Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Marin Volarić

Odabir najinformativnijih genomskih regija za određivanje ishodišnih stanica melanoma metodama strojnog učenja

Diplomski rad

Zagreb 2020.

University of Zagreb

Faculty of Science

Department of Biology

Marin Volarić

Selection of the most informative genomic regions for the determination of melanoma cell-of-origin using machine learning methods

Graduation Thesis

Zagreb 2020.

Ovaj rad,

izrađen u Zavodu za molekularnu biologiju,

pod vodstvom doc. dr. sc. Rose Karlić

predan je na ocjenu

Biološkom odsjeku

Prirodoslovno-matematičkog fakulteta

Sveučilišta u Zagrebu

radi stjecanja zvanja

***magistar molekularne biologije**.*

This work,
completed at the Department of Molecular Biology,
under the guidance of Associate Professor Rosa Karlić, PhD,
was submitted for assessment to
Department of Biology,
Faculty of Science
at University of Zagreb
in order to acquire
Master's degree in molecular biology.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odjsek **Diplomski rad**

ODABIR NAJINFORMATIVNIJIH GENOMSKIH REGIJA ZA ODREĐIVANJE ISHODIŠNIH STANICA MELANOMA METODAMA STROJNOG UČENJA

Marin Volarić

Rooseveltov trg 6, 10000 Zagreb, Croatia

Iako najmanje uobičajeni oblik karcinoma kože melanom je njegov najsmrtonosniji oblik karakteriziran visokom invazivnošću i velikim metastatskim potencijalom. Metastatski potencijal melanoma pokazao se problemom u djelu bolesnika koji nemaju očito mjesto porijekla primarnog tumora, pokazano najjačeg prediktora ponašanja tumora. Novi pristupi temeljeni na metodama strojnog učenja bili su uspješni u identificiranju stanice porijekla u različitih vrsta karcinoma isključivo korištenjem karakterističnih profila putničkih mutacija i somatskih epigenomskih profila. Cilj ovog istraživanja bio je istražiti može li se stanica podrijetla melanoma odrediti s djelom genomskih regija koji su korišteni u prethodnim studijama. Istražili smo može li upotreba analize glavnih komponenti smanjiti broj regija koje modeli strojnog učenja trebaju koristiti da bi se uspješno predvidjelo stanično podrijetlo melanoma. Ovdje pokazujemo da je čak 10% veličine profila korištene u prethodnim istraživanjima dovoljno za predviđanje melanomne stanice podrijetla s velikom točnošću. Nadalje, otkrili smo da najinformativnije regije imaju veći proporcionalni udio prepisane sekvence i imaju relativno malo otkrivenih poznatih mutacija koje su povezane s razvojem melanoma. Ti nalazi otkrivaju potencijalni put do novih istraživanja dijagnostičkog potencijala metoda strojnog učenja ne samo za melanom, već i za druge vrste raka.

(33 stranice, 12 slika, 3 tablice, 37 literaturnih navoda, jezik izvornika: engleski)

Rad je pohranjen u Središnjoj biološkoj knjižnici

Ključne riječi: melanom, strojno učenje, epigenetika, bioinformatika, PCAWG

Voditelj: doc.dr.sc. Rosa Karlić

Ocjenitelji:
1. Dr. sc. Rosa Karlić, doc.
2. Dr. sc. Sven Jelaska, prof.
3. Dr. sc. Tomislav Ivanković, doc.

Zamjena: Dr.sc. Kristian Vlahoviček, prof.

Rad prihvaćen:

BASIC DOCUMENTATION CARD

University of Zagreb

Faculty of Science

Division of Biology

Graduation Thesis

SELECTION OF THE MOST INFORMATIVE GENOMIC REGIONS FOR THE DETERMINATION OF MELANOMA CELL-OF-ORIGIN USING MACHINE LEARNING METHODS

Marin Volarić

Rooseveltov trg 6, 10000 Zagreb, Croatia

Although being the least common form of skin cancer melanoma is by far its deadliest form characterized by high invasiveness and a large metastatic potential. The metastatic potential of melanoma proves to be a problem in a fraction of patients who have no obvious site of primary tumor which is the strongest predictor of tumor behavior. Novel approaches based on machine learning methods have been successful in identifying different cancer types cells-of-origin solely by using cancer passenger mutation and somatic cell epigenomic regional profiles. The aim of this research was to investigate whether melanoma cell-of-origin can be determined with a fraction of genomic regions used in previous studies. We investigated whether the use of principal component analysis can reduce the number of regions that machine learning models need to use to successfully predict melanoma cell-of-origin. Here we show that with even 10% of the profile size used in previous research is enough to predict melanoma cell-of-origin with high accuracy. Moreover, we also found that the best regions have a larger proportional fraction of transcribed sequence and have relatively few of discovered known mutations which are connected with melanoma development. Those findings reveal a potential path to new research into the diagnostic potential of machine learning methods not only for melanoma but for other cancer types.

(33 pages, 12 figures, 3 tables, 37 references, original in: English)

Thesis deposited in the Central biological library.

Key words: melanoma, machine learning, epigenetics, bioinformatics, PCAWG

Supervisor: Dr.sc. Rosa Karlić, Asst. Prof.

Reviewers:

1. Dr.sc Rosa Karlić, Asst. Prof.
2. Dr. sc. Sven Jelaska, Prof.
3. Dr. sc. Tomislav Ivanković, Asst. Prof.

Replacement: Dr.sc. Kristian Vlahoviček, Prof.

Thesis accepted:

Contents

1. Introduction 1
  1.1. Melanoma 1
  1.2. Cancer genomics 2
    1.2.1. Next-generation sequencing 2
    1.2.2. ChIP-seq 3
    1.2.3. Current research in cancer genomics 5
  1.3. Research Goals 7
2. Methods 9
  2.1. Data 9
  2.2. Computational methods 10
    2.2.1. R statistical package 10
    2.2.2. Principal component analysis 10
    2.2.3. Random Forest regression 11
  2.3. Statistical tests 13
    2.3.1. Pearson Product-Moment Correlation 13
    2.3.2. Wilcoxon ranked sum test 13
    2.3.3. Fisher exact sum test 13
3. Results 15
    1.1. Principal component analysis results 15
    1.2. Identifying the optimal regions for model training 18
    1.3. Exploring the predictive quality of the selected regions 22
    1.4. Exploring the genomic background of selected regions 24
4. Discussion 29
5. Conclusion 32
6. References 33

List of abbreviations:

PCA – principal component analysis

PC – principal components

ChIP – chromatin immunoprecipitation

ChIP-seq – chromatin immunoprecipitation followed by sequencing

GC content - guanine-cytosine content

Mb – mega base

MSE – mean squared error

NGS – next generation sequencing

PCAWG - Pan-Cancer Analysis of Whole Genomes

SEdb – Human super enhancer database

UCSC - University of California Santa Cruz

1. Introduction

1.1. Melanoma

Skin cancer is the most common form of human malignancy with a global incidence rising at an alarming rate with an estimated two to three million new cases being reported each year. The 3 most common forms of skin cancer are basal cell carcinoma, squamous cell carcinoma and melanoma. Melanoma is the least common of the three, nonetheless it is credited with the largest number of deaths of all skin cancers, so much in fact that most cancer statistics limit themselves to melanoma deaths being the representative number for all skin cancers [1]. Currently early detection and resection is the best method for curing melanoma, with a success rate of 80%. However, melanoma is a cancer type with a large metastatic potential and spreads very fast if not dealt with in appropriate time. In the case of metastatic melanoma prognosis becomes very poor very fast, with a large refractory rate, a median survival rate of 6 months after diagnoses of metastases in other tissues and a 5-year survival rate of less than 3% [2]. Melanoma arises from occurrence of genetic mutations in melanocytes, specialized pigmented cells that are found predominantly in the skin and eyes, where they produce melanins, the pigments responsible for skin and hair color, which serve to protect our skin from the harmful effects of UV radiation. Melanocytes originate from highly motile neural-crest progenitors that migrate to the skin during embryonic development. In the skin, melanocytes reside in the basal layer, and their homeostasis is regulated by epidermal keratinocytes which communicate intercellularly using a complex network of biochemical pathways. The main cause of mutations in melanocytes is UV radiation [3], the very thing melanocytes have evolved to protect us against. When those mutations happen in critical growth regulatory genes melanocytes lose their ability to control the production of autocrine growth factors, expression of adhesion receptors, as well the control over the internal mechanisms all cells use to stop uncontrolled proliferation the complex intercellular biochemical network is broken and melanocytes are no longer regulated in any meaningful way by keratinocytes. Once the regulatory network is broken the transformation from melanocytes to melanoma begins [4].

Traditionally melanoma detection and diagnosis has been based on pathology, which can be a problem in patients with the recurrent metastatic disease as well as patients with a susceptibility to highly invasive methods (e.g. tissue biopsy). Advent of new and ever cheaper and better sequencing technologies promises new prognostic and diagnostic opportunities.

1.2. Cancer genomics

1.2.1. Next-generation sequencing

The term next-generation-sequencing (NGS) technologies represents a number of different platforms using different sequencing technologies all connected with the same basic principle: performing sequencing of millions of small fragments of DNA in parallel then use bioinformatics analyses to piece together these fragments by mapping the individual reads to the human reference genome.

