NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Cheloshkina Kseniia

MACHINE LEARNING BASED APPROACHES FOR ANALYSIS OF CANCER GENOME BREAKPOINTS

PhD Dissertation Summary

for the purpose of obtaining academic degree

Doctor of Philosophy in Computer Science

Academic Supervisor:
Candidate of Sciences,
Maria Poptsova

Moscow – 2023

The PhD dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor: Maria Poptsova, Candidate of Sciences, Associate Professor at National Research University Higher School of Economics**.**

DISSERTATION TOPIC

Cancer detection and treatment has been a challenge. The reason for that is a complexity of cancerogenesis and heterogeneity of cancer genome mutations. Cancer genomes typically have numerous mutations. First, the cancer genome is characterized by point mutations of a single nucleotide and small, several base pairs deletions and insertions, called indels. Another property of cancer genomes is formation of breakpoints which lead to significant genome rearrangements (insertions, deletions, tandem duplications, translocations) from several dozen to millions of nucleotides. These changes make cancer genome unstable and destroy the mechanisms of normal functioning of the cell such as division, growth and differentiation.

To get insights into cancer mutation processes, detect biomarkers and cancer gene drivers, cancer genome consortiums were created in order to organize collection of cancer genome data. Due to the efforts of The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) hundreds of thousands of cancer breakpoints have been documented for different types of cancers [1], [2]. Recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the ICGC and TCGA reported the integrative analysis of more than 2500 whole-cancer genomes across 38 tumor types [3]. These consortiums made the data publicly available to enable scientists from all over the world conduct cancer research.

In parallel with cancer genome data, omics data became available including whole-genome maps of different epigenetic features (methylation, chromatin accessibility, histone modifications, etc.) and of alternative DNA conformations (Z-DNA, quadruplexes, triplexes, stem-loops). Historically, several scientific branches were formed to study the genome from different perspectives, having the same suffix -omics at the end of the name: genomics, proteomics, metabolomics, transcriptomics. Cumulatively all these scientific branches were named as omics and aimed at getting a comprehensive view of the genome structure and function.

However, despite the large amount of cancer data available, mutagenesis of cancer breakpoints has not yet been sufficiently studied and the quality of prediction of cancer breakpoint prediction models was much lower than for cancer point mutations.

The **purpose of this research** is to study cancer breakpoints mutagenesis using machine learning methods. To achieve the goal the following tasks were set:

1. collect data and analyze state-of-the-art methods for prediction of somatic mutations and breakpoints in cancer;

2. devise rules for identification of cancer breakpoint hotspots and develop and implement a machine learning approach for cancer breakpoint hotspots prediction;

3. propose methods for identification of features predicting likelihood of cancer breakpoint formation;

4. investigate hypothesis of randomness of cancer breakpoint formation;

5. check whether PU-learning methods could improve models' quality.

KEY RESULTS

Key aspects/ideas to be defended:

1. We identified cancer breakpoint hotspots and proposed methodology for their prediction with the help of machine learning methods.

2. The methodology was tested on real data. We developed machine learning models for cancer breakpoint prediction and interpretation of omics data that outperform all previous models.

3. With the developed machine learning approach, we revealed tissue-specific impact of quadruplexes and stem-loops on cancer genome formation.

4. With the developed approach of group-wise and feature-wise importance analysis, we revealed that non-B DNA structures and transcription factors are the major determinants of cancer breakpoint formation in all cancer types.

5. With the developed approach, it was demonstrated that hotspots of higher breakpoints density are more recognizable than the low-density hotspots.

6. We tested two PU-learning (‘‘positive unlabeled”) methods and found that inclusion of hotspots labeling uncertainty into the model could not improve the results.

The personal author contribution is presented by data analysis and visualization, machine learning approach development, code implementations, writing. Maria Poptsova conceptualized the study and assigned tasks.

PUBLICATIONS AND APPROBATION OF RESEARCH

First-tier publications:

1. Cheloshkina K, Poptsova M. Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. BMC cancer. 2019 Dec;19(1):1-7.

2. Cheloshkina K, Poptsova M. Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements. PLoS computational biology. 2021 Mar 1;17(3):e1008749.

3. Cheloshkina K, Bzhikhatlov I, Poptsova M. Cancer Breakpoint Hotspots Versus Individual Breakpoints Prediction by Machine Learning Models. International Symposium on Bioinformatics Research and Applications 2020 Dec 1 (pp. 217-228). Springer, Cham.

Second-tier publications:

1. Cheloshkina K, Poptsova M. Understanding cancer breakpoint determinants with omics data. Integr Cancer Sci Therap. 2020;7(1):10-5761.

2. Cheloshkina K, Bzhikhatlov I, Poptsova M. Randomness in Cancer Breakpoint Prediction. Journal of Computational Biology. 2021 Jun 15.

For all presented papers the author is the first author and was responsible for machine learning design and implementation. Maria Poptsova conceptualized the study and assigned tasks. Both the author and Maria Poptsova participated in writing the papers.

Other publications:

1. Cheloshkina, K. (2021). Ranking Weibull Survival Model: Boosting the Concordance Index of the Weibull Time-to-Event Prediction Model with Ranking Losses. In: Kovalev, S.M., Kuznetsov, S.O., Panov, A.I. (eds) Artificial Intelligence. RCAI 2021. Lecture Notes in Computer Science(), vol 12948. Springer, Cham. https://doi.org/10.1007/978-3-030-86855-0_4

CONTENTS

In this section we will summarize the main results and formulate conclusions.

1. Existing approaches for prediction of cancer genome breakpoints

Cancer genomes are unstable and undergo numerous rearrangements resulting in origination of structural variants such as deletions, insertions, translocations, and copy number variants. Over the last 20 years several consortium cancer genome projects – The Cancer Genome Atlas (TCGA) [1], International Cancer Genome Consortium (ICGC) [2], and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Project [3] – published the information on point and structural mutations in thousands of cancer genomes. Identifying cancer mutation determinants is extremely important for understanding the genomics of the disease, but the heterogeneity of cancer genome mutations presents major difficulties in the analysis of cancer genomes.

In cancer genomics, one of the important tasks is to understand the factors and mechanisms lying behind the mutagenic processes. Below we describe recent studies which use algorithmic and machine learning approaches to predict cancer point mutations and breakpoints.

Algorithmic techniques for cancer point mutation and breakpoint prediction

Before machine learning researchers employed statistical methods to identify dependencies between different genomic factors. Then machine learning became a widely used technique for different tasks such as different target prediction (classification or regression), feature selection, feature dimensionality reduction, and other tasks.

