Genomic diversity and *BCL9L* mutational status in CTC pools predict overall survival in metastatic colorectal cancer

Joao M. Alves[1,2,#,], Nuria Estévez-Gómez[1,2,#], Roberto Piñeiro[3,4], Laura Muinelo-Romay[4,5], Patricia Mondelo-Macía[5], Mercedes Salgado[6], Agueda Iglesias-Gómez[7], Laura Codesido-Prada[7], Astrid Dlez-Martín[7], Joaquin Cubiella[7], David Posada[1,2,*]*

[1]CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[2]Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO.
[3]Translational Medical Oncology Group, Oncomet, Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain
[4]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain.
[5]Liquid Biopsy Analysis Unit, Translational Medical Oncology Group, Health Research Institute of Santiago de Santiago de Compostela (IDIS), Santiago de Compostela, Spain
[6]Department of Oncology, Hospital Universitario de Ourense, Research Group in Gastrointestinal Oncology-Ourense, Ourense, Spain
[7]Department of Gastroenterology Hospital Universitario de Ourense, Research Group in Gastrointestinal Oncology-Ourense, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Ourense, Spain

[#] These authors contributed equally to this work

***Corresponding authors:** Joao M. Alves (jalves@uvigo.gal), David Posada (dposada@uvigo.es).

Keywords: circulating tumor cells, liquid biopsy, intratumor genomic heterogeneity, colorectal cancer, precision medicine.

Competing interests: The authors declare no competing interests.

Abstract

The genomic diversity of circulating tumor cells (CTCs) and its clinical implications remain poorly understood. In this study, we characterized the mutational landscape of CTC pools stemming from 29 metastatic colorectal cancer (mCRC) patients and examined its relationship with disease progression. Our analysis revealed substantial variation in mutational burden among patients, with all CTC pools harboring non-silent mutations in key CRC driver genes. Importantly, higher genomic diversity in CTC pools was significantly associated with reduced overall survival. Furthermore, non-silent mutations in *BCL9L* emerged as a strong predictor of patient survival. Taken together, these findings underscore the potential of CTC genomic profiling as a promising prognostic tool in mCRC and highlight the need for further research into its clinical applications.

1. Introduction

Metastatic colorectal cancer (mCRC) remains one of the leading causes of cancer-related deaths worldwide (Morgan et al. 2023). Despite encouraging improvements in the detection and treatment of early-stage lesions (Zauber et al. 2012; Jodal et al. 2019), mCRC continues to have a dismal 5-year survival rate of roughly 14% (Shin, Giancotti, and Rustgi 2023), highlighting the urgent need for biomarkers that can effectively monitor disease progression, guide treatment decisions, and improve patient outcomes (Zygulska and Pierzchalski 2022; Oh and Joo 2020).

Multiple studies have now shown that tumors with higher levels of genomic diversity are associated with more aggressive disease (Joung et al. 2017), shorter recurrence times (Fernandez-Mateos et al. 2024), and poorer survival outcomes (McDonald et al. 2019). However, measuring genomic diversity from solid tissues often involves invasive and high-risk biopsy procedures. As an alternative, liquid biopsies from peripheral blood and other body fluids, such as saliva and urine, offer a minimally invasive procedure to measure tumoral genomic diversity through the analysis of circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) (Connal et al. 2023). Although most efforts in this regard have focused on the genomic characterization of ctDNA due to its relative ease of isolation (Calabuig-Fariñas et al. 2016), the analysis of CTCs provides a unique opportunity to study intact tumor cells, thus allowing for comprehensive whole-genome profiling of tumors (Lin et al. 2021). Moreover, CTCs are presumably responsible for metastatic seeding and can provide information for both primary and metastatic tumor development (Zhu et al. 2018; Castro-Giner and Aceto 2020).

In this context, CTC enumeration has already been validated as a reliable prognostic marker across several cancer types (Toss et al. 2014; Maly et al. 2019), including CRC (Miller, Doyle, and Terstappen 2010; Tsai et al. 2016). However, not all the CTCs within a patient are genetically identical, and initial sequencing studies have revealed that CTCs can exhibit significant genomic variation within patients (Mishima et al. 2017; Paoletti et al. 2018; Castro-Giner and Aceto 2020). The genomic diversity of a patient's CTC pool may hold clinical significance for several reasons. First, the CTC pool could serve as a proxy for intratumoral genomic diversity, which, as mentioned above, has already been associated with increased cancer aggressiveness. Second, CTCs are considered precursors of metastasis. As such, a genomically diverse CTC population may increase the potential for successful dissemination, colonization, and growth in new environments (Ahmed and Gravel 2018). Nevertheless, the clinical significance of CTC genomic diversity remains largely unexplored.

In this study, we evaluated the clinical relevance of CTC genomic variation in mCRC. Using whole-exome sequencing data stemming from CTC pools collected from 29 mCRC patients, we provide a comprehensive description of the mutational landscape of CTCs and show that the overall genomic variation of CTC pools and the mutational status of the *BCL9L* gene can help predict patient survival.

2. Material & Methods

2.1 Clinical cohort and blood collection

We enrolled 25 mCRC patients diagnosed between October 2017 and January 2022 at the Hospital Universitario de Ourense, Spain, with histologically proven CRC. We collected a 15mL blood sample for each patient, stored in Transfix CTC-TVT tubes (Cytomark, UK) at room temperature. For patients undergoing chemotherapy treatment, blood samples were obtained just before a new chemotherapy cycle. All samples were obtained and collected after written informed consent from all subjects using a protocol approved by the Clinical Ethics Committee of Pontevedra-Vigo-Ourense (2018/301 approved 19/06/2018).

2.2 CTC enrichment and PBMC isolation

We processed all samples within 96 hours of collection using the Parsortix® platform (ANGLE plc, UK), which traps CTCs due to their larger size and lower compressibility than blood cells. For each patient, 10 mL of whole peripheral blood was loaded into a Parsortix microfluidic device. We enriched each sample in disposable Parsortix cassettes with a gap size of 6.5 µm (GEN3D6.5, Angle Inc., Guildford, UK) and at 99 mbar, according to the manufacturer's guidelines. After separation, the captured cells were collected in 200 µL of PBS and stored at −80 °C.

To obtain the peripheral blood mononuclear cell (PBMC) fraction - to be used as "healthy controls" - we took the remaining 5 mL from each blood sample and performed Ficoll-Paque gradient centrifugation. We kept the PBMCs in RNAlater (Ambion, TX, USA) at -80 °C until genomic DNA (gDNA) extraction.

