

# Uso de R para el análisis y creación de informes de secuenciación de exomas en pacientes con asfixia perinatal \*

**Diego Mauricio Gómez Londoño** *Comfamiliar Risaralda, Universidad de Manizales*

**Hernán Felipe García Árias** *Comfamiliar Risaralda*

**Natalia Trujillo Árias** *Comfamiliar Risaralda*

**Jorge Mario Estrada Álvarez** *Comfamiliar Risaralda*

**Natalia Cardona Ramirez** *Comfamiliar Risaralda*

**Juan Alejandro Trujillo Posada** *Universidad de Manizales*

**Gloria-Liliana Porras Hurtado** *Comfamiliar Risaralda*

---

**Resumen:** El análisis de secuenciación de exomas en pacientes con asfixia perinatal requiere herramientas eficientes. Presentamos una solución basada en R, un lenguaje de programación libre y ampliamente utilizado en ciencia de datos. RStudio, un entorno de desarrollo integrado para R, facilita la creación de scripts y la integración con herramientas como Quarto o Markdown para generar documentos dinámicos y reproducibles. Aplicamos filtros y realizamos búsquedas específicas de genes relacionados con el fenotipo de la enfermedad de cada paciente sobre los archivos provenientes del análisis de exomas para seleccionar variantes relevantes y generamos informes individuales para pacientes. El uso de R ofrece ventajas como alta calidad, accesibilidad gratuita, automatización, personalización y capacidad de generar documentos profesionales. Este trabajo demuestra el potencial de R en bioinformática para la investigación en genómica médica y la presentación de los resultados.

**Keywords:** Análisis de datos, Bioinformática, Secuenciación de exomas completos, r markdown,

---

## Introducción

La secuenciación de exomas completos se ha convertido en un enfoque crucial para desentrañar los misterios genéticos detrás de condiciones médicas cruciales, como la asfixia perinatal. Este desafío bioinformático, derivado de la insuficiencia de oxígeno durante el parto, plantea interrogantes genéticos esenciales que requieren atención especializada ([Carlson and Vora, 2017](#)).

A pesar de la creciente importancia de la secuenciación de exomas en medicina, la información genética específica sobre la asfixia perinatal ha sido elusiva, obstaculizando la identificación precisa de variantes genéticas asociadas. Esta brecha en el conocimiento limita el desarrollo de intervenciones personalizadas ([Gorovenko et al., 2010](#)).

Ante esta necesidad crítica, se requieren herramientas eficientes y flexibles para abordar la vasta cantidad de datos generados por la secuenciación de exomas. R, un lenguaje de programación de código abierto, destaca como la elección natural. Su capacidad para manejar grandes conjuntos de datos, realizar análisis estadísticos avanzados y generar informes automatizados lo posiciona como la herramienta ideal para enfrentar el desafío de la secuenciación de exomas en pacientes con asfixia perinatal.

Este proyecto no solo busca llenar el vacío de información genética en la asfixia perinatal mediante la secuenciación de exomas completos, sino también resaltar la eficacia de R en el manejo

---

\* Autor de contacto: [diegomez@comfamiliar.com](mailto:diegomez@comfamiliar.com)

de grandes volúmenes de datos genómicos. Utilizando RMarkdown para la generación y automatización de informes, el objetivo es identificar variantes genéticas asociadas a la asfixia perinatal y presentar un modelo replicable para futuras investigaciones en genómica médica. En última instancia, este estudio aspira a demostrar que, frente a la complejidad de la información genética, R es la herramienta idónea para avanzar hacia una comprensión más profunda de las bases genéticas de la asfixia perinatal.

## Metodología

El desarrollo de una solución integral para abordar el análisis de secuenciación de exomas en pacientes con asfixia perinatal implicó la cuidadosa implementación de un flujo bioinformático respaldado por herramientas avanzadas. La elección central de R, un lenguaje de programación versátil y consolidado en la ciencia de datos, fue fundamental para garantizar la eficacia y reproducibilidad del análisis. El Análisis de los datos se realizó principalmente mediante filtros y búsquedas de genes relacionados con el fenotipo clínico de cada paciente. Se utilizó RMarkdown para la generación y de los documentos de reporte de cada paciente. Se utilizó el software R 4.3.1 sobre una estación de trabajo con Ubuntu 22.04 debido a la necesidad de ejecutar el flujo bioinformático en sistemas basados en Linux, optimizando así la compatibilidad y eficiencia del análisis.

