medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Channel Capacity of Genome-Wide Cell-Free DNA Fragment Length Distribution in Colorectal Cancer

Alexandre Matov[1, 2, †]

[1] Department of Clinical Medicine, Aarhus University Hospital, Århus, Denmark

[2] Present address: DataSet Analysis LLC, 155 Jackson St, San Francisco, CA 94111

[†] Corresponding author:

email: matov@datasetanalysis.com

Key words: Patient Stratification, Disease Classification, Cell-Free DNA, Fragmentation Patterns, Low-Coverage Whole Genome Sequencing, Kullback-Leibler Divergence, Broadcast Channel Capacity, Degraded Broadcast Channel, Achievable Rate Regions, Directed Markov Chain with Absorbing States

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

ABSTRACT

Each piece of cell-free DNA (cfDNA) has a length determined by the exact metabolic conditions in the cell it belonged to at the time of cell death. The changes in cellular regulation leading to a variety of patterns, which are based on the different number of fragments with lengths up to several hundred base pairs (bp) at each of the almost three billion genomic positions, allow for the detection of disease and also the precise identification of the tissue of their origin.

A Kullback-Leibler (KL) divergence computation identifies different fragment lengths and areas of the human genome, depending on the stage, for which disease samples, starting from pre-clinical disease stages, diverge from healthy donor samples. We provide examples of genes related to colorectal cancer (CRC), which our algorithm detected to belong to divergent genomic bins. The staging of CRC can be viewed as a Markov chain and that provides a framework for studying disease progression and the types of epigenetic changes occurring longitudinally at each stage, which might aid the correct classification of a new hospital sample.

In a new look to treat such data as grayscale value images, pattern recognition using artificial intelligence (AI) could be one approach to classification. In CRC, Stage I disease does not, for the most part, shed any tumor in circulation, making detection difficult for established machine learning (ML) methods. This leads to the deduction that early detection, where we can only rely on changes in the metabolic patterns, can be accomplished when the information is considered in its entirety, for example by applying computer vision methods.

INTRODUCTION

CRC is the second most common cancer-related cause of death worldwide (Siegel et al., 2020). Approximately two-thirds of newly-diagnosed patients present with a curable disease (Petrelli et al.,

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2017), but despite curatively intended treatment, up to 40% of them experience relapsed disease after an initial treatment (Jorgensen et al., 2015). In addition, about 86% of Stage I disease do not shed tumor in circulation, which leads to a decreased ability for early detection (Liu et al., 2020). A minimally-invasive analysis based on circulating cfDNA has emerged as a promising nucleic acids biomarker (Cristiano et al., 2019), but this method assumes very homogenous characteristics of the healthy population, which makes it hard to gain regulatory approval. Here, we attempt to consider the processes underlying the evolution of CRC as an electronics communication system and propose to visualize the cfDNA samples as images (Matov, 2024b).

COLORECTAL CANCER AS AN IMAGE

In CRC, patients go through a complex diagnostic paradigm. While healthy, the donors could be characterized as having different stages of co-morbidities. Next, pre-cancerous polyps might lead to adenomas of the colon or the rectum, which also have different stages. About 45% of the patients with advanced adenoma, and oftentimes people with low-grade adenoma or no adenoma at all, develop colorectal neoplasm, which has four stages. The extensive length of the colon makes tissue biopsies and surgical interventions uncertain procedures because of the sheer length of the organ. The physiology of the colon is particular and it is conceivable that changes, reported by the changes in the fragmentation patterns, in the local cellular regulation could offer information on the exact location of early disease. In disease, the variety of cell death fragmentation patterns reflects differential nucleosome packaging, chromatin remodeling and accessibility (van der Pol and Mouliere, 2019). DNA hypomethylation or the loss of a histone in a nucleosome, for instance, lead to a skewed fragment length distribution (Fig. 1A and Fig. 1B). In contrast, in healthy donor samples (Fig. 1C) the distribution is symmetric, centered around 168 bp. One can observe the consistent number of fragments across the genome for the healthy donor sample (Fig. 1C), which is in stark contrast to the copy-number variation seen as bright horizontal streaks in the two Stage IV samples (Fig. 1A and Fig. 1B). The difference image between a healthy

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

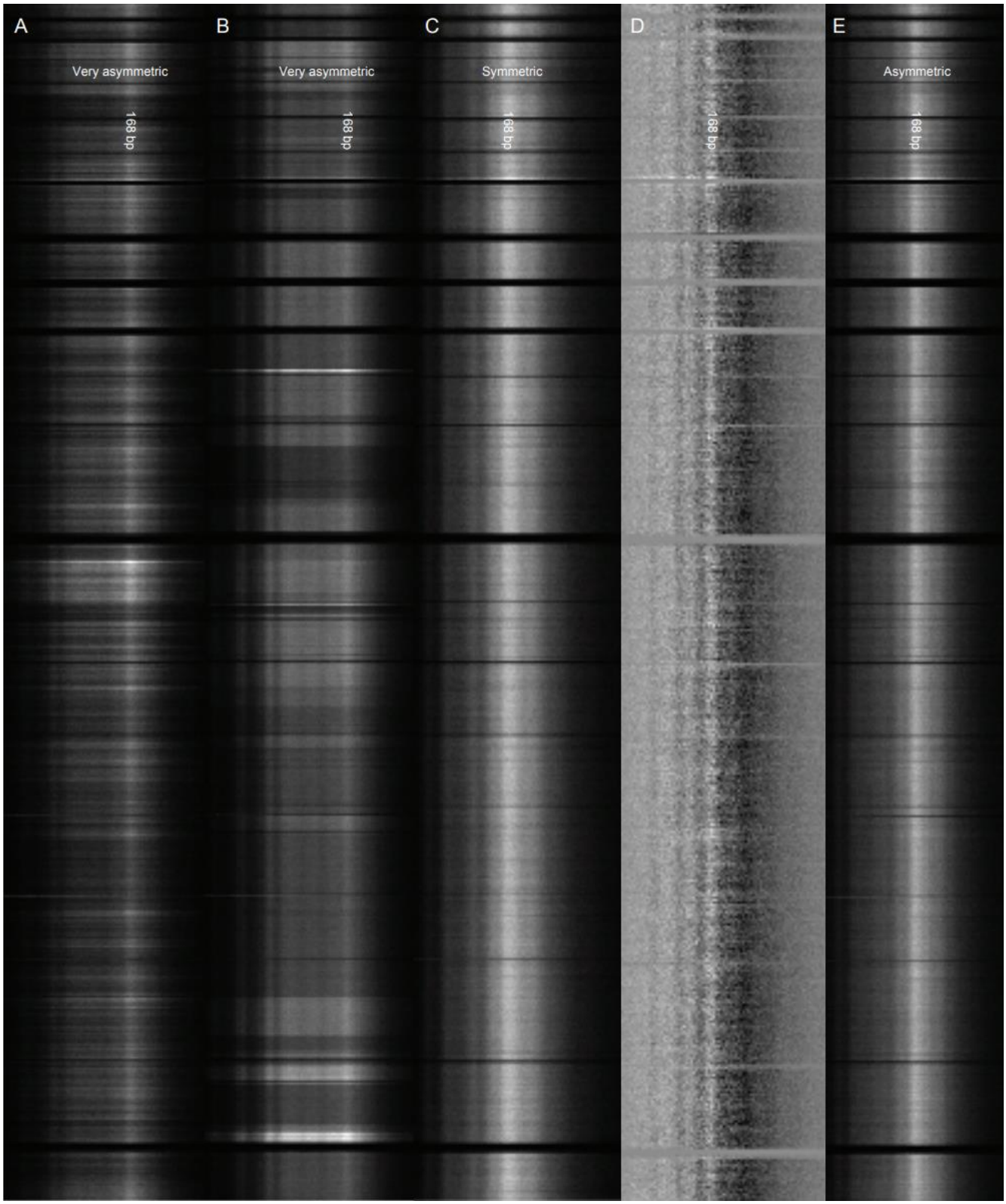
donor sample and a Stage I sample highlights the differences (Fig. 1D) present in cellular regulation and cell death in CRC Stage I (Fig. 1E). This holds true for low- and high-grade adenoma too.