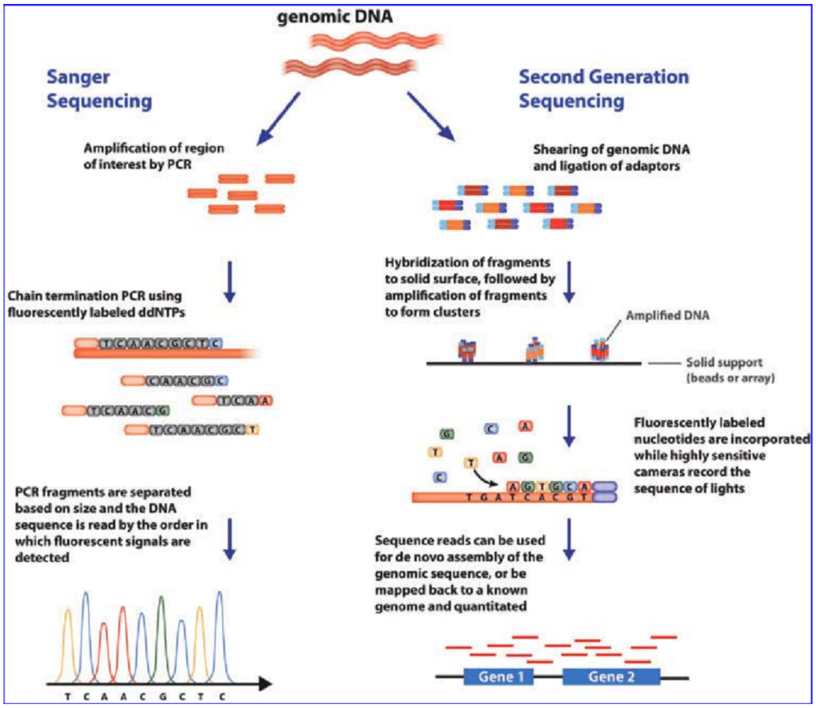
genomic DNA
Sanger Sequencing
Second Generation Sequencing
Amplification of region of interest by PCR
Shearing of genomic DNA and ligation of adaptors
Chain termination PCR using fluorescently labeled ddNTPs
Hybridization of fragments to solid surface, followed by amplification of fragments to form clusters
Amplified DNA
Solid support (beads or array)
T C A A C G C T C
C A A C G C
T C A A
T C A A C G
T C A A C G C T
G C A T T A G C
A G T G C A
T G A T C A C G T
Fluorescently labeled nucleotides are incorporated while highly sensitive cameras record the sequence of lights
PCR fragments are separated based on size and the DNA sequence is read by the order in which fluorescent signals are detected
Sequence reads can be used for de novo assembly of the genomic sequence, or be mapped back to a known genome and quantitated
T C A A C G C T C
Gene 1
Gene 2

Figure 1. Comparison between Sanger and next generation sequencing basic procedures [5].

As mentioned, and, as the name suggests, NGS technologies have revolutionized genomic research [6]. Using NGS, an entire human genome can be sequenced within a single day. This is even more remarkable when compared to the ten years that it took to complete the Human Genome Project using conventional Sanger sequencing [7]. The reason for such big difference in sequencing times can be explained using Figure 1. NGS requires the sequencing procedure to be massively parallelized with multiple different genomic fragments being sequenced at the same time, and that is not simply not possible with conventional Sanger sequencing. Additionally, each of the three billion bases in the human genome is sequenced multiple times. The number of times each base is sequenced is often referred to in bioinformatics as the coverage of the genome. That high coverage of NGS technologies has been able to provide highly accurate data and give insight into unexpected DNA variation.

The fundamental premise of cancer development is that somatic cells over the course of their lifetime acquire different mutations in different regions of their genome. Acquired mutations slowly accumulate in the cell and eventually a trigger is pressed and a cell goes through transformation towards becoming a cancer cell. Although capillary-based cancer sequencing has been able to give us a glimpse into this phenomenon it has been limited to only a selected number of genes and exons without the possibility of giving a full picture[8]. Combining the speed and the (relatively) low price of NGS technologies has allowed whole cancer genomes to be sequenced and mapped with high accuracy giving rise to large databases of known cancer genomes such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) study [9].

1.2.2. ChIP-seq

The application of NGS sequencing to chromatin immunoprecipitation (ChIP) gave rise to ChIP-seq (chromatin immunoprecipitation followed by sequencing), a key technology on the pathway to our better understanding the genomic background of cancer development. ChIP methodologies have been known for a long time [10], but have found very limited use in large sequencing experiments and projects because of their price and a lack of practicality on a large scale. The basic principle of ChIP is to use an antibody that recognizes a TF or histone modification to pull down attached DNA for identifying binding locations, and has been used traditionally as a method for detecting selected promotor regions and histone binding sites. However, the rapid development of NGS technologies has allowed chromatin immunoprecipitation to be followed by sequencing of large genomic regions and even whole genomes giving rise to ChIP-seq technology (Figure 2). ChIP-seq has since become the most common and effective method to identify bound loci genome-wide in vitro and in vivo. The importance of finding and mapping certain DNA-protein interactions and epigenetic modifications lies in the link between DNA-protein interactions and transcriptional regulation. Genome-wide profiling of transcription factor (TF)-binding sites and regions with covalently modified histones has been able to aid in the search for and even discover new cell- or tissue-, species- and disease specific- genomic regions. The RoadMap Epigenomics project focuses on analyzing and collecting ChIP-seq data from different human cell lines with a direct focus on delivering a collection of normal epigenomes that can be used for comparison and integration into number of different studies [11].

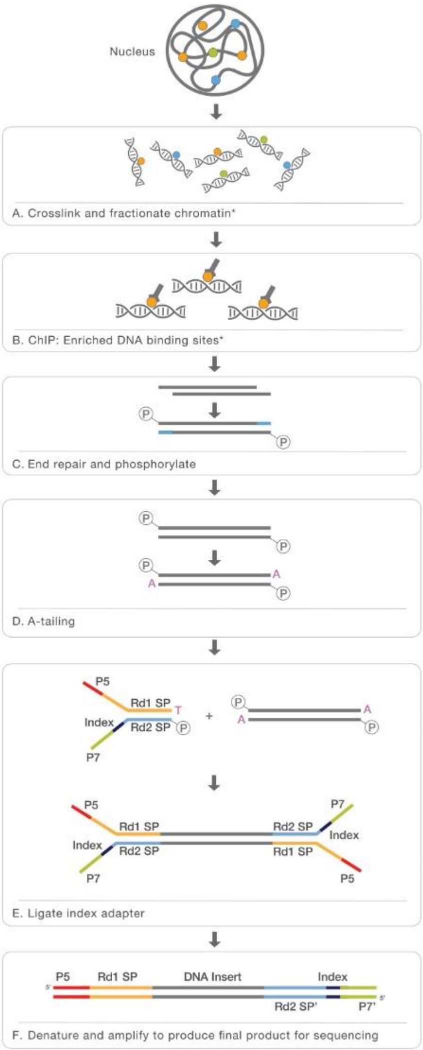
Nucleus
A. Crosslink and fractionate chromatin*
B. ChIP: Enriched DNA binding sites*
P
C. End repair and phosphorylate
P
A
D. A-tailing
P5
Rd1 SP
T
Index
Rd2 SP
P7
+
A
E. Ligate index adapter
P5
Rd1 SP
DNA Insert
Index
5'
Rd2 SP'
P7'
F. Denature and amplify to produce final product for sequencing

Figure 1. Basic workflow of a ChIP-seq experiment [12].

1.2.3. Current research in cancer genomics

Given all aforementioned findings and technologies many researches have shifted the focus of their research towards discovering, annotating and describing the characteristics of the cancer genome and epigenome and using those results for bettering the diagnostic and prognostic tools available. Approaches to these problems have been numerous. Finding and identifying definitive unique patterns of somatic mutations which distinguish certain types of cancer. Based on the large-scale exome and genome-sequencing studies we now have the somatic mutational patterns for many major cancer types including but not limited to:

- chronic myelogenous leukaemia (CML) is characterized by a t(9;22) translocation commonly known as the “Philadelphia chromosome” leading to a BCR–ABL protein fusion, which is now used as a test for the disease [13].
- melanomas is characterized with high rates of C > T and G > A transition mutations due to UV damage which causes photodimerization and faulty repairs [14].
- almost all diagnosed pancreatic ductal adenocarcinomas having mutations in the KRAS gene [15].

The aforementioned PCAWG project aims to expand on those findings by collecting and systematically analyzing cancer genome sequences from more than 2,600 patients across 38 cancer types, characterizing putative non-coding driver events that cannot be found using data from whole-exome sequencing or single-nucleotide polymorphism arrays, creating the largest database of known cancer mutations so far. In addition to those cancer-specific driver mutations on which the cancer is being selected for as they are often connected with an increase in cancerogenic potential, each individual cancer accumulates an order of magnitude larger number of passenger mutations which researchers first presumed are randomly distributed, as cancer is not undergoing any kind of selection involving those mutations and that they preserve sort of a “track record” for a mutagenic process the tumor has undergone [16]. However, detailed investigation has discovered that those passenger mutations do not accumulate at random, rather that the profile of those passenger mutations correlates heavily with a number of different epigenetic modifications [17,18].

When approached with the problem of treating metastatic cancer the first and foremost issue is finding the primary tumor organ or cell-of-origin. The metastatic tumor’s cell-of-origin and histopathology are the strongest determinants of its clinical behavior and therefor a clear pathway to finding the best optimal route to cancer treatment, but in 3-5% of cases patients with a metastatic tumor do not have an obvious cell-of-origin [19]. Determination of the correct cell of origin for a metastatic cancer is a key factor in cancer treatment. There have been several studies which have highlighted the importance of determining the correct cell of origin. For instance, it was shown that patients with the same driver mutation, but appearing in different cancer types (i.e. different cell of origin) will have a different response to treatment [20]. Another study, conducted in a mouse model of glioblastoma, showed that drug sensitivity differed according to the cell-of-origin [21]. Based on this problem a class of emerging approaches aims to classify cancers based on somatic passenger mutation profiles alone, without the need to find and identify all special cancer type identifying genes and regions, which is much more complicated and not reliably testable for all cancer types. The machine-learning models so far have been able to accurately predict the tissue of origin given the epigenetic modifications of many different tissues. Those models use large genomic and epigenomic datasets in order to accurately predict the correct tissue of origin of a given somatic passenger mutational profile [18]. With highlighting the importance of determining correct cell of origin the question of the diagnostic potential for complex machine learning algorithms arises. The main issue in using these models for diagnostics and treatment is the need for whole cancer genome sequencing and assembly. Due to this it is necessary to improve on the existing methods for cell-of-origin determination and to simplify their interpretation in the context of biological systems.

