Reconstruction of mutational pathway progression in cancer

Abstract

Cancer is a complex evolutionary process accompanied by the accumulation of somatic mutations. With the fast development of next generation sequencing, somatic mutations are measured genome-wide on large cohorts. In order to find the temporal progression order of mutations, several computational efforts have been made. These methods either focused on gene-level progression, or relied on pre-defined pathways to find linear progressions. Pathway-based methods have shown some advantages over gene-based methods, and *de novo* discovery of pathways is affected less by the mutational heterogeneity in cancer patients. How to find complex *de novo* pathway progression remains a challenge. In this study, a new method of identifying multiple *de novo* pathways and their temporal progression is introduced. The algorithm iteratively solves the maximum weight submatrix problem to find pathways, and reconstruct progression order of pathways based on association rule discoveries. The performance of the proposed method is illustrated by simulated data on a number of different parameters. Then two somatic mutation datasets, Brain Lower Grade Glioma (LGG) and Glioblastoma multiforme (GBM), are analyzed by the method. In both datasets the pathway progression orders of statistical significance are identified in concordance with biological knowledge. The method is potentially helpful in understanding cancer biology as well as designing targeted gene therapy.

1. Introduction

The development of cancer is an evolutionary process caused by the accumulation of somatic mutations, including single nucleotide variations, copy number variations and other large genomic structural alterations. These genomic changes are accumulated during the individual's lifetime and serve as the main pushing force to the evolution of cancer cells from normal cells. Somatic mutations that occurred during the development of cancer are believed to belong to two categories, i.e. driver mutations and passenger mutations [1]. The former one is functional responsible for cancer development and helps cancer cells gain growth advantage over normal tissues, while the latter one is considered to be neutral.

With the high accuracy and low costs of next-generation sequencing, several consortium projects have been measuring these somatic mutations from large number of samples and wide range of cancers originated from different tissues, such as The Cancer Genome Atlas (TCGA) [2], International Cancer Genome Consortium (ICGC) [3], and Catalogue of Somatic Mutations in Cancer (COSMIC) [4] etc. Given the whole genome or whole exome sequencing results provided by these projects, the first and foremost problem to solve is to identify the driver mutations from the passenger mutations. To answer this question, many efforts have been devoted to the development and improvement of computational tools. Due to the incompleteness of our knowledge in pathways and mutational heterogeneity in cancer patients, *de novo* discovery of cancer mutation pathways is favored against depending completely on prior knowledge. The piloting work in this field was first proposed by Yeang et al. to detect mutually exclusive mutated gene pairs in multiple cancer types based on simple likelihood ratio test [5]. Then Vandin et al. proposed the maximum weight submatrix problem (for a brief description of the problem, see Section 2.1) and used Markov Chain Monte Carlo to solve it [6]. Zhao et al. further improved the efficiency of the algorithm by employing Integer Linear Programming to find the solution [7]. However, previous studies found pathways of several pre-defined number of genes and took the most frequent subset of them as the output. How to determine the number of genes in the pathway is still a remaining challenge for the *de novo* pathway detection.

Given the inferred driver pathways, a subsequent problem is to identify the temporal order of the occurrence of these driver mutations. Finding the temporal progression of cancer mutations can help improve the understanding of mechanisms of cancer biology as well as the development of targeted treatments. However, this is a much more difficult question to answer comparing to only finding pathways based on coverage and mutual exclusivity. Ideally, the dataset should come from the same patients with multiple time-points; and the inferred driver pathways should be very close to the true conditions in order to detect the correct temporal progression of these pathways. In reality, the dataset consists of many single time-point measurements from multiple different patients, also known as *cross-sectional* data. The detected pathways are also a noisy version of the true mutated pathways depending on the accuracy of the employed methods. Both aspects make the problem difficult to solve.

A number of algorithms for detecting the temporal progression of cancer mutations have been developed [8-10]. Different techniques were utilized by these methods, such as Binary Linear Programming, Expectation-Maximization and Bayesian networks. Most of the previous studies focus on the mutational progression on gene levels. However, due to the mutational heterogeneity of cancer patients, the mutational progression signal is probably spread in different genes belonging to the same pathway for different individuals. Thus it is reasonable to assume that temporal progression of cancer mutations target on pathway levels instead of gene levels.

Some recent studies have also been exploring the temporal progression on pathway levels [11,12]. Cheng et al. focused on the known pathway of genes, thus was restricted by the limited knowledge of pathways. Alternatively, Raphael et al. utilized *de novo* pathway discovery algorithm to find pathways and temporal progression orders simultaneously. However they made a strong assumption that the temporal progression of mutations was linear, which is often not true considering the complexity of tumor progressions.

In this paper, I present a new method to find the tumor mutational progressions based on the cross-sectional data. The method is a two-step algorithm. First, *de novo* pathways are found iteratively till the algorithm cannot find any mutational pathways significantly better than random. Permutation is performed to draw random pathways and an empirical procedure is introduced for selecting proper number of genes in each pathway and testing the convergence of the algorithm. Next, tumor progression of mutations is reconstructed based on association rule discovery of the identified pathways. Simulated data is generated to illustrate the performance of the method. Then, the method is employed to analyze somatic mutation data of Glioblastoma multiforme (GBM) and Brain Lower Grade Glioma (LGG) from TCGA for mining new biological discoveries.

2. Materials and Methods

2.1 Brief introduction to Maximum weight submatrix problem

The algorithm *dendrix* (*de novo* driver exclusivity) [6] was proposed by Vandin et al. to *de novo* find a single driver pathway in cross-sectional mutation data. Given an n by m binary mutation matrix A and an integer k > 0, find the n by k column submatrix M of A that maximizes W(M), where W(M) is the weight function describing the trade-off between coverage and exclusivity. $W(M) = |\tau(M)| - \omega(M) = 2|\tau(M)| - \sum_{g \in M}|\tau(g)|$, where $\tau(g) = \{i: A_{ig} = 1\}$ is the set of patients where gene g is mutated, and $\tau(M) = \cup_{g \in M} \tau(g)$ is the union of all genes mutation status in subset M, i.e. the coverage of subset M, and $\omega(M) = \sum_{g \in M}|\tau(g)| - |\tau(M)|$ denotes the overlapped coverage. To maximize the weight function W(M), the coverage is expected to be high while gene overlapped coverage low, indicating that the genes are mutually exclusive as well as highly covered in all patients.

This model was initially solved by MCMC, and later Zhao et al. [7] proposed an exact method Binary Linear Programming to find the solution more efficiently. Thus, although the maximum weight submatrix is NP-hard, it can often be solved efficiently. However, a key problem with the above model is that it assumed the number of genes in the pathway, denoted by k, is given. In tackling with real situations, finding the proper number of genes in the pathway is essential, especially if one wants to find multiple pathways subsequently.

