81963_Auto_Edited.docx

WORD COUNT	6408	TIME SUBMITTED	20-JAN-2023 06:50PM
		PAPER ID	95742906

Name of Journal: *World Journal of Gastrointestinal Oncology*

Manuscript NO: 81963

Manuscript Type: REVIEW

Molecular methods for colorectal cancer screening: progress with next-generation sequencing evolution

NGS-based CRC screening

Salma Abbes, Simone Baldi, Hayet Sellami, Amedeo Amedei, Leila Keskes

Abstract

Currently the colorectal cancer (CRC) represents the third most common malignancy and the second most deadly cancer worldwide, with a higher incidence in developed countries. Like other solid tumors, CRC is a heterogeneous genomic disease in which various alterations, such as point mutations, genomic rearrangements, gene fusions, or chromosomal copy number alterations, can contribute to the disease development. However, because of its orderly natural history, easily accessible onset location and high lifetime incidence, CRC is ideally suited for preventive intervention, but the many screening efforts of the last decades have been compromised by performance limitations and low penetrance of the standard screening tools. Anyway, the advent of next-generation sequencing (NGS) has both facilitated the identification of previously unrecognized CRC features such as its relationship with gut microbial pathogens and revolutionized the speed and throughput of cataloguing CRC-related genomic alterations. Hence, in this review, we summarized the several diagnostic tools used for CRC screening in the past and the present, focusing on recent NGS approaches and their revolutionary role in i) the identification of novel genomic CRC characteristics, ii) the advancement of understanding the CRC carcinogenesis and iii) the screening of clinically actionable targets for personalized medicine.

Key Words: Colorectal Cancer; Gut Microbiota; Colorectal Cancer Screening; Next-Generation Sequencing

Abbes S, Baldi S, Sellami H, Amedei A, Keskes L. Molecular methods for colorectal cancer screening: progress with next-generation sequencing evolution. *World J Gastrointest Oncol* 2023; In press

Core Tip: Nowadays, due to the multitude of host and microbial genetic factors, the optimization of colorectal cancer (CRC) biomarkers remains difficult. Anyway, the advent of next-generation sequencing (NGS) methods has facilitated the identification

of previously unrecognized CRC-related genomic alterations and the CRC relationship with gut microbial composition. Hence, we have summarized the several diagnostic tools used for CRC screening in the past and the present, focusing on the revolutionary role of NGS approaches in i) the identification of novel genomic CRC characteristics, ii) the advancement of understanding the CRC carcinogenesis and iii) the screening of clinically actionable targets for personalized medicine.

INTRODUCTION

Currently the colorectal cancer (CRC) represents the third leading cause of cancer-related deaths in men and women worldwide and the American Cancer Society estimates that the number of new colon and rectum cancer cases in the United States in 2022 will almost be 106,180 and 44,850 respectively (1). Despite the great progress of modern medicine, such as the development of novel therapeutic methods and the advent of new high throughput sequencing technologies, the mortality of CRC patients remains relatively high due to the lack of specific biomarkers and therapies. Nowadays, CRC incidence largely varies across the world, and it appears to be positively correlated with the Human Development Index (HDI). For instance, in 2020, Norway, the Netherlands and Denmark, reported the highest age-standardized incidence rates (respectively 41.9, 41.0, and 40.9 cases per 100,000 persons) while Guinea, Gambia and Burkina Faso showed the lowest age-standardized incidence rates (respectively 3.3, 3.7 and 3.8 cases per 100,000 persons) (2). Usually, these variations reflect differences in the availability of screening services and other factors such as geographic location, environmental factors (*e.g.* polluted surface water sources), economic status, and dietary and lifestyle habits (3). At present, considering the difficulties in implementing significant lifestyle changes or common primary prevention strategies, screening and so the early detection represent the most powerful public health tool to reduce the CRC mortality (4). In general, a good screening marker can be considered by the health community only if it respects specific parameters such as simplicity, safety and accuracy and has a known and suitable cut-off level defined and agreed (5).

Colonoscopy is considered the gold standard test for detecting CRC and promoting effectiveness in reducing its incidence and mortality, however, its high cost, invasiveness and reduced availability of necessary equipment hinder the establishment of organized screening settings, especially in poor countries (6). Therefore, in recent years, considering the emerging evidence that intestinal microbial dysbiosis constitutes a crucial environmental factor in CRC onset and development, massive efforts were focused on next-generation sequencing (NGS) approaches to identify genes and microorganisms that are significantly associated with the malignancy (7). Moreover, metagenomics approaches, considered as a real revolution in the screening and diagnosis of different cancers, are also useful for the identification of novel potential markers for CRC diagnosis (8). Hence, in this review, we summarize the several diagnostic tools used for CRC screening in the past and the present, focusing on recent NGS approaches.

2. Fecal occult blood test

Since the 1970s, stool-based CRC screening was considered a successful non-invasive method with proven effectiveness given by the detection of high-risk polyps and early-stage malignancies that dramatically reduces CRC incidence and death (9) (Figure 1). Fecal occult blood test (FOBT) currently represents the early analysis for CRC screening that is recommended by the National Screening Committee (NSC)(10). This method based on detection of occult blood by measuring the non-protein portion of hemoglobin, the heme group, present in the stool. In particular, the heme present in a stool sample reacts with hydrogen peroxide–based developer to oxidize guaiac-infused paper, resulting in a blue color (4). In general, FOBT has been shown to reduce both the incidence and the risk of CRC death with the advantage to be easier and cheaper than other alternative screening approaches (11) (Table1). Despite this, the FOBT method presents some limitations, such as the low sensitivity for colorectal adenomas that may not bleed or the specificity of the method that can be influenced by diet or drugs (12); hence, Young *et al.*, affirmed that FOBT is only suitable for limited colonoscopy resources with a need to constrain test positivity rate (13).

