

## Article

# Identification of Novel Susceptibility Genes for Early-Onset Colorectal Cancer Through Germline Rare Variant Burden Testing

Ruocen Song <sup>1,†</sup>, Reger R. Mikael <sup>2,†</sup>, Zhongping He <sup>3</sup> , Mehgan Horsnell <sup>4</sup>, Wendy Uylaki <sup>4</sup> , Weimin Meng <sup>5</sup>, Nicola K. Poplawski <sup>6,7</sup> , Bernd Wollnik <sup>8</sup>, Yun Li <sup>8</sup>, Jinghua Feng <sup>9,10</sup>, Hamish S. Scott <sup>9,10,11</sup> , Yufeng Shen <sup>12</sup>, Chen Wang <sup>13</sup>, Rui Yin <sup>5</sup> , Yousong Ding <sup>14</sup>, Xavier Llor <sup>15</sup> , Wendy K. Chung <sup>16,17</sup> , Eric Smith <sup>4,11</sup> , Timothy J. Price <sup>4,11,\*</sup> , Joanne P. Young <sup>4,11,\*</sup> and Xiao Fan <sup>1,18,\*</sup>

<sup>1</sup> Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA

<sup>2</sup> Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA

<sup>3</sup> Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA

<sup>4</sup> Department of Haematology and Oncology, The Queen Elizabeth Hospital, Woodville, SA 5011, Australia

<sup>5</sup> Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA

<sup>6</sup> Adult Genetics Unit, Royal Adelaide Hospital, Adelaide, SA 5000, Australia

<sup>7</sup> Adelaide Medical School, University of Adelaide, Adelaide, SA 5000, Australia

<sup>8</sup> Institute of Human Genetics, University Medical Center Göttingen, 37075 Göttingen, Germany

<sup>9</sup> Centre for Cancer Biology, SA Pathology and University of South Australia, Adelaide, SA 5001, Australia; hamish.scott@sa.gov.au

<sup>10</sup> Department of Genetics and Molecular Pathology, SA Pathology, Adelaide, SA 5000, Australia

<sup>11</sup> Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, SA 5000, Australia

<sup>12</sup> Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032, USA

<sup>13</sup> Department of Biostatistics, Columbia University Irving Medical Center, New York, NY 10032, USA

<sup>14</sup> Department of Medicinal Chemistry, Center for Natural Products, Drug Discovery and Development, University of Florida, Gainesville, FL 32611, USA

<sup>15</sup> Internal Medicine, Yale School of Medicine, Yale University, New Haven, CT 06520, USA

<sup>16</sup> Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115, USA

<sup>17</sup> Harvard Medical School, Boston, MA 02115, USA

<sup>18</sup> University of Florida Health Cancer Center, Gainesville, FL 32610, USA

\* Correspondence: timothy.price@sa.gov.au (T.J.P.); joanne.young@adelaide.edu.au (J.P.Y.); xiaofan@ufl.edu (X.F.)

† These authors contributed equally to this work.



Academic Editor: Luca Roncucci

Received: 12 November 2025

Revised: 1 December 2025

Accepted: 6 December 2025

Published: 9 December 2025

**Citation:** Song, R.; Mikael, R.R.; He, Z.; Horsnell, M.; Uylaki, W.; Meng, W.; Poplawski, N.K.; Wollnik, B.; Li, Y.; Feng, J.; et al. Identification of Novel Susceptibility Genes for Early-Onset Colorectal Cancer Through Germline Rare Variant Burden Testing. *Cancers* **2025**, *17*, 3931. <https://doi.org/10.3390/cancers17243931>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## Simple Summary

The genetic mechanism for colorectal cancer is complex, involving multiple risk genes, each contributing to a small proportion of cases. Our study sought to clarify this complexity, with a focus on early-onset colorectal cancer. We identified several novel colorectal cancer-susceptible genes, advancing our understanding of the disease etiology. These findings lay the groundwork for future functional validation studies and may ultimately inform the expansion of clinical gene panels for colorectal cancer.

## Abstract

**Background:** Colorectal cancer (CRC) is a leading cause of cancer death, and the incidence and mortality rates among young adults are rising. Although a subset of CRC cases presents with a family history, suggesting a hereditary component, the specific genetic underpinnings remain incompletely understood, particularly in early-onset CRC (EOCRC). This study aimed to discover novel risk genes for EOCRC using exome sequencing and gene-based rare variant burden testing. **Methods:** Our cohort consisted of 212 European-ancestry cases (174 diagnosed with CRC and 38 with significant polyps)

from the South Australian Young Onset Colorectal Polyp and Cancer Study (SAYO) and 31,699 unaffected controls from the Simons Foundation Powering Autism Research for Knowledge (SPARK) cohort. After filtering for ancestry, relatedness, variant quality, and population allele frequency, we performed gene-set and individual-gene burden tests using predicted deleterious missense and loss-of-function variants. Statistical significance was assessed using permutation-corrected binomial testing. An independent validation was conducted in the UK Biobank. **Results:** Loss-of-function variants in known CRC tumor suppressor genes were significantly enriched in SAYO cases. Gene-level analyses identified *MEIKIN* as a novel EOCRC susceptibility candidate ( $p$  value =  $1.0 \times 10^{-7}$ ), with supporting enrichment of deleterious missense and loss-of-function variants in distal colon cancer cases from the UK Biobank. Additional genes (*STK25*, *PGBD4*, *DIRAS3*, *ATG3*, *RPS6KA4*, and *DDX42*) demonstrated borderline significance, implicating pathways related to kinetochore assembly, autophagy regulation, and immune signaling. Both predicted gain-of-function and loss-of-function variants contributed to the EOCRC risk, supporting heterogeneous mechanisms of CRC pathogenesis. **Conclusions:** This study identified novel candidate risk genes for EOCRC, underscoring the role of rare variants and expanding our understanding of the genetic architecture of CRC. Future studies should include functional validation and replication studies on other ancestries to confirm and extend these results.

**Keywords:** colorectal cancer; early-onset colorectal cancer; germline variants; rare variant burden test; oncogenes; tumor suppressor genes

## 1. Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed malignancy and the second leading cause of cancer-related mortality worldwide, with an estimated 1.9 million new cases and nearly 904,000 deaths reported in 2022 [1]. While CRC is traditionally associated with aging populations, the incidence of early-onset colorectal cancer (EOCRC) has been rising in recent decades across multiple countries. Although the exact definition of EOCRC varied with age cutoffs of 30, 35, 40, 50, or 55 years in different studies, the incidence of this malignancy has been rising across all of these age groups [2,3]. In adults < 55 years old, the incidence of CRC rose by ~20% from 1994 to 2014 [2]. EOCRC exhibits distinct biological and clinical characteristics, often presenting at more advanced stages and with unique molecular features compared to CRC in older individuals. Despite these alarming trends, the risk factors contributing to EOCRC remain incompletely understood [3].

CRC is traditionally classified as sporadic or hereditary, with nearly 30% of EOCRC patients having a first-degree relative with CRC. Current reports suggest that germline pathogenic/likely pathogenic (P/LP) variants occur in 9% to 26% of EOCRC cases, depending on study design, inclusion criteria, and genetic testing methodologies, with 10% carrying P/LP in genes associated with colorectal and polyposis syndromes and 2.5% harboring P/LP variants in other cancer-related genes [4]. The genetic underpinnings of the majority of EOCRC are still unclear. Identifying genetic risk factors is crucial for mitigating cancer risk, guiding targeted prevention, optimizing therapy selection, assessing individuals at risk, and advancing research on novel therapeutic and preventive strategies.

Advances in sequencing technologies enable comprehensive exploration of rare germline coding variants and their role in disease susceptibility. By aggregating rare deleterious variants into gene-based burden tests, statistical power is enhanced, facilitating the detection of genetic contributors to cancer risk [5]. Here, we conducted a case-control association study using rare variants from exome sequencing data to identify novel EOCRC

risk genes. Given the observed heritability and role of rare variants in cancer susceptibility, we hypothesized that EOCRC risk genes should have an enriched burden of deleterious variants in EOCRC cases.

## 2. Materials and Methods

### 2.1. The South Australian Young Onset Colorectal Polyp and Cancer Study (SAYO) Case–Cohort

The SAYO investigated risk factors and warning signs for CRC/significant polyps in individuals under 55 years old [6]. The study, approved under ethics HREC/14/TQEHLMH/194, has enrolled 270 participants from 2015 to 2022 through interviews and medical record reviews. SAYO participants underwent risk assessments, including evaluations of family history, colorectal polyps, and type 2 diabetes mellitus. Among the 270 participants, 193 were diagnosed with CRC, and 44 had significant polyps only [7]. Given that significant polyps are recognized precursor lesions to CRC and may share overlapping germline susceptibility factors [8,9], the inclusion of patients with polyps provides critical insights into early disease diagnosis and increases the power to detect rare variant associations. Blood samples were collected for exome sequencing [10]. Mismatch repair (MMR) deficiency status was determined as previously described [11].

### 2.2. The Simons Foundation Powering Autism Research for Knowledge (SPARK) Control Cohort

We used the parents of probands with autism in SPARK [12] as controls for our analysis. There was no information about cancer history in these parents. We utilized the SPARK-integrated exome sequencing v2 dataset.

