A weighted exact test for mutually exclusive mutations in cancer

Mark D.M. Leiserson¹, Matthew A. Reyna¹ and Benjamin J. Raphael¹,*

¹Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA

*To whom correspondence should be addressed.

Abstract

Motivation: The somatic mutations in the pathways that drive cancer development tend to be mutually exclusive across tumors, providing a signal for distinguishing driver mutations from a larger number of random passenger mutations. This mutual exclusivity signal can be confounded by high and highly variable mutation rates across a cohort of samples. Current statistical tests for exclusivity that incorporate both per-gene and per-sample mutational frequencies are computationally expensive and have limited precision.

Results: We formulate a weighted exact test for assessing the significance of mutual exclusivity in an arbitrary number of mutational events. Our test conditions on the number of samples with a mutation as well as per-event, per-sample mutation probabilities. We provide a recursive formula to compute P-values for the weighted test exactly as well as a highly accurate and efficient saddle-point approximation of the test. We use our test to approximate a commonly used permutation test for exclusivity that conditions on per-event, per-sample mutation frequencies. However, our test is more efficient and it recovers more significant results than the permutation test. We use our Weighted Exclusivity Test (WExT) software to analyze hundreds of colorectal and endometrial samples from The Cancer Genome Atlas, which are two cancer types that often have extremely high mutation rates. On both cancer types, the weighted test identifies sets of mutually exclusive mutations in cancer genes with fewer false positives than earlier approaches.


Contact: braphael@cs.brown.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

A key challenge in cancer genomics is distinguishing the small number of somatic mutations that drive cancer from the vast majority of mutations that accumulate randomly. The ability to distinguish these driver mutations from the random passenger mutations may lead to better understanding of cancer biology and personalized therapies customized to a tumor’s mutational profile. However, large scale cancer sequencing efforts such as The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network, 2012, 2014; The Cancer Genome Atlas Research Network, et al., 2013) and the International Cancer Genome Consortium (ICGC) have shown that many driver mutations are rare across patient cohorts and thus distinguishing the driver mutations from the passengers by their frequency of occurrence is a difficult problem.

Driver mutations are hypothesized to group into a small number of pathways or hallmarks (Hanahan and Weinberg, 2011), and this hypothesis is a widely accepted explanation for the observed mutational heterogeneity of cancer (Vogelstein et al., 2013). Thus, researchers have developed methods to identify combinations of mutations using varying levels of prior knowledge, from pathway databases (Mootha et al., 2003; Subramanian et al., 2005; Wendl et al., 2011) to protein–protein interaction networks (Ciriello et al., 2012; Leiserson et al., 2015a; Ruffalo et al., 2015; Vandin et al., 2011).

As prior knowledge of pathways and interactions is often noisy or unavailable, de novo methods that do not use prior information are advantageous. The vast number of possible combinations of mutated genes makes complete de novo discovery of combinations computationally and statistically intractable. However, a number of methods (Babur et al., 2015; Ciriello et al., 2012; Constantinescu et al., 2015; Kim et al., 2015, 2016; Leiserson et al., 2013, 2015b; Miller et al., 2011; Szczurek and Beerenwinkel, 2014; Vandin et al., 2012) use the observation that mutations within the same pathway are often mutually exclusive across tumors (Thomas et al., 2007; Yeang et al., 2008). These methods differ in how they score mutual

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com
exclusivity and in how they identify the best scoring set(s) of mutations.

The first type of score for mutual exclusivity is a combinatorial score, such as the scores employed in Miller et al. (2011), Vandin et al. (2012) and Leiserson et al. (2013). For example, in the Dendrix algorithm (Vandin et al., 2012), the score for a set M of mutational events is the difference between the number of samples with a mutation in M (coverage) and the number of mutations in M occurring in more than one sample (coverage overlap). The advantage of a combinatorial score is that it is easy to compute, but it was observed by Leiserson et al. (2015b) and others that the score is often biased towards sets with frequently mutated genes.

The second type of score for mutual exclusivity is a statistical score (Babur et al., 2015; Ciriello et al., 2012; Constantinescu et al., 2015; Kim et al., 2015, 2016; Leiserson et al., 2015b; Ley et al., 2013; Szczurek and Beerenwinkel, 2014). A particularly useful statistical score for exclusivity is based on the exact distribution that conditions on the observed number of mutated samples in each gene (Babur et al., 2015; Leiserson et al., 2015b). For a pair of mutations, such a test is a one-sided Fisher’s exact test for independence (Babur et al., 2015; Leiserson et al., 2015b; Ley et al., 2013). For more than two genes, Leiserson et al. (2015b) generalized the exact test to multi-dimensional contingency tables. They introduced the CoMET algorithm that computes a generalization of Fisher’s exact test for event sets of any size using either an exact tail enumeration algorithm or an approximation. They showed that conditioning on the number of mutations in each event reduces bias towards frequently mutated events compared to combinatorial scores.

Statistical scores that condition only on mutation frequencies do not account for the variation in mutation rate among tumors. It has been observed that the number of mutations in a tumor can vary over several orders of magnitude (Lawrence et al., 2013; Roberts and Gordenin, 2014; Vogelstein et al., 2013). For example, colorectal tumors with microsatellite stability have a median of 66 non-synonymous mutations, but colorectal tumors with microsatellite instability have a median of 777 mutations (Vogelstein et al., 2013). Another example is from The Cancer Genome Atlas Research Network et al. (2013), who classified a subset of TCGA endometrial cancers as ultramutated or hypermutated.

