
# Proyecto Final Curso de Bioinformática y Biología Computacional Código DRBC A001

INFORME FINAL

## DOCTORADO EN BIOLOGÍA COMPUTACIONAL-2025

**Análisis Bioinformático de la Fiebre de Norilsk: Desde la Identificación del Patógeno hasta la Propuesta de Blancos Terapéuticos**

**Presentado por:**

**Pamela Fernández y Margot Paco**

# Contenido

Introducción ..... 2
Objetivos ..... 3
Objetivo General: ..... 3
Objetivos Específicos: ..... 3
Materiales y Métodos ..... 3
Análisis epidemiológico ..... 3
RNA-seq ..... 3
Anotación de secuencias e identificación del agente patógeno ..... 3
Análisis de blancos terapéuticos ..... 4
Red de interacción génica ..... 4
Resultados ..... 5
Análisis epidemiológico ..... 5
RNA-seq ..... 7
Anotación de secuencias e identificación del agente patógeno ..... 8
Análisis de blancos terapéuticos ..... 9
Red de interacción génica ..... 10
Discusión ..... 11
Referencias Bibliográficas ..... 12

# Introducción

En la era actual de la genómica, el volumen y complejidad de los datos biológicos ha superado las capacidades de los métodos tradicionales de análisis. Las herramientas bioinformáticas se han convertido en pilares esenciales para interpretar, visualizar y extraer conocimiento significativo de secuencias genéticas. Estas plataformas permiten desde la comparación y alineamiento de genes hasta la predicción funcional, la reconstrucción filogenética y la exploración de redes biológicas (Tamura, Stecher & Kumar, 2021).

Entre las más destacadas se encuentra BLASTn, una herramienta fundamental para identificar similitudes entre secuencias nucleotídicas, utilizada ampliamente en estudios de anotación funcional y diagnóstico molecular (Zhang et al., 2020). ClustalW, por su parte, sigue siendo una opción robusta para el alineamiento múltiple de secuencias, permitiendo inferencias evolutivas y análisis de conservación entre especies (Sharma & Dubey, 2019). MEGA11 (Molecular Evolutionary Genetics Analysis) complementa este proceso al ofrecer métodos avanzados para construir árboles filogenéticos, estimar tiempos de divergencia y realizar análisis estadísticos de evolución molecular (Tamura et al., 2021). Finalmente, Cytoscape, en su versión más reciente, se posiciona como una plataforma clave para la visualización de redes de interacción genética y funcional, integrando datos de expresión, anotación y conectividad biológica en entornos de análisis multiómico (Omics Tutorials, 2024).

Estas herramientas no solo optimizan la precisión y eficiencia del análisis genómico, sino que también democratizan el acceso al conocimiento, permitiendo que investigadores de diversas disciplinas puedan explorar la arquitectura genética con profundidad y rigor científico.

Gracias a estas herramientas, ha podido presentar el análisis del siguiente caso estudio:

***"En agosto 2020, a un brote infeccioso altamente letal surgido en Norilsk, Siberia, despertó la atención internacional debido a su rápida propagación, sintomatología severa y posible relación con el deshielo del permafrost. Esta enfermedad, conocida como fiebre de Norilsk, se caracteriza por una progresión rápida a etapas críticas, afectando múltiples sistemas del organismo y generando una mortalidad estimada superior al 60%. Su transmisión aparentemente indirecta, junto con una ventana de incubación asintomática, ha favorecido su diseminación internacional, generando una alerta sanitaria global por parte de la Organización Mundial de la Salud."***

Ante la urgencia de comprender el comportamiento del brote y su agente causal, se planteó un enfoque bioinformático integral. Este proyecto combinó análisis epidemiológicos con técnicas de secuenciación de alto rendimiento (RNA-seq), bioinformática estructural, anotación genómica y modelado de redes de interacción. Los datos transcriptómicos provenientes de pacientes infectados fueron procesados para identificar genes diferencialmente expresados, visualizados mediante *volcano plots* y clasificados funcionalmente con ontologías génicas (Gene Ontology).

Paralelamente, se analizaron trece secuencias en formato FASTA mediante búsquedas BLAST y alineamientos múltiples con CLUSTAL Omega, lo que permitió inferir la identidad del agente patógeno. Una vez identificado, se procedió a la anotación de su genoma y a la selección de posibles blancos terapéuticos, los cuales fueron modelados estructuralmente y analizados por técnicas de acoplamiento molecular (*docking*). Finalmente, se infirió una red de interacción génica basada en los genes humanos sobreexpresados, evaluando su estructura mediante métricas como coeficiente de clustering, grado promedio y geodésica.

Este enfoque multidisciplinario permitió abordar la crisis desde distintas escalas (molecular, celular y poblacional), proporcionando herramientas clave para la identificación de estrategias terapéuticas y el entendimiento profundo de las bases genéticas de la enfermedad.

# Objetivos

## Objetivo General:

Investigar el brote infeccioso de la fiebre de Norilsk mediante el uso de técnicas de bioinformática para identificar el agente patógeno, analizar la respuesta génica del huésped y proponer blancos terapéuticos.

## Objetivos Específicos:

- Estimar el número reproductivo básico ($R_0$) y proyectar la evolución del brote hasta la semana 10 en distintas ciudades.
- Analizar los datos de RNA-seq de pacientes infectados para identificar genes sobreexpresados, visualizarlos mediante *volcano plots* y anotarlos funcionalmente.
- Anotar y comparar secuencias virales entregadas utilizando BLAST y CLUSTAL, e identificar el agente patógeno causante del brote.
- Descargar, analizar y anotar el genoma completo del patógeno, identificando genes esenciales como blancos terapéuticos.
- Modelar las estructuras 3D de los blancos, identificar sitios activos y evaluar su interacción con posibles fármacos mediante *docking*.
- Inferir una red de interacción génica a partir de genes humanos sobreexpresados y analizar sus propiedades topológicas.

# Materiales y Métodos

Para el desarrollo del proyecto se emplearon herramientas bioinformáticas ampliamente utilizadas en el análisis de datos ómicos.

## Análisis epidemiológico

Se estimó el $R_0$ con base en los datos de propagación durante las tres primeras semanas del brote. Luego, se proyectó el número de infectados y fallecidos hasta la semana 10 utilizando modelos SIRD ajustados. Tomando los datos de la Tabla, se decidió realizar el modelo SIRD considerando lo siguiente:

a) Que la duración promedio de la enfermedad son 15 días
b) Que el CFR (según texto) está entre el 30 y 60%
c) Se incluyó una calibración de datos de tal manera que los datos reales fueran incorporados en la simulación
d) Se optimizaron los parámetros de beta y gama.