### 2.3. Individual Quality Control—Ancestry and Relatedness Predictions

The SPARK and SAYO relatedness predictions were completed using Kinship-based Inference for Gwas (KING) [13]. If second-degree or closer relatives were detected within the cohort, we retained only one representative per family. Ancestry estimation was conducted using Peddy [14], which projects samples based on principal component analysis (PCA) derived from the 1000 Genomes reference populations and assigns ancestry labels based on population clustering. We visualized the first two principal components colored by the predicted racial and ethnic groups among our SAYO cases. As 94.4% of our cohort are genetically inferred to be of European ancestry, we restricted our analysis to individuals of European ancestry. The same criteria were applied to the SPARK dataset.

### 2.4. Variant Calling and Annotation

SAYO patients underwent exome sequencing using the KAPA HyperPrep Kit, the Roche SeqCap EZ MedExome Enrichment Kit, and the Illumina NextSeq 500 (2 × 150 bp paired-end reads [15,16]. Sequencing reads were aligned to the hg38 reference genome using Burrows-Wheeler Aligner (BWA) version 0.7.17 [17], and variant calling was performed using the Genome Analysis Toolkit (GATK) version 4.1.9.0 [18,19]. For both the SAYO and SPARK datasets, we used Variant Effect Predictor (VEP) version 113.0 [20] to annotate variant types based on Ensembl [21–23] and population frequencies based on gnomAD v3.1 [24] and UK Biobank (UKBB) (accessed August 2025) [25]. We included missense and predicted loss-of-function (LoF) variants for risk gene analysis [26]. LoF variants included premature stop/start-gain, stop/start-loss, frameshift, and splicing variants. To assess the deleteriousness of missense variants, we applied the Variant Effect Scoring Tool version 4 (VEST4) [27], which quantifies the functional impact of variants based on evolutionary conservation and functional parameters.

### 2.5. Variant Quality Control

We applied quality control (QC) metrics to eliminate potential false variant calls and mitigate technical biases between the SAYO and SPARK datasets, ensuring the validity of our analytical approach. As detailed in Table S1, we required the following QC thresholds: a minimum genotype quality of 10, allele depth for alternative alleles of 2, a genotype depth of 7, and allele balance of 0.1. Variant Quality Score Log-Odds (VQSLOD) [19] was used in the SAYO dataset, and Phred-scaled Quality Score (QUAL) [19] was used in the SPARK dataset for variant quality scores. Our analysis was restricted to rare variants with a maximum allele frequency of 0.0001 in both the gnomAD and UKBB databases.

To correct for potential technical differences and ensure that the two datasets are comparable without systematic inflation or deflation of association signals, we utilized rare synonymous and in-frame insertion/deletion (indel) variant rates as neutral benchmarks. Since these variants are generally functionally neutral and less prone to different subpopulations, their rates should be similar across datasets. We adjusted the variant quality and allele balance scores to minimize the difference in variant rates between the two datasets. We also visually evaluated gene-based variant burden using quantile-quantile (QQ) plots of observed versus expected  $p$  values with 95% confidence intervals [28,29].

### 2.6. Known Risk Gene-Set Test

We conducted a gene-set burden test to validate our method by evaluating EOCRC associations with known CRC-related genes. Potentially deleterious missense variants were selected using a VEST4 rank score  $\geq 0.5$ , as this threshold reflects a stronger association with cancer [30]. We excluded LoF variants in oncogenes from the analysis, as those variants are protective against cancer [31]. We first analyzed a panel of 21 CRC risk genes (Table S2) according to the American Society of Clinical Oncology (ASCO) guideline [32]. As only one oncogene exists in this gene set, we expanded our risk gene set to 42 (Table S3) using the Online Mendelian Inheritance in Man (OMIM) database [33]. The OMIM database contained 21 oncogenes and 21 tumor suppressor genes (TSGs). Notably, 13 genes overlapped between the ASCO and OMIM lists. Statistical significance was evaluated using a binomial test, and the ratio of deleterious variant burdens per individual was used to measure burden enrichment.

### 2.7. Individual Risk Gene Test

We identified individual CRC risk genes using a similar approach to our gene-set analysis. As the distribution of VEST4 predicted scores varies significantly across genes, we employed a variable threshold method [34] to optimize thresholds for classifying deleterious missense variants by minimizing  $p$  values from binomial tests. The  $p$  values were then corrected using a permutation test with 10 million iterations, where the case-control status was randomly shuffled. The percentage of iterations with a lower  $p$  value than the original one from the binomial tests was set as the corrected  $p$  value. This approach maximizes statistical power and has been shown to be effective in previous genetic studies [35,36]. We conducted burden analyses based on different mechanisms, including oncogenes (gain-of-function (GoF) variants) and TSGs (LoF variants). Given that missense variants can act as either GoF or LoF, we analyzed the variants in different groups: missense only, LoF only, and combined missense and LoF. Deleterious missense variants were defined using different VEST4 thresholds in individual genes. We first tested the entire SAYO case-cohort (comprising CRC and significant polyps) and then further tested a more homogeneous case-cohort with CRC only. In total, we conducted six analyses per gene. To account for multiple testing across different cohorts and variant types, we applied the Bonferroni correction.

We also displayed variant distributions along proteins, if protein domain information is available from UniProt [37], using lollipop plots.

### 2.8. Validation in an Independent Whole-Genome Sequencing (WGS) Dataset

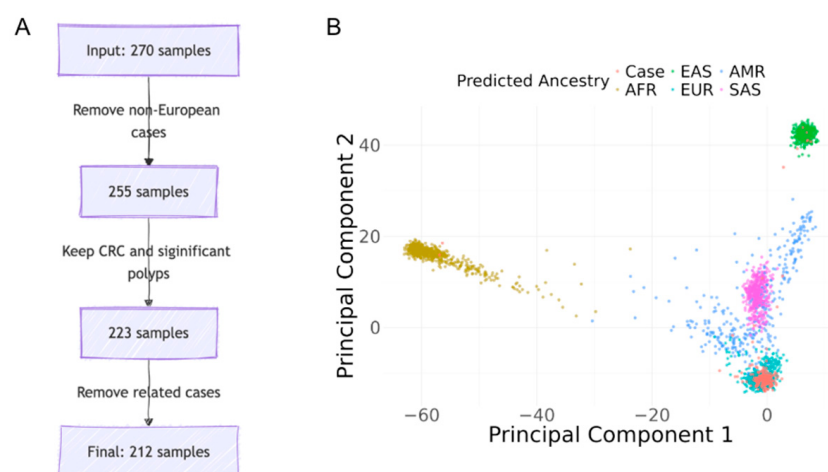
To validate our findings, we performed replication analyses on significant genes using the European-ancestry subset of the UKBB [25] whole-genome sequencing (WGS) [38] cohort. CRC phenotypes were defined based on International Classification of Diseases, 10th Revision (ICD-10) codes [39], with all ICD-9 codes converted to their corresponding ICD-10 equivalents. Since majority of SAYO participants have a distal colon, we identified two case-cohorts: proximal colon cancer includes malignant neoplasm of cecum and ascending/transverse colon (ICD-10 codes of C18.0, C18.2, C18.3, C18.4); distal colon cancer includes malignant neoplasm of descending/sigmoid colon (C18.5, C18.6, C18.7), rectosigmoid junction (C19), and rectum (C20). Participants without any of the codes in C18–20 were employed as controls.

We applied a similar variant QC pipeline (Table S1), restricting to rare variants with a maximum allele frequency of 0.0001 in both gnomAD and UKBB, a genotype quality of greater than 25, an allele depth for alternative alleles of greater than 3, a genotype depth of greater than 10, and a minimum allele balance of 0.2. Variant annotation was performed using the same VEP pipeline. For burden tests, the two CRC sub-cohorts were analyzed separately. We first applied the same VEST4 threshold identified from the SAYO dataset to validate the gene signal and then used the same variable threshold testing to examine the maximal signal from the UKBB dataset.

## 3. Results

### 3.1. Cohort Characteristics

Following the inclusion protocol described in the method, Figure 1A shows the number of individuals excluded under each criterion. Figure 1B shows a PCA plot of the SAYO cohort with predicted ancestry information.



**Figure 1.** Case selection and ancestry predictions. (A) Case selection process: from 270 individuals to the final cohort. (B) Principal component analysis for ancestry predictions in the SAYO cohort.

The entire SAYO cohort ( $n = 270$ ) had a mean age of  $41.9 \pm 10.7$  years, comprising 121 males (44.8%) and 143 females (53.0%), with the remainder not reporting their gender (Table 1).

**Table 1.** Demographics of the SAYO cohort.

| Phenotype         | Age (Years) <sup>a</sup> | Gender (%) <sup>b</sup>        | Self Reported Race (%) <sup>c</sup> |
|-------------------|--------------------------|--------------------------------|-------------------------------------|
| <b>CRC (193)</b>  | 42.6 ± 9.3               | M: 93 (48.2)<br>F: 96 (49.7)   | Caucasian: 180 (93.3)               |
| <b>Polys (44)</b> | 31.4 ± 14.1              | M: 17 (38.6)<br>F: 26 (59.1)   | Caucasian: 43 (97.7)                |
| <b>All (270)</b>  | 41.9 ± 10.7              | M: 121 (44.8)<br>F: 143 (53.0) | Caucasian: 255 (94.4)               |

<sup>a</sup> Mean value ± Standard Deviation. <sup>b</sup> Number (Percentage) of Male (M) and Female (F). <sup>c</sup> Number (Percentage) of Caucasians.