Another useful statistical test for mutual exclusivity conditions on both the number of mutated samples in each event and the number of mutation events in each sample (Ciriello et al., 2012; Kim et al., 2015). Since computing this distribution exactly is not computationally efficient, permutation tests are used. The permutation tests, which compare observed results to a number of samples (~10^8) drawn from a null distribution, are more tractable than computing the P-value exactly on genome-scale data, but the significance of the score is directly limited by the number of permutations. MEMo (Ciriello et al., 2012) computes the significance of the coverage (number of mutated samples) of M using this permutational distribution for sets of any size k. MEMCover (Kim et al., 2015) computes the significance of the exclusivity of pairs to search for exclusivity within, between, and across cancer types. Both MEMo and MEMCover restrict their analysis to sets of genes that interact in a protein-interaction network. WeSME (Kim et al., 2016), which appeared while this paper was under review, computes the significance of exclusivity of pairs of genes with a less expensive approximation to the permutational distribution. To the best of our knowledge, there is no method for quickly computing the significance of mutual exclusivity conditioned on both the observed number of mutations per event and number of mutations per sample.

### 1.1 Contributions

We introduce a weighted test for mutual exclusivity that conditions on the frequency of each mutational event in a set M and also incorporates the probability that each event is mutated in each sample. Our test was inspired by a model derived by Manescu and Keich (2015), who compute the significance of the overlap between two sets of genes weighted by gene length. We introduce the weighted exclusivity test to approximate the fixed event and sample frequency permutation test quickly and accurately by estimating the mutation probabilities from the null distribution of the permutation test. We present a recursive formula for computing the P-value of this test exactly and derive a saddlepoint approximation for arbitrarily sized groups of events. We show that the saddlepoint approximation is both a fast and accurate approximation of the permutational distribution. We also demonstrate that the saddlepoint approximation can be used to rapidly compute the CoMET statistical test, which is a special case of the weighted test where the mutation probabilities for a given event are the same in each sample.

We use our Weighted Exclusivity Test (WExT) software to identify sets of exclusive mutations in hundreds of colorectal and endometrial cancers. Cancer of these types often have extremely high mutation rates (Vogelstein et al., 2013), which make them difficult to analyze when conditioning only on the number of mutated samples per event. However, our weighted statistical test allows us to effectively condition on the number of mutation events per sample, and we identify exclusive patterns of mutations in these cancers that were missed by earlier approaches. We find that the weighted test identifies more biologically interesting sets than CoMET (Leiserson et al., 2015b). We expect that the weighted test for mutual exclusivity will prove useful for many cancer types where defects in DNA damage or environmental exposures, e.g. ultraviolet light, lead to very high mutation rates in some samples.

### 2 Methods

We introduce a new weighted test for mutual exclusivity that incorporates per-event, per-sample mutation probabilities, and we describe how to use particular instances of our test to approximate commonly used tests for mutual exclusivity, which we refer to as the row exclusivity (R-exclusivity) and row-column exclusivity (RC-exclusivity) tests.

First, in Section 2.2, we describe the R-exclusivity and RC-exclusivity tests. Next, in Section 2.3, we introduce our new weighted test for mutual exclusivity, which we call the weighted-row exclusivity (WR-exclusivity) test, that incorporates event and sample mutation frequencies without using permutations. In Section 2.4, we describe how to approximate the R-exclusivity and RC-exclusivity tests with the WR-exclusivity test. Then, in Section 2.5, we provide a recursive formula for computing the WR-exclusivity P-value exactly, and we derive a fast and accurate saddlepoint approximation for the WR-exclusivity P-value. Finally, in Sections 2.6, we describe how we search for exclusive sets, and in Section 2.7, we describe our WExT software.

We summarize the tests and our contributions in this article in Table 1.

### 2.1 Notation

We observe the presence or absence of mutational events across a collection of samples. The presence of an event may reflect a variety of genomic (e.g. the canonical BRAF V600E mutation or deletions in CDKN2A), proteomic and/or epigenomic alterations. In this
Table 1. Three tests for mutual exclusivity, the values that are fixed in each test and different algorithms for computing the P-values associated with the tests

<table>
<thead>
<tr>
<th>Test</th>
<th>P-value</th>
<th>Conditioning</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row (R) exclusivity</td>
<td>$\Phi_R$</td>
<td>Event frequencies</td>
<td>Tail enumeration (CoMEt), saddlepoint approximation</td>
</tr>
<tr>
<td>Row-column (RC) exclusivity</td>
<td>$\Phi_{RC}$</td>
<td>Event and sample frequencies</td>
<td>Permutations</td>
</tr>
<tr>
<td>Weighted-row (WR) exclusivity</td>
<td>$\Phi_{WR}$</td>
<td>Event frequencies and per-event, per-sample weights ($W$)</td>
<td>Recursive formula, saddlepoint approximation</td>
</tr>
</tbody>
</table>

Bold face entries indicate contributions by this manuscript.

study, we analyze single nucleotide variants and small insertion/deletions (indels) by gene. For the clarity of exposition, we will describe these events at the gene level, but our weighted test can accommodate a broader class of mutational events.

Let $\{g_i\}_{i=1}^m$ be a set of $m$ genes and $\{s_j\}_{j=1}^n$ be a set of $n$ samples. For each sample, we observe the presence of one or more mutations in each gene, and we record the presence or absence of mutations in a per-gene, per-sample binary mutation matrix $A \in \{0, 1\}^{m \times n}$, where $A = [a_{ij}]$ with $a_{ij} = 1$ if gene $g_i$ is mutated in sample $s_j$, and $a_{ij} = 0$ otherwise.

Let $M \subseteq \{g_i\}_{i=1}^m$ be a set of $k$ genes. The gene set $M$ has co-occurring mutations in sample $s$ if multiple genes are mutated in that sample, i.e., there exist distinct $g_i, g_j \in M$ such that $a_{is} = 1$ and $a_{js} = 1$. Alternatively, the gene set $M$ has a mutually exclusive mutation in sample $s$, if and only if one gene is mutated in that sample, i.e., there exists $g_i \in M$ such that $a_{is} = 1$ and $a_{js} = 0$ for $g_j \in M \setminus \{g_i\}$. Our goal is to identify sets of genes with statistically significant numbers of mutually exclusive mutations.

Let $r_i = \sum_{j=1}^n a_{ij}$ be the number of samples with mutations in $g_i$, let $c_j = \sum_{i=1}^m a_{ij}$ be the number of genes with mutations in $s_j$, let $z_M$ be the number of samples with co-occurring mutations in $M$ and let $t_M$ be the number of samples with mutually exclusive mutations in $M$.

