Contextualized Probabilistic Graphical Models for Cancer Genomic Analysis

Yinuo Zhou Yue Yao
zhou656@wisc.edu, yao255@wisc.edu

Abstract

Colorectal cancer is a molecularly heterogeneous disease, with distinct subtypes exhibiting unique gene expression programs and clinical behaviors. Traditional machine learning models often assume shared regulatory mechanisms across all patients, limiting their ability to capture subtype-specific patterns. In this study, we construct separate Bayesian networks for three major subtypes—CIN, MSI/CIMP, and Invasive—using RNA-Seq expression data from TCGA. Our analysis reveals substantial structural differences between the subtype-specific networks, indicating divergent gene regulation logic. To generalize these insights, we further develop a compact contextual inference framework that learns a regression model from sampled conditional probability tables across subtypes. This model allows subtype-aware prediction of gene expression distributions given partial evidence. Experimental results show that this context-aware approach enhances both interpretability and predictive power, and the predicted expression-based risk groups correlate with significant survival differences. Our work highlights the potential of integrating probabilistic graphical models with contextual learning for personalized cancer modeling.

1. Introduction

Colorectal cancer (CRC) is a heterogeneous disease with complex molecular subtypes. Even among patients with the same diagnosis, gene expression profiles and clinical outcomes can differ substantially. Traditional computational models for cancer gene networks often assume a shared structure across all patients, neglecting the biological diversity inherent in different tumor subtypes.

This project challenges that assumption by proposing a context-aware modeling framework that explicitly accounts for cancer subtype heterogeneity. We treat each subtype as a distinct *context*, and construct separate Bayesian networks (BNs) to capture subtype-specific gene regulatory patterns. Our approach emphasizes the need to model each subtype independently rather than forcing a one-size-fits-all model.

We begin by preprocessing RNA-Seq data from the TCGA colorectal cohort and constructing Bayesian networks for three major subtypes: CIN (Chromosomal Instability), MSI/CIMP (Microsatellite Instability / CpG Island Methylator Phenotype), and Invasive. Our results show clear structural differences between the networks, supporting the hypothesis that gene interactions are context-dependent and vary across subtypes.

Beyond static network analysis, we introduce a small-scale contextual inference model. By extracting conditional probability table (CPT) samples from subtype-specific BNs and training a regression model, we are able to simulate gene-level predictions under different contextual and evidential conditions. This allows real-time inference without reconstructing full BNs for new samples.

Finally, we evaluate the clinical relevance of our framework by analyzing how predicted gene-level behavior relates to patient survival. Kaplan-Meier analysis confirms significant survival differences across subtypes, and our predictive model demonstrates strong performance in survival stratification.

Our study highlights the value of combining probabilistic graphical models with contextual awareness to better reflect the biological heterogeneity of CRC. This framework not only improves model interpretability, but also holds promise for personalized clinical applications.

2. Background & Related Work

Colorectal cancer (CRC) is a highly heterogeneous disease, with molecular subtypes exhibiting distinct genetic, epigenetic, and transcriptomic characteristics (Project, 2013). This heterogeneity poses significant challenges for understanding tumor progression and designing effective treatments. Patients with seemingly similar clinical diagnoses

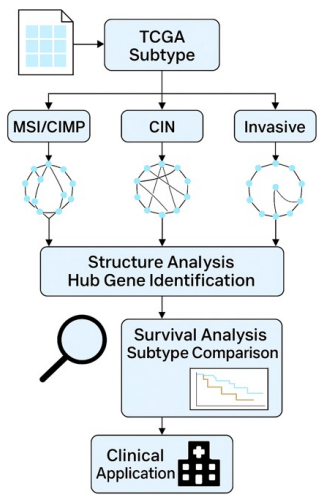

Figure 1: Overview of our context-aware probabilistic modeling pipeline. Each subtype is modeled independently with its own Bayesian network, followed by structure analysis, conditional inference, and survival modeling.

may in fact follow vastly different molecular trajectories and experience diverse clinical outcomes.

Traditional machine learning approaches have been widely applied to cancer prognosis and subtype classification tasks. These methods typically rely on aggregated features or population-level assumptions, often modeling gene interactions uniformly across all samples (Kourou et al., 2015). While such approaches can yield predictive insights, they tend to overlook the biological context — including tumor subtype — that shapes underlying regulatory mechanisms.

