UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES IZTACALA

Análisis de mutagénesis insercional causada por los elementos móviles transponibles *Alu* en paneles de genes de pacientes de América Latina con SHCMO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

PRESENTA:

EDUARDO EMILIO CÓRDOBA GARCÍA

DIRECTOR DE TESIS:

DR. FELIPE VACA PANIAGUA

IZTACALA

LOS REYES IZTACALA, EDO. DE MÉXICO, 2024

AGRADECIMIENTOS

Este trabajo fue realizado en el Laboratorio 13 de la Unidad de Biomedicina de la Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México. Agradezco a mi tutor, el Dr. Felipe Vaca y a los miembros de mi comité asesor: Dr. Héctor Martínez, Dra. Clara Díaz, Dra. Carolina Rodríguez y Dra. Gloria Paniagua. En la parte técnica agradezco a: Dr. Aldo De la Cruz, Dr. Miguel Ruiz, Laura Hernández, Gabriela Reséndiz, Luis Borroel y Fernando Ambriz.

AGRADECIMIENTOS A TÍTULO PERSONAL

A mis padres Zulema y Santos y a mis hermanitos Mariana y Eduardo, por todo el amor y la confianza.

A mis abuelos, tíos y primos por su apoyo incondicional a lo largo de los años.

A mis roomies de Sabino, mis amigos de Iztacala, Xalapa y mis foráneos, personas que nunca voy a olvidar.

A la M.C. Andrea Farías por transmitirme su amor por la ciencia.

A todos aquellos que no dudaron en abrazarme el día que apagaron la luz.

ÍNDICE

I. ABREVIATURAS DE GENES

AKT: Serina/treonina cinasa

ATM: Ataxia Telangiectasia Mutada Serina/treonina cinasa

BRAF: Protooncogén B-Raf, Serina/treonina cinasa

BARD1: Dominio RING 1 asociado a BRCA1

*BRCA1***:** Proteína de susceptibilidad al cáncer de mama tipo 1

*BRCA2***:** Proteína de susceptibilidad al cáncer de mama tipo 2

*BRIP1***:** Helicasa 1 que interactúa con BRCA1

*CDH1***:** Cadherina 1

CDK: Cinasa dependiente de ciclina

CDKN2A: Inhibidor 2A de la cinasa dependiente de ciclina

*CHEK1/2***:** Cinasa de punto de control ½

FANCA/B/C/D/D2/F/G/I/M/L: Grupo A/B/C/D/D2/F/G/I/M/L de complementación de la FA

*EPCAM***:** Molécula de adhesión de células epiteliales

ERBB2/HER2: erb-b2 receptor tirosina cinasa 2/ Receptor 2 del factor de crecimiento epidérmico humano

MLH1/2: MutL Homólogo ½

MSH1/2/6: MutS homólogo 1/2/6

mTOR: Objetivo Mecánico de la rapamicina cinasa

MRN: Complejo proteico compuesto por MRE11, RAD50 y NBN

*PMS1/2***:** PMS1 homólogo 1/2

*NBN***:** Nibrina

*NF1***:** Neurofibromina 1

*PALB2***:** Asociado y localizador de BRCA2

PI3K: Fosfatidilinositol 3-cinasa

PTEN: Fosfatasa y tensina homóloga

*RAD51C/D***:** RAD51C parálogo C/D

*STK11***:** Serina/treonina cinasa 11

*TP53***:** Proteína tumoral p53

II. ABREVIATURA DE SIGLAS

ACGM: Colegio Americano de Genética Médica

ASR: Tasa estandarizada por edad

CM: Cáncer de mama

CO: Cáncer de ovario

CNV: Variaciones en el número de copias

DNA: Ácido desoxirribonuclecio

dsDNA: Ácido desoxirribonuclecio de doble cadena

DSB: Ruptura de doble cadena

ERV: Retrovirus endógeno

FA: Anemia de Fanconi

HGSOC: Carcinoma histológicos seroso de alto grado

HR: Recombinación homóloga

L1 o LINE-1: Elemento intercalado largo-1

MMR: Mecanismo de reparación del DNA por bases mal apareadas

mRNA: Ácido ribonuclecio mensajero

NCCN: National Comprehensive Cancer Network
LTR: Estructuras largas de repetición terminal
NHEJ: Recombinación de extremos no homólogos
SHCMO: Síndrome hereditario de cáncer de mama y ovario
ssDNA: Ácido desoxirribonuclecio de una sola cadena
SVA: SINE-VNTR (número variable de repeticiones en tándem)-*Alu*s
SV: Variantes estructurales
ORF: Marco de lectura abierto
PARP: Poli ADP-ribosa
PB: Pares de bases
VP: Variante patogénica
RPA: Proteínas de replicación A
TE: Elementos móviles transponibles
TSD: Sitio de duplicación objetivo
TPRT: transcripción inversa dirigida
UTR: Región no traducida
VPP: Variante probablemente patogénica
VUS: Variantes de significado incierto
WES: Secuenciación de exoma completo
WGS: Secuenciación de genoma completo

III. LISTA DE FIGURAS

IV. LISTA DE TABLAS

RESUMEN

El cáncer de mama (CM) es la neoplasia maligna con mayor incidencia y mortalidad en mujeres mexicanas, en Colombia ocupa el primer lugar en incidencia y el tercero en mortalidad, en Perú el segundo y séptimo y en Guatemala el segundo y sexto. Por su parte, el cáncer de ovario (CO) es menos frecuente en estas poblaciones, sin embargo, su relevancia radica en su alta letalidad. La tasa de mortalidad de ambos países para dichos padecimientos muestra una tendencia al alta que es consistente con la proyección de casos de cáncer en América Latina para las próximas décadas.

El síndrome hereditario de cáncer de mama y ovario (SHCMO) representa entre el 5-10% y 15-18% de los casos de CM y CO, respectivamente. El SHCMO está asociado a variantes patogénicas germinales en genes de alto riesgo como *BRCA1/2* y *PALB2,* riesgo moderado como *ATM* y *CHEK2*, entre otros.

La frecuencia de variantes patogénicas (variantes de un solo nucleótido) en el SHCMO es del 15-20%. Por lo cual, es necesario estudiar otros mecanismos genéticos involucrados en esta condición genética. Los elementos móviles transponibles (TE), representan el 45% del genoma humano y han sido recientemente estudiados por la asociación de su actividad retrotransponible con múltiples enfermedades, incluyendo el cáncer. En las últimas decadas se han caracterizado inserciones de familias de estos elementos como *Alu*, L1 y SVA en los genes de alto riesgo mencionados, sin embargo, no se ha realizado un estudio a este tipo de inserciones en la población de América Latina.

En este trabajo se identificaron TEs presentes usando dos paneles de genes en un total de 1477 pacientes mexicanas, colombianas, peruanas y guatemaltecas: i) 143 genes de Qiagen y ii) 73 genes de SOPHiA Genetics a través de un análisis *in silico* con tres herramientas bioinformáticas. Ninguno de los eventos detectados con estas herramientas fue un verdadero positivo para la mutagénesis insercional causada por TEs, lo que podría atribuirse al tamaño de la muestra poblacional o una menor prevalencia de estos eventos en las poblaciones. Este resultado es concordante con la baja prevalencia de este tipo de inserciones en enfermedades genéticas.

ABSTRACT

Breast cancer (CM) is the malignant neoplasm with the highest incidence and mortality in mexican women, in Colombia it ranks first in incidence and third in mortality, in Peru second and seventh, and in Guatemala second and sixth. Ovarian cancer (CO) is less frequent in these populations; however, its relevance lies in its high lethality. The mortality rate of both

countries for these diseases shows a high trend that is consistent with the projection of cancer cases in Latin America for the coming decades.

Hereditary breast and ovarian cancer syndrome (SHCMO) accounts for between 5-10% and 15-18% of CM and CO cases, respectively. SHCMO is associated with germline pathogenic variants in high-risk genes such as *BRCA1/2* and *PALB2*, moderate risk such as *ATM* and *CHEK2*, among others.

The frequency of pathogenic variants (single nucleotide variants) in SHCMO is 15-20%. Therefore, it is necessary to study other genetic mechanisms involved in this genetic condition. Mobile transposable elements (TE) represent 45% of the human genome and have recently been studied because of the association of their retrotransposable activity with multiple diseases, including cancer. To date, insertions of families of these elements such as Alu, L1 and SVA have been characterized in the aforementioned high-risk genes; however, to date, no study has been done to determine the prevalence of this type of insertions in the Latin American population.

In this work, TEs were identified using two gene panels in a total of 1477 Mexican, Colombian, Peruvian and Guatemalan patients: i) 143 genes from Qiagen and ii) 73 genes from SOPHiA Genetics through in silico analysis with three bioinformatics tools. None of the events detected with these tools was a true positive for insertional mutagenesis caused by TEs, which may be due to the population sample evaluated and a low prevalence of these events in the populations studied. This result is consistent with the low prevalence of this type of insertions in genetic diseases.

1. INTRODUCCIÓN

1.1 Cáncer

El término cáncer define aquellas células que han adquirido la habilidad de invadir tejidos circundantes de células normales. Un neoplasma es cualquier crecimiento anormal de células, mientras que un tumor es un neoplasma que se asocia con un estado de enfermedad, dado que una población de células relacionadas genéticamente adquiere la habilidad de proliferar de manera anormal (Bunz, 2022).

Durante el desarrollo del cáncer, las células acumulan alteraciones genéticas y epigenéticas en los oncogenes y genes supresores de tumores que pueden tener un origen de línea germinal o somática (Fanciulli, 2010). Los tumores presentan diferentes sellos distintivos que son adquiridos durante su desarrollo como el mantenimiento de la señalización proliferativa, la evasión de los supresores del crecimiento, la resistencia a la muerte celular, la adquisición de la inmortalidad replicativa, la inducción de la angiogénesis y la activación de la invasión y metástasis, así como la inestabilidad del genoma que contribuye a la adquisición de mutaciones y a la progresión de la enfermedad (Hanahan y Weinberg, 2011).

De manera conjunta dichos sellos contribuyen a la alteración de genes involucrados en el crecimiento celular y aunque por sí sola la proliferación de las células tumorales no causa la mutación de los genes que controlan el crecimiento, sí propicia un entorno para que las células mutadas se dividan y adquieran nuevas mutaciones (Roses, 2005).

El desarrollo y la progresión del cáncer involucran su acumulación, ocasionando la salida de la homeostasis y la proliferación. Aquellas mutaciones que brindan una ventaja de crecimiento selectiva a la célula en el proceso se conocen como mutaciones conductoras, aunque también están presentes las pasajeras, mutaciones que no contribuyen directamente a la enfermedad (Wodarz *et al.,* 2018), es así que no todas las alteraciones genómicas contribuyen al desarrollo del cáncer, y esto depende del gen mutado, la naturaleza de la mutación y el potencial replicativo de la célula donde ocurre (Bunz, 2022).

Los tumores evolucionan de lesiones benignas a malignas mediante la adquisición de una serie de mutaciones a lo largo del tiempo, siendo la mayoría de las alteraciones genéticas presentes en un tumor de carácter puntual y una fracción menor alteraciones estructurales (Vogelstein *et al.,* 2013).

En la última década, la incidencia y la mortalidad del cáncer han aumentado considerablemente. En el año 2012 se estimó un total de 14.1 millones de casos y 8.2 millones

de muertes, mientras que en el 2018 fueron 18.1 y 9.6 millones, respectivamente (Torre *et al.,* 2015; Bray *et al.,* 2018).

Durante el año 2020, se estimaron 19.3 millones de casos nuevos de cáncer y 10 millones de muertes en todo el mundo. Se pronostica que para el 2040 los casos nuevos de cáncer llegarán hasta 28.4 millones, lo que supone un aumento del 47% comparado con los valores del 2020 (Sung *et al.,* 2021). De los casos de cáncer estimados en 2020, 1.4 millones ocurrieron en América Latina (Ferlay *et al.,* 2020).

En esta región, existen deficiencias debido a la concentración no equitativa de los servicios oncológicos, los retrasos en el diagnóstico, el acceso inadecuado a nuevos fármacos dirigidos, la falta de ensayos clínicos en oncología y el acceso limitado a pruebas moleculares y genómicas (Gössling *et al.,* 2023). Por ello, es preciso atender dichas problemáticas debido a que la proyección de casos en América Central y Sur para el 2030 estima 1.7 millones de nuevos casos y 1 millón de muertes por cáncer (Ferlay *et al.,* 2021).

Bajo el supuesto de una incidencia constante, tomando en cuenta el envejecimiento y el crecimiento demográfico, para 2040 se estiman hasta 2.4 millones de casos nuevos, con un incremento más marcado en Centroamérica y menor en el Caribe (Piñeros *et al.,* 2022).

1.2 Cáncer de mama (CM)

El CM es una enfermedad que resulta de la acumulación de alteraciones genéticas en los genes supresores de tumores y oncogenes que transforman las células epiteliales mamarias en un fenotipo maligno (Kashyap *et al.,* 2022).

El CM es altamente heterogéneo en términos moleculares, histológicos y clínicos (Rezai *et al.,* 2021). A nivel histológico se clasifica con base en el patrón de crecimiento tumoral e incluye dos subtipos principales: i) carcinoma lobular invasivo (8-10%) y ii) carcinoma ductal invasivo (70-80%). Este último comprende varios subtipos basados en criterios como su histoarquitectura, tipo celular, sitio de secreción y cantidad, así como estudios inmunohistoquímicos (Malik y Masood, 2022). Usualmente los tumores parten de la hiperproliferación ductal y luego se convierten en tumores benignos o bien, carcinomas metastásicos tras la proliferación constante de factores carcinogénicos (Sun *et al.,* 2017).

Clínicamente, el CM se clasifica con base a la expresión de tres receptores celulares: estrógeno, progesterona y HER2 (receptor 2 del factor del crecimiento humano) en los siguientes grupos, que además, poseen un perfil de expresión característico: i) epitelial basal, negativo en tres receptores mencionados ii) *ERBB2* sobreexpresado, positivo a HER2 iii)

normal de mama, con valores normales para los tres casos y iv) epitelio luminal positivo al receptor de estrógenos, que a su vez se divide en los subtipos luminal A y B, donde además de estrógeno y progesterona puede ser positivo a HER2 (Sørlie *et al.,* 2001).

Además de los receptores celulares mencionados, para la clasificación de CM también se toman en cuenta factores asociados con la proliferación medida con el marcador Ki-67 y el grado histológico, una evaluación semicuantitativa útil para el pronóstico que puede ser relacionada con la expresión de los receptores mencionados que es determinada por la evaluación de: i) la diferenciación tumoral, esto es, la proporción del tumor que corresponde a los microtúbulos que se forman ii) los pleomorfismos nucleares, cambios en su forma y tamaño y iii) la proliferación o tasa de mitosis (Eliyatkın *et al.*, 2015; Rezai *et al.,* 2021). Por otra parte, las vías de señalización involucradas en el CM son: los receptores del estrógeno y HER2-neu, vía canónica Wnt/β-catenina, CDK, señalización Notch, Sonic-Hedgehog, cinasa del tumor de mama y la vía PI3K/AKT/mTOR (Bose *et al.,* 2022).

El CM superó al cáncer de pulmón como el más comúnmente diagnosticado en 2020, con un estimado de 2.3 millones de casos nuevos, representando 11.7% de todos los casos de cáncer (Sung *et al.,* 2021). La tasa estandarizada por edad (ASR) resume el comportamiento que se habría observado si la población tuviera una estructura por edades estándar. El ASR (por cada 100,000 habitantes) fue de 40.5 en México, 48.3 en Colombia, 35.9 en Perú y 29.8 en Guatemala, valores de incidencia bajos comparados con Europa y Norteamérica, lo cual se mantiene al comparar con la mortalidad, donde obtuvieron valores de 10.6, 13.1, 9.1 y 7.2 respectivamente (Ferlay *et al.,* 2020).

Con 29,929 nuevos casos, el CM es el cáncer con mayor incidencia en mujeres mexicanas, ocupando a la vez el primer puesto en mortalidad con 7,931. En Colombia se estimaron 15,509 nuevos casos, ocupando también el primer puesto en incidencia, pero el tercero en mortalidad con 4,411. En Perú se estimaron 6,860 nuevos casos, ocupando el segundo lugar en prevalencia y el séptimo en mortalidad con 1,824. En Guatemala se estimaron 2,177 casos nuevos, posicionándolo en el segundo lugar de incidencia y el sexto en mortalidad con 521 (Ferlay *et al.,* 2020). Al respecto, las tasas de mortalidad han aumentado de forma constante en Brasil, Colombia, México y Ecuador, mientras que en el resto del mundo se han observado tendencias a la baja (Piñeros *et al.,* 2022).

Existen dos clasificaciones referentes para los factores involucrados en el aumento del riesgo de padecer CM: i) factores modificables como la nuliparidad, exposición a la terapia hormonal, obesidad, consumo de alcohol, tabaquismo, nula actividad física y una dieta no balanceada (Kashyap *et al.,* 2022) y factores no modificables como el sexo, edad, antecedentes

familiares, mutaciones genéticas, menstruación temprana y menopausia tardía (Sun *et al.,* 2017) y ii) factores extracelulares como la exposición a factores ambientales y factores intracelulares como errores en la síntesis del DNA (Roses, 2005).

Aunque mutaciones y amplificaciones en oncogenes y genes supresores de tumores son clave en el proceso de iniciación y progresión tumoral del CM (Sun *et al.,* 2017), esta etiología se atribuye a la compleja interacción entre estos factores de riesgo (Mayrovitz, 2022), muestra de ello es que el riesgo acumulado de desarrollar CM a los 80 años es del 72% para las portadoras de *BRCA1* y 69% para *BRCA2*, mientras que para el CO es del 44% y 17% respectivamente en los genes mencionados (Kuchenbaecker *et al.,* 2017)

1.3 Cáncer de ovario (CO)

Los tumores primarios de ovario se clasifican en tres categorías principales: tumores epiteliales de ovario, de línea germinal y de los cordones sexuales estromales, siendo la primera categoría la más prevalente. Se ha propuesto que gran parte de los casos tienen un origen extraovárico, esto es, del epitelio de las trompas de falopio, del tejido endometrial de la superficie ovárica, del tracto gastrointestinal o de manera inusual, de la superficie epitelial ovárica (Al Bakir y Gabra, 2014).

Los carcinomas se clasifican en subtipos histológicos: seroso de alto grado (HGSOC), endometrioide, de células claras, mucinoso y seroso de bajo grado. Lesiones en el extremo fimbriado de la trompa de falopio sugieren que el proceso neoplásico empieza con lesiones tubáricas que pasan al ovario donde progresa con agresividad. A pesar de que HGSOC es el subtipo más prevalente, no tiene un modelo de progresión tumoral como sí ocurre con los subtipos restantes, donde la progresión es escalonada (tumores tipo I seguidos de mutaciones en *BRAF* y *KRAS*) (Reid *et al.* 2017).

Debido a su desarrollo silencioso, cerca de tres cuartas partes de casos de cáncer de ovario (CO) se diagnostican en un estadio avanzado, lo cual se dificulta debido a que la manera más fácil para identificar la predisposición a la enfermedad son las pruebas genéticas, que resultan costosas y deben ser adaptadas a las mutaciones más comunes de cada población (Paulo dos Santos *et al.,* 2020).

El CO es uno de los cánceres ginecológicos más comunes, sólo después del cervical y uterino, sin embargo, entre ellos es el que tiene peor pronóstico y la mayor tasa de mortalidad. Por su parte, aunque tiene una prevalencia menor que el CM, es 3 veces más letal, y se

predice que para el año 2040 su mortalidad aumentará de manera significativa (Momenimovahed *et al.,* 2019).

La incidencia de CO en América Latina se mantuvo estable en el rango de edad superior a 60 años para la mayoría de los países, exceptuando Brasil (Paulo dos Santos *et al.,* 2020). Sin embargo, entre 1990 y 2012, la mayor tasa de incidencia de CO ocurrió en Cali (Colombia) y Goiânia, (Brasil) en un rango de edad entre 60-74 años que puntuó 35.4 y 26.4 casos por cada 100,000 habitantes.

El ASR de incidencia en 2020 muestra a la región europea y a Estados Unidos con los valores más altos (>6.8), mientras que Colombia mantiene 7.5, México 6.8, Perú 6.7 y Guatemala 3.4. En tanto a la mortalidad mantienen valores de 4.1, 4.5, 4.0 y 2.2, respectivamente, equiparables con Estados Unidos, pero no con Europa (Ferlay *et al.,* 2020). Aunque en ellos existe mayor incidencia que en América Latina, la mortalidad por CO en esta última es mayor (Paulo dos Santos *et al.,* 2020).

Para el caso de Colombia, en el 2020 la incidencia de CO fue de 2391 casos, ocupando el catorceavo sitio, mientras que, en mortalidad, ocupó el doceavo con 1485 casos. México ocupó los mismos puestos con 4963 y 3038 casos, respectivamente. Perú presentó 1275 nuevos casos posicionándolo en catorceavo lugar de incidencia y el treceavo en mortalidad con 786 casos. Por parte de Guatemala se presentaron 257 nuevos casos y 161 muertes, ocupando el doceavo puesto en incidencia y mortalidad (Ferlay *et al.,* 2020).

La incidencia se ha ligado al estado socioeconómico de las pacientes y también a sus patrones reproductivos. Los países industrializados suelen presentar tasas de incidencia más elevadas, lo cual ha sugerido que se debe a que en ellos las familias son más pequeñas, las dietas son ricas en grasas, la población es de mayor de edad y predominante caucásica. Por su parte, el uso prolongado de anticonceptivos orales, el parto y la lactancia causan una disminución de la ovulación, protegiendo contra el CO (Tollefsbol, 2021).

Dentro de los factores de riesgo para el desarrollo de CO se encuentra la edad avanzada, debido a que la incidencia incrementa significativamente a partir de la quinta década con un pico entre 80-84 años; la historia personal de CM, aumentando hasta 4 veces cuando se diagnostica antes de los 40, la historia familiar de CO, la presencia de mutaciones en *BRCA1/2*, nuliparidad, endometriosis y obesidad (Al Bakir y Gabra, 2014).

Las mutaciones en los genes mencionados y genes en MMR están relacionados con el riesgo genético de CO, vínculo genético conocido como SHCMO, que implica que a la edad de 70

años, 10-40% de las portadoras de las mutaciones desarrollarán neoplasias ováricas (Stewart *et al.* 2019).

1.4 Síndrome hereditario de cáncer de mama y ovario (SHCMO)

Los síndromes son factores genéticos que contribuyen al desarrollo de diferentes tipos de enfermedades, incluyendo el cáncer. Los síndromes hereditarios se caracterizan por la transmisión de mutaciones que predisponen a una alta probabilidad de desarrollar cáncer a través de un progenitor o la asociación con otros tipos de tumores en el individuo o su familia a una edad de diagnóstico más temprana que en el resto de la población (Rezai *et al.,* 2021).

La identificación de dichas mutaciones es clave en las perspectivas clínicas para vigilar y dar seguimiento a los pacientes afectados y a los portadores sanos (Fanale *et al.,* 2022). Varios síndromes hereditarios son causados por mutaciones inactivadoras en genes supresores de tumores que codifican proteínas implicadas en la restricción del crecimiento celular o en la regulación de la reparación del DNA. Aunque en menor medida, también se presentan mutaciones activadoras en oncogenes de proteínas receptoras cinasas o de señalización intracelular promotoras del crecimiento (Aref-Eshghi y Li, 2022).

El modelo de dos golpes propuesto por Knudson (ampliamente aplicable a síndromes de predisposición tumoral) hipotetiza que ambos alelos de un gen tienen que estar inactivos antes de la formación tumoral, con lo cual, la mutación hereditaria no es suficiente por sí sola, sino que se requiere además la inactivación del alelo silvestre (Knudson *et al.,* 1976).

Aunque la predisposición hereditaria a los tumores de mama o de ovario también se da en otras afecciones que predisponen al cáncer, como los síndromes de Li-Fraumeni, Peutz-Jeghers, Cowden y Lynch (Aref-Eshghi y Li, 2022), el síndrome de cáncer de mama y ovario hereditario (SHCMO) se caracteriza por una mayor susceptibilidad a padecer CM a una edad temprana (antes de los 50 años), CM bilateral, CM masculino, o CO a cualquier edad, así como cáncer de trompa de Falopio y tumor peritoneal primario (Fanale *et al.,* 2022).

El CM hereditario se diferencia del esporádico y familiar porque además de lo anterior involucra múltiples individuos afectados en múltiples generaciones, la presencia de otros cánceres menos comunes como el pancreático ligado a *BRCA1/2* y la herencia autosómica dominante de una variante patogénica (VP) o probablemente patogénica (VPP) en un gen de susceptibilidad al cáncer (Bellcross, 2022).

El SHCMO es una enfermedad autosómica dominante con penetrancia incompleta (Kobayashi *et al.,* 2013), lo que implica que la probabilidad del desarrollo de cáncer en los

portadores es variable, incluso en familias con la misma variante (Daly *et al.,* 2021). Además, se trata de un padecimiento con heterogeneidad a nivel del locus, es decir, que variantes en diferentes genes causan el mismo padecimiento. Este síndrome está asociado principalmente a variantes germinales en *BRCA1/2* con una frecuencia del 25%, aunque se han encontrado otros genes con una frecuencia menor tales como *ATM* (1.1%), *ATR* (0.15%), *BARD1* (0.15%), *CHEK2* (2.1%), *PALB2* (1.3%), *PTEN* (0.2%), *RAD51C* (0.29%), *TP53* (0.53%), entre otros (Nielsen *et al.,* 2016; Hoang y Gilks*,* 2018).

En el caso de *BRCA1/2* se presenta una alta penetrancia, esto es, una alta probabilidad de que se produzca una afección clínica cuando está presente el genotipo. Aunque la probabilidad de que los portadores adquieran cáncer es variable, el riesgo excesivo para el cáncer de mama y ovario justifica una detección intensiva (NCCN, 2022).

Una propuesta para explicar la predisposición al cáncer que afecta específicamente a la mama y al ovario es que a pesar de que las proteínas BRCA participan en todos los tipos celulares, en esos epitelios son especialmente vulnerables a la transformación cuando son heterocigotos para mutaciones de *BRCA* (Yoshida y Miki, 2004). Dado que los tejidos de mama y ovario comparten la característica de ser hormonalmente regulados, se ha especulado la interacción de las hormonas con la señalización *BRCA1/2* y su papel en el incremento del estrés oxidativo en el DNA con posibles implicaciones en el incremento de la susceptibilidad a mutaciones (Al Bakir y Gabra *et al.* 2014).

1.5 Criterios de selección de pacientes con SHCMO

La Red Nacional de Oncología Integral (NCCN, por sus siglas en inglés) previo a 2020 se había centrado en criterios de prueba para *BRCA1/2*, debido al incremento del riesgo que confiere para el desarrollo del CM y CO, sin embargo, en el 2021 se modificó dicha guía clínica incluyendo otros genes de susceptibilidad para CM como *CDH1, PALB2, PTEN* y *TP53*) (Daly *et al.,* 2021; NCCN, 2022). Los criterios de la NCCN del 2022 para el SHCMO incluyen a individuos con:

- Historia personal de CM con características específicas:
  - Por edad en el momento de diagnóstico:
    - ≤45 años
    - 46–50 años con cualquiera:
      - Antecedentes familiares desconocidos o limitados
      - Múltiples cánceres de mama primarios

- ≥1 pariente consanguíneo cercano con CM, ovario, páncreas o de próstata a cualquier edad
- ≥56 años
  - ≥1 pariente consanguíneo cercano con cualquiera:
    - ♦ CM con ≤50 años o CM masculino a cualquier edad
    - ♦ CO a cualquier edad
    - ♦ Cáncer pancreático a cualquier edad
    - ♦ Cáncer de próstata a cualquier edad: metastásico, de histología intraductal o en un grupo de riesgo alto o muy alto
  - ≥3 diagnósticos totales de CM en paciente o en pariente consanguíneo cercano
  - ≥2 parientes consanguíneos cercanos con CM o próstata a cualquier edad

➢ A cualquier edad:
- Para ayudar en las decisiones de tratamiento sistémico usando inhibidores PARP para el cáncer de mama metastásico
- Para ayudar en las decisiones de tratamiento adyuvante con olaparib para el CM de alto riesgo, HER-2 negativo
- CM triple negativo
- CM lobular con antecedentes personales o familiares de cáncer gástrico difuso

➢ Por ancestría
- Judíos Ashkenazi

❖ Antecedentes familiares de cáncer:
- ➢ Un individuo afectado (que no cumpla los criterios anteriores) o uno no afectado con un pariente consanguíneo en primer o segundo grado que cumpla alguno de los criterios (exceptuando los de toma de decisiones para tratamiento sistémico).
- ➢ Si el paciente tiene cáncer pancreático o de próstata sólo se debe ofrecer la prueba a familiares en primer grado
- ➢ Individuo afectado o no con >5% de probabilidad de una VP en *BRCA1/2* basada en modelos de probabilidad.

Respecto a los criterios de prueba para los genes de susceptibilidad al CO se menciona lo siguiente dentro de la versión mencionada anteriormente:

❖ Antecedentes personales de CO (incluyendo cáncer de trompa de falopio o cáncer peritoneal) a cualquier edad
❖ Historia familiar del cáncer:

- Un individuo no afectado con un pariente consanguíneo de primer o segundo grado con cáncer epitelial de ovario (incluyendo cáncer de trompa de falopio o cáncer peritoneal) a cualquier edad.
- Una persona no afectada que no cumple con los criterios anteriores, pero con >5% de probabilidad de una variante patogénica *BRCA1/2* basada en modelos de probabilidad.

