UNIVERSITÀ
DI TORINO

Department of Oncology

PhD thesis
Complex Systems for Quantitative Biomedicine

XXXVII cycle

Mutational signatures of colorectal cancers according to distinct computational workflows

Supervisors:
Professor Enzo Medico
Professor Alberto Bardelli
Professor Federica Di Nicolantonio

Candidate:
Paolo Battuello

UNIVERSITÀ
DI TORINO

Department of Oncology

PhD thesis
Complex Systems for Quantitative Biomedicine

XXXVII cycle

Mutational signatures of colorectal cancers according to distinct computational workflows

Supervisors:
Professor Enzo Medico
Professor Alberto Bardelli
Professor Federica Di Nicolantonio

Candidate:
Paolo Battuello

Table of contents:

- Abstract 5
- Background and study design 6
- Results 17
- Discussion and Conclusions 37
- Materials and Methods 43
- Tables 47
- References 53
- Collaborations and side projects 56
- Acknowledgments 58

Abstract

Mutational signatures can be defined as unique patterns of mutations that occur in the genome of an organism. They are the consequence of multiple mutational processes of both endogenous and exogenous origin, such as intrinsic slight infidelity of the DNA replication machinery, defects in the DNA repair machinery or exposure to physical or chemical agents such as UV-light and mutagenic drugs. In the past few years, tumor mutational signatures have become increasingly important in cancer research; however, the absence of standardized methods to perform this analysis limits reproducibility and robustness of the results. Our work aimed to dissect the influence of each variable of the mutational signature computational workflow on the overall result.

We used colorectal cancer (CRC) as a model and monitored the contribution of a defined subset of mutational signatures in a preclinical dataset of 230 CRC cell lines and in a clinical dataset of 152 CRC patients. Results were validated in three independent preclinical and clinical datasets in which different mutational signatures could be monitored: 483 endometrial cancer patients stratified for mismatch repair proficiency, 35 lung cancer patients stratified by smoking status, and 12 patient derived organoids annotated for colibactin exposure.

By evaluating different bioinformatic tools, reference datasets, and input data sizes including whole genome sequencing, whole exome sequencing and a pan-cancer gene panel, we demonstrated significant variability in the results. We report that the use of distinct algorithms and references led to statistically different results, highlighting how arbitrary choices may induce variability in the mutational signature contributions. Furthermore, we found a differential contribution of mutational signatures between exonic, intronic and and intergenic regions. As the applicability of the mutational signature depends on and is limited by the number of mutations present in the samples, we investigated the number of somatic variants required for a reliable mutational signature assignment.

As a conclusion of the project, we developed CoMSCER (Comparative Mutational Signature analysis on Coding and Extragenic Regions), a bioinformatics tool capable of

assessing the impact of multiple parameters on the robustness of the results, allowing researchers to easily perform comparative mutational signature analysis by coupling the results from multiple tools and public reference datasets and to assess mutational signature contributions in coding and non-coding genomic regions. In conclusion, our study provides a comparative framework to elucidate the impact of different computational workflows on mutational signatures.

Background and study design

Genetic instability fuels tumor initiation and progression, and mutations represent the primary source of genetic variation. There is increasing evidence suggesting that a variety of factors can damage DNA and induce specific patterns of mutations in the genome, also known as mutational signatures [1]. The concept of mutational signature was first defined in 2012 with the genomic analysis of a cohort of 21 breast cancer patients [2]. In this study, conducted by Serena Nik-Zainal and colleagues, the authors quantified the contribution of the six classes of base substitutions (C>A, C>G, C>T, T>A, T>C and T>G) in the sequence context in which the mutation occurred, therefore including also the bases immediately 5' and 3' to each mutated nucleotide. This resulted in the classification of mutations into one of 96 possible trinucleotides, based on six classes of base substitution and 16 possible sequence contexts for each mutated base. This assignment results in the *'96 base substitution matrix'*, a matrix that is informative of the frequency of each type of nucleotide variant, and that represents the starting point for many downstream analyses.

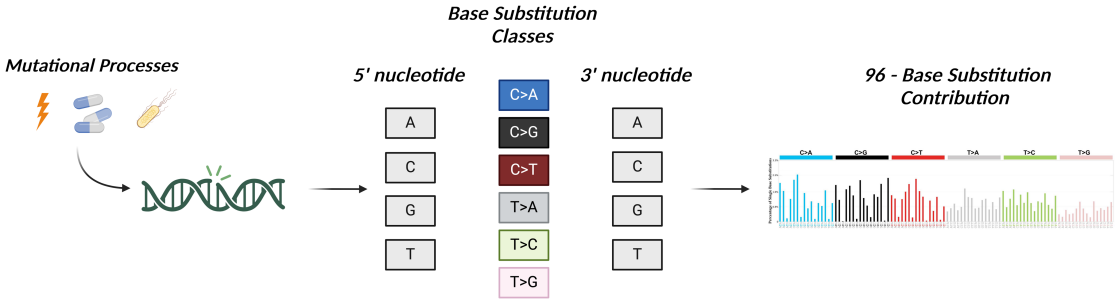


Figure 1: Explicative figure of the classification of single nucleotide variants (SNVs). SNVs generated by a multitude of mutational processes are classified in one of six base substitution classes; 5' and 3' nucleotides are taken into consideration and each variant is classified in one of the 96 base substitution contexts. The last step represents the quantification of the relative contribution of each one of the 96 base substitutions.

In 2013, Alexandrov and colleagues, exploiting a large collection of 507 whole-exome (WES) and 6,535 whole-genome (WGS) sequencing data from 7042 primary cancers of 30 different histology, reported the first set of 21 single-base substitution mutational

signatures [3]. Since 2015, the number of characterized signatures has continued to grow, reaching 86 single base substitution (SBS) signatures in the latest version (v3.4) published on COSMIC (Catalog Of Somatic Mutations In Cancer) in October 2023. In addition to SBS mutational signatures, doublet base substitutions (DBS), small insertions and deletions (ID), copy number (CN) and structural variations (SV) signature have been characterised and released on COSMIC [1, 4-7]. In addition, transcriptional strand bias has been also considered, highlighting mutational processes exhibiting asymmetric number of mutations due to either one of the DNA strands being preferentially repaired or damaged [8]. For many of them the aetiology of the mutational process generating the mutational signature has been also released. The current set of mutational signatures present in COSMIC, is based on an analysis of 10,952 exomes and 1,048 whole-genomes across 40 distinct types of human cancer [7].

Today, mutational signature analysis has become a standard genetic analysis that can inform researchers of a variety of mutational processes that have been active in a cancer cell [9]. For instance, age-related signatures can describe the endogenous mutational process initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine which generates G>T mismatches that when not properly detected and corrected result in the fixation of C>T substitutions. The most known age-related signatures are SBS1 and SBS5, which are defined as clock-like in that the number of mutations correlates with the age of the individual [10-13]. Another example of mutagenic process of endogenous origin is represented by *POLE* related signatures: mutations in the exonuclease domains of the polymerase epsilon gene result in a compromised functionality of the protein that leads to an hypermutated phenotype [14]. This abnormal increase of mutations involves mainly in C>A and C>T substitutions described respectively by SBS10a and SBS10b.

In addition to cell endogenous mutational processes, mutational signatures can also be informative for monitoring exposure to physical and chemical agents. Examples include UV-light related signatures, which are particularly enriched in melanoma and are described by several SBS signatures (SBS7a, SBS7b, SBS7c and SBS7d) [15, 16], DBS1 signature, which exhibits predominantly CC>TT mutations and ID13, which is predominantly characterised by single base T deletions at TT dinucleotides [5]; or

colibactin exposure, described by SBS88 and ID18, which are mainly found in colorectal cancers exposed to *E.coli* bacteria carrying the *pks* pathogenicity island and therefore producing the genotoxic compound colibactin [17-19].

Some mutational processes are known to be involved in a variety of cancer associated features with profound clinical implications and for some of these, mutational signatures may provide a well-suited tool for detection and monitoring [20].

Mismatch repair-deficiency (MMR-d) exemplifies a biological phenomenon with strong clinical implications. The DNA mismatch repair system (MMR) is a system for recognition and repair of various DNA related errors (insertion, deletion, mis-incorporation of bases) that can occur during DNA replication and recombination [21]. MMR is a highly conserved machinery, from prokaryotes to eukaryotes, accounting from proteins able to recognize and repair mismatches. Indeed, studies of the MMR system in prokaryotes organisms such as *E. coli* were pivotal in the characterization of the key protein players. The whole process starts with the recognition of the mismatch: this step is mediated in *E. coli* by the MutS homodimer, which then recruits the MutL homodimer. The formation of the ternary ATP-dependent complex activates the endonuclease activity of the MutH complex, which, together with UvrD helicase and one of several exonucleases (Exo), generates a gap that extends beyond the mismatch. The gap is then filled by DNA polymerase III (pol III) and the remaining nick is sealed by a DNA ligase. In eukaryotes, the process is more complex, and of the five MutS homologues (MSH) identified in human cells, MSH2, MSH3 and MSH6 participate in MMR as heterodimers [22]. With regard to the MutL complex, human cells express four MutL homologues: MLH1, MLH3, PMS1 and PMS2, which act as three distinct heterodimers, of which the MLH1-PMS2 complex plays the most important role in MMR.

Mutations in the MMR genes *MLH1*, *PMS2*, *MSH2* and *MSH6* and promoter hypermethylation of *MLH1* represent the principal causes of microsatellite instability (MSI); this usually results in an elevated mutation rate in microsatellite regions and an hypermutated phenotype [23]. These features often are associated with great sensitivity to immune checkpoint inhibitors (ICB) linked to the potential continuous generation of neoantigens able to boost immune recognition and activation [24]. Consequently, there has been strong interest in assays capable of assessing MSI status: current clinical assays

for detecting these tumours range immunohistochemical staining for concomitant loss of MMR protein pairs to PCR-based assays to determining MSI (e.g. detection of mono-di- nucleotide repeats), and algorithms designed for NGS data, such as mSINGS, MSIsensor, and MSIseq [25-27]. Each of these assays has certain limitations: for example, NGS-based assays can be affected by technical biases, such as varying sequencing coverage of depth, or biological biases, such as differences in tumor content and heterogeneity, which may not always accurately depict the MSI phenotype. Recently, mutational signatures involved in MMR defect have been exploited as MMR classifier: 'MMRDetect' is an example of tool which exploit substitution and indel mutational signatures associated with defects in MMR genes to stratify MMR proficient to MMR deficient tumors [28].

Homologous recombination deficiency (HRD) is another well-suited example of the close link between mutational signatures and clinical application. Nearly twenty years ago, the synthetic lethality between poly (ADP-ribose) polymerase (PARP) and HRD was characterised, paving the way for PARP inhibitors as a strategy for targeting BRCA1-deficient or BRCA2-deficient tumours [29]. Consequently, a plethora of assays for the detection of the HRD phenotype was developed; in addition, six mutational signatures have been associated with HRD including SBS, ID and SV signatures. The first example of bioinformatic algorithm able to exploit those signatures is 'HRDetect', a tool which showed a considerable sensitivity of 98,7% in the prediction of BRCA1 or BRCA2 deficiency in ovarian and breast cancer [30]. The tool was also applied in triple-negative breast cancer (TNBC) in a real-world clinical setting of 254 cases: patients with a high HRDetect score who received adjuvant chemotherapy experienced better invasive disease-free survival and distant relapse-free intervals compared to those with a low HRDetect score [31]. Notably, this improvement was independent of whether the underlying cause of homologous recombination deficiency (HRD), of genetic or epigenetic origin, was identified. In fact, HRD drivers were identified in only about two-thirds of the high HRDetect group, linked to germline or somatic *BRCA1* or *BRCA2* mutations, *BRCA1* or *RAD51C* promoter hypermethylation, or biallelic *PALB2* loss. This indicates that the mutational signature approach could predict outcomes in approximately 30% of tumors with HRD signatures that would be missed by conventional targeted sequencing. Indeed, this represents the first example of a

successful use of mutational signatures in a clinical scenario and highlights their potential in the clinical setting.

Moreover, several SBS signatures and DBS signatures, have been associated to exposure of clinically approved drugs; examples are SBS11 for alkylating agents and SBS31, SBS35, or DBS5 for platinum-based chemotherapy [7]. These signatures can be a surrogate for therapeutic efficacy, as described in the ARETHUSA clinical trial (NCT03519412): in this study, patients with stage IV colorectal cancer underwent a priming phase with temozolomide (TMZ) treatment until disease progression, then a molecular assessment was performed to evaluate the increase in tumor mutational burden (TMB), which, if greater than 20 mut/Mb, made the patient eligible for the immunotherapy phase in which pembrolizumab was administered. Translational findings were described in the *Crisafulli et al.*, where the authors performed mutational signature analysis of tissue biopsies obtained after the priming phase [32]. The results describe a positive correlation between the TMZ exposure (number of cycles) and the relative abundance of SBS signature 11, providing a further example of how mutational signatures can be used to monitor treatment efficacy.

All of these scenarios are evidence of how mutational signatures are entering the clinical setting [20]; the patient stratification and treatment monitoring described above are only a few examples of the applications being explored. Despite its potential, the clinical applications of mutational signatures remain largely unexplored. Its future implementation will depend on further studies proving its effectiveness in enhancing prognosis and/or patient stratification. The common denominator of all these examples is the quantification of the abundance of a single signature or a set of signatures; in these studies, the authors used already characterised and publicly available signatures to perform a 'fitting' analysis, which aims to quantify the presence of previously characterized mutational signatures in a given cancer sample. Currently, from a technical standpoint, there is no gold standard for mutational signature fitting analysis, therefore potentially causing robustness and reproducibility issues. This phenomenon is amplified by the increasing number of publications over the last decade that refer to or include the analysis of mutational signatures. In addition, most articles on signature analysis include 'custom' workflows for mutation calling and signature analysis,

exacerbating such issues. At present, a reference to exploit as a gold standard or a proper benchmark study capable of evaluating different bioinformatics workflows is still missing in the literature.

With this in mind, we designed a study that could address these concerns: we hypothesised that many computational variables could affect the analysis, leading to differences that may reflect possible biological differences or may be purely technical. Accordingly, we decided to identify the main variables of a hypothetical computational workflow and to assess the impact of each single variable on the overall results of the analyses.

Since their initial discovery, more than thirty different bioinformatic tools have been developed to extract de novo mutational signatures or to perform fitting analysis to estimate the prevalence of already characterized signatures in individual samples [33]. The choice of the algorithm to perform mutational signature analysis is obviously a key point of the analysis and, to our knowledge, a proper benchmark study to compare the performance of several available tools is still missing. Secondly, as described above, the number of characterised signatures and online reference databases is constantly increasing. Indeed, the choice of the reference dataset represents another critical step of the workflow. As already highlighted in the study from *Maura* and colleagues the choice of mutational signatures to include in the analysis represents a balance between the possibility of analysing a small number of signatures, leading to potential underestimation of active mutational processes, and the inclusion of a large number of signatures, leading to signal dilution and overfitting [34]. Furthermore, most of the reported signatures and the available tools for mutational signature profiling were characterised and designed to work with WES or WGS data. However, the extent to which next-generation sequencing (NGS) data from targeted gene panels can be used to reliably identify mutational signatures is largely unknown. This becomes particularly relevant especially in the clinical setting, where small NGS panels can be used for clinical diagnosis or predictive purposes. Related to this point, one of the main technical limitations of performing mutational signature analysis is the number of SNVs present in the tumour sample: the fitting procedure is indeed compromised when trying to fit a

low number of mutations to a given reference dataset, a condition that could occur especially when considering small NGS targeted panel.

Having defined several aspects to be addressed, we decided to use colorectal cancer (CRC) as a representative model for conducting our study. CRC represents the third most common malignancy worldwide and has a very well characterised molecular profile [35]. From a genetic standpoint, CRCs can be characterised in into the three subtypes: microsatellite stable (MSS) tumors, characterized by chromosomal instability and usually associated with a proficient mismatch repair machinery (MMRp); microsatellite instable (MSI) tumors, representing a minor fraction of CRCs, generally exhibiting mismatch repair deficiency (MMRd) that leads to a hypermutated phenotype, and *POLE*-mutated CRCs, a small fraction of MSS-MMRp samples (1-2%) harbouring mutations in the exonuclease domain of the DNA polymerase epsilon (POLE), resulting in a hypermutated phenotype [14, 36]. These molecular subtypes are associated with defined clinical features, such as anatomic site, treatment response and prognosis [36]. Notably, hypermutant CRCs, such as MSI and POLE subtypes are characterised by a good responsiveness to immune checkpoint blockade [37-39].

Importantly, these CRC subtypes present distinct genetic profile, with features that can be tracked and quantified by subtype-specific signatures. In details: MSI-MMRd CRCs are enriched in mutational signatures associated with MMRd (COSMIC v3.4: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, SBS44), while MSS-MMRp presenting *POLE* mutated CRCs are enriched in mutational signatures associated with polymerase epsilon exonuclease domain mutations (COSMIC v3.4: SBS10a, SBS10b) [40]. In light of these consideration, we believe that CRC could serve as an appropriate model for investigating the potential of mutational signature-based genetic stratification.

Accordingly, we used as main dataset, an in-house available collection of 230 CRC cell lines encompassing all the main CRC subtypes such as MSS-MMRp (145/230, 63%), MSI-MMRd (78/230, 34%), and POLE-mutated samples (7/230, 3%) (Table 1) [41-43].

