Towards Causal Analysis of Genetic Factors for Colorectal Cancer

David Miller
causalens
Hana Chockler (hana@causalens.com)
causalens https://orcid.org/0000-0003-1219-0713
Andrew Lawrence
causalens
Daniel McNamee
causalens
Nicholas Chia
Mayo Clinic

Brief Communication

Keywords: colorectal cancer, genetic factors, dependencies

Posted Date: October 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-967255/v1

Towards Causal Analysis of Genetic Factors for Colorectal Cancer

David W. Miller[1], Hana Chockler*[1,2], Andrew R. Lawrence[1], Daniel C. McNamee[1], and Nicholas Chia[3,4,5]

[1]causaLens, London, UK
[1]{*david.miller,hana,andrew,dan*}*@causalens.com*
[2]Department of Informatics, King's College London, UK
[3]Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA
[4]Division of Surgical Research, Department of Surgery, Mayo Clinic, Rochester, MN, USA
[5]Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, MN, USA
[3]*chia.nicholas@mayo.edu*

Abstract

We present preliminary results and outline future work for the causal analysis of genetic factors influencing the occurrence and severity of colorectal cancer. The findings are based on publicly available datasets. We show that using relatively simple methods we are able to detect meaningful dependencies, reflecting the current biomedical knowledge about the causes of cancer and related conditions (such as infection and sepsis). We also argue that a deeper analysis, taking into account gene regulation, would require a significantly larger dataset.

1 Introduction and Motivation

Colorectal cancer is common (147,950 new cancers annually) and lethal (over 53,000 deaths each year) [1]. In order to combat this deadly disease, screening colonoscopy, where the goal is to identify and remove precursor legions that may lead to colorectal cancer, has been used for cancer prevention for almost fifty years [2]. The success of this approach rests largely on the recognition that adenomas are causally linked to carcinogenesis. By recognizing the causal lesions and removing them, we effectively prevent cancer from ever forming, leading to an ideal approach to managing an incredibly deadly and common disease.

The power of this causal understanding is clear. However, this approach has limitations. The colorectal cancer incidence rate appears to be increasing in younger individuals (2% annually for ages 50 and under [1]), making the reliance on screening colonoscopy both more challenging and less cost-effective with a broader population to screen. In addition, the only recourse for preventing disease in higher-risk patients has been shorter and shorter screening intervals. Despite these efforts, interval cancers are reported at frequencies of $3.4 - 7.6\%$ [3, 4], even in the face of adequate bowel preparation and seasoned endoscopists [5]. For these reasons, reliable early detection via non-invasive methods, such as blood or stool testing, would significantly improve the long-term prospects of patients with colorectal cancer.

Problem Statement In this paper, we make the first steps towards a causal analysis of the expression of genes and proteins on the manifestation and stage of colorectal cancer. In order to avoid observer bias [6], we adopt a strictly data-driven approach. Therefore, we do not consider any prior knowledge or domain expertise when performing this initial analysis.

*Corresponding author

The data is obtained from The Cancer Genome Atlas (TCGA) project, specifically the Gene Expression by RNAseq (IlluminaHiSeq) data and RPPA Protein Expression data from the Colon and Rectal Cancer (COADREAD) cohort. The TCGA COADREAD data was downloaded via the UCSC Xena platform [7]. The main challenge in analysing this data lies in the high number of candidate RNA expressions and proteins compared to the orders of magnitude smaller number of patients. Additionally, there is not a perfect correspondence between the RNA and protein data tables, which reduces the total number of patients further.

An additional challenge arises from the extreme imbalance in the data, as the dataset does not contain any healthy individuals; therefore, all the observations are from patients with varying severity of colorectal cancer. This extreme imbalance within the dataset can lead to spurious correlations and to missing causal connections [8]; it is inherent to the medical domain, as healthy individuals do not undergo colon resection surgery, and hence any meaningful progress would need to address this imbalance first.

