

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

Diego Barquero Morera¹, Giacomo Fantoni², Gaia Faggini³, Leonardo Golinelli⁴

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 we 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.faggini@studenti.unitn.it

⁴ leonardo.golinelli@studenti.unitn.it

Contents

Introduction	1
1 Biological question	2
2 Data	2
3 Pipeline	2
3.1 Data preprocessing	2
Normalization • Data filtering • Evaluation of pre-processed data	
3.2 Obtaining the list of gene of interest	2
Differential expressed genes between different stages of CRC •	
Vitamin D gene signature in CRC • Intersection and enrichment	
3.3 Fitting to a regression model	3
3.4 Obtaining of the survival markers	3
4 Expected results	3
5 Project management	3
References	3

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 its 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 (1 α ,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 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 and 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

In order to improve the statistical significance of the results, we plan to use multiple datasets each with a specific purpose. To limit batch effects, the idea is to jointly analyze them only when strictly necessary, after an appropriate pre-processing procedure, as described in the pipeline section. We plan to identify the gene signature of vitamin D in CRC using a dataset created for this specific purpose (GSE157982), whereas for the gene signature of CRC (w.r.t. control) we will use a separate, independent dataset (E-MTAB-6698). For the purpose of training and validating the regression model, we will use datasets that include relevant clinical information used for prognosis such as survival time and tumor stage. The specific allocation of the datasets for model training and validation will be decided later on. For more detailed information about the datasets used, please refer to table 1.

3. Pipeline

The project pipeline is described in figure 1. This pipeline is been determine following the work done in [?] and can be separated in four major steps:

1. Data preprocessing
2. Identification of a list of gene of interest.
3. Fitting to a model.
4. Obtaining the survival markers.

3.1 Data preprocessing

3.1.1 Normalization

This pipeline requires lots of samples, so we are retrieving them from different datasets. Because of this a normalization step is necessary: the batch effect could affect the accuracy of our predictions introducing noise in the data. We will try

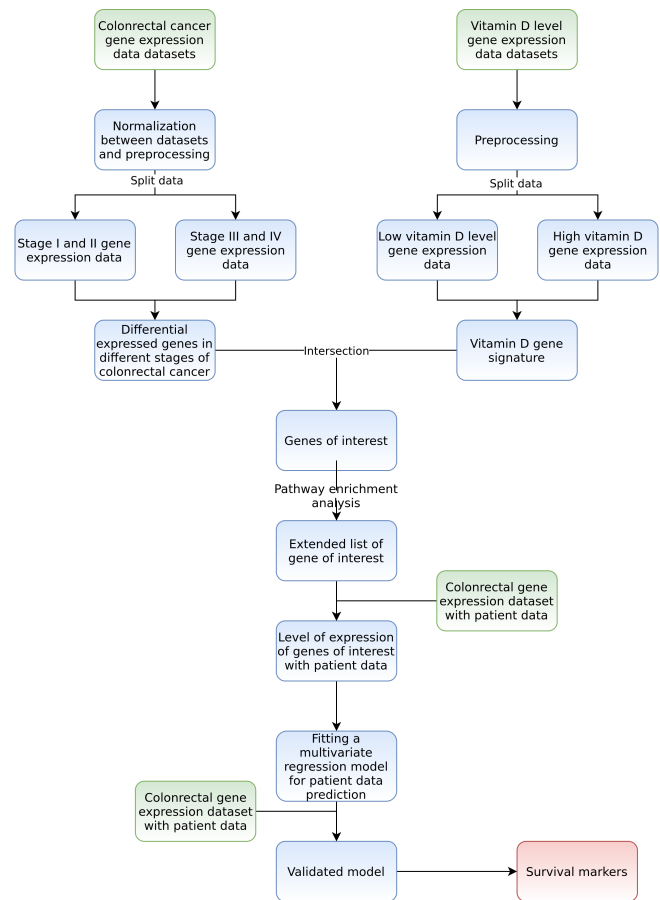


Figure 1. Project pipeline

to normalize the data using different algorithms and evaluate them so to use the one that will leave the least noise.

3.1.2 Data filtering

After normalization we will filter out samples without survival data or with non-comparable data distribution.

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

Input data will be divided 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 or DEGs in different stages of colorectal cancer.

3.2.2 Vitamin D gene signature in CRC

Input data will be divided 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.

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

3.2.3 Intersection and enrichment

Obtained lists of DEGs and vitamin D gene signature will be intersected in order to obtain a subset of genes of interest. This list will be then expanded through 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

The extended list will be used to obtain gene expression level in each sample. After that these data will be used to fit a regression model in order to assess for each gene of interest its correlation with the stratification of the patient.

3.4 Obtaining of the survival markers

After model fitting the list of genes will be reduced so that it contains only genes with a significant impact on patient stratification.

4. Expected results

With this project we expect to identify a list of genes usable as survival markers for colorectal cancer. A list of putative genes is:

