

REPRODUCIBILIDAD EN LOS FLUJOS DE TRABAJO BIOINFORMÁTICOS APLICADOS A GENÓMICA Y ONCOLOGÍA CLÍNICA

Reproducibility in bioinformatics workflows applied to genomics and clinical oncology

 Juan Sebastian Loza Chiriboga ^{(1)*}
juansebaslozascience@gmail.com

 Michael Gustavo Miranda Coello ⁽³⁾
md.gustavo25@gmail.com

 Valeria Alexandra Riera Sampedro ⁽²⁾
valeriariera70@gmail.com

 Dennys Rodrigo Lopez Chavez ⁽⁴⁾
dennysrodrigolopez@gmail.com

⁽¹⁾ Universidad Internacional de Valencia. Facultad de Ciencias de la Salud. Bioinformática. Valencia, España.

⁽²⁾ Universidad Nacional de Chimborazo, Facultad de Ciencias de la Salud, Riobamba, Chimborazo, Ecuador. Código postal: 060106. Correo electrónico: valeria.riera@unach.edu.ec

⁽³⁾ Médico General, Clínica Los Pinos - Hospital General, Quito, Pichincha, Ecuador. Código postal: 170125. Correo electrónico: md.gustavo25@gmail.com

⁽⁴⁾ Odontólogo General, Investigador independiente, Riobamba, Chimborazo, Ecuador. Código postal: 060106. Correo electrónico: dennysrodrigolopez@gmail.com

Autor de correspondencia:

Correo electrónico: juansebaslozascience@gmail.com / Teléfono: +593 98 428 3781

RESUMEN

Introducción: El incremento exponencial de datos genómicos en la oncología clínica ha posicionado a la reproducibilidad bioinformática como un pilar crítico para la validez de diagnósticos y terapias personalizadas. No obstante, la variabilidad en pipelines y la falta de estándares técnicos generan una crisis de consistencia en los resultados. **Objetivo:** Analizar los flujos de trabajo bioinformáticos aplicados a genómica oncológica, evaluando la capacidad de herramientas como Snakemake, Onkopi e iCOMIC para garantizar resultados clínicos reproducibles. **Metodología:** Se realizó una revisión integrativa siguiendo la declaración PRISMA 2020. La búsqueda se ejecutó en Pubmed y ScienceDirect. De 124 registros identificados, 38 estudios cumplieron con los criterios de elegibilidad y fueron analizados mediante una síntesis cualitativa y temática. **Resultados:** Se identificó una crisis de reproducibilidad ligada a la ineficiente documentación de parámetros. De los estudios analizados, 18 describieron prácticas explícitas: el 88,8% (n=16) destacó la documentación detallada, el 77,7% (n=14) el uso de gestores de flujo como Snakemake y el 61,1% (n=11) la implementación de contenedores y control de versiones. Se evaluaron 12 flujos específicos, donde el uso de Snakemake demostró optimizar la trazabilidad y escalabilidad del diagnóstico molecular. **Conclusiones:** La adopción rigurosa de gestores de flujos automatizados y la estandarización de la documentación técnica son indispensables para transitar de la bioinformación investigativa a una genómica clínica auditable y confiable.

Palabras claves: bioinformática, genómica, oncología, reproducibilidad, Snakemake.

ABSTRACT

Introduction: The exponential increase in genomic data in clinical oncology has positioned biological reproducibility as a critical pillar for the validity of personalised diagnoses and therapies. However, variability in pipelines and the lack of technical standards are causing a crisis of consistency in results. **Objective:** The analysis of bioinformatics workflows applied to oncology genomics is essential for evaluating the capacity of tools such as Snakemake, Onkopi and iCOMIC to guarantee reproducible clinical results. **Methodology:** A comprehensive review was conducted in accordance with the PRISMA 2020 statement, incorporating a systematic approach to the analysis of existing literature. The search was conducted using the PubMed and ScienceDirect databases. Of the 124 records identified, 38 studies met the eligibility criteria and were analysed using a qualitative and thematic synthesis. **Results:** It was determined that a reproducibility crisis was occurring, and this was connected to the inadequate documentation of parameters. Of the studies analysed, 18 described explicit practices: 88.8% (n=16) highlighted detailed documentation, 77.7% (n=14) the use of workflow managers such as Snakemake, and 61.1% (n=11) the implementation of containers and version control. Twelve specific workflows were evaluated, and the use of Snakemake was found to optimise the traceability and scalability of molecular diagnostics. **Conclusions:** The rigorous adoption of automated workflow managers and the standardization of technical documentation are essential for transitioning from research bioinformation to auditable and reliable clinical genomics.

