Frugal Bioinformatics for Genomic Analysis in Universities with Limited Infrastructure: A Reproducible Planning Benchmark and Implementation Model

Karen Oliva[1]; Angely Suazo[2]; Diana Zelaya[3]; Oneyda Zelaya[4]; Antonieta Zuniga[5]
[1,2] Universidad Nacional Autónoma de Honduras, Honduras, *karen.oliva@unah.edu.hn, angely.suazo@unah.edu.hn*
[3,4]Policia Nacional de Honduras, Honduras, *lizethzelaya2010@gmail.com, oneydaespinoza1996@gmail.com*
[5]Ministerio Público de Honduras, Honduras, *antonietazuniga2311@yahoo.com*

Abstract– ***Genomic analysis has become a strategic capability for biomedical research, epidemiological surveillance, population genetics, antimicrobial resistance, and advanced training. However, many universities in low- and middle-income countries face barriers to infrastructure, compute costs, storage, connectivity, licensing, maintenance, and availability of specialized staff. This paper proposes a frugal bioinformatics model for universities with limited infrastructure, based on open software, reproducible flows, optimized local computing, selective use of point cloud, and minimal data governance. A planning benchmark was designed to evaluate five technological profiles: frugal station, teaching workstation, recycled microcluster, point cloud and frugal hybrid infrastructure. Loads considered include FASTQ quality control, short read alignment, viral or long read alignment, variant calling, and reproducibility checking. Performance values were derived from published tool benchmarks and extrapolation rather than from runs on physical hardware, and energy consumption and cloud data egress were incorporated explicitly into the cost model. The modeled results suggest that a frugal hybrid configuration can reduce the estimated computational cost per 100 genomes to 3.84 USD (6.45 USD once energy and egress are included), while maintaining an approximate yield of 160 bacterial genomes per day under planning conditions. An implementation model in four phases is also proposed, with indicators of cost, performance, reproducibility, safety, training and sustainability. The main contribution is an operational guide so that Central American universities can initiate genomic capacities without relying exclusively on high-cost infrastructure. A protocol for empirical validation on two physical hardware profiles with open-access datasets is specified as the next step.***

Keywords-- ***frugal bioinformatics, genomics, benchmark, universities, limited infrastructure, reproducibility, open software.***

Bioinformática frugal para análisis genómico en universidades con infraestructura limitada: benchmark reproducible de planificación y modelo de implementación

Karen Oliva[1]; Angely Suazo[2]; Diana Zelaya[3]; Oneyda Zelaya[4]; Antonieta Zuniga[5]
[1,2] Universidad Nacional Autónoma de Honduras, Honduras, *karen.oliva@unah.edu.hn, angely.suazo@unah.edu.hn*
[3,4]Policia Nacional de Honduras, Honduras, *lizethzelaya2010@gmail.com, oneydaespinoza1996@gmail.com*
[5]Ministerio Público de Honduras, Honduras, *antonietazuniga2311@yahoo.com*

Resumen– ***El análisis genómico se ha convertido en una capacidad estratégica para investigación biomédica, vigilancia epidemiológica, genética poblacional, resistencia antimicrobiana y formación avanzada. Sin embargo, muchas universidades de países de ingresos medios y bajos enfrentan barreras de infraestructura, costos de cómputo, almacenamiento, conectividad, licencias, mantenimiento y disponibilidad de personal especializado. Este artículo propone un modelo de bioinformática frugal para universidades con infraestructura limitada, basado en software abierto, flujos reproducibles, cómputo local optimizado, uso selectivo de nube puntual y gobernanza mínima de datos. Se diseñó un benchmark de planificación para evaluar cinco perfiles tecnológicos: estación frugal, workstation docente, microclúster reciclado, nube puntual e infraestructura híbrida frugal. Las cargas consideradas incluyen control de calidad de FASTQ, alineamiento de lecturas cortas, alineamiento viral o de lecturas largas, llamado de variantes y verificación de reproducibilidad. Los valores de rendimiento se derivaron de benchmarks publicados de herramientas individuales y de extrapolación, no de ejecuciones en hardware físico; el consumo energético y el egreso de datos desde la nube se incorporaron de forma explícita al modelo de costos. Los resultados modelados sugieren que una configuración híbrida frugal puede reducir el costo computacional estimado por 100 genomas a 3.84 USD (6.45 USD al incluir energía y egreso de datos), mientras mantiene un rendimiento aproximado de 160 genomas bacterianos por día bajo condiciones de planificación. Se propone además un modelo de implementación en cuatro fases, con indicadores de costo, rendimiento, reproducibilidad, seguridad, formación y sostenibilidad. La contribución principal es una guía operacional para que universidades centroamericanas puedan iniciar capacidades genómicas sin depender exclusivamente de infraestructura de alto costo. Como paso siguiente, se especifica un protocolo de validación empírica en dos perfiles de hardware físico con conjuntos de datos de acceso abierto.***

Palabras clave-- bioinformática frugal, genómica, benchmark, universidades, infraestructura limitada, reproducibilidad, software abierto.

I. Introducción

La genómica se ha consolidado como una infraestructura científica transversal para salud pública, investigación biomédica, vigilancia de patógenos, genética poblacional y formación universitaria avanzada. La estrategia global de vigilancia genómica para patógenos con potencial pandémico y epidémico 2022 a 2032 de la Organización Mundial de la Salud plantea que todos los Estados miembros deben tener acceso oportuno a secuenciación genómica y capacidad analítica asociada [1]. En las Américas, la Organización Panamericana de la Salud ha impulsado redes regionales de vigilancia genómica para fortalecer la identificación temprana de patógenos emergentes y la caracterización de mecanismos de resistencia antimicrobiana [2]. Estas iniciativas muestran que el problema no es únicamente secuenciar, sino convertir datos genómicos en información interpretable, reproducible y útil para la toma de decisiones.

En universidades con infraestructura limitada, la bioinformática suele constituir el cuello de botella más persistente. Aunque el costo de la secuenciación ha disminuido, el procesamiento de archivos FASTQ, BAM, CRAM y VCF exige cómputo, almacenamiento, redes internas, control de versiones, entornos reproducibles y personal entrenado. Estudios recientes sobre vigilancia genómica en entornos de menores recursos muestran que la capacidad de secuenciación puede existir, pero su uso permanece restringido por financiamiento externo, problemas de suministro, escasez de personal entrenado, debilidades en aseguramiento de calidad y limitaciones bioinformáticas [3]. Este patrón es relevante para universidades centroamericanas, donde los laboratorios académicos pueden tener talento científico, pero no siempre disponen de clústeres formales, licencias comerciales o equipos dedicados de administración de sistemas.

