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 a 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 aimed 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 preprocessing of our datasets: a colorectal cancer gene expression data datasets and a vitamin D level gene expression data datasets. 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 will obtain a dataset of differential expressed genes in different stages of colorectal cancer and a dataset of vitamin D gene signature. After that, we will intersect these two datasets in order to obtain our genes of interest. With this list of genes we will be able to perform a pathway enrichment analysis in order to extend our list of genes of interest. At this point, we will be able to compare the level of expression of genes of interest with patient data and fit a regression model for patient data prediction. This regression analysis, once validated will allow us to obtain a list of genes significative for the 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

Introduction

Colorectal cancer (CRC) is the third most common malignant tumor worldwide and is the second one in cancer-related deaths [1]. 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 [2]. There are lot of epidemiological and preclinical studies that indicate a beneficial effect of vitamin D on CRC incidence and mortality [3] [4]. Vitamin D is a fat-soluble vitamin and many genes are related to its metabolism and action [5]. It can be obtained from diet or the endogenous synthesis in the epidermis under sunlight exposure [6]. It has been demonstrated that vitamin D benefits clinical outcome and improves the long-term survival of CRC patients [7]. Moreover, circulating vitamin D may be a CRC biomarker and is deficiency is related to the high incidence of CRC [8]. A better survival outcome in CRC is associated with higher prediagnostic or postdiagnostic serum 25-hydroxyvitamin D concentrations [9]. The most active vitamin D metabolite (1α,25-dihydroxyvitamin D3) 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 [3] [10] [11]. Lots of genes are 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 [12]. Vitamin D and its related genes are correlated with the homeostasis of the intestinal epithelium, regulate immune cells [13]

1. Biological question

The objective of this project is to try to find a way to make prognosis and stratify patients suffering from colorectal cancer by means of their transcriptomic profiles. In particular we are focusing on the gene signature of vitamin D as a prognostic marker. To do so we are leveraging the always increasing gene expression analysis data publicly available. The data will be used to build and validate a statistical model that will be able to stratify patients according to their survival ability. Moreover we are hoping to identify in this process some colorectal cancer survival markers related with vitamin D effects and pathways involved.

2. Data

Where we are looking for data, what data we have chosen and why, how it is going to be splitted. Information about each dataset as in table 1

3. Pipeline

The project pipeline is described in figure 1.

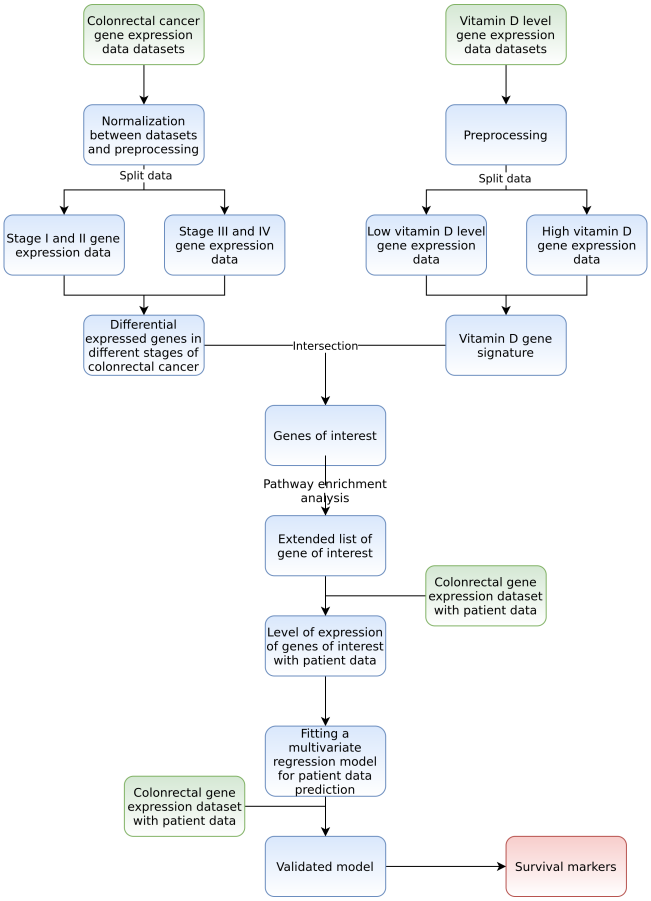

Figure 1. Project pipeline

3.1 Data preprocessing

3.1.1 Normalization

Normalization is done to eliminate the batch effect because we are working with different datasets. We will try different algorithms and evaluate them to discover the better one.

3.1.2 Data filtering

Filter out samples without survival data or withouc comparable data distribution after normalization.

3.1.3 Evaluation of pre-processed data

The evaluation can be done using PCA, linear regression or hierarchical cluster analysis.

3.2 Obtaining the list of gene of interest

3.2.1 Differential expressed genes between different stages of CRC

Split data in stage I and II gene expression data and in stage III and IV gene expression data, in order to obtain differential expressed genes in different stages of colorectal cancer.

3.2.2 Vitamin D gene signature in CRC

Split data in low vitamin D level gene expression data and in high vitamin D level gene expression data, in order to obtain vitamin D gene signature.

3.2.3 Intersection and enrichment

Intersection of the obtained datasets in order to obtain a subset of genes of interest. Pathway enrichment analysis for the identification of the classes of genes over-represented in our subset of genes.

3.3 Fitting to a regression model

Fitting a single variable or multivariate regression model for patient data in order to asses for each gene of interest its correlation with the stratification of the patient.

3.4 Obtaining of the survival markers

Retrieving from the fitted model those gene that presents a significative impact on patient stratification.

4. Expected results

A list of gene usable as survival marker in CRC. According to (cite studies) possible candidates are:

5. Project management

The project will be developed in a dynamic and teamwork driven manner. This is done in order to be able to leverage the particular skills and knowledge of each member. Each

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

Table 1. Datasets used

Gene	Function
CYP24A1	temp
TGFB1	temp
IGFBP2	temp
CYP24A1	temp
CH25H	temp
IGFLR1	temp
DCBLD2	temp
PTPN14	temp
SLC10A2	temp
FGF2	temp

Table 2. Possible survival marker as found in literature

task will be divided into the members in subtasks tailored according to his or hers background. There will be frequent meetings with at least bi-daily cadence to asses progress and resolve problems. Times estimates for each part are outlined in the chart **??**

Acknowledgments