Identifying Simultaneous Rearrangements in Cancer Genomes

Layla Oesper*[1], Simone Dantas†[2], and Benjamin J. Raphael‡[3]

[1]Department of Computer Science, Carleton College, USA.
[2]Institute of Mathematics and Statistics, Fluminense Federal University, Brazil.
[3]Department of Computer Science, Princeton University, USA.

Contents

A Full Proofs Omitted in the Main Text **2**

B Simulation Details **3**
B.1 Chromothripsis Genomes 4
B.2 Step-wise Mutations 4
B.3 Step-wise Mutations along with Branching Evolutions 4

C Additional Results **5**
C.1 Statistical Analysis of Malhotra *et al.* Data 5
C.2 Additional Figures 5

*loesper@carleton.edu
†sdantas@im.uff.br
‡braphael@princeton.edu

A Full Proofs Omitted in the Main Text

(Main Text) Theorem 2. *The fraction of chomothripsis strings of length m derived from a reference genome G composed of n intervals with $\pi(C)$ that is H/T alternating is $\frac{1}{2(m-1)}$.*

In order to prove this theorem we will first need to build up some more notation and observations.

First, we calculate $|G(m,n)|$, the total number of chromothripsis strings of m blocks given a reference genome of n blocks. This is straightforward using the following equation.

$$|G(m,n)| = \prod_{i=0}^{m-1}(2n-2i) = \prod_{i=0}^{m-1} 2(n-i) = 2^m \prod_{i=0}^{m-1}(n-i). \tag{1}$$

Next we calculate $|A(m,n)|$, the number of chromothripsis strings of m blocks given a reference genome of n blocks that are H/T alternating. This computation utilizes the fact that the two cases from Theorem 2 are mutually exclusive and thus the number of instances for each one can be counted separately and then summed together. Furthermore, the selection of any two blocks (telomeres) in a reference genome $G = 1 \dots n$ defines a partition of the remaining $n-2$ blocks into two sets: (1) blocks that lie between the two chosen telomere blocks in G; and (2) blocks that lie outside the chosen telomere blocks in G. In Case 1 from Theorem 2 all non-telomere blocks in the derivative chromosome lie in between the telomeres and in Case 2 they must lie outside the telomeres. For each possible number of blocks that fall between or outside the telomeres (ranging from $m-2$ up to $n-2$) in G we can explicitly count the number of potential telomere pairs and configurations the $m-2$ other blocks. These observations (along with some algebra) allow us to derive the following formula for $|A(m,n)|$.

$$\begin{aligned}
|A(m,n)| &= \sum_{i=m-2}^{n-2}(2)(i+1)\left(\prod_{j=0}^{m-3}(2i-2j)\right) + \sum_{i=m-2}^{n-2}(2)(n-1-i)\left(\prod_{j=0}^{m-3}(2i-2j)\right) \\
&= 2^{m-1}\sum_{i=m-2}^{n-2}(i+1)\left(\prod_{j=0}^{m-3}(i-j)\right) + 2^{m-1}\sum_{i=m-2}^{n-2}(n-1-i)\left(\prod_{j=0}^{m-3}(i-j)\right) \\
&= 2^{m-1}\sum_{i=m-2}^{n-2}(i+1+n-1-i)\left(\prod_{j=0}^{m-3}(i-j)\right) \\
&= 2^{m-1}n\sum_{i=m-2}^{n-2}\left(\prod_{j=0}^{m-3}(i-j)\right)
\end{aligned}$$

Lastly, we prove Lemma A.1 which is used in the proof for (Main Text) Theorem 2.

