Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/16/20240529

A tumor specific Bayesian framework reveals novel molecular subtypes in colorectal cancer

Chenxuan Han

Justin Siena High School, 4038 Maher St, Napa, CA, 94558, United States

louis.han023@gmail.com

Abstract. We applied a tumor-specific Bayesian framework to investigate the drivers and signaling mechanisms behind various subtypes of Colorectal Cancer (CRC), a highly aggressive cancer known to have diverse origins. Our approach aims to identify the cancer drivers that contribute to the development of colorectal cancer (CRC) within individual tumors. By inferring the target differentially expressed genes (DEGs) associated with these drivers, we effectively group patients into distinct molecular subtypes. We employed the tumor-specific causal inference (TCI) model to establish causal associations between somatic genome alterations (SGAs) and differentially expressed genes (DEGs) inside each colorectal cancer (CRC) tumor. Through the process of generalization, we have successfully discovered three distinct mechanism-oriented subtypes of colorectal cancer (CRC) by examining the most statistically significant SGAs and their corresponding target DEGs throughout the CRC cohort. Notably, this subtyping approach stands independently from the previously reported transcriptomic-based molecular subtyping of CRC. Additionally, our analysis successfully grouped patients based on significant prognostic outcomes, outperforming the previously reported subtyping. This research provides valuable insights into understanding the underlying drivers and molecular complexities associated with different CRC subtypes.

Keywords: Bayesian Network, CRC Subtyping, Cancer Drivers, Survival Analysis.

1. Introduction

Colorectal cancer is a complex disease that can be driven by genetic and molecular alterations in the cells that line the colon or rectum. These alterations can lead to changes in cell growth and division, which can contribute to the development and progression of colorectal cancer [1, 2]. Recent research has shown that colorectal cancer is a heterogeneous disease, meaning that it can be subdivided into different molecular subtypes based on the specific genetic and molecular alterations that drive each tumor [3]. This molecular complexity can impact the behavior of cancer and its response to treatment [4]. Despite the various subtyping approaches used in research and clinical practice for colorectal cancer, there are still challenges in identifying the most effective molecular subtypes. This can lead to difficulties in comparing results across studies and developing a standardized approach to molecular subtyping, hindering the identification of the most effective treatment options for patients with colorectal cancer [4, 5].

Colorectal cancer is a heterogeneous disease that can be classified into different subtypes based on various features, such as molecular and pathological characteristics. Several molecular subtyping

DOI: 10.54254/2753-8818/16/20240529

approaches have been proposed for colorectal cancer, including the consensus molecular subtypes (CMS), the molecular subtypes based on somatic copy-number alterations (SCNA), and the molecular subtypes based on gene expression profiles [3]. The CMS classification of CRC into distinct subtypes is based the analysis of gene expression patterns and is extensively employed for the categorization of colorectal cancer into four distinct subtypes: CMS1 (immune), CMS2 (canonical), CMS3 (metabolic), and CMS4 (mesenchymal). CMS1 tumors exhibit a notable presence of immune cells infiltrating the tumor microenvironment. In contrast, CMS2 tumors are distinguished by the activation of the Wnt signaling pathway. CMS3 tumors are characterized by dysregulation of metabolic processes, while CMS4 tumors are marked by the activation of stromal cells. In conclusion, molecular subtyping of colorectal cancer has advanced our understanding of the disease and has the potential to improve clinical outcomes [4].

The present study employed the tumor-specific causal inference (TCI) model for identifying the main driver genes and their corresponding differentially expressed genes (DEGs) in colorectal cancer (CRC) tumors using The Cancer Genome Atlas (TCGA) dataset. [6]. By analyzing the expression patterns of DEGs, we discovered three distinct molecular subtypes of CRC that provide valuable insights into patient prognostic outcomes and are distinct from previous classification methods. Our approach introduces a novel framework for comprehending the mechanisms underlying tumorigenesis in CRC.