1.3. Research Goals

The main goal of this research is to find the optimal genomic regions to use in the predictive models which determine the melanoma cell-of-origin with high accuracy. We will be looking for the regions on which the models trained on chromatin modification values from melanocytes (the correct tissue of origin for melanoma) have the biggest difference in prediction accuracy from other cell line chromatin modification values and to investigate the possible genomic background for the behavior of the model. To achieve those goals, the investigation will be split into 3 different parts:

1. Using principal component analysis to identify the regions which contribute to principal components the most in order to find the optimal number of regions used for prediction.

2. Finding the optimal regions by using the results from 1. and compare the models trained on different subsets of regions.
3. Analysis of known sequence features and known genomic elements located the selected regions.

2. Methods

2.1. Data

The data used in the research is split into two groups: predictor variables (genomic localization of histone modifications) and response variables (mutations profiles of melanoma cell lines). Histone modifications used are H3K4me1, H3K4me3, H3K9me3 and H3K36me3 from 83 different tissues/cell lines, included in ENCODE and RoadMap Epigenomics projects [28]. Mutation profiles of melanoma cell lines are obtained from Pan Cancer Analysis of Whole Genomes (PCAWG) project and publicly available [9]. The human genome is divided into 1 megabase (Mb) regions, excluding regions overlapping centromeres and telomeres, and regions where the fraction of uniquely mappable base pairs is lower than 0.92. In total, 2128 1Mb long genomic regions were used in the analysis. The histone modification data originating from the same cell types is combined and the RPKM value for each of the predefined regions is calculated. The melanoma mutation profile data is composed of 107 different patient samples, all of which are also divided into identical regions. All of this data has been previously processed and the data used in this research has been obtained from the research [18].

Genomic coordinates of super-enhancers data were downloaded from the human super-enhancer database (SEdb) [23]. SEdb is a publicly accessible database with a goal aimed to provide a large number of available resources on human super-enhancers. The database was annotated with potential functions of super-enhancers in the gene regulation. Melanoma super-enhancer genomic positions are included in the database.

In order to find the transcripts and exons located in the genomic regions of interest we used the “GenomicFeatures” package in R [29]. The package is used to retrieve known transcript-related features from the UCSC Genome Bioinformatics [30] and BioMart [31].

Melanoma driver gene data is part of larger research data which focused on finding driver point mutation across 2,658 different genomes in non-coding regions from Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium5 of the International Cancer Genome but data from protein coding point mutations is also included and can be publicly accessed at [32] (under "Supplementary tables") .

We retrieved the sequences of the human genome data from UCSC [30]. Via the "BSgenome" package for R [33]. "BSgenome" package retrieves the latest UCSC human genome and the version we used in our research was the human genome version hg19 (based on GRCh37.19.p13 assembly). Genome data was used to calculate the percent of GC content for each of the 2128 genomic regions analyzed in this study.

2.2. Computational methods

2.2.1. R statistical package

All the analyses in this research were conducted in the R statistical environment, a freely available software and programming environment for statistical computing and graphics. R version used in this research is 3.6.2 [24]. Overlaps between genomic regions and drivers, super-enhancers were calculated using the "GRanges" package [25]. Granges uses formatted genome coordinate range tables to identify overlaps. Genome coordinate ranges are "addresses" of the selected regions or genes in the chromosome given in the format of i.e. [chrX; 1-100] (this means that the selected "gene" is located on the X chromosome, gene starts at the first base of the chromosome and ends at the 100th base of the chromosome).

2.2.2. Principal component analysis

Principal component analysis (PCA) is a classical tool to reduce the dimension of multi variate and high dimensional data, as often seen in gene expression analysis. PCA is often used in those cases to visualize the similarities between the biological samples, and to filter noise. The basic principle of PCA is it that projects highly dimensional data into a new space spanned by the principal components (PC) calculation is of which is based on the variance in the data, trying to use as little possible number of new PCs to explain as much variance as possible with the total number of PCs being equal to N-1 (number of dimensions in the data - 1). All calculated principal components are uncorrelated and orthogonal. The PCs can successfully extract most of the relevant information in the data [26]. In this research the PCA method will be implemented via the “factoMineR” [27] package in R. The package is specifically designed for high-dimensional biological data and is easily manipulated. Main result of PCA which will be used are contributions of different dimension (or regions in this case) which describe how much of the resulting variance explained by PCs is attributed to certain dimensions (regions) in order to give a guided way to identify the most informative regions. Contributions of variables to the principal components are given by the formula:

$$cont_{var}(in\ \%) = \frac{var_{cos^2}}{{cos_{total}}^2} * 100$$

Where $\cos^2$ is equal to the squared coordinates of the variable which are calculated by the PCA algorithm. Total $\cos^2$ value is equal to the sum of $\cos^2$ values of all variables in the component.

2.2.3. Random Forest regression

Random Forest regression is a non-parametric machine learning method developed in early 2000’s which uses ensembles of simple decision trees to capture more complex feature patterns and reduces the chance of overfitting to training data when compared to simple decision trees. The basic principle of the method is to draw a random training set of size n, with replacement, which, when averaged out, on the many trees the algorithm draws ends up approximately 2/3 of the whole data. The remaining 1/3 of the data (often referred to as out-of-bag data) is used to compute the mean squared prediction error of the tree. To calculate the prediction for a given observation the algorithm takes the average of predictions over all trees for all out-of-bag data. The resulting diversity of trees can capture more complex feature patterns than a single decision tree and reduces the chance of overfitting to training data. In this way, the random forest improves predictive accuracy [28].

Here Random Forest algorithm is implemented via “ranger” package in R [28]. The package is designed for high dimensional biological data. We are using forests with 500 trees to predict the mutation densities in the previously constructed 1 Mb regions of melanoma patients. The individual patient samples were divided into ten non-overlapping sets and the total number of mutations in each region was used for model predictions using tenfold cross-validation. Prediction accuracy of each model was measured as rooted mean squared error:

$$MSE = \sqrt{\Sigma(x_{expected} - x_{fitted})^2}$$

Cross-validated prediction accuracies were calculated by the following procedure:

1. Divide the patient data into training set data (9/10) and test set data (1/10).
2. Determine the regions on which the random forest algorithm will be trained (i.e. selected 200 regions).
3. Train the model using the predictors of test set data for regions determined in 2.
4. Use the model to predict the mutation density profiles for ALL regions in the test data given the predictor values of ALL regions from the test set data.
5. Calculate the test MSE of the model by comparing the fitted values and the real values of mutation profile densities.
6. Repeat the procedure 10 times, each time using different sets of training and test data (selected patients mutation profiles).
7. Take the average of the MSE values across the cross-validation data sets and use that metric as a measure of the model accuracy.

This way of calculation of cross-validated MSE has some drawbacks, i.e. the smaller the training set number of different regions the MSE will be higher, as the similarities between patients in genomic regions are greater than similarities between the mutation profiles of different genomic regions (regional profiles are independent of one another, whilst different samples have correlated mutational density profiles in the same genomic regions). Nonetheless this approach gives a glimpse on how the model will behave when tasked to classify unknown data (i.e. if we select 200 regions for training and then use the full set of 2128 regions for testing) which is a far more valuable metric for diagnostic purposes.

2.3. Statistical tests

2.3.1. Pearson Product-Moment Correlation

The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. The formula for the correlation is

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - mx)^2 \sum(y - my)^2}}$$

Where m_x and m_y are the means of x and y variables.

The formally used measure for prediction accuracy is r^2, which is used and displayed on the graphs used in this research.

2.3.2. Wilcoxon ranked sum test

Wilcoxon ranked sum test is a standard non-parametric statistical test used when the distributions of the tested data are not known or does not have a defined expression. The implementation of the test I carried out in base R programming package under the function call "*wilcox.test*" and we use it to measure the p-value of an observation. The null hypothesis is that the distributions of x and y do not differ by a definitive location shift, and the p-value calculated is the probability of seeing that outcome, or to better put it to test whether set of observations x and a set of observations y, where x and y are two independent samples, come from the same distribution and what is the probability of seeing those observations if they do.

2.3.3. Fisher exact sum test

Fisher exact sum test, or Fisher's test for short, is a statistical significance test used in the analysis of contingency tables. Unlike most statistical tests, Fisher's exact test does not use a mathematical function that estimates the probability of a value of a test statistic; instead, you calculate the probability of getting the observed data, and all data sets with more extreme deviations, under the null hypothesis that the proportions are the same under a hypergeometric distribution. The test is implemented in R under the call "*fisher.test()*". The result of the Fisher's test is a p-value for observing the contingency table under the presumption that all of the proportions are conserved.

3. Results

3.1. Principal component analysis results

The first part of the research was to perform and exploratory analysis of the datasets by using principal component analysis on the both predictor and the response variables. The layout of original datasets had to be transposed in order for PCA to explore the variance between the regions and not between different predictors/patients. The results of the PCA for both the predictor and response variables are included in Figure 3.

