



OPEN ACCESS

EDITED BY

John Frederick Pearson,
University of Otago, New Zealand

REVIEWED BY

Dongbo Yang,
The University of Chicago, United States
Hengrui Liu,
University of Cambridge, United Kingdom

*CORRESPONDENCE

Dora Janeth Fonseca-Mendoza
✉ dora.fonseca@urosario.edu.co

†These authors share first authorship

‡These authors have contributed
equally to this work and share
last authorship

RECEIVED 19 February 2025

ACCEPTED 22 April 2025

PUBLISHED 19 May 2025

CITATION

Rodriguez-Salamanca J, Angulo-Aguado M, Orjuela-Amarillo S, Duque C, Sierra-Díaz DC, Contreras Bravo N, Figueroa C, Restrepo CM, López-Cortés A, Cabrera R, Morel A and Fonseca-Mendoza DJ (2025) Integrating next-generation sequencing and artificial intelligence for the identification and validation of pathogenic variants in colorectal cancer. *Front. Oncol.* 15:1568205. doi: 10.3389/fonc.2025.1568205

COPYRIGHT

© 2025 Rodriguez-Salamanca, Angulo-Aguado, Orjuela-Amarillo, Duque, Sierra-Díaz, Contreras Bravo, Figueroa, Restrepo, López-Cortés, Cabrera, Morel and Fonseca-Mendoza. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Integrating next-generation sequencing and artificial intelligence for the identification and validation of pathogenic variants in colorectal cancer

Juliana Rodriguez-Salamanca^{1†}, Mariana Angulo-Aguado^{1†}, Sarah Orjuela-Amarillo¹, Catalina Duque¹, Diana Carolina Sierra-Díaz¹, Nora Contreras Bravo¹, Carlos Figueroa², Carlos M. Restrepo¹, Andrés López-Cortés³, Rodrigo Cabrera¹, Adrien Morel^{1‡} and Dora Janeth Fonseca-Mendoza^{1*‡}

¹School of Medicine and Health Sciences, Center for Research in Genetics and Genomics (CIGGUR), Institute of Translational Medicine (IMT), Universidad del Rosario, Bogotá, Colombia, ²Coloproctology Department, Hospital Universitario Mayor - Méderi - Universidad del Rosario, Bogotá, Colombia, ³Cancer Research Group (CRG), Faculty of Medicine, Universidad de Las Américas, Quito, Ecuador

Background: Colorectal cancer (CRC) is recognized as a multifactorial disease, where both genetic and environmental factors play critical roles in its development and progression. The identification of pathogenic germline variants has proven to be a valuable tool for early diagnosis, the implementation of surveillance strategies, and the identification of individuals at increased cancer risk. Next-generation sequencing (NGS) has facilitated comprehensive multigene analysis in both hereditary and sporadic cases of CRC.

Patients and methods: In this study, we analyzed 100 unselected Colombian patients with CRC to identify pathogenic (P) and likely pathogenic (LP) germline variants, classified according to the guidelines established by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP). Using the BoostDM artificial intelligence method, we were able to identify oncodriver germline variants with potential implications for disease progression. We assessed the model's accuracy in predicting germline variants by comparing its results with the AlphaMissense pathogenicity prediction model. Additionally, a minigene assay was employed for the functional validation of intronic mutations.

Results: Our findings revealed that 12% of the patients carried pathogenic/likely pathogenic (P/LP) variants according to ACMG/AMP criteria. Using BoostDM, we identified oncodriver variants in 65% of the cases. These results highlight the significance of expanded multigene analysis and the integration of artificial intelligence in detecting germline variants associated with CRC. The average overall AUC values for the comparison between BoostDM and AlphaMissense were 0.788 for the entire BoostDM dataset and 0.803 for the genes within our panel, with individual gene AUC values ranging from 0.606 to 0.983. Functional

validation through the minigene assay revealed the generation of aberrant transcripts, potentially linked to the molecular etiology of the disease.

Conclusion: Our study provided valuable insights into the prevalence and frequency of P/LP germline variants in unselected Colombian CRC patients through NGS. Integrating advanced genomic analysis and artificial intelligence has proven instrumental in enhancing variant detection beyond conventional methods. Our functional validation results provide insights into the potential pathogenicity of intronic variants. These findings underscore the necessity of a multifaceted approach to unravel the complex genetic landscape of CRC.

KEYWORDS

next generation sequencing (NGS), pathogenic germline variants, artificial intelligence, minigene assay, functional validation, colorectal cancer

Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed cancer worldwide and ranks as the second leading cause of cancer-related death (1, 2). Worldwide, approximately 1.9 million new CRC cases are reported, resulting in over 900,000 deaths globally (3, 4). Similarly, it has been projected that by the year 2024, approximately 2 million new cancer cases and 611,720 cancer-related deaths will occur in the United States, with an increase in the incidence of 6 of the 10 most common cancers. Notably, the incidence of CRC is estimated to rise by 1% to 2% annually among young adults (under 55 years of age) (5). About 1 in 23 men and 1 in 24 women are projected to be diagnosed with CRC at some point in their lives (6). While CRC is more common in developed countries, mortality rates have decreased due to early screening strategies such as colonoscopy and fecal occult blood tests (7). The incidence and mortality rates of colorectal cancer (CRC) in Colombia have been documented in several studies. According to Carvalho et al., there has been an increasing trend in CRC incidence in Cali, Colombia, with annual percentage increases of 2.8% for men and 3.3% for women between 1983 and 2012 (8). Regarding mortality, Piñeros et al. reported a rising burden of CRC-related deaths, with an estimated annual percentage change of 2% from 1984 to 2008 (9).

Additionally, according to GLOBOCAN, CRC accounts for approximately 11,163 new cases and 5,640 deaths annually in Colombia (3). In 2022, the country recorded 44,371 new cancer cases, of which 3,851 were attributed to CRC (10) (<https://cuentadealcosto.org/higia/>).

Overall, these data highlight the increasing incidence and mortality rates of CRC in our country, underscoring the urgent need for improved preventive strategies and equitable access to healthcare services (11).

CRC is a complex disease in which genetic alterations and environmental risk factors play crucial roles in its development and

progression (12, 13). Approximately 30% to 35% of patients with CRC report a family history of the disease, which can be attributed to genetic factors, common exposures, or both (14, 15). The familial component of CRC includes both hereditary syndromes and non-syndromic familial clustering, which can increase the risk of CRC even in absence of identifiable genetic mutations (16, 17). Only about 5% to 10% of CRC cases are due to high or moderate penetrance genetic variants associated with hereditary cancer syndromes such as Lynch syndrome (*MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*), familial adenomatous polyposis (*APC*), and MUTYH-associated polyposis (18). Patients carrying Pathogenic Variants (PVs) in these genes are subject to appropriate surveillance strategies, as recommended by the National Comprehensive Cancer Network (NCCN) guidelines for hereditary CRC (19).

Recent reports utilizing next-generation sequencing (NGS) to assess germline variants in multiple cancer-related genes have identified non-canonical pathogenic variants (PVs) potentially associated with CRC in both selected and unselected populations (18, 20, 21). Consequently, increasing the detection rate of PVs through NGS will enhance the identification of molecular alterations in CRC patients. In addition to genes associated with syndromes and genetic disorders conferring a high risk of CRC, other high-penetrance genes, such as *AKT*, *ATM*, *BMPRIA*, *BRAF*, *BRCA1/2*, *CHEK2*, *CTLA4*, *KRAS*, *MYO3A*, *PI3KCA*, *PTEN*, *RAS*, *SMAD2*, *SMAD3*, *TCF7L2*, *TGFBR2*, and *TP53*, as well as those classified as moderate- and lower-penetrance genes, should also be considered. This strategy can potentially improve the identification of risk conferred by PVs predisposing to CRC (22–24). The understanding of genetic mutations has been fundamental to the evolution of cancer therapies, enabling the development of targeted treatments that inhibit key oncogenic proteins and driving the creation of antibody-drug conjugates that combine monoclonal antibodies with cytotoxic agents. In addition, it has contributed to the advancement of immunotherapy and cellular therapies such as CAR-T cells. These personalized strategies have improved

treatment efficacy and reduced side effects, representing a milestone in modern oncology (25).

In Colombia, studies focused on germline variants in CRC have primarily concentrated on syndromic genes such as *APC*, *MLH1*, and *TP53* (26–29). However, the lack of comprehensive genomic data and information on population frequency makes it challenging to construct a genomic profile for CRC predisposition in the Colombian population. The aim of this research was to identify molecular variants potentially related to the disease in Colombian patients with unselected CRC and to determine their frequency in CRC-associated genes. The study analyzed 100 CRC-affected patients, generating a virtual gene panel evaluated through NGS, which included 206 genes of interest. Additionally, intronic variants were validated using functional minigene assays.

Materials and Methods

Sampling and Data collection

This study included 100 unselected patients, defined as individuals diagnosed with cancer who have not been pre-screened or stratified based on specific clinical, molecular, or demographic characteristics. All patients had a histopathological diagnosis of colorectal cancer (CRC) and received care at the Coloproctology Service of Méderi Hospital in Bogotá, Colombia. Patients over 18 years of age with a confirmed biopsy of any type of CRC were eligible, invited to participate in the study, and provided informed consent. Sociodemographic and clinical data were collected through interviews and review of clinical records. The variables examined included sex, age, comorbidities (hypertension, diabetes mellitus, chronic obstructive pulmonary disease, cancer, and others), family history of cancer, lifestyle habits, CRC screening tests, height, weight, age at diagnosis, tumor location, lymphovascular infiltration, tumor stage, and metastasis. The sample size was convenience-based, including patients treated at Méderi Hospital during a defined period from 2020 to 2022. All experimental procedures were approved by the Ethics Committee of Universidad del Rosario and were conducted according to the principles outlined in the Declaration of Helsinki (DVO005 1607-CV1436).

NGS - Whole-Exome sequencing

The patient's DNA was extracted, from peripheral blood, using the Quick-DNA 96 plus kit (Zymo Research). The quality and quantity of DNA were determined using the Quantifluor ONE dsDNA system on a GloMax Discover instrument (Promega). Library preparation was carried out with 250 ng of DNA using the MGIEasy FS DNA Library Prep Kit. Enzymatic DNA fragmentation was performed to obtain fragments ranging from 200 to 400 bp, followed by end repair and PCR amplification. Specific regions of interest were captured using the Exome Capture V5 probe and streptavidin beads. Specific primers were employed for enrichment

in the final PCR reaction. For sequencing, the DNA was circularized, and the library was denatured after split oligo ligation, followed by digestion and purification using specific beads. The circularized DNA was used to generate DNBs (nanoballs) through the rolling circle amplification process (<https://en.mgi-tech.com/>). DNBs were quantified and subsequently sequenced on the DNBSeg400 platform. The obtained reads were mapped to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA) and organized using SAMtools (<https://github.com/samtools/samtools>). Duplicate reads were identified and removed using Picard Software (<https://broadinstitute.github.io/picard/>). Coverage and depth analysis were carried out using BAMBA tool, we considered 50X as an acceptable threshold. >93% of total bases called had a Phred-scaled quality score greater than 30 (>Q30).

A minimum of 7 Gb raw data was obtained and the percentage of reads properly mapped was >99.99% (56,446,362-101,662,468) per sample. Average mapping efficiency was >99%, with sequencing depth on target and coverage of target region >50x and >94%, respectively. The average fraction of target covered with >20x was >72%. Coverage uniformity (10x) was ≥90% and the average fraction of target covered with at least 10x, 20x, 50x and 100x was >90%, >72%, >40% and >19%, respectively. The average number of paired ends reads that mapped to the reference genome was 72,351,272 (99.99%).

The library preparation and sequencing were performed by GencellPharma (Bogota, Colombia). The analyzed panel included 206 genes selected based on evidence from case-control association studies, systematic reviews, GWAS and functional validation studies in CRC, considering their biological relevance, the implication in physio pathological processes, and their roles in oncogenesis (this group was considered as candidate genes). Additionally, genes included in diagnostic panels from CGC genetics, ICM Atrys division, Mayo Clinic Laboratories, GENDIA-genetics and molecular biology, CD Genomics disease panel, Centogene, and Invitae were incorporated. The final panel consisted of 102 genes from established diagnostic panels and 104 candidate genes (Table 1).

Bioinformatic Analysis

Variant call format (VCF) files were analyzed using the software VarSeq® (Golden Helix, v 2.3.0). We incorporated the following database annotations: ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), Ensembl (<https://www.ensembl.org/index.html>), RefSeq (<https://www.ncbi.nlm.nih.gov/refseq/>), dbNSFP Functional predictions, dbSNP, REVEL, OMIM Phenotype Ontology (<https://www.omim.org/>), UniProt Variants (<https://www.uniprot.org/>) and gnomAD v2.1 (<https://gnomad.broadinstitute.org/>).

