De novo Discovery of Mutated Driver Pathways in Cancer

Fabio Vandin[*,†] Eli Upfal[*,†] Benjamin J. Raphael[*,†]
vandinfa@cs.brown.edu eli@cs.brown.edu braphael@cs.brown.edu

June 6, 2011

Keywords: cancer genomics; pathways; driver mutations; algorithms.

Abstract

Next-generation DNA sequencing technologies are enabling genome-wide measurements of somatic mutations in large numbers of cancer patients. A major challenge in interpretation of this data is to distinguish functional *driver* mutations important for cancer development from random *passenger* mutations. A common approach for identifying driver mutations is to find genes that are mutated at significant frequency in a large cohort of cancer genomes. This approach is confounded by the observation that driver mutations target multiple cellular signaling and regulatory pathways. Thus, each cancer patient may exhibit a different combination of mutations that are sufficient to perturb these pathways. This mutational heterogeneity presents a problem for predicting driver mutations solely from their frequency of occurrence. We introduce two combinatorial properties, *coverage* and *exclusivity*, that distinguish *driver pathways*, or groups of genes containing driver mutations, from groups of genes with passenger mutations. We derive two algorithms, called Dendrix, to find driver pathways *de novo* from somatic mutation data. We apply Dendrix to analyze somatic mutation data from 623 genes in 188 lung adenocarcinoma patients, 601 genes in 84 glioblastoma patients, and 238 known mutations in 1000 patients with various cancers. In all datasets, we find groups of genes that are mutated in large subsets of patients and whose mutations are approximately exclusive. Our Dendrix algorithms scale to whole-genome analysis of thousands of patients and thus will prove useful for larger datasets to come from The Cancer Genome Atlas (TCGA) and other large-scale cancer genome sequencing projects.

[*] Department of Computer Science, Brown University, Providence, RI.
[†] Center for Computational Molecular Biology, Brown University, Providence, RI.

1 Introduction

Cancer is driven by somatic mutations in the genome that are acquired during the lifetime of an individual. These include single nucleotide mutations and larger copy number aberrations and structural aberrations. With the availability of next-generation DNA sequencing technologies, whole-genome or whole-exome measurements of the somatic mutations in large numbers of cancer genomes is now a reality (Meyerson et al., 2010; International cancer genome consortium, 2010; Mardis and Wilson, 2009). A major challenge for these studies is to distinguish the functional *driver mutations* responsible for cancer from the random *passenger mutations* that have accumulated in somatic cells but that are not important for cancer development. A standard approach to predict driver mutations is to identify recurrent mutations (or recurrently mutated genes) in a large cohort of cancer patients. This approach has identified a number of important cancer mutations (e.g. in *KRAS, BRAF, ERRB2,* etc.), but has not revealed all of the driver mutations in individual cancers. Rather, the results from initial studies (Jones et al., 2008; Ding et al., 2008; The Cancer Genome Atlas Research Network, 2008) have confirmed that cancer genomes exhibit extensive mutational heterogeneity with no two genomes – even those from the same tumor type – containing exactly the same complement of somatic mutations. This heterogeneity results not only from the presence of passenger mutations in each cancer genome, but also because *driver mutations* typically target genes in cellular signaling and regulatory pathways (Hahn and Weinberg, 2002; Vogelstein and Kinzler, 2004). 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 complicates efforts to identify functional mutations by their recurrence across many samples, as the number of patients required to demonstrate recurrence of rare mutations is very large.

An alternative approach to testing the recurrence of individual mutations or genes is to examine mutations in the context of cellular signaling and regulatory pathways. Most recent cancer genome sequencing papers analyze *known* pathways for enrichment of somatic mutations (Jones et al., 2008; Ding et al., 2008; The Cancer Genome Atlas Research Network, 2008), and methods that identify known pathways that are significantly mutated across many patients have been developed (e.g. Efroni et al. (2011); Boca et al. (2010)). Also, algorithms that extend pathway analysis to genome-scale gene interaction networks have recently been introduced (Cerami et al., 2010; Vandin et al., 2010). Pathway or network analysis of cancer mutations relies on prior identification of the groups of genes in the pathways. While some pathways are well-characterized and cataloged in various databases (Jensen et al., 2009; Keshava Prasad et al., 2009; Kanehisa and Goto, 2000), knowledge of pathways remains incomplete. In particular, many pathway databases contain a superposition of all components of a pathway and information regarding which of these components are active in particular cell-types is largely unavailable. These concerns, plus the availability of increasing number of sequenced cancer genomes motivate the question of whether it is possible to *automatically* discover groups of genes with driver mutations, or mutated *driver pathways*, directly from somatic mutation data collected from large numbers of patients.

De novo discovery of mutated driver pathways seems implausible because of the enormous number of possible gene sets to test: e.g. there are more than 10^{26} sets of 7 human genes. However, the current understanding of the somatic mutational process of cancer (McCormick, 1999; Vogelstein and Kinzler, 2004) places two additional constraints on the expected patterns of somatic mutations that significantly reduce the number of gene sets to consider. First, an important cancer pathway should be perturbed in a large number of patients. Thus, given genome-wide measurements of somatic mutations, we expect that 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 (Vogelstein and Kinzler, 2004;

Yeang et al., 2008). There are numerous examples of pairs of mutually exclusive driver mutations including: *EGFR* and *KRAS* mutations in lung cancer (Gazdar et al., 2004), *TP53* and *MDM2* mutations in glioblastoma (The Cancer Genome Atlas Research Network, 2008) and other tumor types, and *RAS* and *PTEN* mutations in endometrial (Ikeda et al., 2000) and skin cancers (Mao et al., 2004). Mutations in the four genes, *EGFR*, *KRAS*, *HER2*, and *BRAF*, from the *EGFR-RAS-RAF* signaling pathway were found to be mutually exclusive in lung cancer (Yamamoto et al., 2008). More recently, statistical analysis of sequenced genes in large sets of cancer samples (Ding et al., 2008; Yeang et al., 2008) identified several pairs of genes with mutually exclusive mutations.

We introduce two algorithms to find sets of genes with the following properties: (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. We define a measure on sets of genes that quantifies the extent to which a set exhibits both criteria. We show that finding sets of genes that optimize this measure is in general a computationally challenging problem. We introduce a straightforward greedy algorithm and prove that this algorithm produces an optimal solution with high probability when given a sufficiently large number of patients and subject to some statistical assumptions on the distribution of the mutations (Section 2.1). Since these statistical assumptions are too restrictive for some data (e.g. they are not satisfied by copy number aberrations) and since the number of patients in currently available datasets is lower than required by our theoretical analysis, we introduce another algorithm that does not depend on these assumptions. We use a Markov Chain Monte Carlo (MCMC) approach to sample from sets of genes according to a distribution that gives significantly higher probability to sets of genes with high coverage and exclusivity. Markov Chain Monte Carlo is a well established technique to sample from combinatorial spaces with applications in various fields (Randall, 2006; Gilks, 1998). For example, MCMC has been used to sample from spaces of RNA secondary structures (Meyer and Miklos, 2007), haplotypes (Bansal et al., 2008), and phylogenetic trees (Yang and Rannala, 1997). In general, the computation time (number of iterations) required for an MCMC approach is unknown, but in our case, we prove that our MCMC algorithm converges rapidly to the stationary distribution.

We emphasize that the assumptions that driver pathways exhibit both high coverage and high exclusivity need not be strictly satisfied for our algorithms to find interesting sets of genes. Indeed, mutual exclusivity is a fairly strong assumption, and there are examples of co-occurring, and possibly cooperative, mutations such as *VHL*/*SETD2*/*PBRM1* mutations in renal cancer (Varela et al., 2011), and *CBF* translocations and kinase mutations in acute myeloid leukemias (Deguchi and Gilliland, 2002). Yeang et al. (2008) suggest a model where mutations in genes from the *same* pathway were typically mutually exclusive and mutations in genes from *different* pathways were sometimes co-occurring. It is also possible that mutations in some genes of an essential pathway are insufficient to perturb the pathway on their own and that other co-occurring mutations are necessary. In this case, 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.

