scientific reports

OPEN

A set cover algorithm identifies minimal circulating tumour DNA sequencing targets for colorectal cancer detection

Kit Moloney-Geany[1,2], Michael A. Black[1], Robert C. Day[1], Parry Guilford[1] & Michael J. Dunnet[1]✉

Advances in high-throughput sequencing and decreasing costs have made cell-free DNA sequencing a promising approach for cancer detection. Sequencing assays require high read depth to detect low-frequency somatic mutations, so cell-free DNA panels must support deep sequencing while still assaying broadly enough to detect as many malignancies as possible. We developed OPTIC (Oncogene Panel Tester for Identifying Cancers), a pipeline employing a set cover algorithm, to identify the minimal set of genomic targets capturing the maximal proportion of tumours. Using three cohorts totalling 2,940 colorectal cancer samples, OPTIC was utilized to design a targeted sequencing panel spanning just 10,975 bases across *APC*, *TP53*, *KRAS*, *BRAF*, *NRAS*, *PIK3CA*, *CTNNB1*, *RNF43*, and *ACVR2A*. Collectively, these loci contain pathogenic mutations in 96.3% of cases. Our pipeline enables compact panel design without compromising sample coverage. This enables higher throughput, greater sequencing depth, and lower costs per-sample in early colorectal cancer detection from cell-free DNA.

Keywords Colorectal cancer, Somatic mutation detection, Sequencing, Gene panel, Coverage, Cell-free DNA

Abbreviations

CRC	Colorectal cancer
CIN	Chromosomal instability
MSI	Microsatellite instability
CIMP	CpG-island methylator phenotype
cfDNA	Cell-free DNA
ctDNA	Circulating tumour DNA
MAF	Mutation annotation format
TCGA	The Cancer Genome Atlas
COAD	Colorectal adenocarcinoma
READ	Rectal adenocarcinoma
COSMIC	Catalogue of somatic mutations in cancer
SNV	Single nucleotide variation

Although the overall incidence of colorectal cancer (CRC) in older adults has decreased in high-income countries, it is rising in developing nations. Additionally, the rate of early-onset disease (≤ 50 years of age) is increasing in most regions of the world[1]. The number of early onset CRC cases is expected to double from 2015 levels by 2030[2]. Global alterations to the genome, such as chromosomal instability (CIN), microsatellite instability (MSI), and the CpG-island methylator phenotype (CIMP) are characteristic phenotypes of CRC, however, most driver mutations are confined to a few biochemical pathways[3,4]. Consequently, a promising method for detecting CRC is by targeting somatic driver mutations within circulating cell-free DNA (cfDNA). Circulating tumour DNA (ctDNA) is the tumour fraction of cfDNA, and is released into the bloodstream by apoptosis, necrosis, or secretion. ctDNA can be collected from a routine blood draw and enables a non-invasive method for the detection and monitoring of cancer[5]. Furthermore, the short half-life of cfDNA (approximately 1–2 h) enables

[1]Department of Biochemistry, University of Otago, Dunedin 9016, Otago, New Zealand. [2]Malaghan Institute of Medical Research, Wellington 6012, New Zealand. ✉email: michael.dunnet@otago.ac.nz

real-time tracking[6], unlike protein-based biomarkers, which can have long lag times[5,7]. Many ctDNA assays employ digital PCR because of its high sensitivity, specificity, and throughput; however, these assays require prior knowledge of patient-specific somatic mutations. The cost of DNA sequencing is continually decreasing[8] and coupled with modern high-throughput sequencing technology and multiplex PCR gene panels, sequencing ctDNA for early CRC detection is an increasingly viable option.

Despite its appeal, several challenges must be overcome if ctDNA is to be reliably detected via DNA sequencing. Firstly, ctDNA molecules often make up less than 1% of all cfDNA in early-stage disease[9]. Therefore, each target locus must be sequenced at a sufficient depth to ensure consistent observation of rare variants. The probability of observing at least *k* variants at a position can be modelled using a binomial distribution, given the sequencing depth at that position. For example, to achieve a 95% probability of observing a variant, a read depth of 2995 is required for a 0.1% frequency variant, 1497 for a 0.2% frequency variant, and 598 for a 0.5% frequency variant, assuming no PCR or sequencing errors. Details on these calculations can be found in Supplementary Eq. 1. Secondly, variant detection can be impaired when the variant allele frequency approaches the background sequencing error rate. The error rate can be reduced with the incorporation of unique molecular identifiers (UMIs), which are short, random sequences of DNA ligated to the starting material[10]. UMIs allow each starting molecule to be uniquely tracked through PCR amplification and sequencing, meaning reads within the same UMI-family—those with identical UMIs that map to the same genomic location—represent PCR clones of the original cfDNA molecule. Consensus calling of reads within the same UMI-family can drastically reduce background noise levels and, as a consequence, lower the limit of detection[10]. However, introducing UMIs requires sequencing multiple PCR clones from each original molecule, further increasing the sequencing burden. Despite this, UMIs are integral to many somatic variant-calling pipelines because of their effectiveness[11–15]. Collectively, this means that effective detection of somatic mutations in cfDNA necessitates a high depth of sequencing due to the low abundance of tumour-derived DNA fragments and redundant sequencing of multiple PCR clones to improve accuracy.

One solution to these high sequencing requirements is to utilize a small sequencing panel, allowing each region to be sequenced to much greater depths with the same amount of total sequencing. While sequencing panels are often selected through literature review alone, we present a novel bioinformatics pipeline that optimizes target gene selection by framing cancer mutations as a set cover problem. Specifically, our pipeline examines cohort-wide tumour mutation profiles and selects the smallest number of target regions needed to ensure all samples are represented. Numerous highly sensitive and specific ctDNA sequencing assays have already been published[16–19]. Therefore, the focus of this work is solely on target selection and is intended to strengthen assay design in the context of early CRC detection, independent of the specific technique, rather than propose a new assay altogether.

We applied our pipeline to 2985 CRC samples from three datasets, identifying target regions across nine genes totalling only 11 kb. Our results show that 90–96% of samples tested contain mutations within our selected targets, suggesting these loci have great potential for detection of CRC from cfDNA samples. Furthermore, while our primary focus was on CRC, we examined mutation profiles from breast invasive ductal carcinoma and non-small cell lung adenocarcinoma samples to assess whether similarly condensed panels could be applied in those contexts.

Materials and methods

Public datasets

We elected to employ tissue-based CRC datasets over plasma-based datasets for the panel design, even though the end use is intended for cfDNA. This was because variant detection in ctDNA is notoriously difficult; there are substantial technical challenges that limit the sensitivity of ctDNA assays, one of which is the need for ultra-deep sequencing coverage[10]. During the undertaking of this work, no cfDNA datasets containing ultra-high depth whole exome sequencing or broad targeted sequencing could be identified. The prevalence of mutated genes in cfDNA is strongly associated with the prevalence of mutated genes in CRC tissue[20], so utilising a tissue-based datasets will still produce a relevant panel. An external validation approach was used for panel construction, whereby the initial panel was developed using one whole exome sequencing dataset and then validated using additional, independent datasets. This approach provides a more generalisable outcome with less risk of overfitting the panel to a single dataset.

We utilized 683 whole exome sequencing files in mutation annotation format (MAF) from The Cancer Genome Atlas (TCGA) colon adenocarcinoma (COAD) data set and 122 from the rectal adenocarcinoma (READ) data set as a discovery cohort. This dataset was chosen for discovery because variants are an aggregation from multiple variant callers which reduces bias associated with any single tool and increases sensitivity to a broad spectrum of variant types. Of the total 805 tumours, 86 were mucinous adenocarcinomas (70 from COAD and 16 from READ) and 719 were non-mucinous adenocarcinomas (613 from COAD and 106 from READ). All MAF files were obtained from the Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/) as masked, aggregated somatic mutation files, produced through the somatic aggregation workflow as described in the data portal. The sample manifest is in Supplementary Table 1.

We used two validation data sets. The first was comprised of whole exome sequencing data from 619 CRC samples in mutation annotation format, published by Giannakis et al.[21] (subsequently referred to as the DFCI data set). Colon was the primary site for 481 samples, rectum for 137, and one sample for which the primary site was not stated. Ninety-one samples were micro-satellite instable (MSI), 438 were micro-satellite stable (MSS), and the remaining 90 samples did not have micro-satellite status available. The CpG island methylator phenotype was categorised as low in 405 samples, high in 95 samples, and was not stated for 119 samples. All MAF files were obtained from cBioPortal (https://www.cbioportal.org/) as masked somatic mutation files. Somatic variant detection had been performed with MuTect[22] using the most up to date GATK best practice guidelines at the time of publication. The second dataset was comprised of targeted sequencing of 1,516 CRC samples using the MSK-IMPACT panel in mutation annotation format, published by Cercek et al.[23] (subsequently referred to as the JNCI dataset). This dataset contained 1083 adenocarcinomas from the colon, 411 from the rectum, and 67 without a reported primary location. Ninety-eight were classified as MSI, 1,357 as MSS, and 67 were not classified. All MAF files were obtained from cBioPortal (https://www.cbioportal.org/) as masked somatic mutation files. Cercek et al.[23] performed variant calling in accordance with the IMPACT-Pipeline. The sample manifests for each validation dataset can be found in Supplementary Tables 2 and 3, respectively.

To examine OPTIC's performance in other solid tumours, we obtained 630 lung adenocarcinoma (LADC) and 1013 breast invasive ductal carcinoma (BIDC) from the MSK-IMPACT clinical sequencing cohort[24]. Samples were obtained as masked somatic mutation files from cBioPortal (https://www.cbioportal.org/). The sample manifest is for lung and breast cancer samples are in Supplementary Tables 4 and 5, respectively.

Bioinformatic workflow

Here we present a bioinformatic pipeline called OPTIC (Oncogene Panel Tester for Identifying Cancers) which aids the generation of small sequencing panels. OPTIC takes MAF files as input and creates a binary array that indicates the presence of somatic mutations in every gene for each sample. The pipeline is designed to operate through a series of iterations, with each step followed by manual inspection. By default, each iteration of OPTIC produces several plots and data files to aid in panel development. Full descriptions of the outputs can be found in the GitHub repository: https://github.com/M-Dunnet/OPTIC.

The OPTIC workflow (Fig. 1) begins with preparing a variant filter file to remove non-pathogenetic variants. Next, an iteration of OPTIC can be run with any of the following options:

(1) Default, which examines all genes from all samples.
(2) Hierarchical clustering, which uses Ward's minimum variance method and Euclidean distance to cluster samples based on somatic mutation patterns within the binary mutation array.
(3) Greedy set coverage, which chooses the fewest number of genes to cover the highest number of samples.
(4) Targeted, in which the user provides a predefined list of genes for OPTIC to examine.

OPTIC is designed to first perform hierarchical clustering to separate tumours by molecular profiles for separate analysis, followed by greedy set coverage to identify optimal genes for a sequencing panel, and then targeted panel assessment. However, this order is not mandatory. Detailed descriptions of each option are provided below.

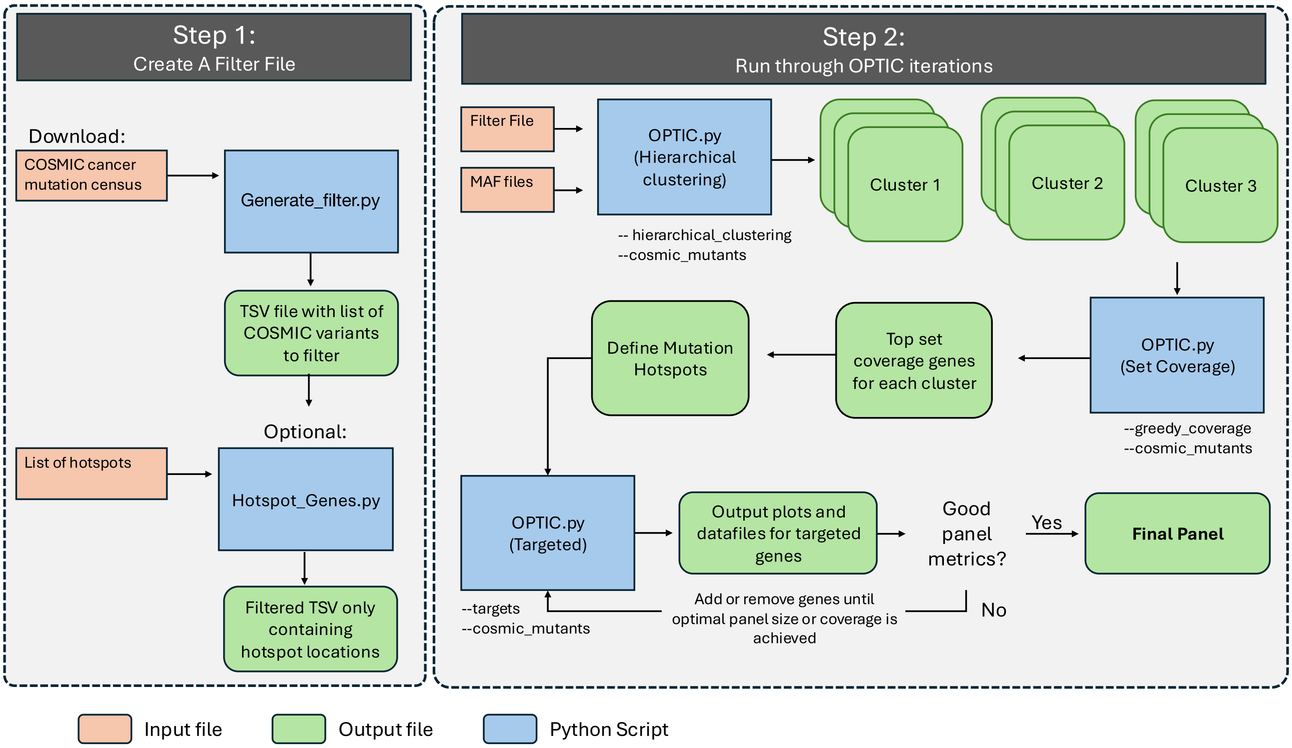
Step 1:
Create A Filter File
Download:
COSMIC cancer mutation census
Generate_filter.py
TSV file with list of COSMIC variants to filter
Optional:
List of hotspots
Hotspot_Genes.py
Filtered TSV only containing hotspot locations
Step 2:
Run through OPTIC iterations
Filter File
MAF files
OPTIC.py (Hierarchical clustering)
-- hierarchical_clustering
--cosmic_mutants
Cluster 1
Cluster 2
Cluster 3
OPTIC.py (Set Coverage)
--greedy_coverage
--cosmic_mutants
Top set coverage genes for each cluster
Define Mutation Hotspots
OPTIC.py (Targeted)
--targets
--cosmic_mutants
Output plots and datafiles for targeted genes
Good panel metrics?
Yes
No
Final Panel
Add or remove genes until optimal panel size or coverage is achieved
Input file
Output file
Python Script