2.2 An iterative algorithm for detecting multiple pathways

Consider an n by m binary mutation matrix A with K pathways embedded in it, each pathway has L_n genes. Here the pathway denotes a set of genes that are mutually exclusive. For simplicity, we further assume these K pathways followed a linear progression order, i.e. if an individual's K_{i+1} pathway has one gene mutated, then K_i pathway of the individual is also mutated, and an individual's maximum mutated pathway is denoted by I, for$1 \leq i \leq I, K_i$is mutated; for $I < i \leq K, K_i$is not mutated; I is chosen uniformly from 1 to K. Thus the pathway can also be viewed as the stage of tumor progression of a patient. See **Figure 1** for an illustration of this setup. For more complex situations such as paralleled branching progression order, they can be derived by mixtures of the linear models. Note that the situation we consider here is an ideal case without perturbance; in simulated data, noise is added by a background mutation rate.

We can show that by maximizing the weight function W(M) at $k = L_n$, the identified pathway is K_1. Since we have the exclusivity and linear progression assumptions, for any $1 \leq i \leq n$, $\omega(K_i) = 0, |\tau(K_i)| \geq |\tau(K_{i+1})|$, thus $W(K_i) \geq W(K_{i+1}), max_i\{W(K_i)\} = W(K_1) = |\tau(K_1)|$. For any other combinations of genes (columns) to form a pathway K', $|\tau(K')| \leq |\tau(K_1)|, \omega(K') \geq 0 = \omega(K_1)$, thus $W(K_1)$ maxes the objective function.

Hence if we know the correct number of genes in each pathway L_n, we can iteratively find the maximum weight submatrix to be the true pathway. In the real situation, L_n is not known. In order to solve this problem, an empirical procedure was proposed, which is similar to gap statistic analysis in k-means clustering to find the most plausible L_n.

Consider the random situation where no pathway signal is embedded in matrix A, but matrix A has a background mutation rate of $\mathrm{p} = (\mathrm{K} + 1)/2\mathrm{M} < 0.5$, which is the same mutation rate when pathways embedded. For any two genes g_1, g_2, the expected number of mutations is $\mathrm{E}[g_1] = \mathrm{Np} = \frac{N(K+1)}{2M}$, thus $\mathrm{E}[\mathrm{W}(g_1)] = \mathrm{Np}$. By putting a randomly selected gene g_2 into the submatrix, $\mathrm{E}[\mathrm{W}(g_1 g_2)] = 2\mathrm{Np}(1 - \mathrm{p})$, when $\mathrm{p} < 0.5, \mathrm{E}[\mathrm{W}(g_1 g_2)] > \mathrm{E}[\mathrm{W}(g_1)]$. More generally, when $g_1$is a set of genes that does not perfectly satisfy coverage and exclusivity, we can always increase the objective function by adding one more gene to the submatrix. Hence by increasing L_n, the objective function W(M) is monotonic increasing.

Let $\Delta_n = W(M_{n+1}) - W(M_n)$ denote the increased value of weight function from L_n =n to L_n =n+1 on real data, $\Delta'_n = (\sum W(M'_{n+1}) - \sum \mathrm{W}(M'_n))/N$ denote the average increased value from N permutated data, and $d_n = \Delta_n - \Delta'_n$ denote the difference of objective function delta values between real and permutated data. Ideally, to find the most plausible L_n in a given range, the algorithm should terminate when increasing L_n, the increased value of objective function is not significantly larger than picking a gene from the permutated matrix, i.e. $d_n = 0$. Empirically it is observed that the objective function increased value in real data decreases to be close to the mean of that from permutated data, but is always above it by some constant number, probably due to the signals from other pathways or internal noise in real data. Hence we find the number of genes in this pathway by $\widehat{L_n} = max_n(\forall i < n, d_i > d_n)$.

The algorithm for iteratively detecting multiple driver pathways is summarized as below.

(1) Given a mutation matrix A and a range $[a, b]$, find the maximum weight submatrix M_n and corresponding its objective function value $\mathrm{W}(M_n)$ for each $L_n \in [a, b]$.
(2) For each L_n, generate N random mutation matrix with background mutation rate $\mathrm{p} = \frac{\mathrm{K+1}}{2M}$, find the maximum weight submatrix M'_n and the average objective function value $\mathrm{W}(M'_n)$.
(3) Find the number of genes in the pathway by $\widehat{L_n} = max_n(\forall i < n, d_i > d_n)$, where $d_n = \Delta_n - \Delta'_n$, $\Delta_n = W(M_{n+1}) - W(M_n)$, and $\Delta'_n = (\sum W(M'_{n+1}) - \sum \mathrm{W}(M'_n))/N$.
(4) Extract M_{Ln} submatrix out of A.
(5) Repeat (1) until $\widehat{L_n}$<2.

2.3 Temporal progression reconstruction based on association rules

Association rule discovery is applied to the pathway detected from the above procedure, in order to determine the temporal order of the pathway mutations. Association rule clustering is first developed in the field of economics, where the goal is to find which sets of items shoppers tend to buy together. For example, two items A and B are bought together 200 times out of 1000 total transaction records. Of these, item C is also bought 100 times along with A and B. Thus a rule 'If A and B, then C' is a rule to be discovered.

An association rule is characterized by two metrics, support and confidence. Support is the number or the proportion of occurrences that obeys the rule. Confidence is the proportion of supported occurrences with respect to the total number of occurrences of the 'if' part. Taking the above example, the support of the rule 'If A and B, then C' is 100/1000=0.1, while the confidence is 100/200=0.5.

In the cancer pathway progression scenario, the pathway of an individual is considered to be mutated if any of the genes in that pathway is mutated. Then the pathway mutation status can be viewed as the shopping items while each individual is a transaction. Two assumptions are made to facilitate the reconstruction of pathway progression: each downstream pathway only has one upstream pathway while an upstream pathway can have multiple downstream pathways; a pathway with more patients mutated comes upstream of pathways with less patients mutated. Confidence information is further used to infer the direct or indirect causal relationships. If pathway A is upstream of pathway B and B is upstream of C, then rules 'If A then B', 'If B then C' as well as 'If A then C' should be observed. Since confidence describes the proportion of supported occurrences with respect to the 'if' part, indirect rules like 'If A then C' are expected to have lower confidence than direct ones. Using the setup in Section 2.2, we have $confidence(K_i \rightarrow K_{i+1}) = \frac{K-1}{K} > confidence(K_i \rightarrow K_{i+2}) = \frac{K-2}{K}$. Hence the upstream of a pathway is defined by the left hand side pathway of the rule with highest confidence and the target pathway on the right hand side.