2. Method

2.1. Data collection and preprocessing

Genomic and transcriptomic data of 430 colorectal cancer (CRC) tumor samples were obtained from the Xena platform, as part of the TCGA project (https://xenabrowser.net/) [7]. The integration of mutation and copy number data using the GISTIC2 algorithm was performed to generate somatic genome alterations (SGAs). In this process, a gene within each tumor sample was classified as altered if it was affected by either a somatic mutation (SM) or/and a somatic copy number alteration (SCNA) event. DEGs were identified by comparing gene expression in tumor cells and normal cells. We regarded gene expression to be significantly different in a tumor if its p value was less than 0.005, if gene expression in normal cells follows a Gaussian distribution. The binary matrix used in this study was created to illustrate the relationship between tumors and genes. In this matrix, a value of 1 indicates an expression change, while a value of 0 indicates no change.

2.2. Tumor-specific causal inference algorithm

Tumor-specific causal inference (TCI) is a Bayesian paradigm that integrates various genomic data types to infer causal links between genome alterations and molecular phenotypic changes for each individual tumor. The objective of TCI is to determine a tumor-specific causal model that has the highest posterior probability, based on a dataset that includes somatic genome alterations (SGAs) and differentially expressed genes (DEGs). TCI scores potential causal arcs between SGAs and DEGs based on their posterior probabilities using a Bayesian framework. Further information regarding the TCI approach can be accessed in the original TCI paper [6].

3. Results

3.1. The TCI Method Identifies Major Colorectal Cancer Driver Genes and Their DEGs Targets

The Bayesian causal discovery methodology, known as TCI, was employed in our study to investigate the functional consequences of SGAs (gene mutations and copy number variations) in the regulation of DEGs (differentially expressed genes) inside individual tumors. By integrating heterogeneous genomic data types, we conducted an analysis on the genomic and transcriptome data of 430 colorectal cancer (CRC) tumors from The Cancer Genome Atlas (TCGA). As a result, we discovered 105 driver genes and 1,005 target differentially expressed genes (DEGs) that are regulated by these drivers (Supplementary Table 1).

DOI: 10.54254/2753-8818/16/20240529

TCI discovered a total of 105 drivers, consisting of many well-known CRC drivers including APC, TP53, KRAS, BRAF, NRAS, PIK3CA, PTEN, CTNNB1, FBXW7, RNF43, SMAD4, ARID1A, and EGFR. For example, APC, RNF43 and CTNNB1 act as key WNT signaling pathway regulators [8-10]; TP53, FBXW7, SMAD4 and ARID1A function as tumor suppressors by regulating cell cycle and DNA damage responses [7, 11-13]. Our method also uncovered genes are not specifically known as colorectal cancer driver genes but have potential functions in cancer development or are associated with other cancer types. For example, ASXL1 mutations is involved in chromatin remodeling as well as transcriptional control in myeloid malignancies such as myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML) [13]. Mutations of the BCOR gene is involved in transcriptional repression and gene regulation in a variety of cancers, including acute myeloid leukemia (AML) and clear cell sarcoma of the kidney [14]. FHIT is a tumor suppressor gene that is involved in DNA repair, cell cycle control, and apoptosis, and is frequently deleted or silenced in cancers such as lung, pancreatic, and esophageal cancer [15]. CTCF is a transcription factor that regulates gene expression and chromatin architecture in a variety of malignancies, including breast cancer, bladder cancer, and glioma [16].

3.2. The Molecular Profiles Derived from TCI Can Predict Substantial Differences in Prognostic Outcomes among Subtypes of Colorectal Cancer Patients

Colorectal cancer is a genetic disorder that exhibits a significant level of diversity among tumors in terms of genomic alterations and molecular/cellular characteristics. The goal is to identify patterns of transcriptomic changes and genomic alterations for subtypes of CRC with shared disease mechanisms. Initially, we attempted to group TCGA CRC tumors into three clusters based on the similarity of their genomic alterations, but no clear patterns or significant differences in patient survival were observed. To explore disease mechanisms and patient outcomes further, we focused on the 1,004 target DEGs of 105 drivers identified by TCI. By employing the gene expression profile of these DEGs as input features, we successfully identified discernible patterns of expression that categorized patients into three distinct groups, as illustrated in Figure 1.

