ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

GRADO EN BIOLOGÍA

TRABAJO DE FIN DE GRADO

"Inserciones *Alu* y heterogeneidad genética de la población gitana del País Vasco".

Autor: Iñigo Marcos Sarobe

Director: Jose Angel Peña García

Leioa, Julio 2015

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

1. RESUMEN

En este trabajo se ha analizado un grupo de 6 inserciones *Alu* autosómicas (*ACE, APO, PV92, TPA25, FXIIIB* y *D1*) en una muestra de 56 individuos de etnia gitana residentes en el País Vasco, con el objetivo de estimar la intensidad de los procesos de microdiferenciación experimentados por esta población y su parentesco genético con otras poblaciones europeas y asiáticas.

Las inserciones *Alu* polimórficas son unos marcadores muy útiles en los estudios de evolución humana, entre otras razones porque se conoce su estado ancestral, que es la ausencia de inserción y porque se producen por un único evento mutacional. Son por ello particularmente interesantes para analizar la heterogeneidad genética de poblaciones originarias de diferentes continentes.

A partir de diferentes referencias bibliográficas se ha elaborado una base de datos de frecuencias alélicas de 95 poblaciones de diferentes continentes. Se ha analizado la heterogeneidad genética por continentes y se han seleccionado para un análisis pormenorizado todas las poblaciones que se encontraban en un transecto que incluye desde la Península Ibérica, lugar de residencia de los gitanos vascos, hasta la India, su presumible lugar de origen.

Se han detectado clinas significativas en 5 de las 6 inserciones, que describen la variación de las frecuencias alélicas a lo largo de este transecto. Utilizando este patrón de variación como referencia, se ha observado que es la deriva el factor que más ha afectado al patrimonio genético de los gitanos vascos, encontrándose una escasa incidencia del mestizaje.

2. ABSTRACT

In this paper, a group of 6 autosomal *Alu* insertions has been analyzed in 56 gypsies living in the Basque Country, in order to estimate the intensity of the microdifferentiation processes experienced by this population and its genetic relationship to other european and asian populations.

Polymorphic *Alu* insertions are robust markers for human evolutionary studies because of the knowledge of its ancient state and the fact that they have an unique mutational mechanism. Therefore they are partyculary interesting for analyze the genetic diversity of populations originated in different continents.

Using different references, a database of allelic frequencies of 95 populations from different continents has been created. The genetic diversity has been analyzed by continent and after that it has been selected all the population included in a transect ranging from the Iberian Peninsula, place of residence of the Basque Gypsies to India, its presumed place of origin.

It has been detected significant clines in 5 of the 6 insertions, which describe the variation of allele frequencies along the transect. Using this variation pattern as reference, it has been found that the genetic drift is the factor with the most impact on the genetic heritage of the Basques gypsies, finding a low impact of admixture.

3. INTRODUCCIÓN

El desarrollo de herramientas y métodos analíticos cada vez más robustos y fiables ha permitido establecer con precisión la proporción de variabilidad intrapoblacional e interpoblacional existente en nuestra especie, así como las potenciales causas de los cambios microevolutivos ocurridos en poblaciones con diferentes orígenes geográficos y socioculturales. De esta manera, los resultados de numerosos estudios antropogenéticos coinciden en que la especie humana es genéticamente homogénea, de forma tal que alrededor del 85-90% de la variabilidad de su actual patrimonio genético es intrapoblacional, y solamente entre 10-15% se puede atribuir a diferencias entre grupos continentales y/o étnicos (Jorde & Wooding, 2004; Shriver & Kittles, 2004). Por ello, cuando el objetivo de un trabajo trata de investigar acerca de la heterogeneidad genética entre poblaciones, es vital seleccionar marcadores que sean capaces de reflejar ese pequeño porcentaje de variabilidad interpoblacional de forma correcta.

Los trabajos que abordan tanto la diversidad genética como las relaciones filogenéticas entre poblaciones humanas a menudo se han llevado a cabo mediante el estudio de marcadores de transmisión uniparental. Es bien conocido que un cierto número de marcadores genéticos ubicados en la región no recombinante del cromosoma Y aportan información específica sobre los patrones de migración de los linajes masculinos. De forma complementaria, los polimorfismos del genoma mitocondrial o mtDNA documentan la historia evolutiva de los linajes maternos de nuestra especie (Lell & Wallace, 2000; Jobling & Tyler-Smith, 2003). Sin embargo, la reconstrucción de la historia de las poblaciones humanas es una cuestión compleja, de modo que un abordaje adecuado de este asunto requeriría del contraste de la información obtenida a partir de diferentes fuentes complementarias de datos. Por esta razón, son también frecuentes los trabajos de genética de poblaciones humanas donde se analiza la variabilidad del DNA nuclear recombinante, es decir, del DNA autosómico (Lell & Wallace, 2000; Jobling & Tyler-Smith, 2003). En opinión de muchos autores, los marcadores autosómicos proporcionan datos muy esclarecedores acerca de la evolución conjunta de los linajes materno y paterno (Jorde *et al*., 1995; Lell & Wallace, 2000; Richards, 2003).

En este trabajo se han utilizado 6 inserciones *Alu* autosómicas como marcadores de los procesos microevolutivos experimentados por la población gitana vasca y su relación con otras poblaciones próximas a ella o a su presumible lugar de origen.

Los elementos *Alu* fueron descubiertos en el año 1979 como un componente de las curvas de renaturalización del DNA humano. Su nombre proviene del hecho de que contiene una secuencia diana tetranucleotídica AGCT para la enzima de restricción *Alu I* (Houck *et al*., 1979). Las inserciones *Alu* componen la mayor familia de elementos móviles del genoma humano (Batzer & Deininger, 2002). Este, como ocurre en el resto de los organismos eucariotas, se divide en dos tipos: el genoma nuclear y el mitocondrial (

Figura 1). En el DNA nuclear se identifica un tipo de secuencias denominadas secuencias no codificantes, que no generan producto genético alguno, lo que no significa que no realicen alguna función (ENCODE Project Consortium, 2012). Asimismo, dentro de estas regiones no codificantes se disciernen las secuencias repetidas, que se pueden encontrar en forma dispersa en el genoma. Dentro

del grupo de secuencias repetidas dispersas se encuentran los *SINEs* (*Short INterspesed Elements*), donde están agrupadas todas las inserciones menores de 500pb. Los *SINEs* más abundantes son las inserciones *Alu*.

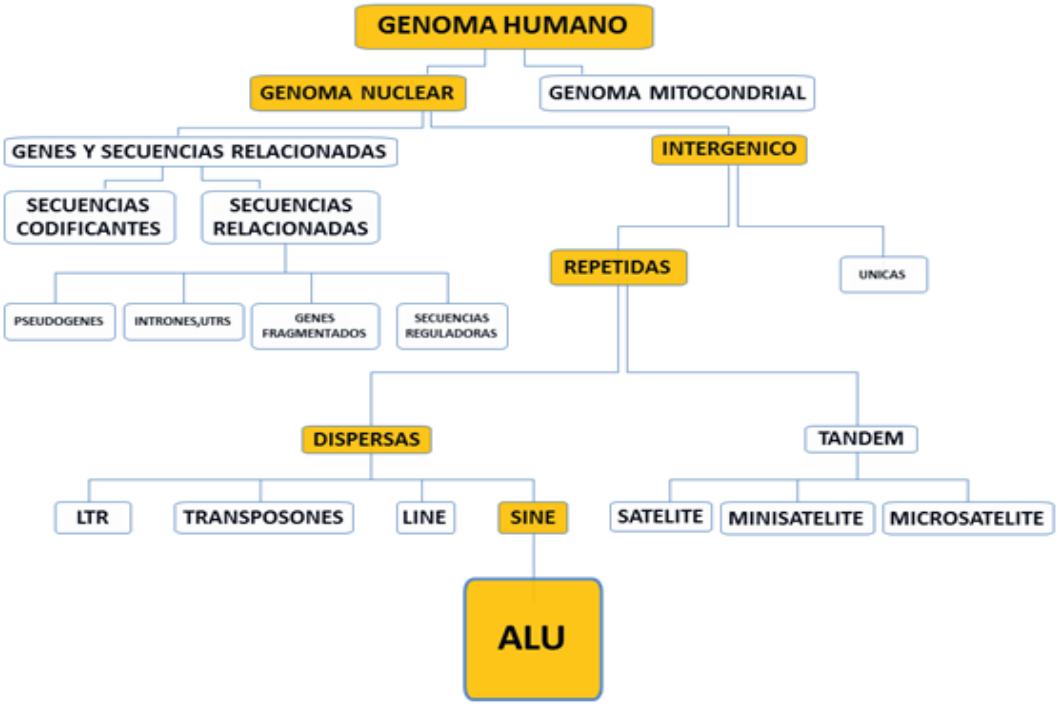

Figura 1: Clasificación de los elementos Alu y otras secuencias del genoma humano.

La estructura de las inserciones *Alu* es dimérica (Figura 2), con una longitud de aproximadamente 300pb que se reparte entre dos unidades o brazos similares (aunque no idénticos), terminados ambos en una secuencia de adeninas (dA) (Weiner *et al*., 1986). El brazo derecho (extremo 3´) contiene una inserción de 31pb que está ausente en el brazo izquierdo (extremo 5´). Sin embargo, es el brazo izquierdo el que contiene secuencias con características funcionales. Así, por ejemplo, en este brazo se encuentra el sitio promotor de la *RNA polimerasa III tipo 2*, en dos regiones denominadas *Caja A* y *Caja B*, cada una de ellas con una longitud de 10pb y situadas en la regiones 10-25 y 70-90, respectivamente (Fuhrman *et al*., 1981; Schimid & Shen, 1985; Quentin, 1992; Knight *et al*., 1996; Novick *et al*., 1996; Rowold & Herrera, 2000). Los elementos *Alu* propiamente dichos no poseen la secuencia terminal de cuatro timinas que actúa como señal de terminación de la *RNA polimerasa III*; sin embargo, esta señal a veces está presente aguas abajo (del inglés *down stream*) de la secuencia *Alu*.

