

## RESEARCH ARTICLE

# Identifying patient-specific flow of signal transduction perturbed by multiple single-nucleotide alterations

Olha Kholod<sup>1,2</sup>, Chi-Ren Shyu<sup>1,5,\*</sup>, Jonathan Mitchem<sup>1,3,4</sup>, Jussuf Kaifi<sup>3</sup>, Dmitriy Shin<sup>1,2</sup>

<sup>1</sup> MU Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65201, USA

<sup>2</sup> Department Pathology and Anatomical Sciences, School of Medicine, University of Missouri, Columbia, MO 65212, USA

<sup>3</sup> Department of Surgery, School of Medicine, University of Missouri, Columbia, MO 65212, USA

<sup>4</sup> Harry S. Truman Memorial Veterans' Hospital, Columbia, MO 65201, USA

<sup>5</sup> Department of Electrical Engineering & Computer Science, University of Missouri, Columbia, MO 65212, USA

\* Correspondence: shyuc@missouri.edu

Received July 17, 2020; Revised September 3, 2020; Accepted September 3, 2020

**Background:** Identifying patient-specific flow of signal transduction perturbed by multiple single-nucleotide alterations is critical for improving patient outcomes in cancer cases. However, accurate estimation of mutational effects at the pathway level for such patients remains an open problem. While probabilistic pathway topology methods are gaining interest among the scientific community, the overwhelming majority do not account for network perturbation effects from multiple single-nucleotide alterations.

**Methods:** Here we present an improvement of the mutational forks formalism to infer the patient-specific flow of signal transduction based on multiple single-nucleotide alterations, including non-synonymous and synonymous mutations. The lung adenocarcinoma and skin cutaneous melanoma datasets from TCGA Pan-Cancer Atlas have been employed to show the utility of the proposed method.

**Results:** We have comprehensively characterized six mutational forks. The number of mutated nodes ranged from one to four depending on the topological characteristics of a fork. Transitional confidences (TCs) have been computed for every possible combination of single-nucleotide alterations in the fork. The performed analysis demonstrated the capacity of the mutational forks formalism to follow a biologically explainable logic in the identification of high-likelihood signaling routes in lung adenocarcinoma and skin cutaneous melanoma patients. The findings have been largely supported by the evidence from the biomedical literature.

**Conclusion:** We conclude that the formalism has a great chance to enable an assessment of patient-specific flow by leveraging information from multiple single-nucleotide alterations to adjust the transitional likelihoods that are solely based on the canonical view of a disease.

**Keywords:** mutational forks; signaling pathways; cancer; single-nucleotide alterations

**Author summary:** Oncogenic mutations act as molecular switches in signaling networks, disrupting the normal functioning of the cells. The precise mechanism that mediates the effects of oncogenic mutations upon various signaling pathways remains elusive. In this paper, we present an improvement of the mutational forks formalism, where literature findings about pathway interactions have been used as a prior to predict the flow of signal transduction impacted by multiple mutations. Our analysis demonstrated the capacity of our method to follow a biologically explainable logic in the identification of crucial signaling routes for patients with lung cancer and skin cancer.

## INTRODUCTION

Signaling pathways play a crucial role in the transformation of external signals into gene regulation mechanisms. In this process, external information is transferred through a series of biochemical reactions, which can be referred to as “signal flow”. Dysregulation of signal flow by a certain mutation may trigger the development of pathological processes, including cancer [1]. It is especially evident in individual patient cases, where the inferred flow can illustrate the deviation from a canonical form of the disease and therefore explain drug resistance or pave the way to new therapeutic regimens.

Various computational methods have been developed for pathway analysis in recent years [2,3]. However, only a small number of papers have focused on the development of models that do not primarily rely on kinetic parameters. To overcome the difficulty in accurately estimating kinetic parameters, pathway topology methods employ various characteristics of a pathway network, such as node degree, betweenness, clustering coefficient, and shortest path length to predict functionally important nodes in the network. For example, Zhao & Liu analyzed topological parameters of complex disease genes and revealed the importance of their location in a biomolecular network [4]. Erten *et al.* developed an algorithm called Vavien for harnessing the topological similarity of proteins in a network of interactions to prioritize candidate disease-associated genes [5]. Ning *et al.* attempted to examine the placement of essential proteins and the network topology by determining the correlation of protein essentiality and reverse nearest neighbor topology (RNN) [6]. However, such methods are limited only to a small network of proteins and do not infer the gene-to-gene direction of signal propagation in biological systems.

To address these limitations, Lee & Cho developed a novel algorithm that can estimate signal flow using only topological information to predict the direction of activity change in various signaling networks [7]. They extended their work to the signal flow control (SFC) method, which is capable of identifying control targets without the information of kinetic parameter values [8]. The limitation of these probabilistic topological methods is that the flow of signal transduction is equally distributed among connected nodes. Therefore, the weights of the edges are roughly approximated from topology and heavily depend on the number of connecting nodes. In addition, only binary interactions have been encoded to the model, such as activation and inactivation.

A separate group of methods deals with inference of perturbation patterns in signaling systems. For example, Santolini & Barabasi proposed a series of network

topology-based models named “DYNamics-Agnostic Network MOdels (DYNAMO)”, which can retrieve the relative magnitude of biological perturbation patterns when lacking knowledge of the kinetic parameters [9]. The DYNAMO series models achieve an average of 65% accuracy in predicting perturbation patterns of both distance-based and biochemical models, which shows promising results. Li & Gao enhanced the distance model in DYNAMO and leveraged graph convolutional networks (GCN) to improve perturbation pattern prediction [10]. However, the limitation of pathway perturbation methods, such as DYNAMO and GCN, is that they do not focus on inferring the gene-to-gene direction of signal propagation, instead pinpointing localities where pathway perturbation occurred.

In this work, we present an improvement of the mutational forks formalism which is capable of handling a combination of single-nucleotide alterations for inferring patient-specific flow. By definition, a mutational fork is a subgraph where signal transduction can flow through alternative paths. We have assigned prior probabilities to the flow of signal transduction through particular entities based on literature findings, rather than approximating it solely from topology. In addition, we employed ontologically-enriched pathway data that preserve the diversity of gene interaction terms, such as phosphorylation-activation, ubiquitination, binding, and degradation. We have also conducted two case studies to show that the proposed algorithm follows a specific, biologically explainable logic, and can thus be employed as hypothesis generation tool in precision medicine settings.

