Revista Ingeniería Biomédica
ISSN 1909-9762 / E-ISSN 1909-9991
Volumen 13 / Número 26 / julio-diciembre de 2019 / pp.71-80
Universidad EIA / Envigado, Colombia

Polimorfismos de inserción de elementos transponibles ligados a cáncer de mama: una prueba de concepto

Nicolas Tobon-Orozco[1, Ψ], Mariana S. Candamil-Cortés[1], Johan S. Piña[1], Simón Orozco-Arias[2,3], Reinel Tabares-Soto[4], Romain Guyot[4,5], Cristian F. Jiménez-Varon[6], Pérez Agudelo J.M.[7]

[1] *Semillero de investigación en bioinformática e inteligencia artificial, Universidad Autónoma de Manizales – Facultad de Ingeniería*

[2] *Departamento Ciencias Computacionales, Universidad Autónoma de Manizales Facultad de Ingeniería*

[3] *Departamento de Sistemas e Informática, Universidad de Caldas – Facultad de Ingeniería*

[4] *Departamento de electrónica y automatización, Universidad Autónoma de Manizales Facultad de ingeniería*

[5] *Institut de Recherche pour le Développement, CIRAD, University Montpellier, France.*

[6] *Departamento Física y matemáticas, Universidad Autónoma de Manizales – Facultad de Ingeniería.*

[7] *Universidad de Manizales, Facultad de ciencias de la salud, escuela de medicina, grupo de investigación médica*

Recibido 2 de febrero de 2019. Aceptado 29 de julio de 2019

***Resumen*—**Los retrovirus endógenos humanos (HERVs) constituyen aproximadamente el 8 % del genoma humano, particularmente están sobreexpresados en algunas células y tejidos del carcinoma de mama que es el más común y la segunda causa de muerte por cáncer en mujeres en todo el mundo. Investigaciones recientes muestran que la familia de retrovirus HERV-K es la de más expresión génica sobre el cáncer de mama. Los elementos HERV-K 108, 109, 113 y 115 son los más comúnmente encontrados en esta enfermedad y están ubicados a nivel genómico en el cromosoma 6, 7, 8, y 19 respectivamente. La liberación de datos genómicos de pacientes patológicamente identificados con esta enfermedad ha permitido avances en aspectos de origen, desarrollo y diagnóstico. Sin embargo, el tratamiento que se da a esta información se hace desde enfoques biológicos y de poco contacto computacional (*in silico*). Dado esto, se plantea la siguiente investigación a escala genómica enfocada en encontrar nuevos polimorfismos de inserción que puedan estar ligados a cáncer de mama, a través de TRACKPOSON, un software diseñado

Ψ Dirección para correspondencia: nicolas.tobono@autonoma.edu.co
DOI: https://doi.org/10.24050/19099762.n26.2019.1401

para la detección de polimorfismos de inserción de elementos transponibles (TIPs). Este pipeline fue utilizado en 3.000 genomas de arroz, pero su funcionalidad fue extrapolada a este proyecto usando el genoma humano de referencia, una base de datos de HERVs y lecturas de secuenciación del genoma de pacientes casos (con cáncer de mama) y controles (personas sin la enfermedad) que fueron obtenidos usando la tecnología Illumina. Finalmente, los resultados obtenidos *in silico* arrojan TIPs asociados a cáncer de mama con su ubicación cromosómica, por lo que este "mimetismo bioinformático", podría ofrecer mejora en los métodos de investigación y de enfoque diagnóstico de esta enfermedad.

***Palabras clave*—** bioinformática, cáncer de mama, HERV-K, retrovirus, polimorfismos de inserción.

Transposable Elements Insertion Polymorphisms Linked To Breast Cancer: A Proof Of Concept

***Abstract*—**Human endogenous retroviruses (HERVs) make up approximately 8% of the human genome, particularly overexpressed in some cells and tissues of breast carcinoma which is the most common and second leading cause of cancer death in women worldwide. Recent research shows that the HERV-K family of retroviruses is the most widely expressed family of genes in breast cancer. HERV-K elements 108, 109, 113, and 115 are the most found in this disease and are located at the genomic level on chromosome 6, 7, 8, and 19, respectively. The release of genomic data from patients pathologically identified with this disease has allowed advances in aspects of origin, development, and diagnosis. However, the treatment given to this information is done from biological approaches and with little computational contact (*in silico*). Thus, the following research at genomic scale is focused on finding new insertion polymorphisms that may be linked to breast cancer, through TRACKPOSON, a software designed for the detection of transposable element insertion polymorphisms (TIPs). This pipeline was used in 3,000 rice genomes, but its functionality was extrapolated to this project using the reference human genome, a database of HERVs and genome sequencing reads of case (breast cancer) and control (disease-free) patients that were obtained using Illumina technology. Finally, the results obtained in silico show TIPs associated with breast cancer with its chromosomal location, so this "bioinformatic mimicry" could offer improvement in the research methods and diagnostic approach of this disease.