Keywords: bioinformatics, genomics, oncology, reproducibility, Snakemake

»» 1. Introducción

La reproducibilidad dentro del ámbito científico ha surgido como uno de los principios fundamentales para la validación y el avance del conocimiento, especialmente en áreas como la genómica aplicada a la oncología, que ha transformado radicalmente la capacidad de comprender el cáncer a nivel molecular, permitiendo el desarrollo de biomarcadores hasta terapias personalizadas. Sin embargo, existen desafíos sustanciales en cuanto a la reproducibilidad de los resultados, debido a la infraestructura, hardware y las versiones específicas de las herramientas empleadas en los pipelines bioinformáticos, además la falta de estandarización de parámetros, la diversidad de algoritmos y la ausencia de criterios consensuados para la interpretación de resultados conducen a la variabilidad en los análisis (1–3).

Para abordar estos desafíos tanto técnicos como metodológicos la comunidad científica ha promovido a la implementación de gestores de flujo de trabajo, como Snakemake, que es ampliamente utilizado en bioinformática, utiliza lenguajes de dominio específico que acrecientan la portabilidad y escalabilidad de los análisis. Cabe destacar que estos sistemas permiten definir con precisión el flujo de datos entre las herramientas, procurando que cada paso sea reproducible y documentado (4).

La reproducibilidad bioinformática en genómica clínica y oncológica conforman el eje central para obtener resultados verificables, robustos y transferibles entre los distintos contextos clínicos e investigativos, por lo cual los flujos de trabajo representan un papel importante ya que son los encargados de la ingestión de datos hasta el análisis y la interpretación de resultados; herramientas como Onkopipec e iCOMIC, se han desarrollado con la finalidad de brindar soluciones estructurales, automatizadas y auditables, facilitando el trabajo con exuberantes volúmenes de datos y estandarización de metodologías (5,6).

A pesar del desarrollo de herramientas potentes, persiste un vacío en la estandarización de criterios que permitan la transferencia de estos flujos entre diferentes centros clínicos e investigativos. Ante este escenario, surge la siguiente pregunta de investigación: ¿Cuáles son las herramientas y estándares de buenas prácticas que permiten mitigar la crisis de reproducibilidad en los flujos de trabajo bioinformáticos aplicados a la genómica oncológica clínica?

Para responder a esta interrogante, el presente estudio tiene como objetivo central analizar el estado actual de la reproducibilidad en los pipeline bioinformáticos, evaluando específicamente el impacto de los gestores de flujo basados en Snakemake y herramientas especializadas como Onkopipec e iCOMIC. La revisión busca identificar los pilares técnicos (documentación, contenedores, control de versiones) necesarios para garantizar la generación de resultados clínicos consistentes y verificables en el ámbito de la medicina de precisión.

»» 2. Metodología

Se llevó a cabo una revisión bibliográfica para explorar cómo se garantiza la reproducibilidad en los flujos de trabajo bioinformáticos utilizados en genómica y oncología clínica, enfocándonos principalmente en herramientas basadas en Snakemake, con énfasis en Onkopipec e iCOMIC, destacando su utilidad y aplicación en este campo.

2.1 Criterios de elegibilidad

Para la selección de los estudios se establecieron criterios de elegibilidad, donde se incluyeron artículos originales, revisiones y reportes técnicos que abordaran la reproducibilidad en bioinformática, publicados entre los años 2017-2025, disponibles en idioma inglés o español y con acceso libre al texto completo. Además, se consideraron aquellos trabajos enfocados en flujos de trabajo bioinformáticos aplicados a la genómica clínica y la oncología.

Por el contrario, se excluyeron las publicaciones anteriores a 2017, los artículos que no trataran aspectos relacionados con la reproducibilidad o los flujos de trabajo bioinformáticos, así como cartas al editor, resúmenes, opiniones y documentos sin revisión por pares. También se descartaron los artículos con acceso restringido o de pago.