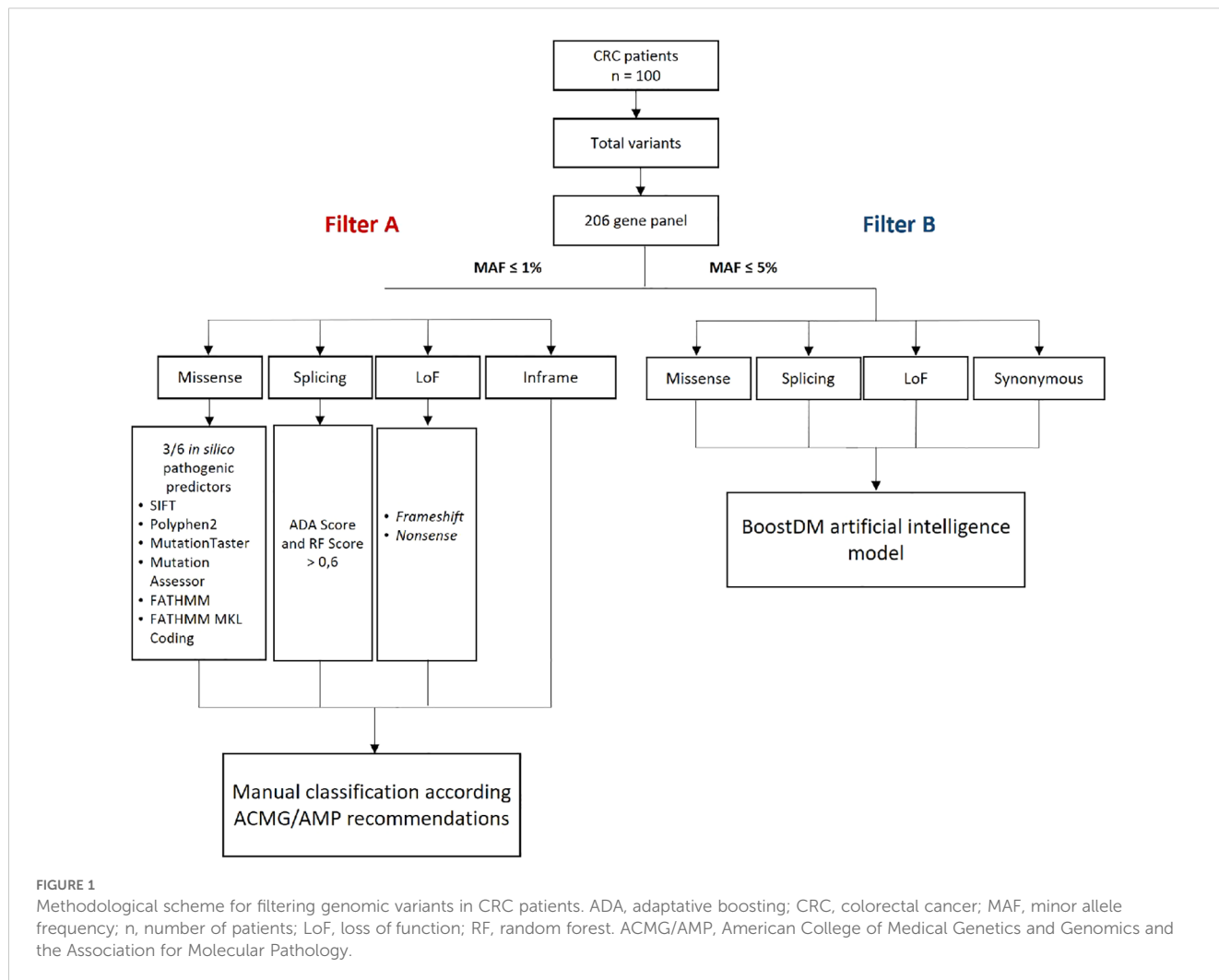
To identify pathogenic/likely pathogenic (P/LP) and oncdriver variants in the 206 CRC-related genes, we applied a bioinformatic analysis using two filtering and classification strategies: (A) manual classification according to the ACMG/AMP recommendations (30, 31), and (B) the BoostDM artificial intelligence system (32) (Figure 1). For strategy A, we prioritized molecular variants with

TABLE 1 Extended multigene panel (n=206 genes).

Molecular diagnosis panel				Gene-candidate panel			
AIP	EGFR	NF2	RUNX1	ACTR1B	EDN1	MAMSTR	SBF2
ALK	EPCAM	NOTCH2	SDHA	ALCAM	EIF3H	MGMT	SEMA4A
APC	EPM2A	NOTCH3	SDHAF2	APE1	ERCC1	MRE11	SF3A3
ATM	FAN1	NRAS	SDHB	ARFGEF2	EXO1	MRE11A	SFMBT1
AXIN2	FH	NTHL1	SDHC	ATF1	FAM109A	MYC	SH2B3
BAP1	FLCN	PALB2	SDHD	ATXN2	FANCC	MYO3A	SHROOM2
BARD1	GALNT12	PDGFRA	SMAD4	B9D2	FANCE	NABP1	SLC15A4
BLM	GATA2	PHOX2B	SMARCA4	BMP2	FEN1	NCAPG	SLC6A18
BMPR1A	GPC3	PIF1	SMARCB1	BMP4	FKBP5	NXN	SMAD6
BRAF	GREM1	PIK3CA	SMARCE1	BMP5	FMN1	OGG1	SMAD7
BRCA1	HOXB13	PMS1	STK11	BORA	FUT2	PIAS1	SMAD9
BRCA2	HRAS	PMS2	SUFU	C11orf53	GLI3	PITX1	SMARCD1
BRIP1	KIT	POLD1	TELO2	CABLES2	GNL1	PLCB1	TBX3
CASR	KRAS	POLE	TERC	CCND2	HHIP	PLGLA	TCF7L2
CCND1	MAX	POT1	TERT	CD44	HNF4A	PNKD	TFEB
CDC73	MEN1	PRKAR1A	TGFBR2	CDKN2B	IL12RB1	POLD3	TLE4
CDH1	MET	PTCH1	TMEM127	CHRD1	KDR	POU5F1B	TMBIM1
CDK4	MITF	PTEN	TP53	COL4A2	KLF5	PRDM1	TMEM59
CDKN1B	MLH1	RAD50	TSC1	COLCA1	LAMA5	PREX1	TNS3
CDKN1C	MLH3	RAD51C	TSC2	COLCA2	LAMC1	PTPN1	TOX2
CDKN2A	MSH2	RAD51D	VHL	COX14	LGR5	PTPN12	TP53BP1
CEBPA	MSH3	RB1	WRN	CRTC3	LIG1	RBBP8	TTC22
CHEK2	MSH6	RBL1	WT1	CTNNB1	LIMA1	RHPN2	VTG1A
CTNNA1	MUTYH	RECQL4	XAF1	DCLRE1C	LIMK2	RPS20	WNT4
DICER1	NBN	RET		DIP2B	LRP1	RTEL1	XRCC2
DIS3L2	NF1	RNF43		DUSP10	MACC1	SATB2-AS1	ZAP70

a minor allele frequency (MAF) of $\leq 1\%$, including loss-of-function (LoF) variants (nonsense, frameshift, and splice site), in-frame, and missense mutations. The potential functional impact of splicing variants was assessed using scores from the adaptive boosting (ADA) and random forest (RF) algorithms (cut-off ≥ 0.6). For missense variants, we prioritized those with positive *in silico* pathogenicity predictions in at least 3 out of 6 predictors integrated into the VarSeq[®] software (SIFT, Polyphen2, Mutation Taster, Mutation Assessor, FATHMM, and FATHMM MKL Coding). Variants classified as functionally relevant were subsequently manually classified following the ACMG/AMP recommendations. The final dataset included variants that met the criteria for P and LP as defined by Hampel et al. (30). For strategy B, all variants with a MAF of less than 5% were analyzed using the BoostDM artificial intelligence model. The final dataset

included variants classified as oncodriver mutations, as described by Muiños et al. (32). Furthermore, we sought to evaluate its potential in predicting the significance of germline variants in colorectal cancer genes. For benchmarking against a high-performing germline variant classifier, we utilized AlphaMissense (33), which demonstrates leading-edge performance in predicting the pathogenicity of missense variants, by integrating structural context and evolutionary conservation. This tool is validated by extensive genetic and experimental benchmarks and can classify a vast majority of missense variants with a high precision score on recognized databases like ClinVar, without being explicitly trained on such data (33). A Python script (S1 Appendix) was developed to assign AlphaMissense classifications (Obtained from https://storage.googleapis.com/dm_alphamissense/AlphaMissense_hg38.tsv.gz) to all variants scored for BoostDM available in the



public repository (<https://www.intogen.org/boostdm/downloads>). Utilizing this script, precision-recall curves were generated, and AUC scores were calculated to evaluate the performance of BoostDM's predictions of variants classified by AlphaMissense. This comparison utilized saturation mutagenesis data, encompassing 81 genes, including 22 that are part of our colorectal cancer gene panel. Genes without AlphaMissense predictions were excluded.

Population Genetic Analysis

For each variant identified through NGS analysis, we assessed the allelic frequency, genotypic frequency, and the Hardy Weinberg equilibrium (HWE) using the SNP-Stats software (<https://www.snpstats.net/start.htm>). Deviation from HWE was determined using χ^2 goodness-of-fit test with 1° of freedom. The allele frequencies of the study were compared to the global and Latin-American populations, obtained from the gnomAD database (<https://gnomad.broadinstitute.org>), and statistical significance was assessed using chi-square (χ^2) test, with significance being established at *p-value* <0.05.

Plasmid constructs

We used patient genomic DNA to amplify the region encompassing the exon nearest to the mutation, along with 300 bp of flanking intronic sequences (upstream and downstream of the exon). Primers were designed according to Putscher et al., 2021 and verified by Primer-BLAST (34). The PCR was performed using the New England Biolabs Q5 master mix (Q5[®] High-Fidelity 2X Master Mix cat: M049), according to the manufacturer's protocol. The PCR products were recombined with the pSpliceExpress vector using Gateway[™] BP Clonase[™] II Enzyme mix (Invitrogen), following the manufacturer's instructions. The sequence of the constructed vector was confirmed through Sanger sequencing.

Cell culture transfection

HCT-116 and HEK-293 cell lines, were cultured in Dulbecco's Modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum and 5% penicillin/streptomycin at 37°C in a 5% CO₂ environment. Both cell lines were seeded at a density of 60,000 cells per well into a 24 well-plate, with triplicates for each experimental

condition. 1 µg of the vector was transfected into the cells and incubated for 48 hours using FuGENE[®] 6 Transfection Reagent (Promega), according to the manufacturer's protocol.

Total RNA extraction and RT-PCR analysis

Total RNA was extracted using the TRIzol[®] Reagent protocol. 500 ng of RNA was used to synthesize cDNA with SuperScript III cDNA first strand (Invitrogen), previous treatment with DNase I Amplification Grade (Sigma-Aldrich). Using primers targeting exons 2 and 3 of rat insulin (whose sequences are located into the pSpliceExpress vector and surrounds the sequences cloned through recombination with Gateway[™] BP Clonase[™] II Enzyme mix). PCR was performed under the following conditions: 95°C for 1 minute, 60°C for 40 seconds, and then 3 minutes at 72°C for 30 cycles. RT-PCR products were visualized on 1.5% agarose gels. The obtained bands were analyzed using densitometry with Image Lab software (BIO-RAD). For each band, obtained in the Wild-Type (WT) and mutant (MUT) RT-PCR assays, intensity measurement was performed, allowing subsequent establishment of differences in the transcripts obtained.

Additionally, the RT-PCR products for both WT and MUT were cloned in the storage vector, pCR4-TOPO, using the TOPO[®] TA Cloning Kit for Sequencing (Invitrogen) following the manufacturer's instructions. After transformation in the One Shot Top 10 chemically competent cells (<https://www.thermofisher.com/>), 10 colonies were randomly selected for Sanger sequencing following plasmid DNA extraction.

Statistical Analysis

We determined the association between non-genetic factors and being a carrier of a Pathogenic, Likely Pathogenic and oncogene variant using a bivariate analysis. Statistical significance was determined at a threshold of *p-value* <0.05.

To compare WT vs MUT for experimental assays, an independent samples t-test (Student's t-test) was applied for comparisons between the two experimental groups and their replicates, using SPSS V.29 and GraphPad Prism 10 software. A statistically significant difference was considered if *p-value* < 0.05.

Results

Demographic and clinicopathological characteristics

This study included 100 unselected patients with a histopathological diagnosis of CRC, and the characteristics of the population are summarized in Table 2. Women constituted the majority, representing 55% (n=55) of the patients. A family history of cancer was reported in 63% of individuals, with 11% of cases involving CRC. Regarding tumor location, there was a higher

prevalence of right-sided colon cancer, followed by sigmoid and rectal cancer (44% and 29%, respectively). The sample analyzed exhibited heterogeneous histology and tumor staging, with a greater proportion of moderately differentiated adenocarcinoma and stage II tumors (30% and 36%, respectively). Most patients did not present metastasis at the time of inclusion in the study (76%). The mean age at CRC diagnosis was 65.5 years, with 12% of patients diagnosed before the age of 50. An association analysis of clinical variables grouped by age at diagnosis revealed statistically significant differences in family history of CRC (*p*<0.05) (Table 2).

Germline mutation landscape of Colombian CRC patients

Molecular variants identified following ACMG/AMP recommendations

Following the bioinformatics pipeline described in the methodology (Figure 1) and applying Filter A, which identified pathogenic (P) and likely pathogenic (LP) variants according to ACMG/AMP recommendations, our study detected a total of 248 variants (Figure 2A). Of these, 5 (2%) were classified as pathogenic (P), 8 (3.2%) as likely pathogenic (LP), 153 (61.7%) as variants of uncertain significance (VUS), and 82 (33.1%) as likely benign (LB) (Table 3 and Figure 2A). The detection rate for P/LP variants was 12% (12 patients). A total of 11% of patients carried a single P/LP mutation in a heterozygous state, while one patient was identified with two variants, *MLH1* c.1039delA and *IL12RB1* c.1791 + 2T>G, classified as LP and P, respectively.

Among the identified P/LP variants in the studied population, 38.5% (5/13) were in genes not typically included in routinely used diagnostic panels in clinical practice. These variants corresponded to *ERCC1*: c.702 + 1G>A, *EXO1*: c.1465delA, *IL12RB1*: c.1791 + 2T>G, *OGG1*: c.137G>A, and *SMAD9*: c.781 + 2T>A (Table 3). Of the P/LP variants, 61.5% (8/13) were found in genes commonly selected in molecular diagnostic panels for cancer. Four of these were in high-penetrance genes (*BMPRI1*: c.176T>A, *MLH1*: c.1039delA, *MSH6*: c.3516_3517delAG, and *PTCH1*: c.3241G>A), while one was in the moderate-penetrance gene *BARD1*: c.2229dupT. For eight of the analyzed genes where variants of interest were found, the penetrance has not been clearly defined (*FLCN*, *NOTCH3*, *NTHL1*, *ERCC1*, *EXO1*, *IL12RB1*, *OGG1*, and *SMAD9*) (Table 3 and Figure 2B).

