

Genome Gerrymandering: optimal division of the genome into regions with cancer type specific differences in mutation rates

Adamo Young^{1,†,‡,*}, Jacob Chmura[†], Yoonsik Park[†], Quaid Morris^{†,‡,*}, Gurnit Atwal^{‡,*}

[†]*Department of Computer Science, University of Toronto,
40 St. George Street, Room 7224
Toronto, ON M5S 2E4, Canada*

[‡]*Donnelly Centre for Cellular and Biomolecular Research,
160 College Street, Room 230
Toronto, ON M5S 3E1, Canada*

^{*}*Vector Institute for Artificial Intelligence,
661 University Ave, Suite 710
Toronto, ON M5G 1M1, Canada*

¹*E-mail: adamo.young@mail.utoronto.ca*

The activity of mutational processes differs across the genome, and is influenced by chromatin state and spatial genome organization. At the scale of one megabase-pair (Mb), regional mutation density correlate strongly with chromatin features and mutation density at this scale can be used to accurately identify cancer type. Here, we explore the relationship between genomic region and mutation rate by developing an information theory driven, dynamic programming algorithm for dividing the genome into regions with differing relative mutation rates between cancer types. Our algorithm improves mutual information when compared to the naive approach, effectively reducing the average number of mutations required to identify cancer type. Our approach provides an efficient method for associating regional mutation density with mutation labels, and has future applications in exploring the role of somatic mutations in a number of diseases.

Keywords: Genome Segmentation, Tumour Classification, Dynamic Programming, Information Theory

1. Introduction

Somatic cells are exposed to multiple mutational events throughout their lifetime. The phenotypic effect of these mutations varies, and the aggregate effect of all somatic mutations has been implicated in the development of a number of neurodegenerative diseases and cancer [1], [2]. Somatic mutations are generated by multiple mutational processes ranging from exogenous mutagens to endogenous DNA repair mechanisms. Mutational processes are mechanisms for generating different types of mutations, and their signal in the genome is manifested through different mutational signatures. Single base substitution signatures (SBS) summarize muta-

tional processes that generate single nucleotide variants (SNVs) by grouping mutations based on short-range sequence characteristics such as trinucleotide context [3]. Mutational signatures differ with their relative timing, with cell-extrinsic signatures often occurring earlier in tumorigenesis, and cell-intrinsic signatures occurring in the later stages [4]. Mutational signatures also show varied activity across different cancer types, with certain signatures having a very strong association with specific cancer types (SBS4 in lung cancer) [3].

Mutational processes have differential activity across the genome, and consequently, mutation rates for different signatures varies across the genome[5]. Process-specific, regional mutation rates are influenced by phenomenon at multiple scales [6]. At the megabase-pair level, mutation rate is strongly influenced by chromatin accessibility and replication timing, largely due to the differential activity of mismatch repair mechanisms in these regions [7]. At the level of individual genes, the level of transcription has a strong relationship with mutation density, likely due to the activity of transcription coupled repair mechanisms. On a local-scale, nucleosome occupancy is associated with an enrichment for a number of mutational processes including UV damage (SBS7), and oxidative damage (SBS17)[8], [9], while regions depleted of nucleosomes are enriched for mutations likely caused by tobacco (SBS4) [10]. This association between mutation rate, chromatin accessibility, and mutational processes give rise to a relationship between cancer-type and regional mutation density. Recently, we used the association of regional mutation density and cancer-type to develop a deep-learning classifier that uses regional mutation density to differentiate between 24 cancer types with an accuracy of 91% [11]. This work suggests that the relationship between mutation density and chromatin state is stronger than previously suggested, and provides motivation for further investigating the relationship between mutational processes and genome organization.