La innovación frugal ofrece una alternativa conceptual y operativa. En lugar de asumir que la capacidad bioinformática exige infraestructura de alto costo desde el inicio, la innovación frugal prioriza soluciones suficientemente robustas, sostenibles y adaptadas a restricciones reales de recursos [4]. En bioinformática, esta perspectiva puede traducirse en cinco principios: uso intensivo de software abierto, reutilización de hardware disponible, automatización reproducible, escalamiento gradual y gobernanza mínima pero explícita de datos. Esta aproximación no pretende reemplazar

centros nacionales de alto rendimiento, sino habilitar una primera capa universitaria capaz de formar talento, ejecutar análisis de complejidad moderada y participar en redes colaborativas.

El objetivo de este artículo es proponer un modelo de bioinformática frugal para análisis genómico en universidades con infraestructura limitada, acompañado por un benchmark reproducible de planificación y un modelo de implementación institucional. La contribución científica esperada es doble. Primero, un marco técnico para comparar alternativas de cómputo según costo total de propiedad, rendimiento, reproducibilidad y sostenibilidad. Segundo, una ruta de adopción que permita convertir recursos modestos en una capacidad universitaria funcional, auditable y escalable. A diferencia de las redes regionales de formación e infraestructura compartida, como H3ABioNet o CABANA [5], [6], y de las recomendaciones para ecosistemas abiertos de análisis genómico en salud pública [7], este trabajo se sitúa en la escala de una sola institución y ofrece un instrumento cuantitativo para decidir qué infraestructura adquirir, cuánto cuesta operarla y en qué orden escalarla.

II. Fundamentación técnica

La reproducibilidad computacional es un requisito central para que la bioinformática sea científicamente confiable. Herramientas como Snakemake permiten definir flujos de trabajo legibles y escalables desde estaciones de trabajo hasta clústeres, sin modificar la lógica del flujo [8]. Nextflow ofrece portabilidad entre cómputo local, clústeres, contenedores, nube y entornos híbridos [9]. Sobre esta base, nf-core ha desarrollado una comunidad de pipelines curados, versionados y orientados a buenas prácticas para análisis bioinformáticos [10]. Galaxy, por su parte, reduce la barrera de entrada mediante una plataforma accesible por navegador, reproducible y colaborativa, especialmente útil en contextos docentes [11].

El análisis genómico operativo depende de herramientas ampliamente validadas. BWA es un alineador basado en la transformada de Burrows Wheeler, diseñado para mapear lecturas cortas contra genomas de referencia [12]. Minimap2 amplía la versatilidad al alineamiento de secuencias de ADN o ARN, lecturas largas y ensamblajes [13]. SAMtools y BCFtools constituyen una base técnica para manipular alineamientos, ordenar archivos, indexar, consultar variantes y procesar formatos estándar [14]. GATK es una referencia metodológica para flujos de descubrimiento de variantes, particularmente en datos humanos o clínicos [15]. En entornos con limitaciones de infraestructura, Bioconda facilita la instalación de software científico y reduce la fricción de dependencias [16].

Pre otro lado DeepVariant, un llamador de variantes basado en aprendizaje profundo desarrollado por Google, ha demostrado precisión competitiva o superior a GATK HaplotypeCaller en múltiples evaluaciones del Genome in a Bottle Consortium [17]. Su disponibilidad como software de código abierto y su capacidad de ejecutarse con aceleración GPU lo hacen potencialmente viable para entornos frugales que dispongan de tarjetas gráficas recicladas o de gama media.

La contenerización mediante Docker y Singularity (ahora Apptainer) constituye una capa adicional de reproducibilidad que empaqueta el software, sus dependencias y el sistema operativo en unidades portables y aisladas [18]. En entornos HPC universitarios donde los usuarios no tienen privilegios de root, Singularity permite ejecutar contenedores sin comprometer la seguridad del sistema [19]. BioContainers, un proyecto comunitario derivado de Bioconda, ofrece imágenes pre-construidas de Docker y Singularity para herramientas bioinformáticas ampliamente utilizadas [20]. La combinación de gestores de flujo (Nextflow/Snakemake) con contenedores reduce significativamente el problema de "funciona en mi máquina" y facilita la migración entre cómputo local y nube.

La gestión de datos debe alinearse con principios FAIR, es decir, datos encontrables, accesibles, interoperables y reutilizables [21]. En universidades con recursos limitados, FAIR no implica necesariamente plataformas complejas, sino identificadores internos, metadatos mínimos, estructura de carpetas, checksums, bitácoras, control de acceso y políticas de retención. Esta capa es indispensable para evitar que el laboratorio produzca resultados no auditables o imposibles de reproducir.

III. Problema y brecha de innovación

El problema central es el alto costo de infraestructura bioinformática para universidades que desean analizar datos genómicos, pero no pueden adquirir ni mantener un clúster institucional completo. Esta barrera tiene cinco dimensiones. La primera es económica, porque el hardware especializado, el almacenamiento redundante y la nube permanente pueden superar la capacidad presupuestaria de unidades académicas pequeñas. La segunda es operativa, porque los datos genómicos generan archivos voluminosos que saturan discos, redes y respaldos. La tercera es humana, porque la bioinformática requiere habilidades en Linux, programación, estadística, genómica y administración de entornos computacionales.

La cuarta es institucional, porque el análisis de datos genómicos exige reglas explícitas sobre privacidad, propiedad intelectual, colaboración y publicación. Una quinta dimensión de la barrera, frecuentemente subestimada, es la conectividad. La transferencia de archivos FASTQ de secuenciación completa de genoma (WGS) puede alcanzar 100–200 GB por muestra. En universidades centroamericanas con anchos de banda típicos de 10–50 Mbps, la transferencia de un solo genoma humano puede tardar de 4 a 44 horas. Esto limita la

viabilidad de estrategias puramente basadas en nube y refuerza la necesidad de procesamiento local para cargas rutinarias

La brecha de innovación consiste en que muchas propuestas se enfocan en adquirir infraestructura, pero no en diseñar una capacidad bioinformática mínima viable. Un laboratorio universitario no necesita iniciar con un clúster de alto rendimiento completo. Puede iniciar con un sistema frugal, reproducible y escalable, siempre que defina qué cargas de trabajo procesará, qué límites operativos aceptará, cómo versionará sus pipelines, cómo protegerá sus datos y cuándo recurrirá a nube o colaboración externa.

