC	94	0,40
Tri13	rs1931712	GGAAAGTCTGTGGGTTAAAGCT	AGCATTAGAAATAAATAGCCAACAGA	120	1,50
Tri15	rs7689445	ACAAAGGCTATGGAGAGAAGGG	TCCATTTCTAGTTCTTTGCAAAATCT	105	2,00
Tri16	rs2052215	ATCTGGCCATTTGATTATTTGCCT	AATTTACAAACCTGGGGAGGGG	65	0,40
Tri17	rs1078462	CATGAGACCCCTGGAGCCG	GCCTCAGTCTACTAAAGTGCT	100	0,27
Tri18	rs6500733	CTGGAGGATTCAGCACATTCT	GCGACGCAGAATTAGTAACAGG	120	1,50
Tri19	rs2063200	CATTGTACTTGCTGAATGTATCTGA	ATTCTCATCTGATATAAACCTGGGT	76	1,20
Tri20	rs7662015	GATAAACCTCACCCTGGGAAAAC	GCATTGAACAGTTCTGCCAGTG	108	0,35
Tri21	rs2189958	GGAAAGCCAAACTACCAACAAG	AACTAGTTTCAGTGCACTCAGT	114	5,00
Tri22	rs2249926	TGAAAGACAAAAGGGGAGGAAA	ACGTTTTCAGTTCACCTACAAACA	100	1,20
Tri23	rs4301041	AGCCATGTAGACTGCTTTAAAATT	ATCCACTTTTAAAATGAGAACCCC	120	3,80
Tri24	rs6592125	CTGTCCTGGCAGTCTGTCA	AGATTTGCCTCTGTTTCCCTCA	81	0,80
Tri25	rs10129337	AATCCCTTGGCATCTAAGACCT	GAACTGTGGCCACTCGACTTC	87	0,30
Tri26	rs943444	GTTCTGCTACCCTTTCTCGGAT	CTAAGGGGAATCTCTCTGGGGT	88	3,50
Tri27	rs11381500	TAGTGAGTTCGGGTTTGCTTCT	CCTCTGTCACTTAGCAGGACTT	74	0,40
Tri28	rs6940924	GCTCTGTTGCGCACTCTTG	GTGAAAGCCTTTGTGCTGTCTC	65	0,35
Tri29	rs10415586	ATGGCTGTTTGTCCCTGTACAA	CTCTCACCTCCAGAAACTACCC	62	0,17
Tri30	rs6758274	TGACAAGGGCTATAAGGGAGAT	CTCATCTCCACCACCCCAAC	104	1,70
Tri31	rs3859194	TGCAATACTCGATCCTGCTGTG	CTAGAAGTACTGGAGTTGAGCCA	100	0,23

Para la purificación post-PCR se combinaron 2,5 µL de producto de PCR con 1 µL de 1:3 Illustra™ ExoStar™ 1-Step (GE Healthcare). Se incubó a 37°C durante 45 min y seguidamente se inactivó el enzima a 85°C durante 15 min.

La reacción de SBE, con un volumen final de 3 µL, consta de:

- 1,25 µL de SNaPshot® Multiplex Ready Reaction Mix (1:2)
- 0,75 µL de *mix* de sondas de SBE
- 1 µL de producto purificado de PCR

En la Tabla 33, se recogen los diseños de sondas para SBE y su concentración en el *mix* de sondas. Las condiciones de la reacción de SBE comprenden 33 ciclos de 10 s a 96°C, 5 s a 56°C y 30 s a 60°C.

Tabla 33. Diseño de sondas para los 29 SNPs incluidos en el ensayo. Las bases en minúscula corresponden a secuencias utilizadas como modificadores de la movilidad. Los SNPs destacados en gris constituyen uno de los dos ensayos SNaPshot que se describen en la sección 5.2.2.1. CI: Código interno. Orient.: orientación de la sonda con respecto a los datos del Proyecto 1000 Genomas Fase III. Conc.: concentración. Long.: longitud de la sonda.

CI	SNP	Orient.	Conc. (µM)	Long. (pb)	Sonda SBE
Tri1	rs10063649	*Forward*	9,00	58	aaactaggtgccacgtcgtgaaagtctgacaaTGTAGTTTATAATTTTAGATTTAGG
Tri2	rs10023685	*Reverse*	4,00	38	tgaaagtctgacaaTAGTGATAAAATAGTGCTTGGAA
Tri3	rs6512916	*Forward*	3,00	40	cgtgaaagtctgacaaCTATCATTCAAAATTTCTGTACC
Tri4	rs2894257	*Forward*	1,00	42	gtgccacgtcgtgaaagtctgacaaATCGGGGGACCTGCCC
Tri5	rs470767	*Reverse*	5,75	44	cacgtcgtgaaagtctgacaaTATAAGTTACTGACAAACTCCC
Tri6	rs1903613	*Forward*	0,50	46	gtgccacgtcgtgaaagtctgacaaGGCCATCATATTGGACAGCA
Tri7	rs1241304	*Forward*	5,25	48	ggtgccacgtcgtgaaagtctgacaaTTACTGGTTACCATACTCATC
Tri8	rs2328264	*Reverse*	1,20	50	aactaggtgccacgtcgtgaaagtctgacaaACATGGGACATCAAGCAACA
Tri9	rs2407301	*Reverse*	0,25	52	aactaggtgccacgtcgtgaaagtctgacaaACATGGGACATCAAGCAACA
Tri10	rs23595	*Reverse*	0,70	54	taaactaggtgccacgtcgtgaaagtctgacaaAGACATTCCCATCCAGAGAA
Tri11	rs2780786	*Forward*	2,00	56	taaactaggtgccacgtcgtgaaagtctgacaaTAATCTTTAGCTCACTTCCTAG
Tri13	rs1931712	*Reverse*	1,30	60	tgactaaactaggtgccacgtcgtgaaagtctgacaaAAAAGTCAGGTACAAGAGTAGA
Tri15	rs7689445	*Reverse*	0,50	64	actgactaaactaggtgccacgtcgtgaaagtctgacaaCTAGTTCTTTGCAAAATCTTCATA
Tri16	rs2052215	*Forward*	0,50	66	caaaactgactaaactaggtgccacgtcgtgaaagtctgacaaTATTTGCCTCAGAGAAAATGAC
Tri17	rs1078462	*Reverse*	0,90	68	tctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaGGACTGGCAGAATCAGAAC
Tri18	rs6500733	*Forward*	1,20	70	agtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaAGGATTCAGCACATTCTGC
Tri19	rs2063200	*Reverse*	1,00	72	gtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaGTACTTGCTGAATGTATCTGAT
Tri20	rs7662015	*Forward*	1,30	74	gaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaGAAAACCACCTTCCTGATCA
Tri21	rs2189958	*Forward*	3,00	76	agtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaCAGTATATATATTACTTAAAGACAC
Tri22	rs2249926	*Reverse*	1,00	78	gtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaGGAAATAAGATGTACAACTGCA
Tri23	rs4301041	*Forward*	2,30	80	gtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaATAGACAAATGGAAACAGAAGTTA
Tri24	rs6592125	*Reverse*	1,00	82	cgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaATAAGAGACTTATGAGTCAACA
Tri25	rs10129337	*Forward*	1,30	84	cacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaATCAAAAGAGTAGCTATCATCC
Tri26	rs943444	*Forward*	2,40	86	tgccacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaGACTTTACTATATGAAGGCTC
Tri27	rs11381500	*Reverse*	2,00	88	aggtgccacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaCTGTCACTTAGCAGGACTTG
Tri28	rs6940924	*Forward*	2,20	90	actaggtgccacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaCTCTTCTACTTCCCACCAT
Tri29	rs10415586	*Forward*	1,80	92	aactaggtgccacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCTGTACAAGTCAAGCTAAC
Tri30	rs6758274	*Forward*	2,00	94	taaactaggtgccacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaGTAGAAGAGGAAATCCTTCT
Tri31	rs3859194	*Reverse*	1,50	96	gactaaactaggtgccacgtcgtgaaagtctgacaaaactgactaaactaggtgccacgtcgtgaaagtctgacaaCTCTTGGCTTCCAGTCACA

El producto total de la reacción de SBE se combinó con 1 µL de 1:2 Illustra™ Shrimp Alkaline Phosphatase (GE Healthcare) para la purificación, se incubó a 37°C durante 80 min y se inactivó la enzima a 85°C durante 15 min.

Los productos de SBE purificados se prepararon para CE añadiendo 1 µL del producto a 9,5 µL de Hi-Di™ Formamide (AB) y 0,25 µL de GeneScan™ 120 LIZ® Size Standard (AB). La electroforesis se llevó a cabo en un secuenciador ABI Prism 3130*xl* Genetic Analyser, con capilares de 36 cm y polímero POP-4™ bajo condiciones estándar. Los electroferogramas se visualizaron usando AB GeneMapper® ID Software v. 3.2.1.

5.2.2 **Resultados**

5.2.2.1 Optimización del ensayo SNaPshot

Inicialmente, se diseñó una única reacción de SBE que comprendía todos los marcadores del panel. A pesar de que las guías recomiendan una separación mínima entre las diferentes longitudes de sondas de 3-4 pb para una adecuada separación electroforética de los alelos de los SNPs (Sánchez y Endicott 2006), la separación se redujo a 2-3 pb para intentar abarcar los 29 SNPs multialélicos del panel (que requieren el uso de al menos tres de los colores) en un intervalo entre 30-90 pb. El límite inferior de este intervalo se eligió para poder usar indistintamente polímero POP-4™ o POP-7™, dado que el último presenta limitaciones a la hora de separar fragmentos <30-35 pb. El límite superior pretende evitar un aumento desproporcionado del coste de los oligonucleótidos.

En la Tabla 34 se muestran las movilidades finales de los posibles alelos de los SNPs incluidos en el panel. Dado que ciertos alelos presentaron movilidades ligeramente desviadas de lo esperado, sus señales se colapsaron de manera que no fue posible individualizarlas. Por ello, el ensayo SNaPshot fue dividido en dos reacciones –Auto 1 y Auto 2– que contienen cada una un número similar de marcadores –ver Tabla 33– separados entre 4-6 pb. No obstante, se debe tener en cuenta que la PCR inicial se mantiene como una única reacción que engloba los 29 SNPs, permitiendo que el gasto de muestra de ADN permanezca invariable: el producto de PCR es más que suficiente para realizar las dos reacciones SBE.

En la Fig. 79 se muestra un perfil de 9947A realizado con 1 ng de ADN inicial. A pesar de que no es posible alcanzar un balance inter-*loci* perfecto, se aseguró que al menos uno de los alelos de cada SNP presentara una altura >1000 RFUs en condiciones óptimas. El desbalance intra-*loci* es una consecuencia inherente a la tecnología de genotipado: SNaPshot reporta las diferentes bases con diferentes fluorocromos que presentan diferentes intensidades de emisión de manera que los alelos G reportados con dR110 son más altos que los alelos A reportados con dR6G y éstos a su vez más altos que los alelos C y T reportados con dTAMRA™ y dROX™, respectivamente. No obstante, trabajos previos (Sánchez *et al.* 2006, Phillips *et al.* 2007, Freire-Aradas *et al.* 2012, Fondevila *et al.* 2013) indican que existe una correlación entre cada posible genotipo heterocigoto y el desbalance entre ambos alelos medido como ratio de altura de los picos –PHR: *peak height ratio*–.

Tabla 34. Movilidad observada de los posibles alelos de los SNPs incluidos en el ensayo. CI: código interno. Long.: longitud.

CI	SNP	Long. sonda	Long. media observada				Desviación estándar			
			A	C	G	T	A	C	G	T
Tri1	rs10063649	59	60,58	59,62		60,82	0,03	0,00		0,04
Tri2	rs10023685	39		40,24	40,09	41,60		0,04	0,02	0,05
Tri3	rs6512916	41	40,80		39,96	41,66	0,03		0,02	0,09
Tri4	rs2894257	43	44,67	42,72	43,04	44,99	0,07	0,04	0,06	0,65
Tri5	rs470767	45	44,80	44,26	43,69		0,00	0,07	0,08	
Tri6	rs1903613	47	47,90	47,98		49,01	0,03	0,03		0,04
Tri7	rs1241304	49	50,20	49,89	49,21		0,00	0,07	0,06	
Tri8	rs2328264	51	52,83	51,91		52,78	0,04	0,02		0,03
Tri9	rs2407301	53	52,81		52,19	53,48	0,04		0,04	0,04
Tri10	rs23595	55	55,33	54,77	54,51		0,03	0,02	0,03	
Tri11	rs2780786	57	58,16	57,55		58,17	0,03	0,03		0,03
Tri13	rs1931712	61	61,13		60,30	61,47	0,03		0,02	0,03
Tri15	rs7689445	65	64,52			65,38	0,04			0,03
Tri16	rs2052215	67	66,35	66,00	65,58		0,02	0,02	0,03	
Tri17	rs1078462	69	68,81		67,56	69,14	0,02		0,01	0,03
Tri18	rs6500733	71	70,44	70,10	69,41		0,04	0,04	0,03	
Tri19	rs2063200	73	73,85		72,78	74,27	0,00		0,04	0,01
Tri20	rs7662015	75	73,97	73,71		74,60	0,03	0,01		0,03
Tri21	rs2189958	77	75,80	75,75	75,03		0,06	0,00	0,08	
Tri22	rs2249926	79	78,10		77,52	78,95	0,06		0,03	0,03
Tri23	rs4301041	81	80,27	80,18		80,49	0,07	0,08		0,03
Tri24	rs6592125	83		81,22	81,38	82,64		0,00	0,06	0,04
Tri25	rs10129337	85	83,81		82,97	84,18	0,00		0,00	0,00
Tri26	rs943444	87	86,44	86,00	85,69		0,04	0,03	0,00	
Tri27	rs17136244	89	89,05	88,00		88,51	0,05	0,00		0,03
Tri28	rs6940924	91	90,26	89,79	89,16		0,04	0,00	0,04	
Tri29	rs10415586	93	92,01	91,77	91,32		0,04	0,03	0,02	
Tri30	rs6758274	95	94,41	94,21	93,61		0,03	0,04	0,04	
Tri31	rs3859194	97	96,14	95,91	95,70		0,04	0,03	0,02	

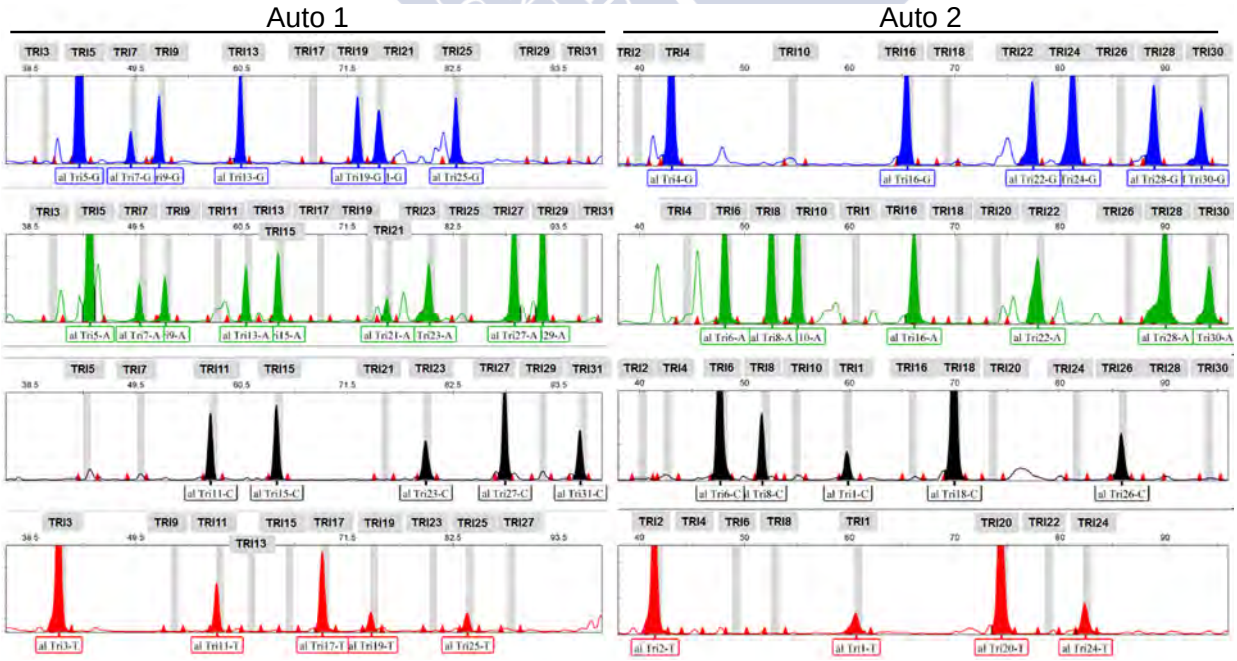


Fig. 79. Electroferogramas –Auto 1 y Auto 2– obtenidos a partir de 1 ng de ADN inicial del control de ADN 9947A.

En la Fig. 80 se recogen (en negro) los valores de PHR que exhiben los diferentes genotipos heterocigotos de los SNPs multialélicos, obtenidos a partir de los análisis de muestras individuales OCE y AMR del panel HGDP-CEPH. La distribución de los patrones PHR de los SNPs multialélicos se corresponde adecuadamente con la que presentan los SNPs bialélicos en otros trabajos (Sánchez *et al.* 2006, Phillips *et al.* 2007, Freire-Aradas *et al.* 2012, Fondevila *et al.* 2013). Así, los heterocigotos CT presentan los valores más estables y balanceados, con valores de PHR en torno a 1/1, con una baja dispersión de los datos. Los heterocigotos AC, AT y GA presentan valores promedio de PHR de 2/1 con una dispersión más acusada que los CT. La máxima dispersión de datos se observa en los heterocigotos GC y, más acusada, GT. Teniendo en cuenta este desbalance, se pueden inferir genotipos correctamente para muestras individuales.