A final challenge is due to the inability to intervene. A common method to discover the underlying causal structure of a system is by performing interventional experimentation, often in the form of randomized control trials. These are ubiquitous in the medical domain, particularly around testing for drug efficacy. However, it is not possible (or ethical) to intervene in this scenario to give someone colorectal cancer or change the concentration of specific RNA molecules or proteins in their cells. Therefore, we must perform this analysis using exclusively observational data.

Related Work There has been a recent rise in the effort to apply AI techniques, and in particular machine learning, to biological data, and specifically to cancer research. The unique challenges posed by the cancer data, namely that it is high-dimensional and low in sample size, as well as the fact that, similarly to other biological data, interventions are not possible or not ethical, motivate different approaches than the traditional causal inference ones.

A recent paper on machine learning for cancer research presents a platform for causal inference suitable for this type of data and its application to colorectal cancer, with the goal of determining causal drivers that differentiate between two subtypes of colorectal cancer [9]. Kalantari et al. [10] use inverse reinforcement learning to gain insights about cancer progression from genome data. Farahmand et al. [11] use causal inference to identify transcriptional regulators.

2 Background and General Approach

The motivation underlying our work is that a principled approach based on *causal counterfactual reasoning* [12] is the correct method for obtaining meaningful results. This approach is based on postulating the existence of an underlying causal model, encompassing the causal dependencies between variables. The process of causal discovery [13] amounts to discovering this model by performing observations and experiments on the variables. Essentially, the existence of a causal dependency between variables is demonstrated by intervening on the candidate cause and observing the changes in the outcome variable. This approach can be extended to interventions on more than one variable, where we observe the effect of changes in several variables on the outcome. If direct interventions are impossible, as they are in this dataset, we can attempt to partially replace it by the analysis of the existing data, where we search for the records that differ only in the values of candidate causes and examine the value of the outcome. This approach requires a large and sufficiently varied dataset. In particular, the number of features (candidate causes) should be small in comparison to the number of records, to guarantee a sufficient variation in their values. In this section, we outline existing approaches to estimate causality from observations and describe the approach we propose for the analysis of the colorectal cancer dataset.

Intractability of causality The computational cost of causality-based methods is high and increases dramatically following even a modest increase in the number of variables (p). For our case p =20,531. As we are unable to intervene, the goal is to perform causal discovery on observational-only data, which in general, is an NP-hard problem [14]. A brief overview highlighting the computational complexities of popular methods follows; please see [13] for a thorough review of causal discovery.

Constraint-based methods, such as the Peter and Clark (PC) [15] and fast causal inference (FCI) [16, 17] algorithms, utilize conditional independence tests to discover the causal structure. At the worst case (dense causal structure), these scale exponentially in the number of variables ($\mathcal{O}(c^p)$). This does not capture the complexity of the conditional independence tests either, which may be expensive to calculate when trying to identify more complex relationships between random variables than a simple linear one.

Score-based methods, which include greedy, e.g., [18], and exhaustive, e.g., [19], search methods, are another key category. These may scale worse than constraint-based methods as some require all permutations to be enumerated, which equals the factorial of the number of variables ($\mathcal{O}(p!)$). Hybrid methods combining constrained-based and score-based methods have also been proposed.

Finally, methods utilizing functional causal models, such as LiNGAM [20] and NOTEARS [21], are also popular but they make assumptions on the types of functional dependencies between variables. To allow for faster convergence, simple linear models are often used, which greatly limits the expressiveness of the models and likely misses complex relationships between variables.

The functional causal model methods also scale poorly with respect to the number of variables. For example, the costs of the DirectLiNGAM [22] and ICA-LiNGAM [20] algorithms scale as at least $\mathcal{O}(np^3+p^4)$, where n is the number of samples/observations. If $p >> n$, as we have in this case as $n = 283$, then twice the number of RNA molecules considered results in sixteen times the computational cost.

One cannot blindly apply popular causal discovery methods to this problem. In order to find meaningful drivers of protein expression, we must restrict the number of RNA candidates from 20,531 to a more manageable quantity.

