

A Bioinformatic Approach to MSI Cancer Research

Nick Napier*, Nico Limogiannis

Department of Computer Science, Wofford College, USA

Copyright © 2016 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Recent advances in genome sequencing, mRNA expression, and other genomic technologies have resulted in several databases containing a wealth of genomic information—so much information that much of it remains unexamined. By using the R programming language, databases can be mined to research cancer without the need for experimentation. Using this approach, the mRNA expression of 9,768 known genes was compared between microsatellite stable (MSS) and microsatellite instable (MSI) colorectal cancers. This computational analysis revealed 5,435 genes with a significant difference ($p < 0.05$) in expression between MSS and MSI forms; many genes showed strongly significant differences ($p < 2 \times 10^{-15}$). In addition, a gene expression signature constructed from these differentially-expressed genes predicted MSI-H (but not MSI-L) status in cancer cell lines with 97% accuracy. Finally, our results potentially associate several pathways with MSI cancers. While much future study will be needed to more closely examine these results, the current study demonstrates the use of bioinformatics in making discoveries based upon existing data and in directing the focus of future experiments.

Keywords Bioinformatics, Colorectal Cancer, MSI, Genomics, Gene Expression

1. Introduction

Colorectal cancer is one of the most prevalent malignant cancers in the world and is thus of great importance in cancer research. Microsatellite instability (MSI) in colorectal cancers has recently been the focus of a great body of research due to its significance in affecting patient outcomes [1]. MSI colorectal cancers possess defective DNA mismatch repair, which leads to frequent slippage mutations, especially in microsatellite regions of the genome. MSI cancers are generally associated with a much better prognosis than microsatellite stable (MSS) cancers [2]. Although it is known that the expression of thousands of genes, cellular phenotype, and patient prognosis all differ greatly between MSI and MSS cancers, relatively little is known about the specific genes involved or how

dysregulation of these genes can lead to such divergent phenotypes and prognoses. Furthermore, colorectal adenocarcinomas exist on a spectrum from MSS to MSI; those that lie between are referred to as MSI-L (and for clarity, fully or mostly MSI cancers are called MSI-H).

One substantial issue that has recently impeded research in the area of MSI/MSS cancer research—as well as genetic research in general—is an overabundance of raw data and a shortage of the resources necessary to process this data. Recent advances in sequencing, microarray, and epigenetic technologies have led to a wealth of genomic data, especially in cancer research. For example, genome sequencing can be done at speeds and prices that were previously unimaginable, and every genome sequence produces approximately three billion base pairs—about 100GB of data [3]. These technological advances have resulted in the creation of databases with abundant data concerning DNA sequence, mutations, mRNA expression, protein expression, epigenetic modifications, and other genomic data. Without organized and greatly accelerated methods for analyzing this “big data”, much of what it reveals about cancer and other diseases will remain overlooked.

The development of computational approaches to data analysis has provided an answer to the problem of big data. As the speed of the production of genomic data has increased, so has the processing speed via more efficient algorithms, faster computer processors, and more compact memory. A bioinformatic program can process and store chunks of data much larger than a human could possibly handle, and it can do so in seconds. Such computing power has allowed such feats as the processing of over 560,000 data points to identify two novel miRNAs, and can be applied to a diverse set of data types for efficient analysis of big data [4].

There is a great need for further understanding of MSI colorectal cancers and the plethora of raw data concerning these cancers. With this overabundance of data, an organized computational approach to mining already-constructed databases could yield many novel insights into colorectal cancer without the need for experimentation and at almost no cost. For this reason, we utilized computational approaches to analyze massive quantities of mRNA expression, DNA sequence, global gene methylation, and other genomic data that is publicly available on The Cancer Genome Atlas

(TCGA) data portal. In the process of analyzing this data, we identified 5,435 genes that were differentially expressed between MSI-H and MSS cancers (MSI-L cancers were excluded to yield clearer data), produced a novel gene expression signature from these differentially-expressed genes, and identified prospective activations of several pathways that may hold potential for further research into the nature of MSI cancers.

2. Methods

2.1. Obtaining Genomic Data

All data were obtained from Munzy et al [5] via TCGA. Affymetrix microarray mRNA expression Z-scores, global methylation β -values, copy number variations (CNVs), mutation data, and MSI status were obtained from TCGA using the CGDS-R package for the programming language R v3.0.1 [6]. All statistical analyses were performed using R v3.0.1, Python v3.4, and Microsoft Excel 2011.

2.2. Analysis of mRNA Expression Differences between MSI Types

The samples obtained from Munzy et al [5] were separated into MSI-H (N=38) and MSS (N=193) groups based upon the MSI statuses reported by Munzy et al [5]. These groups were then compared on a gene-by-gene basis for all available genes (9,768 genes) for differences in mRNA expression using two-sample t-tests or, when appropriate, Welch's t-test. Importantly, MSI-L samples and samples for which data was incomplete were excluded from analyses.