Lemma A.1. $\sum_{i=1}^{n}\prod_{j=0}^{m-1}(i+j) = \frac{1}{m+1}\prod_{j=0}^{m}(n+j)$

Proof. We use proof by induction on the variable n. We start with the base case $n=1$.

$$\sum_{i=1}^{1}\prod_{j=0}^{m-1}(i+j) = \prod_{j=0}^{m-1}(1+j) = \frac{1}{m+1}\prod_{j=0}^{m}(1+j)$$

We now assume the property holds for values up to $n-1$ and want to prove for generic n.

$$\begin{aligned}
\sum_{i=1}^{n}\prod_{j=0}^{m-1}(i+j) &= \sum_{i=1}^{n-1}\prod_{j=0}^{m-1}(i+j) + \prod_{j=0}^{m-1}(n+j) \\
&= \frac{1}{m+1}\prod_{j=0}^{m}(n-1+j) + \prod_{j=0}^{m-1}(n+j) \\
&= \frac{n-1}{m+1}\prod_{j=0}^{m-1}(n+j) + \prod_{j=0}^{m-1}(n+j) \\
&= \frac{n+m}{m+1}\prod_{j=0}^{m-1}(n+j) \\
&= \frac{1}{m+1}\prod_{j=1}^{m}(n+j)
\end{aligned}$$

Therefore we have proven the Lemma for general n.

□

Now we can provide a full proof of (Main Text) Theorem 2

Proof. The fraction of chromothripsis strings of length m derived from a reference genome G composed of n intervals that are H/T alternating is just the the number of such chromothripsis strings that are H/T alternating divided by the total number of such chromothripsis strings.

$$\begin{aligned}
Fraction(\pi(C) \text{ H/T alternating } |m,n) &= \frac{|A(m,n)|}{|G(m,n)|} \\
&= \frac{2^{m-1}n\sum_{i=m-2}^{n-2}\left(\prod_{j=0}^{m-3}(i-j)\right)}{2^m\prod_{i=0}^{m-1}(n-i)} \\
&= \frac{\sum_{i=m-2}^{n-2}\left(\prod_{j=0}^{m-3}(i-j)\right)}{2\prod_{i=1}^{m-1}(n-i)} \\
&= \frac{\sum_{i=1}^{n-m+1}(\prod_{j=0}^{m-3}(i+j))}{2\prod_{i=1}^{m-1}(n-i)} \\
&= \frac{\frac{1}{m-1}\prod_{j=0}^{m-2}((n-m+1)+j)}{2\prod_{i=1}^{m-1}(n-i)} \qquad \text{(by Lemma A.1)} \\
&= \frac{\frac{1}{m-1}\prod_{j=1}^{m-1}(n-j)}{2\prod_{i=1}^{m-1}(n-i)} \\
&= \frac{1}{2(m-1)}
\end{aligned}$$

□

B Simulation Details

In this section we provide further details on how we simulate data.

B.1 Chromothripsis Genomes

When simulating genomes that have undergone a chromothripsis event, we aim to create a genome that has a specified number m of novel adjacencies. We do this with the following procedure. We first construct a reference genome that is partitioned into a specified number of blocks n. We then create a random signed permutation of these blocks. We then construct the chromothripsis genome, as a sequence of these blocks by starting at the beginning of the permutation and counting the number of novel adjacencies we observe in the sequence as we move forward through the permutation. One we have a sequence of blocks with m novel adjacencies (or we have run out of blocks), we return that set of blocks as the chromothripsis string C. More specifically, since the $AF(C)$ value is computed using observed adjacencies, we return the set of m novel adjacencies $\mathcal{A}(C)$ observed when traversing the permutation.

We add noise into this data by adding and removing random adjacencies from $\mathcal{A}(C)$. Removal is done by randomly selecting an adjacencies in $\mathcal{A}(C)$ and removing it from the set. In order to add random adjacencies, we first create a set of possible adjacencies to add. We create this set by looking at the permutation of all n blocks used to create the chromothripsis genome C and add to that set every novel adjacency of blocks that was not included in the random chromothripsis genome. We then select adjacencies from this set to add to $\mathcal{A}(C)$.