We apply our algorithms, called *De novo Driver Exclusivity* (**Dendrix**), to analyze sequencing data from three cancer studies: 623 sequenced genes in 188 lung adenocarcinoma patients, 601 sequenced genes in 84 glioblastoma patients, and 238 sequenced mutations in 1000 patients with various cancers. In all three datasets we find sets of genes that are mutated in large numbers of patients and are mostly exclusive. These sets include genes in the Rb, p53, mTOR, and MAPK signaling pathways, all pathways known to be important in cancer. In glioblastoma, the set of three genes that we identify is associated with shorter survival (Backlund et al., 2003). We also show that the MCMC algorithm efficiently samples multiple sets of six genes in simulated mutation data with thousands of genes and patients. Both the greedy and MCMC algorithms scale to whole-genome analysis of thousands of patients and thus will prove useful for analysis of larger datasets to come from The Cancer Genome Atlas (TCGA) and other large-scale cancer genome sequencing projects.

2 Results

Consider mutation data for m cancer patients, where each of n genes is tested for a somatic mutation (e.g., single nucleotide mutation or copy number aberration) in each patient. We represent the mutation data by a mutation matrix A with m rows and n columns, where each row is a patient, and each column is a gene. The entry A_{ij} in row i and column j is equal to 1 if gene j is mutated in patient i, and it is 0 otherwise (Figure 1). For a gene g, let $\Gamma(g) = \{i : A_{ig} = 1\}$ denote the set of patients in which 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: $\Gamma(M) = \cup_{g \in M} \Gamma(g)$. We say that a set M of genes is *mutually exclusive* if no patient contains more than one mutated gene in M, i.e. $\Gamma(g) \cap \Gamma(g') = \emptyset$ for all $g, g' \in M$. Analogously, we say that an $m \times k$ submatrix M consisting of k columns of a mutation matrix A is *mutually exclusive* if each row of M contains at most one 1. Note that the above definitions also apply when the columns of the mutation matrix A correspond to parts of genes (e.g. protein domains or individual residues). In the results below, we will analyze data using both definitions of the mutation matrix.

Earlier studies (Ding et al., 2008; Yeang et al., 2008) employed straightforward statistical tests to test for exclusivity between pairs of genes. More sophisticated tests for pairwise exclusivity have also been proposed (Bradley and Farnsworth, 2009). However, it is not clear how to extend such pairwise tests to larger groups of genes, particularly because the number of hypotheses grows rapidly as the number of genes in the set increases. Moreover, identification of pairs of mutually exclusive mutated genes is not sufficient for identification of larger sets (as suggested in Yeang et al. (2008)), since mutual exclusion relations are not transitive. For example, consider two patients s_1 and s_2: in s_1, only gene x is mutated; in s_2, genes y, z are mutated. The pairs of genes (x, y) and (x, z) are mutually exclusive, but the pair (y, z) is not. In fact, finding the largest set of genes with mutually exclusive mutations is NP-hard by reduction from maximum independent set (Garey and Johnson, 1990).

Instead, we propose to identify sets of genes (columns of the mutation matrix) that are mutated in a large number of patients and whose mutations are mutually exclusive. We define the following problem.

Maximum Coverage Exclusive Submatrix Problem: *Given an $m \times n$ mutation matrix A and an integer $k > 0$, find a mutually exclusive $m \times k$ submatrix M of k columns (genes) of A with the largest number of non-zero rows (patients).*

We show that this problem is computationally difficult to solve (for proof, see Supplemental Material). Moreover, this problem is too restrictive for analysis of real somatic mutation data. We do not expect mutations in driver pathways to be mutually exclusive because of measurement errors and the presence of passenger mutations. Instead we expect to find a set of genes that are mutated in large number of patients and whose mutations exhibit "approximate exclusivity", meaning that a small number of patients have a mutation in more than one gene in the set. Thus, we aim to find a set M of genes that satisfies the following two requirements: 1. *Coverage:* most patients have at least one mutation in M; 2. *Approximate exclusivity:* most patients have no more than one mutation in M.

There is an obvious trade-off between requiring mutual exclusivity in the set and obtaining low coverage versus allowing greater non-exclusivity in the set and obtaining larger coverage. We introduce a measure on a set of genes that quantifies the tradeoff between coverage and exclusivity. For a set M of genes, we define the coverage overlap $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$. Note that $\omega(M) \geq 0$ with equality holding when the mutations in M are mutually exclusive. To take into account both the coverage $\Gamma(M)$ and the coverage overlap $\omega(M)$ of M we define the weight $W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|$. Note that the weight function $W(M)$ is only one possible measure of the trade-off between coverage and exclusivity (see Methods).

The problem that we want to solve is the following:

Maximum Weight Submatrix Problem: *Given an $m \times n$ mutation matrix A and an integer $k > 0$, find the $m \times k$ column submatrix $\hat{M}$ of A that maximizes $W(M)$.*

Even for small values of k (e.g. $k = 6$) finding the maximum weight submatrix by examining all the possible sets of genes of size k is computationally infeasible: for example, there are $\approx 10^{23}$ subsets of size $k = 6$ of 20000 genes. We show that the Maximum Weight Submatrix Problem is also computationally difficult to solve (for proof, see Supplemental Material) and thus it is likely that there is no efficient algorithm to solve this problem exactly. The problem of extracting subsets of genes with particular properties has also been studied in the context of gene expression data. For example, biclustering techniques are commonly used to identify subsets of genes with similar expression in subsets of patients (Cheng and Church, 2000; Getz et al., 2000; Madeira and Oliveira, 2004; Murali and Kasif, 2003; Segal et al., 2003; Tanay et al., 2002). Other variations, such as finding subsets of genes that preserve order of expression (Ben-Dor et al., 2003) or that cover many patients (Ulitsky et al., 2008; Kim et al., 2010) have been proposed. However, these approaches are not directly applicable to our problem as we seek a set of genes with few co-occurring mutations, while gene expression studies aim to find groups of genes with correlated expression.

We describe our approach considering mutation data at the level of individual genes. However, by adding columns to the mutation matrix, it is possible to apply our method at the subgene level by considering mutations in particular protein domains, structural motifs, or individual residues. (See Section 2.4.1 for an example.)

2.1 A Greedy Algorithm for Independent Genes

A straightforward greedy algorithm for the Maximum Weight Submatrix Problem is to start with the best pair M' of genes and then to iteratively build the set M of genes by adding the best gene (i.e., the one that maximize $W(M)$) until M has k genes (see Methods for the pseudocode of the algorithm). This algorithm is very efficient, but in general there is no guarantee that the set $\hat{M}$ that maximizes $W(M)$ would be identified. However, we show that the greedy algorithm correctly identifies $\hat{M}$ with high probability when the mutation data come from a generative model, that we call *Gene Independence Model* (for proof, see Supplemental Material). In the Gene Independence Model: (1) each gene $g \notin \hat{M}$ is mutated in each patient with probability p_g, independently of all other events, with $p_g \in [p_L, p_U]$ for all g. (2) $W(\hat{M}) \approx m$. (3) each of the genes in $\hat{M}$ is important, so there is no single subset of $\hat{M}$ that has a dominant contribution to the weight of $\hat{M}$. Condition (1) models the independence of mutations for genes that are not in the mutated pathway, and is a standard assumption for somatic single nucleotide mutations (Ding et al., 2008). Condition (2) ensures that the mutations in $\hat{M}$ cover a large number of patients and are mostly exclusive. For a formal definition of *Gene Independence Model*, see Supplemental Material.

Note that in the Gene Independence Model it is possible for the genes in $\hat{M}$ to have observed mutation frequencies that are identical to those of genes not in $\hat{M}$, and thus it is impossible to distinguish the genes in $\hat{M}$ from the genes not in $\hat{M}$ using only the frequency of mutations, for *any* number of patients.

To assess the implications of this for the utility of the greedy algorithm on real data consider the following setting: observed gene mutation frequencies are in the range $[3 \times 10^{-5}, 0.13]$ (derived from a background mutation rate of the order of 10^{-6} (Ding et al., 2008; The Cancer Genome Atlas Research Network, 2008) and the distribution of human gene lengths). If somatic mutations are measured in $n = 20000$ human genes and $k = |\hat{M}| = 10$, then approximately $m = 2400$ patients are required for the greedy algorithm to identify $\hat{M}$ with probability at least $1 - 10^{-4}$. Even if somatic mutations are measured in only a subset of genes (including all the genes in $\hat{M}$) the bound above does not decrease much. For example, assuming $n = 600$ genes are measured (as it is for recent studies (Ding et al., 2008; The Cancer Genome Atlas Research Network, 2008)), including all the $k = 10$ genes in $|\hat{M}|$, approximately $m = 1800$ patients are required to identify $\hat{M}$ with probability at least $1 - 10^{-4}$ using the greedy algorithm. This number of patients is not far from the range that will be soon be available from large-scale cancer sequencing projects (International cancer genome consortium, 2010), but is larger than what is available now. Moreover, we only have shown that the simple greedy algorithm gives a good solution when the mutation data comes from the Gene Independence Model. This model is reasonable for some types of somatic mutations (e.g. single nucleotide mutations) but not others (e.g. copy number aberrations).