Regarding variant types, we observed six loss-of-function (LoF) variants, four missense variants, and three splice-site variants. Notably, two LP variants were novel and had not been previously reported in public databases such as gnomAD (*MLH1*: c.1039delA, p. Thr347Leufs*20, and *EXO1*: c.1465delA, p. Arg489Glyfs*32) (Table 3). All variants classified as P/LP were confirmed through Sanger sequencing.

Molecular Variants identified for BoostDM model

The BoostDM model, developed by Muiños et al. in 2021, is a machine learning-based methodology designed for evaluating the oncogenic potential of mutations. We tested whether this model can

TABLE 2 Demographic and clinicopathological characterization.

Characteristics	Total	Age at diagnosis		P-value	P-value Yates correction
		< 50	≥ 50		
No. of patients	100	12	88		
Gender				0.035*	0.073
Male	45	2	43		
Female	55	10	45		
Age at CRC diagnosis					
Mean (SD)	65,5 (12,6)	41,25 (8,4)	68,8 (8,9)		
Family history of cancer				0.632	0.889
No	32	3	29		
Yes	63	8	55		
Unknown	5	1	4		
Family history of CRC				0.004*	0.017*
No	81	6	75		
Yes	11	4	7		
Unknown	8	2	6		
Primary tumor site					
Right sided colon cancer	44	2	42	0.042*	0.085
Sigmoid	29	6	23	0.087	0.171
Rectum	17	4	13	0.108	0.232
Entire colon	1	0	1	0.711	1
Multiple primary colorectal cancer	9	0	9	0.246	0.533
Tumor histological stage					
Well differentiated adenocarcinoma	7	0	7	0.311	0.682
Moderately differentiated adenocarcinoma	30	4	26	0.788	1
Poorly differentiated adenocarcinoma	1	0	1	0.711	1
Mucinous adenocarcinoma	3	1	2	0.248	0.801
Infiltrating adenocarcinoma	3	1	2	0.248	0.801
Multiple histological stage	56	6	50	0.655	0.892
Cancer Stage				0.484	0.484
0	2	1	1		
I	14	1	13		
II	36	4	32		
III	32	3	29		
IV	15	2	13		
Unknown	1	1	0		
Lymphovascular infiltration				0.188	0.319
Yes	53	8	45		

(Continued)

TABLE 2 Continued

Characteristics	Total	Age at diagnosis		P-value	P-value Yates correction
		< 50	≥ 50		
No	45	3	42		
Unknown	2	1	1		
Metastasis				0.719	1
Yes	22	2	20		
No	76	9	67		
Unknown	2	1	1		
Pathogenic or Likely Pathogenic Variant (ACMG/AMP)				0.140	0.315
Yes	12	3	9		
No	88	9	79		
Oncodriver Variant (BoostDM)				0.156	0.273
Yes	65	10	55		
No	35	2	33		

*Statistical significance.

accurately predict germline variants by evaluating its agreement with the AlphaMissense pathogenicity prediction model, which has shown state-of-the-art performance for predicting germline disease variants (33). We observed average overall AUC values for the comparison between BoostDM and AlphaMissense of 0.788 for the entire BoostDM dataset and 0.803 for the genes within our panel, with AUC values for individual genes ranging from 0.606 to 0.983 (Figure 3) using publicly available saturation mutagenesis data (S1 Appendix). The reliability of BoostDM in our study underscores its potential applicability beyond somatic mutations, suggesting it may serve as an informative tool for germline variant analysis in CRC.

Applying this model facilitated the identification of 68 oncodriver variants across 43 genes (Supplementary Table 1), which were found in 65 of the analyzed patients, resulting in a detection rate of 65%. Notably, 27% of individuals were found to carry more than one mutation (Figure 2C). Of the variants identified through this analysis, 72% were associated with genes included in cancer diagnostic panels, while the remaining 28% were linked to candidate genes. Additionally, 18.3% (8/43) of the genes with oncodriver variants are related to syndromic CRC, and 20.5% (14/68) of the study's variants were identified in these genes. Penetrance has been clearly defined for 18 of the genes in which oncodriver mutations were identified, with 14 being high-penetrance and 4 moderate-penetrance genes (Supplementary Table 1, Figure 2D).

Regarding the types of variants, 82.3% (56/68) were missense, 13.2% (9/68) were loss-of-function (LoF), 1.5% (1/68) were splicing mutations, and 3% (2/68) were synonymous variants. The genes with the most frequently observed variants were *MSH6* (n=5), and *MLH1*, *WRN*, *FANCC*, *KDR*, and *TCF7L2* (each with 3 variants, respectively) (Table 3). Finally, five of the oncodriver variants were novel and had not been previously reported in public databases

(*APC*: c.3663_3665delTTC, *MLH1*: c.1039delA, *RAD51C*: c.659T>C, *CTNNB1*: c.991T>C, and *KDR*: c.2555A>T). All these variants were confirmed through Sanger sequencing.

The comparison between allelic frequencies of P/LP variants identified by Filter A and oncodriver variants identified by Filter B revealed statistically significant differences ($p < 0.05$) when compared to those described for global and Latin American populations in the gnomAD database (Supplementary Table 2). For P/LP variants classified by Filter A, 46% had higher allelic frequencies compared to the global population, while only 23% differed from the Latin American population. This finding highlights distinct genomic profiles specific to our population, with higher frequencies of P/LP variants associated with CRC.

Similarly, 42.6% of the oncodriver variants identified by Filter B showed significant differences in global allelic frequencies, while 70.6% of the allelic frequencies in our patients aligned with those of the Latin American population (Supplementary Table 2).

In vitro assay in splice-site variants

Functional validation was performed for three splicing mutations identified in the genes *ERCCI*: c.702 + 1G>A, *SMAD9*: c.781 + 2T>A, and *IL12RB1*: c.1791 + 2T>G. These variants belong to genes that are not typically included in cancer diagnostic panels, representing novel genetic factors potentially related to CRC.

Confirmation of the generation of aberrant transcripts, validated through post-transfection RT-PCR in HCT-116 and HEK-293 cell models, revealed that the *SMAD9* c.781 + 2T>A mutation induced exon skipping of exon 3, resulting in the loss of 111 bp. The mutant form failed to generate the normal transcript observed in the WT version, leading to complete exon skipping and

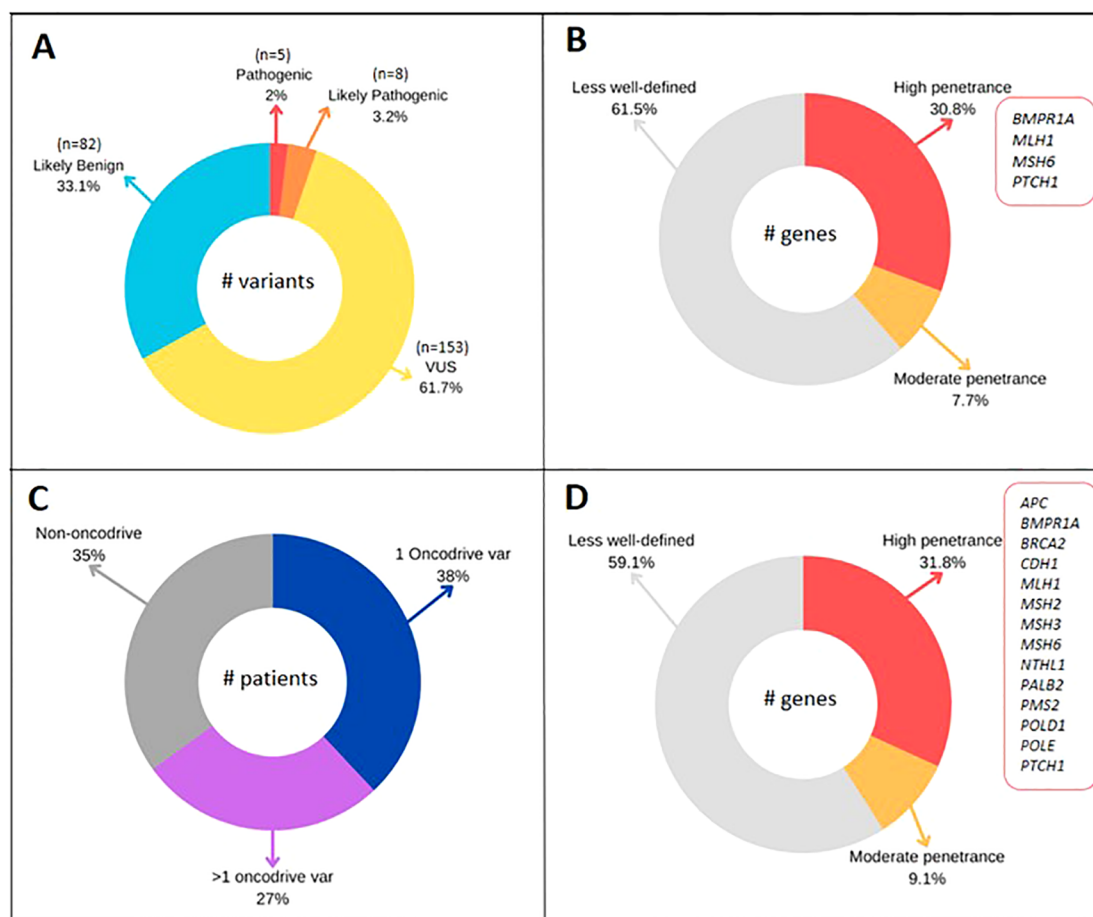


FIGURE 2

(A) Variants classified by ACMG/AMP classification in 100 CRC patients. (B) Penetrance genes of P/LP variants (Filter A). (C) Patients with 1 or >1 oncdriver variant classified with BoostDM model. (D) Penetrance genes of oncdriver variants (Filter B).

the loss of 37 amino acids in the protein ($p < 0.05$). This result was consistent across all three replicates and in both cell lines analyzed. Exon skipping was further confirmed through Sanger sequencing of plasmid DNA obtained from the cloning process (Figure 4A).

Functional analysis of the *ERCC1* c.702 + 1G>A variant revealed the exclusive generation of an aberrant transcript with a loss of 100 bp, corresponding to exon 6 skipping, which result in a frameshift mutation (p. Ser201Argfs*1). The wild-type (WT) allele produced two transcripts: one canonical and one resulting from exon skipping. A statistically significant difference in transcript profiles was observed between the WT and mutant (MUT) alleles in the HCT-116 cell line ($p < 0.05$). This difference was not observed in the HEK-293 Cell line (Figure 4B).

Lastly, the *IL12RB1* c.1791 + 2T>G mutation was associated with the generation of two aberrant transcripts. The first involved partial retention of intron 14 of the *IL12RB1* gene, introducing 98 bp, while the second resulted in exon skipping, eliminating exon 15 (76 bp). The predicted consequences on the protein indicate that these transcripts induce frameshift mutations, p.Thr598Glyfs21 and p.A573Lfs21, respectively (Figure 4C). Although these aberrant transcripts were also observed in the WT version, their expression was significantly reduced compared to the MUT version ($p < 0.05$).

These findings were consistent across both cell models analyzed and in all three replicates. The transcripts were verified through Sanger sequencing of the plasmid DNA obtained from cloning using the TOPO® TA Cloning® Kit for Sequencing (Invitrogen) (Figure 4C).

Discussion

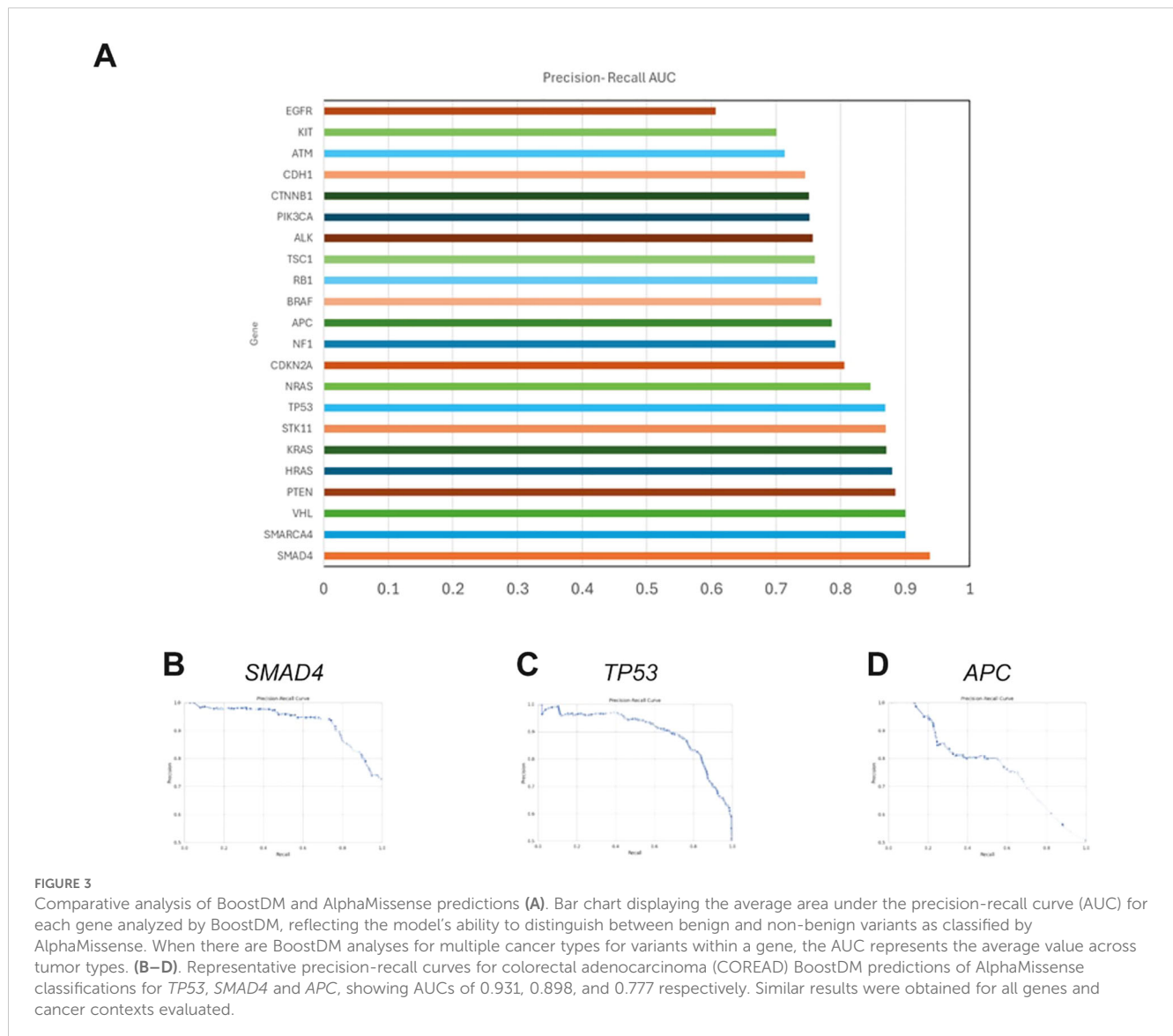
To the best of our knowledge, this study presents, for the first time, a description of the prevalence and spectrum of germline variants in a cohort of Colombian patients with unselected CRC. Our approach utilized next-generation sequencing (NGS) to identify molecular variants within a comprehensive panel of 206 cancer-associated genes (Table 1). This panel includes genes commonly used in diagnostic panels for clinical practice, as well as potential candidate genes identified through an exhaustive literature review. The bioinformatic pipeline implemented enabled the classification of molecular variants according to the ACMG/AMP criteria and the BoostDM artificial intelligence model (Figure 1).

