ORN: Extracting Latent Pathway Activities in Cancer with OR-gate Network

Lifan Liang¹, Kunju Zhu² and Songjian Lu^{1,*}

¹ Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15206, United States

² Clinical Medicine Research Institute, Jinan University, Guangzhou, Guangdong, China

* To whom correspondence should be addressed. Email: songjian@pitt

Present Address: Songjian Lu, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15206, United States

ABSTRACT

Pathway level understanding of cancer plays a key role in precision oncology. In this study, we developed a novel data-driven model, called the OR-gate Network (ORN), to simultaneously infer functional relationships among mutations, patient-specific pathway activities, and gene co-expression. In principle, logical OR gates agree with mutual exclusivity patterns in somatic mutations and bicluster patterns in transcriptomic profiles. In a trained ORN, the differential expression profiles of tumours can be explained by somatic mutations perturbing signalling pathways. We applied ORN to lower grade glioma (LLG) samples in TCGA and breast cancer samples from METABRIC. Both datasets have shown pathway patterns related to immune response and cell cycles. In LLG samples, ORN identified multiple metabolic pathways closely related to glioma development and revealed two pathways closely related to patient survival. Additional results from the METABRIC datasets showed that ORN could characterize key mechanisms of cancer and connect them to less studied somatic mutations (e.g., BAP1, MIR604, MICAL3, and telomere activities), which may generate novel hypothesis for targeted therapy.

INTRODUCTION

In the past decades, the emergence of high-throughput technology has greatly facilitated cancer research. One major insight from these efforts is that although somatic genomic alterations (SGA) vary from patient to patient, they exhibit similar patterns on pathway level (1). That is, different somatic mutations in different samples perturb the same signalling pathway and drive similar tumour phenotype. Based on this insight, many anti-cancer therapies targeting certain pathways have been successfully deployed (2). Unfortunately, most of these therapies are only effective for specific subpopulations. And patient responsiveness is often difficult to predict. For example, despite the effectiveness of trastuzumab in HER2 positive patients, some might suffer relapse with unknown reasons (3).

The current pathway-level understanding of cancer, though valuable, needs further investigation. Several computational methods have been developed to infer pathway activities for individual cancer sample (4, 5). For example, PARADIGM (4) adapted known pathways as Markov networks and inferred pathway activities from the observed multi-omics profiles with the EM algorithm. However, such methods usually relied on prior knowledge about signalling pathways. This may reduce the chance of obtaining novel biological insights, as our current knowledge is incomplete and biased towards well-studied pathways and genes (6). Furthermore, a recent study (7) has shown that databases of biological pathways are different from each other. Using different databases yielded disparate results in pathway analysis.

Therefore, computational researchers have proposed de novo methods (8–10) to identify latent patient pathways (factors) with multi-omics profiles. To ensure interpretability and computational feasibility, these methods assumed linear relationships among different types of omics data. Thus, it may not be sufficient to capture complicated biological relationships. On the other hand, some de novo methods were directly adopted from the machine learning community (11, 12), which may miss some biological features of cellular signalling pathways. For example, mutual exclusivity is a well-known phenomenon of somatic alterations in tumours, which is difficult to be captured by any machine learning model assuming additive effects. In fact, most methods (13–15) modelled mutual exclusivity with statistical tests.

In this work, we presented OR-gate Network (ORN), a de novo approach to infer the latent pathway activities in cancer. Fig. 1 shows how biological pathways (Fig. 1A) can be represented by an ORN (Fig. 1B). Although the order of signalling was ignored, ORN closely follows the causal mechanism from mutations to expressions. Each layer in ORN is connected by logical OR gates such that the output of a node is true if any input to the gate is true. This important property of the ORN agrees with the mutual exclusivity patterns observed in cancer genomics. On the other hand, when modelling the mechanism from latent pathway activities to transcriptomics, a pathway perturbed in a set of patients will cause differential expression of a set of genes. In this way, ORN generates the bicluster patterns a plethora of clustering algorithms aimed to recover in RNA expression data (16).

Since the generative process of ORN is consistent with both mutual exclusivity patterns in cancer genomics and bicluster patterns in cancer transcriptomics, we hypothesized that ORN can recover: (1) relations between pathways and downstream genes they regulate; (2) relations between SGA and the pathways they perturbed; (3) which pathway is dysregulated in which set of patients. To our knowledge, ORN is the first integrative model capable of inferring patient-specific pathway activities while soliciting corresponding SGAs and its regulon.

Figure 1 Figure A is the biologically plausible representation of signaling pathways. Given the difficulty in estimating the order of signal transduction, we replaced the realistic representation by introducing an additional variable indicating the status of pathway dysregulation, resulting in figure B. Genes in the genomic level are connected to pathway status with logical OR relationship. Pathway status is then connected to transcriptomics with the same logical OR relationships.