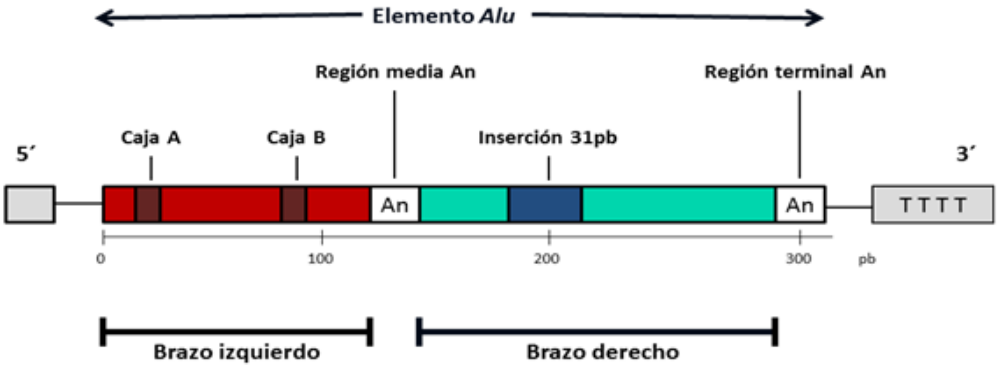

Figura 2: Estructura dimérica del elemento Alu (modificado de Häsler y Strub, 2006).

El origen y posterior diversificación de los elementos *Alu* comprende dos fases bien diferenciadas:

1. Una fase monomérica, que se inició durante la radiación de los mamíferos (Labuda & Zietkiewicz, 1994; Roy-Engel *et al*., 2008), con la aparición del elemento *Alu* ancestral también denominado monómero fósil *Alu* o ***FAM*** (del inglés *Fossil Alu Monomer*) y que concluye con el origen de las secuencias progenitoras de la familia *Alu* diméricas. El monómero fósil *Alu* se origina a partir de una deleción de 141pb (entre la posición 97 y 239) en el dominio central *S* del gen *RNA 7SL*, y la adición en el extremo 3´de una cola de *poli-A*, la cual se piensa que podría facilitar la transcripción inversa de los transcritos de la *RNA polimerasa III*. La diversificación y consecuente evolución de la familia *FAM* originó las familias de monómeros ***FLAM*** (del inglés *Free Left Alu Monomer*) y ***FRAM*** (del inglés *Free Right Alu Monomer*).
2. Una fase dimérica, evolutivamente más reciente, que coincide con la radiación de los primates hace 65 millones de años (Deininger & Daniels, 1986). El progenitor de la familia de los *Alu* diméricos posiblemente fue consecuencia de la fusión de los monómeros *FLAM* y *FRAM*. Los elementos *Alu* diméricos son, por tanto, característicos del Orden Primates.

A partir de la formación del primer *Alu* dimérico tuvo lugar una amplificación masiva, aunque restringida en el tiempo, de retrotransposones *Alu* dentro del genoma humano. Por el contrario, tanto el gen *ARN 7SL* como los *Alu* monoméricos tuvieron una amplificación limitada. La conservación del gen *ARN 7SL* a lo largo de todas las especies sugiere que los precursores monoméricos de las secuencias *Alu* estaban en todos los linajes (Mighell *et al*., 1997).

Las inserciones *Alu* poseen varias características que las hacen idóneas para estudios genético-evolutivos en poblaciones humanas, entre las que cabe destacar, que en estos marcadores es perfectamente posible conocer cuál es la variante original o el estado ancestral, ya que se trata de la ausencia de inserción (Batzer y Deininger, 2002; Salem et al., 2005).

Otra de las características interesantes es su prácticamente nula tasa de inserción de novo. Las inserciones son por tanto originadas por un evento único, de modo que los individuos que comparten una inserción la heredaron de un ancestro común y pueden considerarse alelos idénticos por

descendencia. Esto facilita la verosimilitud de las estimaciones del parentesco entre poblaciones y una discriminación eficaz entre poblaciones de diferentes continentes.

El hecho de que muchas de estas inserciones hayan aparecido en el último millón de años hace que su presencia no necesariamente sea universal, de modo que una parte considerable de ellas se mantenga en estado polimórfico y puedan utilizarse para las comparaciones entre poblaciones, como las seis inserciones utilizadas en este trabajo.

El término **gitanos o personas romaníes** hace referencia a un grupo étnico muy disperso que no tiene ningún estado o nación, habla diferentes idiomas, pertenece a muchas religiones y socio-culturalmente comprende un mosaico de poblaciones divergentes entre sí y separadas de las poblaciones circundantes por estrictas reglas de endogamia (Kalaydjieva *et al.*, 2005).

La ausencia de una historia bien documentada supone que los orígenes de la población romaní sean un tema de debate. Probablemente entraron por el sudeste en Europa, provenientes de Medio Oriente con posterioridad al año 1000 (Gresham et al. 2001) (figura 3). Algunos se denominaron a sí mismos como "nobles egipcios", debido a lo cual en el siglo XV entre sus denominaciones estaban las de egipcios y giptanos, de donde probablemente derivan los términos Gypsy, Gitano e Ijito entre otros (Hancock, 1993). Sin embargo, las evidencias lingüísticas y genéticas indican que la población gitana se originó en la India, posiblemente, en algún punto entre el Norte de la India y Pakistán, incluyendo el estado indio de Rajastán y la región del Punjab (Mendizabal et al. 2011, 2012). En la Península Ibérica habrían penetrado hacia el año 1425 (Lermo et al, 2006).

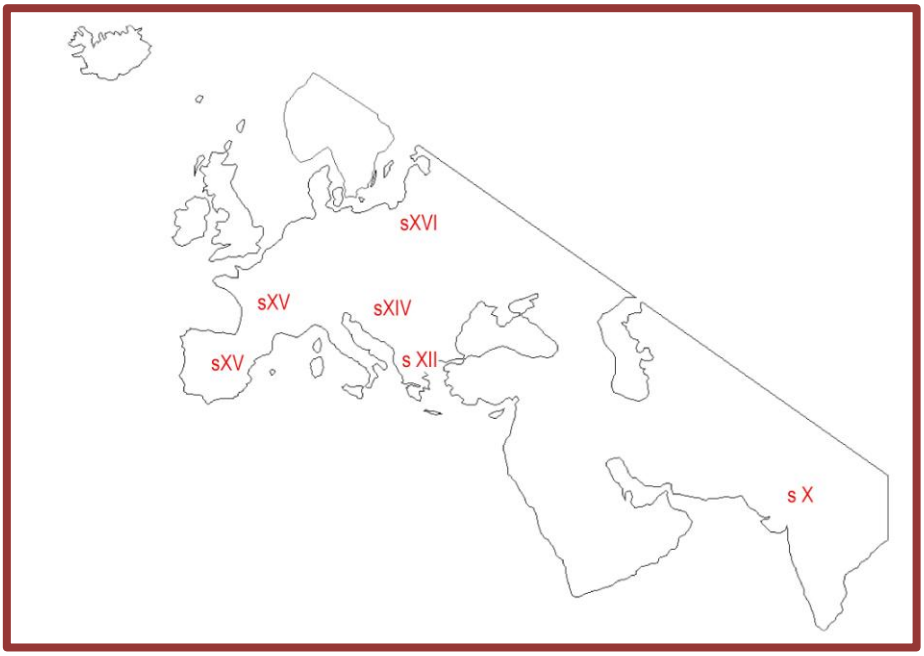

Figura 3: *Principales hitos de la migración de los gitanos.*

El origen indio de la lengua romaní fue propuesto por Samuel Augustini Hortis en su obra "*Zigeuner en Ungaren*" a finales del siglo XVIII. Actualmente, hay un cierto consenso acerca de que

el romaní se originó en el noroeste de la India, pudiendo estar emparentado con el Sánscrito (Matras, 2006).

No hay recuento fiable del tamaño de la población gitana, en parte porque los individuos de esta etnia a menudo evitan su registro en los censos oficiales y por otra parte porque en muchos países como Bélgica, Alemania, Finlandia o Suecia está prohibido recoger información étnica en las estadísticas oficiales. Sin embargo, la actual población romaní de Europa se estima en 10 millones de personas (Comisión Europea, 2004), siendo una de las mayores comunidades ubicada en el centro y oriente de Europa. Según recientes estimaciones, Bulgaria, Hungría y Rumania son los estados miembros de la Unión Europea con mayor número de personas gitanas (Corsi et al., 2008). La población gitana en España está dentro de un rango de entre 650.000 y 750.000 habitantes, con una elevada concentración en la región de Andalucía (FSG, 2008).

4. OBJETIVOS

Con el desarrollo de este trabajo se ha pretendido valorar la heterogeneidad genética de la población gitana del País Vasco en relación a otras poblaciones de su entorno geográfico más próximo, así como de las regiones de su supuesto origen, en el norte de la India. Con ello se pretende obtener información acerca de los procesos de microdiferenciación experimentados por esta población a lo largo de su historia.

5. MATERIAL Y MÉTODOS

5.1. Población de estudio

En este trabajo se ha analizado una muestra de 56 individuos representativos de la población de etnia gitana residente en el País Vasco. Todos ellos eran individuos voluntarios informados y sanos. No se encontraban emparentados hasta donde pudo deducirse por las encuestas realizadas. Además eran residentes estables desde al menos 3 generaciones. Antes de comenzar este trabajo se habían realizado, por otros miembros del equipo, los consentimientos informados, las tomas de muestras, encuestas y las extracciones de ADN, de acuerdo a la legislación vigente. Las muestras que se facilitaron estaban codificadas y tratadas de acuerdo con la Ley Orgánica 15/1999 de protección de datos de carácter personal.

