FUNCTIONAL GENOME ORGANIZATION IN COLORECTAL CACNER: AN INTEGRATIVE DATA SCIENCE APPROACH

ETHICAL CONSIDERATIONS FOR PUBLICLY AVAILABLE GENOME DATA

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Biomedical Engineering

By
Zachary Thomas

November 2, 2020

Technical Team Members:
Alexandra Hickman

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed Zachary Thomas Date: 10/31/2020

Approved: Date:
Chongzhi Zang, Center for Public Health Genomics

Approved: Date:
Catherine D. Baritaud, Department of Engineering and Society

Colorectal cancer is the third most diagnosed cancer in the United States (Thanikachalam et al., 2019). It has been projected that by 2030 the incidence rates for colon and rectal cancer will increase by 90 percent and 124.2 percent for patients 24 to 32 years of age (Bailey et al., 2015). Cancer in general is a complex disease involving multiple layers of malfunctioning systems in the genome, epigenome, and higher-order genome organization. A number of computational frameworks have been developed to study functional regulators at these different levels. For example, Binding Analysis for Regulation of Transcription (BART), was developed to predict functional transcriptional regulators that regulate gene expression (Wang et al., 2018). Like many other data science approaches to studying genomics, BART relies on data collected from the public domain.

The technical project and loosely coupled STS project seek to provide a better understanding of cancer and the data used to develop research tools. More specifically, with the help of Biomedical Engineer Alexandra Hickman, Research Associate Zhenjia Wang, and Chongzhi Zang of the Departments of Public Health Sciences, Biomedical Engineering, and Biochemistry and Molecular Genetics, the technical project aims to develop a computational pipeline for 3D genome analysis. The pipeline will then be applied to colorectal cancer using data collected from the public domain. The loosley coupled STS project focuses on the ethics of consent and data sharing within the genomics community and general public as concrete principles to guide regulatory processes have yet to be well defined. This problem will be analyzed using a social construction of technology approach.

PIPELINE FOR 3-DIMENSIONAL GENOME ANALYSIS

Colorectal cancer is the third leading cause of cancer death in the world and (Rawla et al., 2018). It is estimated that there will be 104,610 new cases of colon cancer and 43,340 cases of

rectal cancer diagnosed in 2020 (Seigal et al., 2020). However, the epigenetics of colorectal cancer are still not well understood.

A number of transcriptional regulators have been shown to promote tumorigenesis, the formation of tumors, in colorectal cancer. CTCC-binding factor (CTCF) is involved in various cancers, such as breast cancer, lung cancer, and prostate cancer (Lai et al., 2020). This transcriptional factor has also been shown to promote the proliferation and chemotherapy resistance of colorectal cancer. Another prominent tumorigenic transcriptional regulator for colorectal cancer is TEAD4. The expression level of TEAD4 is commonly seen in clinical samples of colorectal adenomas. Knockdowns of TEAD4 led to inhibition in colorectal cancer cell proliferation, suggesting that TEAD4 may be a novel biomarker in colorectal tumorigenesis (Tang et al., 2018). Unlike CTCF and TEAD4, there are various transcription factors involved in colorectal cancer that have not yet been identified. The complexity and number of these transcription factors make identification difficult.

Current pipelines for 3D genome analysis that exist are not sufficient to study functional transcriptional regulators in colorectal cancer. These include Chicdiff and Selfish (Ardakany, Ay, Lonardi, 2019; Cairns, Orchard, Malysheva, Spivakov, 2019). Chicdiff is a powerful computational tool for analyzing 3D genomics data, specifically Hi-C data, and the detection of differential interactions. Much like Chicdiff, Selfish is a novel tool for measuring Hi-C replicates' reproducibility and differential chromatin interactions. Chicdiff and Selfish both accomplish differential analysis in a locus-to-locus manner. While these methods are effective at analyzing short, specific regions of the genome, they are not very powerful from a data science perspective. In order to be a practical data science approach, the differential analysis would need to be genome-wide while maintaining high resolution.

The goal of this project is to develop a computational pipeline that combines the functions of (1) drawing Hi-C contact maps and (2) predicting functional transcriptional regulators from Hi-C data. The workflow of the pipeline can be seen in Figure 1. This approach to Hi-C data analysis could be a considerable improvement over Chicdiff and Selfish as it streamlines the analysis of 3D genomic data. In addition to calculating whole-genome differential interactions and predicting functional transcriptional regulators, the proposed pipeline aims to predict CTCF binding sites.

