Vandin *et al. Algorithms for Molecular Biology* 2012, **7**:23
http://www.almob.org/content/7/1/23

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Finding driver pathways in cancer: models and algorithms

Fabio Vandin*, Eli Upfal and Benjamin J Raphael

Abstract

Background: Cancer sequencing projects are now measuring somatic mutations in large numbers of cancer genomes. A key challenge in interpreting these data is to distinguish *driver mutations*, mutations important for cancer development, from *passenger* mutations that have accumulated in somatic cells but without functional consequences. A common approach to identify genes harboring driver mutations is a *single gene test* that identifies individual genes that are recurrently mutated in a significant number of cancer genomes. However, the power of this test is reduced by: (1) the necessity of estimating the *background mutation rate* (BMR) for each gene; (2) the mutational heterogeneity in most cancers meaning that groups of genes (e.g. pathways), rather than single genes, are the primary target of mutations.

Results: We investigate the problem of discovering *driver pathways*, groups of genes containing driver mutations, directly from cancer mutation data and without prior knowledge of pathways or other interactions between genes. We introduce two generative models of somatic mutations in cancer and study the algorithmic complexity of discovering driver pathways in both models. We show that a single gene test for driver genes is highly sensitive to the estimate of the BMR. In contrast, we show that an algorithmic approach that maximizes a straightforward measure of the mutational properties of a driver pathway successfully discovers these groups of genes without an estimate of the BMR. Moreover, this approach is also successful in the case when the observed frequencies of passenger and driver mutations are indistinguishable, a situation where single gene tests fail.

Conclusions: Accurate estimation of the BMR is a challenging task. Thus, methods that do not require an estimate of the BMR, such as the ones we provide here, can give increased power for the discovery of driver genes.

Keywords: Cancer, Somatic Mutations, Driver mutations, Pathways, Background mutation rate, Generative models

Background

Cancer is a disease driven in part by somatic mutations that accumulate during the lifetime of an individual. These mutations include single nucleotide substitutions, small indels, and larger copy number aberrations and structural aberrations. A key challenge in cancer genomics is to distinguish *driver mutations*, mutations important for cancer development, from random *passenger mutations* that have accumulated in somatic cells but do not have functional consequences. Recent advances in DNA sequencing technologies allow the measurement of somatic mutations in large numbers of cancer genomes. Thus, a common approach to identify driver mutations, and the driver genes in which they reside, is to identify genes with recurrent mutations in a large cohort of cancer patients. The standard technique to identify such recurrently mutated genes is to perform a *single gene test*, in which individual genes are tested to determine if their observed frequency of mutation is significantly higher than expected [1-3]. This approach has identified a number of important cancer genes, but has not revealed all of the driver mutations and driver genes in individual cancers.

There are two difficulties with the identification of driver genes by a single gene test of recurrent mutation. First, the test requires a reasonable estimate of the *background mutation rate* (BMR) for each gene, or the rate at which passenger mutations occur in the gene. Obtaining such an estimate is not a straightforward task, as the BMR is not just the rate of somatic mutation per

*Correspondence: vandinfa@cs.brown.edu
Department of Computer Science, and Center for Computational Molecular Biology Brown University, 115 Waterman St., 4th Flr, Providence, RI 02912, USA

nucleotide per cell generation, but also must account for selection and clonal amplification in the somatic evolution of a tumor [1,4]. Second, it is widely observed that there is extensive mutational heterogeneity in cancer, with mutations occurring in different genes in different patients. This mutational heterogeneity is a consequence of both the presence of passenger mutations in each cancer genome, and the fact that driver mutations typically target genes in cellular signaling and regulatory pathways [5,6]. Since each of these pathways contains multiple genes, there are numerous combinations of driver mutations that can perturb a pathway important for cancer. This mutational heterogeneity inflates the number of patients required to distinguish passenger from driver mutations, as rare driver mutations may not be observed at frequencies above the background. An alternative to single gene tests is to test the recurrence of mutations in groups of genes derived from known pathways [7,8] or genome-scale gene interaction networks [9,10]. However, these approaches require prior knowledge of the interactions between genes/proteins, and this knowledge is presently far from complete. Moreover, pathway/network based approaches typically also require an estimate of the BMR.

The availability of somatic mutation data from increasing numbers of cancer patients motivates the question of whether it is possible to identify *driver pathways*, groups of genes with recurrent driver mutations, *de novo*; i.e. without prior knowledge of interactions between genes/proteins. At first glance, this seems implausible because there are an enormous number of possible sets of genes to consider. For example, there are more than 10^{25} sets of 7 human genes. However, we previously showed that mild additional constraints on the expected patterns of somatic mutations considerably reduce the number of gene sets to examine, and make *de novo* discovery of driver pathways possible [11]. These constraints are consistent with the current understanding of the somatic mutational process of cancer [6,12]. In particular, we assume that an important cancer pathway should be perturbed in a large number of patients. Thus, given genome-wide measurements of somatic mutations, we expect that a driver pathway will have high *coverage*: i.e. most patients will have a mutation in some gene in the pathway. Second, a driver mutation in a single gene of the pathway is often assumed to be sufficient to perturb the pathway. Combined with the fact that driver mutations are relatively rare, most patients exhibit only a single driver mutation in a pathway. Thus, we expect that the genes in a pathway exhibit a pattern of *mutually exclusive* driver mutations, where driver mutations are observed in exactly one gene in the pathway in each patient [13].

We emphasize that our assumption of mutual exclusivity holds only for driver mutations in the *same* pathway. It is well known that cancer genomes harbor driver mutations in multiple pathways, and the exclusivity assumption does not preclude the presence of such co-occurring, and possibly cooperative, driver mutations, examples of which are known [14,15]. Indeed, current estimates of the number of driver mutations and number of mutated pathways in a cancer genome are remarkably similar ($\approx$ 10–15 [16,17]) suggesting that the assumption of approximately one driver mutation per pathway is not too strong of an assumption. It is also possible that multiple driver mutations are necessary to perturb a pathway and thus these mutations co-occur in patients. In this situation, there remains a large subset of genes in the pathway whose mutations are exclusive, e.g. a subset obtained by removing one gene from each co-occurring pair. The identification of these subsets of genes can be used as a starting point to later identify the other genes with co-occurring mutations.

Our contribution

This work proposes a mathematical framework to study the problem of *de novo* discovery of driver genes and pathways. We define two generative models of driver mutations in cancer, the D>P model and the D=P model, and study the algorithmic complexity of the discovery problem in each of the models, both analytically and in simulations. The two generative models differ in how conditioning on a genome being from a cancer patient affects the ratio between the driver and passenger mutation probabilities in that genome. While the difference is relatively small, it has a major implication on the practicality of the standard single gene test for identifying the driver genes. In the first model we prove a bound on the number of patients required to detect all driver genes with high probability using a single gene test, while in the second model it is not possible to identify the driver genes using such a test for *any* number of patients.