2.2 Markov Chain Monte Carlo (MCMC) Approach

To circumvent the limitations of the greedy algorithm described above, we developed a Monte Carlo Markov Chain (MCMC) approach that does not require any assumptions about the distribution of the mutation data or about the number of patients. The MCMC approach samples sets of genes, with the probability of sampling a set M proportional to the weight $W(M)$ of the set. Thus, the frequencies that gene sets are sampled in the MCMC method provides a *ranking* of gene sets, where the sets are ordered by decreasing sampling frequency. Thus, in addition to the highest weight set, one may also examine other sets of high weight ("suboptimal" sets) that are nevertheless biologically significant. Moreover, since the MCMC approach does not require any assumptions about independence of mutations in different genes, it is useful for analysis of copy number aberrations (CNAs) which amplify or delete multiple adjacent genes and thus introduce correlated mutations. Both of these advantages will prove useful in analysis of real mutation data below.

The basic idea of the MCMC is to build a Markov chain whose states are the collections of k columns of the mutation matrix A and to define transitions between the states that differ by one gene. We use a Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) to sample sets $M \subseteq \mathcal{G}$ of k genes with a stationary distribution that is proportional to $e^{cW(M)}$ for some constant $c > 0$. At time t, the Markov chain in state M_t chooses a gene w in $\mathcal{G}$ and a gene v inside M_t, and moves to the new state $M_{t+1} = M_t \setminus \{v\} \cup \{w\}$ with a certain probability. In general there are no guarantees on the rate of convergence of the Metropolis-Hasting algorithm to the stationary distribution. However, we prove that in our case the MCMC is rapidly mixing (Section 4.3), and thus the stationary distribution is reached in a practical number of steps by our method. The MCMC algorithm is described in more detail in Methods.

2.3 Results on simulated mutation data

We first tested the ability of the MCMC algorithm to detect the set M^* of maximum weight $W(M^*)$, for different values of $W(M^*)$. We simulated mutation data starting with a set M of 6 genes. For each patient, we mutate a gene (chosen uniformly at random) in M with probability p_1, and if a gene in M is mutated, then with probability p_2 we mutate another gene in M. Thus, p_1 regulates the coverage of M, and p_2 regulates the exclusivity of M. The genes not in M are mutated using a random model based on the observed characteristics of the glioblastoma data (described below). In particular, we simulated both single nucleotide mutations and copy number aberrations (CNAs). For the single nucleotide mutations, genes were mutated in each patient according to the observed frequency of single nucleotide mutations in the glioblastoma data, independently of other genes.[1] We simulated CNAs by permuting the locations of the observed CNAs on the genome while maintaining their lengths. The procedure accounts for the fact that genes that are physically close on the genome might be mutated together in the same CNA, resulting in correlated mutations.

We ran the MCMC algorithm on sets of 6 genes for 10^7 iterations sampling every 10^4 iterations. Figure 2 reports the ratio between frequency $\pi(M)$ at which M is sampled and the maximum frequency $\pi(\max_{\text{other}})$ of any other sampled set. Note that the same value of $W(M)$ is obtained with multiple different settings of the parameters p_1 and p_2. For example, with $p_1 = 0.81$ and $p_2 = 0.04$, the set M has $W(M) = 67$ (in expectation), and is sampled with frequency 3-fold greater than any other set.

The sampling ratio increases dramatically with the weight $W(M)$ of the set.

To test the ability to identify multiple high weight sets of genes, we simulated mutation data starting

[1] For each gene, we used the observed frequency of mutation rather than a fixed background mutation rate to account for the differences in gene mutation frequencies observed in the real data.

with two disjoint sets, M_1 and M_2, each containing 6 genes. For each patient, we mutate genes in M_1 and M_2 using the probabilities p_1 and p_2 as described above. The sets M_1 and M_2 correspond to two pathways with approximate exclusivity. The genes not in M_1 or M_2 were mutated using the random model described above. Table 1 shows the frequencies with which various sets are sampled in the MCMC. M_1 and M_2 are sampled with highest frequency. Moreover the ratio of their frequencies is very close to the ratio of their probabilities in the stationary distribution of the MCMC. If the MCMC is sampling from the stationary distribution for the two sets M and M' the ratio $\frac{\tilde{\pi}(M)}{\tilde{\pi}(M')}$ should be close to $e^{c(W(M)-W(M'))}$. In our simulations, $\frac{\tilde{\pi}(M_2)}{\tilde{\pi}(M_1)} \sim 0.351$, and $e^{c(W(M_2)-W(M_1))} \sim 0.368$.

Finally, we tested the scalability of our method to datasets containing a larger number of genes and varying numbers of patients. We simulated mutation data as described above on 20000 genes and 1000 patients. The results in this case are very close to the ones presented above. In particular, M_1 and M_2 were the two sets sampled with highest frequency, and the frequency of each was larger than 30%. Sets other than M_1 and M_2 were sampled with frequencies less than 1%. We were still able to identify the sets M_1 and M_2 when the number of patients was reduced to 150. M_1 and M_2 were sampled with frequency 13%, much higher than the any other set. Based on these results, we anticipate that our algorithms would be useful on whole-exome sequencing studies with a modest number of patients.

2.4 Results on cancer mutation data

We applied our MCMC algorithm to somatic mutations from high-throughput genotyping of 238 oncogenes in 1000 patients of 17 cancer types (Thomas et al., 2007), and to somatic mutations identified in recent cancer sequencing studies from lung adenocarcinoma (Ding et al., 2008) and glioblastoma multiforme (The Cancer Genome Atlas Research Network, 2008). In the glioblastoma multiforme analysis, we include both copy number aberrations and single nucleotide (or small indel) mutations, while in the lung adenocarcinoma analysis, we consider only single nucleotide (or small indel) mutations. The MCMC algorithm samples sets with frequency proportional to their weights, and so to restrict attention to sets with high weight we report sets whose frequency is at least 1%. We also reduce the size of the mutation matrix by combining genes that are mutated in *exactly* the same patients into larger "metagenes".

2.4.1 Known Mutations in Multiple Cancer Types

We applied the MCMC algorithm to mutation data from Thomas et al. (2007) who tested 238 known mutations in 17 oncogenes in 1000 patients of 17 different cancer types. 298 of patients were found to have at least one of theses mutations and a total of 324 individual mutations were identified. To perform our analysis, we built a mutation matrix with 298 patients and 18 mutation groups. These mutation groups were defined by Thomas et al. (2007), and grouped together mutations that occurred in the same gene, in the same functional domain of the encoded protein (e.g., kinase domain mutations or helical domain mutations of *PIK3CA*), or when a distinct phenotype was correlated with a specific mutation (e.g., the T790M mutation of *EGFR* known to be correlated with resistance to *EGFR* inhibitors). We ran the MCMC algorithm on sets of size k, for $2 \leq k \leq 10$. In each case we ran the MCMC for 10^7 iterations, and sampled a set every 10^4 iterations. All sets sampled with frequency at least 1% in this and all later experiments are reported in Supplement D.

We perform a permutation test to assess the significant of the results: the statistic is the weight $W(M)$ of the set and the null distribution was obtained by independently permuting the mutations for each mutation group among the patients, thus preserving the mutation frequency for each mutation group. We use the observed frequency of mutation rather than a fixed backgound mutation rate because we want to assess the significance of coverage and exclusivity of a set of mutation groups *given* the frequency of mutation of the single mutation groups in the set.[2] We identify a set of of 8 mutation groups (*BRAF_600-601, EGFR_ECD,*

[2] Using the background mutation rate, some mutation groups would be reported as significantly mutated when considered in

EGFR_KD, HRAS, KRAS, NRAS,PIK3CA_HD, PIK3CA_KD)[3] that is altered in $280/298$ of the patients (94%) with at least one mutation and has a total of 295 mutations ($p < 0.01$). The mutated genes are part of well known cancer pathways (Figure 3). There are many sets of size $k = 10$ that contain the set of size $k = 8$ above and also have high weight (see Supplemental Material). In particular, there are two sets of size $k = 10$ that are altered in $287/295$ (95%) of the patients, and have a total of 302 mutations ($p < 0.01$). These two sets include the above 8 mutation groups and (*JAK2, KIT*) and (*FGFR1, KIT*), respectively. We tested each pair of genes for mutual exclusivity with the (one-tailed) Fisher's exact test. No pair of genes show significant mutual exclusivity, with minimum q-value 0.492. Thus, a standard standard test does not report any of the mutation groups identified by our method.