3. Fecal immunochemical test

Different immunoassay methods have been used to measure the development of antibody–globin complexes, including immunochromatography, immunoturbidimetry and ELISA (Table1) (14). For instance, Fecal Immunochemical Test (FIT) is used for the detection of microscopic amounts of blood present in the stool during defecation *via* the utilization of antibodies targeted to globin molecules (Figure 1). The antibodies preferably target lower gastrointestinal bleeding, making FIT easy to use, sensitive to low concentrations of globin and sufficiently flexible to adjust cut-off concentration for positivity (cut-off is usually selected with a risk threshold that would produce a specificity of 96.9% in the study group, matching the specificity of FIT at a cut-off of 20 µg Hb/g faeces) (13, 15). Imperiale *et al*. tested individuals at average risk for CRC having an age comprised between fifty and eighty-four years and documented that FIT detected 48 out of 65 colon cancers, showing a sensitivity of 73.8% and specificity of 96% (16). The same specificity was observed among participants with negative results on colonoscopy, suggesting that FIT had fewer false positive results compared to stool DNA testing (17). On the other hand, weaknesses of FIT tests are the low clinical sensitivity for both cancers (73%, 80%, 82% and 79% for CRC stages I, II, III, and IV, respectively) and advanced adenomas (16%-34%) when used at a low cut-off and the limited detection of upper gastrointestinal bleeds because the hemoglobin undergoes degradation by digestive enzymes with a consequent reduction of the binding to FIT antibodies (18).

4. Flexible sigmoidoscopy and total colonoscopy

Randomized controlled trials showed that the visual inspection of colic mucosa through flexible sigmoidoscopy (FS) decreases the CRC mortality and incidence respectively by 22–31% and 18–23% (Figure1) (19). Overall, FS represents a safe test, but its use is limited to the distal colon and a combined strategy using FS and FOBT/FIT only increase endoscopic workload and reduce patients' participation without solving the problem. Instead, total colonoscopy (TC) allows direct visualization and polyp removal over the whole colon (Figure 1), have a very high sensitivity and specificity for CRC and

is usually used as confirmatory for all other screening strategies (Table1) (20). Anyway, although TC determine a relevant decrease of CRC incidence (66–90%) and mortality (31–65%), many features (*e.g.* invasive, expensive and painful) dramatic reduce its acceptability as a first-line screening test; moreover, proper training programs for endoscopists are necessary, as well as continuous quality assurance (21).

5. Methods-based on Sanger DNA sequencing

As currently well established, CRC development relies upon a stepwise acquisition of several chromosome mutations and the model of the adenoma-carcinoma progression, based on the accumulation of multiple mutations and epigenetic alterations, has been widely accepted (22). Overall, there are two types of mutational events in sporadic CRC: the first concern about 85% of all patients and consists of frequent mutations in APC (23), TP53 (24), KRAS (25), BRAF, TTN, PIK3CA (26), FBXW7 (27), and SMAD4 genes (28); while the second concern 15% of CRC-sporadic patients and is characterized by a high level of hypermethylation of the MLH1 gene, responsible for DNA mismatch repair (29). Additionally, a different complement of mutations in somatic genes has also been described (30).

5.1 Single gene sequencing

Considering their role in resistance to multiple treatment strategies, genotyping of gene mutations currently represents an important diagnostic and therapeutic tool (Figure 2). For instance, a mutation in APC, a tumor suppressor gene highly mutated in 57% of CRC cases and involved in DNA replication and repair processes, has been documented to strongly influence the chemotherapy response (31). Also, SMAD4 gene mutations were observed in 2-20% of CRC cases and were usually associated with poor response to cetuximab treatment (32). In addition, several RAF mutations have been implicated in the induction of genomic instability, driving the proliferation of cancer cells (33) while heterogeneous KRAS mutations have been identified in almost 40% of CRC patients (34) (with a substitution in the G12C position as the most common detected), having a consequent association with anti-EGFR treatment resistance (35). To date, to better represent the cancer heterogeneity using NGS technology, Ye *et al.* have proposed

a protocol for conducting rigorous systematic reviews and meta-analysis on the accuracy of KRAS mutation detection in CRC using non-invasive liquid biopsy samples. (36). Generally, liquid biopsies represent the collection of tumour-derived biomarkers in the blood or other body fluids, such as urine, saliva, stool or cerebrospinal fluid. Circulating tumour DNA (ctDNA), circulating tumour cells (CTCs) and exosomes are the most common tumour-related biomarkers assessed on liquid biopsy so far (37). Moreover, the FDA recently approved a liquid biopsy test to analyze the frequency of KRAS, NRAS and BRAF hotspot mutations in circulating tumor DNA that all could represent good CRC prognostic factors (38).