5.2. Descripción de los marcadores utilizados

Sobre las 56 muestras de ADN se analizaron 6 inserciones *Alu* ubicadas en regiones no codificantes y situadas en diferentes cromosomas (tabla 1). De entre las numerosas inserciones polimórficas del genoma humano, se seleccionaron las 6 que se describen en la tabla 1 por ser las más analizadas en diferentes poblaciones y más específicamente por encontrarse analizadas en un gran número de poblaciones de Medio Oriente, particularmente de La India.

Tabla 1: Ubicación cromosómica (UC), subfamilia (Sf), código NCBI, secuencia de cebadores, tamaño de los fragmentos PCR (TF) y referencia bibliográfica de los elementos Alu analizados.

Locus	UC	Sf	NCBI	Secuencia de cebadores (5´- 3´)	TF (pb)	Referencia
ACE	17q23	Ya5	rs4646994	F: CTGGAGACCACTCCCATCCTTTCA R: GATGTGGCCATCACATTCGTCAGAT	*Alu (+):*490 *Alu (-):*190	Tiret et al., (1992) Batzer et al., (1996ab)
TPA25	8p11.2	Ya8	rs4646972	F:GTAAGAGTTCCGTAACAGGACAGCT R:CCCCACCCTAGGAGAACTTCTCTTT	*Alu (+):*424 *Alu (-):*113	Batzer et al., (1991, 1996ab)
PV92	16q23.3	Ya5 / Ya8	AF302689	F:AACTGGGAAAATTTGAAGAGAAAGT R:TGAGTTCTCAACTCCTGTGTGTTAG	*Alu (++):*789 *Alu (+):*443 *Alu (-):*129	Batzer et al., (1994, 1996ab) Comas et al., (2001)
APO	11q23.3	Ya5	rs3138522	F:AAGTGCTGTAGGCCATTTAGATTAG R:AGTCTTCGATGACAGCGTATACAGA	*Alu (+):*409 *Alu (-):*97	Batzer et al., (1994, 1996ab)
FXIIIB	1q31-q32.1	Ya5	AL353809.20	F:TCAACTCCATGAGATTTTCAGAAGT R:CTGGAAAAAATGTATTCAGGTGAGT	*Alu (+):*700 *Alu (-):*410	Kass et al., (1994) Batzer et al., (1996ab)
D1	3q26.32	Yb8	U12583.1	F:TGCTGATGCCCAGGGTTAGTAAA R:TTTCTGCTATGCTCTTCCCTCTC	*Alu (+):*670 *Alu (-):*333	Batzer et al., (1995) Arcot et al., (1995b)

5.3. Identificación de los genotipos

Con el fin de amplificar las regiones del genoma correspondientes a cada inserción, se han llevado a cabo una serie de reacciones en cadena de la polimerasa (en inglés *Polymerase Chain Reaction*, ***PCR***). En todos los casos se realizaron en un volumen final de 10µl con una cantidad de entre 50 y 100ng de DNA. Las concentraciones de los diferentes componentes fueron modificadas de las referencias originales mostradas en la tabla 1, con el fin de adaptarlas a las condiciones específicas de equipamiento y reactivos (tabla 2). Las reacciones *PCR* se efectuaron en un termociclador *Gene Amp PCR System 9700*, de Applied Biosystems. Las condiciones de los diferentes ciclos para cada inserción se muestran en la tabla 3.

Tabla 2: Concentraciones finales de los reactivos empleados en las reacciones PCR.

Reactivo	Marcador analizado					
	***Alu* ACE**	***Alu* TPA25**	***Alu* PV92**	***Alu* APO**	***Alu* FXIIB**	***Alu* D1**
PCR Buffer	1X	1X	1X	1X	1X	1,3X
MgCl2	3 mM	1,25 mM	1,5 mM	1,5 mM	1,5 mM	2,25 mM
dNTPs	1 mM	1 mM	1 mM	1,5 mM	1,4 mM	1,5 mM
Glicerol	0,5%	-	-	-	-	-
Primer A	0,45 µM	0,5 µM	0,45 µM	0,6 µM	0,6 µM	0,5 µM
Primer B	0,45 µM	0,5 µM	0,45 µM	0,6 µM	0,6 µM	0,5 µM
Taq Pol	0,1 U/µl	0,04 U/µl	0,05 U/µl	0,05 U/µl	0,045 U/µl	0,06 U/µl
ADN	1,4 ng/µl	1,3 ng/µl	1,4 ng/µl	1,4 ng/µl	1,1 ng/µl	1,9 ng/µl

Tabla 3: Características de los ciclos PCR para cada marcador.

Programa PCR		Marcador analizado					
		***Alu* ACE**	***Alu* TPA25**	***Alu* PV92**	***Alu* APO**	***Alu* FXIIB**	***Alu* D1**
Desnaturalización Inicial		94ºC 10 min	94ºC 10 min	94ºC 10 min	94ºC 10 min	94ºC 10 min	94ºC 10 min
Ciclos previos	Desnaturalización	-	94ºC 1 min	-	-	94ºC 1 min	94ºC 0,5 min
	Anneling	-	65ºC 1,5 min	-	-	60ºC 1,5 min	59,5ºC 0,75 min
	Extensión	-	72ºC 1,5 min	-	-	72ºC 1,5 min	72ºC 1,5 min
	Numero de ciclos	-	7	-	-	7	7
Desnaturalización		94ºC 1 min	94ºC 1 min	94ºC 1 min	94ºC 1 min	94ºC 1 min	94ºC 1 min
Anneling		55,5ºC 1,5 min	63ºC 1 min	57,8ºC 1,5 min	47ºC 1,5 min	58ºC 1 min	56,5ºC 1,5 min
Extensión		72ºC 1,5 min	72ºC 1 min	72ºC 1,5 min	72ºC 1,5 min	72ºC 1 min	72ºC 2 min
Numero de ciclos		30	25	30	30	25	25
Extensión Final		72ºC 10 min	72ºC 10 min	72ºC 10 min	72ºC 10 min	72ºC 10 min	72ºC 10 min

Los amplificados PCR fueron sometidos a electroforesis en gel de agarosa 1,5%, a 115 voltios durante 20 minutos, utilizando como colorante Real Safe, de Durviz, siguiendo las instrucciones del fabricante. Los resultados de las electroforesis fueron visualizados bajo luz ultravioleta (*GelVue transiluminador UV - Syngene GVM20)* y finalmente documentados y conservados como imagen digital mediante un sistema *Kodak Digital Science DC120*. En los ensayos se incluyeron controles positivos (homocigotos conocidos para la inserción) y controles negativos, los cuales contenían todos los reactivos de la PCR excepto DNA, con el objetivo de controlar la calidad de la reacción PCR y de la electroforesis. Además, en el gel de agarosa se añadió un marcador de tamaño del DNA, *DNA Molecular Weight Marker V* de Roche, al objeto de identificar el tamaño de las bandas PCR.

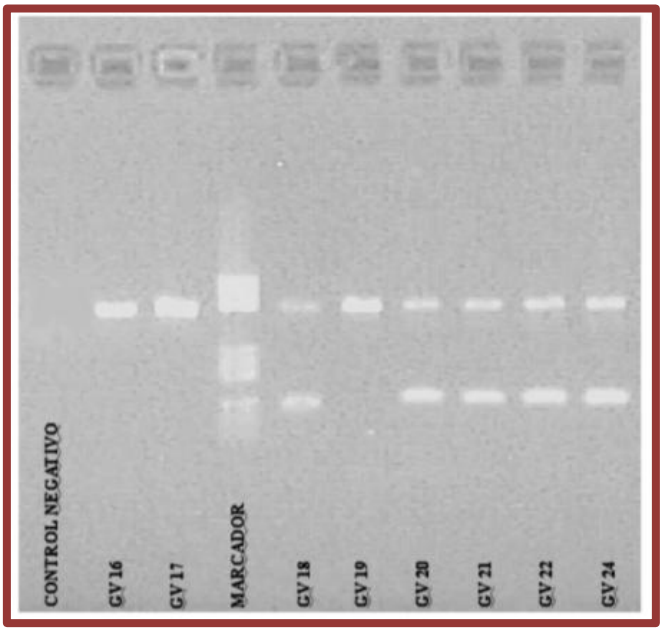

Ilustración 3: Imagen de una electroforesis en gel de agarosa donde se muestran fragmentos amplificados del elemento Alu TPA25. Los individuos heterocigotos presentan dos bandas (inserción y no inserción) mientras que los individuos homocigotos presentan una única banda (inserción o no inserción). En la cuarta posición se observa el marcador de tamaño del DNA.

5.4. Tratamiento estadístico

Se han utilizado una serie de métodos y programas estadísticos, con el fin de facilitar la interpretación de los resultados.

5.4.1. Estimación de frecuencias alélicas, diversidad génica y contenido de información polimórfica

Las frecuencias alélicas y genotípicas para cada uno de los marcadores se calcularon por el método del conteo directo (Nei, 1987), según los cocientes:

$$\textbf{Frecuencia alélica} = n_a / 2N$$

$$\textbf{Frecuencia genotípica} = n_g / N$$

donde **N** = tamaño de la muestra; $\mathbf{n_a}$= frecuencia observada del alelo; $\mathbf{n_g}$ = frecuencia observada del genotipo.

El error estándar (SE) de las frecuencias alélicas se ha obtenido a partir de la formula (Li, 1968):

$$\textbf{Error estándar} = \pm\sqrt{p_i\,(1 - p_i)/2N}$$

donde $\mathbf{p_i}$ es el valor de la frecuencia del i-ésimo alelo y **N** es el tamaño de la muestra.

