



UNIVERSIDAD PERUANA
CAYETANO HEREDIA

AMPLIFICACIONES SERIADAS A TRAVÉS
DE REACCIONES ISOTÉRMICAS CON
RECOMBINASA POLIMERASA Y
EVALUACIÓN DE LA HIPÓTESIS DE LA
PRESIÓN MUTACIONAL
TERMODINÁMICA EN CONDICIONES
MESÓFILAS SIN PRESIÓN DE SELECCIÓN

TESIS PARA OPTAR EL GRADO DE
MAESTRO EN BIOQUÍMICA Y BIOLOGÍA
MOLECULAR

PEDRO FRANCISCO HUERTA ROQUE

LIMA – PERÚ

2026

ASESOR

DR. MIRKO JUAN ZIMIC PERALTA

CO ASESOR

DRA. PATRICIA SHEEN CORTAVARRIA

JURADO DE TESIS

DR. WILFREDO EVANGELISTA FALCON

PRESIDENTE

DR. FELIPE CARLOS YON TORRES

VOCAL

DR. LUIS ESTEBAN CABRERA SOSA

SECRETARIO

DEDICATORIA.

A Dios, por guiarme con su sabiduría (Romanos 11:36).

A mi esposa, por su constante y abnegado apoyo.

A mis padres y hermanos, por sus oraciones.

AGRADECIMIENTOS.

Al equipo científico del Laboratorio de Bioinformática y Biología Molecular quienes me brindaron las herramientas adecuadas para la culminación del trabajo.

FUENTES DE FINANCIAMIENTO.

Este trabajo fue subvencionado por CONCYTEC a través del programa PROCIENCIA en el marco del concurso "E041-2024-03", según contrato o convenio [PE501087234-2024]; nombre de la subvención: Efecto de la "Presión Mutacional Termodinámica" en la Evolución Molecular.



UNIVERSIDAD PERUANA
CAYETANO HEREDIA

DECLARACIÓN DE ORIGINALIDAD

Los egresados:

Nº	APELLIDOS Y NOMBRES
1.	HUERTA ROQUE PEDRO FRANCISCO

Pertenecientes al programa de la **MAESTRÍA EN BIOQUÍMICA Y BIOLOGÍA MOLECULAR**, autores del trabajo titulado: **AMPLIFICACIONES SERIADAS A TRAVÉS DE REACCIONES ISOTÉRMICAS CON RECOMBINASA POLIMERASA Y EVALUACIÓN DE LA HIPÓTESIS DE LA PRESIÓN MUTACIONAL TERMODINÁMICA EN CONDICIONES MESÓFILAS SIN PRESIÓN DE SELECCIÓN**, el cual ha sido elaborado, sustentado y aprobado, según corresponda, para optar por el grado de **MAESTRO EN BIOQUÍMICA Y BIOLOGÍA MOLECULAR** bajo la modalidad de **TESIS**.

En calidad de docentes asesores de la Universidad Peruana Cayetano Heredia:

Nº	APELLIDOS Y NOMBRES DEL DOCENTE	FACULTAD	NIVEL DE ASESORÍA
1.	ZIMIC PERALTA MIRKO JUAN	FACI	ASESOR
2.	SHEEN CORTAVARRIA PATRICIA	FACI	CO ASESOR

Declaramos que el contenido del presente documento es original y que las citas y referencias a otros autores cumplen con las normas académicas establecidas. En ese sentido, hacemos constar que:

- El documento presenta un porcentaje de similitud de **4%**, según el reporte emitido por el software **Turnitin®** (identificador de entrega: **3027587846**; fecha de entrega: **09-09-2026**).
- Tras una revisión detallada del reporte y del contenido del trabajo en cuestión, no se han identificado indicios de plagio.
- Se certifica que el documento respeta los principios de integridad académica y cumple con los requisitos institucionales de originalidad.

Lugar y fecha: **Lima, 09 de septiembre de 2026**

Firma del asesor
Nº DNI: 07620624
ORCID: 0000-0002-7203-8847

Firma del Co-asesor
Nº DNI: 09541127
ORCID: 0000-0002-7118-9301

ÍNDICE

RESUMEN
ABSTRACT

I.	INTRODUCCIÓN	1
II.	OBJETIVOS	15
III.	HIPÓTESIS	17
IV.	MARCO TEÓRICO	20
V.	METODOLOGÍA	28
VI.	RESULTADOS O ARGUMENTACIÓN TEÓRICA	45
VII.	DISCUSIÓN	79
VIII.	CONCLUSIONES	92
IX.	RECOMENDACIONES	93
X.	REFERENCIAS BIBLIOGRÁFICAS	95
XI.	ANEXOS	

RESUMEN

La teoría evolutiva clásica propone que las mutaciones ocurren principalmente de manera aleatoria y que la selección natural determina posteriormente su destino evolutivo. Sin embargo, modelos basados en termodinámica estadística han sugerido que las propiedades fisicoquímicas del ADN podrían introducir sesgos intrínsecos durante el proceso de replicación. En particular, la hipótesis de la presión mutacional termodinámica plantea que las diferencias de estabilidad energética entre los pares de bases influyen en la probabilidad de incorporación de nucleótidos, favoreciendo la formación de pares GC frente a AT en ausencia de presión selectiva. El presente estudio evaluó esta hipótesis mediante un modelo experimental de amplificaciones seriadas del ADN con Recombinasa Polimerasa (RPA) a 37 °C. Se seleccionaron tres genes de *Escherichia coli* con diferentes contenidos iniciales de GC (*evgA*, *ygiV* y *phnH*), los cuales fueron amplificados de manera iterativa, donde cada ronda de amplificación representó una generación del proceso simulado. Los productos obtenidos fueron purificados y secuenciados mediante tecnología Oxford Nanopore, alcanzando profundidades de cobertura entre 300,000× y 7,000,000× por generación. Las lecturas obtenidas fueron procesadas mediante un flujo bioinformático que incluyó análisis de variantes, aplicando filtros estrictos de profundidad, frecuencia alélica y balance por hebra. Las mutaciones detectadas se clasificaron según su efecto sobre el contenido GC y se evaluó la trayectoria composicional a lo largo de las generaciones mediante análisis de correlación y regresión ponderada. La amplificación seriada permitió simular múltiples generaciones y detectar mutaciones acumuladas en las poblaciones de ADN. El análisis del espectro mutacional reveló un sesgo neto hacia

sustituciones que incrementan el contenido de GC (A/T→G/C). Asimismo, el porcentaje de GC mostró una tendencia creciente a lo largo de las generaciones, con pendientes estimadas entre 0.137×10^{-3} y 0.645×10^{-3} puntos porcentuales de GC por generación, concordante con la predicción de la hipótesis de presión mutacional termodinámica bajo condiciones sin presión selectiva. En conclusión, los resultados obtenidos aportan evidencia experimental de que las diferencias en la energía potencial de interacción entre el nucleótido entrante y el entorno local del ADN durante la replicación pueden influir en la dirección de la acumulación de mutaciones, particularmente favoreciendo cambios asociados al incremento del contenido GC.

PALABRAS CLAVES: PRESIÓN MUTACIONAL TERMODINÁMICA, EVOLUCIÓN MOLECULAR, REPLICACIÓN DEL ADN, AMPLIFICACIÓN ISOTÉRMICA, SECUENCIACIÓN NANOPORE.

ABSTRACT

Classical evolutionary theory proposes that mutations occur primarily randomly and that natural selection subsequently determines their evolutionary fate. However, models based on statistical thermodynamics have suggested that the physicochemical properties of DNA could introduce intrinsic biases during the replication process. In particular, the thermodynamic mutational pressure hypothesis states that differences in energy stability between base pairs may influence the probability of nucleotide incorporation, favoring the formation of GC pairs over AT pairs in the absence of selective pressure. This study evaluated this hypothesis using an experimental model of serial DNA amplifications with Recombinase Polymerase (RPA) at 37 °C. Three *Escherichia coli* genes with different initial GC contents (*evgA*, *ygjV*, and *phnH*) were selected and iteratively amplified, with each round of amplification representing a generation of the simulated process. The resulting products were purified and sequenced using Oxford Nanopore technology, achieving coverage depths between 300,000× and 7,000,000× per generation. The resulting reads were processed using a bioinformatics workflow that included variant analysis, applying strict filters for depth, allele frequency, and per-strand balance. Detected mutations were classified according to their effect on GC content, and the compositional trajectory across generations was evaluated using correlation and weighted regression analyses. Serial amplification allowed for the simulation of multiple generations and the detection of accumulated mutations in the DNA populations. Analysis of the mutational spectrum revealed a net bias toward substitutions that increase GC content (A/T→G/C). Similarly, the percentage of GC showed an increasing trend

across generations, with estimated slopes between 0.137×10^{-3} and 0.645×10^{-3} percentage points of GC per generation, consistent with the prediction of the thermodynamic mutational pressure hypothesis under conditions without selective pressure. In conclusion, the results obtained provide experimental evidence that differences in the interaction potential energy between the incoming nucleotide and the local DNA environment during replication can influence the direction of mutation accumulation, particularly favoring changes associated with increased GC content.

KEYWORDS: THERMODYNAMIC MUTATIONAL PRESSURE, MOLECULAR EVOLUTION, DNA REPLICATION, ISOTHERMAL AMPLIFICATION, NANOPORE SEQUENCING.

I. INTRODUCCIÓN

1.1 Contexto teórico e hipótesis de la presión mutacional termodinámica

En la teoría evolutiva clásica se considera que las mutaciones ocurren aleatoriamente, mientras que la selección natural actúa posteriormente como filtro eliminando o favoreciendo aquellas variantes que afectan la adaptación de los organismos [1–5]. Sin embargo, evidencias a nivel genómico sugieren que la aparición de mutaciones podría no ser completamente aleatoria, ni en su frecuencia y distribución. Desde mediados del siglo XX, se ha discutido ampliamente qué fuerzas moldean la composición nucleotídica de los genomas. Observaciones como el sesgo en el uso de codones sinónimos (codon bias) y las diferencias marcadas en el contenido global de GC entre distintos organismos indican que, además de la selección natural, otros factores podrían contribuir a modelar los patrones de evolución molecular. En este contexto surge la hipótesis de la presión mutacional termodinámica, la cual plantea que las propiedades termodinámicas asociadas a los estados locales de incorporación nucleotídica en el ADN, particularmente las diferencias entálpicas efectivas (ΔH), pueden influir en la probabilidad de aparición y fijación de ciertas mutaciones durante la replicación [6–9]. De acuerdo con este planteamiento, la propia estructura del ADN ejercería una influencia direccional sobre el tipo de mutaciones que tienden a acumularse, favoreciendo aquellas que incrementan la estabilidad de la doble hélice. En términos prácticos, ello implicaría que, cuando la presión de selección es mínima o inexistente, la acumulación de mutaciones no sería completamente aleatoria, sino que mostraría una tendencia hacia el incremento del contenido de guanina y citosina (GC) en el genoma [3]. Esta perspectiva introduce un matiz distinto a la interpretación neutralista clásica, al

proponer que los principios termodinámicos podrían orientar la dirección de la evolución molecular junto con, o incluso independientemente de, la selección natural.

El fundamento teórico de esta hipótesis se basa en principios de la termodinámica estadística. Mirko Zimic y colaboradores desarrollaron un modelo conceptual que emplea la distribución de Boltzmann para describir el proceso de incorporación de nucleótidos durante la replicación del ADN [7]. En este modelo, la adición de un nucleótido en el sitio activo de la ADN polimerasa se interpreta como un sistema cercano al equilibrio local, donde los posibles estados corresponden a la incorporación de cualquiera de los cuatro nucleótidos (A, T, C o G) en la cadena en crecimiento. La probabilidad asociada a cada uno de estos estados depende de la energía libre vinculada a la formación del correspondiente par de bases. Debido a que los pares G≡C presentan mayor estabilidad termodinámica —al estar unidos por tres puentes de hidrógeno— en comparación con los pares A=T, que poseen solo dos, las diferencias energéticas entre las posibles interacciones adquieren un papel determinante. Aunque los emparejamientos incorrectos o no canónicos ocurren con muy baja frecuencia, las simulaciones computacionales realizadas por Zimic demostraron que, cuando se producen errores durante la replicación, las condiciones energéticas favorecen la incorporación de nucleótidos G o C frente a A o T [9]. En consecuencia, el proceso mutacional presentaría un sesgo intrínseco hacia el incremento del contenido GC bajo condiciones neutrales. Según este modelo, cuando la presión de selección es reducida o relativamente constante, por ejemplo, en posiciones de baja restricción funcional como la tercera posición de los codones, el sesgo mutacional termodinámico podría manifestarse como un incremento

progresivo del contenido de GC. Este incremento no debe entenderse como indefinido, sino como una aproximación a un estado estacionario o meseta composicional, en el cual el balance entre mutaciones que aumentan el contenido de GC y aquellas que lo reducen tiende a estabilizarse. De hecho, simulaciones *in silico* han mostrado que secuencias sometidas únicamente a esta fuerza termodinámica incrementarán gradualmente su proporción de G+C, aproximándose de forma asintótica a valores elevados, incluso superiores al 90 % GC en ausencia de restricciones funcionales [7]. Cuando se introducen restricciones biológicas, como la necesidad de conservar la funcionalidad proteica y mantener los aminoácidos codificados, el modelo continúa prediciendo un aumento del contenido GC, aunque con valores de equilibrio menores, situados aproximadamente entre 60 % y 75 % [10]. Asimismo, el modelo predice un sesgo en el uso de codones, ya que la sustitución por codones sinónimos con mayor contenido de G y C permitiría incrementar la estabilidad del ADN sin alterar la secuencia de aminoácidos, generando así una preferencia por codones ricos en nucleótidos fuertes. En síntesis, la hipótesis de la presión mutacional termodinámica plantea que la estabilidad fisicoquímica de la molécula de ADN durante el proceso de replicación puede actuar como una fuerza impulsora en la evolución molecular, orientando la acumulación de mutaciones hacia configuraciones energéticamente más estables, caracterizadas por una mayor proporción de pares GC [9].

1.2 Evidencias teóricas y observaciones iniciales

Aunque esta propuesta es relativamente reciente, encuentra apoyo en ciertos patrones observados empíricamente. Análisis filogenéticos basados en comparaciones genómicas han identificado tendencias compatibles con la existencia

de un sesgo mutacional asociado a la estabilidad del ADN. En este contexto, Zimic y colaboradores examinaron secuencias genómicas de organismos pertenecientes al linaje Kinetoplastida, que incluye tripanosomátidos como *Trypanosoma* y *Leishmania*. La elección de este linaje se sustenta en que muchos de estos parásitos se desarrollan en hospedadores que mantienen condiciones fisiológicas relativamente conservadas, como temperatura y pH, además de presentar reproducción predominantemente clonal, ausencia de intrones y suficiente antigüedad evolutiva para acumular mutaciones detectables. Estas características hacen que Kinetoplastida sea un sistema comparativo útil para evaluar patrones composicionales compatibles con el modelo de presión mutacional termodinámica. A partir de la reconstrucción de la historia evolutiva de estos taxones y del análisis de la composición nucleotídica de sus genes, los autores identificaron una tendencia progresiva hacia el incremento del contenido de GC a medida que aumenta la divergencia evolutiva entre las especies [7]. Este aumento resulta particularmente evidente en la tercera posición de los codones, la cual posee un mayor grado de degeneración en el código genético. Tal patrón sugiere que gran parte de estos cambios se produce mediante mutaciones sinónimas que no alteran la secuencia de aminoácidos. Este hallazgo se correlaciona con una creciente preferencia por codones ricos en GC a medida que evolucionan las especies del linaje. Por ejemplo, en tripanosomátidos primitivos predominan codones de leucina o serina con mayor contenido de bases A y T, mientras que en especies más derivadas se aprecia una sustitución progresiva por codones sinónimos enriquecidos en GC [9]. Dichos cambios no necesariamente se explican por una ventaja adaptativa en la traducción, sino que son coherentes con un sesgo mutacional acumulativo a nivel genómico. Por

tanto, la información genómica disponible para estos parásitos constituye una evidencia indirecta que respalda la hipótesis de la presión mutacional termodinámica: cambios observados en la composición de bases (aumento de %GC) y en el uso de codones coinciden con las predicciones del modelo. Bajo esta perspectiva, la acumulación de mutaciones no dependería únicamente de eventos aleatorios modulados posteriormente por la selección natural, sino también de diferencias energéticas locales que podrían favorecer configuraciones moleculares más estables.

1.3 Experimentos previos de amplificación seriada *in vitro*: polimerasa termófila vs. mesófila

Para contrastar directamente estas predicciones en condiciones controladas, diversos investigadores han utilizado modelos experimentales de amplificación seriada *in vitro*. En este tipo de aproximaciones, una secuencia de ADN se somete a múltiples ciclos sucesivos de replicación en ausencia de presiones selectivas externas, de modo que las mutaciones acumuladas reflejen principalmente la fidelidad replicativa y los posibles sesgos intrínsecos de la ADN polimerasa empleada. Uno de los primeros trabajos que aplicó este enfoque fue el desarrollado por Sanson, Brunstein, Briones y colaboradores en 2002, quienes diseñaron un experimento de “filogenia experimental” basado en PCR serial (reacción en cadena de la polimerasa en serie) [11]. En dicho estudio se utilizó como secuencia inicial un fragmento del gen de ARN ribosomal 18S. Aunque esta región cumple una función biológica en condiciones naturales, en el sistema de PCR serial fue amplificada únicamente como molécula de ADN, sin que su función influyera en la propagación de las variantes generadas. Por ello, la secuencia puede considerarse neutral dentro del contexto experimental. Este fragmento fue sometido a 280 ciclos sucesivos de

PCR utilizando la ADN polimerasa Taq, enzima proveniente de *Thermus aquaticus*, un organismo termófilo. Como resultado de este proceso, los autores obtuvieron un conjunto de secuencias derivadas (37 clones secuenciados) que permitieron reconstruir un árbol filogenético experimental, simulando la divergencia de múltiples linajes a partir de un ancestro común bajo condiciones *in vitro* [11]. El análisis posterior de estos datos realizado por Zimic reveló un hallazgo relevante: a lo largo de las generaciones simuladas se acumuló un sesgo mutacional significativo en favor de nucleótidos G y C [10]. En otras palabras, las secuencias descendientes presentaban un contenido de GC superior al de la secuencia ancestral, lo que sugiere que las mutaciones generadas durante los ciclos de PCR no se distribuyeron de manera aleatoria. Sin embargo, este resultado también abrió la posibilidad de una explicación alternativa relacionada con las características del sistema experimental. En particular, se planteó que la ADN polimerasa Taq, al provenir de un organismo adaptado a altas temperaturas, podría favorecer indirectamente una composición nucleotídica más rica en GC, típica de genomas asociados a ambientes termófilos [11]. Bajo esta interpretación, el incremento del contenido de GC observado en el experimento podría reflejar un sesgo específico de la enzima o de las condiciones de PCR a alta temperatura, más que un principio termodinámico general aplicable al ADN. De ser así, la denominada “presión mutacional” hacia GC representaría un efecto derivado de la naturaleza termófila de la polimerasa utilizada, en lugar de una confirmación directa de la hipótesis termodinámica propuesta [7].

Esta posibilidad condujo a una cuestión fundamental: si la tendencia hacia el incremento de GC constituye realmente un principio general del proceso replicativo, ¿debería observarse el mismo patrón cuando la replicación *in vitro* se realiza con una

polimerasa mesófila bajo condiciones de temperatura moderada? Los dominios catalíticos de la Taq polimerasa I presentan homología estructural con los de polimerasas provenientes de organismos mesófilos, como la ADN polimerasa I de *Escherichia coli*, lo que permite diseñar experimentos comparables utilizando enzimas con distinta estabilidad térmica [12]. Si la hipótesis de la presión mutacional termodinámica describe un principio general, cabría esperar que incluso bajo condiciones mesófilas, y en ausencia de selección, se produjera con el tiempo una acumulación preferencial de mutaciones que incrementan el contenido GC, aunque potencialmente a una tasa diferente. Por el contrario, si el aumento de GC observado con Taq fuese simplemente una consecuencia de su origen termófilo, entonces el mismo patrón no debería reproducirse cuando se emplean polimerasas mesófilas en condiciones equivalentes. Para explorar esta posibilidad, Arnáez y colaboradores llevaron a cabo un experimento de amplificación seriada *in vitro* a temperatura moderada (37 °C) [13]. En dicho estudio se utilizó la ADN polimerasa Klenow (un fragmento de la polimerasa I de *E. coli*, de carácter mesófilo) para amplificar de manera serial el gen *pncA* de *Mycobacterium tuberculosis*. El experimento consistió en siete generaciones sucesivas de amplificación, donde cada ronda incluyó 27 ciclos de PCR realizados a 37 °C. Para favorecer la acumulación de errores replicativos se empleó la variante Klenow exo^- , carente de actividad correctora. Tras cada ronda de amplificación, el producto obtenido se clonó en *E. coli*, seleccionando posteriormente un clon al azar como molde para la siguiente generación. Este enfoque experimental, conceptualmente similar al utilizado por Brunstein y colaboradores, pero ejecutado a menor temperatura, tenía como objetivo determinar si emergía algún sesgo composicional en ausencia de condiciones termofílicas. Los

resultados obtenidos por Arnáez fueron preliminares: después de siete generaciones no se detectó un número suficiente de mutaciones que permitiera establecer con claridad una tendencia en la variación del contenido GC. Esta limitación se debió en gran medida al reducido número de generaciones experimentales y a las restricciones logísticas del estudio [13]. A pesar de ello, el trabajo demostró que es posible implementar modelos experimentales de amplificaciones seriadas utilizando polimerasas mesófilas, lo que proporcionó una base metodológica importante para investigaciones posteriores. La ausencia de resultados concluyentes, atribuible al bajo número de mutaciones generadas, puso de manifiesto la necesidad de realizar experimentos más prolongados y con mayor profundidad mutacional para evaluar adecuadamente la hipótesis planteada.

1.4 Planteamiento del estudio y simulaciones *in vitro* mesófilas vs. termófilas

Con el propósito de esclarecer el papel que podría desempeñar la presión mutacional termodinámica, nuestro grupo desarrolló un proyecto de investigación integral (financiado por PROCENCIA) orientado a evaluar experimentalmente esta hipótesis mediante amplificaciones seriadas *in vitro*. En este proyecto se plantearon tres estrategias principales: (1) amplificación seriada en condiciones termófilas mediante PCR utilizando la ADN polimerasa Taq, con el objetivo de reproducir y ampliar los resultados reportados en estudios previos realizados a altas temperaturas; (2) amplificación seriada en condiciones mesófilas mediante PCR empleando la polimerasa Klenow de *Escherichia coli*, lo que permite examinar el proceso replicativo a temperaturas moderadas y extender el enfoque utilizado en el estudio de Arnáez; y (3) amplificación seriada en condiciones mesófilas utilizando la técnica por recombinasa polimerasa (RPA) [14]. Estos tres sistemas experimentales

incorporan ADN polimerasas con diferentes propiedades de termoestabilidad, lo que permite comparar de manera directa cómo influyen tanto la temperatura como las características intrínsecas de cada enzima en la acumulación de mutaciones durante la replicación del ADN. Bajo la hipótesis de la presión mutacional termodinámica, todos estos sistemas (aún operando en condiciones térmicas distintas) deberían eventualmente mostrar una dirección similar en el sesgo mutacional, caracterizada por un incremento progresivo del contenido GC, ya que el principio físico subyacente (la mayor estabilidad energética de los pares $G\equiv C$) sería común a todos los contextos replicativos. No obstante, es posible que la velocidad de acumulación de mutaciones o los patrones específicos de sustitución difieran entre sistemas, debido a factores como la eficiencia de los mecanismos de corrección de errores presentes en cada polimerasa. Por otro lado, si el sesgo hacia GC se observa únicamente en el sistema basado en Taq y no en aquellos que emplean polimerasas mesófilas, ello sugeriría que el incremento de GC detectado previamente podría estar relacionado con propiedades particulares de la enzima termófila o con las condiciones experimentales de PCR a alta temperatura, cuestionando así la aplicabilidad general de la hipótesis.

La presente tesis de maestría forma parte de este proyecto y se centra específicamente en el modelo experimental de amplificación isotérmica basado en RPA. Este enfoque representa un sistema adecuado para estudiar la dinámica mutacional en ausencia tanto de temperaturas extremas como de presiones selectivas externas. La técnica de amplificación por recombinasa polimerasa (Recombinase Polymerase Amplification, RPA) fue seleccionada por diversas razones metodológicas. En este sistema, la replicación del ADN se lleva a cabo a temperatura constante, aproximadamente 37 °C, lo que elimina la necesidad de ciclos térmicos

característicos de la PCR convencional. El proceso emplea la polimerasa derivada de *Bacillus subtilis* (un organismo mesófilo) junto con proteínas accesorias, entre las que se incluyen recombinasas y proteínas de unión a ADN monocatenario, que facilitan la amplificación isotérmica del ADN. Un aspecto particularmente relevante para los objetivos de este estudio es que la polimerasa utilizada en RPA carece de actividad exonucleasa en dirección 5'→3' y 3'→5', lo que implica la ausencia de mecanismos de corrección de errores o *proofreading* [12]. Como consecuencia, los errores de apareamiento generados durante la replicación tienden a mantenerse en las moléculas sintetizadas, favoreciendo la acumulación progresiva de mutaciones. En el diseño del experimento, cada “generación” *in vitro* corresponde a una reacción de RPA de corta duración (20 minutos) realizada a 37 °C, utilizando como molde el producto amplificado en la generación anterior. Durante este intervalo, la reacción de RPA puede generar una amplificación exponencial del fragmento objetivo, alcanzando niveles comparables a los obtenidos en aproximadamente veinte ciclos de PCR convencional [15]. Esto garantiza la obtención de suficiente ADN para iniciar la siguiente ronda de replicación. Gracias a esta eficiencia, es posible realizar múltiples pasajes sucesivos en un período relativamente corto, permitiendo que las mutaciones se acumulen gradualmente a lo largo de las generaciones simuladas. Las poblaciones de secuencias obtenidas en diferentes etapas del experimento (los amplicones evolucionados) fueron posteriormente analizadas mediante secuenciación de nanoporos (Oxford Nanopore), una tecnología que permite leer moléculas completas de ADN e identificar con precisión las mutaciones puntuales ocurridas.

