Computational Tool for AMOVA and Population Genetic Differentiation Applied to Central America

Yolly Molina[1]; Patricia Soriano[2]; Saulo Romero[3]
[1,3]Universidad Nacional Autónoma de Honduras, Honduras, *yolly.molina@unah.edu.hn, saulo.romero@unah.edu.hn*
[2]Ministerio Público de Honduras, Honduras, *paty.soriano@yahoo.com*

Abstract– ***Regional population genetics in Central America is expanding, but many academic groups still depend on fragmented, non-reproducible workflows for analysis of molecular variance, pairwise differentiation and forensic STR statistics. AMOVA CA is presented as an early-stage academic prototype rather than as a substitute for established packages. This paper presents AMOVA CA, a computational prototype designed to import population datasets, validate allelic structure, estimate diversity indices, calculate pairwise FST, execute hierarchical AMOVA and export traceable tables and figures under user-defined parameters. The prototype was validated at two levels, documentary and computational, using summary statistics from published Honduran STR population datasets for Lenca, Tawahka, Pech, Garifuna and Black Caribbean groups. The proposed model integrates quality control, parameter logging, reproducible reporting and data governance principles. The main scientific contribution is an operational, auditable and regionally adaptable computational workflow, with documented permutation and missing-data procedures, that can support academic training, forensic genetics and comparative population studies in Central America.***

Keywords-- ***AMOVA, population genetics, FST, forensic genetics, Central America.***

.

Herramienta computacional para AMOVA y diferenciación genética poblacional aplicada a Centroamérica

Yolly Molina[1]; Patricia Soriano[2]; Saulo Romero[3]
[1,3]Universidad Nacional Autónoma de Honduras, Honduras, *yolly.molina@unah.edu.hn, saulo.romero@unah.edu.hn*
[2]Ministerio Público de Honduras, Honduras, *paty.soriano@yahoo.com*

Resumen– ***La investigación genética poblacional en Centroamérica ha crecido con nuevos conjuntos de datos de marcadores STR, pero la disponibilidad de software académico reproducible para análisis de varianza molecular y diferenciación poblacional sigue siendo limitada. Este artículo presenta AMOVA CA, un prototipo computacional en etapa inicial de desarrollo, orientado a importar datos genotípicos o frecuencias alélicas, validar su estructura, calcular diversidad intrapoblacional, estimar FST pareado, ejecutar AMOVA jerárquico y exportar reportes auditables con parámetros definidos por el usuario. La validación se realizó en dos niveles, documental y computacional, a partir de estadísticas resumidas de datos poblacionales publicados de grupos hondureños Lenca, Tawahka, Pech, Garífuna y Negro Inglés, todos relevantes para investigación regional y genética forense. El prototipo integra control de calidad, registro de parámetros, generación de tablas y figuras, y principios FAIR. El resultado esperado es una herramienta académica abierta, trazable y adaptable, concebida para complementar y no sustituir los programas consolidados, y para fortalecer capacidades regionales de análisis genético mediante procedimientos documentados de permutación y manejo de datos faltantes.***

Palabras clave-- AMOVA, genética poblacional, FST, genética forense, Centroamérica.

I. Introducción

El análisis de varianza molecular, conocido como AMOVA, constituye uno de los marcos estadísticos más utilizados para cuantificar cómo se distribuye la variación genética dentro y entre poblaciones. Su formulación permitió trasladar la lógica del análisis de varianza a matrices de distancias moleculares, generando componentes de varianza y estadísticos análogos a los índices de fijación [1]. En genética poblacional aplicada, estos estimadores se complementan con FST y otros indicadores de diferenciación, que permiten interpretar estructura poblacional, aislamiento, mezcla histórica y pertinencia de bases de frecuencia alélica [2].

En América Latina, iniciativas como el proyecto GHEP-ISFG han promovido la armonización de ejercicios colaborativos en genética forense y la construcción de bases de referencia regionales [3], pero el desarrollo de herramientas computacionales regionales propias ha sido escaso. La mayor parte de los laboratorios latinoamericanos depende de software desarrollado en Europa o Norteamérica, lo cual genera dependencia tecnológica y dificultades de adaptación a contextos institucionales con recursos limitados [3]. Además, plataformas emergentes como PopART, orientadas a análisis de redes de haplotipos, han demostrado que herramientas especializadas con interfaz gráfica accesible pueden tener un impacto significativo en la productividad de grupos de investigación [4].

En Centroamérica, el problema no es únicamente estadístico. Aunque existen herramientas consolidadas, como Arlequin, Genepop y paquetes especializados de R, su uso académico suele depender de formatos rígidos, procesos manuales de conversión, poca trazabilidad de parámetros y documentación limitada para usuarios regionales [5]-[7]. Esta situación dificulta la reproducibilidad, especialmente cuando los resultados se emplean en artículos científicos, tesis, peritajes o reportes institucionales. La reproducibilidad computacional exige registrar cómo se produjo cada resultado, conservar versiones de programas y datos, y evitar manipulaciones manuales no documentadas [8]-[10].