Although BoostDM was originally developed for somatic variants, we selected this model due to its comprehensive *in silico*

TABLE 3 Described pathogenic and likely pathogenic germline variants in CRC patients – Filter A (ACMG/AMP).

Penetrance	Gene panel	Gene	Transcript	Variant	Protein	rs ID	Zygosity	Type	ACMG/AMP classification	Criteria	Count
High	Diagnosis	<i>BMPRIA</i>	NM_004329.3	c.176T>A	p.Leu59Ter	rs1564714834	Het	Missense	Pathogenic	PVS1 + PM2 + PM4 + BP1	1
High	Diagnosis	<i>MLH1</i>	NM_000249.4	c.1039delA	p.Thr347Leufs*20	Not reported	Het	LOF	Likely Pathogenic	PVS1 + PM2	1
High	Diagnosis	<i>MSH6</i>	NM_000179.3	c.3516_3517delAG	p.Arg1172Serfs*4	rs398123232	Het	LOF	Pathogenic	PVS1 + PM2 + PM4	1
High	Diagnosis	<i>PTCH1</i>	NM_000264.5	c.3241G>A	p.Val1081Met	rs587778629	Het	Missense	Likely Pathogenic	PS3 + PM2 + PP3	1
Less well-defined	Diagnosis	<i>FLCN</i>	NM_144997.7	c.1285delC	p.His429Thrfs*39	rs80338682	Het	LOF	Likely Pathogenic	PVS1 + PM2	1
Less well-defined	Diagnosis	<i>NOTCH3</i>	NM_000435.3	c.1345C>T	p.Arg449Cys	rs762734007	Het	Missense	Likely Pathogenic	PM1 + PM2 + PP2 + PP3	1
Less well-defined	Diagnosis	<i>NTHL1</i>	NM_002528.7	c.244C>T	p.Gln82Ter	rs150766139	Het	LOF	Pathogenic	PVS1 + PM2 + PM4	1
Moderate	Diagnosis	<i>BARD1</i>	NM_000465.4	c.2229dupT	p.Asn744Ter	rs1259296823	Het	LOF	Pathogenic	PVS1 + PM1 + PM2 + PM4	1
Less well-defined	Candidate	<i>ERCC1</i>	NM_202001.3	c.702+1G>A	–	rs747911302	Het	Splice	Likely Pathogenic	PVS1 + PM2	1
Less well-defined	Candidate	<i>EXO1</i>	NM_130398.4	c.1465delA	p.Arg489Glyfs*32	Not reported	Het	LOF	Likely Pathogenic	PVS1 + PM2	1
Less well-defined	Candidate	<i>IL12RB1</i>	NM_005535.3	c.1791+2T>G	–	rs554063682	Het	Splice	Pathogenic	PVS1 + PM2 + PP3	1
Less well-defined	Candidate	<i>OGG1</i>	NM_002542	c.137G>A	p.Arg46Gln	rs104893751	Het	Missense	Likely Pathogenic	PS3 + PM2	1
Less well-defined	Candidate	<i>SMAD9</i>	NM_001127217.3	c.781+2T>A	–	rs770716081	Het	Splice	Likely Pathogenic	PVS1 + PM2 + BP4	1

Het, Heterozygous; LOF, Loss of function.



saturation mutagenesis approach, which enables a systematic assessment of the oncogenic potential of all possible variants across cancer-related genes (32). Given the lack of germline-specific tools for oncodriver classification, we aimed to explore the applicability of BoostDM to germline variants. To ensure its relevance in this context, we compared its performance with AlphaMissense, which demonstrated strong concordance and suggested potential utility beyond the somatic mutation context (33). Notably, this strategy has previously been explored and validated in several populations worldwide exposed to driver mutations in lung and gastric cancers (35, 36).

Our cohort consisted of unselected patients, with only 12% being diagnosed with CRC before the age of 50. This contrasts with the majority of studies on germline mutations, which typically focus on individuals with early-onset cancer or those suspected of having hereditary predisposition syndromes related to CRC (18, 23, 37–39). In such cases, CRC susceptibility is often primarily attributed to

germline variants, particularly in hereditary or syndromic CRC, where inheritance typically follows an autosomal dominant pattern.

In addition to the age of CRC onset, our study revealed a higher proportion of affected women (55%) compared to men, which contrasts with previously reported findings. Most studies show a higher incidence of CRC in men (20, 37, 40, 41). This discrepancy may be related to demographic trends in Colombia, where the proportion of women is slightly higher than men (20.9% vs 19.9%) in the age ranges most represented in our population. It has been established that the only age group in which men outnumber women is the youngest (0 to 14 years) (<https://www.statista.com/statistics/789705/population-total-age-gender-colombia/>).

Understanding the genetic factors related to CRC is crucial for supporting the implementation of appropriate surveillance strategies, as recommended by international guidelines, to reduce CRC incidence and mortality in carriers of pathogenic/likely pathogenic (P/LP) germline variants (18). In this context, our

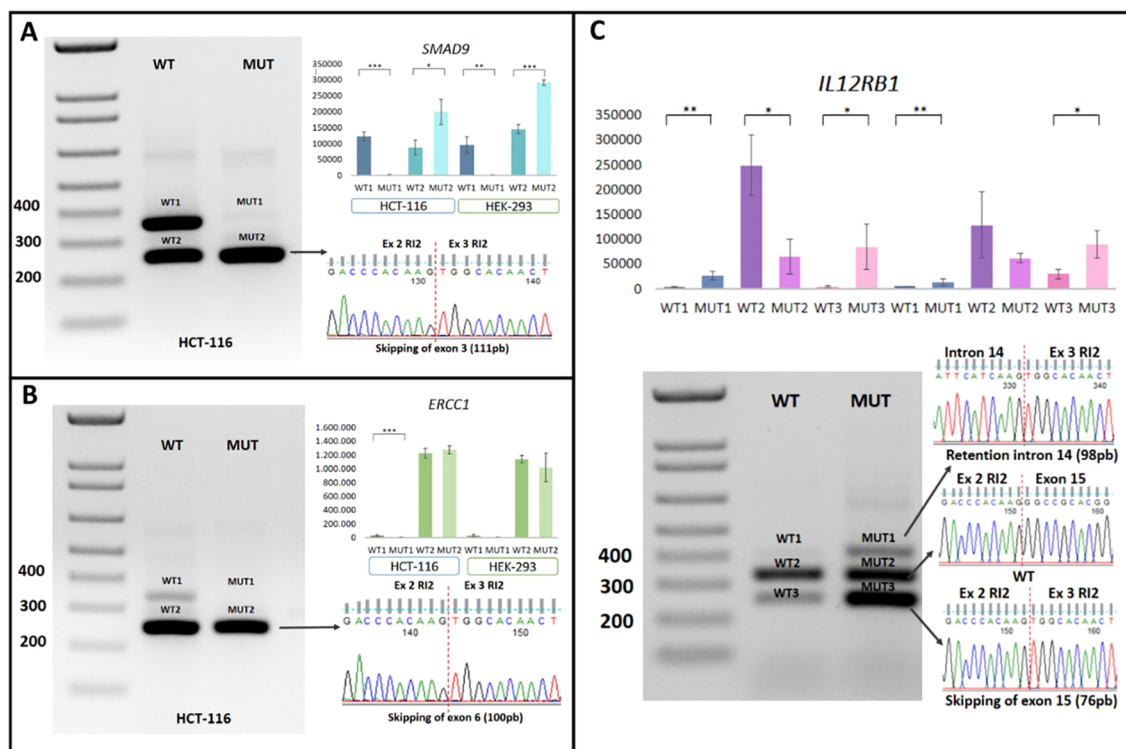


FIGURE 4

Minigene assay on intronic P/LP variants in canonical splicing sites. (A) Functional analysis of *SMAD9*:c.781 + 2T>A in transfected HCT-116 and HEK-293 cells (WT and MUT), with exon skipping confirmed by Sanger sequencing. (B) Functional analysis of *ERCC1*:c.702 + 1G>A transfected HCT-116 and HEK-293 cells (WT and MUT), with exon skipping confirmed by Sanger sequencing. (C) Functional analysis of *IL12RB1*:c.1791 + 2T>G in transfected HCT-116 and HEK-293 cells (WT and MUT), showing exon skipping and an intron retention confirmed by Sanger sequencing. Ex, exon; MUT, mutant; RI2, Rat Insulin 2; WT, Wild-Type *Statistical significance. * $p < 0.05$; ** $p < 0.01$ and *** $p < 0.001$.

study proposes the use of NGS for the analysis of multiple genes, enabling the identification of both canonical and non-canonical pathogenic variants associated with the disease. Non-canonical PVs in cancer predisposition genes, seemingly unrelated to CRC, have proven to be a significant source for identifying the molecular causes in unselected cases (21). Moreover, analyzing unselected populations, as we have done, has been recommended to avoid underestimating the true prevalence of hereditary factors in the global CRC population (18).

Regarding the classification of molecular variants, we applied two approaches. First, Filter A, which followed the ACMG/AMP recommendations (30, 42). This approach is used in both research and routine clinical molecular diagnostics, and it enabled the detection of 13 P/LP variants in 12% of our patients (Table 3).

Our findings demonstrated a higher detection rate compared to other studies analyzing unselected populations, such as those described by Yurgelun et al. (21), which reported a rate of 10% (21). The smaller number of genes analyzed in that study ($n=29$) may explain this difference and supports the advantage of using expanded panels, which potentially enhance diagnostic performance (18, 43). Detection rates of PVs in cases with suspected familial CRC or in patients diagnosed before the age of 50 are substantially higher, with rates ranging from 15.5% to 26.8% (18, 20, 40, 41). In our study, 61.5% of the P/LP variants were associated with genes included in diagnostic panels, while the

remaining 38.5% originated from candidate genes identified through literature review. This significant proportion supports the hypothesis that the analysis of genes not currently included in routine diagnostic panels on various sequencing platforms can increase the detection of P/LP germline variants associated with CRC development.

In the analyzed population, four high-penetrance mutations were identified in the *MLH1*, *MSH6*, *BMPRIA*, and *PTCH1* genes. This finding is significant, as the detection of mutations in some of these genes, particularly *MLH1* and *MSH6*, allows for the implementation of screening, treatment, and follow-up strategies for these high-risk patients, following the recommendations outlined in the NCCN guidelines (National Comprehensive Cancer Network) (<https://www.nccn.org/patients/guidelines/content/PDF/colorectal-screening-patient.pdf>). Germline variants in the mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) are associated with Lynch Syndrome, one of the most common genetic predisposition syndromes, accounting for 2-4% of all CRC cases. Studies have shown that in Latin American and Caribbean countries, there is a higher prevalence of CRC in patients under 50 years of age. Therefore, screening programs and the identification of high-risk individuals carrying pathogenic germline mutations are expected to more effectively reduce CRC incidence (44).

Previous studies have demonstrated that most P/LP variants found in CRC patients are located in MMR genes, which are

associated with DNA repair processes. For instance, Gong et al. (20) reported that out of the 19.1% of P/LP variants identified in their study, 11.3% were from MMR genes. Similarly, Yurgelun et al. (21) demonstrated a high rate of mutations in these genes, and other studies, such as Zhang et al. (39), confirmed these findings, with nearly 50% of genetic variants located in MMR genes (20, 21, 39).

In contrast to these previous findings, our population exhibited only two variants in the MMR genes (*MLH1*: c.1039del and *MSH6*: c.3516_3517delAG), accounting for 15.4% of the total identified. This result indicates a lower frequency of MMR mutations in our study population, which may be attributed to the sample selection criteria. Studies with higher mutation rates in MMR genes typically include patients with a family history of CRC, a diagnosis under the age of 50, and the presence of polyps. In contrast, our study analyzed unselected cases, allowing for the identification of variants in genes not associated with hereditary syndromes.

Additionally, mutations in other genes related to syndromic CRC were identified in our study. One such mutation is the *BMPRIA* mutation (c.176T>A, p.Leu59Ter), which generates a premature termination codon (PTC), leading to the elimination of 473 amino acids. Variants in the *BMPRIA* gene have been associated with juvenile polyposis syndrome. This gene is linked to the AKT signaling pathway and functions as a type I receptor in the TGFβ superfamily (45–48). Beyond its role in regulating epithelial functions in the colon, *BMPRIA* influences critical cellular processes such as growth, differentiation, and apoptosis (46, 48–51). Furthermore, this mutation has been previously reported in a patient with adenomatous polyposis (52). The identified mutation affects the majority of the protein, including its functional activin type I and II domains, as well as the Ser/Thr protein kinase domain. As a result, the protein loses its functionality, its ability to interact with other proteins, and its role in the aforementioned signaling pathways.