2.2 Fuentes de información

La información se recolectó de 2 bases de datos: PubMed y ScienceDirect. Así mismo, se obtuvo información manual de diferentes revistas como Bioinformatics, Briefings in Bioinformatics, BMC Genomics, BMC Bioinformatics, Nature y PLOS Computational Biology. También se recolectó información de repositorios de software como GitHub.

2.3 Estrategia de búsqueda

Aunque el estudio no corresponde a una revisión sistemática, se tomó como referencia metodológica la guía PRISMA 2020 para orientar de manera ordenada el proceso de identificación y selección de los estudios. Para la búsqueda bibliográfica se utilizaron combinaciones de palabras clave como “Reproducibility”, “Bioinformatics”, “Clinical Genomics”, “Pipeline” y “Snakemake”, las cuales fueron integradas mediante operadores booleanos AND, OR y NOT en las bases de datos consultadas.

2.4 Proceso de extracción de datos

La búsqueda inicial permitió identificar un total de 124 registros. Luego de eliminar los duplicados ($n = 18$), se procedió a la evaluación de 106 artículos mediante la lectura de títulos y resúmenes. En esta etapa, 68 estudios fueron excluidos por no cumplir con los criterios de elegibilidad, principalmente por tratar temáticas no relacionadas, corresponder a tipos de publicación no pertinentes o presentar acceso restringido. Finalmente, se seleccionaron 38 estudios para su análisis cualitativo a profundidad, como se observa en la **Figura 1**.

De los artículos seleccionados se extrajo información relacionada con la definición de reproducibilidad, las herramientas bioinformáticas empleadas, el tipo de flujo de trabajo descrito, su aplicación en el ámbito clínico u oncológico y las buenas prácticas de reproducibilidad reportadas. Los datos obtenidos se organizaron en tablas comparativas utilizando hojas de cálculo, lo que facilitó el análisis temático de la información.

2.5 Consideraciones éticas

En el presente estudio no se realizaron tomas de muestras en seres humanos, sin embargo, se llevó un proceso automatizado y transparente para seleccionar y utilizar la bibliografía de interés.

2.6 Limitaciones

La mayoría de estudios y publicaciones sobre reproducibilidad en el campo de la bioinformática es publicada a partir de 2017, por lo que limitó la búsqueda más antigua de bibliografía, por otro lado, una limitación relevante (motivo que llevó a hacer el presente trabajo) es la falta de información sobre flujos de trabajo aplicados a genómica y oncología clínica, por lo que se limitó a utilizar los que actualmente existen.

3. Resultados

De los 38 estudios finalmente incluidos, la mayoría correspondió a artículos originales ($n = 24$), seguidos por revisiones bibliográficas ($n = 10$) y reportes técnicos o metodológicos ($n = 4$). En conjunto, los trabajos analizados se centraron principalmente en tres ejes temáticos: la definición de la reproducibilidad, la identificación de buenas prácticas en el diseño y uso de flujos de trabajo bioinformáticos y la aplicación clínica de pipelines en el ámbito de la genómica y la oncología.

Definición y características de reproducibilidad

La reproducibilidad en flujos de trabajo bioinformáticos se considera una buena práctica que garantiza la seguridad y confianza de los resultados generados. Consiste en la capacidad de utilizar códigos, datos y herramientas de estudios anteriores para poder llegar a los mismos resultados o conclusiones (7).

No existe un concepto universal y estandarizado para definir reproducibilidad aplicada a genómica clínica; es por eso que existen diferentes definiciones. En el estudio de Flier et al., sugieren la reproducibilidad de los métodos como una descripción detallada de la metodología para que sus resultados sean repetidos fácilmente (8). De acuerdo a Gundersen et al., se indica que la reproducibilidad del resultado es la misma del experimento original (9). De la misma manera, Kanwal et al. sugieren que una característica de reproducibilidad es la repetibilidad, es decir, la elaboración repetitiva de un análisis o experimento, y que estos lleguen al mismo resultado. Mientras que Yang et al. definen la replicabilidad como el proceso de emplear los mismos códigos, pero con datos diferentes (10).