B.2 Step-wise Mutations

When simulating genomes that have undergone a sequential accumulation of events we begin by breaking the reference genome into a specified number of blocks n. We then add random rearrangement events that are either deletions, duplications or inversions to the genome. The position of a rearrangement event is randomly determined and its size is also randomly determined. We do however restrict the maximum size of a rearrangement event. We want to restrict the size of events in order to obtain a genome that more realistically resembles what might be seen in real data. For example, if we allowed deletion events to contain any number of blocks, often times we may end up deleting large portions of a genome (or nearly all of it) leading to simulations were we are unable to obtain the desired number of adjacencies. For the simulations presented in this work we set $n = 100$ and restrict the size of events to be no larger than 5 genome blocks/segments.

B.3 Step-wise Mutations along with Branching Evolutions

We also create simulations that incorporate evolutionary branching processes, where each branch may contain a different collection of sequentially obtained mutations. We begin by breaking the reference genome into a specified number of blocks n and start with a single, non-rearranged genome in our set of genomic populations. We then pick a single genome in our set of existing populations that we will modify by adding to it a single randomly generated aberration (using the approach described above). However, before adding this rearrangement to the selection population we first decide if this rearrangement will represent a branch in the evolutionary history of this genome. We randomly add a branch with probability α and this consists duplicating the selected genomic population in the set of existing populations and then adding a random rearrangement to one of the copies of this population. We then consider the entire set of novel rearrangements existing with this population of genomes when constructing $\pi(C)$ and determining its H/T alternating fraction. For the simulations presented in this work we set $n = 100$, restricted the size of events to be no larger than 5 genome blocks/segments and used $\alpha = 0.2$. Once $m = 25$ novel aberrations were created within the set of populations we stopped the simulation procedure. With this value of α and m the expected number of tumor populations per simulation is 6 (the original population plus 5 branches).

C Additional Results

C.1 Statistical Analysis of Malhotra *et al.* Data

We performed additional statistical analysis of the 154 sets of adjacencies reported by Malhotra *et al.* [1] as either one-off (chromothripsis) events (97 sets of adjacencies) or as step-wise (57 sets of adjacencies). We compute the H/T alternating fraction across each of these sets using a Mann-Whitney test and determine that the alternating frequency of the one-off events is statistically higher than that step-wise events ($p = 0.0028$), but with a small to medium effect size ($r = 0.22$).

