Identification and validation of a vitamin D-related prognostic signature in colorectal cancer

Diego Barquero Morera[1], Giacomo Fantoni[2], Gaia Faggin[3], Leonardo Golinelli[4]

Abstract
Colorectal cancer (CRC) is one of the most common malignant carcinomas worldwide with poor prognosis, imposing an increasingly heavy burden on patients. Different studies have shown that vitamin D and vitamin D-related genes play a key role in CRC. In this study we aim to identify and validate vitamin D-related prognostic signature in colorectal cancer. In the first part of our work we will focus on the normalization and the pre-processing of our datasets: a colorectal cancer gene expression dataset and a vitamin D level gene expression dataset. Regarding the first dataset we will split the data in "stage I and II gene expression data" and "stage III and IV" gene expression data; while for the second dataset we will split the data in "low vitamin D level gene expression data" and "high vitamin D gene expression data". Then by the means of a statistical analysis we will obtain a dataset of differentially expressed genes (DEGs) in different stages of CRC and a dataset for the vitamin D gene signature. Using the cox proportional-hazard model we will compute the hazard ratio associated with each of our gene of interest and compare the effect of the genes related to the vitamin-D with respect to the ones differentially expressed on the cancer stages. For each gene related on the vitamin-D found significant and its stage-related counterpart a Kaplan-Meier curve will be computed, to compare the change of overall survival according to its expression level. This regression analysis, once validated will allow us to obtain a list of genes that can be used for stratification and prognosis of patients suffering from colorectal cancer.

[1] *diego.barqueromorera@studenti.unitn.it*
[2] *giacomo.fantoni@studenti.unitn.it*
[3] *gaia.faggin@studenti.unitn.it*
[4] *leonardo.golinelli@studenti.unitn.it*

Contents

1. Introduction

Colorectal cancer (CRC) is the third most common malignant tumor worldwide and is the second one in cancer-related deaths [**?**]. In spite of improvements in the management and treatments of patients with CRC in the last two decades, no satisfactory therapy exists when the surgery is not curative. The poor prognosis and the increasing incidence of CRC have provided strong motivation to construct a predictive model in CRC patients, which will benefit personalized treatment in clinical management [**?**]. There are lot of epidemiological and preclinical studies that indicate a beneficial effect of vitamin D on CRC incidence and mortality [**?**] [**?**]. Vitamin D is a fat-soluble vitamin and many genes are related to its metabolism and action [**?**]. It can be obtained from diet or the endogenous synthesis in the epidermis under sunlight exposure [**?**]. It has been demonstrated that vitamin D benefits clinical outcome and improves the long-term survival of CRC patients [**?**]. Moreover, circulating vitamin D may be a CRC biomarker and its deficiency is related to the high incidence of CRC [**?**]. A better survival outcome in CRC is associated with higher prediagnostic or postdiagnostic serum 25-hydroxyvitamin D concentrations [**?**]. The most active vitamin D metabolite (1α,25-dihydroxyvitamin D_3) inhibits the proliferation and promotes the differentiation of cultured colon carcinoma cells

by mechanisms that include cell cycle arrest at G0/G1 phase, blockade of the Wnt/β-catenin pathway and induction of E-cadherin and other epithelial proteins [**?**] [**?**] [**?**]. Lots of genes related to vitamin D metabolism and action play an essential role in tumors. For example, CYP24A1 an important vitamin D-related gene, is up-regulated in CRC patient and nominated as a promising biomarker [**?**]. Vitamin D and its related genes are correlated with the homeostasis of the intestinal epithelium and regulate immune cells [**?**].

The objective of this project was to find a way to make prognosis and stratify patients suffering from colorectal cancer by means of their transcriptomic profiles. In particular we focused on the gene signature of vitamin D as a prognostic marker, by leveraging the always increasing gene expression data publicly available. The final goal was to identify colorectal cancer survival markers related with vitamin D effects, as well as any pathways involved.

