Laboratory of biological data mining
Project report

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

[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

2. Material and methods

2.1 Data preprocessing

This project requires a lot of data to achieve high level of statistica significance, so we decided to start with an high number of samples, taken from different dataset. A brief description of each dataset considered for the pipeline can be found in table 1.

2.1.1 Sample splitting and filtering

The first step of the project's pipeline is to filter the datasets in order to obtain a list of the samples having all the data necessary for the downstream analyses. So, analysing all of the datasets' metadata we split all the samples in 6 sets according to which clinical data was available. After this operation we retained 2676 samples out of the starting 3812 (70%) divided as in table 2.

Set usage	n° of samples
COX fitting	388
KM curve	157
Vitamine D low	49
Vitamine D high	49
Stage low	1120
Stage high	908

Table 2. Split samples

2.1.2 Dataset normalization

2.2 Differentially expressed genes

2.2.1 DEG between stage of cancer

2.2.2 Vitamin D gene signature

2.2.3 Pathway enrichment

2.3 Survival analysis

3. Results

3.1 Data preprocessing

3.1.1 Sample splitting and filtering

3.1.2 Dataset normalization

3.2 Differentially expressed genes

3.2.1 DEG between stage of cancer

3.2.2 Vitamin D gene signature

3.2.3 Pathway enrichment

3.3 Survival analysis

4. Discussion

References

Dataset name	Sample description	Number of samples
E-MTAB-6698	healthy and tumor colorectal samples	1566
GSE157982	baseline and vit. D-treated CRC rectal samples	98
GSE38832	tumor colorectal samples	122
TCGA-COAD	tumor colorectal samples	438
GSE14333	tumor colorectal samples	290
GSE17536	tumor colorectal samples	177
GSE31595	tumor colorectal samples	37
GSE33113	tumor colorectal samples	96
GSE38832	tumor colorectal samples	122
GSE39084	tumor colorectal samples	70
GSE39582	tumor colorectal samples	585
GSE103479	tumor colorectal samples	156
GSE17537	tumor colorectal samples	55

Table 1. Starting datasets