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

Adamo Young

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 correlates strongly with chromatin features, and 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. This 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.

1. Introduction and Related Work

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 mutational processes that generate single nucleotide variants (SNVs) by grouping mutations based on short-range sequence characteristics such as trinucleotide context [3]. They also 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 change across the genome [5]. Process-specific, regional mutation rates are influenced by phenomena 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, [11]

leveraged this association to develop a deep-learning classifier that only uses regional mutation density to differentiate between 24 cancer types with an accuracy of 91%. 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, including 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 location and cell-type, examining regional mutation density may provide information about the distribution of functional elements across the genome.

Genome segmentation is a well-studied problem, and there are many existing algorithms for segmenting the genome for different purposes. Functional element annotation is a type of genome segmentation that produces labelled, contiguous, non-overlapping segments that are associated with specific functional elements. Many functional annotation algorithms integrate multiple types of genomics data, including histone methylation, DNA accessibility, and TF binding [12] to make segment predictions. One of the early functional segmentation algorithms, ChromHMM [13], models segment labels as the hidden states of a Hidden Markov Model (HMM). ChromHMM combines genomic signals at a resolution of 200 bp (the size of a nucleosome) and binarizes them to produce a sequence of observed binary variables. The transition probabilities of the model are trained with the Baum-Welch algorithm, and the maximum likelihood sequence of hidden states across the genome is extracted with the Viterbi algorithm. The segment boundaries are defined as the locations that correspond to a change in hidden state. Segway [14], [15] is a similar algorithm that uses a Dynamic Bayesian Network (DBN) instead of an HMM. With a DBN, it is possible to explicitly model missing data (instead of rely on smoothing and interpolation like ChromHMM) and better integrate prior beliefs on segment size. Segway models the observed data (ChIP-seq, DNase-seq, and FAIRE-seq from ENCODE [12]) as a mixture of gaussians [15], avoiding the binarization that is used in ChromHMM. Model parameters are fit using Expectation-Maximization on sampled minibatches of the genomic data, and the entire genome is segmented in a similar fashion to ChromHMM, using the Viterbi algorithm. Functional element annotation has proven to be useful for understanding the genome, and has successfully been used to predict functions that were not previously known [14].

Copy number segmentation is another important problem in genomics. Copy number variants (CNVs) are large sections of the genome (>50 bp) that are differentially repeated across members of a population. Copy number measurements can be made with microarray data, including array-based comparative genomic hybridization (aCGH) and SNV arrays, or with whole-genome or whole-exome sequencing data. Circular binary segmentation (CBS [16]) is a simple and efficient change point algorithm for performing copy-number segmentation. The

algorithm assumes that the copy number observations are independently normally distributed, with shared variance (but potentially different means). It circularizes the data by splicing the ends together (i.e. by placing the last measurement next to the first) and then applies a likelihood ratio test at all pairs of points in the circle. The null hypothesis is that the mean of the observations along the shortest path between the points is equal to the mean of observations along the longest path. It greedily selects the pair of points that most strongly reject this null hypothesis, and then recursively applies the test to each resulting segment until no null hypotheses are rejected at any position. The segment labels correspond to the mean of the observations in that segment, which is interpreted as the copy number. SLMsuite [17] presents an alternate approach to copy number segmentation that relies on a shifting level model (SLM). This approach models the observed data as the sum of two latent stochastic processes: a noise process and a mean (copy-number) process. Like with CBS, shifts in the mean are interpreted as segment boundaries. SLMs can be written as a specific kind of HMM [17], which allows them to be fit with a variation of the Baum-Welch algorithm. When using sequencing data, segmentations produced by SLM and CBS can have resolutions as high as 100bp and as low as 1000bp, depending on sequencing depth [17].

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 cancer type. Our algorithm differs from existing segmentation algorithms in its information maximization approach and in the type of data that it uses (cancer SNVs). We show that our optimal segmentation of the genome significantly increases information content between regional mutation density and cancer type over a naive segmentation with an equal number of segments.

2. Methods

2.1. *Overview*

The goal of the algorithm is to split the genome into similarly dense sections that differ in their mutation rates among different cancer types. We do this by labelling somatic mutations (SNVs) 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 distribution of mutations per cancer type differs by as much as possible. Formally, we use a modification of the Bellman K-segmentation algorithm [18] 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 [19]. 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. Con-

sensus 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. *Problem Formulation*

Let K be the desired number of segments in the segmentation (specified by the user), N be the number of distinct genomic positions of mutations in the data, and N_T be the number of tumour types. Let tumour type T, segment B_θ, and chromosome C be categorical random variables. T has support over the tumour types $\{t_1, ..., t_{N_T}\}$, B_θ has support over the segments $b_1, ..., b_K$, and C has support over the chromosomes $\{c_1, ..., c_{22}\}$. The segmentation boundaries θ define a categorical distribution $B_\theta \sim p_\theta(b)$ that represents the probability that a mutation is inside a segment b. For a given choice of K, our algorithm seeks to maximize $I(T; B_\theta)$ by changing the segmentation boundaries θ.

2.4. *Objective Function*

The optimization goal is to find $\arg\max_\theta I(T; B_\theta)$. $I(T; B_\theta)$ and $I(T; B_\theta \mid C)$ can be related in the following equation:

$$I(T; B_\theta) = I(T; B_\theta \mid C) + H(C) - H(C \mid T) - H(C \mid B_\theta) - H(C \mid T, B_\theta) \tag{1}$$

We assume that segments must be contiguous and cannot span multiple chromosomes. This means that C is a function of B_θ, since $\forall b_k : \exists c_i : p(c_i \mid b_k) = 1$. As a result, $H(C \mid B_\theta) = 0$ and $H(C \mid T, B_\theta) = 0$. Thus:

$$\begin{aligned} I(T; B_\theta) &= I(T; B_\theta \mid C) + H(C) - H(C \mid T) \\ &= I(T; B_\theta \mid C) + I(T; C) \\ &\geq I(T; B_\theta \mid C) \end{aligned} \tag{2}$$

