

## CAPÍTULO 4

### BASES DE DATOS EN GENÉTICA FORENSE

DEFINICIÓN DE BASES DE DATOS EN GENÉTICA FORENSE. TIPOS DE  
BASES DE DATOS

*Guillermo Giovambattista, Daniel Estanislao Goszczynski, María Elena  
Fernández, Juan Pedro Lirón, Pilar Peral García*

#### 4.1. Introducción

En la práctica de genética forense de rutina, una vez que las muestras llegan al laboratorio, se realizan una serie de pasos sucesivos durante la resolución de casos forenses: extracción de ADN, tipificación de los marcadores genéticos, análisis de los resultados, estimación de los índices forenses y redacción de informes. Es por esta razón que los genetistas forenses, tanto dedicados a genética forense humana como no humana, deben recurrir a las bases de datos genéticos para estimar estadísticamente el valor de los resultados presentados en los informes forenses y, por lo tanto, determinar el peso de la evidencia en un juicio. Además, es frecuente que el perfil de ADN de una evidencia tenga que ser comparado con uno obtenido previamente, motivo por el cual ha surgido la necesidad de implementar las denominadas **bases datos forenses**, definidas como "el conjunto de programas informáticos (*software*) y soportes físicos (*hardware*) donde se almacena de modo ordenado y coherente la información de los perfiles genéticos, así como todo dato asociado a la muestra/individuo, información que luego puede ser recuperada y comparada de modo automático de acuerdo a parámetros previamente establecidos".

En los últimos años, el desarrollo tecnológico en los métodos de genotipificación y de bioinformática ha permitido la generación de una cantidad de información genética que tiene que almacenarse de modo racional y ordenada para su posterior uso. Es por ello que las bases de datos han tenido un gran desarrollo en genética forense humana, y en menor medida en genética forense no humana. Es común que cada laboratorio de genética forense tenga su propia base de datos, producto de la información acumulada correspondiente a las muestras analizadas a lo largo del tiempo, que es utilizada para la estimación de los índices forenses. Sin embargo, estas bases de datos pueden estar sesgadas y no representar el perfil y la estructuración de la/s población/es de donde proceden las evidencias y referencias que tipifica un determinado laboratorio. Por lo tanto, se recomienda la implementación de muestreos especialmente diseñados para construir bases de datos forenses representativas de las zonas de influencia del laboratorio y que los cálculos estadísticos sean realizados a partir de parámetros poblacionales correspondientes a las poblaciones de origen de las muestras analizadas. Estos muestreos deben incluir todos los tipos raciales presentes en una determinada región geográfica. Así por ejemplo, si se desea implementar una base de datos para una determinada especie de animales domésticos, deben muestrearse y tipificarse animales de todas las razas criadas en esa zona. Por otra parte, ante la necesidad de contar con bases de datos representativas y de libre acceso, se han llevado a cabo diferentes iniciativas nacionales e internacionales. A lo largo del capítulo se mencionarán diferentes ejemplos de bases de datos forenses implementadas hasta el momento.

## **4.2. Clasificación de bases de datos**

Se han implementado diferentes tipos de bases de datos genéticos forenses, que pueden agruparse de diferente forma según el criterio utilizado. Como se mencionó en el capítulo 1, la genética forense puede

abordarse a nivel individual, poblacional/racial o de especie, por lo que una primera división de las bases de datos puede basarse en aquellas que son utilizadas para identificación individual y en las empleadas para la determinación de la especie de origen.

Las **bases de datos para identificación genética individual** contienen perfiles de marcadores genéticos, expresados en números y/o letras, asociados al código de identificación de una persona, animal o muestra. El acceso a los datos debe estar perfectamente controlado, y las conclusiones que se pueden obtener de los mismos van a depender de los programas informáticos que sean autorizados por las diferentes legislaciones. El acceso a dichas bases de datos es más o menos restringido, dependiendo de la trascendencia o la propiedad de los datos almacenados, siendo los más protegidos aquéllos que contienen información sobre la "identidad" de las personas (Lorente et al., 2001, 2002a, 2002b; Álvarez-Cubero et al., 2010).