## RNA-seq

Para el análisis de expresión génica diferencial, se utilizó el lenguaje R (versión 4.5.0) junto con las librerías readr, dplyr, ggrepel, biomaRT y ggplot2. (Ito & Murphy, 2013) Los datos fueron importados desde un archivo CSV, se tomaron los valores de $\log_2$ *fold change* ($\log_2$FC) y los *p-value* ajustados. A fin de facilitar

la interpretación visual de la significancia estadística, se calculó una nueva variable correspondiente a $-\log_{10}$(p-valor ajustado). Posteriormente, los genes fueron clasificados en tres grupos según los siguientes criterios: genes sobreexpresados ($\log_2$FC > 2 y p-ajustado < 0.01), genes subexpresados ($\log_2$FC < –2 y p-ajustado < 0.01), y genes no significativos, que no cumplían con estos umbrales.

Posteriormente, se realizó un análisis de enriquecimiento funcional GO con los genes sobreexpresados obtenidos en el paso anterior usando el paquete *clusterProfiler* en R y anotaciones de *org.Hs.eg.db*.(Wu et al., 2021) Los IDs ENSEMBL se convirtieron a ENTREZ con bitr, y se aplicó enrichGO considerando las categorías BP, CC y MF. Se usó el método de Benjamini-Hochberg (Haynes, 2013) para controlar el FDR (pvalueCutoff < 0.05). Los resultados se exportaron como .csv y se visualizaron con un dotplot y cnetplot.

## Anotación de secuencias e identificación del agente patógeno

Para realizar el alineamiento múltiple de secuencias, se utilizó el algoritmo ClustalW con parámetros ajustados para optimizar la calidad del alineamiento. Se aplicó una penalización de apertura de huecos de 15.00 y una penalización de extensión de 6.66, lo que permite controlar la introducción y prolongación de gaps en las secuencias alineadas, favoreciendo alineamientos más conservadores. La matriz de sustitución seleccionada fue ClustalW (1.4), adecuada para secuencias de ADN, junto con un peso de transición de 0.50, que otorga menor penalización a las sustituciones entre bases químicamente similares (A↔G, C↔T). Además, se estableció un umbral de divergencia del 30%, lo que permite retrasar el alineamiento de secuencias altamente divergentes para evitar distorsiones en las etapas iniciales. Estos parámetros fueron seleccionados para maximizar la precisión del alineamiento y facilitar la posterior interpretación funcional y evolutiva de las secuencias analizadas. El Alineamiento Múltiple se realizó utilizando el Programa Mega.

Para la búsqueda BLASTn se utilizaron parámetros optimizados para secuencias cortas, activando la opción “Short queries” para ajustar automáticamente la sensibilidad del algoritmo. Se estableció un umbral de expectativa **(**Expect threshold**)** de 0.05, lo que indica que solo se reportarán alineamientos con una probabilidad muy baja de ocurrir por azar, aumentando la confiabilidad de los resultados. El tamaño de palabra **(**Word size**)** se configuró en 28, lo cual es adecuado para búsquedas con Megablast, ya que favorece la detección de secuencias altamente similares. El número máximo de secuencias objetivo **(**Max target sequences**)** se fijó en 100, permitiendo una exploración amplia de coincidencias en la base de datos. En cuanto a los parámetros de puntuación, se asignó una puntuación de 1.2 para coincidencias y se utilizó un modelo de penalización lineal para los huecos **(**Gap Costs**),** lo que simplifica el cálculo de alineamientos. Se aplicaron filtros para excluir regiones de baja complejidad, así como repeticiones específicas de *Homo sapiens*, con el fin de evitar alineamientos espurios. Además, se activó la opción “Mask for lookup table only”, lo que permite enmascarar regiones repetitivas solo durante la búsqueda inicial, sin afectar el alineamiento final. Estos parámetros en conjunto aseguran una búsqueda precisa y específica, adecuada para la identificación de homología en secuencias humanas altamente conservadas.

## Análisis de blancos terapéuticos

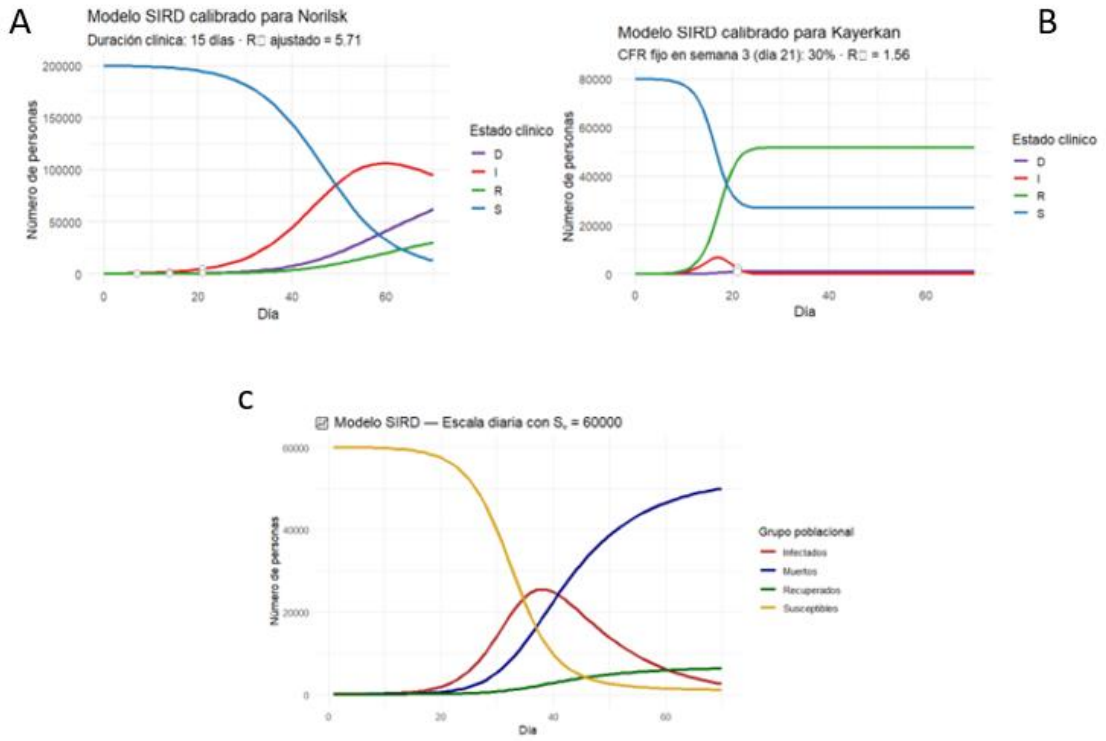
Se seleccionaron las proteínas neuraminidasa (NA) y NS1 como blancos terapéuticos, cuyos modelos tridimensionales se generaron mediante SWISS-MODEL (Biasini et al., 2014) para garantizar estructuras de alta calidad y sin mutaciones. Los sitios activos se identificaron con PockDrug, (Hussein et al., 2015) priorizando los pockets con mayor probabilidad de drugabilidad. Para la NA se utilizó el sitio activo del fármaco como referencia. Para el análisis de potenciales drogas, una base de datos de compuestos naturales peruanos (PeruNPDB) (Barazorda-Ccahuana et al., 2023) fue filtrada en DataWarrior (López-López et al., 2019) usando los criterios de *druglikeness*, toxicidad (mutagénico, tumorigénico, efecto reproductivo, irritante) y patrones PAINS. Para su evaluación se utilizó el método de cribado virtual con PyRx,(Dallakyan & Olson, 2015) evaluando afinidades de unión mediante acoplamiento molecular con 9 repeticiones para cada ligando.

