BROWN UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE &
CENTER FOR COMPUTATIONAL MOLECULAR BIOLOGY

Methods for Identifying Driver Pathways in Cancer

Author:
Mark DM LEISERSON

Research Advisor:
Prof. Benjamin RAPHAEL

Committee:
Prof. Rodrigo FONSECA
Prof. Sohini RAMACHANDRAN
Prof. Eli UPFAL

Wednesday, April 24, 2013

Contents

Author Summary

My research during my first two years at Brown has focused on the problem of identifying driver pathways in cancer. Specifically, I have worked on developing algorithms to discover driver mutations from next-generation sequencing data. For my research comps defense, I present a paper – co-authored with Prof. Raphael and with researchers at Tel-Aviv University – titled "Simultaneous Identification of Driver Pathways in Cancer" which has been accepted for publication by the journal *PLoS Computational Biology*. For this paper, I both developed a new algorithm for identifying driver pathways and applied this approach to data from large cancer sequencing studies.

In addition to my research developing algorithms for the analysis of mutation data, I have also applied my computational expertise via collaborations with biology researchers. Working with Prof. Raphael, Fabio Vandin, and Hsin-Ta Wu, I performed analysis of exclusive mutations in a cohort of acute myeloid leukemia (AML) patients as part of The Cancer Genome Atlas (TCGA) Research Network's AML project. The TCGA's paper on this project was recently accepted for publication by the *New England Journal of Medicine*, and I created a summary figure of my analysis that appears in the main text. I include the relevant reference in Publications below. Also, in collaboration with Prof. Raphael and researchers at Washington University in St. Louis, I have performed network analysis of somatic and germline mutations in ovarian cancer patients. This work is in preparation for submission to *Nature Genetics*, and my analysis is featured in the results.

Publications

1. **Mark DM Leiserson**, Dima Blokh, Roded Sharan, and Benjamin J Raphael. Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comp Bio* (in press).

2. The Cancer Genome Atlas Research Network. The genomic and epigenomic landscape of adult de novo Acute Myeloid Leukemia (AML). *New England Journal of Medicine* (in press).

Presentations

1. **Mark DM Leiserson**, Hsin-Ta Wu, Dima Blokh, Fabio Vandin, Roded Sharan, and Benjamin J Raphael. Methods for Identifying Driver Pathways in Cancer, Poster Presentation, *Beyond the Genome*, Boston, MA: September 2012.

2. **Mark DM Leiserson**, Hsin-Ta Wu, Fabio Vandin, and Benjamin J Raphael. Network Analysis of Mutations Across Cancer Types, Poster Presentation, *The Research in Computational Molecular Biology Conference (RECOMB)*, Beijing, China: April 2013.

3. **Mark DM Leiserson**, Hsin-Ta Wu, Fabio Vandin, and Benjamin J Raphael. Network and Pathway Analysis of Mutations Across Cancer Types, Poster Presentation, *The Biology of Genomes Conference*, Cold Spring Harbor, NY: May 2013.

Simultaneous Identification of Multiple Driver Pathways in Cancer *

Mark D.M. Leiserson[1], Dima Blokh[2], Roded Sharan[2, †], and Benjamin J. Raphael[1, †]

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

[2]*Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel*

[†]*Equal contribution*

Abstract

Distinguishing the somatic mutations responsible for cancer (*driver* mutations) from random, *passenger* mutations is a key challenge in cancer genomics. Driver mutations generally target cellular signaling and regulatory pathways consisting of multiple genes. This heterogeneity complicates the identification of driver mutations by their recurrence across samples, as different combinations of mutations in driver pathways are observed in different samples.

We introduce the Multi-Dendrix algorithm for the simultaneous identification of multiple driver pathways *de novo* in somatic mutation data from a cohort of cancer samples. The algorithm relies on two combinatorial properties of mutations in a driver pathway: high coverage and mutual exclusivity. We derive an integer linear program that finds set of mutations exhibiting these properties. We apply Multi-Dendrix to somatic mutations from glioblastoma, breast cancer, and lung cancer samples. Multi-Dendrix identifies sets of mutations in genes that overlap with known pathways – including Rb, p53, PI(3)K, and cell cycle pathways – and also novel sets of mutually exclusive mutations, including mutations in several transcription factors or other genes involved in transcriptional regulation. These sets are discovered directly from mutation data with *no prior knowledge* of pathways or gene interactions. We show that Multi-Dendrix outperforms other algorithms for identifying combinations of mutations and is also orders of magnitude faster on genome-scale data.

Software available at: http://compbio.cs.brown.edu/software.

Author Summary

Cancer is a disease driven largely by the accumulation of somatic mutations during the lifetime of an individual. The declining costs of genome sequencing now permit the measurement of somatic mutations in hundreds of cancer genomes. A key challenge is to distinguish *driver mutations* responsible for cancer from random *passenger* mutations. This challenge is compounded by the observation that different combinations of driver mutations are observed in different patients with the same cancer type.

One reason for this heterogeneity is that driver mutations target signaling and regulatory pathways which have multiple points of failure. We introduce an algorithm, Multi-Dendrix, to find these pathways solely from patterns of mutual exclusivity between mutations across a cohort of patients. Unlike earlier approaches, we simultaneously find multiple pathways, an essential feature for analyzing cancer genomes where multiple pathways are typically perturbed. We apply our algorithm to mutation data from hundreds of glioblastoma, breast cancer, and lung adenocarcinoma patients. We identify sets of interacting genes that overlap known pathways, and gene sets containing subtype-specific mutations. These results show that multiple cancer

*The work in this section has been accepted for publication by the journal *PLoS Comp Bio*.

pathways can be identified directly from patterns in mutation data, and provide an approach to analyze the ever-growing cancer mutation datasets.

Introduction

Cancer is a disease driven in part by somatic mutations that accumulate during the lifetime of an individual. The declining costs of genome sequencing now permit the measurement of these somatic mutations in large numbers of cancer genomes. Projects such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) are now undertaking this task in hundreds of samples from dozens of cancer types. A key challenge in interpreting these data is to distinguish the functional *driver* mutations important for cancer development from random *passenger* mutations that have no consequence for cancer. The ultimate determinant of whether a mutation is a driver or a passenger is to test its biological function. However, because the ability to detect somatic mutations currently far exceeds the ability to validate experimentally their function, computational approaches that predict driver mutations are an urgent priority. One approach is to directly predict the functional impact of somatic mutations using additional biological knowledge from evolutionary conservation, protein structure, etc. and a number of methods implementing this approach have been introduced (see [1–4]). These methods are successful in predicting the impact of some mutations, but generally do not integrate information across different types of mutations (single nucleotide, indels, larger copy number aberrations, etc.); moreover, these methods are less successful for less conserved/studied proteins.

Given the declining costs of DNA sequencing, a standard approach to distinguish driver from passenger mutations is to identify *recurrent* mutations, whose observed frequency in a large cohort of cancer patients is much higher than expected [5, 6]. Nearly all cancer genome sequencing papers, including those from TCGA [7–10] and other projects [5, 11, 12], report a list of significantly mutated genes. However, driver mutations vary greatly between cancer patients – even those with the same (sub)type of cancer – and this heterogeneity significantly reduces the statistical power to detect driver mutations by tests of recurrence. One of the main biological explanations for this mutational heterogeneity is that driver mutations target not only individual genomic loci (e.g. nucleotides or genes), but also target groups of genes in cellular signaling and regulatory pathways. Consequently, different cancer patients may harbor mutations in different members of a pathway important for cancer development. Thus, in addition to testing individual loci, or genes, for recurrent mutation in a cohort of patients, researchers also test whether groups of genes are recurrently mutated. Since exhaustive testing of all groups of genes is not possible without prohibitively large sample sizes (due to the necessary multiple hypothesis testing correction), current approaches focus on groups of genes defined by prior biological knowledge, such as known pathways (e.g. from KEGG [13]) or functional groups (e.g. from GO [14]), and methods have been introduced to look for enrichment in such pre-defined groups of genes (e.g. [15–17]). More recently, methods that identify recurrently mutated subnetworks in protein-protein interaction networks have also been developed, such as NetBox [18], MeMO [19], HotNet [20], and EnrichNet [21].

Knowledge of gene and protein interactions in humans remain incomplete, and most existing pathway databases and interaction networks do not precisely represent the pathways and interactions that occur in a particular cancer cell. Thus, restricting attention to only those combinations of mutations recorded in these data sources may limit the possibility for novel biological discoveries. Thus algorithms that do not make this restriction – but also avoid the multiple hypothesis testing problems associated with exhaustive enumeration – are desirable. Recently, the RME [22] and De novo Driver Exclusivity (Dendrix) [23] algorithms were introduced to discover *driver pathways* using combinatorial constraints derived from biological knowledge of how driver mutations appear in pathways [24, 25]. In particular, each cancer patient contains a relatively small number of driver mutations, and these mutations perturb multiple cellular pathways. Thus, each driver

pathway will contain approximately one driver mutation per patient. This leads to a pattern of *mutual exclusivity* between mutations in different genes in the pathway. In addition, an important driver pathway should be mutated in many patients, or have high *coverage* by mutations. Thus, driver pathways correspond to sets of genes that are mutated in many patients, but whose mutations are mutually exclusive, or approximately so. We emphasize that the *driver pathways exhibiting patterns of mutually exclusivity and high coverage are generally smaller and more focused* than most pathways annotated in the literature and pathway databases. The latter typically contain many genes and perform multiple different functions; e.g. the "cell cycle" pathway in KEGG contains 143 genes. It is well known that co-occurring (i.e., not exclusive) mutations are observed in these larger, multifunctional biological pathways [25]. The RME and Dendrix algorithms use different approaches to find sets of genes with high coverage and mutual exclusivity: RME builds sets of genes from pairwise scores of exclusivity, while Dendrix computes a single score for the mutual exclusivity of a set of genes, and finds the highest scoring set. The aforementioned MeMO algorithm [19] also considers mutual exclusivity between mutations, but only for pairs of genes that have recorded interactions in a protein-protein interaction network. Thus, MeMO does not attempt to identify driver pathways *de novo* and can only define subnetworks in existing interaction networks. While many of the strongest signals of mutual exclusivity are between genes with known interactions, below we show examples in cancer data of mutual exclusive mutations between genes with no known direct iterations.

The two existing *de novo* algorithms, RME and Dendrix, consider the detection of only a *single* driver pathway from the pattern of mutual exclusivity between mutations. However, it is well known that mutations in several pathways are generally required for cancer [26]. There is little reason to assume that mutations in different pathways will be mutually exclusive, and in contrast may exhibit significant patterns of co-occurrence across patients. Multiple pathways may be discovered using these algorithms by running the algorithm iteratively, removing the genes found in each previous iteration, and such an approach was employed for Dendrix [23]. However, such an iterative approach is not guaranteed to yield the optimal set of pathways.

Here we extend the Dendrix algorithm in three ways. First, we formulate the problem of finding exclusive, or approximately exclusive, sets of genes with high coverage as an integer linear program (ILP). This formulation allows us to find optimal driver pathways of various sizes directly – in contrast to the greedy approximation and Markov Chain Monte Carlo algorithms employed in Dendrix. Second, we generalize the ILP to *simultaneously* find *multiple* driver pathways. Third, we augment the core algorithm with additional analyses including: examining gene sets for subtype-specific mutations, summarizing stability of results across different number and size of pathways, and imposing greater exclusivity of gene sets.

We apply the new algorithm, called Multi-Dendrix, to four somatic mutation datasets: whole-exome and copy number array data in 261 glioblastoma (GBM) patients from The Cancer Genome Atlas (TCGA) [7], whole-exome and copy number array data in 507 breast cancer (BRCA) patients from TCGA [8], 601 sequenced genes in 84 patients with glioblastoma multiforme (GBM) from TCGA [7] and 623 sequenced genes in 188 patients with lung Adenocarcinoma [27]. In each dataset Multi-Dendrix finds biologically interesting groups of genes that are highly exclusive, and where each group is mutated in many patients. In all datasets these include groups of genes that are members of known pathways critical to cancer development including: Rb, p53, and RTK/RAS/PI(3)K signaling pathways in GBM and p53 and PI(3)K/AKT signaling in breast cancer. Multi-Dendrix successfully recovers these pathways solely from the pattern of mutual exclusivity and *without any prior information* about the interactions between these genes. Moreover, Multi-Dendrix also identifies mutations that are mutually exclusive with these well-known pathways, and potentially represent novel interactions or crosstalk between pathways. Notable examples include mutual exclusivity between: mutations in PI(3)K signaling pathway and amplification of PRDM2 (and PDPN) in glioblastoma; mutations in p53, GATA3 and cadherin genes in breast cancer.

Finally, we compare Multi-Dendrix to an alternative approach of iteratively applying Dendrix [23] or RME [22], two other algorithms that search for mutually exclusive sets. We show that these iterative ap-

proaches typically fail to find an optimal set of pathways on simulated data, while Multi-Dendrix finds the correct pathways even in the presence of a large number of false positive mutations. On real cancer sequencing data, the groups of genes found by Multi-Dendrix include more genes with known biological interactions. Moreover, Multi-Dendrix is orders of magnitude faster than these other algorithms, allowing Multi-Dendrix to scale to the latest whole-exome datasets on hundreds of samples, which are largely beyond the capabilities of Dendrix and RME. Multi-Dendrix is a novel and practical approach to finding multiple groups of mutually exclusive mutations, and complements other approaches that predict combinations of driver mutations using biological knowledge of pathways, interaction networks, protein structure, or protein sequence conservation.

Results

Multi-Dendrix algorithm

The Multi-Dendrix algorithm takes somatic mutation data from m cancer patients as input, and identifies *multiple* sets of mutations, where each set satisfies two properties: (1) the set has high *coverage* with many patients having a mutation in the set; (2) the set exhibits a pattern of mutual exclusivity where most patients have exactly one mutation in the set. We briefly describe the Multi-Dendrix algorithm here. Further details are provided in the Methods section below.

We assume that somatic mutations have been measured in m cancer patients and that these mutations are divided into n different *mutation classes*. A mutation class is a grouping of different mutation types at a specific genomic locus. In the simplest case, a mutation class corresponds to a grouping of all types of mutations (single nucleotide variants, copy number aberrations, etc.) in a single gene. We represent the somatic mutation data as an $m \times n$ binary *mutation matrix* A, where the entry A_{ij} is defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if gene } j \text{ is mutated in patient } i \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

More generally, a mutation class may be defined for an arbitrary genomic locus, and not just a gene, and may distinguish different types of mutations. For example, one may define a mutation class as single-nucleotide mutations in an individual residue in a protein sequence or in a protein domain. Or alternatively, one may separate different types of mutations in a gene (e.g. single-nucleotide mutations, deletions, or amplifications) by creating separate mutation classes for each mutation type in each gene. We will use this later definition of mutation classes in the results below. For ease of exposition we will assume for the remainder of this section that each mutation class is a gene.

Vandin et al. [23] formulate the problem of finding a set of genes with high coverage and high exclusivity as the *Maximum Weight Submatrix Problem*. Here the weight $W(M) = |\Gamma(M)| - \omega(M)$ of a set M of genes is the difference between the coverage $|\Gamma(M)|$, the number of patients with a mutation in one of the genes in M, and the coverage overlap $\omega(M)$, the number of patients having a mutation in more than one gene in M. Vandin et al. [23] introduce the De novo Driver Exclusivity (Dendrix) algorithm [23] that finds a set M of k genes with maximum weight $W(M)$.

While finding single driver pathways is important, most cancer patients are expected to have driver mutations in multiple pathways. Dendrix used a greedy iterative approach to find multiple gene sets (described below), that is not guaranteed to find optimal gene sets. Identification of multiple driver pathways requires a criterion to evaluate possible collections of gene sets. Appealing to the same biological motivation as above, we expect that each pathway contains approximately one driver mutation. Moreover, since each driver pathway is important for cancer development, we also expect that most individuals contain a driver mutation in most driver pathways. Thus, we expect high exclusivity within the genes of each pathway and

high coverage of each pathway on its own. One measure that satisfies these criteria is to find a *collection* $\mathbf{M} = \{M_1, M_2, ..., M_t\}$ of gene sets whose sum of weights is maximized.

We define the Multiple Maximum Weight Submatrices problem as the problem of finding such a maximum weight collection. We solve the *Multiple Maximum Weight Submatrix* problem using an integer linear program (ILP), and refer to the resulting algorithm as Multi-Dendrix (see Methods). In addition, the ILP formulation used in Multi-Dendrix uses a modified weight function $W_\alpha(M) = |\Gamma(M)| - \alpha\omega(M)$, where $\alpha > 0$ is a parameter that adjusts the tradeoff between finding sets with higher coverage $\Gamma(M)$ (more patients with a mutation) versus higher coverage overlap $\omega(M)$ (greater non-exclusivity between mutations). We use this parameter in the breast cancer dataset below. In contrast, Dendrix was limited to $\alpha = 1$.