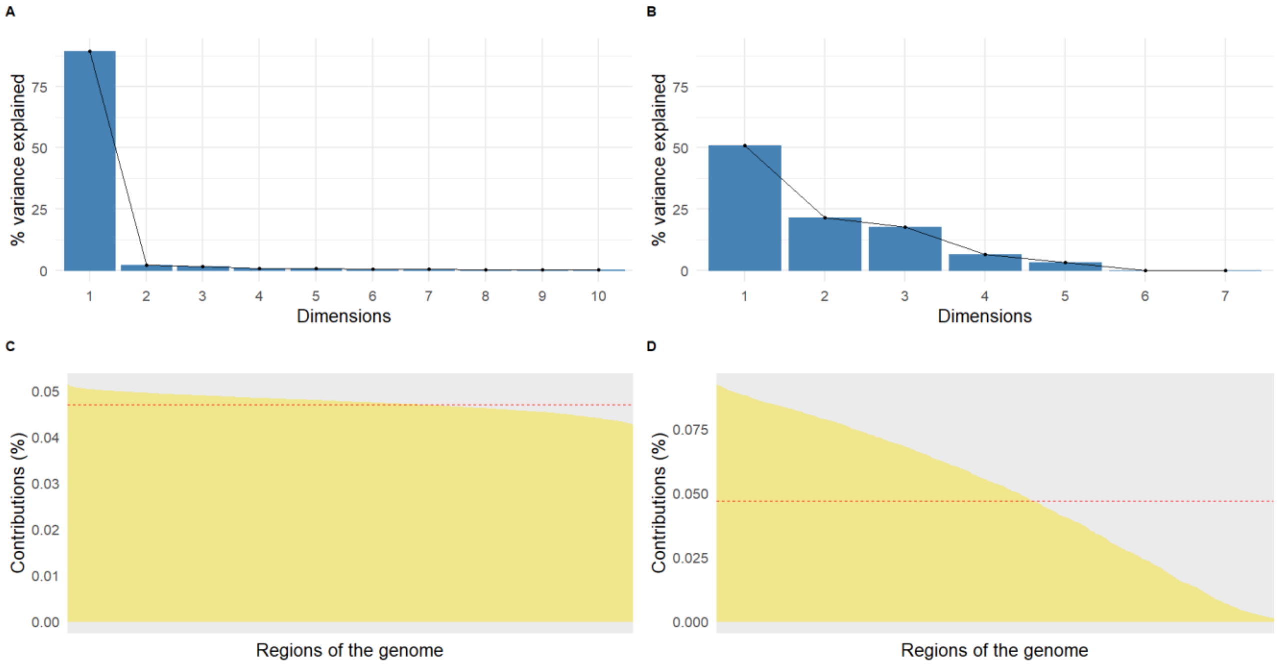

Figure 3. Scree and contribution plots of analyzed data sets. *Figure 3.A is the scree plot of response variable PCA. On the Y axis is the percentage of total variance explained by the principal component, and on the X axis are principal components ordered by their relative percentage. Figure 3.B is also a scree plot but of predictor variable PCA. Figures 3.C and 3.D represent relative contributions of genomic regions to the variance explained by the first largest principal component from A and B respectively.*

From the scree plots we can see that most of the response variable can be explained by only 2 PCs, with up 87% of the variation in the mutation profile data explained in only one dimension and the rest of the PCs contributing only a fraction of that to the overall variance explained percentage. In the case of predictor variables, the variance in regions has to be explained in a bigger number of PCs with 3 different PCs totaling the percentage variance explained of ~90 %. The contribution of individual regions to variance in PCs in the case of response variables Figure 3.C differs only very slightly between the ones that contribute most and the ones that contribute the least with maximal difference <10% of the mean value for the individual contributions. In the case of predictor variables (Figure 3.D) the difference is much more pronounced with a large drop in contributions throughout the different regions of the genome.

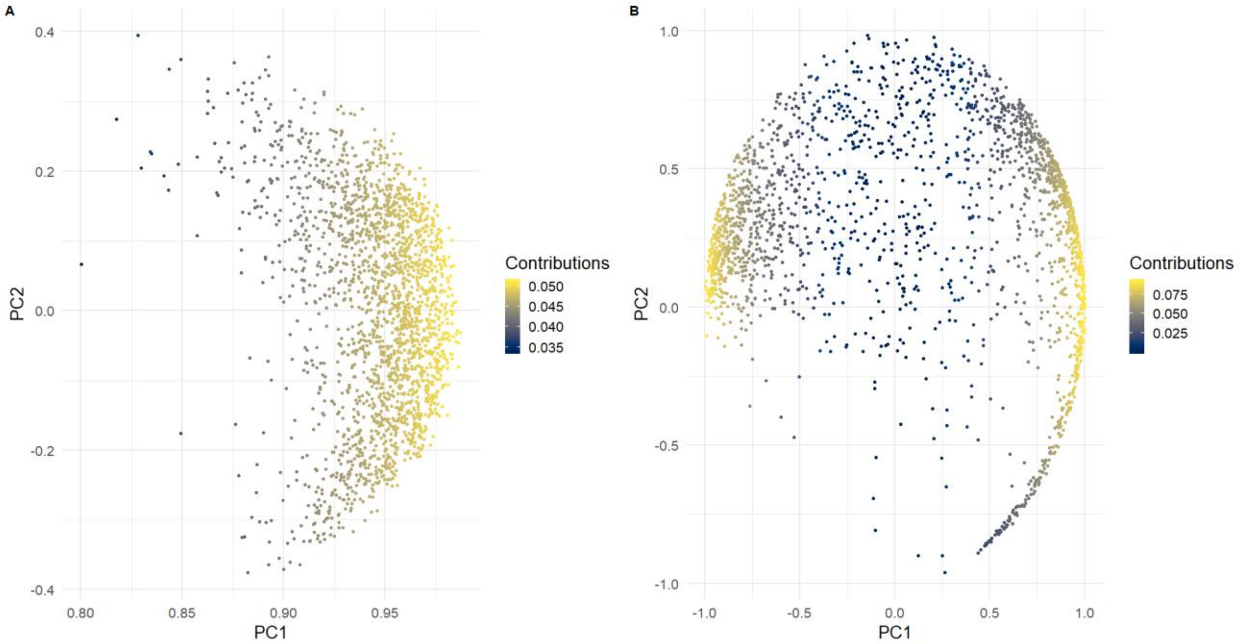

Figure 4. PCA biplots of the predictor and response variables. *Figure 4.A is a standard PC1/PC2 loading vector biplot for PCA of response variables, with values on the X and Y axis corresponding to the value of the loading vector with different colors representing different relative contributions to the % of variance explained. Figure 4.B is the PC1/PC2 loading vector biplot for PCA of predictor variables, colored with regard to different relative contributions. Note that the contributions in this case are much more diverse as can be presumed from Figure 3.D where the slope is much more pronounced when compared to Figure 3.C*

PCA biplots in both cases show that there no particular clustering of data in those dimensions with a highly pronounced ellipsoid shape in both cases. The conservation of the ellipsoid shape can be presumed as some fundamental property of the data, such as interconnectedness of all epigenetic variations and the relative bounds of the data and low variance regions in both the predictors and the response variables. The investigation of causes for this ellipsoid shape is beyond the scope of this research.

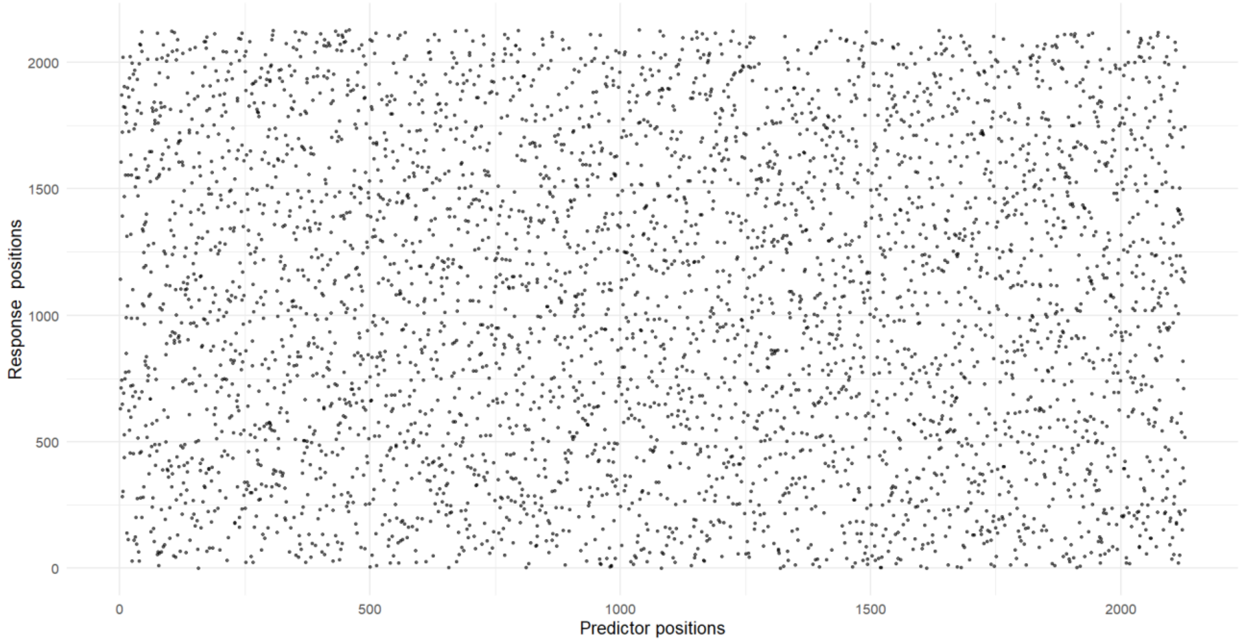

Figure 5. Position vs. position dot-plot of ordered PCA contributions. *The figure represents is a dot-plot of contribution ordered regions of response and predictor PCA results. Each dot represents a positional match e.g. region chr1.11 region is 100. in predictor contribution ordering and 600 in response variable ordering so the dot is placed on the (100,600) place on the graph**.***

We then explored the relationship between the ordering of regions using the results of PCA of either predictor or response variables. From Figure 5. we can see that there are no particular regions which have any kind of clustering, rather that the positioning is pretty uniformly and randomly distributed throughout the graph. If the orderings were related, we would expect to see particular clusters or diagonal elements existing on the graph. Having no clusters or diagonal elements in this case is a favorable option because it means that there are no region-specific biases between predictor and response variables which could mean possible biases in model training and testing.