Finally, one patient presented a heterozygous variant in a gene known for its high penetrance and its correlation with metastatic potential in CRC (53). The identified variant is a missense change in the *PTCH1* gene (c.3241G>A, p.Val1081Met). *PTCH1* plays a crucial role in the Hedgehog signaling pathway, which is involved in tumorigenesis, regulation of proliferation, angiogenesis, and stem cell renewal. These processes have been linked to oncogenic mechanisms related to CRC (54). The discovery of this variant is highly relevant, as previous studies have shown that *PTCH1* could serve as a potential biomarker for distinguishing between CRC cases with high and low metastatic risk, with an inverse correlation to *PTCH1* protein expression levels (53). Additionally, this variant has been reported in patients with hereditary cancer predisposition syndromes, with functional evidence indicating protein loss of function (55), further supporting the pathogenicity criteria for this mutation.

For other P/LP variants identified according to ACMG/AMP recommendations, the exact definition of penetrance is not available. However, due to their pathogenicity and clear involvement in molecular signaling pathways related to the CRC pathophysiology, they are relevant in its molecular etiopathology.

46% of the variants corresponded to loss-of-function (LoF) variants and altered the function of *FLCN*, *NOTCH3*, *NTHL1*, and *BARD1* genes.

BARD1 encodes a protein that interacts with the N-terminal region of the *BRCA1* gene. The *BARD1/BRCA1* complex plays a critical role in DNA repair, recombination, and cell cycle control (56–60). The *BARD1* c.2229dupT variant generates a premature termination codon (PTC), affecting the BCRT 2 domain, which is involved in protein-protein interactions and has been linked to DNA repair, recombination, and cell cycle control—mechanisms critical in the development of malignant neoplasms (61, 62).

On the other hand, the *FLCN* gene is associated with Birt-Hogg-Dubé syndrome and increases the risk of CRC (63). Pathogenic variants in this gene have recently been identified in patients with early-onset CRC (64–68). The *FLCN* c.1285delC variant induces a frameshift (p.His429Thrfs*39), resulting in the loss of the terminal DENN domain, which is important for the AMPK and mTOR signaling pathways (69).

NOTCH3 is a cell surface receptor that plays a crucial role in signaling pathways that regulate epithelial cell proliferation, polarity/adhesion, and apoptosis (70–72). One patient was a carrier of the *NOTCH3* c.1345C>T, p.Arg449Cys variant, classified as likely pathogenic. Previous research has established a connection between variants in this gene and susceptibility to CRC, particularly in men (73).

The *NTHL1* gene is associated with familial adenomatous polyposis and hereditary cancer predisposition syndrome, with an autosomal recessive inheritance pattern. It is involved in base excision repair (BER), the primary repair pathway for oxidative DNA damage (22, 74–76). The *NTHL1* c.244C>T, p.Gln82Ter variant induces the loss of the HhH domain, which is crucial for DNA binding and, consequently, for the proper functioning of the BER pathway (77).

Our study identified that approximately 40% of the P/LP variants classified through the ACMG/AMP criteria (Filter A described in the methodology) were located in genes not typically analyzed in the context of clinical-molecular diagnosis. Our finding supports the need to implement molecular analyses, using NGS, that allow the simultaneous analysis of hundreds of genes potentially related to CRC. 60% of these variants were located in genomic sites involved in the splicing process and correspond to the *ERCC1* (c.702 + 1G>A), *IL12RB1* (c.1791 + 2T>G), and *SMAD9* (c.781 + 2T>A) mutations. These genes participate in the signaling pathways of cellular damage repair and processes related to carcinogenesis (78–80). Mutations in these genes could, therefore, alter these signaling pathways and increase susceptibility to CRC (81, 82). However, to determine the effect of these intronic variants on carcinogenesis, further studies must be performed at several levels such as transcriptomic, post-transcriptional modifications, and proteomics in the tumoral context. To the best of our knowledge, these studies have not yet been conducted.

Our study also identified heterozygous P/LP mutations in the *EXO1* and *OGG1* genes, both of which are involved in repair and recombination processes (83–88). When these genes are affected by

mutations, replication and post-replication processes, including checkpoint activation, are disrupted, potentially leading to genomic instability and cancer development (84, 89, 90).

Collectively, our findings indicate that germline P/LP variants predominantly cluster in genes associated with DNA repair and regulatory pathways, such as *BARD1*, *BMPRIA*, *ERCC1*, *EXO1*, *MLH1*, *MSH6*, *NTHL1*, and *OGG1* (comprising 60% of the identified variants). These pathways are critical for maintaining genomic stability (91). Consequently, germline variants in genes linked to these signaling pathways can have detrimental effects, increasing susceptibility to CRC.

In addition to searching for pathogenic germline variants using strategies commonly employed in routine molecular diagnostics for identifying individuals at high risk of developing CRC, we utilized an artificial intelligence method. BoostDM is a machine learning-based methodology designed for *in silico* mutagenesis of genes associated with cancer development. This innovative approach systematically evaluates all possible changes within a gene or protein to identify cancer-causing factors (32, 92). Variants with detrimental effects are categorized as oncdrivers. We identified a total of 68 oncdriver variants in 65% of the patients, with 27% presenting more than one oncdriver variant. Of these, 72% were found in genes commonly used in clinical diagnostics, while the remaining 28% were in candidate genes.

The application of BoostDM to germline variant analysis addresses a critical gap in current interpretation strategies. Most existing germline-focused tools, including AlphaMissense, are designed to assess pathogenicity but do not distinguish between driver and passenger mutations—a distinction that can provide deeper insight into cancer predisposition mechanisms (33). BoostDM offers a complementary approach by prioritizing variants with high oncogenic potential, regardless of their previous clinical annotation (32). This feature is particularly useful in unselected populations, where novel or non-canonical germline variants are frequently detected. While BoostDM is not intended to replace clinically validated models, its application can improve variant prioritization for downstream functional assays and support comprehensive molecular profiling in hereditary cancer research (Figure 3).

A consensus was established to support the efficacy of the BoostDM model in our analysis, we analyzed its predictions with those from AlphaMissense, a tool recognized for its accuracy in classifying pathogenic germline variants (Figure 3). This comparison aimed to ascertain BoostDM's reliability in the germline context, despite its original design for somatic variant analysis. Our data supports the potential applicability of BoostDM beyond its initial somatic mutation focus and suggests its utility for germline variant analysis in cancer genes. These findings, like those with the previous filter, demonstrate enhanced diagnostic performance with the incorporation of massive sequencing methods such as NGS in the analysis of cancer-related mutations. Recently, Garg et al., (93), indicated that the use of NGS increases the identification of mutations by approximately 36% compared to single-gene analysis, highlighting the clear benefit of using massive molecular analysis strategies (93).

Of the total germline variants classified as oncdriver by BoostDM, a higher proportion of changes in genes related to CRC predisposition syndromes was identified compared to filter A. Significantly, 20.6% of molecular variants were found in these genes, and it was possible to identify them in genes related to Lynch Syndrome; *MSH6* (n=5) and *MLH1* (n=3). As mentioned earlier, these findings allow for advising patients and their relatives following international guidelines, potentially improving the early detection of at-risk individuals and the management of carriers of such variants (94). Additionally, this model allowed the identification of other variants in genes related to syndromic CRC, such as *APC*, *BMPRIA*, *MSH2*, *MSH3*, *NTHL1*, and *PMS2*. These molecular changes, along with those observed in genes related to DNA damage repair signaling pathways, demonstrate the ability of this artificial intelligence algorithm to detect oncdriver variants (76, 85, 86). Variations in these genes have been reported in large cohorts of patients with highly significant associations with the development of CRC, such as *MSH2* (OR: 18.1), *MLH1* (OR: 8.6), and *APC* (OR: 49.4), supporting the importance of germline variant identification studies in an unselected population to contribute to the generation of genetic assessment policies and variant interpretation in CRC (95).

In this context, based on our results, we believe that the oncdriver variants identified particularly those not detected by conventional prediction algorithms hold potential for inclusion in future clinical prediction panels. However, their inclusion requires further validation, and we propose the following next steps: a) replication in larger and independent cohorts of CRC patients, b) functional validation of candidate variants to confirm their biological relevance, and c) integration of these variants into multigene risk models to evaluate their predictive value in clinical practice. Additionally, drawing on international experiences such as the Personalized OncoGenomics (POG) program in Canada (96), it would be necessary to define a framework for identifying, evaluating, and reporting research-based germline findings within the clinical infrastructure of a publicly funded healthcare system. In this context, since the variants identified by BoostDM are potentially oncogenic, prioritization could be given to those associated with moderate to high penetrance cancer susceptibility genes, or variants in cancer predisposition genes known to influence tumor phenotype and evolution.

Recent studies have recognized the involvement of synonymous variants as causal factors in Mendelian and multifactorial diseases, including cancer (97–103). Interestingly, the BoostDM model identified two synonymous variants in the *CDH1* (c.1710T>C) and *NFI* (c.3498C>T) genes as oncdrivers. This finding is particularly intriguing because it involves genes related to neurofibromatosis and hereditary diffuse gastric cancer. The *NFI* gene is a tumor suppressor that contributes to cancer development and has been associated with gastrointestinal tract adenocarcinoma (104). Studies by Seminog and Goldacre (105) determined that patients with pathogenic mutations in the *NFI* gene have a higher risk of colon cancer compared to the general population, supporting our findings in the patient analyzed (105). It is important to note that, although the variant in the *NFI* gene is classified as an

oncogene by Filter B, it does not meet the criteria to be classified as pathogenic or likely pathogenic (P/LP) according to ACMG/AMP guidelines. Therefore, this variant is not reported as associated with neurofibromatosis.

On the other hand, pathogenic germline variants in the *CDH1* gene are responsible for over 20% of hereditary diffuse gastric cancer. However, consistent with our findings, recent studies have established that carrying a variant in this gene generates a higher predisposition to colorectal polyps, suggesting a potential association between *CDH1* variants and CRC risk. These data are relevant for cancer risk assessment and patient counseling with variants in this gene (106). The molecular mechanisms associated with the pathogenicity of synonymous variants have been explored, but it is generally considered that they mostly affect splicing, generating aberrant transcripts (e.g., exon skipping). A recent study analyzing this type of variants in over 3000 cancer samples demonstrated that around 6–8% of all driver variants in oncogenes are synonymous (107). In this context, we can highlight the importance of incorporating additional methods for detecting variants related to cancer development, as ACMG/AMP criteria may exclude these variants. However, the findings must be approached with caution and supported by functional validation studies.

Additionally, the contribution of oncogene variants in genes usually used in diagnostic panels, combined with extended molecular analysis of candidate genes, has a significant impact, as demonstrated by our results in which variants were identified in *FANCC*, *KDR*, and *TCF7L2*. *FANCC* is involved in DNA repair and transcription processes and has been associated with an increased risk of CRC (108–110). On the other hand, *KDR* encodes one of the two receptors for VEGF. This is a type III tyrosine kinase receptor involved in the proliferation, survival, migration, and tubular morphogenesis of endothelial cells and has been associated as a prognostic marker in CRC (111, 112). Finally, *TCF7L2* plays a key role in the Wnt signaling pathway, and variants in this gene have also been found in patients with CRC (113, 114). Taken together, the findings derived from the BoostDM model highlight a high capacity for identifying oncogene mutations, supporting recent claims about the use of artificial intelligence as a tool that could enhance precision and effectiveness in cancer prediction, diagnosis, and treatment (115).

The population-genetic analysis carried out in our study identified, for the first time in the country and in Latin America, the frequency of pathogenic, likely pathogenic, and oncogene variants in 206 genes related to CRC. Comparing allele frequencies with data obtained from the gnomAD database allowed us to establish statistically significant differences between the global and Latin American populations. The greatest differences were observed in non-Latin American populations, demonstrating that approximately 46% of the variants identified by us are presented at higher or lower frequencies. In contrast, only 23% to 29% of the variants in this study were different from other Latin American populations. This finding is relevant as it underscores the need for studies like the present one, where genomic

characterization is performed in populations typically underrepresented in public databases.

Taken together, the findings of this study suggest the need to conduct expanded analyses in the Colombian population using whole-exome sequencing (WES) or other high-throughput sequencing methods (e.g., whole-genome sequencing) to establish population-specific variant databases. Our results support this recommendation based on the following observations: (a) 13% of the variants identified were classified as pathogenic or likely pathogenic (P/LP), (b) 38.5% of the P/LP variants were found in genes not currently included in standard diagnostic panels, and (c) we propose a novel classification approach capable of identifying oncogene variants not detected by conventional algorithms.

Finally, we conducted a functional validation analysis of pathogenic variants identified in genes not conventionally involved in molecular diagnostic genetic panels to contribute new knowledge about their potential involvement in CRC etiology. In this context, intronic variants in canonical splicing sites were evaluated using minigene assay. This type of variant is recognized for its potential effect on splicing and the generation of aberrant transcripts through mechanisms such as exon skipping, intron retention, or pseudogene generation (116).

Our findings revealed that the *SMAD9*: c.781 + 2T>A variant leads to exon skipping, resulting in the loss of 37 amino acids. This molecular effect potentially impacts the protein's function, crucial in the TGF β signaling pathway. This pathway has been associated with the regulation of pro-oncogenic processes, such as invasion, epithelial-mesenchymal transition, and the promotion of angiogenesis (117). SNPs in this gene have been identified in CRC patients and associated with an increased risk, mainly due to its role in the TGF β signaling pathway (38, 118–121).