5.2 Multi-target stool DNA test

Multi-target stool DNA (MT-sDNA) test allows to identify specific gene mutations in human tumor DNA cells separately from the more abundant microbial DNA in the stool (Figure 2). During the last few years, several key technological advances have led to increasingly accurate approaches to stool DNA testing, including i) the use of a DNA preservative swab for stool collection, ii) the improvement of the target capture and amplification methods and iii) the identification of new informative marker panels (39). Zou *et al.*, have produced a methyl-binding domain (MBD) protein bounded to a column of nickel-agarose resin to increase the assay sensitivity for detecting methylated DNA markers in stool (40). Subsequently, multiple prototypes of MT-sDNA test were commercialized but only two were approved in August 2014 by FDA for screening people at average risk for CRC and having more than fifty years (29). To date, both the American Cancer Society and United Services Preventive Services Task Force (USPSTF) affirmed that MT-sDNA test can be repeated every three years to provide a decrease in CRC incidence and mortality with an acceptable cost and have approved this test for screening people aged forty-five to forty-nine years old (41, 42).

Moreover, Heigh *et al.*, performed a targeted single assay test with aberrant methylation of BMP3 alone and detected sessile serrated polyps (SSPs) with a sensitivity of 66% and a specificity of 91% (43) but, although additional biomarkers can be used by including multiple targets that reach 21-target MT-sDNA test, no increase in the sensitivity or

specificity was observed (44). In general, most studies agree that MT-sDNA is effective to detect CRC with only a few exceptions; in fact, Imperiale and colleagues detected 60 out of 65 colon cancers by MT-sDNA test with an estimated sensitivity of 92.3% and a specificity of 90%, confirming that MT-sDNA test is more sensitive than FIT, especially for the detection of lesions with high-grade dysplasia or sessile serrated polyps (≥ 1cm) (16). Overall, the method sensitivity varied from 62% to 91% for cancer and from 27% to 82% for advanced adenomas, with a specificity of 93% to 96% in people with normal findings on colonoscopy (45). Surely, the advancement of genetic knowledge in CRC and their related mutational events would improve the efficiency and the sensibility of MT-sDNA test by increasing the target DNA genes. Nowadays, MT-sDNA tests include quantitative molecular assays for KRAS mutations, NDRG4 and BMP3 methylation, β-actin and include eleven different DNA sequences commonly seen in colon polyps/cancers (46). Therefore, as confirmed by a retrospective study conducted by Weiser *et al*. on 368.494 subjects, MT-sDNA test represents the most recommended CRC screening tool because of its widespread accessibility and higher sensitivity compared with other previously described methods such as FIT and FOBT (Table1) (47).

5.3 Droplet Digital PCR

Droplet digital PCR (ddPCR) is recognized as an established and trustworthy approach for clinical cancer research due to its high sensitivity (almost 74% for CRC) in comparison to traditional standard procedures, even in degraded samples (48) (Figure 2). This method consists of an enrichment strategy that allows the detection of low-level mutations by amplification of single DNA molecules without the need for standard reference curves and it is considered much easier, faster, and less error-prone than real-time qPCR (49). Nowadays, ddPCR is commonly used for detecting rare alleles as molecular markers in plasma samples of pre- and post-operative CRC patients not only because of its high sensitivity for detecting tumor DNA (even only a very small fraction or degraded DNA) but also to monitor disease progression and the emergence of drug resistance (50). Through this method, Taly *et al*. have documented seven common mutations in codons 12 and 13 of the KRAS oncogene from plasma samples of CRC

patients, demonstrating the clinical utility of multiplex ddPCR to screen multiple mutations with a sensitivity sufficient to detect mutations in circulating DNA obtained by non-invasive blood collection (51). In the same context, ddPCR platforms using OncoBEAM technology demonstrated a high sensitivity for plasma detection of KRAS mutations (52) and, overall, ddPCR has been largely applied to the detection and quantification of mutated genes including the aforementioned KRAS (53) but also BAT26 (54), ITGA6, ITGA6A (55) and hypermethylated GRIA4, VIPR2 (56) and VIM (57) from both circulating tumor DNA (ctDNA) or fecal DNA of CRC patients. Recently, Garrigou *et al*. proposed the screening of modifications in methylated ctDNA as a biomarker to monitor tumor evolution of CRC patients at different stages and concluded that it could be a universal approach to follow tumor burden of CRC patients, as compared with mutated ctDNA, which requires previous tumor mutation identification (58). To summarize, although there are many advantages of ddPCR including the high sensitivity and their large range of target mutations, its major limitation is represented by the lower availability of primer/probe sets (Table1) (59).

5.4 The Idylla approach

The Idylla system (Biocartis, Mechelen, Belgium) consists of a cartridge-based fully automated medical device able to perform an innovative technology that consists of a conventional TaqMan reporter system and a novel chemistry known as PlexPCR (amplicons containing a small region with a sequence different from that of target DNA) simultaneously with a PlexZyme (specific amplicon sequence-matched reporter probe) that allows multiplexing of numerous gene targets in one assay (60) (Figure 2). Hence, thanks to its ability to easily detect a wide range of CRC-related mutations, the Idylla approach can be easily implemented in pathology laboratories to reduce turnaround time (61) and it currently represent a feasible and validated test for KRAS, NRAS and EGFR mutations in FFPE (Formalin-Fixed Paraffin-Embedded) tissues (62) and for BRAF hotspot mutation in plasma samples (63). In addition, the Idylla system can be also used to confirm within a few hours uncertain outcomes of doubtful NGS results and/or in case of scarce tissue material; for instance, Zwaenepoel *et al*.,

evaluated the clinical performance of the Idylla method in 330 CRC samples and demonstrated that this technology was able to give results in less than 2.5 h with only two invalid results (64). Many authors tested the full panel of CRC gene targets (BRAF, KRAS and NRAS) and found that the concordance between Idylla and NGS was 100% for BRAF and KRAS mutations and 94% for NRAS (65). Therefore, this methodology is highly accurate to detect frequent mutations and minimize the contamination risk, in addition to reducing cost per test when compared with NGS or some conventional PCR assays. However, rare and/or complex genomic variants, which are not included in the reference ranges, cannot be detected by the Idylla system and so continuous improvement of its biomarkers panel is necessary to guarantee efficient diagnosis (66).