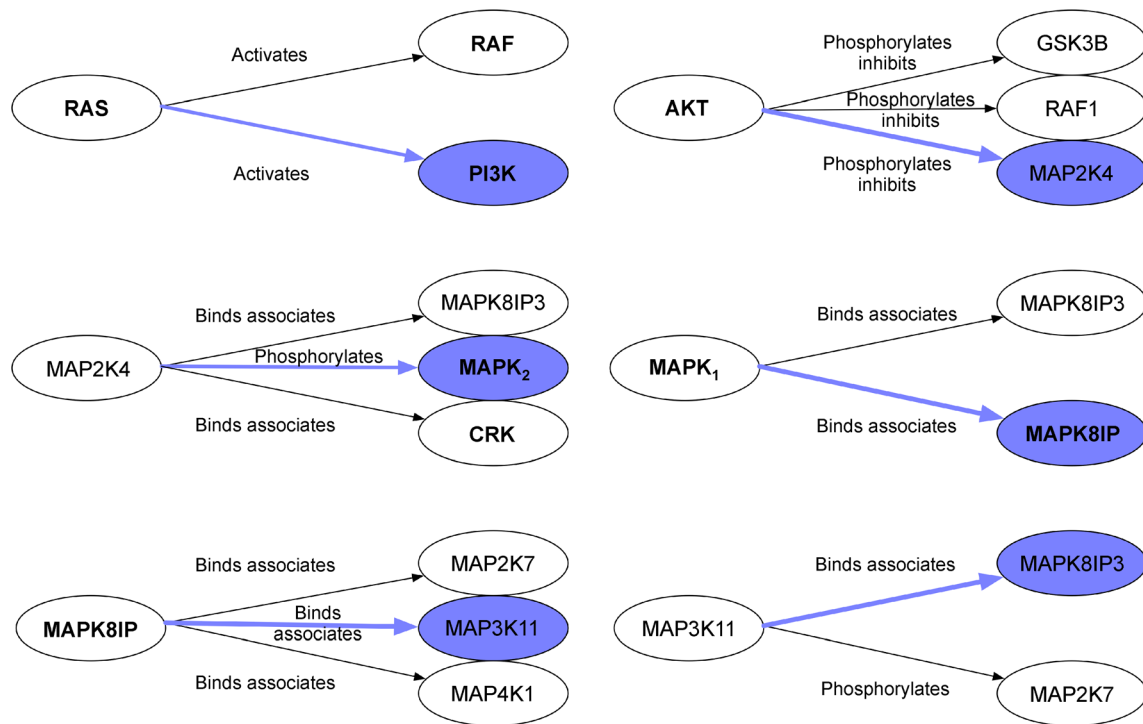
## RESULTS

### Summarization of extracted mutational forks

After setting the *EGFR* gene as a start gene and the *CCND1* gene as an end gene, we identified 178 unique linear paths between these gene entities. The length of the linear paths varied between eight and eleven gene entities. By combining frequently occurring linear paths, we have extracted eleven mutational forks from the RDF knowledge network using a custom script. Out of these eleven, six mutational forks were selected as having distinct frequencies for each transition (Fig. 1). Gene entities corresponding to protein families are highlighted in bold.

### Case study 1: lung adenocarcinoma

Table 1 presents an example of TC calculations for the KRAS mutational fork in lung adenocarcinoma patients. Detailed TC calculations for the rest of the mutational forks have been provided in the Supplementary Materials.



**Figure 1. Extracted mutational forks for lung adenocarcinoma patients.** The high likelihood route highlighted in orchid, along with corresponding shoulder gene. Protein families are highlighted in bold (see subsequent Table).  $MAPK_1$  refers to either  $MAPK1$  or  $MAPK3$ ;  $MAPK_2$  refers to either  $MAPK8$  or  $MAPK9$  or  $MAPK10$ .

### KRAS mutational fork

Interestingly, the mutational status of the BRAF gene increases the possibility of transitioning by the KRAS-PIK3CA path (see Block 1 in Table 1). While BRAF and PIK3CA mutations occur simultaneously, the possibility of transitioning by the KRAS-PIK3CA path is higher than in the other two cases (see Block 2 in Table 1). When junction gene KRAS is not mutated, a transition by the KRAS-PIK3CA path is more likely. However, when junction gene KRAS is mutated, the differences between various TC values are more pronounced.

Using functional classification of network-attacking mutations (NAMs) [11], the KRAS mutational fork represents an example of dysregulation of network dynamics, specifically the “node inactivation” scenario. Inactivating mutations on the essential residues of BRAF kinase lead to the interruption of information flow through this node. Thus, the phosphorylation of subsequent BRAF substrate—MAP2K1 protein is disabled, which in turn allows PIK3CA protein kinase through generation of secondary messenger PIP3 to recruit AKT1 and PDPK1 proteins to the cellular membrane, thus activating the PI3K-AKT axis [12,13].

### AKT1 mutational fork

Notably, mutation in the MAP2K4 gene alone increases the chance of transitioning by the AKT1-MAP2K4 route (see Block 1 in Supplementary Table S1). While MAP2K4 and RAF1 mutations occur conjointly, the possibility of transitioning by the AKT1-MAP2K4 path is higher than in other cases (see Block 2 in Supplementary Table S1). Similarly, conjoint mutations in the AKT1, RAF1 and MAP2K4 genes increase the chance of transitioning by the AKT1-MAP2K4 route, comparing to other combinations of mutations (see Block 3 in Supplementary Table S1). Overall, the mutational status of junction gene AKT1 does not profoundly impact transition by the AKT1-MAP2K4 route.

Following functional classification of NAMs, the AKT1 mutational fork represents an example of dysregulation of network dynamics, specifically the “node activation” scenario. It has been shown that AKT1 regulates the storage of glucose in the form of glycogen by phosphorylating GSK3B at the “Ser-9” residue, resulting in inhibition of its kinase activity [14]. In addition, AKT1 phosphorylates RAF1 at the “Ser-259” residue and negatively regulates its activity as well [15].