***Keywords*—** bioinformatics, breast cancer, HERV-K, retrovirus, insertion polymorphisms.

Polimorfismos de inserção de elementos transponíveis ligados ao cancro da mama: uma prova de conceito

***Resumo*—**Os retrovírus endógenos humanos (HERV) constituem aproximadamente 8% do genoma humano, particularmente superexpressos em algumas **células e** tecidos do carcinoma da mama, que é a causa mais comum e a segunda principal causa de morte por cancro nas mulheres em todo o mundo. Pesquisas recentes mostram que a família HERV-K de retrovírus é a família de genes mais amplamente expressa no cancro da mama. Os elementos HERV-K 108, 109, 113, e 115 são os mais frequentemente encontrados nesta doença e estão localizados a nível genómico nos cromossomas 6, 7, 8, e 19 respectivamente. A divulgação de dados genómicos de pacientes patologicamente identificados com esta doença permitiu avanços em aspectos de origem, desenvolvimento e diagnóstico. No entanto, o tratamento dado a esta informação é feito a partir de abordagens biológicas e com pouco contacto computacional (em silico). Tendo isto em conta, propõe-se a seguinte investigação à escala genómica, centrada na descoberta de novos polimorfismos de inserção que podem estar ligados ao cancro da mama, através do TRACKPOSON, um software concebido para a detecção de polimorfismos de inserção de elementos transponíveis (TIPs). Esta conduta foi utilizada em 3.000 genomas de arroz, mas a sua funcionalidade foi extrapolada para este projecto utilizando o genoma humano de referência, uma base de dados de HERVs e leituras de sequenciação de genoma de pacientes com casos (com cancro da mama) e controlos (pessoas sem a doença) que foram obtidos utilizando a tecnologia Illumina. Finalmente, os resultados obtidos em silico mostram DICAS associadas ao cancro da mama com a sua localização cromossómica, pelo que esta "mímica bioinformática" poderia oferecer uma melhoria nos métodos de investigação e abordagem diagnóstica desta doença.

***Palavras-chave*—**bioinformática, cancro da mama, HERV-K, retrovírus, polimorfismos de inserção.

I. Introducción

Los elementos transponibles (TEs) son componentes no estáticos del genoma, pues tienen la capacidad de moverse de una ubicación cromosómica a otra [1]–[3], específicamente, en los seres humanos, existe un tipo de estos elementos que se denominan retrovirus endógenos humanos (HERVs) y constituyen aproximadamente el 8 % del genoma humano [4]. Los HERVs poseen una estructura genómica conformada por: 5'LTR □ gag □ pro □ pol □ env □ 3'LTR, donde cada uno de los genes (gag, pro, pol y env) [5], es capaz de codificar para diferentes proteínas que cumplen roles específicos [6] que son necesarios estructural y funcionalmente para el retrovirus [7].

Se ha demostrado que existe una sobreexpresión de los HERVs de tipo K en algunos tipos de cáncer como el melanoma [8], próstata [9], páncreas [10], mama [11] y ovario [12]; en particular, este proyecto se enfoca en el cáncer de mama, ya que es la segunda causa de muerte por cáncer entre las mujeres después del cáncer de pulmón [13]. Un estudio realizado para la identificación y clasificación de HERVs presentes en 512 pacientes que padecen cáncer de mama, reporta que, en el subtipo basal, se observa la prevalencia de HERV-K108, HERV-K109, HERV-K113 y HERV-K115 en los cromosomas 6, 7, 19 y 8 respectivamente [11].

La diversificación del genoma y la variación adaptativa del organismo, se han convertido en la principal razón para análisis genéticos, en los que se han encontrado que los polimorfismos de inserción (TIPs) podrían ser la expresión más frecuente en el genoma que den razón de la existencia de estos fenómenos [14]. Sin embargo, la complejidad del análisis *in vivo* de TIPs es alta, por la cual el uso de programas bioinformáticos resulta necesario para analizar *in silico* la variación de TIPs en el comportamiento de una enfermedad mediada por agentes genéticos, lo que traduce en un enfoque genómico potencial para la investigación en este tipo de procesos [15].

Actualmente existen algunos softwares de identificación de TIPs (algunos de ellos evaluados en [16]) tales como Jitterbug [17] y BreakDancer [18] que resultan ineficientes al momento de analizar masivamente datos de gran demanda de espacio en disco. Además, realizan un mapeo de todas las lecturas de secuenciación en una secuencia del genoma de referencia, el cual representa un gasto computacional significativo. Por esta razón Carpentier *et al.* implementaron el pipeline TRACKPOSON [14], que realiza un mapeo de todas las lecturas en cada familia de TEs, obtenidas mediante secuencia de consenso, logrando así, optimizar el gasto computacional de manera parcial. Este software fue desarrollado para la detección de TIPs en 3.000 genomas de arroz, pero su funcionalidad podría ser extrapolada a datos humanos derivados de genómica del cáncer, utilizando un genoma humano de referencia, una base de datos de HERVs y lecturas de secuenciación del genoma tanto de pacientes casos como controles.

El presente proyecto está enfocado en la realización de una prueba de concepto que permita identificar actividad polimórfica de retrovirus relacionada a cáncer de mama, pasando de los estudios actuales que están enfocados en genes o regiones específicas, a una escala genómica. Además de aproximar locus de posibles genes o marcadores asociados, apuntando a futuras aplicaciones bioinformáticas a gran escala y ofreciendo mejora en los métodos de investigación y de enfoque diagnóstico de esta enfermedad.