DNA FRAGMENTATION

A way to classify samples is to compute the relative entropy for each fragmentation length by building a probability density function based on the fragments from all genomic positions. Such distributions consist of about 40 million fragments for each whole genome sequencing sample, which can be binned in genomic regions of, for instance, five mega bases (MB). In this scenario, for each fragment length, we could compare histograms of about 45,000 values from each region of the 23 chromosomes (Fig. 2). A KL divergence computation based on an adaptive minimax rate-optimal estimator (Han et al., 2016) of the changes in disease from healthy state(s) to precancerous lesion(s) to malignant tumor(s) can be presented as a heterogeneous directed Markov chain with absorbing states (Bremaud, 1999). Considering this Markov chain as a degraded broadcast channel, considerations regarding the capacity region (Han) and its estimation for a cohort of CRC samples (Bergmans, 1973) could aid the classification of a new patient sample in the clinic. To build the boundary of the capacity region for each stage, we assumed that the two peaks in the DNA fragmentation length KL divergence histogram (similar to Fig. 2, but for all cohorts, including the adenoma cohorts) provide the (X, Y) coordinates for which the boundary intersects with the X-axis (divergence value for the first peak, 0) and the Y-axis (0, divergence value for second peak) (boundaries of the capacity regions not shown). As different boundaries of the capacity regions (Cover, 1998) are defined by the probability density function of the cohort of each disease stage, a new sample will be classified according to the boundary it falls within. A new sample for an already existing patient which has become an outlier for the disease stage it has been assigned, and falls outside the boundary (Bergmans, 1974), will be classified in the next (or more advanced) disease stage. Oppositely, after a surgical resection or drug treatment, when tumors are removed or recede, and a new sample is collected and classified after an intervention, if it falls within an

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

earlier disease stage boundary of the capacity region, it will be re-classified as a lower burden disease or an earlier disease stage, according to the boundary it falls within.



medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 1. DNA fragmentation patterns in whole genome sequencing datasets. For advanced disease, the distributions are very asymmetric and the histograms are negatively skewed.

(A) CRC Stage IV. (B) CRC Stage IV (another patient). (C) Healthy donor sample, no co-morbidities. (D) A difference image between the images shown on (C) and (E). (E) CRC Stage I.

The genomic bins (of 5MB) are on the Y axis. Chromosome 1 is at the top of the image. The DNA fragment length is on the X axis, from left to right. Pixels with brighter intensity correspond to bins with a higher number of fragments. The peak in intensity is for each sample the vertical streak at 168 bp.