For any mutation matrix $B$, let $B(M)$ be the submatrix of $B$ with rows corresponding to the gene set $M$, and let $t_{B(M)}$ be the number of mutually exclusive mutations in $B(M)$. We will use $t_M = t_{A(M)}$.

### 2.2 Permutation tests for mutual exclusivity

We describe two different permutation tests for mutual exclusivity. First, the row-exclusivity ($R$-exclusivity) test finds the probability $\Phi_R(M)$ of observing at least $t_M$ mutually exclusive mutations in a gene set $M$ given that each $g_i \in M$ is mutated in $r_i$ samples. We describe this test as the row-exclusivity test because it conditions on the row sums of the mutation matrix.

Formally, we define $\Omega_R$ to be the set of mutation matrices with the same row sums as $A$. Let $E_R = \{ B \in \Omega_R : t_{B(M)} \geq t_M \}$ be the set of mutation matrices with at least $t_M$ mutually exclusive mutations in $M$. Then

$$\Phi_R(M) = \frac{|E_R|}{|\Omega_R|}$$

is the $P$-value of the $R$-exclusivity test.

Since the $R$-exclusivity test only conditions on the row sums of $A$, we can consider each row of $A$ independently. This implies that to compute $\Phi_R(M)$, we use only the rows corresponding to $M$. Thus, for $k = 2$, the $P$-value $\Phi_R(M)$ is equal to the $P$-value from the one-sided Fisher’s exact test, which computes the tail probability by summing the hypergeometric probability of $2 \times 2$ contingency tables with fixed margins. The hypergeometric probability of each contingency table is the proportion of matrices in $\Omega_R$ that gives a contingency table with those margins. Note also that, when $k = 2$, the probability of observing $t_M$ or more mutually exclusive mutations is equal to the probability of observing $t_M$ or more co-occurring mutations. Leiserson et al. (2015) generalized this test to $k > 2$ genes as part of the CoMEt algorithm.

The row-column-exclusivity ($RC$-exclusivity) test finds the probability $\Phi_{RC}(M)$ of observing at least $t_M$ mutually exclusive mutations in a gene set $M$ given that each $g_i \in M$ is mutated in $r_i$ samples and each $s_j$ is mutated in $c_j$ genes. We describe this test as the row-column-exclusivity test because it conditions on the row and column sums of the mutation matrix.

Formally, we define $\Omega_{RC}$ to be the set of mutation matrices with the same row and column sums as $A$. Let $E_{RC} = \{ B \in \Omega_{RC} : t_{B(M)} \geq t_M \}$ be the set of mutation matrices with at least $t_M$ mutually exclusive mutations in $M$. Then

$$\Phi_{RC}(M) = \frac{|E_{RC}|}{|\Omega_{RC}|}$$

is the $P$-value of the $RC$-exclusivity test. Since $\Omega_{RC}$ depends on the row and column sums of $A$, we cannot consider the rows of $A$, or even $A(M)$, independently.

The $RC$-exclusivity test is related to the co-occurrence and mutual exclusivity tests used in Ciriello et al. (2012) and Kim et al. (2015, 2016). Ciriello et al. (2012) use permutation tests that sample matrices from $\Omega_{RC}$ instead of exclusivity as the test statistic, while Kim et al. (2015, 2016) limit to pairs of genes. Both Ciriello et al. (2012) and Kim et al. (2015) use permutation tests that sample matrices from $\Omega_{RC}$, so their $P$-values are limited by the number of draws (e.g., both use $10^5$ permutations).

### 2.3 Weighted exact test for mutual exclusivity

We introduce a new weighted test for mutual exclusivity. The weighted-row-exclusivity ($WR$-exclusivity) test finds the probability $\Phi_{WR}(M)$ of observing at least $t_M$ mutually exclusive mutations in a gene set $M$ given that $g_i \in M$ is mutated in $r_i$ samples and a per-gene, per-sample mutation probability matrix $W$ that prescribes weights with the presence or absence of individual mutations. We describe this test as the weighted-row-exclusivity test because it conditions on the row sums of the mutation matrix and a mutation probability weight matrix.

For our model, we assume that $\{X_{ij}\}_{i=1}^m$ is a set of mutually independent Bernoulli random variables for each gene $g_i$ with success probabilities $W = [w_{ij}]$, i.e.,

$$\Pr(X_{ij} = \ell) = \begin{cases} w_{ij}, & \text{if } \ell = 1, \\ 1 - w_{ij}, & \text{if } \ell = 0, \end{cases}$$

where $w_{ij}$ is the probability that gene $g_i$ is mutated in sample $s_j$. Let $T_M$ be a random variable with $T_M = 1$ if $s_j$ has a mutually exclusive mutation in a gene set $M$ and $T_M = 0$ otherwise. Therefore, $Y_i = \sum_{j=1}^n X_{ij}$ is a Poisson binomial distributed variable for the
number of mutations in \( g_i \) and \( T_M = \sum_{j=1}^n T_{Mj} \) is a test statistic for mutual exclusivity indicating the number of mutually exclusive mutations in \( M \). We want to find the tail probability (commonly referred to as the P-value) of observing \( t_M \) mutually exclusive mutations in \( M \) given that \( g_i \) is mutated in \( r_i \) samples. The WR-exclusivity P-value \( \Phi_{WR}(M) \) is the probability of observing at least \( t_M \) mutually exclusive mutations in a gene set \( M \) under this model with

\[
\Phi_{WR}(M) = \Pr(T_M \geq t_M | Y_M = \tilde{r}_M)
\]  

(4)

where \( Y_M = \{Y_i\}_{i \in M} \) and \( \tilde{r}_M = \{r_i\}_{i \in M} \).

Note that, for any gene \( g_i \) the assumption that \( Y_i = r_i \) implies that

\[
\sum_{j=1}^n {w}_{ij} = \sum_{j=1}^n E[X_{ij}] = E\left[\sum_{j=1}^n X_{ij}\right] = E[Y_i] = r_i
\]  

(5)

by the definitions of \( \{X_{ij}\}_{i,j=1}^n \) and \( Y_i \).