II. Metodología

De manera inicial se realizó una búsqueda bibliográfica en la cual se identificó el tipo de cáncer más relevante sobre el cual se extrapolaría el software para la identificación de TIPs [19], teniendo en cuenta que el objetivo de esta investigación era la aplicación de la técnica propuesta por [14] en genomas humanos relacionados con cáncer.

Obtención de datos

Se ejecutó una búsqueda sistemática en la base de datos SRA (NCBI) usando la cadena de búsqueda presentada en (1) donde se obtuvieron 512 resultados con lecturas de secuenciación de pacientes con cáncer de mama. El listado obtenido se filtró para almacenar aquellos que fueron obtenidos por la tecnología Illumina y que las secuencias estuvieran disponibles para descargar, quedando aproximadamente 300 conjunto de datos de lecturas de secuenciación. Para ejecutar esta prueba de concepto se seleccionaron únicamente el 10 % de las secuencias debido a la gran cantidad de información que conlleva la secuenciación de genomas humanos con un peso por paciente entre 150 y 200 gigas.

((((((((((breast cancer) AND Homo sapiens[Organism]) AND PAIRED[Layout])) AND GENOMIC[Source])) AND WGS[Strategy]))) NOT exome)) (1)

El siguiente script de (2), muestra la manera secuencial en que se descargaron los datos del NCBI a través del programa fastq-dump [20].

```
#! /bin/bash
for i in `cat archivo.txt`
do                                                    (2)
   fastq-dump -X 30000000 --split-files $i -v
done
```

Para el caso de pacientes sanos (controles) se usaron datos secuenciados del proyecto "*Understanding the genetic architecture of schizophrenia in Chinese population*" [15], en el que las lecturas fueron de pacientes con esquizofrenia sin antecedentes relacionados a cualquier otra enfermedad, de los cuales se utilizaron 30.

Tanto para los pacientes casos como controles se usó como máximo 30 millones de lecturas, debido a la gran demanda en almacenamiento que necesitaba cada paciente, sin embargo, fueron necesarias 1.1 TeraBytes para almacenar las lecturas, específicamente 642 GB para lecturas correspondientes a controles y 458 GB para los casos.

De manera posterior se elaboraron las librerías (bases de datos con las secuencias) tanto de los HERVs y cromosomas relacionados a cáncer de mama, y el consenso de información de los 60 pacientes que fueron analizados que respectivamente se encuentran en el dato suplementario S1.

Finalmente se descargó el genoma humano de referencia (3.3 Gb) en formato fasta de (https://www.ncbi.nlm.nih.gov/genome/51?genome_assembly_id=582967) y su anotación en formato GFF para comparaciones de localidad, posicionamiento y existencia de TIPs con respecto a las lecturas de cada grupo de pacientes.

Software y requisitos para ejecución

El software TRACKPOSON [14], es un pipeline compuesto por un conjunto de programas bioinformáticos. Este pipeline realiza la identificación de TIPs para TEs mediante la técnica de creación de secuencia consenso para cada tipo de TE que se analizará, hace división del genoma de referencia en ventanas para comparación, lo que explica que TRACKPOSON no hace énfasis en todas las variaciones estructurales del organismo, de manera preferencial identifica la actividad transposicional de TIPs pertenecientes a un tipo específico de TEs.

Para su ejecución fue necesaria la instalación de varios programas que son prerrequisito: Bowtie 2 [21], Samtools [22], Bedtools [23], NCBI-Blast [24], Emboss (versión 6.6.0), Perl (versión 3.20.3) y Bash (4.1.2).

Junto a los elementos anteriores, se tuvieron en cuenta las siguientes medidas para ejecutar TRACKPOSON:

1. Crear una secuencia consenso por cada tipo de HERV.

2. Contar cuantas copias tiene cada tipo de HERV en el genoma de referencia.

3. Crear ventanas de 10 kb del genoma de referencia.

4. Formatear la base de datos de NCBI-Blast con el genoma de referencia.

5. Formatear la base de datos de TEs con Bowtie2.

Cada uno de los pasos anteriores generó una ruta que se especificó en TRACKPOSON ya que fueron usadas durante todo el proceso de análisis. Posteriormente los datos de entrada a TRACKPOSON son los archivos en formato fastq de las lecturas de pacientes caso y control en sus dos formas, *forward* y *reverse*, teniendo un total respectivamente de 60 archivos por clasificación de pacientes.

Hardware y recursos informáticos

El pipeline se ejecutó en un servidor de 24 procesadores (CPUs) Intel(R) Xeon(R) CPU E5645, 47 Gb de memoria RAM y sistema CentOS Linux versión 7.2.1511**.** Este servidor está gestionado por el calendarizador Slurm [25].