MATERIAL AND METHODS

Data preprocessing

Before applying ORN to real-world datasets, we need to transform genomic profiles and transcriptomic profiles into binary data. On the genomic side, we used two types of data: (1) non-silent gene-level simple somatic mutations (SNV) dataset; (2) gene-level copy number variation (CNV). An element in the CNV matrix was set to 1 if its original value was ≥ 2 . Any element with an absolute value less than 2 was set to 0. We then combined CNV and SNV data into a binary event matrix, that is, if the alteration of gene i was observed in either SNV or CNV in sample j , then the ij th element in the binary event matrix was set to 1.

For the transcriptomic profiles, we first remove genes with median expression counts lower than 10 across all samples. Then the Z scores provided by the CBioPortal platform were binarized. An element in the Z score matrix was set to 1 if its absolute value was greater than 1.96, otherwise 0.

To further filter the genes on the genomics side, we applied Multitask Lasso implemented in Scikit-learn (17). The genomic profiles were used as independent variables and the status of differential expression as targets. Somatic mutations with nonzero coefficients were retained as the input for our proposed algorithm.

Model formulation

Before describing the full model, we need to first illustrate the key operation unit, leaky OR-gate. When there are several possible causes $X_1, X_2, \dots, X_n$ of an effect variable Y , the probability of Y should be:

$$Pr(Y = 1|X_1, X_2, \dots, X_n) = OR(X_1, X_2, \dots, X_n) = 1 - \prod_{i=1}^n [1 - Pr(X_i = 1)Pr(X_i \rightarrow Y)]$$

This is different from the definition in (18) with the introduction of $Pr(X_i \rightarrow Y)$. This term represents the probability that X causes Y . It allows us to infer the layered causal structure of ORN.

However, there are situations where X cannot fully capture the cause of Y . For example, pathway perturbation in cancer cells may be caused by methylation changes, not just somatic mutations. To deal with unmodeled causes, we introduced the leaky parameter, which is similar with the intercept in linear regression:

$$P_0 = Pr(Y = 1|X_1 = 0, X_2 = 0, \dots, X_n = 0)$$

Hence the leaky OR gate function becomes:

$$Pr(Y = 1|X_1, X_2, \dots, X_n) = 1 - (1 - P_0) \prod_{i=1}^n [1 - Pr(X_i = 1)Pr(X_i \rightarrow Y)]$$

During implementation, P_0 can be modelled by adding a row into the binary event matrix with all ones. For convenience, we will denote the vectorized OR gate function as $OR(X, \delta)$ in all the following content, where X represents the vector $[Pr(X_i = 1)]$, and δ represents the vector $[Pr(X_i \rightarrow Y)]$.

With the leaky OR-gate operation, we connected SGA to latent pathways and then to DEGs, as illustrated in Fig. 1B. After vectorization, the procedure of ORN is illustrated in Fig. 2. Let us denote S as the number of samples, M the number of genes in genomics, P the number of pathways, G the number of genes in transcriptomics. Let Mut be the $S \times M$ binary event matrix of somatic mutation, $Path$ be the $S \times P$ pathway activity matrix, and $Expr$ the $S \times G$ RNA expression matrix. Let U denote the $M \times P$ causal relationship matrix between SGA and pathways. Let Z denote the $P \times G$ causal relationship matrix between pathways and DEG. For the s th sample, the activity of the p th pathway is:

$$Path_{s,p} = OR(Mut_{s,\cdot}, U_{\cdot,p})$$

And the status of the g th DEG in the s th sample is:

$$Expr_{s,g} = OR(Path_{s,\cdot}, Z_{\cdot,g})$$

In this formulation, patient-specific pathway activities, $Path$, were represented as the logical units within the hidden layer.

Figure 2 The source omics data for ORN consists of quantified matrices of single nucleotide variation (SNV), copy number variation (CNV), and gene expression (RNAseq). SNV and CNV were combined, binarized, and filtered on the genes' side to produce a binary event matrix. As for RNAseq, we calculate robust Z score for each gene in each sample. We assumed Logical OR relations when binary events led to pathway dysregulation and, in turn, led to differential expression. ORN algorithm aimed to infer: (1) the relationship between somatic mutations and signaling pathways; (2) the relationship between signaling pathways and differential expressions. With the output from ORN, we can recover the latent pathways that were perturbed by somatic mutations and caused differential expression.