The procedure of reconstructing the progression order of pathways is summarized as below.

(1) Initialize an empty network *net*.
(2) Given l detected pathways, order the pathways by the total number of patients mutated, $P_{(1)}, P_{(2),} \cdots, P_{(l)}$. $P_{(1)}$ is connected to start and push them into net.
(3) Make the n by l pathway mutation matrix B. For each individual, the binary mutation status of a pathway is 1 if any of the genes in that pathway is mutated, otherwise 0.
(4) Apply association rule discovery to B.
(5) For i in 2 to l, find the rules with right hand side equal to $P_{(i)}$ and left hand side in net. Connect $P_{(i)}$ to the pathway on left hand side with highest confidence. If no rules are available, swap $P_{(i)}$ and $P_{(i+1)}$.
(6) Connect all pathways or stop if maximum iterations are reached.

2.4 Simulated data

An n by m mutation matrix was simulated with various parameters to illustrate the performance of the proposed method as previous study [12]. Different number of samples (n=100, 200, 300) and genes (m=500, 1000, 1500) were considered for simulation. In each matrix, K=5 or 10 stages were embedded in the matrix, each with L_n =5 or 10 genes. For each individual, the progression stage was chosen uniformly from 1 to K, and the gene to be mutated in the pathway was chosen uniformly. Then the background mutation was added by probability q=0.001, 0.005, 0.01 to randomly flip any entry in the matrix. These mutation rates were in the range of mutation rates of passenger mutations observed in real data. For each set of parameters, ten independent trials were performed in order to evaluate the accuracy repeatedly.

2.5 Biological data

Glioblastoma multiforme (GBM) and Brain Lower Grade Glioma (LGG) somatic mutation datasets from TCGA were analyzed by the proposed method. The level 2 DNA-Seq data was first converted to a binary mutation matrix. The entry A_{ij} was 1 if the gene j was mutated in the individual i, otherwise 0. A gene was considered to be mutated if there were any alterations recorded in that gene, regardless of the certain mutation types (single nucleotide mutation, insertion or deletion, etc.). A bootstrap procedure of resampling individual with replacement was performed 30 times, and results by applying the same algorithm were recorded to analyze the robustness of the identified temporal progression order.

3. Results

Simulated data

3.1 Multiply pathway identification

A number of simulations were generated to evaluate the performance of the proposed method. We started by 100 samples and 500 genes with 5 pathways, each had 10 genes embedded. Each individual was chosen uniformly from 5 or 10 stages. In this set of parameters, the expected mutation rates for the five pathways were 1, 0.8, 0.6, 0.4 and 0.2 for 5 stages, and 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2 and 0.1 for 10 stages, respectively. Then noise was added through background mutation rate of 0.001 and 0.01 to reflect different noise level. Since each pathway had 10 genes, the mutation rate for a single gene in the pathways was decreased by 10 folds comparing to the pathway level, hence a noise level of 0.01 was almost comparable to signal strength, and they had an expected error number larger than 1 for each individual.

For each set of parameters, ten independently repeated trials were performed to estimate the accuracy and its confidence interval. The accuracy was measured by sensitivity and precision for each pathway, where $precision = \frac{TP}{TP+FP}$ and $sensitivity = \frac{TP}{TP+FN}$. These two metrics measured both how correctly the algorithm could find the pathway genes, as well as how many genes should the algorithm include in the pathway. To determine the number of genes in each pathway, L_n from 2 to 12 was tested as described in Section 2.2. The mean and standard deviation of the sensitivity and precision for the repeated trials were reported.

As shown in **Figure 2**, both sensitivity and precision decreased as the coverage of the pathway went down. Also the standard errors were also increasing as the accuracy decreased. Interestingly the precision was always above the sensitivity in both setups, indicating that the proposed method was conservative, which was a desirable property in dealing with noisy clinical data. Specifically, when there were 5 pathways embedded, the precision and sensitivity were both above 0.75 for the first three stages with gene mutation rate 0.1, 0.08, 0.06, respectively. When gene mutation rate further went down to 0.04 and 0.02, accuracy dropped rapidly while the largest variable precision was observed when gene mutation rate 0.02. The overall trend of the accuracy was the same when 10 pathways were embedded. However, note that the accuracy of pathways with mutation rate 0.08 and 0.06 was lower in 10 stages than that in 5 stages, indicating signals were weaker when more stages were present in the data, probably due to the complex inter-stage associations.

Next the effect of more non-stage genes to the proposed method was analyzed. Fixing the number of samples 100 and background mutation rate 0.001, the number of genes increased to 1000 and 1500. Thus the expected error number for each individual was increased by 2- and 3-folds, respectively. The decrease in accuracy was little, if any, when number of genes increased for both 5 stages and 10 stages (**Table 1, 2**). Hence when fixed background mutation rate is low, adding more non-progression genes does not decrease the accuracy significantly.

When increasing the background mutation rate, it was expected that the accuracy would drop, especially for pathways with low gene mutation rate. In this setup, the background mutation rate was increased to 0.01 to analyze the mutually offset effects with increased background noise and increased sample size. Note that the noise level 0.01 was the same to the last progression stage, thus the last progression stage was impossible to be detected. As shown in **Table 1**, when there were only 100 samples and 500 genes, the overall accuracy was low. When sample size increased to 200, progression stages with mutation rate larger or equal to 0.06 were identified correctly, which was consistent with the simulation results of 100 samples and 0.001 background noise level. When adding non-progression genes to 1000, the accuracy decreased. Hence when background noise level is high, adding more non-progression genes could potentially reduce the performance of the algorithm.

3.2 Reconstruction of pathway temporal progression

Having obtained the inferred pathways, the next step was to reconstruct the progression order of the pathways. The performance for reconstructing the temporal progression order was first illustrated by the ideal case, where sample size was 200, number of genes was 500 and background noise 0.001 with 10 embedded stages. In this set of parameters, most of the pathways were identified correctly. The accuracy was measured as the proportion of correctly detected progression orders with respect to all true progression orders (sensitivity) and all detected results (precision).

The proposed method performed with high accuracy in the ideal case. In the first step, the identified pathways had good matching with the true conditions (**Table 1**) except for the last stage. Then the sensitivity and precision were computed for the second step, finding the progression orders. Sensitivity had mean 0.87 with standard deviation 0.067; precision had mean 0.825 with standard deviation 0.092. The overall accuracy was satisfying, given the last stage was almost never identified correctly. If the last stage was disregarded, sensitivity achieved 0.967 on average, indicating that the proposed method was capable of finding the correct progression order when the pathways were identified with high accuracy in the first step.