C.2 Additional Figures

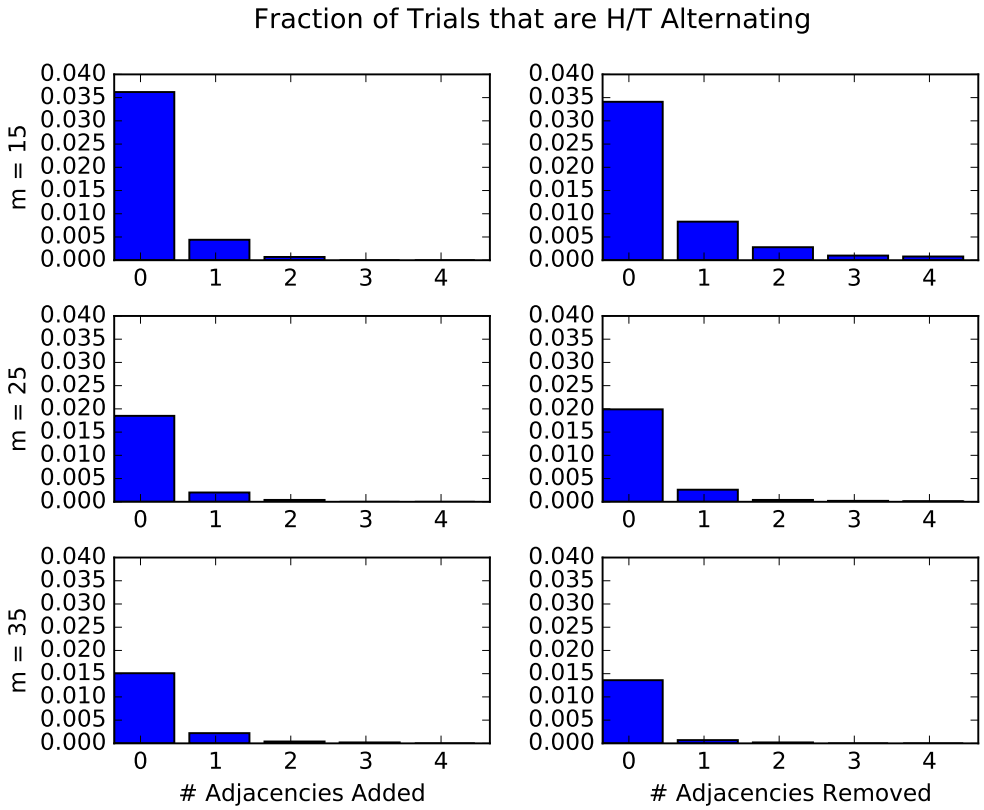

Figure C.1: For each value of m (number of novel adjacencies in the chromothripsis genome) we created $10,000$ random chromothripsis genomes. We then incorporated noise into the data by randomly adding and removing adjacencies and then determined what fraction of the resulting datasets were H/T alternating.

References

[1] Ankit Malhotra, Michael Lindberg, Gregory G Faust, Mitchell L Leibowitz, Royden A Clark, Ryan M Layer, Aaron R Quinlan, and Ira M Hall. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res*, 23(5):762–76, May 2013. doi: 10.1101/gr.143677.112.