The objective function to optimize for ORN is the overall log likelihood of the observed $Expr$ given estimated $Expr$:

$$LL(Expr) = \sum_{s \leq S, g \leq G} [Expr_{sg} \log \widehat{Expr}_{sg} + (1 - Expr_{sg}) \log (1 - \widehat{Expr}_{sg})]$$

where $\widehat{Expr}_{sg}$ is the probability of differential expression estimated by the model, in other words, the last layer of ORN. Since $Path$ and DEG can be computed by U and Z , they are the only parameters we need to estimate. Similar latent variable models, such as LDA (19), are usually computational expensive. However, we found that the layered structure of ORN is similar with the neural network architect. Thus, the powerful backward propagation approach, which is essential to all deep learning models, can be used to estimate ORN parameters. First, we need to reparameterize U and Z to μ and ζ , such that the parameters are not bounded within $[0,1]$:

$$U_{mp} = \text{sigmoid}(\mu)$$

$$Z_{pg} = \text{sigmoid}(\zeta)$$

Then we updated μ and ζ during gradient descent. The gradient of ζ with respect to $LL(Expr)$ is:

$$\frac{\partial LL}{\partial \zeta_{pg}} = Z_{pg}(1 - Z_{pg}) \sum_{n \leq N} \left[\frac{Path_{sp} \left(1 - \frac{Expr_{sg}}{\widehat{Expr}_{sg}} \right)}{1 - Path_{sp} Z_{pg}} \right]$$

Using the chain rule, we can derive the gradient of μ with respect to $LL(Expr)$:

$$\frac{\partial LL}{\partial \mu_{mp}} = \sum_{s \leq S} \frac{\partial LL}{\partial Path_{sp}} \cdot \frac{\partial Path_{sp}}{\partial U_{mp}} = \sum_{s \leq S} \frac{\partial LL}{\partial Path_{sp}} \cdot \frac{Mut_{sm} U_{mp} (1 - U_{mp}) (1 - Path_{sp})}{1 - Mut_{sm} U_{mp}}$$

where the computation of $\partial LL / \partial Path_{sp}$ is symmetric to $\partial LL / \partial Z_{sp}$.

To control the sparsity of parameters, we assumed U and Z are samples from the Beta distribution. Thus, the gradients above need to be modified. For example, the partial derivative of ζ should be modified as:

$$\frac{\partial LL}{\partial \zeta_{pg}^*} = \frac{\partial LL}{\partial \zeta_{pg}} + (\alpha - 1)(1 - \zeta) + (\beta - 1)\zeta$$

where α and β are the hyperparameters for Beta distribution. For all the experiments in this study, we set $\beta = \alpha = 0.95$.

Simulation and evaluation

Synthetic data was generated by following the noisy OR gate process. First, somatic mutations and the two relationship matrices were generated with Bernoulli distribution. Then $Path$ and DEG were generated by performing noisy OR-gate computation with $P_0 = 0$. To simulate the mutual exclusivity patterns observed in real data, we performed post pruning. That is, when several mutations belonging to the same pathway took place in the same sample, all but one of them were removed.

The artificial neural network (NN) was used as a baseline to evaluate ORN's efficacy of inferring pathway activities. To ensure the neural network model is comparable with ORN, it had one hidden layer, and the activation function was sigmoid. In this way, the values of hidden neurons were also

within [0,1]. We ran NN and ORN on 20 synthetic datasets and computed their reconstruction error and the average cosine similarity of pathways. Reconstruction error was computed as:

$$\text{error} = \frac{|\widehat{\text{Expr}}_{sg} - \text{Expr}_{sg}|}{S \times G}$$

We further proposed Jaccard score, a more stringent criterion, to evaluate how ORN's performance changes in various settings. To compute Jaccard score, we first need to compute Jaccard similarity for U and Z respectively:

$$\text{sim}(U_{\cdot p}^*) = \max_{p' \leq P} [\text{Jaccard}(U_{\cdot p}^*, U_{\cdot p'}^*)]$$

$$\text{sim}(Z_{p \cdot}^*) = \max_{p' \leq P} [\text{Jaccard}(Z_{p \cdot}^*, Z_{p' \cdot}^*)]$$

where U^* and Z^* are the true relationship matrix in synthetic data. $\text{Jaccard}(A, B)$ is computed as:

$$\text{Jaccard}(A, B) = A \cdot B / (A + B - A \cdot B)$$

Then the Jaccard score for one dataset is:

$$\text{Jaccard score} = \sum_{p \leq P} \text{sim}(U_{\cdot p}^*) \text{sim}(Z_{p \cdot}^*) / P$$

When applied to real world datasets, we extracted pathway activities together with its upstream SGAs and downstream regulon from the two relationship matrices. For example, if $U_{mp} > 0.5$, then we concluded that mutation of gene m could disrupt pathway p. If $Z_{pg} > 0.1$, then we concluded that the disruption of pathway p could cause differential expression of gene g. In this way, the set of genes affecting the same pathway were grouped into an upstream module, while the set of genes regulated by the same pathway were grouped into a downstream module. Note that in real data analysis, the cutoff for elements in Z was the top 5% value among all genes in the pathway p.