2.3 Whole-genome amplification of CTC pools

Given the large collection volume (~200 µl), we initially performed gDNA extraction of the CTC pools using the QIAamp DNA Blood Mini Kit (Qiagen, Germany) before performing whole-genome amplification (WGA) with the Ampli1™ kit (Menarini Silicon Biosystems, Italy). We carried out the WGA starting with 1 µl of gDNA and included positive (10 ng/µl REPLIg human control kit, Qiagen, Germany) and negative controls (DNase/RNase free water). We worked in a laminar-flow hood to avoid contamination and used a dedicated set of pipettes and UV-irradiated plastic materials. We evaluated the quality of the amplified product with the Ampli1 QC Kit (Menarini Silicon Biosystems, Italy), a PCR-based assay to establish DNA integrity. Samples that produced at least two of the four expected PCR fragments were selected for the following steps: increasing the total double-stranded DNA (dsDNA) with the Ampli1 ReAmp/ds kit (Menarini Silicon Biosystems, Italy) and WGA adaptor removal. The latter was carried out by incubating a mixture of 5 µl of NEBuffer 4 10X (New England Biolabs, MA, USA), 1 µl of MseI 50U/µl (New England Biolabs, MA, USA), 19 µl of nuclease-free water and 25 µl of dsDNA at 37 °C for 3 h, with a final step of enzyme inactivation at 65 °C for 20 min. Finally, we purified the samples with 1.8X AMPure XP beads (Agencourt, Beckman Coulter, CA, USA), quantified the DNA yield with Qubit 3.0 fluorometer (Thermo Fisher Scientific, MA, USA), and checked the amplicon size distribution with the D1000 ScreenTape System in a 2200 TapeStation platform (Agilent Technologies, CA, USA).

2.4 PBMCs gDNA isolation

In parallel, we used the QIAamp DNA Blood Mini Kit (Qiagen, Germany) to extract gDNA from the PBMCs. We estimated DNA yield using the Qubit 3.0 fluorometer (Thermo Fisher Scientific, MA, USA) and DNA integrity with the Genomic DNA ScreenTape Assay (Agilent Technologies, CA, USA).

2.5 Whole-exome sequencing

Sequencing libraries were constructed at the Spanish National Center for Genomic Analysis (CNAG; http://www.cnag.crg.eu) with the SureSelect XT and Agilent Human Exon v5 kits (Agilent Technologies, CA, USA). In total, 25 amplified CTC pools and 25 PBMCs gDNA samples were whole-exome sequenced (WES) at 100X and 60X, respectively, on an Illumina NovaSeq 6000 (PE100) at CNAG. In addition, we also included matched CTC pools and PBMCs WES datasets previously generated in our lab from four mCRC patients ("PP") (Alves et al. 2022) (Sequence Read Archive accession code PRJNA886718).

2.6 Data processing

After trimming amplification and sequencing adapters from the raw FASTQ files, we aligned the sequencing reads to the Genome Reference Consortium Human Build 37 (GRCh37) using the MEM algorithm in the BWA software (Li 2013). Following the GATK's standardized best-practices pipeline (Van der Auwera et al. 2013), we filtered out reads with low mapping quality, performed a local realignment around indels, and removed PCR duplicates.

2.7 Somatic variant calling

We identified somatic single nucleotide variants (SNVs) and short insertions and deletions (indels) for each patient using the paired-sample variant-calling approach implemented in MuTect2 software (Benjamin et al. 2019) (i.e., CTC pools (tumor) + PBMCs (healthy control)). We then applied GATK FilterMutectCalls to remove calls in problematic sequence contexts ("--orientation-bias-artifact-priors") or due to potential cross-sample contamination ("--contamination"). Genotypes supported by < 10 total reads, with < 2 alternative reads, or with a variant allele frequency (VAF) ≤ 0.075 were set as missing, and variants composed only of missing genotypes were removed. We annotated the variants with Annovar (v.20200608) (Wang, Li, and Hakonarson 2010). In addition, and for each patient, we estimated somatic copy number gains, losses, loss-of-heterozygosity (LOH) events, tumor purity, and global ploidy status with Sequenza (Favero et al. 2015) under default settings.

2.8 MSI status

Furthermore, we identified the microsatellite stability status of each patient with MSIsensor-pro (Jia et al. 2020), following the recommended 'best practices' for paired-sampling analysis. We classified patients with more than 30% of microsatellites mutated as microsatellite instable (MSI), while all other patients were categorized as microsatellite stable (MSS), as in Heide *et al.* (2022).

2.9 Mutational signatures

We ran the SIGNAL web tool (https://signal.mutationalsignatures.com/) (Degasperi et al. 2020) under default parameters to identify the single base substitution (SBS) signatures active in each patient (Alexandrov et al. 2020). As recommended by the authors, SBS fitting was performed using candidate SBS signatures from CRC.

2.10 Estimation of genomic diversity in CTC pools

The genomic diversity of a CTC pool can be measured in multiple ways depending on the type and features of the variants selected. Here, we computed 15 diversity metrics, including mutational burden (MB), the proportion of aberrant genome (PAG) (Schumacher et al. 2017), the mutant allele tumor heterogeneity score (MATH) (Mroz and Rocco 2013), number of SNVs and indels at CRC driver genes, number of copy number alterations (CNAs) at CRC driver genes, and mutation status at recurrently mutated CRC driver genes. Since our cohort only included one MSI patient, all downstream analyses were restricted to the MSS cases (N=28).

2.10.1 Mutational burden (MB)

For each patient, we measured the CTC pool mutational burden (MB) as the total number of SNVs that passed our filtering thresholds, normalized by the size of the exome captured. We report MB in units of number of SNVs / megabase.

2.10.2 Proportion of aberrant genome (PAG)

We also measured, for each patient, the proportion of the autosomal genome affected by copy number gains, losses, and LOH events.

2.10.3 Mutant allele tumor heterogeneity (MATH)

We estimated for each patient the CTC-pool mutant-allele tumor heterogeneity (MATH) score (Mroz and Rocco 2013), which is based on the distribution of SNV and indel allelic frequencies and calculated as:

$$MATH = 100 \times \frac{median\ absolute\ deviation\ (VAF)}{median\ (VAF)}$$

Where VAF is the variant allele frequency of SNVs and indels. Since MATH scores are sensitive to unreliable VAF at sites with insufficient sequencing depth (Makrooni, O'Sullivan, and Seoighe 2022), we only considered SNVs and indels with ≥ 25 total reads and ≥ 5 alternative reads.

2.10.4 Number of SNVs and indels at CRC driver genes (CRC-mut)

We additionally measured the total number of the non-silent SNVs and indels (i.e., mutations that alter the amino acid sequence of a protein and may affect its function) overlapping with CRC driver genes at the IntOGen portal (release 2023) (Gonzalez-Perez et al. 2013).

