



**UNIVERSITA
DI TORINO**

Department of Oncology

PhD in Complex Systems for Quantitative Biomedicine

XXXVII cycle

**Transcriptome-wide gene expression outlier
analysis pinpoints therapeutic vulnerabilities
in colorectal cancer**

Supervisors:

Professor Enzo Medico

Professor Alberto Bardelli

Candidate:

Gaia Grasso

Summary

	Abstract	2
1	Introduction	3
1.1	Precision medicine in cancer	3
1.2	Gene expression outliers	5
1.3	Colorectal Cancer	8
2	Results	13
2.1	Multi-omics profiling of 226 CRC cell lines	13
2.2	Landscape of gene expression outliers in CRC cell lines	14
2.3	Method validation based on the identification of gene expression abnormalities associated with clinically relevant features of CRC cell lines	16
2.4	Aberrant DNA methylation in promoter regions underlies outlier expression values that are recurrently found in CRC cell lines	19
2.5	Outlier expression values are associated with genomic rearrangements	21
2.6	Analysis of overexpressed enzyme genes pinpoints therapeutic targets	22
2.7	Characterization of underexpressed genes unveils synthetic lethal vulnerabilities in CRC cells	23
2.8	Aberrantly expressed genes in both CRC cell lines and tissue samples obtained from colorectal tumours	25
3	Discussion	59
4	Materials and Methods	67
	Acknowledgements	79
	References	80
	Tables	87

Abstract

Multiple strategies are continuously being explored to expand the drug target repertoire in solid tumours.

We devised a novel computational workflow for transcriptome-wide gene expression outlier analysis that allows the systematic identification of both overexpression and underexpression events in cancer cells. Here it was applied to expression values obtained through RNA sequencing in 226 colorectal cancer (CRC) cell lines that were also characterized by whole exome sequencing and microarray-based DNA methylation profiling.

We found cell models displaying an abnormally high or low expression level for 3,533 and 965 genes, respectively.

Gene expression abnormalities that have been previously associated with clinically relevant features of CRC cell lines were confirmed.

Moreover, by integrating multi-omics data, we identified both genetic and epigenetic alterations underlying outlier expression values. Importantly, our atlas of CRC gene expression outliers can guide the discovery of novel drug targets and biomarkers.

As a proof of concept, we found that CRC cell lines lacking expression of the MTAP gene are sensitive to treatment with a PRMT5-MTA inhibitor (MRTX1719). Finally, other tumour types may also benefit from this approach.

1. Introduction

1.1 Precision medicine in oncology

Over the past two decades, the advent of precision medicine has brought about a significant revolution in the field of modern healthcare¹. The theoretical concept has emerged as a shift from the traditional approaches to medical treatment, which were designed to fit all patients, to a more tailored approach that considers the individual characteristics of each patient¹. This incorporates genetic, environmental, and lifestyle factors to develop personalised therapeutic strategies, that, for each individual patient, aims to enhance efficacy while minimizing adverse effects¹. The evolution of precision medicine has been significantly driven by the advent of next-generation sequencing (NGS) technologies and by the substantial advances in computational approaches for the analysis of the large data sets generated by these technologies. These developments have facilitated a notable expansion and deepening of our knowledge of the human genome¹.

The concept of precision medicine is currently applied in a range of medical disciplines and for diverse purposes. Firstly, it enables precise molecular diagnoses, for example of rare genetic disorders, or in the management of infectious diseases, where it has revolutionized pathogen identification. Secondly, it underlies pharmacogenomics studies which seek to identify biomarkers of response to different drugs. More recently, polygenic risk scores have been used to stratify patients, for example in cardiology, where the effects of multiple genetic variants are combined in order to predict an individual's risk of developing coronary artery disease¹.

The inherently heterogeneous nature of cancer has long been a challenge for the 'one-size-fits-all' treatment approach^{2,3}. Traditional chemotherapy, administered without patient stratification, has a limited efficacy and often elicit significant side effects, due to its non-specific nature, affecting cells that are both healthy and neoplastic^{2,3}. In this context, oncology is one of the medical fields that has been most revolutionised by precision medicine.

Firstly, the identification and characterization of somatic mutations has promoted the design of drugs that specifically inhibit mutated proteins within oncogenic signalling pathways that drive tumour progression^{4,5}. For example, the discovery of the *BRAF* V600E mutation has resulted in the development of vemurafenib, an inhibitor that has markedly enhanced the survival rate of patients with melanoma^{4,5}. Similarly, the identification of mutations in the *EGFR* gene has led to the development of targeted therapies against the hyperactive form of the receptor, thus allowing a significant improvement in the treatment of lung cancer^{4,5}.

Secondly, the analysis of somatic mutations has been crucial for the identification of biomarkers in signalling pathways downstream of targeted proteins, like the discovery that activating mutations in RAS/RAF proteins render the targeting of the upstream EGFR ineffective^{6,7}.

In addition to mutations, other molecular alterations have proven useful in identifying clinically relevant features of human tumours and cancer cell-specific vulnerabilities^{2,3}. A paradigmatic example is *ERBB2* gene amplification which occurs at variable frequency across multiple tumour types, such as breast and colorectal cancer (CRC), where it leads to sensitivity to HER2 inhibition^{8,9}. In a similar way, TRK inhibitors have been approved for use across a broad range of solid tumours that harbour genetic rearrangements involving *NTRK* genes¹⁰. Importantly, in both cases, a genetic alteration leads to gene overexpression that results in aberrant protein activity.

On the other hand, during tumour evolution, the expression of multiple genes is lost due to homozygous deletion or epigenetic silencing, leading to the emergence of clinically significant features^{11,12}. For instance, in sporadic CRC, the silencing of the *MLH1* gene, as a result of promoter hypermethylation, determines loss of DNA mismatch repair function and induces microsatellite instability¹³.

Similarly, in glioblastoma, the hypermethylation of the *MGMT* gene promoter leads to its loss of expression, thereby making tumours more sensitive to temozolomide (TMZ), a potent alkylating agent that is an essential component of the current standard protocol for glioblastoma treatment¹⁴.

Intriguingly, gene silencing can also be associated with synthetic lethal targets. This phenomenon occurs when the simultaneous loss of function of two distinct genes or pathways results in cell death, whereas the loss of either gene or pathway individually

does not affect cell survival¹⁵. For example, sensitivity to PARP inhibitors is markedly increased in cancer cells that lack RAD51C, which plays a crucial role in DNA repair. The RAD51C deficiency impairs homologous recombination (HR), which is one of the two DNA repair mechanisms that are involved in the repair of double strand breaks (DSBs). On the other hand, PARP inhibitors disrupt the alternative repair pathway, non-homologous end joining (NHEJ): this combined impairment leads to excessive DNA damage and cell death¹⁵.

In addition to the vulnerabilities associated with the overexpression or silencing of individual genes, the concept of transcriptional addiction has recently emerged as a critical framework for understanding cancer cell dependencies. During tumorigenesis, cancer cells accumulate several mutations, some of which will inevitably affect gene expression. The principle of transcriptional addiction describes how cancer cells develop a complex reliance on these dysregulated transcriptional programs, which means that they become dependent on specific transcription factors and co-factors that are upregulated or constitutively active and drive the expression of genes involved in cell growth, survival, and metastasis. Additionally, this implies that the aforementioned alterations at the transcriptional level, and thus any resulting vulnerabilities, cannot be directly predicted by genetic changes¹⁶.

Overall, these examples highlight that methods for systematically identifying aberrant expression levels in cancer cells may be helpful in rapidly finding genes that deserve an in-depth characterization, with the ultimate goal of discovering potential drug targets and biomarkers. In particular, given the observation that multiple molecular alterations result in significant changes in gene expression, the analysis of tumours from a transcriptional perspective may prove a more convenient and efficient approach than performing independent analyses of different omics data.

1.2 Gene expression outliers

In the context of statistical analysis, observations that fail to align with a global pattern are usually referred to as “outliers”, often manifesting as unusually small or large values. They can arise from measurement errors, execution errors, or the inherent

variability within a population. Although outliers are often regarded as noise or errors that could distort the results, they may also contain crucial information about unique characteristics of the data¹⁷.

In gene expression studies, outliers are samples that exhibit an abnormally high or low expression level for a specific gene, significantly diverging from the rest of the dataset. In most analytical contexts, the existence of outlier samples is simply accounted for to avoid misleading conclusions. However, it is important to remember that the abnormal expression of a gene may highlight unique biological phenomena^{17,18}.

For example, the study of gene expression outliers has been exploited to understand the genetic basis of diseases. In particular, Xin Li and colleagues reported that extreme gene expression levels are frequently driven by rare variants located in regulatory regions of the genome, including enhancers, promoters, and untranslated regions. These variants can disrupt normal regulatory mechanisms, leading to significant deviations in gene expression that may contribute to disease susceptibility and phenotypic diversity across different tissues. The study demonstrates the value of analysing gene expression outliers in genetic association studies to predict the effects of rare variants, where the very low number of patients makes it unfeasible to employ standard modelling methods¹⁸.

In addition, in a pioneering study, the outlier analysis of pan-cancer gene expression data has been exploited to identify genes that, when altered, play a role in cancer development¹⁹. In particular, this analysis allowed to reveal in prostate cancer the recurrent over-expression of two genes in the ETS family of transcription factors, ERG and ETV1, due to recurrent fusions involving these genes and the androgen-responsive promoter of TMPRSS2¹⁹.

Several studies have also exploited the identification of gene expression outliers to establish which tumour samples or preclinical models were characterized by overexpression of druggable genes^{20–22}. For example, the search for kinase expression outliers among cancer preclinical models led to the identification of cell lines that were functionally and pharmacologically addicted to overexpressed kinases^{20,23}.

While the objectives of studies focusing on gene expression outliers are sometimes conceptually similar, a variety of techniques are used to define them. Some methods use sophisticated statistical techniques that assess the distribution and spread of multi-dimensional data, aiming to detect values that deviate markedly from expected patterns¹⁷. Other, more simple approaches use unidimensional data to take into account the deviation of each single sample from central measures, such as the mean or median expression, often using arbitrary thresholds¹⁷. For instance, the use of the z-score enables the identification of outliers, which are defined as samples situated at the periphery of a normal distribution curve. In the case of data that do not follow a normal distribution, alternative methods apply rules based on data percentiles or interquartile ranges to identify extreme values, like the popular Tukey's rule. In this case, outlier calling is based on gene-specific thresholds since, given the characteristic distribution of expression values of each single gene, outliers are found as values more than 1.5 interquartile ranges (IQR) above the third quartile (Q3) or below the first quartile (Q1)¹⁷.

A common feature shared by these methods is that the identification of abnormally expressed genes is based on the comparison of samples within a singular dataset, which must have a considerable size to yield reliable results. Alternatively, each individual sample can be compared with a reference dataset. For instance, Vivian *et al.* proposed a method that employs gene expression data from normal tissues to generate a consensus background distribution for each individual gene, which is then utilized to identify gene overexpression in individual cancer samples²⁴.

Differently, Kothari *et al.*, using transcriptomic data from both cancer and normal samples of 25 different tissues, focused on the comparison of the expression levels of multiple kinase genes in each individual sample, *i.e.*, they looked for “outlier genes in individual samples” rather than “samples displaying outlier expression of candidate genes”²⁵. Importantly, they found that cancer cell lines are sensitive to the inhibition of outlier kinases. Therefore, despite methodological differences, their conclusion was in line with that of the studies that looked for outlier samples.

We reasoned that, although successful, the scope of previous studies was limited to gene overexpression and often restricted to specific gene subsets, usually kinases. Kinases have been a focal point because they play critical roles in regulating various

cellular processes, including cell signalling, growth, and differentiation, which are often dysregulated in cancer. Moreover, their structure makes them ideally suited for targeted therapies, as overexpressed kinases can be inhibited by either antibodies or small molecules²⁶.

However, this narrow focus may overlook other significant factors: we hypothesized that additional vulnerabilities could be uncovered by extending the gene expression outlier analysis to the entire coding genome and by concomitantly profiling gene underexpression and overexpression events.

From now on, in this work we will refer to samples showing overexpression of an individual gene as “positive outliers”, while “negative outliers” are samples characterized by the underexpression of a single gene.

1.3 Colorectal cancer

In the present study, focused on CRC, which represents the third most frequent tumour type worldwide and the second leading cause of cancer-related deaths. It accounts for approximately 10% of all cancer diagnoses and cancer-related deaths worldwide, and was associated with 1.8 million cases and 900,000 deaths in 2020²⁷.

Despite the introduction of population screening programs, that has led to a decline in CRC-associated deaths by enabling earlier diagnosis and surgical intervention, five-year overall survival rates following surgical resection are 99% for stage I, 68%-83% for stage II, and 45%-65% for stage III CRC. However, 15%-30% of patients present with metastatic CRC (stage IV), and 20%-50% of those with initially localized tumours eventually develop metastases, commonly to the liver, lung, or peritoneum. At this metastatic stage, the 5-year overall survival rate remains around 15%, highlighting a significant clinical challenge²⁸.

One of the reasons for such a low survival rate in patients is the significant inter-patient heterogeneity from a molecular point of view. These solid tumours develop in the colon or rectum from pre-malignant adenomatous lesions through a multistep tumorigenesis process involving both genetic and epigenetic changes²⁹. In particular, there are three primary evolutionary pathways for CRC development. The traditional

adenoma-adenocarcinoma sequence, which accounts for approximately 60-70% of cases, begins with genetic loss of *APC*, followed by mutations in genes such as *DCC*, *RAS*, *SMADs*, and *TP53*, and is characterized by Chromosomal Instability (CIN) with chromosomal aberrations. The serrated adenoma pathway, responsible for 10-20% of cases, features a CpG Island Methylator Phenotype (CIMP), *BRAF* mutations, and often Microsatellite Instability (MSI) due to epigenetic silencing of the *MLH1* gene. The third pathway, observed in about 5% of cases, involves microsatellite instability resulting from inherited mutations in mismatch repair (MMR) genes like *MSH2*, *MLH1*, *PMS1*, and *PMS2*, followed by a second somatic event in the same gene²⁹.

Additionally, rare genetic alterations such as *HER2* amplifications, *ROS1* and *ALK* gene fusions, and *NTRK* rearrangements can be found in patients, regardless of the evolutionary pathway followed by the tumour^{30,31}.

This highlights the necessity for identifying biomarkers to stratify patients into distinct prognostic or treatment categories, tailored to the specific biology of each tumour^{29,32}.

In clinical practice, the primary curative treatment for localized, non-metastatic CRC is surgery^{29,32}. Adjuvant chemotherapy is typically administered to high-risk stage II and stage III patients to reduce the risk of relapse^{29,32}. Metastatic CRC, however, is not effectively treated with current therapeutic options, despite the introduction of fluoropyrimidine-based regimens such as FOLFOX (5-FU, leucovorin, oxaliplatin) and FOLFIRI (5-FU, leucovorin, irinotecan)^{29,32}.

The addition of molecularly targeted agents into the standard of care of metastatic CRC has enhanced treatment efficacy when combined with or administered after traditional chemotherapy and has paved the way for a first patient molecular stratification. First-line therapy often involves anti-VEGF agents (*e.g.*, bevacizumab) for tumours exhibiting angiogenesis-related markers or anti-EGFR monoclonal antibodies (*e.g.*, cetuximab, panitumumab) for *RAS/RAF* wild-type tumours. A recent therapeutic advancement for patients harbouring *BRAF* mutations involves the incorporation of a *BRAF* inhibitor (encorafenib) in combination with anti-EGFR therapy as a second-line treatment option^{28,29,32}.

Nevertheless, there are additional targets for which inhibitors have been approved. Recently, a number of trials have tested different therapeutic regimens targeting *HER2* in *ERBB2*-amplified, *KRAS* wild-type, mCRC patients. These include dual *HER2* blockade

with the monoclonal antibody trastuzumab and the small kinase inhibitor lapatinib, combination of two anti-HER2 monoclonal antibodies (trastuzumab and pertuzumab), and treatment with trastuzumab conjugated with a topoisomerase 1 inhibitor (deruxtecan)³²⁻³⁴. Moreover, despite being less than 1%, patients with NTRK fusion positive CRC can benefit from the treatment with two first generation NTRK inhibitors (larotrectinib and entrectinib) that have been approved in a tumour-agnostic manner³⁵.

Finally, the introduction of KRAS G12C inhibitors, (*e.g.*, adagrasib, initially approved for NSCLC patients and only recently for mCRC ones, in combination with cetuximab), represents a significant advancement in the treatment of CRC. Although only 3% of CRC patients are diagnosed with the *KRAS* G12C mutations, inhibitors targeting this oncogene pave the way for a new therapeutic option for a significant proportion of patients (40-50%) in whom the presence of these previously untreatable mutations has precluded access to other therapies³⁶.

Despite the existence of these numerous approved targeted therapies for specific patient subsets, several issues still need to be addressed. Firstly, 40 to 50% of patients present with tumours that lack any targetable alteration³². Secondly, there is currently no biomarker available to identify patients who will benefit from chemotherapy, which still represents the most common treatment option for CRC patients³². Moreover, the efficacy of targeted therapies is not only limited to specific subsets of patients, but many of these ultimately develop resistance, resulting in treatment failure³².

In the last few years, new insights into cancer and the immune system have led to the development of immune checkpoint inhibitors, that have transformed the landscape of therapeutic options for numerous types of cancer. However, colorectal cancer is one of the tumours in which immunotherapy has been demonstrated to be less effective, as only a restricted subset of patients, characterised by MSI, display remarkable response to PD1-blocking antibodies (pembrolizumab and nivolumab), that are for this reason approved in this setting. On the other hand, the vast majority of patients with MSS tumours do not benefit from immunotherapy^{28,37}.

Therefore, there is an urgent need to identify novel biomarkers and therapeutic targets in CRC in order to best control tumour growth in an ever-increasing proportion of

patients, overcome resistance, and improve long-term outcomes for a broader population of patients^{29,32}.

In the last decade, studies on CRC preclinical models, recapitulating the genomic, phenotypic and drug response diversity that is observed in patients, have stimulated the design of successful clinical trials testing inhibitors of oncogenic drivers found in small but clinically relevant fractions of CRC patients, including those mentioned above^{29,32}. Further exploration of existing and novel preclinical models may reveal additional therapeutic hits.

Moreover, recent research has increasingly employed gene expression profiles to more accurately predict outcomes and treatment responses, using various analytical approaches^{29,32}. Recent developments include various gene expression assays designed to predict recurrence risk in early-stage CRC patients. These assays, such as Oncotype DX®, ColoPrint®, OncoDefender-CRC™, and GeneFx™, use different methodologies to assess risk and guide treatment decisions²⁹.

The Consensus Molecular Subtypes (CMS) classification provides a more detailed stratification of patients by identifying four subtypes with distinct biological and clinical characteristics: CMS1 (immune), CMS2 (canonical), CMS3 (metabolic), and CMS4 (mesenchymal). Of note, tumours classified as CMS4, characterized by a mesenchymal phenotype and extensive stromal infiltration, tend to demonstrate inferior outcomes in comparison to other subtypes. Although the CMS classification has shown considerable promise in predicting treatment response and survival, particularly in mCRC, its clinical utility requires further validation³⁸.

The Colorectal Cancer Intrinsic Subtypes (CRIS) classification provides an additional layer of granularity to the CMS framework by identifying five distinct molecular subtypes within CRC, each associated with specific molecular features and therapeutic vulnerabilities. Notably, CRIS classification emphasizes the role of tumour cell-intrinsic properties, as opposed to the more microenvironment-focused CMS, offering insights into subtype-specific drug sensitivities^{38,39}.

Moreover, recent studies have also explored the use of gene expression data in predicting drug responses. Several studies, for example, employed different models to select features based on the gene expression profile of the NCI-60 panel of tumour cell

lines, or, in combination with drug sensitivity data, the Cancer Cell Line Encyclopedia (CCLE) database⁴⁰. Other approaches instead explored the differentially expressed genes between sensitive and resistant cell lines⁴⁰.

Furthermore, patient-derived xenograft (PDX) expression data have been employed to develop a drug response prediction model (called PDXGEM) based on concordant co-expression patterns between the PDXs and the primary tumours in patients⁴¹. Evan *et al.* instead used only expression data from patients' primary tumour tissues (from TCGA) to predict their responses to the most annotated chemotherapeutics in that dataset (5-Fluorouracil and Gemcitabine): the machine learning model presented in this study demonstrated an accuracy of up to 86% in predicting responses⁴².

An alternative strategy, that couples the recent interest in gene expression profiles, with the growing number of available CRC preclinical models, has led to the exploitation of gene expression outlier identification in different studies, aimed at establishing which models were characterised by overexpression of druggable genes^{20,22}.

For example, a previous study by our group investigated the gene expression profiles of 151 CRC cell lines to identify gene expression outliers, aiming to uncover new dependencies of these cell lines that might not be apparent through other analyses. The study focused on tyrosine kinase genes for which targeted agents had already been approved or were undergoing clinical trials. The results revealed that a number of them, including ALK, FGFR2, NTRK1/2, and RET, were significantly overexpressed in a subset of RAS/BRAF wild-type cell lines, that were resistant to EGFR inhibitor therapies. This identified a crucial role for these kinases in driving resistance, as well as a potential critical vulnerability. This hypothesis was tested with a drug sensitivity assay, whose results demonstrated that cell lines that were outliers for a kinase, were particularly sensitive to the corresponding inhibitor, proving that they were functionally and pharmacologically addicted to overexpressed kinases²⁰.

Over the last years, we have significantly expanded our collection of CRC cell lines for which multi-omics data, including RNA sequencing (RNA-seq) data, has been collected. In this study, this resource has been exploited to develop a novel computational workflow for transcriptome-wide gene expression outlier analysis, allowing us to

identify CRC cell lines that are positive or negative outliers for any protein coding gene that is expressed in these models. Importantly, this broader analysis enabled us to uncover both known and novel molecular drivers and therapeutic targets. Notably, we validated the functional significance of an underexpressed non-kinase gene, demonstrating its potential as a biomarker for targeted therapy.

2. Results

2.1 Multi-omics profiling of 226 CRC cell lines

In this study we analysed 226 CRC cell lines collected from different sources or established in our laboratory starting from biopsy or surgical samples, as well as from patient-derived xenografts (Fig. 1A and Table 1)^{20,22}. To check the genetic identity of the cell lines and to assess whether they are representative of CRC subtypes with clinical relevance, we performed STR profiling and assessment of the microsatellite status. STR profiling highlighted a small subset of genetically identical cell lines, including models previously reported as being established from the same patient (STR profiles not reported)²⁰. In addition, we found that 148 (65%) models were MSS, while 76 (34%) displayed microsatellite instability (Fig. 1A and Table 1). Two models derived from the same patient (HROC147_T0M1 and HROC147_MET) displayed an intermediate profile and their MSS or MSI status was not uniquely attributed.

Next, data obtained by WES, RNA-seq and DNA methylation microarray experiments were analysed. Genetic alterations associated with CRC onset and progression were found at high frequency, including activating mutations in *KRAS*, *NRAS* and *BRAF* genes, as well as mutations in *APC* and *TP53* genes that likely determine protein loss of function (Fig. 1A and Table 1). Furthermore, a gene expression-based hierarchical clustering confirmed that CRC cell lines that originated from the same patient (as identified by the same STR profile) preferentially clustered together (Fig. 1B)²⁰. We next assessed whether gene expression-based CRC subtypes (CMS and CRIS) found in tumour samples could be detected in preclinical mode^{43–45}. In total, 189 and 178 CRC cell lines were assigned to a specific CMS or CRIS class, respectively (Fig. 1A and Table 1).

We also confirmed some known correspondences between CMS and CRIS classifications (Fig. 2A). For instance, samples classified as CMS2, which are known to be characterised by high WNT and MYC signalling, are also predominantly classified as CRIS-C, a subtype characterised by high WNT activity as well (Fig. 2A). Similarly, the majority of samples classified as CRIS-B, a subtype linked to EMT (epithelial-

mesenchymal transition), are also classified as CMS4, the mesenchymal subtype (Fig. 2A)^{38,39}. These observations validate our analysis.

Finally, we took advantage of DNA methylation microarray data to classify CRC cell lines with respect to CIMP that is a distinguishing feature of the serrated pathway in colorectal tumorigenesis (Fig. 1A, Fig. 2B and Table 1)⁴⁶.

2.2 Landscape of gene expression outliers in CRC cell lines

We have devised a computational workflow to systematically identify overexpression and underexpression events in cancer cells (Fig. 3A). Initially, transcriptome-wide positive and negative gene expression outliers are mapped by applying the Tukey's rule to the expression profile of individual coding genes. We reasoned that this set of gene expression outliers may include functional events, such as alterations associated with oncogenic dependency or cancer cell-specific vulnerabilities, alongside with neutral events that reflect technical or biological noise. Therefore, a multi-filter strategy is implemented to select extreme gene expression outliers, based on the hypothesis that they are more likely to be functionally relevant in cancer cells. First, when multiple samples are scored as positive or negative outliers for an individual gene, only the furthest outliers are kept. In detail, the positive outliers of any gene can be divided in two groups, those that are nearer to the maximum expression value and those that instead are nearer to the highest value that is not an outlier. The former are the furthest outliers of that gene. Similarly, the furthest negative outliers of any gene are those nearer to the minimum expression value than to the lowest value obtained after excluding outliers. Next, additional filters are performed taking into consideration both differential and absolute expression values (see section 4.11).

When this strategy was applied to gene expression data from our collection of 226 CRC cell lines, positive and negative gene expression outliers were identified for 12,744 and 6,928 genes, respectively, out of a total of 16,392 analysed genes (Fig. 3A). This included 5,001 genes for which both positive and negative gene expression outliers were found. After filtering, we retained 3,533 and 965 genes that had extreme positive or negative gene expression outliers, respectively (Fig. 3A and Table 2).

For example, the application of the Tukey's method resulted in the pinpointing of 34 CRC cell lines as positive outliers for the *SERPINA4* gene (Fig. 3B). The number of outliers progressively reduced at each filtering step and in the end 3 CRC cell lines were selected as "extreme positive outliers" for the *SERPINA4* gene (Fig. 3B). Similarly, 2 CRC cell lines were scored as "extreme negative outliers" for the *CNKSR1* gene starting from 10 samples that were considered as negative outliers based on the application of the Tukey's method (Fig. 3C).

Of note, only 138 genes had both extreme positive outliers and extreme negative outliers. We also observed that in most cases only one or two CRC cell lines were scored as extreme outliers for individual genes, but we found two subsets of genes for which extreme positive or negative outliers were identified at a relatively high frequency (Fig4).

Globally, every CRC cell line was scored as extreme positive outlier for at least one gene and at least one underexpressed gene was found in virtually all samples (223 out of 226). The median number of overexpressed genes per cell model was 24, while the median number of underexpressed genes per cell line was 7.

We also characterized each CRC cell line by taking into account the corresponding amount of aberrantly expressed genes to compute positive, negative or total outlier burden values (Fig4).

To strengthen the validity of high outlier burden values in a few samples, we ruled out the possibility that they were due to technical artefacts. In particular, we reasoned that a major confounding factor could have been that the 226 CRC cell lines were sequenced in 14 different batches, but we observed that the samples with the highest outlier burden values are distributed across different sequencing batches, even in the presence of two sequencing batches for which a mild deviation was observed in the Principal Component Analysis (PCA) result (data not shown).

We therefore looked for correlations with known molecular features of CRC cell lines (Fig. 5A-I). While the total outlier burden is not affected by the MSS or MSI status of CRC cell lines (Figure 5A), we observed that it significantly changes between CRC cell lines that belong to different transcriptional subtypes (Figure 5B and 5C).

In particular, we found that the positive outlier burden is significantly higher in CMS4 CRC cell lines than in samples classified in other CMS subtypes (Figure 5D). Consistently

with that, when the CRIS classification is considered, the CRIS-D subtype is the one with the highest positive outlier burden values (Figure 5E).

To dig into the high positive outlier burden of the CMS4 samples, we classified the overexpressed genes found in each CRC cell line based on whether they are exclusively found in that CRC cell line, also found in other CRC cell lines of the same transcriptional subtype or found in several CRC cell lines independently from the transcriptional subtype. We found that CMS4 CRC cell lines are also characterized by a significantly higher fraction of overexpressed genes whose overexpression only occurs in CMS4 samples (Figure 5F).

When we restricted the analysis to extreme negative outliers, we found that the negative outlier burden changes based on the CIMP status of CRC cell lines (Fig. 5G). In particular, it is slightly higher in CIMP-H samples than in samples belonging to other DNA methylation-based classes (Fig. 5H).

2.3 Method validation based on the identification of gene expression abnormalities associated with clinically relevant features of CRC cell lines

To validate our approach, we checked whether transcriptional alterations that are associated with known clinically relevant features of CRC cell lines are recapitulated by our atlas of extreme gene expression outliers.

First, we considered a group of tyrosine kinase (TK) genes whose overexpression in CRC leads to oncogene addiction (*ALK*, *ERBB2*, *FGFR2*, *NTRK1*, *NTRK2* and *RET*)^{23,47–49}. Importantly, the extreme positive outliers identified for these TK genes included cell models that we found to be highly sensitive to the corresponding kinase inhibitors (Fig. 6A)^{23,47–49}. In line with a previous result and consistently with the fact that they are genetically identical cell lines, *NTRK1* overexpression was found in all CRC cell lines of the KM12 family (KM12, KM12C, KM12SM and KM12L4), although the effect of NTKR1 inhibition has been previously tested only in the KM12 CRC cell line (Fig. 6A)⁴⁷.

Similarly, few CRC cell lines in which we have not previously assessed the effect of kinase inhibitors were pinpointed as extreme positive outliers for the *NTRK2* and *RET* genes (Fig. 6A).

Moreover, 3 extreme positive outliers for the gene *ERBB2* were identified: CRC0080_XL, LM701_A_XL, and IRCC72A_XL (Fig. 6A). Two of the models, CRC0080_XL and LM701_A_XL, were previously found to be sensitive to HER2 inhibition with Lapatinib plus Trastuzumab^{23,49}. More in detail, our CRC cell line collection includes two models, LM0201_A_XL and LM0701_A_XL, that were derived from the same patient but from different liver metastases. The first one was derived from a metastatic lesion that exhibited no change in *ERBB2* copy number and did not respond to anti-HER2 treatment⁴⁹. The LM0701_A_XL model was derived from a metastatic lesion in which the genetic amplification of the *ERBB2* gene was identified and which responded to Lapatinib plus Trastuzumab, a previous finding that aligns with our result⁴⁹.

Since the third model, IRCC72A_XL, had never been screened before, we decided to test its responsiveness to HER2 inhibition (Fig. 6B-C). We demonstrated its sensitivity to Lapatinib and Trastuzumab, alone or in combination (Fig. 6B-C). The identification of a *RAS/RAF* WT cell line exhibiting an *ERBB2* amplification represents an intriguing result, as it is both a relatively uncommon, yet significant, mechanism of resistance to EGFR inhibition and a targetable vulnerability, as we have shown (Fig. 1A, 6B-C)⁴⁷.

Regarding underexpression events, we noted *MLH1* and *MGMT*, whose status has a known predictive or prognostic role in CRC patients, in the subset of genes for which a relatively high number of CRC cell lines have been recognized as extreme negative outliers (Fig. 4)^{50,51}.

MLH1 loss of expression was detected in 31 CRC cell lines and, as expected, microsatellite instability was observed in all *MLH1* extreme negative outliers (Fig. 7A). On the other hand, 41 CRC cell lines were scored as *MGMT* extreme negative outliers. Of note, a subset of CRC cell lines labelled as *MGMT* extreme negative outliers was previously tested with TMZ: while MSS and *MGMT*-deficient CRC cell lines are mostly sensitive to TMZ, the MSI status confers resistance to TMZ regardless of *MGMT* silencing (Fig. 7B)⁵².

We then compared our workflow with an alternative method for detecting aberrantly expressed genes. Specifically, we selected OUTRIDER since it was specifically developed to use RNA-seq data, distinguishing it from other methods that were originally designed for microarray data⁵³. Moreover, it has both relevant features in common with our method and interesting differences. Like our method, OUTRIDER is

capable of detecting both positive and negative outliers without being restricted to specific classes of genes. On the other hand, OTRIDER provides a statistical test to determine whether a sample is an outlier for a particular gene after fitting a probabilistic model that accounts for the typical distribution of RNA-seq read counts. This model is designed to handle unusually high variation in gene expression (overdispersion) and a high number of genes with no detectable expression (zero inflation), commonly observed in RNA-seq data. Additionally, OTRIDER includes methods to correct for count variations caused by confounding factors, such as batch effects or other sources of technical noise, ensuring that the detected outliers are true biological aberrations rather than artifacts of data processing. In detail, it bases the data correction on an autoencoder, a neural network that compresses gene expression data into a smaller set of latent variables, and then reconstructs it from this compact representation. The complexity of the model is controlled by the number of latent variables, which determines how detailed or general the representation of the data will be⁵³.

In our analysis, we tested how different configurations of OTRIDER impact on the reliability of its result. We first applied OTRIDER's default settings, which use an automatically estimated optimal encoding dimension, which for our dataset was estimated to be 14⁵³. Comparing the results obtained with our workflow, OTRIDER found fewer outlier samples for a lower number of genes: specifically, 2,049 overexpressed genes were found in 148 CRC cell lines, while 836 underexpressed genes were identified in 157 models (Fig. 8 A-B).

To assess the impact of model complexity on outlier detection, we experimented with an autoencoder that used a more simplified representation of the data, compressing it into half latent variables (7), with the aim of focusing on the most significant variations in gene expression to identify the most robust outliers. Additionally, a more complex model with double the variables (28) was tested to determine whether this could uncover more subtle outliers. This analysis demonstrated that increasing model complexity leads to the detection of more outliers, at the risk of identifying noise as signal (Fig. 8 A-B).