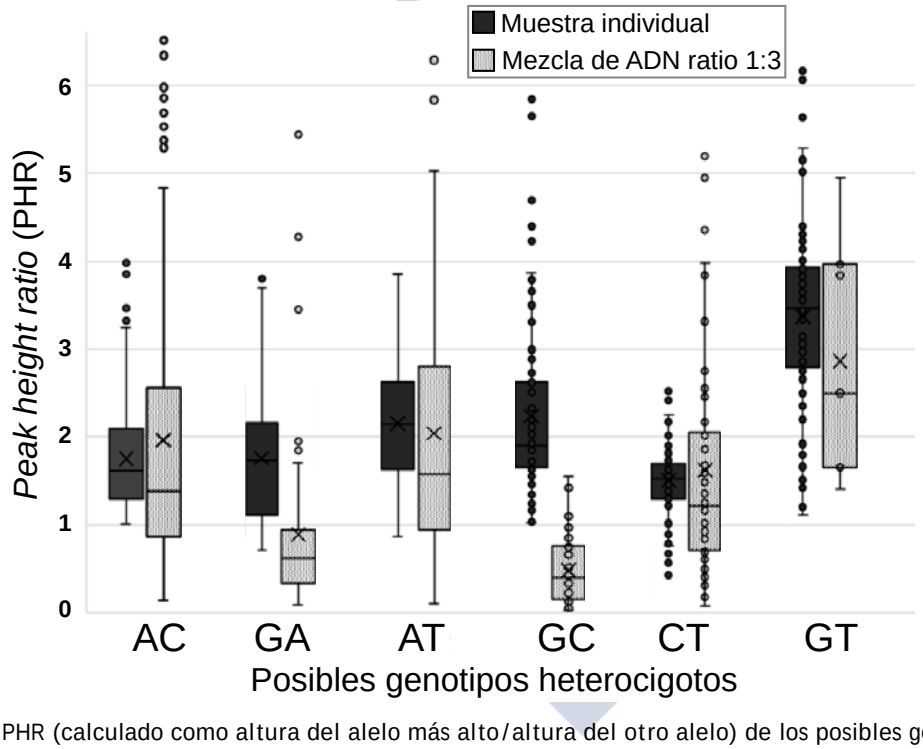


Fig. 80. PHR (calculado como altura del alelo más alto/altura del otro alelo) de los posibles genotipos heterocigotos de los SNPs multialélicos incluidos en el ensayo. Se representan los valores observados para las 92 muestras individuales del panel HGDP-CEPH frente a los de las mezclas de ADN de ratio 1:3.

A pesar del cuidadoso diseño *in-silico* de la PCR no se pudieron evitar ciertas interacciones entre los *primers*, que produjeron una serie señales electroforéticas inespecíficas. Todas estas señales se producen con una movilidad estable y fuera de las posiciones alélicas de los SNPs incluidos en el ensayo –ver Tabla 35–, por lo que pueden ser fácilmente descontadas del perfil. Una de las posibles causas de las interacciones es un exceso de los *primers* de algún SNP en la PCR. Se deben realizar más pruebas para identificar los *primers* implicados; no obstante, si se realizara una PCR por separado de los SNPs de Auto 1 y Auto 2 –duplicando el gasto de muestra inicial– no aparecen picos inespecíficos.

Tabla 35. Características de movilidad de las señales inespecíficas y del alelo del SNP más próximo.
CI: código interno. Long.: longitud.

SNP más próximo			Señales inespecíficas					
			Long. media observada			Desviación estándar		
CI	SNP	Long. Alelo	A	C	G	A	C	G
Tri3	rs6512916	40,80	41,97			0,07		
Tri2	rs10023685	40,09			41,58			0,06
Tri4	rs2894257	44,67	45,84			0,07		
Tri5	rs470767	44,80	45,84			0,07		
Tri6	rs1903613	47,98		45,89			0,05	

5.2.2.2 Características de los SNPs seleccionados y del panel

Las estimaciones de frecuencias alélicas de los 29 SNPs en las 5 poblaciones continentales (AFR, EUR, EAS, OCE y AMR) se listan en la Tabla 36 y se representan en la Fig. 81. Considerando una frecuencia del alelo minoritario –MAF: *minor allele frequency*– de 0,05 todos los SNPs del panel presentan un carácter multialélico en EUR; disminuyendo al ~90% si se consideran los 3 grupos poblaciones (AFR, EUR y EAS) de los que se disponía de datos de frecuencias cuando se realizó la selección y al ~50% si se incluyen los 5 grupos poblacionales. Se debe tener en cuenta que el pequeño tamaño muestral de OCE y AMR puede haber impedido la detección de alelos a baja frecuencia en dichas poblaciones.

Tabla 36. Estimaciones de las frecuencias alélicas de los SNPs multialélicos incluidos en el ensayo para 5 grupos poblacionales. Las frecuencias de AFR, EUR y EAS se recopilaron a través de los datos del Proyecto 1000 Genomas, excluyendo las poblaciones *admixed* ACB y ASW. Las frecuencias de OCE Y AMR se calcularon a partir de los datos obtenidos del análisis del panel HGDP-CEPH.
CI: código interno. Ref.: alelo de referencia. Al.1,2,3: resto de alelos del SNP en orden alfabético.

CI	SNP	Ref./Al. 1, 2, 3	AFR				EUR				EAS				OCE				AMR			
			Ref.	Al.1	Al.2	Al.3	Ref.	Al.1	Al.2	Al.3	Ref.	Al.1	Al.2	Al.3	Ref.	Al.1	Al.2	Al.3	Ref.	Al.1	Al.2	Al.3
Tri1	rs10063649	A/C/T	0,34	0,36	0,30	0,00	0,40	0,30	0,30	0,00	0,15	0,20	0,65	0,00	0,20	0,30	0,50	0,00	0,25	0,07	0,68	0,00
Tri2	rs10023685	C/A/G	0,19	0,55	0,25	0,00	0,10	0,48	0,42	0,00	0,08	0,44	0,48	0,00	0,04	0,67	0,29	0,00	0,07	0,66	0,27	0,00
Tri3	rs6512916	G/A/T	0,36	0,26	0,37	0,00	0,41	0,21	0,38	0,00	0,41	0,12	0,46	0,00	0,17	0,06	0,77	0,00	0,19	0,08	0,73	0,00
Tri4	rs2894257	G/A/C/T	0,37	0,05	0,19	0,40	0,30	0,13	0,32	0,25	0,39	0,04	0,31	0,26	0,66	0,22	0,12	0,00	0,15	0,51	0,33	0,02
Tri5	rs470767	C/G/T	0,66	0,18	0,16	0,00	0,56	0,31	0,13	0,00	0,60	0,18	0,22	0,00	0,71	0,02	0,27	0,00	0,46	0,48	0,06	0,00
Tri6	rs1903613	A/C/T	0,53	0,26	0,21	0,00	0,46	0,33	0,20	0,00	0,54	0,33	0,13	0,00	0,06	0,42	0,52	0,00	0,52	0,40	0,07	0,00
Tri7	rs1241304	G/A/C	0,39	0,14	0,47	0,00	0,34	0,20	0,46	0,00	0,22	0,38	0,40	0,00	0,06	0,46	0,48	0,00	0,33	0,03	0,64	0,00
Tri8	rs2328264	T/A/G	0,48	0,24	0,29	0,00	0,31	0,23	0,46	0,00	0,33	0,10	0,57	0,00	0,48	0,31	0,21	0,00	0,63	0,11	0,25	0,00
Tri9	rs2407301	C/A/T	0,52	0,26	0,22	0,00	0,34	0,42	0,24	0,00	0,20	0,38	0,42	0,00	0,21	0,33	0,46	0,00	0,26	0,34	0,40	0,00
Tri10	rs23595	T/C/G	0,12	0,57	0,30	0,00	0,52	0,23	0,24	0,00	0,50	0,24	0,27	0,00	0,13	0,58	0,27	0,00	0,57	0,22	0,19	0,00
Tri11	rs2780786	C/A/T	0,41	0,46	0,13	0,00	0,42	0,27	0,32	0,00	0,16	0,53	0,30	0,00	0,17	0,54	0,29	0,00	0,25	0,23	0,52	0,00
Tri13	rs1931712	C/A/T	0,34	0,42	0,23	0,00	0,27	0,25	0,48	0,00	0,11	0,46	0,43	0,00	0,40	0,54	0,06	0,00	0,00	0,17	0,83	0,00
Tri15	rs7689445	T/A/G	0,43	0,37	0,20	0,00	0,34	0,39	0,27	0,00	0,30	0,53	0,17	0,00	0,38	0,31	0,31	0,00	0,34	0,65	0,01	0,00
Tri16	rs2052215	C/A/G	0,17	0,46	0,36	0,00	0,29	0,47	0,25	0,00	0,35	0,45	0,20	0,00	0,40	0,19	0,40	0,00	0,16	0,60	0,25	0,00
Tri17	rs1078462	C/A/T	0,21	0,60	0,20	0,00	0,32	0,42	0,26	0,00	0,35	0,37	0,29	0,00	0,24	0,32	0,44	0,00	0,02	0,51	0,47	0,00
Tri18	rs6500733	G/A/C	0,28	0,27	0,45	0,00	0,54	0,29	0,17	0,00	0,48	0,11	0,41	0,00	0,24	0,02	0,74	0,00	0,26	0,32	0,42	0,00
Tri19	rs2063200	A/C/T	0,01	0,98	0,01	0,00	0,20	0,71	0,08	0,00	0,21	0,79	0,00	0,00	0,04	0,96	0,00	0,00	0,06	0,93	0,01	0,00
Tri20	rs7662015	C/A/T	0,53	0,32	0,15	0,00	0,24	0,34	0,42	0,00	0,36	0,20	0,44	0,00	0,50	0,12	0,38	0,00	0,19	0,18	0,63	0,00
Tri21	rs2189958	A/C/G	0,42	0,35	0,23	0,00	0,26	0,05	0,69	0,00	0,24	0,16	0,60	0,00	0,00	0,39	0,61	0,00	0,00	0,02	0,85	0,13
Tri22	rs2249926	T/A/C	0,13	0,13	0,75	0,00	0,42	0,11	0,47	0,00	0,34	0,37	0,29	0,00	0,65	0,06	0,29	0,00	0,21	0,13	0,67	0,00
Tri23	rs4301041	A/C/T	0,09	0,54	0,38	0,00	0,31	0,24	0,44	0,00	0,02	0,00	0,98	0,00	0,02	0,37	0,62	0,00	0,27	0,22	0,51	0,00
Tri24	rs6592125	A/C/G	0,32	0,19	0,49	0,00	0,39	0,34	0,26	0,00	0,43	0,44	0,12	0,00	0,21	0,79	0,00	0,00	0,24	0,76	0,00	0,00
Tri25	rs10129337	A/G/T	0,19	0,35	0,46	0,00	0,26	0,29	0,45	0,00	0,16	0,51	0,33	0,00	0,19	0,40	0,40	0,00	0,03	0,46	0,51	0,00
Tri26	rs943444	A/C/G	0,29	0,10	0,61	0,00	0,26	0,35	0,39	0,00	0,39	0,41	0,20	0,00	0,33	0,54	0,13	0,00	0,33	0,58	0,09	0,00
Tri27	rs17136244	A/G/T	0,50	0,09	0,42	0,00	0,49	0,07	0,44	0,00	0,28	0,20	0,52	0,00	0,62	0,04	0,35	0,00	0,15	0,05	0,80	0,00
Tri28	rs6940924	A/C/G	0,14	0,02	0,85	0,00	0,41	0,08	0,50	0,00	0,11	0,00	0,89	0,00	0,00	0,00	1,00	0,00	0,21	0,00	0,79	0,00
Tri29	rs10415586	A/C/G	0,25	0,25	0,50	0,00	0,22	0,51	0,26	0,00	0,28	0,52	0,21	0,00	0,58	0,35	0,08	0,00	0,42	0,42	0,17	0,00
Tri30	rs6758274	G/A/C	0,21	0,36	0,43	0,00	0,34	0,38	0,28	0,00	0,37	0,34	0,29	0,00	0,19	0,67	0,13	0,00	0,21	0,46	0,33	0,00
Tri31	rs3859194	C/G/T	0,23	0,31	0,46	0,00	0,24	0,47	0,29	0,00	0,14	0,48	0,39	0,00	0,10	0,27	0,63	0,00	0,03	0,68	0,28	0,00

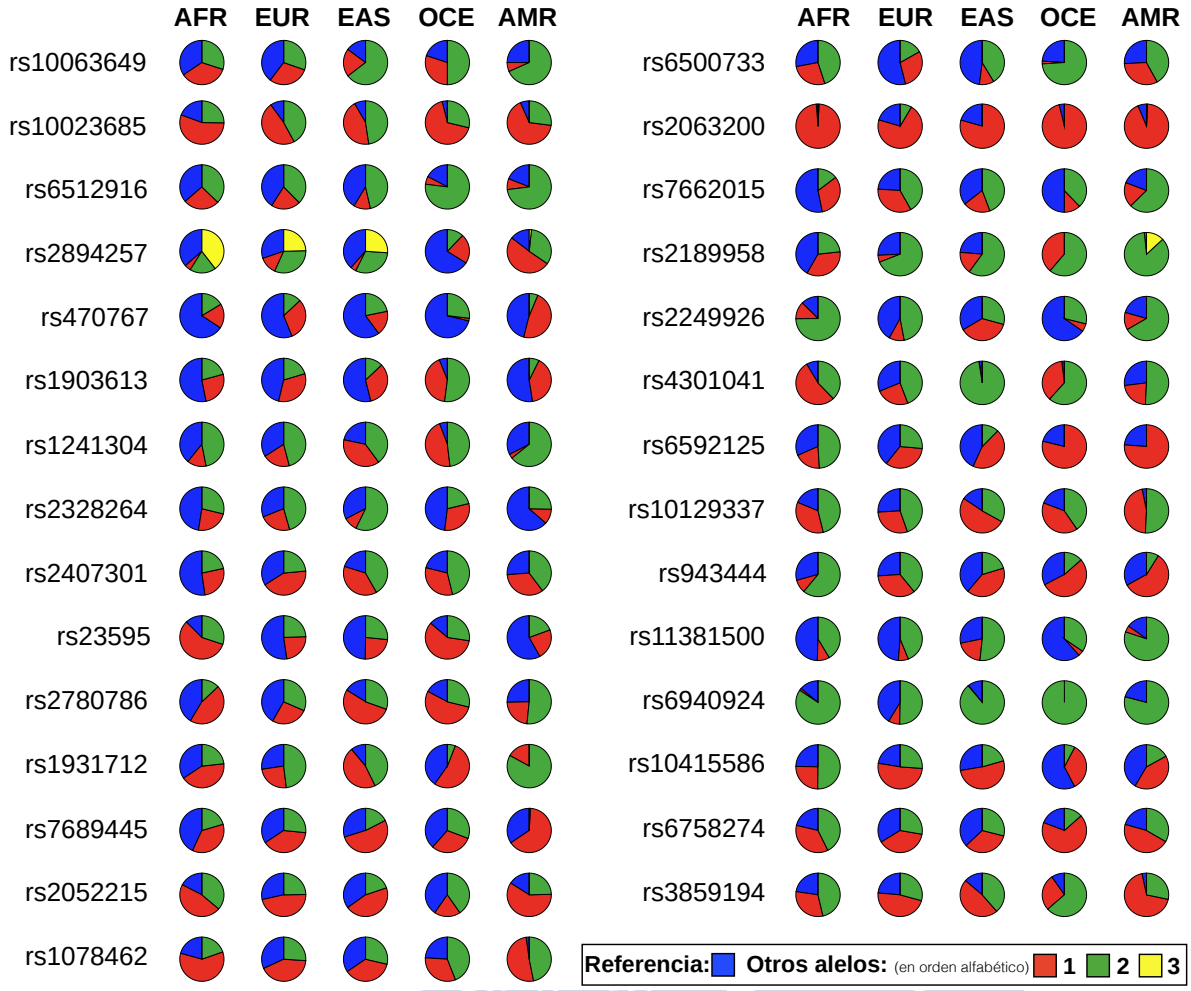


Fig. 81. Representación gráfica de las frecuencias alélicas recogidas en la Tabla 36.

Los SNPs se encuentran adecuadamente distribuidos por el genoma, con más de 1 Mb de distancia entre marcadores sinténicos, de manera que pueden ser utilizados como marcadores independientes –ver Fig. 82–.

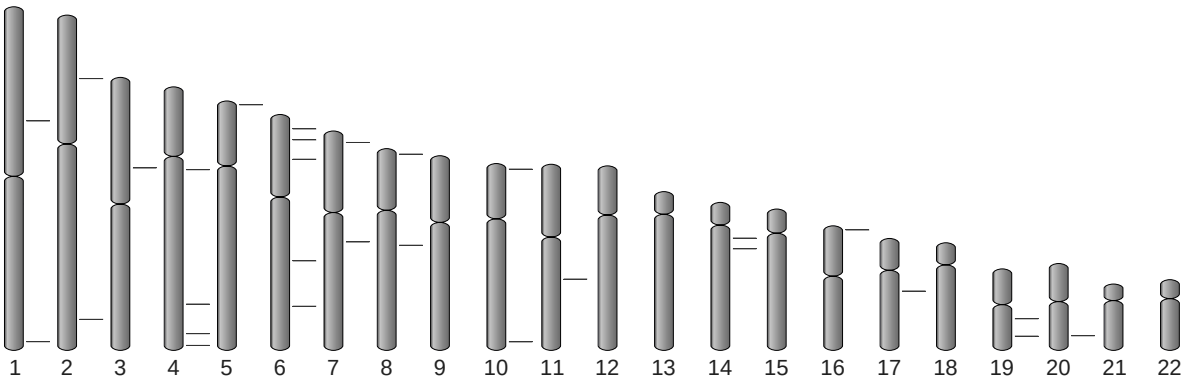


Fig. 82. Ideograma en el que se representan los 22 cromosomas autosómicos y las posiciones de los SNPs multialélicos incluidos en el ensayo.

La máxima heterocigosidad teórica de los marcadores bialélicos es de 0,50; valor que se eleva a 0,67 en marcadores trialélicos y 0,75 en tetraalélicos. Por ello, los SNPs seleccionados muestran valores de heterocigosidad muy altos en todas las poblaciones, cercanos o superiores a 0,50; tal y como se recoge en la Tabla 37. Como excepción, los SNPs rs2063200, rs4301041 y rs6940924 presentan valores de heterocigosidad considerablemente bajos en algunas de las poblaciones, ya que fueron seleccionados en base a la alta heterocigosidad que presentan en EUR.