Based on genetic inference, 255 participants (94.4%) were of European ancestry, aligning closely with the 96.7% who self-identified as Caucasian. The CRC group (n = 193) had a mean age of 42.6 ± 9.3 years, comprising 93 (48.2%) males and 96 (49.7%) females, with 180 (93.3%) self-identifying as Caucasian. The polyp group (n = 44) had a mean age of 31.4 ± 14.1 years, comprising 17 (38.6%) males and 26 (59.1%) females, with 43 (97.7%) identifying as Caucasian. Individuals with non-CRC conditions, such as appendiceal neoplasms, were removed. Excluding related individuals and those of non-European ancestry, the final cohort consisted of 212 cases (174 with CRC and 38 with significant polyps) and 31,699 controls from SPARK. Among all 174 CRC cases, 83 (47.7%) reported family histories of CRC, and 143 (82.2%) reported family histories of any cancer.

### 3.2. Cleaned Datasets

To mitigate batch effects and technical artifacts, we used rates of rare synonymous and in-frame indel variants as metrics to calibrate quality filters. Variants were retained with VQSLOD greater than 2 in SAYO and QUAL greater than 38 [40], and allele balance greater than 0.1 in SAYO and SPARK (Table S1). The mean number of synonymous variants per individual was 26.43 in the SAYO cohort and 26.65 in the SPARK cohort (Figure S1B), resulting in a nonsignificant ratio of 0.99 (*p* value of 0.27). For in-frame indel variants, the variant ratio was 1.07 (*p* value of 0.30). A QQ plot testing rare synonymous and in-frame indel variant rates for individual genes showed no significant deviation between the observed and expected *p* values, suggesting no significant technical bias (Figure S1A).

### 3.3. Association of Rare Variants in Known CRC Risk Genes

We identified 36 unique predicted deleterious variants in the ASCO or OMIM-defined CRC risk genes carried by 33 European-ancestry cases (26 with CRC and 7 with significant polyps), shown in Table S4. Three CRC cases carry two deleterious variants in different genes. Notably, two individuals had only one heterozygous variant in the *MUTYH* gene, which is an autosomal recessive gene. In individuals who had not progressed to CRC, serrated-type polyps were detected in every case. The average age at diagnosis for CRC cases with at least one deleterious variant in a known CRC risk gene is 39.0 years, indicating a trend towards a younger age, but it is not significantly different from the entire European-ancestry CRC population (41.7 years, *p* value = 0.06).

The gene-based burden test revealed that LoF variants in known CRC TSGs were significantly enriched in the SAYO dataset compared to the SPARK dataset. In ASCO-listed genes, the relative risk (RR) for LoF variants was 7.21 (*p* value of  $5.8 \times 10^{-6}$ ), and in OMIM-listed genes, the RR was 3.16 (*p* value of 0.01) (Table 2), reinforcing the strong link between LoF variants and CRC risk across multiple gene sets.

**Table 2.** Deleterious variant burden in 21 and 42 CRC-associated genes from ASCO and OMIM.

| Gene Set | Variant Type                         | # of Variants per Individual in SAYO | # of Variants per Individual in SPARK | Relative Risk | <i>p</i> Value       |
|----------|--------------------------------------|--------------------------------------|---------------------------------------|---------------|----------------------|
| ASCO     | LoF                                  | 0.04                                 | 0.01                                  | 7.21          | $5.8 \times 10^{-6}$ |
|          | Missense with VEST4 $\geq 0.5$       | 0.06                                 | 0.05                                  | 1.05          | 0.77                 |
|          | Missense with VEST4 $\geq 0.5$ + LoF | 0.10                                 | 0.06                                  | 1.66          | 0.03                 |
| OMIM     | LoF                                  | 0.03                                 | 0.01                                  | 3.16          | 0.01                 |
|          | Missense with VEST4 $\geq 0.5$       | 0.09                                 | 0.09                                  | 1.00          | 1.00                 |
|          | Missense with VEST4 $\geq 0.5$ + LoF | 0.11                                 | 0.10                                  | 1.20          | 0.38                 |

# means number.

When combining LoF variants with missense variants predicted deleterious by VEST4 ( $\geq 0.5$ ), the RR decreased (ASCO: 1.66, *p* value of 0.03; OMIM: 1.20, *p* value of 0.38). Notably, missense variants alone showed no significant enrichment in OMIM genes (RR = 1.00, *p* value = 1.00). This suggests that most known CRC susceptibility genes were driven by large-effect-size LoF variants, and the contribution from the missense variants is minimal. Furthermore, the limited representation of oncogenes in the ASCO dataset (*n* = 1) likely contributed to the low RR observed for missense variants with VEST4  $\geq 0.5$  (*p* value of 0.77), reflecting a potential bias toward TSG content. Collectively, these findings, particularly in detecting LoF variant enrichment among CRC TSGs, underscore the effectiveness of our burden analysis approach.

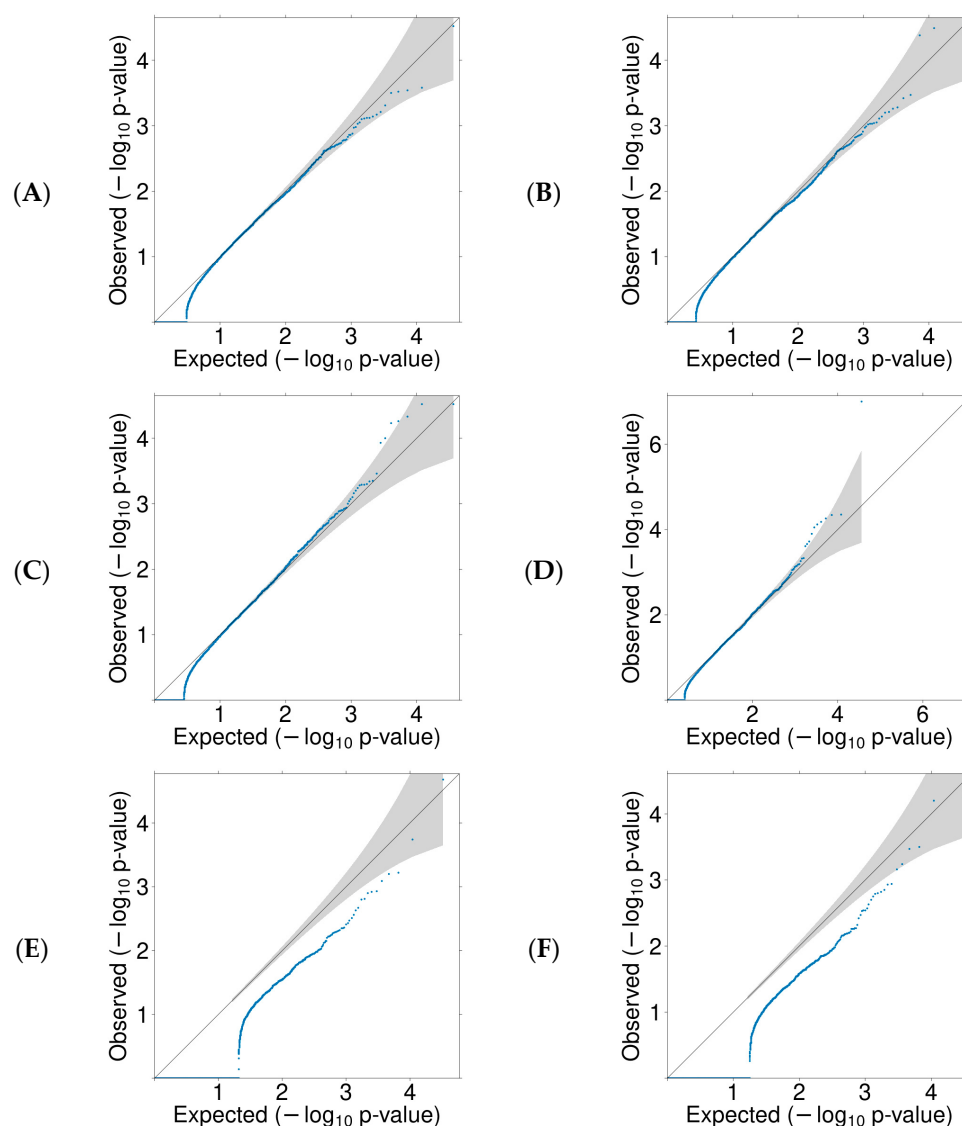
### 3.4. Novel Individual CRC-Risk Gene Discovery

We conducted a rare variant burden analysis to discover novel susceptibility genes for EOCRC based on different mechanisms, including oncogenes (gain-of-function (GoF) variants) and TSGs (LoF variants). Given that missense variants can act as either GoF or LoF, we analyzed the variants in different groups: missense only, LoF only, and combined missense and LoF. VEST4 thresholds defined deleterious missense variants in individual genes. One set of analyses is based on the entire SAYO dataset (comprising both CRC and polyp cases), and the other set is based on the confirmed CRC cases. Figure 2 presents QQ plots for the burden test for all six analyses. The diagonal line represents the null hypothesis of uniformly distributed *p* values, while genes above the diagonal line indicate an enriched burden of deleterious variants in cases compared with controls.

Using a Bonferroni-corrected significance threshold (*p* value of  $2.7 \times 10^{-6}$ ) and a borderline significant threshold (*p* value of  $5.5 \times 10^{-5}$ ), seven genes showed statistically significant or borderline significant associations in at least one test: *MEIKIN*, *STK25*, *PGBD4*, *DIRAS3*, *ATG3*, *RPS6KA4*, and *DDX42*. *MEIKIN* was consistently associated with CRC across most tests and cohort subsets. Additional genes showed variant-type- or cohort-specific potential associations, as summarized below (Table 3). All individuals who had not progressed to CRC have serrated-type polyps.

