Universitat Rovira i Virgili
Máster en Genética, Física y Química Forense

Trabajo de Fin de Máster

MitoID: Desarrollo de una herramienta bioinformática para el análisis y reporte estandarizado de variantes en ADN mitocondrial humano a partir de secuencias FASTA en contextos académicos y forenses

Kevin Andrés Raza Santos

Tutor: Santi Garcia Vallvé. Departamento de Bioquímica y Biotecnología. URV

Tarragona - España

2025

Resumen

El estudio de mutaciones del ADN mitocondrial (ADNmt) es una herramienta forense indispensable en el establecimiento de linajes maternos e identificación humana, por su elevada abundancia idealmente en muestras altamente degradadas. Sin embargo, su análisis presenta desafíos significativos, incluyendo la detección fiable frente a artefactos como las secuencias mitocondriales nucleares (NUMTs), las heteroplasmias de bajo nivel, y persistentes inconsistencias en la estandarización de nomenclaturas. A pesar de la existencia de diversas herramientas bioinformáticas robustas, el panorama actual carece de una solución integrada, accesible y visualmente completa, óptima para fines formativos y de investigación

Para cubrir esta brecha, se desarrolló MitoID, una herramienta bioinformática de código abierto y ejecución local, diseñada para el análisis estandarizado y la comprensión visual de variantes en ADNmt humano a partir de archivos FASTA. Su pipeline en Python integró el alineamiento contra la Secuencia de Referencia de Cambridge Revisada (rCRS), se logró una correcta identificación de SNPs e indels. Se prestó especial atención al riguroso filtrado de artefactos como la posición 3107N, se integró lógicas específicas de regiones de interés y particularidades del ADNmt, como las regiones hipervariables (HVS), motivos repetitivos AC en np 515-525 y C-stretch en HVS-I y HVS-II. Crucialmente, MitoID estandarizó las variantes bajo nomenclaturas HGVS, Mitomaster y EMPOP/SWGDAM. Además, generó informes detallados y un visualizador interactivo que permite a los usuarios observar la posición de cada mutación en relación con las regiones codificantes o de control del genoma, ofreciendo una perspectiva didáctica de los perfiles genéticos

MitoID contribuye a la estandarización y accesibilidad del análisis de ADNmt, ofreciendo una alternativa complementaria y pedagógica. Su diseño y disponibilidad como software de código abierto lo convierten en un recurso valioso para la formación e investigación, facilitando la comprensión de las variantes mitocondriales

Link de acceso: https://github.com/Andres-ADN/MitoID.git

Palabras clave: ADN mitocondrial, genética forense, bioinformática, detección de variantes ADNmt, visualización de datos, Python.

Abstract

The study of mitochondrial DNA (mtDNA) mutations constitutes an indispensable forensic tool in establishing maternal lineages and human identification, due to its high abundance, ideally in highly degraded samples. However, its analysis presents significant challenges, including reliable detection against artifacts like nuclear mitochondrial insertions (NUMTs), low-level heteroplasmies, and persistent inconsistencies in nomenclature standardization. Despite the existence of diverse robust bioinformatics tools, the current landscape lacks an integrated, accessible, and visually comprehensive solution, optimal for training and research purposes.

To bridge this gap, MitoID was developed, an open-source, locally executable bioinformatics tool designed for the standardized analysis and visual comprehension of mtDNA variants from FASTA files. Its Python pipeline integrated alignment against the revised Cambridge Reference Sequence (rCRS), achieving correct identification of SNPs and indels. Special attention was paid to rigorous artifact filtering, such as the ambiguous 3107N position, and specific logic for regions of interest and mtDNA peculiarities was incorporated, including hypervariable regions (HVS), AC repetitive motifs at np 515-525, and C-stretches in HVS-I and HVS-II. Crucially, MitoID standardized variants under HGVS, Mitomaster, and EMPOP/SWGDAM nomenclatures. Furthermore, it generated detailed reports and an interactive variant visualizer that allows users to observe the position of each mutation in relation to the genome's coding or control regions, offering a didactic perspective on genetic profiles.

MitoID contributes to the standardization and accessibility of mtDNA analysis, providing a complementary and pedagogical alternative. Its design and availability as open-source software make it a valuable resource for training and research, facilitating the understanding of mitochondrial variants.

Link to access: https://github.com/Andres-ADN/MitoID.git

Key words: Mitochondrial DNA, Forensic genetics, Bioinformatics, mtDNA variant detection, Data visualization, Python.

Índice de contenido

1 Introducción 1

1.1 ADN mitocondrial 1

1.2 Secuencia de referencia de cambridge (rCRS) y su contexto forense 5

1.3 Estándares de Nomenclatura y Bases de Datos en el Análisis Forense de ADNmt 5

1.4 Necesidad de una Nueva Herramienta Bioinformática para el ADNmt Forense ... 6

2 Objetivos 10

2.1 Objetivo general 10

2.2 Objetivos específicos 10

3 Metodología 10

3.1 Entorno de desarrollo 10

3.2 Elección de Librerías esenciales para el tratamiento de ADNmt. 11

3.3 Datos de referencia y definición de características del genoma. 11

3.4 Metodología de Análisis del ADNmt en MitoID 12

3.4.1 Preprocesamiento y Alineamiento de Secuencias 12

3.4.2 Detección y Filtrado de Variantes 14

3.4.3 Estandarización, Anotación de variantes 14

3.4.4 Manejo de diferencias de estilos 15

3.4.5 Manejo de Regiones Homopoliméricas y Casos Especiales 17

3.4.6 Anotación del Locus Genómico de las Variantes 18

3.4.7 Visualización de resultados: 18

3.5 Flujo de trabajo 19

4 Resultados y discusión 20

5 Conclusiones: 28

6 Bibliografía 29

7 Anexos 32

Índice de figuras

figura 1. Análisis comparativo de herramientas bioinformáticas clave para el ADNmt forense 8
Figura 2. Diagrama de flujo de lógicas de funcionamiento MitoID..... 19
figura 3 Salida principal de MitoID en formato HTML..... 21
figura 4 Inserción HVS-II salida de MitoID..... 23
figura 5. Inserción HVS-I salida de MitoID..... 23
figura 6. deleción 294 salida de MitoID..... 24
figura 7. Región repetida AC salida MitoID..... 24
figura 8. Comparación EMPOP inserción compleja..... 25
figura 9. Comparación EMPOP deleción compleja..... 25
figura 10 Salida del visualizador de variantes de MitoID en formato HTML 26
figura 11 zoom de la salida del visualizador de variantes sobre una variante..... 27

Índice de tablas

Tabla 1. Características esenciales de rCRS ADNmt utilizadas en el diseño de MitoID 1
Tabla 2 Uso de ADNmt en Casos forenses relevantes..... 3
Tabla 3. Representación de variantes de ADNmt en diferentes estándares de nomenclatura. 17
Tabla 4. Regiones de interés y casos especiales en el análisis de ADNmt..... 17

1 Introducción

1.1 ADN mitocondrial

El ADNmt humano es una molécula circular de doble cadena, de 16.569 pb presente en cientos o incluso miles de copias por célula (a excepción de los glóbulos rojos)[1]. Esta elevada abundancia le otorga resistencia a la degradación y lo convierte en un marcador robusto para la recuperación de material genético de muestras biológicas con bajo contenido de ADN o altamente degradadas, como tallos de cabello sin raíz, huesos antiguos y dientes[2]. Su herencia es estrictamente matrilineal, lo que lo hace invaluable para establecer linajes y relaciones de parentesco a lo largo de generaciones, además de simplificar la reconstrucción de su historia evolutiva debido a la ausencia de recombinación entre individuos relacionados.

La organización del ADNmt se distingue por estructurarse únicamente de regiones codificantes[1], entre estos, 37 genes que codifican 13 proteínas esenciales para la fosforilación oxidativa, 22 ARN de transferencia (ARNt) y 2 ARN ribosomales (ARNr), como se detalla de manera explícita en la siguiente tabla.

Tabla 1. Características esenciales de rCRS ADNmt utilizadas en el diseño de MitoID

Gen	**Región de control (D-loop)**	
	Extensión / Localización	Nombre común (Nombre para MitoID)
Región Hipervariable II	57 – 372 (H/L)	HV2
Región Hipervariable III	438 – 574 (H/L)	HV3
Región Hipervariable I	16024 – 16383 (H/L)	HV1
Gen	Genes codificantes de proteínas	
	Extensión / Localización	Nombre común (Nombre para MitoID)
MT-ND1	3307 – 4262 (H)	NADH deshidrogenasa, subunidad 1 (ND1)
MT-ND2	4470 – 5511 (H)	NADH deshidrogenasa, subunidad 2 (ND2)
MT-CO1	5904 – 7445 (H)	Citocromo c oxidasa, subunidad 1 (CO1)
MT-CO2	7586 – 8269 (H)	Citocromo c oxidasa, subunidad 2 (CO2)
MT-ATP8	8366 – 8572 (H)	ATP sintasa, Fo subunidad 8 (ATP8)
MT-ATP6	8527 – 9207 (H)	ATP sintasa, Fo subunidad 6 (ATP6)
MT-CO3	9207 – 9990 (H)	Citocromo c oxidasa, subunidad 3 (CO3)
MT-ND3	10059 – 10404 (H)	NADH deshidrogenasa, subunidad 3 (ND3)
MT-ND4L	10470 – 10766 (H)	NADH deshidrogenasa, subunidad 4L (ND4L)
MT-ND4	10760 – 12137 (H)	NADH deshidrogenasa, subunidad 4 (ND4)

MT-ND5	12337 – 14148 (H)	NADH deshidrogenasa, subunidad 5 (ND5)
MT-ND6	(14149...14673) (L)	NADH deshidrogenasa, subunidad 6 (ND6)
MT-CYB	14747 – 15887 (H)	Citocromo b (CYTB)
Gen	Genes de ARNt	
	Extensión / Localización	Nombre común (Nombre para MitoID)
MT-TF	577 – 647 (H)	Fenilalanina (Phe)
MT-TV	1602 – 1670 (H)	Valina (Val)
MT-TL1	3230 – 3304 (H)	Leucina (Leu-UUR)
MT-TI	4263 – 4331 (H)	Isoleucina (Ile)
MT-TQ	(4329...4400) (L)	Glutamina (Gln)
MT-TM	4402 – 4469 (H)	Metionina (Met)
MT-TW	5512 – 5579 (H)	Triptófano (Trp)
MT-TA	(5587...5655) (L)	Alanina (Ala)
MT-TN	(5657...5729) (L)	Asparagina (Asn)
MT-TC	(5761...5826) (L)	Cisteína (Cys)
MT-TY	(5826...5891) (L)	Tirosina (Tyr)
MT-TS1	(7446...7514) (L)	Serina (Ser-UCN)
MT-TD	7518 – 7585 (H)	Ácido Aspártico (Asp)
MT-TK	8295 – 8364 (H)	Lisina (Lys)
MT-TG	9991 – 10058 (H)	Glicina (Gly)
MT-TR	10405 – 10469 (H)	Arginina (Arg)
MT-TH	12138 – 12206 (H)	Histidina (His)
MT-TS2	12207 – 12265 (H)	Serina (Ser-AGY)
MT-TL2	12266 – 12336 (H)	Leucina (Leu-CUN)
MT-TE	(14674...14742) (L)	Ácido Glutámico (Glu)
MT-TT	15888 – 15953 (H)	Treonina (Thr)
MT-TP	(15956...16023) (L)	Prolina (Pro)
Gen	Genes de ARNr	
	Extensión / Localización	Nombre común (Nombre para MitoID)
MT-RNR1	648 – 1601 (H)	ARN ribosomal 12S (12S)
MT-RNR2	1671 – 3229 (H)	ARN ribosomal 16S (16S)

Nota: genes codificantes representados con "(localización)" son solapamientos que se toman en cuenta para integrar lógicas que prioricen la anotación de las variantes en la región más funcionalmente relevante cuando una mutación cae en un solapamiento.