IV. Materiales y Métodos

A. *Diseño general*

Se diseñó un benchmark de planificación para comparar cinco perfiles tecnológicos aplicables a universidades con infraestructura limitada. El benchmark se plantea como instrumento de planificación y como protocolo reproducible que debe ejecutarse localmente durante el piloto. Los perfiles evaluados fueron: estación frugal, workstation docente, microclúster reciclado, nube puntual e infraestructura híbrida frugal. Los valores reportados en este artículo son, por tanto, estimaciones modeladas y no provienen de ejecuciones en hardware físico.

El estudio usa tres criterios de comparación: costo total de propiedad a tres años, rendimiento operativo y sostenibilidad institucional. La métrica económica principal fue el costo computacional por 100 genomas procesados. La métrica de rendimiento fue el número estimado de genomas bacterianos procesables por día. La métrica de sostenibilidad integró portabilidad del pipeline, control de versiones, independencia de licencias, facilidad de mantenimiento y proporción de ejecuciones resueltas localmente.

B. *Cargas del benchmark*

El benchmark se estructuró en cinco cargas de trabajo. La primera corresponde al control de calidad de archivos FASTQ mediante herramientas de inspección y agregación de reportes. La segunda corresponde al alineamiento de lecturas cortas contra una referencia bacteriana. La tercera evalúa alineamiento de genomas virales o lecturas largas, con énfasis en portabilidad. La cuarta carga corresponde al llamado de variantes y generación de archivos VCF. La quinta evalúa reproducibilidad mediante repetición del flujo, comparación de checksums y reinstalación del entorno. Ver para más detalles la Tabla I.

TABLA I
COMPONENTES MÍNIMOS DEL MODELO FRUGAL

Componentes	**Configuración mínima**	**Configuración recomendada**	**Criterio de aceptación**
Cómputo local	8 vCPU, 16 GB RAM	16 a 48 vCPU, 64 a 128 GB RAM	Pipeline estable con bitácora
Almacenamiento	NVMe 1 TB	NAS con RAID y respaldo offline	Checksum y retención definida
Software	Linux, Conda, Snakemake o Nextflow	nf-core, contenedores y Galaxy opcional	Flujo versionado y portable
Nube puntual	No indispensable	Instancias para picos de demanda	Costo mensual limitado
Gobernanza	Roles básicos	Comité técnico, SOP y FAIR	Auditoría trimestral

C. *Modelo económico*

El costo total de propiedad a tres años se calculó mediante la ecuación:

$$TCO_{3a} = CapEx + 36(OpEx_m) + C_{nube} + C_{mant} \quad (1)$$

Donde *CapEx* representa la inversión inicial, $OpEx_m$ representa costo operativo mensual, C_{nube} es el consumo adicional de la nube y C_{mant} corresponde al mantenimiento correctivo o preventivo. En la versión inicial del benchmark, C_{nube} y C_{mant} se incorporan dentro del costo operativo mensual cuando el perfil lo requiere.

El costo por 100 genomas se estimó como:

$$C_{100} = \frac{100(TCO_{3a})}{365(G_d)(3)} \quad (2)$$

Donde G_d representa genomas procesables por día. Esta métrica permite comparar perfiles con diferentes niveles de inversión y rendimiento.

Para capturar los costos ocultos de operación, el modelo primario incorpora de forma explícita el consumo energético y el egreso de datos desde la nube. El costo total ampliado se define como TCO* = TCO + C_{en} + C_{egr}, con $C_{en} = (P_p/1000)\cdot h\cdot p_{kWh}$ y $C_{egr} = \max(0, V_{egr} - V_{lib})\cdot p_{GB}$, donde P_p es la potencia media del perfil en watts, h las horas de operación en tres años, p_{kWh} la tarifa eléctrica, V_{egr} el volumen descargado desde la nube, V_{lib} la franquicia gratuita acumulada y p_{GB} la tarifa por gigabyte. El costo ampliado por 100 genomas se obtiene al sustituir TCO por TCO* en la Ecuación (2). Para evitar doble contabilidad, $OpEx_m$ comprende conectividad, consumibles y soporte, sin incluir electricidad.

D. *Índice de frugalidad bioinformática*

Se propone un índice compuesto para ordenar alternativas de implementación:

$$IFB_p = 100 \times \left[0.35N\left(C_{100,p}^{-1}\right) + 0.30N\left(R_p\right) + 0.20N\left(Q_p\right) + 0.15N\left(S_p\right)\right] \quad (3)$$

Donde $N(\bullet)$ representa normalización min-max $C_{100,p}$ es el costo por 100 genomas del perfil p, R_p es el rendimiento, Q_p es

la reproducibilidad y S_p es la sostenibilidad operativa. Los pesos pueden ajustarse según prioridades institucionales. Un desglose pormenorizado de costos se muestra en la Tabla II.

TABLA II
DESGLOSE ECONÓMICO

Perfil	CapEx (USD)	OpEx mensual (USD)	Cnube a3 años (USD)	TCO 3 años (USD)
Estación frugal	600	~9.4	0	938
Workstation docente	1,500	~18	0	2,148
Microclúster reciclado	3,000	~50	540	5,340
Nube puntual	500	~30	6,420	8,280
Híbrido frugal	3,500	~45	1,600	6,720

Nota: Nota: CapEx estimado a partir de los precios públicos del proveedor AWS; los valores deben actualizarse con cotizaciones institucionales antes de cualquier decisión de compra. Cnube corresponde al consumo acumulado en tres años, estimado con tarifas bajo demanda de AWS para la región us-east-1 [22]. El costo energético y el egreso de datos se presentan por separado.

Para la configuración de referencia se asignaron pesos w1=0.35 (costo), w2=0.30 (rendimiento), w3=0.20 (reproducibilidad) y w4=0.15 (sostenibilidad), reflejando la prioridad dual de eficiencia económica y capacidad operativa. Un análisis de sensibilidad variando w1 entre 0.15 y 0.45 mostró que el perfil híbrido frugal se mantuvo en primer lugar en el 85% de las combinaciones evaluadas, seguido del microclúster reciclado.