While the association between mutation rate and genome organization at the megabase scale is well characterized, genome organization can be studied at a much higher resolution. Genome organization is strongly influenced by a large variety of chromatin marks that include histone modifications, histone variants, and chromatin accessibility. Combinations of chromatin marks in specific spatial contexts can be grouped into functional elements. These elements include promoters, enhancers, transcribed, repressed and repetitive regions, and vary across different cell-types. As there is a strong relationship between mutation density and cell-type, examining regional mutation density may provide information about the distribution of functional elements across the genome.

In this work, we investigate the relationship between regional mutation density and genome organization by segmenting the genome based solely on the differential activity of mutational processes. To do so, we present a novel, information theoretic algorithm for associating mutation density with cell-type. Using this algorithm, we show that an optimal segmentation of the genome significantly increases information content between mutation density and cell-type, and we explore the relationship between optimal segmentation and functional elements.

2. Methods

2.1. Overview

The goal of Genome Gerrymandering is to split the genome into sections that differ in their mutation rates among different cancer types. We do this by labelling somatic mutations by cancer type in a large cohort of cancer samples. We then map these mutations to a single set of reference genome coordinates and partition this meta-cancer genome into sections so that the count of mutations per cancer type differs by as much as possible. Formally, we use a modification of Bellman’s K-segmentation algorithm [12] to maximize the mutual information between the segment assignment of a mutation and its cancer type label.

2.2. Data

All patients who donated to the Pan-cancer Analysis of Whole genomes (PCAWG), data set consented to international data sharing and secondary analysis of their genomes [13]. Permission to reanalyze these data was granted by the University of Toronto’s Research Ethics Board.

Variant calls were downloaded from Synapse (<https://www.synapse.org/#!/Synapse:syn2351328/wiki/62351>); the syn numbers that follow refer to Synapse data set IDs. Consensus Somatic SNV (syn7118450) file covers 2778 whitelisted samples from 2583 donors. Tumour histological classifications were reviewed and assigned by the PCAWG Pathology and Clinical Correlates Working Group (annotation version 6, August 2016; syn7253568). Kataegis events and SNV files containing the all SNVs caused by kataegis events were provided by the PCAWG Evolution and Heterogeneity Working Group (annotation version 10, August 2018) and were downloaded from (<https://www.synapse.org/#!/Synapse:syn12978907>).

We additionally made use of variants from 1178 tumour whole-genomes described in [3]. These data comprise 11 tumour types that overlap with PCAWG types collected from a variety of published studies, non-PCAWG donors in the ICGC data portal (<http://dcc.icrg.org>), and donors present in the COSMIC database (<http://cancer.sanger.ac.uk/cosmic>).

2.3. Algorithm Overview

The Genome Gerrymandering algorithm is a dynamic programming algorithm that finds a segmentation that maximizes the conditional mutual information $I(T; B | C)$, where tumour type (T), segment location (B), and chromosome (C) are all categorical random variables. Let K be the desired number of segments (provided by the user), N be the number of distinct mutation positions in the genome. For a given choice of K , our algorithm seeks to maximize $I(T; B | C)$ by changing the segmentation boundaries θ between mutations. The segmentation boundaries define a categorical distribution $B | C \sim p_{\theta}(b | c)$ that represents the probability that a mutation is inside a segment b given that it is on chromosome c . While our overall goal is to optimize $I(T; B)$, we choose to condition on C to reduce run time. We assume that the segments must be contiguous and cannot span multiple chromosomes.

2.4. Objective Function

It is known from the definition of conditional mutual information that

$$I(T; B | C) = H(T | C) + H(B | C) - H(T, B | C) \quad (1)$$

$$I(T; B | C) = \sum_c p(c) I(T; B | C = c) \quad (2)$$

Since $T \perp\!\!\!\perp B | C$ (T is conditionally independent of B given C), the value of $H(T | C)$ does not change with respect to θ . It is therefore clear that

$$\begin{aligned} \arg \max_{\theta} I(T; B | C) &= \arg \max_{\theta} [H(B | C) - H(T, B | C)] \\ &= \arg \max_{\theta} -H(T | B, C) \\ &= \arg \max_{\theta} - \sum_c p(c) H(T | B, C = c) \end{aligned} \quad (3)$$