Elaborado por: (El autor, 2025)

La región de control (CR) o D-loop, es una porción no codificante crucial, que destacada por su alta tasa de mutación y su importancia para la discriminación de haplotipos en contextos

forenses y poblacionales[3]. Dentro de la CR se encuentran los segmentos hipervariables (HVS-I, HVS-II y HVS-III), los cuales concentran gran parte de la variabilidad genética del ADNmt[4], misma que se manifiesta principalmente como polimorfismos de un solo nucleótido (SNPs) que consisten en la sustitución de un único nucleótido (clasificados como transiciones o transversiones) e inserciones/deleciones (Indels) que implican la adición o pérdida de uno o más nucleótidos.

En este sentido, un haplotipo de ADNmt (o perfil de ADNmt) se define como la combinación única de las mutaciones nucleotídicas observadas a lo largo de las diferentes regiones de la secuencia de ADNmt de un individuo[5]. Dada la característica de herencia uniparental y la ausencia de recombinación entre individuos relacionados, los haplotipos se perpetúan a lo largo de los linajes maternos, y la acumulación de mutaciones específicas en estos linajes permite la formación de un haplogrupo[1], es decir, un grupo de individuos que comparten un ancestro materno común[6]. Esta singularidad y la identificación de haplogrupos son esenciales para la identificación humana y el establecimiento de relaciones de parentesco en el ámbito de las ciencias forenses[7].

El ADN mitocondrial (ADNmt) en contextos forenses, ha tenido especial relevancia en casos particulares conocidos a lo largo de los últimos 40 años, algunos de estos casos se recopilan en la Tabla 2.

Tabla 2 Uso de ADNmt en Casos forenses relevantes

Objetivo del análisis	**Muestras / regiones estudiadas**	**Muestras utilizadas como ID**	**Resultados clave**
Identificación de restos esqueléticos de un niño desaparecido[8]	Restos esqueléticos de un niño de 3 años, encontrados en el desierto (HV1 y HV2 de la región de control)	Madre del niño desaparecido	El ADNmt de los restos y el de la madre eran idénticos con los del niño desaparecido[8]. Este fue uno de los primeros casos que demostró el valor del ADNmt para vincular restos biológicos con individuos desaparecidos y su potencial en casos de agresión sexual[1]

Identificación de los restos del Zar Nicolás II de Rusia[9]	restos humanos atribuidos al Zar Nicolás II (posición 16,169)	Restos de su hermano, Georgij Romanov	Este caso destacó la utilidad de la heteroplasmia para distinguir entre individuos relacionados maternalmente y apoyar la identificación de restos humanos[2,9], especialmente en muestras desafiantes
Víctimas de la Dictadura Argentina[10]	340 restos esqueléticos de personas desaparecidas	No especificado directamente	El propósito era la identificación de víctimas de violaciones a los derechos humanos[10]
Víctimas de la guerra civil española[11]	252 restos esqueléticos de personas encontradas en fosas comunes	Una base de datos de 186 familiares de posibles víctimas aportada para la identificación mediante relación parental	Se logró la identificación de 87 de los 252 restos, El ADNmt resultó ser una herramienta valiosa para la identificación humana en este contexto, donde los restos a menudo estaban degradados[2,11]
Casos de Personas Desaparecidas con Restos Altamente Degradados[7]	Restos humanos altamente degradados (protocolo de 11 SNP y HV1 y HV2 de la región de control)	Familias de referencia	El propósito era excluir o confirmar la identidad, y clasificar elementos esqueléticos mezclados, se demostró la utilidad de los SNP de ADNmt para distinguir muestras con tipos HV comunes y como una opción más eficiente que la secuenciación adicional de la CR en muestras degradadas[7]

Nota: la tabla muestra el propósito del ADNmt durante accidentes masivos o casos con ADN altamente degradado, donde el establecimiento de linajes o relaciones parentales es el principal objetivo de su uso en el contexto forense con fines identificativos.

Un fenómeno particular a estudiar del ADNmt es la heteroplasmia, que se define como la presencia de más de una secuencia de ADNmt en un individuo (puntual o de longitud)[12]. Puede ser un desafío, pero también un elemento valioso para aumentar el poder de discriminación del análisis forense[13]. En ese sentido las herramientas bioinformáticas deben

ser capaces de detectar y cuantificar la heteroplasmia de bajo nivel, siendo que, la asignación de haplogrupos, basada en patrones de variantes específicas, no solo sirve como una herramienta de control de calidad, sino que también ofrece consideraciones filogeográficas y puede proporcionar pistas de investigación valiosas[13,14], por ejemplo, la heteroplasmia fue clave en la confirmación de la identidad de los restos del Zar Nicolás II de Rusia[15] (Tabla 2).

1.2 Secuencia de referencia de cambridge (rCRS) y su contexto forense

La primera secuencia completa del ADNmt humano, conocida como la Secuencia de Referencia de Cambridge (CRS) o "secuencia de Anderson", fue descrita en 1981 y numerada arbitrariamente de 1 a 16,5697[2]. En 1999, la CRS fue secuenciada nuevamente, dando origen a la Secuencia de Referencia de Cambridge Revisada (rCRS), se ha consolidado como el estándar internacional para la comparación y el reporte de haplotipos de ADNmt en el ámbito forense y científico[5], disponible en NCBI como (NC_012920.1). La rCRS se ha mantenido como referencia por su estabilidad y para evitar complejidades asociadas a la traducción de datos preexistentes a nuevas notaciones[7].

Un aspecto crítico en la rCRS, que exige particular atención en el análisis bioinformático, es la ambigüedad en la posición 3107, representada por una 'N'[5]. Esta particularidad puede generar interpretaciones erróneas o artefactos en la detección de variantes, como deleciones (m.3107delN) o sustituciones (m.N3107X)[5], si no se gestiona de forma adecuada durante el procesamiento de datos, es por eso que la necesidad de filtrar estos artefactos es fundamental para la fiabilidad de los perfiles genéticos.

Aun con el surgimiento de tecnologías de secuenciación más avanzadas, la práctica forense estándar en muchos laboratorios todavía emplea la secuenciación Sanger y el análisis de secuencias consenso en formato FASTA[1,3]. Este formato, valorado por su simplicidad y amplia disponibilidad, es fundamental para comparar las secuencias de las muestras con la rCRS y obtener perfiles de variantes precisos para análisis de rutina. Por lo tanto, el diseño de herramientas que soporten esta entrada de datos es esencial para la práctica forense actual y para la estandarización de los resultados[16].

1.3 Estándares de Nomenclatura y Bases de Datos en el Análisis Forense de ADNmt

La estandarización de la nomenclatura y el uso de bases de datos comunes son fundamentales para la comunicación clara y la comparabilidad de los resultados en la genética forense.

HGVS (Human Genome Variation Society) proporciona un marco internacional como el estándar para la convención global para describir variantes genéticas, y su aplicación es crucial para una representación precisa de las mutaciones[5]. En esta misma línea de investigación, las directrices del SWGDAM (Scientific Working Group on DNA Analysis Methods)[17] complementan estos estándares de HGVS, proporcionando recomendaciones para el reporte de variantes de ADNmt con nomenclaturas específicas[17] que son usadas por bases de datos como Mitomaster y EMPOP (EDNAP Mitochondrial DNA Population Database) son referencias clave[2]. EMPOP, utiliza fielmente la anotación de variantes propuesta por la SWGDAM, esta aproximación se fundamenta en los patrones de mutación establecidos de la filogenia mitocondrial[18]. lo que permite estimar tasas de mutación posicionales y una interpretación más fundamentada de la evidencia, determinando la frecuencia del haplotipo en bases de datos de población relevantes[18].

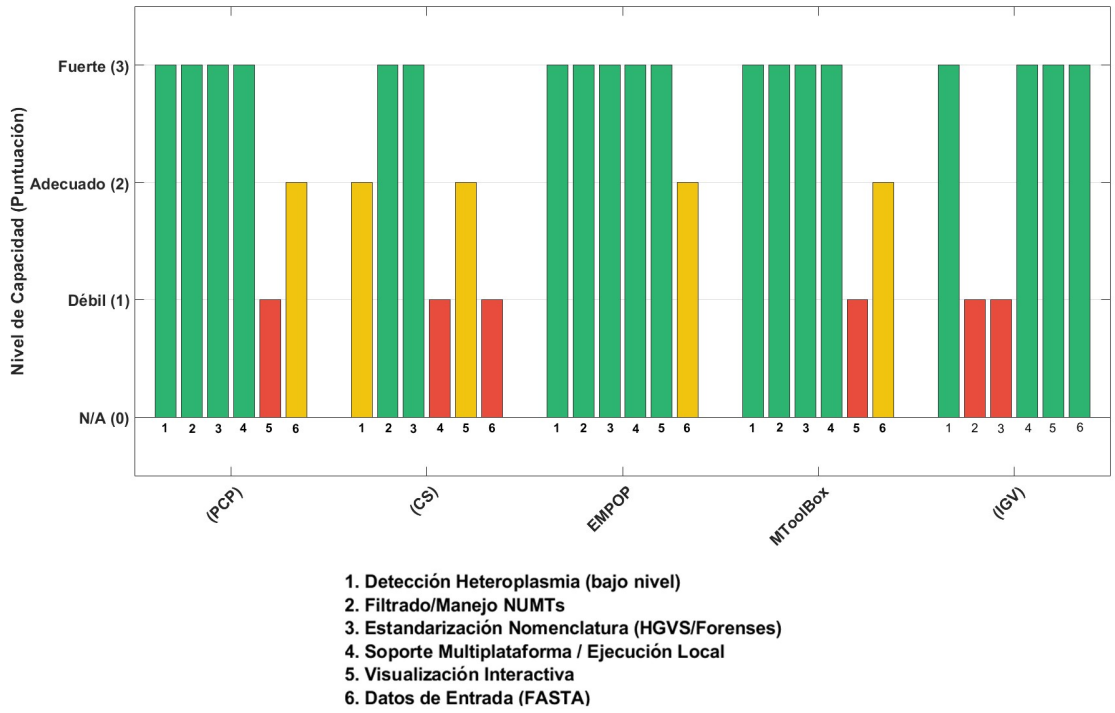
Mientras que Mitomaster posee capacidades clave que abordan las complejidades de la variación del mtDNA[19,20], utiliza una versión "modified Cambridge Reference Sequence" (mCRS). Una característica distintiva de esta mCRS es que, aunque ha sido corregida de errores, mantiene las "errores de inserción-deleción" para preservar el sistema de numeración de nucleótidos original, Su enfoque se centra en interpretar nuevas secuencias de ADNmt con confianza al compararlas con datos extensos ya existentes en la literatura, por esto, utiliza un sistema de nomenclatura de variantes diferente al de EMPOP, con una base de datos con información sobre variantes de mtDNA publicadas demasiado amplia y útil, principalmente en los contextos clínicos y forenses[20].

1.4 Necesidad de una Nueva Herramienta Bioinformática para el ADNmt Forense

La gran cantidad y diversidad de datos generados en el análisis de ADNmt requieren herramientas bioinformáticas especializadas para su análisis e interpretación[4]. Si bien existen diversas herramientas en el panorama actual, algunas de las cuales son líderes en el ámbito forense y de investigación. Sin embargo, al examinar sus capacidades de manera integral, se hace evidente que ninguna herramienta existente cubre de forma exhaustiva todas las necesidades críticas para un flujo de trabajo forense completo y accesible, especialmente para estudiantes e investigadores que requieren una comprensión profunda de la manipulación de variantes.