Tabla 37. Valores de heterocigosidad que presentan los SNPs incluidos en el ensayo en las 5 poblaciones definidas continentalmente. Los SNPs destacados en gris presentan buenos valores en EUR pero considerablemente bajos en otras. CI: código interno. Cr.: cromosoma. Posición: en base al genoma de referencia GRCh37. Ref.: alelo de referencia. Al. 1, 2, 3: resto de alelos del SNP. Anc.: alelo ancestral.

CI	SNP	Cr.	Posición	Ref./ Al. 1, 2 ,3	Anc.	Heterocigosidad				
						AFR	EUR	EAS	OCE	AMR
Tri1	rs10063649	5	2790659	A/C/T	C	0,665	0,661	0,519	0,620	0,466
Tri2	rs10023685	4	157534249	C/A/G	G	0,592	0,583	0,572	0,462	0,482
Tri3	rs6512916	20	52421497	G/A/T	T	0,659	0,645	0,598	0,375	0,424
Tri4	rs2894257	6	32433276	G/A/C/T	C	0,670	0,728	0,684	0,502	0,611
Tri5	rs470767	6	10461089	C/G/T	C	0,509	0,573	0,557	0,421	0,557
Tri6	rs1903613	8	70073449	A/C/T	C	0,607	0,633	0,583	0,550	0,557
Tri7	rs1241304	14	25869209	G/A/C	C	0,607	0,633	0,646	0,553	0,480
Tri8	rs2328264	6	18409003	T/A/G	T	0,635	0,641	0,557	0,629	0,520
Tri9	rs2407301	8	3982953	C/A/T	C	0,612	0,649	0,640	0,635	0,657
Tri10	rs23595	19	36204831	T/C/G	G	0,564	0,613	0,626	0,577	0,588
Tri11	rs2780786	1	242964135	C/A/T	A	0,604	0,655	0,596	0,597	0,616
Tri13	rs1931712	10	129356075	C/A/T	T	0,648	0,634	0,591	0,544	0,281
Tri15	rs7689445	4	60104910	T/A/G	–	0,640	0,659	0,603	0,663	0,462
Tri16	rs2052215	7	8327961	C/A/G	C	0,623	0,639	0,634	0,637	0,560
Tri17	rs1078462	2	46064705	C/A/T	A	0,563	0,654	0,663	0,646	0,522
Tri18	rs6500733	16	2993288	G/A/C	G	0,646	0,595	0,588	0,394	0,653
Tri19	rs2063200	6	106188553	A/C/T	C	0,039	0,446	0,328	0,074	0,134
Tri20	rs7662015	4	179037990	C/A/T	C	0,592	0,651	0,636	0,591	0,539
Tri21	rs2189958	7	80504737	A/C/G	C	0,649	0,455	0,558	0,475	0,252
Tri22	rs2249926	4	187534542	T/A/C	C	0,409	0,591	0,664	0,486	0,497
Tri23	rs4301041	3	65720967	A/C/T	C	0,562	0,646	0,048	0,487	0,620
Tri24	rs6592125	11	83261542	A/C/G	A	0,622	0,658	0,600	0,334	0,363
Tri25	rs10129337	14	33698620	A/G/T	T	0,629	0,647	0,603	0,637	0,529
Tri26	rs943444	1	82843292	A/C/G	G	0,533	0,658	0,641	0,585	0,546
Tri27	rs11381500	10	4031769	A/G/T	A	0,574	0,564	0,613	0,500	0,332
Tri28	rs6940924	6	139073363	A/C/G	G	0,266	0,568	0,196	0,000	0,330
Tri29	rs10415586	19	48779290	A/C/G	–	0,624	0,617	0,613	0,541	0,626
Tri30	rs6758274	2	220665364	G/A/C	G	0,643	0,661	0,663	0,492	0,634
Tri31	rs3859194	17	38125796	C/G/T	G	0,637	0,637	0,605	0,516	0,452

En la Fig. 83 se representa la RMP –*random match probaility*– acumulada del ensayo para las 5 poblaciones definidas continentalmente. El panel presenta una informatividad universal a pesar de que no se alcanzan niveles de informatividad equivalentes para todas las poblaciones. El incremento de los valores de RMP se produce con una pendiente prácticamente constante en cada población, subrayando una informatividad similar de los marcadores incluidos y el balance del panel. La pérdida de la información de cualquier

marcador disminuye mínimamente el poder de discriminación del panel, aportando robustez a los análisis de ADN degradado o *low level*.

Sin embargo, los SNPs rs2063200, rs4301041 y rs6940924 representan puntos de inflexión en algunas poblaciones, manteniendo la pendiente en EUR. Como consecuencia, se produce un desbalance del nivel de informatividad del panel en las diferentes poblaciones: los valores de RMP acumulada se elevan hasta $5{,}42 \times 10^{-20}$ en EUR, distanciándose del resto de poblaciones en 2 órdenes de magnitud para AFR ($2{,}52 \times 10^{-18}$) y EAS ($5{,}44 \times 10^{-18}$) y en 5 para OCE ($2{,}48 \times 10^{-15}$) y AMR ($4{,}61 \times 10^{-15}$). Estos valores globales de informatividad son menores que los que se obtienen con los kits de STRs comúnmente utilizados o con los 52 ID-SNPs del panel del consorcio SNP*for*ID (Sánchez *et al.* 2006). No obstante, se acercan a los valores que proporcionan los 13 STRs del CODIS (Butler 2006).

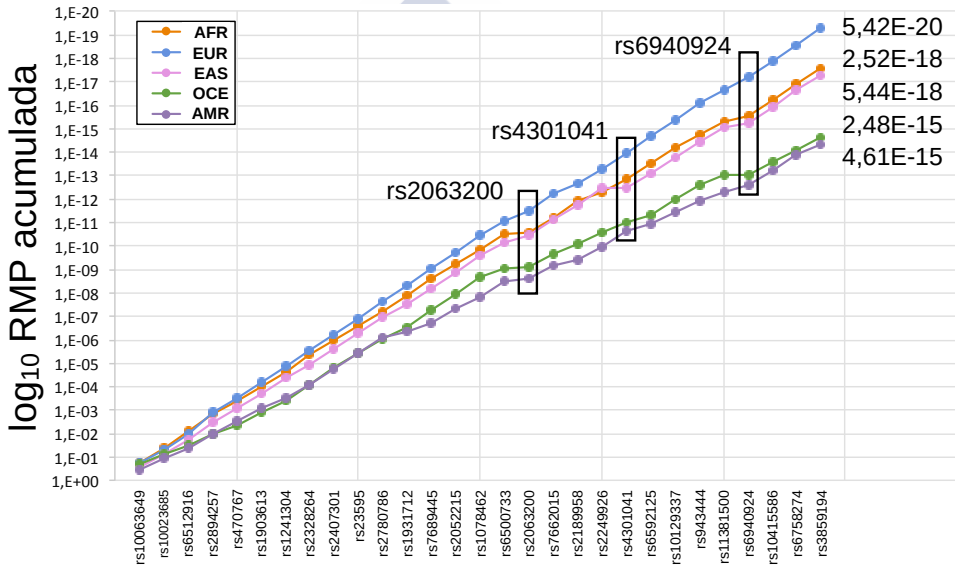


Fig. 83. Representación del aumento de la RMP del panel a medida que se incorporan los marcadores. Se muestran los valores finales para cada población y se destacan los SNPs que presentan valores de heterocigosidad considerablemente bajos en algunas de las poblaciones.

5.2.2.3 Análisis de mezclas de ADN

Durante el análisis exploratorio de las 4 mezclas de ADN de ratio 1:3 se consideraron dos aspectos: la capacidad de detección de mezclas y la capacidad de deconvolución de las mismas.

En primer lugar, todas las mezclas pudieron ser detectadas, ya que se encontraron más de dos alelos en al menos el 30% de los SNPs. En la Tabla 38 se muestra la probabilidad de encontrar más de dos alelos en cada SNP en una mezcla de dos individuos de cada una de las poblaciones, así como la probabilidad de encontrar al menos un SNP que presente más de 2 alelos en el panel. La combinación de los 29 SNPs permite obtener probabilidades >99,8% de

encontrar al menos un SNPs que muestre más de 2 alelos en el perfil en todas las poblaciones, permitiendo identificar la mezcla como tal. La probabilidad acumulada es más alta en EUR, en la que todos los SNPs presentan patrones multialélicos. Los SNPs con probabilidad 0 no presentan patrones multialélicos en dicha población, no obstante, pueden ser informativos si se producen mezclas entre individuos de diferentes poblaciones.

Tabla 38. Considerando una mezclas de 2 individuos de la misma población, se recoge la probabilidad de encontrar más de dos alelos (como indicador de la presencia de una mezcla de ADN) en cada uno de los SNPs y la probabilidad acumulada de encontrar más de dos alelos en al menos uno de los SNPs del perfil.

SNPs	Probabilidad de encontrar >2 alelos en una mezcla de 2 individuos				
	AFR-AFR	EUR-EUR	EAS-EAS	OCE-OCE	AMR-AMR
rs10063649	0,441	0,433	0,237	0,360	0,137
rs10023685	0,324	0,241	0,208	0,090	0,141
rs6512916	0,429	0,396	0,281	0,092	0,132
rs2894257	0,463	0,604	0,496	0,209	0,339
rs470767	0,231	0,270	0,283	0,044	0,167
rs1903613	0,346	0,376	0,277	0,157	0,184
rs1241304	0,306	0,375	0,399	0,154	0,080
rs2328264	0,389	0,396	0,229	0,376	0,215
rs2407301	0,354	0,408	0,385	0,383	0,426
rs23595	0,258	0,357	0,375	0,246	0,286
rs2780786	0,294	0,422	0,315	0,323	0,361
rs1931712	0,406	0,387	0,261	0,151	0,000
rs7689445	0,385	0,428	0,330	0,437	0,022
rs2052215	0,350	0,396	0,375	0,376	0,279
rs1078462	0,291	0,419	0,437	0,406	0,068
rs6500733	0,409	0,318	0,255	0,043	0,419
rs2063200	0,001	0,148	0,000	0,000	0,006
rs7662015	0,300	0,411	0,379	0,274	0,264
rs2189958	0,407	0,111	0,279	0,000	0,022
rs2249926	0,142	0,262	0,438	0,131	0,210
rs4301041	0,208	0,406	0,001	0,052	0,365
rs6592125	0,359	0,427	0,282	0,000	0,000
rs10129337	0,365	0,409	0,319	0,376	0,089
rs943444	0,208	0,426	0,387	0,284	0,204
rs11381500	0,220	0,189	0,351	0,098	0,069
rs6940924	0,022	0,206	0,002	0,000	0,000
rs10415586	0,373	0,362	0,353	0,184	0,351
rs6758274	0,393	0,433	0,438	0,209	0,380
rs3859194	0,390	0,391	0,303	0,197	0,077
Probabilidad de encontrar >2 alelos en al menos un SNP del panel	0,9999867	0,9999982	0,9999778	0,9988110	0,9981195

En segundo lugar, la identificación de los alelos de los componentes mayoritario y minoritario no resultó factible en ninguno de los casos. En la Fig. 80 se incluyen los PHR de los heterocigotos observados en las mezclas de ADN (en gris). Incluso para las mezclas estudiadas de ratio poco extremo (1:3) la mayoría de los heterocigotos observados no pueden ser distinguidos de los que se presentan en muestras individuales, por lo que no es posible identificar los alelos de los componentes minoritario y mayoritario.

5.2.3 **Discusión**

Se presenta un ensayo SNaPshot que comprende 29 ID-SNPs multialélicos analizados mediante una única PCR tipo *multiplex* y 2 reacciones de SBE. A pesar de que el diseño inicial contemplaba una única reacción SBE, durante la optimización del ensayo se puso de manifiesto la necesidad de separar los SNPs en dos reacciones para individualizar correctamente las señales alélicas. No obstante, el ensayo mantiene la amplificación simultánea de los 29 SNPs en una única PCR inicial que proporciona más que suficiente producto para las 2 reacciones SBE, de manera que el consumo de muestra permanece invariable.

El panel presenta características adecuadas en cuanto a nivel de discriminación: a pesar de que el número de marcadores incorporados es relativamente bajo, cada uno de los SNPs incluidos presenta niveles de heterocigosidad muy altos, elevando la RMP del ensayo hasta valores equivalentes a los de los 13 STRs del CODIS. Así, el uso de SNPs multialélicos permite solventar en parte una de las limitaciones de los marcadores típicamente bialélicos: el bajo poder de discriminación y, consecuentemente, el alto número de marcadores que es necesario combinar para alcanzar niveles de informatividad adecuados.

Asimismo, los marcadores bialélicos presentan limitaciones a la hora de detectar mezclas de ADN. Este panel de ID-SNPs multialélicos constituye una herramienta simple y eficaz para la detección de mezclas. En este sentido, el panel presenta probabilidades de encontrar al menos un SNP con más de dos alelos en el perfil de una mezcla de dos individuos de la misma población >99% para todas las poblaciones. No obstante, debido a las características de la tecnología SNaPshot la deconvolución de mezclas, incluso mezclas sencillas, no parece factible.

Se debe de tener en cuenta que la PCR inicial del panel ha sido diseñada con la intención de ser adaptada fácilmente a tecnologías de MPS (Daniel *et al.* 2015), en las que existe una alta correlación entre la proporción del alelo en la muestra y la frecuencia de las lecturas del mismo. Así, mediante el uso de *software* estadísticos en principio diseñado para STRs, se podrían llegar a realizar inferencias a partir de datos del análisis de SNPs en MPS en mezclas de ADN (Gill *et al.* 2015).

6. Discusión final

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6. Discusión final

La comunidad forense debe investigar y desarrollar herramientas que permitan expandir las aplicaciones de la genética forense, haciendo uso de todos los recursos disponibles para solventar las limitaciones actuales y aportar información útil para las investigaciones policiales y judiciales. Por ello, los trabajos presentados en esta memoria para optar al grado de Doctor pretenden, por una parte, evaluar la capacidad de las nuevas tecnologías genómicas, tratando de establecer las ventajas e inconvenientes de su uso; y, por otra parte, presentar *multiplexes* optimizados de nuevos marcadores para la predicción de ancestralidad biogeográfica y el análisis de mezclas de ADN.

- **Bloque I: ID-SNPs en MPS**

En el primer bloque de la memoria se recogen los trabajos de validación interna de dos paneles de SNPs de identificación para Ion PGM™: HID-Ion AmpliSeq™ Identity v. 2.2 y Qiagen SNP-ID. Los análisis mediante nuevas tecnologías genómicas exigen, por lo general, altas cantidades iniciales de ADN de buena calidad. En un contexto forense, estas condiciones limitan su aplicabilidad a casos en los que se dispone de muestras de referencia, como análisis de parentesco (Lareu *et al.* 2012). Sin embargo, las plataformas de MPS presentan ciertas características, como una mayor sensibilidad y flexibilidad, que las hacen adaptables al análisis de muestras forenses.

La evaluación de ambos paneles permite subrayar consideraciones sobre el uso de la plataforma Ion PGM™ en específico y sobre la aplicabilidad de las tecnologías MPS en general, dado que, a excepción de la metodología de secuenciación, comparten una serie de características definitorias.

En primer lugar, una de las principales ventajas de las plataformas MPS para el análisis forense es la alta capacidad de *multiplexing*. En este sentido, los dos paneles analizados reúnen conjuntos de marcadores previamente establecidos en la comunidad forense, combinando en una única reacción *multiplex* más de 100 SNPs. Este número es muy superior a los ~50 SNPs necesarios para alcanzar un poder de discriminación comparable a ~16 STRs (Krawczak 1999, Gill 2001, Ayres 2005) y confiere una alta robustez a los análisis.

En segundo lugar, la estimación del número de muestras que se pueden analizar simultáneamente en un mismo *run* de Ion PGM™ depende del tipo de chip y del *coverage* mínimo que se pretenda obtener para cada SNP del panel. Los trabajos presentados refuerzan las estimaciones realizadas previamente en otros estudios (Bentley *et al.* 2008, Quail *et al.* 2012, Daniel *et al.* 2015), que apuntan a un *coverage* mínimo de 15-20x para la obtención de genotipos fiables a partir de muestras de referencia. Para muestras de ADN *low level*, degradado o mezclas de ADN, el analista debe apuntar a valores de *coverage* mínimos más altos, tratando de minimizar los efectos estocásticos que se producen durante la amplificación

de este tipo de muestras. Así, la optimización del número de muestras se debe realizar atendiendo a cada panel, chip y tipo de muestra; por lo que supondrá un importante reto si se adaptan estas tecnologías a la rutina.

Se debe tener en cuenta que en la PCR de captura se producen diferencias en la eficiencia de amplificación de los SNPs, de manera que una pequeña proporción de los mismos se encuentran infrarrepresentados respecto al resto. No obstante, ambos paneles presentan un grado de homogeneidad del *coverage* promedio de cada SNP razonable, teniendo en cuenta el alto número de SNPs amplificados simultáneamente, y con valores más de 10 veces por encima del umbral de 20x. El panel HID-Ion AmpliSeq™ Identity presenta los valores de *coverage* promedio más bajos en los Y-SNPs que incluye, indicando cierta correlación entre la dotación genética y el *coverage*. El menor *coverage* de los Y-SNPs y la variación de los niveles promedio de *coverage* de cada SNP refuerzan la idea de que la optimización del número de muestras por *run*, tipo de chip y tipo de muestra debe ser llevada a cabo individualmente para cada panel. Los resultados de comparación del *coverage* observado con el esperado indican que se debe optar por una metodología conservadora, cargando un número de muestras menor al que sugieren las guías de la casa comercial.