Since there is no ground truth for real data analysis, we run Gene Ontology (GO) enrichment analysis on the downstream modules to characterize the functional impacts of the latent pathways and its SGA modules.

Public Data

We performed ORN on the high-throughput data of 514 lower grade glioma patients. It was downloaded from The Cancer Genome Atlas (TCGA). After preprocessing, 511 samples remained with 15501 DEGs and 670 SGAs. The number of latent pathways was set to 10.

To evaluate how different cancers may have perturbed the same pathways, we also applied ORN to the METABRIC dataset (20) downloaded from CBioPortal. It contained 2173 samples of somatic variant profiles and 1904 samples of RNA-seq. After preprocessing, we retained a binary event matrix

with 1092 genes and a binary DEG matrix with 24256 genes. Both matrices had 1866 samples. ORN was trained with 15 latent pathways.

RESULTS

ORN was effective in recovering OR-gate relationships

Synthetic data were generated according to the generative process described in Methods. The number of pathways was set to 5; The number of samples, SGAs, and DEGs were all set to 1000. This was referred to as the standard setting.

We proposed Jaccard score and reconstruction error to evaluate the performance. Jaccard score can measure the concordance between inferred relationship matrices and the ground truth. Details of calculation was described in Methods.

From the standard setting, each condition was changed separately to see how they affected the performance. As shown in Fig. 3, ORN has achieved almost perfect recovery (>99%) in the standard setting. However, ORN's performance dropped over 20% when the number of samples dropped from 1000 to 300, or the number of mutations increased from 1000 to 3000 mutations. Note that when the number of DEGs were reduced to 500, the performance remained the same.

Figure 3 In the standard setting, ORN has recovered latent pathways with almost perfect accuracy. When the number of samples decreased to 300 or the number of mutations increase to 3000, the median Jaccard score has decreased to 73% and 81% respectively. Reducing the number of DEGs did not affect the performance of ORN.

ORN provided more insights than neural network

We cannot identify similar algorithms that only used high-throughput data to infer pathway activities. However, we found that artificial neural networks (NN) with sigmoid activation function can also produce binary values in the hidden layer. In addition, both ORN and NN used backward propagation to optimize parameters. Thus, we designed a neural network architecture similar with ORN and used it as a baseline.

In the synthetic experiment, although NN converged to comparable reconstruction error as ORN, its accuracy was in pathway recovery was only around 50%. This showed that NN is less capable of capturing all the signals in the data.

Similar results were observed when we applied NN to the glioma dataset (described in the next section). The relationship matrix estimated with NN is much more redundant than ORN. GO enrichment analysis of the downstream modules (see Supp. Table 1) also showed that NN could only capture less than 5 major aspects with 10 hidden neurons, while ORN can cover different biological aspects of glioma with each latent pathway. As shown in Fig. 5, the relationship matrix learned by NN contains much redundancy, while each pathway in ORN regulated different sets of genes with few overlaps. This indicated that the biological mechanism from somatic mutations to transcriptomic profiles could be more accurately characterized by the OR-gate logic imposed by ORN rather than conventional non-linear relationships.

Figure 4 Boxplot on the left showed the distribution of prediction error of NN and ORN across 20 synthetic experiments. Boxplot on the right showed the distribution of cosine similarity between the inferred pathways and ground truth.

Figure 5 Heatmap representation of the relationship matrix between pathways and differential expression after row normalization. Relationship matrix generated by a neural network (NN) contains many redundant signals, while ORN automatically pushes for sparsity. Each latent pathway in ORN has uniquely caused a subset of genes to express differentially.

ORN detected pathways closely related to patient survival

After applying ORN to the lower grade glioma dataset. Pathway activities showed that pathway 6 and pathway 7 had significant impacts on patient survival (Fig 6). We performed Gene Ontology (GO) enrichment analysis on the top DEGs in these pathways (see supplement table 2). The downstream module of pathway 6 is mostly related to DNA processing activities. Cancer samples with this pathway dysregulated probably have compromised genomic instability (21), leading to worse survival. Its upstream module includes CDK13 (22), H3F3A (23), IDH1 (24), PTEN (25), SNRPE (26) that are closely related to DNA repair or DNA replication.