La diversidad genética (***GD***, del inglés *Gene Diversity*), también conocida como la heterocigosidad esperada (h), representa la probabilidad de que dos alelos o haplotípos tomados al azar en una muestra sean diferentes. Es una de las medidas de variabilidad genética más comúnmente

empleadas y, por tanto, una de las más útiles para análisis comparativos entre muestras poblacionales de distintas zonas geográficas, diferentes grupos étnicos, troncos lingüísticos, etc. (Nei, 1987).

El contenido de información polimórfica (PIC, del inglés *Polimorphism Information Content*) es uno de los diversos parámetros por el cual se puede cuantificar el grado de polimorfismo de un determinado marcador genético. El valor PIC depende esencialmente del número de alelos del locus examinado y de las frecuencias registradas en la población de estudio, como puede apreciarse en la fórmula:

$$\mathbf{h} = \mathbf{1} - \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{p_i}^{\mathbf{2}} \right)$$

$$\mathbf{PIC} = \mathbf{1} - \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p_i}^{\mathbf{2}} \right) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n-1}} \cdot \sum_{\boldsymbol{j=i+1}}^{\boldsymbol{n}} \mathbf{2}\boldsymbol{p_i}^{\mathbf{2}}\boldsymbol{p_i}^{\mathbf{2}}$$

donde **h** es la heterocigosidad esperada por locus (cuya expresión matemática se corresponde con la diversidad genética de Nei (1987), **n** es el número de alelos para cada marcador genético y $\mathbf{p_i}$, $\mathbf{p_j}$ la frecuencia de los alelos i-ésimo y j-ésimo en la población estudiada.

Estos parámetros fueron estimados mediante el programa strs2 (Peña, comunicación personal).

5.4.2. Equilibrio de Hardy-Weinberg

El principio del equilibrio genético, postulado de forma independiente por Hardy y Weinberg establece que, bajo ciertas condiciones, las frecuencias génicas y genotípicas poblacionales, así como los caracteres que ellos determinan, se mantienen constantes de generación en generación (Hardy, 1908; Weinberg, 1908).

Cuando todas las premisas necesarias para que una población pueda alcanzar el equilibrio genético se cumplen, entonces las frecuencias genotípicas esperadas se podrán calcular como:

Frecuencia esperada de homocigotos = p^2 y q^2

Frecuencia esperada de heterocigotos = 2pq

donde **p + q = 1.**

Con el fin de evaluar la existencia de equilibrio Hardy-Weinberg en nuestra población, se utilizó el programa Arlequin 3.5 (Excoffier y Lischer, 2010).

5.4.3. Distancia genética entre poblaciones

Si se asume que cuanto más parecidas resulten ser las frecuencias genéticas de dos poblaciones, el grado de microdiferenciación entre ellas será menor, puede utilizarse un coeficiente que combinando las frecuencias alélicas de distintos *loci* de lugar a una estimación de la distancia genética que existe entre dos o más poblaciones. En este trabajo se ha calculado la distancia R, basada en el coeficiente de parentesco condicional (Harpending y Jenkins, 1973), mediante el programa GeDis 2.0 (Peña *et al*., 2009). Esta distancia se basa en la probabilidad de que un alelo tomado en un

individuo al azar de una población sea idéntico por descendencia a otro alelo del mismo locus en un individuo tomado al azar de otra población.

$$r_{ij} = \sum_{k} \frac{(P_{ik} - \overline{P_k})(P_{jk} - \overline{P_k})}{\overline{P_k}(1 - \overline{P_k})}$$

donde p_{ik}, p_{jk} son las frecuencias del alelo k en las poblaciones i y j y $\overline{P_k}$ es la frecuencia promedio del alelo k.

A partir de la matriz R de parentesco se construye una distancia genética mediante la fórmula:

$$d_{ij} = r_{ii} + r_{jj} - 2r_{ij}$$

5.4.4. Análisis de escalamiento multidimensional

El propósito principal del análisis de escalamiento multidimensional (MDS, del inglés *Multidimensional Scaling*) es la obtención de una representación euclídea de una serie de poblaciones en un espacio multidimensional (Torgerson, 1952; Krustal, 1964). Con la aplicación del *MDS* se pretende solucionar la dificultad que supone, dependiendo del número de poblaciones que se hayan incluido en el análisis, la interpretación de la matriz R de distancias. Los *MDS* tienden a generar una relación monótona creciente entre las medidas originales de proximidad (distancias génicas) y las distancias observadas en la configuración espacial resultante. La relación obtenida puede ser valorada en función del coeficiente de estrés. El cálculo de los MDS se realizó mediante el programa Past 3.0.7 (Hammer et al, 2001)

5.4.5. Test de Mantel

Este método valora la significación de la correlación entre dos matrices. Dado que los elementos de una matriz son interdependientes, no es posible conocer la distribución esperada de probabilidades de un coeficiente de correlación calculado entre matrices. Por ello, mediante un procedimiento de permutación iterativo se obtiene una distribución empírica nula del coeficiente de correlación (Mantel, 1967). Su cálculo se realizó mediante el programa Past 3.0.7 (Hammer et al, 2001)

5.4.6. Clinas de frecuencias alélicas

Se denomina clina a la variación asociada a una componente geográfica de una característica determinada genéticamente. Su origen puede deberse a variaciones en la eficacia biológica de diferentes genotipos a lo largo de un territorio en respuesta a factores ambientales, siempre que estos sean duraderos y estables en el tiempo. También pueden ser generadas por otros procesos, como la migración, cuando dos poblaciones con orígenes diferentes se asientan en puntos distantes de una región, de modo que la migración de individuos procedentes de ambas hacia la zona intermedia y el correspondiente mestizaje crearan una variación gradual (Hartl y Clark, 1997).

Para la detección de potenciales clinas, se han calculado los coeficientes de correlación entre las frecuencias alélicas y la posición geográfica de las poblaciones, respecto a un sistema de coordenadas

móviles que rota 360 grados. En sucesivas iteraciones se rota el eje un grado, se proyectan las posiciones geográficas de las poblaciones perpendicularmente sobre el eje y se calcula la correlación entre las coordenadas y las frecuencias. Cuando se encuentra un eje que presenta una correlación significativa, su orientación es considerada como una clina. A menudo se observan varias correlaciones significativas en grados consecutivos; en ese caso, el algoritmo selecciona aquella orientación para la cual la significación es máxima (Pérez-Miranda et al. 2003, 2004). La detección de clinas se ha realizado mediante el programa GeDis 2.0 (Peña et al. 2009).

5.4.7. Test exacto de diferenciación poblacional

El test exacto (Raymond y Rousset, 1995) es el estadístico más utilizado para comprobar las diferencias entre poblaciones. Es análogo al test exacto de Fisher en una tabla de contingencia 2x2, pero extendido a una tabla de contingencia rxk. Se calculó mediante el programa Arlequin 3.5 (Excoffier y Lischer, 2010).

6. RESULTADOS

6.1. Frecuencias alélicas

Las frecuencias de las seis inserciones *Alu* tipificadas en la muestra de gitanos vascos se muestran en la tabla 4.

Tabla 4: Frecuencias de las inserciones Alu, con sus errores estándares (±SE), diversidad genética (GD) y contenido de información polimórfica (PIC) en una muestra de gitanos vascos. 2N: número de cromosomas analizados.

Alu locus	2N	Frecuencia	SE	GD	PIC
TPA25	112	0,411	±0,024	0,484	0,367
ACE	74	0,284	±0,048	0,406	0,324
APO	98	0,990	±0,010	0,020	0,020
PV92	112	0,339	±0,046	0,448	0,348
FXIIIB	98	0,439	±0,050	0,492	0,371
D1	102	0,500	±0,051	0,500	0,375

Se observa un rango muy amplio de frecuencias, siendo APO la más cercana a la fijación y ACE la que presenta una menor frecuencia.

El grado de variabilidad genética de la muestra se evaluó mediante el cálculo de GD. Como cabría esperar, se observa el valor más bajo de diversidad en la inserción con un valor más extremo (APO). Por el contrario, los valores más elevados de diversidad se encuentran en aquellas inserciones que muestran una frecuencia de inserción más próxima a 0,5. Este es el caso de D1, que con una frecuencia de inserción precisamente de 0,5 muestra un grado de variabilidad máxima. Los valores del

contenido de información polimórfica de cada inserción muestran unos resultados similares. En conjunto, siendo marcadores bialélicos no ofrecen unos valores PIC elevados.

6.2. Equilibrio de Hardy-Weinberg

Excepto TPA25, todos los marcadores se encuentran en equilibrio de Hardy-Weinberg, como se observa en la tabla 5.

Tabla 5: Test de equilibrio Hardy-Weinberg para las seis inserciones Alu analizadas en este estudio. Het. obs: heterocigosidad observada; Het. esp: heterocigosidad esperada; P: probabilidad; SD: desviación estándar.

Marcador	Het. obs.	Het. esp.	P	SD
TPA25	0,821	0,484	0,000	0,000
ACE	0,405	0,406	1,000	0,000
APO	0,020	0,020	1,000	0,000
PV92	0,357	0,448	0,138	0,000
FXIIIB	0,429	0,492	0,390	0,000
D1	0,373	0,500	0,091	0,000

6.3. Comparaciones con otras poblaciones

Se han recopilado de la bibliografía las frecuencias alélicas para las seis inserciones *Alu* de una serie de poblaciones europeas, africanas y asiáticas. Se ha añadido a esta base de datos una población ancestral hipotética, en la que todas las frecuencias de inserción serían 0.