Assumptions Two common assumptions made when performing causal discovery are: (1) Causal Markov Condition and (2) Causal Faithfulness Condition [15]. We make these assumptions in our analysis. Without going into the formal mathematical definitions, when combined it implies "that two [random] variables are directly causally related *if and only if* they are not conditionally independent given any subset of the remaining variables" [19]. This subset can be the empty set. Therefore, a common first step of constraint-based causal discovery is to measure if two random variables are independent given the empty set. We utilize Spearman's rank correlation as this independence test. It makes no assumption on the underlying probability distributions of the random variables and only assumes a possible monotonic dependency between the variables, which is a natural assumption for this data. The null hypothesis is that the joint distribution factorizes, i.e., $p(x, y) = p(x)p(y)$, which means x and y are independent. Calculating a p-value from the null distribution represents the likelihood of calculating the measured correlation coefficient given that the null hypothesis is true. Therefore, a really small p-value means it is unlikely that the null hypothesis is true and we can assume the two random variables are dependent.

Proposed approach We performed several types of analysis as detailed below in § 3. We calculated Spearman's rank correlation between all pairs of RNA and protein expressions focusing on the ataxia-telangiectasia mutated (ATM) protein as our primary example. The ATM protein kinase has been extensively studied for its role in the DNA damage response, and there is increasing evidence that ATM plays an important role in other cellular processes, including carbon metabolism. Carbon metabolism is highly disregulated in cancer due to the increased need for cellular biomass. Some therapeutic strategies for cancer involve the development of ATM inhibitors. We also computed an intercorrelation between RNA expressions associated with a given protein (specifically, ATM) and examined the resulting clusters. The expectation is that the clusters provide a smaller space to perform more expensive causal discovery techniques. Finally, we analysed the bimodal activations, which infer whether an RNA expression is on or off in a given record, based on the bimodal graph of the correlation with the protein.

3 Detailed Methodology and Results

3.1 Data characteristics

The TCGA Gene Expression dataset contains the log-normalised gene-level transcription estimates of 20,531 distinct RNA molecules for 434 distinct individuals, while the Protein Expression dataset contains the RBN-

normalized RPPA values of 131 distinct proteins for 464 individuals.

We join the datasets by individual, however despite being from the same cohort of patients, there is not a complete overlap between individuals in the two. Only 283 individuals were present in both datasets, reducing the overall amount of data available for analysis by roughly a third.

3.2 Bivariate filtering by independence between RNA and protein expression

As a first measure to infer relationships between RNA expression and protein expression, we look at the bivariate correlations calculated across individuals as a statistical test to infer independence. We calculate Spearman's rank correlation [23] and associated significance between all pairs of RNA and protein expressions. Spearman's rank correlation between two random variables can be measured by calculating the Pearson Correlation [24] between the ranks of the values in the observations of each random variable.

For each of the 131 protein expressions in the dataset we look at which RNA molecules expressions exhibit a statistically significant correlation. We then use this significance as a filter for RNA molecules independent to the protein, in order to reduce the number of features to be included in the problem.

Transforming from raw values to ranks reduces the impact of outliers and non-linearity on the Pearson's R correlation metric, making Spearman Rank Correlation a more robust measure for use in noisy datasets.

Table 1 shows the percentile of significant RNA relationship counts across proteins, depending on the p-value chosen as a cut-off between dependence and independence. For example, using a p-value cutoff of 0.01 (1e-2), the median (50th percentile) protein has 2492 significant relationships amongst the $20,531$ possible RNA molecules.

percentile p-value	0.05	0.25	0.50	0.75	0.95
1e-1	3610	5104	6498	7774	10110
1e-2	697	1526	2492	3841	6254
1e-3	118	450	901	2172	4240
1e-4	18	118	318	1193	3038
1e-5	2	30	102	678	2262

Table 1: Percentiles of significant RNA relationship counts, given p-value, across all proteins

In statistical analysis it is common practice to use p-values to determine the reliability of a correlation between two variables. In practice, a limit of 0.05 is often used as a cutoff, representing a maximum 5% chance that a relationship is caused by random noise rather than by a true correlation. However, the limitations of this approach are significant and well documented, especially when managing datasets with a large number of variables.