As for pathway 7, we found that PTEN, H3F3A, and POM121L12 were shared by the upstream modules in both pathways. However, the top 300 DEGs caused by the two pathways have no genes in common. GO enrichment analysis showed that downstream modules are related to neutrophil activities, Ras signal transduction, and viral genome replication. We conjectured that cancer samples with pathway 7 dysregulated exhibited viral infection and its immune response. Since virus infection can drive glioma formation (27), This subgroup of patients may be more likely to progress to malignancy and worse survival.

Figure 6 LGG patients with pathway 6 (A) and pathway 7 (B) dysregulated have worse overall survival. X-axis is in the unit of month; Y-axis represents the proportion of each subgroup. 102 patients' pathway 6 were dysregulated, 100 patients' pathway 7 were dysregulated. Both groups have 62 samples in common.

ORN characterized common mechanisms in cancer

As for the results in the METABRIC dataset, upstream modules of six latent pathways were almost identical and hence merged into one union (pathway 0, 1, 2, 7, 9, 14 in supplement table 3). This module contained well-known oncogenes such as TP53, PTEN, PIK3CA, and MAP3K1. The corresponding pathway was dysregulated across all samples (probability above 0.5). This pathway likely represented the common cause of breast cancer. GO enrichment analysis (see supplement table 3) showed that its corresponding downstream module mostly involved in mitosis such as G1/S transition of mitotic cell cycle phase transition (GO:0044772). Pathway 5 also exhibited similar downstream effects but had different SGAs. Most notably, MIR604 was in the upstream modules of pathways. Studies have shown that polymorphism of MIR604 was related to the development of hepatocellular carcinoma (28) and the metastasis of colorectal cancer (29). MIR604 was differentially expressed in breast cancer (30). Still, to our knowledge, the impact of MIR604 mutation in breast cancer has not been investigated.

In the glioma dataset, we found pathway 4 to be closely related to immune response. Downstream modules of pathway 4 were enriched for various immune response, including cytokine-mediated signalling and toll signalling. This subgroup of patients may not be responsive to immune therapy. Within this pathway was the mutation of interferon alpha 21 (IFNA21), which played an important role in inflammatory response and toll signalling. IFNs were also identified as major factors of patient response to various cancer therapies (31). Moreover, we found that PTK6 and SRMS within the same upstream module. The product of two genes work closely together as intracellular kinases (32) and promotes invasive prostate cancer (33). However, they are rarely studied in the context of glioma and immune response.

Like glioma, one particular pathway in breast cancer captured abnormal immune response in a subgroup of cancer samples. The downstream module in pathway 3 is related to immune response, including T cell activation (GO:0042110), regulation of immune response (GO:00507006),

inflammatory response (GO:0006954). The upstream module included CDC20, COLEC12, MED8, MPL, SOX5, and OTUD1. CDC20 was known to be related to T cell activation. COLEC12's protein product is associated with innate immunity (34). SOX5 was shown to be related to B cell proliferation (35). Another interesting gene was MED8. Studies showed that MED8 was important to regulate resistance against bacteria in plants (36). Meanwhile, MED8 was implicated in renal cell carcinoma. However, it is rarely investigated in the case of breast cancer and innate immunity. As for OTUD1, a recent study (37) has shown that its induction by RNA virus may inhibit innate immune response.

ORN detected pathway dysregulation specific to lower grade glioma

Although not related to patient survival, other pathways in glioma samples also captured different aspects of molecular characteristics. For example, pathway 0 is closely related to the biosynthesis of cholesterol, steroid, and alcohol, while cholesterol metabolism has recently been studied as a potential therapeutic target (38). In addition, downstream modules of pathway 0 contained differentially expressed genes enriched for central nervous system development. In the corresponding upstream module, we identified SZT2 (39) and TIAM1 (40) to be closely related to nervous system development. Other mutations, such as CPAMD8 and RUBP1, exhibited mutual exclusivity and similar expression patterns. Yet these two genes have not been studied in terms of central nervous system development.

As for the breast cancer samples, the upstream module of pathway 8 contained several well-known driver mutations, including KRAS, APC, and ARID1A. As shown in Fig 7, most genes in this upstream module exhibited mutual exclusivity, while these mutations caused differential expression of a similar set of genes. The downstream module was enriched for telomere and t-circle formation, which were well-known factors for cancer initiation and tumour survival (41). In the upstream module, the relation between telomere and APC (42), KRAS (43), ARID1A (44), PRKG1 (45) were reported. Although mutually exclusive to the four genes above, we have not found research linking BAP1, MIR604, and MICAL3 to telomere activities.