Simulated data

We compare Multi-Dendrix to iterative versions of Dendrix [23] and RME [22] on simulated mutation data with both driver mutations implanted in pathways in a mutually exclusive manner and random passenger mutations. The goal of these simulations is to compare Multi-Dendrix to other algorithms that identify mutually exclusive genes on straightforward datasets that contain *multiple* mutually exclusive sets. We generate mutation data for $m = 160$ patients and $n = 360$ genes as follows. We select a set of four pathways $\mathbf{P} = (P_1, P_2, P_3, P_4)$ with each P_i containing four genes. We select the coverage $\Gamma(P_i)$ uniformly from the following intervals: $[0.75m, 0.9m]$, $[0.6m, 0.75m]$, $[0.45m, 0.6m]$, $[0.3m, 0.45m]$, respectively. The size of this dataset and the varying coverages of the pathways model what is observed in real data (see § Somatic Mutation data) and is consistent with models of mutation progression where driver mutations accumulate in pathways [28]. For each pathway P_i, we select $|\Gamma(P_i)|$ patients at random and add a driver mutation to exactly one gene from the set P_i. Thus, the driver mutations in each pathway are mutually exclusive. We then add passenger mutations by randomly mutating genes in each patient with probability, q, the *passenger mutation probability*.We used values of q similar to our estimates for q on the TCGA GBM and Lung cancer data sets (in § Somatic Mutation data below), which were $q = 0.001$ and $q = 0.0005$, respectively. We emphasize that these simulations do not model all of the complexities of somatic mutations in cancer e.g. gene-specific and patient-specific mutation rates, genes present in multiple pathways, etc.

Since the Dendrix and RME algorithms are designed to find single pathways, we compared Multi-Dendrix to iterative versions of these methods that return multiple gene sets. For Dendrix we used the iterative approach described in [23]: apply Dendrix to find a highest scoring gene set, remove those genes from the dataset, and apply Dendrix to the reduced dataset, repeating these steps until a desired number t of gene sets are found. We will refer to this algorithm as Iter-Dendrix. Thus, Iter-Dendrix returns a collection $\mathbf{P} = (P_1, P_2, \ldots, P_t)$ of t gene sets such that $W(P_1) \geq W(P_2) \geq \cdots \geq W(P_t)$. We implemented the analogous iterative version of RME, and will refer to this algorithm as Iter-RME. We compared the collection $\mathbf{M}$ of gene sets found by each algorithm to the planted pathways $\mathbf{P}$, computing the symmetric difference $d(\mathbf{P}, \mathbf{M})$ between $\mathbf{M}$ and $\mathbf{P}$ as described in Methods.

Table 1 shows a comparison of Multi-Dendrix, Iter-Dendrix, and Iter-RME on simulated mutation data for different values of q. Note that we do not show comparisons to Iter-RME for $q \geq 0.005$ as Iter-RME did not complete after 24 hours of runtime for *any* of the 1000 simulated mutation data sets. While the RME publication [22] analyzed mutation matrices with thousands of genes and hundreds of patients, this analysis (and the released RME software) required that mutations were presented in at least 10% of the samples, greatly reducing the number of genes/samples input to the algorithm. In fact, a threshold of 10% will remove nearly all genes in current whole-exome studies (see § Comparison of Multi-Dendrix and RME).

For $0.0005 \leq q \leq 0.015$, Multi-Dendrix identifies collections of gene sets that were significantly closer ($p < 0.01$) to the planted pathways $\mathbf{P}$ than the collections found by either Iter-Dendrix and Iter-RME. These results demonstrate that Multi-Dendrix outperforms other methods, even when the passenger mutation probability q is more than 15 times greater than the value estimated from real somatic mutation

data. For $q \leq 0.0001$, the differences between Multi-Dendrix and Iter-RME were not significant.

We also compared the runtimes of each algorithm on the simulated datasets. Multi-Dendrix was several orders of magnitude faster than Iter-Dendrix and Iter-RME on all datasets (Table 2). Note that as the passenger mutation probability q increases, the number of recurrently mutated passenger genes increases. Multi-Dendrix scales much better than Iter-RME and maintains a significant advantage over Iter-Dendrix, completing all simulated datasets in less than 5 seconds.

We evaluated how the runtime of Multi-Dendrix scales to larger datasets. Using the same passenger mutation probabilities $0.0001 \leq q \leq 0.02$ listed above, we calculated the average runtime in seconds of Multi-Dendrix for ten simulated mutation matrices with $m = 100, 200, 400, 800, 1600, 3200, 6400, 12800, 22000$ genes and $n = 1000$ patients, more than the number of patients to be measured in any cancer study from TCGA. In each case, we run Multi-Dendrix only on the subset of genes that are mutated in more than the expected number nq of samples. For the largest dataset with $m = 22000$ genes, the average number of genes input to Multi-Dendrix for the highest and lowest passenger mutation probabilities are ~ 9700 and ~ 2100, respectively. (Table S1 shows the average number of input genes for varying m and q.) The average runtime for this largest dataset is under one hour (average of 54.4 minutes). Figure S1 shows the runtimes for varying m and q.

The Multi-Dendrix Computational Pipeline

We incorporate the Multi-Dendrix algorithm into a larger pipeline (Figure 1) that includes several additional pre- and post-processing tasks including: (1) Building mutation matrices for input into Multi-Dendrix; (2) Summarizing Multi-Dendrix results over multiple values for the parameters t, the number of gene sets, $k_{\min}$ the minimum size of a gene set, and $k_{\max}$ the maximum size of a gene set; (3) Evaluating the statistical significance of results; (4) Examining Multi-Dendrix results for mutually exclusive sets resulting from subtype-specific mutations. We describe these steps briefly below, with further details in the Methods and Supporting Information.

First, we build mutation matrices A from somatic mutation data. We use several steps to process single-nucleotide variant (SNV) data, copy number variant (CNV) data, and to combine both types of data. Second, in contrast to simulated data, on real data we do not know the correct values of the parameters t, $k_{\min}$, and $k_{\max}$. Thus, we consider a reasonable range of values for these parameters and summarize the results over these parameters into *modules*. We build a graph, where the nodes are individual genes (or mutation classes) and edges connect genes (respectively mutation classes) that appear in the same gene set for more than one value of the parameters. We weight each edge with the fraction of parameter values for which the pair of genes appear in the same gene set. The resulting edge-weighted graphs provide a measure of the stability of the resulting gene sets over different parameter values. By choosing a minimum edge weight, we partition the graph into connected components, or *modules*. One may choose to use these modules as the output of Multi-Dendrix.

Third, we evaluate the statistical significance of our results using two measures. Since the collection $\mathbf{M}$ with high weight $W'(\mathbf{M})$ may not be surprising in a large mutation matrix A, the first measure evaluates the significance of the score $W'(\mathbf{M})$ maximized by Multi-Dendrix. We evaluate whether the weight $W'(\mathbf{M}^*)$ of the maximum weight collection $\mathbf{M}^*$ output by Multi-Dendrix is significantly large compared to an empirical distribution of the maximum weight sets from randomly permuted mutation data. We generate random mutation data using the permutation test described in [19]. This test permutes the mutations among the genes in each patient, preserving both the number of mutated genes in each patient and the number of patients with a mutation in each gene while perturbing any patterns of exclusivity between mutated genes. Note that this permutation test requires running Multi-Dendrix many times to determine statistical significance for a single parameter setting. Thus, the runtime advantages of Multi-Dendrix compared to Iter-Dendrix and Iter-RME are very important in practice on real datasets.

Next, we evaluate whether the collection $\mathbf{M}^*$ output by Multi-Dendrix contains more protein-protein interactions than expected by chance by applying our *direct interactions test* on a PPI network constructed from the union of the KEGG and iRefIndex PPI networks. The direct interactions test computes a statistic ν of the difference in the number of interactions *within* and *between* gene sets in $\mathbf{M}^*$, and compares the observed value of ν to an empirical distribution on 1000 permuted PPI networks (full details of the test are in § Evaluating known interactions). These permuted networks account for the observation that many genes that are frequently mutated in cancer also have large degree in the interaction network – either due to biological reasons or ascertainment bias. We use an interaction network to assess biological function rather than known pathways (e.g. KEGG pathways or GSEA sets) because most of these pathways are relatively large, while the gene sets found by Multi-Dendrix that exhibit exclusivity tend to be much smaller, each containing only a few genes.

Finally, we examine possible correlations between the mutually exclusive sets reported by Multi-Dendrix and particular subsets of samples. A number of cancers are divided into subtypes according to pathology, cytogenetics, gene expression, or other features. Since mutations that are specific to particular subtypes will be mutually exclusive, disease heterogeneity is an alternative explanation to pathways for observed mutually exclusive sets. For example, [29] report four subtypes of GBM based on gene expression clusters, and show that several mutations – including IDH1, PDGFRA, EGFR, and NF1 – have strong association with individual subtypes. Unfortunately, if the subtypes are unknown there is no information for Multi-Dendrix, Dendrix, RME, or other algorithms that analyze mutual exclusivity to distinguish between mutual exclusivity resulting from subtypes and mutual exclusivity resulting from pathways or other causes. If subtypes are known, two possible solutions are to analyze subtypes separately, or to examine whether patterns of mutual exclusivity are associated to these subytpes. We annotate results by known subtypes as a post-processing step in Multi-Dendrix.

Somatic Mutation data

We applied Multi-Dendrix and Iter-Dendrix to four somatic mutation matrices: (1) copy number variants (CNVs), small indels, and non-synonymous single nucleotide variants (SNVs) measured in 601 genes in 84 glioblastoma multiformae (GBM) patients [7]; (2) indels and non-synonymous single nucleotide variants in 623 sequenced genes in 188 Lung Adenocarcinoma patients [27]; (3) CNVs, small indels, and non-synonomous SNVs measured using whole-exome sequencing and copy number arrays in 261 GBM patients [7]; and (4) CNVs, small indels, and non-synonymous SNVs measured in 507 BRCA patients. We will refer to these datasets as GBM(2008), Lung, GBM, and BRCA below. We removed extremely low frequency mutations and known outliers from these datasets as described in Methods. After this processing, the GBM(2008) dataset contained mutation and CNV data for 46 genes in 84 patients; the Lung dataset contained somatic mutation for 190 genes in 163 patients; the GBM dataset contained mutation and CNV data for 398 genes in 261 patients; and the BRCA dataset contained mutation and CNV data for 375 genes in 507 patients. We focus here on presenting results from the latter two datasets because they are the latest whole-genome/exome datasets and most representative of the datasets that are now being produced and will be analyzed now and in the coming years. Results with the first two older and smaller datasets from targeted sequencing are described in the Supporting Information.

We compute $2 \leq t \leq 4$ gene sets, each of minimum size $k_{\min} = 3$ and maximum size ranging from $3 \leq k_{\max} \leq 5$. We summarize the results over these 9 different parameter values into modules using the procedure described above.

Mutually exclusive sets in Glioblastoma (GBM)

We applied Multi-Dendrix and Iter-Dendrix to the GBM dataset, considering EGFR amplification as a separate event (see Methods). The algorithms report the same results over all values of the parameters except $k_{\max} = 5$, where Iter-Dendrix includes the IRF5 gene in a gene set with RB1, CDK4(A), and CDKN2A/CDKN2B(D), and MSL3. However, Multi-Dendrix is significantly faster running in 142 seconds compared to 37,786 seconds (over 10 hours) for Iter-Dendrix.

We summarize the results of these different parameter choices by connecting genes that appear in the same gene set at least twice, resulting in four modules (Figure 2). These four modules include all the genes (except ERBB2) that are: (1) members of the three signaling pathways highlighted in the TCGA GBM study [7], and (2) are mutated in $> 5\%$ of the samples. The weight $W'(M)$ of all collections found by Multi-Dendrix on the GBM dataset are significant $P < 0.0001$) and the direct interactions statistic ν of these four modules is also significant ($P = 0.002$). Three of the four modules also contain a significant number of interactions ($P < 0.05$). In addition to these four modules, two additional mutation classes, CNTNAP2 and deletion of 10q26.3, each appear in one choice of parameters for Multi-Dendrix. Since these are not part of a larger module, they were not further analyzed. Figure S2 shows a combined mutation matrix with all four modules.

The first module includes the amplification of CDK4, mutation of RB1, and a deletion that includes both CDKN2A and CDKN2B. This module is mutated in 87.7% (229 / 261) of the samples, and are discovered for all parameter choices. These four genes are members of the RB signaling pathway (as annotated in [7]) involved in G1/S progression ($P = 9.7 \times 10^{-14}$ by Bonferonni-corrected hypergeometric test): CDKN2A and CDKN2B inhibits CDK4, which in turn inhibits RB1. In addition for 7/9 parameter choices, this module includes mutations in MSL3. MSL3 is a member of the MSL (male-specific lethal) complex that has a major role in dosage compensation in *Drosophila*. While this complex is conserved in mammals, the specific function of human MSL3 is unknown. However, the MSL complex also includes the histone acetyltransferase MOF which is involved in cell cycle regulation of p53 and may play a role in cancer [30]. Thus, the mutual exclusivity of mutations in MSL3 and the other well-known members of the RB signaling pathway is intriguing and deserves further study. This module contains two interactions ($P = 0.005$).

The second module includes mutations and deletion of PTEN, mutations in PIK3CA, mutations in PIK3R1, mutations in IDH1, and an amplification that includes PDPN and PRDM2. The module is mutated in 62.8% (164/261) of the samples. PTEN, PIK3CA, and PIK3R1 are all members of the RTK/RAS/PI(3)K signaling pathway (as annotated in [7]) involved in cellular proliferation ($P = 3.2 \times 10^{-11}$). IDH1 is not a known member of this pathway; moreover, IDH1 is preferentially mutated in the proneural subtype of GBM [29]. Deletions in PTEN are also associated with the proneural subtype of GBM, although they are not considered a defining feature of this subtype (as IDH1 mutations are) and do not result in a gene expression signature [29]. However, there are no reports that PTEN, PIK3CA, or PIK3R1 mutations are subtype specific, and thus the mutual exclusivity of IDH1 and the remaining genes in this set is not simply explained by subtypes. PRDM2 is not known to be part of the RTK/RAS/PI(3)K signaling pathway. PRDM2 is a member of the histone methyltransferase superfamily, interacts with the RB protein [31], and is proposed as a tumor suppressor in colorectal cancer [32]. PDPN is used as a molecular marker for glioma, due to its association with clinical outcomes [33]. Our results suggest that PDPN and PRDM2 may have an undiscovered role in GBM as well. This module contains three interactions ($P = 0.001$).

The third module includes mutations in TP53, the amplification of MDM2, the amplification of MDM4, mutations in NLRP3, and the deletion involving AKAP6 and NPAS3. This module is mutated in 57.8% (151/261) of the samples, and appears for every parameter choice for $t > 2$. TP53, MDM2, and MDM4 are members of the p53 signaling pathway ($P = 9.7 \times 10^{-14}$), a critical and frequently altered pathway in GBM involved in senescence and apoptosis. NPAS3 is a transcription factor expressed in the brain and implicated in psychiatric disorders including schizophrenia [34, 35]. In addition, NPAS3 was recently shown to act as

a tumor suppressor in astrocytomas, with a possible role in glioblastoma progression and proliferation [36]. This module contains three interactions ($P = 0.001$).

The fourth module includes mutations in EGFR, the amplification of PDGFRA, and the deletion of RB1. This module is mutated in 45.6% (119/261) of samples, and appears for $t = 4$. EGFR and PDGFRA are members of the RTK/RAS/PI(3)K signaling pathway ($P = 8.9 \times 10^{-9}$), and RB1 is a member of the RB signaling pathway. While EGFR and PDGFRA both interact with RAS, there are no reported direct interactions between these three proteins. In addition, mutations in these three genes are significantly associated with two of the expression subtypes reported in [29]: mutations in EGFR and the deletion of RB1 are associated with the Classical GBM subtype, and the amplification of PDGFRA is significantly associated with the Proneural subtype. Thus, it appears that the mutual exclusivity discovered by Multi-Dendrix is a result of subtype-specific mutations, despite PDGFRA and EGFR being a member of the same biological pathway.

In summary, we see that subtype-specific mutations provide an alternative explanation for observed mutual exclusivity and confound the identification of driver pathways. However, on the GBM data subtype-specific mutations are a minor feature in the data, and Multi-Dendrix successfully identifies *de novo* portions of three critical signaling pathways in GBM.

