Towards building a predictor for response to cancer treatment using tumor mutational signatures.

Stefano Pellegrini
Faculty of SCIENCE
University of Copenhagen
(MLQ211)

I. Abstract

Mutational signatures are characteristic patterns of somatic mutations that can reflect the presence of particular DNA damage, which may correlate with the activity or inactivity of biological pathways in cancer cells. As a consequence, they can potentially be used in a clinical setting to support cancer diagnosis or as predictor for the response to cancer treatment.

Cancer patients with exhausted treatment options undergo genomic screening to enroll in Fase1 clinical trials and thus could benefit from mutational signature analysis.

In this study, we applied mutational signature analysis on 111 tumor samples to explore the signature landscape of Fase1 clinical trial patients with various cancer subtypes, while a subset of those samples was used to build machine learning models that predict response to PARP inhibitors.

We were able to stratify tumors according to the mutational processes underlying their mutational signatures and their genomic profiles, whereas, due to limited size of the cohort, we did not succeed to accurately predict the response to cancer treatment. Nevertheless, we hope that this study can help to form the basis for the clinical implementation of mutational signatures in precision medicine.

II. Introduction

It is only in the past decade, and with the advent of next generation sequencing (NGS), that molecular testing became a common procedure to support classical methods for cancer diagnostic and treatment guidance [32]. Due to this knowledge, nowadays cancer care has improved significantly and most patients receive a combinations of treatments, such as surgery with chemotherapy and/or radiation therapy [1]. Moreover, cancer treatments options also includes immunotherapy, targeted therapy, or hormone therapy [1]. Patients with exhausted treatment options may decide to undergo cancer clinical trials, where efficacy and safety of new approaches for cancer treatment are compared to current treatments [1]. Not all patients can enroll in a clinical trial and specific requirements must be met. In this regard, precision medicine can provide support to the selection of candidates, as well as the choice of specific treatments, based on the genetic understanding of their diseases [1].

Recently, it has been shown that known mutagenic processes, arising through endogenous and exogenous factors, leave a distinctive pattern of mutations, which have been termed mutational signatures [5]. Also, it is known that patients with tumor cells lacking certain functional pathways, such as homologous recombination repair (HRR) and DNA mismatch repair (MMR), may benefit from specific molecularly targeted drugs [32]. The correct functioning of the HRR pathway depends on the presence of functional tumor suppressor genes such as BRCA1 and BRCA2 genes [31] [32]. Harmful mutations in these genes may produce unrepaired DNA damage that can result in tumorigenesis [9]. PRPA inhibitors, a group of pharmacological inhibitors of the enzyme poly ADP ribose polymerase (PARP), have shown promising results in treating tumors with defective HRR due to mutations in BRCA1/2 genes, by decreasing the DNA damage response and causing the tumor cells death [32]. The DNA mismatch repair (MMR), which is another important DNA repair mechanism, is a system for detecting and repairing errors arising during DNA replication [31]. Impaired MMR leads to a hypermutable phenotype called microsatellite instability (MSI) and the accumulation of mutations in cancer-related genes, which may lead to cancer [31]. Interestingly, inhibitors of the programmed death 1 (PD1) immune checkpoint have shown to be effective on colorectal and pancreatic cancers with MMR deficiency [32]. Therefore, the detection of mutational signatures patterns associated with deficiency in HRR and MMR can support the identification of patients who would benefit from these therapies. Moreover, different other therapies have shown to be more effective on tumors presenting specific genomic profiles. Furthermore, the detection of characteristic DNA damage may also reflect the presence of particular cancers. Consequently the mutational signatures offer a general novel approach to support cancer diagnosis and treatment guidance [32].

This preclinical study was intended to investigate the application of mutational signatures analysis as decision making support for the selection of potential candidates and treatments for Fase1 clinical trials. These patients, who did not respond to standard chemotherapy, may benefit from alternative treatments whose efficacy is correlated with the presence of specific genomic profiles in tumor cancer cells [32]. We explored the potential use of the mutational signatures analysis in discovering such profiles in a Fase1 clinical trial and, ultimately, we attempted to build a model for the prediction of the response to olaparib, a PARPi drug for cancer treatment.

III. Materials and methods

A. Mutational signatures analysis

1) Data: The input used for the exploratory mutational signatures analysis was variant call format (VCF) files containing somatic variant calls from 111 tumor samples (Fig. 11 on Supplementary Materials) from patients enrolled in Fase1 clinical trials. and relative patients metadata. Samples were collected at Rigshospitalet and underwent whole exomes sequencing (WES) followed by *in silico* processing at the Center of Genomic Medicine (Rigshospitalet), using a genome analysis toolkit (GATK) based pipeline [21]. The pipeline, as standard practice, subtracts germline calls from paired blood samples subtracted from somatic calls. The relative patients metadata included the samples sequencing date, tissue tumor mutational burden (TMB) and patients age and gender. The TMB corresponds to the total number of nonsynonymous mutations in the protein-coding regions of genes in a tumor and it is studied as a biomarker for immunotherapy [6]. A small subset of the cohort, which have been annotated by independent methods in the laboratory, was used as validation to assess proper functioning of the the mutational signature analysis.

2) Mutational signatures and COSMIC: The mutational signatures are a characteristic combination of somatic mutations arising from specific mutational processes [5]. The catalogue of somatic mutations in cancer (COSMIC) is the world's largest curated database of somatic mutation information relating to human cancers [13]. COSMIC includes a collection of three different variant classes of mutational signatures: single base substitution (SBS), doublet base substitution (DBS) and small insertion and deletion (ID) signatures. The set of mutational signatures has been extracted using SigProfiler, from 4,645 whole-genome and 19,184 exome sequences of most types of cancer [5]. The current analysis used the exome SBS COSMIC v3 set [2].

The mutational signatures analysis was mainly based on the MutationalPattern R/Bioconductor package, which allows for the characterization and visualization of mutational patterns in VCF data [7]. We used the human reference genome hg19 [3] to retrieve the sequence context of the SBS and MutationalPattern allowed us to construct a mutation matrix with the 96 trinucleotide changes. Then, we computed the cosine similarity between the SBS COSMIC signatures and the samples mutation matrix. The cosine similarity is measured by the cosine of the angle between two non-zero vectors and determines if the two vectors are oriented in the same direction [18]. The package computes the cosine similarity between two mutation profiles A and B with n mutation types as follows:

$$\mathrm{sim}(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

[7]. Also, a *de novo* mutational signatures extraction from the mutation count matrix was performed using the non-negative matrix factorization (NMF) method introduced by Stratton et al. [24]. We used the MutationalPattern estimation of the cophenetic correlation coefficient (Fig 13 on Supplementary Materials) of the *de novo* reconstructed profiles to obtain the optimal number of signatures to be extracted. Since a common approach for deciding the rank value is to choose the smallest rank for which the cophenetic correlation coefficient starts decreasing [14], we choose to extract 4 signatures. To link mutational processes with mutational patterns, the *de novo* extracted signatures were compared to the COSMIC signatures, and the relative contribution of the *de novo* extracted signatures to the reconstructed profiles was analysed.

3) Data mining: In order to visualize patterns in the data, I used principal component analysis (PCA) and hierarchical clustering. PCA is a mathematical method used to represent, or project, high dimensional data into low dimensional space [26], Hierarchical clustering is a method used to group observations into clusters based on the distance between data points. A common metric for hierarchical clustering is the Euclidean distance, which is defined as the straight-line distance between two points in Euclidean space [33]. The pheatmap [20] R package was used to build the heatmaps produced in this study.

B. Model building

1) Predictor and target variables: The input used for building the models is a subset of the cohort used for the mutational signatures profiles analysis. It is composed of 15 cancer samples, most of them obtained from reproductive organs (Fig. 12 on Supplementary Materials), whose response to olaparib (PARPi cancer drug) was available. The treatment response has been annotated into four categories: stable disease (SD), partial response (PD), complete response (CR) and progressive disease (PD). The models take as input the pairwise cosine similarity between the mutation profile of the samples and the 65 COSMIC signatures, as well as the tissue TMB, and the patient's gender and age. We chose not to include the cancer type as a model feature because the samples are distributed across its classes in an imbalanced setting. Even if in the literature several methods are proposed to deal with this problem [4] [19], due to the number of features far exceeding the number of samples, we decided to omit this categorical variable to avoid a further reduction of the already limited model generalization ability. As already mentioned, the target variable that the model attempts to predict is the response of the cancer treatment. We considered SD, PR, and CR as positive responses and PD as no response, to turn the multinomial classification problem into a binary classification one. The binary encoding classification is used to encode the positive response into ones and the no response into zeros.

2) Preprocessing: In the preprocessing phase we centered and scaled the predictor variables prior to cross validation, so to obtain features with zero mean and unit standard deviation. This process, called standardization, is a common preprocessing technique used in several machine learning methods. Particularly, it is necessary when the input variables are not on the same scale, such as the tissues TMB, the patients

age and the cosine similarity predictor variables used in this experimental study. After standardization we performed PCA and multidimensional scaling of the data to two dimensions. Ideally, the validation data can be used to find the optimal dimensions for the dimensionality reduction technique, but in our case the choice was more arbitrary and it was simply guided by the possibility of plotting the data in two dimensional space. Nevertheless, PCA is not only useful for the visualization of multidimensional data, but it is a powerful preprocessing technique able to extract relevant information and reveal patterns that might otherwise be hidden in noisy and confusing datasets [28].

3) Logistic regression and random forest: Two supervised learning methods, logistic regression and random forest, were used to build the model. The logistic regression is a generalized linear regression model used to predict a binary dependent variable [29]. We used our implementation of the statistical model using stochastic gradient descent algorithm to estimate the optimal parameters. Random forest is a class of ensemble learning algorithms that consist of a large number of decision trees operating as a committee [27]. Each decision tree is trained on a dataset that is obtained by random sampling with replacement from the original training set [27]. To make a prediction, for a certain observation, each decision tree predicts a class and the class with the majority of votes become the model's final decision [27]. We used the scikit-learn [25] implementation of the random forest with default parameters, and to evaluate the quality of the classifier, due to the element of randomness, we performed a hundred independent predictions and we computed the average accuracy across all iterations.