2.10.5 Number of CNAs at CRC driver genes (CRC-CNA)

We estimated the number of copy number alterations (CNAs), which includes gains, losses, and LOH events in the CTC pools overlapping with CRC driver genes at IntOGen (release 2023).

2.10.6 Mutation status at CRC driver genes

For CRC driver genes with recurrent non-silent SNVs/Indels in more than four patients, we additionally considered their mutation status (i.e., unmutated vs. mutated) as a diversity metric.

2.11 Correlation of CTC genomic diversity with overall survival

We correlated the above diversity metrics with overall survival (OS). To select which metrics should be included in a multivariate Cox proportional hazards (CPH) analysis, we assessed them individually in a univariate CPH analysis. We used the raw values for the continuous metrics. For the discrete ones (i.e., mutation status and gender), we used "unmutated" and "female" as reference groups, respectively.

Additionally, we created a binary version of each continuous metric (i.e., bMB, bAge, bLoT, bPAG, bMATH, bCRC-mut, and bCRC-CNA) for inclusion in a univariate analysis by splitting the patients into two groups given an optimal threshold. This threshold was determined using CutoffFinder (Budczies et al. 2012), which fits a CPH model to the explanatory (diversity metric) and response (OS) variables, identifying the optimal cutoff as the value that provides the most statistically significant split based on a log-rank test (**Table S1**).

Adopting a similar strategy to Fernandez-Mateos *et al.* (2024), we only included diversity metrics with a p-value ≤ 0.1 in the multivariate analysis. Importantly, when the continuous and binary versions of a metric returned p-values ≤ 0.1, only the continuous version was included in the multivariate analysis. All CPH

analyses were performed using the survival (Therneau 2001) and survminer (Kassambara, Kosinski, and Biecek 2016) R packages.

2.12 Evaluating custom and commercial cancer gene panels for CTC genomic diversity estimation

To identify a cost-effective strategy for the genomic characterization of CTC pools, we constructed and evaluated two customized gene panels, each consisting of a subset of the most frequently mutated genes in our cohort. These panels were tested to determine their ability to provide diversity estimates comparable to those derived from WES. The first panel, CRC-P22, comprised 22 genes harboring > 15 mutations across patients. The second panel, CRC-P55, included 55 genes with > 10 mutations across patients. In addition to these custom panels, we assessed the performance of two commercially available NGS panels, the Pillar® oncoReveal Solid Tumor v2 Panel (48 genes; Pillar Biosciences, MA, USA) and the Oncomine Precision Assay (50 genes; Thermo Fisher Scientific, MA, USA). A list of the genes from the four panels is provided in **Table S2**.

We extracted the SNVs and indels for each panel and calculated the MATH scores for each patient as in Section 2.10.3. Importantly, given the much smaller number of variants in these panels, we did not apply any filter based on read counts, as previously done. Finally, we assessed the clinical relevance of these panels by correlating their derived MATH scores with OS using a univariate CPH analysis. As before, we additionally tested a binary MATH (bMATH) version by dichotomizing patients at optimal threshold values determined using CutoffFinder.

2.13 Declaration of AI-assisted copy editing

During the preparation of this work we have used ChatGPT 4.0 turbo to improve language and readability. After using this tool/service, we reviewed and edited the content as needed and took full responsibility for the content of the publication.

3. Results

3.1 Genomic characterization of CTC pools from the mCRC cohort

A brief overview of the cohort is shown in **Fig. 1A**. The comprehensive clinical data for this cohort, including tumor and treatment details, is available in **Table S3**. Ten (34.5%) patients were female, and 19 (65.5%) were male, with a mean age at the time of mCRC diagnosis of 66.28 ± 16.33 years.

After sequencing, we identified one patient (PP5) exhibiting microsatellite instability (MSI) (**Fig. 1B**). As expected, this patient displayed a substantially higher SNV and indel burden (**Fig. 1C**). Specifically, for PP5, we called a total of 1,224 SNVs and 5,509 indels. In contrast, we estimated 116-429 SNVs and 331-2,043 indels for the other patients.

We identified a large number of mutations in known CRC driver genes, including non-synonymous changes and stop codon gains in *APC*, *KRAS,* and *PIK3CA*. Notably, the number of mutations in CRC driver genes varied significantly within the cohort, ranging from two mutations in patients P1, P3, P10, and P18 (MSS) to 33 mutations in patient PP5 (MSI) (**Fig. 1D**). We also inferred a substantial number of CNAs, with considerable differences among patients (**Fig. 1E**). Consistent with previous reports (Heide et al. 2022), the MSI patient PP5 exhibited a predominantly diploid copy number profile.

We found ten single-base substitution signatures, including signatures commonly observed in CRC, such as SBS1 (age-related), SBS3 (homologous recombination deficiency) and SBS88 (associated with exposure to *E.coli* bacteria) (**Fig. 1F**). Nevertheless, there was a significant contribution of signatures of unknown etiology - SBS8, SBS93, SBS97, SBS121 and SBS123 - all of which are relatively infrequent in CRC.