Next, we study a weight function on sets of genes that quantifies the coverage and exclusivity properties of a driver pathway. We introduced this function in [11], and showed that finding sets with high weight provides an alternative approach for identifying driver mutations. Here, we prove that for both generative models, when mutation data from enough patients is available, the weight function is monotone in the number of discovered driver genes and is maximized by the driver pathway. Based on this observation we prove that a simple greedy algorithm identifies the driver pathways with high probability. This improves the result in [11], where we showed that the discovery problem is NP-hard for arbitrary mutation data and that a greedy algorithm performs well under different conditions that did not arise from a generative model of the data. We also show that our earlier Markov Chain Monte Carlo (MCMC) approach for identifying the driver pathways rapidly converges to the driver pathway in both generative models, thus improving the convergence result of [11] that considered arbitrary mutation data. These results show that we can identify driver pathways *without* an estimate of the background mutation rate (BMR), giving a more reliable and robust solution for the problem.

We complement our analytical results with experiments on simulated and real cancer sequencing data. For the first D>P model, we compare the number of patients required to identify driver genes using the single gene test with the number required using the greedy algorithm that maximizes the weight function. We show that the number of patients is similar when a perfect estimate of the BMR is available, but that the greedy algorithm requires a smaller number of patients when the estimate of the BMR deviates from its real value. For the second D=P model, we empirically verify that the single gene test cannot identify the driver genes even when a huge number of patients are analyzed, while the greedy algorithm correctly identifies all the driver genes. Finally, we test the performance of the greedy algorithm on mutation data from recent cancer sequencing studies, and show that the greedy algorithm can be used to identify the set of maximum weight on these datasets, even if the data is not guaranteed to satisfy the assumptions of our models. Our analytical and experimental results help characterize the limitations of detecting driver genes and pathways under reasonable models of somatic mutation.

In the remainder of this paper we consider the case in which the mutation matrix contains only one driver pathway. However, our results can be generalized to the case of multiple disjoint driver pathways. In particular the following iterative procedure identifies all driver pathways using our algorithms: after identifying a driver pathway, remove its genes from the mutation matrix, and look for driver pathways in the reduced mutation matrix.

Methods

Stochastic models for somatic mutations in cancer

In this section we introduce two stochastic models for somatic mutations in cancer. In both models driver mutations occur in *sets* of genes, which we refer to as *driver pathways*. Passenger mutations occur randomly across all genes. We assume that mutations have been measured in n genes in a collection of m cancer patients, and represent the somatic mutations as a $m \times n$ binary mutation matrix A. The entry A_{ig} in row i and column g is equal to 1 if gene g is mutated in patient i, and it is 0 otherwise. Let $\mathcal{G}$ be the set of all columns (genes). In both models, we assume that the mutation matrix contains a *driver pathway*: a subset $\mathcal{D} \subseteq \mathcal{G}$ of genes, with $|\mathcal{D}| = k$, such that in each patient *exactly one* of the genes of $\mathcal{D}$ contains a driver mutation. Thus, a driver pathway $\mathcal{D}$ exhibits the properties of high *coverage* – every patient has a mutation in a gene in $\mathcal{D}$ – and *mutual exclusivity* – no patient has a driver mutation in more than one gene in $\mathcal{D}$. In both models, random *passenger* mutations occur at random in all genes, including genes in $\mathcal{D}$. The difference between the two models is in the relative mutation rates in driver and passenger genes.

Following the hypothesis that cancer is triggered by a mutation in a driver gene, the sample of cancer patients can be viewed as a subset of a larger initial population. The genome of each member of the initial population was subject to random mutations, where each gene was mutated independently, and our sample is the subset of the initial population with a driver mutation in a gene of $\mathcal{D}$.

The first stochastic model captures the above intuition by modeling the distribution of mutations in patients as independent with fixed probability q, conditioning on having a driver mutation. The mutation matrix A is generated by the following process: in each row (patient) we choose one gene $d \in \mathcal{D}$ uniformly at random to contain the driver mutation, and set the corresponding entry A_{id} to 1. All other entries at that row are set to 1 with probability $q < 1$ and to 0 otherwise, and all events are independent. We call the parameter q the *passenger mutation probability*, as it is the probability that a gene contains a passenger mutation. We emphasize that q is greater than the BMR, since it is the probability that a *gene* has a passenger mutation. For example, estimates of the BMR are typically $\approx 10^{-5} - 10^{-6}$, and since the length of most genes is around 10^4, we have that $q \approx 10^{-1} - 10^{-2}$. We denote this model as the D>P model.

A possible limitation of the D>P model is that it implies a conditional distribution in which driver genes have higher expected frequency of mutation than the passenger genes (thus the name D>P model) in a cohort of patients. In practice the driver pathway $\mathcal{D}$ could contain dozens of genes, and some of them may have rare driver mutations. Thus the expected frequency of mutation of some genes in $\mathcal{D}$ may be indistinguishable from the expected frequency of mutation of some passenger genes. To examine this situation we introduce a second model, which we call the D=P model, in which all genes in $\mathcal{G}$ are mutated with the same probability in the patients, regardless of whether they are driver or passenger genes. Of course, this is a "worst case" model, as any cancer cohort with a reasonable number of patients will have some driver genes mutated at appreciable frequency. Nevertheless, we study the D=P model to consider the limits of driver pathway identification. The mutation matrix A in the D=P model is generated by the following process: in row (patient) i an entry A_{id} is chosen uniformly at random for $d \in \mathcal{D}$ and is set to 1. All other entries $A_{id'}$ for $d' \in \mathcal{D}$ are set to 1 with probability $r = \frac{qk-1}{k-1}$, and all entries A_{ig}, for $g \in \mathcal{G} \setminus \mathcal{D}$ are set to 1 with probability q. All events are independent. We require $q \geq \frac{1}{k}$ so that r is a proper probability. Note that for any $g \in \mathcal{G}$ the probability that g is mutated is the same since for $d \in \mathcal{D}$, $\frac{1}{k} + (1 - \frac{1}{k})r = q$.

Note that both models differ from a simple *binomial* model, where each entry of A is mutated independently with a fixed probability. Since we condition on each patient having at least one mutation in $\mathcal{D}$, the entries of A corresponding to genes in $\mathcal{D}$ are not independent. In what follows, we let $\Gamma(g) = \{i : A_{ig} = 1\}$ denote the set of patients in which a gene g is mutated. Similarly, for a set M of genes, let $\Gamma(M)$ denote the set of patients in which at least one of the genes in M is mutated: $\Gamma(M) = \cup_{g \in M}\Gamma(g)$.