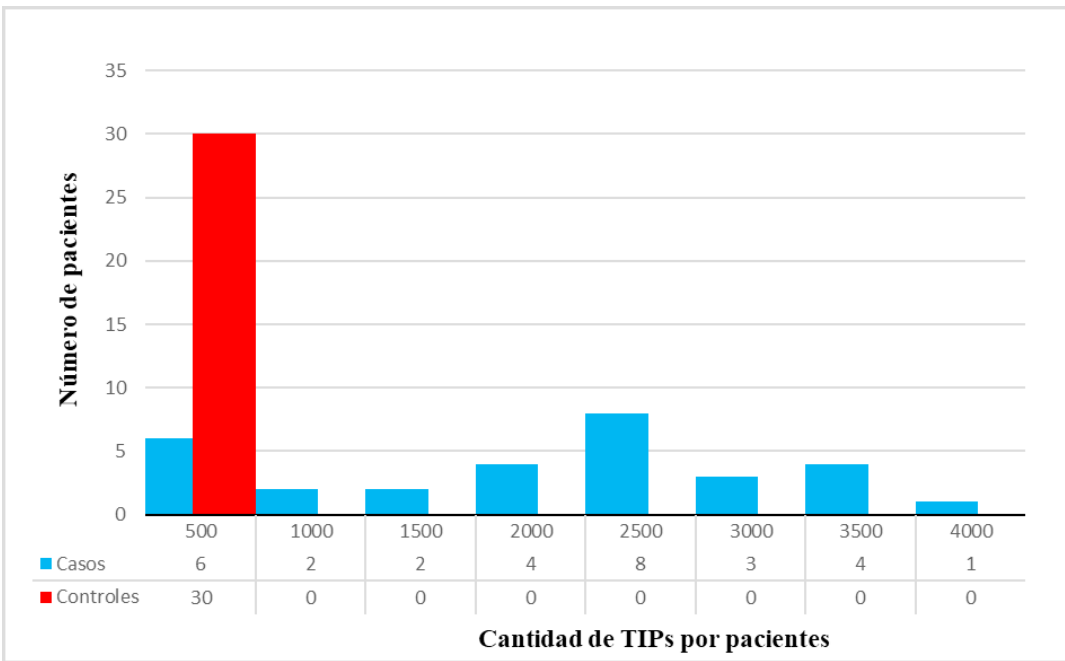

Fig. 1. Cantidad de TIPs en pacientes casos y controles. Se identifican los pacientes control (rojo) y casos (azul), el eje X representa la cantidad de TIPs y el eje Y el número de pacientes que corresponde.

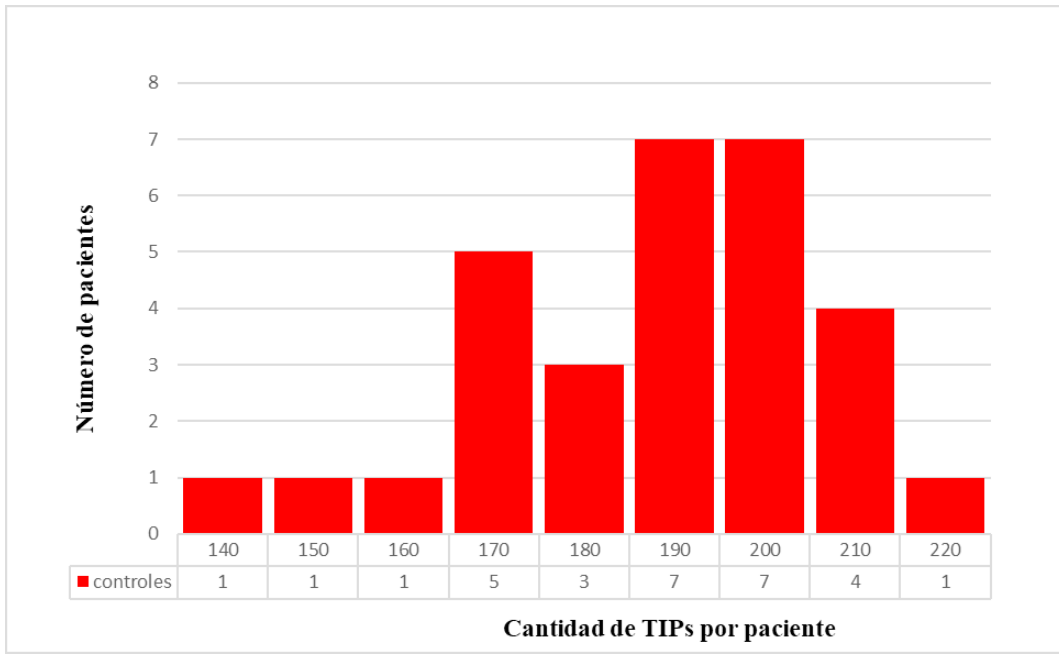

Fig. 2. Cantidad de TIPs en controles. El eje X representa la cantidad de TIPs y el eje Y el número de pacientes relacionado.