The last inequality holds since mutual information is always non-negative. Thus we have shown $I(T; B_\theta \mid C)$ is a lower bound for $I(T; B_\theta)$. Since $I(T; C)$ does not depend on θ, the following equality holds:

$$\arg\max_\theta I(T; B_\theta) = \arg\max_\theta I(T; B_\theta \mid C) \tag{3}$$

The advantage of maximizing $I(T; B_\theta \mid C)$ is that it admits a decomposition of the optimization problem into subproblems for each chromosome. Recalling the definition of conditional mutual information:

$$I(T; B_\theta \mid C) = H(T \mid C) + H(B_\theta \mid C) - H(T, B_\theta \mid C) \tag{4}$$

The value of $H(T \mid C)$ does not change with respect to θ. It is therefore clear that:

$$\begin{aligned}\arg\max_{\theta} I(T; B_\theta \mid C) &= \arg\max_{\theta} [H(B_\theta \mid C) - H(T, B_\theta \mid C)] \\ &= \arg\max_{\theta} \sum_c p(c) \left[H(B_\theta \mid C = c) - H(T, B_\theta \mid C = c)\right] \\ &= \arg\max_{\theta} \sum_c p(c) \left(-H(T \mid B_\theta, C = c)\right)\end{aligned} \tag{5}$$

Intuitively, maximizing this objective works by minimizing joint uncertainty about tumour type and segment, $H(T, B_\theta \mid C)$, while encouraging equal segment size by maximizing uncertainty about segment, $H(B_\theta \mid C)$. Since $H(T \mid C)$ remains unchanged throughout optimization, it can be computed afterwards to find $I(T; B_\theta \mid C)$.

We can greatly simplify the problem by fixing the total number of segments in each chromosome, and optimizing θ subject to this constraint. We heuristically assume that each chromosome c has a number of segments K_c that is proportional to that chromosome's size (in bp) relative to the rest of the genome (not including sex chromosomes). This assumption introduces bias, but allows for the optimization subproblems to be solved independently (and in parallel) for each chromosome.

Let θ_c parameterize the boundaries of the segments on chromosomes c, with $\theta = (\theta_1, ..., \theta_{22})$. Then $-H(T \mid B_{\theta_c}, C = c)$ can be written as follows:

$$\begin{aligned}-H(T \mid B_{\theta_c}, C = c) &= -\sum_b p(b \mid c) H(T \mid B_{\theta_c} = b, C = c) \\ &= \sum_b p(b \mid c) \sum_t p(t \mid b, c) \log p(t \mid b, c)\end{aligned} \tag{6}$$

To compute $I(T; B_{\theta^*})$ once the optimal boundaries θ^* have been found, we can simply ignore the random variable C (which was introduced to allow for parallel optimization by chromosome) and use the following equation:

$$I(T; B_{\theta^*}) = H(T) - H(T|B_{\theta^*}) \tag{7}$$

The class probabilities of the categorical distributions (C, T, and B_θ) are estimated from the mutation count data. Details on estimators can be found in Section 2.6.

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. Additionally, 32 samples that were identified as hypermutators were removed entirely from the dataset, as well as all mutations on the sex chromosomes (see code, Section 5, for details).

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. *Estimators*

After preprocessing, the mutation data is in the form of a grouped count matrix A, which is N by N_T. We split this matrix into matrices $A_1, ..., A_{22}$ (one for each chromosome), where each A_c is N_c by N_T. Let $A_c[n_1, t_1]$ be the count of mutations at position n_1 of tumour type t_1, 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 along each axis, $n_1 \leq n_2$ and $t_1 \leq t_2$.

For each chromosome c, the corresponding count matrix A_c can be used to estimate probabilities for the categorical conditional distributions $\hat{p}(t, b|c)$, which are then used to compute the conditional entropy in Equation 5.

Let b be a segment with starting point n_1 and end point n_2, $n_1 \leq n_2$. Maximum likelihood (ML) estimates for categorical distributions are simply the count fractions, as follows:

$$\hat{p}_{ML}(c) = \frac{A_c[:, :]}{A[:, :]} \tag{8}$$

$$\hat{p}_{ML}(t, b \mid c) = \frac{A_c[n_1 : n_2, t]}{A_c[:, :]} \tag{9}$$

The conditional marginals $\hat{p}_{ML}(b \mid c)$ and $\hat{p}_{ML}(t \mid c)$ can be computed from $\hat{p}_{ML}(t, b \mid c)$ by marginalizing. The corresponding plug-in ML entropy estimator $\hat{H}_{ML}$ is computed using the ML estimates for the class probabilities.

$$\hat{H}_{ML}(B_\theta \mid C = c) = -\sum_b \hat{p}_{ML}(b \mid c) \log \hat{p}_{ML}(b \mid c) \tag{10}$$

$$\hat{H}_{ML}(T, B_\theta \mid C = c) = -\sum_{t,b} \hat{p}_{ML}(t, b \mid c) \log \hat{p}_{ML}(t, b \mid c) \tag{11}$$

It has been shown that plug-in ML estimator for entropy is optimal in the high-sample regime $n \gg p$, where n is the number of samples and p is the size of the support of the discrete random variable. However, in the case where $n \ll p$, this estimator is known to be biased, with a tendency to severely underestimate entropy [20]. For our distributions, $n \gg p$ (often two orders of magnitude), thus the ML estimator seemed like the better choice.