2. Material and methods

2.1 Data preprocessing

To achieve a better statistical significance, a high number of samples was originally collected from different available public datasets on gene expression in CRC. These were downloaded from the databases Gene Expression Omnibus (GEO) [1] and The Cancer Genome Atlas (TCGA) [2]. In total, 21 datasets from GEO and 1 from TCGA were obtained, and their metadata manually curated and standardized (table 1). It is worth noting that only the dataset GSE157982 had the gene signature of Vitamin D in CRC, necessary for its further analysis.

2.1.1 Sample splitting and filtering

All these starting samples will be filtered so that they have all the data necessary for our analysis. Furthermore we will split them in different sets, so that every step in our pipeline will have enough data to be statistically significant. So we define 4 different sets. The first one will be used to find differentially expressed genes between different stages of cancer. The second one will be used to find the vitamin D gene signature. The third one will be used to fit cox proportional-hazard model. The fourth one will be used to plot the Kaplan-Meier curves.

2.1.2 Dataset normalization

Having decided to use different dataset for this analysis, a normalization step has been introduced to remove batch effects. To do so we performed on the raw micro-array data different type on normalization. Firstly we tried with Robust Multiarray Analysis *RMA* and then with Frozen Robust Multiarray analysis *fRMA*. The logaritmic value of all the gene expression levels has been computed and used downstream in the analysis. We divided the normalized data in two sets to be used downstream. The first was composed of datasets GSE17536 and GSE17637 and will be used to fit the Cox proportional-hazards model. The second was composed of datasets GSE31595, GSE33113, GSE38832, GSE39084 and GSE39582. This second set has been normalized in two ways: we first normalized all the samples together to compute a single data matrix used downstream to plot the Kaplan-Meier curves and then we implemented a bootstrap procedure. In this procedure we sampled from each dataset 600 individuals 10 times so that in each matrix we obtained balanced data with respect to the cancer stage.

2.2 Differentially expressed genes

After having removed the batch effect and having obtained all the necessary datasets we used them to find two sets of differentially expressed genes or *DEGs*. The first set contains DEGs found in different stages of cancer and the second the vitamin-D gene signature.

2.2.1 DEG between stage of cancer

To find DEgs between the stage of cancer we used 10 data matrices of 600 samples. This data matrices have been computed so that the samples were balanced between high and low stage. We identify stage low of cancer as stages 0, 1 and 2 and stage high as stages 3 and 4. For each of the 10 subsets we used LIMMA to find a ranking of the DEGs. The resulting gene lists have been sorted in increasing order by p-value. We aggregated the resulting list using Borda count, so to obtain a ranking for each of the 12644 genes. This ranking will be used to compare prediction power of the vitamin-D signature with respect to the top DEGs between stages.

2.2.2 Vitamin D gene signature

In order to obtain the list of differentially expressed genes for the vitamin D signature, the state-of-the-art R package "Deseq2" was used on the vitamin D dataset, which contains counts of RNA-seq expression data from rectal biopsies of CRC patients pre-and post- vitamin D supplementation. Transcript IDs were converted to Gene Symbols through "BioMart". For some of the transcripts, a Gene Symbol was not found. The statistically relevant genes were then selected using the adjusted p-value automatically computed by the Deseq2 package. DEGs were then manually expanded using the protein-protein interaction network database STRING. Enrichment analysis was performed on the original list using the web interface EnrichR.

2.3 Survival analysis

The survival analysis has been used to determine if both sets of DEGs we found upstream are responsible for a change in the probability of survival. This analysis has been performed in two steps. The first uses a cox proportional-hazards model and the second involves building the Kaplan-Meier curves for the significative gene found by cox on a different dataset.

2.3.1 Cox proportional-hazards model

Cox-proportional-hazards model is used to determine for each of the 12644 gene if they are significant in changing the overall survival and the corresponding hazard ratio *HR*. The hazard ratio determines how a gene is associated with the length

Dataset name	Number of samples	Dataset name	Number of samples
GSE39582	585	GSE33113	96
TGCA-COAD	438	GSE9348	82
GSE14333	290	GSE13067	74
GSE17536	177	GSE39084	70
GSE26682	176	GSE23878	59
GSE103479	156	GSE17537	55
GSE13294	155	GSE18088	53
GSE20916	145	GSE4183	53
GSE38832	122	GSE31595	37
GSE18105	111	GSE4107	22
GSE157982	98	GSE15960	18

Table 1. Distribution of the samples in the starting datasets. In total, 3072 samples from 22 datasets were obtained.

of survival. For each gene the optimal cutpoint for the level of expression will be computed using the *cutp* function. Only the genes with $pvalue \leq 0.05$ were considered for computing the Kaplan-Meier curves.