Then let us move on to more realistic conditions. As illustrated in Section 3.1, pathways with lower gene mutation rate were not identified with high accuracy, so it was not rational to evaluate the progression order for those pathways. The threshold was set as precision>0.5 and sensitivity>0.5 so that pathways below the threshold were not considered. Thus when the number of genes at 500 and 10 stages embedded, for background noise level 0.001, we consider the first four pathways at sample size 100; for background noise level 0.01, we consider the first six pathways at sample size 200. The difference in sample size was to compensate for different noise level, as shown in Section 3.1. For the first setup, the mean sensitivity was 0.806 with standard deviation 0.242. For the second setup, the mean sensitivity was 0.883 with standard deviation 0.137. The pooled confidences were further compared between the true positive and false positive progression orders detected (**Figure 3**). Unsurprisingly, the confidence of true positives was significantly larger than that of the false positives (Wilcoxon one-sided test, p-value=0.0014). Hence, the proposed method could find the progression order with decent sensitivity under the simulated scenario similar to real conditions, and the significant difference between true positive and false positive progressions detected could potentially be used as the signal to filter the incorrectly identified results.

Biological data

3.3 Brain Lower Grade Glioma (LGG)

As an application to generate more biological findings, the proposed method was first applied to Brain Lower Grade Glioma (LGG). Somatic mutation data were downloaded from TCGA database. The dataset contained 286 samples with 5961 genes. These genes were mutated at least in one sample, and IDH1 was mutated most frequently in 221 (77%) patients. The mean and median of the number of mutated genes in all patients were 34.2 and 28 respectively.

To determine the number of genes in each pathways, the range 2 to 12 was tested in each iteration. Then bootstrap was performed to test the robustness of the detected pathways. If the detected pathways were robust, then the genes in a given pathway should co-occur/overlap significantly more in the bootstrap results than random, hence the statistical significance given a pathway could be derived by permuting the occurrences for each gene and compare the union of all gene occurrences to the observed number in bootstrap.

The algorithm detected altogether 5 pathways, with 4 pathways significant (p-value<0.01) and subsequently connected to each other. The fifth pathway detected was not significant (p-value=0.089) and not connected to the rest pathways. The four significant pathways were IDH1, TP53, Unknown; ATRX, CIC, COL6A3, EGFR, NF1, NOTCH1, PIK3CA, TTN; APOB, ARID1A, BCOR, FUBP1, HMCN1, IDH2, MUC16, PTEN, RYR2, SMARCA4 and C3, DNAH5, FAT2, FLG, MUC17, MYH2, NEB, PIK3R1, PKHD1, ZBTB20 (**Table 3**). Among these, a number of oncogenes could be spotted, such as TP53, PTEN, EGFR, etc. Notably, IDH1 and IDH2, previously reported as the survival marker for LGG [13], were also included in the pathways.

The mutual exclusivity could be clearly observed in **Figure 4**, and the coverage was decreasing as the stage progressing. The first pathway contained only three genes, IDH1, TP53 and Unknown. The Unknown genes comprised a number of genomic loci that had no known gene annotation. These genes showed good exclusivity with TP53, indicating the potential role in tumorigenesis. The coverages of the pathways were 0.954, 0.808, 0.444, 0.315 and 0.182, respectively. The detected progression order was a linear order from root to pathway 1, 2, 3, and 4. The confidence for these progression orders was 0.954, 0.831, 0.459 and 0.346.

In order to understand the biological meaning of the pathways, Gene Ontology (GO) enrichment analysis for pathways was performed using DAVID [14], and the GO terms from the top two clusters detected with FDR<0.1 were extracted. The first pathway was not considered since it only contained two known genes and served as the source for tumorigenesis. Interestingly, a specific developmental trace was unveiled (**Figure 5**). The second pathway was enriched by GO terms related to the growth of cells, rendering a growth advantage to tumor cells over normal tissues. Moving on to the third pathway, the enriched GO terms were related to ion binding and blood vessel development, although the significance was slightly below the threshold (FDR>0.1). The enrichment signal in the last pathway was further weaker, but consisted of terms with cytoskeleton, indicating their potential role in tumor metastasis. This clearly depicted a progression trace of mutation origination, cell proliferation, tumor tissue formation and tumor metastasis. The progression order as well as the genes in each pathway will be of great potential interests to the targeted medicines.

3.4 Glioblastoma multiforme (GBM)

Next, the somatic mutation dataset of Glioblastoma multiforme (GBM) was download from TCGA database. In this dataset, there were 282 individuals and 9507 genes. The mean and median number of mutated genes for the patients were 74.8 and 72. The gene with maximum number of patients mutated was TTN (88, 31.2%).

From the above brief summary statistics, it could be expected that the pathways and the corresponding progression were possibly more complicated than those in LGG, since more genes were involved and the average mutated genes elevated. Indeed, thirteen pathways were identified by the proposed method in GBM dataset. Among them four pathways did not pass the statistical significance test. The detailed pathway genes, pathway coverage and statistical significance were listed in **Table 4**.

After pruning the statistically insignificant pathways, the rest of the pathways were utilized to reconstruct the progression orders. The resulting progression of these pathways exhibited some paralleled branching features (**Figure 6**). As the same illustrated in Section 3.3, the genes in each pathway were analyzed by GO enrichment. The enrichment analysis results did not have a clear pattern, and several downstream pathways had no significantly enriched terms (FDR>0.1). However, there was an interesting pattern between the two branching progressions when pooling the genes of the same branch together. In both branches, the direct downstream of the origination was enriched with Epidermal Growth Factor (EGF)-like regions (FDR<0.1). After that, one branch was mainly enriched by calcium ion binding and other metal ion binding as well cell adhesion. On the other hand, another branch with only two stages was related to extracellular matrix and cell projection. These two braches concerned about cell adhesion and cell motility respectively, which both led to tumor metastasis eventually. The proposed method showed two different pathological pathways for the GBM development. It was not clear if the branching was due to the many subtypes of GBM patients, and it could be of future interests to investigate the relationships of tumor subtypes and mutational pathway progressions.

4. Conclusions & Discussions

In this study a new two-step method was proposed to find the pathway progression order from cross-sectional data. The first step was to find multiple pathways from the binary mutation matrix, and the second step was to infer progression orders of the identified pathways based on association rules. The proposed method worked with high accuracy on simulated data with a number of different parameters. Then it was applied to two real datasets LGG and GBM. The findings in concordance with biological knowledge indicated that the method could be of potentially great interests to understanding cancer biology and applying targeted gene therapies in the future.