The first comprehensive study of mutation densities was conducted at 1 Mb scale and included data on gene expression, replication timing, heterochromatin (H3K9me3 signal) and DNA mismatch repair state (measured via microsatellite instability status) [6]. Applying Mann-Whitney test it was shown that mismatch repair impacts the mutation rate variation having 72% of genome windows with significant difference in mutation frequencies. Besides, PCA analysis of 1 Mb mutation densities in samples demonstrated the difference between cancer types by performing Mann-Whitney test on principal components for different cancers.

Comprehensive analysis of non-coding point mutations together with indels specifically in 212 gastric cancer genomes was done in [7]. To define the most informative epigenetic features for modeling the somatic mutation rate, the authors used LASSO logistic regression to regress binary mutation status of each genome position on mean signal of features. Additionally, logistic regression was used to model patient specific mutation probabilities which were processed with the Poisson binomial model with the purpose to identify mutation hotspots. The authors identified 34 mutation hotspots, of which 11 were located in CTCF binding sites. Wilcoxon rank-sum test showed that distance from CBS hotspot to the nearest SCNA breakpoint is shorter in mutated than non-mutated tumors.

In [8] vicinities of breakpoints were investigated for the presence of non-B DNA structures. Considering the distance between G-quadruplex forming motifs and breakpoints in fragile regions, the association between presence of G-quadruplex and breakpoint regions was found in almost 70% of genes involved in rearrangements in lymphoid cancers.

Analysis of almost 700 000 cancer breakpoints from 26 cancer types revealed enrichment of the breakpoint regions with G-quadruplex forming sequences using Mann-Whitney test [9]. Applying the same method, the authors showed association of breakpoints hotspots with hypomethylated states.

Comprehensive statistical analysis of translocation and deletion breakpoints in cancer genomes confirmed significant association of breakpoints with non-B DNA structures for a large data set (around 20 000 of translocations and 46 000 of deletions) [10]. Specifically, with Student's t-tests for differences in the number of potential non-B DNA structures in the regions of translocation and deletion breakpoints, it was revealed that repeats were frequently found at the translocation breakpoints and poly-A sites were found more frequently at the deletion breakpoints.

Statistical analysis of enrichment of DNA protein binding and open chromatin was done for a set of 147 samples comprising 8 cancer types and 14600 structural mutations [11]. It was based on 457 ENCODE protein binding ChIP-seq experiments, 125 DNase I and 24 FAIRE experiments. The study presented enrichment of protein binding and open chromatin in the vicinity of breakpoints using a two-tailed t-test for the log odds ratios of nearby and distant regions (or their difference).

Genome-wide comparison of UV-induced DNA lesion distribution with epigenetic feature distributions was performed for skin cancer [12]. Examining the deviations of UV-induced DNA lesion distribution from genome median for different chromatin states and correlations with histone modifications, it was shown that heterochromatin is more susceptible to UV-induced DNA lesions.

Machine learning methods for cancer point mutation and breakpoint prediction

One of the first comprehensive machine learning studies of cancer point mutation was performed in [4] using data on histone modifications, CTCF binding sites, Pol-II binding sites, recombination rate, replication timing, nucleosome positioning, gene density, and conservation level. With linear regression the authors predicted point mutation densities at 1Mb scale and achieved predictive power of 55% R2. With feature importance analysis, it was revealed that one single feature – the histone modification H3K9me3 – explains 40% of cancer point mutation variation.

In another work [5] Random Forest model explained up to 86% of cancer point mutation density variance across different cancer types. The study revealed that usage of cell type-specific epigenomic features (chromatin accessibility, histone modifications and replication timing) improves the prediction quality. The modeling also showed that the reverse task can be solved – mutation density profiles can be used to detect type of cancer.

Predictive modeling of cancer point mutations appeared to be much more effective than modeling breakpoints. In [13] the authors used linear regression and Random Forest to predict both cancer point mutations and breakpoint density in 500 kB genome windows using non-B DNA structures, histone marks and replication timing, as the combined sets or separately. Depending on the type of cancer combination of non-B DNA structures and epigenetic marks could explain 43-76% of the variance while using all considered features could give only 10% of explained variance for all cancers with an exception of 18% for breast cancer.

Linear regression analysis of translocation breakpoint frequencies for blood cancer and solid cancers combined [14] using chromatin density, gene density and CTCF-binding site densities gave 18% - 39% adjusted R2 where chromatin density was identified as the most significant factor.

Association of breakpoints with gene-rich regions was also studied in [15] where the number of breakpoints was linearly regressed on the number of genes. The authors achieved 40% R2 and demonstrated that association remains highly significant for both recurrent and nonrecurrent chromosome abnormalities.

In [18] the authors proposed an ML-approach for prediction of double-strand breaks (DSB), that were generated by DSBcapture [16] and BLESS [17] methods. Using such features as densities of histone marks, DNase-seq, DNA shape parameters, CTCF and p63 binding sites at 1 kb scale Random Forest model predicted whether the site is double-strand break with 97% ROC AUC. This high prediction power can be explained by the biased method of DSB generation with restriction enzyme EcoRV.

In [19] the authors studied the dependence of gene expression and methylation of CpG islands on nearby breakpoints. Using linear regression for the purpose, they revealed that the vicinity of the breakpoint in up to ± 1 Mb region changes methylation.

2. Novel machine learning approach for cancer breakpoints prediction

Previous attempts to predict cancer breakpoints demonstrated that this task is not straight-forward. That is why we aimed at development of a machine learning pipeline for cancer breakpoint prediction which will resolve all the issues and drawbacks of the previous approaches.

Data preprocessing: choice of the aggregation level

For analysis [20] we used publicly available data from the International Cancer Genome Consortium (ICGC) Data Portal (release 25) [21]. The data comprised 10 cancer types (breast, bone, brain, blood, prostate, skin, pancreatic, liver, ovary, uterus) with 2 234 samples and 487 425 breakpoints in total where the most samples belonged to breast cancer (644 samples) while only 72 and 16 samples accordingly were related to the brain and uterus cancers.

To predict cancer breakpoints, we considered the most prevalent non-B DNA structures – stem-loops and quadruplexes – which are the known sources of chromosome instability. As our primary goal was to develop and test a machine learning approach for the task of cancer breakpoints prediction, we limit the set of features to these two potentially important factors. Human genome annotations with stem-loops (three types of length: 6-15, 15-30, 16-50) were downloaded from the DNA punctuation project [22] while annotation of the genome with G-quadruplexes was done by applying regular expression [23]. The input data (target as well as features) were presented in table format where each row corresponds to one genomic object instance with designation of start and end position of this instance. To discover the major patterns in data we needed to perform data aggregation.

The whole genome was split in non-overlapping disjoint windows of a specific length. Since there is no intuition about the optimal window length, we discovered 6 different length options (hereinafter aggregation levels): 10, 20, 50, 100, 500 kb and 1 Mb. Then the data were aggregated to these windows by mapping each instance to its window by position, calculating density for breakpoints and coverage for features. Breakpoint density in a window is the number of breakpoints located in a window divided by the total number of breakpoints in the genome. Feature coverage in a given window was calculated as the total length of all structures in the window (without overlaps) divided by the window size.