Considerando las funcionalidades esenciales para el análisis de ADNmt, como la detección precisa de heteroplasmia de bajo nivel, el filtrado efectivo de NUMTs, la estandarización de la nomenclatura (HGVS y formatos forenses como EMPOP y Mitomaster), el soporte multiplataforma con ejecución local, la visualización interactiva de resultados y la compatibilidad con formatos de entrada comunes como FASTA, se observa un patrón. Herramientas comerciales como Converge Forensic Analysis Software (CS)[21], aunque robustas en la detección de variantes y el manejo de NUMTs, a menudo presentan limitaciones en la detección de heteroplasmia de bajo nivel y no ofrecen soporte multiplataforma o ejecución local fuera de sus ecosistemas propietarios. Por otro lado, pipelines de código abierto como Precision Caller Pipeline (PCP)[4], que destacan por su optimización en la detección de heteroplasmia y el filtrado de NUMTs, suelen carecer de una visualización interactiva integrada, esencial para una interpretación pedagógica, y su soporte de entrada FASTA para el análisis directo de la secuencia completa es limitado. Herramientas de base de datos como EMPOP[22], fundamentales para el control de calidad filogenético y la estandarización de nomenclatura, sobresalen en visualización de datos poblacionales, pero no son primariamente plataformas de análisis de variantes de secuencias de consulta en FASTA. De manera similar, MToolBox[23], un pipeline automatizado y de código abierto con fuertes capacidades en detección y estandarización, presenta limitaciones en la visualización interactiva, una característica de creciente demanda para la validación intuitiva de resultados. Finalmente, aunque herramientas como Integrative Genomics Viewer (IGV)[24] son indispensables para la visualización manual y la confirmación de variantes, no son soluciones integrales que automaticen todo el proceso de detección y estandarización de forma autónoma. Este análisis subraya la necesidad de una herramienta que integre de manera sinérgica todas las capacidades, como se puede visualizar en la figura 1.

figura 1. Análisis comparativo de herramientas bioinformáticas clave para el ADNmt forense



Nota: Los números en la base de cada barra indican el criterio evaluado, el color de cada barra representa el nivel de capacidad de la herramienta para el criterio correspondiente: **Verde**: Capacidad Fuerte (3 puntos); **Amarillo**: Capacidad Adecuada (2 puntos); **Rojo**: Capacidad Débil (1 punto); **Gris Claro**: No aplica o no es el foco (0 puntos), información basada en la recopilación bibliográfica de los autores de las aplicaciones[4,21,22,24].

Además de las herramientas destacadas en este análisis, el ecosistema bioinformático del ADNmt incluye otras soluciones valiosas, aunque con enfoques más especializados o con ciertas limitaciones para un uso generalizado en todos los contextos forenses y académicos. Por ejemplo, Mitopore[13] y MitoSAlt[13] son reconocidas por su alta precisión en la detección de SNVs y deleciones a gran escala, respectivamente, pero pueden carecer de un paso explícito para la exclusión de NUMTs o de reporte directo de heteroplasmia. MitoMut[25] se especializa en la detección y cuantificación de deleciones, particularmente a bajos niveles de heteroplasmia. MitoRS[26] y GATK Mutect2[27] ofrecen detección de variantes de baja frecuencia, aunque con consideraciones sobre falsos positivos o adaptaciones del ADN nuclear. Herramientas como HaploGrep[22] y mtDNA-Server[6] son primordiales para la clasificación de haplogrupos y la detección de contaminación, mientras que MitoVisualize[28]

se enfoca en la visualización de variantes en estructuras de ARN. Finalmente, soluciones como Ion Torrent Suite Software (TSS)[21], aunque proporcionadas por fabricantes, a menudo utilizan referencias modificadas o umbrales de detección que limitan su interoperabilidad y la detección de variantes de bajo nivel con herramientas de código abierto. Esta diversidad de herramientas, cada una con un nicho específico o con ciertas barreras de acceso y usabilidad, refuerza la brecha en la disponibilidad de una solución integral, accesible y adaptada a las necesidades formativas y de investigación forense.

Ante esta situación, el presente trabajo se enfoca en el desarrollo de MitoID, una herramienta bioinformática de código abierto y ejecución local. Esta solución se propone como una alternativa accesible y visualmente intuitiva a las plataformas existentes, buscando unificar funcionalidades diversas para el análisis y anotación de variantes de ADN mitocondrial de manera rápida, fiable y completa. Entre sus capacidades clave, que permita la detección de Polimorfismos de un Solo Nucleótido (SNPs) e inserciones/deleciones (indels) mediante el alineamiento de secuencias contra la Secuencia de Referencia de Cambridge Revisada (rCRS), incorporando el filtrado de artefactos relevantes. Además, que ofrezca la visualización del alineamiento de los nucleótidos y su posición respecto a los genes del ADNmt, facilitando una comprensión más profunda de los perfiles genéticos.

2 Objetivos

2.1 Objetivo general

Desarrollar e implementar una herramienta bioinformática de escritorio, denominada MitoID, que sea de código abierto y funcione offline, para el análisis estandarizado de variantes en secuencias de ADN mitocondrial humano a partir de archivos en formato FASTA, orientada a su aplicación en contextos académicos y forenses.

2.2 Objetivos específicos

- Construir un pipeline bioinformático en Python dedicado procesamiento de secuencias de ADNmt en formato FASTA, que incluya la carga de datos, el alineamiento con la secuencia de referencia (NC_012920.1), la extracción de variantes (SNPs e indels), con un sistema de filtrado de artefactos establecidos, y la anotación de los locus genómicos afectados por cada variante.
- Implementar la normalización y formateo variantes detectadas según los estándares de nomenclatura HGVS, el perfile de variantes tipo Mitomaster y el perfil de variantes tipo EMPOP, incluyendo el manejo de particularidades como las regiones “hotspots”.
- Desarrollar un sistema para la generación de informes detallados en formatos HTML y PDF que presenten de manera clara y visual los resultados del formateo de variantes y alineamiento (incluyendo un informe dedicado al visualizador de variantes), y asegurar la validación, accesibilidad y reproducibilidad de la herramienta MitoID mediante su distribución como software de código abierto con la documentación necesaria (README.md) en una plataforma como GitHub.

3 Metodología

3.1 Entorno de desarrollo

El desarrollo de MitoID se llevó a cabo en un entorno multiplataforma, utilizando Python 3.12 como lenguaje de programación principal. La elección de Python se fundamentó en su amplia adopción en el ámbito bioinformático, la riqueza de sus librerías especializadas y su versatilidad para la manipulación de datos y la automatización de procesos. Para la gestión del entorno y las dependencias, se empleó el sistema de Anaconda, y para asegurar la

compatibilidad con ciertas librerías y facilitar el desarrollo en un entorno Linux-like bajo Windows, se trabajó con Windows Subsystem for Linux

3.2 Elección de Librerías esenciales para el tratamiento de ADNmt.

Las librerías principales de Python, esenciales para las funcionalidades de MitoID, incluyen:

- **Biopython**: librería de código abierto, pilar fundamental para el manejo y análisis de datos biológicos computacionales. Proporciona un conjunto de herramientas y módulos para abordar una amplia gama de problemas bioinformáticos[29], entre los cuales se utilizó para los siguientes propósitos:
    - **Manipulación de secuencias**: Bio.SeqIO permitió la lectura y escritura de diversos formatos de archivos de secuencias biológicas, como FASTA y GenBank, junto con el módulo SeqRecord que aumentan los objetos Seq (secuencias leídas) con propiedades como el nombre, identificador y descripción, y pueden contener anotaciones de características de la secuencia
    - **Alineamiento de secuencias**: el módulo Bio.Align fue empleado para realizar alineamientos de secuencias de ADNmt
    - **Procesamiento de características genómicas**: el módulo Bio.SeqFeature se encarga de la extracción de información anotada en archivos GenBank, como la ubicación y tipo de genes, ARNt, ARNr y la D-loop
- **Pandas**: Empleada para el manejo eficiente de datos tabulares, lo que facilita la creación y manipulación de DataFrames que sirven como base para la generación de informes detallados.
- **HGVS**: estándar global para describir y comunicar variantes en secuencias de ADN, ARN y proteínas en genómica clínica y de investigación[30], facilitó la normalización y el parseo de las descripciones formales de las variantes.
- **WeasyPrint**: Utilizada para la conversión programática de los informes generados en formato HTML a documentos PDF, garantizando una salida profesional y portable.

3.3 Datos de referencia y definición de características del genoma.

Para la adquisición y preparación de las secuencias utilizadas en el desarrollo y validación, se emplearon diversas fuentes y herramientas. Las secuencias de prueba de ADNmt, incluyendo haplotipos específicos y mutaciones descritas en la literatura, se recolectaron de

repositorios públicos como NCBI o de bases de datos de referencia como Mitomaster[19] (Most Frequent Variants in Mitomap) y el manual de usuario de EMPOP[18]. Para la edición de secuencias de casos especiales o aquellas mencionadas en la bibliografía, se utilizó el software MEGA 11.

La Secuencia de Referencia de Cambridge Revisada (rCRS), con número de acceso NC_012920.1, constituyó la referencia principal para el alineamiento y la anotación de variantes. Esta secuencia fue obtenida en formatos FASTA (NC_012920.1_rCRS.fasta) y GenBank (NC_012920.1_rCRS.gb).

La definición de las características genómicas de la rCRS es esencial para la anotación de variantes, incluyendo su tipo y nombres comunes, características que se detallan en Tabla 1, se realizó procesando el fichero de GenBank para obtener la información necesaria sobre los genes, ARNt, ARNr y la D-loop.

3.4 Metodología de Análisis del ADNmt en MitoID

3.4.1 Preprocesamiento y Alineamiento de Secuencias

Posterior al procesamiento de secuencias gracias a los módulos integrados de Biopython, para asegurar la integridad de los datos procesados, se incorporaron mecanismos de manejo de errores que notifican al usuario y abortan el proceso en caso de que los archivos FASTA de referencia o de consulta no sean encontrados o estén malformados.

El corazón de la detección de variantes en MitoID reside en el algoritmo de alineamiento de secuencias, donde se comparan las secuencias denominadas durante el procesamiento de carga como “query_seq” a la secuencia de caso de estudio y como “rcrs_seq” a la secuencia de referencia, donde para identificar correspondencias exactas y diferencias, se analiza la variabilidad en la longitud y la calidad de las secuencias de entrada, MitoID implementa una lógica adaptable para la selección del modo de alineamiento:

- **Alineamiento Global (Needleman-Wunsch):** Este modo se emplea cuando la longitud de la secuencia de consulta es significativamente comparable a la de la rCRS (es decir, si la secuencia de consulta posee al menos el 80% de la longitud total de la rCRS). En este escenario, el algoritmo intenta alinear la totalidad de ambas secuencias, buscando la máxima similitud a lo largo de toda su extensión.

- **Alineamiento Local (Smith-Waterman):** Este modo se emplea si la longitud de la secuencia de consulta es inferior al 80% de la rCRS y con un límite mínimo establecido, ya que no se puede definir un alineamiento óptimo con menos longitud que 200pb. Este algoritmo está diseñado para encontrar las regiones de mayor similitud entre las dos secuencias, incluso si solo una porción de la consulta se alinea bien con la referencia, lo cual es particularmente útil en el análisis forense donde las muestras pueden estar fragmentadas.