## Red de interacción génica

Se utilizaron herramientas como STRINGDB (Szklarczyk et al., 2025) para construir redes de interacción con los genes humanos sobreexpresados. Las redes fueron visualizadas en Cytoscape, (Kohl et al., 2011) y se calcularon propiedades con la extensión Analyzer Network, para obtener la anotación de los genes que conforman la geodésica se utilizó Python.

# Resultados

## Análisis epidemiológico




**Las Figuras 1. muestran la proyección epidemiológica a las 10 semanas realizando una simulación en base al modelo SIRD.** La Figura 1A muestra las curvas dinámicas de las categorías S (susceptibles), I (infectados), R (recuperados) y D (fallecidos), simuladas con una duración clínica de 15 días y un valor de $R_0$ ajustado de 5.71. La simulación incorpora puntos de calibración empíricos en los días 14 y 22, utilizados para validar la coherencia del modelo con datos clínicos reales. Las tasas de recuperación y letalidad son producto del ajuste paramétrico que preserva la consistencia con observaciones poblacionales. La Figura 1B, Curvas simuladas para las categorías S (susceptibles), I (infectados), R (recuperados) y D (fallecidos) bajo una duración clínica de 15 días. El modelo fue calibrado con un $R_0$ de 1.56 y una tasa de letalidad (CFR) ajustada al 30% en la semana 3 (día 21). La dinámica reflejada muestra una progresión contenida de la infección, con estabilización temprana de las tasas de recuperación y mortalidad. Los datos empíricos integrados permiten evaluar el comportamiento poblacional bajo condiciones de transmisión moderada. Y finalmente, la Figura 1C muestra la representación temporal del modelo SIRD aplicado a una población inicial de 60,000 personas, con escala diaria. Las curvas muestran la dinámica de los grupos poblacionales: Susceptibles (amarillo), Infectados (rojo), Recuperados (verde) y Fallecidos (azul), conforme avanza la transmisión de la enfermedad. El modelo refleja una progresión continua en las tasas de infección y recuperación, útil para interpretar

escenarios de transmisión en tiempo real y calibrar medidas sanitarias en función de los desplazamientos poblacionales. El Ro Calculado fue de 4.03.

El valor de $R_0$ (número básico de reproducción) representa cuántas personas, en promedio, puede contagiar un individuo infectado en una población completamente susceptible. Este parámetro influye directamente en la velocidad de propagación y en la magnitud del brote. Para estimar su impacto, se utilizó el modelo SIRD (Susceptible-Infectado-Recuperado-Fallecido), ampliamente reconocido por su capacidad de representar dinámicas poblacionales en contextos epidémicos (Kermack & McKendrick, 1927).

Al analizar los resultados, se observan comportamientos contrastantes entre las tres ciudades simuladas:

- Norilsk presenta un $R_0$ de 5.71, lo que indica una alta transmisibilidad. La curva epidémica muestra un crecimiento exponencial acelerado, con más de 94,000 infectados en 70 días. Aunque la tasa de letalidad es del 39 %, la proporción recuperada sugiere cierta capacidad de compensación sanitaria. Este tipo de brote podría generar una rápida saturación hospitalaria y requerir intervenciones urgentes.
- En Kayerkan, el $R_0$ estimado fue de 1.56, lo que refleja una transmisión controlada, cercana al umbral epidémico. El número de infectados y fallecidos se mantiene contenido, incluso al extrapolar al día 70. Esto sugiere una buena capacidad de respuesta clínica y puede servir como referencia para estrategias preventivas efectivas.
- Por otro lado, Dudinka muestra un $R_0$ de 4.03, lo que indica una transmisión significativa. Sin embargo, el bajo porcentaje de recuperados frente a una letalidad del 95.3 % sugiere un colapso funcional o ineficiencia en los sistemas de mitigación, posiblemente derivado de supuestos clínicos extremos o baja capacidad de respuesta.

El modelo SIRD permite simular estas dinámicas con base en parámetros clínicos como duración de la enfermedad, tasas de recuperación y mortalidad, y se ha consolidado como una herramienta útil para evaluar escenarios de intervención y planificación sanitaria (Brauer et al., 2019). Cabe señalar que esta simulación también se realizó usando el modelo SIR, sin embargo, encontramos que el modelo no se ajustaba con la semana 2 y 3, y por tanto la proyección a la semana 10 daba resultados incoherentes, por esta razón nos quedamos con el modelo SIRD.

**Observación Kayerkan:** Visualizando la simulación de Kayerkan, nos podemos dar cuenta que, en el caso de los datos teóricos de muertos, se ajusta los datos de la semana 1 y 2, sin embargo, el dato de muertos a la semana 3 no logra ajustarse a pesar de que fue insertado en el script Rstudio (ver Figura 2).

```
# ──────────────────────────────────────────────────────────
# 🔬 Datos empíricos (Norilsk – semana 3)
S0 <- 80000
I0 <- 10
R0_ini <- 0
D0 <- 0
dias <- 70
duracion_enfermedad <- 15

# Datos observados específicos para calibración
observado <- data.frame(
  dia = c(21),
  Infectados = c(2300),
  Muertos = c(1000)
)
```

**Figura 2. Captura de pantalla del script Rstudio de los datos ingresados para la ciudad de Kayerkan.**

## RNA-seq

El análisis de expresión génica permitió identificar un conjunto de genes significativamente regulados entre las condiciones comparadas. Aplicando un umbral de significancia de $p < 0.01$ y $\log_2$Fold Change > 2 (sobreexpresión) o < –2 (subexpresión), se detectaron varios genes con cambios relevantes en sus niveles de expresión. Entre los genes sobreexpresados destaca JUN, mientras que entre los subexpresados resaltan UCP2, ALDOC, GFRA1, EPHX2, RCSD1, C21orf58 y MOCS1. Previamente se analizó con un corte $\log_2$Fold Change > 1 y > -1 (revisar material suplementario).