*MEIKIN* showed significant enrichment in all variant-based tests, including combined (*p* value of  $1.0 \times 10^{-7}$ ), missense variants-only (*p* value of  $4.2 \times 10^{-5}$ ), and LoF-only (*p* value of  $6.3 \times 10^{-5}$ ) analyses. The Bonferroni method identified *MEIKIN* as the only significant risk gene. Four individuals in the SAYO cohort harbored predicted deleterious variants in *MEIKIN* (2 missense variants, 2 LoFs). Two were CRC cases, and the other two had significant polyps. Three of them—one CRC case and two with significant polyps—had sessile serrated adenomas. The two individuals with significant polyps had a family history of CRC or polyps; one CRC patient reported a family history of Melanoma. The phenotypic heterogeneity may explain why *MEIKIN* has been overlooked in previous genetic studies. One CRC case also carried a frameshift variant in a known CRC risk gene, *POLE* (detailed

in Table S5). *POLE*-associated CRC risk is primarily conferred by pathogenic missense variants; thus, the contribution of this LoF variant remains uncertain.



**Figure 2.** QQ plots for missense and/or LoF variants burden test. Each blue dot represents a gene, and the 95% confidence intervals are shaded. (A) missense in CRC, (B) missense in CRC and significant polyps, (C) missense and LoF in CRC, (D) missense and LoF in CRC and significant polyps, (E) LoF in CRC, (F) LoF in CRC and significant polyps.

**Table 3.** Top EOCRC risk genes identified in case–control burden analysis.

| Gene           | # of Variants per Individual in Cases | # of Variants per Individual in Controls | Relative Risk | VEST4 Threshold | <i>p</i> Value       | Most Significant Analysis          |
|----------------|---------------------------------------|------------------------------------------|---------------|-----------------|----------------------|------------------------------------|
| <i>MEIKIN</i>  | 0.019                                 | 0                                        | 150.5         | 0.11            | $1.0 \times 10^{-7}$ | LoF and missense in CRC and polyps |
| <i>STK25</i>   | 0.017                                 | $2.2 \times 10^{-4}$                     | 55.0          | 1.00            | $2.1 \times 10^{-5}$ | LoF in CRC                         |
| <i>PGBD4</i>   | 0.019                                 | $5.4 \times 10^{-4}$                     | 28.7          | 0.23            | $2.9 \times 10^{-5}$ | missense in CRC and polyps         |
| <i>DIRAS3</i>  | 0.017                                 | $1.9 \times 10^{-4}$                     | 61.1          | 0.94            | $3.0 \times 10^{-5}$ | LoF and missense in CRC            |
| <i>ATG3</i>    | 0.019                                 | $6.0 \times 10^{-4}$                     | 26.2          | 0.51            | $3.2 \times 10^{-5}$ | missense in CRC and polyps         |
| <i>RPS6KA4</i> | 0.017                                 | $1.9 \times 10^{-4}$                     | 61.1          | 0.86            | $4.7 \times 10^{-5}$ | LoF and missense in CRC            |
| <i>DDX42</i>   | 0.017                                 | $1.9 \times 10^{-4}$                     | 61.1          | 0.88            | $5.4 \times 10^{-5}$ | LoF and missense in CRC            |

# means number.

*STK25* showed an association with LoF variants, suggesting its potential role as a TSG in EOCRC. We identified three LoF variants in CRC cases (Table S5, Figure 3). Notably, one CRC case also carried a predicted deleterious missense variant in *MLH3*, a gene listed among known CRC-associated genes in the OMIM database but not in the ASCO guidelines. Based on recent ClinGen expert panel reviews, *MLH3* currently has only limited evidence supporting its role in hereditary CRC predisposition. Therefore, while its contribution cannot be ruled out entirely, the *STK25* LoF variants could potentially explain the germline risk factor in this case. Two individuals had first-degree family histories of CRC and/or polyps, and all three had family histories of other cancer types.

*PGBD4* was identified as a potential oncogene based on its association with missense variants only. Four deleterious missense variants and no LoF were identified in the SAYO dataset, while both missense and LoF were found in SPARK. Two individuals with *PGBD4* variants were diagnosed with CRC. Table S5 provides detailed information on individual variants. Three already have one deleterious missense variant in known CRC risk genes—*MCC*, *MSH6*, and *FGFR3*. These three genes are primarily implicated in CRC through somatic variants in tumors or Lynch syndrome, rather than germline missense variants.

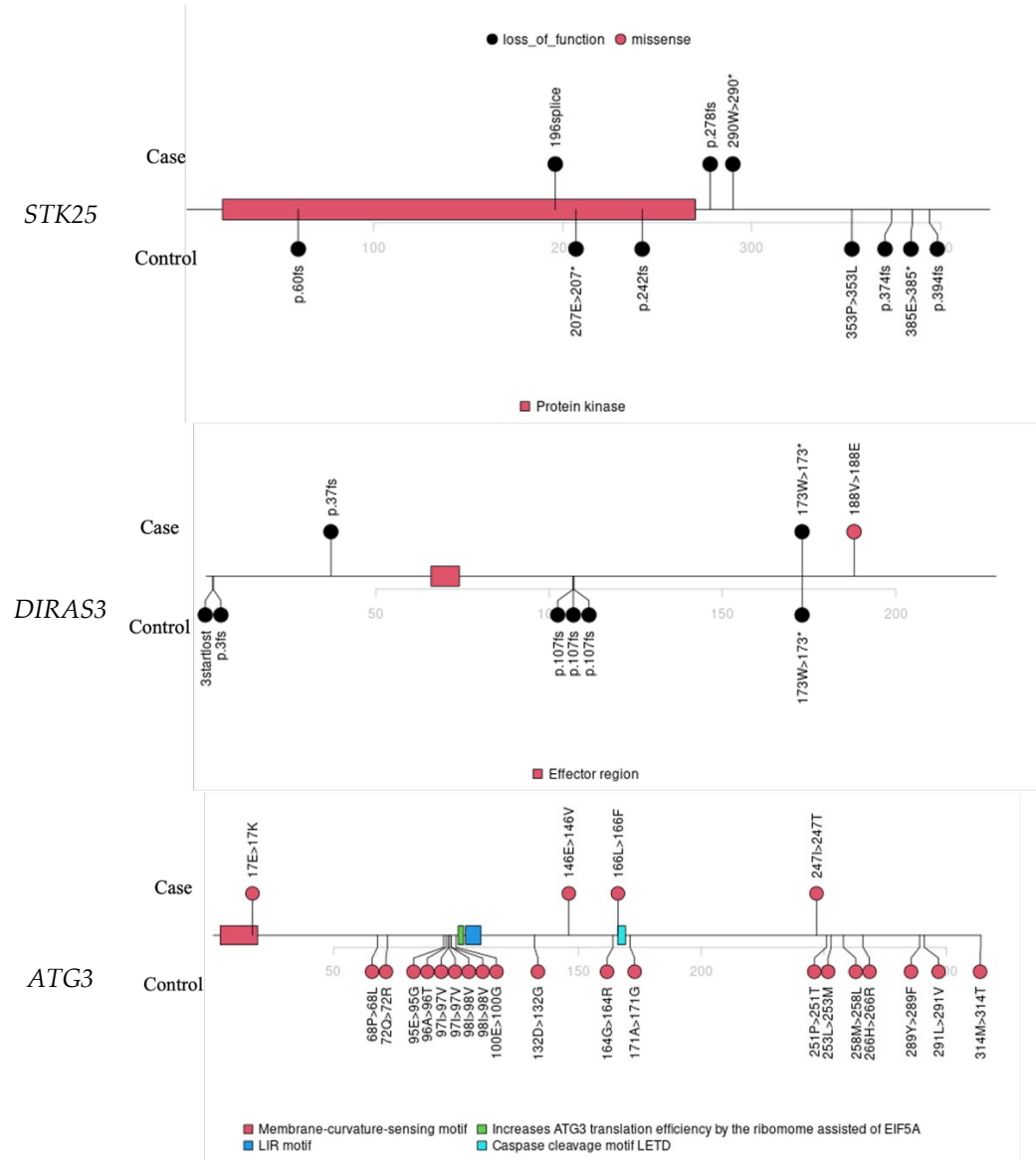
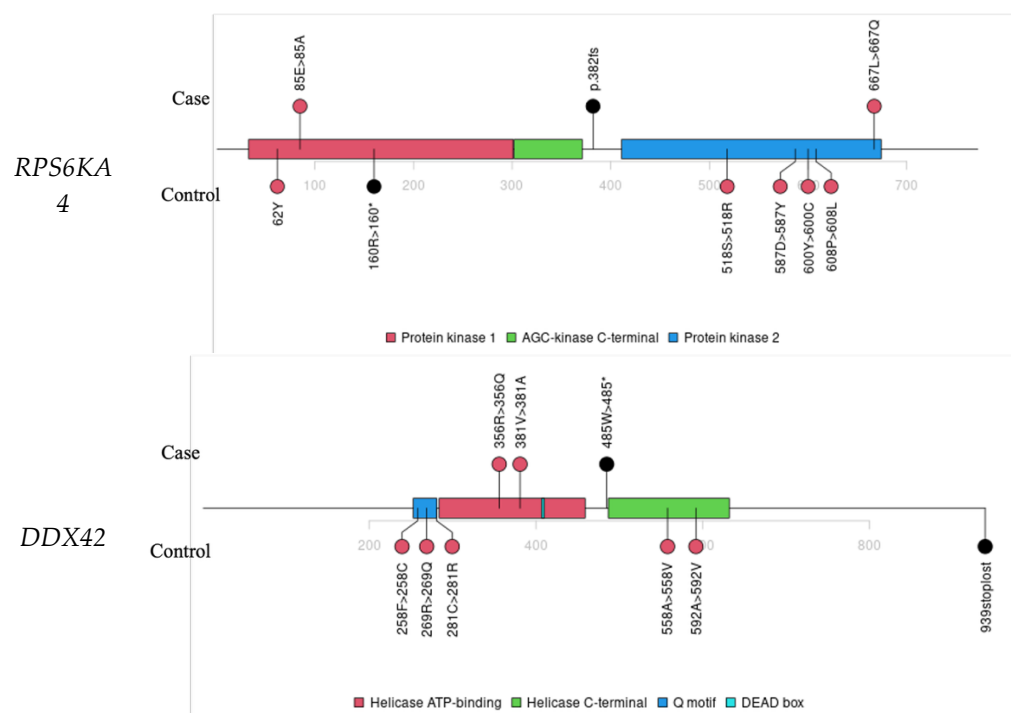


Figure 3. Cont.