El desarrollo de este trabajo se encuentra bajo el proyecto de investigación “Sistema de monitoreo Automático para la evaluación clínica de infantes con lateraciones neurológicas motoras mediante el análisis de volumetría cerebral y patrón de la marcha” que es financiado por el Ministerio de Ciencia, Tecnología y Educación. Adicionalmente, el proyecto se encuentra avalado por el comité de ética de Comfamiliar Risaralda y se cuenta con consentimiento informado de los pacientes.

## Resultados

### *Proceso de Análisis de Datos*

La ejecución de scripts en R ha simplificado el procesamiento de datos complejos de la secuenciación de exomas. La integración de funciones específicas ha permitido un filtrado preciso de variantes, considerando criterios como la frecuencia alélica, la localización en genes relevantes y la presencia en bases de datos de enfermedades genéticas. Este enfoque ha agilizado el análisis, reduciendo posibles errores y generando resultados más confiables (Figura 1).












hpo_366665	14309 obs. of 7 variables	
ID_raw_vcf	2282091 obs. of 92 variables	
resumen_hpo	7 obs. of 24 variables	
resumen_patho	7 obs. of 21 variables	
vcf_paciente	300320 obs. of 105 variables	
vcf_paciente_1	127677 obs. of 105 variables	
vcf_paciente_2	104395 obs. of 105 variables	
vcf_paciente_3	615 obs. of 105 variables	
vcf_paciente_hpo_1	615 obs. of 111 variables	
vcf_paciente_hpo_2	7 obs. of 111 variables	
vcf_paciente_patho	7 obs. of 105 variables	

Figure 1: Dataframes creados con distintos filtros.

La aplicación de filtros fue esencial para seleccionar las variantes más relevantes. Criterios como la frecuencia alélica, la localización en genes neuromusculares o relacionados con el fenotipo clínico, y la presencia en bases de datos de enfermedades genéticas se implementaron mediante un Rscript. Este script genera un dataframe conciso (inicialmente 105 columnas, finalmente 21 o 24) como se presenta parcialmente en la figura 2 y un conjunto limitado de genes, facilitando la revisión y selección de variantes genéticas por parte de los médicos genetistas.


gene_name	nt_change	aa_change	effect	impact	CLNSIG	CLNDN
ITPKB	c.964G>A	p.Ala322Thr	missense_variant	MODERATE	Pathogenic	Myeloproliferative_neoplasm_unclassifiable
EHBP1	c.1290+30064G>A		intron_variant	MODIFIER	Pathogenic	Prostate_cancer_hereditary_12
SLC1A4	c.964C>T	p.Arg322*	stop_gained	HIGH	Pathogenic/Likely_pathogenic	Spastic_tetraplegia-thin_corpus_callosum-progressiv...
CTBP2	c.86A>T	p.Asn29Ile	missense_variant	MODERATE	Pathogenic	Myoepithelial_tumor Hereditary_pancreatitis Trypsin...
CTBP2	c.2287G>A	p.Val763Met	missense_variant	MODERATE	Pathogenic	Pulmonary_artery_atresia
KLK2	c.748C>T	p.Arg250Trp	missense_variant	MODERATE	Pathogenic	Acute_myeloid_leukemia
CBS	c.833T>C	p.Ile278Thr	missense_variant	MODERATE	Pathogenic	Connective_tissue_disorder Homocystinuria not_pro...

Showing 1 to 7 of 7 entries, 21 total columns

Figure 2: Dataframe resumido con variantes genéticas de interés.

La generación de informes de análisis se realizó con RMarkdown, proporcionando documentos estructurados y detallados por paciente (figura 3). Estos informes incluyeron datos sobre el paciente, tipo de estudio, tabla con variantes reportadas o resultados positivos, información clínica respaldada en literatura relevante sobre las variantes, y detalles técnicos del proceso de secuenciación por un laboratorio externo. Este enfoque integrado aseguró presentar los resultados de manera clara y completa.

La principal ventaja de RMarkdown es la automatización del proceso de generación de informes. La integración de código R en el documento facilita la reproducción de resultados, asegurando coherencia entre hallazgos y presentación. RMarkdown permite la inclusión dinámica de gráficos, tablas y referencias bibliográficas, mejorando la calidad y eficacia en la comunicación de los resultados.