6. Methods based on NGS technologies

Since the 2000s, and in coincidence with the emergence and development of new high-throughput sequencing technologies, many analyses have been undertaken to examine genetic susceptibility to diseases through genome-wide association studies (GWAS). Regarding CRC, Zanke *et al.*, using a multistage genetic association approach comprising 7.480 CRC patients and 7.779 controls, recognized a wide association of markers in chromosomal region 8q24, the same site where the SMAD7 gene is located (67). In addition, a GWAS performed by Broderick *et al.*, consisting of the genotyping of 550,163 tag SNPs (Single Nucleotide Polymorphisms) in 940 individuals with familial CRC and 965 controls, identified three SNPs in the SMAD7 gene (68). Subsequently, Tomlinson *et al.* confirmed these results and elucidated other markers in chromosomal regions of 8q23.3 and 10p14 at which common variants can influence the risk of CRC developing (69). In the last years, NGS-based diagnostics essays are increasingly adopted especially with the dropping sequencing costs. In the early stage, sequencing technologies were used to target driver genes known to contribute to CRC, but recently larger chromosomal regions have been targeted exploiting the potential of these technologies in multigene sequencing by using a very low amount of biological material from liquid or tissue biopsy samples. In this step, many efforts have been focused to standardize sequencing procedures and data analyses and then generating databases

aiming to store the sequencing information and contribute to future use by clinicians and research communities to better quality care (70). Early in 2010, The Cancer Genome Atlas project (TCGA) conducted a genome-scale analysis of samples obtained from 276 CRC patients, analyzing exome sequence, DNA copy number, promoter methylation and messenger RNA and microRNA expression and concluded that 16% of CRC samples were found to be hyper-mutated, 77% of patients displayed one or both breakpoints leading to translocation in an intergenic region and 7% of patients reported a translocation involving TTC28 gene (an inhibitor of tumor cell growth) located on chromosome 22 (71). Furthermore, the Pan-Cancer Analysis of Whole Genomes (PCAWG), the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) projects recently described 2,658 whole-genomes of tumor samples and their matching normal tissues, not only of CRC but of 38 different cancer types, providing insights into the nature and timing of the many mutational processes that shape large and small-scale somatic variation in the cancer genome (65).

To date, according to the improvement of NGS approaches, different sequencing platforms have been developed (Illumina, Ion Torrent, SOLiD, PacBio and Nanopore) that, overall, are classified in terms of maximum output, reads per run, accuracy, run time, amount of nucleic acids necessary for analysis and reads length; in particular, they can generate short (e.g. SOLiD, Ion Torrent, Roche 454 pyrosequencing, Illumina) or long reads (e.g. PacBio, Nanopore). In particular, they can generate short (*e.g.* SOLiD, Ion Torrent, Illumina) or long reads (*e.g.* PacBio, Nanopore) and while short reads sequencing does not exceed 300 base pairs and is more suitable for CRC diagnosis, a long reads sequencing determine a better coverage of the genome and it is more adaptable for large deletions/insertions determination or chromosomal rearrangement (72). Considering that both short-read and long-read sequencing have their benefits and flaws, depending on the experimental aim, it is important to remark that when somatic alterations in oncogenes and tumor suppressor genes are stable throughout the tumor clonal evolution, chromosomal alterations and copy number variation (CNV) could be lost during cancer progression (73).

In addition, the CRC represents one of the most interesting fields of NGS application because of its great quantity of activating mutation; in fact, next-gen techniques enable the identification of novel mutations/altered genes or genomic rearrangements allowing the discovery of new possible treatments (74). In general, there are three more commonly NGS-based methods used for CRC studies which are custom panel, whole genome (WGS) or exome (WES) sequencing and third-generation sequencing approaches. In general, larger-scale mutations were identified by whole-genome sequencing of tumor DNA while point mutations were identified by targeted sequencing (Table1).

Custom panel sequencing

During the last decade, several pipelines based on NGS approaches have been developed and additional somatic mutations and chromosomic aberration were detected in CRC samples (Figure 3). To simplify routine adoption of NGS tools, Zheng *et al*. considered a custom-designed panel of genes of only 2.2 Mb (exons and partial introns of cancer driver of more than 600 genes) and deduced a 9-loci model for detecting microsatellite instability with 100% sensitivity and specificity compared with microsatellite instability (MSI-PCR), and 84.3% overall concordance with immunohistochemistry (IHC) staining (75). Many authors have undertaken to sequence simultaneously many driver genes including low allele frequencies using NGS technologies and emphasize the importance of the fine classification of mutational status as some cancers were associated with poor prognosis treatment (76). In this regard, the comprehension of the wide heterogeneity of CRC lesions seems to be an extremely important point for tracing the therapeutic approach of the patient and developing effective strategies for early CRC detection and prevention. To date, liquid biopsy samples have been more investigated than tumor tissue sample because of their non-invasiveness and their better representation of cancer heterogeneity (77). In this context, Myint *et al*., developed a multiregional NGS approach from circulating cell-free DNA (cfDNA) using a customized targeted CRC panel consisting of all coding exons of 116 genes, 22 genes recurrently amplified/deleted, 51 copy number regions, 121 MSI

