Journal of Research High School

Vol. 2023 (2) 347 – 352
ISSN 2688-3651 [online]

Investigate the Link Between the hMLH1 Gene and Microsatellite Instability (MSI) in Colorectal Carcinoma (CRC)

Sareena Garg,[1 *] Jeru Manuel[2]

[1]Notre Dame High School, San Jose, CA, USA
[2]MENTORCONNECT
*Corresponding Author: sgarg24@ndsj.org

Advisor: Dr. Jeru Manuel, jerumm@gmail.com

Received June 19, 2023; Revised September 26, 2023; Accepted, November 1, 2023

Abstract

Colorectal carcinoma is one of the most common cancers in the world, with the primary causal factor being genomic instability. Microsatellite Instability is an indicator of an increased tendency of genome alterations which may be caused by defective mismatch repair pathway. *hMLH1, is* a critical *m*ismatch repair pathways gene known to be mutated in Colorectal Carcinoma. Aim of the study was to better understand the relationship between *hMLH1* and Microsatellite Instability in Colorectal Carcinoma using publicly available datasets. An *in silco* study was performed using a dataset retrieved from cBioportal and the Cancer genome atlas. The results demonstrated that low *hMLH1* expression had significantly higher MSI as expected. However, it interestingly showed an inverse association between *hMLH1* and fraction of genome alterations. This possibly highlights the complexity of the link between *hMLH1* and Microsatellite Instability.

Keywords: Colorectal carcinoma, Microsatellite instability, MSI, hMLH1 gene, fraction of genome altered

1. Introduction

Colorectal Carcinoma (CRC) is cancer of the colon or rectum, arising from precancerous polyps, adenomatous polyps, or serrated polyps that form over several years. It is the third most common cancer and the fourth most common cause of cancer-related deaths globally (Herman et al., 1998). CRC, like most cancers is caused by mutations in critical genes like DNA repair mechanisms resulting in alterations of the genome. Depending on the origin of the mutation, CRC is classified as sporadic, inherited, or familial (Marmol et. al., 2018). Genomic instability, the increased tendency of genome alteration during cell division, is an important feature underlying CRC. Microsatellites are short non-coding repeating sequences throughout the genome. They occur at thousands of locations and have a higher mutation rate. When instable, the specific genes that monitor genomes for errors are unable to correct thus leading to instability. Microsatellite Instability (MSI) which is caused by a hypermutable phenotype, due to loss of DNA repair mechanisms, are one of the underlying mechanisms that causes this instability (Kawakami et al., 2015). The mutations can affect non-coding regions as well as codifying microsatellites, resulting in the reading frames of oncogenes or tumor suppressor genes being altered and tumors developing. Fraction of genomic alterations (FGA) includes measuring the percentage of copy number altered chromosome regions due to cancer-driven mutations, gene fusions, amplification, deletion, and post-transitional modifications.

Mismatch repair (MMR) pathways play a vital role in identifying and repairing mismatched bases during DNA replication and genetic recombination in normal and in cancerous cells (Sameer et al., 2014). Defects in MMR are also known to cause subsequent high MSI which leads to the accumulation of a mutation load. MMR genes mutated in tumors with MSI include *hMSH2, hMLH1, hMSH6,* h*PMS1* and *hPMS2.* MLH1/MSH2 phenotype constitutes a pathologically and clinically distinct subtype of sporadic CRC (Richman, 2015). These markers are important to define therapeutic strategy in CRC. Several clinical trials have demonstrated that MMR deficiency or high MSI is significantly associated with long-term immunotherapy-related responses and better prognosis in CRC (Richman, 2015). *hMLH1* has an impact on the fraction of genome alterations, because the lower the range the less altered the genome is and the higher the range the more altered the genome is, therefore the *hMLH1* gene expression leads to MSI which is a biomarker for CRC (Marmol et al., 2017). The goal of this study was to understand better the expression of *hMLH1* and corresponding MSI. It was predicted that CRC patients may have a lower expression of *hMLH1* and a corresponding higher MSI. By analyzing publicly available datasets on the *hMLH1* expression, MLH1/MSH2 phenotype, the MSI status and genomic alteration in CRC patients, the link between *hMLH1* and MSI was better understood.

2. Materials And Methods

The *in silco* study was performed using the publicly available TCGA dataset: Colorectal Adenocarcinoma (cBioportal for Cancer Genomics, n.d.), licensed by National Institute of Health (NIH).