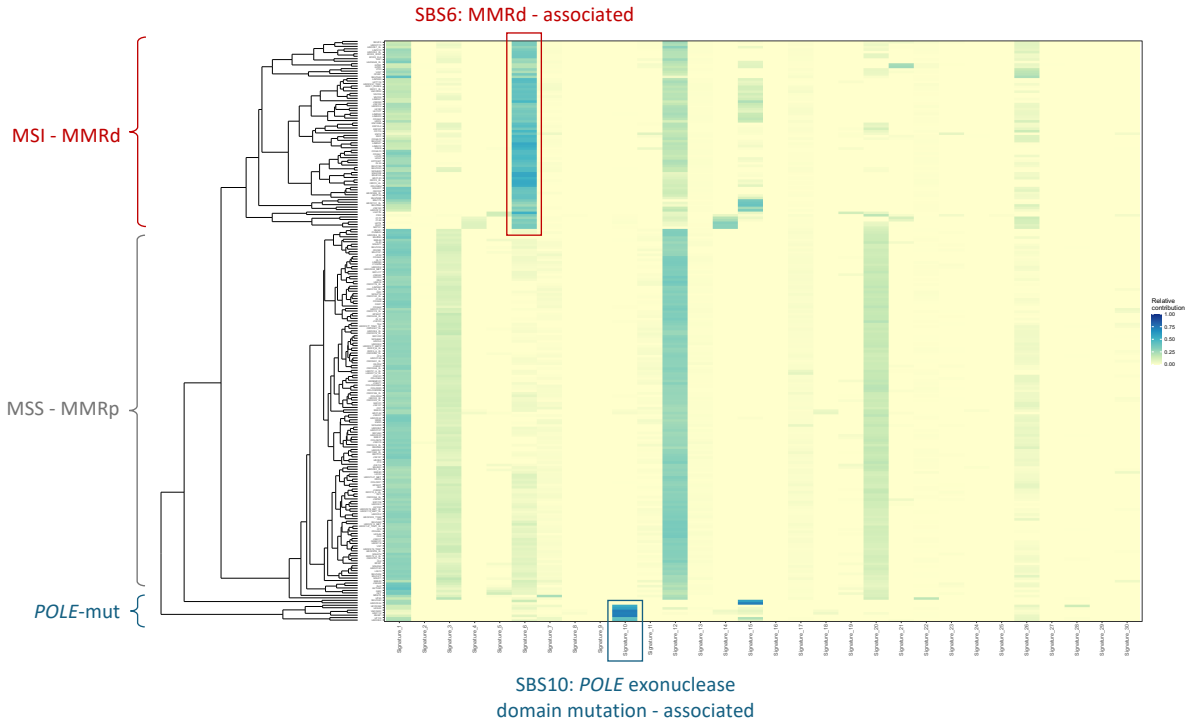


Figure 2: Representative heatmap of the mutational signature enrichment in the CRC preclinical dataset using COSMIC v2 as reference dataset. SBS6 enrichment is indicative of the MSI-MMRp cell lines, SBS10 enrichment is indicative of the *POLE* mutated cell lines.

We further reasoned that our results should be confirmed in a second dataset. Therefore, we downloaded an external clinical dataset from The Cancer Genome Atlas (TCGA) including 152 CRC patients, specifically: 59 MSS-MMRp (59/152, 39%), 88 MSI-MMRp (88/152, 58%), 5 *POLE*-mut (5/152, 3%) CRCs (Table 2) [44].

Moreover, we reasoned that defining a hypothetical standard bioinformatic pipeline would be valuable to dissect the main variables of the analysis. We identified three key parameters: the NGS workflow, spanning from WGS to WES or a smaller NGS gene targeted panel; the bioinformatics tool, and the reference catalogue of mutational signatures for the signature fitting analysis.

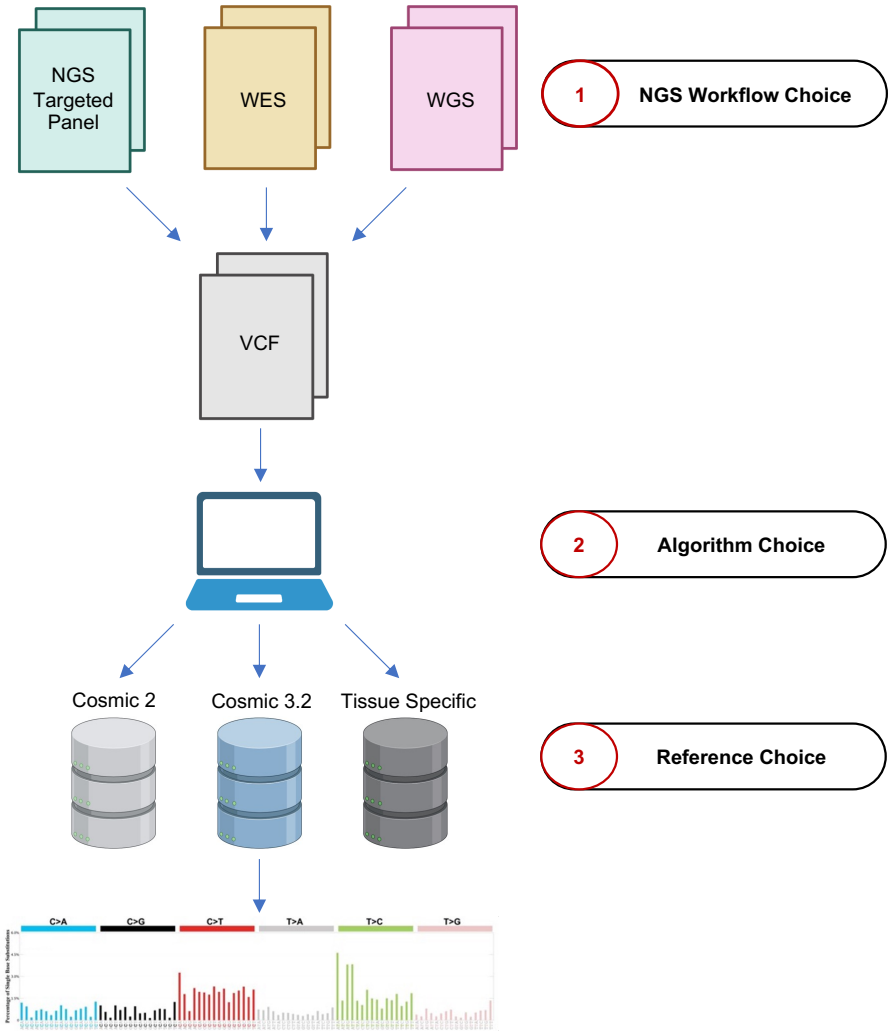


Figure 3: Representative figure of the identification of the three main variables of a hypothetical mutational signature analysis computational workflow: the NGS workflow choice, evaluating NGS targeted panel, WES and WGS data; the algorithm choice, evaluating five different bioinformatics tools selected from the literature; the reference choice, evaluating three different sized reference SBS signature datasets. NGS: next generation sequencing; WES: whole exome sequencing; WGS: whole genome sequencing; VCF: Variant Call Format.

The first parameter concerns the genomic size covered by the sequencing data: we exploited the availability of multiple sequencing data (WES and WGS) for many of the cell lines present in the preclinical dataset. Starting from the WES we also reduced in-silico the covered regions to obtain the region targeted by the TruSight Oncology 500 targeted gene panel (TSO-500 from Illumina©) as a representative example of NGS targeted gene panel data. This allowed us to compare the results of the mutational

signature analysis for each CRC sample using three different NGS workflows. The second parameter is the algorithm: in order to properly chose a subset of bioinformatics tools to evaluate, we decided to perform a literature systematic review from the publicly available repository PubMed Central (details reported in the methods section) with the aim of identifying the most used bioinformatics tool for performing mutational signature fitting. This analysis led us to identify the five most commonly used tools: MutationalPatterns [45], deconstructSigs [46], signature.tools.lib [47], SigProfilerAssignment [48] and SignatureAnalyzer [1]. As a final step in workflow, we evaluated how the use of different reference datasets of varying size and signature would affect the fitting analysis. We used three different datasets as reference: COSMIC v2, as a representative small reference dataset containing 30 SBS signatures; COSMIC3.2, as a large dataset containing 78 SBS signatures; and a CRC-specific dataset containing a subset of 28 SBS signatures expected to be found in CRCs [1].

We then established two different readouts to properly assess the performance of the fitting analysis. The first readout was defined as the '*biological readout*', and we exploited the genetic stratification of CRC to evaluate it. Specifically, we measured how mutational signatures could correctly stratify CRC MSS-MMRp, MSI-MMRd and POLE-mutated samples. We performed this analysis by quantifying the total contribution of MMRd and POLE mutation-associated signatures.

The second readout was defined as the '*mathematical readout*', assessing how fitted mutational signatures recapitulate the mutational landscape of individual samples. This was measured by calculating the cosine similarity between the mutational profile of each sample and the profile reconstructed using the fitted mutational signatures. We considered 0.9 as a threshold for reliability as reported in the literature [49].

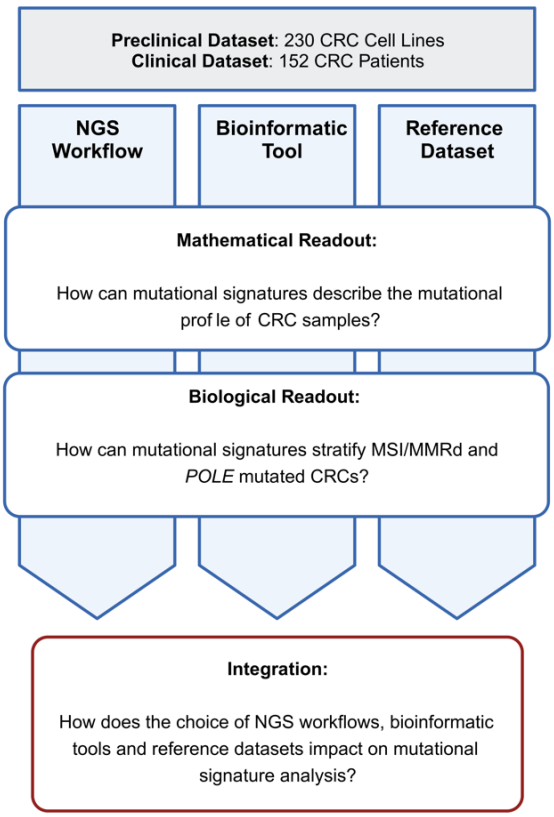


Figure 4: Schematic representation of the two readouts used for mutational signature fitting evaluation. Starting from the clinical and preclinical datasets, each variable of the analysis was evaluated with both the mathematical and the biological readout. NGS: next generation sequencing; MSI: microsatellite instability; MMRd: mismatch repair deficiency.

Finally, we thought that our evaluation workflow could be of interest to the scientific community for assessing the robustness of the analysis and the performance of signature fitting. Indeed, to make our workflow accessible, we developed CoMSCER (Comparative Mutational Signature analysis on Coding and Extragenic Regions), a bioinformatics tool capable of assessing the impact of multiple parameters on the robustness of the results and recapitulating the main analysis of the study.

Results

Evaluation of NGS workflows on mutational signature analysis

First, we investigated the impact of the NGS workflow, focusing on how different sequencing data, namely WGS, WES and the targeted pan-cancer panel TSO-500, might affect mutational signatures. As previously described, we evaluated both the mathematical and biological readout, preforming the analysis on the preclinical dataset, taking advantage of the multiple sequencing data that were available for each of the CRC cell lines. We selected the Illumina TruSight Oncology 500 (TSO-500) panel for two key reasons: its pan-cancer gene coverage, which is not specific to colon cancer and thus avoids overestimating the number of detected mutations, and its broad genomic scope, covering over 1.9 Mb making it suitable for comprehensive mutation profiling.

When considering the mathematical readout, cosine similarity reached the reliability threshold of 0.9 with all three NGS types of data, supporting the technical feasibility of the analysis spanning from WGS to gene-targeted panels. However, the three outcomes were significantly different when compared using the Wilcoxon rank test.

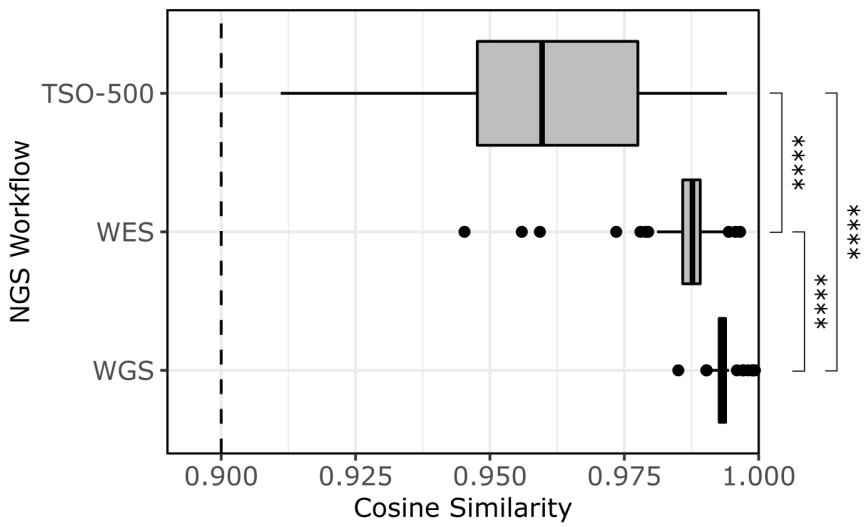


Figure 5: Distribution of cosine similarity values in the preclinical CRC dataset, using WGS, WES and the TSO-500 pan-cancer panel. Dashed line represents the cosine similarity threshold. NGS: next generation sequencing; TSO-500: TruSight Oncology 500; WES: whole exome sequencing; WGS whole genome

sequencing. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Concerning the biological readout, to properly measure how the different variables affected the stratification of CRC molecular subtypes, we defined ΔMMR as the difference between the median contribution of MMRd-associated signatures between MSS-MMRp and MSI-MMRd samples. Similarly, ΔPOLE was defined as the difference between the median contribution of POLE-associated signatures between POLE wild-type MSS-MMRp and POLE-mutated MSS-MMRp samples. The median ΔMMR was > 0 in all cases, indicating that the use of TSO-500, WES and WGS data types allows significant stratification of MSS-MMRp from MSI-MMRd CRCs. A comparable scenario was observed when considering sample stratification based on *POLE* mutational status, as determined by the median ΔPOLE parameter.

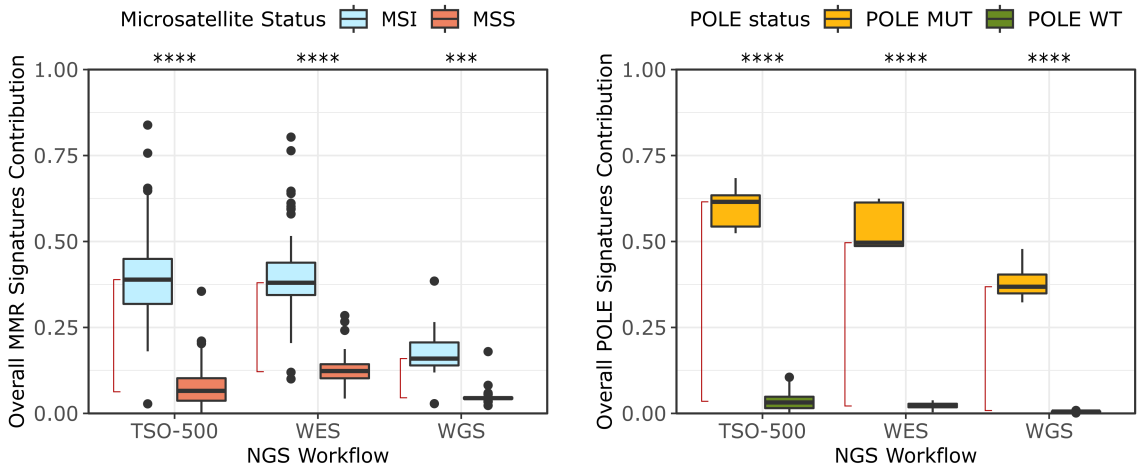


Figure 6: On the left side: boxplots representing the overall contribution of MMR associated mutational signatures in MSI-MMRd and MSS-MMRp of the CRC preclinical samples; red line represents ΔMMR. On the right side: boxplots representing the overall contribution of POLE mutation-associated signatures in POLE-mutated and POLE wild-type CRC preclinical samples; red line represents ΔPOLE. TSO-500: TruSight Oncology 500; WES: whole exome sequencing; WGS whole genome sequencing; MMR: mismatch repair. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Although the median ΔMMR was higher than 0 with all three sequencing workflows, the value was unexpectedly lower using WGS data (0.11 WGS < 0.26 WES < 0.32 TSO-500).

We hypothesized that this could be due to the dilution of the MMR signature signal with larger genomic sources such as WGS. To test this hypothesis, we investigated the contribution of different classes of mutational signatures in our subset of MSI-MMRd CRC cell lines. Specifically, we considered: MMRd related signatures, aspecific signatures often referred to as *'flat signatures'* (SBS: 3-5-8-40-89) defined, as previously proposed, as signatures in which the 96-mutational profile shows relatively even contribution of each trinucleotide context (<0.05%) [34], artefact-associated signatures and mutational signatures linked to unrelated biological processes.

This analysis showed an increased signal for *'flat signatures'* in WGS data, thus suggesting a possible explanation for the observed dilution of the MMR signatures. Additionally, to elucidate the possible source of the increase of *'flat signatures'* signal, we asked whether distinct genomic regions might contribute unequally to the mutational signature profile. To investigate this, we performed mutational signature analysis considering mutations derived from the exonic, intronic and extragenic regions extracted from WGS. The results showed the absence of *'flat signatures'* in the exonic regions, consistent with the findings from WES, while intronic and extragenic regions showed an increase in the contribution of *'flat signatures'* by 11,6% and 36%, respectively, supporting our hypothesis.