### 2.4 Approximating the permutation tests with the weighted exclusivity test

Each of the sets \( \Omega_R \) and \( \Omega_{RC} \) underlying the R-exclusivity and RC-exclusivity tests, respectively, determines a per-gene, per-sample weight matrix \( W = [w_{ij}] \) by considering the probability \( w_{ij} \) of observing a mutation in gene \( g_i \) in sample \( s_i \), i.e.

\[
W = \frac{1}{|\Omega_k|} \sum_{B \in \Omega_k} B
\]  

(6)

where \( \Omega \in \{\Omega_R, \Omega_{RC}\} \). Since both \( \Omega_R \) and \( \Omega_{RC} \) fix the number of mutated samples per gene, the weight matrix \( W \) in (6) with \( \Omega = \Omega_R \) and \( W_{RC} \) to be the weight matrix with \( \Omega = \Omega_{RC} \).

For the R-exclusivity test, each row of \( B \in \Omega_R \) can be considered separately, so (6) for the set \( \Omega_R \) is given by \( W_R = [w_{ij}] \), with \( w_{ij} = \frac{1}{n} \).

However, for the RC-exclusivity test, each row \( B \in \Omega_{RC} \) cannot be considered separately, so, to the best of our knowledge, there is no closed-form expression for (6) for the set \( \Omega_{RC} \). Therefore, we generate an empirical weight matrix \( W_{RC} = [w_{ij}] \) for \( \Omega_{RC} \) by drawing \( N \) matrices \( \Omega_{RC} \), uniformly at random from \( \Omega_{RC} \) and computing (6) with \( \Omega_{RC} \) instead of \( \Omega_{RC} \). We assume that there is a nonzero probability that a gene is mutated in a sample, and thus set \( w_{ij} = \frac{1}{2n} \) when no mutation in gene \( g_i \) is observed in sample \( s_i \) in \( \Omega_{RC} \).

Estimating \( W_{RC} \) in this way gives an accurate approximation of \( \Phi_{RC}(M) \) using relatively small values of \( N \).

### 2.5 Computing the weighted exclusivity test

Our weighted test for mutual exclusivity requires computing the tail probability in (4), which can be computationally expensive. We compute the tail probability using two different strategies: a recursive formula and a saddlepoint approximation.

#### 2.5.1 Recursive formula for the weighted exclusivity test

We present a recursive formula for computing the tail probability in (4) exactly for sets \( M \) of any size \( k \). Assuming that \( \{Y_i\}_{i=1}^n \) are mutually independent, we can write (4) as

\[
\Phi_{WR}(M) = \frac{\Pr(T_M \geq t_M, Y_M = \tilde{r}_M)}{\Pi_{i \in M} \Pr(Y_i = r_i)}
\]  

(7)

Without loss of generality, let \( M = \{1, \ldots, k\} \). We first find the joint probability in the numerator of (7) using a recursive formula, where \( \Pr(T_M \geq t_M, Y_M = \tilde{r}_M) = F(t_M, r_1, \ldots, r_k, n) \) is computed by the recurrence relation

\[
F(t, x_1, \ldots, x_k, j) = \sum_{w_{ij} \in \{0,1\}} \prod_{i=1}^k q_{ij} F(w_{ix}(t), y_{ix}(x_1), \ldots, y_{ix}(x_k), j - 1),
\]  

(8)

where

\[
q_{ij} = \begin{cases} p_i & \text{if } \ell = 1, \\ 0 & \text{otherwise,} \end{cases}
\]

\[
w_{ix}(t) = \begin{cases} t - 1 & \text{if } \sum_{i=1}^k p_i = 1, \\ t & \text{otherwise,} \end{cases}
\]

and

\[
y_{ix}(x) = \begin{cases} x - 1 & \text{if } \ell = 1, \\ x & \text{otherwise.} \end{cases}
\]

The base cases for (8) are

\[
1, \text{ if } t = x_1 = \cdots = x_k = j = 0,
\]

\[
0, \text{ if } \min\{t, x_1, \ldots, x_k, j\} < 0.
\]

We then find the marginal probabilities in the denominator of (7) using dynamic programming, which is a standard method for computing the Poisson-Binomial probability mass function (Hong, 2013).

#### 2.5.2 Saddlepoint approximation for the weighted exclusivity test

We derive a saddlepoint approximation (Butler, 2007) for computing the conditional tail probability in (4). This approach is inspired by Manescu and Keich (2015), who derive a saddlepoint approximation for an enrichment test for differentially expressed genes in Gene Ontology categories weighted by gene lengths. The saddlepoint approximation is specifically designed to provide a quick and accurate approximation of the tail probability. We present the key equation in (10) and provide a full derivation for \( k = 3 \) in the supplement. The saddlepoint approximation is given by

\[
\Pr(T_M \geq t_M | Y_M = \tilde{r}_M) \approx 1 - \Phi(\tilde{u}) - \phi(\tilde{u}) \left( \frac{1}{\tilde{u}} - \frac{1}{\tilde{v}} \right),
\]  

(10)

where \( \Phi \) and \( \phi \) are, in this setting, the cumulative distribution and density functions, respectively, of the standard normal distribution, and \( \tilde{u} \) and \( \tilde{v} \) are defined as follows.

Without loss of generality, let \( M = \{1, \ldots, k\} \). First, for \( \lambda \in \mathbb{R}^{k+1} \), let \( M_{Y_M \mid T_M}(\lambda) = E\left[\sum_{i=1}^k Y_{ix} \lambda_i T_M\right] \) be the joint moment generating function of \( \{Y_i\}_{i \in M} \) and \( T_M \), and let \( K_{Y_M \mid T_M}(\lambda) = \log M_{Y_M \mid T_M}(\lambda) \) be the corresponding cumulant generating function.