En la figura 4 se muestra un análisis de escalamiento multidimensional realizado a partir de una matriz de distancias R para estas poblaciones; se agrupan en una elipse roja las poblaciones del África sursahariana, en una elipse azul las poblaciones europeas (entre las que se incluyen las norteafricanas y de Próximo Oriente) y en una elipse amarilla las poblaciones de Medio y Extremo Oriente. Se observa una distribución acorde al modelo *Out of Africa.* Las poblaciones más próximas a la hipotética población ancestral son las poblaciones africanas. Estas a su vez muestran un cierto solapamiento con las europeas y un poco más alejadas y con una gran heterogeneidad aparecen las poblaciones de Medio y Extremo Oriente. En este análisis, la población gitana vasca (marcada con una elipse verde) aparece en el grupo de poblaciones europeas, si bien en la periferia.

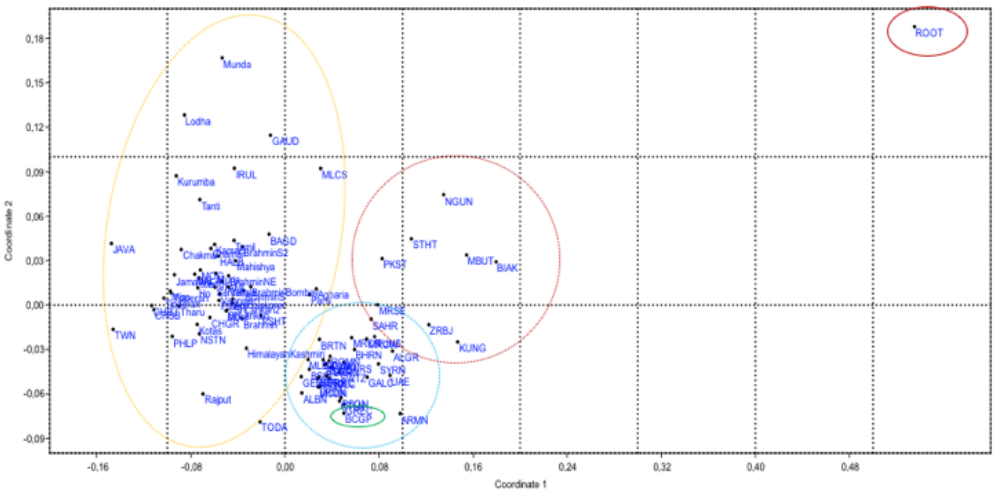

Figura 4: Análisis de escalamiento multidimensional (MDS) obtenido a partir de la matriz R de Harpending y Jenkins para poblaciones de Africa, Asia y Europa. Stress: 0,16. ROOT: poblción ancestral hipotética.

Una vez comprobada la eficacia de los marcadores utilizados para discriminar entre poblaciones de diferentes continentes, se redujeron las poblaciones de la base de datos a aquellas situadas en Europa, Próximo y Medio Oriente, eliminando del análisis aquellas situadas en África y al este de la India, ya que con toda probabilidad se encuentran alejadas de un potencial lugar de origen de los gitanos. Las poblaciones incluidas configuran un territorio a modo de transecto, que se muestra en la figura 5.

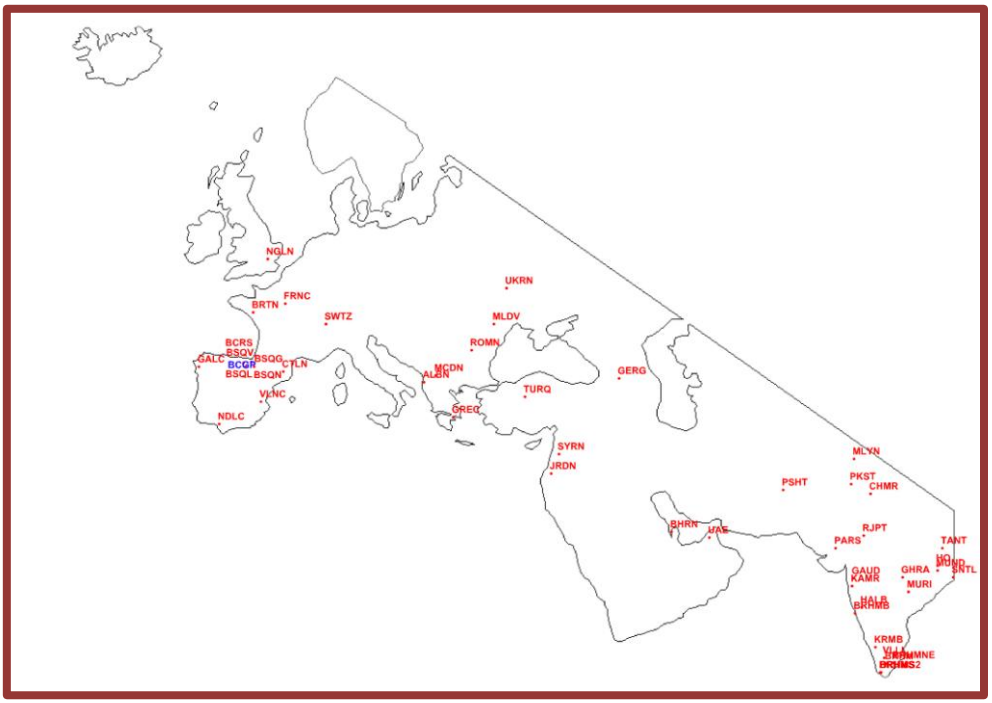

Figura 5: Poblaciones seleccionadas de Europa, Próximo y Medio Oriente.

Se realizó un nuevo análisis de escalamiento multidimensional con estas poblaciones, el cual se muestra en la figura 6. Se observan dos grandes grupos poblacionales, con un grado diferente de heterogeneidad. El grupo englobado por una línea amarilla hace referencia a las poblaciones de Medio Oriente; se trata de un grupo poblacional muy heterogéneo, ya que abarca gran parte de la superficie del gráfico, a pesar de que no incluye un número mayoritario de poblaciones (22 sobre 48). Por otro lado, las poblaciones de Próximo Oriente y Europa, englobadas en una elipse de color azul, aparecen agrupadas y mostrando una menor heterogeneidad.

Se realizó un test de Mantel, comparando la matriz de distancias genéticas y la matriz de distancias geográficas para este grupo de poblaciones, observándose una correlación significativa (R: 0,391, p: 0,000). En consecuencia, puede decirse que en esta región y para estos marcadores, geografía y genética siguen unos patrones similares, aunque lógicamente no puede descartarse una cierta influencia de otros factores.

La población de gitanos vascos (BCGP) aparece nuevamente en un punto intermedio entre ambos grupos, quedando relativamente próxima de las poblaciones de Emiratos Árabes (UAE) y Baréin (BHRN) en Próximo Oriente y de Parsis (PARS) y Agharias (GHRA) en Medio Oriente.

El test exacto de diferenciación poblacional reveló que las poblaciones con menor número de inserciones *Alu* con diferencias estadísticamente significativas con los gitanos vascos son las de Ucrania, con ninguna y Guipúzcoa, Inglaterra, Grecia, Emiratos Árabes, Jordania, Georgia, Parsis y Agharias, con una diferencia significativa.

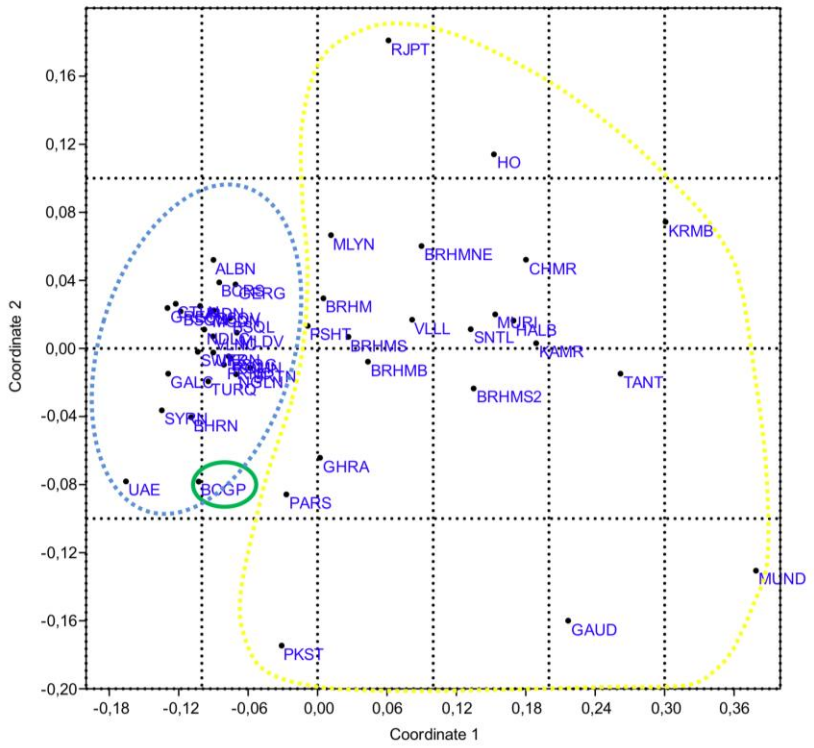

Figura 6: Análisis de escalamiento multidimensional (MDS) obtenido a partir de la matriz R de Harpending y Jenkins para poblaciones de Europa, Próximo y Medio Oriente. Stress: 0,10.

Con el fin de valorar la influencia de las diferentes inserciones *Alu* en la posición genética de los gitanos vascos, se muestran en la figura 7 los valores de sus frecuencias en relación al rango observado para el conjunto de poblaciones europeas y asiáticas. Se observa que se encuentran en unos valores medios para PV92 y FXIIIB, próximos a un extremo en TPA25, ACE y D1 y claramente en el extremo superior para APO.

