Transcriptomic Data Analysis of Racial Differences in Colorectal Cancer

RICCARDO GIULIANI, University of Trento
QUINTINO FRANCESCO LOTITO, University of Trento
ALICE MASSACCI, University of Trento
ALESSANDRO SARTORI, University of Trento

Colorectal cancer is recognized worldwide as one of the most common diseases. Motivated by geographical differences in the rates of CRC diagnosis worldwide, in this study we investigate the role of ethnicity in the etiology of CRC.

In this work, we collected a dataset of White and Asian cancer patients, in which every patient is described by the expression level of its genes and by its ethnicity. We performed feature selection to extract the most significative genes when classifying between ethnicities, requiring that they also differ in terms of expression levels between White and Asian patients. From this list, we were able to isolate a set of genes related to cancer by using the database OncoSearch. A large subset of these genes resulted to be part or closely related to the Notch signaling pathway, a pathway involved in many biological responses. By comparing the expression levels of these genes in White and Asian patients, we found that they were upregulated in Western patients (around 1.3 fold upregulation). We also performed network analysis of the gene co-expression networks of the two ethnicities, which confirmed our previous findings. By using CIBERSORTx, we were able to predict the immune cell populations for each patient, and from this analysis we found an increased level of Macrophages in White patients compared to Asian. These results can be further again attributed to the upregulation of the Notch pathway in the White ethnicity.

Our findings suggest differences in Notch signaling among racially-distinct CRC patients that may contribute to the more aggressive clinical behavior of White patients, and motivate further study.

1 INTRODUCTION

Colorectal cancer – the development of cancer from the colon or rectum – is recognized worldwide as one of the most common diseases. It is the second most frequently diagnosed cancer in women (614,000 cases, 9.2% of the total female population) and the third most frequently diagnosed cancer in men (746,000 cases, 10% of the total male population) [1]. As reported in fig. 1, geography appears to be an important factor in the rates of CRC diagnosis worldwide, with a lower incidence of CRC in Africa and Asia and higher in Europe, North America, and Australia.

Besides the environmental component, which is surely correlated with the variations of incidence and mortality of CRC in certain parts of the world, it is now commonly believed that ethnicity also plays an important role in the etiology of CRC. In this direction, a recent study [2], performed on CRC samples from 104 Taiwanese patients, provides a comprehensive characterization of the transcriptomic alterations underlying ethnically specific CRC.

To further investigate the role of ethnicity in the etiology of CRC, we collected transcriptomic data of patients with different ethnicity, focusing only on Asian and White patients. While data about White patients are more common and we were able to collect ~250 samples from TCGA, data about Asian patients are more rare (~130 samples collected, merging data from TCGA and from GEO). The lack of a large amount of accessible data about Asian patients shows again how the impact of ethnicity, especially in Oriental countries, is largely unexplored.

After having collected and validated our dataset, we built a set of tools to analyze ethnicity-based differences in CRC patients' gene expressions.

Authors' addresses: Riccardo Giuliani, University of Trento, riccardo.giuliani@studenti.unitn.it, 213785; Quintino Francesco Lotito, University of Trento, quintino.lotito@studenti.unitn.it, 215032; Alice Massacci, University of Trento, alice.massacci@studenti.unitn.it, 222088; Alessandro Sartori, University of Trento, alessandro.sartori-1@studenti.unitn.it, 215062.

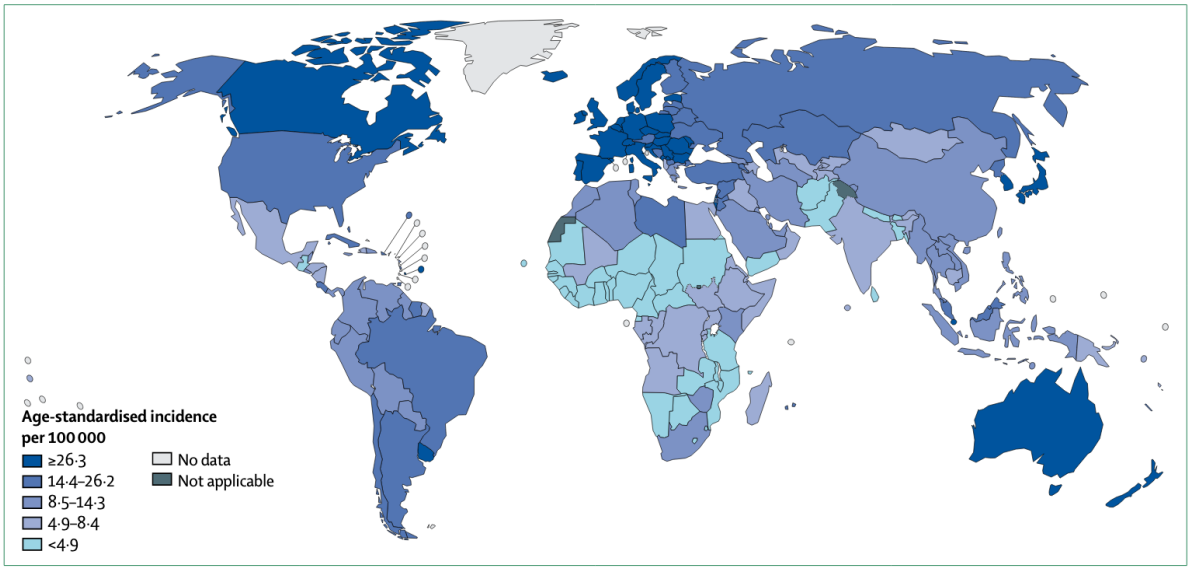


Fig. 1. Distribution of estimated age-standardised incidence rates of colorectal cancer worldwide in 2012. Data from GLOBOCAN 2012.

First of all, we developed a pipeline to correctly normalize the three different datasets. We performed feature selection in order to get the most significative genes defining differences in tumoral patients belonging to Asian and White ethnicity and analyzed these genes by using softwares and databases such as OncoSearch.

Following the same idea of selecting the most significative genes, we also performed network analysis on the Gene Co-expression Networks (GCNs) built by separating the patients based on their ethnicity. We built two GCNs, one for the White patients and one for the Asian patients, and ranked the nodes of the networks (i.e., the genes) based on their betweenness centrality, which is a measure of how "important" a node is in a network. Once again, we used OncoSearch and similar platforms to analyze the results.

We also provided a more detailed classification of CRC tumors based on intrinsic gene expression profiles by assigning each sample to a biologically distinct Consensus Molecular Subtype (CMS) and studied the differences in the subtypes distribution between patients of different ethnicities. We exploited the software CMScaller for this purpose.