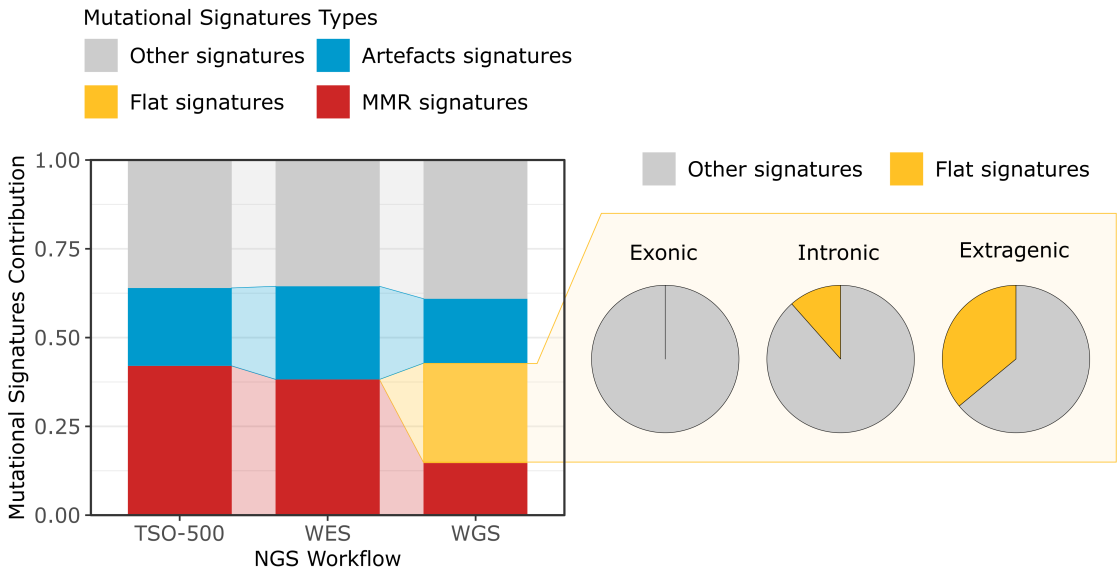


Figure 7: On the left, median contribution of artefact-driven, MMRd-associated and 'flat signatures' in the MSI-MMRd CRC cell lines. On the right, pie charts of the contribution of 'flat signatures' in exonic, intronic

and extragenic regions from WGS data. TSO-500: TruSight Oncology 500; WES: whole exome sequencing; WGS whole genome sequencing; MMR: mismatch repair.

To further support and extend these results, we evaluated the median ΔMMR in each specific genomic region. As highlighted in Figure 8, the median ΔMMR between MSI-MMRd and MSS-MMRp CRC cell lines of the extracted exonic regions aligns closely with that observed from WES data.

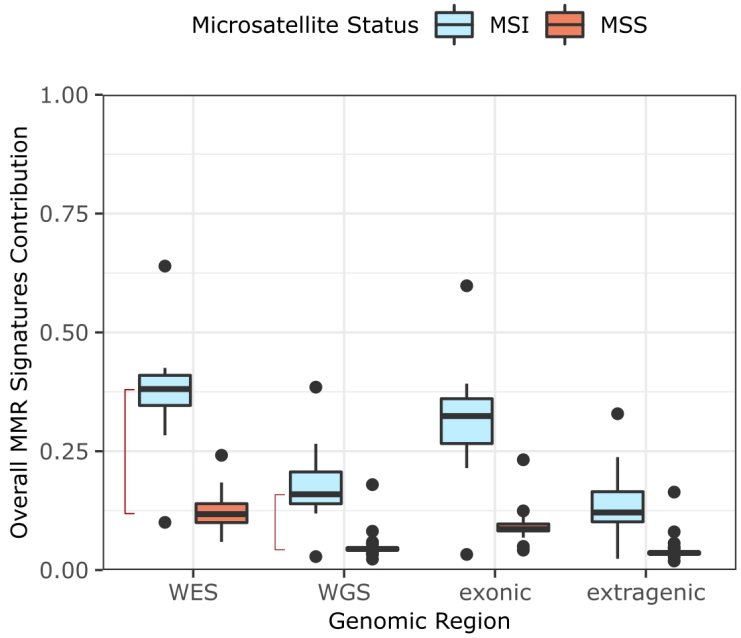


Figure 8: Comparison of MMR signatures contribution in WES, WGS, and exonic and extragenic regions extracted from WGS data. WES: whole exome sequencing; WGS whole genome sequencing; MMR: mismatch repair.

Overall, these results indicate that mutational signature analysis may be feasible not only using WES and WGS data but also large pan-cancer NGS panels such as the TSO-500. Notably, increasing the genomic size evaluated in the analysis was only partially helpful in improving signature accuracy.

Evaluation of the bioinformatics tool on mutational signature analysis

We conducted a literature systematic review on PubMed Central (PMC) using as search key "mutational signatures". From the initial 831 entries, 128 manuscripts were available for download. From this pool, we identified 70 papers that referenced algorithms for fitting mutational signatures that were available for download. From this list, we selected the top 5 most referenced tools: MutationalPatterns (MP) [45], deconstructSigs (DS) [46], signature-tools.lib (STL) [47], SigProfilerAssignment (SPA) [48] and SignatureAnalyzer (SA) [1].

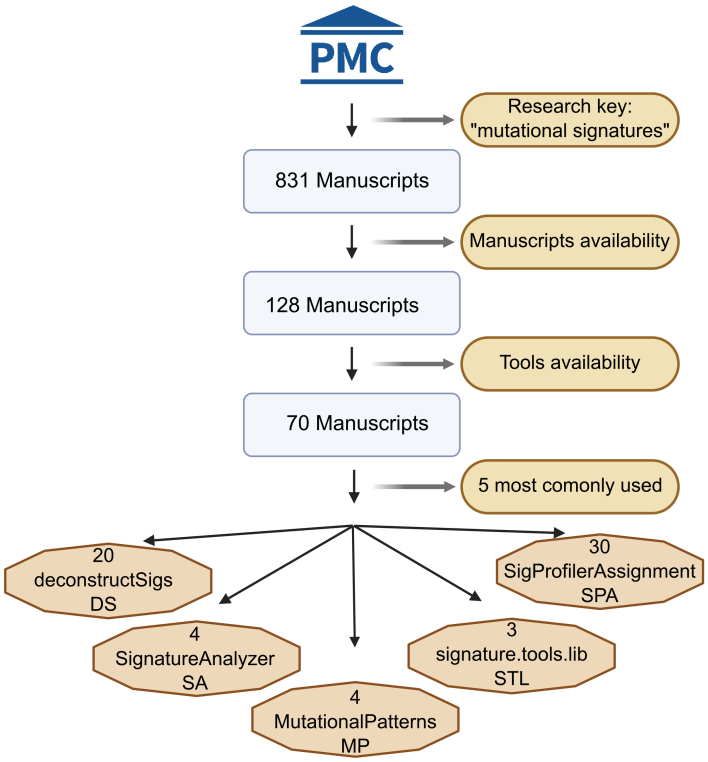
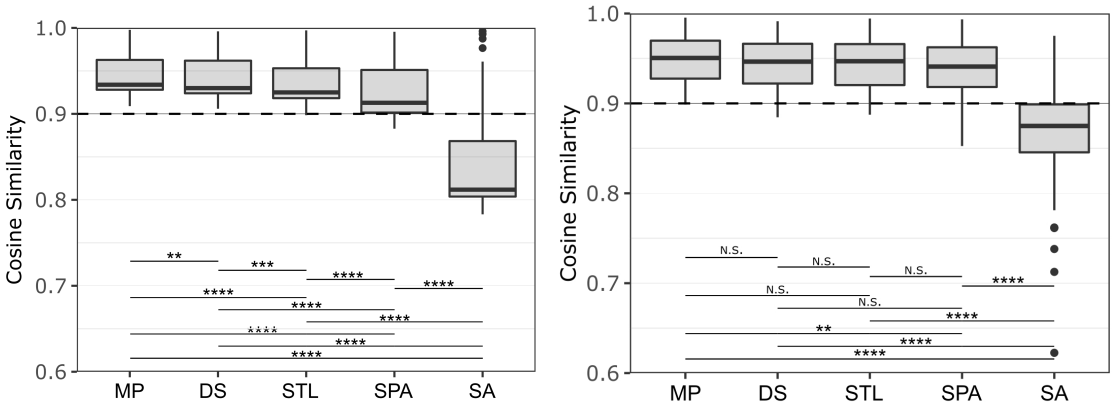


Figure 9: Graphical representation of the systematic review utilized for the identification of the 5 most used tools. PMC: PubMed Central.

We next performed mutational signature fitting using the five bioinformatic tools on the CRC preclinical and clinical datasets. In the preclinical dataset, 4 out of 5 tools achieved a median cosine similarity of 0.9, although the differences in cosine similarity distribution among the five software were statistically significant. We highlighted that, among the five tools evaluated, SPA and SA did not allow the assignment (cosine

similarity < 0.9) of more than 20% of the samples (1/230, 0.4% with STL; 48/230, 20.9% with SPA and 217/230, 94.3% with SA). Notably, only MP and DS allowed mutational signature fitting for all 230 samples. Results from the clinical dataset were comparable: 4 out of 5 software reached a median value of cosine similarity above the technical reliability threshold, with only limited samples not reaching the threshold. Similar to what we observed in the preclinical dataset, cosine similarity distributions were significantly different. Also in this case, multiple samples were not assigned by different tools: 1/152, 0.66% with MP, 10/152, 6.6 % with DS, 12/152, 7.9% with STL, 25/152, 16.4% with SPA 116/152, 76.3% with SA. Of note, the trend between the median value of cosine similarity among the five different algorithms was maintained across the preclinical and clinical datasets.



Tool	Assigned samples (CS >= 0.9)	NOT assigned samples (CS < 0.9)	Fractional abundance of NOT assigned samples
MP	230	0	0,0%
DS	230	0	0,0%
STL	229	1	0,4%
SPA	182	48	20,9%
SA	13	217	94,3%

Tool	Assigned samples (CS >= 0.9)	NOT assigned samples (CS < 0.9)	Fractional abundance of NOT assigned samples
MP	151	1	0,7%
DS	142	10	6,6%
STL	140	12	7,9%
SPA	127	25	16,4%
SA	36	116	76,3%

Figure 10: Distribution of cosine similarity values obtained with MutationalPatterns (MP), deconstructSigs (DS), signature-tools.lib (STL), SigProfilerAssignment (SPA) and signatureanalyzer (SA) and number and percentage of samples not reaching the 0,9 value of cosine similarity in the preclinical dataset (on the left side) and in the clinical dataset (on the right side). Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Next, we evaluated the ability of each bioinformatic tool to correctly stratify MSS-MMRp and MSI-MMRd tumors. In the preclinical dataset, the MMR deficiency signature contribution between MSS-MMRp and MSI-MMRd samples was significantly different with all the five software (Wilcoxon rank sum test, p<2e-16). Nevertheless, SPA proved to have the highest MMRd signature fitting ability as indicated by the highest median MMR signature contribution obtained in MSI-MMRd samples with this tool.

Furthermore, to properly compare the tools performance in discriminating MSS-MMRp and MSI-MMRd tumors, we analysed the ΔMMR distribution between MSI-MMRd and MSS-MMRp samples. This analysis highlighted significant differences between the contribution of MMR signatures in MSS-MMRp and MSI-MMRd using different algorithms. Notably, SPA provided the highest median separation between the two subtypes (MP=0.34, DS=0.28, STL=0.31, SPA=0.67, SA=0.28).

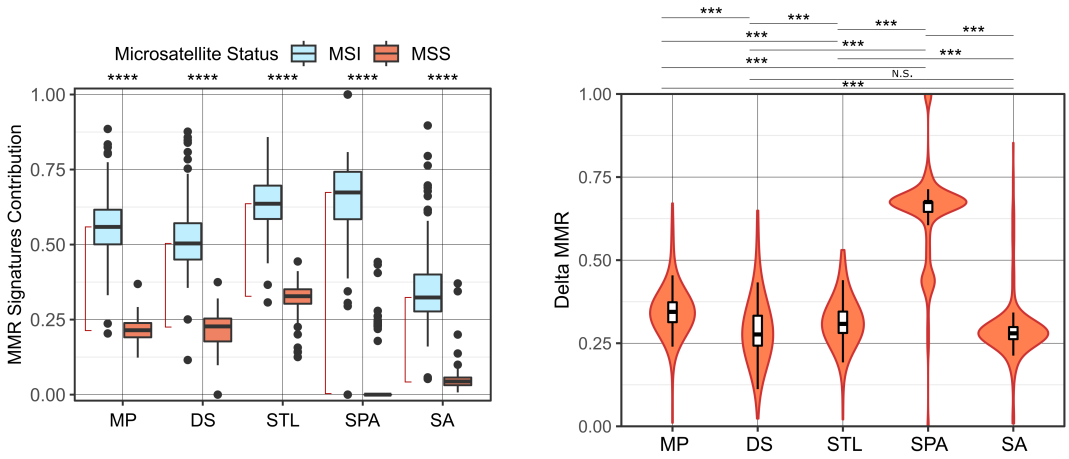


Figure 11: On the left side: overall contribution of MMR associated signatures in MSI-MMRd and MSS-MMRp CRC cell lines using MutationalPatterns (MP), deconstructSigs (DS), signature-tools.lib (STL), SigProfilerAssignment (SPA) and signatureanalyzer (SA) in the preclinical dataset. On the right side: distribution of ΔMMR values according to the indicated algorithms in the preclinical dataset. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Finally, to evaluate how mutational signatures stratify CRC POLE-mutated phenotype, we considered the POLE-related signature SBS10A and SBS10B as ΔPOLE distribution. In the preclinical datasets, a significant difference was reported for all 5 algorithms.

Considering POLE related signatures contribution, SPA showed again the highest values (MP=0.59, DS=0.58, STL=0.59, SPA=0.7, SA=0.57) as shown in the figure 12.

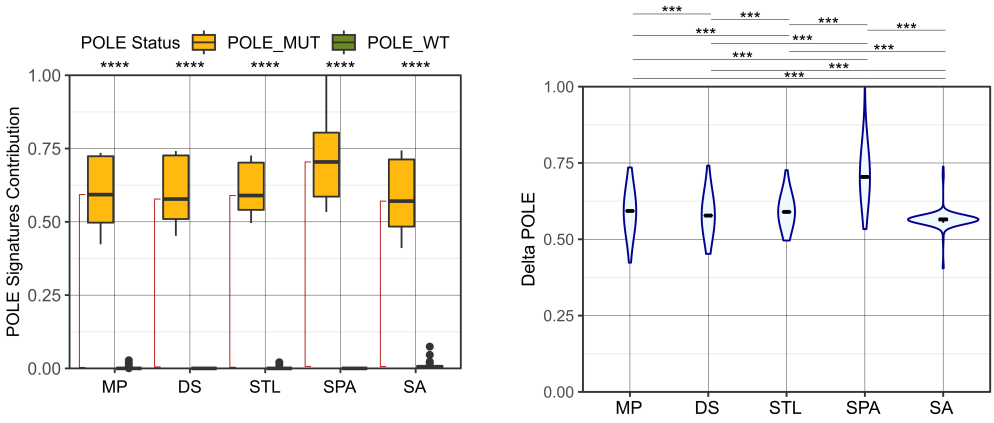


Figure 12: On the left side: overall contribution of POLE-associated signatures in MSS POLE-wild type and MSS POLE-mutated CRC cell lines using MutationalPatterns (MP), deconstructSigs (DS), signature-tools.lib (STL), SigProfilerAssignment (SPA) and signatureanalyzer (SA) in the preclinical dataset. On the right side: distribution of ΔPOLE values according to the indicated algorithms in the preclinical dataset. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

The analysis of the clinical dataset showed similar results: genetic stratification of MSI-MMRd and MSS-MMRp patients was statistically significant for all algorithms (ΔMMR MP=0.61, DS=0.67, STL=0.77, SPA=1, SA=0.26).

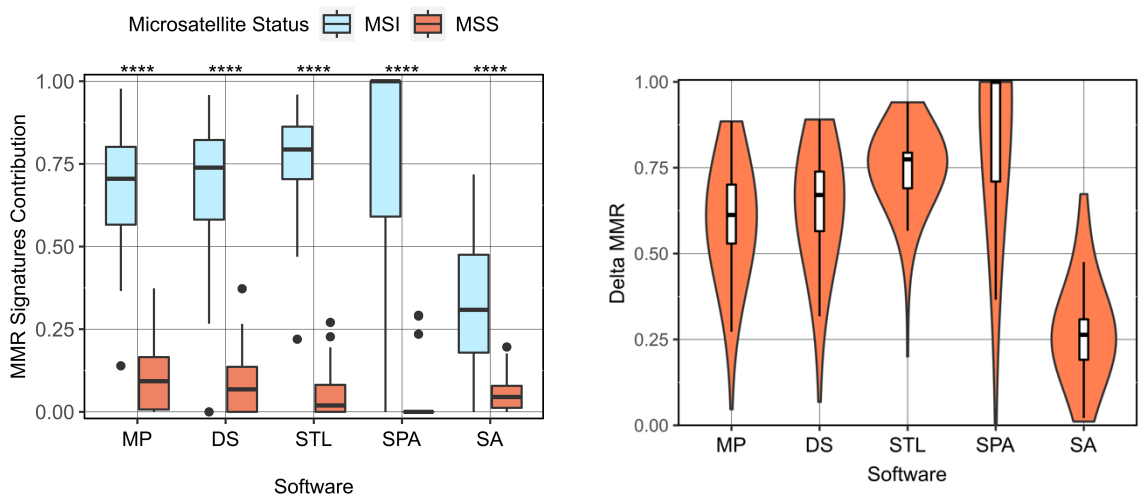


Figure 13: On the left side: overall contribution of MMR associated signatures in MSI-MMRd and MSS-MMRp CRC patients using MutationalPatterns (MP), deconstructSigs (DS), signature-tools.lib (STL), SigProfilerAssignment (SPA) and signatureanalyzer (SA) in the clinical dataset. On the right side: distribution of ΔMMR values according to the indicated algorithms in the clinical dataset. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Concordant results were also obtained for POLE stratification (ΔPOLE clinical dataset: MP=0.59, DS=0.68, STL=0.77, SPA=0.71, SA=0.85).