Av. Circunvalar No 3-01 Pereira

<b>Numero del paciente:</b>	<b>Código:</b>
#####	#####
<b>ID Paciente:</b>	<b>Sexo:</b>
#36665	xxxxx
<b>Tipo de muestra:</b>	<b>Médico resultante:</b>
Sangre periférica	Gloria Liliana Perea Hurtado
<b>Fecha de nacimiento:</b>	
DD/MM/AAAA	

**Prueba realizada:** Análisis de exoma completo mediante NGS

**Información clínica:** Paciente con síndrome perinatal, pérdida de audición sensoriomotor bilateral, Debilidad muscular distal, Deterioro en las extremidades, Atrofia muscular/Trombosis arterial.

**Resultado:** Positivo

Gen	Variantes	Cantidad	Clasificación	Herencia
SLC1A4	c.964C>T	Heterocigota	Patogénica/Probablemente patogénica	AR

#### Interpretación

El gen SLC1A4 codifica el transportador de aminoácidos neutros DEPENDIENTE de Na<sup>+</sup> (ASCT1), que transporta L-serina, L-alanina, L-cisteína y L-treonina. En el cerebro, la L-serina es sintetizada por los astrocitos y es transportada a las células neuronales por el transportador SLC1A4/detractor. La L-serina se considera un aminoácido esencial para las neuronas.

#### Información Adicional

Este resultado debe discutirse con el médico tratante en un contexto de asesoría genética, para obtener más información sobre este resultado y los próximos pasos sugeridos para una evaluación adicional. El seguimiento clínico aún puede estar justificado. Este resultado debe interpretarse dentro del contexto de resultados de laboratorio adicionales, antecedentes familiares y hallazgos clínicos.

#### Metodología

La toma de las muestras y la secuenciación genética se realizó en GenCoil Pharma, donde el ADN genómico obtenido de la muestra sanguínea se extrajo para regiones objetivo usando un protocolo basado en hibridación, y se secuenció usando la tecnología Illumina. A menos que se indique lo contrario, todas las regiones objetivo se secuenciaron con 50x de profundidad o se complementaron con análisis adicionales. Las lecturas se alinearon con una secuencia de referencia

(GRCh37). Los promotores, las regiones sin traducir y otras regiones no codificantes no son evaluadas.

El análisis de variantes genéticas se realizó mediante un algoritmo bioinformático interno en Salud Comfamiliar, donde la identificación e interpretación se realizaron bajo el contexto clínicamente relevante del paciente.

#### Limitaciones

Este ensayo alcanza una sensibilidad y especificidad >99% para las variantes de un solo nucleótido y las inserciones y deleciones <15 pb en longitud.

Puede que no sea posible resolver completamente ciertos detalles sobre las variantes, como el mosaicismo, la fase o la ambigüedad de mapeo. A menos que se garantice explícitamente, este ensayo no cubre los cambios de secuencia en el promotor, los exones no codificantes y otras regiones no codificantes. Este informe refleja el análisis de una muestra de ADN genómico extraído. En casos muy raras (como receptores hematológicos circulantes, trasplante de médula ósea, transfusión de sangre reciente o contaminación de células maternas), el ADN analizado puede no representar el genoma constitucional del paciente.

#### Descargo de responsabilidad

Los estudios de ADN no constituyen una prueba definitiva para las alteraciones seleccionadas en todos los individuos. Se debe tener en cuenta que hay posibles fuentes de error. Los errores pueden resultar de la contaminación por trans, errores técnicos raras, variantes genéticas raras que interfieren con el análisis, desarrollos tecnológicos recientes y sistemas de clasificación alternativos. Esta prueba ha sido desarrollada con fines de investigación.

1

2

Figure 3: Documento de reporte para análisis genético.

## Conclusión

La implementación exitosa de scripts en R y Markdown no solo ha mejorado la eficiencia y la precisión en el análisis de datos genómicos, sino que también ha establecido un modelo valioso para la generación de informes en el campo de la bioinformática y la genómica médica. Este enfoque no solo representa un avance en la investigación actual sobre la asfixia perinatal, sino que también sienta las bases para futuros proyectos que busquen integrar análisis de datos genómicos de manera eficaz y reproducible.

## References

- Carlson, Laura M and Neeta L Vora. 2017. "Prenatal diagnosis: screening and diagnostic tools." *Obstetrics and Gynecology Clinics* 44(2):245–256.
- Gorovenko, NG, ZI Rossokha, SV Podolskaya, VI Pokhylko and Gunilla A Lundberg. 2010. "The role of genetic determinant in the development of severe perinatal asphyxia." *Cytology and genetics* 44(5):294–299.