**Figure 3.** Lollipop plots for variants in *STK25*, *DIRAS3*, *ATG3*, *RPS6KA4*, and *DDX42*.

Other notable findings include *DIRAS3*. We identified two LoF variants and one deleterious missense variant in *DIRAS3* in the SAYO dataset (detailed in Table S5), suggesting a TSG, consistent with previous findings [41,42]. All three individuals carrying deleterious *DIRAS3* variants were diagnosed with CRC. The *DIRAS3* signal in controls is driven by a recurrent missense variant, as shown in Figure 3, which typically suggests a benign effect. The association signal could be further improved with more accurate predictors of missense deleteriousness.

*ATG3* had four deleterious missense variants and no LoF variants in the SAYO dataset, while both variant types were identified in SPARK, so it is identified as a potential oncogene. Two individuals with the *ATG3* deleterious variant were diagnosed with CRC; the other two had significant polyps (Table S5). Three had family histories of CRC and/or significant polyps, and all four had family histories of non-CRC cancer. Four *ATG3* missense variants were identified in controls, primarily localized to two protein regions (Figure 3). This highlights the importance of predicting the functional effects of variants in novel risk gene discovery.

*RPS6KA4* was associated with LoF variants and is a potential TSG. We identified one LoF and two deleterious missense variants in the SAYO dataset. All three individuals are confirmed CRC cases (Table S5). One CRC case also has a deleterious missense variant in *POLE*, a TSG [43] for CRC [33]. In Figure 3, the *RPS6KA4* variants in cases and controls do not overlap, suggesting that real pathogenic variants may localize in specific protein regions.

*DDX42* showed significant enrichment of risk variants in CRC cases (Table S5), with a relative risk of 61.1 and a  $p$  value of  $5.4 \times 10^{-5}$  at a VEST4 threshold of 0.88, as shown in Table 3. In Figure 3, all three variants in cases cluster within biologically critical regions: two missense variants fall in the Helicase ATP-binding domain, and one stop-gained variant lies near the Helicase C-terminal domain, where truncation is likely disruptive.

The 17 CRC cases carrying deleterious variants in our seven candidate risk genes had an average age of diagnosis of 41.9 years, which was not significantly different from the overall CRC cohort ( $p$  value of 0.92), suggesting that these variants may contribute to CRC

risk without markedly altering age of onset. However, four individuals with an additional deleterious variant in a known CRC risk gene were diagnosed at a significantly younger average age of 31.5 years ( $p$  value of 0.04), suggesting a possible additive genetic effect. A family history of CRC or significant polyps was reported in 11 of the 17 CRC cases. Of the 17 CRC cases, 7 had a family history of CRC, and 14 had a family history of any cancer. A statistical comparison using a binomial test revealed no significant differences between the entire CRC cohort ( $p$  values of 0.6 and 1.0). Notably, although these variants were identified in individuals of European ancestry, no carriers were observed among the non-European cases in SAYO ( $n = 25$ ).

We further examined tumor MMR immunohistochemistry (IHC) status to explore the underlying pathways in these carriers, which was available for 14 of the 17 cases. Among these, 12 tumors were MMR-proficient, showing intact expression of *MLH1*, *PMS2*, *MSH2*, and *MSH6*. Two cases exhibited MMR deficiency: one showed loss of *MLH1* and was clinically diagnosed with Lynch syndrome (germline *MLH1*: T117M), while the other (a 53-year-old woman) had concurrent loss of *MLH1* and *PMS2* without detectable germline variants in either gene. A somatic pathogenic *BRAF* V600E variant was identified in this woman, suggesting *MLH1* promoter hypermethylation [44], although methylation testing was not available to confirm this. These findings indicate that most variant heterozygotes had MMR-proficient tumors, supporting the involvement of non-MMR pathways in EOCRC susceptibility in this cohort.

### 3.5. Validation of *MEIKIN* in the UKBB WGS Cohort

We extended our analysis to the UKBB WGS dataset to evaluate *MEIKIN* effects in an independent population-scale cohort. The UKBB includes 502,242 participants, of whom 472,462 remained after restricting the analysis to individuals of European ancestry. Cancer phenotypes were defined using ICD-10 codes, with all ICD-9 codes converted to their corresponding ICD-10 equivalents: proximal colon cancer ( $n = 2816$ ), distal colon cancer ( $n = 2557$ ), rectosigmoid junction cancer ( $n = 619$ ), and rectal cancer ( $n = 2596$ ). The average age at diagnosis was 66.51, 63.98, 62.12, and 63.38 years for proximal colon, distal colon, rectosigmoid junction, and rectal cancer, respectively. Under-55 cases comprise 9.7%, 15.9%, 21.9%, and 18.7% of each group (Table S6).

Following application of the same rare variant filters as in the SAYO cohorts, a total of 127 unique rare coding variants were retained in *MEIKIN*. *MEIKIN* LoF variants were enriched in distal colon cases, demonstrating a relative risk of 2.46 and a  $p$  value of 0.086 in Table S7. Applying the deleteriousness threshold identified in the SAYO dataset ( $VEST4 \geq 0.109$ ) and the variable-threshold approach ( $VEST4 \geq 0.24$ ), missense variants were insignificantly enriched in distal colon cases. When combining missense and LoF variants, the variable-threshold approach identified a borderline significance with the relative risk of 1.98 and a  $p$  value of 0.048, as shown in Table S7. Analyzing distal colon cases in individuals younger than 55 years, the variable-threshold approach revealed an insignificant enrichment, with a relative risk of 8.23 and a  $p$  value of 0.217. No significant enrichment was observed in proximal colon cancer. Although full replication of the *MEIKIN* association across colorectal subtypes was not achieved, these results suggest that *MEIKIN* missense and LoF variants are potentially associated with distal colon cancer.

## 4. Discussion

Our gene-based rare-variant analysis identified one novel candidate risk gene, *MEIKIN*. It is a kinetochore protein and plays a crucial role in meiosis [45–47]. In addition to the testis, *MEIKIN* is not highly expressed in any other somatic tissues. However, ectopic activation of meiotic genes has been increasingly recognized as a contributor to oncogenesis, disrupting

normal cellular processes, chromosome cohesion, kinetochore function, and segregation fidelity, contributing to cancer hallmarks [48–52]. Based on STRING analysis [53], *MEIKIN* is implicated in chromosome segregation and cell cycle regulation [45], through interactions with key proteins such as SGO1 [54], SGO2 [54], PLK1 [55], and ESPL1 [56], all of which are known to play critical roles in cancer-associated pathways (Figure S2). Moreover, we demonstrated the enriched *MEIKIN* missense and LoF variants in the distal colon cancer cases in the European-ancestry subset of the UKBB WGS cohort. This replication in an independent, population-scale dataset reinforces the robustness of our analytic framework and provides additional evidence that rare LoF variation in meiotic regulators may contribute to EOCRC susceptibility. Notably, our SAYO datasets defined case status based on clinical diagnoses of CRC, whereas the UKBB relies on ICD billing codes. The UKBB primarily consists of older adults; analyses focused on early-onset diseases were limited. Differences in disease ascertainment and age composition in cases between the two cohorts may partially explain the lack of statistical significance. Nonetheless, our replication in the distal colon cancer cases aligns with the fact that EOCRC is more likely to occur in the distal colon.