We also used OTRIDER with an alternative data correction approach based on PCA⁵³. While the PCA-based correction detected the highest number of positive outliers

(3,703), it also missed more negative outliers (555), suggesting that while it may be sensitive to genes with high expression, it might also overcorrect and miss genes with low expression (Fig. 8 A-B).

As the number of identified outliers is an insufficient criterion for evaluating the quality of the method, we proceeded to assess the performance of OUTRIDER in identifying the molecular features that we had previously presented.

Importantly, OUTRIDER failed in identifying several of the clinically relevant and previously validated features of CRC cell lines that we mentioned above.

In detail, OUTRIDER recognized all models with known *ERBB2* overexpression as outliers for the *ERBB2* gene, irrespective of the selected configuration. However, only a variable number of models between one (using PCA-based correction, 14 latent variables) and three (using default settings) out of five with *NTRK1* overexpression were selected as positive outliers by OUTRIDER. In another example, only the OUTRIDER model that used PCA-based correction and 14 latent variables was able to pinpoint at least one of three models with *NTRK2* overexpression and the only model with *FGFR2* overexpression that are responsive to the pharmacological inhibition of the corresponding TK proteins. Moreover, OUTRIDER never detected any of the four models with *RET* overexpression that are sensitive to the corresponding TK inhibitor (Fig. 8 C-H).

Furthermore, according to the OUTRIDER output, *MLH1* and *MGMT* loss of expression were not identified in CRC cell lines, regardless of the chosen model configuration (Fig. 8 I-J).

In summary, the strategy that we have devised was able to correctly identify gene expression abnormalities that have been previously associated with clinically relevant features of CRC cell lines, thus supporting its effectiveness. In addition, it outperformed an alternative algorithm for gene expression outlier analysis.

2.4 Aberrant DNA methylation in promoter regions underlies outlier expression values that are recurrently found in CRC cell lines

Epigenetic modifications critically control gene expression levels⁵⁴.

Therefore, we assessed whether extreme gene expression outliers are characterized by aberrant DNA methylation in promoter regions of overexpressed or underexpressed genes.

First, DNA methylation microarray data analysis confirmed high DNA methylation levels in *MLH1* or *MGMT* promoter regions in CRC cell lines recognized as extreme negative outliers for each of these genes (Fig. 7).

Then, the analysis was extended to all genes for which at least two CRC cell lines were pinpointed as extreme positive or negative outliers, thus including both overexpression and underexpression events. Globally, this revealed 198 genes for which underexpression in extreme negative outliers was associated with high DNA methylation levels in promoter regions (promoter hypermethylation) (Table 3). For example, few CRC cell lines (HROC278_MET, HROC278_TOM1, HROC334 and KP363T) were selected as extreme negative outliers for the *RAD51C* gene. In previous works, we found that these cells are highly sensitive to inhibition of PARP and other proteins involved in the DNA damage response^{55,56}. Here we observed that in these models *RAD51C* silencing is associated with promoter hypermethylation, a feature previously detected in cancer types other than CRC (Fig. 9)⁵⁷.

In addition, we found 102 genes for which the reduction of DNA methylation levels in promoter regions (promoter hypomethylation) likely contributed to overexpression in extreme positive outliers (Table 3).

We also examined the position of each individual CpG within the promoter regions by taking into account CpG islands, shores, shelves and open sea. For both extreme positive outliers and extreme negative outliers, we observed that the distribution in the different regions differs between CpG sites that are or not differentially methylated between outlier and other samples (chi-square p-value = 0.00023 and 7.6e-12, respectively). More in general, in both cases the majority of differentially methylated CpG sites is located within CpG islands (Fig. 10A and Table 3).

In addition, we found that CIMP-H samples are characterized by a slightly higher fraction of underexpressed genes whose silencing is associated with promoter hypermethylation (Figure 5I). This is consistent with the global genome hypermethylation that characterizes this subset of CRC samples⁵⁸. In the presence of the CIMP phenotype, multiple genes are expected to be silenced through promoter

hypermethylation, while, in CIMP-negative samples, this event should be sporadic and thus involve a lower number of genes.

Finally, we considered the extreme negative outliers of genes for which we found a correspondence between underexpression and promoter hypermethylation and tested whether they were significantly enriched for CIMP-positive samples (CIMP-H and CIMP-L). Using this strategy, we were able to distinguish 23 genes associated with CIMP and 138 genes that were silenced through promoter hypermethylation in CRC cell lines independently from CIMP (Fig. 10B and Table 4).

2.5 Outlier expression values are associated with genomic rearrangements

Large genetic abnormalities, such as gene amplifications, gene deletions and gene fusions, are a potential source of gene expression outliers^{23,47,59}. Therefore, we assessed their impact in our collection of CRC extreme expression outliers (Table 2).

First, we identified 174 genes whose overexpression was associated with gene amplification in at least some extreme positive outliers and 104 genes whose underexpression was concomitant with genetic deletion (Fig. 11A).

In line with the concept that tumour suppressor genes are frequently lost or inactivated during colorectal tumorigenesis, *PTEN* and *SMAD4* were among the recurrently underexpressed and deleted genes in CRC cell lines^{60,61}. More in details, 5 CRC cell lines were scored as *PTEN* extreme negative outliers and exhibited genetic deletion at the genome locus that includes *PTEN* (HROC277_MET2, KP283T, OUMS23, SNU479 and SW1222) (Fig. 11B). Additionally, consistent with the fact that the HROC277_T0M1 and HROC277_MET2 CRC cell lines were derived, respectively, from the primary tumour and a metastatic lesion obtained from the same patient, *PTEN* genetic deletion was also found in the HROC277_T0M1 model that had very low *PTEN* expression.

A total of 14 CRC cell lines were scored as *SMAD4* extreme negative outliers and 9 of them were also characterized by *SMAD4* genetic deletion (COLO678, HROC239_T0M1, HROC284_MET, HROC300_T2M1, HROC59, IRCC10_A_HL, KP363T, SNU1181 and SNU1411) (Fig. 11C).

Furthermore, these copy number alterations were distinguished in focal and non-focal amplifications or deletions by analysing co-occurring overexpression or underexpression events due, respectively, to co-amplification or co-deletion of genes that are located in the same genomic regions. We found that 74% of amplified genes were attributable to non-focal events, 11% to focal events and 15% exhibited diverse amplification patterns across different samples. For deleted genes, 35% were due to non-focal events, 33% to focal events and 32% to different events observed in different samples.

For example, 3 CRC cell lines were found as *ERBB2* extreme positive outliers (CRC0080_XL, IRCC72_A_XL, LM0701_A_XL) and all of them were also characterized by overexpression and amplification of 6 additional genes (*CDK12*, *GRB7*, *MED1*, *MIEN1*, *PGAP3* and *STARD3*) that are positioned in the same genomic region as the *ERBB2* gene (chromosome 17q12) (Fig. 11D).

Next, RNA-seq data were leveraged to identify somatic fusion transcripts in CRC cell lines and we checked whether the presence of each fusion transcript matched with the overexpression of the 3' partner gene. It is indeed possible that its expression comes under the control of a stronger promoter if the fusion transcript results from a genomic rearrangement. We found 24 gene fusions associated with the overexpression of 23 genes in 30 CRC cell lines (Fig. 12A). We observed that the majority of them involves as 5' partner a gene that is normally expressed in CRC cells and as 3' partner a gene that is not expressed in the majority of models (median FPKM < 1) (Fig. 12B-C). We also confirmed the presence of actionable gene fusions that we functionally characterized in previous studies, namely *EML4-ALK*, *LMNA-NTRK1* and *TMP3-NTRK1*^{47,48}. In addition, the *PTPRK3-RSPO3* gene rearrangement was found in two previously characterized models (SNU1411 and CRC0781_XL)^{23,62} and also found in the HROC300_T2M1 cell line. Finally, the *MTMR3-APOH* fusion was detected in 4 CRC cell lines (CX1, HT29, KM20 and WIDR) that were found as *APOH* extreme positive outliers, thus further validating the approach as these cell lines were derived from the same patient.

2.6 Analysis of overexpressed enzyme genes pinpoints therapeutic targets

As discussed above, the identification of overexpressed genes may unveil new drug targets. Here we focused on genes predicted to encode proteins endowed with enzymatic activity, since they are ideally suited for pharmacological inhibition. Starting from 20,375 human proteins extracted from the Swiss-Prot UniProtKB database, we found 3,704 enzyme genes expressed in CRC cell lines (Fig 13A). Among these, we identified at least one overexpression event for 759 enzyme genes.

As a first step, we applied the classification scheme proposed by the Illuminating the Druggable Genome (IDG) initiative for target prioritization⁶³. In brief, human proteins have been categorised into four groups (Tclin, Tchem, Tbio and Tdark) according to the target development level (TDL), that is based on the extent of their clinical, chemical, and biological characterization. Tclin proteins are those that are associated with approved drugs that have well-defined mechanisms of action, making them the most clinically relevant. Tchem proteins interact with potent small molecules that are used in preclinical research, though they are not yet linked to approved drugs. Tbio proteins are supported by substantial biological data, such as disease associations and experimental evidence, but are not targeted by existing drugs or small molecules. Finally, Tdark proteins are the least characterized, with minimal known information, representing potential but largely unexplored targets in drug discovery⁶³.

Overexpressed enzyme genes were distributed in all TDL categories, including enzymes for which clinically approved drugs or small molecules are already available: 73 and 218 genes indeed belonged to the Tclin and Tchem category, respectively (Fig. 13B).

Next, we noticed that the list of overexpressed enzyme genes included many kinase genes. Therefore, we decided to focus on kinases, starting from 495 genes obtained from the integration of the kinases in Wilson et al. and in the Swiss-Prot UniProtKB database (Fig 14A)⁶⁴. Among these, we identified at least one overexpression event for 70 kinase genes (Fig 14B)⁶⁴.

Overexpressed kinase genes were also found across all TDL categories, but the TK family is the only one found in the Tclin category that, as expected, included the clinically relevant kinases that we discussed above (Fig. 14B).

We also found cell models bearing overexpression of additional TK genes that are included in the Tclin class (*FES*, *FGFR3*, *LCK*, *MET* and *SYK*) and other kinases for which small molecules are available (Tchem category) (Fig. 14B).

2.7 Characterization of underexpressed genes unveils synthetic lethal vulnerabilities in CRC cells

Several studies have previously reported synthetic lethal interactions between metabolic enzymes^{65,66}. Intriguingly, we identified 274 enzyme genes that are underexpressed in one or multiple CRC cell lines (extreme negative outliers). In particular, seven CRC cell lines (C10, CAR1, COLO678, CRC0558_XL, CRC0781_XL, HROC284_MET and RW2982) were identified as extreme negative outliers for the methylthioadenosine phosphorylase (*MTAP*) gene. In the human genome, the *MTAP* gene is located at chromosome 9p21.3 adjacent to two tumour suppressor genes (*CDKN2A* and *CDKN2B*) and co-deletion is frequently observed in cancer cells (Fig. 15A)^{67,68}. In fact, we found that *MTAP*, *CDKN2A* and *CDKN2B* were co-deleted in *MTAP* extreme negative outliers (Fig. 15 A-B), with *CDKN2A* and *CDKN2B*-deleted CRC cell lines exhibiting very low or zero expression values (Fig. 15 C-D).

MTAP is a key enzyme in the adenine and methionine salvage pathway and when deleted results in the accumulation of 5-methylthioadenosine (MTA). MTA competes with the methyl donor S-adenosylmethionine (SAM) for binding to the PRMT5 enzyme, leading to its partial inhibition (Fig 16 A). Recent studies also reported that the dysregulated metabolic state of MTAP-deficient cancer cells gives rise to a targetable vulnerability^{67,68}. The viability of MTAP-deficient cells is indeed impaired by inhibition of the protein arginine methyltransferase 5 (PRMT5) enzyme whose activity can be reduced both indirectly and directly, through MAT2A inhibition or by targeting the PRMT5 MTA complex that selectively appears in the absence of MTAP^{69–72}.

Therefore, 6 *MTAP*-deleted CRC cell lines were selected for biological characterization, along with 5 CRC cell lines (HROC46, C125PM, IRCC72_A_XL, LIM2099 and OXCO1) in which the *MTAP* gene is normally expressed. Since the *MTAP*-deleted cell line RW2982 grows very poorly, it was excluded from further experiments. Loss of MTAP protein expression in *MTAP*-deleted cell lines was confirmed by Western blot (Data not shown). Then, we assessed whether inhibitors that determine reduction of PRMT5 activity are selectively active in *MTAP*-deleted CRC cell lines (Fig 16 A). When the cell models were treated with increasing concentrations of an inhibitor of the metabolic

enzyme MAT2A (AGI-24512), we observed a modest growth reduction in 5 out of 6 *MTAP*-deleted CRC cell lines while viability of control cell lines was not affected (Fig 16 A,B). In line with what can be observed in the drug response curves, the difference between *MTAP*-deleted and wild-type CRC cell lines was statistically significant when two relevant concentrations were considered: 0.74 μ M (one-way ANOVA $p=0.00798$) and 2.22 μ M (one-way ANOVA $p=0.00536$). We then treated the same subset of CRC cell lines with an inhibitor of the PRMT5:MTA complex (MRTX1719), a recently developed drug that has shown promising results in solid tumours (Fig 16A)^{70,72}. Notably, treatment with MRTX1719 was much more effective in limiting the growth of *MTAP*-deleted CRC cell lines as compared with wild-type cells (Fig. 16C). The difference was statistically significant at multiple drug concentrations: 9.14nM (one-way ANOVA $p=0.00207$), 27.43nM (one-way ANOVA $p=9.8e^{-5}$), 82.3nM (one-way ANOVA $p=2.09e^{-5}$), 246.9 nM (one-way ANOVA $p=0.000126$).

2.8 Aberrantly expressed genes in both CRC cell lines and tissue samples obtained from colorectal tumours

To further assess the clinical relevance of our findings in CRC preclinical models, we checked whether extreme positive and negative outliers are found in human tumour samples for the same genes that are aberrantly expressed in cell lines. To this end, we obtained gene expression data of 611 primary colorectal tumours from The Cancer Genome Atlas (TCGA) project⁶⁰.

Positive and negative outliers were found for 14,913 and 7,890 genes, respectively, when the Tukey's rule was applied to the expression profile of each individual gene.

Before proceeding with the selection of the extreme gene expression outliers in the TCGA dataset, it should be noted that direct application of the same method to gene expression data from either cell models or tumour samples may be technically challenging to the inherent features of the different sample types. While in the case of cell models expression signals solely derive from tumour cells, bulk RNA-seq data from intact tumour tissues inevitably reflect the average of expression signals from multiple cell types, thus potentially confounding the outlier identification. Importantly, a previous study has shown that the tumour purity is not uniform among TCGA samples,

that it is usually lower in samples with high tumour mutational burden and that different purity estimates may be obtained for the same sample when different methods are applied⁷³.

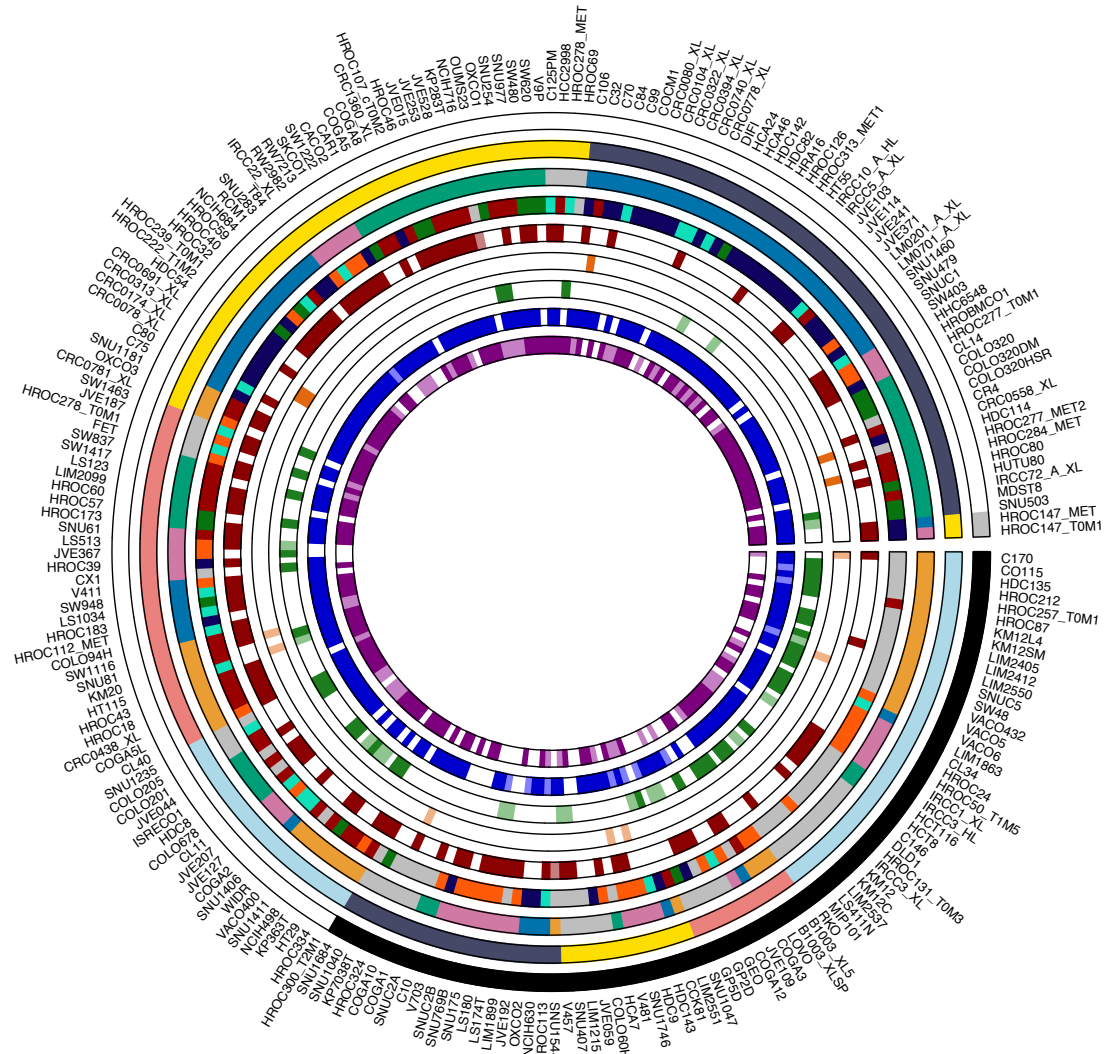
For this reason, we examined the tumour purity of the 611 primary CRC TCGA samples that we considered in outlier calling. Specifically, it was previously estimated using five alternative methods: ESTIMATE takes into account the gene expression profiles of several immune and stromal genes, ABSOLUTE uses somatic copy-number data, LUMP (leukocytes unmethylation for purity) averages several non-methylated immune-specific CpG sites, IHC is based on image analysis of haematoxylin and eosin stain slides and CPE (consensus measurement of purity estimation) is the median purity level after normalizing values from all the previous methods⁷³. As the requisite data for this analysis was unavailable for all samples, the analysis was conducted solely on 544 of the 611 primary CRC samples.

The distribution of the calculated purity estimates indicates that tumour purity is typically lower in MSI than MSS samples, with the exception of the IHC measurement (Fig 17A)⁷³. In the context of gene expression outlier identification, ABSOLUTE should deserve greater consideration since it provides an actual estimate of tumour cell prevalence in tissue samples in the form of a copy-number average. The average tumour purity across 544 out of 611 primary CRC samples was $64.2\% \pm 16.7\%$ (mean $\pm$ sd), indicating significant contributions from non-tumor cells in the gene expression measurements (Fig 17A).

In line with the previous considerations, when we selected the furthest outliers for each gene in the TCGA dataset, based on the same criterium used before, we observed that the deviation of positive and negative outliers from the other samples was usually less prominent than in the CRC cell line dataset (Wilcoxon test p-value $< 2.2e-308$ in both cases) (Fig. 17 B-C). Therefore, we chose alternative thresholds to pinpoint the extreme outliers in the TCGA dataset, making sure that the fraction of extreme positive or negative outliers compared to the total number of positive and negative outliers was comparable to that obtained in the CRC cell line dataset (differential expression > 2 in the case of positive outliers and > 1 in the case of negative outliers). In this way, we identified 6,005 and 3,083 genes for which CRC samples were recognized as extreme positive or negative outliers, respectively. Importantly, in the TCGA dataset

we found extreme positive outliers for 2,005 genes whose overexpression was also observed in CRC cell lines (56.75%), while we confirmed that the underexpression of 437 genes for which extreme negative outliers were found in the CRC cell line dataset also occur in human tumours (45.28%).

A



1. MSS/MSI

- MSS (65%)
- MSI (34%)
- NA (1%)

2. CIMP

- CIMP-H (24%)
- CIMP-L (18%)
- CIMP3 (27%)
- CIMP4 (31%)

3. CMS

- CMS1 (18%)
- CMS2 (28%)
- CMS3 (15%)
- CMS4 (22%)
- NA (16%)

4. CRIS

- CRIS-A (17%)
- CRIS-B (20%)
- CRIS-C (21%)
- CRIS-D (10%)
- CRIS-E (11%)
- NA (21%)

5. KRAS

- Hotspot mutation (47%)
- Other missense mutation (1%)

6. NRAS

- Hotspot mutation (2%)
- Other missense mutation (3%)

7. BRAF

- Hotspot mutation (18%)
- Other missense mutation (8%)

8. APC

- Nonsense mutation, frameshift indel (79%)
- Missense mutation, in-frame indel (5%)

9. TP53

- Hotspot mutation, nonsense mutation, frameshift indel (57%)
- Other missense mutation, in-frame indel (19%)

B

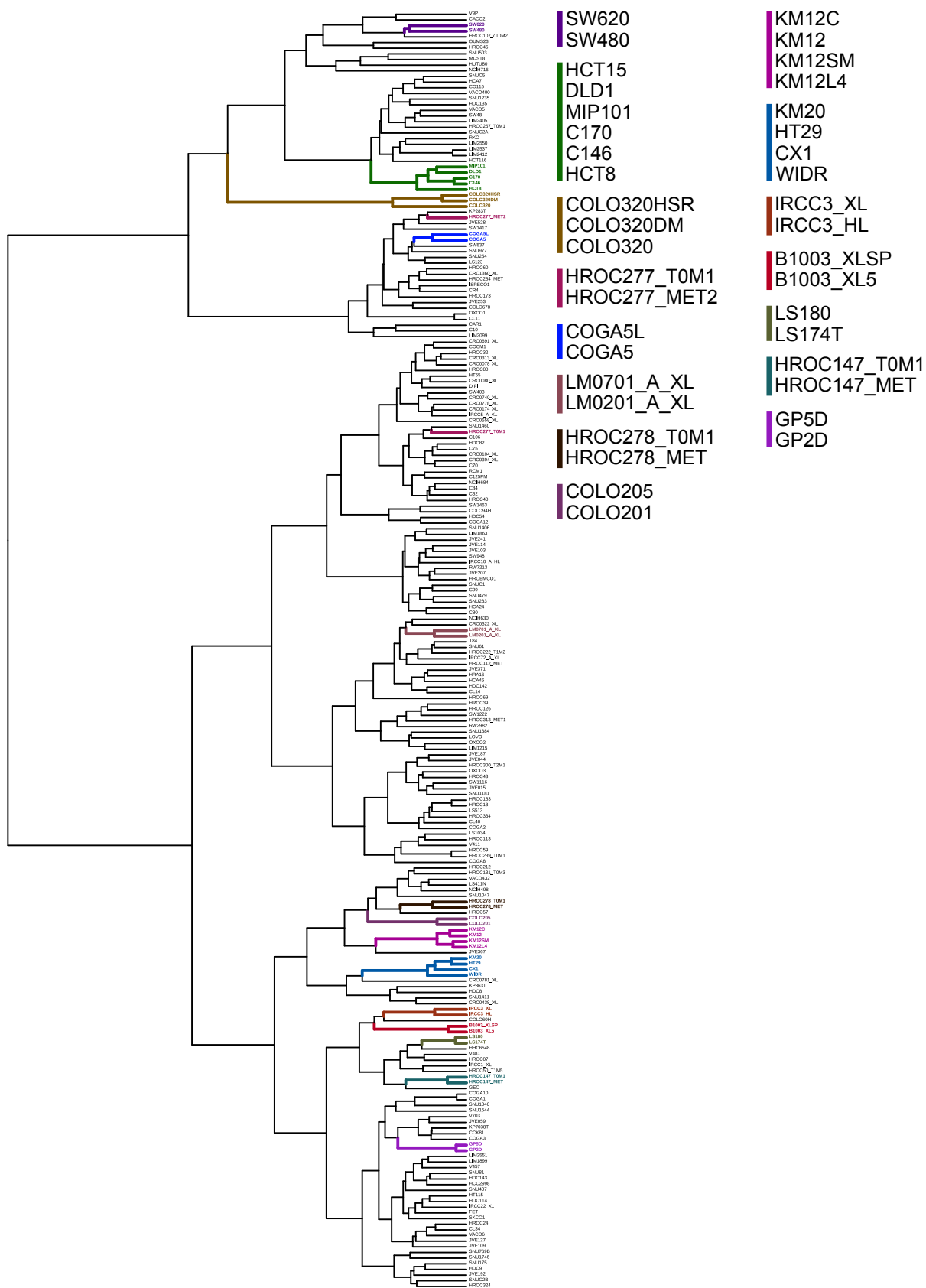


Figure 1: Genetic, transcriptomic and epigenetic features of 226 colorectal cancer (CRC) cell lines.

(A) Circos plot delineating clinically relevant molecular features of 226 CRC cell lines. Each layer corresponds to a different genetic, transcriptomic or epigenetic feature, organized from the outside to the inside in the following order: microsatellite status as microsatellite stable (MSS) or instable (MSI); *KRAS* mutational status; *NRAS* mutational status; *BRAF* mutational status; *APC* mutational status; *TP53* mutational status; consensus molecular subtypes (CMS) as CMS1, CMS2, CMS3 or CMS4; CRC intrinsic subtypes (CRIS) as CRIS-A, CRIS-B, CRIS-C, CRIS-D or CRIS-E; classification based on the CpG island methylator phenotype (CIMP) as CIMP-H, CIMP-L, CIMP3 or CIMP4. In the layers corresponding to CMS and CRIS gene expression-based subtypes, CRC cell lines that cannot be confidently assigned to a single subtype (FDR>5%) were labelled as NA (not available).

(B) Dendrogram depicting the hierarchical clustering of 226 CRC cell lines based on their gene expression profiles. Groups of genetically identical cell lines, i.e., cell lines with same short tandem repeats (STR) profile, that derived from a same individual are highlighted in different colours and their names are reported on the right in the order by which they occur in the dendrogram, from top to bottom.

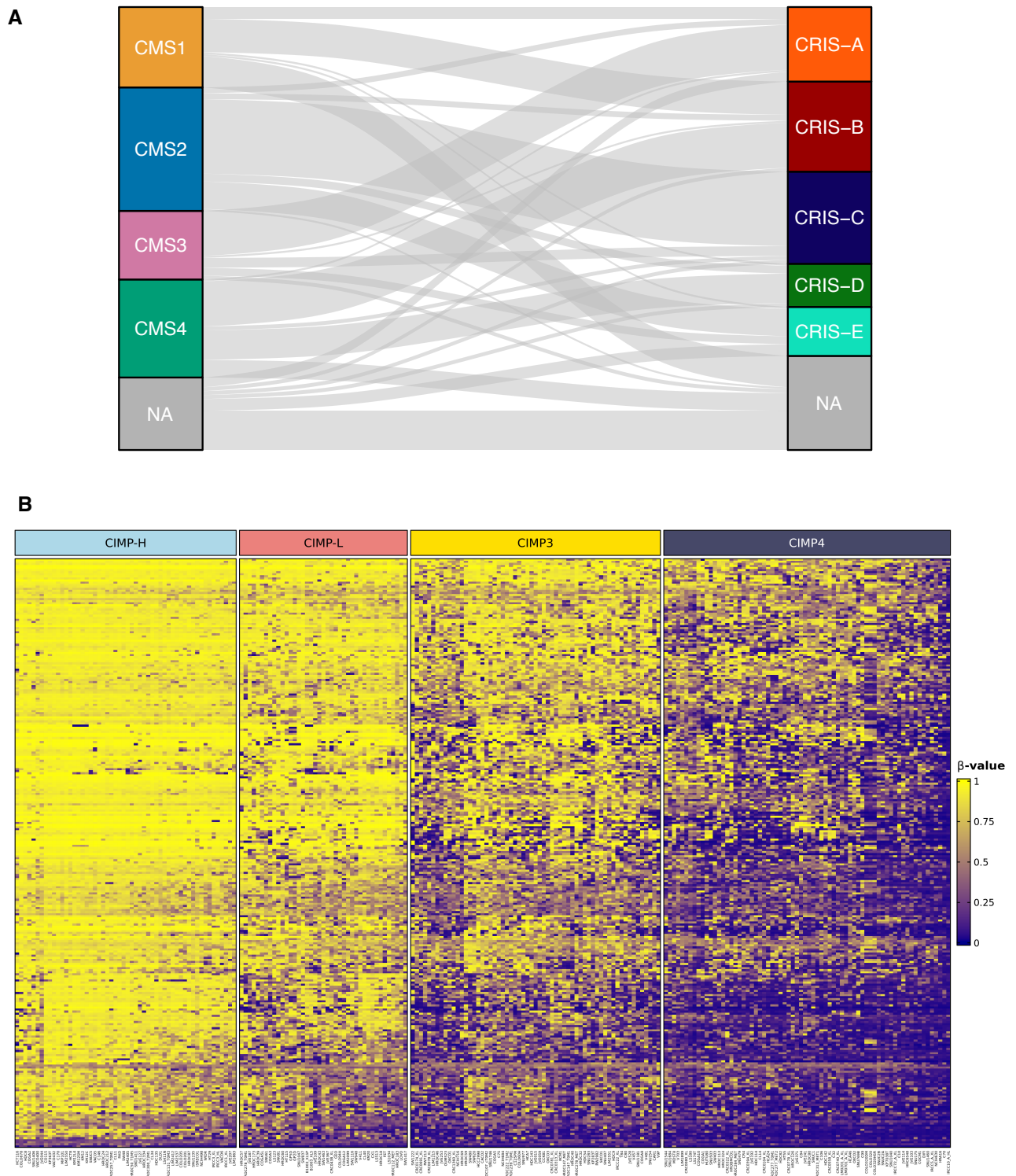


Figure 2: Molecular classifications of 226 colorectal cancer (CRC) cell lines.

(A) Alluvial plot depicting correspondences between CMS and CRIS gene expression- based subtypes identified in 226 CRC cell lines. In both classifications, CRC cell lines that cannot be confidently assigned to a single subtype (FDR>5%) were labelled as NA (not available).

(B) Heatmap depicting DNA methylation levels of 289 CIMP associated probes in 226 CRC cell lines. DNA methylation levels were measured as β -values and are represented using a colour scale from

dark blue (low DNA methylation level) to yellow (high DNA methylation level). DNA methylation-based subtypes were identified through unsupervised clustering and they were associated to different CIMP classes based on their median DNA methylation level (CIMP-H, CIMP-L, CIMP3 and CIMP4, from high to low median DNA methylation level), as reported above the heatmap. Samples are split according to CIMP classes and, in each slice, they are arranged based on hierarchical clustering done using the Euclidean distance metric and average linkage. The probes are arranged based on hierarchical clustering done using the Euclidean distance metric and average linkage.

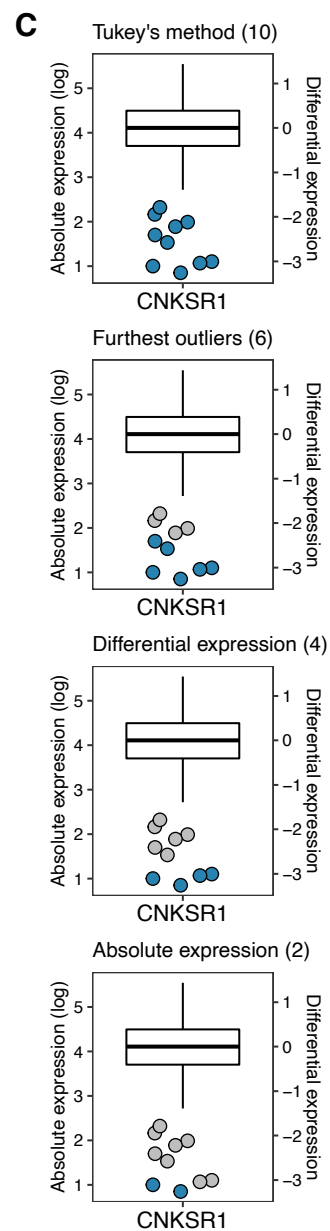
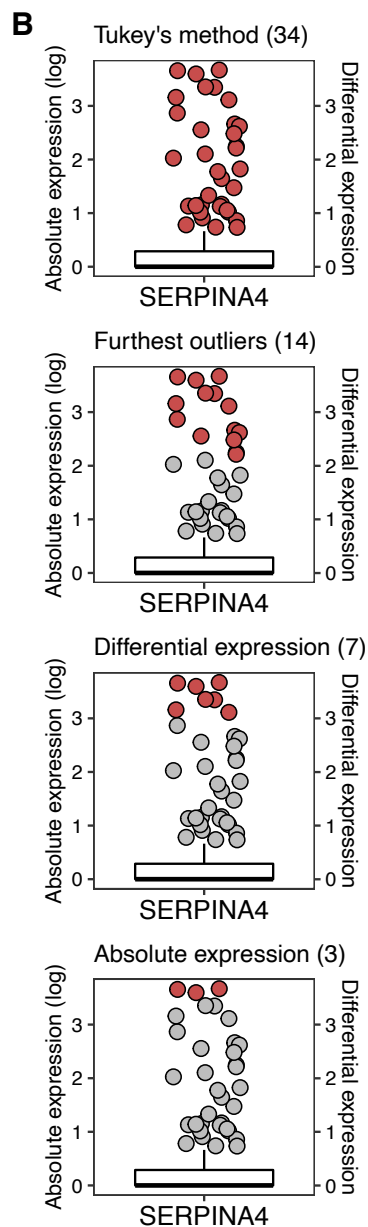


Figure 3: Identification of overexpression and underexpression events in 226 colorectal cancer (CRC) cell lines.