Es necesario que las pruebas genéticas que se deriven de dicha selección tengan un seguimiento, ya que de ello depende si deberán considerarse otros miembros de la familia para realizarlas, las decisiones en el tratamiento de las pacientes, la vigilancia de componentes del síndrome genético, el pronóstico y la respuesta terapéutica (Bose *et al.,* 2022).

1.6 Principales genes asociados a la predisposición al SHCMO

La NCCN recomienda que se realicen pruebas en genes de alto, moderado y bajo riesgo para el padecimiento de CM, CO, cáncer pancreático, entre otros (Tabla 1).

Sin embargo, esto no resulta ser suficiente para identificar todos los casos con SHCMO, debido a que se ha encontrado que menos de la mitad de las pacientes con CM presentan una VP/VPP con las guías clínicas actuales y que el resto se distribuyen en otros genes de riesgo desconocido, por esa razón se sugieren pruebas en un panel de genes más amplio (Beitsch *et al.,* 2019). Pese a esta problemática, VP/VPP en genes de alta penetrancia como *BRCA1/2, PALB2* y *TP53* son identificadas comúnmente en pacientes que cumplen los criterios de la NCCN u otros (Bellcross, 2022).

Tabla 1. Gestión de riesgo de cáncer basado en resultados de pruebas genéticas. Adaptado de NCCN, 2022.

Gen	CM	CO	Cáncer pancreático y otros tipos de cáncer
ATM	15-40%*	<3%*	~5%–10%*Próstata: NE
BARD1	15%–40* (con predisposición a triple negativo)	NA	Otros: NE
BRCA1	>60%**	39%–58%**	≤5%*
BRCA2	>60%** (con predisposición a ER+)	13%–29%**	5%–10%**
BRIP1	ND	>10%*	Otros:NE

CDH1	41-60%*	NA	
CDKN2A	NA	NA	>15%** Melanoma: 28-76%*
CHEK2	15-40%*	NA	Colon: NE
MSH2, MSH1, MSH6, PMS2, EPCAM	<15%[#]	*MSH2, MSH1*:>10%* *MSH6*: ≤13% *PMS2*: <3%[#] *EPCAM*: <10%[#]	<5%–10%*
NBN	NA	ND[#]	Otros: NE
NF1	15%–40%*	NA	
PALB2	41%–60%*	3%–5%*	5%–10%[#] Otros: NE
PTEN	40%–60%*	NA	
RAD51C	15%–40%*	>10%*	Otros: NE
RAD51D	15%–40%*	>10%*	Otros: NE
STK11	40-60%*	NA	>15%* Cáncer de ovario no epitelial: >10%*
TP53	>60%*	NA	5%–10%[#]

*NA: Sin asociación con incremento del riesgo; NE: Evidencia insuficiente en el incremento del riesgo; ND: Datos insuficientes; [#]Nivel de evidencia limitado, tamaño muestral pequeño. *Fuerte nivel de evidencia, basado en al menos un estudio de caso/control, incluidos casos determinados por laboratorios comerciales o sin controles de la misma población; **Muy fuerte nivel de evidencia, estudios de cohortes prospectivos en un entorno poblacional han demostrado el riesgo.*

El análisis genético de aquellas que presentan el fenotipo del síndrome, negativas a VPs en *BRCA1/2* ha demostrado que pueden identificarse VP en otros genes de la vía de señalización *BRCA* que incluyen a *RAD51C, RAD51D, BRIP1, PALB2, BARD1, NBN, MRE11A* (Al Bakir y Gabra, 2014), genes de susceptibilidad de la vía de Fanconi (*FANCD2, FANCA* y *FANCC*), MMR (*MLH1, MSH2, PMS1, PMS2* y *MSH6*), de reparación del DNA por HR (*ATM, ATR* y *CHK1/2*) y supresores de tumores (*TP53, SKT11 y PTEN*).

Adicionalmente, se han identificado otros genes que actúan en distintas vías de señalización con penetrancia moderada como *CDK1, RAD50* y *FANCJ* y baja como *FGFR2, LSP1, MAPK3K1, TGFB1, TX3, VEGF, PGR* y *KRAS* (Kobayashi *et al.,* 2013). En estos casos, las estimaciones de dicha penetrancia y las recomendaciones para su manejo son menos conocidas, pero con paneles más amplios es más probable encontrar VUS, VP o VPP que aporten más evidencia al respecto (Bellcross, 2022).

1.7 Mecanismos de reparación relacionados con SHCMO

Diferentes lesiones al DNA desencadenan vías particulares de respuesta al daño como: i) la reparación por escisión de bases (BER, por sus siglas en inglés), ii) la reparación por escisión de nucleótidos (NER, por sus siglas en inglés), iii) el mecanismo de reparación del DNA por bases mal apareadas (MMR, por sus siglas en inglés), iv) reparación por recombinación homóloga (HR, por sus siglas en inglés), v) recombinación de extremos no homólogos (NHEJ, por su sigla en inglés), vi) la unión de extremos mediada por microhomología (Voutsadakis y Stravodimou, 2023).

Casi todos los casos de genes relacionados con el SHCMO codifican supresores de tumores que participan en la reparación del DNA por HR, el MMR y la reparación de entrecruzamientos a través de la vía de la anemia de Fanconi (Nielsen *et al.,* 2016).

1.7.1 Recombinación homóloga

La reparación del daño por ruptura de doble cadena (DSB, por sus siglas en inglés) se realiza a través de la HR, un mecanismo que utiliza la cromátida hermana no dañada como molde de reparación en las fases tardías S/G2 del ciclo celular, involucrando proteínas detectoras de las roturas (ATM/ATR), reparadores del daño (BRCA2 y RAD51) y mediadores para la unión (CHK2 y BRCA1) (Macedo *et al.,* 2019).

En la presinapsis, un conjunto de nucleasas corta ambos lados de la DSB para generar proyecciones 3’ ssDNA y permitir la nucleación de filamentos de la nucleoproteína recombinasa Rad51, de esta manera, el filamento de nucleoproteína RAD51-ssDNA busca una secuencia homóloga no dañada en el genoma. Una vez encontrada, durante la sinapsis se forma un loop de desplazamiento cuya cadena invasora sirve como primer para la síntesis de DNA, luego el cromosoma intacto es restaurado en la postsinapsis (Hanaoka y Sugasawa, 2016).

El corte ocurre de 5’ a 3’ en los extremos de la DSB mediante una nucleasa que deja libres colas monocatenarias de DNA recubiertas por RPAs (proteínas de replicación A) (Altieri *et al.,* 2008). Tras reconocer el DSB, ATM y ATR reclutan otras proteínas como BRCA1, que sirven como sitio de acoplamiento del complejo MRN (MRE11, RAD50 y NBN), que crea extensiones de cadena única en la DSB. El dominio de unión al DNA C-terminal de BRCA2 se une al ssDNA-dsDNA en la lesión, desplazando las RPAs (Venkitaraman *et al.,* 2014).

BRCA2 en cooperación con PALB2 facilita que RAD51 se acople a las proyecciones mencionadas (estabilizan el filamento de RAD51 decrementando su actividad ATPasa) que

lo habilitan para invadir las cadenas hermanas y así, usarla como molde para crear extensiones de DNA que luego serán ligadas para reparar la ruptura (Voutsadakis y Stravodimou, 2023).

El CM y CO familiar está asociado a la pérdida de la función en genes modificadores de la HR como *BRCA1/2*, *PALB2*, *ATM*, *RAD51C* y *RAD51D* (Lord y Ashworth, 2012). Las DSBs son aberraciones estructurales típicas encontradas en células deficientes de BRCA, lo que sugiere un importante papel de la HR en la supresión tumoral, debido a que las células, al no contar con BRCA1/2 o PALB2, reparan sus errores con mecanismos como la unión de extremos no homólogos (NHEJ), propenso a aneuploidías y segregaciones cromosómicas (Macedo *et al.,* 2019).

Durante el proceso de carcinogénesis muchas proteínas de la HR son inactivadas selectivamente en las células tumorales, por lo que pueden ser sensibilizadas a agentes quimioterapéuticos que dañan el DNA, sin embargo, una HR funcional también es requerida para la supervivencia de las células cancerosas, por ello los esfuerzos farmacológicos están orientados en la disrupción de la HR u otra vías que incrementen la eficacia de la terapia y al tiempo sensibilicen a los cánceres resistentes (Hanaoka y Sugasawa, 2016).

1.7.2 Reparación de errores de apareamiento (MMR)

La MMR es un sistema para la reparación de inserciones erróneas, deleciones y mala incorporación de bases que surgen durante la replicación del DNA (Nielsen *et al.,* 2016). Las células tumorales deficientes en MMR tienen una frecuencia de mutaciones más alta que las normales y exhiben inestabilidad microsatelite, esto es, la reducción de la fidelidad de los elementos repetitivos del DNA durante la replicación (D'Andrea, 2008).

Defectos genéticos en el sistema de MMR del DNA incluyen sustituciones de base e inserciones y deleciones (Indels). El complejo MutSα (MSH2 y MSH6) o el MutSβ (MSH2-MSH3) reconocen los errores en el apareamiento de una sola base, uniéndose y reclutando el complejo MutLα (MLH1-PMS2), que conduce a la discriminación de la cadena, tras esto, ocurre la remoción de los errores mediante una exonucleasa, resíntesis y ligadura (Kobayashi *et al.,* 2013). Defectos en MSH y MLH/PMS causan predisposición al síndrome de Lynch (Hanaoka y Sugasawa, 2016).

El mecanismo mediante el cual mutaciones en genes de MMR predisponen al desarrollo tumoral difiere de la pérdida de heterocigosidad que se observa en el SHCMO (Shulman, 2010), sin embargo, existe un solapamiento funcional de las vías MMR y FA-BRCA: BRCA1

forma parte de un complejo de múltiples unidades proteicas que incluye proteínas envueltas en la reparación del daño; existe una interacción funcional entre FANCJ y MutLα para el establecimiento de enlaces cruzados entre cadenas de DNA; FANC2 se requiere para la unión MSH2- MLH1, que participa en la ubiquitinación de FANCD2, lo que lleva al reclutamiento de ATR, activando CHK1 y TP53; MutS y MutLα son necesarios para reclutar a ATR a la lesión en el DNA (Kobayashi *et al.,* 2013).

Cuando RAD51C no se empareja con la cromátida hermana o el DNA heteroduplex contiene un exceso de nucleótidos mal emparejados, el MMR suprime la HR. En ausencia de MMR, la precisión de HR para la reparación de DSB se reduce por errores en la elección del molde o durante la extensión de la polimerasa en la hebra rota, lo cual en conjunto permite la acumulación de cambios genéticos en el SHCMO (Nielsen *et al.,* 2016).

1.7.3 Vía de anemia de Fanconi

Los productos de los genes *BRCA* también están envueltos en la reparación de otro tipo de daños como son los enlaces cruzados entre cadenas de DNA (ICL, por sus siglas en inglés), donde la vía de la anemia de Fanconi (FA) contiene un paso inevitable de HR (Motegi *et al.,* 2019).

Los ICLs se producen mediante la unión covalente de hebras opuestas de la hélice de DNA, bloqueando la progresión de la horquilla de replicación. El complejo nuclear FANCore (FANCA/B/C/D/F/G/M/L) con actividad de ubiquitina E3 ligasa actúa sobre el complejo heterodímero ID (FANCD2-FANCI) que se reubica en el DNA dañado de un modo dependiente de BRCA1 y ATR para promover la ruptura nucleolítica de los sitios 3’ y 5’ del DNA y así desenrollar los ICL, posteriormente, se induce la acción de polimerasas durante la síntesis y se reestablece la integridad del genoma mediante HR (Bogliolo y Surrallés, 2015).

BRCA1/2, BRIP1, RAD51C y *PALB2* son participantes en la reparación de los ICLs, lo que permite apreciar lo amplia que es la red funcional de los factores implicados en el SHCMO. Por su parte, otros genes como *BLM, RECQL, FANCC* y *FANCM* se han asociado con susceptibilidad al CM, pero no se han asociado mutaciones con predisposición al CO (Nielsen *et al.,* 2016).

1.8 Mutaciones comunes dentro del síndrome

Para definir una mutación patogénica de línea germinal, el Colegio Americano de Genética Médica (ACGM, por sus siglas en inglés) usa el término de VP o VPP. Estimaciones previas sugerían que 1/400 individuos en la población general era portador de una VP en *BRCA1/2*,

sin embargo, estudios más recientes basados en secuenciación de exoma completo (WES, por sus siglas en inglés) predicen que se trata de 1/200, mientras que esta incidencia aumenta para el caso de la comunidad de judíos ashkenazi (1/40) (Bellcross *et al.,* 2022).

Se ha reportado que el SHCMO explica entre 15-18% de los casos de CO, aunque en otros estudios se reporta hasta 80% (Al Bakir y Gabra, 2014), por su parte, explica entre el 5-10% de los casos de CM (Silva *et al.,* 2014). A nivel molecular, el SHCMO se define mediante la identificación de variantes patogénicas de línea germinal en genes de alto riesgo (*BRCA1/2* y *PALB2*), riesgo moderado (*ATM* y *CHEK2*), entre otros genes de bajo riesgo (Nakamura *et al.,* 2021).

Aproximadamente, el 30% de los casos de pacientes con SHCMO son *BRCA* negativos y muestran mutaciones en otros genes. En la región codificante y no codificante de dichos genes se han encontrado alrededor de 2000 mutaciones, siendo las más prevalentes los cambios de marco de lectura por inserciones/deleciones, codones de paro, mutaciones no sinónimas y defectos en el corte y empalme del mRNA que puede llevar a proteínas no funcionales (Malik y Masood, 2022).

Aunque el SHCMO explica ~10% de casos de CM a partir de mutaciones en genes de riesgo alto y moderado, en la mitad de los casos clínicos de dichos pacientes se desconoce aún la base genética, no se explica por variantes de un solo nucleótido o, por el contrario, pacientes con VP en genes del síndrome no tienen una historia familiar sugerente (Sessa *et al.,* 2022).

El 70-80% de las VPs o VPPs generan codones de paro, truncan la proteína codificada o reducen su expresión a través del decaimiento del mRNA mediado por mutación terminadora (NMD, por sus siglas en inglés) (Yoshida, 2021). Es destacable que cerca del 20-30% de las pacientes con el síndrome tienen mutaciones puntuales o rearreglos estructurales que derivan en variaciones en el número de copias (CNV, por sus siglas en inglés) asociadas con un mayor riesgo de padecer CM (45-70%) y CO (20-40%) (Silva *et al.,* 2014).

Además de las VP/VPP germinales en genes de susceptibilidad, se han identificado variantes de significado clínico incierto (VUS, por sus siglas en inglés) en 10-20% de las pacientes sometidas a evaluación genética de *BRCA1/2* (Fanale *et al.* 2021). Los portadores de VPs en *BRCA1* y *BRCA2* se caracterizan por tener un riesgo de entre 15-45% y 10-20%, respectivamente de desarrollar CO y de 50-85% para CM, de ahí la necesidad de analizar otros genes involucrados en la predisposición de dichos tumores (Bono *et al.,* 2021).

Se ha observado que la edad promedio en que el cáncer se desarrolla es menor en pacientes donde hay mutaciones en *TP53* y *PTEN* respecto a *BRCA1/2*, por lo cual es necesario prestar atención en la edad donde la vigilancia comienza. Por su parte, se ha observado una relación entre los sitios donde ocurre la mutación y el riesgo de SHCMO, pues en pacientes portadoras de mutaciones en *BRCA1/2* se han identificado que VPs en regiones de CO adyacentes al exón 11 confieren mayor riesgo al padecimiento que los casos donde hay VPs en otros sitios de ambos genes (Yoshida, 2021).

Los tipos más comunes de mutaciones deletéreas en *BRCA1/2* son pequeñas deleciones o inserciones de cambio de marco de lectura, mutaciones sin sentido y en el sitio de empalme. Las regiones genómicas de ambos genes están compuestas por una gran cantidad de elementos repetitivos del DNA, 42% y 20% de secuencias, respectivamente, por lo cual, se han observado reordenamientos mediados por ellos (Macedo *et al.,* 2019).

Los casos restantes, donde aparentemente no hay una causa de la heredabilidad del síndrome pueden explicarse a partir de VPs y CNVs asociados al SHCMO en genes no analizados rutinariamente en la secuenciación, susceptibilidad poligénica, variantes en elementos reguladores o del splicing, silenciamiento epigenético e inserciones de elementos móviles transponibles (TEs, por sus siglas en inglés), que han cobrado relevancia en el estudio de la predisposición genética en el cáncer en los últimos años (Klein *et al.,* 2023).

1.9 Elementos móviles transponibles

Los TEs son fragmentos de DNA capaces de cambiar su localización o replicarse dentro del genoma de un huésped, sin embargo, difieren en sus mecanismos de transposición y en su capacidad para transponerse en otra localización cromosómica (Klein y Anderson, 2022).

Los TEs se subdividen en dos clases principales definidas por su movilidad. Los TEs de clase I son conocidos como retrotransposones y comprenden elementos que cambian su localización mediante un mecanismo de “copiar y pegar” que involucra un intermediario de RNA. Por su parte, los de clase II son los transposones de DNA que se movilizan con un mecanismo de “cortar y pegar” (Hancks y Kazazian, 2016)

De acuerdo con su contenido de estructuras largas de repetición terminal (LTR) y sus regiones flanqueantes con promotores, poliadenilación y potenciadores, se categorizan en LTR y non-LTR (Lee *et al.,* 2022). En el genoma humano, los únicos TEs actualmente activos conocidos son los non-LTR, que incluyen: i) elementos largos intercalados (LINEs), ii) elementos cortos intercalados (SINEs) y iii) SINE-VNTR-*Alu*s (SVAs), donde VNTR es un número variable de repeticiones en tándem (Pfaff *et al.,* 2022).

1.9.1 Estructura de elementos retrotransponibles

Teniendo en cuenta que los genes codificantes para proteínas representan 1.5% del genoma, el estudio de los TEs, que representan 45% del genoma humano, es crucial para entender la evolución del genoma, diversidad genética, regulación de genes y enfermedades (Lee *et al.,* 2022). Una pista para explicar muchos cambios que resultan en deleciones, inserciones y rearreglos cromosómicos son las secuencias repetitivas de DNA que flaquean puntos de ruptura. Los TEs patogénicos representan aproximadamente entre 0.03 y 0.1% de las variantes responsables de enfermedades genéticas, sin embargo, esto puede tratarse de una subestimación por la dificultad de su detección con pruebas moleculares rutinarias (Eyries *et al.,* 2022).

Los elementos LINE-1 o L1 (~6 kb) son LINEs autónomos, codifican su propia maquinaria enzimática para movilizarse a sí mismos y a otros TEs como los *Alu* y SVAs (Rishishwar *et al.,* 2017). A pesar de que representan 17% del genoma humano, sólo 80-100 copias conservan dicha capacidad (De Brakeleer *et al.* 2020). Representan 17% del genoma (516,000 copias), mientras que los elementos *Alu* el 11% (1,090,000 copias) y los SVA el 0.2% (2,700 copias) (Richardson *et al.,* 2015). Por su parte, los elementos L1 codifican dos proteínas esenciales: ORF1p y ORF2p, las cuales son, respectivamente, una proteína de unión al RNA con actividad chaperona de ácidos nucleicos y una proteína con dominios multifuncionales que proveen una actividad endonucleasa y transcriptasa reversa (Ade *et al.,* 2013) (Figura 1).

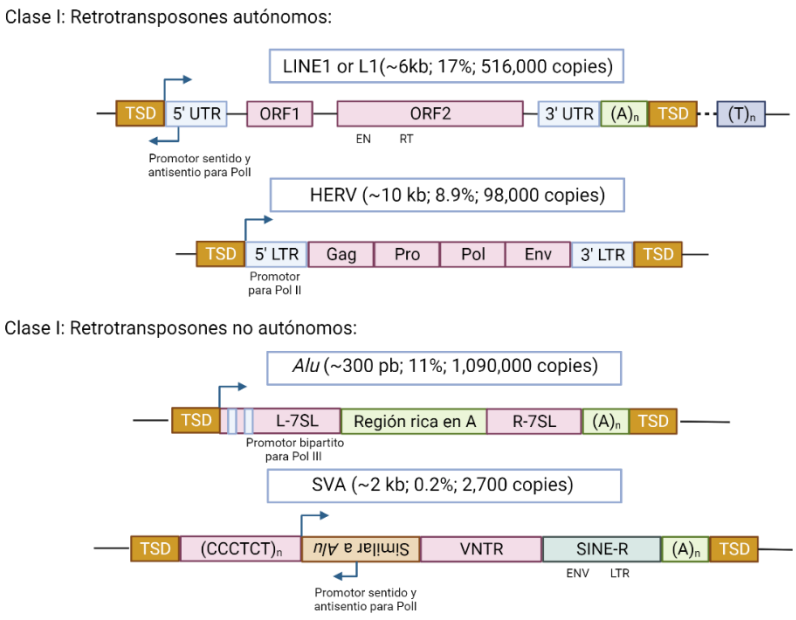


Figura 1. Estructura de retrotransposones activos en el genoma humano.

ORF1= Marco de lectura abierto 1, ORF2= Marco de lectura abierto 2, UTR= Regiones no traducidas, TSD= Sitio de duplicación objetivo, Pol III= Polimerasa III, EN= Endonucleasa, RT= Retrotranscriptasa, $(A)_n$=Sitio de número variable de adeninas, $(T)_n$= Región terminadora de timinas, Gag= codifica para una poliproteína susceptible de ser digerida por proteasas virales, Pro=una proteasa para procesar polipéptidos virales como las proteínas GAG y POL, Pol= transcriptasa inversa e integrasa, Env= proteína transmembrana implicada en el reconocimiento de receptores flanqueados por LTRs, LTR=Repetición terminal larga, L-7SL= Componente izquierdo de RNA de la partícula de reconocimiento de señales 7SL1, R-7SL= Componente derecho de RNA de la partícula de reconocimiento de señales 7SL1, $(CCCTCT)_n$= Región variable de repeticiones CCCTCT, VNTR=Número variable de repeticiones en tándem, SINE-R= Región similar a elemento nuclear intercalado corto

Los elementos *Alu* son los SINEs más abundantes, con una longitud ~300 pb (Konkel *et al.* 2010). Estructuralmente se componen de dos monómeros no idénticos derivados del gen 7SLRNA unidos por una región rica en adenina, además de una cola poli-A en el extremo 3' (Ade *et al.,* 2013). Cuentan con dos elementos internos del promotor de la RNA pol III en el monómero izquierdo. Debido a la falta de señales terminadoras, los transcritos cuentan con una región única más allá de la cola poli-A (Roy-Engel, 2012) y su transcripción termina en sitios cercanos ricos en timinas (Deininger, 2011).

Los SVAs cuentan con una estructura que consiste de una repetición hexamérica de CCTCT, un elemento parecido a *Alu*, un conjunto de repeticiones variable de nucleótidos en tándem (VNTR) ricos en GC, una secuencia SINE que comparte homología con HERVK-10 (un transposón LTR inactivo), un sitio de unión del factor de especificidad de clivaje y poliadenilación y un tracto poli-(A) (Richardson *et al.,* 2015).

1.9.2 Principales familias activas de elementos *Alu* en el genoma humano

Se reconocen 3 subfamilias principales: *Alu*J es la más vieja, seguido de los miembros intermedios *Alu*S y las inserciones más jóvenes *Alu*Y (Konkel *et al.,* 2010). Miembros de las subfamilias *Alu*S y *Alu*J, activas desde hace 50 millones de años son los responsables de la gran mayoría de las copias presentes en el genoma humano y aunque se consideran no funcionales, incapaces de retrotransposición, se han reportado elementos *Alu* de la subfamilia *Alu*Sq/Sp insertados *de novo* en genes como *BRCA1,* aunque aún no se ha ligado a un impacto funcional en el SHCMO (Teugels *et al.,* 2005).

Familias como L1PA1, SVA y varias subfamilias *Alu*Y muestran polimorfismos en la población humana, lo que indica su actividad reciente. Por su parte HERVK es el único linaje de retrovirus endógenos (ERVs) que exhibe inserciones polimórficas en la población humana (Kojima, 2018).

1.9.3 Mecanismo de inserción

Los L1 se transponen mediante la transcripción inversa dirigida (TPRT, por sus siglas en inglés), que comienza cuando el RNA del L1 se transcribe mediante la RNA pol II (Figura 2).

Una vez en el citoplasma los dos ORFs codificados por el L1 son traducidos y dichas proteínas se unen al RNA de L1 para formar el complejo ribonucleoproteico L1 RNP que es devuelto al núcleo. Ahí, ORF2p corta la hebra inferior en la secuencia consenso 5'-TTTTAA-3' del DNA en el sitio TA y el grupo 3'-OH expuesto es utilizado como primer para la transcripción reversa (Pfaff *et al.,* 2022), aunque en otros casos la región de los nucleótidos expuestos es referida como el primer (Roy-Engel, 2012, Pócza *et al.,* 2021).

De este modo, la cola poli-A del mRNA intermediario del elemento L1 (para *Alu* y SVA es el mismo principio) se une complementariamente con el TTTT escindido y es transcrito en reversa por la actividad de ORF2p a cDNA (Konkel *et al.,* 2010). Después, se corta la hebra superior para integrarlo y se sintetiza la cadena complementaria de DNA (Pfaff *et al.,* 2022).

El RNA de la secuencia *Alu* es transcrito en el núcleo por la acción de la polimerasa III, donde forma una RNP al asociarse con las proteínas PABP (de unión a poli-A), SRP9 y SRP14 (de reconocimiento de señales), que se ha propuesto que participan para dirigirlo hacia los ribosomas donde se está traduciendo el L1 RNA, para así reclutar a ORF2p (Ade *et al.,* 2013). PABP se ha propuesto como el responsable de poner en cercanía al RNP con ORF2p al unirse al complejo CAP del L1 RNA que se está traduciendo (Roy-Engel, 2012).

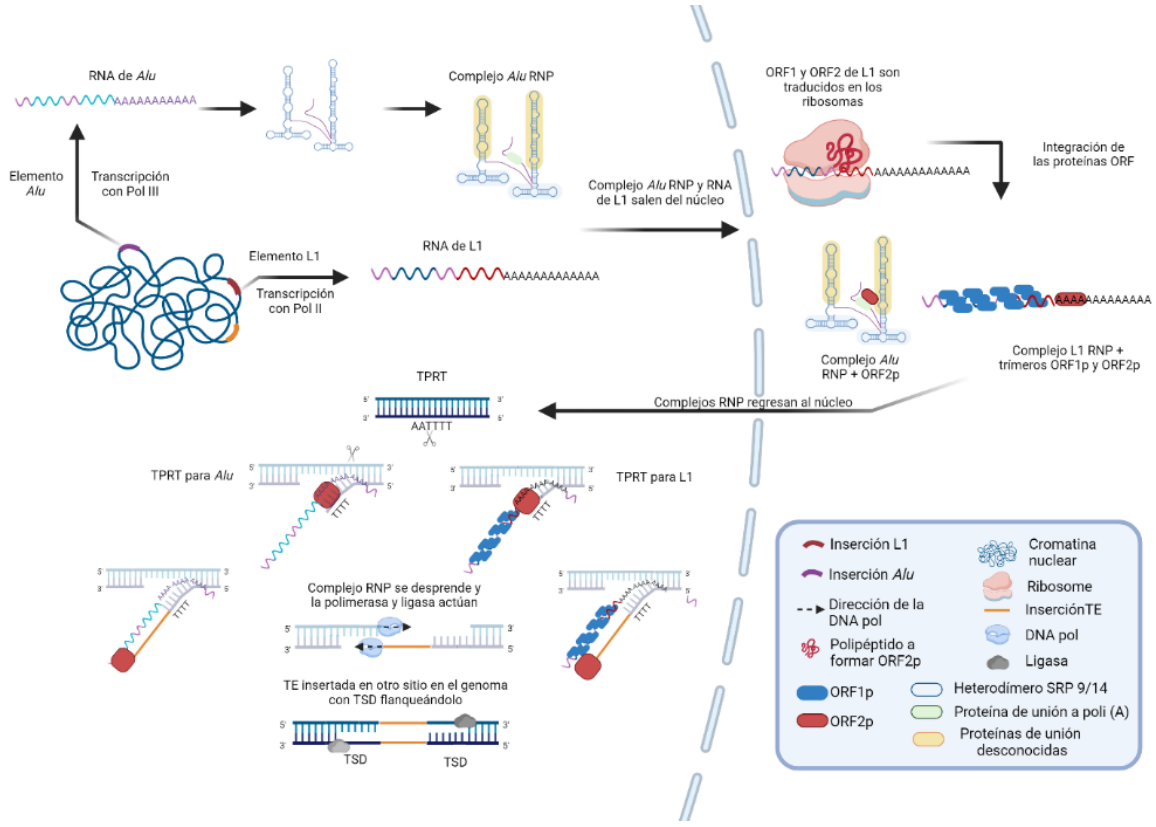


Figura 2. Mecanismo de TPRT para elementos Alu y L1.
TPRT= Transcripción inversa dirigida. RNP=Ribonucleoproteína, Pol II=Polimerasa II, Pol III= Polimerasa III, ORF1= Marco de lectura abierto 1, ORF2= Marco de lectura abierto 2, ORF1p= Proteína del marco de lectura abierto 1, ORF2p= Proteína del marco de lectura abierto 2, DNA pol= Polimerasa de DNA, SRP=Partícula de reconocimiento de señales 9/14.