The *ERCC1* c.702 + 1G>A mutation was classified as an oncogene by BoostDM, and Likely Pathogenic according to the ACMG classification criteria, as detailed in Table 3 and Supplementary Table 1. Based on both classification frameworks, this variant can be considered pathogenic. Furthermore, the ACMG PVS1 criterion supports the hypothesis of a splicing alteration. Our minigene assay results demonstrated exon skipping without the production of a canonical transcript in the HCT116 cell line, consistent with *In Silico* predictions and bioinformatics based classifications. This mutation is predicted to generate a truncated protein (p.Ser201Argfs*1), which lacks the C-terminal HhH2 domain of ERCC1 (residues 220–297) required for dimerization with XPF (122, 123). The ERCC1-XPF complex is a structure-specific endonuclease involved in nucleotide excision repair (NER), interstrand crosslink (ICL) repair, and double-strand break (DSB) repair. Consequently, its impairment significantly compromises DNA repair capacity, increasing the risk of cancer (124).

However, in our functional validation, exon skipping was also observed in the wild-type plasmid, although the canonical transcript remained detectable in HCT116 cell line. This result is difficult to explain, as although *ERCC1* has been reported to produce isoforms through alternative splicing, these transcript variants differ mainly in their untranslated regions or result from

intron retention (125). In this context, additional studies may be required to confirm this finding.

The variant in *IL12RB1*: c.1791 + 2T>G had the most significant impact, resulting in the generation of two aberrant transcripts: exon 15 skipping (76 bp), and intron 14 partial retention (98 bp). Both effects caused a reading frame shift and the generation of a premature stop codon, potentially affecting the structural integrity of the protein. *IL12RB1* is an interleukin receptor that plays a role in DNA damage repair pathways (82, 126). Variants in this gene in CRC patients provide convincing evidence of its role in predisposition to this neoplasia (126, 127).

In agreement with recent reports, we confirmed that the pSplice Express vector, used in combination with the Gateway recombination system, is an effective approach for evaluating splicing alterations. Nearly twenty studies in the literature have successfully applied this strategy, reporting relevant findings related to exon skipping, intron retention, cryptic splice sites, among others. Notably, this approach is suitable for assessing mutations in canonical splice sites as well as deep intronic variants, further supporting its technical robustness (121, 126–130).

Study limitations

The present study has several noteworthy limitations, which could be summarized as technical and biological. The technical limitations are related to the NGS methodology employed, based on short-read sequencing, leading to difficulties in the variant detection in exons with high GC content, highly homologous, repetitive, or low complexity regions, and pseudogenes resulting in false-positive or false-negative variant calls due to inaccurate sequence alignment and variant calling in these challenging regions (128–131). Another technical limitation correlates with the design of the exome probes and the use of the exome capture kits. It has been reported that they can lead to biases in variant detection due to uneven coverage of some coding regions, resulting in systematic insufficiencies in sequencing depth for certain genes, leading to false negative pathogenic variant calls (132, 133).

The NGS methodology employed falls short in identifying pathogenic variants within deep promoter or intronic regions. Furthermore, this study did not investigate Copy Number Variations (CNVs), which have been documented as potential causative factors in Colorectal Cancer (CRC). Concerning the biological limitations, we assessed germline variants from peripheral blood from a cohort of Colombian patients with unselected CRC. This methodology only allowed us to evaluate the genomics of a bigger landscape that requires a multi-omic approach involving transcriptomics, proteomics, and metabolomics to better understand CRC and the correlation between germline and somatic variants in this intricate network (134–136). Functional validation studies focused on splicing variants' impact on mRNA, yet protein validations were not conducted. It is also critical to recognize that, while the concordance analysis between BoostDM and AlphaMissense provided insights into BoostDM's applicability to germline variants, both models have inherent limitations in

predicting variant pathogenicity. Further study is required to determine functional consequences of the identified variants. Specifically, the reliance on computational predictions without functional validation may not capture the complete biological impact of certain variants. The BoostDM artificial intelligence analysis also identified variants in genes with undetermined penetrance, necessitating cautious interpretation. Additionally, no functional validation was conducted for synonymous variants proposed as oncdrivers.

We performed a minigene assay, a widely used methodology to evaluate the effect of intronic variants on splicing. This assay allows assessment of the variant's impact on splicing; however, its potential oncogenic effect can be further investigated using complementary approaches such as proliferation, invasion, and apoptosis assays, among others.

Conclusion

In conclusion, our findings played a crucial role in delineating the germline mutational landscape among unselected CRC patients within the Colombian population, using a comprehensive multigene panel that includes genes from both established diagnostic panels and candidate genes. The importance of evaluating genes typically omitted from routine diagnostic procedures was evident, shedding light on the potential oversight of a substantial proportion of P/LP variants. This omission is primarily due to the limited availability of information regarding their association with CRC. Therefore, expanding the number of genes potentially related to the etiology of the disease could help us to understand how germline variants contribute to increased susceptibility to CRC. Understanding the genetic predisposition for CRC is essential for early diagnosis, prevention, and patient treatment. While improvements in sequencing technologies and the emergence of advanced artificial intelligence bioinformatics platforms, such as the BoostDM model, have expanded our tools for understanding the mechanisms by which variants affect genes or proteins, including synonymous variants that are often not considered pathogenic, it is necessary to further support the molecular involvement with functional validation analyses.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Universidad del Rosario's Ethics Committee (DVO005 1607-CV1436). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written

informed consent to participate in this study. Ethical approval was not required for the studies on animals in accordance with the local legislation and institutional requirements because only commercially available established cell lines were used.

Author contributions

JR: Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. MA: Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. SO: Formal Analysis, Software, Writing – original draft, Writing – review & editing. CD: Formal Analysis, Software, Writing – original draft, Writing – review & editing. DS: Methodology, Validation, Writing – original draft, Writing – review & editing. NC: Formal Analysis, Software, Writing – original draft, Writing – review & editing. CF: Methodology, Validation, Writing – original draft, Writing – review & editing. CR: Methodology, Validation, Writing – original draft, Writing – review & editing. AL: Methodology, Writing – original draft, Writing – review & editing. RC: Methodology, Validation, Writing – original draft, Writing – review & editing. AM: Conceptualization, Formal Analysis, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. DF: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This project was supported by the Universidad del Rosario (Grant ABN062) and Hospital Universitario Mayor MÉDERI (Grant QAN-BG276).

References

- Morgan E, Arnold M, Gini A, Lorenzoni V, Cabasag CJ, Laversanne M, et al. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut*. (2023) 72:338–44. doi: 10.1136/gutjnl-2022-327736
- Chen J, Zheng Y, Wang H, Zhang D, Zhao L, Yu D, et al. Cause of death among patients with colorectal cancer: a population-based study in the United States. *Aging (Albany NY)*. (2020) 12:22927–48. doi: 10.18632/aging.104022
- Globocan. International Agency for research on cancer - WHO(2022). Available online at: <https://gco.iarc.fr/today/> (Accessed April 2, 2025).
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2021) 1:1–41. doi: 10.3322/caac.21660
- Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. (2024) 74:12–49. doi: 10.3322/caac.21820
- Ferlitsch M, Reinhart K, Pramhas S, Wiener C, Gal O, Bannert C, et al. Sex-specific prevalence of adenomas, advanced adenomas, and colorectal cancer in individuals undergoing screening colonoscopy. *JAMA*. (2011) 306:1352. doi: 10.1001/jama.2011.1362
- Ahmad R, kumar SJ, Wunnava A, Al-obeed O, Absulla M, Srivastava SK. Emerging trends in colorectal cancer: Dysregulated signaling pathways (Review). *Int J Mol Med*. (2021) 47:1–25. doi: 10.3892/ijmm.2021.4847
- Carvalho TC de, Borges AK daM, Koifman RJ, Silva IF da. Time trends in colorectal cancer incidence in four regions of Latin America: 1983–2012. *Cad Saude Publica*. (2021) 37:e00175720. doi: 10.1590/0102-311x00175720
- Piñeros M, Gamboa O, Hernández-Suárez G, Pardo C, Bray F. Patterns and trends in cancer mortality in Colombia 1984–2008. *Cancer Epidemiol*. (2013) 37:233–9. doi: 10.1016/j.canep.2013.02.003
- Cuenta de alto costo (CAC)/HIGIA(2022). Available online at: <https://cuentadealtocosto.org/higia/cancer-morbimortalidad-demografico/> (Accessed April 2, 2025).
- Carioli G, Bertuccio P, Malvezzi M, Boffetta P, Levi F, Negri E, et al. Cancer mortality predictions for 2021 in Latin America. *Eur J Cancer Prev*. (2022) 31:217–27. doi: 10.1097/CEJ.0000000000000702
- Lewandowska A, Rudzki G, Lewandowski T, Strykowska-Góra A, Rudzki S. Risk factors for the diagnosis of colorectal cancer. *Cancer Control*. (2022) 29:1–15. doi: 10.1177/10732748211056692

Acknowledgments

We would like to thank Hospital Universitario Mayor Méderi for patients' support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2025.1568205/full#supplementary-material>