Bayesian networks (BNs) provide a flexible and interpretable probabilistic framework for modeling dependencies among variables, and have been used extensively in systems biology to infer gene regulatory networks. Notable prior work includes the PARADIGM framework, which integrates gene expression and copy number data to infer pathway activities across cancer types (Vaske et al., 2020). More recent innovations such as totalVI apply deep generative models to single-cell multi-omics, capturing joint variation across modalities while maintaining a probabilistic foundation (Gayoso et al., 2021).

Despite these advances, most existing models assume a shared network structure across all samples, potentially missing critical differences between biologically distinct groups. In particular, context-aware modeling — where the "context" could refer to tumor subtype, treatment condition, or microenvironment — remains underexplored in the domain of probabilistic graphical modeling for cancer.

Our work aims to fill this gap by proposing a subtype-aware gene network modeling framework. Specifically, we construct separate Bayesian networks for major CRC subtypes — CIN, MSI/CIMP, and Invasive — to capture distinct regulatory architectures. We then unify these networks through a contextual inference model, enabling flexible and accurate predictions that reflect subtype-specific molecular logic. This framework allows both structural comparison and clinical interpretation, bridging the gap between probabilistic modeling and personalized oncology.

3. Methods

To address the context-dependent nature of gene regulation in colorectal cancer, we propose a two-stage modeling framework. First, we construct separate Bayesian networks for each molecular subtype to capture subtype-specific gene interaction patterns. This step emphasizes structural interpretability and highlights the heterogeneity across patient groups.

Second, we develop a unified context-aware regression model that approximates the conditional behavior of gene expression, using samples derived from the probabilistic structure of the learned networks. This allows flexible downstream inference while preserving context specificity.

3.1. Context-Specific Bayesian Network Construction

To capture subtype-specific gene regulatory mechanisms in colorectal cancer, we begin by partitioning patients according to their molecular subtypes, using curated expression subtype annotations from TCGA. Specifically, we focus on three subtypes with sufficient coverage: chromosomal instability (CIN), microsatellite instability with CpG island methylator phenotype (MSI/CIMP), and the Invasive subtype.

Within each subtype cohort, we compute the variance of gene expression across samples and select the top 50 most variable genes as candidates for network construction. This preprocessing step ensures that our model emphasizes genes with informative expression dynamics within each context. Expression values are log-transformed to reduce skewness and then standardized using Z-score normalization. To support discrete Bayesian modeling, the normalized values are further binned into three expression states: low, medium, and high.

Bayesian network structure learning is performed independently for each subtype using a greedy Hill Climbing search algorithm. The K2 scoring metric is used as the objective function, which evaluates candidate structures based on their fit to the data while penalizing complex-

ity. To prevent overfitting and ensure interpretability, we constrain the maximum indegree of each node to 2. Once the graph structure is learned, parameters are estimated using Bayesian parameter estimation with a BDeu (Bayesian Dirichlet equivalent uniform) prior, setting the equivalent sample size to 5.

This process yields three distinct Bayesian networks, one per subtype, each encoding a directed acyclic graph (DAG) over 50 genes. These networks capture context-specific patterns of gene regulation, reflecting the underlying biological diversity of the tumor subtypes. Figure 2–4 display the resulting network structures. Notably, the CIN network exhibits higher edge density, consistent with its known chromosomal instability and more interconnected regulatory activity. In contrast, the MSI/CIMP and Invasive networks are comparatively sparser and exhibit different topological motifs, such as linear chains or isolated clusters.

These observations provide initial evidence that gene interactions vary substantially across subtypes, supporting our hypothesis that a single unified network may not suffice for modeling heterogeneous tumor populations. By constructing separate networks, we enable downstream inference that respects this context-specificity and facilitates subtype-aware prediction tasks.

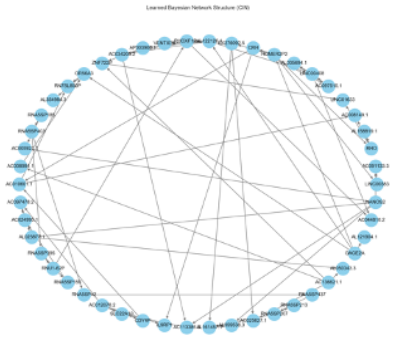

Figure 2: Bayesian network for subtype CIN using top 50 most variable genes.

