Learning the Relationship Between Variants, Metabolic Fluxes and Phenotypes

Dongin Kim[*,†,‡] **Seokjin Han**[*,‡] **Seong-Hyeuk Nam**[*] **Tae Yong Kim**[*]

Abstract

In the dynamic field of cancer research, the fusion of genetics and machine learning presents a groundbreaking opportunity to understand the intricate links between genomic variations and cancer phenotypes. Despite the wealth of genetic data, translating it into actionable insights remains a challenge, particularly in the context of complex cellular metabolism. VaMP (**Va**riants, **M**etabolic Fluxes & **P**henotypes), our novel approach, addresses this by integrating neural networks and metabolic models, offering a comprehensive framework for deciphering cancer biology. The developed method leverages a novel end-to-end neural network architecture, integrating genome-scale metabolic models (GSMs) to capture the intricate relationship between genetic variations and cellular phenotypes. VaMP's encoder maps genetic variants to metabolic fluxes through a series of carefully designed steps utilizing neural network and GSM, while the decoder predicts phenotype probabilities using the differences between input and reference fluxes. The training data comprise mutation information and phenotypes, eliminating the need for explicit metabolic flux data preparation. Validation experiments on five cancer types demonstrate VaMP's ability to identify significant genes and metabolic signatures. Further utilizing SCREENER and co-occurrence analyses, the assessment reveals VaMP's capacity to anticipate established gene-disease relationships. The identified metabolic signatures are robustly substantiated by diverse literature grounded in experimental studies. Furthermore, an in-depth exploration of the outcomes from five VaMP models involved the correlation analysis for variants and fluxes to establish connections between significant genes and cancer-causing variants. Overall, VaMP can be used as a promising tool for unraveling the complex interplay between genetic alterations and cancer phenotypes, with implications for understanding disease mechanisms and identifying novel therapeutic targets.

[*]Standigm Inc., Seoul, South Korea
[†]Corresponding author: dongin.kim@standigm.com
[‡]These authors contributed equally to this work.

1 Introduction

In the current landscape of cancer research, the convergence of genetic studies and machine learning presents an unprecedented opportunity to decode the intricate relationship between genomic variations and cancer phenotypes [1–3]. The genomic era has witnessed an exponential increase in the availability of genetic data, offering a comprehensive catalog of mutations associated with various cancers [4–6]. However, translating this wealth of information into actionable insights remains a formidable challenge.

Traditional methodologies for interpreting genetic variants often fall short in capturing the nu-

anced interplay of alterations within cellular systems, particularly in the context of metabolism [7, 8]. The dynamic and context-dependent nature of metabolic networks adds an additional layer of complexity to deciphering the consequences of genetic mutations. As machine learning continues to demonstrate its prowess in unraveling complex patterns in biological data [9, 10], there is a compelling need for specialized tools that bridge the gap between genetic variations and metabolic phenotypes.

As a response to this pressing need, we developed VaMP, leveraging cutting-edge neural network architectures and genome-scale metabolic models (GSMs) [11]. By providing a holistic framework that integrates genetic information with metabolic fluxes, VaMP aims to enhance our understanding of cancer biology. The urgency of this endeavor is underscored by the potential impact on precision medicine, where a deeper comprehension of the molecular underpinnings of cancer can pave the way for personalized therapeutic strategies.

Moreover, in the broader context of machine learning research, VaMP contributes to the evolving landscape of interpretable models in genomics. As black-box algorithms pose challenges in elucidating the rationale behind predictions [10], VaMP offers a transparent and interpretable approach, shedding light on the functional consequences of genetic variants within the realm of cellular metabolism.

2 VaMP: Variants, Metabolic Fluxes & Phenotypes

VaMP is an end-to-end neural network for predicting the phenotype from given patient's variants, where the latent space represents the metabolic flux calculated from the GSM (see Figure 1).

Note that the training data for VaMP is the mutation information and the phenotypes, so there is no need to prepare training data for the metabolic flux.

2.1 Encoder

VaMP maps variants to the flux via the following steps (see Figure 2):

i) For each genes, compute gene damage from the variants.
ii) Derive reaction damages from gene damages, where the relationship is encoded as the gene product association format in the GSM.
iii) Solve the minimization of metabolic adjustment (MOMA) [12] to get the flux, where the reference flux v_0 is the solution of the original GSM and the flux bound l and u are changed according to the reaction damage r.

2.1.1 Input Data

VaMP requires the number of each variant consequence (defined in [13]) of every genes as an input. In other words, $\mathbf{X} \in \mathbb{N}_0^{C \times G}$ where C is the number of consequence types and G is the number of genes in the GSM.