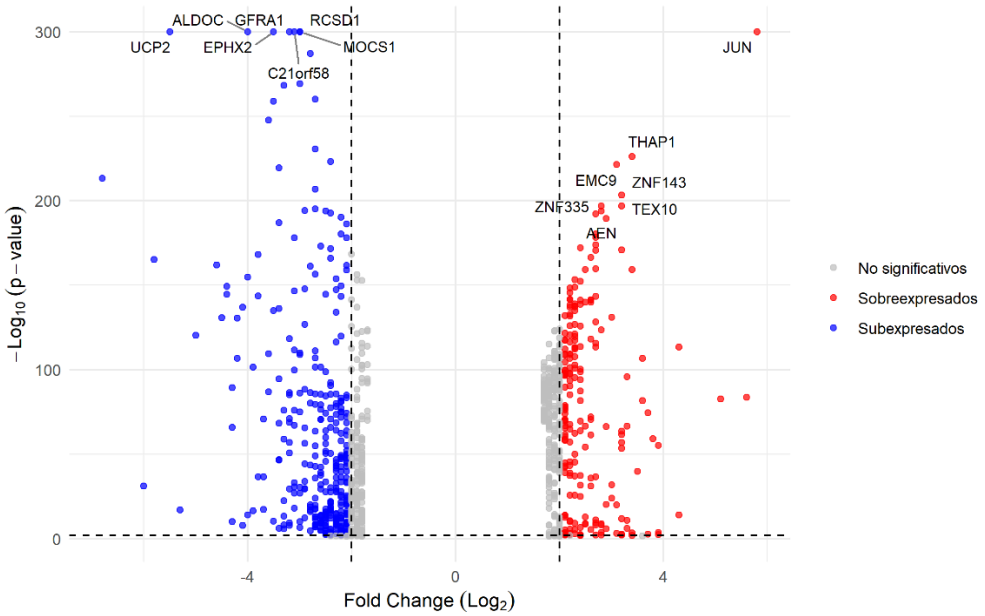



**Figura 3. Volcano plot que representa los genes diferencialmente expresados obtenidos del análisis de RNA-seq.** El eje X muestra el $\log_2$ del cambio en los niveles de expresión ($\log_2$Fold Change), y el eje Y representa la significancia estadística ($-\log_{10}$p). Los genes sobreexpresados con significancia estadística $p < 0.01$ y $\log_2$FC > 2 se muestran en rojo, mientras que los subexpresados con $\log_2$FC < –2 en azul. Los genes no significativos se representan en gris.

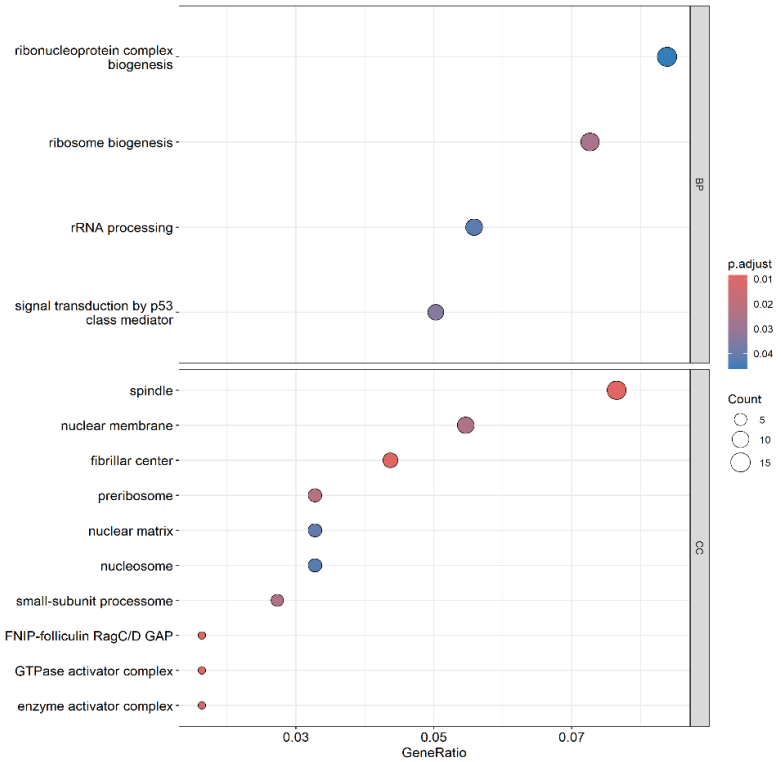



**Figura 4. Dotplot del análisis de enriquecimiento funcional GO a partir de genes sobreexpresados.** Se muestran los términos significativamente enriquecidos en las categorías de procesos biológicos (BP) y componentes celulares (CC). El tamaño de los puntos representa la cantidad de genes asociados a cada término GO, mientras que el color indica el nivel de significancia (p.adjust).

El análisis de enriquecimiento funcional GO mostró términos significativamente representados entre los genes sobreexpresados, principalmente en las categorías de procesos biológicos (BP) y componentes celulares (CC). En BP, destacaron “ribonucleoprotein complex biogenesis”, “ribosome biogenesis” y “rRNA processing”, con altos valores de GeneRatio y baja significancia ajustada (p.adjust < 0.05). También se identificó “signal transduction by p53 class mediator” como término relevante. En CC, los términos más enriquecidos fueron “spindle”, “nuclear membrane” y “fibrillar center”, todos con un número considerable de genes asociados. Los términos que más destacan son biogénesis de ribosomas, procesamiento de rRNA, huso mitótico y membrana nuclear. (ver Figura 4) Además, con $\log_2$FC > 1, emergieron términos de Funciones Moleculares (ver Figura S1).

## Anotación de secuencias e identificación del agente patógeno

Todos los resultados de este apartado se encuentran como resultados complementarios (revisar material suplementario). A través de múltiples búsquedas BLASTn, se comparó la secuencia incógnita con bases de datos genómicas de diversos organismos, incluyendo virus humanos (Influenza A, Dengue, Ébola), bacterias (Salmonella enterica), genes humanos (INS, IGF2, TH, NEK2), e inmunoglobulinas. Los resultados fueron evaluados considerando tres criterios clave: *E-value*, porcentaje de identidad y cobertura de la consulta.