Fig. 1. OPTIC Workflow. Orange boxes represent input files, green represents output files, and blue represents specific OPTIC scripts. Command line arguments for each OPTIC script are listed under the blue boxes. In some cases the output file from one script produces data that is used in another.

Generating a filter file

The filter removes passenger mutations and variants not clinically relevant to the disease using data from the Cancer Mutation Census in the Catalogue of Somatic Mutations in Cancer (COSMIC)[25]. The `Generate_filter.py` script transforms the cancer mutation census into a format readable by OPTIC and only variants within this file are analysed. Non-pathogenic variants, as determined via ClinVar classification and/or COSMIC's variant classification[26], can be excluded. The filter can be used directly with OPTIC or refined with 'Hotspot_Genes.py' to exclude variants outside mutation hotspots in specific genes. For instance, targeting BRAF V600E (chr7:140753336, GRCh38) excludes other BRAF variants while leaving other genes unaffected.

During filtering, OPTIC checks both the mutation's genomic position and amino acid change, only requiring a match in one to avoid discarding relevant mutations due to mapping or transcript discrepancies. Such discrepancies can arise from 0- vs. 1-based indexing, indels in repeat regions, or differences in reference transcripts. The transcript ID for variants used by OPTIC is in the filter file, listed in column two. If no filter file is provided, all mutations are considered.

Hierarchical clustering

Hierarchical clustering separates samples into multiple clusters based on molecular subtype; for example, hypermutated and non-hypermutated CRCs. Clustering enables groups of samples with different mutation profiles to be examined separately, allowing for more precise identification of unique patterns within each group. Furthermore, it may enable the detection of low-frequency subtypes driven by specific genes not easily identifiable in large sample sizes. All plots and data files are generated for each generated cluster.

Greedy set coverage algorithm

Finding the minimum number of genes required to cover the maximum number of CRC samples is an example of the set cover problem. Here, the universal set U represents CRC samples, and each subset $S_{1,} S_{2,\dots,} S_n$ corresponds to a gene that is mutated in specific samples within U. The objective is to find the smallest number of genes (subsets) whose mutations collectively cover all the CRC samples in U. OPTIC uses a greedy algorithm to address this problem. It begins by selecting the gene with the highest mutation rate across all samples. Subsequent genes are chosen based on their ability to cover the maximum number of previously uncovered samples. When the option for set coverage is used, the order of genes in the mutation matrix and set coverage plot is determined by the number of new samples covered, rather than by total mutation frequency.

Targeted OPTIC

Users can provide OPTIC with a text file containing a list of genes using HUGO gene nomenclature to limit the analysis to that gene set. This enables users to examine how a specific set of genes perform, or to check metrics with the addition or removal of specific genes.

Defining mutation hotspot regions

OPTIC does not explicitly define mutation hotspots. However, it generates a file containing counts of all gene variants, which can be used to identify regions with a high concentration of mutations, potentially indicating mutation hotspot regions.

In our analyses, we used known mutation hotspot data from cBioPortal, which defines a mutation hotspot as an amino acid position in a protein-coding gene mutated more frequently than would be expected in the absence of positive selection[27]. We supplemented this data with additional frequently mutated regions from the variant counts file. These mutation hotspots were defined as any regions where more than 50% of the variants from that gene were within 50 base-pairs of one another. We acknowledge that this definition may not capture diffuse mutation patterns present in some genes and may lead to underrepresentation of hotspots in such cases.

Statistical analysis and reproducibility

Hierarchical clustering of CRC samples in OPTIC was performed directly on publicly accessible MAF data. All other OPTIC analyses were performed with an OPTIC filter file containing all COSMIC tier three and above variants from COSMIC's Cancer Mutation Census (v101).

Additional statistical analyses outside of OPTIC were performed in R (v4.3.1)[28]. Co-occurrence testing of gene pairs was performed using Fisher's Exact Test using the fisher.test() function in R. Results were evaluated using p-values, confidence intervals, and odds ratios, with all tests run at 95% confidence. Linear regression was calculated using the lm() function in R with confidence intervals at 95%. Calculations for the binomial cumulative distribution function were performed with the SciPy package (v.1.11.4) in Python (v.3.9.6).

Gene mutation analysis by stage was performed with the mafCompare() function from the maftools package (v2.22.0) for R[29].

Results

Separation of hypermutated and non-hypermutated colorectal cancers

We aimed to identify target regions for single nucleotide variant (SNV) and insertion/deletion (INDEL) detection by examining exonic regions of CRC-associated genes, with the goal of detecting the maximum number of CRC cases using the minimum number of assessed genomic regions. For our discovery dataset, we used 805 MAF files from the publicly available TCGA COAD and READ data sets[30]. CRC can be either hypermutated or non-hypermutated, which have large differences in molecular phenotype[31]. We separated non-hypermutated (675 samples) (Fig. 2A) and hypermutated (130 samples) (Fig. 2B) CRCs using OPTIC's hierarchal clustering, without the use of a filter file. Given the large clustering distance between the hypermutated and non-hypermutated CRCs, we kept these two groups separate for further analyses.

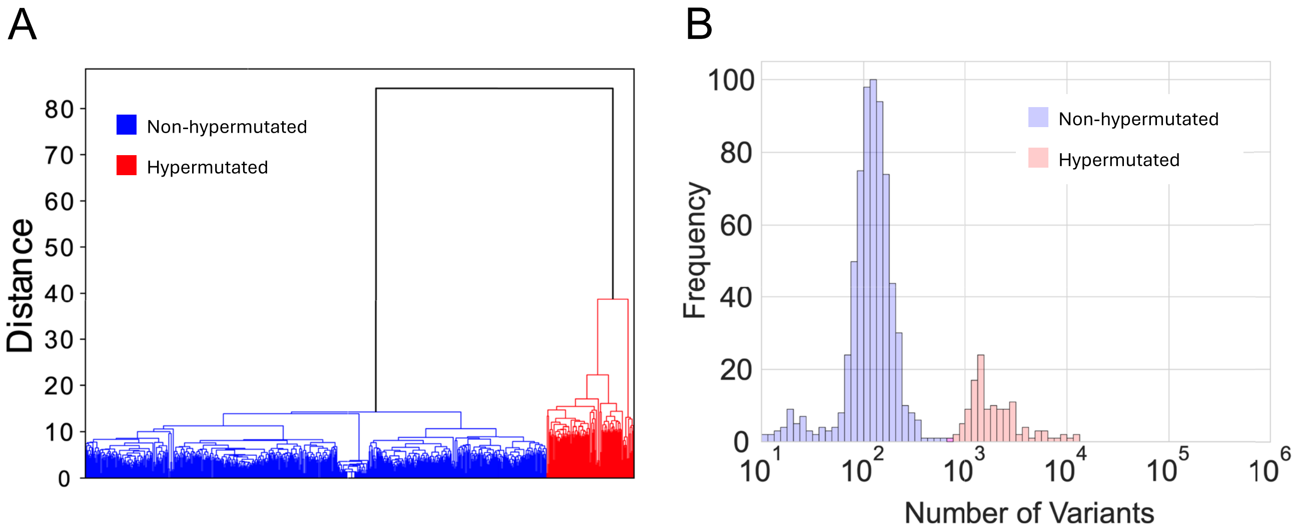

Fig. 2. Separating non-hypermutated from hypermutated colorectal cancers. (**A**) Dendrogram of all samples in the TCGA discovery dataset clustered by all somatic mutations. Blue: non-hypermutated cluster, Red: hypermutated cluster. (**B**) Histogram of the number of total variants per sample. Blue: non-hypermutated cluster, Red: hypermutated cluster.

For the remainder of our analyses, we used an OPTIC generated filter file that only contained variants catalogued as tier three or higher in the COSMIC cancer mutation census to prioritize driver mutations and exclude passenger mutations[25]. Classification at tier three requires either recurrent variants in known cancer genes, evidence of positive selection within the tumour, or for the variants to be classified as pathogenic in ClinVar cancer-related diseases. Full classification details for all tiers of mutation can be found at https://cancer.sanger.ac.uk/cosmic. Of the 805 CRC samples, 59 (7.3%) did not contain any variants after filtering. All these samples were within the non-hypermutated group.

The top four most mutated genes in the non-hypermutated cluster were *APC* (71.0%), *TP53* (56.3%), *KRAS* (41.3%), and *PIK3CA* (16.3%). These genes had much higher mutation frequencies than all other genes in the cluster. *FBXW7* was the fifth most mutated at 7.4%. Nevertheless, the top ten most mutated genes have all previously implicated with CRC progression previously[18] (Fig. 3A, left) (Supplementary Table 4). We observed several mutational patterns in this subgroup. *KRAS, NRAS,* and *BRAF* mutations were mutually exclusive, as were *APC* and *CTNNB1*. In contrast, *PIK3CA, FBXW7, AMER1*, and *SMAD4* mutations tended to co-occur with *KRAS* (Supplementary Fig. 1A). These patterns have all been previously described[33–38]. Furthermore, hierarchical clustering of this subgroup, resolved into four additional clusters, highlighted that the mutational landscape is dominated by five unique patterns: (1) *APC*, *TP53*, and *KRAS* mutations, (2) *APC* and *KRAS* mutations, (3) *APC* and *TP53* mutations, (4) *TP53* and *KRAS* mutations, (5) *APC* mutations alone (Supplementary Fig. 1B).

In the hypermutated sub-group *ACVR2A* (51.5%) was the most mutated gene, in line with previous reports[39,40]. *BRAF* (45.4%), *APC* (44.6%), and *RNF43* (43.9%) were also frequently mutated in the hypermutated subgroup, consistent with hypermutated tumours arising commonly from the serrated pathway[41–43] (Fig. 3A, right) (Supplementary Table 6).

Set coverage analysis of CRC samples

Greedy set coverage analysis of the non-hypermutated CRCs showed that 90.1% of all samples (98.7% of samples with variants after filtering) contained mutations in one or more of *APC, TP53, KRAS, BRAF, NRAS, ARID1A*, and *PIK3CA* (Fig. 3B,C). The remaining samples could each be represented by several mutant genes unique to that specific case; these genes were excluded from consideration. Interestingly, *BRAF*, *NRAS*, and *ARID1A* gave greater sample coverage than *PIK3CA* despite a lower combined mutation frequency. This observation is likely due to the association of *PIK3CA* mutations with *KRAS* mutations[34], while *KRAS*, *BRAF* and *NRAS* mutations are mutually exclusive[35]. Several genes with strong associations to CRC and with overall high mutational frequencies were not selected by the algorithm, namely *FBXW7* (7.4% mutation frequency, 5th most mutated), *AMER1* (7.1% mutation frequency, 6th most mutated), and *SMAD4* (5.2% mutation frequency, 7th most mutated). All three genes had a statistically significant tendency towards co-occurrence with *KRAS* (*FBXW7*: OR = 2.13, 95% CI 1.14–4.05; *AMER1*: OR = 2.62, 95% CI 1.37–5.16; *SMAD4*: OR = 2.70, 95% CI 1.26–6.09). *FBXW7* and AMER1 also tended towards co-occurrence with *APC* and *PIK3CA* (*APC*/*FBXW7*: OR = 1.94, 95% CI 0.91–4.46; *APC*/*AMER1*: OR = 2.14, 95% CI 0.96–5.39; *PIK3CA*/*FBXW7*: OR = 2.11, 95% CI 1.01–4.20; *PIK3CA*/*AMER1*: OR = 2.24, 95% CI 1.07–4.50), suggesting mutations in these genes regularly co-occur with other highly mutated genes and are redundant in terms of sample coverage (Supplementary Fig. 1, Supplementary Table 7).