2.4.2 Lung adenocarcinoma

We next analyzed a collection of 1013 somatic mutations identified in 623 sequenced genes from 188 lung adenocarcinoma patients from the Tumor Sequencing Project (Ding et al., 2008). In total, 356 genes were reported mutated in at least one patient. We ran the MCMC algorithm for sets of size $2 \leq k \leq 10$. When $k = 2$, the pair (*EGFR, KRAS*) is sampled 99% of the time. This pair is mutated in 90 patients with a coverage overlap $\omega(M) = 0$ indicating mutual exclusivity. When $k = 3$, the triplet (*EGFR, KRAS, STK11*) is sampled with frequency 8.4%. For $k \geq 4$, no set is sampled with frequency greater than 0.1%. The pairs (*EGFR, KRAS*) and (*EGFR, STK11*) are the most significant pairs in the mutual exclusivity test performed in (Ding et al., 2008), and thus it is not surprising that we also identify them. However, the pair (*KRAS, STK11*) is not reported as significant using their statistical test. Thus, the coverage and mutual exclusivity of the triplet (*EGFR, KRAS, STK11*) is a novel discovery.

We performed a permutation test as described in Section 2.4.1 to compare the significance of (*EGFR, KRAS*) and (*EGFR, KRAS, STK11*). The p-values obtained are 0.018 and 0.005, respectively. Thus, the triplet (*EGFR, KRAS, STK11*) is at least significant as the pair (*EGFR, KRAS*). The three genes *EGFR, KRAS and STK11* are all involved in the regulation of mTOR (Fig. 4), whose dysregulation has been reported as important in lung adenocarcinoma (Ding et al., 2008). In particular, *STK11* downregulates the mTOR pathway, and mTOR activation has been reported as significantly more frequent in tumors with gene alterations in either *EGFR* or *KRAS* (Conde et al., 2006). This supports the hypothesis that all three genes are upstream regulators of mTOR, explaining their observed exclusivity of mutations.

To identify additional gene sets, we removed the genes *EGFR, KRAS, STK11* and ran the MCMC algorithm again on the remaining genes. We sample the pair (*ATM, TP53*) with frequency 56%, and compute that the weight of the pair is significant ($p < 0.01$). *ATM* and *TP53* are known to directly interact (Khanna et al., 1998) and both genes are involved in the cell cycle checkpoint control (Chehab et al., 2000). Moreover this genes have no known role in mTOR regulation (Fig. 4) consistent with the observation that their mutations are not exclusive with those in the triplet above. Note that the pair (*ATM, TP53*) was not sampled with high frequency before removing *EGFR, KRAS*, and *STK11*. The reason is that the coverage of (*ATM, TP53*) is not as high as other pairs in the triplet: for example, the pair (*EGFR, KRAS*) covers 90 patients (with a coverage overlap of 0), while the pair (*ATM, TP53*) covers 76 patients (with a coverage overlap of 1). Although the exclusivity of both sets is high, their coverage is low ($< 60\%$), suggesting these gene sets are not complete driver pathways. We hypothesize that the coverage is low because: (i) somatic mutations were measured in only a small subset of genes; (ii) only single nucleotide mutations and small indels in these genes were measured, and other types of mutation (or epigenetic changes) might occur in the "unmutated"

isolation (because of their significant coverage). Thus, larger sets of mutation groups containing these individually significant mutation groups would also be reported as significant, even if the pattern of mutations in the set is not surprising after conditioning on the observed frequency of mutations of single mutation groups.

[3] The suffix of the mutation group identifies the positions of mutations in the gene, as in *BRAF_600-601*, or the mutated functional domain of the encoded protein, that is ECD for extracellular domain mutations, KD for kinase domain, and HD for helical domain, as described in Thomas et al. (2007)

patients. Either of these would reduce the coverage or imply that mutations in a superset of these genes were not measured.

We examined the overlap between the patients with mutations in (*ATM, TP53*) and those with mutations in (*EGFR, KRAS, STK11*). We found that the overlap was not significantly different from the expected number in a random dataset, suggesting that mutations in these two sets are not exclusive. This is consistent with our model, in which the two sets are part of two different pathways. While neither of these sets is mutated in $> 60\%$ of the patients, this does not imply that they are not part of important cancer pathways, for the same reasons regarding incomplete measurements outlined above.

2.4.3 Glioblastoma multiforme

We also applied the MCMC algorithm to 84 glioblastoma multiforme (GBM) patients from The Cancer Genome Atlas (The Cancer Genome Atlas Research Network, 2008). Somatic mutations in these patients[4] were measured in 601 genes. A total of 453 somatic single nucleotide mutations were identified and 223 genes were reported mutated in at least one patient. In addition, array copy number data was available for each of these 601 genes in every patient. We recorded a gene as somatically mutated in a patient if it was part of a focal copy number aberration identified in (The Cancer Genome Atlas Research Network, 2008), discarding copy number aberrations for which the sign of aberration (i.e., amplification or deletion) was not the same in at least 90% of the samples. Note that copy number aberrations (even focal aberrations) typically encompass more than one gene, and the boundaries of such aberrations vary across patients. Since we only collapse genes into "metagenes" if they are mutated in exactly the same patients, we will not collapse all of the genes in a focal copy number aberrations into a "metagene" if the genes in the aberrations vary across patients. Thus, the genes in overlapping, but not identical, aberrations will remain separate in our analysis. If our algorithm selects any of these genes in a high weight set, it might select the gene (or genes) that is altered in the largest number of patients, a behavior that is similar to "standard" copy number analysis methods that select the minimum common aberration. We ran the MCMC algorithm sets of sizes k ($2 \leq k \leq 10$) for 10^7 iterations, and sampling one set every 10^4 iterations.

For $k = 2$, the pair of genes sampled with the highest frequency is (*CDKN2B, CYP27B1*), sampled with frequency 18%. For $k = 3$, the most frequently sample set is (*CDKN2B, RB1, CYP27B1*), sampled with frequency 10%. The second most sampled pair (frequency 11%) was *CDKN2B* and a metagene containing six genes[5], and the second most sampled triplet (frequency 6%) was *CDKN2B, RB1*, and the same metagene. Moreover, the mutational profile of *CYP27B1* was nearly identical to a metagene: *CYP27B1* is mutated in all of the same patients as the metagene plus one additional patient with a single nucleotide mutations in *CYP27B1*. Because of this one extra mutation, *CYP27B1* was not merged into the metagene. Further, the six genes in the metagene are adjacent on the genome and are mutated by a copy number aberration (amplification) in all patients. This amplification also affects *CYP27B1* which is adjacent to these genes. The amplification was previously reported and the presumed target of the amplification is the gene *CDK4* (Wikman et al., 2005). Thus it is likely thus that the triplet (*CDKN2B, RB1, CDK4*) is the triplet of interest and the somatic mutation in *CYP27B1* identified in one patient does not have a biological impact. This example shows one of the advantages of the MCMC method: it allows one to identify additional "suboptimal" genes sets of high weight, and those whose weight is close to the highest. We performed a permutation test as described in Section 2.4.1 to compare the significance of (*CDKN2B, CDK4*) and (*CDKN2B, CDK4, RB1*). The p-values obtained are 0.1 and $< 10^{-2}$ respectively. Therefore the triplet (*CDKN2B, CDK4, RB1*) is at least as significant as the pair (*CDKN2B, CDK4*). *CDKN2B, RB1*, and *CDK4* are part of the *RB1* signaling pathway (Figure 5), and abnormalities in these genes are associated with shorter survival in glioblastoma

[4] Mutations were measured in 91 patients, but we removed 7 patients that were identified as hypermutated in (The Cancer Genome Atlas Research Network, 2008). These patients have higher observed mutation rate, presumably due to defective DNA repair.

[5] Genes in the metagene are *TSFM,MARCH9, TSPAN31, FAM119B, METTL1, CDK4, CENTG1*.

patients (Backlund et al., 2003). Thus, our method identifies a triplet of genes with a known association to survival rate directly from the somatic mutation data.