Cabe señalar que la presencia o ausencia de mecanismos de corrección de errores en las ADN polimerasas puede influir tanto en la frecuencia de acumulación de mutaciones como en el tipo de variantes que persisten, ya que la actividad de *proofreading* podría corregir determinados apareamientos incorrectos de manera diferencial según el tipo de *mismatch*, el contexto de secuencia y las propiedades de la enzima. En este sentido, el sistema RPA empleado en el presente estudio, basado en una polimerasa derivada de *Bacillus subtilis* sin actividad exonucleasa correctora 3'→5', permite reducir la influencia de este filtro de corrección y evaluar la acumulación mutacional generada durante amplificaciones seriadas *in vitro* bajo condiciones mesófilas. En consecuencia, si se observa una tendencia hacia el incremento del contenido GC en este sistema, ello sería compatible con la hipótesis de presión mutacional termodinámica; sin embargo, su interpretación debe considerar también los posibles sesgos propios de la polimerasa utilizada y de las condiciones experimentales. Por el contrario, si tal incremento no se detecta bajo condiciones de temperatura moderada, ello apoyaría la hipótesis alternativa de que el sesgo hacia GC observado previamente en experimentos con Taq podría estar asociado a propiedades particulares de las condiciones termofílicas en las que opera esta enzima. En este contexto, el presente estudio evalúa experimentalmente dicha hipótesis mediante un modelo controlado de amplificaciones seriadas en condiciones mesófilas. De esta manera, se busca determinar si las propiedades termodinámicas inherentes a la molécula de ADN pueden orientar la dirección de las mutaciones en ausencia de selección natural. Los resultados obtenidos complementarán la evidencia previa derivada tanto de modelos teóricos como de estudios experimentales.

1.5 Pregunta de Investigación

El problema central que aborda esta tesis se origina en una pregunta aún no resuelta dentro del campo de la evolución molecular: ¿las mutaciones se distribuyen realmente de manera neutral en términos de su dirección composicional, o existe una fuerza fisicoquímica inherente al sistema molecular que, independientemente de la selección natural, pueda influir en el tipo de mutaciones que tienden a acumularse a lo largo del tiempo? La hipótesis de la presión mutacional termodinámica propone que, aun cuando los apareamientos no canónicos o *mismatches* ocurren con baja frecuencia durante la replicación del ADN, las diferencias en estabilidad energética entre los estados finales del sistema pueden favorecer la incorporación y posterior fijación de nucleótidos asociados a configuraciones energéticamente más estables. En particular, dado que las interacciones que involucran guanina y citosina presentan mayor estabilidad termodinámica que aquellas asociadas a adenina y timina, el modelo predice que, en ausencia de presiones selectivas significativas, la acumulación neta de mutaciones tendería a incrementar el contenido de GC en las secuencias de ADN.

Sin embargo, una crítica plausible a esta hipótesis plantea que el sesgo hacia GC observado en algunos sistemas experimentales podría no reflejar un principio físico general, sino más bien características particulares de la enzima replicativa utilizada. Por ejemplo, una polimerasa procedente de organismos termófilos podría haber evolucionado para mantener la estabilidad de genomas sometidos a altas temperaturas, generando así un sesgo composicional específico. Con el fin de abordar esta posible explicación alternativa, el presente estudio se enfoca en un sistema experimental mesófilo que evita condiciones térmicas extremas. Para ello se emplea

un modelo de amplificación isotérmica seriada mediante RPA a 37 °C, utilizando una polimerasa derivada de un organismo mesófilo y manteniendo deliberadamente la ausencia de presiones selectivas externas. Bajo estas condiciones, la trayectoria mutacional observada reflejará principalmente los sesgos intrínsecos asociados al proceso de replicación y a las propiedades fisicoquímicas del sistema molecular.

1.5.1 Pregunta de investigación principal

En un modelo experimental basado en amplificaciones seriadas por RPA, realizado bajo condiciones mesófilas (37 °C) y sin presión selectiva externa, ¿puede detectarse un cambio direccional y estadísticamente significativo hacia el aumento del porcentaje de GC en secuencias derivadas de *Escherichia coli*, tal como predice la hipótesis de la presión mutacional termodinámica?

1.5.2 Preguntas derivadas: operacionales y analíticas

Con el objetivo de hacer empíricamente evaluable la pregunta principal, se plantean las siguientes interrogantes específicas:

Direccionalidad composicional: ¿El porcentaje de GC del amplicón muestra un incremento consistente conforme aumenta el número de generaciones experimentales (variable independiente), bajo condiciones controladas?

Espectro mutacional: ¿El conjunto de sustituciones observadas presenta un sesgo neto hacia mutaciones que incrementan el contenido de GC (por ejemplo, A/T→G/C) en comparación con aquellas que lo reducen (G/C→A/T)?

Tasa mutacional: ¿Cuál es la tasa de mutación observable en cada generación del experimento y cómo se distribuyen dichas mutaciones a lo largo de la región amplificada?

Generalidad entre secuencias: ¿La tendencia composicional detectada (en caso de existir) se manifiesta de manera similar en diferentes genes de *E. coli* amplificados bajo el mismo esquema experimental (por ejemplo, *evgA*, *ygiV* y *phnH*) o depende principalmente de la composición inicial de cada secuencia?

Neutralidad funcional aproximada: En un sistema sin presión selectiva externa, ¿las mutaciones que alcanzan frecuencias relativamente altas se concentran preferentemente en posiciones más tolerantes desde el punto de vista funcional (como las mutaciones sinónimas) lo que sería consistente con una dinámica evolutiva dominada por procesos mutacionales más que adaptativos?

II. OBJETIVOS

2.1 Objetivo general

Evaluar experimentalmente, mediante amplificación seriada *in vitro* por RPA bajo condiciones mesófilas controladas y sin presión selectiva externa, si la acumulación de mutaciones se asocia con un incremento neto y estadísticamente significativo del contenido de GC en el ADN.

2.2 Objetivos específicos

Los objetivos específicos han sido planteados con el fin de cubrir de manera integral las distintas etapas del proceso experimental (desde el diseño del sistema hasta el análisis e interpretación de los patrones mutacionales observados) garantizando la trazabilidad del estudio.

2.2.1 Implementar un modelo experimental *in vitro* bajo condiciones mesófilas utilizando RPA.

Diseñar un sistema de amplificación isotérmica seriada a 37 °C en el que cada reacción de RPA representa una generación del proceso simulado. Durante el experimento se mantendrán constantes las condiciones fisicoquímicas del sistema (incluyendo temperatura, volumen de reacción y composición de los reactivos) y se evitará la introducción de presiones selectivas externas, de manera que las variaciones observadas reflejan principalmente la dinámica mutacional intrínseca al proceso replicativo.

2.2.2 Evaluar la consistencia del cambio en el contenido de GC entre fragmentos génicos con distinta composición inicial.

Comparar la variación del contenido de GC en fragmentos génicos de *Escherichia coli* amplificados bajo el mismo esquema experimental, con el fin de determinar si el cambio observado constituye un patrón común y no un efecto dependiente de una secuencia particular.

2.2.3 Caracterizar la acumulación mutacional y el espectro de sustituciones a lo largo de las generaciones experimentales.

Identificar, mediante secuenciación de nanoporos de alta cobertura, las mutaciones presentes en las secuencias obtenidas, estimar su frecuencia y persistencia entre generaciones, y calcular la tasa mutacional observada. Asimismo, se distinguirán transiciones y transversiones, evaluando si existe un sesgo consistente hacia mutaciones que incrementan el contenido de GC.

2.2.4 Evaluar la direccionalidad mutacional hacia el incremento de GC como predicción central de la presión mutacional termodinámica.

Analizar el balance entre mutaciones que incrementan el contenido de GC y aquellas que lo reducen, con el fin de determinar si el patrón mutacional observado concuerda con la predicción teórica de que, en ausencia de selección, las diferencias de estabilidad energética favorecen la acumulación de nucleótidos G y C.

2.2.5 Cuantificar la trayectoria del porcentaje de GC en función del número de generaciones.

Calcular el contenido de GC en las poblaciones de secuencias obtenidas en cada generación experimental y evaluar su tendencia a lo largo del tiempo mediante

análisis estadísticos apropiados, como regresiones o pruebas de tendencia. En este análisis se considera explícitamente el número de generaciones como variable independiente y el porcentaje de GC como variable dependiente.

III. HIPÓTESIS

3.1 Hipótesis de investigación: fundamento conceptual

Dentro del marco conceptual de la presión mutacional termodinámica, el proceso de incorporación de nucleótidos durante la replicación del ADN puede interpretarse como una transición entre estados caracterizados por diferentes energías. En consecuencia, los posibles estados finales del sistema no presentan la misma probabilidad de ocurrencia. En ausencia de presión selectiva, la dinámica del sistema tendería a favorecer aquellas configuraciones que poseen mayor estabilidad energética. Dado que los pares de bases que incluyen guanina y citosina presentan una mayor estabilidad termodinámica que los pares adenina-timina, el modelo predice que, cuando se acumulan múltiples eventos de replicación y mutación, la composición nucleotídica tenderá progresivamente a enriquecerse en GC. Además, si este sesgo corresponde a una propiedad fisicoquímica general del proceso de replicación del ADN (y no únicamente a una adaptación particular de polimerasas provenientes de organismos termófilos), entonces debería ser posible detectarlo también en condiciones mesófilas que emplean enzimas replicativas de origen mesófilo, como ocurre en el sistema de amplificación por RPA utilizado en este estudio.

3.2 Hipótesis general (H1)

En un modelo de amplificaciones seriadas por RPA realizado bajo condiciones mesófilas (37 °C) y en ausencia de presión selectiva, la acumulación de mutaciones presentará un sesgo direccional hacia la incorporación/fijación de nucleótidos G y C. Como consecuencia, el porcentaje de GC en fragmentos génicos de *Escherichia coli* aumentará de manera estadísticamente significativa a medida que se incrementa el número de generaciones experimentales.

3.3 Hipótesis específicas (derivadas y verificables)

3.3.1 H1a - Dirección del espectro mutacional: La frecuencia acumulada de mutaciones que incrementan el contenido de GC (A/T→G/C) será superior a la de aquellas que lo disminuyen (G/C→A/T), generando un balance neto positivo hacia nucleótidos G y C.

3.3.2 H1b - Consistencia entre secuencias: La tendencia hacia el aumento del porcentaje de GC se observará de forma coherente en diferentes genes amplificados bajo el mismo esquema experimental (por ejemplo, *evgA*, *ygiV* y *phnH*) incluso cuando dichas secuencias difieran en su composición inicial o en su contexto local.

3.3.3 H1c - Compatibilidad con ausencia de selección: En ausencia de presión selectiva externa, las variantes que alcanzan mayores frecuencias tenderán a localizarse en regiones donde, desde una perspectiva termodinámica, existe una mayor probabilidad de que ocurran desajustes de apareamiento (mismatches) durante el proceso de amplificación.

3.4 Hipótesis nula (H0)

En un sistema de evolución *in vitro* mesófilo basado en amplificación por RPA y sin presión de selección, el contenido de GC no presentará una tendencia direccional en función del número de generaciones. En estas condiciones, las mutaciones se acumularán sin mostrar un sesgo composicional neto hacia G o C ($A/T \rightarrow G/C \approx G/C \rightarrow A/T$), y cualquier variación observada en el porcentaje de GC será atribuible a fluctuaciones estocásticas asociadas al muestreo o al propio proceso de replicación.

IV. MARCO TEÓRICO

4.1 Modelo termodinámico basado en la distribución de Boltzmann para estimar la probabilidad de ingreso de nucleótidos libres al sitio de replicación del ADN

La hipótesis de presión mutacional termodinámica plantea que la probabilidad de incorporación de un nucleótido durante la replicación del ADN puede aproximarse mediante la distribución de Boltzmann. Para ello, se considera como sistema de análisis una vecindad local del ADN alrededor del sitio de replicación, en la cual los nucleótidos libres disponibles en el medio compiten por ocupar una posición específica de la cadena naciente.

En este planteamiento, el sitio de incorporación se representa como una posición X . Dicha posición puede quedar ocupada por adenina, timina, citosina o guanina. Por tanto, los estados posibles del sistema local pueden representarse como $X = A$, $X = T$, $X = C$ o $X = G$ (Figura 1).

Cada una de estas posibilidades corresponde a una configuración distinta de la vecindad molecular del ADN. Es decir, el sistema no tiene la misma energía si el sitio X es ocupado por adenina que si es ocupado por guanina, citosina o timina. Por ello, a cada estado se le asigna una energía configuracional específica:

E_A = energía de la vecindad cuando $X = A$

E_T = energía de la vecindad cuando $X = T$

E_G = energía de la vecindad cuando $X = G$

E_C = energía de la vecindad cuando $X = C$

Estas energías representan la estabilidad energética de la vecindad del ADN después de que un nucleótido específico ha ingresado al sitio de replicación.

De acuerdo con la distribución de Boltzmann, la probabilidad de que un sistema se encuentre en un estado energético E_i se expresa como:

$$P_{(E_i)} = \frac{1}{Z} e^{-G/kT}, \text{ donde } G = H - TS$$

$$P_{(E_i)} = \frac{1}{Z} e^{S/k} e^{-H/kT}$$

El número de configuraciones posibles se relaciona directamente con la entropía del sistema, según la relación estadística $S = k \ln \Omega_{(E_i)}$, por lo tanto:

$$P_{(E_i)} = \frac{1}{Z} \Omega_{(E_i)} e^{-H/kT}$$

Esta ecuación indica que la probabilidad de un estado depende de dos componentes. El primero es el factor exponencial, $e^{-H/kT}$, que favorece a los estados de menor energía. El segundo es $\Omega_{(E_i)}$, que representa cuántas configuraciones posibles puede adoptar el sistema para ese estado energético.

Como primera aproximación, el número de configuraciones accesibles para cada nucleótido es proporcional a la cantidad de nucleótidos libres disponibles en el volumen local del sistema. Considerando un volumen V constante para los cuatro nucleótidos, el número de configuraciones accesibles puede expresarse como:

$$\Omega(E_A) = \alpha [dATP] V$$

$$\Omega(E_G) = \alpha [dGTP] V$$

$$\Omega(E_C) = \alpha [dCTP] V$$

$$\Omega(E_T) = \alpha [dTTP] V$$

donde $[dATP]$, $[dGTP]$, $[dCTP]$ y $[dTTP]$ representan las concentraciones de los desoxinucleótidos trifosfato libres, V indica el volumen del sistema local y α es una

constante de proporcionalidad relacionada con el acomodo molecular del nucleótido en el sitio X. Al reemplazar estas expresiones en la distribución de Boltzmann, se obtiene la probabilidad de ocupación del sitio X por cada nucleótido:

$$P(X = A) = (1 / Z) \cdot \alpha [dATP] V \cdot e^{-H/kT}$$

$$P(X = G) = (1 / Z) \cdot \alpha [dGTP] V \cdot e^{-H/kT}$$

$$P(X = C) = (1 / Z) \cdot \alpha [dCTP] V \cdot e^{-H/kT}$$

$$P(X = T) = (1 / Z) \cdot \alpha [dTTP] V \cdot e^{-H/kT}$$

Para comparar directamente la probabilidad de ocupación del sitio X entre distintos nucleótidos, se pueden definir probabilidades relativas tomando como referencia uno de ellos. Esta normalización no modifica la relación entre los estados, sino que permite expresar cuántas veces más o menos probable es la ocupación del sitio por un nucleótido respecto al nucleótido elegido como referencia. Supongamos que, para una configuración determinada de la vecindad del ADN, se toma como referencia la ocupación del sitio X por guanina. En ese caso, las probabilidades relativas pueden definirse como:

$$P'(X = G) = \frac{P(X=G)}{P(X=G)} = 1$$

$$P'(X = A) = \frac{P(X=A)}{P(X=G)} = \left(\frac{A}{G}\right) e^{-(\Delta H)/kT}$$

$$P'(X = T) = \frac{P(X=T)}{P(X=G)} = \left(\frac{T}{G}\right) e^{-(\Delta H)/kT}$$

$$P'(X = C) = \frac{P(X=C)}{P(X=G)} = \left(\frac{C}{G}\right) e^{-(\Delta H)/kT}$$

Estas expresiones muestran que la preferencia relativa por un nucleótido determinado depende de dos factores principales: la disponibilidad relativa del nucleótido libre en el medio y la diferencia energética entre los estados finales comparados. Al considerar concentraciones equivalentes para los cuatro nucleótidos, la comparación entre estados queda determinada principalmente por el término energético.

Bajo condiciones de presión y volumen constantes, la variación del término PV es nula [$\Delta(PV)=0$]. Por ello, la expresión de la variación de la entalpía $\Delta H=\Delta U+\Delta(PV)$ se reduce a $\Delta H=\Delta U$. La energía interna U asociada a cada estado puede descomponerse en una contribución cinética y una contribución potencial dependiente del nucleótido incorporado $U_N = E_{\text{cinética}} + E_{\text{potencial}} (X=N)$. Por lo tanto,

$$\Delta U = \Delta E_{\text{cinética}} + \Delta E_{\text{potencial}}, \text{ a } T=\text{cte}$$

$$\Delta U = \Delta E_{\text{potencial}}$$

La energía potencial está determinada por las interacciones locales entre el nucleótido incorporado y sus vecinos inmediatos en la cadena de ADN, así como por las interacciones entre los nucleótidos vecinos entre sí. Por ello, al comparar dos estados que difieren únicamente en la identidad del nucleótido que ocupa el sitio X, gran parte de las contribuciones comunes del sistema se cancelan. En consecuencia, la diferencia energética relevante queda dominada por las interacciones específicas del nucleótido incorporado con su entorno inmediato (principalmente las energías por puente de hidrógeno y electrostáticas). Así, para dos nucleótidos N y M, la diferencia de energía entre ambos estados puede expresarse como:

$$E_N - E_M \approx \sum E_{\text{int}}(N-y_i) - \sum E_{\text{int}}(M-y_i)$$

donde y_i representa los nucleótidos que forman parte de la vecindad inmediata del sitio de incorporación, y E_{int} corresponde a la energía de interacción entre la base incorporada y el entorno molecular adyacente. Por lo tanto, las diferencias energéticas relevantes no dependen del sistema completo, sino de las interacciones locales que cambian al variar la identidad de la base que ocupa el sitio (X).

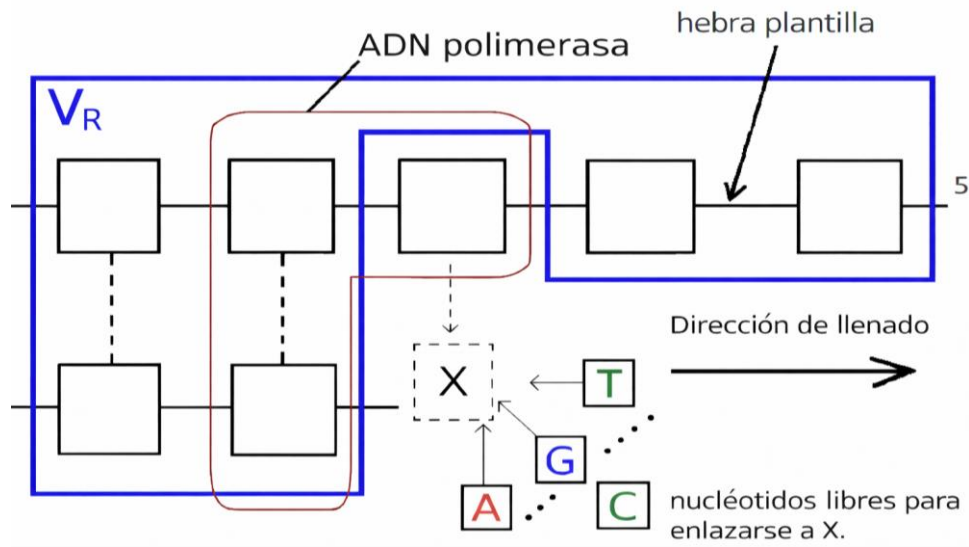


Figura 1. Modelo termodinámico de la replicación del ADN. V_R indica la vecindad de referencia. La diferencia energética entre los estados del sistema se aproxima adecuadamente al considerar una vecindad de 5 nucleótidos al sitio X. Para mayores detalles del modelo revisar el Anexo 1.

No obstante, es importante precisar el alcance de esta aproximación: el modelo considera principalmente términos entálpicos, como las energías de interacción molecular, mientras que los componentes entrópicos (por ejemplo, el efecto hidrofóbico) no son incorporados explícitamente. Por ello, aunque el número de configuraciones accesibles se aproxima mediante su proporcionalidad con la concentración de nucleótidos trifosfato libres en un volumen determinado, esta simplificación representa una primera aproximación del modelo. Una formulación más rigurosa requeriría estimar con mayor precisión los estados accesibles para cada configuración o emplear una descripción basada en la energía libre de Gibbs, en la que se integren explícitamente las contribuciones entálpicas y entrópicas del sistema.

4.2 Amplificación isotérmica del ADN por recombinasa polimerasa

La amplificación isotérmica por recombinasa polimerasa (RPA) constituye una herramienta adecuada para modelar un proceso de replicación del ADN bajo condiciones controladas. Esta técnica genera copias de ADN a temperatura constante utilizando tres proteínas principales (recombinasa UvsX, proteínas de unión a ADN monocatenario SSB y el fragmento grande de la ADN Polimerasa I de *Bacillus subtilis*, Bsu). El proceso de RPA comienza cuando la proteína recombinasa uvsX se une a los cebadores en presencia de ATP y un agente de aglomeración (polietilenglicol de alto peso molecular), formando un complejo recombinasa-cebador. Este complejo encuentra una secuencia homóloga en el ADN de doble cadena y genera una separación local entre la cadena desplazada que se estabiliza mediante SSB y la cadena objetivo que se hibrida con el cebador. Finalmente, la recombinasa se desensambla del complejo y la ADN polimerasa Bsu se une al

extremo 3'-OH del cebador para la elongación de la cadena y la formación de una nueva cadena complementaria [14,15]. La repetición cíclica de este proceso permite obtener una amplificación exponencial del gen objetivo (Figura 2).

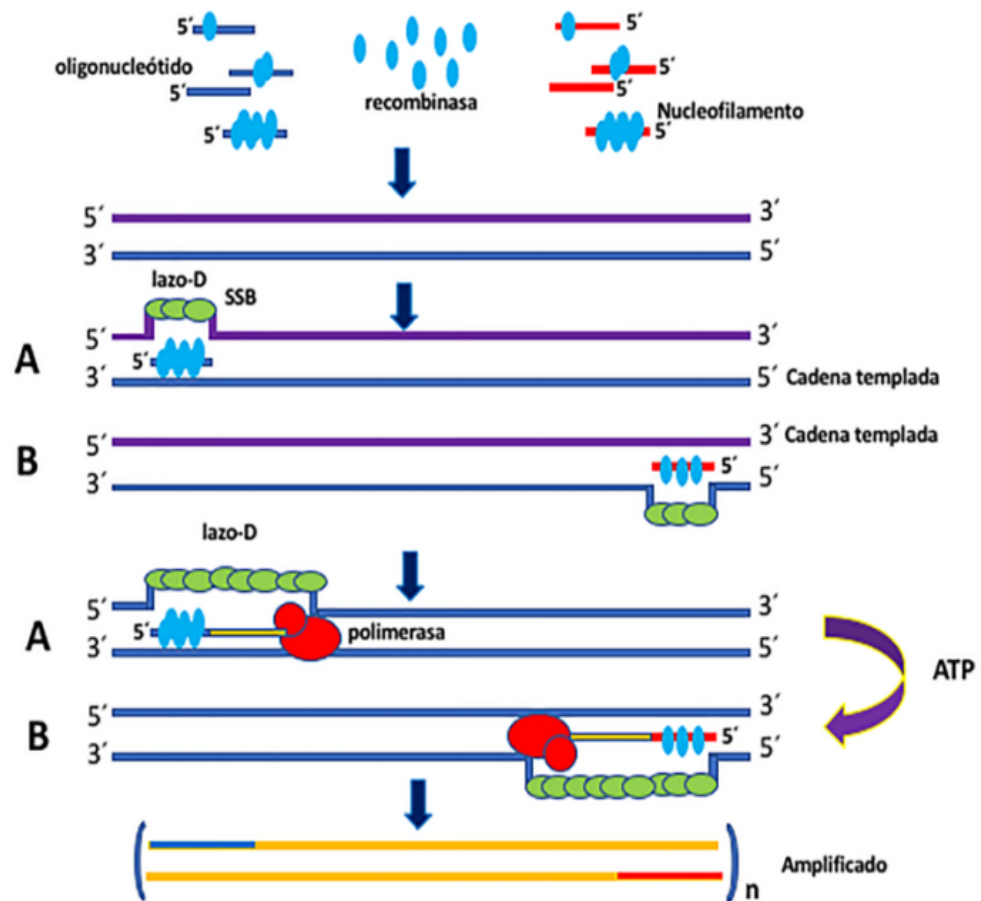


Figura 2. Proceso de amplificación de ADN mediante RPA [16].

Este sistema de replicación permite la síntesis eficiente de múltiples copias de ADN en un intervalo corto de tiempo y sin requerir ciclos térmicos. Cuando la amplificación se realiza de manera iterativa, utilizando como molde el producto de la amplificación previa, se establece un esquema de replications sucesivas que puede simular un proceso de evolución molecular. En este modelo, cada generación

corresponde a una ronda de amplificación. Durante las rondas sucesivas, los errores inherentes al proceso de replicación y la incorporación diferencial de nucleótidos generan variación genética en el conjunto de copias de ADN. Dado que el sistema carece de presiones selectivas, las mutaciones que persisten a lo largo de las generaciones reflejan principalmente las propiedades fisicoquímicas del proceso de replicación, así como los sesgos intrínsecos asociados a la incorporación de nucleótidos.

La caracterización de las poblaciones generadas en cada generación se realiza mediante secuenciación, lo que permite identificar y cuantificar las mutaciones acumuladas a lo largo del proceso. Al reducir la complejidad del sistema biológico y mantener constantes las condiciones externas, el modelo permite explorar si emergen patrones direccionales en la acumulación de mutaciones que puedan atribuirse a principios termodinámicos fundamentales.

V. METODOLOGÍA

5.1 Diseño de la investigación

El presente estudio corresponde a una investigación de tipo experimental básica. El modelo consiste en amplificaciones seriadas de ADN utilizando la técnica de RPA bajo condiciones mesófilas y en ausencia de presión selectiva externa. A través de este sistema se busca explorar los principios termodinámicos que podrían influir en la generación y acumulación de mutaciones durante la replicación del ADN, así como su posible impacto en los patrones mutacionales. El diseño experimental se definió a partir de los siguientes componentes:

- **Variable independiente:** número de generaciones experimentales
- **Variable dependiente:** porcentaje de contenido GC en las secuencias analizadas.
- **Material genético:** se amplificaron tres genes de *Escherichia coli*: el factor de transcripción *evgA*, la proteína de membrana *ygjV* y la subunidad de la liasa carbono-fósforo *phnH*. Estas secuencias fueron obtenidas a partir de células NovaBlue (*E. coli* K-12).
- **Estrategia de simulación:** para cada gen se realizaron rondas sucesivas de amplificación isotérmica, cada una con una duración de 20 minutos a 37 °C.
- **Condiciones experimentales:** el experimento se desarrolló sin introducir presiones selectivas externas. Durante todas las reacciones se mantuvieron constantes las condiciones fisicoquímicas del sistema, incluyendo temperatura, volumen de reacción y concentraciones de nucleótidos y enzimas.

