FACULTAD DE CIENCIAS AGRARIAS

UNIVERSIDAD NACIONAL DE ROSARIO

IDENTIFICACIÓN DE VARIANTE GENÉTICA CAUSAL PARA SÍNDROMES DE CÁNCER COLORRECTAL HEREDITARIO: SECUENCIACIÓN MASIVA EN PARALELO Y APLICACIÓN DE HERRAMIENTAS BIOINFORMÁTICAS

Dra. ANDREA CONSTANZA MAYORDOMO

TRABAJO FINAL PARA OPTAR AL TÍTULO DE ESPECIALISTA EN BIOINFORMÁTICA

DIRECTOR: Dr. Adrián Turjaski

CO- DIRECTOR: Dr. Javier Murillo

AÑO

2023

IDENTIFICACIÓN DE VARIANTE GENÉTICA CAUSAL PARA SÍNDROMES DE CÁNCER COLORRECTAL HEREDITARIO: SECUENCIACIÓN MASIVA EN PARALELO Y APLICACIÓN DE HERRAMIENTAS BIOINFORMÁTICAS

Dra. ANDREA CONSTANZA MAYORDOMO

Lic. en Biología Molecular – Universidad Nacional de San Luis

Este Trabajo Final es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido previamente presentado para la obtención de otro título en esta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en_____, durante el período comprendido entre _____, bajo la dirección de _____.

Nombre y firma del autor

Nombre y firma del director

Nombre y firma del Co-director

Defendida.....de 20__.

Agradecimientos

Quiero agradecer a las Facultades de Ciencias Bioquímicas y Farmacéuticas y de Ciencias Agrarias de la Universidad Nacional de Rosario por la posibilidad de formarme académicamente. A la comisión académica y a todos los profesores o profesionales que brindaron su conocimiento en las diferentes materias de la especialidad.

A mis directores Adrian Turjaski y Javier Murillo por darme un lugar en su grupo de trabajo para poder realizar mi trabajo final y la paciencia, sobre todo.

Al grupo de ProCanHe, especialmente a Walter Pavicic quien me ayudó a descubrir este nuevo camino académico en la bioinformática.

A todos mis compañeros de la especialidad, especialmente a Maribel Vallasciani quien me brindó una amistad y un espacio donde quedarme en las cursadas.

A mi familia y amigos por estar acompañándome en todos mis proyectos nuevos.

"Sólo podemos dar lo que ya hemos dado. Sólo podemos dar lo que ya es del otro... ¡Qué misterio es una dedicatoria, una entrega de símbolos!"

Jorge Luis Borges

Presentaciones a congresos

- ✓ ***"Identification of the predisposing genetic variant for inherited colorectal cancer syndromes in Argentina***" Andrea C. Mayordomo; Jonathan J. Zaiat; Belén Cerliani; Tamara Piñero; Romina Cajal; Mariana Coraglio; Karina Collia Ávila; Alejandro Gutiérrez; Carlos Vaccaro; Marcelo Marti; Walter Pavicic; Javier I. Murillo; Adrián Turjanski. **1st Congress of Women in Bioinformatics and Data Science LA 2020**.

- ✓ ***"Identification of the predisposing genetic variant for inherited colorectal cancer syndromes in Argentina***" Andrea C. Mayordomo; Jonathan J. Zaiat; Belén Cerliani; Tamara Piñero; Romina Cajal; Mariana Coraglio; Karina Collia Ávila; Alejandro Gutiérrez; Carlos Vaccaro; Marcelo Marti; Walter Pavicic; Javier I. Murillo; Adrián Turjanski. **9° Jornada Integral de Investigación y 18° Premio Prof. Dr. José Tessler Año 2021.**

Abreviaturas

ACMG: Colegio Americano de Genética Médica y Genómica

ADN: ácido desoxirribonucleico

APC: Adenomatous poliposis coli

ARN: ácido ribonucleico

BWA: Burrows-Wheeler Aligner

CCHNP: Cáncer Colorrectal Hereditario No Polipósico

CCR: cáncer colorrectal

CNVs: variantes del número de copias

GRCh37: Genome Reference Consortium Human Build 37

InDel: Variante de inserción-deleción

LP: variante probablemente patogénica

MLPA: amplificación de sondas dependiente de ligandos múltiples

MMR: sistema de reparación por mal apareamiento de bases

MUTHY: gen mutY DNA glycosylase

NGS: técnicas de secuenciación de nueva generación

P: variante patogénica

PAF: Poliposis adenomatosa familiar

PAFA: Poliposis adenomatosa familiar atenuada

PCR: reacción en cadena de la polimerasa

PCR-RFLP: fragmentos de restricción de longitud polimórfica

POLD1: DNA polymerase delta 1, catalytic subunit

POLE: gen DNA polymerase epsilon, catalytic subunit

PPAP: Poliposis Asociada a la actividad reparadora de las Polimerasas

Pro.Can.He: Programa de Cáncer Hereditario

REM: Registro de Epidemiología Molecular

RT: transcripción reversa

SL: Síndrome de Lynch

SNV: variante puntual

VUS: variante de significado incierto

Resumen

El cáncer colorrectal (CCR) tiene una elevada incidencia y mortalidad a nivel mundial y en Argentina es la segunda causa de muerte por cáncer (10,6%). Los Síndromes de CCR hereditario se dividen en Cáncer Colorrectal Hereditario No Polipósico (CCHNP) y síndromes de Poliposis, siendo el síndrome de Lynch (SL) y la Poliposis Adenomatosa Familiar (PAF) las formas de CCR hereditario más comunes. La utilización de las técnicas de secuenciación de nueva generación (NGS) para guiar la prevención, el diagnóstico y el tratamiento de enfermedades basadas en los genes individuales de una persona o una familia, su medio ambiente y su estilo de vida, se conoce con el nombre de medicina de precisión. A pesar de este nuevo paradigma de la Medicina de Precisión son pocos los reportes y la aplicación de estas técnicas en América Latina. Particularmente, el grupo de investigación REM-ProCanHe del cual formo parte, lleva adelante desde 1996 el desarrollo de investigación clínica para la identificación temprana de CCR en *pos* de alcanzar la medicina de precisión. La vinculación internacional de este grupo con especialistas en la temática de la universidad de Helsinki (Finlandia) ha permitido realizar técnicas de secuenciación masiva en paralelo en muestras argentinas. El objetivo general del presente trabajo fue aplicar herramientas bioinformáticas para realizar un análisis preciso y rápido con el objetivo de identificar a nivel germinal la variante causal asociada con aumento de susceptibilidad a desarrollar CCR hereditario, partiendo de resultados genómicos derivados de secuenciación de nueva generación. Se estudiaron 21 casos de pacientes clínicamente diagnosticados con síndrome de Poliposis provenientes del Registro de Poliposis Adenomatosa Familiar del Hospital de Gastroenterología *Dr. Carlos Bonorino Udaondo*. Se analizaron de manera secuencial por i) secuenciación con el método de Sanger del exón 15 del gen APC (directamente relacionado con PAF); ii) luego con la técnica de MLPA para evaluar presencia de grandes rearreglos; iii) finalmente, para aquellas muestras aun negativas sin alteración genética-causal identificada (mediante métodos i y ii) se realizó la secuenciación del exoma completo. Como resultado, a través de la técnica de secuenciación por Sanger, en 6 casos identificamos la variante genético-causal en el gen APC, siendo todas variantes novel, las cuales podrían ser verificadas a nivel funcional por otras técnicas o incorporar mayor información genética ya sea de muestras de la familia en estudio u otras muestras independientes que presenten la misma variante. Además, para otros 4 casos y mediante la secuenciación del exoma completo pudimos identificar la variante genético-causal candidata para predisposición en genes con asociación establecida para la patología. Específicamente, 2 casos presentaron alteración en genes relacionados con la patología, pero por el fenotipo clínico no pudieron confirmarse como genético-causal. Los otros 2 casos presentaron variantes en dos genes que se relacionan con el funcionamiento homeostático del intestino por lo cual necesitamos mayor evidencia sobre los mismos para poder adjudicarlos como genético-causal. Po otra parte, no encontramos genes que se relacionan directamente con el fenotipo en 7 casos los cuales todavía siguen en estudio. A partir de los datos obtenidos en este trabajo, se evidenciaron los parámetros, los pasos a seguir para llevar el resultado obtenido desde el secuenciador hasta el archivo VCF, como así también el diseño de un *pipeline* para la priorización de variante. Se espera que las herramientas y procedimientos obtenidos en el presente trabajo contribuyan de manera significativa en la medicina de precisión posibilitando el desarrollo de nuevas estrategias para el estudio del Síndrome de CCR hereditario.

Abstract

Colorectal cancer (CRC) has a high incidence and mortality worldwide and, in Argentina, it is the second cause of death from cancer (10.6%). Hereditary CRC Syndromes are divided into Hereditary Non-Polyposis Colorectal Cancer (HNPCC) and Polyposis syndromes, with Lynch syndrome (LS) and Familial Adenomatous Polyposis (FAP) being the most common forms of hereditary CRC. The use of NGS techniques to guide the prevention, diagnosis, and treatment of disease based on a person's or family's individual genes, environment, and lifestyle is known as precision medicine. Despite this new paradigm of precision medicine, there are few reports and applications of these techniques in Latin America. In particular, the REM-ProCanHe research group, of which I am a member, has been carrying out precision medicine for the early detection of CRC since 1996. The international linkage of this group with specialists in the subject from the University of Helsinki (Finland) has made it possible to carry out massive parallel sequencing techniques in Argentine samples. Therefore, the general objective of this work was to apply bioinformatic tools to perform a precise and rapid analysis to identify at the germinal level the causal variant associated with increased susceptibility to develop hereditary CRC, based on genomic results derived from new sequencing techniques. 21 samples were taken from patients diagnosed with polyposis syndromes from the Family Adenomatous Polyposis Registry of the *Dr Carlos Bonorino Udaondo* Hospital of Gastroenterology, first performing the sequencing by the Sanger method of exon 15 of the APC gene (directly related to PAF), then by the MLPA technique and finally those negative samples for both methods we performed whole exome sequencing. We were able to find, through the Sanger sequencing technique, 6 samples with a genetic-causal variant in the APC gene with a “novel” variant, which could be verified at a functional level by other techniques or incorporate more genetic information either from samples of the family under study or other independent samples that present the same variant. Moreover, by whole exome sequencing technique: a) 4 samples with a genetic-causal variant in genes related to the pathology were found; b) 2 samples in genes related to the pathology that cannot be identified as genetic-causal because of the clinical phenotype; c) 2 samples present variants in two genes that are related to the homeostatic functioning of the intestine, for which we need more evidence about them to be able to adjudicate them as genetic-causal. However, we did not find genes that are directly related to the phenotype in 7 samples which are still under study. From the data obtained in this work, the parameters were evidenced, and the steps to follow to take the result obtained from the sequencer to the VCF file, as well as the design of a pipeline for variant prioritization. Therefore, based on the tools obtained in this work, will allow us to apply for precision medicine in our research group, enabling new strategies to study and analyze cases of hereditary CRC syndrome.

Índice

1-Introducción

1.1 Síndromes de cáncer colorrectal hereditario

El cáncer colorrectal (CCR) tiene una elevada incidencia y mortalidad a nivel mundial. En general, el CCR ocupa el tercer lugar en términos de incidencia, pero el segundo en términos de mortalidad. Dentro de América, la argentina se ubica entre los países con tasas de incidencias altas (74/100.000 habitantes) y, a su vez, el CCR es la segunda causa de muerte por cáncer en nuestro país (10,6%) (Sung et al., 2021).

El 95% de los CCR son adenocarcinomas, estando precedidos en la mayoría (80-90%) por lesiones preneoplásicas: los pólipos adenomatosos o adenomas (Bujanda, Cosme, Gil, & Arenas-Mirave, 2010; Sack & Rothman, 2000). Actualmente, se estima que 1 en 18 individuos (5,5%) desarrollará CCR en su vida. En la mayoría de los casos el CCR ocurre de manera esporádica, y como forma familiar o hereditaria en 1/3 de los pacientes. Estos últimos, engloban a los individuos que presentan riesgo moderado o alto de desarrollo de la enfermedad; haciendo necesario fomentar los estudios para mejorar su identificación, diagnóstico y tratamiento.

Los Síndromes de CCR hereditario se dividen en Cáncer Colorrectal Hereditario No Polipósico (CCHNP) y síndromes de Poliposis, siendo el síndrome de Lynch (SL) y la Poliposis Adenomatosa Familiar (PAF) las formas de CCR hereditario más comunes. Se asocian con alteración patogénica en los genes del sistema de reparación por mal apareamiento de bases (MMR del inglés, los que incluyen los genes *MLH1*, *MSH2*, *MSH6* y *PMS2*), o el gen poliposis adenomatosa coli (*APC*), respectivamente (L. J. C. G. Valle & Hepatology, 2017). Además, el SL es una condición hereditaria que incrementa la probabilidad de presentar, conjuntamente al CCR, cáncer de endometrio, ovario, mama, estómago, entre otros (Dominguez-Valentin et al., 2020). De igual manera, PAF aumenta el riesgo de presentar tumores de estómago, intestino delgado, páncreas, vías biliares, hígado, entre otras manifestaciones clínicas extracolónicas (por ejemplo, osteoma, tumor desmoide, etc.) (Byrne & Tsikitis, 2018).

1.2 Poliposis

Dentro de los síndromes de poliposis, PAF es el más frecuente y es la segunda forma más común de predisposición a desarrollar CCR familiar (Burt et al., 2004; Galiatsatos, Foulkes, & ACG, 2006). Se divide en dos fenotipos ligeramente diferentes, PAF clásica y PAF atenuada; siendo la presencia de alteraciones germinales en el gen *APC*, la causa principal de

ambos síndromes. Ambos son heredados de un modo autosómico dominante (Jasperson, Tuohy, Neklason, & Burt, 2010)

Los pacientes diagnosticados con PAF clásica presentan desde cientos hasta miles de pólipos adenomatosos colorrectales que, de no mediar un diagnóstico y tratamiento precoz, desarrollaran CCR en el 100% de casos (Groden et al., 1991; Nishisho et al., 1991), es decir, presenta una heredabilidad de alta penetrancia. Si bien la variante genética es portada desde el nacimiento, el desarrollo de pólipos suele iniciarse en la pubertad, mientras que los síntomas suelen aparecer en la tercera década de la vida y el desarrollo de CCR entre los 30 y los 35 años, lo cual es significativamente más joven que los CCR esporádicos (Syngal et al., 2015). El cambio genético responsable se detecta entre el 70 y el 90% de los casos. Sin embargo, entre el 15 y el 40% de los pacientes con PAF clínica e histológicamente certificada no tienen antecedentes familiares, produciéndose los mismos por alteraciones de *novo* (Cairns et al., 2010; Hegde, Ferber, Mao, Samowitz, & Ganguly, 2014; Leoz, Carballal, Moreira, Ocaña, & Balaguer, 2015).

Otra variante fenotípica y menos severa de la poliposis clásica, denominada atenuada o atípica (PAFA), se caracteriza por su aparición a una edad más avanzada, por la presencia de menor cantidad de adenomas colorrectales y ocurre en aproximadamente un 8% de las familias con diagnóstico de PAF. El riesgo de padecer CCR es superior al 80% a lo largo de la vida (Burt et al., 2004; Newton et al., 2012; Nieuwenhuis et al., 2007).

Otro fenotipo encontrado de síndrome de PAF es la Poliposis Asociada al gen *mutY DNA glycosylase* (*MUTYH*) o PAM, la cual presenta una herencia autosómica-recesiva causada por alteraciones bialélicas en el gen. Se asemeja a la forma atenuada de poliposis y se asocia con la aparición de numerosos pólipos (15-100) a nivel colorrectal. A su vez, presenta un riesgo de desarrollar CCR del 80% y la edad promedio de aparición es aproximadamente a los 40 años (Cleary et al., 2009; Lubbe, Di Bernardo, Chandler, & Houlston, 2009).

Otro síndrome de poliposis de tipo atenuado descripto recientemente, y asociado con predisposición a desarrollar cáncer colorrectal y/o de endometrio, es la Poliposis Asociada a la actividad reparadora de las Polimerasas o PPAP (del inglés, Polymerase Proofreading-Associated Polyposis). Alteraciones de línea germinal en el gen *DNA polymerase delta 1, catalytic subunit* (*POLD1*) se asocian con el desarrollo de cáncer endometrial, mientras que las mutaciones en el gen *DNA polymerase epsilon, catalytic subunit* (*POLE*) están asociadas con cáncer colorrectal (Briggs & Tomlinson, 2013).

Finalmente, mediante el empleo de técnicas de secuenciación de última generación (NGS por sus siglas en inglés), en los últimos años se han detectado diferentes alteraciones genéticas asociadas con formas poco frecuentes de poliposis adenomatosa colorrectal como Síndrome de poliposis serrada, síndrome de Peutz-Jeghers, síndrome de poliposis juvenil y síndromes de tumor harmartoma PTEN (Dinarvand et al., 2019; L. Valle et al., 2019).

1.3 Diagnóstico de Síndromes de poliposis

El estudio diagnóstico inicial de poliposis incluye endoscopia y/o colonoscopía, la cual permite conocer el número y distribución de los pólipos, el tamaño de los mismos y determinar la realización de una polipectomía (cirugía para extraer un pólipo). A partir de esta última, se realiza la histología que permite una clasificación de los pólipos. Junto con la historia familiar se puede recrear un familigrama, que permite entender el tipo de herencia: dominante, recesiva de la condición o esporádica. A su vez, se observa la presencia de manifestaciones extracolónicas que ayudan al diagnóstico. Finalmente, con el diagnóstico determinado se realiza el estudio genético correspondiente para buscar la alteración genética que causa el fenotipo, en caso que sea hereditaria se lleva a cabo en las personas asintomáticas relacionadas al probando, que se encuentran en riesgo, en el contexto del asesoramiento genético. Los pacientes con poliposis deben ser vistos en intervalos regulares en un centro interdisciplinario especializado. En la **figura 1** se resumen los criterios diagnósticos (Aretz, 2010).

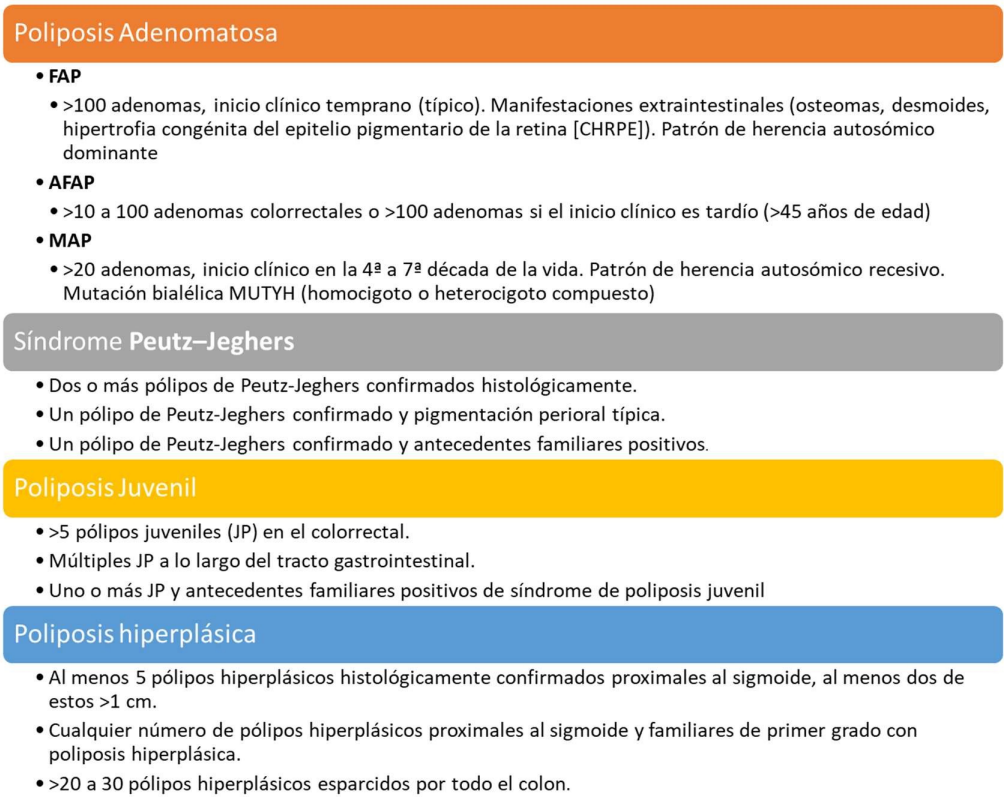

Figura 1. Criterio diagnóstico para la clasificación de síndrome de poliposis

1.4 Impacto de la biología molecular en el diagnóstico clínico

La demanda de nuevas técnicas de diagnóstico en medicina tiene su origen en los problemas clínicos que requieren con urgencia un método de diagnóstico rápido, idealmente con una alta sensibilidad y especificidad, que pueda ser accesible a la población. La investigación básica junto con el desarrollo tecnológico, ha permitido el diseño de instrumentos y métodos para responder a estas necesidades clínicas. El advenimiento de la técnica molecular: reacción en cadena de la polimerasa (PCR), ha permitido, principalmente, el diagnóstico de enfermedades infecciosas (50-60%).