La necesidad de una herramienta regional se vuelve más evidente ante la reciente publicación de bases poblacionales hondureñas para marcadores STR autosómicos. La estandarización de 23 loci con PowerPlex Fusion 6C, plataforma validada para genética forense y compatible con estándares ampliados, ha permitido caracterizar poblaciones Lenca, Tawahka, Pech, Garífuna y Negro Inglés [11]-[12], [13]-[17]. Estos estudios muestran que las poblaciones indígenas y afrodescendientes de Honduras no deben analizarse como una sola referencia genérica, ya que las diferencias en heterocigosidad, riqueza alélica y FST pueden afectar la interpretación de estructura genética y la ponderación de evidencia forense.

El presente trabajo propone una herramienta computacional denominada AMOVA CA. Su finalidad es resolver una brecha operativa: la limitada disponibilidad de software académico reproducible, abierto y contextualizado para investigación genética poblacional regional. El producto científico esperado es un prototipo funcional y una estrategia de validación con datos poblacionales publicados, orientada a fortalecer

investigación universitaria, formación metodológica y colaboración interinstitucional en Centroamérica.

El alcance del prototipo debe delimitarse con precisión para evitar expectativas equivocadas. AMOVA CA es un prototipo académico en etapa inicial y no una alternativa a Arlequin, Genepop o a los paquetes de R, cuya cobertura de métodos, madurez y base de usuarios son muy superiores. En su versión actual cubre cuatro funciones: importación y validación estructural de datos, estadísticos de diversidad por locus, FST pareado y AMOVA jerárquica de hasta tres niveles con prueba por permutación, además de la exportación de tablas, figuras y una bitácora de parámetros. Queda fuera de este alcance el análisis de desequilibrio de ligamiento, el equilibrio de Hardy-Weinberg por locus, la asignación bayesiana de individuos, las pruebas de cuellos de botella, los marcadores de linaje y los datos de secuencia. Tampoco sustituye la revisión estadística especializada ni la autorización ética requerida para trabajar con genotipos individuales. Su aporte es la trazabilidad del flujo, no la ampliación del repertorio metodológico.

II. Fundamento metodológico

El prototipo se apoya en tres familias de resultados. La primera corresponde a diversidad intrapoblacional, que incluye número de alelos, heterocigosidad observada, heterocigosidad esperada, contenido de información polimórfica, poder de discriminación y probabilidad de exclusión. Para cada locus, la heterocigosidad esperada puede expresarse de forma general como la Fórmula (1):

$$He = 1 - \Sigma\, pk^2 \qquad (1)$$

Donde pk representa la frecuencia del alelo k en una población. La segunda familia corresponde a diferenciación entre poblaciones. Una formulación interpretativa simple de FST se basa en la diferencia relativa entre la heterocigosidad total y la heterocigosidad promedio dentro de poblaciones conforme la Fórmula (2):

$$FST = (HT - HS) / HT \qquad (2)$$

La tercera familia corresponde a AMOVA. En una jerarquía mínima, con individuos anidados en poblaciones y poblaciones anidadas en grupos, el componente análogo a ΦST puede resumirse como la Fórmula (3):

$$\Phi ST = \sigma^2 entre / (\sigma^2_{entre} + \sigma^2_{dentro}) \qquad (3)$$

La implementación clásica de AMOVA emplea una matriz de distancias euclidianas al cuadrado calculada a partir de las diferencias en el número de repeticiones entre alelos STR, bajo el modelo de mutación por pasos (Stepwise Mutation Model, SMM). Alternativamente, para aplicaciones forenses con marcadores altamente polimórficos, el modelo de alelos infinitos (Infinite Alleles Model, IAM) que asigna distancia binaria (igual o diferente) puede ser preferible, especialmente cuando las tasas de mutación entre loci son heterogéneas [18].

El prototipo ofrece al usuario la selección explícita del modelo de distancia, ya que la elección entre SMM e IAM puede alterar la magnitud de los componentes de varianza y los valores de Φ-estadísticos resultantes [1], [19].

Aunque las ecuaciones (1) a (3) son simplificaciones útiles para presentar el modelo, la implementación computacional debe conservar la estructura completa de distancias, grados de libertad, cuadrados medios, componentes de varianza y permutaciones, tal como se detalla en la sección III. Por ello, AMOVA CA no se concibe como una calculadora aislada, sino como un flujo reproducible que une importación, validación, análisis, visualización y exportación. Ver Fig. 1 para más detalle.

Desde el punto de vista de investigación abierta, el diseño adopta principios FAIR, ya que los datos y resultados deben ser encontrables, accesibles, interoperables y reutilizables cuando las condiciones éticas lo permitan [9]. En genética humana, esto exige una lectura cuidadosa. La apertura no implica publicar información identificable ni ignorar derechos colectivos. El prototipo, por tanto, prioriza datos agregados, seudonimización, bitácoras de acceso y exportación de metadatos, con advertencias explícitas cuando el usuario intenta analizar perfiles individuales sin documentación ética suficiente.



Fig. 1 Arquitectura funcional del prototipo AMOVA CA.

III. Materiales y métodos

Se diseñó un prototipo modular con enfoque de software académico reproducible. El módulo de entrada acepta matrices de genotipos diploides, tablas de frecuencias alélicas y archivos derivados de formatos comunes en genética poblacional. El módulo de validación examina presencia de loci, codificación de alelos, valores faltantes, consistencia de

tamaños muestrales y correspondencia entre población, individuo y grupo jerárquico. El módulo estadístico calcula diversidad por locus, FST pareado, escenarios de theta y AMOVA con permutaciones configurables; su implementación se detalla a continuación.