Mutually exclusive sets in Breast Cancer (BRCA)

We applied Multi-Dendrix and Iter-Dendrix to the BRCA dataset. We found that for most values of the parameters t and $k_{\max}$, the results combined the most frequently mutated genes into a single gene set despite the fact that these genes had high coverage overlap (Figures S4 and S5). That is, for a gene set M, high coverage $|\Gamma(M)|$ was outweighing a high coverage overlap $\omega(M)$ in the weight function $W(M) = |\Gamma(M)| - \omega(M)$ optimized by Multi-Dendrix. To enforce greater mutual exclusivity, we increased the coverage overlap penalty to $\alpha = 2.5$ from its default value of $\alpha = 1$.

Using $\alpha = 2.5$, Multi-Dendrix identifies four distinct modules (Figure 3). These four modules overlap with three pathways known to be important in BRCA: p53 signaling, PI(3)K/AKT signaling, and cell cycle checkpoints. In addition, they include genes recently identified by [8] as important BRCA genes. These modules cover a smaller proportion of samples than our results on the GBM dataset even though they include the same number of genes, suggesting greater mutational heterogeneity or disease heterogeneity (i.e. subtypes) in the breast cancer dataset. Indeed, breast cancers are commonly divided into four major subtypes: Luminal A/B, Basal, and HER2 type. We annotate the subtype-specific mutations below, and Table S6 lists the significant associations between mutations and subtypes. The weight $W'(M)$ of all collections found by Multi-Dendrix on the BRCA dataset are significant ($P < 0.01$), and the direct interactions statistic ν of these four modules is also significant ($P < 0.001$). In addition, one module is significantly enriched for interactions ($P < 0.05$). In addition to these four modules, five additional genes (SF3B1, CCDC150, COL23A1, C20orf26, PCDHA5) each appear in one choice of parameters for Multi-Dendrix. Since these are not part of a larger module, they were not further analyzed. Figure S3 shows a combined mutation matrix with all four modules.

The first module includes the deletion of PTEN, mutations in PIK3CA, PIK3R1, AKT1, HIF3A, and the amplification of genomic region 12p13.33. These six mutation classes are mutated in 61% (308/507) of the samples, and this module is discovered for all values of the parameters. PTEN, PIK3CA, PIK3R1, and AKT1 form the core of the PI(3)K/AKT signaling pathway, as annotated in [8] ($P < 10^{-15}$): AKT1 is known to interact with PTEN, PIK3R1, and PIK3CA, while PTEN inhibits PIK3CA and PIK3R1. These genes constitute five of the eight genes mutated in the PI(3)K/AKT signaling pathway, as reported by [8]. One of the other mutation classes in this module is the amplification of 12p in 82 samples. This amplification event has been reported before in BRCA, although the likely target of this amplification is unknown [37]. This module contains six interactions ($P < 0.001$).

The second module includes mutations in the genes TP53, CDH1, GATA3, CTCF, and GPRIN2. These

five genes are mutated in 56% (283 / 507) of the samples, and are discovered for each parameter choice. TP53 is a member of the p53 signaling pathway, while GATA3, CDH1, and CTCF are well-known for their role in breast cancer and are involved in metastasis and proliferation. The loss or downregulation of CDH1, the E-cadherin gene located on 16q22.1, is implicated in breast cancer invasion and proliferation (reviewed in [38]). CTCF, also located on 16q22.1, has been found to act as a tumor suppressor in breast cancer through mechanisms similar to CDH1 [39]. GATA3 is a transcription factor that regulates immune cells, and has long been known to be involved in breast cancer tumorigenesis [40]. Recently, a novel role for GATA3 was discovered, whereby GATA3 suppresses breast cancer metastasis through inhibition of E-cadherin promoters [41]. This module contains zero known interactions.

The third module includes the deletion of MAP2K4 and mutations in MAP3K1, PPEF1, SMARCA4, and WWP2. These five mutations occur in 44.4% (225/507) of the samples, and this module is identified when the number t of gene sets is at least 3. MAP3K1 and MAP2K4 are both members of the p38-JNK1 stress kinase pathway as reported in [8], and are involved in the regulation of apoptosis. Both MAP3K1 and MAP2K4 are serine/threonine kinases, while PPEF1 is a serine/threonine phosphatase. The targets of PPEF1 are unknown, although there are reports that PPEF2, another member of the gene family, does interact with the p38-JNK pathway through the gene ASK1 [42]. SMARCA4 is also a known cancer gene, and has been shown to have a role as a tumor suppressor in lung cancer [43]. This module contains one interaction ($P = 0.179$).

The fourth module includes the amplification of CCND1 and mutations in MAP2K4, RB1, and GRID1. These four mutations occur in 36.3% (184/507) of the samples, and this module is discovered when the number t of gene sets is 4. CCND1 and RB1 encode interacting proteins that play a role cell cycle progression. CCND1 encodes the cyclin-D1 protein that inhibits the retinoblastoma protein encoded by RB1 via hyperphosphorylation. The hyperphosphorylation of *RB* inactivates its role as a tumor suppressor, and thus mutations that target either CCND1 or RB1 are thus important for proliferation in cancer [44]. Mutations in MAP2K4, as discussed above for the third module, target the p38-JNK1 pathway, and MAP2K4 is not known to have any interactions with CCND1 or RB1. This module contains one interaction ($P = 0.1$).

We separately ran Multi-Dendrix on mutation data restricted to the 224 Luminal A patients, the 124 Luminal B patients, the 93 Basal-like patients, and the 58 HER2-enriched patients (as annotated in [8]) using different values of α for the different sized datasets (see Supporting Information § BRCA subtypes for details). The gene pair {TP53, GATA3} from the Multi-Dendrix modules is also identified by Multi-Dendrix when restricting to basal-like or luminal B samples. The same is true for the gene pair {AKT1, PIK3CA} in HER2-enriched and luminal A subtypes. Thus, the mutual exclusivity between mutations in these pairs of genes is not a result of subtype-specific mutations. On the other hand, the mutual exclusivity between 12p13.33 amplification and PIK3CA mutation appears to be an effect of subtypes, as these aberrations are associated with basal-like and luminal A subtypes, respectively (Table S6), and these two aberrations are not grouped into the same module on any Multi-Dendrix run of individual subtypes. These results show that mutual exclusivity can result from mutational heterogeneity within pathways, disease heterogeneity with subtype-specific mutations, or both.

Discussion

We introduce an algorithm Multi-Dendrix that simultaneously finds multiple cancer driver pathways using somatic mutation data from a collection of patients. Multi-Dendrix finds combinations of somatic mutations solely from the combinatorial pattern of their mutations with *no prior knowledge of pathways or interactions between genes*. On simulated data, Multi-Dendrix outperforms iterative versions of Dendrix and RME, two previous algorithms for identifying single driver pathways. Multi-Dendrix finds optimal groups of mutually exclusive mutations even in the presence of significant noise where the passenger mutation rate is 15 times

greater than observed in real data. Multi-Dendrix is orders of magnitude faster than iterative versions of Dendrix and RME and scales to the analysis of mutations in thousands of genes from hundreds of cancer patients. Finally, Multi-Dendrix finds optimal solutions over a range of sizes for the individual gene sets, while Dendrix examines only a fixed size gene set for each run.

We apply Multi-Dendrix to multiple cancer datasets, including glioblastoma and breast cancer data from The Cancer Genome Atlas, two of the most comprehensive somatic mutation datasets from high-throughput sequencing data. Multi-Dendrix finds multiple driver pathways of interacting genes/proteins including portions of the p53, RTK/RAS/PI(3)K, PI(3)K/AKT, and Rb signaling pathways. At the same time, we identify additional genes whose mutations are mutually exclusive with mutations in these pathways. Intriguingly, these additional genes are transcription factors or other nuclear proteins involved in regulation of transcription (SMARCA4, NPAS3, PRDM2, and MSL3). In general, gene transcription is the downstream "output" of signaling pathways, suggesting that mutations in these downstream targets may be substituting for mutations in the upstream (and presumably more general) signaling pathways.

Our results demonstrate the advantages of simultaneously finding sets of mutually exclusive genes. However, there can be multiple explanations for observed mutual exclusivity in a set of genes including both pathways (i.e., interactions between genes) as well as disease heterogeneity (i.e., cancer subtypes). Interpretation of Multi-Dendrix results should consider these various explanations. To facilitate these analyses, we include additional steps in the Multi-Dendrix pipeline to compare the results sets of mutations to known gene interactions and/or known subtypes in the samples.

Our Multi-Dendrix analysis shows an interesting feature of mutual exclusivity between SNVs and deletions in tumor suppressor genes. In the glioblastoma data, we find that SNVs in PTEN and the deletion of PTEN are mutually exclusive. It is not clear whether this is a genuine biological feature of the mutational process, where either SNVs or deletions (but not both) are present in different samples, or merely an artifact of the mutation calling. Regarding the latter, it is possible that SNVs are more challenging to detect in hemizygous samples, where the other allele has been deleted.

There are several extensions that might further improve the Multi-Dendrix algorithm. First, the ILP used in Multi-Dendrix finds optimal solutions effectively, but is not guaranteed to rigorously examine suboptimal solutions. This is in contrast to the Markov Chain Monte Carlo (MCMC) approach used by Dendrix that samples suboptimal solutions in proportion to their weight. Extending the MCMC approach to simultaneous discovery of multiple pathways, or perhaps using the ILP to initialize a sampling procedure are interesting directions for future research. Second, the weight function used by the Multi-Dendrix algorithm does not explicitly incorporate co-occurrence of mutations between genes in different gene sets. Instead, the weight function is highest for gene sets with high coverage and approximate exclusivity, which may co-occur in many patients simply due to high coverage (a trivial example is when all gene sets have full coverage, and thus all mutations co-occur across gene sets). As noted in [25], co-occurrence of mutations is known to be important in large biological pathways, and thus algorithms that explicitly optimize for collections of gene sets where mutations between gene sets co-occur a surprising amount may have an advantage in identifying key components of these larger biological pathways.

We anticipate that Multi-Dendrix will be useful for analyzing somatic mutation data from different types of cancer and with larger cohorts of patients, such as those now being generated by The Cancer Genome Atlas (TCGA) and other large-scale cancer sequencing projects. However, mutual exclusivity and high coverage are not the only criteria for selecting driver mutations, and thus the model optimized by Multi-Dendrix has some limitations. As noted above, it is well-known that each patient has multiple driver mutations and there are many examples of co-occurring mutations. Some of these co-occurring mutations are clearly in different pathways, although this depends on one's definition of a pathway. Large, multifunctional "pathways" such as the cell-cycle pathway do indeed exhibit co-occurring mutations; on the other hand, co-occurring mutations appear to be less common in directly interacting genes/proteins. In addition, the coverage, or frequency, of important driver mutations may vary considerably, according to the

stage at which patients are sequenced, or extensive disease heterogeneity. Private mutations that are unique to a single individual in a study might be driver mutations, but will not exhibit a strong signal of mutual exclusivity. Multiple approaches are required to prioritize somatic mutations for further experimental study. Multi-Dendrix is a useful complement for analysis of significantly mutated genes [6, 45, 46], functional impact [1–4], pathway/network analysis [18–22], and other approaches.

Methods

Weight function for gene sets

Given a mutation matrix A, and a set M of genes (or equivalent mutation classes), Vandin et al. [23] define the weight function $W(M)$ as follows. For a gene g, the *coverage* $\Gamma(g) = \{i : A_{ig} = 1\}$ is the set of patients in which gene g is mutated. Similarly, for a set M of genes, the *coverage* is $\Gamma(M) = \cup_{g \in M} \Gamma(g)$. M is *mutually exclusive* if $\Gamma(g) \cap \Gamma(g') = \emptyset$, for all $g, g' \in M$, $g \neq g'$. A gene set in A is a column submatrix of A with high coverage and approximate exclusivity. Since increasing coverage may be achieved at the expense of decreasing exclusivity, [23] define a weight $W(M)$ on set M of columns of A that quantifies both coverage and exclusivity of M. In particular, they define the *coverage overlap* of M as $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$. Note that $\omega(M) = 0$ when M is mutually exclusive. Then $W(M)$ is the difference between the coverage and coverage overlap of M:

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

Note that for a mutually exclusive submatrix M, $\Gamma(M) = \sum_{g \in M} \Gamma(g)$ and therefore $W(M) = \Gamma(M)$.

Vandin et al. [23] introduce the Maximum Weight Submatrix Problem defined for an integer $k > 0$ as the problem of finding the $m \times k$ submatrix M of A that maximizes a weight $W(M)$, show that this problem is NP-hard and derive the De novo Driver Exclusivity (Dendrix) algorithm to solve it. Dendrix is a Markov Chain Monte Carlo (MCMC) algorithm that samples sets of k genes in proportion to their weight W. While the MCMC algorithm is not guaranteed to find a gene set of optimal W, the authors showed that applying the method to real somatic mutation data produces gene sets with high coverage and approximate exclusivity, and that the Markov chain rapidly converges to a stationary distribution.

An ILP for the Maximum Weight Submatrix Problem

We formulate the *Maximum Weight Submatrix Problem* as an integer linear program, which we refer to as Dendrix$_{ILP}(k)$. Given a mutation matrix A and gene set size k, Dendrix$_{ILP}(k)$ finds a gene set M^* with largest weight $W(M^*)$. A gene set M is determined by a set of indicator variables, one for each gene j,

$$I_M(j) = \begin{cases} 1 & \text{if gene } j \text{ is a member of gene set } M, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

To compute the weight function $W(M)$ in (2), it is necessary to compute the coverage $\Gamma(M)$. To do this, we define an indicator variable for each patient i,

$$C_i(M) = \begin{cases} 1 & \text{if gene set } M \text{ is mutated in patient } i, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Then, $Dendrix_{ILP}(k)$ is defined as follows:

$$\text{maximize} \sum_{i=1}^{m} \left(2 \cdot C_i(M) - \sum_{j=1}^{n} I_M(j) \cdot A_{ij} \right) \tag{5a}$$

subject to

$$\sum_{j=1}^{n} I_M(j) = k \tag{5b}$$

$$\left(\sum_{j=1}^{n} A_{ij} \cdot I_M(j) \right) \geq C_i(M), \tag{5c}$$

$$\text{for } 1 \leq i \leq m.$$

Note that the last constraint (5c) only forces $C_i(M) = 0$ when all genes in M are not mutated (i.e. $\sum_{j=1}^{n} A_{ij} \cdot I_M(j) = 0$), but does not force $C_i(M) = 1$ when at least one gene in M is mutated as required by (4). However, in the latter case the objective function will be maximized when $C_i(M) = 1$ and thus (4) is satisfied.

Note that removing equation (5b) from Dendrix$_{ILP}(k)$ above produces a gene set with optimal W for any value of k, i.e. a gene set M with maximum $W(M)$ of any size. We will refer to this variant of the ILP as Dendrix$_{ILP}$. In addition, we can set bounds on the size of the gene set M, $k_{\min} \leq |M| \leq k_{\max}$, by replacing equation (5b) with

$$k_{\min} \leq \sum_{j=1}^{n} I_M(j) \leq k_{\max}. \tag{6}$$

We will refer to this variant as Dendrix$_{ILP}(k_{\min}, k_{\max})$.

Multiple Maximum Weight Submatrices Problem

We define the following problem.

Multiple Maximum Weight Submatrices Problem. *Given an $m \times n$ mutation matrix A and an integer $t > 0$, find a collection $\mathbf{M} = \{M_1, M_2, \ldots, M_t\}$ of $m \times k$ column submatrices that maximizes $W'(\mathbf{M}) = \sum_{\rho=1}^{t} W(M_\rho)$.*

Note that this problem is NP-hard, as stated above for the case $t = 1$. Further note that while the weight $W'(\mathbf{M})$ is increased by greater mutual exclusivity between mutations *within* a gene set, there is no restriction on the mutations *between* gene sets. Moreover, collections with large weight $W'(\mathbf{M})$ will also tend to have larger coverage $\Gamma(M_\rho)$ of each individual gene set ρ. Thus, optimal solutions will tend to produce a collection with the property that many patients have a mutation in more than one gene set; or alternatively, there may be pairs or larger sets of co-occurring mutations, a phenomenon that has been observed in cancer [25].