2.3.2 Kaplan-Meier curves

For each one of the significative genes found by cox a Kaplan-Meier curve was computed. For each gene we divided the samples in a "low" set if their expression level was less than the cutpoint found by the cutp function and in a "high" set if their expression level was more than the cutpoint. The gene were still considered downstream if the log-rank p-value was still ≤ 0.05. For each of this genes a Kaplan-Meier curve was computed for the sample in the low set and one for the ones in the high set. Furthermore this genes have been intersected with the enriched vitamin D signature and the DEG ranking.

3. Results

3.1 Data preprocessing

3.1.1 Sample splitting and filtering

The sets used to find the vitamin-D gene signature and the DEGs between stages of cancer needed only the expression level of each gene, while the other ones needed information about the time of survival and the status of the sample. We retained data that had all the information we needed, so we obtaine 1117 samples from the starting 3394, the 33%. Table 2 contains information on how we divided all the remaining samples of each dataset in our sets.

Set	n° of samples
COX fitting	232
KM curve	1000
Vitamine D gene signature	90
Stage DEGs	1000

Table 2. Split samples

3.1.2 Dataset normalization

After filtering the best normalization result were obtained using the fRMA algorithm, as can be seen in figures 1 and in 2

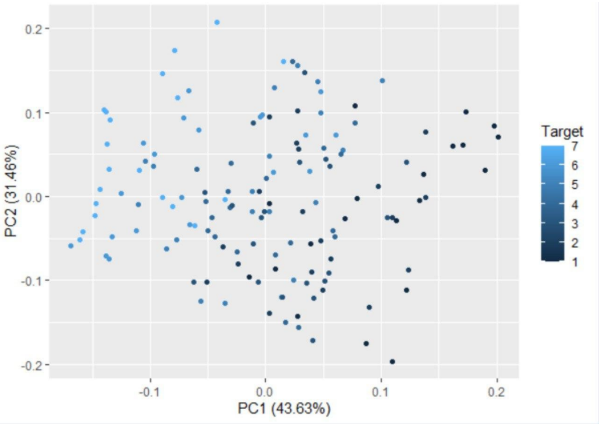

Figure 1. PCA on dataset normalized with RMA, each color represent a dataset

3.2 Differentially expressed genes

3.2.1 DEG between stage of cancer

After having computed the p-value by limma and having aggregated those results using borda count we obtained a ranking list of the 12644 genes used as input for cox. This ranking will be used to find the most differentially expressed genes between stage high and stage low of colorectal cancer and compare how they perform with respect to the one found in the vitamin D gene signature in predicting a patient risk.

3.2.2 Vitamin D gene signature

A serum level of vitamin D was chosen as a threshold for stratifying the two groups of patients (high vs low vitamin D level). This is because for a small subset of patients in the post-supplementation group, the serum levels were relatively low, whereas for another subset of patients in the pre-supplementation group, levels were already high. The chosen cutoff (figure 3) eventually led to better results than the classification based on the available label.

Principal component analysis on the full expression data matrix did not yield segregated clustering of the two groups (figure 4.