5.2 Selección de secuencias de partida y caracterización composicional

Para evaluar la direccionalidad mutacional hacia un aumento del contenido de GC, se seleccionaron tres regiones codificantes de *Escherichia coli* con longitudes similares (entre 552 y 615 pb), pero con diferentes proporciones iniciales de GC. Esta estrategia permitió establecer un gradiente composicional inicial (%GC bajo, intermedio y alto), adecuado para comparar la magnitud y la tasa de acumulación neta de mutaciones que incrementan GC (AT→GC) frente a aquellas que lo reducen (GC→AT). Las secuencias seleccionadas correspondieron a los genes *evgA*, *ygjV* y *phnH*, cuyos contenidos iniciales de GC fueron 40.0%, 49.6% y 60.9%, respectivamente. Debido a que el experimento se desarrolló bajo condiciones *in vitro*, sin influencia de la selección natural, estos genes no fueron seleccionados con el propósito de evaluar su función biológica ni su historia evolutiva natural.

En todos los casos se trabajó exclusivamente con regiones codificantes dentro del rango de amplificación definido, con el fin de evitar que las comparaciones entre genes estuvieran influenciadas por la presencia de segmentos no codificantes. Adicionalmente, se verificó que las secuencias seleccionadas no presentaran una composición extremadamente sesgada de nucleótidos, sino que mostraran una distribución relativamente homogénea del contenido de GC a lo largo del amplicón. Este criterio fue considerado importante para reducir la posibilidad de que la señal composicional estuviera dominada por segmentos locales atípicos. Para cada gen se registraron en la Tabla 1 la secuencia inicial (5'→3'), la longitud del fragmento en pares de bases y su porcentaje inicial de GC. Estos parámetros sirvieron como referencia para el diseño de los cebadores (*primers*) utilizados en las amplificaciones posteriores.

5.3 Diseño de *primers*

El diseño de los *primers* se realizó siguiendo las recomendaciones descritas en la guía de usuario de los kits TwistAmp® DNA Amplification (TwistDx Inc., Cambridge, Reino Unido). En el caso de la amplificación por recombinasa polimerasa (RPA), los cebadores suelen ser más largos que los utilizados en PCR convencional. Por esta razón, se consideraron *primers* con longitudes aproximadas de 30 a 35 nucleótidos y con un contenido de GC entre 30 % y 70 %. Asimismo, se evitó la presencia de tramos extensos de guaninas o citosinas consecutivas, ya que estos pueden favorecer la formación de estructuras secundarias o promover hibridaciones inespecíficas durante la amplificación.

A diferencia de la PCR tradicional, la técnica RPA no requiere un control estricto de la temperatura de fusión (T_m) de los *primers*. Esto se debe a que los procesos de hibridación y elongación están mediados por complejos proteicos que incluyen recombinasas y polimerasas, los cuales permiten la amplificación sin necesidad de ciclos térmicos. En consecuencia, el diseño de los *primers* se enfocó principalmente en optimizar la estabilidad, especificidad y eficiencia de amplificación bajo condiciones isotérmicas cercanas a 37 °C. Otro aspecto considerado durante el diseño experimental fue la longitud adecuada de los amplicones para la reacción de RPA. Ensayos preliminares, así como reportes asociados al sistema TwistAmp®, indican que fragmentos mayores a aproximadamente 500 pb suelen amplificarse con menor eficiencia y reproducibilidad. Por esta razón, los fragmentos seleccionados se diseñaron dentro de un rango aproximado de 405–464 pb, manteniendo la condición de ser fragmentos codificantes representativos. Además, se verificó que los segmentos elegidos

conservaran aproximadamente el porcentaje de GC presente en las secuencias originales, evitando que el recorte del fragmento modifique artificialmente la composición nucleotídica. Este criterio permitió que las comparaciones posteriores de la trayectoria composicional y del cambio en el contenido de GC entre los distintos genes reflejaran características propias de las secuencias analizadas y no efectos derivados de la selección del amplicón (Tabla 1).

Tabla 1. Características composicionales de las secuencias génicas y amplicones

Gen	Longitud del gen/amplicon	%GC del gen/amplicon	Secuencia (5'→3')*
<i>evgA</i>	615/405 pb	40/42.7%	ATGAACGCAATAATTATTGATGACCATCCTCTTGCTATCG CAGCAATTCGTAATTTATTGATCAAAAACGATATTGAAA TCTTAGCAGAGTTGACTGAAGGCGGAAGTGCCGTTGAGC GGGTGGAAACACTTAAGCCTGATATCGTCATCATTGATG TCGATATCCCCGGAGTTAACGGTATCCAGGTGTTAGAAA CGCTGAGGAAGCGCCAATATAGCGGAATTATTATTATCG TCTCCGCTAAAAATGACCATTTTACGGGAAACATTGTG CTGATGCTGGCGCTAATGGTTTCGTGAGTAAAAAAGAAG GCATGAACAATATCATTGCGGCTATTGAAGCTGCAAAAA ATGGCTACTGCTATTTCCCCTTCTCTCAACCGGTTTGT TGGAAGTTTAACGTCCGACCAGCAAAAACCTCGACTCCTT ATCGAAACAAGAAATTAGTGTCATGCGGTATATTCTTGA TGGCAAGGATAATAATGACATTGCTGAAAAAATGTTTCAT CAGCAACAAAACCTGTCAGCACTTATAAAAGTCGCCTGAT GGAAAAATTAGAATGTAAATCACTGATGGATCTTTACAC ATTCGCACAACGTAACAAAATCGGCTAA
<i>ygiV</i>	552/464 pb	49.6/48.7%	ATGACCGCGTATTGGCTGGCCCAGGGCGTGCGGTGTCATC GCCTTTCTGATTGGTATCACAACATTTTCAATCGTGACG AACGTCGCTTCAAAAAGCAGCTTTCGGTCTATAGCGCCG TTATTGGCGTACATTTTTTCTTCTGGGCACTATCCCGC TGCTGCCAGTGCCATCCTTAATGCCATTCTGATACATTGATT ACCTTACGCACGCGCAGCTTATGGGTAATGGCGATTTTT ATTGTGCTGACTGGCGGAATTGGCCTCGCGAAGTTCCAT CATCCTGTGCAACTATTGCCGGTTATCGGCACGATTGTCA GTACCTGGGCGCTATTCTGCTGTAAAGGGCTGACCATGC GCTGCGTAATGTGGTTTTCAACGTGTTGCTGGGTGATTCA CAACTTCTGGGCGGGGTCGATAGGCGGCACGATGATTGA GGGTAGTTTTCTGCTTATGAATGGCCTGAATATCATTTCGT TTCTGGCGGATGCAGAAAAGGGGAATTGATCCGTTTAAA GTTGAGAAAACCCCTTCCGCCGTAGACGAAAGGGGTAA
<i>phnH</i>	585/443 pb	60.9/59.8%	ATGACCCTGGAAACCGCTTTTATGCTTCCCGTGCAAGGAT GCCCAGCACAGTTTTTCGTCGCCTGTAAAGGCCATGAGC GAGCCGGGCGTGATTGTGCGCCCTGCATCAGCTCAAACGC GGCTGGCAACCGCTGAATATCGCCACCACCAGCGTGCTG CTGACGCTGGCCGATAACGACACGCCGGTGTGGCTTTCT ACCCCATTAATAACGATATCGTCAACCAGAGCCTGCGT TTTCATACCAACGCGCCGCTGGTCAAGCAGCCGGAACAG GCGACCTTCGCGGTGACGGATGAGGCGATTTCAGCGAA CAGCTCAACGCCCTTTCCACCGGCACCGCGTTGCGCCG GAAGCGGGTGCGACGCTGATTTTACAGGTGCGCCAGCCTG AGCGGCGGACGCATGTTGCGCCTTACTGGTGCGGGTATT GCCGAAGAACAATGATCGCTCCGCAGCTGCCGGAGTG CATTCTGCACGAACCTACCGAGCGCCCGCATCCGTTCCC GCTCGGCATCGACCTGATCCTGACCTGTGGCGAGCGCCT GCTGGCTATTCCGCGAACCCTCATGTGGAGGTGTGCTG A

* La región sombreada corresponde al amplicón

La predicción inicial de los *primers* se llevó a cabo utilizando la herramienta bioinformática Primer-BLAST, priorizando candidatos con alta especificidad para *Escherichia coli* y con baja probabilidad de amplificación inespecífica fuera de la región objetivo. Posteriormente, las secuencias candidatas fueron analizadas mediante OligoAnalyzer, con el propósito de descartar *primers* con tendencia elevada a formar dímeros, estructuras tipo horquilla (*hairpins*) u otras configuraciones secundarias que pudieran afectar la eficiencia de la reacción. Finalmente, los *primers* seleccionados fueron sintetizados por la empresa Macrogen, y sus características se presentan en las Tablas 2.

Tabla 2. Características de los primers diseñados para los experimentos RPA

Gen	Primers	%GC
<i>evgA</i>	F: 5'-CAAAAACGATATTGAAATCTTAGCAGAGTTGAC-3' R: 5'-AGAATATACCGCATGACACTAATTTCTTGTTTC-3'	F: 33.33 R: 33.33
<i>ygiV</i>	F: 5'-ATTGGTATCACAAACATTTTTCAATCGTGAC-3' R: 5'-TTAAACGGATCAATTCCCCTTTTCTGCATC-3'	F: 33.33 R: 40.00
<i>phnH</i>	F: 5'-ATGACCCTGGAAACCGCTTTTATGCTTCCC-3' R: 5'-ATTCGTTCTTCGGCAATACCCGCACCAGTAA-3'	F: 50.00 R: 48.39

5.4 Extracción de ADN

El ADN genómico fue aislado a partir de células competentes NovaBlue (*Escherichia coli* K-12) utilizando el kit ZymoBIOMICS™ DNA Miniprep (Zymo Research Inc., California, EE.UU.), siguiendo las recomendaciones del fabricante.

De manera resumida, la muestra celular se transfirió al tubo de lisis ZR BashingBead™, al cual se añadieron 750 µL de solución de lisis. Posteriormente, la mezcla se centrifugó a $10\,000 \times g$ durante 1 minuto. El sobrenadante obtenido se trasladó al filtro Zymo-Spin™ III-F, contenido en un tubo colector, y se centrifugó nuevamente a $8\,000 \times g$ por 1 minuto. El filtrado resultante se mezcló con 1200 µL del buffer de unión de ADN y se transfirió a una columna Zymo-Spin™ IICR colocada en un tubo de recolección. Después de centrifugar a $10\,000 \times g$ durante 1 minuto, el filtrado fue descartado. A continuación, se añadieron 400 µL del buffer de lavado 1, seguido de centrifugación y eliminación del líquido filtrado. Posteriormente, se aplicaron 700 µL del buffer de lavado 2, repitiendo el mismo procedimiento de centrifugación y descarte del filtrado. Se realizó un lavado adicional agregando 200 µL del buffer de lavado 2, seguido de centrifugación bajo las mismas condiciones ($10\,000 \times g$ durante 1 minuto). Para la elución del ADN, se añadieron 100 µL de agua libre de nucleasas directamente sobre la columna. Tras una incubación de 10 minutos, la muestra fue centrifugada nuevamente a $10\,000 \times g$ durante 1 minuto. Con el fin de mejorar la pureza del ADN obtenido, el eluido fue transferido a un filtro Zymo-Spin™ III-HRC previamente preparado y centrifugado a $16\,000 \times g$ durante 3 minutos. El ADN recuperado mediante este procedimiento se consideró adecuado para los ensayos posteriores.

5.5 Amplificación con recombinasa polimerasa (RPA)

Las amplificaciones se realizaron utilizando el protocolo correspondiente al kit TwistAmp™ Basic. Inicialmente, la mezcla de reacción para RPA se preparó en un tubo de 1.5 mL, siguiendo las proporciones indicadas en la Tabla 3. El volumen total de reacción (42.5 µL) se añadió posteriormente al tubo que contenía el pellet

lioofilizado del kit. La mezcla fue homogeneizada mediante pipeteo repetido y posteriormente centrifugada por breves segundos. Luego, se agregaron 1 μL de ADN templado y 2.5 μL de acetato de magnesio (MgOAc , 280 mM). Previa estandarización de la temperatura de incubación, se seleccionó 37°C como condición para los ensayos posteriores. La reacción final, con un volumen total de 50 μL , se incubó a esta temperatura durante 20 minutos.

Tabla 3. Mezcla inicial de RPA

Mix RPA	Cantidades
Primer forward (10 μM)	2.4 μL
Primer reverse (10 μM)	2.4 μL
Buffer de rehidratación	29.5 μL
H ₂ O	12.2 μL
Total	42.5 μL

La verificación de la amplificación se realizó mediante electroforesis en gel de agarosa al 2 %, aplicando un voltaje de 80 V durante 50 minutos.

5.6 Purificación de los productos de amplificación

Los productos obtenidos tras la reacción de RPA se purificaron utilizando el kit Zymoclean Gel DNA Recovery Kit. Luego de la electroforesis, el fragmento de ADN correspondiente se extrajo del gel de agarosa con un bisturí estéril y se colocó en un tubo de microcentrífuga de 1.5 mL. El fragmento de gel fue pesado y se añadió el buffer de disolución de agarosa en un volumen equivalente a tres veces el volumen del gel. Posteriormente, la mezcla se incubó a 55 °C durante 10 minutos para permitir la completa disolución de la matriz de agarosa. La solución resultante se transfirió a una columna Zymo-Spin™ ubicada en un tubo colector y se centrifugó a 10 000 $\times$ g

durante 1 minuto. El filtrado fue descartado y se añadieron 200 μL del buffer de lavado, seguido de centrifugación bajo las mismas condiciones. Finalmente, se añadieron 20 μL del buffer de elución, la columna se colocó en un tubo limpio de 1.5 mL y se centrifugó nuevamente a $10\,000 \times g$ durante 1 minuto para recuperar el ADN purificado. La concentración del ADN eluido fue determinada utilizando un NanoDrop (Thermo Fisher Scientific Inc., Massachusetts, EE.UU.). Tras cada etapa de purificación se obtuvieron alícuotas destinadas tanto al secuenciamiento como a las rondas posteriores de amplificación por RPA.

Como alternativa de purificación, también se utilizó un sistema basado en perlas magnéticas (Agencourt AMPure XP), que permite la selección de fragmentos de ADN de aproximadamente 200 pb o mayores. Para este procedimiento se añadieron 0.3 μL de perlas por cada 1 μL de muestra, mezclando la solución mediante pipeteo. La mezcla se incubó durante 5 minutos en un agitador Hula a temperatura ambiente para optimizar la recuperación del ADN. Posteriormente, el tubo se colocó en una placa magnética durante 5 minutos para separar las perlas de la solución. El sobrenadante fue retirado y se realizaron dos lavados con etanol al 70 %, cada uno con incubación de 30 segundos. Luego de eliminar el etanol, las perlas se dejaron secar durante 5 minutos a temperatura ambiente. Finalmente, el ADN se eluyó añadiendo 20 μL de buffer de elución, seguido de incubación a 37 °C durante 5 minutos. Tras una nueva separación magnética, el sobrenadante que contenía el ADN purificado se transfirió a un tubo limpio para continuar con las amplificaciones posteriores.

5.7 Análisis mutacional

El análisis de las mutaciones acumuladas en los productos de amplificación se realizó mediante secuenciación con tecnología Nanopore, utilizando el dispositivo PromethION 2 Solo (Oxford Nanopore Technologies Ltd., Cambridge, Reino Unido), en combinación con química V14 y celdas de flujo R10.4.1. Este sistema está optimizado para alcanzar una precisión modal de lectura cruda Q20+, equivalente a una exactitud superior al 99% por lectura individual. Esta precisión fue considerada en el análisis de variantes, debido a que los errores residuales de secuenciamiento pueden influir en la detección de mutaciones de baja frecuencia.

La tecnología Nanopore se basa en el paso de moléculas individuales de ADN a través de nanoporos incrustados en una membrana polimérica conductora. Al aplicarse un voltaje constante en una solución electrolítica, las moléculas de ADN, cargadas negativamente, atraviesan el nanoporo desde el compartimento cis hacia el trans, generando cambios característicos en la corriente iónica. Estas variaciones son interpretadas mediante algoritmos computacionales de *basecalling*, lo que permite determinar la secuencia de nucleótidos en tiempo real [16]. Este enfoque proporciona la información necesaria para identificar variantes nucleotídicas y analizar los cambios en la composición de GC a lo largo de las generaciones experimentales.

5.7.1 Preparación de librerías

Una vez purificados los amplicones, se procedió a la preparación de las librerías de secuenciación para adaptar las moléculas de ADN al sistema Nanopore. Para ello se empleó el Native Barcoding Kit 96 V14 SQK-NBD114.96 (Oxford Nanopore Technologies Ltd., Cambridge, Reino Unido), el cual permite la incorporación de

códigos de barras únicos para la secuenciación simultánea de múltiples muestras [17].

El procedimiento tuvo una duración aproximada de 140 minutos. Las muestras se ajustaron inicialmente a una concentración de 5 ng/μL en un volumen de 14 μL. Debido a que los fragmentos analizados correspondían a amplicones específicos, no fue necesario realizar un proceso previo de fragmentación del ADN. El protocolo incluyó la preparación de extremos y la adición de adeninas (dA-tailing) mediante el reactivo NEBNext® Ultra™ II End Repair/dA-Tailing, seguida de la ligación de adaptadores con códigos de barras únicos que contienen colas dT. Posteriormente, las muestras barcodificadas fueron agrupadas y se incorporó el adaptador de secuenciación correspondiente. Finalmente, la librería preparada se almacenó en tubos Eppendorf a 4 °C hasta su carga en la plataforma de secuenciación.

5.7.2 Secuenciación

Previo al inicio de la corrida de secuenciación, las celdas de flujo fueron cebadas utilizando una mezcla compuesta por 1170 μL de Flow Cell Flush (FCF) y 30 μL de Flow Cell Tether (FCT). Posteriormente, la librería preparada se cargó en la celda de flujo utilizando una mezcla de 100 μL de Sequencing Buffer (SB), 68 μL de Library Beads (LIB) y 32 μL de librería de ADN. Una vez preparada la celda de flujo, la secuenciación fue iniciada mediante el software MinKNOW, que controla el funcionamiento del dispositivo, la adquisición de datos y el proceso de llamada de bases (*basecalling*) en tiempo real. Para lograr el objetivo de identificar mutaciones y distinguir claramente entre diferentes fragmentos de ADN se empleó una profundidad de secuenciación aproximada de 100,000× [18].

5.7.3 Análisis bioinformático

El proceso de *basecalling* de las lecturas generadas durante el secuenciamiento se realizó utilizando el software Dorado v1.2.0 (github.com/nanoporetech/dorado). Posteriormente, las lecturas obtenidas fueron sometidas a un proceso de demultiplexación, así como a la eliminación de barcodes y adaptadores, empleando el módulo demux del mismo entorno de análisis. Con el objetivo de evaluar el rendimiento global del experimento de secuenciación y definir criterios apropiados para el filtrado de las lecturas, se examinó la calidad de los reads demultiplexados mediante NanoPlot v1.46.2 (<https://github.com/wdecoster/NanoPlot>) [19]. Este análisis permitió caracterizar parámetros como la longitud de lectura, la distribución de calidades y otras métricas relevantes para el control de calidad de los datos. Posteriormente, los archivos FASTQ correspondientes a cada generación experimental fueron alineados contra la secuencia de referencia de *Escherichia coli* K-12 [20] utilizando el programa Minimap2 v2.30 (<https://github.com/lh3/minimap2>) [21]. Para los análisis posteriores se conservaron únicamente alineamientos primarios de alta confianza, correspondientes a lecturas con $\text{MAPQ} \geq 20$ (probabilidad de error de alineamiento $\leq 1\%$) y cuyo alineamiento cubría entre 250 y 500 pb con respecto a la secuencia de referencia.

Para caracterizar la composición nucleotídica por posición y detectar sustituciones puntuales, los archivos de cada generación se procesaron mediante un flujo basado en samtools mpileup. Se obtuvieron conteos de nucleótidos (calidad de base ≥ 20) de los alineamientos forward y reverse para calcular la profundidad total (DP) y las frecuencias de A, C, G y T por orientación. A partir de estos datos se identificó el alelo alternativo dominante (ALT) en cada posición como la base

distinta a la referencia con mayor número de observaciones. Para reducir ruido y errores sistemáticos, se aplicaron filtros de profundidad mínima ($DP \geq 50,000$), frecuencia alélica mínima ($AF \geq 0.0005$) y balance por hebra, exigiendo que el alelo alternativo presente soporte en ambas orientaciones (como mínimo el 20% de los conteos debe provenir de la hebra minoritaria). Con el fin de identificar mutaciones con dinámica compatible con señal biológica a lo largo de generaciones seriadas, se realizó un análisis temporal sobre el conjunto de variantes candidatas. La selección final se basó en filtros simultáneos de persistencia (sin ausencias desde su aparición), crecimiento neto ($\Delta AF \geq 0.0003$) y estabilidad temporal (fracción de caídas sucesivas ≤ 0.2). Las variantes que cumplieron estos criterios se emplearon para evaluar la direccionalidad mutacional hacia GC.

Para sustentar el umbral mínimo de frecuencia alélica empleado en la selección de variantes candidatas, se estimó el ruido basal empírico por tipo de sustitución a partir de los alelos alternativos no candidatos presentes en los datos procesados. Por ejemplo, para estimar el ruido basal de $A \rightarrow G$, se incluyeron las posiciones con A como base de referencia y se contabilizaron las lecturas que presentaban G como base alternativa, excluyendo previamente las variantes candidatas identificadas por posición y sustitución exacta. La probabilidad basal empírica (ruido) se calculó como $p_0 = \Sigma ALT_i / \Sigma DP_i$, donde ΣALT_i corresponde a la suma de lecturas con la base alternativa no candidata y ΣDP_i a la suma de las profundidades de las posiciones evaluadas para ese mismo cambio. Así, si para $A \rightarrow G$ se observaran 50 lecturas alternativas no candidatas sobre una profundidad acumulada de 165 000 lecturas, el ruido basal sería $50/165\,000 = 0.000303$. Este valor corresponde principalmente a errores propios del secuenciamiento Nanopore, originados durante la generación de

la señal eléctrica o el *basecalling*. La comparación de las variantes candidatas frente a este fondo permitió sustentar el umbral $AF \geq 0.0005$, aplicado junto con criterios de profundidad mínima, soporte por ambas hebras, persistencia temporal y crecimiento neto de la frecuencia alélica.

Para evaluar el balance neto entre mutaciones cada sustitución se categorizó según su efecto en GC asignando $\Delta GC = +1$ para transiciones $AT \rightarrow GC$, $\Delta GC = -1$ para $GC \rightarrow AT$ y $\Delta GC = 0$ para cambios neutrales. Para cada generación se calcularon (i) la suma de frecuencias alélicas de mutaciones que favorecen GC ($\Sigma AF_{\{AT \rightarrow GC\}}$), (ii) la suma de frecuencias de mutaciones que favorecen AT ($\Sigma AF_{\{GC \rightarrow AT\}}$) y (iii) un índice de sesgo neto definido como $\Sigma(\Delta GC \cdot AF)$. Este índice permite determinar si el balance de mutaciones acumuladas se orienta hacia el aumento de GC, en concordancia con la predicción teórica bajo ausencia de selección.

El contenido GC por generación se estimó a partir de la secuencia de referencia y de las frecuencias alélicas de mutaciones persistentes detectadas en cada población. Primero, se calculó el contenido inicial de GC (GC_0) del amplicón, definido como la fracción de bases G y C respecto a la longitud total (L) de la secuencia. Posteriormente, el balance neto de mutaciones ponderado, expresado como $Ng = \Sigma(AF \cdot \Delta GC)$, se incorporó a la composición inicial mediante $GC_{pred} = GC_0 + Ng/L$. El valor resultante se expresó como porcentaje (%GC) para cada generación y se utilizó en los análisis posteriores de tendencia temporal. La tendencia de %GC se evaluó mediante correlación de Spearman y regresión ponderada por mínimos cuadrados (WLS), considerando la generación como variable independiente y el %GC como variable dependiente, con ponderación basada en la profundidad total.

5.8 Análisis de suficiencia del número de generaciones

Con el fin de evaluar si el número de generaciones experimentales fue suficiente para permitir la detección de mutaciones, se incorporó un análisis de suficiencia basado en la tasa efectiva de error de la polimerasa *Bsu*. Dado que cada generación experimental corresponde a una reacción completa de RPA, el número acumulado de oportunidades efectivas de copia se estimó como:

$$C_{\text{cum}} = C_{\text{efectivo}} \times G$$

donde C_{efectivo} representa el número aproximado de eventos efectivos de copia por reacción de RPA y G el número de generaciones evaluadas. Estudios previos han demostrado que la RPA puede producir amplificación exponencial a temperatura constante y alcanzar límites de detección comparables a la PCR en tiempo real en aproximadamente 20 min [22]. Por ello, se empleó $C_{\text{efectivo}} = 20$ como aproximación operativa.

La frecuencia alélica esperada por sitio se estimó mediante:

$$AF_{\text{esperada}} \approx \mu \times C_{\text{cum}}$$

donde μ corresponde a la tasa efectiva de error por base incorporada. Se empleó $\mu = 1 \times 10^{-5}$ como un valor aproximado para representar la tasa de error de la polimerasa *Bsu* en el modelo, basado en el orden de magnitud reportado para polimerasas de la familia A y considerando la ausencia de actividad correctora 3'→5' [23]. Con este enfoque se evaluó si las variantes esperadas podrían superar el umbral mínimo de frecuencia alélica utilizado en el análisis bioinformático. Por ejemplo, para *evgA*, considerando siete generaciones y $C_{\text{efectivo}} = 20$, se obtiene $C_{\text{cum}} = 20 \times 7 = 140$. Por

lo tanto, la frecuencia esperada por sitio sería: $AF_{\text{esperada}} \approx 10^{-5} \times 140 \approx 0.0014$. Este valor supera el umbral mínimo aplicado de ($AF \geq 0.0005$). Con una profundidad mínima de 50,000 lecturas, dicho umbral equivale a 25 lecturas alternativas, mientras que el valor esperado corresponde aproximadamente a 70 lecturas alternativas. Por tanto, este análisis permite sustentar que el número de generaciones evaluadas fue suficiente para detectar variantes persistentes, siempre que estas superaran además los criterios de profundidad, balance por hebra, persistencia, crecimiento neto y estabilidad temporal.