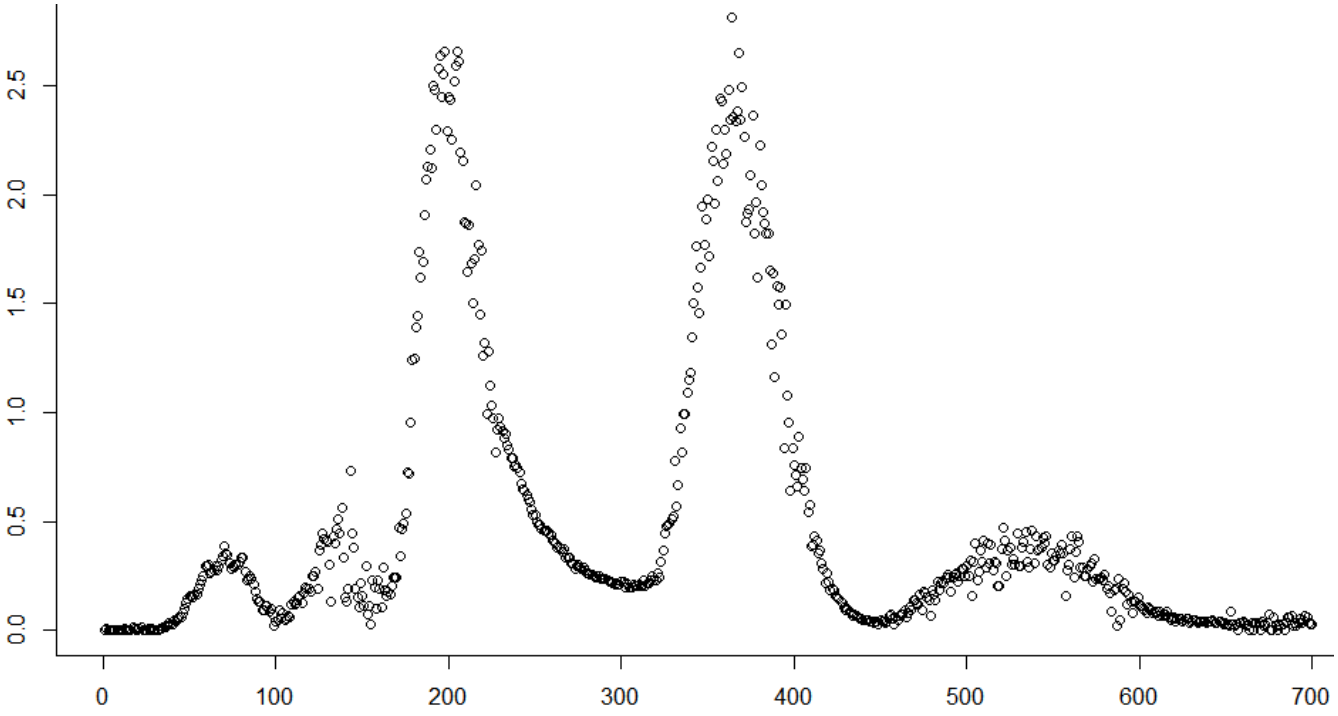


Figure 2. DNA fragmentation length KL divergence (in bits) from a CRC patient cohort of a healthy donor cohort.

The CRC cohorts consist of 42 patients in stages I-IV. The healthy donor cohort consists of 26 samples. The KL divergence (in bits) is on the Y axis. The maximal value is 2.8 bits divergence for DNA fragment length 364 bp. The second highest divergence value is 2.6 bits for fragment length 198 bp. The fragment length (in bp) is on the X axis, from left to right.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Observe that the two distributions are very similar, and consequently their divergence is low, for fragment length 168 bp.

The KL divergences from different cancer cohorts of healthy donor samples exhibit distributions, which are multi-modal, with different peaks being present for different cancers, defining potentially unique signatures. The two highest peaks on Fig. 3, for 364 base bp and 198 bp, are the result of significant differences in the median number of fragments in the genomic bins for these fragment lengths. While for healthy samples we measured that most genomic bins consist of about 20 fragments with length 364 bp (see the peak in the green histogram on Fig. 3), the CRC samples exhibit a very different distribution in which most genomic bins consist of less than 10 fragments with length 364 bp (see the peak in the pink histogram on Fig. 3).

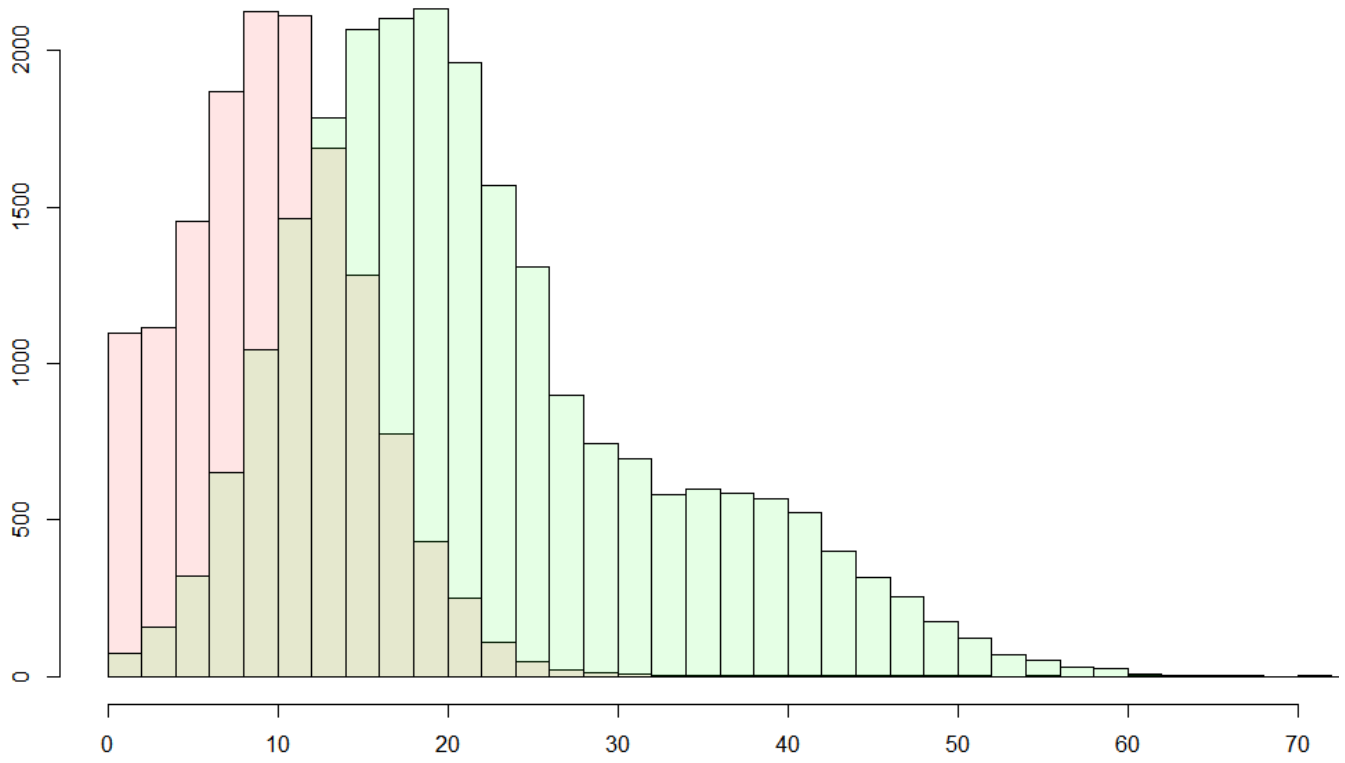


Figure 3. Probability density functions of all genomic bins of a CRC patient cohort (in pink color) and a healthy donor cohort (in green color) for fragment length 364 bp.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The CRC cohorts consist of 42 patients in Stages I-IV. The healthy cohort consists of 26 donor samples. The frequency, or the number of bins, is on the Y axis. The number of fragments in each bin, or per bin, is on the X axis.

Observe that the healthy samples have roughly 2.2-fold more fragments per bin for DNA fragment length 364 bp. These two histograms, and the corresponding differences in disease, for fragment length 198 bp are very similar.