Buenas prácticas en reproducibilidad

Un total de 18 estudios describieron de manera explícita buenas prácticas orientadas a garantizar la reproducibilidad. Entre los aspectos reportados con mayor frecuencia destacaron la documentación detallada de los flujos de trabajo ($n = 16$), el uso de gestores de flujos como Snakemake ($n = 14$), la implementación de control de versiones y contenedores ($n = 11$) y la compartición abierta de código y datos ($n = 9$). Estos elementos fueron sistematizados en la Figura 1 y Tabla 1, donde se sintetizan los cinco pilares de la reproducibilidad computacional y su aplicación en el ámbito de la bioinformática (11–17):

Tabla 1: Se describen diferentes recomendaciones de buenas prácticas en bioinformáticas tomadas de varios autores.

Pilares	Características
Programación Literaria	Combinación de texto y código que sea entendible. Posibilitan el origen de los resultados y permiten ahorrar tiempo. Además, se acompañan de figuras, gráficos, explicaciones y tablas.
Control de versiones de código	Compartir el código garantiza la reproducibilidad y transparencia. Git es un sistema de control de versiones distribuido (DVCS), cuyas ventajas están en evitar la pérdida de código, facilitación de gestión y colaboración de trabajos, mantenimiento del historial, así como ayudan a la corrección de errores de código.
Control de entorno computacional	Para llevar un orden adecuado, se recomienda registrar e informar las versiones de software y paquetes empleados. Docker es un sistema de contenedores popular que permite crear, ejecutar y gestionar. Cabe recalcar en este aspecto de genómica clínica que el Colegio Americano de Patólogos (CAP) recomienda el uso de contenedores como una buena práctica para el control de calidad en laboratorios diagnóstico y análisis bioinformáticos.
Intercambio persistente de datos	En genómica clínica es importante garantizar la reproducibilidad a partir del intercambio de datos bajo políticas y organismos internacionales, es así que existen diferentes repositorios de intercambio de datos biológicos como el European Nucleotide Archive (ENA), Gene Expression Omnibus (GEO) o Sequence Read Archive (SRA).
Documentación	La documentación del proceso de un flujo de trabajo es necesaria para que otros investigadores entiendan y puedan replicar los análisis realizados. Como se indicó anteriormente, se recomienda adjuntar el material de apoyo como protocolos, versiones de software, códigos, etc.

Nota: Adaptado de Ziemann, M. 2023.

Flujos de trabajo bioinformáticos

Un total de 12 estudios describieron flujos de trabajo bioinformáticos aplicados a genómica clínica, de estos, 7 estudios utilizaron Snakemake como gestor principal. Un flujo de trabajo bioinformático consiste en una secuencia de pasos estructurada y automatizada para analizar y procesar grandes volúmenes de datos de origen biológico, como por ejemplo secuencias de ácido desoxirribonucleico (ADN), ácido ribonucleico (ARN), proteínas, etc. Cada paso del flujo de trabajo se encarga de ejecutar una tarea específica y genera archivos que serán utilizados en pasos posteriores (18). Es así que un flujo de trabajo reproducible, suele implementar herramientas de gestión de flujos, como por ejemplo Snakemake, muy comúnmente empleada en bioinformática. Snakemake permite que los trabajos sean reproducibles y escalables; está basado en el lenguaje de programación de Python y su estructura básica general consta de reglas (rules), archivo Snakefile y grafo acíclico dirigido (DAG) (19).

Los flujos de trabajo en Snakemake son ejecutados mediante reglas que se especifican en el archivo Snakefile. De manera general, su sintaxis consiste en archivos de entrada, archivos de salida y el comando Shell donde se generan las instrucciones. Una regla se considera como la unidad básica en los flujos de trabajo de Snakemake, estas indican las instrucciones de lo que hay que hacer, de dónde y qué recursos utilizar (20). A continuación, en la Tabla 2 se muestra la sintaxis general, mencionada previamente (20).

Tabla 2: Sintaxis general de Snakemake, donde se visualizan las reglas generadas a partir de archivos de entrada y de salida.

Sintaxis
<pre>rule myrule: input: "path/to/inputfile", "path/to/other/inputfile", output: "path/to/outputfile", "path/to/another/outputfile", shell: "Somecommand {input} {output}"</pre>

Fuente: Snakemake Documentation, Snakefiles and Rules, disponible en línea.

