A latent variable model for evaluating mutual exclusivity and co-occurrence between driver mutations in cancer

Ahmed Shuaibi[†1,2], Uthsav Chitra[†1], and Benjamin J. Raphael[1]

[1]Department of Computer Science, Princeton University
[2]Lewis-Sigler Institute for Integrative Genomics, Princeton University

Abstract

A key challenge in cancer genomics is understanding the functional relationships and dependencies between combinations of somatic mutations that drive cancer development. Such *driver* mutations frequently exhibit patterns of *mutual exclusivity* or *co-occurrence* across tumors, and many methods have been developed to identify such dependency patterns from bulk DNA sequencing data of a cohort of patients. However, while mutual exclusivity and co-occurrence are described as properties of driver mutations, existing methods do not explicitly disentangle functional, driver mutations from neutral, *passenger* mutations. In particular, nearly all existing methods evaluate mutual exclusivity or co-occurrence at the gene level, marking a gene as mutated if any mutation – driver or passenger – is present. Since some genes have a large number of passenger mutations, existing methods either restrict their analyses to a small subset of suspected driver genes – limiting their ability to identify novel dependencies – or make spurious inferences of mutual exclusivity and co-occurrence involving genes with many passenger mutations. We introduce DIALECT, an algorithm to identify dependencies between pairs of *driver* mutations from somatic mutation counts. We derive a latent variable mixture model for drivers and passengers that combines existing probabilistic models of passenger mutation rates with a latent variable describing the unknown status of a mutation as a driver or passenger. We use an expectation maximization (EM) algorithm to estimate the parameters of our model, including the rates of mutually exclusivity and co-occurrence between drivers. We demonstrate that DIALECT more accurately infers mutual exclusivity and co-occurrence between driver mutations compared to existing methods on both simulated mutation data and somatic mutation data from 5 cancer types in The Cancer Genome Atlas (TCGA).

Availability: DIALECT is available online at https://github.com/raphael-group/dialect.

Contact: braphael@princeton.edu

[†]These authors contributed equally.

1 Introduction

Cancer is an evolutionary process driven by a small number of somatic *driver* mutations against a larger background of random and functionally neutral (or slightly deleterious) *passenger* mutations [28, 80, 49]. Distinguishing driver mutations from passenger mutations and understanding the function of driver mutations is critical for understanding cancer progression and for developing targeted cancer therapies [25]. To this end, large-scale sequencing projects such as the International Cancer Genome Consortium (ICGC) [32, 81] and The Cancer Genome Atlas (TCGA) [51, 9, 44, 37, 76, 5] have measured somatic mutations in large cohorts of tumor samples, allowing for the systematic analysis of driver mutations across many different cancer types.

Beyond the prioritization of individual driver mutations and genes, another important problem in cancer genomics is understanding the functional relationships and dependencies between *combinations* of driver mutations. For example, it has been empirically observed that certain pairs or sets of driver mutations are *mutually exclusive*, meaning that these driver mutations are observed in the same tumor sample less frequently than expected by chance [78]. A widely held explanation for such observed mutual exclusivity is that driver mutations are grouped into a small number of biological pathways, such that a single driver mutation is sufficient to perturb a pathway in a tumor. Combined with the relatively small number of driver mutations in a single tumor, two driver mutations rarely occur in the same pathway. For example, driver mutations in the *KRAS* and *BRAF* genes – two oncogenes in the Ras/Raf/MAP-kinase signaling pathway – have been observed to be mutually exclusive across large cohorts of colorectal cancer samples [18, 7]. Another explanation for mutual exclusivity is synthetic lethality where a pair of mutations – but not the individual mutations – result in cell death [56, 34]. On the other hand, some pairs or sets of driver mutations are *co-occurring*, meaning that they are observed in the same tumor sample more often than expected, e.g. the *VHL/SETD2/PBRM1* mutations in renal cancer [73]. Co-occurrence between driver mutations is observed to be much rarer than mutual exclusivity [10] and may result from some pathways requiring multiple mutations to be perturbed [72].

Numerous computational methods have been developed over the past decade to identify pairs (or larger sets) of genes with mutually exclusive or co-occurring mutations (reviewed by [63, 70, 53]). Importantly, although dependency relationships such as mutual exclusivity and co-occurrence are often described as properties of individual driver mutations, the typical practice is to analyze these dependencies at the *gene* level, treating all observed nonsynonymous single-nucleotide mutations in a gene identically [52, 72, 41, 43, 15, 10, 42, 68, 36, 16, 35, 2, 45]. (Some methods also analyze larger alterations such as copy number aberrations (CNAs) or DNA methylation changes [59, 41, 10], but we restrict our attention to single nucleotide somatic mutations, which are the vast majority of somatic mutations analyzed by existing methods.) There are three major reasons why mutual exclusivity and co-occurrence analysis is typically performed at the gene level. First, it is often unknown *a priori* which somatic mutations are driver mutations and which are passenger mutations, and the classification of mutations as drivers or passengers remains an active area of research [63]. Second, beyond a small number of mutational hotspots [74], individual genomic positions are mutated infrequently in the available cohorts of hundreds to thousands of patients. Third, it is computationally intractable to analyze all combinations of somatic mutations in a cohort, as most cancers are estimated to contain 1,000-20,000 somatic mutations [48].

Methods for identifying dependencies between driver mutations at the gene level do not explicitly account for passenger mutations. Instead, existing methods typically aggregate all somatic mutations in a gene – both drivers *and* passengers – into a single mutational event. Most of these methods use *ad hoc* procedures to restrict analysis to a small subset of genes that are predicted to be driver genes. However, requiring such prior knowledge substantially limits the ability of these methods to identify novel sets of mutually exclusive or co-occurring driver mutations. On the other hand, if existing methods are used to analyze larger lists of genes, then these methods will identify many *spurious* dependencies involving non-driver mutations. For example, we show that existing methods often identify mutual exclusivity involving mutations in the genes *TTN* or *MUC16*, two genes which are hypothesized to not carry any driver mutations and instead have large numbers of passenger mutations due to their length (>60,000 base-pairs) and high background mutation rates [40]. This empirical observation suggests that separately modeling driver and passenger mutations is a promising approach for identifying dependencies between drivers.

Separately, there is a large line of work on identifying individual driver genes from somatic mutation data (e.g. [69, 20, 40, 75, 67, 21, 27, 30, 4, 55, 26, 3, 13, 12]). Some of these algorithms implicitly (or explicitly) model the number the number of passenger mutations inside each gene, i.e. a *background mutation rate model*, and they identify individual genes whose number of observed somatic mutations is significantly greater than expected under the background mutation model. Critically, such algorithms do not identify genes like *TTN* or *MUC16* as driver genes,