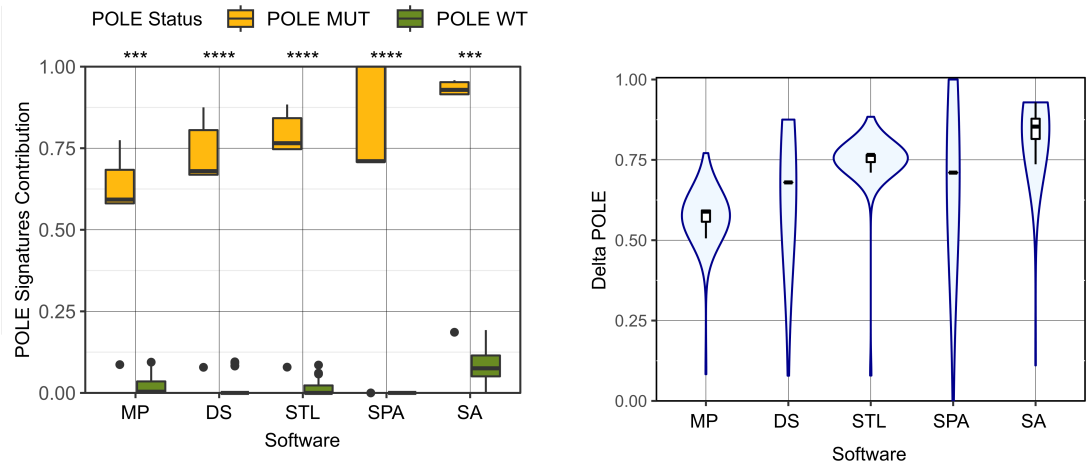


Figure 14: On the left side: overall contribution of POLE-associated signatures in MSS POLE-wild type and MSS POLE-mutated CRC patients using MutationalPatterns (MP), deconstructSigs (DS), signature-tools.lib (STL), SigProfilerAssignment (SPA) and signatureanalyzer (SA) in the clinical dataset. On the right side: distribution of ΔPOLE values according to the indicated algorithms in the clinical dataset. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Having observed a significant variability in tool performance, we decided to investigate whether these discrepancies were specific to our CRC datasets or whether they extended to other tumor types. To address this, we expanded the analysis to an independent dataset comprising samples from endometrial tumors of 483 patients. This dataset was downloaded from TCGA, encompassing both MMRp and MMRd patients. These additional analyses confirmed the consistency of our initial findings, highlighting that, depending on the tool of choice, more than 30% of the samples achieved a cosine similarity value below the 0.9 threshold (Figure 15). Finally, to further validate these divergent software performances, we evaluated different biological readouts. We focused on a dataset of 35 lung tumors from TCGA and classified them based on smoking status therefore exploiting smoking associated SBS signatures a preclinical dataset of PDOs annotated for colibactin exposure [17]. The results of these analyses further corroborated the findings from the CRC dataset, highlighting the tool-dependent variability in the detection of mutational signatures.

UCEC Dataset C2				
Tool	**Assigned samples (CS >= 0.9)**	**NOT assigned samples (CS < 0.9)**	**Fractional abundance of NOT assigned samples**	**Median DeltaMMR**
MP	337	146	30,2%	0,62
DS	309	174	36,0%	0,64
STL	321	162	33,5%	0,69
SPA	272	211	43,7%	1
SA	185	298	61,7%	0,34
UCEC Dataset C3				
Tool	**Assigned samples (CS >= 0.9)**	**NOT assigned samples (CS < 0.9)**	**Fractional abundance of NOT assigned samples**	**Median DeltaMMR**
MP	438	45	9,3%	0,47
DS	342	141	29,2%	0,46
STL	406	77	15,9%	0,53
SPA	344	139	28,8%	0,62
SA	221	262	54,2%	0,43
LUNG Dataset C2				
Tool	**Assigned samples (CS >= 0.9)**	**NOT assigned samples (CS < 0.9)**	**Fractional abundance of NOT assigned samples**	**Median DeltaSMOKER**
MP	21	14	40,0%	0,38
DS	16	19	54,3%	0,4
STL	21	14	40,0%	0,38
SPA	11	24	68,6%	0,71
SA	5	30	85,7%	0,61
LUNG Dataset C3				
Tool	**Assigned samples (CS >= 0.9)**	**NOT assigned samples (CS < 0.9)**	**Fractional abundance of NOT assigned samples**	**Median DeltaSMOKER**
MP	26	9	25,7%	0,25
DS	15	20	57,1%	0,27
STL	23	12	34,3%	0,3
SPA	19	16	45,7%	0,44
SA	2	33	94,3%	0,61
Colibactin Signature in CRC Dataset C3				
Tool	**Assigned samples (CS >= 0.9)**	**NOT assigned samples (CS < 0.9)**	**Fractional abundance of NOT assigned samples**	**Median DeltaCOLIBACTIN**
MP	12	0	0,0%	0,28
DS	10	2	16,7%	0,28
STL	12	0	0,0%	0,26
SPA	12	0	0,0%	0,32
SA	NA	NA	NA	NA

Figure 15: Table representing the number and percentage of samples that reach the threshold of cosine similarity and value of ΔMMR in endometrial cancer dataset ('UCEC dataset' from TCGA), lung cancer dataset ('LUNG dataset' from TCGA) and a preclinical dataset of PDOs. Each line represents a specific software. CS: cosine similarity; MP: MutationalPatterns; DS: decostructSigs; STL: signature.tools.lib; SPA: SigProfilerAssignment; SA: SignatureAnalyzer.

Evaluation of the reference mutational signatures on mutational signature analysis

Following the same strategy as above, we assessed how the mutational signature reference impacts mutational signature fitting and CRC molecular stratification. We selected three distinct references: COSMIC v2 (C2), COSMIC v3.2 (C3), and a CRC tissue-specific signature catalogue (TS), each containing a different number of mutational signatures (30 in C2, 72 in C3 and 26 in TS). Specifically, the CRC tissue-specific signatures were selected from literature, including signatures characterised in CRC patients from the analysis of large pan-cancer datasets of WGS and WES [1].

Cosine similarity analysis showed median values above the reliability threshold with all references, with higher values corresponding to larger references. Differences were statistically significant in both the preclinical (Wilcoxon rank sum test, C2 vs TS, C2 vs C3, C3 vs TS, respectively p=1.2e-13, p<2.2e-16, p<2.2e-16) and the clinical dataset.

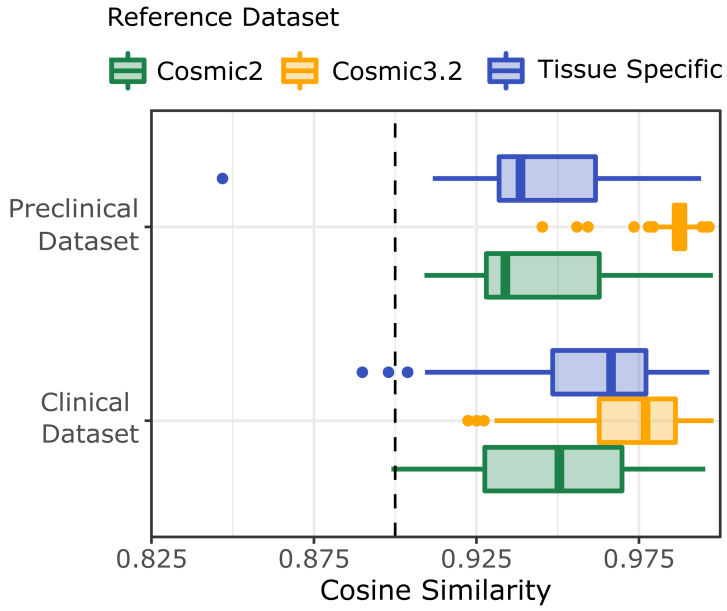


Figure 16: Distribution of cosine similarity using three different signatures references in the clinical and preclinical datasets.

With respect to the ability to define CRC molecular subsets, all references obtained a significant ΔMMR, thus allowing proper identification of MSS-MMRp and MSI-MMRd (Wilcoxon rank sum test, C2vsTS, C2vsC3, C3vsTS, p<2e-16) even if minor differences

were present (preclinical dataset: ΔMMR C2=0.34, ΔMMR C3= 0.26, ΔMMR TS=0.27; clinical dataset: ΔMMR C2=0.61, ΔMMR C3=0.41, ΔMMR TS=0.56).

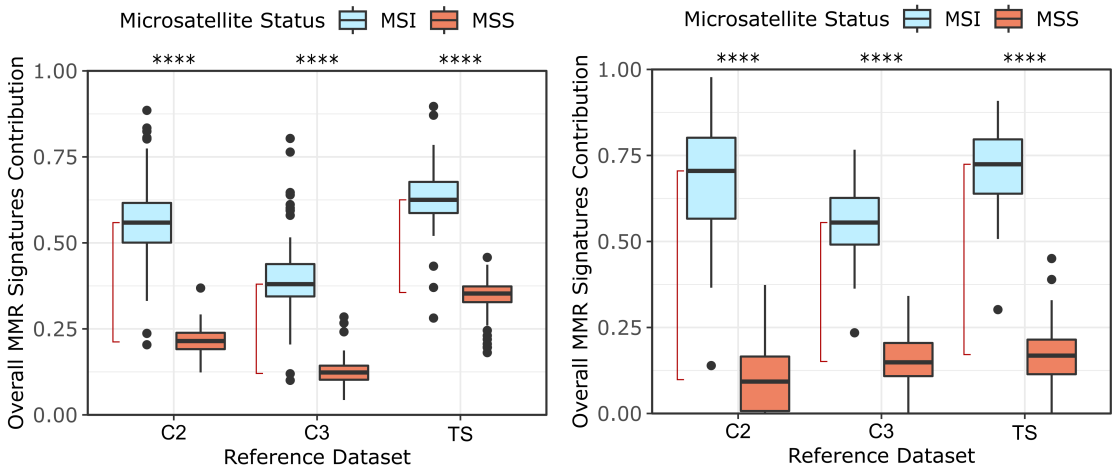


Figure 17: On the left side: contribution of MMRd-associated signatures in the CRC cell line dataset using three different signature references; Red line represents ΔMMR. On the right side: contribution of MMRd-associated signatures in the clinical dataset using three different references; Red line represents ΔMMR. C2: Cosmic v2; C3: Cosmic v3.2; TS: CRC tissue specific. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

To further investigate if the reference choice could alter the contribution of a distinct mutational signatures associated with MMR deficiency, we compared the contribution of each signature associated with MMR deficiency in the MSI-MMRd cohort of the preclinical dataset. Of note, a certain variability was present, particularly in case of SBS6 (46 % in C2, 13% in C3 and 24% in TS), SBS15 (3% in C2, 19% in C3 and 9% in TS) and SBS26 (5% in C2, 0% in C3 and 25% in TS).

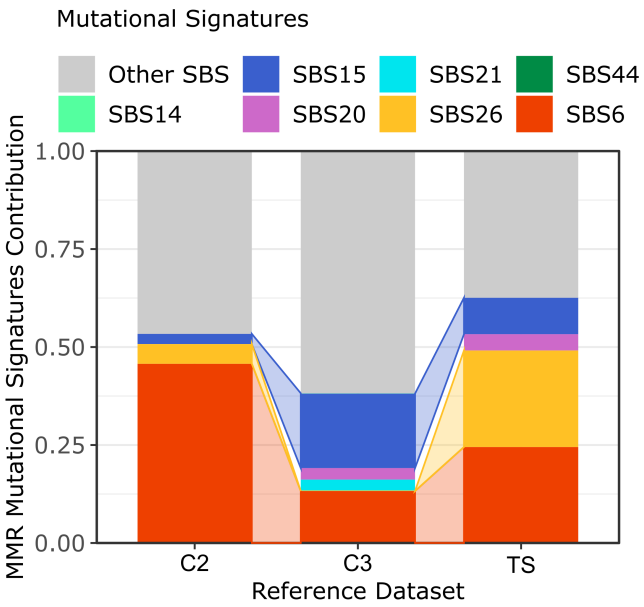
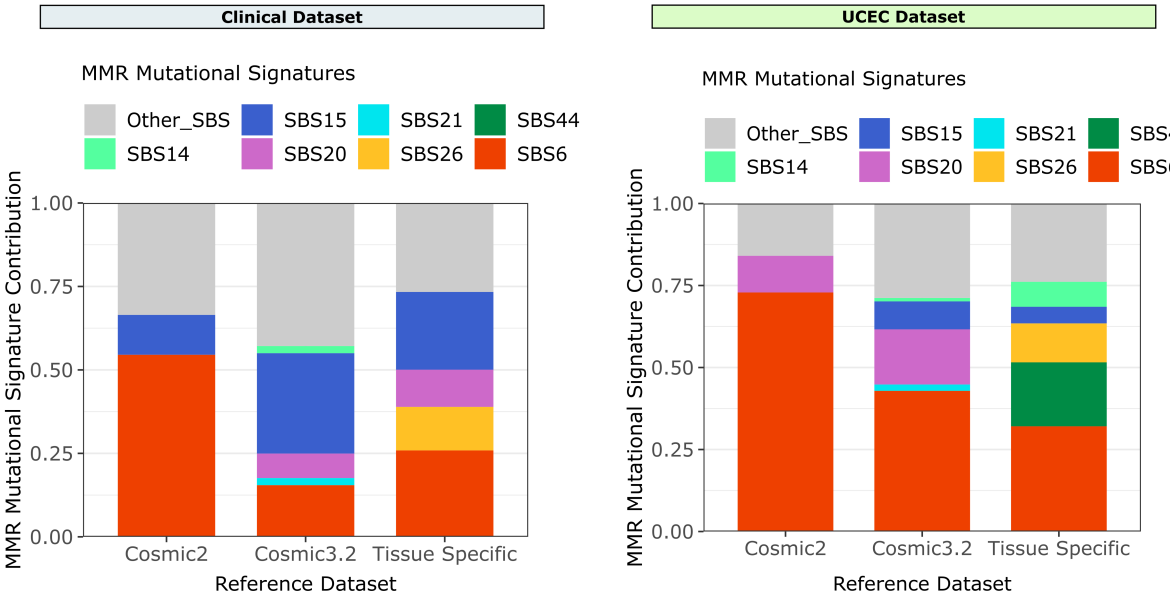


Figure 18: Normalised contribution of single MMR-associated signatures in the MSI-MMRd subset of the CRC cell line dataset. C2: Cosmic v2; C3: Cosmic v3.2; TS: CRC tissue specific.

Comparable results were obtained when we evaluated the contribution of specific MMRd signature in the clinical dataset. Furthermore, we investigated whether the emergence of individual MMRd SBS signatures was associated with CRC-specific datasets. Therefore, we performed the analysis in an independent dataset of 167 endometrial cancers annotated for MSI-MMRd status. Even in this scenario, the use of different mutational signature references led to changes in the contribution of individual signatures: SBS6 decreased from 73% in C2 to 43% and 32% respectively in C3 and in the TS references, respectively; while SBS21 appeared only in C3, SBS26 and SBS44 appeared only using the TS reference.



Clinical Dataset							
Mutational Signature Reference	SBS6	SBS14	SBS15	SBS20	SBS21	SBS26	SBS44
C2	55%	0%	12%	0%	0%	0%	NA
C3	15%	2%	30%	7%	2%	0%	0%
TS	26%	NA	23%	11%	0%	13%	0%

UCEC Dataset							
Mutational Signature Reference	SBS6	SBS14	SBS15	SBS20	SBS21	SBS26	SBS44
C2	73%	0%	0%	11%	0%	0%	NA
C3	43%	1%	9%	17%	2%	0%	0%
TS	32%	8%	5%	NA	NA	12%	19%

Figure 19: On the left side: normalised contribution of single MMR-associated signatures in the MSI-MMRd subset of the clinical dataset. On the right side: normalised contribution of single MMR-associated signatures in the MSI-MMRd subset of the endometrial cancer (UCEC project from TCGA).

Finally, we considered *POLE* genetic stratification: in both CRC datasets, all references led to effective discrimination of *POLE*-mutated from *POLE* wild-type CRCs (Wilcoxon rank sum test, C2vsTS, C2vsC3, C3vsTS, p<2e-16) (preclinical dataset: respectively ΔPOLE=0.59, ΔPOLE=0.47, ΔPOLE=0.51; clinical dataset: ΔPOLE=0.59, ΔPOLE=0.59, ΔPOLE=0.61).

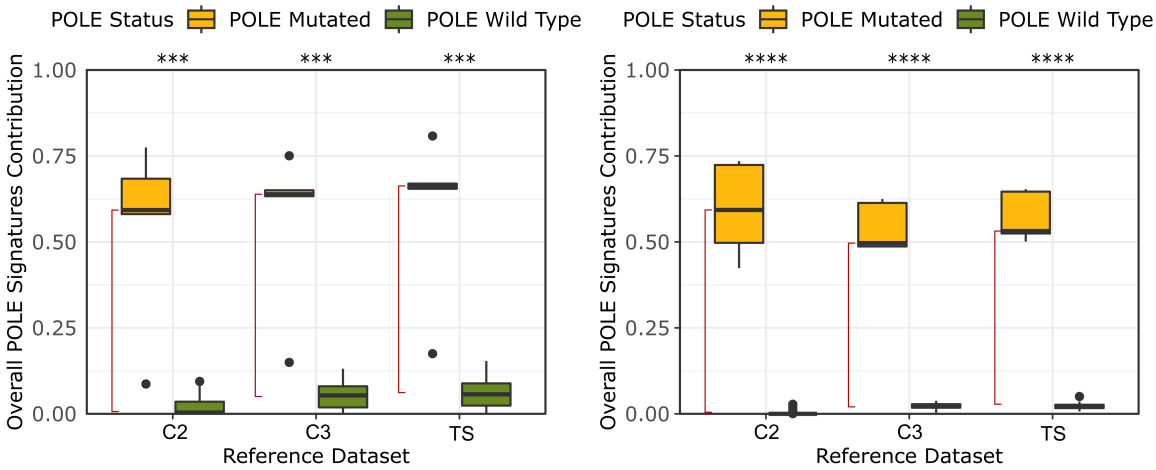


Figure 20: On the left side: Contribution of POLE-associated signatures in the clinical dataset using three different references; Red line represents ΔPOLE. On the right side: Contribution of POLE associated signatures in the preclinical dataset using three different references; Red line represents ΔPOLE. C2: Cosmic v2; C3: Cosmic v3.2; TS: CRC-tissue specific. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

In summary, the size and composition of the mutational signature reference can impact the molecular stratification of CRC samples, specifically when distinct mutational signatures are considered.