Our results show that the main differences between the two ethnicities are related to different expression levels of key genes in the Notch signaling pathway. Specifically, it appears that in White patients these genes are upregulated compared to Asians'. This pathway is involved in many functional activities inside cells and further analysis will be necessary to study the effects of this altered pathway on CRC patients from different ethnicities.

Besides this introductory section, this document is organized as follows. In the Data section, we describe the data collected and survey the preprocessing steps performed as well as the difficulties we faced with respect to data collection. In the Methods section, we enumerate and technically describe the methods and the softwares we employed for the analysis of the data. In the Results and Discussion sections, we examine the results we obtained and their biological implications. In the Conclusions section, we give a summary of our work and of what we have learned while working to this project, as well as some possible further analyses. In the Appendix,

in the What Did Not Work section, we present a brief overview of which methods have been tested and not ended up being included in the main report.

2 DATA

In this section, we survey the data collection and preprocessing steps of our project. The task of setting up a good dataset for the project appeared to be the most difficult step overall. We overcame the scarcity of information by merging datasets coming from different sources. This, however, has not been trivial, since we needed to handle differences in format, scale and normalization of the data that might have come from different transcriptomic technologies. We validated the results of the proposed normalization by performing Principal Component Analysis.

2.1 Data description

Since analyzing the disparities in CRC among ethnic groups is the primary aim of our project, we needed to collect data related to patients coming from different geographical areas, focusing in particular to Asian and White ethnicities. To overcome the scarcity of data available in public databases belonging to Asian patients, we collected clinical and gene expression information of primary colorectal cancers from two different sources of data, i.e., *The Cancer Genome Atlas (TCGA)* database, and an aggregation of two independent *GEO* datasets, Microarray and RNA-Seq gene expression respectively. The TCGA dataset contains the gene expression profiling of 271 (246 White, 25 Asian) single colorectal adenocarcinomas samples, while the GSE154548 and GSE101896 cohorts consist of 40 (Korean) and 90 (Japanese) tumor samples respectively.

HTSeq-Count gene expression files were available for download from the TCGA data portal in a tab-separated format with one Ensembl gene ID column and one mapped reads column for each gene. The GSE154548 expression matrix was provided in `.txt` format having Ensembl gene identifiers as row names and Samples as columns. Expression profiles from the GSE101896 microarray dataset were available as both raw fluorescence CEL files and preprocessed sample-probe series matrix.

The available clinical metadata for the TCGA samples included diagnoses (tumor tissue site and pathologic stage, vital status, age at diagnosis, days to last follow up, days to recurrence), treatments (days to treatment, treatment id, therapeutic agents) and demographic (gender, race, ethnicity) information. The available clinical metadata for the GEO samples were limited to age at the time of surgical resection of the tumor, gender and race.

2.2 Preprocessing

In order to use the data we collected, we had to perform some preprocessing steps to make datasets comparable.

TCGA quantifies the gene expression levels using HTSeq, which generates a count of the reads mapped to each gene as non-normalized integer counts. In the GSE154548 dataset, although the same HTSeq software is used for extracting read counts and generate the expression matrix, values are normalized into CPMs (counts per million), by scaling by the total number of reads and further transformed into $\log_2(\text{CPM} + 1)$.

We encountered major problems in combining the TCGA and the GSE154548 datasets, raised by the difficulty in understanding the type of normalization actually applied to the GSE154548 dataset. After reversing the logarithm, expression values did not appear to be CPM-normalized as instead stated by the owner of the data. Calculating the sum for each column of the expression matrix, did not in fact result to be one million, thus denying the CPM-normalization.

Given the small number of patients, we decided to re-generate ourselves the expression matrix starting from the deposited FASTQ. We searched for the PRJNA646641 BioProject in ENA (European Nucleotide Archive) and downloaded the paired-end FASTQ files for each patient. Quantification was performed using Kallisto [3]. Kallisto avoids base-to-base alignment of the reads, which is a time-consuming step. Instead, it performs a procedure

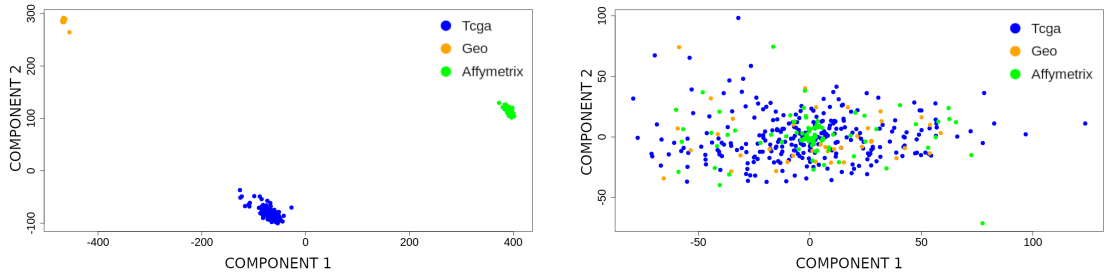
called pseudoalignment. Pseudoalignment requires processing a transcriptome file to create a transcriptome index. The reference transcriptome file for Homo Sapiens was downloaded in FASTA format from the Ensembl ftp site. After building the index, we ran the quantification algorithm using the `quant` command, which resulted in a tab-separated output file containing the transcript abundance estimates. This was repeated for all the patients, of which the resulting abundance files were eventually merged together to form a single expression matrix.

We divided the preprocessing pipepeline in 2 main parts: the first aimed at merging and normalizing the two RNA-seq experiments, while the second aimed at merging the new composite dataset with the Affymetrix one.

2.2.1 First Phase: Merging RNA-Seq Data. We found difficulties in merging data coming from two different RNA-seq experiments, even though they represent the same disease conditions. Both datasets are composed of around 60000 genes. PCA visualization brought us evidences of a strong batch effect, fig. 2a. Before applying the actual normalization procedure, we merged together the 25 TCGA Asians with the Korean dataset and performed a two tailed t-test to filter out the noisiest rows (p-value adjusted with BH method < 0.01). We then added the left-out TCGA set of only White patients and applied the *combatR* [4] software employing an empirical Bayes approach to adjust the batch effect between the two datasets.