We solve the *Multiple Maximum Weight Submatrix* problem using an integer linear program (ILP). We define an ILP, called Multi-Dendrix, using the same definitions of $C_i(M)$ and $I_M(i)$ as above:

$$\text{maximize} \sum_{\rho=1}^{t} \sum_{i=1}^{m} \left(2 \cdot C_i(M_\rho) - \sum_{j=1}^{n} I_{M_\rho}(j) \cdot A_{ij} \right) \tag{7a}$$

subject to

$$\left(\sum_{j=1}^{n} I_{M_\rho}(j) \cdot A_{ij}\right) \geq C_i(M_\rho), \tag{7b}$$

for $1 \leq i \leq m, 1 \leq \rho \leq t$,

$$\sum_{\rho=1}^{t} I_{M_\rho}(j) \leq 1, 1 \leq j \leq m. \tag{7c}$$

We solve this ILP using CPLEX v12.3 using default parameters. It is obvious that for a mutation matrix A, the collections $\mathbf{M}$ and $\mathbf{I}$ of gene sets produced by Multi-Dendrix and Iter-Dendrix, respectively satisfy $W'(\mathbf{M}) \geq W'(\mathbf{I})$. Multi-Dendrix can also produce sets $\mathbf{M}$ with strictly larger weight than the sets $\mathbf{I}$ produced by Iter-Dendrix. There are a number of ways this might occur. First, there may be multiple gene sets I_ρ with maximum weight on iteration ρ, and only one of these is extended by the iterative algorithm. Second, the maximum weight gene set I_ρ selected by Iter-Dendrix in the ρ^{th} iteration may not be a member of the optimal $\mathbf{M}$; i.e. $\mathbf{M}$ may contain gene sets that are suboptimal when considered in isolation. Third, when $k_{\min} < k_{\max}$ Multi-Dendrix may select gene sets with fewer than $k_{\max}$ genes if this maximizes the weight $W'(\mathbf{M})$ of the whole collection. We find that all of these cases occur on real somatic mutation data.

Multi-Dendrix can be extended to allow genes to be members of more than one gene set. Such genes may be involved in multiple biological processes. We define the parameter Δ to be the maximum number of gene sets a gene can be a member of, and the parameter τ to be the number of overlaps allowed per gene set. Then we replace Eq. 7c with $\sum_{\rho=1}^{t} I_{M_\rho(j)} \leq \Delta, \forall j$ and add the constraint $\sum_{j=1}^{n} \sum_{\rho' \neq \rho} I_{M_\rho}(j) \cdot I_{M_{\rho'}}(j) \leq \tau, 1 \leq \rho \leq t$. Note that the product in the second group of constraints must be encoded using additional indicator variables and thus the number of additional constraints grows rapidly as more overlaps between gene sets are allowed. While this extension is implemented in the Multi-Dendrix package, we did not use it for the results presented herein.

Simulations

We ran Dendrix with default parameters using $4 \times 10^4 n$ iterations for the MCMC method, where n is the number of genes. We also used the default parameters for RME. We ran each algorithm with the true values $t = 4$ and $k_{\min} = k_{\max} = 4$. We compared the collection $\mathbf{M}$ of gene sets found by each algorithm to the planted pathways $\mathbf{P}$, computing the difference between $\mathbf{M}$ and $\mathbf{P}$ as follows:

$$d(\mathbf{P}, \mathbf{M}) = \underset{\pi}{\text{argmin}} \sum_{\rho=1}^{|\mathbf{P}|} |P_{\pi(\rho)} \triangle M_\rho|,$$

where π is a permutation of $(1, 2, 3, 4)$ and $\triangle$ is the symmetric difference between sets.

Construction of mutation matrices

We have two pipelines for building mutation matrices from somatic mutation data: one for the whole-exome datasets and the second for the targeted gene sequencing datasets, GBM(2008) and lung adenocarcinoma. Here we describe the pipeline for the whole-exome datasets. See Supporting Information for further details on the targeted gene sequencing data processing.

Whole-exome and copy number aberration data preparation For the whole-exome datasets, we extracted non-silent somatic mutations from MAF files and copy number aberrations from GISTIC 2.0 [47] downloaded from the TCGA portal:

- GBM: http://gdac.broadinstitute.org/runs/analyses__2012_09_13/data/GBM/20120913/.
  - MAF file: gdac.broadinstitute.org_GBM.Mutation_Assessor.Level_4.2012091300.0.0/GBM.maf.annotated.
  - GISTIC wide peaks: gdac.broadinstitute.org_GBM.CopyNumber_Gistic2.Level_4.2012091300.0.0.tar.gz.
- BRCA: https://tcga-data.nci.nih.gov/docs/publications/brca_2012/.
  - MAF file: "Somatic MAF archive [tar.gz] (public access)".
  - Segment data: "Merged segment file".
  - GISTIC wide peaks: "Genes in focal peaks [xslx]".

We applied the following filters to remove genes from the analysis:

- 10,443 genes in the GBM dataset and 11,428 genes in the BRCA dataset mutated in fewer than 5 samples.
- 209 genes in the GBM dataset and 474 genes in the BRCA dataset whose coding regions are longer than 6Kb and had fewer mutations than expected using a binomial test with an estimate of 10^{-6} for the background mutation rate.
- 94 genes from both datasets that are observed to have unusually high rates of somatic mutation in exome-sequencing data, but are likely artifacts resulting from replication timing, active transcription, and other factors (as reported by M. Lawrence here: http://1.usa.gov/RBtuz7). These include olfactory receptors, "cub and sushi" proteins, and TTN.
- 10 additional genes (KRTAP5-5, HEATR7B2, EML2, CC2D1A, DUS3L, GABRA6, FLG, HYDIN, SUSD3, CROCCL1) that also appear to be artifacts as they were observed to have higher than average mutation rates, but with no known role in cancer.

For copy number data, we used the "wide peaks" in the GISTIC output as the copy number aberrations in each sample, with copy number ratio thresholds of 0.1 and -0.1 for determining if a segment is amplified or deleted, respectively. We performed the following additional filtering of the copy number data.

- We combine copy number aberrations that co-occur in the same sample $> 50\%$ of the time and are within 10Mb of one another into a single larger "meta" gene.
- We restricted analysis to 249 CNVs in the GBM data and 204 CNVs in the BRCA data that appeared in wide peaks with 20 or fewer genes or were given a peak label by GISTIC.
- For BRCA, we removed segments longer than 10Mb.
- Remove wide peaks containing known fragile sites at PARK2 [48], WWOX [49], RPL5, and FAM19A5.

Note that in many cases, the wide peaks include multiple genes which we combine into a single "meta" gene for Multi-Dendrix analysis. We manually selected the target gene for the following metagenes:

- For GBM, we selected CDK4, CDKN2A and CDKN2B, PTEN, PDGFRA, MDM2, MDM4, PIK3R1, and RB1 as the targets of their respective aberrations.

- For BRCA, we used the GISTIC peak labels as the target for each aberration.
- In the GBM dataset, we treated EGFR as a separate event and did not include it in the Multi-Dendrix analysis. EGFR amplification is the most common amplification, and second most common mutation (after CDKN2A/B deletion). There is a significant co-occurrence between EGFR amplifications and SNVs ($p = 9.4 \times 10^{-5}$ by Fisher's exact test), although EGFR amplification occurs in 74 more samples than EGFR SNV.

After this filtering, we analyze a total of 398 mutation classes in GBM and 375 mutation classes in BRCA (i.e. separating SNVs and CNVs in individual genes or metagenes).

Evaluating known interactions

We assess whether the mutually exclusive gene sets found by Multi-Dendrix are enriched for interacting genes using the following *direct interactions test*. We use two analogous tests: one test for collections, and one test for for individual gene sets.

For a collection $\mathbf{P} = (P_1, \ldots, P_t)$, let $\alpha(\mathbf{P})$ be the fraction of possible interactions between pairs of genes *within* each gene set P_i that are observed and $\beta(\mathbf{P})$ be the fraction of possible interactions between pairs of genes in *different* pairs of gene sets P_i, P_j that are observed. We define $\nu(\mathbf{P}) = \alpha(\mathbf{P}) - \beta(\mathbf{P})$. Thus, a large value $\nu(\mathbf{P})$ indicates a collection with many interactions between genes in the same gene set and few interactions between genes in different gene sets. This measure models our assumptions that mutual exclusivity should be strongest between genes that directly interact. Moreover, we also expect few interactions between genes in different gene sets reported by Multi-Dendrix. Thus, the measure ν is a more strict measure of the topology of interactions than merely counting the total number of observed interactions. Similarly, we assess an individual gene set by counting the number of interactions among genes in that gene set. Let $\rho(P)$ be the number of interactions among genes in P.

We assess the statistical significance of ν and ρ by comparing the observed value to the empirical distribution of ν and ρ on an ensemble of permuted networks with the same degree distribution. We permute the network by swapping edges between pairs of genes, so as to preserve the distribution of edges within the network. We perform $Q \times |E|$ edge swaps, where E is the edges in the network and Q is some constant (we use $Q = 100$ as recommended for permuting graphs with fixed degree distribution in [50]). This network permutation corrects for the observation that many genes that are frequency mutated in cancer also have many interactions in human PPI networks. Since frequently mutated genes typically appear in Multi-Dendrix results, the number of interactions between genes in Multi-Dendrix results might be higher than a *random set* of genes of the same size. When performing the direct interactions test, we use a protein-protein interaction (PPI) network constructed from the union of interactions reported in KEGG [13] and iRefIndex 9.0 [51] containing 236,060 interactions among 15,257 proteins. We calculate p-values from 1000 permuted versions of this combined network.

By examining only direct interactions between genes in our sets, we might miss cases where genes do not directly interact, but have a common interacting partner (e.g. EGFR, PDGFRA, and NF1 all share RAS as an interacting partner in the fourth module of the GBM results). Average pairwise distance is another commonly used metric for assessing whether groups of genes are clustered on an interaction network, and this metric might identify such cases. However, the tradeoff is that the diameter (average pairwise distance) of most biological interaction networks is small, and thus many genes are close on the network. We found that average pairwise distance was not a strict enough measure for examining mutually exclusive gene sets (data not shown). Finally, note that counting the number of interactions between genes in a PPI network is only an approximate measure of biological function. Current interaction networks have large number of false positive interactions – particularly when interactions from high-throughput experiments are included

– as well as an unknown number of false negatives. In addition, there is a problem of ascertainment bias as cancer genes are some of the most studied human genes and many of their interactions have been assessed.

Acknowledgments

The authors thank Hsin-Ta Wu for his assistance in visualizing mutation matrices, and the anonymous reviewers for their helpful suggestions.

References

1. Gonzalez-Perez A, Lopez-Bigas N (2012) Functional impact bias reveals cancer drivers. Nucleic acids research : 1–10.

2. Adzhubei IA, Scmidt S, Peshkin L, Ramensky VE, Gerasimoa A, et al. (2010) A method and server for predicting damaging missense mutations. Nature methods 7: 248–249.

3. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research 39: e118.

4. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4: 1073–81.

5. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314: 268–274.

6. Getz G, Hofling H, Mesirov J, Golub T, Meyerson M, et al. (2007) Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 317: 1500.

7. The Cancer Genome Atlas Research Network (2012) Comprehensive molecular portraits of human breast tumours. Nature 490: 61-70.

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

9. The Cancer Genome Atlas Research Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330–7.

10. The Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–15.

11. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475: 101–5.

12. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, et al. (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature 486: 400–4.

13. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109–114.

14. Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 25-29.

15. Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, et al. (2011) PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics (Oxford, England) 27: 1595–602.

16. Lin J, Gan CM, Zhang X, Jones S, Sjöblom T, et al. (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome research 17: 1304–18.

17. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G (2010) Patient-oriented gene set analysis for cancer mutation data. Genome biology 11: R112.

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

19. Ciriello G, Cerami E, Sander C, Schultz N (2012) Mutual exclusivity analysis identifies oncogenic network modules. Genome Res 22: 398–406.

20. Vandin F, Upfal E, Raphael B (2011) Algorithms for detecting significantly mutated pathways in cancer. Journal of Computational Biology 18: 507-22.

21. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A (2012) EnrichNet: network-based gene set enrichment analysis. Bioinformatics (Oxford, England) 28: i451–i457.

22. Miller C, Settle S, Sulman E, Aldape K, Milosavljevic A (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC medical genomics 4: 34.

23. Vandin F, Upfal E, Raphael B (2012) De novo discovery of mutated driver pathways in cancer. Genome Res 22: 375–385.

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

25. Yeang C, McCormick F, Levine A (2008) Combinatorial patterns of somatic gene mutations in cancer. The FASEB Journal 22: 2605-22.

26. Hanahan D, Weinberg R (2011) Hallmarks of cancer: the next generation. Cell 144: 646–74.

27. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, et al. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455: 1069–1075.

28. Gerstung M, Eriksson N, Lin J, Vogelstein B, Beerenwinkel N (2011) The temporal order of genetic and pathway alterations in tumorigenesis. PLoS ONE 6: e27136.

29. Verhaak RGW, Hoadley Ka, Purdom E, Wang V, Qi Y, et al. (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell 17: 98-110.

30. Rea S, Xouri G, Akhtar A (2007) Males absent on the first (MOF): from flies to humans. Oncogene 26: 5385–94.

31. Buyse IM, Shao G, Huang S (1995) The retinoblastoma protein binds to RIZ, a zinc-finger protein that shares an epitope with the adenovirus E1A protein. Proceedings of the National Academy of Sciences of the United States of America 92: 4467–71.

32. Chadwick RB, Jiang GL, Bennington Ga, Yuan B, Johnson CK, et al. (2000) Candidate tumor suppressor RIZ is frequently involved in colorectal carcinogenesis. Proceedings of the National Academy of Sciences of the United States of America 97: 2662–7.

33. Ernst A, Hofmann S, Ahmadi R, Becker N, Korshunov A, et al. (2009) Genomic and expression profiling of glioblastoma stem cell-like spheroid cultures identifies novel tumor-relevant genes associated with survival. Clinical cancer research 15: 6541–50.

34. Erbel-Sieler C, Dudley C, Zhou Y, Wu X, Estill SJ, et al. (2004) Behavioral and regulatory abnormalities in mice deficient in the NPAS1 and NPAS3 transcription factors. Proceedings of the National Academy of Sciences of the United States of America 101: 13648–53.

35. Huang J, Perlis RH, Lee PH, Rush AJ, Fava M, et al. (2010) Cross-Disorder Genomewide Analysis of Schizophrenia, Bipolar Disorder, and Depression. American Journal of Psychiatry 167: 1254–1263.

36. Moreira F, Kiehl TR, So K, Ajeawung NF, Honculada C, et al. (2011) NPAS3 demonstrates features of a tumor suppressive role in driving the progression of Astrocytomas. The American journal of pathology 179: 462–76.

37. van der Groep P, van der Wall E, van Diest PJ (2011) Pathology of hereditary breast cancer. Cellular oncology (Dordrecht) 34: 71–88.

38. Cowin P, Rowlands TM, Hatsell SJ (2005) Cadherins and catenins in breast cancer. Current opinion in cell biology 17: 499–508.

39. Green AR, Krivinskas S, Young P, Rakha Ea, Paish EC, et al. (2009) Loss of expression of chromosome 16q genes DPEP1 and CTCF in lobular carcinoma in situ of the breast. Breast cancer research and treatment 113: 59–66.

40. Usary J, Llaca V, Karaca G, Presswala S, Karaca M, et al. (2004) Mutation of GATA3 in human breast tumors. Oncogene 23: 7669–78.

41. Yan W, Cao QJ, Arenas RB, Bentley B, Shao R (2010) GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. The Journal of biological chemistry 285: 14042–51.

42. Kutuzov Ma, Bennett N, Andreeva AV (2010) Protein phosphatase with EF-hand domains 2 (PPEF2) is a potent negative regulator of apoptosis signal regulating kinase-1 (ASK1). The international journal of biochemistry & cell biology 42: 1816–22.

43. Medina PP, Romero Oa, Kohno T, Montuenga LM, Pio R, et al. (2008) Frequent BRG1/SMARCA4-inactivating mutations in human lung cancer cell lines. Human mutation 29: 617–22.

44. Kaye FJ (2002) RB and cyclin dependent kinase pathways: defining a distinction between RB and p16 loss in lung cancer. Oncogene 21: 6908–14.

45. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, et al. (2012) MuSiC: identifying mutational significance in cancer genomes. Genome research 22: 1589–98.

46. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics 173: 2187–98.

47. Mermel C, Schumacher S, Hill B, Meyerson M, Beroukhim R, et al. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology 12: R41.

48. Mitsui J, Takahashi Y, Goto J, Tomiyama H, Ishikawa S, et al. (2010) Mechanisms of genomic instabilities underlying two common fragile-site-associated loci, PARK2 and DMD, in germ cell and cancer cell lines. American journal of human genetics 87: 75–89.

49. Bednarek AK, Laflin KJ, Daniel RL, Liao Q, Hawkins KA, et al. (2000) WWOX, a Novel WW Domain-containing Protein Mapping to Affected in Breast Cancer Advances in Brief. Cancer research 60: 2140–2145.