2.1 Data processing

The dataset was sorted to specifically test the hypothesis by applying criteria that made sure no patients had information missing. Only patients with the following variables were included: sex type, cancer stage, cancer type, MSI type, MLH1 silencing (an epigenetic modification that prevents the expression of *hMLH1*), and vital status. The sorted dataset was matched to the original clinical data to make sure all patient IDs were the same. Out of the 276 patients originally, 202 patients remained. All files chosen were imported to Microsoft Excel and then sorted and characterized based on frequency distribution. The frequency distributions across the variables in the dataset are shown in Table 1.

2.2 Data analysis

The file was imported into RStudio – 1.4.1106 and DATAtab for further statistical analysis (DATAtab, n.d.). Histogram and Box & whisker plots were created through Excel and cBioportal (cBioportal for Cancer Genomics, n.d.).

Shapiro-Wilk normality distribution test was conducted to check that continuous variables in the datasets followed normal distribution. T-tests were considered initially to check for the correlation between datasets, but results from Shapiro-Wilk test showed that the datasets being considered did not have normal distribution, and so the Mann-Whitney U test, a non-parametric test that does not need uniform distribution, was used to test for correlation between the continuous variable datasets. The null hypothesis for the Mann-Whitney test states that there is no difference between the datasets if the p-value > 0.05 for 5% significance, and the alternate hypothesis states that there is a significant difference between the datasets if the p-value < 0.05. Chi-square test was used to test the independence of categorical datasets. Chi-square test has the same null and alternate hypothesis as the Mann-Whitney U test.

All instruments used allowed for a graphical view of the results and indicated the difference in the means. Larger difference in means meant more significant differences between variables.

3. Results

3.1 Characteristics of the dataset

Once data sorting was completed there were 202 patients that satisfied the selection criteria, the details are shown in Table 1. MSI high was defined when two of five markers showed instability in the genome. MSI low was determined when only one MSI marker showed instability and the rest showed stability. Note that *hMLH1* refers to the gene, while MLH1 refers to the protein created by *hMLH1* gene. Units for MLH1 expression are RPKM or Reads per kilo million mapped reads in a library. (Biostars, n.d.)

$$RPKM\ of\ a\ gene = \frac{Number\ of\ reads\ of\ a\ gene\ \times 10^{3}\ \times 10^{6}}{Total\ number\ of\ mapped\ reads\ from\ given\ library\ \times\ gene\ length}$$

Reads in the above equation for MLH1 expression scoring refer to immunohistochemical data scored regarding staining intensity (negative, weak, moderate or strong) and fraction of stained cells (<25%, 25-75% or >75%). (The Human Protein Atlas, n.d.). Any patient with Tumor stage 1 or 2 in the dataset were put into the Cancer Stage category Low while any patient with Tumor stage 3 or 4 cancer was put into the Cancer Stage category High. The data with the variables sex type, cancer stage, cancer type, MSI type, MLH1 silencing, and vital status was grouped for certain tests in the report based on the median of the MLH1 expression.

It was observed in the data, 100% of patients had cancer with ~44% having high stage cancer, and ~58% having colon cancer, yet surprisingly only 9.4% of the patients had died. A reason for higher survival could be the stable microsatellite stability status observed in 68.8% of the patients.

Table 1. Cohort characteristics		
Total Sample Size = 276. Selected Samples = 202.		
Variables (n)	No. of Patients	Frequency (%)
Sex (202)		
Female	97	48.0%
Male	105	52.0%
Cancer Stage (202)		
High	89	44.1%
Low	113	55.9%
Cancer Type (202)		
Colon	118	58.4%
Colorectal	34	16.8%
Rectal	50	24.8%
MSI Type (202)		
MSS	139	68.8%
Low MSI (MSI-L)	35	17.3%
High MSI (MSI-H)	28	13.9%
MLH1 Silencing (202)		
0 (Not Silenced)	177	87.6%
1 (Silenced)	25	12.4%
Vital Status (202)		
Alive	183	90.6%
Dead	19	9.4%

3.2 MSI type and *hMLH1* expression

MSI was seen in ~30% of the patients among which ~44% had high MSI. Figure 1, a box and whisker plot graph, shows the MLH1 expression on the y axis and the three MSI types on the x-axis. The graph shows a significant difference in MLH1 expression between the Low MSI (MSI-L) and High MSI (MSI-H) samples. MSI-L samples had a mean of 20.9 RPKM and a standard deviation of 6.7, while MSI-H samples had a much smaller mean of 8.0 RPKM but a larger standard deviation of 8.0 because of a few outliers.