La validación se concibió en dos niveles. El primer nivel fue documental y verificó que los datos poblacionales seleccionados correspondieran a estudios publicados, con DOI, población definida, tamaño muestral, plataforma de genotipado y parámetros estadísticos reportados. El segundo nivel fue computacional y evaluó si el prototipo podía reproducir la estructura lógica de análisis, registrar parámetros y generar salidas auditables. La selección de fuentes incluyó estudios de Lenca, Tawahka, Pech, Garífuna y Negro Inglés, debido a que representan grupos indígenas y afrodescendientes de Honduras con datos STR recientes [13]-[17].

Para evitar una interpretación empírica excesiva, la validación se definió como prueba de funcionamiento y concordancia metodológica, no como una nueva estimación poblacional definitiva. Las inferencias de estructura fina requieren acceso a genotipos completos, autorización ética, armonización estricta de loci y revisión por expertos. En este manuscrito, los valores comparativos publicados se usan para demostrar aplicabilidad regional, construir tablas de control y mostrar cómo un flujo reproducible puede reducir errores de transcripción y dependencia de operaciones manuales. El diseño y sus componentes se especifican en la Tabla I.

TABLA I
DISEÑO FUNCIONAL DEL PROTOTIPO AMOVA CA

Módulo	Función	Entrada	Salida	Control reproducible
Importación	Lectura de CSV, Excel, GENEPOP y Arlequin.	Datos crudos	Objeto validado	Formato y versión.
Calidad	Revisión de loci, alelos, faltantes y metadatos.	Objeto validado	Advertencias	Bitácora.
Diversidad	Ho, He, Na, PIC, PD y CE.	Frecuencias	Tabla por locus	Fórmulas.
Diferenciación	FST pareado y theta.	Poblaciones	Matriz	Semilla y parámetros.
AMOVA	Varianza jerárquica.	Grupos	Phi y p	Permutaciones.
Reporte	Exportación de resultados.	Resultados	DOCX, XLSX, PNG	Hash y referencias.

El prototipo se implementó en Python 3.11 utilizando las librerías NumPy 1.26 para cálculo matricial, pandas 2.1 para manipulación de datos tabulares, SciPy 1.11 para pruebas estadísticas y permutaciones, y matplotlib 3.8 para generación de figuras. La lectura de formatos Genepop y Arlequin se realizó mediante parsers desarrollados ad hoc con expresiones regulares validadas contra archivos de ejemplo publicados. El control de versiones se gestionó mediante Git y el código se alojó en un repositorio institucional. Las pruebas unitarias se ejecutaron con pytest 7.4 y cubrieron las funciones de importación, cálculo de diversidad y generación de matrices FST. El entorno de ejecución fue documentado mediante archivos requirements.txt y se proporcionó un archivo Dockerfile para garantizar reproducibilidad en distintas plataformas [20].

La prueba de significancia de los Φ-estadísticos se implementó por permutación no paramétrica, con esquemas específicos para cada nivel de la jerarquía. Para el componente dentro de poblaciones se permutan los individuos entre poblaciones de un mismo grupo; para el componente entre poblaciones dentro de grupos se permutan las poblaciones completas entre grupos, conservando su tamaño; y para el componente entre grupos se permutan los individuos entre grupos sin restricción. En cada réplica se recalculan los cuadrados medios y los componentes de varianza a partir de la matriz de distancias original, que no se recalcula, lo que reduce el costo computacional. El valor p se obtiene como la proporción de réplicas cuyo estadístico iguala o supera el observado, con la corrección $(b + 1) / (B + 1)$, donde b es el número de réplicas iguales o mayores y B el número de permutaciones; esta corrección evita valores p iguales a cero. El número de permutaciones es configurable, con 1,023 como valor por defecto y 9,999 como opción recomendada para publicación. La aleatorización usa el generador PCG64 de NumPy inicializado con una semilla que el usuario fija o que el sistema genera y registra en la bitácora, de modo que cualquier corrida puede repetirse con resultados idénticos. Para las comparaciones pareadas múltiples, el usuario puede aplicar corrección de Bonferroni o de tasa de descubrimientos falsos, y la opción elegida queda documentada en el reporte.

El manejo de datos faltantes sigue reglas explícitas y reportadas. Los valores vacíos, los ceros y los códigos convencionales de ausencia se normalizan a un valor faltante único, y los genotipos parciales, con un solo alelo presente, se tratan como faltantes en ese locus. El prototipo no imputa alelos. La exclusión opera en dos umbrales configurables: se descarta el locus cuya proporción de datos faltantes supera el 5 % en cualquier población, y se descarta el individuo con más del 10 % de loci faltantes. Con los datos restantes, el cálculo de diversidad y de FST se realiza por eliminación por pares, es decir, cada locus aporta su propio tamaño muestral efectivo, mientras que la AMOVA exige una matriz de distancias completa y por tanto se ejecuta sobre el conjunto de loci sin faltantes, salvo que el usuario autorice el modo de distancias promediadas sobre los loci disponibles en cada par. El reporte incluye, para cada corrida, el número de loci e individuos excluidos, el tamaño muestral efectivo por locus y una advertencia cuando la exclusión supera el 10 % de los datos originales, condición en la que los resultados se marcan como preliminares.