50. Milo R, Kashtan N, Itzkovitz S, Newman MEJ, Alon U (2003) On the uniform generation of random graphs with prescribed degree sequences. eprint arXiv:cond-mat/0312028 .

51. Razick S, Magklaras G, Donaldson I (2008) iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9: 405.

Figures

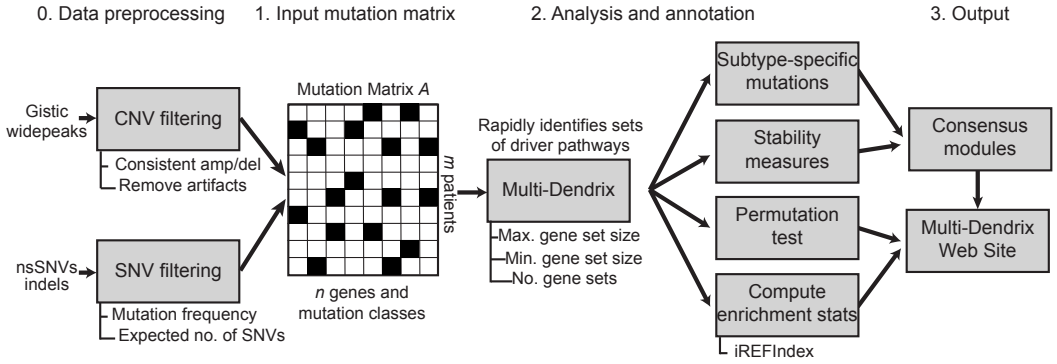

Figure 1: **The Multi-Dendrix pipeline.** Multi-Dendrix analyzes integrated mutation data from a variety of sources including single-nucleotide mutations and copy number aberrations. Multiple gene set are identified using a combinatorial optimization approaches. The output is analyzed for subtype-specific mutations and summarized across multiple values of the parameters: t, number of gene sets, and $k_{\max}$, maximum size per gene set.

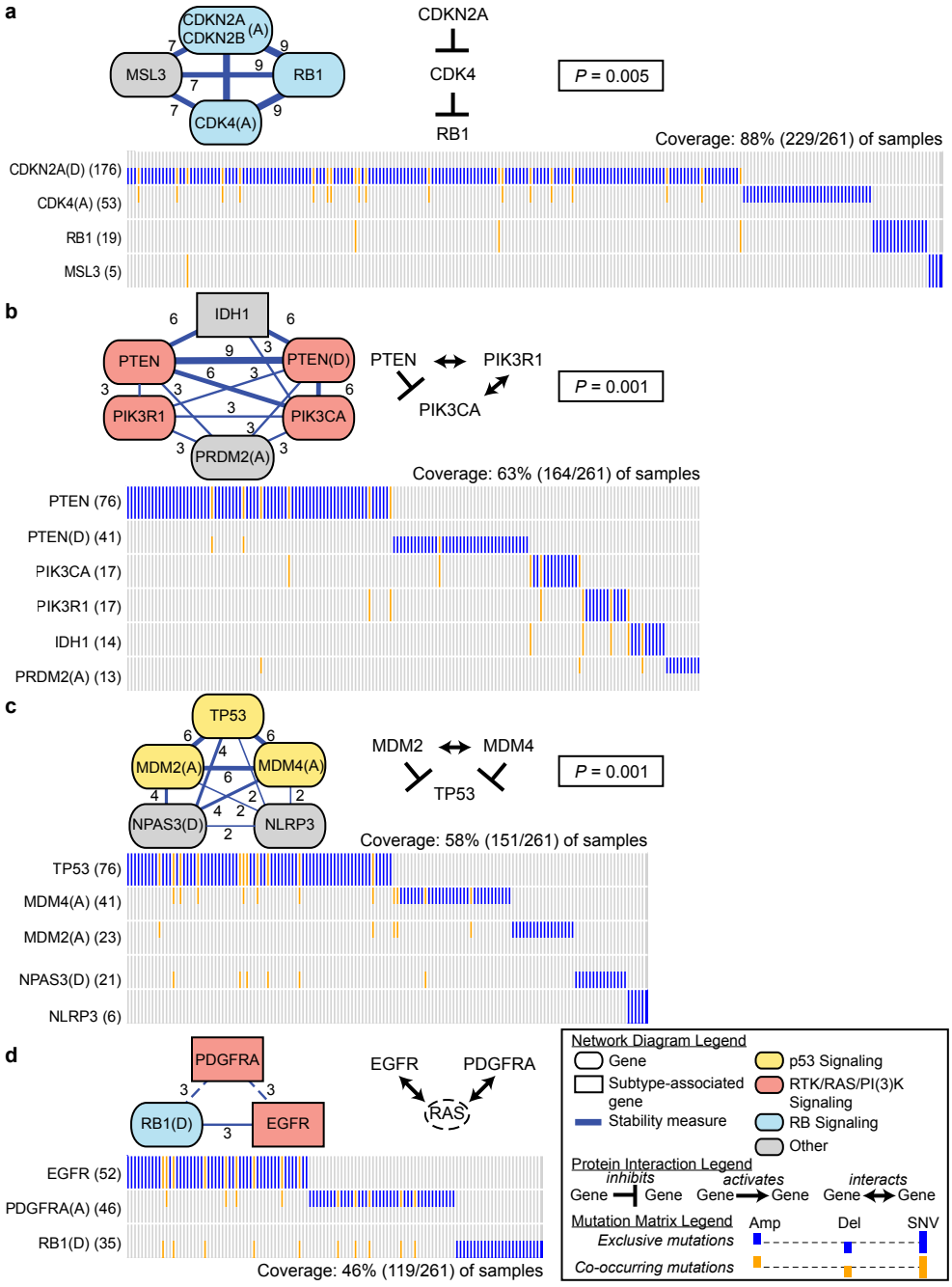

Figure 2: **Multi-Dendrix results on the GBM dataset.** (Left) Nodes represent genes in four modules found by Multi-Dendrix using $t = 2, \ldots, 4$ gene sets of minimize size $k_{\min} = 3$ and maximum size $k_{\max} = 3, \ldots, 5$. Genes with "(A)" appended are amplification events, genes with "(D)" appended are deletion events, and genes with no annotation are SNVs. Edges connect genes that appear in the same gene set for more than one value of the parameters, with labels indicating the fraction of parameter values for which the pair of genes appear in the same gene set. Color of nodes indicates membership in three signaling pathways noted in [7] as important for GBM: RB, p53, and RTK/RAS/PI(3)K signaling. Shape of nodes indicates genes whose mutations are associated with specific GBM subtypes, and dashed edges connect genes associated with different subtypes. The direct interactions statistic ν of this collection of gene sets is significant ($P = 0.002$). (Middle) Known interactions between proteins in each set and p-value for observed number of interactions. (Right) Mutation matrix for each of four modules with mutual exclusive (blue) and co-occurring mutations (orange).

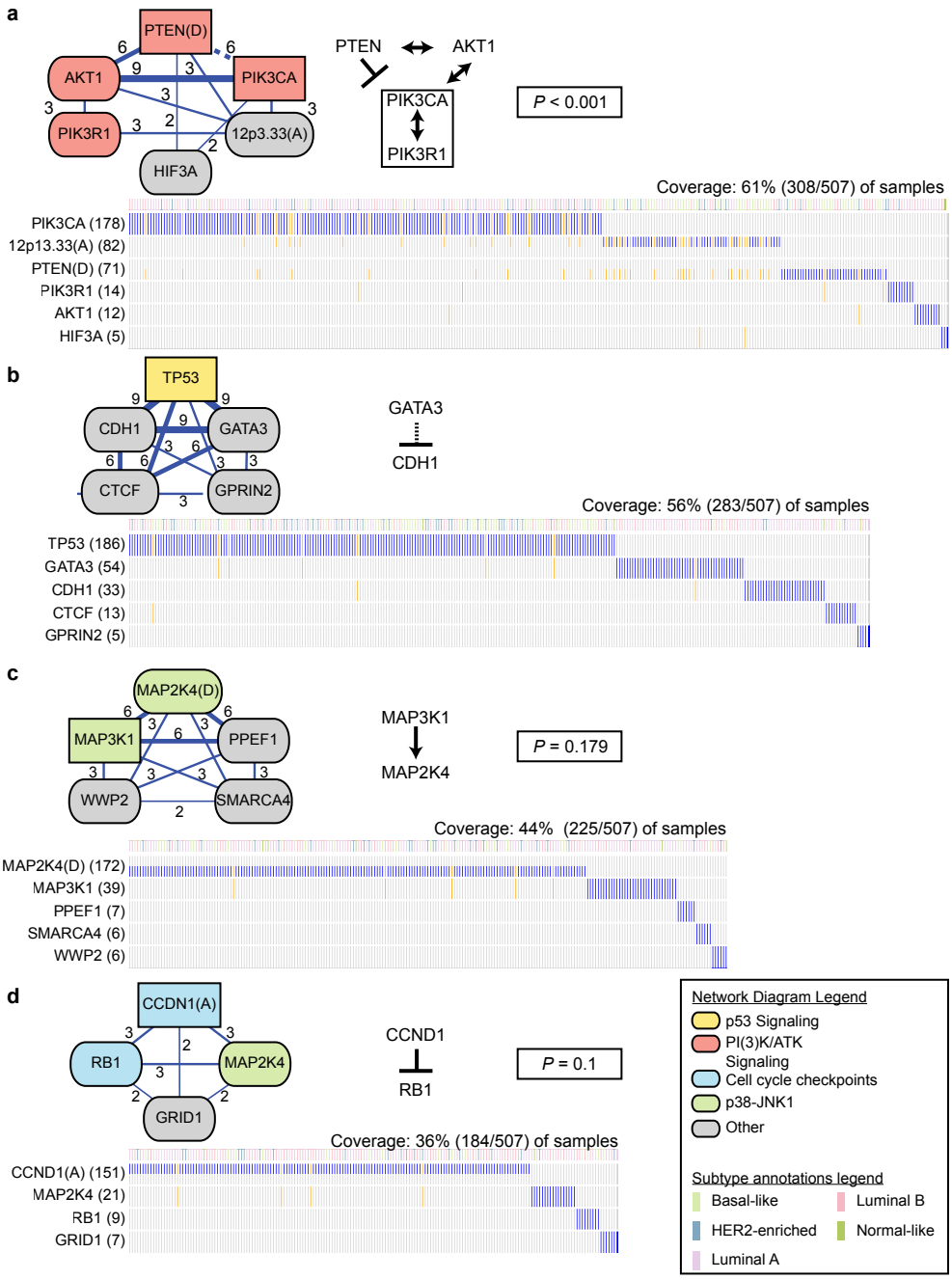

Figure 3: **Multi-Dendrix results on the BRCA dataset.** Graphical elements are as described in Figure 2 caption, except for the following. Color of nodes indicates membership in four signaling pathways noted in [8] as important for BRCA: p53 signaling, PI(3)K/AKT signaling, cell cycle checkpoints, and p38-JNK1. The top row of each mutation matrix annotates the subtype of each patient. The regulatory interaction between GATA3 and CDH1 is shown as a dashed line. The direct interactions statistic ν of this collection of gene sets is significant ($P < 0.001$).

Tables

	Avg. distance d from planted pathways		
q	Multi-Dendrix	Iter-RME	Iter-Dendrix
0.0	0.02 ± 0.19	$\mathbf{0.01 \pm 0.12}$	0.30 ± 0.86
0.0001	*0.02 ± 0.18*	***0.01 ± 0.16***	*0.30 ± 0.86*
0.0005	$\mathbf{0.04 \pm 0.23}$	0.10 ± 0.40	0.35 ± 0.89
0.001	***0.10 ± 0.35***	*0.32 ± 0.60*	*0.44 ± 1.01*
0.005	$\mathbf{0.44 \pm 0.71}$	–	0.75 ± 1.07
0.01	$\mathbf{1.03 \pm 1.00}$	–	1.20 ± 1.15
0.015	$\mathbf{1.68 \pm 1.16}$	–	1.78 ± 1.26
0.02	$\mathbf{2.17 \pm 1.24}$	–	2.21 ± 1.29

Table 1: A comparison of the algorithms on simulated mutation data with varying passenger mutation probability q. Italicized rows correspond to values of q approximated from real cancer datasets. Each entry is mean (μ) and standard deviation (σ) (across 1000 simulations) of the distance $d(\mathbf{P}, \mathbf{M})$ between the planted set of pathways $\mathbf{P}$ and the collections $\mathbf{M}$ found by each algorithm. The minimum distance $d = 0$ indicates an algorithm found the planted pathways exactly, while the maximum distance $d = 16$ indicates that an algorithm did not find *any* of the genes in the planted pathways. Bold text indicates the top performing algorithm for each value of q. Multi-Dendrix is the top performer for all values of q except the smallest $q = 0.0001$. The differences between Multi-Dendrix and both Iter-Dendrix and Iter-RME are statistically significant ($p < 0.01$) for $0.0005 \leq q \leq 0.01$. For $q > 0.001$, Iter-RME did not complete after 24 hours of runtime.

	Avg. runtime (secs)		
q	Multi-Dendrix	Iter-RME	Iter-Dendrix
0.0001	***0.28***	*19.22*	*609.79*
0.0005	**0.28**	17.36	621.22
0.001	***0.50***	*123.16*	*610.21*
0.005	**1.46**	$>$ 86, 400	672.43
0.01	**2.60**	$>$ 86, 400	711.55
0.015	**4.06**	$>$ 86, 400	730.85
0.02	**4.93**	$>$ 86, 400	727.82

Table 2: A comparison of the runtimes of Multi-Dendrix, Iter-RME, and Iter-Dendrix on simulated mutation data with varying passenger mutation probability q. Runtimes for each algorithm are reported as the mean runtime of 10 runs for each value of q. Note that Multi-Dendrix has runtimes under 5 seconds for all datasets, and is orders of magnitude faster than the other methods for each q. For Iter-RME, we report a runtime of $>$86,400 seconds for $q \geq 0.005$ as Iter-RME did not complete within one day. Simulations were performed on machines running 64-bit Debian Linux with Xeon 2.8GHz processors and a maximum of 8GB of available memory.

Supporting Information

GBM(2008) and Lung datasets

Somatic mutation processing in targeted gene sequencing studies.

For the GBM(2008) and lung adenocarcinoma datasets, we used all genes with non-synonymous single-nucleotide mutations or small indels in at least two patients. For GBM(2008), we also included copy number variants (CNVs), adding a CNV mutation type for a gene if the gene had a CNV of the same type (amplification or deletion) in at least 90% of the patients with a CNV. If a set of genes was mutated in the same patients, we merged these genes into a single metagene. We manually merged CYP27B1 into the metagene containing CDK4, as described in [23].

Multi-Dendrix results

We ran Multi-Dendrix and Iter-Dendrix on the GBM(2008) and Lung datasets using the same parameters our analysis of the GBM and BRCA datasets: minimum gene set size $k_{\min} = 3$, maximum gene set size $k_{\max} = 3 \ldots 5$, number t of gene sets $t = 2 \ldots 4$, for a total of 9 parameter combinations. We analyze Multi-Dendrix's results below, and present a comparison of Multi-Dendrix's and Iter-Dendrix's results in § Comparison of Multi-Dendrix and Iter-Dendrix results.

GBM(2008). Figure S6 shows Multi-Dendrix's results on the GBM(2008) dataset. Multi-Dendrix finds four modules that contain known cancer genes. In addition eight genes appear in the results for one or two parameter choices, but with different groups of genes. Two modules contain triples that are consistent across all parameter choices: {CDKN2B, RB1, CDK4}, and {CDKN2A, TP53, DTX3}. Note that due to different copy number processing, the genes CDKN2A and CDKN2B appear as different mutation classes in the GBM(2008) data, while they were merged in a metagene in the GBM data presented above. Beyond this difference, the triple {CDKN2B, CDK4, RB1} is the same as the triple in the first module found by Multi-Dendrix on the GBM dataset, and contains two known interactions (for further analysis of this module, see § Mutually exclusive sets in Glioblastoma (GBM)). For six of the nine parameter choices, this module also contains the well-known cancer gene ERBB2, a member of the RTK pathway important for glioblastoma [7].

The second triple, {CDKN2A, TP53, DTX3 }, contains two frequently mutated GBM cancer genes that interact: TP53 and CDKN2A. For six of the nine parameter choices, this module also contains CDC123, although CDC123 is mutated in only two samples. The third module contains both EGFR and NF1, two genes identified by [29] to be associated with the Classical and Mesenchymal subtypes, respectively. These two genes account for 52 of the 70 mutated samples in this module, and thus the exclusivity of mutations in this module is likely due to subtype-specific mutations. Interestingly, the two genes associated with the Proneural subtype, IDH1 and PDGFRA, are not found in any of the results (IDH1 was not included in the list of targeted genes for the GBM(2008) dataset). The weight W' of all collections identified by Multi-Dendrix are significant ($p < 0.0001$).