Las **bases de datos genéticos para la determinación del origen especie- específico** de muestras biológicas almacenan secuencias de ADN, especialmente aquellas basadas en la secuenciación de regiones conservadas de ADNmt (por ejemplo, Cyt b, COI). Las principales bases públicas disponibles en la actualidad que pueden ser utilizadas para la determinación del origen especie específico son GenBank ([www.ncbi.nlm.nih.gov/genbank/](http://www.ncbi.nlm.nih.gov/genbank/)) y BOLD (*Barcoding of Life Database*, <http://www.barcodeoflife.org/>). Aunque el GenBank es el repositorio primario de secuencias de ADN, esta base de datos no cuenta con software especializados para el almacenamiento, la organización y el análisis de los datos asociados al origen de las secuencias, y por lo tanto, que certifique la especie de origen de la muestra secuenciada. Por el contrario, en BOLD, la unidad básica de información es el registro del espécimen, el cual es asociado a la información de la secuencia de ADN. El registro del espécimen incluye imágenes y la posición satelital de la localidad donde fue colectado, entre otros datos. En el capítulo 9 se discuten y comparan las características de estas bases de datos en el

contexto de la identificación especie específica. Las bases de datos de secuencia de ADN, por ejemplo las basadas en regiones variables del ADNmt (lazo D; *D-loop*), también pueden ser utilizadas para identificación individual o para la determinación de la raza o población de origen, aunque con un bajo poder estadístico.

Las bases de datos genéticas también pueden clasificarse de acuerdo a la clase de marcadores genéticos que almacenan:

**Bases de datos de microsatélites (STRs).** Estas bases de datos almacenan los genotipos (perfiles genéticos) de los microsatélites correspondientes a los paneles de identificación genética, y representan la diversidad genética autosómica y la estructura genética de las poblaciones. Los genotipos se ingresan generalmente con números, que representan el número de repeticiones o los pares de bases, o en algunos casos con códigos de letras (por ejemplo, equinos). Esta nomenclatura es estandarizada internacionalmente. En el caso de los animales domésticos, el panel de microsatélites y la nomenclatura es recomendada por la ISAG (<http://www.isag.us/>). Un ejemplo de bases de datos basadas en STRs es la Unidad CODIS. Esta base maneja los sistemas CODIS (*Combined DNA Index System*) y NDIS (*National DNA Index System*) y desarrolla y provee soporte técnico al programa CODIS (<http://www.fbi.gov/about-us/lab/biometric-analysis/codis>). En marzo de 2014 esta base de datos contenía más de 10 millones de perfiles de ADN basados en el panel CODIS de STRs. Actualmente, la eficiencia del programa es medida por la cantidad de crímenes que ha ayudado a resolver.

**Base de datos de linajes maternos y paternos.** Este tipo de base de datos genéticos almacena información de haplotipos mitocondriales o del cromosoma Y, ya sea producto del análisis de microsatélites, SNPs, o ambos. Dado que los marcadores genéticos con herencia materna o paterna no tienen una distribución geográfica al azar, los perfiles de las evidencias tienen más probabilidad de compartir características con los individuos de la misma región (Roewer et al., 2007). Es por esta razón

que aunque tienen poco poder estadístico para la identificación de individuos, son de gran utilidad para determinar su origen racial. Las bases de datos de ADNmt son de gran utilidad en genética forense, ya que generalmente esta clase de marcadores pueden ser obtenidos a partir de rastros o en muestras altamente degradadas.

Un ejemplo de bases de datos de linajes es YHRD (*Y Chromosome Haplotype Reference Database*). Se trata de la mayor base de datos de haplotipos del cromosoma Y humano; está ubicada en el Instituto de Medicina Legal y Ciencias Forenses (Berlín, Alemania, [www.yhrd.org](http://www.yhrd.org)) y contiene más de 115.000 haplotipos pertenecientes a 850 poblaciones humanas (Willuweit y Roewer, 2007). YHRD es de gran utilidad para inferir la población de origen de una muestra (linaje paterno). Además, los haplotipos del cromosoma Y basados en STRs son de particular utilidad para tipificar ADN de machos purificados a partir de muestras forenses que presentan mezcla.