En tercer lugar, la visualización de las secuencias en IGV y la aplicación de umbrales para diferentes parámetros de calidad de las secuencias –*coverage* promedio del SNP, porcentaje de incorporaciones erróneas, sesgo de cadena por SNP o por alelo y balance de lecturas de los alelos– permiten identificar una pequeña proporción de SNPs atípicos. Por un lado, estos SNPs atípicos pueden presentar un alto riesgo de *no-call* debido al sesgo de lecturas de las cadenas o un bajo *coverage* promedio, especialmente si se aplican parámetros de *coverage* mínimo rigurosos para la obtención de genotipos. Por otro lado, estos SNPs pueden presentar un alto riesgo de producir genotipos discordantes, debido a que presentan incorporaciones erróneas de nucleótidos alélicos o no alélicos en la posición del SNP, desbalance de los alelos, artefactos tipo Indel o sesgo de cadena de las lecturas de los alelos. La exclusión o pérdida de información de esta pequeña proporción de SNPs tendría efectos mínimos sobre el poder de discriminación de los paneles. Además, la identificación de estos SNPs permite excluirlos de los análisis de muestras comprometidas, elevando el grado de rigor –*stringency*– de dichos análisis. Se debe destacar que un diseño de validación compacto, como el presentado en el estudio del panel Qiagen SNP-ID, permite identificar con fiabilidad estos SNPs atípicos.

En los estudios presentados se observa que los SNPs atípicos con alto riesgo de producir genotipos discordantes presentan comúnmente tractos homopoliméricos o regiones repetitivas adyacentes a la posición del SNP. Los resultados concuerdan con los estudios que indican que la metodología de secuenciación del Ion PGM™ presenta limitaciones a la hora de resolver tractos homopoliméricos, debido a la pérdida de la proporcionalidad entre el número de nucleótidos añadidos y la intensidad de la señal detectada (Loman *et al.* 2012, Ratan *et al.* 2013). Así, se deben tener en cuenta las características de la secuencia contexto durante el diseño de nuevos paneles o al buscar sustitutos en futuras revisiones del panel. Además,

durante el diseño de *primers* se debe evitar que las regiones de *annealing* presenten polimorfismos que puedan afectar a la correcta amplificación del SNP. En la mayoría de los casos se pueden realizar correcciones manuales de los genotipos, atendiendo a que las incorporaciones erróneas no representan uno de los posibles alelos del SNP o a que una de las cadenas no se ve afectada por alineamientos erróneos. Para evitar estas correcciones manuales, los *software* de genotipado necesitan ciertas mejoras que permitan una mayor flexibilidad, como que se puedan establecer los umbrales de los parámetros de genotipado, los posibles alelos o la cadena a partir de la que se debe inferir el genotipo individualmente para cada SNP. Incluir estas mejoras en el *software* permitiría automatizar las correcciones de los genotipos; en caso contrario, estos SNPs tendrían que ser excluidos de los paneles, dado que las correcciones manuales no son deseables ni factibles en la rutina forense. No obstante, se debe señalar que los genotipos obtenidos mediante la plataforma Ion PGM™ presentan niveles muy altos de concordancia con los listados en las bases de datos y los obtenidos mediante otros métodos de genotipado, indicando una alta fiabilidad global del genotipado.

En cuarto lugar, las validaciones de ambos paneles permiten constatar que se pueden expandir los límites de sensibilidad de la plataforma. Mientras las casas comerciales indican cantidades de ADN inicial óptimas de 10-20 ng, los análisis de ADN *low level* revelan que se pueden analizar cantidades más de 40 veces menores, perdiéndose en algún caso la información para ~10 SNPs, lo que produce un efecto mínimo sobre el poder de discriminación global. En las muestras de ADN *low level* se acentúan las diferencias en *coverage* promedio entre los SNPs, por lo que se debe apuntar a mayores niveles de *coverage* mínimo que para las muestras de referencia. Asimismo, los análisis preliminares indican que se pueden obtener perfiles a partir de muestras de ADN degradado, aunque los límites de la plataforma deben ser explorados en mayor profundidad. El ADN extraído a partir de algunas muestras forenses (p. ej. restos esqueléticos) puede contener inhibidores, de manera que, para realizar un estudio adecuado e informativo de las capacidades de la plataforma, se debe caracterizar el grado de inhibición y degradación de las muestras.

En cuanto a los análisis de mezclas de ADN, los resultados indican que una ventaja de las plataformas de MPS en el análisis de SNPs, en comparación con SNaPshot, es la capacidad de detección de mezclas de ADN. La alta correlación entre la proporción de lecturas de cada alelo y su representación en la muestra original permite detectar mezclas de ADN en SNPs bialélicos a través del desbalance de las frecuencias de lecturas de los alelos. Una vez que se detecta la presencia de una mezcla de ADN se deben reconfigurar los parámetros del *software* de análisis de manera que se disminuya el umbral para asignar genotipos heterocigotos, evitando un mayor número de *drop-outs*. Además, la adaptación de métodos inicialmente diseñados para análisis de STRs permitirá realizar inferencias estadísticas a partir de los análisis de SNPs en MPS (Gill *et al.* 2015).

En definitiva, ambos trabajos proporcionan argumentos favorables al uso de las plataformas de MPS para el análisis de SNPs, ya que se solventarían las dos principales limitaciones del análisis de SNPs bialélicos mediante SNaPshot: aumentan la capacidad de

multiplex, necesaria para alcanzar niveles de poder de discriminación adecuados, y permiten una detección más fiable de las mezclas de ADN. Además, los ensayos presentan unos niveles de sensibilidad adecuados para la mayoría de muestras. No obstante, la implantación de estas metodologías puede suponer un reto y es necesario realizar reajustes del *software* que permitan adecuarlo a las necesidades del análisis forense: más flexible y, necesariamente, menos opaco.

- **Bloque II: Ancestralidad biogeográfica**

En el segundo bloque de la memoria se recogen los trabajos sobre predicción de ancestralidad biogeográfica. La comunidad forense cuenta con herramientas ya establecidas para esta aplicación, que han demostrado su utilidad en las investigaciones policiales (Phillips *et al.* 2009). Como indica el estudio de Taboada-Echalar *et al.* (2013), estas herramientas pueden presentar sesgos de inferencia de las proporciones de coancestralidad cuando se analizan individuos *admixed*. Para solventarlo, se diseñó un panel de 128 marcadores denominado EUROFORGEN AIM-SNP (Phillips *et al.* 2014a), que presenta un poder de diferenciación equivalente para cada una de las cinco poblaciones definidas continentalmente. Para que la comunidad pueda beneficiarse de las ventajas de este panel, es necesario llevarlo a la práctica, adaptándolo a metodologías de genotipado que permitan el análisis del tipo de muestras comúnmente encontradas en los casos forenses. En el primero de los trabajos, se adapta el panel a SNaPshot y, en el segundo, a la plataforma de MPS Ion PGM^{TM}.

La adaptación del panel a SNaPshot permite su implementación en todos los laboratorios de rutina forense ya que la metodología de análisis precisa de los mismos instrumentos que el análisis de STRs: un termociclador y un equipo de electroforesis capilar. No obstante, debido a las limitaciones de *multiplexing* inherentes a SNaPshot, se requiere disminuir el número de marcadores del panel original. En el nuevo panel, G-AIMs Nano, se incluyen 31 SNPs que representan la mayoría de los marcadores más informativos del panel EUROFORGEN AIM-SNP.

La reducción del número de marcadores conlleva, inevitablemente, una reducción del poder de diferenciación de las poblaciones, que se produjo de una forma desproporcionada para EAS. Este hecho se puede solventar mediante la adición de 1-2 marcadores específicamente informativos para esta población. No obstante, la búsqueda de marcadores informativos para cada población está limitada por la propia estructura genética de las poblaciones: como consecuencia de la historia evolutiva humana los pares de poblaciones EAS-AMR y EAS-OCE presentan unos niveles bajos de diferenciación.

A pesar de esto, las diferentes metodologías aplicadas para los análisis poblacionales indican que las inferencias de proporciones de coancestralidad de poblaciones *admixed* obtenidas mediante el panel G-AIMs Nano se corresponden adecuadamente con las del panel original EUROFORGEN AIM-SNP. Además, las LRs de asignación de ancestralidad de 5 ADNs control de ancestralidad conocida obtenidas con este panel presentan mejores resultados que los paneles de 34 SNPs (Fondevila *et al.* 2013) y 46 Indels (Pereira *et al.* 2012) ya establecidos en la comunidad.

El ensayo SNaPshot es altamente sensible, ya que se obtienen perfiles completos con 0,063 ng de ADN inicial, y robusto, ya que se obtienen predicciones con LRs>1000 aunque se pierda la información de hasta la mitad de los marcadores más informativos del panel. Estas características favorecen el análisis de ADN *low level* y degradado. Además, el panel G-AIMs Nano incluye 3 SNPs trialélicos, que permiten detectar mezclas de ADN en SNaPshot si los perfiles muestran 3 alelos para al menos un marcador. No obstante, los paneles de predicción de ancestralidad biogeográfica son herramientas que se utilizan cuando los perfiles de STRs no aportan información de utilidad mediante comparaciones, por lo que esta información ya estaría disponible de antemano. Por otra parte, la selección compacta de marcadores que se presenta en este panel constituye un complemento ideal para paneles de MPS en los que se quiera incluir un método sencillo de predicción de ancestralidad geográfica a escala continental.

La adaptación del panel EUROFORGEN AIM-SNP a MPS constituye uno de los primeros trabajos de adaptación de paneles personalizados para uso forense mediante *primers* AmpliSeq™ para Ion PGM™. A pesar de que las capacidades *multiplex* de la plataforma son superiores al número de marcadores incluidos en el panel, durante el diseño de la PCR de captura 3 de los 128 SNPs del panel original tuvieron que ser sustituidos, ya que se sitúan en regiones repetitivas que conllevan problemas de unión inespecífica de los *primers*. A pesar de la alta tasa de conversión del panel, este hecho apunta a la necesidad de seleccionar conjuntos de posibles marcadores sustitutos durante el diseño de paneles para MPS. Se debe tener en cuenta que en paneles de predicción de características externas visibles ciertos SNPs no pueden ser sustituidos por otros que presenten una informatividad semejante. No obstante, estos SNPs suelen presentarse en regiones codificantes, que presentan habitualmente regiones contexto complejas y no repetitivas.

Además, los análisis de concordancia de los genotipos y calidad de las secuencias obtenidas indican que el SNP rs2080161 debe ser excluido del panel y que los genotipos de rs595961, rs6875659 y, en algunas muestras de individuos AFR, rs12402499 deben ser corregidos manualmente. Estos SNPs presentan ciertas características en sus secuencias contexto –Indels y tractos homopoliméricos en las regiones adyacentes a la posición del SNP– que podrían haber sido identificadas previamente, acentuando la necesidad de realizar un escrutinio profundo y detallado de las secuencias contexto durante la selección de SNPs para MPS. No obstante, la sustitución de 3 de los SNPs originales del panel y la exclusión del SNP rs2080161 provocan un efecto mínimo sobre el balance original del poder de diferenciación de las 5 poblaciones definidas continentalmente. Así, el panel adaptado permite realizar estimaciones de proporciones de coancestralidad no sesgadas en individuos *admixed*.

A pesar de que el panel original fue diseñado para análisis poblacionales en base a 5 grupos de referencia, las diferentes metodologías aplicadas indican que se puede diferenciar la población SAS de EUR. El poder de diferenciación de SAS es menor que el del resto de grupos poblacionales, por lo que se pueden producir sesgos en las estimaciones de proporciones de coancestralidad. Se debe tener en cuenta que cuando se diseñó el panel

original EUROFORGEN AIM-SNP las poblaciones SAS no estaban recogidas en el Proyecto 1000 Genomas. La expansión de las bases de datos permitirá identificar nuevos AIM-SNPs específicamente informativos para esta población y, dado que las altas capacidades *multiplexing* de los sistemas MPS lo permiten, futuras revisiones del panel pueden abordar una diferenciación balanceada de SAS mediante la adición de nuevos AIM-SNPs.

Los resultados de la validación forense del panel indican una adecuada sensibilidad forense. En primer lugar, los genotipos de las muestras de ADN *low level* presentan una concordancia total con hasta 100 pg de ADN inicial, mientras que con cantidades iniciales de ADN inferiores se pruducen *no-calls*, *drop-outs* y *drop-ins*. En segundo lugar, la capacidad de detección de mezclas de la plataforma permite diferenciar las mezclas de ADN de individuos de ancestralidades diferentes de las de individuos *admixed*. En este sentido, la inclusión de 6 SNPs trialélicos en el panel proporciona un método adicional para la identificación de mezclas de ADN. No obstante, el *software* de análisis debe ser modificado de manera que, cuando se detecte una mezcla, permita la identificación automática de un tercer o cuarto alelo en los SNPs que se indiquen como multialélicos.

- **Bloque III: Mezclas de ADN**

En el tercer bloque de la memoria se recogen dos trabajos en los que se presentan ensayos de nuevos marcadores que facilitan el análisis de mezclas de ADN, uno de los principales retos de la genética forense. En este sentido, la comunidad forense está realizando un gran esfuerzo en desarrollar metodologías que permitan realizar inferencias estadísticas a partir de los perfiles de mezclas. Los trabajos aquí recogidos pretenden aportar soluciones sobre la primera fase limitante del análisis de mezclas: la interpretación del perfil.

Los STRs, como marcadores multialélicos, permiten una fácil detección de las mezclas de ADN: encontrar más de dos alelos en varios sistemas es un indicador fiable de que estamos ante una mezcla de ADN y no un perfil individual. No obstante, el uso de STRs presenta una limitación importante: la presencia de picos *stutter*. Los picos *stutter* son un fenómeno inherente a la amplificación de muchos de los STRs que se analizan comúnmente y habitualmente presentan una repetición menos que los alelos reales de los que provienen. Así, estos artefactos aparecen en posiciones alélicas, dificultando la interpretación de los perfiles de STRs: es necesario identificar los picos *stutter* como tales y no confundirlos con un alelo del componente minoritario de la mezcla o con la señal combinada de ambos.

En el primer trabajo se presenta un panel que combina 9 nuevos STRs pentaméricos y dos marcadores específicos de cromosoma Y: el STR DYS391 y el Y-Indel rs2032678. Los nuevos STRs pentaméricos presentan una actividad *stutter* reducida debido a la mayor longitud de su unidad de repetición, con tasas más de 3 veces menores que las de STRs tetraméricos de tamaño equivalente. Esta característica representa una ventaja en los análisis de mezclas de ADN, como demostraron los análisis comparativos exploratorios: teniendo en cuenta la baja tasa *stutter* de los STRs pentaméricos es más sencillo identificar los picos *stutter* de los perfiles como tales. Además, en los casos analizados el número de picos por encima del umbral de detección para cada STR permite inferir correctamente el número

mínimo de contribuyentes de la mezcla; mientras que los STRs tetraméricos presentan picos *stutter* por encima del umbral de detección que pueden llevar a sobreestimar el número mínimo de contribuyentes si no se descuentan apropiadamente.

Los nuevos STRs pentaméricos presentan unos niveles de poder de discriminación menores que los STRs comúnmente utilizados en genética forense, debido a su menor variabilidad. No obstante, el panel se diseña como una herramienta complementaria a los kits de STRs autosómicos convencionales en casos que así lo requieran. En este sentido, además de los análisis de mezclas de ADN, el panel puede aportar información en casos en los que el poder de discriminación de los kits convencionales no sea suficiente, como el análisis de pedigríes complejos.

En el análisis de STRs mediante MPS, los *stutter* también constituyen una importante limitación en el análisis de mezclas de ADN, al ser fenómenos dependientes de la PCR. La capacidad de estas plataformas para diferenciar isoalelos facilitará la interpretación de los perfiles complejos. No obstante, dadas las altas capacidades *multiplex* que presentan estas metodologías, analizar simultáneamente conjuntos de marcadores con diferentes propiedades permitirá ampliar las ventajas de la plataforma, así como la información obtenida para cada tipo de muestra. Por ello, la búsqueda de nuevos marcadores con características de amplificación diferenciales sigue siendo de especial interés en la comunidad forense.

Cuando se analiza ADN degradado, los marcadores bialélicos –SNPs e Indels– presentan ventajas ya que, al ser polimorfismos más cortos, se pueden diseñar amplicones más pequeños, elevando la tasa de éxito de amplificación. No obstante, los marcadores bialélicos presentan importantes limitaciones a la hora de detectar mezclas, limitaciones que, en el caso de los SNPs, se ven magnificadas por las características de la metodología de genotipado más extendida: SNaPshot.

El segundo de los trabajos de este bloque presenta un ensayo SNaPshot para 29 SNPs multialélicos, principalmente trialélicos dado su mayor abundancia relativa en el genoma. El uso de SNPs multialélicos permite, por una parte, elevar la informatividad de cada marcador incluido en el *multiplex*, de manera que se alcanzan niveles de poder de discriminación adecuados con un menor número de marcadores. Por otra parte, la aparición en los perfiles de más de dos alelos para un mismo marcador permite la detección de mezclas de ADN.

El ensayo SNaPshot presenta características adecuadas para su uso en genética forense y presenta buenos niveles de RMP en las 5 poblaciones definidas continentalmente, a pesar de que no se alcanza un balance total del poder de discriminación entre ellas. Se debe tener en cuenta que varios de los marcadores son bialélicos para al menos una de las poblaciones estudiadas, de manera que la probabilidad de encontrar al menos un SNP que presente más de 2 alelos en una mezcla de 2 individuos es más baja en algunas poblaciones, llegando a alcanzar aún así probabilidades por encima del 99,8% en todos los casos. Así, se confirma que estos marcadores son una herramienta sencilla y eficaz para la detección de mezclas que contengan ADN degradado.

La total implementación de los 29 marcadores en SNaPshot permite su uso a pequeña escala en aquellos laboratorios que no tengan acceso a plataformas de MPS. No obstante, durante la selección de los SNPs del panel y de los *primers* de PCR, se puso un especial cuidado en evitar regiones repetitivas o tractos homopoliméricos. Estas medidas pretenden favorecer la adaptación del panel a MPS, aprovechando las características de alta flexibilidad de estas plataformas, tal y como ha demostrado un estudio previo (Daniel *et al.* 2015). En estas plataformas existe una alta correlación entre la proporción del alelo en la muestra y la frecuencia de las lecturas del mismo. Así, mediante el uso de *software* estadísticos en principio diseñados para STRs, se podría llegar a realizar inferencias a partir datos del análisis de SNPs en MPS en mezclas de ADN (Gill *et al.* 2015).