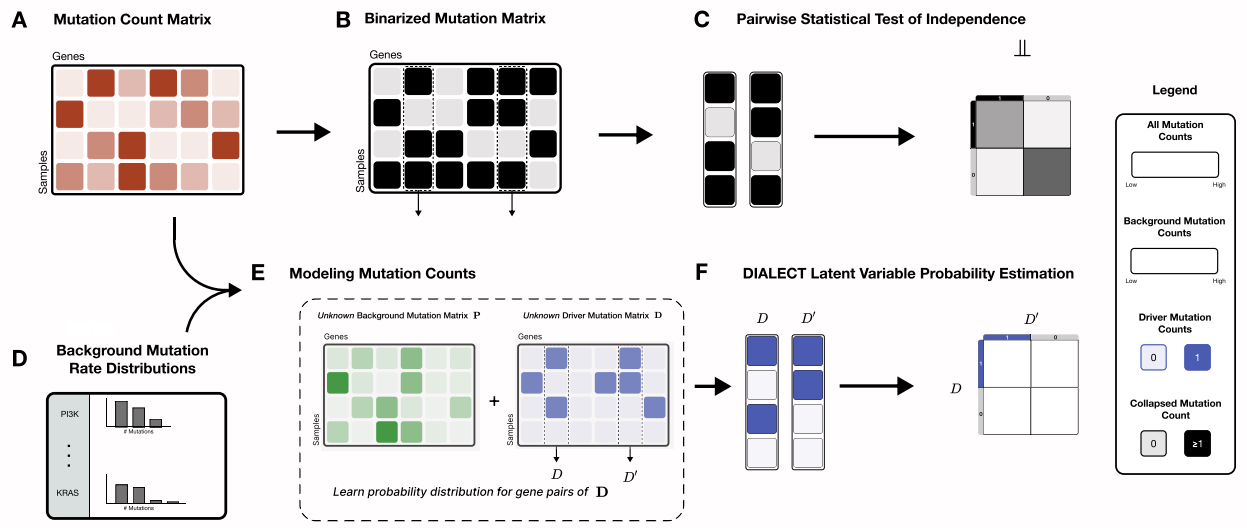

Figure 1: Overview of DIALECT. **(A)** From DNA sequencing data, one obtains a count matrix $C = [c_{ij}]$ indicating the number of nonsynonymous somatic mutations in genes across tumor samples. **(B)** Existing methods for identifying mutually exclusive driver mutations first create a binarized count matrix $X = [x_{ij}] = [1_{\{c_{ij}>0\}}]$ and **(C)** test for independence between pairs of genes. By binarizing the somatic mutation counts, these methods conflate driver mutations versus random, passenger mutations. **(D)** Separately, several algorithms estimate background mutation rate distributions, or the distribution of the number of passenger mutations inside a gene, in order to identify individual driver genes. **(E)** DIALECT explicitly models the distribution of somatic mutation counts $C_i = P_i + D_i$ and $C'_i = P'_i + D'_i$ for two genes as a sum of passenger mutations P_i, P'_i, respectively, and latent variables D_i, D'_i, respectively, indicating the presence or absence of driver mutations. DIALECT incorporates background mutation rate distributions $\mathbb{P}(P_i)$ learned by prior approaches. **(F)** DIALECT learns the parameters $\tau = (\tau_{00}, \tau_{01}, \tau_{10}, \tau_{11})$ of the driver mutation distribution $\mathbb{P}(D_i, D'_i)$ which describes dependencies between drivers including mutual exclusivity and co-occurrence.

as they derive background mutation models using genomic features correlated with increased passenger mutation rates including gene length, replication timing, and synonymous mutation rate [40]. However, these algorithms only model the distribution of passenger mutations inside individual genes, and have not been used to model the distribution of *driver* mutations inside pairs or larger sets of genes.

We introduce a new algorithm, Driver Interactions and Latent Exclusivity or Co-occurrence in Tumors (DIALECT), to identify pairs of genes with mutually exclusive and co-occurring *driver* mutations. We derive a latent variable model for dependencies between driver mutations in a pair of genes, which combines existing probabilistic models of background mutation rates with latent variables that describe the presence or absence of driver mutations in each gene. Importantly, by incorporating existing background mutation rate models, we identify combinations of driver mutations *de novo*; unlike existing approaches, we do not need ad hoc heuristics to analyze small subsets of previously studied driver genes. We derive an expectation-maximization (EM) algorithm to learn the parameters of our model, which describe the rates of mutual exclusivity and co-occurrence between a pair of driver mutations. We use DIALECT to identify dependencies in simulated data and to identify pairs of genes with mutually exclusive driver mutations in real somatic mutation data across 5 cancer subtypes. We show that DIALECT has improved statistical power and lower false positive rate compared to existing methods.

2 Methods

We derive a latent variable model for evaluating mutual exclusivity and co-occurrence between driver mutations in a pair of genes. We assume we are given as input a count matrix $C = [c_{ij}] \in \mathbb{R}^{N \times G}$ indicating the number of non-synonymous somatic mutations in G genetic loci (e.g. genes) across N tumor samples. We aim to test whether each pair (j, j') of genes has mutually exclusive driver mutations. For ease of notation, we omit the subscripts j and focus our exposition on a single pair of genes, where the first gene has somatic mutation counts $\boldsymbol{c} = [c_i] \in \mathbb{R}^N$ and

the second gene has somatic mutation counts $\boldsymbol{c}' = [c'_i] \in \mathbb{R}^N$.

Let C_i and C'_i be random variables indicating the number of somatic mutations observed in two genes, respectively, in tumor sample $i = 1, \ldots, N$. We assume the somatic mutation count C_i (resp. C'_i) in each sample i is equal to the sum of two independent random variables: (1) the number P_i (resp. P'_i) of *passenger* mutations in sample i, and (2) an indicator variable $D_i \in \{0, 1\}$ (resp. $D'_i \in \{0, 1\}$) describing the presence or absence of a *driver* mutation in the gene in sample i, i.e.

$$C_i = P_i + D_i \quad \text{and} \quad C'_i = P'_i + D'_i. \tag{1}$$

We note that we assume that there is at most one driver mutation in a gene in a given sample, which is a reasonable assumption in many cases[1].

We aim to estimate the joint distribution $\mathbb{P}(D_i, D'_i)$ of driver mutations, which describes *dependencies* between driver mutations, i.e. when the random variables D_i and D'_i are *not independent*. For example, mutual exclusivity (ME) corresponds to $\mathbb{P}(D'_i = 1 \mid D_i = 1) < \mathbb{P}(D'_i = 1)$ while co-occurrence (CO) corresponds to $\mathbb{P}(D'_i = 1 \mid D_i = 1) > \mathbb{P}(D'_i = 1)$. (Note that if D_i and D'_i are independent, then $P(D'_i = 1 \mid D_i = 1) = P(D'_i = 1)$.)

We emphasize that existing methods do not model the distribution $\mathbb{P}(D_i, D'_i)$ of driver mutations. Instead, these methods first binarize the somatic mutation counts, forming the matrix $\boldsymbol{X} = [x_{ij}]$ where $x_{ij} = 1_{\{c_{ij}>0\}}$, and then analyze the binarized mutation counts $\boldsymbol{x} = [x_i] \in \{0, 1\}^N$ and $\boldsymbol{x}' = [x'_i] \in \{0, 1\}^N$ for a pair of genes, respectively (Figure 1A-C). Typically, each binarized counts x_i (resp. x'_i) is modeled as a sample of a random variable X_i (resp. X'_i), and one aims to test whether the random variables X_i and X'_i are independent. For example, a classical approach for testing CO and ME is Fisher's exact test, which tests for independence by using a hypergeometric model for the entries of a 2×2 contingency table formed from the binarized counts $(x_i, x'_i)_{i=1}^N$.

The key challenge in estimating the distribution $\mathbb{P}(D_i, D'_i)$ of driver mutations is that we only observe the total number C_i, C'_i of somatic mutations in a sample and *not* the number P_i, P'_i of passenger mutations (or equivalently the value of D_i, D'_i). Although the number P_i of passenger mutations is unknown, many methods have been developed to predict driver genes [69, 20, 40, 75, 67, 21, 27, 30, 4, 55, 26, 3, 13, 12] and some of these implicitly (or explicitly) estimate the *distribution* $\mathbb{P}(P_i)$ of the number P_i of passenger mutations – sometimes called a *background mutation rate* (BMR) distribution (Figure 1D). Note that distributions $\mathbb{P}(P_i)$ may differ across samples $i = 1, \ldots, N$ for a variety of reasons, e.g. some tumor samples being hypermutators [65]. In the next section, we show how to use the BMR distributions $\mathbb{P}(P_i)$ to estimate the distribution of driver mutations.

2.1 Driver distribution for a single locus

We start by studying the simple problem of estimating the driver mutation distribution $\mathbb{P}(D_i)$ in a *single* genetic locus. We will then demonstrate that our approach readily extends to learning the distribution of driver mutations in a pair (or any larger combination) of genetic loci.

We make the simplifying assumption that the driver mutation random variables D_i are *independent and identically distributed* (i.i.d.) across all tumor samples $i = 1, \ldots, N$, i.e. the probability of a locus having a driver mutation does not depend on the specific tumor sample. This assumption is motivated by many standard models of tumor growth, where the probability of a cell receiving a driver mutation does not depend on which other mutations are present in the cell [8, 23]. The assumption that a particular driver mutation is identically distributed across tumor samples may not always hold, but we demonstrate below that this assumption allows for tractable estimation of the distribution $P(D_i)$ of driver mutations and works well in practice. Under this assumption, the driver mutations D_i are each independently distributed according to a Bernoulli distribution $\text{Bern}(\pi)$ with a shared parameter π, representing the *driver mutation rate* across all samples $i = 1, \ldots, N$.