Variantes de la misma técnica como PCR-RFLP (fragmentos de restricción de longitud polimórfica) permite la detección de polimorfismos genéticos, RT (transcripción reversa)-PCR permite detectar virus de ARN, entre otras variantes.

La PCR ha sido imprescindible para ciertos diagnósticos de enfermedades, pero algunas patologías o condiciones requieren conocer los nucleótidos que constituyen un fragmento de ADN en particular (gen) que se relaciona con la patología. De allí que la técnica de secuenciación por Sanger es una técnica utilizada para el diagnóstico de ciertas enfermedades

hereditarias en los genes que se relacionan directamente con la patología o condición a abordar, con longitudes entre 700 y 1000 pares de bases (pb). Es importante destacar que la secuenciación por Sanger permitió alcanzar hitos tan relevantes para la genética humana como la secuenciación del primer genoma humano en el año 2001 (Steward, 2001; Venter et al., 2001), o la caracterización del primer haplotipo humano por el consorcio HapMap (Altshuler, Donnelly, & Nature, 2005).

Luego, las técnicas de secuenciación masiva de próxima generación (NGS), utilizada en paneles de genes, exomas y genoma completo ha permitido obtener mayor información en poco tiempo sobre una patología/condición en particular (Del Vecchio et al., 2017; Kim, Lee, & Chung, 2013), principalmente debido a que el costo por base secuenciada ha disminuido a lo largo de los años, como se puede observar en la **figura 2** (KA, 2021). La secuenciación por Sanger pasó a ser la técnica *gold standard* para validar el resultado de las mismas. Todos los resultados obtenidos por NGS no son de criterio diagnóstico, sino que son utilizados para acompañar el diagnóstico del médico o es parte de la investigación médica.

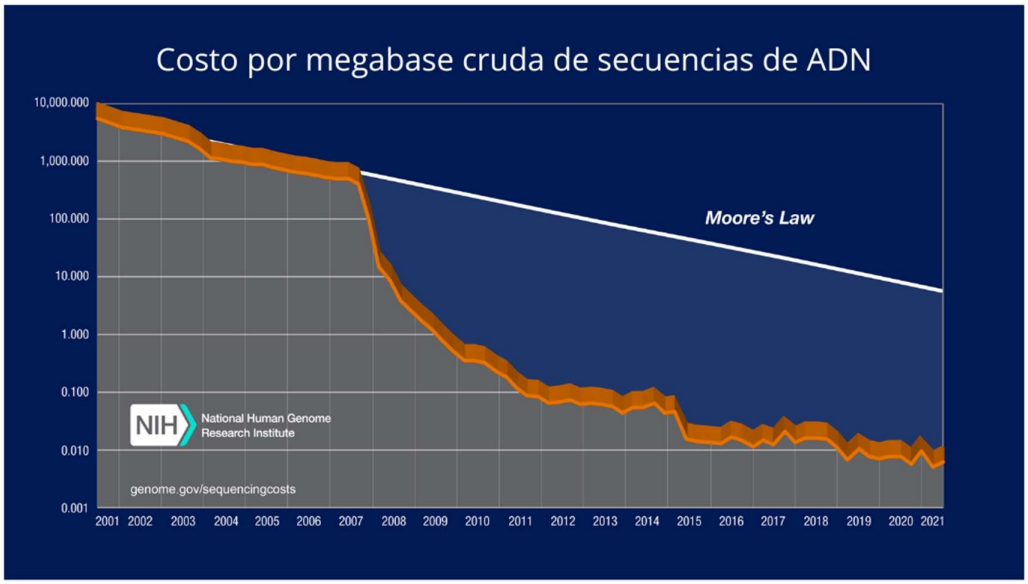

Figura 2. **Costo de Secuenciación por megabase**. En el eje de las Y se utiliza una escala logarítmica. Se observa una disminución pronunciada en el año 2008 que es el momento en que se realizó la transición de la técnica de Sanger por NGS. El gráfico también muestra datos hipotéticos que reflejan la Ley de Moore, la cual describe una tendencia a largo plazo en la industria del hardware informático que implica la duplicación del "poder de cómputo" cada dos años. www.genome.gov/sequencingcostsdata (KA, 2021).

1.5 Uso de paneles de genes en CCR hereditario

Existe una amplia batería de paneles de genes para realizar un diagnóstico de CCR en el mercado (Lorans, Dow, Macrae, Winship, & Buchanan, 2018) pero, hasta el momento, no hay ninguno establecido como óptimo, ya que en algunos casos se incluyen genes con baja y/o alta penetrancia o genes que no tienen una evidencia biológica demostrada sobre estos síndromes de CCR hereditario.

Recientemente, el Grupo Colaborativo de las Américas en Cáncer Gastrointestinal Hereditario publicó una declaración de posicionamiento sobre las pruebas de paneles de genes múltiples, recomendando un conjunto de 11 genes primordiales en la cual se incluyen *MLH1*, *MSH2*, *MSH6*, *PMS2*, *EPCAM*, *APC*, *BMPR1A*, *MUTYH*, *PTEN*, *STK11* y *SMAD4*. A su vez, en este escrito también se recomienda un conjunto adicional de 16 genes los cuales tienen las siguientes características: Genes adicionales: (a) aumento del riesgo de cáncer colorrectal (CCR) de bajo a moderado (*ATM*, *CHEK2*, *TP53*); (b) datos preliminares pero limitados sobre el riesgo de CCR (*GREM1*, *POLD1*, *POLE*, *AXIN2*, *NTHL1*, *MSH3*, *GALNT12*, *RPS20*); (c) genes con variantes patogénicas encontradas en pacientes con CCR que son procesables con respecto a otros cánceres, pero no se ha probado la causalidad del CRC (*BRCA1*, *BRCA2*, *CDKN2A*, *PALB2*). La adición de estos genes al panel se puede considerar en circunstancias específicas (Heald et al., 2020).

En la **figura 3** (modificada de (Olkinuora, Peltomäki, Aaltonen, & Rajamäki, 2021)) se pueden observar los genes que han sido relacionados directamente con los diferentes síndromes de CCR hereditario y aquellos que aún se encuentran en estudio. Esta línea de tiempo tiene en cuenta la tecnología que permitió identificarlos y la temporalidad de los mismos. Como se puede observar, a partir de la llegada de la técnica de NGS se han encontrado un mayor número de genes que han sido identificados como causales del fenotipo estudiado.

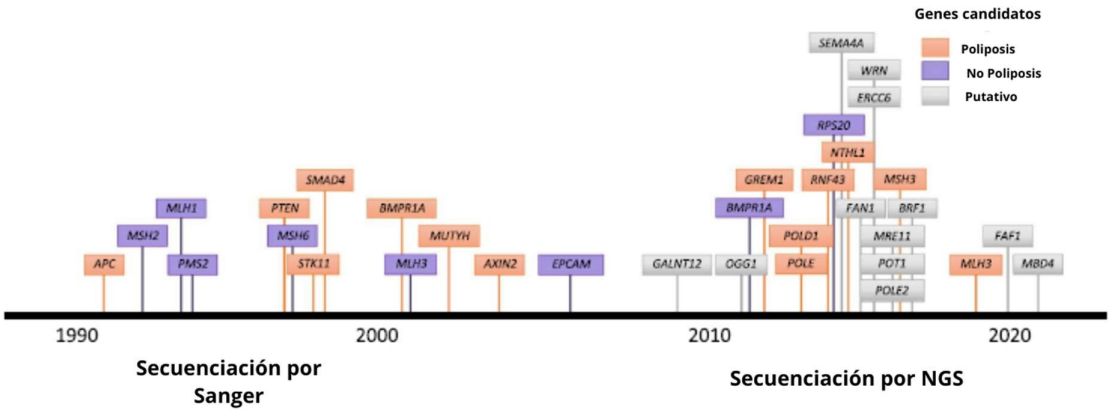

Figura 3. Línea de tiempo de los genes descubiertos para síndromes CCR hereditarios. Los genes en rosa son los relacionados a PAF, en violeta los asociados a HPNCC y en gris se encuentran en estudio. Adaptado de (Olkinuora et al., 2021).

1.6 Clases de variantes genéticas

A partir del Proyecto del Genoma Humano (1990 – 2003) el cual fue una gran iniciativa de colaboración internacional se mapeó y secuenció el genoma humano por primera vez, obteniéndose de esta manera el primer genoma de referencia humano. A medida que se fue adquiriendo mayor información genética se fue ensamblando con mayor precisión e información las diferentes partes del genoma, logrando así diferentes versiones del mismo llegando hasta la versión actual GRCh38.p14 liberada este año 2022 (PRJNA31257).

Una variante -alteración- genética se puede obtener contrastando una secuencia de ADN con la referencia. De esta manera podemos obtener diferentes tipos o clases de variantes genéticas (**figura 4**), las cuales pueden ser: de un solo nucleótido o puntuales (SNVs), son variaciones de la secuencia de ADN en las que se altera un solo nucleótido (A, T, G o C); variantes de inserción-eliminación (indels) ocurren cuando uno o más pares de bases están presentes en algunos genomas, pero ausentes en otros. Están generalmente compuestos de sólo unas pocas bases, pero pueden tener más de 80 kb de longitud. Una variante de inversión es aquella en la que se invierte el orden de los pares de bases en una sección definida de un cromosoma. Las variantes del número de copias (CNVs) ocurren cuando son idénticas o casi idénticas (Frazer, Murray, Schork, & Topol, 2009).

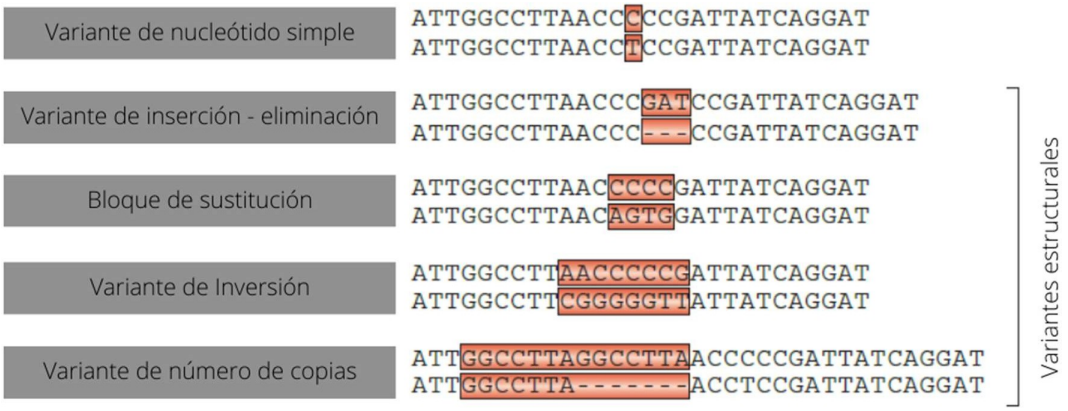

Figura 4. Clases de variantes genéticas. En esta figura se muestran los tipos de variantes encontradas cuando se realiza una comparación entre una secuencia desconocida contra una secuencia de referencia. Las variantes pueden clasificarse en: variantes de nucleótido simple, de inserción-eliminación, bloque de sustitución, variante de inversión y variante de número de copias. Las cuatro últimas son conocidas como variantes estructurales. Figura modificada de (Frazer et al., 2009)

1.7 Interpretación de las variantes y clasificación

El Colegio Americano de Genética Médica y Genómica (ACMG, por sus siglas en inglés) desarrolló previamente una guía para la interpretación de variantes de secuencia. Este presenta un sistema de puntuación con cinco clases: benigna, probablemente benigna, variante de significado desconocido o incierto (VUS), probablemente patogénica y patogénica (**figura 5**). Para llegar a la clasificar una variante en una clase específica se tiene en cuenta la frecuencia de la misma obtenida de bases de datos poblacionales y clínicas, la evidencia demostrada en trabajos científicos publicados, la herencia genética, la historia clínica, el uso de herramientas *in silico* para demostrar el efecto sobre el gen, entre otros (Richards et al., 2015).

En todos los genes relacionados directamente con PAF y CCHNP antes detallados, se han descripto una o más variantes patogénicas en línea germinal que presentan alta penetrancia y, por ello, son informadas como genético-causales de estos síndromes (fenotipo). Por el contrario, en otros casos que presentan una clínica severa aún no se han identificado genes causales o variantes en genes relacionados que puedan asociarse directamente con la sintomatología del paciente (Balmaña et al., 2016).

Figura 5. Clasificación de la variante según ACMG. Esta figura permite observar la clasificación en código de colores del rojo al verde creando un gradiente desde una variante patogénica a una benigna.

1.8 Medicina de precisión en América Latina

La utilización de las técnicas de NGS para guiar la prevención, el diagnóstico y el tratamiento de enfermedades basadas en los genes individuales de una persona o una familia, su medio ambiente y su estilo de vida, se conoce con el nombre de medicina de precisión.

A pesar de este nuevo paradigma de la Medicina de Precisión son pocos los reportes y la aplicación de estas técnicas en América Latina. Sólo Brasil (en la región de Sao Paulo), Chile y Argentina han desarrollado programas de detección temprana de CCR basados en estudios epidemiológicos y aplicación de estudios moleculares y genéticos con base en NGS. Particularmente, el grupo de investigación REM-ProCanHe del cual formo parte, lleva adelante desde 1996 el desarrollo de investigación clínica para la identificación temprana de CCR en *pos* de alcanzar la medicina de precisión.

Sin embargo, estos programas se encuentran aún en fases iniciales de implementación y cubren solo grandes áreas urbanas. Cuba, Ecuador, México, Puerto Rico y Uruguay también han iniciado la implementación de dichos programas, pero aún en etapas menos avanzadas. Las circunstancias económicas de los países de América Latina y El Caribe, junto con los sistemas de salud a menudo débiles, los recursos humanos de salud limitados y la baja implementación de políticas locales en materia de salud y desarrollo del área en el sistema de investigación aplicada en medicina, en particular para detección del CCR hereditario, son algunos de los factores que inhiben el desarrollo y la transferencia de programas de detección, tratamiento y seguimiento de pacientes con cáncer en estos países (Schreuders et al., 2015).

1.9 Antecedentes del grupo REM-ProCanHe

Esta temática presentada tiene una gran complejidad debido a los diferentes fenotipos que presenta la condición clínica abordada, lo cual requiere implementar diferentes estrategias para ayudar a optimizar el tratamiento y manejo clínico de los pacientes con estos síndromes y sus familiares en riesgo. En este último tiempo, los estudios se han focalizado en los genes directamente relacionados o en la búsqueda de genes candidatos genético-causales para estos síndromes.

Desde el año 2019 el grupo REM-Pro.Can.He del cual formo parte, a través del trabajo de investigación desarrollado por el Dr. Walter Pavicic, director de mi beca posdoctoral, en conjunto con el Dr. Carlos A. Vaccaro, ambos referentes en el estudio de Síndromes de CCR Hereditarios, en colaboración estrecha con el equipo liderado por la Dra. Päivi Peltomäki de la Universidad de Helsinki (Biomedicum Helsinki) ha secuenciado exomas completos junto con estudios de MLPA (para analizar presencia de grandes rearreglos genómicos) un total de 38 casos de pacientes argentinos. A su vez, la colaboración internacional con las Dras. Mev Dominguez-Valentin (Instituto de Investigación del cáncer, Oslo, Noruega) y Alexandra Martins han permitido realizar análisis funcionales de variantes identificadas como VUS (Piñero et al., 2020). Además, existe una vinculación a nivel nacional con el grupo clínico dirigido por el Dr. Alejandro Gutiérrez, del Hospital de Gastroenterología *Dr. Carlos Bonorino Udaondo* (CABA, ARG). Este último grupo está ligado a la atención, diagnóstico y seguimiento de pacientes con síndrome de Poliposis; posee el mayor número de casos a nivel nacional y el registro de casos PAF más grande e importante de Latinoamérica.

Como consecuencia de i) contar con una base de datos con una gran cantidad de variantes genéticas e información clínicas, ii) la posibilidad de realizar estudios genéticos especializados como la técnica de exoma completo por NGS (producto de la colaboración internacional), y iii) la compra de una computadora de características informáticas suficientes para desarrollar análisis bioinformáticos avanzados, se plantearon los siguientes objetivos para el presente trabajo.

2-Objetivos

2.1- OBJETIVO GENERAL

Aplicar herramientas bioinformáticas para realizar un análisis preciso y rápido con el objetivo de identificar a nivel germinal la variante causal asociada con aumento de susceptibilidad a desarrollar CCR hereditario, partiendo de resultados genómicos derivados de secuenciación de nueva generación.

2.2- OBJETIVOS ESPECÍFICOS

- ✓ Obtener a partir de datos alineados (BAM) aquellos datos procesados vinculados a variantes con datos biológicos (VCF)
- ✓ Priorizar genes que se relacionan directa o indirectamente con la patología para la búsqueda de variantes
- ✓ Clasificar y anotar las variantes encontradas
- ✓ Construir un pipeline adecuado para el análisis bioinformático a partir de los resultados obtenidos

3-Materiales y Métodos

3.1-Muestras

Se recolectó sangre periférica de 21 pacientes clínicamente diagnosticados con síndromes de poliposis provenientes del registro de poliposis adenomatosa familiar del hospital de Gastroenterología *Dr. Carlos Bonorino Udaondo*. Este registro se encuentra coordinado por médicos especialistas, las muestras siguen los protocolos y consensos determinados internacionalmente para cada síndrome, así como también los estándares éticos impuestos por el Hospital. Este proyecto se encuentra enmarcado en el protocolo principal aprobado en la institución beneficiaria titulado "Estudio molecular de marcadores de susceptibilidad para cáncer colorrectal hereditario y tumores asociados. Incidencia y estratificación en la población argentina" (Protocolo N.º 5469). Las muestras fueron congeladas hasta su posterior utilización en el Biobanco de muestras del Registro de Epidemiología Molecular (REM) del grupo Pro.Can.He, debidamente incorporadas a tal registro (anonimizadas) cumpliendo con todos los requerimientos éticos pertinentes a tal fin; las mismas fueron conservadas junto con el respectivo diagnóstico clínico y patológico.

3.2 Extracción de ADN

Se obtuvo el material genético mediante el protocolo, presentado a continuación, para aumentar la obtención del mismo a partir de muestras congeladas.

Día 1

1. Se colocaron 200 µl de sangre descongelada en un tubo de 1,5 ml, manteniendo las mismas en hielo.

2. Se agregó 800 µl de Buffer A (ver anexo).
3. Las muestras fueron centrifugadas durante 30 minutos a 3.500 rpm a 4ºC. Luego se descartó el sobrenadante.
4. Se agregó 1 ml de buffer A y luego el pellet fue resuspendido con vortex
5. Las muestras fueron centrifugadas por 30 minutos a 3.500 rpm a 4ºC. Posteriormente, se descartó el sobrenadante.
6. Se repitieron los pasos 4 y 5 hasta obtener una solución transparente.
7. Se agregó 350 µl de buffer de digestión (ver anexo). Luego se agregó a cada muestra una concentración final de 10mg/ml de proteinasa K. Finalmente, el pellet fue resuspendido con vortex.
8. Las muestras fueron incubadas durante 2 horas a 55ºC. Posteriormente, las muestras se dejaron durante toda la noche (ON) a 37ºC.