**Table 1 TC for KRAS mutational fork in lung adenocarcinoma patients**

Transition	Description	TC	
Block 1	KRAS-BRAF	$TC(T_{JA}(J_{AB})   M(J) = 1, M(A) = 0, M(B) = 0)$	0.0416
		$TC(T_{JA}(J_{AB})   M(J) = 0, M(A) = 1, M(B) = 0)$	0.0448
		$TC(T_{JA}(J_{AB})   M(J) = 0, M(A) = 0, M(B) = 1)$	0.0442
	KRAS-PIK3CA	$TC(T_{JB}(J_{AB})   M(J) = 1, M(A) = 0, M(B) = 0)$	0.8807
		$TC(T_{JB}(J_{AB})   M(J) = 0, M(A) = 1, M(B) = 0)$	<b>0.9539</b>
		$TC(T_{JB}(J_{AB})   M(J) = 0, M(A) = 0, M(B) = 1)$	0.942
Block 2	KRAS-BRAF	$TC(T_{JA}(J_{AB})   M(J) = 1, M(A) = 1, M(B) = 0)$	0.0389
		$TC(T_{JA}(J_{AB})   M(J) = 1, M(A) = 0, M(B) = 1)$	0.0383
		$TC(T_{JA}(J_{AB})   M(J) = 0, M(A) = 1, M(B) = 1)$	0.0435
	KRAS-PIK3CA	$TC(T_{JB}(J_{AB})   M(J) = 1, M(A) = 1, M(B) = 0)$	0.8294
		$TC(T_{JB}(J_{AB})   M(J) = 1, M(A) = 0, M(B) = 1)$	0.8217
		$TC(T_{JB}(J_{AB})   M(J) = 0, M(A) = 1, M(B) = 1)$	<b>0.9260</b>
Block 3	KRAS-BRAF	$TC(T_{JA}(J_{AB})   M(J) = 1, M(A) = 1, M(B) = 1)$	0.0231
	KRAS-PIK3CA	$TC(T_{JB}(J_{AB})   M(J) = 1, M(A) = 1, M(B) = 1)$	0.4913

However, mutations in the MAP2K4 gene allow avoiding of negative regulation from AKT1 phosphorylation, thus enabling direct activation of downstream targets, such as MAPK8, MAPK9 and MAPK10 [16].

#### MAP2K4 mutational fork

Interestingly, mutations in junction gene MAP2K4 can alone increase chances of transitioning by the MAP2K4-MAPK8 path (see Block 1 in Supplementary Table S2). Conjoint mutations in MAP2K4 and CRK genes increase the possibility of transitioning by the MAP2K4-MAPK8 path (see Block 2 in Supplementary Table S2). Furthermore, conjoint mutations in the MAP2K4, MAPK8 and CRK genes increase chances of transitioning by the MAP2K4-MAPK8 path as well (see Block 3 in Supplementary Table S2). Overall, the mutational status of junction gene MAP2K4 can profoundly impact transition by the MAP2K4-MAPK8 path.

Based on the functional classification of NAMs, the MAP2K4 mutational fork represents an example of network rewiring, specifically the “downstream rewiring” scenario. Evidently, MAP2K4 mutations shift the signaling network structure to the MAP2K4-MAPK8 route by “rewiring” downstream interactions with MAPK8IP3, MAPK8 and CRK proteins. Such network behavior can be explained by drifts in the peptide specificity upon mutation of the determinants of specificity (DoS) in the MAP2K4 kinase. Therefore, activated MAPK8 phosphorylates a number of transcription factors, primarily components of AP-1 such as JUN, JDP2 and ATF2, and thus regulates AP-1 transcriptional activity [17].

#### MAPK8IP1 mutational fork

Notably, a mutation in the junction gene MAPK8IP1 alone increases chances of transitioning by the MAPK8IP1-MAP3K11 path (see Block 1 in Supplementary Table S3). While MAPK8IP1 and MAP2K7 mutations occur conjointly, a transition by the MAPK8IP1-MAP3K11 path is more probable (see Block 2 in Supplementary Table S3). A similar TC value has been obtained when the MAPK8IP1 and MAP3K11 genes were mutated simultaneously (TC = 0.403 in both cases). Furthermore, conjoint mutations in the MAPK8IP1, MAP3K11 and MAP2K7 genes increase a chance of transitioning by the MAPK8IP1-MAP3K11 path as well (see Block 3 in Supplementary Table S3). Overall, the mutational status of junction gene MAPK8IP1 profoundly impact the transition of the MAPK8IP1-MAP3K11 path. Unfortunately, since MAPK8IP1 is not a phosphoprotein, we were unable to derive functional implications of its single-nucleotide mutations for corresponding junction genes.

#### MAP3K11 mutational fork

Interestingly, mutation in the shoulder gene MAP2K7 alone increases the chance of transitioning by the MAP3K11-MAPK8IP3 path (see Block 1 in Supplementary Table S4). A similar TC value has been obtained, while the junction gene MAP3K11 mutated alone (TC = 0.5080). Moreover, when both genes are mutated (see Block 2 in Supplementary Table S4), the chance of transitioning by the MAP3K11-MAPK8IP3 path is higher than the remaining cases. Overall, the mutational status of

the junction gene MAP3K11 profoundly impacts the transition by the MAP3K11-MAPK8IP3 path.

Using the functional classification of NAMs, the MAP3K11 mutational fork represents an example of dysregulation of network dynamics, specifically the “node inactivation” scenario. Inactivating mutations on the essential residues of the MAP2K7 kinase lead to an interruption of information flow through this node. Thus, the direct activation of the MAP2K7 downstream substrates, such as MAPK8, MAPK9 and MAPK10, is disabled, which in turn allows the MAPK8IP3 scaffold protein to selectively mediate JNK signaling by aggregating specific components of the MAPK cascade to form a functional JNK signaling module [17].

#### MAPK8 mutational fork

Notably, mutation in the junction gene MAPK8 alone increases chances of transitioning by the MAPK8-MAPK8IP1 path (see Block 1 in Supplementary Table S5). Furthermore, while junction gene MAPK8 conjointly mutated with shoulder gene MAPK8IP1, the possibility of the MAPK8-MAPK8IP1 path is the highest (see Block 2 in Supplementary Table S5). Overall, the mutational status of junction gene MAPK8 impacts the transition by the MAPK8-MAPK8IP1 route.

From the functional classification of NAMs, the MAPK8 mutational fork represents an example of network rewiring, specifically the “downstream rewiring” scenario. Evidently, MAPK8 mutations shift the signaling network structure to the MAPK8-MAPK8IP1 route by “rewiring” downstream interactions with MAPK8IP1 and MAPK8IP3 proteins. It has been shown that cytoplasmic MAPK8IP1 causes inhibition of JNK-regulated activity by retaining JNK in the cytoplasm and inhibiting JNK phosphorylation of c-Jun [18].