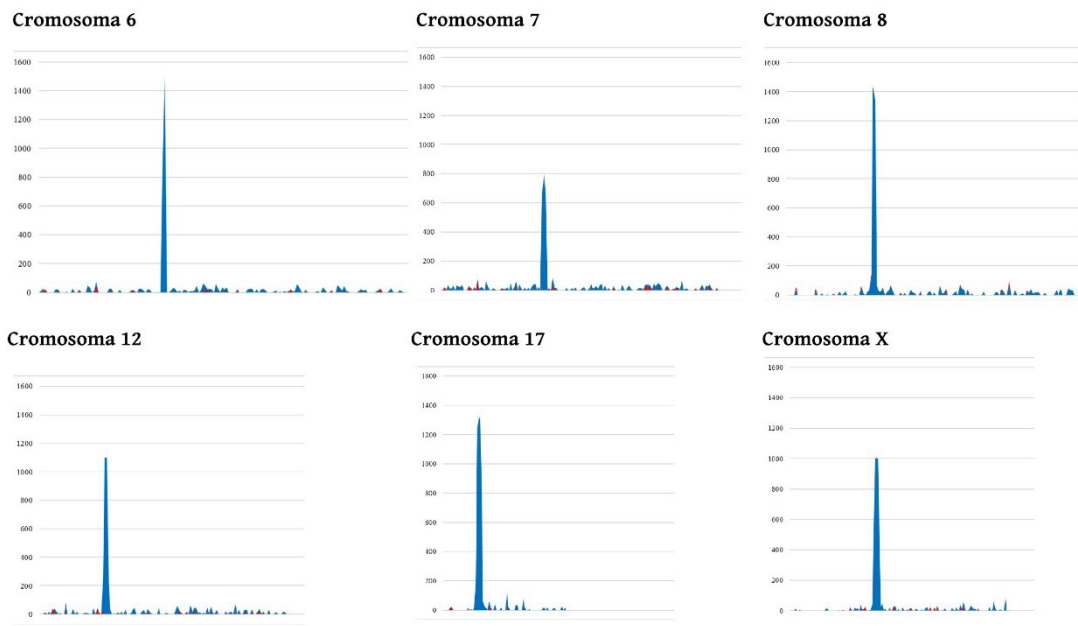

Fig. 3. Distribución de la cantidad de TIPs relacionados a HERVs en los cromosomas donde se observó mayor actividad transposicional. El eje X representa la longitud cromosómica en un orden desde 1,00 E+6 hasta 1,5 E+8, casos y controles están en color azul y rojo respectivamente. El eje Y representa la cantidad de TIPs a lo largo de la longitud cromosómica correspondiente.

III. Resultados

De acuerdo con la metodología descrita, se obtuvo una matriz de presencia/ausencia de TIPs para cada genoma, lo que sirvió para encontrar el número de polimorfismos tanto para casos como controles y se agruparon en rangos, de lo que se obtuvo la distribución de frecuencias para los pacientes casos y controles expuesta en la Fig. 1. Lo que de manera preliminar permite aproximar la relación entre el cáncer de mama y la presencia de actividad polimórfica relacionada a HERVs

presente en cada grupo, ello por la distribución atípica de los casos sobre todo el rango (0 a 4000) de la gráfica, mientras que los controles solo tienen una distribución de cantidad de TIPs en un rango corto de (0 a 500).

Por otra parte, la Fig. 2 muestra en detalle la distribución de pacientes control sobre el rango (0-500) y la cantidad de TIPs asociada, esto sugiere una actividad transposicional (polimórfica) mucho mayor en pacientes casos en relación con los controles.

TRACKPOSON requiere de gran cantidad de tiempo para generar resultados, debido a que requiere de un alto espacio en disco y además de recursos computacionales de altas prestaciones para una sola ejecución, puesto que para la cantidad de información utilizada y el hardware del cual se dispuso, demanda alrededor de 7 u 8 horas en ejecución por paciente, debido a que almacena los resultados del NCBI-Blast en su formato estándar y los procesa usando un script en Perl; esto demuestra que, aunque es uno de los algoritmos más rápidos para detectar inserciones de TEs [16], su rendimiento se ve afectado cuando es usado en experimentos en grandes volúmenes de datos, lo que podría sugerir cambios o retos futuros en la creación de herramientas relacionadas.

Finalmente, se agruparon las frecuencias de los TIPs para ambos grupos (casos y controles), de acuerdo con su posición estimada en cada uno de los cromosomas, en 300 rangos. Cabe recordar que TRACKPOSON dividió el genoma de referencia en ventanas de a 10.000 Kb, por lo tanto, las posiciones de cada TIP estaban dadas en intervalos de la longitud de la ventana. Este agrupamiento se realizó con el objetivo de encontrar secciones de los cromosomas en donde es más frecuente la inserción de los retrovirus usados en esta investigación. La figura 3 muestra la distribución por frecuencias de la presencia de TIPs en seis de los 23 cromosomas humanos, en donde claramente los picos de actividad transposicional (en términos de TIPs) es más significativa, específicamente los cromosomas 6, 7, 8, 12, 17 y X. sin embargo en el dato suplementario S2 se presenta la gráfica de distribución para cada uno de los 23 cromosomas.

IV. Discusión

El análisis de TIPs es una ruta eficiente para obtener información acerca de las dinámicas evolutivas en el genoma de los organismos, específicamente a entender como muchas enfermedades han aparecido y derivado en generaciones [14], [26], [27]. Específicamente el estudio de HERVs alrededor de investigaciones que buscan solución a preguntas de origen, mutación, permanencia, resistencia e interacciones biológicas de enfermedades como el cáncer de mama, han permitido desde los laboratorios desarrollar métodos efectivos para combatir este tipo de enfermedades [28]. Sin embargo esta ruta de investigación ha reportado resultados a escala micro, es decir desde análisis *in vivo* y con una selección de datos reducida, debido a que la mayoría de análisis deben estar supervisados y analizados de manera manual, aunque softwares como TRACKPOSON [14] han innovado en la manera de buscar TIPs asociados a una condición específica, en este caso extrapolado en cáncer de mama, el tiempo y gasto computacional requerido para los análisis son elevados y resulta ineficiente para obtener datos simultáneos con la obtención del genoma y su análisis a nivel *in silico.* Lo que sugiere una mejora a nivel informático del software y una optimización a través de programación paralela, debido a que se ha demostrado que esta técnica puede arrojar ventajas en otras tareas en la genómica [29], con el fin de poder usar genomas de cualquier organismo (incluso de cualquier tamaño genómico).