Día 2

1. Se agregó a cada tubo un volumen (500 µl) de LiCl 5 M. Posteriormente, se mezcló por inversión durante 1 minuto,
2. Se agregaron dos volúmenes (1 ml) de mezcla SEVAG. Seguidamente, las muestras fueron agitadas en agitador rotatorio durante 30 min.
3. Las muestras fueron centrifugadas a 14.000 rpm durante 15 min. Posteriormente, se recuperó el sobrenadante de la fase liquida (contiene el ADN).
4. Se agregó a cada tubo 2 volúmenes (1 ml) de isopropanol frío (o en su defecto etanol absoluto). Posteriormente, se mezcló por inversión.
5. Luego las muestras fueron centrifugadas a 14.000 rpm durante 15 minutos. Posteriormente, se descartó el sobrenadante por inversión.
6. El pellet de DNA fue lavado con 1 volumen (500 µl) de Etanol 70% frío. Posteriormente, se centrifugaron las muestras a 14.000 rpm por 15 min. Finalmente, se descartó el sobrenadante.
7. Los tubos fueron tapados con Parafilm perforado. Posteriormente, fueron secados al aire en bloque térmico (temperatura aproximada: 50ºC, durante 20 min.).
8. El ADN obtenido fue re-suspendido en buffer TE. Posteriormente, las muestras fueron guardadas a - 20ºC hasta su primer uso.

Las muestras, junto con las alícuotas de trabajo se encuentran codificadas y almacenadas en Freezer -80ºC en el Biobanco de muestras REM-Pro.Can.He.

El promedio del rendimiento obtenido con el siguiente protocolo fue el siguiente:

- Concentración de ADN: entre 400/200 ng/µl
- Absorbancia 260/280: entre 1,72 y 1,98 (cuanto más cercano a 2 menos impurezas proteicas)
- Absorbancia 260/230: entre 2,6 y 1,45

3.3 Obtención de la secuencia del exón 15 del gen *APC*

Este paso fue realizado en el laboratorio del Instituto de Medicina Molecular Finlandés (FIMM) en Helsinki, Finlandia (en colaboración con el grupo de la Dra. Peltomäki). Las muestras de ADN fueron enviadas vía encomienda hasta el destino antes mencionado. Allí, fueron amplificadas mediante PCR de punto final utilizando los cebadores descriptos en la tabla 1 para obtener los productos de amplificación del exón 15 del gen *APC* (NM_000038.5).

El exón 15 comprende >75 % de la secuencia codificante del gen y es el Hotspot de las mutaciones somáticas y de la línea germinal (Laurent-Puig, Béroud, & Soussi, 1998). Posteriormente, estos productos fueron luego purificados utilizando Exo-SapIt.

Tabla 1: Cebadores utilizados específicos del exón 15 del gen APC

Nombre	Secuencia	Tamaño del producto
APCEx15_1 (Fw)	5´- TACTGCATACACATTGTGACCTT -3´	1039 pb
APCEx15_2 (Rv)	5´- CGAGGGTTTCATTTGACCTC -3´	
APCEx15_3 (Fw)	5´- CAATTTCACTAAGTCGGAAAATTCA -3´	1084 pb
APCEx15_4 (Rv)	5´- CTGCTTCCTGTGTCGTCTGA -3´	
APCEx15_15 (Fw)	5´- GGCATTATAAGCCCCAGTGA -3´	584 pb
APCEx15_6 (Rv)	5´- ACTTGGTTTCCTTGCCACAG -3´	
APCEx15_5 (Fw)	5´- GCTCCATCCAAGTTCTGCAC -3´	603 pb
APCEx15_16 (Rv)	5´- CATGGTTTGTCCAGGGCTAT -3´	

3.4 Secuenciación

La secuenciación fue realizada en el laboratorio FIMM. La secuenciación por Sanger del exón 15 del gen *APC* se realizó utilizando el analizador genético AbiPrism 3100 (Applied Biosystems, California, Estados Unidos). En aquellas muestras donde no se encontró la variante

genético-causal en dicho exón, se realizó la captura de exoma completo utilizando el Kit *SeqCap EZ MedExome Target Enrichment* (Roche, Indiana, Estados Unidos), y posteriormente la secuenciación masiva en paralelo del mismo utilizando la plataforma Illumina HiSeq 2000 (Illumina, California, Estados Unidos).

3.5 Análisis de datos obtenidos por secuenciación de Sanger

Datos obtenidos de la secuenciación del exón 15 del gen *APC*, utilizando el conjunto de cebadores resumidos en la tabla 1, fueron analizadas utilizando el visualizador gratuito de cromatograma, *Chromas* (Technelysium Pty Ltd, Australia); junto con la herramienta gratuita *Indigo* (Gear-genomics, Heidelberg, Alemania), el cual es un buscador de variantes puntuales (SNP) y de deleción/inserción, para realizar la búsqueda y corroboración de variantes genético-causales de las muestras.

3.6 Procesamiento de datos de NGS

En aquellas muestras que no se encontraron variantes genético-causales se procedió a la secuenciación masiva en paralelo del exoma completo, los datos fueron enviados en formato BAM y VCF desde FIMM hasta nuestro laboratorio. Para el procesamiento se utilizó el esquema de pasos (*pipeline*) ya establecido en el grupo del Dr. Turjanski (director del presente trabajo), el cual se basa en las buenas prácticas de GATK del Broad Institute (McKenna et al., 2010).

Se realizó el control de calidad de las lecturas, mapeo, alineamiento y llamado a las variantes a través de un script ad-hoc bash. Luego se procede a la anotación de las mismas y, finalmente, al análisis y priorización de las variantes en el contexto de la clínica. Utilizando, para este último paso, la plataforma web _B Platform (Bitgenia, Buenos Aires, Argentina), la cual permite integrar la información de cada variante junto con la información de calidad de la corrida y filtrar por paneles de genes. Las herramientas utilizadas siguiendo este pipeline, junto con el tipo de archivo que se va obteniendo en pasos intermedios se resume en la figura 6. Existen sentidos bidireccionales (obtención de archivos anteriores) en algunos de los pasos, los cuales fueron simplificados en la esquematización considerando sólo el sentido unidireccional.

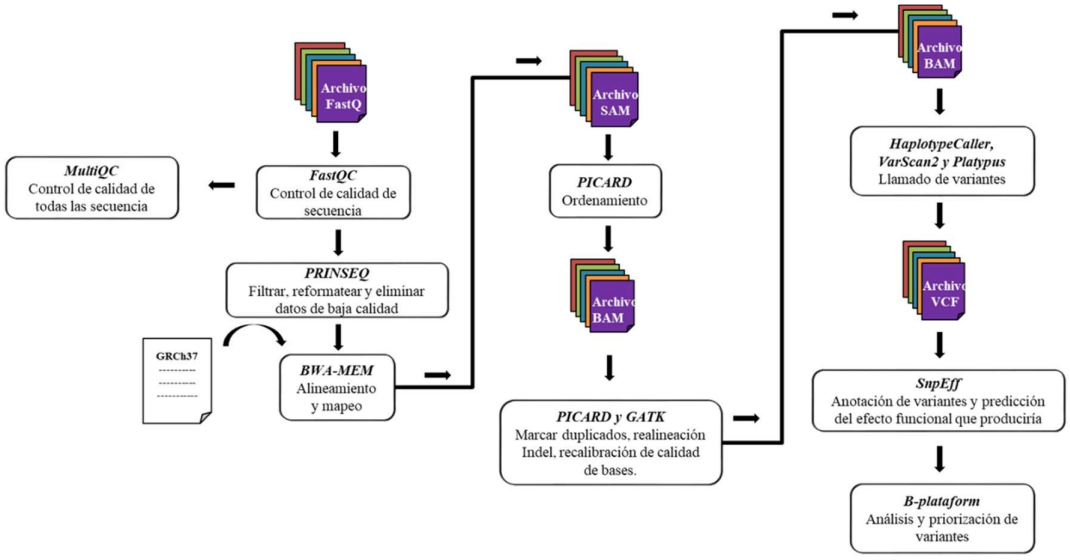

Figura 6. Pipeline del procesamiento de los datos de NGS.

3.6.1 Control de Calidad

Los datos de NGS fueron obtenidos mediante la tecnología *paired-end*, es decir que de un mismo fragmento amplificado se obtienen lecturas correspondientes a ambos extremos, separados por una distancia conocida. Los archivos enviados por el laboratorio FIMM fueron en formato BAM. Por lo cual mediante el uso de las herramientas de samtool pudimos obtener desde el archivo BAM dos archivos FASTQ por muestra denominados "R1.fastq" y "R2.fastq", según la dirección de la lectura.

Este formato ". fastq" es un archivo de texto con una forma común de almacenar lecturas de secuencia, cada lectura está representada por cuatro líneas. La primera comienza con un carácter '@', junto a un identificador de secuencia. La segunda línea contiene la información de la secuencia de letras sin procesar, en el alfabeto de cuatro letras habitual para identificar a los nucleótidos (ATCG). La tercera línea comienza con un carácter "+", el cual divide la información de la secuencia de la información de calidad de cada base. La cuarta línea representa las puntuaciones de calidad. Nos informa respecto a la probabilidad de que un llamado de base sea incorrecto. Cada símbolo representado con el alfabeto ASCII corresponde a un nivel de calidad en escala Phred, esta última determinada por el secuenciador.

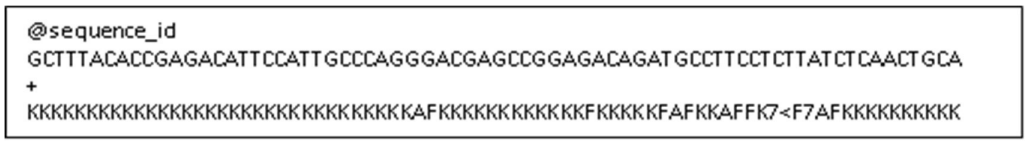

Figura 7. Representación del archivo ". fastq". Ejemplo del archivo obtenido luego de la secuenciación por NGS, en la cual se observa la información de la secuencia más la calidad de la misma.

En este paso se utilizó el programa FastQC (Babraham Institute) con el fin de obtener estadísticas en relación a la cantidad de lecturas totales, el tamaño promedio, el contenido de C-G, la cantidad de lecturas repetidas y el contenido de adaptadores que no se hayan filtrado en los pasos previos, entre otros parámetros. Finalmente, se utilizó el programa MultiQC (Ewels, Magnusson, Lundin, & Käller, 2016) que permite aunar todos los archivos de análisis obtenidos con FastQC en un solo archivo.

En el paso de filtrado, se utilizó el programa PRINSEQ (Schmieder & Edwards, 2011) como parámetro en este paso se decidió eliminar las lecturas que tengan un mínimo de calidad de 20 en valores de Phred score (equivalente a eliminar lecturas con menos del 99% de precisión en las bases o que se encuentra 1 error cada 100 bases secuenciadas).

3.6.2 Mapeo y alineamiento

Una vez obtenidas las lecturas y controlada su calidad, fue necesario ordenarlas y conocer su ubicación en el genoma. Este proceso se conoce como mapeo, se establece su posición correspondiente en relación al genoma de referencia, en este trabajo se utilizó el genoma humano de referencia versión 37 (Genome Reference Consortium Human Build 37, GRCh37). Al ubicar cada lectura en su posición correspondiente, se pudo comparar cada base del genoma de referencia con aquellas presentes en las lecturas que se superponen en esa región. Este proceso se conoce como alineamiento, en el cual se comparan dos secuencias, estableciendo una correspondencia de las bases de una con las de la otra (lecturas en relación a la referencia).

Para el proceso de alineamiento y mapeo utilizamos el software de Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2010). Se obtiene un archivo ".sam" (del inglés, Sequence Alignment Map) como resultado de este. El archivo ".sam" consiste en un archivo de texto

tabulado con un encabezado que contiene información general del secuenciamiento y alineamiento, así como, líneas correspondientes a cada lectura alineada. Cada línea está dividida en campos tales como nombre del fragmento, nombre de la referencia, posición de mapeo, calidad de mapeo, código CIGAR (código de cómo se encuentra alineado a la referencia), la longitud del fragmento y su secuencia, entre otros (Tabla 2).

Tabla 2: Campos obligatorios en el formato SAM/BAM

Campo	Descripción
QNAME	Identificador de la lectura (read) secuenciado
FLAG	Indicador binario de mapeo
RENAME	Nombre de la secuencia de referencia
POS	Posición del mapeo
MAPQ	Calidad del mapeo
CIGAR	Código. Indicador compacto de mapeo
RNEXT	Nombre de la secuencia de referencia de la siguiente lectura
PNEXT	Posición de la siguiente lectura
TLEN	Longitud del templado
SEQ	Secuencia del templado
QUAL	Calidad obtenida en código ASCII

Los archivos SAM tienen la característica de ser desordenados y en texto plano (el alineador realiza una línea por cada secuencia analizada según vaya apareciendo), y para que pueda ser correcta y rápidamente "leído" por futuros programas que trabajan de forma secuencial es necesario un proceso de "ordenado" por coordenadas. Para este procesamiento se utilizó el programa PICARD el cual permite convertir el archivo ".sam" a su versión binaria e indexada en archivo ".bam", el cual ocupa menor espacio, pero no es posible acceder a su información por un procesador de texto (no es "human-readable" o legible por humanos).

3.6.3 Procesamiento

Primero se verifican las estadísticas generales del mapeo, como la cantidad de lecturas que alinearon correctamente contra la referencia, con la herramienta "flagstat" de Samtools.

Los múltiples procesos de PCR realizados, tanto en la generación de la librería como en la secuenciación en sí, pueden generar duplicados, y estos pueden contribuir a una profundidad de lectura ficticia. Las lecturas duplicadas pueden generar sesgos en el llamado de variantes, corriendo el riesgo de tener una sobrerrepresentación de las secuencias en algunas áreas. Los

duplicados por lo general no son eliminados, pero son identificados y marcados con una bandera, “FLAG” en los archivos SAM o BAM para que los programas llamadores de variantes (pasos posteriores) no los tomen en cuenta a la hora de establecer los genotipos para cada variante. El programa PICARD mediante la opción “MarkDuplicates” se procedió a marcar los duplicados de PCR del archivo “.bam”.

Posteriormente al mapeo, se realiza además una realineación local alrededor de los indels encontrados. Este procedimiento ocurre en dos pasos, primero, se utiliza la opción “RealignerTargetCreator” de Genome Analysis Toolkit (GATK, versión 3.5 - McKenna et al. 2010), en el cual el programa identifica los intervalos que deben ser realineados y en un segundo paso, mediante la opción “IndelRealigner” del mismo paquete de herramientas, se determina la secuencia consenso óptima y se realinean las lecturas.

Por último, la recalibración del nivel de calidad de base (BQSR), es un paso de pre-procesamiento de datos que detecta errores sistemáticos cometidos por el equipo de secuenciación cuando estima la precisión de cada llamada de base. Los puntajes de calidad de base son estimaciones de error por base emitidas por los equipos de secuenciación y expresan cuán seguro estaba el equipo de llamar a la base correcta cada vez. Los puntajes de calidad son expresados por el equipo en escala de Phred, y son importantes debido a que nuestros algoritmos de llamado de variante (utilizados en pasos posteriores) dependen en gran medida del puntaje de calidad asignado a las llamadas de base individuales en cada secuencia de lectura. Este paso también se realiza en dos partes, primero generando un modelo de la covariación de los datos provistos con un set de variantes conocidas, utilizando la opción “BaseRecalibrator” y luego, se ajusta la calidad de las bases en la muestra basándose en el modelo creado mediante la opción “PrintReads”, ambos comandos son herramientas de GATK.

3.6.4 Llamado de variantes

Una vez mapeadas y alineadas las lecturas contra la referencia, pueden relacionarse las diferencias entre el consenso obtenido de la secuenciación y del genoma de referencia (GRh37) y “llamar las variantes” (en inglés, Variant Calling), es decir, determinar los sitios que la muestra difieren con la referencia.

La naturaleza de estas variantes puede ser de cambios de un único nucleótido (Single Nucleotide Polymorphism; SNP), inserciones o deleciones (Insertions, Deletions; Indel) o

variantes estructurales (Structural Variants; SV) como las variantes del número de copias (Copy Number Variations; CNVs).

Los programas llamadores de variantes analizan las regiones donde se presentan variaciones con respecto al genoma de referencia y seleccionan aquellas que cumplan con determinados criterios que las hagan elegibles como "variantes verdaderas", tales como: calidad de la base secuenciada, calidad de mapeo y el número de lecturas independientes que den evidencia a favor de su presencia.

Para el proceso de llamado de variantes, utilizamos la herramienta "HaplotypeCaller" de GATK, siguiendo las prácticas recomendadas del Broad Institute (Van der Auwera et al., 2013). La herramienta "HaplotypeCaller" es capaz de llamar a los SNPs e indels simultáneamente a través del ensamblaje local de novo de haplotipos en una región activa. Obteniéndose un archivo de formato tipo VCF (por sus siglas Variant Call Format), introducido por el consorcio responsable del proyecto 1000 Genomas (Auton et al., 2015).

El formato VCF contiene la información sobre el polimorfismo (SNP, inserción, deleción, etc), más lo que se denomina meta-información. Dentro de los campos más importantes se encuentra el número de cromosoma (CHROM), posición dentro del cromosoma (POS), identificador de la variante en dbSNP (ID), nucleótido en el genoma de referencia (REF), nucleótido en la muestra secuenciada (ALT), calidad de la secuencia en escala phred (QUAL), filtro personalizable para selección de variantes (FILTER), información adicional (INFO), formato de la información sobre las lecturas y su genotipo (FORMAT), un campo por cada muestra (SAMPLE) (Danecek et al., 2011).

3.6.5 Anotación y predicción de los efectos biológicos

Para anotar las variantes, se utilizó el programa de anotación estructural y predicción de efectos "SnpEff" (Cingolani et al., 2012), el cual agrega al campo "INFO" del archivo ".vcf", información pertinente sobre la posición donde se encuentra la variante (cerca o dentro de un gen, en la zona regulatoria, en un exón o intrón), el efecto que produce (cambios sinónimos o missense, pérdida de codón de iniciación o aparición de codón stop prematuros, pérdida de sitios dadores o aceptores de splicing, etc) y, le asigna un valor de efecto de patogenicidad (LOW, MEDIUM o HIGH) para zonas codificantes o modificante (MODIFIER) para variantes en zonas inciertas.

En este procesamiento se utilizaron bases de datos biológicas para añadir información relevante a la variante de frecuencia poblacional, conservación evolutiva, base de datos clínicas y farmacológicas, entre otras. Estas bases de datos fueron: Clinvar (Landrum et al., 2014), ExAC (Exome Aggregation Consortium) (Karczewski et al., 2017), gnomAD (Gudmundsson et al., 2022), dbSNP, Hapmap (Gibbs et al., 2003), GWASCat (Ferrarini et al., 2015), VarType, PhastCons (Siepel et al., 2005), CADD (Combined Annotation Dependent Depletion) (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019), PharmGKB (Barbarino, Whirl-Carrillo, Altman, & Klein, 2018) e InterVar (incluye ACMG) (Q. Li & Wang, 2017).

3.7 Análisis de los archivos VCF

Para este punto utilizamos la herramienta _B Platform (Bitgenia) que permite priorizar y clasificar variantes, fue diseñada por el grupo del Dr. Adrian Turjaski. Es una herramienta muy flexible que permite la utilización de filtros que corresponden a frecuencias poblacionales, genes específicos, herramientas *in silico* que permiten predecir el impacto fenotípico de la variante, entre otras.

3.8 Herramientas para verificar variantes indel *in silico*

Mutation Taster emplea un clasificador Bayesiano para predecir el potencial de patogenicidad de una variante. Este utiliza el resultado de todas las pruebas y las características de la variante y calcula las probabilidades de que la misma sea causante de la enfermedad o un polimorfismo. Para esta predicción, se estudiaron las frecuencias de todas las características individuales de mutaciones/polimorfismos de enfermedades conocidas en un gran conjunto de datos compuesto por > 390,000 mutaciones de enfermedades conocidas de HGMD Professional y > 6,800,000 polimorfismos tanto SNPs como Indel obtenidos de 1000 Genomes Project (Schwarz, Rödelsperger, Schuelke, & Seelow, 2010).

CRAVAT es una herramienta web que permite asociar información de diferentes bases de datos poblacionales y clínicas para que sea más fácil y eficiente la interpretación de la variante. Devuelve un score de patogenicidad que va de 0 (benigna) a 1 (patogénica), permitiendo linkear tanto el gen con bases de datos como la posición de la variante (Christopher Douville et al., 2013).

MutPred-Indel es un software basado en el aprendizaje automático que integra datos genéticos y moleculares para razonar probabilísticamente sobre la patogenicidad de las variantes indels. El modelo proporciona predicción de patogenicidad y una lista clasificada de alteraciones moleculares que pueden afectar al fenotipo. MutPred-Indel es un conjunto empaquetado de 100 redes neuronales de retroalimentación (*feed-forward neural networks*), cada una entrenada en un subconjunto equilibrado de variantes patógenas y polimorfismos (Pagel et al., 2019).