3.2. Identifying the optimal regions for model training

Next part of the exploratory analysis was finding out how does the prediction accuracy behave in accordance to the number of regions used in modelling of the number of mutations in specific regions of melanoma genomes, with the goal being finding the optimal number of different genomic regions with which to begin the search for the best ones, and the result of the search are included in Figure 6.

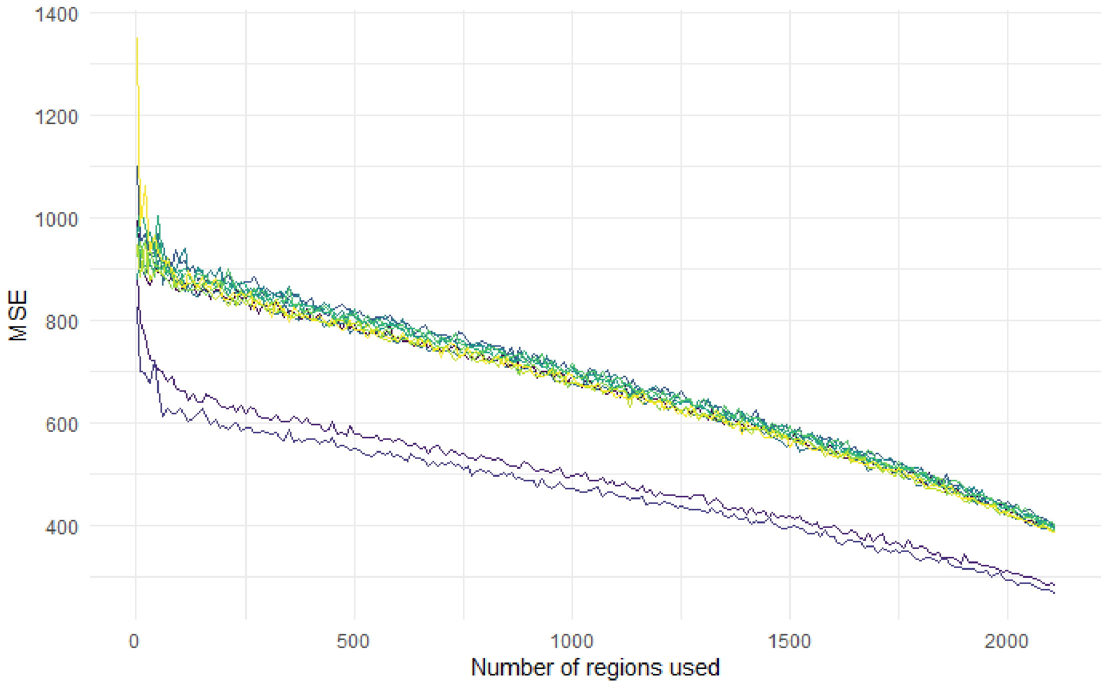

Figure 6. Dependence of sample sizes used for model training on MSE. *Regions used for model training and predictions are randomly sampled via the generic sample() function. The 2 lines below represent both melanocyte cell lines and are clearly separated from the rest of the cell*

samples. Grouped lines above are randomly sampled are 20 different cell-of-origin. Here used in the calculation is a size increase step of 10, meaning that number of regions sampled increased by 10 in each calculation, which resulted in a 10-fold increase in speed, but the resulting MSE values became more unstable. It can be seen that y-values after the initial drop begin to stabilize around 200 regions being used.

From the plot it is visible that the behavior of the prediction accuracy vs. total number of regions is predictable. We can see that the relative predictive power is lower when the total number of genomic regions is lower. When the number of regions reaches 200-220 the predictive power stabilizes and begins to linearly drop off, in the case of correct cell of origin being used for model training while the drop is less pronounced in other cell lines used. Using that number of regions in ordered search is therefore favored as any differences between models using the correct cell-of-origin is maximized when using ~200 regions.

After finding the proposed optimal number of regions to start our search for the best regions, the next step was starting the ordered search based on ordering from Figure 3.C and Figure 3.D. Using a sliding window of a size 200 as established from Figure 6., with a side-step of 10 regions, meaning that for each step of calculation the window is moved 10 regions down the ordered list, with 10-fold cross validated MSE the results from Figure 7. were obtained.

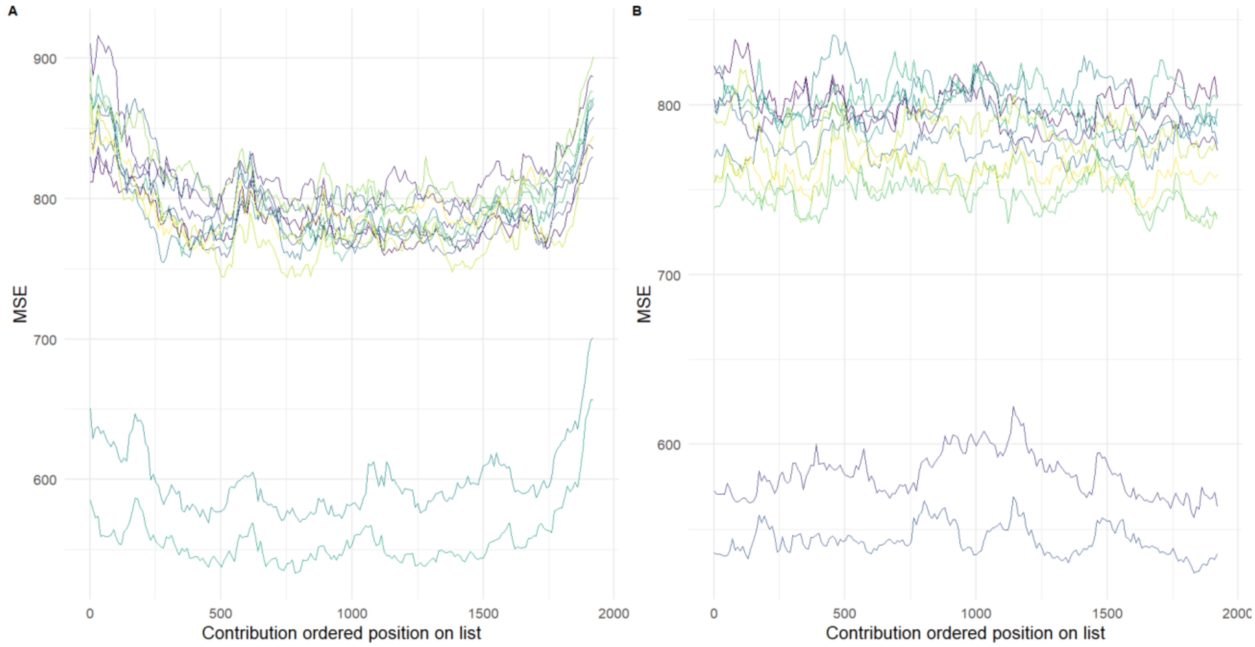

Figure 7. Optimal number of selected windows moved over ordered positions. *Figure 7.A are response variable PCA contribution ordered regions; Figure 7.B are predictor variable PCA contribution ordered regions. On Y-axis is the MSE for a window which starts at the x-axis marked tick and continues for next 200 regions from the ordering. Bottom 2 lines on both graphs are 2 samples from melanocyte cell lines used for model training, above lines are 30 randomly sampled cell samples from other tissues.*

Using response variable ordering results in a more divergent graph which has bigger differences between individual regions, while using predictor variable ordering results in a graph which is more uniform, with less pronounced differences between different regions, as well as a smaller total difference between melanocyte and other tissue samples. To quantify the difference between melanocyte and the other cell samples and to identify the regions with the biggest predictive power we calculated the difference between the average MSE values of the melanocyte cell samples and the average MSE values of the other cell samples.

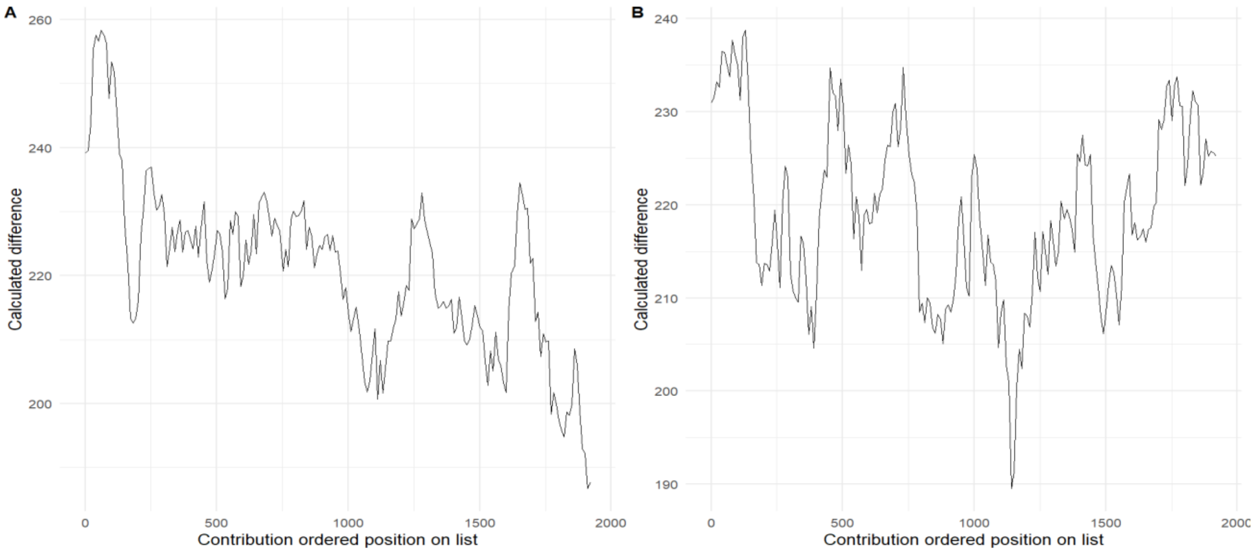

Figure 8. Average difference between melanocyte contribution ordered MSEs and other cell lines MSEs using the same ordered contribution ordered lists. *Figure 6.A represents the calculated differences as explained above for response variable PCA contribution orderings; Figure 6.B represents the calculated differences for predictor variables PCA contribution ordering. Y-axis represents the values of the calculated differences and x-axis representing the relative positioning in the respective orderings.*

As can be seen from the Figure 8.A, for the response variable orderings the differences are more pronounced with a pronounced peak in the beginning, although the mean calculated differences are approximately the same (for Figure 8.A the mean is 220.76, and for Figure 8.B the mean is 219.06). Those regions which fall within the pronounced peak of the window which starts at the 70th region of the ordering (meaning it includes regions from 70-270, z-score of the values obtained under the presumption of normal distribution is equal to 2.6705 which translates to a p-value = 0.00265) are those regions we hoped to identify, regions which enable us the biggest predictive power, or the biggest separation in model accuracy between the correct cell-of-origin and all the rest. The next step of the research is confirming that those regions separate between correct cell-of-origin equally effective as models which include all genomic regions. In the rest of the research we have used the results from response variable contribution orderings because the predictor variable contribution orderings provided no statistically significant difference ($P<0.01$) between different regions.