Para asegurar la precisión y la pertinencia biológica del alineamiento en el contexto del ADN mitocondrial humano, se configuró una matriz de sustitución personalizada. Esta matriz se diferencia de las genéricas (como PAM o BLOSUM, optimizadas para proteínas o secuencias de ADN con mayores divergencias evolutivas) para adaptarse a la alta conservación general del genoma mitocondrial humano y la naturaleza de las variaciones que se buscan. La matriz de sustitución empleada asigna las siguientes puntuaciones:

- **Coincidencia (Match):** +3.0, para maximizar la alineación de bases idénticas.
- **No Coincidencia (Mismatch):** −3.0, para desincentivar sustituciones.
- **Bases Ambiguas ('N') vs. Bases Canónicas (A, T, C, G):** +1.0, para evitar penalizar excesivamente posiciones inciertas y permitir el flujo del alineamiento.
- **'N' vs. 'N':** +1.0, dado que no representa una divergencia significativa.

Adicionalmente para el alineamiento, mediante una matriz de sustitución se definieron las penalizaciones por la introducción y extensión de huecos (gaps):

- **Penalización por Apertura de Gap (Open Gap Score):** −7.0, un valor relativamente alto para desfavorecer la introducción de nuevos gaps, dado que las inserciones y deleciones son eventos menos frecuentes que las sustituciones en el ADNmt.
- **Penalización por Extensión de Gap (Extend Gap Score):** −2.0, un valor menor que la apertura para reflejar la tendencia biológica de que, si un evento de inserción o deleción ocurre, es más probable que afecte a varias bases contiguas.

3.4.2 Detección y Filtrado de Variantes

La lógica principal de la detección radicó en la comparación posicional exhaustiva de estas dos secuencias alineadas. el software desarrollado recorrió sistemáticamente el alineamiento, examinando cada par de bases enfrentadas evaluando los siguientes escenarios:

- **Correspondencia Directa de Bases**: Si la base de la rCRS y la base de la secuencia query en una posición dada son idénticas, no se considera una variante.
- **Diferencia de Bases (No-Coincidencia)**: Cuando la base de la rCRS difiere de la base de la secuencia query, se registra una variación genética, base para la detección de polimorfismos de un solo nucleótido (SNPs). Incluso si una de las bases es ambigua ('N'), el sistema lo reconoce como una diferencia potencial para su posterior manejo. Para cada SNP detectado, se realiza la tipificación de transiciones y transversiones, clasificando las sustituciones según el intercambio de purinas/pirimidinas.
- **Presencia de Gaps**: Un guion ('-') en la secuencia de referencia indica una inserción en la secuencia query, mientras que un guion en la secuencia query denota una deleción respecto a la rCRS. La lógica implementada agrupa guiones contiguos para identificar eventos de cambio de longitud de múltiples bases.

Esta identificación inicial generó una lista de "variantes crudas", las cuales son posteriormente sometidas a un filtrado de artefactos específicos para mejorar la precisión de la detección como es el caso del manejo de la posición 3107, esta posición en la rCRS está representada por una 'N' con el fin de preservar la numeración histórica del genoma mitocondrial tras una deleción detectada en la secuencia original de Cambridge. Dada esta particularidad, las variaciones en esta región pueden generar interpretaciones erróneas o artefactos en la detección de variantes.

3.4.3 Estandarización, Anotación de variantes

La nomenclatura de variantes de ADNmt fue un aspecto fundamental para asegurar la consistencia y comparabilidad de los resultados en genética forense, por este motivo la estandarización de variantes se realizó por la librería HGV que define los siguientes lineamientos:

- Prefijo: Se utiliza el prefijo "m." para indicar variantes en ADN mitocondrial.

- Coordenadas: Las posiciones se especifican como coordenadas ordinales de secuencia.
- Regla del desplazamiento 3' (justificación 3' rule): Para todas las descripciones de variantes, la posición más 3' posible de la secuencia de referencia se asigna arbitrariamente como el lugar del cambio.
- Sustituciones: Describen el reemplazo de un nucleótido por otro (ej., pos REF>ALT). MitoID gestiona la notación de transiciones y transversiones bajo esta categoría. Las sustituciones que involucran dos o más nucleótidos consecutivos se describen como deleción/inserción (InDels).
- Deleciones (del): Indican la ausencia de uno o más nucleótidos. Se especifican por la posición o rango seguido de "del".
- Inserciones (ins): Describen la adición de uno o más nucleótidos donde la inserción no es una copia de una secuencia inmediatamente 5'. Se usa el rango de posiciones adyacentes seguido de "ins" y la secuencia insertada.
- Duplicaciones (dup): Indican la inserción de una copia de uno o más nucleótidos directamente al extremo 3' de la copia original de esa secuencia. Se describe la posición o rango seguido de "dup". Si un cambio puede describirse como duplicación, debe priorizarse sobre una inserción.
- Deleción-Inserción (delins): Representan el reemplazo de uno o más nucleótidos por uno o más diferentes, cuando no pueden clasificarse como sustitución o duplicación. Se utiliza la posición o rango seguido de "delins" y la secuencia reemplazante.
- Secuencias Repetidas: Describen un segmento de uno o más nucleótidos (la unidad de repetición) presente varias veces consecutivas. Esta notación es particularmente relevante para las variaciones en los tractos de polibases del ADNmt.

3.4.4 Manejo de diferencias de estilos.

El manejo de diferencias de estilos se fundó en la base de nomenclatura propuesta por SWGDAM[17] que sigue lo siguientes lineamientos:

- Sustituciones: Se describen utilizando las convenciones IUPAC con letras mayúsculas (ej., 16089C).

- Inserciones: Se describen anotando el sitio inmediatamente anterior a la inserción (respecto a la hebra ligera de la rCRS), seguido de un punto y un '1' para la primera base insertada, con numeración secuencial para bases subsiguientes (ej., 315.1C). Las inserciones no deben alterar la numeración subsiguiente de la secuencia.
- Deleciones: Se describen indicando el sitio delecionado seguido de un guion (-) o las palabras "del" o "DEL" (ej., 249-, 249del, 249DEL).
- Posicionamiento de Indels (Regla del 3'): Las indels deben colocarse en el extremo 3' de la hebra ligera, a menos que la filogenia sugiera lo contrario.
- Heteroplasmia de Punto (PHP): Las mezclas de bases se representan con los códigos IUPAC extendidos en mayúsculas (ej., A/G = R, C/T = Y).
- Heteroplasmia de Longitud (LHP): Las mezclas entre bases delecionadas/no delecionadas o insertadas/no insertadas se indican con letras minúsculas (ej., 249a para A y una deleción en 249; 152c para una mezcla de C y una deleción en 152). SWGDAM no recomienda el uso de variantes de longitud comunes (p. ej., después de las posiciones 16193, 309 y 573) para comparaciones o búsquedas, ya que pueden ser ignoradas por las bases de datos.

Con el fin de llegar a un consenso entre nomenclaturas, la propuesta fue la salida de las variantes formateadas cumpliendo con las recomendaciones tanto de HSVG, como de Miotomaster y SWGDAM en formato PDF, siguiendo la estructura planteada en la Tabla 3:

Tabla 3. Representación de variantes de ADNmt en diferentes estándares de nomenclatura.

Tipo de Variante	Descripción (rCRS)	Formato HGVS (Ejemplo)	Formato MITOMASTER (Ejemplo)	Formato EMPOP (Ejemplo)
SNP Transición	A>G en pos. 73	NC_012920.1:m.73A>G	A73G	73G
SNP Transversión	C>A en pos. 146	NC_012920.1:m.146C>A	C146A	146A
Deleción	Deleción de T en 249	NC_012920.1:m.249delT	T249d	249DEL
Inserción	Inserción de C después de 315	NC_012920.1:m.315_316insC	G316CC	315.1C
Deleción Larga (Indel)	Deleción de 6 pb (105-110)	NC_012920.1:m.105_110del	GGAGCC105-	105DEL… 110DEL
Inserción Larga (Indel)	Inserción de 10 bases en 291	NC_012920.1:m.291_292insAAAAAAAAAA	C296CAAAAAAAAAA	291.1T… 291.10A

Elaborado por: (El autor, 2025)

3.4.5 Manejo de Regiones Homopoliméricas y Casos Especiales

Con el fin de priorizar la identificación y el manejo de variantes en regiones hotspot y casos especiales del genoma mitocondrial, conocidos por su alta tasa de mutación o por presentar desafíos específicos en su anotación debido a su naturaleza repetitiva o ambigua. Se define una lista para tener en cuenta de estas regiones requieren un tratamiento particular en el análisis y reporte de perfiles de ADNmt que se detallan en la Tabla 4.

Tabla 4. Regiones de interés y casos especiales en el análisis de ADNmt

Caso Especial/ Zona	Extensión (rCRS 1-based)	Descripción y Relevancia Forense
Tractos Poliméricos de C (HV1, HV2)	HVS-I: 16183-16194; HVS-II: 302-310[5,18,30,31]	Regiones con alta variabilidad de longitud (heteroplasmia de longitud). Su notación específica es crucial[7,14,32] (ej., 16189C, 309.1C/309.2C). SWGDAM no recomienda el uso de variantes de longitud comunes en estas ubicaciones para comparaciones o búsquedas directas

AC Regiones repetitivas (HVIII)	515-525	Motivo de repetición AC que debe mantenerse. Propensa a indels[6,21], Mitomaster tiene un arreglo especial en esta secuencia, debido a cambios de la RSRS50[33].
Regiones repetitivas C	455 451 456 463 460 464 573 567 574 960 955 961 5899 5894 5900 8276 8271 8277 8285 8280 8286	Tractos de C alrededor de las posiciones 16193, 309, 463 y 573, y el tracto de T alrededor de la posición 455 deben ser ignoradas para bases de datos forenses[18]. Al igual que en la región codificante, las variantes de longitud alrededor de las posiciones 960, 5899, 8276 y 8285[18].
Deleciónes np	249[34], 290, 291[2], 105-110[18]	Deleción común (ej., 249DEL) que debe mantenerse en el alineamiento filogenético para una notación consistente[18].

Elaborado por: (El autor, 2025)

3.4.6 Anotación del Locus Genómico de las Variantes

Cada variante detectada por MitoID es contextualizada asignándole la región o gen específico del genoma mitocondrial que afecta. Esta anotación se realiza comparando la posición de la variante con las características ya bien definidas en la Tabla 1 y también se añade el atractivo visual de la representación del alineamiento de cada variante en el informe final para una mayor interactividad.

3.4.7 Visualización de resultados:

Para complementar la información tabular y numérica, MitoID incluye un Track Viewer (Visualizador de alineamientos) como herramienta de visualización gráfica de variantes en el contexto del genoma mitocondrial. Este módulo, presenta las variantes marcadas directamente sobre una representación esquemática de las diferentes pistas del genoma, incluyendo las posiciones de cada mutación facilitando la interpretación visual de su posición y relación con las características genómicas anotadas. Esto mejora la comprensión de los perfiles genéticos y el impacto de las mutaciones, ofreciendo una perspectiva intuitiva que es especialmente valiosa para fines educativos y de revisión rápida.

Las variantes detectadas por MitoID se marcan en la visualización con líneas verticales de diferentes colores y estilos, permitiendo diferenciar visualmente entre SNPs (ej., rojo discontinuo), inserciones (ej., azul punteado) y deleciones (ej., naranja-punto-guion). Esta

codificación visual facilita la identificación rápida de la naturaleza de la mutación y su ubicación precisa dentro del genoma mitocondrial y sus características funcionales.

3.5 Flujo de trabajo

El siguiente flujo de trabajo (Figura *2*) Ilustra visualmente las etapas principales del procesamiento, donde cada fase representa una serie de operaciones bioinformáticas interconectadas.