The distinct per-disease peaks in the divergence from healthy samples (Tab. 1, col. 2) and the divergence of between-cancers signatures (Tab. 1, col. 3) are the result of differential epigenetic regulation of the different cancer types and can be used as diagnostic biomarkers for the detection of disease and identification of the tissue of origin. We measured the divergence in fragment lengths also between the different stages of CRC (Stage I – IV), including between cohorts of samples of pre-cancerous polyps (data not shown). Our approach to the identification of discriminative biomarkers in disease does not require any user-defined input metrics and it is data-driven. Further divergence analysis of the genomic bins only (for the most divergent fragment lengths - the peaks in the histogram on Fig. 2) demonstrated the ability of the method to pinpoint the areas of the human genome involved in pathogenesis and drug resistance/susceptibility.

HIERARCHICAL CLUSTERING

Hierarchical clustering of patient and healthy samples based on fragments with length 364 bp resulted in a few false positive (21.7% FP), but, importantly for the detection of sub-clinical disease, no false negative (Fig. 4) (0% FN). It suggested that the depletion of fragments from di-nucleosome-protected DNA in genomic regions associated with disabled antioxidant program (Barrett et al., 2013) in samples from healthy donors might be indicating pathogenesis and early CRC (Fig. 4). This poses the question whether the three healthy donors from the DELFI (Cristiano et al., 2019) study associated in our analysis with CRC have since the time of blood draw been diagnosed with CRC.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

One of the CRC-related genes, *AXIN2*, mutations in which have been associated with mismatch repair errors (Liu et al., 2000), falls within one of the most divergent genomic bins (Fig. 4). It has also been shown to interact with Glycogen synthase kinase-3β (Lejeune et al., 2006), which regulates microtubules in migrating cells (Kumar et al., 2009). Two other genomic bins that are the most divergent (Fig. 4) from the healthy cohort in CRC are those containing *RAP2B* (Yi et al., 2019) and *GPX3 (Zhang et al., 2020)*. The second gene is involved in the detoxification (reduction by GPX3, produced mainly in the kidneys (Avissar et al., 1994)) of soluble reactive oxygen species (Dixon and Stockwell, 2014). Interestingly, this new result indicates the presence of a divergence in the metabolic/redox patterns in CRC.

Cancer cohort	**Healthy cohort**	**Colorectal cancer cohort**
Colorectal cancer	364, 205	-
Ovarian cancer	359, 208	248, 175
Pancreatic cancer	203, 359	111, 269
Gastric cancer	122, 347	193, 138
Breast cancer	177, 333	176, 280
Bile duct cancer	200, 351	111, 164
Lung cancer	193, 343	121, 198

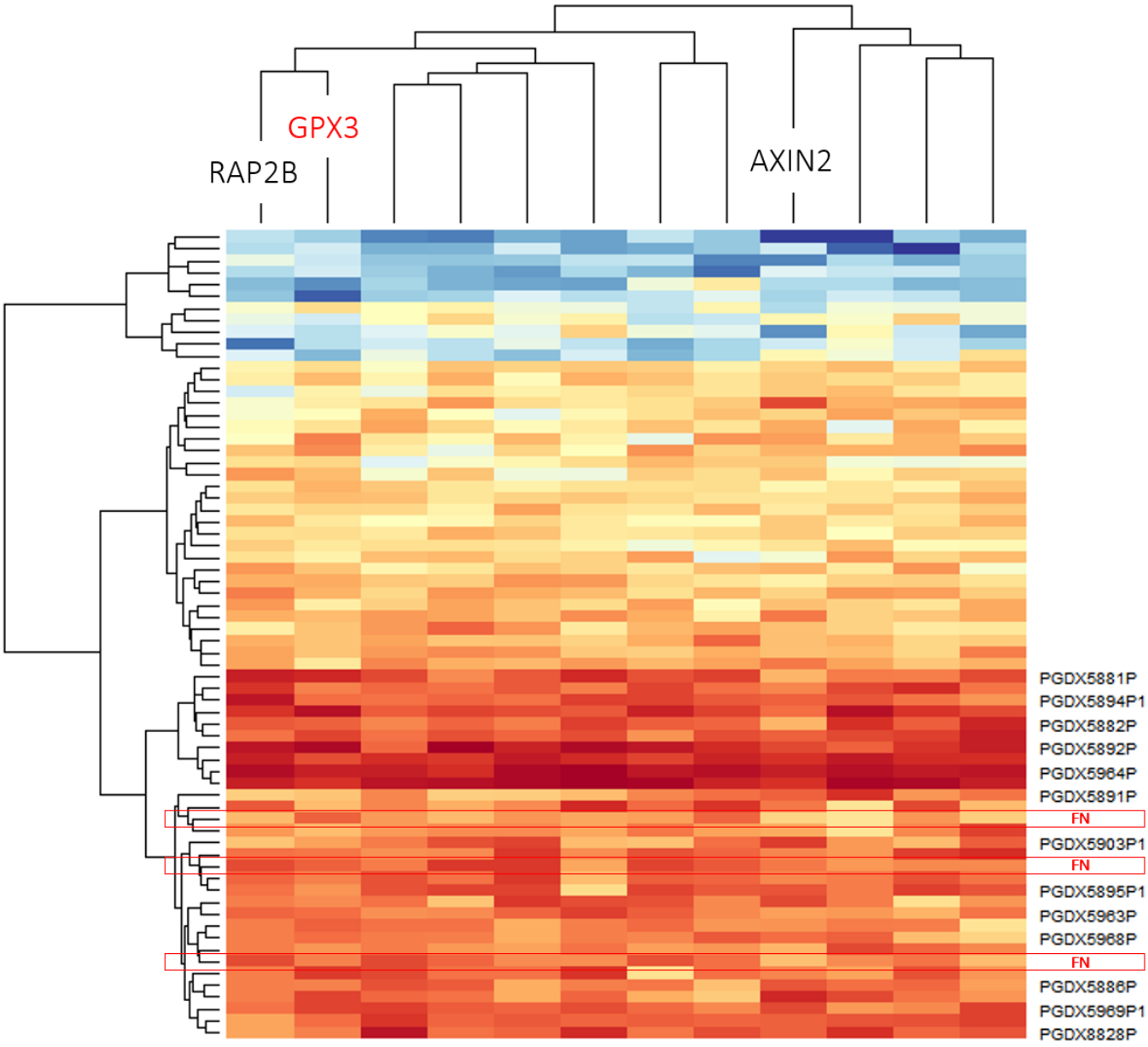
Table 1. DNA fragment lengths (in bp) for the two highest peaks (the primary peak is listed first and the secondary is listed second) of the KL divergence from seven cancers (see the list in the left column) of healthy donor samples (see the peak fragment lengths in the middle column) and of CRC (see the peak fragment lengths in the right column).