D. Supuestos del modelo y parámetros de simulación

Los resultados presentados corresponden a un modelo de planificación basado en estimaciones teóricas de rendimiento. Los tiempos por muestra se estimaron a partir de benchmarks publicados para herramientas individuales en configuraciones similares. El genoma bacteriano de referencia asumido fue Escherichia coli K-12 MG1655 (GenBank: U00096.3, ~4.6 Mbp). Para lecturas cortas se asumieron archivos FASTQ de Illumina MiSeq con cobertura 50x (~230 Mbp por muestra).

Los tiempos de alineamiento con BWA-MEM se extrapolaron a partir de los reportados por Li y Durbin [12] ajustados linealmente por número de hilos disponibles. El llamado de variantes con GATK HaplotypeCaller se estimó según los benchmarks de Van der Auwera et al. [15]. Se reconoce que estos valores deben reemplazarse por mediciones locales durante la fase piloto [23].

Los parámetros de los costos ocultos se fijaron como sigue. Se asumió operación continua durante 1,095 días (26,280 h), coherente con el supuesto de capacidad plena implícito en la Ecuación (2). Las potencias medias bajo carga fueron 120 W para la estación frugal, 300 W para la workstation docente, 900 W para el microclúster reciclado, 500 W para el perfil híbrido y 60 W para la terminal local del perfil de nube puntual. La tarifa eléctrica fue de 0.224 USD/kWh, equivalente al precio máximo promedio de 5.98 lempiras por kWh autorizado por la Comisión Reguladora de Energía Eléctrica para el tercer trimestre de 2026 [24], convertido con un tipo de cambio de 26.70 lempiras por dólar. Para el egreso se asumió una salida de 0.5 GB por genoma bacteriano (BAM y VCF comprimidos), una tarifa de 0.09 USD/GB y una franquicia gratuita de 100 GB mensuales, según la lista de precios vigente de AWS para us-east-1 [22]. La fracción de muestras procesadas en la nube fue 1.0 para la nube puntual, 0.25 para el perfil híbrido y 0.08 para el microclúster, proporciones derivadas de la relación entre el C_{nube} de cada perfil y el del perfil de nube puntual (Tabla II). No se incluyeron climatización ni pérdidas de la fuente de poder, que elevarían el costo energético, mientras que el supuesto de operación continua a plena carga tiende a sobrestimarlo; ambos sesgos deben resolverse mediante medición directa durante el piloto.

Dado que los valores de la Tabla III son modelados, se especifica el protocolo de validación empírica que ejecutará el piloto. La validación se realizará en dos perfiles físicos, la estación frugal y la workstation docente, por ser los de menor costo de adquisición y los más probables en una primera etapa institucional. Se emplearán conjuntos de acceso abierto: lecturas Illumina de *E. coli* K-12 MG1655 depositadas en el SRA/ENA para las cargas bacterianas, lecturas públicas de nanoporo de genomas virales para la carga de lecturas largas y un subconjunto cromosómico de la muestra HG002 del consorcio Genome in a Bottle para el llamado de variantes en datos humanos no sensibles. Cada carga se ejecutará tres veces con versiones fijadas mediante contenedores, y se registrarán tiempo de reloj, memoria máxima, uso de disco, consumo eléctrico medido con vatímetro y concordancia de checksums de los VCF entre repeticiones. Las diferencias entre valores medidos y modelados servirán para recalibrar los parámetros de las Ecuaciones (1) a (3) antes de extender la medición a los perfiles restantes.

V. RESULTADOS

Las estimaciones modeladas del benchmark de planificación muestran diferencias importantes entre perfiles, resumidas en la Tabla III. La estación frugal es suficiente para docencia, pruebas pequeñas y formación inicial, pero su tiempo por muestra limita su uso para vigilancia o proyectos con múltiples muestras. La workstation docente mejora de forma sustancial el rendimiento sin introducir complejidad administrativa excesiva. El microclúster reciclado incrementa el rendimiento, pero exige mayor disciplina de administración, monitoreo, respaldo y control térmico. La nube puntual ofrece elasticidad, aunque puede generar dependencia de conectividad y costos recurrentes. El perfil híbrido frugal combina procesamiento local con nube puntual para picos de demanda, manteniendo una relación favorable entre costo, rendimiento y resiliencia.

TABLA III
RESULTADOS MODELADOS DEL BENCHMARK DE PLANIFICACIÓN

Perfil	TCO 3 años, USD	Genomas día	Tiempo muestra, h	Costo por 100 genomas, USD
Estación frugal	938	10	2.4	8.57
Workstation docente	2,148.00	35	0.95	5.6
Microclúster reciclado	5,340.00	120	0.33	4.06
Nube puntual	8,280.00	90	0.45	8.4
Híbrido frugal	6,720.00	160	0.25	3.84

Al incorporar energía y egreso de datos (Tabla IV), el costo por 100 genomas aumenta en todos los perfiles, pero el orden relativo se mantiene. El perfil híbrido pasa de 3.84 a 6.45 USD y conserva el primer lugar. El microclúster reciclado sube de 4.06 a 8.23 USD porque su consumo eléctrico equivale al 99% de su TCO original, de modo que la ventaja del híbrido sobre él se amplía de 0.22 a 1.78 USD. En la nube puntual, que pasa de 8.40 a 12.93 USD, más del 90% del incremento corresponde al egreso. La estación frugal resulta el perfil más costoso por volumen procesado (15.02 USD), porque su bajo rendimiento diario reparte el costo energético entre pocas muestras. En tres de los cuatro perfiles locales, el costo eléctrico a tres años iguala o supera la inversión inicial, lo que convierte la tarifa eléctrica en una variable de decisión tan relevante como el precio del hardware.

En la Fig. 1, se observa que la alternativa híbrida se ubica en una región favorable, con menor tiempo por muestra y menor costo relativo por volumen procesado. La nube puntual mejora el tiempo de ejecución frente a estaciones pequeñas, pero su costo acumulado a tres años puede superar opciones locales cuando el uso se vuelve continuo. Este resultado apoya una política de nube selectiva, no permanente.