IV. Resultados

El prototipo AMOVA CA quedó estructurado como una canalización de análisis con cinco salidas principales: resumen de calidad, estadísticas por locus, matriz de diferenciación, tabla AMOVA y reporte reproducible. La Fig. 1 muestra la arquitectura general. La Fig. 2 resume el flujo de validación, desde la incorporación de fuentes publicadas hasta la generación de reportes listos para revisión.

El diseño reduce tres fuentes habituales de error. Primero, evita conversiones manuales repetidas entre hojas de cálculo y programas especializados. Segundo, conserva los parámetros usados en cada corrida, incluyendo población, loci, método de distancia, número de permutaciones y semilla aleatoria. Tercero, separa los datos agregados de los perfiles individuales, lo que facilita una gobernanza compatible con investigación humana y con protección de comunidades indígenas y afrodescendientes.

El grado de intervención del usuario es amplio y queda registrado en cada corrida. En la importación, el usuario define el separador, la codificación de alelos, las columnas de individuo, población y grupo, y el umbral de datos faltantes. En el modelo, elige entre distancias SMM e IAM, define la jerarquía de hasta tres niveles, selecciona los loci que se incluyen y decide si el FST se estima por el método de Weir y Cockerham o por la formulación basada en heterocigosidades. En la inferencia, fija el número de permutaciones, la semilla, el nivel de significancia y la corrección por comparaciones múltiples. En la visualización, selecciona el tipo de gráfico entre barras comparativas, mapa de calor de distancias y diagrama de componentes de varianza, además de la paleta, los rótulos, el idioma de las figuras y la resolución de salida. Todas estas decisiones se almacenan en un archivo de configuración reutilizable, lo que permite reejecutar un análisis completo o compartirlo con otro usuario sin repetir la parametrización manual. Los valores por defecto están definidos para que un usuario sin experiencia obtenga una corrida válida, pero ninguno queda oculto.

La Tabla II permitió verificar que los cinco conjuntos poblacionales seleccionados son adecuados para validación regional. Los estudios Lenca, Tawahka, Pech, Garífuna y Negro Inglés usaron 23 loci STR autosómicos y muestras de 100 individuos no relacionados, lo que favorece comparabilidad metodológica. Aun así, el prototipo conserva advertencias sobre diferencias de población, muestreo, disponibilidad de genotipos completos y posible necesidad de armonización previa.

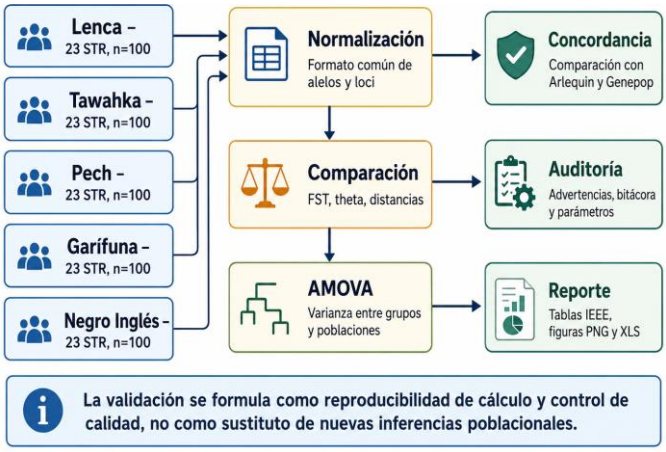


Fig. 2 Flujo de validación con datos poblacionales publicados.

TABLA II
DIVERSIDAD STR MEDIA REPORTADA EN LAS POBLACIONES USADAS PARA LA VALIDACIÓN

Población	He media	Alelos/locus
Lenca [13]	0.7425	8.91
Tawahka [14]	0.7385	8.61
Garífuna [15]	0.7893	9.43
Pech [16]	0.6681	6.43
Negro Inglés [17]	0.8150	10.48

Nota: He: heterocigosidad esperada. Para los estudios poblacionales se consideró una muestra (n) de 100 individuos no relacionados y sanos, y 23 loci. Los valores provienen de los estudios citados en la primera columna y se usan como referencia de contraste, no como estimaciones nuevas.

La Fig. 3 sintetiza la diversidad STR media reportada para las poblaciones incluidas. La comparación no busca ordenar poblaciones por valor biológico, sino demostrar que los perfiles de diversidad son suficientemente distintos para justificar bases específicas y análisis de diferenciación. Negro Inglés y Garífuna muestran valores medios más altos de heterocigosidad esperada y riqueza alélica, consistentes con la contribución afroatlántica reportada en sus estudios. Pech presenta menor heterocigosidad media y menor número de alelos por locus, lo que refuerza la importancia de no sustituir su referencia por bases genéricas.

El módulo de diferenciación incorporó como prueba los escenarios FST publicados para Pech frente a otras referencias. El estudio Pech reporta FST de 0.018 frente a Lenca, 0.022 frente a Tawahka, 0.038 frente a mestizo hondureño y 0.089 frente a una referencia caucásica NIST [16]. Estos valores permiten validar la lógica de matriz simétrica, selección de theta y generación de advertencias. La Fig. 4 muestra cómo el prototipo traduce esos resultados en escenarios operativos. En una aplicación académica, el usuario puede conservar el valor exacto de FST, exportar intervalos de confianza y documentar la razón de un theta conservadora. Más detalles en la Tabla III y la Tabla IV.