4) Validation technique: In this classification task, the limited size of the cohort convinced us to choose not to split the dataset into test, train and validation sets, so, to use more data points for the training process and obtain better parameters. Since the evaluation of a model accuracy on the training set would result in a biased quality score, we used the scikit-learn [25] implementation of the leave one-out cross validation (LOO-CV), to estimate the model performance on unseen data. LOO-CV is a special case of K-fold cross validation, with the number of folds K equal to N, the number of samples present in the data [23]. N cross validation iterations are performed and, at each round, N 1 data points are used for training the model, while 1 point is used for the prediction. Finally, the average error across all iterations is computed. Compared to other methods, LOO-CV allows to obtain a less biased and optimistic assessment of the performance of a model using limited sample size, but due to its computational complexity it is not a recommended method for large samples [23].

Also, it is worth mentioning that the the lack of training-test split (holdout method [22]), due to limited sample size, convinced us not to perform features selection nor hyperparameter tuning. In fact, as already mentioned, we believe that these important steps in classifier development should be performed on a separated dataset, and the model generalization ability should be always evaluated on previously unseen data.

5) Performance evaluation: To evaluate the model accuracy, we made two ROC curves showing the performance of the logistic regression and random forest models on data preprocessed by PCA and data in full dimensional space. A ROC curve shows the performance of a binary classifier, by comparing sensitivity versus specificity for every possible classification threshold [12]. The sensitivity, or true positive rate (TPR), or recall, measures the proportion between the number of correct positive predictions and the total number of positives ones:

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{2}$$

[17]. The specificity, or true negative rate (TNR), measures the proportion between the number of correct negative predictions divided by the total number of negatives ones:

$$TNR = \frac{TN}{FP + TN} = \frac{TN}{N} \tag{3}$$

[17]. The fallout (1 - specificity), or false positive rate (FPR), measures the proportion between the number of incorrect positive predictions and the total number of negatives ones:

$$FPR = \frac{FP}{FP + TN} = 1 - TNR \tag{4}$$

[17]. The precision, or positive predicted values (PPV), measures the proportion between the number of correct positive predictions and the total number of positive ones:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

[17]. The area under the ROC curve (AUC) is a derived summary measure for the accuracy of a binary classifier [12]. While the standard definition of accuracy is simply the proportion between the number of correct predictions and the total number of predictions, the AUC can be interpreted as the probability that a randomly selected positive responding sample is ranked as more likely to have a positive response than a randomly selected sample which had no response [12]. Lastly, a diagonal line can be drawn to show the performance of a random classifier.

IV. Results

A. The mutational signature landscape of Fase1 clinical trial patients

1) Original mutation profiles: COSMIC SBS signatures are classified based on the 6 base substitutions C>A, C>G, C>T, T>A, T>C and T>G and the flanking 5' and 3' bases [5]. The frequency of these substitutions are considered in trinucleotide context, generating 96 contexts that are used for the SBS signatures classification, and represent the mutational signatures profile [5]. The mutation profiles of the most relevant COSMIC signatures for this analysis is shown in Fig. 1 and we will now briefly introduce their proposed aetiology.

SBS1, SBS5 and SBS40 are associate with normal mutational decay due to spontaneous deamination of methylated nucleotides, a natural consequence of aging [10]. SBS2 (not

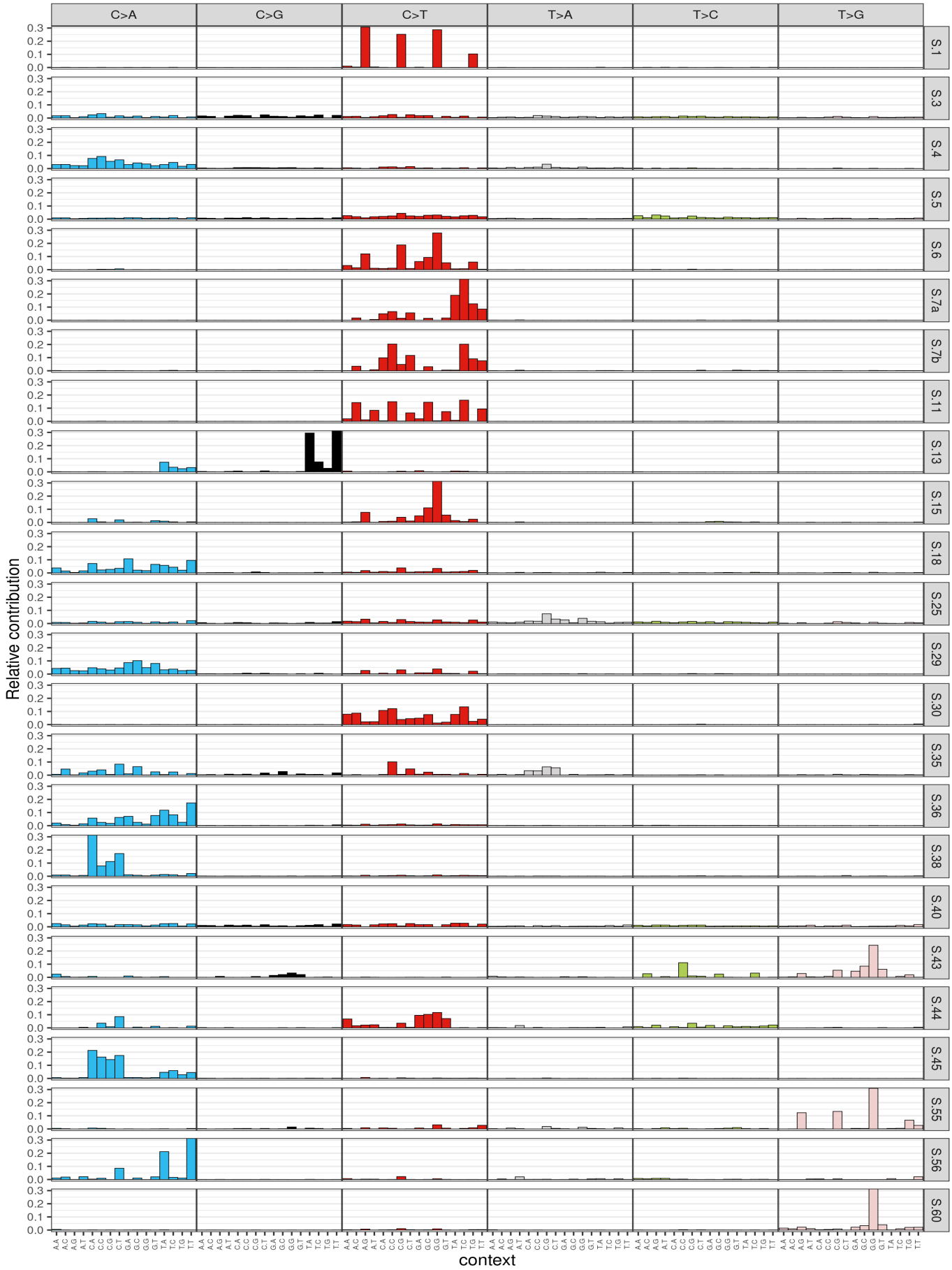

Fig. 1. **Mutation profiles of the SBS COSMIC [13] signatures most relevant to the current analysis.** This subset was chosen because it includes the signatures that better described the mutational patterns found in the data. The plot shows the frequency of the 6 base substitutions C>A, C>G, C>T, T>A, T>C and T>G, each in 16 nucleotides contexts.

shown in Fig. 1) and SBS13 are usually found in the same samples and are associated with activation of AID/APOBEC cytidine deaminases which might be caused by viral infection or tissue inflammation [10]. SBS3 is associated with defective homologous recombination-based (HRD) DNA damage repair and it is associated with germline and somatic BRCA1 and BRCA2 mutations in breast, pancreatic, and ovarian cancers [10]. SBS4 is associated with tobacco smoking and SBS29 with tobacco chewing [10]. SBS6, SBS15, SBS20 and SBS44 are associated with defective DNA mismatch repair (MMR-D) and subsequent microsatellite instability (MSI) [10]. SBS7a, SBS7b and with minor evidence SBS38, are associated with exposure to ultraviolet light in melanoma cancer [10]. SBS18 and SBS36 are associated with reactive oxygen species induced DNA damage [10]. SBS30 is associated with base excision repair [10], while SBS25, SBS31 and SBS35 are associated with prior chemotherapy treatment [10]. SBS45, SBS43, SBS55 and SBS60 are possible artefacts introduced during sequencing, while SBS8 (not shown in the Fig. 1) and SBS11 aetiologies are unknown [10]. SBS3, SBS5 and SBS8 (also SBS40 and SBS25 to some extent) are flat signatures and due to their mutation profiles are mathematically challenging to deconvolute [5].

The relative contribution of the 6 base substitutions, termed mutational spectrum, was extracted from the 13 types of cancer analysed in this study (Fig. 2). Given that ovarian and prostate cancers represented the majority of cancers in our samples (Fig. 11), it was expected that they also contributed to the largest total mutation burden. Moreover, lung and colon cancers shown the most characteristic mutational spectrum. In fact, lung cancer was the only tumor that shown a high frequency of C>A base substitution. While colon cancer was the only tumor showing a frequency of C>T at CpG sites larger than the frequency of C>T at other sites.

In order to associate the samples mutation profiles with known mutational signatures, we used the single base substitution signatures list COSMIC set, introduced by Alexandrov et al. [5]. We calculated the pairwise cosine similarity (Materials and Methods) between the COSMIC signatures and the samples mutation profiles. It is possible to observe that the average cosine similarity is low on the signatures located on the right side and it incrementally increases moving on the opposite side of the plot. In fact, on the left side, we can observe the presence of a cluster of four flat signatures (SBS3, SBS5, SBS25 and SBS40) presenting the largest cosine similarity within the set. Another small cluster is present in the upper side of the plot, including three signatures (SBS43, SBS55 and SBS60) that are known to be possible artefacts introduced during sequencing [5].