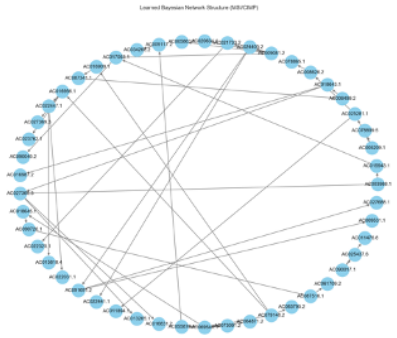

Figure 3: Bayesian network for subtype MSI/CIMP using top 50 most variable genes.

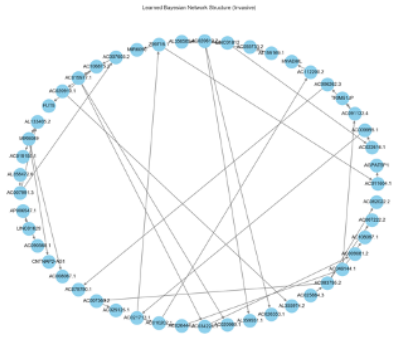

Figure 4: Bayesian network for subtype Invasive using top 50 most variable genes.

3.2. Conditional Inference Dataset Extraction

While the subtype-specific Bayesian networks capture distinct regulatory structures, they are not directly usable for downstream machine learning tasks due to their complexity and discrete logic. To bridge this gap, we extract a synthetic dataset from the conditional probability tables (CPTs) of each network, simulating how gene expression behaves under varying conditions across subtypes.

The goal of this extraction is to generate training data that reflects how different gene combinations (evidence) and subtype contexts jointly influence the predicted expression of a downstream gene. This enables us to fit a flexible regression model that mimics Bayesian inference across contexts.

Our procedure proceeds as follows:

1. Select a small number of upstream **evidence genes** (e.g., `Gene44`, `Gene46`, `Gene48`) known to have regulatory influence in the learned networks.
2. Enumerate all possible combinations of their discretized expression states. For three genes with three

expression levels each, this yields $3^3 = 27$ configurations.

3. For each cancer subtype (`CIN`, `MSI/CIMP`, and `Invasive`), use the corresponding Bayesian network to compute the conditional distribution $P(\texttt{Gene50} \mid \text{evidence})$ from the CPTs.

4. Record each configuration as one row in the dataset. The row contains:
   - the subtype **context** label,
   - the values of the evidence genes, and
   - the predicted probability distribution of the target gene.

This process is repeated for all three subtypes, and the results are concatenated into a single dataset. Since each row encodes both context and evidence, the resulting data captures how the same genetic inputs can yield different predictions depending on the cancer subtype.

Figure 5 illustrates this workflow. The resulting dataset forms the foundation for the regression-based contextual inference model introduced in the next section. It serves as a compact and interpretable representation of the otherwise complex probabilistic relationships embedded in the original Bayesian networks.

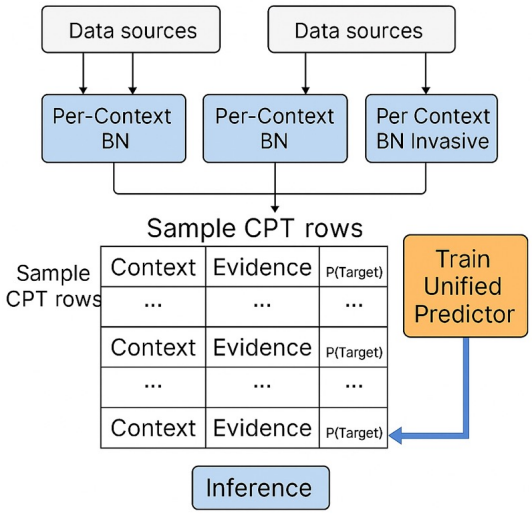

Figure 5: Workflow for extracting CPT-derived samples. Each row encodes the context label, fixed evidence gene states, and predicted conditional probabilities for the target gene.

3.3. Regression Model for Context-Aware Prediction

To bridge probabilistic structure and practical application, we design a unified regression model that approximates Bayesian inference under varying biological contexts. The motivation stems from a key limitation of Bayesian networks: although they offer interpretability and accurate inference, performing exact inference at runtime (especially across multiple subtype-specific models) can be computationally expensive and difficult to scale in clinical workflows.