La metodología se organizó en tres etapas principales: diseño experimental, amplificación seriada por RPA y secuenciación/análisis bioinformático, como se resume en la Figura 3.

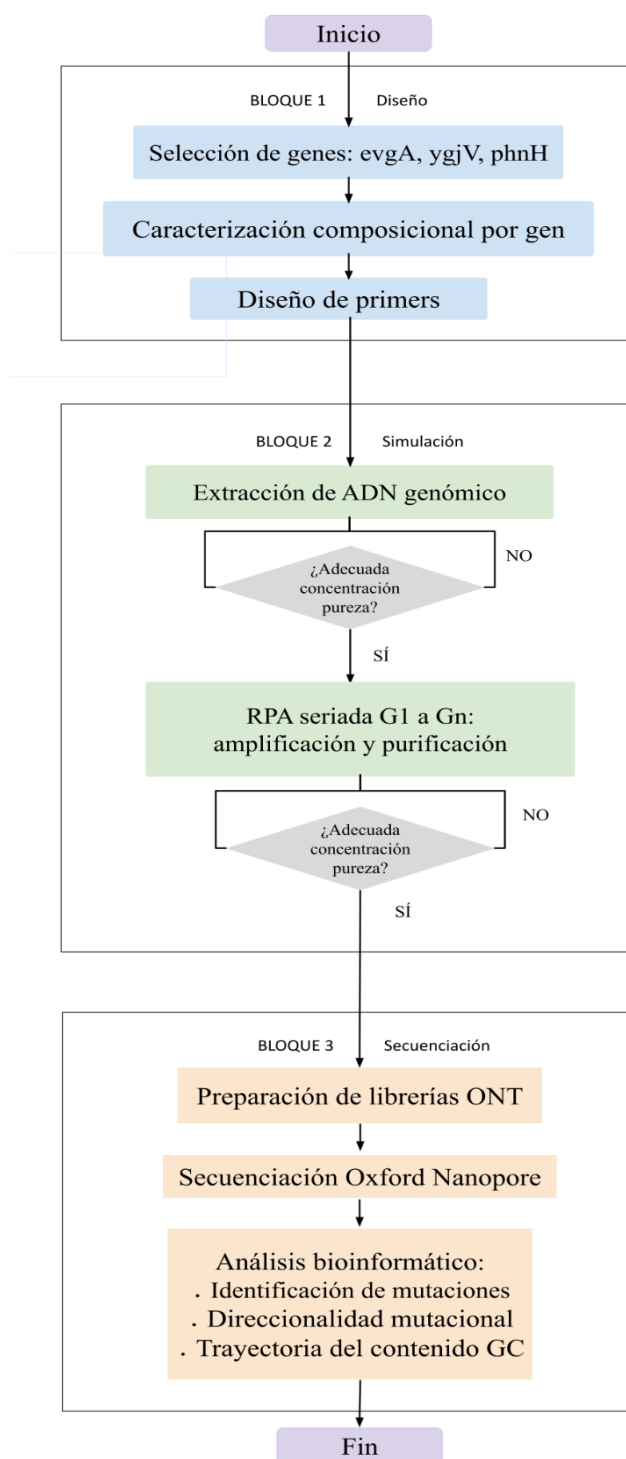


Figura 3. Flujograma general de la metodología experimental y análisis de mutaciones.

VI. RESULTADOS O ARGUMENTACIÓN TEÓRICA

6.1 Implementación del modelo experimental basado en amplificaciones seriadas de fragmentos génicos de *E. coli*

El ADN genómico extraído de *Escherichia coli* K-12 Novablué presentó adecuada integridad y pureza para su empleo en procesos de amplificación isotérmica. Se obtuvo una concentración de 282.7 ng/μL, con relaciones de absorbancia $A_{260}/A_{280} = 1.88$, compatible con ausencia de contaminación proteica, y $A_{260}/A_{230} = 1.99$, indicativa de mínima interferencia por compuestos orgánicos.

Antes de iniciar las amplificaciones seriadas, se optimizó la reacción de RPA evaluando distintas temperaturas dentro del rango de 25 °C a 42 °C. Esta etapa permitió identificar la temperatura más adecuada para obtener un producto específico del fragmento objetivo, evidenciado por una banda definida y con baja presencia de productos secundarios o *smearing*. A 30 °C no se detectó una banda visible en la región esperada, mientras que a 42 °C se observó una mayor tendencia a la aparición de productos inespecíficos. En contraste, a 37 °C se obtuvo una banda definida, compatible con el tamaño esperado del amplicón. Por esta razón, dicha temperatura fue seleccionada como condición estándar para las amplificaciones seriadas posteriores (Figura 4).

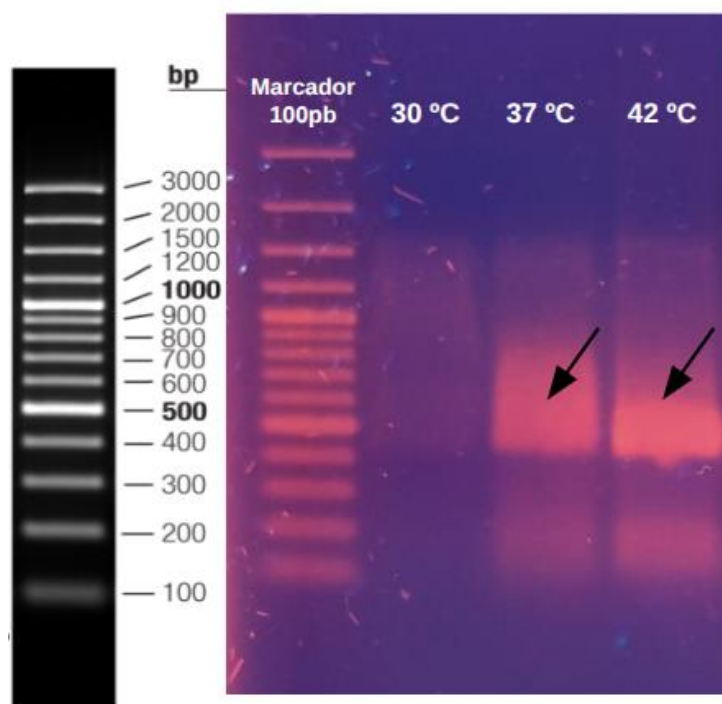


Figura 4. Electroforesis en gel de agarosa de productos amplificados (gen *pnhH*) mediante RPA bajo diferentes temperaturas de reacción dentro del rango de 25 °C a 42 °C. La selección de la temperatura óptima se realizó considerando la presencia de una banda compatible con el tamaño esperado del amplicón, la intensidad de la señal y la menor formación de productos inespecíficos o smearing. La condición de 37 °C mostró el mejor balance entre eficiencia y especificidad, por lo que fue empleada como temperatura estándar en las amplificaciones seriadas posteriores.

A partir del material genético extraído y bajo la condición optimizada de 37 °C, se realizaron amplificaciones sucesivas del gen *evgA* mediante RPA, obteniéndose un total de siete generaciones experimentales (G1–G7). La integridad de estos productos se evaluó mediante electroforesis en gel de agarosa al 2%. En cada generación se identificó una banda única y bien definida de aproximadamente 405 pares de bases que corresponde al tamaño teórico del amplicón (Figura 5). Sin embargo, a partir de la octava ronda (G8) no fue posible continuar la serie

generacional, debido a que los productos amplificados no presentaron una banda específica, sino un *smearing* difuso.

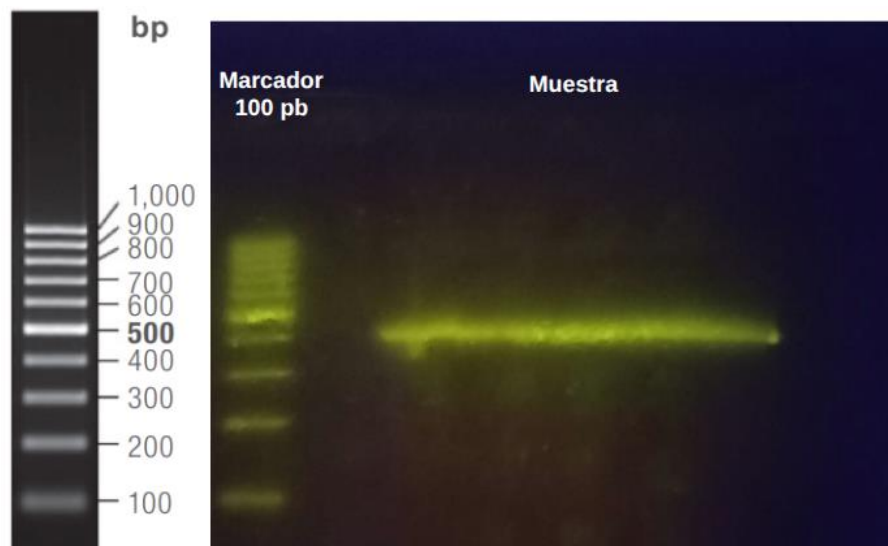


Figura 5. Electroforesis en gel de agarosa al 2 % de la generación 3 del gen *evgA*. A la izquierda se muestra el marcador de peso molecular *GeneRuler 100 bp DNA Ladder* (Thermo Fisher Scientific), y a la derecha la muestra amplificada, donde se observa una banda principal de aproximadamente 405 pb, concordante con el tamaño teórico del amplicón. El ADN presentó una concentración de 35 ng/ μ l, con relaciones de absorbancia $A_{260}/_{280} = 1.95$ y $A_{260}/_{230} = 0.60$.

Con el objetivo de mejorar la pureza de los amplicones y reducir la presencia de productos inespecíficos, en los genes *ygjV* y *phnH* se incorporó un paso de purificación mediante beads magnéticos. Este procedimiento permitió remover *primers* libres, fragmentos cortos amplificados, nucleótidos residuales y proteínas, además de facilitar una selección controlada de fragmentos por tamaño. En el caso del gen *ygjV*, las rondas sucesivas de amplificación mediante RPA permitieron obtener un total de diez generaciones, en las cuales se observaron bandas únicas y nítidas correspondientes al amplicón esperado de aproximadamente 464 pb. No obstante, a partir de la generación 11 se evidenció la aparición predominante de

bandas inespecíficas y zonas de *smearing* difuso (Figura 6). A pesar de este cambio en el patrón electroforético, la serie se prolongó hasta la generación 30, con el propósito de disponer de un mayor número de generaciones para el análisis de acumulación de mutaciones.

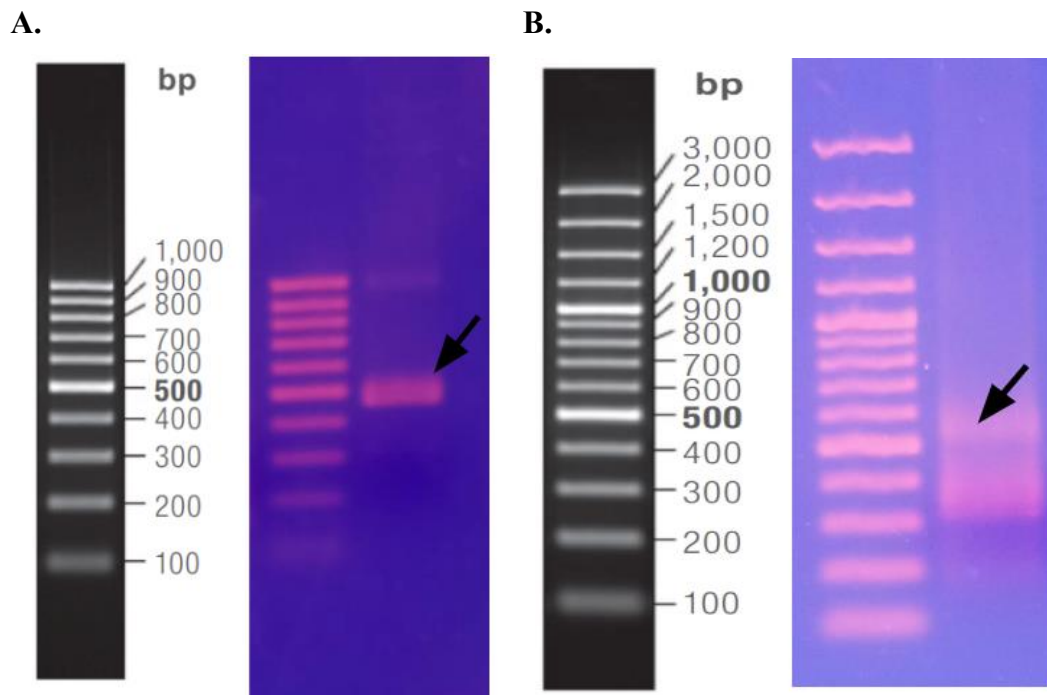


Figura 6. Electroforesis en gel de agarosa al 2 % de los productos de amplificación del gen *ygjV*. A) En la generación 7 se observa una banda específica que corresponde al amplicon de 464 pb (flecha). La muestra presentó una concentración de 90 ng/ μ l y relaciones de absorbancia $A_{260}/_{280} = 2.0$ y $A_{260}/_{230} = 1.8$, compatibles con una adecuada calidad del ADN. B) En la generación 11 se observa la dispersión del material amplificado y la presencia de un patrón de *smearing* (flecha).

En el gen *phnH*, se observó un patrón similar, ya que a partir de la generación 7 se registró la aparición de *smearing* difuso en los geles de electroforesis (Figura 7). No obstante, el proceso continuó hasta la generación 25, con la finalidad de contar con un número suficiente de generaciones para la evaluación de la composición nucleotídica. De manera general, en los tres genes analizados (*evgA*, *ygjV* y *phnH*) se

observó una pérdida progresiva de bandas específicas después de varias generaciones sucesivas.

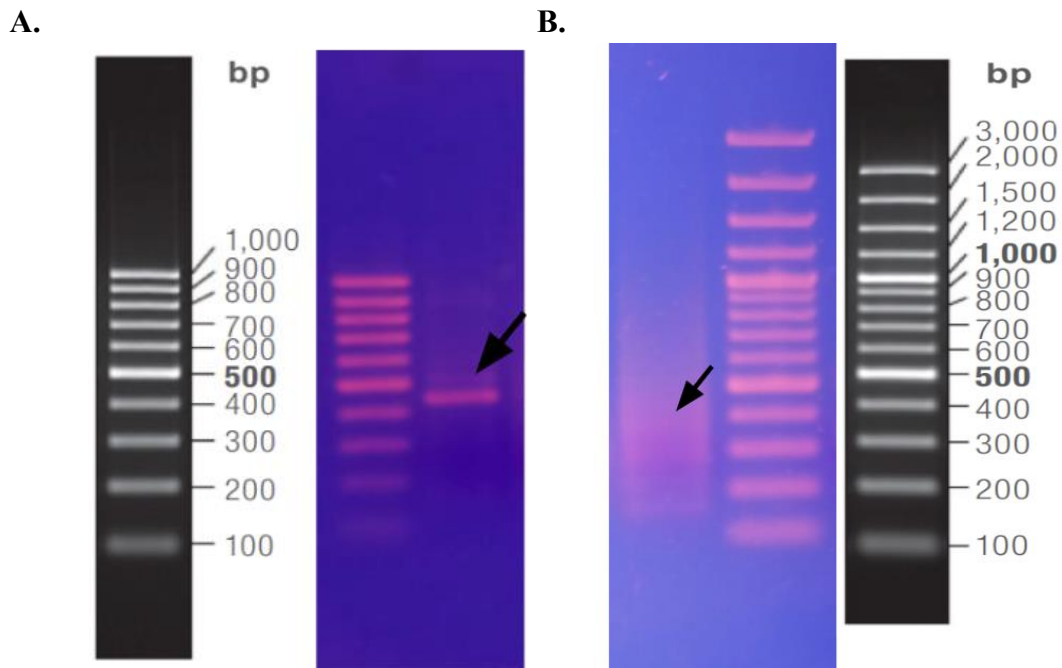


Figura 7. Electroforesis en gel de agarosa al 2 % del gen *phnH*. A) Generación 4, en la que se observa una banda específica de aproximadamente 443 pb (flecha), que concuerda con el tamaño teórico del amplicón. La muestra presentó una concentración de 34 ng/ μ l, con relaciones de absorbancia $A_{260}/_{280} = 2.3$ y $A_{260}/_{230} = 2.0$, indicativas de alta pureza. B) Generación 7, donde se evidencia aparición de smearing (flecha) y pérdida de banda específica.

6.2 Secuenciación por Oxford Nanopore: rendimiento, cobertura y calidad de lecturas

6.2.1. Gen *evgA*

El análisis de calidad de las lecturas obtenidas mediante secuenciación Nanopore para las generaciones G1 a G7 se realizó a partir de los reportes generados con NanoPlot (Tabla 4). En las generaciones iniciales (G1–G5), las distribuciones de longitud mostraron valores similares, con medianas entre 403 - 404 pb y N50 de 405

pb, lo cual indica que la mayoría de las lecturas corresponden a amplicones completos y de tamaño cercano al esperado. En contraste, en las generaciones G6 y G7 se registró una disminución de la longitud mediana y del N50, que evidencia una mayor dispersión en los tamaños de lectura. Este cambio concuerda con el patrón observado en los geles de electroforesis, donde se aprecia una pérdida progresiva de especificidad. La calidad media (Q) se mantuvo dentro de un rango estrecho (19.6–20.8) a lo largo de todas las generaciones, con porcentajes de bases $Q \geq 20$ superiores al 69%. Por tanto, pese al incremento de la variabilidad en longitud en las generaciones tardías, las lecturas conservaron una calidad adecuada para el análisis de mutaciones puntuales y composición nucleotídica.

El alineamiento de las lecturas de cada generación contra la secuencia de referencia indica que en las generaciones iniciales (G1–G5) predominaron alineamientos primarios de alta confianza (Tabla 5), lo que concuerda con la homogeneidad observada en las distribuciones de longitud. No obstante, a partir de G6 se evidenció una disminución en la eficiencia de mapeo, coincidente con el incremento en la heterogeneidad de los productos amplificados.

Tabla 4. Métricas de calidad de las lecturas del gen *evgA*

Generación	Número de lecturas	Bases totales (Mb)	Longitud media (pb)	Longitud mediana (pb)	N50 (pb)	Calidad media (Q)	Q $\geq$ 20 (%)
1	5219440	1810.87	346.9	403	405	20.8	77.3
2	7108974	2546.63	358.2	404	405	20.6	75.2
3	5449466	1926.28	353.5	404	405	20.8	77.3
4	5643635	2023.19	358.5	404	405	20.6	75.7
5	4506265	1586.05	352.0	403	405	20.8	77.2
6	4803616	1430.1	297.7	310	365	20.0	72.4
7	3051136	1230.39	403.3	399	442	19.6	69.1

Tabla 5. Métricas de alineamiento del gen *evgA*

Generación	Lecturas totales	Lecturas mapeadas	Primarios	Secundarios	Suplementarios	No mapeados	Lecturas mapeadas (%)
1	5442463	5020120	4813503	76	206541	422343	92.24
2	7277257	7146604	6992323	88	154193	130653	98.20
3	5562135	5295385	5192326	365	102694	266750	95.20
4	5752378	5487061	5393654	325	93082	265317	95.39
5	4599912	3898961	3821027	317	77617	700951	84.76
6	4893249	1919038	1863603	229	55206	2974211	39.22
7	3089506	385324	374620	133	10571	2704182	12.47

6.2.2. Gen *ygjV*

Las métricas de calidad de lectura (Tabla 6) evidencian que en las generaciones iniciales (G1–G7) las distribuciones de longitud son altamente homogéneas, con medianas cercanas al tamaño teórico del amplicón (463–465 pb), valores estables de N50 y calidades medias próximas a Q20, con más del 70% de las bases presentando Q $\geq$ 20. Sin embargo, a partir de la generación G8, tanto la longitud

mediana como el N50 disminuyeron de forma gradual, conforme aumentó la dispersión en los tamaños de lectura. De manera similar, la calidad media descendió hacia valores cercanos a Q18, y el porcentaje de bases con $Q \geq 20$ se estabilizó en torno al 50–56%, reflejando un deterioro progresivo de la calidad global de las lecturas.

Este comportamiento también se vio reflejado en las métricas de alineamiento (Tabla 7). En G1–G7, más del 95% de las lecturas se alinearon de forma primaria contra la secuencia de referencia. A partir de G8, la fracción de lecturas mapeadas disminuyó de manera sostenida. En las generaciones finales (G13–G30), la proporción de lecturas alineadas no superó el 10%, predominando las lecturas no mapeadas. Estos resultados indican que la población de moléculas amplificadas se diversifica estructuralmente conforme aumenta el número de rondas de replicación.

Tabla 6. Métricas de calidad de las lecturas del gen *ygiV*

Generación	Número de lecturas	Bases totales (Mb)	Longitud media (pb)	Longitud mediana (pb)	N50 (pb)	Calidad media (Q)	Q$\geq$20 (%)
1	2372013	1024.49	431.9	463	464	19.8	73.8
2	1574452	684.63	434.8	463	464	19.6	72.9
3	1241905	529.89	426.7	463	464	19.8	73.1
4	2201090	921.07	418.5	462	463	19.5	70.9
5	1145318	491.85	429.4	463	464	19.5	72.2
6	1500575	658.15	438.6	464	465	19.6	71.8
7	1128844	498.88	441.9	465	466	19.4	70.9
8	946319	394.66	417.0	465	487	19.3	68.1
9	1173973	443.52	377.8	337	464	18.9	63.5
10	1026005	379.48	369.9	337	389	18.6	59.9
11	1729506	630.72	364.7	339	366	18.2	55.6
12	1138364	405.84	356.5	341	365	18.2	54.1
13	1827490	642.65	351.7	350	365	18.1	51.9
14	875587	299.24	341.8	338	362	18.2	52.3
15	1446539	473.62	327.4	336	352	18.1	51.7
16	1147387	374.35	326.3	336	353	18.1	51.3
17	932406	288.62	309.5	331	348	18.1	52.9
18	1333483	414.87	311.1	331	344	18.1	52.0
19	1514500	484.63	320.0	333	348	18.0	52.0
20	1599631	484.63	303.6	318	337	18.2	52.9
21	1696706	493.70	291.0	293	334	18.2	53.8
22	1406405	385.14	273.8	266	304	18.2	54.5
23	1909861	523.40	274.0	266	309	18.4	55.8
24	1481745	406.20	274.1	264	311	18.3	54.8
25	1241070	343.29	276.6	274	313	18.3	55.7
26	1623215	483.56	297.9	293	327	18.5	56.1
27	2321525	633.89	273.0	266	303	18.4	56.6
28	1363730	400.91	294.0	278	325	18.5	56.9
29	1273572	395.33	310.4	304	348	18.5	56.6
30	1238596	75.48	316.4	303	353	18.5	56.2

Tabla 7. Métricas de alineamiento del gen *ygjV*

Generación	Lecturas totales	Lecturas mapeadas	Primarios	Secundarios	Suplementarios	No mapeados	Lecturas mapeadas (%)
1	2387872	2379243	2366127	10	13106	8629	99.64
2	1582744	1575198	1569614	4	5580	7546	99.52
3	1255554	1235593	1225094	8	10491	19961	98.41
4	2232669	2029100	2000844	33	28223	203569	90.88
5	1162201	1129620	1116341	20	13259	32581	97.20
6	1521970	1461723	1444604	21	17098	60247	96.04
7	1154147	1101287	1080371	87	20829	52860	95.42
8	966520	755439	739611	104	15724	211081	78.16
9	1190315	617522	605176	74	12272	572793	51.88
10	1036775	394360	387144	53	7163	642415	38.04
11	1744344	396654	387460	104	9090	1347690	22.74
12	1144852	130622	127506	56	3060	1014230	11.41
13	1833927	92471	90053	83	2335	1741456	5.04
14	878229	22429	21901	27	501	855800	2.55
15	1449512	25561	25252	12	297	1423951	1.76
16	1150177	21862	21683	4	175	1128315	1.90
17	934760	12172	12107	1	64	922588	1.30
18	1335894	10397	10353	2	42	1325497	0.78
19	1517815	7514	7464	2	48	1510301	0.50
20	1602357	6989	6953	0	36	1595368	0.44
21	1700675	58696	57582	2	1112	1641979	3.45
22	1408851	10823	10713	1	109	1398028	0.77
23	1913547	6964	6888	4	72	1906583	0.36
24	1484603	9443	9121	0	322	1475160	0.64
25	1243278	3121	3103	0	18	1240157	0.25
26	1626684	1622	1604	0	18	1625062	0.10
27	2325618	893	872	1	20	2324725	0.04
28	1367452	561	546	0	15	1366891	0.04
29	1276981	1099	1089	0	10	1275882	0.09
30	1210897	680	660	0	20	1210217	0.06

6.2.3. Gen *phnH*

Las métricas de calidad (Tabla 8) indican que en las generaciones iniciales (G1–G5) las distribuciones de longitud son homogéneas, con valores de mediana cercanos al tamaño teórico del amplicón (443 pb) y N50 comprendidos entre 443 y 444 pb. Asimismo, las calidades medias se mantuvieron próximas a Q20, con porcentajes de bases $Q \geq 20$ superiores al 62%, lo que es compatible con una amplificación eficiente durante las primeras rondas de replicación *in vitro*. A partir de la generación G6, se observó una disminución progresiva tanto de la longitud mediana como del N50, acompañada de un aumento en la dispersión de tamaños de lectura. De forma similar, la calidad media descendió gradualmente hasta valores cercanos a Q18, y el porcentaje de bases $Q \geq 20$ se situó alrededor del 50%.

El alineamiento de las lecturas contra la secuencia de referencia (Tabla 9) evidenció una alta proporción de alineamientos primarios en las primeras generaciones (G1–G3), seguida de una disminución progresiva en la eficiencia de mapeo en generaciones posteriores. La reducción simultánea de la calidad de lectura y de la proporción de alineamientos primarios es consistente con el incremento de productos de amplificación heterogéneos en rondas sucesivas de RPA.