### Case study 2: skin cutaneous melanoma

#### KRAS mutational fork

Interestingly, the mutational status of the junction gene KRAS increases the possibility of transitioning by the KRAS-PIK3CA path (see Block 1 in Supplementary Table S6). While KRAS and PIK3CA mutations occur simultaneously, the possibility of transitioning by the KRAS-PIK3CA path is higher than in the other two cases (see Block 2 in Supplementary Table S6). These results differ from the KRAS mutational fork in lung adenocarcinoma, where mutations in the shoulder gene BRAF plays more profound role in transitioning by the KRAS-PIK3CA axis.

#### AKT1 mutational fork

Notably, mutation in the shoulder gene MAP2K4 alone increases the chance of transitioning by the AKT1-MAP2K4 route (see Block 1 in Supplementary Table S7). While MAP2K4 and GSK3B mutations occur conjointly, the possibility of transitioning by the AKT1-MAP2K4 path is higher than in other cases (see Block 2 in Supplementary Table S7). Similarly, conjoint mutations in the AKT1, GSK3B and MAP2K4 genes increase the chance of transitioning by the AKT1-MAP2K4 route, comparing to other combinations of mutations (see Block 3 in Supplementary Table S7). Unlike the AKT1 mutational fork in lung adenocarcinoma, where combination of MAP2K4 and RAF1 mutations increase the possibility of transitioning by the AKT1-MAP2K4 route, mutations in the GSK3B gene are more likely to define the flow of signal towards AKT1-MAP2K4 axis.

#### MAP2K4 mutational fork

Interestingly, mutations in the junction gene MAP2K4 can alone increase chances of transitioning by the MAP2K4-MAPK8 path (see Block 1 in Supplementary Table S8). Conjoint mutations in MAP2K4 and CRK genes, as well as in another pair of genes – MAPK8 and CRK – increase the possibility of transitioning by the MAP2K4-MAPK8 path (see Block 2 in Supplementary Table S8). Furthermore, conjoint mutations in MAP2K4, MAPK8 and CRK genes increase chances of transitioning by the MAP2K4-MAPK8 path as well (see Block 3 in Supplementary Table S8). Overall, the mutational status of junction gene MAP2K4 can profoundly impact transition by the MAP2K4-MAPK8 path. In comparison to the MAP2K4 mutational fork in lung adenocarcinoma, CRK gene is widely implicated into transitioning by the MAP2K4-MAPK8 axis.

#### MAPK8IP1 mutational fork

Notably, a mutation in the shoulder gene MAP2K7 alone increases chances of transitioning by the MAPK8IP1-MAP3K11 path (see Block 1 in Supplementary Table S9). While MAPK8IP1 and MAP2K7 mutations occur conjointly, a transition by the MAPK8IP1-MAP3K11 path is more probable (see Block 2 in Supplementary Table S9). Furthermore, conjoint mutations in the MAPK8IP1, MAP2K7 and MAP4K1 genes increase a chance of transitioning by the MAPK8IP1-MAP3K11 path as well (see Block 3 in Supplementary Table S9). These results differ from the MAPK8IP1 mutational fork in lung adenocarcinoma, where the junction gene MAPK8IP1 plays more profound role in transitioning by the MAPK8IP1-MAP3K11 axis.

### MAP3K11 mutational fork

Interestingly, mutation in the shoulder gene MAP2K7 alone increases the chance of transitioning by the MAP3K11-MAP2K7 path (see Block 1 in Supplementary Table S10). Moreover, when MAP3K11 and MAPK8IP3 genes are mutated concomitantly (see Block 2 in Supplementary Table S10), the chance of transitioning by the MAP3K11-MAP2K7 path is higher than the remaining cases. Unlike the MAP3K11 mutational fork in lung adenocarcinoma, where signal flow through the MAP3K11-MAPK8IP3 route, the predominant transition for the MAP3K11 mutational fork using melanoma dataset is the MAP3K11-MAP2K7 axis.

### MAPK8 mutational fork

Notably, mutation in the junction gene MAPK8 alone increases chances of transitioning by the MAPK8-MAPK8IP1 path (see Block 1 in Supplementary Table S11). Furthermore, while junction gene MAPK8 conjointly mutated with shoulder gene MAPK8IP1, the possibility of the MAPK8-MAPK8IP1 path is the highest (see Block 2 in Supplementary Table S11). Overall, the mutational status of junction gene MAPK8 impacts the transition by the MAPK8-MAPK8IP1 route. These findings are consistent with the MAPK8 mutational fork in lung adenocarcinoma.

## DISCUSSION

In this paper, we have presented the mutational forks formalism for inferring patient-specific flow of signal transduction based on multiple single-nucleotide alterations. The fundamental idea behind the mutational forks formalism is an application of a Bayesian “flip” principle to infer the most probable transitions of signaling flow given the mutational profile of the case. In this setup, the marginal *prior probability* of a transition is computed based on its frequency as the transition appears in literature-derived knowledge bases such as KEGG. This quantity, *i.e.*, the *initial belief* of the transition, reflects a chance of the transition from a *canonical* view of the disease, without accounting for any patient-specific data. The *normalized conditional probability* of a mutation given the transition is then computed and represents the *support* the mutation provides to the transition. In this scheme, the Bayesian “flip” enables the assessment of a chance of the transition based on mutations in a specific patient cohort. In other words, the initial probability of the transition is adjusted by the mutational profile of the cohort. The resulting quantity, *i.e.*, the *posterior probability of the transition given mutations*, may not strictly follow the mathematics of the probabilistic systems.

Therefore, we call it the *transitional confidence* (TC) and compute it for a set of network junctions, which we call *mutational forks* (MF), in context of this study.

The proposed method provides a more comprehensive view of pathway deregulation in cancer compared to existing pathway tools [8–11]. For example, our approach uses literature findings about pathway interactions as a prior to predict the flow of signal transduction, rather than approximating it from the topology of the interactome. In addition, the method expands semantics of gene-to-gene interaction terms to 27 interactions, including *binds*, *associates*, *phosphorylates*, *activates*, *phosphorylates*, *inhibits*, etc. Finally, our method helps to pinpoint mutated genes that are directly involved in transitioning by a specific pathway route. For example, mutations in BRAF gene increases the possibility of signal flow through the KRAS-PIK3CA axis in lung adenocarcinoma, while mutations in KRAS gene have more profound impact in transitioning by the same route in melanoma cohort.