Modelo de detección de variantes MitoID

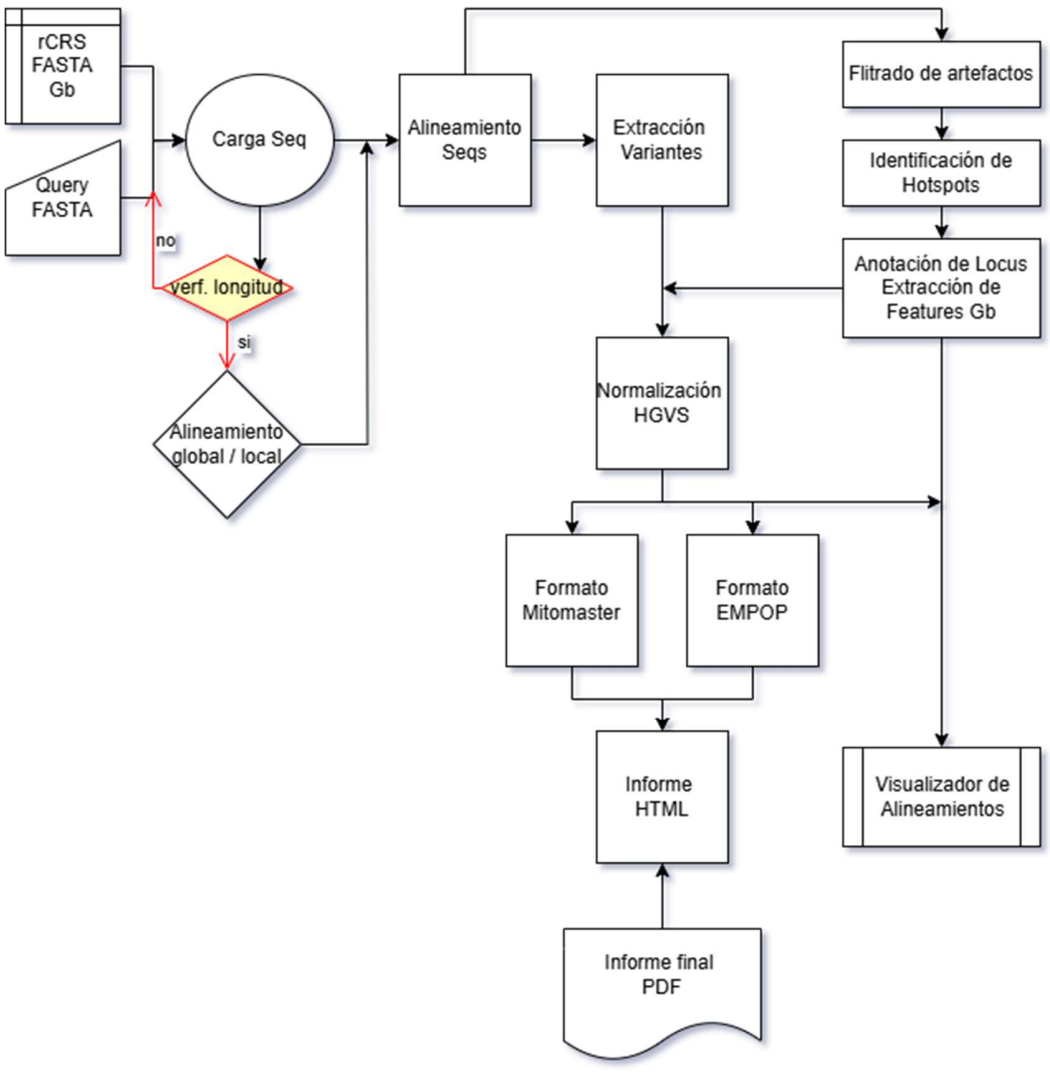


Figura 2. Diagrama de flujo de lógicas de funcionamiento MitoID

Elaborado por: (El autor.)

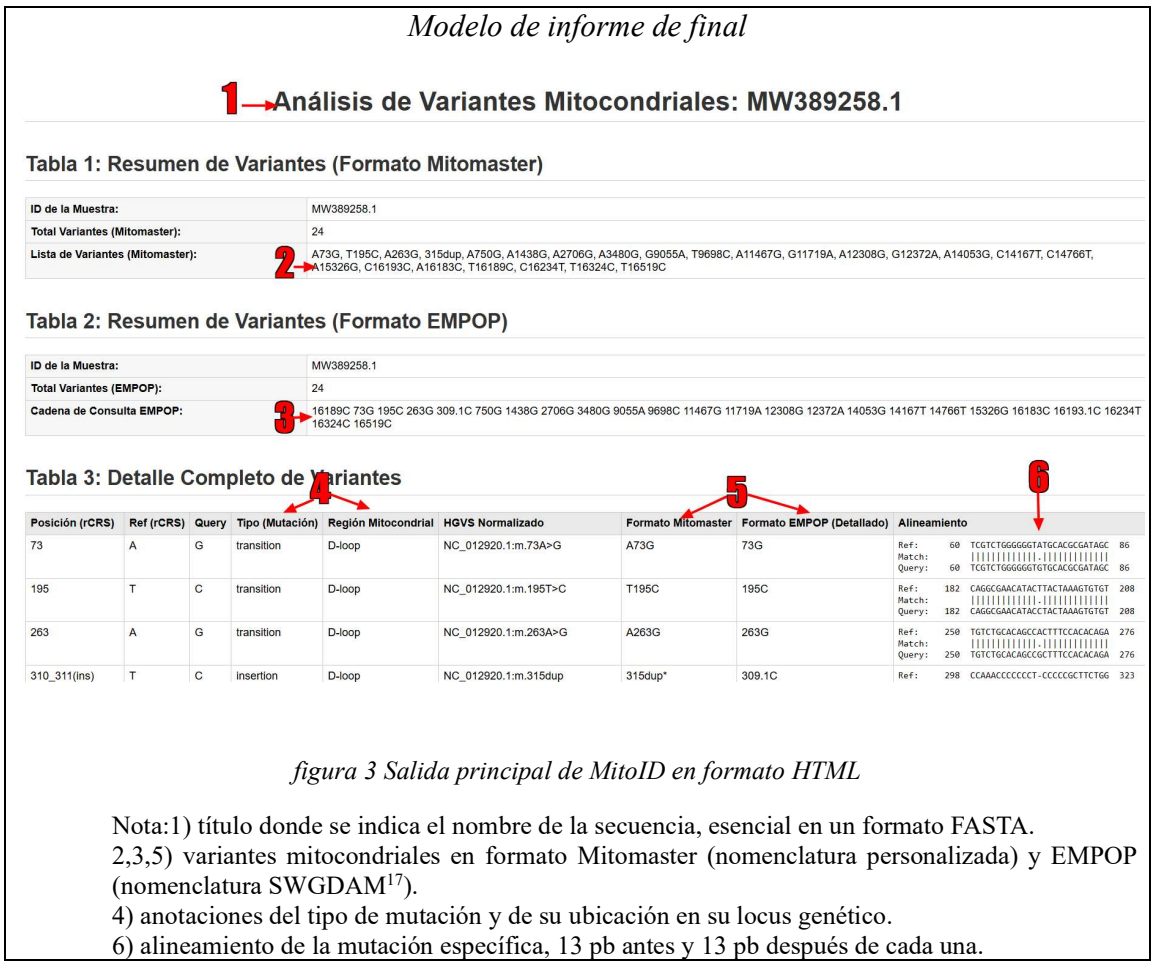
4 Resultados y discusión

El primer resultado relevante del proyecto es la correcta implementación de un sistema de carga de secuencias y alineamiento de Query vs rCRS que no genere errores, como se puede ver reflejado en Anexo 1, prints en la consola de Jupyter Notebook que advierten del tamaño de la secuencia o de su incorrecta carga o alineamiento, e incluso el alineamiento completo se puede observar un ejemplo en Anexo 2 como función descomentada del alineamiento de las secuencias si así el usuario lo desease.

Esto demuestra que la implementación de los algoritmos de alineamiento global (Needleman-Wunsch) y local (Smith-Waterman) en MitoID son efectivos en la identificación de secuencias homólogas de ADNmt respecto a la rCRS (NC_012920.1). La elección adaptable del modo de alineamiento, basada en la longitud de la secuencia de consulta (global si es ≥ 80% de la rCRS, local en caso contrario), permitió un manejo eficiente de muestras tanto completas como fragmentadas, un aspecto crítico en genética forense donde las muestras pueden estar degradadas. La matriz de sustitución personalizada, con puntuaciones optimizadas para la alta conservación del genoma mitocondrial humano, contribuyó a la precisión del alineamiento

Los resultados definitivos del proyecto se pueden ver reflejados como salidas de MitoID dentro de la misma carpeta donde se lo esté ejecutando, los dos archivos a manera de sistema de reporte generados lucen de la siguiente manera (imagen) donde el PDF imprime solo las características esenciales por decisión de optimización de espacio.

Modelo de informe de final



1 → Análisis de Variantes Mitocondriales: MW389258.1

Tabla 1: Resumen de Variantes (Formato Mitomaster)

ID de la Muestra:	MW389258.1
Total Variantes (Mitomaster):	24
Lista de Variantes (Mitomaster):	A73G, T195C, A263G, 315dup, A750G, A1438G, A2706G, A3480G, G9055A, T9698C, A11467G, G11719A, A12308G, G12372A, A14053G, C14167T, C14766T, A15326G, C16193C, A16183C, T16189C, C16234T, T16324C, T16519C

2 →

Tabla 2: Resumen de Variantes (Formato EMPOP)

ID de la Muestra:	MW389258.1
Total Variantes (EMPOP):	24
Cadena de Consulta EMPOP:	16189C 73G 195C 263G 309.1C 750G 1438G 2706G 3480G 9055A 9698C 11467G 11719A 12308G 12372A 14053G 14167T 14766T 15326G 16183C 16193.1C 16234T 16324C 16519C

3 →

Tabla 3: Detalle Completo de Variantes

4 5 6

Posición (rCRS)	Ref (rCRS)	Query	Tipo (Mutación)	Región Mitocondrial	HGVS Normalizado	Formato Mitomaster	Formato EMPOP (Detallado)	Alineamiento
73	A	G	transition	D-loop	NC_012920.1:m.73A>G	A73G	73G	Ref: 60 TCGTCTGGGGGGTATGCACGCGATAGC 86 Match: |||||||||||.|||||||||||| Query: 60 TCGTCTGGGGGGTGTGCACGCGATAGC 86
195	T	C	transition	D-loop	NC_012920.1:m.195T>C	T195C	195C	Ref: 182 CAGGCGAACATACTTACTAAAGTGTGT 208 Match: |||||||||||.|||||||||||| Query: 182 CAGGCGAACATACCTACTAAAGTGTGT 208
263	A	G	transition	D-loop	NC_012920.1:m.263A>G	A263G	263G	Ref: 250 TGTCTGCACAGCCACTTTCCACACAGA 276 Match: |||||||||||.|||||||||||| Query: 250 TGTCTGCACAGCCGCTTTCCACACAGA 276
310_311(ins)	T	C	insertion	D-loop	NC_012920.1:m.315dup	315dup*	309.1C	Ref: 298 CCAAACCCCCCCT-CCCCCGCTTCTGG 323

figura 3 Salida principal de MitoID en formato HTML

Nota:1) título donde se indica el nombre de la secuencia, esencial en un formato FASTA.
2,3,5) variantes mitocondriales en formato Mitomaster (nomenclatura personalizada) y EMPOP (nomenclatura SWGDAM[17]).
4) anotaciones del tipo de mutación y de su ubicación en su locus genético.
6) alineamiento de la mutación específica, 13 pb antes y 13 pb después de cada una.

Cada informe incluye un resumen de variantes en formato Mitomaster y EMPOP, y una tabla detallada de variantes. La tabla detallada proporciona información exhaustiva, incluyendo la posición de la variante en la rCRS, las bases de referencia y consulta, el tipo de mutación, la región mitocondrial afectada, el formato Mitomaster/EMPOP, y un snippet de alineamiento contextual. Este snippet muestra una ventana de bases alrededor de cada variante con símbolos que indican coincidencias, no-coincidencias o gaps, facilitando una revisión visual precisa de cada evento.