As we can see from Table 1, selecting by strength of correlation or significance requires extremely low p-value cutoff to filter to a reasonable numbers of relevant RNA expression variables.

The issue here is two-fold. Firstly, and most widely understood, by simple definition we can expect 1% of the relationships that survive the significance cutoff will not be true at all, and will only appear to be correlated by chance, given noise in our dataset. Secondly, and more importantly, an unknown proportion of the remaining 99% relationships will not describe true causal relationships (where expression of the RNA molecule in question drives Protein expression), but instead describe confounded relationships (where both expression of the protein and of the RNA molecule are driven by a third variable). An example of a confounded relationship would be where both the RNA and protein are produced in separate parts of the same physiological pathway, however, it is also possible that both RNA and protein expression are driven by some other, unmeasured, variable.

Using a p-value cutoff of 0.01 we found that the expressions of 100 different proteins were significantly correlated with the expressions of more than 1000 different RNA molecules. Conservatively, we can expect 990 'true' correlations per protein in this case. While the pathways that include generating proteins may be complex, we cannot claim to have produced a practical range of RNA candidates to investigate in detail.

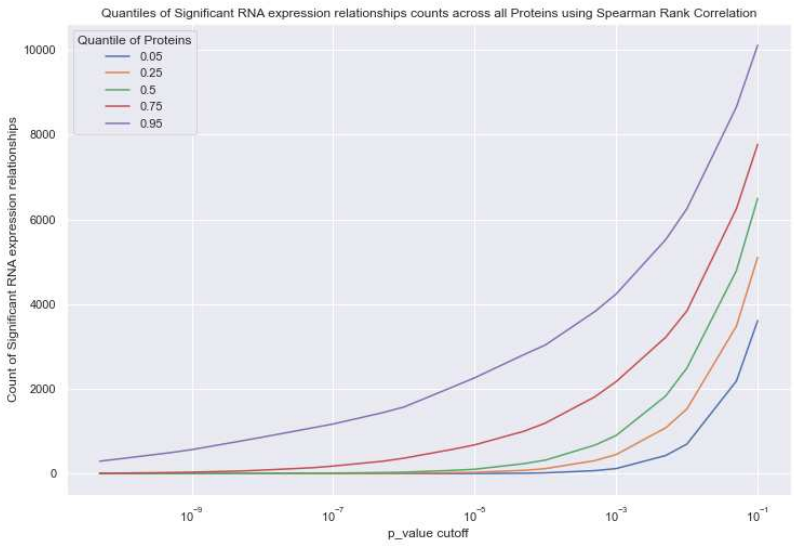

Figure 1: Percentiles of significant RNA relationship counts, given p value, across all proteins

A straightforward next step might be to decrease the p-value cutoff further, to limit the set of variables that we need to consider. However, this risks filtering out true relationships that are have weaker correlations, either due to a complex relationship or because they are subject to noise or small sample size effects.

At this point, simple Bivariate Filtering using expression correlations has reached the limits of its use and we are forced to consider more sophisticated methods to remove spurious correlations.

3.3 Intercorrelation between RNA expressions associated with a given protein

To investigate this problem further, we narrowed our focus to the expression of individual proteins. For the purposes of this communication we discuss results relating to expression of ATM, as mentioned in § 2.

For the purposes of illustration, we focus on the intercorrelation between the expressions of 50 RNA molecules that are most correlated with ATM expression.