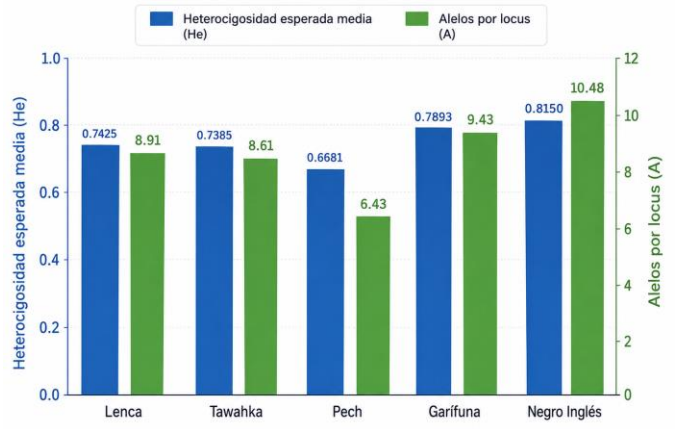


Fig. 3 Diversidad STR (Short Tandem Repeats) comparativa en poblaciones hondureñas publicadas.

TABLA III
MÉTRICAS DE VALIDACIÓN DEL PROTOTIPO

Componente	Prueba	Criterio	Resultado esperado
Datos	Campos obligatorios	Campos críticos presentes	Carga válida o advertencia.
Diversidad	Cálculos por locus	Diferencia ≤ 0.0001	Concordancia aritmética.
FST	Matriz pareada	Diagonal 0 y simetría	Matriz válida.
AMOVA	Semilla fija	Salida repetible	Reproducibilidad.
Reporte	Exportación	XLSX y PNG	Salida sin edición manual.

TABLA IV
ESCENARIOS FST PUBLICADOS PARA LA VALIDACIÓN DEL MÓDULO DE DIFERENCIACIÓN

Comparación	FST	IC 95% inferior	IC 95% superior	Uso operativo sugerido
Pech vs. Lenca	0.018	0.012	0.025	Theta 0.02 para comparación indígena.
Pech vs. Tawahka	0.022	0.015	0.030	Theta 0.02 para comparación indígena.
Pech vs. mestizo hondureño	0.038	0.029	0.048	Theta 0.03 en etnicidad incierta.
Pech vs. NIST Caucásico	0.089	0.078	0.101	Theta 0.05 a 0.10 en casos internacionales.

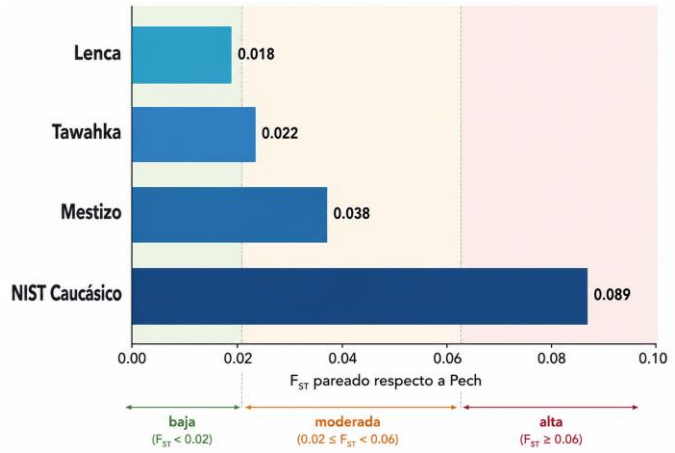


Fig. 4 Escenarios FST pareado derivados de datos publicados para la población Pech.

V. DISCUSIÓN

El valor científico del prototipo no reside en reemplazar herramientas consolidadas, sino en integrarlas conceptualmente dentro de un flujo regional más transparente. Arlequin y Genepop han sido fundamentales para genética poblacional, pero muchos usuarios universitarios continúan enfrentando barreras de formato, reproducibilidad y documentación [5], [6]. Un prototipo orientado a Centroamérica puede funcionar como puente entre docencia, investigación y práctica forense, siempre que conserve compatibilidad con estándares reconocidos y permita comparar resultados con programas de referencia.

A diferencia de Arlequin, que requiere archivos de entrada en formato propietario ".arp" y no genera reportes exportables directamente, AMOVA CA acepta formatos estándar (CSV, Excel) y produce salidas listas para manuscrito, , como las que ilustran las Fig. 3 y 4 y la Tabla IV. Comparado con los paquetes de R como pegas y hierfstat, el prototipo ofrece una interfaz gráfica que reduce la barrera de entrada para usuarios sin experiencia en programación, un factor relevante en universidades centroamericanas donde la formación en bioinformática es incipiente [21]. No obstante, la flexibilidad analítica de R y la comunidad de soporte de Arlequin representan ventajas que el prototipo no suplanta en su fase actual, y que definen los límites de alcance descritos en la sección I. Por ello, la estrategia de validación incluye exportar resultados en formatos compatibles con ambas herramientas para permitir verificación cruzada.