Inferring a minimum number of mutations for reliable mutational signature analysis

The discrepancy observed in the WGS based analysis between its high technical reliability and its lower effectiveness to stratify CRC samples when compared to smaller size NGS workflows was unexpected (Figure 7). To further investigate this aspect, we inferred the minimum number of mutations required to achieve a reliable mutational signature fitting. Specifically, using both the CRC cell lines and the clinical dataset, we randomly sampled from 5 to 95% of all the mutations in each sample. Next, to establish the minimum number of mutations required to obtain technically robust results, we evaluated the cosine similarity in each subset. In the CRC preclinical dataset, 323 mutations were needed to reach the cosine similarity reliability threshold. The value plunged to 64 mutations for the clinical dataset.

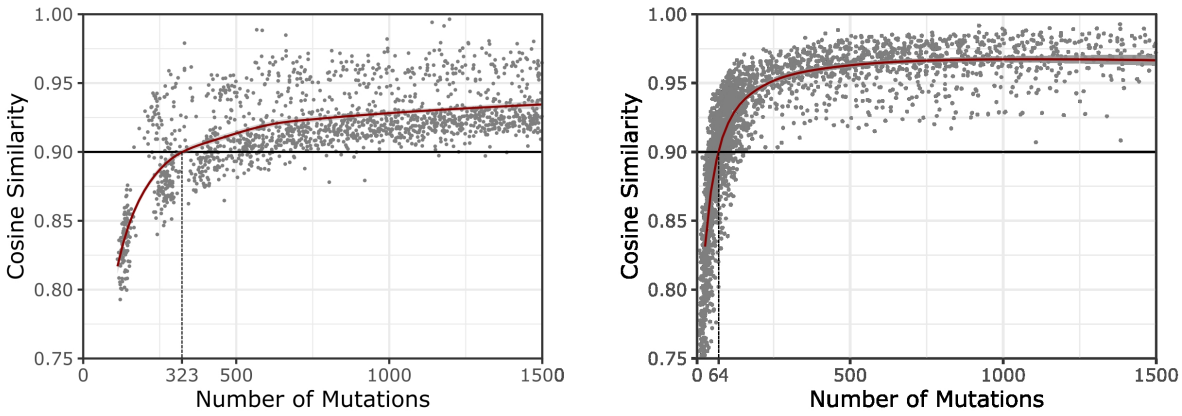


Figure 21: On the left side: sampling experiment on the CRC cell line dataset showing that at least 323 mutations are required to reach the threshold of cosine similarity for performing the analysis. On the right side: sampling experiment on the clinical dataset showing that at least 64 variants are required to reach the threshold of cosine similarity for performing the analysis.

We reasoned that this discrepancy could be related to the specific features of the two datasets. Indeed, whilst the clinical datasets contain CRC samples matched with healthy tissue, the preclinical CRC dataset lacks a non-malignant control line. To understand the impact of this discrepancy, we investigated to what extent the use of a matched normal affect mutational signature calling by decreasing the background originating from germinal variants and sequencing artefacts. For this purpose, we established a

'metanormal' obtained from 21 PBMCs of CRC patients and performed the mutational calling of the entire CRC cell line dataset using the metanormal as a normal sample [32]. In this instance, the number of mutations required to reach the cosine similarity threshold decreased from 323 to 145 (-55%).

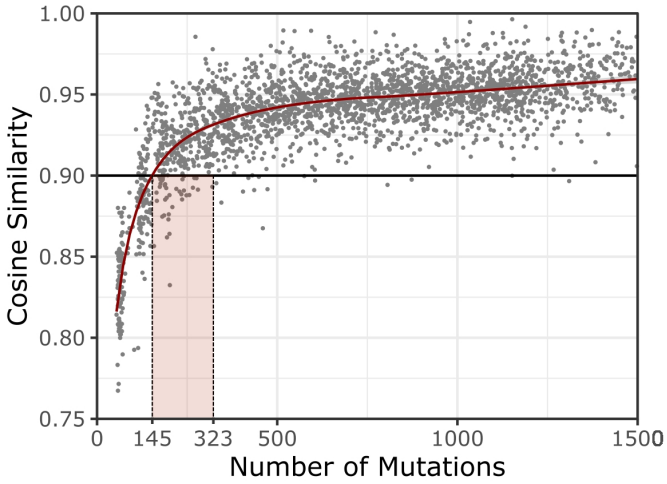


Figure 22: Sampling experiment in the CRC cell line dataset using a metanormal as matched normal in the variant calling analysis. The red interval is showing the decrease of total number of mutations that are needed for reaching the threshold of cosine similarity without and with the use of a metanormal as healthy control (from 323 to 145 -55%).

Finally, we investigated how the use of a metanormal could impact the occurrence of mutational signatures associated with artefacts: the overall signal of artefact SBS signatures dropped from 0.30 to 0.15 (-50%), thus confirming the effectiveness of this approach.

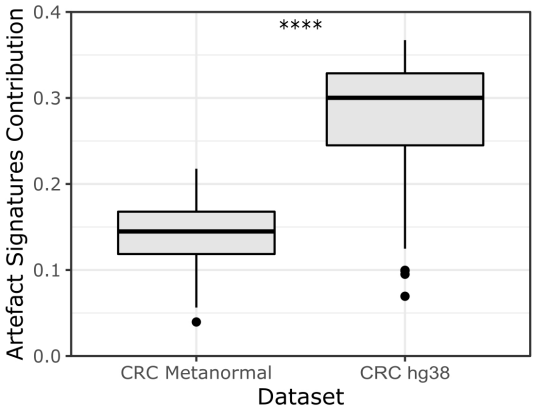


Figure 23: Boxplot highlighting the 50% reduction in the contribution of mutational signatures associated with artefacts in the CRC cell line dataset using a metanormal as the matched normal sample. Wilcoxon rank sum test was performed '*', '**', '***', '****' footnotes were used to mark significance level, respectively $P < 0.05$, $P < 0.01$, $P < 0.001$, $P < 0.0001$.

Comparative Mutational Signature analysis on Coding and Extragenic Regions (CoMSCER)

Finally, we reasoned that a bioinformatic tool which comprehensively and systematically performs the above-mentioned analyses in multiple datasets originating from distinct tumor types is not available and could be exploited by researchers to assess analysis robustness by evaluating several computational workflow performances. To address this knowledge gap, we developed CoMSCER, a freely available bioinformatic tool. By specifying the SBS mutational signatures of interest (e.g., MMRd, treatment induced), and two given conditions (e.g. MMRp vs MMRd, pre- vs post-treatment), CoMSCER evaluates the mathematical and biological readouts from multiple bioinformatic tools, reference datasets and differential signature contribution between coding and non-coding regions, providing researchers with information on the consistency of their analysis (https://github.com/pbattuello/CoMSCER).

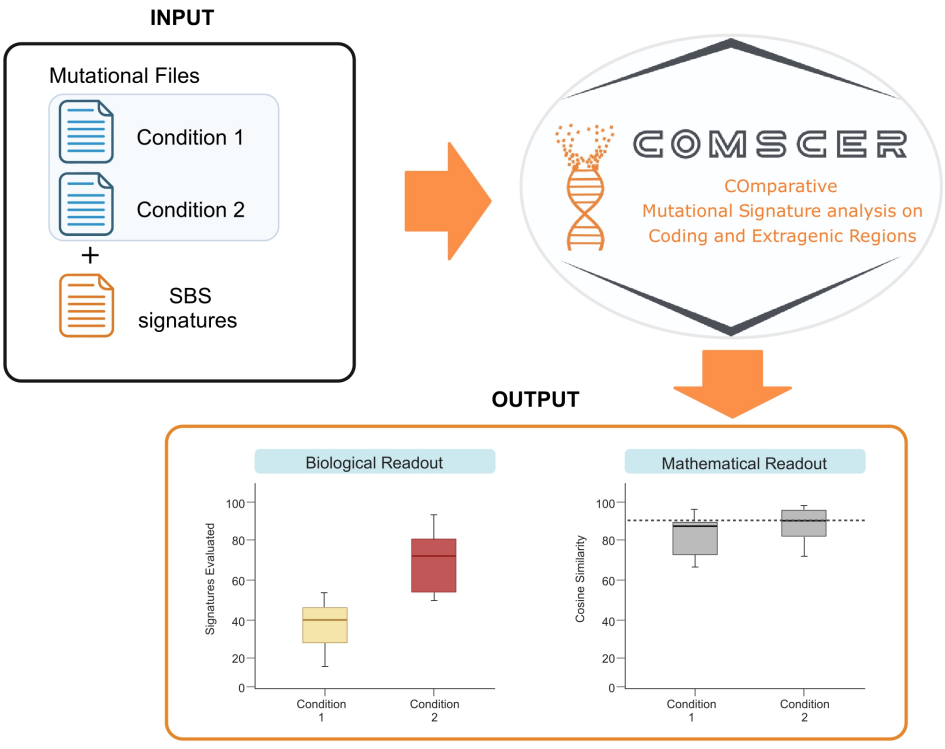


Figure 24: The CoMSCER workflow is graphically represented: given two conditions and a list of SBS signatures to analyse, the tool is able to perform both biological and mathematical evaluation.

Discussion and Conclusion

Assessing the mutational signatures that characterize cancer genomes has biological and clinical implications, as reported in breast, colorectal, and melanoma tumors [30, 32, 50]. Currently, standardized methods to perform mutational signature analysis unfortunately are not available causing a lack of reproducibility and robustness of the results. In addition, there are no comparative studies or tools available to help in identifying the most appropriate bioinformatics workflow for a specific cancer type. Therefore, to improve the reproducibility and robustness of mutational signature calls, the implementation of standardised workflows is needed as well as computational tools able to identify the influence of distinct variables of the analysis. In light of these needs, we decided to contribute by analysing the main variables that could impact mutational signature analyses.

In this context, we used CRC as a model to investigate how different methodological approaches affect the determination of mutational signatures. The choice of using CRC as a model was driven mainly by CRC genetics landscape: CRC presents a well described molecular background, with the MSS subtype being the most representative (80-85% of patients), generally characterised with low level of tumor mutational burden (TMB). Conversely, the MSI subtype is representative only for a small proportion of CRC (10-15%), a percentage that decrease to 5% if we consider only stage IV CRC. MSI tumors usually show a malfunction in the MMR machinery, most often caused by genetic mutations or promoter methylation of one of the MMR genes. This is associated with a high tumor mutational burden and in most cases, a good response to immunotherapy treatment. Finally, a small subset of MSS tumor, accounting for about 1-3% are characterised by mutations in the regions coding for the exonuclease domain of the polymerase epsilon gene (*POLE*) resulting in an hypermutated phenotype and a good response to immunotherapy treatment. The main reason why this stratification was particularly suitable for our study is that several mutational signatures have been associated with specific CRC subtypes. Specifically, seven distinct SBS mutational signatures are distinctive of MSI-MMRd tumors, being indeed associated with mismatch repair deficiency, while two distinct SBS mutational signatures are linked to mutations in the exonuclease domain of *POLE* gene. This represent the core of our work: being able

to stratify MSS tumor from MSI tumor based on the differential enrichment of these specific mutational signature allowed us to define a measure for mutational signatures performance.

This choice was further supported by the great availability of CRC preclinical samples that our lab has been collecting over the years [41-43]. The CRC preclinical dataset represents a valuable resource, including genetic annotation for 230 CRC cell lines encompassing microsatellite status and genetic alteration on CRC genes. Moreover, the presence of multiple NGS data, including WGS and WES, allowed us to properly compare and evaluate technical differences due to NGS data.

In addition, we reasoned that a second dataset would be crucial for a first validation; in this regard we downloaded from TCGA a clinical dataset of 152 CRC patients selected for having available a genetic annotation including microsatellite and *POLE* status. This cohort of patients was pivotal for several reasons: it allowed us to confirm data obtained in preclinical models in a proper clinical dataset; moreover, TCGA samples have matched normal PBMC, a condition not present when analysing CRC cell lines, allowing us to perform comparative studies in this regard.

Having two datasets of different origin and characteristics available, we designed an experimental procedure aiming to assess the influence of the main variable of a hypothetical mutational signature analysis workflow. Starting from the different type of NGS data, including NGS targeted panel, WES and WGS, we then benchmarked different bioinformatics tools and evaluated how the choice of different sized reference signature dataset could affect the overall result.

In order to properly evaluate different mutational signature workflows focusing on parameters that can affect the results, we decided to perform two complementary assessments: a) a mathematical evaluation, in which we calculated how accurately mutational signatures recapitulate the genetic landscape of cancer samples; b) a biological evaluation, in which we evaluated the identification of the MSI-MMRd/MSS-MMRp and the *POLE*-mutant status of CRC samples. Specifically, for the mathematical evaluation we used the cosine similarity as a measure of robustness: we assessed the cosine similarity of each sample between the original 96 base substitution profile and the same profile reconstructed using the contribution of the fitted mutational

signatures. This analysis resulted in a value spanning from 0 to 1 describing the similarity of these two 96-values vectors, in which 0 correspond to highly different profiles while 1 indicate equality. In brief, the higher is the cosine similarity the better the fitted mutational signatures are able to describe the mutational profile of the sample and provide reliable results; in this view we set 0.9 as a cut-off based on the current literature [49]. Conversely, the biological evaluation was performed with the aim of measuring the exploitability of mutational signature analysis in CRC. By considering all MMR associated SBS signatures, we defined the ΔMMR value as the difference between the median contribution of MMR associated signatures between MSS-MMRp and MSI-MMRd samples. In the same manner, we defined the ΔPOLE value for assessing the difference between MSS-MMRp *POLE*-wt and MSS-MMRp *POLE*-mut CRC sample.

As a further validation, we conducted our analysis on three independent datasets encompassing tumors of different histologies and molecular characteristics, including a cohort of endometrial cancer patients, a cohort of lung cancer patients and a dataset of CRC PDOs.

Concerning the NGS data type, our results showed that the use of WGS data may not improve the ability to stratify biologically relevant CRC subtypes, highlighting the importance of appropriate experimental design for mutational signature analysis. In particular, we found that focusing on the coding regions for mutational signature fitting improved CRC stratification. Given the enrichment in coding sequences of the currently available NGS targeted panels, this finding becomes particularly relevant from a clinical perspective. Accordingly, we found that performing mutational signature fitting using large pan-cancer targeted gene panels for CRC subtypes stratification is technically effective, reliable, and robust in terms of biological outcomes. Furthermore, we found differential contribution of *'flat-signatures'* when performing mutational signature analysis on exonic, intronic and extragenic regions. We speculated that this could be the effect of both technical reasons as for example differential filtering of germinal variants from coding and non-coding or also biologically related, as for example due to differential effectiveness of the DNA repair systems between coding and extragenic, or different selective pressure of mutations occurring in protein-coding genes.

In addition, we found that the choice of the algorithm led to statistically different results. In this regard, our study has limitations: for pragmatic reasons, we focused on five of the most used algorithms for performing mutational signature analysis, however more than 30 different tools are currently present in literature (as of July, 2023). Furthermore, we selected a specific version of each of the 5 software and we cannot exclude that the results could slightly differ depending on the versions. Overall, we found that MP was the best choice in the CRC cell lines. In contrast, SPA was the preferred choice for CRC molecular stratification. To extend the benchmarking to a broader context, we further compared the tools with respect of other aetiological and molecular tumor features; these included neoplasms with distinct DNA repair deficiencies, tumors associated with tobacco smoke and colibactin exposure such as samples from endometrial and lung cancer patients and a preclinical dataset of CRC PDOs. These extended analyses further confirmed that the level of performance of MP exceeded that of other tools we evaluated.

Lastly, the mutational signature reference is also relevant to the outcome of the analysis and should be chosen depending on the biological question. According to our results, reducing the number of signatures in the reference improved the stratification of CRC subtype (MSI-MMRd, MSS-MMRp, *POLE*-mutated), suggesting that TS or C2 repositories might be a better choice compared to C3 once ascertained that they contain all the signatures to be investigated in a particular experimental setting. Additionally, we have shown how the contribution of specific signatures varies depending on the mutational signature reference. This point becomes particularly relevant when evaluating the contribution of a single signature or of a small subset of signatures that are linked to a specific aetiology. This is the prototypical situation that can arise in a clinical setting when particular signatures are used to detect or monitor a certain phenotype; this is exemplified by the case of the ARETHUSA clinical trial with the monitoring of the alkylating agent exposure associated SBS11, in this case a single SBS related signature was monitored highlighting the important of accounting of signature enrichment fluctuations.

The number of mutations represents a key point among mutational signatures limitations: this has implications mainly with performing mutational signature analysis

of small NGS targeted gene panels, but also when the analysis is on other tissue sources, as cfDNA or samples low tumor content. Therefore, in order to investigate this aspect, we performed random sampling of mutations for each sample of both preclinical and clinical datasets, obtaining subsets spanning from 5% to 95% of the original mutation lists. On these samples we perform mutational signature analysis and cosine similarity evaluation, aiming to define a specific number of mutations above which the fitting analysis could provide robust results.

Our study showed that the threshold for a reliable analysis depends on both the quantity and quality of mutations, considering artefacts and germline mutations. This was particularly evident when comparing the results from the two datasets. In the clinical dataset, characterised by matched tumour-normal samples, the number of mutations required to achieve a cosine similarity of 0,9 was 64. The preclinical dataset, on the other hand, showed a cut-off of 323 mutations, underlying a significant difference between the datasets. The impact of germline mutations was further confirmed in the preclinical dataset when a 'metanormal' sample obtained from a pool of healthy PBMC WES data was used as matched normal sample for the variant calling analysis, resulting in a reduction from 323 to 145 mutations. Relatedly, we observed a 50% reduction in artefacts associated signature levels when using only unmatched somatic variants, suggesting the importance of matched normal or 'metanormal' samples to enhance mutational signature profiling. This result extends the problem of reproducibility and robustness of the analysis far beyond the mere fitting procedure but to include mutation calling and appropriate filtering steps. Proper germline mutation filtering is one example of all mutations that are contributing to the dilution of biologically driven signals, but sequencing artefact, formalin-induced artefacts, and many others should be carefully taken into account.