2.1.2 Gene Damage

From the variant consequence counts, VaMP computes a damage for each gene: $\Delta = \beta^\top \log(\mathbf{X} + 1) \in \mathbb{R}^G$, where β is a C-dimensional parameter.

Notice that each element of β should be positive since we assume that every variants may cause functional damage. Furthermore, because we assume that different types of consequences will have different severities, further constraints are added: $\beta_k \geq \beta_{k+1}$ for all $k \in \{1, \ldots, C-1\}$.

To establish this constraint, we derive β from $\beta^{(0)} \in \mathbb{R}^C$ as follows:

$$\beta_k = \sum_{i=k}^{C} \left\{ \mathrm{ELU}\left(\beta_i^{(0)}\right) + 1 \right\}$$

where ELU is defined as

$$\mathrm{ELU}(x) = \begin{cases} x & \text{if } x > 0, \\ \exp(x) - 1 & \text{if } x \leq 0. \end{cases}$$

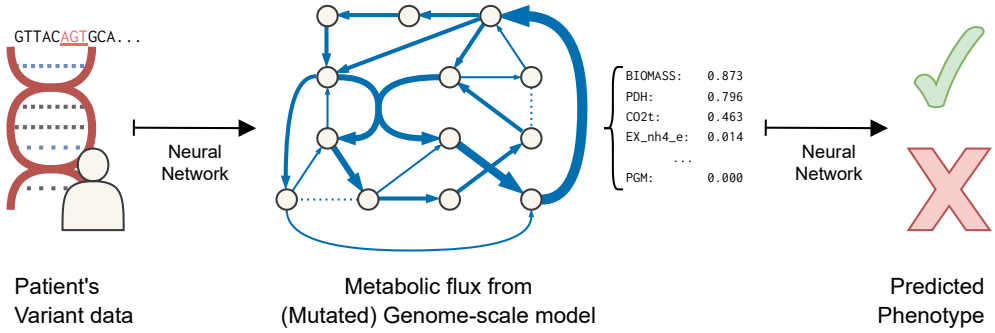

Figure 1: Overview of VaMP

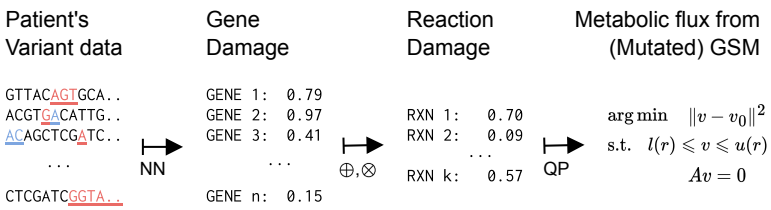

Figure 2: Encoder of VaMP

2.1.3 Reaction Damage

Reaction damages are derived from the gene damages, combined with the gene product association encoded in the GSM. Gene product association refers to the mapping of genes to their associated metabolic reactions, where the association is written as a boolean expression using `or` and `and`, implying one of the unit are required to catalyze the reaction (*e.g.*, they encode isozymes) and all units are required to catalzye the reaction (*e.g.*, they are subunits of enzymatic complex).

Usually, the logical expression in GSM is represented as a positive disjunctive normal form (DNF). In other words, the logical expression can be formulated as a nested list of genes, where the inner-level genes are connected by `and` and the outer-level conjunctions are connected by `or`. For instance, `[[A, B], [A, C, E]]` is interpreted to `(A and B) or (A and C and E)`.

In terms of damage, `or` and `and` in the association can be interpreted as $\min$ and $\max$ operations, respectively. If every genes involved in a reaction are in an `or` expression, then even if some genes are heavily damaged, another gene can replace it. Therefore, the actual damage to the reaction can be modeled via $\min$ function. Similarly, if every genes involved in a reaction are in an `and` expression, then a large damage to one gene will affect the reaction by that much, even if all the other genes are normal. Therefore, the actual damage to the reaction can be modeled via $\max$ function.

Formally, let $\delta \in \mathbb{R}^R$ be the reaction damage, where R is the number of reactions in the GSM. Define $\alpha : \{1, \ldots, R\} \to \mathcal{P}(\mathcal{P}(\{1, \ldots, G\}))$ be the gene product association map for each relation. From the DNF $\alpha(r) = \{A_1, \ldots, A_T\}$, δ_r can be computed as follows:

$$\delta_r = \min \{\max \{\Delta_i, i \in A_j\}, j = 1, \ldots, T\}$$

However, this expression is difficult to parallelize on a GPU, so VaMP uses $\mathrm{softmax}$ to efficiently

approximate the $\min/\max$ expressions.