En el campo de genómica clínica, los flujos de trabajo permiten transformar datos crudos de secuenciación a información biológica y clínicamente relevante. Un flujo de trabajo típico inicia con la evaluación de la calidad de las lecturas provenientes del secuenciador, continua con el alineamiento de las lecturas a un genoma de referencia, seguido por la identificación y anotación de variantes genéticas y culmina con la generación de un reporte clínico que facilita un diagnóstico genético eficaz y preciso (21).

En términos prácticos, un flujo de trabajo genómica se realiza mediante herramientas bioinformáticas de manera sistemática ordenada. Los datos crudos de secuenciación, se almacenan en archivos de formato FASTQ, seguido, la evaluación y filtrado de calidad de adaptadores se realiza con herramientas como FastQC, fastp o MultiQC. El alineamiento de lecturas al genoma de referencia se ejecuta con Bowtie2 o BWA, en donde se generan archivos SAM (*Sequence Alignment Map*) /BAM (*Binary Aligned Mapped*), los cuales contienen las posiciones exactas de cada fragmento alineado (22). Seguidamente, en la etapa de post-alineamiento, los archivos se ordenan para eliminar elementos duplicados mediante la herramienta Picard MarkDuplicates para luego iniciar el llamado de variantes con FreeBayes o GATK (23).

Finalmente, se producen archivos en formato VCF (*Variant Call Format*), los mismos que contienen variantes de interés detectadas como inserciones, deleciones o SNPs (*Single Nucleotide Polimorphism*). Estos resultados son útiles para el diagnóstico y toma de decisiones en medicina personalizada (24). A continuación, se describen brevemente dos flujos de trabajo bioinformáticos aplicados a oncología clínica:

Onkopipe

Es un sistema de secuenciación basada Snakemake útil en el análisis de variantes clínicas en medicina de precisión e identificación de biomarcadores genéticos. Su aplicación permite detectar variantes de número de copias (CNV), variantes estructurales (SV) y variantes de nucleótido único (SNV) (3). El flujo de trabajo de Onkopipe consta de tres etapas:

- 1. Preprocesamiento de datos crudos:** Este primer paso, inicia con el control de calidad, donde se eliminan las bases de baja calidad mediante la herramienta TrimGalore que incorpora Cutadapt y FastQC. A continuación, con BWA-MEM se alinean las lecturas al genoma de referencia humano generando así archivos SAM, los cuales se convierten a BAM mediante Picard y GATK. Una vez convertidos, se ordenan e indexan con Picard MarkDuplicates y Samtools y se puntúa la calidad de base con GATK.
- 2. Descubrimiento de variantes:** El llamado de variantes de SNV se realiza con GATK Mutect2. Para SNV, la detección de duplicaciones o eliminaciones de secciones de genoma se ejecuta CNVkit. Mientras que, para SV, la detección de alteraciones genómicas como inserciones, eliminaciones, inversiones y translocaciones se utiliza LUMPY-SV.
- 3. Post-procesamiento y uso clínico:** Una vez realizado el llamado de variantes de interés, la concatenación de las mismas da como resultado un archivo unificado de salida VCF. En la Figura 3, se muestra el flujo de trabajo de Onkopipe ya descrito anteriormente (3):

iCOMIC

iCOMIC es una herramienta para el análisis de datos ómicos de cáncer. Su flujo permite analizar datos de secuenciación de genoma completo (DNA-Seq) y transcriptoma (RNA-Seq). Sus algoritmos incorporados permiten predecir la patogenicidad de las mutaciones causantes de cáncer, así

como diferenciar genes supresores de tumores y oncogenes. El flujo de trabajo se resume de la siguiente manera (6):

- Entrada de datos crudos en formato FASTQ
- Control de calidad con FastQC
- Alineación de secuencias contra el genoma de referencia mediante BWA-MEM, GEM- Mapper o Bowtie2.
- Llamado de variantes con Samtools, GATK, Haplotype-Caller, FreeBayes y GATK Mutect2.
- Anotación de variantes con ANNOVAR y SnpEff
- Finalmente, para análisis en cáncer se utilizan NBDriver para identificación de mutaciones y cTaG para la predicción de genes supresores de tumores.

Como se muestra en la **Figura 3**, el flujo de trabajo de iCOMIC, descrito:

»»» 4. Discusión

La documentación de los flujos de trabajo utilizados en genómica clínica y oncológica sigue siendo limitada. En los últimos años, la reproducibilidad de estos análisis ha adquirido gran relevancia, pues aplicar buenas prácticas garantiza la obtención de resultados confiables. Además, promueve la estandarización de los procesos y la solidez de los hallazgos, aspectos esenciales para avanzar en la precisión y calidad de los estudios genómicos (9).