To estimate $I(B; T)$ with the learned segmentation boundaries, we needed to estimate the associated unconditioned entropies (see Equation 7). We compared using the ML estimator $\hat{p}_{ML}(t, b)$ and an alternate estimator called the James-Stein (JS) shrinkage estimator [20] which is known to have lower variance in the low sample regime. It is a plug-in estimator that works as follows:

$$t = \frac{1}{p} = \frac{1}{K_c N_T} \tag{12}$$

$$\lambda = \frac{1 - \sum_{t,b} \hat{p}_{ML}(t, b)^2}{(n - 1) \sum_{t,b} (t - \hat{p}_{ML}(t, b))^2} \tag{13}$$

$$\hat{p}_{JS}(t, b) = \lambda t + (1 - \lambda) \hat{p}_{ML}(t, b) \tag{14}$$

$$\hat{H}_{JS}(T, B_\theta) = -\sum_{t,b} \hat{p}_{JS}(t, b) \log \hat{p}_{JS}(t, b) \tag{15}$$

Note that we only used the JS estimator for $\hat{I}_{JS}(T; B_\theta)$, to evaluate the optimal segmentation boundaries θ^* and compare them with the naive boundaries θ^{**}. It was never used during op-

timization. In practice the JS entropy estimations were nearly identical to the ML estimations (see Section 3.1).

2.7. *Optimization with Dynamic Programming*

It is possible to efficiently maximize the objective function for each chromosome (Equation 6) with dynamic programming. Let $\phi_c(j,k)$ be a function that outputs a segmentation $\{b_1,...,b_k\}$ defined entirely over mutation positions $1,...,j$ on chromosome c, $j \leq N_c$. Let s_c, S_c and S'_c be functions defined as follows:

$$s_c(b) = p(b \mid c) \sum_t p(t \mid b,c) \log p(t \mid b,c) \tag{16}$$

$$S_c(j,k) = \max_{\phi_c} \sum_{b \in \phi_c(j,k)} s_c(b) \tag{17}$$

$$S'_c(j,k) = \arg\max_{\phi_c} \sum_{b \in \phi_c(j,k)} s_c(b) \tag{18}$$

We call $s_c(b)$ the score of a segment b. Intuitively, $S_c(j,k)$ is the score of a best possible segmentation with k (non-empty) segments over the first j mutations, and $S'(j,k)$ are the boundaries that define that segmentation. Note that maximizing Equation 6 with respect to θ_c is equivalent to finding $S_c(N_c,K_c)$ and $S'_c(N_c,K_c)$.

For notational consistency, let $f_c(a,b)$ be the function that outputs a segment consisting of mutations from a to b, inclusive. The base case, where $k=1$, corresponds to the trivial segmentation (where all mutations are put in one segment $f_c(1,j)$).

$$S_c(j,1) = p(f_c(1,j) \mid c) \sum_t p(t \mid f_c(1,j),c) \log p(t \mid f_c(1,j),c) \tag{19}$$

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

$$S_c(j,k) = \max_{i=k-1,...,j-1} S_c(i,k-1) + s_c(f_c(i+1,j)) \tag{20}$$

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, Equation 20 states that $S_c(j,k)$ 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 contiguous 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 parallelized across chromosomes. Once S'_c is known for each chromosome, $I(T;B \mid C=c)$ can be computed using Equation 4, while $I(T;B)$ can be estimated directly from the count matrix using Equation 7 as described in Section 2.6. Note that this algorithm is a variation of the Bellman K-segmentation algorithm [18] that has been modified to maximize a mutual information objective.

Algorithm 1: Genome Gerrymandering Algorithm

Input : A_c, an N_c by N_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

$S_c \leftarrow \texttt{InitNegInf}(N_c, K_c)$
$S'_c \leftarrow \texttt{InitNegInf}(N_c, K_c)$
for $i = P$ **to** N_c **do**
    $S_c[i, 1] \leftarrow \texttt{ComputeScore}(A_c, 1, i)$
for $k = 2$ **to** K_c **do**
    for $j = k$ **to** N_c **do**
        for $i = k - 1$ **to** $j - 1$ **do**
            if $S_c[j, k] < S_c[i, k-1] + \texttt{ComputeScore}(A_c, i+1, j)$ **then**
                if $j - i - 1 \geq P$ & $i - 1 \geq P$ **then**
                    $S_c[j, k] \leftarrow S_c[i, k-1] + S_c[j, 1] - S_c[i, 1]$
                    $S'_c[j, k] \leftarrow i$
return S_c, S'_c

2.8. *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 $I(T; B \mid C)$ and $I(T; B)$ using mutations from only one dataset and the segmentation boundaries 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 outlined in Section 2.8. 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.

Table 1: Summary of Segmentations

Type	Trained on	Evaluated on	$\hat{I}_{ML}(T;B \mid C)$	$\hat{I}_{ML}(T;B)$	$\hat{I}_{JS}(T;B)$
Naive	N/A	Train	0.0237	0.0253	0.0253
Naive	N/A	Test	0.0322	0.0349	0.0349
Naive	N/A	Both	0.0245	0.0262	0.0262
Optimal	Train	Train	0.0318	0.0335	0.0335
Optimal	Train	Test	0.0408	0.0436	0.0435
Optimal	Both	Both	0.0327	0.0345	0.0345

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 N_T tumour types, it takes $\log_2(N_T)$ 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 $\log_2(N_T)/I(T;B)$ mutations from the same sample is a lower bound on the average number of mutation needed to identify its tumour type.

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

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 [21]
If $B_2 \to T \to B_1$ forms a Markov chain, then $I(T;B_2 \mid B_1) \leq I(T;B_2)$ by the data processing inequality [21]

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

*In a slight abuse of notation, we drop the θ from B_θ for clarity

Thus $B_2 \to T \to B_1$.
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 Visualizations*

The segmentations for each chromosome are shown in Figures 4, 5, 6, and 7. The precise boundaries of the segments can be found on the github repo (see Section 5). When comparing the optimal and naive segmentations side-by-side, it is interesting to note the differences. Perhaps unsurprisingly, regions that have a low mutation count, such as $120-140$ Mb on chromosome 1 (Figure 4a) and $45-70$ Mb on chromosome 9 (Figure 5c), have been grouped together in the optimal segmentation. These large, sparse segments appear in every chromosome and seem to roughly correspond to boundaries of the centromeres. There are also segments that are mutation-dense but have low conditional tumour entropy, such as the segment around 80 Mb on chromosome 2 (Figure 4b) or the segment around 35 Mb on chromosome 22 (Figure 7d). Such segments may be of particular diagnostic interest.