Una vez en el núcleo ocurre la TPRT, la endonucleasa corta en los extremos del sitio objetivo liberando el extremo 3' con nucleótidos TT para que la cola poli-A del elemento *Alu* complementaria pueda unirse. Tras la transcripción reversa de la secuencia de RNA, las brechas son llenadas por la polimerasa, la cadena de RNA se lisa e intercambia por DNA con la acción 3' → 5' de la polimerasa.

Una vez participa la ligasa, la secuencia *Alu* termina insertada con la duplicación del sitio objetivo (TSD) flanqueándola (Pócza *et al.,* 2021), esto es, la duplicación de la hebra sobresaliente de ~15 pb en ambos lados de la inserción (Chu *et al.,* 2021). En el proceso la cola única en 3' del RNA del *Alu* es eliminada. Muchos mRNAs primarios de genes contienen múltiples secuencias *Alu*, particularmente en la región 3'UTR, afectando por tanto mecanismos reguladores de la expresión genética (Ade *et al.,* 2013).

1.9.4 Efecto de los TEs

Una visión de los TEs como simbiontes funcionales involucra su participación en la generación de nuevas señales que participan en la regulación o de nuevas secuencias codificantes. Como ejemplo, los TEs pueden proveer nuevas secuencias potenciadoras para los TFs y contribuyen al mantenimiento de la arquitectura genómica al proveer sitios de unión para la proteína CTCF, que es clave en el establecimiento de los dominios topológicamente asociados (Colonna y Fanti, 2022).

Se ha postulado que es gracias a la acumulación de mutaciones puntuales que los elementos *Alu* pueden volverse potenciadores, tal es el caso de los que intervienen en el ciclo celular a través del reclutamiento del factor de transcripción C de la RNA Pol III, lo que desencadena cambios en la cromatina y acetilación de las histonas, resultando en cambios en la expresión genética en *cis*. Por su parte, el RNA de los elementos *Alu* puede reducir la transcripción de la RNA Pol II en *trans* (Payer *et al.,* 2021).

Cuando las secuencias *Alu* son transcritas independientemente o como parte de otros RNA se pueden formar estructuras que regulan el inicio de la traducción, como es el caso de los IR*Alu*s que potencialmente regulan la retención nuclear del RNA y la formación de circRNA (Chen y Yang, 2017).

En tanto a sus efectos en detrimento del hospedero se han identificado aproximadamente 100 inserciones de Tes asociados con enfermedades humanas, muchas de las cuales corresponden a genes de predisposición al cáncer como es el caso de *BRCA1, BRCA2, APC, MSH2, NF1* y *ATM* (Bouras *et al.,* 2021). A pesar de que no tienen capacidad codificante, la inserción de elementos *Alu* en regiones críticas del genoma puede alterar la transcripción, corte y empalme, y la traducción (Pócza *et al.,* 2021).

El nivel de actividad es variable entre pacientes con el mismo tipo de cáncer y entre distintos tipos. Aproximadamente 42% del gen *BRCA1* se compone de elementos *Alu*, volviéndolo un gen objetivo para la recombinación homóloga no alélica, sobre todo eventos *Alu-Alu*. Por su parte, es importante saber que no son los únicos, ya que se han reportado inserciones somáticas L1 en pacientes con CM (Steely *et al.,* 2021). La integración de un TE en un exón codificante puede originar un codón de paro prematuro o un cambio en la secuencian del ORF. De manera alternativa, es posible que su integración en potenciadores exónicos del splicing cause que ciertos exones sean omitidos. El splicing del mRNA puede alterarse por TEs intrónicas en sitios de splicing (SS) e incluso en casos donde no se afecta directamente los sitios donadores o aceptores (Payer y Burns, 2019).

La inserción y la remoción de TEs no suele ser precisa debido a que puede afectar las secuencias circundantes a ellos, causando duplicaciones y reordenamientos. Aunado a esto, los TEs pueden causar SVs, incluso después de haber perdido su capacidad para retrotransponerse. Esto lo logran mediante la recombinación entre las regiones homólogas dispersas por ellos en regiones distantes, lo que ocasiona deleciones, duplicaciones e inversiones (Bourque *et al.,* 2018). Algunos TEs generan sitios de splicing crípticos y por lo tanto pueden introducir patrones nuevos del splicing del mRNA. Finalmente, TEs insertados por TPRT pueden proveer SSs (Payer y Burns, 2019). Existen más de 3200 elementos *Alu* polimórficos reportados, es decir, que presentan una frecuencia alélica >5%. Aunque parezca contraintuitivo se ha reportado un subconjunto intrónico que puede alterar el splicing de exones cercanos (Payer *et al.,* 2021)

El impacto en la estructura genómica de los TEs conduce a modificaciones potenciales de secuencias homólogas en 5’-UTR, 3’-UTR y regiones intrónicas. Además, pueden generar entrecruzamientos desiguales que originan CNVs, incluyendo desde un exón hasta todo el gen (Solassol *et al.,* 2019).

Además de la fuente de daño asociada a la inserción, existen otros mecanismos por los cuales los TEs pueden llegar a causar daño celular. Tal es el caso de la desrepresión de los promotores de LTR y L1, que pueden ser los responsables de la activación de oncogenes; la

actividad endonucleasa de ORF2p puede ser responsable de rupturas en el DNA e inestabilidad genómica. La acumulación de los transcritos de RNA derivados de TEs puede desencadenar la respuesta inmune innata que provoque enfermedades autoinmunes e inflamación (Bourque *et al.*, 2018).

1.9.5 Aspectos que limitan los efectos de los TEs

Más del 99.9% de las copias de L1 en el genoma humano están fijas y no pueden moverse debido a mutaciones y truncamientos. Se estima que un genoma humano promedio contiene de 80 a 100 L1 activos y un pequeño número de secuencias muy activos responsables de la mayor parte de la retrotransposición humana (Zhou *et al.* 2020). Esto implica ciertas limitaciones para su estudio ya que el genoma de referencia no incluye los TEs humanos de todas las poblaciones (Bourque *et al.,* 2018).

Los L1 están usualmente transcripcionalmente reprimidos, pero cambios epigenéticos que ocurren en los tumores pueden promover su expresión y retrotransposición. La mayoría de los casos representan mutaciones pasajeras sin efecto en el desarrollo de cáncer, pero pueden promover otros tipos de SVs de línea germinal y somáticas adicionales a la inserción del L1 (Rodríguez-Martin *et al.,* 2020). La ausencia general de L1s en genes supresores de tumores ha llevado a la investigación de dichos elementos desde su papel en la retrotransposición de otros elementos involucrados en la progresión del cáncer (Hancks y Kazazian, 2016).

El potencial funcional de los elementos *Alu* depende de su localización genómica y de las características de la secuencia, ya que su distribución en el genoma humano tiende a regiones ricas en genes, lo que no ocurre con elementos L1 (Chen y Yang, 2017).

Pese a que los TEs constituyen la fuente principal de nuevos elementos reguladores en los genomas de primates (Kojima *et al.,* 2023), el silenciamiento de la mayoría de los elementos *Alu* en el genoma se explica gracias al contexto de la cromatina en donde se localizan o a que estos son enriquecidos con dinucleótidos CpG, siendo una fuente de metilación del DNA en regiones intrónicas e intergénicas (Chen y Yang, 2017).

Muy pocos elementos *Alu* son capaces de generar copias, aunque la mayoría hace transcritos, tras insertarse en un nuevo locus son degradados con rapidez en una escala temporal evolutiva por el acortamiento del tracto de poli-A, la acumulación de bases heterogéneas en esta última y la acumulación de mutaciones en el elemento, lo que en conjunto limita su capacidad para la formación del complejo ribonucleoproteico que participa en la retrotransposición (Deininger, 2011).

La acumulación de mutaciones tras la retrotransposición altera el promotor interno de la Pol III, la estabilidad de la estructura del RNA (y por tanto su habilidad para unirse a proteínas involucradas en la retrotransposición) y ocasiona que la secuencia poli-A se acorte con rapidez (Ade *et al.,* 2013).

1.10 Secuenciación masiva y plataformas

La secuenciación del DNA es un método empleado para determinar el orden de las bases nucleotídicas en el DNA de un paciente, que en este contexto tiene la finalidad de asociar genotipos con caracteres fenotípicos.

La historia de la secuenciación se divide en tres generaciones: La primera de ellas involucra la secuenciación por síntesis (Sanger) y por escisión (Maxam-Gilbert); la segunda comprende principalmente tres tecnologías: pirosecuenciación, plataforma 454 (Roche); terminadores reversibles, plataforma Solexa (Illumina), y secuenciación por ligación, plataforma ABI/SOLiD (Life Technologies) (Verma *et al.,* 2016). Por su parte, la secuenciación de tercera generación ocurre a partir de moléculas individuales de DNA, es decir, no se requiere amplificar bibliotecas de fragmentos (Delseny *et al.,* 2010).

A diferencia de Sanger, donde primero se sintetizan una serie de copias parciales de DNA y luego son analizadas, en la secuenciación por síntesis tras amplificación se amplifican los fragmentos de DNA en clústers, luego se desnaturalizan y distribuyen en microarreglos que se introducen en una celda de flujo donde ocurre la secuenciación, un primer se extiende cíclicamente uno o varios nucleótidos a la vez y la secuencia se lee en cada paso de síntesis del DNA (Delseny *et al.,* 2010).

En Sanger se requiere un DNA molde purificado en cada reacción, mientras que las tecnologías de secuenciación de próxima generación (NGS) o secuenciación de alto rendimiento que comprenden la segunda y tercera generación, generan datos de secuencias a partir de un gran conjunto de fragmentos de DNA (Wang y Cheng, 2019). Su principal característica es el uso de métodos masivamente paralelos, esto es, que muchas muestras de DNA se secuencian una al lado de la otra en el mismo aparato, lo cual ha implicado una gran miniaturización y detectores de fluorescencia muy sensibles (Clark *et al.,* 2019).

1.10.1 Secuenciación de segunda generación (por síntesis tras amplificación)

El proceso de pirosecuenciación parte de la fragmentación de DNA que se desnaturaliza para obtener cadenas únicas a las que son ligadas adaptadores en ambos extremos. Cada fragmento se une a una microperla y es amplificado mediante PCR en emulsión (cada una se

aísla en una gota con una mezcla de reacción PCR en emulsión de aceite), para generar millones de copias de un fragmento único de DNA (Verma *et al.,* 2016).

Cuando la emulsión se rompe, las perlas se depositan en pozos de una placa de titulación (contenida en una celda de flujo) donde ocurre la secuenciación gracias a una DNA polimerasa. Cuando un nucleótido se añade al primer, una molécula de pirofosfato se libera y es convertida mediante una sulfurilasa en ATP a ser usado en la producción de una señal quimioluminiscente por la reacción de la luciferasa (Delseny *et al.,* 2010). La pirosecuenciación es más adecuada para la detección de variantes que ocurren en una pequeña región seleccionada, sobre todo cuando se trata de variaciones de baja frecuencia (Bayrak-Toydemir y Wooderchak-Donahue, 2014).

En la plataforma Solexa el proceso también implica la fragmentación del DNA genómico, adaptadores se ligan en ambos extremos; en la celda de flujo uno de ellos se hibrida con un un oligo complementario anclado covalentemente a la superficie. El extremo opuesto se hibrida a otro oligonucleótido cercano formando un puente; con DNA polimerasa y dNTPs, el ssDNA se amplifica, de modo que en las siguientes rondas de PCR la cadena molde y la sintetizada se desnaturalizan para reiniciar la amplificación hasta tener en cada canal de la celda clústers densos de miles de cadenas de DNA sentido (solo queda éstas) y antisentido (Verma *et al.,* 2016).

Primers de secuenciación y nucleótidos modificados se añaden a la celda, cada uno tiene un bloqueador, una molécula terminadora fluorescente reversible, de modo que sólo un nucleótido puede añadirse en cada ciclo, la cámara captura la imagen de la celda para determinar la longitud de onda de la etiqueta fluorescente. Las moléculas que no se incorporan se lavan cada ciclo, secuenciando miles de fragmentos simultáneamente. La longitud de lectura de las plataformas de Illumina es arriba de 300 pb con ≤ 0.1% errores para ≥ 85% de las bases (Wang y Cheng, 2019). A la secuencia final de cada clúster se le llama lectura, hay decenas de millones de clústers en una celda y más de una lectura por cada clúster (Clark *et al.,* 2019).

La secuenciación por ligadura utiliza una emPCR para la amplificación clonal, perlas se inmovilizan en una superficie de vidrio, lo cual es posibilitado gracias a una modificación en el extremo 3’ del DNA objetivo. La reacción de secuenciación en la plataforma SOLiD depende de secuencias de ocho nucleótidos usadas por la sonda de detección y de la actividad de la DNA ligasa, y es completada en cinco partes con cinco primers universales, una base desplazada después del primero, logrando que todas las bases son leídas dos veces (Ari y Arikan, 2016)

SOLiD y 454 se han vuelto obsoletas debido a su alto costo por base y su alta tasa de errores (Wang y Cheng, 2019). La ventaja principal de la NGS es el incremento en su rendimiento y la sustitución del proceso de clonación por la amplificación de los fragmentos mediante PCR, que en los casos más recientes se elimina porque el DNA se secuencia directamente. Sin embargo, también tienen limitaciones, pues muchas tecnologías producen lecturas cortas, lo que en ocasiones implica volver a secuenciar un genoma para detectar mutaciones puntuales, hay problemas con las regiones repetidas para ensamblarlas sin ambigüedad y cada paso en la construcción de las librerías puede introducir sesgos y artefactos (Delseny *et al.,* 2010).

1.11 Metodología empleada para la detección de elementos *Alu*

Muchos de los métodos empleados en pruebas de cáncer hereditario no pueden detectar de manera confiable los TEs debido a limitaciones técnicas. Los métodos disponibles para la detección de TEs en cáncer hereditario tienen ciertas limitaciones, por ejemplo, las sondas empleadas en la hibridación genómica comparativa de matrices (aCGH) solo se unen a secuencias objetivo conocidas y no pueden detectar una insertada, además, la amplificación de sonda dependiente de ligadura múltiple (MLPA) sólo puede detectar inserciones cuando están en o cerca del sitio de ligación de la sonda (Qian *et al.,* 2017).

Otra de las dificultades recae en que existe una mayor cantidad de herramientas enfocadas en el estudio de las SNVs respecto a los TEs, donde además existe una baja precisión de los métodos *in silico* y una baja cantidad de genomas estudiados mediante secuenciación de lecturas largas (Kojima *et al.*, 2023).

Mientras los avances en la tecnología de secuenciación han mejorado la detección de VPs en *BRCA1/2*, esto no siempre resulta informativo para todos los casos de CM y OC con alta sospecha clínica, dichos casos pueden explicarse por VPs en otros genes de predisposición al cáncer e incluso puede que los individuos portan variantes como los TEs no detectadas en un primer momento (Deuitch *et al.,* 2020). Aunado a ello, con otras metodologías como la secuenciación Sanger, la identificación de elementos *Alu* en secuencias codificantes es un reto, por lo que estas alteraciones genéticas están menos documentadas que las puntuales (Solassol *et al.,* 2019).

1.11.1 Métodos analíticos

Secuenciación Sanger

Previo a la secuenciación se necesita una PCR para amplificar la región de interés, tras esto, la reacción emplea un primer de DNA único y DNA polimerasa para lograr una amplificación

lineal en vez de la exponencial (Bayrak-Toydemir y Wooderchak-Donahue, 2014). Esta técnica también es conocida como secuenciación de terminación de cadena, ya que está basada en el uso de una cantidad limitante de dideoxinucleótidos (ddNTPs) marcados con fluorescencia que carecen de grupo hidroxilo 3', de modo que el enlace fosfodiéster entre C3'OH de la última azúcar y el C5' del siguiente dNTP no se formará, resultando en la terminación de la cadena en dicho punto (Verma *et al.,* 2016).

El resultado es un conjunto de cadenas de DNA recién sintetizadas complementarias a la cadena molde pero que varían en longitud. Tras la purificación, la electroforesis capilar separa las cadenas de DNA por tamaño, el láser excita y detecta diferentes coloraciones asociadas a las bases nucleotídicas marcadas con fluorescencia (Bayrak-Toydemir y Wooderchak-Donahue, 2014).

Reacción en cadena de la polimerasa (PCR)

La PCR es una reacción enzimática dividida en ciclos (normalmente de 25-30), cada uno consistente de tres etapas que resultan en la amplificación exponencial de las moléculas molde. La desnaturalización tiene por objetivo la disociación a alta temperatura de las moléculas de doble cadena, la hibridación de los primers en su sitio complementario con la cadena molde y amplificación del DNA (Farrell, 2010). La especificidad y la eficiencia de la PCR implica que un número muy bajo de moléculas molde puede amplificarse en una gran cantidad de DNA producto, a menudo un microgramo o más que puede emplearse en análisis posteriores (McPherson y Møller, 2006).

Aunque dicha técnica puede detectar nuevas inserciones sin conocer a priori su localización genómica dificulta el análisis simultáneo de múltiples loci (Cardelli *et al.,* 2012). El producto de PCR adecuado para la detección va de 200-500 bases, aunque la longitud del análisis puede ampliarse hasta 10 mega de bases; es útil para analizar el número de repeticiones en la secuencia y CNVs. Tras la amplificación de la región de interés se recurre a la separación electroforética de los productos junto a un control para identificar diferencias en su tamaño.

El análisis de los fragmentos puede ser de utilidad en otros análisis como la pérdida de heterocigosidad, inestabilidad microsatelital y MLPA (Nakamura *et al.,* 2021). Sin embargo, la amplificación por PCR de las regiones de interés puede estar infrarrepresentada o no detectar alelos que contengan TEs debido al tamaño o la pérdida de alelos (Deuitch *et al.,* 2020).

El término “en tiempo real” hace referencia a que el producto de amplificación se monitorea en cada ciclo de reacción sin necesidad de recurrir a un gel de agarosa para saber si fue exitosa, mientras que el término “cuantitativo”, a que permite conocer la cantidad de DNA en la muestra. Cuando se emplea DNA genómico se habla de una qPCR, pero si se realiza a partir de cDNA, entonces se trata de una RT-qPCR (Tamay de Dios *et al.,* 2013).

La RT-PCR es una tecnología mediante la cual las moléculas de RNA son convertidas en secuencias de cDNA por una transcriptasa reversa (RT) seguida de su amplificación con una PCR estándar. Este método se centra en los loci transcripcionalmente activos, pues para que la transcripción y la amplificación ocurran debe existir un transcrito, lo que no ocurriría en caso de que el gen estuviera silenciado transcripcionalmente (Farrell, 2010).

Para verificar que la amplificación ocurrió se emplea el análisis Southern blot, que implica la transferencia de fragmentos de DNA de un gel de agarosa hacia una membrana de nylon mediante transferencia capilar para la posterior hibridación del DNA con una sonda específica (McPherson y Møller, 2006).

Amplificación de la sonda dependiente de ligadura multiplex (MLPA)

En el cáncer este ensayo es empleado en el análisis de línea germinal y somático de deleciones e inserciones en genes particulares y en el análisis de la metilación del DNA como mecanismo de inactivación de genes supresores de tumores (Stuppia *et al.,* 2012).

Esta técnica implica la desnaturalización del DNA, tras esto, se incuba con una mezcla de sondas específicas del gen, que consisten en dos oligonucleótidos inmediatamente adyacentes por cada exón objetivo, cada uno, compuesto por la secuencia a hibridar y la secuencia del primer para PCR. Tras la hibridación, las sondas se ligan y los fragmentos se amplifican con PCR a partir de los primers universales teñidos, cuyos productos serán separados por tamaño mediante electroforesis capilar (Bayrak-Toydemir y Wooderchak-Donahue, 2014).

A pesar de que ofrece hasta 40 objetivos con un alto rendimiento y posibilita la detección de pequeños reordenamientos, no puede detectar la pérdida de heterocigosidad y puede tener problemas por mosaicismos, heterogeneidad tumoral o contaminación con células normales (Stuppia *et al.,* 2012).

Hibridación genómica comparativa con microarreglos (aCGH)

En la aCGH las muestras del DNA de prueba y referencia se marcan con fluoróforos, se desnaturalizan e hibridan con clones que están en la superficie del chip de microarreglos. Así, las proporciones de fluorescencia brindan una medida en cada locus de las CNVs del DNA (Bayrak-Toydemir y Wooderchak-Donahue, 2014), la detección de estas aberraciones va desde unas 1000 bases, que podría corresponder al tamaño de un exón, hasta el nivel cromosómico en la región genómica completa (Nakamura *et al.,* 2021).

Su capacidad de detección de segmentos cromosómicos amplificados, deleciones y reordenamientos que abarcan uno o pocos genes la vuelve útil para descubrir la base molecular de síndromes genéticos (Shinawi y Cheung, 2012).

Secuenciación de próxima secuenciación (NGS)

Es un método masivo de secuenciación paralela que analiza 10G de bases o más. Los paneles de múltiples genes que usan los secuenciadores de lecturas cortas (100-200 bases) como Illumina y Thermo, son ampliamente utilizados con fines clínicos, los de lecturas largas (>10k) como los desarrollados por Oxford Nanopore y Pacbio, son útiles para analizar anomalías estructurales genómicas con fines de investigación (Nakamura *et al.,* 2021).

El primer paso para el análisis de datos obtenidos mediante NGS es combinar las diferentes lecturas de cada clúster. Las lecturas que provienen de DNA genómico se pueden alinear en una secuencia continua de información ya sea comparándolo con un genoma previamente secuenciado o comparando las lecturas buscando la superposición de secuencias (Clark *et al.,* 2019).

Dentro de las tecnologías de análisis de secuencias genómicas, la técnica de secuenciación del exoma completo (WES) se utiliza para dilucidar las causas de enfermedades hereditarias para las que no se ha identificado un gen causante. La principal diferencia respecto a la secuenciación del genoma completo (WGS) radica en que este último proporciona información sobre la secuencia de todas las regiones del genoma mientras que el WES se enfoca solamente en las regiones exónicas, que corresponden menos del 2% del genoma humano (Nakamura *et al.,* 2021)

Aunque podría parecer simple descubrir variantes en el genoma al contar las discordancias en cada posición respecto a un genoma de referencia, existen múltiples fuentes de error: durante la preparación de bibliotecas, errores en la secuenciación y artefactos de mapeo

cuando se alinean las lecturas. Por esa razón se necesitan implementar métodos que los compensen o corrijan, esto es, algoritmos que permitan un procesamiento adecuado de los datos como parte del flujo de trabajo en el llamado de variantes, como pueden ser BWA y GATK (Van der Auwera, 2013).

Laboratorios comerciales ofrecen pruebas genéticas que rondan entre 270-340 USD, pero en muchos casos se limitan a *BRCA1/2* (Bose *et al.,* 2022). El aumento en el uso de paneles de múltiples genes (que seleccionan docenas de ellos entre más de 20,000) se debe a la rentabilidad y comodidad para disponer de pruebas que incluyan los genes asociados a cada tipo de cáncer hereditario (Lee *et al.,* 2019).

Los ensayos de secuenciación de siguiente generación (NGS, por sus siglas en inglés) pueden revelar de forma consistente la presencia de TEs comparado con métodos tradicionales, pero es necesaria la confirmación y la caracterización de las inserciones detectadas con métodos complementarios para la evaluación de su impacto funcional (Qian *et al.,* 2017). Sin embargo, presentan algunas desventajas, pues al leer solo las regiones codificantes se excluyen las variantes que podrían ocurrir en regiones reguladoras, promotoras e intrónicas. Muchos de estos ensayos multigénicos analizan los bordes de las regiones exón-intrón en un rango de 2-5 pb (Nakamura *et al.,* 2021).

Aunado a ello, la identificación de un gran número de VUS requiere avances en la investigación epidemiológica que esclarezcan su contribución al riesgo. Aún con todo, la identificación de variantes genéticas que predisponen al SHCMO es de utilidad en el manejo clínico de las pacientes y sus familiares, ya que de este modo se puede recurrir a medidas de prevención de riesgo, planes de vigilancia y asesoramiento genético (Flores, 2019).

1.11.2 Métodos in silico

Existen herramientas empleadas en la detección de familias conocidas de TEs que por lo tanto necesitan de bibliotecas para realinear las lecturas que no coinciden con el genoma de referencia, aunque en otros casos estas son prescindibles. De acuerdo con el tipo de inserciones que detectan pueden enfocarse en inserciones de línea germinal (que suelen provenir de secuencias genómicas obtenidas de muestras de sangre) e inserciones de novo (secuencias provenientes del tejido de interés y un control) (Chu *et al.,* 2021).

El paradigma algorítmico básico en el que operan las herramientas para el análisis de lecturas cortas pareadas involucra dos patrones de alineación de lecturas cerca de los puntos de ruptura de cada inserción que apuntan a la presencia de un TE: Pares de lectura discordantes

(DR, por sus siglas en inglés) y lecturas divididas o recortadas (SR, por sus siglas en inglés) (Rishishwar *et al.,* 2017).

Las DRs indican que la lectura de un extremo está completamente mapeada con la referencia, pero el otro extremo no lo está, un SR se refiere a un par de lecturas donde una parte está parcialmente mapeada a la referencia. Una lectura pareada (RP, por sus siglas en inglés) provee un sitio de inserción putativo y un intervalo del punto de ruptura basándose en grupos de DRs en los extremos 5' y 3' (Lee *et al.,* 2022).

Sin embargo, existen otras estrategias que se han empleado en la detección de elementos *Alu* polimórficos (Figura 3), tal es el caso de *Alu*Mine, que utiliza métodos para WGS sin alineamiento que se complementan con la llamada de genotipos con inserciones previamente conocidas a partir de lecturas en bruto (Puurand *et al.,* 2019). Otras aproximaciones de algoritmos como T-lex2 se basan en la implementación de módulos que combinan mapeo y cobertura de profundidad para identificar aquellas lecturas que aporten pruebas de la presencia o ausencia de TEs (Fiston-Lavier *et al.,* 2015).

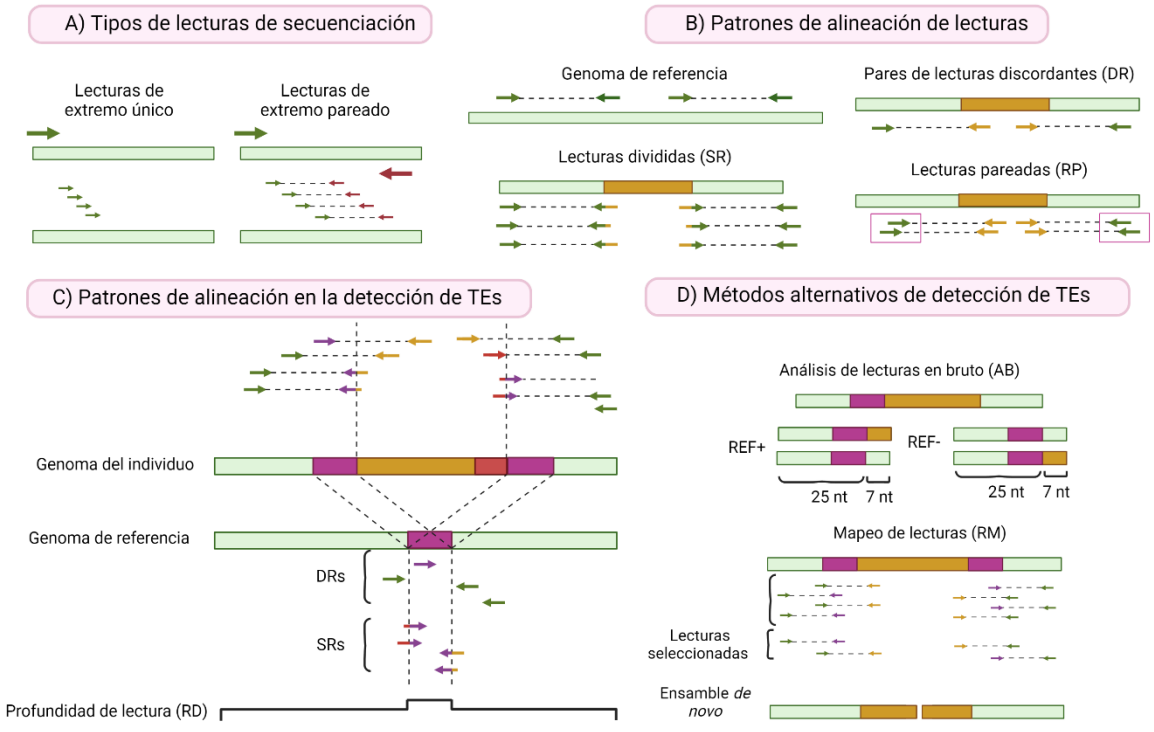


Figura 3. Estrategias para la detección de TEs a partir de lecturas cortas de extremo pareado.
A: Las flechas verdes representan lecturas en dirección sentido, las rojas son antisentido. B: La franja de color naraja representa una TE; cada recuadro rosa corresponde a la agrupación de DRs. C: La franja morada corresponde al TSD, mientras que la roja a la secuencia poli-A; en la parte inferior se muestran los patrones de alineación en el genoma del individuo que permiten inferir la presencia de TE, cuya alineación de lecturas se infiere como se muestra en la parte superior. En la parte inferior se muestra el cambio en la profundidad de lectura característico de una inserción fuera de la referencia. Las líneas discontinuas indican el límite de los TSD. D: Se basa en la formación de una base de datos de parejas de fragmentos de 32 nt para luego filtrar

cuáles apoyan TEs en la referencia (REF+) o en el individuo (REF-). En RM se recurre a un primer filtrado de las lecturas basadas en su calidad y tras su ensamblaje de novo, de acuerdo con su longitud e identidad de secuencia. Figura elaborada en Biorender basada en los esquemas de Fiston-Lavier et al., 2012; Puurand et al., 2019; Chu et al., 2020 y Lee et al., 2022.

Dentro de los algoritmos que se apegan al paradigma tenemos herramientas bioinformáticas como RetroSeq, MELT, Mobster, TEMP, *Alu*-Detect, Tangram, IMGEins, Tea, xTEA, ITIS, TEfinder (Tabla S1). Los cuatro primeros mencionados fueron desarrollados para la detección de inserciones de línea germinal. Herramientas como Tea, TraFiC, MELT, Mobster y xTea anotan características específicas del TE detectado como la subfamilia, tamaño de la inserción, orientación, secuencia de la TSD y la presencia de colas poli (A) (Chu *et al.,* 2020)

Para conocer el desempeño de las herramientas se recurre a medidas como la precisión o valor de predicción positiva, que se calcula como TP/ (TP+FP), donde TP son los verdaderos positivos y FP, los falsos. La recuperación, sensibilidad o tasa de verdaderos positivos, se calcula como TP/ (TP+FN), donde FN son los falsos negativos (sitios conocidos que no fueron predichos en el rango establecido contiguo al sitio de inserción), aunque también suele reportarse basada en PCR (Figura 4). La puntuación F1 es un valor que se calcula con la media armónica de los dos anteriores (Rishishwar *et al.,* 2017).

En un estudio comparativo del desempeño de diversos algoritmos sobre datos de Illumina NA12878 y simulados, MELT y Mobster mostraron los mejores resultados cuando se evaluó su precisión y recuperación. Sin embargo, Retroseq, Tangram y Mobster exhiben mejores métricas de recuperación para L1 simulados que MELT (Kosugi *et al.,* 2019). Empleando el mismo conjunto de datos, tratándose de la evaluación de TEs polimórficos, MELT se destaca con el mejor desempeño tratándose de datos de baja cobertura (5.7x) seguidos de Mobster y Retroseq, sin embargo, todas las herramientas disminuyeron su desempeño cuando se evaluaron con datos de alta cobertura (95.6x), en este caso RetroSeq y Mobster fueron mejores, pero con un alto número de falsos positivos y una precisión baja (Rishishwar *et al.,* 2017).

La detección de polimorfismos también ha sido evaluada con datos de otras especies como *Oryza sativa ssp*. japonica cv. Nipponbare, donde algoritmos como PoPoolationTE2, Teflon y Jitterbug tuvieron los mejores desempeños, al extrapolarse a datos humanos sólo PoPoolationTE2 mantiene los mejores resultados para detectar inserciones heterocigotas, mientras que MELT lo supera para homocigotas (Vendrell-Mir *et al.,* 2019).

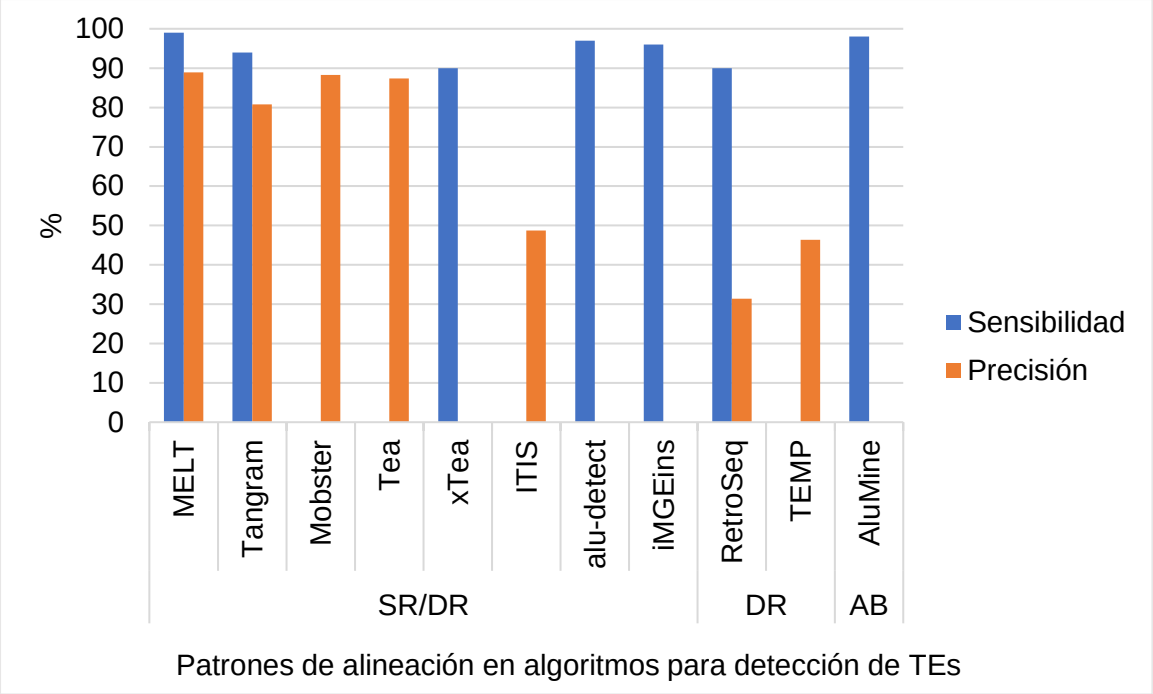


Figura 4. Sensibilidad basada en PCR y precisión en conjunto de datos no simulados de Illumina NA12878 empleados en algoritmos utilizados para la detección de TEs.
DR= Pares de lecturas discordantes, SR=Lecturas divididas y AB=Análisis de lecturas en bruto. Figura basada en lo reportado por Kosugi et al., 2019 y Lee et al., 2022.

Existen bases de datos de inserciones polimórficas, tales como dbRIP, no disponible actualmente, pero que recupera una lista con 1625 elementos *Alu*, 407 L1 y 63 SVA (Wang *et al.,* 2006). La librería unificada tras la tercera fase del proyecto de los 1000 genomas, que recurrió a datos de WGS de 2504 individuos y a nueve algoritmos enfocados en el descubrimiento de SVs sumó un total de 12748 elementos *Alu*, 3048 L1 y 835 SVA (Sudmant *et al.,* 2015).

2. ANTECEDENTES

2.1 Inserciones *Alu* en SHCMO

En individuos sometidos a pruebas genéticas de cáncer hereditario empleando secuenciación Sanger, NGS, qPCR múltiplex y PCR dirigida para la caracterización de TEs, se identificaron 34 elementos *Alu* en 10 genes que afectaron a 211 individuos (0.012% de la cohorte), de los cuales 70 fueron en individuos con SHCMO. Los TEs fueron más comunes en *BRCA2* y *ATM*. Por su parte, 32% de los individuos con una inserción tenían ancestría europea y sólo 14.7% en América Latina y del Caribe. En 5 individuos de esta última población, se identificó una inserción patogénica c.5007_5008ins*Alu*, localizada en el exón 11 de *BRCA2*, que conduce a una mutación de cambio de marco de lectura que causa la omisión del exón y NMD (Qian *et al.,* 2017).

La inserción *Alu* más estudiada en SHCMO es c.156_157ins*Alu*, localizada en el exón 3 de *BRCA2*. Esta fue detectada en 14/208 familias del norte y centro de Portugal sin mutaciones patogénicas en *BRCA1/2* evaluada mediante electroforesis en gel de gradiente desnaturalizante (DGGE) y MLPA. Para la detección de c.156_157ins*Alu* se realizaron dos rondas de PCR. Finalmente, el splicing del RNA se evaluó amplificando la región codificante de *BRCA2* del exón 1 al 6. Con todo lo anterior, se pudo demostrar su efecto patogénico y que la presencia de dicha mutación representa una cuarta parte de las mutaciones deletéreas en *BRCA1/2* de familias con SHCMO en dicha región (Peixoto *et al.,* 2008)

Debido a la migración durante la colonización, c.156_157ins*Alu* tiene una frecuencia baja, aunque relevante en la población brasileña (Felicio *et al.,* 2018), siendo considerada la tercera variante más común. Esta se ha reportado acompañada de p. Ala938Profs*21, p. Tyr3009Serfs*7 y p. Arg3128Ter en 4 pacientes de una cohorte de 95 tras haber evaluado la totalidad de muestras con 2 rondas de PCR, una para la región exónica de *BRCA2* y otra específica para el fragmento *Alu*. Su patogenicidad se atribuye a que saltarse el exón mencionado causa la pérdida del sitio de unión de *PALB2* y *RAD51*, esenciales para la reparación por recombinación homóloga (Da Costa *et al.,* 2020).

Por otro lado, en una familia europea *BRCA* negativa, 26 genes de 2 pacientes fueron analizados mediante NGS. El procesamiento bioinformático se realizó usando el software SOPHiA DDM en sus versiones v4.5.1 y v5.4.0. La caracterización de la mutación c.4475_4476ins*Alu*, localizada en el exón 14 se desarrolló utilizando dos PCRs independientes, posterior a ello, se realizó un análisis RT-PCR con primers abarcando la región entre los exones 13 y 15 que evidenció un splicing anormal que ocasiona un

corrimiento en el marco de lectura y un codón de paro prematuro, lo que permite considerar la inserción como deletérea para la estructura funcional proteica (Bouras *et al.,* 2021).

Un TE patogénico, c.2872_2888delins114*Alu*L2, fue detectado en la región codificante del exón 9 del gen *PALB2* de una paciente con predisposición al CM sin variantes patogénicas previamente identificadas por análisis de datos mediante NGS. Utilizaron scripts de R (v.3.5.3) para detectar puntos de ruptura y el algoritmo Trinity (v.2.8.4) para predecir las secuencias consenso alrededor de ellas.

Los contigs reconstruidos se filtraron y las secuencias seleccionadas se anotaron con BLAST (2.0.9+) y RepeatMasker (Smit *et al.,* 2013-2015) para luego crear una representación con IGV-snapshot-automator. Aunque 15 variantes detectadas en 55 muestras fueron validadas con una PCR y secuenciación Sanger, sólo se observó un efecto patogénico con la variante mencionada debido a sus efectos en el truncamiento proteico (p. Gln958Valfs*38) (Eyries *et al.,* 2022).

En una paciente de 53 años con ancestría italiana, noruega y alemana se detectó una duplicación del exón 13 de *PALB2* utilizando el Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) que fue confirmada con MLPA y cuya interrupción en el empalme del mRNA asociada al SHCMO se caracterizó mediante RT-PCR, clonación y secuenciación del DNA. Mediante RepeatMasker se encontraron 8 elementos *Alu*Sg en el intrón 12, además de 5 elementos *Alu*Sc8 en la región 3’ UTR de *PALB2*, lo que implica la recombinación homóloga como mecanismo para explicar la duplicación en tándem (Yang *et al.,* 2016).

En paciente emiratí de 40 años con CM unilateral que no mostró VPs en *BRCA1/2* y *PALB2* en las pruebas clínicas iniciales (se reportaron 2 VUS), el análisis con Mobster detectó una nueva inserción SVA patogénica c.7894_7895insSVA en *BRCA2* asociada a SHCMO (Deuitch *et al.,* 2020).

El estudio más reciente al respecto describe una inserción *Alu* en el intrón 54 de *ATM* detectada con Mobster en 2/303 pacientes alemanes donde no se presentó una VPP tras la secuenciación del panel multigénico y en 4/242 pacientes de otra cohorte con SHCMO, con SCRAMble se detectaron 2 casos más dando una frecuencia alélica total de 0.05%. Dicha inserción fue validada mediante PCR y secuenciación Sanger, mientras que su efecto fue evaluado con RT-PCR y el ensayo de minigen, que permitió comprobar su efecto en la omisión del exón 54 (Klein *et al.,* 2023).

2.2 Inserciones *Alu* que predisponen a CM

Empleando datos de WES generados con el kit SeqCap 101 EZ Exome v3.0 de Roche junto con el software RetroSeq sumado a un pipeline propio aplicado a 65 archivos BAM obtenidos a partir de muestras de pacientes con CM, se detectaron 12 elementos *Alu*, 2 L1, 1 SVA y 3 pseudogenes (sólo 2/18 eventos de retrotransposición en regiones codificantes), todos validados mediante dos PCR consecutivas. Al investigar las consecuencias moleculares de 5 inserciones patogénicas candidatas mediante RT-PCR y secuenciación Sanger del RNA se encontró que la inserción en la región 3'UTR de GHR reprimía la expresión del alelo y que la cercana al exón de *PTPN14* (un mediador de la actividad supresora tumoral de P53) puede conducir a la pérdida de la impronta (De Brakeleer *et al.,* 2020).

Se reportaron inserciones de novo de dos elementos *Alu* a partir de ensayos como la prueba de truncamiento de proteínas (PTT) dirigida hacia el exón 11 de *BRCA1* y 10 y 11 de *BRCA2*, aunado al análisis de los exones restantes mediante electroforesis en gel sensible a la conformación (CSGE) en 3/50 familias belgas con una mutación de predisposición al cáncer. Dicha metodología resulta poco sensible pues una inserción se detectó mediante Southern blot (c.156_157ins*Alu*) y en la otra, tras obtener una proteína trunca, se recurrió a primers específicos para la secuencia *BRCA1* y el elemento *Alu* (c.1739_1740ins*Alu*) (Teugels *et al.,* 2005).

3. HIPÓTESIS

Considerando que existen TEs cuya inserción ocurre en genes asociados al SHCMO, este mecanismo de mutagénesis insercional podría estar relacionado con la patogénesis molecular en pacientes de América Latina que cumplen con los criterios de selección del SHCMO y que previamente no presentaron variantes patogénicas en genes de susceptibilidad a tumores hereditarios.

4. OBJETIVO GENERAL

Evaluar la prevalencia de mutagénesis insercional causada por los TEs en genes de susceptibilidad a tumores hereditarios en pacientes de América Latina con SHCMO.

5. OBJETIVOS PARTICULARES

- Establecer el flujo de trabajo bioinformático para la detección de elementos móviles transponibles.
- Realizar el procesamiento bioinformático de las muestras de las pacientes con SHCMO secuenciadas mediante paneles de genes.
- Identificar los elementos móviles transponibles de los datos de secuenciación masiva.
- Validar por secuenciación Sanger los elementos móviles identificados en el análisis bioinformático.

6. JUSTIFICACIÓN

A pesar de los avances en la capacidad de las tecnologías de secuenciación masiva, las guías de diagnóstico para el SHCMO se enfocan en la detección de variantes patogénicas en genes de alta penetrancia (*BRCA1/2, PALB2, TP53),* mediana (*ATM, CHEK*2) y baja penetrancia (*MSH6* y *PMS2)*. Sin embargo, sólo el 15-25% de las pacientes presentan VP en estos genes, 20% VUS, 10% CNV y aproximadamente el 50% de las pacientes son negativas para estas alteraciones genéticas previamente mencionadas. Por lo cual, es de importancia buscar mecanismos alternativos de patogénesis molecular en pacientes con SHCMO que son negativos a variantes patogénicas.

Por otro lado, se tiene conocimiento que en el SHCMO la inserción de TEs en regiones exónicas de genes como *BRCA1/2* y *PALB2* pueden causar la omisión de exones durante el splicing que conduce a productos proteicos no funcionales o al decaimiento de los transcritos por un codón de paro prematuro mediante NMD. Aunque en otras enfermedades los efectos

en la regulación no se limitan solo a su inserción en la región exónica, la dificultad para su detección con pruebas moleculares rutinarias sugiere una subestimación de los TEs patogénicos, ya que a la fecha se sabe que representan sólo 0.03 al 0.1% de las variantes responsables de enfermedades genéticas.

A pesar de la exclusión de las regiones reguladoras, intrónicas y promotoras donde la inserción de un TE podría tener un efecto en la regulación del gen, el análisis que confirme la presencia o ausencia de dichas inserciones en el nivel exónico que comprenden los paneles resulta de utilidad para esclarecer su contribución al riesgo en la población de América Latina, debido a que el número de publicaciones donde se aborda su detección y análisis funcional es preponderantemente europea. Fuera de estudios donde se han identificado elementos *Alu* con efectos patogénicos en una población estadounidense con ancestría de América Latina a la fecha no existen trabajos con un enfoque en la caracterización de TEs de la población mencionada, lo que se propone precisamente este estudio para explorar la posibilidad de que el genotipo se asocie a su presencia.

El presente trabajo tiene el objetivo de identificar elementos móviles transponibles en pacientes de América Latina que son negativas a variantes patogénicas en genes de susceptibilidad a tumores hereditarios, como un mecanismo de patogénesis molecular en el SHCMO.

7. METODOLOGÍA

7.1 Población de estudio

El tipo de estudio que se realizó en la siguiente tesis es de tipo descriptivo, retrospectivo y transversal. Para este fin, se reclutaron un total de 1477 pacientes de América Latina: i) 1225 pacientes mexicanas, ii) 113 colombianas, iii) 90 peruanas y iv) 49 guatemaltecas, que cumplieron con los criterios establecidos por la NCCN v.2022 para el SHCMO, con la aprobación del comité de ética en investigación (ISEM-02092015; INSP-CI:1065; INSP-341) y llevada a cabo bajo los criterios de Helsinki.

A las pacientes reclutadas se les extrajo una muestra de sangre y con ello se realizó la extracción del DNA. La integridad y la calidad del DNA se evaluaron mediante un gel de agarosa y la cantidad mediante fluorometría. La preparación de la biblioteca de secuenciación masiva se realizó usando dos paneles de genes: i) Qiagen de 143 genes para 1163 pacientes y ii) SOPHiA Genetics de 73 genes para 314 pacientes (292 pacientes y 22 pacientes control) (Tabla 2).

En el caso del panel de SOPHiA Genetics, las pacientes control se incluyeron con la finalidad de asegurar homogeneidad durante la secuenciación tanto entre las corridas como dentro de ellas. Se analizaron: i) 5 controles intracorridas, ii) 8 controles intercorridas, iii) 1 control inter e intracorrida y iv) 8 muestras de pacientes con VPs que generan codones de paro en *BRCA2* (GT7 y GT275) y *PALB2* (GT298); de cambio de marco de lectura en *FANCI* (GT55) y *BRCA1* (GT44); con un sitio donador de splicing +1 en *BRCA2* (G291) y de cambio de sentido en *BRCA2* (GT278) y *BRCA1* (GT35).

Tabla 2. Muestras de 1477 pacientes de LACAM analizadas mediante paneles de genes de SOPHiA y Qiagen

<table>
<tr><th rowspan="2">Panel</th><th rowspan="2">País</th><th colspan="7">ID de la corrida</th><th colspan="2" rowspan="2">Total</th></tr>
<tr><th>C1</th><th>C2</th><th>SG1</th><th>SG2</th><th>CHCS4</th><th>CHCS5</th><th>C</th></tr>
<tr><td rowspan="2">SOPHiA Genetics</td><td>MEX</td><td>24</td><td>27</td><td>72</td><td>72</td><td>9</td><td>25</td><td>22</td><td>251</td><td rowspan="2">314</td></tr>
<tr><td>COL</td><td colspan="7">C3
63</td><td>63</td></tr>
<tr><td rowspan="5">Qiagen</td><td rowspan="2">MEX</td><td>Tol-IMSS</td><td>Tol- ISSEMYM</td><td>FAS</td><td>GT</td><td colspan="2">IMSS</td><td>IXT</td><td rowspan="2">974</td><td rowspan="5">1163</td></tr>
<tr><td>92</td><td>92</td><td>67</td><td>269</td><td colspan="2">435</td><td>19</td></tr>
<tr><td rowspan="2">COL</td><td colspan="7">Colombia_113
50</td><td rowspan="2">50</td></tr>
<tr><td rowspan="2">PER</td><td colspan="7">PERU_92
90</td><td rowspan="2">90</td></tr>
</table>

	GUA	GUATEMALA_50	49	
		49		

MEX: México, COL= Colombia; PER: Perú; GUA=Guatemala. C1: Corrida 1, C2: Corrida 2, C3: Corrida 3, C: Pacientes control, FAS: Fundación Alma Salvati, GT: Gaby Torres, ID: Nombre del identificador de la corrida, IMSS: Instituto Mexicano del Seguro Social; IXT: Ixtapaluca, Tol-IMSS: Instituto Mexicano del Seguro Social en Toluca, Tol- ISSEMYM: Instituto de Seguridad Social del Estado de México en Toluca.

7.2 Identificación *in silico* de elementos *Alu*

Los archivos crudos de la secuenciación masiva en formato fastq se alinearon con el programa bwa mem (Li y Durbin, 2009) usando el genoma de referencia humano hg19. El preprocesamiento de datos fue realizado con GATK (Van der Auwera *et al.,* 2013) para generar los archivos bam que se utilizaron para la identificación de los elementos móviles transponibles.

La detección de los elementos *Alu* se realizó con el programa MELT (Gardner *et al.,* 2017). Para el filtrado de las variantes se excluyeron aquellas con alguno de los siguientes filtros: s25 (más de 25% de las muestras sin datos), rSD (proporción de DRs de la izquierda con DRs de la derecha mayor a 2 desviaciones estándar), hDP (más del número esperado de DR también son SR), ac0 (no hay individuos en el archivo vcf identificados en la inserción) e lc (dentro de una región de baja complejidad). Otro parámetro empleado fue ASSESS, que aporta evidencia de la inserción de un elemento *Alu* bajo los siguientes criterios: (1) punto de ruptura impreciso debido a distancia mayor a la esperada entre la evidencia, (2) solo evidencia de DR (no hay de SR), (3) solo evidencia a la izquierda del sitio de duplicación objetivo (TSD), (4) TSD apoyado por SRs, (5) mayor calidad posible, por lo cual, aquellas variantes con ASSESS < 5 y SR ≤ 2 fueron excluidas.

Se emplearon algoritmos adicionales centrados en la detección de TEs a partir de DRs como RetroSeq (Keane *et al.,* 2013) y TEfinder (Sohrab *et al.* 2021). En el caso de RetroSeq los archivos bam generados fueron utilizados para el llamado de variantes, donde se consideró como candidatas aquellas con los siguientes valores: i) calidad del genotipo (GQ) ≥28 y confiabilidad de la llamada basada en el número de lecturas cercanas al punto de ruptura (FL, definido por los desarrolladores del algoritmo) ≥7 o bien, ii) FL=6 y GQ≥28.

En FL el valor de 8 representa la mayor confiabilidad, mientras que GQ brinda información sobre el número de lecturas de apoyo. Posteriormente se crearon archivos BED individuales para cada grupo de TEs a partir del archivo BED del genoma completo obtenido desde el UCSC Table Browser (Karolchik *et al.,* 2004), contemplando las siguientes categorías:

transposones de DNA, SINEs, LINEs, LTRs, antiguos y desconocidos de acuerdo con la clasificación empleada en Repbase (Kojima, 2018)

En el caso de TEfinder, tras emplear bwa index para el genoma de referencia humano, se realizó el proceso de alineación con bwa, esta vez obteniendo el archivo bam directamente desde los fastq. Se excluyeron aquellas variantes que incluyeran los filtros weak_evidence (donde el total de lecturas para cada evento de inserción fue menor a 10) y strand_bias (cálculo basado en el recuento de lecturas sentido (F) y reversa (R) que debe caer en los rangos R < F*0.8 o R > F*1.25 para considerarse un TE presente/ausente en dicha posición cromosómica. Por su parte, se incluyeron aquellas con el filtro in_repeat (sitio de inserción coincidió con un sitio anotado como TE en la secuencia del genoma de referencia).

7.3 Visualización de elementos *Alu* identificados *in silico*

Se realizó la visualización individual de los candidatos mediante el programa Integrative Genomics Viewer (IGV) en su versión 2.14.0 (Thorvaldsdóttir *et al.,* 2013) seleccionando la agrupación de alineamientos por cromosoma de coincidencia y el color de alineamientos por tamaño de inserción y orientación por pares, una vista colapsada y habilitando la visualización de SRs.

El archivo BED de repeticiones del genoma completo versión hg19 de RepeatMasker obtenido desde el UCSC Table Browser se adicionó como un track para las visualizaciones. En los casos donde se presentó un patrón de DRs y SRs concordante con un TE, en un panel adicional se observó la región de las SRs coincidentes para determinar si la posición coincidía con un TE del track mencionado.

7.4 Diseño de primers

Una vez identificada los elementos *Alu* candidatas, se realizó el diseño de primers para la validación de dicho elemento tomando en cuenta las siguientes consideraciones: el tamaño del primer debe oscilar entre 15-25 pares de bases, contenido de GC 45-55% (Tamay de Dios *et al.,* 2013), sin pares 4 pares de bases idénticas consecutivas, un rango de temperatura entre cada par de primers no mayor a 3°C, así como complementariedad entre los extremos 5’ y 3' menor o igual a 6.

Además, se evaluó la especificidad de los primers con Primer-BLAST (Ye *et al.,* 2012), se corrió una PCR *in silico* usando la base de datos de UCSC genome browser (Kent *et al.,* 2002), así como los casos donde no se predijera la formación de horquillas mediante

OligoCalc (Kibbe, 2007). Antes de la validación, se estandarizó las condiciones óptimas de los primers.

7.5 Validación de elementos *Alu*

7.5.1 PCR

Se extrajo DNA como se describe anteriormente en tres pacientes: Col-069, Col2-014 y GT245 que se les identificó los elementos Alu por NGS. La PCR se llevó a cabo usando GoTaq Green Master Mix 2X (Promega) en un volumen final de 50 µL usando las siguientes concentraciones finales: Green Master Mix 1X, 0.1 µM primer forward y reverse para *ATR* (Col-069), *RB1* (Col2-014) y 0.2 uM para *MSR1* (GT245), 25 ng de DNA genómico. Las condiciones de la PCR fueron las siguientes: desnaturalización inicial a 95°C por 5 minutos seguido de 35 ciclos de desnaturalización a 95°C por 30 segundos, hibridación de los primers a 60°C para *ATR*, 58°C para *RB1*, 58°C para *MSR1*, y extensión a 72°C en los tres casos por 60 segundos, con una extensión final a 72°C por 5 minutos.

7.5.2 Secuenciación Sanger

Los productos de PCR fueron purificados usando perlas Ampure y se enviaron al Laboratorio de Secuenciación Genómica de la Biodiversidad y de la Salud, Laboratorio Nacional de Biodiversidad del Instituto de Biología de la UNAM para la secuenciación Sanger con un equipo 3730xl. Los electroferogramas fueron analizados con el software ApE v3.1.4 (Davis y Jorgensen et al., 2022)

8. RESULTADOS

8.1 Identificación in silico de elementos *Alu* en el panel de SOPHIA Genetics

En este trabajo, se analizó a 1477 pacientes de 4 países de América Latina. Respecto a las 314 pacientes que fueron secuenciadas con el panel de genes de SOPHiA Genetics (251 mexicanas y 63 colombianas) y analizadas con el algoritmo de MELT, se predijeron 131 TEs en 57 muestras, de las cuales 7 TEs cumplieron con los criterios de inclusión ASSESS=5 y 2 ≤ SRs en 11 muestras. Al respecto, 5 de ellas presentaron un TE en el chr13: 49034102-49034119 en la región intrónica de *RB1* (Tabla 3). En los controles se detectaron 4 TEs en 6 de 22 muestras donde únicamente chr8: 15978150 en la región intrónica de *MSR1* cumplió con los criterios de inclusión.

Tabla 3. Resumen de TEs que cumplieron los criterios de inclusión con el algoritmo MELT.

Muestra	País	Corrida	SR	Filtro	Posición predicha	T	TSD
GT245	MEX	C2	6	lc, ac0	chr8: 15978150 *MSR1* (intrónico)	215	$(A)_{23}$
Col-069	COL	C3	12	lc, ac0	chr3: 142231080 *ATR* (intrónico)	269	$(A)_{18}$
Col2-014	COL	C3	9	lc, ac0	chr13: 49034104 *RB1* (intrónico)	80	dCTCTTTCTTTC
LAM550043_S66	MEX	SG1	20	hDP, lc	chr13: 49034102 *RB1* (intrónico)	79	dCTCTCTTTCTTTC
LAM550018_S51	MEX	SG1	12	hDP, lc	chr13: 49034119 *RB1* (intrónico)	61	TTTT
LAM530017_S30	MEX	SG1	31	hDP, lc	chr13: 49034119 *RB1* (intrónico)	268	TTTT
LAM550001_S107	MEX	SG2	14	lc, ac0	chr13:49034088 *RB1* (intrónico)	71	dTTCTTTCTTTCTTTCTCTCTTTCTTTC
GT121SG_S14	MEX	CHCS4	2	PASS	chr7: 105564465 *CDHR3* (intrónico)	280	AAGACTGAGG
GT33B_S22	MEX	CHCS4	2	PASS	chr7:115865235 *TES* (intrónico)	281	AAAAACACT
GT121SG_S14	MEX	CHCS4	2	hDP	chr10:48419952 Región intergéncia de *RBP3* y *GDF2*	263	ATAAAATTT
GT481_S19	MEX	CHCS4	4	PASS	chr13: 79271601 Región intergéncia de *OBI1-AS1* y *LINC00331*	281	AGAAATGGGCATATTC

*MEX: México, COL= Colombia. T: Tamaño, TSD: Sitio de duplicación objetivo; ASSESS: proporciona una evaluación de la precisión para determinar el punto de ruptura en cada inserción; Ic: Filtro sitios donde la inserción está dentro de +/- 25pb; ac0: Filtra sitios sin un alelo genotipado, hDP: Filtro de divisiones discordantes que se muestra cuando el total de (SRs/DRs)*100 ≥ 55.*

Del total de 57 muestras donde se predijeron TEs con MELT, 50 corresponden a probandos, el resto al grupo control. Por otro lado, 100/131 TEs predichas fueron eventos únicos, de los cuales 98 fueron pacientes y 2 controles entre corridas (GT245 y GT84).

Las pacientes de las corridas 1, 2 y 3 y el grupo control de SOPHiA Genetics (n=136) se analizaron con el algoritmo TEfinder y RetroSeq. TEfinder identificó 14 TEs en 101/136 muestras: 81 pacientes, 8 controles, 3 controles intracorrida, 7 controles intercorrida, 1 control inter en intracorrida y 1 control CNV. No se presentó ningún evento de transposición único con este algoritmo. RetroSeq predijo 335 TEs en 118/136 muestras: 104 pacientes 8 controles intercorridas, 4 controles intracorridas, 1 control inter e intracorrida y 1 control CNV. Con este algoritmo 211/335 TEs predichas correspondieron a inserciones predichas únicas.

No se reportaron coincidencias entre los tres algoritmos (Figura 5). Todos los casos predichos por TEfinder corresponden a TEs que están en el genoma de referencia, lo que explica su alta frecuencia. De las 5 TEs que se presentaron tanto en RetroSeq como en TEfinder, solo 2 de ellas fueron reportadas en la misma paciente: chr11:111963610-111964078, que ambos algoritmos detectaron en Col-090 y Col-072 y chr22:29126286-29126620, donde esto ocurrió para Col-061 (Tabla S1).

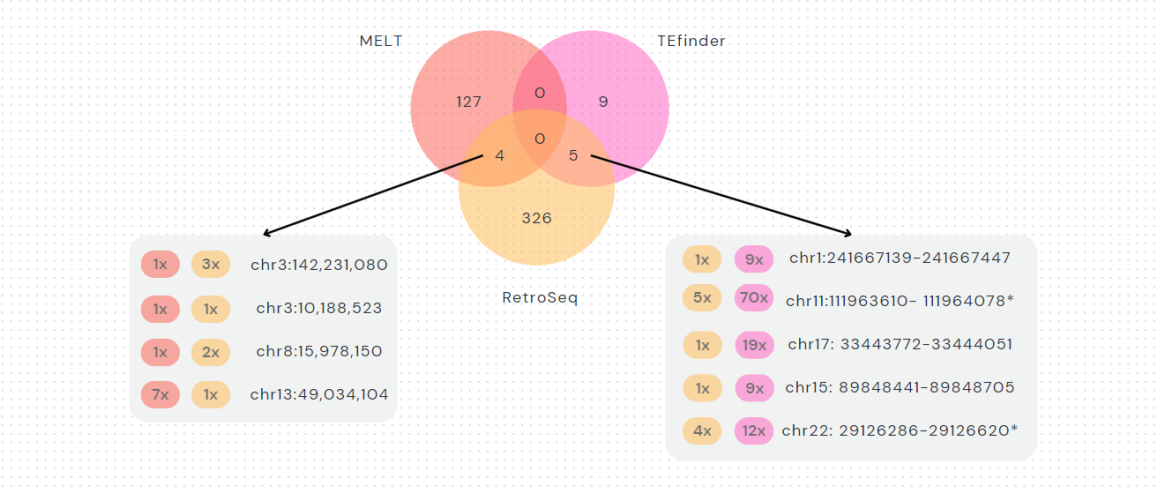


Figura 5. Número de TEs detectados con los algoritmos MELT, RetroSeq y TEfinder.
*En los recuadros puede apreciarse la posición cromosómica de los TEs coincidentes en al menos dos algoritmos. En el lado izquierdo los TEs predichas coincidentes entre MELT y RetroSeq, en el derecho, entre RetroSeq y TEfinder. Se presenta con una x la frecuencia de cada TE predicha. * Al menos una de las muestras es compartida, para más detalles consultar la tabla S1.*

8.2 Identificación in silico de elementos *Alu* en el panel de Qiagen

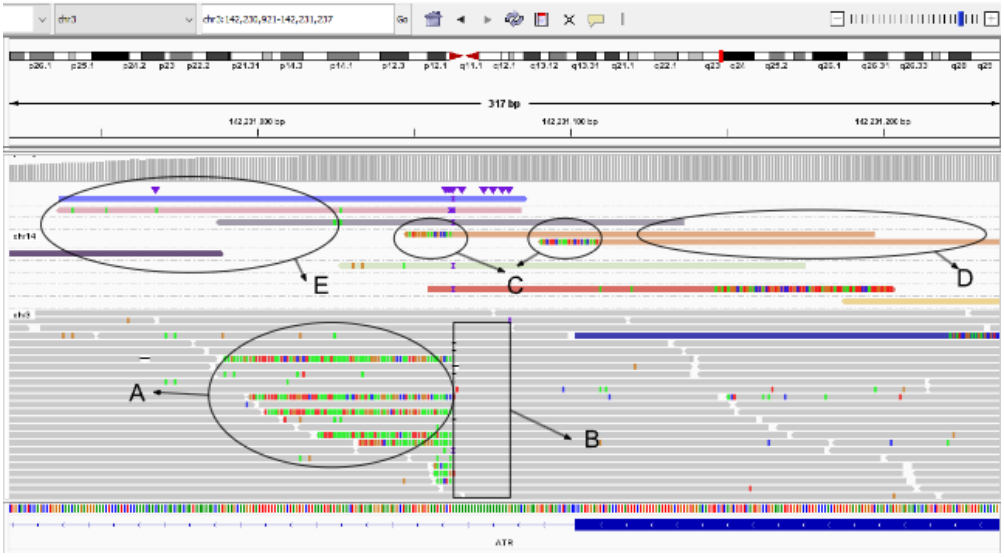
Se analizaron 1163 pacientes usando el panel de Qiagen con el algoritmo de MELT y se descartó el análisis con las herramientas bioinformáticas RetroSeq y TEfinder debido a la gran cantidad de falsos positivos resultantes en los datos de SOPHIA Genetics. Se encontró un TE candidato en la paciente GT276 en la región intrónica de *SOS1* (chr2:39315326), sin embargo, esta variante no cumplió con los criterios de inclusión por lo cual fue descartada.

8.3 Visualización de elementos *Alu* identificados *in silico*

8.3.1 Identificación de elementos transponibles con MELT

De los 7 TEs que fueron seleccionadas por cumplir con los criterios de inclusión, 3 de ellos mostraron soporte en IGV: chr3:142231080 en *ATR* (Col-069), chr8:15978150 en *MSR1* (GT245), y chr13: 49034088- 49034119 en *RB1* (Col-2-014, LAM550043_S66, LAM550018_S51 y LAM530017_S30, LAM550001_S107). Considerando los casos que se presentaron en las corridas C2 y C3 de SOPHiA Genetics, las lecturas con SRs del TE predicho en la paciente Col-069 mantuvieron más SRs coincidentes con el chr14 (C, figura 6a); lo que también se observó en GT245 (C, figura 6b); en el caso de la paciente Col2-014, estas se mostraron principalmente en los cromosomas 12 y 9, mostradas como las lecturas grises y verdes (C, figura 6c).

a)



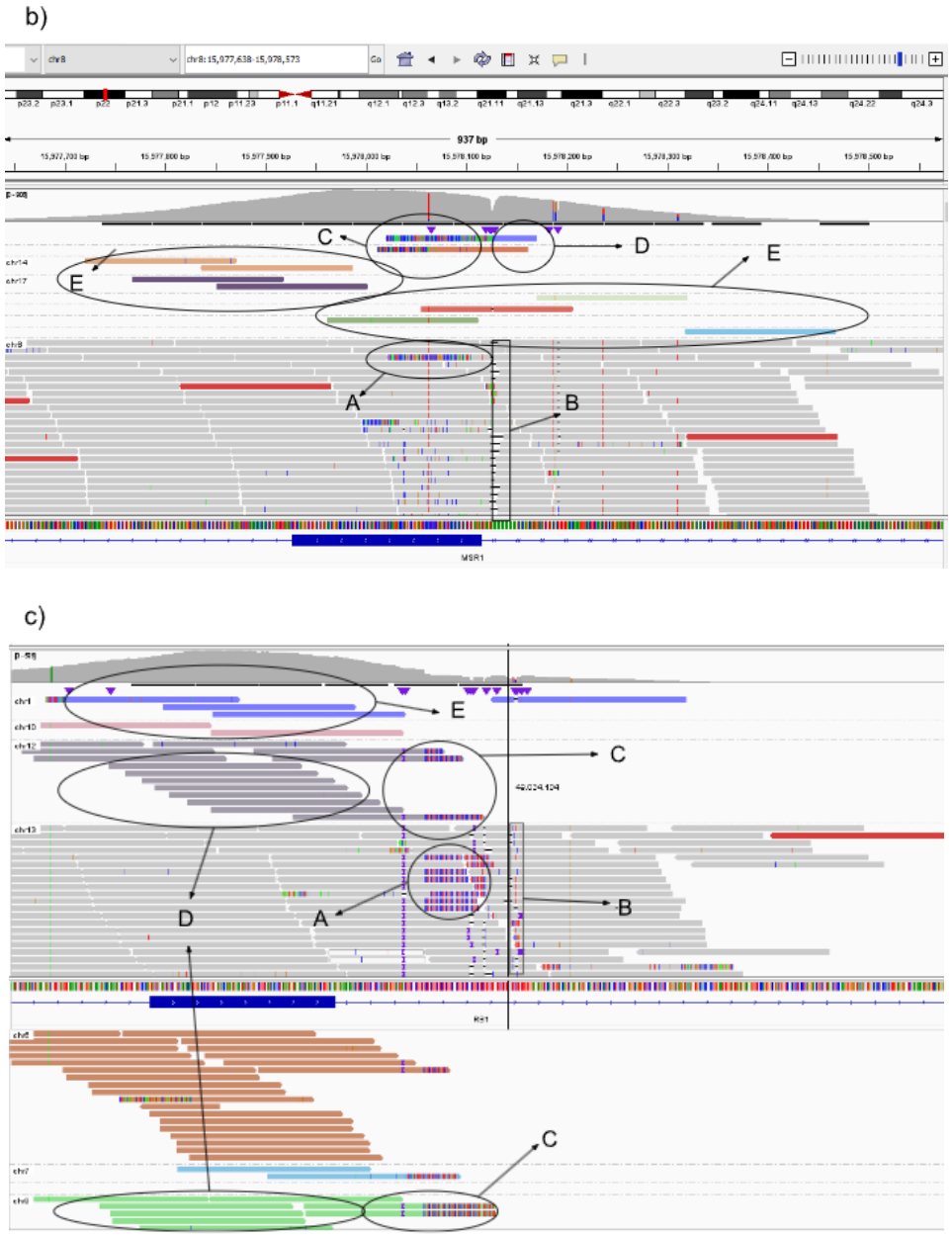


Figura 6. Visualización en IGV de inserciones predichas por MELT en a) ATR (Col-069, chr3: 142231080), b) MSR1 (GT245, chr8:15978150) y c) RB1 (Col2-014, chr13: 49034104).
A: SRs; B: TSD; C: SRs con lecturas concordantes con otro cromosoma; D: DRs de las lecturas con SRs concordantes en otro cromosoma; E: DRs en otro cromosoma.

En el caso de la inserción predicha en la paciente Col-069, el TSD consta de una región poli (A) con múltiples lecturas de SRs flanqueando ambos lados, sobre todo el izquierdo. De acuerdo con la visualización de las repeticiones obtenidas con el UCSC Table Browser, el sitio se encuentra cercano a un TE de la familia *Alu*Sq en el intrón 27.

Al visualizar las coincidencias de dichas SRs en los otros cromosomas, aparecieron en posiciones diferentes: chr14:74450467-74450546, intrón 5 del gen *ENTPD5* (correspondiente a un TE de la familia *Alu*Sz); chr14:78405390, región intergénica entre *ADCK1* y *NRXN3* (correspondiente a TEs de las familias MER5A y MER5B); chr2:13752407-

13752565, en la región intergénica entre NR_038434.1 y LINC00276, coincidente con un elemento de la familia *Alu*Sza; chr5:251106-251739, que remite al exón 12 de *SDHA*, sin TEs coincidentes en dicha posición. En los casos del cromosoma 14, la ventana que muestra la región coincidente no presenta SRs, solo DRs únicos y aislados, mientras que para el cromosoma 2 presenta una sola lectura con SRs, lo que en todos los casos brinda poco apoyo para considerarlo como un verdadero positivo (Figura 7).

Figura 7. Visualización de las lecturas con SRs en chr14 y chr2 concordantes con los del chr3:142, 231,080. La coloración que toman las lecturas resaltadas se muestra en los paneles restantes, donde se aprecian casos únicos de lecturas. El primer panel corresponde a la región en el cromosoma 2 y las dos siguientes al cromosoma 14.

Al comparar con el control entre corridas GT245 se encontró una clara diferencia en el patrón de las SRs, en este último se presentaron menos lecturas en ambos lados (Figura 8). Al tomar en cuenta el control GT298, que corresponde a una paciente con una VP sin sentido a nivel exónico en *PALB2*, también se observó una disminución de las lecturas. Tratándose de muestras pertenecientes a su misma corrida, el patrón de SRs prevaleció en todos los casos, sin embargo, no se mostraron las mismas coincidencias con el cromosoma 14.

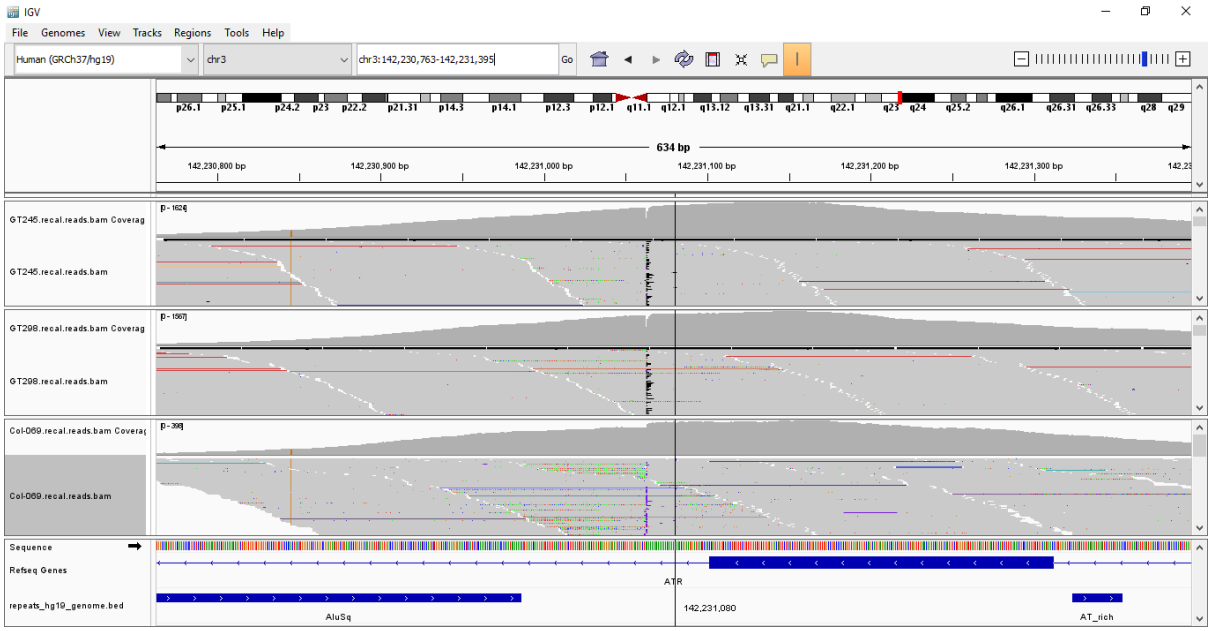


Figura 8. Visualización en IGV de la inserción predicha por MELT en el intrón 27 de ATR (chr3:142, 231, 080) de la paciente Col-069 (inferior), del control entre corridas GT245 (superior), el control GT298 (centro).

Al visualizar la posición del TE predicho en GT245 se pueden apreciar SRs flanqueando al TSD de poli (A) en el intrón 8 de dicho gen. Al compararla con la muestra Col-069 y el control GT298 se puede notar el aumento en el número de lecturas de ambas muestras (Figura 9).

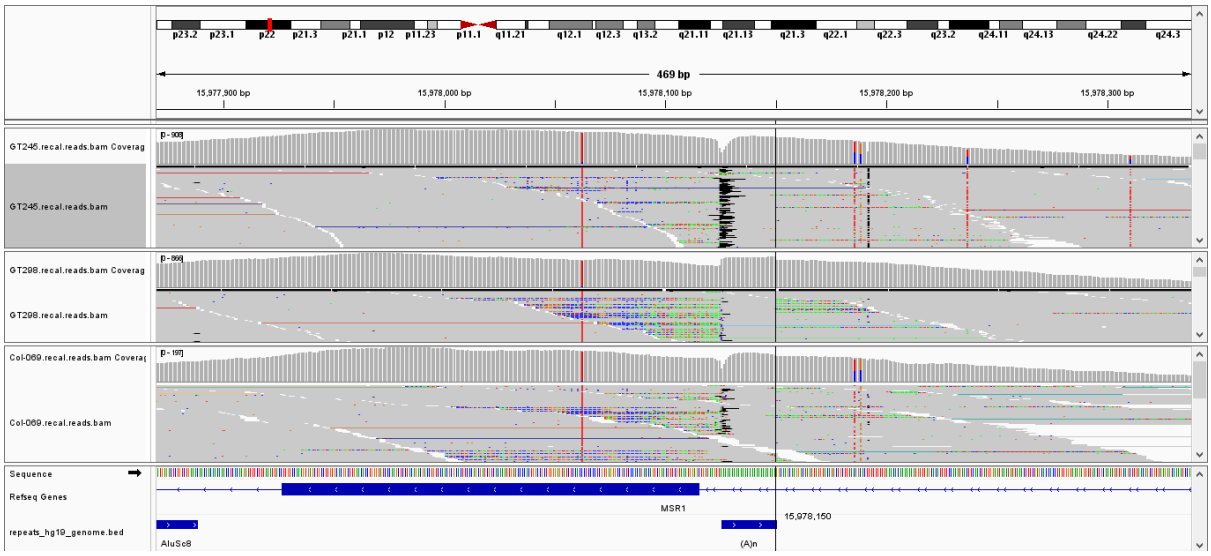


Figura 9. Visualización en IGV de la inserción predicha por MELT en el intrón 8 de MSR1 (chr8:15978150) del control entre corridas GT245 (superior), el control GT298 (con variantes patogénicas sin sentido a nivel exónico en PALB2) y la paciente Col-069 (inferior).

En el caso del TE detectado en *RB1*, el TSD que reporta el algoritmo en la muestra de la paciente Col2-014 está ligeramente alejado de las SRs, el espacio que las separa consta de repeticiones $(TTTC)_n$ (Figura 10). Al buscar las coincidencias de las SRs en otros

cromosomas se presentaron grupos de lecturas con SRs y DRs asociados a los cromosomas 12 y 9, aunque también hubo casos únicos de lecturas que fueron descartadas.

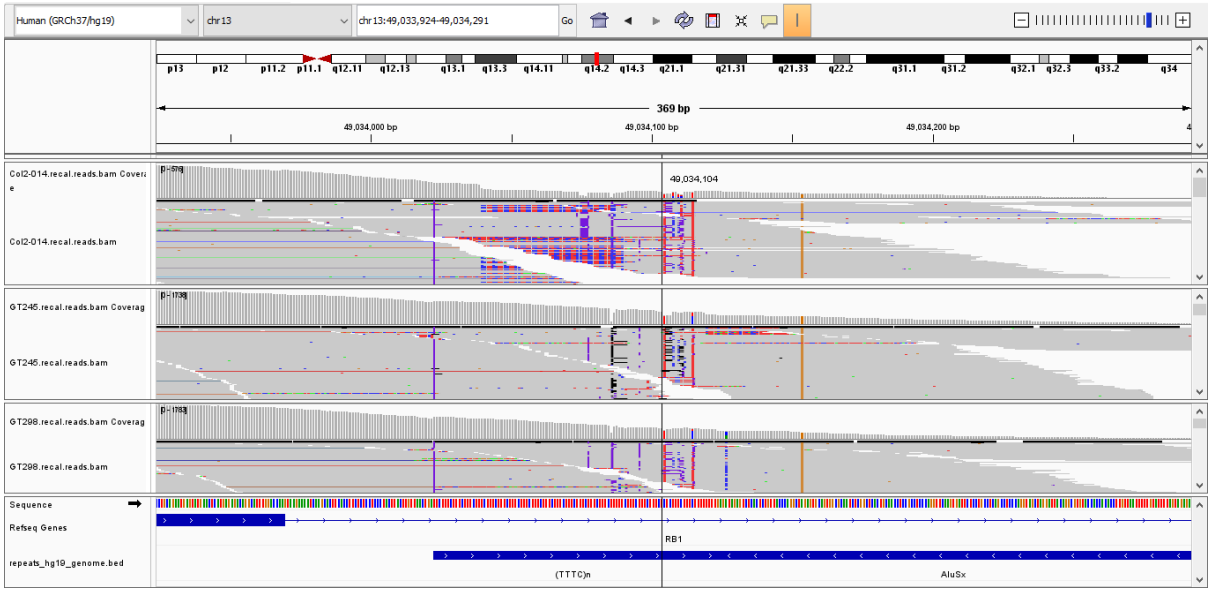


Figura 10. Visualización en IGV de la inserción predicha por MELT en el intrón 20 de RB1 (chr13: 49034104) de la paciente Col2-014 (superior), el control entre corridas GT245 (centro) y control GT298 (con variantes patogénicas sin sentido a nivel exónico en PALB2).
En la parte inferior el archivo bed obtenido del UCSC Table Browser de las repeticiones en hg19 permite apreciar que la inserción corresponde a la ubicación de una región $(TTTC)_n$ de un elemento AluSx

Al buscar la posición genómica de las coincidencias en el cromosoma 12 en el track de repeticiones obtenido del UCSC se encontró que dos de las lecturas coinciden con el intrón 1 del gen *GRIP1* en una región rica en T contigua a un elemento *Alu* de la familia *Alu*Sx (chr12:67116645-67117386), mientras que la otra coincide con el intrón 4 del gen *LGR5*, región donde se encuentran TEs de la familia *Alu*Sx y *Alu*Sg (chr12:71943872-71944609) (Figura 11).

Figura 11. Visualización de las lecturas con SRs en chr12 concordantes con los del chr13: 49034104.

La coloración que toman las lecturas resaltadas se muestra en los paneles restantes, donde se aprecian que las lecturas marcadas con verde y rojo remiten a la misma posición.

En el caso del otro grupo de lecturas con SRs concordantes en el cromosoma 9 los tres casos remitieron a la posición chr9:38,538,557-38,538,925 (región intergénica entre *FAM95C* y *ALDH1B1*), que coincide con un TE de la familia *Alu*Y, que en su extremo terminal consta de repeticiones $(GAAA)_n$ (Figura 12). En los controles GT245 y GT298, las SRs en los cromosomas mencionados no se presentaron.

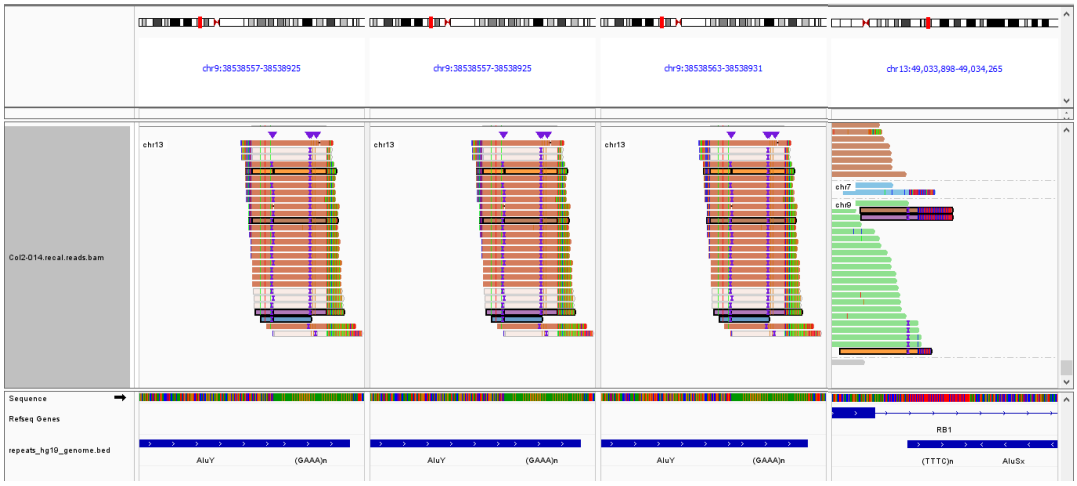


Figura 12. Visualización de las lecturas con SRs en chr9 concordantes con los del chr13: 49034104 (último panel a la derecha).

Por parte de las pacientes restantes donde también se predijo esta inserción no hubo homogeneidad en los cromosomas donde se presentaron las SRs coincidentes: LAM550001_S107 (chr 1 y 12), LAM550018_S51 (chr 8,12 y X), LAM550043_S66 (chr 8 y 20) y LAM530017_S30 (chr 8 y 16).

Respecto al TSD tenemos que solo en LAM550001_S107 predijo una secuencia que comienza justo cuando lo hacen los SRs, mientras que en el resto de los casos esta región fue predicha cercana. Las 4 TEs predichas en pacientes de la corrida CHCS4: GT33B_S22 y GT481_S19, GT121SG_S14 (2 TEs se predijeron en la misma paciente) mostraron muy pocas lecturas en la posición de inserción predicha y casos únicos de lecturas con SRs coincidentes en otros cromosomas. Ninguna de las posiciones remitió a genes considerados dentro del panel correspondiente.

8.3.2 Identificación de elementos transponibles con Retroseq

Respecto a lo reportado en RetroSeq, las muestras GT_504, GT_249, GT55 y GT7 no presentaron TEs candidatas, mientras que 11/335 inserciones predichas en 19 pacientes cumplieron con los criterios de inclusión, de los cuales 4/11 se presentaron en regiones

intergénicas asociadas a familias desconocidas que fueron agrupadas en una categoría adicional a las contempladas en la clasificación de Kojima *et al.,* 2018 que fue nombrada como UNK y 1/11 se presentó en la región chrUn_gl000220, un contig para el que el crosomosa es desconocido.

Al realizar la visualización de dichos casos no se presentó evidencia que soportara dichas inserciones (Tabla S2). Para los 6 casos restantes, una TE se presentó en *MSH6* en la paciente Col-097 (chr2:48029030) cuyo sitio de inserción fue coincidente con un elemento de la familia *Alu*Jb, sin embargo, en la región cromosómica compatible (chr11:82323007-82324563) se presentaron TEs de la familia *Alu*Sx1 y *Alu*Sz6 con solo una lectura de apoyo.

Respecto a las 3 TEs que se presentaron en *VHL*, en ISEM101 (chr3:10189026) se presentó un patrón de DRs diferente del control GT278, además mostró 2 SRs coinicidentes (chr17:73147111-73147151) que cayeron en repeticiones de la familia L1M5 y *Alu*Sx (se esperaba *Alu*Sc8). Para la paciente Col-089 (chr3:10190482) la región coincidente (chr2:32494188-32494346) solo presentó una lectura en apoyo en una repetición de la familia MLT1H2 (se esperaba *Alu*Sx3). Respecto a la paciente Col-090 (chr3:10188607) se encontraron solo DRs cercanas al sitio de inserción que remitían a una región (chr1:14,627,845-14,634,223) coincidente con un elemento de la familia MER5A (se esperaba un elemento *Alu*Sc8).

Por parte de *PMS2* se presentaron 2 TEs: La reportada en la paciente Col-072 (chr7:6036887) correspondiente a un elemento *Alu*Sx tuvo lecturas coincidentes con un elemento de la familia HSATII (chr10:42595902- 42597938). La otra TE coincidente con la familia *Alu*Jo se presentó en 2 pacientes: en Col-077 (chr7:6037045) que mostró lecturas remitiendo a una inserción de familia HERV-Hint (chr5:170,191,631-170,192,649) y en Col-061, que remitió a L1MC4a (chr10:54911896-54912914). En los casos reportados por RetroSeq solo la paciente Col-077 y Col-061 no mostraron la categoría SINE aun cuando se reportó un elemento Alu en el sitio de inserción.

8.3.3 Identificación de elementos transponibles con TEfinder

Con TEfinder, se presentaron 13/15 TEs como candidatas en 19 muestras (16 pacientes y 3 controles), sin embargo, ninguna de las inserciones predichas fue única de la paciente. El TE predicho de la familia *Alu*Sp en chr11: 11196361-11964186 (*SDHD*) fue la predicción más frecuente reportada por este algoritmo pues se presentó en 50 pacientes. Es notable que esto solo es en tanto los casos que cumplieron con los criterios de inclusión, pues de considerar la totalidad de las llamadas tendría un total de 70 apariciones (columna FT tabla S3).

En las 5 pacientes con un TE de la familia *Alu*Jr en chr17:3344377-33444054 (*CHEK2)* se presentó un patrón claro de DRs y TSD y en las 12 pacientes con una TE de la familia Tigger5 predicha en 7576394-7576723 (*TP53*) se presentó un patrón claro de SRs. En los restantes es notable una tendencia a la presencia de DRs dentro del rango que comprende las coordenadas de la inserción, destacando la aparición de familias adicionales como L3b (*FANCI)*, MIRc (*MEN1, FH*), MIRb (*SMAD4, PALB2*), L3 (*MSH2*), L2a (*ERCC2*), L1PA7 (*POT1*), X5A_LINE (*APC*) que concuerdan con las presentes en el track de repeticiones obtenido desde el UCSC Table Browser. Al comparar con el control GT298, en ninguno de los casos se presentó una coincidencia entre las familias de TEs reportadas en el sitio de inserción y en la posición de las lecturas coincidentes (Tabla S3). En tanto a las lecturas compartidas en el sitio de inserción entre las pacientes y el control se encontró que en 6/13 casos se presentó el mismo cromosoma reportado en al menos una ocasión, pero en solo 2 casos se presentó una coincidencia que remitía a la misma posición

8.4 Diseño de primers

Se diseñaron tres pares de primers para la validación de los TEs de las posiciones genómicas: i) chr13: 49034104 (*RB1*); ii) chr3: 142231080 (*ATR*) y iii) chr8: 15978150 (*MSR1*) identificados con el algoritmo MELT, descartando las predicciones restantes al no contar con suficiente evidencia durante la visualización (Tabla 4).

Considerando que el TE predicho en *RB1* fue encontrada en 5 pacientes se optó por el diseño de primers, tomando como referencia sólo una de ellas, la paciente Col2-014. Contemplando un rango de ±215 pb flanqueando el punto de inserción los programas empleados mostraron pares de primers con mejores valores en los criterios de inclusión, por lo que también fue tomado en cuenta un tamaño del producto esperado menor respecto a otros candidatos en todos los casos.

Tabla 4. Juegos de primers seleccionados para los sitios de inserción de las tres TEs predichas.

E	L	tm	gc%	any	3'	S	P
Col-069_*ATR*_chr3_ 142231080_4_215_FWD	20	59.81	50	4	2	CTGGGATTACA GGCCAACAT	737
Col-069_*ATR*_chr3_ 142231080_4_215_REV	20	59.66	55	4	0	GGCCAAGGCA GATAGATCAC	737
Col2-014_*RB1*_chr13_ 49034104 _2_215_FWD	20	60.02	50	6	1	TTGTGAACGCC TTCTGTCTG	517
Col2-014_*RB1*_chr13_ 49034104 _2_215_REV	20	60.1	50	6	2	CAACACTTTGG GAGGCCTTA	517
GT245_*MSR1*_chr8_ 15978150 _2_150_FW	21	58.23	47.62	4	2	AATCTGTGACG TGTCCTCGTA	430

GT245_MSR1_chr8_ 15978150 _2_150_REV	20	59.41	55	4	3	GGCTTGAACAC CTGGGTATC	430

E: Etiqueta del primer; L: Longitud; tm: temperatura de fusión; gc%: contenido de guanina y citosina; any: estabilidad de cualquier emparejamiento de bases del primer consigo mismo; 3': estabilidad de cualquier emparejamiento de bases del extremo 3' del primer consigo mismo; S: Secuencia del primer; P: Producto.

Para el TE predicho en *RB1,* los primers seleccionados se localizaron en el exón e intrón 20, respectivamente. El tamaño del producto del alelo silvestre (WT, por sus siglas en inglés) esperado en este caso fue de 517 pb, sin embargo, de presentarse un elemento *Alu* insertado de acuerdo con la predicción del algoritmo el tamaño del producto aumentaría hasta 597 pb (Figura 13).

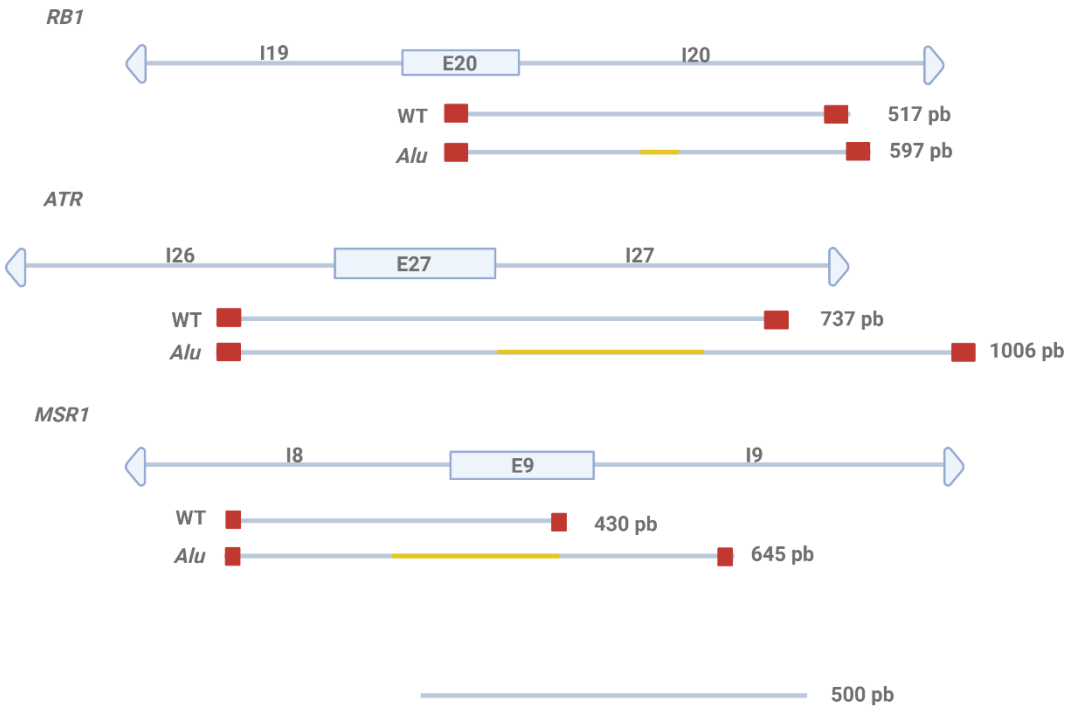


Figura 13. Tamaño esperado para los productos de PCR en RB1, ATR y MSR1 tras el fenómeno de retrotransposición.

En el predicho en *ATR*, los primers se localizaron en el intrón 26 y 27, donde en el caso de inserción se esperaría un producto de 1006 pb, mientras que para *MSR1* esto se dio en el exón 9 y el intrón 8, respectivamente, esperando un producto de 645 pb.

8.5 Validación de elementos *Alu*

8.5.1 PCR

La validación mediante PCR permitió dilucidar que en los TEs predichos para *MSR1* y *RB1* no hay diferencias en el tamaño del producto entre el individuo sano y las pacientes GT245 y Col2-014, respectivamente (Figura 14). Los tamaños de inserción predichos por MELT para

ATR, RB1 y *MSR1* son 1006 pb, 597 pb y 645 pb, respectivamente, se puede descartar la presencia de una TE en uno de los alelos, ya que en tal caso en la paciente en cuestión se apreciaría la banda correspondiente al tamaño mencionado y otra de un menor tamaño para el alelo sin la inserción.

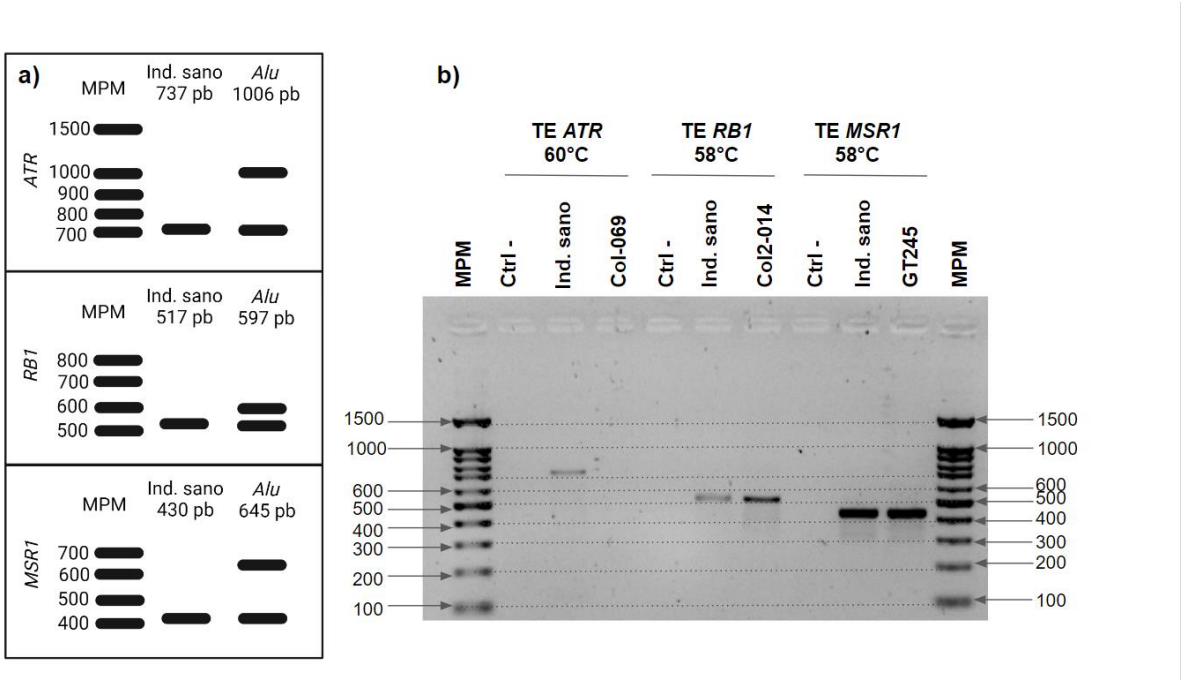


Figura 14. Validación mediante PCR realizada a 58°C para TEs predichas en RB1 y MSR1 y a 60°C en ATR a) Tamaño de bandas esperadas en caso de la transposición, b) electroforesis de los productos de PCR para ATR, RB1 y MSR1. Se incluyó un control negativo y un individuo sano empleando los respectivos primers para cada TE predicha. MPM=marcador de peso molecular de 100 pb (las flechas señalan los pares de bases correspondientes a lo largo del marcador), Ctrl - = control negativo, Ind. sano= individuo sano

Para la paciente Col-069 no se apreció una banda clara del producto en primera instancia. En una nueva estandarización se determinó que la mejor temperatura para la hibridación era 60°C empleando una concentración final para los primers de 0.2 µM. El bandeado mostró un producto de PCR con inespecíficos por debajo del producto esperado de 737 pb, pero con una banda definida correspondiente a este tamaño de amplicón (Figura 15).

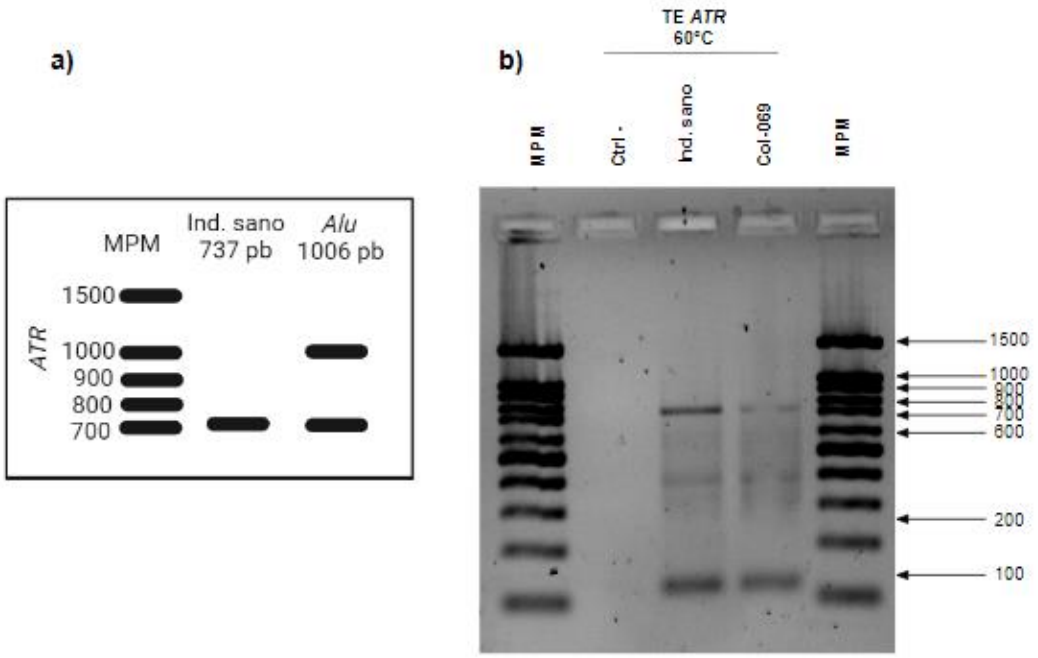


Figura 15. Validación mediante PCR realizada a 60°C para TE predicha en ATR
a) Tamaño de bandas esperadas en caso de la transposición, b) electroforesis de los productos de PCR

Con lo observado en este último gel se determinó que la muestra de la paciente no mostró una banda diferente a la del individuo sano, descartando así la presencia de una TE en la región intrónica de *ATR.*

8.5.2 Secuenciación Sanger

El TSD del TE predicho por MELT correspondiente a la paciente Col2-014 (*RB1*) no pudo ser identificado en el electroferograma debido a una disminución en la calidad de la secuenciación después de la posición nucleotídica 147 para el caso de la cadena sentido y 227 para la antisentido (a partir de donde se observó una región poli (A) hasta la posición 312). Se descartó la presencia de un corrimiento en el marco de lectura al no recuperarse la misma secuencia en la cadena complementaria. Al localizar la posición correspondiente al TSD de manera teórica no hubo evidencia de la presencia de un TE por cambios en las bases reportadas en las posiciones adyacentes.

9. DISCUSIÓN

El SHCMO es una enfermedad autosómica de dominancia incompleta con una alta incidencia y mortalidad en América Latina cuyo carácter hereditario es explicado en la mayor parte de los casos a través de VPs en genes implicados en mecanismos de reparación (Nielsen *et al.,* 2016). Buscando dilucidar la causa en aquellas pacientes donde no se presentan SVs, CNVs o SNVs se ha propuesto el estudio de mecanismos alternativos poco explorados implicados en la expresión génica, tal es el caso de epimutaciones, o bien, de otros procesos como la retrotransposición mediada por TEs (Klein *et al.,* 2023).

Al respecto, los eventos de mutagénesis insercional tanto a nivel exónico como intrónico reportados en el SHCMO ocurren en genes como *BRCA1/2*, *PALB2* y *ATM* causantes de omisiones exónicas y codones de paro prematuros, sin embargo, se trata de estudios enfocados en poblaciones europeas (Peixoto *et al.,* 2008; Yang *et al.,* 2016; Qian *et al.,* 2017; Bouras *et al.,* 2021; Eyries *et al.,* 2022), de ahí que el presente se enfoque en el análisis de muestras de pacientes de América Latina con la finalidad de evaluar la prevalencia de mutagénesis causada por TEs en genes de susceptibilidad a tumores hereditarios.

La ausencia de bibliotecas propias de cada algoritmo responde a no incurrir con faltas en la licencia de bases de datos. Repbase/GIRI ha sido ampliamente utilizada pero solo funcionó de manera gratuita hasta 2014 y a la fecha no tiene en su listado instituciones de América Latina con libre acceso. Dentro de su repositorio cuentan con archivos en formato fasta y embl para hg19 creados a partir de RepeatMasker (Smit *et al.,* 2013-2015).

Ante esto, otras bases como Dfam (Hubley *et al.,* 2016) se han desarrollado para brindar información sobre la localización y secuencia de las principales familias de TEs en los formatos mencionados, sin embargo, dentro de sus desventajas tenemos que solo contemplan esta información para hg38, lo que imposibilita trabajar con versiones anteriores del genoma humano, como en este caso. Otras opciones como dbRIP (Wang *et al.,* 2006) no están disponibles actualmente, por lo que de manera alternativa se ha hecho uso de las repeticiones caracterizadas como parte del proyecto 1KGP (1000 Genomes Project Consortium, 2015) tal como hace MELT o bien, recurrir al Genome Table Browser.

Como parte de las recomendaciones para utilizar MELT, Gardner *et al.*, 2017 consideraron sólo aquellos casos con la etiqueta PASS, sin embargo, en los candidatos identificados en este estudio aparecieron etiquetas en los siguientes filtros: hDP, filtro de divisiones discordantes que se muestra cuando la relación (SRs/DRs)*100 ≥ 55; Ic, sitios de inserción predicha que están a +/- 25 pb de una región de baja complejidad y ac0, relacionado con el

recuento de 0 alelos, esto es, casos sin un alelo genotipado o bien, que al menos una de las muestras difiere del resto al presentarse un alelo alternativo. Al cumplir sólo con los criterios de ASSESS y de SRs, los casos reportados permanecen como candidatos que cumplen parcialmente con las condiciones requeridas para incluirlos como TEs verdaderos.

Las disparidades entre los tres algoritmos se explican tras la determinación de las regiones de búsqueda de los TEs. La misma base de datos fue empleada en Retroseq y TEfinder, usando un archivo BED y uno GTF procedentes del UCSC Table Browser, respectivamente. Para MELT utiliza un mobioma de referencia de las familias *Alu*, SVA y L1, una lista de los TEs prioritarios en la fase III del 1KGP.

Las diferencias en los archivos requeridos como entrada representan una barrera al tratar de implementar múltiples algoritmos para la detección de TEs, más aún, cuando estos programas no han sido optimizados hacia el análisis del genoma humano. TEfinder fue pensado para analizar genomas pequeños, como es el caso de *Fusarium oxysporum*, *Drosophila melanogaster* y *Arabidopsis thaliana*, sin embargo, no aparece dentro de estudios enfocados en la evaluación del rendimiento y precisión entre algoritmos (Rishishwar *et al.*, 2017, Vendrell-Mir *et al.*, 2019; Kosugi *et al.*, 2019; Lee *et al.*, 2022). Dicho algoritmo mostró un mejor rendimiento cuando se especificó la localización de los TEs en el genoma humano (proporcionadas en los archivos BED y GTF), sin embargo, tras esta modificación solo fue posible detectar casos presentes tanto en el genoma de referencia como en el de las pacientes, no TEs particulares.

A pesar de que en las corridas de SOPHiA Genetics la profundidad fue de 500-600X (de acuerdo con Keane *et al.*, 2013, utilizando RetroSeq sobre datos de Illumina HiSeq para un trío CEU del 1KGP de profundidad >75x se obtuvo una sensibilidad en la detección <97%), la gran cantidad de inserciones candidatas que no cumplieron con los filtros de GQ y FL se atribuye a una combinación de factores, ya que el procesamiento se realizó de manera individual, mientras que en el artículo que presenta el algoritmo reportan que la sensibilidad se incrementa cuando múltiples individuos se agrupan.

Es destacable que la referencia que dicho algoritmo admite es un archivo Fasta/BED procedente de Repeatmasker, que a su vez recurre a instancias a las que no se tuvo acceso. Otro punto es que este algoritmo se centra en la búsquda de únicamente DRs, dejando de lado otro tipo de evidencia que brinde mejor soporte. Respecto al filtrado tenemos que los desarrolladores recomiendan dos combinaciones de criterios, lo que amplió la cantidad de TEs predichas considerablemente y por tanto la cantidad de falsos positivos.

La sensibilidad basada en PCR y la precisión en el conjunto de datos no simulados con lecturas cortas de Illumina NA12878 posicionan a MELT con los porcentajes más altos, >99% y 88.9% respectivamente, mientras que en el caso de RetroSeq estos valores bajan hasta >90% y 31.9%.

En el artículo que presenta a TEfinder se sugiere tomar como TEs candidatas aquellas donde no está presente ninguna etiqueta, sin embargo, en el presente estudio se consideró aquellas donde apareció *in_repeat*, lo que aunado a que ninguna inserción fue única para alguna paciente (mostrando por lo tanto una alta frecuencia alélica), permitió descartarlas como TEs implicadas en el síndrome.

Respecto a los TEs reportados tanto en TEfinder como RetroSeq tenemos que, aunque en RetroSeq cada inserción correspondió a una sola paciente, en TEfinder apareció en más de una. Con ello, se destaca que a pesar del bajo desempeño de RetroSeq y TEfinder respecto a MELT, la comparativa entre múltiples algoritmos es imprescindible para evitar el reporte de falsos positivos y un método que permite descartarlos antes de su visualización en IGV, con lo cual tratándose de un gran número de muestras reduce el tiempo de análisis para el usuario.

En tanto a las coincidencias presentes en MELT y RetroSeq tenemos que a excepción de la reportada en *VHL*, los TEs predichas en *RB1*, *MSR1* y *ATR* presentaron valores de SRs y PASS que permitieron incluirlos como parte de los TEs candidatas a ser validadas. Ahora bien, es destacable que en los tres casos los algoritmos predijeron la misma inserción en pacientes diferentes, lo cual indica una potencial inserción previamente caracterizada en el genoma de referencia común dentro de la población.

Las muestras que fueron analizadas empleando un panel de genes de Qiagen utlizaron hibridación por PCR, por esa razón únicamente se incluye la región exónica de los genes de interés, esto no se presenta para las muestras de SOPHiA Genetics donde al usarse hibridación por sondas se incluye parte de la región intrónica. Es en esta última junto con la región intergénica donde está presente la mayor proporción de elementos *Alu* (Chen y Yang *et al.*, 2017), por lo tanto, a pesar de que la proporción de muestras de Qiagen supera a las de SOPHiA no lo hace en TEs candidatas, lo cual se ve reflejado en que solo en una paciente de Qiagen se predijo una inserción.

En artículos donde se han reportado TEs asociados al SHCMO (Qian *et al.*, 2017, De Brakeleer *et al.*, 2020, Bouras *et al.*, 2021, Eyries *et al.*, 2022) no suelen ser especificados los parámetros con los cuales fue realizada la visualización en IGV. Un verdadero positivo

involucra elementos como un TSD definido (un espacio flanqueado por múltiples SRs en uno de sus lados y una región poli (A) en el otro) y SRs (filas que muestran un patrón de coloración de cambios de base consecutivos respecto al genoma de referencia). Al agrupar los alineamientos por cromosoma de coincidencia se puede notar que el patrón de los SRs también se presenta en otra localización cromosómica, lo cual se esperaría dado que los TEs de las familias *Alu*, L1 y SVA, al ser elementos non-LTR, siguen un mecanismo de retrotransposición de copiar y pegar.

La visualización de las tres TEs para las cuales fueron diseñados los juegos de primers no presentaron secuencias poli (A), lo cual se ha reportado previamente en casos de TEs de línea germinal de un intrón de *P2RX2* asociada a osteosarcoma, donde tanto los SRs como los DRs coinciden con un elemento *Alu* en el chr12. En el caso, el sitio donde ocurrió la inserción no muestra una región poli (A) en la visualización, lo que atribuyen a que el elemento fue insertado de manera incompleta (Wang y Liang *et al.*, 2022). Incluso en dicho artículo otra inserción *Alu* en *CDH13* mostró coincidencias de SRs en más de un cromosoma, lo que llevó a anotar la inserción a la familia *Alu* que tuvo más.

Al respecto, la coincidencia de DRs y SRs en cromosomas diferentes también se ha reportado en otros casos como por ejemplo el control positivo de c.1739_1740ins*Alu* en exomas de pacientes con CM (De Brakeleer *et al.*, 2020), por esa razón, el hecho de haber encontrado eventos de transposición en este estudio que remiten a varios cromosomas no es un criterio para descartarlos como candidatos. Por su parte, en otros casos como el estudio de Wang y Liang *et al.*, 2022 la región poli (A) se puede apreciar junto al TSD, patrón que se mantiene al visualizar este tipo de TEs en otros estudios (Bouras *et al.*, 2021; Deuitch *et al.*, 2020). En el caso del presente estudio la ausencia de esta secuencia no se atribuye a una inserción incompleta para las 3tres TEs, sino de falsos positivos, pues, aunque se presentó un TSD y SRs sólo en las pacientes Col2-014 y LAM550001_S107 se presentó una concordancia en la familia AluSx presente en el sitio de inserción y en la posición a la que las lecturas con SRs remitían.

La presencia de un patrón de SRs en la posición del TE predicho en Col-069 permaneció en el resto de las pacientes pertenecientes a la corrida C3, lo que permite descartarlo como una inserción relacionada con el padecimiento por su alta frecuencia alélica o bien, debido a un error sistemático de la tecnología en el procesamiento de ese lote de muestras. Aunado a ello, dado que los casos de lecturas coincidentes en otro cromosoma son aislados, existe poca evidencia de que se haya presentado un evento de retrotransposición *de novo*, además, la posición donde se predijo la inserción en *ATR* no se superpone con ningún elemento *Alu,* siendo el más cercano uno de la familia *Alu*Sq que además no es consistente con las familias

de SINEs coincidentes detectadas en otros cromosomas como es el caso de *Alu*Sz, MER5A, MER5B y *Alu*Sz6, las cuales al tener lecturas únicas apoyando dicho evento de retrotransposición, se descartan como los SINEs responsables de la retrotransposición.

Aunque en la paciente GT245 se mostraron los mejores niveles de evidencia para considerar la inserción como candidata esto no ocurrió para el caso del control negativo GT298 ni las pacientes de la corrida C3, donde incluso se mostró una mayor cantidad de SRs flanqueando el TSD en dicha posición. Esto implica que una gran cantidad de SRs no necesariamente determina que todas las lecturas remitan la misma región cromosómica y más aún, que todas contengan SRs que coincidan con un TE.

Aunado a ello, en esta misma paciente se presentó un decremento en la cobertura en la posición correspondiente al inicio de la TSD que no es consistente con el aumento esperado para las inserciones TE verdaderas (De Brakeleer *et al.*, 2020; Chu *et al.,* 2020). Este descenso no se observó en el control GT298, donde la cobertura mostrada en IGV no presentó una discontinuidad en dicha posición, lo cual sumado a lo anterior refuerza que esta TE predicha sea descartada como un verdadero positivo.

Respecto al TE predicho en Col2-014, se encontró que la familia del elemento *Alu*Sx del sitio de inserción fue consistente con las lecturas del chr 12, sin embargo, este no fue el caso donde se encontraron más lecturas, pues esto ocurrió en el chr 9 con la familia *Alu*Y. Esto implicaría que un elemento *Alu*Sx localizado en la región intrónica de *GRIP1* fue copiado y pegado a la región intrónica de *RB1* en la paciente Col2-014 o bien, que esto fue mediado por uno *Alu*Y, lo cual resulto problemático al diferir de la familia presente en el sitio de inserción.

Al comparar con los TEs reportados por Gardner *et al.,* 2017 disponibles en dbVar no se encontró que alguno de los elementos encontrados en este estudio se halla reportado con anterioridad, tanto en los sitios de inserción como en las regiones cromosómicas coincidentes con sus lecturas, solo se hallaron elementos cercanos que diferían en miles de bases de los predichos.

Respecto a las candidatas reportadas por RetroSeq tenemos que 10 casos ocurrieron en regiones intergénicas, lo cual coincide con la mayor proporción de eventos de retrotransposición esperados para esta región pero que brinda poco soporte a que se trate de verdaderos positivos pues al tratarse de muestras tratadas con técnicas de enriquecimiento dirigido por PCR a las regiones exónicas se espera que los reportes de TEs

predichas *in silico* ocurran en o sean cercanas a las regiones codificantes incluidas en el panel.

Adicionalmente los TEs reportadas en estas regiones coinciden con familias del archivo BED obtenido desde el UCSC Table Browser que están fuera de la clasificación de Kojima *et al.*, 2018. La categoría UNK se añadió con la finalidad de no dejar fuera las posiciones de repeticiones faltantes, sin embargo, esto resalta la necesidad de homogeneidad en los archivos disponibles para correr el algoritmo, pues en este caso el archivo BED suplió al archivo Fasta, cuando lo ideal es implementar ambos.

Entre los casos restantes que remitieron a los genes *VHL*, *PMS2*, *MSH6* y a NR_038958.1, *VHL* no forma parte del panel de genes de SOPHiA Genetics y NR_038958.1 representa una región que transcribe para lncRNA. Aunque el primero sí está presente como parte del panel de genes de Qiagen, las pacientes no fueron analizadas usando este último kit, con lo cual estos reportes corresponden a regiones para las cuales las muestras no fueron enriquecidas.

A pesar de que ni en *VHL* ni *PMS2* se presentaron coincidencias entre la familia de elemento *Alu* en el sitio de inserción y en la región a la que remitían las lecturas coincidentes, los casos de ISEM101 y Col-061, correspondientes a cada gen mencionado respectivamente tuvieron un patrón de lecturas diferente al control, pero en la región coincidente se encontraron lecturas aisladas, lo cual brinda poco apoyo para considerarlas como candidatas. Solo en estas dos pacientes se presentó un patrón de SRs y TSD, lo cual es consistente con que este método está basado solo en DRs pero resulta desventajoso al ser la única evidencia con la que es posible realizar el filtrado de variantes.

Otra desventaja de este algoritmo recae en que el sitio de inserción predicho no coincidía exactamente con la región de DRs o en el caso, de SRs y TSD reportadas, lo que lo vuelve impreciso y poco reproducible además de no proveer un estimado de la longitud de la inserción, con lo cual sería posible asignar una familia dado el TE. Problemáticas similares ya han sido reportadas con anterioridad, pues De Brakeleer *et al.*, 2020 no lograron detectar a c.1739_1740insAlu en *BRCA1* en la muestra del individuo usado como control positivo usando el algoritmo, lo que los llevó a la modificación de la segunda fase de detección del mismo.

La inserción *Alu* predicha por TEfinder en *SDHD* prevalece a lo largo del 36% de las muestras de SOPHiA Genetics, con lo cual se descarta que pueda tener un papel dentro del padecimiento. Aunado a esto, en dbVar no existen registros de un elemento *Alu* en dicho gen,

pero sí se han encontrado mutaciones de línea germinal en paragangliomas, tumores de los ganglios parasimpáticos de crecimiento lento (Pasini y Stratakis, 2009).

Por su parte, la familia *Alu*Jo actualmente se considera inactiva, pues se consiera estuvo activa hace 50 millones de años (Teugels *et al.,* 2005) y no se presentó en ninguno de los antecedentes revisados como una TE patogénica. Por otro lado, Tigger5 corresponde a una familia de transposones de DNA (no requieren de un intermediario de RNA) inactiva dentro del genoma humano (Kojima, 2018). El análisis realizado en este caso permite sentar un precedente en la extrapolación de un algortimo diseñado para analizar genomas de otras especies al genoma humano.

Aunque en otros genomas analizados esta herramienta se reporta con buenos niveles de sensibilidad, resulta limitado concluir que esto se mantendrá en el genoma humano pues no se proporciona un comparativo respecto a la profundidad o la cobertura a la cual se lleva a cabo el análisis. Un rendimiento equiparable no se se vio reflejado para la detección exónica en las muestras de las pacientes del presente estudio.

Aún con todo es necesario realizar un estudio adicional que provea información sobre el desempeño de esta herramienta sobre datos de lecturas cortas tanto para muestras individuales como tomando en cuenta un conjunto de ellas, pues se han reportado casos de herramientas como PoPoolation donde en el genoma humano hay un mejor desempeño respecto al genoma para el cual fue diseñado originalmente para la detección de MITEs (TEs de repetición invertida en miniatura) y LTRs (Vendrell-Mir *et al.,* 2019).

Por último, es importante resaltar que sólo en TEfinder el archivo de salida proporciona un rango de las coordenadas en que se predice la inserción, el número de lecturas que soportan el evento y en el criterio llamado *InsRegion*, el rango que incluye a las DRs que apoyan la inserción. Aunque en el caso de RetroSeq y MELT solo se proporciona un sitio de inserción en el archivo de salida, lo que en primer momento podría sugerir una mejora en la precisión para definir la ubicación del TE, se deja de lado la información respecto a los DRs para dar paso a otros criterios de filtrado basados en el número de SRs y en la calidad del TE predicho.

Es vital adecuar los algoritmos para conjuntar ambos tipos de evidencia en el uso de datos de lecturas cortas, no solo en el flujo de trabajo, también en la información que es proporcionada en la salida para evitar falsos positivos cuya visualización en IGV aumente el tiempo que el usuario dedica al análisis.

Los estudios donde se han reportado TEs en el SHCMO involucran principalmente dos escenarios: la inserción exónica o intrónica en un gen de interés a una CNV mediada por

estos elementos como resultado de la recombinación homóloga entre TEs en los genes *BRCA1, BRCA2, PALB2* y *ATM*, sin embargo, una de las limitantes de dichos estudios recae en que o bien, involucran la detección dentro de una paciente identificada con el síndrome mediante métodos analíticos o requieren de una amplia cohorte de pacientes a la que se aplican métodos bioinformáticos para la detección de TEs candidatos que luego serán validados.

A la fecha el estudio de Qian *et al.,* 2017 permanece como el más amplio al respecto, donde se detectaron 30 TEs patogénicos en una cohorte de más de 1.8 millones de pacientes, aún con todo, sólo una de ellas fue validada con métodos funcionales. Por su parte, es precisamente esa inserción el único caso reportado en la población de América Latina, pues los casos restantes corresponden a poblaciones brasileñas y europeas, sobre todo portuguesas, donde TEs como c.156_157ins*Alu* tienen una lata prevalencia.

Por otro lado, solo existe el reporte de una inserción L2 implicada en el síndrome (Eyries *et al.,* 2022), mientras que en los elementos SVA no son cosegregados, concluyéndose que se trata de inserciones *de novo* (Deuitch et al., 2020), es así como los elementos *Alu,* debido a la gran cantidad de copias existentes en el genoma son encontrados mayoritariamente en los reportes donde se han validado TEs con efectos como omisiones exónicas o codones de paro prematuros.

Tras ser evaluados los principales genes asociados al SCHMO, la validación de los TEs predichas realizada en el presente estudio permite descartar que el fenómeno de retrotransposición sea la causa del desarrollo del síndrome en las pacientes de la presente cohorte, lo que se atribuye al pequeño tamaño muestral comparado con otros estudios.

La validación en *ATR* no presentó diferencias entre la paciente y el individuo sano, de donde se deduce la ausencia de un TE insertado en la región intrónica. Aunque se tiene conocimiento de que este gen confiere susceptibilidad moderada al desarrollo de SHCMO no está considerado como parte de los analizados en las pruebas genéticas de la NCCN en su versión 2022. Tanto *ATR* como *ATM* funcionan como los responsables de fosforilar sustratos en respuesta al daño de DNA causado por un DSB (Macedo *et al.,* 2019).

Inserciones en *ATM,* otro gen asociado al reclutamiento de *BRCA1*, corresponden mayoritariamente a la población europea y conllevan a omisiones exónicas causantes de codones de paro prematuro y efectos en el NMD, además, los TEs exónicos reportados a la fecha asociados a este gen corresponden a la familia *Alu*Y, una de las más activas en el

genoma humano, siendo todos los casos exónicos con una sola excepción intrónica (Qian *et al.,* 2017; Klein *et al.,* 2023).

Aunque estos genes están localizados en cromosomas diferentes se tiene conocimento de que *ATR* y *ATM* son estructuralmente similares en la organización de sus dominios (Maréchа y Zou, 2013). En este caso la predicción *in silico* no es consistente con la familia mencionada, pues las coincidencias remiten a elementos pertenecientes a *Alu*S, lo cual, sumado a la ausencia de diferencias entre el control y la paciente en la validación PCR permite descartar definitivamente la presencia de un elemento insertado.

En otras enfermedades como el osteosarcoma se ha asociado la pérdida de función del *RB1* con el desarrollo del cáncer, además, mutaciones somáticas o deleciones en dicho gen han sido detectadas en CM, OC, cáncer uterino, de vejiga y de pulmón (Engeland, 2022). Aún con todo, en el contexto del SHCMO no se ha reportado a la fecha algún efecto atribuible a una TE.

Aunque el patrón de SRs fue único respecto a muestras de la misma corrida y controles en el TE predicho en *RB1*, los productos amplificados de las pacientes fueron del mismo tamaño que en el control de DNA de un individuo sano, lo que junto a la ausencia de un bandeado doble o cambios en la secuenciación Sanger forward y reverse se infiere la ausencia de una inserción *Alu* en alguno de los alelos de la paciente Col2-014, aspecto que puede ser extrapolado a las pacientes con esta inserción en las corridas restantes.

En el caso de artículos como Bouras *et al.,* 2021, este patrón está presente cuando se realiza la electroforesis de los productos de PCR que flanquean parte del exón e intrón 14 donde está comprendida la inserción *Alu*, además, cuando se realiza la secuenciación Sanger (forward) es posible visualizar la totalidad del elemento, incluyendo la región poli (A) en la región 3’ del mismo.

En otro casos como Klein *et al.,* 2023 está región solo pudo ser inferida debido a las deficiencias en el rendimiento de la polimerasa, por lo cual no se visualiza en el electroferograma, pese a esto, se llega a apreciar un cambio en la secuencia esperada y un TSD definido, aspecto en el que el presente caso está limitado, pues aunque pudo ser identificado, no presento en la secuencia contigua inmediata un patrón que mantuviera diferencias claras respecto al genoma de referencia mostrado en IGV.

Respecto a las lecturas de apoyo encontradas en el sitio de inserción, De Brakeleer *et al.* 2020 muestran un ejemplo donde hay lecturas que coinciden con la región poli (A) en el extremo 3’ del TSD del sitio de inserción de un control positivo en *BRCA1,* con ello se muestra

que no en todos los casos están presentes SRs en ambos lados del TSD tal como han reportado Wang y Liang *et al.,* 2022. No se encontró indicio de que la secuencia predicha en este estudio correspondiera a algún elemento de la familia *Alu*Y, *Alu*Sx o *Alu*Sg, lo que refuerza que ninguna de estas familias está implicada en un proceso de retrotransposición en el gen mencionado.

Finalmente, la muestra de la paciente GT245 fue reportada por Quezada-Urban *et al.,* 2018, con una variante exónica que genera un codón de paro en *PDE11A* que implica un cambio c.C919T en el DNA y un cambio p.R307X proteico. La existencia de una variante a la que pueda ser atribuido el fenotipo en la paciente vuelve improbable que un TE por sí solo pueda ser responsable de efectos asociados al síndrome. Al ser un control entre corridas se trata de una inserción que prevalece a lo largo de las muestras y por su frecuencia no podría vincularse con el desarrollo del padecimiento.

11. CONCLUSIONES

- Entre las tres herramientas bioinformáticas utilizadas para la detección de TEs *in silico*, MELT fue mejor que TEfinder y RetroSeq en términos del número de TEs que cumplieron los criterios de filtrado respecto al total predicho, tiempo de procesamiento y cantidad de requerimientos durante la instalación.

- El panel de genes de SOPHiA Genetics fue más adecuado para este tipo de análisis debido a las diferencias con el panel de Qiagen en el enriquecimiento dirigido

- El análisis de un total de 1477 muestras de pacientes con SHCMO y la validación de los TEs predichos por MELT, ninguno fue un verdadero positivo para la mutagénesis insercional causada por estos elementos lo que se atribuye a una muestra poblacional pequeña o insuficiente respecto a otros estudios y es concordante con la baja prevalencia de este tipo de inserciones en enfermedades genéticas.

- Es necesario ampliar la búsqueda en este tipo de mecanismos en la población de América Latina, debido a que no suelen ser considerados como parte de las guías de diagnóstico ni estudiados dentro de las cohortes de pacientes a las que no se ha asociado alguna mutación implicada en el padecimiento.

12. REFERENCIAS

1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature*, 2015. 526(7571): 68-74.

Ade, C., Roy-Engel, A. M., y Deininger, P. L. (2013). *Alu* elements: an intrinsic source of human genome instability. Current opinion in virology, 3(6), 639–645. DOI: https://doi.org/10.1016/j.coviro.2013.09.002

Adrion, J. R., Song, M. J., Schrider, D. R., Hahn, M. W., y Schaack, S. (2017). Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. *Genome biology and evolution*, 9(5), 1329–1340. DOI: https://doi.org/10.1093/gbe/evx050

Aref-Eshghi, E. y Li, M.M (2022). Hereditary Cancer and Cancer Predisposition Syndromes. Advances in Molecular Pathology, 5 (1): 9-27. DOI: https://doi.org/10.1016/j.yamp.2022.07.002

Ari, Ş., Arikan, M. (2016). Next-Generation Sequencing: Advantages, Disadvantages, and Future. En: Hakeem, K., Tombuloğlu, H., Tombuloğlu, G. (eds.) Plant Omics: Trends and Applications. Springer. DOI: https://doi.org/10.1007/978-3-319-31703-8_5

Al Bakir, M. y Gabra, H. (2014). The molecular genetics of hereditary and sporadic ovarian cancer: implications for the future. British medical bulletin, 112(1), 57–69. DOI: https://doi.org/10.1093/bmb/ldu034

Altieri, F., Grillo, C., Maceroni, M., y Chichiarelli, S. (2008). DNA Damage and Repair: From Molecular Mechanisms to Health Implications. Antioxidants & Redox Signaling, 10(5): 891–938. DOI: 10.1089/ars.2007.1830

Bae, J., Lee, K. W., Islam, M. N., Yim, H. S., Park, H., y Rho, M. (2018). iMGEins: detecting novel mobile genetic elements inserted in individual genomes. BMC genomics, 19 (944):1-11. DOI: https://doi.org/10.1186/s12864-018-5290-9

Beitsch, P. D., Whitworth, P. W., Hughes, K., Patel, R., Rosen, B., Compagnoni, G., Baron, P., Simmons, R., Smith, L. A., Grady, I., Kinney, M., Coomer, C., Barbosa, K., Holmes, D. R., Brown, E., Gold, L., Clark, P., Riley, L., Lyons, S., Ruiz, A., … Nussbaum, R. L. (2019). Underdiagnosis of Hereditary Breast Cancer: Are Genetic Testing Guidelines a Tool or an Obstacle?. Journal of clinical oncology, 37(6): 453–460. DOI: https://doi.org/10.1200/JCO.18.01631

Bellcross C. A. (2022). Hereditary Breast and Ovarian Cancer: An Updated Primer for OB/GYNs. Obstetrics and gynecology clinics of North America, 49(1), 117–147. https://doi-org.pbidi.unam.mx:2443/10.1016/j.ogc.2021.11.005

Bono, M., Fanale, D., Incorvaia, L., Cancelliere, D., Fiorino, A., Calò, V., Dimino, A., Filorizzo, C., Corsini, L. R., Brando, C., Madonia, G., Cucinella, A., Scalia, R., Barraco, N., Guadagni, F., Pedone, E., Badalamenti, G., Russo, A. y Bazan, V. (2021). Impact of deleterious variants in other genes beyond *BRCA1/2* detected in breast/ovarian and pancreatic cancer patients by NGS-based multi-gene panel testing: looking over the hedge. ESMO open, 6(4): 1-9. DOI: https://doi-org.pbidi.unam.mx:2443/10.1016/j.esmoop.2021.100235

Bogliolo, M., y Surrallés, J. (2015). Fanconi anemia: a model disease for studies on human genetics and advanced therapeutics. Current opinion in genetics & development, 33: 32–40. DOI: https://doi-org.pbidi.unam.mx:2443/10.1016/j.gde.2015.07.002

Bose, S.M., Sharma, S.C., Mazumdar, A., Kaushik, R. (eds.) (2022). Breast Cancer: Comprehensive Management. Springer Singapore. DOI: https://doi-org.pbidi.unam.mx:2443/10.1007/978-981-16-4546-4

Bouras, A., Leone, M., Bonadona, V., Lebrun, M., Calender, A. y Boutry-Kryza, N. (2021). Identification and Characterization of New *Alu* Element Insertion in the *BRCA1* Exon 14 Associated with Hereditary Breast and Ovarian Cancer. Genes, 12 (1736): 1-2. DOI: https://doi.org/10.3390/genes12111736

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., y Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome biology*, *19*(1), 199. DOI: https://doi.org/10.1186/s13059-018-1577-z

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., y Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6): 394–424. DOI: https://doi.org/10.3322/caac.21492

Bunz, F. (2022). The Genetic Basis of Cancer. In: Principles of Cancer Genetics. Springer, Cham. https://doi-org.pbidi.unam.mx:2443/10.1007/978-3-030-99387-0_1

Chen, L. L., y Yang, L. (2017). *ALU*ternative Regulation for Gene Expression. Trends in cell biology, 27(7): 480–490. DOI: https://doi-org.pbidi.unam.mx:2443/10.1016/j.tcb.2017.01.002

Chu, C., Zhao, B., Park, P. J., y Lee, E. A. (2020). Identification and Genotyping of Transposable Element Insertions From Genome Sequencing Data. *Current protocols in human genetics*, 107(1), e102. https://doi.org/10.1002/cphg.102

Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A., y Park, P. J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nature Communications, 12(1). DOI:10.1038/s41467-021-24041-8

Clark, D.P., Pazdernik, N.J. y McGehee, M.R. (2019). Molecular Biology. Elsevier. Third edition. DOI: https://doi.org/10.1016/C2015-0-06229-3

Cradelli, M., Marchegiani, F. y Provinciali, M. (2012). *Alu* insertion profiling: Array-based methods to detect *Alu* insertions in the human genome. Genomics, 99(6):340-346. DOI: https://doi-org.pbidi.unam.mx:2443/10.1016/j.ygeno.2012.03.005

Colonna Romano, N., y Fanti, L. (2022). Transposable Elements: Major Players in Shaping Genomic and Evolutionary Patterns. *Cells*, *11*(6): 1048. DOI: https://doi.org/10.3390/cells11061048

D'Andrea, A.D. (2008). DNA Repair Pathways and Human Cancer. En: Mendelsohn, J.P., Howley, M., Israel, M.A., Gray, J.W. y Thompson, C.V. (eds). The Molecular Basis of Cancer (3° edición). W.B. Saunders. DOI:https://doi.org/10.1016/B978-141603703-3.10004-4.

Da Costa E Silva Carvalho, S., Cury, N. M., Brotto, D. B., de Araujo, L. F., Rosa, R. C. A., Texeira, L. A., Plaça, J. R., Marques, A. A., Peronni, K. C., Ruy, P. C., Molfetta, G. A., Moriguti, J. C., Carraro, D. M., Palmero, E. I., Ashton-Prolla, P., de Faria Ferraz, V. E., y Silva, W. A., Jr (2020). Germline variants in DNA repair genes associated with hereditary breast and ovarian cancer syndrome: analysis of a 21 gene panel in the Brazilian population. BMC medical genomics, 13(1): 21. https://doi.org/10.1186/s12920-019-0652-y

Daly, M. B., Pal, T., Berry, M. P., Buys, S. S., Dickson, P., Domchek, S. M., Elkhanany, A., Friedman, S., Goggins, M., Hutton, M. L., CGC, Karlan, B. Y., Khan, S., Klein, C., Kohlmann, W., CGC, Kurian, A. W., Laronga, C., Litton, J. K., Mak, J. S., … Dwyer, M. A. (2021). Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network: JNCCN, 19(1), 77–102. https://doi.org/10.6004/jnccn.2021.0001

David, M., Mustafa, H., y Brudno, M. (2013). Detecting *Alu* insertions from high-throughput sequencing data. Nucleic acids research, 41(17), e169. DOI: https://doi.org/10.1093/nar/gkt612

Davis, M. W., y Jorgensen, E. M. (2022). ApE, A Plasmid Editor: A Freely Available DNA Manipulation and Visualization Program. *Frontiers in bioinformatics*, 2 (818619): 1-15. DOI: https://doi.org/10.3389/fbinf.2022.818619

De Brakeleer, S., De Grève, J. y Teugels, E. (2020). A systematic screen of breast cancer patients' exomes 4 for retrotransposon insertions reveals disease 5 associated genes. bioRxiv. DOI: https://doi.org/10.1101/2020.06.04.123240

Deininger, P. (2011) *Alu* elements: know the SINEs. *Genome Biology,* 12, 236. DOI: https://doi.org/10.1186/gb-2011-12-12-236

Delseny, M., Han, B., y Hsing, Y. I. (2010). High throughput DNA sequencing: The new sequencing revolution. *Plant Science*, 179 (5): 407–422. DOI:10.1016/j.plantsci.2010.07.019

Deuitch, N., Li, ST., Courtney, E., Shaw, T., Dent, R., Tan, V., Yackowski, L., Torene, R., Berkofsky-Fessler, W. y Ngeow, J. (2020). Early-onset breast cancer in a woman with a germline mobile element insertion resulting in *BRCA2* disruption: a case report. Human Genome Variation, 7 (24). DOI: https://doi.org/10.1038/s41439-020-00111-z

Disdero, E., y Filée, J. (2017). LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA*, 8 (5). DOI: https://doi.org/10.1186/s13100-017-0088-x

Eliyatkın, N., Yalçın, E., Zengel, B., Aktaş, S. y Vardar, E. (2015). Molecular Classification of Breast Carcinoma: From Traditional, Old-Fashioned Way to A New Age, and A New Way. *The journal of breast health*, 11(2), 59–66. https://doi.org/10.5152/tjbh.2015.1669

Engeland, K. (2022). Cell cycle regulation: p53-p21-RB signaling. Cell Death Differ, 29: 946–960. DOI: https://doi.org/10.1038/s41418-022-00988-z

Eyries, M., Ariste, O., Legrand, G., Basset, N., Guillerm, E., Perrier, A., Duros, C., Cohen-Haguenauer, O., de la Grange, P., y Coulet, F. (2022). Detection of a pathogenic *Alu* element insertion in PALB2 gene from targeted NGS diagnostic data. European journal of human genetics: EJHG, 30(10): 1187–1190. DOI: https://doi.org/10.1038/s41431-022-01064-3

Farrell, R.E. (2010). Chapter 18 - RT-PCR: A science and an art form. En: Farrell, R.E. (ed.) RNA Methodologies (Fourth Edition). Academic Press. DOI: https://doi.org/10.1016/B978-0-12-374727-3.00018-8.

Fanciulli, M., Petretto, E., y Aitman, T. (2010). Gene copy number variation and common human disease. Clinical Genetics, 77(3): 201–213. DOI:10.1111/j.1399-0004.2009.01342.x

Fanale, D., Fiorino, A., Incorvaia, L., Dimino, A., Filorizzo, C., Bono, M., Cancelliere, D., Calò, V., Brando, C., Corsini, L. R., Sciacchitano, R., Magrin, L., Pivetti, A., Pedone, E., Madonia, G., Cucinella, A., Badalamenti, G., Russo, A. y Bazan, V. (2021). Prevalence and Spectrum of Germline *BRCA1* and BRCA2 Variants of Uncertain Significance in Breast/Ovarian Cancer: Mysterious Signals From the Genome. Frontiers in oncology, 11, 682445. DOI: https://doi-org.pbidi.unam.mx:2443/10.3389/fonc.2021.682445

Fanale, D., Pivetti, A., Cancelliere, D., Spera, A., Bono, M., Fiorino, A., Pedone, E., Barraco, N., Brando, C., Perez, A., Guarneri, M. F., Russo, T. D. B., Vieni, S., Guarneri, G., Russo, A., & Bazan, V. (2022). *BRCA1/2* variants of unknown significance in hereditary breast and ovarian cancer (HBOC) syndrome: Looking for the hidden meaning. Critical

reviews in oncology/hematology, 172: 1-16. DOI: https://doi.org/10.1016/j.critrevonc.2022.103626

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I. y Bray, F. (2020). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Disponible en: https://gco.iarc.fr/today, consultado el 16/05/2023.

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., y Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 149:778-789. DOI: https://doi.org/10.1002/ijc.33588

Felicio, P. S., Alemar, B., Coelho, A. S., Berardinelli, G. N., Melendez, M. E., Lengert, A. V. H., …, Palmero, E. I. (2018). Screening and characterization of *BRCA2* c.156_157ins*Alu* in Brazil: results from 1380 individuals from the South and Southeast. Cancer Genetics. doi: 10.1016/j.cancergen.2018.09.001

Fiston-Lavier, A. S., Barrón, M. G., Petrov, D. A., y González, J. (2015). T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic acids research, 43(4), e22. DOI: https://doi.org/10.1093/nar/gku1250

Flores, L.L. (2019). Prevalencia de mutaciones germinales en genes de susceptibilidad de cáncer en pacientes con cáncer de mama y/o ovario y riesgo genético (Tesis de Maestría). Universidad Nacional Autónoma de México. Recuperado de https://repositorio.unam.mx/contenidos/3429782

Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E., Pittard, S., Mills, R.E. The 1000 Genomes Project Consortium y Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Research, 27:1916–1929. DOI: http://www.genome.org/cgi/doi/10.1101/gr.218032.116.

Gössling, G., Rebelatto, T.F., Villarreal-Garza, C., Ferrigno, A.S., Bretel, D., Sala, R., Giacomazzi, J., William, W.N. Jr. y Werutsky, G. (2023). Current Scenario of Clinical Cancer Research in Latin America and the Caribbean. Current Oncology, 30: 653–662. https://doi.org/10.3390/ curroncol30010050

Hanaoka, F. y Sugasawa, K. (eds.) (2016). DNA Replication, Recombination, and Repair: Molecular Mechanisms and Pathology. Springer Tokyo. DOI: https://doi-org.pbidi.unam.mx:2443/10.1007/978-4-431-55873-6

Hanahan, D. y Weinberg, A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5): 646. DOI: 10.1016/j.cell.2011.02.013.

Hancks, D. C., y Kazazian, H. H., Jr (2016). Roles for retrotransposon insertions in human disease. Mobile DNA, 7, 9. https://doi.org/10.1186/s13100-016-0065-9

Hénaff, E., Zapata, L., Casacuberta, J.M. y Ossowski, S. (2015). Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*, 16 (768) DOI: https://doi.org/10.1186/s12864-015-1975-5

Hoang, L. N. y Gilks, B. C. (2018). Hereditary Breast and Ovarian Cancer Syndrome: Moving Beyond *BRCA1* and *BRCA2*. *Advances in anatomic pathology*, 25(2): 85–95. DOI: https://doi.org/10.1097/PAP.0000000000000177

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F., y Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic acids research*, 44 (D1), D81–D89. DOI: https://doi.org/10.1093/nar/gkv1272

Jiang, C., Chen, C., Huang, Z., Liu, R., y Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. BMC bioinformatics, 16(1): 72. DOI: https://doi.org/10.1186/s12859-015-0507-2

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D y Kent WJ. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res., 32 (Database issue): D493-6. http://genome.ucsc.edu/cgi-bin/hgTables

Kashyap, D., Pal, D., Sharma, R., Garg, V. K., Goel, N., Koundal, D., Zaguia, A., Koundal, S. y Belay, A. (2022). Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures. BioMed research international, 2022: 1-11. DOI:https://doi.org/10.1155/2022/9605439

Keane, T. M., Wong, K., y Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics, 29(3): 389–390. DOI: https://doi.org/10.1093/bioinformatics/bts697

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. y Haussler, D. (2002). The human genome browser at UCSC. Genome Res, 12(6):996-1006.

Kibbe, W.A. (2007). OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Res, 35 (webserver issue). Disponible en: http://biotools.nubic.northwestern.edu/OligoCalc.html

Klein, J., Allister, A.B., Schmidt, G., Otto, A., Heinecke, K., Bax-Knoche, J., Beger, C., Becker, S., Bartels, S., Ripperger, T., Bohne, J., Dörk, T., Schlegelberger, B., Hofmann W. y Steinemann, D. (2023). A Novel *Alu* Element Insertion in *ATM* Induces Exon Skipping in Suspected HBOC Patients. Human Mutation. DOI: https://doi.org/10.1155/2023/6623515

Klein, S. P. y Anderson, S. N. (2022). The evolution and function of transposons in epigenetic regulation in response to the environment. *Current opinion in plant biology*, 69 (102277). DOI: https://doi.org/10.1016/j.pbi.2022.102277

Knudson, A. G., Meadows, A. T., Nichols, W. W. yHill, R. (1976). Chromosomal Deletion and Retinoblastoma. *New England Journal of Medicine*, 295(20): 1120–1123. DOI: 10.1056/nejm197611112952007

Kobayashi, H., Ohno, S., Sasaki, Y., y Matsuura, M. (2013). Hereditary breast and ovarian cancer susceptibility genes (review). Oncology reports, 30(3): 1019–1029. DOI: https://doi.org/10.3892/or.2013.2541

Kofler, R., Gómez-Sánchez, D., y Schlötterer, C. (2016). PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Molecular biology and evolution*, 33(10), 2759–2764. DOI: https://doi.org/10.1093/molbev/msw137

Kojima K. K. (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. Mobile DNA, 9 (2). DOI: https://doi.org/10.1186/s13100-017-0107-y

Kojima, S., Koyama, S., Ka, M., Saito, Y., Parrish, E. H., Endo, M., Takata, S., Mizukoshi, M., Hikino, K., Takeda, A., Gelinas, A. F., Heaton, S. M., Koide, R., Kamada, A. J., Noguchi, M., Hamada, M., Biobank Japan Project Consortium, Kamatani, Y., Murakawa, Y., Ishigaki, K., … Parrish, N. F. (2023). Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. *Nature genetics*, 55(6), 939–951. DOI: https://doi.org/10.1038/s41588-023-01390-2

Konkel, M. K., Walker, J. A., y Batzer, M. A. (2010). LINEs and SINEs of primate evolution. *Evolutionary anthropology*, *19*(6), 236–249. DOI: https://doi.org/10.1002/evan.20283

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology*, 20(1): 117. DOI: https://doi.org/10.1186/s13059-019-1720-5

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., …, Andrieu, N. (2017) Risks of Breast, Ovarian, and Contralateral Breast Cancer for *BRCA1* and *BRCA2* Mutation Carriers. JAMA.; 317(23):2402–2416. DOI:10.1001/jama.2017.7112

Li, H. y Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14):1754-60. DOI: 10.1093/bioinformatics/btp324

Li, Y., Salo-Mullen, E., Varghese, A., Trottier, M., Stadler, Z. K. y Zhang, L. (2020). Insertion of an *Alu*-like element in MLH1 intron 7 as a novel cause of Lynch syndrome. Molecular genetics & genomic medicine, 8(12): 716-720. DOI: https://doi-org.pbidi.unam.mx:2443/10.1002/mgg3.1523

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., 3rd, Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., Wheeler, D. A., Gibbs, R. A., Kucherlapati, R.,

Lee, C., Kharchenko, P. V., Park, P. J., y Cancer Genome Atlas Research Network (2012). *Landscape of somatic retrotransposition in human cancers*. *Science* (New York, N.Y.), 337(6097): 967–971. DOI: https://doi.org/10.1126/science.1222077

Lee, W. P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P. y Marth, G. T. (2014). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLoS ONE*, 9(3), e90581. DOI: 10.1371/journal.pone.0090581

Lee, K., Seifert, B. A., Shimelis, H., Ghosh, R., Crowley, S. B., Carter, N. J., Doonanco, K., Foreman, A. K., Ritter, D. I., Jimenez, S., Trapp, M., Offit, K., Plon, S. E. y Couch, F. J. (2019). Clinical validity assessment of genes frequently tested on hereditary breast and ovarian cancer susceptibility sequencing panels. *Genetics in medicine: official journal of the American College of Medical Genetics*, 21(7), 1497–1506. https://doi.org/10.1038/s41436-018-0361-5

Lee, H., Min, J.W., Mun, S. y Han, K. (2022). Human Retrotransposons and Effective Computational Detection Methods for Next-Generation Sequencing Data. Life, 12 (1583). https://doi.org/10.3390/life12101583

Lord, C. J., & Ashworth, A. (2012). The DNA damage response and cancer therapy. Nature, 481(7381), 287–294. https://doi.org/10.1038/nature10760

Macedo, G. S., Alemar, B., & Ashton-Prolla, P. (2019). Reviewing the characteristics of *BRCA* and *PALB2*-related cancers in the precision medicine era. Genetics and Molecular Biology, 42 (1): 215-231. DOI: 10.1590/1678-4685-gmb-2018-0104

McPherson, M.J. y Møller, S.G. (2006). PCR. Taylor y Francis Group. Second edition.