The algorithm still has several future extensions that are not addressed in this work due to the time limit. First, for the empirical procedure used to determine the number of genes in each pathway, it will be more solid if a theoretical proof could be given to bound the performance as well as the algorithm will terminate. In this work, this problem was addressed through extensive simulations and the performance was good overall. Second, the true and false progression orders detected showed a significant difference in confidence as illustrated in Section 3.2, thus a Gaussian mixture model could be fitted into the pooled confidences and further prune false positive pathway connections. Third, all types of mutations (Insertions or deletions, frameshift mutation, synonymous and non-synonymous, etc.) are all considered equally for gene mutation, which apparently have different functional severity. It might improve the algorithm performance if a scoring scheme is available. Fourth, the algorithm can potentially benefit from adding proper prior knowledge about genes and/or patients. For example, removal of hyper-mutated genes or individuals as outliers could potentially enhance the robustness of the algorithm.

With the fast pace of sequencing technologies development, more and more whole-genome, transcriptome and epigenome sequencing results will be available, making it possible to analyze the disease stage and omics alteration on large scale. The idea of the proposed method can be potentially extended to other types of data, and it will be of great interest to generate and compare the temporal progressions for different types of cancers on inter- and intra- platform datasets.

References

1. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463: 191-196.
2. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics 45: 1113-1120.

3. Zhang J, Baran J, Cros A, Guberman JM, Haider S, et al. (2011) International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database : the journal of biological databases and curation 2011: bar026.

4. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, et al. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research 43: D805-811.

5. Yeang CH, McCormick F, Levine A (2008) Combinatorial patterns of somatic gene mutations in cancer. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 22: 2605-2622.

6. Vandin F, Upfal E, Raphael BJ (2012) De novo discovery of mutated driver pathways in cancer. Genome research 22: 375-385.

7. Zhao J, Zhang S, Wu LY, Zhang XS (2012) Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 28: 2940-2947.

8. Gerstung M, Baudis M, Moch H, Beerenwinkel N (2009) Quantifying cancer progression with conjunctive Bayesian networks. Bioinformatics 25: 2809-2815.

9. 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.

10. Tofigh A, Sj̦lund E, Ḥ̦glund M, Lagergren J. A global structural EM algorithm for a model of cancer progression; 2011. pp. 163-171.

11. Cheng YK, Beroukhim R, Levine RL, Mellinghoff IK, Holland EC, et al. (2012) A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. PLoS computational biology 8: e1002337.

12. Raphael BJ, Vandin F (2015) Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. Journal of computational biology : a journal of computational molecular cell biology 22: 510-527.

13. Houillier C, Wang X, Kaloshi G, Mokhtari K, Guillevin R, et al. (2010) IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. Neurology 75: 1560-1566.

14. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4: 44-57.

Tables and Figures

Figure 1. Illustration of cancer mutational pathway and its stage progression. In the ideal case a), the pathways rigorously follow the mutual exclusivity and coverage. If an individual has a gene mutated in a later stage, e.g. pathway 3, then genes in former stages (pathway 2 and 1) must be mutated for progression. In simulated condition b), random noise is added through flipping the entries in mutation matrix, thus possibly disrupts the mutual exclusivity, coverage and/or progression assumptions.

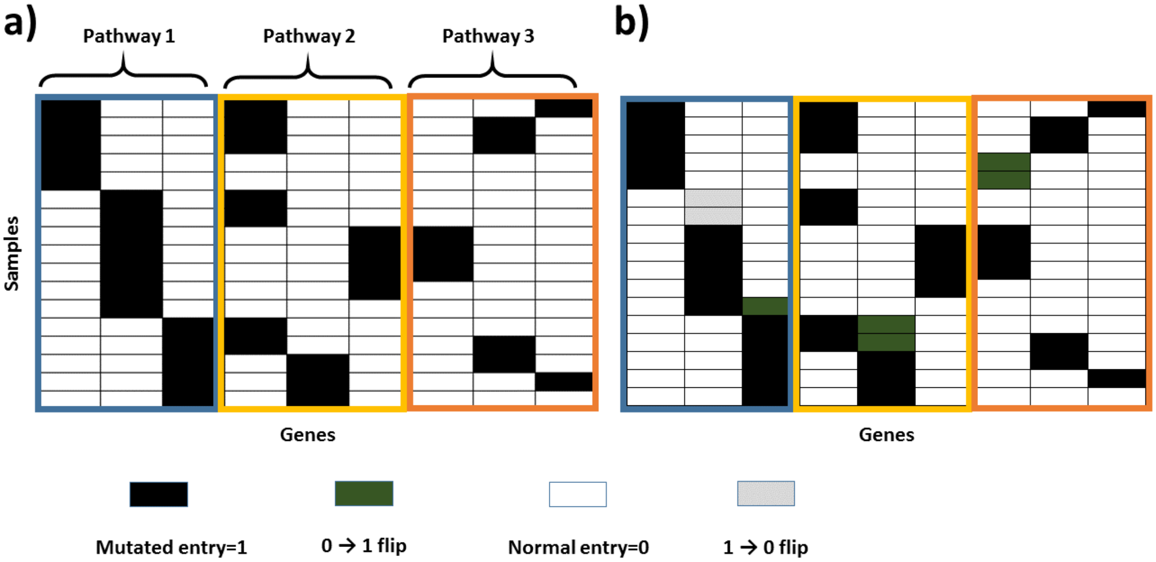

Figure 2. Precision and sensitivity line plot under N=100, M=500, q=0.001 and K=5 (left) and 10 (right). Error bars are derived by independent repeated trials 10 times.