Los alineamientos con virus de la Influenza A mostraron *E-values* extremadamente bajos (hasta 2e-136), porcentajes de identidad superiores al 95%, y coberturas cercanas al 79%, específicamente con el gen NP (nucleoproteína), una región altamente conservada y funcionalmente esencial en el ciclo viral (PAHO, 2020; ECDC, 2021). En contraste, los matches con virus del Dengue tipo 2 también presentaron *E-values* bajos (1e-120) y identidades del 91.99%, pero con menor cobertura (44%). Los resultados relacionados con el virus del Ébola

mostraron *E-values* de 8e-164 y cobertura del 96.69%, pero con un porcentaje de identidad más bajo (37%), lo que reduce la certeza de asignación. Las coincidencias con *Salmonella entérica* y genes humanos presentaron identidades moderadas (28–43%) y valores de *E* más altos o no reportados, lo que sugiere que no corresponden al patógeno en cuestión.

Considerando todos los parámetros, el patógeno que presenta la mayor evidencia de coincidencia biológica con la secuencia incógnita es el virus de la Influenza A, específicamente en su gen de nucleoproteína (NP). Esta conclusión se basa en la combinación de:

En la tabla de Influenza A, los cinco primeros matches presentaron:

- *E-value* = 0.0
- % Identidad entre 94.35 y 94.78 %
- *Query Cover* = 79 %

Estos valores indican coincidencias estadísticamente perfectas (*E-value* cero), con altísima similitud nucleotídica y un alineamiento que cubre la mayor parte de la secuencia. Por contraste, los hits de Dengue 2 mostraron identidades algo inferiores (≈92 %) y coberturas parciales (44 %), y los del virus del Ébola, aunque con *E-value* muy bajo (8e-164) y alta cobertura (96 %), tuvieron identidades muy bajas (37 %). Las comparaciones con *Salmonella* y genes humanos arrojaron identidades moderadas (28–43 %) o se basaron en regiones no virales.

La combinación de *E-value* cero, identidad > 94 % y cobertura alta (79 %) en el gen NP de Influenza A supera claramente el soporte estadístico y biológico de cualquier otro match. Por lo tanto, se concluye que el agente patógeno incógnito es el virus de la Influenza A, y la región analizada corresponde al segmento 5 (nucleoproteína NP), esencial para la encapsulación y ensamblaje (WHO, 2021; Daniels et al., 2020).

## Análisis de blancos terapéuticos

En los análisis realizados, se identificó la proteína JUN como uno de los genes sobreexpresados en el análisis transcriptómico, además de presentar un valor destacado de *betweenness centrality* en la red PPI construida, lo que la posiciona como un nodo crítico en la respuesta del hospedero. Esta observación motivó la búsqueda en la ruta de Influenza A (KEGG hsa05164) (ver Figura Sx), en la cual JUN se involucra en la cascada de señalización MAPKK, en esta cascada se observó que HA (Hemaglutinina), NA (Neuraminidasa) y el canal iónico M2 se encuentran implicadas. Por otra parte, se observó que la proteína NS1 (Non-Structural Protein 1) del virus puede inducir la activación de la vía JNK/JUN. (Nacken et al., 2014; Zhang et al., 2016) Por lo tanto, se seleccionó NS1 como uno de los blancos terapéuticos, junto con NA, dada su relevancia clínica. (Kumari et al., 2023)

**Tabla 1. Anotación estructural y funcional de NS1 (Non-Structural Protein 1) y NA (Neuraminidasa)**

| **Característica** | **NS1**(Carrillo et al., 2014; Engel, 2013; Ji et al., 2021) | **NA** (Air, 2012; Lederhofer et al., 2024) |
|---|---|---|
| **Función principal** | Inhibición de la respuesta inmune del hospedero | Clivaje del ácido siálico para liberar nuevos viriones |
| **Tipo de proteína** | No estructural | Proteína de superficie |
| **Estado oligomérico** | Homodímero | Tetrámero |
| **Dominios principales** (Figuras Sx y Sx), | Dominio de unión a ARN (RBD) Dominio de unión a proteínas celulares (ED) | Dominio catalítico tipo sialidasa |
| **Conformaciones estructurales** | Cerrada, semiabierta, abierta | Forma activa estable con canal hidrofóbico |
| **Sitios de interacción importantes** | Unión a CPSF30, ARN bicatenario | Sitio catalítico (Arg118, Glu119, Asp151, etc.) |
| **Homología** | Conservada entre cepas de IAV tipo A, pero variable en región C-terminal | Alta homología con otras neuraminidasas tipo A y B |
| **Rol en la virulencia** | Suprime el interferón y modula la expresión génica | Facilita la propagación viral |
| **Interés farmacológico** | Blanco emergente para antivirales | Blanco validado con fármacos aprobados |

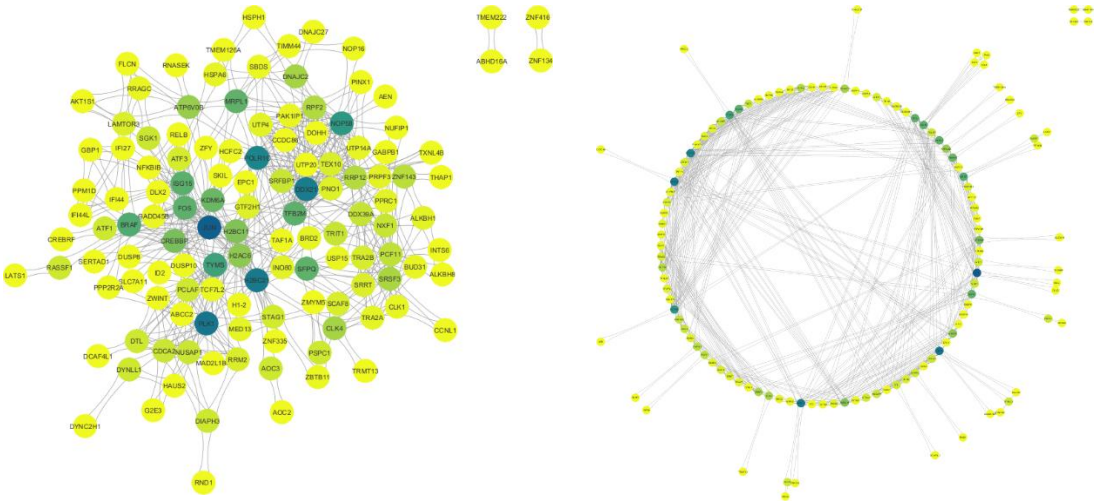
Con el fin de facilitar la visualización y comprensión de la organización estructural de los blancos terapéuticos seleccionados, las representaciones tridimensionales de sus dominios y la topología correspondiente han sido incluidas en el material suplementario (Figuras 8, 9, y 10).