The power of the pathway-based approach is that it may provide clues about the possible mechanisms underlying the differences in observed cancer patients’ survival. Mutational forks may be useful for suggesting therapeutic targets or to select the most appropriate patients for clinical trials. For example, the KRAS G12C mutation is a well-known marker of particular forms of lung cancer that are treatable by the drug AMG 510 (Amgen) [19]. However, some patients with the KRAS G12C mutation have tumors that are resistant to treatment. By considering combinations of mutational events incorporated in the inference of transitional confidence, our approach can identify a cohort of patients with the KRAS G12C mutation that is functionally inactive at the pathway level due to a cumulative silencing effect from other mutations. Thus, the treatment-resistant nature of these patients can be explained by their pathway mutational status and when this pattern is observed, different therapy can be chosen. This approach provides us with the opportunity to avoid toxicity and wasting patient time with predictably ineffective therapy, thus directing clinicians to potentially more efficacious treatment upfront, which is the primary goal of precision health and patient-centered medicine.

We chose lung adenocarcinoma and skin cutaneous melanoma as case studies for several reasons. Both malignancies have high mutational burden [20,21] and it is still unclear which path will be activated given a combination of such mutational events if, for example, a cancer patient has both BRAF and PIK3CA mutations. By knowing the exact signaling path affected by the cumulative impact of multiple mutations, clinicians would be able to tackle treatment resistance more effectively. Although our experiments had been conducted on genomics profiles derived from lung adenocarcinoma and melanoma patients, our method can be

relevant to another types of cancer as well.

There are some limitations to our method. Due to the algorithmic design, our approach employs a limited number of curated signaling pathways relevant to cancer theranostics. Furthermore, our method accounts for the functional impact of single-nucleotide mutations, taking into account the sporadic nature of investigated malignancies. Due to the small scope of our study, we were unable to utilize additional database resources, such as SynMICdb [22], however, we will employ this database in our future work. Since our intention was to show a holistic picture of pathway deregulation for cancer patients given multiple mutational events, we did not differentiate among various phenotypical populations of cancer patients. Finally, the clonal evolution of solid tumors is beyond the scope of this article.

From a data standpoint, Leek *et al.* argued that differences in sample collection and processing across multiple centers could lead to batch effects, or variation in the data due to technical factors that mask relevant biological variation. For example, while considering TCGA data, indel calls were more sensitive to technical artifacts than single-nucleotide variant (SNV) calls [23]. Therefore, to avoid false positive associations we considered only single-nucleotide mutations in this paper.

In the future, we plan to incorporate more signaling pathways and to test the pipeline on a larger cohort of patients. It is possible to extend the analysis to mutational forks that consist of a larger number of genes, given a sufficient amount of computing resources. To reduce the search space for potential gene candidates in a particular mutational fork, we will employ heuristic approaches, including greedy and semi-greedy search, throughout the interactome. For example, the EGFR gene has more than a hundred potential gene-interaction partners, and by applying heuristic searches the analysis could be narrowed down to the most relevant nodes, which in turn would help to address the combinatorial complexity of TC calculations.

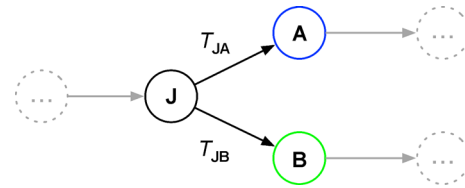
We also plan to redefine the activation status of the gene (*e.g.*, two standard deviations from the mean value of the expression) to account for highly differentially expressed genes. Moreover, we plan to incorporate other types of genomic data (*e.g.*, deletions, insertions) and use TREC collections and MEDLINE for non-canonical literature findings. Finally, we will conduct studies outside of the scope of cancer research to test the generalizability of the proposed method.

## MATERIALS AND METHODS

### Introduction and definitions

This section provides a brief overview of definitions in the

mutational fork formalism (Fig. 2). Here, we use bold font for all pathway entities and italic font for functions and probability distributions. For more details, please refer to our previous publication [24].



**Figure 2.** An example of a dual mutational fork.

**J** stands for junction gene, **A** and **B** denote shoulder genes,  $T_{JA}$  signifies transition of signal flow from the node **J** to the node **A**,  $T_{JB}$  represents transition of signal transduction flow from the node **J** to the node **B**, black arrows depict the direction of signal flow in the fork. Dashed-line circles represent incoming and outgoing nodes in the fork, grey arrows denote the direction of signal flow from incoming and outgoing nodes in the fork.

**J<sub>AB</sub>** – *mutational fork*, a place where signal transduction can flow through alternative paths

**J** – *junction gene*, where signal transduction flow splits

**A**, **B** – *shoulder genes*, the first genes in alternative paths

**JA**, **JB** – *legs of the mutational fork J<sub>AB</sub>*

$T_{JA}(\mathbf{J}_{AB})$  – transition of signal transduction flow to leg

**JA** of the mutational fork **J<sub>AB</sub>**

$T_{JB}(\mathbf{J}_{AB})$  – transition of signal transduction flow to leg **JB** of the mutational fork **J<sub>AB</sub>**

$P(T_{JA}(\mathbf{J}_{AB}))$  – probability of  $T_{JA}(\mathbf{J}_{AB})$

$P(T_{JB}(\mathbf{J}_{AB}))$  – probability of  $T_{JB}(\mathbf{J}_{AB})$

$M(\mathbf{X})$  – mutation of gene **X**

$C(M(\mathbf{X}))$  – mutation condition of gene **X**

$A(\mathbf{X})$  – status of gene **X**

$C(A(\mathbf{X}))$  – activation condition of gene **X**

$P(M(\mathbf{A}) | T_{JA}(\mathbf{J}_{AB}))$  – probability of  $M(\mathbf{A})$  given  $T_{JA}(\mathbf{J}_{AB})$

$P(M(\mathbf{A}) | T_{JB}(\mathbf{J}_{AB}))$  – probability of  $M(\mathbf{A})$  given  $T_{JB}(\mathbf{J}_{AB})$

$P(M(\mathbf{B}) | T_{JB}(\mathbf{J}_{AB}))$  – probability of  $M(\mathbf{B})$  given  $T_{JB}(\mathbf{J}_{AB})$

$P(M(\mathbf{B}) | T_{JA}(\mathbf{J}_{AB}))$  – probability of  $M(\mathbf{B})$  given  $T_{JA}(\mathbf{J}_{AB})$