Fig. 3 shows the PCA plot of the pairwise cosine similarity between the COSMIC signatures and the mutation profiles of a small subset of seven annotated samples. These samples, annotated by independent methods in the laboratory, are used as a form of validation for the mutational signatures analysis. It is possible to observe that the samples clustered according to their annotation (HRD, high TMB, MSI profile and BRCA

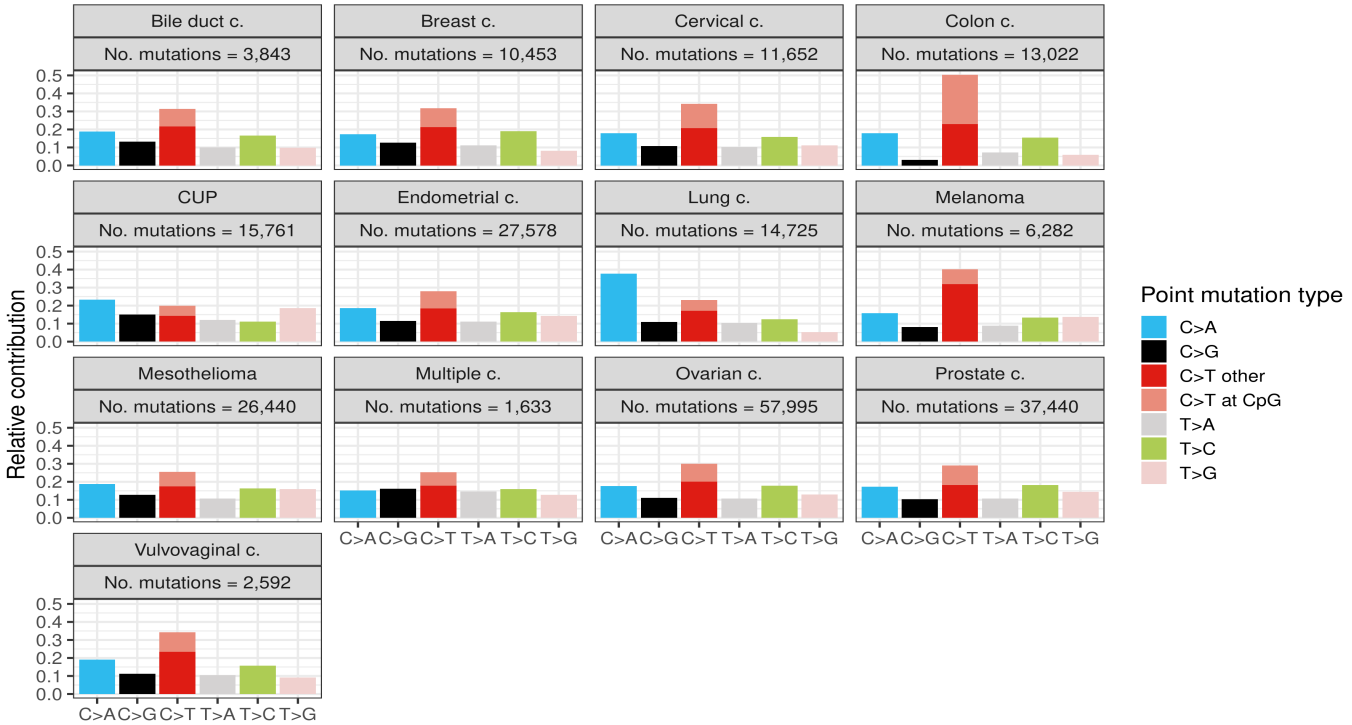

Fig. 2. **Samples mutational spectrum [7] divided by cancer type.** The plot shows the total number of nonsynonymous somatic mutations and the frequency of the 6 base substitutions C>A, C>G, C>T, T>A, T>C and T>G found in the different types of cancer analysed in this study. The plot makes a distinction between C>T at CpG sites and other sites.

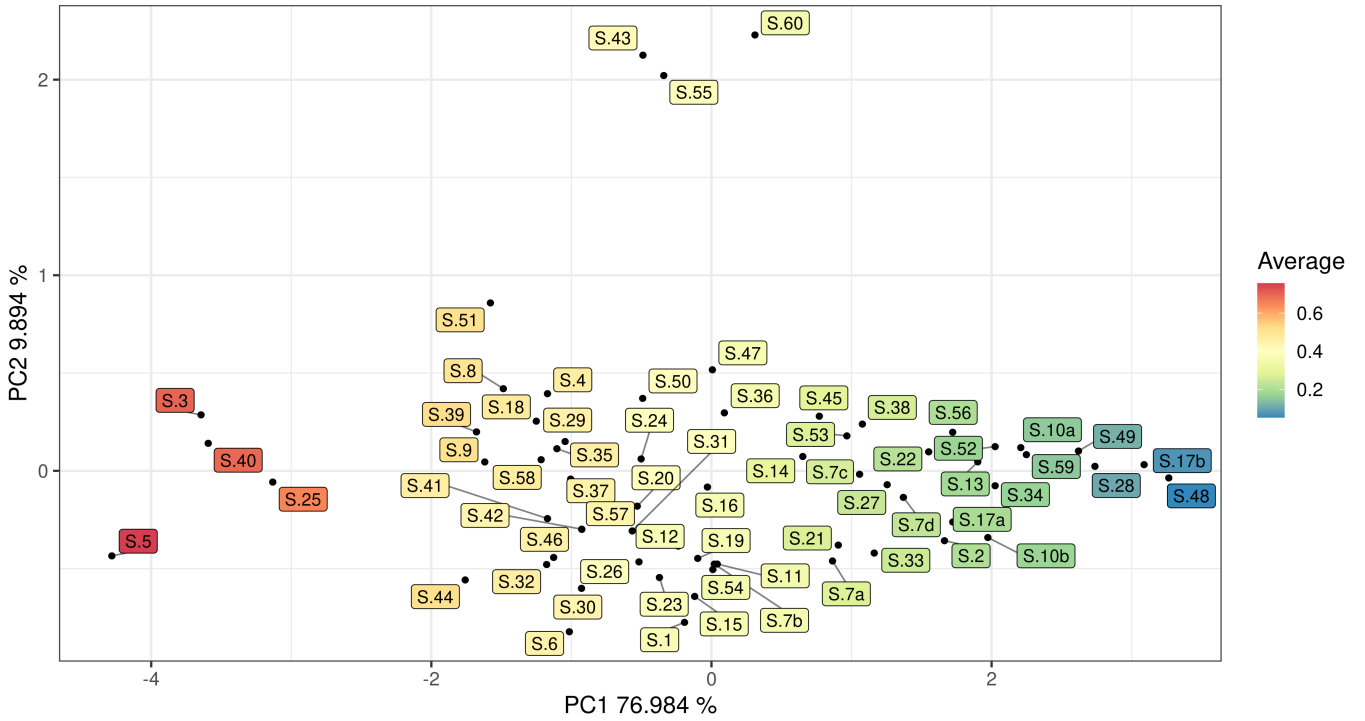

Fig. 3. **COSMIC signatures PCA plot: pairwise cosine similarity between COSMIC signatures and samples mutation profiles.** The COSMIC mutational signatures are projected in the space defined by the two principal components obtained by linear combination of the samples profiles. The signatures are colored by their average cosine similarity across all samples. The variance captured by the fist and second principal component are, respectively, 77% and 9.9%.

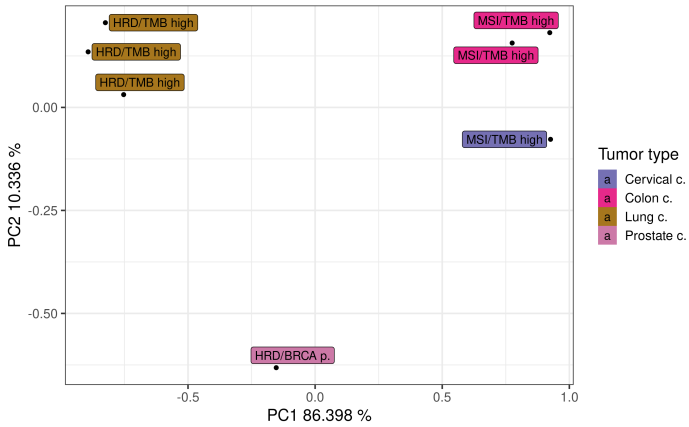

Fig. 4. **Annotated samples PCA plot: pairwise cosine similarity between COSMIC signatures and annotated samples mutation profiles.** Seven annotated samples, from four different tissues, are projected in the space defined by the two principal components obtained by linear combination of the COSMIC mutational signatures. These samples, annotated by independent methods in the laboratory, are used to validate the mutational signatures analysis. *HRD* refers to homologous recombination deficiency, *TMB high* refers to a high tumor mutational burden, *MSI* refers to microsatellite instability due to DNA mismatch repair, and *BRCA p.* refers to the presence of mutations in BRCA1 or BRCA2 tumour suppressor genes [16]. The first and the second principal components captured, respectively, 86.4% and 10.3% of the original variance in the data.

profile) and cancer type. Also, Fig. 6 shows that the expected signature for lung cancer (SBS4, which is associated with tobacco smoking [10]) was detected.

In Fig. 5, that shows the pairwise cosine similarity between all samples mutation profiles and the COSMIC signatures, one can observe that several samples with older sequencing dates clustered separately. Since different sequencing machines were used during the years, it indicates that the utilization of dated DNA sequencer may have an influence on the samples mutation profiles resulting from a mutational signatures analysis.