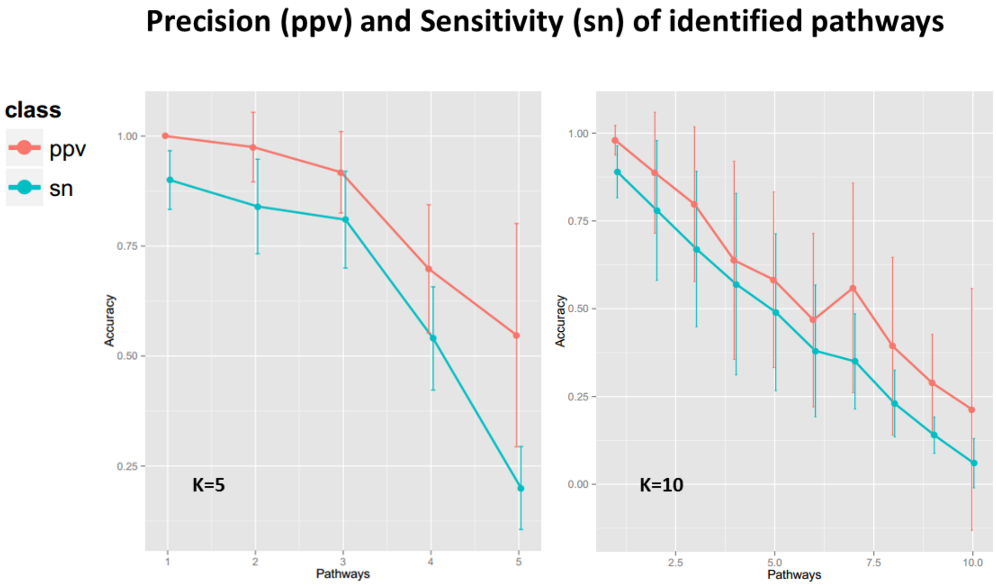

Figure 3. Confidences of true and false identified progression orders. a) Boxplot of confidence of true and false orders. Notch is 95% confidence interval for sample median. b) Histogram of pooled true and false orders confidence. Two Gaussian densities are fitted.

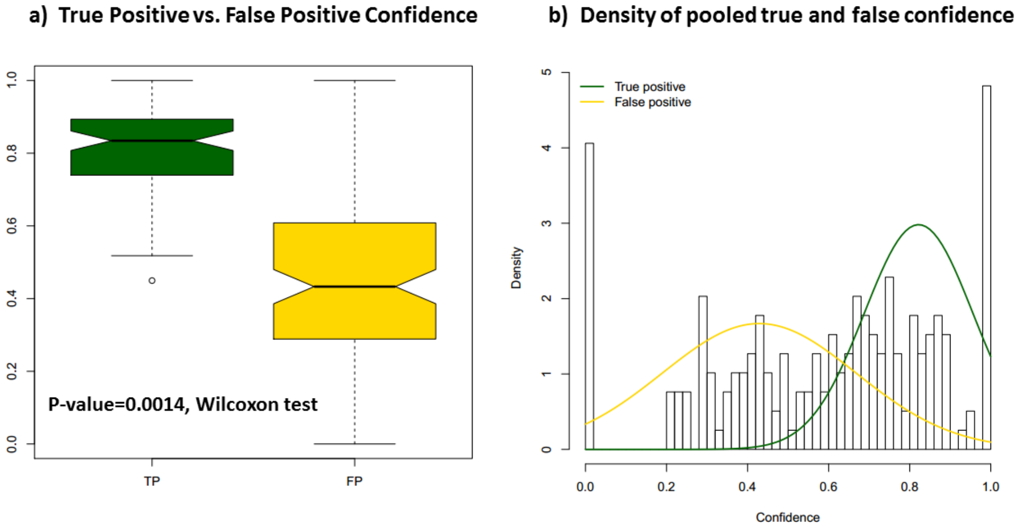

Figure 4. Gene mutation status for each pathway. From top-left to bottom-right are pathway 1 to pathway 5. Black chunks are mutated genes. Sample orders are altered for more clear coverage and mutual exclusivity illustration.

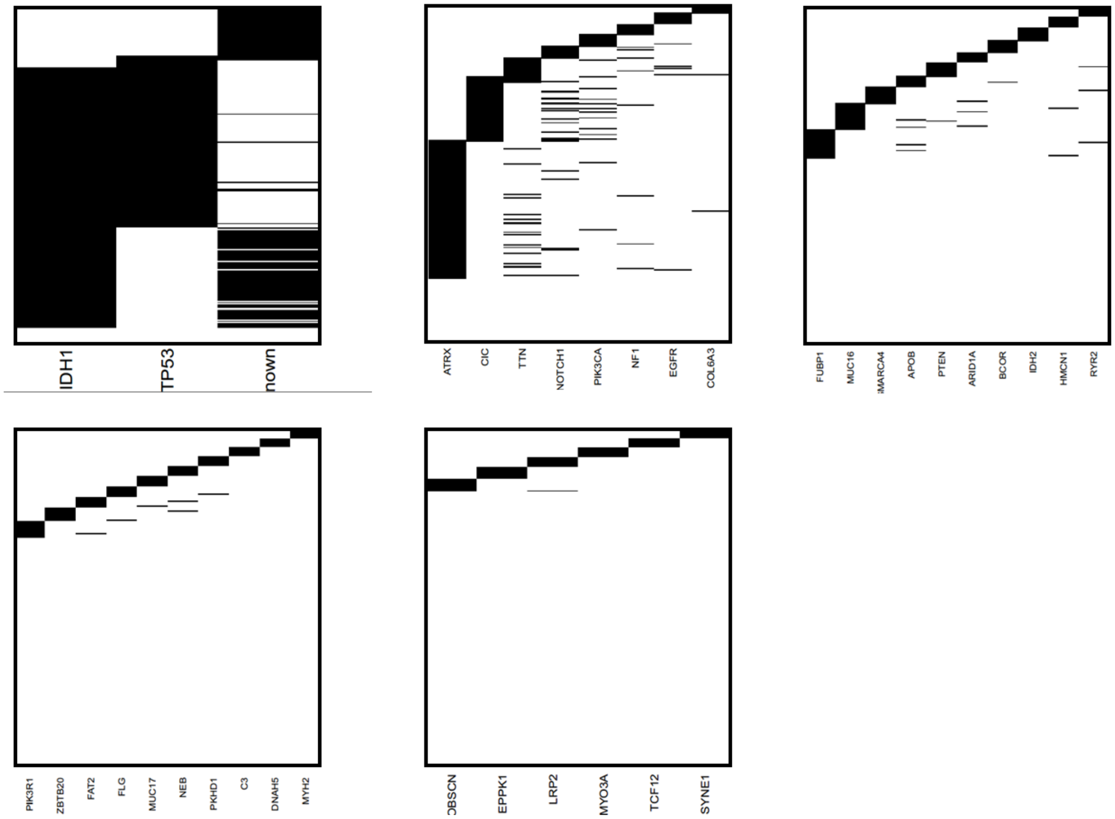

Figure 5. Results for LGG. a) Enriched GO terms for each pathway, negative log10 of FDR. Dashed line represents FDR=0.1. b) Reconstructed progression order for LGG tumor development.