Since cancer is a heterogeneous disease, it is important to mine common patterns persisting in multiple samples. Earlier the notion of recurrent breakpoints was introduced [24] where the authors fixed the set of most frequently discovered cancer breakpoints. Here we defined breakpoint hotspots using data-driven approach: for each cancer type we found values of top 1%, 0.5%, 0.1%, 0.05% and 0.01% of breakpoints density distribution and label windows with breakpoints density equal to or higher than these thresholds as breakpoint hotspots for corresponding ‘‘labeling type”. After analysis of the number of positive examples (breakpoint hotspots) for each cancer type, aggregation level and labeling type we ended up with a total of 236 datasets for modeling while datasets with extremely small number of positive examples or with duplicate labeling were removed from the research. It should be noted that from a machine learning point of view in all mentioned labeling types the task of breakpoint hotspots classification is highly imbalanced which should be taken into account during modeling.

In addition to considered cancer types we composed a general cancer profile (‘‘overall”) using breakpoint densities in each window for each cancer type and worldwide statistics of cancer cases’ distribution by type.

Dealing with extreme class imbalance

To develop a machine learning pipeline accounting for high class imbalance we performed a search over different resampling schemes, machine learning algorithms and class balancing techniques. Since the most important quality metrics were the recall and precision, we used their harmonic mean in the form of F1-score to assess model performance. During the search we aimed at minimal overfitting, maximal performance on unseen data (test/validation) and minimal standard deviation of the scores on unseen data. We applied the next methods to our data:

- Resampling schemes: train-test split (50%), LOOCV (leave-one-out cross-validation), 15-times repeated 3-fold cross-validation. The last method was finally selected as it provides us with a performance distribution with an estimate of the worst, the best and mean quality values.
- Machine learning algorithm: logistic regression, random forest. Random Forest (with stratification) in high class imbalance case demonstrated greater overfitting than logistic regression.
- Class balancing technique: stratification, oversampling, SMOTE. Compared to stratification, oversampling improved the quality of hotspot classification while SMOTE did not show improvement.

We ended up with a logistic regression model with oversampling fitted in 15-times repeated 3-fold cross-validation with z-score normalization of features. This pipeline was applied to each of 236 datasets.

Lift of recall and lift of precision as metrics to assess model performance

To assess model performance in case of class imbalance such metrics as recall and precision are usually used. Since only 0.01 to 1% of samples are represented by positive examples, the classification task is hard. To understand whether it makes sense to use the machine learning model we introduced such derived metrics as lift of recall and lift of precision. If a random choice model selects n% of samples as predicted as '"ones", then the recall of the model will also approximate n% (random selection of n% of samples results in n% of examples of each class). Then we can estimate whether the machine learning model is better than a random choice model by dividing the ML-model recall on the probability percentile (the proportion of the number of marked as ‘‘ones” examples). If the lift of recall is greater than one the machine learning model outperforms the random model, and the higher, the stronger the model is. On the contrary, if the lift of recall is less than one or approaches it, it means that the machine learning model could not capture the relationships. Similarly, the lift of precision is defined as a ratio of model precision to the proportion of positive examples in a sample.

In addition to mean and median ROC AUC on the test set we reported the lift of recall for different thresholds according to fixed probability distribution percentiles 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%.

Choosing the best model for each cancer type

We applied the described pipeline to all 236 datasets for 3 sets of features: only stem-loop features, only quadruplex features and both stem-loop and quadruplexes. It was observed that the impact of these features is tissue-specific: there were no such feature set that demonstrated the best results for all cancer types while the direction of features impact (positive or negative) was preserved. Table 1 presents the best achieved quality metrics among different labeling types, aggregation levels and model specifications for each cancer type. It could be seen that the best performance is observed for breast and bone cancer with ROC AUC of 0.94 and 0.86 and lift of recall of 8 and 10 respectively. At the same time there are cancer types with relatively low-quality metrics: ROC AUC 0.63 and 0.61 and lift of recall 2.05 and 1.71 for prostate and pancreatic cancer.

Cancer type	**Number of samples**	**Number of breakpoints**	**Number of datasets**	**Lift of recall (best model)**	**Median test ROC AUC (best model)**
Brain	72	1 564	20	5,00	67%
Blood	118	2 330	20	4,00	67%
Bone	117	2 546	20	10,00	86%
Uterus	16	6 782	19	4,00	65%

Liver	255	22 324	21	5,71	73%
Prostate	212	48 126	23	2,05	55%
Skin	190	54 688	23	4,00	64%
Ovary	115	71 446	22	6,67	68%
Pancreatic	495	85 769	22	1,71	57%
Breast	644	191 850	23	10,00	94%
Overall			23	6,67	72%

Table 1 Dataset stats: cancer type, number of samples, number of breakpoints, number of datasets for prediction, quality metrics on best models (all aggregation levels and labeling types).

To sum up, the developed machine learning pipeline enabled us to discover tissue-specificity of considered features. Tested on a small set of features the pipeline could be applied to a wider feature set to enrich and expand the research.

3. Cancer breakpoint prediction from omics data

Using omics data for cancer breakpoints prediction

With the development of sequencing technologies omics data became a valuable source of information for machine learning models. Omics is aimed at comprehensive quantification of characteristics that describe DNA from different perspectives (structure, function, etc.). Machine learning approaches that can aggregate multiple factors could really help in understanding cancer breakpoint determinants. Currently, only two groups of factors – histone modifications and non-B DNA structures were tested as predictors on large data sets. Adding other groups from omics experiments into machine learning approach likely will help in finding or stratifying more determinants of cancer breakpoint formation [25].

As there was no previous research, which considered the majority of the available at the time features to predict cancer breakpoints, we performed such a comprehensive study using the developed machine learning approach to get a higher model quality compared to the previous results. The research [26] included such features as non-B DNA structures (quadruplexes, Z-DNA, stem-loops, repeats), histone modifications (HMs), DNA methylation, transcription factor (TF) binding sites, chromatin accessibility (HDNase), chromatin partitioning with topologically associated domains (TADs), and genomic positions (whole genes, exons, introns, 5'- and 3'-UTRs promoters, downstream areas).