regions and 2 gene fusions (RSPO2 and RSPO3) and confirmed the widespread genetic heterogeneity in six adenoma samples, which affected driver genes MMR, APC, PIK3CA, TP53 and SMAD4 (78). Additionally, based on an NGS analysis of a panel of 324 CRC-associated genes, Stahler *et al*. have documented frequent single nucleotide variations (SNVs) in TP53, APC, KRAS, PIK3CA, BRAF, SMAD4 and FBXW7 genes, and copy number alterations (CNAs) in MYC and FLT3 genes (79). Furthermore, Leary *et al*., developed a "personalized analysis of rearranged ends" (PARE) approach, which can identify translocations and copy number alterations in CRC and other solid tumors; in addition, PARE can detect 57 regions containing putative somatic rearrangements, with an average of 14 rearrangements per sample (80). Moreover, also targeted sequencing strategies based on short reads and CNV determination could represent a good strategy for CRC studies. In fact, Gould *et al*. confirmed that a NGS approach using short fragments presents a sensitivity >96% and a specificity >99% for detecting samples with CNVs in the terminal five exons of PMS2 (81). Additionally, Corti *et al*., have developed multiple DNA NGS approaches coupled with the computational and bioinformatics algorithm "IDEA" to target a WES of about 30 Mb, a custom panels of genes IRCC-TP of about 603Kb (frequently mutated genes) and other of 918Kb (intron-exon junction to precisely identify the genomic breakpoint) (82). Currently, IDEA represents a flexible and comprehensive pipeline for the management of CRC patients, also suitable for identifying several genetic alterations from a non-invasive sample (ctDNA) such as single nucleotide variants, insertions and deletions, gene copy-number alterations, and chromosomal rearrangements in KRAS, BRAF, PIK3CA, and ERBB2 genes, that are usually involved in drug-resistance. In general, sequencing of smaller target regions provides greater sequencing depth which allows for better recognition of low gene frequency variation. Hence a customized gene approach is more suitable for clinical oncology laboratories for many advantages such as the simplicity, low cost and fast of the method and the non-need for bioinformatics specialists in the laboratories (Table1).

Whole exome and whole genome sequencing

The contribution of microsatellite instability to the tumor mutational burden (TMB) due to a defective mismatch repair system (dMMR) is considered important in about 15% of CRC patients. According to the phenotype, microsatellite unstable tumors can be divided into two distinct MSI phenotypes: MSI-high (MSI-H) and MSI-low (MSI-L) (83). Recently, considering that the dMMR phenotype is crucial to define the efficacy of immune checkpoint inhibitor treatment, Xiao *et al.*, used the WES to evaluate the immune microenvironment and 2539 microsatellite locus in a group of 98 CRC patients and concluded that the microenvironment of high tumor mutation burden (TMB-H) was significantly more immune-responsive than low TMB (TMB-L) (84). On the other hand, Gurjao *et al.* demonstrated the presence of a novel alkylating mutational signature identified through WES of 900 CRC patients and predicted that KRAS p.G12D, KRAS p.G13D and PIK3CA p.E545K driver mutations were mainly targeted by the alkylating signature in non-hypermutated patients (85). Moreover, Chang *et al.*, performed a WES of DNA obtained from tumor tissues of 32 surgical CRC patients and identified not only the well-known recurrent mutations in APC, TP53, KRAS and FBXW7 genes but also unreported mutations in additional 14 genes (86). Furthermore, many authors confirmed that the WGS largely contributed to determining the significant role of non-coding regions such as enhancers, transcription factor binding sites, promoters and 3'UTR (untranslated region) in CRC carcinogenesis (87). In addition, WGS was used to demonstrate that metastatic lesions are enriched in gene mutations affecting PI3K-AKt signaling, cell adhesion and extracellular matrix processes (88). Finally, Dashti *et al.* have conceived a new pipeline based on a novel concept called 'gene-motif', which identified seven CRC subtypes that can be effectively used to develop a personalized treatment (89). In general, in comparison to WES, the WGS approach has the advantage to increase the overall variant calling accuracy and poor coverage but comes more expensive and requires fresh-frozen tumor material to perform analysis with the highest quality (Figure 3) (Table 1).

Third-generation sequencing

Nowadays, third-generation sequencing of long reads have been developed and represents the most suitable approach for the identification of deletion/duplication breakpoints and of complex SVs and CNV-neutral rearrangements such as inversions and large intronic insertions (90) (Figure 3). Indeed, many studies affirmed that long-read sequencing technologies have potential advantages over existing alternatives especially when pathogenic variants are in complex genomic regions, such as the recurrent PMS2 insertion-deletion variant. Using a locus-specific amplicon template, Watson *et al*. undertook a Nanopore long-read sequencing to assess the CRC diagnostic accuracy of this platform; pairwise comparison between sequencing results derived from short-read NGS and unidirectional Sanger sequencing and the consensus Nanopore dataset revealed 100% sequence identity (91). Furthermore, reads produced by Nanopore oxford technology were able to identify both the 5′ and 3′ junctions and revealed detailed insertion sequence information (92).