Intuitively, maximizing this objective works by minimizing uncertainty about tumour type in a segment, $H(T, B | C)$, while encouraging equal segment size by maximizing uncertainty about segment location within a chromosome, $H(B | C)$. Using equation 2, the problem of maximizing $I(T; B | C)$ naturally decomposes into independent optimization problems $I(T; B | C = c)$ for each chromosome, since θ has no effect on $p(c)$. To allow this decomposition, we assume that $p(c) = N_c/N$ (where N_c is the number of mutations on chromosome c) and assign each chromosome $K_c = p(c)K$ segments (rounded).

Since our random variables are all categorical, it is possible to compute these entropies directly during optimization, using mutation counts to estimate probabilities (see equation 7).

$$\begin{aligned} -H(T | B, C = c) &= - \sum_b p(b | c) H(T | B = b, C = c) \\ &= \sum_b p(b | c) \sum_t p(t | b, c) \log p(t | b, c) \end{aligned} \quad (4)$$

Since $H(T | C = c)$ remains unchanged throughout optimization, it can be computed afterwards to find $I(T; B | C)$. To compute $I(T; B)$ for the final segmentation, the following equation can be used [14]:

$$I(T; B) = I(T; B | C) - H(C | T) - H(C | B) + H(C | T, B) + H(C) \quad (5)$$

$I(T; B | C)$ is given by equation 2, and it is trivial to compute the entropies explicitly from the mutation counts once the segmentation is provided.

2.5. Mutation Preprocessing

PCAWG mutations that were identified as part of a kataegic cluster (see section 2.2) were merged into a single mutation that took the median position of all of the mutations in the cluster.

After kataegic cluster removal, adjacent mutations were placed into groups of at most 100 distinct mutation positions and summed together, to reduce the size of the mutation array and speed up run time. A maximum group size of 3 Kb was chosen to ensure that physically far

away mutations were not placed in the same group, resulting in group sizes of approximately 50 distinct mutation positions on average.

2.6. Optimization with Dynamic Programming

We will show that it is possible to efficiently optimize the objective function with dynamic programming techniques. For each chromosome c , let us define the arrays S_c and S'_c as follows, with θ_k defining segments b_1, \dots, b_k that are contiguous and exist entirely over mutation positions $n = 1, \dots, j \leq N_c$:

$$\begin{aligned} S_c[j, k] &= \max_{\theta_k} \sum_{b=b_1}^{b_k} p(b | c) \sum_{t \in T} p(t | b, c) \log p(t | b, c) \\ S'_c[j, k] &= \arg \max_{\theta_k} \sum_{b=b_1}^{b_k} p(b | c) \sum_{t \in T} p(t | b, c) \log p(t | b, c) \end{aligned} \quad (6)$$

Note that optimizing our objective $-H(T | B, C)$ with respect to θ_K is equivalent to finding $S_c(N_c, K_c)$ and $S'_c(N_c, K_c)$ for each chromosome c . Let A_c be the mutation count array for chromosome c . Let $A_c[n, t]$ be the count of mutations at position n of tumour type t , and let $A_c[n_1 : n_2, t_1 : t_2]$ denote count summing from position n_1 to n_2 and t_1 to t_2 inclusive along each axis, $n_1 \leq n_2$ and $t_1 \leq t_2$.

$$S_c[j, 1] = p(b | c) \sum_t p(t | b, c) \log p(t | b, c) = \frac{A_c[1 : j, :]}{A_c[:, :]} \sum_t \frac{A_c[1 : j, t]}{A_c[1 : j, :]} \log \frac{A_c[1 : j, t]}{A_c[1 : j, :]} \quad (7)$$

When $k > 1$, the following recursive relationship holds:

$$S_c[j, k] = \max_{i=k-1, \dots, j-1} S_c[i, k-1] + S_c[j, 1] - S_c[i, 1] \quad (8)$$