Lung. Figure S7 shows Multi-Dendrix's results on the Lung dataset. In contrast to the other datasets, the output of Multi-Dendrix varies widely with the choice of parameters (e.g. STK11 has 8 edges with weight ≥ 0.2 in the summarization graph). There are two large modules: one of size four and one of size thirteen, as well as 10 additional genes that appear with another gene for only one choice of parameters (and thus are nodes of degree 1 in the summarization graph).

The module of size four includes the triple TP53, ATM, and PAK4 found for all values of parameters. This triple is nearly perfectly exclusive; it covers 79 patients (i.e. $\Gamma(M) = 79$) with a coverage overlap $\omega(M) = 1$. TP53 and ATM are both tumor suppressors that interact in response to DNA damage, and

have been shown previously to be recurrently mutated in lung adenocarcinoma [52]. PAK4 is involved in both cellular survival and angiogenesis, and has been implicated in tumor progression in a variety of cancers [53], though we did not find reports of its role in lung adenocarcinoma. While PAK4 is mutated in only three patients in the Lung dataset, it is still significantly exclusive with ATM ($p = 0.0008$ by Fisher's exact test), and may deserve further inquiry.

The module of size 13 includes two gene triples that are grouped together for most parameter choices. The triple {EGFR, KRAS, EPHB1} contains the well-known interaction between EGFR and KRAS, two proteins whose mutations are important for lung adenocarcinoma [27]. Similar to EGFR, EPHB1 is a receptor tyrosine kinase, and mutations in this gene (as in mutations in EGFR) were reported to be associated with higher copy number and mRNA expression levels in [27]. The triple {STK11, NF1, NTRK1} contains two well-known cancer genes: STK11 and NF1. Mutations in STK11 have long been a hallmark of lung adenocarcinoma, but the discovery of mutations in the tumor suppressor NF1 was more recent [27]. While the genes in this triple have no known interactions, the triple is nearly perfectly exclusive (coverage $\Gamma(M) = 54$; coverage overlap $\omega(M) = 1$) which indicates that lung adenocarcinoma patients need inactivating mutations in just one of these genes. Thus, they may have an uncharacterized relationship in lung adenocarcinoma.

While Multi-Dendrix's results on the Lung dataset are promising, we did not conduct further analysis or comparison of these results because of the inconsistency across parameter choices. We note that a likely explanation for this inconsistency is that the processed data includes a relatively small number of 163 samples and only includes mutations from targeted sequencing of 190 genes, missing other mutations and copy number aberrations in these samples.

Comparison of Multi-Dendrix and Iter-Dendrix results

We compare the gene sets found by Multi-Dendrix and those found by Iter-Dendrix on the GBM(2008), GBM, and BRCA datasets, computing the overlap between the results of the algorithms and comparing the number of known protein-protein interactions in the results.

Overlap

We compared the distances between collections found by Multi-Dendrix and Iter-Dendrix, using the symmetric difference function $d(\mathbf{M}, \mathbf{I})$ defined in § Simulated data. Table S4 reports these distances. There are many differences between the two algorithms on the GBM(2008) and BRCA datasets. On the GBM dataset, there is only one difference: Iter-Dendrix includes the gene IRF5 in the gene set that includes RB1, CDK4(A), and CDKN2A/CDKN2B(D) for $k_{\max} = 5$.

Interactions

Tables S2 and S3 show the results of the direct interactions test on collections found by Multi-Dendrix and Iter-Dendrix (for details of the test, see § Evaluating known interactions). In total, across all three datasets, nearly all collections found by Multi-Dendrix are significantly enriched for interactions (24/27 collections, $p < 0.05$), slightly more than are found by Iter-Dendrix (20/27 collections, $p < 0.05$). The largest difference is on the BRCA dataset, where all collections found by Multi-Dendrix have lower p-values than Iter-Dendrix. In addition, Multi-Dendrix finds all nine collections significantly enriched for interactions ($p \leq 0.05$), while Iter-Dendrix finds just two. This difference is amplified by the fact Iter-Dendrix finds five collections with $p \geq 0.15$ (by comparison the largest p-value for Multi-Dendrix on BRCA is 0.01). Thus, overall Multi-Dendrix finds collections of gene sets that are more enriched for interactions.

We do not compare the enrichment for interactions of individual gene sets found by Multi-Dendrix and Iter-Dendrix. Iter-Dendrix is a greedy algorithm and is thus guaranteed to find the same gene sets multiple

times as we vary t. Thus, it is more reasonable to evaluate the stable modules for enrichment for interactions, as presented in the § Mutually exclusive sets in Glioblastoma (GBM) and § Mutually exclusive sets in Breast Cancer (BRCA).

Comparison of Multi-Dendrix and RME

We compared Multi-Dendrix and Iter-RME on the GBM(2008), GBM, and BRCA datasets. We ran RME with default parameter values, including the default minimum mutation frequency of $\geq 10\%$. After removing genes mutated in fewer than 10% of samples, the GBM(2008), GBM, and BRCA datasets contains 18, 10, and 28 genes, respectively. As a result, using the same parameters as above, RME can find at most three gene sets in the GBM dataset of the minimum size 3, since there are only 10 genes with mutation frequencies $\geq 10\%$ the GBM dataset. We ran RME for up to $t = 4$ gene sets of size $k_{\max} \in [2, 5]$ so that RME could find $t = 4$ gene sets in the GBM and BRCA data. Note that unlike Dendrix, RME has a parameter for setting the maximum gene set size, equivalent to $k_{\max}$, and returns gene sets of size $[2, k_{max}]$. Thus, at each iteration, we chose the gene set $P, 2 \leq |P| \leq k_{\max}$, such that P had the largest RME algorithmic significance score.

For all values of the parameters t and $k_{\max}$, Iter-RME finds collections where each gene set has size 2. This is due to a combination of RME's algorithmic significance score, which is highest for gene sets of size 2, and also because there are only 10-28 genes in each dataset. It is difficult to compare Multi-Dendrix and Iter-RME using the tests presented in Section , as Iter-RME finds only 4 gene sets of size two in the GBM(2008), GBM, and BRCA datasets. It is unclear how interesting these pairs are: only three pairs of genes from all datasets interact in either the iRefIndex or KEGG protein-protein interaction networks, despite these pairs being chosen from the most highly mutated genes in each dataset (Table S7). Thus, the runtime requirements of RME limit its use to only a small subset of of highly mutated genes, and precludes RME from finding interesting gene sets of size greater than 2 on the tested datasets. In contrast, with the same parameters, Multi-Dendrix finds larger gene sets of up to 5 genes, many of which show enrichment for interactions (see Results).

Subtype analysis

GBM subtypes

We annotate the Multi-Dendrix results for associations with known subtypes. [29] derived four subtypes of GBM from gene expression data, and these classifications have become the standard in TCGA analysis. However, we were unable to directly use the sample annotations from [29] for the GBM mutation dataset since there are only 39 samples in common between the two datasets. Instead, we identify whether the gene sets produced by Multi-Dendrix contain any of the subtype-specific genes (IDH1, PDGFRA, EGFR, and NF1) reported in [29]. Figure 2 shows that one of the four modules found by Multi-Dendrix contains three of these four genes (all except IDH1). These genes are all members of the RTK/RAS/PI(3)K signaling pathway as annotated in [7], although they do not interact in iREFIndex or KEGG.

We also considered associations between subtypes defined by other automated clustering of gene expression data from TCGA. We downloaded GBM subtypes generated from consensus hierarchical clusters of mRNA expression data from the Broad's Firehose website. This dataset contains subtype information for 529 samples, 224 of which appear in our GBM mutation dataset, and maps patients to one of three subtypes ("1", "2", or "3"). Note that [29] identified four subtypes, and thus there is no one-to-one correspondence between the two subtype classifications. We determined if mutations in a gene were associated with subtype using Fisher's exact test. Table S5, lists all significant associations ($P < 0.01$ following Bonferroni multiple hypothesis correction). Notably, Consensus 1 and Consensus 2 have no significant associations, while the Consensus 3 subtype has all eight of the the significant associations. These include the markers IDH1

and PDGFRA in the Proneural subtype of [29]. The other two subtype markers found by [29], EGFR and NF1, which characterize the Classic and Mesenchymal subtype, do not appear associated with any of the Consensus clusters. Because the overlap between these clusters and those of [29] was imperfect, we used the genes that [29] reported to have subtype-specific mutations.

BRCA subtypes

We ran Multi-Dendrix on BRCA mutation data restricted to patients from each of the subtypes detailed in [8]. Because the number of patients within each subtype is not distributed evenly, we varied Multi-Dendrix's exclusivity parameter α for each subtype (Luminal A: $\alpha = 2$; Luminal B: $\alpha = 1.5$; HER2-enriched: $\alpha = 1$; Basal-like: $\alpha = 1$). We discuss these results in § Mutually exclusive sets in Breast Cancer (BRCA). In addition, we report all significant associations (significance level: $p < 0.01$) between genes or mutation classes and BRCA subtypes in Table S6.

Multi-Dendrix results without SNV filtering

We tested Multi-Dendrix on the GBM and BRCA mutation matrices constructed without any filtering of SNV date. Thus, we included all 10,987 (respectively 12,248) genes with at least one non-synonymous SNV in any sample. We also included the copy number aberrations output from GISTIC as described in § $Construction of mutation matrices$. The resulting mutation matrices included 11,023 mutation classes in 261 GBM samples and 12,281 mutation classes in 507 BRCA samples. We report the stable modules identified by Multi-Dendrix here. The modules identified on the mutation data without SNV processing have symmetric distance Δ of 7 and 36 from the modules identified from the mutation data with SNV processing. In the GBM data, the differences are minor: for all four modules reported in the main text, at least four genes are in common with the corresponding module found with the unfiltered mutation data. In particular, the Multi-Dendrix results with the unfiltered data contain the following changes:

- LMNB2 and SUSD3 appear in the module containing CDKN2A(D), CDK4(A), RB1, and MSL3.
- KRTAP5-5 replaces NLRP3 in the module containing TP53, MDM4(A), MDM2(A), and NPAS3(D).
- HMCN1 replaces PIK3R1 for two parameter choices in the module containing PIK3CA, PTEN(D), PTEN, IDH1, and PRDM2(A).
- MDN1 appears in the module containing EGFR, PDGFRA(A), and RB1(D).

On the BRCA data, the differences are larger: for three of the four modules reported in the main text, only two of the genes in those modules are placed in the same module when Multi-Dendrix is run on the unfiltered mutation data. However, the fourth module (TP53, GATA3, CDH1, CTCF, and GPRIN2) remains exactly the same. In particular, the Multi-Dendrix results with the unfiltered data contain the following changes:

- HUWEI, SVIL, and LPHN3 replace AKT1, PIK3R1, 12p13.33(A), and HIF3A in the module containing PTEN(D) and PIK3CA.
- LAMA2 and USP34 replace MAP2K4 and GRID1 in the module containing CCND1(A) and RB1.
- TLN1, MYH7, DNAH1, and NOTCH4 replace SMARCA4, PPEF1, and WWP2 in the module containing MAP2K4(D) and MAP3K1.

Many of the genes that appear only in the modules found with unfiltered data are long genes or genes with high background mutation rates that are more likely to be mutated in more samples than other genes.

The fifteen genes that only appear in modules on the unfiltered GBM and BRCA data have an average length of greater than 8500, and an average BMR of greater than 5×10^{-6} (BMRs were calculated from SNV data from twelve cancers). Such outlier genes are a potential source of false positives, not only for Multi-Dendrix, but also for any method that attempts to identify cancer genes by their recurrence across samples [6]. Multi-Dendrix identified the stable modules in 6,863 and 14,491 seconds, respectively.

References

52. Greulich H (2010) The genomics of lung adenocarcinoma: opportunities for targeted therapies. Genes & cancer 1: 1200-10.

53. Kesanakurti D, Chetty C, Rajasekhar Maddirela D, Gujrati M, Rao JS (2012) Functional cooperativity by direct interaction between PAK4 and MMP-2 in the regulation of anoikis resistance, migration and invasion in glioma. Cell death & disease 3: e445.

4

Supporting Tables

m	**100**	**200**	**400**	**800**	**1600**	**3200**	**6400**	**12800**	**22000**
$q = 0.02$	51.8	93.4	185.4	357.8	722.7	1409.3	2826.6	5663.4	9698.4
$q = 0.0001$	24.2	34.9	54	88.1	169.1	312.9	617.2	1236	2108.4

Table S1: **Average number of genes input to Multi-Dendrix across 100 simulated datasets.** We use a minimum threshold of nq, the expected number of n total samples with a mutation. We report the values for different number m of genes and the lowest and ($q = 0.0001$) highest ($q = 0.02$) passenger mutation probabilities. On all simulated datasets, Multi-Dendrix's average runtime is < 1 hour.

	GBM(2008)			GBM			BRCA		
$k_{\max}$	3	4	5	3	4	5	3	4	5
$t = 2$	0.005	0.026	0.027	< 0.001	< 0.001	< 0.001	0.003	0.007	0.006
$t = 3$	< 0.001	0.021	0.375	< 0.001	0.001	0.001	< 0.001	0.003	0.007
$t = 4$	0.253	0.036	0.314	0.013	0.006	0.002	< 0.001	0.001	0.01

Table S2: p-values for the number of observed protein-protein interactions in Multi-Dendrix results (direct interactions test) for different values of parameters t, the number of gene sets, and $k_{\max}$, the maximum gene set size. The minimum gene set size, $k_{\min} = 3$ for all runs. The p-values were calculated from 1000 permuted networks constructed from the union of the KEGG and iRefIndex PPI networks.

	GBM(2008)			GBM			BRCA		
k	3	4	5	3	4	5	3	4	5
$t = 2$	0.005	0.026	0.032	< 0.001	< 0.001	< 0.001	0.166	0.17	0.184
$t = 3$	< 0.001	0.021	0.046	< 0.001	0.001	0.001	0.05	0.26	0.72
$t = 4$	0.001	0.002	0.043	0.013	0.006	0.002	0.003	0.005	0.103

Table S3: p-values for the number of observed protein-protein interactions in Iter-Dendrix results (direction interactions test) for different values of parameters t, the number of gene set, and $k_{\max}$, the maximum gene set size. The minimum gene set size, $k_{\min} = 3$ for all runs. The p-values were calculated from 1000 permuted networks constructed from the union of the KEGG and iRefIndex PPI networks.

	GBM(2008)			GBM			BRCA		
$k_{\max}$	3	4	5	3	4	5	3	4	5
$t = 2$	0	0	3	0	0	1	6	8	10
$t = 3$	6	6	13	0	0	1	9	12	15
$t = 4$	4	10	15	0	0	1	12	16	20

Table S4: The size of the symmetric difference $d(\mathbf{M}, \mathbf{I})$ for collections $\mathbf{M}, \mathbf{I}$ found by Multi-Dendrix and Iter-Dendrix, respectively.

Gene	Subtype	p-value
IDH1	Consensus 3	1.24E-05
ARID2(D)	Consensus 3	3.66E-05
PDGFRA(A)	Consensus 3	3.79E-05
CDK4(A)	Consensus 3	0.00098192
KIF4B	Consensus 3	0.001198772
BBS1	Consensus 3	0.006889494
TP53	Consensus 3	0.008230181
MET(A)	Consensus 3	0.009878469

Table S5: Significant associations ($p < 0.01$) between mutations in genes (SNVs, amplifications "(A)", or deletions "(D)") and three subtypes from GBM consensus clusters. p-values were calculated using Fisher's exact test with a Bonferroni correction for multiple hypotheses.

Gene	Subtype	p-value
TP53	Basal-like	1.53E-18
12p13.33(A)	Basal-like	2.17E-10
PTEN(D)	Basal-like	3.16E-06
MAP3K1(D)	Basal-like	7.93E-06
5q21.3(D)	Basal-like	0.000109176
PIK3CA(A)	Basal-like	0.000417525
RB1(D)	Basal-like	0.001671689
11p13(A)	Basal-like	0.002899894
FGF2(A)	Basal-like	0.003068744
EPS8L1	Basal-like	0.00649918
MYC(A)	Basal-like	0.007955101
IGF1R(A)	Basal-like	0.008512299
ERBB2(A)	HER2-enriched	5.07E-26
TP53	HER2-enriched	8.64E-09
MIR21(A)	HER2-enriched	4.24E-08
6q21(A)	HER2-enriched	0.001672642
4q13.3(A)	HER2-enriched	0.002513158
SRPR	HER2-enriched	0.003786332
ATP1A4	HER2-enriched	0.007586379
ERBB3	HER2-enriched	0.007586379
PIK3CA	Luminal A	7.39E-06
MAP3K1	Luminal A	9.94E-05
CCND1(A)	Luminal B	1.36E-08
KCNB2	Luminal B	0.001123083
3p25.1(A)	Luminal B	0.001333521

Table S6: Significant associations ($p < 0.01$) between mutations (SNVs, amplifications "(A)", or deletions "(D)") in the four BRCA subtypes. p-values were calculated using Fisher's exact test with a Bonferroni correction for multiple hypotheses.