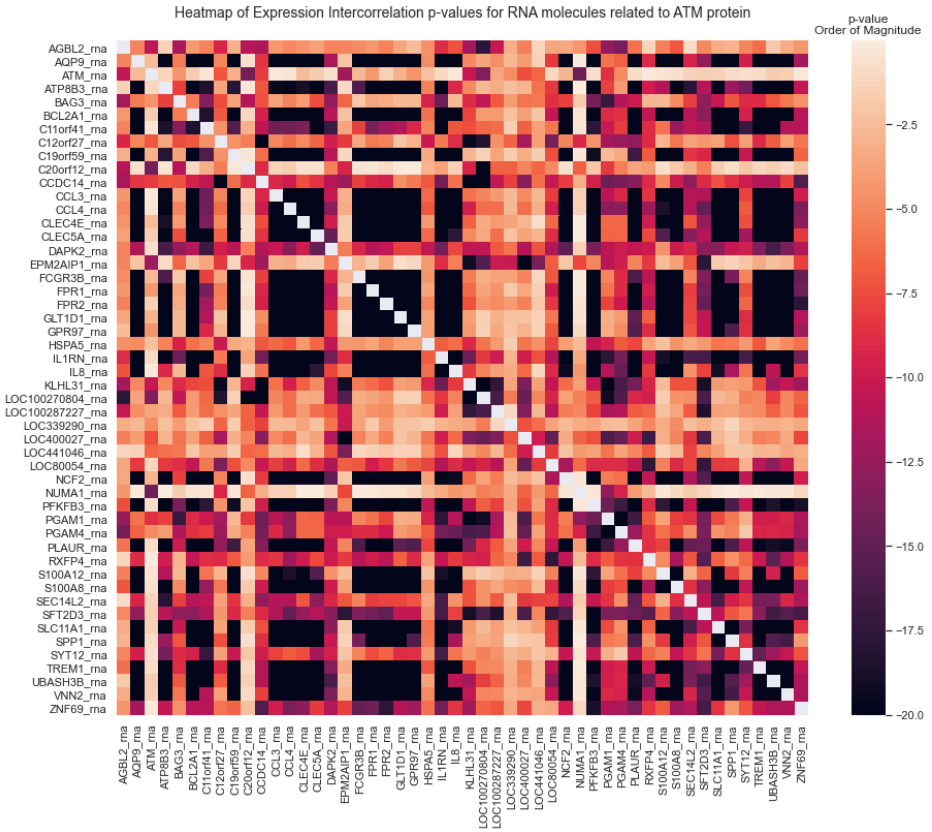

Figure 2: Expression Intercorrelation Heatmap for top 50 RNA molecules related to ATM protein

Figure 2 shows a heatmap of the intercorrelation p-values (log scale) for these RNA molecule expressions. The multiple dark patches show that many RNA molecules' expressions are highly correlated.

Interestingly, we note that many of highly inter-correlated RNA molecule expressions are grouped together into larger blocks. This is due to the fact that the RNA molecule columns are ordered alphabetically, and molecules with similar names appear to share a similar function (i.e. with stronger intercorrelations).

Although we are selecting for RNA expressions that are all strongly correlated with ATM expression, and therefore should expect some overlap between the top most correlated features, we have placed no bounds on the sign or the strength of that correlation, only its significance, and the existence of multiple separate and non-correlating blocks shows that there is a high redundancy in the dataset beyond the RNA molecules' shared expression relationship with the ATM protein.

Further, such blocks of clustered RNA molecules imply that information is being shared within them. This type of relationship hints at a confounder within each block, or a hidden confounder not included in the dataset, which is driving the remaining block members' expressions.

This shows how the Spearman Rank correlations between RNA molecules' expressions can identify similarities between the RNA molecules without any prior data, and implies that such intercorrelations can be used to identify RNA molecules that share a confounder.

3.4 Clustering of RNA expressions for Dimensionality Reduction

In order to use these intercorrelations to reduce the dimensionality of the problem, we propose clustering similar RNA molecules together according to their expressions within the COADREAD Colon and Rectal Cancer dataset.

Agglomerative Clustering, or Hierarchical Clustering, is a method that produces clusters by repeatedly merging the closest individuals or existing clusters according to some method to determine cluster similarity,

based on some distance metric (between individuals) [25]. Agglomerative clustering allows intuitive visualisation of clusters, and the ability to dynamically choose the number of clusters to be generated, based on the value of each new merge.