EMPOP ([www.empop.org](http://www.empop.org)) es la mayor base de datos de ADNmt y es mantenida por el Instituto de Medicina Legal de Innsbruck, Austria. El repositorio posee en la actualidad más de 34.000 secuencias de la región hipervariable de la región control del ADNmt, gracias a la colaboración de más de 100 instituciones de 63 países (Parson y Dür, 2007). Esto pone en evidencia la importancia de las colaboraciones internacionales para la implementación de estas bases de datos.

**Base de datos de secuencias de ADN.** Mientras que las dos primeras permiten la identificación de individuos y del origen racial, estas bases de datos genéticas son utilizadas principalmente para la determinación de la especie de origen de una muestra. Almacenan secuencias de ADN, siendo las regiones conservadas del ADNmt las más utilizadas. Como se mencionó anteriormente, los dos principales repositorios que incluyen secuencias que pueden ser utilizadas para identificación especie-específica son GenBank y BOLD. Por otra parte, también existen bases de datos específicas, como por ejemplo, la perteneciente a la Unidad de ADN Mitocondrial (mtDNAU);

<http://www.fbi.gov/about-us/lab/biometric-analysis/mtdna>) del FBI (*Federal Bureau of Investigation*), que analiza y almacena secuencias de ADNmt humanas para resolver diferentes tipos de crímenes.

De acuerdo a la procedencia de las muestras, las bases de datos pueden dividirse en:

**Bases de datos poblacionales.** Estas bases de datos almacenan perfiles genéticos (genotipos) de individuos representativos de la/s población/es de pertenencia de los individuos involucrados en un delito. Permiten obtener las frecuencias génicas y otros parámetros poblacionales necesarios para el cálculo del poder de exclusión y los índices forenses (ver capítulos 5 y 6).

Ejemplos de bases de datos poblacionales en animales domésticos son los trabajos realizados por Halverson y Basten (2005), Himmelberger y colaboradores (2008) y Kanthaswamy y colaboradores (2009). Estos estudios incluyeron el análisis de 18 microsatélites de la región control del ADNmt en las poblaciones de las razas más populares de EE.UU, según los registros de pedigrís de la Asociación Canina Norteamericana (AKC, por sus siglas en inglés, *American Kennel Club* <http://www.akc.org/>), así como de perros mestizos distribuidos a lo largo de EE.UU. La base de datos resultante es de acceso público y contiene información sobre los loci, las frecuencias génicas, la distribución de la variación genética entre las poblaciones analizadas, las estimaciones de la probabilidad de coincidencia entre muestras y los coeficientes de consanguinidad (*STR DNA Internet DataBase*, <http://www.cstl.nist.gov/strbase/>). Entre las razas incluidas se encuentran tanto las que se usan comúnmente como mascotas como aquéllas que han sido causales de agresiones a humanos y animales. Es por ello que estas bases de datos poblacionales pueden ser utilizadas para resolver casos de robo de mascotas y ataques a humanos o animales.

Los gatos son uno de los animales domésticos más comunes en todo el mundo, razón por la que es usual encontrar pelos de esta especie en la escena del crimen, los que pueden ser usados para relacionar a la

mascota con la víctima y/o el sospechoso. Con el fin de establecer una base de datos para usos forenses, Grahn et al. (2011) analizaron el lazo D mitocondrial de 25 poblaciones correspondientes a 26 razas de gatos distribuidas en todo el mundo.

**Bases de datos de inteligencia.** Esta clase de bases de datos almacena la información genética de muestras de individuos conocidos, generalmente acusados o culpables de algún delito, así como de donantes muestreados al azar en la población. Su aplicación permite comparar los genotipos de las evidencias recolectadas en el lugar del hecho con los perfiles almacenados en la base de datos y de esta forma identificar al autor de un ilícito. El primer uso de este tipo de bases de datos para resolver un crimen ocurrió en el Reino Unido en el año 1995. La frase "*cold hit*" se refiere a la coincidencia (*match*) entre el perfil de ADN de una evidencia y el de una muestra almacenada en la base de datos. Los "cold hit" son de gran utilidad para la identificación del sospechoso durante una investigación. Dependiendo del contenido de estas bases de datos, podemos diferenciar:

**Bases de datos forenses criminales.** Almacenan perfiles genéticos procedentes de personas que han sido procesadas o condenadas, así como de indicios biológicos encontrados en la escena del crimen. En algunos casos pueden considerarse perfiles genéticos de víctimas conocidas, con el objeto de facilitar la resolución de delitos. Su característica principal es que algunas de las muestras y datos considerados se obtienen sin el consentimiento de las personas implicadas.