Results

Finding recurrently mutated genes

The standard approach to identify the driver genes is to identify recurrently mutated genes, i.e. those genes whose observed frequency of mutations is significantly higher than the expected *passenger mutation probability*[1-3]. This approach assumes a prior knowledge or a good estimate of the passenger mutation probability, the parameter q in our models. In particular if gene $g \in \mathcal{G}$ is not in the driver pathway $\mathcal{D}$, then the number of patients in which g is mutated among a collection of m cancer patients is described by a binomial random variable $B(m, q)$ with success probability q. If we know the value of q for each gene $g \in \mathcal{G}$ we can compute the probability $p_g = \Pr[B(m, q) \geq |\Gamma(g)|]$ of observing gene g mutated in at least $|\Gamma(g)|$ patients assuming $g \notin \mathcal{D}$ (i.e., p_g is the p-value for g). This approach is combined with a multi-hypothesis test to identify a list $\mathcal{O}$ of genes, each mutated in significantly more patients than expected. The pseudocode for such a test is given in Algorithm 1 RMG. In Algorithm 1 RMG we use Bonferroni correction for multiple hypothesis testing, that is we include in $\mathcal{O}$ the genes for which $p_g \leq \frac{\alpha}{n}$, for a fixed value α; the Bonferroni correction guarantees that the probability of reporting in $\mathcal{O}$ any gene not in $\mathcal{D}$ is bounded by α. Other corrections, like Benjamini-Hochberg [18] to control the *False Discovery Rate*, are possible. The results of this section also apply to these other corrections.

Algorithm 1 RMG

Pseudocode of the algorithm for finding recurrently mutated genes, based on a single-gene test.

Input: An $m \times n$ mutation matrix A, a probability q that a gene contains a passenger mutation in a patient, a significance level α.

Output: Set $\mathcal{O}$ of recurrently mutated genes.

```
$\mathcal{O} \leftarrow \emptyset$;
for $g \in \mathcal{G}$ do
    $\Gamma(g) \leftarrow \{i : A_{ig} = 1\}$;
    $p_g \leftarrow \Pr[B(m, q) \geq |\Gamma(g)|]$;
    if $p_g \leq \frac{\alpha}{n}$ then $\mathcal{O} \leftarrow \mathcal{O} \cup \{g\}$;
return $\mathcal{O}$;
```

We first analyze the D>P model of Section "Stochastic models for somatic mutations in cancer". We start by showing that if q is known and the number of patients is sufficiently large, then Algorithm 1 RMG outputs all the driver genes with high probability.

Theorem 1. *Suppose an $m \times n$ mutation matrix A is generated by the D>P model with $\mathcal{D} = k$, the family wise error rate of the test is $\alpha = \frac{1}{2n^{\varepsilon}}$ and Algorithm 1 RMG outputs $\mathcal{O}$. If $m \geq \frac{2k^2(1+\varepsilon)}{(1-q)^2} \ln 2n$ for a constant $\varepsilon > 0$, then $Pr[\mathcal{O} \neq \mathcal{D}] \leq \frac{1}{n^{\varepsilon}}$.*

Proof. The p-value calculations and the Bonferroni correction in Algorithm 1 RMG guarantee that the probability that any gene $g \notin \mathcal{D}$ is included in the output set $\mathcal{O}$ is bounded by $\alpha = \frac{1}{2n^{\varepsilon}}$. It remains to prove that if $m \geq \frac{2k^2(1+\varepsilon)}{(1-q)^2} \ln 2n$ the probability that any $d \in \mathcal{D}$ is not included in $\mathcal{O}$ is bounded by $\frac{1}{2n^{\varepsilon}}$.

Consider a gene $d \in \mathcal{D}$. Let $X_i = 1$ if gene d is mutated in patient i, and $X_i = 0$ otherwise. Note that for $i \neq j$, X_i and X_j are independent. Let X be the number of patients in which d is mutated. We have $X = \sum_{i=1}^{m} X_i$. To compute $\mathbf{E}[X_i]$ we observe that a driver gene is mutated with probability 1 when it contains the driver mutation, and with probability q otherwise. Since the gene d containing the driver mutation is chosen uniformly at random among all the k genes in $\mathcal{D}$, we have $\mathbf{E}[X_i] = \frac{1}{k} + \left(1 - \frac{1}{k}\right) q$. Thus $\mathbf{E}[X] = \sum_{i=1}^{m} \mathbf{E}[X_i] = m(\frac{1}{k} + \left(1 - \frac{1}{k}\right) q) > mq$. Let $t = \frac{1}{k}\left(\frac{1-q}{2}\right)$. By the Chernoff-Hoeffding bound:

$$\begin{aligned}\Pr[X \leq \mathbf{E}[X] - tm] &= \Pr[X \leq m\mathbf{E}[X_i] - tm] \\ &\leq e^{-\frac{2m^2t^2}{m}} \leq \frac{1}{2n^{1+\varepsilon}}.\end{aligned}$$

Since $|\mathcal{D}| < n$, by union bound we have:

$$\begin{aligned}&\Pr[\exists d \in \mathcal{D} \text{ mutated in} \leq (\mathbf{E}[X] - tm) \text{ patients}] \\ &\qquad \leq n\frac{1}{2n^{1+\varepsilon}} = \frac{1}{2n^{\varepsilon}}.\end{aligned}$$

Thus with probability at least $1 - \frac{1}{2n^{\varepsilon}}$ all genes in $\mathcal{D}$ are mutated in at least $\mathbf{E}[X] - tm$ patients. Let $B(m, q)$ be a binomial random variable with parameters m,q. Using the Chernoff-Hoeffding bound we can upper bound the p-value p_d that Algorithm 1 RMG derives for $d \in D$:

$$p_d \leq \Pr[|B(m, q) - mq| \geq tm] \leq e^{-2\frac{t^2m^2}{m}} \leq \frac{1}{2n^{1+\varepsilon}}.$$

Thus, with probability at least $1 - \frac{1}{2n^{\varepsilon}}$ for any $d \in \mathcal{D}$ the number of patients with a mutation in d is such that its p-value satisfies $p_d < \alpha/n$ and thus it is included in the output set $\mathcal{O}$.

Theorem 1 shows that in the D>P model an estimate of the passenger mutation probability q and a sufficient number of patients are enough to identify the driver genes. This is not the case in the D=P model. It is easy to see that in D=P model the expected number of rows in which a column g is mutated is the same for all $g \in \mathcal{G}$, that is for all $g \in \mathcal{G}$ we have $\mathbf{E}[\,|\Gamma(g)|] = qm$. In fact, the number $|\Gamma(d)|$ of patients in which a gene $d \in \mathcal{D}$ is mutated and the number $|\Gamma(g)|$ of patients in which gene $g \notin \mathcal{D}$ is mutated are both binomial random variables $B(m, q)$. We thus have the following. □

Fact 1. *Under the D=P model, the probability distribution of $|\Gamma(d)|$ for $d \in \mathcal{D}$ and $|\Gamma(g)|$ for $g \notin \mathcal{D}$ are the same. Thus Algorithm 1 RMG cannot identify the genes in $\mathcal{D}$ for any number of patients m.*

Finding recurrently mutated driver pathways

In this section we analyze a method that identifies the set $\mathcal{D}$ of driver genes with no prior information on the passenger mutation probability q, and works for both the D>P and D=P models. The method relies on a weight function $W(M)$, defined on sets of genes, first introduced in [11]. The measure W quantifies the extent to which a set simultaneously exhibits both: (i) high *coverage*: most patients have at least one mutation in the set; (ii) high *exclusivity*: nearly all patients have no more than one mutation in the set.