Then, the distribution $\mathbb{P}(C_i)$ of somatic mutation count C_i in sample i is given by

$$\begin{aligned} \mathbb{P}(C_i = c_i) &= \mathbb{P}(C_i = c_i \mid D_i = 0)\mathbb{P}(D_i = 0) + \mathbb{P}(C_i = c_i \mid D_i = 1)\mathbb{P}(D_i = 1) \\ &= \mathbb{P}(P_i = c_i)(1 - \pi) + \mathbb{P}(P_i = c_i - 1)\pi, \end{aligned} \tag{2}$$

where we use that passenger mutations P_i and driver mutations D_i are independent in the second equation. We set $\mathbb{P}(P_i = -1) = 0$ for notational simplicity, so that the probability of zero somatic mutations in a loci is given

[1] One notable exception are tumor suppressor genes where both copies of the gene are typically inactivated ("two hit hypothesis"). However, it is common for one of these mutations to be a copy number aberration.

by $\mathbb{P}(C_i = 0) = \mathbb{P}(P_i = 0)(1-\pi)$. Thus, the log-likelihood $\ell_C(\pi) = \log \mathbb{P}(C_1, \ldots, C_N; \pi)$ of the observed somatic mutation counts $\boldsymbol{c}$ for a gene is given by

$$\ell_C(\pi) = \log \mathbb{P}(C_1 = c_1, C_2 = c_2, \ldots, C_N = c_N; \pi) = \sum_{i=1}^{N} \log \big(\mathbb{P}(P_i = c_i)(1-\pi) + \mathbb{P}(P_i = c_i - 1)\pi\big). \tag{3}$$

Given observed mutation counts $\boldsymbol{c}$ and BMR distributions $\mathbb{P}(P_1), \ldots, \mathbb{P}(P_N)$, we compute the driver mutation rate π that maximizes the log-likelihood $\ell_C(\pi)$ of the observed data:

$$\widehat{\pi} = \underset{\pi \in [0,1]}{\operatorname{argmax}}\, \ell_C(\pi) = \underset{\pi \in [0,1]}{\operatorname{argmax}} \sum_{i=1}^{N} \log \left(\mathbb{P}(P_i = c_i)(1-\pi) + \mathbb{P}(P_i = c_i - 1)\pi\right). \tag{4}$$

The maximum likelihood problem (4) is challenging to solve exactly as it is often a *non-convex* optimization problem, depending on the form of the background distributions $\mathbb{P}(P_i)$. We solve this optimization problem by making the observation that the mutation count distribution (2) may be viewed as a *latent variable model*, where the unobserved, binary driver mutations D_i are the *latent variables* and the somatic mutation counts C_i are distributed according to a mixture of two distributions, $\mathbb{P}(P_i)$ and $\mathbb{P}(P_i - 1)$.

The standard approach for computing an MLE for a latent variable model is the *expectation maximization (EM)* algorithm [6]. Thus, we solve (4) using the EM algorithm, whose steps we describe below.

E-step. Given an estimated driver mutation rate $\pi^{(t)}$ at iteration t, we compute the *responsibility* $z_i^t = \mathbb{P}(D_i \mid C_i = c_i; \pi^{(t)})$, i.e. the probability of the latent variable $D_i = 1$ being equal to 1 conditioned on the observed mutation count C_i, for each sample $i = 1, \ldots, N$ as

$$\begin{aligned} z_i^{(t)} &= \mathbb{P}(D_i = 1 \mid C_i = c_i; \pi^{(t)}) \\ &= \frac{\mathbb{P}(D_i = 1; \pi^{(t)}) \cdot \mathbb{P}(C_i = c_i \mid D_i = 1; \pi^{(t)})}{\mathbb{P}(D_i = 1; \pi^{(t)}) \cdot \mathbb{P}(C_i = c_i \mid D_i = 1; \pi^{(t)}) + \mathbb{P}(D_i = 0; \pi^{(t)}) \cdot \mathbb{P}(C_i = c_i \mid D_i = 0; \pi^{(t)})} \\ &= \frac{\pi^{(t)} \cdot \mathbb{P}(P_i = c_i - 1)}{\pi^{(t)} \cdot \mathbb{P}(P_i = c_i - 1) + (1 - \pi^{(t)}) \cdot \mathbb{P}(P_i = c_i)}. \end{aligned} \tag{5}$$

M-step. Given the responsibility $z_i^{(t)}$ for each sample i, we estimate the driver mutation rate $\pi^{(t+1)}$ for iteration $t+1$ as

$$\pi^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} z_i^{(t)}. \tag{6}$$

2.2 Driver distribution for a pair of loci

We next extend the approach presented above to estimate the distribution $\mathbb{P}(D_i, D_i')$ of a *pair* of driver mutations. We start by observing that the driver mutations $(D_i, D_i') \in \{0,1\}^2$ are distributed according to a *bivariate* Bernoulli distribution. A bivariate Bernoulli distribution is specified by four parameters [17]:

1. the probability $\tau_{00} = \mathbb{P}(D_i = 0, D_i' = 0)$ that neither locus has a driver mutation;
2. the probability $\tau_{10} = \mathbb{P}(D_i = 1, D_i' = 0)$ that first locus has a driver mutation;
3. the probability $\tau_{01} = \mathbb{P}(D_i = 0, D_i' = 1)$ that the second locus has a driver mutation; and
4. the probability $\tau_{11} = \mathbb{P}(D_i = 1, D_i' = 1)$ that both loci have driver mutations,

where one of the parameters is redundant since $\tau_{00} + \tau_{10} + \tau_{01} + \tau_{11} = 1$. We note that the bivariate Bernoulli distribution $\mathbb{P}(D_i, D_i')$ is equivalent to a *categorical* distribution on binary strings $00, 01, 10, 11$ with corresponding probabilities $\tau_{00}, \tau_{01}, \tau_{10}, \tau_{11}$.

The parameters $\tau = (\tau_{00}, \tau_{01}, \tau_{10}, \tau_{11})$ of the bivariate Bernoulli distribution $\mathbb{P}(D_i, D_i')$ describe whether there is a *statistical interaction* [71] between the driver mutation D_i in the first locus and the driver mutation $D_{i'}$ in the second locus. If $\tau_{11}\tau_{00} < \tau_{01}\tau_{10}$, then the driver mutations are more likely to be mutually exclusive across samples than not

(i.e. a *negative* interaction) while if $\tau_{11}\tau_{00} > \tau_{01}\tau_{10}$, then the driver mutations are more likely to co-occur across samples than not (i.e. a *positive* interaction). Driver mutations D_i and D'_i are independent (i.e. no interaction) if and only if $\tau_{11}\tau_{00} = \tau_{01}\tau_{10}$.

More concisely, the interaction between driver mutations is quantified by the *log-odds ratio* $L = \log\left(\frac{\tau_{01}\tau_{10}}{\tau_{00}\tau_{11}}\right)$, which has previously been previously used to measure ME and CO for binarized mutations [38, 60, 14, 58]. The sign $\mathrm{sgn}(\ell)$ of the log-odds ratio ℓ determines the type of interaction: a positive log-odds ratio $L > 0$ describes ME between the driver mutations D_i, D'_i while a negative log-odds ratio $L < 0$ describes CO.

Following a similar derivation as in the previous section, the distribution $\mathbb{P}(C_i, C'_i)$ of mutation counts is given by

$$\begin{aligned}\mathbb{P}(C_i = c_i, C'_i = c'_i) = \mathbb{P}(P_i = c_i, P'_i = c'_i)\tau_{00} + \mathbb{P}(P_i = c_i - 1, P'_i = c'_i)\tau_{10} \\ + \mathbb{P}(P_i = c_i, P'_i = c'_i - 1)\tau_{01} + \mathbb{P}(P_i = c_i - 1, P'_i = c'_i - 1)\tau_{11},\end{aligned} \tag{7}$$

and the log-likelihood $\ell_{C,C'}(\tau) = \mathbb{P}(C_1 = c_1, C'_1 = c'_1, \ldots, C_N = c_N, C'_N = c'_N; \tau)$ is equal to

$$\begin{aligned}\ell_{C,C'}(\tau) &= \log \mathbb{P}(C_1 = c_1, \ldots, C'_N = c'_N; \tau) \\ &= \sum_{i=1}^{N} \log\Big((\mathbb{P}(P_i = c_i)\mathbb{P}(P'_i = c'_i)\tau_{00} + \mathbb{P}(P_i = c_i - 1)\mathbb{P}(P'_i = c'_i)\tau_{10} \\ &\quad + \mathbb{P}(P_i = c_i)\mathbb{P}(P'_i = c'_i - 1)\tau_{01} + \mathbb{P}(P_i = c_i - 1)\mathbb{P}(P'_i = c'_i - 1)\tau_{11}\Big).\end{aligned} \tag{8}$$

Given observed mutation counts $\boldsymbol{c}, \boldsymbol{c}'$ for a pair of genes and passenger mutation distributions $\mathbb{P}(P_1), \ldots, \mathbb{P}(P'_N)$ across N tumor samples, we compute the parameters $\tau_{00}, \tau_{01}, \tau_{10}, \tau_{11}$ of the driver mutation distribution that maximize the log-likelihood of the observed data:

$$\begin{aligned}(\widehat{\tau}_{00}, \widehat{\tau}_{01}, \widehat{\tau}_{10}, \widehat{\tau}_{11}) = \underset{\tau_{00},\tau_{01},\tau_{10},\tau_{11}}{\operatorname{argmax}} \quad & \sum_{i=1}^{N} \log\big(\mathbb{P}(P_i = c_i)\mathbb{P}(P'_i = c'_i)\tau_{00} + \mathbb{P}(P_i = c_i - 1)\mathbb{P}(P'_i = c'_i)\tau_{10} \\ & \quad + \mathbb{P}(P_i = c_i)\mathbb{P}(P'_i = c'_i - 1)\tau_{01} + \mathbb{P}(P_i = c_i - 1)\mathbb{P}(P'_i = c'_i - 1)\tau_{11}\big) \\ \text{subject to} \quad & \tau_{00} + \tau_{01} + \tau_{10} + \tau_{11} = 1, \\ & 0 \le \tau_{00}, \tau_{01}, \tau_{10}, \tau_{11} \le 1.\end{aligned} \tag{9}$$

The maximum likelihood problem (9) is difficult to solve as, for many background distributions $\mathbb{P}(P_i)$, it a non-convex optimization problem over a three-dimensional simplex. Thus, similar to the previous section, we solve (9) using the EM algorithm, whose steps we briefly describe below.