**Bases de datos forenses civiles.** Su único fin es la identificación de personas desaparecidas, lo cual se hace comparando el ADN de los individuos no identificadas con el de los familiares. Por sus características humanitarias, se requiere que los familiares colaboren de modo voluntario luego de firmar un consentimiento informado siguiendo las consideraciones del protocolo de Helsinki (<http://www.wma.net/es/30publications/10policies/b3/>), hecho que las

diferencia de las bases de datos criminales. Así por ejemplo, podemos mencionar el programa Fénix (<http://www.guardiacivil.org/prensa/actividades/fenix/index.jsp>) y la iniciativa DNA-PROKIDS (<http://www.dna-prokids.org/>).

El Programa Fénix fue implementado para la identificación mediante ADN de cadáveres y restos humanos procedentes de guerras, problemas socio-políticos y desastres en masa. El programa contiene dos bases de datos independientes. Una de ellas es la *Base de Datos de Referencia* que contiene información de microsátélites de ADN nuclear y del cromosoma Y, así como secuencias de ADNmt de parientes. La otra base de datos se denomina *Base de Datos Dubitada* y contiene todos los perfiles genéticos de los restos encontrados que no han podido ser identificados por métodos clásicos (identificación de huellas dactilares, parámetros antropológicos, odontología, rayos X). La comparación los perfiles genéticos depositados en ambas bases ha permitido la identificación de cientos de restos (Lorente et al., 2001; <http://www.guardiacivil.org/prensa/actividades/FÉNIX/presentacion.jsp>).

Estrategias similares se han implantado en todo el mundo: Argentina (Banco Nacional de Datos; <http://www.mincyt.gov.ar/ministerio/banco-nacional-de-datos-geneticos-bndg-23>), Brasil (*Missing person Database-gen*; da Silva et al., 2009), la antigua Yugoslavia (ICMP, *International Commission of Missing Persons*), Colombia (FÉNIX-Colombia", Laboratorio de Genética del Centro Técnico de Investigación, Fiscalía General de la Nación), Chile (el Servicio Médico Legal) y EE.UU (*Texas Missing Persons DNA Database*, UNT Health Science Center).

La Iniciativa DNA-PROKIDS es un proyecto internacional diseñado para luchar contra el tráfico de seres humanos a través de la identificación genética de las víctimas y sus familiares, especialmente en niños (<http://www.dna-prokids.org/>). Este proyecto se inició en España y se ha ampliado a China, India, Indonesia, Filipinas, Nepal, Sri Lanka, Tailandia, Brasil, Guatemala y México. DNA-PROKIDS; el mismo emplea los marcadores genéticos de rutina, como los STR incluidos en el

CODIS, ADNmt y STR del cromosoma Y. Actualmente se están desarrollando paneles de SNPs que permitirán determinar características fenotípicas de los individuos, así como su procedencia geográfica y étnica. Sobre la base de este tipo de información genética se podrán crear bases de datos que permitirán la identificación de menores víctimas de tráfico humano y la reunificación de éstos con sus familias.