Fig. 6 shows the heatmap of the pairwise cosine similarity between COSMIC signatures and samples profiles. It is possible to observe that, except for one cluster, the samples within the clusters are not homogeneous with respect to their cancer type. Also, the flat signatures are similar to most cancer types and are present in all clusters except the first two, starting from the top. This group includes signatures associated with HRD and mutated BRCA genes (SBS3), patients age (SBS5 and SBS40) and chemotherapy treatment (SBS25). The first cluster contains samples that present a strong similarity to signatures known to be associated with sequencing artefacts (SBS60, SBS43 and SBS55) [10]. The second and the third clusters include samples with old sequencing date. The former doesn't show any strong similarity to any COSMIC signatures, which may be related due to the use of dated sequencing technology. The fourth cluster includes three lung cancers samples and other cancer types. The lung cancers, which was verified to be HRD/TMB high, show a strong similarity with the smoking and HRD signatures (SBS4 and SBS3), while a similarity with the flat signatures is present in all samples of that cluster.The fifth cluster presents a similarity with the signatures

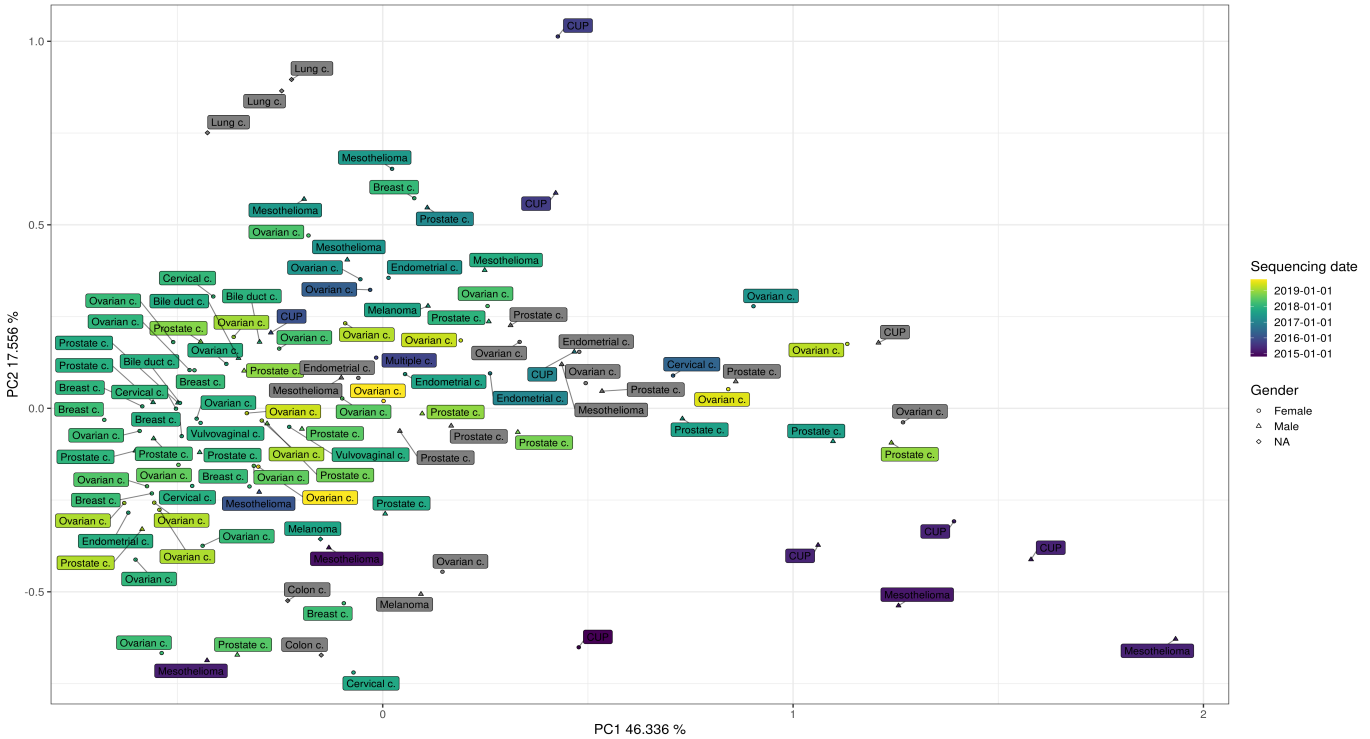

Fig. 5. **Samples (complete cohort) PCA plot: pairwise cosine similarity between COSMIC signatures and all samples mutation profile.** The figure shows the samples in the space defined by the two principal components obtained by linear combination of the COSMIC mutational signatures. The color is mapped to the sequencing date, ranging from dark purple for the oldest (late 2014/2015) sequenced samples and yellow for the newly sequenced ones (2019). Gray color represents samples which sequencing date was not annotated. The cancer type is shown in the labels, while the gender is mapped to the shape of the data points. The variance captured by the fist and second principal component are, respectively, 46.3% and 17.5%.

known to be associated with AID/APOBEC activity (SBS2 and SBS13), while the two melanoma samples of the sixth cluster show a strong similarity to the signatures associated with exposure to ultraviolet light (SBS7a and SBS7b) [10]. The seventh cluster includes a prostate, an ovarian cancer, and the two colon cancer samples. They all have a very high TMB and show a strong similarity to signatures associated to MSI profile (SBS44, SBS6, SBS15), and to an age-related signature (SBS1) [10]. It is worth mentioning that the colon and ovarian samples were verified to be MSI/TMB high, therefore, it provides a confirmation of the detected mutational processes. The eighth and ninth clusters are larger than the others. The smaller one has higher similarity to sequencing artefacts signatures and weaker to clock/HRD signatures. The larger one displays a high burden of clock (SBS5, SBS40), HRD (SBS3) and prior chemotherapy treatment (SBS25). As all the patients in this study were in Fase1, they had all undergone some form of chemotherapy before the tumor sample was collected. Moreover, the prostate sample verified to be HRD/BRCAness is included in this cluster.

2) De novo extraction: After analysing the similarity between the COSMIC signatures and the original samples mutation profiles, we used MutationalPattern [7] to perform a *de novo* mutational signatures extraction and to reconstruct the samples mutation profiles by fitting it to the *de novo* extracted signatures. Four signatures were extracted and their mutation profiles is shown in Fig. 14 (Supplementary Materials).

The similarity between the 4 *de novo* extracted signatures and the COSMIC signatures is shown in the heatmap in Fig. 15 (Supplementary Materials). The extracted signature A presents a strong similarity with signatures associated with MSI profile (SBS6, SBS15 and SBS44) and patient age (SBS1). Signatures B has a strong similarity with the signature associated with tobacco smoking (SBS4). Signature C only presents a weak similarity with signatures associated with sequencing artefact (SBS60, SBS43 and SBS55). Signature D shows a similarity with signatures associated with patient age (SBS5 and SBS40), HRD and BRCA profile (SBS3), and prior chemotherapy treatment (SBS25).

By fitting the samples original mutation matrix to the *de novo* extracted signatures, we obtained the samples reconstructed profiles. The relative contribution of the 4 signatures to the reconstructed profiles is shown in the heatmap of Fig. 16 (Supplementary Materials). Most of the clusters were similar to the ones obtained from the cosine similarity matrix. Thus, with minor differences, the results from analyzing the *de novo* extracted signatures roughly recapitulate the analysis based on original mutation profiles.

B. Building a predictor for response to PARP inhibitors

The models were built using 15 samples from the Fase1 cohort, whose metadata on the response to olaparib, a PARP inhibitor drug, was available. Fig. 7 shows the samples tis-

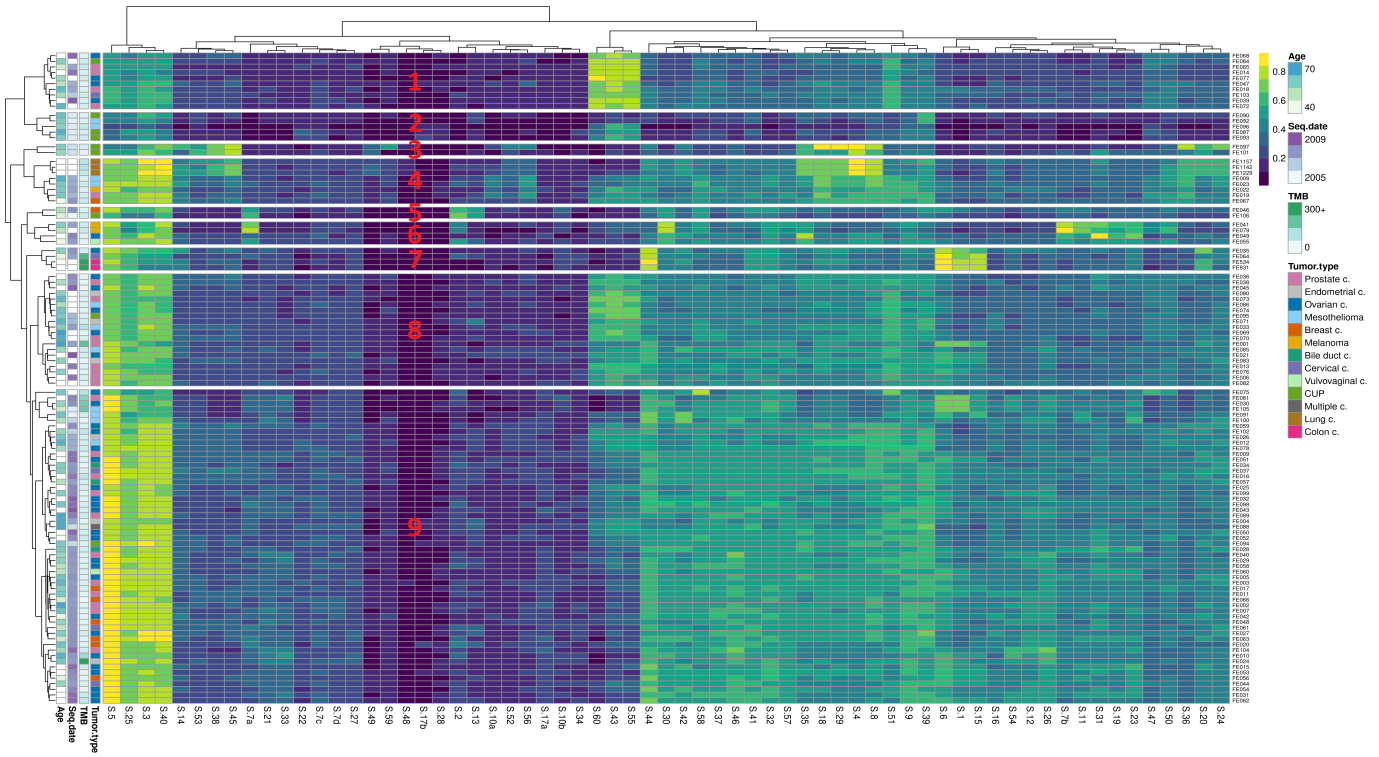

Fig. 6. **Heatmap of pairwise cosine similarity between COSMIC signatures and samples mutation profiles.** The signatures, shown on the bottom of the figures, correspond to the columns, while the samples, shown on the right, correspond to the rows. The color of each cell is mapped to the cosine similarity between a certain signature and sample mutation profile, ranging from dark purple for low similarity, to yellow for high similarity. Two hierarchical trees, located on the top and on the left of the figure, have ordered, respectively, the signatures and the samples according to their within similarity. On the left side of the plot, there are four annotation columns showing samples patient age, sequencing date, tissue TMB and cancer type. The plot was generated with pheatmap [20] R package and cutree function was used to split the samples in 9 clusters, which are numbered from top to bottom to help the reader in the visualization.