All samples in the hypermutated cluster contained mutations in at least one of *ACVR2A, BRAF, APC, RNF43*, and *CTNNB1* (Fig. 3B,D). The first four genes are the most frequently mutated within the subgroup, while

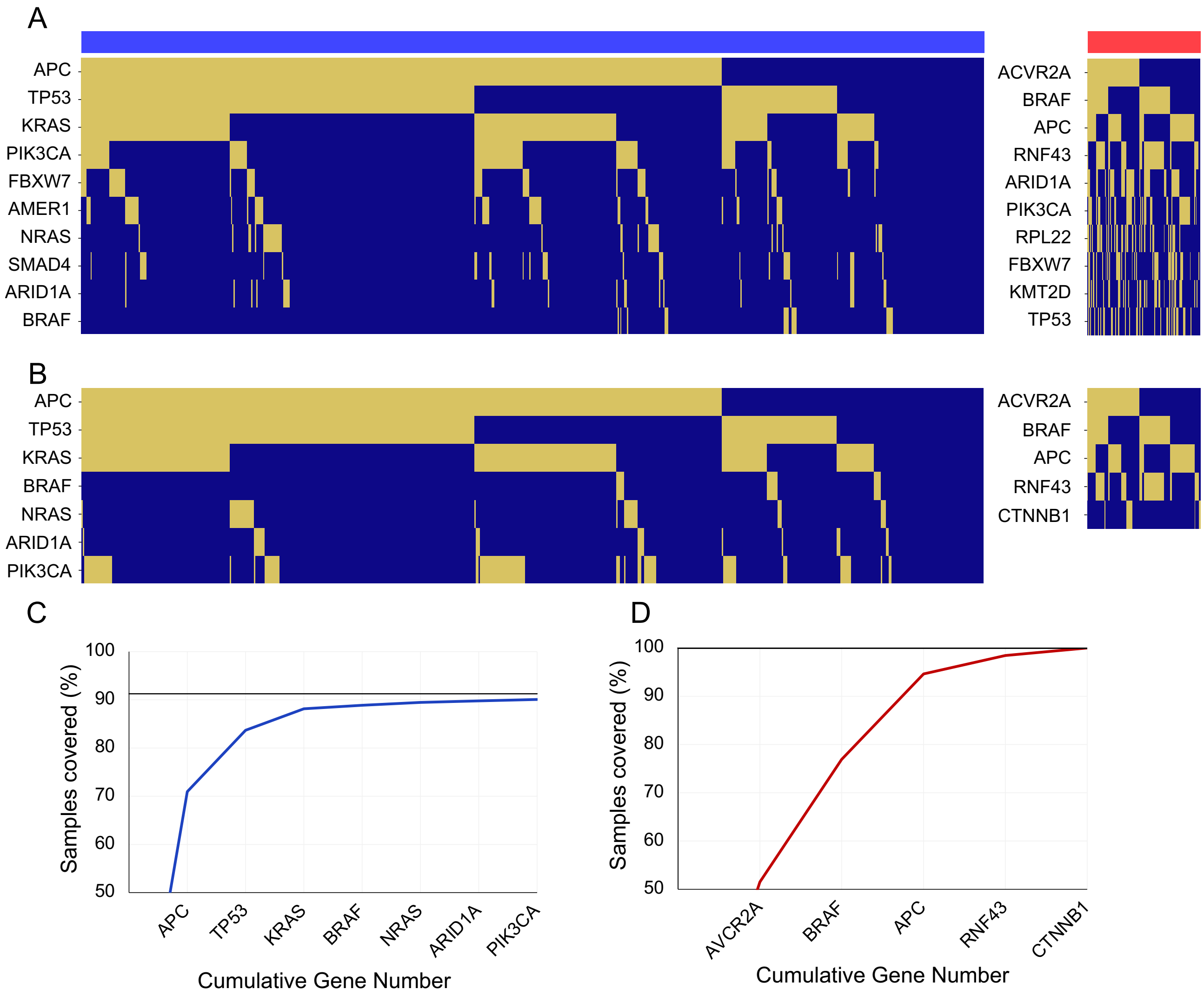

Fig. 3. TCGA mutation matrices based on total mutation frequency or set coverage analysis. (**A**) Mutation matrix for the non-hypermutated (left, blue bar) and hypermutated (right, red bar) clusters ordered by overall mutation frequency. Dark blue bars indicate WT alleles, while yellow indicates a somatic mutation in that gene. (**B**) Mutation matrices as in (**A**) for non-hypermutated (left) and hypermutated (right) clusters, with genes ordered by total coverage increase as determined by the greedy set coverage algorithm from OPTIC. (**C**) Cumulative sample coverage for the top seven genes in the non-hypermutated cluster. Genes are ordered by the set coverage algorithm. Horizontal black line indicates the best achievable maximum when samples with no mutations in the OPTIC filter file are excluded. (**D**) As in (**C**) for the hypermutated cluster.

CTNNB1 has a much lower mutation frequency. C*TNNB1's* inclusion is likely driven by its well-established mutual exclusivity with *APC*[33].

Mutation hotspots and removal of large genes

To reduce the total target region size, we evaluated the impact of only employing mutation hotspots in *KRAS*, *BRAF, NRAS, PIK3CA*, and *CTNNB1*. We compiled a list of known mutation hotspots from cBioPortal[44]. Furthermore, we defined additional hotspots from the TCGA dataset based on unique variant counts, revealing that most *ACVR2A* and *RNF43* mutations were c.1310del (84.5%) and c.1976del (85.2%), respectively. All mutation hotspots are listed in Supplementary Table 8. The majority of variants in both non-hypermutated and hypermutated samples were concentrated in these hotspots (Fig. 4A,B). When only hotspot regions within hotspot-mutated genes were employed, the overall sample coverage was not impacted (Fig. 4C). Therefore only variants in mutation hotspots were employed for *KRAS, BRAF, NRAS, PIK3CA*, *CTNNB1, ACVR2A* and *RNF43*.

The primary goal of making a small panel is to reduce overall sequencing requirements. It is therefore necessary to consider both gene number and gene size. We sought to eliminate large genes that offered minimal improvements in sample coverage. For each gene, we examined the percentage increase in coverage for non-hypermutated (Fig. 4D) and hypermutated (Fig. 4E) CRCs relative to the size of the targeted regions. For genes with mutation hotspots in Supplementary Table 8, only these hotspot regions were considered as part of gene

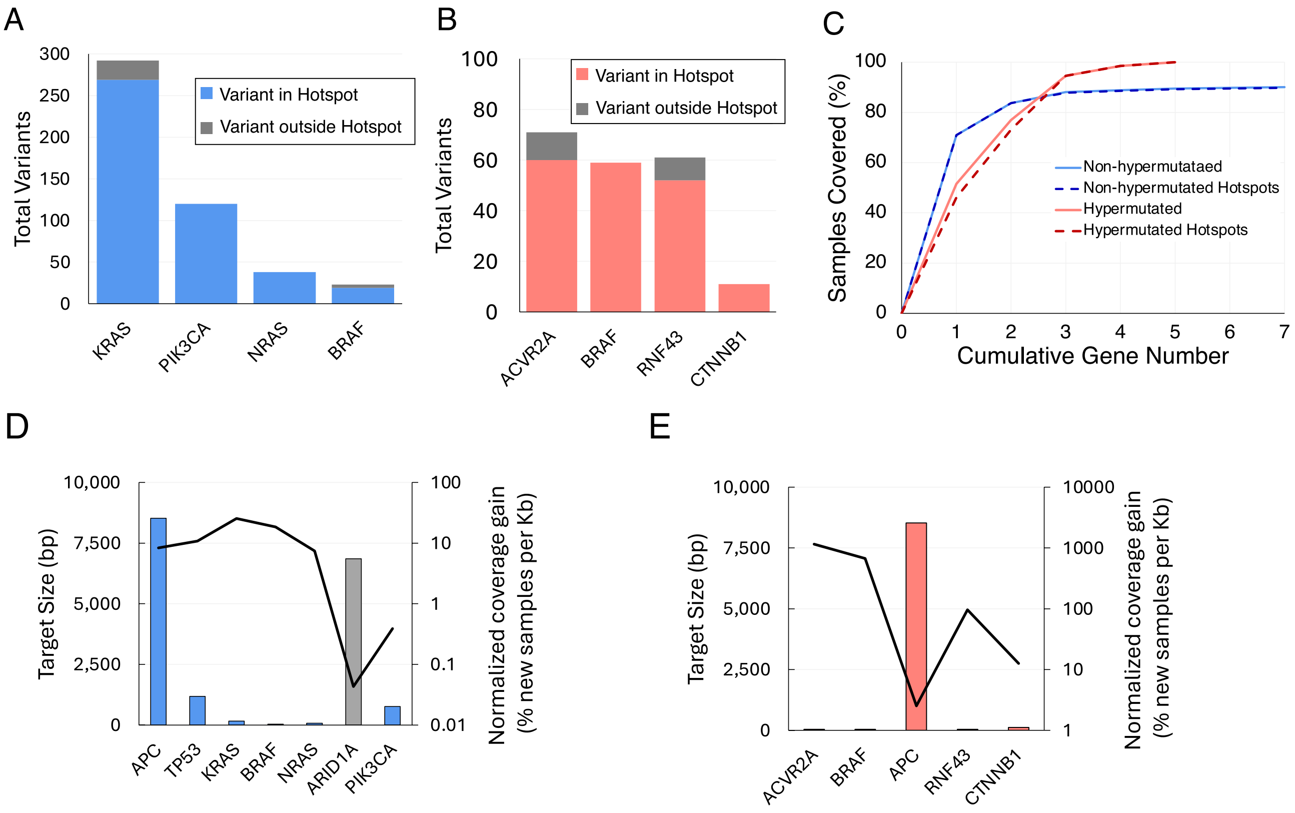

Fig. 4. Mutation hotspot region and gene size analysis for the TCGA discovery dataset. (**A**) Proportion of variants from selected genes within mutation hotspot locations in non-hypermutated colorectal cancers (CRCs). (**B**) As in (**A**) for hypermutated genes. (**C**) Difference in total set coverage when using all variants (solid line) compared to variants only with mutation hotspot regions (dotted line) for non-hypermutated CRCs (blue) and hypermutated CRCs (red). (**D**) Total gene size (bars) and normalized coverage gain (percent of new samples covered divided by target size in kilobases) (line) for the top genes in the non-hypermutated group. Coloured bars indicate a selected gene, grey bars indicate an excluded gene. The secondary Y-axis has a logarithmic scale (**E**) As in (**D**), for top genes in the hypermutated group.

size; otherwise, the full coding sequence was used. Among the non-hypermutated CRCs, only *APC* and *ARID1A* had particularly large target regions, with lengths of 8,529 and 6,855 bases, respectively. We chose to exclude *ARID1A* from the list of target loci due to its minimal impact, with only a 0.04% increase in sample coverage normalised to gene length (0.3% overall coverage increase), while *APC* was retained because it provided the most significant sample coverage across all genes. The genes used for the hypermutated cluster (*ACVR2A, BRAF, APC, RNF43*, and *CTNNB1*) were kept because they either had small, targetable mutation hotspots, were already present in the non-hypermutated target genes, or both.

The final target gene list consisted of *APC, TP53, KRAS, PIK3CA, NRAS, BRAF, ACVR2A, RNF43*, and *CTNNB1* (Fig. 5A). *APC, TP53, KRAS*, and *PIK3CA* are frequently mutated in the non-hypermutated cluster (Table 1; Fig. 5A, left). *CTNNB1* and *NRAS* are mutated much less frequently, but very few bases are required to cover their mutation hotspots. Of all *CTNNB1* mutations, 17 out of 18 (94.4%) occurred without concurrent APC mutations, in line with the ability of mutant β-catenin to drive CRC progression in the absence of mutated *APC*[45]. *ACVR2A*, *RNF43* and *BRAF* were frequently mutated in the hypermutated cluster. *PIK3CA, CTNNB1*, and *NRAS* mutations were also more frequent in this cluster compared to the non-hypermutated cluster, while the *APC, TP53,* and *KRAS* mutation frequencies were reduced, albeit still occurring at a high rate (Table 1; Fig. 5A, right). In the non-hypermutated samples, 89.5% of CRCs contained mutations within the target loci (97.6% of CRCs with variants after filtering) (Fig. 5B). In hypermutated CRCs, 100% contained mutations in the target loci (Fig. 5C), resulting in a combined coverage of 91.2% across all samples (98.4% of CRCs with variants after filtering) from the TCGA dataset (Fig. 5D).

Validation with whole exome and targeted sequencing datasets

We externally validated the panel with the DFCI dataset[21] comprising 619 whole exome CRC sequences (Supplementary Fig. 2). We separated the CRCs based on overall mutation patterns with hierarchical clustering, as with the TCGA dataset, into non-hypermutated and hypermutated subgroups. After clustering, there were 507 and 112 CRCs in the non-hypermutated and hypermutated groups, respectively. After applying the filter file, 35 CRCs contained zero variants, all of which were in the non-hypermutated subgroup. In the non-hypermutated cluster, *APC*, *TP53*, and *KRAS* were highly mutated (Supplementary Fig. 2A, left; Table 1), while *ACVR2A*,

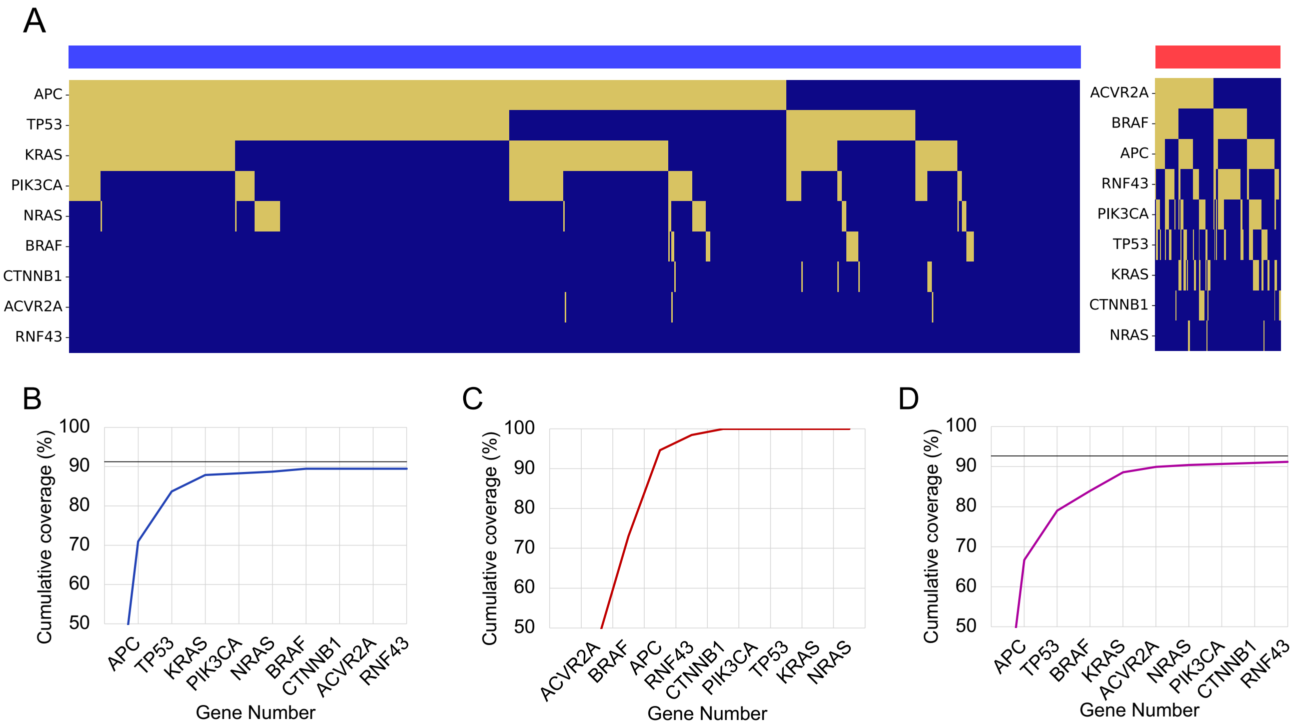

Fig. 5. Nine-gene panel metrics (**A**) Mutation matrices as in Fig. 3 for the nine-gene panel in non-hypermutated (Left) and hypermutated (right) colorectal cancers (CRCs) from the TCGA discovery dataset, ordered alphabetically. (**B**) Cumulative sample coverage for non-hypermutated CRCs. Genes are ordered by total mutation frequency. Horizontal black line indicates the best achievable maximum when samples with no mutations in the OPTIC filter file are excluded (**C**) As in (**B**) for hypermutated CRCs. (**D**) As in (**B**) for all CRCs combined.

	Non-hypermutated		**Hypermutated**		**Total**	
Gene	**TCGA**	**DFCI**	**TCGA**	**DFCI**	**TCGA**	**DFCI**
ACVR2A	0.4	2.3	46.2	27.7	7.8	6.95
APC	71.0	58.6	44.6	21.4	66.7	51.9
BRAF	2.8	9.86	45.4	60.7	9.7	19.1
CTNNB1	1.0	1.0	8.5	2.7	2.2	1.3
KRAS	41.3	30.2	21.5	9.8	38.1	26.5
NRAS	5.6	4.1	3.1	1.8	5.2	3.7
PIK3CA	16.3	15.3	34.6	25.9	19.3	16.8
RNF43	0.0	1.8	40.0	40.1	6.5	8.7
TP53	56.3	50.6	25.4	22.3	51.3	45.7

Table 1. Mutation rates (%) for the nine-gene panel in the TCGA and DFCI datasets. Mutation rates are shown separately for non-hypermutated and hypermutated cases. The Total columns represent the overall mutation rates across all samples, regardless of hypermutation status. Values indicate the percentage of samples in each dataset with a somatic mutation in the corresponding gene.

BRAF, *RNF43*, and *PIK3CA* were prominent in the hypermutated cluster (Supplementary Fig. 2A, right; Table 1). Although overall mutation patterns were similar, specific gene mutation frequencies differed between the datasets. Except for *BRAF* in hypermutated CRCs and *TP53* in non-hypermutated CRCs, mutation frequencies were generally lower in the DFCI validation set compared to the TCGA dataset (Table 1). The panel covered 89.5% of all CRCs in the dataset, and 96.2% of CRCs with variants after filtering (Supplementary Fig. 2D). Specifically, it included 87.7% of non-hypermutated CRCs (94.2% of non-hypermutated CRCs with variants after filtering) (Supplementary Fig. 2B) and 95.5% of hypermutated CRCs) (Supplementary Fig. 2C).

Sample coverage was lower in the DFCI dataset CRCs. One possible explanation is the over-fitting of our panel to the TCGA dataset. To assess this we first examined the correlation between overall gene mutation frequencies between the two datasets. In the DFCI non-hypermutated CRCs, we observed that the same set of genes (*APC, TP53, KRAS, PIK3CA, FBXW7, AMER1, NRAS,* and *SMAD4*) were consistently mutated as in the TCGA dataset. However, the mutation frequency of *BRAF* was higher (Supplementary Fig. 3A). Likewise in the hypermutated samples, *BRAF, RNF43, PIK3CA, ACVR2A, TP53, ARID1A, FBXW7*, and *APC* were also highly mutated in the DFCI dataset (Supplementary Fig. 3B). We also examined stage-specific mutation frequencies using available staging metadata, but found no consistent differences in pathway level alterations (Supplementary Fig. 4, Supplementary File 1) or individually mutated genes between stages. Sample coverage between stages was also largely the same, with some variation due to low CRC numbers in some groups (Supplementary Fig. 5, Supplementary File 1). This suggests mutation profiles are largely consistent across stages.

Direct comparisons of gene mutation frequencies between datasets showed a high correlation ($R^2 = 0.98$ and 0.74 for non-hypermutated and hypermutated CRCs, respectively). However, in non-hypermutated CRCs there was a 17% reduction in overall mutation frequency for genes in the DFCI data set compared to the TCGA dataset. In hypermutated CRCs there was a 29% reduction (Supplementary Fig. 3C-D). Overall, the same genes contained a similar pattern of mutations, but the overall gene mutation frequency was reduced in the DFCI data set.

To further examine the discrepancies between the two datasets, we reperformed the initial gene selection with set coverage analysis. We reasoned that if overfitting had occurred, re-construction of the panel using the DFCI dataset would result in a different set of genes. This identified *APC, TP53, KRAS, BRAF*, and *PIK3CA* as the most efficient gene set in the non-hypermutated subgroup, where 379 of 438 CRCs (85.2%, or 91.5% of samples with variants after filtering) contained mutations in these genes. CHEK2, *CTNNB1, MTOR*, and *SMAD4* covered a collective 30 additional samples, while the remaining samples with variants after filtering were covered by individual genes (Supplementary Fig. 6A, C). For hypermutated CRCs, *BRAF*, *RNF43, APC, PIK3CA*, and *B2M* were the most efficient genes, where 94.6% of CRCs had mutations at these loci; the remaining 5.6% were covered by individual genes (Supplementary Fig. 6B, D). Notably, *ACVR2A*, despite being the most mutated gene in hypermutated CRCs from the TCGA dataset (Supplementary Table 6), was not among these genes. Its mutation rate in the DFCI hypermutated samples was 28.57%, much lower than reported rates between 50 and 90%[30,46,47]. Overall, there was only a slight disparity between the two datasets, whereby *CHEK2*, *MTOR* and *B2M* were selected in the DFCI dataset, while *NRAS* and *ACVR2A* were selected in the TCGA dataset. The high correlation of gene mutation frequencies between the two datasets, along with the selection of similar genes after set coverage analysis show that the overall mutational landscape is well preserved. We suspect that the overall lower mutation frequencies observed in the DFCI dataset is the primary cause of the reduced sample coverage. One explanation is that the TCGA dataset is comprised of aggregated somatic annotations from several variant callers, while the DFCI dataset is based on only one variant caller (see Materials and Methods for details).

The TCGA MAF files indicate which variant callers were used for each variant. To evaluate the impact of calling strategy on target region selection, we re-analysed the TCGA dataset using only Mutect2-called variants, restricted to the nine genes identified in the original TCGA analysis. Sample coverage decreased slightly, from 91.2 to 89.7%, aligning more closely with the DFCI coverage of 89.2% (Supplementary Fig. 7). Gene-level mutation frequencies also declined across the panel but maintained dataset-specific patterns, with gene from the TCGA dataset showing generally higher frequencies except for *BRAF* (Supplementary Table 9). Thus, variant calling strategy is a confounder to overall sample coverage, while dataset-specific variation remains a dominant factor. We elected to maintain the original panel developed using the TCGA dataset and justified the inclusion of *ACVR2A*, despite it not increasing sample coverage in the DFCI dataset, because it has been identified as a biomarker for MSI-high CRCs in multiple studies[48,49].

In both the TCGA and DFCI datasets, similar sets of genes provided the most efficient coverage of CRCs. However, about 10% of samples in both groups were not covered by these genes (Fig. 6B, Supplementary Fig. 2B). These samples either had sample-specific mutations or did not contain any valid variants after filtering. Uncaptured CRCs may be driven by epigenetic changes or chromosomal copy number alterations. Alternatively, driver mutations, although present in the tumour, may not have been identified during variant calling due to a lack of sequencing depth or tumour impurity. We hypothesised that low sequencing depth could have resulted in missed variants, leading to loss of coverage. To test this, we examined read depth for each variant in the TCGA dataset (Fig. 6A) and compared it to a targeted sequencing dataset with 1516 CRC samples with much greater variant depth published by Cercek et al. (JNCI dataset)[23] (Fig. 6B). Unfortunately, variant depth was not available for the DFCI dataset. The median variant depth of the TCGA variants was 116 (standard deviation = 164.1), while the median depth of the JNCI dataset was 644 (standard deviation = 363.8). Of the 1516 CRC samples in the JNCI dataset, only 33 (2.19%) did not contain any variants after filtering. Nevertheless, 96.3% of samples (99.0% excluding samples with no valid variants) were covered with the final nine gene panel (Fig. 6C-D). Furthermore, the targeted sequencing panel in the JNCI dataset did not include ACVR2A, which we predict would further increase the coverage. The JNCI dataset's higher coverage correlated with fewer no-variant samples and better sample coverage compared to the TCGA and DFCI datasets, suggesting the metrics for the latter datasets are the result of lower sequencing depth. We suspect that with appropriate sequencing depth, the coverage rate for the nine-gene panel should approach full inclusion of all CRC samples. To summarize the panel's performance across the training and validation datasets, Fig. 7 presents an overlay of total sample coverage from each dataset for direct comparison.

Application of the OPTIC workflow to additional solid tumours

This study aimed to identify the smallest set of target regions that maximized CRC detection from cfDNA. However, we also sought to evaluate the performance of the OPTIC pipeline in other solid tumour types beyond CRC. To this end, we utilized somatic variant data from 1013 BIDCs and 630 LADCs obtained from the MSK-IMPACT clinical sequencing cohort[24] to assess its applicability to these cancer types.