$P(T_{JA}(\mathbf{J}_{AB}) | M(\mathbf{A})=1)$  – probability of  $T_{JA}(\mathbf{J}_{AB})$  given  $M(\mathbf{A})=1$

$P(T_{JA}(\mathbf{J}_{AB}) | M(\mathbf{B})=1)$  – probability of  $T_{JA}(\mathbf{J}_{AB})$  given  $M(\mathbf{B})=1$

$P(T_{\mathbf{JB}}(\mathbf{J}_{\mathbf{AB}}) | M(\mathbf{B})=1)$  – probability of  $T_{\mathbf{JB}}(\mathbf{J}_{\mathbf{AB}})$  given  $M(\mathbf{B})=1$

$P(T_{\mathbf{JB}}(\mathbf{J}_{\mathbf{AB}}) | M(\mathbf{A})=1)$  – probability of  $T_{\mathbf{JB}}(\mathbf{J}_{\mathbf{AB}})$  given  $M(\mathbf{A})=1$

We define mutation and activation statuses of a particular gene  $\mathbf{X}$ , given a corresponding mutational event as shown in Eqs. (1) and (2). The gene is considered to be mutated/activated if the corresponding conditions

$$M(\mathbf{X}) = \begin{cases} 1, & \text{if } C(M(\mathbf{X})) \text{ is met} \\ 0, & \text{otherwise} \end{cases}, \text{ value 1 means gene } \mathbf{X} \text{ is mutated;} \tag{1}$$

$$A(\mathbf{X}) = \begin{cases} 1, & \text{if } C(A(\mathbf{X})) \text{ is met} \\ 0, & \text{otherwise} \end{cases}, \text{ value 1 means gene } \mathbf{X} \text{ is active.} \tag{2}$$

### Mutational forks formalism for combinatorial perturbations

Here, we introduce the notion of *transitional confidence* (TC) to characterize transitions from the junction gene to any of the shoulder genes, based on *simultaneous mutations in several genes* in a fork. Consider an example of combinatorial perturbation, where the shoulder genes  $\mathbf{A}$  and  $\mathbf{B}$  are mutated, while the junction gene  $\mathbf{J}$  is not mutated:

$$TC(T_{\mathbf{JA}}(\mathbf{J}_{\mathbf{AB}}) | M(\mathbf{J})=0, M(\mathbf{A})=1, M(\mathbf{B})=1) = \frac{P(M(\mathbf{J})=0, M(\mathbf{A})=1, M(\mathbf{B})=1 | T_{\mathbf{JA}}(\mathbf{J}_{\mathbf{AB}})) * P(T_{\mathbf{JA}}(\mathbf{J}_{\mathbf{AB}}))}{P(M(\mathbf{J})=0, M(\mathbf{A})=1, M(\mathbf{B})=1)} \tag{3}$$

$P(M(\mathbf{J})=0, M(\mathbf{A})=1, M(\mathbf{B})=1 | T_{\mathbf{JA}}(\mathbf{J}_{\mathbf{AB}}))$  – number of patient cases, where shoulder genes  $\mathbf{A}$  and  $\mathbf{B}$  are mutated and activated, while junction gene  $\mathbf{J}$  is not mutated, divided by all cases in the cohort.

$P(M(\mathbf{J})=0, M(\mathbf{A})=1, M(\mathbf{B})=1)$  – number of patient cases, where shoulder genes  $\mathbf{A}$  and  $\mathbf{B}$  are mutated, and junction gene  $\mathbf{J}$  is not mutated regardless of the activation status of gene  $\mathbf{A}$  and  $\mathbf{B}$ , divided by all cases in the cohort.

$P(T_{\mathbf{JA}}(\mathbf{J}_{\mathbf{AB}}))$  – frequency of topological linear motif from junction gene  $\mathbf{J}$  to shoulder gene  $\mathbf{A}$ .

Following the logic in Eq. (3), we can infer a perturbation pattern for any combination of mutations in the mutational fork. Please note that unlike previously described pathway analysis tools [8–12], our approach uses literature findings about pathway interactions as a prior to predict the flow of signal transduction. The TC concept is fundamental in this paper, because it not only enables the identification of gene-to-gene perturbation patterns, but also is essential for determining patient-specific flow through a particular mutational fork. Since

$C(M(\mathbf{X}))$  and  $C(A(\mathbf{X}))$  are met. In this work, the activation condition is set when the expression value of  $\mathbf{X}$  exceeds the average gene expression values for  $\mathbf{X}$  among all patients after a mutational event occurred. The mutation condition  $C(M(\mathbf{X}))$  is enabled when a single-nucleotide polymorphism (SNP) in gene  $\mathbf{X}$  occurred.

the previous version of the method addressed only single mutational events in the inference of transitional confidences [25], we improved the current version by accounting for all possible combinations of mutations in forks and, therefore, captured a larger scope of pathway deregulation events for lung adenocarcinoma patients.

### Data acquisition and preprocessing

#### Pathway data

The pathway data have been acquired from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [26]. These pathway maps have been processed using a two-step curation approach.

The first step of curation consists of removing inconsistencies and errors (e.g., missing or erroneous gene-to-gene interactions) that were present in original KGML files. During this process, all gene entities and their interactions were verified using cited literature to ensure accurate data provenance.

The second step of curation consists of data formalization and data transformation procedures. The data formalization procedure affects gene-to-gene interactions and their directionality. To formalize gene-to-gene interactions, we have replaced simplistic interaction types in KEGG pathway maps by a set of more specific biological relationships (e.g., *binds\_associates*, *phosphorylates\_activates*, *phosphorylates\_inhibits*). Overall, we have modeled 27 gene-to-gene interactions as in our previous work [27]. These interaction terms have been encoded as predicates in RDF triples. To formalize the directionality of interactions, signal transduction is considered to flow from the effectors to regulated gene entities, which correspondingly determine the *Start* and the *End* gene entities in a path. The data transformation procedure translates curated KGML files into traversable RDF *subject-predicate-object* triples.

Finally, we have integrated five different pathway maps, including the MAPK signaling pathway (hsa04010), calcium signaling pathway (hsa04020), JAK-STAT signaling pathway (hsa04630), NSCLC signaling pathway (hsa05223), and PI3K-AKT signaling pathway (hsa04151) into the RDF knowledge network as per our previous work [28]. To address computational complexity, we have introduced the notion of *Protein Family* to represent proteomic isoforms and enable traversal through singular protein family instances (Table 2). A protein family can take an “un-collapsed” form when a detailed analysis of a specific isoform is needed. All gene entities were linked to terms in standard gene and protein databases and ontologies such as GO [29] and Uniprot [30] through the Human Protein Reference Database (HPRD) [31].