Similarly, let \( M_\lambda(\lambda) = E[e^{\lambda V}] \) be the moment generating function of \( Y_i \), and let \( K_\lambda(\lambda) = \log M_\lambda(\lambda) \) be the corresponding cumulant generating function.
Next, let $\mathbf{k}_Y^\prime \mathbf{r}_Y^\prime (\mathbf{z})$ and $\mathbf{k}_Y^\prime \mathbf{r}_Y^\prime (\mathbf{z})$ be the gradient vector and Hessian matrix, respectively, of $\mathbf{k}_Y^\prime \mathbf{r}_Y^\prime (\mathbf{z})$, and let $\mathbf{k}_Y^\prime \mathbf{r}_Y^\prime (\mathbf{z})$ and $\mathbf{k}_Y^\prime \mathbf{r}_Y^\prime (\mathbf{z})$ be the gradient vector and Hessian matrix, respectively, of $\mathbf{k}_Y^\prime \mathbf{r}_Y^\prime (\mathbf{z})$.

Finally, define $\hat{\mathbf{v}}$ by

$$\hat{\mathbf{v}} = \sqrt{2\text{sgn}(\hat{y}_{k+1})} \left( \sum_{i=M} \mathbf{k}_Y^\prime (\hat{x}_i) - \frac{\mathbf{r}_Y^\prime (\mathbf{y}) - \mathbf{y}^T (\hat{x} - \hat{x})}{\sqrt{\mathbf{k}_Y^\prime (\mathbf{x}_i)}} \right)$$

and $\hat{\mathbf{u}}$ by

$$\hat{\mathbf{u}} = 2\sinh \left( \frac{\hat{y}_{k+1}}{2} \right) \left( \frac{\mathbf{k}_Y^\prime (\mathbf{y})}{\mathbf{r}_Y^\prime (\mathbf{x}_i)} \right)$$

where $\hat{x} = (r_1, \ldots, r_k, t_M - 1/2)$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_{k+1})$ with $\hat{y}$ the unique solution for $\mathbf{k}_Y^\prime (\hat{x}_i) = \hat{x}$ and (10) undefined if $\hat{y}_{k+1} = 0$, and $\hat{x} = (x_1, \ldots, \hat{x}_k, 0)$ with $\hat{x}_i$ the unique solution for $\mathbf{k}_Y^\prime (\hat{x}_i) = r_i$. 

2.6 Searching for sets of mutually exclusive mutations

Our goal is to identify sets $\mathbf{M}$ of genes with significantly exclusive mutations, i.e. extremely small $P$-values $\Phi_{WR}(\mathbf{M})$. There has been a considerable amount of work on methods for optimizing scores for mutually exclusive mutations, including Markov chain Monte Carlo methods (Leiserson et al., 2015b; Vandin et al., 2012), integer linear programs (Leiserson et al., 2013; Zhang et al., 2014), greedy algorithms (Babur et al., 2015) and others. These methods have been shown to be able to search datasets of many hundreds of genes for mutually exclusive mutations. Many of these methods can be modified to use our weighted exclusivity test to identify the most significant sets.

Since the focus of this work is on a statistical test for exclusivity, we instead enumerate and test all sets $\mathbf{M}$ of $k$ genes that satisfy the following basic criteria using the R-exclusivity, RC-exclusivity and WR-exclusivity tests:

1. The number $t_M$ of samples with mutually exclusive mutations must be larger than the number $z_M$ of samples with co-occurring mutations, i.e. $t_M > z_M$.
2. Each gene $g_i \in \mathbf{M}$ must have at least one exclusive mutation.

We use the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR). We examine the subset of genes in each dataset with a minimum mutation frequency so that we can enumerate and test all combinations of genes of a certain size in a reasonable amount of data.

2.7 Implementation

We implemented the recursive formula for the WR-exclusivity test in Python and C, and we implemented the saddlepoint approximation for the WR-exclusivity test in Python using the NumPy and SciPy numerical libraries. We implemented the RC-exclusivity test in Python, and we used a bipartite double edge swap algorithm (Gobbi et al., 2014; Milo et al., 2003) that has been shown empirically to sample uniformly from $\Omega_{RC}$. Our code, along with commands and data for reproducing the results and figures in this paper, is available as the WExT software package at http://compbio.cs.brown.edu/projects/wext.

3 Results

We compare the results of the WR-exclusivity test to both the R-exclusivity and RC-exclusivity tests on real data. In general, we can choose any weights to compute WR-exclusivity, but, in this article, we specifically consider weights to allow us to approximate the R-exclusivity and RC-exclusivity tests. We use WExT to discover mutually exclusive sets of mutations in thyroid, colorectal and endometrial cancers, restricting our analysis to mutations at the gene level.

The rest of this section is organized as follows. In Section 3.1, we describe the data used in our experiments. In Section 3.2, we compare the tail enumeration and saddlepoint approximation algorithms for computing the R-exclusivity $P$-values $\Phi_R(\mathbf{M})$, and we show that the saddlepoint approximation provides a fast and accurate approximation for $\Phi_R(\mathbf{M})$. In Section 3.3, we compare the results of the recursive and saddlepoint approximation algorithms for computing the WR-exclusivity $P$-values $\Phi_{WR}(\mathbf{M})$ with the results of the RC-exclusivity test, and we show that $\Phi_{WR}(\mathbf{M})$ is an accurate approximation of $\Phi_{RC}(\mathbf{M})$ using either the recursive or saddlepoint approximation algorithms. In Section 3.4, we show that $\Phi_{WR}(\mathbf{M})$ provides an accurate approximation of $\Phi_{RC}(\mathbf{M})$ even with coarser estimates of the weight matrix $W$.