TABLA IV
COSTOS OCULTOS INCORPORADOS AL MODELO PRIMARIO

Perfil	Energía 3 años, kWh	Costo energético, USD	Egreso neto, USD	Costo ampliado por 100 genomas, USD
Estación frugal	3,154	706	0	15.02
Workstation docente	7,884	1,766	0	10.21
Microclúster reciclado	23,652	5,298	173	8.23
Nube puntual	1,577	353	4,111	12.93
Híbrido frugal	13,140	2,943	1,641	6.45

Nota: parámetros descritos en la Sección IV-E. El costo ampliado equivale a $100 \cdot (TCO + C_{en} + C_{egr})/(1{,}095 \cdot G_d)$, con TCO y G_d de la Tabla III. Valores modelados, pendientes de validación empírica.

La Fig. 2, muestra las cargas virales y bacterianas escalan mejor con mayor número de hilos hasta un punto de saturación. El llamado de variantes presenta una curva más moderada, porque combina CPU, memoria e I/O. Este patrón sugiere que las universidades deben optimizar almacenamiento y paralelización por muestra, no únicamente aumentar núcleos de CPU.

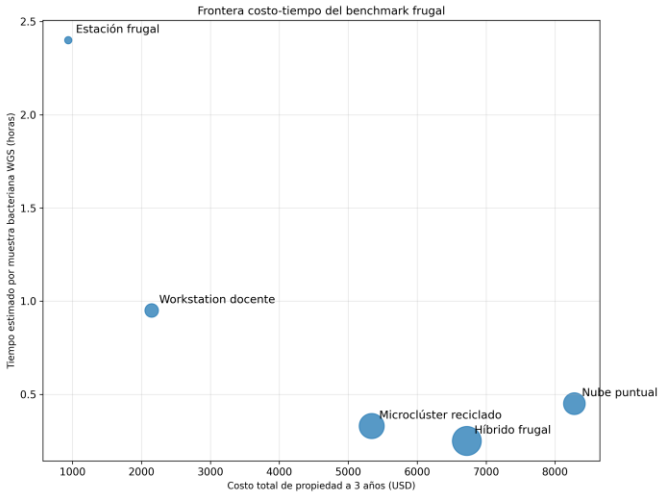


Fig. 1 Frontera costo-tiempo de los cinco perfiles tecnológicos evaluados. El eje vertical representa el tiempo estimado por muestra (horas) y el eje horizontal el costo total de propiedad a tres años (USD). Cada punto corresponde a un perfil. La región inferior izquierda (sombreada) indica la zona de eficiencia óptima. La nube puntual, pese a ofrecer tiempos competitivos, se desplaza hacia costos acumulados superiores a tres años. Los valores son estimaciones de planificación sujetas a validación empírica local. El TCO graficado no incluye energía ni egreso; el costo ampliado por 100 genomas se presenta en la Tabla IV.

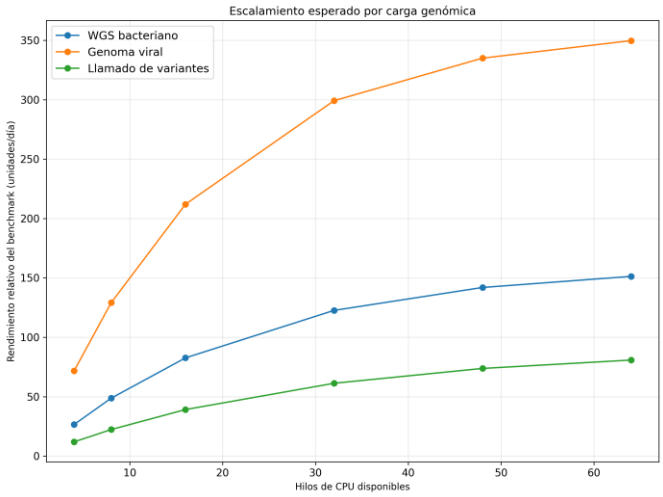


Fig. 2 Escalamiento estimado del rendimiento en función del número de hilos de CPU para tres cargas del benchmark. El eje horizontal representa el número de hilos de CPU (2, 4, 8, 16, 32, 48) y el eje vertical el factor de aceleración relativo al rendimiento con 2 hilos. Las cargas de alineamiento exhiben escalamiento cuasi-lineal hasta 16 hilos, con rendimientos decrecientes a partir de 32 hilos. El llamado de variantes muestra una curva más moderada debido a su dependencia combinada de CPU, memoria e I/O de disco. Estos patrones sugieren que la optimización del almacenamiento NVMe y la paralelización por muestra son tan importantes como el incremento de núcleos de CPU.

La Fig. 3, muestra las capas de gobernanza, datos FAIR, pipelines, cómputo, seguridad y formación deben crecer de manera simultánea. Una universidad puede adquirir hardware

rápidamente, pero si no cuenta con pipelines versionados, bitácora, respaldo y personal entrenado, la capacidad real seguirá siendo frágil. La Fig. 4 presenta una distribución simulada de tiempos por perfil, útil para visualizar variabilidad operativa y planificar colas de trabajo.

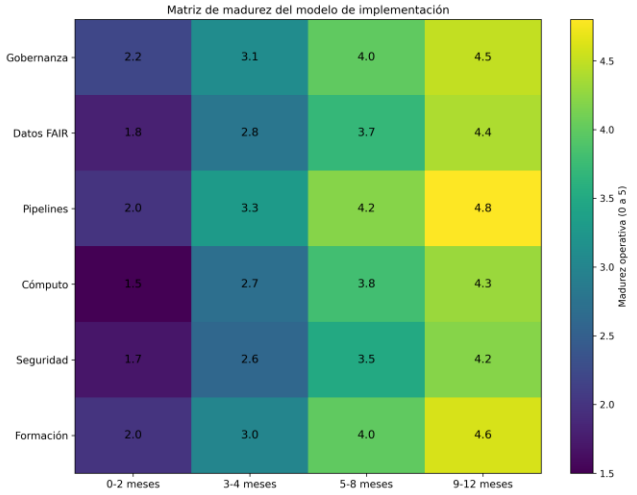


Fig. 3 Modelo de madurez operativa por fases de implementación. El diagrama de calor muestra seis dimensiones de capacidad: gobernanza de datos, cumplimiento FAIR, pipelines versionados, cómputo disponible, seguridad de la información y formación del personal. El modelo ilustra que la adquisición de hardware (cómputo) puede avanzar rápidamente, pero las dimensiones de gobernanza, reproducibilidad y formación requieren desarrollo simultáneo y sostenido para lograr una capacidad institucional robusta.