We conducted a comparison of our three CRC subtypes with the previously reported TCGA CRC 4 CMS molecular subtyping and examined the patient overlap between the two studies. Figure 1 illustrates that CRC patient subgroup 3 significantly aligns with CMS4, whereas patient subgroup 2 primarily correlates with CMS2. Notably, patient subgroup 1 is an intriguing mix of all CMS groups, predominantly comprising CMS1 and CMS3. As we continued our investigation into the genomic alterations of well-known CRC drivers, we discovered a more pronounced enrichment of BRAF and RNF43 in patient subgroup 1 as opposed to subgroup 2 and 3. This finding suggests that these two drivers may play a significant role in regulating the gene expression specific to subgroup 2.

DOI: 10.54254/2753-8818/16/20240529

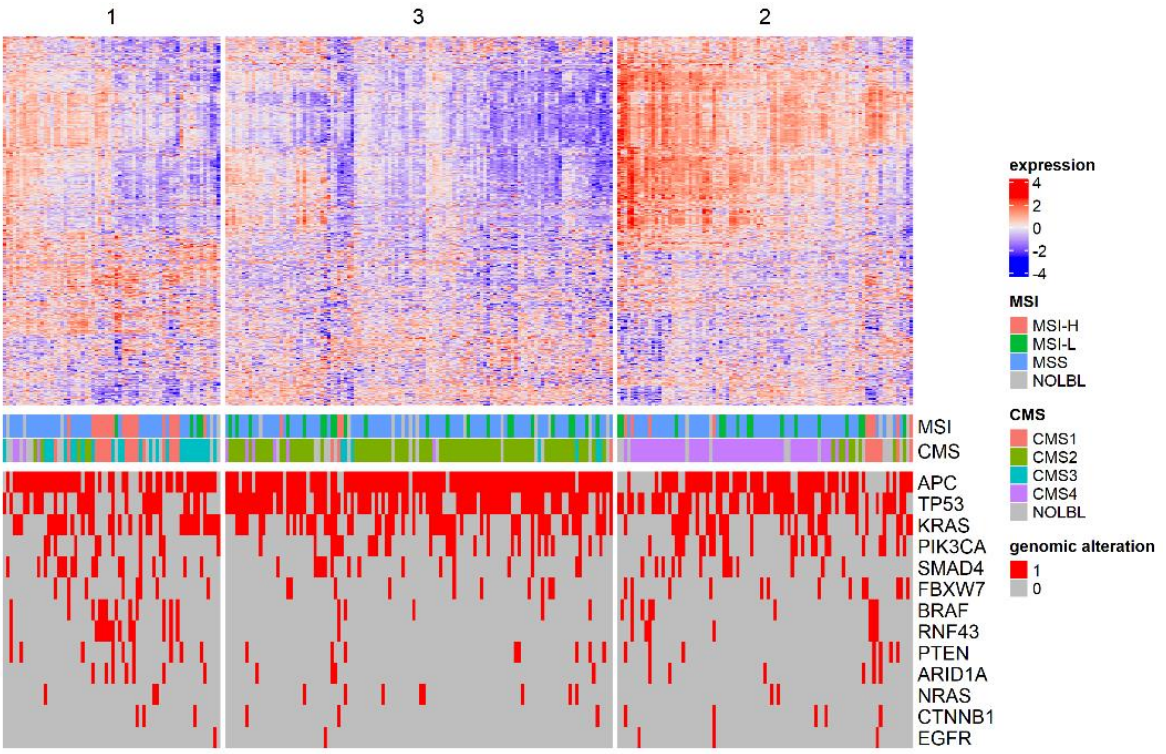

Figure 1. Heatmap of 1,004 gene expression across 430 CRC patients. The patients were classified into three groups through consensus clustering. The heatmap represents genes as rows and patients as columns. The annotation bar displays MSI status and CMS subgroups. The mutation status of well-known CRC drivers is visualized, with red indicating genomic alteration events and grey representing the wild type.