Para evaluar los sitios de regulación de los blancos terapéuticos seleccionados (NS1 y NA), se generaron modelos tridimensionales mediante SWISS-MODEL, seleccionando estructuras con alta calidad y sin mutaciones relevantes. En el caso de la proteína NS1, se identificaron seis cavidades a través de PockDrug, destacando el bolsillo 2 (cadena B) y 3 (cadena A) por su alta probabilidad de drugabilidad (0.95 y 0.84 respectivamente) (ver Tabla S15), por lo que fue seleccionada como sitio potencial de interacción. Para la neuraminidasa (NA), se utilizó como sitio de regulación el centro activo conocido de los antivirales aprobados (oseltamivir, zanamivir, peramivir).

Posteriormente se filtraron 36 compuestos de la base de datos peruNPDB usando DataWarrior, descartando aquellos con propiedades toxicológicas y de drugabilidad. El cribado virtual y acoplamiento molecular destacó el compuesto peruNPDB_064 por mostrar consistentemente las mejores afinidades de unión tanto para NA (-9.2 kcal/mol) como para NS1 (-7.9 kcal/mol), superando a los controles positivos oseltamivir (-6.0 kcal/mol) y EGCG (-7.4 kcal/mol) respectivamente. Estos resultados posicionan a peruNPDB_064 (II-2.3-dihydro-I3', II8-biapigenin) como un candidato promisorio para el desarrollo de inhibidores duales frente a proteínas clave del virus de influenza A. (Datos representados en Material Suplementario).

## Red de interacción génica

Se construyó una red de interacción utilizando los genes sobreexpresados mediante la plataforma STRING y se visualizó en Cytoscape. Para representar la red, se empleó el layout orgánico (organic layout), el cual organiza los nodos en función de la densidad de conexiones, simulando fuerzas de atracción y repulsión entre ellos. Esta disposición permite identificar comunidades funcionales o clústeres de genes relacionados.



A B

**Figura 5. Red de interacción proteína-proteína (PPI) construida a partir de los genes sobreexpresados utilizando la plataforma STRINGdb.** (A) Representación de la red utilizando el algoritmo orgánico, donde los nodos más grandes y oscuros tienen una mayor influencia en el flujo de información dentro de la red. (B) Representación circular de la misma red, permitiendo visualizar la distribución global de interacciones y componentes conectados.

La red de interacción proteína-proteína (PPI), construida a partir de los genes sobreexpresados mediante STRINGdb y visualizada en Cytoscape, estuvo compuesta por 127 nodos y 706 aristas. El análisis topológico reveló un coeficiente de clustering de 0.391, un grado promedio de 5.707, un diámetro de 8 y una geodésica compuesta por 9 nodos que representan el camino más largo entre los caminos más cortos posibles dentro de la red. La longitud promedio del camino fue de 3.346, mientras que la densidad (probabilidad de conectividad entre los nodos) fue de 0.047. Los nodos que conforman la geodésica (RND1, DIAPH3, CDCA2, TYMS, SRFBP1, NOP58, DDX39A, PCF11 e INTS6) presentan una disposición que sugiere una relevancia funcional clave en la red (ver Figura S15). Algunos genes relevantes como PCGF1, parte del complejo epigenético Polycomb PRC1, y ZMYM5, un represor transcripcional, también estuvieron involucrados. Asimismo, CHD4 y MTA2, miembros del complejo NuRD, refuerzan la hipótesis de que procesos de remodelación de cromatina y regulación transcripcional podrían estar implicados en la respuesta del hospedero frente a la infección por Influenza A.

# Discusión

El presente estudio representa un abordaje multidisciplinario para la caracterización del agente causal de la fiebre de Norilsk, combinando análisis epidemiológicos, transcriptómicos y estructurales mediante herramientas bioinformáticas de última generación. La identificación del virus Influenza A como el patógeno responsable se sustenta en alineamientos BLASTn altamente significativos, con *E-values* de 0.0, identidades superiores al 94% y coberturas del 79%, específicamente en el gen de nucleoproteína (NP), una región altamente conservada y funcionalmente esencial en el ciclo viral (WHO, 2021; Daniels et al., 2020).

La robustez de esta conclusión se ve reforzada por la comparación con otros virus como Dengue tipo 2 y Ébola, cuyos alineamientos presentaron menor cobertura o identidad, lo que reduce la certeza de asignación. Además, los matches con *Salmonella entérica* y genes humanos mostraron identidades moderadas y *E-values* más altos, descartando su implicancia como agentes causales.

El uso de herramientas como BLASTn, ClustalW y MEGA permitió no solo la identificación precisa del patógeno, sino también la inferencia evolutiva y la anotación funcional de sus secuencias. Estas plataformas

han demostrado ser esenciales en estudios de vigilancia genómica y diagnóstico molecular, especialmente en contextos de brotes emergentes (Ramírez-Salinas et al., 2020; King et al., 2020).

El análisis mediante Volcano Plot permitió identificar claramente los genes diferencialmente expresados, observándose que con un umbral inicial de $\log_2$FC > 1 se abarcaban casi todos los genes, debido a la ausencia de valores cercanos a cero, lo que contrasta con otros estudios donde predominan genes sin cambio.(Linggi et al., 2021; Thaden et al., 2023; Zhang et al., 2022) Esta distribución sugiere una respuesta celular intensa causada probablemente por la infección viral o el estrés asociado a la enfermedad analizada.(Dexheimer et al., 2023) Por ello, se aplicó un umbral más estricto ($\log_2$FC > 2 y $p$ < 0.01) para enfocarse en genes con cambios altamente significativos. (Rodriguez-Esteban & Jiang, 2017)

Con respecto a los resultados del análisis GO, indican un perfil transcriptómico marcado por la sobreexpresión de genes involucrados en la biogénesis ribosomal, el procesamiento de rRNA, y componentes celulares como el huso mitótico y la membrana nuclear. Además, se observó la activación de rutas asociadas a la señalización mediada por p53, sugiriendo una respuesta regulatoria ante estrés celular. Este conjunto de cambios refleja un estado de alta actividad biosintética y proliferativa, que también ha sido descrito en células sometidas a infección viral. (Bianco & Mohr, 2019; *Genome Guardian P53 and Viral Infections - Sato - 2013 - Reviews in Medical Virology - Wiley Online Library*, n.d.; Limkar et al., 2021)

Del análisis de acoplamiento molecular sobre NS1 y NA, se identificaron compuestos de la base de datos PeruNPDB,(6) como *peruNPDB_064* y *peruNPDB_061*, con afinidades de unión más favorables que los ligandos de referencia EGCG y oseltamivir, respectivamente. Estos resultados sugieren que metabolitos naturales peruanos podrían tener un potencial inhibitorio comparable o superior al de fármacos comerciales.