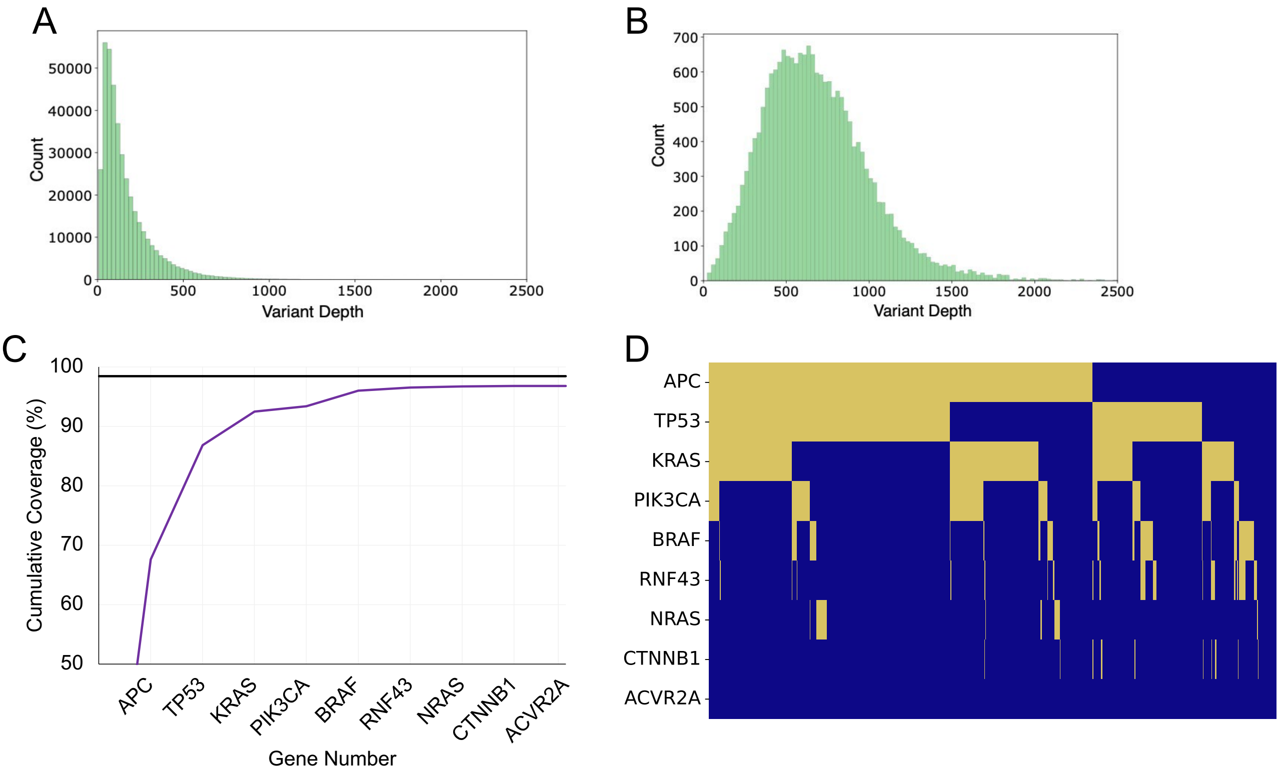

Fig. 6. Validation with the JNCI dataset. (**A**) Histogram of variant depth in the for all variants in TCGA dataset. (**B**) Histogram of variant depth for all variants in the JNCI dataset. (**C**) Cumulative sample coverage of the nine-gene panel in the JNCI dataset. Genes are ordered by total mutation frequency. The black horizontal line represents the maximum coverage when samples with no variants after filtering are excluded. (**D**) mutation matrix (as in Fig. 3) for the nine-gene panel in the JNCI dataset, ordered alphabetically.

Application of the set coverage algorithm to the BIDC cohort revealed that the top ten selected genes were mutated in 80.0% of cases. Notably, four genes—*TP53, PIK3CA, ESR1*, and *CDH1*—accounted for mutations in 71.4% of samples (Supplementary Fig. 8A–B). In the LADC cohort, the top ten genes identified by the algorithm were mutated in 88.6% of samples, while the top five—*TP53, KRAS, EGFR, STK11*, and *BRAF*—were mutated in 84.2% of cases (Supplementary Fig. 8C–D). In both cancer types, as observed previously for CRC, the highest-ranking genes were recurrently mutated in the majority of samples. However, the total proportion of samples with mutations in the top ten genes was lower in BIDC and LADC than in CRC, suggesting that a broader set of target regions may be necessary to achieve comprehensive coverage based on somatic mutations alone. Furthermore, a pattern of diminishing returns was observed with the incremental addition of genes, underscoring the trade-off between maximizing sample coverage and minimizing the number of genomic targets required for efficient panel design.

Discussion

Numerous strategies, encompassing both wet-lab methodologies and computational analysis pipelines, have been developed to improve the detection of circulating tumour DNA (ctDNA) from cell-free DNA (cfDNA) samples. These include approaches aimed at increasing ctDNA capture rates[50], enhancing sequencing accuracy and lowering limits of detection[51–54], as well as the development and validation of assays specifically tailored for ctDNA analysis[17,55]. Given these advancements, the present study focuses exclusively on the selection of target regions. This approach was pursued not with the intention of developing an entirely new assay, but rather to inform and potentially refine the design of existing or future assays by identifying genomic regions with maximal diagnostic relevance.

In this work, we developed a bioinformatic pipeline, OPTIC, designed to help design small, sequencing-efficient panels. Using OPTIC, we developed a highly compact panel for detecting CRC. To generate this panel, we utilized three publicly available datasets: two whole exome sequencing datasets and one extended targeted sequencing panel, comprising a total of 2940 CRC samples. These datasets included a mix of primary anatomical locations, molecular phenotypes, and both mucinous and non-mucinous adenocarcinomas, ensuring the panel's broad applicability across different CRC sub-types.

The intended use for the panel is early CRC detection from ctDNA. Several factors influence the sensitivity of ctDNA assays, such as the variant calling software limits of detection, sequencing depth, and assay-specific metrics like DNA capture efficiency[10,56]; however, our focus was solely on informing target selection. As such, we

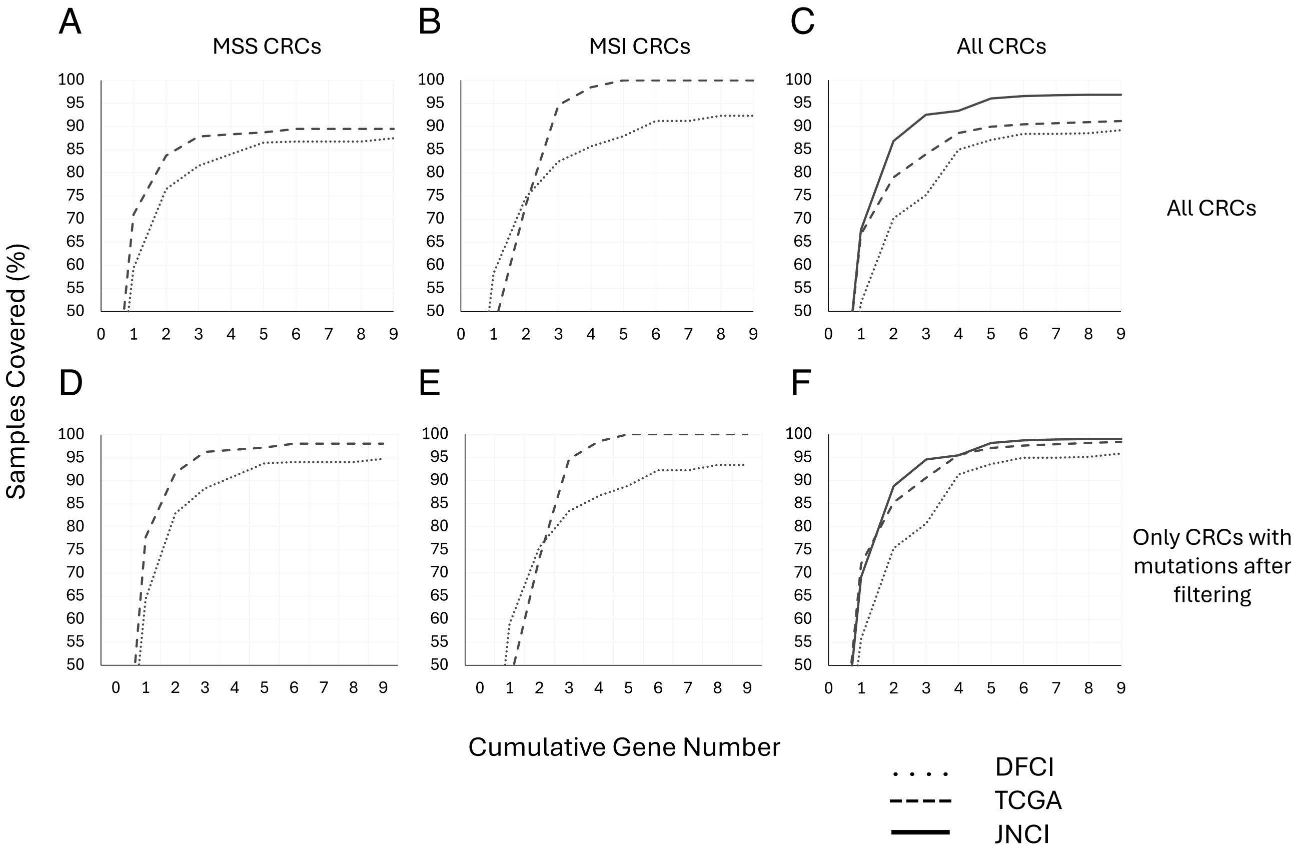

Fig. 7. Sample coverage comparison of the nine-gene panel across datasets. (**A**–**C**) Cumulative sample coverage plots using all samples, including those with no variants after filtering, for: (**A**) non-hypermutated/microsatellite stable (MSS) colorectal cancers (CRCs), (**B**) hypermutated/microsatellite instable (MSI) CRCs, (**C**) all samples. (**D**–**F**)) Cumulative sample coverage plots excluding CRCs with no variants after filtering, for: (**D**) non-hypermutated/MSS CRCs, (**E**) hypermutated/MSI CRCs, (**F**) all samples.

did not perform sensitivity analyses, which are more appropriate in the context of complete assay development and validation.

Nevertheless, we would expect our panel to excel over other large SNV- and INDEL-based panels in a cfDNA context. For instance, even on a low-throughput platform such as the Illumina iSeq100, which generates approximately 8 million paired-end reads per run, our 10 kb panel would yield an estimated average coverage depth of 72,800 × per sample, assuming 100 bp fragment lengths:

$$\frac{8{,}000{,}000\ read\ pairs \times 100\ bp\ fragments}{10{,}975\ bp\ panel\ size} = 72{,}839\ fold\ average\ depth$$

Importantly, most early stage ctDNA samples typically contain between 5 and 16 ng/mL, corresponding to 1500–4800 genome equivalents per mL[56]. In this context, extremely high sequencing depths can be leveraged to increase unique molecular identifier (UMI) family sizes, which are critical for reducing sequencing error rates and improving the limit of detection in UMI-aware variant callers. Although the above example is hypothetical, these depth estimates suggest that nearly every ctDNA molecule incorporated into the sequencing library would be represented in the sequencing output. Ultimately, whether a ctDNA signal is detected depends on (1) the presence of a ctDNA molecule in the original blood draw and (2) retention of material throughout library preparation[10]. The latter is influenced by the chosen enrichment method; for example, PCR-based enrichment can be performed with lower DNA input but tends to be less uniform than hybridisation-based capture[57]. As such, provided enough ctDNA molecules are successfully incorporated into the library, the focused and high-depth coverage achieved through OPTIC-selected targets substantially increases the likelihood of its detection.

OPTIC as a tool for panel construction in cfDNA sequencing

The identification of tumour-derived somatic variants from cfDNA is a promising approach for the early detection of cancer[58]. Shedding of ctDNA from tumours into circulation varies based on cancer stage (where lower stage cancers tend to shed less ctDNA) and by cancer type[59]. Regardless, ctDNA typically comprises less than 1% of the overall cfDNA[58]. As a result, sequencing thousands of independent cfDNA molecules is often required to reliably detect ctDNA. Additionally, reducing noise from sequencing errors and PCR artifacts is required to differentiate true variants from background errors. Strategies such as molecular barcoding-based consensus calling reduce noise levels[10]; however, this relies on the redundant over-sequencing of PCR clones and further increases the sequencing burden. Indeed, this sequencing requirement is one of the biggest barriers preventing the success of ctDNA sequencing in a clinical setting[58]. OPTIC was designed to alleviate this sequencing burden by identifying the smallest number of target regions that can detect as close to all cancer cases as possible. The purpose of this is to maximize the sequencing depth per target region relative to the total bases sequenced, allowing for either greater depth per sequencing run or increased sample multiplexing, resulting in a lower sequencing cost per sample. While we have primarily applied OPTIC to CRC, and to a lesser extent LADC and BIDC, in this study, we envision that it will be a useful panel design tool for many solid cancers.

Rational for a greedy set coverage algorithm

OPTIC relies on multiple iterations of hierarchal clustering, set coverage, and targeted gene analysis so that mutation data can be combined with current disease literature to flexibly create a cfDNA sequencing panel. We chose to utilize a greedy algorithm because they are very computationally efficient and allow for faster processing of each iteration[60]. Greedy algorithms are a heuristic approach that focus on making the best possible decision at each individual step by always choosing the next set with the most previously uncovered elements. This prioritizes immediate, local improvements but does not consider the overall, long-term impact of each choice. When analysing a complete set of elements, for example, if the algorithm could examine every possible case of CRC, greedy algorithms will provide close to optimal solutions, but have the potential to miss the most effective combination of sets. In such cases, exact algorithms, such as a branch and bound algorithm, offer the optimal solution[60]. However, it is impossible to study the complete set of cancer cases; it is only possible to take a cohort of cancer cases to study. Thus, we reasoned that exact algorithms might produce a set that is overfit to the dataset and not necessarily generalizable to the entire population. This is because the exact solutions will include dataset-specific variance, either as biological variance within the samples, or from noise related to false positive and negative variant calls. While overfitting concerns apply to any model, we expected a greedy algorithm's preference for more frequently mutated genes to reduce dataset-specific overfitting while capturing broader disease trends. Indeed, in the context of gene selection, we have observed that the greedy algorithm tends to favour genes with (a) higher mutation counts and, (b) genes with mutations that tend to be mutually exclusive with those in other genes.

Furthermore, a key limitation of exact algorithms is that they only identify the optimal subset of sets that collectively achieve complete coverage of the universe. This focus on a single, complete solution constrains their applicability in contexts where a stepwise or incremental selection process is desirable. In contrast, the greedy algorithm produces an ordered sequence of sets based on their marginal contribution to coverage at each iteration. This property enables a natural ranking of genes by importance and facilitates intermediate evaluations of panel performance at varying sizes. Consequently, greedy selection is better suited for applications such as gene panel design, where practical constraints often necessitate evaluating partial solutions and prioritizing sets by their individual utility.

Finally, greedy algorithms are much more computationally efficient. The time complexity of a greedy algorithm is:

$$O(n \cdot m)$$

where n is the number of elements to cover and m is the number of subsets, and the time taken scales linearly with the size of the dataset. Conversely, the time complexity of a branch and bound method in the worst-case scenario is exponential:

$$O(2^m)$$

In the worst-case scenario all possible combinations are tired, however, in practice this is often faster due to the pruning of less efficient branches. Nevertheless, as time complexity increases exponentially it is unsuitable for large scale analyses with thousands of cancer samples.

Guidance of multimodal ctDNA detection assays

In all three cancer types (CRC, LADC, and BIDC), we observed that the addition of extra target regions resulted in diminishing returns to sample coverage, as represented by a decrease in slope gradient of the sample coverage plots (Fig. 5, Supplementary Fig. 8). While it is theoretically possible to select targets which contain mutations in all samples by continuously adding additional loci, it is cost- and sequencing-inefficient. Recent studies have demonstrated that multimodal strategies for ctDNA detection can offer improved sensitivity compared to approaches relying solely on somatic mutations by integrating SNVs with other genomic features such as aberrant DNA methylation, copy number alterations, and cfDNA fragment length[61,62]. In multimodal strategies that incorporate somatic mutation analysis, the OPTIC pipeline can serve as a complementary tool to refine target selection by prioritizing the most informative SNV regions. This targeted prioritization reduces redundant sequencing, thereby freeing up sequencing capacity for additional molecular modalities. This refinement is particularly relevant to cases where multimodal approaches are likely necessary to achieve comprehensive cancer detection. For instance, set coverage analysis of BIDC and LADC samples showed that the top 10 genes covered only 80.0% and 88.6% of samples, respectively. However, copy number alterations frequently drive many breast cancers. In ER-positive breast cancer, whole chromosome arm amplification of 1q occurs in 60% of cases[63] and focal amplification of 11q13 occurs in 20% of cases[64], while ERBB2 amplification occurs in 60–90% of HER2-positive breast cancers[64]. In this context, prioritisation of only the most informative somatic mutations frees up sequencing capacity for the inclusion of additional copy number alteration targets.

Nevertheless, not all cancer types require multimodal approaches. Cancer Personalized Profiling by deep Sequencing (CAPP-Seq), a ctDNA method based solely on SNV and INDEL detection, achieved a specificity of 96% and a sensitivity of 50% in stage I and 100% in stage II–IV in non-small-cell lung cancers[65]. Those results highlight the potential of streamlined, mutation-only strategies in certain tumour contexts.

Limitations of OPTIC

OPTIC was designed to create a cancer detection panel, not for the complete molecular characterisation of a tumour. The number of somatic mutations varies greatly by cancer type, but most cancers contain hundreds of somatic variants and multiple driver mutations[66]. As a result, additional characterisation such as whole exome sequencing or large-scale targeted gene assays (for example, FoundationOne CDx[67] or MSK-IMPACT[68]) must be carried out to obtain a complete molecular phenotype. However, these assays must be performed on tissue biopsies after diagnosis rather than on plasma-derived cfDNA given their sequencing requirements.

Designing a panel to contain the minimum number of target regions leaves little room for redundancy. Our panel has been designed with early CRC detection from ctDNA in mind. However, in this context the number of tumour-derived molecules is very low, and variant dropout is possible. To mitigate this, additional, redundant genes may be considered to improve the likelihood that at least one informative mutation is detected per sample. In this instance, OPTIC would be required to identify two (or more) mutated genes in each sample more than once before considering that sample as adequately covered.

Redundancy may also be required in highly heterogenous cancers. For example, triple negative breast cancer has the highest mutation rate of all breast cancer subtypes but possesses high levels of intra-tumour heterogeneity[69]. Other than *TP53*, there are very few genes that are consistently mutated, and somatic variants tend to occur in tumour subclones at low frequencies[69]. In this case, panel redundancy is desired because the risk of false negatives is much greater with ultra-low frequency somatic variants. Nevertheless, while panels designed by OPTIC may not be directly suitable in this context, their small sequencing footprint does allow for the addition of more genes with greater flexibility, thus they may be a strong starting point. This concept can also be applied to prognostic markers, whereby OPTIC can select the best genes to detect a tumour, and prognostic genes can be easily added to quickly triage patients in the absence of additional characterisation.

The design of our nine-gene panel is based on tumour tissue data, despite its intended application in a ctDNA context. Although overall mutation patterns correlate strongly between tissue and ctDNA in CRC[20], supporting the relevance of tissue-informed panel design, this does not guarantee that the same mutations will be detectable in plasma. Biological factors such as tumour shedding, anatomical site, and cfDNA fragmentation can influence which variants are released into circulation[70]. In addition, the ctDNA fraction occurs at very low levels compared with background cfDNA and shows high variability even within tumours of the same stage[71]. For example, Bettegowda et al. (2014) report that 78% of localised CRCs (stages I–III) had detectable ctDNA, and in those, the number of mutant molecules per 5 mL of plasma ranged from 3 to 745, but is predominately below 100 copies per 5 mL[71]. At such low molecule counts, detection can become stochastic, depending on whether sufficient ctDNA molecules are captured during extraction and retained during processing and sequencing[9]. While higher sequencing depth improves variant detection, it cannot fully compensate for very low input molecule numbers. Consequently, even when a pathogenic mutation lies within a targeted region, its detectability in plasma depends jointly on ctDNA fraction, molecule recovery, sequencing depth, and error rates[9,10]. Although our panel covers 96% of CRC cases at the tissue level, this figure represents a theoretical maximum. In practice, clinical sensitivity is likely to be lower, particularly in early-stage disease. CfDNA-based studies are required to establish the true per-stage performance of this panel. For comparison, sensitives from existing ctDNA assays are between 43.3 and 73.9% for stage I, 62.3–88.3% for stage II–III, and > 90% for stage IV, metastatic CRC[61,71,72].

Finally, OPTIC, like all analysis tools, relies on high quality input data that is generalisable to the entire population[20]. Low quality or low depth sequencing will impair variant calling, leading to missed variants. During panel construction we used a variant filter file to remove non-pathogenic mutations. This resulted in a small fraction of samples (between 2 and 8%, depending on the dataset) containing zero variants. We have provided evidence that this was, at least in part, due to low sequencing depth. Furthermore, results may vary depending on the variant calling strategy. Conservative variant callers or parameters may miss variants, yielding lower sample coverage but fewer false positives, whereas less conservative approaches produce the opposite. To assess this effect, we compared TCGA data processed with aggregated variant calls versus MuTect2-only calls and benchmarked both against the DFCI dataset, which was generated with MuTect (Supplementary Fig. 7, Supplementary Table 9). As expected, overall sample coverage decreased in the MuTect2-only TCGA dataset, bringing it closer to the DFCI dataset than the aggregated TCGA calls. Nevertheless, gene-level mutation frequencies remained consistent, suggesting that while variant caller choice is a confounder in OPTIC analysis, overall mutation patterns are preserved. High-frequency genes that drive coverage (e.g., *APC*, *TP53*, *KRAS*) appear robust to caller choice, whereas differences emerge at later stages of target selection where incremental coverage gains are small and caller-specific biases are amplified. For example, greedy set coverage analysis consistently selected *APC*, *TP53*, *KRAS*, *BRAF*, *PIK3CA*, *RNF43*, and *CTNNB1* across datasets, but diverged on *ACVR2A* and *NRAS* (TCGA) versus *CHEK2*, *B2M*, and *MTOR* (DFCI). While the absence of *ACVR2A* in DFCI likely reflects dataset-specific biology (notably, its unusually low mutation rate), other differences may have arisen from either variant caller bias or cohort-specific features. This highlights the need for (1) consistent processing across cohorts, or (2) analyses across multiple datasets, to minimize the confounding effects of variant caller choice.

Design considerations with respect to CRC heterogeneity

CRC is a heterogeneous disease where multiple regions of a tumour will have arisen from genetically distant subclones, each with their unique mutation profile[73]. Precancerous adenomas and carcinomas in situ often exhibit multiple subclones with unique driver mutations[74]. These subclones are under selective pressure and evolve in ways that indicate tumour evolution via natural selection; however, upon the transition to colon or rectum adenocarcinoma the canonical driver mutations such as *APC, KRAS,* and *TP53* typically become ubiquitous among all subclones. Tumour heterogeneity at this stage reflects neutral evolution in a treatment naive adenocarcinoma[75]. Given the datasets we examined lacked precancerous lesions or carcinoma in situ, and truncal mutations tend to produce strong signals despite expected tumour heterogeneity, most sub-clonal variants are likely sample-specific. By using a pathogenic variant filter and a greedy algorithm favouring frequently mutated genes, intra-tumour heterogeneity is unlikely to limit panel utility. Nonetheless, heterogeneity underscores the need to select datasets with relevant sample characteristics.

Relevance of the panel genes to colorectal cancer

CRC is a heterogenous disease that can possess several genome-wide molecular phenotypes: CIN, MSI, and CIMP[76]. Despite this, the events that lead to CRC are limited to a small number of signalling pathways: Wnt signalling, MAPK/ERK signalling, PI3K/AKT/mTOR signalling, TGF-β signalling, and the TP53 pathway[3,77]. All genes selected for the panel are members of these signaling pathways and are well-established drivers of CRC pathogenesis. In most CRCs *WNT* signaling upregulation, primarily due to *APC* truncation, drives tumor initiation. Additional *WNT* regulators like *CTNNB1*, *FBXW7*, *AMER1*, and *TCF7L2* also contribute to tumorigenesis, albeit at a lesser frequency[45,76]. *CTNNB1* driver mutations, unlike *FBXW7, AMER1,* and *TCF7L2,* are mutually exclusive with *APC* because they are sufficient for independent constitutive activation of the Wnt signaling pathway[33]. *RNF43*, a RING-type E3 ubiquitin ligase, negatively regulates Wnt signaling by directing the proteasomal degradation of the Wnt receptor Frizzled[78]. However, unlike *CTNNB1* and *APC*, which drive tumor initiation, loss of function *RNF43* mutations typically occur in the serrated pathway late in tumor progression[43]. Mutated *RNF43* regularly co-occurs with *MLH1* hypermethylation[41] and *BRAF* V600E mutations[43], and is mutually exclusive with *APC* truncating mutations[79]. Furthermore, *RNF43* mutations are associated with good prognostic outcomes in serrated CRCs treated with immunotherapy or anti-BRAF/EGFR therapy[80,81], which allows for additional prognostic information to be collected by this assay.

The MAPK/ERK signalling pathway stimulates cell growth and proliferation. In CRC, the pathway frequently acquires gain of function mutations in *KRAS*, *NRAS*, and *BRAF*. Driver mutations in codons 12, 13 and 61 for *KRAS* and *NRAS*, and codon 600 for *BRAF* all enable constitutive activation of the pathway and unregulated cell growth[82,83]. As a result, mutations in these genes are mutually exclusive[84]. The serrated pathway is initiated by upregulation of MAPK/ERK signalling due to mutations in *BRAF*, whereas *KRAS* and *NRAS* are commonly mutated in the traditional progression pathway following upregulation of Wnt signalling[76]. As we observed in the non-hypermutated CRC set coverage analysis (Fig. 3A), all three genes are needed for optimum sample coverage, likely due to their mutually exclusive nature.

Canonically, *TP53* is activated upon genotoxic stress and induces serval downstream responses such as cell cycle arrest, DNA repair, and apoptosis[85]. Mutations in *TP53* can have a wide spectrum of effects, whereby loss of function mutations prevent canonical activation of TP53 target genes and as a result increased cell proliferation and survival[86]. In contrast, gain of function mutations in the DNA binding domain of TP53 can have an oncogenic effect that results in transcription of non-canonical target genes[86]. Gain of function *TP53* mutations have been shown to increase tumour progression and metastasis risk[87]. *TP53* mutations occur more frequently in MSS CRCs than MSI CRCs and are relatively late events that are typically present in carcinomas, but not adenomas[88–90]. This suggests that *TP53* promotes the final progression from adenoma to cancer, making it especially important for distinguishing between adenomas and carcinomas.