2.1.4 Metabolic Flux

After computing the reaction damage δ, VaMP solves the minimization of metabolic adjustment (MOMA) [12] to get the flux v^* of damaged GSM:

$$\begin{aligned} v^* = \arg\min_{v} \ & \|v - v_0\| \\ \text{subject to } & Sv = 0, \\ & l \odot (1 - \text{sigmoid}(\delta)) \leqslant v, \\ & v \leqslant u \odot (1 - \text{sigmoid}(\delta)), \end{aligned}$$

where S is the stoichiometry matrix of GSM, v_0 is the reference flux, and the l and u are the flux bounds. Note that VaMP uses the solution of parsimonious flux balance analysis (pFBA) [14] from the original GSM as the reference flux.

VaMP currently uses OSQP [15] to solve every optimization problems. Note that this is not compatible to the usual automatic differentiation frameworks (*e.g.*, TensorFlow [16], PyTorch [17], ...) so we have to implement the backpropagation algorithm manually. Based on OptNet [18], we've developed a package to support the backpropagation of the specific form of the quadratic programming (see Appendix A for details).

2.2 Decoder

To predict the phenotype probability, VaMP first calculates the difference between the input flux and the reference flux. Then, the differences are fed into a neural network to predict the final probability (see Figure 3).

VaMP first calculates the difference between the flux v^* and the reference flux v_0. Since $Sv^* = Sv_0 = 0$, we can reduce the dimension by using bases of kernel space of S. Finally, the reduced difference is fed into a neural network to predict the final probability.

3 Experiment Setup

3.1 Preprocessing and Training

We used a human GSM from Metabolic Atlas [19], and genetic variants data from Genomics England (GE) [20] and gnomAD [21]. We chose five cancer types (breast, pancreatic, lung, renal, and colorectal cancer) from GE and trained model for each cancer type. We removed variants that not passed quality control filter and not mapped to genes in GSM.

We used Adam optimizer [22] to train VaMP, accompanied with the gradient clipping [23] and OneCycleLR learning rate scheduler [24]. Furthermore, we applied ℓ_1 penalty on the encoder and the decoder respectively.

3.2 Post-hoc Analysis

We followed post-hoc analysis method suggested by [25]. First, we tried to find significant reactions based on reaction fluxes. We used t-test to compute p-values between the normal group and the cancer group, and adjusted using Bonferroni correction. Furthermore, we calculated $\log_2$ FC and choose significant reactions based on adjusted p-values and $|\log_2 \text{FC}|$.

3.3 Pathway enrichment Analysis

After the significant reactions are obtained, we collected all genes related to that reaction. Using these genes, we did pathway enrichment analysis via hypergeometric tests and Benjamini-Hochberg procedure. We identified significantly enriched pathways having adjusted p-value less than 0.01.

3.4 SCREENER Analysis

We used a web service platform named SCREENER[1] [26], which extracts the gene-disease relations from the biomedical literature in real-time

[1] https://ican.standigm.com

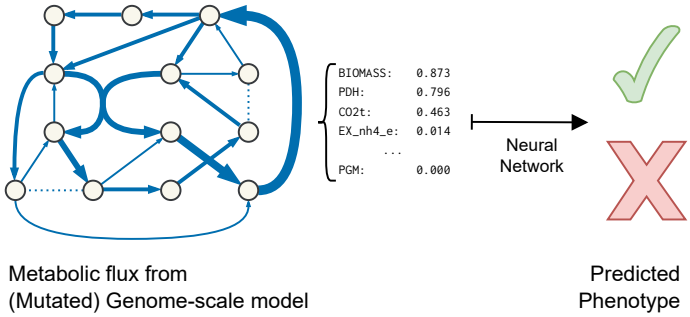

Figure 3: Decoder of VaMP

to identify the pairs of a significant gene and a cancer type having the known relationship based on literature evidence.

3.5 Co-occurrence Analysis

We used Biopython [27], freely available Python tools for computational molecular biology and bioinformatics to count the number of literature in PubMed mentioning each pair of a significant gene and a cancer type simultaneously.

3.6 LOR-correlation Analysis

We followed LOR-correlation analysis method suggested by [25]. We conducted statistical tests to assess the cancer-causing effect of variants based on the flux difference of the significant reactions between cancer and normal GSMs. It was examined by fitting a linear regression line between the log odds ratio (LOR) of variants and fluxes. The LOR for each variant concerning a specific cancer type was calculated as LOR $= \log_{10}((\mathrm{PC}/\mathrm{PN}) / (\mathrm{AC}/\mathrm{AN}))$, where PR, PS, AR, and AS represent the counts of the individual with the variant and the type of cancer (PC), have the variants and are normal (PN), do not have the variant and the type of cancer (AC), and do not have the variant and are normal (AN), respectively. If any of these counts were 0, 0.5 was added to each to ensure non-zero values when computing the logarithm. The fluxes associated with each variant were defined as the set of fluxes in individuals containing that variants. We used t-test to compute p-values between each pair of variants in a significant reaction.