By choosing an appropriate number of clusters, we can group together the RNA molecules that share extremely similar patterns of expression in the dataset, and take a representative, or an aggregate, from each. Applying an initial causal analysis on this smaller dataset will then allow us to eliminate entire clusters thought to be non-causally or spuriously correlated, and focus on smaller clusters of potential causal drivers. Reducing the dimensionality of the dataset makes the application of causal tools feasible in terms of computational costs.

To produce a cluster hierarchy for the top 50 RNA molecules associated with the ATM protein, we use the Ward method to determine cluster similarity, based on a measure of Euclidean distance across the RNA expression per individual within the dataset. Figure 3 shows the dendrogram representing this hierarchy.

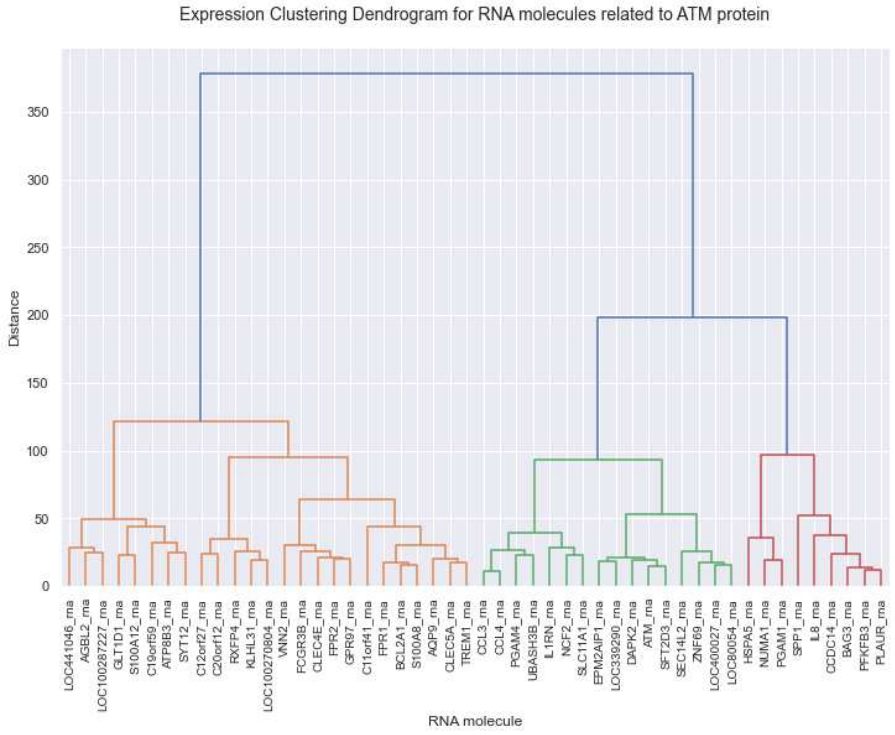

Figure 3: Expression Clustering Dendrogram of top 50 RNA molecules associated with the ATM protein

The inverted-U shaped joins on a Dendrogram show merges between clusters, and the height of the legs of these merges represents the difference between the two clusters. For example, we can see how the CCL3 and CCL4 RNA molecules (green) are considered closer to each other than the nearest neighbouring pair of RNA molecules, PGAM4 and UBASH3B, and that the algorithm only slightly prefers to merge these individuals into pairs before merging the pairs into a cluster of 4.

As with the heatmap of intercorrelation p-values in figure 2, the dendrogram shows the high redundancy between the RNA expression features. A significant proportion of the cumulative distance between consecutive clusters is concentrated between the final 6 merges (i.e. between the top 7 clusters, compared to the total of 50 input features).

While the cumulative distance is not as direct a measure of information compression as calculating cumulative variance of features from the eigenvalues of produced in Principal Component Analysis, it is still a measure of total Ward distances within the cluster hierarchy, which equals the sum of minimum increases of within-cluster variance from the remaining sub-cluster merges.

This density of cumulative distance found within a small number of clusters therefore implies that there is significant room for compression within the RNA expression variables.