3.1 Data

We analyzed non-synonymous single nucleotide variants (SNVs) and small insertions or deletions (indels) in 224 colorectal (COADREAD) (The Cancer Genome Atlas Research Network, 2012), 402 papillary thyroid carcinoma (THCA) (The Cancer Genome Atlas Research Network, 2014), and 248 uterine corpus endometrial carcinoma (UCEC) (The Cancer Genome Atlas Research Network et al., 2013) samples from The Cancer Genome Atlas (TCGA). We analyzed the mutations in the COADREAD and UCEC samples from the TCGA Pan-Cancer project (Weinstein et al., 2013) by downloading the mutations in Mutation Annotation Format (MAF) from Synapse. We downloaded the mutations in THCA from Firehose. We restricted our analysis to non-synonymous mutations, ignoring mutations classified as ‘Silent’, ‘Intron’ ‘3’UTR’, ‘5’UTR’, ‘IGR’, ‘lincRNA’ and ‘RNA’. We also downloaded lists of hypermutator samples for COADREAD and UCEC. We created a list of 35 hypermutator samples in COADREAD listed in the Supplementary Table S3, and 82 hypermutator samples in UCEC listed by The Cancer Genome Atlas Research Network et al. (2013) as samples labeled ‘POLE OR MSI’ in their Supplementary Datafile S1.1. We restrict our analysis to genes mutated in at least 20, 5, and 30 samples in the COADREAD, THCA and UCEC datasets, analyzing 76, 30 and 62 genes in each dataset, respectively.

In general, COADREAD samples have the most mutated genes (median: 78.5), with COADREAD hypermutators with mutations in at least an order of magnitude more genes than non-hypermutators (median for hypermutators: 797; median for non-hypermutators: 69). THCA samples have the fewest mutated genes per sample (median: 12), with no hypermutators, while UCEC has more mutated genes per sample (median: 78.5), with COADREAD hypermutators with mutations in at least an order of magnitude more genes than non-hypermutators (median: 78.5), with COADREAD hypermutators with mutations in at least an order of magnitude more genes than non-hypermutators (median: 78.5), with COADREAD hypermutators with mutations in 62 genes in each dataset, respectively.

3.2 Comparison of methods for computing the R-exclusivity test on real data

First, we investigated the accuracy and speed of the saddlepoint approximation of the R-exclusivity $P$-value $\Phi_R(\mathbf{M})$. We enumerated
triplies according to the procedure described in Section 2.6 in the THCA, COADREAD and UCEC datasets, and computed $\Phi_R(M)$ exactly using the CoMEt software from Leiserson et al. (2015b) as well as approximately using the saddlepoint approximation with $W_R$ given in Section 2.4.

Supplementary Figure S2 shows a comparison of the $P$-values and runtimes given by the two methods, where the weights for the WR-exclusivity test are uniform across samples. On these datasets, the saddlepoint approximation is an extremely accurate approximation of the tail enumeration procedure ($\rho^2 = 0.995$). Additionally, while the median runtimes of the two algorithms are similar, the tail enumeration procedure is much slower for sets with co-occurring mutations while the saddlepoint approximation is largely unaffected. We expect the discrepancy between runtimes to grow for gene sets of larger sizes.

### 3.3 Comparison of methods for computing the WR-exclusivity test on real data

Next, we compared the results of methods for computing WR-exclusivity test with weights $W_R^M$ with the RC-exclusivity test on pairs of genes from the THCA, COADREAD, and UCEC datasets. We chose pairs instead of triples because of the prohibitive cost of computing the recursive formula for $\Phi_R(M)$. We used $N = 10^6$ permutations to compute $\Phi_{RC}(M)$, and also included the tail enumeration procedure for $\Phi_R(M)$ as a control.

Table 2 shows the results of WR-exclusivity test—computed either with the recursive formula or the saddlepoint approximation—are strongly correlated with the RC-exclusivity test (Fig. 2b). The results of the R-exclusivity test are more weakly correlated with the RC-exclusivity test (Fig. 2a), showing that conditioning on the number of mutations in each sample changes the distribution of mutually exclusive mutations. This discrepancy remains when we restrict to gene sets $M$ with $\Phi_{RC}(M) \geq 10^{-4}$, i.e. sets of genes for which the empirical permutation distribution finds at least one mutually exclusive mutation in $M$.

The WR-exclusivity $P$-values computed exactly and with the saddlepoint approximation are highly correlated (Fig. 2c), with a Pearson’s correlation coefficient of 0.996 for all $P$. For smaller $P$-values with $\Phi_R(M) < 10^{-4}$ from either the recursive formula or the saddlepoint approximation, the correlation increases to 0.9999.

The runtime to compute $\Phi_R$ using the recursive formula varies widely because pairs with co-occurring mutations require more computation, but the runtime of the saddlepoint approximation is more consistent. As a result, testing all pairs with the recursive formula requires approximately 2 hours, but testing the same pairs with the saddlepoint approximation requires approximately 30 seconds. Note that the runtime does not include generating the weights $W_R^M$, which requires several minutes.

### 3.4 Approximating the RC-exclusivity test with the WR-exclusivity test

We compared the saddlepoint approximation of the WR-exclusivity test to the RC-exclusivity test using gene triples from the COADREAD dataset, again using the R-exclusivity test as a control. We computed $\Phi_{RC}(M)$ with $N = 10^6$ permutations. We computed the saddlepoint approximation for $\Phi_{WR}(M)$ using $W_R^M$ with $N = 10^4$ draws from $\Omega_{RC}$, which is three orders of magnitude fewer than the number of permutations than we used to compute $\Phi_{RC}(M)$. The $P$-values $\Phi_R(M)$ and $\Phi_{RC}(M)$ are weakly correlated in the tail ($\rho^2 = 0.67$ for $\Phi_{RC}(M) < 0.001$; see Fig. 3a). In contrast, the $\Phi_{WR}(M)$ (saddlepoint) $P$-values provide an accurate approximation of the $\Phi_{RC}(M) P$-values. The RC-exclusivity and WR-exclusivity $P$-values are highly correlated in the tail ($\rho^2 = 0.948$ for $\Phi_{RC}(M) < 0.001$; see Fig. 3b). Moreover, $\Phi_{WR}(M)$ is an accurate estimate of $\Phi_{RC}(M)$ to within one or more digits for most triples and within an order of magnitude for all triples. Furthermore, despite the much smaller number of permutations used to generate $W_R^M$, $\Phi_{WR}(M)$ provides smaller $P$-values than $\Phi_{RC}(M)$ with a greater number of significant predictions, and is much faster than the permutation test.
3.5 Mutually exclusive mutations in thyroid carcinomas

We computed the WR-exclusivity \( P \)-values for all triples of genes that were each mutated in at least 5 of the 402 thyroid carcinomas in the THCA dataset. The WR-exclusivity test identifies 48 triples with significantly exclusive mutations (FDR < 0.001), while the R-exclusivity test identifies 38 triples (FDR < 0.001). The top 25 ranked triples by both tests are identical, which is not surprising since THCA samples have low mutation rates compared to most cancer types [see Vogelstein et al. (2013) and Section 3.1]. In addition, the \( P \)-values for the top ranked triples are all within a few orders of magnitude, demonstrating that the two tests are very similar on this dataset.