The differences between segmentations also become apparent when visualizing the distributions of segment size (Figure 1), mutation count (Figure 2), and conditional entropy of tumour type (Figure 3). Note that by definition, the naive segmentation has segments of equal genomic size (1 Mb), with the exception of segments that appear at the end of each chromosome (which may be less than 1 Mb). The optimal segmentation favours smaller segments, and the segment size distribution (Figure 1) is right-skewed with a heavier tail. The largest optimal segment can be found on chromosome 22 (Figure 7d), and is around 18 Mb in size. When comparing the distributions of conditional tumour entropies, it appears that the optimal segments are somewhat bimodal, compared to the unimodal distribution of the naive segments. Because there are 11 tumour types, we know that $H(T|B = b) \leq \log_2(11) \approx 3.45$ bits, which is achieved when $p(t|b)$ is a uniform distribution. In practice, for both segmentations, the maximum conditional tumour entropy was < 3 bits, although the optimal segmentation's maximum (2.98 bits) was greater than the naive segmentation's (2.86 bits).

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 1 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 1 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 1 Mb segments lacks both biological motivation, and the resolution to capture local variation

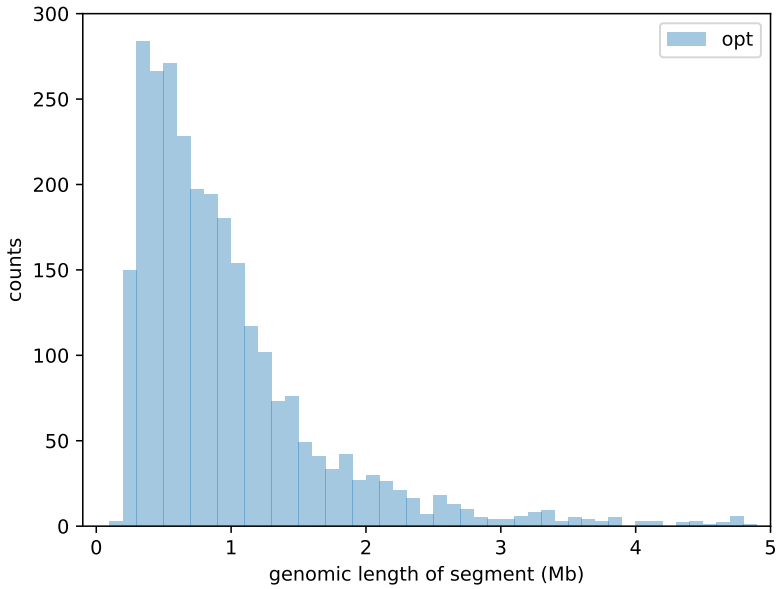


Fig. 1: Distribution of the genomic size (in Mb) of the segments in the optimal segmentation (trained on both datasets). Histogram bin width is 100 Kb. Maximum segment size is 18 Mb.

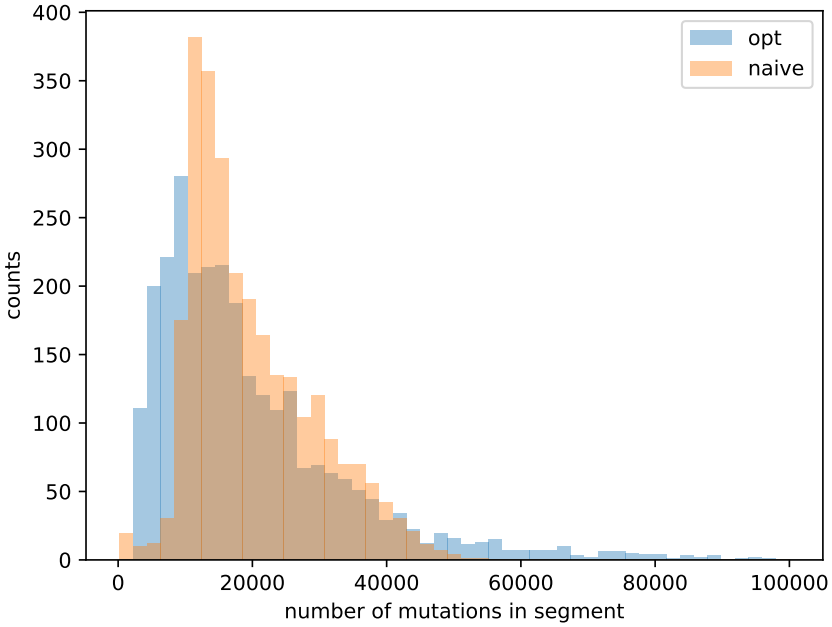


Fig. 2: Distribution of mutation counts per segment of the optimal and naive segmentations (trained on both datasets).

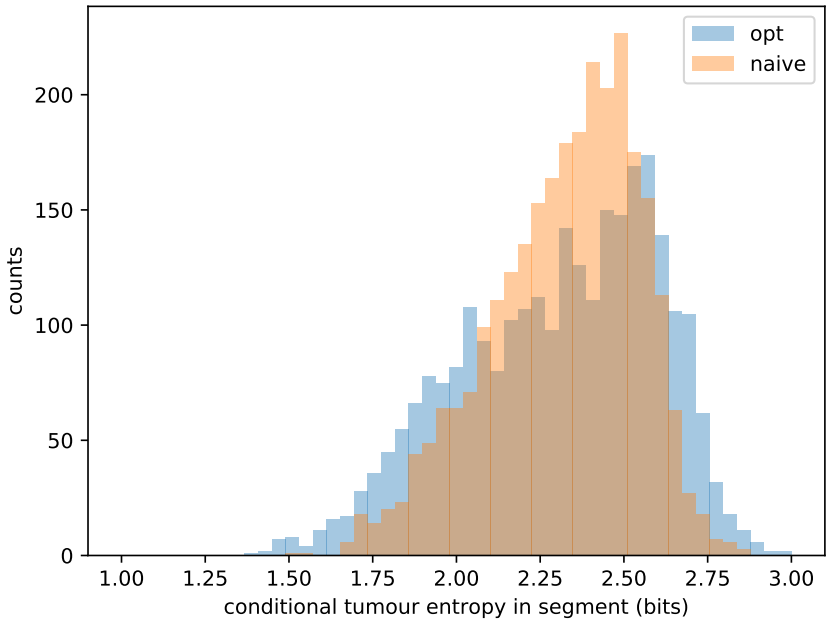


Fig. 3: Distribution of conditional entropy of tumour type $H(T|B = b)$ per segment for the optimal and naive segmentations (trained on both datasets).