Conviene precisar en qué se diferencia el aporte y en qué no. La diferencia no está en los estimadores, que son los mismos de la literatura clásica [1], [2], sino en tres elementos del flujo. El primero es la trazabilidad: cada corrida produce una bitácora con versión del programa, loci incluidos, modelo de distancia, número de permutaciones, semilla y reglas de exclusión aplicadas, de modo que un tercero puede repetir el resultado sin reconstruir decisiones a partir del texto de un artículo. El segundo es la continuidad del flujo: importación, control de calidad, cálculo, figura y reporte ocurren en una misma canalización, lo que elimina los pasos manuales de conversión que la Fig. 2 identifica como fuente de error. El tercero es la adaptación al contexto regional: plantillas de metadatos poblacionales, documentación en español y advertencias éticas orientadas a poblaciones indígenas y afrodescendientes. Lo que el prototipo no ofrece es un repertorio metodológico comparable al de Arlequin o al ecosistema de R, ni una validación cuantitativa frente a sus salidas, que sigue pendiente según se indica en la sección VII.

La disponibilidad de bases poblacionales recientes en Honduras ofrece una oportunidad estratégica. La caracterización de Lenca, Tawahka, Pech, Garífuna y Negro Inglés demuestra que el país posee una diversidad humana que

no se captura adecuadamente con una sola base mestiza o con referencias hispanas generales [13]-[17]. Desde una perspectiva tecnológica, esta diversidad exige herramientas capaces de organizar metadatos poblacionales, documentar supuestos, informar limitaciones y producir resultados auditables. Los perfiles de diversidad de la Figura 3 y de la Tabla II ilustran esa exigencia: la distancia entre el valor de Pech y el de Negro Inglés es suficiente para que el uso de una base única introduzca sesgo en la ponderación de evidencia. Desde una perspectiva de negocios e innovación universitaria, el prototipo puede convertirse en una plataforma de servicios académicos, capacitación, análisis colaborativo y soporte a laboratorios públicos.

La diferenciación genética tiene implicaciones metodológicas y éticas. En contextos forenses, Buckleton y colaboradores muestran que los valores poblacionales de FST para STR pueden variar entre regiones y que su interpretación requiere cautela [22]. En investigación con pueblos indígenas y afrodescendientes, la cautela debe ser aún mayor. El prototipo incorpora advertencias para impedir que los valores de distancia se interpreten como categorías biológicas rígidas. Los escenarios de la Figura 4 y de la Tabla IV muestran cómo se implementa esa cautela: el prototipo conserva el valor estimado de FST y, por separado, sugiere un valor de theta más conservador para uso operativo, con el criterio documentado en la bitácora. Los resultados deben entenderse como medidas estadísticas dependientes de marcadores, muestras, historia demográfica y preguntas de investigación.

Una contribución adicional es la generación de salidas listas para manuscritos y auditoría. El usuario puede exportar tablas en Excel, figuras en PNG y reportes con referencias, parámetros y bitácora. Esto responde a una necesidad práctica en congresos y revistas, donde los revisores suelen solicitar datos reproducibles, trazabilidad de cálculo y claridad visual. La automatización de tablas y figuras reduce errores de transcripción, especialmente en estudios con múltiples loci y comparaciones pareadas.

VI. Gobernanza, sostenibilidad y modelo operativo

La sostenibilidad del prototipo requiere una gobernanza liviana pero explícita. Se propone un modelo de desarrollo universitario con tres capas. La primera es técnica, formada por repositorio de código, documentación, pruebas unitarias y control de versiones. La segunda es científica, formada por un comité metodológico que revise fórmulas, compatibilidad con software de referencia y pertinencia de indicadores. La tercera es ética, formada por reglas de acceso, uso de datos agregados, protección de perfiles individuales y consulta con comunidades cuando corresponda.

Como producto tecnológico, AMOVA CA puede implementarse inicialmente como aplicación local en Python o R, con interfaz gráfica sencilla y motor estadístico documentado. Una segunda fase puede incorporar una versión web institucional para cargar datos agregados y generar reportes. La tercera fase puede integrar repositorios regionales, plantillas de datos y módulos de enseñanza para cursos de genética poblacional, bioestadística, antropología molecular y genética forense. Esta progresión reduce riesgo tecnológico y permite validar cada componente antes de escalar.

El modelo de negocio académico no debe centrarse en comercializar datos genéticos. La propuesta sostenible es ofrecer servicios de capacitación, análisis reproducible, acompañamiento metodológico, generación de reportes y desarrollo de capacidades en universidades públicas. En ese escenario, la universidad conserva liderazgo científico, los laboratorios fortalecen su trazabilidad y los investigadores regionales disponen de herramientas menos dependientes de flujos cerrados o manuales.

En el contexto de pueblos indígenas y afrodescendientes de Honduras, la gobernanza de datos genéticos debe alinearse con los principios de soberanía de datos indígenas CARE (Collective benefit, Authority to control, Responsibility, Ethics), que complementan los principios FAIR al incorporar derechos colectivos sobre el uso de información biológica [23]. En la práctica, esto implica que el prototipo debe incluir campos de metadatos que registren si existe consentimiento comunitario, qué instituciones tienen acceso autorizado y bajo qué condiciones se permite la reutilización de datos agregados. Adicionalmente, el Convenio 169 de la OIT sobre pueblos indígenas y tribales establece obligaciones de consulta previa que deben considerarse cuando los datos genéticos se emplean en contextos forenses o de política pública [24].