Dataset	Iteration	Genes	Interact
GBM(2008)	1	RB1, CDKN2B	No
	2	NF1, EGFR	No
	3	OS9, CDKN2A	No
	4	TP53, MDM2	Yes
GBM	1	CDK4(A), CDKN2A_CDKN2B(D)	Yes
	2	PTEN, PTEN(D)	N/A
	3	MDM4(A), TP53	Yes
	4	EGFR, PDGFRA(A)	No
BRCA	1	TP53, GATA3	No
	2	PIK3CA, PTEN(D)	Yes
	3	MAP2K4(D), FOXA1(A)	No
	4	8p11.23(A), ERBB2(D)	No

Table S7: Gene sets found by Iter-RME in the GBM(2008), GBM, and BRCA datasets (after removing genes with mutation frequency $< 10\%$) for maximum gene set size $k_{\max} = 2, \ldots, 5$ and and number t of gene sets $t = 2, \ldots, 4$. For all values of $k_{\max}$, Iter-RME returned only gene sets of size 2. "Iteration" column denotes the index of each gene set P_i returned in each iteration of RME. Only 4/12 gene sets contain an interacting pair of genes according to the union of the KEGG and iRefIndex protein-protein interaction network.

Supporting Figures

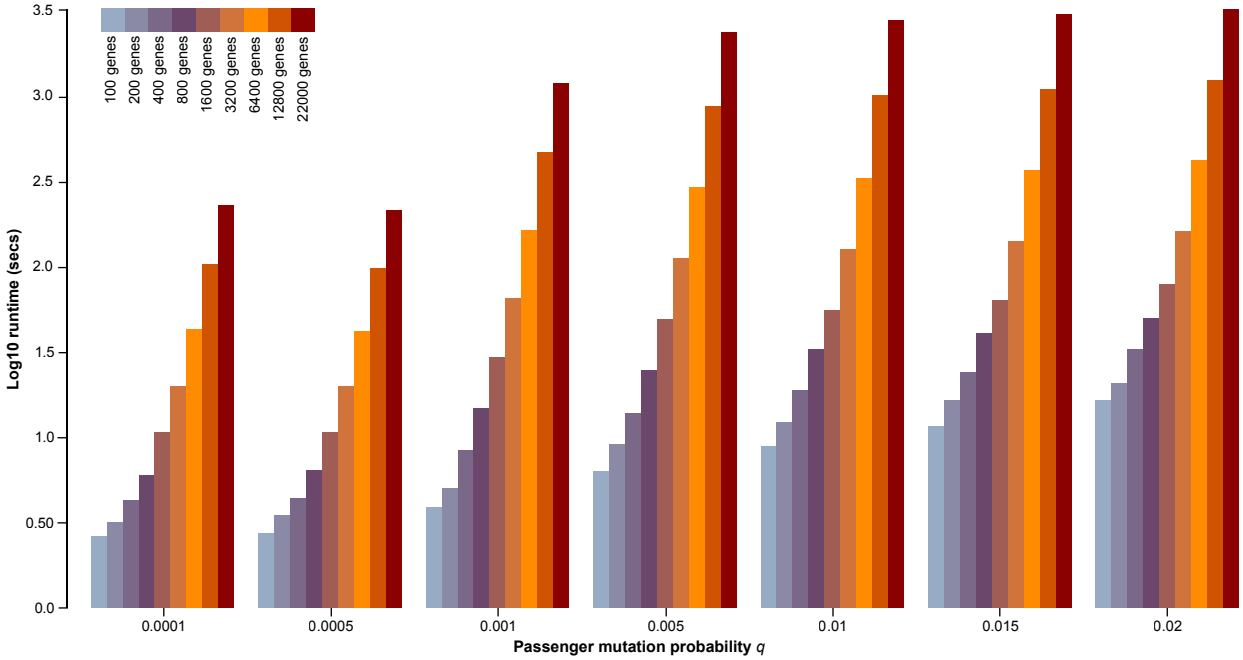

Figure S1: **Multi-Dendrix runtimes on simulated datasets**. Average runtime (secs) of Multi-Dendrix on simulated data with 1000 patients and varying passenger mutation probabilities q and number of genes. For each set of parameters, average runtime across 10 simulations is reported. The average runtime of Multi-Dendrix is under 3500 seconds for each dataset.

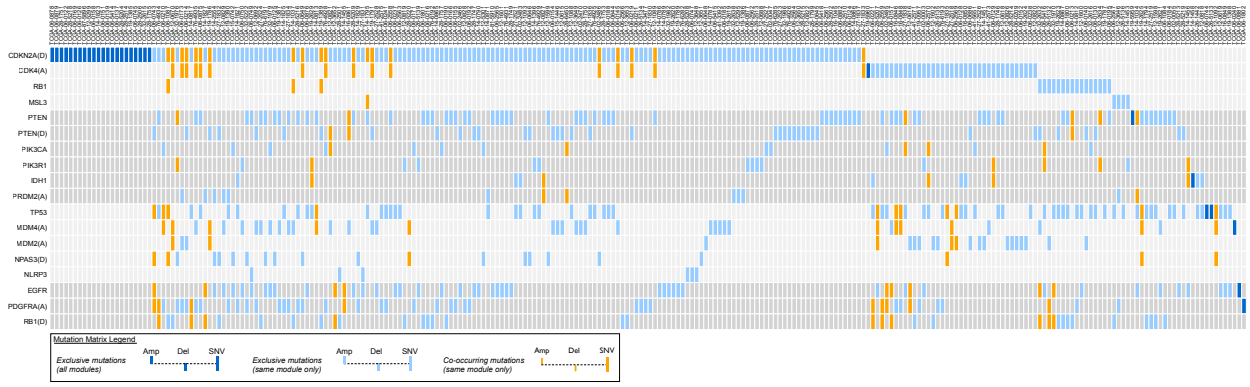

Figure S2: **Multi-Dendrix results on GBM dataset as one mutation matrix**. The four modules are shown in alternating gray backgrounds. For each module, patients are sorted first by those with mutations in only the module (dark blue), then by patients that have just one mutation in the module (light blue), and finally by patients that have co-occurring mutations in that module (orange). Upticks indicate amplification, downticks indicate deletions, and full ticks indicate SNVs. The four modules found by Multi-Dendrix on the GBM dataset cover 98.8% (258/261) of the patients in the GBM mutation data.

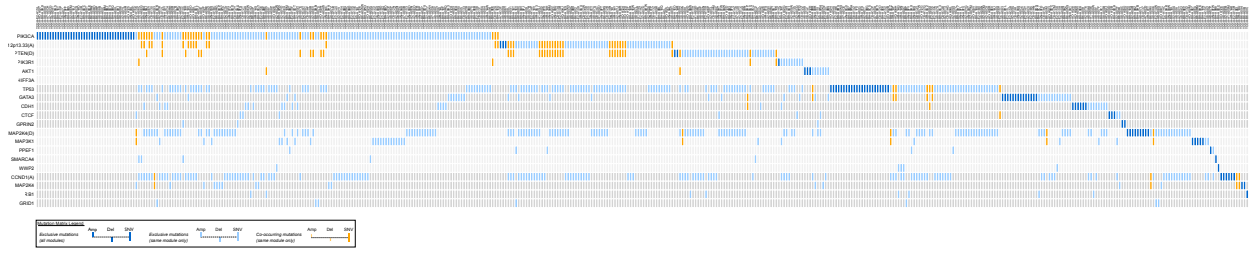

Figure S3: **Multi-Dendrix results on BRCA dataset as one mutation matrix**. Representation is as in Figure S2. The four modules found by Multi-Dendrix cover 91.9% (466/507) of the patients in the BRCA mutation data.

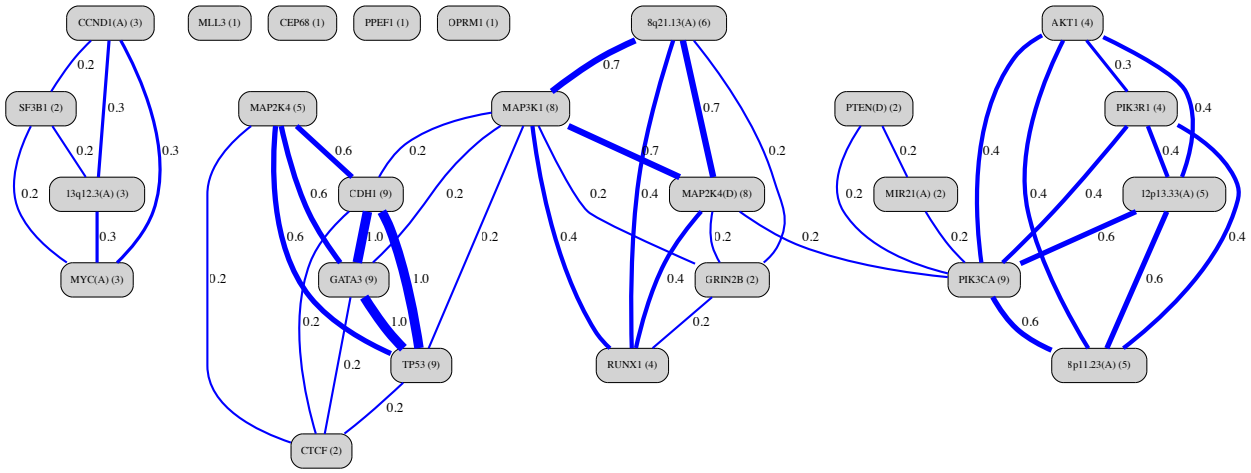

Figure S4: **Multi-Dendrix results on the BRCA dataset with $\alpha = 1$.** Four genes (genes with degree zero) are output for only one choice of parameter values. The remaining genes form two connected components indicating that the collections are variable across different values of the parameters. In comparison Multi-Dendrix with $\alpha = 2.5$ (Figure 3) finds more stable results on this dataset.

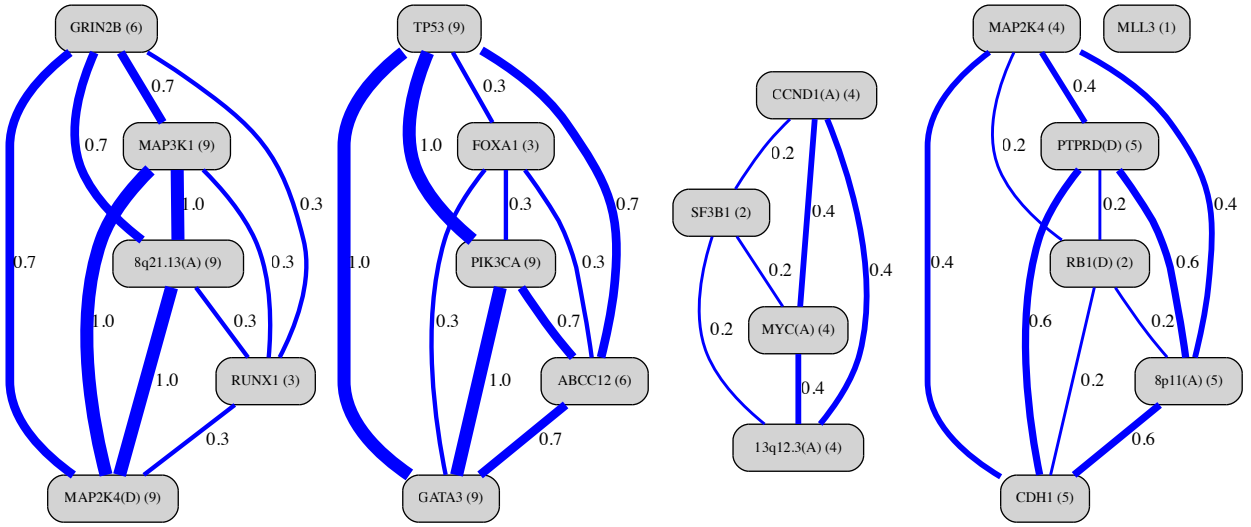

Figure S5: **Iter-Dendrix results on the BRCA dataset.** The modules identified by Iter-Dendrix group the most frequently mutated genes together (e.g. TP53 and PIK3CA) and thus there is a large number of samples with co-occurring mutations in each module. In comparison Multi-Dendrix with $\alpha = 2.5$ (Figure 3) finds more stable with less co-occurrence of mutations within modules on this dataset.