13. Xu P, Sun D, Gao Y, Jiang Y, Zhong M, Zhao G, et al. Germline mutations in a DNA repair pathway are associated with familial colorectal cancer. *JCI Insight*. (2021) 6:1–14. doi: 10.1172/jci.insight.148931
14. Kastrinos F, Samadder NJ, Burt RW. Use of family history and genetic testing to determine risk of colorectal cancer. *Gastroenterology*. (2020) 158:389–403. doi: 10.1053/j.gastro.2019.11.029
15. Stoffel EM, Kastrinos F. Familial colorectal cancer, beyond lynch syndrome. *Clin Gastroenterol Hepatol*. (2014) 12:1059–68. doi: 10.1016/j.cgh.2013.08.015
16. Issaka RB, Chan AT, Gupta S. AGA clinical practice update on risk stratification for colorectal cancer screening and post-polypectomy surveillance: expert review. *Gastroenterology*. (2023) 165:1280–91. doi: 10.1053/j.gastro.2023.06.033
17. Shaikat A, Kahi CJ, Burke CA, Rabeneck L, Sauer BG, Rex DK. ACG clinical guidelines: colorectal cancer screening 2021. *Am J Gastroenterol*. (2021) 116:458–79. doi: 10.14309/ajg.0000000000001122
18. Poliani L, Greco L, Barile M, Buono AD, Bianchi P, Basso G, et al. Canonical and uncanonical pathogenic germline variants in colorectal cancer patients by next-generation sequencing in a European referral center. *ESMO Open*. (2022) 7:100607. doi: 10.1016/j.esmoop.2022.100607
19. National Comprehensive Cancer Network. Clinical practice guidelines in oncology. *Genetic/Familial High-Risk Assessment: Colorectal*. (2023). https://www.nccn.org/professionals/physician_gls/pdf/genetics_colon.pdf (Accessed April 2, 2025).
20. Gong R, He Y, Liu XY, Wang HY, Sun LY, Yang XH, et al. Mutation spectrum of germline cancer susceptibility genes among unselected Chinese colorectal cancer patients. *Cancer Manag Res*. (2019) 11:3721–39. doi: 10.2147/CMAR.S193985
21. Yurgelun MB, Kulke MH, Fuchs CS, Allen BA, Uno H, Hornick JL, et al. Cancer susceptibility gene mutations in individuals with colorectal cancer. *J Clin Oncol*. (2017) 35:1086–95. doi: 10.1200/JCO.2016.71.0012
22. Rebuzzi F, Ulivi P, Tedaldi G. Genetic predisposition to colorectal cancer: how many and which genes to test? *Int J Mol Sci*. (2023) 24:2137. doi: 10.3390/ijms24032137
23. Olkinuora AP, Peltomäki PT, Aaltonen LA, Rajamäki K. From APC to the genetics of hereditary and familial colon cancer syndromes. *Hum Mol Genet*. (2021) 30:R206–24. doi: 10.1093/hmg/ddab208
24. Stodolna A, He M, Vasipalli M, Kingsbury Z, Becq J, Stockton JD, et al. Clinical-grade whole-genome sequencing and 3' transcriptome analysis of colorectal cancer patients. *Genome Med*. (2021) 13:1–15. doi: 10.1186/s13073-021-00852-8
25. Sonkin D, Thomas A, Teicher BA. Cancer treatments: Past, present, and future. *Cancer Genet*. (2024) 286–287:18–24. doi: 10.1016/j.cancergen.2024.06.002
26. Alonso-Espinaco V, Giraldez MD, Trujillo C, van der Klift H, Muñoz J, Balaguer F, et al. Novel MLH1 duplication identified in Colombian families with Lynch syndrome. *Genet Med*. (2011) 13:155–60. doi: 10.1097/GIM.0b013e318202e10b
27. Alvarez-Gomez RM, de la Fuente-Hernandez MA, Herrera-Montalvo L, Hidalgo-Miranda A. Challenges of diagnostic genomics in Latin America. *Curr Opin Genet Dev*. (2021) 66:101–9. doi: 10.1016/j.gde.2020.12.010
28. Palacio-Rúa KA, Isaza-Jiménez LF, Ahumada-Rodríguez E, Ceballos-García H, Muñeton-Peña CM. Genetic analysis in APC, KRAS, and TP53 in patients with stomach and colon cancer. *Rev Gastroenterol Mexico (English Ed)*. (2014) 79:79–89. doi: 10.1016/j.rgmexn.2014.06.006
29. Palacio Rúa KA, Ceballos García H, Ahumada Rodríguez E, Isaza Jiménez LF, Muñeton Peña CM. Molecular characterization of TP53 tumor suppressor gene in colorectal cancer. *Rev Colomb Gastroenterol*. (2013) 28:294–300. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84893405008&partnerID=40&md5=f1544d1cf6ec55dfa1faff2c91c069e0> (Accessed February 11, 2025).
30. Hampel H, Bennett RL, Buchanan A, Pearlman R, Wiesner GL. A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: Referral indications for cancer predisposition assessment. *Genet Med*. (2015) 17:70–87. doi: 10.1038/gim.2014.147
31. Syngal S, Brand RE, Church JM, Giardiello FM, Hampel HL, Burt RW. ACG clinical guideline: Genetic testing and management of hereditary gastrointestinal cancer syndromes. *Am J Gastroenterol*. (2015) 110:223–62. doi: 10.1038/ajg.2014.435
32. Muñíos F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. *Nature*. (2021) 596:428–32. doi: 10.1038/s41586-021-03771-1
33. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (80-)*. (2023) 381:eadg7492. doi: 10.1126/science.adg7492
34. Putscher E, Hecker M, Fitzner B, Lorenz P, Zettl UK. Principles and Practical Considerations for the Analysis of Disease-Associated Alternative Splicing Events Using the Gateway Cloning-Based Minigene Vectors pDESTsplice and pSpliceExpress. *Int J Mol Sci*. (2021) 22:5154. doi: 10.3390/ijms22105154
35. Echeverría-garcés G, Ramos-medina MJ, Vargas R, Cabrera-andrade A, Altamirano-colina A, Freire MP, et al. Gastric cancer actionable genomic alterations across diverse populations worldwide and pharmacogenomics strategies based on precision oncology. *Front Pharmacol*. (2024) 15:1–14. doi: 10.3389/fphar.2024.1373007
36. Echeverría-garc G, Jos M, Ramírez-s D, Altamirano-colina A, Guerrero S, Qui LA. Worldwide analysis of actionable genomic alterations in lung cancer and targeted pharmacogenomic strategies Ariana Gonz a Isaac Armend a Sebasti a es L o es European Society for Medical Oncology. *Heliyon* (2024) 10:2405–8440. doi: 10.1016/j.heliyon.2024.e37488
37. Jiang TJ, Wang F, Wang YN, Hu JJ, Ding PR, Lin JZ, et al. Germline mutational profile of Chinese patients under 70 years old with colorectal cancer. *Cancer Commun*. (2020) 40:620–32. doi: 10.1002/cac2.v40.11
38. Djursby M, Madsen MB, Frederiksen JH, Berchtold LA, Therkildsen C, Willemoe GL, et al. New pathogenic germline variants in very early onset and familial colorectal cancer patients. *Front Genet*. (2020) 11:1–16. doi: 10.3389/fgene.2020.566266
39. Zhang J, Wang X, de Voer RM, Hehir-Kwa JY, Kamping EJ, Weren RDA, et al. A molecular inversion probe-based next-generation sequencing panel to detect germline mutations in Chinese early-onset colorectal cancer patients. *Oncotarget*. (2017) 8:24533–47. doi: 10.18632/oncotarget.15593
40. Pearlman R, Frankel WL, Swanson B, Zhao W, Yilmaz A, Miller K, et al. Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. *JAMA Oncol*. (2017) 3:464–71. doi: 10.1001/jamaoncol.2016.5194
41. Uson PLS, Riegert-Johnson D, Boardman L, Kiesel J, Mountjoy L, Patel N, et al. Germline cancer susceptibility gene testing in unselected patients with colorectal adenocarcinoma: A multicenter prospective study. *Clin Gastroenterol Hepatol*. (2022) 20:e508–28. doi: 10.1016/j.cgh.2021.04.013
42. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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*. (2015) 17:405–24. doi: 10.1111/aos.13972_502
43. Roht L, Tooming M, Rekker K, Roomere H, Toome K, Murumets Ü, et al. The prevalence of germline pathogenic variants in Estonian colorectal cancer patients: results from routine clinical setting 2016–2021. *Front Genet*. (2022) 13:1020543/full. doi: 10.3389/fgene.2022.1020543/full
44. Vaccaro CA, López-Kostner F, Adriana DV, Palmero EI, Rossi BM, Antelo M, et al. From colorectal cancer pattern to the characterization of individuals at risk: Picture for genetic research in Latin America. *Int J Cancer*. (2019) 145:318–26. doi: 10.1002/ijc.v145.2
45. Blatter R, Tschupp B, Aretz S, Bernstein I, Colas C, Evans DG, et al. Disease expression in juvenile polyposis syndrome: a retrospective survey on a cohort of 221 European patients and comparison with a literature-derived cohort of 473 SMAD4/BMPRIA pathogenic variant carriers. *Genet Med*. (2020) 22:1524–32. doi: 10.1016/j.jss.2013.01.015
46. Howe JR, Dahdaleh FS, Carr JC, Wang D, Sherman SK, Howe JR. BMPRIA mutations in juvenile polyposis affect cellular localization. *J Surg Res*. (2013) 184:739–45. doi: 10.1186/s12881-020-01135-6
47. Liu Q, Liu M, Liu T, Yu Y. Familial juvenile polyposis syndrome with a *de novo* germline missense variant in BMPRIA gene: a case report. *BMC Med Genet*. (2020) 21:196. doi: 10.1186/s12881-020-01135-6
48. Waite KA, Eng C. From developmental disorder to heritable cancer: it's all in the BMP/TGF- β family. *Nat Rev Genet*. (2003) 4:763–73. doi: 10.1038/s41559-018-0642-z
49. Cross W, Kovac M, Mustonen V, Temko D, Davis H, Baker A-M, et al. The evolutionary landscape of colorectal tumorigenesis. *Nat Ecol Evol*. (2018) 2:1661–72. doi: 10.1016/S0092-8674(00)81296-9
50. Massagué J. TGF β Signaling: receptors, transducers, and mad proteins. *Cell*. (1996) 85:947–50. doi: 10.1016/S0092-8674(00)81296-9
51. Zhao M, Mishra L, Deng C-X. The role of TGF- β /SMAD4 signaling in cancer. *Int J Biol Sci*. (2018) 14:111–23. doi: 10.7150/ijbs.23230
52. Ngeow J, Heald B, Rybicki LA, Orloff MS, Chen JL, Liu X, et al. Prevalence of germline PTEN, BMPRIA, SMAD4, STK11, and ENG mutations in patients with moderate-load colorectal polypos. *Gastroenterology*. (2013) 144:1402–1409.e5. doi: 10.1053/j.gastro.2013.02.001
53. You S, Zhou J, Chen S, Zhou P, Lv J, Han X, et al. PTCH1, a receptor of Hedgehog signaling pathway, is correlated with metastatic potential of colorectal cancer. *Ups J Med Sci*. (2010) 115:169–75. doi: 10.3109/03009731003668316
54. Raleigh DR, Reiter JF. Misactivation of Hedgehog signaling causes inherited and sporadic cancers. *J Clin Invest*. (2019) 129:465–75. doi: 10.1172/JCI120850
55. Chassaing N, Davis EE, McKnight KL, Niederriter AR, Causse A, David V, et al. Targeted resequencing identifies PTCH1 as a major contributor to ocular developmental anomalies and extends the SOX2 regulatory network. *Genome Res*. (2016) 26:474–85. doi: 10.1101/gr.196048.115
56. Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV. A superfamily of conserved domains in DNA damage- responsive cell cycle checkpoint proteins. *FASEB J*. (1997) 11:68–76. doi: 10.1096/fasebj.11.1.9034168
57. Gerloff DL, Woods NT, Farago AA, Monteiro ANA. BRCT domains: A little more than kin, and less than kind. *FEBS Lett*. (2012) 586:2711–6. doi: 10.1016/j.febslet.2012.05.005
58. Jiang Q, Greenberg RA. Deciphering the BRCA1 tumor suppressor network. *J Biol Chem*. (2015) 290:17724–32. doi: 10.1074/jbc.R115.667931
59. Leung CCY, Glover JNM. BRCT domains. *Cell Cycle*. (2011) 10:2461–70. doi: 10.4161/cc.10.15.16312

60. Wu D, Huang H, Chen T, Gai X, Li Q, Wang C, et al. The BRCA1/BARD1 complex recognizes pre-ribosomal RNA to facilitate homologous recombination. *Cell Discov.* (2023) 9:99. doi: 10.1038/s41421-023-00590-8
61. Billing D, Horiguchi M, Wu-Baer F, Tagliatalata A, Leuzzi G, Nanez SA, et al. The BRCT domains of the BRCA1 and BARD1 tumor suppressors differentially regulate homology-directed repair and stalled fork protection. *Mol Cell.* (2018) 72:127–139.e8. doi: 10.1016/j.molcel.2018.08.016
62. Peña-Guerrero J, Fernández-Rubio C, García-Sosa AT, Nguewa PA. BRCT Domains: Structure, Functions, and Implications in Disease—New Therapeutic Targets for Innovative Drug Discovery against Infections. *Pharmaceutics.* (2023) 15:1839. doi: 10.3390/pharmaceutics15071839
63. Sattler EC, Syunyaeva Z, Reithmair M, Dempke W, Steinlein OK. Colorectal cancer risk in families with Birt-Hogg-Dubé syndrome increased. *Eur J Cancer.* (2021) 151:168–74. doi: 10.1016/j.ejca.2021.04.013
64. Jirka GW, Lefler DS, Russo J, Bashir B. Colon adenocarcinoma and Birt-Hogg-Dubé syndrome in a young patient: case report and exploration of pathologic implications. *Cancer Biol Ther.* (2023) 24:2184153. doi: 10.1080/15384047.2023.2184153
65. Nahorski MS, Lim DHK, Martin L, Gille JJP, McKay K, Rehal PK, et al. Investigation of the Birt-Hogg-Dubé tumour suppressor gene (FLCN) in familial and sporadic colorectal cancer. *J Med Genet.* (2010) 47:385–90. doi: 10.1136/jmg.2009.073304
66. Schmidt LS, Linehan WM. FLCN: The causative gene for Birt-Hogg-Dubé syndrome. *Gene.* (2018) 640:28–42. doi: 10.1016/j.gene.2017.09.044
67. van de Beek I, Glykofridis IE, Wolthuis RMF, Gille HJJP, Johannesma PC, Meijers-Heijboer HEJ, et al. No evidence for increased prevalence of colorectal carcinoma in 399 Dutch patients with Birt-Hogg-Dubé syndrome. *Br J Cancer.* (2020) 122:590–4. doi: 10.1038/s41416-019-0693-1
68. Whitworth J, Skytte A-B, Sunde L, Lim DH, Arends MJ, Happerfield L, et al. Multilocus inherited neoplasia alleles syndrome: A case series and review. *JAMA Oncol.* (2016) 2:373. doi: 10.1001/jamaoncol.2015.4771
69. Ramirez Reyes MJ, Cuesta R, Pause A. Folliculin: A regulator of transcription through AMPK and mTOR signaling pathways. *Front Cell Dev Biol.* (2021) 9:667311/full. doi: 10.3389/fcell.2021.667311/full
70. Huang K, Luo W, Fang J, Yu C, Liu G, Yuan X, et al. Notch3 signaling promotes colorectal tumor growth by enhancing immunosuppressive cells infiltration in the microenvironment. *BMC Cancer.* (2023) 23:55. doi: 10.1186/s12885-023-10526-w
71. Xiu M, Wang Y, Li B, Wang X, Xiao F, Chen S, et al. The role of notch3 signaling in cancer stemness and chemoresistance: molecular mechanisms and targeting strategies. *Front Mol Biosci.* (2021) 8:694141/full. doi: 10.3389/fmolb.2021.694141/full
72. Zhang Y-Q, Liang Y-K, Wu Y, Chen M, Chen W-L, Li R-H, et al. Notch3 inhibits cell proliferation and tumorigenesis and predicts better prognosis in breast cancer through transactivating PTEN. *Cell Death Dis.* (2021) 12:502. doi: 10.1038/s41419-021-03735-3
73. Alanazi IO, Shaik JP, Parine NR, Al Naeem A, Azzam NA, Almadi MA, et al. NOTCH single nucleotide polymorphisms in the predisposition of breast and colorectal cancers in Saudi patients. *Pathol Oncol Res.* (2021) 27:616204. doi: 10.3389/pore.2021.616204/full
74. Bessho T. Nucleotide excision repair 3' endonuclease XPG stimulates the activity of base excision repair enzyme thymine glycol DNA glycosylase. *Nucleic Acids Res.* (1999) 27:979–83. doi: 10.1093/nar/27.4.979
75. Das L, Quintana VG, Sweasy JB. NTHL1 in genomic integrity, aging and cancer. *DNA Repair (Amst).* (2020) 93:102920. doi: 10.1016/j.dnarep.2020.102920
76. Williams SC, Parsons JL. NTH1 is a new target for ubiquitylation-dependent regulation by TRIM26 required for the cellular response to oxidative stress. *Mol Cell Biol.* (2018) 38:e00616-17. doi: 10.1128/MCB.00616-17
77. Carroll BL, Zahn KE, Hanley JP, Wallace SS, Dragon JA, Doublié S. Caught in motion: human NTHL1 undergoes interdomain rearrangement necessary for catalysis. *Nucleic Acids Res.* (2021) 49:13165–78. doi: 10.1093/nar/gkab1162
78. Dai Q, Luo H, Li XP, Huang J, Zhou TJ, Yang Z-H. XRCC1 and ERCC1 polymorphisms are related to susceptibility and survival of colorectal cancer in the Chinese population. *Mutagenesis.* (2015) 30:441–9. doi: 10.1093/mutage/geu088
79. Ngeow J, Yu W, Yehia L, Niazi F, Chen J, Tang X, et al. Exome sequencing reveals germline SMAD9 mutation that reduces phosphatase and tensin homolog expression and is associated with hamartomatous polyposis and gastrointestinal ganglioneuromas. *Gastroenterology.* (2015) 149:886–889.e5. doi: 10.1053/j.gastro.2015.06.027
80. Ni M, Zhang W, Qiu J, Liu F, Li M, Zhang Y, et al. Association of ERCC1 and ERCC2 polymorphisms with colorectal cancer risk in a Chinese population. *Sci Rep.* (2014) 4:4112. doi: 10.1038/srep04112
81. Li Y-K, Xu Q, Sun L-P, Gong Y-H, Jing J-J, Xing C-Z, et al. Nucleotide excision repair pathway gene polymorphisms are associated with risk and prognosis of colorectal cancer. *World J Gastroenterol.* (2020) 26:307–23. doi: 10.3748/wjg.v26.i3.307
82. Pietrzyk L, Torres A, Maciejewski R, Torres K. Obesity and obese-related chronic low-grade inflammation in promotion of colorectal cancer development. *Asian Pacific J Cancer Prev.* (2015) 16:4161–8. doi: 10.7314/APJCP.2015.16.10.4161
83. Bolderson E, Tomimatsu N, Richard DJ, Boucher D, Kumar R, Pandita TK, et al. Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks. *Nucleic Acids Res.* (2010) 38:1821–31. doi: 10.1093/nar/gkp1164
84. Keijzers G, Bakula D, Petr M, Madsen N, Teklu A, Mkrtchyan G, et al. Human exonuclease 1 (EXO1) regulatory functions in DNA replication with putative roles in cancer. *Int J Mol Sci.* (2018) 20:74. doi: 10.3390/ijms20010074
85. Reilly NM, Novara L, Di Nicolantonio F, Bardelli A. Exploiting DNA repair defects in colorectal cancer. *Mol Oncol.* (2019) 13:681–700. doi: 10.1002/1878-0261.12467
86. Rubenstein JH, Tavakkoli A, Koeppe E, Ulintz P, Inadomi JM, Morgenstern H, et al. Family history of colorectal or esophageal cancer in Barrett's esophagus and potentially explanatory genetic variants. *Clin Transl Gastroenterol.* (2020) 11:e00151. doi: 10.14309/ctg.0000000000000151
87. Wu Y, Mensink RGJ, Verlind E, Sijmons RH, Buys CHCM, Hofstra RMW, et al. Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. *Gastroenterology.* (2001) 120:1580–7. doi: 10.1053/gast.2001.25117
88. Zhang Y, He B-S, Pan Y-Q, Xu Y-Q, Wang S-K. Association of OGG1 Ser326Cys polymorphism with colorectal cancer risk: a meta-analysis. *Int J Colorectal Dis.* (2011) 26:1525–30. doi: 10.1007/s00384-011-1258-9
89. Lai C-Y, Hsieh L-L, Tang R, Santella RM, Chang-Chieh CR, Yeh C-C. Association between polymorphisms of APE1 and OGG1 and risk of colorectal cancer in Taiwan. *World J Gastroenterol.* (2016) 22:3372–80. doi: 10.3748/wjg.v22.i12.3372
90. Moreno V, Gemignani F saturation mutagenesis of cancer genes, Landi S, Gioia-Patricola L, Chabrier A, Blanco I, et al. Polymorphisms in genes of nucleotide and base excision repair: risk and prognosis of colorectal cancer. *Clin Cancer Res.* (2006) 12:2101–8. doi: 10.1158/1078-0432.CCR-05-1363
91. Wallace SS, Murphy DL, Sweasy JB. Base excision repair and cancer. *Cancer Lett.* (2012) 327:73–89. doi: 10.1016/j.canlet.2011.12.038
92. Hamamoto R, Koyama T, Kouno N, Yasuda T, Yui S, Sudo K, et al. Introducing AI to the molecular tumor board: one direction toward the establishment of precision medicine using large-scale cancer clinical and biological information. *Exp Hematol Oncol.* (2022) 11:82. doi: 10.1186/s40164-022-00333-7
93. Garg S, Grenier S, Misyura M, Sukhai MA, Thomas M, Kamel-Reid S, et al. Assessing the diagnostic yield of targeted next-generation sequencing for melanoma and gastrointestinal tumors. *J Mol Diagnostics.* (2020) 22:467–75. doi: 10.1016/j.jmoldx.2019.12.008
94. Benson AB, Venook AP, Al-Hawary MM, Arain MA, Chen Y-J, Ciombor KK, et al. Colon cancer, version 2.2021, NCCN clinical practice guidelines in oncology. *J Natl Compr Cancer Netw.* (2021) 19:329–59. doi: 10.6004/jcnccn.2021.0012
95. Fujita M, Liu X, Iwasaki Y, Terao C, Mizukami K, Kawakami E, et al. Population-based screening for hereditary colorectal cancer variants in Japan. *Clin Gastroenterol Hepatol.* (2022) 20:2132–2141.e9. doi: 10.1016/j.cgh.2020.12.007
96. Dixon K, Young S, Shen Y, Thibodeau ML, Fok A, Pleasance E, et al. Establishing a framework for the clinical translation of germline findings in precision oncology. *JNCI Cancer Spectr.* (2020) 4:pkaa045. doi: 10.1093/jncics/pkaa045
97. Bailey SF, Hinz A, Kassen R. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nat Commun.* (2014) 5:4076. doi: 10.1038/ncomms5076
98. Giacometto CJ, Benjamin R, Deng H-W, Rotter JI, Schiller MR. Most synonymous allelic variants in HIV tat are not silent. *Genomics.* (2023) 115:110603. doi: 10.1016/j.ygeno.2023.110603
99. Kaissarian NM, Meyer D, Kimchi-Sarfaty C. Synonymous variants: necessary nuance in our understanding of cancer drivers and treatment outcomes. *JNCI J Natl Cancer Inst.* (2022) 114:1072–94. doi: 10.1093/jnci/djac090
100. Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife.* (2019) 8:e45952. doi: 10.7554/eLife.45952
101. Lin BC, Katneni U, Jankowska KI, Meyer D, Kimchi-Sarfaty C. In silico methods for predicting functional synonymous variants. *Genome Biol.* (2023) 24:126. doi: 10.1186/s13059-023-02966-1
102. Zeng Z, Bromberg Y. Predicting functional effects of synonymous variants: A systematic review and perspectives. *Front Genet.* (2019) 10:914. doi: 10.3389/fgene.2019.00914
103. Zhou X, Zhou W, Wang C, Wang L, Jin Y, Jia Z, et al. A comprehensive analysis and splicing characterization of naturally occurring synonymous variants in the ATP7B gene. *Front Genet.* (2021) 11:592611/full. doi: 10.3389/fgene.2020.592611/full
104. Li Y, Bollag G, Clark R, Stevens J, Conroy L, Fults D, et al. Somatic mutations in the neurofibromatosis 1 gene in human tumors. *Cell.* (1992) 69:275–81. doi: 10.1016/0092-8674(92)90408-5
105. Seminog OO, Goldacre MJ. Risk of benign tumours of nervous system, and of Malignant neoplasms, in people with neurofibromatosis: population-based record-linkage study. *Br J Cancer.* (2013) 108:193–8. doi: 10.1038/bjc.2012.535
106. Passi M, Gamble LA, Samaranyake SG, Schueler SA, Curtin BF, Fasaye G-A, et al. Association of CDH1 germline variants and colon polyp phenotypes in patients