sue TMB and patient age and gender, divided by treatment response. One can see that male patients seems to respond better than female to the cancer treatment. Both tissues TMB and patients age show a balanced distribution between the two classes but also a large variability within each class. The TMB is only slightly larger in the samples with positive response, but one outlier is present in each group. Meanwhile, looking at the median value the patients that responded to the treatment seem to be slightly older than the patients that did not respond.

Fig. 17 (Supplementary Materials) shows the correlation between the continuous predictor variables of the input data. It is possible to observe that the TMB and a subset of signatures show a negative correlation with the three signatures associated with sequencing artefact (SBS43, SBS55 and SBS60), this may be related to the lost of information due to sequencing errors, which might have resulted in a reduction of the profiles similarity to a subset of signatures and the overall TMB. Meanwhile, most of the other input variables show a positive correlation and only few show no correlation.

Fig. 8 shows the coefficients of the logistic regression model trained on data not preprocessed by PCA. It is possible to observe that TMB is the most important predictor, and it shows a positive association with the treatment response. This means that a high tissue TMB is considered by the model an indicator of positive response. The second most important predictor

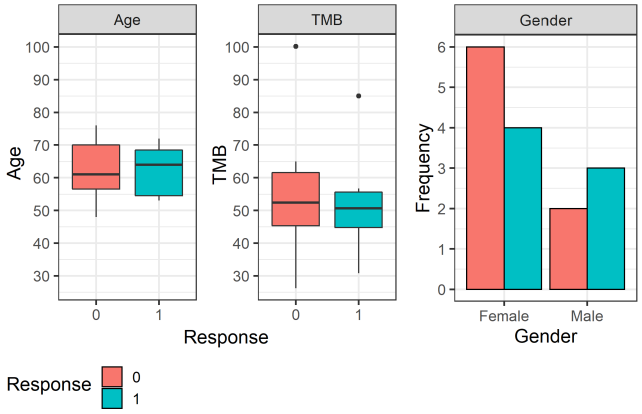

Fig. 7. **Input metadata and response variable.** These three plots show the three metadata features used for the prediction of the cancer treatment response. The metadata is shown in relation to the response target variable. The dark orange color represents samples that did not respond to the treatment, while turquoise represents samples that had a positive response. The box plot on the left shows the patient age, the one in the middle shows the tissue TMB, and the bar graph on the right shows the patient gender.

is the SBS10a which did not appear to be an important signature during the mutational signature exploratory analysis. The third most important predictor is the gender, which has a positive association with the response variable. Since the sex is encoded as 1 for males and 0 for females, it shows a positive

association between being male and having a positive cancer treatment response. This might be due to most cancer types analysed in this study being reproductive cancer and thus sex-specific. The patients age is ranked only ninth and it has a negative association with the response variable.

The performance of both logistic regression and random forest, were tested on the training and test set, for the latter a leave-one-out cross validation was used (LOO-CV). Also, the performance was tested with both, data in full dimensional space and data preprocessed with PCA.

Fig. 9 shows the prediction of the logistic regression on the training data, on the left, and one iteration of the LOO-CV, on the right. The green line represents the decision boundary dividing the samples in the two classes. It is possible to see how a sample, that was correctly classified during training set prediction, is wrongly classified during one iteration of the LOO-CV.

The accuracy of the models, calculated as the ratio between correct predictions and all predictions, is shown in Table 1. We can observe that, as expected, both models have a 100% accuracy on the training set in full space dimensions, while only the random forest has 100% accuracy on the data preprocessed with PCA. On the test set, logistic regression performed better with 60% accuracy on data in full space dimensions and 66.67% accuracy on PCA-preprocessed data. Meanwhile, random forest had an accuracy of 51.33% on the training data non-preprocessed by PCA and 58.93% on PCA-preprocessed data.

Models	Train	LOO-CV	Train PCA	LOO-CV PCA
LR	100%	60%	86.66%	66.67%
RF	100%	51.33%	100%	58.93%

TABLE I
LOGISTIC REGRESSION AND RANDOM FOREST ACCURACY. *The table shows, from left to right, the accuracy of the logistic regression (LR) and random forest (RF) on the training set, on the test set (LOO-CV), on the training set after PCA preprocessing, and on the test set (LOO-CV) after PCA preprocessing.*

Fig. 10 includes two ROC curves, showing the predictive performance of the two models on both, PCA-preprocessed and non preprocessed data. It is possible to observe that the AUCs were larger when the predictions were made on PCA-preprocessed data. However, in both settings the models performances were far from being optimal. As mentioned (Materials and Methods), a ROC curve shows the trade-off between sensitivity and specificity. In the plot on the left (Fig. 10), we can see that the predictive ability of the random forest is good only when the specificity drops below 0.6, therefore, it is very sensitive only when it is poorly specific. Meanwhile the logistic regression shows a more balanced trade-off, its predictive ability is only slightly better than the predictive ability of a random classifier. Overall, the AUC of the two models are very similar with a value of 0.62 for the logistic regression and 0.63 for the random forest. In the plot on the right, we can see that the random forest trade-off between sensitivity and specificity is more balanced than the one observed in the previous ROC curve. But, both models performances are now very similar to the performance of a random classifier.

V. DISCUSSIONS

A. Mutational signatures analysis

We have seen how the mutational signatures can be used to identify common genomic patterns in tumors of Fase1 clinical trial patients. We observed that when we analysed the mutation profiles of the seven verified samples (Fig. 4), we successfully stratified the tumors according to the annotations and cancer types. But, when the analysis was performed using the complete cohort, except for a few tumors, the mutational signatures alone were not able to cluster the samples based on the cancer types. Nevertheless, they provided useful insights in identifying common mutational processes underlying characteristic profiles that may bear relevant clinical value. In fact, exogenous factors, such as tobacco smoking and ultraviolet light exposure were detected in lung cancer and melanoma, and all annotations of the verified samples were assigned correctly. Furthermore, it was possible to stratify the samples (Fig. 6) based on endogenous processes, such as spontaneous deamination of methylated nucleotides, AID/APOBEC cytidine deaminase anti-pathogen response, impaired DNA damage response (DDR) gene function due to HRD and BRCAness, and DNA mismatch repair (MMR) deficiency with consequent MSI. We recall that profiles associated with DNA repair deficiencies, such as HRD/BRCA and MSI profiles, may benefit from treatment with PARP and PD1 inhibitors, respectively. Moreover, recent studies [15] [30] revealed that platinum-based therapy is associated with good prognosis in breast cancer with HRD/BRCA profile, and lethal mutagenesis has been proposed as possible treatment for APOBEC overactivity [32]. Therefore the identification of such profiles can be of great interest for treatment selection and decision making in cancer care, such as candidate selection for clinical trials. Furthermore, the detection of characteristic mutational processes associated to the presence of particular cancers, may be used to support cancer diagnostic procedures.

B. Treatment response prediction

Even if the availability of clinical data has increased year after year, supervised learning methods require annotated samples which are often a relatively scarce and expensive resource [11]. Ideally, the sample size used in a classification task should be large enough to allow the separation of the data into training, validation and test sets (holdout method [22]). Then, the training data could be used for feature selection and learning the model parameters. The validation data could be used for hyperparameter tuning and the previously unseen test data could be used to compute an unbiased estimation of the model generalization performance. In this work, due to limited size of the cohort, we did not use the holdout method to avoid a further reduction of the already limited model learning ability. Moreover, due to limited amount of