VII. Limitaciones

Este estudio presenta varias limitaciones que deben considerarse. Primero, la validación se realizó exclusivamente con datos poblacionales hondureños, lo que limita la generalización inmediata a otras poblaciones centroamericanas como las de Guatemala, El Salvador o Panamá, donde existen grupos indígenas y afrodescendientes con historias demográficas distintas. Segundo, el prototipo fue validado conceptualmente con estadísticas publicadas (frecuencias alélicas y parámetros resumidos), pero no con genotipos individuales completos, lo cual impide una evaluación directa de la robustez del motor estadístico frente a datos crudos con valores faltantes elevados o errores de genotipado. Tercero, no se realizó una comparación cuantitativa formal (benchmarking) contra la salida exacta de Arlequin, Genepop o HIERFSTAT usando los mismos conjuntos de entrada, lo que debería abordarse en fases posteriores. Cuarto, la sostenibilidad del software académico en universidades públicas latinoamericanas enfrenta desafíos conocidos de financiamiento, rotación de personal técnico y mantenimiento a largo plazo, factores que podrían afectar la continuidad del

prototipo [25]-[27].

VIII. Trabajo futuro

El desarrollo futuro incluye tres líneas prioritarias. La primera es la integración de marcadores adicionales, incluyendo Y-STR e InDels de ancestría, que permitirían análisis de linajes paternos y estimación de mezcla continental, ampliando la utilidad del prototipo para antropología molecular y estudios de migración [28]. La segunda es la implementación de análisis de asignación poblacional mediante métodos bayesianos, como los empleados por STRUCTURE, lo cual permitiría al usuario no solo medir diferenciación sino inferir la composición ancestral probable de individuos o grupos [26]. La tercera línea es la incorporación de técnicas de visualización interactiva, como análisis de coordenadas principales (PCoA) y mapas de calor de distancia genética, mediante librerías como Plotly o D3.js, para facilitar la exploración visual de patrones de estructura poblacional en entornos web [29].

IX. Conclusiones

Este artículo presenta AMOVA CA, una herramienta computacional concebida para análisis de varianza molecular y diferenciación genética poblacional en Centroamérica. El prototipo responde a una brecha concreta: la falta de flujos académicos abiertos, reproducibles y adaptados al contexto regional. Su arquitectura integra importación de datos, control de calidad, cálculo de diversidad, FST pareado, AMOVA, visualización y exportación de resultados.

La validación con datos poblacionales hondureños publicados demuestra pertinencia regional y permite estructurar un conjunto inicial de pruebas. Los estudios Lenca, Tawahka, Pech, Garífuna y Negro Inglés ofrecen un escenario idóneo para evaluar formatos, comparabilidad, diversidad STR y diferenciación. La propuesta no reemplaza los programas consolidados ni la revisión estadísticas especializada, ni la autorización ética para uso de genotipos individuales, pero sí proporciona una base operativa para investigación reproducible.

En términos científicos, AMOVA CA puede fortalecer la formación, la colaboración y la publicación de estudios de genética poblacional en universidades públicas. En términos tecnológicos y de negocios, puede evolucionar hacia una plataforma de servicios académicos, capacitación y soporte metodológico. Su principal aporte es convertir cálculos dispersos en un flujo transparente, auditable y adecuado para una región con creciente producción de datos genéticos.

Agradecimientos

Agradecemos el apoyo recibido del Instituto de Investigación en Ciencias Médicas y Derecho a la Salud (ICIMEDES) de la Facultad de Ciencias Médicas (FCM) y la Dirección de Investigaciones Científicas, Humanísticas y Tecnológicas (DICIHT) todas ellas de la Universidad Nacional Autónoma de Honduras (UNAH).

Referencias

[1] L. Excoffier, P. E. Smouse, and J. M. Quattro, “Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data,” *Genetics*, vol. 131, no. 2, pp. 479–491, Jun. 1992, doi: 10.1093/genetics/131.2.479.

[2] B. S. Weir and C. C. Cockerham, “Estimating F-statistics for the analysis of population structure,” *Evolution*, vol. 38, no. 6, pp. 1358–1370, Nov. 1984, doi: 10.1111/j.1558-5646.1984.tb05657.x.

[3] A. Salas et al., "An updated analysis of the mitochondrial DNA macro-haplogroup phylogeny for forensic purposes", *Forensic Sci. Int. Genet.*, vol. 61, Art. no. 102778, Nov. 2022, doi: 10.1016/j.fsigen.2022.102778

[4] J. W. Leigh and D. Bryant, "POPART: Full-feature software for haplotype network construction", *Methods Ecol. Evol.*, vol. 6, no. 9, pp. 1110–1116, Sep. 2015, doi: 10.1111/2041-210X.12410

[5] L. Excoffier and H. E. L. Lischer, “Arlequin suite ver. 3.5: A new series of programs to perform population genetics analyses under Linux and Windows,” *Mol. Ecol. Resour.*, vol. 10, no. 3, pp. 564–567, May 2010, doi: 10.1111/j.1755-0998.2010.02847.x.

[6] F. Rousset, “GENEPOP’007: A complete re-implementation of the GENEPOP software for Windows and Linux,” *Mol. Ecol. Resour.*, vol. 8, no. 1, pp. 103–106, Jan. 2008, doi: 10.1111/j.1471-8286.2007.01931.x.