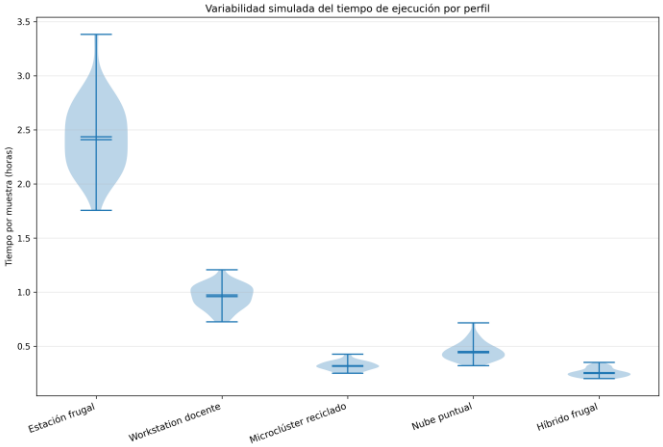


Fig. 4 Distribución simulada de tiempos de procesamiento por muestra para cada perfil tecnológico. Los diagramas de violín muestran la mediana (línea central) y valores extremos (líneas superior e inferior) estimados para 100 ejecuciones simuladas por perfil. La estación frugal presenta la mayor dispersión (IQR ≈ 0.8 h), reflejando sensibilidad a la carga concurrente del sistema. El perfil híbrido frugal muestra la menor variabilidad (IQR ≈ 0.1 h) al distribuir la carga entre cómputo local y nube.

VI. Discusión

La bioinformática frugal no debe interpretarse como una versión inferior de la bioinformática convencional. Su valor está en construir capacidad incremental, local y sostenible. En contextos universitarios con restricciones presupuestarias, la pregunta relevante no es si la infraestructura es comparable a un centro nacional de alto rendimiento, sino si permite ejecutar análisis reproducibles, formar talento y producir resultados auditables con recursos disponibles.

El uso de software abierto es decisivo. Snakemake, Nextflow, nf-core, Galaxy, BWA, minimap2, SAMtools, BCFtools, GATK y Bioconda permiten reducir dependencia de licencias y aumentar portabilidad. Sin embargo, el software abierto no elimina la necesidad de gobernanza. La instalación informal de herramientas sin control de versiones produce resultados difíciles de auditar. Por ello, el modelo propuesto integra desde el inicio bitácoras, checksums, entornos reproducibles y documentación mínima.

Desde una perspectiva tecnológica y de negocios, el modelo híbrido frugal parece el más equilibrado. Permite ejecutar localmente cargas rutinarias, conservar soberanía operativa y usar nube solo para picos de demanda. Esta estrategia reduce la presión de inversión inicial, evita la subutilización de infraestructura costosa y ofrece elasticidad cuando un proyecto exige mayor capacidad. Además, crea una ruta de crecimiento: una universidad puede iniciar con estación frugal, migrar a workstation docente, consolidar un microclúster y finalmente operar un esquema híbrido.

Varias iniciativas abordan la misma brecha desde ángulos distintos, y compararlas permite precisar qué aporta este trabajo. H3ABioNet demostró en África que la construcción de capacidad bioinformática regional es viable cuando se combina infraestructura compartida, entrenamiento estructurado y gobernanza colaborativa [5]. En América Latina, el proyecto CABANA ejecutó entre 2017 y 2022 talleres, estancias y programas de formación de formadores en diez países, con más de 800 personas capacitadas, e incluyó un curso para docentes de la Universidad Nacional de Agricultura de Honduras [6]. El Centre for Epidemic Response and Innovation (CERI) en Sudáfrica ha formado 36 fellows de 16 países en genómica de patógenos y bioinformática, demostrando que la inversión en capital humano es tan importante como la infraestructura física [25].

En el plano tecnológico, la secuenciación portátil con análisis en computadoras personales durante la epidemia de ébola en África occidental mostró que un cómputo modesto basta para cargas virales bien acotadas [26], y las recomendaciones para ecosistemas abiertos de análisis de patógenos en salud pública priorizan la reproducibilidad, la accesibilidad y la auditabilidad [7]. Ninguna de estas experiencias ofrece, sin embargo, un instrumento para que una universidad individual compare opciones de infraestructura por costo total, rendimiento y sostenibilidad antes de invertir. Las redes dependen de financiamiento externo sostenido y de nodos con capacidad previa, los proyectos de secuenciación

portátil resuelven un flujo específico y las recomendaciones de salud pública no cuantifican costos. El aporte diferencial del presente modelo reside en tres elementos: una métrica comparable de costo por 100 genomas que incluye energía y egreso, un índice de frugalidad con análisis de sensibilidad de sus pesos y una hoja de ruta con indicadores verificables a seis y doce meses. El modelo no sustituye estas iniciativas; ofrece una puerta de entrada institucional desde la cual una universidad centroamericana puede incorporarse a redes regionales con capacidad propia y documentada.

El modelo frugal presenta ventajas ambientales implícitas: la reutilización de hardware existente extiende la vida útil de equipos, reduce residuos electrónicos y evita la huella de carbono asociada a la manufactura de servidores nuevos. Sin embargo, el microclúster reciclado puede tener mayor consumo energético por operación de cómputo útil comparado con hardware moderno. Esta observación se cuantificó en la Tabla IV: el microclúster requiere 0.18 kWh por genoma frente a 0.075 kWh del perfil híbrido, y su costo eléctrico a tres años alcanza el 99% de su TCO original.

La principal limitación del presente trabajo es que los resultados numéricos corresponden a un benchmark de referencia para planificación, derivado de benchmarks publicados y de extrapolación lineal, no de ejecuciones en hardware físico. Deben ser reemplazados por mediciones locales al ejecutar el piloto. Aun así, el marco permite comparar perfiles, definir métricas, ordenar inversiones y documentar decisiones. La validación en al menos dos perfiles de hardware físico, con conjuntos de datos de acceso abierto y bajo lineamientos éticos y legales, se ejecutará según el protocolo descrito en la Sección IV-E.