Por otra parte, en la literatura se reportan que los HERVs asociados con el cáncer de mamá, se encuentran en los cromosomas 6, 7, 8 y 19 [9]. En los resultados obtenidos por TRACKPOSON, se hallaron picos elevados (entre 1.200 y 1.600 TIPs), que podrían estar ligados con el cáncer de mama debido a que solo se presentaron en los pacientes caso. Sin embargo, estos resultados deben ser sujetos a mayor validación a través de técnicas estadísticas y experimentos con una fusión de la genómica y la interpretación médica. Estos picos se observaron en los cromosomas 6, 7, 8, 12, 17 y X, (Fig. 3). Los retrovirus reportados en la literatura en los cromosomas 6, 7 y 8 que están correlacionados con el tipo basal de cáncer de mama [11], podrían estar representados en los picos mostrados en los mismos cromosomas de la Fig. 3. De igual manera, se identificó una posible ubicación de TIPs en otros 3 cromosomas no reportados hasta ahora (12, 17 y X), lo anterior podría deberse a variaciones estructurales en el genoma del paciente [14], las cuales no son tratadas en esta prueba de concepto.

Metodologías de análisis de enfermedades humanas como la expuesta en este trabajo con el cáncer de mama, son innovadoras y coherentes con la constante demanda de ser más rápidos que la enfermedad. Esto debido a que, al ser un análisis netamente computacional, se puede estudiar al individuo como tal o en un grupo en tiempos reducidos y determinar su comportamiento a determinados fármacos, si es propenso o no a sufrir alguna patología o simplemente para conocer la predisposición de su organismo a presentar algún fenómeno por cambios del ambiente o por ingestión de algún alimento o medicamento.

El impacto que genera este estudio está a la vanguardia de la era post genómica [30], de la economía cada vez más demandante de procesos médicos y biológicos de bajo costo, pero efectivos, de esta manera se erradica la brecha de la poca interacción de las tecnologías informáticas

y computacionales para el estudio de enfermedades humanas u otro tipo de calificativo medico relacionado a un organismo, sin depender, edad, raza, sexo y lo más importante condición socioeconómica, lo que potencia una calidad de vida innegable a través de las ciencias de la vida y la bioinformática.

V. Conclusión

Es posible usar la técnica propuesta en TRACKPOSON para identificar TIPs que puedan estar asociados al cáncer de mama. Pero debido al gran tamaño de datos que se generan en los procesos de secuenciación de ADN humano, analizar una muestra poblacional significativa es muy costoso computacionalmente tanto en espacio en disco como en tiempo de ejecución. Sin embargo, la metodología expuesta en esta investigación podría abrir el camino para la aproximación *in silico* de la identificación de nuevos polimorfismos causados por retrovirus que estén relacionados con el cáncer de mama.

Agradecimientos

Los autores reconocemos al IRD itrop HPC (South Green Platform) del IRD en Montpellier, Francia por proporcionar recursos de HPC que han contribuido a los resultados de la investigación reportados en este trabajo. URL: https://bioinfo.ird.fr/ - http://www.southgreen.fr. Reconocemos y agradecemos igualmente la colaboración de la Universidad Autónoma de Manizales por permitir el uso de sus instalaciones para el desarrollo de esta investigación dentro del marco del proyecto 589-089.

Referencias

[1]. O. Ridge, J. Genetics, L. Cells, J. Reed, M. Lefort, and E. Sgourakis, "(or recognition)," 1950.

[2]. G. Bourque *et al.*, "Ten things you should know about transposable elements," *Genome Biol.*, vol. 19, no. 1, p. 199, 2018.

[3]. S. Orozco-arias, G. Isaza, R. Guyot, and R. Tabares-soto, "A systematic review of the application of machine learning in the detection and classi fi cation of transposable elements," *PeerJ*, vol. 7, p. 18311, 2019, doi: 10.7717/peerj.8311.

[4]. R. P. Subramanian, J. H. Wildschutte, C. Russo, and J. M. Coffin, "Identification , characterization , and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses," pp. 1–22, 2011.

[5]. E. R. Havecker, X. Gao, and D. F. Voytas, "The diversity of LTR retrotransposons," *Genome Biol.*, vol. 5, no. 6, p. 225, 2004, doi: 10.1186/gb-2004-5-6-225.

[6]. J. Arango-López, S. Orozco-Arias, J. A. Salazar, and R. Guyot, "Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case," in *Advances in Computing*, vol. 735, 2017, pp. 156–170.