For a set of genes, M, we define the coverage overlap $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$. Note that $\omega(M) \geq 0$, with equality if and only if the mutations in M are mutually exclusive. To account for both the coverage, $\Gamma(M)$, and the coverage overlap, $\omega(M)$, we define the weight function of M:

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|.$$

Finding a set M of genes with maximum weight is in general a computationally challenging problem (it is NP-hard in the worst case [11]). Nonetheless, we showed in [11] that under some assumptions on the distribution of mutations in patients, a greedy algorithm will identify the maximum weight set. We also proposed a Markov Chain Monte Carlo (MCMC) approach that samples sets of genes with probability proportional to their weight.

Based on the coverage and exclusivity properties of a driver pathway we expect it has the highest weight among all sets of size k. In this section we formalize this intuition for our generative models and show that under the two models the maximum weight set is easy to compute. We use M_k^* to denote the set of size k with maximum weight (M_k^* may not be unique).

We start with the D>P model. Note that the parameter q controls the expected number of passenger mutations in a set of k passenger genes. Since passenger mutations are relatively rare and k (the number of genes in a driver pathway) is relatively small, we expect that a set of k passenger genes will not have a mutation in the majority of the patients. Thus we assume that the probability $1 - (1-q)^k$ that a set of k passenger genes contains at least one mutation in a patient is less than a constant $a < \frac{1}{2}$. Since $1 - (1-q)^k \approx qk$ we have $q \leq \frac{a}{k}$. For ease of exposition in what follows we set $a = \frac{1}{4}$, so that $q \leq \frac{1}{4k}$.

Let $M_{k,\ell} \subset \mathcal{G}$ be a set of k genes with exactly ℓ genes of $\mathcal{D}$, that is $M_{k,\ell} = \{d_1, d_2, \ldots, d_\ell\} \cup \{g_1, \ldots, g_{k-\ell}\}$ with $d_j \in \mathcal{D}$ for $1 \leq j \leq \ell$, and $g_j \in \mathcal{G} \setminus \mathcal{D}$ for $1 \leq j \leq k - \ell$. We first prove that $\mathbf{E}[\,W(M_{k,\ell})]$ is monotone in ℓ.

Lemma 1. *Let $q \leq \frac{1}{4k}$. For $0 \leq \ell \leq k-1$: $\mathbf{E}[\,W(M_{k,\ell+1})] \geq \mathbf{E}[\,W(M_{k,\ell})] + \frac{m}{2k}$.*

Proof. Let M be any subset of $\mathcal{G}$, and let $\mathbf{E}[\,W(M)] = \sum_{i=1}^{m} \mathbf{E}[\,T_i]$, where T_i is the "contribution" of patient i to $W(M)$, i.e. $T_i = 2 - \ell$ if $\ell > 0$ genes of M are mutated in i, and 0 otherwise. Note that T_i is the difference of two (dependent) random variables: $T_i = Y_i - Z_i$, where: $Y_i = 0$ if no gene of M is mutated in i, and 2 otherwise; $Z_i = \ell$ if $\ell \geq 0$ genes of M are mutated in i.

Now consider $M_{k,\ell}$ that contains a subset L of ℓ elements of $\mathcal{D}$. Consider the event E_i="one of the genes of L is driver in patient i", and $\bar{E}_i$ its complement. We have $\mathbf{E}[\,T_i] = \mathbf{E}[\,T_i|E_i]\Pr[E_i] + \mathbf{E}[\,T_i|\bar{E}_i]\Pr[\bar{E}_i]$. For $M_{k,\ell}$, when E_i holds, we have that $Y_i = 2$ (because one gene of L is mutated) and $Z_i = 1 + B(k-1, q)$, where $B(k, q)$ is a binomial random variable with parameters k,q. When E_i does not hold, we have that $Y_i = 2$ with probability $1 - (1-q)^k$ and $Z_i = B(k, q)$ (since each gene of $M_{k,\ell}$ is mutated independently with probability q). Thus for $M_{k,\ell}$ we have

$$\begin{aligned}\mathbf{E}[\,T_i] =& \mathbf{E}[\,T_i|E_i]\Pr[E_i] + \mathbf{E}[\,T_i|\bar{E}_i]\Pr[\bar{E}_i]\\ =& (2 - (1 + q(k-1)))\frac{\ell}{k} + (2(1-(1-q)^k)\\ &- qk)\left(1 - \frac{\ell}{k}\right).\end{aligned}$$

Thus $\mathbf{E}[M_{k,\ell}] = m((2 - (1 + q(k-1)))\frac{\ell}{k} + (2(1-(1-q)^k) - qk))\left(1 - \frac{\ell}{k}\right)$.

Analogously for $M_{k,\ell+1}$ we have $\mathbf{E}[\,T_i] = (2 - (1 + q(k-1)))\frac{\ell+1}{k} + (1 - \frac{\ell+1}{k})(2(1-(1-q)^k) - qk)$ and $\mathbf{E}[M_{k,\ell+1}] = m((2-(1+q(k-1)))\frac{\ell+1}{k} + (2(1-(1-q)^k) - qk))(1 - \frac{\ell+1}{k})$.

Thus we have:

$$\begin{aligned}
&\mathbf{E}[\,W(M_{k,\ell+1})] - \mathbf{E}[\,W(M_{k,\ell})]\\
&\quad = m\left(\frac{2-(1+q(k-1))}{k} - \frac{2(1-(1-q)^k)-qk}{k}\right)\\
&\quad = m\left(-\frac{1}{k}+\frac{q}{k}+\frac{2}{k}(1-q)^k\right)\\
&\quad \geq m\left(-\frac{1}{k}+\frac{q}{k}+\frac{2}{k}-2q\right)\\
&\quad = m\left(\frac{1}{k}+\frac{q}{k}-2q\right)\\
&\quad \geq m\frac{1}{2k}.
\end{aligned}$$

where the first inequality follows from $(1-q)^k \geq 1-qk$, and the last inequality follows from $q \leq \frac{1}{4k}$ and $q > 0$.