Figure 7 Illustration of pathway 8 in breast cancer samples. Cancer samples sorted by pathway activities (middle figure) was the X-axis shared by the three subplots. The upper figure showed the mutation event of the upstream module (cutoff 0.5), while the bottom figure showed the heatmap of differential expression of the downstream module (roughly top 1% related). The top figure showed patterns of mutual exclusivity, while the bottom showed a strong correlation between pathway activities and differential expression.

DISCUSSION

In this study, we proposed ORN, the first de novo method to infer latent pathway activities from genomic profiles and transcriptomic profiles of cancer patients. Compared with the traditional neural network method, ORN provided much more insights into how the somatic mutations function

together. It also implied that the OR-gate logic assumed in ORN is suitable to the mechanism from mutations to differential expressions.

In the real data analysis with lower grade glioma and breast cancer, ORN has recovered major mechanisms consistent with current knowledge, such as abnormal DNA repair ability and immune response. Glioma patients with these dysregulated pathways had lower survival rates. ORN further revealed mechanisms specific to cancer types, such as the steroid metabolism and nervous system development in glioma. We identified several somatic mutations that might be related to certain malfunctions in cancer cells, worthy of further biological investigation. However, ORN requires in-depth analysis to obtain useful insights. In the future, we will try to develop statistical tests to automatically return meaningful genes in both upstream and downstream modules.

For the METABRIC dataset, none of the latent pathways generated by ORN is significantly related to patient survival. We conjectured that the reason was two-fold: (1) Somatic mutations may not be the only source of variance of RNA expression. Sharma, et al showed that copy number alterations, epigenetic changes, transcription factors, and microRNAs collectively explain, on average, only 31–38% and 18–26% expression variation (46); (2) compared with glioma samples, cellular constitution in breast cancer samples was probably more diverse. To handle the first issue, we need to include more data sources in a principled way. In the future, epigenetic profiles may be included to inform the coregulation of RNA expression. To deal with the second issue, future research needs to incorporate reliable complete deconvolution algorithms in the data preprocessing step.

For some genes in the upstream module, we failed to find evidence supporting their associations to the downstream module. Although some of them were likely to provide novel molecular insights, many were false positive. Upon closer investigation, we believed there are two major sources of false positives: (1) passenger mutations that exclusively occur in highly mutated samples. For example, ACSS1 was in most upstream modules of glioma because it only occurred in highly mutated samples, which had most pathways dysregulated. (2) passenger mutations within the same copy number variation event as driver mutations. When a set of genes mutated in almost the same set of patients, it is likely that only one of them contributed to the pathway dysregulation.

During analysis, we also found that different upstream modules may share a subset of somatic mutations. For example, pathway 6 and pathway 7 in glioma have PTEN in common. Thus, it is possible that hierarchical structures of SGA functions can be inferred from the overlapping upstream modules. In the future, we may provide more convenient visualization utilities to analyse the hierarchies among upstream modules.

We proposed ORN to infer latent pathway activities from high-throughput profiles of cancer samples. Application of ORN in lower grade glioma and breast cancer detected latent pathways closely related to patient survival. ORN also connected somatic mutations to key mechanisms of cancer, such as DNA repair and innate immune response. Although some mutations' function (e.g., MIR604) was not supported by literature, they were mutually exclusive to well-known driver mutations and caused differential expression in a similar subset of genes. We encouraged biological researchers to use ORN to infer personalized pathway activities and generate novel hypotheses for targeted therapy.

DATA AVAILABILITY

Processed data from METABRIC can be downloaded at:

https://www.cbioportal.org/study/summary?id=brca_metabric

Processed data of TCGA-LLG can be downloaded at:

[https://xenabrowser.net/datapages/?cohort=TCGA%20Lower%20Grade%20Glioma%20\(LGG\)](https://xenabrowser.net/datapages/?cohort=TCGA%20Lower%20Grade%20Glioma%20(LGG))

Python implementation of ORN is available at: <https://github.com/LifanLiang/ORN>

FUNDING

This work was partially supported by the National Institutes of Health [R00LM011673, U54HG008540]. The project used the Hillman Cancer Bioinformatics Services and the UPMC Hillman Cancer Center Developmental Funding that are supported in part by the National Cancer Institute [P30CA047904].