[7]. M. Zhang and J. Q. Liang, "Expressional activation and functional roles of human endogenous retroviruses in cancers," no. November 2018, pp. 1–11, 2019, doi: 10.1002/rmv.2025.

[8]. R. H. Yolken, H. Karlsson, F. Yee, N. L. Johnston-Wilson, and E. F. Torrey, "Endogenous retroviruses and schizophrenia," *Brain Res. Rev.*, vol. 31, no. 2–3, pp. 193–199, 2000, doi: 10.1016/S0165-0173(99)00037-5.

[9]. G. Ding, F. Liu, C. Feng, J. Xu, and Q. Ding, "Asociación entre los polimorfismos de genes de mieloperoxidasa y la susceptibilidad a cáncer de próstata: Un estudio caso-control en la población de nacionalidad china," *Actas Urol. Esp.*, vol. 37, no. 2, pp. 79–82, Feb. 2013, doi: 10.1016/j.acuro.2012.03.020.

[10]. M. Li *et al.*, "Downregulation of Human Endogenous Retrovirus Type K (HERV-K) Viral env RNA in Pancreatic Cancer Cells Decreases Cell Proliferation and Tumor Growth," vol. 23, no. 43, 2017, doi: 10.1158/1078-0432.CCR-17-0001.

[11]. G. L. Johanning *et al.*, "Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype.," *Sci. Rep.*, vol. 7, no. 1, p. 41960, Dec. 2017, doi: 10.1038/srep41960.

[12]. L. Cegolon, C. Salata, E. Weiderpass, P. Vineis, G. Palù, and G. Mastrangelo, "Human endogenous retroviruses and cancer prevention : evidence and prospects," 2013.

[13]. C. E. Desantis, "Breast Cancer Statistics , 2017 , Racial Disparity in Mortality by State," vol. 67, no. 6, pp. 439–448, 2017, doi: 10.3322/caac.21412.

[14]. M. C. Carpentier *et al.*, "Retrotranspositional landscape of Asian rice revealed by 3000 genomes," *Nat. Commun.*, vol. 10, no. 1, 2019, doi: 10.1038/s41467-018-07974-5.

[15]. L. N. Al-Eitan, B. H. Al-Ahmad, and F. A. Almomani, "The Association of IL-1 and HRAS Gene Polymorphisms with Breast Cancer Susceptibility in a Jordanian Population of Arab Descent: A Genotype-Phenotype Study.," *Cancers (Basel).*, vol. 12, no. 2, Jan. 2020, doi: 10.3390/cancers12020283.

[16]. P. Vendrell-Mir, F. Barteri, M. Merenciano, J. González, J. M. Casacuberta, and R. Castanera, "A benchmark of transposon insertion detection tools using real data," *Mob. DNA*, vol. 10, no. 1, pp. 1–19, 2019, doi: 10.1186/s13100-019-0197-9.

[17]. E. Hénaff, L. Zapata, J. M. Casacuberta, and S. Ossowski, "Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution," *BMC Genomics*, vol. 16, no. 1, p. 768, Dec. 2015, doi: 10.1186/s12864-015-1975-5.

[18]. X. Fan, T. E. Abbott, D. Larson, and K. Chen, "BreakDancer: Identification of genomic structural variation from paired-end read mapping," *Curr. Protoc. Bioinforma.*, vol. 45, no. 1, pp. 15–16, 2014.

[19]. N. Bannert, H. Hofmann, A. Block, and O. Hohn, "HERVs New Role in Cancer : From Accused Perpetrators to Cheerful Protectors," vol. 9, no. February, pp. 1–8, 2018, doi: 10.3389/fmicb.2018.00178.

[20]. S. Sherry, "NCBI SRA Toolkit Technology for Next Generation Sequence Data." Plant and Animal Genome.

[21]. B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, vol. 9, no. 4, p. 357, 2012.

[22]. H. Li *et al.*, "The Sequence Alignment / Map format and SAMtools," vol. 25, no. 16, pp. 2078–2079, 2009, doi: 10.1093/bioinformatics/btp352.

[23]. A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[24]. S. F. Altschup, W. Gish, T. Pennsylvania, and U. Park, "Basic Local Alignment Search Tool 2Department of Computer Science," pp. 403–410, 1990.

[25]. M. Jette and M. Grondona, "SLURM: Simple Linux Utility for Resource Management," *Clust. Conf. Expo CWCE*, vol. 2682, pp. 44–60, 2003, doi: 10.1007/10968987.

[26]. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase Update, a database of eukaryotic repetitive elements," *Cytogenet. Genome Res.*, vol. 110, no. 1–4, pp. 462–467, 2005, doi: 10.1159/000084979.

[27]. K. Ruprecht *et al.*, "Human Endogenous Retrovirus Family HERV-K (HML-2) RNA Transcripts Are Selectively Packaged into Retroviral Particles Produced by the Human Germ Cell Tumor Line Tera-1 and Originate Mainly from a Provirus on Chromosome 22q11 . 21 □ †," vol. 82, no. 20, pp. 10008–10016, 2008, doi: 10.1128/JVI.01016-08.

[28]. R. Contreras-galindo *et al.*, "Human Endogenous Retrovirus K (HML-2) Elements in the Plasma of People with Lymphoma and Breast Cancer □ †," vol. 82, no. 19, pp. 9329–9336, 2008, doi: 10.1128/JVI.00646-08.