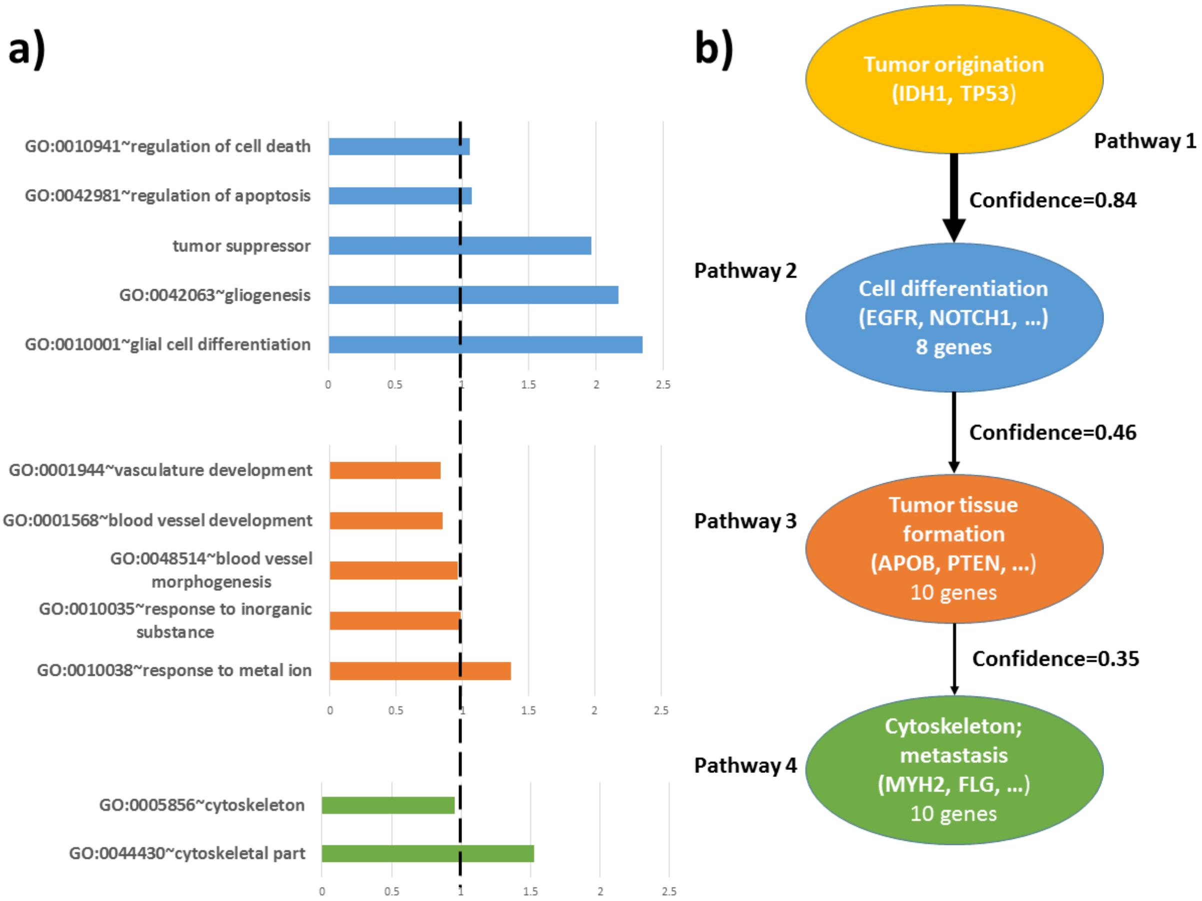

Figure 6. Results for GBM. a) Enriched terms for each pathway, negative log10 of FDR. Dashed line represents FDR=0.1. Enriched terms are only shown for FDR<0.01 for the pooled pathway 2 and pathway 4 to 8, for better plotting and illustration. b) Reconstructed progression order for LGG tumor development. Pathway 4, 5, 6, 7, 8 are pooled together as the downstream of pathway 2.

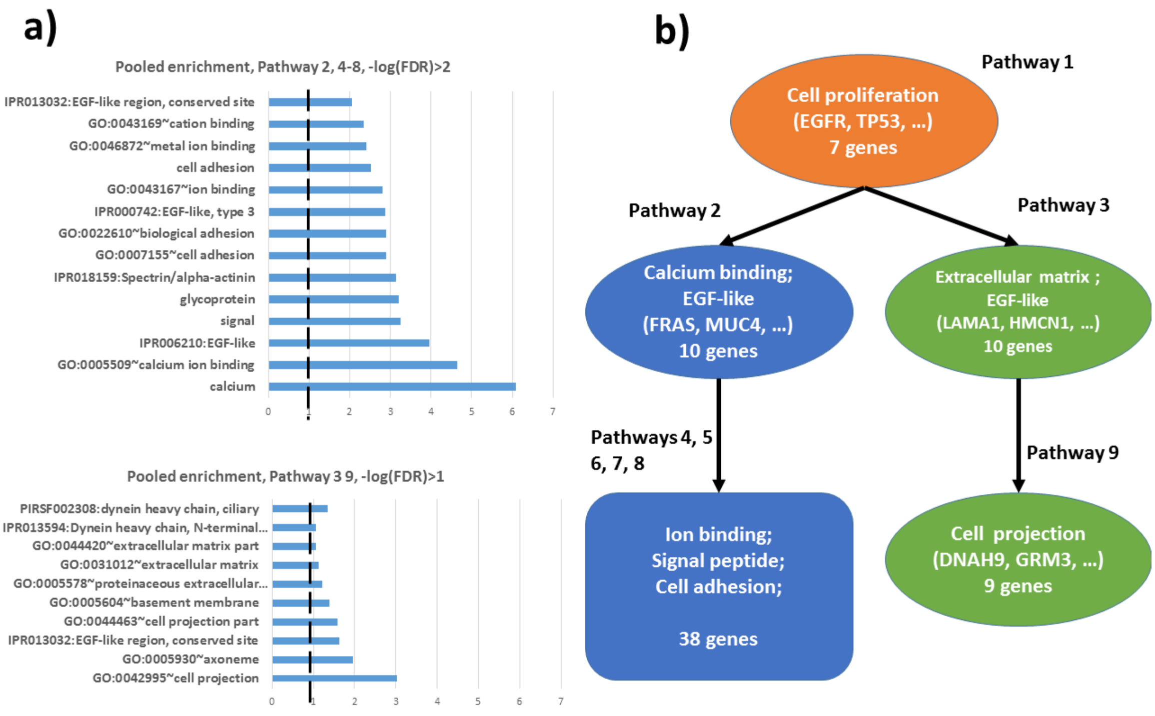

Table 1. Evaluation of five simulated pathway embedded (mean, standard error)