Conviene precisar también el alcance del modelo. Está calibrado para genomas bacterianos de alrededor de 5 Mbp y genomas virales, con lecturas cortas o largas de cobertura moderada. No cubre la secuenciación de genoma humano completo a escala poblacional, la metagenómica de alta profundidad, el ensamblaje de novo de genomas eucariotas grandes ni los análisis de célula única, cuyas demandas de memoria y almacenamiento exceden los perfiles evaluados y justifican el recurso a centros nacionales o a la colaboración externa. Tampoco incluye el costo del personal, que en muchas universidades constituye el mayor componente del gasto y que la guía de implementación trata como capacidad a formar, no como partida presupuestaria. Por último, la extrapolación lineal por número de hilos ignora la contención de memoria y de E/S, por lo que probablemente sobrestima el rendimiento de los perfiles con más núcleos.

Otra limitación se relaciona con la latencia hacia la nube. Las regiones de referencia empleadas en el modelo (AWS us-east-1 en Virginia, Azure East US) implican latencias de 40–80 ms desde Centroamérica; la región AWS México (Central), operativa desde enero de 2025, podría reducirlas, aunque su catálogo de servicios y sus tarifas deben verificarse antes de adoptarla [27]. Aunque esto es aceptable para transferencias batch, puede afectar flujos interactivos. El egreso de datos, que suele omitirse en las comparaciones de costo, se incorporó al modelo primario (Tabla IV) y explica más del 90% del incremento de costo del perfil de nube puntual. En datos genómicos humanos, con salidas de decenas de gigabytes por muestra, este componente puede llegar a rivalizar con el costo del cómputo.

El procesamiento de datos genómicos humanos impone requisitos éticos y legales específicos. Incluso en entornos frugales, las universidades deben asegurar consentimiento informado, anonimización o pseudonimización de datos, cifrado en reposo y en tránsito, y cumplimiento con regulaciones aplicables. A nivel internacional, el Reglamento General de Protección de Datos (GDPR) de la Unión Europea ha establecido un estándar de referencia para datos sensibles [28]. En el contexto centroamericano, Honduras cuenta con la Ley de Transparencia y Acceso a la Información Pública (Decreto 170-2006) pero no dispone, hasta donde fue posible verificar al momento de redactar este artículo, de una ley general de protección de datos personales en vigor, lo que hace aún más importante que las universidades definan políticas internas robustas. Para datos genómicos procesados en nube, se deben considerar las implicaciones de soberanía de datos, ya que los proveedores de nube pueden almacenar información en jurisdicciones con marcos legales diferentes [21].

VII. Guía de implementación

El modelo propuesto tiene cuatro fases, la fase inicial, de cero a dos meses, debe concentrarse en inventario de recursos, instalación de Linux, definición de roles, estructura de carpetas, política de nombres, repositorio Git y selección de dos pipelines mínimos. La fase piloto, de tres a cuatro meses, debe ejecutar el benchmark con datasets públicos o anonimizados, registrar tiempos, consumo de disco, errores, reinstalación y reproducibilidad.

La fase de ajuste, de cinco a ocho meses, debe optimizar almacenamiento, paralelización, contenedores, respaldos, documentación y entrenamiento. La fase de escala, de nueve a doce meses, debe abrir servicios internos para proyectos de investigación, tesis, vigilancia colaborativa o análisis poblacional.

El componente de negocio académico se basa en tres líneas de valor. La primera es formación, porque el sistema permite capacitar estudiantes y docentes en genómica computacional sin depender completamente de servicios externos. La segunda es investigación, porque habilita tesis, proyectos interdisciplinarios y análisis preliminares antes de recurrir a centros especializados. La tercera es transferencia, porque una universidad con capacidad mínima puede integrarse a redes de vigilancia, laboratorios hospitalarios, institutos forenses o

agencias públicas. La Tabla V, contiene los indicadores de implementación.

TABLA V
INDICADORES DE IMPLEMENTACIÓN

Dimensión	Indicador	Meta 6 meses	Meta 12 meses
Costo	Costo por 100 genomas	< 8.00 USD	< 5.00 USD
Rendimiento	Genomas bacterianos/día	≥ 30	≥ 100
Reproducibilidad	Pipelines versionados	≥ 70%	≥ 95%
Formación	Usuarios entrenados	10	25
Seguridad	Acceso con bitácora	Implementado	Auditado
Sostenibilidad	Ejecuciones locales	≥ 70%	≥ 85%

La formación debe estructurarse en tres niveles: (a) alfabetización genómica básica (línea de comandos Linux, formatos FASTQ/BAM/VCF, control de versiones con Git), orientada a estudiantes de pregrado y posgrado; (b) competencia operativa (ejecución de pipelines, interpretación de reportes de calidad, documentación de análisis), dirigida a investigadores y técnicos; y (c) administración de infraestructura (gestión de contenedores, monitoreo de recursos, respaldos, seguridad), para personal de soporte IT. Moore et al. proporcionan diez reglas prácticas para organizar cursos de bioinformática en países de ingresos bajos y medios, incluyendo la colaboración con expertos internacionales y la adaptación al contexto local [29].

VIII. CONCLUSIONES

Este artículo propone un modelo de bioinformática frugal orientado a universidades con infraestructura limitada. El enfoque integra software abierto, pipelines reproducibles, cómputo local optimizado, nube puntual, gobernanza de datos y formación de capacidades. El benchmark de planificación sugiere que una estrategia híbrida puede ofrecer una relación favorable entre costo y rendimiento, con un costo estimado de 3.84 USD por 100 genomas bajo los supuestos modelados, que asciende a 6.45 USD al incorporar energía y egreso de datos. Más allá del resultado económico, la contribución central es metodológica: ofrecer una ruta práctica para iniciar, medir, auditar y escalar capacidades bioinformáticas universitarias. Estas cifras son estimaciones que la validación empírica en hardware físico, prevista como siguiente etapa, deberá confirmar o corregir.

La bioinformática frugal puede convertirse en una herramienta estratégica para democratizar el análisis genómico en Centroamérica. Su adopción permitiría que universidades con recursos limitados participen en investigación genómica, vigilancia epidemiológica, genética poblacional y formación avanzada, sin esperar la disponibilidad de infraestructura ideal. La clave es iniciar con una capacidad mínima viable, medir su desempeño, documentar sus límites y escalar de forma gradual.