Our study extends previous work by identifying additional candidate genes with diverse biological functions that may contribute to EOCRC susceptibility [57,58], including *STK25*, *PGBD4*, *DIRAS3*, *ATG3*, *RPS6KA4*, and *DDX42*. *STK25* is critical in CRC progression by regulating autophagy, metabolism, and epithelial–mesenchymal transition (EMT) [59–61]. Its downregulation enhances autophagy via the *JAK2/STAT3* pathway, while higher *STK25* levels are associated with a better prognosis [59]. *STK25* also inhibits CRC cell proliferation and glycolysis through interaction with *GOLPH3* and modulation of the mTOR pathway [60]. Additionally, *STK25* promotes EMT by interacting with *LIMK1*, contributing to increased invasion and poorer outcomes [61]. The enrichment of LoF variants in *STK25* among cases supports its putative role as a CRC tumor suppressor. While little is known about *PGBD4*, other members of the PGBD gene family have been more extensively characterized. Notably, *PGBD5* has been shown to drive site-specific oncogenic mutations in human tumors, implicating it in tumorigenesis [62]. Although the biological role of *PGBD4* itself is undercharacterized, its enrichment signal in EOCRC and its close homology with the oncogenic *PGBD5* gene suggest potential functional relevance. This highlights the diverse and impactful roles that PGBD family genes may play in the development and progression of cancer. *DIRAS3* plays a crucial role in inhibiting RAS/MAPK signaling, which is often dysregulated in many cancers. *ATG3* was upregulated in CRC tissues and cell lines compared to normal counterparts, with experimental evidence showing that *ATG3* knockdown significantly suppresses cancer cell proliferation and invasion [63]. The oncogenic effects of *ATG3* are primarily mediated through an autophagy-dependent pathway, evidenced by the counteractive effects of autophagy blockade on cancer progression [63]. Additionally, the interaction of *ATG3* with *ATG12* suggests a broader role in the autophagy machinery [64,65]. We conducted AlphaFold 3D structural modeling of the *ATG3* missense variants. Two variants in solvent-exposed regions may affect protein interactions, and the other two in buried positions may affect protein folding or local stability. *RPS6KA4* is involved in tumor suppression through epigenetic regulation and interaction with key signaling pathways [66]. Furthermore, high-throughput LoF screening has identified *RPS6KA4* as a possible regulator of p53 activity, reinforcing its involvement in key cancer-associated signaling pathways [67]. *DDX42* encodes a DEAD-box RNA helicase with RNA-chaperone activity that participates in pre-mRNA splicing, RNA remodeling, and maintenance of genome stability [68]. It interacts with *ASPP2*, a co-activator of p53, thereby modulating apoptosis, and has been identified as a PARP1 interactor, further linking *DDX42* to DNA damage response [69]. *DDX42* also functions as an intrinsic inhibitor of retroviruses and

LINE-1 retrotransposons, restricting aberrant nucleic acid species that can promote genomic instability [70]. In contrast, overexpression of *DDX42* in hepatocellular carcinoma activates the PI3K/AKT signaling pathway, enhancing proliferation, radioresistance, and sorafenib resistance, suggesting context-dependent oncogenic activity [71].

Our study has several strengths. Our analytic approach was validated by the detection of established associations with LoF variants in known TSGs, confirming its robustness. The identification of *ATG3*, *STK25*, and *DIRAS3* aligns with prior research implicating autophagy and tumor suppression mechanisms in CRC development. The use of stringent quality control metrics, a tailored burden test, and 10 million permutations provided high confidence in the robustness of our association findings. Prioritizing predicted deleterious variants further enriched our results for potentially functional contributors to disease risk. The identification of additional candidate genes with diverse biological functions expands the current spectrum of CRC predisposition loci, opening new avenues for biological and clinical investigation.

Nevertheless, certain limitations should be acknowledged. First, our results provide statistical evidence rather than definitive evidence of causality. Functional studies are crucial for validating the impact of these variants. Future studies should focus on experimentally characterizing the candidate genes through gene editing, cellular modeling, and transcriptomic profiling to elucidate their roles in the initiation and progression of CRC. Second, while our sample size was sufficient to detect strong associations, larger and more diverse cohorts will be needed to identify lower-penetrance risk genes and enhance generalizability across different ancestries. Finally, small-effect contributors, including common variants [72–74], lifestyle factors, and environmental exposures, were not assessed within our rare-variant association tests.

Clinically, the discovery of novel genetic risk factors could enhance early detection and personalized screening for individuals and families at elevated risk of EOCRC. If validated, these genes may be incorporated into multi-gene panels, improving risk stratification beyond known CRC susceptibility genes. Moreover, insights into the molecular pathways affected by these genes, particularly those involving autophagy, metabolism, and immune signaling, could inform the development of new therapeutic targets. As EOCRC incidence continues to rise globally, understanding the contribution of rare germline variation will be essential to refining both preventive and treatment strategies.

In summary, our findings broaden the landscape of EOCRC-associated genes and highlight the potential of rare variant burden analysis to uncover biologically and clinically relevant risk loci. Continued functional validation and integration with clinical phenotypes will be key to translating these discoveries into precision oncology for EOCRC.

## 5. Conclusions

Using exome sequencing and rare variant burden testing in a well-curated EOCRC cohort, we confirmed significant enrichment of LoF variants in established tumor suppressor genes and identified *MEIKIN* as a novel candidate susceptibility gene, with supportive evidence in an independent population dataset for distal colon cancer. We also identify *STK25*, *PGBD4*, *DIRAS3*, *ATG3*, *RPS6KA4*, and *DDX42* as additional candidates with variant-class-specific signals. Collectively, these findings underscore a meaningful contribution of rare germline variants—both LoF and GoF—to the genetic architecture of EOCRC.

These results expand the landscape of heritable CRC risk and suggest practical avenues for improving genetic risk assessment. Pending replication and functional validation, the genes highlighted here could inform panel design, enable earlier identification of high-risk individuals, and guide mechanistic studies of pathways such as kinetochore biology, autophagy, and immune signaling. Larger, ancestrally diverse cohorts and experimental

follow-up will be essential to translate these discoveries into precision screening and prevention strategies.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers17243931/s1>, Figure S1: Post quality control for rare synonymous variants and inframe indels in SAYO and SPARK; Figure S2: Protein-Protein Interaction Network of MEIKIN-Associated Genes; Table S1: Quality control thresholds for variant filtering in SAYO and SPARK cohorts; Table S2: 21 colorectal cancer risk genes based on the ASCO guidelines; Table S3: 42 colorectal cancer risk genes based on the OMIM guidelines; Table S4: Individual variants information for known CRC risk genes; Table S5: Individual variants information for novel CRC risk-susceptible genes; Table S6: Burden test results for *MEIKIN* in the UKBB WGS cohort using a fixed VEST4 threshold across cancer types; Table S7: Burden test results for *MEIKIN* in the UKBB WGS cohort using variable VEST4 thresholds across cancer types.

**Author Contributions:** R.S.: Formal Analysis, Visualization, Writing—Original Draft Preparation. R.R.M.: Conceptualization, Investigation, Methodology, Project Administration, Writing—Original Draft Preparation. Z.H.: Formal Analysis. M.H.: Sample collection, Data curation, Writing—Review and Editing. W.U.: Data curation, Software, Writing—Review and Editing. W.M.: Data acquisition. N.K.P.: Resources, Writing—Review and Editing. B.W.: Resources, Writing—Review and Editing. Y.L.: Formal analysis, Writing—Review and Editing. J.F.: Bioinformatics, Writing—Review and Editing. H.S.S.: Resources, Writing—Review and Editing. Y.S.: Writing—Review and Editing. C.W.: Writing—Review and Editing. R.Y.: Writing—Review and Editing. Y.D.: Writing—Review and Editing. X.L.: Resources, Writing—Review and Editing. W.K.C.: Resources, Writing—Review and Editing. E.S.: Supervision, Writing—Review and Editing. T.J.P.: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing—Review and Editing. J.P.Y.: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing—Review and Editing. X.F.: Formal Analysis, Investigation, Methodology, Supervision, Funding Acquisition, Writing—Original Draft Preparation. All authors have read and agreed to the published version of the manuscript.

**Funding:** X.F. is supported for this work by the National Institutes of Health (grant number R00HG011490). T.J.P. and J.P.Y. are supported by Cancer Council SA [grant number 1138776]. R.R.M. was supported by the University of Adelaide International Scholarship and the Hans-Jürgen & Marianne Ohff Research Grant. Y.D. is supported for this work by the National Institutes of Health (grant number U01CA265719).

**Institutional Review Board Statement:** The request to use the SPARK dataset for our study was approved by the WCG IRB. SPARK is actually a publicly available dataset that we have applied to. We do not work with their IRB team. Their review team approved our request to use their data and provided deidentified data on their platform.