E-step. Given the estimated driver mutation probabilities $\tau^{(t)} = \left(\tau_{00}^{(t)}, \tau_{01}^{(t)}, \tau_{10}^{(t)}, \tau_{11}^{(t)}\right)$ at iteration t, we compute the responsibility $z_{i,uv}^{(t)} = \mathbb{P}(D_i, D'_i \mid C_i = c_i, C'_i = c'_i; \tau^{(t)})$ for each driver mutation probability $\tau_{uv}^{(t)}$ and sample $i = 1, \ldots, N$ as

$$z_{i,uv}^{(t)} = \frac{\tau_{uv}^{(t)} \cdot \mathbb{P}(P_i = c_i - u) \cdot \mathbb{P}(P'_i = c'_i - v)}{\sum_{(x,y)\in\{0,1\}^2}\left(\tau_{xy}^{(t)} \cdot \mathbb{P}(P_i = c_i - x) \cdot \mathbb{P}(P'_i = c'_i - y)\right)} \tag{10}$$

M-step. Given the estimated responsibilities $z_i^{(t)} = \left(z_{i,00}^{(t)}, z_{i,01}^{(t)}, z_{i,10}^{(t)}, z_{i,11}^{(t)}\right)$ at iteration t, we compute the estimated driver mutation probabilities $\tau_{uv}^{(t+1)}$ at iteration $t+1$ as

$$\tau_{uv}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{i,uv}^{(t)}. \tag{11}$$

2.3 Testing for statistical significance

We test the null hypothesis H_0 that the driver mutations D_i, D'_i are independent against the alternative hypothesis H_1 that the driver mutations D_i, D'_i are not independent. We perform this test using the likelihood ratio test (LRT),

whose test statistic is equal to the following scalar multiple of the difference between the log-likelihoods under the null hypothesis H_0 and alternative hypothesis H_1:

$$\lambda = -2\left(\left(\ell_C(\widehat{\pi}) + \ell_{C'}(\widehat{\pi}')\right) - \ell_{C,C'}(\widehat{\tau})\right), \tag{12}$$

where $\widehat{\pi}, \widehat{\pi}'$ are the estimated driver mutation rates assuming that driver mutations are independent, which are computed by solving (4), and $\widehat{\tau} = (\widehat{\tau}_{00}, \widehat{\tau}_{01}, \widehat{\tau}_{10}, \widehat{\tau}_{11})$ are the estimated parameters of the driver mutation distribution $P(D_i, D_i')$ computed by solving (9). We compute a p-value assuming that the LRT statistic λ follows a χ^2-distribution with one degree of freedom, which holds asymptotically by Wilks' theorem [77]. We say a pair of genes has ME or CO driver mutations if the p-value is less than a threshold ϵ.

2.4 DIALECT

We implement the EM algorithm for the latent variable model described above in an algorithm called Driver Interactions and Latent Exclusivity or Co-occurrence in Tumors (DIALECT, Figure 1). Given a mutation count matrix C (Figure 1A) and estimated BMR distributions $\mathbb{P}(P_i), \mathbb{P}(P_i')$ for each gene (Figure 1D), DIALECT estimates the pairwise driver mutation parameters $\widehat{\tau}$ by solving (9) for each pair of genes, and estimates the individual driver mutation rates $\widehat{\pi}$ by solving (4) for each individual gene (Figure 1E-F). DIALECT identifies mutually exclusive (resp. co-occurring) pairs as those with p-value less than a threshold ϵ (see previous section) and with a positive log-odds ratio $L = \log\left(\frac{\widehat{\tau}_{10}\widehat{\tau}_{01}}{\widehat{\tau}_{00}\widehat{\tau}_{11}}\right) > 0$ (resp. negative log-odds ratio $L < 0$). We emphasize that the BMR distributions $\mathbb{P}(P_i)$ used by DIALECT may be estimated using one of several methods, e.g. [40, 75, 67].

3 Results

3.1 Simulations

We evaluated the ability of DIALECT to identify dependencies between mutations, including mutual exclusivity and co-occurrence, in simulated somatic mutation data.

Data. We simulated somatic mutation counts $(c_i)_{i=1}^N, (c_i')_{i=1}^N$ for a pair of genes with lengths l and l', respectively, in nucleotides following equation (1). The passenger mutation count P_i (resp. P_i') in sample i is drawn from a binomial distribution $\text{Binom}(l, \mu)$ (resp. $\text{Binom}(l', \mu')$) where μ (resp, μ') is a per-nucleotide mutation rate. Such binomial distributions are often used in background mutation rate (BMR) models [40]. We drew each driver mutation (D_i, D_i') from a bivariate Bernoulli distribution with parameters $\tau = (\tau_{00}, \tau_{01}, \tau_{10}, \tau_{11})$, where we choose the parameters τ to describe either mutual exclusivity or co-occurrence of driver mutations.

Mutual exclusivity. We first assessed DIALECT in identifying *mutually exclusive* driver mutations. We compared DIALECT with two approaches for identifying mutual exclusivity from binarized mutations: Fisher's exact test [22], a classical statistical test of independence; and MEGSA [31], a recent method for identifying mutually exclusive driver mutations.

We simulate somatic mutation counts $(C_i)_{i=1}^N, (C_i')_{i=1}^N$ across $N = 1000$ samples with the following parameter choices. The driver mutation distribution $\mathbb{P}(D_i, D_i')$ has parameters $\tau_{11} = 0$, i.e. no co-occurrence between drivers, and $\tau_{01} = \tau_{10} = \tau$, where τ represents the rate of mutual exclusivity between driver mutations. To specify the passenger count distributions, we use gene lengths $l = l' = 10000$ and we use nucleotide mutation rate $\mu = 10^{-6}$ for the first gene, which was chosen so that the probability $\mathbb{P}(P_i > 0) \approx 0.01$ of this gene having more than one passenger mutation matches the median probability $\mathbb{P}(P_i > 0)$ across all genes in real data. In order to model how power varies with the presence of passenger mutations, we vary the nucleotide mutation rate μ' of the second gene such that that the BMR probability $\mathbb{P}(P_i' > 0)$, or the probability of the second gene having more than one passenger mutation, varies between 0.01 and 0.10. We assume there are no hypermutated samples, i.e. samples i with mutation factor $s_i > 1$.

We run DIALECT with the true BMR distributions $\mathbb{P}(P_i), \mathbb{P}(P_i')$ for each sample $i = 1, \ldots, N$. Since the power and specificity improves with an increasing number N of samples, we choose the p-value threshold ϵ based on the number N of samples: if $N \geq 1000$ then we set the p-value threshold to be $\epsilon = 0.05$, while if $N < 1000$ then we