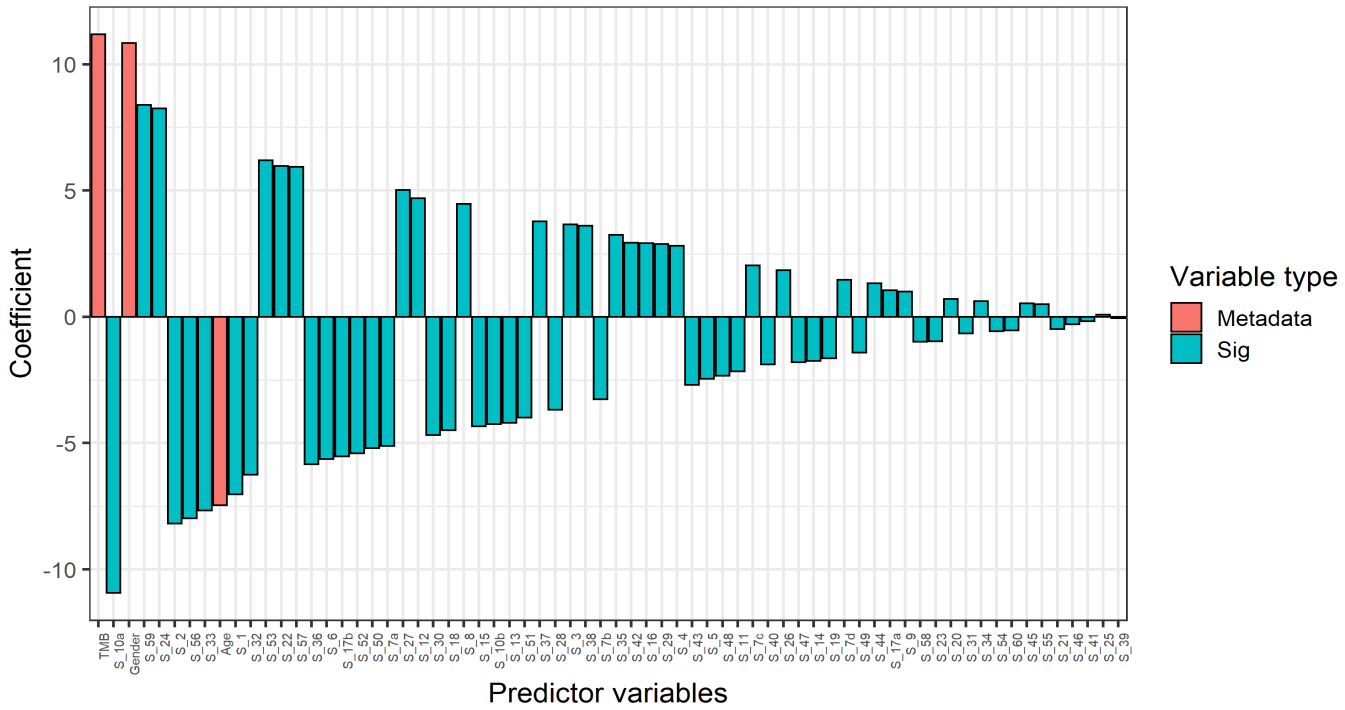

Fig. 8. **Logistic regression coefficients.** The bar graph shows the coefficients of the linear regression model sorted by their absolute value. The predictor variables located on the left are the most important, while the sign of the coefficient shows if the association between the predictor variable and the treatment response is positive or negative.

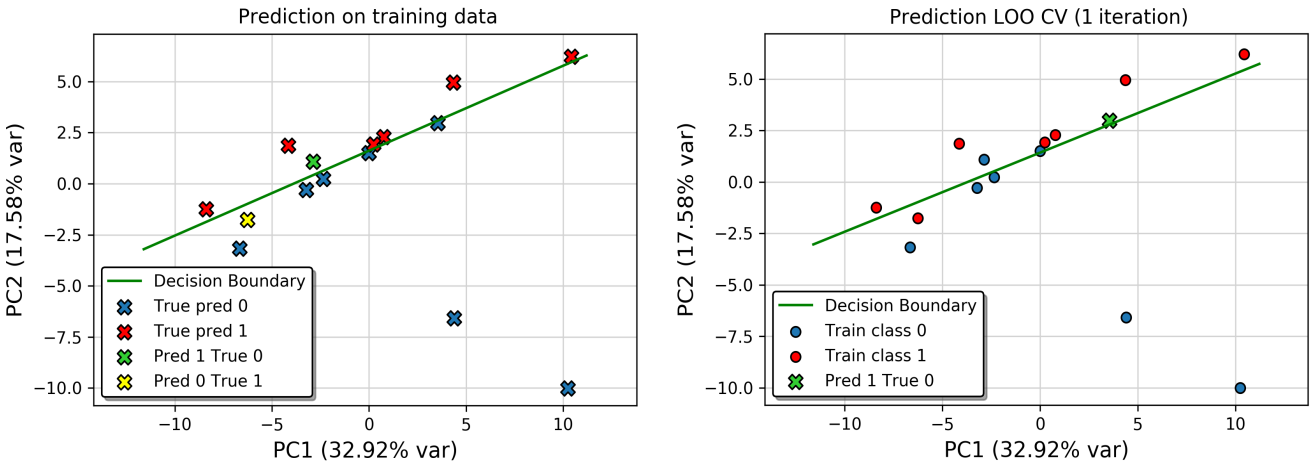

Fig. 9. **Logistic regression prediction after PCA.** The plot on the left shows the logistic regression prediction on the training data, the one on the right shows one iteration of the LOO-CV. PCA and MDS to two dimensions were performed as a preprocessing step. The green line represents the decision boundary. In the plot on the left, training and prediction is performed on all data points, while in the plot on the right, the training is performed on 14 data points and the prediction is performed on the remaining one. The circular shape indicates samples that are used for training only, while the x-shape represents samples whose class was predicted by the model. Yellow and green x-shaped data points represent wrong model predictions. The variance captured by the fist and second principal component are, respectively, 32.9% and 17.6%.

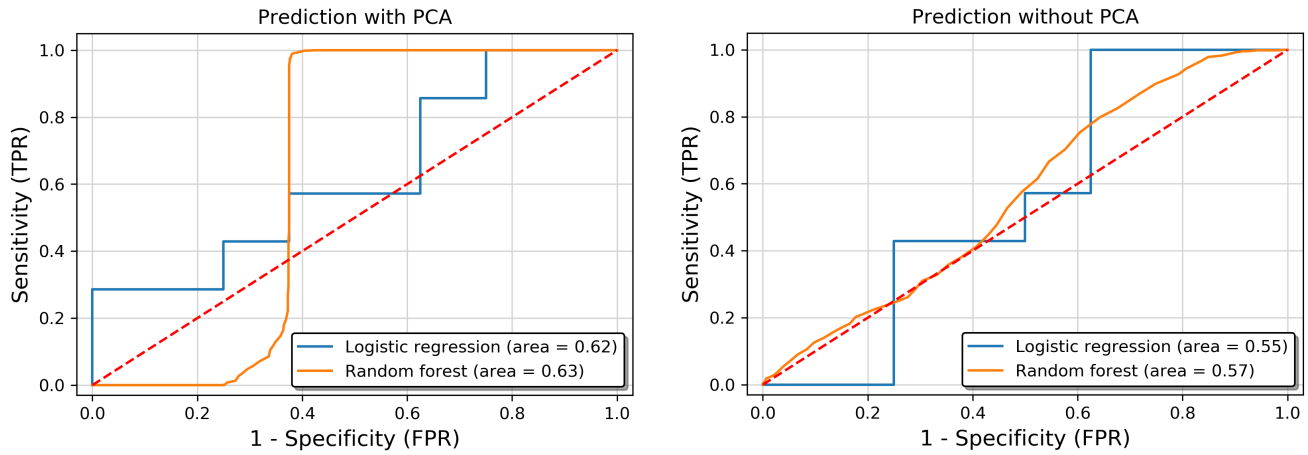

Fig. 10. **ROC curves for prediction on PCA-preprocessed data and non preprocessed data.** The two ROC curves show the test performance of the logistic regression and random forest classifiers, for both, data preprocessed with PCA and data in full dimensional space. The orange and the blue lines represent the performance of the two models, at different classification thresholds. The y-axis shows the true positive rate or sensitivity, while the x-axis shows the false positive rate or fallout (1 - specificity). The area under the ROC curve (AUC) represents a derived summary measure for the accuracy of the models [12]. Lastly, the red dashed diagonal line shows the performance of a random classifier.

time for model optimization, standardization and PCA were performed prior to cross validation. Therefore, in doing so we introduced some bias in the evaluation of the model performance, overestimating its generalization ability. Since both models performed poorly on test data, and given that we introduced some bias overestimating the already poor generalization performance, we can't state that the models built in this study achieved a significant predictive ability on previously unseen data.

Moreover, the exploration of the input revealed potential problems in the data used for building the models. We have seen (Fig. 7) that patients age and tissues TMB distributions were similar between the two groups and shown a great variability within groups. Therefore, due to the small effect size, the small sample size and the large variance, we were expecting a low statistical power from these two variables. Nevertheless, the TMB was considered as the most important predictor for the logistic regression classifier (Fig. 8). One possible explanation is that the presence of outliers in the small dataset influenced the regression analysis, which is very sensitive to data points that differ significantly from other observations [8]. Moreover, the high correlations among predictor variables could have further biased the estimation of the model weights. Therefore, we believe that the logistic regression coefficients (Fig. 8) are unreliable, and that outliers detection and features selection are essential steps to achieve an useful interpretability of the model parameters.

VI. Conclusions

This study has shown that the mutational signatures can be used to stratify Fase1 clinical trial patients according to the mutational processes underlying their mutational signatures and their associated genomic profiles. The identification of such profiles can support cancer diagnosis, treatment selection, and genomic screening of Fase1 clinical trial candidates. Moreover, we built a model for the prediction of the response to PARP inhibitors, but we did not succeed to achieve good generalization performance. We identified the major weak points, such as the limited sample size, and, consequently, the limited model learning ability and lack of separate datasets to perform feature selection and hyperparameter tuning. Therefore, we hope to be able to repeat the study with a larger Fase1 cohort.

The field of mutational signatures analysis is growing rapidly and it offers new promising approaches to support cancer care. Even so, further research is needed to define new reference mutational signatures and to gain insights into their mutational processes. Anyhow, we hope that this work can help to form the basis for the clinical implementation of mutational signatures in precision medicine.

VII. Acknowledgement

I am grateful to Olga Østrup and, especially, to Maria Anna Misiakou for giving me the opportunity to do this project and for providing supervision at the Center for Genomic Medicine at Rigshospitalet. Also, I woud like to thank professor Albin Gustav Sandelin for providing guidance as principal supervisor at the Copenhagen University.

VIII. Supplementary Materials

Additional information in support of this study is included in the next few pages.