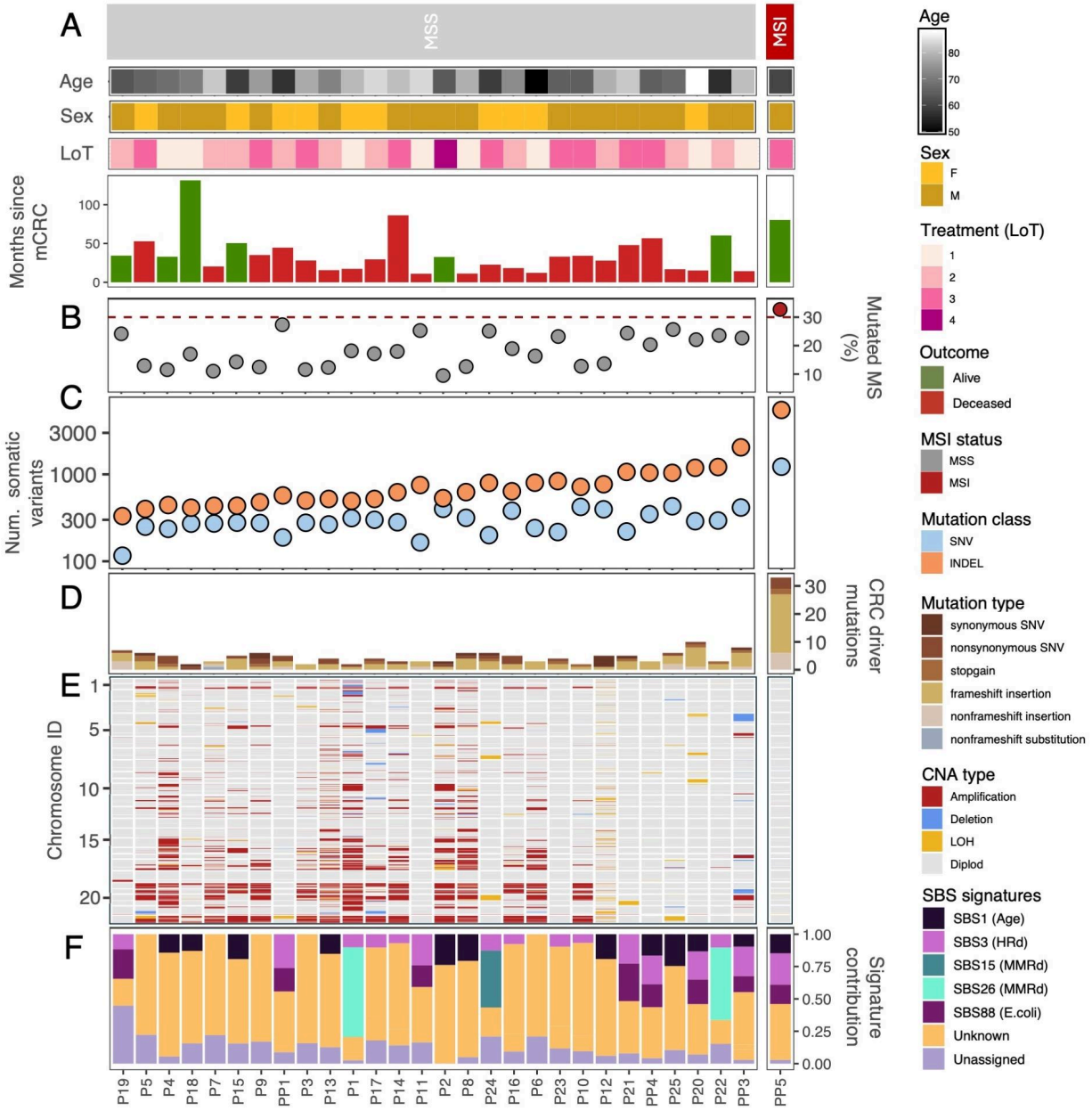

Figure 1. Clinical and genomic characterization of mCRC cohort. mCRC patients are split based on MS status and ordered according to their total number of somatic variants. **A.** Age, sex, number of lines of treatment (LoT), and months since mCRC diagnosis. **B.** Proportion of mutated microsatellites. Data points are colored by the MS status (grey = "MSS", red = "MSI"). The dashed red line depicts the threshold for MSI classification (i.e., 30% mutated microsatellites). **C.** Total number of somatic mutations identified with MuTect2, including SNVs (light blue) and indels (orange). **D.** Total number of mutations affecting known CRC driver genes in each patient. Bars are colored according to the mutation type. **E.** Genome-wide copy number profiles for each patient. Only autosomes are shown, with chromosomes ordered

from 1 to 22. Genomic regions are colored according to CNA type. **F**. Contribution of the different mutational signatures. HRd = homologous recombination deficiency; MMRd = mismatch repair deficiency. SBS signatures with unknown etiology - SBS8, SBS93, SBS97, SBS121, and SBS123 - were collapsed into a single "Unknown" category.

3.2 CTC pools show high levels of genomic variation

We observed high levels of genomic variation in the CTC pools (**Fig. 2A**), with substantial variation among patients. The MB estimates ranged from 0.95 to 2.89 SNVs/Mb, PAG values from 0.05% to 40.5%, and MATH scores between 32.38 and 67.63.

In addition to the substantial variability in the number of non-silent SNVs in CRC driver genes observed among patients (**Fig. 1D**), we also identified several genes that were recurrently mutated across multiple patients (**Fig. 2B**). In particular, ten driver genes were independently mutated in at least four patients, including *ARID1A* (N=4), *BCL9L* (N=4), *BIRC6* (N=5), *FAT3* (N=5), *FAT4* (N=5), *KMT2C* (N=5), *PARP4* (N=4), *TAF1L* (N=7), *TSC2* (N=4), and *ZFHX3* (N=5). Similarly, CNAs encompassing CRC drivers were also widespread (**Fig. 2C**), with *ARID1A* (N=16), *MYH9* (N=17), and *PPP2R1A* (N=16) having the highest frequency of copy gains, while *EP300* (N=7) and *TSC2* (N=6) had the highest number of deletions and LOH events.

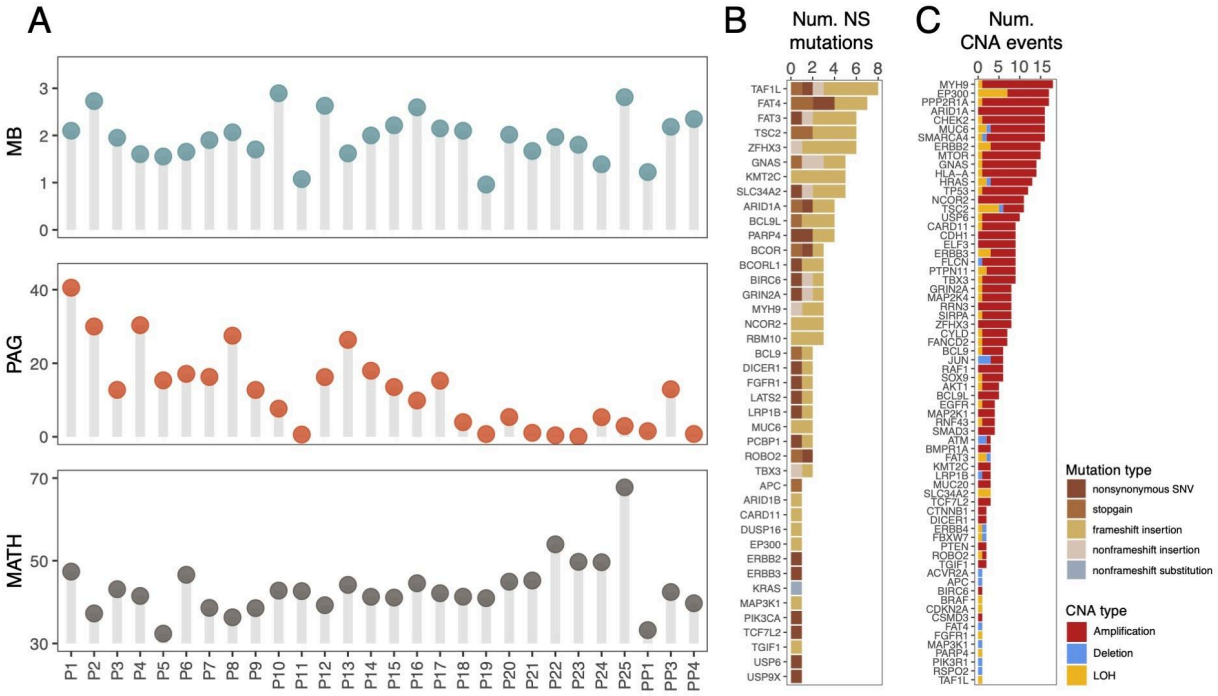

Figure 2. Genomic diversity in CTC pools. A. Lollypop plots depict the mutational burden (MB), the proportion of aberrant genome (PAG), and the mutant allele tumor heterogeneity (MATH) score for each patient. **B.** Number of non-silent (NS) SNVs and indels affecting known CRC driver genes. Bars are colored according to mutation type. **C.** Number of CNAs overlapping known CRC driver genes. Bars are colored according to CNA type.