Las bases de datos de genética forense están mucho más desarrolladas en humanos que animales. Sin embargo, en esta última área de aplicación existen bases de datos equivalentes a las antes mencionadas:

**Bases de datos de los servicios de filiación.** Para la inscripción de reproductores en los libros genealógicos de las principales especies de animales de producción y de compañía, tales como bovinos, equinos, ovinos y perros, se requiere certificar la paternidad mediante análisis de ADN. Por ello, los laboratorios que realizan rutinariamente estos análisis poseen extensas bases de datos que permiten verificar la paternidad declarada o comparar un perfil genético con los previamente almacenados en la base de datos para resolver una filiación. Originalmente se almacenaban datos de grupos sanguíneos y de polimorfismos bioquímicos, y actualmente se incluyen los microsatélites. Estas bases también podrían ser empleadas como bases de datos poblacionales o de inteligencia. En Argentina, la mayor base de datos de este tipo, con varios miles de perfiles genéticos, se encuentra en el laboratorio de Genética Aplicada de la Sociedad Rural Argentina (SRA, <http://www.sra.org.ar/laboratorio/>), aunque no es de acceso libre. En el caso particular de los perros, la base de datos de esta especie es mantenida por el Laboratorio de Genética de Animales Domésticos del IGEVET (UNLP-CONICET LA PLATA), FCV- UNLP ([http://igevet.laplata-conicet.gov.ar/index.php?title=P%C3%A1gina\\_principal](http://igevet.laplata-conicet.gov.ar/index.php?title=P%C3%A1gina_principal)).

**Bases de datos para resolución de abigeato.** El robo de ganado o abigeato, así como el robo de mascotas, es un delito común en muchos

países. La creación de bases de datos poblacionales y de inteligencia puede ser de utilidad para resolver este tipo de delitos. En el año 2001, la Facultad de Ciencias Veterinarias de la UNLP y la provincia de Buenos Aires pusieron en marcha el Programa Provincial de Identificación Genética para la Prevención y Resolución de Casos de Abigeato en Ganado Mayor (Bovinos y Equinos). Este programa, que posteriormente fue extendido a otras regiones del país, contempló tanto la resolución de estos casos mediante análisis de ADN, como la implementación de bases de datos poblacionales que permitan la estimación de los índices forenses. Por otra parte, la custodia de muestras permite contar con muestras de referencia.

Finalmente, cabe mencionar que en el área de salud también se han implementado bases de datos que incluyen perfiles genéticos asociados a registros médicos y ambientales. Ejemplos de estos repositorios, denominados biobancos, son el Biobank del Reino Unido (<https://www.ukbiobank.ac.uk/>) y el de EE.UU (Henderson et al., 2013; <http://usbiobankstudy.web.unc.edu/>).

#### **4.3. Consideraciones finales**

Las bases de datos forenses han crecido ininterrumpidamente durante los últimos años, especialmente en la especie humana (Levitt, 2007). Así por ejemplo, países como China poseen aproximadamente 16 millones de perfiles genéticos humanos incorporados en sus bases de datos, Estados Unidos unos 10 millones de perfiles y el Reino Unido aproximadamente 6 millones de perfiles, siendo estos países los que mantienen las mayores bases de datos de perfiles de ADN humanos. Estos valores representan un alto porcentaje de la población: el 10% de la población del Reino Unido, el 0,9% de Alemania y el 0,8% de los Países Bajos están tipificados (Levitt, 2007).

Pero, ¿cuáles son los beneficios y riesgos de expandir las bases de datos genéticos? Desde el punto de vista técnico, el incremento de las bases de datos aumenta significativamente la probabilidad de que ocurran coincidencias (*match*) entre perfiles de ADN correspondientes a muestras de evidencias y de individuos inocentes (voluntarios) almacenados en las bases de datos (Gill et al., 2006). Estos eventos ya se han observado en repetidas ocasiones, tanto en bases de datos humanas como de animales domésticos, y ha llevado a recomendar un aumento en el número de los marcadores genéticos analizados y a utilizar métodos de cálculos más conservadores.

Por otra parte, el desarrollo de las bases de datos forenses ha hecho necesario discutir los criterios de inclusión y retención de muestras y perfiles de ADN, la eficiencia para la resolución de casos forenses, y el efecto sobre la privacidad de las personas. En las bases de datos forenses no humanas, este último punto de discusión (central para las bases de datos humanas) es reemplazado por la cuestión de la propiedad de las muestras y, por lo tanto, de los perfiles de ADN.