**Informed Consent Statement:** Written informed consent was obtained from all subjects in the SAYO study.

**Data Availability Statement:** Exome sequencing data from the SAYO study will be made available upon request.

**Acknowledgments:** We thank the participants in the South Australian Young Onset Colorectal Polyp and Cancer (SAYO) study and the Simons Foundation Powering Autism Research for Knowledge (SPARK) for their contributions. We are grateful to all of the families in SPARK, the SPARK clinical sites, and the SPARK staff. We appreciate obtaining access to the genetic data on SFARI Base. Approved researchers can obtain the SPARK population dataset described in this study by applying at <https://base.sfari.org>. We appreciate obtaining access to recruit participants through SPARK research match on the SFARI Base. The data was accessed on 23 May 2024.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2024**, *74*, 229–263. [\[CrossRef\]](#) [\[PubMed\]](#)
- Siegel, R.L.; Miller, K.D.; Fedewa, S.A.; Ahnen, D.J.; Meester, R.G.S.; Barzi, A.; Jemal, A. Colorectal cancer statistics, 2017. *CA A Cancer J. Clin.* **2017**, *67*, 177–193.
- Mikaeel, R.R.; Price, T.J.; Smith, E.; Drew, P.A.; Uylaki, W.; Horsnell, M.; Young, J.P. Colorectal cancer in Australian young adults. *Mathews J. Cancer Sci.* **2019**, *4*, 18. [\[CrossRef\]](#)
- Alvarez, M.D.; Quintana, I.; Terradas, M.; Mur, P.; Balaguer, F.; Valle, L. The inherited and familial component of early-onset colorectal cancer. *Cells* **2021**, *10*, 710. [\[CrossRef\]](#)
- Ivarsdottir, E.V.; Gudmundsson, J.; Tragante, V.; Sveinbjornsson, G.; Kristmundsdottir, S.; Stacey, S.N.; Halldorsson, G.H.; Magnusson, M.I.; Oddsson, A.; Walters, G.B.; et al. Gene-based burden tests of rare germline variants identify six cancer susceptibility genes. *Nat. Genet.* **2024**, *56*, 2422–2433. [\[CrossRef\]](#)
- Mikaeel, R.R.; Symonds, E.L.; Kimber, J.; Smith, E.; Horsnell, M.; Uylaki, W.; Rico, G.T.; Hewett, P.J.; Yong, J.; Tonkin, D.; et al. Young-onset colorectal cancer is associated with a personal history of type 2 diabetes. *Asia-Pac. J. Clin. Oncol.* **2021**, *17*, 131–138.
- Kimber, J.; Symonds, E.; Uylaki, W.; Horsnell, M.; Drew, P.; Smith, E.; Mikaeel, R.; Hardingham, J.; Tomita, Y.; Jesudason, D.; et al. Exploring the associations between colorectal polyps and type 2 diabetes mellitus in a colonoscopy clinic population. *ESMO Gastrointest. Oncol.* **2024**, *4*, 100053. [\[CrossRef\]](#)
- Leggett, B.; Whitehall, V. Role of the serrated pathway in colorectal cancer pathogenesis. *Gastroenterology* **2010**, *13*, 2088–2100. [\[CrossRef\]](#)
- Win, A.K.; Jenkins, M.A.; Dowty, J.G.; Antoniou, A.C.; Lee, A.; Giles, G.G.; Buchanan, D.D.; Clendenning, M.; Rosty, C.; Ahnen, D.J.; et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. *Cancer Epidemiol. Biomark. Prev.* **2017**, *26*, 404–412.
- Molmenti, C.L.; Kolb, J.M.; Karlitz, J.J. Advanced colorectal polyps on colonoscopy: A trigger for earlier screening of family members. *Am. J. Gastroenterol.* **2020**, *115*, 311–314. [\[CrossRef\]](#)
- Mikaeel, R.R.; Young, J.P.; Li, Y.; Smith, E.; Horsnell, M.; Uylaki, W.; Rico, G.T.; Poplawski, N.K.; Hardingham, J.E.; Tomita, Y.; et al. Survey of germline variants in cancer-associated genes in young adults with colorectal cancer. *Genes Chromosomes Cancer* **2022**, *61*, 105–113. [\[CrossRef\]](#) [\[PubMed\]](#)
- Feliciano, P.; Daniels, A.M.; Snyder, L.G.; Beaumont, A.; Camba, A.; Esler, A.; Gulsrud, A.G.; Mason, A.; Gutierrez, A.; Nicholson, A.; et al. SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron* **2018**, *97*, 488–493. [\[CrossRef\]](#) [\[PubMed\]](#)
- Manichaikul, A.; Mychaleckyj, J.C.; Rich, S.S.; Daly, K.; Sale, M.; Chen, W.-M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **2010**, *26*, 2867–2873. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pedersen, B.S.; Quinlan, A.R. Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *Am. J. Hum. Genet.* **2017**, *100*, 406–413. [\[CrossRef\]](#)
- Van Kets, V.; Kitzman, J.; Snyder, M.; Shendure, J.; Gray, P. Kapa Hyper Prep: A next-generation kit for fast and efficient library construction from challenging DNA samples. In Proceedings of the Advances in Genome Biology and Technology (AGBT) Meeting, Marco Island, FL, USA, 12–15 February 2014.
- Paijmans, J.L.A.; Baleka, S.; Henneberger, K.; Taron, U.H.; Trinks, A.; Westbury, M.V.; Barlow, A. Sequencing single-stranded libraries on the Illumina NextSeq 500 platform. *arXiv* **2017**, arXiv:1711.11004.
- Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [\[CrossRef\]](#)
- Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33.
- DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; Del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [\[CrossRef\]](#)
- McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. *Genome Biol.* **2016**, *17*, 122. [\[CrossRef\]](#)
- Flicek, P.; Amode, M.R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; et al. Ensembl 2014. *Nucleic Acids Res.* **2014**, *42*, D749–D755.
- Hunt, S.E.; McLaren, W.; Gil, L.; Thormann, A.; Schuilenburg, H.; Sheppard, D.; Parton, A.; Armean, I.M.; Trevanion, S.J.; Flicek, P.; et al. Ensembl variation resources. *Database* **2018**, *2018*, bay119. [\[CrossRef\]](#)
- Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; et al. Ensembl 2020. *Nucleic Acids Res.* **2020**, *48*, D682–D688. [\[CrossRef\]](#)

24. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alfoldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*, 434–443. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **2015**, *12*, e1001779. [\[CrossRef\]](#) [\[PubMed\]](#)
26. MacArthur, D.G.; Balasubramanian, S.; Frankish, A.; Huang, N.; Morris, J.; Walter, K.; Jostins, L.; Habegger, L.; Pickrell, J.K.; Montgomery, S.B.; et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **2012**, *335*, 823–828. [\[PubMed\]](#)
27. Carter, H.; Douville, C.; Stenson, P.D.; Cooper, D.N.; Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genom.* **2013**, *14*, S3. [\[CrossRef\]](#)
28. Sofer, T.; Zheng, X.; Laurie, C.A.; Gogarten, S.M.; Brody, J.A.; Conomos, M.P.; Bis, J.C.; Thornton, T.A.; Szpiro, A.; O'Connell, J.R.; et al. Variant-specific inflation factors for assessing population stratification at the phenotypic variance level. *Nat. Commun.* **2021**, *12*, 3506. [\[CrossRef\]](#)
29. Wang, G.T.; Peng, B.; Leal, S.M. Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. *Am. J. Hum. Genet.* **2014**, *94*, 770–783. [\[CrossRef\]](#)
30. Douville, C.; Masica, D.L.; Stenson, P.D.; Cooper, D.N.; Gyax, D.M.; Kim, R.; Ryan, M.; Karchin, R. Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel). *Hum. Mutat.* **2016**, *37*, 28–35.
31. Baeissa, H.M.; Benstead-Hume, G.; Richardson, C.J.; Pearl, F.M. Mutational patterns in oncogenes and tumour suppressors. *Biochem. Soc. Trans.* **2016**, *44*, 925–931. [\[CrossRef\]](#)
32. Tung, N.; Ricker, C.; Messersmith, H.; Balmaña, J.; Domchek, S.; Stoffel, E.M.; Almhanna, K.; Arun, B.; Chavarri-Guerra, Y.; Cohen, S.A.; et al. Selection of germline genetic testing panels in patients with cancer: ASCO guideline. *J. Clin. Oncol.* **2024**, *42*, 2599–2615. [\[CrossRef\]](#)
33. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; Valle, D.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2005**, *33* (Suppl. 1), D514–D517. [\[PubMed\]](#)
34. Lee, S.; Abecasis, G.R.; Boehnke, M.; Lin, X. Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **2014**, *95*, 5–23. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Zhu, N.; Swietlik, E.M.; Welch, C.L.; Pauciulo, M.W.; Hagen, J.J.; Zhou, X.; Guo, Y.; Karten, J.; Pandya, D.; Tilly, T.; et al. Rare variant analysis of 4241 pulmonary arterial hypertension cases from an international consortium implicates FBLN2, PDGFD, and rare de novo variants in PAH. *Genome Med.* **2021**, *13*, 80. [\[PubMed\]](#)
36. Zhu, N.; LeDuc, C.A.; Fennoy, I.; Laferrère, B.; Doege, C.A.; Shen, Y.; Chung, W.K.; Leibel, R.L. Rare predicted loss of function alleles in Bassoon (BSN) are associated with obesity. *npj Genom. Med.* **2023**, *8*, 33. [\[CrossRef\]](#)
37. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699. [\[CrossRef\]](#)
38. The UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* **2025**, *645*, 692. [\[CrossRef\]](#)
39. World Health Organization. International Statistical Classification of Diseases and Related Health Problems: 10th Revision (ICD-10). 1992. Available online: <https://icd.who.int/browse10/2019/en> (accessed on 5 December 2025).
40. Gudmundsson, S.; Singer-Berk, M.; Watts, N.A.; Phu, W.; Goodrich, J.K.; Solomonson, M.; Genome Aggregation Database Consortium; Rehm, H.L.; MacArthur, D.G.; O'Donnell-Luria, A. Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **2022**, *43*, 1012–1030.
41. Bildik, G.; Liang, X.; Sutton, M.N.; Bast, R.C., Jr.; Lu, Z. DIRAS3: An imprinted tumor suppressor gene that regulates RAS and PI3K-driven cancer growth, motility, autophagy, and tumor dormancy. *Mol. Cancer Ther.* **2022**, *21*, 25–37. [\[CrossRef\]](#)
42. Lu, Z.; Yang, H.; Sutton, M.N.; Yang, M.; Clarke, C.H.; Liao, W.S.-L.; Bast, R.C. ARHI (DIRAS3) induces autophagy in ovarian cancer cells by downregulating the epidermal growth factor receptor, inhibiting PI3K and Ras/MAP signaling and activating the FOXo3a-mediated induction of Rab7. *Cell Death Differ.* **2014**, *21*, 1275–1289.
43. Ma, X.; Dong, L.; Liu, X.; Ou, K.; Yang, L. POLE/POLD1 mutation and tumor immunotherapy. *J. Exp. Clin. Cancer Res.* **2022**, *41*, 216. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Weisenberger, D.J.; Siegmund, K.D.; Campan, M.; Young, J.; Long, T.I.; Faasse, M.A.; Kang, G.H.; Widschwendter, M.; Weener, D.; Buchanan, D.; et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet.* **2006**, *38*, 787–793. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Kim, J.; Ishiguro, K.-I.; Nambu, A.; Akiyoshi, B.; Yokobayashi, S.; Kagami, A.; Ishiguro, T.; Pendas, A.M.; Takeda, N.; Sakakibara, Y.; et al. Meikin is a conserved regulator of meiosis-I-specific kinetochore function. *Nature* **2015**, *517*, 466–471. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Maier, N.K.; Ma, J.; Lampson, M.A.; Cheeseman, I.M. Separase cleaves the kinetochore protein Meikin at the meiosis I/II transition. *Dev. Cell* **2021**, *56*, 2192–2206.e8. [\[CrossRef\]](#)