3.3 MATH score and *BCL9L* status predict survival outcome in mCRC

The univariate analyses indicated that patient age is significantly associated with shorter OS, both when analyzed as a continuous variable (Hazard Ratio [HR]=1.07, 95% CI: 1.02–1.12, p=0.00894) and when dichotomized (HR=3.67, 95% CI: 1.39–9.67, p=0.00859) (**Fig. 3A**). Patients in the High-bMATH group

(HR=3.45, 95% CI: 1.36–8.75, p=0.00895) also had poorer OS compared to those in the corresponding "Low" groups.

Notably, a subgroup of patients (n=4) with non-silent mutations in *BCL9L* also displayed significantly worse OS (HR=3.71, 95% CI: 1.18–11.7, p=0.02). Importantly, *BCL9L* has been linked to increased genomic heterogeneity by promoting aneuploidy tolerance in CRC (López-García et al. 2017). In our cohort, patients with *BCL9L* mutations exhibited a higher fraction of CNAs, as reflected by elevated PAG estimates (**Fig. S1A**), but this did not translate into higher MATH scores (**Fig. S1B**). Similarly, patients (n=5) harboring non-silent mutations in *FAT4* displayed significantly worse OS (HR=2.80, 95% CI: 0.966–8.09, p=0.058).

A multivariate analysis including patient age, bMATH, as well as *BCL9L* and *FAT4* mutation status, confirmed that both high bMATH and mutated *BLC9L* status were significant prognostic factors, with HRs of 3.6 (95% CI=1.22–10.8, p-value=0.020) and 6.0 (95% CI=1.64–22.2, p-value=0.007), respectively (**Fig. 3B**). Accordingly, the survival curves indicate that patients in the Low-bMATH group survive much longer (median survival time: 52.9 months) than their High-bMATH counterparts (median survival time: 18.3 months) (**Fig. 3C**). Similarly, patients carrying *BCL9L* mutations experienced a notable reduction in OS (median survival time: 14.5 months) compared to those without them (median survival time: 33.1 months).

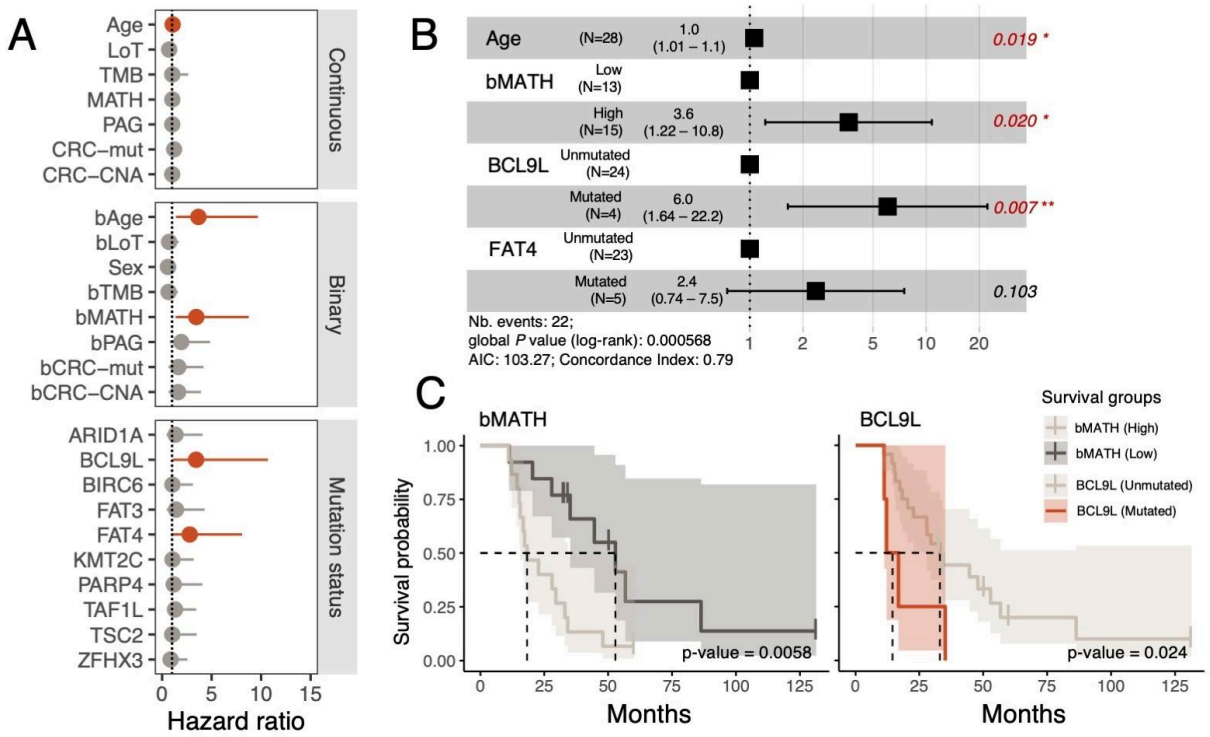

Figure 3. Impact of CTC genomic diversity on overall survival. A. Univariate CPH analysis for OS using all diversity metrics. Dots are the point estimate of the Hazard ratio, while the lines represent the 95% confidence interval (CI). The "Low" group was used as a reference in the binary version of all continuous variables. Hazard ratios with p-values < 0.1 are colored in red. **B.** Multivariate CPH analysis of survival time for the significant metrics and covariates in the univariate CPH analysis shown in A. The forest plot shows a 95% CI of HRs and the covariate p-values. Significant p-values are colored in red. **C.** Kaplan–Meier survival curves and p-values for bMATH and *BCL9L* on the mCRC cohort. The shaded area around each survival curve depicts the 95% CI.