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

1. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
2. Seebacher,N.A., Stacy,A.E., Porter,G.M. and Merlot,A.M. (2019) Clinical development of targeted and immune based anti-cancer therapies. *J Exp Clin Cancer Res*, **38**, 156.
3. Goutsouliak,K., Veeraraghavan,J., Sethunath,V., De Angelis,C., Osborne,C.K., Rimawi,M.F. and Schiff,R. (2020) Towards personalized treatment for early stage HER2-positive breast cancer. *Nat. Rev. Clin. Oncol.*, **17**, 233–250.
4. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–45.
5. Hua,X., Hyland,P.L., Huang,J., Song,L., Zhu,B., Caporaso,N.E., Landi,M.T., Chatterjee,N. and Shi,J. (2016) MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. *Am. J. Hum. Genet.*, **98**, 442–455.
6. Haynes,W.A., Tomczak,A. and Khatri,P. (2018) Gene annotation bias impedes biomedical research. *Sci. Rep.*, **8**, 1362.
7. Mubeen,S., Hoyt,C.T., Gemünd,A., Hofmann-Apitius,M., Fröhlich,H. and Domingo-Fernández,D. (2019) The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front. Genet.*, **10**, 1203.
8. Gerstung,M., Pellagatti,A., Malcovati,L., Giagounidis,A., Porta,M.G.D., Jädersten,M., Dolatshad,H., Verma,A., Cross,N.C.P., Vyas,P., *et al.* (2015) Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.*, **6**, 5901.

9. Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
10. Yang,Z. and Michailidis,G. (2016) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, **32**, 1–8.
11. Li,Y., Wu,F.-X. and Ngom,A. (2018) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinformatics*, **19**, 325–340.
12. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
13. Szczurek,E. and Beerenwinkel,N. (2014) Modeling mutual exclusivity of cancer mutations. *PLoS Comput. Biol.*, **10**, e1003503.
14. Deng,Y., Luo,S., Deng,C., Luo,T., Yin,W., Zhang,H., Zhang,Y., Zhang,X., Lan,Y., Ping,Y., *et al.* (2019) Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Brief. Bioinformatics*, **20**, 254–266.
15. Leiserson,M.D.M., Reyna,M.A. and Raphael,B.J. (2016) A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics*, **32**, i736–i745.
16. Xie,J., Ma,A., Fennell,A., Ma,Q. and Zhao,J. (2019) It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Brief. Bioinformatics*, **20**, 1449–1464.
17. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.
18. Oniško,A., Druzdzal,M.J. and Wasyluk,H. (2001) Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. *International Journal of Approximate Reasoning*, **27**, 165–182.
19. Blei,D.M., Ng,A.Y. and Jordan,M.I. (2003) Latent dirichlet allocation. *Journal of machine Learning research*.
20. Pereira,B., Chin,S.-F., Rueda,O.M., Volland,H.-K.M., Provenzano,E., Bardwell,H.A., Pugh,M., Jones,L., Russell,R., Sammut,S.-J., *et al.* (2016) The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.*, **7**, 11479.
21. Negrini,S., Gorgoulis,V.G. and Halazonetis,T.D. (2010) Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.*, **11**, 220–228.
22. Blazek,D., Kohoutek,J., Bartholomeeusen,K., Johansen,E., Hulinkova,P., Luo,Z., Cimermancic,P., Ule,J. and Peterlin,B.M. (2011) The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.*, **25**, 2158–2172.
23. Tagami,H., Ray-Gallet,D., Almouzni,G. and Nakatani,Y. (2004) Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell*, **116**, 51–61.