SIFT es una herramienta basada en la homología de secuencias que clasifica las sustituciones de aminoácidos intolerantes de las tolerantes y predice si una sustitución de aminoácidos en una proteína tendrá un efecto fenotípico. SIFT se basa en la premisa de que la evolución de las proteínas se correlaciona con la función de las proteínas. Las posiciones importantes para la función deben conservarse en un alineamiento de la familia de proteínas, mientras que las posiciones sin importancia deben aparecer diversas en un alineamiento (Hu & Ng, 2013).

3.9 Herramientas para verificar variantes puntuales *in silico*

CADD v1.0 es un software que utiliza un modelo para clasificar las variantes utilizando aproximadamente 8.6 mil millones de sustituciones en el genoma humano de referencia, el cual provee dos tipos de resultados RawScore estos provienen directamente del modelo y son interpretables como el grado en que el perfil de anotación para una variante dada sugiere que es probable que la variante sea "observada" (valores negativos) frente a "simulada" (valores positivos). Valores altos indican que una variante es más probable que sea simulada (o "no observada") y, por lo tanto, más probable que tenga efectos deletéreos. Y el score PHRED mayor o igual a 10 indica que se predice que estas son las sustituciones más perjudiciales del 10% que puede hacer al genoma humano, un puntaje de 20 mayor o igual indica el 1% más perjudicial y así sucesivamente (Kircher et al., 2014).

DANN se basa en redes neuronales profundas (*Deep learning*) el rango de valores es de 0 a 1, con 1 dado a las variantes que se prevé que sean las más dañinas (Quang, Chen, & Xie, 2015).

3.10 Otras herramientas bioinformáticas utilizadas

Utilizamos una herramienta adicional del grupo de la empresa Bitgenia denominada *Genecov app* (Bitgenia)la cual permite conocer la cobertura del gen que se está consultando y

devuelve las variantes anotadas en dbSNP que no están dentro de esta cobertura. Los parámetros que necesita la aplicación es el archivo ".bed" del kit de exoma completo utilizado y la lista de genes.

Además, utilizamos la herramienta FASTQ Screen creada por Babraham Institute para corroborar que la procedencia de la muestra original no se encuentre contaminada por otros organismos como por ejemplo virus o bacterias (Wingett & Andrews, 2018).

Otra herramienta utilizada para hacer los gráficos lollipop fue Mutation Mapper de cBioportal que permite agregar el gen en nomenclatura HUGO con las mutaciones, produciendo un gráfico con los dominios de una proteína en particular marcando sobre ella las mutaciones (Vohra & Biggin, 2013).

4- RESULTADOS Y DISCUSIÓN

4.1 Resultados del experimento

4.1.1 Análisis de datos de secuenciación

Dentro del pipeline descrito en la sección 3.6 de materiales y métodos se puede visualizar la calidad de las secuencias obtenidas en los archivos FASTQ. Para ello utilizamos el software FASTQC, que presenta diferentes parámetros como: calidad de la secuencia por base o por secuencia, contenido de GC, bases que no pudieron ser reconocidas, tamaño de las secuencias, duplicaciones, entre otras. Se realizó el análisis con el software MultiQC que permite aunar todos los análisis de FASTQC en un solo archivo.

La **figura 8** muestra los resultados obtenidos en base a la calidad obtenida por base (**Figura 8A**) y el promedio de la calidad de las lecturas por muestra (**Figura 8B**). El eje Y de cada gráfico defina por el Phred score permite visualizar las calidades a través del siguiente código de color: muy buena calidad (verde), calidad razonable (rosa) y baja calidad (rojo). Las gráficas muestran que los datos de secuenciación poseen, en general, muy buena calidad.

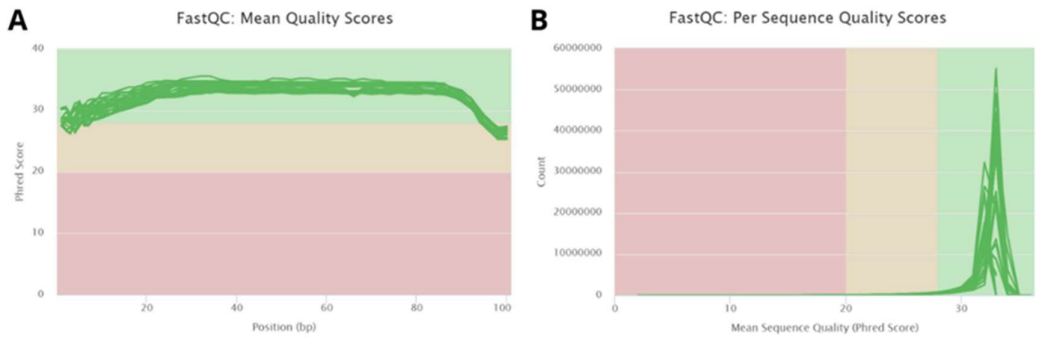

Figura 8. Calidad de los resultados utilizando MultiQC. A. Calidad de la secuencia por base representa la calidad teniendo en cuenta la posición de las bases vs su puntaje en escala Phred. **B.** Calidad de secuencia obtenida representa el número de lecturas obtenidas y en el otro eje su promedio de calidad en escala Phred.

El contenido de GC también es una medida de calidad del producto obtenido. El pico esperado en la curva es único para cada genoma en particular, para *Homo Sapiens* se espera que el contenido de GC siga una distribución normal y centrada alrededor del 40%. En la **figura 9** podemos observar una distribución bimodal de la curva con picos a los 40% y 60% respectivamente en todas las muestras.

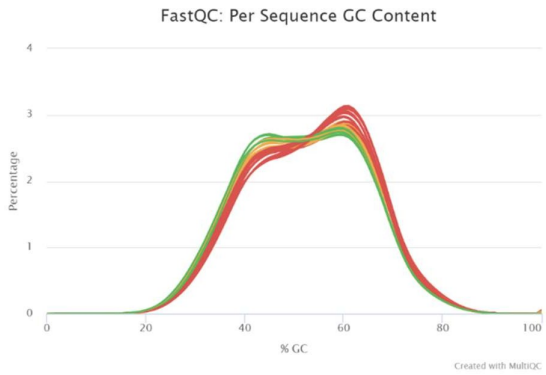

Figura 9. Curva de contenido de GC. Composición de Guanina/Citocina (GC) por base de los productos obtenidos

Para descartar contaminación de las muestras utilizamos la herramienta FASTQ Screen (Wingett & Andrews, 2018) que permite chequear las lecturas contra otros genomas. Para ello elegimos dos casos en forma aleatoria. Para realizar el análisis utilizamos el método más

completo que permite chequear contra genomas de diferentes organismos como así también vectores y adaptadores. La **figura 10** muestra los gráficos obtenidos con códigos de colores: celeste (existe un hit en el genoma); azul (múltiples hits en el mismo genoma); rosa (un hit en diferentes genomas); rojo (múltiples hits en diferentes genomas). En nuestro caso obtuvimos solo colores celestes y azules en el organismo *homo sapiens,* indicando que las lecturas son específicas de este organismo, en menor proporción obtenemos diferentes hits en colores rosa y rojo compartida con rata (rat) y ratón (mouse) correspondiente a lecturas con información conservada entre estas especies.

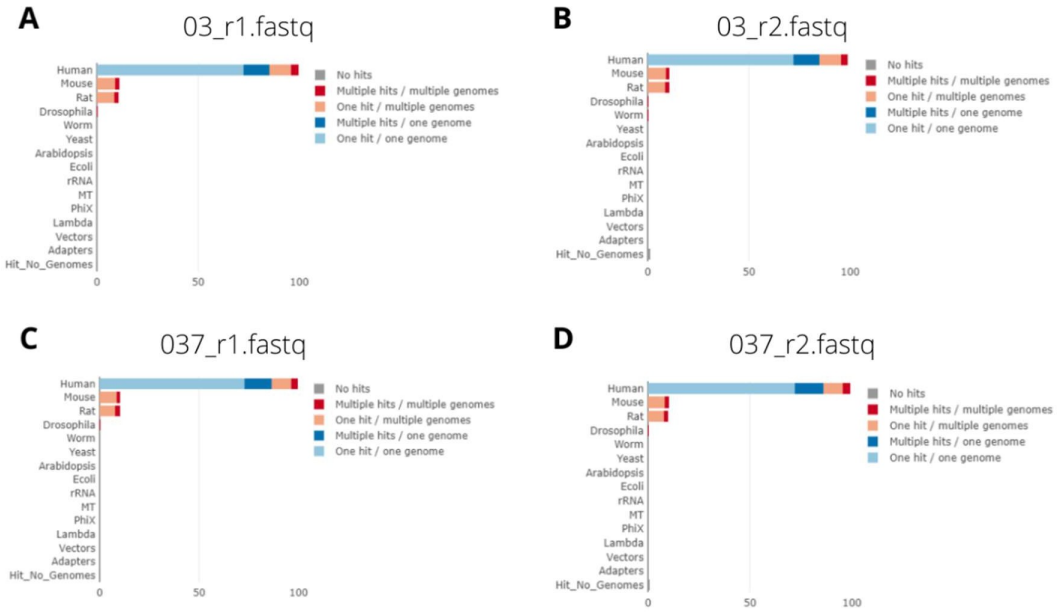

Figura 10. Verificación de contaminación con la herramienta FASTQ Screen. **A-B)** Son los resultados de las lecturas r1 y r2 una de las muestras aleatorias **C-D)** Son los resultados de las lecturas r1 y r2 de otra muestra aleatoria.

Los resultados obtenidos sugieren que la distribución bimodal corresponde a un artefacto de la técnica utilizada. Resultados bibliográficos (Bernardi, 2000; Lamolle & Musto, 2018) sugieren que este fenómeno puede explicarse por la diferencia en la composición de GC del genoma humano, donde los intrones contienen en su mayoría un bajo porcentaje de GC y los exones, en su mayoría una alto contenido del mismo. Otro trabajo (Wang, Shashikant, Jensen, Altman, & Girirajan, 2017) publicado recientemente, utiliza diferentes plataformas de

secuenciación de exoma, demuestra una falta de uniformidad de la cobertura dentro de un exón (local) como los exones de un gen (global) llevando a porcentajes elevados en las curvas GC observando curvas bimodales. Habiendo descartado la posibilidad de contaminación de las muestras en estudio, la razón exacta de la distribución bimodal requiere un análisis adicional que escapa los objetivos planteados en este trabajo y proveen una nueva línea de investigación a futuro.

Otros tres factores que pueden establecer la calidad del producto obtenido son: a) Contenido de N: bases que no pudieron ser leídas por alguna interferencia técnica que lleva a no determinar la letra que continua o empieza en esa secuencia; b) Los niveles de duplicaciones: permite reconocer fallas en los diferentes pasos de la técnica como duplicaciones debido a la técnica de PCR, errores en el equipo de secuenciación, entre otros; c) La distribución del tamaño de las lecturas: permite reconocer fallas en el largo de la secuencia obtenida, ya que el mismo debería ser uniforme para todas las lecturas.

En la **figura 11** resumimos los resultados de nuestro experimento a partir de los factores de calidad antes mencionados. Como podemos observar en la parte **A** de la gráfica, solo una de las muestras (010_r2) presenta una mínima curva en la gráfica de contenido de N las demás lecturas no muestran bases indeterminadas en sus secuencias. La parte **B** de la gráfica muestra uniformidad en las curvas sugiriendo que no hay sobrerrepresentación de alguna secuencia por duplicación. La parte C de la gráfica muestra similitud en la longitud de las lecturas, aunque la cantidad de lecturas por muestra difiere un poco.

Tomando en cuenta los resultados del análisis de calidad efectuado, podemos estar seguros que los datos obtenidos en el experimento de secuenciación del exoma completos de estos casos, almacenados en archivos de tipo VCF son íntegros y poseen información de buena calidad.

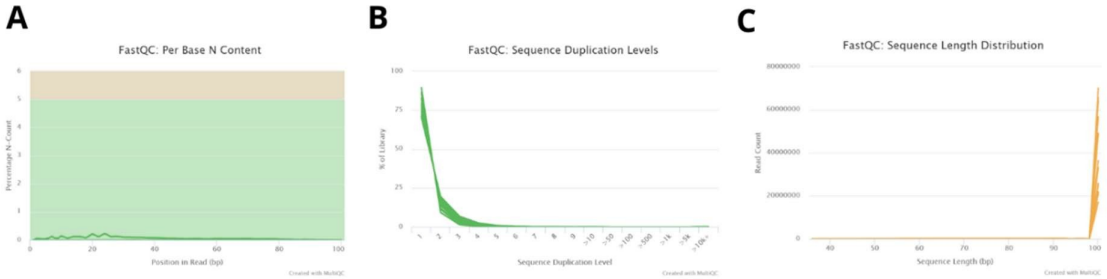

Figura 11. Otros factores determinantes de la calidad. **A.** Contenido de N, bases nitrogenadas no determinadas. **B.** Niveles de duplicación, producto de artefactos que pueden sobrerrepresentar una secuencia. **C.** Distribución del tamaño de las secuencias obtenidas.

4.1.2- Generalidades de las variantes anotadas

Luego del chequeo de calidad del punto anterior se realizó el *pipeline* completo, es decir hasta obtener el archivo VCF con las variantes anotadas. Para el análisis de este archivo se utilizó el software _B-Platform (Bitgenia), que permite hacer una búsqueda rápida de las variantes utilizando filtros específicos y también describe la estadística general del archivo VCF obtenida de cada muestra en particular.

Esta estadística se utilizó para comparar los resultados de cada muestra secuenciada. La tabla 3 resume los datos obtenidos a nivel de cantidad de polimorfismos encontrados por cada muestra, las variantes que son conocidas o reportadas en bases de datos (para el detalle de las usadas en el pipeline, ver sección 3.6.5), la profundidad promedio de cada variante y los diferentes genes que poseen las mismas.

Tabla 3: Características de los polimorfismos encontrados de cada muestra

Muestra	Variantes	Variantes conocidas	Profundidad promedio	Genes
001	35349	35197	112	11994
003	35705	35544	151	12040
006	35912	35734	210	12226
008	36356	36130	197	12278
010	35641	35486	101	12036
018	34946	34812	153	11862
019	36059	35901	177	12084
024	36599	36403	194	12248
027	35151	34990	67	11951
029	35099	34902	75	11912
037	35051	34930	60	11942
041	34964	34812	78	11897
046	35907	35744	74	12077
048	34634	34489	52	11822
049	35792	35630	62	12071

Podemos observar en la tabla 3 que las últimas 7 filas muestran una profundidad del polimorfismo baja respecto de las otras muestras, esto puede deberse a: la calidad de estas muestras (las mismas se realizaron a partir de sangre congelada), como también al experimento de secuenciación en sí mismo, el cual haya producido una disminución en la calidad de las lecturas y las restricciones impuestas por el pipeline respecto de la calidad las haya descartado.

A su vez, esta herramienta permite realizar una clasificación del impacto de la variante a nivel proteico (alto, moderado, modificador o bajo), el efecto de la misma en el transcripto (*stop-gained, frame shift* o *stop-lost)* y la clase de polimorfismo a nivel de ADN (*missense, nonsense, silent* o *none*). Como puede verse en la **figura 12**, el análisis es similar en cada muestra, tanto en la cantidad de polimorfismos encontrados como en la clasificación de los efectos, clases e impacto que las variantes producen. Sin encontrar ninguna muestra discrepante en esta estadística general.

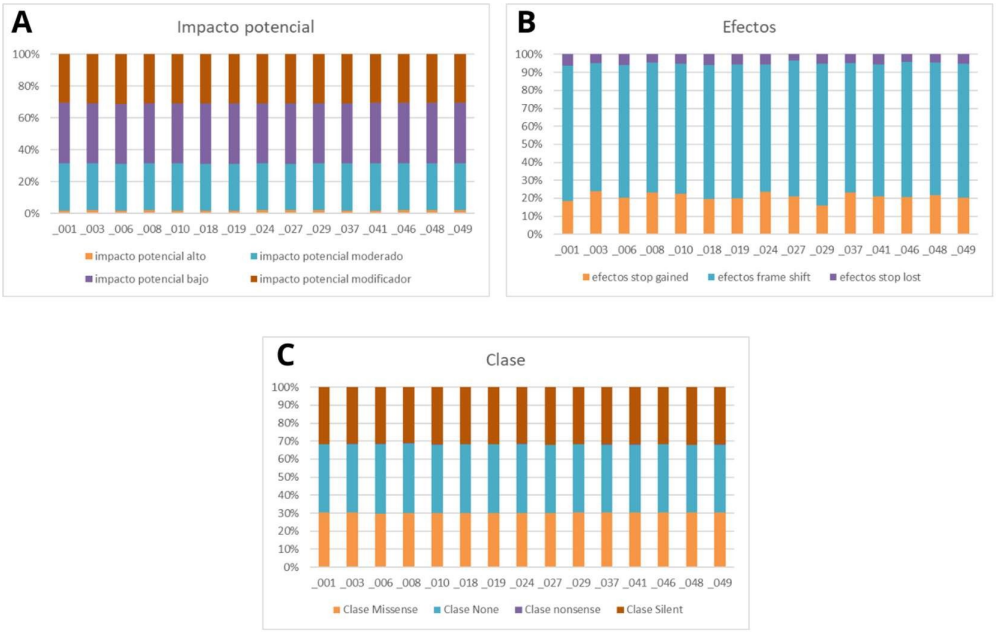

Figura 12. Clasificación de los polimorfismos. A El impacto se clasifica en alto, moderado, bajo y modificador **B** El efecto se clasifica en *stop gained*, *frame shift* y *stop lost*. **C** La clase se clasifica en *missense*, *silent*, *nonsense* y *none*.

4.1.4- Búsqueda de la variante genético-causal

Como se mencionó previamente en materiales y métodos, el software utilizado permite realizar la búsqueda de variantes empleando diferentes filtros, entre ellos podemos nombrar tipo de variante, calidad de la variante, frecuencia, predicción del daño según diferentes softwares (ej. Mutation Taster), genes (se pueden restringir a genes específicos), consecuencia, todos por separados o usando filtros en conjunto.

Para evitar hallazgos incidentales utilizamos un pipeline de priorización de variables. Comenzando con la creación de dos paneles de genes: a) genes relacionados directamente con

PAF y CCHNP en caso de encontrar casos mixtos; b) genes relacionados directa o indirectamente con PAF, CCHNP y cáncer colorrectal.

Los genes del primer panel son los siguientes: *APC*, *BMPR1A*, *EPCAM*, *MLH1*, *MSH2*, *STK11*, *MSH6*, *MUTYH*, *PMS2*, *PTEN*, *SMAD4*, *RNF43*, *GREM1*, *MSH3*, *NTHL1*, *POLD1*, *POLE*. Los primeros 11 se encuentran establecidos como se explicó en la sección 1.5 de la introducción y luego agregamos genes que se relacionan con otros tipos de poliposis en caso de incorporar muestras con otros fenotipos. Los genes del segundo panel se construyeron consultando bibliografía que permitía incorporar genes relacionados de manera directa e indirecta con el fenotipo obteniendo un total de 187 genes (ver anexo 6.2).