Tabla 8. Métricas de calidad de las lecturas del gen *phnH*

Generación	Número de lecturas	Bases totales (Mb)	Longitud media (pb)	Longitud mediana (pb)	N50 (pb)	Calidad media (Q)	Q≥20 (%)
1	1520694	652.31	429.0	443	443	20.1	76.7
2	1604463	683.25	425.8	443	443	20.0	76.8
3	1078082	410.86	381.1	442	443	19.8	73.8
4	805086	314.86	391.1	441	444	19.3	68.7
5	972633	369.97	380.4	425	444	18.8	62.2
6	847021	294.14	347.3	324	435	18.3	55.5
7	850880	287.68	338.1	311	400	18.2	53.0
8	928395	329.27	354.7	328	410	18.1	51.2
9	816092	281.36	344.8	315	402	18.0	50.3
10	874131	274.74	314.3	282	378	18.0	50.6
11	1233871	400.27	324.4	290	376	17.9	48.7
12	687062	224.76	327.1	290	379	18.0	51.4
13	489966	194.23	396.4	366	455	17.9	48.3
14	736838	246.40	334.4	299	394	17.9	49.2
15	302558	99.05	327.4	292	382	17.7	47.9
16	990606	334.71	337.9	309	396	17.8	48.7
17	819306	268.50	327.7	290	381	18.0	51.3
18	646043	221.70	343.2	323	412	17.9	49.9
19	950081	274.00	288.4	261	335	18.3	54.3
20	955439	334.43	350.0	327	408	18.1	52.0
21	814,871	292.78	359.3	341	424	18.2	53.9
22	970788	340.92	351.2	330	411	18.4	54.7
23	1160778	408.69	352.1	335	418	18.5	56.2
24	988366	343.06	347.1	332	416	18.4	55.5
25	1093330	332.01	303.7	278	353	18.6	57.6

Tabla 9. Métricas de alineamiento del gen *phnH*

Generación	Lecturas totales	Lecturas mapeadas	Primarios	Secundarios	Suplementarios	No mapeados	Lecturas mapeadas (%)
1	1547149	1540640	1516394	22	24224	6509	99.58
2	1628327	1602892	1582414	42	20436	25435	98.44
3	1099874	912108	893107	36	18965	187766	82.93
4	846087	607284	577041	34	30209	238803	71.78
5	1015815	485340	460470	42	24828	530475	47.78
6	875925	183484	172294	22	11168	692441	20.95
7	872144	70243	66150	13	4080	801901	8.05
8	952650	25152	23547	7	1598	927498	2.64
9	838005	4220	3897	8	315	833785	0.50
10	893038	543	512	1	30	892495	0.06
11	1260797	287	263	1	23	1260510	0.02
12	702255	142	125	1	16	702113	0.02
13	510085	112	109	0	3	509973	0.02
14	755467	165	150	1	14	755302	0.02
15	309684	87	84	0	3	309597	0.03
16	1012401	436	412	0	24	1011965	0.04
17	836914	191	185	1	5	836723	0.02
18	660287	256	203	0	53	660031	0.04
19	965517	178	164	0	14	965339	0.02
20	979363	189	180	1	8	979174	0.02
21	834723	185	165	1	19	834538	0.02
22	993819	130	122	0	8	993689	0.01
23	1187505	145	140	0	5	1187360	0.01
24	1007867	109	106	0	3	1007758	0.01
25	1110121	124	100	0	24	1109997	0.01

6.3 Identificación de mutaciones acumuladas: frecuencia alélica, persistencia y dinámica temporal

Los valores de ruido basal empírico variaron según el tipo de sustitución, evidenciando que el error técnico no fue homogéneo entre cambios nucleotídicos (Tabla 10). Las probabilidades basales más elevadas correspondieron a A→G (3.91×10^{-4}) y T→C (3.84×10^{-4}), mientras que los valores más bajos se observaron para A→C (2.90×10^{-5}) y T→G (4.00×10^{-5}). En todos los casos, los valores estimados de ruido basal p_0 por tipo de sustitución fueron inferiores al umbral mínimo de frecuencia alélica utilizado para la selección de variantes candidatas ($AF \geq 5.0 \times 10^{-4}$).

Tabla 10. Ruido basal empírico por tipo de sustitución

Sustitución	p_0 ruido basal
A→C	2.90×10^{-5}
A→G	3.91×10^{-4}
A→T	6.46×10^{-5}
C→A	1.34×10^{-4}
C→G	1.10×10^{-4}
C→T	2.98×10^{-4}
G→A	3.05×10^{-4}
G→C	8.32×10^{-5}
G→T	1.42×10^{-4}
T→A	2.88×10^{-4}
T→C	3.84×10^{-4}
T→G	4.00×10^{-5}

6.3.1. Gen *evgA*

Se identificaron 26 mutaciones clasificadas como Señales Biológicas, las cuales mostraron frecuencias alélicas bajas y persistentes a lo largo de generaciones consecutivas (Tabla 11). La distribución temporal mostró que una fracción importante de mutaciones emergió tempranamente: 5/26 variantes fueron detectadas desde la G1, manteniéndose sin ausencias hasta la G7. De forma similar, 8/26 mutaciones aparecieron a partir de la G2, 8/26 desde la G3, mientras que un número menor surgió en etapas tardías (2/26 desde la G4 y 3/26 desde la G5).

En cuanto al tipo de sustitución, el conjunto de variantes persistentes se caracterizó por un predominio marcado de mutaciones que incrementan el contenido GC. Del total de señales identificadas, 23/26 (88.5%) corresponden a cambios clasificados como AT→GC, principalmente transiciones T→C (15 mutaciones) y A→G (8 mutaciones). En contraste, se detectaron únicamente 2/26 (7.7%) mutaciones en dirección opuesta (GC→AT), correspondientes a las sustituciones 190 C→T y 396 G→A. Adicionalmente, se registró 1/26 (3.8%) mutación neutral respecto al contenido GC (34 T→A).

En términos cuantitativos, las frecuencias iniciales de las mutaciones persistentes se ubicaron entre 5.0×10^{-4} y 1.62×10^{-3} , mientras que las frecuencias finales oscilaron entre 8.44×10^{-4} y 8.20×10^{-3} . Dentro del conjunto AT→GC, se observó que varias sustituciones presentaron incrementos netos cercanos o superiores a 5×10^{-4} , destacándose entre ellas las mutaciones 277 T→C ($\Delta AF = 0.000705$), 176 T→C ($\Delta AF = 0.000652$), 154 T→C ($\Delta AF = 0.000625$), 258 A→G ($\Delta AF = 0.000580$) y 314 T→C ($\Delta AF = 0.000583$). De forma complementaria, otras mutaciones

AT→GC exhibieron incrementos más moderados, pero persistentes, como 70 T→C, 211 T→C, 76 T→C y 304 T→C, con ΔAF en torno a $3 \times 10^{-4} - 4 \times 10^{-4}$.

También se observaron señales con incrementos relativamente más altos. En particular, la sustitución 34 T→A destacó como la variante con mayor expansión dentro del intervalo evaluado ($AF_{inicio} = 0.00162$; $AF_{final} = 0.00820$; $\Delta AF = 0.00658$). Asimismo, dentro de las mutaciones direccionalmente relevantes para el sesgo composicional, se identificaron incrementos notables en 190 C→T ($\Delta AF = 0.00163$) y 396 G→A ($\Delta AF = 0.00094$), ambas clasificadas como eventos GC→AT. Estos datos constituyen la base para los análisis posteriores de sesgo composicional y tendencia del contenido GC por generación.

Tabla 11. Mutaciones persistentes detectadas en *evgA*

Mutación	Gen inicio	Gen final	AF inicio	AF final	ΔAF	Dirección GC
190 C→T	1	7	0.0006247998	0.0022568698	0.00163207	GC→AT
176 T→C	1	7	0.0010823139	0.0017346005	0.0006522866	AT→GC
154 T→C	1	7	0.0008294652	0.0014543273	0.0006248621	AT→GC
214 T→C	1	7	0.0005685109	0.0011191879	0.0005506770	AT→GC
63 A→G	1	7	0.0005474259	0.0010829927	0.0005355668	AT→GC
277 T→C	2	7	0.0006000991	0.0013052953	0.0007051962	AT→GC
314 T→C	2	7	0.0006152629	0.0011984496	0.0005831867	AT→GC
150 A→G	2	7	0.0005639967	0.0010898008	0.0005258041	AT→GC
311 T→C	2	7	0.0005726975	0.0010718840	0.0004991865	AT→GC
299 T→C	2	7	0.0007845217	0.0012194624	0.0004349407	AT→GC
211 T→C	2	7	0.0005603586	0.0009423156	0.0003819570	AT→GC
70 T→C	2	7	0.0005646492	0.0009422171	0.0003775679	AT→GC
304 T→C	2	7	0.0006351357	0.0009526905	0.0003175548	AT→GC
258 A→G	3	7	0.0005649186	0.0011450781	0.0005801595	AT→GC
204 A→G	3	7	0.0005444855	0.0011037771	0.0005592916	AT→GC
297 A→G	3	7	0.0005381688	0.0010609712	0.0005228024	AT→GC
245 A→G	3	7	0.0005506197	0.0010713146	0.0005206949	AT→GC
87 T→C	3	7	0.0005281543	0.0010218454	0.0004936911	AT→GC
51 T→C	3	7	0.0005401347	0.0010325784	0.0004924437	AT→GC
76 T→C	3	7	0.0005599791	0.0009414285	0.0003814494	AT→GC
316 T→C	3	7	0.0006844820	0.0010314632	0.0003469812	AT→GC
34 T→A	4	7	0.0016206376	0.0082042367	0.0065835991	Neutral
279 A→G	4	7	0.0005039734	0.0008548658	0.0003508924	AT→GC
396 G→A	5	7	0.0007470000	0.0016820090	0.0009350090	GC→AT
194 T→C	5	7	0.0006001124	0.0009874382	0.0003873258	AT→GC
184 A→G	5	7	0.0005337416	0.0008443600	0.0003106184	AT→GC

6.3.2. Gen *ygiV*

La identificación de variantes puntuales a partir de conteos por posición se realizó para las primeras 11 generaciones (Tabla 12). En este intervalo, 8 mutaciones se clasificaron como Señales Biológicas ya que cumplen los criterios simultáneos de persistencia, crecimiento neto y estabilidad dinámica. Estas variantes presentaron frecuencias alélicas bajas, con valores iniciales comprendidos entre 5.0×10^{-4} y 8.0×10^{-4} , y frecuencias finales entre 8.3×10^{-4} y 1.68×10^{-3} , lo que evidencia incrementos netos medibles a lo largo de generaciones consecutivas. La distribución temporal mostró que 2 mutaciones se detectaron desde la generación 1 (268 A→G y 75 T→C), mientras que 3 mutaciones emergieron en la generación 3 (390 T→C, 177 A→G y 210 T→C). Adicionalmente, se identificaron señales tardías, con aparición en generaciones posteriores (170 T→C en generación 5; 284 T→C en generación 8; y 17 T→C en generación 9), las cuales se mantuvieron hasta la generación final del intervalo.

El conjunto de mutaciones persistentes estuvo compuesto exclusivamente por eventos AT→GC (8/8; 100%), incluyendo 6 transiciones T→C y 2 transiciones A→G. Por lo tanto, no se registraron mutaciones persistentes en dirección GC→AT ni cambios neutrales respecto al contenido GC dentro del intervalo analizado. El análisis de crecimiento evidenció que la mutación 268 A→G presentó el mayor incremento neto de frecuencia alélica ($\Delta AF = 0.000871$), alcanzando la frecuencia final más alta del conjunto ($AF_{final} = 0.001676$). En contraste, las señales de aparición tardía mostraron incrementos más moderados, como 17 T→C ($\Delta AF = 0.000302$) y 170 T→C ($\Delta AF = 0.000320$), lo cual es coherente con el menor número

de generaciones disponibles para su acumulación desde el momento de aparición. Estos resultados muestran que el gen *ygiV* acumula un conjunto limitado de mutaciones persistentes, compuesto exclusivamente por sustituciones AT→GC.

Tabla 12. Mutaciones persistentes detectadas en el gen *ygiV*

Mutación	Gen inicio	Gen final	AF inicio	AF final	ΔAF	Dirección GC
268 A→G	1	11	0.0008046726	0.0016759464	0.0008712738	AT→GC
75 T→C	1	11	0.0005517898	0.0010806438	0.0005288540	AT→GC
390 T→C	3	11	0.0005183303	0.0010842481	0.0005659178	AT→GC
177 A→G	3	11	0.0005549971	0.0010916075	0.0005366104	AT→GC
210 T→C	3	11	0.0005188157	0.0009190179	0.0004002022	AT→GC
170 T→C	5	11	0.0005055884	0.0008257736	0.0003201852	AT→GC
284 T→C	8	11	0.0005042048	0.0009177192	0.0004135144	AT→GC
17 T→C	9	11	0.0005822200	0.0008845000	0.0003022800	AT→GC

6.3.3. Gen *phnH*

La búsqueda de mutaciones acumuladas se enfocó en el intervalo generacional 1–6, debido a que en este rango el número de lecturas primarias alineadas que cumplían los filtros de calidad se mantuvo por encima de 100000 reads. En total se identificaron 4 Señales Biológicas (Tabla 13). Dos de estas variantes se detectaron desde la G1 (244 A→G y 237 T→C), mientras que las otras aparecieron en generaciones posteriores (328 T→C en la G2 y 192 T→C en la G3), manteniéndose todas hasta la generación G6. En términos cuantitativos, las frecuencias iniciales se ubicaron entre 5.8×10^{-4} y 1.09×10^{-3} , mientras que las

frecuencias finales oscilaron entre 8.9×10^{-4} y 1.54×10^{-3} , evidenciando incrementos netos comprendidos entre 3.1×10^{-4} y 6.2×10^{-4} .

En cuanto al tipo de sustitución, el conjunto completo de señales persistentes corresponde a cambios AT→GC (4/4; 100%), incluyendo tres transiciones T→C y una transición A→G. No se detectaron mutaciones persistentes clasificadas como GC→AT ni cambios neutrales respecto al contenido GC dentro del intervalo analizado. Entre las señales identificadas, la sustitución 244 A→G presentó el mayor incremento neto de frecuencia alélica ($\Delta AF = 0.000616$), seguida por 237 T→C ($\Delta AF = 0.000578$). Los resultados indican que el gen *phnH* exhibe un patrón restringido de mutaciones persistentes orientadas hacia el aumento de bases G/C.

Tabla 13. Mutaciones persistentes detectadas en el gen *phnH*

Mutación	Gen inicio	Gen final	AF inicio	AF final	ΔAF	Dirección GC
244 A→G	1	6	0.0009153625	0.0015308749	0.0006155124	AT→GC
237 T→C	1	6	0.0005809526	0.0011590426	0.0005780900	AT→GC
328 T→C	2	6	0.0010867962	0.0015378880	0.0004510918	AT→GC
192 T→C	3	6	0.0005785597	0.0008909757	0.0003124160	AT→GC

6.4 Balance neto de mutaciones y evaluación de direccionalidad mutacional

Al comparar los tres genes analizados, se observó un patrón consistente de direccionalidad mutacional hacia el incremento del contenido GC, aunque con magnitudes y dinámicas temporales diferenciadas entre fragmentos. En los tres casos, la suma de frecuencias alélicas asociada a mutaciones que favorecen GC ($\Sigma AF \{AT \rightarrow GC\}$) predominó ampliamente sobre la contribución de eventos $GC \rightarrow AT$ ($\Sigma AF \{GC \rightarrow AT\}$), lo que resultó en valores positivos del índice neto $\Sigma(\Delta GC \cdot AF)$ a lo largo de los intervalos evaluados.

El gen *evgA* presentó el mayor sesgo neto acumulado, con un incremento final–inicial de 0.01887, evidenciando una acumulación progresiva de sustituciones $AT \rightarrow GC$ durante el periodo de análisis (Figura 8). En contraste, *ygiV* mostró un sesgo positivo más moderado (0.007123), aunque con una trayectoria temporal sostenida y sin aportes apreciables de mutaciones $GC \rightarrow AT$, lo que sugiere una direccionalidad composicional estable, pero de menor intensidad relativa (Figura 9). Finalmente, el gen *phnH* exhibió el sesgo neto más bajo (0.003622), en concordancia con su menor número de generaciones evaluables y su conjunto más limitado de mutaciones persistentes (Figura 10). Aun así, la señal se mantuvo consistentemente positiva y dominada por sustituciones $AT \rightarrow GC$, indicando que incluso en un fragmento con menor acumulación detectable, el balance mutacional conserva la misma orientación.

En conjunto, los tres genes analizados presentaron un balance neto positivo hacia GC, lo que respalda la existencia de una direccionalidad mutacional consistente con la predicción central de la presión mutacional termodinámica. Las diferencias

cuantitativas observadas entre *evgA*, *ygiV* y *phnH* pueden atribuirse a características intrínsecas de cada fragmento amplificado, tales como la composición nucleotídica y el contexto de secuencia. Estos factores influyen tanto en la probabilidad de aparición de determinados tipos de sustitución como en la dinámica de acumulación de variantes a lo largo de las generaciones evaluadas.

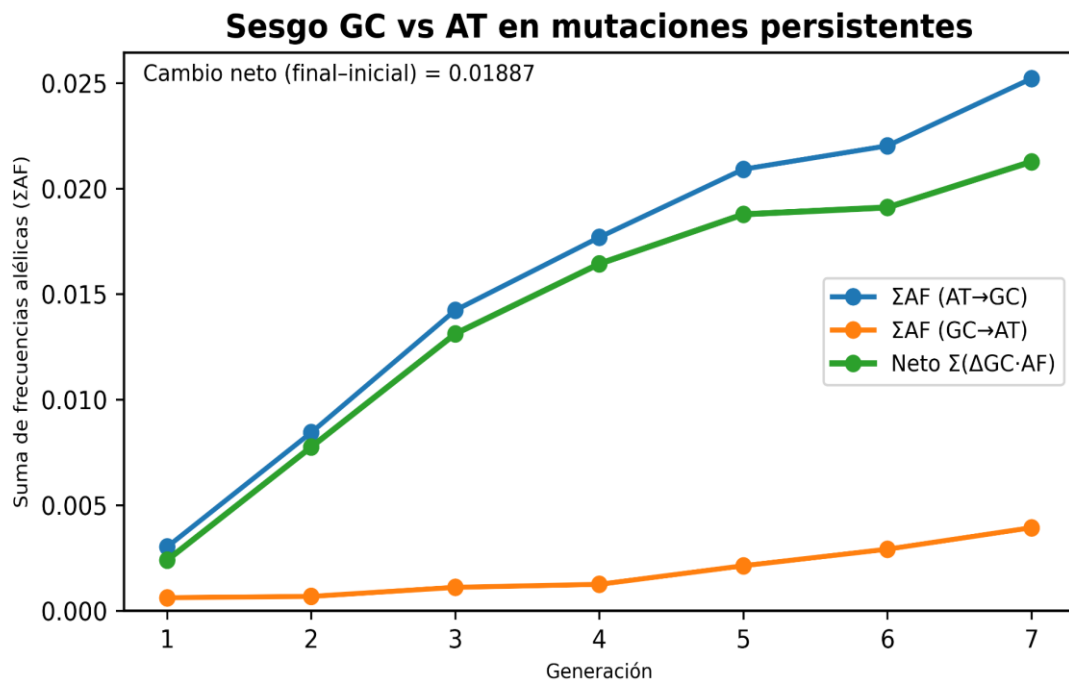


Figura 8. Sesgo GC vs AT en mutaciones persistentes del gen *evgA* (G1–G7). En el intervalo evaluado, $\Sigma AF_{\{AT \rightarrow GC\}}$ mostró un incremento sostenido desde aproximadamente 3.0×10^{-3} en G1 hasta 2.5×10^{-2} en G7, mientras que $\Sigma AF_{\{GC \rightarrow AT\}}$ se mantuvo considerablemente menor, aumentando de $\sim 5 \times 10^{-4}$ a $\sim 4.0 \times 10^{-3}$. Como resultado, el índice neto $\Sigma(\Delta GC \cdot AF)$ permaneció positivo en todas las generaciones y se incrementó de $\sim 2.5 \times 10^{-3}$ en G1 a $\sim 2.1 \times 10^{-2}$ en G7, con un cambio neto final–inicial de 0.01887.

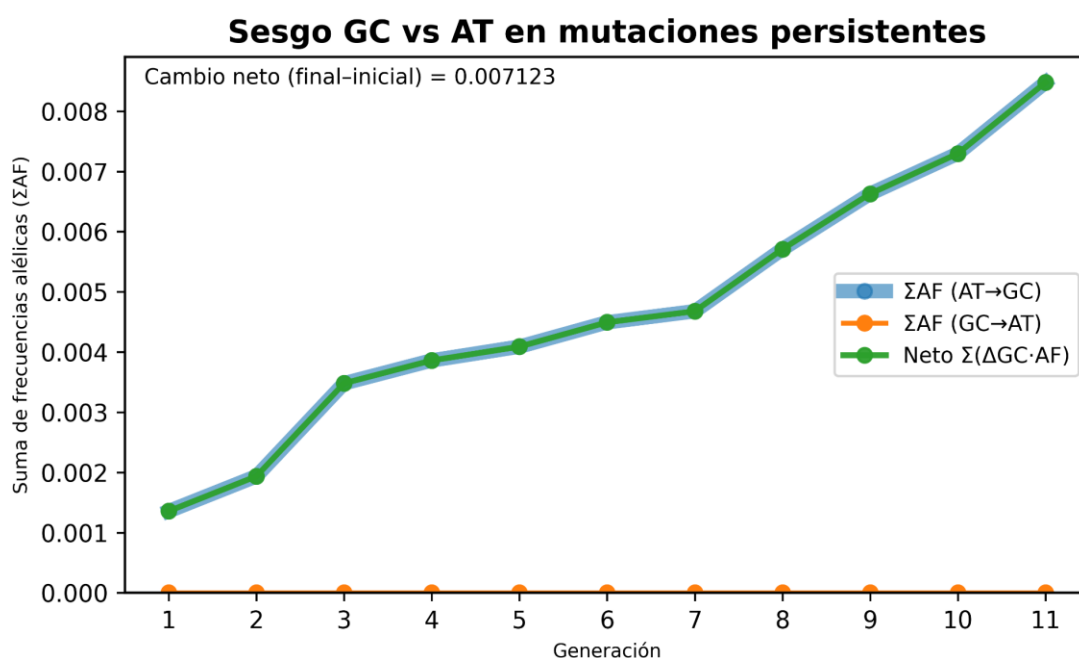


Figura 9. Sesgo GC vs AT en mutaciones persistentes del gen *ygjV* (G1–G11). El balance mutacional estuvo dominado por sustituciones $AT \rightarrow GC$ durante todo el intervalo analizado. La suma de frecuencias alélicas asociada a este tipo de cambios aumentó desde aproximadamente 1.4×10^{-3} en G1 hasta cerca de 8.5×10^{-3} en G11. Por el contrario, no se detectó un aporte relevante de mutaciones $GC \rightarrow AT$, cuya suma en cero en todas las generaciones. En consecuencia, el índice neto $\Sigma(\Delta GC \cdot AF)$ fue consistentemente positivo y registró un incremento acumulado de 0.007123 entre la primera y la última generación.

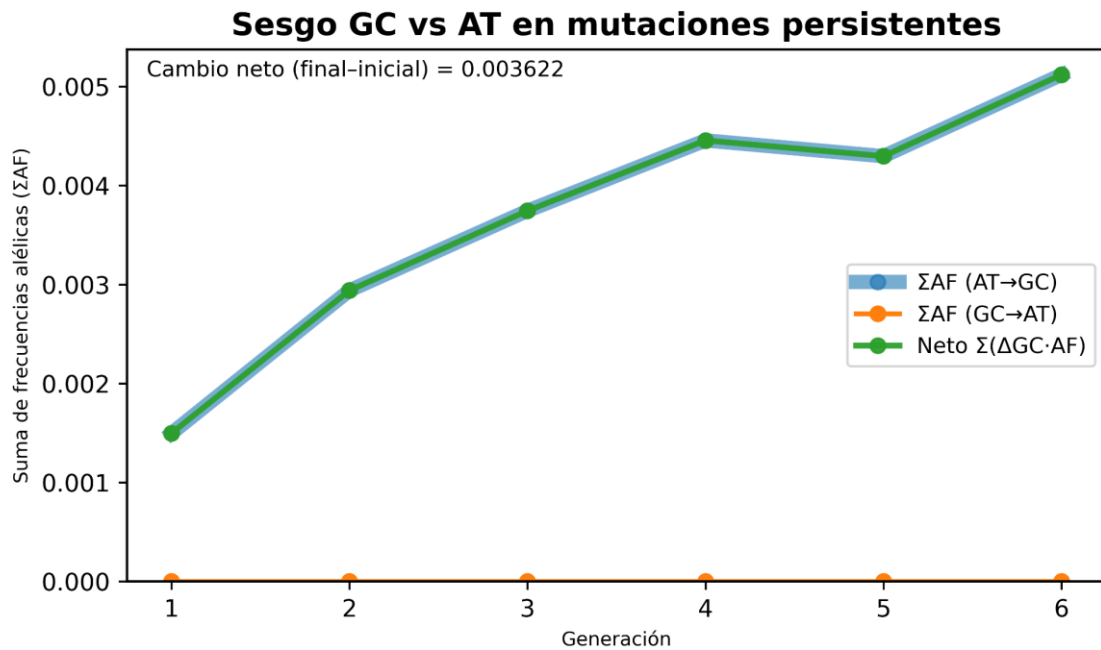


Figura 10. Sesgo GC vs AT en mutaciones persistentes del gen *phnH* (G1–G6). La señal acumulada estuvo dominada por sustituciones $AT \rightarrow GC$, cuya suma de frecuencias alélicas aumentó desde aproximadamente 1.5×10^{-3} en la G1 hasta alrededor de 5.1×10^{-3} en la G6. En cambio, las mutaciones $GC \rightarrow AT$ se mantuvieron en cero durante todo el intervalo. Por lo tanto, el índice neto $\Sigma(\Delta GC \cdot AF)$ permaneció positivo y registró un incremento acumulado de 0.003622 entre la generación inicial y final.

6.5 Trayectoria del contenido GC a lo largo de generaciones

En el intervalo evaluado del gen *evgA* (G1–G7), el porcentaje de GC (%GC) mostró un incremento progresivo a lo largo de las generaciones (Figura 11). El análisis de tendencia confirmó esta dinámica mediante una correlación de Spearman perfecta ($p = 1.000$), indicando una relación monotónica estrictamente positiva entre el número de generación (variable independiente) y el %GC (variable dependiente). Asimismo, la regresión ponderada por mínimos cuadrados (WLS) mostró una pendiente positiva de 0.00064 %GC por generación, con un intervalo de confianza al 95% entre 0.00038 y 0.00091, y un ajuste elevado (R^2 ponderado = 0.890). Esta tendencia resultó estadísticamente significativa ($p = 0.00143$).