Supplementary Table S1 shows the top triples, which include many known thyroid cancer genes. The top five triples include seven genes, five of which are well-known cancer genes with known roles in thyroid cancer (The Cancer Genome Atlas Research Network, 2014): \textit{BRAF}, \textit{HRAS}, \textit{NRAS}, \textit{EIF1AX} and \textit{ATM}. The other two genes are \textit{BDP1} and \textit{TG}, both of which may play a role in cancer. Woiwode et al. (2008) describe a role for \textit{BDP1} in AKT signaling, which was also noted in TCGA thyroid publication (The Cancer Genome Atlas Research Network, 2014), although \textit{BDP1} has greater than 11 000 nucleotides in its coding sequence, so it may also accumulate many passenger mutations. \textit{TG} is the thyroglobulin gene, and is used as a tumor marker in papillary thyroid carcinoma, which is the same subtype of thyroid cancer analyzed in TCGA.

3.6 Mutually exclusive mutations in colorectal cancers and endometrial carcinomas

We expect that the difference between the R-exclusivity and WR-exclusivity tests would be more pronounced on cancer types with higher and highly variable mutation rates. Thus, we computed \( P \)-values on triples of genes from colorectal cancers (COADREAD) and endometrial carcinomas (UCEC). We find that the WR-exclusivity test predicts more biologically interesting triples than the R-exclusivity test. The WR-exclusivity test identifies 5290 and 6835 triples (many of which overlap) with significantly mutually exclusive mutations (FDR < 0.001) in the 224 COADREAD and 248 UCEC samples, respectively. In contrast, the R-exclusivity test computes 4 and 130 triples, respectively, with significantly mutually exclusive mutations (FDR < 0.001).

Compared to the R-exclusivity test results, the highest ranked triples by the WR-exclusivity test include fewer long genes that tend to
accumulate random, passenger mutations—especially in samples with high mutation rates (Tables 3 and 4).

On COADREAD, the WR-exclusivity test identifies ten different genes in the five most significant triples (Table 3). Nine of these genes are well-known cancer genes—BRAF, KRAS, NRAS, AOV2RA, PIK3CA, TP53, ATM, TGFBR2 and ARID1A—while the 10th gene (ABCA12) is known to have an association with colorectal cancers (Hlavata et al., 2012). The R-exclusivity test results are similar—two of the top five triples identified by the WR-exclusivity test are in the top five triples identified by the R-exclusivity test—but the R-exclusivity test does not identify ARID1A, TGFBR2, KRAS or NRAS. Further, the three additional genes identified by the R-exclusivity—APC, FAT2 and WDFY3—are all in the top 600 longest genes in the human genome (at least 9,560 nucleotides in coding sequence). While mutations in APC are well-known to play a role in colorectal cancers, there is currently little evidence for the roles of FAT2 or WDFY3 in cancer, and it is likely that these long genes have accumulated many passenger mutations, particularly in hypermutated samples. Also of note is the fact that the number of hypermutator samples that contain mutations in the top triples from the WR-exclusivity test are not appreciably different from the number of hypermutator samples that contain mutations in the top triples from the R-exclusivity test. This demonstrates that the WR-exclusivity test is not systematically excluding hypermutator samples from consideration, but rather weighting the contribution of these samples appropriately in evaluating the significance of mutual exclusivity.

On UCCEC, the differences between the R-exclusivity and WR-exclusivity tests are even more pronounced. The WR-exclusivity test identifies seven genes in the top five most significant triples (Table 4). These include six genes with known roles in cancer—CTTNB1, TP53, RPL22, KRAS, CTCF and ML14—with only one gene—RYRI—with likely spurious mutations. In contrast, the top five triples ranked by the R-exclusivity test include PTEN, and TP53—two well-known cancer genes—but also five genes with no known role in cancer that all have greater than 11,000 nucleotides in their coding sequences: CACNA1E, LAMA2, RYRI2, NBEA and FAT4. Further, none of the top five triples identified by the WR-exclusivity test are in the top twenty R-exclusivity triples. Finally, the R-exclusivity triples include many more mutations in hypermutator samples (ranging from mutations in 75 to 78 of the 81 hypermutators, versus 40 to 47 for the WR-exclusivity triples). This further demonstrates how the results of the R-exclusivity test are skewed by hypermutator samples, while the WR-exclusivity test incorporates the contribution of these samples appropriately in evaluating the significance of mutual exclusivity.