2.2.2 Second Phase: Merging RNA-Seq Data and Affymetrix Data. Once again we relied on combatR's functionalities to normalize the two sets of patients. Looking at the boxplots of the two RNA-Seq experiments in fig. 3a, a difference in median signaling levels was clearly visible. We applied median normalization to fix this issue prior to apply combatR. We eventually filtered out rows with the least variance (RowVariance > 0.1) to remove genes that provide poor information about differences between datasets. The resulting dataset contains 19811 genes.

By looking at the new PCA reported in fig. 2b, we were now able to confirm the absence of the batch effect. Moreover, the boxplots reported in fig. 3b, confirmed that the distributions for each patient are now comparable.



(a) Starting data

(b) Normalized data

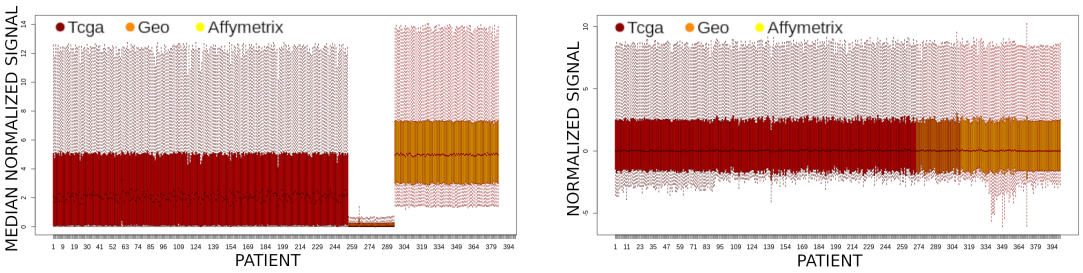
Fig. 2. Differences in PCA of initial data (a) and after-normalization (b). In (a) it is clearly visible that a strong batch effect makes data far from comparable

3 METHODS

In this section, we enumerate and describe more technically the methods and the softwares we used for the analysis of the data.

3.1 Feature Selection

After having successfully performed normalization, we needed to apply a feature selection strategy in order to find the most important features to work with. Feature selection helps to focus only on the relevant variables in a



(a) Starting data

(b) Normalized data

Fig. 3. Differences in the boxplots of the initial data (a) and after-normalization (b). (a) explicits very well the differences in scale between the two RNA-Seq experiments.

dataset by for instance eliminating collinear variables or reducing the noise, therefore allowing the software or the machine learning model to receive as input the most relevant signals.

3.1.1 Mutual Information. The criterion we ended up using to perform feature selection is *mutual information*. Mutual information has been successfully adopted in filter feature-selection methods to assess both the relevancy of a subset of features in predicting the target variable and the redundancy with respect to other variables. Since we were interested in differences at ethnicity level, in our case we considered the genes as the set of features and the ethnicity as the target variable. In this way, we were able to identify the most important genes to discriminate the ethnicity.

To perform feature selection with mutual information we exploited the Python package *scikit-learn*.

3.1.2 Statistical and Biological Validation. An important parameter that needs to be selected when using feature selection by mutual information is k, i.e., the size of the set of selected features. In our case, we wanted to reduce the number of features from around 20000 to around 200.

The performance of selecting k as a parameter can be measured by training a classifier which aims at correctly classifying whether a patient (defined by the k feature) is Asian or White. We trained a Support Vector Machine classifier, using again the Python library scikit-learn.

The classifier allowed us to validate the parameter k, meaning to understand if k features were able to correctly classify a sample. The classifier has been validated by a k-fold cross validation (with 10 folds).

We ended up selecting $k = 150$ genes, which yields an accuracy of about 90% but, for further analysis, we also considered $k = 300$, which yields an accuracy of about 87%.

To biologically validate the extracted features, and therefore the extracted genes, we relied on external databases such as OncoSearch. OncoSearch is a Web-based engine that searches Medline abstracts for sentences that mention gene expression changes in cancers, with queries that specify whether a gene expression level is up-regulated or down-regulated, whether a certain type of cancer progresses or regresses along with such gene expression change and the expected role of the gene in the cancer.

OncoSearch allowed us to understand if the extracted genes were involved in oncogenesis and to understand how such genes affect cancers.

Further analysis have been carried out using EnrichNet and Gene Ontology, which are web-services for enrichment analysis.

3.2 Gene Co-expression Network Analysis

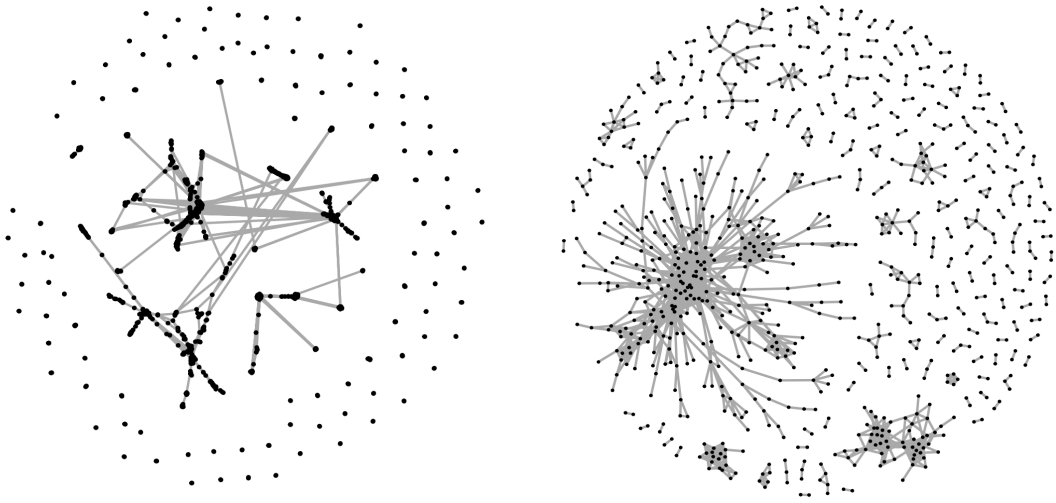
Gene Co-expression Networks (GCNs) are transcript–transcript association networks, generally reported as undirected graphs, where genes are connected when an appreciable co-expression association between them exists. GCNs are built from gene expression data by calculating co-expression values in terms of pairwise gene correlation (Pearson's) score and choosing a significance threshold [5].

For all the following analysis we used R and the package igraph, which is a collection of network analysis tools.

3.2.1 GCNs Creation. The aim was to create two different GCNs, one which represented White patients and one which represented Asian patients. To do so, we divided the dataset in two sub-datasets: one with only Asian patients, and one with only White patients. Each patient is described by the whole list of ~20000 genes. Building the correlation matrix with 20000 genes is really expensive from a computational point of view, and we advise to perform some feature selection also in this case. However, since our previous method was heavily based on feature selection, and since we wanted to compare against that method, for our purposes we decided to keep the original list of genes.