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 0.087 bits on the held-out test set. This roughly corresponds to, at a minimum, a $0.0087/0.0436 \approx 20\%$ 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 mutation-based cancer-type identification, our result suggests that an optimal genome segmentation may provide a greater association between regional 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. Genome Gerrymandering is, in essence, performing a regional clustering of mutations; so standard heuristics 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 validation set. We note that the algorithm provides optimal solutions for all values of k less than the input K, which makes selecting the number of segments easier in practice.

In summary, this paper presents a general purpose, information theoretic algorithm for

finding an optimal genomic segmentation. Our algorithm could help identify genomic regions of interest based on large differences in mutation density among cancer types. Furthermore, the algorithm could be run using subsets of mutations that are of specific interest. For example, using only mutations from mutation signatures that are involved in hypermutation may allow 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).

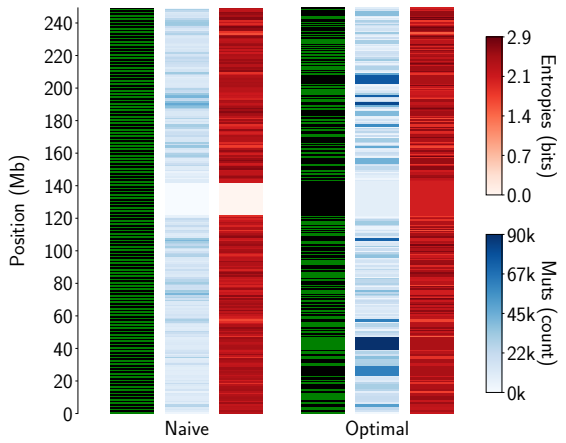
Some algorithmic improvements are possible and there are some clear directions for further investigation. Currently, the algorithm does not take in explicit information about copy-number or mutational signature activities, both of which are important features influencing mutation rate. The utility of 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

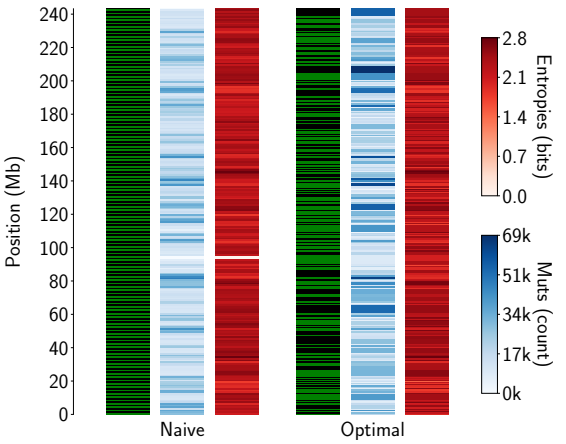
The source code for our work is available at https://github.com/adamoyoung/MutSeg. Any questions or concerns can be addressed to adamo.young@mail.utoronto.ca.

6. Acknowledgements

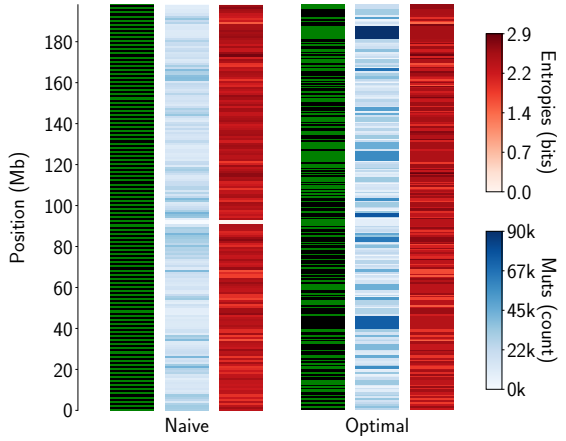
I would like to thank my supervisor Quaid Morris for guidance throughout the project, as well as the rest of the Morris lab for feedback on this work. Gurnit Atwal specifically helped with data preparation and biological motivation. I would also like to thank my co-supervisor Hannes Röst for his support throughout my Master's degree. The experiments were made possible by a compute allocation from Compute Canada. Many thanks to Wei Jiao at OICR for help with accessing the data. Finally, I would like to thank the Department of Computer Science and the Vector Institute for financially supporting me for the duration of the program.