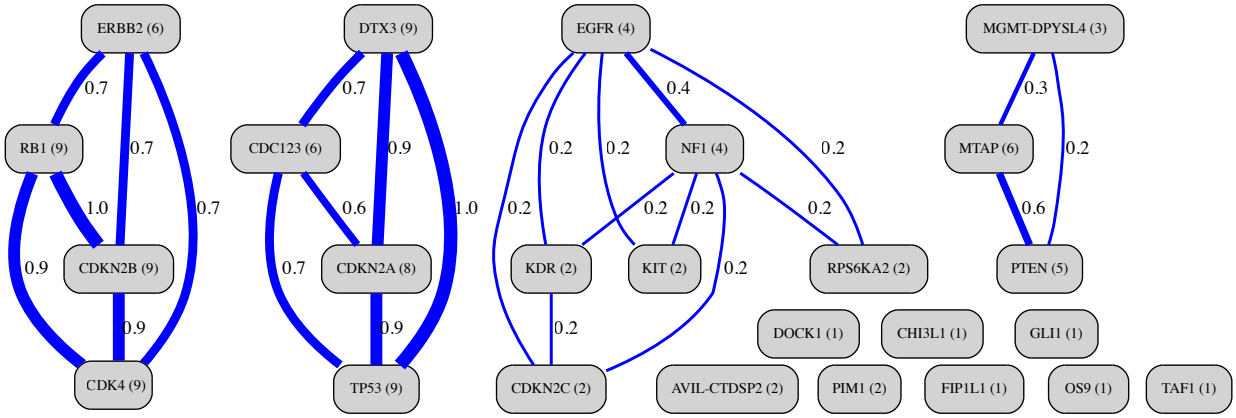

Figure S6: **Multi-Dendrix results on the GBM(2008) dataset.**

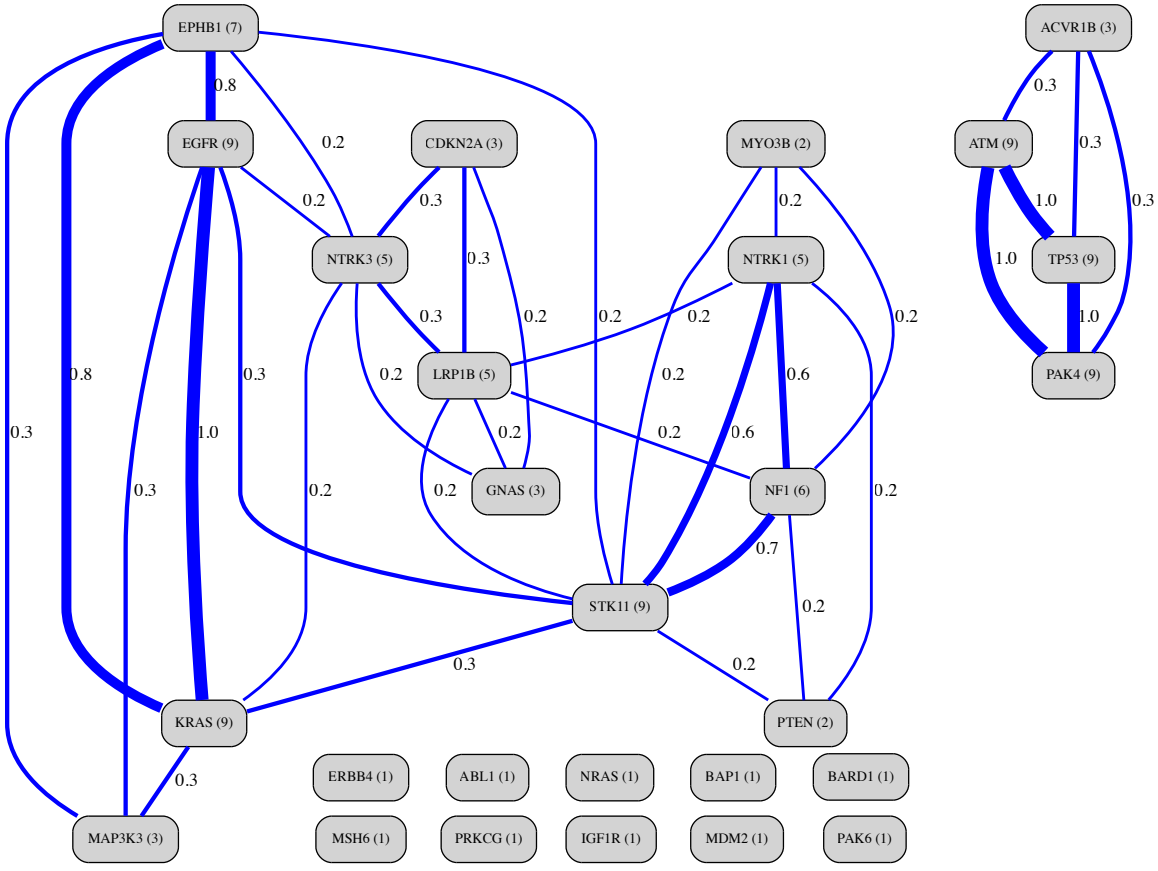

Figure S7: **Multi-Dendrix results on the Lung dataset.**

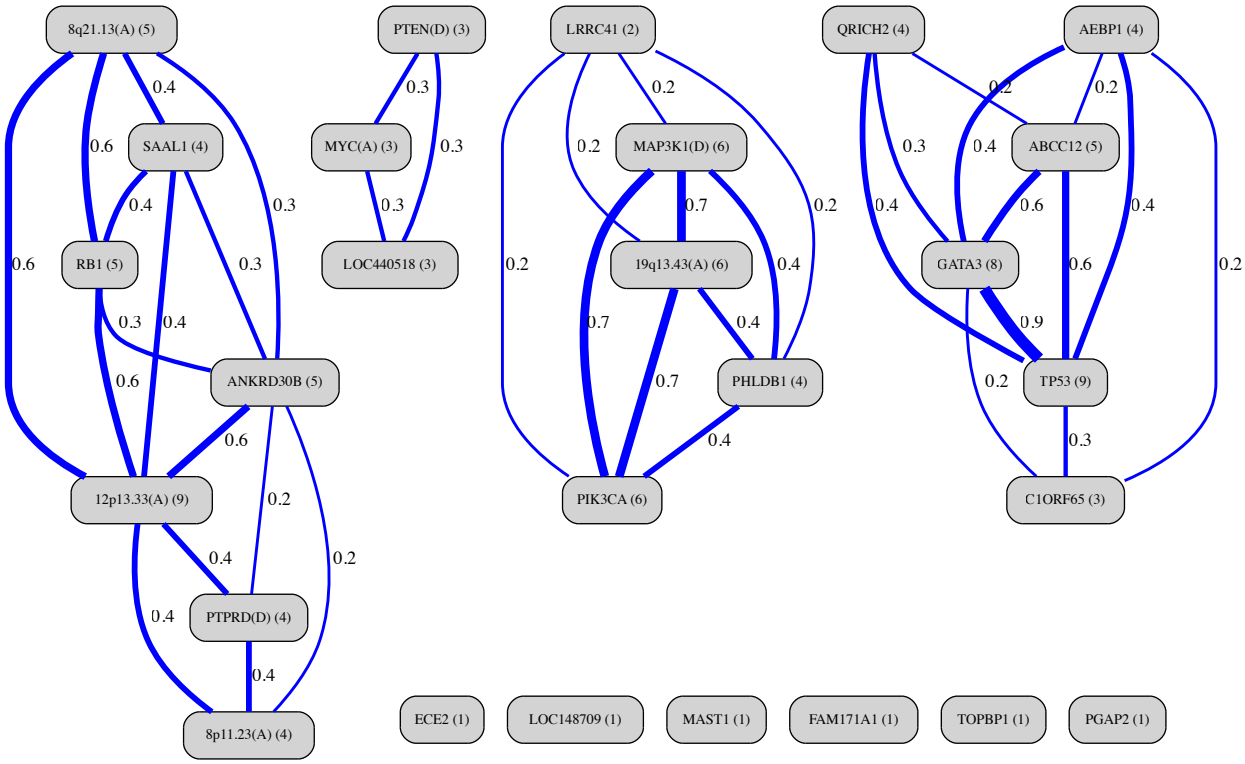

Figure S8: **Multi-Dendrix results on the BRCA dataset restricted to Basal-like patients.**

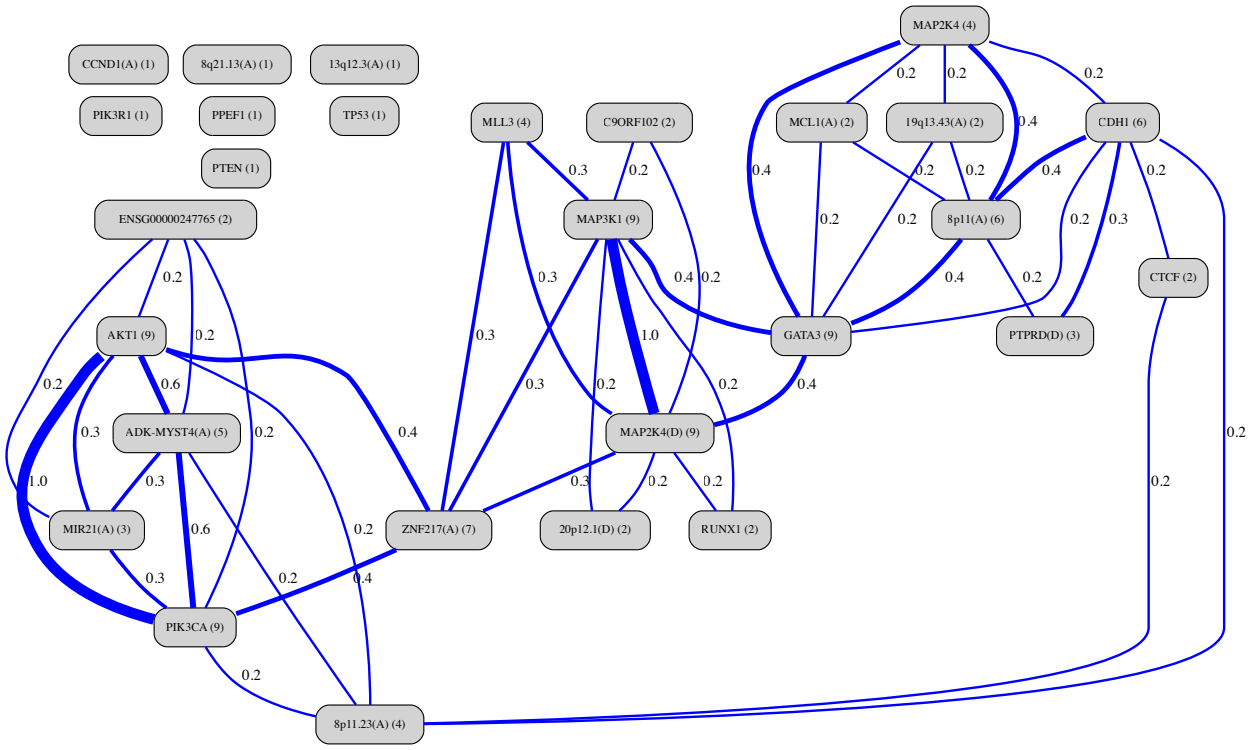

Figure S9: **Multi-Dendrix results on the BRCA dataset restricted to Luminal A patients.**

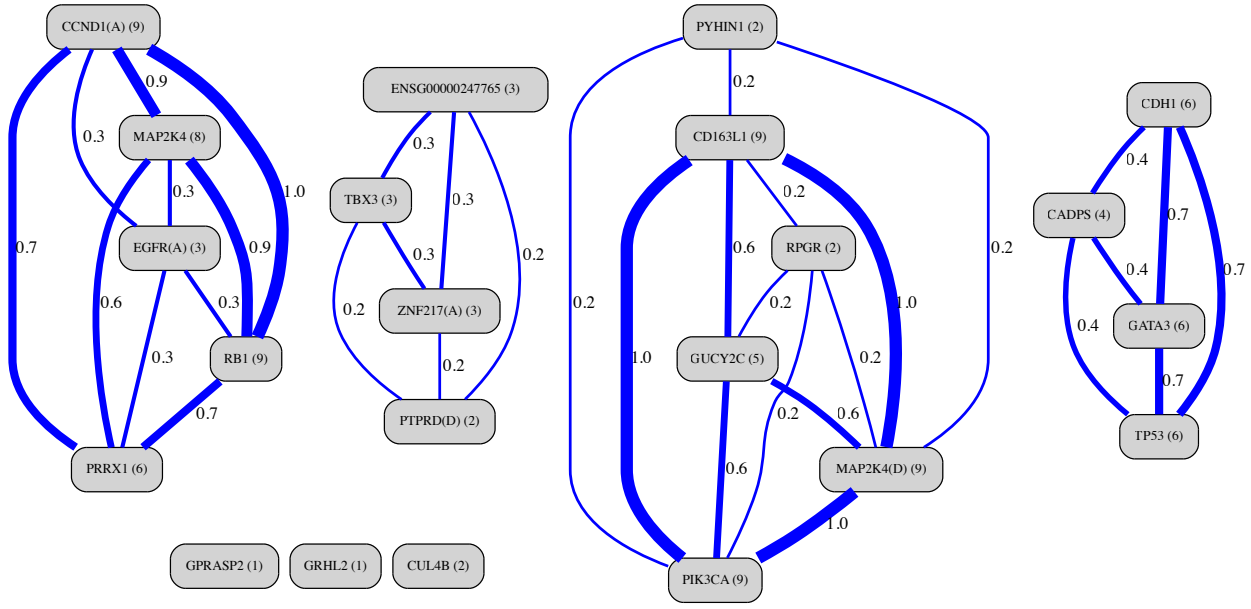

Figure S10: **Multi-Dendrix results on the BRCA dataset restricted to Luminal B patients.**

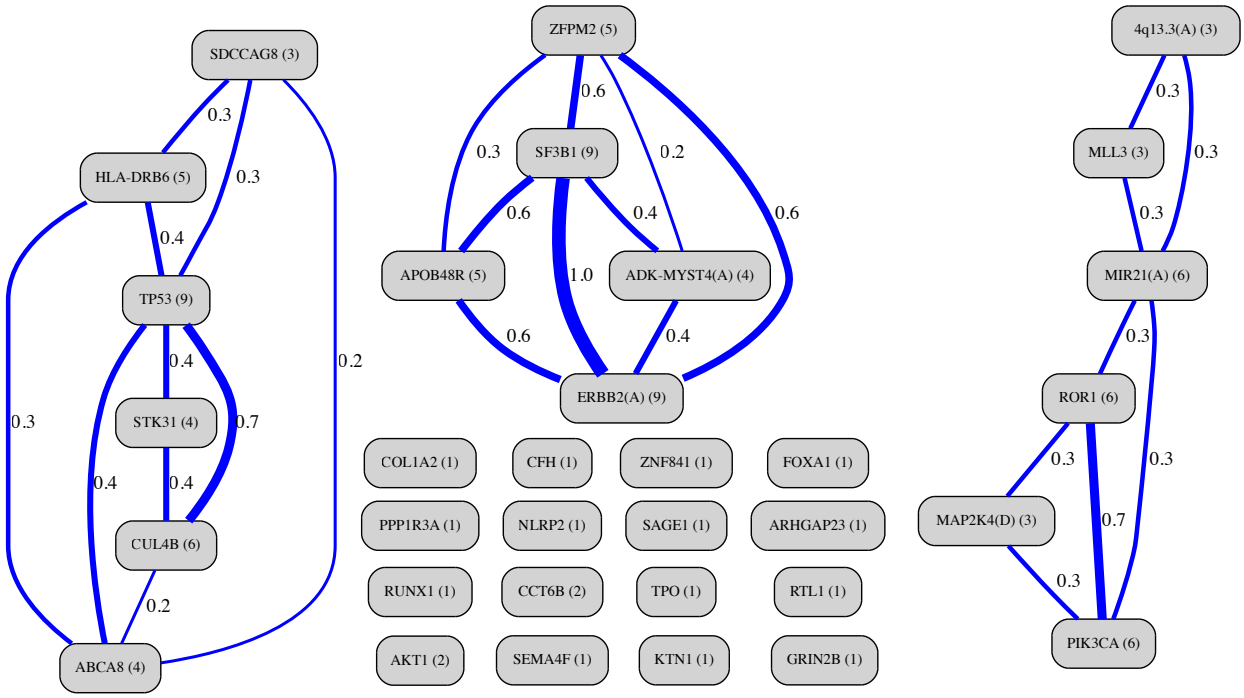

Figure S11: **Multi-Dendrix results on the BRCA dataset restricted to HER2-enriched patients.**

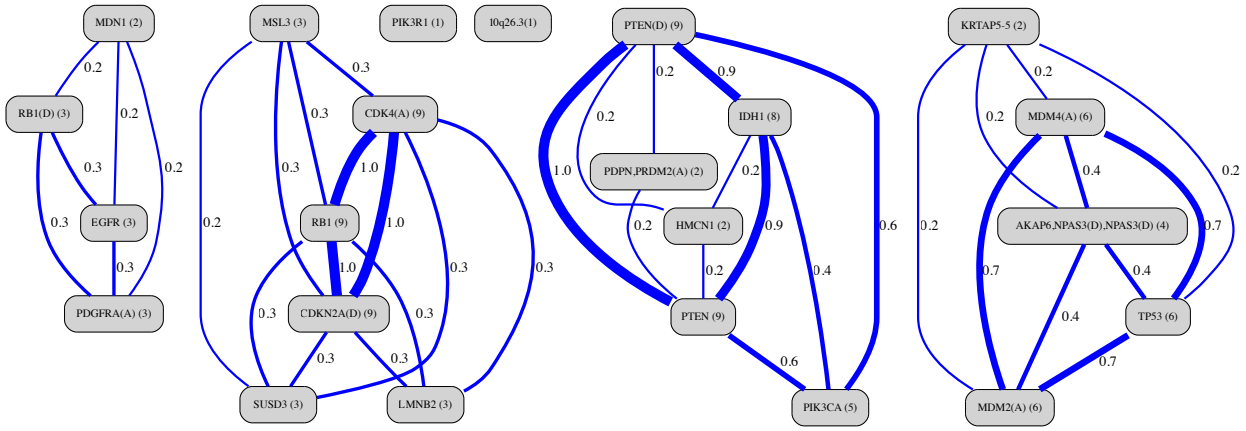

Figure S12: **Multi-Dendrix results on the unfiltered GBM mutation data.**

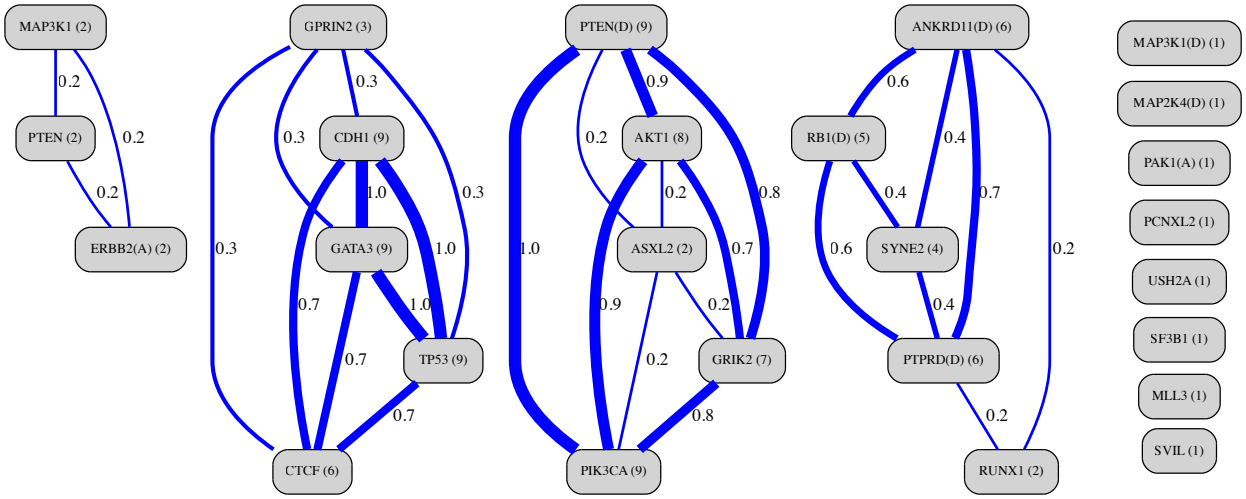

Figure S13: **Multi-Dendrix results on the unfiltered BRCA mutation data.**