More completely, when applied to the entire COADREAD Colon and Rectal Cancer dataset, we find that reducing the 20,500 RNA expression features to 4,100 agglomerative clusters (20% of the dataset's original size), retains 50.2% of the cumulative distance between consecutive clusters.

3.5 Bimodal Activations

Bimodal Activations are a widely used transformation in analysis of RNA expression.

The transform assumes that globally, across all RNA molecules and all individuals in a dataset, there is a bimodal distribution of expression levels. This is interpreted as defining a cutoff above or below which an RNA molecule is considered 'active' or 'inactive' respectively. Around the cutoff there is also often assumed to be a grey area of uncertain or noisy expressions that are ignored by the transform, which are to be excluded.

We attempted to expand our analysis of Bivariate Filtering to include Bimodal Activations of the RNA expressions in the dataset. We use Point Bi-Serial Correlations to measure the strength of relationship between RNA Bimodal Activations and Protein Expression and, as with the Spearman Rank Correlations, examine the top 50 most correlated RNA molecules for illustration. Point Bi-Serial Correlation measures the relationship between the state of a binary variable and the value of a continuous variable, and is a special case of Pearson Correlation.

In order to define the cutoffs for bimodal activation quantitatively we train a Mixed Gaussian Model on the distribution of RNA expressions across all RNA molecules and all individuals in the dataset. By asserting that the Bimodal Activations are described by 3 Gaussian distributions (the distribution of 'inactive' expressions, the distribution of 'active' individuals, and a distribution for noise between the peaks), we calculate cutoffs as the crossover points of these distributions, with expressions above 6.35 counting as 'active' (28%), expressions equal to or below 0.05 counting as 'inactive' (57%). Approximately 15% of the data falls between these cutoffs, and is discarded as noise.

In addition to the proportion discarded as noise, many RNA features are found to be either 'active' or 'inactive' across all individuals in the dataset, and therefore have no relationship with protein expression. This results in only 2878 RNA molecules out of the initial 20,000 having a calculable expression correlation with the ATM protein, reducing the available RNA features by 86% before Bivariate Filtering could be applied. This loss of data significantly impacted the analysis, with only 1 out of the top 50 most significant RNA molecules according to Spearman Rank Expression Correlations even having a calculable Point Bi-Serial Correlation using Bimodal Activations.

To address this problem, we repeated the analysis without discarding noisy data and use a rule of thumb cutoff expression of 6.00 to divide between the 'active' (41%) and 'inactive' (59%) expressions. This still results in only 8389 RNA molecules out of the initial 20000 having a calculable expression correlation with the ATM protein, and only 28 out of the top 50 most significant RNA molecules according to Spearman Rank Expression Correlations even having a calculable Point Bi-Serial Correlation.

Reducing a continuous variable (RNA expression) to a binary one (RNA Bimodal Activation) clearly eliminates a large amount of information from the dataset, and such transformations should be used with care. In a dataset of this small size, and in a problem where we a searching for relationships amongst a large number of features, this information loss is unmanageable.

3.6 Towards Causal Analysis

As mentioned in § 2, analysis based upon causal counterfactual reasoning requires a large and sufficiently varied dataset. The COADREAD dataset used in this preliminary analysis does not meet these criteria. A natural next step would be to combine several datasets from the TCGA for different types of cancer in order to provide a larger and more balanced dataset.

The analysis performed in § 3.4 identified clusters of RNA molecules which allows us to move forward using causal discovery methods, as referenced in § 2, for this reduced dimensionality problem. We will decompose the problem in two ways: (1) independent data clusters, and (2) a coarse-to-fine approach. The first method decomposes the model into loosely related components, see [26] for a formal definition of decomposable models.

For the coarse-to-fine approach, we can learn a high-level causal model where each node represents a cluster. The definition of each cluster node can be a single representative from the cluster, the centroid, or a linear/non-linear combination of all elements, which can be found using classical dimensionality reduction techniques. Once the structure between clusters has been learned, we can zoom in on the local structure within clusters. The dependencies between clusters can also be further refined.