24. Wu,F., Chai,R.-C., Wang,Z., Liu,Y.-Q., Zhao,Z., Li,G.-Z. and Jiang,H.-Y. (2019) Molecular classification of IDH-mutant glioblastomas based on gene expression profiles. *Carcinogenesis*, **40**, 853–860.
25. Ho,J., Cruise,E.S., Dowling,R.J.O. and Stambolic,V. (2019) PTEN Nuclear Functions. *Cold Spring Harb. Perspect. Med.*, 10.1101/cshperspect.a036079.
26. Li,Z. and Pützer,B.M. (2008) Spliceosomal protein E regulates neoplastic cell growth by modulating expression of cyclin E/CDK2 and G2/M checkpoint proteins. *J. Cell Mol. Med.*, **12**, 2427–2438.
27. McFaline-Figueroa,J.R. and Wen,P.Y. (2017) The viral connection to glioblastoma. *Curr. Infect. Dis. Rep.*, **19**, 5.
28. Cheong,J.Y., Shin,H.D., Cho,S.W. and Kim,Y.J. (2014) Association of polymorphism in microRNA 604 with susceptibility to persistent hepatitis B virus infection and development of hepatocellular carcinoma. *J. Korean Med. Sci.*, **29**, 1523–1527.
29. Boni,V., Zarate,R., Villa,J.C., Bandrés,E., Gomez,M.A., Maiello,E., Garcia-Foncillas,J. and Aranda,E. (2011) Role of primary miRNA polymorphic variants in metastatic colon cancer patients treated with 5-fluorouracil and irinotecan. *Pharmacogenomics J*, **11**, 429–436.
30. Zhang,S., Zhang,X., Fu,X., Li,W., Xing,S. and Yang,Y. (2018) Identification of common differentially-expressed miRNAs in ovarian cancer cells and their exosomes compared with normal ovarian surface epithelial cell cells. *Oncol. Lett.*, **16**, 2391–2401.
31. Budhwani,M., Mazziere,R. and Dolcetti,R. (2018) Plasticity of Type I Interferon-Mediated Responses in Cancer Therapy: From Anti-tumor Immunity to Resistance. *Front. Oncol.*, **8**, 322.
32. Serfas,M.S. and Tyner,A.L. (2003) Brk, Srm, Frk, and Src42A form a distinct family of intracellular Src-like tyrosine kinases. *Oncol. Res.*, **13**, 409–419.
33. Wozniak,D.J., Kajdacsy-Balla,A., Macias,V., Ball-Kell,S., Zenner,M.L., Bie,W. and Tyner,A.L. (2017) PTEN is a protein phosphatase that targets active PTK6 and inhibits PTK6 oncogenic signaling in prostate cancer. *Nat. Commun.*, **8**, 1508.
34. Ma,Y.J., Hein,E., Munthe-Fog,L., Skjoedt,M.-O., Bayarri-Olmos,R., Romani,L. and Garred,P. (2015) Soluble Collectin-12 (CL-12) Is a Pattern Recognition Molecule Initiating Complement Activation via the Alternative Pathway. *J. Immunol.*, **195**, 3365–3373.
35. Rakhmanov,M., Sic,H., Kienzler,A.-K., Fischer,B., Rizzi,M., Seidl,M., Melkaoui,K., Unger,S., Moehle,L., Schmit,N.E., *et al.* (2014) High levels of SOX5 decrease proliferative capacity of human B cells, but permit plasmablast differentiation. *PLoS One*, **9**, e100328.
36. An,C. and Mou,Z. (2013) The function of the Mediator complex in plant immunity. *Plant Signal. Behav.*, **8**, e23182.
37. Zhang,L., Liu,J., Qian,L., Feng,Q., Wang,X., Yuan,Y., Zuo,Y., Cheng,Q., Miao,Y., Guo,T., *et al.* (2018) Induction of OTUD1 by RNA viruses potentially inhibits innate immune responses by promoting degradation of the MAVS/TRAF3/TRAF6 signalosome. *PLoS Pathog.*, **14**, e1007067.
38. Pirmoradi,L., Seyfizadeh,N., Ghavami,S., Zeki,A.A. and Shojaei,S. (2019) Targeting cholesterol metabolism in glioblastoma: a new therapeutic approach in cancer therapy. *J. Investig. Med.*, **67**, 715–719.

39. Tsuchida,N., Nakashima,M., Miyauchi,A., Yoshitomi,S., Kimizu,T., Ganesan,V., Teik,K.W., Ch'ng,G.S., Kato,M., Mizuguchi,T., *et al.* (2018) Novel biallelic SZT2 mutations in 3 cases of early-onset epileptic encephalopathy. *Clin. Genet.*, **93**, 266–274.
40. Miller,M.B., Yan,Y., Eipper,B.A. and Mains,R.E. (2013) Neuronal Rho GEFs in synaptic physiology and behavior. *Neuroscientist*, **19**, 255–273.
41. Jafri,M.A., Ansari,S.A., Alqahtani,M.H. and Shay,J.W. (2016) Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Med.*, **8**, 69.
42. Deng,Y., Chan,S.S. and Chang,S. (2008) Telomere dysfunction and tumour suppression: the senescence connection. *Nat. Rev. Cancer*, **8**, 450–458.
43. Liu,W., Yin,Y., Wang,J., Shi,B., Zhang,L., Qian,D., Li,C., Zhang,H., Wang,S., Zhu,J., *et al.* (2017) Kras mutations increase telomerase activity and targeting telomerase is a promising therapeutic strategy for Kras-mutant NSCLC. *Oncotarget*, **8**, 179–190.
44. Zhao,B., Lin,J., Rong,L., Wu,S., Deng,Z., Fatkhutdinov,N., Zundell,J., Fukumoto,T., Liu,Q., Kossenkova,A., *et al.* (2019) ARID1A promotes genomic stability through protecting telomere cohesion. *Nat. Commun.*, **10**, 4067.
45. Lee,J.H., Cheng,R., Honig,L.S., Feitosa,M., Kammerer,C.M., Kang,M.S., Schupf,N., Lin,S.J., Sanders,J.L., Bae,H., *et al.* (2013) Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: the Long Life Family Study. *Front. Genet.*, **4**, 310.
46. Sharma,A., Jiang,C. and De,S. (2018) Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations. *Nucleic Acids Res.*, **46**, 4370–4381.