Feature transformations

The research was limited to one aggregation level - 100 kB - that is the most prevalent in the literature and one of the best performing according to our results. In addition to feature coverage used earlier - hereinafter ''local features" – we checked whether different feature transformations could improve model quality: binary flags of feature presence, indicators of local (1-10 neighbors) and global coverage maximum in the window and distant features (coverage of 1Mb window). Addition of presence flags gave no uplift while training the model using only these features demonstrated significant quality drop (~ -0.13 ROC AUC in mean) which could be only partially compensated by addition of maximum indicators (~ -0.03 ROC AUC in mean) which indicates that models benefit from considering exact coverage values. However, addition of distant features to local demonstrated mean 0.03 ROC AUC uplift in general although the effect slightly differs across cancer types. This quality gain could be explained by the fact that combination of distant and local features enables the model to estimate ''anomaly score" of each window regarding its nearest environment. Based on the results, these features were added to final feature set for building models. Moreover, as shown below in Chapter 4, distant features entered the top important features for the models.

Results

In the research [26] we slightly changed the machine learning pipeline. First, we replaced 15-times repeated 3-fold cross-validation resampling scheme with 30-times repeated train-test split with 30% of data in the test sample. Secondly, we replaced logistic regression model with random forest model as the feature set increased and became more diverse. Using this machine learning pipeline and omics data presented as local and distant features we built models for cancer breakpoints hotspots (99% / 99.5% / 99.9%) on 100 kb aggregation level.

Final models' quality for each cancer type is presented in Table 2. The obtained results could be compared to two papers for cancer breakpoints prediction: our previous work [20] and the study of Georgakopoulos-Soares et al. [13].

Compared to our previous work [20] – as shown in Table 2 – models on omics data demonstrated higher quality for all except bone cancer considering both median test ROC AUC and lift of recall. Previously median test ROC AUC exceeded 70% only for bone cancer while omics-based models achieved 69-86% median test ROC AUC for all cancer types except skin and bone cancer. In addition, median lift of recall also significantly increased: median uplift of this metric over all cancer types is 2,6 in absolute or +77,5%, although PR AUC remained low ranging from 0,3% to 4,8%.

Cancer type	**Omics model Lift of recall (best model, 0.03 probability percentile)**	**Omics model Median test ROC AUC (best model)**	**Omics model Mean test PR AUC (best model)**	**Non-B DNA model, 100 kb Lift of recall (best model)**	**Non-B DNA model, 100 kb Median test ROC AUC (best model)**
Blood	5,7	75%	0,3%	2,5	65%
Bone	5,1	64%	1,8%	6,0	80%
Brain	8,0	75%	0,5%	5,0	67%
Breast	7,6	86%	0,6%	6,7	65%
Liver	7,8	73%	0,6%	4,0	66%
Ovary	5,0	69%	2,9%	2,1	59%
Pancreatic	16,7	76%	4,8%	1,7	57%
Prostate	4,3	73%	0,4%	2,0	56%
Skin	2,6	57%	1,5%	2,2	56%
Uterus	4,0	69%	1,3%	4,0	62%

Table 2 Quality of models predicting cancer breakpoints hotspots on omics data and comparison with previous results [20] for models with 100kb aggregation level.

Another comparable study was the study of Georgakopoulos-Soare et al. [13] where the authors considered the regression task of cancer breakpoints density prediction by non-B DNA structures and histone modifications. Using the data provided by the authors we reproduced their research and then added novel features (omics data, aggregated as local and distant feature coverage) to predict density of the same 500 kb genome windows, keeping feature and label transformations the same as in original work [13]. Original random forest model achieved a maximum 18% R-squared while addition of novel data to the training set and addition of other omics data as predictors improves this metric up to 34%.

4. Approach for analysis of omics data feature importance

As in the research [26] we used many diverse (by genomic function) features, the key task is to perform feature importance analysis to determine key factors affecting cancer breakpoint hotspot formation. For this purpose we first assessed group feature importance and then conducted research in order to find the most influential individual features.

Group feature importance

Features, used in the research [26], could be grouped into several groups by their origin: non-B DNA structures (non-B), histone modifications (HMs), DNA methylation, transcription factor (TF) binding sites, chromatin accessibility (HDNase), topologically associated domains (TADs), and genomic positions. To find the most important feature group we used the developed pipeline to get machine learning models trained on each feature group separately for each cancer type. To rank feature groups we calculated the maximum value of the mean lift of recall at 0.03 probability percentile over all models for each labeling type and cancer type and then scaled the mean lift of recall value of each feature group model by that value. As for 99% and 99.5% labeling types the results are slightly different, we average this scale coefficient over these labeling types for each type of cancer. The results (Fig. 1) show that the best feature group significantly (by 0.25) outperforms others for almost all cancer types, and this value even reaches 0.5 for 3 cancers (blood, pancreas and prostate). This outperforming group was TFs for 5 cancers (liver, skin, prostate, ovary, breast), non-B DNA structures for 2 cancers (brain and bone) and other feature groups for one cancer type. Analysis of top-3 feature groups by the ranking revealed that non-B DNA features appeared in it for all cancer types and TFs – for 8 cancer types. Changing the evaluation metric from the lift of recall to ROC AUC led to similar results. Additionally, from trained models we extracted importance of features for the best performing feature groups – non-B DNA and TFs. It was found that among non-B DNA features G-quadruplexes and direct repeats demonstrated the highest importance while short tandem repeats and Z-DNA also do make their contribution. Concerning transcription factors, the most important features are less similar to each other for different cancer types but nevertheless CTCF, GABPA, RXRA, SP1, MAX and NR2F2 are more frequently included in top features than other transcription factors.