with hereditary diffuse gastric cancer. *Gastro Hep Adv.* (2023) 2:244–51. doi: 10.1016/j.gastha.2022.10.006

107. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell.* (2014) 156:1324–35. doi: 10.1016/j.cell.2014.01.051

108. Esteban-Jurado C, Franch-Expósito S, Muñoz J, Ocaña T, Carballal S, López-Cerón M, et al. The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. *Eur J Hum Genet.* (2016) 24:1501–5. doi: 10.1038/ejhg.2016.44

109. Laitman Y, Boker-Keinan L, Berkenstadt M, Liphshitz I, Weissglas-Volkov D, Ries-Levavi L, et al. The risk for developing cancer in Israeli ATM, BLM, and FANCC heterozygous mutation carriers. *Cancer Genet.* (2016) 209:70–4. doi: 10.1016/j.cancergen.2015.12.006

110. Martínez-Romero J, Bueno-Fortes S, Martín-Merino M, Ramirez de Molina A, De Las Rivas J. Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genomics.* (2018) 19:857. doi: 10.1186/s12864-018-5193-9

111. Dong G, Guo X, Fu X, Wan S, Zhou F, Myers RE, et al. Potentially functional genetic variants in KDR gene as prognostic markers in patients with resected colorectal cancer. *Cancer Sci.* (2012) 103:561–8. doi: 10.1111/j.1349-7006.2011.02194.x

112. Toledo RA, Garralda E, Mitsi M, Pons T, Monsech J, Vega E, et al. Exome sequencing of plasma DNA portrays the mutation landscape of colorectal cancer and discovers mutated VEGFR2 receptors as modulators of antiangiogenic therapies. *Clin Cancer Res.* (2018) 24:3550–9. doi: 10.1158/1078-0432.CCR-18-0103

113. Lu Y, Kweon S-S, Cai Q, Tanikawa C, Shu X-O, Jia W-H, et al. Identification of novel loci and new risk variant in known loci for colorectal cancer risk in east asians. *Cancer Epidemiol Biomarkers Prev.* (2020) 29:477–86. doi: 10.1158/1055-9965.EPI-19-0755

114. Testa U, Pelosi E, Castelli G. Colorectal cancer: genetic abnormalities, tumor progression, tumor heterogeneity, clonal evolution and tumor-initiating cells. *Med Sci.* (2018) 6:31. doi: 10.3390/medsci6020031

115. Sebastian AM, Peter D. Artificial intelligence in cancer research: trends, challenges and future directions. *Life.* (2022) 12:1991. doi: 10.3390/life12121991

116. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci.* (2011) 108:11093–8. doi: 10.1073/pnas.1101135108

117. Maslankova J, Vecurkovska I, Rabajdova M, Katuchova J, Kicka M, Gayova M, et al. Regulation of transforming growth factor- β signaling as a therapeutic approach to treating colorectal cancer. *World J Gastroenterol.* (2022) 28:4744–61. doi: 10.3748/wjg.v28.i33.4744

118. Chen L, Ye L, Hu B. Hereditary colorectal cancer syndromes: molecular genetics and precision medicine. *Biomedicines.* (2022) 10:3207. doi: 10.3390/biomedicines10123207

119. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet.* (2019) 51:76–87. doi: 10.1038/s41588-018-0286-6

120. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun.* (2019) 10:1–15. doi: 10.1038/s41467-019-09775-w

121. Yuan Y, Bao J, Chen Z, Villanueva AD, Wen W, Wang F, et al. Multi-omics analysis to identify susceptibility genes for colorectal cancer. *Hum Mol Genet.* (2021) 30:321–30. doi: 10.1093/hmg/ddab021

122. Tsodikov OV, Enzlin JH, Schärer OD, Ellenberger T. Crystal structure and DNA binding functions of ERCC1, a subunit of the DNA structure-specific endonuclease XPF-ERCC1. *Proc Natl Acad Sci.* (2005) 102:11236–41. doi: 10.1073/pnas.0504341102

123. Tripsianes K, Folkers G ABE, Das D, Odijk H, Jaspers NGJ, et al. The structure of the human ERCC1/XPF interaction domains reveals a complementary role for the two proteins in nucleotide excision repair. *Structure.* (2005) 13:1849–58. doi: 10.1016/j.str.2005.08.014

124. Gregg SQ, Robinson AR, Niedernhofer LJ. Physiological consequences of defects in ERCC1-XPF DNA repair endonuclease. *DNA Repair (Amst).* (2011) 10:781–91. doi: 10.1016/j.dnarep.2011.04.026

125. Zhang G, Xue P, Cui S, Yu T, Xiao M, Zhang Q, et al. Different splicing isoforms of ERCC1 affect the expression of its overlapping genes CD3EAP and PPP1R13L, and indicate a potential application in non-small cell lung cancer treatment. *Int J Oncol.* (2018) 52:2155–65. doi: 10.3892/ijo.2018.4347

126. Terradas M, Capellá G, Valle L. Dominantly inherited hereditary nonpolyposis colorectal cancer not caused by MMR genes. *J Clin Med.* (2020) 9:1954. doi: 10.3390/jcm9061954

127. Chubb D, Broderick P, Dobbins SE, Frampton M, Kinnersley B, Penegar S, et al. Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nat Commun.* (2016) 7:11883. doi: 10.1038/ncomms11883

128. Guha S, Reddi HV, Aarabi M, DiStefano M, Wakeling E, Dungan JS, et al. Laboratory testing for preconception/prenatal carrier screening: A technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet Med.* (2024) 26:101137. doi: 10.1016/j.gim.2024.101137

129. Freeman TM, England G, Wang D, Harris J. Genomic loci susceptible to systematic sequencing bias in clinical whole genomes. *Genome res.* (2020) 30:415–26. doi: 10.1101/gr.255349.119.5

130. Chinn IK, Chan AY, Chen K, Chou J, Dorsey MJ, Hajjar J, et al. Diagnostic interpretation of genetic studies in patients with primary immunodeficiency diseases: A working group report of the Primary Immunodeficiency Diseases Committee of the American Academy of Allergy, Asthma & Immunology. *J Allergy Clin Immunol.* (2020) 145:46–69. doi: 10.1016/j.jaci.2019.09.009

131. Miya F, Kato M, Shiohama T, Okamoto N, Saitoh S, Yamasaki M, et al. A combination of targeted enrichment methodologies for whole-exome sequencing reveals novel pathogenic mutations. *Sci Rep.* (2015) 5:1–7. doi: 10.1038/srep09331

132. Guo MM, Duan XN, Cui S, Tian FG, Cao XC, Geng CZ, et al. Circulating High-Molecular-Weight (HMW) adiponectin level is related with breast cancer risk better than total adiponectin: A case-control study. *PLoS One.* (2015) 10:e0129246. doi: 10.1371/journal.pone.0129246

133. Barbitoff YA, Polev DE, Glotov AS, Serebryakova EA, Shcherbakova IV, Kiselev AM, et al. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci Rep.* (2020) 10:2057. doi: 10.1038/s41598-020-59026-y

134. Liu H, Guo Z, Wang P. Genetic expression in cancer research: Challenges and complexity. *Gene Reports.* (2024) 37:102042. doi: 10.1016/j.genrep.2024.102042

135. Liu H, Li Y, Karsidag M, Tu T, Wang P. Technical and biological biases in bulk transcriptomic data mining for cancer research. *J Cancer.* (2025) 16:34–43. doi: 10.7150/jca.100922

136. Kar SP. A new frontier for cancer genetics: identification of germline-somatic associations. *Cancer Res.* (2023) 83:1165–6. doi: 10.1158/0008-5472.CAN-23-0152