**Table 2** Example of protein families

Protein family	Members
RAS	HRAS, KRAS, NRAS
RAF	ARAF, BRAF, RAF1
PI3K	PI3KCA, PI3KCB, PI3KCD, PI3KR1, PI3KR2, PI3KR3
MAP2K	MAP2K1, MAP2K2
MAPK <sub>1</sub>	MAPK1, MAPK3
AKT	AKT1, AKT2, AKT3
MAPK <sub>2</sub>	MAPK8, MAPK9, MAPK10
CRK	CRK, CRKL
MAPK8IP	MAPK8IP1, MAPK8IP2

## Genomics data

The lung adenocarcinoma patients’ profiles have been extracted from TCGA Pan-Cancer Atlas (TCGA-LUAD) at the cBioPortal platform [25]. Whole-exome sequencing (WES) has been performed on tumor and germ-line DNA with a mean coverage 97.6× and 95.8×, respectively. The mean somatic mutation rate across the lung adenocarcinoma cohort was 8.87 mutations per megabase (Mb) of DNA. The single-nucleotide mutation rate was 6.86 per Mb. Seventy-five percent of somatic mutations identified by WES were present in the RNA transcriptome when the locus in question was expressed (minimum 5×) [32]. The data consists of 566 patients, each represented by 114 attributes.

The skin cutaneous melanoma patients’ profiles have been extracted from TCGA Pan-Cancer Atlas (TCGA-SKCM) at the cBioPortal platform [25]. The melanoma cohort is focused on metastatic cases (11.6% regional skin cutaneous or subcutaneous metastatic tissue, 56.4% regional metastatic lymph node, 25.1% distant or

unspecified metastatic tissue) because melanoma is most often discovered after it has metastasized [33]. WES was performed on paired tumor and germline normal genomic DNA, including primary and metastatic melanomas with a mean exon coverage of 87×, adequate for detecting a single-nucleotide variant at an allelic fraction of 0.3 with a power of 80% [34]. The data consists of 448 patients, each represented by 114 attributes.

The following attributes were inputted to the mutational fork:

- number of cases, where gene X (or Y or Z) is mutated. This attribute requires information about specific alteration (*e.g.*, Pro369Gln);
- number of cases, where gene X (or Y or Z) is mutated and activated. This attribute requires information about specific alteration (*e.g.*, Pro369Gln) and the corresponding FPKM value;
- number of cases, where gene X (or Y or Z) is not mutated;
- topological frequency for transition from X to Y (or X to Z). This attribute requires information derived from gene-to-gene interaction pairs, for example, how frequent a specific gene pair appears in the set of extracted linear paths.

Whole exome sequencing is not required, and a combination of several candidate gene sequencings is acceptable. We have considered both non-synonymous and synonymous mutations for the downstream analysis, because they have significant impact on translation or co-translational protein folding and play an important role in many diseases including cancer [35]. In addition, we have averaged the expression values across cancer patients for each gene in order to measure the activation effect from the mutational event.

## Mutational forks extraction from RDF knowledge network

To reflect the direction of signal propagation, surface receptors have been encoded as *start genes*, while nuclear transcriptional factors have been encoded as *end genes* [36]. For this study, we chose the EGFR gene as a start gene, which corresponds to the cellular receptor and the CCND1 gene as the end gene, which corresponds to the nuclear target. The objective behind selection of these genes was that activation of the epidermal growth factor receptor (EGFR)-tyrosine kinases is a key reason for lung adenocarcinoma progression [37]. On the other hand, deregulation of CCND1 has been implicated in the pathogenesis of lung adenocarcinoma and is associated with poor prognosis [38]. However, other genes relevant to lung adenocarcinoma pathogenesis can be selected as well, including multiple-input, multiple-output options. The k-shortest path graph algorithm has been applied to

extract a set of *linear paths* from start to end genes. Finally, a custom script has identified frequently-occurring linear paths and combined them into mutational forks in a form of *dual* and *triple* forks.

### Identification of patient-specific signaling routes

We have performed three blocks of TC calculations for dual forks and four blocks of TC calculations for triple forks:

- Block 1: one mutational event occurred, either in the junction gene or one of the shoulder genes;
- Block 2: two mutational events occurred simultaneously, either in the junction gene and one of the shoulder genes or in two shoulder genes;
- Block 3: three mutational events occurred conjointly, either in the junction gene and two of the shoulder genes (in the case of dual and triple forks) or in three shoulder genes (in the case of triple forks only). A special case of this scenario has occurred when all genes are mutated in the triple fork.

We then calculated the TC for each block using Eq. (3).

### SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at <https://doi.org/10.1007/s40484-020-0227-0>.

### ACKNOWLEDGEMENTS

We appreciate the critical input of members from the Translational and Cancer Bioinformatics (TCBI) Laboratory at the University of Missouri-Columbia. This work has been supported by the MU Institute for Data Science and Informatics (IDSI) to O.K., by the Department of Pathology & Anatomical Sciences, School of Medicine, University of Missouri-Columbia to D.S and O.K. and in part by Career Development Award #IK2-BX004346-01A1 (JBM) from the United States (U.S.) Department of Veterans Affairs Biomedical Laboratory Research and Development Service. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Department of Veterans' Affairs.

### COMPLIANCE WITH ETHICS GUIDELINES

The authors Olha Kholod, Chi-Ren Shyu, Jonathan Mitchem, Jussuf Kaifi and Dmitriy Shin declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