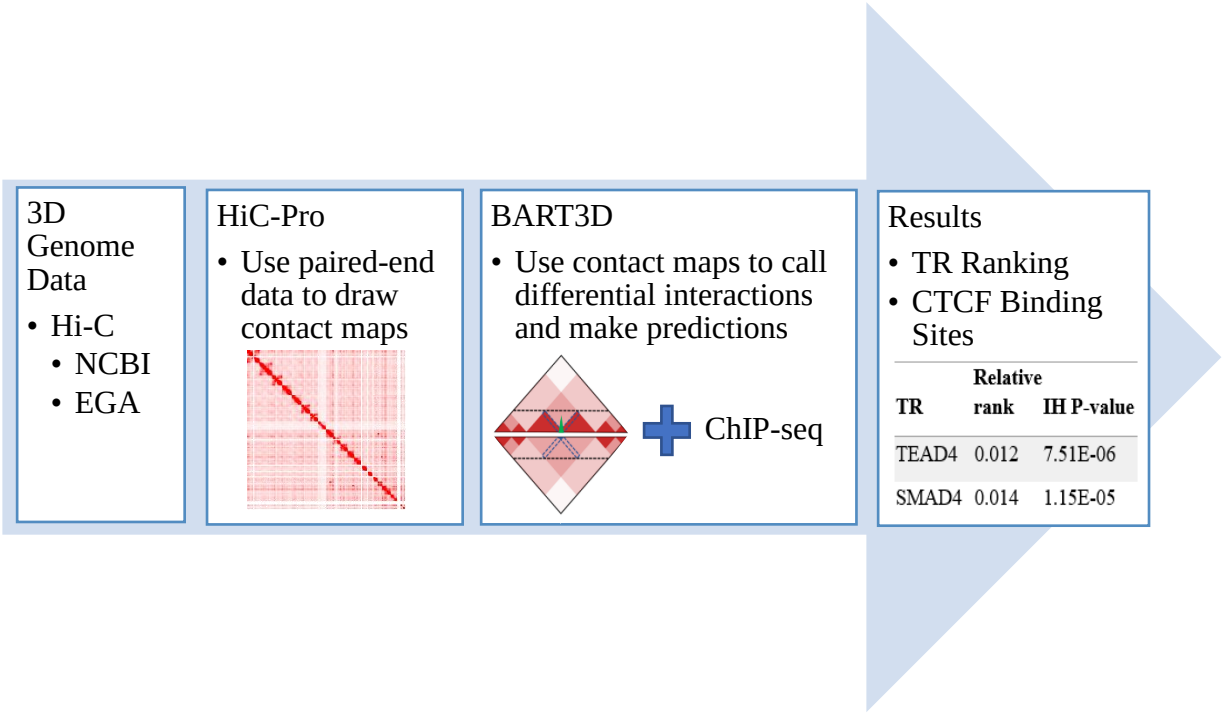

Figure 1: Work flow describing pipeline function. The figure displays the steps the proposed pipeline will combine. First, it takes in 3D genome data. Then it will use Hi-C Pro (Servant et al., 2015) to draw contact matrices. Finally, those matrices will be used in BART3D to call differential interactions and make functional transcriptional regulator predictions. (Thomas, 2020)

In order to build and test the pipeline, a number of computational tools must be used. As illustrated in Figure 1, HiC-Pro will be used to draw 3D genome interaction maps (Servant et al., 2015). Second, BART3D, recently developed by the Zang Lab at the University of Virginia, will

be used to make functional transcriptional regulator predictions. All of these tools will be run on Rivanna, the University of Virginia's high-performance computing environment. Finally, in order to apply the pipeline to colorectal cancer, normal colon and colon tumor Hi-C samples will be collected from the public domain and used as input to the pipeline. For example, Suhas S. P. Rao from the Center of Genome Architecture at the Baylor College of Medicine studied cohesion and its impact on chromosomal loop domains (Rao et al., 2017). Rao and his colleagues performed their experiments on HCT-116 cells, a colorectal cancer cell line. These Hi-C data sets, along with other publicly available Hi-C data sets, will be used in the construction and verification of the proposed pipeline.