La red de interacción génica construida muestra características estructurales propias de sistemas biológicos robustos, con conectividad eficiente y la presencia de nodos clave que podrían desempeñar funciones regulatorias esenciales durante la respuesta del hospedero frente al virus de la influenza A. La existencia de módulos y caminos cortos entre genes sugiere una organización funcional que favorece la rápida propagación de señales, destacando la relevancia de ciertos genes como posibles blancos terapéuticos.

A través del presente trabajo evidenciamos cómo la bioinformática permite integrar múltiples niveles de análisis desde datos genómicos y transcriptómicos hasta redes de interacción y modelado estructural para abordar problemas complejos en biología y salud. Este enfoque multidisciplinario refleja la esencia del curso, al capacitar en el uso crítico de tecnologías bioinformáticas aplicadas a problemas reales en biología computacional.

Los documentos y scripts utilizados se encuentran en:
https://github.com/MargotPC/Final_Project_DRBC_A001_USS

# Referencias Bibliográficas

Air, G. M. (2012). Influenza neuraminidase. Influenza and Other Respiratory Viruses, 6(4), 245–256. https://doi.org/10.1111/j.1750-2659.2011.00304.x

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., & Supuran, C. T. (2021). Natural products in drug discovery: Advances and opportunities. Nature Reviews Drug Discovery, 20(3), 200–216. https://doi.org/10.1038/s41573-020-00114-z

Barazorda-Ccahuana, H. L., Ranilla, L. G., Candia-Puma, M. A., Cárcamo-Rodriguez, E. G., Centeno-Lopez, A. E., Davila-Del-Carpio, G., Medina-Franco, J. L., & Chávez-Fumagalli, M. A. (2023). PeruNPDB: The Peruvian Natural Products Database for in silico drug screening. Scientific Reports, 13(1), Article 1. https://doi.org/10.1038/s41598-023-34729-0

Bianco, C., & Mohr, I. (2019). Ribosome biogenesis restricts innate immune responses to virus infection and DNA. eLife, 8, e49551. https://doi.org/10.7554/eLife.49551

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., & Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research, 42(W1), W252–W258. https://doi.org/10.1093/nar/gku340

Brauer, F., Castillo-Chavez, C., & Feng, Z. (2019). *Mathematical Models in Epidemiology*. Springer. https://doi.org/10.1007/978-1-4939-9828-9

Carrillo, B., Choi, J.-M., Bornholdt, Z. A., Sankaran, B., Rice, A. P., & Prasad, B. V. V. (2014). The Influenza A Virus Protein NS1 Displays Structural Polymorphism. Journal of Virology, 88(8), 4113–4122. https://doi.org/10.1128/jvi.03692-13

Daniels, R., Ermetal, B., Rattigan, Á., & McCauley, J. (2020). *Influenza virus characterisation – Summary Europe, July 2020*. European Centre for Disease Prevention and Control. https://www.ecdc.europa.eu/sites/default/files/documents/influenza-virus-characterisation-july-2020.pdf

Dallakyan, S., & Olson, A. J. (2015). Small-Molecule Library Screening by Docking with PyRx. In J. E. Hempel, C. H. Williams, & C. C. Hong (Eds.), Chemical Biology: Methods and Protocols (pp. 243–250). Springer. https://doi.org/10.1007/978-1-4939-2269-7_19

Dexheimer, P. J., Pujato, M., Roskin, K. M., & Weirauch, M. T. (2023). VExD: A curated resource for human gene expression alterations following viral infection. G3: Genes|Genomes|Genetics, 13(10), jkad176. https://doi.org/10.1093/g3journal/jkad176

Engel, D. A. (2013). The influenza virus NS1 protein as a therapeutic target. Antiviral Research, 99(3), 409–416. https://doi.org/10.1016/j.antiviral.2013.06.005

European Centre for Disease Prevention and Control. (2021). *Seasonal influenza 2020–2021 Annual Epidemiological Report*. https://www.ecdc.europa.eu/sites/default/files/documents/AER-seasonal-influenza-2020-final.pdf

Farooq, Q. ul A., Shaukat, Z., Aiman, S., Zhou, T., & Li, C. (2020). A systems biology-driven approach to construct a comprehensive protein interaction network of influenza A virus with its host. BMC Infectious Diseases, 20(1), 480. https://doi.org/10.1186/s12879-020-05214-0

Genome guardian p53 and viral infections—Sato—2013—Reviews in Medical Virology—Wiley Online Library. (n.d.). Retrieved July 22, 2025, from https://onlinelibrary.wiley.com/doi/abs/10.1002/rmv.1738

Haynes, W. (2013). Benjamini–Hochberg Method. In Encyclopedia of Systems Biology (pp. 78–78). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_1215

Hussein, H. A., Borrel, A., Geneix, C., Petitjean, M., Regad, L., & Camproux, A.-C. (2015). PockDrug-Server: A new web server for predicting pocket druggability on holo and apo proteins. Nucleic Acids Research, 43(W1), W436–W442. https://doi.org/10.1093/nar/gkv462

Ito, K., & Murphy, D. (2013). Application of ggplot2 to Pharmacometric Graphics. CPT: Pharmacometrics & Systems Pharmacology, 2(10), 79. https://doi.org/10.1038/psp.2013.56

Ji, Z., Wang, X., & Liu, X. (2021). NS1: A Key Protein in the "Game" Between Influenza A Virus and Host in Innate Immunity. Frontiers in Cellular and Infection Microbiology, 11. https://doi.org/10.3389/fcimb.2021.670177

Kermack, W. O., & McKendrick, A. G. (1927). *A Contribution to the Mathematical Theory of Epidemics*. Proceedings of the Royal Society A, 115(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

King, J., Harder, T., Beer, M., & Pohlmann, A. (2020). *Rapid multiplex MinION nanopore sequencing workflow for Influenza A viruses*. BMC Infectious Diseases, 20(648). https://doi.org/10.1186/s12879-020-05367-y