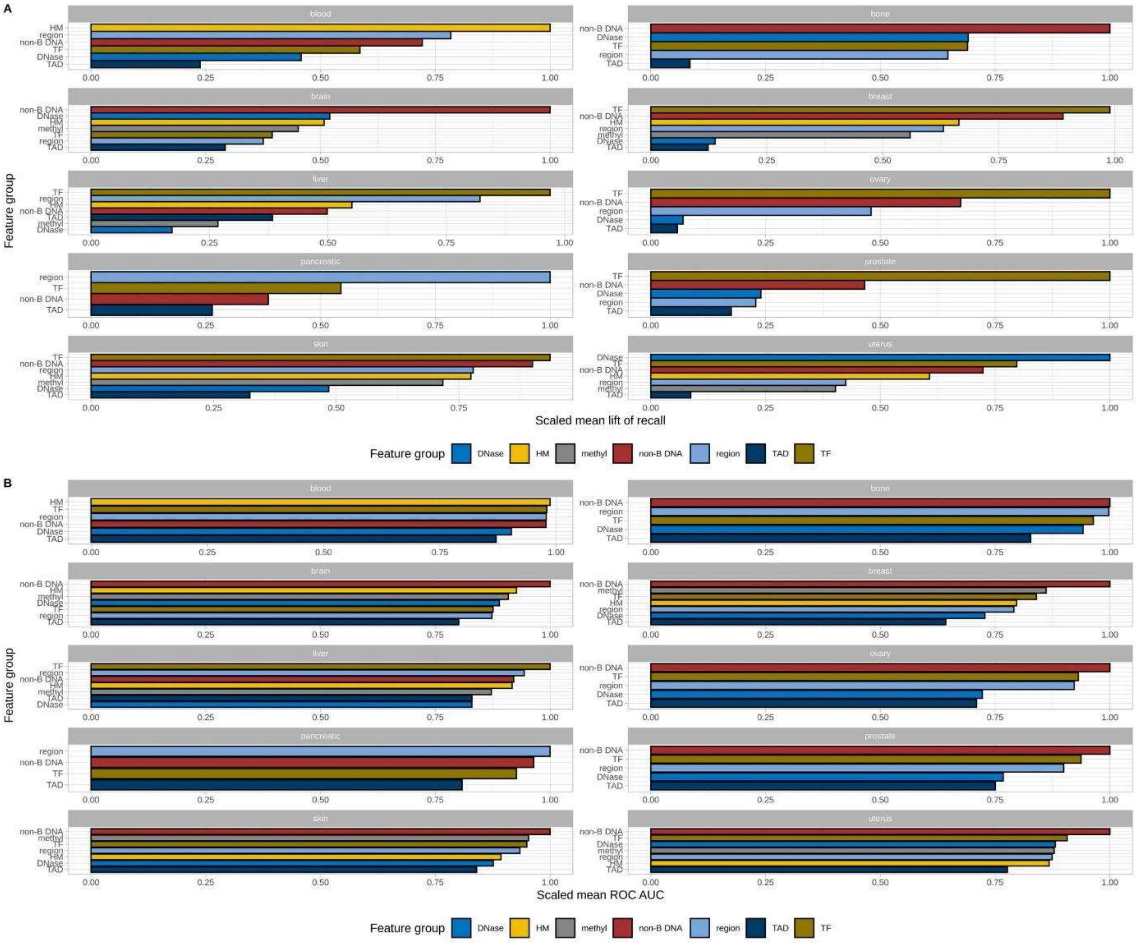

Figure 1. Feature group ranging by model performance using two different metrics – lift of recall and ROC AUC. **A.** Mean lift of recall for 0.03 probability percentile for each cancer type and feature group scaled and averaged for 99% and 99.5% labeling types. **B.** Mean ROC AUC for each cancer type and feature group scaled and averaged for 99% and 99.5% labeling types.

Individual feature importance

To select the most influential features for prediction of breakpoints hotspots for each cancer type we performed the Boruta feature selection method. Finally, we found 50 important features with 5-23 features for one cancer type (Fig. 2). The list included features mostly from non-B DNA, TFs and genomic region groups which correlates with the results of group feature importance analysis.

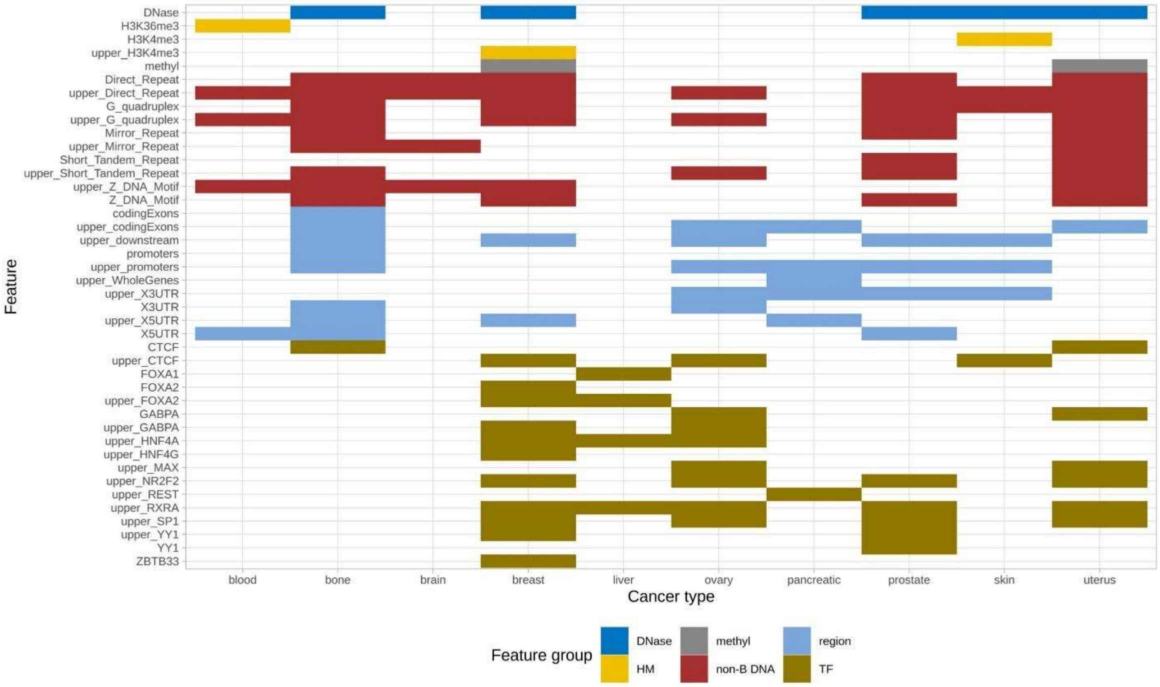

Figure 2. Boruta selected feature sets by cancer type and feature group.

In our previous research [27] it was shown that there is no single breakpoint density threshold (labeling type) for hotspots identification that demonstrates the best results for all cancer types. Moreover, it was demonstrated that the higher the breakpoints density threshold the higher the variance of machine learning models predicting corresponding hotspots. Hence, we performed the Boruta feature selection procedure on 99% labeling type as it provides the most stable results.

For each of the 30 train-test dataset splits of each cancer type the next Boruta feature selection algorithm was applied. The method is iterative, and each iteration considers only the set of ‘‘important” features defined at the previous step while in the first iteration all features are taken into account. On each iteration the set of shadow features is added to the set of important features. Shadow features present the shuffled features' values so that each real important feature has one of its shuffled versions in the dataset. Then random forest model is trained on the extended dataset and feature importance measure (mean decrease accuracy) and its z-score are calculated. The new set of important features is composed of those real features which have z-score higher than maximum z-score of all shadow features. The algorithm moves to the next iteration if there are more than 5 features in the important feature set and less than 10 iterations passed.

This algorithm selected top features for each cancer type based on 99% labeling type. Comparing the quality of models on these features with models on all available features revealed three cancer types (pancreatic, prostate and breast cancer) with slightly lower quality. For these cancer types we performed forward feature selection and found one, one and two features respectively, which addition to feature sets led to comparable quality. The selected top features for cancer type are presented on Fig. 2.