3.4 CTC diversity from gene panels can also predict patient survival

To explore a practical and low-cost alternative for CTC genomic profiling, we designed two gene panels – CRC-P22 and CRC-P55 – each comprising the most frequently mutated genes in our cohort (**Fig. 4A**). The MATH scores derived from the CRC-P55 panel were highly correlated with those obtained from the WES data, whereas the CRC-P22 panel showed weaker concordance (**Fig. 4B-C).** Importantly, MATH estimates from the CRC-P55 panel were significantly associated with worse OS (**Fig. 4D**), both as a continuous variable (HR = 1.04, 95% CI: 1.01–1.07, p = 0.0223) and as a binary variable (HR = 4.83, 95% CI: 1.66–14.0, p = 0.00377). Patients in the high-bMATH group exhibited a significantly shorter median OS (20.4 months) compared to those in the low-bMATH group (52.9 months) (**Fig. 4E**). On the other hand, neither of the commercial panels showed a significant association between MATH scores and OS (**Fig. S2**).

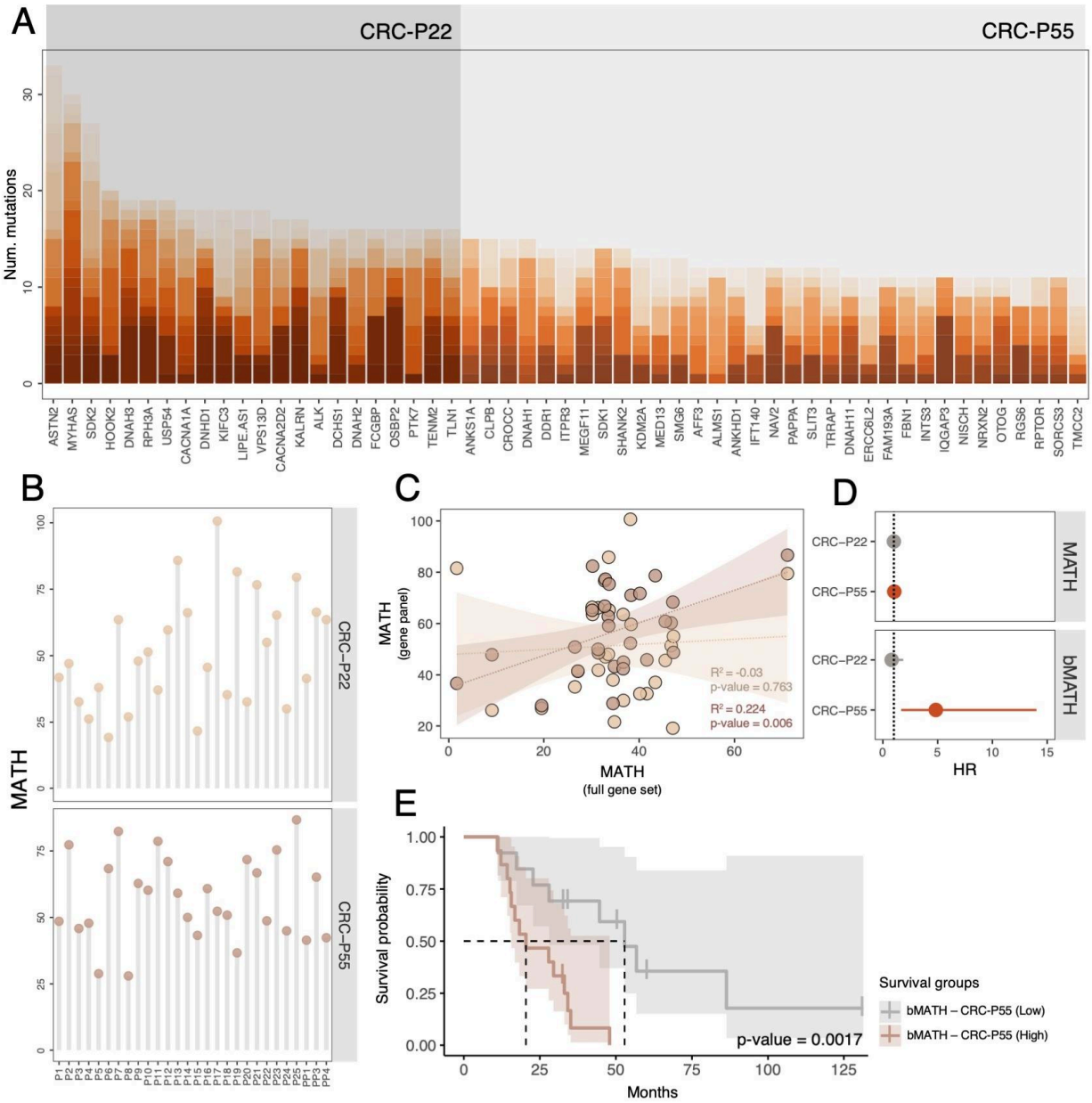

Figure 4. Optimized gene panel analysis. A. Genes included in the CRC-P22 and CRC-P55 panels were sorted according to mutation counts. Different color shades represent the different patients. **B.** MATH scores for each panel

and patient. **C.** Scatter plot describing the similarity of MATH scores between gene panels and WES. Solid lines represent the best fit from regression analysis. R^2 scores and *p-values* are shown on the bottom right side of the plot. Lighter and darker shades of brown highlight the MATH scores from CRC-P22 and CRC-P55, respectively. **D.** Univariate CPH analysis of survival time and MATH and bMATH for both gene panels. Dots are the Hazard ratio (HR) point estimate, while the lines represent the 95% CI. In the bMATH, the "Low" group was used as a reference. P-values < 0.05 are colored in red. **E.** Kaplan–Meier survival curves and p-values for bMATH of CRC-P55 gene panel. The shaded area around each survival curve depicts the 95% CI.

4. Discussion

The prognostic value of CTC enumeration has already been established, with higher CTC counts linked to aggressive disease across multiple cancer types, including CRC (Tsai et al. 2016). However, CTC populations can be genomically heterogeneous (Alves et al. 2022), a factor that, in the case of primary tumors, has already been linked to poor prognosis (Joung et al. 2017; Fernandez-Mateos et al. 2024; McDonald et al. 2019).

Here, we found that, in mCRC, CTC populations with higher MATH scores - a measure of genomic diversity - or with non-silent mutations in *BCL9L* are significantly associated with poorer overall survival. To our knowledge, this is the first study showing an association between CTC genomic diversity and cancer outcome. Intuitively, larger overall genetic variation, as measured by the MATH scores, likely reflects the presence of distinct subclones, some of which could proliferate more rapidly, have a higher propensity for metastasis, or better evade cancer drugs (McGranahan and Swanton, 2017). In this context, genomic diversity estimates derived from CTC pools may not only serve as prognostic biomarkers but might also contribute to real-time monitoring of disease progression (through longitudinal sampling). Increasing CTC heterogeneity over time could signal the emergence of resistant clones, potentially allowing for earlier clinical intervention. Conversely, a decline in diversity following treatment might suggest a favorable response.