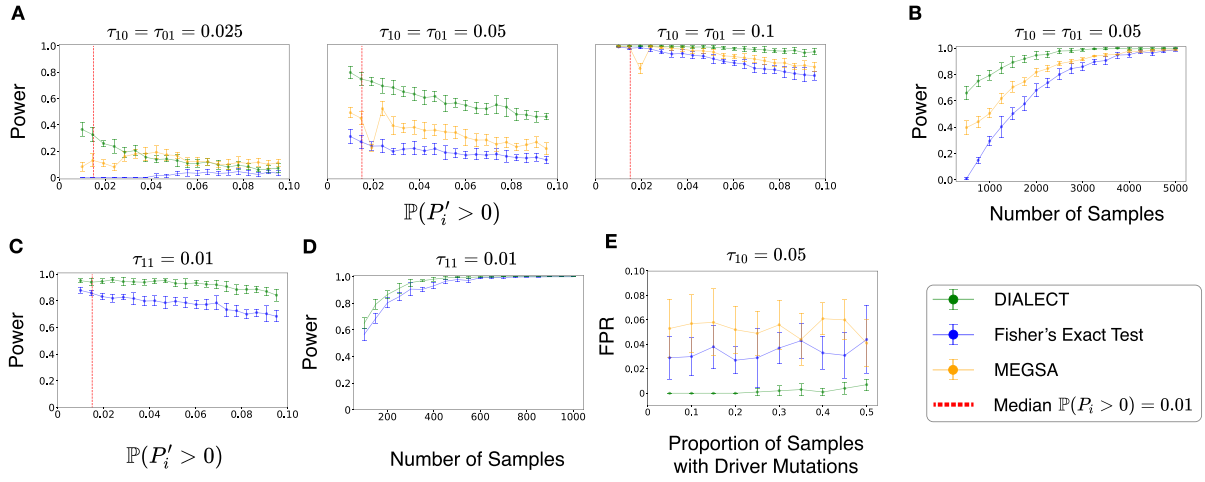

Figure 2: Statistical power and false positive rate for detecting dependencies between driver mutations in simulated data. **(A)** Power (sensitivity) of DIALECT, Fisher's exact test, and MEGSA for identifying mutually exclusive driver mutations from $N = 1000$ tumor samples, for different choices of the rate τ of mutual exclusivity of driver mutations and different probabilities $\mathbb{P}(P_i' > 0)$ of a gene having passenger mutations. Dashed red line indicates median estimated passenger mutation probability across all genes. **(B)** Power of DIALECT, Fisher's exact test, and MEGSA versus number N of samples, which we vary from 100 to 5000, in detecting mutually exclusive driver mutations. **(C)** Power (sensitivity) of DIALECT and Fisher's exact test for identifying co-occurring driver mutations with co-occurrence rate $\tau_{11} = 0.01$ from $N = 300$ tumor samples, for different probability $\mathbb{P}(P_i' > 0)$ of having passenger mutations. **(D)** Power of DIALECT and Fisher's exact test versus number N of samples in detecting co-occurring driver mutations. **(E)** False positive rate versus percentage of samples with driver mutations for $\tau_{10} = 0.05$ across $N = 1000$ samples.

set the p-value threshold to $\epsilon = 0.001$. For Fisher's exact test, a gene pair was identified as mutually exclusive if the resulting p-value was less than 0.05. For MEGSA, a gene pair is identified as mutually exclusive if the MEGSA p-value, i.e. the MEGSA LRT statistic under the χ^2-distribution, is less than 0.10.

We observe (Figure 2A) that DIALECT has greater power compared to Fisher's exact test and MEGSA across a range of driver mutual exclusivity rates τ and BMR probabilities $\mathbb{P}(P_i' > 0)$. In particular, DIALECT has substantially larger power than Fisher's exact test and MEGSA when the gene pairs have small rates τ of mutually exclusivity ($\tau \leq 0.05$) and there are a small number of passenger mutations ($\mathbb{P}(P_i' > 0) \leq 0.01$) — parameters which describe many pairs of driver genes in real data. For these parameter choices, we also performed a *power analysis* and assessed the number of samples needed to achieve a given statistical power. We found (Figure 2B) that $N > 1000$ samples are needed for DIALECT to achieve power > 0.75, while $N > 2500$ samples are needed for Fisher's exact test and MEGSA to achieve the same power. We emphasize that most large cohort studies only measure $N = 100 - 1000$ samples, meaning that DIALECT, as well as existing approaches like Fisher's exact test, may not have sufficient power to detect gene pairs with small rates τ of mutual exclusivity. Nevertheless, our simulations demonstrate that for sufficiently large cohort sizes, DIALECT more accurately identifies pairs of mutually exclusive driver mutations compared to standard approaches.

Co-occurrence. We next evaluated DIALECT in identifying *co-occurring* driver mutations. We compared DIALECT with Fisher's exact test [22] which tests for co-occurrence in binarized mutations between a pair of genes. We do not compare to MEGSA as it only identifies genes with mutually exclusive mutations. We simulated somatic mutation counts $(C_i)_{i=1}^N, (C_i')_{i=1}^N$ for $N = 300$ tumor samples where (1) the passenger mutation count distributions $\mathbb{P}(P_i), \mathbb{P}(P_i')$ are distributed as previously described and (2) the driver mutation distribution $\mathbb{P}(D_i, D_i')$ has parameters $\tau_{11} = 0.01$ and $\tau_{01} = \tau_{10} = 0$.

We observe that DIALECT has greater power compared to Fisher's exact test across a range of BMR probabilities $\mathbb{P}(P_i' > 0)$ (Figure 2C) and number N of samples (Figure 2D). We emphasize that a much smaller number N of samples are needed to achieve a power of 1 for identifying co-occurring mutations ($N \approx 600$, Figure 2D) compared

to identifying mutually exclusive mutations ($N \approx 5000$, Figure 2B), reflecting that co-occurrence is easier to detect than mutual exclusivity. This analysis demonstrates that for small cohort sizes, DIALECT more accurately identifies co-occurring driver mutations than existing approaches.

False positive rate. We assessed the false positive rate (FPR, i.e. $1-$specificity) of DIALECT and other methods by simulating somatic mutations for a driver gene (i.e. a gene with driver mutations, i.e. $D_i = 1$ for some samples i) and a passenger gene with no driver mutations (i.e. $D'_i = 0$) and a large number P_i of passenger mutations. Following the simulation set-up described previously, we set the passenger mutation distribution parameters as $l = 10000$, $\mu = 10^{-6}$ for the driver gene and $l' = 100000$ and $\mu' = 10^{-5}$ for the passenger mutation. The distribution $P(D_i, D'_i)$ of driver mutations has parameters $\tau_{11} = \tau_{01} = 0$, and $\tau_{10} = \pi$, where π represents the driver mutation rate for the driver gene. Furthermore, in this simulation we assume driver mutations are not identically distributed across samples; instead, we draw driver mutations D_i, D'_i for a ρ fraction of all N samples selected uniformly at random, where we vary ρ between 0.05 and 0.5, and set $D_i = D'_i = 0$ for the other $(1-\rho)N$ samples.

We find (Figure 2E) that DIALECT consistently exhibits lower FPR (i.e. higher specificity) than the existing methods across different proportions ρ of samples with driver mutations. In particular, DIALECT achieves FPR close to zero when $\rho < 0.4$, which is larger than the mutation rate of nearly all driver genes, while Fisher's exact test and MEGSA have FPR above 0.02. We emphasize that even relatively small FPRs result in the inference of many spurious dependencies in real data analyses. For example, using an algorithm with FPR $= 0.01$ – which is lower than the FPRs of Fisher's exact test and MEGSA but larger than DIALECT's FPR – to identify dependencies between all pairs of $G = 100$ genes will result in $0.01 \cdot \binom{G}{2} \approx 50$ spurious dependencies. We also emphasize that these results show that DIALECT is robust to model mis-specification, since DIALECT assumes driver mutations are identically distributed across tumor samples while our simulated driver mutations are not identically distributed. Such behavior is hypothesized to occur in some cancer types; for example, [70] observed that certain driver mutations are more likely to occur in colorectal cancer subtypes with lower overall mutation loads.

3.2 Analysis of mutations in TCGA

We next evaluated DIALECT using somatic mutation data from The Cancer Genome Atlas (TCGA) [76]. We used DIALECT to identify mutual exclusivity, as mutual exclusivity between driver mutations is observed more often than co-occurrence [10, 43]. We compared DIALECT to two state-of-the-art statistical tests for identifying mutual exclusivity: Fisher's exact test [22] and DISCOVER [10]. Fisher's exact test implicitly assumes that each sample is identically distributed, while DISCOVER performs a statistical test where genes have different, sample-specific mutation rates (the DISCOVER test is also asymptotically equivalent to the test used by [42]). However, both Fisher's exact test and DISCOVER use *binarized* mutations as input, and thus do not distinguish between driver mutations and passenger mutations. Since DIALECT analyzes missense mutations and nonsense mutations in a gene separately (since these mutation types often have different background mutation rates), we additionally ran DISCOVER with somatic counts separated into gene events including only nonsynonymous missense mutations (indicated by *GENE_M*) and only nonsense mutations (indicated by *GENE_N*). We denote these results using DISCOVER*. For DISCOVER and DISCOVER* (resp. Fisher's exact test), a gene pair was identified as mutually exclusive if the resulting q-value (resp. p-value) was less than 0.05.

Data. We analyzed non-synonymous mutations from tumor samples in 5 different cancer types from TCGA. Each cancer type contains 100-1000 tumor samples. We obtained the somatic mutation data in Mutation Annotation Format (MAF) from the TCGA PanCancer project, available through cBioPortal [24]. We separately analyzed missense and nonsense mutations, appending gene names with $_M$ for missense mutations and $_N$ for nonsense mutations, and we excluded mutations classified as 'Silent', 'Intron', '3' UTR', '5' UTR', 'IGR', 'lincRNA', and 'RNA'. For computational efficiency, we restricted our analysis to the 500 most frequently mutated genes across samples – a criterion that is typically used in other mutual exclusivity analyses – yielding a total of $124,750$ gene pairs that we analyze. We obtained background mutation rate distributions $\mathbb{P}(P_i)$ for each gene and mutation type (missense, nonsense) using CBaSE [V1.2] [75]. We emphasize that DIALECT could also be run with other methods for estimating background mutation rate distributions such as MutSigCV2 [40] or Dig [67].