4 Result

4.1 Identification of significant genes for five cancer types

We trained VaMP models of five cancer types to predict the cancer phenotype based on the metabolic flux adjusted by the information of variant consequences of metabolic genes. Then, we performed post-hoc analysis on each VaMP model of five cancer types to identify significant reactions of which fluxes are remarkably discriminating between cancer and normal phenotype, that is the cancer-causing reactions. During this process, we calculated $\log_2$ FC and adjusted p-values for every reaction in GSM and selected significant reactions using the threshold value of 0.5 for $|\log_2 \mathrm{FC}|$ and 0.01 for adjusted p-values. Additionally, the reactions with the mean flux values below 0.1 in both cancer and normal GSMs are excluded to avoid the influence by insignificant flux values. Furthermore, we extracted all genes which have variant information and are also related to the significant reactions in GSM simultaneously and these genes are defined as significant genes causing the cancer phenotype. Using these genes, we did pathway enrichment analysis using KEGG pathways to explore overall metabolic

pathways where significant genes are located (see Figure 4–Figure 8). As a result, the pathways in purine metabolism, carbohydrate digestion and absorption, starch and sucrose metabolism, adipocytokine signaling and thermogenesis are significantly enriched in five types of cancers.

4.2 Evaluation of identified significant genes based on literature evidence

After post-hos analysis, we further performed SCREENER analysis and co-occurrence analysis to evaluate the association of these significant genes with each cancer type. SCREENER is an AI-based text-mining model that learns the document-level relations between genes and diseases using an attention mechanism and identifies biological connections between target genes and diseases [26]. We did SCREENER analysis by counting the number of PubMed literature identified to have biological relationship between each pair of a significant gene and a cancer type by SCREENER model. Co-occurrence analysis was performed by counting the number of PubMed literature which mentioned each pair of significant gene and cancer type simultaneously. Based on these analysis, we compared the significant genes with SCREENER or co-occurrence evidence among five cancer types (see Figure 9 and Figure 10). Of total 146 significant genes, 87 genes (60%) have literature evidence at least with one cancer types by SCREENER analysis and 125 genes (86%) by co-occurrence analysis, meaning VaMP models can predict the existing known targets at a fairly reasonable level. The list of PubMed IDs for the pairs of significant genes and cancer types which are turned out to have literature evidence by SCREENER and co-occurrence analysis (see Supplementary File 1) could serve as supporting evidences to affirm the performance of VaMP. In SCREENER analysis, the more cancer types a gene is associated with, the more literature evidence there is while this correlation is not clear in co-occurrence analysis as SCREENER analysis. Because SCREENER captures semantic relationship of gene and disease in literature, SCREENER analysis is more reliable than co-occurrence analysis considering the fact that co-occurrence analysis just detects the coexistence of a pair of gene and disease in literature. If a gene is associated with various cancer types, it might be well studied and known previously and the correlation above reflects this phenomenon.

4.3 Metabolic signatures of significant genes by post-hoc analysis

Next, significant genes with post-hoc analysis results ($\log_2$ FC and adjusted p-values) and SCREENER and co-occurrence count information for five VaMP models are further studied. (see Figure 11–Figure 15). The y-axis value represents each significant gene, and the number of PubMed literature calculated through SCREENER and co-occurrence analysis is shown in two bar plots. The lower x-axis value of the heatmap with divided cells indicates the reaction IDs of the significant reactions to which a significant gene is related and the upper x-axis value indicates the metabolic subsystems in which the reactions participate. The lower and upper triangle of the rectangle in the heatmap represent $\log_2$ FC and adjusted p-value of each significant reaction respectively.