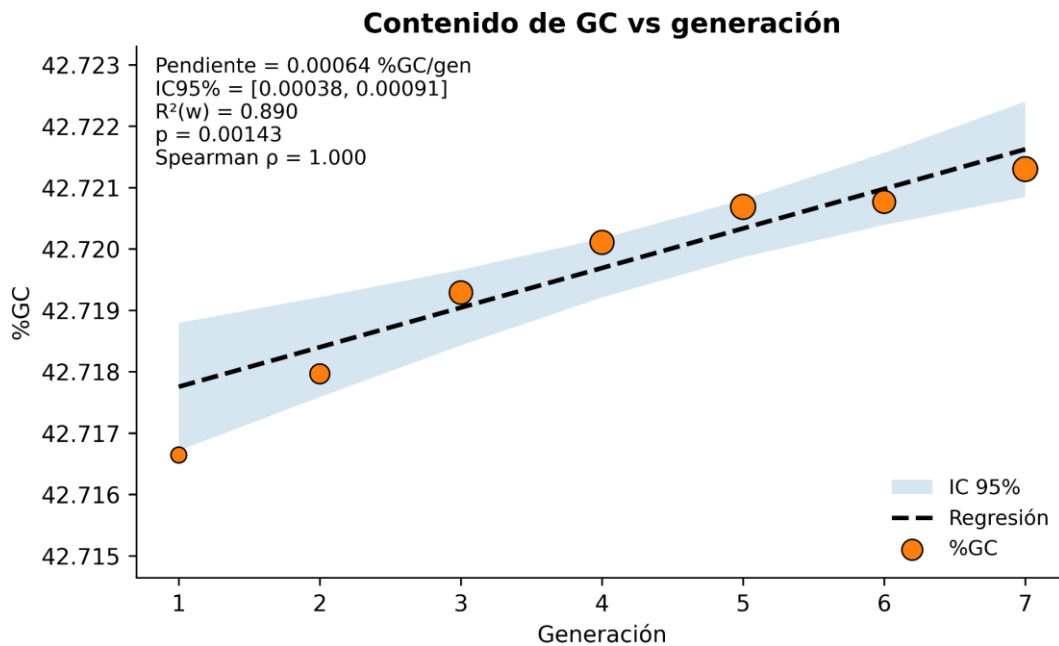


Figura 11. Contenido GC (%GC) a lo largo de generaciones en el gen *evgA*

En el gen *ygiV*, el porcentaje de GC (%GC) estimado por generación mostró una tendencia ascendente sostenida a lo largo del intervalo evaluado (G1–G11). La relación entre el número de generación y el %GC fue estrictamente positiva, evidenciada por una correlación de Spearman $\rho = 1.000$, lo que indica un incremento monótonico del contenido GC conforme avanzan las rondas de replicación (Figura 12). Este comportamiento fue corroborado mediante regresión ponderada por mínimos cuadrados (WLS), la cual arrojó una pendiente de 0.00014 %GC por generación, con un IC95% = [0.00012, 0.00016]. El ajuste del modelo fue alto (R^2 ponderado = 0.959) y la tendencia resultó altamente significativa ($p = 1.49 \times 10^{-7}$).

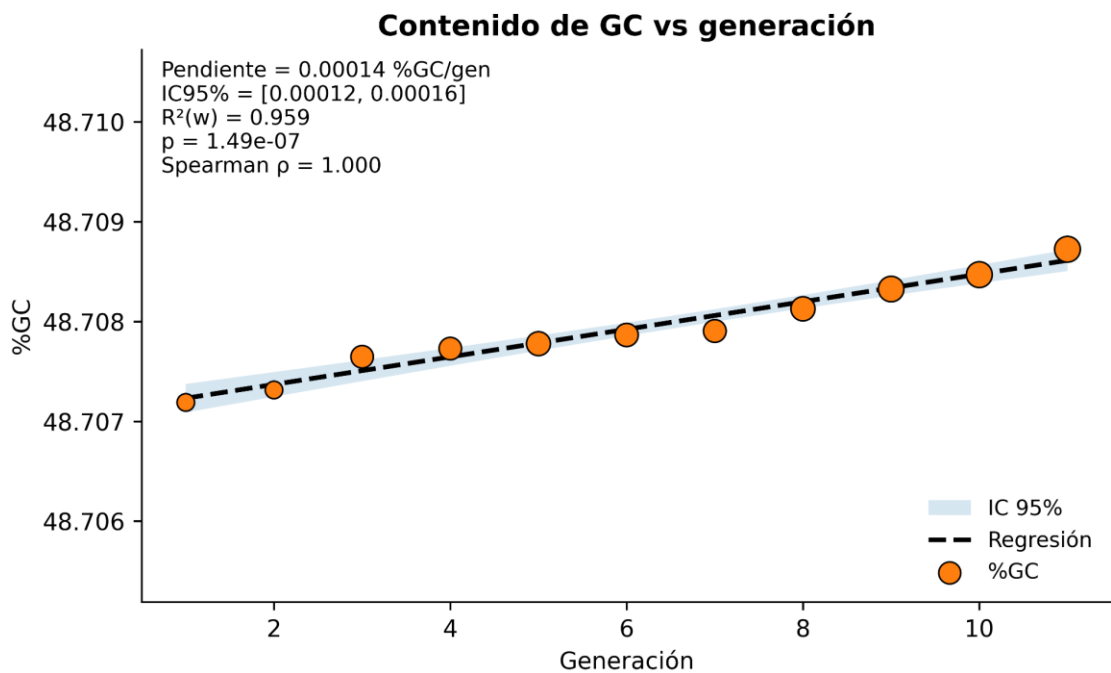


Figura 12. Contenido GC (%GC) a lo largo de generaciones en el gen *ygiV*

Para *phnH*, el contenido de GC estimado por generación evidenció un aumento leve pero consistente dentro del intervalo analizado (G1–G6). Los valores de %GC se mantuvieron dentro de un rango estrecho, pero con una progresión general ascendente, reflejada en una correlación de Spearman positiva y alta ($\rho = 0.943$), lo cual indica que el %GC tiende a incrementarse conforme avanza el número de generaciones (Figura 13). La regresión ponderada (WLS) confirmó una pendiente positiva de 0.00014 %GC por generación, con un intervalo de confianza del 95% [0.00006, 0.00021]. El modelo explicó una fracción importante de la variabilidad observada (R^2 ponderado = 0.872) y la tendencia fue estadísticamente significativa ($p = 0.00642$).

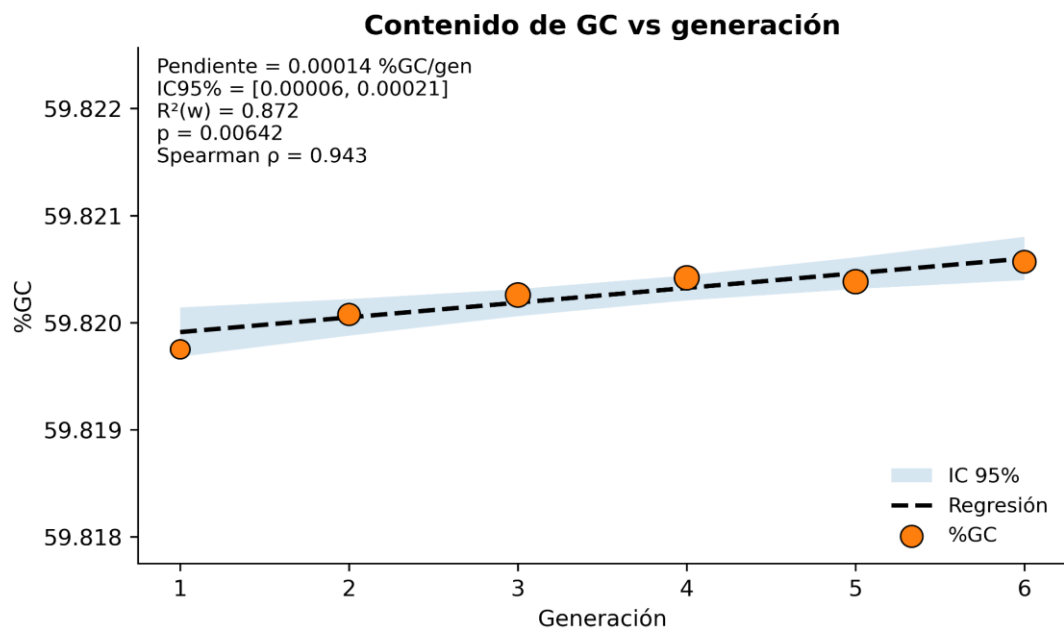


Figura 13. Contenido GC (%GC) a lo largo de generaciones en el gen *phnH*

6.6 Distribución de mutaciones según posición

La distribución por posición codónica para el gen *evgA* mostró un patrón altamente estructurado. La tercera base del codón concentró 11/26 variantes (42.3%), y todas fueron sinónimas (11/11), es decir, sin cambio en el aminoácido codificado. En contraste, las mutaciones ubicadas en la segunda base corresponden a 9/26 eventos (34.6%) y fueron exclusivamente *missense* (9/9), mientras que las de la primera base representaron 6/26 (23.1%), también todas *missense* (6/6). Es decir, las sustituciones sinónimas se concentraron en la tercera posición codónica, mientras que los cambios en las primeras posiciones del codón se asociaron a modificaciones aminoacídicas (Tabla 14).

Tabla 14. Mutaciones en *evgA* agrupadas por posición

Posición	Total	Sinónimas	Missense	AT→GC	GC→AT	Neutrales
1ª base	6	0	6	5	1	0
2ª base	9	0	9	8	1	0
3ª base	11	11	0	10	0	1
Total	26	11	15	23	2	1

Las mutaciones identificadas en el gen *ygjV* se distribuyeron en las tres posiciones del codon. En la tercera base se registraron 4/8 mutaciones (50%) y todas fueron sinónimas (4/4; 100%), consistentes con cambios en la posición más tolerante del código genético. En contraste, las mutaciones localizadas en la primera y segunda posición (1 y 3 mutaciones, respectivamente) fueron exclusivamente *missense* (100%), reflejando que las sustituciones en posiciones menos permisivas tendieron a asociarse con cambios aminoacídicos (Tabla 15).

Tabla 15. Mutaciones en *ygjV* agrupadas por posición

Posición	Total	Sinónimas	Missense	AT→GC	GC→A T	Neutrales
1ª base	1	0	1	1	0	0
2ª base	3	0	3	3	0	0
3ª base	4	4	0	4	0	0
Total	8	4	4	8	0	0

En el gen *phnH*, las mutaciones se distribuyeron entre la primera y la tercera posición. En la tercera base codónica se observaron 2 mutaciones sinónimas (100%), coherentes con cambios en una posición más degenerada. En contraste, las 2 mutaciones localizadas en la primera base fueron de tipo *missense* (100%), generando sustituciones aminoacídicas. No se detectaron variantes persistentes en la segunda posición codónica (Tabla 16).

Tabla 16. Mutaciones en *phnH* agrupadas por posición

Posición	Total	Sinónimas	Missense	AT→GC	GC→AT	Neutrales
1ª base	2	0	2	2	0	0
2ª base	0	0	0	0	0	0
3ª base	2	2	0	2	0	0
Total	4	2	2	4	0	0

La coexistencia de mutaciones sinónimas y no sinónimas sugiere que el sistema experimental permitió la aparición de sustituciones con potencial efecto funcional que no dependen de presiones selectivas. Además, la persistencia de múltiples mutaciones que incrementan el contenido de bases GC en las tres posiciones del codón, pero primordialmente en la tercera posición del codón, respaldan la

influencia de la presión mutacional termodinámica en la preferencia del uso de codones.

6.7 Contexto local de secuencia asociado a mutaciones

Con el fin de explorar si las mutaciones persistentes identificadas se asocian a un entorno de secuencia particular, se reconstruyó el contexto local de cada sustitución clasificada como Señal Biológica. Para ello, se extrajo desde la secuencia de referencia un motivo de 5 nucleótidos (± 2 bases respecto a la posición mutada), permitiendo evaluar la composición inmediata alrededor del sitio de sustitución. Posteriormente, se construyó un *sequence logo* para las mutaciones con sustituciones AT→GC (incremento de GC). En este tipo de representación, el eje X indica la posición relativa respecto a la mutación (0 corresponde a la base mutada), mientras que el eje Y representa la frecuencia relativa de cada nucleótido observado en esa posición dentro del conjunto de mutaciones. La altura de cada letra (A, C, G o T) es proporcional a su frecuencia, y las letras se apilan de manera que la suma total en cada posición corresponde a 1 (100%).

En *evgA*, se observó que el conjunto de mutaciones persistentes que favorecen GC (AT→GC) se encuentra mayoritariamente inmerso en un entorno local enriquecido en A/T, particularmente en las posiciones flanqueantes (-2, -1, +1, +2). En el logo generado para esta clase (n=23), las bases A y T dominan de forma consistente las posiciones laterales, indicando que las sustituciones AT→GC tienden a emerger y mantenerse en regiones localmente ricas en AT. En la posición central (0), la base predominante corresponde principalmente a T, lo cual es consistente con la alta frecuencia de sustituciones tipo T→C (Figura 14A).

El conjunto de mutaciones persistentes AT→GC en *ygjV* fue de tamaño moderado (n=8), y el patrón observado sugiere un contexto local menos uniforme. El análisis global de las posiciones laterales indica que la frecuencia de bases A/T tiende a ser mayor que la de bases G/C en el entorno inmediato del sitio mutado. Este comportamiento sugiere que las sustituciones AT→GC ocurren mayoritariamente dentro de segmentos donde el contexto local mantiene predominio AT, aunque con variabilidad dependiente de la posición (Figura 14B).

En *phnH* (n=4) el *sequence logo* de mutaciones persistentes sugiere que las sustituciones AT→GC detectadas se ubican en entornos donde la frecuencia de bases A/T tiende a ser mayor que la de bases G/C (Figura 14C). Incluso en un gen con menor número de eventos persistentes, la direccionalidad hacia GC se manifiesta principalmente en regiones donde existen sitios AT disponibles para conversión. Este patrón sugiere que el sesgo mutacional hacia GC observado en los tres genes se produce principalmente en regiones localmente ricas en AT.

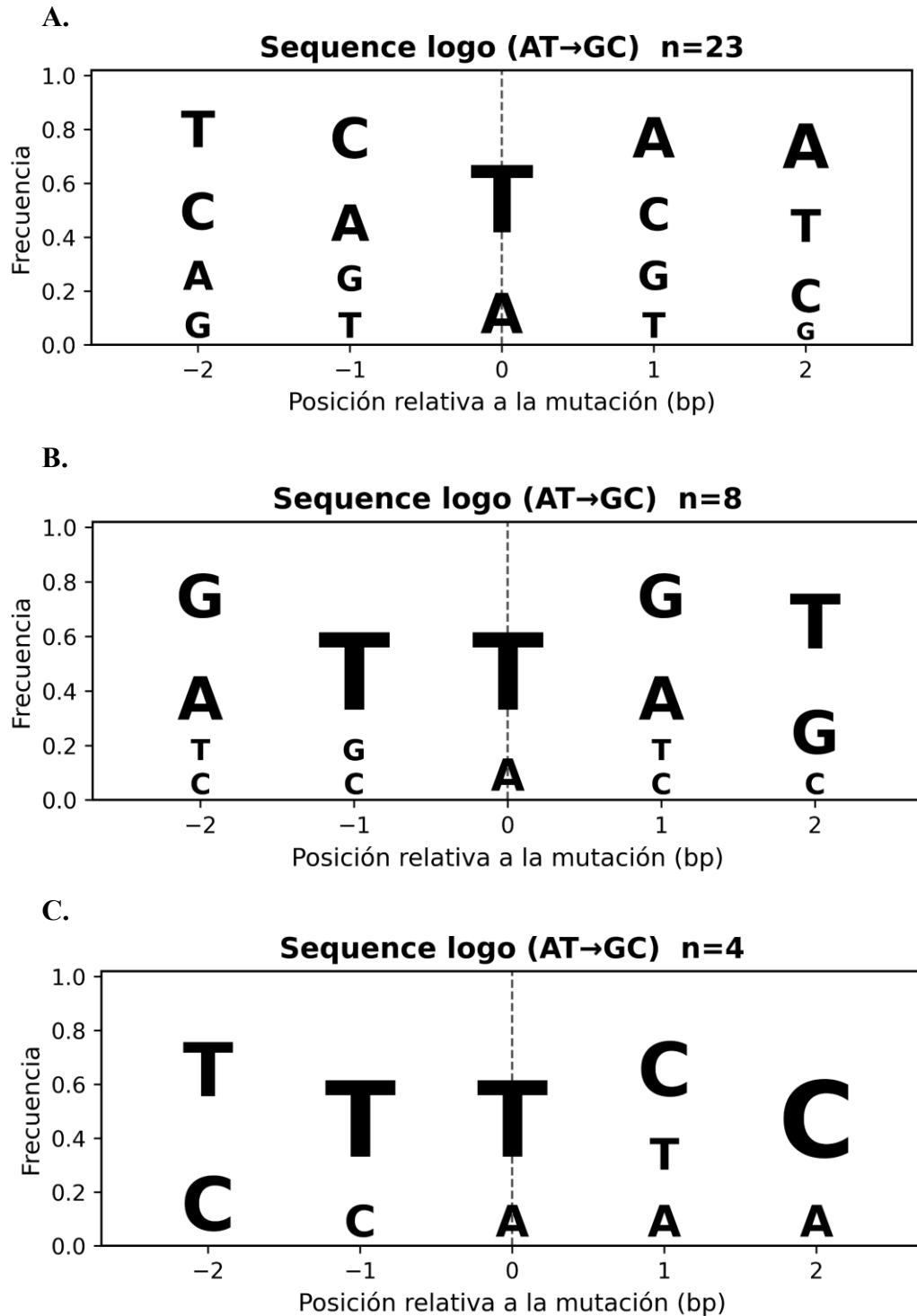


Figura 14. Sequence logos (± 2 bp; 5-mer) de mutaciones persistentes AT→GC en (A) *evgA* ($n=23$), (B) *ygjV* ($n=8$) y (C) *phnH* ($n=4$). La línea discontinua indica el sitio mutado (posición 0) y la altura de cada base representa su frecuencia relativa.

6.8 Análisis comparativo de tendencias del %GC por gen

Para evaluar la trayectoria del contenido de GC entre los genes analizados, se estimaron pendientes de %GC por generación mediante un enfoque de regresión ponderada (WLS), utilizando como pesos la cobertura total. Este procedimiento permitió reducir la influencia de puntos con menor soporte de lectura, aumentando la robustez de la estimación de tendencia. Dado que el número de generaciones disponibles no fue homogéneo entre genes y que los pesos podían diferir sustancialmente, los pesos fueron normalizados dentro de cada gen, garantizando que cada serie contribuyera de forma comparable al modelo global. Posteriormente, se aplicó un procedimiento de bootstrap ($n = 20\,000$ remuestreos) para obtener intervalos de confianza al 95% tanto para las pendientes individuales como para las diferencias entre pendientes.

La trayectoria de *evgA* presentó la pendiente más alta, con un incremento estimado de 0.645×10^{-3} unidades de %GC por generación, y un intervalo de confianza al 95% positivo (IC95%: 0.442×10^{-3} a 0.995×10^{-3}). En contraste, *ygiV* y *phnH* mostraron pendientes considerablemente menores y de magnitud muy similar (0.138×10^{-3} ; 0.137×10^{-3}), con intervalos de confianza positivos en ambos casos (IC95%: 0.1118×10^{-3} a 0.1569×10^{-3} ; IC95%: 0.0771×10^{-3} a 0.2227×10^{-3}).

El análisis de comparaciones (Tabla 17) confirmó que la pendiente de *evgA* es significativamente mayor que la de *ygiV* ($\Delta = 0.610 \times 10^{-3}$; $p = 4.06 \times 10^{-8}$) y *phnH* ($\Delta = 0.601 \times 10^{-3}$; $p = 8.81 \times 10^{-6}$). En cambio, no se detectaron diferencias significativas entre las pendientes de *ygiV* y *phnH* ($\Delta = -8.94 \times 10^{-6}$; $p = 0.992$), sugiriendo que ambos fragmentos presentan tasas de cambio composicional

comparables. Aunque el sesgo mutacional AT→GC se observa en los tres genes, *evgA* presenta una pendiente significativamente mayor porque contiene una mayor disponibilidad y concentración local de sitios A/T susceptibles de conversión, lo que incrementa la tasa efectiva de acumulación neta de GC. Este patrón es consistente con la idea de que la presión mutacional termodinámica puede expresarse con distinta intensidad según el contexto de secuencia, disponibilidad de sitios mutables y composición inicial.

Tabla 17. Comparación de pendientes ponderadas entre genes

Contraste	Diferencia de pendientes	IC95% (límite inferior)	IC95% (límite superior)	p-value
<i>evgA</i> – <i>ygjV</i>	0.5066×10^{-3}	0.2995×10^{-3}	0.8602×10^{-3}	0.0003
<i>evgA</i> – <i>phnH</i>	0.5072×10^{-3}	0.2815×10^{-3}	0.8647×10^{-3}	0.0003
<i>ygjV</i> – <i>phnH</i>	6.73×10^{-7}	-8.73×10^{-5}	6.30×10^{-5}	0.9527

VII. DISCUSIÓN

7.1 Sesgo mutacional observado y coherencia con la hipótesis

Los resultados obtenidos en este estudio aportan evidencia compatible con la hipótesis de la presión mutacional termodinámica. En el modelo experimental de amplificación seriada mediante RPA, realizado bajo condiciones mesófilas (37 °C) y en ausencia de presión selectiva, se detectó un patrón direccional de acumulación mutacional que condujo a un incremento progresivo del contenido de GC a lo largo de las generaciones experimentales. Todas las secuencias analizadas mostraron aumentos estadísticamente significativos en el porcentaje de guanina y citosina conforme avanzaba el proceso evolutivo simulado. De hecho, los análisis de regresión indicaron tendencias lineales positivas en el %GC por generación, con pendientes significativamente diferentes de cero. Aunque el cambio observado por cada ronda de replicación es relativamente pequeño, su consistencia y significancia estadística sugieren que no corresponde a fluctuaciones aleatorias, sino a la manifestación de un sesgo sistemático en la incorporación o fijación de mutaciones. Resulta particularmente relevante que el incremento se haya observado en los tres genes analizados, aun cuando presentan distintos contextos de secuencia y composición inicial. Asimismo, las diferencias en la magnitud de las pendientes indican que el sesgo AT→GC no se expresa con la misma intensidad en todos los fragmentos. En particular, *evgA*, al presentar menor contenido inicial de GC, tendría una mayor disponibilidad relativa de sitios A/T susceptibles de conversión hacia G/C, lo que explicaría su mayor tasa de acumulación neta de GC. En conjunto, la dirección concordante del cambio y la variación de su magnitud según el contexto de secuencia respaldan la interpretación de que el sistema de replicación isotérmica

en condiciones mesófilas favorece una acumulación no aleatoria de cambios nucleotídicos. En términos generales, incluso bajo temperaturas moderadas, la evolución neutra del ADN evidenció una tendencia hacia configuraciones que aumentan la estabilidad de la doble hélice, lo cual coincide con las predicciones del modelo de presión mutacional termodinámica.

Un aspecto adicional que respalda esta interpretación es que la mayor parte de las mutaciones detectadas contribuyó directamente al aumento del contenido de GC. Las sustituciones persistentes observadas con baja frecuencia corresponden principalmente a cambios como A→G o T→C, así como a otras sustituciones que incrementan la proporción relativa de nucleótidos G o C en la secuencia. Este patrón direccional coincidió con el análisis composicional global que evidenció el aumento del %GC y difícilmente puede explicarse como resultado de errores aislados de replicación. El hecho de que ciertas mutaciones se hayan mantenido y acumulado a lo largo de múltiples generaciones indica que no se trata simplemente de errores transitorios de la polimerasa, sino de un proceso mutacional con una dirección definida. Desde una perspectiva poblacional, puede interpretarse que las variantes con mayor contenido de GC aumentaron gradualmente su frecuencia dentro de la población de moléculas de ADN amplificadas durante cada ronda experimental. Dado que el sistema utilizado elimina la influencia de la selección natural (todas las moléculas replican por igual independientemente de su secuencia), la ventaja relativa de estas variantes solo puede explicarse por factores fisicoquímicos. En particular, las mutaciones que incrementan la estabilidad de apareamiento en la doble hélice (pares G≡C) parecen presentar una mayor probabilidad de fijación que aquellas que la reducen. En consecuencia, los resultados experimentales

concuerdan con el modelo teórico según el cual la estabilidad termodinámica del ADN puede generar un sesgo mutacional hacia nucleótidos GC a lo largo del tiempo evolutivo. En términos prácticos, esto implica que la composición nucleotídica de un gen puede modificarse gradualmente siguiendo principios termodinámicos asociados al proceso de replicación, incluso en ausencia de presiones adaptativas externas.

Otro resultado particularmente relevante es que la mayoría de las mutaciones identificadas correspondieron a cambios sinónimos localizados en la tercera posición del codón. Dado que el sistema experimental se desarrolló *in vitro* y sin presión selectiva externa, esta distribución no debe interpretarse como resultado de una menor restricción funcional sobre dicha posición, sino como un patrón compatible con sesgos mutacionales dependientes del contexto molecular de la secuencia. En este sentido, la mayor frecuencia de cambios hacia GC en la tercera posición podría relacionarse con el arreglo local de bases, la composición inicial de los codones y las diferencias de energía potencial de interacción entre bases durante la incorporación nucleotídica. Así, aunque la ocurrencia de mutaciones mantiene un componente estocástico, los resultados sugieren que su distribución y dirección composicional pueden estar influenciadas por propiedades fisicoquímicas locales del ADN [10]. Al no existir limitaciones estrictas para la aparición de cambios sinónimos, las mutaciones favorecidas desde el punto de vista termodinámico (aquellas que incrementan el contenido de GC) pueden acumularse con mayor facilidad, mientras que las mutaciones opuestas permanecen menos representadas. Este patrón es coherente con observaciones en genomas naturales, donde el aumento del contenido GC suele manifestarse con mayor intensidad en posiciones

sinónimas, como se ha descrito en tripanosomátidos [24]. En conjunto, los resultados obtenidos son compatibles con la existencia de un sesgo mutacional asociado a la presión mutacional termodinámica, capaz de modificar la composición de bases en poblaciones de secuencias sometidas a amplificaciones seriadas *in vitro*.