Tabla 4: Cobertura de los genes y sus variantes

Gen	Cobertura: Región codificante	Cobertura: Región UTR	Variantes Patogénicas	Variantes patogénicas cubiertas	Variantes patogénicas no cubiertas
RNF43	100.0%	43.7%	2	2	0
SMAD4	100.0%	19.4%	236	235	1
PTEN	98.3%	20.0%	602	584	18
MLH1	100.0%	89.9%	1032	1032	0
GREM1	100.0%	18.3%	0	0	0
MSH6	99.9%	64.8%	883	879	4
STK11	100.0%	53.7%	166	166	0
APC	99.8%	79.8%	874	869	5
MSH2	100.0%	90.8%	1056	1043	13
EPCAM	100.0%	65.4%	17	17	0
MSH3	100.0%	85.1%	5	5	0
NTHL1	98.3%	88.3%	2	2	0
POLD1	99.3%	96.6%	4	4	0
PMS2	95.1%	88.2%	385	363	22
POLE	100.0%	88.3%	24	24	0
BMPR1A	100.0%	14.7%	94	94	0
MUTYH	100.0%	91.2%	211	211	0

Del primer panel de genes realizamos el análisis de cobertura obtenido del experimento de secuenciación, utilizando para ello el archivo ".bed" del kit empleado (brinda información de las coordenadas/posición en los cromosomas). Como se puede observar en la tabla 4 obtuvimos datos de cobertura y variantes reportadas como patogénicas que están o no cubiertas por este experimento: **a)** *APC* obtuvo una cobertura del 99,8% pudiendo detectar a partir de

este kit 869 variantes junto con 5 que no son posibles de detectar (NM_001127511.2(APC): c.-191T>C; NM_001127511.2(APC): c.-192A>T; NM_001127511.2(APC): c.[-125delA;-195A>C]; NM_001127511.2(APC): c.-192A>G), las cuales se encuentran reportadas en la base de datos ClinVar pero sin estrellas, es decir que no hay suficiente evidencia para adjudicarlas como genético-causales; **b)** BMPR1A una cobertura del 100% detectando 94 variantes; **c)** EPCAM una cobertura del 100% detectando 17 variantes; **d)** MLH1 una cobertura del 100% detectando 1032 variantes; **e)** MSH2 una cobertura del 100% detectando 1043 junto con 13 que no pueden ser detectada por este experimento las cuales se encuentran reportadas en ClinVar solo con 3 estrellas dos de las variantes (NM_000251.2(MSH2): c.212-478T>G; NM_000251.2(MSH2):c.942+2T>G) y las demás con dos, una o ninguna estrella es decir que no hay evidencia para asociarlas como genético-causal; **f)** STK11 una cobertura del 100% detectando 166 variantes; **g)** MSH6 una cobertura del 99,9% detectando 879 variantes junto con 4 que no pudieron ser detectadas con este experimento solo dos se encuentran reportadas en ClinVar con una estrella; **h)** MUTHY una cobertura del 100% detectando 211 variantes; **i)** PMS2 una cobertura del 95,2% detectando 363 variante junto con 22 que no pudieron ser detectadas por este experimento se encuentran reportadas en la base de datos ClinVar y solo 6 de ellas con dos estrellas (NM_000535.7(PMS2): c.354-2A>G; NM_000535.7(PMS2): c.163+1G>A; NM_000535.7(PMS2): c.163+1G>A; NM_000535.7(PMS2): c.251-2A>T; NM_000535.7(PMS2): c.353+2T>C; NM_000535.7(PMS2): c.354-2A>G); **j)** PTEN cobertura del 98,3% detectando 584 variante junto con 18 que no pudieron ser detectadas por esta técnica las cuales se encuentra reportadas en ClinVar con una o ninguna estrella; **k)** SMAD4 cobertura del 100% detectando 235 variante junto con una variante reportada en clinVar con dos estrellas que no pudo ser detectada por esta técnica (NM_005359.6(SMAD4): c.1228_1229del (p.Gln410fs)); **l)** RNF43 cobertura del 100% detectando 2 variantes; **ll)** GREM1 cobertura del 100%; **m)** MSH3 cobertura del 100% detectando 5 variantes; **n)** NTHL1 cobertura 98,3% variantes detectadas 2; **ñ)** POLD1 cobertura del 99,3% detectando 4 variantes; **o)** POLE cobertura del 100% detectando 24 variantes. A partir de estos resultados pudimos corroborar la cobertura de los genes que se relacionan directamente con la patología obteniendo valores mayores al 95%, como así también conocer la información de aquellas variantes que quedan sin ser detectadas en estos genes. De esta manera, en caso que la evidencia clínica asuma un rol esencial de un gen particular se puede buscar con cebadores específicos dichas variantes utilizando la técnica de secuenciación por Sanger.

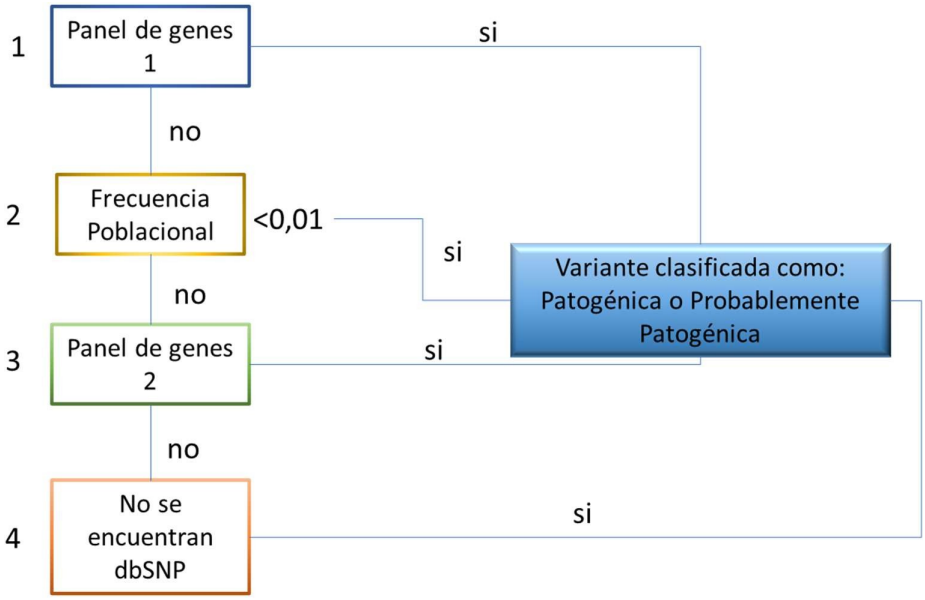

Figura 13. Resumen de priorización de variantes. Los números 1, 2, 3 y 4 son los pasos sucesivos si no se encuentra la variante genética-causal en el paso anterior.

En la **figura 13** se resumen los pasos planteados para priorizar la búsqueda de variantes. Primero se elige el panel de genes 1 (directamente relacionados con la patología) de esta manera el programa devuelve una lista de variantes sólo en estos genes seleccionados. La información que obtenemos es la posición de la variante en el genoma, el gen, el cambio, el impacto/efecto, el criterio ACMG, la frecuencia, la evidencia (datos de OMIM, ClinGen, GeneCard, gnomAD, entre otras) y la información de la muestra (si se encuentra en una o varias muestras la misma variante, si es una herencia dominante o recesiva y la profundidad de la misma). Si a partir de esta búsqueda encontramos una variante clasificada como patogénica o probablemente patogénica es decir con evidencia sobre la misma podemos afirmar que encontramos la variante genética causal y posteriormente debemos validarla por secuenciación de Sanger. En caso contrario podemos dejar seleccionado el panel 1 y agregar información de la frecuencia poblacional permitiendo de esta manera filtrar por frecuencias bajas <0.01 para priorizar las variantes que no se encuentran en la población.

Si no encontramos una variante con esta nueva búsqueda, cambiamos el panel de genes por el 2, dejando la frecuencia poblacional seteada y de esta manera realizar la búsqueda. Hallando una variante con esta búsqueda necesitamos corroborar por secuenciación de Sanger, pero a su vez, tenemos varios interrogantes: a) ¿El gen se encuentra relacionado directamente

con la patología?; b) ¿Hay evidencia biológica de la misma?; c) ¿Existen reportes que encuentran la misma mutación?; d) ¿En la base de datos ClinVar o en la herramienta Varsome la clasifica de la misma forma? Si estas preguntas son fáciles de responder podemos estar en presencia de la variante genético-causal, en caso contrario debemos estudiar la variante con otros métodos para corroborar su patogenicidad.

Por último, si no encontramos la variante de esta manera podemos hacer una búsqueda en las variantes donde no hay información en la base de datos dbSNP, es decir no está reportada y en esta búsqueda podemos mantener el panel de genes 2 o tomar todos los datos. Por ello, si encontramos alguna variante en un gen debemos corroborar tanto su patogenicidad analizando marcadores moleculares y genéticos en la muestra específica como su asociación con el fenotipo clínico.

4.2 Casos clínicos

4.2.1 Consideraciones generales

Luego de obtener el ADN de las muestras se sigue una serie de pasos para detectar la variante genético-causal: a) Secuenciación por el método de Sanger del exón 15 del gen *APC*; b) Análisis de grandes deleciones e inserciones por MLPA para el gen *APC* (este análisis no está descripto en este trabajo); c) Método de secuenciación por NGS del exoma completo de las muestras negativas.

Del conjunto de muestras analizadas encontramos seis muestras que presentaron variantes en el exón 15 del gen *APC*, las cuales están resumidas en la tabla 5. Las mismas no se encuentran reportadas en las bases de datos como: dbSNP, ClinVar, gnomAD, entre otras, por lo cual se definen como "*novel*", utilizamos la herramienta web Varsome (Kopanos et al., 2018) para clasificarlas según el criterio ACMG.

Tabla 5: Resumen de las variantes obtenidas por secuenciación de Sanger

M	Pólipo	AF	Posición	Cambio de nucleótidos	Cambio de Aminoácidos	Efecto	ACMG
02	FAP (>100)	SI	5:112175094	- → C	Ile1269Asnfs*7	Frameshift	**P**
05	FAP (>100)	NO	5:112175389	TCAGACACCCAAAA → -	Gln1367Serfs*8	Frameshift	**P**
13	FAP (>100)	SI	5:112175389	TT → T	Phe773Leufs*4	Frameshift	**P**
23	FAP (>100)	SI	5: 112175475	- → A	Ser1395Lysfs*3	Frameshift	**LP**
32	FAP (>100)	SI	5: 112137009	- → C	His255Glnfs*2	Frameshift	**LP**
33	FAP (>100)	SI	5:112175650	T → -	Asn1455Ilefs*18	Frameshift	**LP**

4.2.1.1- Caso 02

La posición de la variante encontrada es GRCh37:5:112175094 el cambio a nivel p.Ile1269NfsTer7 (NM_000038.6: c.3803dupC). La herramienta *Varsome* la clasifica como patogénica. No se han reportado hasta el momento cambios genéticos en la misma posición, aunque se han evidenciado variantes que llevan un cambio en la misma posición del aminoácido 1269.

En una cohorte de pacientes italianos diagnosticados con poliposis con o sin CCR, reportan un probando con la alteración a nivel germinal en la posición c.3805_3806delAT (p.Ile1269MetfsTer6) (Marabelli et al., 2016). A su vez, otro estudio reporta una familia (FAP 60) de procedencia italiana la cual presentó una deleción de 12 pb a nivel germinal en la posición c.3805 llevando a una proteína truncada (Pezzi et al., 2009). Otro estudio realizado en una cohorte de pacientes polacos reportan una variante novel a nivel germinal en la posición c.3807_3808delAT en el exón 15 p.Ile1269fs (Plawski and Slomski, 2008). Además se reportan dos variantes en un estudio de pacientes alemanes diagnosticados con FAP, una en la familia 221 en el exón 15 posición c.3806insT y otra en la familia 382 en el exon 15 posición c.3805delA ambos en el mismo sitio 1269 de la proteína (W Friedl et al., 2001). Al igual que en otro estudio también en una cohorte de pacientes alemanes reportan un caso en la misma posición c.3805delA p.Ile1269fs (Waltraut Friedl & Aretz, 2005). Finalmente, en una cohorte de pacientes chinos se reporta una variante c.3807_3808delAT, p.Ile1269Metfs*6 pero en este caso la base de datos ClinVar la reporta como patogénica con dos estrellas (dos o más fuentes reportaron la variante, pero no se encuentra revisada y validada por un panel de expertos) (Zhang et al., 2017).

4.2.1.2- Caso 05

La posición de la variante encontrada es GRCh37:5:112175389 el cambio a nivel proteico p.Gln1367SerfsTer3 (NM_000038.6: c.4098_4111del). Utilizando la herramienta Varsome se clasifica teniendo en cuenta los criterios de ACMG como patogénica. Hasta el momento no hay otra variante reportada en el mismo sitio aunque un estudio realizado en una cohorte de paciente FAP o AFAP del hospital AC Camargo (Sao Pablo) se identificó una variante como novel en la posición c.4097dupC (p.Gln1367Serfs*8) en el probando (ID_23) con fenotipo FAP severa (Cruz-Correa et al., 2017).

4.2.1.3- Caso 13

La posición de la variante encontrada es GRCh37:5:112175389, siendo la alteración proteica en la posición p.Phe773LeufsTer4 (NM_000038.6:c.2319delT), utilizando la herramienta Varsome se la clasificó como patogénica siguiendo los criterios de ACMG, no hay reportes hasta el momento de una variante en el mismo sitio.

Sin embargo, en un estudio reportado para una cohorte de pacientes alemanes se encuentró la variante c.2318_ 2319delTT p.Phe773fs en la misma posición del aminoácido 773 (Waltraut Friedl & Aretz, 2005). A su vez, otro trabajo realizado en pacientes australianos con diagnóstico de CCR, informan la mutación a nivel somático en el probando 270 (c.2318_2319delTT) (Christie et al., 2013).

4.2.1.4- Caso 23

La posición de la variante encontrada es GRCh37:5:112175475, en la posición a nivel proteico p.Ser1395LysfsTer3 (NM_000038.6: c.4183dupA), Varsome la clasificó como probablemente patogénica siguiendo los criterios de ACMG, la cual no se encuentra reportada hasta el momento.

En un estudio realizado en 50 familias diagnosticadas con FAP o AFAP reportaron una variante NM_000038.6: c.4184delGT (F1396X) en el probando número 2478 a nivel germinal (Armstrong, Davies, Guy, Frayling, & Evans, 1997).

4.2.1.5- Caso 32

La posición de la variante encontrada es GRCh37:5:112137009, el cambio a nivel proteico es p.His255ProfsTer2 (NM_000038.6:c.763dupC), Varsome la clasificó como probablemente patogénica según el criterio de ACMG, no hay reportes de la misma variante hasta el momento.

En un estudio realizado en pacientes checoslovacos con diagnóstico de FAP y AFAP, reportaron la variante novel c.759_787del29 p.His255ArgfsX11 en el probando S7 (Stekrova et al., 2007). Además, en otro estudio realizado en una cohorte de pacientes chinos reportan la mutación c.763_764insA (p.His255GlnfsX2) (Cai, Zhang, & Zheng, 2008).

4.2.1.6- Caso 33

La posición de la variante encontrada es GRCh37:5:112175650, el cambio a nivel proteico es p.Asn1455IlefsTer18 (NM_000038.6:c.4359delT), la herramienta Varsome la clasificó como probablemente patogénica siguiendo el criterio de ACMG, la cual no ha sido reportada a nivel germinal pero si fue encontrada a nivel somático en adenocarcinoma y reportada en la base de datos Cosmic (COSM4168331).

4.2.1.7- Efecto de las variantes genéticas sobre *APC*

Para entender el efecto de las mismas primero debemos conocer la estructura de la proteína codificada por el gen *APC,* es decir los dominios de la misma. Para poner en contexto las funciones y los sitios de unión de esta proteína hemos utilizado el gráfico publicado por Fodde y cols (Fodde, Smits, & Clevers, 2001). Como podemos observar en la **figura 14A** esta proteína tiene sitios de unión para la proteína B-catenina, cuya vía de acción son los procesos de regeneración de tejido, proliferación, diferenciación, entre otras (Mantilla, Suárez Mellado, Duque Jaramillo, & Navas, 2015). A su vez, posee sitios que le permiten formar homodímeros, puede unirse a microtúbulos, tiene señales que le permiten ingresar al núcleo de la célula como salir hacia el citosol, sitios de unión a otras proteínas, etcétera.

Como podemos observar en la **figura 14B** las variantes encontradas llevan a una proteína truncada perdiendo los sitios de unión, señales de ingreso y egreso del núcleo, entre otros. Según los criterios de ACMG la clasificación biológica de las variantes identificadas es patogénica y probablemente patogénica, además no se encuentran reportadas en ninguna de las

bases de datos poblacionales o bases clínicas de referencia. En consecuencia, podemos catalogarlas como variantes *noveles*.

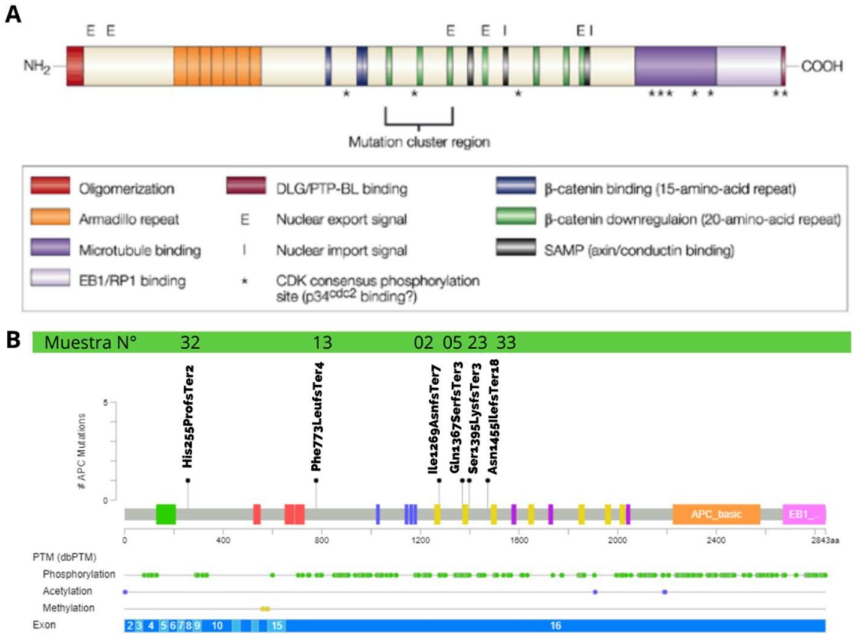

Figura 14. Representación de las variantes encontradas. **A** estructura del gen APC con sus dominios identificados tomado de (Fodde et al., 2001) **B.** Utilizando la herramienta de cBioportal, se realizó un gráfico tipo lollipop sobre el gen APC, para marcar las mutaciones encontradas.

Luego utilizamos herramientas *in silico* para clasificar las variantes obtenidas. Debido a que son variantes tipo *frameshift* y producen un codón stop, no todas las herramientas pueden ser utilizadas para clasificar las mismas, por lo cual realizamos una búsqueda de aquellas herramientas que nos permitieran analizar la variante. La tabla 6 resume las herramientas utilizadas junto con las conclusiones/clasificaciones de cada una de ellas.

Tabla 6: Resumen del análisis *in silico* de las variantes obtenidas

Muestra	Mutation Taster	CRAVAT Impacto de patogenicidad (VEST indels)	MutPred Indel score	SIFT
02	Variante deletérea	0.971	0.66496	Perjudicial
05	Variante deletérea	0.74	0.51186	Perjudicial
13	Variante deletérea	0.749	0.69953	Perjudicial
23	Variante deletérea	0.947	0.63913	Perjudicial
32	Variante deletérea	0.729	0.66131	Perjudicial
33	Variante deletérea	0.937	0.57662	Perjudicial

La herramienta Mutation Taster permite realizar la búsqueda tanto en cromosoma como por transcripto, en este caso utilizamos la búsqueda por transcripto además de clasificarla nos da información de otras bases de datos como HGMD. Los resultados obtenidos fueron: **a)** 02: proteína truncada, afecta las características de la proteína, frameshift, cambia la secuencia de aminoácidos, existen tres variantes encontradas en la misma posición (HGMD CD058146, HGMD CD995185, HGMD CI011217); **b)** 05: cambia la secuencia de aminoácidos, afecta las características de la proteína, existen tres variantes encontradas en la base de datos HGMD en la mis ma posición (HGMD CI130131, HGMD CM920054, HGMD CM940072); **c)** 13: cambia la secuencia de aminoácidos, frameshift, afecta las características de la proteína y lleva a una proteína truncada, existe una variante encontrada en HGMD (HGMD CI106513); **d)** 23: cambio de la secuencia de aminoácidos, frameshift, afecta las características de la proteína, produce una proteína truncada y hay dos variantes encontradas en el mismo sitio reportadas en HGMD (HGMD CD972010, HGMD CP995074); **e)** 32: cambio de la secuencia de aminoácidos, frameshift, degradación del ARN mensajero mediada por mutación terminadora (NMD, mecanismo celular de vigilancia que evita la expresión de proteínas truncadas o erróneas), afecta las características de la proteína y hay dos variantes encontradas en la misma posición en HGMD (HGMD CD041143, HGMD CI130113); **f)** 33: cambio en la secuencia de aminoácidos, frameshift, afecta las características de la proteína, proteína truncada, hay una variante reportada en la misma posición en HGMD (HGMD CM080064).