(A) Graphical representation of the computational pipeline. The number of genes for which positive or negative outliers(s) were found after each step of the pipeline is shown on the left and the right side, respectively.

(B) Identification of extreme positive outliers for the *SERPINA4* gene. Each boxplot shows the distribution of *SERPINA4* expression values. Red dots are samples selected as positive outliers after each step of the pipeline. Their number is also reported between brackets in the title of each subpanel.

(C) Identification of extreme negative outliers for the *CNKSR1* gene. Each boxplot shows the distribution of *CNKSR1* expression values. Blue dots are samples selected as negative outliers after each step of the pipeline. Their number is also reported between brackets in the title of each subpanel.

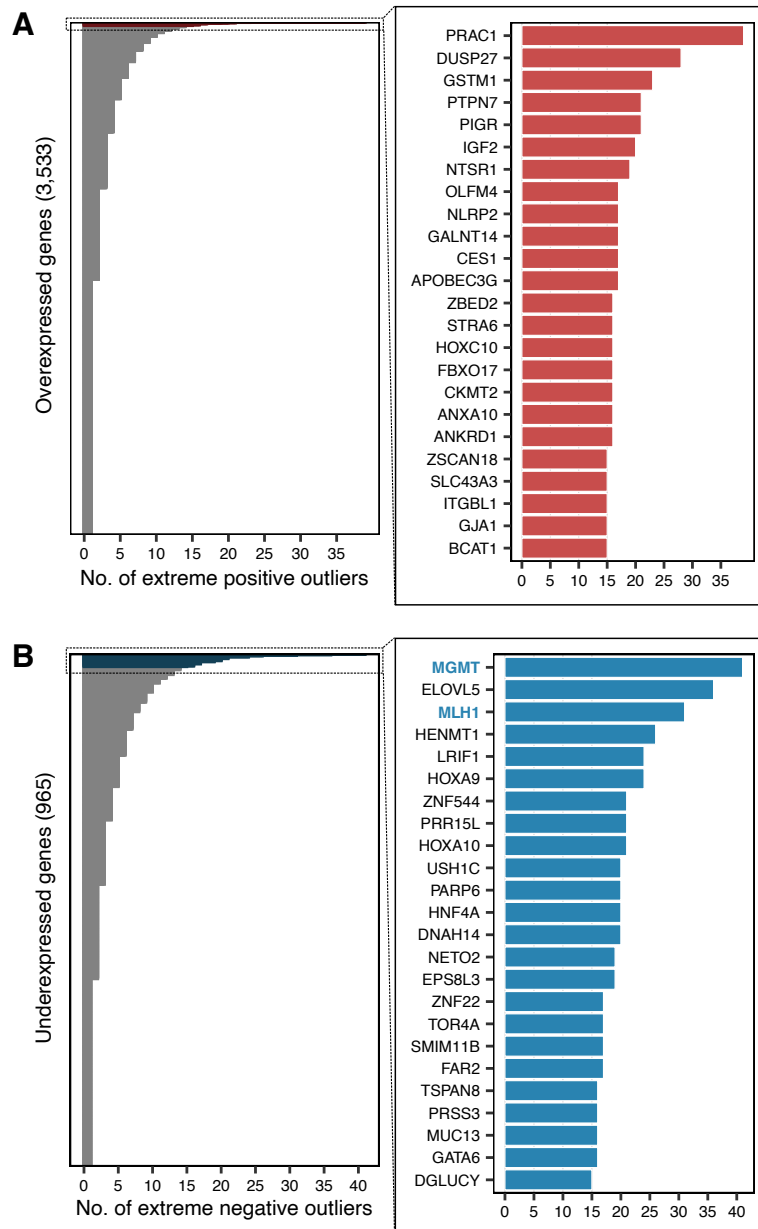


Figure 4: Gene expression outlier frequency among 226 colorectal cancer (CRC) cell lines.

(A) In the left subpanel, the number of extreme positive outliers is shown for each overexpressed gene. A detailed representation is reported for genes with most elevated outlier frequency in the right subpanel (number of outlier samples > 15).

(B) In the left subpanel, the number of extreme negative outliers is shown for each underexpressed gene. A detailed representation is reported for genes with most elevated outlier frequency in the right subpanel (number of outlier samples > 15).

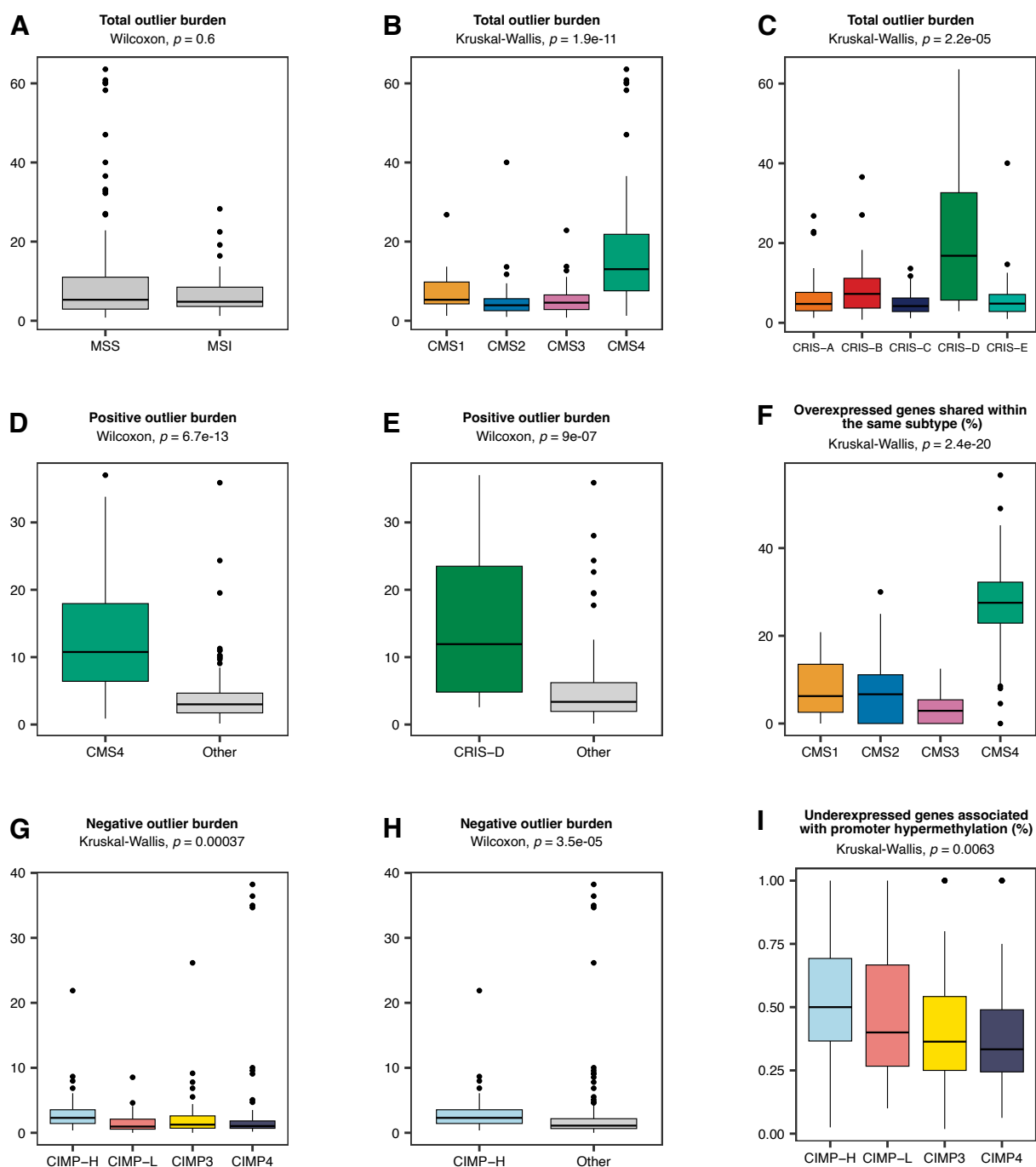


Figure 5: Outlier burden values are associated with molecular features of colorectal cancer (CRC) cell lines.

(A) Boxplots showing the distribution of total outlier burden values in samples stratified according to the microsatellite status defined as microsatellite stability (MSS) or instability (MSI).

(B) Boxplots showing the distribution of total outlier burden values in samples stratified according to consensus molecular subtypes (CMS).

(C) Boxplots showing the distribution of total outlier burden values in samples stratified according to CRC intrinsic subtypes (CRIS).

(D) Boxplots showing the distribution of positive outlier burden values in samples stratified according to whether they belong to the CMS4 subtype or to other CMS subtypes.

(E) Boxplots showing the distribution of positive outlier burden values in samples stratified according to whether they belong to the CRIS-D subtype or to other CRIS subtypes.

(F) Boxplots showing the distribution of the fraction of overexpressed genes whose overexpression only occurs within the same subtype in samples stratified according to the CMS transcriptional classification.

(G) Boxplots showing the distribution of negative outlier burden values in samples stratified according to CpG island methylator phenotype (CIMP).

(H) Boxplots showing the distribution of negative outlier burden values in samples stratified according to whether they belong to the CIMP-H subtype or to other DNA methylation-based subtypes.

(I) Boxplots showing the distribution of the fraction of underexpressed genes whose silencing is associated with promoter hypermethylation in samples stratified according to the CIMP classification.

In each boxplot the box corresponds to the interquartile range (IQR) and the median is represented by the centre line. Whiskers maximally extend to $1.5 \times \text{IQR}$ and outliers are shown. P values that are shown above each graph were calculated by a Wilcoxon test when two sample groups are present, or by a Kruskal-Wallis test when multiple sample groups exist.

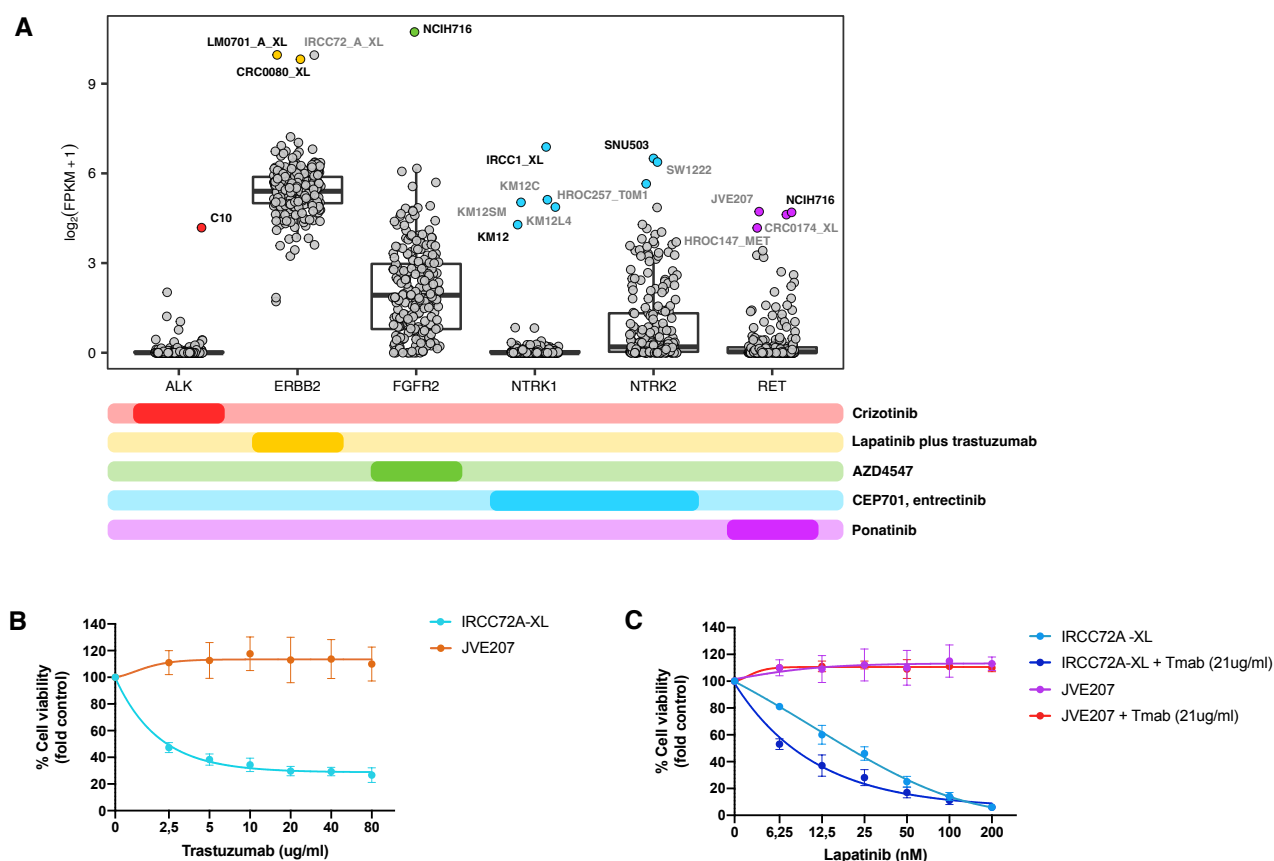
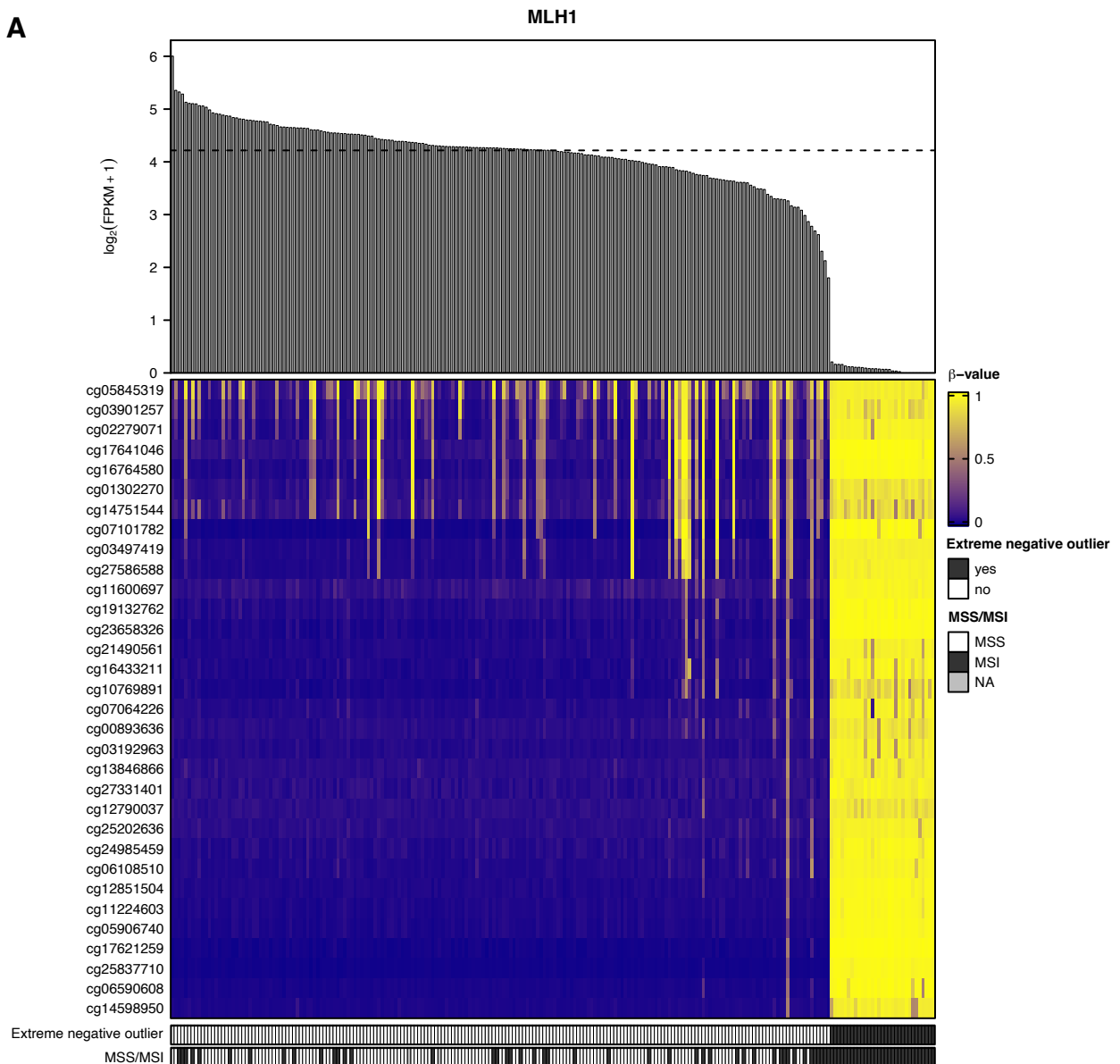


Figure 6: Extreme positive outliers for TK genes included sensitive cell models to the corresponding kinase inhibitors

(A) Boxplots depicting the distribution of expression values of tyrosine kinase (TK) genes whose overexpression has been previously associated with oncogene addiction in CRC cells. For each TK gene, extreme positive outliers are highlighted with dots of a particular colour, while other samples are reported as grey dots. The name of each CRC cell line selected as extreme positive outlier is also indicated (in grey those that were not included in a previous drug screening). The bottom panel summarizes the matching between overexpressed TK genes and kinase inhibitors whose effect on cell viability has been previously assessed.

(B-C) Drug screenings in *ERBB2* amplified and wild-type CRC cell lines. Cell viability was assessed by measuring ATP content after 6 days of treatment with increasing concentration of Trastuzumab (B) or Lapatinib, alone or in combination with trastuzumab (C). Orange, pink and red lines correspond to the cell line JVE207, negative control, in which the *ERBB2* gene is normally expressed, while lines in shades of blue correspond to the cell line IRCC72A_XL, extreme positive outlier for *ERBB2*. Cell viability measured for each treatment condition was normalized to viability of untreated cells. Final data are an average of at least three individual biological replicates.

A



B

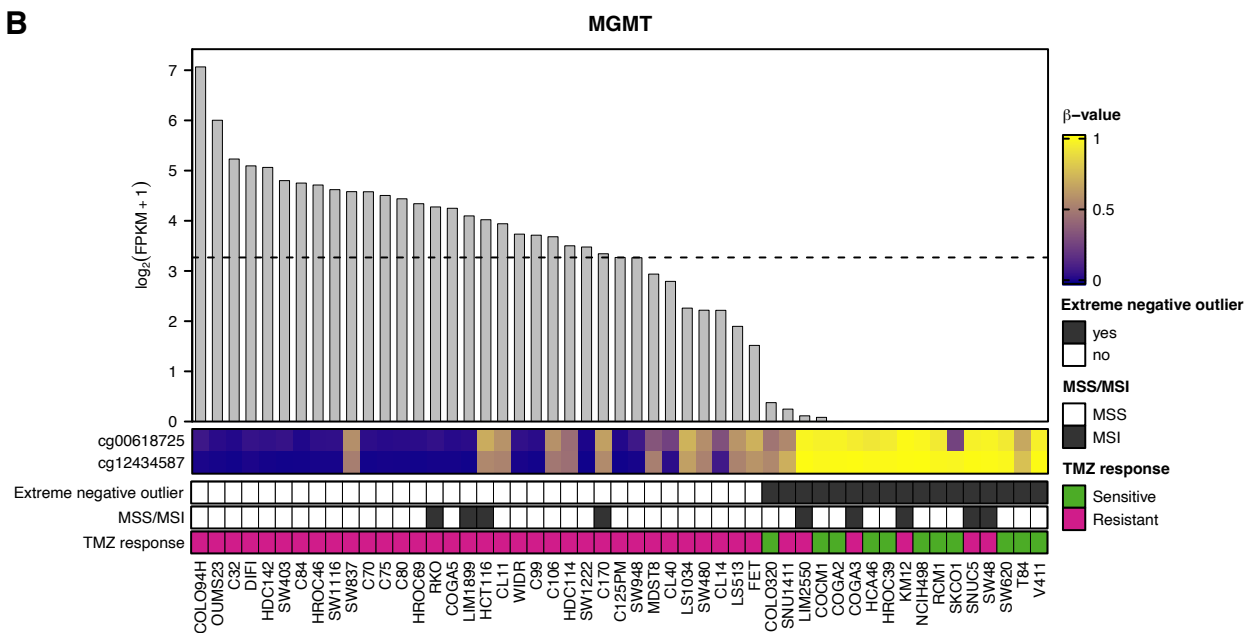
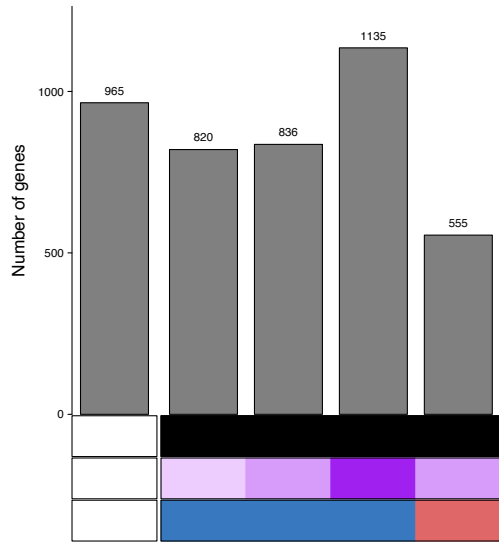


Figure 7: Gene underexpression events associated with clinically relevant features of colorectal cancer (CRC) cell lines.

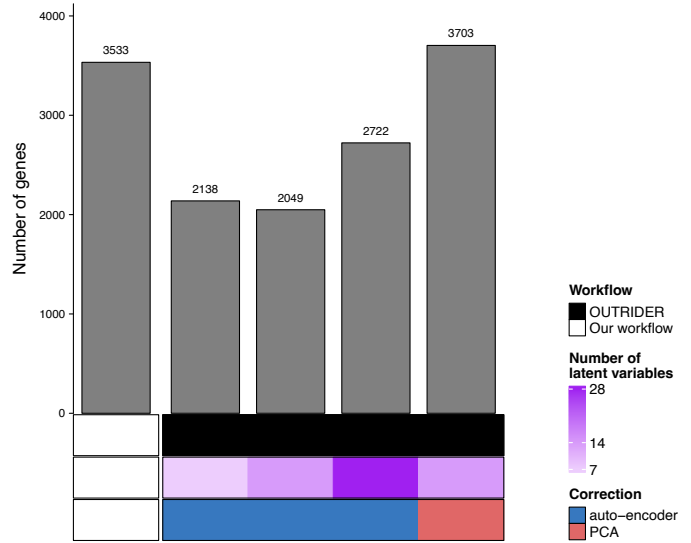
(A) Heatmap depicting DNA methylation levels of promoter probes that were found differentially methylated between *MLH1* extreme negative outliers and other samples. DNA methylation levels were measured as β -values and are represented using a colour scale from dark blue (low DNA methylation level) to yellow (high DNA methylation level). *MLH1* expression profile is shown above the heatmap and sample ordering is the same in the two graphs, from high to low *MLH1* expression levels. Annotation bars below the heatmap indicate – from top to bottom – samples that were scored as *MLH1* extreme negative outliers and the microsatellite status of each sample defined as microsatellite stability (MSS) or microsatellite instability (MSI).

(B) Heatmap showing DNA methylation levels of promoter probes that were found differentially methylated between *MGMT* extreme negative outliers and other samples. DNA methylation levels were measured as β -values and are represented using a colour scale from dark blue (low DNA methylation level) to yellow (high DNA methylation level). *MGMT* expression profile is shown above the heatmap and sample ordering is the same in the two graphs, from high to low *MGMT* expression levels. Annotation bars below the heatmap indicate – from top to bottom – samples that were scored as *MGMT* extreme negative outliers, the microsatellite status of each sample and sample response to temozolomide (TMZ). Only CRC cell lines that were previously tested for TMZ sensitivity are shown in the figure.

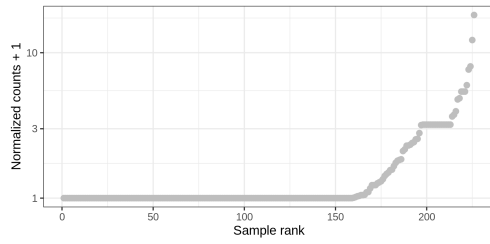
A Underexpressed genes



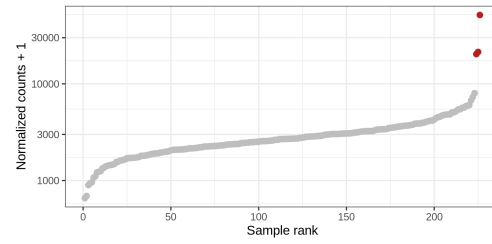
B Overexpressed genes



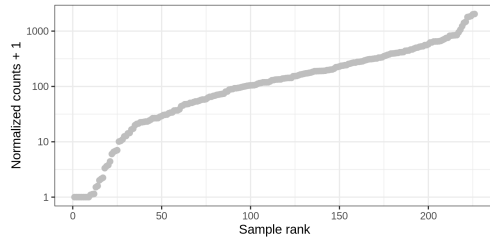
C Expression rank plot: ALK



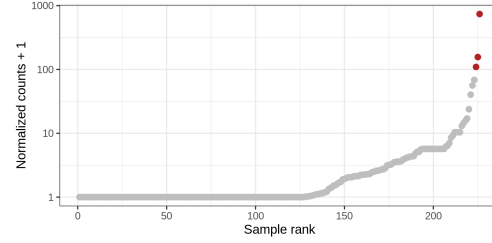
D Expression rank plot: ERBB2



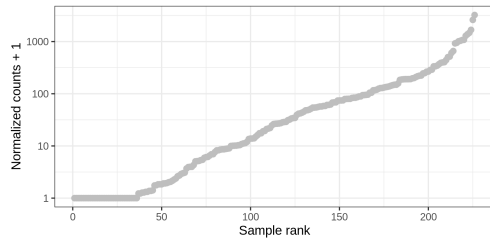
E Expression rank plot: FGFR2



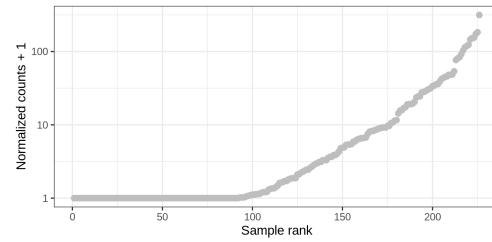
F Expression rank plot: NTRK1



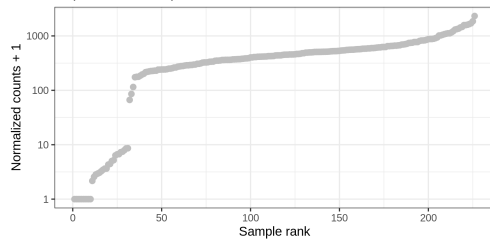
G Expression rank plot: NTRK2



H Expression rank plot: RET



I Expression rank plot: MLH1



J Expression rank plot: MGMT

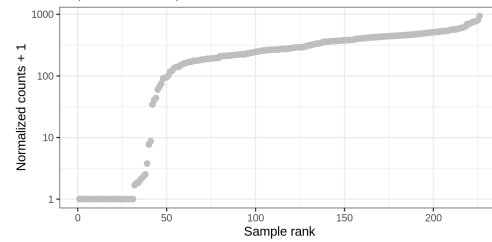


Figure 8: OUTRIDER is outperformed by our approach

(A-B) Barplots depicting the number of genes for which at least a positive (A) or negative (B) outlier sample was detected. Bars obtained by testing different OUTRIDER options (black) are separated from the bar obtained with our approach (white). The annotation bars below the barplots represent the different OUTRIDER options: the number of latent variables is shown in a colour scale from light purple (low number of variables) to dark purple (high number of variables) and the option used for the correction (auto-encoder in blue and PCA in red).

(C-J) Expression rank plots generated by Outlier in RNA-Seq Finder (OUTRIDER) for a selection of relevant genes. In each panel, the controlled counts of a single gene are shown in ascending order. Points corresponding to samples recognized as outliers by OUTRIDER are red. These graphs were obtained with the default OUTRIDER settings (14 latent variables and auto-encoder-based correction)

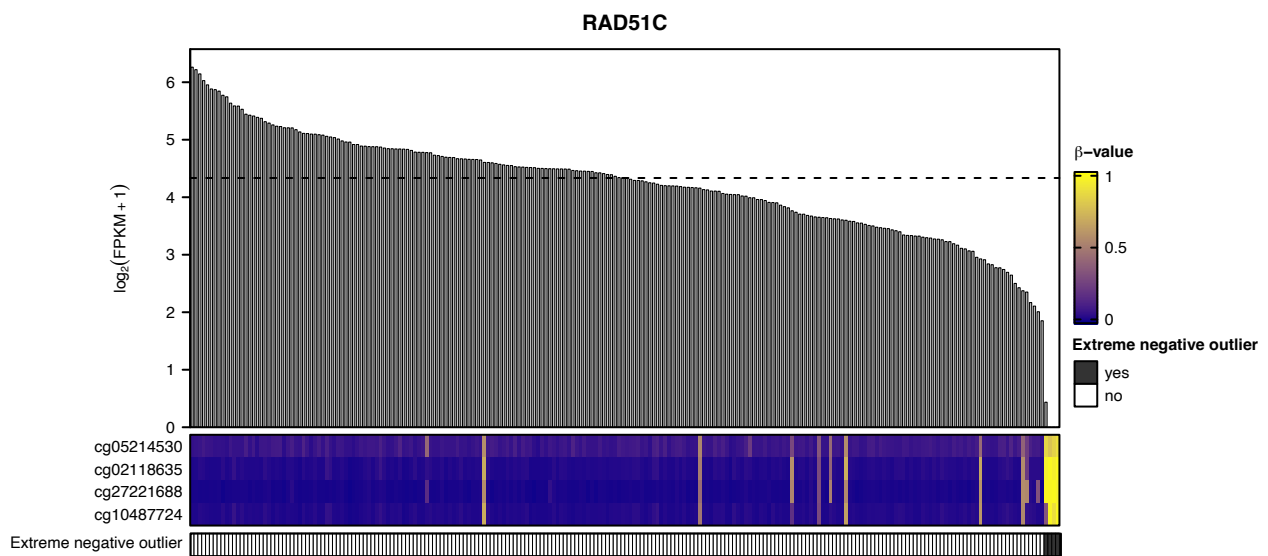


Figure 9: Promoter hypermethylation in *RAD51C* extreme negative outliers.

Heatmap depicting DNA methylation levels of promoter probes that were found differentially methylated between *RAD51C* extreme negative outliers and other samples. DNA methylation levels were measured as β -values and are represented using a colour scale from dark blue (low DNA methylation level) to yellow (high DNA methylation level). *RAD51C* expression profile is shown above the heatmap and sample ordering is the same in the two graphs, from high to low *RAD51C* expression levels. The annotation bar below the heatmap indicates samples that were pinpointed as *RAD51C* extreme negative outliers.