Existe literatura sobre flujos de trabajo reproducibles en otras áreas de la bioinformática y ciencias básicas, sin embargo, en genómica clínica es limitada, esto debido a que institutos y empresas privadas resguardan la privacidad de los datos y resultados obtenidos por su alta sensibilidad (25). Es así que en el estudio de Dodani y Talhouk publicado en la revista Clinical Cancer Research, indican que la baja reproducibilidad y ausencia de estándares bioinformáticos generan resultados poco consistentes en el estudio de la microbiota vaginal y cáncer ginecológico (26).

De la misma manera, en el estudio de Shifang y Jürgen enfocado a genómica clínica, se encontró que la falta de análisis reproducibles es responsable del fracaso farmacológico en ensayos clínicos de enfermedad de hígado graso no alcohólico (esteatohepatitis alcohólica) (27). Por otro lado, la revisión de la literatura muestra que actualmente no existe una definición universal y estandarizada de reproducibilidad bioinformática; en el trabajo de Pan et al. se menciona la ausencia

de un único concepto de reproducibilidad aplicada a la detección de variantes genéticas (28).

En contraste, existen varias definiciones de reproducibilidad bioinformática; en el estudio de Baykal et al. la definen como la cualidad de las herramientas bioinformáticas para mantener resultados coherentes en todas las réplicas y que además es vital para generar conocimiento científico y médico (29). Por otro lado, en el estudio publicado por Keenum et al. la definen como la capacidad de ejecutar independientemente un análisis previamente publicado y llegar así a resultados y conclusiones que sean comparables (30).

Por consiguiente, se destacan varios estudios donde la falta de reproducibilidad llevó al fracaso de los trabajos. Ziemann et al. describen un estudio llamado “Genomic signatures to guide the use of chemotherapeutics” el cual fue retractado en 2011 por que no fue posible reproducir los experimentos de validación para predecir respuesta a tratamiento de quimioterapia. De la misma manera, en el trabajo de Kanwal et al. mencionan que en un estudio publicado en 2012 donde se revisaron 299 artículos que citaron al proyecto “1000 Genomes”, solo cuatro aplicaron el flujo de trabajo completo (7).

Considerando el enfoque clínico, el Colegio Americano de Patólogos (CAP) cuenta con una guía de recomendación que abarca las fases de una prueba de NGS (Next Generation Sequencing) que incluyen diseño, optimización, validación y gestión de calidad, es decir desde la correcta selección de genes, elección de una prueba que garantice la cobertura de secuenciación, validación en la extracción de ADN e informe clínico final. La gestión de calidad monitoriza el flujo de trabajo para generar correctamente la reproducibilidad en análisis genómico (31).

En este contexto, Snakemake se ha consolidado como uno de los flujos de trabajo más representativos en el ámbito de la bioinformática. Su diseño intuitivo, flexible y robusto facilita la automatización de tareas complejas, garantizando la reproducibilidad, trazabilidad e interoperabilidad de los análisis bioinformáticos. Asimismo, herramientas como Onkopipe e iCOMIC han sido desarrolladas con el propósito de optimizar y simplificar los procesos que abarcan desde el control de calidad de las secuencias hasta la identificación de variantes genéticas y la generación de reportes clínicos finales, contribuyendo así a mejorar la eficiencia

y precisión en los entornos de genómica clínica y oncológica (32).

La evidencia analizada es heterogénea y predominantemente descriptiva, con limitaciones en términos de estandarización y validación clínica. Asimismo, este estudio, al tratarse de una revisión bibliográfica, se basó exclusivamente en literatura publicada, lo que podría haber excluido trabajos no divulgados por restricciones asociadas a la confidencialidad de los datos clínicos. Aunque el énfasis en herramientas basadas en Snakemake se justifica por su amplia adopción en genómica clínica, la ausencia de comparaciones experimentales directas con otros gestores como Nextflow o CWL, podría introducir sesgos (25).