Next we show that for sufficiently large number of patients m, the random value $W(M_{k,\ell})$ is concentrated near its expectation. □

Theorem 2. *Suppose an $m \times n$ mutation matrix A is generated by the* D>P *model with $|\mathcal{D}| = k$ and $q \leq \frac{1}{4k}$. For $m \geq 8k^3(k+\varepsilon)\ln n$, and for $0 \leq \ell \leq k-1$, $Pr[\,\exists M_{k,\ell}$ s.t. $|W(M_{k,\ell}) - \mathbf{E}[\,W(M_{k,\ell})| \geq \frac{m}{4k}] \leq \frac{1}{n^{\varepsilon}}$.*

Proof. Let $M = \{g_1, g_2, \ldots, g_k\}$ be a set of k genes. Consider the sequence $\mathcal{X}$ of random variables $X_{1,1}, X_{1,2}, \ldots, X_{1,k}, X_{2,1}, X_{2,2}, \ldots, X_{2,k}, \ldots, X_{m,k}$, with $X_{i,j} = 1$ if gene g_j is mutated in patient i, and $X_{i,j} = 0$ otherwise. (Note that $X_{i,j}$ are not mutually independent since at least one gene in the driver pathway $\mathcal{D}$ is mutated in each patient.) The random variable $W(M)$ is determined by the sequence $\mathcal{X}$. Now consider the *Doob martingale* (see [19]) $Z_{i,j} = \mathbf{E}[\,W(M)|X_{1,1}, \ldots, X_{i,j}]$, $0 \leq i \leq m$, $1 \leq j \leq k$. Note that $Z_{m,k} = W(M)$, and $\mathbf{E}[\,Z_{m,k}] = \mathbf{E}[\,W(M)]$. Since changing the value of any of the random variables in $\mathcal{X}$ changes $W(M)$ by at most 1 and there are km such random variables, by Azuma-Hoeffding inequality we have that, for all $t > 0$:

$$\Pr[\,|W(M) - \mathbf{E}[\,W(M)]\,| \geq t] \leq 2e^{-\frac{2t^2}{km}}.$$

Setting $t = m/(4k)$, and summing over all $\binom{n}{k}$ possible choices of the set M gives the result.

Combining the results of Lemma 1 and Theorem 2 we have □

Corollary 1. *If $m \geq 8k^3(k+\varepsilon)\ln n$, then $Pr[\,M_k^* \neq \mathcal{D}] \leq \frac{1}{n^{\varepsilon}}$.*

Corollary 1 shows that with sufficient number of patients the set $\mathcal{D}$ can be identified by finding the set of maximum weight, without an estimate of the passenger mutation probability q We previously showed in [11] that with an arbitrary mutation distribution identifying the set of maximum weight is NP-Hard. However, a corollary of Theorem 2 shows that in the D>P model computing a set of maximum weight is easy.

Corollary 2. *If $m \geq 8k^3(k+\varepsilon)\ln n$ and $q \leq \frac{1}{4k}$, Algorithm 2* GreedyWeight *that computes the weight function of up to $O(nk)$ sets finds M_k^* with failure probability $\leq \frac{1}{n^{\varepsilon}}$.*

Proof. The pseudocode for Algorithm 2 GreedyWeight is given below. Theorem 2 guarantees that if g^* is inserted in M, it is in $\mathcal{D}$, and that when a gene $g \in M \setminus \mathcal{D}$ is considered, it will be switched with a gene $g' \in \mathcal{D} \setminus M$. □

Algorithm 2 GreedyWeight

Pseudocode of the greedy algorithm for finding the set M of maximum weight $W(M)$.

Input: An $m \times n$ mutation matrix A, integer $k > 0$.

Output: Set M^* of maximum weight $W(M^*)$.

```
M ← $k$ random columns from $A$;
$M^* \leftarrow M$;
for $g \in M$ do
    $g^* \leftarrow \arg\max_{g' \in \mathcal{G} \setminus M^*}\{W(M^* \setminus \{g\} \cup \{g'\})\}$;
    if $W(M^* \setminus \{g\} \cup \{g^*\}) > W(M^*)$ then
        $M^* \leftarrow M^* \setminus \{g\} \cup \{g'\}$;
return $M^*$;
```

We now consider the D=P model. Analogously to what we proved under the D>P model, we prove that maximizing the weight function W identifies the driver pathway $\mathcal{D}$ when mutation data from enough patients is available.

Theorem 3. *Suppose an $m \times n$ mutation matrix A is generated by the* D=P *model with $|\mathcal{D}| = k$. If $m \geq \frac{k^3(k+\varepsilon)}{2(1-q)^{2k+2}}\left(\frac{k-1}{k}\right)^{2k}\ln n$, then $Pr[\,M_k^* \neq \mathcal{D}] \leq \frac{1}{n^{\varepsilon}}$.*

Proof. Consider $\ell \geq 1$. As in the proof of Lemma 1, we have that for any pair of sets $M_{k,\ell}, M_{k,\ell+1}$ of size k containing ℓ and $\ell+1$ elements of $\mathcal{D}$ respectively, we have: $\mathbf{E}[\,W(M_{k,\ell+1}] \geq \mathbf{E}[\,W(M_{k,\ell})] + m\frac{2}{k}(1-q)^{k+1}\left(\frac{k}{k-1}\right)^k$. Using the Azuma-Hoeffding inequality with $t = m\frac{1}{k}(1-q)^{k+1}\left(\frac{k}{k-1}\right)^k$ we have that $Pr[\,\exists M_{k,\ell}$ s.t. $|W(M_{k,\ell}) - \mathbf{E}[\,W(M_{k,\ell})| \geq t] \leq \frac{1}{n^{\varepsilon}}$ for all $M_{k,\ell}$. The theorem follows combining these two properties.

We prove that a simple greedy algorithm, similar to Algorithm 2 GreedyWeight that we proposed for the D>P model, identifies the set M_k^* of maximum weight under the D=P model. □

Corollary 3. *If* $m \geq \frac{k^3(k+\varepsilon)}{2(1-q)^{2k+2}}\left(\frac{k-1}{k}\right)^{2k}\ln n$, *a greedy algorithm that computes the weight function of up to* $O(n^2)$ *sets finds* M_k^* *with failure probability* $\leq \frac{1}{n^{\varepsilon}}$.

Proof. Start with an arbitrary set M of k genes. By the proof of Theorem 3, if M already contains at least one gene of $\mathcal{D}$, Algorithm 2 `GreedyWeight` produces the set $\mathcal{D}$ in output with failure probability $\leq \frac{1}{n^{\varepsilon}}$. Thus we only need to make sure that the initial set M includes at least one gene of $\mathcal{D}$. To do this, take an arbitrary pair g_1, g_2 of genes in M, and find the pair $(g_3, g_4) \in \mathcal{G} \setminus M$ that maximizes $W(M \setminus \{g_1, g_2\} \cup \{g_3, g_4\})$. Then exchange g_1, g_2 for g_3, g_4 if $W(M \setminus \{g_1, g_2\} \cup \{g_3, g_4\}) > W(M)$. Running Algorithm 2 `GreedyWeight` from the resulting set M gives the result.