For $k \geq 4$, no set is sampled with frequency $\geq 0.2\%$. We remove the set (*CDKN2B, CDK4, RB1*) from the analysis, and ran the MCMC algorithm again. The pair (*TP53, CDKN2A*) is sampled with frequency 30% ($p < 0.01$). This pair is part of the p53 signaling pathway (Figure 5). As discussed in Section 2.4.2, the fact that this pair is sampled with high frequency only after removing (*CDKN2B, CDK4, RB1*) is likely due to the fact that not all genes and mutations in the pathways have been measured, resulting in different coverage for the two pathways. Finally, removing both (*CDKN2B, CDK4, RB1*) and (*TP53, CDKN2A*) we identify the pair (*NF1, EGFR*) sampled with frequency 44% ($p < 0.01$). *NF1* and *EGFR* are both part of the RTK pathway (Figure 5), that is involved in the proliferation, survival, and translation processes.

3 Discussion

We introduce two algorithms for finding mutated driver pathways in cancer *de novo* using somatic mutation data from many cancer patients. Our algorithms, called *De novo Driver Exclusivity* (**Dendrix**), find sets of genes that are mutated in many samples (high *coverage*) and that are rarely mutated together in the same patient (high *exclusivity*). These properties model the expected behavior of driver mutations in a pathway, or a "sub-pathway". We define a weight on sets of genes that measures how well a set exhibits these two properties. We show that finding the set M of genes with maximum weight is computationally difficult, derive conditions under which a greedy algorithm gives optimal solutions, and develop a Markov Chain Monte Carlo (MCMC) algorithm to sample sets of genes in proportion to their weight. Further, we prove that the Markov chain converges rapidly to the stationary distribution.

We applied our MCMC approach to three recent cancer sequencing studies: lung adenocarcinoma (Ding et al., 2008), glioblastoma (The Cancer Genome Atlas Research Network, 2008), and multiple cancer types (Thomas et al., 2007). In the latter dataset we identify a group of 8 mutations in 6 genes that are present at least once in a large fraction of patients and are largely exclusive. In the first two datasets, we identified groups of 2-3 genes with those properties. These gene sets include members of well-known cancer pathways including the Rb pathway, the p53 pathway, and the mTOR pathway. In the glioblastoma data, the mutations in the 3 genes that we identify have been previously associated with shorter survival (Backlund et al., 2003). Notably, we discover these pathways *de novo* from the mutation data without any prior biological knowledge of pathways or interactions between genes. However, it is also important to note that some of the genes that were measured in these datasets were selected because they were known to have a cancer phenotype, and thus there is some ascertainment bias in the finding that individual genes (or groups of genes) are mutated in many samples.

The results on the Thomas et al. (2007) data and on simulated data illustrate that our algorithm is able to identify relatively large sets of genes with high coverage and high exclusivity. However, in the lung adenocarcinoma and glioblastoma data, the sizes of gene sets that we identify is relatively modest. It is not yet possible to conclude whether this is real phenomenon or a consequence of limited data. For example, the numbers of patients and genes in these studies is relatively small, and the types of mutations that were measured was not comprehensive. For example, we examined only single nucleotide (and small indel) mutations in lung adenocarcinoma, and these plus copy number aberrations in the glioblastoma data. Other types of mutations, such as rearrangements, or even epigenetic changes could alter the function or expression of genes. In addition, considering mutation data at the level of individual genes might reduce the power to distinguish driver mutations from passenger mutations. Thus, it would be interesting to analyze the other datasets at "subgene" resolution to distinguish mutations at particular amino acid residues. We have shown that our algorithms are useful at a finer scale of resolution by introducing additional columns to the mutation matrix that correspond to protein domains, structural motifs, or other parts of a protein sequence.

The algorithms we presented assumed the availability of reasonably accurate mutation data. While the ability to measure somatic mutations from next-generation DNA sequencing data or microarrays is becoming more routine, there remain challenges in the identification of somatic mutations from these data with the incorrect prediction of somatic mutations (false positives) and the failure to identify genuine mutations (false negatives) (Meyerson et al., 2010). One particular source of false negatives is the heterogeneity of many tumor samples, which often include both normal cell admixture and subpopulations of tumor cells with potentially different sets of mutations. False negatives are a particular problem with samples with low tumor cellularity. Although the algorithms we propose are able to handle some false positives and false negatives, high rates of these errors would reduce the exclusivity and coverage, respectively, of a driver pathway. Moreover, this problem will be compounded if the genes in a driver pathway are mutated *only* in a subpopulation of tumor cells.

Our algorithms could be improved in several ways. First, we could include additional information in the scoring of mutations and gene sets. In the present analysis, we considered each mutation to have one of two states (mutated or normal). Extending our techniques to use additional information about the functional impact, or expression status, of each mutation is an interesting open problem. Second alternative weight functions $W(M)$ could be considered. For example, the inclusion of patient-specific mutation rates might provide a more refined way to analyze hypermutated patients. However, we note that some of our analytical results (e.g. the rapid mixing of the MCMC algorithm) relied on the particular form of the weight function $W(M)$ and these results would also require modification to maintain similar performance. Finally, the performance of our algorithm in complex situations involving multiple, overlapping high weight sets of genes requires further analysis. It is not yet clear whether such complex situations arise in cancer mutation data.

Our algorithms will be useful for analysis of whole genome or whole exome sequencing data from large sets of patients, and we anticipate that with these comprehensive datasets it will be possible to identify larger sets of driver genes. Such datasets will soon be available from The Cancer Genome Atlas (TCGA) and other large-scale cancer sequencing projects. We expect that the *de novo* techniques introduced here will complement existing methods for assessing enrichment of mutations in known pathways. As larger cancer datasets become available, it will be interesting to compare the exclusive gene sets identified by our techniques to known cancer pathways. A key questions in the analysis of these larger datasets is whether mutual exclusivity of driver mutations in genes in the same pathway is a widespread phenomenon, or whether it is a feature of particular genes, pathways, or cancer types. We anticipate that our algorithms will be helpful in addressing this question. In addition, it would be interesting to extend these ideas to other types of cancer genomics data, such as epigenetic alterations and structural aberrations. Finally, an intriguing future direction is to generalize these techniques to analyze combinations of (rare) germline variants in genetic association studies.

4 Methods

4.1 Complexity of the problem

The problems we are interested in are the Maximum Coverage Exclusive Submatrix Problem and the Maximum Weight Submatrix Problem (see Results for their definition). We show that these problems are computationally difficult (for proof, see Supplemental Material).

Theorem 1. *The Maximum Coverage Exclusive Submatrix Problem is NP-hard.*

Theorem 2. *The Maximum Weight Submatrix Problem is NP-hard.*

Note that our weight $W(M)$ is only one possible measure of the trade-off between coverage and exclusivity For example, another approach is to minimize the maximum number of genes that co-occur in a patient. The associated problem remains computationally difficult as shown in (Kuhn et al., 2005) (with additional generalizations in (Dom et al., 2006)).

4.2 A Greedy algorithm and Gene Independence Model

We propose the following greedy algorithm for the Maximum Weight Submatrix problem:
Greedy(k):

```
1. $M = \{g_1, g_2\} \leftarrow$ pair of genes that maximizes $W(\{g_1, g_2\})$.
2. For $i = 3, ..., k$ do:
    (a) Let $g^* = \arg\max_g W(M \cup \{g\})$.
    (b) $M \leftarrow M \cup \{g^*\}$.
3. return $M$.
```

The time complexity of the algorithm is $O\left(n^2 + kn\right) = O\left(n^2\right)$. We analyze the performance of the algorithm on mutation matrices generated from the following Gene Independence Model.

Definition 1. *Let A be an $m \times n$ mutation matrix such that $\hat{M}$ is the maximum weight column submatrix of A and $|\hat{M}| = k$. The matrix A satisfies the Gene Independence Model if and only if*

1. *Each gene $g \notin \hat{M}$ is mutated in each patient with probability p_g, independently of all other events, with $p_g \in [p_L, p_U]$ for all g.*
2. *$W(\hat{M})$ is $\Omega\left(|\mathcal{S}|\right)$, i.e. $W(\hat{M}) = rm$ for a constant $r, 0 < r \leq 1$;*
3. *For all ℓ, any subset $M \subset \hat{M}$ of cardinality $|M| = \ell$ satisfies: $W(M) \leq \frac{\ell+d}{k} W(\hat{M})$, for a constant $0 \leq d < 1$.*

We show that the greedy algorithm above will produce the optimal solution with high probability for any mutation matrix generated from the Gene Independence Model, when the number of rows (patients) is sufficiently large.