La validación de MitoID se llevó a cabo mediante el análisis de secuencias de ADNmt obtenidas de repositorios públicos, complementadas con secuencias modificadas para simular casos especiales y las salidas de MitoID se compararon en el modo “alineamiento” de la base de datos de EMPOP como método de verificación final, como se puede observar

en los Anexos. Se utilizaron 4 secuencias de GenBank (MZ387958.1, MW389258.1, MW389269.1, MW389271.1) para evaluar la precisión del alineamiento y la estandarización de variantes comunes, mientras que 6 secuencias modificadas fueron empleadas para evaluar el manejo de regiones hotspots y variaciones complejas.

A continuación, se presentan ejemplos específicos de variantes detectadas por MitoID, su estandarización y su alineamiento:

- **SNPs (Transiciones y Transversiones):**
    - **Transición A>G en la posición 73** (73G): Esta variante, localizada en la región D-loop (HVS-II), se detectó como una transición de adenina a guanina y se formateó correctamente como 73G en EMPOP y A73G en Mitomaster
    - **Transversión T>A en la posición 789** (789A): Ubicada en la región 12S(RNR1), esta transversión de timina a adenina es una variante común en fue anotada como T789A en formato Mitomaster, y 789A en EMPOP conforme a las recomendaciones de SWGDAM.
- **Inserciones:**
    - **Inserción de C en la posición 309 (309.1C 309.2C):** Este es un caso de inserción de dos citosinas consecutivas en un tracto homopolimérico de la HVS-II. MitoID la detectó y la reportó como **309.1C 309.2C** en EMPOP, siguiendo la convención de numeración secuencial para bases insertadas después del sitio de anclaje. En Mitomaster, se anotó como **C309CCC**.
    - **Inserción en la posición 310 (315dup):** Aunque EMPOP la reportó como **315.1C**, MitoID la detectó como una inserción en 310_311, y el formato Mitomaster la representó como **315dup*** (con un asterisco indicando que es un hotspot). Este tipo de inserción es un ejemplo de cómo MitoID maneja los casos especiales, ya que SWGDAM indica que los C-stretch en HVII deben interpretarse como 310C si el T en esa posición no está presente.

Deleciones:

- **Deleción GC en la posición 513 (523DEL 524DEL):** Esta deleción de dos bases (GC) fue detectada en la posición 513 y, según las directrices de EMPOP/SWGDAM para la región del motivo repetitivo AC (np 515-525), se formateó como dos deleciones consecutivas: **523DEL 524DEL** en EMPOP y **GC512d** en Mitomaster.

Para las pruebas exhaustivas con el conjunto de 6 secuencias de prueba modificadas intencionalmente para simular estas variantes problemáticas. Los resultados obtenidos con MitoID demostraron su capacidad para:

- **Manejo de regiones C-repetitivas en HVS-I y HVS-II:**
    - **Mod_2 (Inserción en HVS-II C-stretch):** Una secuencia modificada para simular una inserción en la región homopolimérica de C (C-stretch) de HVS-II (posiciones 302-310) fue correctamente identificada por MitoID. Específicamente, para una inserción en la posición 302-303, la salida en formato EMPOP fue 309.1C 309.2C. Esto concuerda con la convención de EMPOP y las directrices de SWGDAM para inserciones, las cuales indican que se debe anotar el sitio inmediatamente anterior a la inserción, seguido de un punto y un número secuencial para cada base insertada (ej., 315.1C). Este formato es crucial para la estandarización de inserciones en regiones de longitud variable, como el C-stretch de HVII**.**

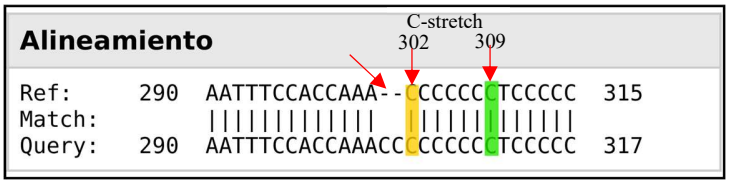


figura 4 Inserción HVS-II salida de MitoID

    - **Mod_3 (Inserción en 16193):** Una secuencia modificada para simular una inserción en el C-stretch de HVS-I (posiciones 16183-16194) fue correctamente identificada por MitoID, para una inserción en la posición 16183-16184 con una salida 16193.1C 16193.2C en EMPOP. Esto concuerda con la convención de EMPOP y las directrices de SWGDAM para inserciones en posiciones de la región HVS-I.

figura 5. Inserción HVS-I salida de MitoID

- **Deleción 249:**
    - **Mod_4 (Deleción 249 y otras):** Una secuencia modificada para simular una deleción en la posición 249. MitoID la identificó y reportó como 249DEL en EMPOP. Esta es una deleción común y su correcto reporte es fundamental para el análisis forense. SWGDAM resalta la importancia de mantener la deleción 249DEL filogenéticamente establecida para evitar falsas exclusiones en las bases de datos.

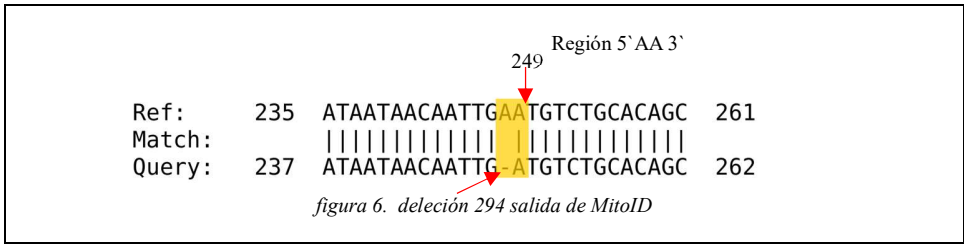


figura 6. deleción 294 salida de MitoID

- **Región 513-524 (Motivo Repetitivo AC):**
    - Como se mostró en la sección 4.2.2 con la secuencia MZ387958.1, MitoID maneja las deleciones en esta región crítica, reportándolas como 523DEL 524DEL en EMPOP. Esto es consistente con las directrices de SWGDAM que establecen el mantenimiento del motivo de repetición AC en HVIII (np 515-525).

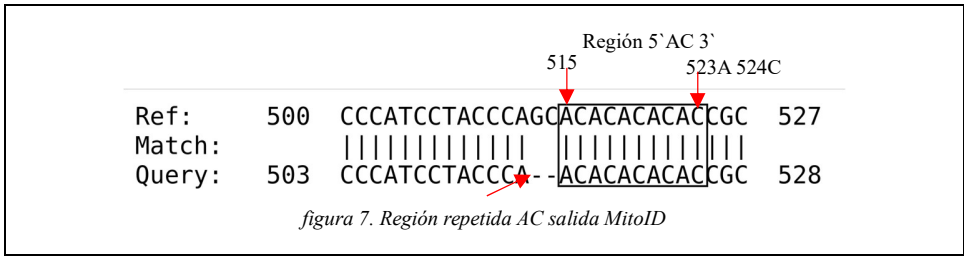


figura 7. Región repetida AC salida MitoID

- **Deleciones y Anotaciones Complejas:**
    - **Mod_5 (Inserción Compleja 291.x):** Una secuencia con una inserción compleja alrededor de la posición 291 (que en EMPOP se descompone en múltiples inserciones 291.1A, 291.2C, etc.) fue exitosamente desglosada y reportada por MitoID como una serie de variantes 291.x en EMPOP. Esto demuestra la capacidad de la herramienta para manejar inserciones complejas que a menudo causan inconsistencias en la nomenclatura manual.

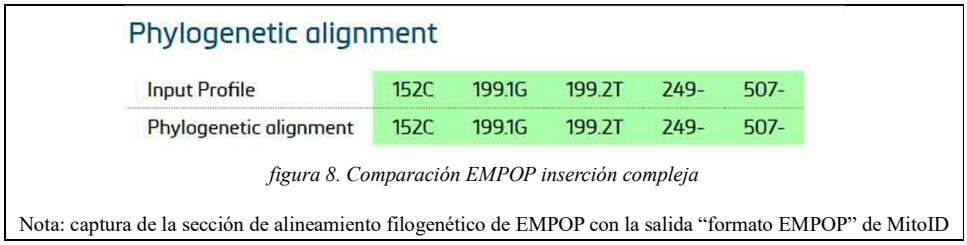


figura 8. Comparación EMPOP inserción compleja

Nota: captura de la sección de alineamiento filogenético de EMPOP con la salida "formato EMPOP" de MitoID

- **Mod_6 (Deleción en 8270-8292):** Este caso, que implica una deleción grande y compleja, es un desafío debido a las diferentes interpretaciones de los alineadores. MitoID generó una serie de 8207DEL-8215DEL en formato EMPOP para esta deleción. Si bien SAM 2[18] (matriz de alineamiento utilizada por EMPOP) puede tener un comportamiento diferente en la anotación de deleciones consecutivas muy grandes, la representación de MitoID como una secuencia de deleciones en cada posición se alinea con la lógica de EMPOP de desglosar el evento en deleciones individuales para la consulta. Aunque la salida de MitoID es detallada, la propia documentación de SWGDAM reconoce que las "deleciones comunes" (ej., después de 16193, 309, 573) pueden ser ignoradas por las bases de datos para comparaciones directas.

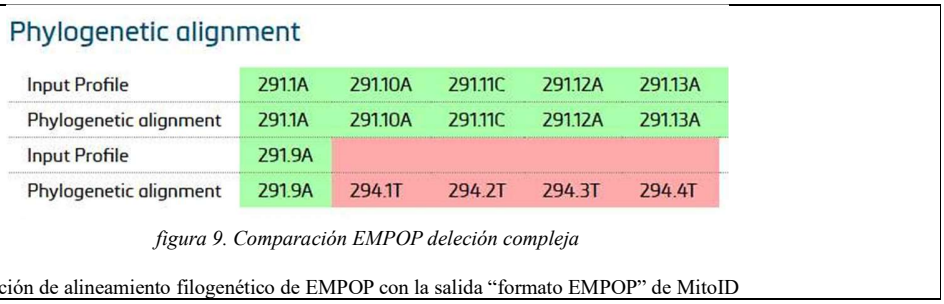


figura 9. Comparación EMPOP deleción compleja

Nota: captura de la sección de alineamiento filogenético de EMPOP con la salida "formato EMPOP" de MitoID

Los resultados para el anclaje de un visualizador de alineamientos son representados en figura 10 y figura 11

Las variantes se visualizan como marcadores gráficos (puntos y triángulos) directamente superpuestos en la pista de la Secuencia Query. Cada marcador está conectado mediante una línea vertical a la línea base del rCRS o a las pistas genómicas inferiores, permitiendo una rápida correlación posicional. La codificación visual es intuitiva:

- **Sustituciones:** Representadas por marcadores circulares (rojo).
- **Inserciones:** Indicadas por triángulos hacia arriba (azul).
- **Deleciones:** Señaladas por triángulos hacia abajo (naranja)