Note that if $j < k$, $S_c[j, k]$ is undefined since it is impossible to divide j mutations into k non-empty segments. In plain terms, $S_c[j, k]$ is equal to the best scoring segmentation that uses $k-1$ segments plus the score that results from grouping the remaining mutations into a single segment. This relationship allows the problem to be optimized with dynamic programming using algorithm 1. Our algorithm also allows the user to specify a minimum segment size P , which can reduce overfitting by preventing the creation of very small segments. The algorithm has $\mathcal{O}(N_c^2 K_c)$ time and space complexity for each chromosome c , and can be run in parallel. Once S'_c is known for each chromosome, $I(T; B | C = c)$ can be computed for each chromosome using equation 1, which allows subsequent computation of $I(T; B | C)$ and $I(T; B)$. Note that this algorithm is very similar to the Bellman K -segmentation algorithm [12], with the main difference being the information objective function.

Algorithm 1: Genome Gerrymandering Algorithm**Input** : A_c , an N_c by $|T|$ array of mutations in chromosome c **Input** : K_c , number of desired segments for chromosome c **Input** : P , minimum segment size**Output:** S_c , segmentation scores (optimal score found at $S_c[N_c, K_c]$)**Output:** S'_c , optimal segmentation traceback array

```

1  $S_c \leftarrow \text{InitNegInf}(N_c, K_c)$ 
2  $S'_c \leftarrow \text{InitNegInf}(N_c, K_c)$ 
3 for  $i = P$  to  $N_c$  do
4    $S_c[i, 1] \leftarrow \text{ComputeFromCounts}(A_c, i)$ 
5 for  $k = 2$  to  $K_c$  do
6   for  $j = k$  to  $N_c$  do
7     for  $i = k - 1$  to  $j - 1$  do
8       if  $S_c[j, k] < S_c[i, k - 1] + S_c[j, 1] - S_c[i, 1]$  then
9         if  $j - i - 1 \geq P$  &  $i - 1 \geq P$  then
10           $S_c[j, k] \leftarrow S_c[i, k - 1] + S_c[j, 1] - S_c[i, 1]$ 
11           $S'_c[j, k] \leftarrow i$ 
12 return  $S_c, S'_c$ 

```

Dynamic Programming Algorithm Overview:For each chromosome c , maximize $-H(T|B, C = c)$ with respect to boundary placement θ_K score $_c(i, j) = -p(b|c) \sum_t p(t|b, c) \log p(t|b, c)$, where segment b contains mutations i to j

Create table S_c of dimensions $N_c \times K_c$:




$S_c[j, 1] = \text{score}_c(1, j)$
 \dots
 $S_c[j, k] = \text{maximum total score for } k \text{ segments, for mutations 1 to } j$
 $= \max\{S_c[i, k - 1] + \text{score}_c(i + 1, j) \mid \forall i \in \mathbb{N}, k - 1 \leq i < j\}$
 $S_c[N_c, K_c] = \text{maximum mutual information for segmentation}$

Example for $S_c[8, 4]$:


Calculate Difference in Mutual Information:

Naive Segments :



Optimal Segments :




Fig. 1. Visualization of Algorithm and Analysis on a Toy Dataset

2.7. Data Split

Before fitting the model to the data, roughly 30% of samples from both the PCAWG and the alternate dataset were selected to be a held-out test set. Note that only 11 tumour types were present in both datasets: the other tumour data was excluded for our experiments. The samples were sorted based on tumour type and then split by sample ID, with the additional restriction that samples coming from the same patient donor could not be in both sets. This allowed for a balance of number of samples of each tumour type across datasets. To evaluate generalization performance, we fit segmentations on the training set and evaluated them on both sets by computing $I(T; B | C)$ and $I(T; B)$ using mutations from only one dataset and the segmentation boundaries θ that were optimized on the training data. We also computed a segmentation on all of the data to try to get the best result possible.