Finally, to further improve the useability of our results and to help researchers to test the performance of multiple workflows in their setting, we developed CoMSCER, a bioinformatic tool which streamlines mutational signature analysis by evaluating the impact of multiple variables on the mutational signature profile. Specifically, by enabling users to quickly access parallel analyses using multiple algorithms and various mutational signature references, it can provide valuable insights into the reliability and

consistency of the results. Moreover, CoMSCER provides information on the most appropriate reference and tool which would reduce the frequency by which samples are excluded due to cosine similarity values. Finally, CoMSCER can evaluate how mutational signature profiling might vary across different genomic regions, whether coding or extragenic, in order to shed light on potential differences as highlighted in our study.

In conclusion, our study demonstrates that there is a significant degree of variability in mutational signature analysis that is attributable to differences in the bioinformatic tools, reference datasets, and input data types employed. Our findings illustrate that arbitrary decisions in computational workflows can result in statistically different outcomes, potentially impacting the accuracy of mutational signature enrichment. Furthermore, we discovered that mutational signatures exhibit distinct patterns between coding and non-coding regions, and that a minimum number of mutations is necessary for reliable analysis. To address these challenges, we developed CoMSCER, a bioinformatics tool designed to shed light reproducibility of mutational signature analysis across various genomic regions and datasets.

The study was published this May 2024, on *'Briefings in Bioinformatics'* (ISSN 1477-4054 - Oxford University Press).
Battuello, Paolo et al. "Mutational signatures of colorectal cancers according to distinct computational workflows." *Briefings in bioinformatics* vol. 25,4 (2024): bbae249. doi:10.1093/bib/bbae249.

Materials and Methods

Datasets

The main datasets include a preclinical dataset comprising a collection of 230 CRC cell lines (Table1) maintained as previously reported [41-43, 51], and a publicly available clinical datasets from Genomic Data Commons (GDC) data portal repository under the TCGA project (TCGA-COAD) (Table2) [44]. The validated datasets include a cohort of 483 endometrial cancer patients (TCGA-UCEC), 35 lung cancer patients (TCGA-LUAD and TCGA-LUSC), and 12 CRC PDOs .

Genetic analysis

Maxwell® RSC Blood DNA Kit was used for DNA extraction from cell lines and the preparation was performed following the manufacturer's protocol. Starting from 400 ng of DNA from cell lines, WGS libraries were prepared using Nextera™ DNA Flex Library Preparation Kit according to the manufacturer's protocol. For the preclinical dataset, fastq files were generated from Illumina Novaseq6000 and processed using the genomic analysis workflow as previously described [41]. BWA-mem algorithm was used to map reads to the human genome version 38 and PCR duplicates were removed using the RMDUP function in the SAMtools [52, 53]. Mutations supported only by alteration in the first/last read position were filtered and strand bias correction was applied as previously described [54]. Starting from mutational files containing genetic alterations, only genetic alterations with fractional abundance>=10% were used for mutational signatures analysis. VCF files of samples in the clinical dataset (Table2) and UCEC cohort were downloaded and filtered for the availability of clinical information concerning microsatellite and *POLE* status. "MAF" files from the GDC lung cancer dataset were downloaded and filtered for genetic alterations with fractional abundance>=10% and clinical annotation concerning smoking status.

Mutational signature analysis using genomic data of different size

Mutational signature fitting analysis was performed using R (version 4.1.2), the 'MutationalPatterns' version 3.4.0 package and COSMIC v3.2 as a signatures reference in three different datasets: 230 WES CRC samples, 63 WGS and 230 NGS targeted panel

sequencing (Table 3). Concerning NGS targeted panel sequencing, TSO-500 from Illumina was chosen due to its large applicability in clinical settings and for its large size (523 genes). The TSO-500 dataset was created *in silico* from WES data upon mutations extraction based on the coding region of TSO-500 gene list. Mutational fitting was performed using 'fit_to_signatures' function with standard setting. Cosine similarity was assessed with the R function 'cos_sim_matrix' from MutationalPatterns package between the original mutational matrix (from SigProfilerMatrixGenerator) and the *reconstructed* matrix obtained using custom script publicly available on Github (https://github.com/pbattuello/MutationalSignatures). Cosine similarity distribution was plotted with 'ggplot2' R-package. Each mutational signature contribution was normalized ranging from 0 to 1, representing the percentage of mutations assigned to that specific mutational signature. As a percentage, this contribution resulted to be normalized also to the genomic size of the reference dataset: whether it was WGS, WES, or TSO-500. Normalised contributions for the mutational signatures reported on COSMIC with *'defective DNA mismatch repair'* as aetiology (SBS: 6-15-20-21-25-26-44) were taken into consideration and used for sample stratification. SBS10a-b were used instead for *POLE*-mutated sample stratification. 'Flat signatures' (SBS: 3-5-8-40-89) were defined, as previously proposed, as signatures in which the 96-mutational profile shows relatively even contribution of each trinucleotide context (<0.05%). ΔMMR was defined as the difference between the median contribution of MMRd-associated signatures between MSS-MMRp and MSI-MMRd samples. In the same manner, ΔPOLE was defined as the difference between the median contribution of POLE-associated signatures between *POLE* wild-type MSS-MMRp and *POLE*-mutated MSS-MMRp samples.

Metanormal Creation

The metanormal sample was created from WES data from 21 peripheral blood mononuclear cells (PBMCs) as previously reported [32]. For the metanormal generation, an equal number of reads were randomly taken from each of the samples and merged in a single fastq file. All the genetic analysis was repeated as described in the previous section using the metanormal sample as a matched normal.

Systematic review of bioinformatic tools to analyse mutational signatures

We conducted a literature systematic review from the publicly available repository PubMed Central (PMC) database (https://www.ncbi.nlm.nih.gov/pmc/), using as the searching key *'mutational signatures'* in the title or the abstract section. The literature search cut-off date was July 31st, 2023. From the SigProfiler suite SignatureProfilerAssignment was chosen as the most recent tool for mutational signature fitting analysis. SomaticSignatures tool was not available for fitting analysis. The 5 tools with most occurrences were included in the manuscript analysis unless the software was not available for use. Table 4 provides a comprehensive overview.

Mutational Signature Analysis – Algorithms comparison

Starting from mutational call files from WES, mutational matrices were generated using SigProfilerMatrixGenerator version 1.1.31. Then, mutational signature fitting was evaluated using five algorithms from current literature: 'signature.tools.lib' version 2.1.2, 'SignatureAnalyzer' version 0.0.8, 'SigProfilerAssignment' version 0.0.7, 'deconstructSigs' version 1.9.0 and 'MutationalPatterns' version 3.4.0. All algorithms were run in standard settings or following authors guidelines to minimise differences due to arbitrary settings and highlight differences due to different fitting approaches. Cosine similarity was calculated between the original mutational profile and the one reconstructed upon mutational signature fitting analysis using `cos_sin_matrix()` function from `MutationalPatterns` R-package. 230 cell lines from CRC cell bank and 132 samples (20 samples annotated as MSI-L were excluded from the analysis) from the clinical dataset were used in this analysis. Based on both mathematical and biological evaluations `MutationalPatterns` was chosen as the tool most suited for CRC samples and therefore used in the other results and as part of CoMSCER analysis.

Mutational Signature Analysis - Reference evaluation

Mutational signature analysis was performed on WES data using `MutationalPatterns` and COSMIC v2, v3.2 and CRC-specific as reference dataset.

Inferring a minimum number of mutations

We performed random sampling by 5% using the '*shuf*' function version 8.30 from 'GNU coreutils' for each sample of the two datasets (Table1, Table2). 19 subgroups of mutations (from 5% to 95% using 5% interval) were identified for each sample; five different replicates were created for each subset and mutational signatures fitting analysis was performed for each subset as described in the previous methods section. Cosine similarity was calculated for each sample as reported and the median value was plotted using R-package ggplot2 version 3.3.5.

Statistical Analysis

Statistical analysis was performed using R version 4.1.2. The individual statistical tests are specified in the results section and figure legends. Wilcoxon rank sum test was performed using R function '*wilcox.test*' and "*", "**", "***", "****" footnotes were used to mark significance level, respectively $p<0.05$, $p<0.01$, $p<0.001$, $p<0.0001$.

Data/Code availability

All the code and data necessary to reproduce the study are available on GitHub repository (https://github.com/pbattuello/MutationalSignatures). NGS data are available at the European bioinformatics institute in the European Nucleotide Archive (ENA) with PRJEB33045, PRJEB33640, PRJEB57691 and PRJEB61897 accession codes. Cell lines were selected based on the availability of genomic data from NGS (Table3). Compared to the datasets we reported previously, additional cell lines WGS were included in the current cohort. Idea tool for mutational calling pipeline is available at (https://bitbucket.org/irccit/idea/src/master/) [55]. CoMSCER is available at https://github.com/pbattuello/CoMSCER.

Table 1:

CELL_LINE	Microsatellite	*POLE* STATUS	*APC*	*TP53*	*KRAS*	*BRAF*	*PIK3CA*
B1003_XL5	MSI	WT	WT	MUT	WT	MUT	MUT
B1003_XLSP	MSI	WT	WT	MUT	WT	MUT	MUT
C10	MSI	WT	WT	WT	WT	WT	WT
C106	MSS	WT	MUT	MUT	MUT	WT	WT
C125PM	MSS	WT	MUT	MUT	MUT	WT	WT
C146	MSI	WT	MUT	MUT	MUT	WT	MUT
C170	MSI	WT	MUT	MUT	MUT	WT	MUT
C32	MSS	WT	MUT	MUT	WT	WT	WT
C70	MSS	WT	MUT	WT	WT	WT	WT
C75	MSS	WT	MUT	MUT	WT	WT	WT
C80	MSS	WT	MUT	MUT	MUT	WT	WT
C84	MSS	WT	MUT	MUT	MUT	WT	WT
C99	MSS	WT	WT	WT	WT	WT	WT
CACO2	MSS	WT	MUT	MUT	WT	WT	WT
CAR1	MSS	WT	WT	MUT	WT	WT	WT
CCK81	MSI	WT	MUT	MUT	WT	MUT	MUT
CL11	MSS	WT	MUT	MUT	MUT	WT	WT
CL14	MSS	WT	MUT	MUT	WT	WT	WT
CL34	MSI	WT	MUT	MUT	WT	MUT	MUT
CL40	MSS	WT	MUT	MUT	MUT	WT	WT
CO115	MSI	WT	MUT	WT	WT	MUT	WT
COCM1	MSS	WT	MUT	MUT	WT	WT	MUT
COGA1	MSI	WT	MUT	WT	WT	WT	MUT
COGA10	MSI	WT	MUT	WT	WT	WT	MUT
COGA12	MSI	WT	MUT	WT	WT	WT	WT
COGA2	MSS	WT	MUT	MUT	MUT	WT	WT
COGA3	MSI	WT	WT	WT	MUT	WT	WT
COGA5	MSS	WT	MUT	WT	MUT	WT	WT
COGA5L	MSS	WT	MUT	WT	MUT	WT	WT
COGA8	MSS	WT	MUT	MUT	WT	WT	WT
COLO201	MSS	WT	MUT	MUT	WT	MUT	WT
COLO205	MSS	WT	MUT	MUT	WT	MUT	WT
COLO320	MSS	WT	MUT	MUT	WT	WT	WT
COLO320DM	MSS	WT	MUT	MUT	WT	WT	WT
COLO320HSR	MSS	WT	MUT	MUT	WT	WT	WT
COLO60H	MSI	WT	MUT	WT	MUT	WT	WT
COLO678	MSS	WT	MUT	WT	MUT	WT	WT
COLO94H	MSS	WT	MUT	WT	MUT	WT	MUT
CR4	MSS	WT	MUT	MUT	MUT	WT	WT
CRC0078_XL	MSS	WT	MUT	MUT	WT	WT	WT
CRC0080_XL	MSS	WT	WT	MUT	WT	WT	WT
CRC0081_XL	MSS	WT	MUT	WT	WT	WT	WT
CRC0104_XL	MSS	WT	MUT	MUT	WT	WT	MUT
CRC0174_XL	MSS	WT	MUT	MUT	WT	WT	WT
CRC0186_XL	MSS	WT	MUT	WT	WT	WT	WT
CRC0313_XL	MSS	WT	MUT	MUT	WT	WT	WT
CRC0322_XL	MSS	WT	MUT	MUT	WT	WT	WT
CRC0394_XL	MSS	WT	MUT	WT	WT	WT	WT
CRC0438_XL	MSS	WT	WT	MUT	MUT	WT	WT
CRC0558_XL	MSS	WT	WT	MUT	WT	WT	WT
CRC0691_XL	MSS	WT	MUT	MUT	MUT	WT	WT
CRC0740_XL	MSS	WT	MUT	MUT	WT	WT	MUT
CRC0778_XL	MSS	WT	WT	WT	MUT	WT	NA
CRC0781_XL	MSS	WT	MUT	MUT	WT	MUT	NA
CRC1360_XL	MSS	WT	MUT	MUT	MUT	WT	WT
CW2	MSI	WT	MUT	WT	MUT	WT	MUT
CX1	MSS	WT	MUT	WT	WT	MUT	MUT
DIFI	MSS	WT	MUT	MUT	WT	WT	WT
DLD1	MSI	WT	MUT	MUT	MUT	WT	MUT
FET	MSS	WT	MUT	MUT	MUT	WT	WT
FHC	MSS	WT	WT	WT	WT	WT	WT
GEO	MSI	WT	MUT	MUT	MUT	WT	WT
GP2D	MSI	WT	MUT	WT	MUT	MUT	MUT
GP5D	MSI	WT	MUT	WT	MUT	MUT	MUT
HCA24	MSS	MUT	MUT	WT	WT	WT	MUT
HCA46	MSS	WT	MUT	MUT	WT	WT	MUT

CELL_LINE	Microsatellite	*POLE* STATUS	*APC*	*TP53*	*KRAS*	*BRAF*	*PIK3CA*
HT115	MSS	MUT	MUT	MUT	WT	MUT	MUT
HT29	MSS	WT	MUT	WT	WT	MUT	MUT
HT55	MSS	WT	MUT	MUT	WT	MUT	WT
HUTU80	MSS	WT	WT	WT	WT	WT	WT
IRCC1_XL	MSI	WT	MUT	MUT	WT	WT	WT
IRCC1_XLRES	MSI	WT	MUT	MUT	WT	WT	WT
IRCC10_A_HL	MSS	WT	MUT	WT	WT	WT	WT
IRCC3_HL	MSI	WT	MUT	MUT	MUT	MUT	WT
IRCC3_XL	MSI	WT	MUT	MUT	MUT	WT	WT
IRCC5_A_XL	MSS	WT	MUT	MUT	WT	WT	WT
IRCC5_B_XL	MSS	WT	MUT	MUT	WT	WT	WT
IRCC72_A_XL	MSS	WT	MUT	MUT	WT	WT	WT
ISRECO1	MSS	WT	MUT	MUT	WT	WT	WT
JVE015	MSS	WT	MUT	MUT	MUT	WT	MUT
JVE017	MSS	WT	MUT	MUT	WT	MUT	WT
JVE044	MSS	WT	MUT	WT	MUT	WT	WT
JVE059	MSI	WT	MUT	WT	WT	WT	WT
JVE103	MSS	WT	MUT	MUT	WT	WT	WT
JVE109	MSI	WT	MUT	MUT	WT	MUT	WT
JVE114	MSS	WT	MUT	MUT	WT	WT	WT
JVE127	MSS	WT	WT	MUT	WT	MUT	WT
JVE187	MSS	WT	MUT	WT	MUT	WT	MUT
JVE192	MSI	WT	WT	WT	MUT	WT	MUT
JVE207	MSS	WT	MUT	MUT	WT	MUT	WT
JVE241	MSS	WT	MUT	MUT	MUT	WT	WT
JVE253	MSS	WT	MUT	MUT	MUT	WT	WT
JVE367	MSS	WT	WT	WT	WT	MUT	WT
JVE371	MSS	WT	MUT	WT	MUT	WT	MUT
JVE528	MSS	WT	MUT	MUT	MUT	WT	WT
KM12	MSI	WT	MUT	MUT	WT	MUT	WT
KM12C	MSI	WT	MUT	MUT	WT	WT	WT
KM12L4	MSI	WT	MUT	MUT	WT	WT	WT
KM12SM	MSI	WT	MUT	MUT	WT	WT	WT
KM20	MSS	WT	MUT	WT	WT	MUT	MUT
KP283T	MSS	WT	MUT	WT	MUT	WT	WT
KP363T	MSS	WT	WT	MUT	WT	MUT	MUT
KP7038T	MSI	WT	MUT	WT	WT	WT	WT
LIM1215	MSI	WT	WT	WT	WT	WT	WT
LIM1863	MSS	WT	MUT	MUT	WT	WT	WT
LIM1899	MSI	WT	WT	MUT	MUT	WT	WT
LIM2099	MSS	WT	WT	WT	MUT	WT	WT
LIM2405	MSI	WT	MUT	WT	WT	MUT	WT
LIM2412	MSI	WT	MUT	MUT	WT	MUT	MUT
LIM2537	MSI	WT	MUT	MUT	WT	MUT	MUT
LIM2550	MSI	WT	MUT	MUT	MUT	WT	WT
LIM2551	MSI	WT	WT	MUT	WT	MUT	MUT
LM0201_A_XL	MSS	WT	MUT	WT	WT	WT	WT
LM0701_A_XL	MSS	WT	MUT	WT	WT	WT	WT
LOVO	MSI	WT	MUT	WT	MUT	WT	WT
LS1034	MSS	WT	MUT	MUT	MUT	WT	WT
LS123	MSS	WT	MUT	WT	MUT	WT	WT
LS174T	MSI	WT	WT	WT	MUT	MUT	MUT
LS180	MSI	WT	WT	WT	MUT	MUT	MUT
LS411N	MSI	WT	MUT	MUT	WT	MUT	WT
LS513	MSS	WT	WT	WT	MUT	MUT	WT
MDST8	MSS	WT	MUT	WT	WT	MUT	WT
MIP101	MSI	WT	MUT	MUT	MUT	WT	MUT
NCIH498	MSS	WT	MUT	WT	MUT	WT	MUT
NCIH630	MSI	WT	MUT	MUT	WT	WT	WT
NCIH684	MSS	WT	MUT	MUT	MUT	WT	WT
NCIH716	MSS	WT	WT	WT	MUT	WT	WT
OUMS23	MSS	WT	MUT	MUT	WT	MUT	WT
OXCO1	MSS	WT	MUT	MUT	WT	MUT	WT
OXCO2	MSI	WT	MUT	WT	WT	WT	WT
OXCO3	MSS	WT	MUT	MUT	MUT	WT	WT
RCM1	MSS	WT	MUT	MUT	MUT	WT	WT