La herramienta CRAVAT nos permite realizar la búsqueda de las mutaciones en conjunto a diferencia de la herramienta anterior que analiza por variante individual, en este caso los resultados obtenidos es un score de patogenicidad que va de 0 (benigna) a 1 (patogénica), brindando datos del tipo de variante, información del gen, bases de datos poblacionales,

COSMIC y pubmed que relacionan tanto la información del gen como de las variantes (C. Douville et al., 2016). En el análisis obtuvimos valores con un p>0.05 es decir que son patogénicas o probablemente patogénica según esta herramienta: **a)** 02: 0,971 con un p=0,0001; **b)** 05: 0,74 con un p=0,0038; **c)** 13: 0,749 con un p=0,0017; **d)** 23: 0,947 con un p=0,0001; **e)** 32: 0,729 con un p=0,0024; **f)** 33: 0,937 con un p=0,0002.

Otra herramienta que utilizamos fue MutPred-Indel también se carga por variante individual y se debe tener la secuencia en formato fasta de la proteína tanto *wild type* como la mutada. El puntaje que brinda es el promedio de los puntajes de todas las redes neuronales en MutPred-Indel, un umbral de puntuación de 0,50 sugeriría patogenicidad (con un 10% de falso positivos), a su vez, brinda valores de probabilidad (p<0,05) por propiedad/característica, por ejemplo, lugares de fosforilación, unión a proteínas, etc. Los resultados fueron: **a)** 02: valor de score de 0.66496 con valores de p según las características fosforilación (p=0.0001475); ADP-ribosilacion (p=0.00028854); desorden VSL2B (p=0.00066867); loop intracelular (p=0.00072127); Anclaje glicosilfosfatidilinositol (GPI) (p=0.00079421); **b)** 05: valor de score de 0.51186 con valores de p para fosforilación (p=0.02677); **c)** 13: valor de score de 0.69953 con valores de p según las características fosforilacion (p=7.3748e-05); ADP-ribosilacion (p=0.00028854); desorden VSL2B (p=0.00030394); loop intracelular (p=0.00036063); factor B (p=0.00047964); **d)** 23: valor de score de 0.63913 con valores de p según las características fosforilación (p=0.00022124); Anclaje glicosilfosfatidilinositol (GPI) (p=0.00031769); desorden VSL2B (p=0.00085103); loop intracelular (p=0.00086552); Accesibilidad superficial (p=0.0012318); **e)** 32: valor de score de 0.66131 con valores de p según las características Fosforilación (p=0); desorden VSL2B (p=6,0788e-05); Bucle intracelular (p=0,00014425); factor B (p=0,00020556); Bucle(p=0.00022363); **f)** 33: valor de score de 0.57662 con valores de p según las características fosforilación (p=0,00036874); desorden VSL2B (p=0,00085103); Bucle intracelular (p=0,00086552); ADP-ribosilación (p=0,00086562); Ácido pirrolidona carboxílico (p=0,001167). La herramienta SIFT nos permite realizar la búsqueda de hasta 100 variantes al mismo tiempo, en este caso todas las variantes genéticas analizadas resultaron en un efecto deletéreo para la proteína.

4.2.1.8- Clínica de los pacientes

Para evidenciar el fenotipo de los pacientes teniendo en cuenta los resultados genéticos encontrados, en la tabla 7 resumimos algunas de las variables tomadas en cuenta. Como se

puede observar todos los pacientes presentan un diagnóstico clínico de PAF, y como excepción del caso 05, todos presentan antecedente familiar ligado a la patología en estudio. Al estar incluidos en un programa de cáncer hereditario se puede diagnosticar y hacer el seguimiento tanto del probando como de su grupo familiar.

El fenotipo de los pacientes es bastante diferente. Considerando el efecto deletéreo a nivel proteico de las alteraciones genéticas identificadas en los pacientes 02, 05, 23 y 33, a pesar de estar localizadas en regiones adyacentes se evidencia que estos presentan manifestaciones extracolónicas diferentes; uno con diagnóstico de CCR y otro sin manifestaciones de cáncer. En el caso de la alteración genética identificada como causal de efecto deletéreo más severo en la proteína, esto es el truncamiento al principio de la proteína *APC*, el paciente portador (caso 032) no presenta CCR ni manifestaciones extracolónicas.

Varias publicaciones han evaluado la posible asociación entre la posición de la alteración genética dentro del gen con el fenotipo clínico, ya sea el tipo de poliposis (atenuada, intermedia y severa), como también las manifestaciones extracolónicas. En caso de la poliposis severa (más de 1000 pólipos colorrectales) suele observarse en pacientes con variantes entre los codones 1250 y 1464 (exón 15) (de Oliveira et al., 2019), en cuyo caso los pacientes 02, 05, 23 y 33 podría presentar este tipo de poliposis dado que la mutación cae dentro de ese rango. En caso de AFAP, generalmente se atribuye a variantes ubicadas en los extremos del gen APC, o en el exón 9 (Rossanese, Marson, Ribeiro, Coy, & Bertuzzo, 2013). Tumores desmoides parecen limitarse a pacientes con cambios en el mapeo entre los codones 1403 y 1578. Se ha observado que las manifestaciones extracolónicas (tumores desmoides y pólipos gástricos) están causadas por variantes situadas entre los codones 1445 y 1578 o entre los codones 1395 y 1493 (Fearnhead, Britton, & Bodmer, 2001), en este trabajo el paciente 23 presenta pólipos gástricos como también desmoides y la variante se encuentre en la posición 1395 dentro de este rango.

Tabla 7: Resumen de los fenotipos

Probando	Poliposis	Edad al diagnóstico	Manifestaciones extracolónicas	CCR
02	PAF	49	Adenoma duodeno	NO
05	PAF	24	-	SI
13	PAF	34	-	NO
23	PAF	35	Polipos gástricos Desmoide	NO
32	PAF	48	-	NO
33	PAF	14	-	NO

4.2.2 Análisis de muestras negativas para *APC*

Como se explicó anteriormente las muestras que resultaron negativas en la secuenciación de Sanger y estudio por MLPA fueron analizadas posteriormente por secuenciación NGS de exoma completo. Las variantes encontradas en las muestras se resumen en la tabla 8 como podemos observar en algunos casos son genes relacionados directamente con la patología y en otros son genes del panel 2, es decir directa e indirectamente relacionados. En el caso de las

muestras 01, 03, 08, 18, 29, 46 y 49 no pudimos encontrar variantes genéticas de predisposición (causales) en genes relacionados con la patología que pudieran explicar el fenotipo del paciente o nuevos genes candidatos.

Tabla 8: Resumen de los resultados obtenidos

M	Pólipo	AF	Gen	Posición	Cambio de nucleótido	Cambio de aminoácido	Efecto	Cigosidad	Cob.
01	FAP (>100)	NO	-	-	-	-	-	-	-
03	AFAP (20)	SI	-	-	-	-	-	-	-
06	Mixed (20)	NO	SMAD4	18:48584569	C → T	p. Gln248*	*stop-gained*	HET	99/216
08	FAP (>100)	NO	-	-	-	-	-	-	-
18	FAP (>100)	NO	-	-	-	-	-	-	-
19	FAP (>100)	NO	BMPR1A	10:88683182	TC → T	p.Pro465fs	*frameshif*	HET	85/185
24	FAP (>100)	NO	MUTYH	1:45799144	G → A	p.Arg97*	*stop-gained*	HET	57/133
				1:45797186	T → TCC	p.Glu410fs	*frameshif*	HET	126/277
27	Juvenil (>100)	NO	SMAD4	18:48586236	G → A	p.Trp302*	*stop-gained*	HET	19/58
29	AFAP (<100)	NO	-	-	-	-	-	-	-
37	FAP (>100)	SI	NUDT7	16:77756501	G → T	p.Glu8*	*stop-gained*	HET	30/66
41	AFAP (20)	NO	MUC17	7:100676569	C → G	p.Tyr624*	*stop-gained*	HET	73/169
46	Mixed (20)	SI	-	-	-	-	-	-	-
48	AFAP (20)	NO	MSH6	2:48028145	C → T	p.Thr1008Ile	*missense*	HET	29/65
49	AFAP (30)	NO	-	-	-	-	-	-	-

4.2.2.1 Variantes encontradas en el gen *SMAD4*

4.2.2.1.1 Caso 06

La posición de la variante encontrada es 18:48584569 sobre el gen *SMAD4*, el cambio a nivel ADN es C→T y a nivel proteico p.Gln248* (c.742C>T). El efecto es *stop-gained* con una buena cobertura de 99/216 en heterocigosis (**figura 15**), el criterio de ACMG la cataloga como patogénica. Esta variante se encuentra reportada en ClinVar con una estrella.

Un trabajo reporta una variante con pérdida de función en la posición p.Gln249Ter a nivel germinal con diagnóstico de síndrome polipósico juvenil. Posee antecedentes familiares los cuales presentaron cáncer colorrectal y pólipos, pero no se realizó la verificación de la variante en sus familiares. El probando también presentó múltiples pólipos en el estómago (Pantelis et al., 2016). La misma mutación fue encontrada (p.Gln249Ter) a nivel somático en la posición chr18:g.48584572C>T (GRCh37), reportada en la base de datos de cáncer ICGC

Somatic, dentro del proyecto COCA-CN (cáncer colorrectal) el número de identificación del donante es MU81649743.

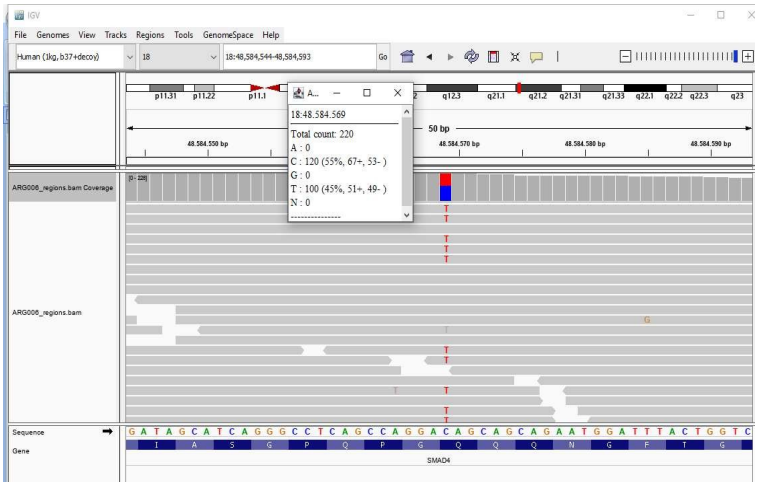

Figura 15. Cobertura de la variante encontrada. Visualización del archivo BAM de la muestra 006 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.1.2 Caso 27

La posición de la variante encontrada es 18:48586236 y a nivel proteico p.Trp302* (c.905G>A). El efecto es *stop-gained* con una cobertura moderada de 19/58 en heterocigosis (**figura 16**), el criterio de ACMG la cataloga como patogénica. Esta variante se encuentra reportada en ClinVar con una estrella. A su vez, en la base dbSNP contiene número de identificación rs1555686071.

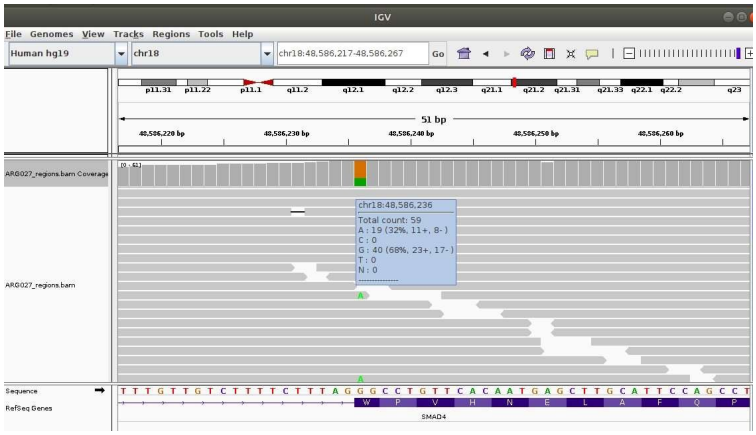

Figura 16. Cobertura de la variante encontrada. Visualización del BAM de la muestra 27 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.2.3 Análisis *in silico* de las variantes encontradas

Para interpretar el efecto de las variantes sobre la proteína realizamos un gráfico tipo lollipop para representar las mutaciones y a su vez reconocer los dominios de esta proteína. Este gen cumple dos funciones, como factor de transcripción ya que el dominio MH1 le permite unirse al ADN y como supresor de tumores ya que interacciona con proteínas por medio de homo o hetero dimerización a través de su dominio MH2. Como podemos observar ambas mutaciones llevan a una proteína truncada perdiendo el dominio MH2.

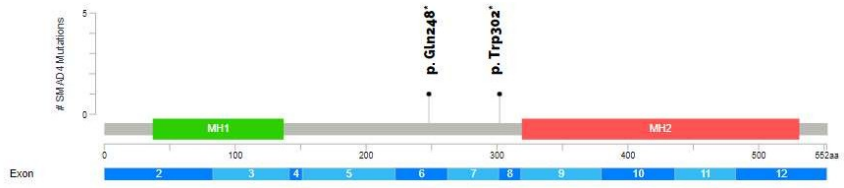

Figura 17. Representación de las variantes sobre *SMAD4*. Estructura de la proteína con sus respectivos dominios, debajo se grafican los exones de la misma y en lollipop las variantes.

La herramienta _B Platform permite asociar información de diferentes predictores *in silico* con la variante encontrada, los cuales se encuentran resumidos en la tabla 9 para cada una de las variantes: a) 06: mutation taster predice un cambio de secuencia de aminoácidos, afecta las características de la proteína y NMD llevando a clasificarla como causante de enfermedad, DANN da un valor cercano a 1 por lo tanto también la clasifica como patogénica; b) 27: mutation taster predice un cambio de secuencia de aminoácidos, afecta las características de la proteína, puede producir un cambio en sitio de splicing y NMD llevando a clasificarla como causante de enfermedad, al igual que los demás predictores la clasifican como patogénica.

Tabla 9: Resumen del análisis de las variantes *in silico*

Predictor	Muestra 06	Muestra 27
Intervar	Patogénica	Patogénica
CADD	Posible patogénica	Posible patogénica
DANN	0.998	0.994
Mutation Taster	Causante de enfermedad	Causante de enfermedad

4.2.2.2 Caso 10

La posición de la variante encontrada es 5:112154999 sobre el gen *APC* y a nivel proteico E425Gfs*4 (c.1271dupA). El efecto es frameshift con una buena cobertura de 43/89

en heterocigosis (**figura 18**), el criterio de ACMG la cataloga como probablemente patogénica. Esta variante no se encuentra reportada en bases de datos, por lo cual es una variante *novel*.

Evaluando la literatura, se encuentra una variante en la misma posición de la proteína p.E425* reportada en la base de datos Cancer Genome Atlas Network en la muestra tumoral identificada como TCGA-AF-6655 a nivel somático (Nature, 2012). Una línea celular Cellosaurus DiFi (CVCL_6895) (expasy), presenta la variante p.E425fs*15 a nivel proteico (Russo et al., 2018). La mutación p.E425* ha sido reportada a nivel germinal en otros trabajos en pacientes con cáncer colorrectal (Lagarde et al., 2010; Olschwang, Laurent-Puig, Groden, White, & Thomas, 1993).

La evaluación *in silico* de la variante arroja: **a)** mutation taster cambio en la secuencia de aminoácidos, *frameshift*, afecta las características de la proteína, cambio en sitio de splicing y NMD por lo tanto la clasifica como causante de la patología y se encuentra una variante en la misma posición en HGMD (HGMD CM930021); **b)** Intervar posiblemente patogénica; **c)** MutPred-Indel con un score de 0.71012 y valor de p para las características de fosforilación (p=0,0001475); amidación de anclaje GPI (p = 0,00015884); trastorno VSL2B (p=0,00018236); Bucle intracelular (p=0,00021638); Bucle(p=0.00027954).

A partir de los datos clínicos del paciente observamos que fue diagnosticado con PAF y CCR a la edad de 29 años.

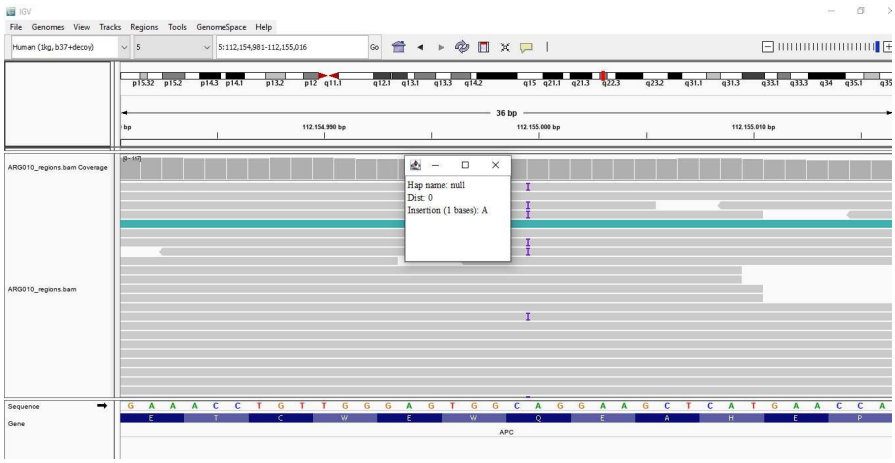

Figura 18. Cobertura de la variante encontrada. Visualización del BAM de la muestra 010 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.3 Caso 19

La posición de la variante encontrada es 10:88683182 dentro del gen *BMPR1A* y a nivel proteico p.Pro465Argfster33 (c.1394delC), con una buena cobertura de 85/185 (**figura 19**). El efecto es *frameshift* en heterocigosis, el criterio de ACMG la cataloga como probablemente patogénica. Esta variante no se encuentra reportada en bases de datos, por lo cual es una variante *novel*.

Hay dos variantes reportadas por ClinVar con dos estrellas en posiciones aledañas a la encontrada en este probando a nivel germinal clasificadas en Varsome como patogénicas. Una de ellas en la posición 10:88683388 con un cambio en la proteína p.Trp504Ter (NM_004329.2:c.1511G>A) que es reportada en dos probandos uno con síndrome de poliposis juvenil y otro con síndrome predisponente a cáncer hereditario, clasificada como patogénica. La segunda reportada en la posición 10:88683357 con un cambio en la proteína p.Arg494Ter (NM_004329.2: c.1480C>T, COSM197910) reportada en dos probandos uno síndrome predisponente a cáncer hereditario y otro sin descripción, clasificada como probablemente patogénica.

Esta variante analizada *in silico*: **a)** Mutation taster predice cambios en la secuencia de aminoácidos, frameshift, afecta las características de la proteína y proteína truncada por lo cual la clasifica como causante de la enfermedad; **b)** CADD v2.1 le asigna un valor de rawdata de 2.56 y Phred 22.7 la cual también la asigna como probablemente patogénica; **c)** SIFT lo clasifica como deletérea.

A nivel clínico, el caso se diagnostica con PAF a la edad de 25 años presentando cáncer de colon a los 39 años, y sin antecedentes familiares para síndrome hereditario. En consecuencia la clínica característica esperada para la alteración en la función de este gen *BMPR1A* es aquella asociada a síndrome de poliposis juvenil cuyo pólipo se clasifica como harmatomatoso (morfología característica) (Arévalo et al., 2012) y desarrollo a muy temprana edad. Este resultado no es concordante con la clínica que presenta el paciente, y por lo tanto no se puede inclinar la clasificación clínica de la variante como genético-causal.

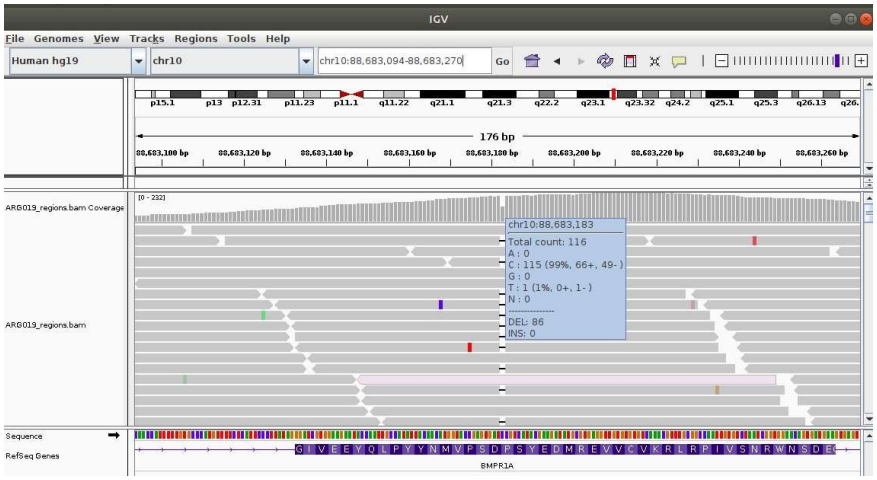

Figura 19. Cobertura de la variante encontrada. Visualización del BAM de la muestra 19 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.4 Caso 24

Se identifican dos variantes genéticas localizadas en el gen *MUTYH* el cual tiene una herencia recesiva, a diferencia de los demás genes antes mencionados, por ello son necesarias dos alteraciones simultáneas bialélicas para que se observe el fenotipo.