Mutual exclusivity. DIALECT identified between 5 and 14 gene pairs in each of the five different cancer types. In contrast, DISCOVER, DISCOVER*, and Fisher's exact test reported a higher number of pairs across all cancer

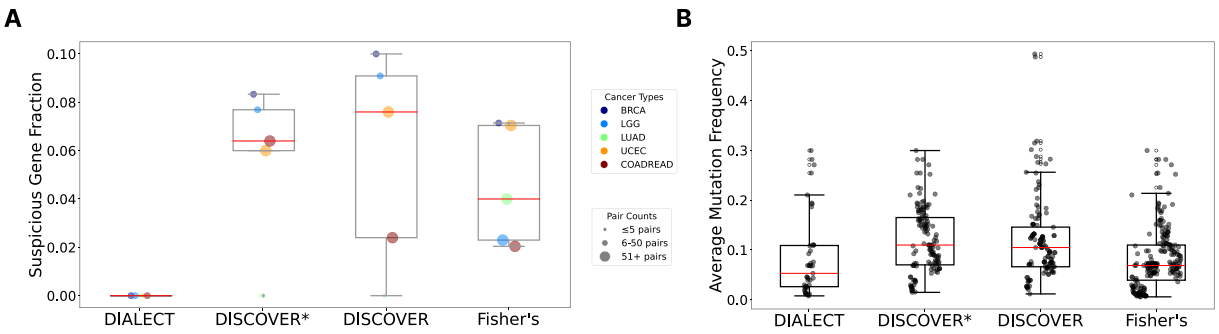

Figure 3: Comparison of pairs of genes identified by DIALECT, DISCOVER, and Fisher's exact test for 5 cancer subtypes in The Cancer Genome Atlas (TCGA). (A) Suspicious gene fractions, or the fraction of gene pairs where at least one gene is in a list of "suspicious" genes that are likely not driver genes, as annotated in [40], for DIALECT, DISCOVER, DISCOVER*, and Fisher's exact test. DISCOVER* is a variant of DISCOVER that is run separately on missense and nonsense mutations, similar to DIALECT. We select all gene pairs with q-value less than 0.05 for DISCOVER, DISCOVER*, and Fisher's exact test. **(B)** The average mutation frequency of the two genes in each gene pair identified by DIALECT, DISCOVER, DISCOVER*, and Fisher's exact test.

subtypes, including over 300 pairs for colon adenocarcinoma and rectum adenocarcinoma (COADREAD) and uterine corpus endometrial carcinoma (UCEC). This pattern suggests that these methods may be prone to identifying interactions between genes with high numbers of mutations, many of which are likely passengers. Thus, for each method, we next evaluated the fraction of "suspicious" genes, or genes that are likely not driver genes as annotated by [40], in the mutually exclusive pairs identified by each method. Such suspicious genes have high numbers of passenger mutations, and are commonly identified or removed from the analyses by existing mutual exclusivity methods. We find that DIALECT does not identify pairs with suspicious genes, while 5-10% of the pairs identified by DISCOVER, DISCOVER*, and Fisher's exact test contain suspicious genes (Figure 3A). As another assessment, we find that DIALECT identifies gene pairs with lower average mutation frequencies compared to gene pairs identified by DISCOVER, DISCOVER*, and Fisher's exact test (Figure 3B). Genes with high mutation frequencies are often falsely identified by other methods, and contribute to the larger number of gene pairs identified by these methods. These analyses indicate that DIALECT does not identify mutual exclusivity between likely passenger genes with large numbers of mutations, in contrast DISCOVER, DISCOVER*, and Fisher's exact test which often identify suspicious or highly mutated genes.

Focusing on breast cancer, the largest cohort in the dataset with $N = 1084$ patients, we observed (Table 1) that the gene pairs with the highest rates of mutual exclusivity, i.e. the pairs with largest log-odds estimated by DIALECT, are comprised of genes that are reported as drivers in breast cancer. Pairs such as CDH1_N:TP53_M (DIALECT p-value $= 0.002$) and AKT1_M:PIK3CA_M (DIALECT p-value $= 0.015$) have been found to reflect distinct functional modules within breast cancer, e.g. *TP53*, *CDH1*, *AKT1*, and *PIK3CA* are all known breast cancer driver genes [57, 37, 62].

In contrast, DISCOVER* and Fisher's Exact Test identify spurious pairs that contain at least one "suspicious" gene. In particular, both DISCOVER* and Fisher's exact test identify the pair AKT1_M:TTN_M. *TTN* has many random passenger mutations due to its extraordinary length and likely does not contain any driver mutations [39, 40]. The identification of the suspicious gene *TTN* by Fisher's exact test agrees with its low specificity as we demonstrated in simulations (Figure 2E).

DISCOVER and DISCOVER* are particularly prone to identifying interactions between genes with high mutation rates, an issue exacerbated in types like COADREAD and UCEC which exhibit higher background mutation rates. In particular, COADREAD and UCEC samples typically exhibit a higher number of mutated genes per sample (median of 78.5 genes per sample for COADREAD and 57.5 genes per sample for UCEC) [42]. DISCOVER and DISCOVER* report over 500 significant pairs in COADREAD and over 1000 pairs in UCEC. In contrast, DIALECT identifies a far more selective 8 and 5 mutually exclusive pairs for COADREAD (Table S2) and UCEC (Table S3), respectively.

DIALECT also identifies novel mutual exclusivity between driver mutations that were not identified by existing methods. In particular, DIALECT identifies mutual exclusivity between STAB2_M:TP53_M. This pair was not identified by DISCOVER* or Fisher's exact test (Figure 4, Table 1) due to the low mutation rate of *STAB2*. *STAB2* overexpression has been observed to cause increased tumor metastasis rates [29] and poor tumor prognosis [79],

and may explain the observed mutual exclusivity between missense mutations in *TP53* and *STAB2*. These examples demonstrate how by modeling driver and passenger mutations separately, DIALECT is able to identify novel driver mutations and mutual exclusivity relations that are missed by current approaches.

DIALECT		DISCOVER*		Fisher's Exact Test	
Pair	LLR	Pair	q-value	Pair	p-value
CDH1_N:TP53_M	14.728	PIK3CA_M:TP53_M	$4.45 * 10^{-7}$	CDH1_N:TP53_M	$7.46 * 10^{-4}$
TP53_M:TP53_N	12.132	TP53_M:TP53_N	$9.57 * 10^{-6}$	PIK3CA_M:TP53_M	$1.08 * 10^{-3}$
PIK3CA_M:TP53_N	11.153	CDH1_N:TP53_M	$2.13 * 10^{-5}$	TP53_M:TP53_N	$1.39 * 10^{-3}$
AKT1_M:PIK3CA_M	10.463	PIK3CA_M:TP53_N	$4.98 * 10^{-5}$	PIK3CA_M:TP53_N	$1.56 * 10^{-3}$
PIK3CA_M:TP53_M	9.933	AKT1_M:PIK3CA_M	$4.44 * 10^{-4}$	AKT1_M:PIK3CA_M	$1.84 * 10^{-3}$
MAP3K1_N:TP53_M	8.877	MAP3K1_M:TP53_M	$3.54 * 10^{-3}$	MAP3K1_N:TP53_M	$1.08 * 10^{-2}$
NCOR1_N:TP53_M	7.049	MAP3K1_N:TP53_M	$5.24 * 10^{-3}$	MAP3K1_M:TP53_M	$1.61 * 10^{-2}$
ARID1A_N:TP53_M	6.239	FOXA1_M:TP53_M	$6.88 * 10^{-3}$	FOXA1_M:TP53_M	$2.43 * 10^{-2}$
FOXA1_M:TP53_M	5.813	AKT1_M:**TTN_M**	$1.01 * 10^{-2}$	NCOR1_N:TP53_M	$2.82 * 10^{-2}$
MYH9_M:TP53_M	4.750	MYH9_M:TP53_M	$1.92 * 10^{-2}$	CBFB_M:TP53_M	$3.58 * 10^{-2}$
MAP3K1_M:TP53_M	4.728	NCOR1_N:TP53_M	$3.78 * 10^{-2}$	MYH9_M:TP53_M	$3.66 * 10^{-2}$
CBFB_M:TP53_M	3.898	AHNAK2_M:TP53_M‡	$4.44 * 10^{-2}$	AKT1_M:**TTN_M**	$4.34 * 10^{-2}$
STAB2_M:TP53_M‡	3.676			GREB1L_M:TP53_M‡	$4.55 * 10^{-2}$
AKT1_M:TP53_N	3.519			ARID1A_N:TP53_M	$4.55 * 10^{-2}$

Table 1: Mutually exclusive pairs of mutations identified by DIALECT, DISCOVER*, and Fisher's Exact Test on TCGA breast cancer (BRCA) data. Higher LLR, lower q-values, and lower p-values indicate stronger mutual exclusivity. Suspicious genes are shown in bold. Pairs uniquely identified by a method are shown with ‡.