We created the two correlation (Pearson's) matrices, one from the White dataset and one from the Asian dataset, and discarded uncorrelations and weak correlations, i.e., we set a threshold of 0.75.

The resulting matrices were considered as the adjacency matrices of our networks.



(a) Gene co-expression network for White patients (b) Gene co-expression network for Asian patients

Fig. 4. Gene co-expression networks of the patients based on their ethnicity. For the sake of visualization, we excluded isolated nodes. We found differences in the number of nodes in the networks, after the filtering step. The network of White patients preserves an higher number of nodes, around 2000, while the other network preserved 850 genes. The topology of the network, however, is the same: we have a giant component and a lot of really small disconnected communities made of 3, 4 or 5 nodes. This behavior is more graphically noticeable in the Asian network.

3.2.2 GCNs Analysis. When dealing with networks, a lot of analysis can be done. We focused on analysing the most "central" genes in the networks, and understand if there are any similarities or differences between the two networks.

The centrality of the nodes, or the identification of which nodes are more "important" than others, has been a key issue in network analysis, since there is not a unique indicator of importance of a node in a network [6]. In this work, we follow the literature in GCNs analysis and use the *betweenness centrality* to rank the nodes.

The betweenness centrality is a measure based on shortest paths, it considers the number of shortest paths that pass through the node. Intuitively, it is correlated with how well a node connects distant communities, which in our case grouped genes having similar biological functions.

After having ranked the nodes, we performed the same analysis as before, exploiting OncoSearch and the other previously mentioned softwares.

3.3 Consensus Molecular Subtypes Analysis

In 2015 Guinney et al. in the CRC Subtyping Consortium analysed gene expression profiles and tumor characteristics of more than 4000 patients, and classified colorectal cancers into four biologically distinct and clinically relevant consensus molecular subtypes (CMS) [7]: CMS1, MSI-immune; CMS2, epithelial and canonical; CMS3, epithelial and metabolic; and CMS4, mesenchymal. CMS1 (14%) displays microsatellite instability with mutations in genes encoding DNA mismatch-repair proteins, along with increased immune cell infiltrates. CMS2 (37%) is characterized by high somatic copy number alteration (SCNA) and WNT and MYC activation. CMS3 (13%) is defined by dysregulation of metabolic pathways including carbohydrate and fatty acid oxidation. CMS4 (23%) involves the upregulation of epithelial mesenchymal transformation (EMT) pathways, stromal invasion, angiogenesis and transforming growth factor$-\beta$ (TGF$-\beta$) activation. This leaves ~13% of the tumors that cannot be assigned to a consensus subtype, as they have mixed gene expression signatures with characteristics of multiple CMS. The CMS subtypes are not classified only by molecular features, but also by clinical behavior and prognosis, with CMS1 having the worse survival after relapse and CMS4 having the worse relapse-free survival and overall survival. Given the lower CRC mortality among Asian populations compared to CRC mortality in Western countries, we hypothesized a heterogeneous CMS prevalence among our datasets. In order to verify our hypothesis, we inferred the consensus subtype for each patient by using the CMScaller R package [8]. The CMScaller classifier is based on a set of cancer cell-specific, subtype-enriched gene markers, and is suitable for both microarray and RNA-Seq data. We applied the CMScaller function to each one of our three non-normalized expression matrices independently.

3.4 Deconvolution Analysis by CIBERSORTx

The tumor microenvironment, which is composed of molecules such as immune cells and mesenchymal cells, is the cell environment in which the tumor is located. A pan-cancer immunogenomic analysis revealed that numerous tumor-infiltrating lymphocytes associated with adaptive immunity are associated with a good prognosis, including activated CD8 T cells, resting memory CD4 T cells and effector memory CD4 T cells [9]. To investigate the differences in immune cell infiltration within the tumor microenvironment between White and Asian samples, we scored immune expression signatures and determined the fraction of immune associated cell types. For each independent dataset, we employed the CIBERSORTx algorithm [10] which uses a set of barcode gene expression values (LM22, a validated "signature matrix" of 547 genes) for the deconvolution of 22 types of infiltrating immune cells. These transcriptomic markers are transcripts specifically expressed by a given cell population and not by the others. CIBERSORTx has been successfully validated, and used for determining immune cell landscapes and their relations to treatment response in breast and liver cancers [11, 12]. The format of our datasets are Ensembl identifiers, but because LM22 uses HUGO gene symbols we used biomaRt for the conversion. RNA-Seq data required no further preprocessing. Conversely, Affymetrix CEL files belonging to the Japanese dataset needed to be transformed into a tabular format suitable for analysis with CIBERSORTx. For this purpose we downloaded a CDF (Chip Description File) compatible with the HGU133 Plus 2.0 microarray platform from BrainArray, and

run an R script provided by the CIBERSORTx website to generate the normalized expression dataset that is compatible with CIBERSORTx. Eventually, gene expression data was uploaded to the CIBERSORTx web portal (https://cibersortx.stanford.edu/), with the algorithm run using the default signature matrix at 100 permutations as recommended to achieve statistical rigor.

4 RESULTS AND DISCUSSION

In this section, we survey the results of our analysis and discuss their biological implications.

4.1 Cancer Studies

We started by looking at possible correlation with cancer in the list of genes extracted via mutual information, matching with the cancer database provided by OncoSearch. In this way we were able to select 9 genes involved in cancer. Analysing them with enrichment analysis software provided by GO website, EnrichNet and EnrichR Fig. 5. (biological process analysis) we found out they were mainly involved (6 out of 9) in regulation of cell death and most of them are related to NOTCH1 signaling pathway metabolism, crucial for maintaining progenitor cell population as well as the balance between cell proliferation, differentiation and apoptosis [13].