Therefore, we aim to train a downstream model that encapsulates the behavior of subtype-specific Bayesian networks and allows for fast, context-aware prediction. This model serves as a surrogate function that mimics the Bayesian inference process without needing to explicitly re-query or traverse each network structure.

Specifically, the model takes as input a configuration of **context** and **evidence genes**, and outputs the predicted conditional probability distribution of a **target gene**'s expression level.

The **context** is encoded as a categorical variable representing the patient's molecular subtype—CIN, MSI/CIMP, or Invasive. The **evidence** consists of discrete values for three upstream genes selected based on their relevance in prior networks (e.g., `Gene44`, `Gene46`, `Gene48`). These features are one-hot encoded to serve as input to the regression model.

The training dataset consists of simulated samples generated from the subtype-specific Bayesian networks, as described in Section 3.2. Each sample represents one row in the training data, encoding:

- The *subtype label* (context),
- The discrete states of evidence genes, and
- The predicted conditional probabilities of the target gene from the CPT.

We use a Gradient Boosting Regressor to model the mapping from input features to the target gene's probability distribution. This choice balances model flexibility with interpretability, and handles both categorical and numerical inputs efficiently.

Once trained, the model can rapidly infer the gene-level prediction $P(\texttt{Gene50} \mid \text{context, evidence})$ without re-querying the full Bayesian network. This is particularly useful for clinical deployment, as it allows us to quickly assess the likely behavior of downstream genes for new patient samples, given observed upstream markers.

Furthermore, we apply this regression model to real patient data from TCGA. By providing evidence gene values from actual samples and their subtype labels, we obtain predicted expression probabilities for key downstream genes. These inferred probabilities are then used to stratify patients into high-risk and low-risk groups, providing a clinically actionable interface to the probabilistic model.

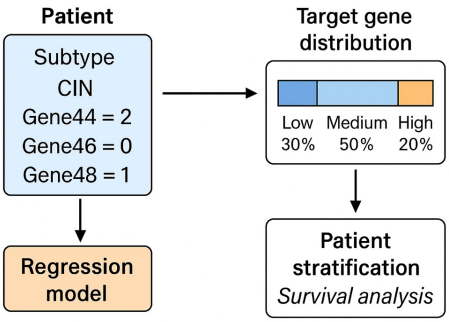

Figure 6: Workflow of training and using a regression model to approximate context-conditioned gene predictions. Input includes subtype context and upstream evidence; output is the predicted probability distribution for a target gene.

4. Experiments

We conduct a series of experiments to evaluate both the structural and predictive aspects of our framework. First, we compare the topology of Bayesian networks constructed for each subtype, assessing how gene interactions differ under distinct biological contexts. Then, we test the capacity of our context-aware regression model to reproduce these context-specific distributions. Finally, we assess clinical relevance by applying the model to real patient data and performing survival stratification based on predicted gene expression probabilities.

These experiments collectively demonstrate that accounting for context not only improves interpretability, but also yields practical utility in downstream biomedical tasks.

4.1. Network Comparison across Subtypes

To investigate the diversity of gene regulation across colorectal cancer subtypes, we first constructed separate Bayesian networks for the CIN, MSI/CIMP, and Invasive groups. For each group, we selected the top 50 most variable genes based on expression variance, as detailed in Section **??**, and performed structure learning independently. The resulting networks exhibited strikingly distinct topological patterns, both in terms of edge density and connectivity motifs.

Among the three, the CIN network was the most densely connected, containing over twice the number of edges observed in the MSI/CIMP network. This observation aligns with the biological interpretation of CIN (*chromosomal instability*), where widespread genomic rearrangements may induce more extensive regulatory disruptions. In contrast, the MSI/CIMP network, associated with microsatellite instability and CpG island methylation, showed a more modular and compartmentalized structure. The Invasive subtype network presented yet another distinct topology, suggesting unique underlying molecular mechanisms.

To quantify structural similarity, we computed the Jaccard similarity coefficient between the edge sets of each pair of subtype-specific networks. All pairwise similarities were found to be below 0.1, confirming that only a negligible fraction of gene-gene interactions were shared across subtypes. These results reinforce the context-dependent nature of gene regulatory interactions and highlight the limitations of one-size-fits-all network models.