Theorem 3. *Suppose $\varepsilon > 0$ and A is an $m \times n$ mutation matrix generated from the Gene Independence Model that satisfies*

$$m \geq \left(1 + \frac{\varepsilon}{2}\right) \log n \times \max \left\{ \left(\frac{2r}{k} - 2(p_U - p_L^2)\right)^{-2}, \left(\frac{r(1-d)}{k} - p_U + \frac{4rp_L}{k}\right)^{-2} \right\}. \tag{1}$$

Then the greedy algorithm identifies the $m \times k$ column submatrix $\hat{M}$ with maximum weight $W(\hat{M})$ with probability at least $1 - 2n^{-\varepsilon}$.

For proof of Theorem 3 see Supplemental Material.

4.3 Markov Chain Monte Carlo (MCMC) algorithm

The basic idea of MCMC is to build a Markov chain whose states are the possible configurations and to define transitions between states according to some criterion. If the number of states is finite and the transitions are defined such that the Markov chain is ergodic, then the Markov chain converges to a unique stationary distribution. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) gives a general method for designing transition probabilities that gives a desired stationary distribution on the state space. However, the Metropolis-Hastings method does not guarantee fast convergence of the chain, which is a necessary condition for practical use of this method. In fact, if the chain converges slowly then it may take an impractically long time before the chain samples from the desired distribution. Defining transition probabilities so that the chain converges rapidly to the stationary distribution remains a challenging task.

Despite significant progress in recent years in developing mathematical tools for analyzing the convergence time (Randall, 2006), our ability to analyze useful chains is still limited, and in practice most MCMC algorithms rely on simulations to provide evidence of convergence to stationarity (Gilks, 1998).

We use a Metropolis-Hastings algorithm to sample sets $M \subseteq \mathcal{G}$ of k genes with a stationary distribution that is proportional to $e^{cW(M)}$ for some $c > 0$, and we show that the resulting chain converges rapidly.

Initialization: Choose an arbitrary subset M_0 of k genes in $\mathcal{G}$ (the set of all genes).
Iteration: for $t = 1, 2, \ldots$ obtain M_{t+1} from M_t as follows:

1. Choose a gene w uniformly at random from $\mathcal{G}$.
2. Choose v uniformly at random from M_t.
3. Let $P(M_t, w, v) = \min[1, e^{cW(M_t - \{v\} + \{w\}) - cW(M_t)}]$. [6]
4. With probability $P(M_t, w, v)$ set $M_{t+1} = M_t - \{v\} + \{w\}$, else $M_{t+1} = M_t$.

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. We are able to analyze the mixing time of the chain because we do not restrict the set of states that the chain can visit, focusing instead on the desired stationary probabilities of the various states.

Let $P^t_{I,M}$ be the transition probability from initial state I to state M in t steps of the Markov chain. We measure the distance between the distribution of the chain at time t and the stationary distribution by the *variation distance* between the two distribution:

$$\Delta_I(M) = \frac{1}{2} \sum_{M \in \mathcal{M}_k} |P^t_{I,M} - \pi(M)|.$$

The ϵ*-mixing time* of the chain is

$$\tau(\epsilon) = \max_I \min\{t \mid \Delta_I(M) \leq \epsilon\}.$$

A chain is *rapidly mixing* if $\tau(\epsilon)$ is bounded by a polynomial in the size of the problem ($m = |\mathcal{S}|$ and $n = |\mathcal{G}|$ in our case) and $\log \epsilon^{-1}$.

We show that there is a non-trivial interval of values for c for which the chain is rapidly mixing (for proof, see Supplemental Material). Our proof uses a *path coupling* argument (Bubley and Dyer, 1997). In path coupling we define coupling only on pairs of adjacent states in the Markov chain. Let M_t and M'_t be the states of two copies of the Markov chain at time t, and assume that $M_t = M'_t + \{z\} - \{y\}$ (thus, the two states are adjacent in the Markov chain). We use the following coupling: assume that the first chain chooses $w \in \mathcal{G}$ and $v \in M_t$ in computing the transition to M_{t+1}. The second chain uses the same w, and if $v \in M_t \cap M_{t+1}$ it also uses the same v. Otherwise, if in the first chain $v = y$, then the second chain uses $v = z$. If $P(M'_t, w, v) \leq P(M_t, w, v)$ and the first chain performs a switch then the second chain performs a switch with probability $P(M'_t, w, v)/P(M_t, w, v)$. If $P(M'_t, w, v) \geq P(M_t, w, v)$ then the second chain performs a switch whenever the first chain does, and when the first chain did not perform a switch the second chain switches with probability $P(M'_t, w, v) - P(M_t, w, v)$. Our analysis applies the following simple version of path coupling adapted to our setting (see (Bubley and Dyer, 1997) and (Mitzenmacher and Upfal, 2005)):

[6] For ease of notation in this section given sets A and B we denote their difference by $A - B = \{x \mid x \in A \text{ and } x \notin B\}$, and their union by $A + B = \{x \mid x \in A \text{ or } x \in B\}$.

Theorem 4. *Let $\phi_t = |M_t - M'_t|$, and assume that for some constant $0 < \beta < 1$, $E[\phi_{t+1}| \ \phi_t = 1] \leq \beta$, then the mixing time*

$$\tau(\epsilon) \leq \frac{k \log(k\epsilon^{-1})}{1-\beta}.$$

Using the above, we prove the following convergence result for our chain.

Theorem 5. *The MCMC is rapidly mixing for some $c > 0$.*

Theorem 5 gives a range of values of c where the resulting chain will converge rapidly. We explored different values of c, and use the $c = 0.5$, which we found empirically to give the best tradeoff between the exploration of different sets and the convergence to sets with high weight $W(M)$ on simulated data. We use $c = 0.5$ for both the experiments on both simulated data and real cancer mutation data described below.

4.3.1 Extension to Multiple Sets of Mutated Genes

There are multiple capabilities that a cell has to acquire in order to become a cancer cell; for example, Hahn and Weinberg (2002) describe 6 capabilities. Thus, we expect that a small number of pathways will be mutated, and in each pathway the mutations in the corresponding genes will have both high exclusivity and high coverage. We aim to recover sets of genes in each of these pathways. If the sets of genes in each pathway are disjoint, then an iterative procedure will suffice: once we identify a set M with high weight, we remove the genes in M from the analysis and look for high weight sets in the reduced mutation matrix. Thus, if two sets M_1 and M_2 of genes are disjoint and have high weight then the iterative procedure finds both, because exclusivity is required only within and not between sets. If instead M_1 and M_2 have genes in common, then removing one of the them could remove part of another. If the intersection is small, we will still be able to identify the remaining part of the other set. The problem of identifying two sets M_1 and M_2 of genes that both have high exclusivity and high coverage (but with no exclusivity between them) and have a number of genes in common is an interesting open problem.

4.4 Cancer data

In all tumor patients we consider, we use both single nucleotide mutations and small indels reported in the original studies (Ding et al., 2008; The Cancer Genome Atlas Research Network, 2008; Thomas et al., 2007). For glioblastoma patients, we also consider focal copy number aberrations identified in the original study (The Cancer Genome Atlas Research Network, 2008), discarding copy number aberrations for which the sign of aberration (i.e., amplification or deletion) was not the same in at least 90% of the samples.

We reduce the size of the mutation matrix by combining genes that are mutated in *exactly* the same patients into larger "metagenes". For example, suppose there exists a set $S = \{g_1, g_2\}$ of two genes that are mutated in the same set of patients. Two sets X and Y with $X \backslash Y = \{g_1\}$ and $Y \backslash X = \{g_2\}$ satisfy $W(X) = W(Y)$. Thus, both sets have the same probability. The same result holds when $|S| > 2$. To improve the efficiency of the MCMC sampling procedure we replace a maximal set of genes $T = \{g_1, g_2, \dots\}$ that are mutated in the same patients with a single "metagene" g_T whose mutations are the same patients. Copy number aberrations typically encompass more than one gene, and the boundaries of such aberrations vary across patients. Since we only collapse genes into "metagenes" if they are mutated in exactly the same patients, we will not collapse all of the genes in a copy number aberrations into a metagene if the genes in the metagene vary across patients.

4.5 Software

A Python implementation of **Dendrix** (*De novo Driver Exclusivity*) is available at `http://cs.brown.edu/people/braphael/software.html`.