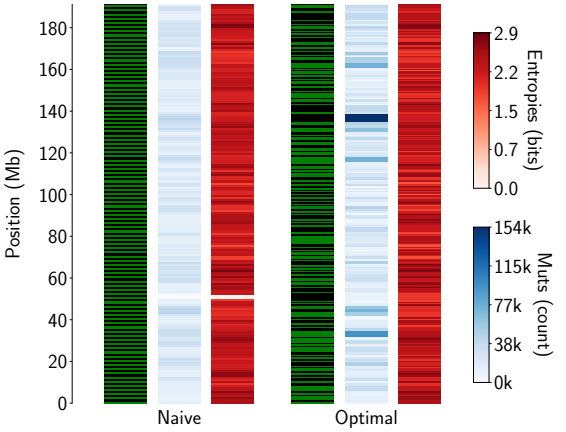
(a) Chromosome 1



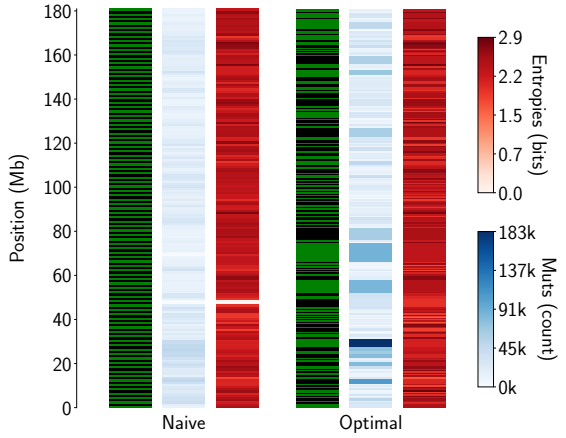
(b) Chromosome 2



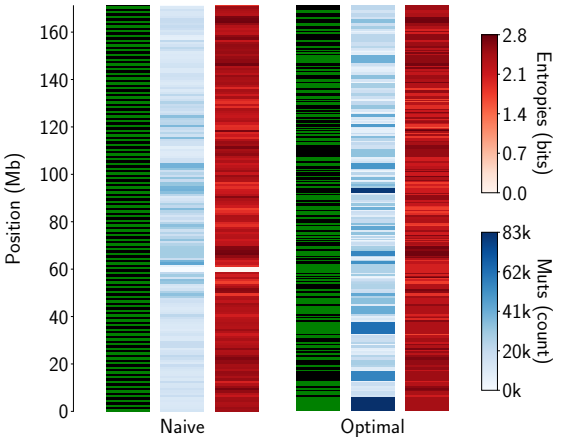
(c) Chromosome 3



(d) Chromosome 4

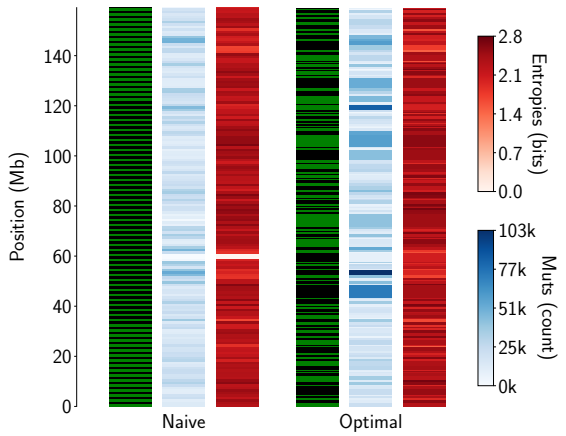


(e) Chromosome 5

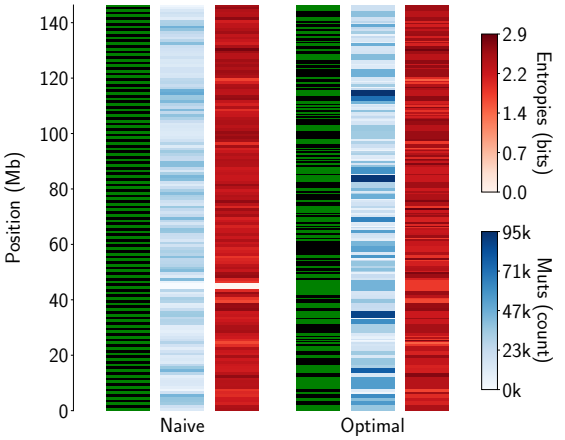


(f) Chromosome 6

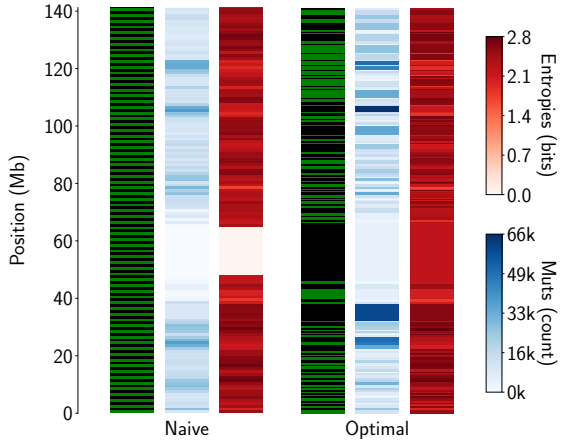
Fig. 4: Chromosome segmentations, 1-6. Entopies shown are $H(T|B = b, C = c)$.