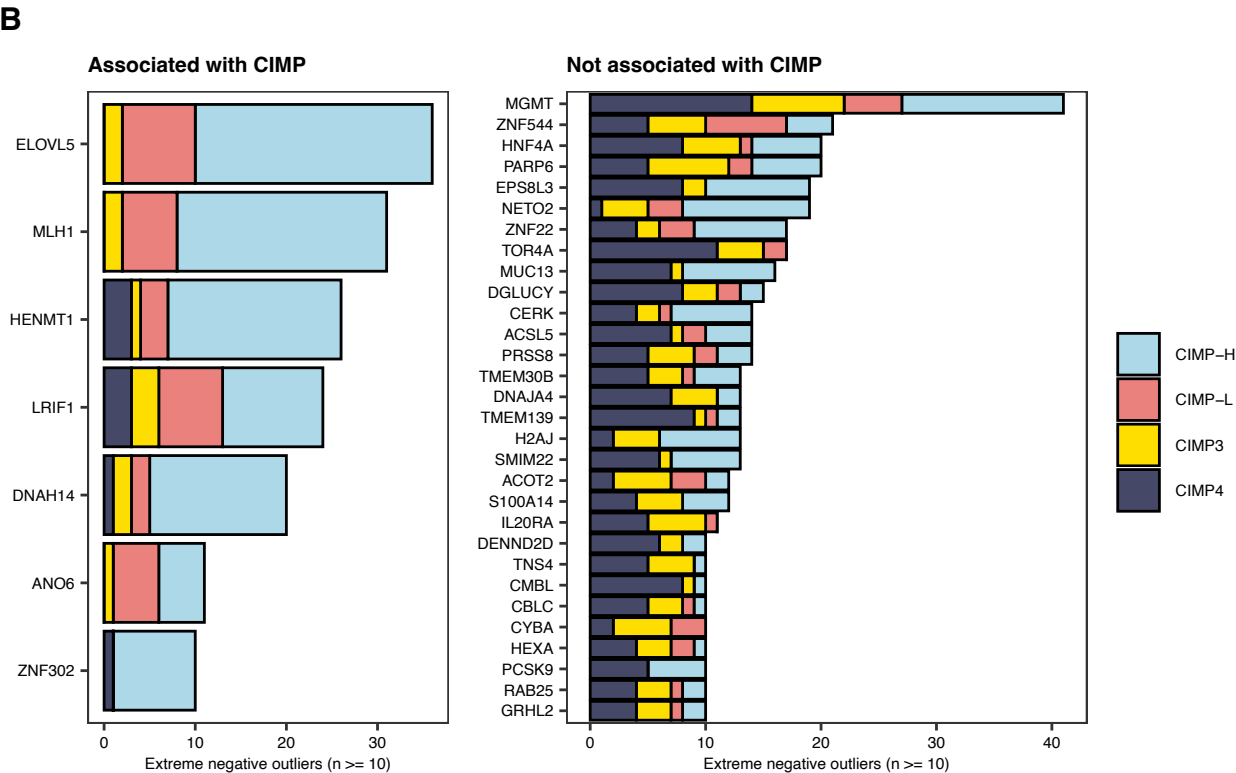
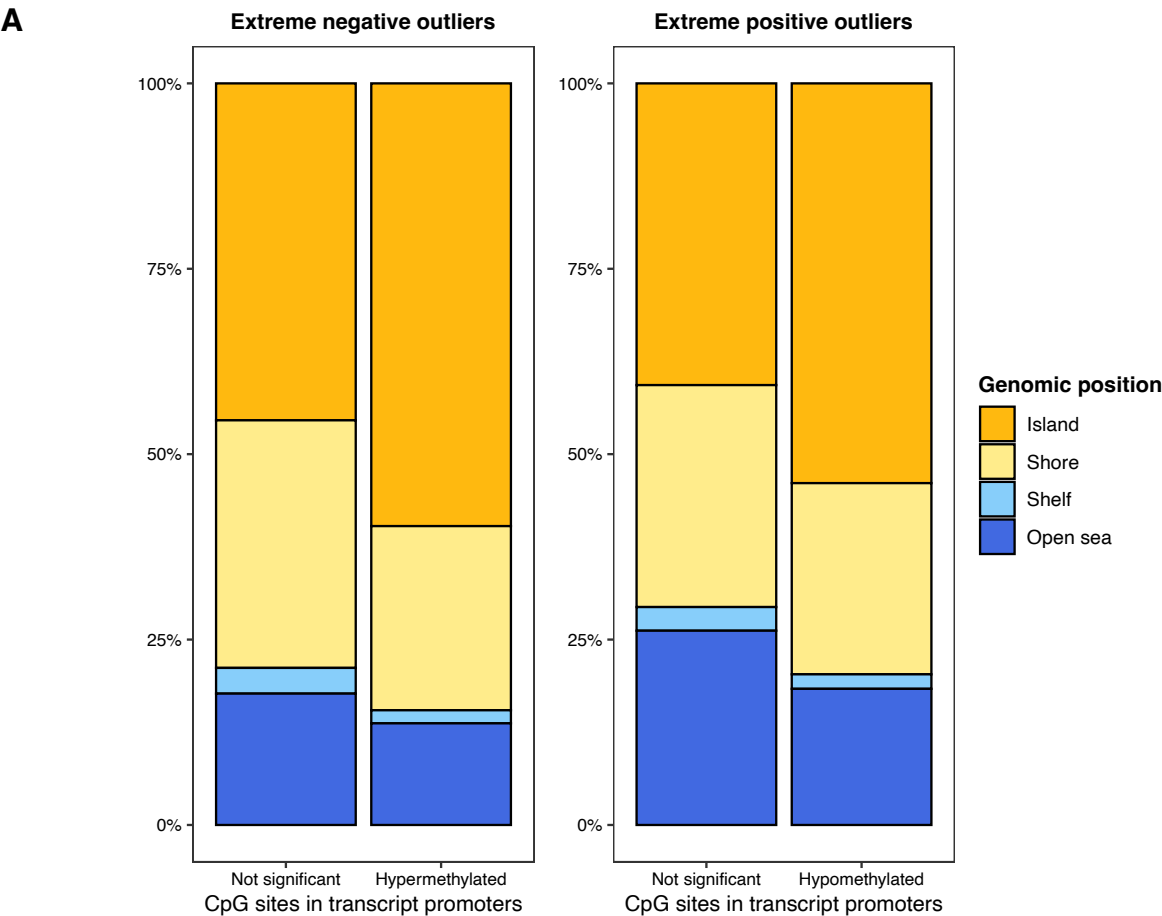


Figure 10: CpG site genomic position and sample CIMP status have an impact on methylation patterns in extreme outliers.

(A) The genomic position of CpG sites within transcript promoters that were tested for differential methylation was determined with respect to CpG islands. Islands, shores, shelves and open sea were defined as detailed in the method section. In the left panel, stacked bar graph depicting the proportion of analysed CpG sites in different genomic regions for both not significant CpG sites and those that are significantly hypermethylated in extreme negative outliers. In the right panel, stacked bar graph depicting the proportion of analysed CpG sites in different genomic regions for both not significant CpG sites and those that are significantly hypomethylated in extreme positive outliers.

(B) Genes whose underexpression events were associated with promoter hypermethylation were classified as CIMP associated (left subpanel) or not CIMP associated (right subpanel) based on the enrichment of CIMP positive samples (CIMP-H and CIMP-L) in extreme negative outliers. Each bar depicts the CIMP classes that were attributed to the extreme negative outliers of a single gene. Only genes for which at least 10 colorectal cancer (CRC) cell lines were identified as extreme negative outliers are shown in the figure.

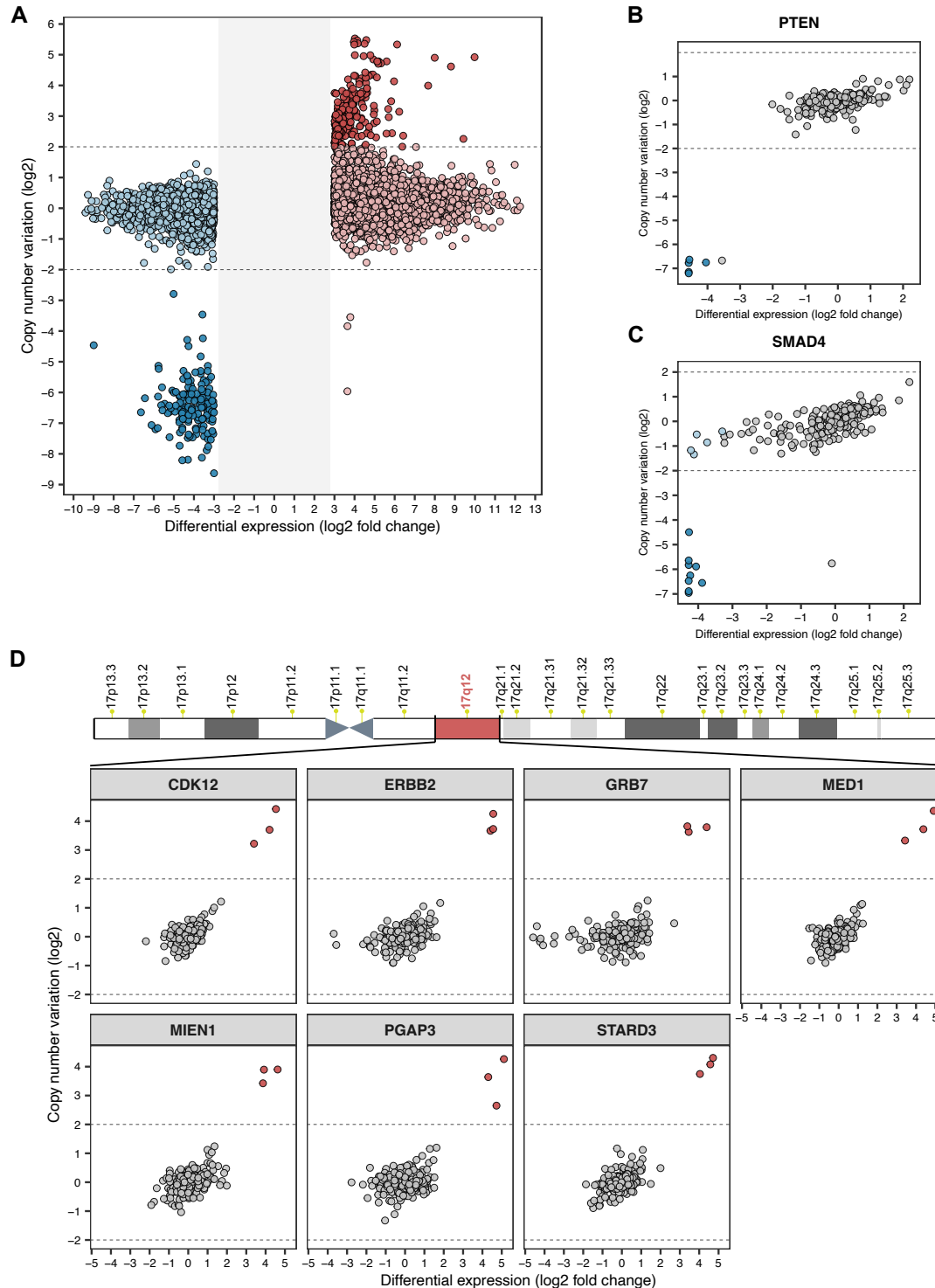


Figure 11: Gene amplifications and deletions leading to overexpression and underexpression events.

(A) Scatter plot of all the extreme positive (red) and negative (blue) gene expression outliers that were identified. The outliers of each gene are defined by differential expression values (x-axis), measured as log2 fold change with respect to the median gene expression, and log2-transformed copy number variation (CNV) values (y-axis). The grey area in the middle corresponds to the interval where extreme positive and negative outliers cannot be located according to the filters that were

applied. Horizontal dashed lines correspond to thresholds used in calling gene amplifications and deletions. Outliers in which either gene overexpression was associated with gene amplification or gene underexpression was associated with gene deletion are shown in darker colours.

(B) Scatter plot of *PTEN* differential expression values, measured as log2 fold change with respect to the median expression, and CNV values in 226 colorectal cancer (CRC) cell lines. Samples in which both *PTEN* underexpression and *PTEN* genetic deletion were found were coloured in dark blue.

(C) Scatter plot of *SMAD4* differential expression values, measured as log2 fold change with respect to the median expression, and CNV values in 226 CRC cell lines. *SMAD4* extreme negative outliers are blue coloured and a darker colour is used for samples in which both *SMAD4* underexpression and *SMAD4* genetic deletion were found.

(D) Scatter plots of differential expression values, defined as log2 fold change with respect to the median gene expression, and log2-transformed CNV values in 226 CRC cell lines for genes that are co-overexpressed and co-amplified with *ERBB2* due to genome proximity. Extreme positive outliers are red coloured. The ideogram of chromosome 17 with G-banding pattern is shown at the top and the genomic region in which the selected genes are located (chromosome 17q12) is highlighted in red.

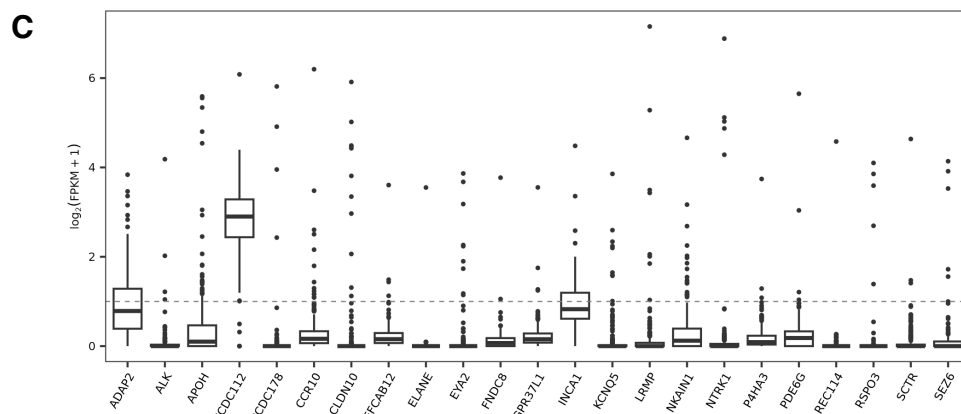
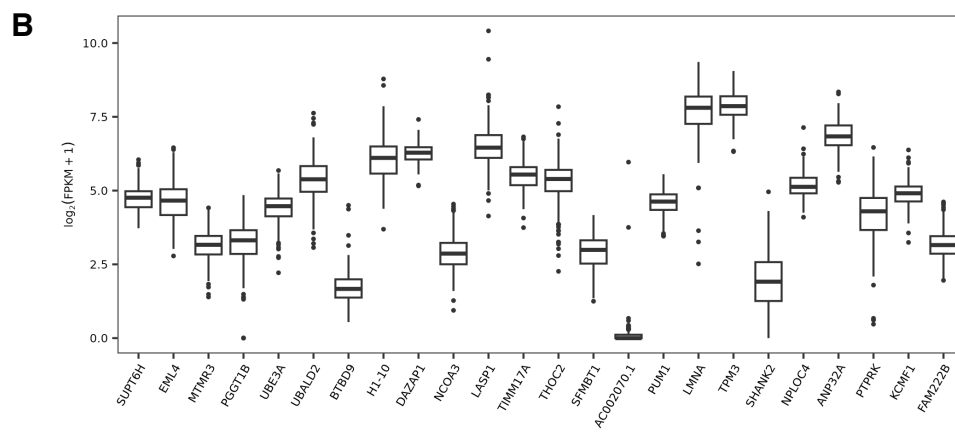
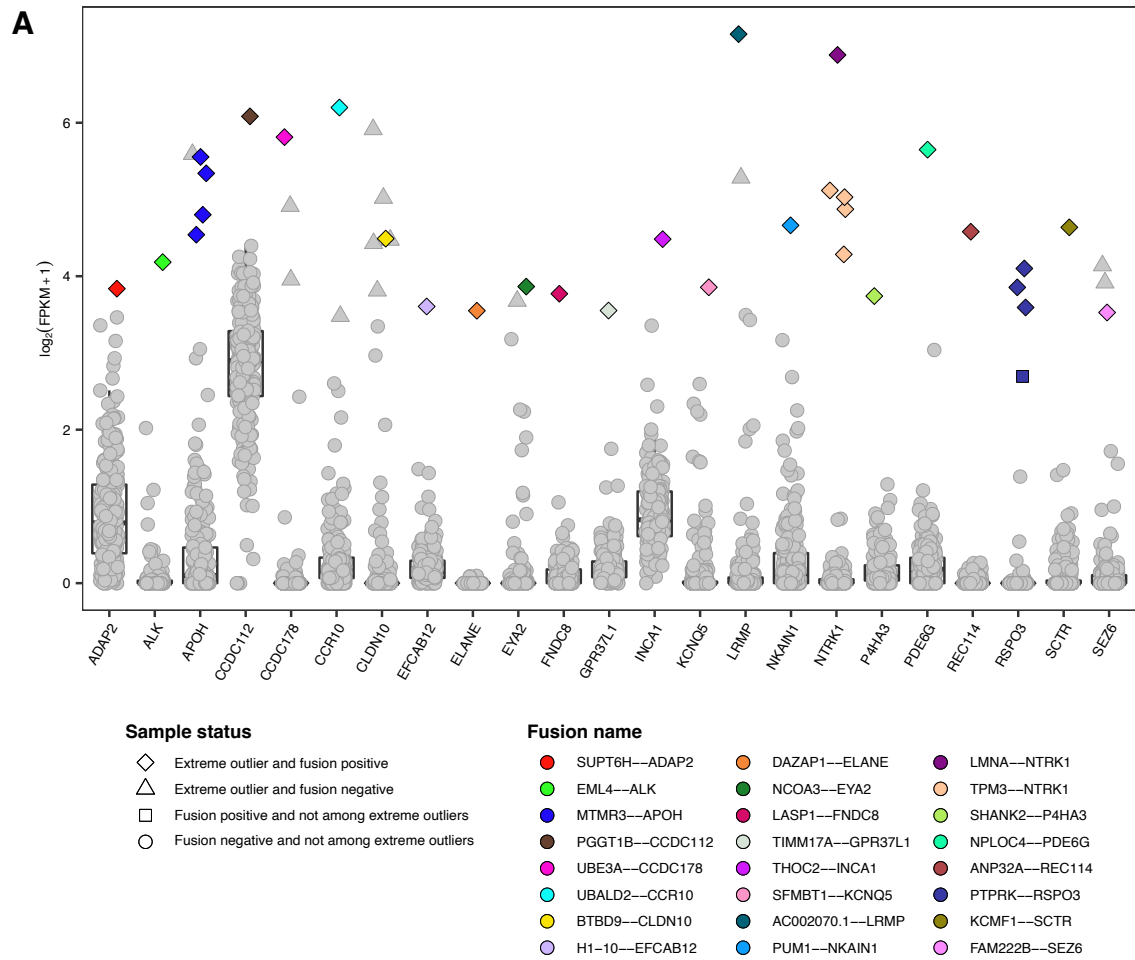
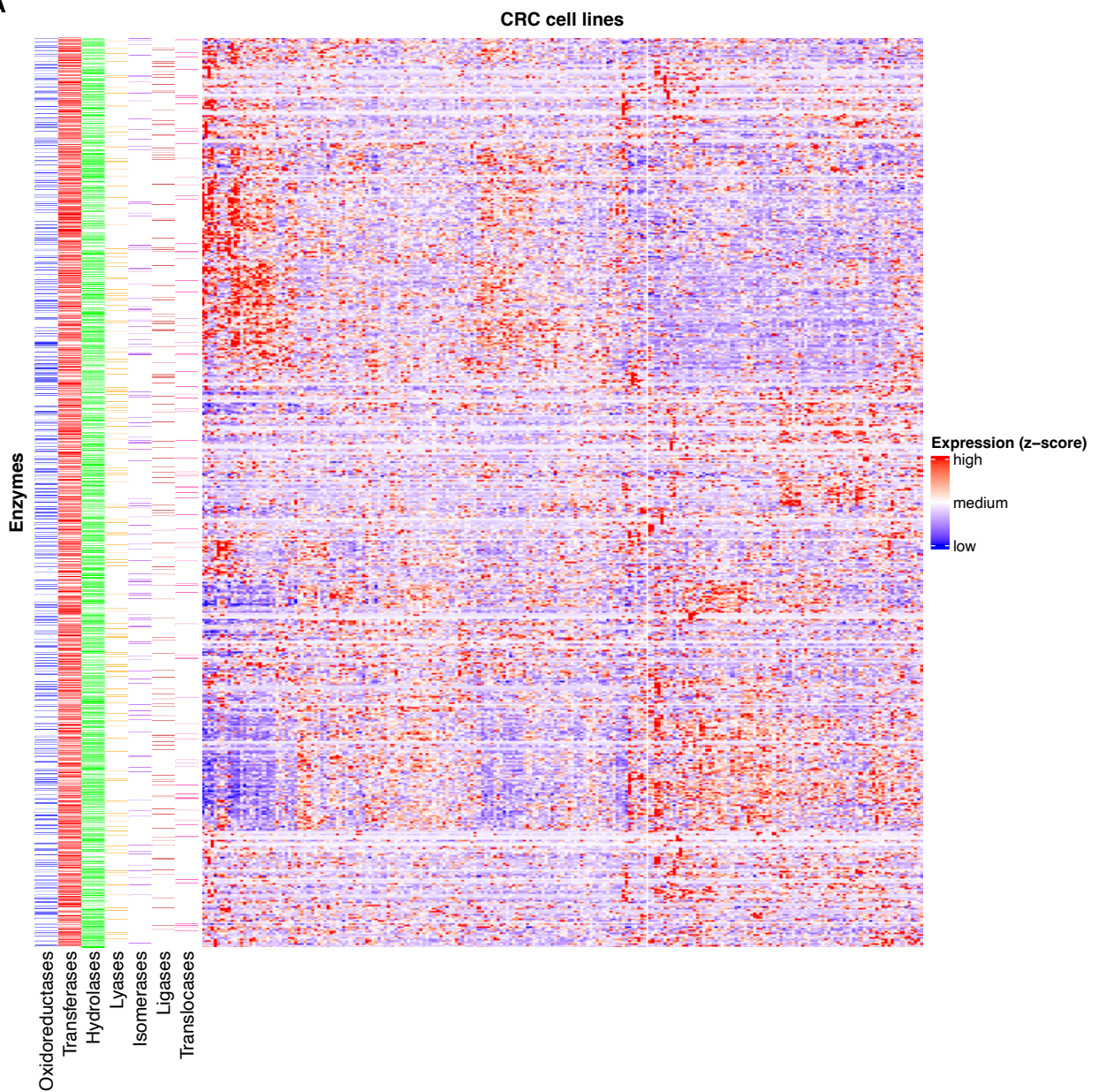


Figure 12: Somatic fusion transcripts associated with the overexpression of 3' partner genes in extreme positive outliers.

(A) Boxplots depicting the distribution of expression values for genes whose overexpression was associated with a somatic fusion transcript in extreme positive outliers. Different point shapes are used to distinguish colorectal cancer (CRC) cell lines based on their gene-specific status. In addition, fusion-positive extreme positive outliers are shown in different colours according to the somatic fusion transcript that was identified.

(B-C) Boxplots depicting the distribution of expression values of genes that are involved as 5' (B) or 3' (C) partner genes in somatic fusions transcripts associated with gene overexpression in extreme positive outliers. The box corresponds to the interquartile range (IQR) and the median is represented by the center line. Whiskers maximally extend to $1.5 \times \text{IQR}$ and outliers are shown. In (C) the dashed grey line corresponds to the threshold used to identify the silenced genes (median FPKM < 1).

A



B

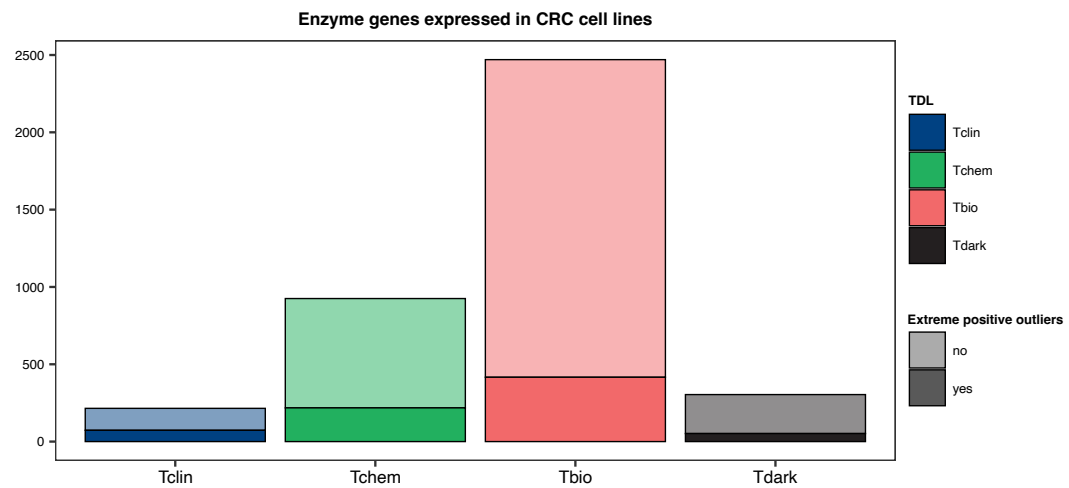
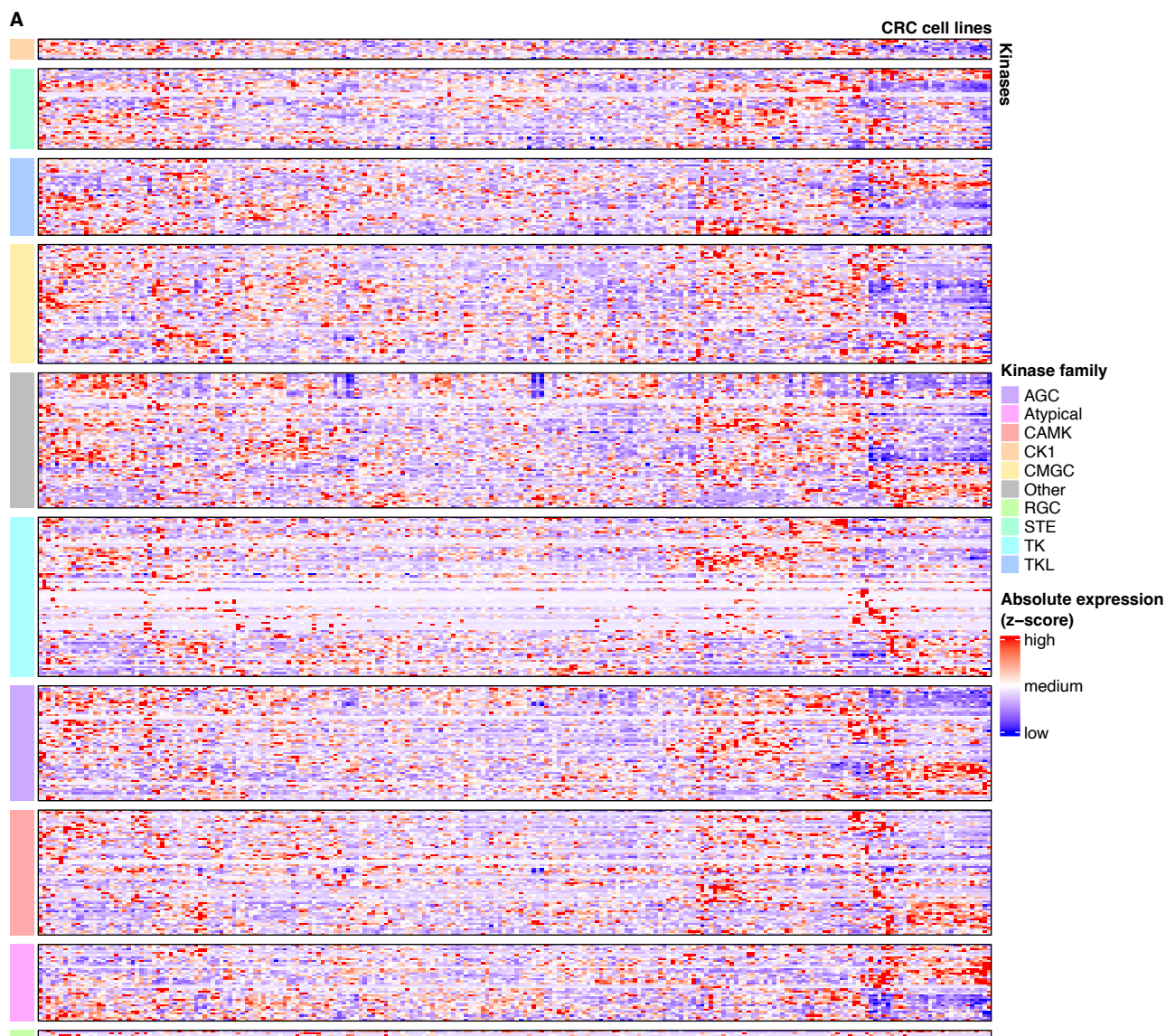


Figure 13: Target Developmental Level (TDL) categories applied to overexpressed enzyme genes.

(A) Heatmap showing absolute expression (z-scores of FPKM values) of 3,704 kinase genes (rows) across 226 colorectal cancer (CRC) cell lines (columns); low expression corresponds to z-score < -2 and colour blue, medium expression to $-2 < \text{z-score} < 2$ and colour shifting to white, high expression to z-score > 2 and colour red. Bars on the left show the top level translation of the Enzyme Commission number (EC number) of each enzyme (EC1=Oxidoreductases, in blue; EC2=Transferases, in red; EC3=Hydrolases, in green; EC4=Lyases, in orange; EC5=Isomerases, in purple; EC6=Ligases, in brown; EC7=Translocases, in pink).

(B) Enzyme genes expressed in our colorectal cancer (CRC) cell line collection were categorized in four classes according to their TDL, from high (Tclin) to low (Tdark) depth of clinical, biological and chemical investigation. Each bar corresponds to the number of expressed enzyme genes for each TDL class and the fraction of genes for which extreme positive outliers were found is highlighted with darker colours.



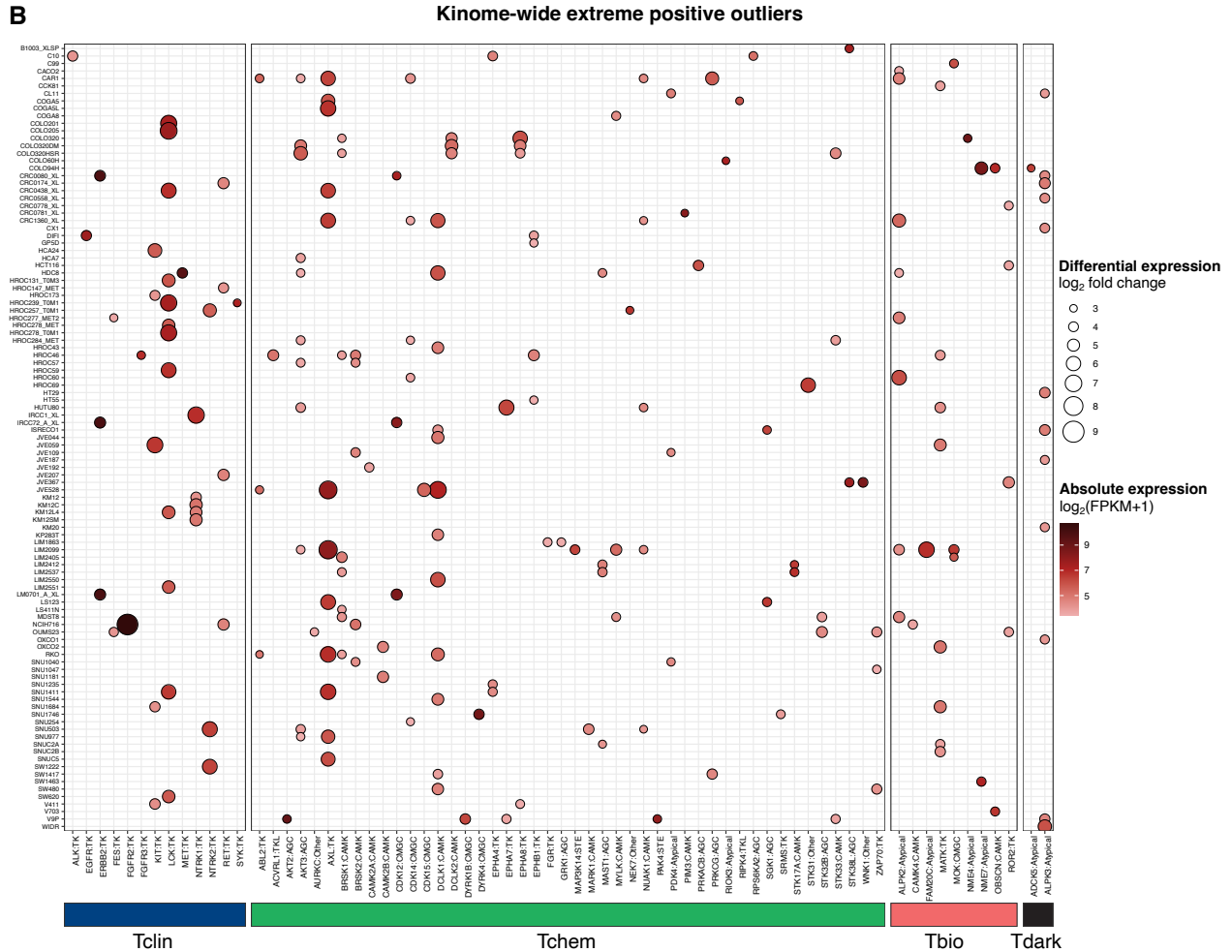


Figure 14: Target Developmental Level (TDL) categories applied to overexpressed kinase genes.

(A) Heatmap showing absolute expression (z-scores of FPKM values) of 558 kinase genes (rows) across 226 colorectal cancer (CRC) cell lines (columns); low expression corresponds to z-score < -2 and colour blue, medium expression to -2 < z-score < 2 and colour shifting to white, high expression to z-score > 2 and colour red. Bars on the left show the ePK (eukaryotic Protein Kinase) group to which each kinase belongs (**AGC**: APKC (Protein kinase A), GPK (Protein kinase G), CPK (Protein kinase C); **CAMK**: Calcium/Calmodulin-dependent Protein Kinases; **CK1**: Casein Kinase 1 family, CMGC: CDK (Cyclin-dependent kinase), MAPK (Mitogen-Activated Protein Kinase), GSK3 (Glycogen Synthase Kinase 3), CDK-like kinases; **TK**: Tyrosine Kinases; **TKL**: Tyrosine Kinase-Like kinases; **STE**: Named after STERile kinases in yeast, includes kinases that are part of the MAPK signaling; **RGC**: Receptor Guanylate Cyclases; **Atypical**: kinases that do not fit into the conventional ePK groups; **Other**: A miscellaneous category for kinases that do not belong to the above families)

(B) Bubble chart of extreme positive outliers identified for kinase genes (kinome-wide extreme positive outliers). The outliers of each kinase gene are defined by absolute gene expression (FPKM value) and differential expression (log₂ fold change with respect to the median gene expression). Overexpressed kinase genes are categorized in the four TDL classes and reported along with the name of the kinase families.