Si bien guías internacionales, como las del College of American Pathologists, proporcionan marcos normativos robustos, su implementación en la práctica local sigue siendo inestable, lo que refuerza la necesidad de una adopción progresiva, así como de futuros estudios comparativos y guías de implementación estandarizadas.

»» 5. Conclusiones

La reproducibilidad bioinformática es la capacidad de que los flujos de trabajo sean robustos, replicables, trazables, escalables y transparentes a partir de mismos parámetros y muestras para así llegar a los mismos resultados. Actualmente existe una crisis de reproducibilidad en el campo de la genómica clínica, lo que imposibilita la generación de resultados e informes clínicos consistentes, así mismo no se cuenta con una definición estandarizada y universal de reproducibilidad en genómica clínica. Es importante que los flujos de trabajo en genómica clínica adopten las buenas prácticas en reproducibilidad.

Los hallazgos principales de esta revisión demuestran que la combinación de una documentación detallada, el uso de sistemas de control de versiones, la contenerización de software y la automatización de los análisis constituyen los pilares fundamentales para construir un ecosistema de genómica clínica robusto y reproducible.

»» 6. Agradecimientos

Los autores expresan su agradecimiento a la Facultad de Salud Pública, Carrera de Medicina, por brindar el espacio y las facilidades necesarias

para la difusión del conocimiento generado en la presente revisión.

7. Declaración de conflicto de interés

Los autores declaran no tener ningún conflicto de interés relacionado con el desarrollo o la publicación de este trabajo.

8. Declaración de responsabilidad

El contenido y las conclusiones presentadas en este estudio son de exclusiva responsabilidad de los autores.

9. Fuentes de apoyo

La presente investigación fue autofinanciada por los autores, sin recibir apoyo económico de instituciones públicas o privadas.

10. Referencias bibliográficas

- Canzoneri R, Lacunza E, Abba MC. Genómica y bioinformática como pilares de la medicina de precisión en oncología. *Medicina (B Aires)*. 2019;79(6/1):587–92.
- Del Pozo A. Bioinformática y gestión de datos ómicos en diagnóstico genético. *An Pediatr (Engl Ed)*. 2025;103:504013. doi:10.1016/j.anpedi.2025.504013
- Yang J, Beißbarth T, Dönitz J. Onkopipe: A Snakemake Based DNA-Sequencing Pipeline for Clinical Variant Analysis in Precision Medicine. *Stud Health Technol Inform [Internet]*. el 12 de septiembre de 2023 [citado el 1 de noviembre de 2025];307:60–8. Disponible en: <https://pubmed.ncbi.nlm.nih.gov/37697838/>
- Fernandez Isern G. Herramientas informáticas para la bioinformática [Internet]. Fundació Universitat Oberta de Catalunya (FUOC); 2023 [citado 2025 Nov 1]. Available from: <https://eines-informatiques.recursos.uoc.edu/workflows/es/>
- Yang J. An automated data integration platform for interpreting genomic data and reporting treatment options in molecular tumor boards [Internet]. 2024 [citado 2025 Nov 1]. Available from: <https://ediss.uni-goettingen.de/handle/11858/15432>
- Anilkumar Sithara A, Maripuri DP, Moorthy K, Amirtha Ganesh SS, Philip P, Banerjee S, et al. iCOMIC: a graphical interface-driven bioinformatics pipeline for analyzing cancer omics data. *NAR Genom Bioinform*. 2022;4(3). Available from: <https://pubmed.ncbi.nlm.nih.gov/35899080/>
- Ziemann M, Poulain P, Bora A. The five pillars of computational reproducibility: bioinformatics and beyond. *Brief Bioinform*. 2023;24(6):bbad375. doi:10.1093/bib/bbad375
- Flier JS. The Problem of Irreproducible Bioscience Research. *Perspect Biol Med [Internet]*. el 1 de junio de 2022 [citado el 1 de noviembre de 2025];65(3):373–95. Disponible en: <https://muse.jhu.edu/pub/1/article/863666>
- Gundersen OE. The fundamental principles of reproducibility. *Philos Trans A Math Phys Eng Sci*. 2021;379(2197). doi:10.1098/rsta.2020.0210
- Kim YM, Poline JB, Dumas G. Experimenting with reproducibility: a case study of robustness in bioinformatics. *Gigascience*. 2018;7(7):giy077. doi:10.1093/gigascience/giy077
- Ziemann M, Poulain P, Bora A. The five pillars of computational reproducibility: bioinformatics and beyond. *Brief Bioinform [Internet]*. el 22 de septiembre de 2023 [citado el 1 de noviembre de 2025];24(6):1–13. Disponible en: <https://dx.doi.org/10.1093/bib/bbad375>
- Mattevi S, Mazzarotto F, Martini P. Allele-specific expression analysis: pipelines, applications, challenges, and unmet needs. *Comput Biol Med*. 2025;196:110890
- Kadri S, Sboner A, Sigaras A, Roy S. Containers in bioinformatics: applications, practical considerations, and best practices in molecular pathology. *J Mol Diagn*. 2022;24(5):442–54.
- Vallet N, Michonneau D, Tournier S. Toward practical transparent verifiable and long-term reproducible research using Guix. *Sci Data*. 2022;9(1):1–9.
- Git Project. Git documentation [Internet]. [citado 2025 Nov 1]. Available from: <https://git-scm.com/docs/gi>