2.3. Construction and Evaluation of MSI-H Gene Signature

Genes were selected from the 5,435 differentially expressed genes for use in the gene expression signature using the formula:

$$\frac{X}{S}$$

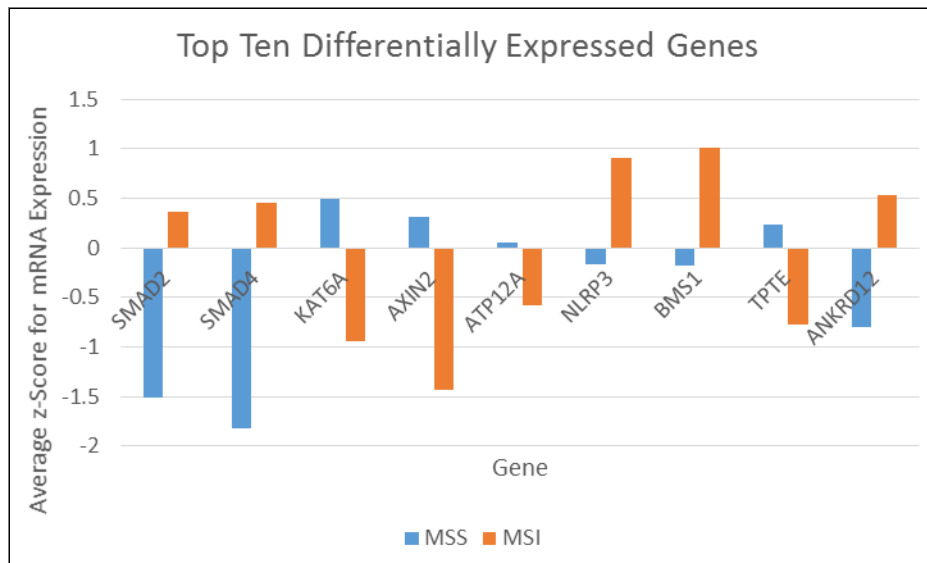
where x is the sample mean expression for a gene in MSI cancers and s is the standard deviation for gene. The genes which had the greatest values for this analysis were used in the expression signature.

The MSI status predicted by this novel signature and the actual MSI status were then compared for colorectal cancer cell line data obtained from the Cancer Cell Line Encyclopedia (CCLE) database via the TCGA data portal. The success rate of this signature was defined as the percentage of CCLE MSI statuses that were correctly predicted by our signature.

3. Results

3.1. Analysis of mRNA Expression Differences between MSI Types

Out of the 9,768 genes examined, 5,435 genes were found to have a significant difference in mRNA expression between MSI-H and MSS samples ($p \leq 0.05$). Many of these genes possessed extremely small p-values (Figure 1).



Gene	SMAD2	SMAD4	KAT6A	AXIN2	ATP12A	NLRP3	BMS1	TPTE	ANKRD12
p-Value	1.70E-18	1.64E-16	7.28E-11	1.75E-09	1.43E-08	1.89E-08	2.35E-08	9.78E-08	2.77E-07

Figure 1. Relative average expression levels of the ten genes that were most strongly associated with differential mRNA expression between MSI-H and MSS cancers, as indicated by p-values, shown in a table below the chart. N=38 MSI-H and 193 MSS samples for each gene.

3.2. Construction and Evaluation of MSI-H Gene Signature

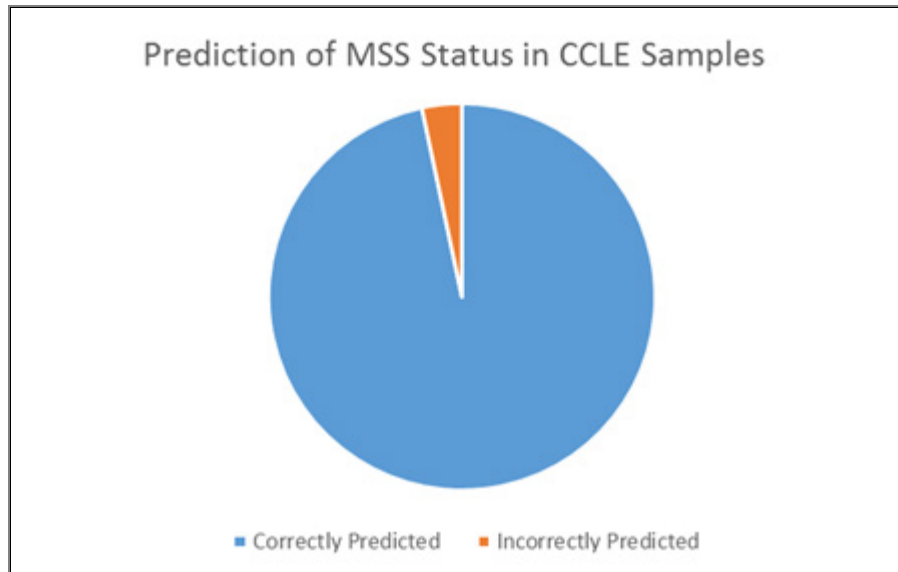


Figure 2. The novel gene signature consisted of ten genes. This signature correctly predicted the MSI status of 97% (58 out of 60) of the CCLE samples.

Gene	GSK3B	CTNNB1	APC	AXIN2	WNT	FZD10
p-value	3.70E-08	8.90E-07	6.70E-06	1.80E-09	0.054	0.047

Figure 3. p-Values for differential mRNA expression between MSS and MSI-H for several key regulators of the Wnt signaling pathway. Note that Wnt itself does not show a significant difference at our alpha level.

4. Conclusions

The benefits of using a computational approach in this research are immediately apparent in the ability to develop novel results without performing any experimentation. The speed with which data could be accessed was also a clear benefit. For example, the final R program was designed to query TCGA for data on chunks of 200 genes at a time; this process performed all 9,768 gene assays in approximately ~4 hours. This speed far outpaces manual access of TCGA to get assay data: such an endeavor would at best take ~5-10 minutes for a single gene assay. Furthermore, the use of programming allowed for the rapid and efficient application of a wide range of statistical analyses on gene data, as well as the organization and graphical representation of the results of these statistical analyses. This combination of efficient statistical analysis, organization, and graphical representation of gene data allowed for the rapid detection of patterns in a massive sea of data.

Our analysis identified 5,435 genes that were expressed differently between MSI-H and MSS cancers. Amongst these genes were many that have previously been identified as important to cancer and MSI status, such as *KRAS* and *TGFBR2*. The 97% accuracy exhibited by the expression signature constructed from a subset of these genes was fantastic given the ease with which this accuracy was attained. Unfortunately, the 3% of samples that were inaccurately identified exhibit the limitations of the method

used. Given the extremely low p-values of some of the signatures used (e.g. $p < 2.0 \times 10^{-27}$), we expected the accuracy of the test to have been much higher. In addition, our use of cell lines rather than patient samples and the exclusion of MSI-L samples from our study would have inflated the accuracy somewhat, making our actual accuracy lower than 97%. A similar study that included MSI-L samples and did not use cell lines obtained accuracies of ~90% using a 64-gene signature [7]. It is apparent from these inaccuracies that although mRNA expression is a significant factor in determining MSI status, other factors such as protein-level expression regulation must also affect MSI status. Future attempts at creating diagnostic profiles should identify and include these factors.

Comparison of our 5,435 differentially-expressed genes with known cancer pathways revealed several pathways which were highly enriched with differentially-expressed genes. Notably, the *Wnt* pathway contained numerous highly significantly different genes and is known to lead to several phenotypic characteristics that are seen in MSI cancers, such as DNA repair malfunction and epithelial cell characteristics (data not shown)[8]. This identification of potential targets for future research demonstrates a final benefit of computational approaches to research: out of the thousands of pathways and genes that could be involved in MSI status, several weeks of computational work eliminated thousands of possibilities and implicated a small set of possibilities at virtually no cost. The patterns and associations that are found

in computational studies may or may not be important, but they allow future research to be focused upon a small set of possibilities— thereby reducing cost, accelerating discoveries, and offering hope for future cancer research.

Acknowledgements

The authors would like to thank our faculty mentor Dr. Mingli Yang and Drs. Timothy Yeatman and Jack Pledger of the Gibbs Cancer Center & Research Institute for their extensive mentoring and guidance. The authors are also indebted to Dr. Angela Shiflet of Wofford College for making this research opportunity possible.

REFERENCES

- [1] Timmerman B, Kerick M, Roehr C, Fischer A, Isau M, *et al.* (2010 Dec 22). Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PloS One*, 5(12) doi: 10.1371/journal.pone.0015661
- [2] Xavier S, Van Cutsem E, Tejpar S, Prenen H, De Hertogh G. (2014). MSI versus MSS sporadic colorectal cancers: Morphology, inflammation, and angiogenesis revisited. *J Clin Oncol*, 32.
- [3] Puckelwartz M, Pesce L, Nelakuditi V, Dellefave-Castillo L, Golbus J, *et al.* (2014). Supercomputing for the parallelization of whole genome analysis. *Bioinformatics*, 30(11), 1508-1513.
- [4] Kuo TY, Hsi E, Yang IP, Tsai PC, Wang JY, *et al.* (2012). Computational Analysis of mRNA Expression Profiles Identifies MicroRNA-29a/c as Predictor of Colorectal Cancer Early Recurrence. *PloS One*, 7(2), doi:10.1371/journal.pone.0031587
- [5] Munzy D, Bainbridge M, Chang K, Dinh H, Drummond J, *et al.* (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330-7.
- [6] Cerami E, Jianjiong G, Dogrusoz U, Gross B, Sumer S, *et al.* (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5), 401-4.
- [7] Tian S, Roepman P, Popovici V, Michaut M, Majewski I, *et al.* (2012). A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol*, 228, 586-595.
- [8] Kanehisa Laboratories. (2015 Mar 29). KEGG Pathway Database. *Kyoto Encyclopedia of Genes and Genomics*. Retrieved from <http://www.kegg.jp/>
- [9] R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Retrieved from <http://www.R-project.org/>.