Using a threshold for the adjusted p-value of 0.05, 33 differentially expressed genes were selected from the raw

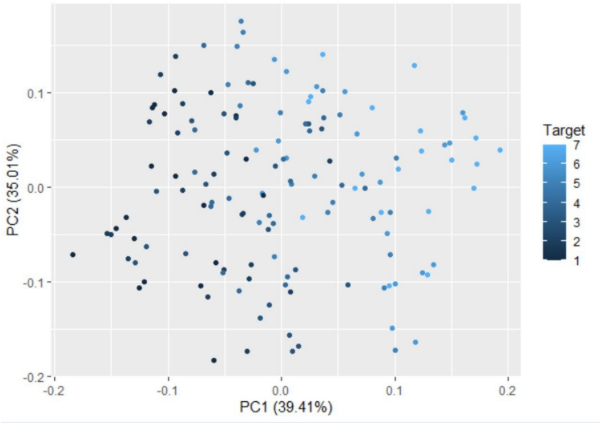

Figure 2. PCA on dataset normalized with fRMA, each color represent a dataset

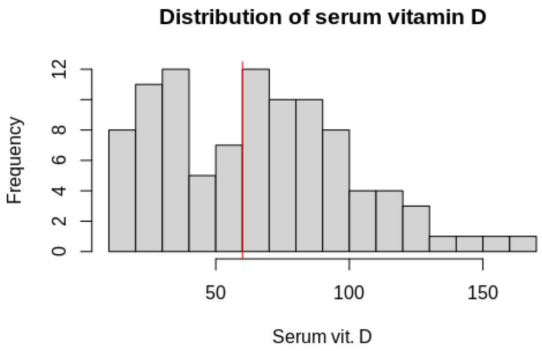

Figure 3. Distribution of serum vitamin D

output of the Deseq2 package.

3.2.3 Pathway enrichment

The top 100 differentially expressed transcripts for which a Gene Symbol could be found were given as input to EnrichR. Some of the most interesting results of the enrichment analysis on the top 100 genes of the vitamin-D gene signature in colorectal cancer are represented by the statistically significant (q-value < 0.05) enrichment of the PTEN signalling pathway (indipendently highlighted by 2 different databases: 'BioPlanet 2019' and 'BioCarta 2016'), and pathways associated with various integrins and cell motility in cancer ('Elselvier Pathway Collection' database), such as the alpha-4-beta-7 integrin signalling. PTEN (figure 5) is a negative regulator of the Akt\PkB pathway, which promotes cell proliferation and survival, and is often altered in cancer cells [a]. Alpha-4-beta-7 (coded by the ITGB7 gene) is an integrin involved in the recruiting of lymphocytes in the lymphoid tissue of the gut [b]. The content of tumor-infiltrating lymphocytes of different malignancies, including CRC, is strongly linked to better long term and short term survival [c].

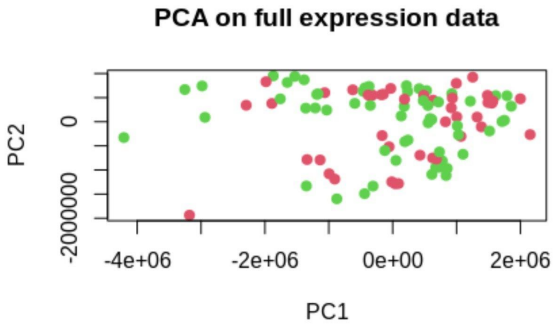

Figure 4. PCA on vitamin D dataset

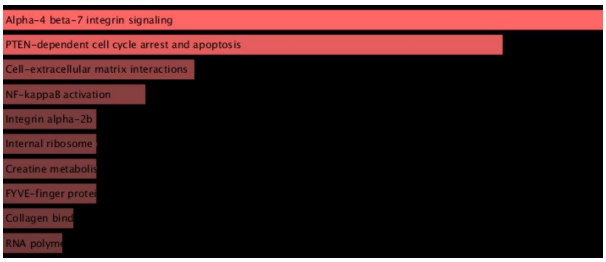

Figure 5. Pathway enrichment results

3.3 Survival analysis

3.3.1 Cox proportional-hazards model

Cox proportional-hazards model identified from the initial 12644 genes 943 genes, with an average HR of $1.49 \cdot 10^{18}$, a minimum of $2.7 \cdot 10^{-17}$ and a maximum of $1.4 \cdot 10^{21}$.