16. Cadwallader L, Gabhann FM, Papin J, Pitzer VE. Advancing code sharing in the computational biology community. *PLoS Comput Biol*. 2022;18(6):e1010193
17. Granger BE, Perez F. Jupyter: thinking and storytelling with code and data. *Comput Sci Eng*. 2021;23(2):7–14.
18. O'Brien J, Mitchell C, Auerbach S, Doonan L, Ewald J, Everett L, et al. Bioinformatic workflows for deriving transcriptomic points of departure: current status, data gaps, and research priorities. *Toxicol Sci*. 2025;203(2):147–59.
19. Mölder F, Jablonski KP, Letcher B, Hall MB, van Dyken PC, Tomkins-Tinch CH, et al. Sustainable data analysis with Snakemake. *F1000Res*. 2025;10:33.
20. Snakemake developers. Snakemake documentation [Internet]. [citado 2025 Nov 2]. Available from: <https://snakemake.readthedocs.io/en/stable/index.html>
21. Baykal PI, Łabaj PP, Markowetz F, Schriml LM, Stekhoven DJ, Mangul S, et al. Genomic reproducibility in the bioinformatics era. *Genome Biol*. 2024;25(1):1–15.
22. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–90.
23. Broad Institute. Picard tools [Internet]. [citado 2025 Nov 2]. Available from: <https://broadinstitute.github.io/picard/>
24. Rapti M, Zouaghi Y, Meylan J, Ranza E, Antonarakis SE, Santoni FA. CoverageMaster: comprehensive CNV detection and visualization from NGS short reads for genetic medicine applications. *Brief Bioinform*. 2022;23(2):bbac049.
25. Manolio TA, Fowler DM, Starita LM, Haendel MA, MacArthur DG, Biesecker LG, et al. Building bridges between basic and clinical genomic research. *Cell*. 2017;169(1):6–12.
26. Dodani DD, Talhouk A. Assessing the reproducibility crisis in vaginal microbiome studies for clinical applications in endometrial cancer. *Clin Cancer Res*. 2024;30(5 Suppl):A023.
27. Tang S, Borlak J. Genomics of human NAFLD: lack of data reproducibility and high interpatient variability in drug target expression as major causes of drug failures. *Hepatology*. 2024;80(4):901–15.
28. Pan B, Ren L, Onuchic V, Guan M, Kusko R, Bruinsma S, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. *Genome Biol*. 2022;23(1):2.
29. Baykal PI, Łabaj PP, Markowetz F, Schriml LM, Stekhoven DJ, Mangul S, et al. Genomic reproducibility in the bioinformatics era. *Genome Biol* [Internet]. el 9 de agosto de 2024 [citado el 9 de noviembre de 2025];25(1):213. Disponible en: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03343-2>
30. Keenum I, Jackson SA, Eloë-Fadrosch E, Schriml LM. A standards perspective on genomic data reusability and reproducibility. *Front Bioinform*. 2025;5:1572937.
31. Cokelaer T, Cohen-Boulakia S, Lemoine F. Reprohackathons: promoting reproducibility in bioinformatics through training. *Bioinformatics*. 2023;39(Suppl_1):i11–20.
32. Kulkarni N, Alessandrì L, Panero R, Arigoni M, Olivero M, Ferrero G, et al. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinform*. 2018;19(10):5.