The proposed pipeline will be tested on colorectal cancer data collected from the public domain. Using Hi-C data as input, the pipeline will output putative transcriptional regulators of colorectal cancer and putative CTCF binding sites. By applying the proposed pipeline to colorectal cancer, we hope to demonstrate its serviceability to the cancer research community. This serviceability is the result of streamlining Hi-C data analysis and identifying possible tumorigenic promoters with one tool. These putative tumorigenic promoters could then be used as targets for cancer therapies. As previously stated, there are thousands of transcriptional regulators. Ranking putative transcriptional regulators narrows that list and provides a concrete basis for experimental initiatives to delve deeper into the roles of significant transcription factors. The pipeline will be detailed in a paper to be submitted to *Bioinformatics*.

ETHICS OF PUBLIC GENOME DATA

Data science approaches to the analysis of any topic rely on data availability, and typically, a lot of it. This so called "Big Data" can be used to make decisions based on statistics and trends, it can also provide insight into the cause of these trends. Genomic data is no different,

providing statistical likelihoods of low frequency events, like cancers, drug treatment side effects, and drug-drug interactions on a per patient level (Francis, 2014, p. 111). The availability of genomic data also proves crucial for the development, maintenance, and progression of bioinformatics tools like BART (Wang et al., 2018). For this reason, a number of international policies demand immediate release of all sequencing data without explicit patient consent (Johnson, Slade, Giubilini, Graham, 2020, p. 150). However, this public release of genome data introduces a number of concerns including patient privacy and questions of data ownership; whether or not genomic data sets belong to the patient or the researcher. For example, in 1989, a member of the Havasupai Tribe of Arizona had donated DNA samples for genetic research on type II diabetes. Yet, in 2003, she found that her DNA samples were being used for other, nondiabetic research studies. After a six-year trial, a settlement was reached that included monetary compensation but no legal ramifications for the researchers that misused the DNA samples (Garrison, 2013, p. 202).

The complication exists in not only the lack of agreement among researchers' views on the usefulness and privacy of such personal data sets but also in the public's perception of genomics data. To many, this is an issue of balancing the positive and negative aspects of publicly releasing genomics data. The table in Figure 2 summarizes popular beliefs of why or why not genomics data should be shared. It is very clear to the research community that genomic data carries a substantial amount of weight in regard to potential research applications as well as personal privacy (Johnson, Slade, Giubilini, Graham, 2020). This weight has been leveraged by both proponents and opponents to genomic data sharing. The advocates claim that this weight comes with an ethical obligation to share the data due to its potential research benefit while the detractors claim that the weight necessitates further privacy protections (Johnson, Slade,

Giubilini, Graham, 2020; Fisher & Layman, 2018). While the research community is well-informed of the weight of genomic data sets, the public is not so well educated. There have been substantial reports of parents freely sharing their children's genomic information obtained from at-home DNA testing kits like the ones provided by 23andMe (Bala, 2020). This lack of awareness of the ramifications for sharing genetic information highlights the issue with the public's perception of genetic data. In terms of at-home DNA testing, there are a number of issues that skew the public's perception of the genome. First, the testing kits are returned without mediation by a doctor or genetic counselor. This limits the ability of the user to extrapolate key information that might reveal privacy concerns from the results. Second, the kits allow the user to freely post their results in the form of whole-genome sequences. This allows for potential identification even if they posted anonymously. Finally, the data contains information about the user who took the test, but also their closely related family members. This allows for interpolation of genetic information between family members should the data be re-identified.

Reasons Genomics Data…	
Should be publicly available	**Should NOT be publicly available**
Owners of genomics data have an ethical obligation to share data for potential societal benefits	Protect privacy of individuals as many genomics data sets are potentially identifiable
Address inequalities in access to the benefits of genomic services	Lack of consensus on technical approaches to facilitate data sharing
Some genetic data is not personal or sensitive (not identifying or unique)	Access to and control of genetic information makes it possible for others to have power over an individual's life or future.
Genomic data is not intrinsically identifiable	Data storage location could have accidental data release or be hacked

Figure 2: Reasons for and against publicly available genomics data. This table summarizes popular ideas regarding genomics data and their public accessibility (Thomas, 2020)

The STS topic seeks to build an understanding of the ethical concerns, within the research community and broader public, of publicly available genomics data. It will then aim to propose recommendations for ethical data collection and distribution within the research community and commercial DNA testing companies. Health data of individuals contain significant amounts of personally identifiable information and regulations regarding their collection, storage, and distribution must be safe and well-defined.