Thus under the `D=P` model we identify the driver pathway $\mathcal{D}$ by maximizing $W(M)$. Recall that Algorithm 1 `RMG` cannot find driver genes under this model (Section "Finding recurrently mutated genes", Fact 1). Also note that when $q \leq 1/2$ and the probability $(1-q)^k$ that a set of k genes in $\mathcal{G} \setminus \mathcal{D}$ is not mutated in a patient is greater than $\frac{1}{2}\left(\frac{k-1}{k}\right)^k$ (this occurs when passenger mutations are relatively rare, for example when $q \approx 1/k$) the bound on m in Corollary 3 is the same as the bound in Corollary 2. That is, the weight W identifies the set $\mathcal{D}$ under both the `D>P` and `D=P` models with the same number of patients.

For completeness, we also analyze the Monte-Carlo Markov Chain approach proposed in [11] to sample sets of genes with distribution exponentially proportional to their weight. The pseudocode for the sampling procedure used by the Monte-Carlo Markov Chain approach is given in Algorithm 3 `MCMC-Sampling`. It is easy to verify that the chain is ergodic with a unique stationary distribution $\pi(M) = \frac{e^{cW(M)}}{\sum_{R \in \mathcal{M}_k} e^{cW(R)}}$, where $\mathcal{M}_k = \{M \subset \mathcal{G} | |M| = k\}$. The efficiency of this algorithm depends on the speed of convergence of the Markov chain to its stationary distribution. □

Algorithm 3 `MCMC-Sampling`

Pseudocode of the sampling procedure for the MCMC algorithm.

Input: Current state $M^{(t)}$

Output: Next state $M^{(t+1)}$

```
$w \leftarrow$ gene chosen uniformly at random from $\mathcal{G}$;
$v \leftarrow$ gene chosen uniformly at random from $M^{(t)}$;
$P(M^{(t)}, w, v) \leftarrow \min[1, e^{cW(M^{(t)} \setminus \{v\} \cup \{w\}) - cW(M^{(t)})}]$;
With probability $P(M^{(t)}, w, v)$ set
    $M^{(t+1)} \leftarrow M^{(t)} \setminus \{v\} \cup \{w\}$, otherwise $M^{(t+1)} \leftarrow M^{(t)}$;
```

In [11], we show that there is a non-trivial interval of values for c for which the chain is rapidly mixing without assuming any generative model for the mutation matrix. Applying the analysis in [11] to the `D>P` and `D=P` models requires $0 < c < \frac{1}{k}$. However, applying Lemma 1 and 2 under the `D>P` model, and Theorem 3 under the `D=P` model we show that for any $c > 0$ the process rapidly converges to the set $\mathcal{D}$.

Theorem 4. *Suppose an* $m \times n$ *mutation matrix* A *with* $|\mathcal{D}| = k$ *is generated by the* `D>P` *model with* $q \leq \frac{1}{4k}$, *or the* `D=P` *model with* $q \leq \frac{1}{2}$ *and* $(1-q)^k \geq \frac{1}{2}\left(\frac{k-1}{k}\right)^k$. *For* $m \geq 8k^3(k+\varepsilon)\ln n$ *and any* $c > 0$, *the MCMC converges to the set* $\mathcal{D}$ *in* $O(nklogk)$ *iterations with probability* $\geq 1 - \frac{1}{n^{\varepsilon}}$.

Proof. As stated above, the analysis of [11] applied to the `D>P` and `D=P` models gives the result for $0 < c < \frac{1}{k}$. We now prove that the result holds for $c \geq \frac{1}{k}$. The theorem follows by combining the two cases.

Consider the MCMC and assume there is no time step t such that the chain transitions from a set $M_{k,\ell+1}$ containing $\ell + 1$ genes in $\mathcal{D}$ to a set $M_{k,\ell}$ containing ℓ genes in $\mathcal{D}$. Note that if the MCMC is in a state containing $\ell < k$ genes of $\mathcal{D}$, it will transition to a state with $\ell + 1$ genes of $\mathcal{D}$ (that is, $w \in D$ and $v \notin \mathcal{D}$ are chosen) with probability $\geq 1/(kn)$. From a coupon collector analysis we have that, if the MCMC never transitions from a set $M_{\ell+1}$ containing $\ell + 1$ genes in $\mathcal{D}$ to a set M_ℓ containing ℓ genes in $\mathcal{D}$, the MCMC converges to the set $\mathcal{D}$ after $2kn\ln(2kn)$ steps with probability at least $1 - (2n)^{-1}$.

We now bound the probability that the MCMC moves from a set $M_{\ell+1}$ to a set M_ℓ in the $2kn\ln(2kn)$ steps before reaching state $\mathcal{D}$. Given the choice of m, from Theorem 2 and Theorem 3 we have that the probability that in a particular step the MCMC moves from a set $M_{\ell+1}$ to a set M_ℓ is bounded by $e^{-c\frac{m}{k}}$. The theorem follows by union bound on the $2kn\ln(2kn)$ steps and from the bounds of m and c. □

Experimental results: simulated data

In this section we compare the single gene test (Algorithm 1 `RMG`) with the driver pathway approach (using the weight function $W(M)$) to detect the set of driver genes using mutation data simulated using the `D>P` and the `D=P` model. In particular, we use Algorithm 2 `GreedyWeight` of Section "Finding recurrently mutated driver pathways" to identify the set M_k^* of maximum weight, where $k = |\mathcal{D}|$.

We first consider the `D>P` model, generating mutation data with $k = |\mathcal{D}| = 20$, $n = 10000$, and for different values of q. In particular we considered $q \in \{0.0125; 0.0075; 0.001\}$. We set $\alpha = 0.005$ for Algorithm 1 `RMG` which corresponds to $\varepsilon = 0.5$. To compare the performance of the two algorithms, we measured the minimum number of patients required to detect the driver pathway $\mathcal{D}$ over a range of estimates of the passenger mutation probability q. Specifically, let $E_{s(q)}$ = "estimate $s(q)$ of q is used by Algorithm 1 RMG". Let $m_{R,x}(s(q)) = \min_m\{\Pr[\mathcal{O} = \mathcal{D}|E_{s(q)}] > x\}$ be the minimum number of patients required for Algorithm 1 RMG to output $\mathcal{O} = \mathcal{D}$ with probability $> x$ over all $m \times n$ mutation matrices generated by the model when the estimate $s(q)$ is used. Similarly, let $\mathcal{P}$ be the output of Algorithm 2 GreedyWeight. Let $m_{G,x} = \min_m\{\Pr[\mathcal{P} = \mathcal{D}] > x\}$ be the minimum number of patients required for Algorithm 2 GreedyWeight to output $\mathcal{D}$ with probability $> x$ over all $m \times n$ mutation matrices generated by the model. Recall that $m_{G,x}$ does not depend on $s(q)$ by Corollary 2.