HCA7	MSI	WT	MUT	MUT	WT	WT	WT
HCC2998	MSS	MUT	MUT	MUT	MUT	WT	WT
HCT116	MSI	WT	WT	WT	MUT	WT	MUT
HCT8	MSI	WT	MUT	MUT	MUT	WT	MUT
HDC114	MSS	MUT	MUT	MUT	WT	WT	MUT
HDC135	MSI	WT	MUT	WT	WT	MUT	WT
HDC142	MSS	WT	MUT	MUT	WT	MUT	MUT
HDC143	MSI	WT	MUT	MUT	WT	WT	WT
HDC54	MSS	WT	MUT	MUT	WT	WT	WT
HDC8	MSS	WT	MUT	MUT	MUT	WT	WT
HDC82	MSS	WT	MUT	MUT	WT	WT	WT
HDC9	MSI	WT	MUT	MUT	WT	WT	MUT
HHC6548	MSS	WT	MUT	MUT	MUT	WT	WT
HRA16	MSS	WT	MUT	WT	WT	WT	WT
HROBMCO1	MSS	WT	MUT	MUT	MUT	WT	MUT
HROC107_T0M2	MSS	WT	MUT	MUT	MUT	WT	MUT
HROC112_MET	MSS	WT	MUT	MUT	WT	WT	WT
HROC113	MSI	WT	WT	MUT	MUT	WT	MUT
HROC126	MSS	WT	MUT	MUT	WT	WT	WT
HROC131_T0M3	MSI	WT	MUT	MUT	WT	MUT	WT
HROC173	MSS	WT	MUT	MUT	MUT	WT	MUT
HROC18	MSS	WT	MUT	MUT	MUT	WT	MUT
HROC183	MSS	WT	MUT	MUT	MUT	WT	WT
HROC212	MSI	WT	MUT	WT	WT	MUT	WT
HROC239_T0M1	MSS	WT	MUT	MUT	MUT	WT	WT
HROC24	MSI	WT	MUT	WT	WT	MUT	WT
HROC257_T0M1	MSI	WT	MUT	MUT	WT	MUT	MUT
HROC277_MET2	MSS	WT	MUT	MUT	MUT	WT	WT
HROC277_T0M1	MSS	WT	MUT	MUT	MUT	WT	WT
HROC278_MET	MSS	WT	WT	MUT	WT	MUT	WT
HROC278_T0M1	MSS	WT	WT	MUT	WT	MUT	WT
HROC284_MET	MSS	WT	MUT	WT	WT	WT	WT
HROC285_T2M2	MSI	WT	MUT	MUT	MUT	WT	MUT
HROC300_T2M1	MSS	WT	WT	MUT	MUT	WT	WT
HROC313_MET1	MSS	WT	MUT	MUT	MUT	WT	WT
HROC32	MSS	WT	MUT	MUT	MUT	WT	WT
HROC324	MSI	WT	MUT	MUT	MUT	WT	WT
HROC334	MSS	WT	MUT	MUT	MUT	WT	WT
HROC39	MSS	WT	MUT	WT	MUT	WT	WT
HROC40	MSS	WT	MUT	MUT	MUT	WT	WT
HROC43	MSS	WT	MUT	MUT	MUT	WT	WT
HROC46	MSS	WT	MUT	WT	MUT	WT	WT
HROC50_T1M5	MSI	WT	MUT	MUT	WT	MUT	WT
HROC57	MSS	WT	MUT	MUT	WT	MUT	WT
HROC59	MSS	WT	MUT	WT	MUT	WT	MUT
HROC60	MSS	WT	MUT	WT	MUT	WT	WT
HROC69	MSS	MUT	MUT	MUT	WT	WT	MUT
HROC80	MSS	WT	MUT	MUT	MUT	WT	WT
HROC87	MSI	WT	MUT	WT	WT	MUT	WT

RKO	MSI	WT	MUT	WT	WT	MUT	MUT
RW2982	MSS	WT	MUT	WT	MUT	WT	WT
RW7213	MSS	WT	MUT	MUT	MUT	WT	WT
SKCO1	MSS	WT	MUT	WT	MUT	WT	WT
SNU1040	MSI	WT	MUT	MUT	WT	WT	WT
SNU1047	MSI	WT	MUT	WT	WT	WT	WT
SNU1181	MSS	WT	MUT	MUT	MUT	WT	WT
SNU1235	MSS	WT	MUT	MUT	WT	MUT	WT
SNU1406	MSS	WT	WT	MUT	WT	MUT	WT
SNU1411	MSS	WT	WT	MUT	MUT	MUT	WT
SNU1460	MSS	WT	WT	WT	WT	WT	WT
SNU1544	MSI	WT	MUT	WT	MUT	WT	MUT
SNU1684	MSI	WT	MUT	WT	MUT	WT	WT
SNU1746	MSI	WT	MUT	WT	MUT	WT	WT
SNU175	MSI	WT	MUT	WT	MUT	WT	WT
SNU254	MSS	WT	MUT	MUT	MUT	WT	MUT
SNU283	MSS	WT	WT	WT	WT	WT	WT
SNU407	MSI	WT	WT	WT	MUT	MUT	MUT
SNU479	MSS	WT	MUT	MUT	WT	WT	WT
SNU503	MSS	WT	MUT	MUT	WT	MUT	WT
SNU61	MSS	WT	MUT	WT	MUT	WT	WT
SNU769B	MSI	WT	MUT	WT	WT	WT	WT
SNU81	MSS	MUT	MUT	MUT	MUT	WT	MUT
SNU977	MSS	WT	MUT	WT	WT	WT	WT
SNUC1	MSS	WT	WT	MUT	WT	WT	WT
SNUC2A	MSI	WT	WT	MUT	MUT	WT	MUT
SNUC2B	MSI	WT	WT	MUT	MUT	WT	MUT
SNUC5	MSI	WT	MUT	MUT	WT	MUT	MUT
SW1116	MSS	WT	MUT	MUT	MUT	WT	WT
SW1222	MSS	WT	MUT	WT	MUT	WT	WT
SW1417	MSS	WT	MUT	MUT	WT	MUT	WT
SW1463	MSS	WT	MUT	MUT	MUT	WT	WT
SW403	MSS	WT	MUT	MUT	MUT	WT	MUT
SW48	MSI	WT	MUT	WT	WT	MUT	MUT
SW480	MSS	WT	MUT	WT	MUT	WT	WT
SW620	MSS	WT	MUT	WT	MUT	WT	WT
SW837	MSS	WT	MUT	MUT	MUT	WT	WT
SW948	MSS	WT	MUT	MUT	MUT	WT	MUT
T84	MSS	WT	MUT	WT	MUT	WT	MUT
V411	MSS	WT	MUT	WT	MUT	WT	WT
V457	MSI	WT	MUT	WT	MUT	MUT	WT
V481	MSI	WT	MUT	WT	MUT	WT	MUT
V703	MSI	WT	WT	MUT	MUT	WT	WT
V9P	MSS	WT	WT	MUT	WT	WT	WT
VACO400	MSS	MUT	MUT	MUT	WT	MUT	MUT
VACO432	MSI	WT	MUT	WT	WT	MUT	MUT
VACO5	MSI	WT	MUT	MUT	WT	MUT	MUT
VACO6	MSI	WT	WT	WT	WT	MUT	WT
WIDR	MSS	WT	MUT	WT	WT	MUT	MUT

Table1: Genetic characterization of the CRC preclinical dataset: the dataset includes 230 CRC cell lines genetically annotated. The table reports sample name, microsatellite status (MSI or MSS), and the status of few CRC genes, POLE status (MUT samples indicate hypermutant CRC cell lines), APC status (MUT samples indicate loss of function CRC cell lines), TP53, KRAS, BRAF and PIK3CA status (WT vs MUT).

Table 2:

Sample	Microsatellite Status	*POLE* Status
TCGA-5M-AAT6	MSI	WT
TCGA-A6-2672	MSI	WT
TCGA-A6-2686	MSI	WT
TCGA-A6-3809	MSI	WT
TCGA-A6-5661	MSI	WT
TCGA-A6-5665	MSI	WT
TCGA-A6-6653	MSI	WT
TCGA-A6-A565	MSI	WT
TCGA-AA-3492	MSI	WT
TCGA-AA-3663	MSI	WT
TCGA-AA-3672	MSI	WT
TCGA-AA-3713	MSI	WT
TCGA-AA-3715	MSI	WT
TCGA-AA-3811	MSI	WT
TCGA-AA-3815	MSI	WT
TCGA-AA-3821	MSI	WT
TCGA-AA-3833	MSI	WT
TCGA-AA-3845	MSI	WT
TCGA-AA-3864	MSI	WT
TCGA-AA-3877	MSI	WT
TCGA-AA-3947	MSI	WT
TCGA-AA-3949	MSI	WT
TCGA-AA-3950	MSI	WT
TCGA-AA-3966	MSI	WT
TCGA-AA-A01P	MSI	WT
TCGA-AA-A01R	MSI	WT
TCGA-AA-A022	MSI	WT
TCGA-AA-A02R	MSI	WT
TCGA-AD-5900	MSI	WT
TCGA-AD-6889	MSI	WT
TCGA-AD-6895	MSI	WT
TCGA-AD-6964	MSI	WT
TCGA-AD-A5EJ	MSI	WT
TCGA-AM-5821	MSI	WT
TCGA-AU-6004	MSI	WT
TCGA-AY-6197	MSI	WT
TCGA-AZ-4313	MSI	WT
TCGA-AZ-4615	MSI	WT
TCGA-AZ-6598	MSI	WT
TCGA-CK-4951	MSI	WT
TCGA-CK-5913	MSI	WT
TCGA-CK-5916	MSI	WT
TCGA-CK-6746	MSI	WT
TCGA-CM-4743	MSI	WT
TCGA-CM-4746	MSI	WT
TCGA-CM-5861	MSI	WT
TCGA-CM-6162	MSI	WT
TCGA-CM-6171	MSI	WT
TCGA-CM-6674	MSI	WT
TCGA-D5-6530	MSI	WT
TCGA-D5-6540	MSI	WT
TCGA-D5-6927	MSI	WT
TCGA-D5-6928	MSI	WT
TCGA-D5-6930	MSI	WT
TCGA-DM-A1HB	MSI	WT
TCGA-F4-6570	MSI	WT
TCGA-F4-6703	MSI	WT
TCGA-F4-6856	MSI	WT
TCGA-G4-6302	MSI	WT
TCGA-G4-6304	MSI	WT
TCGA-G4-6309	MSI	WT
TCGA-G4-6320	MSI	WT
TCGA-G4-6586	MSI	WT
TCGA-G4-6588	MSI	WT
TCGA-G4-6628	MSI	WT
TCGA-NH-A5IV	MSI	WT
TCGA-QG-A5Z2	MSI	WT
TCGA-WS-AB45	MSI	WT
TCGA-A6-2671	MSS	WT
TCGA-A6-2674	MSS	WT
TCGA-A6-2677	MSS	WT
TCGA-A6-2681	MSS	WT
TCGA-A6-3810	MSS	WT
TCGA-A6-4107	MSS	WT
TCGA-AA-3495	MSS	WT
TCGA-AA-3506	MSS	WT
TCGA-AA-3510	MSS	MUT
TCGA-AA-3530	MSS	WT
TCGA-AA-3666	MSS	WT
TCGA-AA-3673	MSS	WT
TCGA-AA-3675	MSS	WT
TCGA-AA-3678	MSS	MUT
TCGA-AA-3679	MSS	WT
TCGA-AA-3681	MSS	WT
TCGA-AA-3684	MSS	WT
TCGA-AA-3685	MSS	WT
TCGA-AA-3693	MSS	WT
TCGA-AA-3695	MSS	WT
TCGA-AA-3696	MSS	WT
TCGA-AA-3812	MSS	WT
TCGA-AA-3814	MSS	WT
TCGA-AA-3818	MSS	WT
TCGA-AA-3831	MSS	WT
TCGA-AA-3837	MSS	WT
TCGA-AA-3841	MSS	WT
TCGA-AA-3846	MSS	WT
TCGA-AA-3848	MSS	WT
TCGA-AA-3850	MSS	WT
TCGA-AA-3851	MSS	WT
TCGA-AA-3867	MSS	WT
TCGA-AA-3875	MSS	WT
TCGA-AA-3956	MSS	WT
TCGA-AA-3968	MSS	WT
TCGA-AA-3971	MSS	WT
TCGA-AA-3975	MSS	WT
TCGA-AA-3976	MSS	WT
TCGA-AA-3977	MSS	MUT
TCGA-AA-3980	MSS	WT
TCGA-AA-3984	MSS	MUT
TCGA-AA-3986	MSS	WT
TCGA-AA-3989	MSS	WT
TCGA-AA-3994	MSS	WT
TCGA-AA-A01I	MSS	WT
TCGA-AA-A01T	MSS	WT
TCGA-AA-A01V	MSS	WT
TCGA-AA-A01X	MSS	WT
TCGA-AA-A01Z	MSS	WT
TCGA-AA-A02H	MSS	WT
TCGA-AA-A02K	MSS	WT
TCGA-AA-A02O	MSS	WT
TCGA-AA-A02Y	MSS	WT
TCGA-AA-A03F	MSS	WT
TCGA-AA-A03J	MSS	WT
TCGA-AY-4071	MSS	WT
TCGA-AZ-4308	MSS	WT
TCGA-AZ-4315	MSS	MUT
TCGA-AZ-4682	MSS	WT
TCGA-CA-5256	MSS	WT
TCGA-CM-4744	MSS	WT
TCGA-CM-4748	MSS	WT
TCGA-CM-4752	MSS	WT
TCGA-CM-5341	MSS	WT
TCGA-A6-3808	MSI-L	WT
TCGA-AA-3502	MSI-L	WT
TCGA-AA-3680	MSI-L	WT
TCGA-AA-3688	MSI-L	WT
TCGA-AA-3819	MSI-L	WT
TCGA-AA-3852	MSI-L	WT
TCGA-AA-3854	MSI-L	WT
TCGA-AA-3855	MSI-L	WT
TCGA-AA-3861	MSI-L	WT
TCGA-AA-3866	MSI-L	WT
TCGA-AA-3930	MSI-L	WT
TCGA-AA-3982	MSI-L	WT
TCGA-AA-A004	MSI-L	WT
TCGA-AA-A00N	MSI-L	WT
TCGA-AA-A010	MSI-L	WT
TCGA-AA-A01S	MSI-L	WT
TCGA-AA-A024	MSI-L	WT
TCGA-AA-A02E	MSI-L	WT
TCGA-AZ-4614	MSI-L	WT
TCGA-CM-4747	MSI-L	WT

Table2: Genetic characterization of the CRC clinical dataset: the dataset includes 152 CRC patients annotated or microsatellite status (MSS vs MSI) and POLE status (MUT samples indicate hypermutant CRCs).

Table 3:

Whole Exome Sequencing	Whole Exome Sequencing	Whole Exome Sequencing	Whole Exome Sequencing	Whole Genome Sequencing
B1003_XL5	GP5D	IRCC72_A_XL	SNU1406	C10
B1003_XLSP	HCA24	ISRECO1	SNU1411	C125PM
C10	HCA46	JVE015	SNU1460	C80
C106	HCA7	JVE017	SNU1544	C84
C125PM	HCC2998	JVE044	SNU1684	CACO2
C146	HCT116	JVE059	SNU1746	CAR1
C170	HCT8	JVE103	SNU175	CL11
C32	HDC114	JVE109	SNU254	COGA2
C70	HDC135	JVE114	SNU283	COGA5
C75	HDC142	JVE127	SNU407	COGA8
C80	HDC143	JVE187	SNU479	COLO94H
C84	HDC54	JVE192	SNU503	CRC0080_XL
C99	HDC8	JVE207	SNU61	CRC0438
CACO2	HDC82	JVE241	SNU769B	CRC0558_XL
CAR1	HDC9	JVE253	SNU81	CRC0778
CCK81	HHC6548	JVE367	SNU977	CRC0781_XL
CL11	HRA16	JVE371	SNUC1	DLD1
CL14	HROBMCO1	JVE528	SNUC2A	FET
CL34	HROC107_cT0M2	KM12	SNUC2B	HCA24
CL40	HROC112_MET	KM12C	SNUC5	HCA46
CO115	HROC113	KM12L4	SW1116	HCC2998
COCM1	HROC126	KM12SM	SW1222	HCT116
COGA1	HROC131_T0M3	KM20	SW1417	HDC8
COGA10	HROC173	KP283T	SW1463	HROBMCO1
COGA12	HROC18	KP363T	SW403	HROC24
COGA2	HROC183	KP7038T	SW48	HROC277_T0M1
COGA3	HROC212	LIM1215	SW480	HROC278_MET
COGA5	HROC239_T0M1	LIM1863	SW620	HROC32
COGA5L	HROC24	LIM1899	SW837	HROC334
COGA8	HROC257_T0M1	LIM2099	SW948	HROC39
COLO201	HROC277_MET2	LIM2405	T84	HROC46
COLO205	HROC277_T0M1	LIM2412	V411	HROC57
COLO320	HROC278_MET	LIM2537	V457	HROC59
COLO320DM	HROC278_T0M1	LIM2550	V481	HROC69
COLO320HSR	HROC284_MET	LIM2551	V703	HROC80
COLO60H	HROC285_T2M2	LM0201_A_XL	V9P	HT29
COLO678	HROC300_T2M1	LM0701_A_XL	VACO400	JVE207
COLO94H	HROC313_MET1	LOVO	VACO432	KM12
CR4	HROC32	LS1034	VACO5	KP363T
CRC0078_XL	HROC324	LS123	VACO6	LIM1215
CRC0080_XL	HROC334	LS174T	WIDR	LIM2099
CRC0081_XL	HROC39	LS180		LIM2537
CRC0104_XL	HROC40	LS411N		LIM2551
CRC0174_XL	HROC43	LS513		LS411N
CRC0186_XL	HROC46	MDST8		LS513
CRC0313_XL	HROC50_T1M5	MIP101		MDST8
CRC0322_XL	HROC57	NCIH498		OUMS23
CRC0394_XL	HROC59	NCIH630		OXCO2
CRC0438_XL	HROC60	NCIH684		OXCO3
CRC0558_XL	HROC69	NCIH716		RW7213
CRC0691_XL	HROC80	OUMS23		SKCO1
CRC0740_XL	HROC87	OXCO1		SNU1181
CRC0778_XL	HT115	OXCO2		SNU1411
CRC0781_XL	HT29	OXCO3		SNU254
CRC1360_XL	HT55	RCM1		SNU81
CW2	HUTU80	RKO		SNU977
CX1	IRCC1_XL	RW2982		SW1417
DIFI	IRCC1_XLRES	RW7213		SW1463
DLD1	IRCC10_A_HL	SKCO1		SW403
FET	IRCC3_HL	SNU1040		SW480
FHC	IRCC3_XL	SNU1047		SW837
GEO	IRCC5_A_XL	SNU1181		SW948
GP2D	IRCC5_B_XL	SNU1235		V411

Table3: NGS data availability of WES and WGS for each sample of the preclinical dataset.