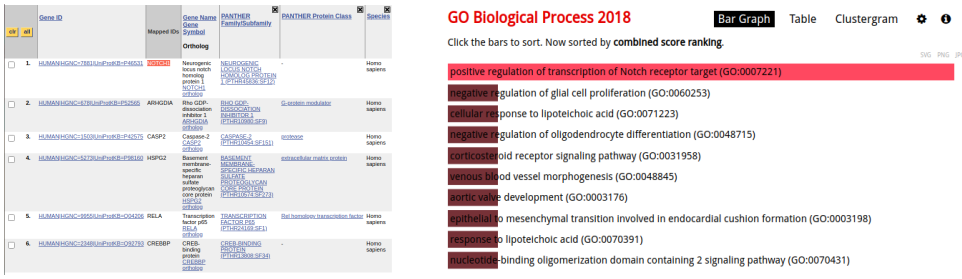


Fig. 5. Enrichment analysis performed on the 9 cancer genes matching with OncoSearch DB. On the right Notch pathway genes related to negative regulation of cell death (GO, panther DB), on the right the various functionalities that Notch pathway can have in a cell (EnrichR analysis)

In particular, displaying boxplots fig. 6 of the selected genes we can clearly see their level of up-regulation in case of White population with respect to the Asian ethnicity. From literature, enhanced levels of NOTCH1 have been correlated with tumor progression and metastasis, related to inhibition of apoptosis promoted by NOTCH1 expression [14].

To further validate our results, we tried to find overlaps between our list of 150 genes and a second list of 2700 genes associated with cancer in Asian ethnicity form Wu et. al studies [2]. Unfortunately we were not able to find common genes. We then repeated the experiment with a bigger list of 300 genes, output from mutual information validated before. In this case we were able to identify 12 common genes. Matching the same list of 300 genes in OncoSearch we had the same results, furtherly proving the role of these 12 genes in cancer regulation.

Analysing these genes with Enrichnet and Gene Ontology, was very interesting to notice the presence of NOTCH3, key player along NOTCH1 in the Notch signaling pathway and other secondary genes belonging to that pathway. This result provided further evidences about the involvement of this pathway in defining tumor among the two populations.

4.1.1 Notch signaling pathway. Despite the fact that the core Notch pathway operates in vastly different developmental and disease contexts, from stem cell regulation and heart morphogenesis to cancers and cardiomyopathies,

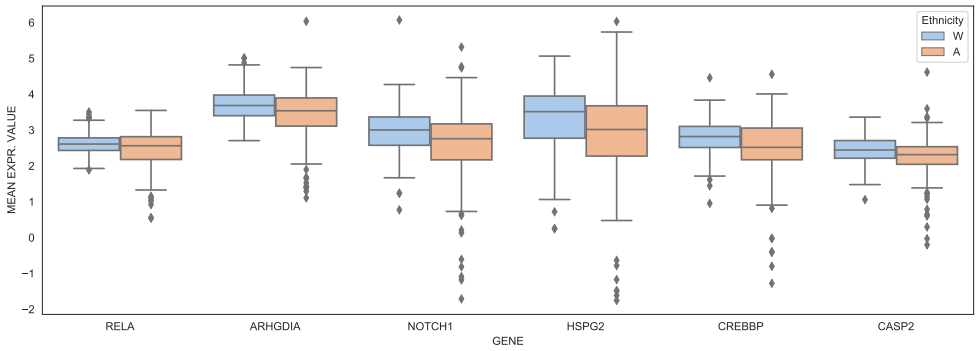


Fig. 6. Differences in expression of genes involved in Notch signalling pathway between white and asian patients, in blue and orange respectively. As noticeable expression levels of these genes are higher in white patients with the respect to asians

it is relatively simple in its operation. Notch ligands are transmembrane proteins receptors from which ligand-mediated activation induces a series of proteolytic cleavages in members of the Notch family of receptors, which release the Notch intracellular domain (NICD). Once released, the NICD enters the nucleus and, together with the DNA-binding protein CBF1–Suppressor of Hairless–LAG1 (CSL; also known as RBPJ) and the co-activator Mastermind (Mam; Mastermind-like transcriptional co-activator 1 (MAML1) in human), stimulates transcription of target genes.

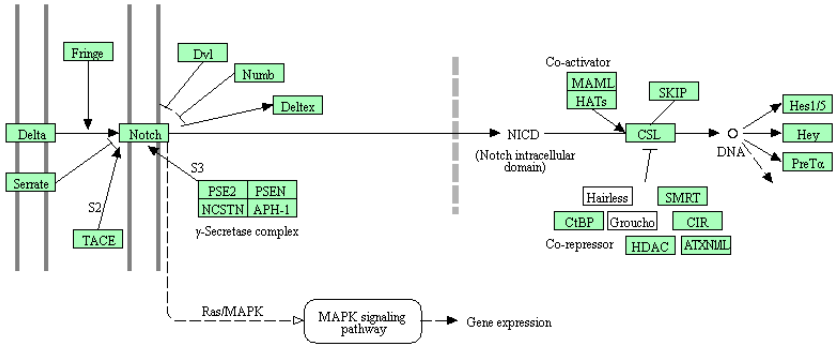


Fig. 7. Main genes involved in Notch signalling pathway

4.1.2 Notch Signaling in the Immune System. NOTCH1 seems involved in the growth and differentiation of mature T cells (CD4 and CD8) [15]. Moreover, Notch signaling is involved in Macrophage activation and its effector functions [16]. Macrophages are a class of immune cells present in high numbers in the microenvironment of solid tumors, at all stages of tumor progression. They are involved in cancer-related inflammation and have emerged as a crucial player in the creation of a tumor microenvironment that supports tumor growth and metastasis [17], in opposition to their traditional role as an innate immune cell, whose function is to eliminate cancer cells. Macrophages are divided into three subtypes (M0, M1 and M2). M0 is the unactivated subtype and does not exhibit inflammatory or tumor-related function. Depending on the activation pathways, M0 can

differentiate into two activated subtypes, M1 and M2, which serve different immune functions. We found that all three subtypes of Macrophages were more infiltrated in White CRC samples compared with Asian samples (Fig. 9a and Fig. 9b), which was consistent with NOTCH1 upregulation among White patients. Conversely there was no correlation between NOTCH1 upregulation and infiltration of mature T cells in the White population, despite insights from the literature suggesting a role of NOTCH1 in modulating T cells immunity.

4.1.3 Immune Characteristics of CRC in GEO and TCGA Datasets. Taking a broader look at all 22 infiltrated immune populations quantified by CIBERSORTx, a wide heterogeneity of immune cells among different datasets was uncovered (Fig. 9a and Fig. 9b). Memory B cells and plasma cells were more infiltrated in Japanese and Korean GEO cohorts compared with both White and Asian TCGA. Conversely, M1 and M2 macrophages were less infiltrated. M0 macrophages were the highest in White TCGA tumors. T cells gamma delta and naïve CD4 T cells were only expressed among Japanese and Asian TCGA tumor samples.