En definitiva, en este bloque se presentan paneles de nuevos marcadores con características que facilitan el análisis de mezclas y son independientes de la metodología de genotipado utilizada, por lo que sus ventajas se mantienen aunque los sistemas de electroforesis capilar sean reemplazados en un futuro por plataformas de MPS.

7. Conclusiones

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7. Conclusiones

7.1 Sobre la validación de paneles de ID-SNPs para Ion PGM™

7.1.1 Validación del panel HID-Ion AmpliSeq™ Identity v. 2.2

1. Los resultados de la evaluación interlaboratorio del panel HID-Ion AmpliSeq™ Identity v. 2.2 en cuanto a sensibilidad forense y concordancia de genotipos proporcionan argumentos favorables para la utilización de las tecnologías de MPS en genética forense.

2. Los datos indican que más de un 80% de los SNPs del panel presentan buenos valores de *coverage* y producen genotipos fiables. Sin embargo, un total de 5 SNPs discordantes (rs2032597, rs2399332, rs1979255, rs1004357 y rs938283) deben ser excluidos del panel.

3. Las estimaciones del número de muestras óptimo para cada tipo de chip deben realizarse en función del panel que se utilice y del *coverage* mínimo que se desea obtener. En los análisis con una baja concentración de ADN –*low level*– se acentúan las diferencias de *coverage* entre los SNPs, por lo que se deben mantener estimaciones conservadoras en el número de muestras que se cargan por *run*.

4. En los análisis con cantidades iniciales de ADN de 100, 50 y 25 pg, se aprecia un incremento del número de *no-calls* (entre 8-12 SNPs). La pérdida de la información de dichos SNPs produce un efecto mínimo sobre los valores de probabilidad de coincidencia al azar –RMP: *random match probability*– que, incluso con la pérdida de la información del 40-50% de los SNPs del panel, alcanzan valores similares a los del panel de STRs GlobalFiler.

5. El aumento de la heterocigosidad, la pérdida del balance de las frecuencias de lecturas de los alelos y, en mezclas con componente masculino, la disminución del *coverage* relativo de los Y-SNPs permiten una detección fiable de las mezclas de ADN.

6. Aunque el uso del *software* Torrent Suite™ es adecuado para el genotipado, necesita importantes mejoras para su aplicación en genética forense. En primer lugar, proporciona poco margen para cambiar umbrales de parámetros de calidad de las secuencias. En segundo lugar, los parámetros recomendados son los de *Germline*, pero es necesario aplicar parámetros *Somatic* para un análisis adecuado de las mezclas de ADN.

7.1.2 **Validación del panel Qiagen SNP-ID**

1. El escrutinio detallado de las secuencias obtenidas para los controles de ADN revela que un total de 12 SNPs (rs4847034, rs1554472, rs4796362, rs1004357, rs733164, rs1821380, rs2270529, rs1029047, rs2399332, rs4606077, rs445251 y rs1523537) presentan un alto riesgo de producir genotipos discordantes. Las características de las secuencias contexto de dichos SNPs incluyen tractos homopoliméricos o regiones repetitivas que provocan alineamientos erróneos de las cadenas. Así, el bajo rendimiento de dichos SNPs es independiente de la metodología de preparación de librerías utilizado.

2. La mayoría de los genotipos pueden ser corregidos manualmente, salvo para rs1029047 que debe ser excluido del panel. Futuras actualizaciones de *software* de análisis que permitan establecer individualmente para cada SNP los umbrales de los parámetros de genotipado, los posibles alelos o la cadena a partir de la que se debe inferir el genotipo, permitirían la corrección automatizada de los genotipos. En caso contrario, deben ser reemplazados o eliminados del panel, ya que la corrección manual de genotipos no es factible en rutina forense.

3. El panel Qiagen SNP-ID incluye 140 SNPs recogidos de dos paneles de SNPs de identificación previamente establecidos en la comunidad: 52 SNPs del panel SNP*for*ID (Sánchez *et al.* 2006) y 92 de *Kiddlab* (Pakstis *et al.* 2010), con 4 SNPs en común entre los dos paneles. Dos pares de SNPs próximos (rs10768550-rs10500617 y rs9606186-rs5746846) no son aplicables como marcadores independientes y presentan haplotipos con una informatividad similar al uso de un único SNP de cada par.

4. La concordancia entre los genotipos obtenidos mediante el panel Qiagen SNP-ID y los listados por la base de datos del Proyecto 1000 Genomas es de un 99,52%, y alcanza el 100% tras la corrección de los genotipos de dos SNPs (rs1004357 y rs5746846) previamente identificados como atípicos. Los datos apuntan a una alta precisión del genotipado y altos valores de *coverage* incluso al analizar cantidades iniciales de ADN entre 0,125-1 ng, muy por debajo de los 20 ng recomendados por el fabricante. Además, los análisis de la muestra de ADN degradado produjeron genotipos para el 100% de los SNPs, concordantes entre réplicas.

5. A pesar de que todos los marcadores incluidos en el panel son bialélicos, el aumento del nivel de heterocigosidad y la distribución atípica de los patrones de frecuencias de lecturas de los alelos permiten detectar la presencia de mezclas de ADN. En estos casos, para una correcta detección del componente minoritario, se debe disminuir el umbral del parámetro que determina la frecuencia mínima para asignar un genotipo heterocigoto.

7.2 SOBRE LOS NUEVOS PANELES PARA PREDICCIÓN DE ANCESTRALIDAD BIOGEOGRÁFICA

7.2.1 G-AIMs Nano

1. Se presenta un ensayo SNaPshot optimizado y validado para su uso en genética forense, que incluye 31 AIM-SNPs recopilados de entre los más informativos del panel EUROFORGEN AIM-SNP.

2. Pese a la reducción del número original de marcadores del panel en un 75%, el balance final del poder de diferenciación acumulado para cada uno de los cinco grupos poblacionales definidos continentalmente permite realizar estimaciones de ancestralidad no sesgadas en individuos *admixed*.

3. El poder de diferenciación de la población EAS es ligeramente menor que el del resto de grupos poblacionales. En futuros ajustes del panel, se podría solventar adicionando uno o dos marcadores específicamente informativos para EAS.

4. El análisis de controles de ADN de ancestralidad conocida indica que las LRs de las asignaciones de ancestralidad en comparaciones de 5 grupos mediante los 31-SNPs de este panel superan o son equivalentes a las obtenidas mediante paneles establecidos en la comunidad forense de 34-SNPs y/o 46 Indels.

7.2.2 Adaptación a Ion PGM™ y validación del panel Global AIM-SNP

1. Un total de 125 de los 128 marcadores originales del set original EUROFORGEN AIM-SNP, más 3 SNPs sustitutos, se incluyeron en una reacción tipo *multiplex* personalizada para Ion PGM™; indicando una alta tasa de conversión de los paneles de SNPs teóricos a MPS.

2. Los resultados del análisis de concordancia del genotipado indican que un total de tres SNPs –rs2080161, rs595961 y rs6875659– presentan discordancias sistemáticamente. La presencia de tractos homopoliméricos cercanos a la posición del SNP produce un alineamiento erróneo de las secuencias y causa las discordancias; por lo que los SNPs podrían haber sido fácilmente identificados y sustituidos durante el proceso de diseño de la *multiplex* personalizada. Además, el SNP rs12402499 presenta un alto número de *no-calls* en las muestras de estudio AFR, debido a la presencia de un Indel específico de población adyacente al SNP. Los genotipos pueden corregirse manualmente, excepto en el caso del SNP rs2080161, que debe ser excluido del panel.

3. La sustitución de 3 SNPs del panel original y la exclusión de rs2080161 produce un efecto mínimo sobre el balance final del poder de diferenciación acumulado para cada uno de los cinco grupos poblacionales definidos continentalmente. Este balance

permitió realizar estimaciones de ancestralidad no sesgadas en individuos de las nuevas poblaciones de la Fase III del Proyecto 1000 Genomas y de 14 nuevas poblaciones de estudio seleccionadas para ampliar el ámbito geográfico disponible a través de las bases de datos.

4. Aunque el panel Global AIM-SNP fue diseñado para realizar inferencias de ancestralidad basadas en 5 grupos poblacionales de referencia definidos continentalmente, las diferentes aproximaciones utilizadas para el análisis poblacional permitieron diferenciar SAS de EUR en la mayoría de los casos. No obstante, los análisis de ancestralidad en base a 6 grupos de referencia pueden presentar un sesgo de infraestimación de la proporción del componente SAS frente al resto de grupos, debido a que presenta un poder de diferenciación marcadamente reducido.

5. La alta correlación entre la proporción de cada alelo en la muestra y el número de lecturas obtenidas permiten diferenciar mezclas de ADN de componentes con diferentes ancestralidades de muestras de individuos *admixed*. Además, los 6 SNPs trialélicos del panel proporcionan un segundo método de detección de mezclas.

7.3 SOBRE EL DESARROLLO DE PANELES DE NUEVOS MARCADORES PARA MEZCLAS DE ADN

7.3.1 **STRs pentaméricos**

1. Se presenta un panel que incluye 9 nuevos STRs pentaméricos y 2 marcadores específicos de cromosoma Y: el STR DYS391 y el Y-Indel rs2032678. El ensayo está optimizado para la obtención de perfiles de buena calidad a partir de muestras de ADN *low level* o extraídas de material esquelético, típicas de casos de rutina forense.

2. La mayoría de los nuevos STRs pentaméricos desarrollados presentan un poder de discriminación menor que el de los STRs tetraméricos comúnmente utilizados en genética forense. No obstante, este ensayo constituye un avance en la búsqueda de *loci* complementarios aplicables cuando los kits de STRs autosómicos convencionales no proporcionen suficiente información.

3. Los 9 STRs pentaméricos presentan una actividad *stutter* más de 3 veces menor que STRs tetraméricos de tamaño equivalente, con niveles del ~2% similares a los de STRs pentaméricos ya establecidos. La baja tasa de *stutter* de los STRs pentaméricos simplifica el análisis de las mezclas de ADN, ya que permite diferenciar más fácilmente los picos *stutter* de posibles alelos de un componente minoritario. Además, en los casos analizados, el número de picos identificados en los STRs pentaméricos proporciona una inferencia correcta del número mínimo de contribuyentes a la mezcla.

4. Dado que los *stutter* son un fenómeno derivado de la PCR, seguirán constituyendo un problema en la interpretación de mezclas aunque los secuenciadores de electroforesis capilar sean reemplazados por sistemas de MPS. Además, el hecho de que los sistemas MPS permitan analizar simultáneamente un mayor número de STRs fomenta el interés en explorar marcadores adicionales con nuevas propiedades como los presentados en este trabajo.

7.3.2 **ID-SNPs multialélicos**

1. Se presenta un panel que comprende 29 ID-SNPs multialélicos, analizados mediante una única PCR tipo *multiplex* y 2 reacciones de SBE con SNaPshot. El uso de SNPs multialélicos permite reducir las desventajas derivadas de su naturaleza típicamente bialélica, manteniendo sus ventajas en cuanto al análisis de ADN degradado.

2. Los 29 ID-SNPs multialélicos se encuentran bien distribuidos por el genoma y presentan niveles de heterocigosidad altos, de manera que el panel alcanza niveles de RMP adecuados en todas las poblaciones. No obstante, la priorización del grupo EUR produce cierto desbalance en el poder de discriminación de las diferentes poblaciones definidas continentalmente.

3. La probabilidad acumulada de encontrar al menos un SNP que presente más de dos alelos en el perfil de una mezcla de ADN de dos individuos de la misma población es mayor del 99,8% en todos los casos, a pesar de que ciertos SNPs no presentan patrones multialélicos en todas las poblaciones. Así, el panel constituye una herramienta eficaz a la hora de detectar mezclas que contengan ADN degradado. No obstante, las características de la tecnología SNaPshot impiden la deconvolución de mezclas, incluso en los casos más sencillos.

4. Durante la selección de SNPs y la optimización de la PCR inicial se evitaron secuencias contexto con tractos homopoliméricos o regiones repetitivas, de manera que se favorezca la adaptación del panel a MPS. Las tecnologías MPS presentan una alta correlación entre la proporción de cada alelo en la muestra original y la frecuencia de lecturas de los alelos de manera que, al menos en casos sencillos, se podría llegar a identificar los alelos de los componentes mayoritario y minoritario.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8. Bibliografía

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8. Bibliografía

Alaeddini R, Walsh SJ y Abbas A (2010). "*Forensic implications of genetic analyses from degraded DNA--a review*." Forensic Sci Int Genet **4**(3): 148-157.

Albrecht U (2006). "*Orchestration of gene expression and physiology by the circadian clock*." J Physiol Paris **100**(5-6): 243-251.

Allocco DJ, Song Q, Gibbons GH, Ramoni MF y Kohane IS (2007). "*Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms*." BMC Genomics **8**: 68.

Amigo J, Salas A, Phillips C y Carracedo A (2008). "*SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access*." BMC Bioinformatics **9**: 428.

Amorim A y Pereira L (2005). "*Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs*." Forensic Sci Int **150**(1): 17-21.

An JH, Choi A, Shin KJ, Yang WI y Lee HY (2013). "*DNA methylation-specific multiplex assays for body fluid identification*." Int J Legal Med **127**(1): 35-43.

Andersen JD, Johansen P, Harder S, Christoffersen SR, Delgado MC, Henriksen ST, Nielsen MM, Sorensen E, Ullum H, Hansen T, *et al.* (2013). "*Genetic analyses of the human eye colours using a novel objective method for eye colour classification*." Forensic Sci Int Genet **7**(5): 508-515.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, *et al.* (1981). "*Sequence and organization of the human mitochondrial genome*." Nature **290**(5806): 457-465.

Antheunisse J (1972). "*Decomposition of nucleic acids and some of their degradation products by microorganisms*." Antonie van Leeuwenhoek **38**(1): 311-327.

Arends MJ, Morris RG y Wyllie AH (1990). "*Apoptosis. The role of the endonuclease*." Am J Pathol **136**(3): 593-608.

Astbury WT (1947). "*X-ray studies of nucleic acids*." Symp Soc Exp Biol **1**: 66-78.

Avery OT, Macleod CM y McCarty M (1944). "*Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III*." J Exp Med **79**(2): 137-158.

Ayres KL (2005). "*The expected performance of single nucleotide polymorphism loci in paternity testing*." Forensic Sci Int **154**(2-3): 167-172.

Babol-Pokora K y Berent J (2008). "*SNP-minisequencing as an excellent tool for analysing degraded DNA recovered from archival tissues*." Acta Biochim Pol **55**(4): 815-819.

Bacher J y Schumm JW (1998). "*Development of Highly Polymorphic Pentanucleotide Tandem Repeat Loci with Low Stutter*." Profiles in DNA **2**: 3-6.

Bamshad M, Wooding S, Salisbury BA y Stephens JC (2004). "*Deconstructing the relationship between genetics and race*." Nat Rev Genet **5**(8): 598-609.

Bashir M y Hassan NH (2016). "*Analysis of 30 Biallelic INDEL Markers Using the Investigator DIPplex Kit*." Methods Mol Biol **1420**: 135-142.

Bauer CM, Niederstatter H, McGlynn G, Stadler H y Parson W (2013). "*Comparison of morphological and molecular genetic sex-typing on mediaeval human skeletal remains*." Forensic Sci Int Genet **7**(6): 581-586.

Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W y Decorte R (2015). "*Improved age determination of blood and teeth samples using a selected set of DNA methylation markers*." Epigenetics **10**(10): 922-930.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, *et al.* (2008). "*Accurate whole human genome sequencing using reversible terminator chemistry*." Nature **456**(7218): 53-59.

Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S y Vilain E (2011). "*Epigenetic predictor of age*." PLoS One **6**(6): e14821.

Børsting C, Sánchez JJ, Hansen HE, Hansen AJ, Bruun HQ y Morling N (2008). "*Performance of the SNPforID 52 SNP-plex assay in paternity testing*." Forensic Sci Int Genet **2**(4): 292-300.

Børsting C, Tomas C y Morling N (2012). "*Typing of 49 autosomal SNPs by single base extension and capillary electrophoresis for forensic genetic testing*." Methods Mol Biol **830**: 87-107.

Børsting C, Fordyce SL, Olofsson J, Mogensen HS y Morling N (2014). "*Evaluation of the Ion Torrent™ HID SNP 169-plex: A SNP typing assay developed for human identification by second generation sequencing*." Forensic Sci Int Genet **12**: 144-154.

Børsting C y Morling N (2015). "*Next generation sequencing and its applications in forensic genetics*." Forensic Sci Int Genet **18**: 78-89.

Branicki W, Brudnik U, Kupiec T, Wolanska-Nowak P y Wojas-Pelc A (2007). "*Determination of phenotype associated SNPs in the MC1R gene*." J Forensic Sci **52**(2): 349-354.

Branicki W, Liu F, van Duijn K, Draus-Barini J, Pospiech E, Walsh S, Kupiec T, Wojas-Pelc A y Kayser M (2011). "*Model-based prediction of human hair color using DNA variants*." Hum Genet **129**(4): 443-454.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U y Pääbo S (2009). "*Primer Extension Capture: Targeted Sequence Retrieval from Heavily Degraded DNA Sources*." J Vis Exp(31): 1573.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J y Rolf B (1998). "*Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat*." Am J Hum Genet **62**(6): 1408-1415.

Brookes C, Bright JA, Harbison S y Buckleton J (2012). "*Characterising stutter in forensic STR multiplexes*." Forensic Sci Int Genet **6**(1): 58-63.

Brown WM, George M y Wilson AC (1979). "*Rapid evolution of animal mitochondrial DNA*." Proc Natl Acad Sci U S A **76**(4): 1967-1971.

Budowle B (2004). "*SNP typing strategies*." Forensic Sci Int **146 Suppl**: S139-142.

Budowle B y van Daal A (2008). "*Forensically relevant SNP classes*." Biotechniques **44**(5): 603-608, 610.

Budowle B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerrieri RA, Luttman JC y McClure DL (2009). "*Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework*." J Forensic Sci **54**(4): 810-821.