Furthermore, it is worth noting that the subtypes displayed notable variations in terms of survival outcomes, with a statistically significant *p*-value of 0.01, as illustrated in Figure 2. This observation suggests that the 1,004 DEGs that are causally regulated by the 105 TCI derived drivers give a succinct and efficient representation of CRC tumor attributes in relation to the biological pathway perturbations. Patients in subgroup 1 had a survival rate of 100%, whereas patients in subgroup 2 demonstrated the most diminished survival rate. The BRAF gene is a prominent oncogene that is commonly implicated in microsatellite instability-high colorectal cancer [17]. BRAF mutations are often associated with Microsatellite Instability-High (MSI-H) colorectal cancer (CRC). Tumors classified as MSI-H exhibit a notable prevalence of mutations occurring in short, repetitive DNA sequences known as microsatellites. These mutations are caused by defects in the DNA mismatch repair system. Patients with colorectal cancer (CRC) who exhibit Microsatellite Instability-High (MSI-H) typically experience a more favorable prognosis in terms of survival as compared to those with Microsatellite Stable (MSS) or Microsatellite Instability-Low (MSI-L) CRC. Notably, as illustrated in Figure 1, BRAF alteration status is associated with MSI-H, but these patients are classified into subgroups 1 and 2 with significant disparities in survival outcome. These results suggest that additional features may contribute to significant disparities in survival outcomes of MSI-H patients between the two subgroups.

DOI: 10.54254/2753-8818/16/20240529

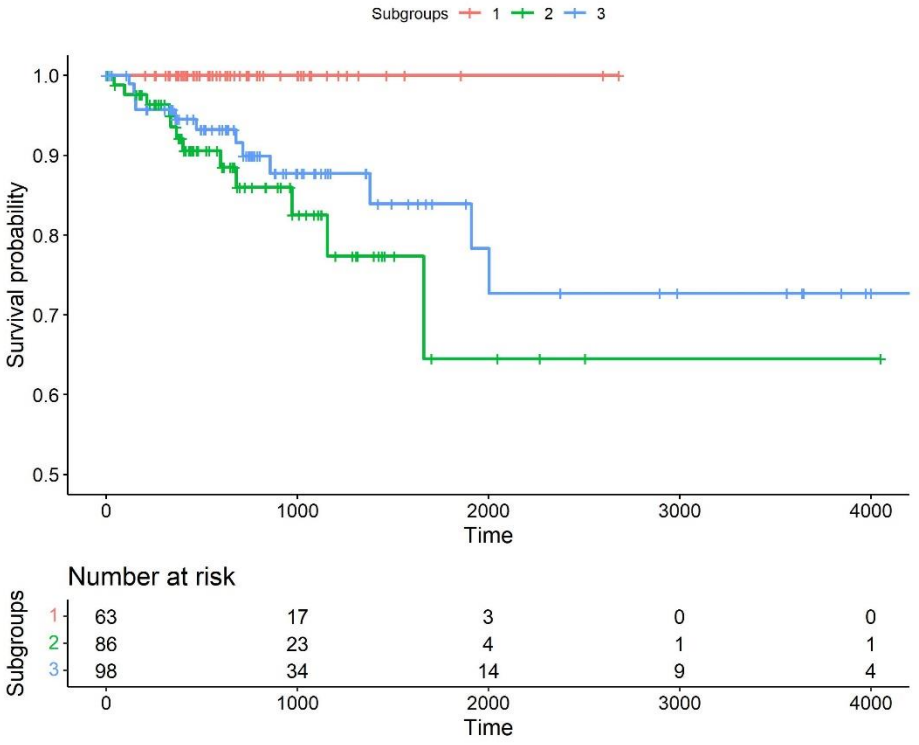

Figure 2. Overall survival of colorectal cancer patients. Kaplan-Meier survival plot illustrates the overall survival of 3 CRC patients. The x-axis represents the survival time in days, while the y-axis shows the survival probability.