Analysis of top features for all cancer types combined revealed that features of only four feature groups (non-B DNA, TFs, genomic regions and HDNase) are included in the top list that consists of features which were selected as important in at least 300 of 3 000 times during Boruta feature selection procedure. The top five features include direct repeats and G-quadruplexes, both local and distant, and transcription factor SP1, followed by Z-DNA, short tandem repeats, mirror repeats, transcription factors RXRA, NR2F2, GABPA, CTCF, genomic regions such as 5' UTR, coding exons, 3' UTR, promoters and downstream areas, and HDNase, which impact differs for cancer types. It is worth to note that the majority of important features are presented on a distant scale.

Results

We proposed an approach for feature importance analysis of omics data. With this approach we revealed that non-B DNA structures and transcription factors are the most influential factors for cancer breakpoint hotspot prediction which is approved by both group and individual feature importance methods. Among the most important individual features the highest contribution came from G-quadruplexes and repeats and CTCF, GABPA, RXRA, SP1, MAX and NR2F2 transcription factors.

5. Approach for breakpoints randomness analysis in cancer genomes

Despite known cancer genome heterogeneity, in order to find regularities in cancer breakpoints formation, in our previous studies we explored recurrent/repeated breakpoints – hotspots. Nevertheless, it is important to understand how well individual breakpoints could be predicted and whether there is a dependency between breakpoint density thresholds and quality of corresponding hotspots detection by machine learning model. To answer this question, we conduct a series of experiments reported in the research [26].

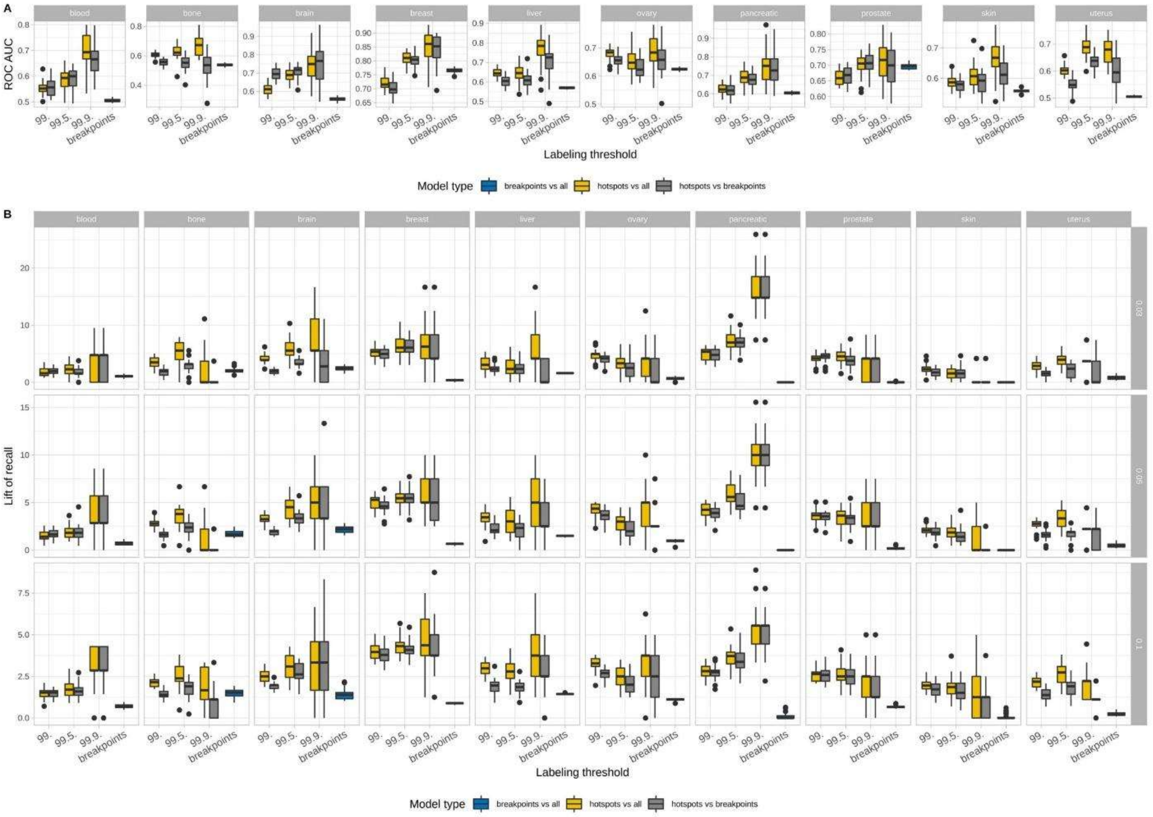

Figure 3. Comparison of models predicting hotspots and individual breakpoints. The lift of recall is presented for 0.03, 0.05 and 0.1 probability percentile.

First, we applied the pipeline to the task of individual cancer breakpoints prediction (Fig.3). It was found that the majority of individual breakpoints are almost indistinguishable from other genomic windows with ROC AUC having values around 50% or slightly higher and the mean lift of recall – in the range of 0 to 2.5. Nevertheless, it is worth to note that for several cancer types (breast, ovary, prostate) ROC AUC could achieve 65–75% but the lift of recall remains very low.

Secondly, we built machine learning models to distinguish breakpoint hotspot regions from other genomic regions with breakpoints. It was revealed that models that can separate hotspots from breakpoints can be as good as models predicting hotspots, since ROC AUC reached more than 70% for brain, liver, pancreas, and prostate cancer and 85% for breast cancer. This fact confirms that hotspots' locations are completely different from individual breakpoints locations when one takes into account only considered features. The results are presented in Fig.3.