3. Results

3.1. Mutual Information Comparisons

We define a naive segmentation as a segmentation that groups mutations into contiguous 1 Mb bins, resulting in 2897 segments across the autosomal genome. To compare our algorithm with this baseline, we computed an optimal segmentation with the same number of segments. To make the comparison fairer for the baseline, we removed segments in our naive segmentation that completely lacked mutations – such segments are not informative and thus do not contribute to mutual information – and adjusted the number of segments in our optimal algorithm accordingly. We computed segmentations on different data splits, as outline in section 2.7. The results for these experiments are summarized in table 1. Each segmentation used a minimum segment size of 90 mutation positions (after grouping). ΔI compares the mutual information of Genome Gerrymandering segmentation to its equivalent naive segmentation: a higher value indicates better performance of our algorithm over the naive baseline. We include $I(T; B|C)$ since this is the objective that our algorithm is directly optimizing, even though improving $I(T; B)$ is our real goal. Note that estimates of mutual information are biased to higher values when done on the test set because it contains fewer samples. However, our $\Delta I(T; B)$ estimates are less biased.

Table 1. Summary of Optimal Segmentations:

Trained on	Evaluated on	$I(T; B C)$	$I(T; B)$	$\Delta I(T; B C)$	$\Delta I(T; B)$
Train	Train	0.0454	0.0477	0.0118	0.0118
Train	Test	0.0583	0.0619	0.0128	0.0128
Both	Both	0.0472	0.0497	0.0121	0.0121

3.2. Interpretation of Mutual Information Gain

$I(T; B)$ represents the average reduction in uncertainty of the value of T , the tumour type of a mutation, when the value of B is observed for that same mutation. If there are M tumour types, it takes $\log_2(M)$ bits of information to specify a tumour. If observing B gives us on average $I(T; B)$ bits of information about the value of T , then $\log_2(M)/I(T; B)$ mutations from the same sample is a lower bound on the average number of mutation needed to identify its tumour type.

Theorem 1. *If $B_1, \dots, B_D \sim B$ are identically distributed with $B_i \perp\!\!\!\perp B_j \mid T$, then $I(T; B_1, \dots, B_D) \leq DI(T; B)$*

Proof. Induction on B_d , consider $I(T; B_1, B_2)$.

$I(T; B_1, B_2) = I(T; B_1) + I(T; B_2 \mid B_1)$ by chain rule of information [14]

If $B_2 \rightarrow T \rightarrow B_1$ forms a Markov chain, then $I(T; B_2 \mid B_1) \leq I(T; B_2)$ by the data processing inequality [14]

$$\begin{aligned} p(b_2, t, b_1) &= p(b_2)p(t \mid b_2)p(b_1 \mid t, b_2) && \text{by chain rule of probability} \\ &= p(b_2)p(t \mid b_2)p(b_1 \mid t) && \text{since } B_1 \perp\!\!\!\perp B_2 \mid T \end{aligned}$$

Thus $B_2 \rightarrow T \rightarrow B_1$.

Thus $I(T; B_1, B_2) \leq I(T; B_1) + I(T; B_2) = 2I(T; B)$ since $B_1, B_2 \sim B$ □

Theorem 1 implies that $DI(T; B)$ is an upper bound on the average information we can get about tumour type from D conditionally independent mutations. Thus having a higher $I(T; B_{opt})$ (where B_{opt} uses the optimized segmentation boundaries) reduces the upper bound on the average number of mutations needed by a factor of $\Delta I(T; B)/I(T; B_{opt})$, where we defined $\Delta I(T; B) = I(T; B_{opt}) - I(T; B_{naive})$.

3.3. Segmentation Size Comparisons

There are two ways to measure segment size: the genomic length (in bp) that the segment spans, shown in Fig. 2, and the number of mutation positions that lie within the segment's boundaries, shown in Fig. 3. In general, the optimal segmentation favours smaller segments over the naive segmentation and is left-skewed with a long tail. Note that segments that are small by one measure are not necessarily small by the other (for example, the expansive mutation sparse regions in the genome).