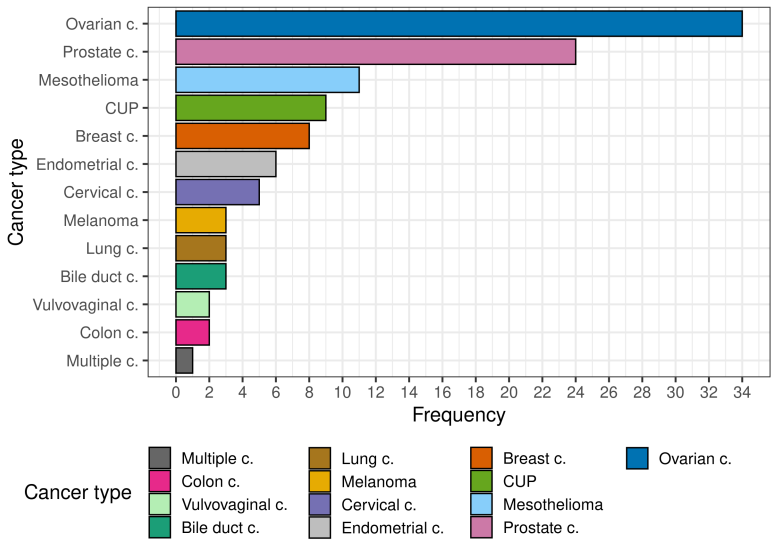

Fig. 11. **Complete cohort cancer types distribution.** The bar chart shows the number of samples for each cancer tissue present in the data used for the mutational signatures exploratory analysis. Carcinoma of unknown primary (CUP) refers to a tumor of unknown origin, while multiple cancer refers to a tumor affecting pancreas, breast and ovarian tissues.

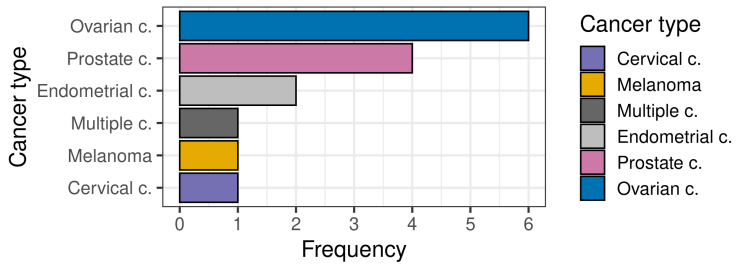

Fig. 12. **Cohort subset cancer types distribution.** The bar chart shows the number of samples, for each cancer tissue, whose response to PARPi drug was available. These samples are used to build the model for the prediction of the cancer treatment response.

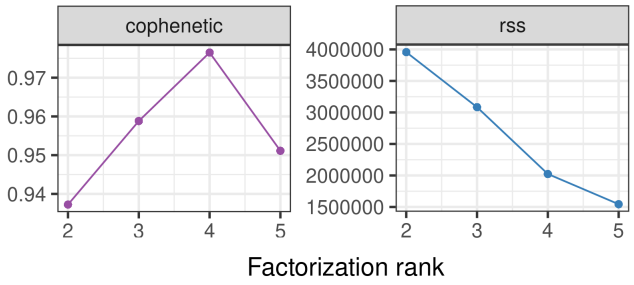

Fig. 13. **NMF de novo rank estimation [7].** The plot shows the estimated cophenetic correlation coefficient of the de novo reconstructed profiles and the residual sum of squares (RSS) between the reconstructed profiles and the original ones, at different rank values. This estimation is used to choose the optimal number of signatures, or rank, to be extracted.

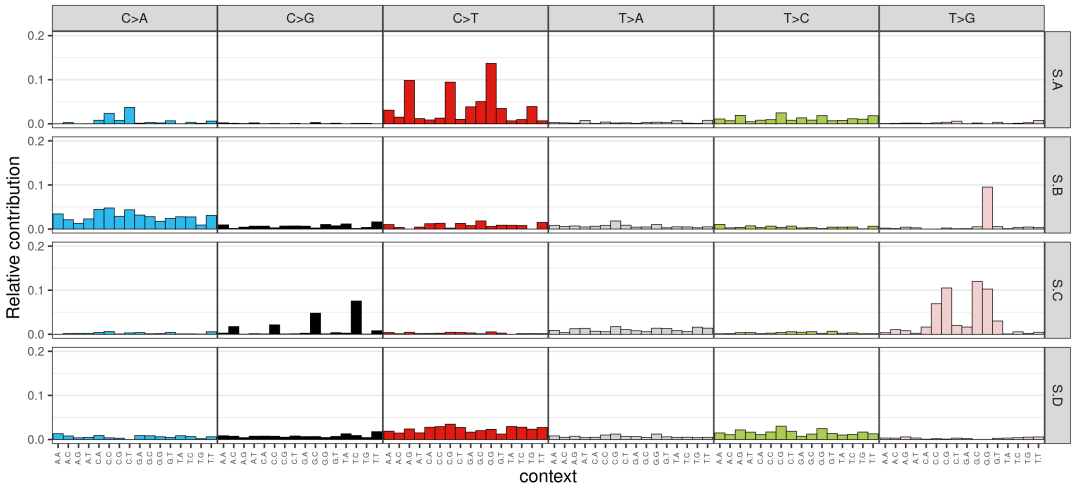

Fig. 14. **De novo extracted signatures mutation profiles [7].** The plot shows the mutation profiles of the four mutational signatures extracted de novo with NMF, or in other words, the frequency of the 6 base substitution used for the signatures classification in the 96 nucleotides contexts.

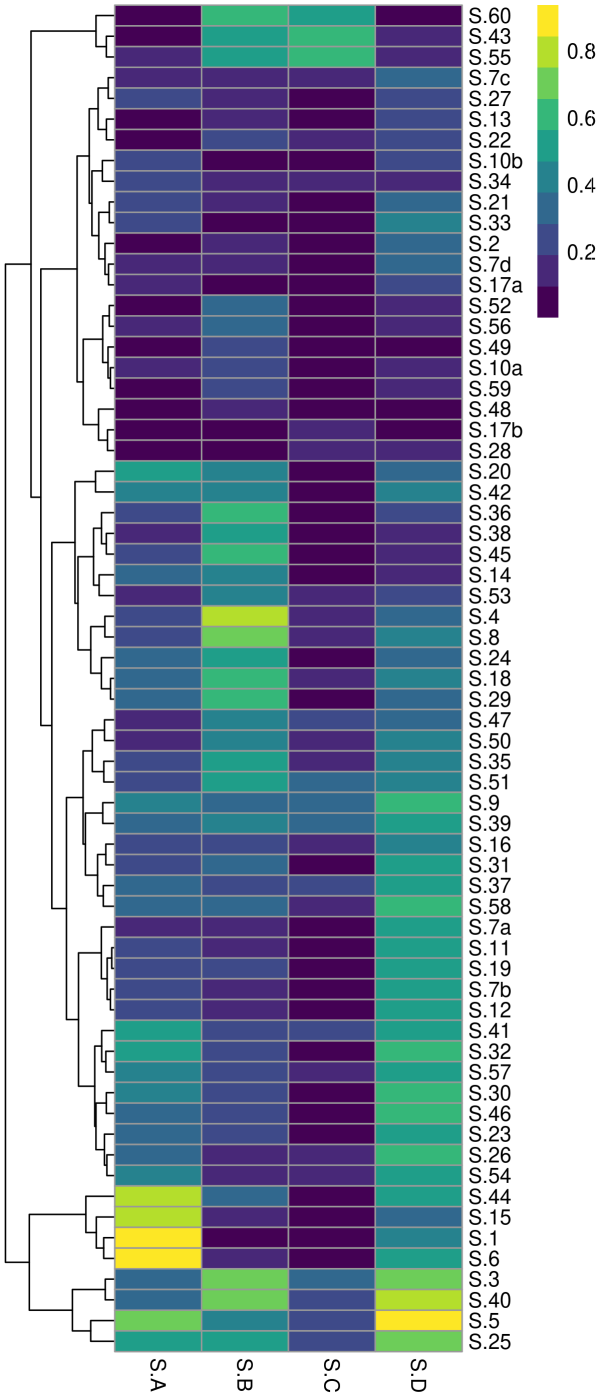

Fig. 15. **Heatmap of pairwise cosine similarity between de novo extracted signatures and COSMIC signatures.** The columns of the heatmap correspond to the de novo extracted signatures, which are shown on the bottom. The rows correspond to the COSMIC signatures and are shown on the right. The color mapping legend is shown on the top right. The hierarchical tree present on the left has ordered the COSMIC signatures according to their similarity. The figure was produced with pheatmap [20].