4. Conclusion and discussion

Combining transcriptome and protein perturbations, we developed and evaluated a novel method for identifying signaling mechanisms in individual CRC tumors. The TCI-derived drivers and their causal DEGs allowed for the classification of CRC tumors into distinct subgroups with substantial survival differences. The combination of transcriptome and genomic alteration data offers a patient-centric perspective on colorectal cancer, hence presenting opportunities for the advancement of tailored treatment.

The TCI method has several advantages over traditional approaches. The TCI framework differentiates between driver and passenger SGAs and makes inferences about causative links at the individual tumor level. This allows researchers to explore disease mechanisms that are specific to each tumor and identify processes that are common across different subtypes of tumors. On the other hand, conventional molecular subtyping can be influenced by non-oncogenic factors and cell origins, which can potentially confound the analysis. Additionally, TCI provides insight into the gene expression profile of DEGs predicted by TCI. This information is essential for clinical decision-making in precision medicine, as it provides the basis for predicting drug sensitivity and repurposing.

References

[1] Markowitz, S.D. and M.M. Bertagnolli, *Molecular origins of cancer: Molecular basis of colorectal cancer*. N Engl J Med, 2009. 361(25): p. 2449-60.

[2] Vogelstein, B., et al., *Genetic alterations during colorectal-tumor development.* N Engl J Med, 1988. 319(9): p. 525-32.

[3] Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer*. Nat Med, 2015. 21(11): p. 1350-6.

DOI: 10.54254/2753-8818/16/20240529

[4] Sadanandam, A., et al., *A colorectal cancer classification system that associates cellular phenotype and responses to therapy.* Nat Med, 2013. 19(5): p. 619-25.
[5] Dienstmann, R., et al., *Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study.* Ann Oncol, 2017. 28(5): p. 1023-1031.
[6] Cai, C., et al., *Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference.* PLoS Comput Biol, 2019. 15(7): p. e1007088.
[7] Akhoondi, S., et al., *FBXW7/hCDC4 is a general tumor suppressor in human cancer.* Cancer Res, 2007. 67(19): p. 9006-12.
[8] Joslyn, G., et al., *Identification of deletion mutations and three new genes at the familial polyposis locus.* Cell, 1991. 66(3): p. 601-13.
[9] Koo, B.K., et al., *Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors.* Nature, 2012. 488(7413): p. 665-9.
[10] Morin, P.J., et al., *Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC.* Science, 1997. 275(5307): p. 1787-90.
[11] Fleming, N.I., et al., *SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer.* Cancer Res, 2013. 73(2): p. 725-35.
[12] Vazquez, A., et al., *The genetics of the p53 pathway, apoptosis and cancer therapy.* Nat Rev Drug Discov, 2008. 7(12): p. 979-87.
[13] Zhao, S., et al., *Roles of ARID1A variations in colorectal cancer: a collaborative review.* Mol Med, 2022. 28(1): p. 42.
[14] Astolfi, A., et al., *BCOR involvement in cancer.* Epigenomics, 2019. 11(7): p. 835-855.
[15] Pekarsky, Y., et al., *FHIT: from gene discovery to cancer treatment and prevention.* Lancet Oncol, 2002. 3(12): p. 748-54.
[16] Debaugny, R.E. and J.A. Skok, *CTCF and CTCFL in cancer.* Curr Opin Genet Dev, 2020. 61: p. 44-52.
[17] Barras, D., *BRAF Mutation in Colorectal Cancer: An Update.* Biomark Cancer, 2015. 7(Suppl 1): p. 9-12.