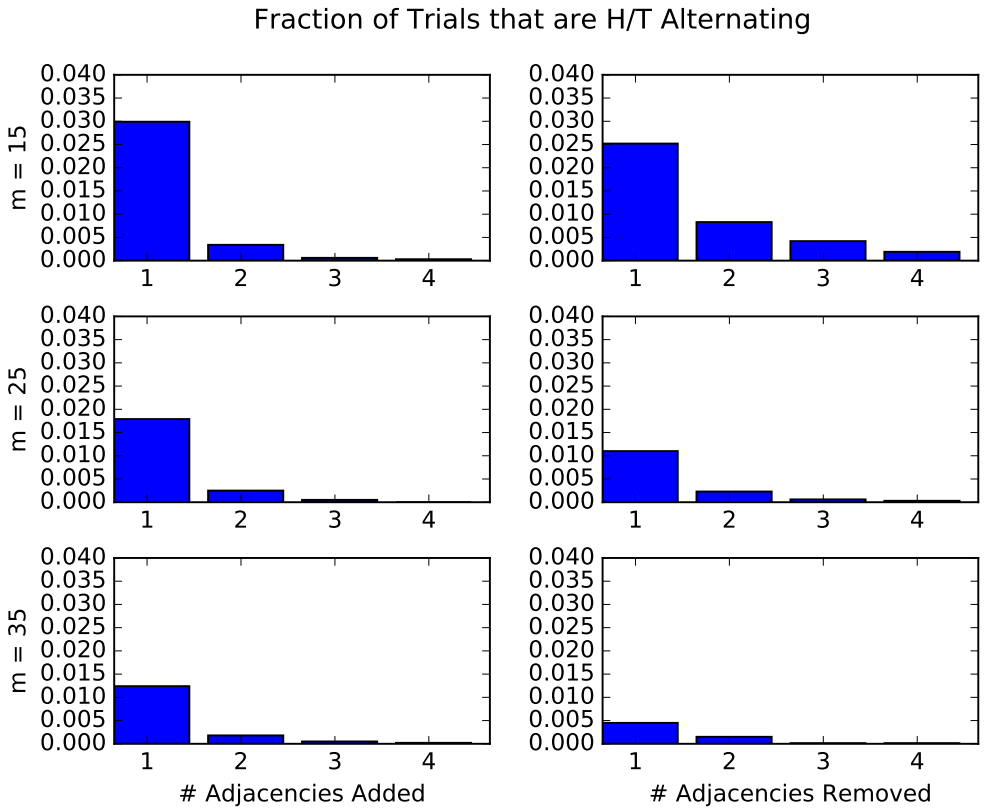

Figure C.2: For each value of m (number of novel adjacencies in the chromothripsis genome) we created $10,000$ random chromothripsis genomes that were initially H/T alternating. We then incorporated noise into the data by randomly adding and removing adjacencies and then determined what fraction of the resulting datasets were H/T alternating.

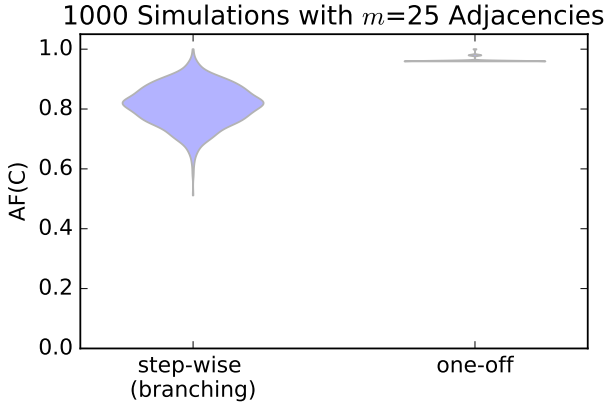

Figure C.3: In error-free simulations, we observe that the H/T alternating fraction $AF(C)$ measure is much higher for one-off (chromothripsis) genomes than genomes that have undergone step-wise sequences of events occurring during branching evolution and have the same number of total novel adjacencies.

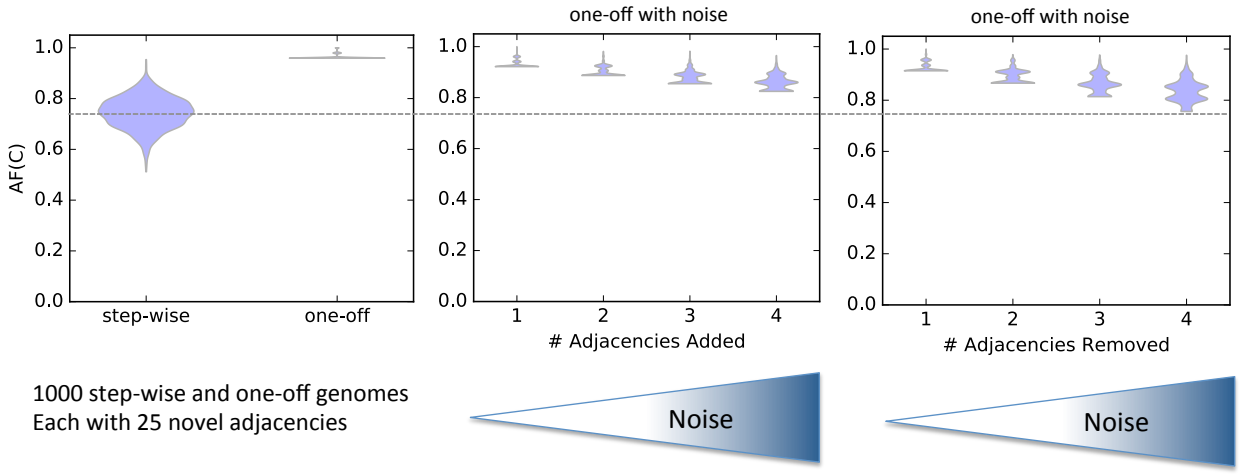

Figure C.4: The H/T alternating fraction $AF(C)$ measure remains higher for one-off (chromothripsis) genomes than genomes that have undergone step-wise sequences of events even as noise is added in terms of added and removed adjacencies.