Gain of function mutations in *PIK3CA* are the most common variant type in the PI3K/AKT/mTOR signalling pathway in CRC[91,92]. Upregulation of this pathway directly promotes proliferation and survival, apoptosis, migration, and metabolism[93]. Mutations in *PIK3CA* regularly co-occur with *KRAS* mutations[34,94], explaining why it was ranked lower in the set coverage analysis compared to genes with lower mutation frequencies (Fig. 3A). However, CRCs with *PIK3CA* mutations but wild type *KRAS* are often less responsive to anti-EGFR therapy and have worse clinical outcomes[84], suggesting that *PIK3CA* is also a valuable prognostic marker. The least characterized gene in the panel, *ACVR2A*, encodes a transmembrane serine-threonine kinase receptor which mediates the functions of activins in the TGF-β signaling pathway. *ACVR2A* has recently emerged as a strong biomarker for MSI-high CRCs[48]. In vitro analysis suggests that *ACVR2A* suppression activates the PI3K/AKT/mTOR pathway and subsequently induces angiogenesis in hypoxic CRCs[48].

Exclusion of other known CRC-related genes

Several highly mutated genes in the TCGA discovery dataset, such as *AMER1*, *SMAD4*, and *FBXW7*, were not selected for the final panel despite their known roles in CRC progression[32,95]. Overall, the exclusion of these genes from the panel is likely due to their tendency to co-occur with other genes rather than display mutual exclusivity, a common feature of the selected panel genes (Supplementary Table 7). Nevertheless, each of these genes can provide valuable prognostic information that may warrant their inclusion in an expanded panel.

AMER1 functions as both a positive and negative regulator of the Wnt signalling pathway[96]; however, its exact role in CRC remains unclear. *AMER1* mutations are not mutually exclusive with other Wnt regulators[97], suggesting that they do not sufficiently drive Wnt signalling enough to promote carcinogenesis, but can act as a secondary driver of Wnt signalling[98,99]. Furthermore, *AMER1* deficient CRCs exhibit a mesenchymal phenotype[95] and have been linked to CRC metastasis through non-Wnt related mechanisms, such as ferroptosis inhibition[100].

SMAD4 is a component of the TGF-β signalling pathway, which is often protective during the early stages of CRC progression[101], but drives epithelial-mesenchymal transition (EMT) in later stage tumours[102]. Loss of SMAD4 typically occurs in the transition from adenoma to malignancy, and is preceded by increased Wnt and MAPK/ERK signalling, frequently occurring alongside *KRAS* or *NRAS* mutations[36,103]. *SMAD4* mutant CRCs are associated with an overall worse prognosis, with worse overall survival, higher tumour, node, and metastasis (TNM) staging, and increased metastasis[103–105].

FBXW7 is the substrate targeting subunit of the ubiquitin ligase complex, and acts as a tumour suppressor gene by targeting several proto-oncogenes for degradation, such as *NOTCH1*, *MYC*, Cyclin E, and mTOR[106,107]. Regardless of hypermutation status, *FBXW7* was frequently mutated in both all three datasets (Supplementary Tables 6, 10 and 11). It is well documented that mutations in *FBXW7* can infer chemoresistance in a large number of cancers[108]. Li et al. demonstrated that ZEB2, which is a key inducer of EMT and chemoresistance, is a target for degradation by FBXW7, and that loss of FBXW7 function leads to the accumulation of ZEB2, enhanced EMT, and acquired chemoresistance[109]. Furthermore, *FBXW7* mutant CRCs are associated with worse overall survival than *FBXW7* wild-type CRCs[109–111].

Incorporating *FBXW7*, *AMER1*, and *SMAD4* could provide more comprehensive prognostic information and may be useful in an expanded panel. However, adding these genes would come at the cost of increased sequencing space, necessitating careful consideration of the balance between the depth of genetic coverage and the practical limitations of sequencing resources.

Conclusions

Here, we have presented a new bioinformatic tool called OPTIC to facilitate the design of small sequencing panels for ctDNA-based cancer detection. At the core of OPTIC is a greedy set cover algorithm that identifies the minimum number of genes needed to capture the majority of CRC samples. Using OPTIC, we have designed a nine gene ctDNA panel for detecting CRC. Between 90 and 96% of CRCs contain mutations in genes present in our panel.

The lower sequencing requirements of the panel should improve cost per sample and allow for more samples to be processed simultaneously in a single sequencing run without reductions in sequencing depth and variant calling accuracy. This enables cfDNA to be assayed at a higher throughput and a lower cost per sample than would be possible with a large panel, which is particularly important in a diagnostic setting where budget constraints or sample volume are considerations. While the nine-gene panel may exclude some frequently mutated genes (as discussed above), it efficiently identifies the most common drivers of CRC. Further sequencing would be necessary for comprehensive tumour characterization and prognostication. Nevertheless, the panel we have proposed provides a valuable tool for CRC detection and preliminary prognostic assessment. Additionally, in assays that integrate multiple molecular modalities, the application of a set coverage algorithm can provide a rational framework for somatic mutation target selection by identifying the point of diminishing returns. This enables a more strategic allocation of sequencing resources, allowing for the reallocation of sequencing capacity to other molecular features within multimodal detection strategies.

Data availability

The public data used in this manuscript were obtained from cBioPortal (https://www.cbioportal.org) and the Genomics Data Commons (GDC) portal (https://portal.gdc.cancer.gov). See the Methods section for details on specific datasets.

Code availability

OPTIC is publicly available from the following GitHub repository: https://github.com/M-Dunnet/OPTIC.

Received: 22 June 2025; Accepted: 15 October 2025
Published online: 20 November 2025

References

1. Morgan, E. et al. Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from GLOBOCAN. *Gut* **72**(2), 338–344 (2023).
2. Bailey, C. E. et al. Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975–2010. *JAMA Surg.* **150**(1), 17–22 (2015).
3. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **28**(6), 479–507 (2011).
4. Pierantoni, C., Cosentino, L., Ricciardiello, L. Molecular pathways of colorectal cancer development: mechanisms of action and evolution of main systemic therapy compunds. Vol. 42, Digestive Diseases. S. Karger AG; 2024. p. 319–24.
5. Zou, D. et al. Circulating tumor DNA is a sensitive marker for routine monitoring of treatment response in advanced colorectal cancer. *Carcinogenesis* **41**(11), 1507–1517 (2020).
6. Kustanovich A, Schwartz R, Peretz T, Grinshpun A. Life and death of circulating cell-free DNA. Vol. 20, Cancer Biology and Therapy. Taylor and Francis Inc.; 2019. p. 1057–67.
7. Yao, W., Mei, C., Nan, X. & Hui, L. Evaluation and comparison of in vitro degradation kinetics of DNA in serum, urine and saliva: A qualitative study. *Gene* **590**(1), 142–148 (2016).
8. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., *et al.* The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biol.* 17(1) (2016).
9. Abbosh C, Birkbak NJ, Swanton C. Early stage NSCLC—Challenges to implementing ctDNA-based screening and MRD detection. Vol. 15, Nature Reviews Clinical Oncology. Nature Publishing Group; 2018. p. 577–86.
10. Roberto TM, Jorge MA, Francisco GV, Noelia T, Pilar RG, Andrés C. Strategies for improving detection of circulating tumor DNA using next generation sequencing. Vol. 119, Cancer Treatment Reviews. W.B. Saunders Ltd; 2023.
11. Xu, C. et al. Smcounter2: An accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers. *Bioinformatics* **35**(8), 1299–1309 (2019).
12. Maruzani R, Brierley L, Jorgensen A, Fowler A. Benchmarking UMI-aware and standard variant callers for low frequency ctDNA variant detection. *BMC Genom.* **25**(1) (2024).
13. Shugay M, Zaretsky AR, Shagin DA, Shagina IA, Volchenkov IA, Shelenkov AA, et al. MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. PLoS Comput Biol. **13**(5) (2017).

14. Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: A sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ. **2016**(5) (2016).
15. Sater, V. et al. UMI-VarCal: A new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. *Bioinformatics* **36**(9), 2718–2724 (2020).
16. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med. [Internet]*. **9**(403) (2017). https://www.science.org
17. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science (1979) [Internet]. 359(6378):20 (2018).
18. Santonja A, Cooper WN, Eldridge MD, Edwards PAW, Morris JA, Edwards AR, et al. Comparison of tumor-informed and tumor-naïve sequencing assays for ctDNA detection in breast cancer. *EMBO Mol. Med.* 15(6) (2023).
19. Min L, Chen J, Yu M, Liu D. Using circulating tumor DNA as a novel biomarker to screen and diagnose colorectal cancer: A meta-analysis. *J. Clin. Med.* 12(2) (2023).
20. Strickler, J. H. et al. Genomic landscape of cell-free DNA in patients with colorectal cancer. *Cancer Discov.* **8**(2), 164–173 (2018).
21. Giannakis, M. et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep.* **15**(4), 857–865 (2016).
22. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**(3), 213–219 (2013).
23. Cercek, A. et al. A comprehensive comparison of early-onset and average-onset colorectal cancers. *J. Natl. Cancer Inst.* **113**(12), 1683–1692 (2021).
24. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**(6), 703–713 (2017).
25. Tate, J. G. et al. COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**(D1), D941–D947 (2019).
26. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(D1).
27. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**(2), 155–163 (2016).
28. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
29. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**(11), 1747–1756 (2018).
30. Muzny, D. M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**(7407), 330–337 (2012).
31. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**(11), 1350–1356 (2015).
32. Ciepiela I, Szczepaniak M, Ciepiela P, Hińcza-Nowak K, Kopczyński J, Macek P, et al. Tumor location matters, next generation sequencing mutation profiling of left-sided, rectal, and right-sided colorectal tumors in 552 patients. *Sci. Rep*. 14(1) (2024).
33. Sparks AB, Morin PJ, Vogelstein B, Kinzler KW. Mutational Analysis of the APC/B-Catenin/Tcf Pathway in Colorectal Cancer. Cancer Res [Internet]. 1998 Mar 15;58:1130–4. Available from: http://aacrjournals.org/cancerres/article-pdf/58/6/1130/2468751/cr0580061130.pdf
34. Janku F, Lee JJ, Tsimberidou AM, Hong DS, Naing A, Falchook GS, et al. PIK3CA mutations frequently coexist with ras and braf mutations in patients with advanced cancers. PLoS One. 2011;6(7).
35. De Roock W, De Schutter J, Biesmans B, Tejpar S, Research Centre V, Leuven K, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. Lancet Oncology [Internet]. 2010;11:753–62
36. Sarshekeh AM, Advani S, Overman MJ, Manyam G, Kee BK, Fogelman DR, et al. Association of SMAD4 mutation with patient demographics, tumor characteristics, and clinical outcomes in colorectal cancer. PLoS One. 12(3) (2017).
37. Korphaisarn, K. et al. High frequency of KRAS codon 146 and FBXW7 mutations in Thai patients with stage II-III colon cancer. *Asian Pac. J. Cancer Prev.* **20**(8), 2319–2326 (2019).
38. Guo, L. et al. Molecular profiling provides clinical insights into targeted and immunotherapies as well as colorectal cancer prognosis. *Gastroenterology* **165**(2), 414-428.e7 (2023).
39. Donehower, L. A. et al. MLH1 -silenced and non-silenced subgroups of hypermutated colorectal carcinomas have distinct mutational landscapes. *J. Pathol.* **229**(1), 99–110 (2013).
40. Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22**(11), 1342–1350 (2016).
41. Yan, H. H. N. et al. RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. *Gut* **66**(9), 1645–1656 (2017).
42. Rustgi AK. BRAF: A Driver of the Serrated Pathway in Colon Cancer. Vol. 24, Cancer Cell. Cell Press; 2013. p. 1–2.
43. Catherine E. Bond, Diane M. McKeone, Murugan Kalimutho, Mark L. Bettington, Sally-Ann Pearson, Troy D. Dumenil, et al. RNF43 and ZNRF3 are commonly altered in serrated pathway colorectal tumorigenesis. Oncotarget. 2016;7(43).
44. Cerami, E. et al. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**(5), 401–404 (2012).
45. Shitoh K, Furukawa T, Kojima M, Konishi F, Miyaki M, Tsukamoto T, et al. Frequent Activation of the-Catenin-Tcf Signaling Pathway in Nonfamilial Colorectal Carcinomas With Microsatellite Instability. 2001; Available from: https://onlinelibrary.wiley.com/doi/https://doi.org/10.1002/1098-2264
46. Müller MF, Ibrahim AEK, Arends MJ. Molecular pathological classification of colorectal cancer. Vol. 469, Virchows Archiv. Springer Verlag; 2016. p. 125–34.
47. Pinheiro, M. et al. Target gene mutational pattern in Lynch syndrome colorectal carcinomas according to tumour location and germline mutation. *Br. J. Cancer.* **113**(4), 686–692 (2015).
48. Wang J, Zhang Z, Liu H, Liu N, Hu Y, Guo W, et al. Identification of 8 candidate microsatellite instability loci in colorectal cancer and validation of the ACVR2A mechanism in the tumor progression. Sci Rep. 2024 Dec 1;14(1).
49. Zhao, L. et al. Microsatellite instability-related ACVR2A mutations partially account for decreased lymph node metastasis in MSI-H gastric cancers. *Onco Targets Ther.* **13**, 3809–3821 (2020).
50. Martin-Alonso C, Tabrizi S, Xiong K, Blewett T, Sridhar S, Crnjac A, et al. Priming agents transiently reduce the clearance of cell-free DNA to improve liquid biopsies. Science (1979). 2024 Jan 1;383(6680):1–10.
51. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA.* **108**(23), 9530–9535 (2011).
52. Verma S, Moore MW, Ringler R, Ghosal A, Horvath K, Naef T, et al. Analytical performance evaluation of a commercial next generation sequencing liquid biopsy platform using plasma ctDNA, reference standards, and synthetic serial dilution samples derived from normal plasma. BMC Cancer. 2020 Oct 1;20(1).
53. Lanman RB, Mortimer SA, Zill OA, Sebisanovic D, Lopez R, Blau S, et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. PLoS One. 2015 Oct 16;10(10).
54. Calapre, L. et al. Locus-specific concordance of genomic alterations between tissue and plasma circulating tumor DNA in metastatic melanoma. *Mol. Oncol.* **13**(2), 171–184 (2019).