4. Discussion

The activity of mutational signatures varies across the genome, and is influenced by a number of factors including chromatin state [6]. Previous work has demonstrated that the mutation rate at the one Mb-scale is strongly correlated with chromatin features from a tumour's cell-of-origin [7]. More recent work has demonstrated that mutation rate in one Mb-segments can be used to accurately identify cancer-type [11]. These works demonstrate the strength of the relationship between regional mutation density and genome organization, but the choice of one Mb segments lacks both biological motivation, and the resolution to capture local variation

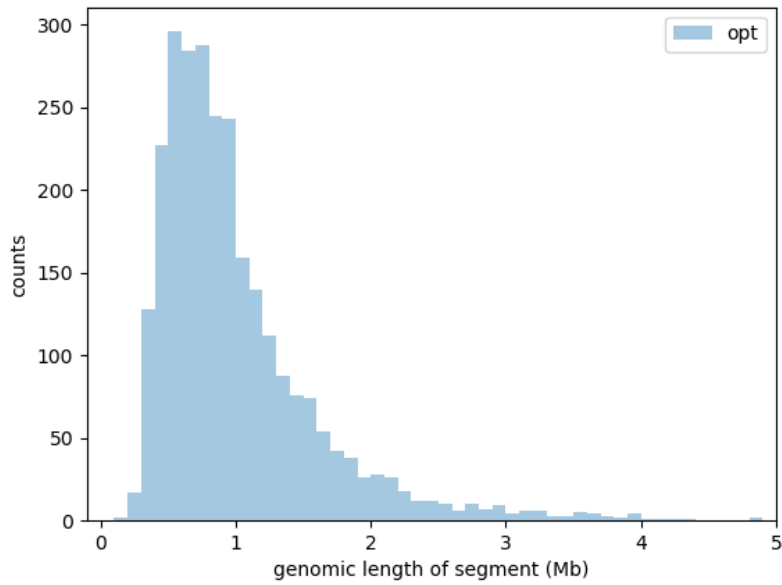


Fig. 2. Distribution of the genomic length (in Mb) of the segments in the optimal segmentation (trained on both datasets). Histogram bin width is 100 Kb. Note that naive segments are not shown here since they all have the same genomic length by definition (1 Mb).

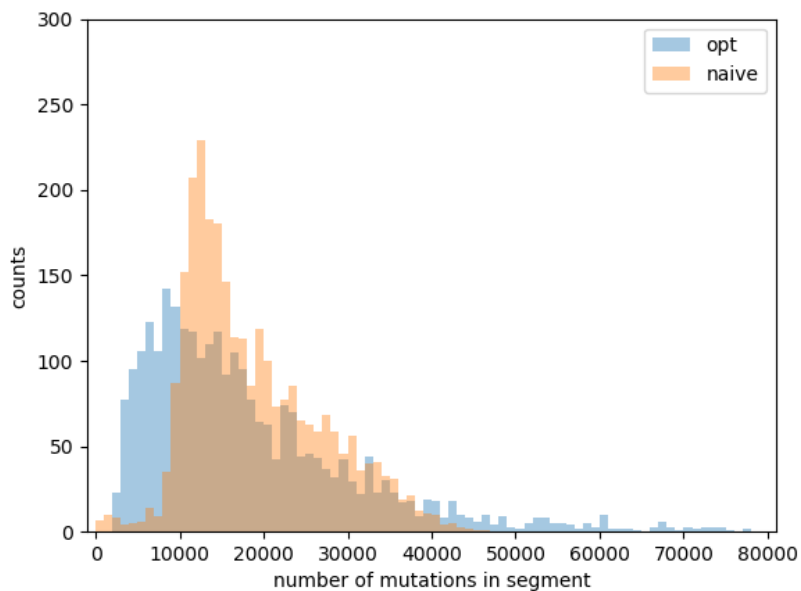


Fig. 3. Distribution of number of mutations in each segment of the optimal and naive segmentations (trained on both datasets). Histogram bin width is 1000 mutations.

in mutation rate. In this work, we present Genome Gerrymandering, an information theoretic algorithm for determining a genomic segmentation that maximizes the mutual information between regional mutation density and cancer type.