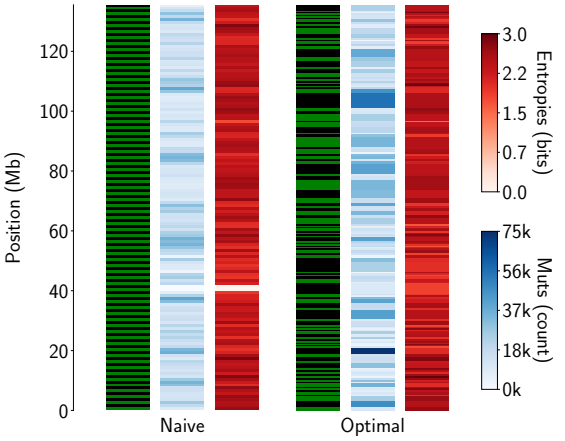
(a) Chromosome 7



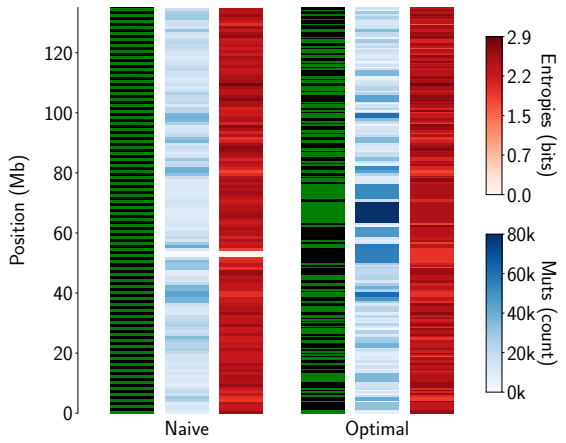
(b) Chromosome 8



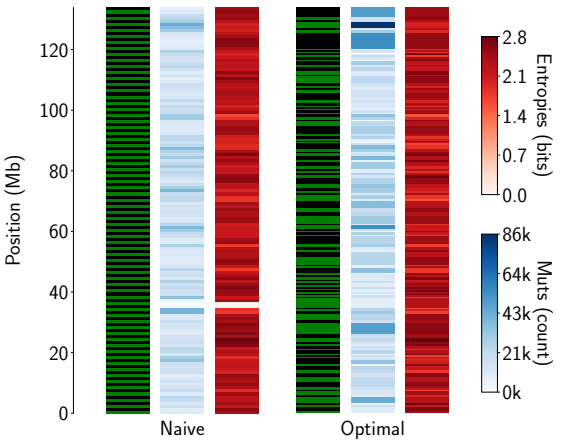
(c) Chromosome 9



(d) Chromosome 10

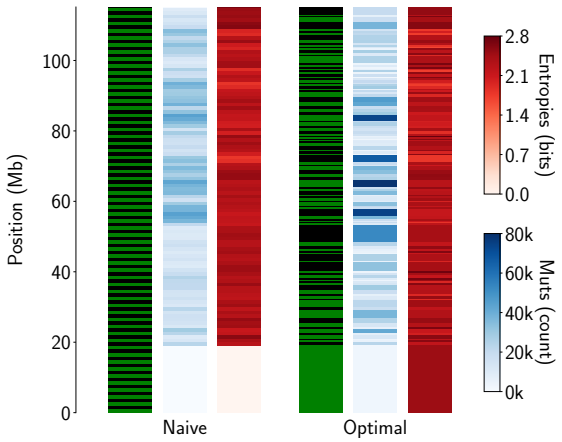


(e) Chromosome 11

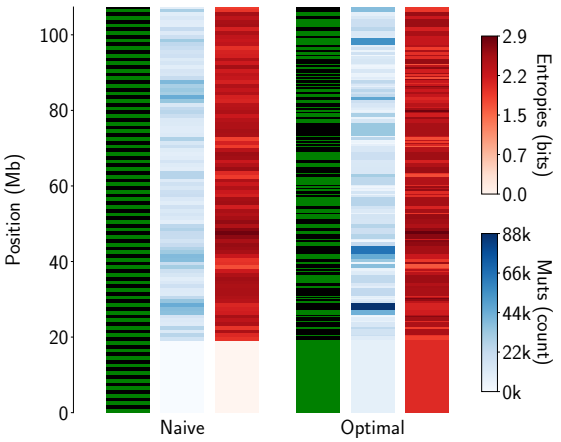


(f) Chromosome 12

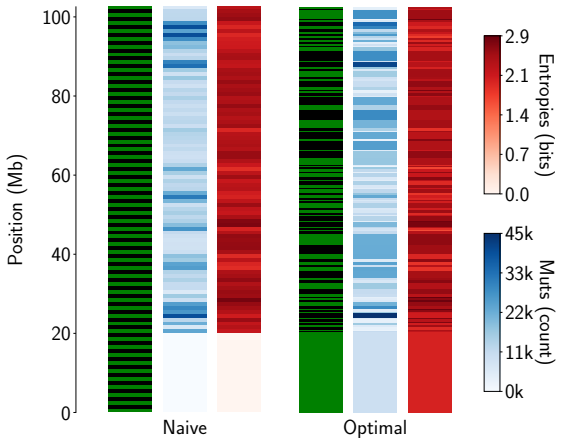
Fig. 5: Chromosome segmentations, 7-12. Entopies shown are $H(T|B = b, C = c)$.