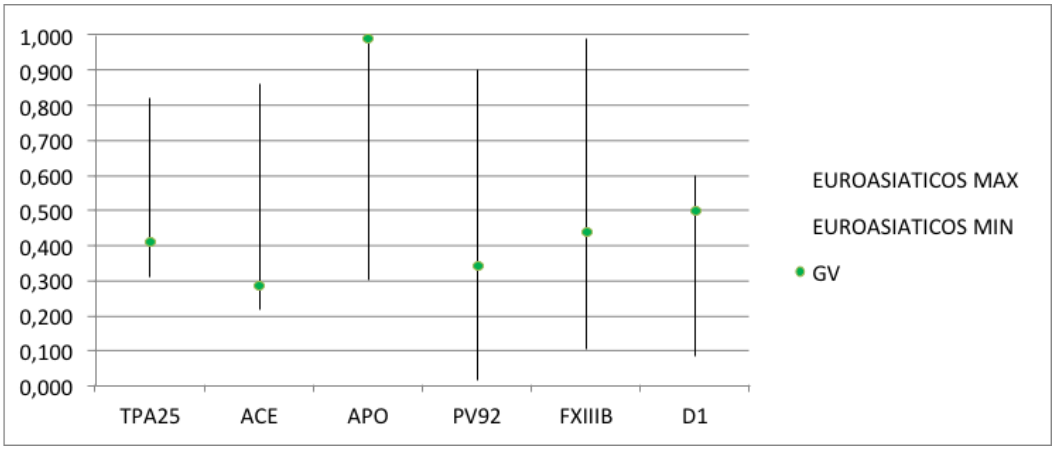

Figura 7: Gráfico de frecuencias de inserción máximas y mínimas de poblaciones europeas y asiáticas en relación con la frecuencia en la población gitana vasca.

6.4. Clinas de frecuencias alélicas

Se ha analizado la base de datos de poblaciones europeas y asiáticas en busca de posibles clinas para los marcadores *Alu* incluidos en este trabajo, encontrándose 5 clinas significativas (figura 8), que corresponden a TPA25, ACE, APO, FXIIIB y PV92. Tan sólo D1 no mostró un gradiente significativo de frecuencias. Puede apreciarse que hay 3 clinas que muestran un gradiente aproximadamente noroeste-sudeste (ACE, FXIIIB y PV92) y una clina con un gradiente opuesto (APO); este grupo de clinas se ajusta razonablemente al transecto definido por las poblaciones seleccionadas. Por último, se observa una clina norte-sur (TPA 25).

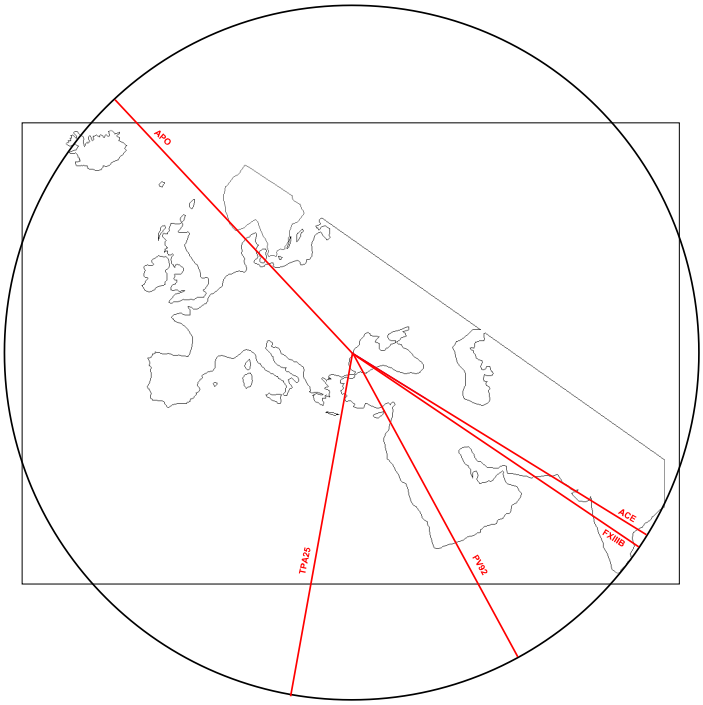

Figura 9: Orientación de las clinas observadas en 6 inserciones Alu en poblaciones de Europa, Próximo y Medio Oriente (ACE, APO, PV92, TPA25 Y FXIIIB).

La clina de TPA 25, por su orientación, no permite una discriminación clara entre poblaciones europeas y asiáticas (figura 9), observándose un solapamiento entre las poblaciones de ambos continentes. Esta clina es significativa al nivel 0,05, pero no al nivel 0,01. En todo caso, los gitanos vascos presentan la tercera más baja frecuencia de inserción.

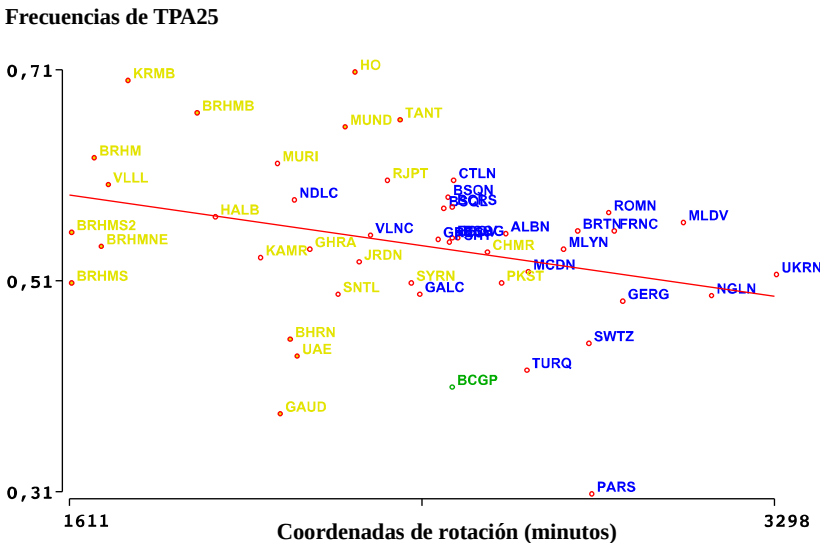

Figura 9: Distribución clinal de las frecuencia de inserción del marcador TPA25. Su orientación es de 190 grados respecto al Norte (R: 0,300; p: 0,035). La línea roja muestra la recta de regresión de las frecuencias alélicas respecto a las coordenadas rotadas.

La inserción ACE presenta un gradiente noroeste-sudeste (122 grados de acimut), de modo que a diferencia de TPA25, este marcador permite *a priori* una discriminación entre poblaciones europeas y asiáticas (figura 10). No obstante, existe un cierto grado de solapamiento. Los valores más altos de inserción de este marcador se observan en poblaciones de Medio Oriente, aunque cabe destacar el amplio rango de frecuencias que muestran estas poblaciones. Las poblaciones europeas y de Próximo Oriente muestras unos valores y un rango de frecuencias menor. En cuanto a los gitanos vascos, son la segunda población, después de Grecia, con la frecuencia de inserción más baja. Se encuentran en el extremo inferior del rango de las poblaciones europeas y muy por debajo del rango de las poblaciones de Medio Oriente.

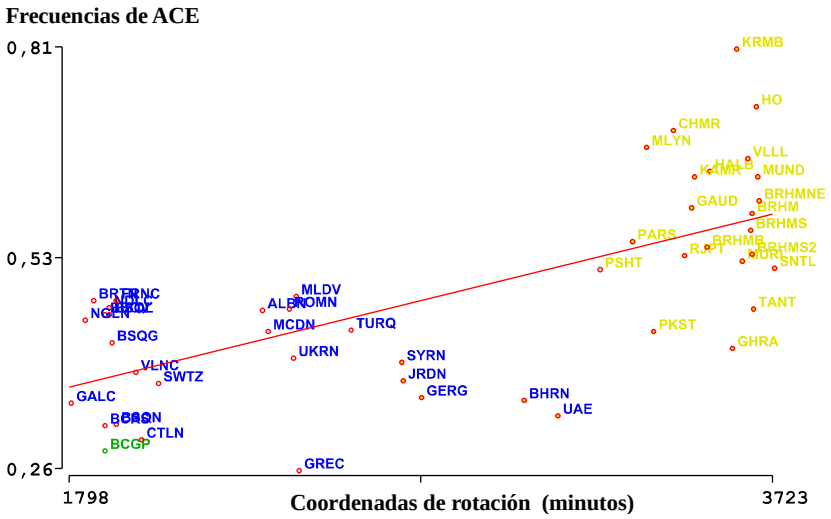

Figura 10: Distribución clinal de las frecuencia de inserción del marcador ACE. Su orientación es de 122 grados respecto al Norte (R: 0,680; p: 0,000). La línea roja muestra la recta de regresión de las frecuencias alélicas respecto a las coordenadas rotadas.

Prácticamente de sentido opuesto a la clina de ACE, APO muestra un gradiente sudeste-noroeste (figura 11). Las poblaciones de Medio Oriente presentan también un amplio rango de frecuencias, en tanto que Europa y Próximo Oriente resultan bastante homogéneas. La frecuencia en los gitanos vascos, igual a la de los franceses (0,990) sólo es superada por los albaneses (1,000).

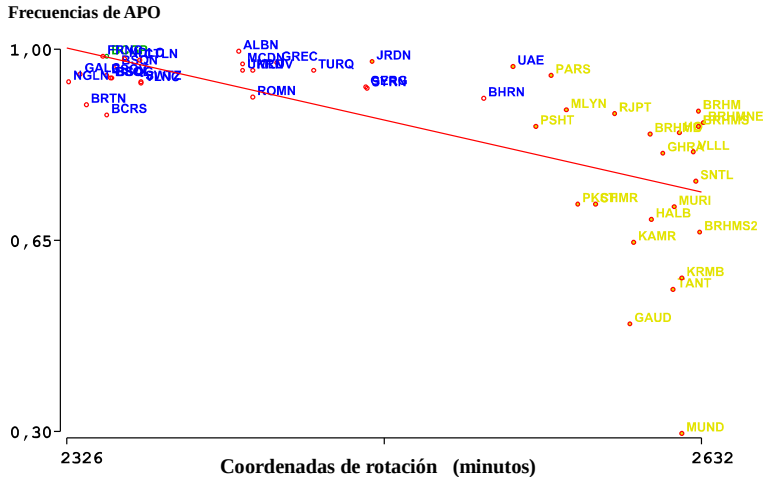

Figura 11: Distribución clinal de las frecuencia de inserción del marcador APO. La clina presenta una orientación sudeste-noroeste. La orientación es de 317 grados respecto del Norte (R: 0,670; p: 0,000). La línea roja muestra la recta de regresión de las frecuencias alélicas respecto a las coordenadas rotadas.