Modelo de Visualizador de alineamientos interactivo

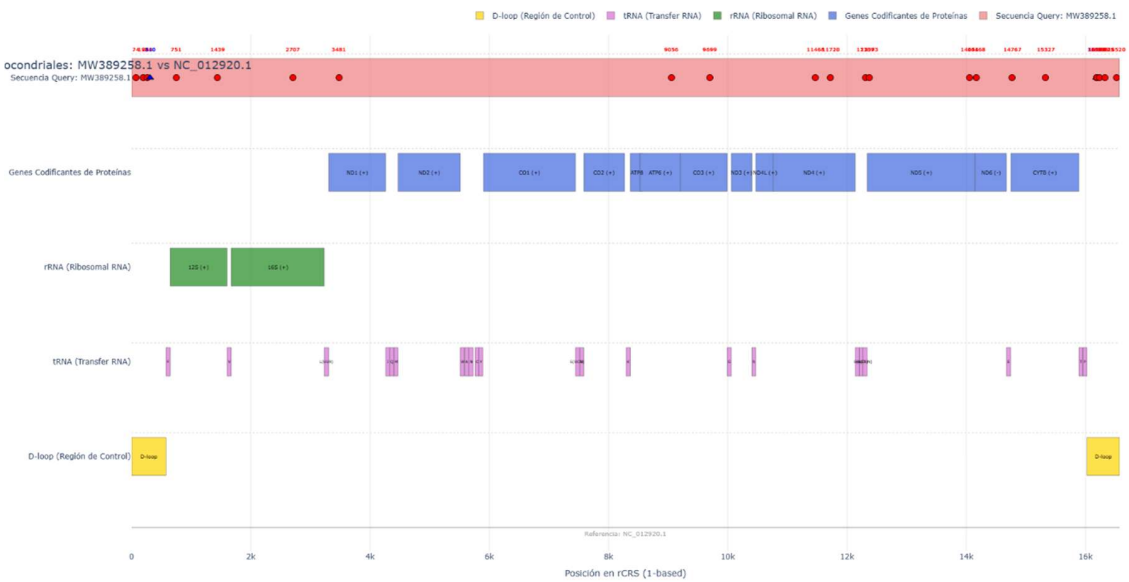


figura 10 Salida del visualizador de variantes de MitoID en formato HTML

Este módulo transforma datos genómicos complejos en una representación visualmente accesible, lo que facilita no solo la validación de los resultados por parte del usuario, sino también su aplicación en la enseñanza y la comunicación de hallazgos a audiencias no expertas. La interactividad de Plotly, que permite hacer zoom, desplazarse y obtener información detallada al pasar el cursor, mejorando la capacidad de interpretar visualmente los resultados del análisis de ADNmt.

Al pasar el cursor sobre un marcador de variante, se despliega un cuadro con datos esenciales, presentados de forma clara, este cuadro incluye:

- Variante: Nomenclatura concisa (ej., A74G, -310C).
- Tipo: Clasificación de la mutación (ej., Transition, Insertion).
- Posición: Coordenada numérica en la rCRS.
- Locus: La región o gen mitocondrial afectada (ej., D-loop, 12S, ND1).

- Región Hipervariable: Novedad: Si la variante cae dentro de una de las Regiones Hipervariables (HVS-I, HVS-II, HVS-III) definidas en la Tabla 1 de Introducción,
- Las nomenclaturas ya establecidas, HGVS, Mitomaster y EMPOP

Modelo de visualización de variantes Plotly específicas

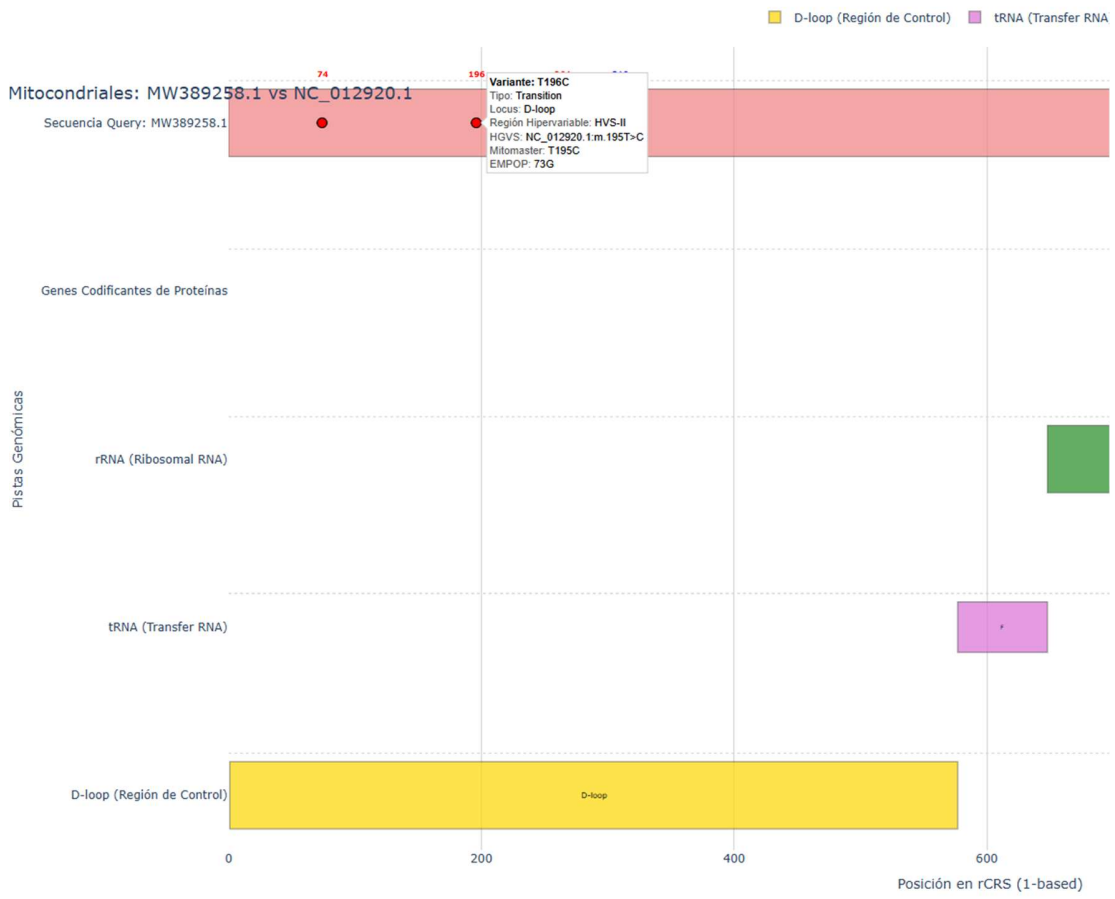


figura 11 zoom de la salida del visualizador de variantes sobre una variante

Nota: al usar el visualizador en HTML es posible realizar zoom y pasar el cursor sobre las mutaciones o incluso genes para obtener su información complementaria

5 Conclusiones:

MitoID opera como una herramienta de ejecución local, diseñada para ser utilizada mediante la línea de comandos o en un entorno de cuaderno interactivo (ej., Jupyter Notebook). Esta modalidad de operación simplifica el acceso y la ejecución para usuarios en contextos académicos y forenses que no siempre disponen de conexión a internet o de complejas infraestructuras de servidor. El proceso de análisis en MitoID se organiza en etapas secuenciales, desde la carga de la secuencia de consulta hasta la generación de informes finales, siguiendo el flujo de trabajo detallado en el Diagrama de flujo de lógicas de funcionamiento MitoID de la sección de Materiales y Métodos.

El sistema de reporte de MitoID satisface las necesidades de comunicación forense y académica, proporcionando informes completos, bien estructurados y visualmente informativos que consolidan todos los resultados del análisis, además de la implementación del Visualizador de variantes en MitoID que mejora la capacidad de interpretar visualmente los resultados del análisis de ADNmt

Las pruebas de funcionamiento demuestran una capacidad suficiente para la detección de Polimorfismos de un Solo Nucleótido (SNPs) e inserciones/deleciones (indels) mediante el alineamiento de secuencias FASTA contra la rCRS, así como para el filtrado de artefactos de secuenciación, donde la capacidad de para procesar y reportar de manera precisa estas variantes complejas y hotspots es un punto de valor crítico. Sin embargo, aunque MitoID maneja eficazmente el alineamiento y las variantes en la mayoría de los hotspots y regiones repetitivas, se identificaron escenarios extremadamente complejos donde la lógica del alineador interno puede diferir de interpretaciones filogenéticas muy específicas, como en el caso de la deleción en la región 8270-8292. Para emular la lógica de detección de herramientas como SAM2 (utilizada por EMPOP) en estos casos raros, sería necesaria una reimplementación significativa de la fase de alineamiento o la integración de VCF/BAM preprocesados, lo cual excedía el alcance de este proyecto.

Aunque el uso actual vía consola/Jupyter es funcional, el desarrollo de una interfaz gráfica de usuario (GUI) más intuitiva mejoraría significativamente la experiencia del usuario final, especialmente para profesores y estudiantes menos familiarizados con entornos de programación.

Finalmente, investigar y desarrollar algoritmos de alineamiento o posprocesamiento más sofisticados sigue siendo una prioridad, para que se puedan resolver las ambigüedades en el modo en que Mitomaster reporta variantes de forma menos estandarizada con referencia a SWGDAM sigue siendo un reto aplicable en un futuro, posiblemente mediante enfoques basados aprendizaje automático o en contacto con los desarrolladores principales de Mitomaster.

6 Bibliografía

(1) Amorim, A.; Fernandes, T.; Taveira, N. Mitochondrial DNA in Human Identification: A Review. *PeerJ* **2019**, *7*, e7314. https://doi.org/10.7717/peerj.7314.

(2) Cavalcanti, P.; Nogueira, T. L. S.; Carvalho, E. F. D.; Silva, D. A. D. Forensic Use of Human Mitochondrial DNA: A Review. *An. Acad. Bras. Ciênc.* **2024**, *96* (4). https://doi.org/10.1590/0001-3765202420231179.

(3) Cappa, R.; De Campos, C.; Maxwell, A. P.; McKnight, A. J. "Mitochondrial Toolbox" – A Review of Online Resources to Explore Mitochondrial Genomics. *Front. Genet.* **2020**, *11*, 439. https://doi.org/10.3389/fgene.2020.00439.

(4) Cortes-Figueiredo, F.; Carvalho, F. S.; Fonseca, A. C.; Paul, F.; Ferro, J. M.; Schönherr, S.; Weissensteiner, H.; Morais, V. A. From Forensics to Clinical Research: Expanding the Variant Calling Pipeline for the Precision ID mtDNA Whole Genome Panel. *Int. J. Mol. Sci.* **2021**, *22* (21), 12031. https://doi.org/10.3390/ijms222112031.

(5) Parson, W.; Gusmão, L.; Hares, D. R.; Irwin, J. A.; Mayr, W. R.; Morling, N.; Pokorak, E.; Prinz, M.; Salas, A.; Schneider, P. M.; Parsons, T. J. DNA Commission of the International Society for Forensic Genetics: Revised and Extended Guidelines for Mitochondrial DNA Typing. *Forensic Sci. Int. Genet.* **2014**, *13*, 134–142. https://doi.org/10.1016/j.fsigen.2014.07.010.

(6) Laricchia, K. M.; Lake, N. J.; Watts, N. A.; Shand, M.; Haessly, A.; Gauthier, L.; Benjamin, D.; Banks, E.; Soto, J.; Garimella, K.; Emery, J.; Genome Aggregation Database Consortium; Rehm, H. L.; MacArthur, D. G.; Tiao, G.; Lek, M.; Mootha, V. K.; Calvo, S. E. Mitochondrial DNA Variation across 56,434 Individuals in gnomAD. *Genome Res.* **2022**, *32* (3), 569–582. https://doi.org/10.1101/gr.276013.121.

(7) Just, R. S.; Leney, M. D.; Barritt, S. M.; Los, C. W.; Smith, B. C.; Holland, T. D.; Parsons, T. J. The Use of Mitochondrial DNA Single Nucleotide Polymorphisms to Assist in the Resolution of Three Challenging Forensic Cases. *J. Forensic Sci.* **2009**, *54* (4), 887–891. https://doi.org/10.1111/j.1556-4029.2009.01069.x.