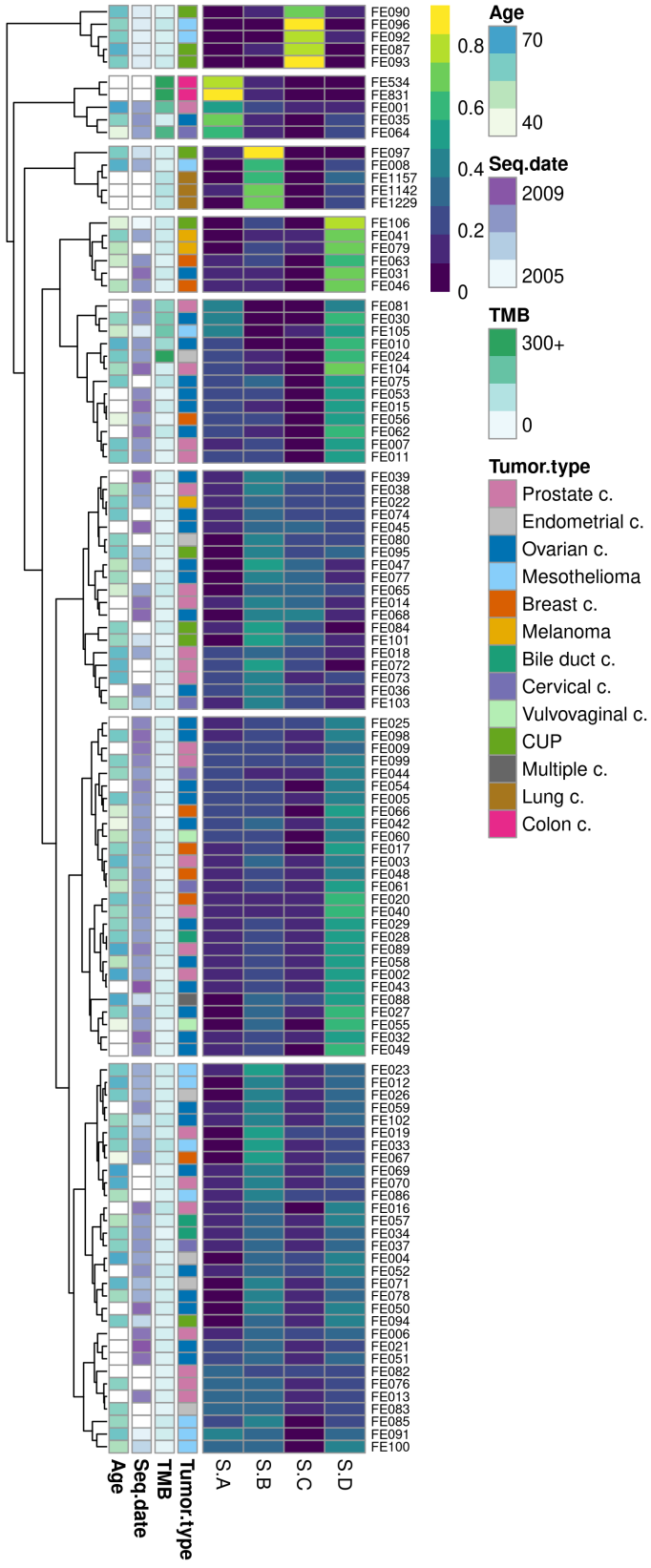

Fig. 16. **Heatmap of de novo extracted signatures relative contribution to reconstructed profiles.** The color of each cell is mapped to the relative contribution of a de novo extracted signature to a sample reconstructed profile, ranging from dark purple for a small contribution, to yellow for a large contribution. The signatures, shown on the bottom of the figure, correspond to the columns, while the samples, shown on the right, correspond to the rows. The hierarchical tree on the left has ordered the samples according to their similarity. The four annotation columns present on the left show the samples patient age, sequencing date, tissue TMB and cancer type. The plot was generated with pheatmap [20] and the samples were split into 8 clusters by the cutree function.

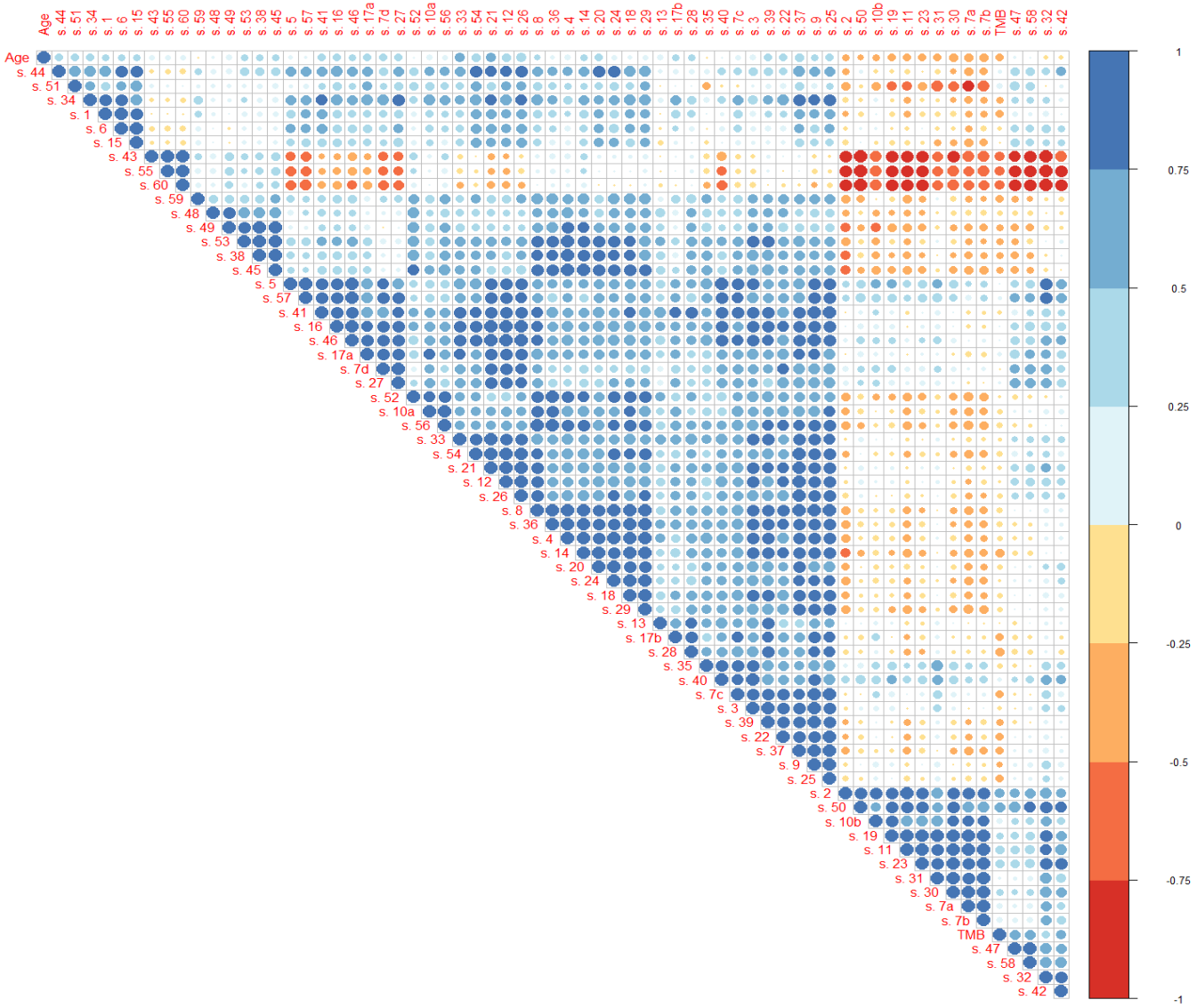

Fig. 17. **Correlogram of predictor continuous variables.** The plot shows the correlation matrix of the continuous input variable used to build the model. In each cell, the spots color and size are mapped to the correlation between two variables. A large spot indicates high correlation, while red and blue colors are used to indicate, respectively, negative and positive correlation.

REFERENCES

[1] Cancer treatments, national cancer institute. https://www.cancer.gov/about-cancer/treatment.

[2] Sigprofiler reference signatures. https://www.synapse.org/#!Synapse:syn12009743.

[3] Genome reference consortium human build 37 (grch37, hg19). National Center for Biotechnology Information, 2009.

[4] E. Adeli, X. Li, D. Kwon, Y. Zhang, and K. Pohl. Logistic regression confined by cardinality-constrained sample and feature selection. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[5] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

[6] L. Berland, S. Heeke, O. Humbert, A. Macocco, E. Long-Mira, S. Lassalle, V. Lespinet-Fabre, S. Lalvée, O. Bordone, C. Cohen, et al. Current views on tumor mutational burden in patients with non-small cell lung cancer treated by immune checkpoint inhibitors. *Journal of thoracic disease*, 11(Suppl 1):S71, 2019.

[7] F. Blokzijl, R. Janssen, R. van Boxtel, and E. Cuppen. Mutationalpatterns: comprehensive genome-wide analysis of mutational processes. *Genome medicine*, 10(1):33, 2018.

[8] F. H. Cutanda. Outliers and robust logistic regression in health sciences. *Revista espanola de salud publica*, 82(6):617–625, 2008.

[9] R. R. da Cunha Colombo Bonadio, R. N. Fogace, V. C. Miranda, and M. d. P. E. Diz. Homologous recombination deficiency in ovarian cancer: a review of its epidemiology and management. *Clinics*, 73, 2018.

[10] A. et al. Cosmic single base substitution signatures. https://cancer.sanger.ac.uk/cosmic/signatures/SBS, 2020.

[11] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo. Predicting sample size required for classification performance. *BMC medical informatics and decision making*, 12(1):8, 2012.

[12] C. M. Florkowski. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83, 2008.

[13] S. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. Teague, P. Futreal, and M. Stratton. The catalogue of somatic mutations in cancer (cosmic). *Current protocols in human genetics*, 57(1):10–11, 2008.

[14] R. Gaujoux and C. Seoighe. A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):367, 2010.

[15] L. Gerratana, V. Fanotto, G. Pelizzari, E. Agostinetto, and F. Puglisi. Do platinum salts fit all triple negative breast cancers? *Cancer treatment reviews*, 48:34–41, 2016.

[16] I. Gorodetska, I. Kozeretska, and A. Dubrovska. Brca genes: the role in genome stability, cancer stemness and therapy resistance. *Journal of Cancer*, 10(9):2109, 2019.

[17] K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[18] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[19] M. Khalilia, S. Chakraborty, and M. Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):51, 2011.

[20] R. Kolde. Pheatmap: pretty heatmaps. *R package version*, 1(2), 2012.

[21] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[22] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004.

[23] A. M. Molinaro, R. Simon, and R. M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[24] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] A. Rago. Exploratory data analysis with principal components analysis. Bioinformatics of High Throughput Analysis, University of Copenhagen, 2020.

[27] A. Sarica, A. Cerasa, and A. Quattrone. Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in aging neuroscience*, 9:329, 2017.

[28] J. Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[29] S. Sperandei. Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1):12–18, 2014.

[30] M. L. Telli, K. M. Timms, J. Reid, B. Hennessy, G. B. Mills, K. C. Jensen, Z. Szallasi, W. T. Barry, E. P. Winer, N. M. Tung, et al. Homologous recombination deficiency (hrd) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clinical cancer research*, 22(15):3764–3773, 2016.

[31] A. Torgovnick and B. Schumacher. Dna repair mechanisms in cancer development and therapy. *Frontiers in genetics*, 6:157, 2015.

[32] A. Van Hoeck, N. H. Tjoonk, R. van Boxtel, and E. Cuppen. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC cancer*, 19(1):1–14, 2019.

[33] Z. Zhang, F. Murtagh, S. Van Poucke, S. Lin, and P. Lan. Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with r. *Annals of translational medicine*, 5(4), 2017.