N=100, Q=0.001	PATHWAY	1	2	3	4	5
SENSITIVITY	M=500	0.9;0.067	0.84;0.107	0.81;0.11	0.54;0.117	0.2;0.094
	M=1000	0.92;0.063	0.92;0.063	0.75;0.097	0.53;0.211	0.16;0.126
	M=1500	0.91;0.057	0.88;0.103	0.67;0.134	0.36;0.151	0.13;0.116
PRECISION	M=500	1;0	0.975;0.079	0.918;0.093	0.698;0.146	0.548;0.254
	M=1000	1;0	0.97;0.067	0.916;0.105	0.733;0.225	0.515;0.391
	M=1500	1;0	0.95;0.085	0.816;0.127	0.517;0.21	0.243;0.292

Table 2a. Sensitivity of ten simulated pathway embedded (mean, standard error)

	SENSITIVITY						
	q=0.001				q=0.01		
PATHWAY	N=100, M=500	N=100, M=1000	N=100, M=1500	N=200, M=500	N=100, M=500	N=200, M=500	N=200, M=1000
1	0.89;0.074	0.79;0.179	0.63;0.442	0.94;0.052	0.32;0.23	0.95;0.053	0.89;0.12
2	0.78;0.199	0.73;0.2	0.57;0.406	0.93;0.048	0.27;0.157	0.91;0.099	0.78;0.132
3	0.67;0.221	0.54;0.207	0.55;0.392	1;0	0.17;0.195	0.88;0.092	0.67;0.183
4	0.57;0.258	0.5;0.24	0.46;0.353	0.95;0.071	0.09;0.16	0.77;0.157	0.5;0.156
5	0.49;0.223	0.4;0.205	0.36;0.32	0.94;0.07	0.06;0.126	0.75;0.201	0.27;0.116
6	0.38;0.187	0.37;0.125	0.32;0.239	0.9;0.094	0.06;0.052	0.6;0.189	0.16;0.126
7	0.35;0.135	0.27;0.116	0.22;0.181	0.84;0.143	0.02;0.042	0.42;0.23	0.06;0.084
8	0.23;0.095	0.25;0.071	0.16;0.117	0.77;0.134	0;0	0.22;0.114	0.05;0.053
9	0.14;0.052	0.09;0.057	0.06;0.084	0.54;0.117	0;0	0.08;0.092	0.01;0.032
10	0.06;0.07	0.02;0.042	0.01;0.032	0.19;0.088	0;0	0.02;0.042	0;0

Table 2b. Precision of ten simulated pathway embedded (mean, standard error)

	PRECISION						
	q=0.001				q=0.01		
PATHWAY	N=100, M=500	N=100, M=1000	N=100, M=1500	N=200, M=500	N=100, M=500	N=200, M=500	N=200, M=1000
1	0.98;0.042	0.901;0.177	0.67;0.464	1;0	0.469;0.351	1;0	0.99;0.032
2	0.887;0.172	0.836;0.198	0.635;0.444	1;0	0.57;0.309	0.99;0.032	0.916;0.134
3	0.798;0.22	0.699;0.278	0.655;0.458	1;0	0.4;0.394	0.978;0.047	0.876;0.163
4	0.638;0.282	0.603;0.274	0.528;0.422	0.979;0.045	0.146;0.261	0.938;0.111	0.742;0.298
5	0.582;0.25	0.45;0.202	0.413;0.333	0.95;0.071	0.088;0.208	0.907;0.144	0.632;0.356
6	0.468;0.247	0.528;0.207	0.446;0.334	0.96;0.052	0.23;0.316	0.689;0.242	0.238;0.158

7	0.559;0.298	0.364;0.165	0.347;0.285	0.919;0.089	0.045;0.096	0.503;0.213	0.062;0.086
8	0.393;0.253	0.42;0.155	0.292;0.242	0.86;0.137	0;0	0.378;0.241	0.221;0.339
9	0.288;0.138	0.154;0.139	0.159;0.313	0.643;0.158	0;0	0.217;0.319	0.01;0.032
10	0.213;0.344	0.133;0.322	0.01;0.032	0.397;0.278	0;0	0.053;0.117	0;0

Table 3. Pathways, progressions and statistics for LGG

Pathway	Genes	Coverage	Pathway p-value	Upstream	Confidence
1	IDH1,TP53,Unknown	0.954	0	root	0.954
2	ATRX,CIC,COL6A3,EGFR,NF1,NOTCH1,PIK3CA,TTN	0.808	0	1	0.831
3	APOB,ARID1A,BCOR,FUBP1,HMCN1,IDH2,MUC16,PTEN,RYR2,SMARCA4	0.444	0	2	0.458
4	C3,DNAH5,FAT2,FLG,MUC17,MYH2,NEB,PIK3R1,PKHD1,ZBTB20	0.315	0	3	0.346
NS	EPPK1,LRP2,MYO3A,OBSCN,SYNE1,TCF12	0.182	0.0888	NA	NA

Table 4. Pathways, progressions and statistics for GBM

Pathway	Genes	Coverage	Pathway p-value	Upstream	Confidence
1	EGFR,FLG,MUC16,PCDHAC2,PTEN,TP53,TTN	0.876	0	root	0.875
2	FRAS1,MUC17,MUC4,NBPF10,PCDHGC5,PCLO,PIK3R1,PKHD1,RYR2,SPTA1	0.719	0	1	0.728
3	FRG1B,HMCN1,HRNR,LAMA1,MST1P9,NF1,OBSCN,PIK3CA,RB1,USH2A	0.609	0	1	0.631
4	APOB,ATRX,COL6A3,DNAH5,GPR98,KEL,TUBBP5,Unknown	0.457	0.0001	5	0.582
5	CNTNAP2,FCGBP,HSD17B7P2,IDH1,NLRP5,RELN,RYR3,SYNE1,TCHH	0.475	0	2	0.497
NS	DNAH2,DNAH3,DNAH8,DOCK5,FLG2,LRP2,MUC5B	0.354	0.2816	NA	NA
6	DSP,FAT2,GRIN2A,HCN1,HEATR7B2,MLL3,WASH3P	0.329	0.0166	4	0.411
7	GABRA6,KRTAP4-11,MACF1,MXRA5,SDK1,SEMG2,SLIT3,TSHZ2	0.343	0.0004	4	0.372
8	BC101079,COL1A2,FGD5,RIMS2,SDHAP2,ZNF814	0.262	0.0002	7	0.309

NS	MLL2,PDGFRA,UGT1A1	0.138	1	NA	NA
NS	PIK3CG,PRDM9,SCN9A	0.134	0.123	NA	NA
9	DNAH11,DNAH9,FBN2,GRM3,POM121L12,RBM47,RYR1,SPAG17,TMEM132D	0.347	0.0326	3	0.389
NS	AHNAK2,DCHS2,LZTR1,MYH2,PCDH11X,POTEC,SLCO6A1,TRPV6,TRRAP,UGT2B10	0.411	0.6993	NA	NA