Las frecuencias de inserción de FXIIIB presentan un gradiente prácticamente idéntico al de ACE, tanto en dirección como en sentido (noroeste-sudeste) (figura 12). Se observa una heterogeneidad mucho más amplia en Asia que en Europa, de modo que ésta queda integrada completamente dentro del rango de aquella. Los gitanos vascos aparecen en un punto intermedio del rango europeo.

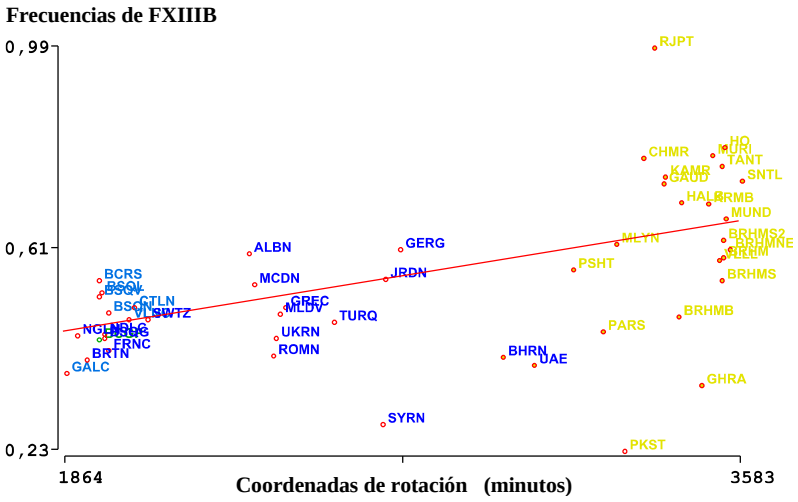

Figura 12: Distribución clinal de las frecuencia de inserción del marcador FXIIIB. La clina presenta una orientación noroeste-sudeste. La orientación es de 124 grados respecto del Norte (R: 0,500; p: 0,000). La línea roja muestra la recta de regresión de las frecuencias alélicas respecto a las coordenadas rotadas.

PV92 muestra una orientación de nor-noroeste a sur-sudeste, más cerca de un paralelo que de un meridiano, pero aun así se ajusta bastante bien a la orientación principal del territorio del transecto de las poblaciones seleccionadas (figura 13). La distribución de frecuencias de PV92 muestra unas diferencias claras entre Europa y Asia, como APO y ACE. Sin embargo, a diferencia de los casos

anteriores, los gitanos vascos presentan una frecuencia superior a todas las poblaciones europeas, incluyéndose dentro del rango de frecuencias asiáticas.

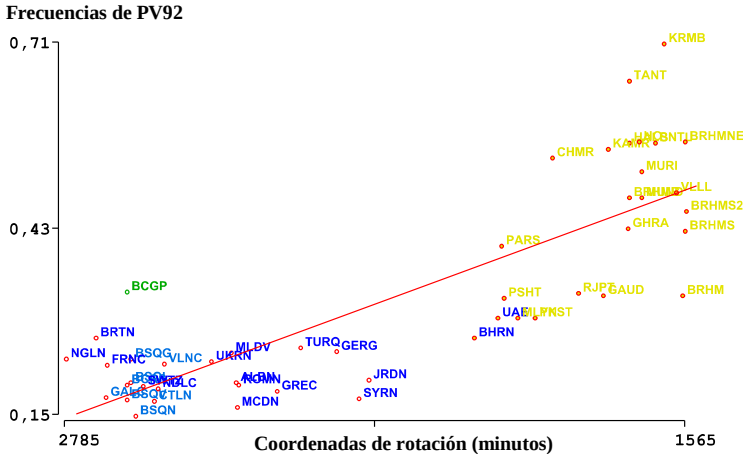

Figura 13: Distribución clinal de las frecuencia de inserción del marcador PV92. La clina presenta una orientación nornoroeste-sursudeste. La orientación es de 151 grados respecto del Norte (R: 0,820; p: 0,000). La línea roja muestra la recta de regresión de las frecuencias alélicas respecto a las coordenadas rotadas.

7. DISCUSIÓN Y CONCLUSIONES

Hace entre 6.000 y 5.000 años, en la estepa situada al norte del Cáucaso, entre el Mar Negro y el Mar Caspio, habitaba un grupo de pastores de la cultura Yamnaya, con poco contacto con otras poblaciones del centro y oeste de Europa. Se originaron presumiblemente a partir de grupos de cazadores recolectores autóctonos y de granjeros provenientes de Próximo Oriente. Un tiempo después, hace unos 4.500 años, protagonizaron una migración masiva hacia el resto del continente, favorecida por su conocimiento de la rueda, los metales y la domesticación de caballos, originando posiblemente la expansión de los idiomas indoeuropeos (Haak et al. 2015). Es la hipótesis Kurgán (sinónimo de Yamnaya), que tradicionalmente ha competido con la hipótesis de los granjeros de Anatolia (Piazza et al, 1995; Ray & Atkinson, 2003) para explicar el origen de los idiomas indoeuropeos. Estos idiomas se encuentran presentes en la mayor parte de Europa y buena parte de Medio Oriente, como la India, entre otras regiones de Asia (Lewis et al. 2015).

Pero además de poblaciones con lenguajes indoeuropeos, que habrían llegado hace unos 3.500 años, en la India también conviven poblaciones dravidianas, probablemente originadas por grupos de granjeros provenientes del este del creciente fértil desplazadas por los indoeuropeos (Thanseem et al. 2006) y poblaciones austroasiáticas como los munda, más emparentadas con otras del sudeste asiático (Riccio et al, 2011). Esta compleja historia es posiblemente la principal causa de la enorme heterogeneidad genética del subcontinente indio.

Los gitanos habrían realizado un viaje similar a los indoeuropeos, pero en sentido contrario. En efecto, como se ha mencionado en la introducción, presumiblemente su origen se encuentra en una región situada entre el noroeste de la India y el norte de Pakistán y su presencia en Europa parece bien

documentada a partir de los siglos XI o XII en los Balcanes, donde habrían permanecido unos dos siglos, antes de dispersarse por todo el continente (Mendizabal et al. 2011). En la Península Ibérica se encontrarían al menos desde 1425, año en que el rey Alfonso de Aragón emitió un salvoconducto que les permitía realizar el Camino de Santiago (Lermo et al, 2006).

Durante este trayecto han podido experimentar numerosos cuellos de botella y eventualmente procesos de mestizaje con las poblaciones que fueran encontrando a su paso. Si bien presentan un fuerte componente endogámico, en algunos casos se ha observado un cierto flujo génico, como en los gitanos húngaros, en los que se ha detectado en proporción relevante un marcador casi exclusivamente europeo (Almos et al, 2008).

En la población gitana vasca fundamentalmente se ha encontrado la huella de la deriva genética. Así, ha mostrado valores en los extremos del rango europeo y asiático para varias inserciones, algo que sólo puede explicarse por deriva. Para TPA25, sólo dos poblaciones asiáticas han mostrado valores más bajos de la inserción, parsis (PARS, 0,309) y gauds (GAUD, 0,385). Para ACE, tan sólo Grecia presenta un valor más bajo. En APO, el valor es sólo superado por los albaneses. Entre las frecuencias de D1, la única inserción que no presenta clina, sólo son más altas las de los kurumba (0,528) y los bramines del nordeste (0,521). Teniendo en cuenta lo extraordinariamente elevada que es la heterogeneidad de las poblaciones de la India, es muy relevante que los gitanos vascos se encuentren en esta situación en 4 de las 6 inserciones analizadas. Por lo demás, FXIIIB presenta un solapamiento completo de la heterogeneidad entre Europa, Próximo y Medio Oriente, de modo que no resulta informativa acerca de las circunstancias de los gitanos vascos. Tan sólo PV92 podría informar acerca de su grado de mestizaje, ya que se encuentran fuera del rango de los europeos y en el rango de las poblaciones indias; sin embargo, por todo lo anterior no tiene sentido estimar una tasa de mestizaje, ya que a todas luces el flujo génico que sin duda ha existido entre gitanos y europeos ha quedado enmascarado por los efectos de la deriva genética.

En un análisis MDS (figura 6), los gitanos vascos aparecen separados de las poblaciones europeas y relativamente próximos a algunas poblaciones de Próximo Oriente (Emiratos Árabes y Baréin) y de Medio Oriente (parsis y agharias). En relación a las poblaciones de Próximo Oriente, es lógica su proximidad, dado que se encuentran a medio camino entre las poblaciones europeas e indias, precisamente en la ruta de los kurganes; por lo demás, es conocido el papel que ha jugado Próximo Oriente como cruce de caminos a lo largo de la historia (Pérez-Miranda et al. 2006). Y en relación a las poblaciones de Medio Oriente, es demasiado aventurado proponerlas como los parientes más próximos de los gitanos, a la luz de la gran influencia de la deriva sobre su patrimonio genético.

En relación a los marcadores utilizados, no cabe dudar de su idoneidad, ya que han mostrado que son capaces de reflejar eficazmente la heterogeneidad entre continentes e incluso el proceso *Out of Africa* (figura 5), algo que en todo caso ha sido ya previamente mostrado (Gómez-Pérez et al (2007).