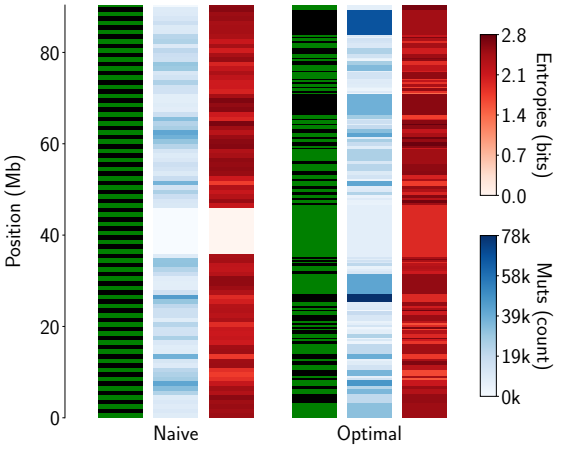
(a) Chromosome 13



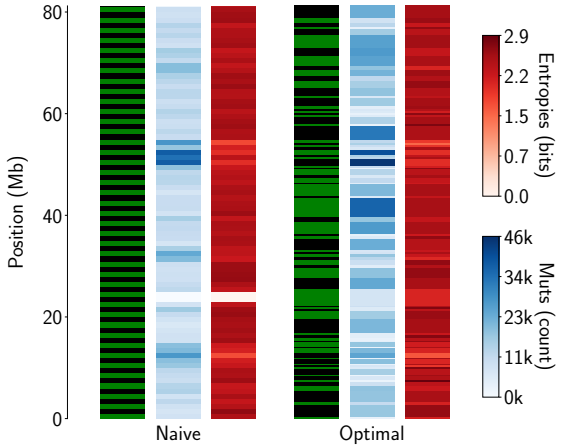
(b) Chromosome 14



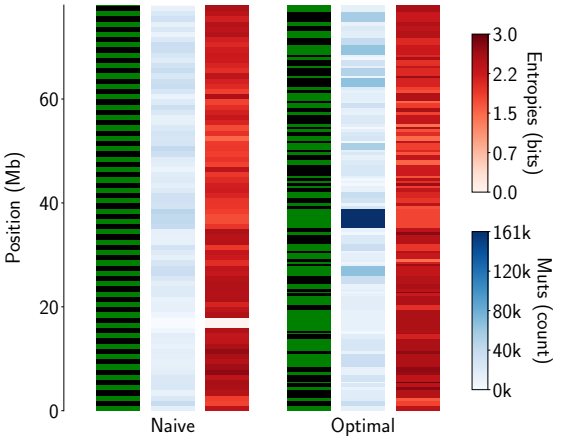
(c) Chromosome 15



(d) Chromosome 16

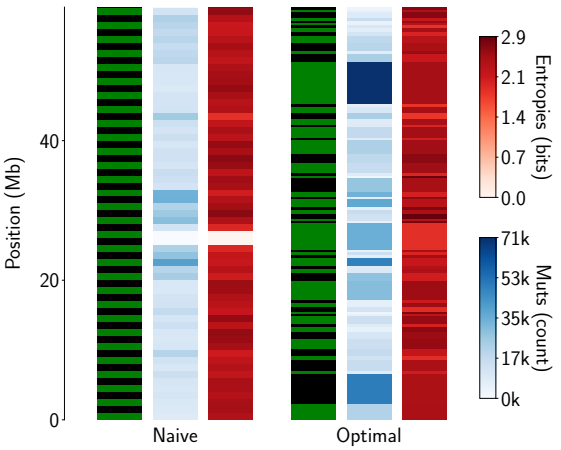


(e) Chromosome 17

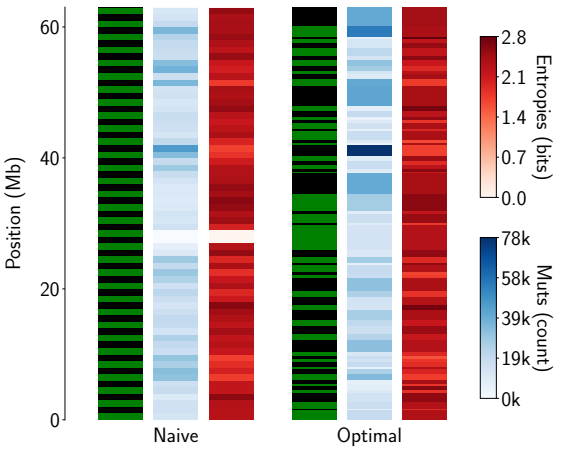


(f) Chromosome 18

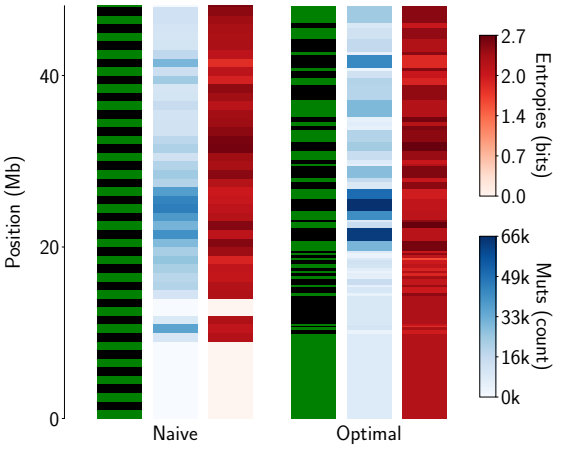
Fig. 6: Chromosome segmentations, 13-18. Entopies shown are $H(T|B = b, C = c)$.



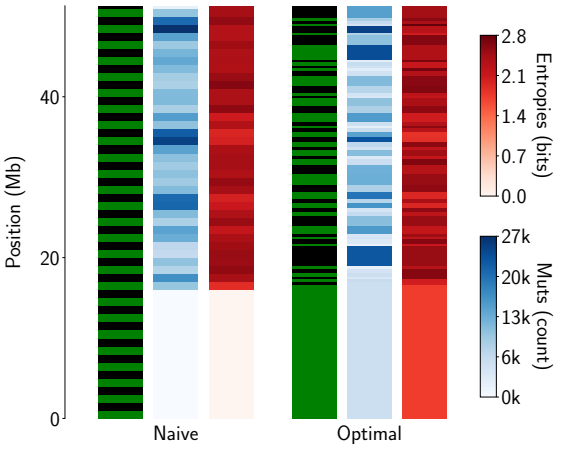
(a) Chromosome 19



(b) Chromosome 20



(c) Chromosome 21



(d) Chromosome 22

Fig. 7: Chromosome segmentations, 19-22. Entopies shown are $H(T|B = b, C = c)$.

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. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29217584http://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. [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. [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. [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. [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. [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. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26308346http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4550472https://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. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27075101http://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. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12379880http://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. [Online]. Available: http://biorxiv.org/content/early/2019/01/22/214494.abstract.

[12] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E.

Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guig? R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. R?der, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guig? T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guig? J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M.

Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, and E. Birney, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[13] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization," *Nat. Methods*, vol. 9, no. 3, pp. 215–216, 2012.

[14] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," *Nat. Methods*, vol. 9, no. 5, pp. 473–476, 2012.

[15] R. C. W. Chan, M. W. Libbrecht, E. G. Roberts, J. A. Bilmes, W. S. Noble, and M. M. Hoffman, "Segway 2.0: Gaussian mixture models and minibatch training," *Bioinformatics*, vol. 34, no. 4, pp. 669–671, Feb. 2018.

[16] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.

[17] V. Orlandini, A. Provenzano, S. Giglio, and A. Magi, "SLMSuite: a suite of algorithms for segmenting genomic profiles," *BMC Bioinformatics*, vol. 18, no. 1, p. 321, 2017.

[18] R. Bellman, "On the approximation of curves by line segments using dynamic programming," *Communications of the ACM*, p. 284, 1961.

[19] P. J. Campbell *et al.*, "Pan-cancer analysis of whole genomes," *bioRxiv*, p. 162 784, 2017. DOI: 10.1101/162784. [Online]. Available: https://www.biorxiv.org/content/10.1101/162784v1.

[20] "Entropy inference and the james-stein estimator, with application to nonlinear gene association networks," *Journal of Machine Learning Research*, vol. 10, pp. 1469–1484, 2009.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, ISBN: 0471241954.