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

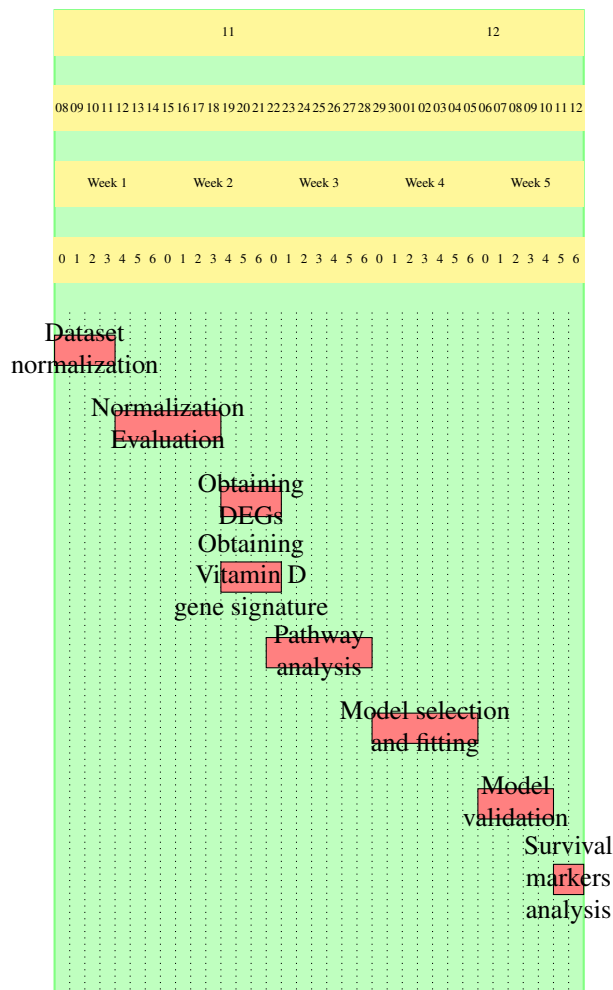
Table 2. Putative Vitamin-D -related CRC prognostic marker genes

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 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 assess progress and resolve problems. Times estimates for each part are outlined in the chart ??.

References

- [1] J Ferlay, M Colombet, I Soerjomataram, T Dyba, G Randi, M Bettio, A Gavin, O Visser, and F Bray. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer*, 103:356–387, November 2018.
- [2] Luping Bu, Fengxing Huang, Mengting Li, Yanan Peng, Haizhou Wang, Meng Zhang, Liqun Peng, Lan Liu, and Qiu Zhao. Identification of vitamin d-related gene signature to predict colorectal cancer prognosis. *PeerJ*, 9(e11430):e11430, May 2021.
- [3] Fábio Pereira, María Jesús Larriba, and Alberto Muñoz. Vitamin D and colon cancer. *Endocr. Relat. Cancer*, 19(3):R51–71, June 2012.
- [4] Edward Giovannucci. Epidemiology of vitamin D and colorectal cancer. *Anticancer Agents Med. Chem.*, 13(1):11–19, January 2013.
- [5] Veronika Fedirko, Hannah B Mandle, Wanzhe Zhu, David J Hughes, Afshan Siddiq, Pietro Ferrari, Isabelle Romieu, Elia Riboli, Bas Bueno-de Mesquita, Fränzel J B van Duynhoven, Peter D Siersema, Anne Tjønneland, Anja Olsen, Vittorio Perduca, Franck Carbonnel, Marie-Christine Boutron-Ruault, Tilman Kühn, Theron Johnson, Aleksandrova Krasimira, Antonia Trichopoulou, Periklis Makrythanasis, Dimitris Thanos, Salvatore Panico,



san Johnston, Albert Tenesa, Harry Campbell, and Malcolm G Dunlop. Plasma vitamin D concentration influences survival outcome after a diagnosis of colorectal cancer. *J. Clin. Oncol.*, 32(23):2430–2439, August 2014.

- [10] David Feldman, Aruna V Krishnan, Srilatha Swami, Edward Giovannucci, and Brian J Feldman. The role of vitamin D in reducing cancer risk and progression. *Nat. Rev. Cancer*, 14(5):342–357, May 2014.
- [11] H G Pálmer, J M González-Sancho, J Espada, M T Berciano, I Puig, J Baulida, M Quintanilla, A Cano, A G de Herreros, M Lafarga, and A Muñoz. Vitamin d(3) promotes the differentiation of colon carcinoma cells by the induction of e-cadherin and the inhibition of beta-catenin signaling. *J. Cell Biol.*, 154(2):369–387, July 2001.
- [12] Hossein Sadeghi and Mohammad Heiat. A novel circular RNA hsa_circ_0060927 may serve as a potential diagnostic biomarker for human colorectal cancer. *Mol. Biol. Rep.*, 47(9):6649–6655, September 2020.
- [13] Jorge Martinez-Romero, Santiago Bueno-Fortes, Manuel Martín-Merino, Ana Ramirez de Molina, and Javier De Las Rivas. Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genomics*, 19(Suppl 8):857, December 2018.

Vittorio Krogh, Carlotta Sacerdote, Guri Skeie, Elisabete Weiderpass, Sandra Colorado-Yohar, Núria Sala, Aurelio Barricarte, Maria-Jose Sanchez, Ramón Quirós, Pilar Amiano, Björn Gylling, Sophia Harlid, Aurora Perez-Cornago, Alicia K Heath, Konstantinos K Tsilidis, Dagfinn Aune, Heinz Freisling, Neil Murphy, Marc J Gunter, and Mazda Jenab. Vitamin d-related genes, blood vitamin D levels and colorectal cancer risk in western european populations. *Nutrients*, 11(8):1954, August 2019.

- [6] Vrinda Saraff and Nick Shaw. Sunshine and vitamin D. *Arch. Dis. Child.*, 101(2):190–192, February 2016.
- [7] Shu Xu, Yao He, Lihui Lin, Peng Chen, Minhu Chen, and Shenghong Zhang. The emerging role of ferroptosis in intestinal disease. *Cell Death Dis.*, 12(4):289, March 2021.
- [8] Stacey Meeker, Audrey Seamons, Lillian Maggio-Price, and Jisun Paik. Protective links between vitamin d, inflammatory bowel disease and colon cancer. *World J. Gastroenterol.*, 22(3):933–948, January 2016.
- [9] Lina Zgaga, Evropi Theodoratou, Susan M Farrington, Farhat V N Din, Li Yin Ooi, Dominik Glodzik, Su-