4 Discussion

We introduce a weighted exact test for the mutual exclusivity of mutations in cancer. We use this test to approximate the permutation test for exclusivity where the number of mutations in each event and each sample are fixed. To do so, we estimate per-event, per-sample mutation probabilities directly from the permutational distribution. We derive a recursive formula and a saddlepoint approximation of the $P$-value of the weighted test for event sets of any size, and we demonstrate the accuracy and efficiency of the saddlepoint approximation on genome-scale mutation datasets. Together, these contributions allow us to overcome the significant computational challenge of finding highly significant sets of mutually exclusive

| Table 3. Five most significant triples identified by the R-exclusivity (top 5) and WR-exclusivity (bottom 5) tests on the COADREAD dataset |
|---|---|---|---|---|---|
| $\Phi_R$ rank | $\Phi_WR$ rank | Triple M | $\Phi_R(M)$ | $\Phi_WR(M)$ | Hypermutator mutations |
| 1 | 2 | ACVR2A, PIK3CA, TP53 | $2.65 \times 10^{-7}$ | $2.54 \times 10^{-18}$ | 31 |
| 2 | 32 | APC, BRAF, PRDM2 | $5.44 \times 10^{-7}$ | $2.30 \times 10^{-13}$ | 33 |
| 3 | 33 | APC, BRAF, WDFY3 | $5.44 \times 10^{-7}$ | $2.43 \times 10^{-13}$ | 32 |
| 4 | 3 | ATM, PIK3CA, TP53 | $5.87 \times 10^{-7}$ | $1.15 \times 10^{-17}$ | 24 |
| 5 | 81 | APC, BRAF, FAT2 | $1.93 \times 10^{-6}$ | $6.48 \times 10^{-32}$ | 35 |
| 6 | 1 | BRAF, KRAS, NRAS | $2.50 \times 10^{-6}$ | $9.95 \times 10^{-39}$ | 26 |
| 1 | 2 | ACVR2A, PIK3CA, TP53 | $2.65 \times 10^{-7}$ | $2.54 \times 10^{-18}$ | 31 |
| 4 | 3 | ATM, PIK3CA, TP53 | $5.87 \times 10^{-7}$ | $1.15 \times 10^{-17}$ | 24 |
| 10 | 4 | ARID1A, TGFBR2, TP53 | $5.89 \times 10^{-6}$ | $1.76 \times 10^{-36}$ | 29 |
| 9 | 5 | ABCA12, TGFBR2, TP53 | $4.29 \times 10^{-6}$ | $1.83 \times 10^{-56}$ | 28 |

Genes in bold are among 600 longest genes (at least 9,560 nucleotides in coding sequence).

| Table 4. Five most significant triples identified by the R-exclusivity (top 5) and WR-exclusivity (bottom 5) tests on the UCCEC dataset |
|---|---|---|---|---|---|
| $\Phi_R$ rank | $\Phi_WR$ rank | Triple M | $\Phi_R(M)$ | $\Phi_WR(M)$ | Hypermutator mutations |
| 1 | 20 | CACNA1E, PTEN, TP53 | $3.11 \times 10^{-12}$ | $5.71 \times 10^{-30}$ | 77 |
| 2 | 21 | LAMA2, PTEN, TP53 | $4.13 \times 10^{-12}$ | $8.05 \times 10^{-30}$ | 77 |
| 3 | 29 | PTEN, RYRI2, TP53 | $4.60 \times 10^{-12}$ | $4.85 \times 10^{-29}$ | 78 |
| 4 | 28 | NBEA, PTEN, TP53 | $8.40 \times 10^{-12}$ | $3.32 \times 10^{-29}$ | 76 |
| 5 | 39 | FAT4, PTEN, TP53 | $1.23 \times 10^{-11}$ | $3.00 \times 10^{-33}$ | 75 |
| 22 | 1 | CTNNB1, RPL22, TP53 | $2.11 \times 10^{-10}$ | $1.20 \times 10^{-41}$ | 47 |
| 44 | 2 | CTNNB1, KRAS, TP53 | $3.05 \times 10^{-9}$ | $1.10 \times 10^{-37}$ | 48 |
| 55 | 3 | CTNNB1, ML14, TP53 | $4.26 \times 10^{-9}$ | $5.78 \times 10^{-36}$ | 42 |
| 57 | 4 | CTCF, CTNNB1, TP53 | $4.84 \times 10^{-9}$ | $3.29 \times 10^{-35}$ | 43 |
| 60 | 5 | CTNNB1, RYRI1, TP53 | $1.14 \times 10^{-7}$ | $9.51 \times 10^{-55}$ | 40 |

Notation as in Table 3.
mutations conditioned on both the observed number of mutations per-event and per-sample.

We then demonstrate the weighted test on three datasets with hundreds of samples from TCGA, including colorectal and endometrial cancers that have high variability in the number of mutations per sample. The weighted test identifies sets of mutually exclusive mutations including known cancer genes in each dataset, and its results include many fewer long genes and mutations in hypermutator samples than do the results of the generalization of Fisher’s exact test from CoMe (Leiserson et al., 2015b).

There are several avenues for improving analyses with the weighted test. First, while we restricted our study to non-synonymous SNVs and indels, one should also test mutual exclusivity between other types of aberrations, such as copy number aberrations and gene fusions. We searched for mutually exclusive mutations by enumerating sets containing the most mutated genes, but the weighted test could easily be used in existing algorithms for optimizing mutual exclusivity scores [e.g. the MCMC from Vandin et al. (2012), the greedy approach from Babur et al. (2015)] or to search for multiple sets simultaneously [e.g. from Leiserson et al. (2015b)]. We estimated the per-event, per-sample mutation probability weights directly from the permutational distribution, but we also anticipate alternative methods for setting the weights that incorporate different event or sample attributes, such as gene length, to further reduce the number of false positives.

The weighted test may be of broader interest beyond searching for mutually exclusive mutations, both in other areas of computational biology and other disciplines. For example, statistical tests of ‘presence–absence’ matrices with fixed row and column sums are a common tool in ecology for looking at species-associations, but can be computationally prohibitive (Miklós and Podani, 2004). The weighted exact test presented here may offer a fast, alternative approach for computing the significance of associations with high accuracy.

Acknowledgments

The authors would like to acknowledge Manescu and Keich (2015) for inspiring this work, and Uri Keich (2015) for generously helping us run their code and providing comments on our manuscript.

Funding

This study was supported by the US National Institutes of Health (NIH) grants (R01HG006569, R01HG007069 and R01CA180776 to B.J.R.), NSF fellowship (GRFP DGE 0228243 to M.D.M.L.) and a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship and an NSF CAREER Award (CCF-1053753 to B.J.R.).

References

https://doi.org/10.7303/syn1710680.4.
Level_3_2016012800.0.tar.gz.
http://www.mayomedicallaboratories.com/test-catalog/
Clinical+and+Int+Interpretive/62800.