En definitiva, puede concluirse que los gitanos residentes en el País Vasco han revelado una notable heterogeneidad genética en relación a otras poblaciones europeas y asiáticas y aunque a la luz

de los resultados obtenidos puede asumirse un origen asiático de esta población, no es posible identificar su origen más concretamente, ya que en sus procesos de microdiferenciación se ha impuesto con creces la deriva a la previsible acción del flujo génico.

8. BIBLIOGRAFÍA

- Almos P Z, Horváth S, Czibula A, Raskó I, Sipos B, Bihari P, Béres J, Juhász A, Janka Z, Kálmán J, (2008). H1 tau haplotype-related genomic variation at 17q21.3 as an Asian heritage of the European Gypsy population. *Heredity (Edinb).* 101(5):416-9.
- Arcot SS, Fontius JJ, Deininger PL & Batzer MA (1995b). Identification and analysis of a "young" polymorphic *Alu* element. *Biochim Biophys Acta*, 1263:99-102.
- Batzer MA & Deininger PL (2002) *Alu* repeats and human genomic diversity. *Nat Rev Genet*, 3:370-379.
- Batzer MA, Arcot SS, Phonney JW, Alegria-Hartman M, Kass DH, Milligan SM, Kimpton C, Gill P, Hochmeister M, Ioannou PA, Herrera RJ, Boudreau DA, Scheer WD, Keats BJB, Deininger PL & Stoneking M (1996b) Genetic variation of recent *Alu* insertions in human populations. *J Mol Evol*, 42:22-29.
- Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E & Zuckerkandl E (1996a). Standardized nomenclature for *Alu* repeats. *J Mol Evol*, 42:3-6.
- Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ & Deininger PL (1991) Amplification dynamics of human-specific (HS) *Alu* family members. *Nucleic Acids Res*, 19:3619-3623.
- Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, Bazan HA, Shaikh TH, Deininger PL & Schmid CW (1995) Dispersion and insertion polymorphism in two small subfamilies of recently amplified human *Alu* repeats. *J Mol Biol*, 247:418-427.
- Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ & Deininger PL (1994) African origin of human specific polymorphic *Alu* insertions. *Proc Natl Acad Sci USA*, 91:12288-12292.
- Comas D, Plaza S, Calafell F, Sajantila A & Bertranpetit J (2001) Recent insertion of an Alu element within a polymorphic human-specific *Alu* insertion. *Mol Biol Evol*, 18:85-88.
- Deininger PL & Daniels GR (1986) The recent evolution of mammalian repetitive DNA elements. *Trends Genet*, 2:76-80
- ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C & Snyder M (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57-74

- Excoffier L & 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*, 10:564-567. http://cmpg.unibe.ch/software/arlequin35/
- Fuhrman SA, Deininger PL, LaPorte P, Friedmannn T & Geiduschek EP (1981) Analysis of transcription of the human *Alu* family ubiquitous repeating element by eukaryotic RNA polymerase III. *Nucleic Acids Res*, 9:6439-6456
- Gómez-Pérez L, Alfonso-Sánchez MA, Pérez-Miranda AM, de Pancorbo MM, Peña JA, 2007, Utilidad de las inserciones *Alu* en los estudios de mestizaje, Antropo, 14, 29-36. www.didac.ehu.es/antropo
- Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen P, Tournev I, de Pablo R, Kuĉinskas V, Perez-Lezaun A, Marushiakova E, Popov V, Kalaydjieva L (2001) Origins and divergence of the Roma (gypsies). *Am J Hum Genet*. 69(6):1314-31.
- Haak W, Lazaridis I, Patterson N, et al (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. 522(7555):207-11.
- Hammer, Ø., Harper, D.A.T., Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4(1): 9pp. http://palaeo-electronica.org/2001_1/past/issue1_01.htm
- Hardy GH (1908) Mendelian proportions in a mixed population. *Science*, 28:49-50.
- Harpending H & Jenkins T (1973) Genetic distance among southern African populations. In: Methods and theories of anthropological genetics. EDK University of New Mexico Press, Albuquerque, USA
- Häsler J & Strub K (2006). *Alu* elements as regulators of genet expression. *Nucleic Acids Res*, 34: 5491-5497.
- Houck CM, Rinehart FP & Schmid CW (1979) A ubiquitous family of repeated DNA sequences in the human genome. *J Mol Biol*, 132:289-306
- Jobling MA & Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*, 4:598-612
- Jorde LB & Wooding SP (2004) Genetic variation, classification and `race´. *Nat Genet*, 36:S28-33
- Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, Soodyall H, Jenkins T & Rogers AR (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet*, 57:523-538
- Kass DH, Aleman C, Batzer MA & Deininger PL (1994) Identification of a human specific *Alu* insertion in the factor FXIIIB gene. *Genetica*, 94:1-8.

- Knight A, Batzer MA, Stoneking M, Tiwari HK, Scheer WD, Herrera RJ & Deininger PL (1996) DNA sequences of *Alu* elements indicate a recent replacement of the human autosomal genetic complement. *Proc Natl Acad Sci*, 93:4360-4364
- Kruskal J (1964) Non metric multidimensional scaling: A numeric method. *Psycometrika*, 29:115-129.
- Labuda D & Zietkiewicz E (1994) Evolution of secondary structure in the family os 7SL-like RNAs. *J Mol Evol*, 39:506-518
- Lell JT & Wallace DC (2000) The peopling of Europe from the maternal and paternal perspectives. *Am J Hum Genet*, 67:1376-1381
- Lermo, J., Román, J., Marrodán, M.D., Mesa, M.S. 2006, Modelos de distribución de apellidos en la población gitana española. Antropo, 13, 69-87. www.didac.ehu.es/antropo
- Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Research 27 (2): 209–220
- Mendizabal I, Lao O, Marigorta UM, Wollstein A, Gusmão L, Ferak V, Ioana M, Jordanova A, Kaneva R, Kouvatsi A, Kučinskas V, Makukh H, Metspalu A, Netea MG, de Pablo R, Pamjav H, Radojkovic D, Rolleston SJ, Sertic J, Macek M Jr, Comas D, Kayser M. (2012) Reconstructing the population history of European Romani from genome-wide data. Curr Biol. 22(24):2342-9
- Mendizabal I, Valente C, Gusmão A, Alves C, Gomes V, Goios A, Parson W, Calafell F, Alvarez L, Amorim A, Gusmão L, Comas D, Prata MJ. (2011) Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. PLoS One. 6(1):e15988.
- Mighell AJ, Markham AF & Robinson PA (1997) *Alu* sequences. *FEBS lett*, 417:1-5
- Nei M (1987) Molecular Evolutionaly Genetics. EDK Columbia University Press, New York, USA.
- Novick GE, Batzer MA, Deininger PL & Herrera RJ (1996) The mobile genetic element *Alu* in the human genome. *Bioscience*, 46:32-41
- Peña JA, Alfonso-Sanchez MA, Pérez-Miranda AM, Garcia-Obregón S & Gómez-Perez L (2009) GeDis: un programa para análisis de datos en antropogenética. *Antropo*, 20:49-59
- Pérez-Miranda AM, Alfonso-Sánchez MA, Peña JA & Calderón R (2003) HLA-DQA1 polymorphism in autochthonous Basques from NAvarre (Spain): genetic position within European and Mediterranean scopes. *Tissue Antigens*, 61:465-474.
- Pérez-Miranda AM, Alfonso-Sánchez Ma, Vidales MC, Calderón R & Peña JA (2004) Genetic polymorphism and linkage disequilibrium of the HLA-DP región in Basques from Navarre (Spain). *Tissue Antigens*, 64:264-275.

- Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, Cavalli-Sforza LL (1995) Genetics and the origin of European languages. *Proc Natl Acad Sci USA*. 92(13):5836-40.
- Quentin Y (1992) Origin of the *Alu* family: a family of Alu-like monomers gave birth to the left and the right arms of the *Alu* elements. *Nucleic Acids Res*, 20:3397-3401
- Riccio ME, Nunes JM, Rahal M, Kervaire B, Tiercy JM, Sanchez-Mazas A (2011) The Austroasiatic Munda population from India and Its enigmatic origin: a HLA diversity study. Hum Biol. 83(3):405-35.
- Richards M (2003) The Neolithic invasion of Europe. *Annu. Rev. Anthropol*. 32:135-162
- Rowold DJ & Herrera RJ (2000) *Alu* elements and the human genome. *Genetica*, 108:57-72
- Roy-Engel AM, Batzer MA & Deininger PL (2008) Evolution of Human Retrosequences: Alu. In: Encyclopedia of Life Sciences (ELS). EDK John Wiley & Sons, Ltd: Chichester, UK
- Salem AH, Raya DA & Batzer MA (2005) Identity by descent and DNA sequence variation of human SINE and LINE elements. *Cytogenet Genome Res*, 108:63-72
- Schmid CW & Shen CJ (1985) The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. In Molecular Evolutionary Genetics edited by R.J. MacIntire. Plenum, New York, USA
- Shriver MD & Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. *Nat Rev Genet*, 5:611-618
- Thanseem I, Thangaraj K, Chaubey G, Singh VK, Bhaskar LV, Reddy BM, Reddy AG, Singh L. (2006) Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet. 7:42.
- Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F & Soubrier F (1992) Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet, 51:197-205.
- Torgerson W (1952) Multidimensional scaling: I. Theory and method. *Psychometriks*, 17:401-419.
- Weinberg W (1908) Über den Nachweis der Vererbung Beim Menschen. *Jahreshefte des Vereins Für vaterländische Naturkunde in Württemberg*, 64:368-382.
- Weiner A, Deininger P & Efstradiatis A (1986) Nonvarial retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem, 55:631-661