When examining such intricate problems like genomic data sharing, it is helpful to use an analytical framework. For this particular problem, a social construction of technology (SCOT) model as developed by Trevor Pinch and Wiebe Bijker will be used to analyze the parties at play (Bijker & Pinch, 1984). As a model, SCOT is made up of four major parts: relevant social groups, interpretive flexibility, closure, and stabilization. Figure 3 illustrates social groups relevant to genomic data sharing and their intended interactions.

With genomic data sharing as the technology of interest, there are several relevant groups that perceive value from it. In this model, each group applies different values to the technology and have either competitive or collaborative problems with the technology. For example, the researchers using the data to build data science driven tools might need to know the age and sex of the patient whose data they are using in order to maintain a robust pipeline. However, the patient might not want that level of detail disclosed. The application of this framework to the technology of genomic data sharing will aid the identification of areas of agreement, disagreement, and potential growth. The STS project will be synthesized in a Public Understanding of Science article discussing the current state of genomic data sharing and offering recommendations for future policies and regulations regarding genomics data not only for researchers, but at-home DNA testing companies as well.

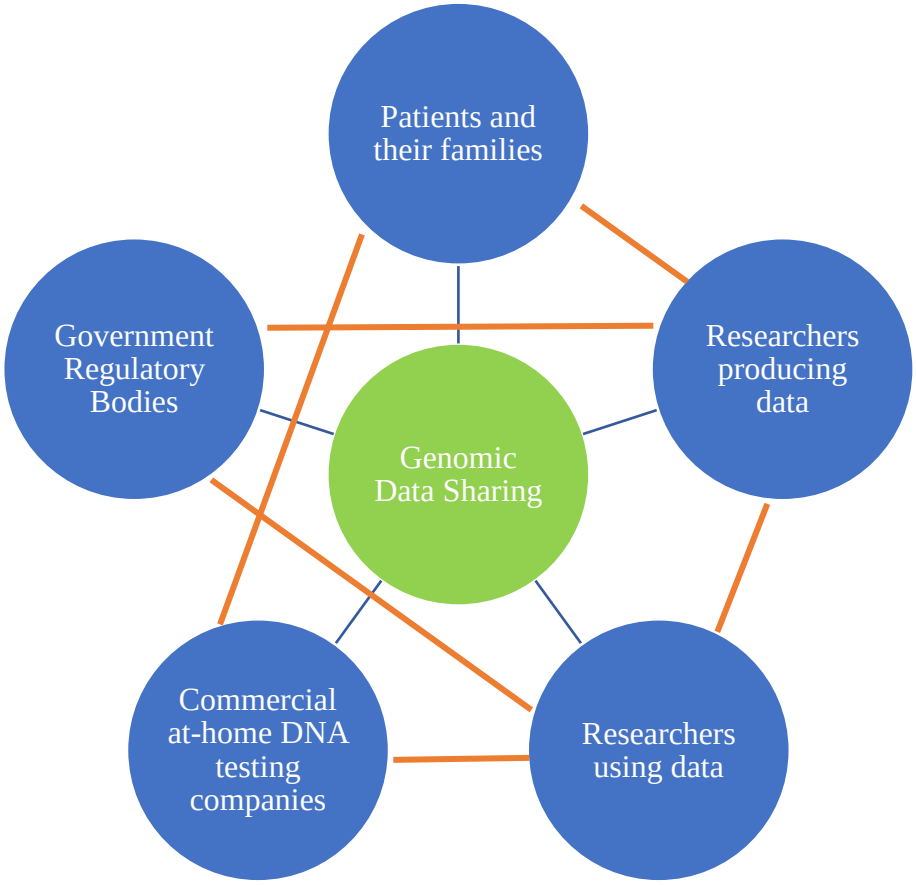

Figure 3: Social Construction of Genomic Data Sharing. This figure illustrates the relevant social groups (blue circles) as well as their intended interactions with each other (orange lines) (Adapted by Zachary Thomas (2020) from Bijker, Bönig, & van Oost, 1984).