4.3.1 Breast cancer

As for breast cancer, top 5 significant genes by SCREENER count include *ACBA1*, *RRM2*, *CYP27A1*, *NT5E* and *SIRT6*. Each gene participates in the significant metabolic reactions related to transport reactions, purine metabolism, bile acid biosynthesis, nucleotide metabolism and nicotinate and nicotinamide metabolism (see Figure 11). As a result of post-hoc analysis, *ACBA1*-related reactions appear to have greater fluxes in cancer GSMs than normal GSMs meaning its up-regulation in breast cancer. The up-regulation pattern identified by VaMP is previously revealed in the study exploring the relationship of breast cancer and *ACBA1* [28]. Meanwhile, *RRM2*, *CYP27A1* and *SIRT6*-related

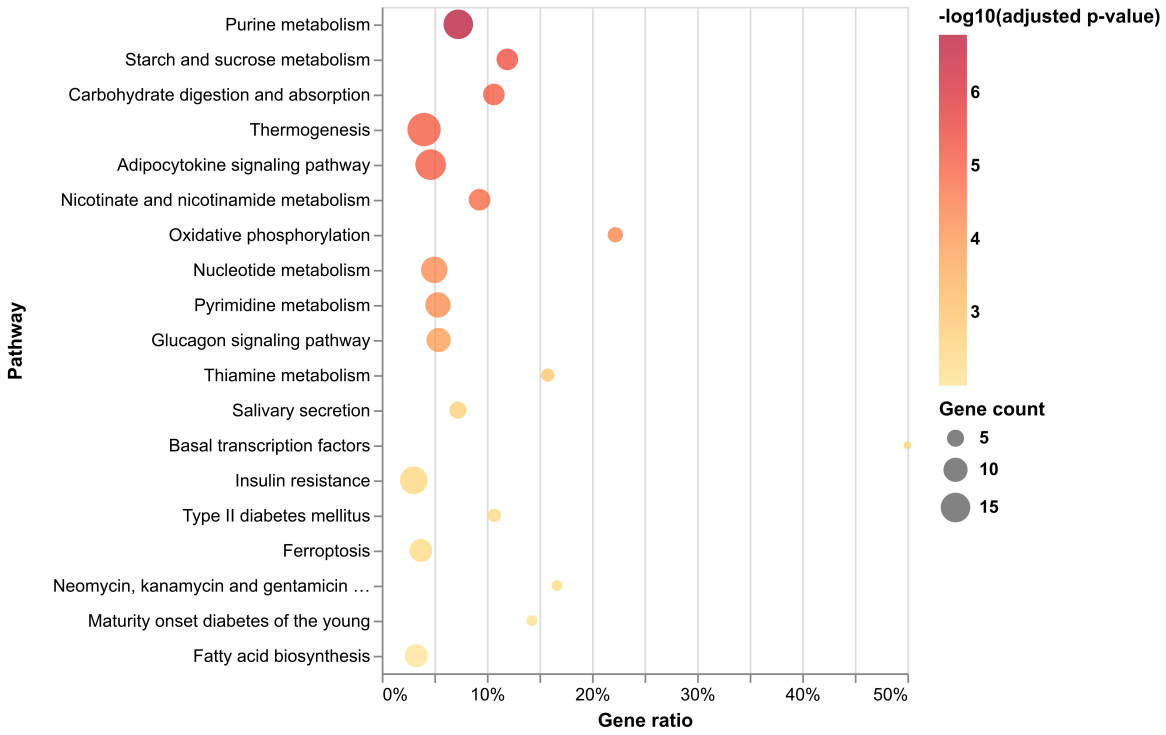

Figure 4: Pathway enrichment analysis of significant genes in breast cancer

reactions show smaller fluxes in cancer GSMs indicating their down-regulation in breast cancer. As for the case of *SIRT6*, metabolic signature identified VaMP is consistent with the study showing breast cancer patient's survival positively correlated with *SIRT6* abundance [29]. In the case of *NT5E*, purine metabolism-related reactions are up-regulated and while nucleotide metabolism-related reaction is up-regulated.

4.3.2 Colorectal cancer

In colorectal cancer, top 5 significant genes are *SLC2A1*, *RRM2*, *PKM*, *SIRT6* and *NT5E*. *SLC2A1* is related to transport reactions and *PKM* is related to purine metabolism (see Figure 12). *SLC2A1*-related reaction is up-regulated and *RRM2*, *PKM* and *SIRT6*-related reactions are down-regulated while *NT5E*-related reactions show the same pattern in the case of breast cancer. The metabolic pattern that *SLC2A1* is up-regulated in colorectal cancer well recapitulates the previous experimental study showing *SLC2A1* is highly expressed in colorectal cancer patients [30] and the down-regulation pattern of *PKM* is also supported by the study describing that the *PKM1*, an isoform of *PKM* is decreased in human colorectal carcinomas as compared to non-cancerous tissue [31].