4.1.4 Network Analysis Results. Another interesting result comes from network analysis. When analyzing the network of White patients, we found again NOTCH3 among the top 5 genes ranked according to their betweenness centrality, validating the involvement of this pathway in a large number of biological processes. Plus, in the top 1% of the nodes, we also found the gene MAP2K2, which is a kinase closely related with Notch pathway fig. 7, and that could be potentially involved in a number of cancer phenotypes.

By analyzing the network of Asian patients, we found the gene CDKN3 among the highest ranked genes. In [18], authors found that CDKN3 had remarkable effects in suppressing colorectal cancer cell proliferation and migration, inducing cell cycle arrest and apoptosis in a colorectal cancer cell line, SW480 cells. The in vitro studies in SW480 cells revealed a unique role of CDKN3 in regulating cellular behavior of colorectal cancer cells, and implied the possibility of targeting CDKN3 as a novel treatment for colorectal cancer.

4.2 Consensus Molecular Subtyping

We performed the CMS classification before realizing that something was wrong within the GSE154548 dataset, and indeed only 5 tumor samples, out of 40, were assigned to a consensus subtype, with 87.5% of the samples designated as "not assigned" (NA) by the algorithm. After Kallisto re-generation of the gene expression matrix from scratch (as described in Preprocessing section), we repeated the CMScaller transcriptome-based classification (Fig. 8). The distribution of CMS among White TCGA patients were CMS1 15%, CMS2 28%, CMS3 15%, CMS4 24%, NA 18%. The distribution of CMS among Asian TCGA patients were CMS1 8%, CMS2 8%, CMS3 8%, CMS4 10%, NA 64%. Frequencies for the Japanese tumor samples were as follows: CMS1 10%, CMS2 20%, CMS3 12%, CMS4 28%, NA 30%. For the Korean tumor samples, frequencies were as follows: CMS1 18%, CMS2 10%, CMS3 5%, CMS4 25%, NA 42%. Compared to CMS classification analysis on the original GSE154548 dataset, classification on the re-generated gene expression matrix performed better. Still, the percentage of "not assigned" samples remained high (42%). The classifier failed because these tumors had molecular features that were mixed. They may reflect a transition phenotype or samples with intratumor heterogeneity. The "not assigned" percentage was even higher (64%) among the 25 Asian TCGA samples, while the White counterpart reported the lowest NA percentage (18%). Yet, they belong to the same TCGA dataset. We venture the hypothesis that high intratumor heterogeneity and mixed gene expression signatures could be a characteristic of CRC in Asian patients. Moreover, while the White population matched reported distributions by the CRC Subtyping Consortium (CRCSC), the distribution of CMS subtypes in the remaining two datasets varied substantially. It should be noticed that the molecular subtypes, as defined by the CRCSC, were mostly derived from a US/European population and could not be representative of other ethnic groups. In this direction a recent study [19] analysed 366 colon cancer samples from Brazil, Canada, Mexico, Thailand, and the US. Their findings suggest that CMS subtype prevalence in CRC differs substantially by geographic region.

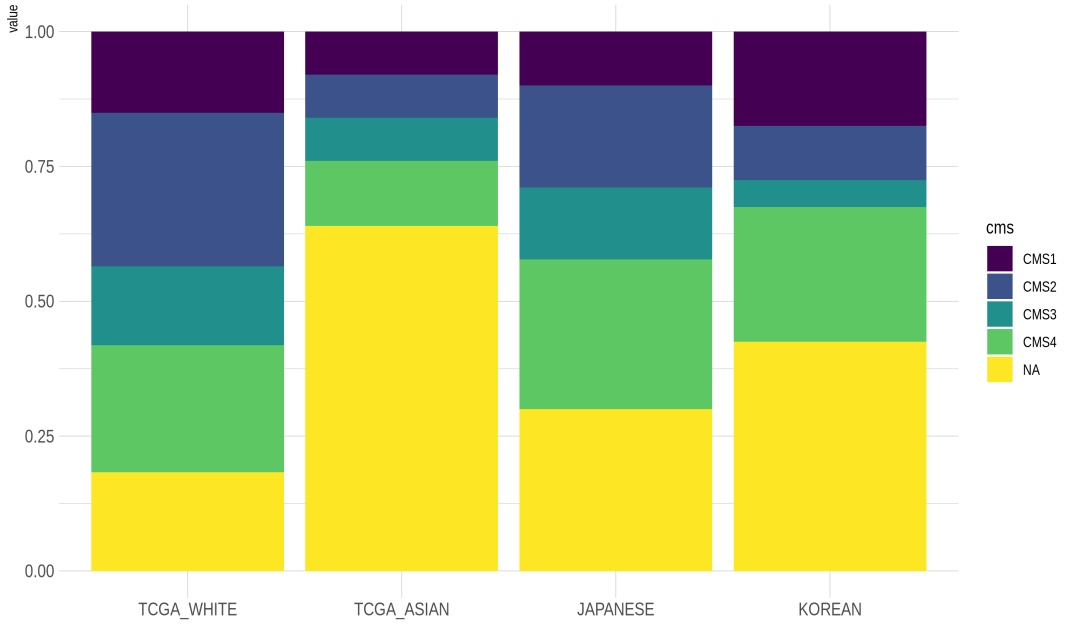
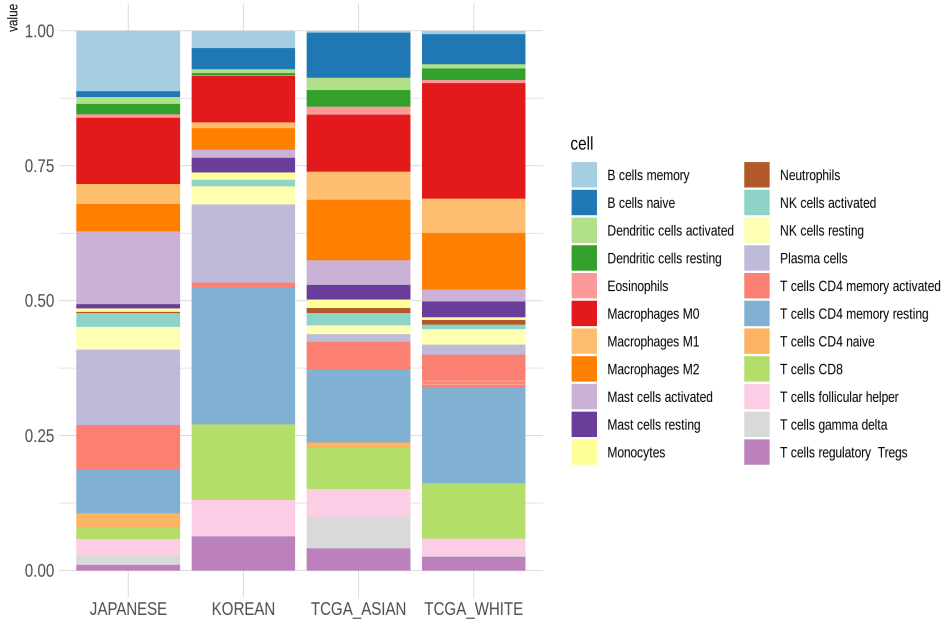
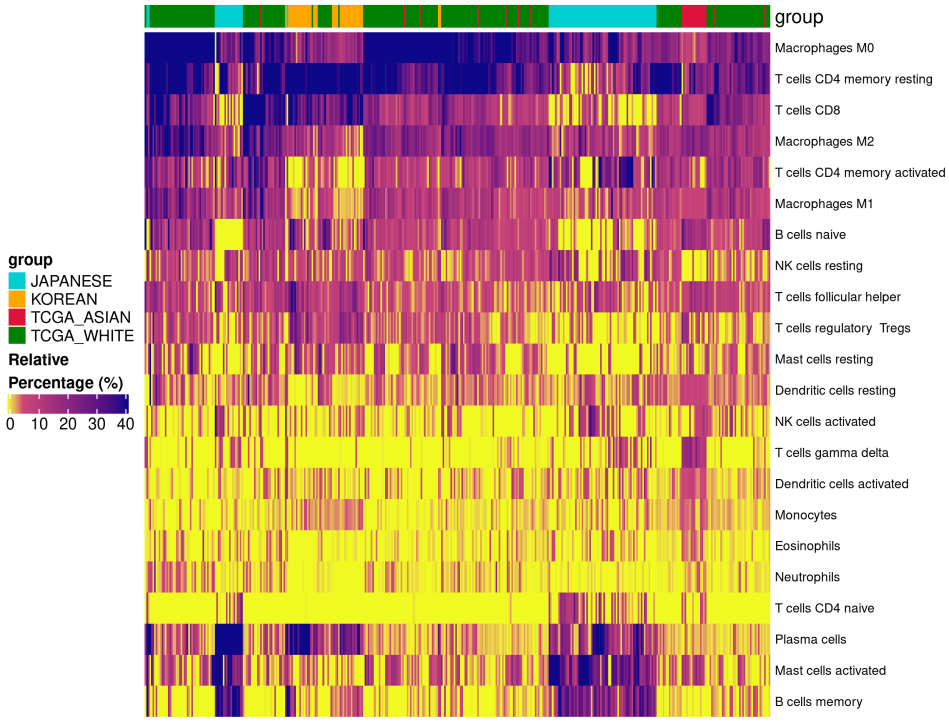