The ability to develop bioinformatics tools relies heavily on publicly available data. While this data has allowed computational methods to delve deeper into the human genome to better understand the epigenetics of cancers, genomic data is still sensitive information and must be handled with extreme care.

Works Cited

Ardakany, A. R., Ay, F., & Lonardi, S. (2019). Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics, 35*, i145-i153. doi:10.1093/bioinformatics/btz362

Bailey, C. E., Hu, C.-Y., You, Y. N., Bednarski, B. K., Rodriguez-Bigas, M. A., Skibber, J. M., … Chang, G. J. (2015). Increasing disparities in age-related incidence of colon and rectal cancer in the United States, 1975-2010. *JAMA Surgery*, *150*(1), 17–22. https://doi.org/10.1001/jamasurg.2014.1756

Bala, N. (2020, January 2). Why are you publicly sharing your child's DNA information? *The New York Times*. Retrieved from https://www.nytimes.com/

Bijker, W., Bönig, J., van Oost, E. (1984). The social construction of technological artefacts. *Zeitschrift für Wissenschaftsforschung, 2*, 39-52.

Bijker, W., & Pinch, T. (1984). The social construction of facts and artifacts: or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science, 14*, 399-441. doi:10.1177/030631284014003004

Cairns, J., Orchard, W. R., Malysheva, V., & Spivakov, M. (2019). Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data. *Bioinformatics, 35*, 4767-4766. doi:10.1093/bioinformatics/btz450

Fisher, C. B., & Layman, D. M. (2018). Genomics, big data, and broad consent: a new ethics frontier for prevention science. *PLOS Biology, 6*, 871-879. doi:10.1371/journal.pbio.0060073

Francis, L. P. (2014). Genomic knowledge sharing: A review of the ethical and legal issues. *Applied & Translational Genomics, 3*, 111-115. doi:10.1016/j.atg.2014.09.003

Garrison, N. A. (2013). Genomic justice for the Native Americans: Impact of the Havasupai case on genetic research. *Science, Technology, & Human Values, 38*, 201-223. doi:10.1177/0162243912470009

Johnson, S. B., Slade, I., Giubilini, A., Graham, M. (2020). Rethinking the ethical principles of genomic medicine services. *European Journal of Human Genetics, 28*, 147-154. doi:10.1038/s41431-0195-0507-1

Lai, Q., Li, Q., He, C., Fang, Y., Lin, S., Cai, J., … Liu, S. (2020). CTCF promotes colorectal cancer cell proliferation and chemotherapy resistance to 5-FU via the P53-Hedgehog axis. *Aging, 12*, 16270-16293. doi:10.18632/aging.103648

Rao, S. S. P., Huang, S., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K., … Aiden, E. L. (2017). Cohesin loss eliminates all loop domains. *Cell, 171*, 305-320. doi:10.1016/j.ell.2017.09.026

Rawla, P., Sunkara, T., Barsouk, A. (2019). Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol, 14*, 89-103. doi:10.5114/pg.2018.81072

Servant, N., Varoquaux, N., Lajoir, B. R., Viara, E., Chen, C.-J., Vert, J.-P., … Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology, 16,* 259. doi: 10.1186/s13059-015-0831-x

Siegel, R. L., Miller, K. D., Sauer A. G., Fedewa, S. A., Butterly, L. F., Anderson, F. C., … Jemal, A. (2020). Colorectal cancer statistics, 2020. *A Cancer Journal for Clinicians, 70,* 145-164. doi:10.3322/caac.21601

Tang, J.-Y., Yu, C.-Y., Bao, Y.-J., Chen, L., Chen, J., Yang, S.-L., … Fang, J.-Y. (2018). TEAD4 promotes colorectal tumorigenesis via transcriptionally targeting YAP1. *Cell Cycle, 17,* 102-109. doi:10.1080/15384101.2017.1403687

Thanikachalam, K., & Khan, G. (2019). Colorectal cancer and nutrition. *Nutrients*, *11*. doi:10.3390/nu1101016

Thomas, Z. (2020). *Workflow describing pipeline function*. [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Thomas, Z. (2020). *Reasons for and against publicly available genomics data*. [Figure 2]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Wang, Z., Civelek, M., Miller, C., Sheffield, N., Guertin, M. J., & Zang, C. (2018). BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics, 34,* 2867-2869. doi:10.1093/bioinformatics/bty194