3.3. Exploring the predictive quality of the selected regions

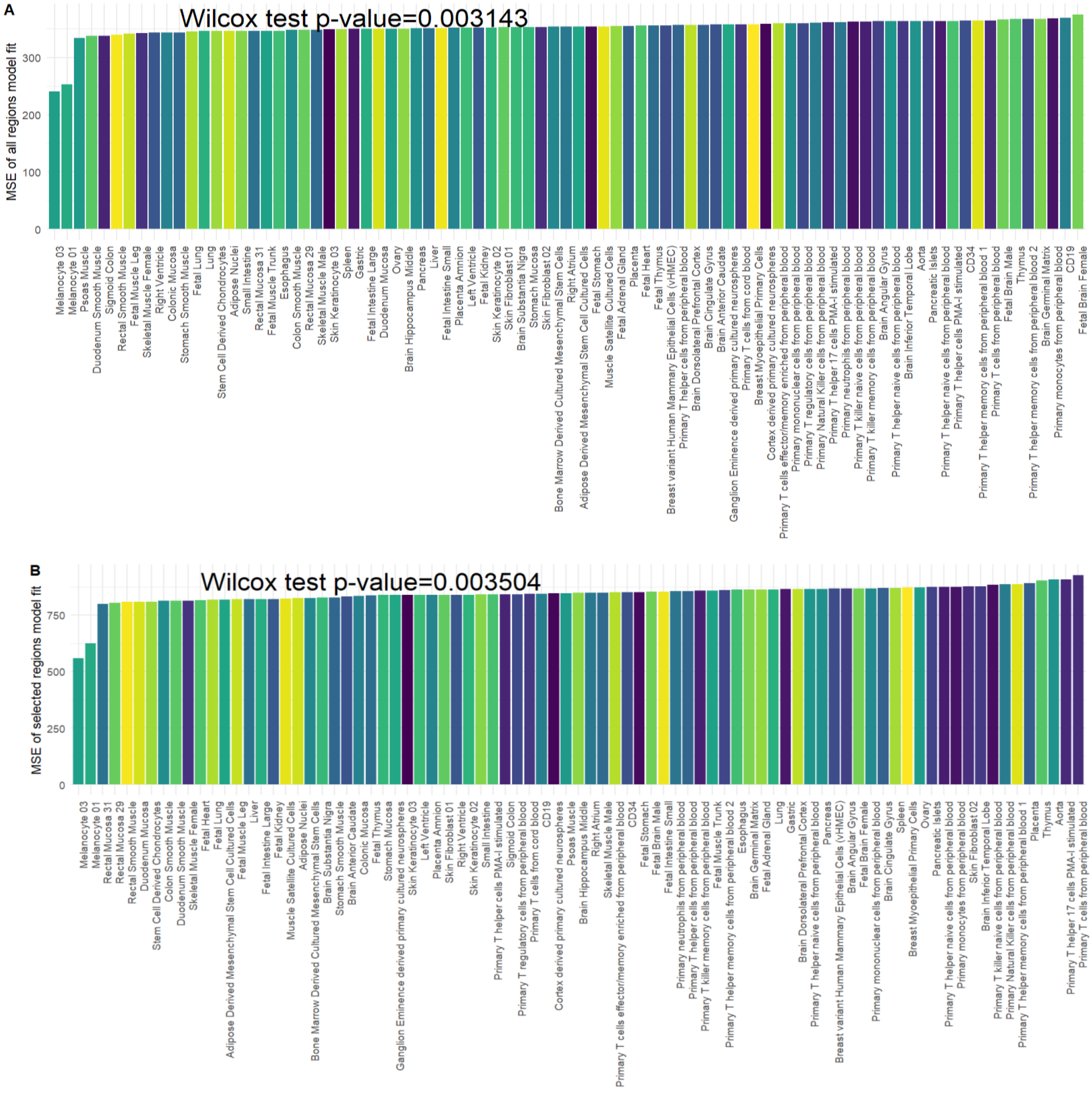

Figure 9. Prediction accuracies for models trained on the selected regions across different tissue samples. *Figure 9.A represents relative MSE of models trained on full set of genomic regions with each bar representing different tissue sample. Figure 9.B represents relative MSE of models*

trained on regions selected in Figure 8.A. On the x-axis are different tissue samples ordered by their size. Both figures have Wilcox rank sum test calculated p-values of the difference between melanocyte cell samples tested versus all other tissue samples. All MSE scores are 10-fold cross validated.

From Figure 9. we can clearly conclude that models trained on selected regions, albeit having a larger total MSE than those trained on full data, can be used to differentiate the correct tissue of origin given the mutation profile of melanoma cancer cell with accuracy which is almost identical to full genomic region, since we have significant (*P*<0.01) difference of the melanocyte MSE using the Wilcox ranked tests. To further explore goodness of fit of the models, graphs of real vs. fitted values were constructed for both melanocyte cell lines and the second-best tissue in both full region models and selected regions models ("Psoas Muscle", "Rectal Mucosa 31" and "Rectal Mucosa 29" samples respectively) and are contained in Figure 10.

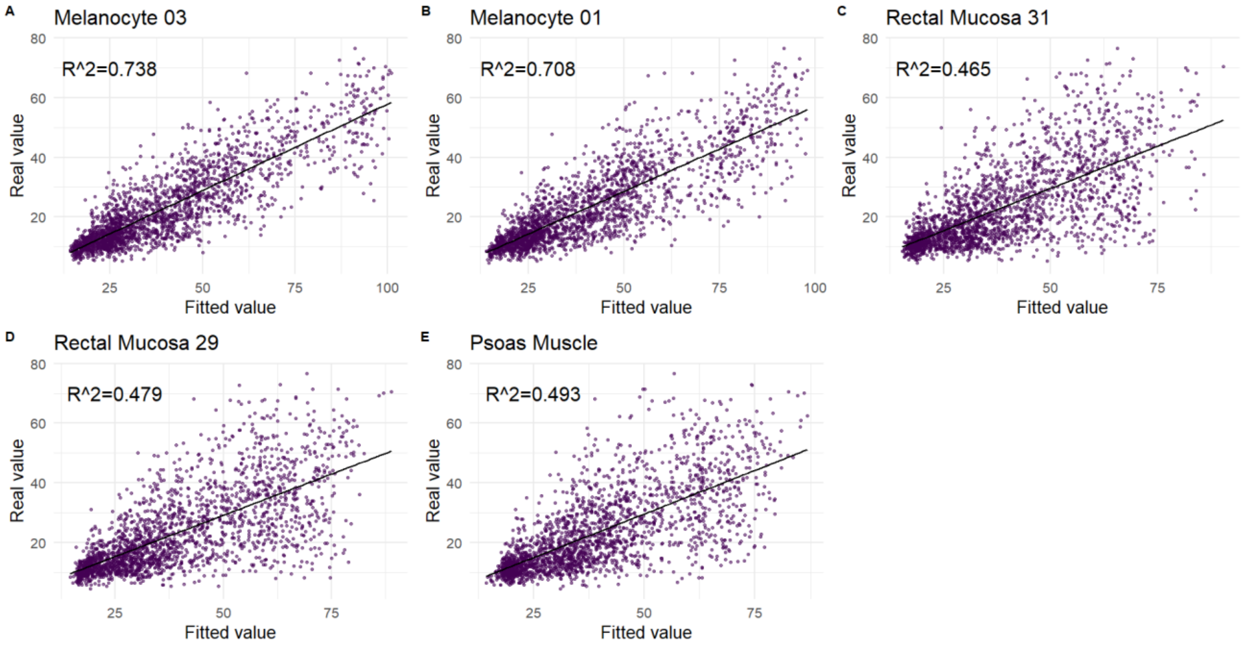

Figure 10. Real (observed) vs. fitted value plots of selected region trained models of selected tissue samples. *R^2 values are calculated as Pearson correlations coefficients squared and displayed for each sample individually measured between observed and fitted values, with higher R^2 meaning better fitted values. The line represents a generalized linear model fit to the data.*

As shown above, the models trained on selected regions and correct cell-of-origin samples are more accurate and correlate much better to observed data in comparison to second-best tissue model fitted values. Biggest errors (or misclassifications) are focused on regions with higher mutation densities and from there stems the better accuracy of correct cell-of-origin as for Figure 9.A and Figure 9.B the spread is in those regions is much lower and the values are closer to the diagonal. In contrast, all tissues are fairly accurate when it comes to predicting lower mutation density values.

3.4. Exploring the genomic background of selected regions

The next step of the research was to try and identify the underlying reasons for such observations, the reason why those selected regions give us better prediction accuracy when compared to the rest. For this it is crucial to first check the distribution of the mutation densities in those regions and compare them to the rest general distribution of the mutation densities across all regions. The reason for this is to determine GC – content of the selected regions because of the known bias of sequencing to produce high read values in GC high regions of the genome.