4.3.3 Lung cancer

Next, top 5 significant genes in lung cancer turn out to be *SIRT6*, *SLC2A1*, *RRM2*, *SLC19A1* and *DCK* and the last two genes are related to transport reactions, purine metabolism and pyrimidine metabolism (see Figure 13). *SIRT6*, *RRM2* and *SLC19A1* are down-regulated and *SLC2A1* are up-regulated in lung cancer. The metabolic relationships of *SIRT2* and *SLC2A1* with lung cancer captured by VaMP are previously validated by two exper-

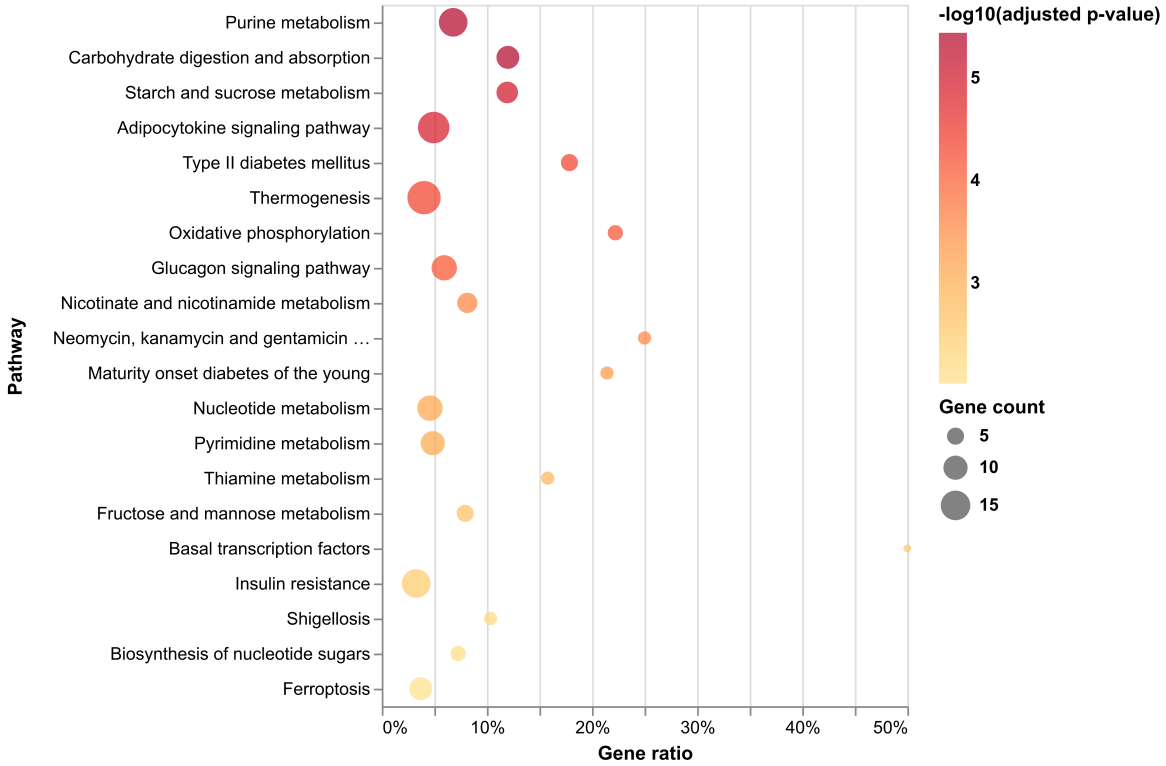

Figure 5: Pathway enrichment analysis of significant genes in colorectal cancer

imental studies, the one describing that *SIRT6* is generally down-regulated and linked to tumorigenesis in non-small cell lung carcinoma [32] and the other observing that strong up-regulation of *SLC2A1* in patients with lung adenocarcinoma and unveiling that *SLC2A1* was significantly correlated with prognosis [33]. *DCK* shows down-regulated pattern in purine metabolism-related reaction while up-regulated pattern is observed in pyrimidine metabolism-related reaction.

4.3.4 Pancreatic cancer

Top 5 significant genes in pancreatic cancer includes *RRM1*, *RRM2*, *NT5E*, *DCK* and *SIRT6* (see Figure 14). *RRM1*, *RRM2* and *SIRT6* appear to be down-regulated in pancreatic cancer and *NT5E* and *DCK* shows the same mixed pattern in the cases of breast and lung cancer. The down-regulation pattern of *SIRT6* in pancreatic cancer identified by VaMP is consistent with the experimental study finding *SIRT6* inactivation accelerates pancreatic ductal adenocarcinoma progression and metastasis *via* up-regulation of *LIN28B*, a negative regulator of the let-7 microRNA [34]. In the case of *NT5E*, among two mixed patterns, the up-regulation pattern in pancreatic cancer is aligned with the study describing *NT5E* gene was almost absent in normal pancreas but strongly up-regulated in pancreatic tumors [35].