Since the MATH scores in this study were derived from whole-exome sequencing - which may be impractical in clinical settings due to high costs and complex bioinformatics workflows (Bertier, Hétu, and Joly 2016) - we explored the potential of smaller gene panels as an alternative. Our findings revealed a strong correlation between MATH scores derived from a 55-gene panel and survival, offering a cost-effective strategy for clinical diagnostics and patient stratification.

Importantly, we also observed that non-silent mutations in *BCL9L* were associated with reduced survival. *BCL9L* has been recently identified as a driver of aneuploidy tolerance in CRC, thereby increasing genomic instability (López-García et al. 2017), which has already been proposed as a prognostic marker in several cancer types (Hosea et al. 2024).

Despite these promising findings, the relatively small sample size of our cohort may limit the generalizability of our conclusions, and larger cohorts will be needed to fully validate the relationship between CTC genomic data and patient outcomes.

In summary, our study demonstrates for the first time a relationship between CTC genomic diversity and overall survival in mCRC, underscoring the potential of genomic profiling of CTCs as a novel prognostic tool in mCRC and highlighting the need for further research into the clinical applications of CTC genomics, which could ultimately improve personalized treatment strategies for cancer patients.

Data availability

Raw exome sequencing data from CTC pools, together with matching healthy samples, have been deposited in the Sequence Read Archive database under the accession code ########. All data supporting the findings of this study are available within the article and its supplementary information files.

Author contributions

D.P. and J.M.A. conceived the study, designed the analyses, and obtained the funding to perform the experiments. M.S. and J.C. supervised sample collection and obtained patient information. A.I.-G., L.C.-P., and A.D.-M. collected the blood samples. R.P., L.M.R., and P.M. performed the CTC enrichment experiments. N.E.G. performed whole-genome amplification and obtained DNA from CTC pools and healthy PBMCs. J.M.A. performed the analyses. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank the Supercomputation Center of Galicia (CESGA; https://www.cesga.es) for providing all the computational resources needed for this study. This work was supported by the Spanish Ministry of Science and Innovation - MICINN (PID2019-106247GB-I00 awarded to DP) and by an AXA Research Fund postdoctoral grant (awarded to J.M.A.). DP receives further support from the Galician government (ED431C 2022/26). JMA is currently supported by an AECC-Investigator grant (INVES20007FERN). R.P. is currently supported by an AECC-Investigator grant (INVES234992PIÑE). L.M.R. is supported by a contract “Miguel Servet” from ISCIII (CP20/00119).

ORCID

0000-0002-9599-5463 (Joao M Alves)

0000-0001-8822-2991 (Astrid I. Díez-Martín)

0000-0001-9479-139X (Roberto Piñeiro)

0000-0002-7281-3137 (Nuria Estévez-Gómez)

0000-0002-7456-7531 (Laura Muinelo-Romay)

0000-0002-9994-4831 (Joaquin Cubiella)

0000-0003-1407-3406 (David Posada)

References

Ahmed, Zafarali, and Simon Gravel. 2018. “Intratumor Heterogeneity and Circulating Tumor Cell Clusters.” *Molecular Biology and Evolution* 35 (9): 2135–44.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Alves, Joao M., Nuria Estévez-Gómez, Monica Valecha, Sonia Prado-López, Laura Tomás, Pilar Alvariño, Roberto Piñeiro, et al. 2022. “Comparative Analysis of Capture Methods for Genomic Profiling of Circulating Tumor Cells in Colorectal Cancer.” *Genomics* 114 (6): 110500.

Benjamin, David, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. 2019. “Calling Somatic SNVs and Indels with Mutect2.” bioRxiv. https://doi.org/10.1101/861054.

Bertier, Gabrielle, Martin Hétu, and Yann Joly. 2016. “Unsolved Challenges of Clinical Whole-Exome Sequencing: A Systematic Literature Review of End-Users’ Views.” *BMC Medical Genomics* 9 (1): 52.

Budczies, Jan, Frederick Klauschen, Bruno V. Sinn, Balázs Győrffy, Wolfgang D. Schmitt, Silvia Darb-Esfahani, and Carsten Denkert. 2012. “Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization.” *PloS One* 7 (12): e51862.

Calabuig-Fariñas, Silvia, Eloísa Jantus-Lewintre, Alejandro Herreros-Pomares, and Carlos Camps. 2016. “Circulating Tumor Cells versus Circulating Tumor DNA in Lung Cancer-Which One Will Win?” *Translational Lung Cancer Research* 5 (5): 466–82.

Castro-Giner, Francesc, and Nicola Aceto. 2020. “Tracking Cancer Progression: From Circulating Tumor Cells to Metastasis.” *Genome Medicine* 12 (1): 31.

Connal, Siobhan, James M. Cameron, Alexandra Sala, Paul M. Brennan, David S. Palmer, Joshua D. Palmer, Haley Perlow, and Matthew J. Baker. 2023. “Liquid Biopsies: The Future of Cancer Early Detection.” *Journal of Translational Medicine* 21 (1): 118.

Degasperi, Andrea, Tauanne Dias Amarante, Jan Czarnecki, Scott Shooter, Xueqing Zou, Dominik Glodzik, Sandro Morganella, et al. 2020. “A Practical Framework and Online Tool for Mutational Signature Analyses Show Inter-Tissue Variation and Driver Dependencies.” *Nature Cancer* 1 (2): 249–63.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. 2015. “Sequenza: Allele-Specific Copy Number and Mutation Profiles from Tumor Sequencing Data.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 26 (1): 64–70.

Fernandez-Mateos, Javier, George D. Cresswell, Nicholas Trahearn, Katharine Webb, Chirine Sakr, Andrea Lampis, Christine Stuttle, et al. 2024. “Tumor Evolution Metrics Predict Recurrence beyond 10 Years in Locally Advanced Prostate Cancer.” *Nature Cancer* 5 (9): 1334–51.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P. Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. 2013. “IntOGen-Mutations Identifies Cancer Drivers across Tumor Types.” *Nature Methods* 10 (11): 1081–82.

Heide, Timon, Jacob Househam, George D. Cresswell, Inmaculada Spiteri, Claire Lynn, Maximilian Mossner, Chris Kimberley, et al. 2022. “The Co-Evolution of the Genome and Epigenome in Colorectal Cancer.” *Nature* 611 (7937): 733–43.

Hosea, Rendy, Sharon Hillary, Sumera Naqvi, Shourong Wu, and Vivi Kasim. 2024. “The Two Sides of Chromosomal Instability: Drivers and Brakes in Cancer.” *Signal Transduction and Targeted Therapy* 9 (1): 75.