Acknowledgements

We thank the anonymous reviewers for helpful suggestions that improved the manuscript. This work is supported by NSF grant IIS-1016648, the Department of Defense Breast Cancer Research Program, the Alfred P. Sloan Foundation, and the Susan G. Komen Foundation. BJR is also supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

List of Figures

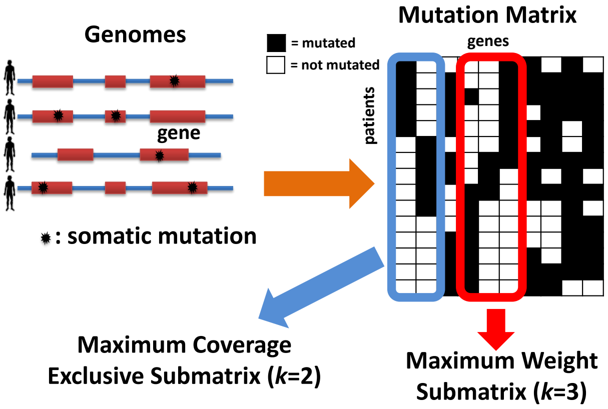

Figure 1: Somatic mutations in multiple patients are represented in a mutation matrix. Gene sets are identified as exclusive submatrices or high weight submatrices.

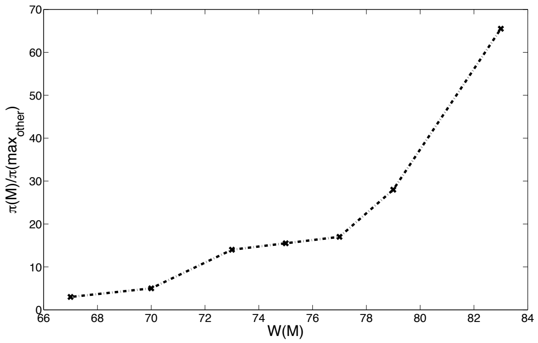

Figure 2: Ratio between the sampled frequency $\pi(M)$ of the maximum weight set, and the maximum frequency $\pi(\max_{\text{other}})$ of any other set in the sample for different values of $W(M)$.

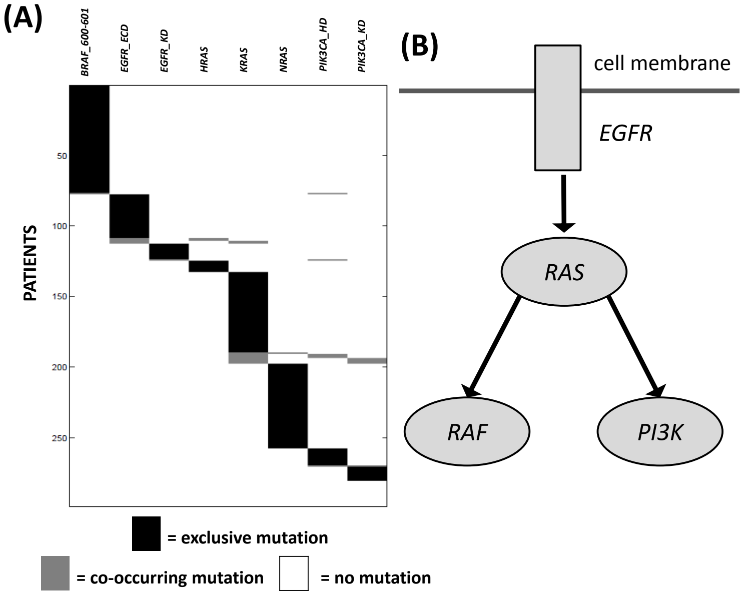

Figure 3: (A) High weight submatrix of 8 genes in the somatic mutations data from multiple cancer types (Thomas et al., 2007). Black bars indicate exclusive mutations, while gray bars indicate co-occurring mutations. (B) Location of identified genes in known pathway. Interactions in pathway are as reported in Ding et al. (2008).

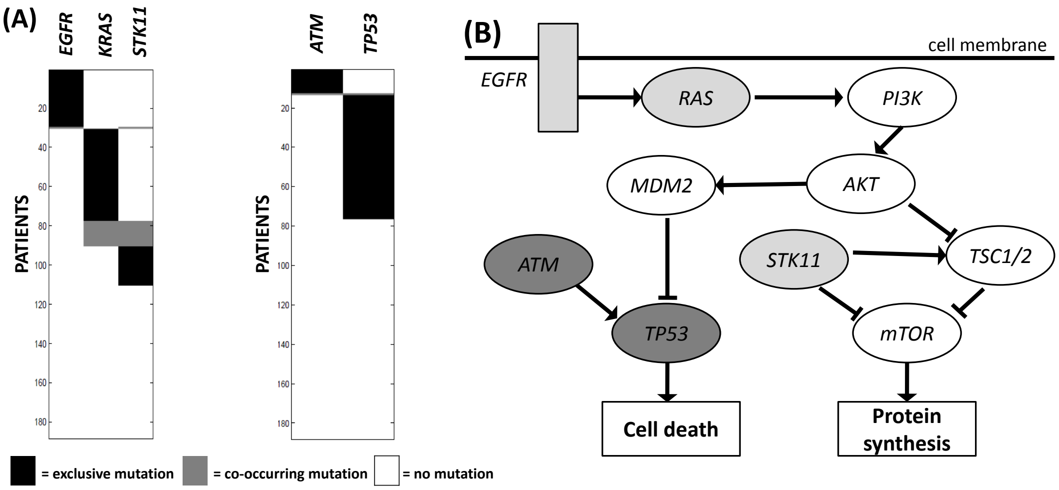

Figure 4: (A) High weight submatrices of two and three genes in the lung adenocarcinoma data. Black bars indicate exclusive mutations, while gray bars indicate co-occurring mutations. Rows (patients) are ordered differently for each submatrix, to illustrate exclusivity and co-occurrence. (B) Location of gene sets in known pathways reveals that the triplet of genes codes for proteins in the mTOR signalling pathway (light gray nodes) and the pair (ATM, TP53) corresponds to interacting proteins in the cell cycle pathway (dark gray nodes). Interactions in pathway are as reported in Ding et al. (2008).

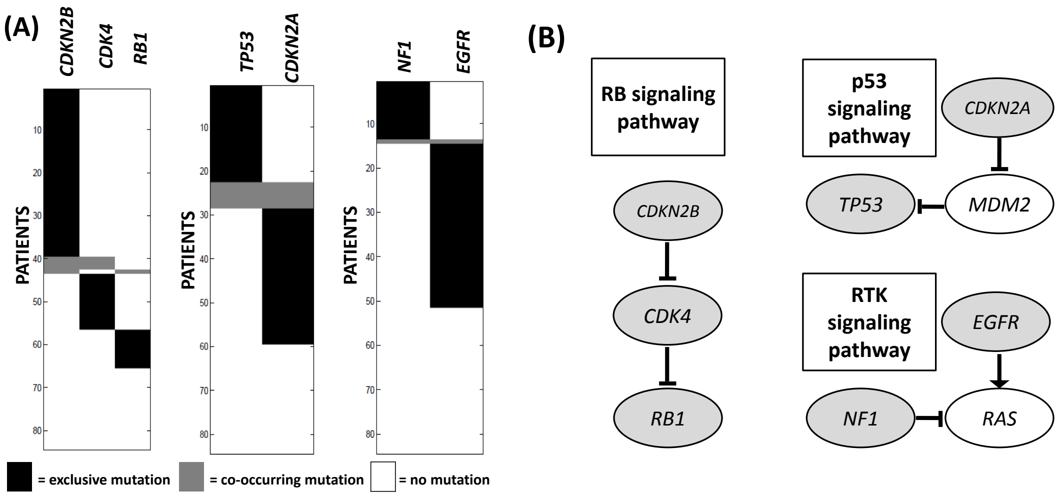

Figure 5: (A) High weight submatrices of two and three genes in the glioblastoma data. Black bars indicate exclusive mutations, while gray bars indicate co-occurring mutations. Rows (patients) are ordered differently for each submatrix, to illustrate exclusivity and co-occurrence. (B) Location of identified genes in known pathways. Interactions in pathways are as reported in The Cancer Genome Atlas Research Network (2008).

	M_1	M_2	$\max_{\text{other}}$	$\text{avg}_{\text{other}}$
$\tilde{\pi}(\cdot)$	24.5	8.6	0.9	1.6×10^{-4}
$W(\cdot)$	80	78	73	56

Table 1: MCMC results on simulated data. $\tilde{\pi}(M_i)$ is the frequency of M_i, $\tilde{\pi}(\max_{\text{other}})$ is the maximum frequency with which set different from M_1 and M_2 is sampled, and $\tilde{\pi}(\text{avg}_{\text{other}})$ is the average frequency with which a set different from M_1 and M_2 is sampled.