Figure 1 shows the values of $m_{R,0.99}(s(q))$ and $m_{G,0.99}$ as a function of $s(q)$. We varied $s(q)$ starting from $s(q) = q$ (i.e., q is perfectly estimated) and gradually increased $s(q)$ while maintaining $s(q) < 1/k$. The latter condition assures that $s(q)$ is strictly smaller than the expected probability of mutation of any gene in $\mathcal{D}$, a necessary condition for Algorithm 1 RMG to be able to identify $\mathcal{D}$. To compare $m_{R,0.99}$ and $m_{G,0.99}$ we generated 100 mutation matrices for each $m_i = i \times 100$ patients for $1 \leq i \leq 52$ and obtained an empirical estimate of $m_{R,0.99}$ and $m_{G,0.99}$.[a] For a fixed q, Figure 1 shows that $m_{R,0.99}(s(q))$ is monotonically increasing with $s(q)$, and that as expected both $m_{R,0.99}(s(q))$ and $m_{G,0.99}$ decrease using lower values of q. For $q = 0.0125$ and $q = 0.0075$, when the estimate of q is perfect Algorithm 2 GreedyWeight requires more patients than Algorithm 1 RMG to correctly identify the set $\mathcal{D}$, but when the estimate $s(q)$ is larger than the true value of q, $m_{R,0.99}(s(q))$ increases and becomes much larger than $m_{G,0.99}$. (Typically, an overestimate of q is used so that the test for recurrent genes in conservative [20]). Note that even when $s(q) = q$, $m_{G,0.99}$ is close to $m_{R,0.99}(q)$, and that for $q = 0.001$, $m_{G,0.99} < m_{R,0.99}(s(q))$ even when $s(q) = q$, while the bounds in Theorem 1 and in Corollary 2 give $\frac{m_{G,0.99}}{m_{R,0.99}(q)} \geq 1000$ for all the parameters we used. Similar results were obtained when comparing $m_{R,0.95}(s(q))$ and $m_{G,0.95}$; i.e. the minimum number of patients for which Algorithm 1 RMG and Algorithm 2 GreedyWeight report the driver set $\mathcal{D}$ at least 95% of the time(data not shown).

We also considered the case $s(q) < q$ where the estimate of q is smaller than its true value. In this case, some genes not in $\mathcal{D}$ (false positives) are eventually reported by Algorithm 1 RMG. For example, with $m = 1000$ patients and $q = 0.0125$, when $s(q) = q$, the correct set of genes (with no false positives) were reported. However, when $s(q) = 0.8q$ Algorithm 1 RMG reports false positives in approximately 16% of the datasets. In contrast, Algorithm 2 GreedyWeight does not suffer from this problem, since it does not require an estimate $s(q)$ of q.

We now consider the D=P model, generating mutation data with $k = |\mathcal{D}| = 20$, $q = 0.05$ and $n = 120$. As stated in Fact 1, Algorithm 1 RMG cannot identify the genes in $\mathcal{D}$ for any number m of patients; we checked this property for values of m up to 10^7. For Algorithm 2 GreedyWeight we again estimated $m_{G,0.99}$ as described above generating 100 mutation matrices for each $m_i = i \times 1000$ patients for $1 \leq i \leq 100$, and obtained that $m = 95000$ patients suffices for GreedyWeight to correctly output exactly the genes in $\mathcal{D}$, while the bound of Corollary 3 gives that more than 4×10^5 patients are required for the parameters we used.

In the above experiments we provided the correct parameter k in input to the Algorithm 2 GreedyWeight. In practice, the exact value of k is not known. However, when the number m of patients satisfies the bound of Corollary 2 (resp., Corollary 3) in the D>P (resp., D=P) model, then the weight $W(\mathcal{D})$ of the set $\mathcal{D}$ is greater than

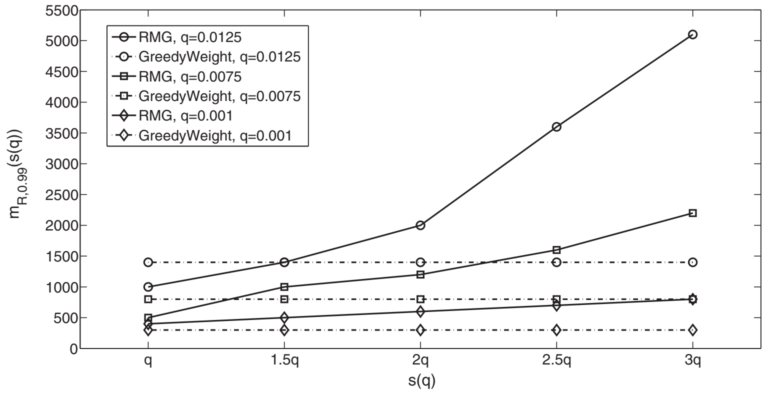

Figure 1 Comparison of Algorithm 1 RMG and Algorithm 2 GreedyWeight. Comparison between the estimate of the number of patients $m_{R,0.99}(s(q))$ required to identify the driver pathway $\mathcal{D}$ with Algorithm 1 RMG, for different estimates $s(q)$ of the probability q and different values of q, and the number of patients $m_{G,0.99}$ required to identify $\mathcal{D}$ with Algorithm 2 GreedyWeight.

the weight of *any* other set of genes. We therefore implemented a modified version of the greedy algorithm that takes as input an upper bound $k_{\max}$ on the size of $\mathcal{D}$, runs Algorithm 2 GreedyWeight for all values k with $2 \leq k \leq k_{\max}$ and outputs the set (of any size) of maximum weight found in the different runs. We repeated the experiments above for the D>P model with $n = 10000$ and $q = 0.0125$ and for the D=P model using $k_{\max} = 22$ for this algorithm, and obtained the same estimates of $m_{G,0.99}$ reported above. This show that even when the exact value of k is not known, Algorithm 2 GreedyWeight can correctly identify $\mathcal{D}$.

Experimental results: cancer sequencing data

Finally, we tested Algorithm 2 GreedyWeight on mutation data coming from three different cancer sequencing studies, as described in [11]. In particular we analyzed cancer mutation data from: lung adenocarcinoma [21], glioblastoma [3], and multiple cancer types [22]. The mutation matrices were prepared using the same procedure described in [11]. Since not all genes have been assayed for mutations in these studies, there is no guarantee that the assumptions of our models hold for these datasets. In addition, the number of mutated patients in the studies is small compared to the bounds our analytical and empirical results suggest for Algorithm 2 GreedyWeight to find the set of maximum weight. Nonetheless, for each of the three datasets we attempted to use Algorithm 2 GreedyWeight to find the set of maximum weight we reported in [11], using the parameter k given by the size of the sets found in [11].