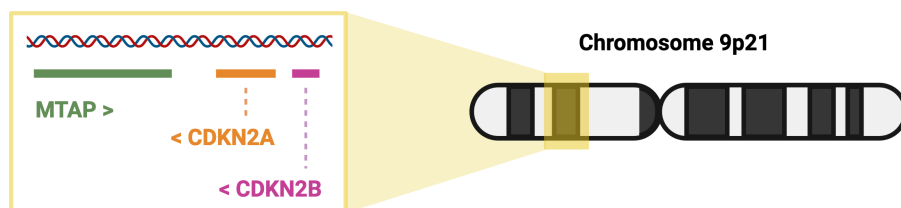
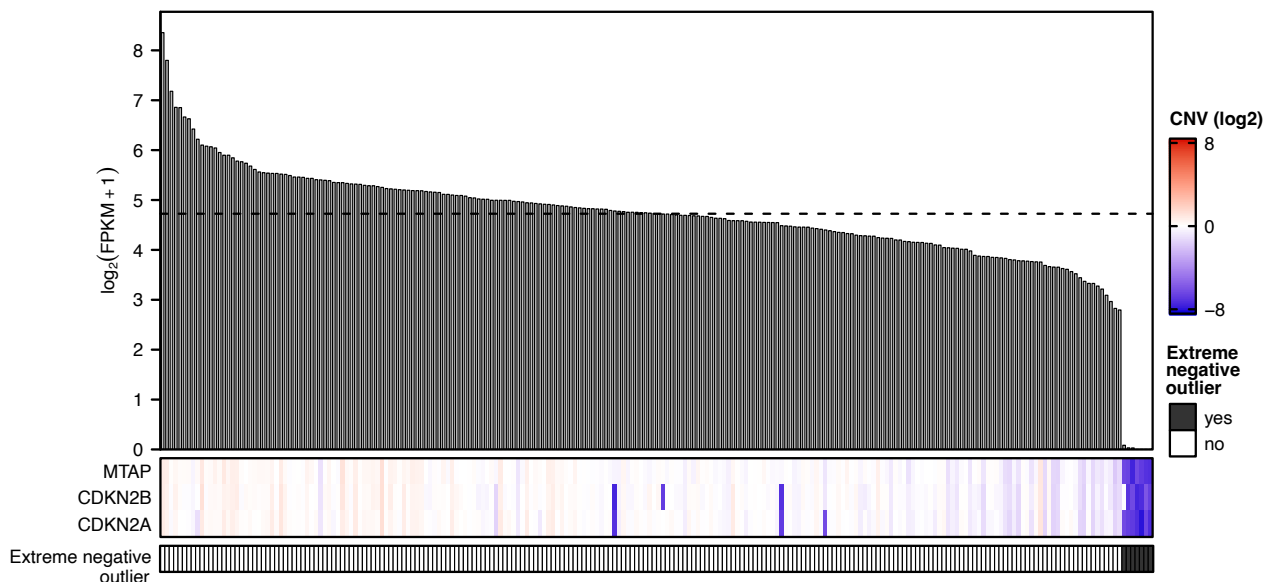
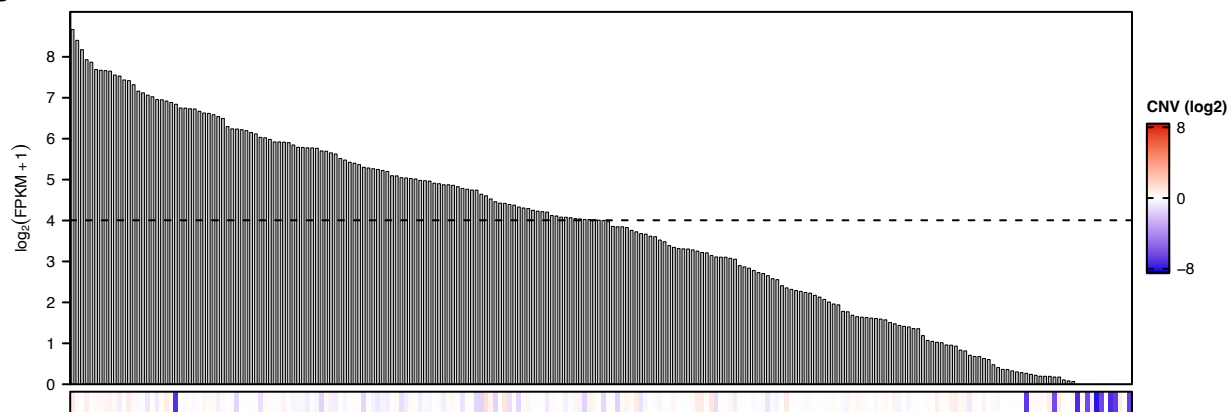
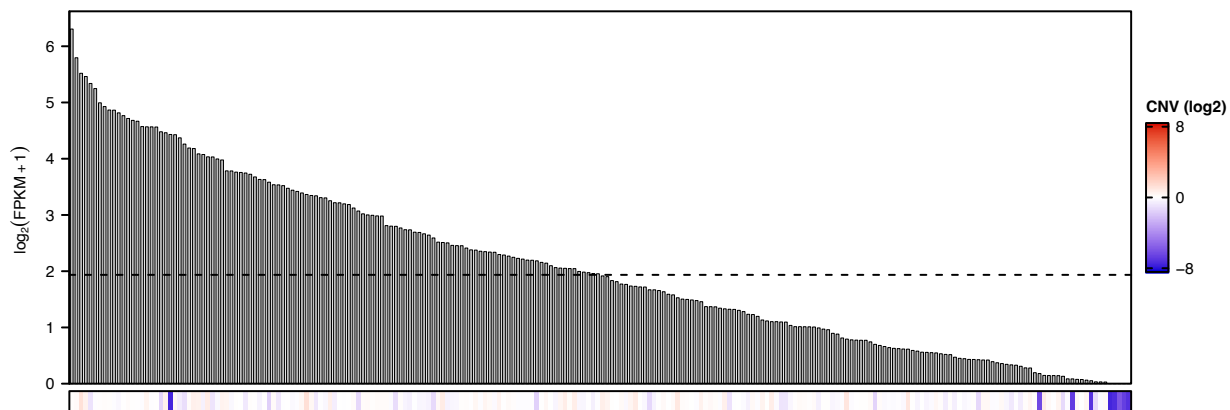
A**B****MTAP****C****CDKN2A****D****CDKN2B**

Figure 15: MTAP, CDKN2A and CDKN2B deletions are associated with very low or zero expression values in colorectal cancer (CRC) cell lines

(A) Graphical representation indicating the position of *MTAP*, *CDKN2A* and *CDKN2B* genes on chromosome 9.

(B) Heatmap showing log2-transformed copy number variation (CNV) values of *MTAP*, *CDKN2A* and *CDKN2B* genes in 226 CRC cell lines. *MTAP* expression profile is shown above the heatmap and sample ordering is the same in the two graphs, from high to low *MTAP* expression levels. The annotation bars below the heatmap indicates samples that were identified as *MTAP* extreme negative outliers.

(C) Heatmap depicting log2-transformed copy number variation (CNV) values of *CDKN2A* in 226 CRC cell lines. *CDKN2A* expression profile is shown above the heatmap and sample ordering is the same in the two graphs, from high to low *CDKN2A* expression levels.

(D) Heatmap depicting log2-transformed copy number variation (CNV) values of *CDKN2B* in 226 CRC cell lines. *CDKN2B* expression profiles are shown above the heatmap and sample ordering is the same in the two graphs, from high to low *CDKN2B* expression levels.

Log2-transformed CNV values are represented using a colour scale from red (gene amplification) to blue (gene deletion) and white indicates copy number neutrality.

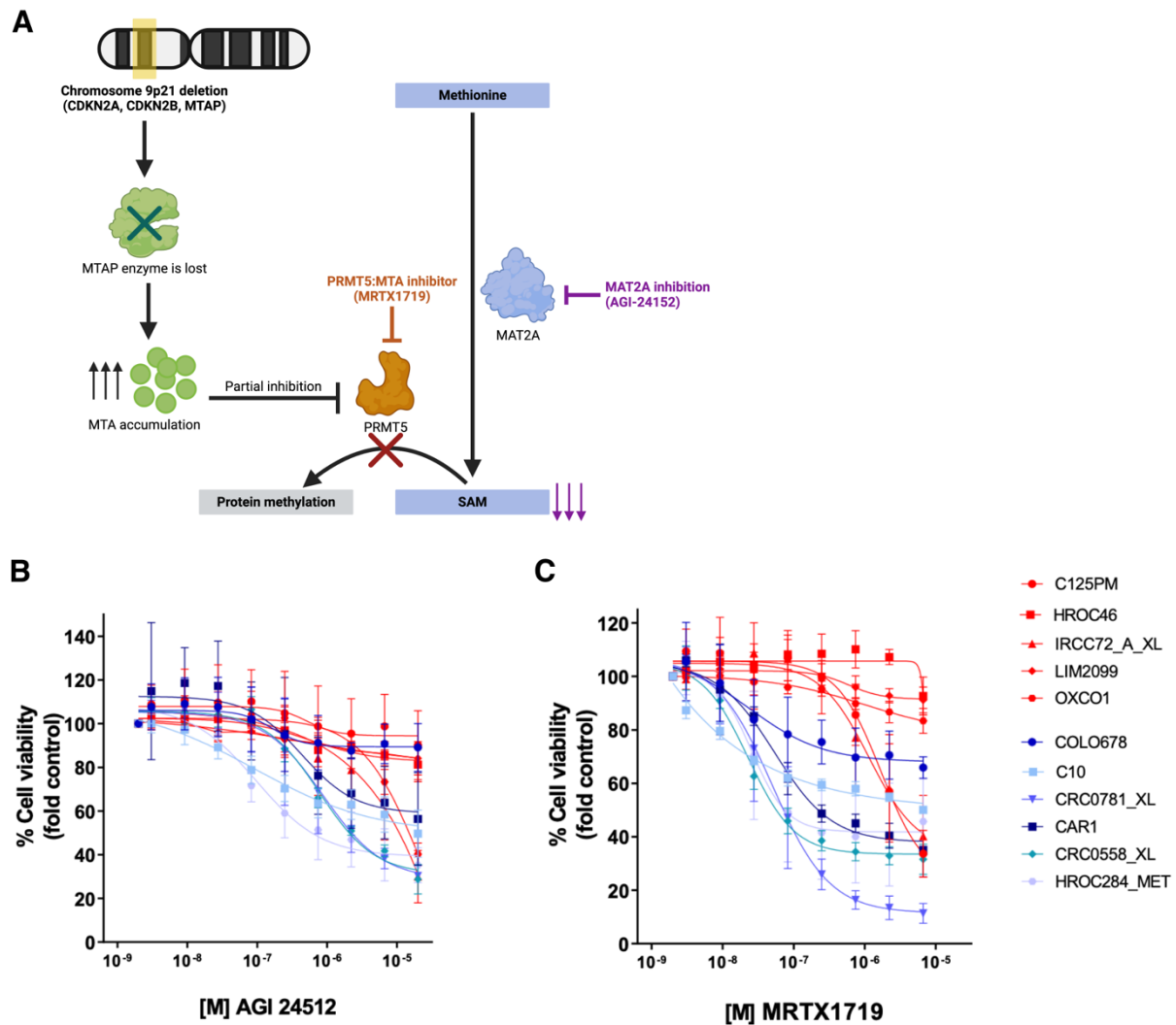


Figure 16: AGI-24512 and MRTX1719 screening in MTAP-deleted and wild-type colorectal cancer

(A) Graphical representations showing the mechanisms of action of AGI-24512 and MRTX1719.

(B-C) Cell viability was assessed by measuring ATP content after 7 days of treatment with increasing concentration of AGI-24512 (B) or MRTX1719 (C). Red lines are negative controls in which the MTAP gene is normally expressed. Cell viability measured for each treatment condition was normalized to viability of DMSO-treated cells.

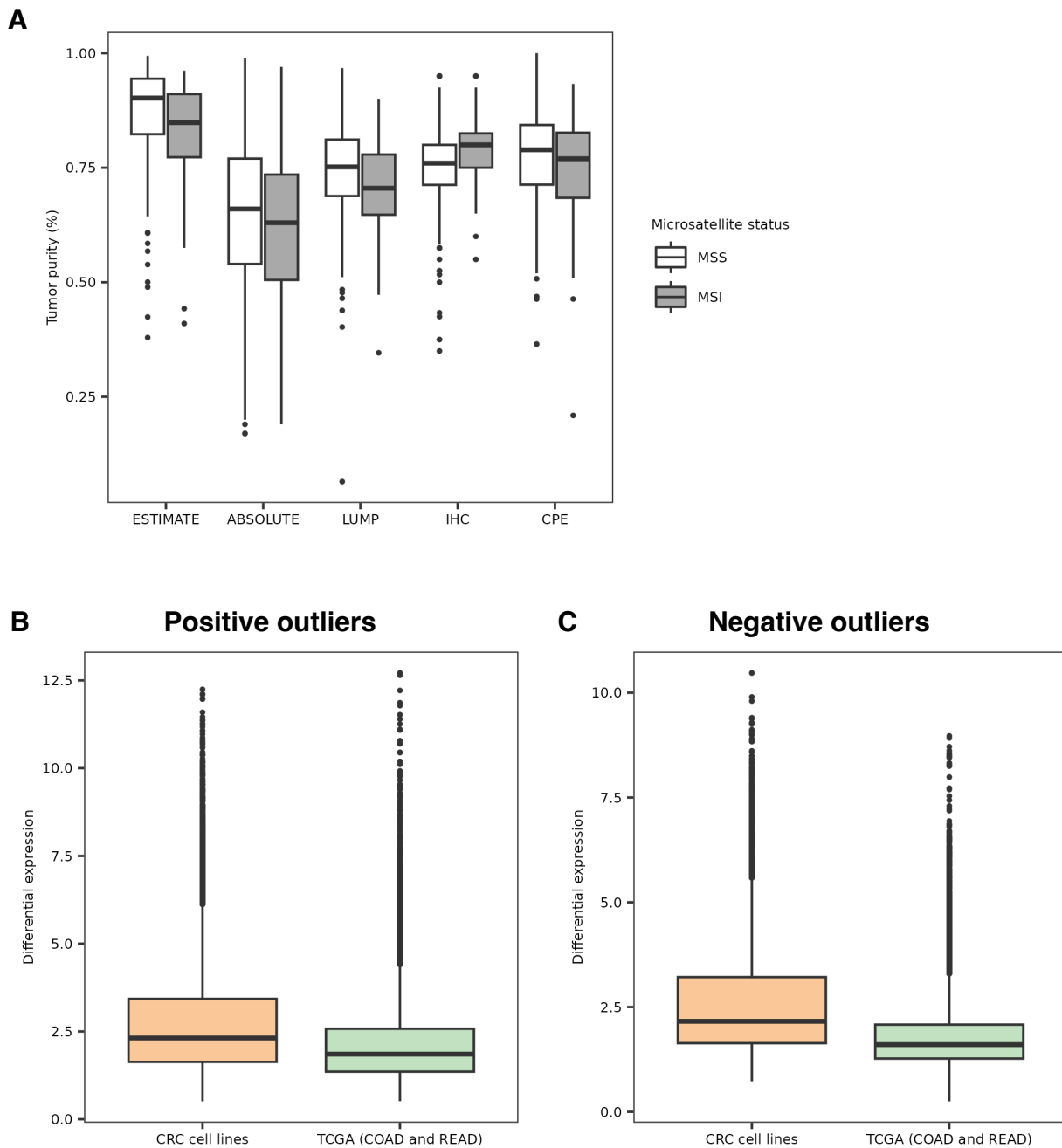


Figure 17: Positive and negative outliers found in The Cancer Genome Atlas (TCGA) dataset are usually less prominent than those found in the colorectal cancer (CRC) cell line dataset

(A) Boxplots depicting the distribution of tumor purity estimates across 611 primary CRC samples from the TCGA. Four methods were considered: ESTIMATE that is based on gene expression signals from immune and stromal genes, ABSOLUTE that exploits copy number data, LUMP (leukocytes unmethylation for purity), IHC and CPE (consensus measurement of purity estimations). Tumor purity values were obtained from Aran et al., Nature Communications (2015). The annotation of the microsatellite status was obtained from Ding et al., Cell (2018).

(B-C) Boxplots depicting the distribution of differential expression values (log2 fold change with respect to the median gene expression) after the selection of the furthest positive (B) or negative (C)

outliers for each gene in the CRC cell line dataset or in the TCGA dataset. The box corresponds to the interquartile range (IQR) and the median is represented by the center line. Whiskers maximally extend to $1.5 \times \text{IQR}$ and outliers are shown.

3. Discussion

Several molecular alterations that have been associated with clinically relevant features of human tumours lead to gene expression alterations, such as *ERBB2* amplification and *MGMT* promoter hypermethylation^{8,9,14}. Furthermore, cancer cells can become highly dependent on dysregulated transcriptional programs independently from the genetic alterations that have induced the gene expression changes¹⁶.

Here we propose a novel computational workflow to identify and prioritize transcriptome-wide gene expression outliers as a way to rapidly find genes whose alterations could give rise to therapeutic vulnerabilities in cancer cells (Fig. 18).

Specifically, our novel strategy goes beyond the simple use of the Tukey's rule for outlier calling¹⁷. We indeed decided to implement a multi-step filtering strategy that does not require the comparison with an independent dataset and was designed to progressively focus on samples with the most extreme deviations in gene expression compared to the overall distribution in the dataset, thus enriching for biologically and clinically relevant outliers (Fig. 18).

In addition, extending the analysis beyond small groups of druggable genes such as kinases and searching both gene overexpression (positive outliers) and gene underexpression (negative outliers) events are crucial improvements with respect to previous studies^{20–22,25}.

Furthermore, it is noteworthy that, while we employed additional multi-omics data to validate and better characterize our findings (Fig. 18), the gene expression outlier analysis relies solely on bulk gene expression data to identify potential vulnerabilities, although they can have different origins at the molecular level. This provides a practical advantage, especially in scenarios where access to different data types might be limited.

Finally, the application of our approach is not restricted to specific tumour types. In this study we decided to focus on CRC for which precision oncology has lagged behind with respect to other solid tumours (Fig. 18)⁷⁴.

Over the last years, we have assembled and annotated a panel of 226 CRC cell lines that well recapitulates the heterogeneity and molecular characteristics that are observed in CRC patients^{20–22}.

The MSI prevalence is usually lower in CRC patients (5-15% depending of the disease stage)⁷⁵ as compared to our collection of CRC cell lines (34%). This discrepancy is most likely related to the higher capacity of MSI cell lines to adapt to in vitro growth conditions probably due to their higher genetic instability⁷⁶.

Notably, for each CRC cell line we have obtained multi-omics data that provides an integrated genetic, epigenetic and transcriptional characterization which was not available for such a large dataset of CRC models.

When RNA-seq expression data were analysed, we found 3,533 and 965 genes with extreme positive and negative gene expression outliers, respectively. Importantly, this analysis correctly identified several gene expression abnormalities of CRC cell lines that have been previously associated with clinically relevant features^{23,47–49}. This observation supports the effectiveness of the novel computational approach we developed. On the contrary, an alternative method for recognizing expression outliers in RNA-seq data (OUTRIDER⁷⁷) failed in detecting many of them.

We chose OUTRIDER for this comparison as it was more aligned with our study's objectives due to its capacity to handle RNA-seq data and systematically identify both overexpression and underexpression events by examining the entire transcriptome. Other tools, such as KAOS (Kinase Automatic Outliers Search) and ZODET (z-score outlier detection), were considered but ultimately not selected. KAOS is less interesting as it is designed to detect genes that are overexpressed in a small fraction of samples within a specific tissue, is focused on kinase genes and its algorithm is based on the assumption that the input values are normally distributed, which cannot be given for granted in the case of log2-transformed RNA-seq expression levels. On the other hand, ZODET was designed to detected outliers based on microarray expression data and requires the comparison with a reference population of healthy individuals to identify outlier cancer samples, which does not fit with the dataset we analysed^{53,78,79}.

Looking for correlations of outlier burden values with known molecular features of the CRC cell lines, we observed a significantly higher positive outlier burden in CMS4 CRC

cell lines than in samples classified in other CMS subtypes. After ruling out the possibility that these values arose from technical artefacts, we discovered that CMS4 CRC cell lines also have a higher proportion of overexpressed genes that are distinctive of the CMS4 subtype. This suggests that the high positive outlier burden that is observed in CMS4 samples likely reflects the activation of distinctive biological processes.

Similarly, we found that CIMP-H CRC cell lines displayed a higher negative outlier burden than samples classified within other DNA methylation-based classes. In addition, CIMP-H samples are characterized by a higher fraction of underexpressed genes associated with promoter hypermethylation, supporting the idea that high values of negative outlier burden in these cells are the consequence of widespread DNA methylation.

Despite these results may indicate that the targets identified by our approach could be biased towards specific subsets of the CRC population (those for which highest outlier burden values were observed), we would like to highlight that every CRC cell line in the dataset was identified as an outlier for at least one gene. Furthermore, the observation that the median number of overexpressed genes per model was 24, and the median number of underexpressed genes per model was 7, further supports the idea that the proposed strategy is effective in identifying therapeutic vulnerabilities of CRC cells, without being limited by the transcriptional or epigenetic status of the models.

We then exploited the multi-omics data to search genetic and epigenetic alterations associated with extreme expression values (Fig. 18).

Specifically, we have considered alterations of DNA methylation levels at promoter regions, gene amplifications, gene deletions and the presence of fusion transcripts. Different genomic rearrangement-independent mechanisms were reported to describe the origin of fusion transcripts⁸⁰. However, given the gene expression profiles and their position in the genome, it is plausible that genomic translocations account for the majority of the fusion transcripts that we identified. This is also in line with a previous pan-cancer study⁸¹.

Overall, about 6% and 3% genes for which extreme positive outliers were found were connected with a genetic or epigenetic alteration, respectively. In addition, about 11% genes for which underexpression events were identified were linked to a genetic

alteration and about 21% of them were instead associated with an epigenetic alteration. Thus, for 92% and 69% of the genes for which extreme positive or negative outliers were found, respectively, neither genetic alterations nor epigenetic abnormalities could be identified to explain the overexpression or underexpression. One potential explanation, for some overexpressed genes, may be the existence of other chromosomal translocations, that may have remained undetected due to technical limitations. In the absence of whole genome sequencing (WGS) data, the search was limited to fusion transcripts in the available RNA-seq data; moreover, single-end RNA-seq data are suboptimal even for the identification of fusion transcripts^{82,83}.

For example, another form of chromosomal rearrangements that results in the overexpression of one gene involves the fusion of its coding sequence with solely the promoter or enhancer elements of an active gene¹⁹. From the transcriptional point of view, this type of events would be characterized by the isolated overexpression of the gene; this stands in contrast to non-focal amplifications (the most prevalent type of amplifications), where a concomitant overexpression of the nearby genes is expected. In principle, it may be feasible to differentiate between these two genetic aberrations on the basis of expression data. However, we reasoned that this type of predictive classification would not be straightforward, as it is difficult to discern whether nearby genes are overexpressed as well, given that they must be expressed in colon in the first place.

Furthermore, it was previously shown that chromosomal translocations in leukaemia are frequently associated with focal deletions in the vicinity of the breakpoints^{84,85}. It is to be noted that some of the genes that we have designated as focal deleted, may potentially delineate the breakpoint of a translocation.

Moreover, further associations between additional epigenetic features of CRC cell lines and outlier expression values could be identified if data other than DNA methylation would be available⁸⁶.

Nonetheless, our results prove that looking for transcriptome-wide gene expression outliers is an effective strategy to discover features of cancer cells that may have a different origin. They also indicate that many extreme expression values of unknown

functional relevance may not be easily predicted without directly parsing gene expression data.

We acknowledge that the approach that we propose has some limitations. For example, although numerous studies have shown that protein levels are primarily determined by transcript concentrations in steady-state conditions⁸⁷, the translation of extremely high or low gene expression levels into abnormal protein abundance cannot be given for granted due the complexity of gene expression regulation and the influence of other biological processes. Despite this, several studies demonstrated that the “extreme” expression outliers should in most instances lead to abnormal protein expression^{20,22,88,89}.

By mining our atlas of CRC expression outliers, novel findings have emerged.

Firstly, we were interested in identifying new drug targets. We therefore focused on enzymes because their abnormal activity in cancer cells can play a role in escaping from normal physiological constraints on growth and survival. Moreover, enzymes are ideally suited for therapeutic inhibition and many inhibitors have been developed. By therefore annotating overexpressed enzyme genes with the corresponding Target Development Level (TDL)⁹⁰, we pinpointed several proteins whose activity is expected to be dysregulated in CRC cells and for which drugs or active ligands are already available, including several kinases within both the TK subgroup and other families (Fig. 18).

On the other hand, the lack of activity of an enzyme, due to its loss of expression or other mechanisms, may give rise to a synthetic lethal vulnerability, *i.e.*, the inhibition of another protein is lethal only in that particular scenario. A typical example of the “synthetic lethality” concept is the presence of loss-of-function mutations in the *BRCA2* gene, that causes homologous recombination deficiency and thus create a dependency on the non-homologous-end-joining pathway, leading to sensitivity to PARP inhibitors.

Therefore, we reasoned that the exploration of the underexpression events that we found in CRC cell lines can suggest novel synthetic lethal targets.

As a proof of concept, we focused on CRC cell lines recognized as extreme negative outliers for the *MTAP* gene that is co-deleted with the *CDKN2A/CDKN2B* tumour suppressor genes. In our CRC cell line collection, the prevalence of MTAP loss (3%) is

in line with that seen in mCRC patients for other molecular alterations that are targeted by clinically relevant treatments⁷⁴.

We showed that two drugs (AGI-24512 and MRTX1719), whose effect is synthetic lethal with the MTAP enzyme loss, selectively impaired cell viability of the CRC cell lines pinpointed as extreme negative outliers for the *MTAP* gene. The most effective results were obtained with the second-generation inhibitor MRTX1719, which is currently being tested in a phase 1/2 clinical trial in patients with advanced solid tumours with homozygous genetic deletion of the *MTAP* gene^{91,92}.

This class of drugs has been developed to overcome the haematological toxicities seen in patients treated with first-generation PRMT5 inhibitors, that act on both *MTAP*-deleted and WT cells simultaneously. Instead, second-generation inhibitors directly target the PRMT5 in the presence of MTA, thus showing enhanced tolerability in patients⁹³.

A new second-generation PRMT5:MTA inhibitor, AMG-193, has recently been developed with the objective of enhancing the inhibitory potency. It is currently being tested in a phase 1/2 clinical trial (NCT05094336) in patients with *MTAP*-deleted solid tumours, in whom it is showing an enhanced safety and tolerability⁹³. It would be of interest to evaluate this promising inhibitor, in addition to those previously discussed, in our *MTAP*-deleted CRC cell lines.

A further avenue of interest would be to test the consequences of induced DNA damage, cell cycle arrest and increase in alternative mRNA splicing in our CRC cell lines due to the complete loss of PRMT5 function in the presence of MTAP loss, as demonstrated in the AMG-193 in vitro mechanism-of-action tests⁹³.

Many studies have demonstrated that cancer cell lines recapitulate the molecular features observed in human tumours, including gene expression and pharmacological profiles^{20,22,94,95}. However, we cannot exclude that some extreme gene expression outliers are artefacts due to adaptation to in vitro growth conditions.

Therefore, we have verified how many expression abnormalities that we identified in CRC cell lines also occur in colorectal tumours by taking advantage of expression data from 611 TCGA samples.

This required adapting the thresholds used in the selection of the extreme outliers to deal with the observation that the deviation of the furthest positive or negative

outliers of each gene from the other samples is usually less prominent in the TCGA dataset than in the CRC cell line dataset. This likely occurs because gene expression values obtained through bulk RNA-seq only reflect the average of signals from multiple cell types when tumour tissue specimens are analysed.

It should also be noted that two databases encompass samples derived from different stages of CRC progression: CRC cell lines were derived from both primary and metastatic tumours, while only primary tumours were analysed in the TCGA dataset. Despite this, we report that about half of the genes for which extreme positive or negative outliers were found in CRC cell lines are also overexpressed or underexpressed in human tumours. This, together with the capability of our analyses to pinpoint genes whose expression status has relevance in the clinics such as oncogenes that are known therapeutic targets and tumor suppressor genes with diagnostic or prognostic value, provides another validation of the clinical relevance of our atlas of CRC gene expression outliers.

In conclusions, we have devised a novel computational workflow in which a transcriptome-wide analysis is performed to find extreme gene expression outliers displaying the overexpression or the underexpression of single genes. Subsequent characterization of dysregulated genes can be used to guide the discovery of novel drug targets and biomarkers in cancer cells. Specifically, in this study we present an atlas of CRC extreme expression outliers whose further exploration may lead to the identification of new actionable targets. The approach can be also exploited to prioritize candidates for drug development in other tumour types.

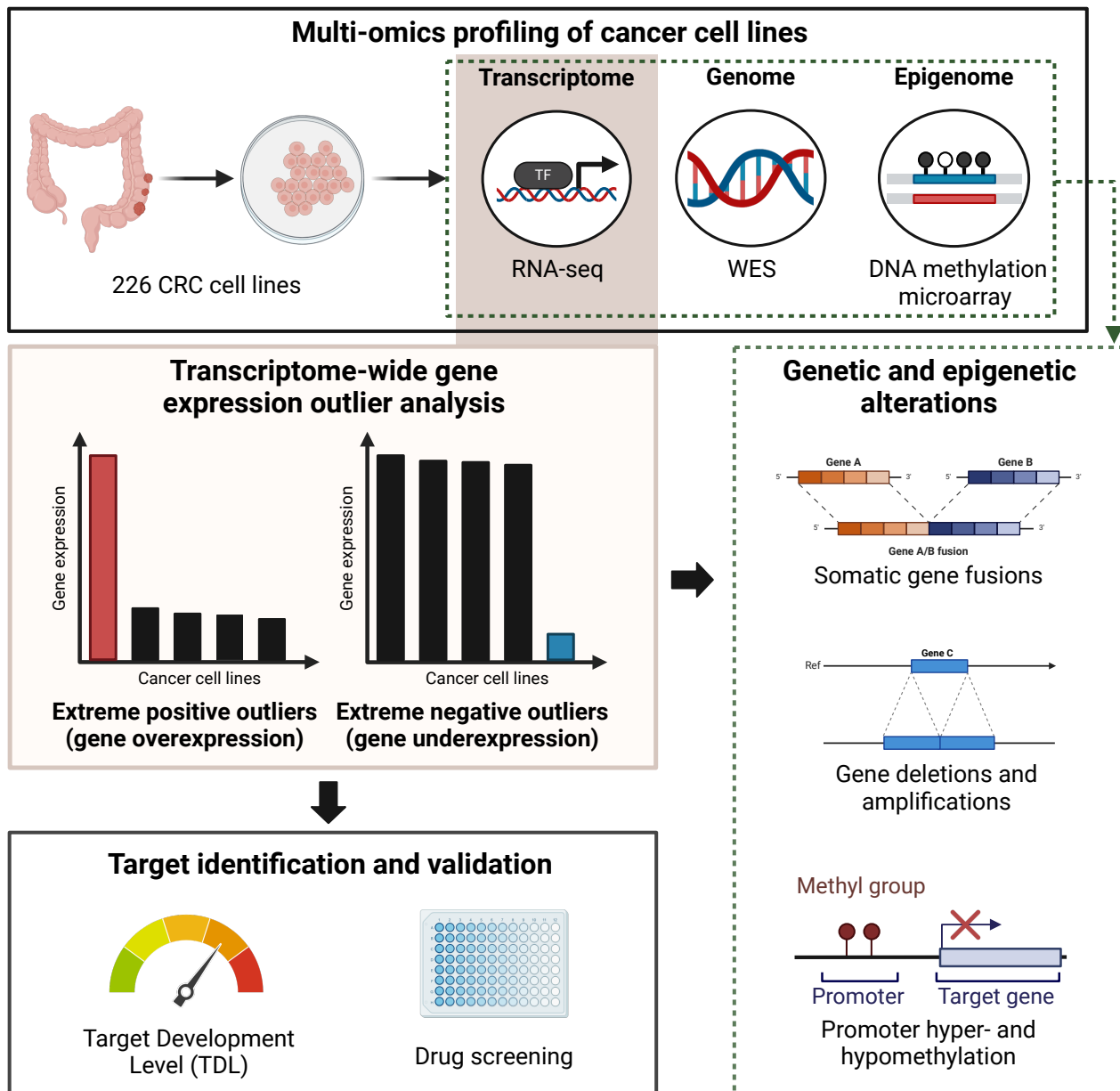


Figure 18: Graphical representation of the thesis workflow

Graphical representation showing the multi-layer characterization of 226 CRC cell lines, that led to the identification of gene expression outliers, some of which were validated with having genetic and epigenetic alterations, and the identification and validation of putative targets.

4. Materials and Methods

4.1 Cell cultures

Cell cultures were supplemented with FBS 10% 2mM L-glutamine, antibiotics (100U/mL penicillin and 100 mg/mL streptomycin) and grown in a 37°C and 5% CO₂ air incubator. Cell cultures were routinely screened for absence of Mycoplasma contamination using the Venor® GeM Classic kit (Minerva biolabs). The identity of each cell line was checked no more than three months before performing the experiments using the PowerPlex® 16 HS System (Promega), through Short Tandem Repeats (STR) tests at 16 different loci (D5S818, D13S317, D7S820, D16S539, D21S11, vWA, TH01, TPOX, CSF1PO, D18S51, D3S1358, D8S1179, FGA, Penta D, Penta E, and amelogenin). Amplicons from multiplex PCRs were separated by capillary electrophoresis (3730 DNA Analyzer, Applied Biosystems) and analysed using GeneMapper v.3.7 software (Life Technologies).

4.2 Microsatellite status

The microsatellite status of 224 out of 226 CRC cell lines was assessed with the MSI Analysis System kit (Promega). The analysis requires a multiplex amplification of seven markers including five mononucleotide repeat markers (BAT-25, BAT-26, NR-21, NR-24, and MONO-27) and two pentanucleotide repeat markers (Penta C and Penta D). The amplification products were analysed by capillary electrophoresis in a single injection (3730 DNA Analyzer, ABI capillary electrophoresis system (Applied Biosystems)). Then, the results were analysed using GeneMapper V5.0 software. The microsatellite status of the remaining 2 CRC cell lines (IRCC22_XL and HROC222_T1_M2) was determined based on whole exome sequencing (WES) data using MSIsensor-pro⁹⁶. The microsatellite status of each CRC cell line is reported in Table 1 as MSS (microsatellite stability) or MSI (microsatellite instability).

4.3 Genomic DNA extraction and sequencing

WES data was available for 149 out of 226 CRC cell lines from previous studies. WES data was therefore obtained for the remaining 77 CRC cell lines. In addition, from a previous study we retrieved WES data from matched normal samples obtained from the same patients from which 25 out of 226 CRC cell lines were derived²². Genomic DNA (gDNA) was extracted from both CRC cell lines and normal samples using Maxwell® RSC Blood DNA kit (AS1400, Promega). Library preparation, sequencing and data demultiplexing were outsourced. Final DNA libraries were sequenced as paired-end 100 bp reads.

4.4 RNA extraction and sequencing

RNA sequencing (RNA-seq) data was available for 140 out of 226 CRC cell lines from previous studies. RNA-seq data was therefore obtained for the remaining 86 CRC cell lines. Total RNA was extracted using Maxwell® RSC miRNA Tissue Kit (AS1460, Promega). The quantification of RNA was performed by Thermo Scientific Nanodrop 1000 (Agilent) and Qubit 3.0 Fluorometer (Life Technologies). RNA integrity was evaluated with the Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano Kit. Total RNA (800 ng) with RNA integrity number (RIN) score between 9 and 10 was used as input to the Illumina TruSeq RNA Sample Prep Kit v2-Set B (48Rxn), according to the manufacturer's protocol. After library preparation, sequencing was performed on NextSeq500 to get single-end 150 bp reads.

4.5 DNA methylation microarray experiments

DNA methylation data was previously obtained for 146 CRC cell lines using the Infinium HumanMethylation450 BeadChip array⁹⁷. The DNA methylation profile of 80 additional CRC cell lines was obtained for this work with the Infinium MethylationEPIC BeadChip microarray, which allowed for over 850,000 methylation sites to be quantitatively interrogated across the genome at single-nucleotide resolution. Preparation and processing of DNA samples were carried out using the Illumina Infinium HD Assay

Methylation Protocol Guide. Briefly, DNA samples (500 ng) were treated with sodium bisulfite using the Zymo EZ-96 DNA Methylation-Lightning Kit (Zymo Research, CA, USA). The bisulfite-converted DNA was used for amplification, fragmentation, precipitation, resuspension, and hybridization on BeadChips. Next, single-base extension and staining were performed following the Infinium HD Methylation automated workflow. Finally, the IScan software analysis system (Illumina, USA) was used for signal acquisition and analysis.

4.6 Identification of somatic mutations

WES reads were aligned to human genome version 38 (hg38) using BWA-mem algorithm⁹⁸. PCR and optical duplicates were annotated and ignored in the subsequent processing. Then, different modules previously developed by our group were exploited for identification of single nucleotide variants (SNVs) and indels (IDEA workflow)⁹⁹. Bases with Phred quality score lower than 20 (Q20) were discarded. In addition, only SNVs supported by at least 10 altered reads and whose mutant allele frequency was higher than 10% were selected. Furthermore, common SNPs from dbSNP version 151 were used to exclude germinal variants and synonymous mutations were ignored in the determination of gene mutational status. Similarly, only indels supported by at least 10 altered reads and whose mutant allele frequency was higher than 10% were considered. KRAS (G12, G13, A59, Q61, K117 and A146), NRAS (G12, G13 and Q61) and BRAF (V600E) hotspots were considered in the annotation of the mutational status of these genes. In parallel, hotspot mutations in TP53 (R175, G245, R248, R249, R273 and R282), nonsense mutations and frameshift indels were considered as mutations that likely determine protein loss of function when annotating TP53 and APC mutational status.

4.7 Copy number analysis

We adapted a previously described strategy to deal with the fact that WES data from matched normal samples are not available for the majority of CRC cell lines⁹⁹. In each CRC cell line, gene copy number (CN) values were computed as the ratio between the

median read depth of each gene and the median read depth of all genes using GENCODE v33 as gene annotation. A principal component analysis (PCA) of gene CN values revealed three different sample clusters that were confirmed through the application of k-means clustering on the first 10 principal components. Based on these findings, a “metanormal” was created for each sample cluster taking for each gene the median CN value across samples and genes whose median CN value was lower than 0.1 in either one of the clusters were excluded from subsequent analysis (genes not captured by WES probes). Then, copy number variation (CNV) values were computed for every gene in every sample as the ratio between the gene CN value in the sample and the gene CN value in the corresponding metanormal. This corrects for systematic differences in sequencing depth among genes reflecting variability in the affinity of WES probes.

To validate the procedure, we exploited 25 CRC cell lines for which WES data from matched normal samples are available²². Importantly, when we compared log2-transformed CNV values obtained using the matched normal or the metanormal we usually found a high correlation (the median Pearson correlation coefficient was 0.78). Previously, in tumour samples, we considered the gene copy number as altered when the log2-transformed CNV values were higher than 1 or lower than -1^{100} . Here, more stringent thresholds were used to select only gene amplifications or homozygous deletions. More in detail, a gene amplification was called when the log2-transformed CNV value was higher than 2, while a gene deletion was identified in samples whose log2-transformed CNV value was lower than -2.

4.8 Gene expression values and molecular subtypes

RNA-seq reads were aligned to hg38 using the splice-aware MapSplice aligner¹⁰¹. Output BAM files were processed using `ubu sam-xlate` and `sam-filter` to translate genomic coordinates into transcriptomic ones and remove reads with indels, large inserts and zero mapping quality before proceeding with transcript and gene quantification using RSEM¹⁰² and GENCODE v33 as gene annotation. Then, starting from RSEM genes results (expected counts and effective lengths) we computed robust FPKM values exploiting the `tximport` R Bioconductor package and the `fpkm` function

included in the DESeq2 R Bioconductor package¹⁰³. The resulting gene expression matrix was subsequently annotated with gene names from the GENCODE annotation file and filtered by applying the following criteria: (a) genes on chrM or chrY pseudoautosomal regions were removed; (b) genes with robust FPKM < 1 in all analysed samples were considered not expressed and removed; (c) only protein coding genes, as defined by the GENCODE annotation file, were selected for subsequent analysis.

In order to check RNA-seq data quality with respect to the presence of CRC cell lines derived from the same individual (same STR profile), hierarchical clustering as implemented by the `hclust` R function was applied to log2-transformed gene expression values using Pearson correlation coefficient to estimate sample-wise distances and the `ward.D2` agglomeration method. Furthermore, prediction gene expression-based subtypes, i.e., consensus molecular subtypes (CMS) and CRC intrinsic subtypes (CRIS), was performed after log2-transformation of gene expression values using `CMScaller` and `CRISclassifier` R packages with default parameters^{43,44}. In both cases, CRC cell lines that could not be confidently assigned to a single subtype (FDR > 5%) were labelled as NA (not available).

4.9 Identification of somatic fusion transcripts

RNA-seq reads were processed using two publicly available tools (STAR-Fusion (v1.10.0) and FusionCatcher (v1.33)) to identify somatic fusion transcripts in CRC cell lines^{82,83}. STAR-Fusion predicted fusions were filtered based on the number of supporting reads (FFPM > 0.1 as suggested by the authors) and fusions thought to be red herrings (i.e., fusions that may not be relevant for cancer and potential false positives) were excluded based on STAR-Fusion automatic annotation. Then, to further control for the presence of artefacts, we required somatic fusion transcripts to be called also by FusionCatcher that was executed with default parameters. Finally, somatic fusion transcripts identified in CRC cell lines in which the amplification of the 3' partner gene was also found were not considered in subsequent analysis, since amplification-associated fusion transcripts likely are a by-product of chromosomal instability, as previously reported¹⁰⁴.

4.10 DNA methylation levels and annotation of the CpG island methylator phenotype (CIMP)

HM450 and EPIC raw data (IDAT files) were processed with a similar workflow using the minfi R Bioconductor package and corresponding annotation packages¹⁰⁵. In each sample, probes with detection p-value > 0.05 were masked in the RGChannelSet. In addition, cross-reactivity probes and probes matching SNPs at the target CpG site were filtered using previously published lists^{106,107}. Then, methylation and unmethylation signals were obtained for each CpG site using the preprocessNoob function and β -values were computed. Results obtained in all samples were subsequently merged in a single β -matrix, thus keeping only CpG sites targeted by both HM450 and EPIC. Furthermore, CpG sites that are located on sex chromosomes or chrM were excluded. For the classification of CRC cell lines with respect to CIMP, we referred to a previously published list of 318 probes that showed significantly higher DNA methylation level in both CIMP-H and CIMP-L tumours compared to non-CIMP tumours (CIMP associated probes)⁴⁶. Starting from the β -matrix obtained as described above, we selected the CIMP associated probes and excluded those containing missing values. Then, we performed unsupervised clustering using recursively partitioned mixture model (RPMM) with maxlevel=2 and finally associated the resulting clusters to CIMP classes based on their median DNA methylation level (CIMP-H, CIMP-L, CIMP3 and CIMP4, for high to low median DNA methylation level).

4.11 Transcriptome-wide gene expression outlier analysis in cancer cell lines

Our pipeline for the identification of transcriptome-wide gene expression outliers relies on two different steps. As a starting point, the Tukey's method is applied to the log2-transformed FPKM values of each single gene. In detail, the first (Q1) and the third (Q3) quartiles are obtained and the interquartile range (IQR) is computed as Q3-Q1. Then, positive outliers are identified as samples whose log2-transformed expression level is higher than $Q3 + 1.5 \times IQR$. Similarly, negative outliers are identified as samples

whose log2-transformed expression level is lower than $Q1 - 1.5 \times IQR$. In the second part of the pipeline, multiple filters are exploited to select the strongest gene expression alterations, assuming that they have a higher likelihood of being functionally relevant in CRC cells.

- (1) The furthest positive and negative outliers of each gene are selected. When multiple samples are identified as positive or negative outliers for the same gene, two groups of outliers can be distinguished. Specifically, the positive outliers can be subdivided into those that are nearest to the maximum expression value and those that are nearest to the highest expression value that is not an outlier. The positive outliers in the former group are the furthest positive outliers for that gene. Similarly, the furthest negative outliers of each gene are those whose distance from the minimum expression level is less than the distance from the lowest expression level that is not an outlier.
- (2) Positive and negative outliers for which the absolute value of the differential expression level is higher than 3 are selected. For each gene expression outlier, the differential expression level is computed as the log2 fold change with respect to the median gene expression using this formula: $\log_2((FPKM+1)/(median+1))$.
- (3) Positive and negative outliers are further filtered based on the absolute expression level. Specifically, only positive outliers that have $FPKM > 10$ are selected, while negative outliers are retained if they have $FPKM < 1$.

We refer to the resulting gene expression outliers as “extreme positive outliers” and “extreme negative outliers” to distinguish them from positive and negative outliers initially obtained by applying the Tukey’s method.

The described pipeline was applied to gene expression data from 226 CRC cell lines for which multi-omics (WES, RNA-seq and DNA methylation microarray) data were available. We reasoned that the gene expression outlier analysis should be done separately for female and male samples in the case of genes on sex chromosomes. However, due to missing information on gender of individuals from which the cell lines were derived and since sex chromosomes are frequently lost in cancer cells¹⁰⁸, genes located on sex chromosomes were excluded from the gene expression outlier analysis.

4.12 Computation of outlier burden values in CRC cell lines

For each CRC cell line, we first considered the number of overexpressed genes (i.e., the genes for which it was selected as an "extreme positive outlier") and the number of underexpressed genes (i.e., the genes for which it was selected as an "extreme negative outlier"). Then, these values were normalized with respect to the number of genes for which each CRC cell line could have been selected as an "extreme outlier" based on the filtering strategy described above. In detail, the "positive outlier burden" of each model is obtained by dividing the number of overexpressed genes by the number of genes that have $\text{FPKM} > 10$ and then multiplying the result by 1,000. Similarly, the "negative outlier burden" of each model is obtained by dividing the number of underexpressed genes by the number of genes that have $\text{FPKM} < 1$ and then multiplying the result by 1,000. The "positive outlier burden" and the "negative outlier burden" of each model are then summed to get the "total outlier burden".

4.13 Identification of aberrantly expressed genes in CRC cell lines with the Outlier in RNA-Seq Finder (OUTRIDER) algorithm

OUTRIDER⁷⁷ was applied to gene expression data from 226 CRC cell lines, starting from RSEM genes counts. In detail, to allow a fair comparison between the different methods, OUTRIDER was used to analyse the expression profile only of those genes that have been considered in the transcriptome-wide gene expression outlier analysis performed with our approach. We tested both the fitting of an autoencoder and a PCA-based correction to control for confounders that could have been present in the data. Since OUTRIDER assumes that samples are independent, in the case of CRC cell lines that are genetically identical, only one model for each group was selected and the others were masked during the autoencoder fitting. The automatically estimated optimal encoding dimension was found to be $q=14$ for this dataset; we also tested models using 7 and 28 latent variables. Outlier samples for each gene were identified as those whose adjusted p-value was lower than 0.05 and then z-scores were used to distinguish overexpressed and underexpressed genes.

4.14 Integration of gene expression outliers with promoter DNA methylation levels

We tested whether outlier gene expression values are associated with DNA hypermethylation or hypomethylation in promoter regions. First, we selected genes for which at least 2 CRC cell lines were identified as extreme positive or negative outliers and from the β -matrix we extracted the CpG sites that are located in the corresponding promoter region defined as 1500bp upstream of transcription start site(s). Then, for each gene we compared the β -values obtained for each CpG site within its promoter region between outliers and remaining samples by performing a Wilcoxon test and FDR values were computed by applying the Benjamini-Hochberg method. Promoter CpG sites were considered differentially methylated between outlier and other samples when both $\text{FDR} < 0.05$ and a high difference in β -values were observed. In the case of extreme positive outliers, the median β -value must be > 0.80 in outlier samples and < 0.20 in other samples; on the contrary, in the case of extreme negative outliers, the median β -value must be < 0.20 in outlier samples and > 0.80 in other samples. Furthermore, the analysed CpG sites were annotated with respect to CpG islands (CGI). Starting from the University of California Santa Cruz (UCSC) annotation track for CpG islands in hg38, we segmented the human genome in CGI, CGI shores (i.e., 2kb flanking a CGI on the left and on the right), CGI shelves (i.e., 2kb flanking outwards a CGI shore) and open sea (i.e., the rest of the genome). This annotation was used to compare the genomic position of CpG sites that are differentially methylated with that of the non-significant ones. Moreover, to classify genes for which we found a significant association between underexpression and promoter hypermethylation in extreme negative outliers in “associated with CIMP” and “not associated with CIMP”, we tested whether outlier samples were enriched for CIMP samples (i.e., CIMP-H and CIMP-L samples) by performing a hypergeometric test and FDR values were computed by applying the Benjamini-Hochberg method. A significant association with the CIMP phenotype was called for genes for which $\text{FDR} < 0.05$ was obtained.

4.15 Enzymes, kinases and target development level classification

The Swiss-Prot database within Universal Protein Resource Knowledgebase (UniProtKB) was downloaded in XML format on 17/11/2021¹⁰⁹. The XML file was then processed with a custom script to select human proteins with the following annotations if available: entry primary accession in UniProtKB, entry name in UniProtKB, recommended protein name, Enzyme Commission (EC) number(s), primary gene name, Ensembl gene ID(s) and Target Development Level (TDL). Swiss-Prot UniProtKB retrieved the TDL classification of human proteins from the Pharos portal that provides access to data collected by NIH Illuminating the Druggable Genome (IDG) program⁹⁰. Then, to focus on genes coding for enzymes, we selected Swiss-Prot UniProtKB entries that were associated with EC numbers, meaning that they have catalytic activity, and we used the resulting list of genes to filter the extreme gene expression outliers. In this way, 3895 genes coding for enzymes were identified and categorized into seven families, based on the first digit of the EC number (oxidoreductases, EC 1; transferases, EC 2; hydrolases, EC 3; lyases, EC 4; isomerases, EC 5; ligases, EC 6; translocases, EC 7), with consideration of the fact that enzymes can catalyse multiple types of reactions and, thus, have multiple EC numbers¹¹⁰. In addition, a further selection was performed considering an independent collection of 535 human kinases that were classified into different families, including both the eukaryotic protein kinase superfamily (TK, TKL, STE, CK1, AGC, CAMK, CMGC, other) and atypical kinases¹¹¹. However, human kinases that according to Swiss-Prot UniProtKB do not have detectable kinase activity were excluded, reducing the number of analysed kinases to 454.

4.16 Western Blot

Whole protein lysates were isolated from cell lines by boiling in SDS buffer (50 mmol/L Tris-HCl (pH 7.5), 150 mmol/L NaCl, and 1% SDS) to extract total cellular proteins. The lysates were quantified by the BCA Protein Assay Reagent kit (Thermo Fisher Scientific) and prepared for western blot using LDS and Reducing Agent (Invitrogen). Western blot analysis was performed with Enhanced Chemiluminescence System (GE Healthcare) and peroxidase-conjugated secondary antibodies (Amersham). The following primary antibodies were used for Western blotting: anti-MTAP (Santa Cruz

42-T; 1:200) and anti-HSP90 (Santa Cruz SC-7947; 1:500). Detection of the chemiluminescent signal was performed with ChemiDoc Imaging System (Bio-Rad).

4.17 In vitro drug screenings

Lapatinib and AGI-24512 were purchased from MedChem Express, trastuzumab was obtained from the Pharmacy at Niguarda Cancer Center in Milan, Italy, while MRTX1719 was kindly provided by Mirati Therapeutics under an academic material transfer agreement. CRC cells were seeded at different densities ($1,5 - 3 \times 10^3$ cells/well) in 100µl medium containing 10% FBS in 96-multiwell plates at day 0. The following day, serial dilutions of the indicated drugs (Lapatinib, trastuzumab, AGI-24512 or MRTX1719) in serum-free medium were added to the cells (ratio 1:1) in technical triplicates, while DMSO-only treated cells were included as controls. Cell viability was assessed after 6 (for lapatinib and trastuzumab) or 7 (for AGI-24512 and MRTX1719) days by measuring ATP content through Cell Titer-Glo® Luminescent Cell Viability assay (Promega) according to the manufacturer's protocol. Luminescence was measured by the Tecan SPARK M10 plate reader. Cell viability measured for each treatment condition was normalized to viability of DMSO-treated controls. Data represent average $\pm$ SD of at least three independent biological replicates. To compare the mean of the reduction of the cell viability between MTAP-deleted and wild-type CRC cell lines, for each inhibitor a one-way ANOVA test has been performed at different drug concentrations with the formula $\text{Viability} \sim \text{Condition} + \text{Error}(\text{Model})$. This has required selecting few relevant doses among those we considered in our screening. In the case of the AGI-24512 inhibitor, we have taken into account the ones closest to 1 µM since this concentration has been previously used in functional assays¹¹². Instead, as for the MRTX1719 inhibitor, we have considered four concentrations (9.14 nM, 27.43 nM, 82.3 nM and 246.9 nM) at which it is expected to be highly selective in inhibiting PRMT5 activity and cell viability in MTAP-deleted cells, based on recently published results in HCT116 isogenic cell lines⁹¹.

In addition, the response to temozolomide (TMZ) was retrieved from a previous publication in which forty-seven CRC cell lines were tested with TMZ in long-term

colony forming assays, from which half-maximal inhibitory concentration (IC50) values were obtained⁸⁹.

4.18 Transcriptome-wide gene expression outlier analysis in CRC samples

Gene expression quantification data (STAR counts) generated by The Cancer Genome Atlas (TCGA) project were downloaded for TCGA-COAD and TCGA-READ samples from Genomic Data Commons (GDC) using the GenomicDataCommons Bioconductor package (Data Release 37). Data corresponding to primary tumors were selected and those corresponding to patients for which RNA-seq has been performed multiple times from the same or different vials were excluded. Then, data were processed as described above using GENCODE v36 as gene annotation to be consistent with the GDC mRNA quantification analysis pipeline. Finally, extreme positive and negative outliers were identified by adapting the previously described pipeline. In brief, after the Tukey's rule had been applied to the expression profile of each individual gene, the furthest outliers were selected for each gene and then further filtered based on differential and absolute expression values. In this case, only positive outliers whose differential expression level was higher than 2 were maintained, while negative outliers were kept if the absolute value of differential expression level was higher than 1. These alternative thresholds were chosen so that the percentage of extreme positive and negative outliers compared to the total number of positive and negative outliers in the TCGA dataset is similar to that observed in the CRC cell line dataset.

Acknowledgements

Most of the data presented in this thesis have been published on:

Mariella E., Grasso G., et al., *Transcriptome-wide gene expression outlier analysis pinpoints therapeutic vulnerabilities in colorectal cancer*, Molecular Oncology, 2024, PMID: 38468448, PMCID: PMC11161737.

We thank members of the Bardelli lab at University of Torino and IFOM for valuable suggestions and critical review of the manuscript.

Part of the results are based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

The Figures 3A, 15A, 17A-B and 18 were created with BioRender.com.

Raw DNA methylation data (IDAT files) previously obtained for CRC cell lines using the Infinium HumanMethylation450 BeadChip array are available in the Gene Expression Omnibus (GEO) with GSE86078 accession code.

Raw DNA methylation data (IDAT files) recently obtained for CRC cell lines using the Infinium MethylationEPIC BeadChip microarray were deposited on GEO with GSE220197 accession code.

WES and RNA-seq data previously obtained for CRC cell lines are available in the European Nucleotide Archive (ENA) with PRJEB33045 and PRJEB33640 accession codes.

Further NGS data for CRC cell lines were deposited in ENA with PRJEB57691 accession code.

References

1. Evans, W., Meslin, E. M., Kai, J. & Qureshi, N. Precision Medicine-Are We There Yet? A Narrative Review of Precision Medicine's Applicability in Primary Care. *J Pers Med* **14**, (2024).
2. Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for precision cancer therapies. *Genome Med* **12**, 8 (2020).
3. Andrei, P. *et al.* Integrated approaches for precision oncology in colorectal cancer: The more you know, the better. *Semin Cancer Biol* **84**, 199–213 (2022).
4. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364**, 2507–2516 (2011).
5. Sharma, S. V, Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat Rev Cancer* **7**, 169–181 (2007).
6. Lièvre, A. *et al.* KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res* **66**, 3992–3995 (2006).
7. Di Nicolantonio, F. *et al.* Wild-type BRAF is required for response to panitumumab or cetuximab in metastatic colorectal cancer. *J Clin Oncol* **26**, 5705–5712 (2008).
8. Siena, S. *et al.* Targeting the human epidermal growth factor receptor 2 (HER2) oncogene in colorectal cancer. *Ann Oncol* **29**, 1108–1119 (2018).
9. Swain, S. M., Shastry, M. & Hamilton, E. Targeting HER2-positive breast cancer: advances and future directions. *Nat Rev Drug Discov* **22**, 101–126 (2023).
10. Harada, G. & Drilon, A. TRK inhibitor activity and resistance in TRK fusion-positive cancers in adults. *Cancer Genet* **264–265**, 33–39 (2022).
11. Cheng, J. *et al.* Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat Commun* **8**, 1221 (2017).
12. Nishiyama, A. & Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet* **37**, 1012–1027 (2021).
13. Taieb, J. *et al.* Deficient mismatch repair/microsatellite unstable colorectal cancer: Diagnosis, prognosis and treatment. *Eur J Cancer* **175**, 136–157 (2022).
14. Strobel, H. *et al.* Temozolomide and Other Alkylating Agents in Glioblastoma Therapy. *Biomedicines* **7**, (2019).
15. Min, A. *et al.* RAD51C-deficient cancer cells are highly sensitive to the PARP inhibitor olaparib. *Mol Cancer Ther* **12**, 865–877 (2013).
16. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168**, 629–643 (2017).
17. Farazi, M. M. R. & Imon, A. H. M. R. Identification of Outliers in Gene Expression Data. *Data Science and SDGs: Challenges, Opportunities and Realities* 135–145 (2021) doi:10.1007/978-981-16-1919-9_11.
18. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).

19. Tomlins, S. A. *et al.* Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer.
20. Medico, E. *et al.* The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat Commun* **6**, 7002 (2015).
21. Sengupta, S. *et al.* Integrative omics analyses broaden treatment targets in human cancer. *Genome Med* **10**, 60 (2018).
22. Lazzari, L. *et al.* Patient-Derived Xenografts and Matched Cell Lines Identify Pharmacogenomic Vulnerabilities in Colorectal Cancer. *Clin Cancer Res* **25**, 6243–6259 (2019).
23. Lazzari, L. *et al.* Patient-Derived Xenografts and Matched Cell Lines Identify Pharmacogenomic Vulnerabilities in Colorectal Cancer. *Clin Cancer Res* **25**, 6243–6259 (2019).
24. Vivian, J., Eizenga, J. M., Beale, H. C., Vaske, O. M. & Paten, B. Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples. *JCO Clin Cancer Inform* 160–170 (2020) doi:10.1200/CCI.19.00095.
25. Kothari, V. *et al.* Outlier Kinase Expression by RNA Sequencing as Targets for Precision Therapy. *Cancer Discov* **3**, 280–293 (2013).
26. Cohen, P., Cross, D. & Jänne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature Reviews Drug Discovery* 2021 20:7 **20**, 551–569 (2021).
27. 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* **71**, 209–249 (2021).
28. Cervantes, A. *et al.* Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol* **34**, 10–32 (2023).
29. Andrei, P. *et al.* Integrated approaches for precision oncology in colorectal cancer: The more you know, the better. *Semin Cancer Biol* **84**, 199–213 (2022).
30. Bertotti, A. *et al.* A Molecularly Annotated Platform of Patient-Derived Xenografts (“Xenopatients”) Identifies HER2 as an Effective Therapeutic Target in Cetuximab-Resistant Colorectal Cancer. *Cancer Discov* **1**, 508–523 (2011).
31. Russo, M. *et al.* Acquired Resistance to the TRK Inhibitor Entrectinib in Colorectal Cancer. *Cancer Discov* **6**, 36–44 (2016).
32. Di Nicolantonio, F. *et al.* Precision oncology in metastatic colorectal cancer - from biology to medicine. *Nat Rev Clin Oncol* **18**, 506–525 (2021).
33. Sartore-Bianchi, A. *et al.* Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial. *Lancet Oncol* **17**, 738–746 (2016).
34. Meric-Bernstam, F. *et al.* Pertuzumab plus trastuzumab for HER2-amplified metastatic colorectal cancer (MyPathway): an updated report from a multicentre, open-label, phase 2a, multiple basket study. *Lancet Oncol* **20**, 518–530 (2019).
35. Wang, H. *et al.* NTRK fusion positive colorectal cancer is a unique subset of CRC with high TMB and microsatellite instability. *Cancer Med* **11**, 2541 (2022).

36. Piazza, G. A., Chandrasekaran, P., Maxuitenko, Y. Y. & Budhwani, K. I. Assessment of KRASG12C inhibitors for colorectal cancer. *Front Oncol* **14**, (2024).
37. Ciardiello, D. *et al.* Immunotherapy of colorectal cancer: Challenges for therapeutic efficacy. *Cancer Treat Rev* **76**, 22–32 (2019).
38. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350–1356 (2015).
39. Isella, C. *et al.* Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat Commun* **8**, 15107 (2017).
40. Dong, Z. *et al.* Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer* **15**, 1–12 (2015).
41. Kim, Y., Kim, D., Cao, B., Carvajal, R. & Kim, M. PDXGEM: Patient-derived tumor xenograft-based gene expression model for predicting clinical response to anticancer therapy in cancer patients. *BMC Bioinformatics* **21**, 1–21 (2020).
42. Clayton, E. A., Pujol, T. A., McDonald, J. F. & Qiu, P. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinformatics* **21**, 1–11 (2020).
43. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep* **7**, 16618 (2017).
44. Isella, C. *et al.* Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat Commun* **8**, 15107 (2017).
45. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350–1356 (2015).
46. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* **22**, 271–282 (2012).
47. Medico, E. *et al.* The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat Commun* **6**, 7002 (2015).
48. Russo, M. *et al.* Acquired Resistance to the TRK Inhibitor Entrectinib in Colorectal Cancer. *Cancer Discov* **6**, 36–44 (2016).
49. Siravegna, G. *et al.* Radiologic and Genomic Evolution of Individual Metastases during HER2 Blockade in Colorectal Cancer. *Cancer Cell* **34**, 148–162.e7 (2018).
50. Taieb, J. *et al.* Deficient mismatch repair/microsatellite unstable colorectal cancer: Diagnosis, prognosis and treatment. *Eur J Cancer* **175**, 136–157 (2022).
51. Pietrantonio, F. *et al.* Activity of temozolomide in patients with advanced chemorefractory colorectal cancer and MGMT promoter methylation. *Annals of Oncology* **25**, 404–408 (2014).
52. Germano, G. *et al.* Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116–120 (2017).
53. Brechtmann, F. *et al.* OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *The American Journal of Human Genetics* **103**, 907–917 (2018).

54. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33**, 245–254 (2003).
55. Arena, S. *et al.* A Subset of Colorectal Cancers with Cross-Sensitivity to Olaparib and Oxaliplatin. *Clin Cancer Res* **26**, 1372–1384 (2020).
56. Durinikova, E. *et al.* Targeting the DNA Damage Response Pathways and Replication Stress in Colorectal Cancer. *Clin Cancer Res* **28**, 3874–3889 (2022).
57. Min, A. *et al.* RAD51C-deficient cancer cells are highly sensitive to the PARP inhibitor olaparib. *Mol Cancer Ther* **12**, 865–877 (2013).
58. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* **22**, 271–282 (2012).
59. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep* **23**, 227–238.e3 (2018).
60. Network, C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
61. Serebriiskii, I. G. *et al.* Comprehensive characterization of PTEN mutational profile in a series of 34,129 colorectal cancers. *Nat Commun* **13**, 1618 (2022).
62. Picco, G. *et al.* Loss of AXIN1 drives acquired resistance to WNT pathway blockade in colorectal cancer cells carrying RSPO3 fusions. *EMBO Mol Med* **9**, 293–303 (2017).
63. Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* **17**, 317–332 (2018).
64. Wilson, L. J. *et al.* New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome. *Cancer Res* **78**, 15–29 (2018).
65. Muller, F. L. *et al.* Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* **488**, 337–342 (2012).
66. Frezza, C. *et al.* Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* **477**, 225–228 (2011).
67. Kryukov, G. V *et al.* MTAP deletion confers enhanced dependency on the PRMT5 arginine methyltransferase in cancer cells. *Science* **351**, 1214–1218 (2016).
68. Mavrakis, K. J. *et al.* Disordered methionine metabolism in MTAP/CDKN2A-deleted cancers leads to dependence on PRMT5. *Science* **351**, 1208–1213 (2016).
69. Kaley, P. *et al.* MAT2A Inhibition Blocks the Growth of MTAP-Deleted Cancer Cells by Reducing PRMT5-Dependent mRNA Splicing and Inducing DNA Damage. *Cancer Cell* **39**, 209–224.e11 (2021).
70. Engstrom, L. D. *et al.* MRTX1719 is an MTA-cooperative PRMT5 inhibitor that exhibits synthetic lethality in preclinical models and patients with MTAP deleted cancer. *Cancer Discov* (2023) doi:10.1158/2159-8290.CD-23-0669.
71. Marjon, K. *et al.* MTAP Deletions in Cancer Create Vulnerability to Targeting of the MAT2A/PRMT5/RIOK1 Axis. *Cell Rep* **15**, 574–587 (2016).

72. Smith, C. R. *et al.* Fragment-Based Discovery of MRTX1719, a Synthetic Lethal Inhibitor of the PRMT5:MTA Complex for the Treatment of MTAP-Deleted Cancers. *J Med Chem* **65**, 1749–1766 (2022).
73. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, (2015).
74. Di Nicolantonio, F. *et al.* Precision oncology in metastatic colorectal cancer - from biology to medicine. *Nat Rev Clin Oncol* **18**, 506–525 (2021).
75. Network, C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
76. Wilding, J. L. & Bodmer, W. F. Cancer cell lines for drug discovery and development. *Cancer Res* **74**, 2377–2384 (2014).
77. Brechtmann, F. *et al.* OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *The American Journal of Human Genetics* **103**, 907–917 (2018).
78. Nuzzo, A. *et al.* KAOS: a new automated computational method for the identification of overexpressed genes. *BMC Bioinformatics* **17**, (2016).
79. Roden, D. L. *et al.* ZODET: software for the identification, analysis and visualisation of outlier genes in microarray expression data. *PLoS One* **9**, (2014).
80. Dorney, R., Dhungel, B. P., Rasko, J. E. J., Hebbard, L. & Schmitz, U. Recent advances in cancer fusion transcript detection. *Brief Bioinform* **24**, (2023).
81. Group, P. T. C. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
82. Nicorici, D. *et al.* FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv* 11650 (2014) doi:10.1101/011650.
83. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* **20**, 213 (2019).
84. Kolomietz, E. *et al.* Primary chromosomal rearrangements of leukemia are frequently accompanied by extensive submicroscopic deletions and may lead to altered prognosis. *Blood* **97**, 3581–3588 (2001).
85. Huntly, B. J. P., Bench, A. & Green, A. R. Double jeopardy from a single translocation: deletions of the derivative chromosome 9 in chronic myeloid leukemia. *Blood* **102**, 1160–1168 (2003).
86. Mehrmohamadi, M., Sepehri, M. H., Nazer, N. & Norouzi, M. R. A Comparative Overview of Epigenomic Profiling Methods. *Front Cell Dev Biol* **9**, (2021).
87. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
88. Siravegna, G. *et al.* Radiologic and Genomic Evolution of Individual Metastases during HER2 Blockade in Colorectal Cancer. *Cancer Cell* **34**, 148-162.e7 (2018).
89. Germano, G. *et al.* Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116–120 (2017).
90. Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* **17**, 317–332 (2018).

91. Engstrom, L. D. *et al.* MRTX1719 is an MTA-cooperative PRMT5 inhibitor that exhibits synthetic lethality in preclinical models and patients with MTAP deleted cancer. *Cancer Discov* (2023) doi:10.1158/2159-8290.CD-23-0669.
92. Smith, C. R. *et al.* Fragment-Based Discovery of MRTX1719, a Synthetic Lethal Inhibitor of the PRMT5:MTA Complex for the Treatment of MTAP-Deleted Cancers. *J Med Chem* **65**, 1749–1766 (2022).
93. Belmontes, B. *et al.* AMG 193, a Clinical Stage MTA-Cooperative PRMT5 Inhibitor, Drives Antitumor Activity Preclinically and in Patients With MTAP-Deleted Cancers. *Cancer Discov* (2024) doi:10.1158/2159-8290.CD-24-0887.
94. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
95. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
96. Jia, P. *et al.* MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability. *Genomics Proteomics Bioinformatics* **18**, 65–71 (2020).
97. Barault, L. *et al.* Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. *Gut* **67**, 1995–2005 (2018).
98. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
99. Corti, G. *et al.* A Genomic Analysis Workflow for Colorectal Cancer Precision Oncology. *Clin Colorectal Cancer* **18**, 91-101.e3 (2019).
100. Crisafulli, G. *et al.* Temozolomide Treatment Alters Mismatch Repair and Boosts Mutational Burden in Tumor and Blood of Colorectal Cancer Patients. *Cancer Discov* **12**, 1656–1675 (2022).
101. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178 (2010).
102. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
103. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
104. Kalyana-Sundaram, S. *et al.* Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. *Neoplasia* **14**, 702–708 (2012).
105. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
106. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208 (2016).
107. Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**, 4 (2013).
108. Rubin, J. B. The spectrum of sex differences in cancer. *Trends Cancer* **8**, 303–315 (2022).

109. Consortium, U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).
110. McDonald, A. G. & Tipton, K. F. Enzyme nomenclature and classification: the state of the art. *FEBS J* **290**, 2214–2231 (2023).
111. Wilson, L. J. *et al.* New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome. *Cancer Res* **78**, 15–29 (2018).
112. Kalev, P. *et al.* MAT2A Inhibition Blocks the Growth of MTAP-Deleted Cancer Cells by Reducing PRMT5-Dependent mRNA Splicing and Inducing DNA Damage. *Cancer Cell* **39**, 209–224.e11 (2021).

Tables

Table 1 - Annotation of the 226 colorectal cancer (CRC) cell lines.

For each cell line we report:

1. CRC cell line name
2. Microsatellite status: microsatellite stability (MSS) or instability (MSI)
3. KRAS mutational status (Hotspot: hotspot mutations, Other: wild-type or other mutations)
4. NRAS mutational status (Hotspot: hotspot mutations, Other: wild-type or other mutations)
5. BRAF mutational status (Hotspot: hotspot mutations, Other: wild-type or other mutations)
6. APC mutational status (WT: wild-type, Nonsense: nonsense mutation, Missense: missense mutations, frameshift: frameshift indel, in-frame: in-frame indel)
7. TP53 mutational status (WT: wild-type, Hotspot: hotspot mutations, Nonsense: nonsense mutation, frameshift: frameshift indel, Other: other missense mutation or in-frame indel)
8. CMS: consensus molecular subtypes (CMS1, CMS2, CMS3 or CMS4). Cell lines that cannot be confidently assigned to a single subtype (FDR>5%) were labelled as NA (not available)
9. CRIS: colorectal cancer intrinsic subtypes (CRIS-A, CRIS-B, CRIS-C, CRIS-D or CRIS-E). Cell lines that cannot be confidently assigned to a single subtype (FDR>5%) were labeled as NA (not available)
10. CIMP: status for the CpG island methylator phenotype

CRC cell line name	Microsatellite status	KRAS	NRAS	BRAF	APC	TP53	CMS	CRIS	CIMP
B1003_XL5	MSI	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-A	CIMP-L
B1003_XLSP	MSI	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-A	CIMP-L
C10	MSI	Other	Other	Other	WT	WT	CMS4	CRIS-A	CIMP4
C106	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP4
C125PM	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-E	CIMP3
C146	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	NA	CIMP-H
C170	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS1	NA	CIMP-H
C32	MSS	Other	Hotspot	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-B	CIMP4
C70	MSS	Other	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-C	CIMP4
C75	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
C80	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP3
C84	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
C99	MSS	Other	Other	Other	WT	WT	CMS2	CRIS-E	CIMP4
CACO2	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP3
CAR1	MSS	Other	Other	Other	WT	Other	CMS4	CRIS-B	CIMP3
CCK81	MSI	Other	Other	Other	Missense, in-frame	Other	CMS2	CRIS-C	CIMP3
CL11	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS4	CRIS-B	CIMP-H
CL14	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP4
CL34	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP-H
CL40	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-A	CIMP-L
CO115	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS1	NA	CIMP-H
COCM1	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
COGA1	MSI	Other	Other	Other	Nonsense, frameshift	WT	NA	NA	CIMP4

COGA10	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP4
COGA12	MSI	Other	Other	Other	Nonsense, frameshift	WT	NA	CRIS-E	CIMP-L
COGA2	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-E	CIMP-H
COGA3	MSI	Hotspot	Other	Other	WT	WT	CMS2	NA	CIMP-L
COGA5	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
COGA5L	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-L
COGA8	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-C	CIMP3
COLO201	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-B	CIMP-H
COLO205	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-E	CIMP-H
COLO320	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP4
COLO320D M	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP4
COLO320HS R	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	NA	CIMP4
COLO60H	MSI	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-A	CIMP3
COLO678	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS4	NA	CIMP-H
COLO94H	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-E	CIMP-L
CR4	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP4
CRC0078_X L	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP3
CRC0080_X L	MSS	Other	Other	Other	WT	Other	CMS2	CRIS-C	CIMP4
CRC0104_X L	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP4
CRC0174_X L	MSS	Other	Hotspot	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
CRC0313_X L	MSS	Other	Hotspot	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
CRC0322_X L	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4

CRC0394_X L	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
CRC0438_X L	MSS	Hotspot	Other	Other	WT	Other	CMS1	CRIS-B	CIMP-L
CRC0558_X L	MSS	Other	Hotspot	Other	WT	Hotspot, Nonsense, frameshift	CMS4	CRIS-C	CIMP4
CRC0691_X L	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
CRC0740_X L	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP4
CRC0778_X L	MSS	Hotspot	Other	Other	WT	WT	CMS2	CRIS-E	CIMP4
CRC0781_X L	MSS	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-A	CIMP3
CRC1360_X L	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
CX1	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	NA	CIMP-L
DIFI	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP4
DLD1	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	NA	CIMP-H
FET	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	CRIS-A	CIMP-L
GEO	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	CRIS-A	CIMP-L
GP2D	MSI	Hotspot	Other	Other	Nonsense, frameshift	WT	NA	NA	CIMP-L
GP5D	MSI	Hotspot	Other	Other	Nonsense, frameshift	WT	NA	CRIS-C	CIMP-L
HCA24	MSS	Other	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-E	CIMP4
HCA46	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
HCA7	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	NA	CIMP3
HCC2998	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-B	CIMP3
HCT116	MSI	Hotspot	Other	Other	WT	WT	CMS4	NA	CIMP-H
HCT8	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS4	NA	CIMP-H
HDC114	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	NA	CIMP4
HDC135	MSI	Other	Other	Hotspot	Missense, in-frame	WT	CMS1	NA	CIMP-H
HDC142	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-D	CIMP4

HDC143	MSI	Other	Other	Other	Nonsense, frameshift	Other	CMS3	CRIS-E	CIMP3
HDC54	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
HDC8	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS4	CRIS-B	CIMP-H
HDC82	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
HDC9	MSI	Other	Other	Other	Missense, in-frame	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP3
HHC6548	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP4
HRA16	MSS	Other	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-C	CIMP4
HROBMCO1	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP4
HROC107_cTOM2	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS4	CRIS-D	CIMP3
HROC112_MET	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP-L
HROC113	MSI	Hotspot	Other	Other	Missense, in-frame	Other	CMS2	CRIS-E	CIMP4
HROC126	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
HROC131_TOM3	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP-H
HROC147_MET	NA	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
HROC147_TOM1	NA	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-C	CIMP3
HROC173	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP-L
HROC18	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-L
HROC183	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP-L
HROC212	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS1	NA	CIMP-H
HROC222_T1M2	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
HROC239_TOM1	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-D	CIMP3

HROC24	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS3	CRIS-A	CIMP-H
HROC257_T OM1	MSI	Other	Other	Hotspot	Missense, in-frame	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
HROC277_ MET2	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP4
HROC277_T OM1	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-C	CIMP4
HROC278_ MET	MSS	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	NA	CRIS-E	CIMP3
HROC278_T OM1	MSS	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	NA	CRIS-E	CIMP-L
HROC284_ MET	MSS	Other	Hotspot	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP4
HROC300_T 2M1	MSS	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-H
HROC313_ MET1	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
HROC32	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP3
HROC324	MSI	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP4
HROC334	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-A	CIMP-H
HROC39	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS3	CRIS-C	CIMP-L
HROC40	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-A	CIMP3
HROC43	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-L
HROC46	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS4	CRIS-D	CIMP3
HROC50_T1 M5	MSI	Other	Other	Hotspot	Nonsense, frameshift	Other	CMS3	CRIS-A	CIMP-H
HROC57	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP-L
HROC59	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-D	CIMP3
HROC60	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP-L
HROC69	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP3

HROC80	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP4
HROC87	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS1	CRIS-B	CIMP-H
HT115	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-E	CIMP-L
HT29	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-H
HT55	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP4
HUTU80	MSS	Other	Other	Other	WT	WT	CMS4	CRIS-D	CIMP4
IRCC1_XL	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP-H
IRCC10_A_HL	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
IRCC22_XL	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS3	CRIS-A	CIMP3
IRCC3_HL	MSI	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP-H
IRCC3_XL	MSI	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP-H
IRCC5_A_XL	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP4
IRCC72_A_XL	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP4
ISRECO1	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS4	NA	CIMP-H
JVE015	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
JVE044	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP-H
JVE059	MSI	Other	Other	Other	Nonsense, frameshift	WT	NA	CRIS-C	CIMP3
JVE103	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
JVE109	MSI	Other	Other	Hotspot	Missense, in-frame	Other	CMS3	CRIS-A	CIMP-L
JVE114	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-C	CIMP4
JVE127	MSS	Other	Other	Hotspot	WT	Other	CMS3	CRIS-A	CIMP-H
JVE187	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	CRIS-A	CIMP-L
JVE192	MSI	Hotspot	Other	Other	Missense, in-frame	WT	CMS3	CRIS-A	CIMP4
JVE207	MSS	Other	Other	Hotspot	Nonsense, frameshift	Other	CMS3	CRIS-E	CIMP-H

JVE241	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP4
JVE253	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
JVE367	MSS	Other	Other	Hotspot	WT	WT	CMS3	CRIS-A	CIMP-L
JVE371	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-B	CIMP4
JVE528	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
KM12	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP-H
KM12C	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-A	CIMP-H
KM12L4	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
KM12SM	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
KM20	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-L
KP283T	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
KP363T	MSS	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-H
KP7038T	MSI	Other	Other	Other	Nonsense, frameshift	WT	NA	NA	CIMP4
LIM1215	MSI	Other	Other	Other	WT	WT	NA	NA	CIMP3
LIM1863	MSI	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-E	CIMP-H
LIM1899	MSI	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS3	NA	CIMP4
LIM2099	MSS	Hotspot	Other	Other	WT	WT	CMS4	CRIS-B	CIMP-L
LIM2405	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS1	NA	CIMP-H
LIM2412	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
LIM2537	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP-H
LIM2550	MSI	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
LIM2551	MSI	Other	Other	Hotspot	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP3

LM0201_A_XL	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
LM0701_A_XL	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
LOVO	MSI	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS1	CRIS-B	CIMP-L
LS1034	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-D	CIMP-L
LS123	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP-L
LS174T	MSI	Hotspot	Other	Other	WT	WT	CMS3	CRIS-A	CIMP4
LS180	MSI	Hotspot	Other	Other	Missense, in-frame	WT	CMS3	CRIS-A	CIMP4
LS411N	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP-H
LS513	MSS	Hotspot	Other	Other	WT	WT	CMS3	CRIS-A	CIMP-L
MDST8	MSS	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS4	CRIS-D	CIMP4
MIP101	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	NA	CIMP-H
NCIH498	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS1	CRIS-D	CIMP-H
NCIH630	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
NCIH684	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP3
NCIH716	MSS	Other	Other	Other	WT	Other	CMS4	NA	CIMP3
OUMS23	MSS	Other	Other	Hotspot	Nonsense, frameshift	Other	CMS4	CRIS-D	CIMP3
OXCO1	MSS	Other	Other	Hotspot	Nonsense, frameshift	Other	CMS4	CRIS-B	CIMP3
OXCO2	MSI	Other	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-C	CIMP4
OXCO3	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP3
RCM1	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-B	CIMP3
RKO	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	NA	NA	CIMP-H
RW2982	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS3	CRIS-E	CIMP3
RW7213	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS3	CRIS-A	CIMP3
SKCO1	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS3	CRIS-A	CIMP3
SNU1040	MSI	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-D	CIMP4

SNU1047	MSI	Other	Other	Other	Missense, in-frame	WT	NA	CRIS-D	CIMP-L
SNU1181	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP3
SNU1235	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	NA	CIMP-H
SNU1406	MSS	Other	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP-H
SNU1411	MSS	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-H
SNU1460	MSS	Other	Other	Other	WT	WT	CMS2	CRIS-C	CIMP4
SNU1544	MSI	Other	Other	Other	Nonsense, frameshift	WT	CMS1	NA	CIMP4
SNU1684	MSI	Hotspot	Other	Other	Nonsense, frameshift	WT	NA	NA	CIMP4
SNU1746	MSI	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS3	CRIS-A	CIMP3
SNU175	MSI	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS3	CRIS-A	CIMP4
SNU254	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS4	CRIS-B	CIMP3
SNU283	MSS	Other	Other	Other	Missense, in-frame	WT	CMS2	CRIS-C	CIMP3
SNU407	MSI	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	NA	NA	CIMP3
SNU479	MSS	Other	Other	Other	Nonsense, frameshift	Other	CMS2	CRIS-A	CIMP4
SNU503	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP4
SNU61	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS3	CRIS-B	CIMP-L
SNU769B	MSI	Other	Other	Other	Nonsense, frameshift	WT	CMS3	CRIS-A	CIMP4
SNU81	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-L
SNU977	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP3
SNUC1	MSS	Other	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS2	CRIS-C	CIMP4
SNUC2A	MSI	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS4	NA	CIMP4
SNUC2B	MSI	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS3	CRIS-A	CIMP4
SNUC5	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H

SW1116	MSS	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS1	CRIS-B	CIMP-L
SW1222	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS3	CRIS-C	CIMP3
SW1417	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP-L
SW1463	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	NA	CRIS-E	CIMP-L
SW403	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP4
SW48	MSI	Other	Other	Other	Missense, in-frame	WT	CMS1	NA	CIMP-H
SW480	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP3
SW620	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP3
SW837	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS4	CRIS-B	CIMP-L
SW948	MSS	Hotspot	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS2	CRIS-E	CIMP-L
T84	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-C	CIMP3
V411	MSS	Hotspot	Other	Other	Nonsense, frameshift	WT	CMS2	CRIS-A	CIMP-L
V457	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	NA	NA	CIMP3
V481	MSI	Hotspot	Other	Other	Nonsense, frameshift	Other	CMS3	CRIS-A	CIMP3
V703	MSI	Hotspot	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS3	CRIS-C	CIMP4
V9P	MSS	Other	Other	Other	WT	Hotspot, Nonsense, frameshift	CMS4	CRIS-D	CIMP3
VACO400	MSS	Other	Other	Other	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
VACO432	MSI	Other	Other	Hotspot	Nonsense, frameshift	WT	CMS1	NA	CIMP-H
VACO5	MSI	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	NA	CIMP-H
VACO6	MSI	Other	Other	Hotspot	WT	WT	CMS1	CRIS-A	CIMP-H
WIDR	MSS	Other	Other	Hotspot	Nonsense, frameshift	Hotspot, Nonsense, frameshift	CMS1	CRIS-B	CIMP-H

Table 2 - Colorectal cancer (CRC) extreme gene expression outliers

Summary of how the number of genes and samples changes along the different steps of the transcriptome-wide gene expression outlier analysis that we performed in CRC cells.

	Outlier samples		Genes with outlier samples	
	Positive	Negative	Positive	Negative
Outlier calling with Tukey's method	226	226	12,744	6,928
Selection of furthest outliers	226	225	12,744	6,928
Filter on differential expression	226	225	3,843	1,273
Filter on absolute expression	226	223	3,533	965

	Median number of outlier samples per gene		Median number of genes for which a sample is an outlier	
	Positive	Negative	Positive	Negative
Outlier calling with Tukey's method	4	3	484.5	68
Selection of furthest outliers	2	2	84.5	37
Filter on differential expression	2	2	28	9
Filter on absolute expression	2	2	24	7

Table 3 - Differential methylation analysis for genes for which at least two colorectal cancer (CRC) cell lines were found as extreme positive (first table) or negative (second table) outliers.

For each target gene we report:

- (1) Gene ID: Ensembl gene ID of the promoter target gene;
- (2) Gene name: gene symbol of the promoter target gene;
- (3) Probe (Genome segment): ID of the differentially methylated probes in a comma-separated field, with genomic position of the CpG site targeted by the probe with respect to CpG islands (Islands, shores, shelves and open sea) between brackets.

Extreme Positive Outliers:

Gene ID	Gene name	Probes (Genome segment)
ENSG00000006128.12	TAC1	cg01287975 (Island), cg17437939 (Island), cg16288089 (Island)
ENSG00000007264.15	MATK	cg07752026 (Island)
ENSG00000026025.16	VIM	cg26983469 (Island), cg23991622 (Island), cg12874092 (Island), cg27313572 (Island)
ENSG00000064195.7	DLX3	cg24383056 (Island)
ENSG00000073737.16	DHRS9	cg06836480 (OpenSea)
ENSG00000075043.18	KCNQ2	cg08003150 (Island)
ENSG00000082126.18	MPP4	cg20169521 (OpenSea)
ENSG00000082397.17	EPB41L3	cg19579167 (Island), cg23564700 (Island), cg16304950 (Island), cg19275261 (Island)
ENSG00000086205.18	FOLH1	cg25346915 (Island)
ENSG00000087076.9	HSD17B14	cg24725522 (Shore)
ENSG00000090661.12	CERS4	cg06686742 (Shore)
ENSG00000100060.18	MFNG	cg11201532 (OpenSea), cg01765930 (OpenSea)
ENSG00000100557.9	CCDC198	cg14102434 (OpenSea), cg15661409 (OpenSea)
ENSG00000101017.14	CD40	cg22232207 (Island), cg06571407 (Island)
ENSG00000102575.13	ACP5	cg03302259 (Island), cg03719155 (Island), cg04566159 (Island)
ENSG00000105472.13	CLEC11A	cg16651537 (Shore)
ENSG00000107742.13	SPOCK2	cg10856045 (Shore), cg10983208 (Shore), cg26435178 (Shore)
ENSG00000110031.12	LPXN	cg02044895 (Shelf)
ENSG00000111728.10	ST8SIA1	cg10333808 (Island)
ENSG00000111863.12	ADTRP	cg26761744 (OpenSea)
ENSG00000114378.17	HYAL1	cg14943722 (OpenSea), cg12930727 (OpenSea)
ENSG00000115526.11	CHST10	cg03853987 (Island), cg16882226 (Island), cg19283506 (Island)
ENSG00000116962.15	NID1	cg02957057 (Island)
ENSG00000120669.16	SOHLH2	cg24710938 (Island)
ENSG00000121068.14	TBX2	cg12163132 (Island)
ENSG00000121413.12	ZSCAN18	cg25784220 (Island), cg22721334 (Island)
ENSG00000122691.13	TWIST1	cg20052718 (Island), cg26818735 (Island)
ENSG00000123364.5	HOXC13	cg17085352 (Shore), cg22830707 (Shore)
ENSG00000123388.4	HOXC11	cg03132556 (Shore), cg07697238 (Shore), cg26386624 (Island), cg08962452 (Island), cg06630413 (Island)
ENSG00000124785.9	NRN1	cg08545493 (Island), cg25829425 (Island), cg13021333 (Island), cg24720571 (Island)
ENSG00000125207.7	PIWIL1	cg02382037 (Island), cg23887609 (Island), cg27147004 (Island), cg13900773 (Island), cg23548151 (Island), cg26677194 (Island), cg24838063 (Island)
ENSG00000128918.15	ALDH1A2	cg07721203 (Island), cg13198321 (Island), cg14805347 (Island)
ENSG00000129194.8	SOX15	cg07488259 (Shore), cg01029592 (Shore)
ENSG00000132639.12	SNAP25	cg16564824 (Shore), cg23328622 (Shore)
ENSG00000133104.14	SPART	cg18755783 (Island), cg00947032 (Island), cg16741041 (Island), cg13996405 (Island)
ENSG00000134802.17	SLC43A3	cg19167683 (Island)
ENSG00000134873.10	CLDN10	cg10305311 (Island), cg16275739 (Island), cg16232183 (Island), cg09469554 (Island)
ENSG00000135220.11	UGT2A3	cg00149061 (OpenSea)
ENSG00000135374.10	ELF5	cg10987788 (OpenSea), cg21017775 (OpenSea), cg11343506 (OpenSea), cg22731981 (OpenSea)
ENSG00000135454.14	B4GALNT1	cg02264922 (Island)
ENSG00000135480.16	KRT7	cg08975803 (Island), cg22813430 (Shore), cg09522147 (Island), cg14537533 (Island), cg25313172 (Island), cg14482313 (Island), cg07967679 (Island), cg09670128 (Island)
ENSG00000135702.14	CHST5	cg03858460 (OpenSea)
ENSG00000136305.11	CIDEB	cg25576711 (Island), cg07760773 (Island), cg15364618 (Island), cg13840968 (Island), cg21886367 (Island), cg10193721 (Island), cg26310551 (Island), cg06796435 (Island), cg07509935 (Island), cg06823034 (Island), cg20007021 (Island)
ENSG00000138829.12	FBN2	cg11965976 (Island), cg25084878 (Island), cg17958315 (Island)
ENSG00000139910.20	NOVA1	cg02014003 (OpenSea), cg07559273 (OpenSea), cg19832521 (OpenSea), cg18488855 (OpenSea), cg20478129 (OpenSea)
ENSG00000140287.11	HDC	cg02329886 (OpenSea)
ENSG00000140931.20	CMTM3	cg08601917 (Island), cg06445928 (Island), cg01434608 (Shore), cg09233013 (Island), cg00355909 (Island), cg08682544 (Island), cg26560414 (Island)
ENSG00000143995.20	MEIS1	cg19572637 (Shore)
ENSG00000145623.13	OSMR	cg15599832 (Shore), cg19609242 (Island), cg17528648 (Island)
ENSG00000145908.12	ZNF300	cg18237551 (Island), cg08580836 (Island), cg02343823 (Island), cg04675542 (Island)

ENSG00000151650.8	VENTX	cg02317400 (Island)
ENSG00000151778.11	SERP2	cg22667178 (Island), cg12379940 (Island)
ENSG00000154027.19	AK5	cg22075924 (OpenSea)
ENSG00000154277.13	UCHL1	cg24715245 (Island), cg15032098 (Shore), cg16142306 (Island)
ENSG00000154760.14	SLFN13	cg24955845 (Island), cg22588307 (Shore), cg16563370 (Shore), cg13443575 (Shore), cg04354393 (Shore)
ENSG00000156127.8	BATF	cg15645309 (OpenSea), cg21531300 (OpenSea), cg14424070 (OpenSea)
ENSG00000159182.5	PRAC1	cg00960395 (Shore), cg14230397 (Shore), cg01543654 (Island), cg20945566 (Shore), cg27170782 (Shore), cg04618333 (Island)
ENSG00000160963.14	COL26A1	cg18762727 (Island), cg14874750 (Island), cg27528510 (Island), cg23424003 (Island)
ENSG00000163064.7	EN1	cg00654888 (Island), cg15724256 (Island), cg17167253 (Island), cg20248516 (Island)
ENSG00000163430.12	FSTL1	cg15109851 (Island)
ENSG00000164742.16	ADCY1	cg26459372 (Island)
ENSG00000165443.12	PHYHIPL	cg02662277 (Island), cg26014391 (Island), cg24563570 (Island), cg23712342 (Island)
ENSG00000165794.10	SLC39A2	cg21853693 (OpenSea)
ENSG00000165863.17	C10orf82	cg07945733 (Island)
ENSG00000166035.11	LIPC	cg01733599 (OpenSea)
ENSG00000166402.8	TUB	cg01084239 (Island), cg02629281 (Island)
ENSG00000166682.13	TMPRSS5	cg18165914 (OpenSea)
ENSG00000167244.21	IGF2	cg26517849 (Island)
ENSG00000167555.14	ZNF528	cg03077453 (Shore), cg12950007 (Island), cg27258025 (Island), cg03091337 (Island)
ENSG00000167874.7	TMEM88	cg21845633 (Shore), cg23404101 (Shore), cg23671708 (Shore), cg02459416 (Shore), cg03854722 (Shore), cg15853812 (Shore), cg03719978 (Shore)
ENSG00000168071.22	CCDC88B	cg08024887 (Shore), cg04422903 (Shore), cg16718999 (Shore)
ENSG00000169439.12	SDC2	cg14538332 (Island), cg13096260 (Island)
ENSG00000169696.16	ASPCR1	cg23928512 (Shelf), cg12374579 (Shelf), cg11780549 (Shore), cg21384021 (Island)
ENSG00000169876.14	MUC17	cg17981101 (OpenSea)
ENSG00000170370.12	EMX2	cg19358493 (Shore)
ENSG00000170627.11	GTSF1	cg06809675 (OpenSea), cg07875328 (OpenSea)
ENSG00000170743.17	SYT9	cg22723056 (Island), cg03226737 (Island), cg18560328 (Island)
ENSG00000172716.16	SLFN11	cg03668718 (Island)
ENSG00000173917.10	HOXB2	cg17807663 (Shore), cg09313705 (Shore), cg25882366 (Shore), cg22807449 (Shore), cg22777724 (Shore), cg23530553 (Shore), cg10788618 (Shore), cg12121660 (Shore), cg26841048 (Shore)
ENSG00000175426.11	PCSK1	cg09786257 (Shore)
ENSG00000175544.13	CABP4	cg20022862 (OpenSea)
ENSG00000178363.5	CALML3	cg24831879 (Shore)
ENSG00000178445.9	GLDC	cg14538238 (Shore), cg11747771 (Island)
ENSG00000178878.12	APOLD1	cg26244225 (Shore), cg07091544 (Shore), cg02813863 (Shore), cg16747828 (Shore), cg10048215 (Shore), cg10635645 (Shore)
ENSG00000181234.9	TMEM132C	cg03530754 (Island)
ENSG00000182809.10	CRIP2	cg01564661 (Shore)
ENSG00000185818.8	NAT8L	cg06913330 (Island), cg21269934 (Island), cg19300568 (Island)
ENSG00000188958.9	UTS2B	cg12336960 (Island)
ENSG00000196628.18	TCF4	cg19597776 (Shore)
ENSG00000196917.6	HCAR1	cg26643476 (OpenSea)
ENSG00000197520.10	FAM177B	cg18665384 (OpenSea)
ENSG00000198203.10	SULT1C2	cg13968390 (OpenSea), cg17287107 (OpenSea), cg25570328 (OpenSea), cg03681335 (OpenSea)
ENSG00000198835.4	GJC2	cg03065165 (OpenSea)
ENSG00000212724.3	KRTAP2-3	cg00345704 (OpenSea)
ENSG00000213366.13	GSTM2	cg06970744 (Shore), cg16670497 (Island), cg03942855 (Island)
ENSG00000229637.4	PRAC2	cg01543654 (Island), cg04618333 (Island)
ENSG00000235098.8	ANKRD65	cg15081886 (Shore), cg15633699 (Shore)
ENSG00000237651.6	C2orf74	cg01648237 (OpenSea), cg05037927 (OpenSea), cg23405039 (OpenSea), cg18158151 (OpenSea), cg19790321 (OpenSea), cg16328106 (OpenSea), cg24757310 (OpenSea)
ENSG00000243709.1	LEFTY1	cg25552768 (Shore), cg14264773 (Shore)
ENSG00000244486.8	SCARF2	cg15243570 (Island), cg14998335 (Island), cg03448766 (Island)
ENSG00000261857.7	MIA	cg25152942 (Shelf), cg24083702 (Shelf)
ENSG00000277586.2	NEFL	cg01614020 (Island)

Extreme Negative Outliers:

Gene ID	Gene name	Probes (Genome segment)
ENSG0000005882.12	PKD2	cg20133730 (Shore)
ENSG0000008323.15	PLEKHG6	cg17181255 (Shore), cg27150896 (Shore), cg24333145 (Shore), cg14783987 (Shore)
ENSG00000012660.14	ELOVL5	cg10369451 (Island), cg21195414 (Island), cg03998871 (Island), cg01865606 (Shore), cg13636408 (Island), cg00024396 (Island), cg15133313 (Island), cg20397543 (Shore), cg24524396 (Island), cg20419072 (Island)
ENSG00000013588.8	GPRC5A	cg19737972 (Island), cg03353445 (Shore), cg10198671 (Island)
ENSG00000016402.13	IL20RA	cg04481779 (Island), cg26864905 (Shore), cg15992935 (Shore), cg22487322 (Shore), cg13714540 (Shore), cg24405068 (Shore), cg27656754 (Shore)
ENSG00000021355.13	SERPINB1	cg00031256 (Island), cg19349369 (Island), cg27056119 (Island), cg08542865 (Island)
ENSG00000037757.14	MRI1	cg14272688 (Island), cg01530988 (Island), cg24115922 (Island), cg18631292 (Island), cg25755428 (Island), cg16188681 (Island), cg16474696 (Shore), cg24025995 (Island)
ENSG00000042493.16	CAPG	cg21654383 (Island), cg02242344 (Shore), cg16838838 (Island), cg09337254 (Shore), cg26476820 (Shelf), cg19627006 (Shelf), cg16437908 (Shore)
ENSG00000051523.11	CYBA	cg04879832 (Island), cg00054525 (Shore), cg26537639 (Island), cg09025501 (Shore), cg03787864 (Island), cg24046411 (Shore), cg08843517 (Shore), cg10246531 (Shore)
ENSG00000052344.16	PRSS8	cg27436259 (OpenSea)
ENSG00000056972.20	TRAF3IP2	cg26904406 (OpenSea)
ENSG00000065833.9	ME1	cg27138281 (Island), cg05387146 (Island), cg16701167 (Shore), cg02366534 (Island), cg01121127 (Island), cg06836736 (Island), cg24881834 (Island)
ENSG00000068079.7	IFI35	cg02452732 (OpenSea)
ENSG00000070087.14	PFN2	cg13715862 (Shore)
ENSG00000076242.15	MLH1	cg25837710 (Island), cg07101782 (Island), cg17621259 (Island), cg06590608 (Island), cg05906740 (Island), cg10769891 (Island), cg14598950 (Island), cg23658326 (Island), cg11224603 (Island), cg16433211 (Island), cg27586588 (Island), cg03497419 (Island), cg12851504 (Island), cg06108510 (Island), cg21490561 (Island), cg07064226 (Island), cg16764580 (Island), cg03192963 (Island), cg24985459 (Island), cg19132762 (Island), cg02279071 (Shore), cg25202636 (Island), cg00893636 (Island), cg13846866 (Shore), cg12790037 (Island), cg27331401 (Island), cg01302270 (Island), cg11600697 (Island), cg14751544 (Shore), cg17641046 (Island), cg03901257 (Shore), cg05845319 (Shore)
ENSG00000078114.18	NEBL	cg11787785 (Island), cg23565942 (Island), cg17426273 (Island), cg27495903 (Island), cg01960926 (Island), cg18192417 (Island), cg06731125 (Island), cg24625388 (Shore)
ENSG00000081923.14	ATP8B1	cg21357291 (Island), cg16310513 (Island)
ENSG00000083307.12	GRHL2	cg23973429 (Island), cg20574490 (Island), cg14202477 (Shore), cg17885542 (Shore), cg25330243 (Island)
ENSG00000084207.18	GSTP1	cg04920951 (Island), cg06928838 (Island), cg09038676 (Island), cg11566244 (Island), cg26250609 (Island), cg05244766 (Shore)
ENSG00000087077.14	TRIP6	cg12507655 (Shore), cg01008256 (Island), cg04742719 (Island)
ENSG00000087253.13	LPCAT2	cg27458265 (Island)
ENSG00000089335.21	ZNF302	cg02991799 (Island), cg09567983 (Island), cg08336861 (Island), cg10181615 (Island), cg13337697 (Island), cg19850565 (Island), cg19508683 (Island), cg09106556 (Island)
ENSG00000090238.12	YPEL3	cg04895065 (Island)
ENSG00000092621.12	PHGDH	cg00426921 (Island), cg19096424 (Island), cg26791905 (Shore)
ENSG00000093144.19	ECHDC1	cg13817182 (Shore)
ENSG00000095261.14	PSMD5	cg13460057 (Island), cg13959647 (Island), cg09586646 (Island), cg09833538 (Island), cg09419670 (Shore)
ENSG00000100379.17	KCTD17	cg03995567 (Island), cg25061914 (Shore)
ENSG00000100422.14	CERK	cg24359836 (Island), cg05698090 (Shore)
ENSG00000100558.9	PLEK2	cg24995657 (Island), cg26225496 (Island), cg18306861 (Island)
ENSG00000101076.16	HNF4A	cg27420224 (OpenSea)
ENSG00000101213.7	PTK6	cg14010720 (Shore)
ENSG00000102890.16	ELMO3	cg25341653 (Island), cg26178184 (Shore), cg19514469 (Island)

ENSG00000103067.14	ESRP2	cg00944421 (Island), cg06723863 (Island), cg04513006 (Island), cg04079760 (Island), cg22190721 (Island), cg08694699 (Island), cg04328477 (Island), cg20264732 (Island), cg08886823 (Island), cg01660934 (Island), cg16562724 (Island)
ENSG00000104413.18	ESRP1	cg16648355 (Island)
ENSG00000104518.11	GSDMD	cg12686110 (Island), cg09317036 (Island), cg07422279 (Island), cg26711732 (Shore), cg23090207 (Island), cg14721632 (Island), cg01134012 (Island), cg18145080 (Island)
ENSG00000104611.12	SH2D4A	cg12794421 (Island)
ENSG00000104870.13	FCGRT	cg07127225 (Island), cg10405557 (Island)
ENSG00000106484.16	MEST	cg16511229 (Island), cg02169391 (Island), cg08077673 (Island), cg15164103 (Shore), cg02490034 (Shore), cg22115706 (Shore), cg15150348 (Shore), cg18100127 (Shore)
ENSG00000106541.12	AGR2	cg12565635 (OpenSea)
ENSG00000106853.20	PTGR1	cg13831329 (OpenSea)
ENSG00000108375.12	RNF43	cg24835159 (OpenSea)
ENSG00000108384.15	RAD51C	cg27221688 (Island), cg10487724 (Island), cg02118635 (Island), cg05214530 (Island)
ENSG00000108511.10	HOXB6	cg19828220 (Shelf), cg10906729 (Shelf), cg18684142 (Shelf)
ENSG00000110719.10	TCIRG1	cg11990562 (Island), cg08162803 (Island), cg11731626 (Island), cg06279471 (Shore)
ENSG00000111319.13	SCNN1A	cg01807820 (Island), cg11785537 (OpenSea), cg03196720 (Island), cg12510724 (OpenSea), cg06029700 (Island), cg23245485 (OpenSea)
ENSG00000111321.11	LTBR	cg07648238 (Island), cg08740698 (Island), cg23079808 (Island), cg26306994 (Island), cg10452531 (Island), cg23245485 (OpenSea), cg19476647 (Island)
ENSG00000111679.17	PTPN6	cg12690127 (Shore), cg25719851 (Shore), cg15975802 (Shore)
ENSG00000112559.14	MDFI	cg17094014 (Island), cg14086013 (Island)
ENSG00000112763.17	BTN2A1	cg23371570 (Shore)
ENSG00000112851.14	ERBIN	cg04823499 (Island), cg05465766 (Island), cg24133276 (Island), cg05458841 (Island), cg22009138 (Shore), cg16560309 (Island), cg01598222 (Island), cg07172256 (Island), cg20274746 (Island)
ENSG00000112992.18	NNT	cg03902565 (Island), cg00584670 (Island), cg26233374 (Island), cg09489281 (Shore), cg12626968 (Island)
ENSG00000114631.11	PODXL2	cg12943155 (Island), cg07558287 (Island), cg10451772 (Island)
ENSG00000118418.14	HMGH3	cg17832805 (Island), cg14080940 (Island)
ENSG00000118640.11	VAMP8	cg05486094 (OpenSea), cg04877910 (OpenSea)
ENSG00000119673.14	ACOT2	cg01498900 (Shore)
ENSG00000120451.10	SNX19	cg05493593 (Island), cg05974695 (Island), cg23642342 (Island), cg12179176 (Island), cg05962950 (Island), cg22079354 (Shore)
ENSG00000120675.6	DNAJC15	cg09677945 (Island)
ENSG00000121310.17	ECHDC2	cg09128567 (Island), cg06600287 (Shore), cg05232215 (Shore)
ENSG00000121931.16	LRIF1	cg11787288 (Island), cg23658744 (Island), cg01011619 (Island), cg09319202 (Island), cg15139482 (Island), cg19641455 (Island), cg16524830 (Island), cg04879281 (Island)
ENSG00000124762.13	CDKN1A	cg24425727 (Shore)
ENSG00000124839.13	RAB17	cg01157280 (OpenSea)
ENSG00000125378.16	BMP4	cg08046044 (Shore)
ENSG00000125611.15	CHCHD5	cg21934564 (Island), cg24706932 (Island), cg27310163 (Island), cg25521086 (Island), cg11528832 (Island), cg26236440 (Island), cg24651706 (Island), cg00499289 (Island)
ENSG00000127399.15	LRR61	cg06975979 (Island)
ENSG00000127824.14	TUBA4A	cg11348442 (Shore)
ENSG00000128833.13	MYO5C	cg27633342 (Shore)
ENSG00000131746.13	TNS4	cg20468081 (OpenSea)
ENSG00000132481.7	TRIM47	cg09253696 (Island), cg06407111 (Shore)
ENSG00000132698.15	RAB25	cg09243900 (OpenSea), cg15896939 (OpenSea)
ENSG00000133678.14	TMEM254	cg10577284 (Island), cg17251097 (Island), cg02599225 (Island), cg05377034 (Shore), cg25315252 (Island)
ENSG00000133943.20	DGLUCY	cg23296408 (OpenSea), cg26288577 (OpenSea), cg12563839 (OpenSea), cg15442728 (OpenSea)
ENSG00000134330.19	IAH1	cg05986044 (Island), cg17568255 (Island), cg27494615 (Island), cg01417553 (Island), cg20556744 (Island), cg12285409 (Island), cg04210100 (Island), cg08047252 (Island)
ENSG00000135069.14	PSAT1	cg13828183 (Island)
ENSG00000135457.10	TFCP2	cg22790973 (Island), cg11297723 (Island), cg18751588 (Island), cg20392585 (Island), cg13113115 (Island), cg19016289 (Island), cg20927059 (Shore)

ENSG00000135821.18	GLUL	cg13086581 (Island), cg03970588 (Island), cg01419479 (Island), cg09470059 (Island), cg13180005 (Island)
ENSG00000137817.17	PARP6	cg25274052 (Island), cg27337534 (Shore), cg02883230 (Shore)
ENSG00000138771.16	SHROOM3	cg01477861 (Shore)
ENSG00000139178.11	C1RL	cg01726775 (OpenSea)
ENSG00000139629.16	GALNT6	cg25360385 (Shore)
ENSG00000140403.12	DNAJA4	cg22935921 (Island), cg05392364 (Island), cg17246382 (Island), cg16358679 (Island), cg09214398 (Island), cg01786994 (Island), cg08594606 (Island), cg11019923 (Island), cg25025992 (Shore), cg23366752 (Island)
ENSG00000141526.16	SLC16A3	cg08296680 (Island), cg04184297 (Island), cg19590707 (Island)
ENSG00000141738.14	GRB7	cg14263391 (OpenSea), cg08284496 (OpenSea), cg07327299 (OpenSea), cg25038082 (OpenSea), cg03684977 (OpenSea), cg00735611 (OpenSea), cg14398691 (OpenSea), cg14575854 (OpenSea), cg17355719 (OpenSea)
ENSG00000141753.7	IGFBP4	cg11961138 (Island)
ENSG00000142089.16	IFITM3	cg06046490 (Shore)
ENSG00000142273.13	CBLC	cg03292149 (Shore), cg01057656 (Island), cg21211730 (Shore)
ENSG00000142675.18	CNKR1	cg09890400 (OpenSea), cg26412358 (OpenSea), cg17330765 (OpenSea), cg00300879 (OpenSea)
ENSG00000142910.16	TINAGL1	cg15079885 (OpenSea)
ENSG00000144645.14	OSBPL10	cg15840985 (Shore)
ENSG00000145247.12	OCIAD2	cg26731119 (Island), cg23545105 (Island)
ENSG00000145868.16	FBXO38	cg15536575 (Island), cg12647030 (Island), cg05176146 (Island), cg08592146 (Island), cg23459010 (Island), cg02680629 (Island), cg13965321 (Shore), cg17574857 (Island), cg06803897 (Shore), cg25847835 (Shore), cg21638374 (Island)
ENSG00000146281.6	PM20D2	cg20224813 (Island), cg10496915 (Island)
ENSG00000147804.9	SLC39A4	cg14228592 (Island)
ENSG00000148803.12	FUOM	cg19306063 (Island), cg13336642 (Island), cg12251804 (Island), cg18043267 (Island)
ENSG00000149418.11	ST14	cg26021246 (Shore), cg21029769 (Shore), cg23250795 (Shore)
ENSG00000151287.17	TEX30	cg26269109 (Island), cg06245962 (Island), cg06457396 (Island), cg02927949 (Island), cg14767402 (Island), cg20956131 (Island), cg13675958 (Island)
ENSG00000151376.16	ME3	cg17080504 (Island), cg03953500 (Island), cg10160397 (Island), cg00474723 (Island), cg02493602 (Shore)
ENSG00000151689.13	INPP1	cg15635368 (Shore), cg03345805 (Shore), cg25753817 (Shore)
ENSG00000153130.17	SCOC	cg02653612 (Shore), cg24810741 (Island), cg04220088 (Shore), cg07089235 (Shore), cg17666096 (Island), cg23743449 (Shore), cg04401710 (Shore), cg02945007 (Island), cg17333211 (Shore), cg05576648 (Shore), cg02182440 (Shore)
ENSG00000153406.13	NMRAL1	cg07160937 (Island), cg10063637 (Island), cg00688539 (Shore)
ENSG00000156711.17	MAPK13	cg05859264 (Shore)
ENSG00000157992.12	KRTCAP3	cg21248554 (Island), cg20102877 (Island), cg12000995 (Island), cg26034919 (Shore), cg17158414 (Island), cg11618577 (Island), cg12648201 (Island), cg04845466 (Island)
ENSG00000159166.15	LAD1	cg25947945 (Island), cg10095226 (Island), cg21964551 (Island), cg02838178 (Island), cg27425612 (Island), cg23715582 (Island), cg26539023 (Island), cg07561547 (Island)
ENSG00000159176.14	CSRP1	cg13554744 (Island)
ENSG00000159228.13	CBR1	cg27307781 (Island), cg01962969 (Shore)
ENSG00000159445.13	THEM4	cg03147210 (Island), cg23231729 (Island), cg24540521 (Island), cg05927068 (Island)
ENSG00000161267.12	BDH1	cg00177496 (Island)
ENSG00000162639.16	HENMT1	cg21144922 (Island), cg27587125 (Island), cg24737783 (Island)
ENSG00000162777.17	DENND2D	cg25722212 (Shore), cg14519917 (Island)
ENSG00000163435.16	ELF3	cg26328757 (OpenSea)
ENSG00000163694.15	RBM47	cg16565913 (Island), cg15967525 (OpenSea), cg27076223 (Shore), cg10019467 (Shore), cg08949974 (Shore), cg25117091 (OpenSea), cg20435896 (Shore)
ENSG00000163701.19	IL17RE	cg18738581 (OpenSea)
ENSG00000163728.11	TTC14	cg20296001 (Island), cg01335781 (Island), cg16496895 (Island), cg14930059 (Island), cg10459387 (Shore)
ENSG00000163993.7	S100P	cg27027375 (OpenSea), cg07210669 (OpenSea), cg14140379 (OpenSea), cg14900031 (OpenSea), cg25083732 (OpenSea), cg14323984 (OpenSea)
ENSG00000164078.13	MST1R	cg18129748 (Shore), cg03332271 (Island), cg27360567 (Shore)
ENSG00000164237.9	CMBL	cg05593349 (Island), cg13323825 (Island), cg10178203 (Shore)
ENSG00000164403.14	SHROOM1	cg19793809 (Island)

ENSG00000164904.18	ALDH7A1	cg01814098 (Island), cg08675664 (Island), cg25476565 (Shore)
ENSG00000164930.12	FZD6	cg09042026 (Island), cg08798921 (Island), cg21975377 (Island), cg24115571 (Shore)
ENSG00000165105.10	RASEF	cg10319505 (Island), cg12969644 (Island)
ENSG00000165512.5	ZNF22	cg14539394 (Island), cg23456144 (Shore)
ENSG00000165795.23	NDRG2	cg13141192 (Island)
ENSG00000166145.14	SPINT1	cg22491225 (Island), cg12426344 (Island), cg26531804 (Island), cg27046604 (Island)
ENSG00000166199.13	ALKBH3	cg03342084 (Island), cg20009641 (Island), cg22637507 (Island), cg15593337 (Island)
ENSG00000166557.13	TMED3	cg05104048 (Island), cg04963535 (Island), cg21164303 (Island), cg05014452 (Shore)
ENSG00000166801.15	FAM111A	cg06770568 (Shore)
ENSG00000167608.11	TMC4	cg25736482 (OpenSea)
ENSG00000167644.12	C19orf33	cg27174108 (OpenSea), cg00412772 (OpenSea)
ENSG00000167785.9	ZNF558	cg03663659 (Shore)
ENSG00000167880.7	EVPL	cg18657389 (OpenSea)
ENSG00000167925.16	GHDC	cg06306927 (OpenSea), cg23509272 (OpenSea)
ENSG00000168404.13	MLKL	cg26634130 (Shore)
ENSG00000168758.11	SEMA4C	cg07166409 (Shore)
ENSG00000168765.17	GSTM4	cg01564068 (Island)
ENSG00000168785.8	TSPAN5	cg27654505 (Island), cg00622677 (Island), cg18823647 (Shore)
ENSG00000169174.11	PCSK9	cg17167852 (Island), cg09072162 (Island), cg19249107 (Island), cg05118916 (Island), cg14993491 (Shore)
ENSG00000169220.18	RGS14	cg17509989 (Shelf), cg16006841 (Shelf)
ENSG00000170190.16	SLC16A5	cg17733616 (Island), cg04305621 (Island), cg09300114 (Island), cg04529938 (Island)
ENSG00000170430.10	MGMT	cg12434587 (Island), cg00618725 (Island)
ENSG00000171130.18	ATP6V0E2	cg12022895 (Island), cg27425719 (Island), cg26039848 (Island), cg04817271 (Island), cg00934312 (Island), cg03574289 (Island), cg16480132 (Island), cg15933823 (Island), cg08767710 (Island), cg18396041 (Island), cg03496713 (Island), cg24227081 (Shore), cg04600299 (Island), cg06961439 (Shore)
ENSG00000171208.10	NETO2	cg06364315 (OpenSea), cg04660816 (OpenSea), cg01122672 (OpenSea), cg07973470 (OpenSea), cg04203758 (OpenSea), cg18867659 (OpenSea)
ENSG00000171345.13	KRT19	cg08966188 (Shore)
ENSG00000172164.15	SNTB1	cg08598221 (Shore)
ENSG00000172818.10	OVOL1	cg08296831 (Island), cg01097872 (OpenSea), cg13229961 (OpenSea)
ENSG00000172936.14	MYD88	cg06239037 (Island), cg15348640 (Island), cg27043630 (Island), cg06069310 (Island), cg01353464 (Island)
ENSG00000173221.14	GLRX	cg03852144 (OpenSea), cg16677191 (OpenSea)
ENSG00000173530.6	TNFRSF10D	cg22783363 (Island), cg04134048 (Shore), cg15244327 (Shore)
ENSG00000173702.7	MUC13	cg22907034 (OpenSea), cg14497879 (OpenSea)
ENSG00000174652.19	ZNF266	cg01329093 (Island)
ENSG00000174791.11	RIN1	cg05049335 (Shore)
ENSG00000175701.10	MTLN	cg09068086 (Island)
ENSG00000175793.12	SFN	cg21950166 (Shore), cg14333454 (Shore), cg16549596 (Shore)
ENSG00000176153.12	GPX2	cg09643186 (OpenSea), cg13844922 (OpenSea)
ENSG00000176920.13	FUT2	cg20441902 (Shore), cg08873673 (Shore), cg15197218 (Shelf)
ENSG00000177119.16	ANO6	cg10609068 (Shore), cg21705221 (Island), cg07054062 (Island), cg13582818 (Shore), cg22441683 (Island)
ENSG00000177311.11	ZBTB38	cg13029400 (OpenSea)
ENSG00000178078.11	STAP2	cg12229841 (OpenSea), cg15797834 (OpenSea)
ENSG00000178096.9	BOLA1	cg06605933 (Shore), cg20164915 (Island)
ENSG00000178234.13	GALNT11	cg05712610 (Shore), cg27434954 (Island), cg16586442 (Island), cg02159402 (Island), cg15146720 (Island), cg14618638 (Shore), cg05375487 (Shore)
ENSG00000178467.18	P4HTM	cg17527142 (Island), cg13965724 (Island), cg25264370 (Island), cg15034779 (Island), cg21197973 (Island), cg16035189 (Island), cg05321398 (Shore), cg03791385 (Island), cg07707611 (Island), cg20856064 (Island), cg06679760 (Island), cg13432286 (Shore), cg21574435 (Island), cg09367648 (Island)
ENSG00000178718.6	RPP25	cg09619786 (Island), cg09549240 (Shore)
ENSG00000178826.11	TMEM139	cg15652212 (Shelf), cg26511386 (Shelf), cg22531018 (Shelf), cg21852589 (Shelf)
ENSG00000181481.14	RNF135	cg09811317 (Shore), cg25821304 (Shore), cg18213979 (Shore), cg15258416 (Shore), cg11857033 (Shore), cg20469139 (Shore), cg18459618 (Shore)

Table 4 - Enrichment of samples that are positive for the CpG island methylator phenotype (CIMP) among the extreme negative outliers of a single gene.

For each tested gene we report:

- (1) Gene ID: Ensembl gene ID;
- (2) Gene name: gene symbol;
- (3) # outliers: number of extreme negative outliers;
- (4-7) CIMP-H, CIMP-L, CIMP3, CIMP4: number of extreme negative outliers for each CIMP class;
- (8) p-value: hypergeometric test p-value;
- (9) FDR: False Discovery Rate.

Gene ID	Gene name	# outliers	CIMP-H	CIMP-L	CIMP3	CIMP4	p-value	FDR
ENSG00000151287.17	TEX30	9	9	0	0	0	0	0
ENSG00000087253.13	LPCAT2	8	7	1	0	0	0	0
ENSG00000163728.11	TTC14	6	6	0	0	0	0	0
ENSG00000118418.14	HMG3	4	3	1	0	0	0	0
ENSG00000037757.14	MRI1	3	3	0	0	0	0	0
ENSG00000078114.18	NEBL	3	3	0	0	0	0	0
ENSG00000159228.13	CBR1	3	1	2	0	0	0	0
ENSG00000178096.9	BOLA1	3	3	0	0	0	0	0
ENSG00000255302.5	EID1	3	3	0	0	0	0	0
ENSG00000104870.13	FCGRT	3	3	0	0	0	0	0
ENSG00000112851.14	ERBIN	2	2	0	0	0	0	0
ENSG00000145868.16	FBXO38	2	0	2	0	0	0	0
ENSG00000178234.13	GALNT11	2	2	0	0	0	0	0
ENSG00000165795.23	NDRG2	2	2	0	0	0	0	0
ENSG00000189190.10	ZNF600	2	2	0	0	0	0	0
ENSG00000012660.14	ELOVL5	36	26	8	2	0	2,0166E-14	3,69038E-12
ENSG00000076242.15	MLH1	31	23	6	2	0	4,70606E-12	8,56504E-10
ENSG00000162639.16	HENMT1	26	19	3	1	3	2,86459E-07	5,1849E-05
ENSG00000185842.15	DNAH14	20	15	2	2	1	4,82039E-06	0,00086767
ENSG00000177119.16	ANO6	11	5	5	1	0	5,0803E-05	0,009093742
ENSG00000121931.16	LRIF1	24	11	7	3	3	0,00010925	0,019446493
ENSG00000089335.21	ZNF302	10	9	0	0	1	0,000129099	0,022850604
ENSG00000148803.12	FUOM	9	8	0	0	1	0,000325751	0,057332167
ENSG00000181666.18	ZNF875	9	4	4	1	0	0,000325751	0,057332167
ENSG00000171208.10	NETO2	19	11	3	4	1	0,000752389	0,130915764
ENSG00000153130.17	SCOC	8	5	2	0	1	0,00081625	0,141211165
ENSG00000171130.18	ATP6V0E2	7	5	1	1	0	0,002031348	0,349391892
ENSG00000114631.11	PODXL2	9	4	3	1	1	0,004740238	0,810580693
ENSG00000092621.12	PHGDH	6	4	1	1	0	0,00502131	0,853622731
ENSG00000170430.10	MGMT	41	14	5	8	14	0,213583269	1
ENSG00000198131.14	ZNF544	21	4	7	5	5	0,107980042	1
ENSG00000137817.17	PARP6	20	6	2	7	5	0,478264443	1
ENSG00000198758.10	EPS8L3	19	9	0	2	8	0,230066597	1
ENSG00000173702.7	MUC13	16	8	0	1	7	0,175340047	1
ENSG00000101076.16	HNF4A	20	6	1	5	8	0,662860543	1
ENSG00000133943.20	DGLUCY	15	2	2	3	8	0,835447429	1
ENSG00000198113.3	TOR4A	17	0	2	4	11	0,993744286	1
ENSG00000052344.16	PRSS8	14	3	2	4	5	0,579693572	1
ENSG00000267795.6	SMIM22	13	6	0	1	6	0,272306388	1
ENSG00000246705.4	H2AJ	13	7	0	4	2	0,119962889	1
ENSG00000178826.11	TMEM139	13	2	1	1	9	0,873395553	1
ENSG00000140403.12	DNAJA4	13	2	0	4	7	0,961461413	1
ENSG00000182107.7	TMEM30B	13	4	1	3	5	0,486271977	1
ENSG00000165512.5	ZNF22	17	8	3	2	4	0,013354117	1
ENSG00000189334.9	S100A14	12	4	0	4	4	0,622332146	1
ENSG00000016402.13	IL20RA	11	0	1	5	5	0,98010894	1
ENSG00000083307.12	GRHL2	10	2	1	3	4	0,671593723	1
ENSG00000132698.15	RAB25	10	2	1	3	4	0,671593723	1
ENSG00000169174.11	PCSK9	10	5	0	0	5	0,197106212	1
ENSG00000213614.9	HEXA	10	1	2	3	4	0,671593723	1
ENSG00000051523.11	CYBA	10	0	3	5	2	0,671593723	1
ENSG00000142273.13	CBLC	10	1	1	3	5	0,868946004	1
ENSG00000164237.9	CMBL	10	1	0	1	8	0,967622145	1
ENSG00000119673.14	ACOT2	12	2	3	5	2	0,387518628	1
ENSG00000131746.13	TNS4	10	1	0	4	5	0,967622145	1
ENSG00000181481.14	RNF135	9	0	2	7	0	0,80974032	1
ENSG00000112992.18	NNT	9	3	1	2	3	0,307255098	1
ENSG00000197142.10	ACSL5	14	4	2	1	7	0,36170985	1
ENSG00000163993.7	S100P	9	4	1	1	3	0,119047946	1
ENSG00000124839.13	RAB17	9	0	0	2	7	0,993445901	1
ENSG00000104413.18	ESRP1	8	2	1	1	4	0,453178837	1
ENSG00000204610.13	TRIM15	8	1	0	0	7	0,917187673	1
ENSG00000172818.10	OVOL1	8	2	0	2	4	0,729799489	1
ENSG00000162777.17	DENND2D	10	2	0	2	6	0,868946004	1
ENSG00000100422.14	CERK	14	7	1	2	4	0,072757861	1
ENSG00000164078.13	MST1R	7	0	0	2	5	0,979484276	1

ENSG00000167880.7	EVPL	7	0	0	3	4	0,979484276	1
ENSG00000120675.6	DNAJC15	8	2	3	1	2	0,060233211	1
ENSG00000163701.19	IL17RE	7	2	0	0	5	0,626066744	1
ENSG00000163694.15	RBM47	7	1	0	0	6	0,870340627	1
ENSG00000133678.14	TMEM254	7	5	0	2	0	0,022961082	1
ENSG00000106541.12	AGR2	7	3	0	1	3	0,327937203	1
ENSG00000176920.13	FUT2	7	3	0	2	2	0,327937203	1
ENSG00000159166.15	LAD1	6	1	0	1	4	0,800548089	1
ENSG00000176153.12	GPX2	7	2	0	2	3	0,626066744	1
ENSG00000021355.13	SERPINB1	6	0	0	0	6	0,963892326	1
ENSG00000149418.11	ST14	6	1	0	1	4	0,800548089	1
ENSG00000175701.10	MTLN	6	2	0	1	3	0,498296941	1
ENSG00000142910.16	TINAGL1	6	1	0	1	4	0,800548089	1
ENSG00000253313.6	C1orf210	6	1	1	1	3	0,498296941	1
ENSG00000167785.9	ZNF558	6	2	2	2	0	0,048874086	1
ENSG00000151376.16	ME3	6	1	0	0	5	0,800548089	1
ENSG00000170190.16	SLC16A5	6	1	0	1	4	0,800548089	1
ENSG00000185101.13	ANO9	6	1	0	1	4	0,800548089	1
ENSG00000205336.12	ADGRG1	6	2	1	0	3	0,205491141	1
ENSG00000008323.15	PLEKHG6	6	1	0	1	4	0,800548089	1
ENSG00000013588.8	GPRC5A	6	0	0	1	5	0,963892326	1
ENSG00000121310.17	ECHDC2	6	2	0	2	2	0,498296941	1
ENSG00000127399.15	LRRRC61	6	0	0	0	6	0,963892326	1
ENSG00000188707.6	ZBED6CL	6	0	0	0	6	0,963892326	1
ENSG00000087077.14	TRIP6	7	1	3	0	3	0,113656595	1
ENSG00000142089.16	IFITM3	7	1	1	3	2	0,626066744	1
ENSG00000243449.6	C4orf48	7	4	1	1	1	0,022961082	1
ENSG00000070087.14	PFN2	7	4	1	1	1	0,022961082	1
ENSG00000135821.18	GLUL	6	3	1	0	2	0,048874086	1
ENSG00000065833.9	ME1	5	1	0	0	4	0,699797706	1
ENSG00000141738.14	GRB7	5	1	0	1	3	0,699797706	1
ENSG00000188112.9	C6orf132	5	0	0	1	4	0,936668287	1
ENSG00000188643.11	S100A16	5	1	0	0	4	0,699797706	1
ENSG00000100558.9	PLEK2	5	1	0	0	4	0,699797706	1
ENSG00000104611.12	SH2D4A	5	0	0	0	5	0,936668287	1
ENSG00000164904.18	ALDH7A1	5	1	1	3	0	0,351894041	1
ENSG00000163435.16	ELF3	5	1	0	0	4	0,699797706	1
ENSG00000118640.11	VAMP8	5	1	0	1	3	0,699797706	1
ENSG00000181885.18	CLDN7	5	1	0	0	4	0,699797706	1
ENSG00000167644.12	C19orf33	5	0	0	1	4	0,936668287	1
ENSG00000166199.13	ALKBH3	5	0	1	4	0	0,699797706	1
ENSG00000174791.11	RIN1	5	0	0	0	5	0,936668287	1
ENSG00000196636.8	SDHAF3	5	2	1	2	0	0,101079771	1
ENSG00000108375.12	RNF43	5	3	0	1	1	0,101079771	1
ENSG00000101213.7	PTK6	5	0	0	1	4	0,936668287	1
ENSG00000157992.12	KRTCAP3	5	0	2	0	3	0,351894041	1
ENSG00000196878.15	LAMB3	5	0	0	1	4	0,936668287	1
ENSG00000081923.14	ATP8B1	6	3	0	1	2	0,205491141	1
ENSG00000111319.13	SCNN1A	5	0	0	1	4	0,936668287	1
ENSG00000108511.10	HOXB6	5	1	1	0	3	0,351894041	1
ENSG00000168758.11	SEMA4C	6	2	1	2	1	0,205491141	1
ENSG00000182795.13	C1orf116	5	1	0	0	4	0,699797706	1
ENSG00000108384.15	RAD51C	4	2	1	1	0	0,030080039	1
ENSG00000125611.15	CHCHD5	4	1	1	0	2	0,201405479	1
ENSG00000166145.14	SPINT1	4	0	0	0	4	0,889294171	1
ENSG00000166557.13	TMED3	4	2	0	2	0	0,201405479	1
ENSG00000215845.11	TSTD1	4	0	0	0	4	0,889294171	1
ENSG00000156711.17	MAPK13	4	0	0	1	3	0,889294171	1
ENSG00000103067.14	ESRP2	4	0	0	1	3	0,889294171	1
ENSG00000106484.16	MEST	4	0	0	1	3	0,889294171	1
ENSG00000175793.12	SFN	4	0	0	1	3	0,889294171	1
ENSG00000056972.20	TRAF3IP2	4	0	0	2	2	0,889294171	1
ENSG00000100379.17	KCTD17	4	2	0	2	0	0,201405479	1
ENSG00000178078.11	STAP2	4	1	0	0	3	0,56063624	1
ENSG00000206503.13	HLA-A	4	0	0	1	3	0,889294171	1
ENSG00000168785.8	TSPAN5	5	4	0	0	1	0,012330106	1
ENSG00000120451.10	SNX19	4	2	1	1	0	0,030080039	1

ENSG00000178718.6	RPP25	4	2	0	1	1	0,201405479	1
ENSG00000166801.15	FAM111A	5	1	0	1	3	0,699797706	1
ENSG00000042493.16	CAPG	4	0	0	0	4	0,889294171	1
ENSG00000104518.11	GSDMD	4	0	0	1	3	0,889294171	1
ENSG00000112559.14	MDFI	7	1	0	3	3	0,870340627	1
ENSG00000146281.6	PM20D2	4	2	1	1	0	0,030080039	1
ENSG00000106853.20	PTGR1	4	0	3	1	0	0,030080039	1
ENSG00000132481.7	TRIM47	4	0	0	2	2	0,889294171	1
ENSG00000084207.18	GSTP1	3	0	0	1	2	0,807129688	1
ENSG00000110719.10	TCIRG1	3	0	0	1	2	0,807129688	1
ENSG00000111321.11	LTBR	3	0	0	0	3	0,807129688	1
ENSG00000128833.13	MYO5C	3	1	0	1	1	0,38102086	1
ENSG00000172936.14	MYD88	3	1	0	0	2	0,38102086	1
ENSG00000173530.6	TNFRSF10D	3	0	0	0	3	0,807129688	1
ENSG00000178467.18	P4HTM	3	0	0	3	0	0,807129688	1
ENSG00000204386.11	NEU1	3	2	0	1	0	0,072911399	1
ENSG00000214194.9	SMIM30	3	2	0	0	1	0,072911399	1
ENSG00000265972.6	TXNIP	3	2	0	0	1	0,072911399	1
ENSG00000173221.14	GLRX	4	0	0	2	2	0,889294171	1
ENSG00000168404.13	MLKL	3	0	0	1	2	0,807129688	1
ENSG00000068079.7	IFI35	3	1	0	1	1	0,38102086	1
ENSG00000139178.11	C1RL	3	0	0	2	1	0,807129688	1
ENSG00000151689.13	INPP1	3	0	0	0	3	0,807129688	1
ENSG00000161267.12	BDH1	3	1	0	1	1	0,38102086	1
ENSG00000093144.19	ECHDC1	3	0	0	1	2	0,807129688	1
ENSG00000144645.14	OSBPL10	3	0	0	0	3	0,807129688	1
ENSG00000112763.17	BTN2A1	3	0	0	0	3	0,807129688	1
ENSG00000164930.12	FZD6	3	2	0	1	0	0,072911399	1
ENSG00000203499.11	IQANK1	3	0	1	1	1	0,38102086	1
ENSG00000164403.14	SHROOM1	3	0	0	0	3	0,807129688	1
ENSG00000206075.14	SERPINB5	3	0	0	1	2	0,807129688	1
ENSG00000167608.11	TMC4	3	1	0	1	1	0,38102086	1
ENSG00000165105.10	RASEF	3	2	0	0	1	0,072911399	1
ENSG00000102890.16	ELMO3	3	0	0	0	3	0,807129688	1
ENSG00000174652.19	ZNF266	3	1	0	2	0	0,38102086	1
ENSG00000171345.13	KRT19	3	1	0	0	2	0,38102086	1
ENSG00000124762.13	CDKN1A	3	0	0	0	3	0,807129688	1
ENSG00000147804.9	SLC39A4	3	1	0	0	2	0,38102086	1
ENSG00000172164.15	SNTB1	4	2	1	0	1	0,030080039	1
ENSG00000139629.16	GALNT6	6	2	2	1	1	0,048874086	1
ENSG00000184363.10	PKP3	4	0	0	1	3	0,889294171	1
ENSG00000189143.9	CLDN4	3	1	0	0	2	0,38102086	1
ENSG00000090238.12	YPEL3	2	1	0	0	1	0,175614553	1
ENSG00000095261.14	PSMD5	2	0	1	1	0	0,175614553	1
ENSG00000134330.19	IAH1	2	0	1	1	0	0,175614553	1
ENSG00000135069.14	PSAT1	2	0	1	0	1	0,175614553	1
ENSG00000135457.10	TFCP2	2	1	0	0	1	0,175614553	1
ENSG00000142675.18	CNKSR1	2	1	0	0	1	0,175614553	1
ENSG00000153406.13	NMRAL1	2	0	0	0	2	0,665093412	1
ENSG00000159445.13	THEM4	2	0	0	0	2	0,665093412	1
ENSG00000111679.17	PTPN6	2	1	0	0	1	0,175614553	1
ENSG00000168765.17	GSTM4	2	0	1	0	1	0,175614553	1
ENSG00000159176.14	CSRP1	2	0	0	0	2	0,665093412	1
ENSG00000167925.16	GHDC	2	1	0	0	1	0,175614553	1
ENSG00000197747.9	S100A10	2	1	0	0	1	0,175614553	1
ENSG00000138771.16	SHROOM3	2	1	0	1	0	0,175614553	1
ENSG00000145247.12	OCIAD2	2	0	0	1	1	0,665093412	1
ENSG00000177311.11	ZBTB38	2	0	0	1	1	0,665093412	1
ENSG00000182903.16	ZNF721	3	0	1	0	2	0,38102086	1
ENSG00000127824.14	TUBA4A	2	1	0	0	1	0,175614553	1
ENSG00000141526.16	SLC16A3	2	0	0	0	2	0,665093412	1
ENSG00000005882.12	PDK2	2	1	0	0	1	0,175614553	1
ENSG00000169220.18	RG514	2	1	0	0	1	0,175614553	1
ENSG00000141753.7	IGFBP4	2	1	0	1	0	0,175614553	1
ENSG00000125378.16	BMP4	2	1	0	0	1	0,175614553	1