[7] J. Goudet, “HIERFSTAT, a package for R to compute and test hierarchical F-statistics,” *Mol. Ecol. Notes*, vol. 5, no. 1, pp. 184–186, Mar. 2005, doi: 10.1111/j.1471-8286.2004.00828.x.

[8] R. D. Peng, “Reproducible research in computational science,” *Science*, vol. 334, no. 6060, pp. 1226–1227, Dec. 2011, doi: 10.1126/science.1213847.

[9] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, Art. no. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

[10] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, “Ten simple rules for reproducible computational research,” *PLOS Comput. Biol.*, vol. 9, no. 10, Art. no. e1003285, Oct. 2013, doi: 10.1371/journal.pcbi.1003285.

[11] M. G. Ensenberger et al., “Developmental validation of the PowerPlex Fusion 6C System,” *Forensic Sci. Int. Genet.*, vol. 21, pp. 134–144, Mar. 2016, doi: 10.1016/j.fsigen.2015.12.011.

[12] D. R. Hares, “Selection and implementation of expanded CODIS core loci in the United States,” *Forensic Sci. Int. Genet.*, vol. 17, pp. 33–34, Jul. 2015, doi: 10.1016/j.fsigen.2015.03.006.

[13] A. Zuniga et al., “Population genetic data for 23 STR loci (PowerPlex Fusion 6C™ kit) genetic markers in the Lenca ethnic group in Honduras,” *Leg. Med. (Tokyo)*, vol. 71, Art. no. 102504, Nov. 2024, doi: 10.1016/j.legalmed.2024.102504.

[14] A. Zuniga et al., “Population genetic data for 23 STR loci of Tawahka ethnic group in Honduras,” *Forensic Sci.*, vol. 5, no. 4, Art. no. 72, Dec. 2025, doi: 10.3390/forensicsci5040072.

[15] A. Zuniga et al., “Population genetic data for 23 STR loci of the Garifuna ethnic group in Honduras,” *Genes (Basel)*, vol. 17, no. 4, Art. no. 402, Mar. 2026, doi: 10.3390/genes17040402.

[16] A. Zuniga et al., “Population genetic data for 23 STR loci of the Pech ethnic group in Honduras,” *Genes (Basel)*, vol. 17, no. 4, Art. no. 422, Apr. 2026, doi: 10.3390/genes17040422.

[17] A. Zuniga et al., “Population genetic data for 23 STR loci of the Black Caribbean ethnic group in Honduras,” *Genes (Basel)*, vol. 17, no. 5, Art. no. 496, Apr. 2026, doi: 10.3390/genes17050496.

[18] M. Slatkin, "A measure of population subdivision based on microsatellite allele frequencies," *Genetics*, vol. 139, no. 1, pp. 457–462, Jan. 1995, doi: 10.1093/genetics/139.1.457.

[19] S. Michalakis and L. Excoffier, "A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci," *Genetics*, vol. 142, no. 3, pp. 1061–1064, Mar. 1996, doi: 10.1093/genetics/142.3.1061.

[20]F. Mölder et al., "Sustainable data analysis with Snakemake", *F1000Research*, vol. 10, Art. no. 33, Jan. 2021, doi: 10.12688/f1000research.29032.2.
[21]T. Jombart, "adegenet: A R package for the multivariate analysis of genetic markers", *Bioinformatics*, vol. 24, no. 11, pp. 1403–1405, Jun. 2008, doi: 10.1093/bioinformatics/btn129.
[22]J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, and B. S. Weir, "Population-specific FST values for forensic STR markers: A worldwide survey," *Forensic Sci. Int. Genet.*, vol. 23, pp. 91–100, Jul. 2016, doi: 10.1016/j.fsigen.2016.03.004.
[23]S. R. Carroll et al., "The CARE Principles for Indigenous Data Governance", *Data Sci. J.*, vol. 19, no. 1, Art. no. 43, Nov. 2020, doi: 10.5334/dsj-2020-043.
[24]International Labour Organization, "Indigenous and Tribal Peoples Convention (No. 169)", Geneva, Switzerland, Jun. 1989. [Online]. Available: https://www.ilo.org/convention/C169 Accessed: Sep. 24, 2026
[25]G. Wilson et al., "Best practices for scientific computing", *PLOS Biol.*, vol. 12, no. 1, Art. no. e1001745, Jan. 2014, doi: 10.1371/journal.pbio.1001745.
[26]J. E. Howison and J. D. Herbsleb, "Scientific software production: Incentives and collaboration", in *Proc. ACM Conf. Comput. Supported Cooperative Work (CSCW)*, San Antonio, TX, USA, 2011, pp. 513–522, doi: 10.1145/1958824.1958904.
[27]A. Caliebe, M. Krawczak, and S. Willuweit, "Y-chromosomal STR mutation rates: Theoretical considerations and empirical results", *Forensic Sci. Int. Genet.*, vol. 44, Art. no. 102162, Jan. 2020, doi: 10.1016/j.fsigen.2019.102162.
[28]J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data", *Genetics*, vol. 155, no. 2, pp. 945–959, Jun. 2000, doi: 10.1093/genetics/155.2.945.
[29]M. Bostock, V. Ogievetsky, and J. Heer, "D³: Data-driven documents", *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011, doi: 10.1109/TVCG.2011.185.