47. Ishiguro, K.I.; Matsuura, K.; Tani, N.; Takeda, N.; Usuki, S.; Yamane, M.; Sugimoto, M.; Fujimura, S.; Hosokawa, M.; Chuma, S.; et al. MEIOSIN directs the switch from mitosis to meiosis in mammalian germ cells. *Dev. Cell* **2020**, *52*, 429–445.e10. [[CrossRef](#)]
48. Sou, I.F.; Hamer, G.; Tee, W.W.; Vader, G.; McClurg, U.L. Cancer and meiotic gene expression: Two sides of the same coin? *Curr. Top. Dev. Biol.* **2023**, *151*, 43–68.
49. Lingg, L.; Rottenberg, S.; Francica, P. Meiotic genes and DNA double strand break repair in cancer. *Front. Genet.* **2022**, *13*, 831620. [[CrossRef](#)]
50. Bruggeman, J.W.; Koster, J.; van Pelt, A.M.M.; Speijer, D.; Hamer, G. How germline genes promote malignancy in cancer cells. *Bioessays* **2023**, *45*, 2200112.
51. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [[CrossRef](#)]
52. McFarlane, R.J.; Wakeman, J.A. Meiosis-like functions in oncogenesis: A new view of cancer. *Cancer Res.* **2017**, *77*, 5712–5716. [[CrossRef](#)]
53. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [[PubMed](#)]
54. Jallepalli, P.V.; Lengauer, C. Chromosome segregation and cancer: Cutting through the mystery. *Nat. Rev. Cancer* **2001**, *1*, 109–117. [[CrossRef](#)] [[PubMed](#)]
55. Strebhardt, K.; Ullrich, A. Targeting polo-like kinase 1 for cancer therapy. *Nat. Rev. Cancer* **2006**, *6*, 321–330. [[CrossRef](#)] [[PubMed](#)]
56. Finetti, P.; Guille, A.; Adelaide, J.; Birnbaum, D.; Chaffanet, M.; Bertucci, F. ESPL1 is a candidate oncogene of luminal B breast cancers. *Breast Cancer Res. Treat.* **2014**, *147*, 51–59. [[CrossRef](#)]
57. Matejcic, M.; Shaban, H.A.; Quintana, M.W.; Schumacher, F.R.; Edlund, C.K.; Naghi, L.; Pai, R.K.; Haile, R.W.; Levine, A.J.; Buchanan, D.D.; et al. Rare variants in the DNA repair pathway and the risk of colorectal cancer. *Cancer Epidemiol. Biomark. Prev.* **2021**, *30*, 895–903.
58. Terradas, M.; Capellá, G.; Valle, L. Dominantly inherited hereditary nonpolyposis colorectal cancer not caused by MMR genes. *J. Clin. Med.* **2020**, *9*, 1954. [[CrossRef](#)]
59. Chen, J.; Gao, P.; Peng, L.; Liu, T.; Wu, F.; Xu, K.; Chen, L.; Tan, F.; Xing, P.; Wang, Z.; et al. Downregulation of STK25 promotes autophagy via the Janus kinase 2/signal transducer and activator of transcription 3 pathway in colorectal cancer. *Mol. Carcinog.* **2022**, *61*, 572–586. [[CrossRef](#)]
60. Wu, F.; Gao, P.; Wu, W.; Wang, Z.; Yang, J.; Di, J.; Jiang, B.; Su, X. STK25-induced inhibition of aerobic glycolysis via GOLPH3-mTOR pathway suppresses cell proliferation in colorectal cancer. *J. Exp. Clin. Cancer Res.* **2018**, *37*, 144.
61. Sun, X.; Li, S.; Lin, H. LIMK1 interacts with STK25 to regulate EMT and promote the proliferation and metastasis of colorectal cancer. *J. Oncol.* **2022**, *2022*, 3963883. [[CrossRef](#)]
62. Henssen, A.G.; Koche, R.; Zhuang, J.; Jiang, E.; Reed, C.; Eisenberg, A.; Still, E.; MacArthur, I.C.; Rodríguez-Fos, E.; Gonzalez, S.; et al. PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat. Genet.* **2017**, *49*, 1005–1014.
63. Huang, W.; Zeng, C.; Hu, S.; Wang, L.; Liu, J. ATG3, a target of miR-431-5p, promotes proliferation and invasion of colon cancer via promoting autophagy. *Cancer Manag. Res.* **2019**, *11*, 10275–10285. [[CrossRef](#)] [[PubMed](#)]
64. Radoshevich, L.; Murrow, L.; Chen, N.; Fernandez, E.; Roy, S.; Fung, C.; Debnath, J. ATG12 conjugation to ATG3 regulates mitochondrial homeostasis and cell death. *Cell* **2010**, *142*, 590–600. [[CrossRef](#)] [[PubMed](#)]
65. Metlagel, Z.; Otomo, C.; Takaesu, G.; Otomo, T. Structural basis of ATG3 recognition by the autophagic ubiquitin-like protein ATG12. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18844–18849. [[CrossRef](#)] [[PubMed](#)]
66. Uhlmann-Schiffler, H.; Jalal, C.; Stahl, H. Ddx42p—A human DEAD box protein with RNA chaperone activities. *Nucleic Acids Res.* **2006**, *34*, 10–22. [[CrossRef](#)]
67. Uhlmann-Schiffler, H.; Kiermayer, S.; Stahl, H. The DEAD box protein Ddx42p modulates the function of ASPP2, a stimulator of apoptosis. *Oncogene* **2009**, *28*, 2065–2073. [[CrossRef](#)]
68. Bonaventure, B.; Rebendenne, A.; Valadão, A.L.C.; Arnaud-Arnould, M.; Gracias, S.; de Gracia, F.G.; McKellar, J.; Labaronne, E.; Tauziet, M.; Vivet-Boudou, V.; et al. The DEAD box RNA helicase DDX42 is an intrinsic inhibitor of positive-strand RNA viruses. *EMBO Rep.* **2022**, *23*, e54061. [[CrossRef](#)]
69. Liu, Z.; Yuan, J.; Liu, F.; Zeng, Q.; Wu, Z.; Yang, J. DDX42 Enhances Hepatocellular Carcinoma Cell Proliferation, Radiation and Sorafenib Resistance via Regulating GRB2 RNA Maturation and Activating PI3K/AKT Pathway. *J. Cell. Mol. Med.* **2025**, *29*, e70793. [[CrossRef](#)]
70. Savage, S.R.; Yi, X.; Lei, J.T.; Wen, B.; Zhao, H.; Liao, Y.; Jaehnig, E.J.; Somes, L.K.; Shafer, P.W.; Lee, T.D.; et al. Pan-cancer proteogenomics expands the landscape of therapeutic targets. *Cell* **2024**, *187*, 4389–4407.e15. [[CrossRef](#)]
71. Llanos, S.; Efeyan, A.; Monsech, J.; Dominguez, O.; Serrano, M. A high-throughput loss-of-function screening identifies novel p53 regulators. *Cell Cycle* **2006**, *5*, 1880–1885.

72. Law, P.J.; Timofeeva, M.; Fernandez-Rozadilla, C.; Broderick, P.; Studd, J.; Fernandez-Tajes, J.; Farrington, S.; Svinti, V.; Palles, C.; Orlando, G.; et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat. Commun.* **2019**, *10*, 2154. [\[CrossRef\]](#)
73. Thomas, M.; Su, Y.-R.; Rosenthal, E.A.; Sakoda, L.C.; Schmit, S.L.; Timofeeva, M.N.; Chen, Z.; Fernandez-Rozadilla, C.; Law, P.J.; Murphy, N.; et al. Combining Asian and European genome-wide association studies of colorectal cancer improves risk prediction across racial and ethnic populations. *Nat. Commun.* **2023**, *14*, 6147. [\[CrossRef\]](#)
74. Laskar, R.; Qu, C.; Huyghe, J.; Harrison, T.; Hayes, R.; Cao, Y.; Campbell, P.; Steinfelder, R.; Talukdar, F.; Brenner, H.; et al. Genome-wide association studies and Mendelian randomization analyses provide insights into the causes of early-onset colorectal cancer. *Ann. Oncol.* **2024**, *35*, 523–536. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.