References

Backlund, L. M., Nilsson, B. R., Goike, H. M., Schmidt, E. E., Liu, L., Ichimura, K., and Collins, V. P. (2003). Short postoperative survival for glioblastoma patients with a dysfunctional Rb1 pathway in combination with no wild-type PTEN. *Clin. Cancer Res.*, 9:4151–4158.

Bansal, V., Halpern, A. L., Axelrod, N., and Bafna, V. (2008). An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.*, 18:1336–1346.

Ben-Dor, A., Chor, B., Karp, R. M., and Yakhini, Z. (2003). Discovering local structure in gene expression data: The order-preserving submatrix problem. *Journal of Computational Biology*, 10(3/4):373–384.

Boca, S. M., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., and Parmigiani, G. (2010). Patient-oriented gene set analysis for cancer mutation data. *Genome Biol.*, 11:R112.

Bradley, J. R. and Farnsworth, D. L. (2009). Testing for Mutual Exclusivity. *Journal of Applied Statistics*, 36:1307–1314.

Bubley, R. and Dyer, M. (1997). Path coupling: A technique for proving rapid mixing in markov chains. In *FOCS '97: Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, page 223, Washington, DC, USA. IEEE Computer Society.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, 5:e8918.

Chehab, N. H., Malikzay, A., Appel, M., and Halazonetis, T. D. (2000). Chk2/hCds1 functions as a DNA damage checkpoint in G(1) by stabilizing p53. *Genes Dev.*, 14:278–288.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *ISMB*, pages 93–103.

Conde, E., Angulo, B., Tang, M., Morente, M., Torres-Lanzas, J., Lopez-Encuentra, A., Lopez-Rios, F., and Sanchez-Cespedes, M. (2006). Molecular context of the EGFR mutations: evidence for the activation of mTOR/S6K signaling. *Clin. Cancer Res.*, 12:710–717.

Deguchi, K. and Gilliland, D. G. (2002). Cooperativity between mutations in tyrosine kinases and in hematopoietic transcription factors in AML. *Leukemia*, 16:740–744.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer, C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. A., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. A., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. A., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. A., Meyerson, M., and Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455:1069–1075.

Dom, M., Guo, J., Niedermeier, R., and Wernicke, S. (2006). Minimum membership set covering and the consecutive ones property. In *SWAT*, pages 339–350.

Efroni, S., Ben-Hamo, R., Edmonson, M., Greenblum, S., Schaefer, C. F., and Buetow, K. H. (2011). Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS ONE*, 6:e14437.

Garey, M. R. and Johnson, D. S. (1990). *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Gazdar, A. F., Shigematsu, H., Herz, J., and Minna, J. D. (2004). Mutations and addiction to EGFR: the Achilles 'heal' of lung cancers? *Trends Mol Med*, 10:481–486.

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, 97:12079–12084.

Gilks, W. (1998). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Hahn, W. C. and Weinberg, R. A. (2002). Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer*, 2:331–341.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Ikeda, T., Yoshinaga, K., Suzuki, A., Sakurada, A., Ohmori, H., and Horii, A. (2000). Anticorresponding mutations of the KRAS and PTEN genes in human endometrial cancer. *Oncol. Rep.*, 7:567–570.

International cancer genome consortium (2010). International network of cancer genome projects. *Nature*, 464:993–998.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, 37:D412–416.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S. M., Fu, B., Lin, M. T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., and Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321:1801–1806.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37:D767–772.

Khanna, K. K., Keating, K. E., Kozlov, S., Scott, S., Gatei, M., Hobson, K., Taya, Y., Gabrielli, B., Chan, D., Lees-Miller, S. P., and Lavin, M. F. (1998). ATM associates with and phosphorylates p53: mapping the region of interaction. *Nat. Genet.*, 20:398–400.

Kim, Y., Wuchty, S., and Przytycka, T. (2010). Simultaneous Identification of Causal Genes and Dys-Regulated Pathways in Complex Diseases. In *Research in Computational Molecular Biology*, pages 263–280. Springer.

Kuhn, F., von Rickenbach, P., Wattenhofer, R., Welzl, E., and Zollinger, A. (2005). Interference in cellular networks: The minimum membership set cover problem. In *COCOON*, pages 188–198.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform*, 1:24–45.

Mao, J., To, M., Perez-Losada, J., Wu, D., Del Rosario, R., and Balmain, A. (2004). Mutually exclusive mutations of the Pten and ras pathways in skin tumor progression. *Genes & development*, 18(15):1800.

Mardis, E. R. and Wilson, R. K. (2009). Cancer genome sequencing: a review. *Hum. Mol. Genet.*, 18:R163–168.

McCormick, F. (1999). Signalling networks that cause cancer. *Trends Cell Biol.*, 9:M53–56.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyer, I. M. and Miklos, I. (2007). SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, 3:e149.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, 11:685–696.

Mitzenmacher, M. and Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA.

Murali, T. M. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput*, pages 77–88.

Randall, D. (2006). Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science and Engineering*, 8(2):30–41.

Segal, E., Battle, A., and Koller, D. (2003). Decomposing gene expression into cellular processes. In *Pacific Symposium on Biocomputing*, pages 89–100.

Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. In *ISMB*, pages 136–144.

The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8.

Thomas, R. K., Baker, A. C., Debiasi, R. M., Winckler, W., Laframboise, T., Lin, W. M., Wang, M., Feng, W., Zander, T., MacConaill, L., Macconnaill, L. E., Lee, J. C., Nicoletti, R., Hatton, C., Goyette, M., Girard, L., Majmudar, K., Ziaugra, L., Wong, K. K., Gabriel, S., Beroukhim, R., Peyton, M., Barretina, J., Dutt, A., Emery, C., Greulich, H., Shah, K., Sasaki, H., Gazdar, A., Minna, J., Armstrong, S. A., Mellinghoff, I. K., Hodi, F. S., Dranoff, G., Mischel, P. S., Cloughesy, T. F., Nelson, S. F., Liau, L. M., Mertz, K., Rubin, M. A., Moch, H., Loda, M., Catalona, W., Fletcher, J., Signoretti, S., Kaye, F., Anderson, K. C., Demetri, G. D., Dummer, R., Wagner, S., Herlyn, M., Sellers, W. R., Meyerson, M., and Garraway, L. A. (2007). High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.*, 39:347–351.

Ulitsky, I., Karp, R. M., and Shamir, R. (2008). Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In *RECOMB'08: Proceedings of the 12th annual international conference on Research in computational molecular biology*, pages 347–359, Berlin, Heidelberg. Springer-Verlag.

Vandin, F., Upfal, E., and Raphael, B. (2010). Algorithms for Detecting Significantly Mutated Pathways in Cancer. In *Research in Computational Molecular Biology*, pages 506–521. Springer.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M. L., Teague, J., Bignell, G., Butler, A., Cho, J., Dalgliesh, G. L., Galappaththige, D., Greenman, C., Hardy, C., Jia, M., Latimer, C., Lau, K. W., Marshall, J., McLaren, S., Menzies, A., Mudie, L., Stebbings, L., Largaespada, D. A., Wessels, L. F., Richard, S., Kahnoski, R. J., Anema, J., Tuveson, D. A., Perez-Mancera, P. A., Mustonen, V., Fischer, A., Adams, D. J., Rust, A., Chan-on, W., Subimerb, C., Dykema, K., Furge, K., Campbell, P. J., Teh, B. T., Stratton, M. R., and Futreal, P. A. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469:539–542.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.*, 10:789–799.

Wikman, H., Nymark, P., Vayrynen, A., Jarmalaite, S., Kallioniemi, A., Salmenkivi, K., Vainio-Siukola, K., Husgafvel-Pursiainen, K., Knuutila, S., Wolf, M., and Anttila, S. (2005). CDK4 is a probable target gene in a novel amplicon at 12q13.3-q14.1 in lung cancer. *Genes Chromosomes Cancer*, 42:193–199.

Yamamoto, H., Shigematsu, H., Nomura, M., Lockwood, W. W., Sato, M., Okumura, N., Soh, J., Suzuki, M., Wistuba, I. I., Fong, K. M., Lee, H., Toyooka, S., Date, H., Lam, W. L., Minna, J. D., and Gazdar, A. F. (2008). PIK3CA mutations and copy number gains in human lung cancers. *Cancer Res.*, 68:6913–6921.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol. Biol. Evol.*, 14:717–724.

Yeang, C., McCormick, F., and Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*.