7.2 Comparación con estudios previos y validez de la hipótesis

Los resultados obtenidos en este estudio bajo condiciones mesófilas aportan nueva evidencia que complementa y amplía las observaciones reportadas en investigaciones anteriores, fortaleciendo el marco interpretativo de la hipótesis de la presión mutacional termodinámica. En particular, nuestros hallazgos respaldan y generalizan los resultados del experimento clásico realizado por Brunstein et al. (2002), quienes observaron un incremento del contenido de GC tras realizar 280 ciclos de PCR utilizando la ADN polimerasa Taq [11]. A partir de aquel estudio surgió una cuestión fundamental: si el sesgo hacia GC observado era consecuencia de las propiedades termófilas de la polimerasa Taq (y de las altas temperaturas asociadas a la PCR) o si reflejaba un principio más general del proceso de replicación del ADN. La simulación experimental realizada en el presente trabajo responde directamente a esa interrogante: el aumento del contenido de GC también se produce bajo condiciones moderadas de temperatura, lo que indica que el fenómeno no depende exclusivamente de la actividad de la polimerasa Taq ni de un entorno termofílico. La observación de un patrón equivalente utilizando una polimerasa de origen mesófilo (la polimerasa de *Bacillus subtilis* empleada en el sistema RPA) sugiere que la tendencia hacia un aumento del contenido de GC podría ser una característica inherente al proceso de replicación del ADN y no una

peculiaridad de una enzima específica. Este resultado refuerza la hipótesis de la presión mutacional termodinámica como un mecanismo subyacente común que podría explicar los patrones observados en ambos sistemas experimentales. Es importante señalar que, aunque las tasas de acumulación de mutaciones pueden variar entre sistemas (por ejemplo, la polimerasa Taq operando a 72 °C podría generar mutaciones a una mayor velocidad que la polimerasa de *B. subtilis* a 37 °C), la dirección del cambio composicional observado fue similar en ambos casos. En otras palabras, el sesgo hacia el incremento del contenido de GC se manifiesta tanto en condiciones termofílicas como mesofílicas, siempre que el sistema experimental sea *in vitro* y no imponga presión de selección sobre las variantes generadas. Bajo estas condiciones, las mutaciones pueden acumularse a través de rondas sucesivas de amplificación sin estar condicionadas por una ventaja adaptativa.

Nuestro estudio también extiende las observaciones preliminares reportadas por Arnáez et al., quienes exploraron esta hipótesis a 37 °C utilizando la polimerasa Klenow, aunque su estudio estuvo limitado por el reducido número de generaciones alcanzadas [13]. En aquel caso no fue posible analizar suficientes mutaciones como para establecer una tendencia clara en el contenido de GC. En contraste, el presente trabajo logró extender el número de generaciones analizadas, lo que permitió detectar y confirmar estadísticamente la tendencia composicional hacia el incremento de GC. De esta manera, aportamos la evidencia empírica que faltaba para evaluar la hipótesis en condiciones mesófilas. En el futuro sería interesante desarrollar experimentos adicionales que incluyan simulaciones evolutivas prolongadas utilizando la polimerasa Klenow, con el fin de comparar directamente

tres sistemas experimentales: PCR con polimerasa Taq en condiciones termofílicas, PCR mesófila con polimerasa Klenow y amplificación isotérmica mediante RPA. No obstante, la concordancia cualitativa entre nuestros resultados y los de Brunstein et al. sugiere que el fenómeno es reproducible independientemente del método de amplificación utilizado. Tanto en sistemas de PCR a alta temperatura como en amplificaciones isotérmicas a temperaturas moderadas, el ADN puede acumular mutaciones que incrementan su contenido de GC, siempre que dichas mutaciones no sean corregidas ni sometidas a presiones selectivas externas. Esto apunta a una propiedad intrínseca del ADN y de sus reglas de apareamiento de bases, más que a factores contextuales externos.

Desde una perspectiva más amplia, nuestros hallazgos experimentales se alinean con patrones evolutivos observados en otros contextos y ofrecen un marco conceptual para interpretarlos. Por ejemplo, se ha documentado que numerosos microorganismos termófilos presentan genomas con un contenido de GC más elevado en comparación con especies mesófilas [25]. Tradicionalmente, esta característica se ha interpretado como una posible adaptación funcional, ya que un mayor contenido de GC podría conferir mayor estabilidad estructural al ADN y al ARN en ambientes de alta temperatura [25]. Sin embargo, la hipótesis de la presión mutacional termodinámica plantea una explicación complementaria: parte de este enriquecimiento en GC podría originarse por sesgos mutacionales asociados a diferencias termodinámicas locales durante la incorporación nucleotídica, y no exclusivamente por selección adaptativa. En este sentido, la selección ejercida por altas temperaturas y el sesgo mutacional termodinámico no serían procesos necesariamente excluyentes, sino mecanismos que podrían actuar de manera

conjunta o reforzarse en determinados contextos. Los resultados obtenidos bajo condiciones experimentales controladas y sin presión selectiva externa apoyan la posibilidad de que el entorno termodinámico influya en la direccionalidad de la acumulación mutacional. En organismos mesófilos naturales, donde las condiciones ambientales no favorecen necesariamente el incremento del contenido de GC, este mismo sesgo podría existir, pero coexistir con otros procesos mutacionales o presiones selectivas que actúan en sentido contrario. Un ejemplo representativo se observa en ciertas bacterias endosimbiontes con genomas reducidos, las cuales presentan un elevado contenido de AT debido a la pérdida de mecanismos de reparación del ADN [26]. En estos casos, procesos mutacionales como la desaminación espontánea de citosinas, que favorecen conversiones GC→AT, pueden dominar la dinámica evolutiva del genoma [26]. Este contraste ilustra que la evolución genómica real resulta del equilibrio entre múltiples fuerzas mutacionales (incluyendo efectos termodinámicos, procesos químicos y características del sistema replicativo) junto con las presiones selectivas propias de cada organismo y ambiente. En el contexto de nuestro experimento, se buscó aislar deliberadamente el componente termodinámico del sesgo mutacional mediante un sistema controlado en el que se minimizaron otras influencias. Bajo estas condiciones, dicha fuerza por sí sola es capaz de inducir un desplazamiento gradual de la composición nucleotídica hacia configuraciones más estables, es decir, hacia un mayor contenido de GC. En sistemas biológicos naturales, esta fuerza probablemente actúa en combinación con otros procesos evolutivos; dependiendo de su intensidad relativa, podría contribuir a moldear lentamente la composición

genómica o quedar equilibrada por otros sesgos mutacionales y presiones selectivas presentes en el organismo.

7.3 Limitaciones del estudio y trabajo futuro

Si bien el presente trabajo aporta evidencia compatible con la hipótesis de la presión mutacional termodinámica, es necesario considerar tanto los alcances como las limitaciones del enfoque experimental empleado. En primer lugar, el modelo experimental de amplificaciones seriadas basado en RPA, a pesar de haber alcanzado un mayor número de generaciones que el reportado en estudios previos, presentó dificultades técnicas conforme avanzaban las rondas de amplificación. En particular, se observó que después de 7 generaciones la reacción comenzó a perder especificidad, evidenciándose la aparición de productos inespecíficos y fragmentos de menor tamaño, fenómeno que se manifestó como *smearing* en los geles de agarosa. La acumulación progresiva de estos subproductos redujo la confiabilidad de las amplificaciones posteriores y limitó el número de generaciones que podían analizarse con precisión. En consecuencia, en algunos genes fue necesario restringir el análisis composicional a las primeras 6–11 generaciones, etapa en la cual los alineamientos contra la secuencia de referencia permanecían suficientemente robustos. Este comportamiento probablemente refleja una limitación inherente al sistema RPA cuando se realizan múltiples rondas consecutivas de amplificación. A diferencia de una reacción convencional, en este modelo el producto de cada generación se utiliza como molde para la siguiente. Por ello, dímeros de *primers* o productos parcialmente amplificados pueden transferirse entre rondas y amplificarse progresivamente. Debido al carácter exponencial de la reacción, estos subproductos pueden competir con el fragmento esperado y generar mayor

heterogeneidad, visible como *smearing* o bandas inespecíficas en los geles de agarosa. Esta interpretación es concordante con reportes previos que describen en RPA la formación de artefactos dependientes de *primers* y productos inespecíficos [27]. Desde un punto de vista práctico, este fenómeno restringe la posibilidad de extender indefinidamente las amplificaciones seriadas. Para futuras investigaciones, más que repetir la optimización de las condiciones de reacción, sería recomendable implementar estrategias más selectivas orientadas a reducir el arrastre de subproductos. Entre ellas, se propone evaluar un reinicio periódico de la serie mediante captura híbrida del amplicón con un conjunto de sondas biotiniladas internas y perlas magnéticas de estreptavidina [28]. Otra mejora metodológica importante sería incorporar réplicas experimentales independientes para cada gen, bajo condiciones idénticas de amplificación. En el presente estudio, cada secuencia inicial siguió una única trayectoria mutacional; por tanto, repetir el experimento en paralelo para cada gen permitiría evaluar la reproducibilidad del incremento de GC y cuantificar con mayor precisión la variabilidad entre series experimentales. Esta aproximación permitiría diferenciar tendencias mutacionales consistentes de fluctuaciones aleatorias propias de la amplificación seriada, fortaleciendo así la robustez estadística del modelo experimental.

Asimismo, es importante reconocer que nuestro estudio se centró en fragmentos génicos relativamente cortos, de aproximadamente 400 pb, analizados en condiciones controladas *in vitro*. Por ello, la extrapolación directa de estos resultados a la evolución de genomas completos en organismos vivos debe realizarse con cautela. En los sistemas biológicos reales intervienen múltiples procesos adicionales, incluyendo mecanismos de reparación del ADN, la acción de

distintas polimerasas replicativas, presiones selectivas funcionales y otros factores evolutivos como la deriva genética o sesgos mutacionales alternativos. Una línea prometedora de investigación futura, contemplada dentro del proyecto general, consiste en examinar secuencias obtenidas de experimentos de evolución a largo plazo en laboratorio, como el conocido experimento de 60,000 generaciones de *E. coli* desarrollado por Lenski [29]. Aunque en estos sistemas la selección natural sigue actuando (ya que las poblaciones bacterianas se adaptan a su entorno), es posible analizar mutaciones neutrales acumuladas para determinar si existe algún sesgo composicional consistente con las predicciones de la presión mutacional termodinámica. De manera complementaria, los estudios de perfiles de equilibrio mutacional en virus podrían constituir un punto comparativo relevante, ya que permiten estimar la composición nucleotídica esperada a partir del espectro de sustituciones observado. Para ello, el análisis debería centrarse preferentemente en mutaciones putativamente neutrales, como las sustituciones sinónimas ubicadas en sitios de cuatro veces degenerados, donde el cambio nucleotídico no altera la secuencia aminoacídica. De este modo, sería posible evaluar si el equilibrio composicional estimado se desplaza en una dirección compatible con las predicciones del modelo termodinámico. La detección de patrones composicionales similares en estos contextos reforzaría la relevancia del fenómeno en procesos evolutivos reales. Por el contrario, la ausencia de dicha señal podría indicar que, en presencia de selección y de otros sesgos mutacionales, el efecto termodinámico queda enmascarado o se manifiesta de manera tenue.

Los datos obtenidos también permiten realizar una estimación preliminar de la magnitud del efecto observado. Bajo las condiciones experimentales utilizadas, la

tasa aproximada de cambio en el contenido de GC fue del orden de 10^{-3} % por generación en fragmentos de alrededor de 400 pb, considerando un sistema de replicación sin actividad de corrección de errores. Este resultado plantea nuevas preguntas sobre cómo podría variar dicha tasa bajo diferentes contextos experimentales. Por ejemplo, sería interesante determinar si el cambio composicional ocurre con mayor rapidez cuando se utilizan polimerasas termófilas como Taq a 72 °C, donde la energía térmica disponible es mayor, o si se reduce en presencia de enzimas con actividad de *proofreading*. Estas cuestiones podrían abordarse mediante experimentos comparativos de amplificaciones seriadas *in vitro* basados en PCR con Taq, PCR con Klenow y RPA isotérmica. Si bien el modelo teórico sugiere que el equilibrio final de GC debería ser independiente de la enzima utilizada (siempre que no exista selección), es probable que las dinámicas temporales del proceso difieran entre sistemas. Explorar estas diferencias permitiría comprender mejor la velocidad y las condiciones en las que opera la presión mutacional termodinámica. Asimismo, sería interesante investigar hasta qué punto este fenómeno depende del contexto de la secuencia analizada. En principio, el sesgo mutacional podría manifestarse de manera distinta en regiones del genoma con composiciones iniciales diferentes o con propensión a formar estructuras secundarias particulares. En este estudio se utilizaron secuencias derivadas de *E. coli* con un contenido GC moderado cercano al 50 %. Experimentos futuros podrían evaluar secuencias inicialmente ricas en AT para determinar si la presión hacia GC produce cambios más pronunciados en etapas tempranas de la evolución experimental. De manera similar, analizar secuencias ya ricas en GC permitiría

evaluar si el efecto se atenúa conforme la composición nucleotídica se aproxima a un posible estado de equilibrio.

En resumen, las principales limitaciones están asociadas a la duración limitada de las simulaciones (condicionada por la pérdida progresiva de especificidad en la reacción RPA), al incremento de heterogeneidad en los productos amplificados y a la ausencia de réplicas evolutivas independientes para cada gen analizado. No obstante, a pesar de estas restricciones, los resultados obtenidos proporcionan evidencia experimental consistente con la hipótesis de la presión mutacional termodinámica en condiciones mesófilas. Entre los hallazgos más relevantes destacan: (i) el predominio de mutaciones persistentes que transforman bases AT en GC, (ii) un balance mutacional global positivo hacia nucleótidos G y C, y (iii) tendencias ascendentes estadísticamente significativas del porcentaje de GC a lo largo de las generaciones analizadas. Este estudio demuestra, que existe una fuerza mutacional sesgada que favorece configuraciones más estables de la molécula de ADN. Estos resultados abren nuevas posibilidades para investigar con mayor precisión la magnitud y el papel de este fenómeno en distintos contextos biológicos. De este modo, se establecen las bases para reconsiderar la evolución molecular desde una perspectiva que integre, junto con la selección natural, las influencias derivadas de las propiedades físicas y químicas del ADN. Las próximas investigaciones podrán determinar en qué medida la presión mutacional termodinámica ha contribuido a moldear los genomas a lo largo de la historia evolutiva y cómo interactúa con otras fuerzas evolutivas en distintos organismos. Comprender estas dinámicas podría tener implicancias no solo en biología evolutiva teórica, sino también en áreas aplicadas como la biotecnología y la

medicina evolutiva (por ejemplo, en la predicción de trayectorias mutacionales en patógenos bajo diferentes condiciones ambientales). En última instancia, los resultados aquí presentados sugieren que las mutaciones no ocurren completamente al azar, sino que pueden reflejar, al menos en parte, las condiciones fisicoquímicas del entorno molecular en el que se producen, ampliando así nuestra comprensión de los mecanismos fundamentales de la evolución.

VIII. CONCLUSIONES

- Se implementó un modelo experimental *in vitro* basado en amplificaciones isotérmicas seriadas por RPA, que permitió generar poblaciones de amplicones por generación y evaluar su composición nucleotídica mediante secuenciación de Oxford Nanopore.
- En los tres genes analizados (*evgA*, *ygjV* y *phnH*) se observó un predominio marcado de sustituciones persistentes en dirección AT a GC, mientras que las sustituciones GC a AT fueron escasas o no detectables en los intervalos con mapeo confiable.
- El balance mutacional neto, calculado a partir de las frecuencias alélicas de mutaciones persistentes, fue positivo hacia GC en los tres genes, lo que indica una direccionalidad composicional consistente durante las primeras siete generaciones.
- El porcentaje de GC estimado por generación mostró tendencias positivas y estadísticamente significativas en *evgA*, *ygjV* y *phnH* dentro de los intervalos evaluados, lo que respalda que el sesgo hacia GC se acumula de manera detectable en escalas generacionales cortas.
- La distribución por posición codónica mostró que una fracción relevante de mutaciones se concentró en la tercera posición del codón y fue sinónima, patrón coherente con la acumulación de señal mutacional en posiciones con menor restricción bajo ausencia de selección dirigida.
- El deterioro progresivo del mapeo y la aparición de productos inespecíficos en generaciones tardías evidencian una limitación experimental inherente a la amplificación seriada por RPA, por lo que los análisis de direccionalidad

deben interpretarse principalmente en los intervalos tempranos con alta proporción de lecturas primarias alineadas.

- En conjunto, los resultados obtenidos de las amplificaciones seriadas mediante RPA en condiciones mesófilas (37 °C) son compatibles con la hipótesis de presión mutacional termodinámica, al mostrar una direccionalidad mutacional neta hacia el incremento del contenido de guanina y citosina (GC) en los fragmentos génicos evaluados.

IX. RECOMENDACIONES

- Para reforzar la validez estadística, sería recomendable realizar en paralelo múltiples experimentos de amplificaciones seriadas bajo condiciones experimentales equivalentes para los genes analizados. La implementación de réplicas independientes permitiría evaluar con mayor precisión la reproducibilidad del incremento en el contenido de GC y estimar la variabilidad existente entre distintas trayectorias mutacionales. La observación de patrones convergentes entre estas líneas experimentales respaldaría la interpretación de que el sesgo hacia GC constituye una propiedad intrínseca del proceso replicativo. En cambio, la aparición de divergencias entre réplicas podría aportar información relevante sobre la contribución de factores estocásticos o sobre el impacto de mutaciones particulares que modifiquen la dinámica evolutiva del sistema.
- Otra línea de investigación especialmente relevante consiste en reproducir el experimento utilizando distintas polimerasas mesófilas además de la empleada en este estudio. La comparación sistemática de los resultados

obtenidos con diferentes enzimas permitiría distinguir en qué medida el sesgo mutacional observado responde a principios termodinámicos generales del sistema ADN-replicación o si depende parcialmente de propiedades específicas de cada polimerasa. Este tipo de análisis comparativo también podría tener implicancias prácticas en el ámbito de la biotecnología, ya que permitiría seleccionar enzimas de replicación en función del sesgo composicional que se desee favorecer o evitar durante procesos de amplificación *in vitro*.

X. REFERENCIAS BIBLIOGRÁFICAS

1. Choudhuri S. Fundamentals of molecular evolution. In: Choudhuri S, editor. Bioinformatics for Beginners. Amsterdam: Elsevier; 2014. p. 27–53. doi:10.1016/B978-0-12-410471-6.00002-5
2. Genetic mutation [Internet]. Nature.com. [citado el 27 de febrero de 2025]. Disponible en: <https://www.nature.com/scitable/topicpage/genetic-mutation-441/>
3. Brown TA. Mutation, repair and recombination. Nueva York, NY, Estados Unidos de América: Wiley-Liss; 2002.
4. Errors in DNA replication [Internet]. Nature.com. [citado el 27 de febrero de 2025]. Disponible en: <https://www.nature.com/scitable/topicpage/dna-replication-and-causes-of-mutation-409/>
5. DNA and mutations [Internet]. Berkeley.edu. [citado el 27 de febrero de 2025]. Disponible en: <https://evolution.berkeley.edu/dna-and-mutations/>
6. Taylorandfrancis.com. [cited 2025 May 12]. Available from: https://taylorandfrancis.com/knowledge/Engineering_and_technology/Bio_medical_engineering/Selective_pressure/
7. Zimić MJ. Una simulación teórica de la evolución de secuencias de ADN basada en la estabilidad termodinámica [Tesis de posgrado]. Perú: Universidad Peruana Cayetano Heredia; 1998.
8. Uni-frankfurt.de. [cited 2025 Jun 3]. Available from: https://itp.uni-frankfurt.de/~gros/Vorlesungen/TD/9_Canonical_ensemble.pdf

9. Zimić MJ, Guerra D, Arévalo J. DNA thermodynamic pressure: a potential contributor to genome evolution. *Trans R Soc Trop Med Hyg.* 2002 Apr;96 Suppl 1:S15-20. doi: 10.1016/s0035-9203(02)90046-5.
10. Zimić MJ. La presión mutacional termodinámica en el mecanismo de la replicación del ADN y sus efectos en la evolución molecular biológica [Tesis de Licenciatura]. Lima, Perú: Universidad Nacional de Ingeniería; 2023.
11. Sanson GF, Kawashita SY, Brunstein A, Briones MR. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. *Mol Biol Evol.* 2002 Feb;19(2):170-8. doi: 10.1093/oxfordjournals.molbev.a004069.
12. Neb.com. [cited 2025 Mar 12]. Available from: <https://www.neb.com/en/tools-and-resources/selection-charts/dna-polymerase-selection-chart>
13. Arnáez A. Estudio de la acumulación de mutaciones puntuales durante el proceso de replicación del ADN y evaluación de la hipótesis de la presión termodinámica mediante un modelo evolutivo *in vitro* que utiliza un PCR de baja temperatura [Tesis de pregrado]. Perú: Universidad Peruana Cayetano Heredia; 2014.
14. Daher RK, Stewart G, Boissinot M, Bergeron MG. Recombinase Polymerase Amplification for Diagnostic Applications. *Clin Chem.* 2016 Jul;62(7):947-58. doi: 10.1373/clinchem.2015.245829.
15. Tan M, Liao C, Liang L, Yi X, Zhou Z, Wei G. Recent advances in recombinase polymerase amplification: Principle, advantages,

- disadvantages and applications. *Front Cell Infect Microbiol.* 2022 Nov 28;12:1019071. doi: 10.3389/fcimb.2022.1019071.
16. González-González R, García Medel P, Cordoba F, Peralta-Castro A, Castro-Lara AY, Díaz-Quezada C, et al. La tecnología de amplificación mediada por recombinasas y polimerasas para detectar al SARS-Cov2. *BioTecnología.* 2020;24(3):29-38.
 17. Oxford Nanopore Technologies [Internet]. Oxford Nanopore Technologies. 2022 [cited 2025 Mar 13]. Available from: <https://nanoporetech.com/es/document/ligation-sequencing-amplicons-native-barcoding-v14-sqk-nbd114-96>
 18. Decoding sequencing depth and coverage [Internet]. Cd-genomics.com. [cited 2025 Mar 13]. Available from: <https://www.cd-genomics.com/resource-sequencing-depth-and-coverage.htm>
 19. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics.* 2023 May 4;39(5):btad311. doi: [10.1093/bioinformatics/btad311](https://doi.org/10.1093/bioinformatics/btad311)
 20. Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H, Horiuchi T. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol Syst Biol.* 2006;2:2006.0007. doi: [10.1038/msb4100049](https://doi.org/10.1038/msb4100049)
 21. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018 Sep 15;34(18):3094-3100. doi: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191).

22. Wang J, Liu L, Wang J, Sun X, Yuan W. Recombinase polymerase amplification assay—a simple, fast and cost-effective alternative to real time PCR for specific detection of feline herpesvirus-1. *PLoS One*. 2017;12(1):e0166903. doi:10.1371/journal.pone.0166903.
23. Chen CY. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Front Microbiol*. 2014;5:305. doi:10.3389/fmicb.2014.00305.
24. Alonso G, Guevara P, Ramirez JL. Trypanosomatidae codon usage and GC distribution. *Mem Inst Oswaldo Cruz*. 1992 Oct-Dec;87(4):517-523.
25. Hu EZ, Lan XR, Liu ZL, Gao J, Niu DK. A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics*. 2022 Feb 9;23(1):110. doi: 10.1186/s12864-022-08353-7.
26. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*. 2010;6(9):e1001115. doi:10.1371/journal.pgen.1001115
27. Wu H, Zhao P, Yang X, Li J, Zhang J, Zhang X, et al. A recombinase polymerase amplification and lateral flow strip combined method that detects *Salmonella enterica* serotype Typhimurium with no worry of primer-dependent artifacts. *Front Microbiol*. 2020;11:1015. doi:10.3389/fmicb.2020.01015.
28. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-

generation sequencing. Nat Methods. 2010 Feb;7(2):111-8. doi: 10.1038/nmeth.1419.

29. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature. 2016;536(7615):165-170. doi:10.1038/nature18959.

XI. ANEXOS

ANEXO 1: Teoría de la presión mutacional termodinámica

La evolución molecular se refiere a los cambios que ocurren en la composición genética de los organismos a lo largo del tiempo. Estas modificaciones tienen su origen principalmente en mutaciones que alteran la secuencia del ADN. Dichas mutaciones pueden surgir como consecuencia de errores espontáneos durante el proceso de replicación o debido a la acción de agentes mutagénicos externos, como compuestos químicos o radiación ionizante, que interactúan con el ADN y modifican nucleótidos específicos. En la mayoría de los casos, las mutaciones corresponden a cambios puntuales que implican la sustitución, inserción o eliminación de uno o varios nucleótidos. No obstante, también pueden ocurrir alteraciones a mayor escala que afectan segmentos completos de cromosomas. En estos casos, regiones del ADN pueden ser eliminadas, duplicadas, invertidas o desplazadas a nuevas posiciones dentro del genoma mediante eventos de translocación.

Los errores cometidos durante la replicación constituyen una de las principales fuentes de mutaciones puntuales. Las ADN polimerasas poseen una elevada precisión al seleccionar nucleótidos complementarios durante la síntesis de nuevas hebras de ADN. Sin embargo, aún bajo condiciones normales, estas enzimas pueden incorporar nucleótidos incorrectos con una frecuencia aproximada de un error por cada 100 000 nucleótidos replicados. Durante la síntesis del ADN, los nucleótidos trifosfato libres pueden ingresar al sitio activo de la polimerasa incluso cuando no son perfectamente complementarios a la secuencia molde. Si este nucleótido incorrecto logra formar un enlace fosfodiéster y el sistema de reparación

no corrige el error, se genera un desajuste de apareamiento (mismatch) en la doble hélice recién sintetizada. En posteriores ciclos de replicación, dicho desajuste puede fijarse como una mutación permanente. Como consecuencia, la secuencia resultante difiere de la cadena original y puede actuar como molde para replicaciones futuras, dando lugar a nuevas moléculas de ADN que conservan la mutación inicial. De esta manera, los errores de replicación modifican gradualmente la composición nucleotídica de las secuencias genómicas a lo largo de múltiples generaciones, contribuyendo al proceso evolutivo al generar nuevas variantes genéticas.

Una vez generadas, estas variantes genéticas están sujetas a la acción de la selección natural. Las condiciones ambientales (como la disponibilidad de recursos, la temperatura, la presión, la presencia de oxígeno o agua y la exposición a sustancias potencialmente tóxicas) influyen de manera diferencial en la supervivencia y reproducción de los organismos. Dado que los individuos presentan diferentes composiciones genéticas y, en consecuencia, distintos fenotipos, algunos de ellos estarán mejor adaptados al ambiente que otros. Aquellos organismos cuyas mutaciones o características genéticas reducen su fitness o aptitud biológica presentan menores probabilidades de sobrevivir y reproducirse, lo que puede conducir eventualmente a su desaparición. Por el contrario, las variantes genéticas que confieren ventajas adaptativas permiten a los organismos sobrevivir con mayor éxito y transmitir esos rasgos a su descendencia. En este sentido, las presiones selectivas del entorno actúan sobre las variaciones genéticas existentes y contribuyen a moldear la evolución de los genotipos en las poblaciones.