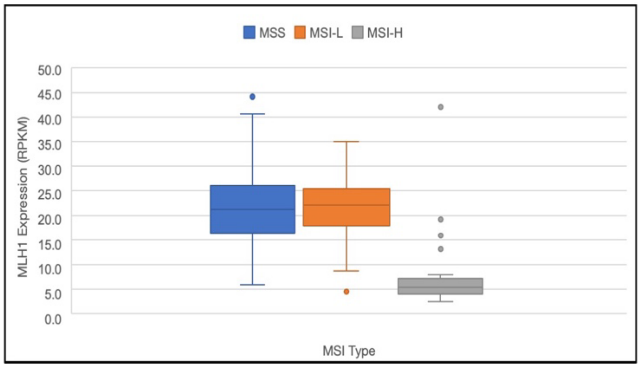

Figure 1: Compares the MLH1 expression with the three MSI types – MSS, MSI-L, and MSI-H.

To test the hypothesis statistically, further analysis was done to first verify that the MSI-H and MSI-L samples were statistically independent, and subsequently the relationship between MLH1 expression and MSI type was checked.

Running the Shapiro-Wilk normality test indicated that MSI-L dataset was normally distributed, while MSI-H dataset was not. Mann-Whitney U test, a non-parametric test, revealed that the difference between MSI-L and MSI- H with respect to MLH1 expression was highly statistically significant, p-value=<.001, r=0.7. If p-value < .05 for 5% significance, MSI-L and MSI-H are considered to be significantly different.

The estimated median of the 206 patients with MLH1 expression was 20.03 RPKM. Based off this, the patients were characterized in to two categories of high and low MLH1 expression for further analysis. In Figure 2, Group A has low MLH1 expression values, while Group B has high MLH1 expression values. This graph shows that patients with low MLH1 expression (Group A) had significantly high MSI-H, while patients with high MLH1 expression had practically no MSI-H.

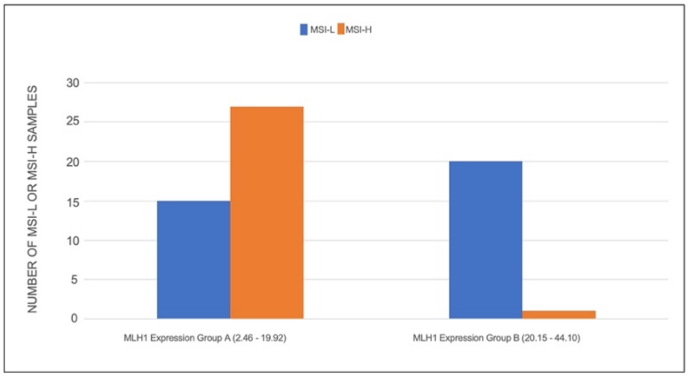

Figure 2: Comparison of number of samples with MSI-L and MSI-H between two groups categorized by MLH1 Expression.

A Chi-square test was performed between MLH1 Expression Groups and MSI Status, and a statistically significant relationship was found, p = <.001. The Chi2 test is therefore significant and the null hypothesis that the two groups are independent is rejected.

This validates the hypothesis that high MSI samples are significantly different from low MSI samples, and that there is a significant relationship between MLH1 expression and MSI.

3.3 High correlation between MLH1 expression and MLH1 Silencing

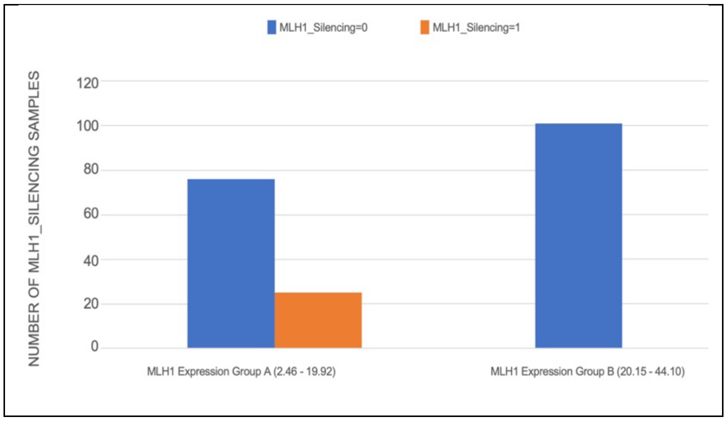

Figure 3: Compares number of samples in low and high MLH expression groups for the two cases of MLH1_Silencing.

MLH1_Silencing=0 represents patients who don't have their *hMLH1* gene silenced, while MLH1_Silencing=1 represents the patients who have their *hMLH1* gene silenced. The Relationship between MLH1 expression and MLH1_Silencing is explored in Figure 3, which shows that the presence of MLH1_Silencing corresponds to a reduction in MLH1 expression and therefore there are no samples of Group B with MLH1_Silencing=1. As expected, the samples are present in both groups when MLH1_Silencing=0.

Results of Chi-square test to test for independence between MLH1 Expression Groups and MLH1_Silencing showed that there was a statistically significant relationship between them, p = <.001. The calculated p-value of <.001 is lower than the defined significance level of 5%. The Chi-square test is therefore significant and the null hypothesis that they were independent was rejected.

This result is significant because it validates the relationship between MLH1 Silencing and *hMLH1* expression.

3.4 Patients with low *hMLH1* expression have a less altered genome

The impact of *hMLH1* on genome alterations was further analyzed. The Mann-Whitney U test showed that the difference between MLH1 Expression Group A (2.46 - 19.92) and MLH1 Expression Group B (20.15 - 44.10) with respect to FGA was statistically significant, p=.003. As demonstrated in Figure 4, it was found that lower expression of *hMLH1* gene demonstrates lower Fraction Genome Altered (FGA). Group A has a mean of 0.22 and a standard deviation of 0.18, while Group B has a higher mean of 0.31 and a higher standard deviation of 0.2. This potentially indicates an adverse impact of higher MLH1 expression and was contradictory to the hypothesis.

4. Discussion

CRC, owing to its high incidence and mortality rate, is of big concern especially in the developed world. Based on phenotype researchers have classified it biologically into two types: 1. MSI and 2. MSS but chromosomally unstable. MSI tumors can help doctors identify the potential cause of the tumor. It also helps with determining the therapy module. The latest developments indicate that immunotherapy, a very new treatment regimen had higher efficacy when there is High MSI.

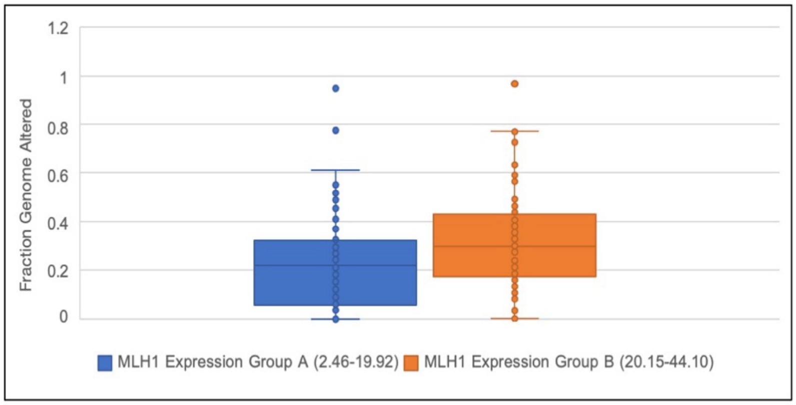

Figure 4: Comparison of Fraction of Genome Altered with MLH1 Expression Groups.

It is also known that MLH1 silencing is one of the causes for MSI. However, only about 30% of CRC tumors are seen with MSI, and among them only 50% have high MSI (Marmol et al., 2017). There is always a need to better understand the underlying molecular cause of MSI, as the complexity of CRC is quite significant. Owing to the interest in studying CRC, the aim was to explore the publicly available dataset to study the link between MSI and *hMLH1* expression and its impact on genomic alterations. The goal of this study was to analyze the impact of low *hMLH1* expression on MSI and its role in genomic alterations in CRC. The cohort was validated, and it was found that the sorted dataset demonstrated a similar frequency of MSI that aligns with the literature. On performing further analysis, it was identified that *hMLH1* expression and MLH1 silencing were significantly associated. Low *hMLH1* expression demonstrated high MSI, whereas the patients with high expression of *hMLH1* showed low MSI. These findings validated prior studies about the role of MLH1 and MSI. An important and interesting observation was the inverse association between *hMLH1* expression and FGA. This contradicts the expected findings and throws open a lot of questions which require research to better understand the implication. It however emphasis the complexity of molecular biological process and highlights the need to explore further the basic biological questions using public datasets. However, that was not within the scope of this research project. But the future does look exciting to unravel these further with advancement of better technology and improved data analysis methods.

5. Conclusion

The purpose of this study was to analyze the impact of low *hMLH1*expression on MSI in CRC. From the analysis it was concluded that patients with low *hMLH1* expression have significantly higher MSI. This led us to explore the idea that MSI-H can show a trend toward a better prognosis for CRC and reviving *hMLH1* expression could restore MMR activity which could eventually result in lower MSI. This means that less people will have their life threateningly affected by colorectal carcinoma and immunotherapy can allow patients to recover from cancer in a smaller timeframe. Looking ahead, it may be valuable to research if methylation of *hMLH1* in pre-malignant adenomatous polyps is an early event in carcinogenesis of CRC. Studying methylation is advantageous, as it could yield potential targets for treatment to combat tumor progression.

Acknowledgment

I would like to thank Skoolmentor, late Nutan Malhotra, and especially my mentor Dr. Manuel for teaching me, guiding me throughout this process and helping me with proofreading of this paper.

References

Biostars. (n.d.). *Blog:Gene expression units explained: RPM, RPKM, FPKM and TPM.* https://www.biostars.org/p/273537/#:~:text=Gene%20expression%20units%20explained%3A%20RPM%2C%20RPKM%2C%20FPKM%20and%20TPM&text=In%20RNA%2Dseq%20gene%20expression,units%20from%20mapped%20sequence%20data.

Bloom, A. (2023) *What Is Microsatellite Instability?* MD Anderson Cancer Center. https://www.mdanderson.org/cancerwise/what-is-microsatellite-instability-MSI.h00-159617067.html

cBioportal for Cancer Genomics. (n.d.). *Colorectal Adenocarcinoma (TCGA, Nature 2012).* https://www.cbioportal.org/study/summary?id=coadread_tcga_pub

DATAtab. (n.d.). *Statistics Calculator.* https://datatab.net/statistics-calculator/charts

Fleming, M., et al. (2012). Colorectal carcinoma: Pathologic aspects. *Journal of Gastrointestinal Oncology*, 3(3):153-73. doi: 10.3978/j.issn.2078-6891.2012.030. PMID: 22943008; PMCID: PMC3418538.

Gelsomino, F., et al. (2016). The evolving role of microsatellite instability in colorectal cancer: A review. *Cancer Treatment Reviews*, 51:19-26. doi: 10.1016/j.ctrv.2016.10.005. Epub 2016 Oct 27. PMID: 27838401.

Herman, J. G., et al. (1998). Incidence and functional consequences of *hMLH1* promoter hypermethylation in colorectal carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95 (12) 6870-6875 https://doi.org/10.1073/pnas.95.12.6870

Kawakami, H., Zaanan, A., and Sinicrope, FA. (2015). Microsatellite instability testing and its role in the management of colorectal cancer. *Current Treatment Options in Oncology*, 16(7):30. doi: 10.1007/s11864-015-0348-2. PMID: 26031544; PMCID: PMC4594190.

Mármol, I., et al. (2017). Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *International Journal of Molecular Sciences*, 18(1):197. doi:10.3390/ijms18010197. PMID: 28106826; PMCID: PMC5297828.

Richman, S., (2015). Deficient mismatch repair: Read all about it (Review). *International Journal of Oncology*, 47(4):1189-202. doi:10.3892/ijo.2015.3119. PMID: 26315971; PMCID: PMC4583524.

Sameer, A. S., et al. (2014). Mismatch Repair Pathway: Molecules, Functions, and Role in Colorectal Carcinogenesis. *European Journal of Cancer Prevention*, vol. 23, no. 4, pp. 246–57. JSTOR, https://www.jstor.org/stable/48504328.

The Cancer Genome Atlas (TCGA) (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330–337. https://doi.org/10.1038/nature11252

The Human Protein Atlas. (n.d.). *Help & FAQ. How is protein expression scored?* https://www.proteinatlas.org/about/help#4

Yurgelun, M. B., et al. (2017). Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer. *Journal of Clinical Oncology*, 35(10):1086-1095. doi:10.1200/JCO.2016.71.0012. PMID: 28135145; PMCID: PMC5455355.