[29]. S. Orozco-Arias, R. Tabares-Soto, D. Ceballos, and R. Guyot, "Parallel Programming in Biological Sciences, Taking Advantage of Supercomputing in Genomics," in *Advances in Computing*, vol. 735, A. Solano and H. Ordoñez, Eds. Zurich: Springer, 2017, pp. 627–643.

[30]. O. Lecompte, J. D. Thompson, F. Plewniak, J.-C. Thierry, and O. Poch, "Multiple alignment of complete sequences (MACS) in the post-genomic era," *Gene*, vol. 270, no. 1, pp. 17–30, 2001, doi: https://doi.org/10.1016/S0378-1119(01)00461-9

Datos suplementarios

Dato suplementario 1

Lista de acceso para casos

Identificador	Cantidad de lecturas	Cantidad de bases	Tamaño	Estudio o Bioproyecto
ERR232239	21,851,875	3.3G	1.6Gb	Barber LJ *et al.*, "Comprehensive genomic analysis of a BRCA2 deficient human pancreatic cancer.", *PLoS One*, 2011;6(7):e21639
ERR232240	23,189,082	3.5G	1.7Gb	
ERR232241	22,981,389	3.5G	1.7Gb	
ERR232242	22,621,836	3.4G	1.6Gb	
ERR232243	22,639,118	3.4G	1.6Gb	
ERR232244	22,261,938	3.4G	1.5Gb	
ERR232245	21,527,631	3.3G	1.4Gb	
ERR232246	10,516,200	1.6G	1.1Gb	
ERR232247	10,430,921	1.6G	1.1Gb	
ERR232248	10,198,798	1.6G	1.1Gb	
ERR232249	9,719,346	1.5G	1.1Gb	
ERR232250	9,473,836	1.4G	1Gb	
ERR232251	9,749,436	1.5G	1.1Gb	
ERR232252	9,571,910	1.5G	1Gb	
SRR1513864	176,642,889	35.3G	19.7Gb	**PRJNA253369**
SRR1513865	133,306,847	26.7G	15.3Gb	
SRR3090651	60,374,706	11.8G	7.5Gb	**PRJNA308098**
SRR3090701	67,475,052	13.2G	8.5Gb	
SRR3090702	67,950,524	13.3G	8.6Gb	
SRR3090705	53,908,355	10.6G	6.8Gb	
SRR3090707	37,496,500	7.3G	4.7Gb	
SRR3090723	43,821,190	8.6G	5.5Gb	

SRR944977	184,296,718	36G	21.3Gb	**PRJNA213134**
SRR944978	181,844,287	36.4G	21.6Gb	
SRR944979	182,902,360	36.6G	21.8Gb	
SRR944980	181,022,543	36.2G	21.5Gb	
SRR944981	181,130,440	36.2G	21.5Gb	
SRR944982	185,812,676	37.2G	22.1Gb	
SRR944983	182,795,197	36.6G	21.7Gb	
SRR944984	182,335,586	36.5G	21.6Gb	

Lista de acceso para controles

Identificador	Cantidad de lecturas	Cantidad de bases	Tamaño	Estudio o Bioproyecto
SRR9649373	370,350,202	111.1G	44.5Gb	PRJNA551447
SRR9649374	331,534,072	99.5G	39.4Gb	
SRR9649375	331,534,072	99.5G	39.4Gb	
SRR9649376	311,053,284	93.3G	39.4Gb	
SRR9649377	301,783,741	90.5G	38.6Gb	
SRR9649378	298,399,259	89.5G	37.8Gb	
SRR9649379	322,309,329	96.7G	40.3Gb	
SRR9649380	326,703,844	98G	41.5Gb	
SRR9649381	319,430,885	95.8G	40.6Gb	
SRR9649382	313,457,184	94G	38.8Gb	
SRR9649383	316,140,982	94.8G	38.8Gb	
SRR9649384	353,893,374	106.2G	45Gb	
SRR9649385	378,484,273	113.5G	47.7Gb	
SRR9649386	340,592,921	102.2G	42.6Gb	
SRR9649387	321,505,765	96.5G	39.8Gb	
SRR9649388	361,660,610	108.5G	43.2Gb	
SRR9649389	388,987,793	116.7G	46.1Gb	
SRR9649390	327,405,862	98.2G	41.4Gb	
SRR9649391	316,923,845	95.1G	39.5Gb	
SRR9649392	348,472,163	104.5G	43.6Gb	
SRR9649393	357,743,831	107.3G	42.5Gb	
SRR9649394	315,378,602	94.6G	38.6Gb	
SRR9649395	817,072,486	122.6G	58.8Gb	
SRR9649396	303,008,395	90.9G	38.3Gb	
SRR9649397	340,865,126	102.3G	42.9Gb	
SRR9649398	385,029,280	115.5G	47.5Gb	
SRR9649399	314,613,420	94.4G	41.4Gb	
SRR9649400	310,583,203	93.2G	39.4Gb	
SRR9649401	339,111,152	101.7G	43.9Gb	
SRR9649402	338,916,948	101.7G	41.6Gb	
SRR9649403	327,357,465	98.2G	42.4Gb	

Dato suplementario 2