Table 4:

Paper	Software	NOTE
PMC3990474.txt		REVIEW
PMC3588146.txt	SigProfiler	
PMC3776390.txt	SigProfiler,SignatureAnalyzer	
PMC3428862.txt		NA
PMC3641671.txt		NA
PMC3891052.txt		NA
PMC3615080.txt		REVIEW
PMC4587544.txt	SigProfiler	
PMC4568299.txt	SigProfiler	
PMC4487409.txt		NA
PMC4930745.txt		REVIEW
PMC4936490.txt	Custom 1	
PMC4783858.txt	SigProfiler	
PMC4137149.txt	SigProfiler	
PMC4636053.txt		NA
PMC5458139.txt		EDITORIAL
PMC5833945.txt	SigProfiler	
PMC5957269.txt	SigProfiler	
PMC5980794.txt	SomaticSignatures	
PMC5562447.txt		NA
PMC5966039.txt	Custom 1	
PMC5854542.txt	SigProfiler	
PMC5886988.txt	SigProfiler,DeconstructSigs	
PMC5516531.txt	Custom 2	
PMC5413372.txt		NA
PMC5292934.txt		NA
PMC5849393.txt	DeconstructSigs	
PMC5972025.txt	DeconstructSigs	
PMC5744871.txt	SomaticSignatures	
PMC5905700.txt	SigProfiler	
PMC5581702.txt		NA
PMC5512575.txt		EDITORIAL
PMC5461196.txt	SigProfiler	
PMC5768571.txt	SigProfiler	
PMC5511565.txt	DeconstructSigs	
PMC5699513.txt		REVIEW
PMC5021589.txt		NA
PMC5164915.txt		NA
PMC6917478.txt		NA
PMC6726428.txt	DeconstructSigs	
PMC6038908.txt	MutationalPatterns	
PMC6141049.txt	SigProfiler	
PMC6044419.txt		REVIEW
PMC6837891.txt	hdp	
PMC6637375.txt	SigProfiler	
PMC6900933.txt		NA
PMC6887544.txt	SigProfiler,SignatureAnalyzer	
PMC6905203.txt		NA
PMC6878125.txt	TrackSig	
PMC6858873.txt		NA
PMC6774882.txt	SigProfiler	
PMC6731024.txt	SigProfiler	
PMC6816465.txt		REVIEW
PMC6887557.txt		NA
PMC6876854.txt		REVIEW
PMC6561810.txt	SomaticSignatures	
PMC6591080.txt		NA
PMC6499643.txt		NA
PMC6310222.txt		NA
PMC6726436.txt	DeconstructSigs	
PMC6119118.txt	Custom	
PMC6193541.txt	SigProfiler	
PMC6613387.txt	SignatureAnalyzer	
PMC6168352.txt	Custom	

Paper	Software	NOTE
PMC6192263.txt	SignatureAnalyzer	
PMC6084431.txt	SomaticSignatures	
PMC6261707.txt		NA
PMC6372067.txt		NA
PMC6501568.txt		NA
PMC6681830.txt		PERSPECTIVE
PMC6169740.txt	SigProfiler	
PMC6810600.txt	SigProfiler	
PMC7334101.txt	DeepMS	
PMC7611134.txt	SigProfiler,MutationalPatterns,DeconstructSigs	
PMC7611045.txt	Custom	
PMC7704768.txt	SigProfiler	
PMC7048622.txt	HRdetect	
PMC7610456.txt	SigProfiler	
PMC7518240.txt	DeconstructSigs	
PMC7260192.txt	HRdetect	
PMC7367727.txt	DeconstructSigs	
PMC7501190.txt	SigProfiler	
PMC7415493.txt		NA
PMC7061455.txt	DeconstructSigs	
PMC8639789.txt	musicatk	
PMC8260632.txt	DeconstructSigs	
PMC8254772.txt	SignatureEstimation	
PMC8387086.txt	DeconstructSigs	
PMC8896908.txt	DeconstructSigs	
PMC8026670.txt	mmsig	
PMC8758513.txt	Custom	
PMC8678141.txt	DeconstructSigs	
PMC8240481.txt	DeconstructSigs	
PMC8102307.txt	SigProfiler	
PMC8783930.txt	SigProfiler	
PMC8102372.txt	SigProfiler	
PMC8230734.txt	DeconstructSigs	
PMC8058239.txt		NA
PMC8166422.txt	DeconstructSigs	
PMC9070557.txt	SigProfiler	
PMC9082009.txt		NA
PMC10011881.txt		de novo
PMC10181095.txt		de novo
PMC10239365.txt	SignatureEstimation	
PMC10240575.txt	SigProfiler	
PMC10291629.txt		NA
PMC10530398.txt		NA
PMC10643023.txt		NA
PMC7613262.txt		de novo
PMC7613712.txt		de novo
PMC7614988.txt	DeconstructSigs	
PMC8516064.txt		NA
PMC8627473.txt		NA
PMC8752466.txt		NA
PMC8826492.txt		NA
PMC8957077.txt		NA
PMC9018481.txt	SigMA	
PMC9199381.txt	SigMA	
PMC9283007.txt	DeconstructSigs	
PMC9356994.txt	DeconstructSigs	
PMC9357465.txt	MutationalPatterns	
PMC9432807.txt	MutationalPatterns	
PMC9433963.txt	HRdetect	
PMC9478565.txt		NA
PMC9585706.txt		NA
PMC9589925.txt		NA
PMC9627127.txt	DeconstructSigs	
PMC9700387.txt		NA

Table4: Results of the literature systematic review: 'REVIEW/PERSPECTIVE/EDITORIAL' were excluded, only research articles were considered in the analysis; 'NA' indicates that the tool for performing mutational signature analysis was not specified in the paper; 'de novo' indicates that the authors performed 'de novo' mutational signature extraction while we only considered mutational signature fitting.

References

1. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.* Nature, 2020. **578**(7793): p. 94-101.
2. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-93.
3. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
4. Petljak, M., et al., *Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis.* Cell, 2019. **176**(6): p. 1282-1294 e20.
5. Chen, J.M., C. Ferec, and D.N. Cooper, *Patterns and mutational signatures of tandem base substitutions causing human inherited disease.* Hum Mutat, 2013. **34**(8): p. 1119-30.
6. Steele, C.D., et al., *Signatures of copy number alterations in human cancer.* Nature, 2022. **606**(7916): p. 984-991.
7. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.
8. Alexandrov, L.B., et al., *Deciphering signatures of mutational processes operative in human cancer.* Cell Rep, 2013. **3**(1): p. 246-59.
9. Koh, G., et al., *Mutational signatures: emerging concepts, caveats and clinical applications.* Nat Rev Cancer, 2021. **21**(10): p. 619-637.
10. Moore, L., et al., *The mutational landscape of human somatic and germline cells.* Nature, 2021. **597**(7876): p. 381-386.
11. Alexandrov, L.B., et al., *Clock-like mutational processes in human somatic cells.* Nat Genet, 2015. **47**(12): p. 1402-7.
12. Cagan, A., et al., *Somatic mutation rates scale with lifespan across mammals.* Nature, 2022. **604**(7906): p. 517-524.
13. Rahbari, R., et al., *Timing, rates and spectra of human germline mutation.* Nat Genet, 2016. **48**(2): p. 126-133.
14. Palles, C., et al., *Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas.* Nat Genet, 2013. **45**(2): p. 136-44.
15. Hayward, N.K., et al., *Whole-genome landscapes of major melanoma subtypes.* Nature, 2017. **545**(7653): p. 175-180.
16. Saini, N., et al., *The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts.* PLoS Genet, 2016. **12**(10): p. e1006385.
17. Pleguezuelos-Manzano, C., et al., *Mutational signature in colorectal cancer caused by genotoxic pks(+) E. coli.* Nature, 2020. **580**(7802): p. 269-273.
18. Boot, A., et al., *Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types.* Genome Res, 2020. **30**(6): p. 803-813.
19. Lee-Six, H., et al., *The landscape of somatic mutation in normal colorectal epithelial cells.* Nature, 2019. **574**(7779): p. 532-537.

20. Crisafulli, G., *Mutational Signatures in Colorectal Cancer: Translational Insights, Clinical Applications, and Limitations.* Cancers (Basel), 2024. **16**(17).
21. Jiricny, J., *The multifaceted mismatch-repair system.* Nat Rev Mol Cell Biol, 2006. **7**(5): p. 335-46.
22. Li, G.M., *Mechanisms and functions of DNA mismatch repair.* Cell Res, 2008. **18**(1): p. 85-98.
23. Sinicrope, F.A. and D.J. Sargent, *Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications.* Clin Cancer Res, 2012. **18**(6): p. 1506-12.
24. Rousseau, B., et al., *The Spectrum of Benefit from Checkpoint Blockade in Hypermutated Tumors.* N Engl J Med, 2021. **384**(12): p. 1168-1170.
25. Salipante, S.J., et al., *Microsatellite instability detection by next generation sequencing.* Clin Chem, 2014. **60**(9): p. 1192-9.
26. Niu, B., et al., *MSIsensor: microsatellite instability detection using paired tumor-normal sequence data.* Bioinformatics, 2014. **30**(7): p. 1015-6.
27. Huang, M.N., et al., *MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations.* Sci Rep, 2015. **5**: p. 13321.
28. Zou, X., et al., *A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage.* Nat Cancer, 2021. **2**(6): p. 643-657.
29. Couch, F.J., K.L. Nathanson, and K. Offit, *Two decades after BRCA: setting paradigms in personalized cancer care and prevention.* Science, 2014. **343**(6178): p. 1466-70.
30. Davies, H., et al., *HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures.* Nat Med, 2017. **23**(4): p. 517-525.
31. Chopra, N., et al., *Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer.* Nat Commun, 2020. **11**(1): p. 2662.
32. Crisafulli, G., et al., *Temozolomide Treatment Alters Mismatch Repair and Boosts Mutational Burden in Tumor and Blood of Colorectal Cancer Patients.* Cancer Discov, 2022. **12**(7): p. 1656-1675.
33. Omichessan, H., G. Severi, and V. Perduca, *Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance.* PLoS One, 2019. **14**(9): p. e0221235.
34. Maura, F., et al., *A practical guide for mutational signature analysis in hematological malignancies.* Nat Commun, 2019. **10**(1): p. 2969.
35. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
36. Andrei, P., et al., *Integrated approaches for precision oncology in colorectal cancer: The more you know, the better.* Semin Cancer Biol, 2022. **84**: p. 199-213.
37. Andre, T., et al., *Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer.* N Engl J Med, 2020. **383**(23): p. 2207-2218.
38. Diaz, L.A., Jr., et al., *Pembrolizumab versus chemotherapy for microsatellite instability-high or mismatch repair-deficient metastatic colorectal cancer (KEYNOTE-177): final analysis of a randomised, open-label, phase 3 study.* Lancet Oncol, 2022. **23**(5): p. 659-670.

39. Lenz, H.J., et al., *First-Line Nivolumab Plus Low-Dose Ipilimumab for Microsatellite Instability-High/Mismatch Repair-Deficient Metastatic Colorectal Cancer: The Phase II CheckMate 142 Study.* J Clin Oncol, 2022. **40**(2): p. 161-170.
40. Cornish, A.J., et al., *The genomic landscape of 2,023 colorectal cancers.* Nature, 2024. **633**(8028): p. 127-136.
41. Medico, E., et al., *The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets.* Nat Commun, 2015. **6**: p. 7002.
42. Lazzari, L., et al., *Patient-Derived Xenografts and Matched Cell Lines Identify Pharmacogenomic Vulnerabilities in Colorectal Cancer.* Clin Cancer Res, 2019. **25**(20): p. 6243-6259.
43. Durinikova, E., et al., *Targeting the DNA Damage Response Pathways and Replication Stress in Colorectal Cancer.* Clin Cancer Res, 2022. **28**(17): p. 3874-3889.
44. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
45. Blokzijl, F., et al., *MutationalPatterns: comprehensive genome-wide analysis of mutational processes.* Genome Med, 2018. **10**(1): p. 33.
46. Rosenthal, R., et al., *DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution.* Genome Biol, 2016. **17**: p. 31.
47. Degasperi, A., et al., *A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies.* Nat Cancer, 2020. **1**(2): p. 249-263.
48. Diaz-Gay, M., et al., *Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment.* Bioinformatics, 2023. **39**(12).
49. Islam, S.M.A., et al., *Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor.* Cell Genom, 2022. **2**(11): p. None.
50. Kim, Y.S., M. Lee, and Y.J. Chung, *Two subtypes of cutaneous melanoma with distinct mutational signatures and clinico-genomic characteristics.* Front Genet, 2022. **13**: p. 987205.
51. Rospo, G., et al., *Evolving neoantigen profiles in colorectal cancers with DNA repair defects.* Genome Med, 2019. **11**(1): p. 42.
52. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
53. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
54. Crisafulli, G., et al., *Whole exome sequencing analysis of urine trans-renal tumour DNA in metastatic colorectal cancer patients.* ESMO Open, 2019. **4**(6).
55. Corti, G., et al., *A Genomic Analysis Workflow for Colorectal Cancer Precision Oncology.* Clin Colorectal Cancer, 2019. **18**(2): p. 91-101 e3.

Collaborations and side projects

Over my PhD, working as a bioinformatician in a large multidisciplinary group with time established external collaboration, I had the opportunity to being involved in a variety of side projects, some of which concluded with publications while other still ongoing.

In this section, I report a short description of the collaborations occurred over my PhD:

- **Temozolomide priming project**: this represented the starting point of my work in the mutational signature field. The project was leaded by Giovanni Crisafulli and published in 2022 in Cancer Discovery (Crisafulli G, Sartore-Bianchi A, Lazzari L, et al. Temozolomide Treatment Alters Mismatch Repair and Boosts Mutational Burden in Tumor and Blood of Colorectal Cancer Patients. Cancer Discov. 2022;12(7):1656-1675. doi:10.1158/2159-8290.CD-21-1434). Being the first time that the lab was facing this type of analysis, I did a large benchmarking work of several tools and various cancer model focused on specific mutational signature detection.
- **Chemotherapy priming project:** this project, leaded by Pietro Paolo Vitiello, is focused of monitoring the immune system activation and involvement after a priming phase with several chemotherapeutics agents. The manuscript containing the results of this project, is currently under second revision in Cancer Cell, allowed me to strengthen my skills in performing genetic analysis on murine preclinical models. Specifically, I was mainly involved in performing mutational calling, tumor mutational burden assessment, mutational signatures fitting, after the priming phase with several agents. Then, followed an *in-vivo* analysis, aiming to characterized the dynamics of the acquired mutations in immune-competent and in immune-deficient murine models.
- **Cell free DNA release project:** in this project, leaded by Valeria Pessei and Marco Macagno, published in 2024 in Genome Medicine (Pessei V, Macagno M, et al. DNA demethylation triggers cell free DNA (cfDNA) release in colorectal cancer cells. Genome Med. 2024;16(1):118. Published 2024 Oct 9. doi:10.1186/s13073-024-01386-5) we focused on studying different aspects of DNA release. Specifically, I was mainly involved for performing the comparison between

several genetic features of DNA extracted from the nuclei and DNA from the supernatant.

In addition to these projects, I was involved in several minor project, involving also external collaborations, for which I participated mainly as a bioinformatician performing genetic analysis.

Acknowledgments

I would like to express my sincere thanks to some people who have been very important during my PhD.

First of all, I would like to thank Alberto Bardelli who, as my PI, has given me the opportunity to work in an incredible group for the last five years, first as an undergraduate and then as a Ph.D. student. I would also like to thank Federica Di Nicolantonio as my external PhD supervisor for the valuable scientific feedbacks. Furthermore, I would like to thank the whole Bradelli's group and especially the bioinformaticians, starting with Giovanni Crisafulli, my supervisor since my first steps in bioinformatics and genetics. A big thank you to him for believing in me and in this thesis until its publication. A big thank you to Vittorio, Giorgio, Elisa and Gaia for sharing these years with me, you have taught me a lot.

Last but not least, a big thank you to Eleonora, Vito and Pietro Paolo for sharing this journey with me, for discussing scientific and everyday issues, for entertaining me outside of work and for helping to create wonderful memories! Thank you all.