Kohl, M., Wiese, S., & Warscheid, B. (2011). Cytoscape: Software for Visualization and Analysis of Biological Networks. In M. Hamacher, M. Eisenacher, & C. Stephan (Eds.), Data Mining in Proteomics: From Standards to Applications (pp. 291–303). Humana Press. https://doi.org/10.1007/978-1-60761-987-1_18

Kumari, R., Sharma, S. D., Kumar, A., Ende, Z., Mishina, M., Wang, Y., Falls, Z., Samudrala, R., Pohl, J., Knight, P. R., & Sambhara, S. (2023). Antiviral Approaches against Influenza Virus. Clinical Microbiology Reviews, 36(1), e00040-22. https://doi.org/10.1128/cmr.00040-22

Lederhofer, J., Tsybovsky, Y., Nguyen, L., Raab, J. E., Creanga, A., Stephens, T., Gillespie, R. A., Syeda, H. Z., Fisher, B. E., Skertic, M., Yap, C., Schaub, A. J., Rawi, R., Kwong, P. D., Graham, B. S., McDermott, A. B., Andrews, S. F., King, N. P., & Kanekiyo, M. (2024). Protective human monoclonal antibodies target conserved sites of vulnerability on the underside of influenza virus neuraminidase. Immunity, 57(3), 574-586.e7. https://doi.org/10.1016/j.immuni.2024.02.003

Limkar, A. R., Lack, J. B., Sek, A. C., Percopo, C. M., Druey, K. M., & Rosenberg, H. F. (2021). Differential Expression of Mitosis and Cell Cycle Regulatory Genes during Recovery from an Acute Respiratory Virus Infection. Pathogens, 10(12), Article 12. https://doi.org/10.3390/pathogens10121625

Linggi, B., Jairath, V., Zou, G., Shackelton, L. M., McGovern, D. P. B., Salas, A., Verstockt, B., Silverberg, M. S., Nayeri, S., Feagan, B. G., & Vande Casteele, N. (2021). Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis. Scientific Reports, 11(1), 18243. https://doi.org/10.1038/s41598-021-97366-5

López-López, E., Naveja, J. J., & Medina-Franco, J. L. (2019). DataWarrior: An evaluation of the open-source drug discovery tool. Expert Opinion on Drug Discovery, 14(4), 335–341. https://doi.org/10.1080/17460441.2019.1581170

Nacken, W., Anhlan, D., Hrincius, E. R., Mostafa, A., Wolff, T., Sadewasser, A., Pleschka, S., Ehrhardt, C., & Ludwig, S. (2014). Activation of c-jun N-Terminal Kinase upon Influenza A Virus (IAV) Infection Is Independent of Pathogen-Related Receptors but Dependent on Amino Acid Sequence Variations of IAV NS1. Journal of Virology, 88(16), 8843–8852. https://doi.org/10.1128/JVI.00424-14

Omics Tutorials. (2024). *Mastering Cytoscape for multi-omic network analysis*. https://www.omicstutorials.org/cytoscape2024

Pan American Health Organization. (2020). *Influenza Regional Update EW 50*. https://www.paho.org/sites/default/files/2020-12/2020-phe-influenza-report-50.pdf

Ramírez-Salinas, G. L., et al. (2020). *Bioinformatics design and experimental validation of influenza A virus multi-epitopes that induce neutralizing antibodies*. Archives of Virology, 165, 891–911. https://doi.org/10.1007/s00705-020-04537-2

Rodriguez-Esteban, R., & Jiang, X. (2017). Differential gene expression in disease: A comparison between high-throughput studies and the literature. BMC Medical Genomics, 10(1), 59. https://doi.org/10.1186/s12920-017-0293-y

Santos, J. D., et al. (2024). *INSaFLU-TELEVIR: an open web-based bioinformatics suite for viral metagenomic detection and routine genomic surveillance*. Genome Medicine, 16(61). https://doi.org/10.1186/s13073-024-01334-3

Sharma, V., & Dubey, D. (2019). *Comparative analysis of multiple sequence alignment tools for viral genome annotation*. VirusDisease, 30, 515–522. https://doi.org/10.1007/s13337-019-00537-8

Szklarczyk, D., Nastou, K., Koutrouli, M., Kirsch, R., Mehryary, F., Hachilif, R., Hu, D., Peluso, M. E., Huang, Q., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2025). The STRING database in 2025: Protein networks with directionality of regulation. Nucleic Acids Research, 53(D1), D730–D737. https://doi.org/10.1093/nar/gkae1113

Talukdar, P. D., & Chatterji, U. (2023). Transcriptional co-activators: Emerging roles in signaling pathways and potential therapeutic targets for diseases. Signal Transduction and Targeted Therapy, 8(1), 427. https://doi.org/10.1038/s41392-023-01651-w

Tamura, K., Stecher, G., & Kumar, S. (2021). *MEGA11: Molecular Evolutionary Genetics Analysis version 11*. Molecular Biology and Evolution, 38(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

Thaden, J. T., Ruffin, F., Gjertson, D., Hoffmann, A., Fowler, V. G., & Yeaman, M. (2023). 240. Transcriptional signatures differentiate pathogen- and treatment-specific host responses in patients with bacterial bloodstream infections. Open Forum Infectious Diseases, 10(Suppl 2), ofad500.313. https://doi.org/10.1093/ofid/ofad500.313

World Health Organization. (2021). *Review of global influenza circulation, late 2019 to 2020, and the impact of the COVID-19 pandemic on influenza circulation*. https://www.who.int/publications/i/item/who-wer-9625-241-264

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3). https://doi.org/10.1016/j.xinn.2021.100141

Yang, C.-R., King, C.-C., Liu, L.-Y. D., & Ku, C.-C. (2020). *FluConvert and IniFlu: a suite of integrated software to identify novel signatures of emerging influenza viruses*. BMC Bioinformatics, 21(316). https://doi.org/10.1186/s12859-020-03650-y

Zhang, J., Wang, J., Liu, Q., & Chen, G. (2020). *Applications of BLAST in bioinformatics*. Journal of Bioinformatics and Computational Biology, 18(5), 2050032. https://doi.org/10.1142/S021972002050032X

Zhang, S., Jiang, H., Gao, B., Yang, W., & Wang, G. (2022). Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network. Frontiers in Cell and Developmental Biology, 9. https://doi.org/10.3389/fcell.2021.811585

Zhang, S., Tian, H., Cui, J., Xiao, J., Wang, M., & Hu, Y. (2016). The c-Jun N-terminal kinase (JNK) is involved in H5N1 influenza A virus RNA and protein synthesis. Archives of Virology, 161(2), 345–351. https://doi.org/10.1007/s00705-015-2668-8