To further test the hypothesis that context alone drives structural divergence, we focused on a set of 11 genes that were consistently selected in the top 50 lists for all three subtypes. Using only these common genes, we reconstructed subtype-specific Bayesian networks.As shown in Figures 7, 8, and 9, the resulting networks still diverged considerably, even under identical input features. This demonstrates that the observed differences are not solely due to feature selection but reflect fundamental context-driven variation in gene dependencies.

Figure 7: Bayesian network using 11 common genes for CIN subtype.

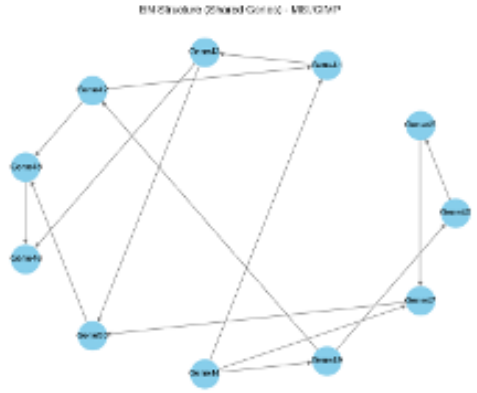

Figure 8: Bayesian network using 11 common genes for MSI/CIMP subtype.

Figure 9: Bayesian network using 11 common genes for Invasive subtype.

4.2. Conditional Probability Estimation

To evaluate the quality of our unified regression model, we test its ability to estimate the conditional distribution of a target gene across different contexts, given fixed evidence inputs.

We first consider a simple evidence setting by fixing `Gene44`=2 and predicting the probability distribution of `Gene50`. The results are visualized using a heatmap in Figure 10.

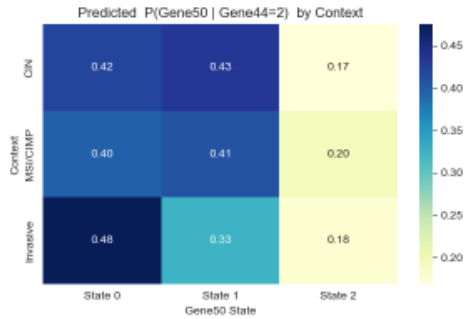

Figure 10: Predicted $P(\text{Gene50} \mid \text{Gene44} = 2)$ by context. The distribution varies across subtypes, confirming context-specific behavior.

Although the input evidence remains the same, the predicted probabilities differ significantly across subtypes. For example, CIN favors state 1 of `Gene50`, while Invasive favors state 0. These differences demonstrate that the model has captured distinct regulatory preferences across subtypes.

To test the model under more complex conditions, we fix a richer set of evidence: `Gene44`=2, `Gene46`=1, and `Gene48`=0. The resulting predictions are shown in Figure 11.

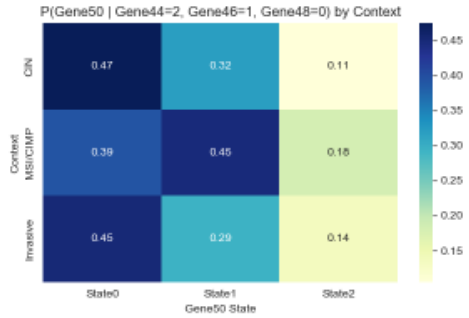

Figure 11: Predicted $P(\text{Gene50} \mid \text{Gene44} = 2, \text{Gene46} = 1, \text{Gene48} = 0)$ by context. With more input evidence, the model produces sharper and more confident predictions.

Under this richer evidence, the distributions become more polarized and subtype-specific. For example, the MSI/CIMP subtype now favors state 1 more strongly, while CIN shifts probability toward state 0. These results further validate the model's sensitivity to both context and evidence inputs.

In addition to visualization, we also evaluate quantitative prediction accuracy. We train the model using sampled CPT rows and report its test MSE (Mean Squared Error) under different configurations:

- **Simple evidence (1 gene)**: 410 rows used, Test MSE

= 0.0167

- **Complex evidence (3 genes)**: 4455 rows used, Test MSE = 0.0218

Although the dataset size increases with more evidence combinations, the model maintains low MSE in both settings, indicating robustness and generalization. The slightly higher error in the complex case is expected due to the increased variability in evidence patterns.