(8) Budowle, B.; Wilson, M. R.; DiZinno, J. A.; Stauffer, C.; Fasano, M. A.; Holland, M. M.; Monson, K. L. Mitochondrial DNA Regions HVI and HVII Population Data. *Forensic Sci. Int.* **1999**, *103* (1), 23–35. https://doi.org/10.1016/S0379-0738(99)00042-0.

(9) Gill, P.; Ivanov, P. L.; Kimpton, C.; Piercy, R.; Benson, N.; Tully, G.; Evett, I.; Hagelberg, E.; Sullivan, K. Identification of the Remains of the Romanov Family by DNA Analysis. *Nat. Genet.* **1994**, *6* (2), 130–135. https://doi.org/10.1038/ng0294-130.

(10) Corach, D.; Sala, A.; Penacino, G.; Iannucci, N.; Bernardi, P.; Doretti, M.; Fondebrider, L.; Ginarte, A.; Inchaurregui, A.; Somigliana, C.; Turner, S.; Hagelberg, E. Additional

Approaches to DNA Typing of Skeletal Remains: The Search for “Missing” Persons Killed during the Last Dictatorship in Argentina. *ELECTROPHORESIS* **1997**, *18* (9), 1608–1612. https://doi.org/10.1002/elps.1150180921.

(11) Baeta, M.; Núñez, C.; Cardoso, S.; Palencia-Madrid, L.; Herrasti, L.; Etxeberria, F.; De Pancorbo, M. M. Digging up the Recent Spanish Memory: Genetic Identification of Human Remains from Mass Graves of the Spanish Civil War and Posterior Dictatorship. *Forensic Sci. Int. Genet.* **2015**, *19*, 272–279. https://doi.org/10.1016/j.fsigen.2015.09.001.

(12) Syndercombe Court, D. Mitochondrial DNA in Forensic Use. *Emerg. Top. Life Sci.* **2021**, *5* (3), 415–426. https://doi.org/10.1042/ETLS20210204.

(13) Barresi, M.; Dal Santo, G.; Izzo, R.; Zauli, A.; Lamantea, E.; Caporali, L.; Ghezzi, D.; Legati, A. Bioinformatics Tools for NGS-Based Identification of Single Nucleotide Variants and Large-Scale Rearrangements in Mitochondrial DNA. *BioTech* **2025**, *14* (1), 9. https://doi.org/10.3390/biotech14010009.

(14) Coble, M. D.; Just, R. S.; O’Callaghan, J. E.; Letmanyi, I. H.; Peterson, C. T.; Irwin, J. A.; Parsons, T. J. Single Nucleotide Polymorphisms over the Entire mtDNA Genome That Increase the Power of Forensic Testing in Caucasians. *Int. J. Legal Med.* **2004**, *118* (3), 137–146. https://doi.org/10.1007/s00414-004-0427-6.

(15) Gill, P.; Ivanov, P. L.; Kimpton, C.; Piercy, R.; Benson, N.; Tully, G.; Evett, I.; Hagelberg, E.; Sullivan, K. Identification of the Remains of the Romanov Family by DNA Analysis. *Nat. Genet.* **1994**, *6* (2), 130–135. https://doi.org/10.1038/ng0294-130.

(16) Shen, L.; Attimonelli, M.; Bai, R.; Lott, M. T.; Wallace, D. C.; Falk, M. J.; Gai, X. MSeqDR mvTool: A Mitochondrial DNA Web and API Resource for Comprehensive Variant Annotation, Universal Nomenclature Collation, and Reference Genome Conversion. *Hum. Mutat.* **2018**, *39* (6), 806–810. https://doi.org/10.1002/humu.23422.

(17) SWGDAM. Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories, 2024. https://www.swgdam.org/files/ugd/4344b0_f400dd61044a4a328b95362f46fcbf4c.pdf.

(18) Institute of Legal Medicine; University of Innsburk. EMPOP, 2019. https://empop.online/.

(19) Brandon, M. C.; Ruiz-Pesini, E.; Mishmar, D.; Procaccio, V.; Lott, M. T.; Nguyen, K. C.; Spolim, S.; Patil, U.; Baldi, P.; Wallace, D. C. MITOMASTER: A Bioinformatics Tool for the Analysis of Mitochondrial DNA Sequences. *Hum. Mutat.* **2009**, *30* (1), 1–6. https://doi.org/10.1002/humu.20801.

(20) Lott, M. T.; Leipzig, J. N.; Derbeneva, O.; Xie, H. M.; Chalkia, D.; Sarmady, M.; Procaccio, V.; Wallace, D. C. mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinforma.* **2013**, *44* (1). https://doi.org/10.1002/0471250953.bi0123s44.

(21) Lee, S. E.; Kim, G. E.; Kim, H.; Chung, D. H.; Lee, S. D.; Kim, M.-Y. Comparison of Two Variant Analysis Programs for Next-Generation Sequencing Data of Whole Mitochondrial Genome. *J. Korean Med. Sci.* **2023**, *38* (36), e297. https://doi.org/10.3346/jkms.2023.38.e297.

(22) Liu, Y.-Y.; Harbison, S. A Review of Bioinformatic Methods for Forensic DNA Analyses. *Forensic Sci. Int. Genet.* **2018**, *33*, 117–128. https://doi.org/10.1016/j.fsigen.2017.12.005.

(23) Attimonelli, M.; Preste, R.; Vitale, O.; Lott, M. T.; Procaccio, V.; Shiping, Z.; Wallace, D. C. Bioinformatics Resources, Databases, and Tools for Human mtDNA. In *The*

Human Mitochondrial Genome; Elsevier, 2020; pp 277–304. https://doi.org/10.1016/b978-0-12-819656-4.00012-7.
(24) Huszar, T. I.; Gettings, K. B.; Vallone, P. M. An Introductory Overview of Open-Source and Commercial Software Options for the Analysis of Forensic Sequencing Data. *Genes* **2021**, *12* (11), 1739. https://doi.org/10.3390/genes12111739.
(25) Elder, C. S.; Welsh, C. E. MitoMut: An Efficient Approach to Detecting Mitochondrial DNA Deletions from Paired-End Next-Generation Sequencing Data. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*; ACM: Niagara Falls NY USA, 2019; pp 177–182. https://doi.org/10.1145/3307339.3342158.
(26) Marquis, J.; Lefebvre, G.; Kourmpetis, Y. A. I.; Kassam, M.; Ronga, F.; De Marchi, U.; Wiederkehr, A.; Descombes, P. MitoRS, a Method for High Throughput, Sensitive, and Accurate Detection of Mitochondrial DNA Heteroplasmy. *BMC Genomics* **2017**, *18* (1), 326. https://doi.org/10.1186/s12864-017-3695-5.
(27) Diroma, M. A.; Lubisco, P.; Attimonelli, M. A Comprehensive Collection of Annotations to Interpret Sequence Variation in Human Mitochondrial Transfer RNAs. *BMC Bioinformatics* **2016**, *17* (S12), 338. https://doi.org/10.1186/s12859-016-1193-4.
(28) Lake, N. J.; Zhou, L.; Xu, J.; Lek, M. MitoVisualize: A Resource for Analysis of Variants in Human Mitochondrial RNAs and DNA. *Bioinformatics* **2022**, *38* (10), 2967–2969. https://doi.org/10.1093/bioinformatics/btac216.
(29) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.
(30) Hart, R. K.; Fokkema, I. F. A. C.; DiStefano, M.; Hastings, R.; Laros, J. F. J.; Taylor, R.; Wagner, A. H.; Den Dunnen, J. T. HGVS Nomenclature 2024: Improvements to Community Engagement, Usability, and Computability. *Genome Med.* **2024**, *16* (1), 149. https://doi.org/10.1186/s13073-024-01421-5.
(31) Li, M.; Schönberg, A.; Schaefer, M.; Schroeder, R.; Nasidze, I.; Stoneking, M. Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes. *Am. J. Hum. Genet.* **2010**, *87* (2), 237–249. https://doi.org/10.1016/j.ajhg.2010.07.014.
(32) Bandelt, H.-J.; Van Oven, M.; Salas, A. Haplogrouping Mitochondrial DNA Sequences in Legal Medicine/Forensic Genetics. *Int. J. Legal Med.* **2012**, *126* (6), 901–916. https://doi.org/10.1007/s00414-012-0762-y.
(33) Lott, M. T.; Leipzig, J. N.; Derbeneva, O.; Xie, H. M.; Chalkia, D.; Sarmady, M.; Procaccio, V.; Wallace, D. C. mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinforma.* **2013**, *44* (123), 1.23.1-26. https://doi.org/10.1002/0471250953.bi0123s44.
(34) Lee, H. Y.; Chung, U.; Park, M. J.; Yoo, J.; Han, G.; Shin, K. Differential Distribution of Human Mitochondrial DNA in Somatic Tissues and Hairs. *Ann. Hum. Genet.* **2006**, *70* (1), 59–65. https://doi.org/10.1111/j.1529-8817.2005.00217.x.

7 Anexos

Anexo 1 Mensajes de error durante la carga y alineamiento de secuencias

```
 mensajes de error/advertencia ---

--- Errores de carga de FASTA ---
Error: No se encontró FASTA en: 'ruta/no_existente.fasta'
Error: FASTA 'archivo_mal_formato.fasta' mal formateado o con >1 secuencia.
Error al leer 'archivo_protegido.fasta': Permission denied

--- Advertencia de ID de FASTA (simulada) ---
Advertencia: El ID ('otra_secuencia_id') no coincide con el esperado ('NC_012920.1').

--- Errores de carga de GenBank (simulados) ---
Error: No se encontró el archivo GenBank rCRS en: 'ruta/genbank_no_existente.gb'
Error cargando el archivo GenBank rCRS 'genbank_corrupto.gb': Bad GenBank format

--- Errores/Advertencias de alineamiento---
No se encontraron alineamientos.
Ocurrió un error durante el alineamiento: Alignment process interrupted

--- Errores/Advertencias del Track Viewer  ---
Error TV: No se encontró el archivo FASTA query en: 'query_tv_no_existe.fasta'
Error TV: El archivo FASTA query 'query_tv_mal_formato.fasta' no contiene una única secuencia o está mal formateado.
Error TV: Ocurrió un error inesperado al leer query FASTA 'query_tv_error.fasta': File locked.
Error TV: No se encontró el archivo GenBank rCRS en: 'rcrs_tv_no_existe.gb'
Error TV: Cargando el archivo GenBank rCRS 'rcrs_tv_error.gb': Corrupted file.
Abortando Track Viewer: No se pudo cargar rCRS GenBank.
Advertencia TV: No se pudo cargar la secuencia query. Se mostrará solo rCRS.
No se encontraron alineamientos para TV.
```

Anexo 2 Selección del Tipo de alineamiento y visualización del alineamiento parcial

```
Realizando alineamiento global...

Alineamiento completado. Puntuación: 49486.0
Número de alineamientos óptimos encontrados: 8

Alineamiento Global Completo:
target            0 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
                  0 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query             0 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT

target           60 CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
                  60 ||||||||||||.||||||||||||||||||||||||||||||||||||||||||||||
query            60 CGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC

target          120 GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
                 120 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query           120 GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT

target          180 ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
                 180 ||||||||||||||||||||.||||||||||||||||||||||||||||||||||||||
query           180 ACAGGCGAACATACTTACTGAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA

target          240 ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
                 240 ||||||||||||||||||||||.||||||||||||||||||||||||||||||||||||
query           240 ACAATTGAATGTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA

target          300 AA-CCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAA
                 300 ||-||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query           300 AACCCCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAA

target          359 AACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCA
                 360 ||||||||||||||||.||||||||||||||||||||||||||||||||||||||||||
query           360 AACAAAGAACCCTAATACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCA
```