Las variantes se ubicaron una en la posición 1:45799144 y a nivel proteico p.Arg97* (c.289C>T), con una buena cobertura de 57/133 (**figura 20A**). El efecto es *stop gained* en heterocigosis, el criterio de ACMG la cataloga como patogénica. Esta variante se encuentra reportada en ClinVar con dos estrellas y en la base de datos dbSNP (rs138775799). Mientras que la otra en la posición 1:45797186 y a nivel proteico p.Glu410GlyfsTer43 (c.1227_1228dupGG), con una buena cobertura de 126/277 (**figura 20B**). El efecto es *frameshift* en heterocigosis, el criterio de ACMG la cataloga como probablemente patogénica. Esta variante se encuentra reportada en ClinVar con dos estrellas y en la base de datos dbSNP (rs587780078).

La clasificación de herramientas *in silico* de estas variantes obtenidas fueron: a) p.Arg97* (c.289C>T): mutation taster la clasifica como causante de enfermedad y agrega una variante reportada en la misma posición en HGMD (HGMD CM030378), InteVar y DANN como patogénica; b) p.Glu410GlyfsTer43 (c.1227_1228dupGG): InterVar la clasifica como probablemente patogénica y mutation taster con cambio en la secuencia de aminoácidos, frameshift, afecta las características de la proteína, cambio en el sitio de splicing, NMD

clasificándola como causante de enfermedad y reporta dos variantes en la misma posición en HGMD (HGMD CI041961; HGMD CI068635).

A nivel clínico el probando presenta un diagnóstico de PAF a la edad de 37 años y CCR sin antecedentes familiares.

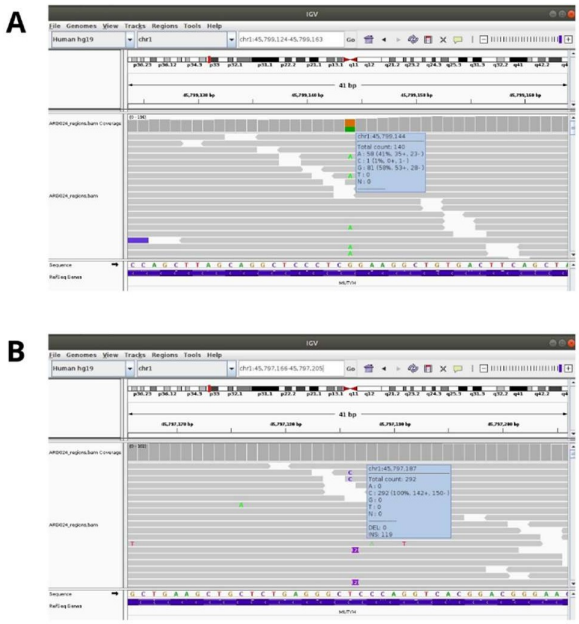

Figura 20. Cobertura de la variante encontrada. A. Variante en la posición 1:45799144C>T **B.** Variante en la posición 1:45797186insGG. Visualización del BAM de la muestra 24 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.5 Caso 37

La posición de la variante encontrada es 16:77756501 sobre el gen *NUDT7* y a nivel proteico p.Glu8* (c.22G>T), con una cobertura moderada de 30/66 (**figura 21**). El efecto es *stop gained* en heterocigosis, el criterio de ACMG la cataloga como significado incierto. Esta variante se encuentra reportada en bases de datos con ID rs182579196.

Esta variante se encuentra reportada en ClinVar con una estrella clasificada como benigna no posee una descripción sobre su procedencia. Varsome la clasifica como probablemente benigna. Un trabajo el cual estudió 96 casos familiares independientes diagnosticados con CCR de origen Finlandes, pudieron identificar 11 genes candidatos predisponentes de CCR (*UACA*, *SFXN4*, *TWSG1*, *PSPH*, ***NUDT7***, *ZNF490*, *PRSS37*, *CCDC18*, *PRADC1*, *MRPL3* y *AKR1C4*) (Gylfe et al., 2013), la variante encontrada sobre el gen *NUTD7*

fue c.111T>A cuya muestra es de un probando de sexo masculino de 72 años diagnosticado con cáncer en colon transverso. A su vez, otro trabajo que estudia pacientes con CCR severo testeados con el panel de genes de síndrome de Lynch, reportan 14 genes con un rol potencial a la susceptibilidad de CCR (*AXIN1*, *BMP4*, *CCDC18*, ***NUDT7***, *PICALM*, *PTPRJ*, *SLC5A9*, *TLR2*, *TWSG1*, *UBAP2*, *USP6NL* y *ZFP14*) (Hansen et al., 2017). También se encuentra reportada la variante en cáncer de hígado y vejiga (2 probandos), como mutación somática (MU4609066).

El análisis *in silico* de esta variante InterVar la considera de significado incierto, CADD como benigna, DANN como patogénica y mutation taster como causante de enfermedad.

La proteína codificada por este gen es miembro de la familia de hidrolasas Nudix. Las hidrolasas de Nudix eliminan los metabolitos de nucleótidos potencialmente tóxicos de la célula y regulan las concentraciones y la disponibilidad de muchos sustratos de nucleótidos, cofactores y moléculas de señalización diferentes (Gasmi & McLennan, 2001). Por lo cual esta variante podría tener relevancia biológica en este probando.

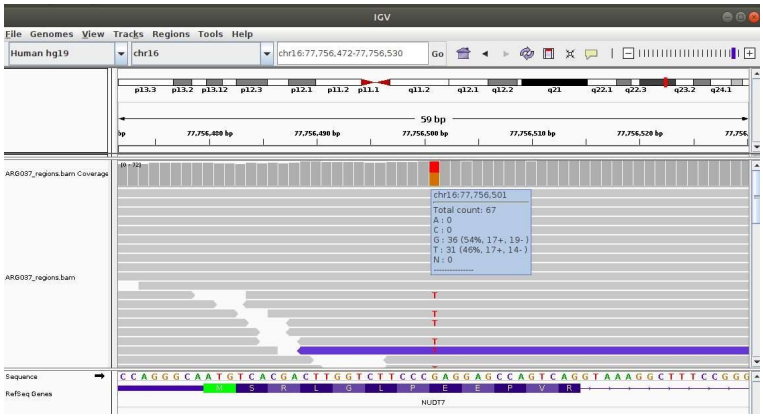

Figura 21. Cobertura de la variante encontrada. Visualización del BAM de la muestra 37 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.6 Caso 41

La posición de la variante encontrada es 7:100676569 dentro del gen *MUC17* y a nivel proteico p.Tyr624* (c.1872C>G), con una buena cobertura de 73/169 (**figura 22**). El efecto es *stop gained* en heterocigosis, el criterio de ACMG la cataloga como significado incierto. Esta variante se encuentra reportada en dbSNP (rs745936745).

El análisis *in silico* de esta variante permite mediante InterVar clasificarla como de significado incierto, CADD como benigna, DANN y mutation taster como patogénica.

Varios estudios realizados en líneas celulares provenientes de cáncer colorrectal utilizando la técnica NGS RNAseq de pólipos serratos (tipo AFAP), (Delker et al., 2014; Senapati et al., 2010), demuestran una disminución de la producción de MUC17. Esta mucina cumple un rol crítico en el mantenimiento de la barrera intestinal, por lo tanto, este gen podría tener una relevancia biológica en esta patología. El probando presenta un diagnóstico de PAFA a los 56 años de edad sin presencia de manifestaciones extracolónicas o CCR.

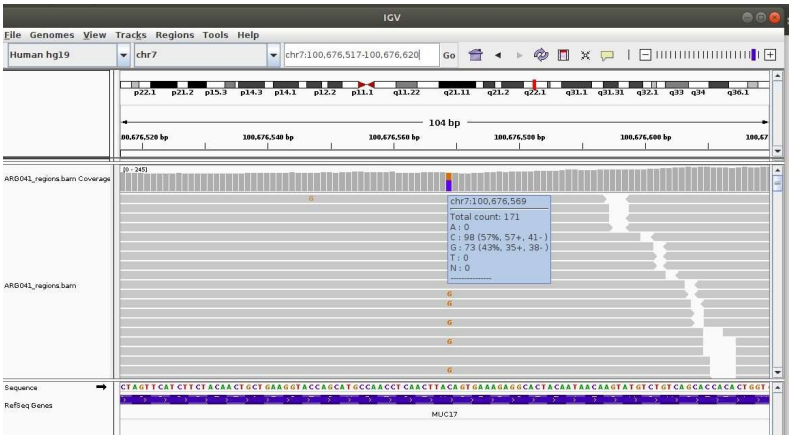

Figura 22 Cobertura de la variante encontrada. Visualización del BAM de la muestra 41 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.7 Caso 48

La posición de la variante encontrada es 2:48028145 dentro del gen *MSH6* y a nivel proteico p.Thr1008Ile (c.3023C>T), con una cobertura moderada de 29/65 (Fig. 23). El efecto es *missense* en heterocigosis, el criterio de ACMG la cataloga como significado incierto.

Hay dos trabajos que reportan la variante uno de ellos es la presentación de una herramienta bioinformática CoDP la cual clasifica esta variante como benigna (Terui, Akagi, Kawame, & Yura, 2013) y otra solo la reporta en una muestra de Argentina sin clasificar (Rossi et al., 2017). Sin embargo, en el año 2020 se realizó un trabajo de caracterización molecular de diferentes variantes clasificadas como de significado incierto, y entre las variantes analizadas, se evaluó la alteración que identificamos en *MSH6*. En el trabajo se observó una disminución en la expresión y/o actividad de la proteína, reclasificando la misma como patogénica (Drost

et al., 2020). En consecuencia, esta variante genética en *MSH6* podría ser la candidata genética-causal para este caso.

El análisis *in silico* de esta variante la clasifica por InterVar como significado incierto, CADD como benigna, mutation taster como causante de enfermedad y se encuentra reportada en ClinVar como de significado incierto con una estrella.

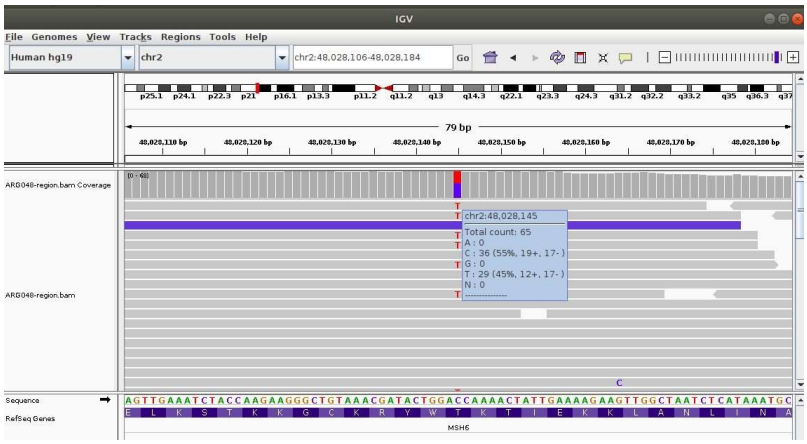

Figura 23. Cobertura de la variante encontrada. Visualización del BAM de la muestra 48 utilizando la herramienta IGV, obteniendo casi un 50 % de la representación de cada variante (referencia y alternativa)

4.2.2.8 Casos negativos

Como describimos anteriormente en las muestras 01, 03, 08, 18, 29, 46 y 49 no pudimos encontrar variantes en genes relacionados con la patología que pudieran explicar el fenotipo del paciente o nuevos genes candidatos que nos permitan relacionar de alguna manera el fenotipo.

Esto nos lleva a plantear diferentes caminos. Por un lado, en los pacientes 03, 08 y 46 que presentar CCR, se puede realizar el análisis de firmas mutacionales mediante el estudio con la técnica de NGS en muestras de sangre (germinal) y tumoral (somática), con el análisis en conjunto de ambos datos para obtener una firma o patrón característico que al comparar con las firmas mutacionales reportadas en la base de datos COSMIC permite inferir la vía o los genes potenciales involucrados en el fenotipo (Kucab et al., 2019).

Por otro lado, podríamos buscar dentro de las variantes que quedaron por fuera del experimento (no cubiertas) en genes que mediante la evidencia clínica podrían sospecharse como posibles candidatos.

5-Conclusiones

- ✓ La ejecución del análisis bioinformático permitió delinear y establecer el conjunto de etapas requeridas para completar el estudio, partiendo desde el dato crudo derivado del proceso de secuenciación masiva hasta la obtención del archivo VCF. Asimismo, determinar los parámetros involucrados en cada etapa y el proceso global

- ✓ El análisis bioinformático de porcentajes GC mediante la herramienta FAST Screen, ejecutada sobre las secuencias obtenidas, en conjunto con el gráfico obtenido permitió aprender el funcionamiento de esta herramienta y, más aún, su incorporación al pipeline desarrollado para nuestros estudios.

- ✓ Tomando la información obtenida a partir de la bibliografía disponible diseñamos dos tipos de paneles de genes relacionados directa o indirectamente con el fenotipo de las patologías abordadas en el estudio. De esta manera, en la ejecución del análisis diseñado, logramos establecer dos metas significativas, primero la no obtención de hallazgos incidentales, y segundo poder buscar de manera específica los genes y las variantes genéticas involucradas en los fenotipos estudiados.

- ✓ Diseñamos un pipeline de priorización de variables que nos permite hacer una búsqueda exhaustiva y direccionada en las muestras estudiadas

- ✓ Resolvimos más del 60% (14/21) de los casos incorporados al estudio y establecimos prioridades y variables para definir aquellos subconjuntos de casos que serán evaluados a futuro, con estudios complementarios y en mayor detalle.

Tanto las herramientas empleadas como el pipeline desarrollado, permitirá a nuestro grupo acercarnos hacia el objetivo de aplicar procedimientos ligados con la medicina de precisión, tanto en el área de investigación clínica como en los procedimientos ejecutados como parte de la clínica médica diaria. Esto posibilitará establecer nuevas estrategias para identificar, diagnosticar y estudiar un número mayor de casos con Síndrome de CCR hereditario.

6-Anexo

6.1 Soluciones para extracción de ADN

- **Solución A** (para 1 litro de solución)
  Sacarosa 0,32 M 109,5 gr
  MgCl 2 0,005M 1 gr
  Tris HCl 0,01 M (pH: 8) 10 ml (stock 1M)
  Tritón X 100 1% 10 ml

- **Proteinasa K** 10mg/ml (conservar a -20ºC)

- **Buffer TE**:
  Tris HCl 10 Mm
  EDTA 1 mM pH 8.0

- **Buffer de digestión**:
  NaCl 100 mM
  Tris-HCl 50 mM
  SDS 1%
  EDTA 50 mM pH 8.0

- **Cloruro de Litio** 5 M

- **Mezcla SEVAG**: Cloroformo / alcohol isoamílico (24:1)

6.2 Panel de genes 2: genes directa e indirectamente relacionados

Nombre	Localizacion	Herencia	Nombre	Localizacion	Herencia	Nombre	Localizacion	Herencia
ADAR	1q21.3	AD/AR	GAB1	4q31.21	AR	POLQ	3q13.33	-
AKR1C4	10p15.1	AR	GALNT12	9q22.33	-	PRADC1	2p13.2	-
AKT1	14q32.33	-	GFI1	1p22.1	AD	PREX1	20q13.13	-
ALPK2	18q21.31-q21.32	-	GLT1D1	12q24.33	-	PRSS37	7q34	-
ARHGAP5	14q12	-	GPR143	Xp22.2	XLR/XL	PSMC3IP	17q21.2	AR
ARSD	Xp22.33	-	GSK3B	3q13.33	-	PSPH	7p11.2	AR
ATM	11q22.3	AR/AD/SMu	H2BW2	Xq22.2	-	PSRC1	1p13.3	-
AXIN1	16p13.3	-	HELQ	4q21.23	-	PTCH1	9q22.32	AD
AXIN2	17q24.1	AD	HHH	13q14.11	AR	PTPRJ	11p11.2	-
BAP1	3p21.1	AD	IGF2	11p15.5	AD	RAD51D	17q12	-
BARD1	2q35	AD/SMu	IQGAP1	15q26.1	-	RANBP2	2q13	AD
BCAM	19q13.32	AR	KIF26B	1q44	-	RB1	13q14.2	AD/SMu
BIRC6	2p22.3	-	KLLN	10q23.31	-	RPS20	8q12.1	-
BLM	15q26.1	AR	KRAS	12p12.1	AD	RPS6KL1	14q24.3	-
BMP4	14q22.2	AD	LAMA2	6q22.33	AR	RUNX3	1p36.11	-
BRAF	7q34	AD	LAMA5	20q13.33	-	RYR2	1q43	AD
BRCA1	17q21.31	AR/AD/SMu	LIG1	19q13.33	AR	SEMA4A	1q22	AR/AD
BRCA2	13q13.1	AR/AD/SMu	LRP1B	2q22.1-q22.2	-	SETD6	16q21	-
BRD9	5p15.33	-	LRP6	12p13.2	AD	SFXN4	10q26.11	AR
BRF1	14q32.33	AR	LZTFL1	3p21.31	AR	SH2B3	12q24.12	-
BUB1	2q13	-	MAP1B	5q13.2	AD	SHROOM2	Xp22.2	-
BUB1B	15q15.1	AR/AD	MAP3K1	5q11.2	AD	SHROOM3	4q21.1	-
BUB3	10q26.13	-	MAPRE2	18q12.1-q12.2	AD	SIGLEC10	19q13.41	-
C11ORF53	11q23.1	-	MCC	5q22.2	-	SMAD7	18q21.1	-
CACNA1G	17q21.33	AD	MCM8	20p12.3	AR	SMAD9	13q13.3	AD
CCDC148	2q24.1	-	MCM9	6q22.31	AR	SMARCA4	19p13.2	AD
CCDC18	1p22.1	-	MLH1	3p22.2	AR/AD	SMO	7q32.1	AR
CDC27	17q21.32	-	MLH3	14q24.3	AD/SMu	SNCAIP	5q23.2	-
CDH1	16q22.1	AD/SMu	MMP9	20q13.12	AR	SOCS1	16p13.13	AD
CDKN1B	12p13.1	AD	MPG	16p13.3	-	SOX9	17q24.3	AD
CDKN2A	9p21.3	AD	MRGPRX4	11p15.1	-	SRC	20q11.23	AD
CHEK2	22q12.1	AD/SMu	MRPL3	3q22.1	AR	STK11IP	2q35	-
CNTN6	3p26.3	-	MSH2	2p21-p16.3	AR/AD	TCEAL6	Xq22.2	-
COLCA1	11q23.1	-	MSH6	2p16.3	AR/AD/SMu	TDG	12q23.3	-
COLCA2	11q23.1	-	MUC17	7q22.1	-	TESK2	1p34.1	-
CSMD3	8q23.3	-	MYC	8q24.21	-	TFDP2	3q23	-
CTNNB1	3p22.1	AD	MYH11	16p13.11	AR/AD	TGFB1	19q13.2	AR/AD
DCAF12L2	Xq25	-	MYH6	14q11.2	AD	TGFBR2	3p24.1	AD
DIP2B	12q13.12	AD	NBN	8q21.3	AR	TMC2	20p13	-
DKK1	10q21.1	-	NDFIP1	5q31.3	-	TP53	17p13.1	AD/SMu
DOCK11	Xq24	-	NDFIP2	13q31.1	-	TTN	2q31.2	AR/AD
DSC2	18q12.1	AD/AR	NEIL1	15q24.2	-	TUBB6	18p11.21	AD
EDA2R	Xq12	-	NEIL2	8p23.1	-	TWSG1	18p11.22	-
EDRF1	10q26.2	-	NEUROG1	5q31.1	-	UACA	15q23	-
ENG	9q34.11	AD	NFATC1	18q23	-	UNC5C	4q22.3	-
EPCAM	2p21	AD/AR	NFKBIZ	3q12.3	-	UNG	12q24.11	AR
EPHB2	1p36.12	AR	NLRP1	17p13.2	AR/AD	VEGFB	11q13.1	-
EPHB4	7q22.1	AD	NOS1	12q24.22	-	WDR78	1p31.3	-
EPHB6	7q34	-	NOTCH3	19p13.12	AD	XRCC4	5q14.2	AR
EPHX1	1q42.12	-	NUDT7	16q23.1	-	XRN1	3q23	-
EVI5	1p22.1	-	OBSCN	1q42.13	-	YES1	18p11.32	-
FAN1	15q13.3	AR	PALB2	16p12.2	AD/SMu	ZBTB2	6q25.1	-
FBXW7	4q31.3	-	PCDH17	13q21.1	-	ZNF490	19p13.2-p13.13	-
FOCAD	9p21.3	-	PDE4D	5q11.2-q12.1	AD	ZSWIM7	17p12	AR
FOXM1	12p13.33	-	PIEZO1	16q24.3	AR/AD			
FOXO1	13q14.11	-	PIK3C2G	12p12.3	-			
FOXO3	6q21	-	PIK3CA	3q26.32	-			
FOXP3	Xp11.23	XLR	PMS1	2q32.2	-			
FSTL5	4q32.2	-	PMS2	7p22.1	AR			

7-Referencias

Altshuler, D., Donnelly, P., & Nature, I. H. C. J. (2005). A haplotype map of the human genome. *437*(7063), nature04226.