Since the output of Algorithm 2 GreedyWeight depends on the choice of the initial random set (the set M on Line 1 of Algorithm 2), we run Algorithm 2 GreedyWeight 100 times (i.e., starting from 100 different random initial sets). For the mutation data from multiple cancer types, Algorithm 2 GreedyWeight *always* reports the set of maximum weight; for the mutation data from the gliblastoma study, the set of maximum weight is reported by Algorithm 2 GreedyWeight in 58% of the runs. For the lung adenocarcinoma mutation data, Algorithm 2 GreedyWeight reports the set of maximum weight in 43% of the runs, and no other set is reported more frequently. These results show that Algorithm 2 GreedyWeight can be used to identify genes in driver pathways on data from cancer sequencing studies containing a modest number of patients.

Conclusions

We investigate the problem of detecting recurrently mutated genes and pathways using two simple generative models of driver mutations in cancer: the D>P model and the D=P model. In the D>P model, the driver mutation probability is larger than the passenger mutation probability. We prove a bound on the number of patients required to detect all driver genes with high probability using a single gene test of recurrence. In the D=P model, the driver mutation probability and passenger mutation probability cannot be distinguished, and thus it is impossible to identify driver genes using the single gene test for *any* number of patients. We prove that under either model, the weight function on sets of genes that we defined in [11] is maximized by a driver pathway. Thus, with mutation data from enough patients, it is possible to identify driver pathways *without* an estimate of the passenger mutation probability q. In particular, we show that a simple greedy algorithm finds driver pathways with high probability. We also show that an MCMC approach converges rapidly. We present results on simulated data showing that the greedy algorithm successfully identifies the driver pathway with fewer patients than the single gene test when the estimate of q deviates from its real value. Finally, we show that the greedy algorithm can find driver genes and driver pathways in real cancer sequencing data containing a modest number of patients.

In practice, any test that identifies driver genes by recurrent mutations requires a good estimate of the passenger mutation probability q. An underestimate of q leads to false positive predictions of driver genes, while an over estimate (i.e. a conservative estimate to minimize false positives) increases the number of patients required to find driver genes. The passenger mutation probability is derived from the background mutation rate (BMR), which is difficult to measure as it depends on a number of parameters whose values are not easily determined. There has been extensive discussion in the community about appropriate ways to estimate the BMR and find recurrently mutated genes [1,4]. Methods that do not require an estimate of the BMR, as the ones we provide here, can give increased power for the discovery of driver genes. However, further study of more sophisticated mutation models is necessary. For example, we assume a constant passenger mutation probability q across all genes, but models that allow q to vary by gene would be useful in applications and warrant further investigation.

Consent

Written informed consent was obtained from the patient for publication of this report and any accompanying images.

Endnotes

[a] We use the empirical estimates of $m_{R,0.99}(q)$ and $m_{G,0.99}$ only to compare the performance of Algorithm 1 RMG and Algorithm 2 GreedyWeight, and to show how $m_{R,0.99}$ and $m_{G,0.99}$ vary by changing the parameter q. Therefore we do not need extremely accurate estimates of of $m_{R,0.99}(q)$ and $m_{G,0.99}$, that would require the generation

of more mutation matrices and the inclusion of more values of m_i.

Competing interests

FV and BJR declare that they have no competing interests. EU has financial interest in Nabsys Inc.

Authors' contributions

All authors contributed to the analytical part of the work. FV wrote the software and ran the experiments. All authors read and approved the final manuscript.

Acknowledgements

We thank the anonymous reviewers for helpful suggestions that improved the manuscript. A preliminary version of this work was presented at WABI 2011. This work is supported by NSF grants IIS-1016648 and CCF-1023160, the Alfred P. Sloan Foundation, and in part by the MIUR of Italy under project AlgoDEEP port. 2008TFBWL4. BJR is also supported by an NSF CAREER Award (CCF-1053753) and a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Received: 3 January 2012 Accepted: 26 July 2012
Published: 6 September 2012

References

1. Sjoblom T, *et al*: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314:**268–274.
2. Ding L, *et al*: **Somatic mutations affect key pathways in lung adenocarcinoma.** *Nature* 2008, **455:**1069–1075.
3. The Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**(7216):1061–1068.
4. Getz G, Hofling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES: **Comment on "The consensus coding sequences of human breast and colorectal cancers".** *Science* 2007, **317:**1500.
5. Hahn WC, Weinberg RA: **Modelling the molecular circuitry of cancer.** *Nat Rev Cancer* 2002, **2:**331–341.
6. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10:**789–799.
7. Efroni S, Ben-Hamo R, Edmonson M, Greenblum S, Schaefer CF, Buetow KH: **Detecting cancer gene networks characterized by recurrent genomic alterations in a population.** *PLoS ONE* 2011, **6:**e14437.
8. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G: **Patient-oriented gene set analysis for cancer mutation data.** *Genome Biol* 2010, **11:**R112.
9. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: **Automated network analysis identifies core pathways in glioblastoma.** *PLoS ONE* 2010, **5:**e8918.
10. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer.** *J Comput Biol* 2011, **18:**507–522.
11. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer.** *Genome Res* 2012, **22**(2):375–385.
12. McCormick F: **Signalling networks that cause cancer.** *Trends Cell Biol* 1999, **9:**M53–M56.
13. Yeang C, McCormick F, Levine A: **Combinatorial patterns of somatic gene mutations in cancer.** *FASEB J* 2008, **22**(8):2605–2622.
14. Varela I, *et al*: **Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma.** *Nature* 2011, **469:**539–542.
15. Deguchi K, Gilliland DG: **Cooperativity between mutations in tyrosine kinases and in hematopoietic transcription factors in AML.** *Leukemia* 2002, **16:**740–744.
16. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314:**268–274.
17. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban R H, *et al*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321:**1801–1806.
18. Benjamini Y, Hochberg Y: **Controlling the false discovery rate.** *J R Stat Soc* 1995, **57:**289–300.
19. Mitzenmacher M, Upfal E: *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. New York: Cambridge University Press; 2005.
20. Parmigiani G, *et al*: **Response to Comments on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers".** *Science* 2007, **317**(5844):1500. [http://www.sciencemag.org/content/317/5844/1500.4.abstract].
21. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, *et al*: **Somatic mutations affect key pathways in lung adenocarcinoma.** *Nature* 2008, **455:**1069–1075.
22. Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, MacConaill L, Macconaill LE, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong KK, Gabriel S, Beroukhim R, Peyton M, Barretina J, Dutt A, Emery C, Greulich H, Shah K, Sasaki H, Gazdar A, Minna J, *et al*: **High-throughput oncogene mutation profiling in human cancer.** *Nat Genet* 2007, **39:**347–351.

doi:10.1186/1748-7188-7-23
Cite this article as: Vandin *et al.*: Finding driver pathways in cancer: models and algorithms. *Algorithms for Molecular Biology* 2012 **7**:23.