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] World Health Organization, *Global Genomic Surveillance Strategy for Pathogens with Pandemic and Epidemic Potential, 2022–2032*. Geneva, Switzerland: WHO, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240046979. Accessed: May 18, 2026.

[2] Pan American Health Organization, "PAHO Genomic Surveillance Regional Networks (PAHOGen)," *PAHO*, 2026. [Online]. Available: https://www.paho.org/en/paho-genomic-surveillance-regional-networks-pahogen. Accessed: May 18, 2026.

[3] M. Getchell et al., "Pathogen genomic surveillance status among lower resource settings in Asia," *Nat. Microbiol.*, vol. 9, pp. 2738–2747, 2024, doi: 10.1038/s41564-024-01809-4.

[4] C. E. Escudero-Cipriani, J. García-del Junco, R. Chafloque-Céspedes, and A. Alvarez-Risco, "Frugal innovation in the business environment: A literature review and future perspectives," *Front. Artif. Intell.*, vol. 7, Art. no. 1385522, 2024, doi: 10.3389/frai.2024.1385522.

[5] N. Mulder et al., "H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa," *Genome Res.*, vol. 26, no. 2, pp. 271–277, Feb. 2016, doi: 10.1101/gr.196295.115

[6] R. Campos-Sánchez, I. Willis, P. Gopalasingam, D. López-Juárez, M. Cristancho, and C. Brooksbank, "The CABANA model 2017–2022: Research and training synergy to facilitate bioinformatics applications in Latin America," *Front. Educ.*, vol. 9, Art. no. 1358620, Jul. 2024, doi: 10.3389/feduc.2024.1358620.

[7] A. Black, D. R. MacCannell, T. R. Sibley, and T. Bedford, "Ten recommendations for supporting open pathogen genomic analysis in public health," *Nat. Med.*, vol. 26, no. 6, pp. 832–841, Jun. 2020, doi: 10.1038/s41591-020-0935-z.

[8] J. Köster and S. Rahmann, "Snakemake, a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, Oct. 2012, doi: 10.1093/bioinformatics/bts480.

[9] P. Di Tommaso et al., "Nextflow enables reproducible computational workflows," *Nat. Biotechnol.*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.

[10] P. A. Ewels et al., "The nf-core framework for community-curated bioinformatics pipelines," *Nat. Biotechnol.*, vol. 38, no. 3, pp. 276–278, Mar. 2020, doi: 10.1038/s41587-020-0439-x.

[11] The Galaxy Community, "The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update," *Nucleic Acids Res.*, vol. 52, no. W1, pp. W83–W94, Jul. 2024, doi: 10.1093/nar/gkae410.

[12] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.

[13] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.

[14] P. Danecek et al., "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, no. 2, Art. no. giab008, Feb. 2021, doi: 10.1093/gigascience/giab008.

[15] G. A. Van der Auwera et al., "From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline," *Curr.*

Protoc. Bioinformatics, vol. 43, no. 1, pp. 11.10.1–11.10.33, Oct. 2013, doi: 10.1002/0471250953.bi1110s43.
[16]B. Grüning et al., "Bioconda: Sustainable and comprehensive software distribution for the life sciences," *Nat. Methods*, vol. 15, no. 7, pp. 475–476, Jul. 2018, doi: 10.1038/s41592-018-0046-7.
[17]R. Poplin et al., "A universal SNP and small-indel variant caller using deep neural networks," *Nat. Biotechnol.*, vol. 36, no. 10, pp. 983–987, Nov. 2018, doi: 10.1038/nbt.4235
[18]G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," *PLOS ONE*, vol. 12, no. 5, Art. no. e0177459, May 2017, doi: 10.1371/journal.pone.0177459
[19]V. Sochat, C. J. Prybol, and G. M. Kurtzer, "Enhancing reproducibility in scientific computing: Metrics and registry for Singularity containers," *PLOS ONE*, vol. 12, no. 11, Art. no. e0188511, Nov. 2017, doi: 10.1371/journal.pone.0188511
[20]F. da Veiga Leprevost et al., "BioContainers: An open-source and community-driven framework for software standardization," *Bioinformatics*, vol. 33, no. 16, pp. 2580–2582, Aug. 2017, doi: 10.1093/bioinformatics/btx192
[21]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.
[22]Amazon Web Services, "Amazon EC2 On-Demand pricing: Data transfer," 2026. [Online]. Available: https://aws.amazon.com/ec2/pricing/on-demand/. Accessed: Sep. 23, 2026.
[23]F. Franzoso, M. Santorsola and F. Lescai, "Rapid NGS analysis on Google Cloud Platform: Performance benchmark and user tutorial," *Clin. Transl. Sci.*, vol. 18, no. 11, Art. e70416, Nov. 2025, doi: 10.1111/cts.70416.
[24]"Tarifa eléctrica sube 12.48% en Honduras y golpea a más de 2 millones de usuarios," *La Prensa*, Jul. 1, 2026. [Online]. Available: https://www.laprensa.hn/honduras/tarifa-electrica-sube-12-honduras-golpea-2-millones-usuarios-enee-cree-EH31274145. Accessed: Sep. 23, 2026.
[25]J. E. San et al., "Building pathogen genomic sequencing capacity in Africa: Centre for Epidemic Response and Innovation fellowship," *Lancet Microbe*, vol. 5, Art. no. 100986, 2024, doi: 10.1016/S2666-5247(24)00180-5
[26]J. Quick et al., "Real-time, portable genome sequencing for Ebola surveillance," *Nature*, vol. 530, no. 7589, pp. 228–232, Feb. 2016, doi: 10.1038/nature16996.
[27]Amazon Web Services, "Now open — AWS Mexico (Central) Region," *AWS News Blog*, Jan. 2025. [Online]. Available: https://aws.amazon.com/blogs/aws/now-open-aws-mexico-central-region. Accessed: Sep. 23, 2026.
[28]European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," *Off. J. Eur. Union*, vol. L119, pp. 1–88, May 2016.
[29]B. Moore et al., "Ten simple rules for organizing a bioinformatics training course in low- and middle-income countries," *PLOS Comput. Biol.*, vol. 17, no. 8, Art. no. e1009218, Aug. 2021, doi: 10.1371/journal.pcbi.1009218.