El análisis detallado del proceso de replicación del ADN resulta esencial para comprender uno de los mecanismos fundamentales mediante los cuales se generan

nuevas variantes genéticas que impulsan la diversidad biológica y la adaptación evolutiva. Con este objetivo, en 1998 Zimic y colaboradores desarrollaron un enfoque computacional basado en principios termodinámicos para estudiar el mecanismo de replicación del ADN y estimar la probabilidad de aparición de mutaciones durante la síntesis de nuevas cadenas. Desde la perspectiva de la termodinámica estadística, los sistemas biológicos no se encuentran estrictamente en equilibrio térmico, ya que intercambian continuamente materia y energía con su entorno. Sin embargo, en escalas espaciales y temporales muy pequeñas, es posible aproximar ciertos procesos moleculares (como la incorporación de un nucleótido en el sitio activo de la ADN polimerasa) mediante modelos que asumen condiciones cercanas al equilibrio local. En estos escenarios, el sistema puede describirse mediante el concepto de ensamble canónico, el cual representa un sistema que no intercambia partículas con el entorno y que se encuentra en equilibrio térmico con un reservorio de calor a temperatura constante. Bajo estas condiciones, la distribución de Boltzmann permite estimar la probabilidad de que el sistema adopte distintos estados energéticos.

Aplicando este marco teórico, Zimic y colaboradores realizaron simulaciones en las que se modelaba la generación de mutaciones puntuales a partir de la formación de pequeños “huecos” en la cadena de ADN, los cuales posteriormente eran ocupados por nucleótidos libres seguidos de un proceso de reparación potencialmente imperfecto. En este modelo, la ocupación de dichos huecos está determinada por las fuerzas intermoleculares entre el nucleótido entrante y el entorno molecular del sitio vacío. Posteriormente, el mismo grupo desarrolló un modelo más cercano al proceso real de replicación, simulando la probabilidad de

incorporación de nucleótidos libres en el sitio activo de la ADN polimerasa durante la síntesis de ADN. En este enfoque, la probabilidad de incorporación de cada nucleótido se modula a través de la distribución de Boltzmann aplicada al ensamble canónico, lo que permite evaluar cómo las diferencias en energía libre entre distintos estados de apareamiento influyen en el proceso replicativo. A este fenómeno, mediante el cual las propiedades termodinámicas del sistema pueden sesgar la probabilidad de incorporación de determinados nucleótidos, los autores lo denominaron presión mutacional termodinámica.

Como se ha mencionado previamente, la probabilidad de incorporación de nucleótidos libres en el sitio activo de la ADN polimerasa puede estimarse utilizando la distribución de Boltzmann. En este marco teórico, la probabilidad de que un nucleótido específico sea incorporado durante el proceso de replicación depende del estado energético del sistema molecular en la vecindad del sitio activo. Por ejemplo, la probabilidad de que una adenina sea incorporada está asociada a la probabilidad de que la región cercana al sitio de replicación adopte un estado energético correspondiente a E_A , que representa la energía del arreglo atómico-molecular en ese entorno cuando se incorpora una adenina. Esta relación puede expresarse de la siguiente manera:

$$P(X=A) = P(E_A) = \alpha [A] V \frac{1}{Z} e^{-\beta E_A - a N}$$

donde N representa el número de partículas que pueden intercambiarse en el sistema (en este caso nucleótidos trifosfato), considerado cercano a cero en este modelo. El parámetro β corresponde a $1/KT$, mientras que $a = \mu/KT$, donde μ es el

potencial químico del sistema. En esta expresión, V representa el volumen, K la constante de Boltzmann, T la temperatura absoluta y Z la función de partición.

Para facilitar el análisis, el modelo puede simplificarse calculando probabilidades relativas entre los distintos nucleótidos posibles. Esto se logra dividiendo cada probabilidad por la correspondiente a uno de los nucleótidos seleccionado como referencia. Si se elige arbitrariamente la probabilidad de incorporación de guanina como factor de normalización, la probabilidad relativa de incorporación de adenina puede expresarse como:

$$P'(X=A) = \frac{[A]}{[G]} e^{-\beta (E(X=A) - E(X=G))}$$

De esta forma, se observa que las probabilidades relativas dependen únicamente de la diferencia de energía entre dos configuraciones posibles del sistema. Estas diferencias energéticas están determinadas principalmente por las interacciones de puentes de hidrógeno y por las interacciones electrostáticas, las cuales varían según el nucleótido que ingresa al sitio activo de la polimerasa. Es importante señalar que las variaciones en la energía total de la molécula completa de ADN, ocasionadas por la incorporación de un nucleótido en el sitio activo, son comparables a las diferencias energéticas presentes en la vecindad inmediata del sitio de replicación. Esto se debe a que las interacciones intermoleculares disminuyen significativamente con la distancia, por lo que los nucleótidos ubicados lejos del sitio activo contribuyen poco al valor total de la energía del sistema. De hecho, los cálculos realizados mostraron que las interacciones electrostáticas dentro de una vecindad aproximada de cinco nucleótidos permiten estimar de manera adecuada las interacciones entre el nucleótido entrante y el entorno total de la

molécula de ADN. Esta región puede considerarse como una vecindad espacio-temporal donde el sistema alcanza una aproximación de equilibrio termodinámico local o cuasi-equilibrio, entendida como una condición en la que la temperatura, el volumen efectivo, la disponibilidad relativa de nucleótidos libres y el potencial químico del entorno permanecen aproximadamente constantes durante el evento de incorporación. En estas condiciones, las configuraciones moleculares alternativas pueden ser comparadas en función de sus diferencias energéticas y ponderadas mediante la distribución de Boltzmann, permitiendo estimar la probabilidad de incorporación de nucleótidos durante la replicación.

La aplicación de este modelo indica que la probabilidad de incorporación de un nucleótido depende directamente de la energía potencial del sistema, la cual incluye tanto la contribución de los puentes de hidrógeno como las interacciones electrostáticas. En este contexto, Zimic calculó las energías asociadas a los pares de bases complementarios de Watson-Crick (pares canónicos), así como aquellas correspondientes a pares no canónicos. Los resultados de estos cálculos mostraron que la formación de pares de bases no canónicos ocurre con una probabilidad extremadamente baja. No obstante, cuando se consideran grandes números de eventos replicativos, incluso estas probabilidades reducidas pueden generar desajustes de apareamiento (*mismatches*). En condiciones donde no existe presión de selección, tales mutaciones tienden a favorecer la incorporación de guanina (G) o citosina (C) en lugar de adenina (A) o timina (T) (Figura 15). Según las simulaciones computacionales realizadas (tanto en secuencias codificantes como no codificantes) este sesgo mutacional conduce a un aumento progresivo del contenido de GC en el genoma a lo largo del tiempo evolutivo. En el marco de la

hipótesis de la presión mutacional termodinámica, este fenómeno se interpreta como una consecuencia de la tendencia del sistema hacia configuraciones de menor energía potencial, es decir, hacia estados estructuralmente más estables de la molécula de ADN.

(eV)		(Kcal/mol)	
A-T =	-0.34	-7.84	
G-C =	-0.44	-10.14	
C-A =	0.21	4.84	
T-G =	0.13 (I)	-0.40 (II)	2.99 (I) -9.22 (II) **
T-C =	-0.14	-3.22	
A-G =	-0.27	-6.22 *	
T-T =	0.02 (I)	-0.16 (II)	0.46 (I) -3.68 (II) **
C-C =	-0.01 (I)	-0.14 (II)	-0.23 (I) -3.22 (II) **
A-A =	0.23	5.30 *	
G-G =	0.54	12.45 *	

Figura 15. Energías de interacción por puentes de hidrógeno correspondientes a pares de bases canónicos y no canónicos. La formación de los pares A-T y G-C presenta una estabilidad termodinámica favorable. En contraste, entre las combinaciones no canónicas solo los pares T-C y A-G muestran energías de interacción relativamente favorables. (*) (**) En estos casos, la doble hélice de ADN tendría que adoptar configuraciones energéticas poco probables o estructuralmente desfavorables para que dichas interacciones se estabilicen.

Con el objetivo de evaluar la hipótesis de la presión mutacional termodinámica, las predicciones derivadas de las simulaciones computacionales fueron comparadas con información evolutiva obtenida a partir de secuencias genómicas de distintas especies pertenecientes al linaje Kinetoplastidia. Este grupo de organismos constituye un modelo adecuado para examinar la hipótesis debido a que muchas de sus especies son parásitos intracelulares, lo que implica que experimentan presiones selectivas relativamente limitadas y estables. En el árbol filogenético construido a partir del gen del ARN ribosomal 18S de diversos kinetoplástidos (incluyendo

tripanosomátidos, *Leishmania*, *Crithidia* y otros organismos relacionados) se ha propuesto una cascada evolutiva (Figura 16) en la cual se observa una tendencia progresiva hacia el incremento del contenido de GC en las secuencias codificantes analizadas.

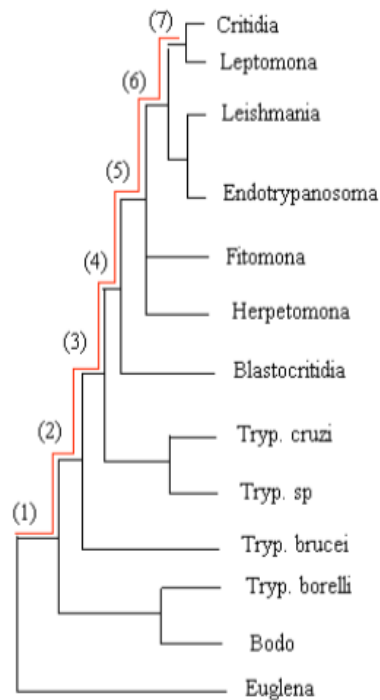


Figura 16. La cascada evolutiva representada mediante doble línea está compuesta por una serie de ancestros comunes hipotéticos actualmente extintos (1), (2), (3)...(7). A pesar de que estos ancestros ya no existen, es posible identificar organismos actuales que se encuentran evolutivamente más próximos a cada uno de estos nodos ancestrales. Por ejemplo, *Bodo* o *Trypanosoma borelli* se consideran los organismos más cercanos al ancestro (1); *T. brucei* se aproxima al nodo (2); *T. cruzi* o *Trypanosoma sp.* al nodo (3); y así sucesivamente hasta el nodo (7), cuyo representante evolutivo más cercano corresponde a *Crithidia* o *Leptomonas*.

El análisis de la variación del contenido de GC a lo largo de esta trayectoria evolutiva revela que las bases G y C se acumulan en las tres posiciones del codón, aunque este incremento es particularmente pronunciado en la tercera posición

codónica. Esta tendencia se explica porque dicha posición presenta mayor degeneración en el código genético y, por lo tanto, está sometida a una presión selectiva menor que las dos primeras posiciones. El incremento del contenido de GC tanto en el genoma completo como en las posiciones codificantes sugiere además la presencia de un sesgo en el uso de codones (codon bias). Este fenómeno surge porque la única forma de aumentar el contenido genómico de GC sin modificar la secuencia de aminoácidos consiste en reemplazar codones por otros equivalentes que contengan una mayor proporción de G y C. Un ejemplo de este patrón se observa en la distribución de codones que codifican para leucina y serina en especies como *T. brucei*, *T. cruzi*, *Leishmania* y *Crithidia*, donde se aprecia una tendencia evolutiva hacia la utilización de codones con mayor contenido GC. En conjunto, estos datos genómicos respaldan la idea de que la presión mutacional termodinámica podría influir en los patrones evolutivos observados, tanto en la composición de bases como en las preferencias de uso de codones.

ANEXO 2: Pipeline de alineamiento de lecturas Nanopore

```
#!/bin/bash
#####
# Descripción:
# Este script procesa lecturas de secuenciación Oxford Nanopore en formato
# FASTQ. Para cada muestra:
# 1) alinea las lecturas contra una secuencia de referencia,
# 2) genera un archivo BAM ordenado,
# 3) filtra alineamientos primarios de alta calidad,
# 4) indexa el BAM final para su análisis posterior.
#
# Programas requeridos:
# - minimap2
# - samtools
#
# Entradas:
# - ref.fasta
# - pass_barcode*.fastq
#
# Salidas:
# - <sample>_gene.bam
# - <sample>_gene.bam.bai
#####

# Parámetros principales del análisis
REF="ref.fasta"      # archivo de referencia
THREADS=8            # número de hilos de procesamiento
MIN_MAPQ=20          # calidad mínima de mapeo aceptada

# Verificar que exista la referencia
[ -f "$REF" ] || { echo "ERROR: no se encuentra $REF"; exit 1; }

# Procesar cada archivo FASTQ
for fq in pass_barcode*.fastq; do
    [ -e "$fq" ] || { echo "ERROR: no se encontraron archivos FASTQ"; exit 1; }

    sample="${fq%.fastq}"

    # 1. Alineamiento de lecturas Nanopore contra la referencia
    # minimap2 genera los alineamientos y samtools los ordena en formato BAM
    minimap2 -t "$THREADS" -ax map-ont "$REF" "$fq" \
        | samtools sort -@ "$THREADS" -o "${sample}.bam"

    # 2. Filtrado de alineamientos de alta calidad
    # Se conservan únicamente:
    # - lecturas mapeadas
    # - alineamientos primarios
    # - calidad de mapeo >= 20
    samtools view -b -F 0x4 -F 0x900 -q "$MIN_MAPQ" \
        "${sample}.bam" > "${sample}_gene.bam"

    # 3. Indexación del BAM filtrado
    # Permite acceder rápidamente al archivo en análisis posteriores
    samtools index "${sample}_gene.bam"
done
```

ANEXO 3: Pipeline de filtrado de alineamientos según la longitud cubierta

```
#!/usr/bin/env bash
#####
# Descripción:
# Este script filtra archivos BAM para conservar únicamente alineamientos
# cuya longitud cubierta sobre la referencia, calculada a partir del CIGAR,
# se encuentre entre 250 y 500 pb.
#
# Requisitos:
# samtools
# awk
#
# Entradas:
# *.bam
#
# Salidas:
# <sample>_fil250.bam
#####

# Parámetros del filtrado
THREADS=8
MINLEN=250
MAXLEN=500

# Procesar cada archivo BAM del directorio
for bam in *.bam; do
    base="${bam%.bam}"
    out="${base}_fil250.bam"

    # Extraer alineamientos con encabezado y filtrar por longitud cubierta
    samtools view -@ "$THREADS" -h "$bam" \
    | awk -v MIN="$MINLEN" -v MAX="$MAXLEN" '
    function ref_span(cig, n, op, rest, span, token) {
        span = 0
        rest = cig
        while (match(rest, /^[0-9]+[A-Z=]/)) {
            token = substr(rest, RSTART, RLENGTH)
            n = substr(token, 1, length(token)-1) + 0
            op = substr(token, length(token), 1)
            if (op=="M" || op=="=" || op=="X" || op=="D" || op=="N") span += n
            rest = substr(rest, RLENGTH + 1)
        }
        return span
    }
    /^@/ { print; next }
    { s = ref_span($6); if (s >= MIN && s <= MAX) print }
    ' \
    | samtools view -@ "$THREADS" -b -o "$out" -

    # Indexar el BAM filtrado
    samtools index -@ "$THREADS" "$out"
done
```

ANEXO 4: Pipeline para el cálculo de frecuencias nucleotídicas por posición según la orientación de las lecturas (forward y reverse)

```
#!/usr/bin/env bash
#####
# Descripción:
#   Calcula frecuencias de nucleótidos (A, C, G y T) por posición de la
#   referencia a partir de archivos BAM filtrados, separando los conteos
#   según la orientación de las lecturas (forward y reverse).
# Entradas:
#   ref.fasta
#   pass_barcode*_gene_fil250.bam
# Salida:
#   basefreq_strand_longQ20.tsv
#####

set -euo pipefail

REF="ref.fasta"
OUT="basefreq_strand_longQ20.tsv"
MIN_MAPQ=20
MIN_BASEQ=20
MAX_DEPTH=100000

[ -f "$REF" ] || { echo "ERROR: no encuentro referencia $REF"; exit 1; }
[ -f "$REF.fai" ] || samtools faidx "$REF"

echo -e "sample\tpos\tref_base\tA_fwd\tC_fwd\tG_fwd\tT_fwd\tA_rev\tC_rev\tG_rev\tT_rev\tdp\tfA\tfC\tfG\tfT" > "$OUT"

# Genera mpileup y cuenta A/C/G/T por posición
count_bases() {
    local bam="$1"
    samtools mpileup -aa -d "$MAX_DEPTH" -q "$MIN_MAPQ" -Q "$MIN_BASEQ" -f "$REF" "$bam" \
    | awk 'BEGIN{OFS="\t"}
    function clean(b, i,c,n,num,rest){
        gsub(/\^./,"",b); gsub(/\$/,"",b)
        while (match(b, /[+-][0-9]+/)) {
            rest=substr(b,RSTART+1); num=""
            for(i=1;i<=length(rest);i++){ c=substr(rest,i,1); if(c~/[0-9]/) num=num c; else break }
            n=num+0
            b=substr(b,1,RSTART-1) substr(b,RSTART+1+length(num)+n)
        }
        gsub(/\*/,"",b)
        return b
    }
    {
        pos=$2; ref=toupper($3); bases=clean($5)
        A=C=G=T=0
        for(i=1;i<=length(bases);i++){
            c=substr(bases,i,1)
            if(c=="." || c=="") c=ref
            c=toupper(c)
            if(c=="A") A++; else if(c=="C") C++; else if(c=="G") G++; else if(c=="T") T++
        }
        print pos,ref,A,C,G,T
    }
}'

for BAM in pass_barcode{01..12}_gene_fil250.bam; do
    [ -e "$BAM" ] || continue
    sample="${BAM%.bam}"

    # Conteos forward y reverse por separado
    count_bases <(samtools view -h -F 16 "$BAM" | samtools view -b -) | sort -k1,1n > "${sample}.fwd.tmp"
    count_bases <(samtools view -h -f 16 "$BAM" | samtools view -b -) | sort -k1,1n > "${sample}.rev.tmp"

    # Integración de conteos y cálculo de cobertura total y frecuencias
    join -t '$\t' -a1 -a2 -e 0 \
    -o '0,1.2,1.3,1.4,1.5,1.6,2.3,2.4,2.5,2.6' \
    "${sample}.fwd.tmp" "${sample}.rev.tmp" \
    | awk -v OFS="\t" -v S="$sample" '
    {
        pos=$1; ref=$2
        Af=$3+0; Cf=$4+0; Gf=$5+0; Tf=$6+0
        Ar=$7+0; Cr=$8+0; Gr=$9+0; Tr=$10+0
        A=Af+Ar; C=Cf+Cr; G=Gf+Gr; T=Tf+Tr
        DP=A+C+G+T
        fA=(DP>0?A/DP:0); fC=(DP>0?C/DP:0); fG=(DP>0?G/DP:0); fT=(DP>0?T/DP:0)
        print S,pos,ref,Af,Cf,Gf,Tf,Ar,Cr,Gr,Tr,DP,fA,fC,fG,fT
    }' >> "$OUT"

    rm -f "${sample}.fwd.tmp" "${sample}.rev.tmp"
done
```

ANEXO 5: Pipeline para la identificación de bases alternativas dominantes con soporte en ambas cadenas

```
#!/usr/bin/env bash
#####
# Descripción:
#   Identifica la base alternativa dominante (ALT) por posición a partir
#   de una tabla de frecuencias nucleotídicas separadas por cadena.
#   El script filtra por profundidad mínima, frecuencia alélica mínima
#   y soporte en ambas orientaciones.
#
# Entrada:
#   basefreq_strand_longQ20.tsv
#
# Salida:
#   alt_dominante_strand_filteredQ20.tsv
#####

set -euo pipefail

IN="basefreq_strand_longQ20.tsv"
OUT="alt_dominante_strand_filteredQ20.tsv"

MIN_DP=50000          # profundidad mínima total por posición
MIN_AF=0.0005         # frecuencia alélica mínima de la base alternativa
MIN_MINOR_STRAND_FRAC=0.2 # fracción mínima de soporte en la cadena minoritaria

echo -e "sample\tpos\tref_base\talt_base\tDP\talt_count\tAF_alt\talt_fwd\talt_rev\tminor_frac" > "$OUT"

awk -F'\t' -v OFS="\t" \
-v MIN_DP="$MIN_DP" -v MIN_AF="$MIN_AF" -v MIN_MINOR_STRAND_FRAC="$MIN_MINOR_STRAND_FRAC" '
NR==1 { next } # omitir encabezado
{
    sample=$1; pos=$2; ref=toupper($3)
    Af=$4+0; Cf=$5+0; Gf=$6+0; Tf=$7+0
    Ar=$8+0; Cr=$9+0; Gr=$10+0; Tr=$11+0
    DP=$12+0

    if (DP < MIN_DP) next

    # Conteos totales por base
    A=Af+Ar; C=Cf+Cr; G=Gf+Gr; T=Tf+Tr

    # Seleccionar la base no referencial con mayor conteo
    alt=""; altc=-1; altf=0; altr=0
    if (ref!="A" && A>altc) {alt="A"; altc=A; altf=Af; altr=Ar}
    if (ref!="C" && C>altc) {alt="C"; altc=C; altf=Cf; altr=Cr}
    if (ref!="G" && G>altc) {alt="G"; altc=G; altf=Gf; altr=Gr}
    if (ref!="T" && T>altc) {alt="T"; altc=T; altf=Tf; altr=Tr}

    if (alt==" " || altc==0) next

    # Calcular frecuencia alélica y soporte en ambas cadenas
    AF = altc / DP
    if (AF < MIN_AF) next

    minor = (altf < altr ? altf : altr)
    minor_frac = minor / altc
    if (minor_frac < MIN_MINOR_STRAND_FRAC) next

    printf "%s\t%s\t%s\t%s\t%d\t%.10f\t%.6f\t%.6f\n", \
        sample,pos,ref,alt,DP,altc,AF,altf,altr,minor_frac
}
' "$IN" >> "$OUT"
```

ANEXO 6: Pipeline para la identificación de mutaciones mediante persistencia temporal de variantes

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
#####
# Descripción:
#   Clasifica mutaciones según su persistencia temporal en una ventana de
#   generaciones. Para cada mutación se evalúa:
#   - aparición en la ventana analizada,
#   - persistencia hasta la generación final,
#   - crecimiento neto de AF,
#   - estabilidad temporal.
# Entrada:
#   alt_dominante_strand_filteredQ20.tsv
# Salida:
#   mutaciones_clasificadasQ2011.tsv
#####
import pandas as pd
import numpy as np
INFILE = "alt_dominante_strand_filteredQ20.tsv"
OUTFILE = "mutaciones_clasificadasQ2011.tsv"
GEN_MIN = 1
GEN_MAX = 11
MIN_AF_START = 0.0005
ALLOW_GAP = 0
MIN_DELTA_AF = 0.0003
ALLOW_DROP = 0.4
MIN_PRESENCE_LEN = 2

# Cargar datos y extraer generación desde el nombre de muestra
df = pd.read_csv(INFILE, sep="\t", dtype=str)
df["gen"] = pd.to_numeric(df["sample"].str.extract(r"barcode(\d+)")[0], errors="coerce")
df = df.dropna(subset=["gen"]).copy()
df["gen"] = df["gen"].astype(int)
df = df[(df["gen"] >= GEN_MIN) & (df["gen"] <= GEN_MAX)].copy()

# Convertir AF y construir identificador de mutación
df["AF_alt"] = pd.to_numeric(df["AF_alt"].str.replace(",", ".", regex=False), errors="coerce")
df = df.dropna(subset=["AF_alt", "pos", "ref_base", "alt_base"]).copy()
df["ref_base"] = df["ref_base"].str.upper().str.strip()
df["alt_base"] = df["alt_base"].str.upper().str.strip()
df["mut_id"] = df["pos"].astype(str) + "_" + df["ref_base"] + ">" + df["alt_base"]
gens_window = np.arange(GEN_MIN, GEN_MAX + 1, dtype=int)
results = []
for mut, d in df.groupby("mut_id", sort=False):

    # Serie temporal completa de AF; generaciones ausentes se rellenan con 0
    g2af = d.groupby("gen")["AF_alt"].max()
    af_series = np.array([float(g2af.get(g, 0.0)) for g in gens_window], dtype=float)
    present = af_series >= MIN_AF_START

    if not present.any():
        continue

    # Subserie desde la primera aparición hasta el final
    first_idx = int(np.argmax(present))
    gen_inicio = int(gens_window[first_idx])
    af_tail = af_series[first_idx:]
    present_tail = present[first_idx:]

    AF_inicio = float(af_tail[0])
    AF_final = float(af_tail[-1])
    delta_AF = float(AF_final - AF_inicio)

    gaps_observados = int((~present_tail).sum())
    diffs = np.diff(af_tail)
    n_drops = int(np.sum(diffs < 0))
    drop_frac = n_drops / max(len(diffs), 1)
    tail_len = len(af_tail)
```

```

# Clasificación según persistencia, crecimiento y estabilidad
status = "PROBABLE_ERROR"
motivo = ""

if tail_len < MIN_PRESENCE_LEN:
    motivo = "APARECE_MUY_TARDE"
elif gaps_observados > ALLOW_GAP:
    motivo = "SE_PIERDE_EN_EL_TAIL"
elif delta_AF < MIN_DELTA_AF:
    motivo = "NO_CRECE"
elif drop_frac > ALLOW_DROP:
    motivo = "DEMASIADAS_BAJADAS"
else:
    status = "SEÑAL_BIOLOGICA"
    motivo = "CUMPLE_PERSISTENCIA_Y_CRECE"
results.append({
    "mut_id": mut,
    "pos": d["pos"].iloc[0],
    "ref": d["ref_base"].iloc[0],
    "alt": d["alt_base"].iloc[0],
    "gen_inicio": gen_inicio,
    "gen_final": GEN_MAX,
    "AF_inicio": AF_inicio,
    "AF_final": AF_final,
    "delta_AF": delta_AF,
    "n_gens_desde_aparicion": tail_len,
    "gaps_observados": gaps_observados,
    "drop_frac": drop_frac,
    "n_drops": n_drops,
    "status": status,
    "motivo": motivo
})
out = pd.DataFrame(results)
out.sort_values(["status", "gen_inicio", "delta_AF", "AF_final"],
                ascending=[True, True, False, False], inplace=True)
out.to_csv(OUTFILE, sep="\t", index=False)
print(f"Clasificación completada: {OUTFILE}")

```