4 Discussion

We introduce DIALECT, a method for identifying dependencies between pairs of *driver* mutations from somatic mutations counts. DIALECT explicitly models the observed somatic mutation counts as a sum of driver mutations and passenger mutations, in contrast to nearly all other methods which conflate drivers with passengers in a gene by *binarize* the mutation events in a gene. DIALECT models the distribution of driver mutations using a latent variable model while accounting for passenger mutations by incorporating existing background mutation rate (BMR) models. We derive an expectation maximization (EM) algorithm to estimate the parameters of our model which describe the degree of mutual exclusivity or co-occurrence between driver mutations. We demonstrate that DIALECT has improved performance compared to the standard mutual exclusivity and co-occurrence tests on simulated and real data.

Our approach for jointly modeling passenger and driver mutations can be readily extended in several directions. First, there are many methods for modeling BMRs, with each method having different strengths and weaknesses. In large-scale cancer studies, a standard practice is to form a "consensus" list of driver genes using BMRs estimated by different methods. Likewise, we imagine that it would be beneficial to run DIALECT with different BMR models in order to form a consensus list of mutually exclusive driver mutations. Second, although DIALECT allows for sample-specific BMRs (as demonstrated in simulations), existing tools do not readily output sample-specific BMRs for real data. Thus it would be useful to evaluate DIALECT using accurate sample-specific BMRs on a large-scale cohort. Similarly, DIALECT assumes that each tumor sample has an equal probability of a driver mutation, and we show in simulations that DIALECT has large power even when this assumption does not hold (i.e. when there is *model mis-specification*). Nevertheless, it may be useful to derive a more general model that incorporates sample-specific driver probabilities. Third, in the present work we used DIALECT to identify mutual exclusivity between driver mutations in real data, which provides a signal that the driver mutations perturb different biological pathways. Preliminary analysis suggests that there is no statistically significant co-occurrence in the TCGA data consistent with previous studies [10], but further analysis of this issue is necessary. Finally, we believe that our novel approach for separately modeling driver and passenger mutations would be advantageous for other problems in cancer genomics, particularly for learning cancer progression models (CPMs) which describe patterns in driver mutation accumulation over time [46, 64, 19, 1, 11, 54, 66, 47, 33].

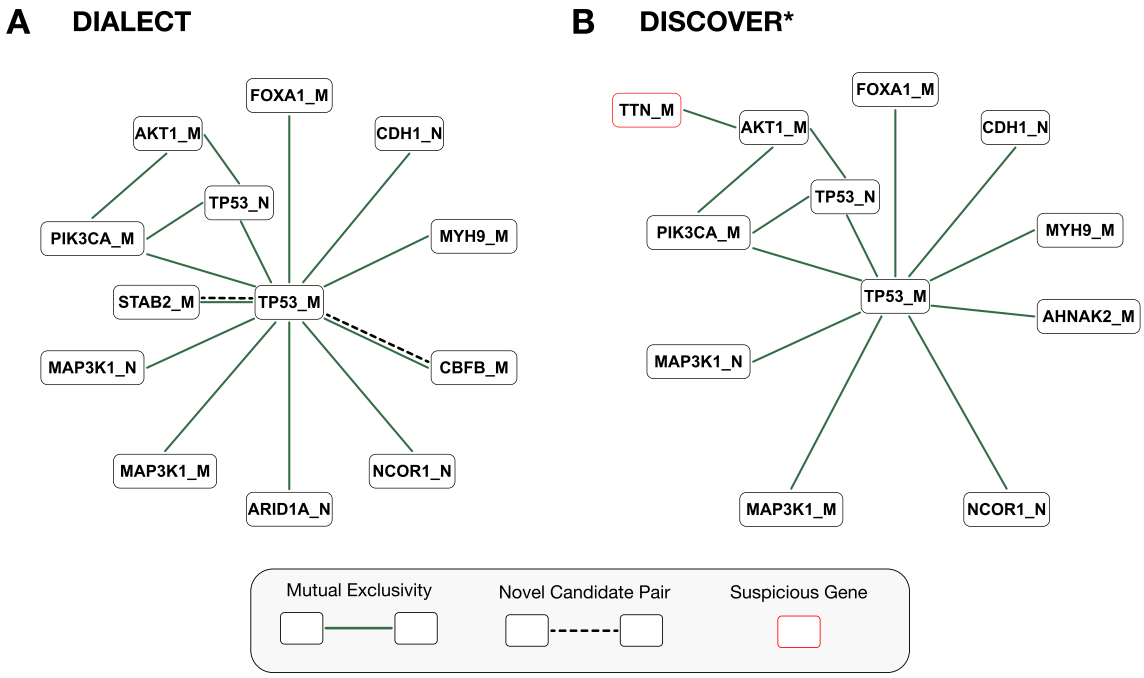

Figure 4: Mutually exclusive pairs of genes detected by DIALECT and DISCOVER* in breast cancer (BRCA). (A) Network of mutually exclusive gene pairs identified by DIALECT, where nodes represent genes, solid edges indicate mutual exclusivity between driver mutations, and dashed edges indicate novel gene pairs not identified in prior literature. **(B)** Network of mutually exclusive gene pairs identified by DISCOVER*. Red highlighted node indicates "suspicious" gene as annotated by [40].

5 Acknowledgments

This research is supported by NIH/NCI grants U24CA248453 and U24CA264027 to B.J.R. U.C. was supported by NSF GRFP DGE 2039656 and the Siebel Scholars program. We thank Donate Weghorn for modifying CBaSE to output sample-specific background mutation distributions, and we thank Madelyne Xiao for work on a previous iteration of the model.

References

[1] F. Angaroni, K. Chen, C. Damiani, G. Caravagna, A. Graudenzi, and D. Ramazzotti. Pmce: efficient inference of expressive models of cancer evolution with high prognostic power. *Bioinformatics*, 38(3):754–762, 2022.

[2] Ö. Babur, M. Gönen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome biology*, 16:1–10, 2015.

[3] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[4] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah. Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, 13:1–14, 2012.

[5] A. J. Bass, V. Thorsson, I. Shmulevich, S. M. Reynolds, M. Miller, et al. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209, 2014.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[7] J. L. Bos. The ras gene family and human carcinogenesis. *Mutation Research/Reviews in Genetic Toxicology*, 195(3):255–271, 1988.

[8] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[9] C. W. Brennan, R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, 2013.

[10] S. Canisius, J. W. Martens, and L. F. Wessels. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome biology*, 17(1):1–17, 2016.

[11] G. Caravagna, A. Graudenzi, D. Ramazzotti, R. Sanz-Pamplona, L. De Sano, G. Mauri, V. Moreno, M. Antoniotti, and B. Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences*, 113(28):E4025–E4034, 2016.

[12] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–6667, 2009.

[13] H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, and R. Karchin. Identifying mendelian disease genes with the variant effect scoring tool. *BMC genomics*, 14:1–16, 2013.

[14] K. Chaudhary, O. B. Poirion, L. Lu, S. Huang, T. Ching, and L. X. Garmire. Multimodal meta-analysis of 1,494 hepatocellular carcinoma samples reveals significant impact of consensus driver genes on phenotypes. *Clinical Cancer Research*, 25(2):463–472, 2019.

[15] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

[16] S. Constantinescu, E. Szczurek, P. Mohammadi, J. Rahnenführer, and N. Beerenwinkel. Timex: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics*, 32(7):968–975, 2016.

[17] B. Dai, S. Ding, and G. Wahba. Multivariate bernoulli distribution. 2013.

[18] H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, et al. Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–954, 2002.

[19] L. De Sano, G. Caravagna, D. Ramazzotti, A. Graudenzi, G. Mauri, B. Mishra, and M. Antoniotti. Tronco: an r package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*, 32(12):1911–1913, 2016.

[20] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, et al. Music: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–1598, 2012.

[21] F. Dietlein, D. Weghorn, A. Taylor-Weiner, A. Richters, B. Reardon, D. Liu, E. S. Lander, E. M. Van Allen, and S. R. Sunyaev. Identification of cancer driver genes based on nucleotide context. *Nature genetics*, 52(2):208–218, 2020.

[22] R. A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, 85(1):87–94, 1922.

[23] J. Foo, L. L. Liu, K. Leder, M. Riester, Y. Iwasa, C. Lengauer, and F. Michor. An evolutionary approach for identifying driver mutations in colorectal cancer. *PLoS computational biology*, 11(9):e1004350, 2015.

[24] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269):pl1–pl1, 2013.

[25] L. A. Garraway. Genomics-driven oncology: framework for an emerging paradigm. *Journal of Clinical Oncology*, 31(15):1806–1814, 2013.

[26] A. Gonzalez-Perez and N. Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic acids research*, 40(21):e169–e169, 2012.

[27] Y. Han, J. Yang, X. Qian, W.-C. Cheng, S.-H. Liu, X. Hua, L. Zhou, Y. Yang, Q. Wu, P. Liu, et al. Driverml: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic acids research*, 47(8):e45–e45, 2019.

[28] D. Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022.

[29] Y. Hirose, E. Saijou, Y. Sugano, F. Takeshita, S. Nishimura, H. Nonaka, Y.-R. Chen, K. Sekine, T. Kido, T. Nakamura, et al. Inhibition of stabilin-2 elevates circulating hyaluronic acid levels and prevents tumor metastasis. *Proceedings of the National Academy of Sciences*, 109(11):4263–4268, 2012.

[30] J. P. Hou and J. Ma. Dawnrank: discovering personalized driver genes in cancer. *Genome medicine*, 6:1–16, 2014.

[31] X. Hua, P. L. Hyland, J. Huang, L. Song, B. Zhu, N. E. Caporaso, M. T. Landi, N. Chatterjee, and J. Shi. Megsa: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. *The American Journal of Human Genetics*, 98(3):442–455, 2016.

[32] T. J. C. Hudson, W. Anderson, A. Aretz, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[33] S. Ivanovic and M. El-Kebir. Modeling and predicting cancer clonal evolution with reinforcement learning. *Genome Research*, pages gr–277672, 2023.

[34] W. G. Kaelin Jr. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698, 2005.

[35] Y.-A. Kim, D.-Y. Cho, P. Dao, and T. M. Przytycka. Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics*, 31(12):i284–i292, 2015.

[36] Y.-A. Kim, S. Madan, and T. M. Przytycka. Wesme: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics*, 33(6):814–821, 2017.

[37] D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[38] J. Kuipers, A. L. Moore, K. Jahn, P. Schraml, F. Wang, K. Morita, P. A. Futreal, K. Takahashi, C. Beisel, H. Moch, et al. Statistical tests for intra-tumour clonal co-occurrence and exclusivity. *PLoS computational biology*, 17(12):e1009036, 2021.

[39] A. Laddach, M. Gautel, and F. Fraternali. Titindb—a computational tool to assess titin's role as a disease gene. *Bioinformatics*, 33(21):3482–3485, 2017.

[40] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[41] M. D. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS computational biology*, 9(5):e1003054, 2013.

[42] M. D. Leiserson, M. A. Reyna, and B. J. Raphael. A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics*, 32(17):i736–i745, 2016.

[43] M. D. Leiserson, H.-T. Wu, F. Vandin, and B. J. Raphael. Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology*, 16(1):1–20, 2015.

[44] T. Ley, C. Miller, L. Ding, B. Raphael, A. Mungall, A. Robertson, K. Hoadley, T. Triche Jr, P. Laird, J. Baty, et al. Cancer genome atlas research network genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22):2059–2074, 2013.

[45] S. Liu, J. Liu, Y. Xie, T. Zhai, E. W. Hinderer, A. J. Stromberg, N. L. Vanderford, J. M. Kolesar, H. N. Moseley, L. Chen, et al. Mescan: a powerful statistical framework for genome-scale mutual exclusivity analysis of cancer mutations. *Bioinformatics*, 37(9):1189–1197, 2021.

[46] L. O. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antoniotti, and B. Mishra. Inferring tree causal models of cancer progression with probability raising. *PloS one*, 9(10):e108358, 2014.

[47] X. G. Luo, J. Kuipers, and N. Beerenwinkel. Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees. *Nature Communications*, 14(1):3676, 2023.

[48] I. Martincorena and P. J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[49] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020.

[50] O. Martínez-Sáez, N. Chic, T. Pascual, B. Adamo, M. Vidal, B. González-Farré, E. Sanfeliu, F. Schettini, B. Conte, F. Brasó-Maristany, et al. Frequency and spectrum of pik3ca somatic mutations in breast cancer. *Breast Cancer Research*, 22(1):1–9, 2020.

[51] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[52] C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, and A. Milosavljevic. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC medical genomics*, 4:1–11, 2011.

[53] M. Mina, A. Iyer, and G. Ciriello. Epistasis and evolutionary dependencies in human cancers. *Current Opinion in Genetics Development*, 77:101989, 2022.

[54] M. Mohaghegh Neyshabouri, S.-H. Jun, and J. Lagergren. Inferring tumor progression in large datasets. *PLoS computational biology*, 16(10):e1008183, 2020.

[55] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas. Oncodrivefml: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*, 17:1–13, 2016.

[56] N. J. O'Neil, M. L. Bailey, and P. Hieter. Synthetic lethality and cancer. *Nature Reviews Genetics*, 18(10):613–623, 2017.

[57] D. Ostroverkhova, T. M. Przytycka, and A. R. Panchenko. Cancer driver mutations: predictions and reality. *Trends in Molecular Medicine*, 29(7):554–566, 2023.

[58] M. Ozcan, J. Janikovits, M. von Knebel Doeberitz, and M. Kloor. Complex pattern of immune evasion in msi colorectal cancer. *Oncoimmunology*, 7(7):e1445453, 2018.

[59] T. Y. Park, M. D. Leiserson, G. W. Klau, and B. J. Raphael. Superdendrix algorithm integrates genetic dependencies and genomic alterations across pathways and cancer types. *Cell genomics*, 2(2), 2022.

[60] B. Pereira, S.-F. Chin, O. M. Rueda, H.-K. M. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S.-J. Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*, 7(1):11479, 2016.

[61] A. Petitjean, M. Achatz, A. Borresen-Dale, P. Hainaut, and M. Olivier. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, 26(15):2157–2165, 2007.

[62] B. K. Rajendran and C.-X. Deng. Characterization of potential driver mutations involved in human breast cancer by computational approaches. *Oncotarget*, 8(30):50252, 2017.

[63] B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine*, 6(1):1–17, 2014.

[64] B. J. Raphael and F. Vandin. Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. *Journal of Computational Biology*, 22(6):510–527, 2015.

[65] S. A. Roberts and D. A. Gordenin. Hypermutation in human cancer genomes: footprints and mechanisms. *Nature Reviews Cancer*, 14(12):786–800, 2014.

[66] R. Schill, S. Solbrig, T. Wettig, and R. Spang. Modelling cancer progression using mutual hazard networks. *Bioinformatics*, 36(1):241–249, 2020.

[67] M. A. Sherman, A. U. Yaari, O. Priebe, F. Dietlein, P.-R. Loh, and B. Berger. Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nature Biotechnology*, 40(11):1634–1643, 2022.

[68] E. Szczurek and N. Beerenwinkel. Modeling mutual exclusivity of cancer mutations. *PLoS computational biology*, 10(3):e1003503, 2014.

[69] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, 2016.

[70] J. van de Haar, S. Canisius, K. Y. Michael, E. E. Voest, L. F. Wessels, and T. Ideker. Identifying epistasis in cancer genomes: a delicate affair. *Cell*, 177(6):1375–1383, 2019.

[71] T. J. VanderWeele and M. J. Knol. A tutorial on interaction. *Epidemiologic methods*, 3(1):33–72, 2014.

[72] F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–385, 2012.

[73] I. Varela, P. Tarpey, K. Raine, D. Huang, C. K. Ong, P. Stephens, H. Davies, D. Jones, M.-L. Lin, J. Teague, et al. Exome sequencing identifies frequent mutation of the swi/snf complex gene pbrm1 in renal carcinoma. *Nature*, 469(7331):539–542, 2011.

[74] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr, and K. W. Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[75] D. Weghorn and S. Sunyaev. Bayesian inference of negative and positive selection in human cancers. *Nature genetics*, 49(12):1785–1788, 2017.

[76] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[77] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

[78] C.-H. Yeang, F. McCormick, and A. Levine. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB journal*, 22(8):2605–2622, 2008.

[79] J. Yong, L. Huang, G. Chen, X. Luo, H. Chen, and L. Wang. High expression of stabilin-2 predicts poor prognosis in non-small-cell lung cancer. *Bioengineered*, 12(1):3426–3433, 2021.

[80] N. Zahir, R. Sun, D. Gallahan, R. A. Gatenby, and C. Curtis. Characterizing the ecological and evolutionary dynamics of cancer. *Nature genetics*, 52(8):759–767, 2020.

[81] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, et al. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, 2011:bar026, 2011.