We measured the divergence in fragment lengths also between the different stages of CRC (Stage I – IV), including between cohorts of samples of pre-cancerous polyps (data not shown).

At least 8% of the fragments in each cohort belong to diverging populations of bins (divergence histograms not shown).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The clustering algorithm separates the samples into healthy donors (upper half of Fig. 4) and patient samples (lower half of Fig. 4). Remarkably, there are no patient samples grouped with the healthy donors, which indicates no false positive selections. There are, however, three false negative selections (labeled with FN in red color on Fig. 4) within the patient cluster (PGDX labels on Fig. 4). This result is promising in the context of the detection of sub-clinical and early disease; it could be further verified whether these three donor have developed CRC since the time of the blood draw.



medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 4. Hierarchical clustering of a cohort of 42 CRC patients, Stages I-IV (lower half of the image, labeled with PGDX followed by a number – some labeled are not displayed) and 26 healthy samples (upper half of the image, no label, with the exception of the three false negative, clustered with the patient samples, labeled with FN in red color) (Y axis).

The clustering is accomplished based on considering the 12 (out of 595) top most divergent 5 MB genomic bins (divergence histogram not shown) for DNA fragment length 364 bp (X axis).

Observe that the area of the genome containing *AXIN2*, a gene implicated in CRC, is among the most divergent.

It is conceivable that predictions regarding disease progression can be achieved using generative methods. Generative methods that produce novel samples from high-dimensional data distributions are finding widespread use, for example in speech synthesis. Generative adversarial network (GAN) models (Karras et al., 2017) consist of two CNNs, style-based generator and discriminator, which converge upon reaching Nash equilibrium and can be used to generate synthetic samples. Such sampler posterior distributions can be then used to reduce the complexity and improve the numerical convergence of predicting disease progression for any new sample using the Bayes formula, given there is available longitudinal data and transitional states for several patients and healthy donors. This approach may allow attempting to induce changes in reversal (Mullis, 1968) to the physiological deterioration occurring in disease, if detected early in a pre-clinical stage.

FERROPTOSIS AND ANGIOGENESIS

A possible mechanism leading to the differential fragmentation is the increase of ferroptotic cell death in CRC. The decrease of DNA fragments with lengths 198 bp, 364 bp, and to a lesser extend 521 bp (resulting in the three divergence peaks on Fig. 2) could be due to a partial switch from apoptosis in normal physiology to ferroptosis in CRC – thus suppressing the three apoptotic peaks of fragments from mono- (147+2x26, histones plus linkers), di- (2x147+3x23, histones plus linkers), and tri- (3x147+3x27,

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

histones plus linkers) nucleosome-protected DNA (Fig. 2). The reason behind our new results is that during apoptosis the cleavage of nuclear DNA results into fragments with length proportional to nucleosome size and resulting in patterned fragmentation. In somatic tissues, the apoptotic cleavage of DNA results in fragments of about 195 bp in lengths and multiples thereof (Ivanov et al., 2015). Ferroptosis, however, is characterized by non-patterned DNA fragmentation and not characterized by caspase-dependent cleavage and, thus, our analysis results demonstrate that there is a generation of less fragments with lengths 198 bp, 364 bp, and 521 bp.

Ferroptosis has a dual role in cancer. It plays a role in tumor initiation, tumorigenesis, which depends on inflammation-associated immunosuppression triggered by ferroptotic damage (Chen et al., 2021) and later, during treatment, in tumor suppression (Hangauer et al., 2017). Erastin, for instance, was discovered during a synthetic lethality screen with oncogenic RAS cells (Dolma et al., 2003). It lowers cysteine and, thus, the cells stop the synthesis of antioxidants/glutathione, and this activates voltage-dependent anion channels (VDAC) by reversing tubulin's inhibition on VDAC2/3 (Yagoda et al., 2007). VDAC is a mitochondrial protein, which is a novel target for anti-cancer drugs. Our analysis shows that for the genomic bin where *VDAC2* falls has an increase in the number of fragments with length 198 bp in CRC (data not shown), which offers a quantitative strategy for drug selection (Lemasters, 2017).

Within the fragments with length 198 bp, we also measured a divergence in CRC in the genomic bin in which falls *TCF7L2* (data not shown), which has been implicated in promoting migration and invasive behavior of human CRC cells (Wenzel et al., 2020). Interestingly, when we analyze the data for CRC Stage IV only (without Stages I-III), an additional peak appears on the divergence histogram from healthy donor samples at 129 bp (data not shown). Within the fragments with length 129 bp in Stage IV, for a few of the patients, there are about three-fold more fragments than the average in healthy samples in the genomic bin that covers the area of the human genome containing *VEGFC*. This gene has been

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

associated with disseminated epithelial tumor cells to regional lymph nodes (Van Trappen and Pepper, 2002). Thus, it can serve as an endothelial marker likely correlating with angiogenesis due to metastasis or minimal residue disease after a curative-intent surgery and can also offer a strategy for the selection of combination therapy using zaltrap or eylea (Muro et al., 2020).

DISCRETE BROADCAST CHANNEL

Visible in the difference image on Fig. 1D is the appearance of periodic vertical streaks around 90 to 150 bp. Visible are also horizontal streaks, likely resulting from errors in gene copy-number amplification in CRC. These examples demonstrate the patterns appearing in disease in comparison to normal physiology. Such grayscale images, in the thousands, can be utilized to train a generative transformer (Ren et al., 2024), or another large language model. This classification approach will ensure all epigenetic changes occurring in disease have been taken into account.

One possible avenue for classification is to derive the maximum likelihood estimate of the parameters of Markov Chain Monte Carlo sampling for time series prediction and supervised Bayesian learning (Chandra et al., 2019). Further, let the appearance of co-morbidities in healthy donor samples be a stationary Markov chain and CRC Stages I-IV denote its noisy version as a Hidden Markov Process (HMP), when corrupted by a discrete memoryless channel (DMC) (Marton, 1979), with channel capacity equal to the maximum of the KL divergence. The DMC (Gamal and Meulem, 1981) is completely characterized by the channel transition matrix, also known as the confusion matrix (Lee et al., 2015). Consider the HMP given by a binary symmetric channel with corrupted symmetric binary Markov source. One can approximate the entropy rate of a HMP via approximations of the stationary distribution of a related Markov process with high precision-complexity trade-off (Ordentlich and Weissman, 2011). Therefore, KL divergence can serve as a prognostic biomarker (Zhong et al., 2020)

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

based on longitudinal data prior to diagnosis, i.e., samples collected periodically from the same healthy donor earlier in life.