Malik, S.S y Masood, N. (eds.) (2022). Breast Cancer: From Bench to Personalized Medicine. Springer Singapore. DOI: https://doi-org.pbidi.unam.mx:2443/10.1007/978-981-19-0197-3

Maréchal, A. y Zou, L. (2013). DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harbor perspectives in biology*, 5(9), a012716. DOI: https://doi.org/10.1101/cshperspect.a012716

Mayrovitz, H.N. (ed) (2022). Breast Cancer. Brisbane (AU): Exon Publications. DOI: https://doi.org/10.36255/exon-publications-breast-cancer-etiology

Momenimovahed Z, Tiznobaik A, Taheri S. y Salehiniya H. (2019). Ovarian cancer in the world: epidemiology and risk factors. *Int J Womens Health*, 11 (287-299). doi:10.2147/IJWH.S197604

Motegi, A., Masutani, M., Yoshioka, K., y Bessho, T. (2019). Aberrations in DNA repair pathways in cancer and therapeutic significances. Seminars in Cancer Biology, 58, 29–46. doi: 10.1016/j.semcancer.2019.02.005

National Comprehensive Cancer Network (NCCN). (2022). Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic. NCCN Clinical Practice Guidelines in Oncology. Versión 2.2022.

Nakamura, S., Aoki, D. y Miki, Y. (eds.) (2021). Hereditary Breast and Ovarian Cancer. Springer Singapore. DOI: https://doi.org/10.1007/978-981-16-4521-1_1

Nielsen, F., van Overeem Hansen, T. y Sørensen, C. (2016). Hereditary breast and ovarian cancer: new genes in confined pathways. Nature Reviews Cancer, 16: 599–612. DOI: https://doi.org/10.1038/nrc.2016.72

Pasini, B. y Stratakis, C. A. (2009). SDH mutations in tumorigenesis and inherited endocrine tumours: lesson from the phaeochromocytoma-paraganglioma syndromes. Journal of internal medicine, 266(1), 19–42. https://doi.org/10.1111/j.1365-2796.2009.02111.x

Payer, L. M., y Burns, K. H. (2019). Transposable elements in human genetic disease. Nature reviews. Genetics, 20(12), 760–772. DOI: https://doi.org/10.1038/s41576-019-0165-8

Payer, L. M., Steranka, J. P., Kryatova, M. S., Grillo, G., Lupien, M., Rocha, P. P., y Burns, K. H. (2021). *Alu* insertion variants alter gene transcript levels. *Genome research*, 31(12), 2236–2248. DOI: https://doi.org/10.1101/gr.261305.120

Peixoto, A., Santos, C., Rocha, P., Pinheiro, M., Príncipe, S., Pereira, D., Rodrigues, H., Castro, F., Abreu, J., Gusmão, L., Amorim, A., y Teixeira, M. R. (2009). The c.156_157ins*Alu BRCA2* rearrangement accounts for more than one-fourth of deleterious *BRCA* mutations in northern/central Portugal. Breast cancer research and treatment, 114(1), 31–38. DOI: https://doi.org/10.1007/s10549-008-9978-4

Pfaff, A. L., Singleton, L. M., y Kõks, S. (2022). Mechanisms of disease-associated SINE-VNTR-*Alu*s. Experimental biology and medicine, 247(9): 756–764. https://doi.org/10.1177/15353702221082612

Piñeros, M., Laversanne, M., Barrios, E., Cancela, M. C., de Vries, E., Pardo, C. y Bray, F. (2022). An updated profile of the cancer burden, patterns and trends in Latin America and the Caribbean. Lancet regional health. Americas, 13: 1-13. DOI: https://doi.org/10.1016/j.lana.2022.100294

Pócza, T., Grolmusz, V. K., Papp, J., Butz, H., Patócs, A. y Bozsik, A. (2021). Germline Structural Variations in Cancer Predisposition Genes. Frontiers in genetics, 12: 1-11. DOI: https://doi.org/10.3389/fgene.2021.634217

Puurand, T., Kukuškina, V., Pajuste, F.-D., y Remm, M. (2019). *Alu*Mine: alignment-free method for the discovery of polymorphic *Alu* element insertions. Mobile DNA, 10(1). DOI: 10.1186/s13100-019-0174-3

Qian, Y., Mancini-DiNardo, D., Judkins, T., Cox, H. C., Brown, K., Elias, M., Singh, N., Daniels, C., Holladay, J., Coffee, B., Bowles, K. R. y Roa, B. B. (2017). Identification of

pathogenic retrotransposon insertions in cancer predisposition genes. Cancer genetics, 216-217:159–169. DOI: https://doi.org/10.1016/j.cancergen.2017.08.002

Quezada-Urban, R., Díaz Velásquez, C. E., Gitler, R., Rojo Castillo, M. P., Sirota Toporek, M., Figueroa Morales, A., Moreno García, O., García Esquivel, L., Torres Mejía, G., Dean, M., Delgado Enciso, I., Ochoa Díaz López, H., Rodríguez León, F., Jan, V., Garzón Barrientos, V. H., Ruiz Flores, P., Espino Silva, P. K., Haro Santa Cruz, J., Martínez Gregorio, H., Rojas Jiménez, E. A. y Vaca Paniagua, F. (2018). Comprehensive Analysis of Germline Variants in Mexican Patients with Hereditary Breast and Ovarian Cancer Susceptibility. *Cancers*, *10* (10), 361. DOI: https://doi.org/10.3390/cancers10100361

Reid, B. M., Permuth, J. B., & Sellers, T. A. (2017). Epidemiology of ovarian cancer: a review. Cancer biology & medicine, 14(1), 9–32. DOI: https://doi.org/10.20892/j.issn.2095-3941.2016.0084

Rezai, M., Kocdor, M.A. y Canturk, N.Z. (eds.) (2021). Breast Cancer Essentials: Perspectives for Surgeons. Springer, Cham. DOI: https://doi-org.pbidi.unam.mx:2443/10.1007/978-3-030-73147-2

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., y Moran, J. V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology spectrum*, 3(2): 1-63. DOI: https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014

Rishishwar, L., Mariño-Ramírez, L. y Jordan, I. K. (2016). Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, 18(6): 908-918. DOI:10.1093/bib/bbw072

Rodríguez-Martín, C., Cidre, F., Fernández-Teijeiro, A., Gómez-Mariano, G., de la Vega, L., Ramos, P., Zaballos, A., Monzón, S. y Alonso, J. (2016). Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. Journal of Human Genetics, 61(5), 463–466. doi:10.1038/jhg.2015.173

Roses, D.F. (ed.) (2005). Breast Cancer. Elsevier, New York. DOI: https://doi.org/10.1016/B978-0-443-06634-4.X5001-8

Roy-Engel A. M. (2012). LINEs, SINEs and other retroelements: do birds of a feather flock together?. Frontiers in bioscience (Landmark edition), 17(4), 1345–1361. DOI: https://doi.org/10.2741/3991

Sessa, C., Balmaña, J., Bober, S. L., Cardoso, M. J., Colombo, N., Curigliano, G., Domchek, S. M., Evans, D. G., Fischerova, D., Harbeck, N., Kuhl, C., Lemley, B., Levy-Lahad, E., Lambertini, M., Ledermann, J. A., Loibl, S., Phillips, K. A., Paluch-Shimon, S. y ESMO Guidelines Committee (2023). Risk reduction and screening of cancer in hereditary breast-ovarian cancer syndromes: ESMO Clinical Practice Guideline.

Annals of oncology: official journal of the European Society for Medical Oncology, 34(1): 33–47. DOI: https://doi.org/10.1016/j.annonc.2022.10.004

Shinawi, M y Cheung, S.W. (2008). The array CGH and its clinical applications. Drug Discovery Today, 13 (17-18): 760-770. DOI: https://doi.org/10.1016/j.drudis.2008.06.007.

Shulman L. P. (2010). Hereditary breast and ovarian cancer (HBOC): clinical features and counseling for *BRCA1* and *BRCA2*, Lynch syndrome, Cowden syndrome, and Li-Fraumeni syndrome. Obstetrics and gynecology clinics of North America, 37(1). DOI: https://doi.org/10.1016/j.ogc.2010.03.003

Silva, F. C., Lisboa, B. C., Figueiredo, M. C., Torrezan, G. T., Santos, E. M., Krepischi, A. C., Rossi, B. M., Achatz, M. I., y Carraro, D. M. (2014). Hereditary breast and ovarian cancer: assessment of point mutations and copy number variations in Brazilian patients. BMC medical genetics, 15, 55. https://doi.org/10.1186/1471-2350-15-55

Smit, A.F.A., Hubley, R. y Green, P. (2013-2015). RepeatMasker *Open-4.0*. Disponible en: http://www.repeatmasker.org

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., y Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), 10869–10874. DOI: https://doi.org/10.1073/pnas.191367098

Solassol, J., Larrieux, M., Leclerc, J., Ducros, V., Corsini, C., Chiésa, J., Pujol, P. y Rey, J. M. (2019). *Alu* element insertion in the MLH1 exon 6 coding sequence as a mutation predisposing to Lynch syndrome. Human mutation, 40(6): 716–720. DOI: https://doi-org.pbidi.unam.mx:2443/10.1002/humu.23725

Steely, C. J., Russell, K. L., Feusier, J. E., Qiao, Y., Tavtigian, S. V., Marth, G., y Jorde, L. B. (2021). Mobile element insertions and associated structural variants in longitudinal breast cancer samples. Scientific reports, 11(1): 13020. https://doi-org.pbidi.unam.mx:2443/10.1038/s41598-021-92444-0

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., and Lancet, D. 2016. The GeneCards suite: from gene data mining to disease genome sequence analyses. Current Protocols in Bioinformatics, 54:1.30.1-1.30.33. DOI: 10.1002/cpbi.5

Stuppia, L., Antonucci, I., Palka, G., y Gatta, V. (2012). Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases.

International journal of molecular sciences, 13(3), 3245–3276. https://doi.org/10.3390/ijms13033245

Stewart, C., Ralyea, C., y Lockwood, S. (2019). Ovarian Cancer: An Integrated Review. Seminars in Oncology Nursing, 35(2): 151–156. doi: 10.1016/j.soncn.2019.02.001

Sohrab, V., López-Díaz, C., Di Pietro, A., Ma, L. J. y Ayhan, D. H. (2021). TEfinder: A Bioinformatics Pipeline for Detecting New Transposable Element Insertion Events in Next-Generation Sequencing Data. Genes, 12(2), 224. DOI: https://doi.org/10.3390/genes12020224

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526 (7571): 75–81. DOI: https://doi.org/10.1038/nature15394

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P. y Zhu, H. P. (2017). Risk Factors and Preventions of Breast Cancer. International journal of biological sciences, 13(11): 1387–1397. DOI: https://doi.org/10.7150/ijbs.21635

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. y Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A cancer journal for clinicians, 71(3), 209–249. DOI: https://doi.org/10.3322/caac.21660

Tamay de Dios, L., Ibarra, C. y Velasquillo, C. (2013). Fundamentos de la reacción en cadena de la polimerasa (PCR) y de la PCR en tiempo real. Tecnología en salud, 2(2): 70-78.

Teugels, E., De Brakeleer, S., Goelen, G., Lissens, W., Sermijn, E., y De Grève, J. (2005). De novo *Alu* element insertions targeted to a sequence common to the *BRCA1* and *BRCA2* genes. Human mutation, 26(3), 284.DOI: https://doi.org/10.1002/humu.9366

Thung, D.T., de Ligt, J., Vissers, L.E. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol 15, 488 (2014). https://doi.org/10.1186/s13059-014-0488-x

Tollefsbol, T. (ed.) (2021). Translational Epigenetics, Epigenetics and Reproductive Health. Academic Press. DOI: https://doi.org/10.1016/B978-0-12-819753-0.00016-7

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics, 14(2), 178–192. https://doi.org/10.1093/bib/bbs017

Thung, D.T., de Ligt, J., Vissers, L.E. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol 15, 488 (2014). https://doi.org/10.1186/s13059-014-0488-x

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., y Jemal, A. (2015). Global cancer statistics, 2012. CA: A cancer journal for clinicians, 65(2), 87–108. DOI: https://doi-org.pbidi.unam.mx:2443/10.3322/caac.21262

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. y DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics, 43(1110): 11.10.1–11.10.33. DOI: https://doi.org/10.1002/0471250953.bi1110s43

Vendrell-Mir, P., Barteri, F. y Merenciano, M. (2019). A benchmark of transposon insertion detection tools using real data. Mobile DNA 10, 53. https://doi.org/10.1186/s13100-019-0197-9

Venkitaraman A. R. (2014). Cancer suppression by the chromosome custodians, *BRCA1* and *BRCA2*. Science, 343 (6178): 1470–1475. DOI: https://doi.org/10.1126/science.1252230

Verma, M., Kulshrestha, S. y Puri, A. (2016). Genome Sequencing. Bioinformatics: 3–33. DOI: 10.1007/978-1-4939-6622-6_1

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. y Kinzler, K.W. (2013) Cancer genome landscapes. Science, 339 (6127): 1546-58. DOI: 10.1126/science.1235122.

Voutsadakis, I.A., y Stravodimou, A. (2023). Homologous Recombination Defects and Mutations in DNA Damage Response (DDR) Genes Besides *BRCA1* and *BRCA2* as Breast Cancer Biomarkers for PARP Inhibitors and Other DDR Targeting Therapies. Anticancer research, 43(3): 967–981. DOI: https://doi-org.pbidi.unam.mx:2443/10.21873/anticanres.16241

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M. A., y Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Human mutation, 27(4): 323–329. https://doi.org/10.1002/humu.20307

Wang, H. y Chen, R. (2019). Whole-exome sequencing and whole-genome sequencing. In: Gao, X.R. (ed.). Genetics and Genomics of Eye Disease: Advancing to Precision Medicine. Elsevier. DOI: https://doi.org/10.1016/C2017-0-04270-2

Wang, C., y Liang, C. (2022). The insertion and dysregulation of transposable elements in osteosarcoma and their association with patient event-free survival. *Scientific reports*, *12*(1), 377. https://doi.org/10.1038/s41598-021-04208-5

Wodarz, D., Newell, A. C., y Komarova, N. L. (2018). Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. *Journal of the Royal Society, Interface*, 15(143), 20170967. DOI: https://doi.org/10.1098/rsif.2017.0967

Wu, J., Lee, W. P., Ward, A., Walker, J. A., Konkel, M. K., Batzer, M. A., y Marth, G. T. (2014). Tangram: a comprehensive toolbox for mobile element insertion detection. BMC genomics, 15(1), 795. https://doi.org/10.1186/1471-2164-15-795

Yang, C., Arnold, A. G., Trottier, M., Sonoda, Y., Abu-Rustum, N. R., Zivanovic, O., Robson, M. E., Stadler, Z. K., Walsh, M. F., Hyman, D. M., Offit, K., y Zhang, L. (2016). Characterization of a novel germline PALB2 duplication in a hereditary breast and ovarian cancer family. Breast cancer research and treatment, 160(3): 447–456. DOI: https://doi.org/10.1007/s10549-016-4021-7

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. y Madden, T. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 13:134.

Yoshida, R. (2021). Hereditary breast and ovarian cancer (HBOC): review of its molecular characteristics, screening, treatment, and prognosis. Breast Cancer, 28:1167–1180. DOI: https://doi.org/10.1007/s12282-020-01148-2

Zhuang, J., Wang, J., Theurkauf, W., y Weng, Z. (2014). TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Research*, 42(11): 6826–6838. DOI:10.1093/nar/gku323

Zhou, W., Emery, S.B., Flasch, D.A., Wang, Y., Kwan, K.Y., Kidd, J.M., Moran y J.V., Mills, R.E. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Research*, 48 (3):1146–1163. DOI: https://doi.org/10.1093/nar/gkz1173

13. ANEXOS

Tabla S1. Posiciones cromosómicas (hg19) de los TEs predichos coincidentes entre MELT, TEfinder y RetroSeq contemplando 114 muestras y 24 controles

RetroSeq	MELT	Gen	Observaciones sobre la visualización en IGV
chr13:49033990 (Col-075)	chr13:49034104 (n=7)	*RB1*	Patrón de SRs flanqueando a un TSD presente en todos los casos reportados por MELT, pero ausente en Col-075.
chr3:142231003-142231057 (Col2-003, Col-088, Col2-010)	chr3:142231080 (Col-069)	*ATM*	Patrón de SRs flanqueando a un TSD consistente en todas las muestras.
chr8:15978117-15978165 (Col2-012, Col-054)	chr8:15978150 (GT245)	*MSR1*	Patrón de SRs flanqueando a un TSD consistente en todas las muestras.
chr3:10188563 (Col-085)	chr3:10188523 (GT479_S18)	*VHL*	Patrón de SRs y DRs ausente cercano al sitio de inserción predicho en mabas muestras.
RetroSeq[b]	TEfinder	Gen	Observaciones sobre la visualización en IGV
chr11:111963682 (Col-072 [a]) chr11:111963806-111963869 (Col-091, Col-058, Col-086, Col-090 [a])	chr11:111963610-111964078 n=70	*SDHD*	Patrón de SRs flanqueando a un TSD bien definido en los casos de RetroSeq y en los controles intracorrida GT48, GT84 (parte de los casos de TEfinder), GT272 y GT298 (controles).
chr1: 241667138 (Col-095)	chr1:241667139-241667447 n=9	*FH*	No hay SRs flanqueando a un TSD en Col-095, sin embargo, tanto en esta muestra como en GT48 y GT276 (parte de los casos de TEfinder) se presentan múltiples DRs en ambos lados del punto de inserción predicho por RetroSeq.
chr22: 29126275-29126324 (Col-061 [a], Col-065, Col-075, Col-104)	chr22: 29126286-29126620 n=12	*CHEK2*	Hay un patrón de SRs del lado izquierdo cercano al sitio de inserción predicho por RetroSeq en los cuatro casos que prevalece en los controles GT291 y GT44 (parte de los casos de TEfinder).
chr17:33443869 (Col-086)	chr17: 33443772-33444051 n=19	NR_037714.1, *RAD51L3*	No hay SRs flanqueando a un TSD en Col-086, tanto en esta muestra como en los controles inter corrida GT48 y GT245 (parte de los casos de TEfinder) se presentan múltiples DRs en ambos lados del punto de inserción predicho por RetroSeq.
chr15:89848454 (Col-053)	chr15: 89848441-89848705 n=9	*FANCI*	No hay SRs flanqueando a un TSD en Col-053. En el control inter corrida BdeN17 (parte de los casos de TEfinder) también se presentan múltiples DRs en ambos lados del punto de inserción predicho por RetroSeq y algunos SRs dispersos.

Para MELT y RetroSeq se contempló un rango ±100 pb respecto a la posición predicha para considerarla como una sola mutación, en estos algortimos los rangos incluyen las posiciones detectadas cuando hay al menos dos muestras. Dado que TEfinder provee un rango para el TE predicho, los rangos son superiores a ±100 pb. Cuando no se especifica el nombre de la muestra entre paréntesis se incluye la frecuencia de apariciones de dicha TEs.[a] Casos donde el TE predicho apareció en la misma muestra en los dos algoritmos. [b] Aquellos casos donde se reportan dos inserciones es debido a que las posiciones en un rango ±100 pb fueron considerados como un solo TE predicho en RetroSeq y MELT, pero no en TEfinder, donde el algoritmo no reportó una sola inserción sino un rango que comprende dicho sitio. Cuando en RetroSeq y MELT se presenta un rango es para contemplar las posiciones predichas en más de una muestra.

Tabla S2. 11 TEs predichas por RetroSeq en 18 pacientes que cumplieron con los criterios de inclusión

Muestra	chr	Posición	INFO	GT	GQ	FL	SP	Gen	Información sobre visualización en IGV
Col2-010	chr1	121484972	UNK	0/1	36	6	36	NE	Patrón de DRs consistente con el control GT298.
Col-097	chr2	48029030	SINE	0/1	27	7	0	*MSH6*	Se presenta un claro patrón de SRs y TSD compartido con el control GT298 coincidente con un elemento de la familia *Alu*Jb
Col-090	chr3	10188607	SINE	0/1	20	7	21	*VHL*	No hay un patrón de SRs consistente en paciente y control GT298. Hay DRs cercanos, pero no son cercanos al sitio de inserción predicho.
ISEM101	chr3	10189026	SINE	0/1	85	7	51	*VHL*	En ISEM101 hay un patrón de DRs diferente del de control GT278 y Col-090 con SRs coincidente con cromosoma 17.
Col-089	chr3	10190482	SINE	0/1	22	8	203	*VHL*	Mismo patrón de SRs en GT298; SR coincidente en chr2 único.
SG063	chr3	11911991	SINE	0/1	343	7	0	NE	Patrón de varias lecturas con SRs que comparte con control entre corridas GT245, pero no con GT298 ni GT272.
Col-077[a]	chr7	6037045	LINE	0/1	25	8	395	*PMS2*	Patrón de SRs y TSD definido que también aparece en control GT272.
Col-061[a]	chr7	6037124	LINE	0/1	32	6	177	*PMS2*	El SR coincidente con chr10 es único.
Col-072	chr7	6036887	SINE	0/1	23	8	323	*PMS2*	En control GT298 la mayor cantidad de coincidencias ocurre en chr17, pero éstas no caen en sitio con repeticiones.
SIC DNA	chr19	27732162	UNK	0/1	35	6	25	NE	Sitio con patrón de DRs compartido con control GT272.
ISEM93[b]	chr19	27732444	UNK	1/1	44	6	2	NE	No hay un patrón claro de SRs o DRs cercano al sitio de inserción predicho.
GT_505[b]	chr19	27732448	UNK	1/1	40	6	6	NE	Sitio con patrón de DRs compartido con control GT272 cercano al sitio de inserción predicho.
GT_503[b]	chr19	27732460	UNK	0/1	46	6	4	NE	Sitio con patrón de DRs compartido con control GT272 cercano al sitio de inserción predicho.
GT84[b]	chr19	27732467	UNK	0/1	30	6	3	NE	No hay coinciencias de SRs en otros cromosomas, solo hay DRs compartidos con chr1, pero que aparece también en control.
GT_440[b]	chr19	27732495	UNK	1/1	38	6	6	NE	No hay un patrón claro de SRs o DRs cercano al sitio de inserción predicho.

GT_506 [b]	chr19	27732501	UNK	0/1	34	6	14	NE	No hay un patrón claro de SRs o DRs cercano al sitio de inserción predicho.
GT_416 [b]	chr19	27732444	UNK	1/1	26	2	14	NE	No hay un patrón claro de SRs o DRs cercano al sitio de inserción predicho.
GT_414	chrUn_gl000220	119536	SINE	0/1	33	8	10	NR_038958.1	No hay SRs coincidentes con otro cromosoma, ni un patrón claro de SRs y TSD en GT_414 ni en GT298.

[a]Estas pacientes mostraron la misma inserción predicha en el cromosoma 7 en el gen PMS2. [b]Estas pacientes mostraron la misma inserción predicha en el cromosoma 19 en una regipon cromosómica indeterminada.

Tabla S3. 13 TEs predichas por TEfinder que cumplieron con los criterios de inclusión

Muestra	F	FT	chr	Ins		Fam	Gen	DRs	FR, RR	InsRegion	Información sobre visualización en IGV
GT_474, GT44, GT_409, ISEM109, GT_476, GT66, GT_414, GT_415, GT_511, ISEM110, ISEM93, GT_506, GT_509, GT_418, GT_423, Col-069, GT_408, Col-087, ISEM100, GT_516, ISEM114, Col-101, GT278, Col-104, GT_411, GT_420, ISEM108, GT84, Col2-008, Col2-004, Col-073, GT_416, ISEM111, GT48, GT_407, ISEM97, ISM 20.1, GT_417, Col2-009, Col-053, Col-077, Col-072, Col-056, EX6B, Col2-007, Col-090, ISEM115, Col-067, Col-094	50/138	70/138	chr11	11196361	11964186	*AluSp*	*SDHD*	123	61,62	111963606 111964234	Patrón de SRs en ambos lados del TSD en múltiples lecturas prevalente en control GT298.
HP6	1/138	2/138	chr5	112177221	112177479	X5A_LINE	*APC*	11	6,5	112177156 112177482	SRs dispersas, solo se presentan DRs en la región.

GT_476, GT_420, GT_406, ISEM93, Col-052	5/138	15/138	chr7	124475181 124475456	L1PA7	*POT1*	19	10,9	124475128 124475459	Se muestran DRs dispersos y una posible inserción que se comparte entre las muestras y el control GT298.
GT_504, GT_516, GT_423	3/138	6/138	chr16	23632587 23632756	MIRb	*PALB2*	15	6,9	23632477 23632789	SRs dispersas, solo se presentan DRs en la región.
GT48, GT291, GT_474	3/138	9/138	chr1	241667143 241667468	MIRc	*FH*	37	20,17	241667131 241667505	DRs circundantes a una deleción prevalente en muestras y controles.
GT_415, GT_516, GT_509, ISEM94	4/138	12/138	chr22	29126286 29126623	*Alu*Jr	*CHEK2*	81	42,39	29126277 29126660	SRs en varias lecturas del lado izquierdo del TSD, mientras que en el derecho son muy pocos, por su parte, existen DRs circundantes.

GT_420, GT_408, GT_414, GT_509, GT_423, GT276	6/138	19/138	chr17	3344377 33444054	MIRb	NR_037714.1	21	12,9	33443682 33444073	No hay un patrón de SRs claro, solo DRs circundantes.
GT_505, ISEM97, GT_406, ISEM112, ISEM103, GT66, GT_506, SG063, GT_420, GT_509, ISEM93, Col-092, GT_511, GT_423	14/138	25/138	chr19	45861996 45862277	L2a	*ERCC2*	22	13,9	45861967 45862304	SRs no están alrededor de un TSD, se notan distintos DRs tanto en muestras como en GT298.
GT_420, SG063, GT_409	3/138	5/138	chr2	47637163 47637352	L3	*MSH2*	15	9,6	47637107 47637394	Sólo DRs presentes en la región.
GT_509, ISEM99, ISEM102, ISEM110, GT_415	5/138	10/138	chr18	48574955 48575127	MIRb	*SMAD4*	14	7,7	48574807 48575158	SRs y DRs dispersos en la región.
GT_410, ISEM114, GT_423, GT_474, ISEM110, ISEM104, GT278, ISEM101, Col-062, SG063, GT420, GT35	12/138	20/138	chr11	64574679 64575140	MIRc	*MEN1*	57	33,24	64574675 64575173	DRs presentes a lo largo de la región, SRs dispersas.

GT_516, GT_473, GT48, ISEM111, ISEM113, IMSS04_06, GT298, ISEM100, ISEM109, ISEM105, Col-064, Col-075	12/138	35/138	chr17	7576394 7576723	Tigger5	*TP53*	32	16,16	7576389 7576733	Patrón claro de SRs en ambos lados de un TSD con múltiples lecturas del lado izquierdo.
ISEM101, GT_420, ISEM109, ISEM97	4/138	9/138	chr15	89848471 89848696	L3b	*FANCI*	14	8,6	89848447 89848714	SRs dispersos y DRs circundantes en la región

F=Frecuencia de la inserción predicha en muestras con filtro in_repeat; FR= Frecuencia de la inserción predicha respecto al total de muestras; Ins= Posición de inicio y final de la inserción; InsRegion= posición de inicio y final de la región de inserción; DRs= Total de lecturas discordantes, FR= Lecturas sentido, RR= Lecturas antisentido. Se incluyen sólo los casos que presentaron el filtro in_repeat. DRs, FR y RR responden a la primera muestra presentada en la columna correspondiente.