### REFERENCES

1. Sever, R. and Brugge, J. S. (2015) Signal transduction in cancer. *Cold Spring Harb. Perspect. Med.*, 5, a006098
2. Khatri, P., Sirota, M. and Butte, A. J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Comput. Biol.*, 8, e1002375
3. Zhang, Q., Li, J., Xue, H., Kong, L. and Wang, Y. (2016) Network-based methods for identifying critical pathways of complex diseases: a survey. *Mol. Biosyst.*, 12, 1082–1089
4. Zhao, X. and Liu, Z. P. (2019) Analysis of topological parameters of complex disease genes reveals the importance of ation in a biomolecular network. *Genes (Basel)*, 10, 143
5. Erten, S., Bebek, G. and Koyutürk, M. (2011) Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J. Comput. Biol.*, 18, 1561–1574
6. Ning, K., Ng, H. K., Srihari, S., Leong, H. W. and Nesvizhskii, A. I. (2010) Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. *BMC Bioinformatics*, 11, 505
7. Lee, D. and Cho, K. H. (2018) Topological estimation of signal flow in complex signaling networks. *Sci. Rep.*, 8, 5262
8. Lee, D. and Cho, K. H. (2019) Signal flow control of complex signaling networks. *Sci. Rep.*, 9, 14289
9. Santolini, M. and Barabási, A. L. (2018) Predicting perturbation patterns from the topology of biological networks. *Proc. Natl. Acad. Sci. USA*, 115, E6375–E6383
10. Li, D. and Gao, J. (2019) Towards perturbation prediction of biological networks using deep learning. *Sci. Rep.*, 9, 11941
11. Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., Perryman, L., Cox, T. R., Zivanovic, N., Palmeri, A., *et al.* (2015) Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, 163, 202–217
12. Brennan, D. F., Dar, A. C., Hertz, N. T., Chao, W. C., Burlingame, A. L., Shokat, K. M. and Barford, D. (2011) A Raf-induced allosteric transition of KSR stimulates phosphorylation of MEK. *Nature*, 472, 366–369
13. Zhao, L. and Vogt, P. K. (2008) Class I PI3K in oncogenic cellular transformation. *Oncogene*, 27, 5486–5496
14. MacAulay, K. and Woodgett, J. R. (2008) Targeting glycogen synthase kinase-3 (GSK-3) in the treatment of Type 2 diabetes. *Expert Opin. Ther. Targets*, 12, 1265–1274
15. Zimmermann, S. and Moelling, K. (1999) Phosphorylation and regulation of Raf by Akt (protein kinase B). *Science*, 286, 1741–1744
16. Matsuura, H., Nishitoh, H., Takeda, K., Matsuzawa, A., Amagasa, T., Ito, M., Yoshioka, K. and Ichijo, H. (2002) Phosphorylation-dependent scaffolding role of JSAP1/JIP3 in the ASK1-JNK signaling pathway. A new mode of regulation of the MAP kinase cascade. *J. Biol. Chem.*, 277, 40703–40709
17. Yao, K., Cho, Y. Y., Bergen, H. R. 3rd, Madden, B. J., Choi, B. Y., Ma, W. Y., Bode, A. M. and Dong, Z. (2007) Nuclear factor of activated T3 is a negative regulator of Ras-JNK1/2-AP-1 induced cell transformation. *Cancer Res.*, 67, 8725–8735
18. Nihalani, D., Wong, H. N. and Holzman, L. B. (2003) Recruitment of JNK to JIP1 and JNK-dependent JIP1 phosphorylation regulates JNK module dynamics and activation. *J. Biol. Chem.*, 278, 28694–28702
19. Canon, J., Rex, K., Saiki, A. Y., Mohr, C., Cooke, K., Bagal, D., Gaida, K., Holt, T., Knutson, C. G., Koppada, N., *et al.* (2019) The

- clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature*, 575, 217–223
20. Forschner, A., Battke, F., Hadaschik, D., Schulze, M., Weißgraeber, S., Han, C. T., Kopp, M., Frick, M., Klumpp, B., Tietze, N., *et al.* (2019) Tumor mutation burden and circulating tumor DNA in combined CTLA-4 and PD-1 antibody therapy in metastatic melanoma—results of a prospective biomarker study. *J. Immunother. Cancer*, 7, 180
  21. Alexander, M., Galeas, J. and Cheng, H. (2018) Tumor mutation burden in lung cancer: a new predictive biomarker for immunotherapy or too soon to tell? *J. Thorac. Dis.*, 10, S3994–S3998
  22. Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., Backofen, R. and Diederichs, S. (2019) A pan-cancer analysis of synonymous mutations. *Nat. Commun.*, 10, 2569
  23. Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11, 733–739
  24. Kholod, O., Lyu, Z., Mitchem, J., Tonellato, P., Joshi, T., and Shin, D. (2019) Mutational forks: inferring deregulated flow of signal transduction based on patient-specific mutations, pp. 2063–2068. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
  25. Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, 6, p11
  26. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45, D353–D361
  27. Thanintorn, N., Wang, J., Ersoy, I., Al-Taie, Z., Jiang, Y., Wang, D., Verma, M., Joshi, T., Hammer, R., Xu, D., *et al.* (2016) RDF Sketch Maps—knowledge complexity reduction for precision medicine analytics. *Pac. Symp. Biocomput.*, 21, 417–428
  28. Shin, D., Arthur, G., Popescu, M., Korkin, D. and Shyu, C. R. (2014) Uncovering influence links in molecular knowledge networks to streamline personalized medicine. *J. Biomed. Inform.*, 52, 394–405
  29. The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47, D330–D338
  30. UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47, D506–D515
  31. Goel, R., Harsha, H. C., Pandey, A. and Prasad, T. S. (2012) Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst.*, 8, 453–463
  32. Pan-Cancer Genome Atlas Research Network. (2018). Available at: [https://www.cbioportal.org/study/summary?id=luad\\_tcga\\_pan\\_can\\_atlas\\_2018](https://www.cbioportal.org/study/summary?id=luad_tcga_pan_can_atlas_2018) (Accessed: August 23, 2020)
  33. Guan, J., Gupta, R. and Filipp, F. V. (2015) Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma. *Sci. Rep.*, 5, 7857
  34. Pan-Cancer Genome Atlas Research Network. (2018). Available at: [https://www.cbioportal.org/study/summary?id=skcm\\_tcga\\_pan\\_can\\_atlas\\_2018](https://www.cbioportal.org/study/summary?id=skcm_tcga_pan_can_atlas_2018) (Accessed: August 23, 2020)
  35. Cheng, N., Li, M., Zhao, L., Zhang, B., Yang, Y., Zheng, C. H. and Xia, J. (2020) Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief. Bioinform.*, 21, 970–981
  36. Sever, R. and Glass, C. K. (2013) Signaling by nuclear receptors. *Cold Spring Harb. Perspect. Biol.*, 5, a016709
  37. Liu, T. C., Jin, X., Wang, Y. and Wang, K. (2017) Role of epidermal growth factor receptor in lung cancer and targeted therapies. *Am J Cancer Res*, 7, 187–202
  38. Brennan, P., Hainaut, P. and Boffetta, P. (2011) Genetics of lung-cancer susceptibility. *Lancet Oncol.*, 12, 399–408