If we consider the transformations occurring in the DNA fragmentation patterns within a cohort of healthy samples so that all become alike to disease samples, we can view pathogenesis and disease progression of a population as a broadcast channel with memory and present it in terms of Gaussian multi-user parallel broadcast channels with identical code words (Matov, 1999). The achievable rate for the capacity of a degraded broadcast channel (in bits) is a function of the logarithm of the signal-to-noise ratio of the transmission signal and depends on the quality of the transmission medium. Next, the achievable rates for the capacity region of a family of parallel broadcast channels is given by the union of the overall capacity in each channel (Tse, 1999). Hence, any divergence in the fragmentation patterning within the healthy cohort (baseline dataset or parallel broadcast channels) would create an elevated noise floor and, thus, an overall unsolvable stochastic heterogeneity.

Each time a new patient sample is presented for classification, it would not be compared to samples collected from the same person longitudinally, in which most of the fragmentation pattern would be very similar to the previously collected healthy samples. Instead, it would be compared to a variety of fragmentation patterns in a whole cohort of (different people) healthy donors. Such comparisons will always generate noisy baseline datasets, even in the case of a very large healthy donor population, because a very few outliers will affect the overall quality of the healthy baseline dataset. For this reason, such population approach, because of the intrinsic variability in human, inherently impedes the ability of all currently utilized methods to detect disease in its early stage.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

CONCLUSIONS

In the specific case of CRC, this conclusion means that any non-neoplastic gastrointestinal (GI) metabolic divergence in samples from the healthy donor cohort considered as baseline during classification would impair the ability of the traditional ML classifiers to reliably detect neoplastic transformation (Wan et al., 2019). Oppositely, an image-based classification, as we propose here, could be in a position and better equipped to successfully detect and delineate both the fragmentation patterns resulting from tumors and those resulting from non-lethal, transient conditions, such as inflammation (Horiuchi et al., 2024). If this holds true, it would impact, besides CRC diagnosis, our ability to correctly diagnose other cancers of the GI tract as well.

We propose the utilization of the described here DNA fragmentation analytics to generate embeddings (Vaswani et al., 2017) for a generative transformer network in which the tokens are the KL divergence metrics of the samples. The embeddings can include DNA fragment length patterns before and after disease diagnosis. We will retrain a transformer network with a new set of tokens - with genetic mutations rather than words and with tumor burden values rather than sentences of human speech, i.e., we will retrain a large language model with disease progression datasets. This will allow training of the network to classify each new sample and further make predictions regarding disease prognosis. Additional tokens will be the texture metrics of patient-derived organoids (Matov, 2024a; Matov, 2024b; Matov et al., 2016) such as shape descriptors and other morphology data (Galletti et al., 2013; Galletti et al., 2014; Matov, 2024g), which could be obtained from circulating tumor cells (Matov and Modiri, 2024; Sung et al., 2013; Sung et al., 2012; Tasaki et al., 2014). The embeddings can further include cytoskeletal dynamics (Matov, 2024c; Matov, 2024d; Matov, 2024e; Matov, 2024f; Matov, 2025a; Matov, 2025b) before and after treatment with different drugs. This will allow training of the network to classify as sensitive or resistant each organoid and further identify subclasses with different mechanisms

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

of resistance. Further, after training a generative transformer network with organoid texture and live-cell dynamics datasets, we will be able to model disease progression as well.

In the long run, the most reliable way to perform early detection will be to personalize the process by aggregating longitudinal baseline datasets for each individual. We have attempted to do that in an effort of detecting lung cancer in urine samples (Matov, 2024h). Analysis of microRNA profiles extracted from the urine of healthy donors longitudinally indicates a relative consistency in their levels. This suggests that tracking the levels of nucleic acids in body fluids longitudinally may allow for the delineation of organ- and tissue-specific patterns of changes in dedicated panels of disease-associated biomarkers and, thus, anticipate early disease and inform therapy.

MATERIALS AND METHODS

Data Processing

Whole genome sequencing DELFI raw data from (Cristiano et al., 2019) was processed to extract and bin all available DNA fragments using: https://github.com/Hogfeldt/ctDNAtool.

Data Analysis

All analysis programs for fragmentomics analysis and graphical/image representation of the results were developed in R and Python. The KL divergence method used is described and validated in (Han et al., 2016). The computer code is available for download at: https://github.com/amatov/FragmentomicsSubclinicalDisease.

ACKNOWEDGEMENTS

I thank Claus Andersen for genome data and his feedback on the shortcomings of the current whole genome sequencing analysis methods.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares no conflicts of interest.

Funding Statement

No funding was received to assist with the preparation of this manuscript.

REFERENCES

Avissar, N., D.B. Ornt, Y. Yagil, S. Horowitz, R.H. Watkins, E.A. Kerl, K. Takahashi, I.S. Palmer, and H.J. Cohen. 1994. Human kidney proximal tubules are the main source of plasma glutathione peroxidase. *The American journal of physiology*. 266:C367-375.

Barrett, C.W., W. Ning, X. Chen, J.J. Smith, M.K. Washington, K.E. Hill, L.A. Coburn, R.M. Peek, R. Chaturvedi, K.T. Wilson, R.F. Burk, and C.S. Williams. 2013. Tumor suppressor function of the plasma glutathione peroxidase gpx3 in colitis-associated carcinoma. *Cancer Res*. 73:1245-1255.

Bergmans, P. 1973. Random Coding Theorem for Broadcast Channels with Degraded Components. *IEEE Transactions on Information Theory*. 19(2):197 - 207.

Bergmans, P. 1974. A Simple Converse for Broadcast Channel with Additive White Gaussian Noise. *IEEE Transactions on Information Theory*. 20 (2):279 - 280.

Bremaud, P. 1999. Gibbs Fields and Monte Carlo Simulation. *Markov Chains*. 31:253–322.

Chandra, R., K. Jain, R.V. Deo, and S. Cripps. 2019. Langevin-gradient parallel tempering for Bayesian neural learning. *Neurocomputing*. 359:315-326.

Chen, X., R. Kang, G. Kroemer, and D. Tang. 2021. Broadening horizons: the role of ferroptosis in cancer. *Nat Rev Clin Oncol*. 18:280-296.

Cover, T. 1998. Comments on Broadcast Channels. *IEEE Transactions on Information Theory*. 44 (6):2524-2530.

Cristiano, S., A. Leal, J. Phallen, J. Fiksel, V. Adleff, D.C. Bruhm, S. Jensen, J.E. Medina, C. Hruban, J.R. White, D.N. Palsgrove, N. Niknafs, V. Anagnostou, P. Forde, J. Naidoo, K. Marrone, J. Brahmer, B.D. Woodward, H. Husain, K.L. van Rooijen, M.W. Ørntoft, A.H. Madsen, C.J.H. van de Velde, M. Verheij, A. Cats, C.J.A. Punt, G.R. Vink, N.C.T. van Grieken, M. Koopman, R.J.A. Fijneman, J.S. Johansen, H.J. Nielsen, G.A. Meijer, C.L. Andersen, R.B. Scharpf, and V.E. Velculescu. 2019. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. 570:385-389.

Dixon, S.J., and B.R. Stockwell. 2014. The role of iron and reactive oxygen species in cell death. *Nature chemical biology*. 10:9-17.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Dolma, S., S.L. Lessnick, W.C. Hahn, and B.R. Stockwell. 2003. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer Cell*. 3:285-296.

Galletti, G., K. Cleveland, A. Matov, J.E. Hayes, R.J. Klein, D.C. Hassane, L.V. Nicacio, and P. Giannakakou. 2013. Clinical and preclinical evaluation of taxane sensitivity in gastric cancer (GC): Relevance of GC histology. *Journal of Clinical Oncology*. 31:37-37.

Galletti, G., K. Cleveland, C. Zhang, A. Gjyrezi, A. Matov, D. Betel, M.A. Shah, and P. Giannakakou. 2014. Elucidating the molecular basis of intrinsic taxane resistance in gastric cancer. *Cancer Research*. 74:897.

Gamal, A., and E. Meulem. 1981. A Proof of Marton's Coding Theorem for the Discrete Memoryless Broadcast Channel. *IEEE Transactions on Information Theory*. 27 (1):120 - 122.

Han, T.S. 1981. The Capacity Region for the Deterministic Broadcast Channel with a Common Message. *IEEE Transactions on Information Theory*. 27(1):122 - 125.

Han, Y., J. Jiao, and T. Weissman. 2016. Minimax Rate-optimal Estimation of KL Divergence between Discrete Distributions. *International Symposium on Information Theory and its Applications*. 2016:256-260.

Hangauer, M.J., V.S. Viswanathan, M.J. Ryan, D. Bole, J.K. Eaton, A. Matov, J. Galeas, H.D. Dhruv, M.E. Berens, S.L. Schreiber, F. McCormick, and M.T. McManus. 2017. Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. *Nature*. 551:247-250.

Horiuchi, D., H. Tatekawa, T. Oura, S. Oue, S.L. Walston, H. Takita, S. Matsushita, Y. Mitsuyama, T. Shimono, Y. Miki, and D. Ueda. 2024. Comparing the Diagnostic Performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and Radiologists in Challenging Neuroradiology Cases. *Clinical neuroradiology*.

Ivanov, M., A. Baranova, T. Butler, P. Spellman, and V. Mileyko. 2015. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC genomics*. 16 Suppl 13:S1.

Jorgensen, M.L., J.M. Young, and M.J. Solomon. 2015. Optimal delivery of colorectal cancer follow-up care: improving patient outcomes. *Patient related outcome measures*. 6:127-138.

Karras, T., T. Aila, S. Laine, and J.J.A. Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. abs/1710.10196.

Kumar, P., K.S. Lyle, S. Gierke, A. Matov, G. Danuser, and T. Wittmann. 2009. GSK3β phosphorylation modulates CLASP–microtubule association and lamella microtubule attachment. *The Journal of Cell Biology*. 184:895-908.

Lee, B., T. Moon, S. Yoon, and T. Weissman. 2015. DUDE-Seq: Fast, flexible, and robust denoising for targeted amplicon sequencing. *PLoS ONE*. 12.

Lejeune, S., F. Guillemot, J.P. Triboulet, S. Cattan, C. Mouton, N. Porchet, S. Manouvrier, and M.P. Buisine. 2006. Low frequency of AXIN2 mutations and high frequency of MUTYH mutations in patients with multiple polyposis. *Human mutation*. 27:1064.

Lemasters, J.J. 2017. Evolution of Voltage-Dependent Anion Channel Function: From Molecular Sieve to Governator to Actuator of Ferroptosis. *Frontiers in oncology*. 7:303.

Liu, M.C., G.R. Oxnard, E.A. Klein, C. Swanton, and M.V. Seiden. 2020. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of oncology : official journal of the European Society for Medical Oncology*. 31:745-759.

Liu, W., X. Dong, M. Mai, R.S. Seelan, K. Taniguchi, K.K. Krishnadath, K.C. Halling, J.M. Cunningham, L.A. Boardman, C. Qian, E. Christensen, S.S. Schmidt, P.C. Roche, D.I. Smith, and S.N. Thibodeau. 2000. Mutations in AXIN2 cause colorectal cancer with defective mismatch repair by activating beta-catenin/TCF signalling. *Nature genetics*. 26:146-147.

Marton, A. 1979. Coding Theorem for the Discrete Memoryless Broadcast Channel. *IEEE Transactions on Information Theory*. 25 (3):306 - 311.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Matov, A. 1999. Capacity Region Characterization of Multi-User Parallel Gaussian Broadcast Channel. *Swiss Federal Institute of Technology Lausanne*:1-41.

Matov, A. 2024a. Analysis of Multiple Myeloma Drug Efficacy. *medRxiv*:2024.24311450.

Matov, A. 2024b. Colorectal Cancer DNA Fragmentation Patterns as Image Datasets. *SSRN*:2024.4909309.

Matov, A. 2024c. Computational Analysis of Treatment Resistant Cancer Cells. *medRxiv*:2024.24312813.

Matov, A. 2024d. Microtubule Regulation in Cancer Cells *SSRN*:2024.5072927.