Jia, Peng, Xiaofei Yang, Li Guo, Bowen Liu, Jiadong Lin, Hao Liang, Jianyong Sun, Chengsheng Zhang, and Kai Ye. 2020. “MSIsensor-pro: Fast, Accurate, and Matched-Normal-Sample-Free Detection of Microsatellite Instability.” *Genomics, Proteomics & Bioinformatics* 18 (1): 65–71.

Jodal, Henriette C., Lise M. Helsingen, Joseph C. Anderson, Lyubov Lytvyn, Per Olav Vandvik, and Louise Emilsson. 2019. “Colorectal Cancer Screening with Faecal Testing, Sigmoidoscopy or Colonoscopy: A Systematic Review and Network Meta-Analysis.” *BMJ Open* 9 (10): e032773.

Joung, Je-Gun, Bo Young Oh, Hye Kyung Hong, Hisham Al-Khalidi, Faisal Al-Alem, Hae-Ock Lee, Joon Seol Bae, et al. 2017. “Tumor Heterogeneity Predicts Metastatic Potential in Colorectal Cancer.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 23 (23):

7209–16.
Kassambara, Alboukadel, Marcin Kosinski, and Przemyslaw Biecek. 2016. “Survminer: Drawing Survival Curves Using ‘ggplot2.’” *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.survminer.
Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.
Lin, Danfeng, Lesang Shen, Meng Luo, Kun Zhang, Jinfan Li, Qi Yang, Fangfang Zhu, et al. 2021. “Circulating Tumor Cells: Biology and Clinical Significance.” *Signal Transduction and Targeted Therapy* 6 (1): 404.
Makrooni, Mohammad A., Brian O’Sullivan, and Cathal Seoighe. 2022. “Bias and Inconsistency in the Estimation of Tumour Mutation Burden.” *BMC Cancer* 22 (1): 840.
Maly, Vilem, Ondrej Maly, Katarina Kolostova, and Vladimir Bobek. 2019. “Circulating Tumor Cells in Diagnosis and Treatment of Lung Cancer.” *In Vivo* 33 (4): 1027–37.
McDonald, Kerry-Ann, Tsutomu Kawaguchi, Qianya Qi, Xuan Peng, Mariko Asaoka, Jessica Young, Mateusz Opyrchal, et al. 2019. “Tumor Heterogeneity Correlates with Less Immune Response and Worse Survival in Breast Cancer Patients.” *Annals of Surgical Oncology* 26 (7): 2191–99.
McGranahan, Nicholas, and Charles Swanton. 2017. “Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.” *Cell* 168 (4): 613–28.
Miller, M. Craig, Gerald V. Doyle, and Leon W. M. M. Terstappen. 2010. “Significance of Circulating Tumor Cells Detected by the CellSearch System in Patients with Metastatic Breast Colorectal and Prostate Cancer.” *Journal of Oncology* 2010:617421.
Mishima, Yuji, Bruno Paiva, Jiantao Shi, Jihye Park, Salomon Manier, Satoshi Takagi, Mira Massoud, et al. 2017. “The Mutational Landscape of Circulating Tumor Cells in Multiple Myeloma.” *Cell Reports* 19 (1): 218–24.
Morgan, Eileen, Melina Arnold, A. Gini, V. Lorenzoni, C. J. Cabasag, Mathieu Laversanne, Jerome Vignat, Jacques Ferlay, Neil Murphy, and Freddie Bray. 2023. “Global Burden of Colorectal Cancer in 2020 and 2040: Incidence and Mortality Estimates from GLOBOCAN.” *Gut* 72 (2): 338–44.
Mroz, Edmund A., and James W. Rocco. 2013. “MATH, a Novel Measure of Intratumor Genetic Heterogeneity, Is High in Poor-Outcome Classes of Head and Neck Squamous Cell Carcinoma.” *Oral Oncology*. https://doi.org/10.1016/j.oraloncology.2012.09.007.
Oh, Hyung-Hoon, and Young-Eun Joo. 2020. “Novel Biomarkers for the Diagnosis and Prognosis of Colorectal Cancer.” *Intestinal Research* 18 (2): 168–83.
Paoletti, Costanza, Andi K. Cani, Jose M. Larios, Daniel H. Hovelson, Kimberly Aung, Elizabeth P. Darga, Emily M. Cannell, et al. 2018. “Comprehensive Mutation and Copy Number Profiling in Archived Circulating Breast Cancer Tumor Cells Documents Heterogeneous Resistance Mechanisms.” *Cancer Research* 78 (4): 1110–22.
Schumacher, Sarah, Christoph Bartenhagen, Martin Hoffmann, Daniel Will, Johannes C. Fischer, Stephan E. Baldus, Christian Vay, et al. 2017. “Disseminated Tumour Cells with Highly Aberrant Genomes Are Linked to Poor Prognosis in Operable Oesophageal Adenocarcinoma.” *British Journal of Cancer* 117 (5): 725–33.
Shin, Alice E., Filippo G. Giancotti, and Anil K. Rustgi. 2023. “Metastatic Colorectal Cancer: Mechanisms and Emerging Therapeutics.” *Trends in Pharmacological Sciences* 44 (4): 222–36.
Therneau, Terry M. 2001. “Survival: Survival Analysis.” *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.survival.
Toss, Angela, Zhaomei Mu, Sandra Fernandez, and Massimo Cristofanilli. 2014. “CTC Enumeration and Characterization: Moving toward Personalized Medicine.” *Annals of Translational Medicine* 2 (11): 108.
Tsai, Wen-Sy, Jinn-Shiun Chen, Hung-Jen Shao, Jen-Chia Wu, Jr-Ming Lai, Si-Hong Lu, Tsung-Fu Hung, et al. 2016. “Circulating Tumor Cell Count Correlates with Colorectal Neoplasm Progression and Is a Prognostic Marker for Distant Metastasis in Non-Metastatic Patients.” *Scientific Reports* 6 (1): 1–8.
Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 43:11.10.1–11.10.33.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38 (16): e164.

Zauber, Ann G., Sidney J. Winawer, Michael J. O’Brien, Iris Lansdorp-Vogelaar, Marjolein van Ballegooijen, Benjamin F. Hankey, Weiji Shi, et al. 2012. “Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths.” *The New England Journal of Medicine* 366 (8): 687–96.

Zhu, Zhongyi, Si Qiu, Kang Shao, and Yong Hou. 2018. “Progress and Challenges of Sequencing and Analyzing Circulating Tumor Cells.” *Cell Biology and Toxicology* 34 (5): 405–15.

Zygulska, Aneta L., and Piotr Pierzchalski. 2022. “Novel Diagnostic Biomarkers in Colorectal Cancer.” *International Journal of Molecular Sciences* 23 (2). https://doi.org/10.3390/ijms23020852.