Budowle B, Ge J, Chakraborty R, Eisenberg AJ, Green R, Mulero J, Lagace R y Hennessy L (2011). "*Population genetic analyses of the NGM STR loci*." Int J Legal Med **125**(1): 101-109.

Burger J, Hummel S, Hermann B y Henke W (1999). "*DNA preservation: a microsatellite-DNA study on ancient skeletal remains*." Electrophoresis **20**(8): 1722-1728.

Butler JM, Buel E, Crivellente F y McCord BR (2004). "*Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis*." Electrophoresis **25**(10-11): 1397-1412.

Butler JM (2005). *Chapter 15 - STR genotyping issues*. Forensic DNA Typing: biology, technology and genetics of STR markers. San Diego, Academic Press**:** 373-388.

Butler JM (2006). "*Genetics and genomics of core short tandem repeat loci used in human identity testing*." J Forensic Sci **51**(2): 253-265.

Butler JM, Coble MD y Vallone PM (2007). "*STRs vs. SNPs: thoughts on the future of forensic DNA testing*." Forensic Sci Med Pathol **3**(3): 200-205.

Butler JM (2012a). *Chapter 5 - Short Tandem Repeat (STR) Loci and Kits*. Advanced Topics in Forensic DNA Typing: Methodology. San Diego, Academic Press**:** 99-139.

Butler JM (2012b). *Chapter 7 - Quality Assurance and Validation*. Advanced Topics in Forensic DNA Typing: Methodology. San Diego, Academic Press**:** 167-211.

Butler JM (2012c). *Chapter 8 - DNA Databases: Uses and Issues*. Advanced Topics in Forensic DNA Typing: Methodology. San Diego, Academic Press**:** 213-270.

Butler JM (2012d). *Chapter 11 - Low-Level DNA Testing: Issues, Concerns, and Solutions*. Advanced Topics in Forensic DNA Typing: Methodology. San Diego, Academic Press**:** 311-346.

Butler JM (2012e). *Chapter 12 - Single Nucleotide Polymorphisms and Applications*. Advanced Topics in Forensic DNA Typing: Methodology. San Diego, Academic Press**:** 347-369.

Butler JM (2015). *Chapter 3 - STR Alleles and Amplification Artifacts*. Advanced Topics in Forensic DNA Typing: Interpretation. San Diego, Academic Press**:** 47-86.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, *et al.* (2002). "*A human genome diversity cell line panel*." Science **296**(5566): 261-262.

Chaitanya L, Walsh S, Andersen JD, Ansell R, Ballantyne K, Ballard D, Banemann R, Bauer CM, Bento AM, Brisighelli F, *et al.* (2014). "*Collaborative EDNAP exercise on the IrisPlex system for DNA-based prediction of human eye colour*." Forensic Sci Int Genet **11**: 241-251.

Chen HD, Chang CH, Hsieh LC y Lee HC (2005). "*Divergence and Shannon information in genomes*." Phys Rev Lett **94**(17): 178103.

Chien A, Edgar DB y Trela JM (1976). "*Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus*." J Bacteriol **127**(3): 1550-1557.

Cho S, Ge J, Seo SB, Kim K, Lee HY y Lee SD (2014). "*Age estimation via quantification of signal-joint T cell receptor excision circles in Koreans*." Leg Med (Tokyo) **16**(3): 135-138.

Churchill JD, Schmedes SE, King JL y Budowle B (2016). "*Evaluation of the Illumina((R)) Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling*." Forensic Sci Int Genet **20**: 20-29.

Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, Pearson LN, McEvoy B, Bauchet M, Zaidi AA, Yao W, *et al.* (2014a). "*Modeling 3D facial shape from DNA*." PLoS Genet **10**(3): e1004224.

Claes P, Hill H y Shriver MD (2014b). "*Toward DNA-based facial composites: preliminary results and validation*." Forensic Sci Int Genet **13**: 208-216.

Cole TJ (2003). "*The secular trend in human physical growth: a biological view*." Econ Hum Biol **1**(2): 161-168.

Collins MJ, Nielsen–Marsh CM, Hiller J, Smith CI, Roberts JP, Prigodich RV, Wess TJ, Csapò J, Millard AR y Turner–Walker G (2002). "*The survival of organic matter in bone: a review*." Archaeometry **44**(3): 383-394.

Colonna V, Pagani L, Xue Y y Tyler-Smith C (2011). "*A world in a grain of sand: human history from genetic data*." Genome Biol **12**(11): 234.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW y Pritchard JK (2009). "*The role of geography in human adaptation*." PLoS Genet **5**(6): e1000500.

Crick F (1970). "*Central dogma of molecular biology*." Nature **227**(5258): 561-563.

Daniel R, Santos C, Phillips C, Fondevila M, van Oorschot RAH, Carracedo Á, Lareu MV y McNevin D (2015). "*A SNaPshot of next generation sequencing for forensic SNP analysis*." Forensic Sci Int Genet **14**: 50-60.

Darwin C y Wallace A (1858). "*On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection*." Zool J Linn Soc **3**(9): 45-62.

Darwin CR (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London, John Murray.

de Knijff P, Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, *et al.* (1997). "*Chromosome Y microsatellites: population genetic and evolutionary aspects*." Int J Legal Med **110**(3): 134-149.

Decorte R (2010). "*Genetic identification in the 21st century--Current status and future developments*." Forensic Sci Int **201**(1-3): 160-164.

Didenko VV, Ngo H y Baskin DS (2003). "*Early necrotic DNA degradation: presence of blunt-ended DNA breaks, 3' and 5' overhangs in apoptosis, but only 5' overhangs in early necrosis*." Am J Pathol **162**(5): 1571-1578.

Dixon LA, Murray CM, Archer EJ, Dobbins AE, Koumi P y Gill P (2005). "*Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes*." Forensic Sci Int **154**(1): 62-77.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, *et al.* (2010). "*Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays*." Science **327**(5961): 78-81.

Earl D y vonHoldt B (2012). "*STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method*." Conserv Genet Resour **4**(2): 359-361.

Eckert KA y Hile SE (2009). "*Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome*." Mol Carcinog **48**(4): 379-388.

Egyed B, Brandstatter A, Irwin JA, Padar Z, Parsons TJ y Parson W (2007). "*Mitochondrial control region sequence variations in the Hungarian population: analysis of population samples from Hungary and from Transylvania (Romania)*." Forensic Sci Int Genet **1**(2): 158-162.

Ellegren H (2000). "*Heterogeneous mutation processes in human microsatellite DNA sequences*." Nat Genet **24**(4): 400-402.

Ellegren H (2004). "*Microsatellites: simple sequences with complex evolution*." Nat Rev Genet **5**(6): 435-445.

Ellis JA, Stebbing M y Harrap SB (2001). "*Polymorphism of the androgen receptor gene is associated with male pattern baldness*." J Invest Dermatol **116**(3): 452-455.

Elsharawy A, Forster M, Schracke N, Keller A, Thomsen I, Petersen BS, Stade B, Stahler P, Schreiber S, Rosenstiel P, *et al.* (2012). "*Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing*." BMC Genomics **13**: 417.

Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, *et al.* (2008). "*Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture*." Am J Hum Genet **82**(1): 57-72.

Enoch MA, Shen PH, Xu K, Hodgkinson C y Goldman D (2006). "*Using ancestry-informative markers to define populations and detect population stratification.*" J Psychopharmacol **20**(4 Suppl): 19-26.

Evanno G, Regnaut S y Goudet J (2005). "*Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study*." Mol Ecol **14**(8): 2611-2620.

Excoffier L y Lischer HE (2010). "*Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows*." Mol Ecol Resour **10**(3): 564-567.

Fan H y Chu J-Y (2007). "*A Brief Review of Short Tandem Repeat Mutation*." Genomics Proteomics Bioinformatics **5**(1): 7-14.

Feng D y Lazar MA (2012). "*Clocks, metabolism, and the epigenome*." Mol Cell **47**(2): 158-167.

Fleming RI y Harbison S (2010). "*The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids*." Forensic Sci Int Genet **4**(4): 244-256.

Fondevila M, Phillips C, Naverán N, Cerezo M, Rodríguez A, Calvo R, Fernández LM, Carracedo Á y Lareu MV (2008). "*Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples*." Forensic Sci Int Genet Suppl Ser **1**(1): 26-28.

Fondevila M, Phillips C, Santos C, Pereira R, Gusmao L, Carracedo A, Butler JM, Lareu MV y Vallone PM (2012). "*Forensic performance of two insertion-deletion marker assays*." Int J Legal Med **126**(5): 725-737.

Fondevila M, Phillips C, Santos C, Freire Aradas A, Vallone PM, Butler JM, Lareu MV y Carracedo A (2013). "*Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies*." Forensic Sci Int Genet **7**(1): 63-74.

Franklin RE y Gosling RG (1953). "*Molecular configuration in sodium thymonucleate*." Nature **171**(4356): 740-741.

Freire-Aradas A, Fondevila M, Kriegel AK, Phillips C, Gill P, Prieto L, Schneider PM, Carracedo Á y Lareu MV (2012). "*A new SNP assay for identification of highly degraded human DNA*." Forensic Sci Int Genet **6**(3): 341-349.

Freire-Aradas A, Ruiz Y, Phillips C, Maronas O, Sochtig J, Tato AG, Dios JA, de Cal MC, Silbiger VN, Luchessi AD, *et al.* (2014). "*Exploring iris colour prediction and ancestry inference in admixed populations of South America*." Forensic Sci Int Genet **13**: 3-9.

Freire-Aradas A, Phillips C, Mosquera-Miguel A, Giron-Santamaria L, Gomez-Tato A, Casares de Cal M, Alvarez-Dios J, Ansede-Bermejo J, Torres-Espanol M, Schneider PM, *et al.* (2016). "*Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system*." Forensic Sci Int Genet **24**: 65-74.

Freitas NS, Resque RL, Ribeiro-Rodrigues EM, Guerreiro JF, Santos NP, Ribeiro-dos-Santos A y Santos S (2010). "*X-linked insertion/deletion polymorphisms: forensic applications of a 33-markers panel*." Int J Legal Med **124**(6): 589-593.

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo JH, Koki G, Hodgson JA, *et al.* (2008). "*The genetic structure of Pacific Islanders*." PLoS Genet **4**(1): e19.

Frudakis T (2010). *Chapter 9. Direct Method of Phenotype Inference*. Molecular Photofitting: Predicting Ancestry and Phenotype Using DNA. San Diego, Academic Press**:** 497-596.

Frumkin D, Wasserstrom A, Davidson A y Grafit A (2010). "*Authentication of forensic DNA samples*." Forensic Sci Int Genet **4**(2): 95-103.

Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, Hidalgo-Miranda A, Contreras AV, Figueroa LU, Raska P, *et al.* (2012). "*Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas*." PLoS Genet **8**(3): e1002554.

Garcia-Diaz M y Kunkel TA (2006). "*Mechanism of a genetic glissando: structural biology of indel mutations*." Trends Biochem Sci **31**(4): 206-214.

Gates KS (2009). "*An overview of chemical processes that damage cellular DNA: spontaneous hydrolysis, alkylation, and reactions with radicals*." Chem Res Toxicol **22**(11): 1747-1760.

Gettings KB, Lai R, Johnson JL, Peck MA, Hart JA, Gordish-Dressman H, Schanfield MS y Podini DS (2014). "*A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population*." Forensic Sci Int Genet **8**(1): 101-108.

Gibb AJ, Huell AL, Simmons MC y Brown RM (2009). "*Characterisation of forward stutter in the AmpFlSTR SGM Plus PCR*." Sci Justice **49**(1): 24-31.

Gibbons A (2011). "*Anthropology. A new view of the birth of Homo sapiens*." Science **331**(6016): 392-394.

Gill P, Jeffreys AJ y Werrett DJ (1985). "*Forensic application of DNA fingerprints*." Nature **318**(6046): 577-579.

Gill P (2001). "*An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes*." Int J Legal Med **114**(4-5): 204-210.

Gill P, Fereday L, Morling N y Schneider PM (2006). "*The evolution of DNA databases--recommendations for new European STR loci*." Forensic Sci Int **156**(2-3): 242-244.

Gill P, Phillips C, McGovern C, Bright JA y Buckleton J (2012). "*An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees*." Forensic Sci Int Genet **6**(4): 477-486.

Gill P, Haned H, Eduardoff M, Santos C, Phillips C y Parson W (2015). "*The open-source software LRmix can be used to analyse SNP mixtures*." Forensic Sci Int Genet Suppl Ser **5**: e50-e51.

Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X y Moreno V (2007). "*SNPassoc: an R package to perform whole genome association studies*." Bioinformatics **23**(5): 644-645.

Goodsell DS (2001). "*The molecular perspective: ultraviolet light and pyrimidine dimers*." Oncologist **6**(3): 298-299.

Goodwin W, Linacre A y Hadi S (2011). *An Introduction to Forensic Genetics*. Hoboken-New Jersey, Wiley-Blackwell.

Götherström A, Collins MJ, Angerbjörn A y Lidén K (2002). "*Bone preservation and DNA amplification*." Archaeometry **44**(3): 395-404.

Grandell I, Samara R y Tillmar AO (2016). "*A SNP panel for identity and kinship testing using massive parallel sequencing*." Int J Legal Med: 1-10.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, *et al.* (2010). "*A draft sequence of the Neandertal genome*." Science **328**(5979): 710-722.

Greer CE, Peterson SL, Kiviat NB y Manos MM (1991). "*PCR amplification from paraffin-embedded tissues. Effects of fixative and fixation time*." Am J Clin Pathol **95**(2): 117-124.

Griffith F (1928). "*The Significance of Pneumococcal Types*." J Hyg (Lond) **27**(2): 113-159.

Grimes EA, Noake PJ, Dixon L y Urquhart A (2001). "*Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype*." Forensic Sci Int **122**(2-3): 124-129.

Grubwieser P, Muhlmann R, Berger B, Niederstatter H, Pavlic M y Parson W (2006). "*A new "miniSTR-multiplex" displaying reduced amplicon lengths for the analysis of degraded DNA*." Int J Legal Med **120**(2): 115-120.

Guillen M, Lareu MV, Pestoni C, Salas A y Carracedo A (2000). "*Ethical-legal problems of DNA databases in criminal investigation*." J Med Ethics **26**(4): 266-271.

Haeckel E (1866). *Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Allgemeine Anatomie der Organismen*. Berlin, Reimer.

Hallgrimsson B, Mio W, Marcucio RS y Spritz R (2014). "*Let's face it--complex traits are just not that simple*." PLoS Genet **10**(11): e1004724.

Hamblin MT y Di Rienzo A (2000). "*Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus*." Am J Hum Genet **66**(5): 1669-1679.

Hamilton JB (1951). "*Patterned loss of hair in man; types and incidence*." Ann N Y Acad Sci **53**(3): 708-728.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, *et al.* (2013). "*Genome-wide methylation profiles reveal quantitative views of human aging rates*." Mol Cell **49**(2): 359-367.

Hanson EK, Lubenow H y Ballantyne J (2009). "*Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs*." Anal Biochem **387**(2): 303-314.

Hares DR (2015). "*Selection and implementation of expanded CODIS core loci in the United States*." Forensic Sci Int Genet **17**: 33-34.

Heilmann S, Kiefer AK, Fricker N, Drichel D, Hillmer AM, Herold C, Tung JY, Eriksson N, Redler S, Betz RC, *et al.* (2013). "*Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology*." J Invest Dermatol **133**(6): 1489-1496.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G y Przeworski M (2011). "*Classic selective sweeps were rare in recent human evolution*." Science **331**(6019): 920-924.

Hershey AD y Chase M (1952). "*Independent functions of viral protein and nucleic acid in growth of bacteriophage*." J Gen Physiol **36**(1): 39-56.

Horvath S (2013). "*DNA methylation age of human tissues and cell types*." Genome Biol **14**(10): R115.

Höss M, Jaruga P, Zastawny TH, Dizdaroglu M y Pääbo S (1996). "*DNA damage and DNA sequence retrieval from ancient tissues*." Nucleic Acids Res **24**(7): 1304-1307.

Hoyle R (1998). "*Forensics. The FBI's national DNA database*." Nat Biotechnol **16**(11): 987.

Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN y Hammer MF (2016). "*Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies*." Genome Res **26**(3): 291-300.

Huang Y, Yan J, Hou J, Fu X, Li L y Hou Y (2015). "*Developing a DNA methylation assay for human age prediction in blood and bloodstain*." Forensic Sci Int Genet **17**: 129-136.

Inagaki S, Yamamoto Y, Doi Y, Takata T, Ishikawa T, Imabayashi K, Yoshitome K, Miyaishi S y Ishizu H (2004). "*A new 39-plex analysis method for SNPs including 15 blood group loci*." Forensic Sci Int **144**(1): 45-57.

Ingman M, Kaessmann H, Paabo S y Gyllensten U (2000). "*Mitochondrial genome variation and the origin of modern humans*." Nature **408**(6813): 708-713.

Jakobsson M y Rosenberg NA (2007). "*CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure*." Bioinformatics **23**(14): 1801-1806.

Jeffreys AJ, Wilson V y Thein SL (1985). "*Hypervariable /`minisatellite/' regions in human DNA*." Nature **314**(6006): 67-73.

Jobling MA, Hurles M y Tyler-Smith C (2004a). *Chapter 3 - The diversity of the human genome*. Human Evolutionary Genetics: Origins, Peoples & Disease. New York, Garland Science**:** 45-86.

Jobling MA, Hurles M y Tyler-Smith C (2004b). *Chapter 4 - Discovering and assaying genome diversity*. Human Evolutionary Genetics: Origins, Peoples & Disease. New York, Garland Science**:** 45-86.

Jobling MA, Hurles M y Tyler-Smith C (2004c). *Chapter 5 - Processes shaping diversity*. Human Evolutionary Genetics: Origins, Peoples & Disease. New York, Garland Science**:** 45-86.

Jobling MA, Hurles M y Tyler-Smith C (2004d). *Chapter 8 - Origins of modern humans*. Human Evolutionary Genetics: Origins, Peoples & Disease. New York, Garland Science**:** 45-86.

Kaiser C, Bachmeier B, Conrad C, Nerlich A, Bratzke H, Eisenmenger W y Peschel O (2008). "*Molecular study of time dependent changes in DNA stability in soil buried skeletal residues*." Forensic Sci Int **177**(1): 32-36.

Kalinowski ST (2011). "*The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure*." Heredity (Edinb) **106**(4): 625-632.

Kayser M y Schneider PM (2009). "*DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations*." Forensic Sci Int Genet **3**(3): 154-161.

Kayser M y de Knijff P (2011). "*Improving human forensics through advances in genetics, genomics and molecular biology*." Nat Rev Genet **12**(3): 179-192.

Kayser M (2015). "*Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes*." Forensic Sci Int Genet **18**: 33-48.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler y David (2002). "*The Human Genome Browser at UCSC*." Genome Res **12**(6): 996-1006.

Kersbergen P, van Duijn K, Kloosterman AD, den Dunnen JT, Kayser M y de Knijff P (2009). "*Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans*." BMC Genet **10**: 69.

Kidd J, Friedlaender F, Speed W, Pakstis A, De La Vega F y Kidd K (2011). "*Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples*." Investig Genet **2**(1): 1-13.

Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SL, Karoma NJ, Kungulilo S, Kim JJ, Lu RB, Odunsi A, *et al.* (2006). "*Developing a SNP panel for forensic identification of individuals*." Forensic Sci Int **164**(1): 20-32.

Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR y Kidd JR (2014). "*Progress toward an efficient panel of SNPs for ancestry inference*." Forensic Sci Int Genet **10**: 23-32.

King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C y Jobling MA (2007). "*Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy*." Eur J Hum Genet **15**(3): 288-293.

Koboldt DC, Ding L, Mardis ER y Wilson RK (2010). "*Challenges of sequencing human genomes*." Brief Bioinform **11**(5): 484-498.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA y Mayrose I (2015). "*Clumpak: a program for identifying clustering modes and packaging population structure inferences across K*." Mol Ecol Resour: doi: 10.1111/1755-0998.12387. [Epub ahead of print].

Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, *et al.* (2009). "*Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America*." Hum Mutat **30**(1): 69-78.

Kossel A (1886). "*Weitere Beiträge zur Chemie des Zellkerns*." Zeitschrift für Physiologische Chemie **10**: 248-264.

Krawczak M (1999). "*Informativity assessment for biallelic single nucleotide polymorphisms*." Electrophoresis **20**(8): 1676-1681.

Krawczak M, Cooper DN, Fandrich F, Engel W y Schmidtke J (2012). "*How to distinguish genetically between an alleged father and his monozygotic twin: a thought experiment*." Forensic Sci Int Genet **6**(5): e129-130.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.* (2001). "*Initial sequencing and analysis of the human genome*." Nature **409**(6822): 860-921.

Landsteiner K (1900). "*Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe*." Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten **27**: 357–362.

Lango-Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, *et al.* (2010). "*Hundreds of variants clustered in genomic loci and biological pathways affect human height*." Nature **467**(7317): 832-838.

Lao O, van Duijn K, Kersbergen P, de Knijff P y Kayser M (2006). "*Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry*." Am J Hum Genet **78**(4): 680-690.

Lao O, Vallone PM, Coble MD, Diegoli TM, van Oven M, van der Gaag KJ, Pijpe J, de Knijff P y Kayser M (2010). "*Evaluating Self-declared Ancestry of U.S. Americans with Autosomal, Y-chromosomal and Mitochondrial DNA*." Human Mutation **31**(12): E1875-E1893.

Lareu MV, García-Magariños M, Phillips C, Quintela I, Carracedo Á y Salas A (2012). "*Analysis of a claimed distant relationship in a deficient pedigree using high density SNP data*." Forensic Sci Int Genet **6**(3): 350-353.

LaRue BL, Ge J, King JL y Budowle B (2012). "*A validation study of the Qiagen Investigator DIPplex(R) kit; an INDEL-based assay for human identification*." Int J Legal Med **126**(4): 533-540.

Lasken RS y Egholm M (2003). "*Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens*." Trends Biotechnol **21**(12): 531-535.

Lazaruk K, Wallin J, Holt C, Nguyen T y Walsh PS (2001). "*Sequence variation in humans and other primates at six short tandem repeat loci used in forensic identity testing*." Forensic Sci Int **119**(1): 1-10.

Lech K, Liu F, Ackermann K, Revell VL, Lao O, Skene DJ y Kayser M (2016). "*Evaluation of mRNA markers for estimating blood deposition time: Towards alibi testing from human forensic stains with rhythmic biomarkers*." Forensic Sci Int Genet **21**: 119-125.

Lee HY, Park MJ, Yoo JE, Chung U, Han GR y Shin KJ (2005). "*Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans*." Forensic Sci Int **148**(2-3): 107-112.

Lee HY, Jung SE, Oh YN, Choi A, Yang WI y Shin KJ (2015a). "*Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study*." Forensic Sci Int Genet **19**: 28-34.

Lee HY, An JH, Jung SE, Oh YN, Lee EY, Choi A, Yang WI y Shin KJ (2015b). "*Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers*." Forensic Sci Int Genet **17**: 17-24.

Levene PA (1919). "*The structure of yeast nucleic acid*." J Biol Chem **40**(2): 415-424.

Lewontin RC (1995). "*The Apportionment of Human Diversity*." Evol Biol **6**: 381-398.

Li C, Zhang S, Que T, Li L y Zhao S "*Identical but not the same: The value of DNA methylation profiling in forensic discrimination within monozygotic twins*." Forensic Sci Int Genet Suppl Ser **3**(1): e337-e338.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, *et al.* (2008). "*Worldwide human relationships inferred from genome-wide patterns of variation*." Science **319**(5866): 1100-1104.

Li R, Brockschmidt FF, Kiefer AK, Stefansson H, Nyholt DR, Song K, Vermeulen SH, Kanoni S, Glass D, Medland SE, *et al.* (2012). "*Six novel susceptibility Loci for early-onset androgenetic alopecia and their unexpected association with common diseases*." PLoS Genet **8**(5): e1002746.

Lindahl T (1993). "*Instability and decay of the primary structure of DNA*." Nature **362**(6422): 709-715.

Lindenbergh A, van den Berge M, Oostra RJ, Cleypool C, Bruggink A, Kloosterman A y Sijen T (2013). "*Development of a mRNA profiling multiplex for the inference of organ tissues*." Int J Legal Med **127**(5): 891-900.

Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC y Kayser M (2009). "*Eye color and the prediction of complex phenotypes from genotypes*." Curr Biol **19**(5): R192-193.

Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, Wollstein A, Lao O, de Bruijne M, Ikram MA, *et al.* (2012). "*A genome-wide association study identifies five loci influencing facial morphology in Europeans*." PLoS Genet **8**(9): e1002932.

Liu F, Hendriks AE, Ralf A, Boot AM, Benyi E, Savendahl L, Oostra BA, van Duijn C, Hofman A, Rivadeneira F, *et al.* (2014). "*Common DNA variants predict tall stature in Europeans*." Hum Genet **133**(5): 587-597.

Liu F, Hamer MA, Heilmann S, Herold C, Moebus S, Hofman A, Uitterlinden AG, Nothen MM, van Duijn CM, Nijsten TE, *et al.* (2016). "*Prediction of male-pattern baldness from genotypes*." Eur J Hum Genet **24**(6): 895-902.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J y Pallen MJ (2012). "*Performance comparison of benchtop high-throughput sequencing platforms*." Nat Biotech **30**(5): 434-439.

Londin ER, Keller MA, Maista C, Smith G, Mamounas LA, Zhang R, Madore SJ, Gwinn K y Corriveau RA (2010). "*CoAIMs: a cost-effective panel of ancestry informative markers for determining continental origins*." PLoS One **5**(10): e13443.

Lou C, Cong B, Li S, Fu L, Zhang X, Feng T, Su S, Ma C, Yu F, Ye J, *et al.* (2011). "*A SNaPshot assay for genotyping 44 individual identification single nucleotide polymorphisms*." Electrophoresis **32**(3-4): 368-378.

Lowe AL, Urquhart A, Foreman LA y Evett IW (2001). "*Inferring ethnic origin by means of an STR profile*." Forensic Sci Int **119**(1): 17-22.

Mannucci A, Sullivan KM, Ivanov PL y Gill P (1994). "*Forensic application of a rapid and quantitative DNA sex test by amplification of the X-Y homologous gene amelogenin*." Int J Legal Med **106**(4): 190-193.

Manta F, Caiafa A, Pereira R, Silva D, Amorim A, Carvalho EF y Gusmao L (2012). "*Indel markers: genetic diversity of 38 polymorphisms in Brazilian populations and application in a paternity investigation with post mortem material*." Forensic Sci Int Genet **6**(5): 658-661.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, *et al.* (2005). "*Genome sequencing in microfabricated high-density picolitre reactors*." Nature **437**(7057): 376-380.

Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, de Cal MC, Ruiz Y, Fondevila M, Carracedo Á, *et al.* (2014). "*Development of a forensic skin colour predictive test*." Forensic Sci Int Genet **13**: 34-44.

Maxam AM y Gilbert W (1977). "*A new method for sequencing DNA*." Proc Natl Acad Sci U S A **74**(2): 560-564.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* (2010). "*The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*." Genome Res **20**(9): 1297-1303.

Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, Zhu G, Gordon SD, Ferreira MA, Wright MJ, Henders AK, *et al.* (2009a). "*Common variants in the trichohyalin gene are associated with straight hair in Europeans*." Am J Hum Genet **85**(5): 750-755.

Medland SE, Zhu G y Martin NG (2009b). "*Estimating the heritability of hair curliness in twins of European ancestry*." Twin Res Hum Genet **12**(5): 514-518.

Meissner C y Ritz-Timme S (2010). "*Molecular pathology and age estimation*." Forensic Sci Int **203**(1-3): 34-43.

Mengel-From J, Børsting C, Sánchez JJ, Eiberg H y Morling N (2010). "*Human eye colour and HERC2, OCA2 and MATP*." Forensic Sci Int Genet **4**(5): 323-328.

Meselson M y Stahl FW (1958). "*THE REPLICATION OF DNA IN ESCHERICHIA COLI*." Proc Natl Acad Sci U S A **44**(7): 671-682.

Miescher F (1871). " *Ueber die chemische Zusammensetzung der Eiterzellen*." Medizinisch-chemische Untersuchungen **4**: 441-460.

Mizuta R, Araki S, Furukawa M, Furukawa Y, Ebara S, Shiokawa D, Hayashi K, Tanuma S y Kitamura D (2013). "*DNase gamma is the effector endonuclease for internucleosomal DNA fragmentation in necrosis*." PLoS One **8**(12): e80223.

Moorthie S, Mattocks CJ y Wright CF (2011). "*Review of massively parallel DNA sequencing technologies*." Hugo J **5**(1-4): 1-12.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G y Erlich H (1986). "*Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.*" Cold Spring Harb Symp Quant Biol **51 Pt 1**: 263-273.

Musgrave-Brown E, Ballard D, Balogh K, Bender K, Berger B, Bogus M, Børsting C, Brion M, Fondevila M, Harrison C, *et al.* (2007). "*Forensic validation of the SNPforID 52-plex assay*." Forensic Sci Int Genet **1**(2): 186-190.

Nachman MW y Crowell SL (2000). "*Estimate of the Mutation Rate per Nucleotide in Humans*." Genetics **156**(1): 297-304.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, *et al.* (2011). "*Sequence-specific error profile of Illumina sequencers*." Nucleic Acids Res **39**(13): e90-e90.

Nakayashiki N, Takamiya M, Shimamoto K, Aoki Y y Hashiyada M (2009). "*Investigation of the methylation status around parent-of-origin detectable SNPs in imprinted genes*." Forensic Sci Int Genet **3**(4): 227-232.

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, *et al.* (2009). "*An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels*." BMC Genet **10**: 39.

Nielsen R, Paul JS, Albrechtsen A y Song YS (2011). "*Genotype and SNP calling from next-generation sequencing data*." Nat Rev Genet **12**(6): 443-451.

Nirenberg MW y Matthaei JH (1961). "*The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides*." Proc Natl Acad Sci U S A **47**: 1588-1602.

Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B y Shriver MD (2007). "*Genetic evidence for the convergent evolution of light skin in Europeans and East Asians*." Mol Biol Evol **24**(3): 710-722.

Norwood OT (1975). "*Male pattern baldness: classification and incidence*." South Med J **68**(11): 1359-1365.

Nyholt DR, Gillespie NA, Heath AC y Martin NG (2003). "*Genetic basis of male pattern baldness*." J Invest Dermatol **121**(6): 1561-1564.

O'Connor KL, Hill CR, Vallone PM y Butler JM (2011). "*Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples*." Forensic Sci Int Genet **5**(5): 538-540.

Onogi A, Nurimoto M y Morita M (2011). "*Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods*." BMC Bioinformatics **12**(1): 1-16.

Opel KL, Chung DT, Drabek J, Tatarek NE, Jantz LM y McCord BR (2006). "*The application of miniplex primer sets in the analysis of degraded DNA from human skeletal remains*." J Forensic Sci **51**(2): 351-356.

Pääbo S (2003). "*The mosaic that is our genome*." Nature **421**(6921): 409-412.

Pakstis AJ, Speed WC, Kidd JR y Kidd KK (2007). "*Candidate SNPs for a universal individual identification panel*." Hum Genet **121**(3-4): 305-317.

Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR y Kidd KK (2010). "*SNPs for a universal individual identification panel*." Hum Genet **127**(3): 315-324.

Pardo-Seco J, Martinon-Torres F y Salas A (2014). "*Evaluating the accuracy of AIM panels at quantifying genome ancestry*." BMC Genomics **15**: 543.

Park JL, Kwon OH, Kim JH, Yoo HS, Lee HC, Woo KM, Kim SY, Lee SH y Kim YS (2014a). "*Identification of body fluid-specific DNA methylation markers for use in forensic science*." Forensic Sci Int Genet **13**: 147-153.

Park JL, Park SM, Kwon OH, Lee HC, Kim JY, Seok HH, Lee WS, Lee SH, Kim YS, Woo KM, *et al.* (2014b). "*Microarray screening and qRT-PCR evaluation of microRNA markers for forensic body fluid identification*." Electrophoresis **35**(21-22): 3062-3068.

Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmao L, Hares DR, Irwin JA, King JL, *et al.* (2016). "*Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements*." Forensic Sci Int Genet **22**: 54-63.

Parsons TJ, Huel R, Davoren J, Katzmarzyk C, Milos A, Selmanovic A, Smajlovic L, Coble MD y Rizvic A (2007). "*Application of novel "mini-amplicon" STR multiplexes to high volume casework on degraded skeletal remains*." Forensic Sci Int Genet **1**(2): 175-179.

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW y Drineas P (2007). "*PCA-correlated SNPs for structure identification in worldwide human populations*." PLoS Genet **3**(9): 1672-1686.

Paternoster L, Zhurov AI, Toma AM, Kemp JP, St Pourcain B, Timpson NJ, McMahon G, McArdle W, Ring SM, Smith GD, *et al.* (2012). "*Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position*." Am J Hum Genet **90**(3): 478-485.

Pegoraro M y Tauber E (2008). "*The role of microRNAs (miRNA) in circadian rhythmicity*." J Genet **87**(5): 505-511.

Pemberton TJ, Sandefur CI, Jakobsson M y Rosenberg NA (2009). "*Sequence determinants of human microsatellite variability*." BMC Genomics **10**: 612.

Pemberton TJ, DeGiorgio M y Rosenberg NA (2013). "*Population structure in a comprehensive genomic data set on human microsatellite variation*." G3 (Bethesda) **3**(5): 891-907.

Pereira L, Alshamali F, Andreassen R, Ballard R, Chantratita W, Cho NS, Coudray C, Dugoujon JM, Espinoza M, Gonzalez-Andrade F, *et al.* (2011). "*PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile*." Int J Legal Med **125**(5): 629-636.

Pereira R, Phillips C, Alves C, Amorim A, Carracedo A y Gusmao L (2009). "*A new multiplex for human identification using insertion/deletion polymorphisms*." Electrophoresis **30**(21): 3682-3690.

Pereira R, Phillips C, Pinto N, Santos C, dos Santos SE, Amorim A, Carracedo A y Gusmao L (2012). "*Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing*." PLoS One **7**(1): e29684.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, *et al.* (2007). "*Diet and the evolution of human amylase gene copy number variation*." Nat Genet **39**(10): 1256-1260.

Pesole G, Gissi C, De Chirico A y Saccone C (1999). "*Nucleotide substitution rate of mammalian mitochondrial genomes*." J Mol Evol **48**(4): 427-434.

Phillips C, Lareu V, Salas A y Carracedo A (2004). "*Nonbinary single-nucleotide polymorphism markers*." Int Congress Ser **1261**: 27-29.

Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, *et al.* (2007). "*Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs*." Forensic Sci Int Genet **1**(3–4): 273-280.

Phillips C, Rodriguez A, Mosquera-Miguel A, Fondevila M, Porras-Hurtado L, Rondon F, Salas A, Carracedo A y Lareu MV (2008a). "*D9S1120, a simple STR with a common Native American-specific allele: forensic optimization, locus characterization and allele frequency studies*." Forensic Sci Int Genet **3**(1): 7-13.

Phillips C, Fondevila M, García-Magariños M, Rodriguez A, Salas A, Carracedo Á y Lareu MV (2008b). "*Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers*." Forensic Sci Int Genet **2**(3): 198-204.

Phillips C, Prieto L, Fondevila M, Salas A, Gómez-Tato A, Álvarez-Dios J, Alonso A, Blanco-Verea A, Brión M, Montesino M, *et al.* (2009). "*Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation*." PLoS ONE **4**(8): e6583.

Phillips C, Fernandez-Formoso L, Garcia-Magarinos M, Porras L, Tvedebrink T, Amigo J, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Freire-Aradas A, *et al.* (2011). "*Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel*." Forensic Sci Int Genet **5**(3): 155-169.

Phillips C, Ballard D, Gill P, Sydercombe-Court D, Carracedo A y Lareu MV (2012a). "*The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data*." Forensic Sci Int Genet **6**(3): 354-365.

Phillips C, García-Magariños M, Salas A, Carracedo Á y Lareu MV (2012b). "*SNPs as Supplements in Simple Kinship Analysis or as Core Markers in Distant Pairwise Relationship Tests: When Do SNPs Add Value or Replace Well-Established and Powerful STR Tests?*" Transfus Med Hemother **39**(3): 202-210.

Phillips C, Fernandez-Formoso L, Gelabert-Besada M, Garcia-Magarinos M, Santos C, Fondevila M, Carracedo A y Lareu MV (2013a). "*Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing*." Electrophoresis **34**(8): 1151-1162.

Phillips C, Freire Aradas A, Kriegel AK, Fondevila M, Bulbul O, Santos C, Serrulla Rech F, Perez Carceles MD, Carracedo A, Schneider PM, *et al.* (2013b). "*Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries*." Forensic Sci Int Genet **7**(3): 359-366.

Phillips C, Kind S, Fernandez-Formoso L, Gelabert-Besada M, Carracedo A y Lareu MV (2013c). "*Global population variability in Promega PowerPlex CS7, D6S1043, and Penta B STRs*." Int J Legal Med **127**(5): 901-906.

Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Børsting C, Johansen P, Fondevila M, *et al.* (2014a). "*Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set*." Forensic Sci Int Genet **11**: 13-25.

Phillips C, Fernandez-Formoso L, Gelabert-Besada M, Garcia-Magarinos M, Amigo J, Carracedo A y Lareu MV (2014b). "*Global population variability in Qiagen Investigator HDplex STRs*." Forensic Sci Int Genet **8**(1): 36-43.

Phillips C, Gelabert-Besada M, Fernandez-Formoso L, Garcia-Magarinos M, Santos C, Fondevila M, Ballard D, Syndercombe-Court D, Carracedo A y Lareu MV (2014c). "*"New turns from old STaRs": enhancing the capabilities of forensic short tandem repeat analysis*." Electrophoresis **35**(21-22): 3173-3187.

Phillips C (2015). "*Forensic genetic analysis of bio-geographical ancestry*." Forensic Sci Int Genet **18**: 49-65.

Phillips C, Amigo J, Carracedo Á y Lareu MV (2015). "*Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data*." Forensic Sci Int Genet **19**: 100-106.

Pickrell JK y Reich D (2014). "*Toward a new history and geography of human genes informed by ancient DNA*." Trends Genet **30**(9): 377-389.

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A y Lareu MV (2013). "*An overview of STRUCTURE: applications, parameter settings, and supporting software*." Front Genet **4**: 98.

Pospiech E, Karlowska-Pik J, Marcinska M, Abidi S, Andersen JD, van den Berge M, Carracedo A, Eduardoff M, Freire-Aradas A, Morling N, *et al.* (2015). "*Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans*." Forensic Sci Int Genet **19**: 280-288.

Pritchard JK, Stephens M y Donnelly P (2000). "*Inference of population structure using multilocus genotype data*." Genetics **155**(2): 945-959.

Prodi DA, Pirastu N, Maninchedda G, Sassu A, Picciau A, Palmas MA, Mossa A, Persico I, Adamo M, Angius A, *et al.* (2008). "*EDA2R is associated with androgenetic alopecia*." J Invest Dermatol **128**(9): 2268-2270.

Pulker H, Lareu MV, Phillips C y Carracedo A (2007). "*Finding genes that underlie physical traits of forensic interest using genetic tools*." Forensic Sci Int Genet **1**(2): 100-104.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP y Gu Y (2012). "*A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers*." BMC Genomics **13**(1): 1-13.

R Core Team (2014) "*R: A language and environment for statistical computing*." http://www.R-project.org/.

Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S y Schuster SC (2013). "*Comparison of sequencing platforms for single nucleotide variant calls in a human sample*." PLoS One **8**(2): e55089.

Real Academia Española (2014). *Forense*. Diccionario de la lengua española**:** 23.ª edición.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, *et al.* (2010). "*Genetic history of an archaic hominin group from Denisova Cave in Siberia*." Nature **468**(7327): 1053-1060.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G y Mesirov JP (2011). "*Integrative genomics viewer*." Nat Biotech **29**(1): 24-26.

Rogalla U, Rychlicka E, Derenko MV, Malyarchuk BA y Grzybowski T (2015a). "*Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples*." Forensic Sci Int Genet **14**: 42-49.

Rogalla U, Rychlicka E, Derenko MV, Malyarchuk BA y Grzybowski T (2015b). "*Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples*." Forensic Sci Int Genet **14**: 42-49.

Romanini C, Catelli ML, Borosky A, Pereira R, Romero M, Salado Puerto M, Phillips C, Fondevila M, Freire A, Santos C, *et al.* (2012). "*Typing short amplicon binary polymorphisms: supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains*." Forensic Sci Int Genet **6**(4): 469-476.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M y Nyren P (1996). "*Real-time DNA sequencing using detection of pyrophosphate release*." Anal Biochem **242**(1): 84-89.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA y Feldman MW (2002). "*Genetic structure of human populations*." Science **298**(5602): 2381-2385.

Rosenberg NA, Li LM, Ward R y Pritchard JK (2003). "*Informativeness of Genetic Markers for Inference of Ancestry*." Am J Hum Genet **73**(6): 1402-1422.

Rosenberg NA (2004). "*distruct: a program for the graphical display of population structure*." Mol Ecol Notes **4**(1): 137-138.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, *et al.* (2011). "*An integrated semiconductor device enabling non-optical genome sequencing*." Nature **475**(7356): 348-352.

Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, *et al.* (2013). "*Further development of forensic eye color predictive tests*." Forensic Sci Int Genet **7**(1): 28-40.

Saiki R, Gelfand D, Stoffel S, Scharf S, Higuchi R, Horn G, Mullis K y Erlich H (1988). "*Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase*." Science **239**(4839): 487-491.

Salas A, Phillips C y Carracedo A (2006). "*Ancestry vs physical traits: the search for ancestry informative markers (AIMs)*." Int J Legal Med **120**(3): 188-189; author reply 190.

Sánchez JJ y Endicott P (2006). "*Developing multiplexed SNP assays with special reference to degraded DNA templates*." Nat Protocols **1**(3): 1370-1378.

Sánchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, *et al.* (2006). "*A multiplex assay with 52 single nucleotide polymorphisms for human identification*." Electrophoresis **27**(9): 1713-1724.

Sanger F, Nicklen S y Coulson AR (1977). "*DNA sequencing with chain-terminating inhibitors*." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sankararaman S, Patterson N, Li H, Pääbo S y Reich D (2012). "*The date of interbreeding between Neandertals and modern humans*." PLoS Genet **8**(10): e1002947.

Santos C, Fondevila M, Ballard D, Banemann R, Bento AM, Børsting C, Branicki W, Brisighelli F, Burrington M, Capal T, *et al.* (2015). "*Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise*." Forensic Sci Int Genet **19**: 56-67.

Santos C, Phillips C, Fondevila M, Daniel R, van Oorschot RAH, Burchard EG, Schanfield MS, Souto L, Uacyisrael J, Via M, *et al.* (2016). "*Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region*." Forensic Sci Int Genet **20**: 71-80.

Santos NP, Ribeiro-Rodrigues EM, Ribeiro-Dos-Santos AK, Pereira R, Gusmao L, Amorim A, Guerreiro JF, Zago MA, Matte C, Hutz MH, *et al.* (2010). "*Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel*." Hum Mutat **31**(2): 184-190.

Schneider PM, Bender K, Mayr WR, Parson W, Hoste B, Decorte R, Cordonnier J, Vanek D, Morling N, Karjalainen M, *et al.* (2004). "*STR analysis of artificially degraded DNA-results of a collaborative European exercise*." Forensic Sci Int **139**(2-3): 123-134.

Seo SB, King JL, Warshauer DH, Davis CP, Ge J y Budowle B (2013). "*Single nucleotide polymorphism typing with massively parallel sequencing for human identification*." Int J Legal Med **127**(6): 1079-1086.

Sheehan MJ y Nachman MW (2014). "*Morphological and population genomic evidence that human faces have evolved to signal individual identity*." Nat Commun **5**: 4800.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD y Church GM (2005). "*Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome*." Science **309**(5741): 1728-1732.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R y Ferrell RE (1997). "*Ethnic-affiliation estimation by use of population-specific DNA markers*." Am J Hum Genet **60**(4): 957-964.

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM y Jones KW (2004). "*The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs*." Hum Genomics **1**(4): 274-286.

Sijen T (2015). "*Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers*." Forensic Sci Int Genet **18**: 21-32.

Sims D, Sudbery I, Ilott NE, Heger A y Ponting CP (2014). "*Sequencing depth and coverage: key considerations in genomic analyses*." Nat Rev Genet **15**(2): 121-132.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH y Hood LE (1986). "*Fluorescence detection in automated DNA sequence analysis*." Nature **321**(6071): 674-679.

Sobrino B, Brion M y Carracedo A (2005). "*SNPs in forensic genetics: a review on SNP typing methodologies*." Forensic Sci Int **154**(2-3): 181-194.

Söchtig J, Phillips C, Maroñas O, Gómez-Tato A, Cruz R, Alvarez-Dios J, Cal M-ÁC, Ruiz Y, Reich K, Fondevila M, *et al.* (2015). "*Exploration of SNP variants affecting hair colour prediction in Europeans*." Int J Legal Med **129**(5): 963-975.

Southern EM (1975). "*Detection of specific sequences among DNA fragments separated by gel electrophoresis*." J Mol Biol **98**(3): 503-517.

Spinney L (2008). "*Eyewitness identification: line-ups on trial*." Nature **453**(7194): 442-444.

Subramanian S, Mishra RK y Singh L (2003). "*Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions*." Genome Biol **4**(2): R13-R13.

Syvänen AC (1999). "*From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms*." Hum Mutat **13**(1): 1-10.

Syvänen AC (2001). "*Accessing genetic variation: genotyping single nucleotide polymorphisms*." Nat Rev Genet **2**(12): 930-942.

Szibor R, Krawczak M, Hering S, Edelmann J, Kuhlisch E y Krause D (2003). "*Use of X-linked markers for forensic purposes*." Int J Legal Med **117**(2): 67-74.

Taboada-Echalar P, Álvarez-Iglesias V, Heinz T, Vidal-Bralo L, Gómez-Carballa A, Catelli L, Pardo-Seco J, Pastoriza A, Carracedo Á, Torres-Balanza A, *et al.* (2013). "*The genetic legacy of the pre-colonial period in contemporary Bolivians*." PLoS ONE **8**(3): e58980.

Takahashi M, Kato Y, Mukoyama H, Kanaya H y Kamiyama S (1997). "*Evaluation of five polymorphic microsatellite markers for typing DNA from decomposed human tissues--correlation between the size of the alleles and that of the template DNA*." Forensic Sci Int **90**(1-2): 1-9.

The Genomes Project Consortium (2012). "*An integrated map of genetic variation from 1,092 human genomes*." Nature **491**(7422): 56-65.

The Genomes Project Consortium (2015). "*A global reference for human genetic variation*." Nature **526**(7571): 68-74.

Thomson R, Pritchard JK, Shen P, Oefner PJ y Feldman MW (2000). "*Recent common ancestry of human Y chromosomes: evidence from DNA sequence data*." Proc Natl Acad Sci U S A **97**(13): 7360-7365.

Tillmar AO y Mostad P (2014). "*Choosing supplementary markers in forensic casework*." Forensic Sci Int Genet **13**: 128-133.

Tillmar AO y Phillips C (2017). "*Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets*." Forensic Sci Int Genet **26**: 58-65.

Tilstone WJ, Savage KA y Clark LA (2006). *Forensic Science: An Encyclopedia of History, Methods, and Techniques*. Santa Barbara - California, ABC-CLIO.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, *et al.* (2001). "*Haplotype diversity*

and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance." Science **293**(5529): 455-462.

Tomas C, Sánchez JJ, Castro JA, Børsting C y Morling N (2010). "*Forensic usefulness of a 25 X-chromosome single-nucleotide polymorphism marker set*." Transfusion **50**(10): 2258-2265.

Trindade-Filho A, Ferreira S y Oliveira SF (2013). "*Impact of a chromosome X STR Decaplex in deficiency paternity cases*." Genet Mol Biol **36**(4): 507-510.

van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JF y de Knijff P (2016). "*Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system*." Forensic Sci Int Genet **24**: 86-96.

Van Neste C, Van Nieuwerburgh F, Van Hoofstat D y Deforce D (2012). "*Forensic STR analysis using massive parallel sequencing*." Forensic Sci Int Genet **6**(6): 810-818.

van Oorschot RA, Ballantyne KN y Mitchell RJ (2010). "*Forensic trace DNA: a review*." Investig Genet **1**(1): 14.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.* (2001). "*The Sequence of the Human Genome*." Science **291**(5507): 1304-1351.

Vidaki A, Daniel B y Syndercombe-Court D (2013). "*Forensic DNA methylation profiling--potential opportunities and challenges*." Forensic Sci Int Genet **7**(5): 499-507.

Vigilant L, Pennington R, Harpending H, Kocher TD y Wilson AC (1989). "*Mitochondrial DNA sequences in single hairs from a southern African population*." Proc Natl Acad Sci U S A **86**(23): 9350-9354.

Voelkerding KV, Dames SA y Durtschi JD (2009). "*Next-Generation Sequencing: From Basic Research to Diagnostics*." Clin Chem **55**(4): 641-658.

von Wurmb-Schwark N, Malyusz V, Simeoni E, Lignitz E y Poetsch M (2006). "*Possible pitfalls in motherless paternity analysis with related putative fathers*." Forensic Sci Int **159**(2-3): 92-97.

Walsh PS, Fildes NJ y Reynolds R (1996). "*Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA*." Nucleic Acids Res **24**(14): 2807-2812.

Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M y Ballantyne KN (2011a). "*Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence*." Forensic Sci Int Genet **5**(5): 464-471.

Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O y Kayser M (2011b). "*IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information*." Forensic Sci Int Genet **5**(3): 170-180.

Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W y Kayser M (2013). "*The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA*." Forensic Sci Int Genet **7**(1): 98-115.

Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, Maeda H, Ishikawa T, Sijen T, de Knijff P, *et al.* (2014). "*Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage*." Forensic Sci Int Genet **9**: 150-161.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, *et al.* (2007). "*Genetic variation and population structure in native Americans*." PLoS Genet **3**(11): e185.

Watson JD y Crick FH (1953). "*Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*." Nature **171**(4356): 737-738.

Weber JL, David D, Heil J, Fan Y, Zhao C y Marth G (2002). "*Human Diallelic Insertion/Deletion Polymorphisms*." Am J Hum Genet **71**(4): 854-862.

Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J y Rolf B (2014). "*Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing*." Forensic Sci Int Genet **9**: 42-46.

Wei Y-L, Wei L, Zhao L, Sun Q-F, Jiang L, Zhang T, Liu H-B, Chen J-G, Ye J, Hu L, *et al.* (2015). "*A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents*." Int J Legal Med: 1-11.

Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, Bauerschlag DO, Jockel KH, Erbel R, Muhleisen TW, *et al.* (2014). "*Aging of blood can be tracked by DNA methylation changes at just three CpG sites*." Genome Biol **15**(2): R24.

Welch L, Gill P, Tucker VC, Schneider PM, Parson W, Mogensen HS y Morling N (2011). "*A comparison of mini-STRs versus standard STRs--results of a collaborative European (EDNAP) exercise*." Forensic Sci Int Genet **5**(3): 257-258.

Werrett DJ (1997). "*The National DNA Database*." Forensic Sci Int **88**(1): 33-42.

Westen AA, Matai AS, Laros JF, Meiland HC, Jasper M, de Leeuw WJ, de Knijff P y Sijen T (2009). "*Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples*." Forensic Sci Int Genet **3**(4): 233-241.

Westen AA, Grol LJ, Harteveld J, Matai AS, de Knijff P y Sijen T (2012). "*Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM*." Forensic Sci Int Genet **6**(6): 708-715.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, *et al.* (2008). "*The complete genome of an individual by massively parallel DNA sequencing*." Nature **452**(7189): 872-876.

Whitaker JP, Clayton TM, Urquhart AJ, Millican ES, Downes TJ, Kimpton CP y Gill P (1995). "*Short tandem repeat typing of bodies from a mass disaster: high success rate and characteristic amplification patterns in highly degraded samples*." Biotechniques **18**(4): 670-677.

Wiegand P y Kleiber M (2001). "*Less is more--length reduction of STR amplicons using redesigned primers*." Int J Legal Med **114**(4-5): 285-287.

Wilkins MH, Stokes AR y Wilson HR (1953). "*Molecular structure of deoxypentose nucleic acids*." Nature **171**(4356): 738-740.

Wong Z, Wilson V, Patel I, Povey S y Jeffreys AJ (1987). "*Characterization of a panel of highly variable minisatellites cloned from human DNA*." Ann Hum Genet **51**(Pt 4): 269-288.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, *et al.* (2014). "*Defining the role of common variation in the genomic and biological architecture of adult human height*." Nat Genet **46**(11): 1173-1186.

Wright S (1951). "*The genetical structure of populations*." Ann Eugen **15**(4): 323-354.

Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, Kittles R, Shigeta R, Silva G, Patel PI, Belmont JW, *et al.* (2005). "*Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine*." Hum Genet **118**(3-4): 382-392.

Zaumsegel D, Rothschild MA y Schneider PM (2013). "*A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry*." Forensic Sci Int Genet **7**(2): 305-312.

Zhang Z y Gerstein M (2003). "*Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes*." Nucleic Acids Res **31**(18): 5338-5348.

Zubakov D, Liu F, van Zelm MC, Vermeulen J, Oostra BA, van Duijn CM, Driessen GJ, van Dongen JJ, Kayser M y Langerak AW (2010). "*Estimating human age from T-cell DNA rearrangements*." Curr Biol **20**(22): R970-971.

Zubakov D, Kokmeijer I, Ralf A, Rajagopalan N, Calandro L, Wootton S, Langit R, Chang C, Lagace R y Kayser M (2015). "*Towards simultaneous individual and tissue identification: A proof-of-principle study on parallel sequencing of STRs, amelogenin, and mRNAs with the Ion Torrent PGM*." Forensic Sci Int Genet **17**: 122-128.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA