UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN

Desarrollo de un pipeline para el análisis de las firmas mutacionales en muestras de pacientes con anemia de Fanconi.

TESIS

QUE PARA OBTENER EL TÍTULO DE:

Licenciada en Bioquímica diagnóstica

PRESENTA

Enya Enara Martínez Torres

Asesor: Dra. Sara Frias Vázquez
Asesor externo: M. en C. Ulises Ehatl Juárez Figueroa

Cuautitlán Izcalli, Estado de México, 2023

VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
SECRETARÍA GENERAL
DEPARTAMENTO DE TITULACIÓN

ASUNTO: VOTO APROBATORIO

DR. DAVID QUINTANAR GUERRERO
DIRECTOR DE LA FES CUAUTITLAN
PRESENTE

ATN: DRA. MARÍA DEL CARMEN VALDERRAMA BRAVO
Jefa del Departamento de Titulación
de la FES Cuautitlán.

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de tesis y examen profesional.**

Desarrollo de un pipeline para el análisis de las firmas mutacionales en muestras de pacientes con anemia de Fanconi.

Que presenta la pasante: **Enya Enara Martínez Torres**
Con número de cuenta: **417003609** para obtener el título de: **Bioquímica Diagnóstica**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO**.

ATENTAMENTE
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cuautitlán Izcalli, Méx. a 14 de noviembre de 2022.

PROFESORES QUE INTEGRAN EL JURADO

	NOMBRE	FIRMA
PRESIDENTE	Dra. Sara Frías Vázquez	
VOCAL	Dra. Dolores Molina Jasso	
SECRETARIO	Q.F.B. Alejandro Gutiérrez García	
1er. SUPLENTE	M. en C. María Llasbeth Hernández Calderón	
2do. SUPLENTE	LBD. Larisa Andrea González Salcedo	

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional

MCVB/cga*

Agradecimiento

El presente trabajo fue realizado con el apoyo y beca para realizar mi tesis, del proyecto UNAM PAPIIT IN205120 titulado: “Búsqueda de alteraciones genómicas y epigenéticas relacionadas con envejecimiento celular en pacientes con anemia de Fanconi” bajo la dirección de la Dra. Sara Frías Vazquez.

Este proyecto fue apoyado por Recursos Fiscales INP, proyecto INP2020/012.

Agradecimientos a título personal

A la Universidad Nacional Autónoma de México, la máxima casa de estudios, que por cuatro años fue mi casa también y me acobijó con una segunda familia.

A mis asesores el M. en C. Ulises Ehatl Figueroa y la Dra. Sara Frías Vazquez, por su apoyo, contribuciones, enseñanza y su paciencia. Gracias por involucrarme en este mundo de conocimiento.

A mis sinodales, quiénes fueron también mis profesores en el transcurso de la licenciatura, personas de excelencia y docentes llenos de vocación. Gracias porque con cada enseñanza reafirmaron mi pasión por esta carrera.

A todos mis profesores y personas que formaron parte de mi vida académica y que también influyeron en mi desarrollo personal.

A todos, siempre, muchas gracias.

Dedicatorias

A mis papás Rosa y Sergio, quienes sabían que mi nombre era Enya desde mucho antes de nacer, sin ustedes esta historia no podría ser contada.
A mi tío Paco, y mi Yaya, que le ayudaron mucho a esa niña preguntona, y que siempre tienen amor que dar a su versión ahora adulta.
A ustedes cuatro, este logro también es suyo, porque me han dado todo lo que han podido y mucho MÁS.

A mis hermanos Yashua y Samara, que siempre me motivan a ser mejor y me levantan el ánimo con su amor, su paciencia y su sentido del humor. Gracias por el apoyo eterno, ser la hermana mayor ha sido más fácil con ustedes como mis hermanos.

A mis amigos, por ser parte de este camino con curvas y baches, por nunca dejar que la distancia signifique lejanía. Gracias por ayudarme a ser más abierta con el mundo y mis emociones, ustedes sacan mi lado más optimista.

A mi compañero de vida y mejor amigo, mein Schatz, my Cornerstone, hemos cursado por momentos difíciles que se hicieron fáciles, y el mérito no es solo mío. Gracias por ser parte de la aventura.

A todos los amo, infinito y para siempre.

Índice

Índice de figuras

Índice de tablas

Abreviaturas

Abreviatura	Significado
A	Adenina
AC	Aberraciones cromosómicas
ADN	Ácido desoxirribonucleico
AF	Anemia de Fanconi
AE	ArrayExpress
ARN	Ácido ribonucleico
BAM	Mapa de alineación binaria
BD	Base de datos
BWA	Burrows-Wheeler aligner
BWT	Burrows-Wheeler transform
C	Citosina
ChIP	Inmunoprecipitación de cromatina
COSMIC	Catálogo de mutaciones somáticas en cáncer
DBS	Sustitución doble de base
EBI	European Bioinformatics Institute
FAAP	Proteínas asociadas a Anemia de Fanconi
FA/BRCA	Vía de la anemia de Fanconi
G	Guanina
GEO	Gene Expression Omnibus
GG-NER	Reparación de escisión de nucleótidos global
HR	Recombinación Homóloga
HTS	Secuenciación de alto rendimiento
ICL	Inter strand crosslinks/ Enlaces covalentes cruzados
INDEL	Inserciones o deleciones de uno o más nucleótidos
LMA	Leucemia mieloide aguda
MDA	Malondialdehído
MMR	Reparación mismatch/ reparación por apareamiento erróneo
NCBI	National Center for Biotechnology Information
NER	Reparación de escisión de nucleótidos
NGS	Secuenciación de nueva generación
NNMF	Factorización matricial no negativa
PCR	Reacción en cadena de la polimerasa
SAM	Mapa de alineación de secuencias
SBS	Sustitución de un solo nucleótido
SNP	Polimorfismo de nucleótido único
SNV	Variante de nucleótido único
SRA	Sequence Read Archive
SSE	Suma de los cuadrados del error
T	Timina
TC-NER	Reparación de escisión de nucleótidos acoplada a la transcripción
TLS	Síntesis de translesión

1.- Introducción

1.1.- Anemia de Fanconi

La anemia de Fanconi (AF) es un síndrome de inestabilidad cromosómica, que presenta una herencia autosómica recesiva en 20 de los 22 genes involucrados (gen *FANCB* herencia ligada al cromosoma X y gen *FANCR* herencia autosómica dominante), es una enfermedad rara con una incidencia de 1-5 por millón de nacidos vivos y una esperanza de vida promedio de 20 años, aunque algunos pacientes son diagnosticados hasta la tercera década de vida o incluso después (Auerbach, 2015). Actualmente, se han identificado mutaciones de línea germinal en 22 genes específicos (*FANCA, FANCB, FANCC, FANCD1/BRCA2, FANCD2, FANCE, FANCF, FANCG/XRCC9, FANCI, FANCJ/BRIP1, FANCL, FANCM, FANCN/PALB2, FANCO/RAD51C, FANCP/SLX4, FANCQ/ERCC4, FANCR/RAD51, FANCS/BRCA1, FANCT/UBE2T, FANCU/XRCC2, FANCV/REV7 y FANCW/RFWD3*), cada uno de los cuales representa un subgrupo de anemia de Fanconi que a nivel clínico presentan heterogeneidad fenotípica (Río *et al.*, 2018).

En lo que concierne a la heterogeneidad fenotípica, existen características clínicas con una mayor prevalencia, siendo las más frecuentes la falla medular, las alteraciones del desarrollo físico (malformaciones congénitas) y la predisposición al desarrollo de tumores sólidos (García-de Teresa et al., 2016). Sin embargo, se ha podido determinar otras alteraciones en los pacientes con AF con una menor prevalencia, pero importantes para el diagnóstico, estas se encuentran agrupadas en dos acrónimos. El primero es VACTERL-H (anormalidades vertebrales, atresia anal, enfermedades cardíacas, fístula traqueoesofágica, atresia esofágica, malformaciones renales, extremidades superiores e hidrocefalia); se ha reportado en la literatura que entre el 5 y el 30% de los pacientes con AF presentan al menos tres de las ocho anomalías congénitas descritas como VACTERL-H. Recientemente, se identificaron otras anomalías frecuentes en AF denominadas con el acrónimo PHENOS (pigmentación de la piel, cabeza pequeña, ojos pequeños, sistema nervioso, otología, estatura corta) que ha sido complemento de VACTERL-H para el diagnóstico (Alter & Giri, 2016; Faivre *et al*., 2000, 2005) y se ha observado al menos cuatro de seis de estas características en 5% de pacientes con AF.

La evaluación de estas características clínicas ha ayudado a poder correlacionar que los pacientes con pérdida completa de una proteína FANC o estados hemicigotos o heterocigotos de genes *FANCB, FANCD2* respectivamente, se asocian a un fenotipo grave, con mayor frecuencia de asociaciones con anomalías congénitas, mientras que los pacientes con una proteína alterada presentan un fenotipo menos agresivo, con un inicio tardío de la anemia aplásica, mayor supervivencia y una menor prevalencia de complicaciones (Faivre *et al*., 2000; Fiesco-Roa *et al*., 2019).

Las principales complicaciones de la AF son la anemia aplásica; el síndrome mielodisplásico y la leucemia mieloide aguda con una probabilidad 6000 y 700 veces mayor, respectivamente, de desarrollarse en pacientes AF en comparación con la población general; y tumores sólidos específicos (Bhandari *et al*., 2021). La AF está asociada al desarrollo de cáncer debido a los defectos en la reparación del ADN que presentan, aumentando la susceptibilidad de las células somáticas a mutágenos y, en consecuencia, una mayor tasa de mutaciones; defectos en control en el ciclo celular, lo que conducen a una proliferación celular anormal; inestabilidad cromosómica, que se considera como promotor de la evolución del cáncer; y aumento de citoquinas proinflamatorias que podría resultar en daño adicional del ADN y la muerte celular (Wegman-Ostrosky & Savage, 2017).

Por lo anteriormente mencionado, el diagnóstico de AF basado en las manifestaciones clínicas es complejo y requiere de pruebas de diagnóstico como la evaluación de rupturas cromosómicas en linfocitos-T de individuos sanos y pacientes AF (Frohnmayer Lynn *et al*., 2020), para confirmarlo. Se trata de una prueba de aberraciones cromosómicas (AC), en la cual las células de los pacientes son expuestos a agentes alquilantes bifuncionales (Mitomicina C o Diepoxibutano) que generan la formación de enlaces covalentes cruzados o ICLs (*Inter strand crosslinks*, por sus siglas en inglés), que en células de individuos sanos pueden ser reparados por la vía FA/BRCA pero en los pacientes con AF que tienen deficiente esta vía, no pueden reparar este tipo de lesiones y generan daño a nivel cromosómico que se evalúa por la presencia de las AC (García-de Teresa & del Castillo, 2014; Molina et al., 2022). Esta deficiencia en la vía FA/BRCA confiere inestabilidad cromosómica a la célula, de manera que en comparación con las células control, en células de pacientes AF en metafase se observa un mayor número de rupturas cromosómicas y formación de figuras radiales, que son producto de la unión entre dos o más rupturas de doble cadena de cromátidas no hermanas (Fig. 1, señaladas con flechas) (Ben Haj Ali *et al*., 2019).

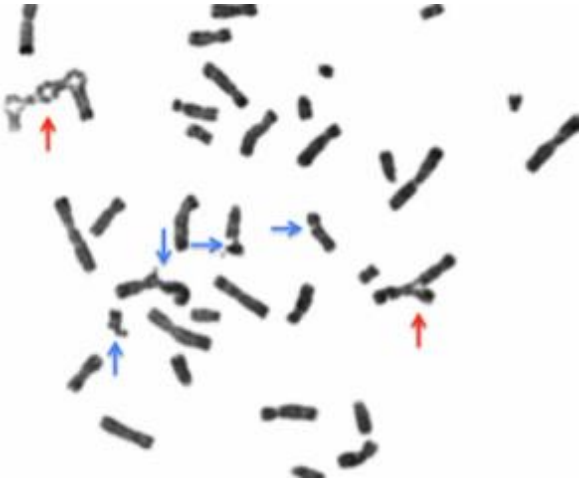


Fig. 1 Metafase después de exposición a Mitomicina C.
Flechas azules señalan rupturas cromosómicas y flechas rojas figuras radiales
(García-de Teresa et al., 2019)

1.2.- Vía de reparación FA/BRCA

Los productos proteicos de los 22 genes de AF, junto con las proteínas asociadas a AF (FAAPs), interactúan en una red bioquímica común para reparar ICLs, conocida como la vía FA/BRCA, que está implicada en el funcionamiento correcto de varios procesos celulares. Esta vía actúa en conjunto con otros mecanismos de la reparación del ADN como la reparación por escisión de nucleótidos (NER), la síntesis translesión (TLS) y la recombinación homóloga (HR), y con otras vías de reparación del ADN, en menor medida (Niraj *et al*., 2019; Rodríguez & D'Andrea, 2017). Las 22 proteínas se organizan funcionalmente en tres grupos (Fig. 2) (Helbling-Leclerc *et al*., 2021):

I. Proteínas que constituyen el complejo central AF ("*core*") o rio arriba (proteínas asociadas FAAP).
II. Proteínas del complejo FANCD2-FANCI
III. Proteínas rio abajo que permiten la incisión del ADN y la reparación de las lesiones.

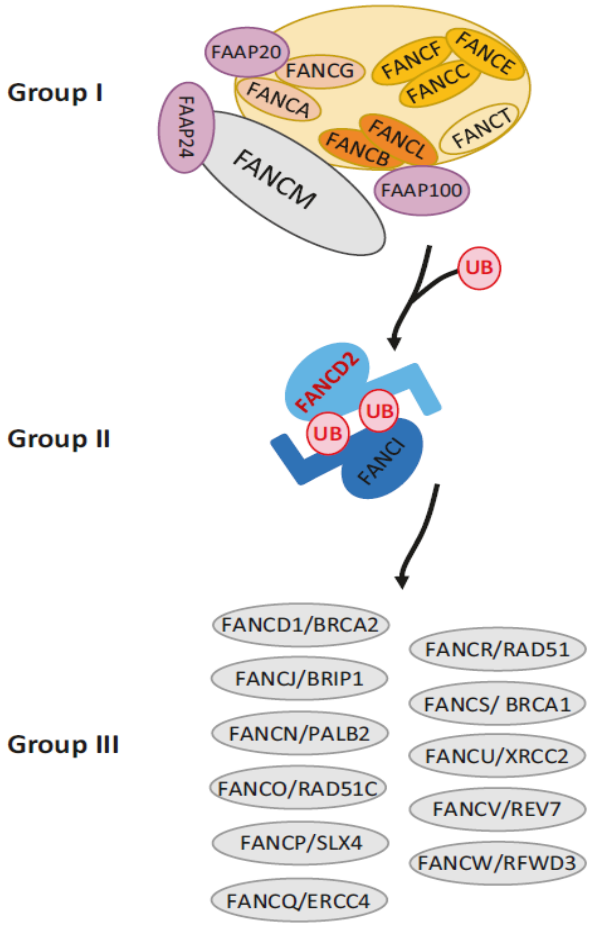


Fig. 2 Representación esquemática de los grupos de proteínas FANC. (Helbling-Leclerc et al., 2021)

1.2.1.- Formación y consecuencias de enlaces covalentes cruzados (ICLs)

El ADN se compone por cuatro bases nitrogenadas, dos purinas: adenina (A) y guanina (G) y dos pirimidinas timina (T) y citosina (C), estas bases se unen A con T y C con G por medio de dos y tres puentes de hidrógeno, respectivamente. Este tipo de enlaces confieren estabilidad suficiente al ADN para mantener la unión de las cadenas complementarias, sin embargo, se requiere su separación al momento en que se lleve a cabo la replicación del ADN. Los ICLs son lesiones que unen covalentemente dos bases de las cadenas complementarias del ADN, es decir, generan el tipo de enlace químico de mayor estabilidad (Fig. 3), y se forman como consecuencia de la exposición a sustancias químicas con dos grupos electrofílicos reactivos (Clauson *et al*., 2013). Existen compuestos químicos exógenos que generan ICLs, los cuales son usados principalmente como quimioterapéuticos, estos son la mostaza nitrogenada, el cisplatino, la mitomicina C y el psoraleno. También existen agentes endógenos inductores de ICLs, que surgen del metabolismo celular normal como el metabolismo de componentes de la dieta o hidrólisis espontánea de purinas; el metabolismo de los aminoácidos y las poliaminas; y a través de la peroxidación de los lípidos se generan diversos aldehídos que son altamente reactivos y poseen la capacidad de formar ICLs (Housh *et al*., 2021; Kee & D'Andrea, 2010).

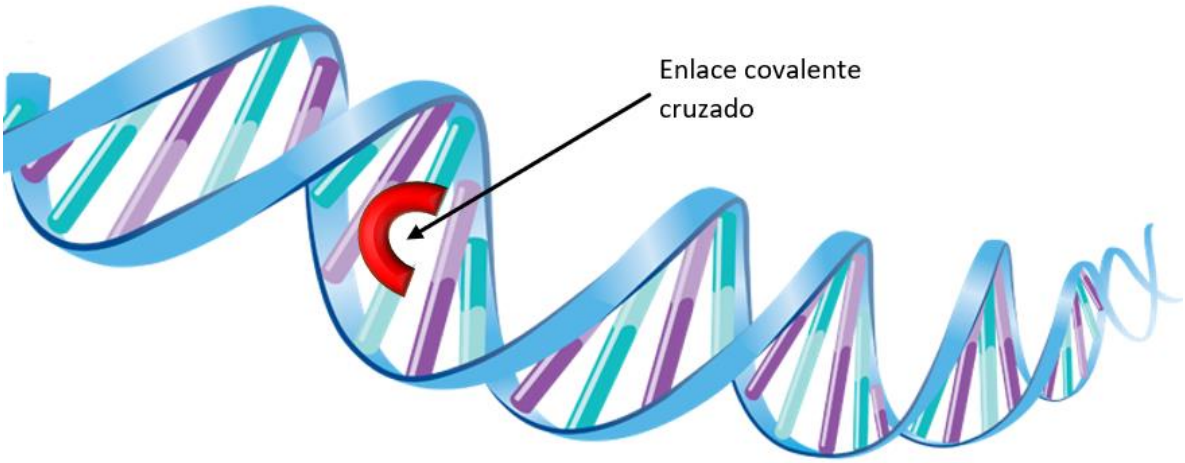


Fig. 3 Representación de un enlace covalente cruzado.
Imagen de elaboración propia.

Los ICLs representan una obstrucción de la maquinaria de transcripción y replicación, debido a que impiden la separación de las dos cadenas de ADN, resultando nocivos para las células que se encuentran en división. Sin embargo, las células cuentan con mecanismo de reparación específicos para esta lesión como es la vía FA/BRCA (Juárez-Figueroa *et al*., 2018; Rodríguez & D'Andrea, 2017).

1.2.2- Reparación de los ICLs

La reparación de ICL mediada por la vía FA/BRCA se produce principalmente en la fase S, cuando la replicación es detenida (Fig. 4). El primer paso es el reconocimiento de la lesión por la proteína UHRF1 y el complejo FANCM-MHF1-MHF2, posteriormente se reclutan 11 proteínas FANC adicionales que forman el complejo central AF. Este complejo monoubiquitiniza al complejo FANCD2 y FANCI, por medio de su actividad de ubiquitina ligasa. La monoubiquitinación del complejo ID2 activa la función endonucleolítica de FANCP/SLX4 y XPF/FANCQ, complejo que realiza una incisión para desenganchar el ICL. El desenganche genera tres lesiones intermediarias: una ruptura de cadena sencilla que se repara por síntesis de translesión; un aducto que se repara por NER y una ruptura de doble cadena que se repara por recombinación homóloga (HR). Si estas lesiones no se reparan o son reparadas erróneamente, pueden conducir a la inestabilidad genómica, como consecuencia del colapso de la horquilla de replicación que resultan en eventos de recombinación aberrante, como fusiones cromosómicas o figuras radiales características de los pacientes con AF. Finalmente, la ubiquitina carboxi-terminal hidrolasa 1 (USP1) en conjunto con UAF1 desubiquitinizan el complejo ID2 para completar la vía del AF (García-de Teresa et al., 2020; Juárez-Figueroa et al., 2018; Kee & D'Andrea, 2010; Niraj et al., 2019; Palovcak et al., 2017; Wang & Gautier, 2010). En el caso de las células que no se dividen, la reparación ocurre por NER (nucleotide excision repair, por sus siglas en inglés) y en regiones activamente transcritas por NER acoplada a la transcripción (Niraj *et al*., 2019).

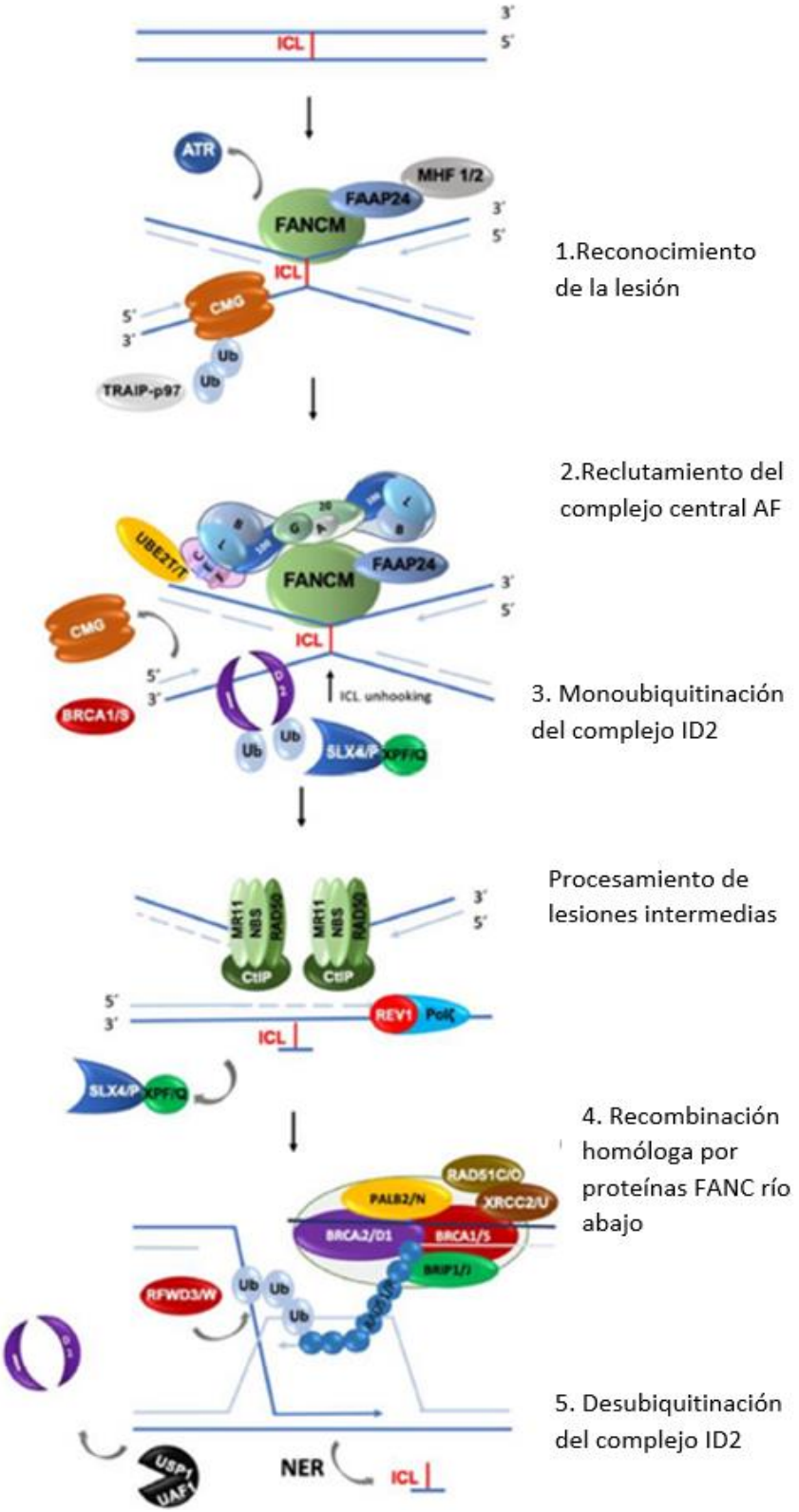


Fig. 4 Reparación de ICL por medio de la ruta FA/BRCA

(García-de Teresa et al., 2020)

1.3.- Consecuencias de la deficiencia vía FA/BRCA: cáncer, inestabilidad genómica y senescencia

Se ha observado que la presencia de mutaciones germinales en genes de proteínas involucradas en el proceso de reparación de daño al ADN, como ocurre en la AF, predispone al desarrollo del cáncer por la adquisición de un "fenotipo mutador"; este evento propicia la evasión de los *checkpoints* y mutación en genes esenciales en la señalización y proliferación (Palovcak *et al*., 2017). A partir de las evidencias anteriores, se sugiere que el estudio de la vía FA/BRCA, al estar involucrada en el mantenimiento de estabilidad genómica, representaría una base para el entendimiento de procesos carcinogénicos.

Un ejemplo especifico del proceso de inestabilidad genómica y sus consecuencias son los pacientes Fanconi con mutaciones bialélicas en *BRCA2* (- / -), ellos presentan un riesgo a desarrollar tumores sólidos así como de malignidad mieloide de mal pronóstico (Myers *et al*., 2012), posiblemente contribuyan en la aparición de glioblastoma en pacientes pediátricos (Dodgshun *et al*., 2016)y, de manera general, se asocia a fenotipos graves de cáncer de forma prematura y malformaciones congénitas (Feben *et al*., 2017). Adicionalmente, se ha apreciado que las mutaciones en estado heterocigoto de genes de la vía FA/BRCA contribuyen al riesgo de cáncer en personas sin AF. Esto genera una fuerte asociación entre los defectos genéticos en la vía FA/BRCA y el cáncer.

Para fortalecer esta premisa, existen estudios en los que se observa que una mutación de la línea germinal en un alelo del gen *BRCA2* aumenta el riesgo de cáncer de mama a casi un 50% y un 15% a cáncer de ovario (Nalepa & Clapp, 2018); además de estos tipos de cáncer, se han observado pacientes con mutaciones monoalélicas de *BRCA2* que muestran elevado riesgo de cáncer de páncreas y próstata (Venkitaraman, 2019).

La inestabilidad genómica incrementa la cantidad de mutaciones espontaneas en cada ciclo de replicación y se ha determinado que puede ser por: inestabilidad de microsatélites, inestabilidad nucleotídica y, por último, la inestabilidad cromosómica, la cual se manifiesta en metafases de pacientes con AF como AC, este tipo de inestabilidad se relaciona con la transformación oncogénica de las células (Palovcak *et al*., 2017).

Asimismo, se ha propuesto que las enfermedades monogénicas con inestabilidad cromosómica, como la anemia de Fanconi, cursan con envejecimiento acelerado (Fig. 5) debido a la característica clínica de presentar prematuramente patologías clásicas de edades avanzadas como falla medular y cáncer (Brosh et al., 2017; Che et al., 2018; García-de Teresa et al., 2020). En otros estudios se ha desarrollado una hipótesis que propone a la AF como síndrome asociado a la senescencia que, paradójicamente, presenta un impulso de selección de células pre-leucémicas que evaden el crecimiento y entran en la fase S del ciclo celular por diferentes mecanismos (Helbling-Leclerc *et al*., 2019, 2021).

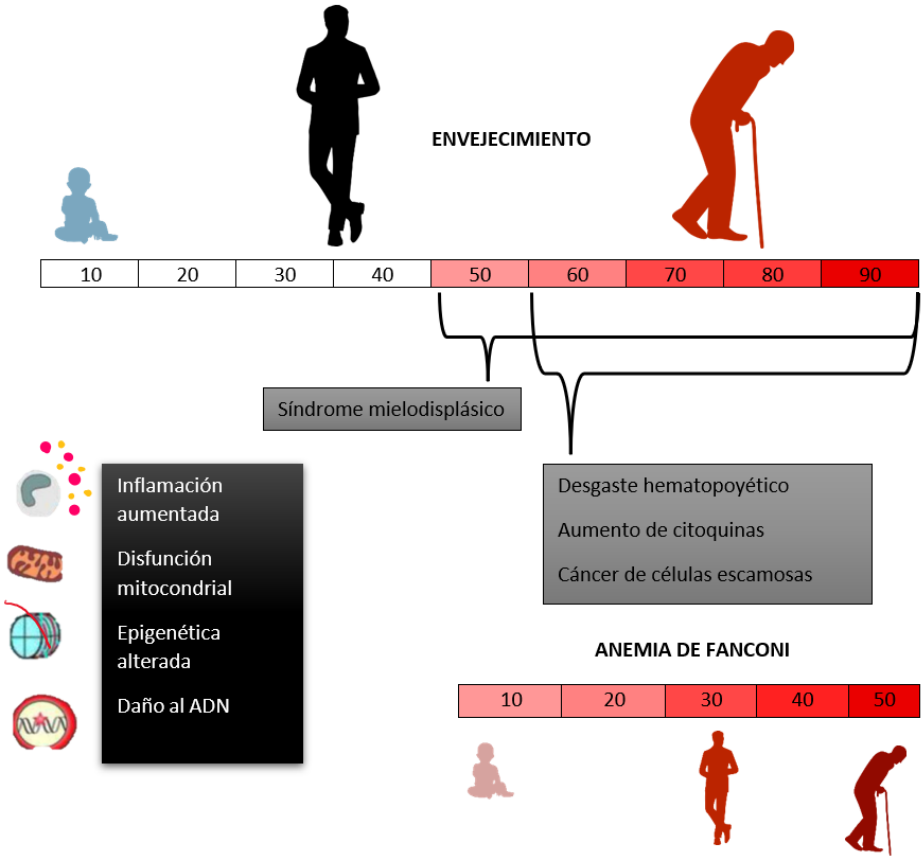


Fig. 5 Factores que propician envejecimiento en AF. Adaptado de (Brosh et al., 2017)

En épocas recientes se ha observado como común denominador del envejecimiento el daño acumulado, es decir que ocurre como consecuencia de diferentes afectaciones al genoma como el acortamiento de telómeros, alteraciones epigenéticas y, principalmente, inestabilidad genómica (López-Otín *et al*., 2013). Estos efectos se pueden estudiar por medio de citogenética; sin embargo, existen nuevas tecnologías que permiten el estudio del genoma completo usando biología molecular y más recientemente bioinformática, disciplinas que en conjunto permiten el estudio del daño específico y acumulado por medio de firmas

mutacionales. para ello es indispensable partir de secuencias de genoma o exoma completo las cuales se obtienen por medio de la secuenciación de nueva generación.

1.4.- Secuenciación de nueva generación

La secuenciación de ADN es una técnica de laboratorio que se basa en biología molecular y química, la cual es utilizada para determinar el orden exacto de las bases nitrogenadas (A, C, G y T), esta técnica ha pasado por diferentes cambios en el transcurso del tiempo hasta llegar a la secuenciación de nueva generación (NGS). El mayor avance que ofrece la NGS es la capacidad de producir un enorme volumen de datos a bajo costo ampliando el ámbito de la experimentación más allá de determinar el orden de las bases y provee mejoras a la práctica clínica.

Su aplicación en oncología ha aumentado debido a la premisa de que el cáncer se basa en mutaciones somáticas adquiridas, por tanto, el estudio a profundidad del genoma puede suponer ciertos beneficios (Behjati & Tarpey, 2013; Metzker, 2010; Shendure & Ji, 2008).

La principal plataforma utilizada en la secuenciación es Illumina, que se basa en una técnica conocida como "amplificación en puente" y admite una variedad de protocolos, incluida la secuenciación genómica, secuenciación del exoma y dirigida, secuenciación metagenómica, de ARN, ChIP-seq, y metiloma (Slatko *et al*., 2018). En esta tecnología se implementaron terminadores reversibles, lo que permite obtener información de manera cíclica. En Illumina, la secuenciación del ADN se realiza mediante análogos de nucleótidos marcados con fluorescencia que actúan como terminadores reversibles de la reacción de amplificación, como se detalla en los siguientes pasos:

1. Amplificación de los fragmentos de ADN mediante el método de PCR en puente.
2. Se añaden los cuatro nucleótidos terminales reversibles simultáneamente y son incorporados a la cadena complementaria.
3. Cada que se incorpora un nucleótido (cada uno marcado de color distinto), se emite una señal de fluorescencia, se obtienen las imágenes, se eliminan los tintes fluorescentes (Fig. 6) (Garrido-Cardenas *et al*., 2017).
4. Se obtiene una gran cantidad de datos que necesitan ser analizados para poder obtener resultados donde la bioinformática toma un papel muy importante en el procesamiento y análisis de los datos generados por NGS.

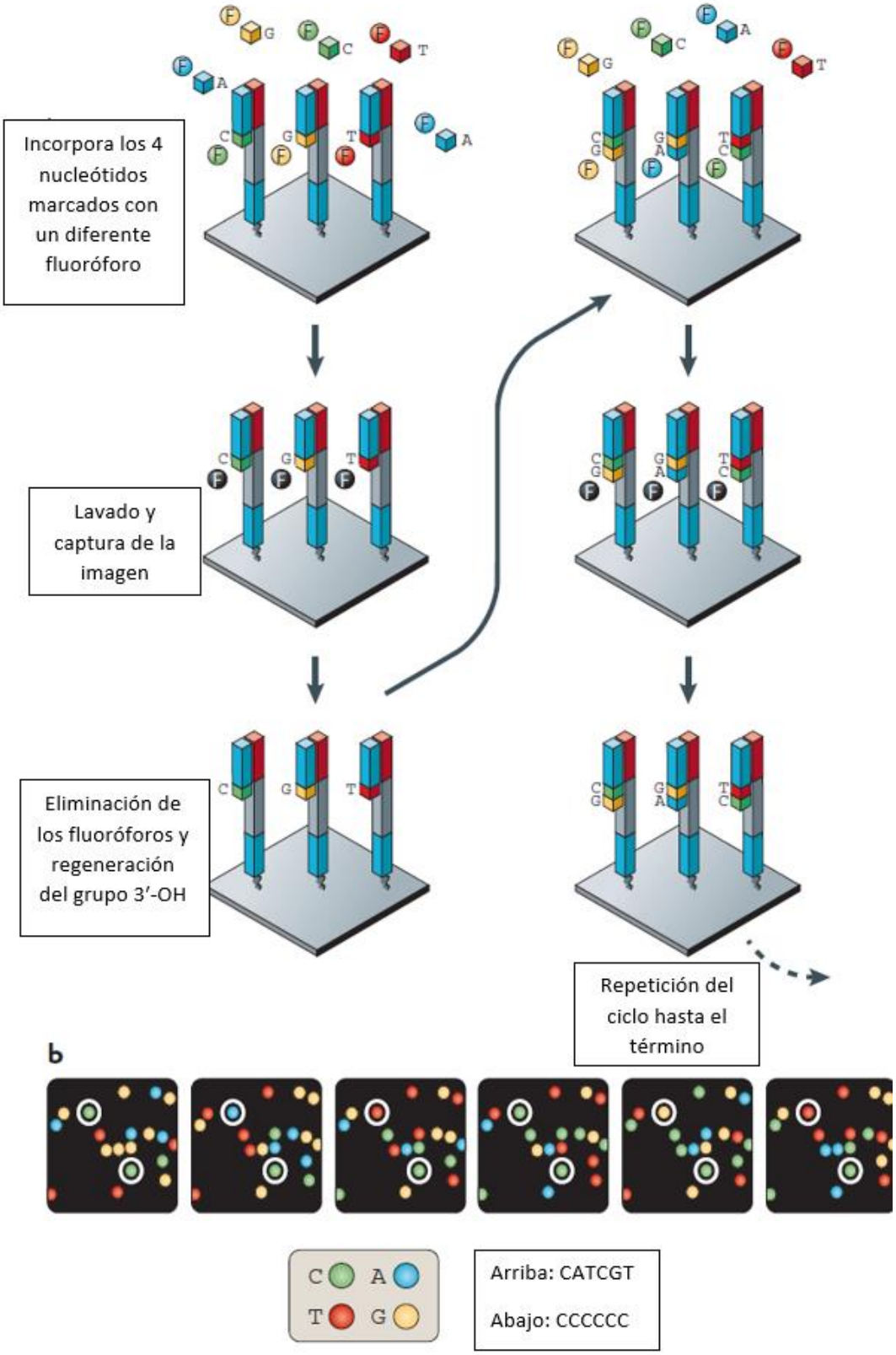


Fig. 6 Método de terminación reversible cíclica de cuatro colores de Illumina Modificado de (Metzker, 2010)

1.5.- Bioinformática como herramienta para investigación

La bioinformática es una ciencia multidisciplinaria que estudia la información biológica a partir de la teoría de la información, la computación y las matemáticas, surgió como consecuencia del enriquecimiento de datos biológicos, los cuales se obtuvieron principalmente por el proyecto del genoma humano y la colaboración de otras compañías privadas. Es una ciencia que, por su naturaleza versátil, puede ser aplicada en una amplia gama de experimentos. El estudio de los genes, proteínas y transcritos, no se limita al conocimiento de ellos, sino que considera la mejora del diagnóstico de enfermedades, terapia y pronóstico; en general, su aplicación principal está en campos como biología y medicina (Bayat, 2002; Cañedo Andalia & Arencibia Jorge, 2004).

Para realizar análisis bioinformáticos se requiere del uso de *software* e internet, a pesar de que son herramientas de fácil acceso, no implica que el análisis de datos crudos sea sencillo. El código utilizado en estas herramientas no es exactamente una receta que se pueda repetir, se suele tomar como base, pero en realidad se trata de una red de trabajo conjunto entre herramientas. Parte de esta "red" puede ser descifrada usando guías que son los *pipelines* (flujos de trabajo) que se desarrollan en herramientas bioinformáticas convencionales, apoyadas por el grado de "*expertise*" del usuario. Como punto de partida se puede considerar datos de secuenciación de ARN, inmunoprecipitación de cromatina (ChIP), metagenoma o, en el caso del presente trabajo, exoma y/o genoma completos (Bayat, 2002; Davis-Turak *et al*., 2017). Conforme al transcurso del tiempo, los datos obtenidos de genes y proteínas han sido conjuntados en repositorios llamados bases de datos (BD) y que generalmente se encuentran en internet. Sin embargo, no ha sido posible concentrar toda la información conocida en una base de datos única, lo cual eleva el valor del aprendizaje en análisis bioinformático dirigido a estudios particulares. Algunas de las BD más importantes para la obtención de datos de NGS son:

- *Gene Expression Omnibus* (GEO) y *Sequence Read Archive* (SRA) de NCBI (*National Center for Biotechnology Information*)
- *ArrayExpress* (AE) de EMBL-EBI (*European Bioinformatics Institute*)

Los análisis bioinformáticos basados en la NGS están diseñados para convertir las señales en datos, los datos en información interpretable y la información en conocimiento procesable. Este proceso puede conceptualizarse como análisis primario, secundario y terciario (Fig. 7)

1° Procesar datos crudos en formato FASTQ de los instrumentos de secuenciación para convertirlas en datos de bases de nucleótidos y lecturas cortas.
2° Alineación con una secuencia de referencia o el ensamblaje *de novo* de las lecturas de nucleótidos de la NGS y la posterior detección de variantes.

3° Contexto a la información generada durante un experimento de NGS al asociar el perfil genómico específico de la muestra con anotaciones descriptivas (Oliver *et al*., 2015).

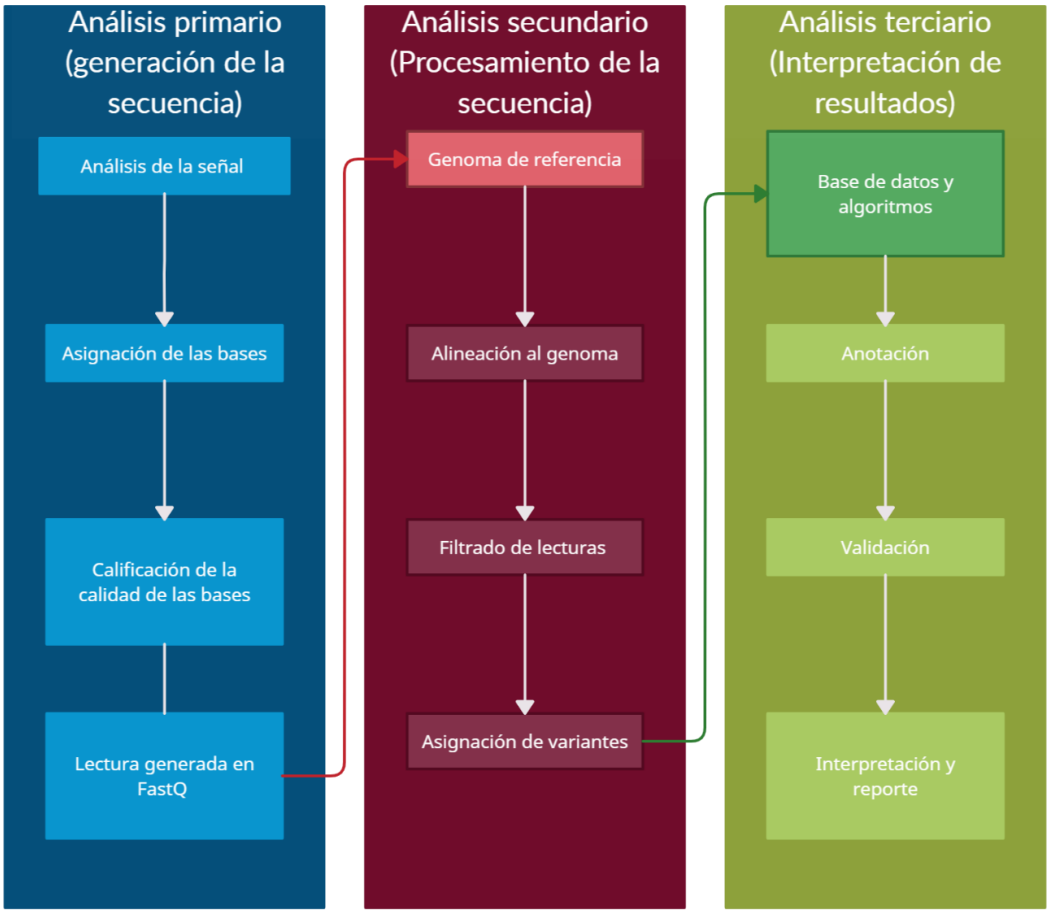


Fig. 7 Flujo de trabajo en el análisis bioinformático de datos NGS. Imagen de elaboración propia

1.5.1.- Bases de datos (GEO y SRA)

Las variaciones genéticas del genoma humano y su contribución a cambios fenotípicos han sido un tópico principal a lo largo de investigaciones asociadas a medicina y biología. A partir de la secuenciación del genoma humano y su disponibilidad pública, se han identificado variaciones en el ADN por medio de diferentes tecnologías y han permitido evaluar sus implicaciones en enfermedades humanas (Rawal *et al*., 2017). Existen bases de datos de interés bioinformático, de diferentes tipos. Por ejemplo: literarias, nucleotídicas, de proteínas, de ARN, de enfermedades; entre otros contenidos.
Gene Expression Omnibus (GEO) es una base de datos respaldada por el Centro Nacional de Información Biotecnológica (NCBI) que acepta datos brutos y procesados con descripciones del diseño experimental, los atributos de las muestras y la metodología usada (Clough & Barrett, 2016). Estos datos se encuentran usualmente en archivos de tipo SRA, que es el

principal archivo de datos de secuenciación de alto rendimiento, que archiva los datos de secuenciación en bruto y la información de alineación de las plataformas de secuenciación como *Illumina Genome Analyzer*®. La BD SRA acepta datos de todo tipo de proyectos de secuenciación; por ello, es posible descargar archivos de este tipo y convertirlos a FASTQ por medio de SRA *toolkit* con los cuales se realizará el *pipeline*.

2.- Antecedentes

2.1.- Firmas mutacionales en cáncer

Una firma mutacional es un patrón característico de diferentes mutaciones somáticas, que se puede observar a lo largo de un proceso mutacional del ADN asociado frecuentemente a diferentes tipos de cáncer (Alexandrov, Nik-Zainal, Wedge, Aparicio, *et al*., 2013). Las firmas mutacionales han sido clasificadas esencialmente en tres categorías: sustitución de base sencilla; de doble sustitución de base; e INDELS (pequeñas inserciones o deleciones). En la primera categoría, se considera que cada pirimidina puede sufrir dos transversiones, cambio de pirimidina por purina, y una transición cambio de pirimidina por pirimidina, es decir T>A,T>G,T>C,C>A,C>G y C>T, un total de seis variantes nucleotídicas simples, (Fig. 8), expresadas en contexto de trinucleótido, es decir, considerando que base se encuentra inmediatamente en 5' y 3', con cuatro posibilidades para cada posición (cuatro bases nucleotídicas) y cada mutación posible colocada en una caja negra (Fig. 9); generando un total de 96 combinaciones (4*6*4) que son expresadas en el eje "X" de una firma mutacional, mientras el eje "Y" representa el porcentaje o cantidad de mutaciones globales (Fig. 10).

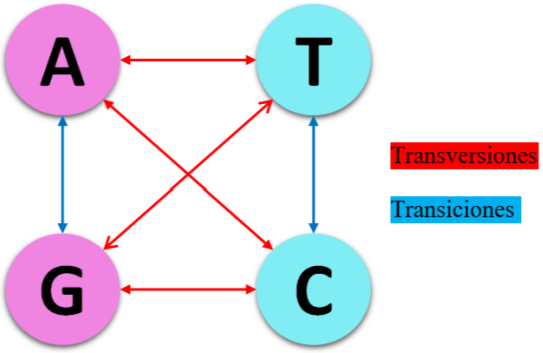


Fig. 8 Transiciones y transversiones de purinas y pirimidinas
Imagen de elaboración propia.

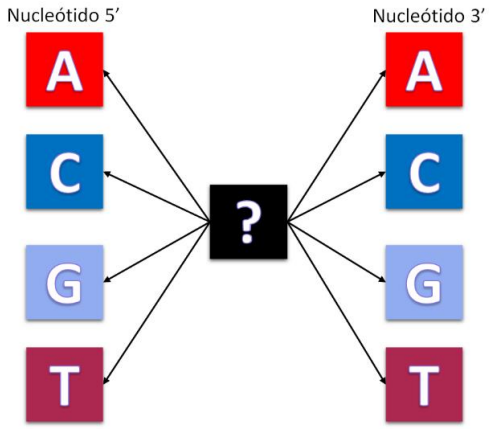


Fig. 9 Combinación de mutaciones en el contexto de trinucleótido.
Imagen de elaboración propia.

La disposición del porcentaje de cada mutación específica presente en diferentes muestras y tipos de cáncer han permitido identificar más de 30 firmas de sustitución de nucleótido sencillo (SBS) (Fig. 11). Estas firmas se pueden encontrar en el catálogo de mutaciones somáticas en cáncer (COSMIC) y cada firma presenta validación experimental y/o estadística según corresponda. Sin embargo, hay que considerar que a pesar de tener evidencia que las respalda estas firmas hablan de los procesos mutacionales que ocurren en el transcurso de una enfermedad. Para entender esto, se tiene el ejemplo de las firmas SBS4 y las de tipo SBS7. La firma SBS4 está asociada a consumo de tabaco y se puede observar en ella que existe una prevalencia de mutaciones tipo C>A las cuales son el reflejo de la formación de aductos de guanina, daño que causan los compuestos aromáticos policíclicos que se encuentran en el tabaco. Por otra parte, las firmas SBS7a y SBS7b, que se han encontrado casi exclusivamente en melanoma, muestran una gran afinidad por mutaciones de cambio C>T las cuales reflejan la formación de dímeros de pirimidina como consecuencia de los daños de la luz ultravioleta.

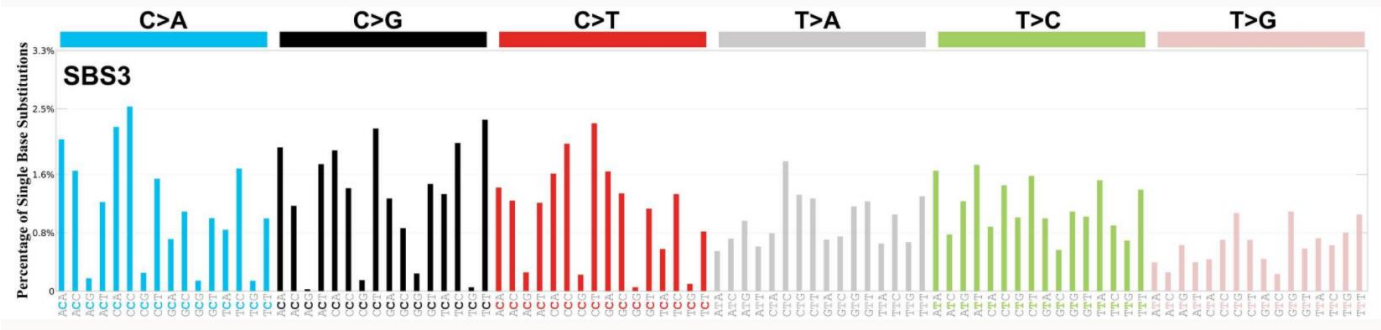


Fig. 10 Esquema de la firma mutacional SBS3

(COSMIC | Mutational Signatures, n.d.)

Cabe señalar que no todas las firmas presentan predominancia de ciertos tipos de mutaciones, existen también firmas que tienen más o menos la misma proporción para cada categoría siendo algunas firmas más fáciles de identificar los procesos a los que se les asocia en comparación de otras.

Fig. 11 Firmas validadas de sustitución de nucleótido único.
En recuadro azul las firmas asociadas a exposición de la luz UV y en amarillo la asociada a consumo de tabaco.
(Alexandrov et al., 2020)

En esta categoría se encuentran firmas como la SBS1 y SBS5 que han exhibido fuerte correlación con la edad en la mayoría de los tipos de cáncer de la niñez y la edad adulta, esto cimenta la hipótesis de que una proporción de las mutaciones de este tipo se adquieren con relativa constancia durante la vida del paciente, probablemente en tejidos somáticos normales (Alexandrov, Nik-Zainal, Wedge, Aparicio, *et al*., 2013).

Por otra parte, se encuentra la firma SBS3, que se asocia a defectos en la reparación por recombinación homóloga por mutaciones en *BRCA1* y *BRCA2* (Alexandrov *et al*., 2020).

Para el diseño de firmas de doble sustitución de base existe 16 posibles bases dobles de origen (4*4 bases) donde AT, TA, CG y GC son su propio complementario, las doce restantes pueden representarse como seis posibles dobletes de la cadena. Por lo tanto, hay 4+6=10 bases dobles de origen. Las bases dobles, AT, TA, CG y GC pueden ser sustituidos por seis posibilidades, mientras los dobletes restantes, pueden ser sustituidos por nueve posibilidades

que se forman por la combinación de tres posibles bases nitrogenadas que sustituyan a la de origen y diferentes de esta (3*3). Por lo tanto, en total hay 4*6 + 6 *9 = 78 combinaciones posibles para esta firma (Fig. 12).

Cabe destacar que no todas las firmas tienen una propuesta de etiología asociada, por ejemplo, las firmas DBS4, 8, 9 y 11 se consideran de etiología desconocida, mientras la DBS7 y la DBS10 están asociadas a defectos en reparación *mismatch* (Fig. 13).

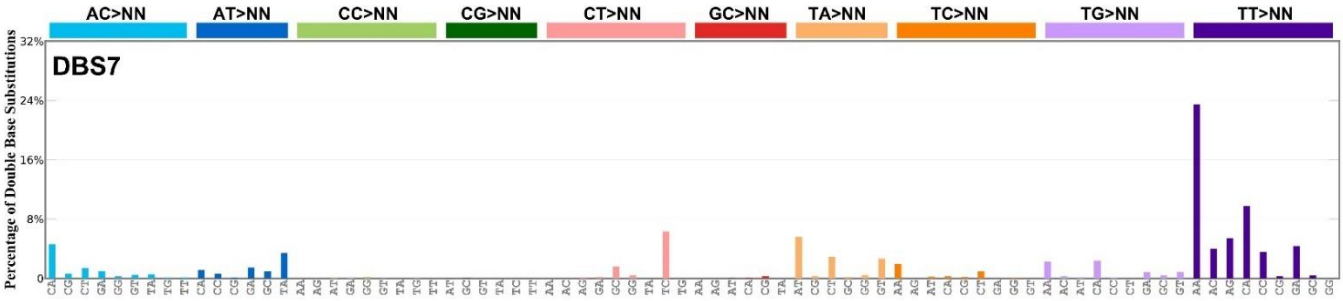


Fig. 12 Esquema de la firma mutacional DBS7.
(COSMIC | Mutational Signatures, n.d.)

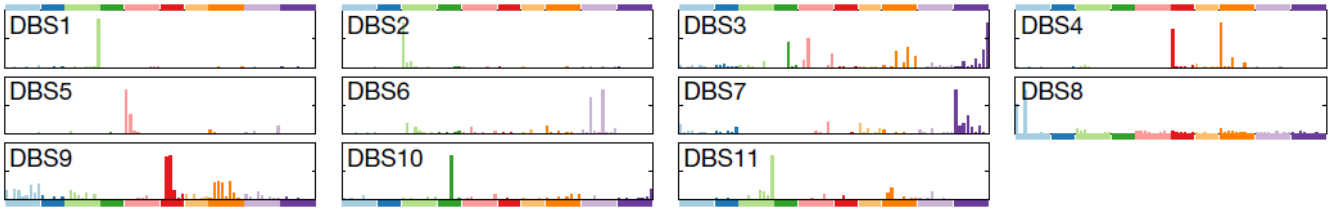


Fig. 13 Firmas validadas de doble sustitución
(Alexandrov et al., 2020)

La tercera categoría, INDELS, se categoriza según el número de bases que son insertadas o eliminadas, esta categoría es menos intuitiva pues se consideraron ciertas regiones repetitivas y si estas poseen microhomología. Se plantean 83 tipos diferentes que surgen de 5 categorías: 1 bp eliminado, 1 bp insertado, > 1bp eliminado, > 1 bp insertado y deleciones con microhomología; cada una se desglosa en la longitud de homopolímero, unidades repetidas o longitud de la microhomología dependiendo el caso (Fig. 14).

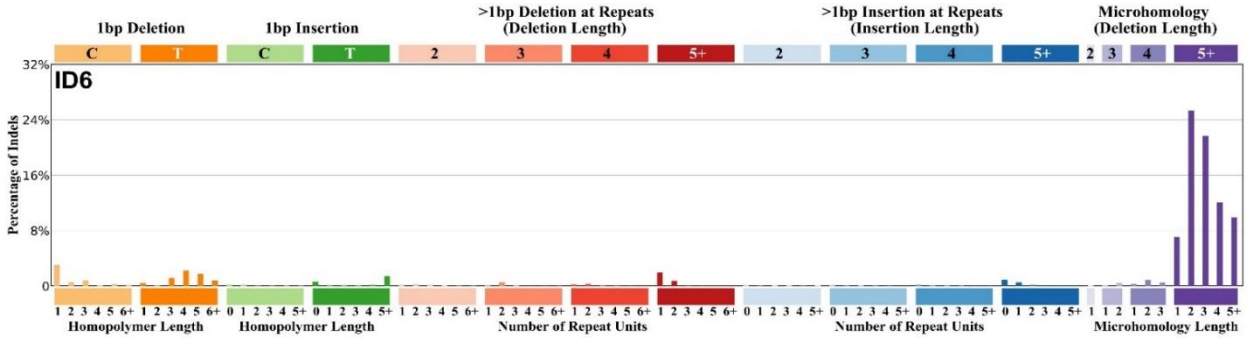


Fig. 14 Esquema de firma ID6
(COSMIC | Mutational Signatures, n.d.)

De ellas, la firma ID6 presenta mutaciones con microhomología y se asocia a defectos por recombinación homóloga mediada por microhomología (Fig. 15)

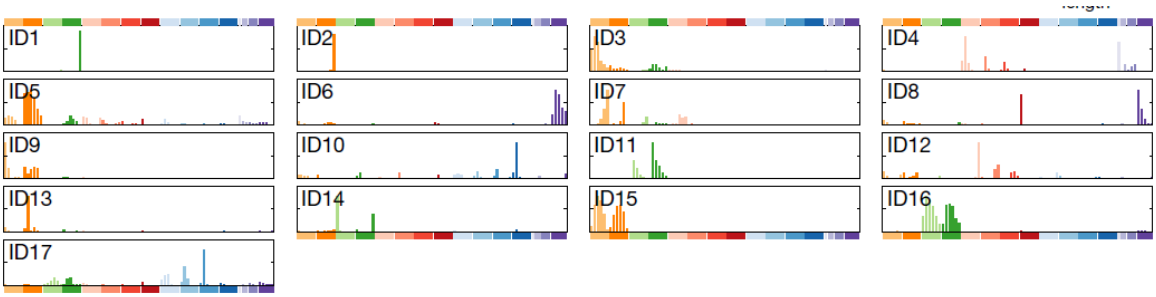


Fig. 15 Firmas validadas de INDELS (Alexandrov et al., 2020)

En la mayoría de los tipos de cáncer, se presentan al menos dos firmas mutacionales, y parece que el número de firmas se asocia con la complejidad de los procesos mutacionales en cada tipo, además algunas firmas presentan mayor especificidad hacía ciertos tipos de cáncer, como las firmas SBS4 y SBS7 que se encuentran principalmente en adenocarcinoma de pulmón y melanoma respectivamente, mientras otras están presentes en un grupo más homogéneo de ellos, como las SBS1 y SBS5 (Alexandrov *et al*., 2020). Además, se ha observado que aquellos cánceres con exposiciones crónicas presentan una mayor tasa de mutaciones totales en comparación con las de cáncer de alta incidencia en jóvenes, esta variación en la prevalencia de las mutaciones es atribuible a las diferencias en cuanto a la duración del linaje celular entre el óvulo fecundado y la célula cancerosa (Alexandrov, Nik-Zainal, Wedge, Aparicio, *et al*., 2013).

El repertorio de firmas mutacionales proporciona una base para la investigación de la etiología de las diferencias geográficas y temporales en la incidencia del cáncer; las representaciones gráficas de las firmas mutacionales permiten visualizar que tipo de mutaciones son predominantes o si se presentan de manera más o menos proporcional, o si estas se encuentran presentes mayoritariamente en la hebra transcrita o en la no transcrita, y esto relacionarlo a los mecanismos biomoleculares de los eventos mutacionales. Estos se podrían estudiar en tejido sanos o no neoplásicos y tener aplicaciones clínicas y de salud pública para comprender los procesos mutacionales que subyacen a la carcinogénesis (Alexandrov *et al*., 2020; Alexandrov, Nik-Zainal, Wedge, Aparicio, *et al*., 2013).

2.2.- Estudio de daño acumulado por medio de firmas mutacionales.

Para procesar, filtrar e interpretar los datos de secuenciación para obtener firmas mutacionales, es necesario el uso de métodos estadísticos y matemáticos, además de poseer conocimientos de la biología y genómica del cáncer o afección en estudio (Maura *et al.*, 2019). El repertorio de firmas mutacionales nos proporciona una base para la investigación

de la etiología de los procesos mutacionales, permitiendo comprender los mecanismos de estos que subyacen a la patología (Alexandrov *et al*., 2020).

Estos algoritmos de análisis de firmas mutacionales producen una matriz M≈ P*E, tomando un valor predefinido de firmas *n* a partir de la factorización de M para diversos valores y determinando la mejor *n* la cual permitirá la reconstrucción del error más pequeña (Fig. 16). Esto se explica con el ejemplo de una fiesta dónde los invitados serían las firmas mutacionales y estas tienen diferentes conversaciones, las cuales son grabadas con micrófonos colocados por toda la habitación; la cual sería la muestra de cáncer; y de ellos se obtienen grabaciones, que son análogas al catálogo de mutaciones, dónde algunas conversaciones son más comprensibles por tener mayor volumen o proximidad a los micrófonos, esto sería análogo a la exposición a los procesos mutacionales; y finalmente hay que considerar el ruido ambiental o la música de fondo que sería equivalente al error. Entonces aquellas conversaciones que mejor se puedan distinguir serán aquellas en dónde exista el menor ruido ambiental posible y las que mayor volumen tengan (Alexandrov, Nik-Zainal, Wedge, Campbell, *et al*., 2013; Omichessan *et al*., 2019).

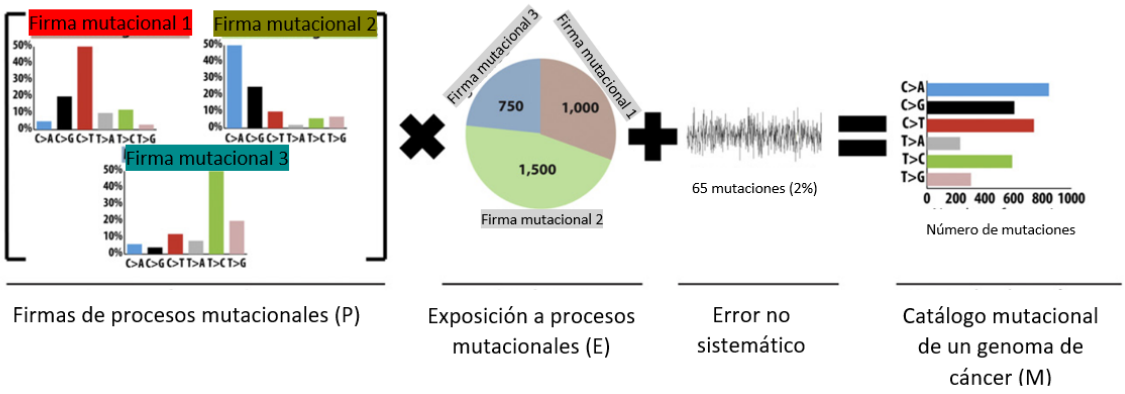


Fig. 16 Representación gráfica de la ecuación M≈P*E
Adaptado de (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013)

Las firmas qué mejor puedan distinguirse en una muestra de cáncer pueden proveer información sobre el tipo de sustituciones de base, así como el porcentaje de su incidencia. Estas en el conjunto de mutaciones de sustitución simple conforman un patrón con el cual se identifica daño genómico acumulado y específico. Además, en el caso de extracciones *de novo*, permiten una asociación única a una enfermedad en concreto como se ha observado con la fuerte correlación de firmas SBS7 a melanoma. En este caso en muestras de pacientes AF, que es una enfermedad no-cáncer pero que presenta altas incidencias de algunos tipos de cáncer podría presentar patrones de firmas únicas o firmas que tengan sentido biológico con relación a cáncer de los tipos más frecuentes que desarrollan estos pacientes. El tipo de daño observable a su vez puede expresar diferentes niveles de exposición, por tanto, en pacientes AF e individuos control de la misma edad se presentaría diferente cantidad de mutaciones

haciendo posible la detección de envejecimiento prematuro asociado al daño acumulado e inestabilidad genómica.

3.-Justificacion

Las firmas mutacionales revelan el curso de exposiciones a agentes genotóxicos exógenos o endógenos, así como deficiencias en las vías de reparación del ADN. Mas de 30 firmas mutacionales se han reportado recientemente en tejidos de diverso origen con cáncer, donde se ha observado exposición genotóxica o envejecimiento. Sin embargo, aún no se conoce si en otro tipo de enfermedad se puedan encontrar patrones específicos de firmas y donde nuestro interés está enfocado en pacientes con AF y explorar la posibilidad de encontrar firmas únicas para esta enfermedad, lo cual puede hipotetizarse ya que es un padecimiento con alteración en la vía FA/BRCA de reparación del daño al ADN. La relevancia del estudio se enfoca principalmente a dos eventos: (1) alta heterogeneidad fenotípica y genotípica y (2) el impacto de las enfermedades que pueden llegar a desarrollar estos pacientes, como se desarrollará a lo largo del presente trabajo.

En los pacientes con Anemia de Fanconi se conocen las mutaciones de los genes y mecanismos de herencia que causan la enfermedad, pero se desconocen las consecuencias de los procesos mutacionales intrínsecos de la enfermedad los cuales pueden estar relacionados con la severidad de las manifestaciones clínicas, la progresión del desarrollo de neoplasias o con envejecimiento prematuro. El estudio de firmas mutacionales en pacientes con AF puede ayudar a entender si existen firmas que puedan ser asociadas con la enfermedad. Con base en aquellas que se puedan observar, podría inferirse que procesos mutacionales se presentan a lo largo de la enfermedad y si estos se asocian a edad, a eventos de exposición crónica exógena o endógena, así como a determinar si la acumulación de daño genómico en AF es diferente que en individuos normales.

4.- Hipótesis

Con el desarrollo del pipeline bioinformático se encontrará una firma específica en los pacientes AF, así mismo, se encontrarán las firmas SBS1 y SBS5 que están relacionadas con envejecimiento y que pueden estar implicadas en el proceso de acumulación de daño y posible desarrollo a neoplasias presentes en AF y la firma SBS3 relacionada a deficiencia en la reparación del ADN por recombinación homóloga por la deficiencia intrínseca de la enfermedad.

5.- Objetivos

5.1.- Objetivo general

Desarrollar un pipeline bioinformático, que permita identificar las firmas mutacionales presentes en pacientes con Anemia de Fanconi, para determinar si existe envejecimiento prematuro.

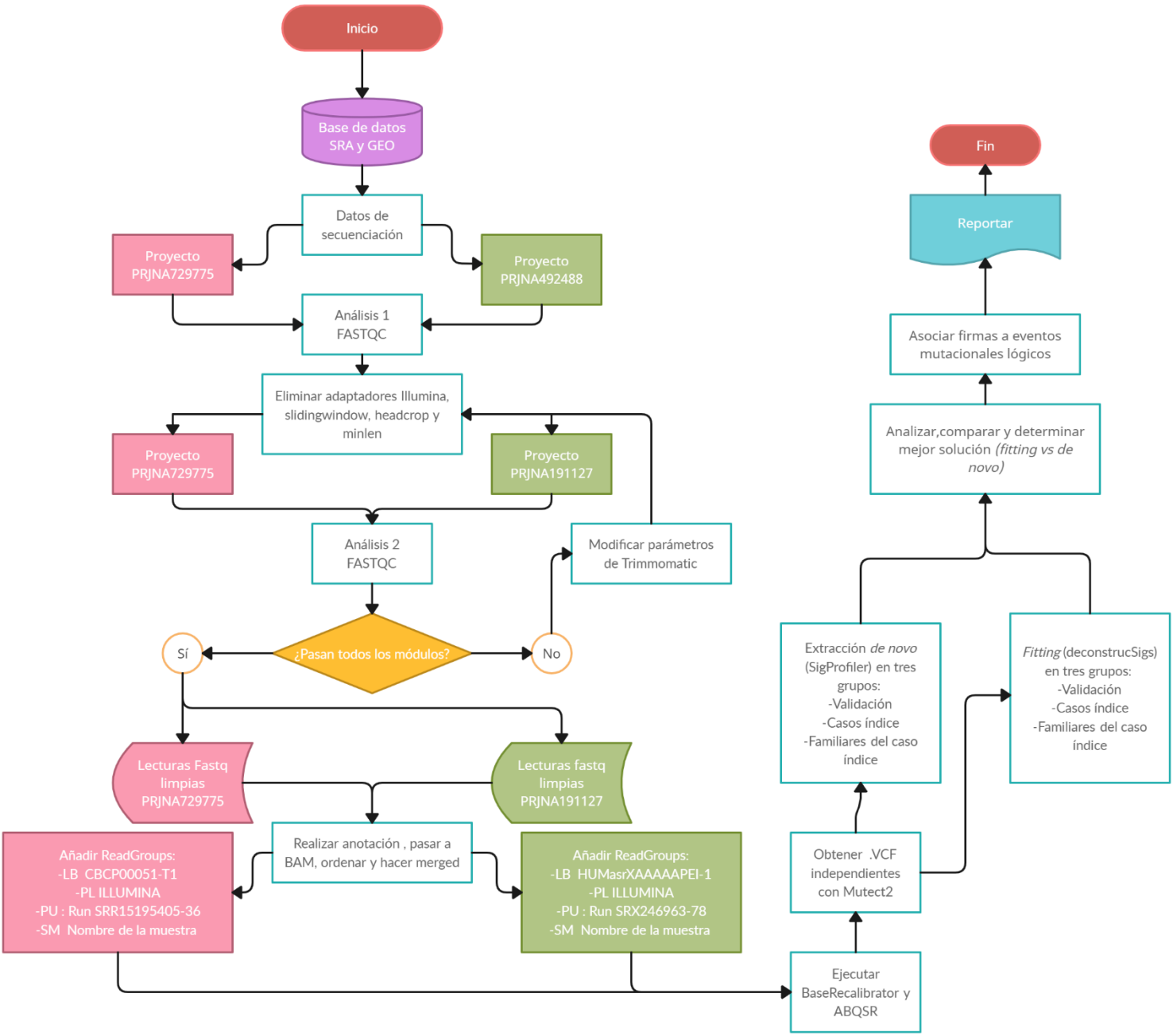
5.2.- Objetivos particulares

- Realizar el pre-procesamiento de datos de WES.
- Realizar el proceso de anotación de las lecturas para obtener el catálogo de variantes de nucleótido único (SNV).
- Desarrollar el pipeline de análisis de firmas mutacionales utilizando *SigProfiler* y *SignatureAnalyzer*.
- Analizar e identificar la presencia de firmas mutacionales reportadas o firmas específicas de los pacientes con anemia de Fanconi estudiados.

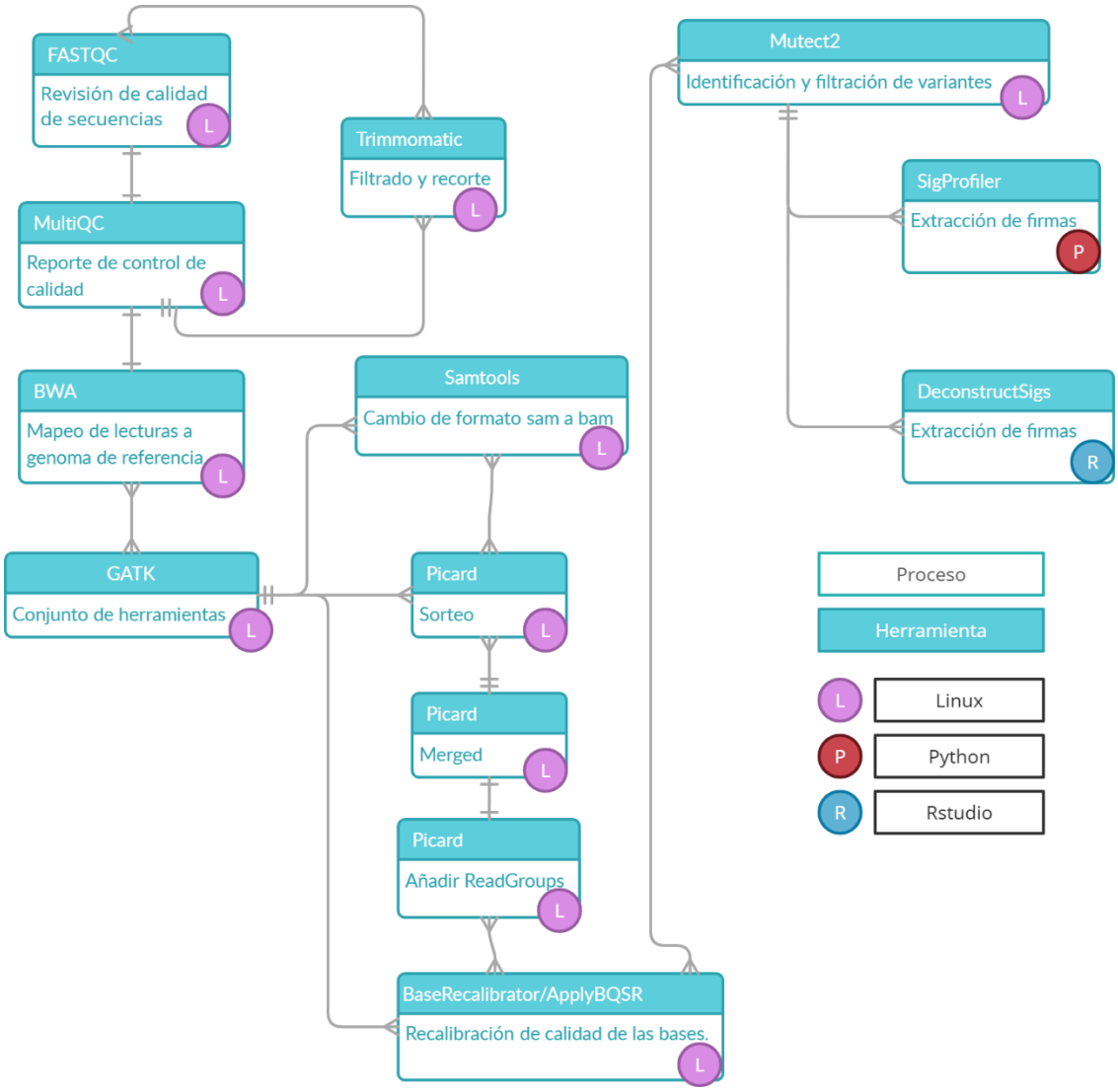
6.- Metodología

Para más detalles y acceso al pipeline elaborado, se creó un repositorio de Github al cual se puede acceder en la liga: https://github.com/ehatl/Mutational_Signature_pipeline o consultando los anexos del presente trabajo.

6.1.- Diagrama de flujo de trabajo.



6.2.- Diagrama metodológico



Para más información técnica de cada herramienta y lenguajes utilizados en el presente trabajo ver anexos I-XIII.

6.3.- Búsqueda de secuencias de exoma o genoma completo.

La búsqueda de estas secuencias se realizó en SRA con "Fanconi anemia" como palabra clave, buscando secuenciación de genoma o exoma en *Homo sapiens* y tipo de archivo FASTQ, que es el formato al que se transforman los archivos de salida de secuenciación de nueva generación. Se encontró un proyecto con el identificador PRJNA191127, el cual contenía 16 secuencias de exoma de pacientes AF (casos índices) y familiares directos. Por otra parte, se seleccionó el proyecto PRJNA729775, que contenía más de 200 secuencias de exoma obtenidos de tejido de cáncer de mama, de estas se seleccionaron cuatro muestras para

realizar la validación del pipeline, por ser provenientes de un cáncer cuyas firmas mutacionales han sido estudiadas. Cada secuencia se descargó por medio del identificador SRR y posteriormente convertidas a archivos FASTQ con la herramienta de NCBI llamada SRA toolkit.

6.4.- Análisis FASTQC

Una vez obtenidas las secuencias en el formato FASTQ, se realizó el análisis de estas con la herramienta FASTQC. Se evaluaron los módulos y se determinaron los parámetros para realizar el proceso de *trimming* posteriormente. El objetivo principal de este primer análisis fue conocer los posibles contaminantes de las secuencias, como son los adaptadores utilizados en secuenciación Illumina o los demás parámetros evaluados por el programa y si eran aptas para análisis posteriores. Para una visualización conjunta de todas las muestras analizadas en FASTQC, se utilizó la herramienta MultiQC. Este programa grafica las diferentes muestras en una sola gráfica para cada módulo evaluado, de esta manera se realizó un análisis más práctico, se comparó que muestras presentaron mejores condiciones y se seleccionaron las modificaciones pertinentes. Para más información ver anexo II.

6.5.- Trimming

Posterior al análisis en FASTQC, se procedió a seleccionar los parámetros que se utilizaron en Trimmomatic. En este caso se observó que el adaptador TruSeq3-PE-2.fa se encontraba en las lecturas, por lo cual se eliminó de ellas con la opción correspondiente. Así mismo se añadieron otras opciones para determinar el umbral de calidad para aceptar o rechazar una base de la secuencia, esto con la finalidad de mejorar la calidad de las bases secuenciadas. Además, se eliminaron las primeras 10 bases de la secuencia, debido a que generalmente son bases que tienen una mala calidad y no tienen una adecuada proporción de %AT y %GC. La selección de estos parámetros y su ejecución, permitieron obtener muestras filtradas y recortadas, en las cuales se preservó una calidad óptima para proseguir el análisis. Como resultado de correr el Trimmomatic, se generaron cuatro archivos de salida: dos archivos pareados y dos archivos no pareados. Para más información ver anexo III.

6.6.- Análisis FASTQC después de la limpieza

A los cuatro archivos de salida de cada muestra analizada, se les repitió el análisis redactado en el apartado 6.3 con la finalidad de corroborar que las muestras después del *trimming* dejaron de presentar interferencias significativas en cuanto a su calidad en los diferentes módulos y continuar con el análisis.

6.7.- Anotación y manejo de archivos SAM-BAM

Para realizar la anotación de las lecturas limpias, se realizó un mapeo con la herramienta BWA de los archivos de salida del *trimming* (en formato FASTQ) contra el genoma GRCh38 como referencia. Las secuencias resultantes se obtienen en formato SAM, pasan de ser cuatro archivos a tres, debido a que se anotan juntos los archivos pareados y de manera individual los no pareados. Con la herramienta de *Samtools* se convierten en archivos BAM. Primero se ordenaron por coordenadas, después se fusionaron para obtener un archivo BAM único y, finalmente, se les agregaron "*readgroups*" los cuales son características adicionales para un mejor rastreo e identificación de la muestra, además de ser parte de las buenas prácticas de GATK. Para mayor entendimiento de la programación de estas tareas ver anexo IV y V.

6.8.- Recalibración y llamado de variantes

Después de obtener los archivos BAM que contienen los *readgroups*, se realizó la recalibración de las muestras con *BaseRecalibrator* y posteriormente se aplicó con *ApplyBQSR*. Esto con la finalidad de detectar errores sistemáticos que ocurren durante la secuenciación al estimar la precisión de cada base y ajustar la puntuación. Una vez que se realizan los ajustes, se realizó el llamado de las variantes con Mutect2, herramienta especializada para variantes somáticas. El archivo de salida final es un .VCF que contiene las columnas con los datos necesarios para ser archivos de entrada de programas para la extracción de firmas mutacionales. Sin embargo, parte de las variantes que detecta muestran leyendas cómo evidencia débil, contaminación, etc. Por ello se filtraron aquellas variantes que estaban marcadas con la leyenda "*PASS*". Esto se realizó con comandos para filtrar en la terminal como se describe en el anexo XIV.

6.9.- Obtención de firmas mutacionales

Las diferentes firmas mutacionales se extraen de una gran serie de datos de secuenciación de genoma y exoma completos, por medio de factorización matricial no-negativa (NNMF). Este método tiene como principal utilidad encontrar una representación lineal de los datos, que no sean negativos. En el caso de las firmas de un proceso mutacional *n* se obtiene una matriz *K*N* de la siguiente manera:

$$P = \begin{pmatrix} p_1^1 & \dots & p_n^1 \\ \vdots & & \\ p_1^k & \dots & p_n^k \end{pmatrix}$$

Esto expresa la frecuencia relativa esperada de un tipo de mutaciones *k* ocurran en un genoma expuesto a un evento mutacional *n.* P es la matriz de firmas, compuesta por tipos de mutación como filas y firmas como columnas. La intensidad de las exposiciones a mutaciones se

representa en una matriz *N*G* que expresa la exposición de muestra dada *g* a un evento mutacional *n*. E es la matriz de exposición a procesos mutacionales, que tiene a las firmas como filas y muestras como columnas

$$E = \begin{pmatrix} e_1^1 & \dots & e_g^1 \\ \vdots & & \\ e_1^n & \dots & e_g^n \end{pmatrix}$$

Por último, la anotación de la matriz de la colección de muestras se representa como *K*G*. M es la matriz de catálogo, conformada por mutaciones como filas y muestras como columnas

$$M = \begin{pmatrix} m_1^1 & \dots & m_g^1 \\ \vdots & & \\ m_1^k & \dots & m_g^k \end{pmatrix}$$

Esto se logra gracias al uso de diferentes algoritmos como *SigProfiler* (extracción *de novo*) y *deconstructSigs* (extracción *fitting*) paquetes de Python y R respectivamente.

Se realizó la extracción *de novo* utilizando los archivos .VCF filtrados resultantes en la herramienta de *sigProfiler*. Esta herramienta utiliza archivos de entrada de tipo VCF, se indica el número mínimo y máximo de firmas que se desean extraer, así como el número de iteraciones que debe realizar para obtenerlas. Como archivos de salida se generan dos carpetas principales "All_solutions" la cual propone el número de soluciones según se haya indicado el número de firmas que se deseaban extraer y "Suggested_solution" que es precisamente la solución que mejor explica el set de muestras e incluye lo siguiente:

- SBS96_De-Novo_Solution: En esta carpeta se visualizaron gráficos para cada firma identificada que representa la proporción de las mutaciones para esa firma, son gráficos de todas las diferentes firmas *de novo*. Se generó un gráfico de actividad que muestra en cuales muestras está presente una firma *de novo* dada y en qué proporción (Fig. 17). Por último, permite visualizar una gráfica que presenta la media de mutaciones por megabase (Fig. 18).

- COSMIC_SBS96_Decomposed_Solution: En esta carpeta se pueden visualizar esencialmente los mismos gráficos, con la diferencia de que se simplifican a la solución que mejor explica el set de muestras. Adicionalmente muestra el gráfico de descomposición de la firma *de novo* en las firmas del catálogo que la forman (Fig. 19). De esta manera, *SigProfiler* permite una orientación tipo *fitting* que fue comparado y respaldado por un extractor diferente basado en este enfoque (*deconstructSigs*).

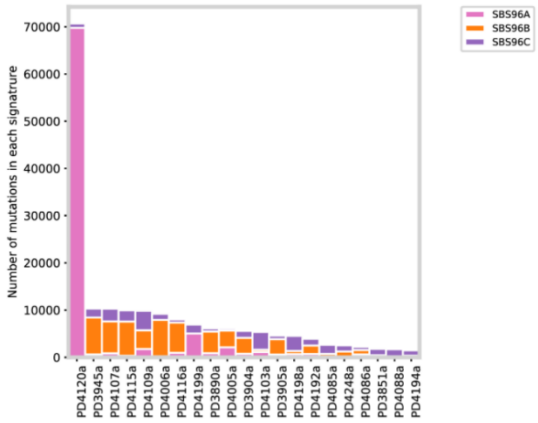


Fig. 17 Gráfica de ejemplo de actividad.

En el eje X se muestra el identificador de cada muestra y en el eje de las Y el número de mutaciones aportadas por cada firma de novo, las cuales están identificadas por diferente color.

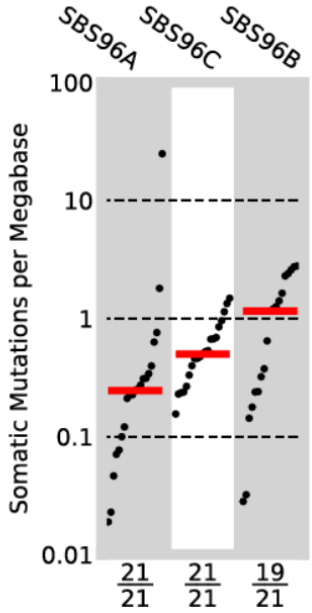


Fig. 18 Gráfica de ejemplo de carga mutacional de la muestra

El eje Y es las mutaciones somáticas por megabase y el eje X es el número de muestras con la firma indicada sobre el número de muestras totales. Los nombres de las columnas son las firmas mutacionales y el gráfico está ordenado por la media de mutaciones somáticas por megabase.

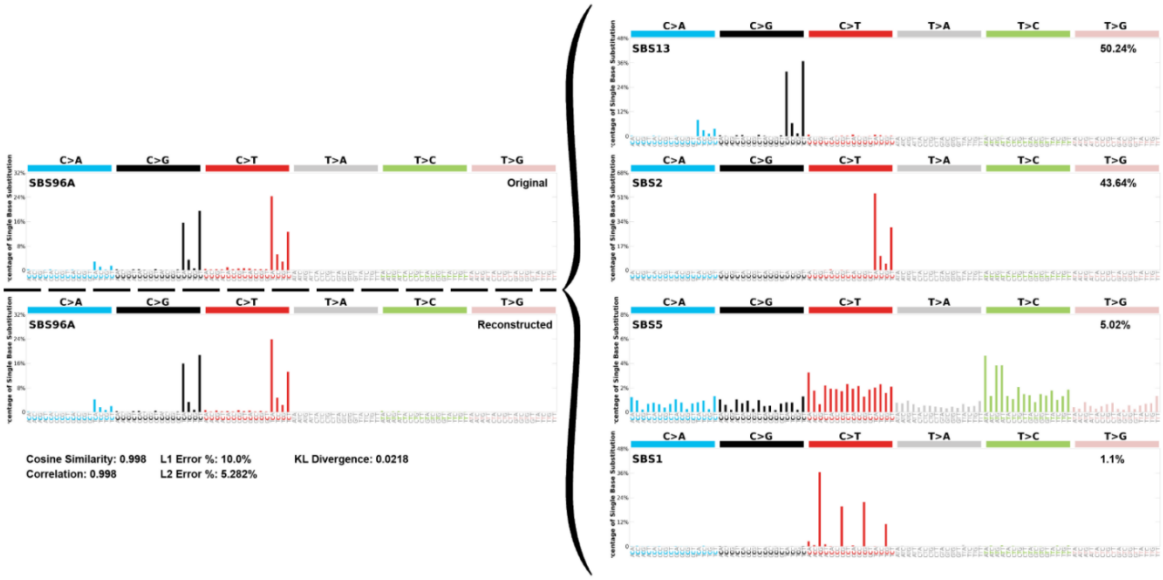


Fig. 19 Ejemplo de un gráfico de descomposición de firmas.
De lado izquierdo se puede observar la firma de novo original y la reconstrucción de esta con los datos de similitud de coseno (valores cercanos a 1 indican similitud), la correlación, los porcentajes de error L1 y L2 y la divergencia KL. Del lado derecho vemos la descomposición de la firma en las firmas propuesta por COSMIC.

Posteriormente, se obtuvo el archivo de entrada para el programa *deconstructSigs*, para realizar la extracción tipo *fitting*, a partir de la matriz obtenida. El archivo de entrada inicial no es un archivo .VCF a diferencia de *SigProfiler*, pero los datos pueden obtenerse de uno debido a que se requiere un archivo con la misma información, especificada en columnas como:

- Sample.id: Identificador de la muestra
- Chr: Cromosoma en el que se encuentra la variante
- Pos: posición en la que se encuentra la variante
- Ref: La base nucleotídica en esa posición en el gen de referencia
- Alt: La base por la cual cambió en la muestra estudiada.

A partir de estos datos se genera el archivo de entrada que contiene el cálculo de la fracción de mutaciones encontradas de cada tipo en los 96 contextos. Posteriormente basándose en la factorización matricial no negativa (NNMF) determina para cada firma, la solución que muestra la menor suma de los cuadrados del error (SSE) entre la muestra tumoral dada y el perfil tumoral reconstruido descartando aquellas que presenten < 6 % de la fracción de mutaciones encontradas para evitar falsos positivos (Rosenthal *et al*., 2016).

Los archivos de salida obtenidos son gráficos similares a los del catálogo de COSMIC, presentan la proporción de mutaciones por tipo de mutación en cada uno de los 96 contextos, añadiendo un encabezado que contiene la fracción de las firmas extraídas (Fig. 20), esta fracción además puede ser representada en un gráfico de pastel (Fig. 21).

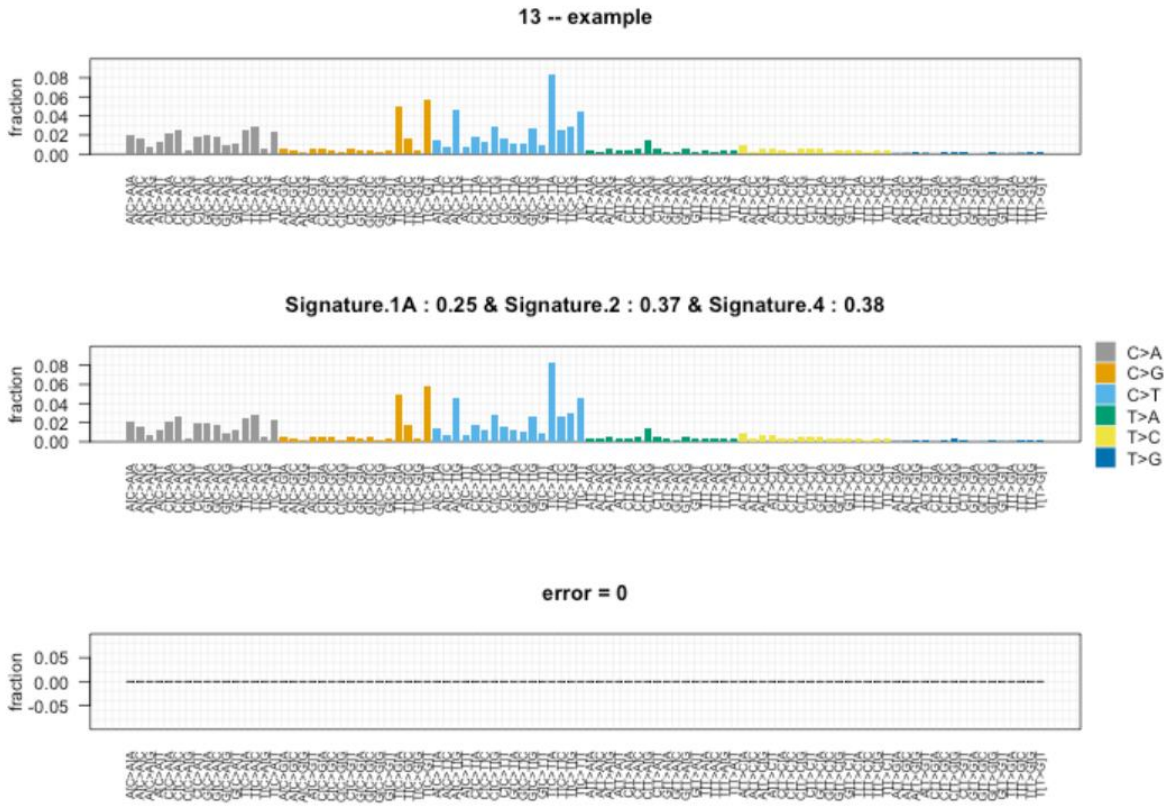


Fig. 20 Gráfico de ejemplo de firma extraída en deconstructSigs.
El primer gráfico representa la firma extraída original, debajo la gráfica de la firma reconstruida indicando las firmas presentes y la fracción en la que se encuentran. Finalmente, el gráfico del error SSE que debe ser cercano o igual a cero.

Se analizaron los gráficos obtenidos de *SigProfiler* y se determinó la firma que mejor explicó el set de muestras, así como la descomposición de esta para conocer las posibles firmas presentes. Por otra parte, se observaron y compararon con los gráficos de *deconstructSigs* y observaron las similitudes y diferencias entre ambos enfoques. Adicionalmente se analizó la viabilidad biológica de los resultados obtenidos de estas herramientas matemáticas.

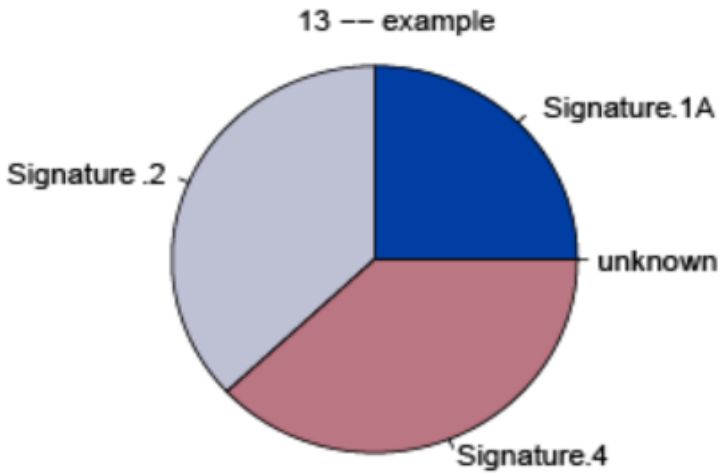


Fig. 21 Gráfico de pastel de ejemplo generado en deconstructSigs
El gráfico se realiza con las fracciones de firmas determinadas del gráfico anterior. La fracción de firmas desconocidas corresponde a mutaciones que no pudieron determinarse o bien se determinaron en proporción menor a 6%.

7.- Resultados

7.1.1- Resultados de multiQC de los archivos FASTQ antes y después del trimming del proyecto PRJNA729775 (pacientes cáncer de mama)

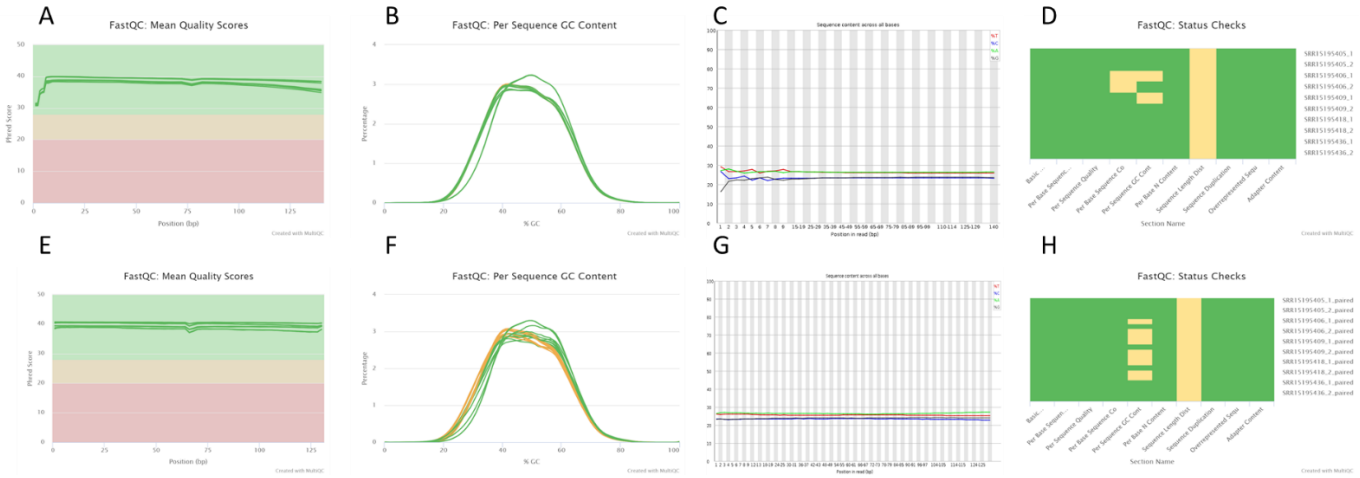


Fig. 22 Resultados del FASTQC antes y después del proceso de trimming con MultiQC. Histograma de las puntuaciones medias de calidad de las lecturas (A, E). Contenido promedio de GC (B, F). Gráfico del contenido de bases en las secuencias (C,G). Gráfica del estado general de las muestras en los diferentes módulos evaluados por FASTQC (D, H).

En la Fig. 22 se observa el resultado del proceso de evaluación de los archivos FASTQC, las imágenes de la parte superior son los datos crudos que se descargaron, como se puede apreciar en las imágenes las primeras posiciones presentan un poco de ruido (inherente al proceso de secuenciación). Las secuencias presentan buena calidad al encontrarse en el rango de puntuación superior a 30. La mayoría de las secuencias se rigen bajo una distribución normal respecto al contenido de GC con excepción de los *forwards* de las muestras SRR15195406 y SRR15195409, las cuales presentan una advertencia. El contenido de bases de las secuencias señala una advertencia debido a que el contenido de las bases no es completamente proporcional entre GC y AT.

A pesar de que tienen una calidad óptima se decidió realizar el proceso de *trimming*. En el panel inferior de la imagen se observan los archivos que salieron después de realizar el proceso de *trimming* con el programa *Trimmomatic* y se aprecia que la calidad de las muestras mejora en la parte inicial. Se observó mejor proporción del contenido de las bases después de realizar el *trimming*, sin embargo, el contenido de GC se modifica para algunos archivos de lecturas pareadas, a pesar de esta advertencia es posible continuar el proceso de análisis.

7.1.2.- Firmas mutacionales generadas por SigProfiler del proyecto PRJNA729775 (pacientes cáncer de mama)

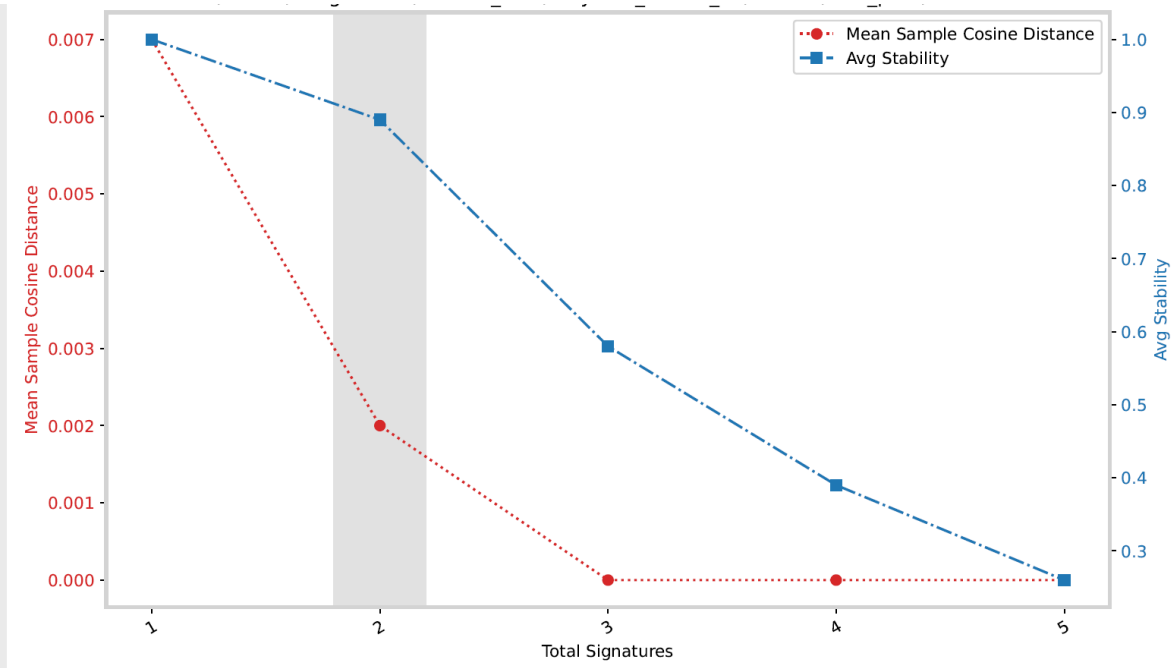
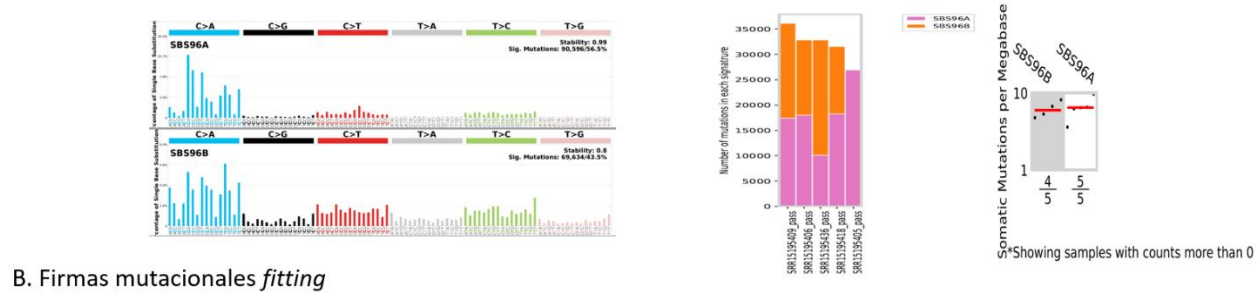


Fig. 23 Selección recomendada de número de firmas para análisis en cáncer de mama Se observa la recomendación del número de firmas adecuadas para el análisis de las muestras del proyecto PRJNA729775 de cáncer de mama. El grafico indica que después de evaluar los diferentes parámetros, la extracción de dos firmas es la mejor solución que se ajusta a muestras del presente proyecto.

A. Firmas mutacionales *de novo*



B. Firmas mutacionales *fitting*

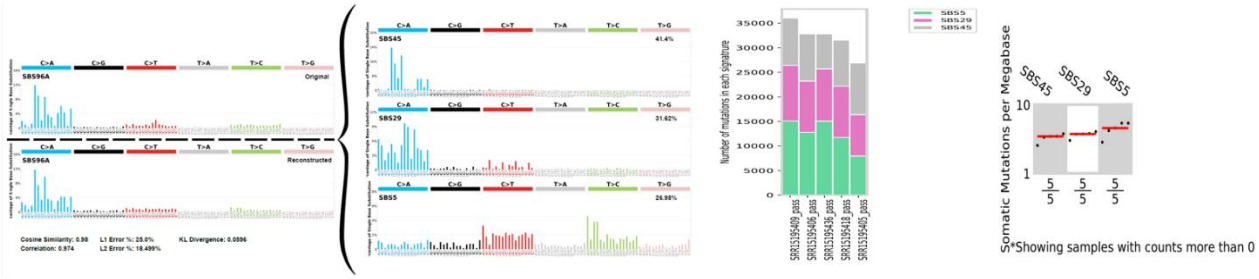


Fig. 24 Firmas mutacionales obtenidas con SigProfiler en cáncer de mama

En la figura Fig. 24 se observan los resultados del análisis de firmas mutacionales generado por *SigProfiler*. En el panel superior (A) se presentan los principales datos obtenidos al extraer una firma *de novo* para las muestras de cáncer de mama, en la primera figura de izquierda a derecha se observan las dos posibles firmas que explican el proceso mutacional de las muestras, las cuales poseen alto contenido de mutaciones C>A; posteriormente, en el panel central el número de mutaciones que explican cada firma en cada una de las muestras (SBS96A en rosa y SBS96B en naranja) y; en el último panel la media de mutaciones somáticas por megabase y la proporción en que se encuentra cada una de las firmas propuestas en las diferentes muestras. En el panel inferior (B) se presentan los principales datos obtenidos al extraer una firma de tipo *fitting*, en la primera figura de izquierda a derecha se observa las dos firmas extraídas *de novo* junto a su reconstrucción según las firmas del catálogo de COSMIC, las cuáles son en este caso la firma SBS45, SBS29 y SBS5; posteriormente, en el panel central se presenta el número de mutaciones de cada una de estas firmas en cada una de las muestras (SBS45 gris, SBS29 rosa y SBS5 en verde) y; en el último panel la media de mutaciones por megabase correspondiente a cada firma del catálogo y el número de muestras en que se encuentran en las diferentes muestras.

7.1.3.- Firmas mutacionales generadas por deconstructSigs del proyecto PRJNA729775 (pacientes con cáncer de mama)

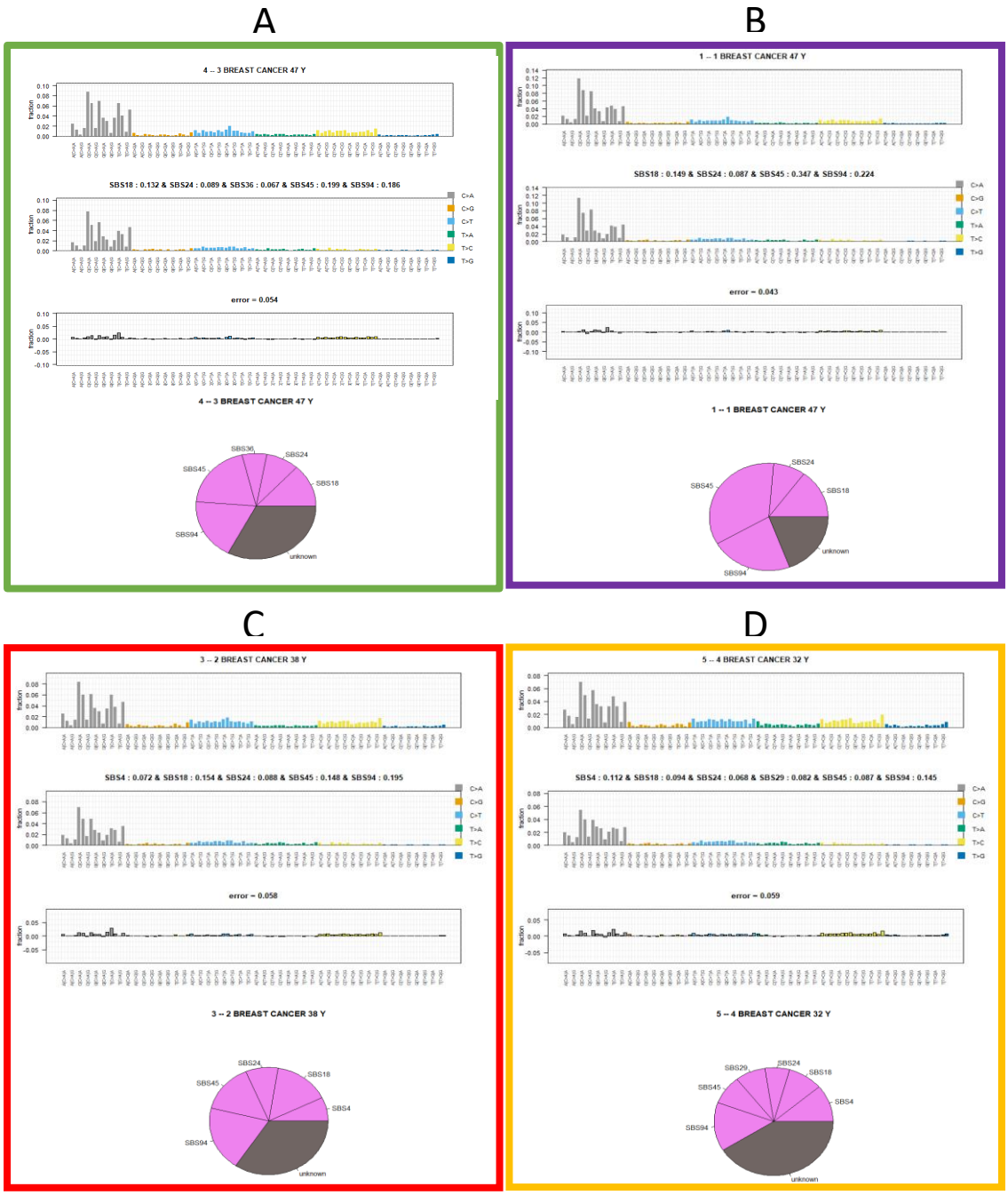


Fig. 25 Gráficos obtenidos en deconstructSigs de las muestras de cáncer de mama. Resultados de firmas mutacionales en tejido de cáncer de mama de pacientes con 47 años (A, B), resultados de firmas mutacionales en tejido de cáncer de mama de pacientes con 38 y 32 años respectivamente (C, D).

En la figura 25, en los cuatro paneles separados por marco de color, se observa de arriba hacia abajo histograma de la firma mutacional extraída; firma mutacional reconstruida con las firmas que posiblemente la compongan, incluyen la leyenda de la fracción en que se encuentran; gráfica de error y; gráfico de pastel que representa en color violeta los porcentajes en los que se encuentra una firma, en gris se representa la porción desconocida o que no se pudo asociar a alguna firma. Las muestras del panel verde y morado presentan en mayor proporción SBS45, mientras que las muestras del panel rojo y amarillo presentan mayor proporción de SBS94, como se desglosa en la siguiente Tabla 1.

Tabla 1 Proporciones obtenidas por firma con el programa de deconstructSigs

Muestra	Edad	Firmas	Proporción
Breast cancer 1	47	SBS45	0.347
		SBS94	0.224
		SBS18	0.149
		SBS24	0.087
Breast cancer 2	38	SBS94	0.195
		SBS18	0.154
		SBS45	0.148
		SBS24	0.088
		SBS4	0.072
Breast cancer 3	47	SBS45	0.199
		SBS94	0.186
		SBS18	0.132
		SBS24	0.089
Breast cancer 4	32	SBS94	0.145
		SBS4	0.112
		SBS18	0.094
		SBS45	0.087
		SBS29	0.082
		SBS24	0.068

7.2.- Resultados del proyecto de muestras de pacientes Fanconi y familiares

7.2.1.- Resultados de MultiQC de los archivos FASTQ antes y después del trimming del proyecto PRJNA191127

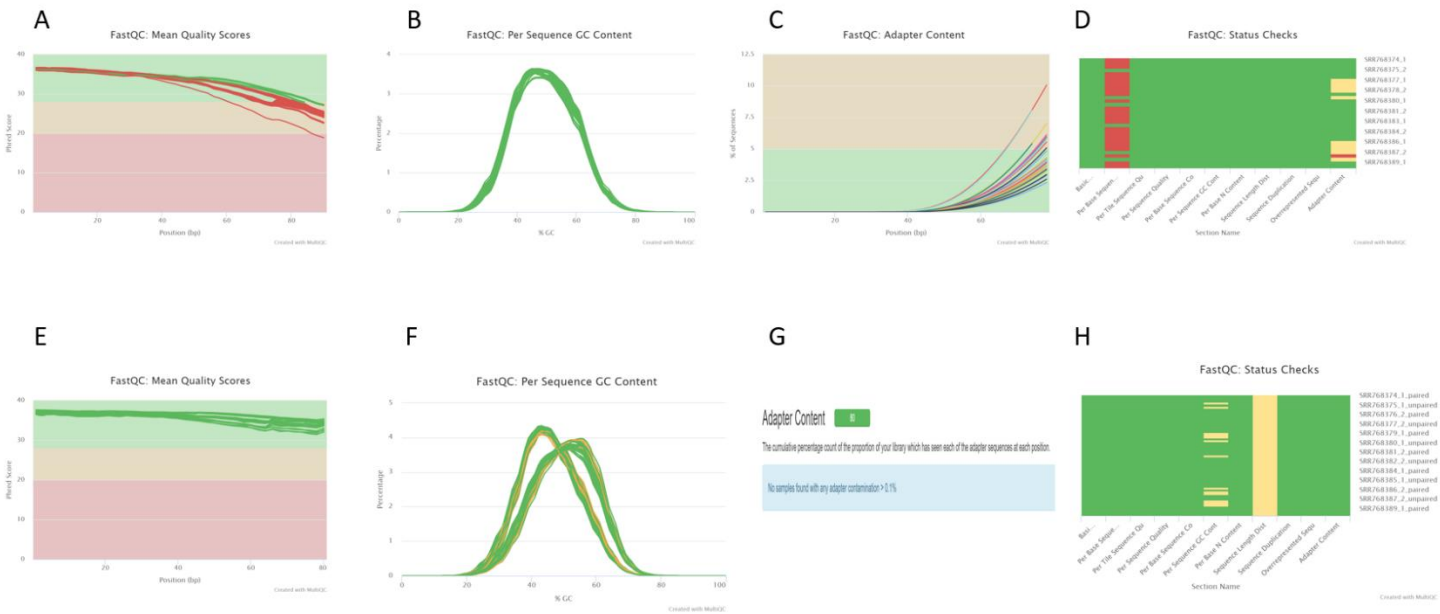


Fig. 26 Resultados del FASTQC antes y después del proceso de trimming con MultiQC. Histograma de las puntuaciones medias de calidad de las lecturas (A, E). Contenido promedio de GC (B, F). Gráfico del contenido de adaptadores (C, G). Gráfica del estado general de las muestras en los diferentes módulos evaluados por FASTQC (D, H).

En la Fig. 26 se observa el resultado del proceso de evaluación de los archivos FASTQC, las imágenes de la parte superior son los datos crudos que se descargaron del proyecto PRJNA191127, como se puede apreciar en el primer panel la mayoría de las secuencias no presentan buena calidad al encontrarse en el rango de puntuación inferior a 30 (fracción amarilla de la gráfica). Todas las secuencias se rigen bajo una distribución normal respecto al contenido de GC. El contenido de los adaptadores presentó un porcentaje superior al 5% en la mayoría de las secuencias, haciéndolas inadecuadas.

Para mejorar la calidad de estas secuencias, se realizó el proceso de trimming y se obtuvieron los resultados presentados en el panel inferior de la imagen, se aprecia que la calidad de las muestras mejora en todas y ahora cumplen con la puntación de calidad superiores a 30. El contenido de GC se modifica para algunos archivos de lecturas pareadas, a pesar de esta advertencia es posible continuar el proceso de análisis debido a que las secuencias ahora indican un contenido de adaptadores inferior a 0.1%, es decir que son muestras limpias en comparación a los datos crudos.

7.2.2.- Resultados obtenidos en SigProfiler de casos índice

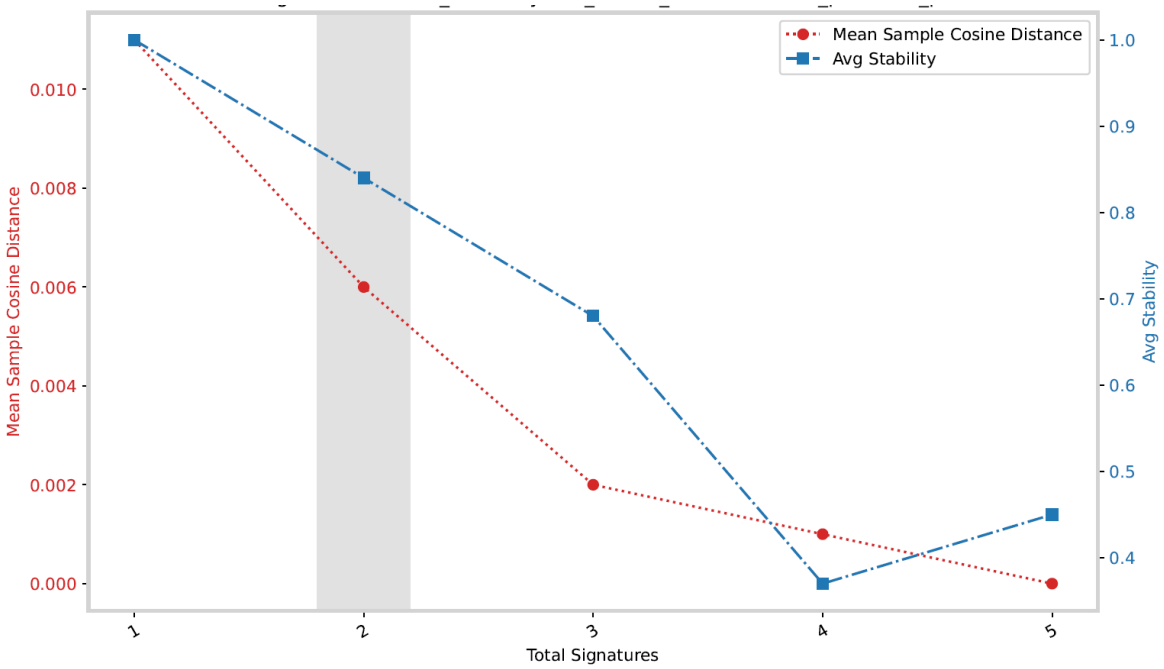


Fig. 27 Selección de firmas en casos índice

Se observa la recomendación del número de firmas adecuadas para el análisis de las muestras del proyecto PRJNA191127. El grafico indica que después de evaluar los diferentes parámetros, dos firmas son las que se ajustan a estas muestras.

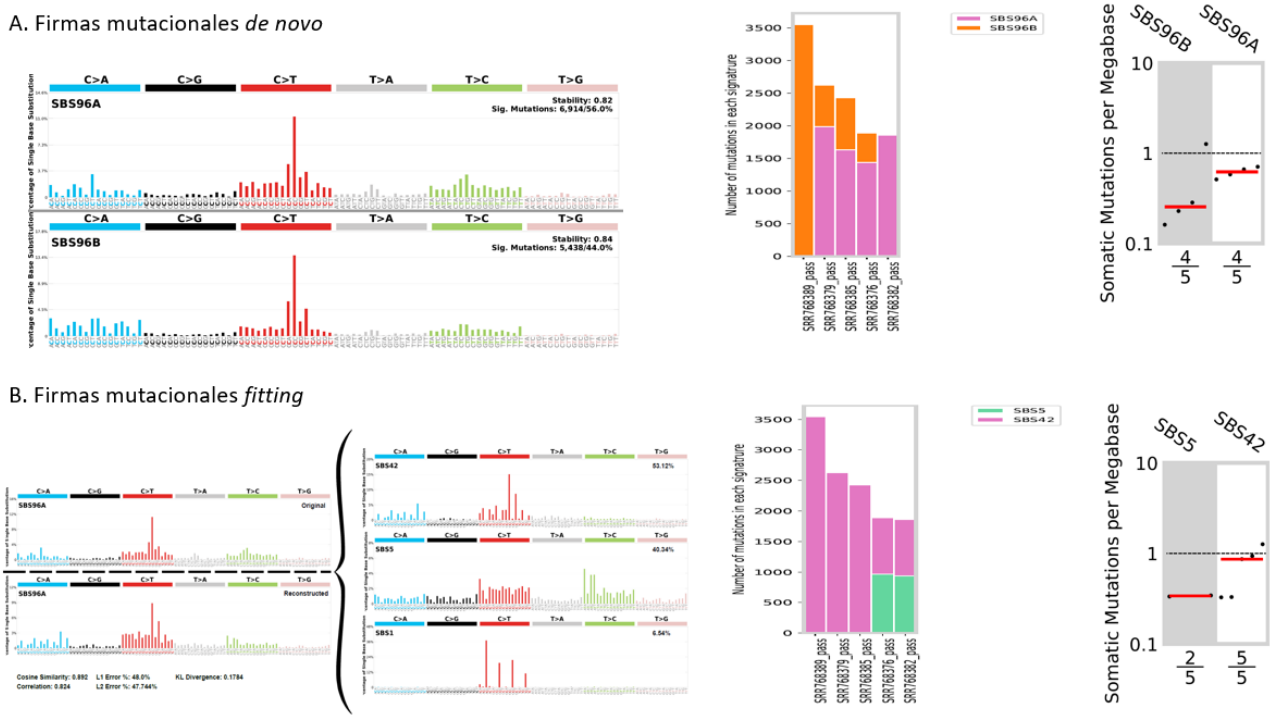


Fig. 28 Firmas mutacionales obtenidas con SigProfiler de los pacientes FA

En la figura Fig. 28 se observan los resultados del análisis de firmas mutacionales generado por *SigProfiler*. En el panel superior (A) se presentan los principales datos obtenidos al extraer una firma *de novo* para las muestras de los casos índice Fanconi, en la primera figura de izquierda a derecha se observan las dos posibles firmas que explican el proceso mutacional de las muestras, ambas presentan en mayor proporción mutaciones C>T; posteriormente, en el panel central el número de mutaciones que explican cada firma en cada una de las muestras (SBS96A en rosa y SBS96B en naranja) y; en el último panel la media de mutaciones somáticas por megabase y la proporción en que se encuentra cada una de las firmas propuestas en las diferentes muestras. En el panel inferior (B) se presentan los principales datos obtenidos al extraer una firma de tipo *fitting*, en la primera figura de izquierda a derecha se observa las dos firmas extraídas *de novo* junto a su reconstrucción según las firmas del catálogo de COSMIC, las cuáles son en este caso la firma SBS42 y SBS5; posteriormente, en el panel central se presenta el número de mutaciones de cada una de estas firmas en cada una de las muestras (SBS42 en rosa y SBS5 en verde) y; en el último panel la media de mutaciones por megabase correspondiente a cada firma del catálogo y el número de muestras en que se encuentran en las diferentes muestras.

7.2.3.- Resultados obtenidos en SigProfiler de familiares de pacientes AF del proyecto PRJNA191127

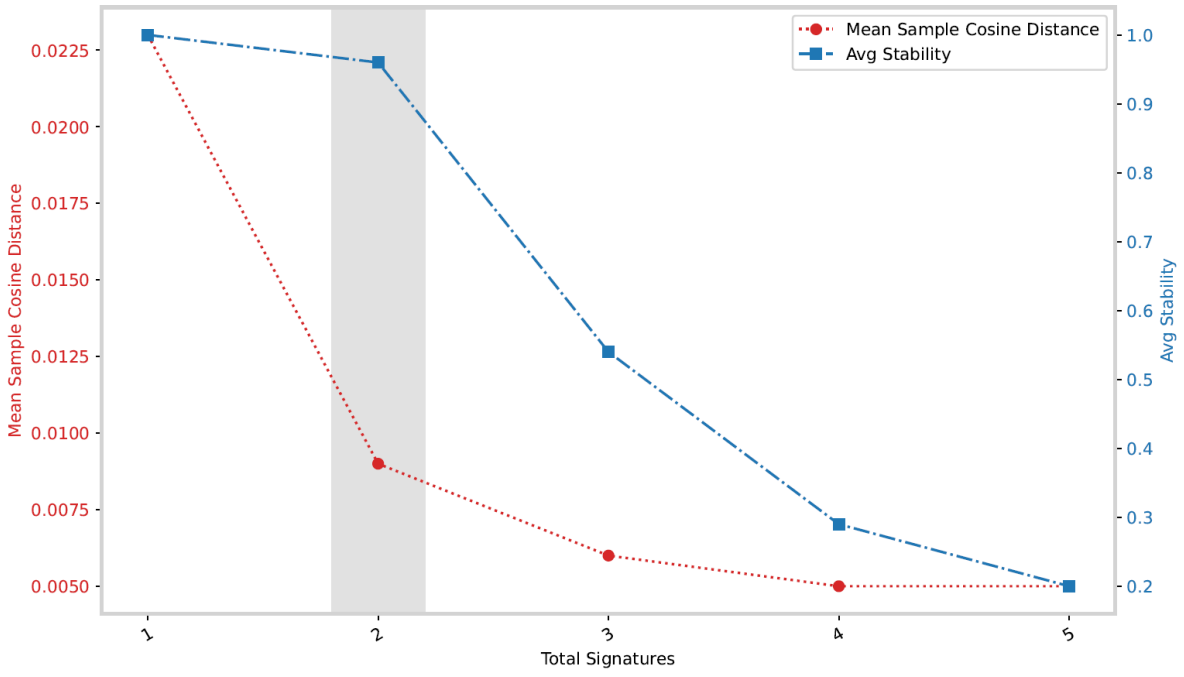


Fig. 29 Selección de firmas en familiares

Se observa la recomendación del número de firmas adecuadas para el análisis de las muestras del proyecto PRJNA191127 de los familiares. El gráfico indica que después de evaluar los diferentes parámetros, dos firmas son las que se ajustan a las muestras del presente trabajo.

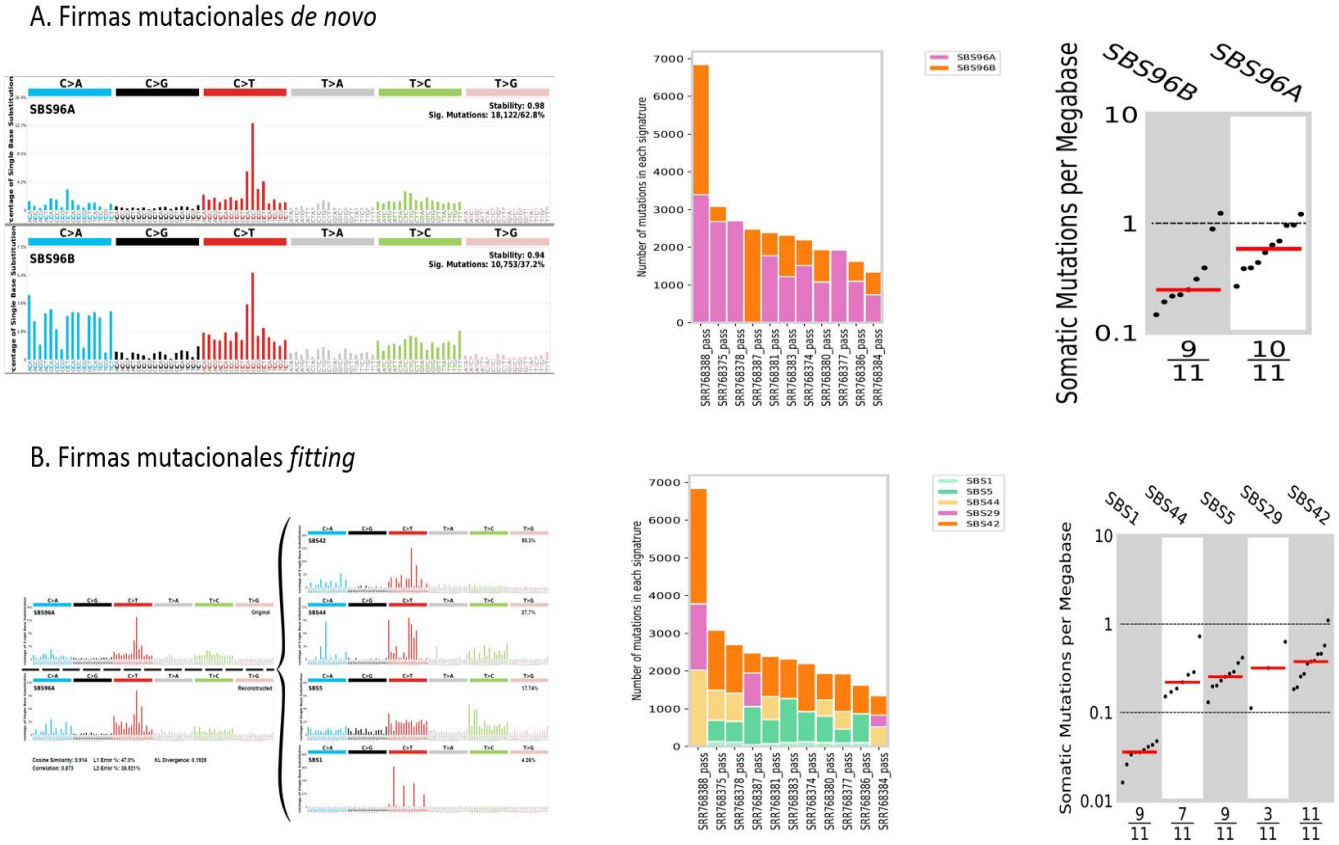


Fig. 30 Firmas mutacionales obtenidas con SigProfiler de familiares de pacientes AF del proyecto PRJNA191127.

En la figura Fig**.** 30 se observan los resultados del análisis de firmas mutacionales generado por *SigProfiler*. En el panel superior (A) se presentan los principales datos obtenidos al extraer una firma *de novo* para las muestras de los casos índice Fanconi, en la primera figura de izquierda a derecha se observan las dos posibles firmas que explican el proceso mutacional de las muestras, ambas presentan en mayor proporción mutaciones C>T; posteriormente, en el panel central el número de mutaciones que explican cada firma en cada una de las muestras (SBS96A en rosa y SBS96B en naranja) y; en el último panel la media de mutaciones somáticas por megabase y el número de muestras en que se encuentra cada una de las firmas propuestas en las diferentes muestras. En el panel inferior (B) se presentan los principales datos obtenidos al extraer una firma de tipo *fitting*, en la primera figura de izquierda a derecha se observa las dos firmas extraídas de *novo* junto a su reconstrucción según las firmas del catálogo de COSMIC, las cuáles son en este caso la firma SBS42, SBS44, SBS29, SBS5 y SBS1; posteriormente, en el panel central se presenta el número de mutaciones de cada una de estas firmas en cada una de las muestras (SBS42 en naranja, SBS44 en amarillo, SBS29 en rosa y SBS5 en verde) y; en el último panel la media de mutaciones por megabase correspondiente a cada firma del catálogo y el número de muestras en que se encuentran en las diferentes muestras.

7.2.4.- Resultados de deconstructSigs de casos de pacientes AF del proyecto PRJNA191127

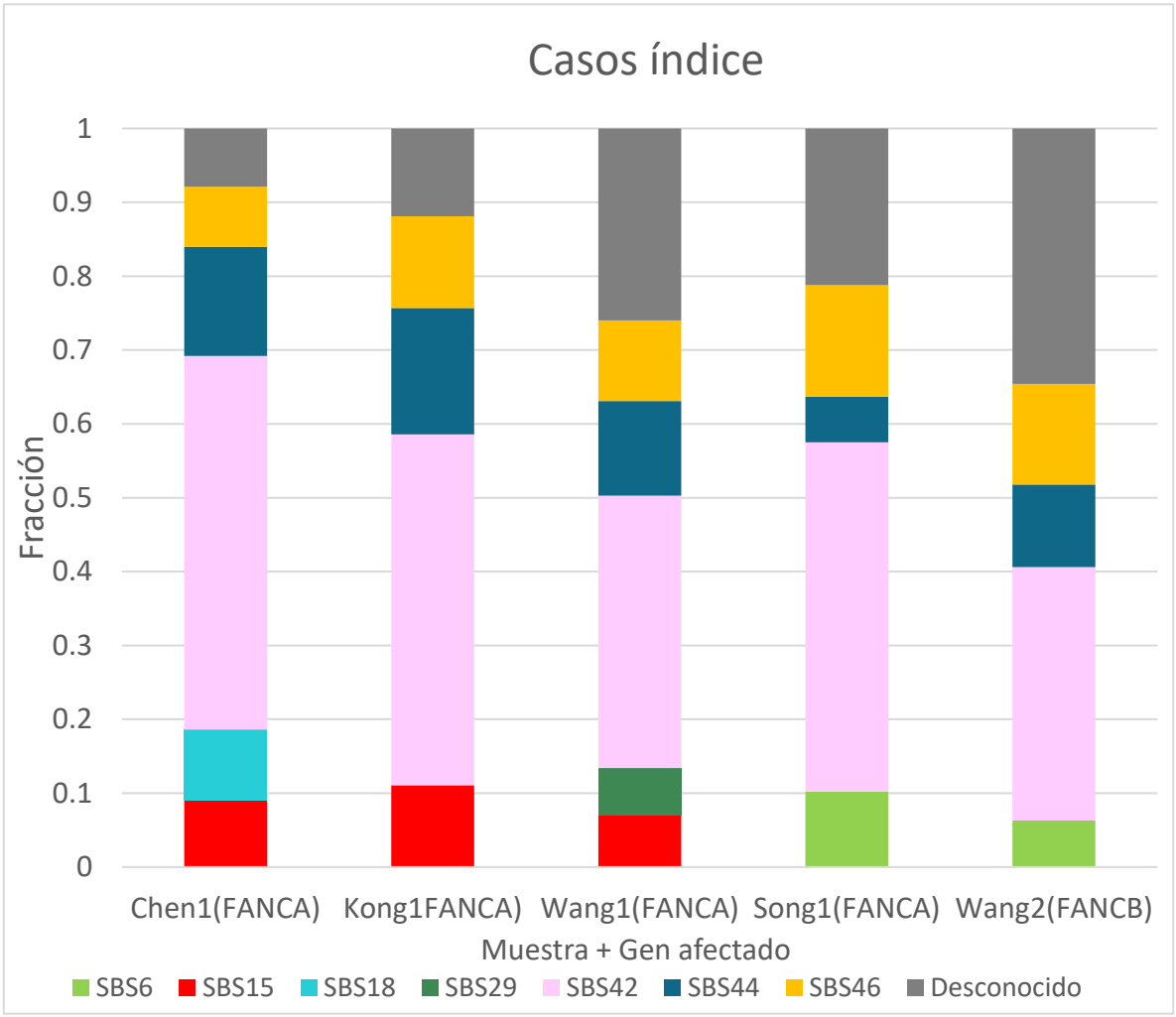


Fig. 31 Resultados obtenidos de los pacientes FA en deconstructSigs. Cada color representa una firma diferente como se observa en la leyenda. En el eje Y se presenta la fracción en que se encuentra una firma y en el eje X se observan las muestras por identificador del individuo y entre paréntesis el subtipo de Fanconi que se trate.

En la Fig. 31 se muestra los resultados para cada paciente Fanconi identificado como caso índice, se puede observar que todas las muestras tienen en común la firma SBS42, SBS44 y la SBS46 mientras que el resto de las firmas varía ligeramente. Se puede observar en los individuos FANCA poseen en común a la firma SBS15 mientras los dos restantes (FANCA y FANCB) poseen la SBS6. Sólo el individuo identificado como Wang1 posee la firma SBS29. Únicamente el individuo identificado como Chen1 posee la firma SBS18. El resto de las mutaciones que no se pudieron asociar a firmas específicas están representadas como desconocidas.

7.2.5.- Resultados de deconstructSigs de familiares de los pacientes AF del proyecto PRJNA191127

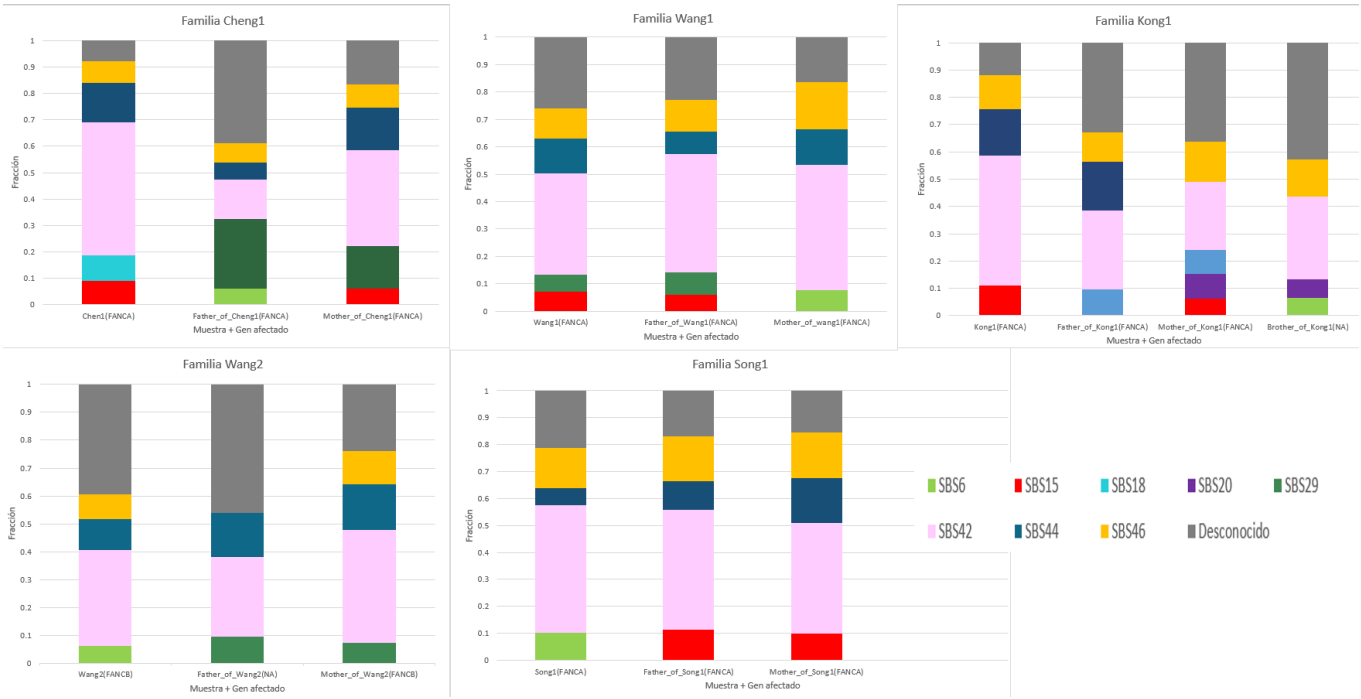


Fig. 32 Resultados obtenidos de familiares de pacientes AF del proyecto PRJNA191127 en deconstructSigs.
Cada gráfica posee el encabezado de la familia que representa. Cada color indica una firma diferente como se observa en la leyenda. En el eje Y se presenta la fracción en que se encuentra una firma y en el eje X se observan las muestras por identificador del individuo.

En la Fig. 32 se muestra los resultados de las diferentes muestras agrupadas por familia, se puede observar que todas las muestras tienen en común la firma SBS42 mientras que el resto de las firmas varía ligeramente. La mayoría de los pacientes presenta la firma 44 con excepción de la madre y hermano de Kong1. La mayoría de los pacientes presenta la firma 46 con excepción del padre de Wang2. La firma SBS15 está presente en Song1 y su madre; Wang1 y su padre; Kong1 y su madre y; ambos padres de Song1. La firma SBS6 está presente en el padre de Chen1; la madre de Wang1; el hermano de Kong1; en Wang2 y en Song1. La firma SBS18 únicamente la presenta Chen1. La firma SBS20 está presente en la madre y el hermano de Kong1. La firma SBS29 está presente en el padre y madre de Chen1; en el padre y madre de Kong1; en Wang1 y su padre y; en ambos padres de Wang2. El resto de las mutaciones que no se pudieron asociar a firmas específicas está representado como desconocido.

8.- Discusión

El estudio de firmas mutacionales se ha convertido en un estándar de los estudios genómicos porque generalmente permite un panorama más amplio de las mutaciones que ocurren a lo largo del genoma en comparación a estudios que se enfocan en pocos *drivers*, que proporcionan una ventaja selectiva de crecimiento, y por lo tanto promueven el desarrollo del cáncer (Alexandrov, Nik-Zainal, Wedge, Aparicio, *et al*., 2013; Koh *et al*., 2021). Las firmas mutacionales han sido estudiadas exclusivamente en cáncer, en principio porque el primer estudio fue realizado en 21 muestras de cáncer de mama (Nik-Zainal *et al*., 2012), además con el conocimiento de las limitaciones en los estudios de mutaciones *driver*, se optó por estudiar un panorama más amplio incluyendo mutaciones *passengers*, como lo son la mayoría de las mutaciones en cáncer. El estudio de mutaciones *passengers* estaba rodeado por un paradigma que las catalogaba como "eventos neutros", y se consideraba a las mutaciones *drivers* como las únicas relevantes, a las cuales los modelos matemáticos estiman que se requieren de cinco a ocho para el desarrollo del cáncer, pero que son superadas en número por mutaciones *passenger* (Pon & Marra, 2015). Posteriormente, dicho paradigma comenzó a cambiar debido a hallazgos de algunos estudios que sugieren que el acumulo de mutaciones *passengers* tiene cierto efecto "*driver*" o simplemente aportan a la progresión del cáncer (Kumar *et al*., 2020; Salvadores *et al*., 2019).

La evaluación de firmas mutacionales requiere de una serie de buenas prácticas en el manejo de datos de secuenciación. Esto solo es posible a partir del desarrollo de un pipeline (o flujo de trabajo) bioinformático con el cual se determina la calidad de lecturas, calibración de la anotación, generación de resultados y su posterior evaluación e interpretación. En el presente proyecto se realizó un pipeline bioinformático para la evaluación de firmas mutacionales en muestras de pacientes con anemia de Fanconi que se caracterizan por tener deficiencia en la vía de reparación FA/BRCA la cual se encarga del mantenimiento y estabilidad genómica del ADN, por lo que evaluar las firmas en estos pacientes podría a) tener un mejor entendimiento de los procesos mutacionales que se desarrollan en el transcurso de esta enfermedad, b) determinar el tipo y la cantidad de eventos mutacionales que presentan los pacientes AF, c) investigar si presentan firmas de envejecimiento y d) ser precursor del estudio de firmas en otras enfermedades no-cáncer.

8.1.- Validación del pipeline con muestras de cáncer de mama del proyecto PRJNA729775.

Para evaluar el correcto funcionamiento del pipeline, se realizó la validación con muestras obtenidas del proyecto PRJNA729775. En la descripción de este proyecto se indica que se buscaron deficiencias de la recombinación homóloga en subtipos de cáncer de mama de la población taiwanesa. La primera consideración que se debe hacer es que este análisis se realizó con 232 secuencias de WES provenientes de tejido de cáncer de mama y tejido normal adyacente con el cual los autores diseñaron su propio panel de normales (Wu *et al*., 2021).

El motivo de selección de este set de datos está asociado al estudio de firmas mutacionales en cáncer de mama, como modelo pionero y a la gran cantidad de información relacionada a este.

De los resultados obtenidos en *SigProfiler* para este set de datos, se observó en la solución *de novo* dos firmas propuestas SBS96A y SBS96B, debido a la estabilidad que presentan y la actividad de cada una en las diferentes muestras (Fig. 24A), la firma SBS96A parece ser la que mejor explica el set de muestras debido a que se encuentra presente en todas las muestras y posee una estabilidad de 0.99 en comparación con la SBS96B con estabilidad de 0.8. Esta firma se descompone en las firmas SBS45, SBS29 y SBS5. Se debe señalar que en el artículo del proyecto se plantea que gracias a la elaboración de un panel de normales propio, se eliminaron variantes falsas positivas, principalmente aquellas que fueron consecuencia del uso de 8-oxoguanina (Wu *et al*., 2021), usada en la secuenciación del genoma (Costello *et al*., 2013). Esto explica que la firma SBS45, validada como artefacto, predomine en el análisis. Por otra parte, el artículo señala que las firmas *de novo* extraídas se asemejan más a la firma SBS1 (asociada al envejecimiento) y la SBS2 (actividad de APOBEC). Sin embargo, en este trabajo se extrajeron la firma SBS5 (asociada a envejecimiento) y SBS29 (asociada a masticar tabaco). Ambas firmas se han encontrado como firmas secundarias presentes en cáncer de endometrio, mismo tipo de cáncer que ha sido asociado a defectos en BRCA al igual que el cáncer de mama (Ashley *et al*., 2019).

No obstante, al realizar la revisión de presencia de firmas en diferentes tipos de cáncer propuesta en (Alexandrov *et al*., 2020), se puede observar que la firma SBS5 se encuentra con mayor frecuencia y en mayor proporción en cáncer de mama, que la firma SBS3. En este punto se debe mencionar que las firmas no tienen la misma probabilidad de ser extraídas, esto se debe a que las firmas que tienen patrones más marcados son más sencillas de determinar por el programa, por lo cual existe cierto sesgo para extraer las firmas SBS3, SBS5 o SBS8, por ejemplo. Adicionalmente, las firmas SBS3 y SBS5 son aquellas que presentan menor correlación estadísticamente significativa entre los programas *deconstructSigs* y *SigProfiler* (Koh *et al*., 2021; Rosenthal *et al*., 2016). Desde otra perspectiva, la firma SBS3 se ha reportado en 7% de carcinosarcoma uterino como firma predominante primaria y 21% como firma secundaria, aunque ninguno de los casos estaba asociado a defectos en genes asociados a recombinación homóloga. El cáncer de ovario seroso de alto grado y el cáncer de mama basal o triple negativo son fuertemente asociados a defectos en la vía de reparación por recombinación homóloga, sin embargo, la fracción de cáncer de mama atribuida a mutaciones de *BRCA1* y *BRCA2* es menor al 5%, de manera que no se esperaría encontrar la firma SBS3 con alta frecuencia en muestras de cáncer de mama en general, en cambio sí se espera una alta proporción en los casos de cáncer de mama asociados a mutaciones en *BRCA1* y *BRCA2*. En el caso de que se requiera detectar con mayor sensibilidad la deficiencia de genes asociados a reparación por recombinación homóloga se pueden utilizar *softwares* complementarios como *HRDetect*, *CHORD* y *SigMA* que limitan la búsqueda de firmas mutacionales a aquellas que se asocian a estos genes

además de proporcionar guías de tratamiento (Davis-Turak *et al*., 2017; Koh *et al*., 2021; Van Hoeck *et al.*, 2019).

Uno de los hallazgos más interesantes es que los resultados obtenidos de la muestra de cáncer de mama, en pacientes de 47 años, son similares entre sí, en el sentido de la proporcionalidad de las firmas extraídas. Sin embargo, la firma SBS5 no logró ser extraída, como se pronosticaba, debido a la dificultad del programa para extraer firmas de comportamiento plano, y como se ha observado en otros estudios la omisión de la firma SBS3 por ejemplo hace que sea reasignada como firma SBS5 u SBS8 pero obteniendo una reconstrucción del error (SSE) diferente (Maura *et al*., 2019; Rosenthal *et al.*, 2016), esto quiere decir que al forzar al programa a extraer firmas de comportamiento plano la asignación puede resultar ambigua. Por otra parte, la firma SBS29 (asociada al hábito de masticar tabaco) solo se observó en una muestra, la correspondiente a tejido de cáncer de mama en una paciente de 32 años.

Aparecen 3 firmas de manera constante, la firma SBS94 en mayor proporción para 2 muestras del set de datos, esta firma no tiene etiologías propuestas pues fueron recientemente identificadas (Islam *et al*., 2021). Sin embargo, una característica en común de las firmas SBS94 y SBS29 son el sesgo transcripcional que se encuentra en mutaciones C>A, mismo tipo de mutaciones que predominan en ambas firmas y ambas han sido encontradas en cáncer colorrectal, lo cual podría indicar una similitud conveniente para su extracción en ambos programas. En este proyecto se trabajó con secuencias de exoma completo que permite observar más detalladamente el sesgo transcripcional y el sesgo a su vez, es un reflejo de las vías de reparación asociadas a las transcripción que están implicadas en estos eventos mutacionales, por ello también es la importancia de añadir el contexto de las secuencias de 5’ a 3’, en términos generales se puede usar datos de exoma sin ningún problema (Alexandrov, Nik-Zainal, Wedge, Campbell, *et al*., 2013; Koh *et al*., 2021).

Por otra parte, ocurre algo similar entre la firma SBS18 que está asociada a exposición de especies reactivas de oxígeno, y la firma SBS24 que está asociada a exposición a aflatoxina. Ambas firmas presentan nuevamente una mayor proporción de mutaciones del tipo C>A, siendo congruente con una de las primeras investigaciones, la realizada en 21 genomas de cáncer de mama, dónde se planteó la relación entre la mutagénesis y la transcripción al encontrarse un sesgo transcripcional de cadena, para las mutaciones C>A/G>T en la mayoría de las muestras de cáncer de mama, siendo predominantes las mutaciones C>A características de este cáncer (Nik-Zainal *et al*., 2012).

Al comparar los resultados obtenidos en *SigProfiler* con *deconstructSigs* (Fig. 25), se pueden observar algunas diferencias, principalmente la extracción de más firmas con *deconstructSigs*, en comparación con *SigProfiler*, esto se ha atribuido a que existen beneficios de analizar muestras de forma individual en *deconstructSigs* y que es posible

detectar procesos mutacionales activos en un pequeño número de muestras (Rosenthal et al., 2016).

Para esta primera parte de la validación del pipeline se da por sentado que este es funcional, si bien las muestras de cáncer de mama han estado fuertemente asociadas a cierto tipo de firmas se debe entender que una de las limitaciones del estudio de las firmas mutacionales es debido a que en principio se tratan de herramientas matemáticas cuyos resultados deben de ser evaluados con respecto de los conocimientos biológicos, es decir que la viabilidad de que las firmas estén presentes no solamente se basa en ser soluciones de la factorización matricial no negativa sino también que tengan un sentido biológico. La viabilidad biológica de la firma SBS94, SBS18 y SBS24 solo podría descartarse realizando estudios enfocados a detectar la presencia de este tipo de mutaciones.

8.2.- Aplicación del pipeline en muestras de pacientes AF del proyecto PRJNA191127

En este set de datos, se logró obtener con el programa *SigProfiler* la solución *de novo* de dos firmas propuestas SBS96A y SBS96B, debido a la estabilidad que presentan y la actividad de cada una en las diferentes muestras, ambas poseen cierta predominancia por mutaciones de tipo transición de citosina a adenina C>A, esta predominancia se observa en las firmas extraídas por el programa (Fig. 28A), la firma SBS96A parece ser la que mejor explica el set de muestras. Esta firma según la solución *fitting* de *SigProfiler* (Fig. 28B) está compuesta por la firma SBS42, SBS5 y SBS1; sin embargo, la firma que se observa en mayor proporción y de manera constante en la gráfica de actividad es la firma SBS42. Esta firma está asociada a exposición por alcanos y fue observada por primera vez en trabajadores de una imprenta japonesa que estaban expuestos a cierto tipo de compuestos que favorecían estas mutaciones, además está altamente presente en adenocarcinoma biliar y en principio podría parecer una firma con cierta asociación a la etnia puesto que este cáncer tiene incidencia de 2.1 por 100,000 en la población occidental mientras en la oriental es de más del doble (5.2 por cada 100,000) (Mimaki *et al*., 2016); por inferencia de los nombres que llevan las muestras, probablemente se trate de pacientes orientales. Por otra parte, las firmas SBS5 y SBS1 son firmas asociadas al envejecimiento (Alexandrov *et al*., 2015).

En el caso de *deconstructSigs* se observa una gran variedad de firmas, siendo la SBS42, la única que comparten ambos programas, siendo además la que prevalece entre los pacientes. La firma SBS44 se relaciona con defectos en reparación *mismatch* y la SBS46 (igual que la SBS45), está validada como firma artefacto en secuenciación de exoma completo (Fig. 31).

La firma SBS42, asociada a exposición de haloalcanos, es la de mayor relevancia en el presente proyecto. Un ejemplo de haloalcano bien documentado es el tetracloruro de carbono (CCl_4), el cual ha sido utilizado para inducir modelos de hepatotoxicidad y estudiar la

aplicación hepatoprotectora de diversos agentes (Izzo *et al*., 2021). El CCl_4 posee dos mecanismos principales: la unión de enlaces covalentes y la peroxidación lipídica, ambos por consecuencia del metabolismo por el citocromo CYP2E1 que dan lugar a un radical libre CCl_3* (Weber *et al*., 2003). Hay que recordar que, de manera general, la peroxidación lipídica forma especies reactivas de carbonilo como el malondialdehído (MDA). El MDA puede reaccionar con la guanina, la adenina y la citosina para producir ICLs. Otros productos de la peroxidación lipídica son los aldehídos insaturados como la acroleína y el crotonaldehído que pueden reaccionar con las bases nitrogenadas, se conjugan, se ciclan y forman monoaductos (Housh *et al*., 2021; Lopez-Martinez *et al*., 2016). Los aldehídos en anillo abierto de un sitio abásico 3' son capaces de formar ICLs al reaccionar con el grupo amino exocíclico de los residuos de citosina (Clauson *et al*., 2013). Teniendo estas consideraciones, se puede decir que los efectos de haloalcanos pueden imitar procesos metabólicos endógenos a los que son hipersensibles los pacientes con AF, debido a su defecto bioquímico basal.

Los ICLs son reparados por la vía FA/BRCA, que durante el desenganche por la función endonucleolítica de FANCP/SLX4 y XPF/FANCQ, genera tres lesiones intermediarias, una de ellas son los aductos. Estos suelen ser reparados por NER a nivel global (GG-NER) o NER acoplada a la transcripción (TC-NER). Esta última es una subvariante que repara los daños en el ADN que se producen durante la transcripción, utilizando la Pol II elongada para escanear la cadena transcrita y "reconocer" el daño que bloquea la transcripción, posteriormente se reclutan las endonucleasas de reparación ERCC1-XPF y XPG las cuales escinden la hebra dañada en sentido 5' y 3' con respecto al daño, liberando un fragmento de ADN monocatenario y seguido las ADN polimerasas rellenan el hueco (Duan *et al*., 2021). Esto es relevante debido a que recientemente se ha propuesto a ERCC1 y XPF como genes asociados a AF debido a defectos encontrados en estos pacientes, además de encontrar interacción directa entre la vía FA/BRCA y NER (García-de Teresa et al., 2020; Juárez-Figueroa et al., 2018; Kashiyama et al., 2013; Mouw & D'Andrea, 2014; Rodríguez & D'Andrea, 2017).

En lo que concierne a firmas mutacionales, TC-NER posee un papel importante debido a que aquellos daños como la exposición a luz UV, que genera dímeros de pirimidina, o consumo de tabaco, que producen aductos voluminosos de guanina, son representados por firmas mutacionales con mayor cantidad de sesgo transcripcional, es decir, no se encuentra la misma cantidad de mutaciones de un tipo y contexto específico en la cadena transcrita con respecto a la no transcrita. Este sesgo es consecuencia directa del tipo de reparación para estos daños guiados por TC-NER (Alexandrov, Nik-Zainal, Wedge, Aparicio, *et al*., 2013). La firma SBS42 es una firma que presenta sesgo transcripcional en C>A y en C>T posee más mutaciones G que C en las cadenas no transcritas de los genes, es coherente con el daño a la guanina y la reparación mediante la TC-NER. Por ello, probablemente más que una

exposición directa a una sustancia, la firma SBS42 en el estudio de este set de datos es indicador biológico de una elevada presencia de aductos (Keppler *et al*., 2000).

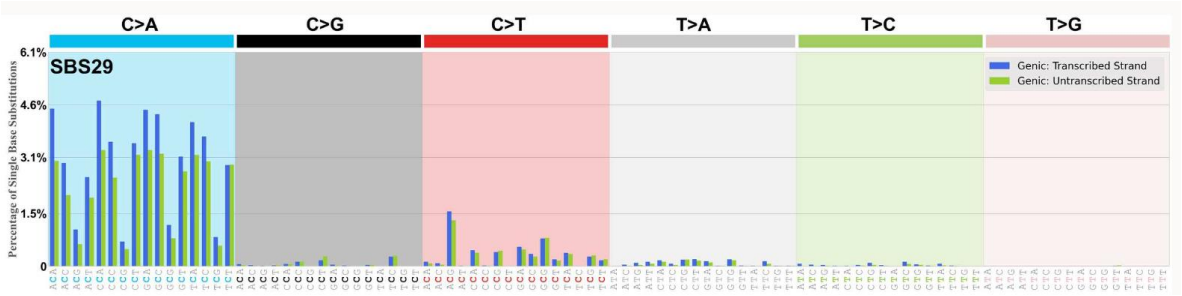


Fig. 33 Representación del sesgo transcripcional en la firma SBS29
En el caso de algunos contextos trinucleotídicos en C>A se puede observar aumento de mutaciones en la hebra transcrita con respecto de la no transcrita.

En el caso de SBS1 y SBS5, sólo fueron extraídas en el programa *SigProfiler*. En la gráfica de actividades por firmas de catálogo, de los pacientes AF se puede observar que en 2 de los 5 pacientes la proporción de SBS5 es de aproximadamente el 50% del total de mutaciones, mientras que en el caso de los familiares de los pacientes AF la proporción de SBS5 fue heterogénea y se presentó en 9 de los 11 individuos. Con esta información se puede cotejar las diferencias que existen entre pacientes no Fanconi y AF con relación al envejecimiento acelerado, debido a que SBS5 se ha observado en alta correlación con la edad en diferentes muestras es de tumores, de manera que es una firma que se espera observar generalmente en todas las personas con mayor o menor proporción acorde a la edad. Sin embargo, en pacientes pediátricos cómo los pacientes AF no se esperaría observar una cantidad tan alta como la que pudiera presentar un adulto sano. A pesar de que los familiares de estos pacientes AF presentan al menos una mutación en un gen Fanconi (con excepción del padre de Wang2 y el del hermano de Kong1, que no presentan ninguna mutación asociada a Fanconi), ninguno sobrepasa el 20% de mutaciones asociadas a SBS5. Por lo tanto, la cantidad de SBS5 presentada en los diferentes individuos además de ser indicador de envejecimiento acelerado se puede asociar también a peores pronósticos de cáncer o mayor daño acumulado (Alexandrov *et al*., 2015; Chong *et al*., 2021).

Respecto a la firma SBS29 asociada a masticar tabaco, el cual posee compuestos aromáticos policíclicos, acarrea como consecuencia la formación de aductos de guanina y su reparación se realiza por NER acoplada a la transcripción. La firma SBS29, asociada a exposición a tabaco masticado, es frecuente en cáncer de células escamosas de la boca, una de las neoplasias con mayor incidencia en AF (Boot *et al*., 2018; Datta & Brosh, 2019). La presencia de estas firmas en los pacientes AF, no necesariamente significa que los pacientes estén expuestos a haloalcanos *per se* o bien mastiquen tabaco, sino que presentan cierto tipo de daño cuyas consecuencias son traducidas en el mismo tipo de firma.

Por otra parte, se observa en *deconstructSigs,* como una constante, otras firmas como las 6, 15, 20 y 44 que están todas asociadas a defectos en reparación *mismatch* (MMR) y que suelen extraerse de forma conjunta. Actualmente, no es posible hacer una distinción entre una firma u otra en relación con su participación en los defectos de MMR, por lo cual ambas tienen la misma explicación biológica. La vía FA/BRCA está relacionada con la actividad de la vía de reparación de *mismatch*; existen estudios que han demostrado la interacción directa entre proteínas de ambas vías. Por ejemplo, MMR activa FA/BRCA por la interacción entre a) MLH1 y FANCD2-FAN1, BRCA1, y FANCJ y b) MSH2 y SLX4/FANCP, cuando no están coordinadas las rutas se desencadena la toxicidad de MSH2 (proteína MMR), la pérdida de MMR puede contribuir al desarrollo de tumores en AF, además, la deficiencia de MSH2 suprime la actividad anticáncer y pro-envejecimiento del acortamiento de telómeros. Por otra parte, la proteína MSH2, además de interactuar con la vía de reparación por recombinación homóloga, también posee interacción con la vía de NER, más específicamente con la proteína ERCC1 (Peng *et al*., 2014; Spies & Fishel, 2015; Williams *et al*., 2011).

8.3.- Limitaciones del estudio

Algunas de las limitantes del presente proyecto son:

1. No tener acceso completo al historial clínico de los pacientes.
2. La tendencia a presentar fenotipos hiper-mutadores, pues el hecho de generar varias mutaciones de manera descontrolada puede proveer una asignación errónea de firmas.
3. Los programas utilizados están mejor estandarizados para diferentes objetivos. *SigProfiler* se recomienda que se realice con al menos 200 muestras para una mejor extracción *de novo*, mientras el programa *deconstructSigs* se considera mejor estandarizado para usar un menor número de muestras (Rosenthal *et al*., 2016).

De manera general los resultados obtenidos representan un nuevo enfoque en el estudio de firmas mutacionales en enfermedades no-cáncer que pudiera promover el interés en investigaciones futuras para el entendimiento de los eventos mutacionales en exoma o genoma completos que consideren aquellas que son "*passenger*" y que estas, a su vez, puedan permitirnos plantearnos nuevas hipótesis sobre las posibles causas del desarrollo de cáncer en la enfermedad estudiada.

9.- Conclusiones

1. Se desarrolló un pipeline bioinformático basado en las buenas prácticas del preprocesamiento de datos.
2. Se propone considerar los resultados del presente trabajo obtenidos por extracción de tipo *fitting* en *deconstructSigs*, como la mejor predicción de las firmas de AF; el programa se recomienda para set pequeños.

3. En los pacientes con AF, se encontraron las firmas: SBS5 asociada a envejecimiento; SBS42 asociada a la exposición de haloalcanos; SBS44, SBS15 y SBS6 asociadas a defectos en reparación *mismatch*; SBS29 asociada a tabaco masticado; y SBS18 asociada a daño por especies reactivas de oxígeno.
4. No se identificó la presencia de las firmas SBS1 asociada a envejecimiento, ni la firma SBS3 relacionada a deficiencia en la reparación del ADN por recombinación homóloga, es posible que se deba a que las muestras sean de pacientes que, por su edad, aún no acumulan la cantidad de daño suficiente para ser detectadas.
5. Se encontró una firma *de novo*, con alto contenido en mutaciones de tipo C>T y sesgo transcripcional marcado, lo cual sugiere defectos en la reparación de ICL's y acumulación de aductos que son reparados por TC-NER. Se recomienda reproducir el pipeline con un set más grande de muestras de AF para validarla.

10.- Anexos

Anexo I. Linux

Al adentrarse en el mundo de la bioinformática es común encontrar, por ejemplo, proyectos con miles de muestras con datos de alto rendimiento que requieren una enorme cantidad de potencia de cálculo. Por tanto, el uso de plataformas de computación de alto rendimiento (HPC) es un requisito esencial, y normalmente, funcionan con sistemas Unix/Linux. El hecho de tener el mismo sistema evita problemas de reproducibilidad o de incompatibilidad de *software*, siendo una de sus fortalezas el alto grado de portabilidad, que permite a los usuarios registrar y compartir flujos de trabajo (*pipelines*) en repositorios privados y públicos (Ayyildiz & Piazza, 2019; Kulkarni *et al*., 2018).
Linux es una interfaz de texto para ordenador basada en Unix que permite controlar el ordenador usando comandos ingresados con un teclado en lugar de controlar las interfaces gráficas de usuario. Es un *software* libre por lo que puede ser utilizado, estudiado, modificado, copiado y redistribuido sin restricciones ("The Linux Command Line for Beginners," n.d.).

Los investigadores que no están familiarizados con la línea de comandos de Unix, generalmente se enfrentan a un aprendizaje forzado. El conocimiento de las herramientas bioinformáticas es el principal cuello de botella en el análisis de datos generados por NGS. Además, los programas utilizados para el procesamiento de archivos de secuenciación están diseñados para ejecutarse en un entorno Unix/Linux, como lo es Ubuntu el subsistema Linux que puede ser utilizado en Windows. Después de obtener los datos de la secuencia hay una variedad de pipelines para el análisis de NGS que utilizan herramientas de código abierto. Esto incluye, por ejemplo, el análisis de calidad de la secuencia (FASTQC), recorte y filtrado (Trimmomatic), y búsqueda de pequeñas inserciones, deleciones y SNVs (SAMtools, Mutect2, etc).

Anexo II. FASTQC

FASTQC es una aplicación ampliamente adoptada en el control de calidad de archivos FASTQ, ya que resume la calidad de la lectura por posición, informa al usuario del contenido de adaptadores en las secuencias, notifica las frecuencias de tetrámeros y muchos otros aspectos presentes en los datos de la secuencia en bruto (Brown *et al*., 2017).

El formato FASTQ fue inventado a principios de siglo en el *Wellcome Trust Sanger Institute* por Jim Mullikin. Este es el formato de salida de la secuenciación en Illumina. Dentro de este formato se aprecian cuatro líneas.

1. Una línea de título "@" que a menudo contiene sólo un identificador de registro.
2. Una segunda línea con la secuencia, la cual utiliza los mismos caracteres del formato FASTA.
3. La tercera línea contine un signo “+” que señala el final de las líneas de secuencia y el comienzo de la cadena de calidad.
4. Y en la última línea se observa la calidad expresada en caracteres imprimibles ASCII con la misma longitud que la cadena de secuencia.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

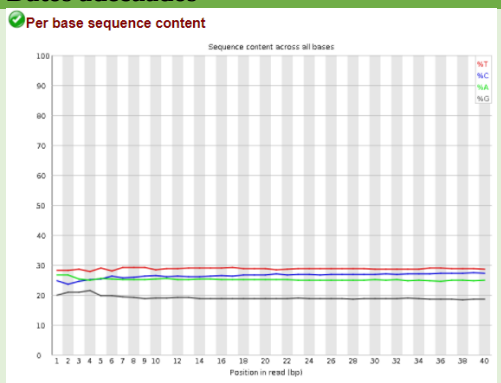
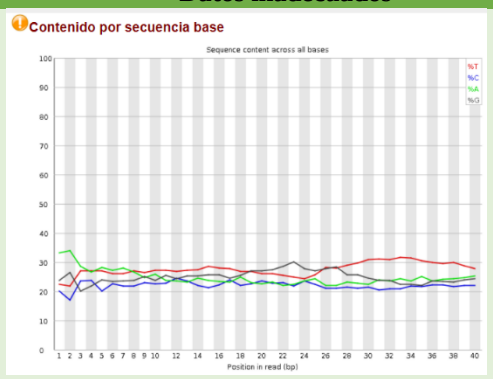
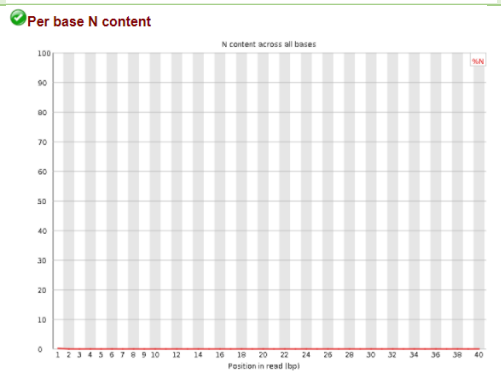
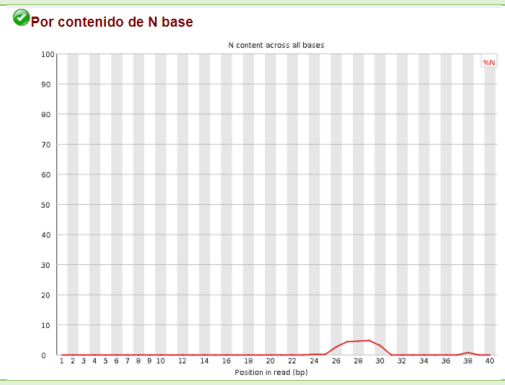
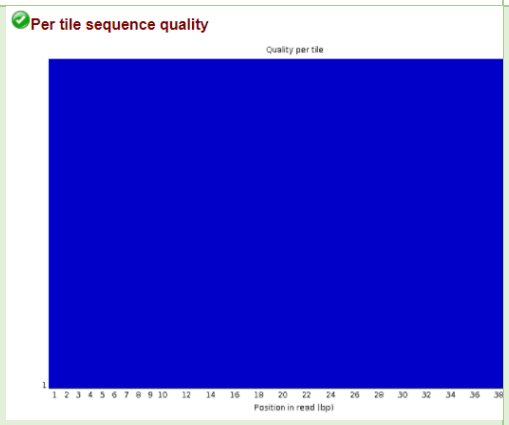
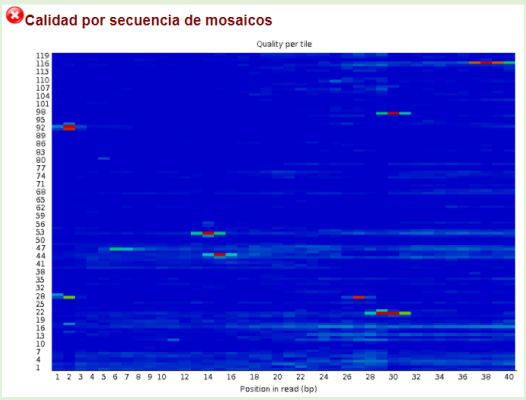
Fig. 34 Formato FASTQ
(Cock et al., 2010)

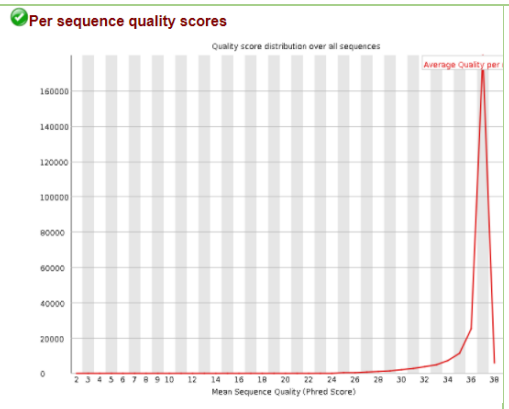
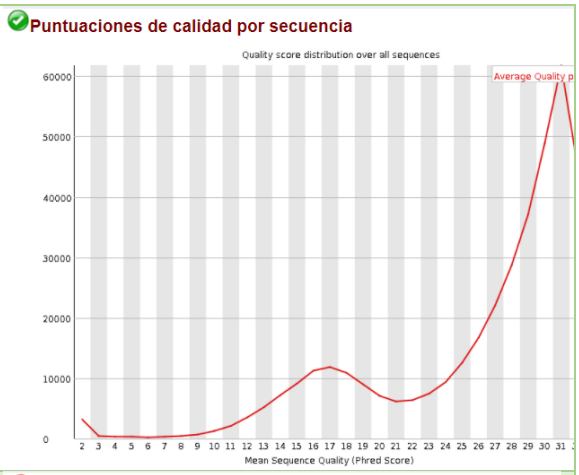
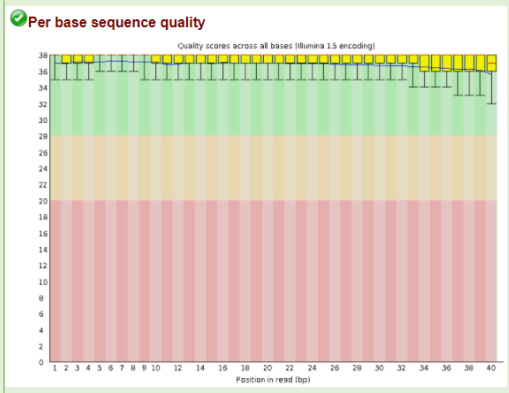
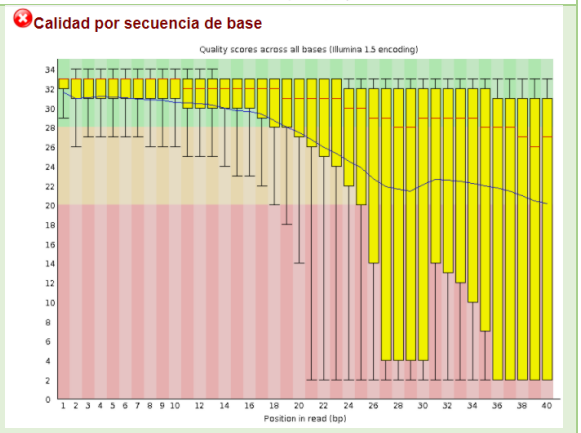
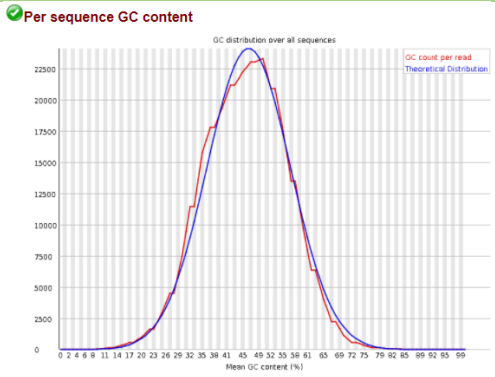
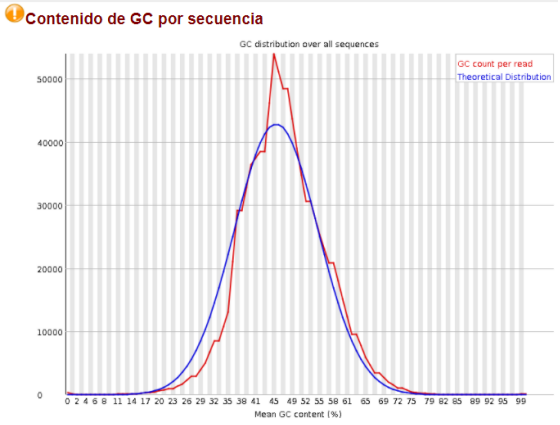
El *software* PHRED es el encargado de leer archivos planos de secuenciación de ADN, genera bases y asigna un valor de calidad a cada base generada en términos de la probabilidad de error estimada (Cock *et al*., 2010).

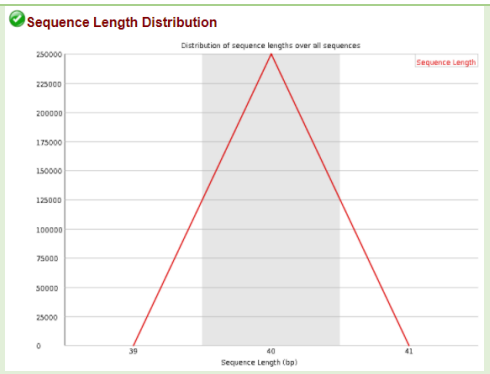
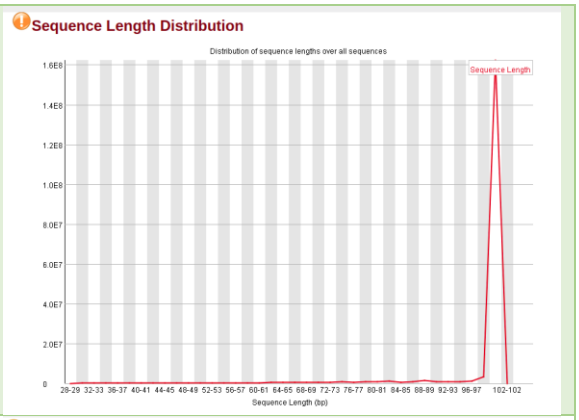
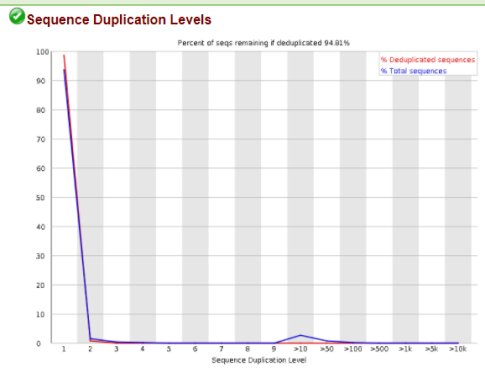
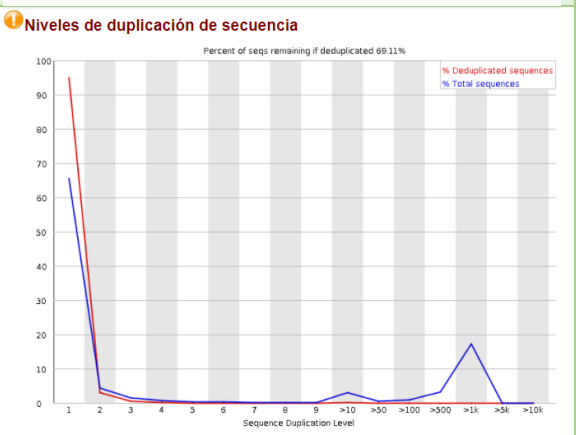
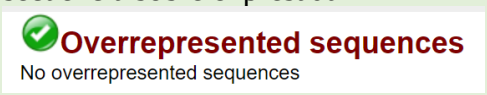
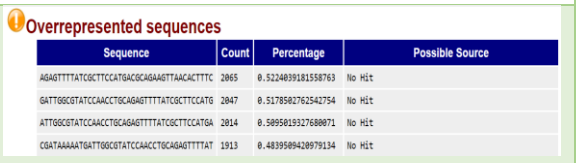
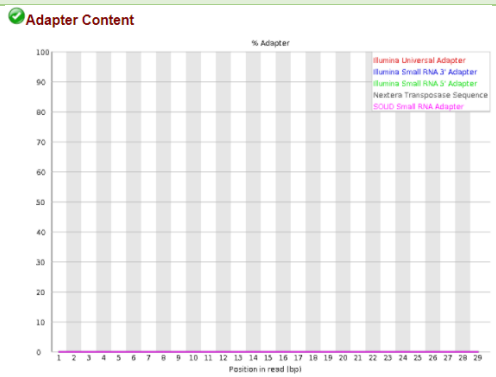
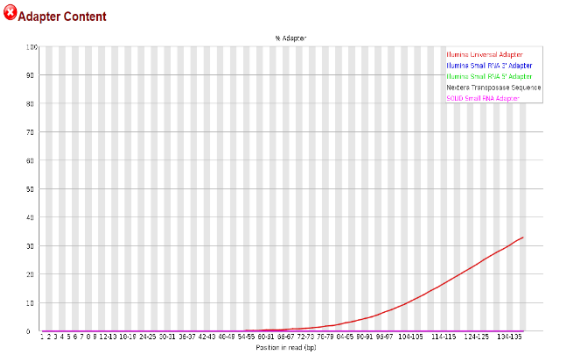
Los informes de análisis de FASTQC son el estándar para varias herramientas de control de calidad, basándose en su resultado como criterio para proceder con los pasos posteriores o, alternativamente, para filtrar, recortar o, en última instancia, descartar los datos (de Sena Brandine & Smith, 2019).

FASTQC presenta diez módulos de análisis que resumen el contenido de un archivo de secuenciación:

Tabla 2 Descripción de módulos de FASTQC

Módulo y descripción	Datos adecuados	Datos inadecuados
Calidad de la secuencia por base: Se debe observar una proporción constante entre el contenido de %A=%T y %G=%C. En una muestra aleatoria se espera que haya poca o ninguna diferencia entre las diferentes bases, por lo que las líneas de este gráfico deberían ser paralelas entre sí.		
Contenido de N por base: No se espera que pase de 0, de lo contrario señala problemas en la secuenciación para realizar el llamado correcto de una base.		
Calidad de la secuencia por mosaico: El gráfico muestra la desviación de la calidad media de cada mosaico. Los colores están en una escala de "frío a caliente", siendo los colores fríos las posiciones en las que la calidad estaba en o por encima de la media; por tanto, un módulo adecuado no debería mostrar colores cálidos.		

Por puntuaciones de calidad de la secuencia: La distribución de la calidad promedio de lectura debe ajustarse al rango superior del gráfico.		
Calidad de la secuencia por base: Es normal que la puntuación de calidad media comience siendo más baja en las primeras 5-7 bases y que luego aumente. Las lecturas con muy buena puntuación se encontrarán en la zona verde. • **La caja amarilla representa el rango de percentiles 25-75%** • **Las líneas negras (bigotes) representan los valores 10 y 90%** • **La línea azul representa la calidad media**		
Por contenido de GC de la secuencia: Se espera ver una distribución tipo normal. Si ve un sesgo de GC que cambia en diferentes bases, esto podría indicar una secuencia sobrerrepresentada que está contaminando su biblioteca.		

Distribución de la longitud de la secuencia: Cada uno de los fragmentos que se secuencia se le denomina "read" o lectura. Y estos varían de tamaño por diferentes factores del mismo proceso de secuenciación. Este módulo indica el tamaño en bases de dichas secuencias.		
Niveles de duplicación de secuencias: Se espera que casi el 100% de las secuencias sean únicas, es decir, sin duplicados.		
Secuencias sobrerrepresentadas: Ninguna secuencia debería presentarse mayor al 0.1% en WES o WGS. En caso de un panel, podría ser normal.	No debería aparecer ninguna secuencia sobre expresada 	
Contenido de adaptadores: Lo ideal es que no sean detectables, pero esto a menudo puede ocurrir. La presencia de estos puede dar lugar a análisis subóptimos.		

Cada uno de estos módulos debe ser analizado para determinar los ajustes a los parámetros de *trimming* (recorte) que sean necesarios para obtener secuencias de mejor calidad.

Anexo III. Trimmomatic

Trimmomatic es una herramienta que se utiliza para mejorar las calidades de datos generados con la plataforma Illumina, donde se realiza el proceso de *trimming* o recorte de bases de mala calidad, eliminación de lecturas de mala calidad y eliminación de adaptadores. Esto permite evitar análisis erróneos como consecuencia de la presencia de secuencias de mala calidad en los datos de secuenciación de nueva generación (NGS).

Trimmomatic utiliza dos métodos para detectar secuencias técnicas dentro de las lecturas. El primero, es "modo simple", funciona encontrando una coincidencia aproximada entre la lectura y la secuencia técnica proporcionada por el usuario; y el segundo modo, denominado "modo palíndromo", tiene como objetivo la detección de este escenario común de "lectura de adaptador", en el que el fragmento de ADN secuenciado es más corto que la longitud de la lectura, y da lugar a la contaminación del adaptador en el extremo de las lecturas (Bolger et al., 2014).

Hay dos modos principales del programa: modo de extremo emparejado y modo de extremo único. En ambos se mantendrá la correspondencia de las lecturas y también utilizará la información contenida en ellas para encontrar mejor los adaptadores o *primers* de PCR. En el caso de extremo emparejado se requieren dos archivos de entrada y cuatro archivos de salida, dos para la salida 'emparejada' donde ambas lecturas sobrevivieron al procesamiento, y dos para la salida correspondiente "no emparejada" donde una sobrevivió, pero la lectura asociada no lo hizo. Por otra parte, en el modo de extremo único, se especifican un archivo de entrada y uno de salida (Fig. 35) (*Trimmomatic Manual: V0.32*, n.d.).

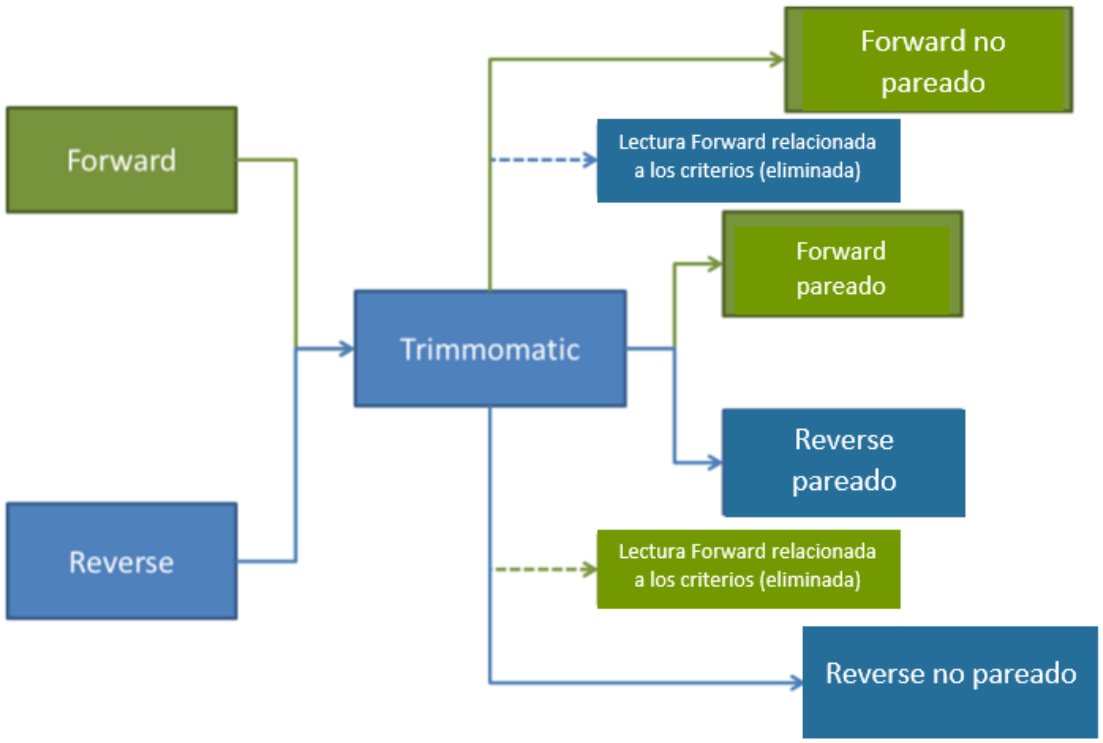


Fig. 35 Flujo de lecturas en Trimmomatic en modo de extremo emparejado. Modificado de TrimmomaticManual_V0.32

Anexo IV. Burrows-Wheeler Aligner BWA

BWA es un *software* que permite la alineación de lecturas cortas con un genoma humano de referencia de forma rápida y precisa generando un archivo de salida en formato SAM.
Se basa en la transformación de Burrows-Wheeler (Fig. 36), algoritmo en el cual se construyen matrices a partir de una cadena de datos, esta cadena está compuesta por letras de alfabeto y tiene un terminador que es el símbolo $, la cadena de datos tiene un orden inicial

pero el algoritmo permuta el orden de los datos y posteriormente los ordena lexicográficamente, de esta manera se obtienen datos de salida más fáciles de codificar.

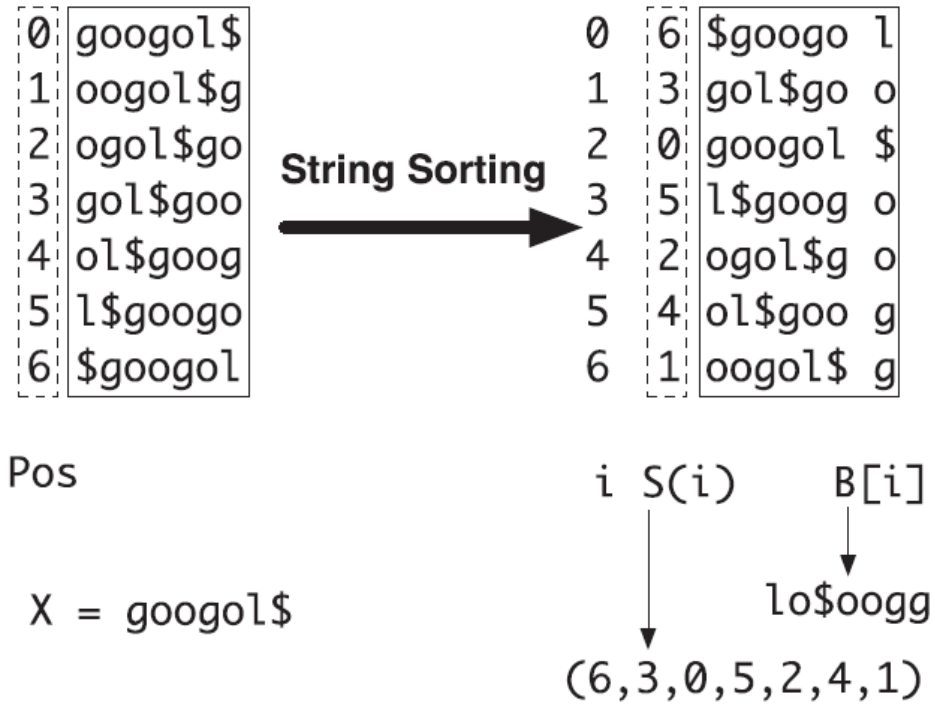


Fig. 36 Construcción de la matriz de sufijos.
Siendo X la cadena de datos googl$, hacemos el supuesto de que los datos forman un círculo, así según el terminador $ podemos obtener 7 cadenas distintas, estas se ordenan lexicográficamente y el orden en el que se obtienen las cadenas es la matriz (Li & Durbin, 2009).