Fig. 8. Distribution of colorectal cancer patients according to Consensus Molecular Subtypes classified with CMScaller

Unexpectedly, CMS4 was the predominant subtype in both the Korean and Japanese datasets. CMS4 corresponds to worse OS or DFS survival. However, Wu et al. [2] observed that whereas the CMS4 and stem-like subtypes in the TCGA dataset were prone to death, such was not the case in their Taiwanese dataset. Unfortunately, survival data was missing for both the Korean and Japanese GEO datasets, and we couldn't perform Kaplan–Meier plot for overall survival analysis.



(a)



(b)

Fig. 9. Inferred composition of 22 immune cell subsets in GEO and TCGA datasets. Both the stacked bar plot figure and the heatmap were generated using the CIBERSORTx output estimates, normalized to sum up to one and thus can be interpreted directly as cell fractions for comparison across different immune cell types and datasets

5 CONCLUSIONS

In this project we dealt with "real" data and all the difficulties they involve.

We posed the goal of identifying differences from gene expression in CRC patients belonging to White and Asian ethnicity, similarly to what has been done by Wu et al. [2]. Specifically, we focused on comparing CRC patients belonging to two different ethnicities in order to find the most important differences in cancer-related gene expression levels.

In the first part of the project, we spent several hours finding an appropriate method to normalize the three different datasets we selected for the analysis (two RNA-Seq and one Affymetrix).

We performed feature selection by mutual information to extract the most significative genes when classifying between ethnicities, requiring the genes to also differ in terms of expression levels between White and Asian patients. From this list of genes we were able to isolate 12 cancer-related genes by using OncoSearch database. Surprisingly, 8 of these genes (NOTCH1, NOTCH3, RELA, CREBBP) are part or (MAP2K2, CASP2, HSPG2, ARHGDIA) closely related to Notch signaling pathway, a pathway involved in many biological responses including regulation of cell death and cellular development. By comparing expression levels of these genes in White and Asian patients, it was clearly visible they were upregulated in Western patient (around 1,3 fold upregulation).

Network analysis performed on White patients confirmed the role of these two genes in the regulation of a larger number of genes downstream, suggesting their ability to control and regulate CRC phenotype in diseased patients.

Looking at literature, it resulted that this pathway is highly conserved across Eukarya [20], suggesting that in healthy patients its expression levels should be very similar. Therefore our results are probably significant at clinical level.

By using CIBERSORTx, we were able to predict the immune cell populations for each patient. From this analysis we found an increased level of Macrophages in White patients compared to Asian, these results can be attributed to the upregulation of Notch pathway in White ethnicity.

Currently, in literature there are many evidences on how Notch pathway upregulation is involved in colorectal cancer, but we found no article focusing on comparing the effect of this pathway in patients from different ethnicity.

From our analysis, we hypothesise Notch signaling pathway is fundamental in regulating CRC in different ethnicities. Based on what is presented in literature, we hypothesise that the higher expression levels of this pathway in White patients could be correlated with a higher severity of the disease that can be furtherly linked with increased of CRC lethality. Moreover, other studies related to different types of cancer (e.g., breast cancer) have investigated and supposed a relation between ethnicity, mortality and the notch signaling pathway [21]. Of course further studies to ascertain the effects of Notch pathway in CRC patient mortality will be necessary to validate our results.

Finally, our study comes with some limitations. Our analyses and derived hypotheses are set without considering the effects of aging and sex on patients' gene expression levels and the differences in expression level between healthy and cancer patients. A more comprehensive study is need to verify the validity of our results.

REFERENCES

[1] Center for Disease Control and Premise. Colorectal cancer statistics, 2020.