4.3.5 Renal cancer

Lastly, *SLC2A1*, *CPT1A*, *ECHS1*, *ACSL1* and *HADH* are placed at top 5 significant genes in renal cancer (see Figure 15). *CPT1A* is related to fatty acid

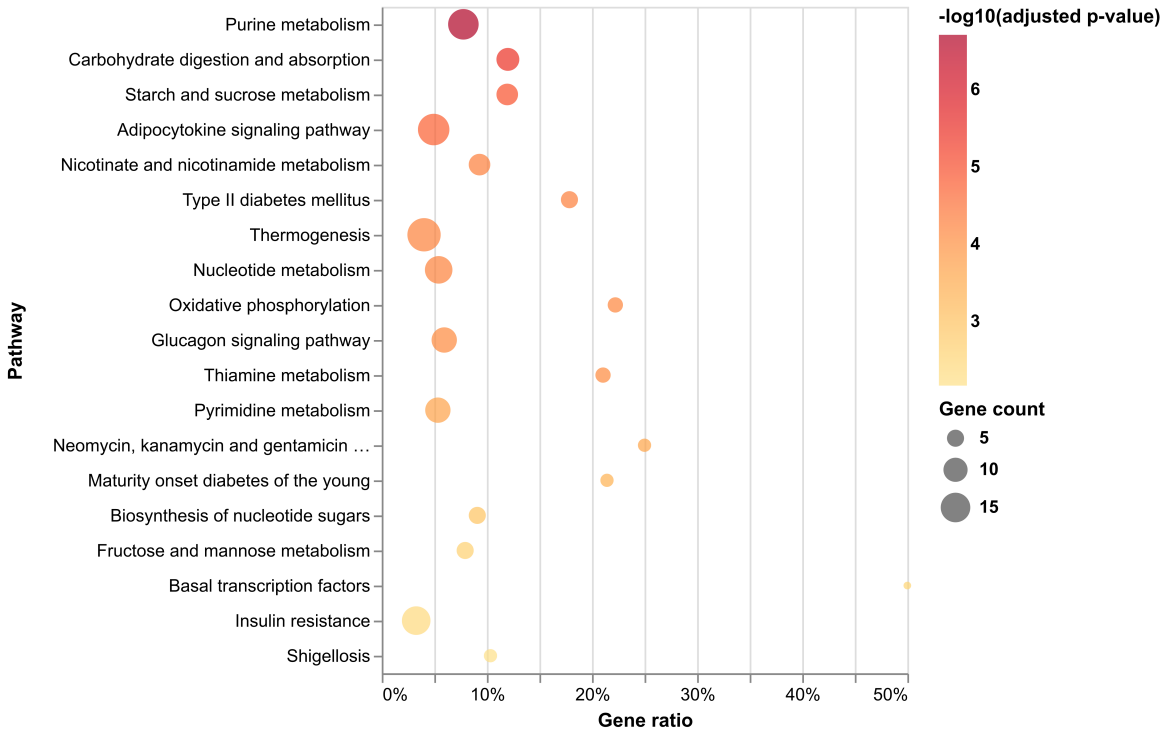

Figure 6: Pathway enrichment analysis of significant genes in lung cancer

oxidation and carnitine shuttle reactions and *ECHS1* is related to beta-oxidation of unsaturated fatty acids reaction. *ACSL1* participates in the significant reactions related to fatty acid activiation/oxidation, leukotriene metabolism and eicosanoid metabolism and *HADH* is related to beta-oxidation of unsaturated fatty acids reaction. According to post-hoc analysis, *SLC2A1* shows up-regulated pattern while *ECHS1* and *HADH* have down-regulated pattern in renal cancer. These metabolic patterns of *SLC2A1*, *ECHS1* and *HADH* in renal cancer identified by VaMP recapitulate the previous experimental studies well [36–39]. In the case of *CPT1A*, it is up-regulated in carnitine shuttle-related reactions but shows mixed pattern in fatty acid oxidation reactions. As for *ACSL1*, it shows up-regulated pattern in fatty acid activiation/oxidation-related reactions while it is down-regulated in leukotriene and eicosanoid metabolism-related reactions. Notably, the conflicting pattern of *ACSL1* in different metabolic subsystems (up-regulation in fatty acid metabolism and down-regulation in leukotriene and eicosanoid metabolism) precisely recapitulate the study showing that the *ACSL1* high-expression subgroup in clear cell renal cell carcinoma had enriched fatty acid metabolism-related pathways and high expression of ferroptosis-related gene while the *ACSL1* low-expression subgroup exhibited higher immune and microenvironment scores, elevated expression of immune checkpoints PDCD1, CTLA4, LAG3, and TIGIT [40]. Considering the fact that eicosanoids and leukotrienes are lipid-based signaling molecules that play a unique role in innate immune responses [41], the down-regulation pattern of *ACSL1* in eicosanoid and leukotriene metabolism captured by VaMP can be reasonably related to the influence of *ACSL1* low expression on immune system in renal cancer as described by the previous