Our algorithm increases the mutational information between segmentation and cancer type by 0.0128 bits on the held-out test set. This roughly corresponds to, at a minimum, a $0.0128/0.0619 \approx 20\%$ reduction in the average number of mutations required to discriminate between cancer types. As the relationship between chromatin state and mutation density is the primary feature driving cancer-type identification based on mutations, our result suggests that an optimal genome segmentation may provide a greater association between mutation density and genome organization than previously reported [7]. Interpreting the association between an optimal segmentation and genome segmentation will benefit from an in-depth investigation of functional elements contained within the intervals produced by our algorithm, and investigation into the association between our segmentation and 3-D genome topology.

In this work, we choose K , the number of segments, to be equal to 2897 for direct comparison with naive 1 Mb segmentation of the autosomes. One can optimize the choice of K through a variety of methods. For example, Genome Gerrymandering is, in essence, performing a regional clustering of mutations; so standard metrics for choosing K in clustering algorithms could be used. An alternative approach might be to set K such that the mutual information is maximized on a held-out validation set. Regardless of the method used, because Genome Gerrymandering provides optimal solutions for all values of K up to the maximum K input to the algorithm, selecting the number of clusters would require less compute than other clustering methods because Genome Gerrymandering does not need to be rerun for other potential numbers of clusters less than K .

In summary, this paper presents a general purpose, information theoretic algorithm for finding an optimal genomic segmentation. While this work focused on associating genomic segmentation with cancer type, improving the cancer type identification based on regional mutation density is just one goal of this work. It could also help to identify genomic regions of interest based on large differences in mutation density among cancer types. The algorithm can also be run using only subsets of mutations that are of specific interest. For example, the algorithm may be run using mutations from mutation signatures involved in hypermutation, allowing us to identify regions of the genome most affected by these mutational processes. Ultimately, our algorithm is generic in the sense that it could be applied to regionally cluster mutations based on any discrete phenotypic label (for example, sex).

Some algorithmic improvements are possible and there are some clear directions for further investigation. Currently, the algorithm does not take in explicit information about copy-number or mutational signature activities, both of which are important features influencing mutation rate. The utility of the optimal segmentation for identifying cancer type has not yet been fully explored. Future studies can also investigate the biological significance of the identified genomic segments; in particular their association with genome topology, functional elements, chromatin state, and local mutation signature exposure.

5. Code availability

The source code for our work is available at <https://github.com/adamoyoung/MutSeg>. The dynamic programming algorithm was implemented in C, while the pre-processing and analysis scripts were implemented in Python 3.

6. Acknowledgements

Quaid Morris is supported by an Associate Investigator Award from the Ontario Institute for Cancer Research (OICR), and he is a Canada CIFAR AI chair. The experiments in this project were made possible by a compute allocation from Compute Canada. Many thanks to Wei Jiao at OICR for help with accessing the data. We would also like to acknowledge the support of the NVIDIA Compute the Cure gift to Quaid Morris.