55. Li W, Huang X, Patel R, Schleifman E, Fu S, Shames DS, et al. Analytical evaluation of circulating tumor DNA sequencing assays. Sci Rep. 2024 Dec 1;14(1).
56. Chen H, An Y, Wang C, Zhou J. Circulating tumor DNA in colorectal cancer: biology, methods and applications. Vol. 16, Discover Oncology. Springer Science and Business Media B.V.; 2025.
57. Singh, R. R. *Target Enrichment Approaches for Next-Generation Sequencing Applications in Oncology* Vol. 12 (Multidisciplinary Digital Publishing Institute (MDPI), 2022).
58. Medina, J. E. et al. *Cell-Free DNA Approaches for Cancer Early Detection and Interception* Vol. 11 (BMJ Publishing Group, 2023).
59. Sánchez-Herrero E, Serna-Blasco R, Robado de Lope L, González-Rumayor V, Romero A, Provencio M. Circulating Tumor DNA as a Cancer Biomarker: An Overview of Biological Features and Factors That may Impact on ctDNA Analysis. Vol. 12, Frontiers in Oncology. Frontiers Media S.A.; 2022.
60. Wang LT, Chang YW, Cheng KT (Tim). Electronic Design Automation. Elsevier; 2009. 173–234 p.
61. Nguyen, V. T. C. et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. *Elife* **11**, 12 (2023).
62. Lennon AM, Buchanan AH, Kinde I, Warren A, Honushefsky A, Cohain AT, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention
63. Shahrouzi P, Forouz F, Mathelier A, Kristensen VN, Duijf PHG. Copy number alterations: a catastrophic orchestration of the breast cancer genome. Vol. 30, Trends in Molecular Medicine. Elsevier Ltd; 2024. p. 750–64.
64. Razavi, P. et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* **34**(3), 427-438.e6 (2018).
65. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**(5), 548–554 (2014).
66. Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W. & Quackenbush, J. Cancer subtype identification using somatic mutation data. *Br. J. Cancer.* **118**(11), 1492–1501 (2018).
67. Milbury CA, Creeden J, Yip WK, Smith DL, Pattani V, Maxwell K, et al. Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. PLoS One. 2022 Mar 1;17(3 March).
68. Cheng, D. T. et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagnostics.* **17**(3), 251–264 (2015).
69. Asleh K, Riaz N, Nielsen TO. Heterogeneity of triple negative breast cancer: Current advances in subtyping and treatment implications. Vol. 41, Journal of Experimental and Clinical Cancer Research. BioMed Central Ltd; 2022.
70. Andersen, L. et al. Exploring the biology of ctDNA release in colorectal cancer. *Eur. J. Cancer.* **1**, 207 (2024).
71. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of Circulating Tumor DNA in Early-and Late-Stage Human Malignancies [Internet]. Available from: https://www.science.org
72. Klein, E. A. et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**(9), 1167–1177 (2021).
73. Chan DKH, Buczacki SJA. Tumour heterogeneity and evolutionary dynamics in colorectal cancer. Vol. 10, Oncogenesis. Springer Nature; 2021.
74. Saito T, Niida A, Uchi R, Hirata H, Komatsu H, Sakimura S, et al. A temporal shift of the evolutionary principle shaping intratumor heterogeneity in colorectal cancer. Nat Commun. 2018 Dec 1;9(1).
75. Uchi R, Takahashi Y, Niida A, Shimamura T, Hirata H, Sugimachi K, et al. Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. PLoS Genet. 2016 Feb 1;12(2).
76. Nguyen, L. H., Goel, A. & Chung, D. C. Pathways of colorectal carcinogenesis. *Gastroenterology* **158**(2), 291–302 (2020).
77. Pierantoni C, Cosentino L, Ricciardiello L. Molecular Pathways of Colorectal Cancer Development: Mechanisms of Action and Evolution of Main Systemic Therapy Compunds. Vol. 42, Digestive Diseases. S. Karger AG; 2024. p. 319–24.
78. Koo, B. K. et al. Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors. *Nature* **488**(7413), 665–669 (2012).
79. Giannakis, M. et al. RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat. Genet.* **46**(12), 1264–1266 (2014).
80. Elez, E. et al. RNF43 mutations predict response to anti-BRAF/EGFR combinatory therapies in BRAF V600E metastatic colorectal cancer. *Nat. Med.* **28**(10), 2162–2170 (2022).
81. Huang K, Ding S, Chen K, Guo C, Wu M, Zhang R, et al. RNF43 mutation as a predictor of immunotherapeutic efficacy in colorectal cancer [Internet]. Vol. 13, Am J Cancer Res. 2023. Available from: www.ajcr.us/
82. Li, Y., Xiao, J., Zhang, T., Zheng, Y. & Jin, H. Analysis of KRAS, NRAS, and BRAF Mutations, Microsatellite Instability, and Relevant Prognosis Effects in Patients With Early Colorectal Cancer: A Cohort Study in East Asia. *Front. Oncol.* **28**, 12 (2022).
83. Nandan MO, Yang VW. An update on the biology of RAS/RAF mutations in colorectal cancer. Vol. 7, Current Colorectal Cancer Reports. 2011. p. 113–20.
84. De Roock W, De Schutter J, Biesmans B, Tejpar S, Research Centre V, Leuven K, et al. Eff ects of KRAS, BRAF, NRAS, and PIK3CA mutations on the effi cacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. Lancet Oncology [Internet]. 2010;11:753–62. Available from: www.thelancet.com/oncology
85. Ozaki T, Nakagawara A. Role of p53 in cell death and human cancers. Vol. 3, Cancers. 2011. p. 994–1013.
86. Sabapathy K, Lane DP. Therapeutic targeting of p53: All mutants are equal, but some mutants are more equal than others. Vol. 15, Nature Reviews Clinical Oncology. Nature Publishing Group; 2018. p. 13–30.
87. Sabapathy K. The contrived mutant p53 oncogene - Beyond loss of functions. Vol. 5, Frontiers in Oncology. Frontiers Media S.A.; 2015.
88. Pan P, Li J, Wang B, Tan X, Yin H, Han Y, et al. Molecular characterization of colorectal adenoma and colorectal cancer via integrated genomic transcriptomic analysis. Front Oncol. 2023;13.
89. Jiri J, Marketa U, Arnoud B, Petr H, Petra B, Anna S, et al. Mutational analysis of driver genes defines the colorectal adenoma: in situ carcinoma transition. Sci Rep. 2022 Dec 1;12(1).
90. Wolff, R. K. et al. Mutation analysis of adenomas and carcinomas of the colon: Early and late drivers. *Genes Chromosomes Cancer.* **57**(7), 366–376 (2018).
91. Leiphrakpam PD, Chowdhury S, Wang J, Black JD, Are C. The role and therapeutic implications of PI3K signaling pathway in cancer. Vol. 123, Journal of Surgical Oncology. John Wiley and Sons Inc; 2021. p. 39–41.
92. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, et al. High Frequency of Mutations of the PIK3CA Gene in Human Cancers. Science (1979). 2004 Apr 23;304(5670):554.
93. Cathomas G. PIK3CA in colorectal cancer. Vol. 4 MAR, Frontiers in Oncology. Frontiers Research Foundation; 2014.
94. Nosho, K. et al. PIK3CA mutation in colorectal cancer: Relationship with genetic and epigenetic alterations. *Neoplasia* **10**(6), 534–541 (2008).
95. Sanz-Pamplona, R. et al. Exome sequencing reveals AMER1 as a frequently mutated gene in colorectal cancer. *Clin. Cancer Res.* **21**(20), 4709–4718 (2015).
96. Tanneberger, K. et al. Amer1/WTX couples Wnt-induced formation of PtdIns(4,5)P2 to LRP6 phosphorylation. *EMBO J.* **30**(8), 1433–1443 (2011).
97. Peng, H. et al. Specific mutations in APC, with prognostic implications in metastatic colorectal cancer. *Cancer Res. Treat.* **55**(4), 1270–1280 (2023).

98. Brunet Guasch M, Feeley N, Soriano I, Thorn S, Tomlinson I, Nicholson MD, et al. Quantifying "just-right" APC inactivation for colorectal cancer initiation. bioRxiv preprint https://doi.org/10.1101/2024.07.10.602868 [Internet]. 2024.
99. Castro-Giner F, Ratcliffe P, Tomlinson I. The mini-driver model of polygenic cancer evolution. Vol. 15, Nature Reviews Cancer. Nature Publishing Group; 2015. p. 680–5.
100. Lei S, Chen C, Han F, Deng J, Huang D, Qian L, et al. AMER1 deficiency promotes the distant metastasis of colorectal cancer by inhibiting SLC7A11- and FTL-mediated ferroptosis. Cell Rep. 2023 Sep 26;42(9).
101. Zhao, M., Mishra, L. & Deng, C. X. The role of TGF-β/SMAD4 signaling in cancer. *Int. J. Biol. Sci.* **14**(2), 111–123 (2018).
102. Batlle E, Massagué J. Transforming Growth Factor-β Signaling in Immunity and Cancer. Vol. 50, Immunity. Cell Press; 2019. p. 924–40.
103. Fang T, Liang T, Wang Y, Wu H, Liu S, Xie L, et al. Prognostic role and clinicopathological features of SMAD4 gene mutation in colorectal cancer: A systematic review and meta-analysis. BMC Gastroenterol. 2021 Dec 1;21(1).
104. Yoo, S. Y. et al. Clinicopathological characterization and prognostic implication of SMAd4 expression in colorectal carcinoma. *J. Pathol. Transl. Med.* **53**(5), 289–297 (2019).
105. Mizuno, T. et al. SMAD4 gene mutation predicts poor prognosis in patients undergoing resection for colorectal liver metastases. *Eur. J. Surg. Oncol.* **44**(5), 684–692 (2018).
106. Yeh CH, Bellon M, Nicot C. FBXW7: A critical tumor suppressor of human cancers. Vol. 17, Molecular Cancer. BioMed Central Ltd.; 2018.
107. Shen W, Zhou Q, Peng C, Li J, Yuan Q, Zhu H, et al. FBXW7 and the Hallmarks of Cancer: Underlying Mechanisms and Prospective Strategies. Vol. 12, Frontiers in Oncology. Frontiers Media S.A.; 2022.
108. Yan L, Lin M, Pan S, Assaraf YG, Wang Z wei, Zhu X. Emerging roles of F-box proteins in cancer drug resistance. Vol. 49, Drug Resistance Updates. Churchill Livingstone; 2020.
109. Li N, Babaei-Jadidi R, Lorenzi F, Spencer-Dene B, Clarke P, Domingo E, et al. An FBXW7-ZEB2 axis links EMT and tumour microenvironment to promote colorectal cancer stem cells and chemoresistance. Oncogenesis. 2019 Mar 1;8(3).
110. Shang W, Yan C, Liu R, Chen L, Cheng D, Hao L, et al. Clinical significance of FBXW7 tumor suppressor gene mutations and expression in human colorectal cancer: a systemic review and meta-analysis. BMC Cancer. 2021 Dec 1;21(1).
111. Liu H, Wang K, Fu H, Song J. Low expression of the ubiquitin ligase FBXW7 correlates with poor prognosis of patients with colorectal cancer [Internet]. Vol. 11, Int J Clin Exp Pathol. 2018. Available from: www.ijcep.com/

Author contributions

K.M.-G. and M.D. developed the methodology, conducted the investigation, and performed the formal analysis. M.D curated the data, developed the software, and prepared the original draft, including figure preparation. M.D, M.B., R.D., and P.G. conceptualised the project. M.B., R.D., and P.G. supervised the project and provided resources. P.G. additionally oversaw project administration. All authors were involved in manuscript editing and reviewed the final manuscript.

Funding

This work was supported by Hope Without Fear and Healthier Lives—National Science Challenge.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24719-9.

Correspondence and requests for materials should be addressed to M.J.D.