Esto permite por ejemplo que, al comparar secuencias cortas con respecto de un genoma de referencia, esto suceda más rápidamente. Sería sencillo plantear que la búsqueda de una secuencia corta se tome completa y se busque a lo largo del genoma de referencia, sin embargo, eso sería más tardado y lo que en realidad sucede al usar programas basados en BWT es que la secuencia se alinea contra fragmentos más cortos y ordenados del mismo genoma de referencia. Esto se denomina *index* o índice y al igual que en un libro al buscar un capítulo permiten que la alineación reduzca el origen potencial de una secuencia de consulta dentro del genoma, ahorrando tiempo y memoria.

Otros aspectos prácticos de su uso se desglosan a continuación:

1. Las bases no A/C/G/T se denominan nucleótidos aleatorios, aunque con una posibilidad mínima debido a que las lecturas son relativamente largas.

2. Mapeo por pares, primero encuentra las posiciones de todos los resultados, los ordena según las coordenadas cromosómicas y luego hace un barrido lineal a través de todos los resultados posibles para emparejar los dos extremos.
3. Se puede determinar el número máximo de desajustes o gap.
4. Permite generar puntuaciones de calidad de mapeo para cada alineación (Li & Durbin, 2009).

Anexo V. SAMTOOLS

SAMtools es una biblioteca y un paquete de *softwares* que permite analizar y manipular archivos de alineaciones en formato SAM o BAM.
El formato de mapa de alineación de secuencias (SAM) es un formato de alineación común que soporta todos los tipos de secuencias y es el formato en el cual se obtienen los archivos procesados en BWA. Para mejorar el rendimiento, (Li & Durbin, 2009) diseñaron el formato *Binary Alignment/Map* (BAM), que es la representación binaria de SAM y mantiene exactamente la misma información con menor tamaño de almacenamiento y mejor velocidad de procesamiento.
Algunas de sus funciones son convertir otros formatos de alineación, ordenar y fusionar alineaciones, eliminar duplicados de PCR, generar información por posición en el formato *pileup*, reconocer SNPs e INDELS cortos, y mostrar alineaciones en un visor basado en texto.

Anexo VI. GATK

Los archivos de salida de BWA que se encuentran en formato SAM se convierten en archivos de entrada de *Genome Analysis Toolkit* (GATK). Esta herramienta se desarrolló originalmente como un marco de programación que permitía el desarrollo de nuevas herramientas de análisis del genoma y ha evolucionado hasta convertirse en una meta-herramienta pues se puede utilizar de forma inmediata para realizar pipelines, debido a que conjunta diferentes programas como *Samtools, Picard, Mutect2* etc (Van der Auwera *et al*., 2013). GATK permite el procesamiento y el control de calidad de los datos de secuenciación de alto rendimiento, es decir, toma en cuenta diferentes fuentes de error y así superar sistemáticamente a otros programas usados en el llamado de variantes. Por ello es el estándar de la industria para la identificación de SNPs e INDELS en datos de exomas y genomas completos generados por NGS. Una de las funciones más relevantes que considera la fuente de error es la Recalibración de la Puntuación de la Calidad de las Bases (BQSR), en la que las puntuaciones de la calidad de las bases asignadas por el secuenciador se corrigen con puntuaciones determinadas empíricamente a partir de los datos del grupo de lectura utilizando variantes validadas; estas puntuaciones recalibradas reflejan con mayor precisión la verdadera fiabilidad del llamado de bases, corrigiendo así los sesgos que presentan las plataformas de secuenciación (McCormick *et al*., 2015).

Generalmente el flujo de trabajo básico que incluye “las buenas prácticas de GATK” es el representado en la Fig. 37.

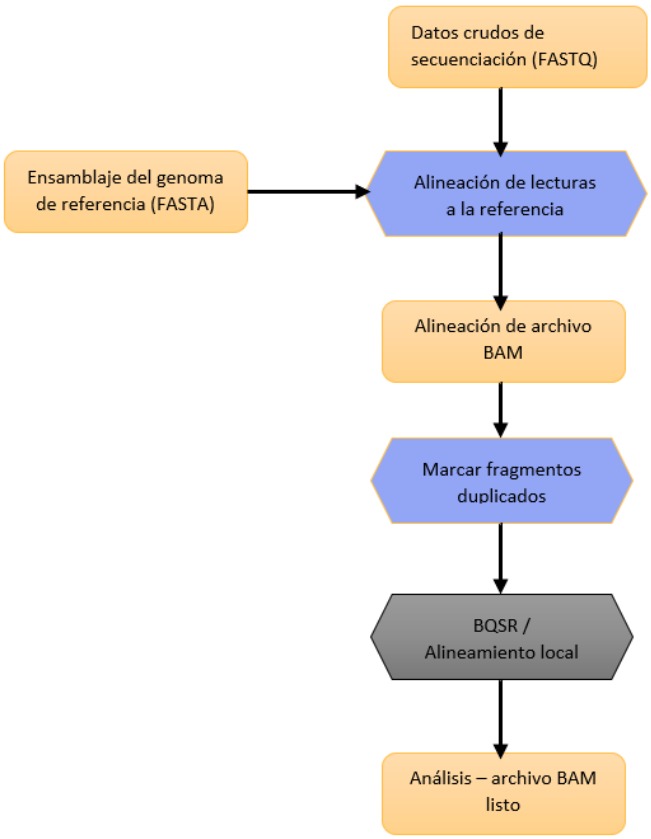


Fig. 37 Pipeline en GATK considerando las buenas prácticas. Modificado de (Koboldt, 2020)

Anexo VII. Picard

Es un conjunto de herramientas de línea de comandos para manipular datos de secuenciación de alto rendimiento (HTS) y formatos como SAM/BAM/CRAM y VCF. Es utilizada para limpiar, clasificar y fusionar los archivos de la misma muestra del paciente y para eliminar las lecturas duplicadas. Identificar y marcar las lecturas duplicadas en un archivo BAM permite excluirlas del análisis posterior. Además, permite añadir *readgroups*, los cuales son clave para la funcionalidad de GATK. El pipeline de GATK no funcionará sin estas etiquetas, pues contienen diferentes metadatos que permiten diferenciar muestras obtenidas en diferentes años, librerías o plataformas (Koboldt, 2020; Turajlic *et al*., 2017; Van der Auwera *et al*., 2013).

Las principales funciones que se usan son:

- SortSam: Esta herramienta ordena el archivo SAM o BAM de entrada por coordenadas, las alineaciones leídas se ordenan primero por el campo del nombre de la secuencia de referencia utilizando el diccionario (etiqueta @SQ). Las alineaciones dentro de estos subgrupos se ordenan en segundo lugar utilizando la posición de mapeo más a la izquierda de la lectura. Después de este esquema de ordenación, los alineamientos se enumeran de forma arbitraria.

- MergeSamFiles: Esta herramienta se utiliza para combinar archivos SAM y/o BAM de diferentes ejecuciones o *readgroups*. Para evitar errores en el procesamiento posterior, es fundamental identificar/etiquetar adecuadamente los *readgroups*. Si diferentes muestras contienen IDs de *readgroups* idénticos, esta herramienta evitará colisiones modificándolos para que sean únicos.

- AddOrReplaceReadGroups: Esta herramienta permite al usuario reemplazar todos los *readgroups* en el archivo INPUT con un único grupo de lectura nuevo y asignar todas las lecturas a este grupo de lectura en el archivo BAM OUTPUT.

Anexo VIII. Mutect2

Mutect2 es parte de las herramientas de GATK que se usa para evaluar mutaciones somáticas este utiliza un modelo probabilístico para la alineación de secuencias por pares para asignar una probabilidad de que cada lectura haya sido secuenciada de cada haplotipo candidato. Esto produce una matriz de probabilidades de lectura-haplotipo para la muestra del tumor y, si está presente, la muestra normal. Entonces calcula en cada una la máxima probabilidad entre los haplotipos que exhiben ese alelo (Benjamin *et al*., 2019). El archivo de entrada para este programa es un archivo BAM ya calibrado que contiene los *readgroups* y el archivo de salida es un VCF (variant calling format) que se utiliza a su vez como archivo de entrada en el programa de extracción de firmas *SigProfiler*. Las funciones principales de esta herramienta son:

- LearnReadOrientationModel: Si alguna de las muestras tiene errores de sustitución que se producen en una sola cadena de ADN antes de la secuenciación, se utiliza el filtro de sesgo de orientación de Mutect2. Primero, con --f1r2-tar-gz se crea una salida con los datos en bruto utilizados para aprender el modelo de sesgo de orientación. El archivo generado se aplica con la función LearnReadOrientationModel.
- GetPileupSummaries: Resumen los recuentos de lecturas que soportan alelos de referencia, alternativos y otros para sitios otorgados.
- CalculateContamination: Calcula la fracción de lecturas procedentes de la contaminación cruzada de la muestra
- FilterMutectCalls: Para obtener el archivo VCF a partir del modelo de orientación aprendido.

Anexo IX. Python

Python es un lenguaje de programación que posee sintaxis clara, comunidad de desarrolladores activa, disponibilidad gratuita y uso extensivo en comunidades científicas como la bioinformática. Python está diseñado para ser visualmente ordenado y sus principales cualidades son:

1. Ser un lenguaje de programación con una sintaxis limpia y semántica sencilla.
2. Con los avances de las ciencias biológicas, los datos científicos son cada vez de mayor tamaño en almacenamiento y heterogéneos. Por eso, al desarrollar y aplicar herramientas informáticas para el análisis de datos, los dos objetivos principales son la escalabilidad que ayuda en el manejo del volumen de datos y las abstracciones robustas que procesan la integración y heterogeneidad de los datos. Estas dos características las maneja adecuadamente el lenguaje Python lo que ayuda en el desarrollo de prototipos de algoritmos para tareas muy sencillas a otros con mayor complejidad que utilizan más recursos computacionales. Otra de las ventajas de este lenguaje es que cuenta con una licencia libre (Ekmekci *et al*., 2016).

Anexo X. R

R es un lenguaje de programación libre, usado principalmente para estadística, se encuentra en constante desarrollo y es apoyado por científicos y programadores. Entre sus principales cualidades se encuentra la diversidad de paquetes que posee en tres repositorios principales: 1. Bioconductor (proporciona herramientas para el análisis y la comprensión de datos genómicos), 2. CRAN y 3. Omegahat. Estos paquetes proporcionan una mayor flexibilidad y crea el vínculo entre varias fuentes de datos de una forma más simple. Además, posee una sintaxis similar a la de otros lenguajes de programación, haciéndolo más digerible (Milano, 2019).

Anexo XI. Extracción *de novo* vs *fitting*.

Los métodos para extracción de firmas pueden agruparse en dos categorías con diferentes objetivos. La primera se utiliza para descubrir nuevas firmas (extracción *de novo*), mientras que la segunda pretende detectar las firmas mutacionales conocidas y validadas en el catálogo mutacional de una muestra determinada (*fitting*). A grandes rasgos ambos enfoques necesitan, en principio, compararse con firmas de referencia como las del catálogo de mutaciones somáticas en cáncer (COSMIC por sus singlas en inglés) sin embargo, por diferentes razones. En el caso de la extracción *de novo*, obtiene medidas como la similitud del coseno para corroborar que una firma se trate o no de la misma a la que está siendo comparada dentro del catálogo y proponer una mejor solución. Por otra parte, en la extracción por *fitting* busca el emparejamiento a una firma específica o grupo de firmas dentro del catálogo que recree o se ajuste a la firma "global" obtenida (Maura *et al*., 2019; Omichessan

et al., 2019). En la extracción *de novo* se busca generar una firma desde factorización matricial no negativa (NNMF), este método matemático por definición genera más de una solución posible, sin embargo, se busca aquella que sea la más aproximada. En el caso de los algoritmos "*fitting*" se asignan a un catálogo de referencia como COSMIC usando exclusivamente el subconjunto de firmas de estas, el objetivo es estimar *E* dadas las matrices *M* y *P* (Fig. 38) (Tate *et al*., 2019).

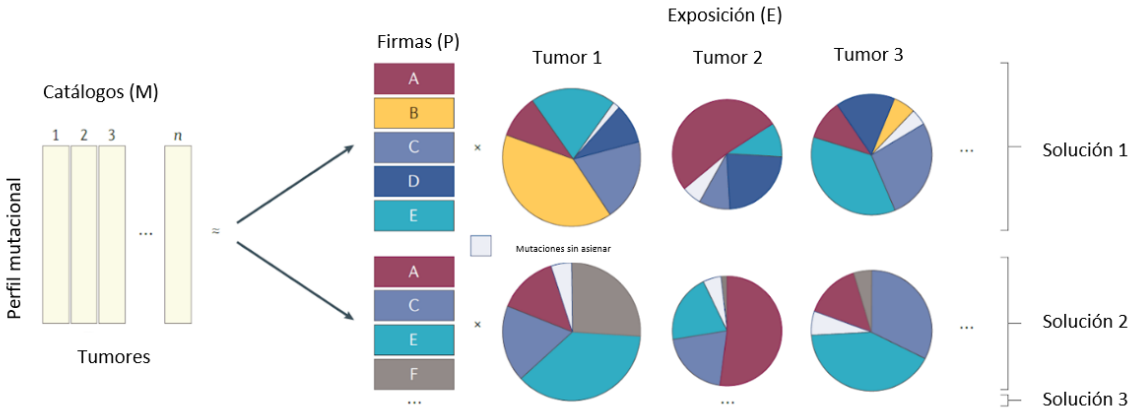


Fig. 38 Representación gráfica de las matrices para obtener firmas. La matriz M representa el catálogo, la matriz P representa la matriz de firmas y, por último, la matriz E que representa la exposición a eventos mutacionales, es aquella que se estima por NNMF, por tanto, se puede obtener más de una solución. Modificado de (Koh et al., 2021).

Con ambas herramientas, *deconstructSigs* y *SigProfiler*, se ha observado un aumento de la especificidad con respecto al número de mutaciones promedio, mientras que la sensibilidad es favorecida en muestras de cáncer que tienen un menor número de firmas características (Omichessan *et al*., 2019).

Anexo XII. *SigProfiler toolkit*

Este conjunto de herramientas bioinformáticas permite la extracción de firmas mutacionales, disponibles en Python y R, a partir de la generación de la matriz mutacional, así como la funcionalidad de apoyo para la esquematización y la simulación de estas. A pesar de que no son las únicas herramientas que lo permiten, cobran relevancia por ser las utilizadas en la creación del catálogo de firmas COSMIC, que posteriormente fue presentado en diversos artículos que respaldan este proyecto, así como otros que proponen la detección de firmas como apoyo al conocimiento de las causas de cáncer o mutaciones no cáncer.

SigProfiler se divide en 4 herramientas esencialmente: *SigProfilerMatrixGenerator*, *SigProfilerExtractor*, *SigProfilerSimulaor* y *SigProfilerPlotting.*

SigProfilerMatrixGenerator, es un marco de Python que crea matrices mutacionales para mutaciones somáticas. Esta herramienta funciona para identificar y categorizar las

mutaciones en función de posibles sustituciones de un solo nucleótido (SBS), sustituciones de doble base (DBS) e inserciones / deleciones (INDEL; además está herramienta se convirtió en un estándar de extracción *de novo* (Bergstrom *et al*., 2019). La extracción *de novo* de firmas mutacionales es un enfoque de aprendizaje automático no supervisado donde se busca determinar actividad de las firmas; es decir, el número de mutaciones aportadas por una firma en una muestra de cáncer. Para realizarlo se basa en la factorización de matrices no negativas (NMF), cuya ventaja principal es su capacidad para producir factores no negativos que forman parte de los datos originales. *SigProfilePlotting* es una herramienta estándar para mostrar y graficar todos los tipos de firmas mutacionales, así como todos los tipos de patrones mutacionales en los genomas del cáncer. *SigProfilerExtractor*, por otra parte, permite la extracción *de novo* de firmas mutacionales a partir de datos generados en un formato matricial, realiza 500 factorizaciones independientes y, para cada repetición, estas se comparan para identificar soluciones resistentes a las fluctuaciones de los datos de entrada y a la falta de unicidad del NMF. La herramienta identifica el número de firmas mutacionales operativas, sus actividades en cada muestra y la probabilidad de que cada firma cause un tipo de mutación específico en una muestra de cáncer. La herramienta hace uso de *SigProfilerMatrixGenerator* y *SigProfilerPlotting*, integrándose perfectamente (Islam *et al*., 2021; *SigProfilerExtractor*, 2019).

A pesar de que permite extraer la o las posibles soluciones *de novo*, también permite ver una "solución sugerida" en el caso de que la firma sea mejor explicada por medio de la descomposición en firmas del catálogo. Sin embargo, no se utiliza como enfoque de tipo *fitting* puesto que está optimizado para la extracción de firmas *de novo* y tiene una mayor sensibilidad al utilizar un número elevado de muestras.

Anexo XIII. *deconstructSigs*

El algoritmo determina la combinación lineal del marco de firmas predefinidas. Es decir, es un algoritmo de tipo "*fitting*". El marco de datos de entrada (T) se genera calculando la fracción de mutaciones encontradas en cada uno de los posibles 96 contextos de trinucleótidos para cada muestra de tumor. La fracción de veces que se observa una mutación en cada uno de los 96 contextos de trinucleótidos (s) para cada firma (k) son datos necesarios para calcular los pesos (W). Para determinar los pesos (W) que mejor recreen T, se adopta un enfoque iterativo. Excluye cualquier firma que contenga un único contexto de trinucleótidos, esto se hace para tener en cuenta el hecho de que algunas firmas se caracterizan casi por completo por mutaciones en contextos trinucleotídicos específicos, y por tanto sin mutaciones encontradas en esos contextos, es poco probable que esas firmas estén presentes. De las firmas restantes se elige una firma mutacional que mejor refleje el perfil mutacional de la muestra minimizando el error de suma cuadrada (SSE). A la primer firma que encuentre como coincidencia se le asigna un peso=1 y posteriormente, se repite con cada firma del catálogo hasta asignar el peso real en que atribuye cada firma a la extraída (Rosenthal et al., 2016).

ANEXO XIV. Todos los Scripts

Descargar archivos en SRA toolkit

1. Comando para descargar los archivos SRA

```
$ prefetch -v RUN
```

2. Comando para dividir archivos fastq en en *forward* y *reverse*

```
$ fastq-dump --outdir /ruta/salida/ --split-files
/ruta/donde/descargué/sra/SRREJEMPLO.sra
```

FASTQC

Debido a que parte importante de este proyecto es simplificar el procedimiento, se diseñó un Script en el editor de texto *Nano*, en el cual se incluye la sintaxis completa de un "ciclo for" para FASTQC que se encargará de realizar la misma acción en bucle hasta finalizar con los archivos FASTQ de interés.

El Script contiene lo siguiente

```
#!/bin/bash

#Ciclo for para correr todos los archivos FASTQ en FastQC

for VARIABLE in $(ls /../Input-Files/FASTQs/sample*)
do
echo Estoy trabajando con ${VARIABLE} y se va a realizar un
Fastqc
fastqc ${VARIABLE} -t 8 -o /../Intermediate-
Files/FASTQC/Before_T/
echo Ya termine con el archivo ${VARIABLE}
done
```

La sintaxis general de un ciclo for es

```
for <VALOR> in <LISTA DE VALORES>
do
#instrucciones
Done
```

En este caso, mi lista de valores se trata del contenido en mi carpeta denominada Input-files y las instrucciones solicitadas son: decir con que archivo está trabajando, realizar en análisis en FASTQC, guardar los archivos de salida en la carpeta denominada Before_T y, finalmente indicar cuando finalice el procedimiento con el archivo determinado.

Una vez realizado el análisis en FASTQC, se revisaron los archivos de salida en donde se pueden visualizar los 10 módulos de análisis relatados en el anexo II

Trimmomatic

Con base en las observaciones pertinentes en FASTQC se realiza una limpieza de las lecturas. Trimmomatic puede ser ejecutado bajo la siguiente sintaxis:

```
java -jar /mnt/d/Programas/Trimmomatic-0.39/trimmomatic-
0.39.jar PE Input1_1.fq Input1_2.fq Output1_1_paired.fq
Output1_1_unpaired.fq Output1_2_paired.fq
Output1_2_unpaired.fq <PARAMETROS>
```

Con java -jar se ejecuta el programa indicando la carpeta donde fue instalado, PE es para indicar el modo Pairend. Este modo requiere dos archivos de entrada (input *forward* y *reverse*); la asignación de 4 archivos de salida (output *forward* pareado, *forward* no pareado, *reverse* pareado y *reverse* no pareado); y la asignación de los parámetros deseados.

Al igual que en el caso de FASTQC, se elaboró un script en el editor de texto nano, para asignar variables que sintetizaran la sintaxis descrita y fuera genérico para poder aplicarlo a diferentes archivos y usarlo dentro de un ciclo for.

```
#!/bin/bash
#Entrada de archivos fastq forward y reverse
F=$1
R=$2

#Renombrar archivos de salida
FP=$(basename ${F}|sed 's/.fq/_paired.fq/')
FU=$(basename ${F}|sed 's/.fq/_unpaired.fq/')
RP=$(basename ${R}|sed 's/.fq/_paired.fq/')
RU=$(basename ${R}|sed 's/.fq/_unpaired.fq/')

#Asignar la carpeta de salida
FILTERFQ='../Intermediate-Files/Filtered-FASTQs'

#Asignar ruta de trimmomatic
trimmomatic='java -jar /mnt/d/Programas/Trimmomatic-
0.39/trimmomatic-0.39.jar PE'

#Correr el trimmomatic
echo Estoy usando Trimmomatic para limpiar mis muestras
${trimmomatic} ${F} ${R} ${FILTERFQ}${FP} ${FILTERFQ}${FU}
${FILTERFQ}${RP} ${FILTERFQ}${RU}
ILLUMINACLIP:/.../Programs/Trimmomatic-0.39/adapters/TruSeq3-
PE-2.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:70
echo Ya termine de usar Trimmomatic
```

Para asignar los archivos de salida se utilizó el comando basename que permite extraer el nombre de un archivo en una ruta dada. En este caso si asigno a la variable

```
F= /mnt/d/MutationalSignatures_Pipeline/Intermediate-Files/Filtered-FASTQs /Sample1_1.fq.gz
```
el comando extrae solo “Sample1_1.fq.gz”
El símbolo | (pipe o tubería) permite encadenar la ejecución de programas, el output del comando basename pasa como el input del comando sed. Este último busca un patrón marcado entre / / y lo sustituye por un nuevo patrón dado. De tal manera, se tiene:

```
sed 's/.fq.gz/_paired.fq.gz/')
```
Busca el patron “.fq.gz” y los sustituye por “_paired.fq.gz” logrando que el resultado final de
```
FP=$(basename ${F}|sed 's/.fq/_paired.fq.gz/') para el archivo /mnt/d/MutationalSignatures_Pipeline/Intermediate-Files/Filtered-FASTQs /Sample1_1.fq
```

Sea

```
Sample1_1_paired.fq.gz
```

Finalmente, se realizó un ciclo for, esta vez sustituyendo las instrucciones por la ejecución del Script realizado.

```
for archivo in $(ls /../Input-Files/FASTQs/*|cut -d ‘_’ -f 1-6|uniq)
do ./02.Trimmomatic.sh ${archivo}’_1.fq.gz’ ${archivo}’_2.fq.gz’
done
```

A los archivos de salida del trimmomatic, se les realiza un segundo análisis FASTQC para corroborar la calidad de las secuencias.

```
#!/bin/bash

#Ciclo for para procesar archivos fastq en el programa FastQC despues del trimmomatic

for VARIABLE in $(ls /../Intermediate-Files/Filtered-FASTQs/*)
do
echo Estoy trabajando con ${VARIABLE} y se va a realizar un Fastqc
fastqc --threads 8 ${VARIABLE} -o /../Intermediate-Files/FASTQC/After_T/
echo Ya termine con el archivo ${VARIABLE}
done
```

Para realizar el alineamiento de secuencias y obtener el merged, se sigue el mismo principio para el cambio de las variables. El cambio más importante es el programa, las opciones y los argumentos que cada uno necesita.

```
#Designar variables para los 4 archivos de salida de
trimmomatic

FP=$1
RP=$2
FU=$3
RU=$4

#Asignar la variable del genoma de referencia
GENREF='/.../References/hg38_v0_Homo_sapiens_assembly38.fasta
'
#Asignar la ruta de BWA
BWA='/.../Programs/bwa-0.7.17/bwa mem -t 8'
#Asignar ruta de salida
BAMs='/../Intermediate-Files/BAMs/'
#Asignar la variable para el archivo de salida de bwa
SAMP=$(basename ${FP}|sed 's/_1_paired.fq.gz/_paired.sam/')
SAMF=$(basename ${FP}|sed 's/_paired.fq.gz/_single.sam/')
SAMR=$(basename ${RP}|sed 's/_paired.fq.gz/_single.sam/')

#Realizar anotación

echo Voy a realizar la anotacion de FP y RP
${BWA} ${GENREF} ${FP} ${RP} -o ${BAMs}${SAMP}
echo Voy a realizar la anotacion del FU
${BWA} ${GENREF} ${FU} -o ${BAMs}${SAMF}
echo Voy a realizar la anotacion del RU
${BWA} ${GENREF} ${RU} -o ${BAMs}${SAMR}
echo Ya termine la anotacion

####Pasar archivos de SAM a BAM

#Asignar variables para el cambio de extension en SAMtools
BAMP=$(echo ${SAMP} |sed 's/\.sam/\.bam/')
BAMF=$(echo ${SAMF} |sed 's/\.sam/\.bam/')
BAMR=$(echo ${SAMR} |sed 's/\.sam/\.bam/')

#Asignar variable de la ruta SAMtools
SAM="/.../Programs/samtools-1.10/samtools"

#Cambio de formato de SAM a BAM
${SAM} view -bS --threads 8 ${BAMs}${SAMP} -o ${BAMs}${BAMP}
${SAM} view -bS --threads 8 ${BAMs}${SAMF} -o ${BAMs}${BAMF}
${SAM} view -bS --threads 8 ${BAMs}${SAMR} -o ${BAMs}${BAMR}

###Realizar ordenamiento de las lecturas

#Asignar ruta para Picard
Picard="java -jar /.../Programs/picard/build/libs/picard.jar"
#Cambio de nombre para generar archivos BAM sorted
BAMsortP=$(echo ${BAMP} |sed 's/\.bam/_sorted\.bam/')
BAMsortF=$(echo ${BAMF} |sed 's/\.bam/_sorted\.bam/')
BAMsortR=$(echo ${BAMR} |sed 's/\.bam/_sorted\.bam/')
```

```
#Realizar ordenamiento con Picard
${Picard} SortSam I=${BAMs}${BAMP} O=${BAMs}${BAMsortP}
SO=coordinate
${Picard} SortSam I=${BAMs}${BAMF} O=${BAMs}${BAMsortF}
SO=coordinate
${Picard} SortSam I=${BAMs}${BAMR} O=${BAMs}${BAMsortR}
SO=coordinate

#Cambio de nombre para el archivo BAM merged
BAMmerged=$(echo ${BAMsortP} | sed
's/_paired_sorted\.bam/_merged\.bam/')

#Realizar merged con Picard
${Picard} MergeSamFiles INPUT=${BAMs}${BAMsortP}
INPUT=${BAMs}${BAMsortR} INPUT=${BAMs}${BAMsort F}
OUTPUT=${BAMs}${BAMmerged}
```

El Script anterior será el ejecutado en el siguiente ciclo for

```
for MUESTRA in $(ls /../Intermediate-Files/Filtered-
FASTQs/*|cut -d '_' -f 1-6|uniq)
do ./05.BWA.sh $MUESTRA'_1_paired.fq.gz'
$MUESTRA'_2_paired.fq.gz' $MUESTRA'_1_unpaired.fq.gz'
$MUESTRA'_2_unpaired.fq.gz'
```

A los archivos de salida merged.bam se les añade los *readgroups*, en caso de ser los mismos para todas las muestras se puede integrar al primer Script, pero si estos son distintos se sugiere añadirlos por separado.

```
###Agregar readgroups al archivo BAMmerged

#Asignar variable al archivo de entrada

BAMmerged=$1

#Asignar ruta de salida
BAMs="/../Intermediate-Files/BAMs/"

#Asignar ruta de Picard
Picard="java -jar /.../Programs/picard/build/libs/picard.jar"
#Asignar nombre al archivo de salida
BAMmergedRG=$(echo ${BAMmerged}|sed
's/_merged\.bam/_merged_RG\.bam/')

${Picard} AddOrReplaceReadGroups -I ${BAMs}${BAMmerged} -O
${BAMs}${BAMmergedRG} -LB library -PL ILLUMINA -PU 1 -SM
sample_name --CREATE_INDEX true
```

Para realizar la recalibración con BaseRecalibrator y ABQSR

```
#!/bin/bash

#Variable para el genoma de referencia
GENREF='/.../References/hg38_v0_Homo_sapiens_assembly38.fasta
'
#Asignar variable al archivo BAM final
BAM=$1

#Variable de known-sites

MILGENOMAS='/.../References/1000G_phase1.snps.high_confidence
.hg38.vcf.gz'
MILLS='/.../References/Mills_and_1000G_gold_standard.indels.h
g38.vcf.gz'

###Recalibración de calidades PHRED con GATK##

#Asignar cambio de nombre para archivo de salida de
BaseRecalibrator

BRC=$(echo ${BAM} | sed
's/_merged_RG\.bam/_recal_data\.table/')

#Asignar ruta del programa
GATK='java -jar /.../Programs/gatk-4.1.8.1/gatk-package-
4.1.8.1-local.jar'

#Ejecutar BaseRecalibrator

echo 'Estoy ejecutando BaseRecalibrator de' ${BAM}

${GATK} BaseRecalibrator -R ${GENREF} -I ${BAM} --known-sites
${MILGENOMAS} --known-sites ${MILLS} -O ${BRC}

#Asignar archivo de salida para ApplyBQSR
ABQSR=$(echo ${BAM} | sed
's/_merged_RG\.bam/_recal_data\.bam/')
#Ajuste con ApplyBQSR
${GATK} ApplyBQSR -R ${GENREF} -I ${BAM} --bqsr-recal-file
${BRC} -O ${ABQSR}

BRC2=$(echo ${BAM} | sed
's/_merged_RG\.bam/_recal_data2\.table/')

#Ejecutar BaseRecalibrator por segunda vez

echo 'Estoy ejecutando BaseRecalibrator por segunda vez'

${GATK} BaseRecalibrator -R ${GENREF} -I ${ABQSR} --known-
sites ${MILGENOMAS} --known-sites ${MILLS} -O ${BRC2}
#Asignar archivo de salida para ApplyBQSR
ABQSR2=$(echo ${BAM} | sed
's/_merged_RG\.bam/_recal_data2\.bam/')
```

```
#Ajuste con ApplyBQSR
${GATK} ApplyBQSR -R ${GENREF} -I ${ABQSR} --bqsr-recal-file
${BRC2} -O ${ABQSR2}
```

El Script anterior se ejecuta en el siguiente ciclo for

```
for MUESTRA in $(ls /../Intermediate-Files/BAMs/*|cut -d '_'
-f 1-6|uniq)
do ./08.BaseRecalibrator.sh $MUESTRA'_merged_RG.bam'
done
```

Para ejecutar Mutect2 para el llamado de variantes

```
#!/bin/bash
#Variable para el genoma de referencia
GENREF='/.../References/hg38_v0_Homo_sapiens_assembly38.fasta
'
#Asignar variable al archivo

#Variable de bases de datos

PON='/.../References/1000g_pon.hg38.vcf.gz'
GNOMAD='/.../References/af-only-gnomad.hg38.vcf.gz'
CNV='/.../References/CNV_and_centromere_blacklist.hg38liftove
r.list'

#Asignar ruta del programa
GATK='java -jar /.../Programs/gatk-4.1.8.1/gatk-package-
4.1.8.1-local.jar'

#Ruta a archivos de salida
TABLE='/../Results/VCFs/VCF_table/'
TAR='/../Results/VCFs/VCF_tar/'
VCFunfiltered='/../Results/VCFs/VCF_unfiltered/'
VCFfiltered='/../Results/VCFs/VCF_filtered/'

#Asignar variable al archivo de entrada
ABQSR2=$1
#Asignar variable a archivos de salida
F1R2=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_f1r2\.tar\.gz/')
UNFILTERED=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_unfiltered\.vcf/')
echo 'Estoy ejecutando Mutec2 de' ${ABQSR2}

${GATK} Mutect2 -R ${GENREF} -L ${CNV} \
-I ${ABQSR2} \
-germline-resource ${GNOMAD} -pon ${PON} --f1r2-tar-gz
${TAR}${F1R2} \
-O ${VCFunfiltered}${UNFILTERED}

#Asignar variables a archivos de salida
```

```
ORIENTATION=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_read-orientation-model\.tar\.gz/')
echo 'Aprendiendo orientacion'

${GATK} LearnReadOrientationModel -I ${TAR}${F1R2} -O
${TAR}${ORIENTATION}
#Asignar variables a archivos de salida
PILEUP=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_getpileupsummaries\.table/')

echo 'Obtener el resumen por GetPileupSummaries'
${GATK} GetPileupSummaries -I ${ABQSR2} -V ${GNOMAD} -L
${CNV} -O ${TABLE}${PILEUP}
#Asignar variables a archivos de salida
SEGMENTS=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_segments\.table/')
CONTAMINATION=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_calculatecontamination\.table/')

echo 'Realizar el calculo de la contaminacion'
${GATK} CalculateContamination -I ${TABLE}${PILEUP} -tumor-
segmentation ${TABLE}${SEGMENTS} -O ${TABLE}${CONTAMINATION}

#Asignar variables a archivo VCF filtrado

FILTERED=$(basename ${ABQSR2}|sed
's/_recal_data2\.bam/_filtered\.vcf/')
echo 'Agregar la orientacion aprendida a las variantes
llamadas'
${GATK} FilterMutectCalls -V ${VCFunfiltered}${UNFILTERED} \
        -R ${GENREF} \
        --tumor-segmentation ${TABLE}${SEGMENTS} \
        --contamination-table ${TABLE}${CONTAMINATION} \
        --orientation-bias-artifact-priors
${TAR}${ORIENTATION} \
        -O ${VCFfiltered}${FILTERED}
```

Este Script se inserta dentro del siguiente ciclo for

```
for MUESTRA in $(ls /.../Intermediate-Files/BAMs/*|cut -d '_'
-f 1-6|uniq)
do ./10-Mutect2.sh $MUESTRA'_recal_data2.bam'
done
```

Script para limpiar el archivo .VCF obtenido en Mutect2 y obtener solo las variantes confirmadas en “PASS”

```
#! /bin/bash
FILTERED=$1

PASS=$(basename ${FILTERED}|sed
's/_filtered\.vcf/_pass\.vcf/')
```

```
grep '#CHROM' ${FILTERED} > /../Results/VCFs/VCF_pass/${PASS}
grep -v '##' ${FILTERED}|grep 'PASS' >>
/../Results/VCFs/VCF_pass/${PASS}
```

Este Script se inserta en el siguiente ciclo for

```
#!/bin/bash

for MUESTRA in $(ls /.../Results/VCFs/VCF_filtered/*|cut -d
‘_’ -f 1-7|uniq)
do ./12-VCFpass.sh ${MUESTRA}'_filtered.vcf'
done
```

Las muestras obtenidas serán los archivos de entrada en SigProfiler, el cual se corre en lenguaje de Python y puede realizarse dentro de la misma terminal Linux con el comando `$ python3`

Y se realiza lo siguiente

```
>>> from SigProfilerMatrixGenerator import install as
genInstall
>>> genInstall.install('GRCh38')
>>> from SigProfilerExtractor import sigpro as sig
>>> def main_function():

>>> sig.sigProfilerExtractor(input_type="vcf", output="/.../
Results/sigProfiler/ ", input_data="/.../
Results/VCFs/VCF_pass/", reference_genome='GRCh38',
exome=False, minimum_signatures=1, maximum_signatures=10,
nmf_replicates=100, cpu=4, make_decomposition_plots=True)
```

Se obtienen diferentes carpetas y gráficos como se explica en el anexo XII. En los resultados obtenidos se encuentran tablas que sirven como archivo de entrada para deconstructSigs.

```
# Asignar variable de entrada
head(sample.mut.ref)
mut.ref.relatives <-
as.data.frame(read.csv(".../Results/TSV/TSV_relatives/All_rel
atives_pass.tsv",header = TRUE,sep="\t"))
head(sample.mut.ref)

library(BSgenome.Hsapiens.UCSC.hg38)

# Convertir a entrada de deconstructSigs
sigs.input.relatives <- mut.to.sigs.input(mut.ref =
mut.ref.relatives,
                                  sample.id = "Sample",
                                  chr = "chr",
                                  pos = "pos",
                                  ref = "ref",
                                  alt = "alt",
```

```
                                    bsg =
BSgenome.Hsapiens.UCSC.hg38)
```

Para la extracción tipo *fitting* se usa deconstructSigs en la terminal de R. deconstructSigs trabaja con el catálogo de COSMIC 2013 (a la fecha de realizado el presente trabajo. Por ello fue necesario cargar el catálogo actualizado, usando el archivo descargable de https://cancer.sanger.ac.uk/signatures/downloads/

```
#Crear referencias con firmas actualizadas
signatures.DBS2020 <-
as.data.frame(read.csv(".../Ref_signatures/
COSMIC_v3.2_DBS_GRCh38.txt"
header = TRUE,sep="\t"))
#Exportado de excel pasar a dataframe
ref.SBS2020 <- as.data.frame(ref_SBS2020,header =
TRUE,sep="\t")

#Referencia que contiene columna como nombre de las filas
cosmic.SBS2020 <-
as.data.frame(tibble::column_to_rownames(ref.SBS2020),header
= TRUE,sep="\t")
```

Una vez creado el catálogo actualizado se continua con el siguiente Script ejemplo para la muestra 1, se puede copiar en R las veces que sea necesario, cambiando la información de la muestra, dentro del mismo Script de R.

```
# Determinar que firmas contribuyen a muestras normalizadas
de casos índice

sample_proband1 = whichSignatures(tumor.ref =
sigs.input.probands,
                                  signatures.ref =
cosmic.SBS2020,
                                  sample.id = 1,
                                  contexts.needed = TRUE)

#Obtener Plots de casos índice

plotSignatures(sample_proband1, sub = 'Wang1')


#Obtener gráficas Pie de casos índice

makePie(sample_proband1, sub = 'Wang1', add.color = "cyan")
```

Este mismo procedimiento se repitió con las muestras de los familiares. En el presente proyecto se trabajó con las muestras de cinco pacientes AF y 11 familiares de ellos, por lo cual en el Script en R se realizó 16 veces para cada muestra.

ANEXO XV. Script maestro

En este Script se conjuntan los Scripts anteriores para realizar un flujo de trabajo continuo.

```
#! /bin/bash

source config-file.sh


### Enumeracion de los Scripts a usar ###
Script_P1='//Scripts/Script01_TrimmToBAM.sh' #Trimmomatic a
bam merged
Script_P2='//Scripts/Script02_VCF.sh' #Llamado de variantes y
obtención de PASS

##### Running Trimmomatic, BWA, converting SAM to BAM, sort
and merge BAMs, addReadgroups ####
#for archivo in $(ls ${InputFolder}*|cut -d '_' -f 1-2|uniq)
#       do
#                echo 'Actualmente estoy en ${PWD} con las
muestas para Trimmomatic'
#               ${Script_P1} ${archivo}'_1.fq'
${archivo}'_2.fq'
#       done
#done

####### Recalibracion de calidades y Mutect2 para generar
archivos VCF y obtener las variantes PASS #######
#find ${BAMs} -name '*_merged-RG.bam' \
#       | xargs -I BAM -P ${NT} ${Script_P2} BAM
```

11.- Referencias bibliográficas

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, *47*(12), 1402–1407. https://doi.org/10.1038/ng.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Getz, G., … Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V, Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell*

Reports, *3*(1), 246–259. https://doi.org/10.1016/j.celrep.2012.12.008

Alter, B. P., & Giri, N. (2016). Thinking of VACTERL-H? Rule out Fanconi Anemia according to PHENOS. *American Journal of Medical Genetics. Part A*, *170*(6), 1520–1524. https://doi.org/10.1002/ajmg.a.37637

Ashley, C. W., Da Cruz Paula, A., Kumar, R., Mandelker, D., Pei, X., Riaz, N., Reis-Filho, J. S., & Weigelt, B. (2019). Analysis of mutational signatures in primary and metastatic endometrial cancer reveals distinct patterns of DNA repair defects and shifts during tumor progression. *Gynecologic Oncology*, *152*(1), 11–19. https://doi.org/10.1016/j.ygyno.2018.10.032

Auerbach, A. D. (2015). Diagnosis of Fanconi anemia by diepoxybutane analysis. *Current Protocols in Human Genetics*, *85*, 8.7.1-8.7.17. https://doi.org/10.1002/0471142905.hg0807s85

Ayyildiz, D., & Piazza, S. (2019). *Introduction to Bioinformatics BT - Microarray Bioinformatics* (V. Bolón-Canedo & A. Alonso-Betanzos (Eds.); pp. 1–15). Springer New York. https://doi.org/10.1007/978-1-4939-9442-7_1

Bayat, A. (2002). Bioinformatics. *BMJ*, *324*(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood - Education and Practice*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Ben Haj Ali, A., Amouri, A., Sayeb, M., Makni, S., Hammami, W., Naouali, C., Dallali, H., Romdhane, L., Bashamboo, A., McElreavey, K., Abdelhak, S., & Messaoud, O. (2019). Cytogenetic and molecular diagnosis of Fanconi anemia revealed two hidden phenotypes: Disorder of sex development and cerebro-oculo-facio-skeletal syndrome. *Molecular Genetics & Genomic Medicine*, *7*(7), e00694. https://doi.org/10.1002/mgg3.694

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. *BioRxiv*, 861054. https://doi.org/10.1101/861054

Bergstrom, E. N., Huang, M. N., Mahto, U., Barnes, M., Stratton, M. R., Rozen, S. G., & Alexandrov, L. B. (2019). SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*, *20*(1), 685. https://doi.org/10.1186/s12864-019-6041-2

Bhandari, J., Thada, P. K., & Puckett, Y. (2021). Fanconi Anemia. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK559133/

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boot, A., Ng, A. W. T., Chong, F. T., Tan, D. S. W., Gopalakrishna Iyer, N., & Rozen, S. G. (2018). Mutational signature analysis of Asian OSCCs reveals novel mutational signature with exceptional sequence context specificity. *BioRxiv*, 368753. https://doi.org/10.1101/368753

Brosh, R. M. J., Bellani, M., Liu, Y., & Seidman, M. M. (2017). Fanconi Anemia: A DNA repair disorder characterized by accelerated decline of the hematopoietic stem cell compartment and other features of aging. *Ageing Research Reviews*, *33*, 67–75. https://doi.org/10.1016/j.arr.2016.05.005

Brown, J., Pirrung, M., & McCue, L. A. (2017). FQC Dashboard: integrates FastQC results

into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics (Oxford, England)*, *33*(19), 3137–3139. https://doi.org/10.1093/bioinformatics/btx373

Cañedo Andalia, R., & Arencibia Jorge, R. (2004). Bioinformática: en busca de los secretos moleculares de la vida. *ACIMED*, *12*(6), 1. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1024-94352004000600002&lng=es&nrm=iso&tlng=en

Che, R., Zhang, J., Nepal, M., Han, B., & Fei, P. (2018). Multifaceted Fanconi Anemia Signaling. *Trends in Genetics : TIG*, *34*(3), 171–183. https://doi.org/10.1016/j.tig.2017.11.006

Chong, W., Wang, Z., Shang, L., Jia, S., Liu, J., Fang, Z., Du, F., Wu, H., Liu, Y., Chen, Y., & Chen, H. (2021). Association of clock-like mutational signature with immune checkpoint inhibitor outcome in patients with melanoma and NSCLC. *Molecular Therapy - Nucleic Acids*, *23*, 89–100. https://doi.org/https://doi.org/10.1016/j.omtn.2020.10.033

Clauson, C., Schärer, O. D., & Niedernhofer, L. (2013). Advances in understanding the complex mechanisms of DNA interstrand cross-link repair. *Cold Spring Harbor Perspectives in Biology*, *5*(10), a012732. https://doi.org/10.1101/cshperspect.a012732

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in Molecular Biology (Clifton, N.J.)*, *1418*, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

COSMIC | Mutational Signatures. (n.d.). Retrieved January 17, 2023, from https://cancer.sanger.ac.uk/signatures/

Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., Fostel, J. L., Friedrich, D. C., Perrin, D., Dionne, D., Kim, S., Gabriel, S. B., Lander, E. S., Fisher, S., & Getz, G. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, *41*(6), e67. https://doi.org/10.1093/nar/gks1443

Datta, A., & Brosh, R. M. J. (2019). Holding All the Cards-How Fanconi Anemia Proteins Deal with Replication Stress and Preserve Genomic Stability. *Genes*, *10*(2), 170. https://doi.org/10.3390/genes10020170

Davis-Turak, J., Courtney, S. M., Hazard, E. S., Glen, W. B. J., da Silveira, W. A., Wesselman, T., Harbin, L. P., Wolf, B. J., Chung, D., & Hardiman, G. (2017). Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Review of Molecular Diagnostics*, *17*(3), 225–237. https://doi.org/10.1080/14737159.2017.1282822

de Sena Brandine, G., & Smith, A. D. (2019). Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Research*, *8*, 1874. https://doi.org/10.12688/f1000research.21142.2

Dodgshun, A. J., Sexton-Oates, A., Saffery, R., & Sullivan, M. J. (2016). Biallelic FANCD1/BRCA2 mutations predisposing to glioblastoma multiforme with multiple oncogenic amplifications. *Cancer Genetics*, *209*(1–2), 53–56. https://doi.org/10.1016/j.cancergen.2015.11.005

Duan, M., Speer, R. M., Ulibarri, J., Liu, K. J., & Mao, P. (2021). Transcription-coupled nucleotide excision repair: New insights revealed by genomic approaches. *DNA Repair*, *103*, 103126. https://doi.org/10.1016/j.dnarep.2021.103126

Ekmekci, B., McAnany, C. E., & Mura, C. (2016). An Introduction to Programming for Bioscientists: A Python-Based Primer. *PLoS Computational Biology*, *12*(6), e1004867. https://doi.org/10.1371/journal.pcbi.1004867

Faivre, L., Guardiola, P., Lewis, C., Dokal, I., Ebell, W., Zatterale, A., Altay, C., Poole, J., Stones, D., Kwee, M. L., van Weel-Sipman, M., Havenga, C., Morgan, N., de Winter, J., Digweed, M., Savoia, A., Pronk, J., de Ravel, T., Jansen, S., … Mathew, C. G. (2000). Association of complementation group and mutation type with clinical outcome in fanconi anemia. European Fanconi Anemia Research Group. *Blood*, *96*(13), 4064–4070.

Faivre, L., Portnoï, M. F., Pals, G., Stoppa-Lyonnet, D., Le Merrer, M., Thauvin-Robinet, C., Huet, F., Mathew, C. G., Joenje, H., Verloes, A., & Baumann, C. (2005). Should chromosome breakage studies be performed in patients with VACTERL association? *American Journal of Medical Genetics. Part A*, *137*(1), 55–58. https://doi.org/10.1002/ajmg.a.30853

Feben, C., Spencer, C., Lochan, A., Laing, N., Fieggen, K., Honey, E., Wainstein, T., & Krause, A. (2017). Biallelic BRCA2 mutations in two black South African children with Fanconi anaemia. *Familial Cancer*, *16*(3), 441–446. https://doi.org/10.1007/s10689-017-9968-y

Fiesco-Roa, M. O., Giri, N., McReynolds, L. J., Best, A. F., & Alter, B. P. (2019). Genotype-phenotype associations in Fanconi anemia: A literature review. *Blood Reviews*, *37*, 100589. https://doi.org/10.1016/j.blre.2019.100589

Frohnmayer Lynn, Ravenhorst Sherri Van, & Wirkkula Leanne. (2020). *Fanconi Anemia Clinical Care Guidelines* (Lynn Frohnmayer, Sherri Van Ravenhorst, & Leanne Wirkkula (Eds.); 5°). The Fanconi Anemia Research Fund. https://www.fanconi.org/images/uploads/other/Fanconi_Anemia_Clinical_Care_Guidelines_5thEdition_web.pdf

García-de Teresa, B., & del Castillo, V. (2014). Diagnóstico clínico y de laboratorio de la anemia de Fanconi. *Acta Pediátrica de México*, *33*(1), 38–43. https://ojs.actapediatrica.org.mx/index.php/APM/article/view/533

García-de Teresa, B., Frias, S., Molina, B., Villarreal, M. T., Rodriguez, A., Carnevale, A., López-Hernández, G., Vollbrechtshausen, L., Olaya-Vargas, A., & Torres, L. (2019). FANCC Dutch founder mutation in a Mennonite family from Tamaulipas, México. *Molecular Genetics & Genomic Medicine*, *7*(6), e710. https://doi.org/10.1002/mgg3.710

García-de Teresa, B., Rodríguez, A., & Frias, S. (2020). Chromosome Instability in Fanconi Anemia: From Breaks to Phenotypic Consequences. *Genes*, *11*(12), 1528. https://doi.org/10.3390/genes11121528

García-de Teresa, B., Rodríguez, A., & Frías, S. (2016). Estudio multidisciplinario del paciente con anemia de Fanconi. *Acta Pediátrica de México; Vol. 37, Núm. 1 (2016)DO - 10.18233/APM37No1pp54-59*. https://ojs.actapediatrica.org.mx/index.php/APM/article/view/1131

Garrido-Cardenas, J. A., Garcia-Maroto, F., Alvarez-Bermejo, J. A., & Manzano-Agugliaro, F. (2017). DNA Sequencing Sensors: An Overview. *Sensors (Basel, Switzerland)*, *17*(3), 588. https://doi.org/10.3390/s17030588

Helbling-Leclerc, A., Dessarps-Freichey, F., Evrard, C., & Rosselli, F. (2019). Fanconi

anemia proteins counteract the implementation of the oncogene-induced senescence program. *Scientific Reports*, *9*(1), 17024. https://doi.org/10.1038/s41598-019-53502-w

Helbling-Leclerc, A., Garcin, C., & Rosselli, F. (2021). Beyond DNA repair and chromosome instability-Fanconi anaemia as a cellular senescence-associated syndrome. *Cell Death and Differentiation*, *28*(4), 1159–1173. https://doi.org/10.1038/s41418-021-00764-5

Housh, K., Jha, J. S., Haldar, T., Amin, S. B. M., Islam, T., Wallace, A., Gomina, A., Guo, X., Nel, C., Wyatt, J. W., & Gates, K. S. (2021). Formation and repair of unavoidable, endogenous interstrand cross-links in cellular DNA. *DNA Repair*, *98*, 103029. https://doi.org/10.1016/j.dnarep.2020.103029

Islam, S. M. A., Wu, Y., Díaz-Gay, M., Bergstrom, E. N., He, Y., Barnes, M., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Vangara, R., Steele, C. D., Otlu, B., … Alexandrov, L. B. (2021). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *BioRxiv*, *2*(11), 100179. https://doi.org/10.1016/j.xgen.2022.100179

Izzo, C., Annunziata, M., Melara, G., Sciorio, R., Dallio, M., Masarone, M., Federico, A., & Persico, M. (2021). The Role of Resveratrol in Liver Disease: A Comprehensive Review from In Vitro to Clinical Trials. *Nutrients*, *13*(3), 933. https://doi.org/10.3390/nu13030933

Juárez-Figueroa, U., Ayala-Zambrano, C., Reyes, P., & Frías, S. (2018). Origen y consecuencias de la inestabilidad genómica Síndromes de inestabilidad cromosómica. *Mensaje Bioquímico*, *42*, 64–80.

Kashiyama, K., Nakazawa, Y., Pilz, D. T., Guo, C., Shimada, M., Sasaki, K., Fawcett, H., Wing, J. F., Lewin, S. O., Carr, L., Li, T.-S., Yoshiura, K., Utani, A., Hirano, A., Yamashita, S., Greenblatt, D., Nardo, T., Stefanini, M., McGibbon, D., … Ogi, T. (2013). Malfunction of nuclease ERCC1-XPF results in diverse clinical manifestations and causes Cockayne syndrome, xeroderma pigmentosum, and Fanconi anemia. *American Journal of Human Genetics*, *92*(5), 807–819. https://doi.org/10.1016/j.ajhg.2013.04.007

Kee, Y., & D'Andrea, A. D. (2010). Expanded roles of the Fanconi anemia pathway in preserving genomic stability. *Genes & Development*, *24*(16), 1680–1694. https://doi.org/10.1101/gad.1955310

Keppler, F., Eiden, R., Niedan, V., Pracht, J., & Schöler, H. F. (2000). Halocarbons produced by natural oxidation processes during degradation of organic matter. *Nature*, *403*(6767), 298–301. https://doi.org/10.1038/35002055

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Koh, G., Degasperi, A., Zou, X., Momen, S., & Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. *Nature Reviews. Cancer*, *21*(10), 619–637. https://doi.org/10.1038/s41568-021-00377-7

Kulkarni, N., Alessandrì, L., Panero, R., Arigoni, M., Olivero, M., Ferrero, G., Cordero, F., Beccuti, M., & Calogero, R. A. (2018). Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, *19*(Suppl 10), 349. https://doi.org/10.1186/s12859-018-2296-x

Kumar, S., Warrell, J., Li, S., McGillivray, P. D., Meyerson, W., Salichos, L., Harmanci, A., Martinez-Fundichely, A., Chan, C. W. Y., Nielsen, M. M., Lochovsky, L., Zhang, Y.,

Li, X., Lou, S., Pedersen, J. S., Herrmann, C., Getz, G., Khurana, E., & Gerstein, M. B. (2020). Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell*, *180*(5), 915-927.e16. https://doi.org/10.1016/j.cell.2020.01.032

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lopez-Martinez, D., Liang, C.-C., & Cohn, M. A. (2016). Cellular response to DNA interstrand crosslinks: the Fanconi anemia pathway. *Cellular and Molecular Life Sciences : CMLS*, *73*(16), 3097–3114. https://doi.org/10.1007/s00018-016-2218-x

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The Hallmarks of Aging. *Cell*, *153*(6), 1194–1217. https://doi.org/https://doi.org/10.1016/j.cell.2013.05.039

Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., Royo, R., Ziccheddu, B., Puente, X. S., Avet-Loiseau, H., Campbell, P. J., Nik-Zainal, S., Campo, E., Munshi, N., & Bolli, N. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications*, *10*(1), 2969. https://doi.org/10.1038/s41467-019-11037-8

McCormick, R. F., Truong, S. K., & Mullet, J. E. (2015). RIG: Recalibration and interrelation of genomic sequence data with the GATK. *G3 (Bethesda, Md.)*, *5*(4), 655–665. https://doi.org/10.1534/g3.115.017012

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Milano, M. (2019). *Computing Languages for Bioinformatics: R. BT - Encyclopedia of Bioinformatics and Computational Biology - Volume 1* (pp. 199–205). https://doi.org/10.1016/b978-0-12-809633-8.20367-1

Mimaki, S., Totsuka, Y., Suzuki, Y., Nakai, C., Goto, M., Kojima, M., Arakawa, H., Takemura, S., Tanaka, S., Marubashi, S., Kinoshita, M., Matsuda, T., Shibata, T., Nakagama, H., Ochiai, A., Kubo, S., Nakamori, S., Esumi, H., & Tsuchihara, K. (2016). Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. *Carcinogenesis*, *37*(8), 817–826. https://doi.org/10.1093/carcin/bgw066

Molina, B., Ramos, S., & Frias, S. (2022). Anemia de Fanconi, Parte 1. Diagnóstico citogenético Fanconi anemia, Part 1. Cytogenetic diagnosis. *Acta Pediátrica de México*, *43*(2), 102–128.

Mouw, K. W., & D'Andrea, A. D. (2014). Crosstalk between the nucleotide excision repair and Fanconi anemia/BRCA pathways. *DNA Repair*, *19*, 130–134. https://doi.org/10.1016/j.dnarep.2014.03.019

Myers, K., Davies, S. M., Harris, R. E., Spunt, S. L., Smolarek, T., Zimmerman, S., McMasters, R., Wagner, L., Mueller, R., Auerbach, A. D., & Mehta, P. A. (2012). The clinical phenotype of children with Fanconi anemia caused by biallelic FANCD1/BRCA2 mutations. *Pediatric Blood & Cancer*, *58*(3), 462–465. https://doi.org/10.1002/pbc.23168

Nalepa, G., & Clapp, D. W. (2018). Fanconi anaemia and cancer: an intricate relationship. *Nature Reviews. Cancer*, *18*(3), 168–185. https://doi.org/10.1038/nrc.2017.116

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K.,

Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979–993. https://doi.org/10.1016/j.cell.2012.04.024

Niraj, J., Färkkilä, A., & D'Andrea, A. D. (2019). The Fanconi Anemia Pathway in Cancer. *Annual Review of Cancer Biology*, *3*, 457–478. https://doi.org/10.1146/annurev-cancerbio-030617-050422

Oliver, G. R., Hart, S. N., & Klee, E. W. (2015). Bioinformatics for clinical next generation sequencing. *Clinical Chemistry*, *61*(1), 124–135. https://doi.org/10.1373/clinchem.2014.224360

Omichessan, H., Severi, G., & Perduca, V. (2019). Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. *PloS One*, *14*(9), e0221235. https://doi.org/10.1371/journal.pone.0221235

Palovcak, A., Liu, W., Yuan, F., & Zhang, Y. (2017). Maintenance of genome stability by Fanconi anemia proteins. *Cell & Bioscience*, *7*, 8. https://doi.org/10.1186/s13578-016-0134-2

Peng, M., Xie, J., Ucher, A., Stavnezer, J., & Cantor, S. B. (2014). Crosstalk between BRCA-Fanconi anemia and mismatch repair pathways prevents MSH2-dependent aberrant DNA damage responses. *The EMBO Journal*, *33*(15), 1698–1712. https://doi.org/10.15252/embj.201387530

Pon, J. R., & Marra, M. A. (2015). Driver and passenger mutations in cancer. *Annual Review of Pathology*, *10*, 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312

Rawal, L., Panwar, D., & Ali, S. (2017). *Bioinformatics Databases: Implications in Human Health BT - Genome Analysis and Human Health* (L. Rawal & S. Ali (Eds.); pp. 109–132). Springer Singapore. https://doi.org/10.1007/978-981-10-4298-0_6

Río, P., Navarro, S., & Bueren, J. A. (2018). Advances in Gene Therapy for Fanconi Anemia. *Human Gene Therapy*, *29*(10), 1114–1123. https://doi.org/10.1089/hum.2018.124

Rodríguez, A., & D'Andrea, A. (2017). Fanconi anemia pathway. *Current Biology*, *27*(18), R986–R988. https://doi.org/https://doi.org/10.1016/j.cub.2017.07.043

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, *17*, 31. https://doi.org/10.1186/s13059-016-0893-4

Salvadores, M., Mas-Ponte, D., & Supek, F. (2019). Passenger mutations accurately classify human tumors. *PLoS Computational Biology*, *15*(4), e1006953. https://doi.org/10.1371/journal.pcbi.1006953

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

SigProfilerExtractor. (2019). https://osf.io/t6j7u/

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/https://doi.org/10.1002/cpmb.59

Spies, M., & Fishel, R. (2015). Mismatch repair during homologous and homeologous recombination. *Cold Spring Harbor Perspectives in Biology*, *7*(3), a022657. https://doi.org/10.1101/cshperspect.a022657

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C.,

Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

The Linux command line for beginners. (n.d.). In *Ubuntu*. https://ubuntu.com/tutorials/command-line-for-beginners

Trimmomatic Manual: V0.32. (n.d.). Trimmomatic

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J. L., Wong, Y. N. S., Rowan, A., Kanu, N., Al Bakir, M., Chambers, T., Salgado, R., Savas, P., Loi, S., Birkbak, N. J., Sansregret, L., Gore, M., Larkin, J., Quezada, S. A., & Swanton, C. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *The Lancet. Oncology*, *18*(8), 1009–1021. https://doi.org/10.1016/S1470-2045(17)30516-8

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., & 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. https://doi.org/10.1002/0471250953.bi1110s43

Van Hoeck, A., Tjoonk, N. H., van Boxtel, R., & Cuppen, E. (2019). Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer*, *19*(1), 457. https://doi.org/10.1186/s12885-019-5677-2

Venkitaraman, A. R. (2019). How do mutations affecting the breast cancer genes BRCA1 and BRCA2 cause cancer susceptibility? *DNA Repair*, *81*, 102668. https://doi.org/10.1016/j.dnarep.2019.102668

Wang, L. C., & Gautier, J. (2010). The Fanconi anemia pathway and ICL repair: implications for cancer therapy. *Critical Reviews in Biochemistry and Molecular Biology*, *45*(5), 424–439. https://doi.org/10.3109/10409238.2010.502166

Weber, L. W. D., Boll, M., & Stampfl, A. (2003). Hepatotoxicity and mechanism of action of haloalkanes: carbon tetrachloride as a toxicological model. *Critical Reviews in Toxicology*, *33*(2), 105–136. https://doi.org/10.1080/713611034

Wegman-Ostrosky, T., & Savage, S. A. (2017). The genomics of inherited bone marrow failure: from mechanism to the clinic. *British Journal of Haematology*, *177*(4), 526–542. https://doi.org/10.1111/bjh.14535

Williams, S. A., Wilson, J. B., Clark, A. P., Mitson-Salazar, A., Tomashevski, A., Ananth, S., Glazer, P. M., Semmes, O. J., Bale, A. E., Jones, N. J., & Kupfer, G. M. (2011). Functional and physical interaction between the mismatch repair and FA-BRCA pathways. *Human Molecular Genetics*, *20*(22), 4395–4410. https://doi.org/10.1093/hmg/ddr366

Wu, C.-H., Hsieh, C.-S., Chang, Y.-C., Huang, C.-C., Yeh, H.-T., Hou, M.-F., Chung, Y.-C., Tu, S.-H., Chang, K.-J., Chattopadhyay, A., Lai, L.-C., Lu, T.-P., Li, Y.-H., Tsai, M.-H., & Chuang, E. Y. (2021). Differential whole-genome doubling and homologous recombination deficiencies across breast cancer subtypes from the Taiwanese population. *Communications Biology*, *4*(1), 1052. https://doi.org/10.1038/s42003-021-02597-x