7. Metagenomics analysis of gut dysbiosis in CRC patients

Genetic factors that concern somatic mutations in KRAS, APC oncogene, p53 tumor suppressor genes or mismatch repair genes, and other chromosomal aberrations explain less than 35% of all diagnosed CRCs and many environmental exposures seem to modulate the cancer risk (93). For instance, metagenomics studies based on 16S rRNA sequencing that has been conducted in the last years have documented the presence of more than a thousand microbial species in the human gastrointestinal tract carrying more than 100 times as many genes than the human genome (94) (Figure 3).

Therefore, considering the high microbial diversity in humans and their contribution to the host health and pathological or malignant conditions, it was suggested that about 20% of the global cancer burden can be linked to microbial agents (95). However, in addition to the several factors that can considerably modify the GM composition (age, gender, nationality, dietary and lifestyle habits, drugs or alcohol abuse) (96), multiple experimental challenges can influence the results of GM studies such as sampling methods and consistency (97), storage sample conditions (98), DNA extraction methods (99), type of primers used and pipelines adopted for data analyses (100). For all these

reasons, it is very hard to define a baseline microbial community for healthy people, especially for the impossibility to obtain biopsy samples from healthy controls and so the consideration of the tumor-adjacent tissue as a healthy control, but many efforts have been expended in the last year to standardize the experimental and analytical methods (101). Nowadays, the two most common metagenomics approaches for GM characterization are shotgun sequencing and metabarcoding. These NGS-based approaches both contains three basic steps that are library preparation, sequencing, and data analysis. Learn the basics of each step and discover how to plan your NGS workflow.

These NGS-based approaches both contain three basic steps that are library preparation, sequencing, and data analysis. Sequencing libraries are typically created by fragmenting DNA and adding specialized adapters to both ends to allow the DNA fragments to bind to the sequencer' flow cell. Thanks to unique barcodes added to each library, that are used to distinguish between the libraries during data analysis, multiple libraries can be pooled together and sequenced in the same run (a process known as multiplexing). During the next sequencing step of the NGS workflow, the sequencer amplifies the DNA fragments, resulting in millions of copies of single-stranded DNA. In detail, chemically modified nucleotides bind to the DNA template strand through natural complementarity, each nucleotide contains a fluorescent tag and a reversible terminator that blocks the incorporation of the next base. The fluorescent signal indicates which nucleotide has been added, and the terminator is cleaved so the next base can bind; after reading the forward DNA strand, the reads are washed away, and the process repeats for the reverse strand. After sequencing, the instrument software identifies nucleotides (a process called base calling) and the predicted accuracy of those base calls. At last, data analysis can be performed with standard analysis tools or with customized pipelines (102). In general, , focusing on the shotgun sequencing and metabarcoding; the first can simultaneously provide functional and taxonomic information about bacteria, fungi, viruses, and a variety of other microorganisms but with a complicated data output (*i.e.* a huge amount of information that can be up to 1.5 terabases per run)

while the second has a less complex data output but provides only taxonomic information about bacterial (16S region sequencing) or fungal (ITS sequencing) composition of the sample (Table1) (103). Thus, metagenome-wide association studies have identified an association between many microbial species/genes markers and CRC, promoting the development of an affordable diagnostic test using both stool or tissue samples (104, 105).

CRC-associated bacteria

Currently, much evidence documented the GM involvement in different diseases, including CRC; in particular recent evidence has demonstrated a bacterial driver-passenger model for CRC initiation and progression and reported that the first epithelial transformations can be supported by certain intestinal bacteria (106). In detail, early in 2010, Tjalsma *et al*., proposed a bacterial driver-passenger model for CRC in which pathogenic driver bacteria interact transiently with host cells to initiate CRC development and then replaced by other passenger bacteria species that were unable to colonize healthy colon tissue but got benefits from altered metabolism of tumors cells (107). To date, Wang *et al*., have identified *Bacillus spp*., *Bradyrhizobium spp*., *Methylobacterium spp*., and *Streptomyces spp*. as a potential driver bacterium while *Fusobacterium spp*., and *Campylobacter spp*. as certain and abundant passenger-bacteria (108). Moreover, Luan *et al*. have characterized the mucosa-adherent fungal microbiota of paired biopsy samples of adenomas and adjacent healthy tissue from 27 subjects by using barcoded high-throughput sequencing that targeted ITS region and reported a different fungal composition in patients with different adenoma stages and identified the phylum Glomeromycota as a possible powerful colorectal marker (109). Consistently, recent findings obtained through the WGS approach demonstrated that also the Ascomycota/Basidiomycota ratio could represent a potential novel marker for early CRC detection (110). Furthermore, Coker *et al*. used a shotgun metagenomics approach to evaluate the role of archeome to colorectal carcinogenesis and found distinct archea clusters in fecal samples from CRC patients, patients with adenomas,