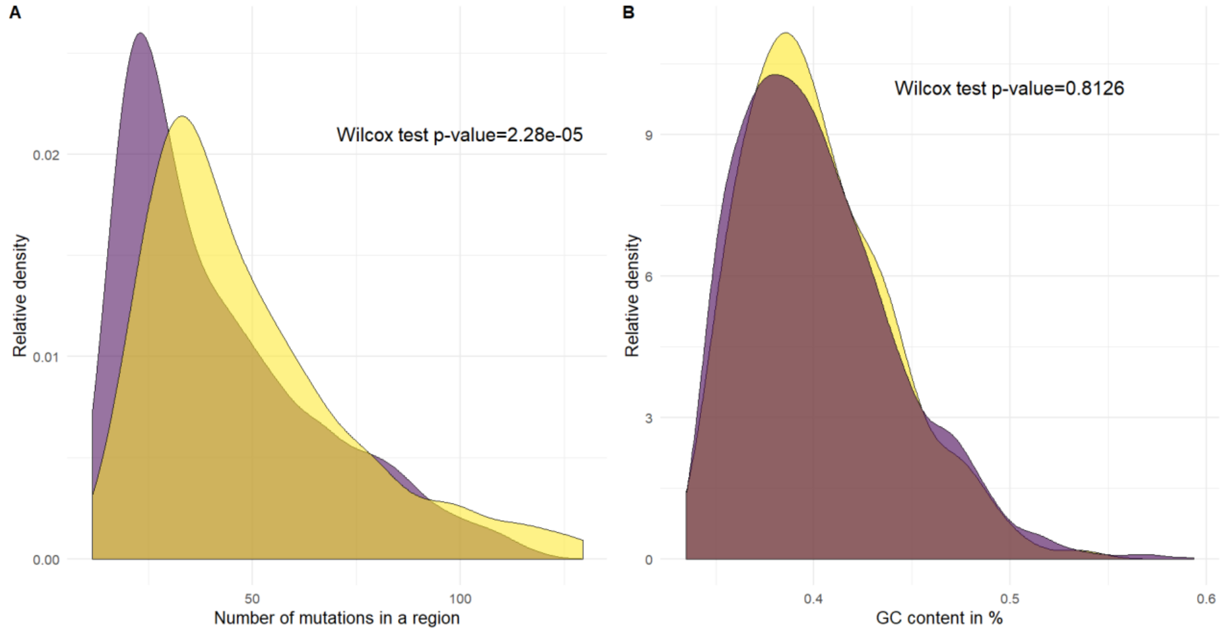

Figure 11. Distributions of mutations densities and GC content in selected regions. *Figure 11.A represents the relative density plot of selected regions aggregated sample mutations (yellow) and all regions aggregated sample mutations (purple). Above displayed is the Wilcoxon rank sum test p-value for difference in distribution means. Figure 11.B represents the relative densities of all regions GC content in % (purple) and selected regions (yellow). Also, result of the Wilcoxon rank sum test is displayed on the graph. GC content was calculated from UCSC hg19 genome data using overlaps with genomic regions used, adding the total number of GC base pairs and dividing with total number of base pairs.*

From the Figure 11.A depicted results we conclude that there is a statistically (Wilcoxon rank sum test p-value<0.01) significant difference between the mutation density distribution in selected regions when compared to all genomic regions mutation density distribution. From Figure 11.B we can conclude that there is no statistically significant difference (Wilcoxon rank sum test p-value=0.81) between the GC content of the selected regions and GC content of the full genomic regions used. This shows that the higher mutation density value obtained from mutation profiles does not stem from the known sequencing GC content bias and as such does not need to be corrected for.

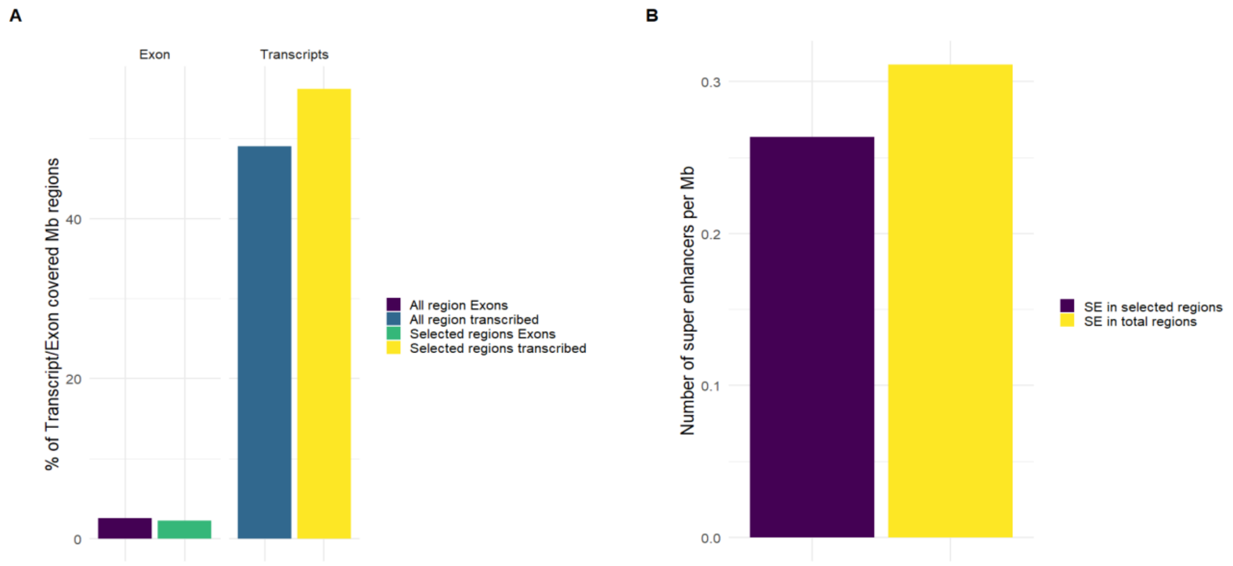

***Figure 12. Calculation of the representation of intraregional genomic elements and super enhancers**. Figure 12.A is the result of overlapping selected and all genomic regions with known genomic annotation of known exons and transcripts from UCSC obtained via "GenomicFeatures" package. Y-axis values are percentages of Mb regions covered in known transcript/exon annotation calculated as:*

$$\% = \frac{\sum Intraregional\ transcrit/exon\ lengths}{Number\ of\ genomic\ regions * 10^6} * 100.$$

Figure 12.B is the result of overlapping selected and total regions versus database of known melanoma linked super enhancers from the aforementioned SEdb. Y-values is the mean number of super enhancers per Mb resulting from the overlap.

Exon content in both groups of selected regions is very close to the general exon content in all regions taken into consideration and in accordance with general experimental data which suggest that around 2% of the genome is made up of exons. Real difference is seen in both cases when examining relative percentage of transcribed regions. When examining the comparison between all regions and selected regions percentages of transcribed total sequence there is a 7% increase in transcription in the selected regions. There is no significant difference in super enhancer densities between selected regions and all regions, with 79 super enhancers found in selected regions vs 662 super enhancers found in all used genomic regions out of total 882 super enhancers found by previous researchers.

***Table 1**. Overlaps between know protein coding drivers and the selected genomic regions. Fisher exact test p-value = 0.6124.*

	OVERLAPS PROTEIN CODING DRIVER	DOES NOT OVERLAP PROTEIN CODING DRIVER
REGIONS IN SELECTED REGIONS	0	200
REGIONS NOT IN SELECTED REGIONS	9	1919

***Table 2.** Overlaps between non-protein driver genes and selected genomic regions. Fisher exact test p-value = 0.8652.*

	OVERLAPS NON-PROTEIN CODING DRIVER	DOES NOT OVERLAP NON-PROTEIN CODING DRIVER
REGIONS IN SELECTED REGIONS	9	191
REGIONS NOT IN SELECTED REGIONS	96	1832

***Table 3**. Combined overlaps for protein and non-protein coding melanoma driver genes in the selected regions. Fisher exact test p-value = 0.7415*

	OVERLAPS EITHER PROTEIN OR NON-PROTEIN CODING DRIVER	OVERLAPS NEITHER PROTEIN NOR NON-PROTEIN CODING DRIVER
REGIONS IN SELECTED REGIONS	9	191
REGIONS NOT IN SELECTED REGIONS	106	1822

The overlap Fisher exact test results (Table 1., Table 2. and Table 3.) imply that there is no statistically significant difference between the presence of identified protein and non-protein coding driver mutations in the selected regions we used in the research.

4. Discussion

The results of PCA on both the response and predictor variable showed that there are potentials to reduce the number of regions used in the models to predict the correct cell-of origin. Both the response and predictor variables can have a large percentage of their variance explained by a relatively few number PCs. PC biplots showed that there were no definitive clustering patterns of the data in the main PCs analyzed, which is useful for downstream analysis as there were no biases toward certain clusters of regions or outliers which could interfere with the predictive power of the models. Contributions of regions to the response variable PCA showed a less pronounced overall difference, meaning that all regions contribute relatively uniformly to the PC. Contributions of the regions to the predictor variable PCA showed more promise because the difference between the most and least contributing regions was larger when compared to the response variable differences.

We have also shown that there is little connection between the contribution orderings of the two PCAs, which means that there are no clusters of regions which contribute to both the predictor and response variables relative variances equally. This finding was important as it showed us that the two contribution-based orderings are mutually independent so if potentially more informative regions for model training were to be discovered we had to search both of orderings separately. Finding the optimal number of regions was an important step forward as it was shown that the melanocyte samples converge to a linear drop in prediction accuracy based on the number of regions used relatively fast. Also, it was shown that even a low number of regions were enough to successfully separate the correct cell-of-origin and there was a low absolute difference in prediction accuracy between the two melanocyte samples we had in our data sets. Although models of prediction of cell-of-origin based on various features of cancer genomes and normal cells were developed previously [18,34,35], to our knowledge this is the first study which uses epigenomic data of normal cells to perform unbiased selection of genomic regions needed for successful cell-of-origin prediction.