Los beneficios de las base de datos son evidentes, ya que en el pasado numerosos crímenes han podido ser resueltos mediante la comparación de los perfiles de ADN obtenidos de las evidencias y de aquellos depositados en los repositorios. Sin embargo, los beneficios de los análisis de ADN en genética forense tendrán necesariamente un costo social y ético, y será necesario maximizar los esfuerzo para minimizarlo (Lewitt, 2007).

#### **4.4. Referencias Bibliográficas**

Álvarez-Cubero M.J., L.J. Mtnez.-Gonzalez; M. Saiz, J.C. Álvarez, J.A. Lorente. 2010. Nuevas aplicaciones en identificación genética New applications in genetic identification Cuadernos de Medicina Forense Cuad. med. forense 16(1-2).

da Silva LAF, Vilaça W, Azevedo D, Majella G, Silva IF, Silva BF. Missing and unidentified persons database. *Forensic Sci Int: Genetics Supplement Series* 2009; 2:255-7.

Gill P., Fereday L., Morling N., Schneider PM. (2006) The evolution of DNA databases recommendations for new European STR loci. *Forensic Sci. Int.* 156, 242-244.

Grahn, R.A., Kurushima, J.D. , Billings, N.C., Grahn, J.C., Halverson, J.L., Hammera, E., Ho, C.K., Kun, T.J., Levy, J.K., Lipinski, M.J., Mwenda, J.M., Ozpinar, H., Schuster, R.K., Shoorijeh, S.J., Tarditi, C.R., Waly, N.E., Wictum, E.J., Lyons, L.A., (2011) “Feline non-repetitive mitochondrial DNA control region database for forensic evidence” en *Forensic Sci Int Genet.* Número 5(1), pp 33-42.

Halverson, J., Basten, C.A., (2005) “PCR multiplex and database for forensic DNA identification of dogs” en *Journal of Forensic Science*, Número 50, pp. 1–12.

Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, Davis AM, Zimmer C, Weiner BJ. Characterizing biobank organizations in the U.S.: results from a national survey, *Genome Medicine* 2013, 5:3 doi:10.1186/gm407.

Himmelberger AL, Spear TF, Satkoski JA, George DA, Garnica WT, Malladi VS, Smith DG, Webb KM, Allard MW, Kanthaswamy S, Forensic Utility of the Mitochondrial Hypervariable Region 1 of Domestic Dogs, in Conjunction with Breed and Geographic Information. *J Forensic Sci* 2008, 53(1), 81-89.

Kanthaswamy S, Bradley KT, Mattila AM, Johnston E, Dayton M, Kinaga J, Erickson BJ-A, Halverson J, Fantin D, DeNise S, Kou A Kanthaswamy, Malladi V, Satkoski J, Budowle B, Smith DG, Koskinen MT, Canine population data generated from a multiplex STR kit for use in forensic casework. *J Forensic Sci*, 2009 54, 829-840.

Levitt M: Forensic databases: benefits and ethical and social costs. *Br Med Bull* 2007, 83:235–248.

Lorente JA, Alvarez JC, Entrala C, Martinez-Espin E, Fernandez-Rosado F, Lorente M, et al. Bases de datos de ADN: Su uso en la investigación criminal y en la identificación civil. Anotaciones prácticas para su desarrollo. *Forénsica* 2002; 1:31-44.

Lorente JA, Entrala C, Alvarez JC, Arce B, Heinrichs B, Lorente M, et al. Identification of missing persons: The Spanish "Phoenix" program. *Croat Med J* 2001; 42:267-70.

Lorente JA, Entrala C, Alvarez JC, Lorente M, Arce B, Heinrich B, et al. Social benefits of non-criminal genetic databases: Missing persons and human remains identification. *Int J Legal Med* 2002; 116:187-90.

Parson W, Dür A: EMPOP - a forensic mtDNA database. *Forensic Sci Int Genet* 2007, 1:88–92.

Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, Lessig R, de Knijff P, Jobling MA, Tyler-Smith C, Krawczak M: Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Hum Genet* 2005, 116:279–291.

Willuweit S, Roewer L, International Forensic Y Chromosome User Group: Y chromosome haplotype reference database (YHRD): update. *Forensic Sci Int Genet* 2007, 1:83–87.34.