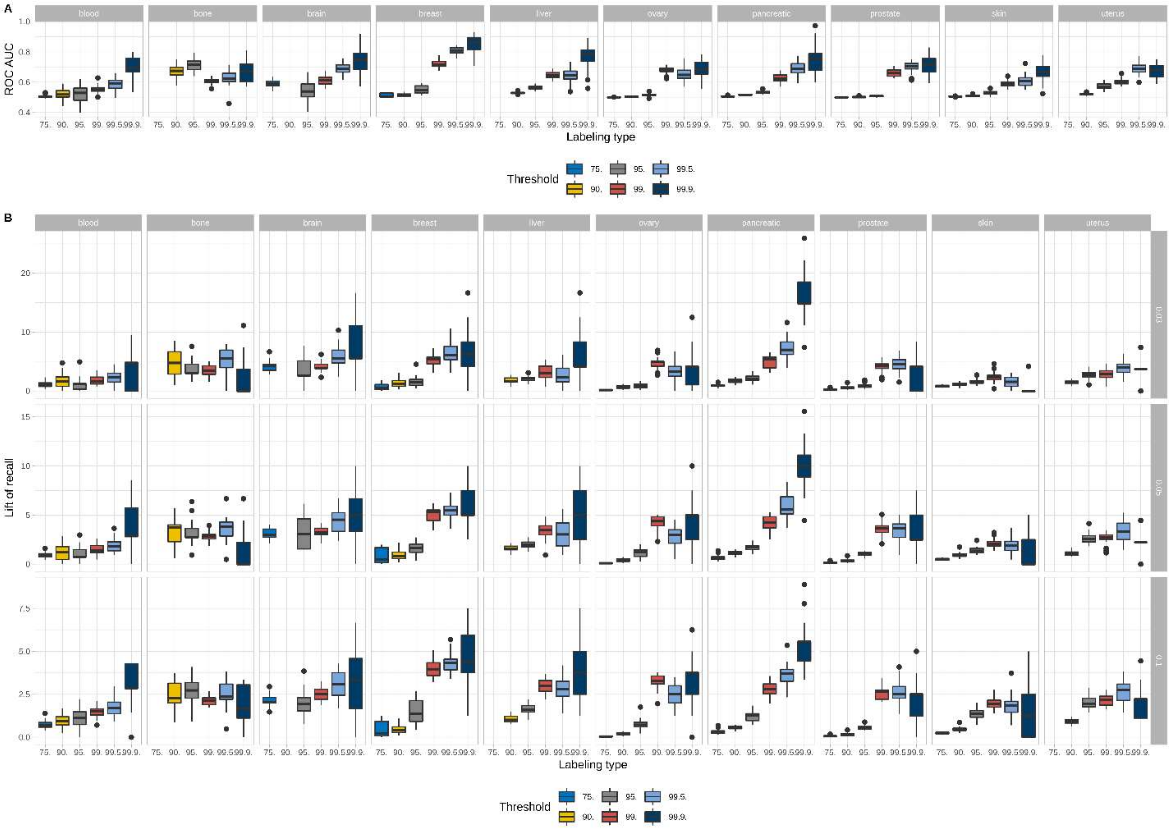

Figure 4. Comparison of different hotspots labeling criteria. The lift of recall is presented for 0.03, 0.05 and 0.1 probability percentile.

Finally, we tested other breakpoint density thresholds for hotspot labeling - 75%, 90%, 95% and built machine learning models to identify genome regions less saturated with breakpoints from other genomic regions (Fig.4). It was shown that the increase in breakpoint hotspot labeling threshold leads to a higher quality of the model recognizing corresponding hotspots. On average, decreasing the labeling threshold by 5% results in 2 times lower mean lift of recall and 15% lower mean ROC AUC. In addition, for low breakpoint density threshold absolute values of ROC AUC reached 60% only for bone cancer while its average value over all cancer types is equal to 54% with the mean lift of recall of 1.6. With this in mind, we could make a conclusion that hotspots of higher breakpoints density differ from other genomic regions to a greater extent than those of lower density.

6. Inclusion of data uncertainty into the model

In case when a target variable heavily relies on the distribution of a particular variable, the data sufficiency problem could arise. If cancer breakpoint data are not representative, then the labeling of breakpoint hotspots may be incorrect (due to the lack of data, breakpoints density may be underestimated, and therefore some regions of the genome will not be labeled as hotspots). In this case, we can incorporate this uncertainty into the model using the PU-learning approach [28, 29].

The PU-learning approach assumes that there are examples marked as positive in the sample, while labels of all other examples are unknown (they can be positive or negative). The task is to assign a label to all unlabeled examples, taking into account known positive labels and all feature distributions. In our published study, we used the PU-learning algorithm [28, 29]. The general idea of the algorithm is as follows. The classification model is initially trained using all the unlabeled examples as negative ones. Next, labels of the unlabeled examples are iteratively updated to convergence. At each iteration, the current model generates predictions for all examples, and based on 10 and 90 percentiles of the probability distribution of positive class examples and on the width of the certainty interval, defines the upper and lower bound. If the probability of an unlabeled example is higher than the upper bound, then the example refers to as a reliable positive example (RP), and if it is below than the lower bound - as a reliable negative example (RN). At each iteration only reliably labeled examples and initial known positive examples are used to train the model, and the process is repeated until convergence. The algorithm has an epsilon hyperparameter that determines the width of the certainty interval. This algorithm was implemented in two modes: RP (it iteratively updates both the set of positive and negative examples) and RN (iteratively updates only the set of negative examples, the set of positive examples is fixed as a set of initial known positive examples).

This approach was applied to all datasets for each type of cancer in the research [30]. Fig. 5 shows for each type of cancer a confidence interval for the average difference in the lift of recall of final models trained in RP and RN modes. As the quality of both models is estimated on the same initial test set, positive difference in lift of recall on a test set means that shifted features' distribution for hotspots (during PU-learning) describes the test sample better than the initial hotspot features' distribution, and additional positive examples during training give a good signal for hotspot detection.

It could be noted that the sign of the difference depends on the number of breakpoints available for the type of cancer. On the one hand, a stable positive effect is observed for the types of cancer that are in the top 5 types of cancer with the smallest number of breakpoints. In particular, the best results are achieved for brain cancer, which has a minimal number of breakpoints. Based on these data, it can be concluded that the inclusion of additional positive examples in the case of noisy data (the markup of the target variable may be noisy due to unrepresentative breakpoints data) helps to improve the quality of the model using PU Learning. On the other hand, a stable negative effect is observed for cancer types that are in the top four types of cancer with the maximum number of breakpoints. This can be explained by the fact that if there is enough data, additional positive examples introduce noise.