References

- [1] M. A. Lodato *et al.*, “Aging and neurodegeneration are associated with increased mutations in single human neurons,” *Science (New York, N.Y.)*, vol. 359, no. 6375, pp. 555–559, 2018, ISSN: 1095-9203. DOI: [10.1126/science.aao4426](https://doi.org/10.1126/science.aao4426). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/29217584><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5831169>.
- [2] I. Martincorena *et al.*, “Universal Patterns of Selection in Cancer and Somatic Tissues.,” *Cell*, vol. 171, no. 5, 1029–1041.e21, 2017, ISSN: 1097-4172. DOI: [10.1016/j.cell.2017.09.042](https://doi.org/10.1016/j.cell.2017.09.042). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/29056346>.
- [3] L. B. Alexandrov *et al.*, “The Repertoire of Mutational Signatures in Human Cancer,” *bioRxiv*, p. 322 859, 2019. DOI: [10.1101/322859](https://doi.org/10.1101/322859). [Online]. Available: <https://www.biorxiv.org/content/10.1101/322859v2>.
- [4] Y. Rubanova *et al.*, “TrackSig: reconstructing evolutionary trajectories of mutation signature exposure,” *bioRxiv*, p. 260 471, 2018. DOI: [10.1101/260471](https://doi.org/10.1101/260471). [Online]. Available: <https://www.biorxiv.org/content/early/2018/02/05/260471>.
- [5] D. Wojtowicz, I. Sason, X. Huang, Y.-A. Kim, M. D. Leiserson, T. M. Przytycka, and R. Sharan, “Hidden markov models lead to higher resolution maps of mutation signature activity in cancer,” *Genome medicine*, vol. 11, no. 1, p. 49, 2019.
- [6] A. Gonzalez-Perez *et al.*, “Leading Edge Review Local Determinants of the Mutational Landscape of the Human Genome,” 2019. DOI: [10.1016/j.cell.2019.02.051](https://doi.org/10.1016/j.cell.2019.02.051). [Online]. Available: <https://doi.org/10.1016/j.cell.2019.02.051>.
- [7] P. Polak *et al.*, “Cell-of-origin chromatin organization shapes the mutational landscape of cancer,” *Nature*, vol. 518, no. 7539, pp. 360–364, 2015, ISSN: 0028-0836. DOI: [10.1038/nature14221](https://doi.org/10.1038/nature14221). [Online]. Available: <http://www.nature.com/doifinder/10.1038/nature14221>.
- [8] P. G. Yazdi and others., “Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact,” *PLOS ONE*, vol. 10, no. 8, A. Imhof, Ed., e0136574, 2015, ISSN: 1932-6203. DOI: [10.1371/journal.pone.0136574](https://doi.org/10.1371/journal.pone.0136574). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/26308346><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4550472><https://dx.plos.org/10.1371/journal.pone.0136574>.
- [9] R. Sabarinathan and others., “Nucleotide excision repair is impaired by binding of transcription factors to DNA,” *Nature*, vol. 532, no. 7598, pp. 264–267, 2016, ISSN: 0028-0836. DOI: [10.1038/nature17661](https://doi.org/10.1038/nature17661). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/27075101><http://www.nature.com/articles/nature17661>.
- [10] J. K. Wiencke, “DNA adduct burden and tobacco carcinogenesis,” *Oncogene*, vol. 21, no. 48, pp. 7376–7391, 2002, ISSN: 0950-9232. DOI: [10.1038/sj.onc.1205799](https://doi.org/10.1038/sj.onc.1205799). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/12379880><http://www.nature.com/articles/1205799>.
- [11] W. Jiao *et al.*, “A deep learning system can accurately classify primary and metastatic cancers based on patterns of passenger mutations,” *bioRxiv*, p. 214 494, 2019. DOI: [10.1101/214494](https://doi.org/10.1101/214494). [Online]. Available: <http://biorxiv.org/content/early/2019/01/22/214494.abstract>.
- [12] R. Bellman, “On the approximation of curves by line segments using dynamic programming,” *Communications of the ACM*, p. 284, 1961.
- [13] P. J. Campbell *et al.*, “Pan-cancer analysis of whole genomes,” *bioRxiv*, p. 162 784, 2017. DOI: [10.1101/162784](https://doi.org/10.1101/162784). [Online]. Available: <https://www.biorxiv.org/content/10.1101/162784v1>.
- [14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, ISBN: 0471241954.