and healthy subjects; the CRC patients showed a significant enrichment of halophilic archaea and depletion of methanogenic archaea (111). Anyway, several metagenomics analyses of CRC patients have documented an over-representation of *Fusobacterium nucleatum* in both tissue or stool samples, in comparison to healthy controls (112). Interestingly, in a large-cohort of 616 participants, Yachida *et al*. have demonstrated that the shift in the GM composition between CRC patients and healthy controls occurred in the very early stages of the CRC development; in particular, the relative abundance of *F. nucleatum* was significantly elevated continuously from intramucosal carcinoma to more advanced stages while *Atopobium parvulum* and *Actinomyces odontolyticus* were significantly increased only in multiple polypoid adenomas and/or intramucosal carcinomas (113). Recently, in addition to *F. nucleatum*, several bacteria have been reported to be enriched in stool or tissue samples of CRC patients compared to healthy ones, such as *Bacteroides fragilis*, *Escherichia coli*, *Streptococcus bovis*, *Enterococcus faecalis*, *Peptostreptococcus anaerobius* and *Lachnoclostridium spp.* (114-118). Moreover, an association between specific bacterial species are and antitumor responses have been reported; for instance, a positive correlation between the abundance of *Bifidobacterium longum* or Ruminococcaceae members and the efficiency of CRC immunotherapy has been documented (119), while *Eubacterium limosum, Ruthenibacterium lactatiformans*, *Fusobacterium ulcerans*, *Bacteroides uniformis*, *Paraprevotella xylaniphila* and *Alistipes senegalensis* resulted to improve the effectiveness of immune checkpoint inhibitors (120).

Therefore, considering that GM composition can be modified by probiotic and prebiotic supplementation, which can help maintain intestinal microbial homeostasis and mitigate dysbiosis, many reports have evaluated their effect on colorectal carcinogenesis. Overall, recent systematic reviews suggested that prebiotics may have a protective effect on the progress of CRC while the administration of certain probiotics in patients with CRC reduced the side effects of chemotherapy, improved the outcomes of surgery, shortened hospital stays, and decreased the risk of death(121, 122). However, the findings are still conflicting and none of them determined changes in bacterial

richness and diversity that are usually reduced in CRC patients, thus further studies are needed to better understand the prebiotics and probiotics effects in CRC patients.

CRC-associated bacterial metabolites

Accumulating evidence suggests that GM modulates the CRC progression, and its metabolites can play a crucial role in this scenario. In fact, the rapid development of technologies such as mass spectrometry and nuclear magnetic resonance have documented that the microbial metabolites are differently abundant between CRC patients and healthy subjects. For instance, lower bile acid hydrolase and β-galactosidase abundances and higher levels of leucine, tyrosine, valine, choline, colibactin, gallocin, formyl methionyl leucyl phenylalanine, *Bacteroides fragilis* toxin and trimethylamine-N-oxide have been associated with CRC development (123-125). Furthermore, the short chain fatty acids (SCFAs) abundances are reduced ì in CRC patients when compared to healthy controls. The SCFAs are the main metabolites produced by the bacterial anerobic fermentation of indigestible polysaccharides, exerting different and fundamental functions for the host. In detail, the total amount of SCFAs resulted significantly lower in fecal and plasma samples of CRC patients compared to both patients with adenomatous polyps and healthy controls, so this analysis could represent a novel potential non-invasive diagnostic tool for CRC (126, 127). In addition, recent investigations have reported that the GM plays a critical role in the effectiveness of anti-CRC treatments, including chemotherapy as well as immunosuppressive agents [186]. It has been reported that the effectiveness of CRC treatment with 5-fluorouracil is enhanced by certain microbial metabolites (128), so the supplementation with probiotics or prebiotics could increase chances of therapeutic success (129).

CONCLUSION

Since the advent of NGS approaches, various molecular techniques for the diagnosis of CRC from invasive or non-invasive sampling have emerged and have significantly

increased the number of known genes and mutations linked to CRC. However, due to the multitude of host and microbial genetic factors and the complexity of the tumor environment, the optimization of a CRC biomarker remains difficult, especially in stool samples, in which the complexity of the lesion environment seems to play a key role [50]. Thus, the development of a biological method to find stable markers in non-invasive samples such as feces or plasma, both sensitive and specific remains an arduous challenge to be carried out. For instance, in the plasma, many types of cells were targeted for CRC screening including, cell-free DNA (cfDNA) and circulating tumor DNA (ctDNA). However, during tumorigenesis and treatment the ctDNA is more strictly tumor specific than cfDNA and proves to be suitable and targeted for CRC diagnosis. In fact, the cfDNA rate shows a remarkable variation level in plasma probably due to DNA released from many other sources than just the tumor cells, including from normal cells dying due to the resection trauma, cells dying due to complications associated with surgery and releasing DNA during the handling of blood samples [61]. In addition, the frequency of mutant DNA alleles in cfDNA is as low as 0.01%, and more than 80% of patients with metastatic CRC show detectable levels of ctDNA higher than that reported for most conventional biomarkers [191]. Furthermore, despite the great progress in metagenomics methods and bioinformatics tools, to date, WES and WGS are still feasible only in expert centers, only limited pieces of genomic information are currently clinically relevant for the care of CRC patients and the list of predictive actionable genomic biomarkers is quite short [86]. However, apart the identification of novel microbial biomarker, different novel molecular biomarkers are under evaluation for CRC screening, such circular RNA (circRNA) and Piwi-interacting RNA (piRNA). These advances in the identification of microbial markers and the improvement of non-invasive diagnostic capabilities and their applications in guiding precision cancer therapies are poised to change the ways to diagnose the CRC and select and monitor the treatments in the close future, also thank to the increasingly adopted precision medicine for care of CRC patients.