4 Conclusions and Future Work

Our findings as described in this paper are aligned with the current understanding of the role of different RNA expressions, thus supporting our claim that a screening based on the blood test results is meaningful. We have also shown that the analysis of bimodal activations is not aligned with the current understanding of the roles and connections between RNA expressions, due to the small size of the dataset after a sizeable subset of records is discarded (those where we cannot deduce with any certainty whether the RNA expression is on or off). We deduce that the bimodal activation analysis is applicable only to sufficiently large datasets.

Our preliminary analysis is based on a relatively small publicly available dataset, hence the findings are limited.

In future work, we plan to perform more complex types of causal analysis and inference on larger datasets, as we predict that they have a potential to uncover previously unknown or not well-understood connections between RNA and occurrence of cancer.

Acknowledgements

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

[1] Rebecca L. Siegel, Kimberly D. Miller, Ann Goding Sauer, Stacey A. Fedewa, Lynn F. Butterly, Joseph C. Anderson, Andrea Cercek, Robert A. Smith, and Ahmedin Jemal. Colorectal cancer statistics. *CA: A Cancer Journal for Clinicians*, 70:145–164, 2020.

[2] Wolff W.I. Colonoscopy: history and development. *American Journal of Gastroenterology*, 84, 1989.

[3] Petrini J.L. Colonoscopy surveillance intervals: we are not there yet. *Gastrointestinal Endoscopy*, 85:1271–1272, 2017.

[4] Singh S., Singh P.P., Murad M.H., Singh H., and Samadder N.J. Prevalence, risk factors, and outcomes of interval colorectal cancers: a systematic review and meta-analysis. *American Journal of Gastroenterology*, 109:1375, 2014.

[5] Nayor J., Saltzman J.R., Campbell E.J., Perencevich M.L., Jajoo K., and Richter J.M. Impact of physician compliance with colonoscopy surveillance guidelines on interval colorectal cancer. *Gastrointestinal Endoscopy*, 85:1263–1270, 2017.

[6] Kamal Mahtani, Elizabeth A. Spencer, Jon Brassey, and Carl Heneghan. Catalogue of bias: observer bias. *BMJ Evidence-Based Medicine*, 23:23–24, 2018.

[7] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature Biotechnology*, 38(6):675–678, 2020.

[8] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 179–186, 1997.

[9] Samuel Budd, Arno Blaas, Adrienne Hoarfrost, Kia Khezeli, Krittika D'Silva, Frank Soboczenski, Graham Mackintosh, Nicholas Chia, and John Kalantari. Prototyping crisp: A causal relation and inference search platform applied to colorectal cancer data. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 517–521, 2021.

[10] John Kalantari, Heidi Nelson, and Nicholas Chia. The unreasonable effectiveness of inverse reinforcement learning in advancing cancer research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):437–445, 2020.

[11] Saman Farahmand, Corey O'Connor, Jill A Macoska, and Kourosh Zarringhalam. Causal Inference Engine: a platform for directional gene set enrichment analysis and inference of active transcriptional regulators. *Nucleic Acids Research*, 47(22):11563–11573, 2019.

[12] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[13] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

[14] David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287—-1330, 2004.

[15] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

[16] Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:211–252, 1999.

[17] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

[18] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[19] Ni Y Lu, Kun Zhang, and Changhe Yuan. Improving causal discovery by optimal bayesian network learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8741–8748, May 2021.

[20] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

[21] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[22] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

[23] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[24] Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[25] Mirco Nanni. Speeding-up hierarchical agglomerative clustering in presence of expensive metrics. In *Advances in Knowledge Discovery and Data Mining, 9th Pacific-Asia Conference, PAKDD*, volume 3518 of *Lecture Notes in Computer Science*, pages 378–387. Springer, 2005.

[26] Dalal Alrajeh, Hana Chockler, and Joseph Y. Halpern. Combining experts' causal judgments. *Artif. Intell.*, 288:103355, 2020.