Although PCA results of predictor variables showed more promise in the beginning of the research, relative differences of melanocyte sample MSEs to other tissue sample MSEs across the ordered contributions were uniform with no particular regions which showed statistically ($P<0.01$) larger difference from the mean prediction accuracy of all windows. On the other hand, response variable PCA result based orderings of the regions have managed to produce statistically significant ($P<0.01$) difference from the mean of the differences under presumed normal distribution.

The selected regions have managed to separate the melanocyte samples from other tissues with relative ease. More fascinating was the fact that the selected regions separate the tissue samples with accuracy almost identical to full genomic regions models meaning that even 10% of the whole genome can be used to successfully determine the correct cell of origin.

Investigation into the genomic background of the results provided mixed results.

On one hand we managed to show that there was no GC-content bias which is commonly associated with NGS experiments. Furthermore, the selected regions had a significantly larger proportion of regions with a high amount of passenger mutations. This result points to the fact that regions with a high number of mutations are more informative for model training than regions with a lower number of mutations. This is in agreement with previous studies which showed that the prediction power of such models depends on the number of mutations in a region [18]. We have also shown that the selected regions also have a higher percentage of their regions transcribed when compared to all regions. This is not surprising, considering that gene expression was previously shown to be related to mutational density [36], and shows that the influence of transcriptional rate of the selected regions on prediction accuracy should be investigated further.

On the other hand, investigation into known melanoma super-enhancers and cancer drivers provided little results that could explain our findings. We found that there were no differences between super enhancer densities in selected regions when compared to the rest of the regions used. Adding to that we also found that there were no differences between the presences of protein and non-protein coding driver mutations in the selected regions when compared to the rest of the regions used. An explanation for this potentially lies in the fact that we were using passenger mutations, which are independent mutations to driver and super-enhancer mutations and are shaped by epigenetic regulation [37].

5. Conclusion

In conclusion, this research has been successful in its goal of finding the most informative regions to be used in melanoma-cell-of origin prediction. We have identified regions which comprise of only 10% of the genomic regions used in previous studies which can successfully identify the melanoma cell-of-origin with a degree of accuracy comparable to larger datasets previously used. Although failing to find the connection between passenger and driver mutations in melanoma this finding successfully opens the question of the diagnostic potential for complex machine learning models in metastatic melanoma cell-of-origin detection. The focus of future studies on this should be identifying the most informative regions of other malignant diseases which show the same potential in order to create a comprehensive guide for metastatic cell-of-origin discovery across multiple cancer types.

6. References

1. USA public cancer statistics (01.07.2020): https://www.cancer.org/research/cancer-facts-statistics/
2. Gray-Schopfer, V., Wellbrock, C., & Marais, R. (2007). Melanoma biology and new targeted therapy. Nature, 445(7130), 851–857. https://doi.org/10.1038/nature05661
3. D'Orazio, J., Jarrett, S., Amaro-Ortiz, A., & Scott, T. (2013). UV Radiation and the Skin. International Journal of Molecular Sciences, 14(6), 12222–12248. https://doi.org/10.3390/ijms140612222
4. Leonardi, G., Falzone, L., Salemi, R., Zanghi, A., Spandidos, D., Mccubrey, J., Candido, S., & Libra, M. (2018). Cutaneous melanoma: From pathogenesis to therapy (Review). International Journal of Oncology. https://doi.org/10.3892/ijo.2018.4287
5. Bunnik, E. M., & Le Roch, K. G. (2013). An Introduction to Functional Genomics and Systems Biology. Advances in Wound Care, 2(9), 490–498. https://doi.org/10.1089/wound.2012.0379
6. Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? Archives of Disease in Childhood - Education & Practice Edition, 98(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340
7. Hood, L., & Rowen, L. (2013). The human genome project: big science transforms biology and medicine. Genome Medicine, 5(9), 79. https://doi.org/10.1186/gm483
8. Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., … Velculescu, V. E. (2006). The Consensus Coding Sequences of Human Breast and Colorectal Cancers. Science, 314(5797), 268–274. https://doi.org/10.1126/science.1133427
9. Official PCAWG project webpage and repository (28.6.2020): https://dcc.icgc.org/pcawg
10. Jackson, V. (1978). Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. Cell, 15(3), 945–954. https://doi.org/10.1016/0092-8674(78)90278-7
11. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330. https://doi.org/10.1038/nature14248
12. Illumina webpage containing ChIP-seq infographic (29.06.2020): https://www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html
13. Kang, Z.-J., Liu, Y.-F., Xu, L.-Z., Long, Z.-J., Huang, D., Yang, Y., Liu, B., Feng, J.-X., Pan, Y.-J., Yan, J.-S., & Liu, Q. (2016). The Philadelphia chromosome in leukemogenesis. Chinese Journal of Cancer, 35(1). https://doi.org/10.1186/s40880-016-0108-0

14. Ikehata, H. (2003). UVA induces C->T transitions at methyl-CpG-associated dipyrimidine sites in mouse skin epidermis more frequently than UVB. Mutagenesis, 18(6), 511–519. https://doi.org/10.1093/mutage/geg030
15. di Magliano, M. P., & Logsdon, C. D. (2013). Roles for KRAS in Pancreatic Tumor Development and Progression. Gastroenterology, 144(6), 1220–1229. https://doi.org/10.1053/j.gastro.2013.01.071
16. Salvadores, M., Mas-Ponte, D., & Supek, F. (2019). Passenger mutations accurately classify human tumors. PLOS Computational Biology, 15(4), e1006953. https://doi.org/10.1371/journal.pcbi.1006953
17. Schuster-Böckler, B., & Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature, 488(7412), 504–507. https://doi.org/10.1038/nature11273
18. Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J. A., & Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature, 518(7539), 360–364. https://doi.org/10.1038/nature14221
19. Fizazi, K., Greco, F. A., Pavlidis, N., Daugaard, G., Oien, K., & Pentheroudakis, G. (2015). Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology, 26, v133–v138. https://doi.org/10.1093/annonc/mdv305
20. Hyman, D. M., Piha-Paul, S. A., Won, H., Rodon, J., Saura, C., Shapiro, G. I., Juric, D., Quinn, D. I., Moreno, V., Doger, B., Mayer, I. A., Boni, V., Calvo, E., Loi, S., Lockhart, A. C., Erinjeri, J. P., Scaltriti, M., Ulaner, G. A., Patel, J., … Solit, D. B. (2018). HER kinase inhibition in patients with HER2- and HER3-mutant cancers. Nature, 554(7691), 189–194. https://doi.org/10.1038/nature25475
21. Jiang, Y., Marinescu, V. D., Xie, Y., Jarvius, M., Maturi, N. P., Haglund, C., Olofsson, S., Lindberg, N., Olofsson, T., Leijonmarck, C., Hesselager, G., Alafuzoff, I., Fryknäs, M., Larsson, R., Nelander, S., & Uhrbom, L. (2017). Glioblastoma Cell Malignancy and Drug Sensitivity Are Affected by the Cell of Origin. Cell Reports, 18(4), 977–990. https://doi.org/10.1016/j.celrep.2017.01.003
22. Human reference epigenome (Roadmap Epigenomics Project) (04.07.2020) http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18927
23. Super enhancer database webpage (06.07.2020): http://www.licpathway.net/sedb/data-browse.php
24. Official R webpage (06.07.2020): https://www.r-project.org/
25. "GenomicRanges" package vignette webpage (01.07.2020): https://bioconductor.org/packages/release/bioc/vignettes/GenomicRanges/inst/doc/GenomicRangesIntroduction.html
26. Yao, F., Coquery, J., & Lê Cao, K.-A. (2012). Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics, 13(1), 24. https://doi.org/10.1186/1471-2105-13-24
27. "factoMineR" package official webpage (04.07.2020): http://factominer.free.fr/

28. Breiman, L. (2001). Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

29. "GenomicFeatures" package vignette webpage (01.07.2020): https://bioconductor.org/packages/release/bioc/html/GenomicFeatures.html

30. Official UCSC genome browser webpage (06.07.2020): http://genome.ucsc.edu/

31. Official "BioMart" package webpage (05.07.2020): http://www.biomart.org/

32. Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J. M., Juul, R. I., Lin, Z., Feuerbach, L., Sabarinathan, R., Madsen, T., Kim, J., Mularoni, L., Shuai, S., Lanzós, A., … Herrmann, C. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature, 578(7793), 102–111. https://doi.org/10.1038/s41586-020-1965-x

33. "BSgenome" package vignette (03.07.2020): https://www.bioconductor.org/packages/release/bioc/html/BSgenome.html

34. Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., de Ridder, J., van Herpen, C., Lolkema, M. P., Steeghs, N., Getz, G., Morris, Q., & Stein, L. D. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nature Communications, 11(1). https://doi.org/10.1038/s41467-019-13825-8

35. Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., Grimes, B., Krysan, K., Yu, M., Wang, W., Alber, F., Sun, F., Dubinett, S. M., Li, W., & Zhou, X. J. (2017). CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biology, 18(1). https://doi.org/10.1186/s13059-017-1191-5

36. Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., … Stratton, M. R. (2009). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature, 463(7278), 191–196. https://doi.org/10.1038/nature08658

37. Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719–724. https://doi.org/10.1038/nature07943

Curriculum Vitae

Osobni podatci

Ime i prezime: Marin Volarić

Datum rođenja: 25.10.1996

Mjesto rođenja: Karlovac

Obrazovanje

2003–2011	Osnovna škola Grabrik, Karlovac
2011–2015	Prirodoslovno-matematička gimnazija. Karlovac
2015–2018	Preddiplomski studij Molekularne biologije. Prirodoslovno-matematički fakultet u Zagrebu

Sudjelovanja u popularizaciji znanosti

2018.	Noć Biologije

Postignuća

2013. Prvo mjesto na državnom natjecanju iz biologije u znanju