Matov, A. 2024e. Mitosis, Cytoskeleton Regulation, and Drug Resistance in Receptor Triple Negative Breast Cancer. *arXiv*:2407.19112.

Matov, A. 2024f. Modulation of the Cytoskeleton for Cancer Therapy *SSRN*:2024.5061950.

Matov, A. 2024g. Quantitative Video Microscopy in Medicine. *SSRN*:2024.4909311.

Matov, A. 2024h. Urinary Biomarkers for Lung Cancer Detection. *medRxiv*:2024.24311186.

Matov, A. 2025a. Microtubule Dysregulation in Prostate Cancer Metastasis to the Vertebrae *SSRN*:2025.50.

Matov, A. 2025b. Quantitative Cell Division in Medicine. *SSRN*:2025.50.

Matov, A., J. de Rooij, J.C. Gatlin, J. Rohrberg, N. Ahmad, R. Aggarwal, J. Simko, A. Goga, C.J. Ryan, and T. Wittmann. 2016. Investigating microtubule growth dynamics in patient-derived metastatic prostate cancer cells. *Cancer Research*. 76:243.

Matov, A., and S. Modiri. 2024. Quantitative Microscopy in Medicine. *medRxiv*:2024.24311304.

Mullis, K.B. 1968. Cosmological Significance of Time Reversal. *Nature*. 218:663.

Muro, K., T. Salinardi, A.R. Singh, and T. Macarulla. 2020. Safety of Aflibercept in Metastatic Colorectal Cancer: A Literature Review and Expert Perspective on Clinical and Real-World Data. *Cancers*. 12.

Ordentlich, E., and T. Weissman. 2011. Bounds on the entropy rate of binary hidden Markov processes. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*.

Petrelli, F., G. Tomasello, K. Borgonovo, M. Ghidini, L. Turati, P. Dallera, R. Passalacqua, G. Sgroi, and S. Barni. 2017. Prognostic Survival Associated With Left-Sided vs Right-Sided Colon Cancer: A Systematic Review and Meta-analysis. *JAMA oncology*. 3:211-219.

Ren, Z., Y. Su, and X. Liu. 2024. ChatGPT-Powered Hierarchical Comparisons for Image Classification. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*:69706 - 69718.

Siegel, R.L., K.D. Miller, and A. Jemal. 2020. Cancer statistics, 2020. *CA: a cancer journal for clinicians*. 70:7-30.

Sung, M.S., A. Gjyrezi, G.Y. Lee, A. Matov, G. Galletti, M. Loftus, Y. Syed, T. Lannin, A. Hristov, C. Mason, S.T. Tagawa, B.J. Kirby, D.M. Nanus, and P. Giannakakou. 2013. Using CTCs to interrogate mechanisms of taxane resistance in the prospective TAXYNERGY clinical trial in prostate cancer. *Cancer Research*. 73:3492.

Sung, M.S., A. Gjyrezi, Y. Syed, T. Lannin, A. Hristov, A. Matov, M. Loftus, G. Galletti, G.Y. Lee, C. Mason, B.J. Kirby, D.M. Nanus, S.T. Tagawa, and P. Giannakakou. 2012. Molecular determinants of taxane activity using circulating tumor cells (CTCs) from castrate-resistant prostate cancer patients. *Understanding Cancer: Genomes to Devices Symposium, New York Academy of Medicine*.

Tasaki, S., M. Sung, A. Matov, G. Galletti, E. Diamond, N. Bander, K. Zhou, S. Tagawa, D. Nanus, and P. Giannakakou. 2014. Analysis of microtubule perturbations and androgen receptor localization in circulating tumor cells from castration resistant prostate cancer patients as predictive biomarkers of clinical response to docetaxel chemotherapy. *Cancer Research*. 74:923-923.

Tse, D.N. 1999. Optimal Power Allocation Over Parallel Gaussian Broadcast Channel. *Proceedings of IEEE International Symposium on Information Theory*:27.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.17.24310568; this version posted January 8, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

van der Pol, Y., and F. Mouliere. 2019. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer cell*. 36:350-368.

Van Trappen, P.O., and M.S. Pepper. 2002. Lymphatic dissemination of tumour cells and the formation of micrometastases. *The Lancet. Oncology*. 3:44-52.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *In* Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA. 6000–6010.

Wan, N., D. Weinberg, T.Y. Liu, K. Niehaus, E.A. Ariazi, D. Delubac, A. Kannan, B. White, M. Bailey, M. Bertin, N. Boley, D. Bowen, J. Cregg, A.M. Drake, R. Ennis, S. Fransen, E. Gafni, L. Hansen, Y. Liu, G.L. Otte, J. Pecson, B. Rice, G.E. Sanderson, A. Sharma, J. St John, C. Tang, A. Tzou, L. Young, G. Putcha, and I.S. Haque. 2019. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC cancer*. 19:832.

Wenzel, J., K. Rose, E.B. Haghighi, C. Lamprecht, G. Rauen, V. Freihen, R. Kesselring, M. Boerries, and A. Hecht. 2020. Loss of the nuclear Wnt pathway effector TCF7L2 promotes migration and invasion of human colorectal cancer cells. *Oncogene*. 39:3893-3909.

Yagoda, N., M. von Rechenberg, E. Zaganjor, A.J. Bauer, W.S. Yang, D.J. Fridman, A.J. Wolpaw, I. Smukste, J.M. Peltier, J.J. Boniface, R. Smith, S.L. Lessnick, S. Sahasrabudhe, and B.R. Stockwell. 2007. RAS-RAF-MEK-dependent oxidative cell death involving voltage-dependent anion channels. *Nature*. 447:864-868.

Yi, L., X. Zhong, Z. Chen, Q. Wang, Y. Yan, J. Wang, and X. Deng. 2019. MicroRNA-147b Promotes Proliferation and Invasion of Human Colorectal Cancer by Targeting RAS Oncogene Family (RAP2B). *Pathobiology : journal of immunopathology, molecular and cellular biology*. 86:173-181.

Zhang, M.L., H.T. Wu, W.J. Chen, Y. Xu, Q.Q. Ye, J.X. Shen, and J. Liu. 2020. Involvement of glutathione peroxidases in the occurrence and development of breast cancers. *Journal of translational medicine*. 18:247.

Zhong, J., R. Liu, and P. Chen. 2020. Identifying critical state of complex diseases by single-sample Kullback-Leibler divergence. *BMC genomics*. 21:87.