Aretz, S. J. D. A. I. (2010). The differential diagnosis and surveillance of hereditary gastrointestinal polyposis syndromes. *107*(10), 163.

Arévalo, F., Aragón, V., Alva, J., Perez Narrea, M., Cerrillo, G., Montes, P., & Monge, E. J. R. d. G. d. P. (2012). Pólipos colorectales: actualización en el diagnóstico. *32*(2), 123-133.

Armstrong, J. G., Davies, D. R., Guy, S. P., Frayling, I. M., & Evans, D. G. R. J. H. m. (1997). APC mutations in familial adenomatous polyposis families in the Northwest of England. *10*(5), 376-380.

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Balmaña, J., Digiovanni, L., Gaddam, P., Walsh, M. F., Joseph, V., Stadler, Z. K., . . . Domchek, S. M. (2016). Conflicting Interpretation of Genetic Variants and Cancer Risk by Commercial Laboratories as Assessed by the Prospective Registry of Multiplex Testing. *J Clin Oncol, 34*(34), 4071-4078. doi:10.1200/jco.2016.68.4316

Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B., & Klein, T. E. (2018). PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdiscip Rev Syst Biol Med, 10*(4), e1417. doi:10.1002/wsbm.1417

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene, 241*(1), 3-17. doi:https://doi.org/10.1016/S0378-1119(99)00485-0

Bitgenia. Retrieved from https://www.bitgenia.com/b-platform/

Briggs, S., & Tomlinson, I. J. T. J. o. p. (2013). Germline and somatic polymerase ϵ and δ mutations define a new class of hypermutated colorectal and endometrial cancers. *230*(2), 148-153.

Bujanda, L., Cosme, A., Gil, I., & Arenas-Mirave, J. I. (2010). Malignant colorectal polyps. *World J Gastroenterol, 16*(25), 3103-3111. doi:10.3748/wjg.v16.i25.3103

Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., . . . Lyon, E. J. G. (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous polyposis. *127*(2), 444-451.

Byrne, R. M., & Tsikitis, V. L. J. A. o. g. (2018). Colorectal polyposis and inherited colorectal cancer syndromes. *31*(1), 24.

Cai, S. R., Zhang, S. Z., & Zheng, S. (2008). [Detection of adenomatous polyposis coli gene mutations in 31 familial adenomatous polyposis families by using denaturing high performance liquid chromatography]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi, 25*(2), 164-167.

Cairns, S. R., Scholefield, J. H., Steele, R. J., Dunlop, M. G., Thomas, H. J., Evans, G. D., . . . Saunders, B. P. J. G. (2010). Guidelines for colorectal cancer screening and surveillance in moderate and high risk groups (update from 2002). *59*(5), 666-689.

Christie, M., Jorissen, R. N., Mouradov, D., Sakthianandeswaren, A., Li, S., Day, F., . . . Sieber, O. M. (2013). Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT/β-catenin signalling thresholds for tumourigenesis. *Oncogene, 32*(39), 4675-4682. doi:10.1038/onc.2012.486

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. J. F. (2012). A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *6*(2), 80-92.

Cleary, S. P., Cotterchio, M., Jenkins, M. A., Kim, H., Bristow, R., Green, R., . . . Lindor, N. J. G. (2009). Germline MutY human homologue mutations and colorectal cancer: a multisite case-control study. *136*(4), 1251-1260.

Cruz-Correa, M., Pérez-Mayoral, J., Dutil, J., Echenique, M., Mosquera, R., Rivera-Román, K., . . . on behalf of the Puerto Rico Clinical Cancer Genetics, C. (2017). Hereditary cancer syndromes in Latino populations: genetic characterization and surveillance guidelines. *Hereditary Cancer in Clinical Practice, 15*(1), 3. doi:10.1186/s13053-017-0063-z

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Sherry, S. T. J. B. (2011). The variant call format and VCFtools. *27*(15), 2156-2158.

de Oliveira, J. C., Viana, D. V., Zanardo, C., Santos, E. M. M., de Paula, A. E., Palmero, E. I., & Rossi, B. M. (2019). Genotype-phenotype correlation in 99 familial adenomatous polyposis patients: A prospective prevention protocol. *Cancer Medicine, 8*(5), 2114-2122. doi:https://doi.org/10.1002/cam4.2098

Del Vecchio, F., Mastroiaco, V., Di Marco, A., Compagnoni, C., Capece, D., Zazzeroni, F., . . . Tessitore, A. J. J. o. T. M. (2017). Next-generation sequencing: Recent applications to the analysis of colorectal cancer. *15*(1), 1-19.

Delker, D. A., McGettigan, B. M., Kanth, P., Pop, S., Neklason, D. W., Bronner, M. P., . . . Hagedorn, C. H. (2014). RNA sequencing of sessile serrated colon polyps identifies differentially expressed genes and immunohistochemical markers. *PLoS One, 9*(2), e88367. doi:10.1371/journal.pone.0088367

Dinarvand, P., Davaro, E. P., Doan, J. V., Ising, M. E., Evans, N. R., Phillips, N. J., . . . medicine, l. (2019). Familial adenomatous polyposis syndrome: an update and review of extraintestinal manifestations. *143*(11), 1382-1398.

Dominguez-Valentin, M., Sampson, J. R., Seppälä, T. T., Ten Broeke, S. W., Plazzer, J.-P., Nakken, S., . . . Sunde, L. J. G. i. M. (2020). Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. *22*(1), 15-25.

Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P. D., . . . Karchin, R. J. B. (2013). CRAVAT: cancer-related analysis of variants toolkit. *29*(5), 647-648.

Douville, C., Masica, D. L., Stenson, P. D., Cooper, D. N., Gygax, D. M., Kim, R., . . . Karchin, R. (2016). Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum Mutat, 37*(1), 28-35. doi:10.1002/humu.22911

Drost, M., Tiersma, Y., Glubb, D., Kathe, S., van Hees, S., Calléja, F., . . . de Wind, N. (2020). Two integrated and highly predictive functional analysis-based procedures for the classification of MSH6 variants in Lynch syndrome. *Genet Med, 22*(5), 847-856. doi:10.1038/s41436-019-0736-2

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047-3048. doi:10.1093/bioinformatics/btw354 %J Bioinformatics

Fearnhead, N. S., Britton, M. P., & Bodmer, W. F. (2001). The ABC of APC. *Human Molecular Genetics, 10*(7), 721-733. doi:10.1093/hmg/10.7.721 %J Human Molecular Genetics

Ferrarini, A., Xumerle, L., Griggio, F., Garonzi, M., Cantaloni, C., Centomo, C., . . . Collino, S. J. P. o. (2015). The use of non-variant sites to improve the clinical assessment of whole-genome sequence data. *10*(7), e0132180.

Fodde, R., Smits, R., & Clevers, H. (2001). APC, Signal transduction and genetic instability in colorectal cancer. *Nature Reviews Cancer, 1*(1), 55-67. doi:10.1038/35094067

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics, 10*(4), 241-251. doi:10.1038/nrg2554

Friedl, W., & Aretz, S. (2005). Familial Adenomatous Polyposis: Experience from a Study of 1164 Unrelated German Polyposis Patients. *Hereditary Cancer in Clinical Practice, 3*(3), 95. doi:10.1186/1897-4287-3-3-95

Friedl, W., Caspari, R., Sengteller, M., Uhlhaas, S., Lamberti, C., Jungck, M., . . . Propping, P. (2001). Can APC mutation analysis contribute to therapeutic decisions in familial adenomatous polyposis? Experience from 680 FAP families. *48*(4), 515-521. doi:10.1136/gut.48.4.515 %J Gut

Galiatsatos, P., Foulkes, W. D. J. O. j. o. t. A. C. o. G., & ACG. (2006). Familial adenomatous polyposis. *101*(2), 385-398.

Gasmi, L., & McLennan, A. G. (2001). The mouse Nudt7 gene encodes a peroxisomal nudix hydrolase specific for coenzyme A and its derivatives. *Biochem J, 357*(Pt 1), 33-38. doi:10.1042/0264-6021:3570033

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., . . . Shen, Y. (2003). The international HapMap project.

Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., . . . Robertson, M. J. C. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. *66*(3), 589-600.

Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., . . . O'Donnell-Luria, A. J. H. m. (2022). Variant interpretation using population databases: Lessons from gnomAD. *43*(8), 1012-1030.

Gylfe, A. E., Katainen, R., Kondelin, J., Tanskanen, T., Cajuso, T., Hänninen, U., . . . Järvinen, H. J. P. g. (2013). Eleven candidate susceptibility genes for common familial colorectal cancer. *9*(10), e1003876.

Hansen, M. F., Johansen, J., Sylvander, A. E., Bjørnevoll, I., Talseth-Palmer, B. A., Lavik, L. A., . . . Drabløs, F. J. C. g. (2017). Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. *92*(4), 405-414.

Heald, B., Hampel, H., Church, J., Dudley, B., Hall, M. J., Mork, M. E., . . . You, Y. N. J. F. c. (2020). Collaborative Group of the Americas on Inherited Gastrointestinal Cancer Position statement on multigene panel testing for patients with colorectal cancer and/or polyposis. *19*(3), 223-239.

Hegde, M., Ferber, M., Mao, R., Samowitz, W., & Ganguly, A. J. G. i. M. (2014). ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *16*(1), 101-116.

Hu, J., & Ng, P. C. (2013). SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One, 8*(10), e77940. doi:10.1371/journal.pone.0077940

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. J. G. (2010). Hereditary and familial colon cancer. *138*(6), 2044-2058.

KA, W. (2021). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., . . . Cummings, B. B. J. N. a. r. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *45*(D1), D840-D845.

Kim, T.-M., Lee, S.-H., & Chung, Y.-J. J. W. j. o. g. W. (2013). Clinical applications of next-generation sequencing in colorectal cancers. *19*(40), 6784.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., & Massouras, A. (2018). VarSome: the human genomic variant search engine. *Bioinformatics, 35*(11), 1978-1980. doi:10.1093/bioinformatics/bty897 %J Bioinformatics

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., . . . Nik-Zainal, S. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell, 177*(4), 821-836.e816. doi:10.1016/j.cell.2019.03.001

Lagarde, A., Rouleau, E., Ferrari, A., Noguchi, T., Qiu, J., Briaux, A., . . . Olschwang, S. (2010). Germline APC mutation spectrum derived from 863 genomic variations identified through a 15-year medical genetics service to French patients with FAP. *J Med Genet, 47*(10), 721-722. doi:10.1136/jmg.2010.078964

Lamolle, G., & Musto, H. (2018). Genoma Humano. Aspectos estructurales. *Anales de la Facultad de Medicina, 5*, 12-28. doi:10.25184/anfamed2018v5n2a10

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. J. N. a. r. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *42*(D1), D980-D985.

Laurent-Puig, P., Béroud, C., & Soussi, T. J. N. a. r. (1998). APC gene: database of germline and somatic mutations in human tumors and cell lines. *26*(1), 269-270.

Leoz, M. L., Carballal, S., Moreira, L., Ocaña, T., & Balaguer, F. J. T. a. o. c. g. (2015). The genetic basis of familial adenomatous polyposis and its implications for clinical practice and risk management. *8*, 95.

Li, H., & Durbin, R. J. B. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *26*(5), 589-595.

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet, 100*(2), 267-280. doi:10.1016/j.ajhg.2017.01.004

Lorans, M., Dow, E., Macrae, F. A., Winship, I. M., & Buchanan, D. D. J. C. c. c. (2018). Update on hereditary colorectal cancer: improving the clinical utility of multigene panel testing. *17*(2), e293-e305.

Lubbe, S. J., Di Bernardo, M. C., Chandler, I. P., & Houlston, R. S. J. J. o. C. O. (2009). Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *27*(24), 3975-3980.

Mantilla, C., Suárez Mellado, I., Duque Jaramillo, A., & Navas, M. C. J. C. M. (2015). Mecanismos de señalización por β-catenina y su papel en la carcinogénesis. *29*(1), 109-127.

Marabelli, M., Molinaro, V., Khouzam, R. A., Berrino, E., Panero, M., Balsamo, A., . . . Ranzani, G. N. (2016). Colorectal Adenomatous Polyposis: Heterogeneity of Susceptibility Gene Mutations and Phenotypes in a Cohort of Italian Patients. *Genetic Testing and Molecular Biomarkers, 20*(12), 777-785. doi:10.1089/gtmb.2016.0198

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . Daly, M. J. G. r. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *20*(9), 1297-1303.

Nature, C. G. A. N. J. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *487*(7407), 330.

Newton, K., Mallinson, E., Bowen, J., Lalloo, F., Clancy, T., Hill, J., & Evans, D. J. C. g. (2012). Genotype–phenotype correlation in colorectal polyposis. *81*(6), 521-531.

Nieuwenhuis, M. H., Mathus–Vliegen, L. M., Slors, F. J., Griffioen, G., Nagengast, F. M., Schouten, W. R., . . . hepatology. (2007). Genotype-phenotype correlations as a guide in the management of familial adenomatous polyposis. *5*(3), 374-378.

Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., . . . Hedge, P. J. S. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *253*(5020), 665-669.

Olkinuora, A. P., Peltomäki, P. T., Aaltonen, L. A., & Rajamäki, K. J. H. M. G. (2021). From APC to the genetics of hereditary and familial colon cancer syndromes. *30*(R2), R206-R224.

Olschwang, S., Laurent-Puig, P., Groden, J., White, R., & Thomas, G. (1993). Germ-line mutations in the first 14 exons of the adenomatous polyposis coli (APC) gene. *Am J Hum Genet, 52*(2), 273-279.

Pagel, K. A., Antaki, D., Lian, A., Mort, M., Cooper, D. N., Sebat, J., . . . Radivojac, P. (2019). Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS Comput Biol, 15*(6), e1007112. doi:10.1371/journal.pcbi.1007112

Pantelis, D., Hüneburg, R., Adam, R., Holzapfel, S., Gevensleben, H., Nattermann, J., . . . Kalff, J. C. J. I. j. o. c. d. (2016). Prophylactic total gastrectomy in the management of hereditary tumor syndromes. *31*(12), 1825-1833.

Pezzi, A., Roncucci, L., Benatti, P., Sassatelli, R., Varesco, L., Di Gregorio, C., . . . Ponz De Leon, M. (2009). Relative role of APC and MUTYH mutations in the pathogenesis of familial adenomatous polyposis. *Scandinavian Journal of Gastroenterology, 44*(9), 1092-1100. doi:10.1080/00365520903100481

Piñero, T. A., Soukarieh, O., Rolain, M., Alvarez, K., López-Köstner, F., Torrezan, G. T., . . . Pavicic, W. H. (2020). MLH1 intronic variants mapping to + 5 position of splice donor sites lead to deleterious effects on RNA splicing. *Familial Cancer, 19*(4), 323-336. doi:10.1007/s10689-020-00182-5

Quang, D., Chen, Y., & Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics, 31*(5), 761-763. doi:10.1093/bioinformatics/btu703

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res, 47*(D1), D886-d894. doi:10.1093/nar/gky1016

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med, 17*(5), 405-424. doi:10.1038/gim.2015.30

Rossanese, L. B. D. Q., Marson, F. A. D. L., Ribeiro, J. D., Coy, C. S. R., & Bertuzzo, C. S. (2013). APC germline mutations in families with familial adenomatous polyposis. *Oncol Rep, 30*(5), 2081-2088. doi:10.3892/or.2013.2681

Rossi, B. M., Palmero, E. I., López-Kostner, F., Sarroca, C., Vaccaro, C. A., Spirandelli, F., . . . Dominguez-Valentin, M. (2017). A survey of the clinicopathological and molecular characteristics of patients with suspected Lynch syndrome in Latin America. *BMC Cancer, 17*(1), 623. doi:10.1186/s12885-017-3599-4

Russo, M., Lamba, S., Lorenzato, A., Sogari, A., Corti, G., Rospo, G., . . . Bardelli, A. (2018). Reliance upon ancestral mutations is maintained in colorectal cancers that heterogeneously evolve during targeted therapies. *Nature Communications, 9*(1), 2287. doi:10.1038/s41467-018-04506-z

Sack, J., & Rothman, J. (2000). *Colorectal Cancer: Natural History and Management*.

Schmieder, R., & Edwards, R. J. B. (2011). Quality control and preprocessing of metagenomic datasets. *27*(6), 863-864.

Schreuders, E. H., Ruco, A., Rabeneck, L., Schoen, R. E., Sung, J. J., Young, G. P., & Kuipers, E. J. (2015). Colorectal cancer screening: a global overview of existing programmes. *Gut, 64*(10), 1637-1649. doi:10.1136/gutjnl-2014-309086

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. J. N. m. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *7*(8), 575-576.

Senapati, S., Ho, S. B., Sharma, P., Das, S., Chakraborty, S., Kaur, S., . . . Batra, S. K. (2010). Expression of intestinal MUC17 membrane-bound mucin in inflammatory and neoplastic diseases of the colon. *J Clin Pathol, 63*(8), 702-707. doi:10.1136/jcp.2010.078717

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res, 15*(8), 1034-1050. doi:10.1101/gr.3715005

Stekrova, J., Sulova, M., Kebrdlova, V., Zidkova, K., Kotlas, J., Ilencikova, D., . . . Kohoutova, M. (2007). Novel APC mutations in Czech and Slovak FAP families: clinical and genetic aspects. *BMC Medical Genetics, 8*(1), 16. doi:10.1186/1471-2350-8-16

Steward, C. (2001). International Human Genome Sequencing Consortium Nature 409, 860−921.

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

Syngal, S., Brand, R. E., Church, J. M., Giardiello, F. M., Hampel, H. L., & Burt, R. W. J. T. A. j. o. g. (2015). ACG clinical guideline: genetic testing and management of hereditary gastrointestinal cancer syndromes. *110*(2), 223.

Terui, H., Akagi, K., Kawame, H., & Yura, K. (2013). CoDP: predicting the impact of unclassified genetic variants in MSH6 by the combination of different properties of the protein. *J Biomed Sci, 20*(1), 25. doi:10.1186/1423-0127-20-25

Valle, L., de Voer, R. M., Goldberg, Y., Sjursen, W., Försti, A., Ruiz-Ponte, C., . . . Nordling, M. J. M. a. o. m. (2019). Update on genetic predisposition to colorectal cancer and polyposis. *69*, 10-26.

Valle, L. J. C. G., & Hepatology. (2017). Recent discoveries in the genetics of familial colorectal cancer and polyposis. *15*(6), 809-819.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., . . . Thibault, J. J. C. p. i. b. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *43*(1), 11.10. 11-11.10. 33.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Holt, R. A. J. s. (2001). The sequence of the human genome. *291*(5507), 1304-1351.

Vohra, S., & Biggin, P. C. (2013). Mutationmapper: a tool to aid the mapping of protein mutation data. *PLoS One, 8*(8), e71711. doi:10.1371/journal.pone.0071711

Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S., & Girirajan, S. (2017). Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep, 7*(1), 885. doi:10.1038/s41598-017-01005-x

Wingett, S. W., & Andrews, S. J. F. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *7*.

Zhang, J., Wang, X., de Voer, R. M., Hehir-Kwa, J. Y., Kamping, E. J., Weren, R. D. A., . . . van Kessel, A. G. (2017). A molecular inversion probe-based next-generation sequencing panel to detect germline mutations in Chinese early-onset colorectal cancer patients. *Oncotarget, 8*(15), 24533-24547. doi:10.18632/oncotarget.15593