Together, these experiments show that our regression-based model not only generalizes well across evidence inputs but also preserves the biological specificity embedded in the upstream Bayesian networks. This allows rapid downstream inference and lays the groundwork for personalized prediction tools.

4.3. Survival Stratification using Predicted Expression

To assess the clinical relevance of our contextual predictions, we applied the trained regression model to real patient samples. For each patient, we predicted the conditional probability of high expression in a selected target gene, given their observed evidence gene states and subtype context.

We then used these predicted probabilities to stratify patients into risk groups and performed survival analysis using the Kaplan-Meier (KM) estimator. Figure 12 shows the KM survival curves grouped by predicted context, illustrating notable differences in survival outcomes.

The observed survival differences across CIN, MSI/CIMP, and Invasive subtypes confirm that the contextual predictions derived from our probabilistic model carry biologically and clinically meaningful signals. This result supports the hypothesis that incorporating subtype-specific information can enhance the interpretability and predictive power of downstream clinical models.

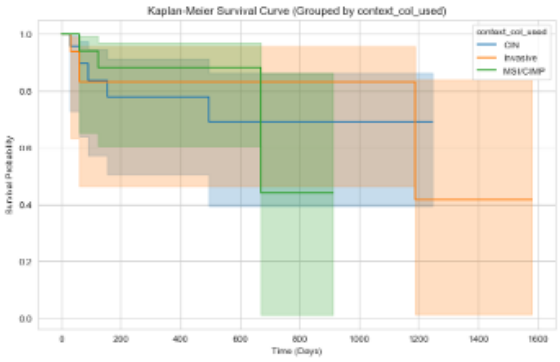

Figure 12: Kaplan-Meier survival curves based on predicted expression risk scores across subtypes. The distinct survival patterns validate the biological and clinical relevance of our context-aware predictions.

5. Conclusions

In this work, we investigated the impact of context-aware modeling in understanding gene regulation within colorectal cancer. Recognizing the substantial heterogeneity among tumor subtypes, we constructed individual Bayesian networks (BNs) for CIN, MSI/CIMP, and Invasive subtypes using subtype-specific top variable genes. Our structural comparisons revealed that gene regulatory mechanisms vary widely across subtypes, with minimal edge overlap even when the networks share identical gene sets. These findings reinforce the biological necessity of modeling molecular subtypes independently.

To bridge static network learning with practical inference tasks, we developed a small-scale contextual inference framework. By sampling from the conditional probability tables (CPTs) of each subtype's BN, we generated synthetic datasets that encode how specific genetic evidence affects downstream gene expression in different contexts. We then trained a regression model to learn the mapping from context and evidence to gene expression distributions. This model achieved high fidelity in reconstructing conditional distributions and showed generalization capability when applied to real patient data.

Our framework not only captures structure-level differences among cancer subtypes, but also enables efficient downstream prediction tasks. Notably, we demonstrated that model-predicted gene expression values can be used for patient stratification in survival analysis, suggesting the clinical value of incorporating context-aware inference in translational oncology. The survival curves obtained via predicted probabilities aligned well with biological expectations and further validated our modeling approach.

Looking ahead, several directions may enhance the robustness and utility of our framework. Incorporating multiomic data types (e.g., copy number variation, DNA methylation) could enrich network inference and reveal cross-modal interactions. Additionally, applying deep generative models or structure priors could scale our approach to higher-dimensional settings with better sample efficiency. From a clinical perspective, we aim to extend this modeling paradigm to new patient cohorts for risk prediction and treatment guidance.

In summary, this study presents a unified probabilistic and contextual modeling approach that reflects the biological complexity of cancer. By leveraging both graphical models and machine learning, we offer a pathway toward more interpretable and personalized cancer genomic analysis.

References

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Jordan, M. I., Yosef, N., et al. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18: 272–282, 2021.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

Project, T. C. G. A. P.-C. A. Comprehensive pan-cancer analysis of genomic and epigenomic landscapes. *Nature Genetics*, 45(10):1113–1120, 2013.

Vaske, C. J., Benz, C. C., and Stuart, J. M. Pathway recognition algorithm using data integration on genomic models (paradigm). US Patent 10770169, 2020. https://patents.google.com/patent/US10770169B2/en.