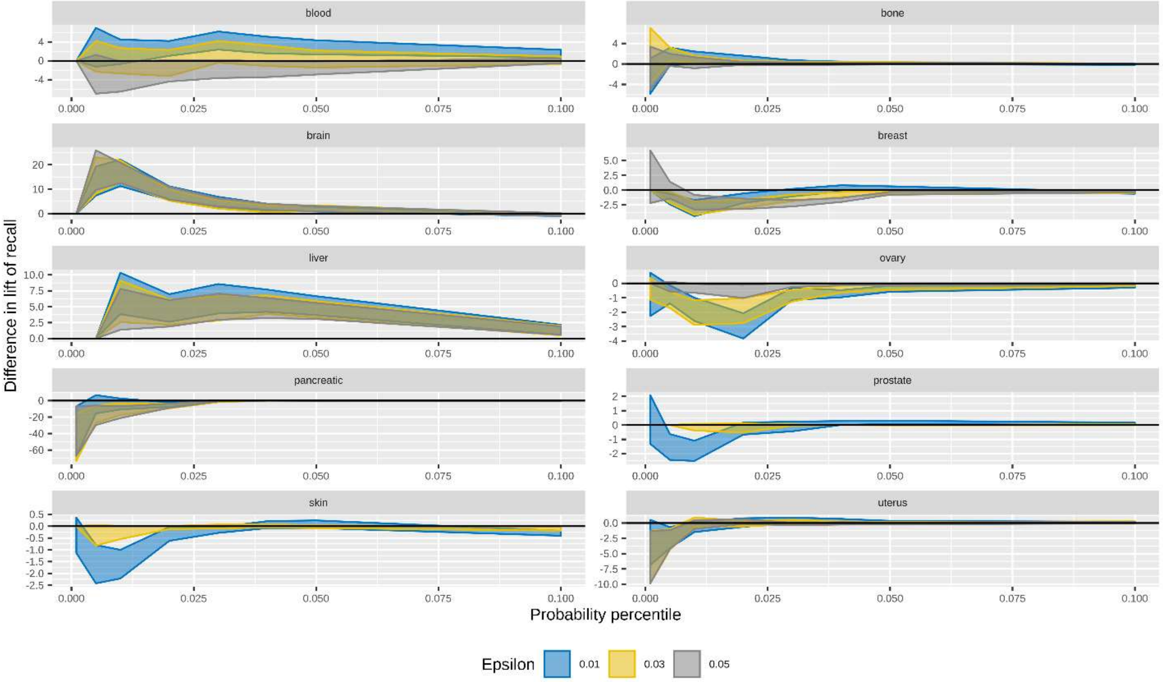

Figure 5. Confidence interval for the mean difference in the lift of recall for RP and RN mode for PU learning algorithm for different probability percentiles.

Nevertheless, PU-learning methods did not improve prediction quality compared to classic binary classification models for the considered probability percentiles (0.03, 0.05 and 0.1): lift of recall is almost identical for all cancer types.

CONCLUSION

In this dissertation we developed a machine learning approach to investigate cancer breakpoint hotspot mutagenesis. The dissertation work made contribution to the field of cancer mutagenesis revealing the key factors associated with cancer breakpoint regions and exploring the problem of randomness in cancer formation.

Collected breakpoint data on 10 common cancer types together with a large set of omics data enabled us to perform comprehensive analysis of cancer breakpoint hotspots. We designed, developed and implemented the machine learning pipeline that was applied to the task of cancer breakpoint prediction. We performed feature importance analysis and revealed two feature groups – non-B DNA structures and TF binding sites, that are important for all cancer types. We designed and implemented a set of experiments aimed at investigation of the degree of randomness of cancer breakpoint hotspot formation. Finally, we compared sample sizes of breakpoint data stratified by cancer types with the results obtained with PU learning methods applied to the data in different setups.

The developed machine learning models trained on omics data demonstrated, to our knowledge, the best performance for the cancer breakpoint hotspot prediction.

REFERENCES

1. Nakagawa, H., et al. (2015). "Cancer whole-genome sequencing: present and future." Oncogene 34(49): 5943-5950.
2. Nakagawa, H. and M. Fujita (2018). "Whole genome sequencing analysis for cancer genomics and precision medicine." Cancer Sci 109(3): 513-522.
3. Consortium, I. T. P.-C. A. o. W. G. (2020). "Pan-cancer analysis of whole genomes." Nature 578(7793): 82-93.
4. Schuster-Böckler B, Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488: 504-507.
5. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, et al. (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518: 360- 364.
6. Supek F, Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature 521: 81-84.
7. Guo YA, Chang MM, Huang W, Ooi WF, Xing M, et al. (2018) Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun 9: 1520.
8. Katapadi VK, Nambiar M, Raghavan SC (2012) Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. Genomics 100: 72-80.
9. De S, Michor F (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat Struct Mol Biol 18: 950-955.

10. Bacolla A, Tainer JA, Vasquez KM, Cooper DN (2016) Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res 44: p. 5673-88.

11. Grzeda KR, Royer-Bertrand B, Inaki K, Kim H, Hillmer AM, et al. (2014) Functional chromatin features are associated with structural mutations in cancer. BMC Genomics 15: 1013.

12. García-Nieto PE, Schwartz EK, King DA, Paulsen J, Collas P, et al. (2017) Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. EMBO J 36: 2829-2843.

13. Georgakopoulos-Soares I, Morganella S, Jain N, Hemberg M, Nik-Zainal S (2018) Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res 28: 1264-1271.

14. Lin CY, Shukla A, Grady JP, Fink JL, Dray E, et al. (2018) Translocation Breakpoints Preferentially Occur in Euchromatin and Acrocentric Chromosomes. Cancers (Basel) 10. [Crossref]

15. Mitelman F, Johansson B, Mertens F, Schyman T, Mandahl N (2019) Cancer chromosome breakpoints cluster in gene-rich genomic regions. Genes Chromosomes Cancer 58: 149-154. [Crossref]

16. Lensing SV, Marsico G, Hänsel-Hertsch R, Lam EY, Tannahill D, et al. (2016) DSBCapture: in situ capture and sequencing of DNA breaks. Nat Methods 13: 855- 857.

17. Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, et al. (2013) Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods 10: 361-365. [Crossref]

18. Mourad R, Ginalski K, Legube G, Cuvier O (2018) Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. Genome Biol 19: 34. [Crossref]

19. Zhang Y, Yang L, Kucherlapati M, Hadjipanayis A, Pantazi A, et al. (2019) Global impact of somatic structural variation on the DNA methylome of human cancers. Genome Biol 20: 209. [Crossref]

20. Cheloshkina, K., Poptsova, M. (2019). Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. BMC cancer, 19(1), 1-17.

21. The Cancer Genome Atlas (TCGA). Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

22. DNA Punctuation Project. Available from: http://www.dnapuncutation.org

23. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33(9):2908–16

24. Mitelman F, Johansson B, Mertens F. Mitelman database of chromosome aberrations and gene fusions in cancer. 2019. http://cgap.nci.nih.gov/Chromosomes/Mitelman

25. Cheloshkina K, Poptsova M (2020) Understanding cancer breakpoint determinants with omics data. Integr Cancer Sci Therap 7: DOI: 10.15761/ICST.1000333

26. Cheloshkina, K., & Poptsova, M. (2021). Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements. PLOS Computational Biology, 17(3), e1008749.

27. Cheloshkina, K., Bzhikhatlov, I., & Poptsova, M. (2020, December). Cancer Breakpoint Hotspots Versus Individual Breakpoints Prediction by Machine Learning Models. In International Symposium on Bioinformatics Research and Applications (pp. 217-228). Springer, Cham.

28. Liu, B., Lee, W.S., Yu, P.S., et al. 2002. Partially supervised classification of text documents. In ICML. 387–394

29. Liu, B., Dai, Y., Li, X., et al. 2003. Building text classifiers using positive and unlabeled examples. In Third IEEE International Conference on Data Mining. IEEE179–IEEE186

30. Cheloshkina, K., Bzhikhatlov, I., & Poptsova, M. (2021) Randomness in cancer breakpoint formation. Journal of Computational Biology.