3.3.2 Kaplan-Meier curves

Of the 943 genes found significative by the cox proportional-hazards model 291 has been found significative on the dataset used to build the Kaplan-Meier curves. Of these 291 4 are found in the enriched vitamin D signature (left side of figure 6). To compare how this vitamin D related gene have an impact on the probability of survival we compared them with the top 4 DEGs identified between stages of cancer and found significative in this step (right side offigures 6).

For each of these gene the respective HR can be found in table 3

Vitamin D gene	HR	Stage DEG	HR
RPS27A	$1.4 \cdot 10^{8}$	SLC22A8	$3.3 \cdot 10^{6}$
TMCO3	$7.9 \cdot 10^{5}$	CDK17	$5.2 \cdot 10^{3}$
CHUK	88	DHRS12	$2.9 \cdot 10^{3}$
KLHL20	$1.6 \cdot 10^{-6}$	AMPD3	$2.6 \cdot 10^{-7}$

Table 3. HR for significative genes

4. Discussion

After having compiled the Kaplan-Meier curves the four genes derived from the consensus stage-based DEGs and the four

gene found from the expanded vitamin D gene signature found statistically significant, we compiled a literature-based characterization for each of those (tables 4 and 5).

Gene	Relevant function
RPS27A	One of the genes encoding for ubiquitin. Misregulated in various cancers, including colorectal [n]. Its upregulation in CRC may promote cancer cell proliferation and inhibition of apoptosis [o].
TMCO3	Probable Na(+) / H(+) antiporter. Linked to unfavorable outcomes in cancer [p]. A patent exists for prognosis and treatment methods of CRC based on this gene [q].
CHUK	Ser/Thr protein kinase involved in the (indirect) activation of NF-kB [f]. Regulates cyclin D1. [g] Decreased activity is linked to cancer [h].
KLHL20	[i] Mediates ubiquitination of DAPK1 thereby downregulating interferon-mediated apoptosis. [l] Mediates ubiquitination of PML thereby promoting resistance to hypoxia and cancer progression through HIF1a signalling. [m].

Table 4. Statistically significant genes in the expanded vitamin-D gene signature

Four DEGs were found to be related to the enrichment of vitamin D in CRC, namely RPS27A, TMCO3, CHUK, KLHL20 (in descending order of relevance, according to their HR). The HR values of these 4 genes were compared to the HR of the top 4 DEGs related to CRC's stage; also in descending order: SLC22A8, CDK17, DHRS12, AMPD3. The HR value is higher for the vitamin D counterpart of the analyzed DEGs, which implies that different concentrations of this vitamin influences gene expression such that a larger impact on prognosis of CRC patients is present (compared to the different expression of genes according to the stage of the disease).

Only for the most relevant vitamin D related DEG (SLC22A8), 3 drugs were found to target it according to the literature, namely ATALUREN, ELX-02, MT-3724. Similarly, for the two most relevant stage related DEGs (SLC22A8, CDK17), several drugs were found in the literature: PROBENECID (for SLC22A8); AZD-5438, PHA-793887, RONICICLIB, AT-7519 (for CDK17).

Further analysis on these DEGs is of interest, to verify their role in CRC evolution and prognosis, and potentially improve the treatment of CRC patients in the future. This work can be further expanded considering a multiple regression analysis that considers the combined effect of the vitamin-D related genes.

Gene	Relevant function
SLC22A8	Integral membrane protein involved in the sodium-dependent excretion of potentially toxic organic anions. Expression specific to kidney and brain [e].
CDK17	Cyclin-dependent protein serine/threonine kinase. Involved in the regulation of transcription involved in G1/S transition of mitotic cell cycle (source: GO biological process). Expressed in many tissues (no specificity) [b]. It is the target of 4 CDK inhibitors.
DHRS12	Oxidoreductase. Linked to poor prognosis in ovarian cancer [c] and suppression of proliferation and metastasis in osteosarcoma [d].
AMPD3	AMP deaminase in erythrocytes. In mice, mutations on this gene reduce levels of naive CD4+ and naive CD8+ cells in peripheral blood but not in lymphoid tissue. This is most likely due to a signalling mechanism triggered by the mutated phenotype of the red blood cells. [a].