81963_Auto_Edited.docx

ORIGINALITY REPORT

18%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	285 words — 4%
2	www.illumina.com Internet	147 words — 2%
3	pubmed.ncbi.nlm.nih.gov Internet	69 words — 1%
4	Arnaud Uguen, Giancarlo Troncone. "A review on the Idylla platform: towards the assessment of actionable genomic alterations in one day", Journal of Clinical Pathology, 2018 Crossref	46 words — 1%
5	link.springer.com Internet	43 words — 1%
6	www.nature.com Internet	43 words — 1%
7	www.researchgate.net Internet	40 words — 1%
8	stm.sciencemag.org Internet	26 words — < 1%
9	www.spandidos-publications.com Internet	

25 words — < 1%

10 prism.ucalgary.ca
Internet
24 words — < 1%

11 Ruth Vera, Vicente Alonso, Javier Gállego, Encarnación González et al. "Current controversies in the management of metastatic colorectal cancer", Cancer Chemotherapy and Pharmacology, 2015
Crossref
23 words — < 1%

12 discovery.ucl.ac.uk
Internet
22 words — < 1%

13 Giovanni Ponti, Marco Manfredini, Aldo Tomasi. "Non-blood sources of cell-free DNA for cancer molecular profiling in clinical pathology and oncology", Critical Reviews in Oncology/Hematology, 2019
Crossref
21 words — < 1%

14 Jingjie Zhang, Yonghui Yu, Jing Wang. "Protein Nutritional Support: The Classical and Potential New Mechanisms in the Prevention and Therapy of Sarcopenia", Journal of Agricultural and Food Chemistry, 2020
Crossref
21 words — < 1%

15 www.acepnow.com
Internet
21 words — < 1%

16 Sonia Garrigou, Geraldine Perkins, Fanny Garlan, Corinne Normand et al. "A Study of Hypermethylated Circulating Tumor DNA as a Universal Colorectal Cancer Biomarker", Clinical Chemistry, 2016
Crossref
20 words — < 1%

17	Carino Gurjao, Rong Zhong, Koichiro Haruki, Yvonne Y. Li et al. "Discovery and features of an alkylating signature in colorectal cancer", Cancer Discovery, 2021 Crossref	18 words —	< 1%
18	Kai Zheng, Hua Wan, Jie Zhang, Guangyu Shan et al. "A novel NGS-based microsatellite instability (MSI) status classifier with 9 loci for colorectal cancer patients", Journal of Translational Medicine, 2020 Crossref	17 words —	< 1%
19	backend.orbit.dtu.dk Internet	17 words —	< 1%
20	Simone Baldi, Marta Menicatti, Giulia Nannini, Elena Niccolai et al. "Free Fatty Acids Signature in Human Intestinal Disorders: Significant Association between Butyric Acid and Celiac Disease", Nutrients, 2021 Crossref	16 words —	< 1%
21	ir.vanderbilt.edu Internet	16 words —	< 1%
22	www.science.gov Internet	16 words —	< 1%
23	Arndt Stahler, Sebastian Stintzing, Jobst C. von Einem, Christoph B. Westphalen et al. "Single-nucleotide variants, tumour mutational burden and microsatellite instability in patients with metastatic colorectal cancer: Next-generation sequencing results of the FIRE-3 trial", European Journal of Cancer, 2020 Crossref	15 words —	< 1%
24	Artur T. Kopylov, Ekaterina V. Ilgisonis, Alexander A. Moysa, Olga V. Tikhonova et al. "Targeted	15 words —	< 1%

Quantitative Screening of Chromosome 18 Encoded Proteome in Plasma Samples of Astronaut Candidates", Journal of Proteome Research, 2016
Crossref

25 Leandro Güttlein, María R Luca, Federico Esteso, Cristóbal Fresno et al. " Liquid biopsy for and mutation testing in advanced colorectal cancer patients: the Argentinean experience ", Future Oncology, 2022 — 15 words — < 1%
Crossref

26 Xin Yi, Jianhui Ma, Yanfang Guan, Rongrong Chen, Ling Yang, Xuefeng Xia. "The feasibility of using mutation detection in ctDNA to assess tumor dynamics", International Journal of Cancer, 2017 — 15 words — < 1%
Crossref

27 Yan Lin, De-Xia Kong, You-Ni Zhang. "Does the Microbiota Composition Influence the Efficacy of Colorectal Cancer Immunotherapy?", Frontiers in Oncology, 2022 — 14 words — < 1%
Crossref

28 Theodore Chiang. "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24", Nature Genetics, 08/2007 — 13 words — < 1%
Crossref

29 mercuric.eu — 13 words — < 1%
Internet

30 translational-medicine.biomedcentral.com — 13 words — < 1%
Internet

31 Kiyoshi Yamaguchi, Rika Kasajima, Kiyoko Takane, Seira Hatakeyama et al. "Application of targeted nanopore sequencing for the screening and determination of — 12 words — < 1%

structural variants in patients with Lynch syndrome", Journal of Human Genetics, 2021
Crossref

32 Qi-You Huang, Fei Yao, Chuan-Ren Zhou, Xiao-Ying Huang, Qiang Wang, Hui Long, Qing-Ming Wu. "Role of gut microbiome in regulating the effectiveness of metformin in reducing colorectal cancer in type 2 diabetes", World Journal of Clinical Cases, 2020 — 12 words — < 1%
Crossref

33 Richard Colling, Lai Mun Wang, Elizabeth Soilleux. "Validating a fully automated real-time PCR-based system for use in the molecular diagnostic analysis of colorectal carcinoma: a comparison with NGS and IHC", Journal of Clinical Pathology, 2017 — 12 words — < 1%
Crossref

34 pesquisa.bvsalud.org — 12 words — < 1%
Internet

35 www.semanticscholar.org — 12 words — < 1%
Internet

36 www.tandfonline.com — 12 words — < 1%
Internet

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS