universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Tumor detection and risk stratification in Ewing sarcoma via methylation based liquid biopsy analysis“

verfasst von / submitted by

Nikolaus Mandlburger, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Inform.Univ. Dr. Claudia Plant
Mitbetreut von / Co-Supervisor:	

Tumor detection and risk stratification in Ewing sarcoma via methylation based liquid biopsy analysis

Abstract

Liquid biopsy is an emerging method in cancer diagnosis and monitoring. While many liquid biopsy approaches in cancer diagnosis rely on cancer specific mutations, such genetic approaches are only of limited use for many pediatric cancers due to their low mutational burden. Epigenetic approaches represent an alternative and have already proven useful in liquid biopsy assays. In this study, I applied machine learning to DNA methylation data obtained from cell-free DNA in the blood to develop minimally invasive biomarkers for Ewing sarcoma, a pediatric bone cancer with unmet clinical needs. I achieved sensitive tumor detection at tumor fractions as low as 0.55% with 100% sensitivity at 95% specificity. Reasonable performance was also achieved for metastasis detection and relapse prediction tasks, suggesting potential applicability of methylation information as prognostic biomarker in Ewing sarcoma. The applied machine learning framework also includes a newly developed feature selection approach termed clusterMRMR, which might be of general relevance also for classification tasks outside the field of biology. When tasked to select features informative for Ewing versus control classification, features yielded by clusterMRMR were found to be enriched in genomic regions with known relevance in hematopoietic cells, which lends support to the hypothesis that such regions might be candidates for pan-cancer markers. Finally, I was able to confirm that enzymatic methylation sequencing does not introduce bias to the length distribution of cell-free DNA fragments. This opens up the possibility to simultaneously perform both methylation and fragmentation-based analysis on enzymatic methylation sequencing data in the future.

1. Introduction

1.1 Liquid biopsy and its applications

In liquid biopsy, analytes are sampled from a body fluid (mostly blood) with the aim to derive information that is usually accessible only via a tissue biopsy. In blood-based liquid biopsies, such analytes can be circulating cell-free DNA and RNA (cfDNA/cfRNA), circulating tumor cells (CTCs) as well as proteins, metabolites and tumor informed platelets[1]. Liquid biopsies can be applied in a variety of different fields, ranging from prenatal testing[2] over tissue damage detection[3] and graft rejection monitoring[4] to detection of cancer[5–9] and monitoring of minimal residual disease[10,11], with some liquid biopsy assays already in clinical use[2,12–14]. Most blood-based liquid biopsy approaches

rely on the analysis of cfDNA, which is not contained within cells but freely circulating in the blood[1]. cfDNA is released into the bloodstream by dying cells, but the details of its release mechanisms are still poorly understood[1,15]. cfDNA is believed to be mainly released from apoptotic or necrotic cells but also secretion seems plausible[16,17]. In healthy individuals, the majority of cfDNA in the blood is contributed by cells of the hematopoietic system, mainly by many different kinds of white blood cells[18,19]. In different diseases, pathological contributions to the cfDNA mixture are observed. For example in cancer, tumor cells are known to release cfDNA[20,21]. As cfDNA reflects the genomic and epigenomic state of its cell-of-origin[19,22,23], analysis of the circulating tumor DNA (ctDNA) can be used to detect or characterize a tumor while avoiding invasive tissue biopsies.

To date, most liquid biopsy approaches for cancer rely on the genetic analysis of cfDNA such as studying pathognomonic or recurrent genetic mutations for a certain cancer type[24–26]. An example of an FDA approved assay is the MSK-IMPACT hybrid capture based test for profiling of actionable cancer target genes[14]. If, however, a tumor does not show a sufficient number of mutations, genetic analysis cannot provide the desired insights. This is the case for early-stage cancers when the tumor has not yet acquired a large number of mutations[27] or for pediatric cancers, which generally feature very few mutations[28]. In such settings, epigenetics-based analysis of cfDNA can be applied instead.

1.2 Epigenetic-based approaches for liquid biopsy

Several studies in the recent years have demonstrated that epigenetic information derived from cfDNA is useful for a variety of tasks such as cancer detection and classification[7–9] or tissue deconvolution[18,29], i.e., quantifying which tissue contributes which amount of cfDNA to the mixture. Several kinds of epigenetic information can be used in cfDNA analysis, two of the most promising ones are cfDNA fragmentation-based measures and DNA methylation.

Fragmentation based measures (fragmentomics) exploit the fact that cfDNA fragmentation is a non-random process that depends on the chromatin state of the corresponding cells-of-origin of cfDNA. Chromatin configuration is strongly linked to cell identity because it reflects which regulatory elements are accessible and which genes are expressed in a certain cell. Open chromatin indicates active regulatory elements or transcription. In regions of open chromatin, DNA tends to get fragmented and degraded more strongly than in regions of closed chromatin. Therefore, regions of open chromatin show lower coverage with shorter fragment sizes in cfDNA data[9,22,30,31]. Fragmentomics data has been successfully applied for detecting cancer and classifying cancer types from cfDNA samples in several recent studies[8,9].

DNA methylation provides epigenetic information carried by cfDNA fragments. DNA methylation occurs mainly at the 5’ carbon atom of the pyrimidine ring of cytosine in the context of CG dinucleotides (CpG sites). DNA methylation plays a vital role in regulating gene expression and maintaining cell identity. A high degree of methylation usually causes transcriptional silencing by impairing binding of activating transcription factors or by favoring binding of chromatin remodeling complexes that lead to heterochromatin formation[32]. DNA methylation is a stable indicator of cell identity and therefore has also been used for cfDNA analysis. In previous studies, cfDNA methylation data has enabled not only cancer detection and classification[7] but also tissue deconvolution[18,29]

Some single locus epigenetic liquid biopsy tests exist, with one already being in clinical use[33]. However, the plethora of information potentially contained in epigenetic data also holds great promises for the use of machine learning approaches taking a multitude of potentially informative

sites into account. Indeed, several studies have already successfully applied machine learning on cfDNA fragmentomics or methylation data for tumor detection and classification[7–9].

1.3 Ewing sarcoma

As mentioned previously, epigenetics-based analysis is especially useful for cancers with low mutational burden. Ewing sarcoma is a useful model-disease for such cancers as it is known to harbor particularly few genetic alterations. Ewing sarcoma is the second most frequent pediatric bone cancer, with peak incidence at 15 years. It is a very aggressive cancer with survival rates of 70-80% for patients with localized disease at diagnosis and ~30% for patients with metastatic disease. The state-of-the-art treatment for Ewing sarcoma involves intense multiagent chemotherapy as well as radiation therapy and surgery. This harsh scheme comes with considerable treatment induced morbidity and secondary malignancies for Ewing sarcoma survivors. Despite the intense treatment, a considerable fraction of patients suffer relapse later on in their life which is associated with poor prognosis[34]. Therefore, predicting the relapse risk of patients would be highly relevant for reducing comorbidity by de-escalating treatment for low-risk patients while giving intense treatment to relapse-prone patients. Currently, treatment intensity is not personalized, partially due to the lack of a good prognostic biomarker. Current candidates for prognostic biomarkers include copy number variations and loss-of-function mutations in certain tumor suppressor genes such as TP53. Furthermore, recent studies have suggested that ctDNA levels at diagnosis carry prognostic value[9,35]. To date, for none of the mentioned biomarkers enough clinical evidence has been gathered in order to be used to reliably inform treatment selection[26].

1.4 Liquid biopsy approaches for Ewing sarcoma

A recent study by the Peneder et al. has shown that cfDNA fragmentomics features are useful for detecting ctDNA from Ewing sarcoma cells, differentiating between Ewing sarcoma and other cancer types and for quantifying ctDNA fractions in cfDNA samples of Ewing sarcoma patients[9]. While this study exploited characteristic, tumor-specific patterns of DNA accessibility, Ewing sarcoma is also characterized by marked alterations in DNA methylation[36]. To make use of this additional layer of tumor-related information, it seems promising to extend liquid biopsy analysis approaches to include cfDNA methylation. This could improve the detection of Ewing sarcoma, or possibly even lead to the discovery of prognostic biomarkers. Eventually, integrating cfDNA methylation and fragmentomics data might improve detection and classification performance even more.

Up to now, bisulfite sequencing has been the gold standard method for methylation analysis, but it comes with the considerable disadvantages of being harsh and leading to DNA fragmentation and degradation[37–41]. As a result, bisulfite sequencing data is not compatible with fragmentomics analysis. Recently, a study by Erger et al. has shown that a newly developed protocol that uses non-destructive enzymatic reactions to convert unmethylated (but not methylated) cytosines to uracils (enzymatic methylation sequencing - EM-seq) can be applied to cfDNA to yield methylation data while avoiding biasing the fragment length distribution. Hence, EM-seq data could potentially be used for combined whole genome DNA methylation and fragmentomics analyses[41]. Therefore, in our current project, EM-seq was our method of choice to obtain methylation data from cfDNA samples of

a cohort of Ewing sarcoma patients, patients suffering from other cancer types and healthy control individuals.

1.5 Aims of my Master's thesis

Using these data, I wanted to investigate whether certain classifications based on cfDNA methylation features are possible. On the one hand I looked into how efficiently I could differentiate between Ewing sarcoma and healthy control samples, a task that could be clinically relevant for screening or differential diagnosis. Furthermore, I explored the possibilities to classify Ewing sarcoma as metastatic or non-metastatic, and to predict relapse events for patients based on cfDNA samples at the timepoint of initial diagnosis (before therapy). Both metastasis detection and relapse prediction are tasks that could be used to guide treatment decisions in a clinical setting.

Scarcity of samples and high dimensional feature spaces posed considerable challenges for machine learning - challenges which I addressed not only by extensively comparing different classification algorithms but also by developing feature selection approaches customizable to the given tasks. Various approaches to data simulation enabled benchmarking of the applied methods as well as determining lower limits of cancer detection. Finally, I examined potential biases introduced by EM-seq to fragment length distribution to determine whether EM-seq data will indeed be suitable for fragmentomics based analysis in the further course of the project.

In the following sections, I describe the available data and technical aspects of my experiments in (Materials and Methods), present my findings (Results), critically discuss, and place them into the context of current scientific literature (Discussion). Finally, I summarize my findings and highlight potential future directions (Conclusions and Outlook).

2. *Materials and Methods*

2.1 Cohort description

My Master's thesis project was done in the context of a large liquid biopsy project currently ongoing in the Tomazou lab (supported by the Vienna Science and Technology Fund - WWTF), which aims to analyze material from at least 200 patients with Ewing sarcoma. I joined the project from the very beginning; I used the first batch of sequenced samples for my analysis. These included 40 samples of cell-free DNA (cfDNA) isolated from blood plasma. Among these 40 samples are 15 from healthy controls, including 5 technical replicates, 19 samples from Ewing sarcoma patients and 6 samples derived from patients suffering from other types of sarcomas (osteosarcoma and rhabdomyosarcoma, n=3 each). Blood samples from sarcoma patients were collected at initial diagnosis (before start of therapy). Ewing sarcoma samples were collected by clinicians from four different treatment centers (St.Anna Kinderspital - Austria, Universitätsklinikum Erlangen – Germany, Institut Curie - France, Oslo University Hospital - Norway). At least 5 years of clinical follow up are available for the patients included in this study; among the collected clinical variables are metastasis status at diagnosis and relapse during follow up. For details on clinical variables see Supplemental Table 1 and Supplemental Table 2.

All samples were obtained with informed consent and with approval by the following review boards: Ethics Committee of the Medical University of Vienna (1292/2018), CPP SUD-EST IV, CPP 14/070, EE2012 study (reference number A 14-419), CPP ILE DE FRANCE III, CPP 3272, MAPPYACT study (reference number 2015-A00464-45), CPP ILE DE FRANCE IV, CPP 56-14, NGSKids study (reference number 2014-A00701-46), Ethics Committee of the "Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität Münster" (2008–391-f-A; EudraCT 2008–003658-13 EWING2008), and Ethics Committee for Medical Research in Southeastern Norway (17866).[9]

cfDNA was extracted from plasma using the QIAsymphony Circulating DNA Kit with the QIAsymphony SP (Qiagen) instrument or the QIAampMinElute cfDNA Kit (Qiagen) for manual isolation according to the manufacturer's recommendations[42,43]. Library preparation and enzymatic cytosine conversion was performed using the NEBNext Enzymatic Methyl-seq kit from New England Biolabs[44]. Applicability of enzymatic methylation conversion for methylation profiling of cfDNA has previously been demonstrated in a study of Erger et al., 2020[41]. Libraries were sequenced at a median coverage of 10x on an Illumina Novaseq 6000 machine at the biomedical sequencing facility (BSF) of the Center for Molecular Medicine (CeMM), Vienna. Laboratory work was performed by Daria Pajak (CCRI, Tomazou lab) and Amelie Nemc (CeMM, Bock lab).

Fastp (v. 0.23.2)[45] with default settings was applied to the raw fastq sequencing reads to perform quality filtering, (adapter) trimming and reporting of quality metrics.

In the following, this data will be referred to as "cfDNA data".

2.2 Reference data

Methylation profiles of healthy cells

An atlas of whole genome scale methylation profiles of healthy human cells was provided in a 2022 study of Loyfer et al.[18] The atlas contains 207 samples of 39 different cell types throughout the body, covering all cell types likely contributing to cfDNA in the blood, such as cells of the hematopoietic system, different kinds of leukocytes, liver hepatocytes and many more[19]. Methylation profiles of this dataset were obtained via whole genome bisulfite sequencing (WGBS) at an average coverage of 32x. Prior to sequencing, cells were FACS sorted based on cell type-specific markers, therefore the resulting methylation profiles are expected to reflect the methylation pattern of the corresponding cell types reliably.[18]

The data was obtained from GEO database in beta file format using the accession ID GSE186458 (raw sequencing reads were not publicly available).

Methylation profiles of Ewing sarcoma and mesenchymal stem cells

Methylation profiles of Ewing sarcoma tissue samples and mesenchymal stem cells were generated by Sheffield et al. in a 2017 study[36]. The methylation profiles have been obtained using the reduced representation bisulfite sequencing (RRBS) method, covering only CpG rich regions of the genome[46] (~5-10% of the genome). This dataset comprises 140 Ewing sarcoma tissue samples, each from a different patient as well as 32 samples of low passage primary mesenchymal stem cell lines.[36] In addition, whole genome bisulfite sequencing (WGBS) data of Ewing sarcoma tissue of three different patients was available[9,36]. Data was generated previously in the Tomazou lab and was obtained in the form of raw sequencing reads. It is available via the gene expression omnibus database (GEO) through the accession ID GSE88826.

Methylation profiles of other cancers

RRBS scale methylation profiles of 105 non-Ewing sarcoma cancers were obtained from a publicly available database (sequencing read archive, SRA). The following cancer types were included: Acute promyelocytic leukemia (APL, N=18), chronic lymphocytic leukemia (CLL, N=43), chondrosarcoma (CHS, N=21), glioblastoma (GB, N=3), prostate cancer (PCA, N=7), metastatic castration resistant prostate cancer (MCRPC, N=6), benign prostate tissue (BPCA, N=7). Data was obtained as raw sequencing reads. The accession IDs of these samples can be found in supplementary table 2 of the 2017 paper by Sheffield et al.[36]

2.3 Mapping of converted reads, methylation calling and extraction

The gemBS pipeline (v. 3.5.1)[47] was used to map bisulfite/enzymatically converted reads to a human reference genome and to perform methylation calling as well as extracting data into bed format[47]. The workflow was conducted using the gemBS commands prepare, index, map, call and extract. The hg38 build of the human genome (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz) was used as a reference for mapping and all further analyses. Furthermore, pUC19 and phage lambda sequences were included as additional references for over- and under conversion control[48,49].

For specific parameters set in different sections of the config file see Appendix.

2.4 cfDNA data – assessment of mapping quality and insert size distribution

Reports on mapping quality are generated by gemBS for individual files, additional measures were obtained via picard CollectWgsMetrics (v. 2.27.1)[50] and samtools flagstats [51] (v. 1.14) both with default parameters and summarized using multiqc (v. 1.11)[52]. Insert size distributions were obtained via picard InsertSizeMetrics with histogram output enabled (--H) and width parameter set to 800 bp (--W 800). Visualizations of insert size distributions were generated using a custom python script utilizing matplotlib [53].

2.5 Methylation data filtering and conversion to beta format

After extracting methylation data into gemBS style bed format, genomic CpG positions were filtered for sites residing on chromosome 1 to 22 using awk. In the same process the bed file was reformatted into the following columns (without a header): chromosome, position of C (0 based indexing), position of G (0 based indexing), number of methylated reads, number of total reads. These reformatted bed files were subsequently converted into beta file format (binary file format for storing methylation data in the form of methylation proportions of individual CpG sites, also called beta values) using wgbstools (v. 0.1.0)[18] . During conversion to beta format, positions are automatically filtered to sites that are CpG also in the reference genome.

Files in beta format were subsequently used for segmentation, differential methylation testing and data matrix generation with wgbstools.

2.6 Segmentation into blocks

The genome was segmented into genomic regions showing homogenously methylated CpGs (methylation blocks) using the wgbstools segment command. Blocks were restricted to contain at least one CpG (--min_cpg 1) and to be no longer than 2000 base pairs (--max_bp 2000). Segmentation was performed on all 207 samples of the Loyfer et al. methylation atlas together with three WGBS Ewing sarcoma tissue samples and yielded 7,070,348 blocks. The resulting blocks file was indexed using the wgbstools index command.

2.7 Differential methylation analysis/Ewing sarcoma marker identification

Blocks of differentially methylated genomic loci between Ewing sarcoma tissue and other tissues or cell types (hence usable as Ewing sarcoma markers) were identified using the wgbstools find_markers command. No requirements for the minimum number of CpGs, base pair lengths and coverage were set (--min_cpg 1, --min_bp 1, --min_cov 1). The minimum difference threshold between target and background groups was set to 0.1 (--delta 0.1), which means that that the beta values of the target sample most similar to the background group and the background sample most similar to the target group must differ by at least 10 percent points for a feature to be accepted as marker. To allow for some outliers, target quantile and background quantile were set to 0.25 and 0.1 (--tg_quant 0.25, --bg_quant 0.1, these values mean that the 25% of target group datapoints most similar to the background group and the 10% of background group datapoints most similar to the target group are ignored in the delta comparison. These settings for marker detection are chosen to be not very restrictive to avoid filtering out relatively weak, yet informative markers.

Marker finding was performed in two ways using two different sets of samples. Firstly, WGBS and RRBS scale Ewing sarcoma samples were tested against healthy cell types of the Loyfer et al. methylation atlas together with mesenchymal stem cell and other cancer samples (see section 2.2 Reference data). This comparison yielded 929 marker blocks (in the following called “strict marker set”). Secondly, a less restrictive set of markers was generated by testing the three available WGBS-scale Ewing sarcoma samples against healthy cell types, which yielded 74,006 markers (in the following called “broad marker set”). This set of markers is presumably less specific for Ewing sarcoma as the comparison does not include other cancer references, but with this approach markers outside the catchment area of RRBS can be identified. A groups-file indicating which samples belong to background and target group was passed to the find_markers command via the --groups_file flag.

For an overview of block size distributions of the marker sets see Supplemental Figure 1.

2.8 Data simulation

2.8.1 Methylation inspired artificial data

For evaluating different feature selector's abilities to select non-redundant features in a high-noise, high-redundancy setting, I devised a system to simulate data at large scales. In contrast to existing tools for synthetic data generation like scikit-learn's make_classification API[54], my newly developed system produces data for which the redundancy structure is completely known. That means it can straightforwardly be read from the data columns whether a feature is informative or uninformative and to which redundancy group a feature belongs to. In addition, the distribution of features has been modeled to resemble the distributions found in real methylation features of the cfDNA methylation dataset. Redundancy groups are based on a single informative feature. The other members of the redundancy group are duplicate versions of this original feature, altered with random noise (details described below). Therefore, redundancy groups consist of features that are very highly correlated to each other. In the case of methylation data at single CpG resolution, highly correlated neighboring CpGs could constitute such redundancy groups. With the approach I use in my analyses, such highly correlated neighboring CpGs should be grouped into blocks by segmenting the genome (section "2.6 Segmentation into blocks"), but many of the "blocks" actually consist of only one CpG, so in parts the problem might still be given. Besides this relevance for methylation data, highly correlated features could arise from numerous mechanisms in a variety of different application fields of machine learning. Therefore, this type of data simulation also makes sense in terms of evaluating clusterMRMR for application in areas outside biology.

The system comes in the form of a python function which was designed to resemble the make_classification API. The user can specify how many case and control samples to simulate, how many informative features to create (informative with respect to pseudo-Ewing versus pseudo-control classification), how often each informative feature gets duplicated, how strongly duplicate features are altered compared to their original counterparts and how many uninformative noise features to create.

When the function is called, at first informative features get created. Values for pseudo-Ewing and pseudo-control samples are modeled using normal distributions parametrized with mean and standard deviation of Ewing sarcoma and control samples found in informative features of the cfDNA methylation dataset. Starting from the strict Ewing sarcoma marker set (see section "2.7 Differential methylation analysis/Ewing sarcoma marker identification"), features were prefiltered for a minimum information content with respect to the Ewing sarcoma versus control classification task ($p<0.05$ in ANOVA f test, performed using scikit learn's f_classif function). For features that pass the test, mean and standard deviation within Ewing sarcoma samples on the one hand and control samples on the other hand are determined and stored. When simulating a new informative feature, a template feature from the prefiltered feature set is picked at random and its previously determined mean and standard deviation are used to parametrize the normal distribution used to model synthetic data points. To generate duplicate features, informative features get duplicated n times, and each duplicate feature is slightly altered by adding random noise (drawn from a normal distribution with mean = 0 and standard deviation = 0.003 - the default value was chosen to cause mild deviation, judged by a correlation plot) to not match its original counterpart exactly. In addition to these informative features, an arbitrary number of uninformative features with data points drawn from a standard normal distribution can be added. Finally, all features get scaled to range from 0 to 1

as this is the range of methylation proportions (using scikit-learn's MinMaxScaler function with default boundaries of 0 and 1).

It must be noted that with this approach informative features are modeled independently of each other, no covariance between features is modeled. Hence, classification performed on simulated data generated with this approach for a high number of informative features would be considerably easier than the real-life Ewing versus control classification task, because in the real life setting varying tumor fractions and other biological confounders can lead to strong covariance between features. However, this simulation approach is not primarily meant to resemble the Ewing versus control classification task as accurately as possible but is used to benchmark feature selector's ability to pick non-redundant features in a generical way.

2.8.2 *In silico* dilution

Semi-artificial cfDNA mixture samples were created by *in silico* mixing raw sequencing reads of Ewing sarcoma samples with samples from healthy controls (Figure 1). Reads were mixed in different proportions during several dilution steps to simulate samples with various levels of tumor content.

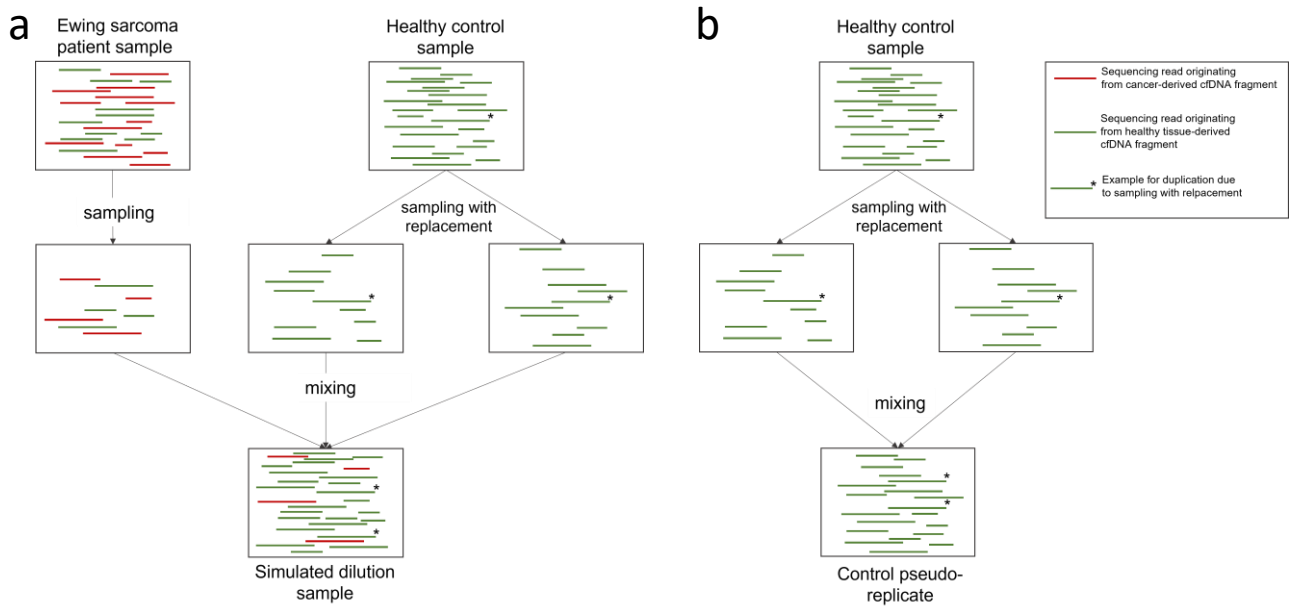


*Figure 1: Illustration of the in silico mixing procedure. Red lines represent sequencing reads originating from cancer-derived cfDNA fragments, green lines represent sequencing reads originating from healthy-tissue derived cfDNA. (a) Generation of a simulated Ewing sarcoma sample with low tumor fraction. Most original Ewing sarcoma samples used for mixing have a tumor fraction of ~50% (~ half of the cfDNA fragments are cancer-derived, therefore ~50% of the depicted sequencing reads in the illustration are red). In silico mixtures were performed using different dilution levels (Ewing sarcoma : control sample ratios of 1:2, 1:10, 1:100 and 1:1000). The mixtures were simulated with the same coverage as the original samples (~10x), the number of reads to be sampled from original Ewing sarcoma and healthy control samples at a certain dilution level was determined via the procedure described in the main text. Sampling of reads from healthy control samples was not done in one step but by sampling half the number of necessary reads twice with replacement. This way, we introduce variability to the simulation and avoid simulated samples being too similar to the healthy samples on which they are based. The three sets of sampled reads are finally merged to yield the simulated dilution sample. Sampling reads from healthy control samples twice with replacement might lead to duplicated reads in the final simulated sample (indicated with * in the illustration), i.e. if a read is sampled in both the first and the second draw, it occurs twice in the final mixture. (b) Generation of a control pseudo-replicate. Control pseudo-replicates are obtained by sampling ~50% of reads from a certain original healthy control sample twice (as previously described). Again, the reasoning behind this is to introduce variability between simulated controls.*

Each Ewing sarcoma cfDNA sample was paired with one healthy control sample to avoid leakage of information between train and test set when performing classification experiments (Figure 2). The number of possible combinations was limited by the 10 available control samples. Tumor content estimations´ values, for the Ewing sarcoma samples used herein, were available in the Tomazou lab. To this end multiple quantification methods were combined: digital droplet PCR (ddPCR) for the EWS-FLI1 fusion, copy number aberrations (ichorCNA[55]), and counts of whole-genome sequencing read spanning the fusion breakpoint[9]. Since tumor content quantifications are expected to be relatively more reliable for samples with high tumor content, Ewing sarcoma samples with high tumor content were preferentially selected for *in silico* mixing. Both metastatic and non-metastatic samples as well as future-relapse and no-future-relapse samples are among the Ewing sarcoma samples used for mixing.

In silico mixtures were performed using four dilution levels with Ewing sarcoma : control sample ratios of 1:2, 1:10, 1:100 and 1:1000. For each sample combination and each dilution level, two replicates were simulated. Subsequently, for every control sample eight pseudo-replicates were generated to obtain equal numbers of simulated controls and simulated mixtures. In total, 160 simulated samples were generated.

Simulated mixtures and control pseudo-replicates were simulated with ~10x coverage, based on the following approximation: The effective genome size of the hg38 genome is 2,913,022,398 bp. Assuming a cfDNA fragment length of 167 bp (the mode of a typical fragment size distribution) and two sequencing reads per fragment (paired-end sequencing), 10*2*2,913,022,398/167=348,864,960 reads must be sampled to reach the desired coverage of 10x. Control pseudo-replicates were generated by sampling half this number twice from the same control sample and merging the two read sets together to yield 10x average coverage. The reasoning behind this procedure is to generate some variability between pseudo-replicates to try to make machine learning on the simulations more difficult and potentially more realistic. For generating simulated Ewing sarcoma mixture samples, the number of reads to achieve the desired ratio is sampled from the original Ewing sarcoma sample of the Ewing sarcoma – control pair. Half the number of reads additionally needed to reach 10x coverage is sampled twice from the control sample of the pair, again to generate variability between replicates of samples with very low Ewing sarcoma contribution. Subsequently the three sets of reads are merged to yield the final simulated sample.

Sampling reads from bam files was conducted via the sambamba (v. 0.8.2) view command[56] where the sampling factor (-s) to achieve desired number of reads is calculated making use of measures obtained by samtools idxstats[51]. Sampled sets of reads were merged via samtools and the resulting bam files were indexed with samtools index using -c flag to produce .csi index files. Methylation calling and extraction from the final bam files as well as filtering and conversion to beta format was done as previously described.

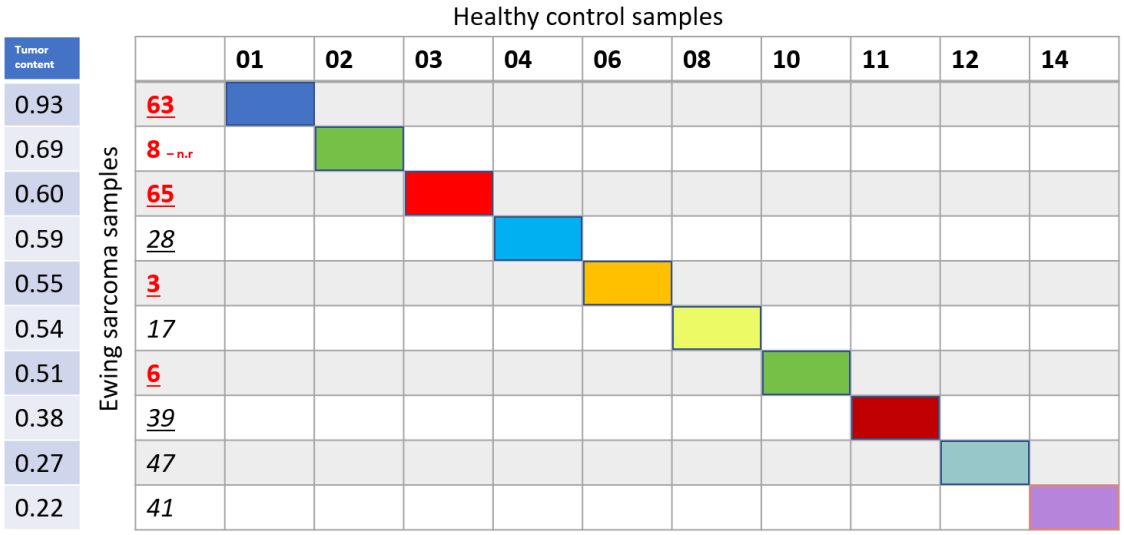


Figure 2: Sample pairings for in silico mixing. Each Ewing sarcoma sample was paired with one healthy control sample to avoid leakage of information between train and test set when performing classification experiments. The number of possible combinations was limited by the 10 available control samples. Tumor content estimations for the original Ewing sarcoma samples (blue column on the lefthand side), available from a previous study of our research group[9], had been obtained by combining estimations from different quantification methods (digital droplet PCR for EWS-FLI1 fusion, copy number aberrations (ichorCNA[55]), and counts of whole-genome sequencing read spanning the fusion breakpoint). Since tumor content quantifications are expected to be relatively more reliable for samples with high tumor content, Ewing sarcoma samples with high tumor content were preferentially selected for mixing. Both metastatic (red font) and non-metastatic (italic black font) samples as well as both future-relapse (underlined) and no-future-relapse (not underlined) samples are among the Ewing samples used for mixing. For one sample no relapse information was available (marked n.r) Mixtures were performed on four dilution levels 1:2, 1:10, 1:100, 1:1000.

2.9 General methylation feature description/data matrix generation

The methylation features used for machine learning are the average methylation proportions of CpGs in blocks. The data tables (features x samples) were produced using the wgbstools beta_to_table command. Minimum coverage parameter was set to 1 (-c 1). Subsequently, a python script utilizing pandas[57] and scikit-learn[54] was used to reformat the data matrix into the conventional samples x features form with a single meaningful index that takes the form chromosome_start_stop. Missing values were replaced with the mean of a marker over all samples using scikit-learn's SimpleImputer.

2.10 Classification, hyperparameter search and features selection

For classification, predominantly scikit-learn implementations of classifier algorithms were used: logistic regression (LogisticRegression)[58], random forest (RandomForestClassifier)[59], naïve bayes[60] (GaussianNB), nearest neighbors classifier (KNeighborsClassifier)[61], support vector machine (SVC)[62,63], and multilayer perceptron (MLPClassifier)[64]. Furthermore, the scikit-learn VotingClassifier was used for building ensemble classifiers composed from algorithmically different classifiers (support vector machines with different kernels, logistic regressions and multi-layer perceptrons).

With this selection of methods, I try to cover a spectrum, from very simple algorithms to those that are also able to recognize more complex patterns. The simplest of the algorithms listed are Naive Bayes and K Nearest Neighbor Classifier. Rather than expecting that these algorithms will yield top performance, I included them due to their simplicity as baseline algorithms to see if other algorithms can improve on them. Logistic regression, support vector machines, and tree-based algorithms such as random forests are well established classification algorithms, furthermore all three algorithms have been successfully used for classification of liquid biopsy data already[65], that is why they are also promising candidates for us. While logistic regressions work with linear decision boundaries, support vector machines and tree-based methods can also use nonlinear decision boundaries and thus recognize more complex patterns[63,66]. Lastly, multilayer perceptrons among the listed algorithms have the greatest flexibility for choosing nonlinear decision boundaries and therefore capture complex patterns[64]. However, since only few training samples are available, there is a risk of overfitting, and it remains to be seen whether they can exploit their potential advantage. Voting classifiers were used to ensemble different algorithms in the hope to improve performance by partly compensating each other's shortcomings, an approach that has previously proven useful for classification tasks based on liquid biopsy data[65,67].

Scikit-learn implementations of select k-best (SelectKBest) and model-based selection (SelectFromModel, based on ExtraTreeClassifier) algorithms were used to benchmark my newly developed feature selection approaches. These two algorithms have been chosen as benchmark because i. selecting features based only on univariate feature importance (k best) is one of the most straightforward ways to perform feature selection and therefore represents a good baseline; ii. Tree based from-model-selection on the other hand makes use of the decision tree's intrinsic redundancy avoidance ability to preferentially select non-redundant features[68]. It therefore represents a more exacting comparative mark for MRMR.

MRMR algorithm was used in an implementation by Samuele Mazzanti[69]. The mrmr_classif python function was incorporated into a wrapper class to be usable in scikit-learn pipelines. As this implementation does not come from a peer reviewed publication, it was compared to a self-implemented version of MRMR to ensure it works properly. The Clustering preselector was used in the implementation described in section "2.13 Clustering preselector".

Scikit-learn pipelines were used to build composite workflows for sequentially applying feature selection and classification steps.

All reported classification performance measures are based on cross-validation. Depending on the specific experiment, grouped cross-validation or repeated stratified cross-validation was used. When sensitivity values at certain levels of specificity are reported, classification thresholds were set based on the test set. I mainly used area under the receiver-operator curve (ROC AUC) as a performance measure in classification experiments. Scikit-learn's roc_curve and roc_auc_score functions were used to determine the ROC curve and its AUC from a classifier's predicted probabilities. Visualizations of ROC curves were generated with custom python scripts. In two cases Matthew's correlation coefficient (MCC) was used as a performance measure.

2.11 Evaluation of lower tumor fraction limit for Ewing sarcoma detection

The *in silico* diluted simulated samples were used to determine the lower limit at which classification works reasonably well. To this end, a 10-fold grouped cross-validation scheme was applied. Grouping

was done with respect to original Ewing sarcoma and control samples making up the simulated mixtures to ensure that original samples never contribute to both train and test set (Figure 2). In every cross-validation fold, the test set consisted of all dilution- and pseudo-replicate samples from a single Ewing-control pair. The training set consisted of all other samples except dilution samples with a tumor content lower than 5%, which were excluded because their inclusion led to impaired classification performance on the test set. A logistic regression classifier (with strong regularization, C=1) using features of the strict marker set was trained on the samples in the training set. Subsequently, two prediction thresholds for the classifier's predicted Ewing sarcoma probability were determined on the test set: The threshold for which test set specificity is 100%, and the threshold for which test set specificity is 95%. These thresholds are equivalent to the highest predicted probability of control samples in the test set and the 95th percentile of predicted probability values of control samples in the test set. Classifications of the Ewing sarcoma samples in the test set were obtained with respect to both thresholds.

Following cross-validation, samples and their predictions were ranked according to their tumor fraction, and sensitivity along the tumor fraction gradient was evaluated using a sliding window approach, where sensitivity was evaluated for groups of five samples at a time. The two sensitivity values found for the two specificity thresholds were plotted against the mean tumor fraction of the five samples. The five-sample window was slid down the tumor fraction gradient in steps of one and the sensitivity values plotted for samples of every window. Sensitivity values obtained at fixed specificity thresholds determined on the test set are equivalent to sensitivity values that can be read from a ROC curve. As all ROC statistics, these measures represent a theoretical optimum (because thresholds are determined on the test set) and should be viewed as optimistic estimates of the achievable sensitivity.

2.12 Feature augmentation approach

I tested if including *in silico* mixed simulated samples into the training process could lead to improved classification on real data in Ewing sarcoma detection tasks. A repeated threefold cross-validation scheme was custom-tailored to this task. The ten groups of simulated samples were randomly split into three folds such that samples belonging to the same group are guaranteed to be in the same fold. Subsequently the original samples that had been used to generate the mixtures were added to the folds that contain their simulated counterparts. Nine Ewing sarcoma samples had not been used to generate mixtures; these were manually arranged in three folds such that the distributions of the estimated tumor fractions are comparable across folds. These folds were then randomly combined with the folds containing the simulated samples. Cross-validation was performed on these combined folds, where in every iteration simulated samples get filtered out of the test set and simulated samples with tumor fractions below a given threshold get filtered out of the training set. This cross-validation scheme was repeated 30 times to account for different combination of folds possibly being differently hard.

To evaluate the effect of feature augmentation on classification performance, the repeated cross-validation scheme was applied with different settings of the minimum tumor fraction threshold. A logistic regression classifier with regularization parameter set to 1 (C=1, strong regularization) was used for classification.

2.13 Clustering preselector

A clustering-based feature preselection algorithm (clustering preselector) was devised primarily to be used in conjunction with MRMR[70,71] feature selection. Developing this approach was motivated by memory limitations of MRMR. Due to its quadratic memory complexity with respect to the number of involved features, MRMR could not be directly applied to a 7 million-dimensional dataset like our available methylation data. To be able to use MRMR nonetheless, the number of features must be reduced upfront to a number manageable by MRMR. The big advantage of MRMR is that it is able to select a jointly informative set of features by considering not only feature informativeness but also redundancy between features. For MRMR to exert its beneficial properties the pre-selected feature set must therefore keep non-redundant features of the full data set. Clustering based pre-selection should achieve this goal.

The first step in the clustering preselector is to filter for features with a minimum information content with respect to a certain classification task using univariate statistical tests (Figure 3(1)). In the current implementation, p values from an ANOVA f test are used as measure for information content. Zhao et al showed in their 2019 paper that MRMR works particularly robustly with the ANOVA f score as an information measure[71]. In order to be as consistent as possible in the joint application of the clustering pre-selector with MRMR, I decided to use the ANOVA f test as an information measure for the clustering pre-selector as well. p values are calculated using scikit-learn's f_classif function[72], features with $p<0.05$ are kept and p values stored for later use.

Subsequently, features get clustered together according to the Euclidian distance between samples using the bisecting k means algorithm[73] as implemented in scikit-learn. That means features that have similar values in the same sample preferentially cluster together

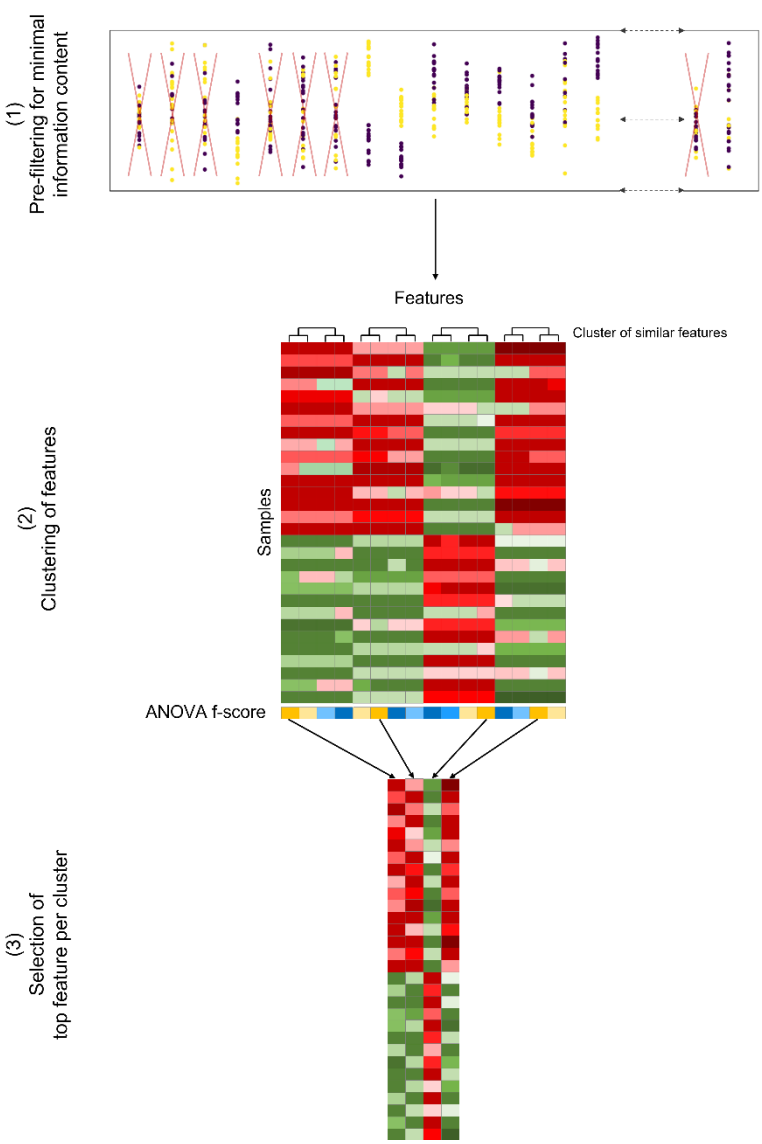


Figure 3: Illustration of the working principle of clustering-based feature (pre-)selection. (1) The first step in the procedure is to filter for features with a minimum information content with respect to the target labeling as measured via ANOVA p value for each feature individually. As a result, features with poor class separation are filtered out (red X in the illustration). (2) The remaining features are clustered so that features that have similar values in the same sample preferentially cluster together. (3) From each cluster, the most informative feature (with the lowest ANOVA p value) is selected and eventually returned.

(Figure 3(2)). Bisecting k means was primarily chosen due to its good memory efficiency. From each cluster, the feature with the lowest ANOVA p-value is chosen (Figure 3(3)). Finally, all chosen features are returned.

The number of clusters to generate is therefore equal to the number of features eventually returned (must be specified by the user). Not only could the choice of the number of selected features affect performance of classification on selected feature sets later on in the workflow, but it also affects runtime performance (considerably faster for few features, for my dataset of 29 samples and several million features; using a single CPU core, selecting several thousand features is a matter of minutes while selecting 25 000 features is a matter of several hours).

The clustering preselector was implemented as a python class, following the fit/transform API of scikit-learn to be used in scikit-learn pipeline objects. Input and output data of clustering preselector objects are pandas data frames. The user can choose from two ways of specifying the number of desired features (=number of clusters): either state the absolute number directly or choose the maximum number of features so that MRMR memory usage does not exceed a stated number of gigabytes. In the following sections the combined clustering preselection + MRMR approach will be referred to as clusterMRMR.

To investigate the identity of the selected features, their genomic distribution as well as enrichment in regions with known biological roles was assessed. The GenomicDistributions R tool (v. 1.4.6)[74] was used to determine the distribution of distances to transcription start sites (calcFeatureDistRefTSS function with default parameters) and the occupancy of certain genomic partitions like introns, exons or intergenic regions (calcPartitionsRef function with default parameters). Genomic distributions for a set of random regions (bedtools random with -n 1000000, v. 2.30.0)[75] were generated as comparison. For enrichment analysis, the LOLA R tool (v. 1.16.0)[76] was used. The features yielded by clusterMRMR were tested for enrichment in region sets of the LOLA core database (hg38 version) as well as in region sets with known relevance in Ewing sarcoma or hematopoietic cells which were eighter generated in our research group or obtained from the ENCODE database[9,77,78] (Ewing sarcoma specific DNAse hypersensitivity sites, EWS-FLI1 binding sites and EWS-FLI1 correlated and anticorrelated enhancers, hematopoietic specific DNAse hypersensitivity sites). Region set of all available methylation blocks (see section “2.6 Segmentation into blocks”) was used as the LOLA universe.

2.14 Metastasis and relapse prediction

Ewing sarcoma samples were classified in two ways: i. samples coming from patients with metastatic versus non metastatic disease; ii. patients suffering relapse versus patients suffering no relapse later on during clinical follow up. Different classifiers were compared and hyperparameter combinations were tested via grid search using three times repeated three-fold stratified cross-validation. Among the tested classifiers were random forests, logistic regressions and support vector machines (scikit-learn implementation as mentioned previously). Comparison and hyperparameter search were conducted on cfDNA data of the strict marker set and on a feature set custom selected for the given tasks using clusterMRMR with a preselector k of 1000 and a final k of 200. ROC AUC was used as classification performance measure in the grid search. For the strict marker set, an additional grid search was performed using Matthew’s correlation coefficient (MCC) as a performance measure. MCC was added in addition to ROC AUC because relapse prediction and metastasis detection are expected to be hard tasks and care must be taken to determine if classification performances are

considerably better than random. A second performance measure at hand might facilitate this distinction.

As control, performance values for hyperparameter combinations were also reported for shuffled label classification for the strict marker set data.

3. Results

3.1 Enzymatic methylation sequencing of cfDNA provides high quality data and allows simultaneous analysis of cfDNA fragmentation patterns and DNA methylation

Various sources of epigenetic information are potentially useful for characterizing cfDNA: besides CpG methylation, an established marker for cell identity, also global and regional fragmentation and coverage patterns can carry epigenetic information because they reflect chromatin accessibility and nucleosome positioning to some degree. Both methylation and fragmentation information have successfully been used for tumor classification, detection, and tissue of origin detection[7–9,22,31].

The gold standard for methylation sequencing is currently whole genome (WGBS) or reduced representation (RRBS) bisulfite sequencing whereas conventional whole genome sequencing (WGS) has mostly been the method of choice in studies analyzing fragmentation patterns (Fragmentomics). A drawback of bisulfite sequencing is that the involved chemical reactions are harsh and lead to considerable degradation and fragmentation of the input DNA, rendering this method unsuitable for studying cfDNA fragmentation patterns[37–40]. However, integrating methylation and fragmentation information might be a promising approach for developing new biomarkers and prediction frameworks, therefore a sequencing protocol yielding both reliable methylation and fragmentation information at the same time would be desirable. Recently, enzymatic methylation sequencing (EM-seq) has been described as a suitable method for obtaining methylation information from cfDNA while retaining the natural fragmentation patterns[41]. While the focus of my thesis work lies on the analysis of methylation patterns, in the further course of the project efforts will be taken to integrate methylation and fragmentation information to improve upon applying the two methods separately. Therefore, EM-seq was chosen for sequencing the cfDNA samples at hand. Consequently, in the process of quality control it was not only necessary to examine sequencing and mapping quality, but also to check if fragment size distributions (i) show the patterns expected for cfDNA from healthy individuals and cancer patients, (ii) are consistent with the distributions obtained via WGS and (iii) are consistent among technical replicates.

Sequencing quality of the samples was consistently good with phred scores of >35 over the whole length of most reads. While a considerable number of reads (~15%) were found to have poor mapping quality, this is not unusual for cytosine converted methylation sequencing data[79]. Most samples showed a median coverage of ~10x. Detailed reports on sequencing and mapping quality can be found in Supplemental Figure 2 and Supplemental Table 3.

Global fragment length distributions of control samples showed the pattern characteristic for cfDNA from healthy plasma: One prominent sharp peak at 166 bp, which corresponds to the length of DNA wrapped around a chromatosome (nucleosome with attached histone 1)[80] and a second weaker and broader peak at ~334 bp, which corresponds to di-nucleosomal fragments[81]. Furthermore, the samples show a characteristic 10 bp periodicity below ~143 bp, which is presumably due to cleavage of DNA wrapped around circulating nucleosomes in the exposed major grooves[80,82](Figure 4a).

Compared to control samples, cancer samples show the expected shift to smaller fragment sizes in their size distributions[8,80](Figure 4b).

The fragment size distributions between different technical replicates proved to be very consistent (Figure 4c). Comparing fragment size distributions of EM-seq samples to size distributions of the same samples previously sequenced via conventional whole genome sequencing (WGS) [9], no deviations that would indicate bias effects of EM-seq were found (Figure 4d).

In summary, I found that EM-seq yielded high quality data that does not only enable methylation-based analyses but also integration of methylation and fragmentomics data in the long run.

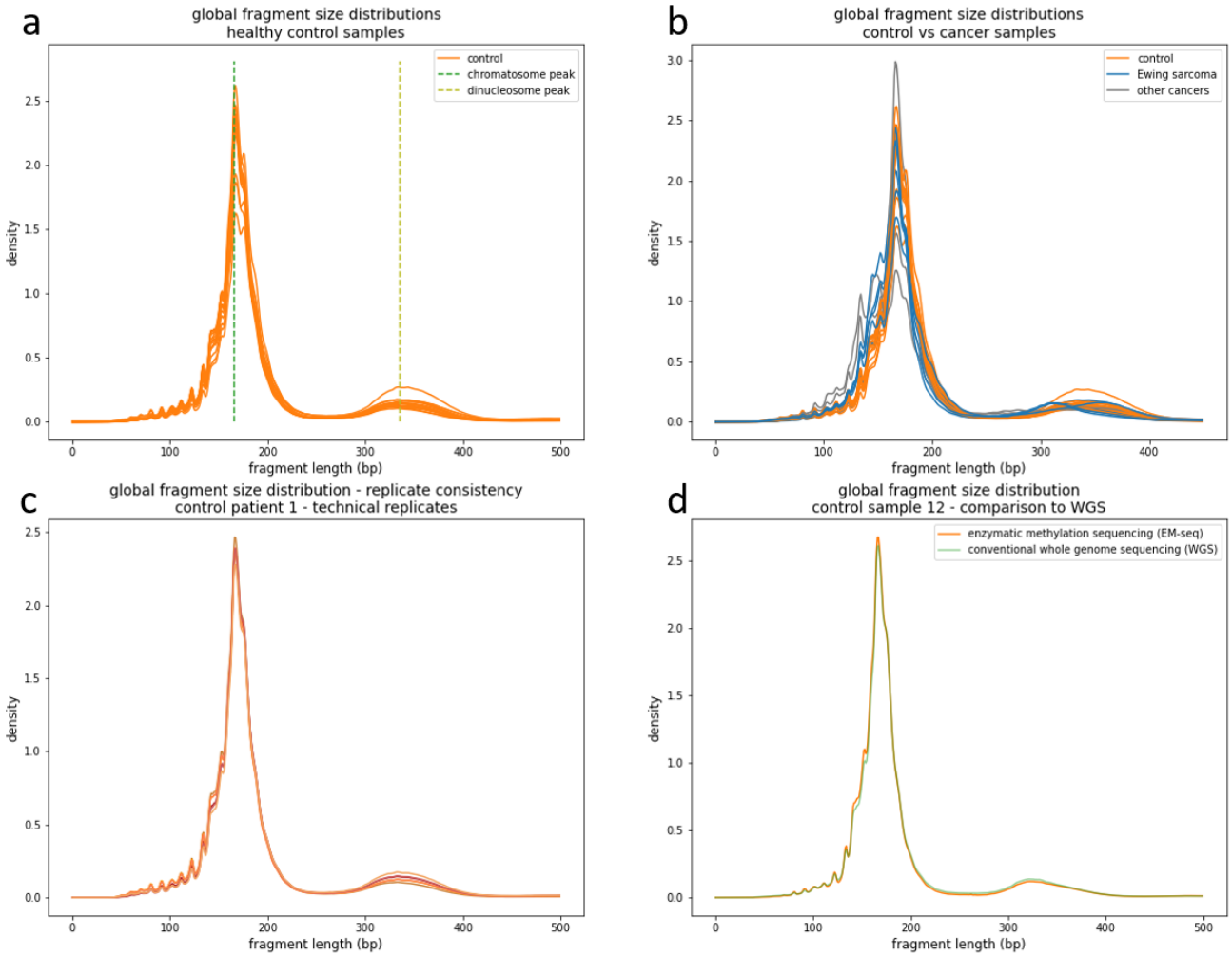


Figure 4: Global fragment size distributions of cfDNA samples: (a) fragment size distributions of healthy controls show the characteristic fragmentation pattern of cfDNA from healthy plasma: a prominent sharp peak at ~166 bp and a smaller and broader peak at ~334 bp, furthermore small peaks with 10 bp periodicity below ~143 bp. (b) As expected, fragment size distributions of cancer samples (blue, grey) are shifted towards smaller sizes compared to healthy controls (orange). (c) As illustrated by the six technical replicates for control subject 1 shown here, fragment size distributions of technical replicate samples are very consistent. (d) The fragment size distributions obtained via EM-seq are consistent with the distributions obtained via conventional WGS of the same samples. Here, the comparison for control sample 12 is shown.

3.2 Differential methylation analysis using tissue reference data yield Ewing sarcoma markers that prove useful for tumor detection

The first major aim of my project was to detect presence of Ewing sarcoma ctDNA based on methylation signatures specific to Ewing sarcoma tumors, in other words to classify cfDNA samples as those corresponding to patients with Ewing sarcoma and to those to healthy control samples. In the human genome approximately 28 million potentially methylated CpG sites exist, and neighboring CpG sites tend to have the same methylation status. Using methylation proportions of CpG sites as features for machine learning posed two fundamental challenges: Firstly, with approximately 28 million CpG sites the feature space is extremely large, which might impair machine learning performance through curse of dimensionality effects[83,84]. Secondly, neighboring CpG sites tend to have the same methylation status, which means that the data is highly redundant.

As a first step to mitigating these effects, the genome was segmented into approximately seven million blocks of homogenously methylated CpGs for all following analyses, and methylation proportions of individual blocks were used instead of methylation proportions of individual CpGs. Using this representation of the methylation data reduced the redundancy from highly correlated neighboring CpGs and at the same time reduced the size of the feature space.

Still, further reduction of the feature space might be beneficial, and in the search for a subset of blocks suitable for identifying Ewing sarcoma contributions to the cfDNA mixture I made use of published datasets of methylation profiles from various tissues and cell types. Comparing Ewing

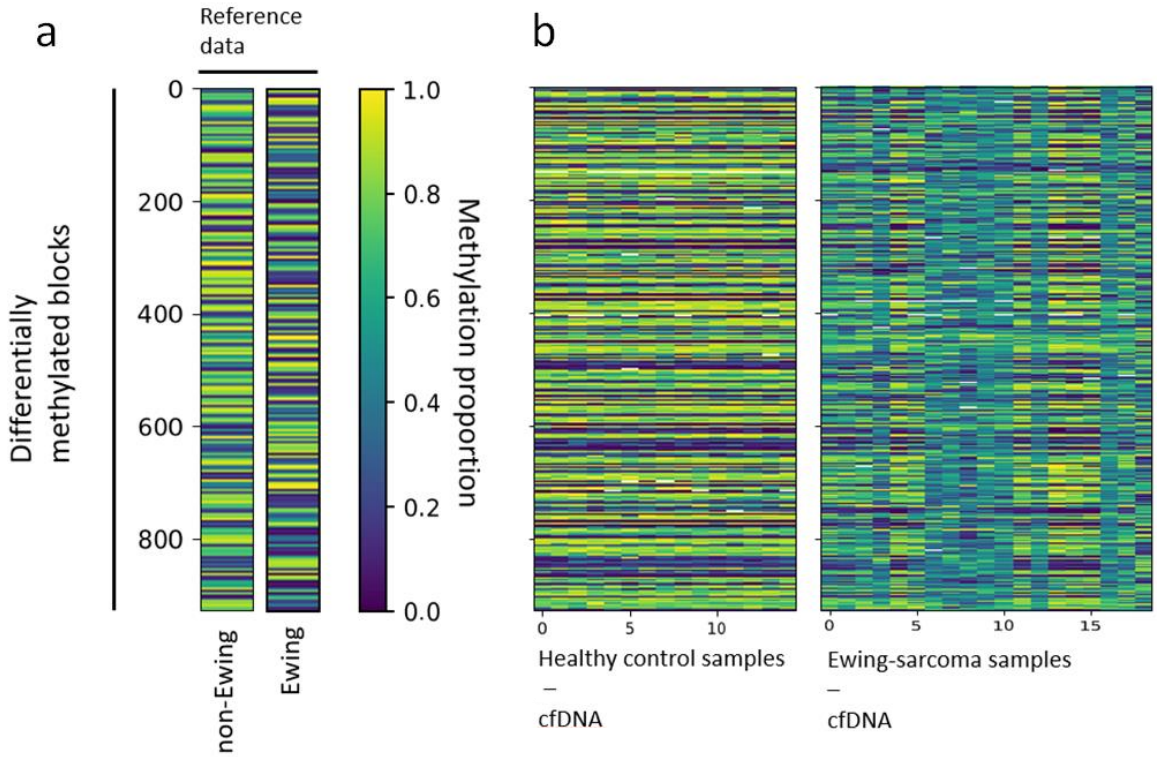


Figure 5: Methylation proportions of various samples at CpGs in strict marker set. (a) mean methylation proportion in markers over Ewing and non-Ewing samples of the tissue and cell type reference dataset. This dataset was used to identify the markers. 536 markers show lower methylation values in Ewing samples compared to non-Ewing samples; 392 markers show higher methylation in Ewing samples (b) methylation proportions in markers (rows) for cfDNA samples (columns) from healthy controls and Ewing sarcoma patients. A striking difference between the two groups is observable, highlighting the potential of these blocks as useful features for cfDNA based detection of Ewing sarcoma. Differences between individual Ewing sarcoma samples are apparent, see Figure 6.

sarcoma methylation profiles against methylation profiles of other cancers[36] and healthy cell types[18], I identified sets of differentially methylated blocks. The methylation proportions of these blocks are promising candidates as features for detecting Ewing sarcoma ctDNA. Two different sets of markers were identified: One strictly selected set at RRBS scale consisting of 928 blocks (strict marker set) and one less strictly selected set at whole genome scale consisting of approximately 70,000 blocks (broad marker set). Visually investing methylation levels of cfDNA samples in the marker blocks, I found clearly visible differences between samples from Ewing sarcoma patients and samples from healthy controls (Figure 5). Striking differences were also observed between individual Ewing sarcoma samples, which can be explained by different levels of tumor content in the cfDNA mixtures of the patient's blood (Figure 6).

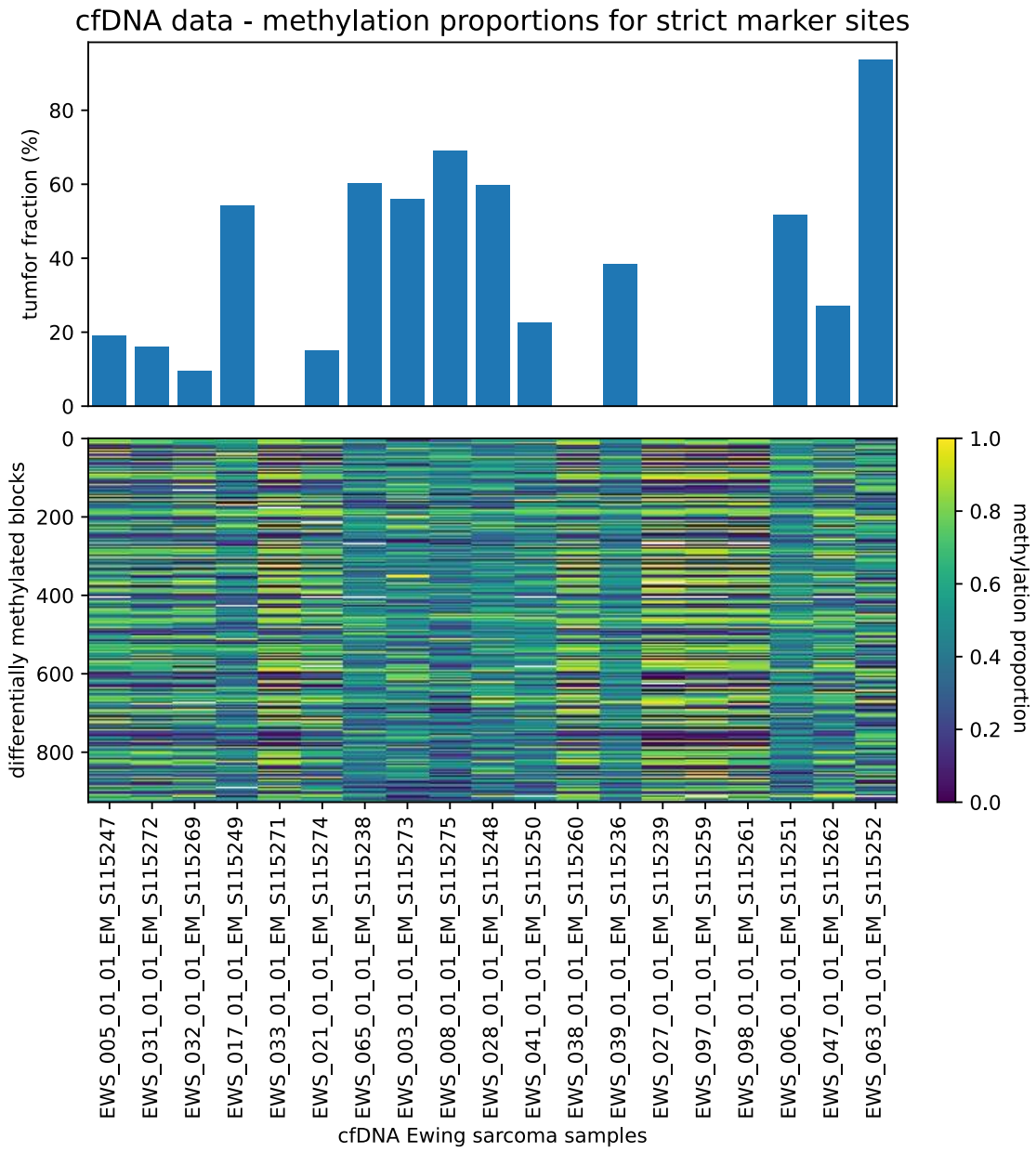


Figure 6: Among the Ewing sarcoma samples there are some that look much more similar to controls samples than others. These differences can be explained by different proportions of tumor DNA in the cfDNA mixtures of the patient's blood. Ewing sarcoma samples with very low tumor fraction show methylation patterns very similar to those of healthy controls. Tumor fractions were determined using a combination of different genetics-based quantification methods in a previous study of our research group[9]. Notably some samples show estimated tumor fraction of 0% although for all samples clinical tumor evidence is given. This can presumably be explained by quantification results not being fully reliable.

At this point I can already conclude that Ewing sarcoma markers could be identified from reference data and that these markers show a striking reflection in the cfDNA data explainable by ctDNA levels.

In the next step I evaluated how useful these markers are for classifying cfDNA samples into coming from patients with Ewing sarcoma or healthy control individuals. To this end, I trained a number of different classifiers on the full 7-million-dimensional methylation data as well as on the methylation data of the broad and the strict marker set. For each dataset, performance comparisons between classifiers and hyperparameter tuning of individual classifiers has been performed. The performance measures of the best performing classifier for each dataset are depicted in Figure 8.

Classification performance was good for all three datasets with ROC AUC values of at least 0.973.

Classifiers trained on strict marker set data performed best, with ROC AUC of up to 0.989 (Figure 7), slightly outperforming classifiers trained on the broad marker set or the full dimensional data. The best performing classifier on the strict marker set was a support vector classifier with strong regularization (C=1) and a polynomial kernel. Logistic regressions with strong regularization (C=1) performed nearly as good as the top performing classifiers in both the strict and the broad marker set data, indicating that they are also good candidates for robust classification.

In summary, I was able to derive Ewing sarcoma specific marker regions from reference datasets; visualizations of methylation proportions of the respective regions for the cfDNA data set showed clearly recognizable differences between Ewing sarcoma and healthy control samples. The marker regions proved useful for classifying cfDNA samples into those corresponding to patients with Ewing sarcoma and those corresponding to healthy individuals, achieving slightly higher performance scores than for classification on the full dataset with ROC AUC values of up to 0.989.

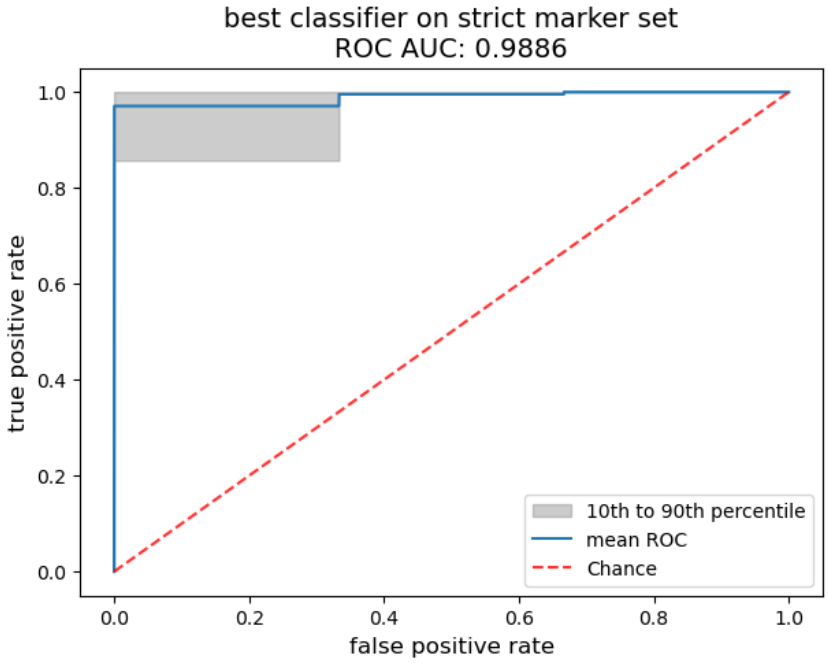


Figure 7: ROC curve displaying the classification performance of the best classifier trained on the strict marker set features as yielded from the hyperparameter search and comparison described in the caption of Figure 8. Best classifier for strict marker set: support vector classifier with regularization parameter C=1 and polynomial kernel, ROC AUC 0.989.

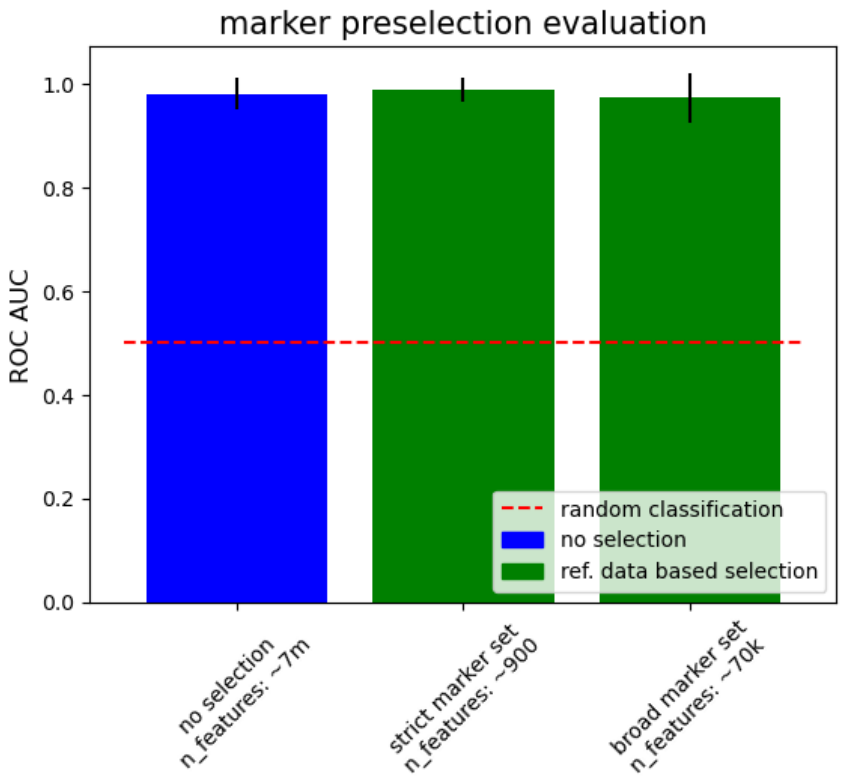


Figure 8: Comparison of classification performance (ROC AUC) between classifiers trained on the two reference-based selected marker sets on the one hand (green bars) and the full dimensional data on the other hand (blue bar) for the classification task of differentiating cfDNA samples from Ewing sarcoma patients from cfDNA samples from healthy control individuals. Performance values are shown for the classifiers that performed best in the hyperparameter search conducted for each of the three datasets. Best estimator for full dimensional dataset: logistic regression with regularization parameter C=0,01 and liblinear solver, ROC AUC 0.980. Best estimator for broad marker set: random forest classifier with maximum depth 1, maximum features 50 and n_estimators 50, ROC AUC 0.973. Best estimator for strict marker set: support vector classifier with regularization parameter C=1 and polynomial kernel, ROC AUC 0.989. A ten times repeated threefold stratified cross-validation scheme was used for hyperparameter search.

3.3 In silico dilution experiments reveal high sensitivity in tumor detection even in samples with low tumor fraction

When considering applying a cfDNA-based classifier for early detection of cancer, it is crucial to know how its performance is related to tumor fraction. That means one needs to determine down to which tumor fraction the classifier yields reliable results. This kind of information is not easily accessible because ground truth on tumor fraction is hard to obtain since the available quantification methods are not completely accurate. To obtain the desired information nevertheless, I generated (very) low tumor fraction samples by diluting high tumor fraction samples *in silico*. To this end, I mixed randomly drawn sequencing reads from Ewing sarcoma cfDNA samples with randomly drawn sequencing reads of control cfDNA samples. The simulated mixtures were subjected to the same methylation calling and post-processing steps as the real samples (see Methods section for further details). This way, I generated artificially diluted samples with tumor fractions down to ~0.02%. The reasoning behind this procedure is that I expect the quantifications of high tumor fraction samples to be relatively more reliable because small absolute changes do not cause big relative error. Furthermore, the fraction of the original sample cannot be above one. Therefore, we know the upper bound of tumor fraction for the simulated sample - for example, a 10x diluted sample cannot

have a tumor fraction above 10%. As the tumor fraction distribution of simulated samples is shifted to lower tumor fraction values compared to the original samples (Figure 9), it must be noted that all classification experiments performed on this artificial cohort are likely harder than on the real cohort because samples with lower tumor fraction are more similar to control samples than samples with higher tumor fraction in all their features.

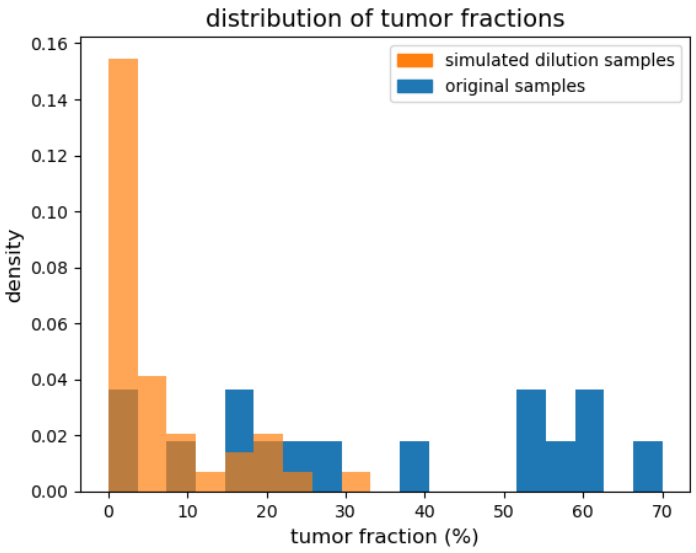


Figure 9: distribution of tumor fraction of original samples (blue) and in silico diluted samples (orange)

The simulated samples were used to determine the lower tumor fraction limit down to which classification works reasonably well. Sensitivity of Ewing sarcoma ctDNA detection was evaluated at two fixed levels of specificity (95% and 100%). These sensitivity values were plotted against the mean tumor fraction of Ewing sarcoma samples in sliding windows of five samples (Figure 10, see Methods section "2.11 Evaluation of lower tumor fraction limit for Ewing sarcoma detection" and "2.8.2 *In silico* dilution" for methodological details).

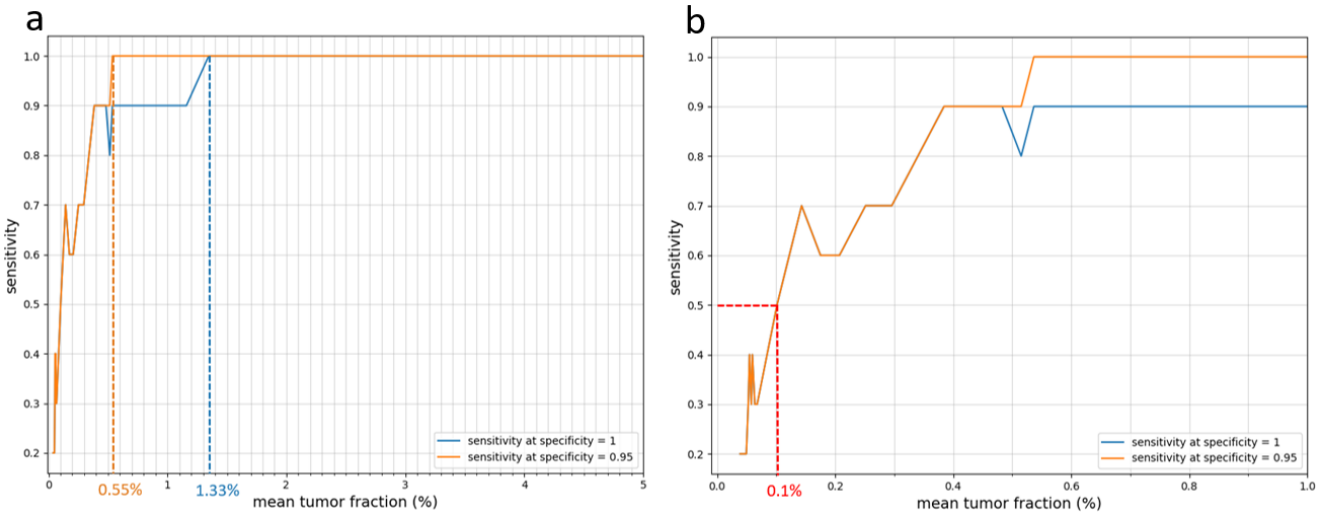


Figure 10: Effect of tumor fraction on a classifier's ability to sensitively classify cfDNA samples from Ewing sarcoma patients. The shown curves were generated by evaluating Ewing sarcoma detection sensitivity at specificity fixed at 100 and 95% for simulated Ewing sarcoma cfDNA samples with known tumor fraction values as obtained via in silico dilution. (a) Perfect sensitivity is achievable down to a tumor fraction of 1.33% at 100% specificity (blue dotted line) or 0.55% at 95% sensitivity (orange dotted line). (b) Down to a tumor fraction of 0.1% a sensitivity of at least 50% can be achieved (red dotted line). A tenfold grouped cross-validation scheme was used to obtain these results. Grouping was done with respect to original Ewing sarcoma and control samples making up the simulated mixtures to ensure that original samples never contribute to both train and test set. Folds are stratified by design as the same number of simulated Ewing samples and pseudo-controls have been simulated for every Ewing sarcoma – control pair. A logistic regression classifier (with strong regularization, C=1) using features of the strict marker set was used to make predictions. Features of the strict marker set were used for training and testing.

The plots in Figure 10 show that sensitive detection is possible down to low tumor fraction levels. At 100% specificity perfect sensitivity can be achieved down to a tumor fraction of 1.33% whereas at 95% specificity at least 0.55% tumor fraction are necessary to achieve 100% sensitivity.

Reducing tumor fraction below these values leads to a steep drop in performance, nevertheless 50% sensitivity is achievable down to a tumor fraction of 0.1%.
Sensitivity values obtained at fixed specificity thresholds determined on the test set are equivalent to sensitivity values that can be read from a ROC curve. As all ROC statistics, these measures represent a theoretical optimum (because thresholds are determined on the test set) and should be viewed as optimistic estimates of the achievable sensitivity.
In summary these results suggest that sensitive detection of Ewing sarcoma ctDNA is possible even at low levels of tumor fraction of around 1%.

3.4 Using simulated data for training does not enhance prediction performance on real data

Using simulated data or altered versions of real data for training ("feature augmentation") is a technique especially known to help mitigating overfitting in several application areas of machine learning. Therefore, I reasoned that using the samples generated by *in silico* dilution for training might improve the prediction performance on real cfDNA samples. I evaluated the effect of adding simulated samples with different minimum tumor fractions to the training set and compared the performance of the resulting classifiers to a classifier trained only on real data. I found that adding simulated data to the training set did not improve performance compared to training on real data only. Including simulated samples with tumor fractions below 1% had adverse effects on classification performance (Figure 11).

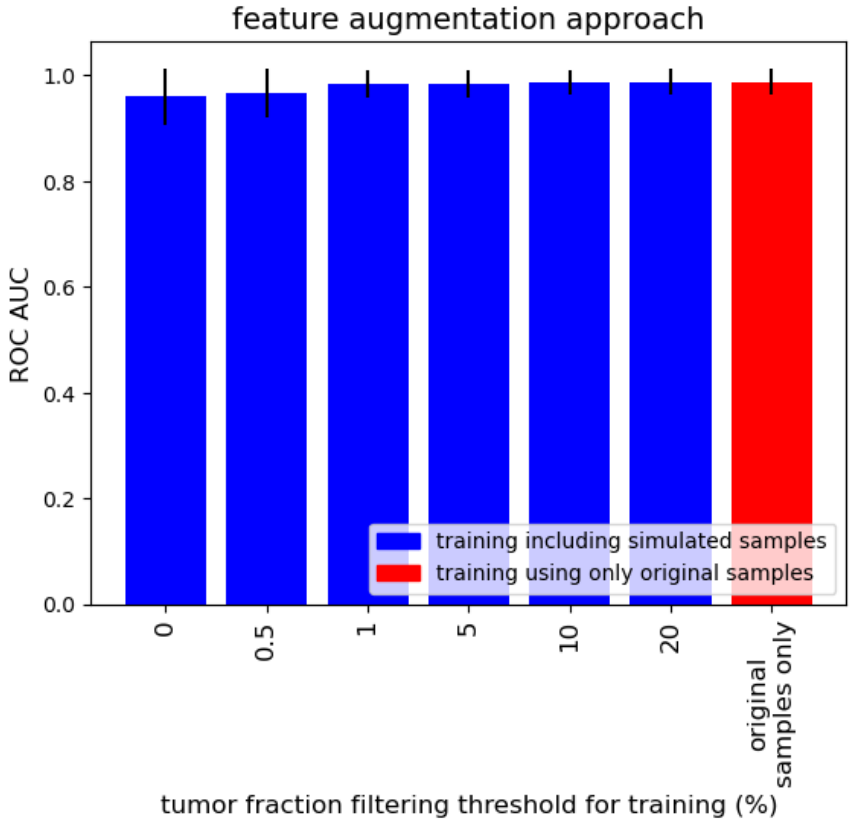


Figure 11: Evaluation of feature augmentation approach for training: adding simulated data to the training process (blue bars) does not result in improved classification performance on real data compared to training on real data only (red bar) for the Ewing sarcoma versus control classification task. Different settings for filtering out very-low tumor fraction simulated samples have been tried (x axis values). Including simulated samples with very low tumor fraction in the training set has adverse effects on classification performance. A custom tailored repeated threefold cross-validation scheme with a logistic regression classifier (with strong regularization, C=1) was used to obtain the performance measures, see Methods section "2.12 Feature augmentation approach". Features of the strict marker set were used for training and testing.

3.5 Selection experiments on methylation inspired artificial data suggest synergetic effects of jointly using clustering preselection and MRMR selection

Custom selecting feature sets for a given classification task is a good way to keep machine learning models interpretable and potentially gain new insights into the underlying problem. For the Ewing sarcoma versus control classification task, I was able to identify a marker set based on reference data, however reference-based selection might not be possible for other potentially interesting classification tasks such as classifying Ewing sarcoma cfDNA samples into metastatic and non-metastatic samples. Therefore, it would be desirable to have a way to custom select good feature sets for arbitrary classification tasks. Good features sets should be jointly informative and not too large to facilitate interpretation and avoid curse of dimensionality effects. An algorithm that aims to select minimal optimal feature sets is maximum relevance – minimum redundancy (MRMR) feature selection[70,71]. MRMR selects features in an iterative fashion considering not only informativeness of candidate features but also redundancy to the already picked features with the aim of yielding a jointly informative set of features in the end. A limiting factor for the use of MRMR is its quadratic memory complexity with respect to number of features. It therefore cannot be directly applied to very high dimensional datasets.

To apply MRMR to such datasets nonetheless, the number of features could be reduced upfront to a number manageable by MRMR. To this end a clustering-based feature pre-selection approach was

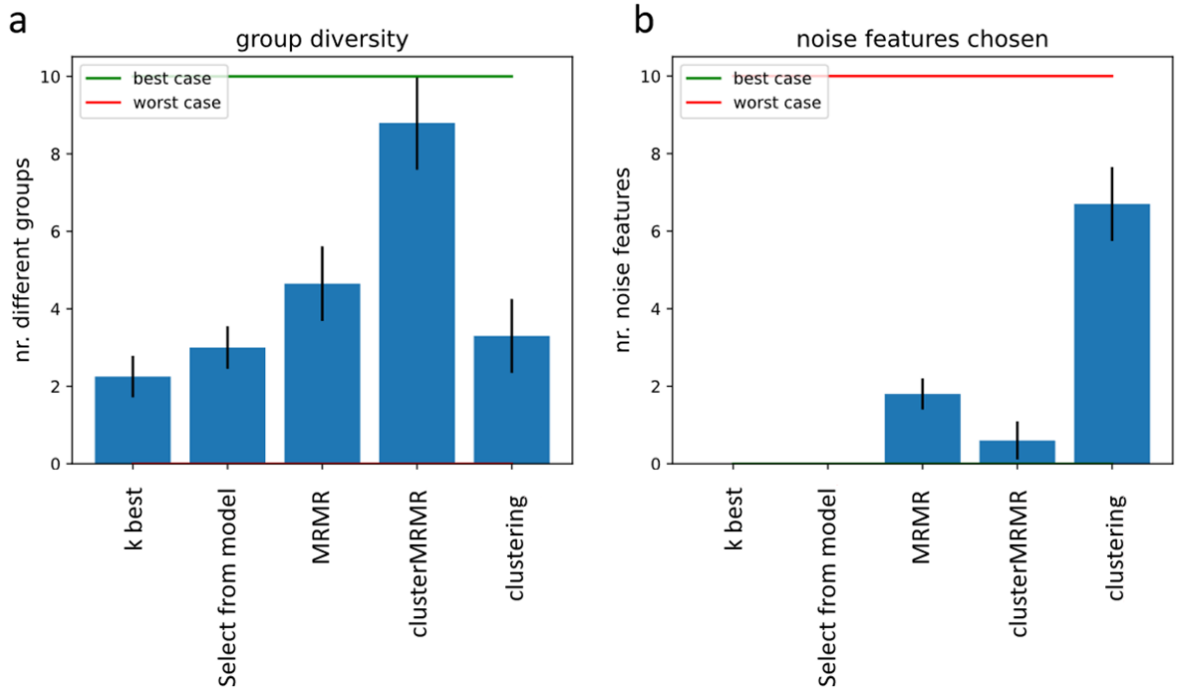


Figure 12: Ability of different feature selectors to select non-redundant features from a methylation inspired artificial data set (see also Methods section "2.8.1 Methylation inspired artificial data"). In a high noise, high redundancy setting the combined clustering pre-selection + MRMR (clusterMRMR) approach performed best in selecting informative features from different predefined redundancy groups while at the same time efficiently avoiding picking non-informative noise features. These results were produced on simulated datasets with 40 informative features from 10 redundancy groups (features of a redundancy group are duplicates of one another altered with slight noise) and 10000 uninformative noise features. 50 pseudo-Ewing samples and 50 pseudo control samples were simulated for each dataset. (a) the number of different redundancy groups among the features selected by each algorithm when tasked to select 10 features. (b) the number of uninformative noise features selected by the different algorithms. Among the tested algorithms are select k best, select from model (using an extra tree classifier as model), MRMR, clustering and clusterMRMR. The shown measures are averages calculated over 20 independently generated datasets.

devised (see Methods section “2.13 Clustering preselector”). Before applying the combined clustering pre-selection + MRMR selection (clusterMRMR) approach to real world problems, I assessed its ability to select non-redundant features on methylation inspired artificial data (see Methods section “2.8.1 Methylation inspired artificial data”) and compared the performance of clusterMRMR with established feature selection algorithms and MRMR and clustering selection on their own.

What I found was that especially in a high noise, high redundancy setting clusterMRMR outperformed the established feature selectors, but also the stand-alone versions of the clustering selector and even MRMR in its ability to select informative, non-redundant features and avoid picking noise features (Figure 12). This came as a surprise as the combination of the two selectors was initially intended as a means to the end of applying MRMR in a very high dimensional setting but now these results even suggest synergetic effects of using the two algorithms in conjunction.

3.6 Hyperparameter tuning revealed best number of features to be used with combined clustering + MRMR selector

The previous results on artificially generated data suggested that using the clustering-based preselection upfront to MRMR selection might not only serve as a means to the end of reducing the number of features down to a number manageable by MRMR but that it might even be synergetic to combine the two methods. While in the previous section such synergy effects have been demonstrated on artificial data in terms of the selectors ability to pick non redundant features, it remains to be shown whether synergy effects are also apparent in terms of classification performance on real data. To this end, I performed a hyperparameter search where I evaluated how varying combinations of the preselector and final k effects the classification performance for the Ewing sarcoma cfDNA samples vs healthy control classification task (Figure 13). The combination of

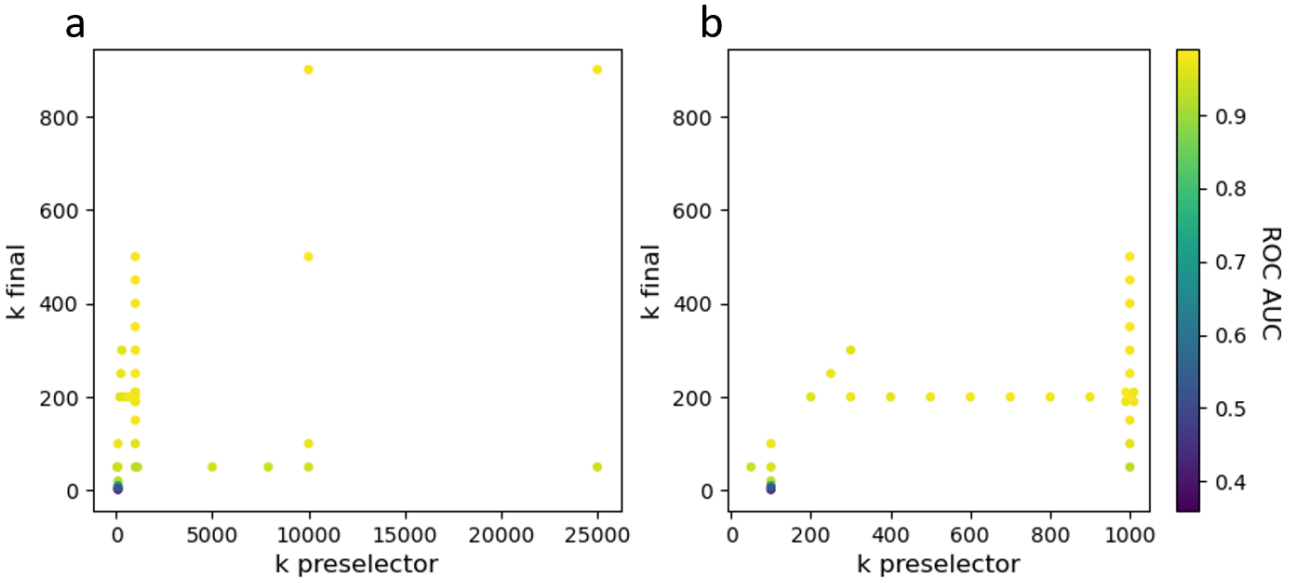


Figure 13: The effect of using different parametrisations of the combined clustering preselection + MRMR selection on classification performance for the Ewing versus control classification task. The best parameter combination found was a preselector k of 1000 and a final k of 200 (ROC AUC of 0.990). Higher numbers did not seem to yield significantly worse performance values. Reasonable performance was achieved down to a final k of 20 (ROC AUC ~ 0.92), below this, performance suffers considerably. A three times repeated threefold stratified crossvalidation scheme with a logistic regression classifier (with strong regularization, C=1) was used to obtain the performance measures. (a) all tested parametrizations (b) zoomed in to the more relevant section.

parameters that yielded the best classification performance was a preselection k of 1000 and a final k of 200 with a ROC AUC of 0.990.

Several parameter combinations with preselection k >= 700 and final k >= 200 yielded ROC AUC values of above 0.980, and with standard deviations of ~0.02 it is reasonable to treat them as equivalently well performing. Increasing the values up to preselection k = 25000 and final k =900 does not adversely affect performance. Down to a final k of 50, ROC AUCs of 0.95 can still be achieved but classification performance begins to suffer considerably if the final k is lowered below 20 (ROC AUCs of ~0.92, 0.72 and 0.52 for k final 20, 10 and 5).

The combination of clustering-based pre-selection with MRMR selections does seem to be beneficial compared to using clustering-based selection only, as selecting 250 or 300 features based only on clustering yields lower classification performance than selecting 200 features MRMR based after clustering-based pre-selection.

These results show that with clusterMRMR one can determine a relatively small set of features that enables very accurate tumor vs. healthy classification of cfDNA samples. Notably, this approach does not rely on reference data or prior knowledge. Still, it remains to be shown how the approach performs compared to other established feature selectors and baselines.

I have applied k best and select from model algorithms as pre-selectors in conjunction with MRMR (Figure 14 b) as well as on their own (Figure 14 a) with the k settings that were found to perform best for the clustering + MRMR selector (preselection k of 1000, final k of 200). I found that the combined clustering + MRMR selection enabled better classification performance than established selectors on their own and in conjunction with MRMR. Classification performance was also found to be better than on the full dataset (ROC AUC of 0.990 vs 0.980) and as good as on the strict marker set data (ROC AUC of 0.990 vs 0.989, see Figure 8). Also, performance measures obtained on the clusterMRMR feature sets are more stable compared to the other selection algorithms as indicated by the smaller error bars. It must be noted that the reported score for clusterMRMR is the test set score of the classifier for the best performing combination of pre-selector and final k. As selecting the best performing hyperparameter combination is a form of training, this procedure leaves room for overfitting the test set to some degree. However, as other hyperparameter combinations similar to the reported optimal combination also yield ROC AUC scores of at least 0.98, it is reasonable to assume that the clustering + MRMR approach indeed outperforms the conventional selectors.

Interestingly, selecting 200 random methylation blocks as features for classification also yielded remarkably good results (ROC AUC of 0.956). This could be explained by a large fraction of the seven million blocks actually being informative, in other words the methylation changes in Ewing sarcoma tumor cells are widespread enough that most genomic locations are to some degree differentially methylated. To see if this explanation is plausible informativeness of features was evaluated according to ANOVA f score and associated p values. 20.5% of total features showed significant information content with respect to Ewing/control labeling and 1.3% of total features showed significant information content after multiple testing correction via Benjamini-Hochberg method. With these numbers, the probabilities of finding at least k significant features among 200 features randomly picked from the full data set were calculated using the hypergeometric distribution (Supplemental Figure 4). When considering features significant without multiple testing correction, one can be confident to find at least 30 of those features in a set of 200 randomly picked features.

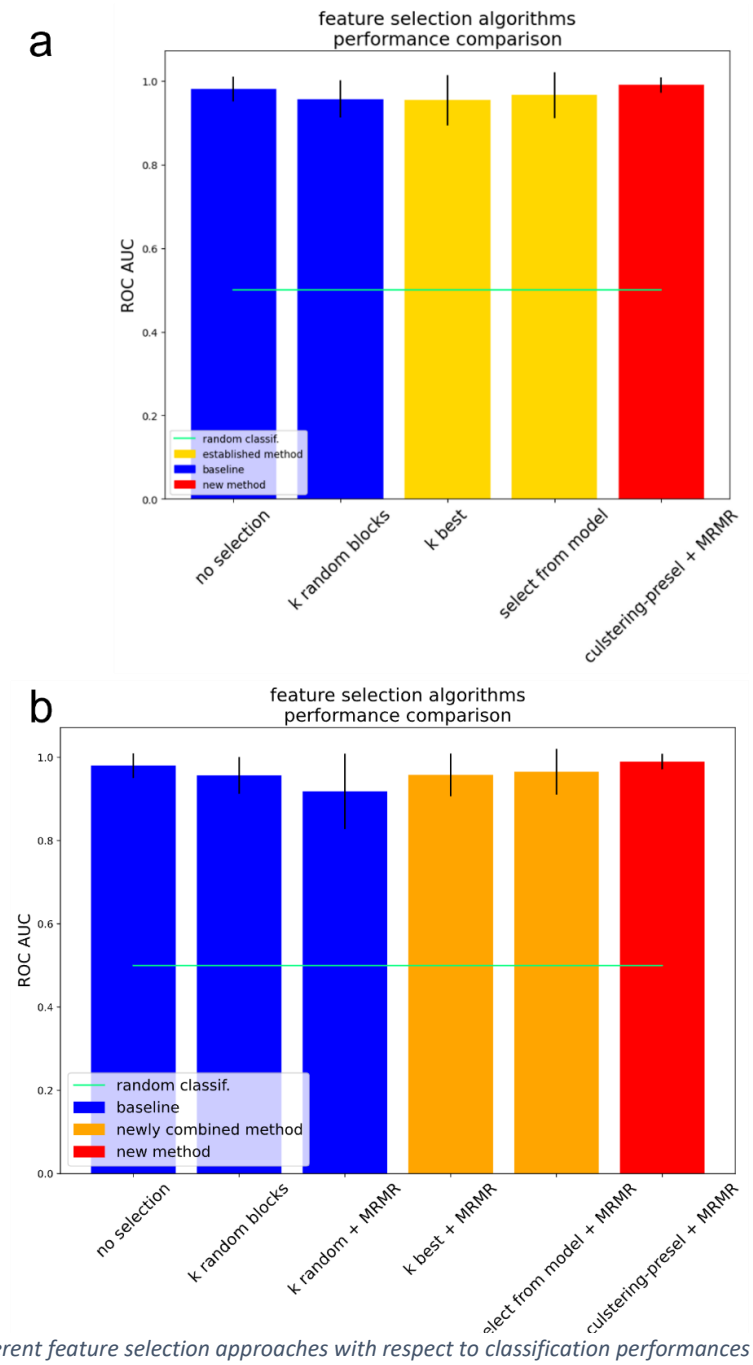


Figure 14: Comparing different feature selection approaches with respect to classification performances on their yielded feature sets for the Ewing versus control classification task. (a) selecting 200 features with clustering + MRMR selection compared to selecting 200 features directly with k best and select from model algorithm. 200 randomly drawn features and classification on the full seven-million-dimensional dataset was used as baseline. Clustering + MRMR selection yielded the best performance (ROC AUC 0.990) and outperformed conventional selectors like k best and select from model (ROC AUC 0.953 and 0.965). (b) preselecting 1000 features using clustering, k best, select from model or by drawing random blocks followed by selecting 200 from these 1000 using MRMR. Also, in this case the clustering + MRMR approach yielded the best results (ROC AUC of 0.990 for clustering preselection vs 0.958 and 0.965 for k best and select from model preselection). Notably, also selecting 200 blocks at random enabled remarkably good classification results (ROC AUC 0.958).

When considering features significant after multiple testing correction, one finds a probability of 0.92 to find at least one such feature and with a probability of ~0.5 one finds at least three in a set of 200 randomly picked features. Overall, it does not seem unplausible that a considerable number of informative features gets picked when selecting 200 features at random.

As a control experiment, selection with shuffled labels was performed and yielded poor performance results for all examples as expected (Supplemental Figure 3).

In addition to determining the performance of classification on custom selected feature sets, another property of interest is how stable clusterMRMR is in choosing features. I therefore investigated which features are chosen in the different training sets of a cross-validation and how consistent the selected features are between different training sets (Figure 15).

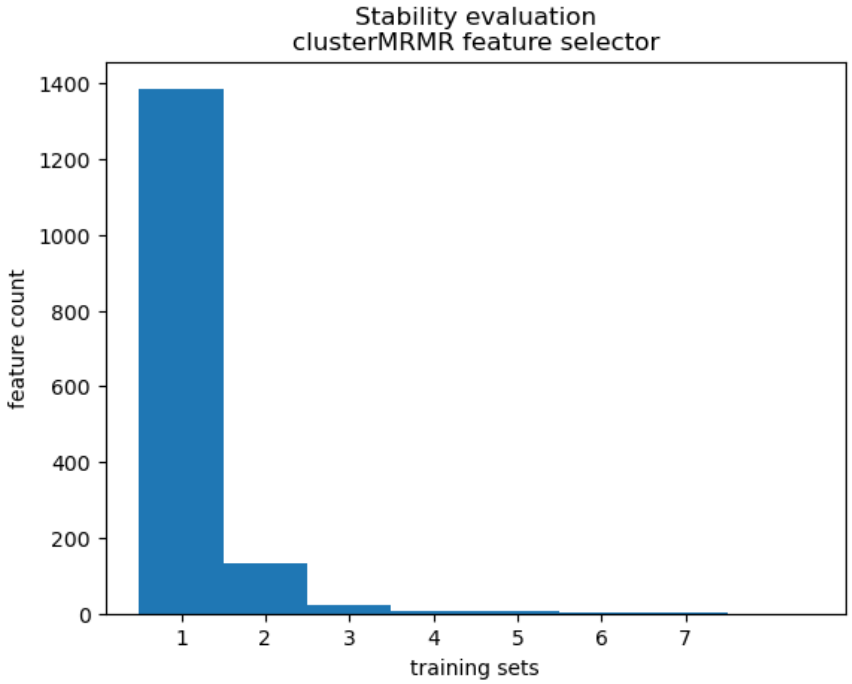


Figure 15: Histogram showing the number of times certain features get selected in the different training sets of three times repeated three-fold cross-validation (selecting informative features with respect to Ewing versus control classification task). ~11% percent of features are selected in more than one training set, ~0.6% of features are selected in more than 4 training sets and one feature is selected in as many as 7 out of 9 training sets.

Most of features are selected in only one of the available training sets, ~11% of overall selected features are selected in more than one training set and ~0.6% of overall selected features are selected in more than four training sets. One feature is selected in as many as 7 of 9 training sets. At first glance these findings might suggest that clusterMRMR is not very stable in selecting features, however one has to keep in mind that there are presumably very many informative features in the full dimensional dataset and that the number of samples is low, therefore it should not come as a surprise that between different training sets the best combination of features can vary considerably. Viewed in this light, a percentage of 11% of features found in multiple training sets does not seem overly low and overall, one should not attest a lack of stability to the algorithm judging from these findings. One of the major advantages of using feature selection in machine learning is that analyzing the identified features can aid the interpretation of the (biological) processes underlying the model. Therefore, I investigated the selected features' general genomic distribution (Figure 16, Figure 17) and their enrichment in genomic region sets from reference databases (Table 1, Table 2). Compared to a control set of random genomic regions, the selected regions showed less intergenic regions and more exonic, intronic and promoter proximal regions (Figure 16). Furthermore, selected regions showed smaller distances to the closest transcription start site (Figure 17).

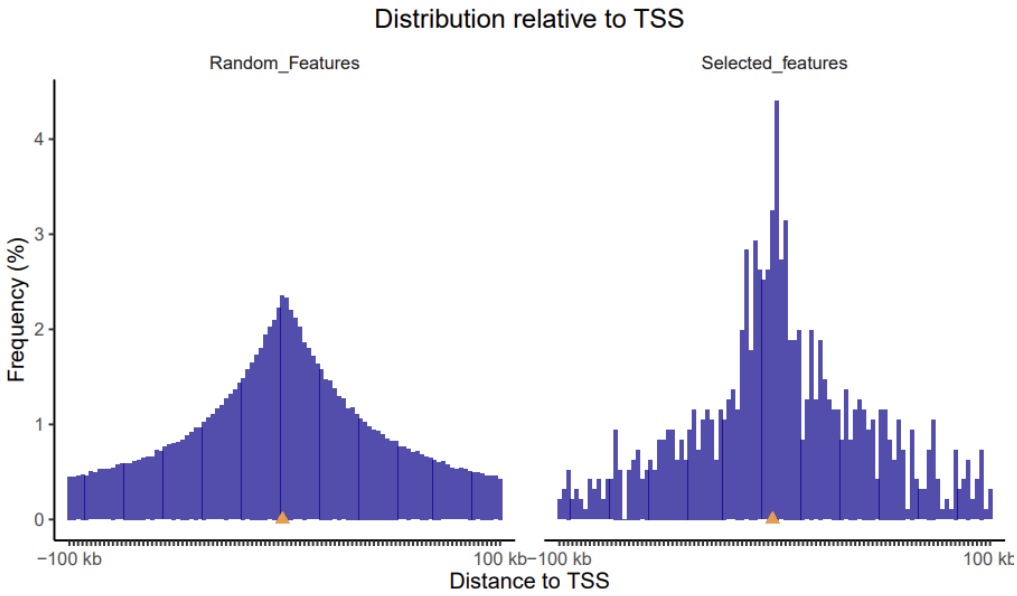


Figure 17: distribution of distances to transcription-start sites (TSS) for random genomic regions (left) and the feature regions selected by clusterMRMR for Ewing versus control classification task. One finds that for selected regions the distribution is shifted towards smaller distances, which aligns well with the findings that more exonic, intronic and promoter proximal regions are found among the selected regions (Figure 16).

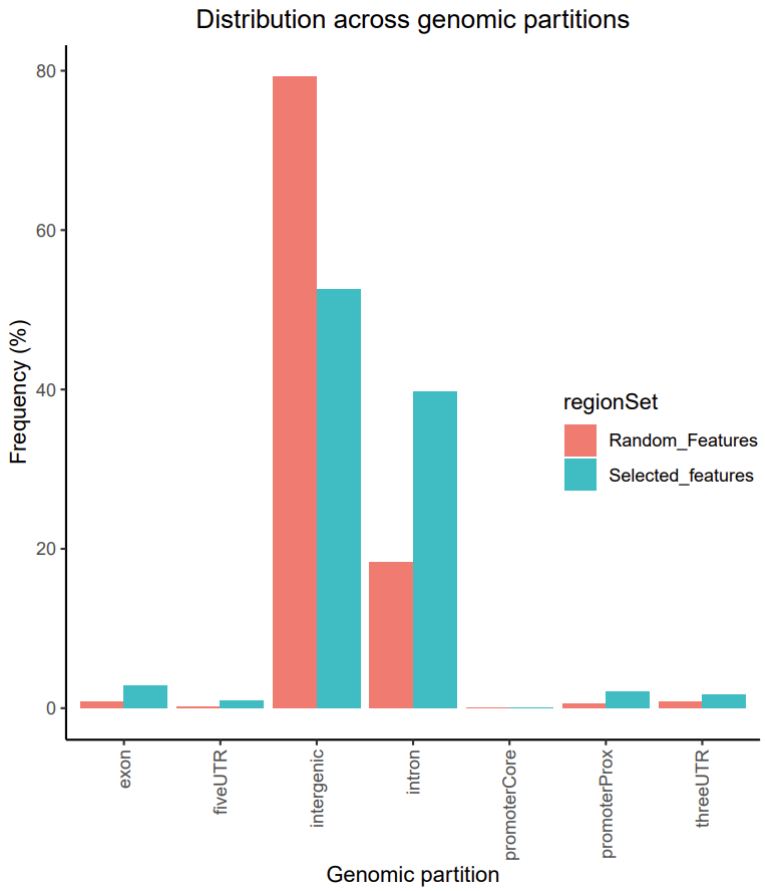


Figure 16: Distribution of the feature regions selected by clusterMRMR for Ewing versus control classification task. Compared to a set of random genomic region, less intergenic regions are found among the selected regions but more exonic, intronic and promoter proximal regions are found. For core promoter regions no difference to random set is recognizable.

The set of overall selected features and the sets of particularly consistently selected features (selected in more than 4 of 9 training sets) have been subjected to enrichment analysis against certain region sets with known relevance in Ewing sarcoma as well as against the large LOLA core database containing a multitude of different genomic region sets [76]. Among the separately tested regions of interest are Ewing sarcoma specific DNAse hypersensitivity sites (DHS), EWS-FLI1 binding sites and EWS-FLI1 correlated and anticorrelated enhancers[9,36]. Together with these Ewing specific regions, regions with general relevance for cfDNA have been analyzed such as hematopoietic or liver specific DHSs[9].

Among these tested region sets, the only one that showed significant enrichment for the selected features was the set of hematopoietic specific DHSs, while Ewing specific region sets showed minor overlap with the selected features but no significant enrichment (Table 1). The set of particularly consistently selected features showed no overlap at all with the regions of interest Supplemental Table 4).

Table 1: Enrichment analysis results for known Ewing sarcoma relevant region sets and region sets with general relevance in cfDNA versus marker regions selected by clusterMRMR. In total 1558 selected marker regions are present, the number of regions in the Ewing sarcoma relevant region sets ranges from 964 to 5611. Ewing sarcoma relevant regions: binding sites of the EWS-FLI1 fusion protein, enhancers correlated or anticorrelated with the presence of the fusion protein and Ewing sarcoma specific DNAse hypersensitivity sites (DHSs). The only region set showing significant enrichment for the selected features was the set of hematopoietic specific DHSs. Ewing specific region sets showed minor overlap with the selected features but no significant enrichment. Top 10 entries of the table are shown.

region set	n overlaps	meanRnk	p value
hematopoietic_specific_DHSs	7	1.00	0.009944
EWS-FLI1_binding_sites	4	2.33	0.161764
ARMS_specific_DHSs	2	3.00	0.291115
EWS-FLI1_correlated_H3K27ac	4	3.33	0.358015
EwS_specific_DHSs	2	4.67	0.575368
EWS-FLI1_anticorrelated_H3K27ac	2	5.33	0.707714
liver_specific_DHSs	0	7.00	1.000000
universal_DHSs	0	7.00	1.000000

Table 2: Enrichment analysis results for marker regions selected by clusterMRMR versus the LOLA core database of genomic regions. The selected marker regions were found to be mainly enrichment for regions occupied with hematopoietic transcription factors and co regulators in peripheral blood cells and acute myelomonocytic leukemia (AML) cell lineages. Among those are the transcription factors are RUNX1, PU1/ SPI1 and ERG and the co regulators EP300 and HDAC1. Top 10 entries of the table are shown

region set	n overlaps	meanRnk	p value
AML cell line carrying inv(16) translocation	71	76.3	1.283942e-12
Peripheral blood (bone marrow mononuclear CD34+)	81	84.7	1.650755e-11
Peripheral blood monocytes separated by leukap...	59	92.7	9.718989e-09
Kasumi-1 (AML FAB M2) with t(8;21)	49	94.0	3.994228e-08
Juvenile acute myeloid leukaemia cell line	73	99.0	1.636652e-08
Kasumi-1 (AML FAB M2) with t(8;21)	31	102.0	1.775961e-06
Kasumi-1 (AML FAB M2) with t(8;21)	40	103.0	1.119760e-06
cistrome_cistrome UPR9 cells PML	42	105.0	1.199650e-06
Macrophages were derived from peripheral blood	44	105.0	1.542067e-06
AML cell line carrying inv(16) translocation	37	105.0	2.143362e-06
cistrome_cistrome UPR9 cells RARA	62	106.0	4.922468e-07

Enrichment analysis against the LOLA core database showed mainly enrichment for regions occupied with hematopoietic transcription factors and co regulators in peripheral blood cells and acute myelomonocytic leukemia (AML) cell lineages. Among those are the transcription factors are RUNX1, PU1/ SPI1 and ERG and the co regulators EP300 and HDAC1 (Table 2). Particularly consistently selected features (selected in at least 4 of 9 training sets) show enrichment in FOXA2 occupied regions of the genome (a transcriptional activator of liver specific genes, also involved in embryonic development) (Supplemental Table 5). Dilution

effects occurring when the tumor contributes to the cfDNA mixture could explain the relevance of blood specific sites for Ewing versus control classification, see Discussion for more details.

In summary, these results suggest that clusterMRMR has selected general blood specific features rather than Ewing sarcoma specific features.

3.7 cfDNA methylation informed machine learning enables metastasis detection and relapse prediction to a certain extent

In the previous section I have demonstrated that clusterMRMR feature sets enabled good classification of Ewing sarcoma and control samples. While serving as a proof of principle, custom selecting a feature set for this task is of limited practical relevance because marker sets selected on reference data do exist for the Ewing versus control classification task. I expected custom feature selection to be more relevant for classification tasks for which no reference-based marker sets are available, such as the classification of Ewing sarcoma cfDNA samples into samples coming from patients with metastatic versus non metastatic disease on the one hand and patient suffering relapse versus patients suffering no relapse later on during clinical follow up on the other hand. If such

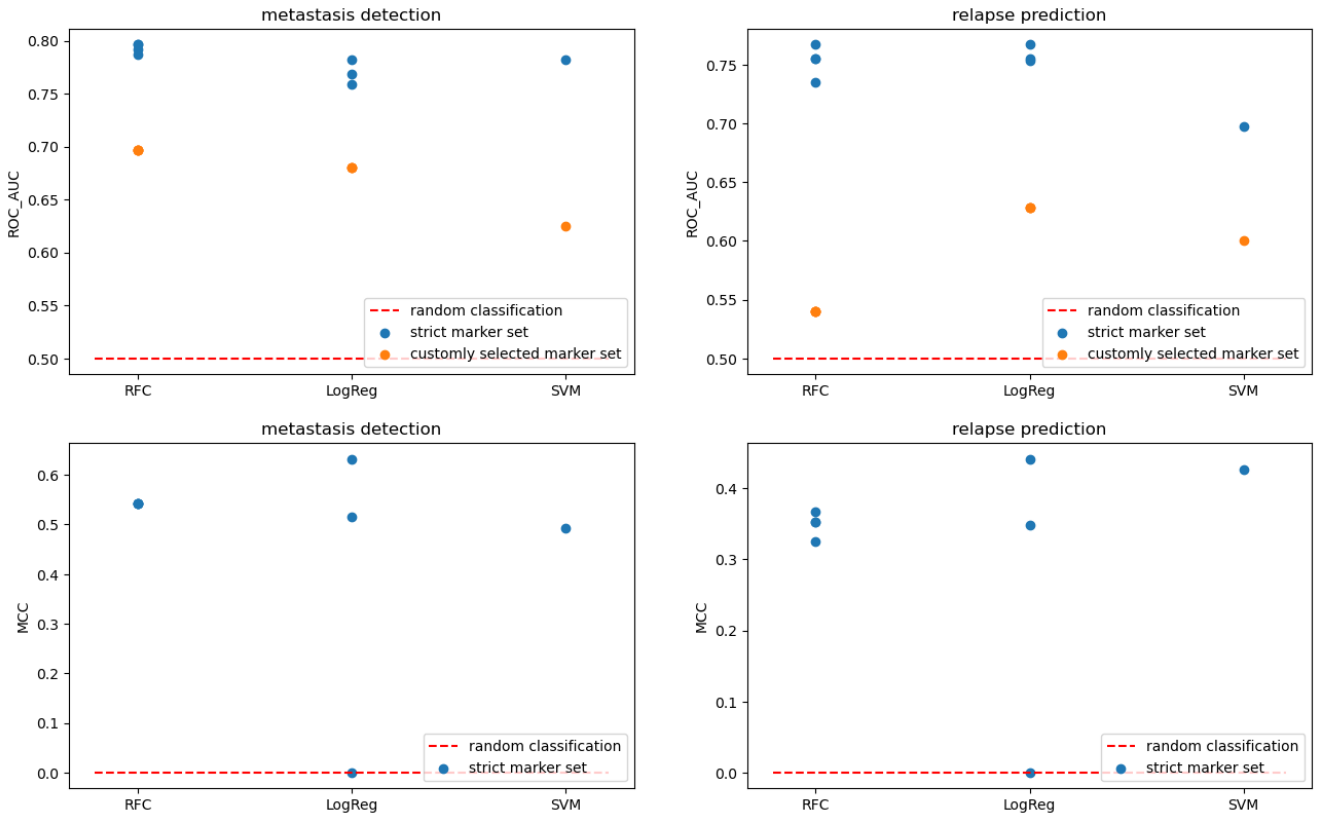


Figure 18: Overview over classification performance values for classification of Ewing sarcoma cfDNA samples into samples coming from patients with metastatic versus non metastatic disease on the one hand (left column) and patient suffering relapse versus patients suffering no relapse later on during clinical follow up on the other hand (right column) as measured by ROC AUC (first row) and Matthew's correlation coefficient (MCC, second row). Among the tested classifiers are random forests (RFC), logistic regressions (LogReg) and support vector machines (SVC). Individual points represent classifier instances with different hyperparameter settings tested in the process of hyperparameter search, blue points represent classifiers trained on strict marker set data, orange points represent classifiers trained on feature sets custom selected for the given classification tasks using clusterMRMR. Interestingly, classifiers trained on custom selected feature sets perform considerably less well than classifiers trained on the strict marker set.

predictions were found to yield reliable results, the gained information could be highly relevant for clinical decision making.

For both tasks I performed classifier comparison and hyperparameter tuning for several classifiers on the strict marker set data on the one hand and features custom selected for the specific tasks using clusterMRMR on the other hand (Figure 18).

For both tasks I found reasonable performance: best estimators from hyperparameter search on strict marker set data showed ROC AUCs of 0.796 for the metastasis detection task and 0.766 for the relapse prediction task, considerably higher than 0.5 expected for random classification. Also, Matthew's correlation coefficient (MCC) values deviated considerably from 0 (expected for random classification). Surprisingly, the classification on feature sets custom selected for the specific tasks performed considerably worse than classification on data of the strict marker set, which was initially selected for a different classification task (the classification of Ewing sarcoma versus control samples).

4. Discussion

4.1 cfDNA methylation enables detection of Ewing sarcoma

My findings show that based on methylation patterns it is possible to differentiate reliably between healthy control cfDNA samples and samples taken from Ewing sarcoma patients at timepoint of diagnosis. It should be mentioned that for all samples collected from patients with Ewing sarcoma clinical evidence confirms the presence of tumor at the point of blood sample collection. The observed performance measures (ROC AUC of up to 0.990 closely match the ones observed for Ewing vs control classification based on fragmentomics data in a previous study of our research group [9], despite the fact that for the current experiments only a fraction of the samples of the 2021 study was available. We expect that as we add more samples in our cohort the performance will increase further. Also compared to other studies aiming at epigenetics based cfDNA classification the current results seem promising: In their 2018 study introducing cfMeDIP-seq[7], Shen et al. reported ROC AUC values from 0.971 to 0.980 for cfDNA methylation-based classification of acute myeloid leukemia, lung and pancreatic cancer. Sensitivity values ranging from 57 to >99% at 98% specificity for the classification of breast, colorectal, lung, ovarian, pancreatic, gastric or bile duct cancer based on fragmentomics features have been reported by Cristiano et al. in their 2019 study[8]. Notably, these studies posed harder classification tasks by not only differentiating samples of specific cancer types from control samples but also from other cancer types. Such classifications will be performed on our cohort as soon as we have a sufficient number of cfDNA samples of non-Ewing cancers available. Despite being not directly comparable to the results of Shen and Cristiano, our results suggest that Ewing sarcoma detection based on cfDNA methylation can keep up with or even outperform the existing cancer detection approaches based on epigenetic cfDNA data.

I must point out that the performance values reported for my methylation-based classification have been obtained on the validation sets used for hyperparameter search. Therefore, it is possible that selecting the classifier with best performing hyperparameter combination might have led to overfitting of the validation set. However, there is reason to believe that the reported performance metrics are not severely over-optimistic, because many different hyperparameter settings yielded good performance results. As soon as the next batch of samples will be sequenced, these samples

will be used as a true test set on which reliable performance estimates will be obtained. The reason why no true test set or nested cross-validation approach has been used in the experiments described here is the scarcity of control samples. As only 10 control samples have been available, in a stratified 3-fold cross-validation a single fold can harbor as little as three control samples. Further reducing this number by separating an extra test set or performing nested cross-validation did not seem sensible to me, which is why I decided to obtain validation set performances only and rely on future batches of samples as true test set.

4.2 *In silico* dilution experiments show low limit of ctDNA detection

Assessing sensitivity of detection for simulated samples with decreasing tumor fraction showed good sensitivity even at low levels of tumor fraction. Reliably detecting even small cancer contributions to the cfDNA mixture is especially relevant for early detection of cancer and monitoring of minimal residual disease. A number of publications have focused specifically on ctDNA detection in ultra-low tumor fraction settings, and few of those have systematically assessed the lower limit of detection of classification frameworks in terms of tumor fraction. The most impressive results have been reported by Widman et al. in their 2022 study[11] with ROC AUC values of 0.84 at simulated tumor fraction levels of as low as 0.0001% for classification of cutaneous melanoma versus healthy control using their newly developed MRD-EDGE signal enrichment platform. This approach however utilizes genetic information and has been applied for cancers with very high mutational load, therefore it is unlikely that this system would yield similar performance also in pediatric sarcomas with very low mutational burden.

We are among the very few to provide a systematically assessed lower limit of detection of a cfDNA epigenetics-based cancer detection framework[7]. I found that at a tumor fraction of 0.55%, sensitivity of 100% could be achieved at a specificity of 95%, and 50% sensitivity at 95% specificity could still be achieved at tumor fractions of 0.1%. These thresholds are low enough to consider our approach for applications like differential diagnosis or even monitoring of minimal residual disease, yet MRD-EDGE in its field of application is orders of magnitude more sensitive. As aforementioned, MRD-EDGE achieves this impressive performance in cancers with high mutational burdens by using a signal enrichment method aligned to mutational information. As mutational information is not promising in Ewing sarcoma, this poses the question of whether comparable signal enrichment approaches are possible for epigenetic data.

Samples of our cohort show a broad distribution of estimated tumor fractions ranging from zero to ~90% (based on genetic evidence). Most of the samples show estimated tumor fractions well above the limit of detection thresholds but some samples also show estimated tumor fractions of zero percent despite clinical tumor evidence, a phenomenon more likely explained by the inaccuracy of the estimation for very low tumor fraction levels than by no ctDNA being present. Nevertheless, the quantifications in combination with the results of the dilution experiment show that low tumor fractions might be the limiting factor for classifying Ewing sarcoma cfDNA samples at the timepoint of diagnosis and pushing the lower limit of detection is expected to be the most promising way to improve classification performance.

Another important conclusion that can be drawn from the results of the dilution experiments is that breadth of sequencing at moderate depths (~10x) combined with a classifier that uses information from multiple genomic locations might be superior to ultra-deep sequencing of single loci biomarkers (e.g. the presence of the disease defining FET::ETS fusion gene). This is because the value of 0.1%

tumor fraction for which appreciable performance is observed is already close to the theoretical limit of detection of ~0.01% tumor fraction for a single-locus molecular biomarker. This limit is imposed by the amount of cfDNA sampled at a single draw of blood. On average 6000 diploid genome equivalents are sampled from a 10 ml blood draw[1,85]. Therefore, 12000 copies of a single locus are present on average, and if one of these copies was mutated this equates to ~0.01%. In other words, at ~0.01% tumor fraction on average only one DNA molecule harboring the tumor specific mutation is contained in the blood sample. Furthermore, the actual number of sampled tumor specific molecules underlies stochastic effects and biases during amplification and library generation, further complicating reliable conclusions based on single-locus-biomarkers.

It has to be noted that the tumor fractions of the simulated samples are calculated based on the estimated tumor fractions of their respective original samples. As this estimation might not be perfect (there is currently no reliable gold-standard method to determine tumor content in cfDNA), also the tumor fractions of the simulated samples are not carved in stone. High tumor fraction samples with relatively more accurate quantifications have been used for mixing to achieve the best possible estimates for the dilution samples. As most original samples show estimated tumor fractions of about 50%, the theoretical upper bounds of tumor fraction for these samples are twice as high as the reported values, because tumor fraction of the original samples cannot possibly be above 100%.

4.3 clusterMRMR proves useful for Ewing sarcoma detection

One focus in terms of machine learning methodology was on the development of a specially adapted feature selection algorithm. In general, what motivated me to look into feature selection was the nature of the problem. With few samples but very many features being available, I expected that curse of dimensionality effects could considerably impede classification performance, therefore reducing the number of features seemed like a promising approach. In addition to optimizing classification performance, I wanted the models to be easy to interpret. Feature selection promised to help with both objectives. It is expected that many genomic loci contain redundant methylation information, therefore it was important to choose a feature selection algorithm that is able to deal with redundancy. My choice fell on the maximum relevance – minimum redundancy (MRMR) feature selection algorithm, which selects a jointly informative set of features by considering both informativeness of features and redundancy to other features. An obstacle was MRMR's quadratic memory complexity with respect to feature number, which prevented directly using MRMR on the methylation dataset with seven million features. To overcome this hindrance, I devised a clustering-based feature pre-selection method to reduce the number of features before using MRMR. I called this combined method clusterMRMR. clusterMRMR showed very favorable properties in simulated data, custom generated to resemble cfDNA methylation data. Also, applied to real cfDNA methylation data, clusterMRMR outperformed several established feature selection algorithms in terms of classification performance on the yielded feature sets. Notably, clusterMRMR is a generically applicable algorithm and in its use not restricted to DNA methylation analysis. If I will be able to demonstrate its superior performance also on non-biological datasets, it could be of general relevance to the field of feature selection.

Ewing sarcoma versus control classification has been performed both based on a set of genomic regions showing Ewing sarcoma specific differential methylation according to a reference dataset as well as on a region set that has been de novo custom selected for the Ewing versus control classification task on the cfDNA data itself using clusterMRMR. Both approaches both yielded good results, with ROC AUC values of up to 0.990 Investigating the identity of the custom selected regions,

I found that they do not considerably overlap with regions a priori known to be relevant in Ewing sarcoma, rather they are enriched in region sets with specific roles in blood cells such as binding sites of hematopoietic transcription factors and transcriptional co-regulators. A possible explanation for not finding enrichment in Ewing sarcoma specific DNAse hypersensitivity sites could be that ctDNA coverage might be particularly low at these regions[9]. The relevance of blood specific sites for Ewing versus control classification can be explained by dilution effects occurring when the tumor contributes to the cfDNA mixture. In a healthy state, the cfDNA of the blood originates mainly from the hematopoietic system, but the stronger the tumor contributes to the cfDNA mixture, the smaller the relative contribution of the hematopoietic system becomes. A classifier can therefore recognize a Ewing sarcoma sample not only by the presence of Ewing sarcoma specific methylation signatures but also by the absence of hematopoietic specific signatures. This means that the custom selected feature set probably would not be suitable for differentiating Ewing versus other non-Ewing cancers cfDNA samples, which is not unexpected because no non-Ewing cancer samples have been included for the selection of the custom feature set, but custom selection of the feature set could be repeated when an appropriate amount of non-Ewing cfDNA samples becomes available. On the other hand, the lack of specificity for a certain cancer type can also hold promises. A previous study of our research group already found genomic regions with relevance in hematopoietic cells to be informative for Ewing versus control classification in terms of fragmentomics measures. This led to the hypothesis that regions with known relevance in hematopoietic cells might serve as pan-cancer markers for epigenetics-based classification. The observed enrichment of custom selected features in hematopoietic region sets fits well with this notion. To lend further support to the hypothesis more cfDNA samples of non-Ewing cancers must be incorporated into the analysis.

Initially, a main motivation for developing the new feature selection approach was to apply it to classification tasks for which no reference-based feature selection was possible, such as metastasis classification or relapse prediction for Ewing samples. By custom selecting feature sets for these tasks, I hoped to both improve classification performance as well as to gain insights into the underlying biological processes by analyzing the identity of the selected features. Opposed to my initial expectations, I found that while the reference-based Ewing specific feature set (which had not been specifically tailored to these tasks) performed surprisingly well for predicting relapse and classifying Ewing samples into metastatic and non-metastatic, custom selected feature sets performed considerably worse for both tasks. Because classification using the latter set of regions was not successful, analysis of region identity to gain biological insights was not expected to lead to useful insights, and therefore omitted. Potentially, a hyperparameter search over different values of the selectors' k parameters could improve performance on the custom selected feature set. However, given the observed drastic difference in performance between the custom feature set and the reference-based feature set, I would not expect that custom selection can outperform the reference-based feature set. Several previous studies have found that the quantity of tumor derived cfDNA itself carries predictive value for relapse and survival[9,35]. Finding the Ewing specific reference-based feature region set to perform best for relapse prediction might suggest that levels of tumor fraction confer most of the information for making the decision rather than methylation signatures specific to relapse-prone tumors.

4.4 EM-seq data for joint methylation and fragmentomics analysis

The previously described experiments show that methylation data derived from cfDNA is highly useful for detecting Ewing sarcoma and identifying relapse prone patients. Recent studies of our

research group and others have shown that also another kind of epigenetic information can be useful for cfDNA based cancer detection and classification, namely cfDNA fragmentation-based metrics (fragmentomics). While harsh bisulfite conversion-based methylation sequencing approaches lead to loss of fragmentomics information, the EM-seq sequencing protocol used in our current study is expected to yield reliable methylation as well as fragmentomics information at the same time. In the near future I aim to integrate methylation and fragmentomics information in the machine learning setup to hopefully achieve synergetic effects on prediction performance. The first step in this direction was to assess whether the EM-seq protocol indeed does not bias the fragment length distribution of cfDNA compared to conventional whole genome sequencing. Indeed, I observed that fragment length distributions obtained from EM-seq closely followed the ones obtained from conventional whole genome sequencing, also the distribution patterns specific for cfDNA from healthy plasma and cancer patient plasma were found as expected. In summary, my results show that data generated via EM-seq is indeed likely suitable for fragmentomics based analyses.

5. Conclusions and outlook

My findings demonstrate that methylation based cfDNA analysis is a promising direction for liquid biopsy-based detection of Ewing sarcoma. Also, metastasis detection and relapse prediction work reasonably well on cfDNA methylation data, although their superiority to other tumor fraction-based approaches remains to be shown. The machine learning framework applied in this study includes a newly developed feature selection approach that not only facilitates the interpretation of tumor detection models but may also prove useful for applications outside of biology (for example on spectral data from chemistry or astronomy, or meteorological data with high time resolution).

After having shown that EM-seq does not seem to bias fragmentation patterns, the next logical step is to jointly use methylation and fragmentomics information for tumor detection and other classification tasks, hoping to achieve synergy effects.

Another promising direction to advance cfDNA-based classification is to first classify individual reads and then classify samples based on the results. This seems plausible because recent studies have suggested that features representing methylation proportion at certain genomic sites (as I used in this study) might not enable sensitive tumor detection in ultra-low tumor fraction settings and that read level approaches could help to further push the lower limit of tumor detection[67,86].

cfDNA methylation data does not only enable tumor detection but is also suitable for tissue deconvolution i.e., quantifying which tissue contributes which amount of cfDNA to the mixture. Recently, it has been shown that damage inflicted by a metastasis to the surrounding tissue can be detected via cfDNA methylation analysis[3]. Therefore, it seems conceivable to use tissue deconvolution to identify unusual contributions to the cfDNA mixture that could indicate the location of a possible metastasis.

In conclusion I can say that our study has clearly highlighted the potential of using cfDNA methylation data in liquid biopsy assays for Ewing sarcoma and that this data also holds great promises for applications beyond tumor detection. I am going to explore these promises in the further course of the project.

Acknowledgments

I would like to give thanks to all people who have supported me throughout the course of this project. Special thanks deserve Daria Pajak and Amelie Nemc for performing the wet-lab work that was necessary to produce the data which my analyses were based on as well as Peter Peneder, Eleni Tomazou and Claudia Plant for their excellent supervision throughout the project. Furthermore, I want to appreciate the work of all clinicians who provided us with the necessary patient blood samples, especially Markus Metzler, Uta Dirksen, Didier Surdez, Kjetil Boye, Caroline Hutter, Gernot Engstler, Heidi Boztug and Michael Dworzak. I want to thank the Austrian red cross for providing control samples of healthy individuals as well as the biomedical sequencing facility at CeMM for performing the sequencing work. Finally, I am grateful to all the patients and their families for allowing us to use their samples for our research.

Bibliography

1. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics 2018 20:2* **20**, 71–88 (2018).

2. Oepkes, D. *et al.* Trial by Dutch laboratories for evaluation of non-invasive prenatal testing. Part I-clinical impact. *Prenat Diagn* **36**, 1083–1090 (2016).

3. Lubotzky, A. *et al.* Liquid biopsy reveals collateral tissue damage in cancer. *JCI Insight* **7**, (2022).

4. Bloom, R. D. *et al.* Cell-Free DNA and Active Rejection in Kidney Allografts. *J Am Soc Nephrol* **28**, 2221–2232 (2017).

5. Analysis of Plasma Epstein–Barr Virus DNA to Screen for Nasopharyngeal Cancer. *New England Journal of Medicine* **378**, 973–973 (2018).

6. Lennon, A. M. *et al.* Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science* **369**, (2020).

7. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature 2018 563:7732* **563**, 579–583 (2018).

8. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385 (2019).

9. Peneder, P. *et al.* Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nature Communications 2021 12:1* **12**, 1–16 (2021).

10. Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nature Medicine 2020 26:7* **26**, 1114–1124 (2020).

11. Widman, A. J. *et al.* Machine learning guided signal enrichment for ultrasensitive plasma tumor burden monitoring. *bioRxiv* 2022.01.17.476508 (2022) doi:10.1101/2022.01.17.476508.

12. Weber, B. *et al.* Detection of EGFR mutations in plasma and biopsies from non-small cell lung cancer patients by allele-specific PCR assays. *BMC Cancer* **14**, (2014).

13. cobas® EGFR Mutation Test v2. https://diagnostics.roche.com/be/en/products/params/cobas-egfr-mutation-test-v2.html.

14. Cheng, D. T. *et al.* Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *Journal of Molecular Diagnostics* **17**, 251–264 (2015).

15. Siravegna, G., Marsoni, S., Siena, S. & Bardelli, A. Integrating liquid biopsies into the management of cancer. *Nature Reviews Clinical Oncology 2017 14:9* **14**, 531–548 (2017).

16. Stroun, M., Lyautey, J., Lederrey, C., Olson-Sand, A. & Anker, P. About the possible origin and mechanism of circulating DNA: Apoptosis and active DNA release. *Clinica Chimica Acta* **313**, 139–142 (2001).

17. Jahr, S. *et al.* DNA fragments in the blood plasma of cancer patients: Quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer research (Chicago, Ill.)* **61**, 1659–1665 (2001).

18. Loyfer, N. *et al.* A human DNA methylation atlas reveals principles of cell type-specific methylation and identifies thousands of cell type-specific regulatory elements. *bioRxiv* (2022) doi:10.1101/2022.01.24.477547.

19. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci USA* **112**, E5503–E5512 (2015).

20. Leon, S. *et al.* Free DNA in the Serum of Cancer Patients and the Effect of Therapy, *Cancer Research* **37**(3):646-50 (1977)

21. Stroun, M. *et al.* Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46**, 318–322 (1989).

22. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).

23. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer 2017 17:4* **17**, 223–238 (2017).

24. Swisher, E. M. *et al.* Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. *Am J Obstet Gynecol* **193**, 662–667 (2005).

25. Kimura, H. *et al.* Detection of Epidermal Growth Factor Receptor Mutations in Serum as a Predictor of the Response to Gefitinib in Patients with Non–Small-Cell Lung Cancer. *Clinical Cancer Research* **12**, 3915–3921 (2006).

26. Shulman, D. S. *et al.* An international working group consensus report for the prioritization of molecular biomarkers for Ewing sarcoma. *npj Precision Oncology 2022 6:1* **6**, 1–11 (2022).

27. Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* **9**, (2017).

28. Alejandro Sweet-Cordero, E. & Biegel, J. A. The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. *Science (1979)* **363**, 1170–1175 (2019).

29. Caggiano, C. *et al.* Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nature Communications 2021 12:1* **12**, 1–13 (2021).

30. Sun, K. *et al.* Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res* **29**, 418–427 (2019).

31. Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nature Communications 2019 10:1* **10**, 1–11 (2019).

32. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology 2019 20:10* **20**, 590–607 (2019).

33. Warren, J. D. *et al.* Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med* **9**, 1–9 (2011).

34. Grünewald, T. G. P. *et al.* Ewing sarcoma. *Nature Reviews Disease Primers 2018 4:1* **4**, 1–22 (2018).

35. Krumbholz, M. *et al.* Quantification of translocation-specific ctDNA provides an integrating parameter for early assessment of treatment response and risk stratification in ewing sarcoma. *Clinical Cancer Research* **27**, 5922–5930 (2021).

36. Sheffield, N. C. *et al.* DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. *Nature Medicine 2017 23:3* **23**, 386–395 (2017).

37. Leontiou, C. A. *et al.* Bisulfite conversion of DNA: Performance comparison of different kits and methylation quantitation of epigenetic biomarkers that have the potential to be used in non-invasive prenatal testing. *PLoS One* **10**, (2015).

38. Tanaka, K. & Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg Med Chem Lett* **17**, 1912–1915 (2007).

39. Worm Ørntoft, M. B., Jensen, S. Ø., Hansen, T. B., Bramsen, J. B. & Andersen, C. L. Comparative analysis of 12 different kits for bisulfite conversion of circulating cell-free DNA. *Epigenetics* **12**, 626–636 (2017).

40. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel)* **5**, (2016).

41. Erger, F. *et al.* CfNOMe - A single assay for comprehensive epigenetic analyses of cell-free DNA. *Genome Med* **12**, 1–14 (2020).

42. QIAamp MinElute ccfDNA Kits. https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/cell-free-dna/qiaamp-minelute-ccfdna-kits/.

43. Krumbholz, M. *et al.* Genomic EWSR1 Fusion Sequence as Highly Sensitive and Dynamic Plasma Tumor Marker in Ewing Sarcoma. *Clin Cancer Res* **22**, 4356–4365 (2016).

44. NEBNext® Enzymatic Methyl-seq Kit | NEB. https://international.neb.com/products/e7120-nebnext-enzymatic-methyl-seq-kit#Protocols,%20Manuals%20&%20Usage_Manuals.

45. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

46. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868 (2005).

47. Merkel, A. *et al.* gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics* **35**, 737–742 (2019).

48. pUC19 Vector | NEB. https://international.neb.com/products/n3041-puc19-vector.

49. Veillard, A. C., Datlinger, P., Laczik, M., Squazzo, S. & Bock, C. Diagenode® Premium RRBS technology: cost-effective DNA methylation mapping with superior coverage. *Nature Methods 2016 13:2* **13**, i–ii (2016).

50. Picard Tools - By Broad Institute. https://broadinstitute.github.io/picard/.

51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

52. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

53. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**, 90–95 (2007).

54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

55. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, (2017).

56. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

57. McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference* 56–61 (2010).

58. Bewick, V., Cheek, L. & Ball, J. Statistics review 14: Logistic regression. *Crit Care* **9**, 112 (2005).

59. Breiman, L. Random Forests. *Machine Learning 2001 45:1* **45**, 5–32 (2001).

60. Lewis, D. D. Naive(Bayes)at forty: The independence assumption in information retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **1398**, 4–15 (1998).

61. Taunk, K., De, S., Verma, S. & Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019* 1255–1260 (2019).

62. Boser, B. E., Guyon, I. M. & Vapnik, V. N. Training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* 144–152 (1992) doi:10.1145/130385.130401.

63. Cristianini, N. & Ricci, E. Support Vector Machines. *Encyclopedia of Algorithms* 928–932 (2008)

64. Delashmit, W., Missiles, L. & Manry, M. Recent Developments in Multilayer Perceptron Neural Networks. *Proceedings MAESC 2005* (2005).

65. Ko, J. *et al.* Machine learning to detect signatures of disease in liquid biopsies – a user's guide. *Lab Chip* **18**, 395–405 (2018).

66. Kirasich, K., Smith, T. & Sadler, B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review* **1**, (2018).

67. Stackpole, M. L. *et al.* Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. *Nature Communications 2022 13:1* **13**, 1–12 (2022).

68. Jain, A. K., Duin, R. P. W. & Mao, J. Statistical pattern recognition: A review. *IEEE Trans Pattern Anal Mach Intell* **22**, 4–37 (2000).

69. GitHub - smazzanti/mrmr: mRMR (minimum-Redundancy-Maximum-Relevance) for automatic feature selection at scale. https://github.com/smazzanti/mrmr.

70. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans Pattern Anal Mach Intell* **27**, 1226–1238 (2005).

71. Zhao, Z., Anand, R. & Wang, M. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. *Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019* 442–452 (2019)

72. sklearn.feature_selection.f_classif — scikit-learn 1.1.3 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html#sklearn.feature_selection.f_classif.

73. Di, J. & Gou, X. Bisecting K-means Algorithm Based on K-valued Selfdetermining and Clustering Center Optimization. *J Comput (Taipei)* 588–595 (2018)

74. Kupkova, K. *et al.* GenomicDistributions: fast analysis of genomic intervals with Bioconductor. *BMC Genomics* **23**, 1–6 (2022).

75. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

76. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).

77. Farlik, M. *et al.* Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep* **10**, 1386–1397 (2015).

78. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882–D889 (2020).

79. Lim, J. Q. *et al.* BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. *Genome Biol* **13**, R82 (2012).

80. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science,* **372**, (2021).

81. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16**, S1 (2015).

82. Han, D. S. C. *et al.* The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. *Am J Hum Genet* **106**, 202–214 (2020).

83. Keogh, E. & Mueen, A. Curse of Dimensionality. *Encyclopedia of Machine Learning and Data Mining* 314–315 (2017) doi:10.1007/978-1-4899-7687-1_192.

84. Berisha, V. *et al.* Digital medicine and the curse of dimensionality. *npj Digital Medicine 2021 4:1* **4**, 1–8 (2021).

85. Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nature Biotechnology 2016 34:5* **34**, 547–555 (2016).

86. Barefoot, M. E. *et al.* Cell-free, methylated DNA in blood samples reveals tissue-specific, cellular damage from radiation treatment. *bioRxiv* 2022.04.12.487966 (2022)

Appendix

Parameter settings for gemBS config files:

[index]:

sampling_rate = 4

[mapping]:

Pipeline informed which sequences to use for under and over conversion estimation, state strandedness of library.
underconversion_sequence = Lambda
overconversion_sequence = pUC19

for EM-seq samples:
non_stranded = False

for Ewing, MSC and other cancer references (as this is setting is universally applicable):

non_stranded = True

[calling]:

mapq_threshold = 10
qual_threshold = 13
reference_bias = 2
left_trim = 5
right_trim = 0
keep_improper_pairs = False
haploid = False
conversion = 0.01, 0.05
contig_pool_limit = 25000000

For RRBS samples (as recommended in the gemBS documentation):

keep_duplicates = True

For WGBS samples:

keep_duplicates = False

[extract]:

Choose the correct output formats and extraction approaches: output methylation data only in gemBS style bed format, map all reads to + strand coordinates and avoid rigorous filtering for genotype certainty.

strand_specific = False
make_bedmethyl = False
make_bigwig = False
phred_threshold = 1
make_cpg =True
make_non_cpg = False

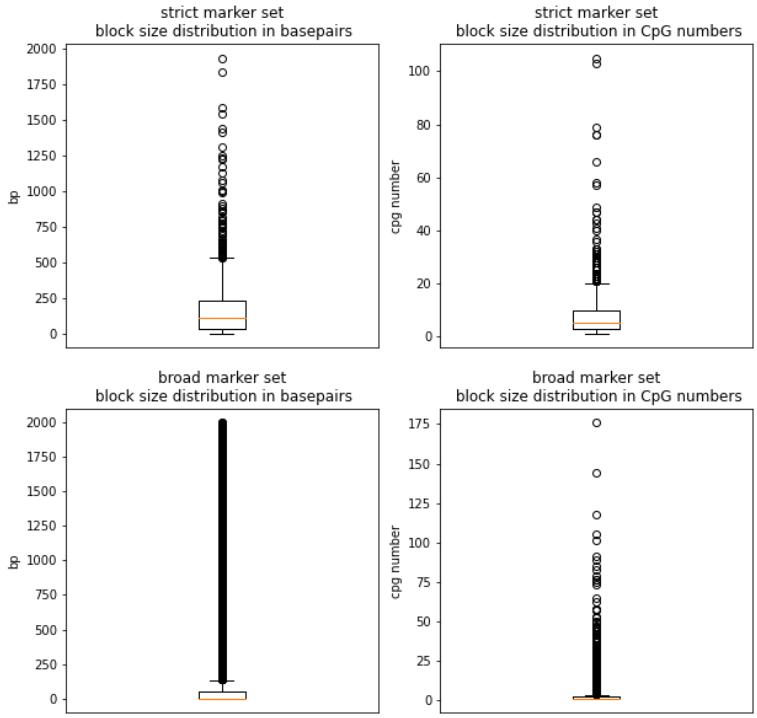
clinical data

Supplemental Table 2: Clinical data for cfDNA samples from cancer patients

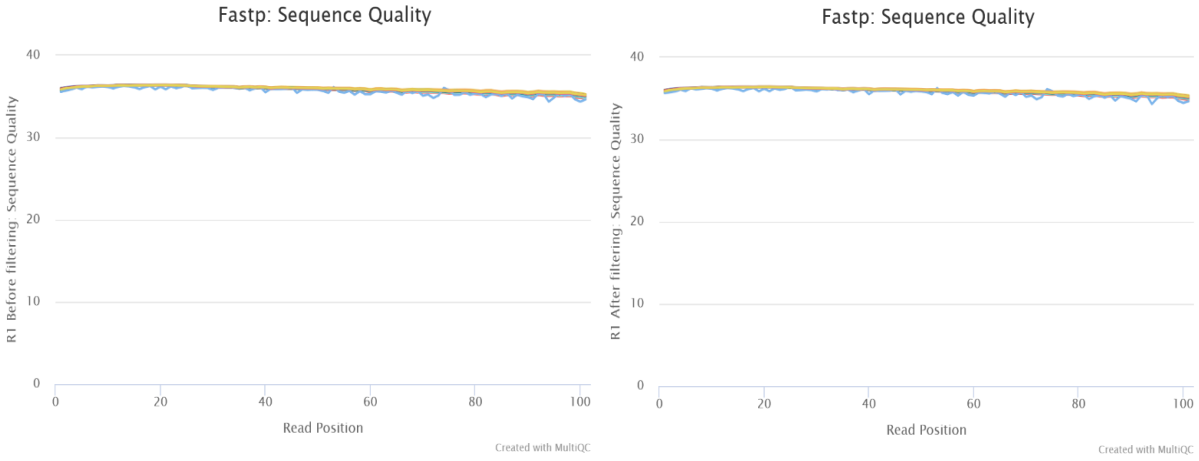
	Sample type	EM-seq name	Center	Gender	Age at diagnosis (years)	Metastases at diagnosis	Patient has relapsed
0	EwS	EWS_003_01_01_EM	GER	female	9.74	metastases	yes
1	EwS	EWS_005_01_01_EM	GER	male	15.93	no_metastases	yes
2	EwS	EWS_006_01_01_EM	GER	male	18.61	metastases	yes
3	EwS	EWS_008_01_01_EM	AUT (Biobank)	female	11.25	metastases	NaN
4	EwS	EWS_017_01_01_EM	FR	female	17.40	no_metastases	no
5	EwS	EWS_021_01_01_EM	FR	male	18.30	metastases	yes
6	EwS	EWS_027_01_01_EM	FR	female	4.10	no_metastases	no
7	EwS	EWS_028_01_01_EM	FR	male	14.90	no_metastases	yes
8	EwS	EWS_031_01_01_EM	GER	male	52.31	no_metastases	yes
9	EwS	EWS_032_01_01_EM	NO	male	8.00	no_metastases	yes
10	EwS	EWS_033_01_01_EM	NO	female	1.00	no_metastases	no
11	EwS	EWS_038_01_01_EM	NO	male	24.00	no_metastases	yes
12	EwS	EWS_039_01_01_EM	NO	male	23.00	no_metastases	yes
13	EwS	EWS_041_01_01_EM	NO	male	22.00	no_metastases	no
14	EwS	EWS_047_01_01_EM	NO	female	25.00	no_metastases	no
15	EwS	EWS_063_01_01_EM	AUT	male	14.20	metastases	yes
16	EwS	EWS_065_01_01_EM	GER	female	17.49	metastases	yes
17	EwS	EWS_098_01_01_EM	GER	male	nan	metastases	no
18	EwS	EWS_097_01_01_EM	GER	female	19.79	no_metastases	no
28	Non-EwS sarcoma	OST_001_01_01_EM	AUT	male	8.99	no_metastases	NaN
29	Non-EwS sarcoma	OST_004_01_01_EM	GER	female	14.09	NaN	yes
30	Non-EwS sarcoma	OST_005_01_01_EM	GER	male	15.54	NaN	yes
31	Non-EwS sarcoma	RMS_010_01_01_EM	GER	male	3.84	NaN	no
32	Non-EwS sarcoma	RMS_003_01_01_EM	AUT	male	5.39	metastases	NaN
33	Non-EwS sarcoma	RMS_006_01_01_EM	GER	male	0.95	NaN	no

Supplemental Table 1: Accessory data for cfDNA samples from healthy control individuals.

	Sample type	EM-seq name	Center	Gender
19	Healthy control	CTR_002_01_01_EM	AUT	male
20	Healthy control	CTR_003_01_01_EM	AUT	female
21	Healthy control	CTR_004_01_01_EM	AUT	male
22	Healthy control	CTR_006_01_01_EM	AUT	female
23	Healthy control	CTR_008_01_01_EM	AUT (Red Cross)	male
24	Healthy control	CTR_010_01_01_EM	AUT (Red Cross)	female
25	Healthy control	CTR_011_01_01_EM	AUT (Red Cross)	male
26	Healthy control	CTR_012_01_01_EM	AUT (Red Cross)	male
27	Healthy control	CTR_014_01_01_EM	AUT (Red Cross)	female
34	Healthy control	CTR_001_04_01_EM	NaN	NaN
35	Healthy control	CTR_001_04_02_EM	NaN	NaN
36	Healthy control	CTR_001_04_03_EM	NaN	NaN
37	Healthy control	CTR_001_04_04_EM	NaN	NaN
38	Healthy control	CTR_001_04_05_EM	NaN	NaN
39	Healthy control	CTR_001_04_06_EM	NaN	NaN



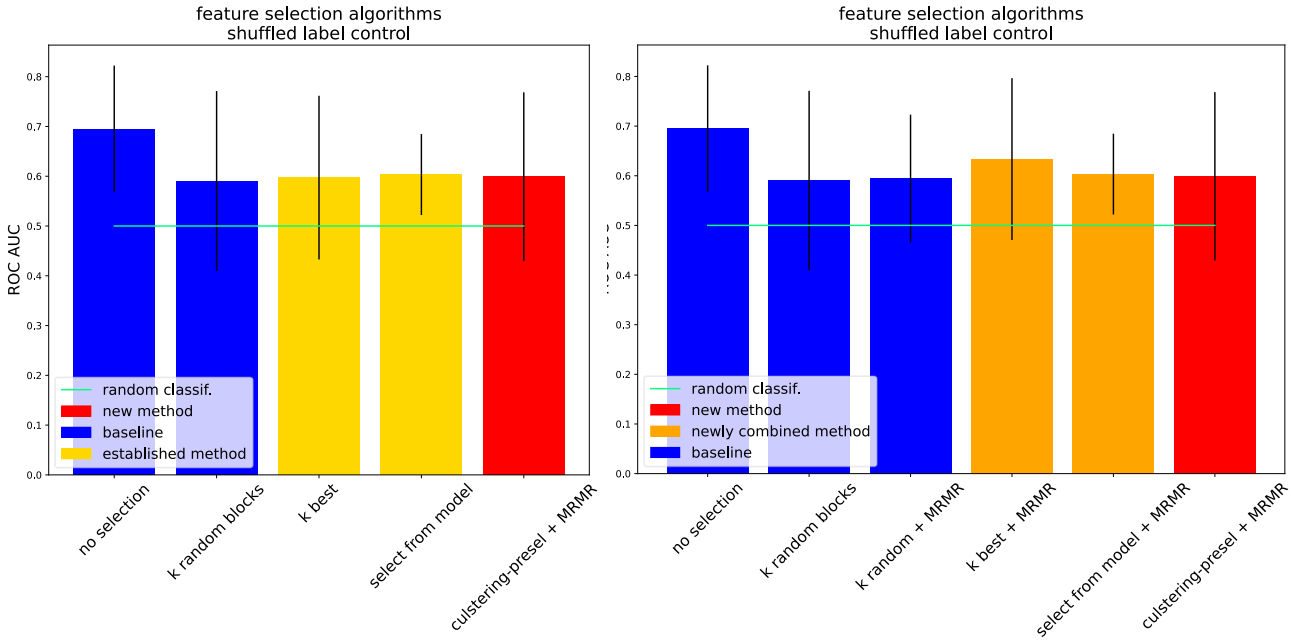
Supplemental Figure 1: Overview of block size distributions for strict and broad Ewing markers, sizes given in base pairs length and number of CpGs. The blocks in the strict marker set show a median base length of 115 bp with 5th and 95th percentile at 1 and 613 bp, median CpG number of 5 with 5th and 95th percentile at 1 and 23 CpGs. The blocks in the broad marker set show a median base length of 1 bp with 5th and 95th percentile at 1 and 404 bp, median CpG number of 1 with 5th and 95th percentile at 1 and 6 CpGs



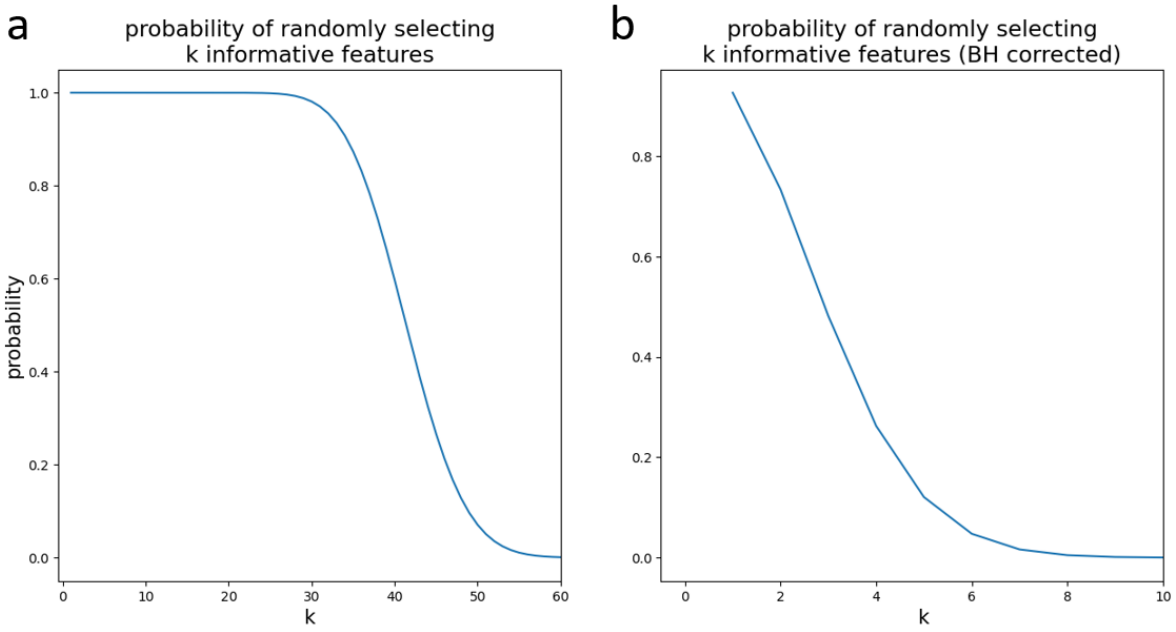
Supplemental Figure 2: Sequencing quality for EM-seq cfDNA samples. Sequencing quality was good with phred scores consistently above 35 along reads even before filtering, most reads (>99%) did pass the quality filter, no outlier samples were recognized. The number of reads passing the filter ranged between ~350 million and ~520 million.

Supplemental Table 3: Mapping statistics for EM-seq cfDNA samples

Sample Name			
CTR_001_04_01_EM_S115242	469.2	168 bp	11.0X
CTR_001_04_02_EM_S115241	428.8	168 bp	10.0X
CTR_001_04_03_EM_S115276	475.1	169 bp	11.0X
CTR_001_04_04_EM_S115264	404.5	168 bp	10.0X
CTR_001_04_05_EM_S115263	334.9	170 bp	8.0X
CTR_001_04_06_EM_S115254	376.6	170 bp	8.0X
CTR_002_01_01_EM_S115244	467.4	169 bp	11.0X
CTR_003_01_01_EM_S115243	438.5	171 bp	11.0X
CTR_004_01_01_EM_S115246	422.4	168 bp	10.0X
CTR_006_01_01_EM_S115266	426.6	173 bp	11.0X
CTR_008_01_01_EM_S115255	398.3	169 bp	9.0X
CTR_010_01_01_EM_S115256	303.4	179 bp	6.0X
CTR_011_01_01_EM_S115253	323.0	173 bp	6.0X
CTR_012_01_01_EM_S115268	386.6	168 bp	9.0X
CTR_014_01_01_EM_S115265	306.8	174 bp	6.0X
EWS_003_01_01_EM_S115273	374.8	165 bp	8.0X
EWS_005_01_01_EM_S115247	396.0	167 bp	8.0X
EWS_006_01_01_EM_S115251	421.4	164 bp	9.0X
EWS_008_01_01_EM_S115275	491.5	163 bp	10.0X
EWS_008_01_02_EM_1to2_CTR_001_02_S115278	429.1	165 bp	9.0X
EWS_008_01_03_EM_1to10_CTR_001_02_S115277	441.7	168 bp	10.0X
EWS_008_01_04_EM_1to101_CTR_001_02_S115280	467.0	169 bp	11.0X
EWS_008_01_05_EM_1to1001_CTR_001_02_S115279	470.4	169 bp	11.0X
EWS_008_01_06_EM_1to10001_CTR_001_02_S115235	420.6	169 bp	10.0X
EWS_008_01_07_EM_1to100001_CTR_001_02_S115234	459.5	169 bp	11.0X
EWS_017_01_01_EM_S115249	320.8	163 bp	7.0X
EWS_021_01_01_EM_S115274	327.9	167 bp	7.0X
EWS_027_01_01_EM_S115239	308.4	173 bp	7.0X
EWS_028_01_01_EM_S115248	393.7	157 bp	8.0X
EWS_031_01_01_EM_S115272	453.4	173 bp	11.0X
EWS_032_01_01_EM_S115269	428.7	166 bp	9.0X
EWS_033_01_01_EM_S115271	442.7	168 bp	10.0X
EWS_038_01_01_EM_S115260	485.9	167 bp	11.0X
EWS_039_01_01_EM_S115236	412.4	164 bp	9.0X
EWS_041_01_01_EM_S115250	470.1	155 bp	10.0X
EWS_047_01_01_EM_S115262	332.7	168 bp	7.0X
EWS_063_01_01_EM_S115252	383.8	167 bp	8.0X
EWS_065_01_01_EM_S115238	425.4	165 bp	10.0X
EWS_097_01_01_EM_S115259	489.0	169 bp	12.0X
EWS_098_01_01_EM_S115261	390.4	171 bp	9.0X
OST_001_01_01_EM_S115257	492.4	165 bp	11.0X
OST_004_01_01_EM_S115258	375.5	169 bp	9.0X
OST_005_01_01_EM_S115270	359.9	168 bp	8.0X
RMS_003_01_01_EM_S115245	419.1	166 bp	8.0X
RMS_006_01_01_EM_S115237	387.1	169 bp	9.0X
RMS_010_01_01_EM_S115267	390.3	167 bp	9.0X



Supplemental Figure 3: Shuffled label control for feature selection procedures. All examples show poor performance as desired.



Supplemental Figure 4: In Figure 12, a remarkably good classification performance has been observed on features randomly drawn from the full data set. This phenomenon could potentially be explained by a large fraction of total features actually being informative. To see if this explanation is plausible, informativeness of features was evaluated according to ANOVA f score and associated p values. 20.5% of total features showed significant information content with respect to Ewing sarcoma/control labeling and 1.3% of total features showed significant information content after multiple testing correction via Benjamini-Hochberg method. With these numbers, the probabilities of finding at least k significant features among 200 features randomly picked from the full data set were calculated using the hypergeometric distribution. (a) When considering features significant without multiple testing correction, one can be confident to find at least 30 of those features in a set of 200 randomly picked features. (b) When considering features significant after multiple testing correction, one finds a probability of 0.92 to find at least one such feature and with a probability of ~0.5 one finds at least three in a set of 200 randomly picked features. It has to be noted that multiple testing correction for ~7 million tests is very restrictive, and the true number of meaningfully relevant features might lie between numbers of significantly informative features with and without correction. Overall, it does not seem unplausible that a considerable number of informative features gets picked when selecting 200 features at random.

Supplemental Table 4: Enrichment results for known Ewing sarcoma relevant region sets and region sets with general relevance in cfDNA versus particularly consistently selected marker by clusterMRMR. Ewing sarcoma relevant regions: binding sites of the EWS-FLI1 fusion protein, enhancers correlated or anticorrelated with the presence of the fusion protein and Ewing sarcoma specific DNAse hypersensitivity sites (DHSs). (a) Features selected in at least 2 of 9 training sets hardly show any overlap with the regions of interest, (b) features selected in at least 4 of 9 training sets do not show any overlap with the regions of interest.

a

region set	n overlaps	meanRnk	p value
ARMS_specific_DHSs	1	1.00	1.0
EWS-FLI1_correlated_H3K27ac	1	1.67	1.0
EWS-FLI1_anticorrelated_H3K27ac	0	3.00	1.0
EWS-FLI1_binding_sites	0	3.00	1.0
EwS_specific_DHSs	0	3.00	1.0
hematopoietic_specific_DHSs	0	3.00	1.0
liver_specific_DHSs	0	3.00	1.0
universal_DHSs	0	3.00	1.0

b

region set	n overlaps	meanRnk	p value
ARMS_specific_DHSs	0	1	1.0
EWS-FLI1_anticorrelated_H3K27ac	0	1	1.0
EWS-FLI1_binding_sites	0	1	1.0
EWS-FLI1_correlated_H3K27ac	0	1	1.0
EwS_specific_DHSs	0	1	1.0
hematopoietic_specific_DHSs	0	1	1.0
liver_specific_DHSs	0	1	1.0
universal_DHSs	0	1	1.0

Supplemental Table 5: Enrichment analysis results for particularly consistently selected marker regions selected by clusterMRMR versus the LOLA core database of genomic regions. (a) features selected in at least 2 of 9 training sets mainly showed enrichment in regions occupied with hematopoietic transcription factors and co regulators in peripheral blood cells and acute myelomonocytic leukemia (AML) cell lineages. Among those are the transcription factors are RUNX1, PU1 and ERG and the co regulators EP300 and HDAC1. (b, first row) Particularly consistently selected features (selected in at least 4 of 9 training sets) show enrichment in genomic regions occupied with FOXA2 (transcriptional activator of liver specific genes, also involved in embryonic development) in human embryonic stem cells. Top 10 entries of the tables are shown.

a

region set	n overlaps	meanRnk	p value
AML cell line carrying inv(16) translocation	10	39.3	0.004452
AML cell line carrying inv(16) translocation	7	43.3	0.012039
AML cell line carrying inv(16) translocation	6	43.3	0.013773
Peripheral blood monocytes separated by leukap...	9	47.7	0.011490
Peripheral blood (bone marrow mononuclear CD34+)	11	49.7	0.007973
Infantile acute megakaryoblastic leukaemia (AM...	4	50.7	0.032409
AML cell line carrying inv(16) translocation	7	51.0	0.019366
ChIP K562 Pol2	5	58.3	0.045866
ChIP K562 NF-E2	3	58.7	0.017402
cistrome_epigenome APL164 cells H3K9K14ac	8	59.0	0.025265
Mobilised CD34+ cells isolated from the	4	60.3	0.060012

b

region set	n overlaps	meanRnk	p value
Day 5 of in vitro differentiation	2	7.00	0.062419
Hematopoietic;Skin	1	7.33	0.030012
Liver	1	8.00	0.041665
ChIP K562 HDAC6_(A301-341A)	1	9.00	0.069682
ChIP K562 SP2_(SC-643)	1	9.67	0.096821
ChIP K562 p300	1	10.30	0.110754
Weak lymphoblast specific	1	11.00	0.111974
Skin;Epithelial	1	11.70	0.131120
ChIP K562 HDAC2_(SC-6296)	1	12.70	0.150467
ChIP K562 HDAC2_(A300-705A)	1	13.30	0.151519
cistrome_cistrome HepG2 liver cells HNF4A	2	14.30	0.145819

Zusammenfassung (Deutsch)

Tumorerkennung und Risikostratifizierung bei Ewing-Sarkom durch methylierungsbasierte Flüssigbiopsieanalyse

Flüssigbiopsie ist eine Methode, die Zusehens in der Diagnose und Überwachung von Krebs Anwendung findet. Während viele Flüssigbiopsie-Ansätze in der Krebsdiagnose auf krebsspezifischen Mutationen beruhen, sind solche genetischen Ansätze für viele pädiatrische Krebsarten aufgrund ihrer geringen Mutationslast nur von begrenztem Nutzen. Epigenetische Ansätze stellen eine Alternative dar und haben sich in mehreren Studien bereits als nützlich erwiesen. In dieser Studie habe ich maschinelles Lernen auf DNA-Methylierungsdaten angewandt, die aus zellfreier DNA im Blut gewonnen wurden, um minimal invasive Biomarker für das Ewing-Sarkom zu entwickeln. Dabei handelt es sich um einen pädiatrischen Knochenkrebs, für den ungelöste klinische Herausforderungen existieren. Ich erreichte Tumordetektion mit 100 % Sensitivität und 95 % Spezifität bei nur 0,55 % Tumoranteil an der zellfreien DNA im Blut. Auch bei der Erkennung von Metastasen und der Vorhersage von Rückfällen wurde eine vielversprechende Leistung erzielt, was auf die potenzielle Anwendbarkeit von Methylierungsinformationen als prognostische Biomarker beim Ewing-Sarkom schließen lässt. Bestandteil der Strategie bezüglich maschinellen Lernens war auch die Entwicklung eines neuen Ansatzes zur Merkmalsauswahl („feature selection“), den ich clusterMRMR genannt habe. Diese könnte auch für Klassifizierungsaufgaben außerhalb der Biologie Anwendung finden. Bei der Auswahl von Merkmalen für die Unterscheidung zwischen Ewing- und Kontrollgruppe zeigte sich, dass die via clusterMRMR ausgewählten Merkmale in genomischen Regionen mit bekannter Relevanz für hämatopoetische Zellen angereichert sind, was die Hypothese stützt, dass solche Regionen Kandidaten für krebsübergreifende Marker sein könnten. Schließlich konnte ich bestätigen, dass die enzymatische Methylierungssequenzierung („enzymatic methylation sequencing“, „EM-seq“) keine Verzerrungen der Längenverteilung der zellfreien DNA-Fragmente verursacht. Dies eröffnet die Möglichkeit, in Zukunft gleichzeitig sowohl Methylierungs- als auch Fragmentierungsanalysen mit EM-seq Daten durchzuführen.

Abstract (English)

Tumor detection and risk stratification in Ewing sarcoma via methylation based liquid biopsy analysis

Liquid biopsy is an emerging method in cancer diagnosis and monitoring. While many liquid biopsy approaches in cancer diagnosis rely on cancer specific mutations, such genetic approaches are only of limited use for many pediatric cancers due to their low mutational burden. Epigenetic approaches represent an alternative and have already proven useful in liquid biopsy assays. In this study, I applied machine learning to DNA methylation data obtained from cell-free DNA in the blood to develop minimally invasive biomarkers for Ewing sarcoma, a pediatric bone cancer with unmet clinical needs. I achieved sensitive tumor detection at tumor fractions as low as 0.55% with 100% sensitivity at 95% specificity. Reasonable performance was also achieved for metastasis detection and relapse prediction tasks, suggesting potential applicability of methylation information as prognostic biomarker in Ewing sarcoma. The applied machine learning framework also includes a newly developed feature selection approach termed clusterMRMR, which might be of general relevance also for classification tasks outside the field of biology. When tasked to select features informative for Ewing versus control classification, features yielded by clusterMRMR were found to be enriched in genomic regions with known relevance in hematopoietic cells, which lends support to the hypothesis that such regions might be candidates for pan-cancer markers. Finally, I was able to confirm that enzymatic methylation sequencing does not introduce bias to the length distribution of cell-free DNA fragments. This opens up the possibility to simultaneously perform both methylation and fragmentation-based analysis on enzymatic methylation sequencing data in the future.