Table 5. 4 top genes in stage DEGs

References

[1] Home - GEO - NCBI.

[2] GDC.

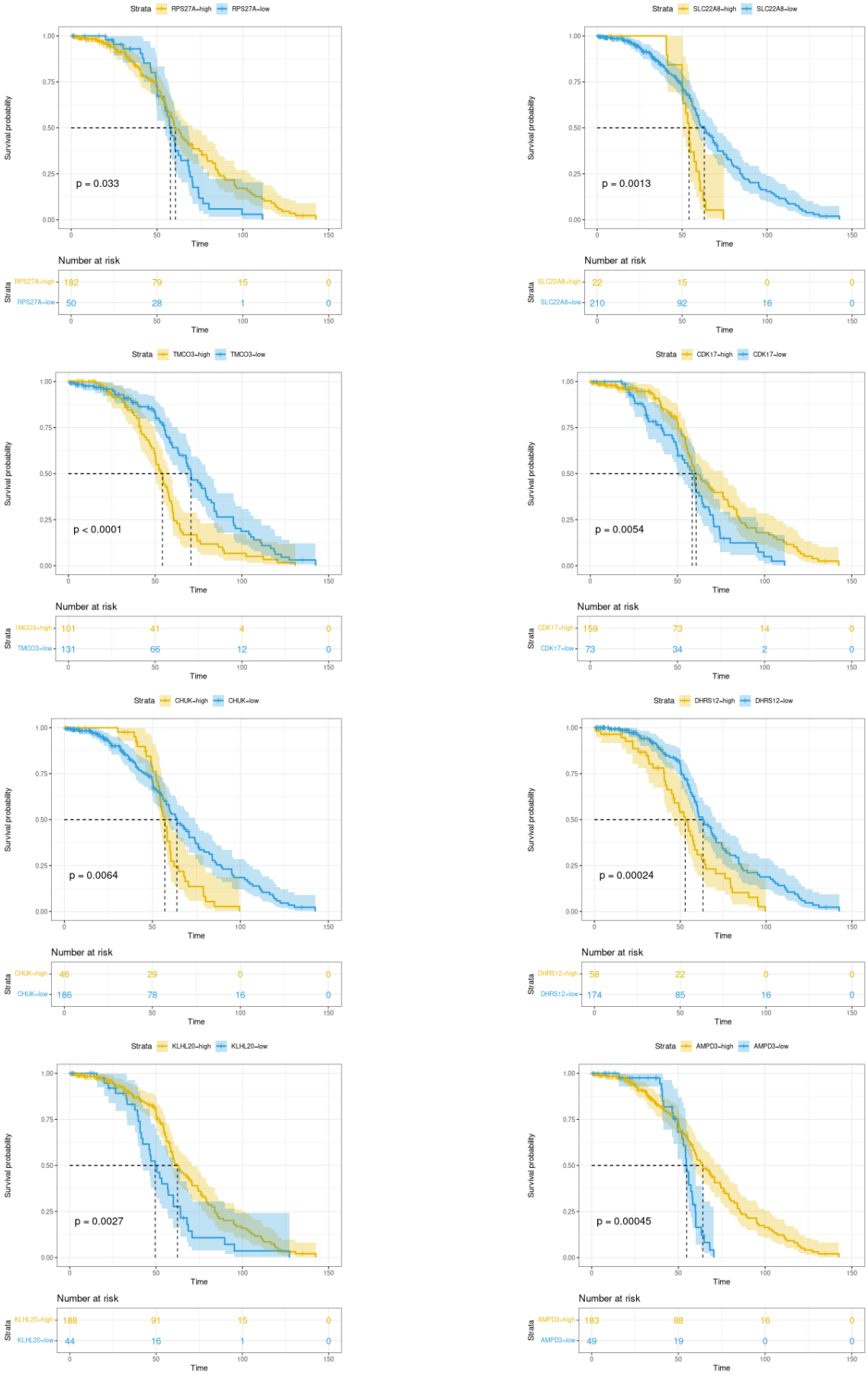

Figure 6. Survival curves, on the left vitamin D related genes, on the right the top stage DEGs