[2] Shao-Min Wu, Wen-Sy Tsai, Sum-Fu Chiang, Yi-Hsuan Lai, Chung-Pei Ma, Jian-Hua Wang, Jiarong Lin, Pei-Shan Lu, Chia-Yu Yang, Bertrand Chin-Ming Tan, and Hsuan Liu. Comprehensive transcriptome profiling of Taiwanese colorectal cancer implicates an ethnic basis for pathogenesis. *Scientific Reports*, 10(1):4526, March 2020.

[3] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.
[4] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *bioRxiv*, 2020.
[5] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 01 2017.
[6] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford; New York, 2010.
[7] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine*, 21(11):1350–1356, 2015.
[8] Peter W Eide, Jarle Bruun, Ragnhild A Lothe, and Anita Sveen. Cmscaller: an r package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Scientific reports*, 7(1):1–8, 2017.
[9] Pornpimol Charoentong, Francesca Finotello, Mihaela Angelova, Clemens Mayer, Mirjana Efremova, Dietmar Rieder, Hubert Hackl, and Zlatko Trajanoski. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell reports*, 18(1):248–262, 2017.
[10] Binbin Chen, Michael S Khodadoust, Chih Long Liu, Aaron M Newman, and Ash A Alizadeh. Profiling tumor infiltrating immune cells with cibersort. In *Cancer Systems Biology*, pages 243–259. Springer, 2018.
[11] Rico D Bense, Christos Sotiriou, Martine J Piccart-Gebhart, John BAG Haanen, Marcel ATM van Vugt, Elisabeth GE de Vries, Carolien P Schröder, and Rudolf SN Fehrmann. Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer. *JNCI: Journal of the National Cancer Institute*, 109(1), 2017.
[12] Nataliya Rohr-Udilova, Florian Klinglmüller, Rolf Schulte-Hermann, Judith Stift, Merima Herac, Martina Salzmann, Francesca Finotello, Gerald Timelthaler, Georg Oberhuber, Matthias Pinter, et al. Deviations of the immune cell landscape between healthy liver and hepatocellular carcinoma. *Scientific reports*, 8(1):1–11, 2018.
[13] Domingos Henrique and François Schweisguth. Mechanisms of notch signaling: a simple logic deployed in time and space. *Development*, 146(3), 2019.
[14] Kaitlyn E Vinson, Dennis C George, Alexander W Fender, Fred E Bertrand, and George Sigounas. The n otch pathway in colorectal cancer. *International Journal of Cancer*, 138(8):1835–1842, 2016.
[15] Joshua D Brandstadter and Ivan Maillard. Notch signalling in t cell homeostasis and differentiation. *Open biology*, 9(11):190187, November 2019.
[16] Tanapat Palaga, Wipawee Wongchana, and Patipark Kueanjinda. Notch signaling in macrophages in the context of cancer immunity. *Frontiers in immunology*, 9:652, 2018.
[17] Roy Noy and Jeffrey W Pollard. Tumor-associated macrophages: from mechanisms to therapy. *Immunity*, 41(1):49–61, 2014.
[18] Cheng Yang and Jun-Jun Sun. Mechanistic studies of cyclin-dependent kinase inhibitor 3 (cdkn3) in colorectal cancer. *Asian Pacific Journal of Cancer Prevention*, 16(3):965–970, 2015.
[19] Krittiya Korphaisarn, Michael Lam, Jonathan M Loree, Erika Ruiz, Samuel Aguiar, and Scott Kopetz. Consensus molecular subtypes in colorectal cancer differ by geographic region., 2020.
[20] Eve Gazave, Pascal Lapébie, Gemma S Richards, Frédéric Brunet, Alexander V Ereskovsky, Bernard M Degnan, Carole Borchiellini, Michel Vervoort, and Emmanuelle Renard. Origin and evolution of the notch signalling pathway: an overview from eukaryotic genomes. *BMC evolutionary biology*, 9(1):1–27, 2009.
[21] Chang-Sheng Chang, Eiko Kitamura, Joan Johnson, Roni Bollag, and Lesleyann Hawthorn. Genomic analysis of racial differences in triple negative breast cancer. *Genomics*, 111(6):1529 – 1542, 2019.

A WHAT DID NOT WORK

In this section, we give a brief overview of which methods have been tested and not ended up being included in the main report of the work.

A.1 Feature Selection

A.1.1 Random Forest. We observed that Random Forest was particularly good (accuracy of 97% with 10-folds cross-validation) in classifying Asian and White patients. However, it shows poor capabilities in distinguishing Asian patients from the TCGA dataset, classifying all of them as White.

Running the algorithm many times, we found that only a small fraction of the most significant genes for classification were actually conserved. This was probably due to the fact that the model tends to overfit the data, and uses noisy genes in the classification process.

We tried to overcome this problem by doing the following. We ran the algorithm 100 times, and we associated a counter to each gene and at each iteration we increased the counter for that specific gene if it was considered by the predictive model.

- 7020 genes were counted more than 10 times
- 3157 genes were used at least 1/4 of the times
- 2141 genes were used at least 1/3 of the times
- 882 genes were used more than 1/2 of the times

Running Random Forest again, but performing feature selection and considering only the most important genes derived by the method explained above, and performing cross validation, the model still suffers from the problem of not being able to correctly classify Asian patients from TCGA. This revealed again overfitting problems and we decided to instead opt for other methods.

A.1.2 Genetic Algorithm. One other method that suffered the same issues as the previous was the use of Support Vector Machines with a subset of genes selected with a genetic algorithm. Essentially, at initialization time a number of random sets of genes is created, with sizes spanning from 2 to 500 genes (larger sets were not considered if not during the first tests since the system always tended to converge to sets of 5 to 30 elements). At this point, an iteration begins which trains and tests a different SVM on each random set, and assigns to each one of them a corresponding fitness value. This value is equal to accuracy average of 5 different train and test procedures, each one using one of 5 different predefined seeds for the random splitting of the dataset in a train (80%) and test (20%) subset. Once all the random sets have a fitness assigned, they are sorted accordingly and a generation is considered complete. To generate the next generation, the least performing half of the sets is discarded and the best half is recombined with itself to give birth to new elements of the population. Recombination consists of a matching phase, where for each sample the best match is chosen, a crossover step where the genes of the two are combined following a predefined criteria, and a final mutation procedure where (to avoid overfitting) random mutations have a chance to be introduced in the form of added or removed genes from the set. This iteration of selection and recombining is repeated until no more improvement can be observed in the accuracy of the best sets.

As anticipated, this approach suffered the very same issues of Random Forests, leading to a high accuracy that hid bad generalization problems.