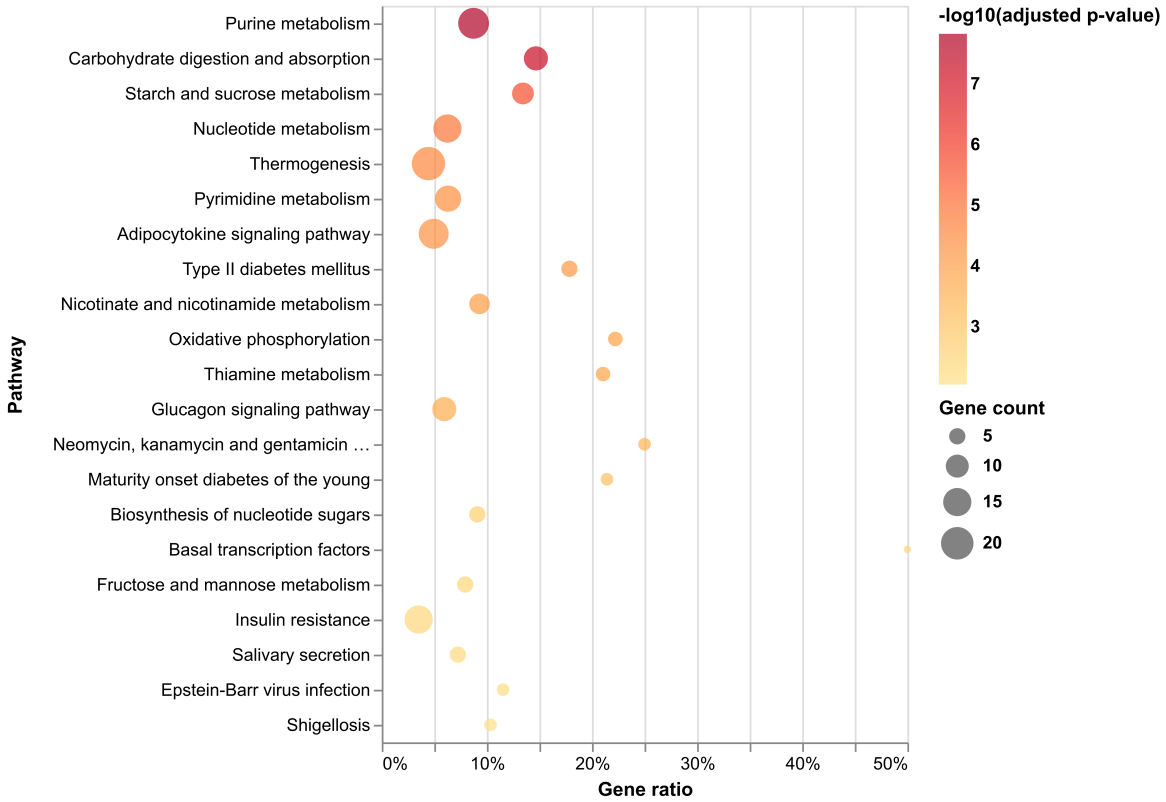

Figure 7: Pathway enrichment analysis of significant genes in pancreatic cancer

study.

Additionally, the pattern of *RRM1/2* and *SLC2A1* appearing to prevail across five cancer types can be explained by the previous study showing these two genes' pan-cancer role as an oncogene or a cancer biomarker [42, 43].

Overall, smaller adjusted p-values are observed in significant genes which have SCREENER evidence (SCREENER count > 0) implying that five VaMP models successfully recapitulate previously known relationships of cancer-causing genes and the cancer types. Therefore, other significant genes with no SCREENER evidence but low adjusted p-values can be potential novel cancer-causing genes and also be therapeutic targets for cancer treatment.

4.4 Linking significant genes to cancer-causing variants

For identifying what type of variant contribute to causing the certain type of cancer, the correlation between the fluxes of significant reactions and their related significant genes' variants are further analyzed by LOR-correlation (see §3.6). LOR-correlation plot of five cancer types and adjusted p-values for each pair of variants in significant reactions can be found in Supplementary File 2. Based on these data, the greater LOR means more frequent presence of the variant in the cancer patients. Therefore, the greater the difference in flux between a variant with a larger LOR value and a variant with a smaller LOR value, the more likely it is to contribute to causing cancer.

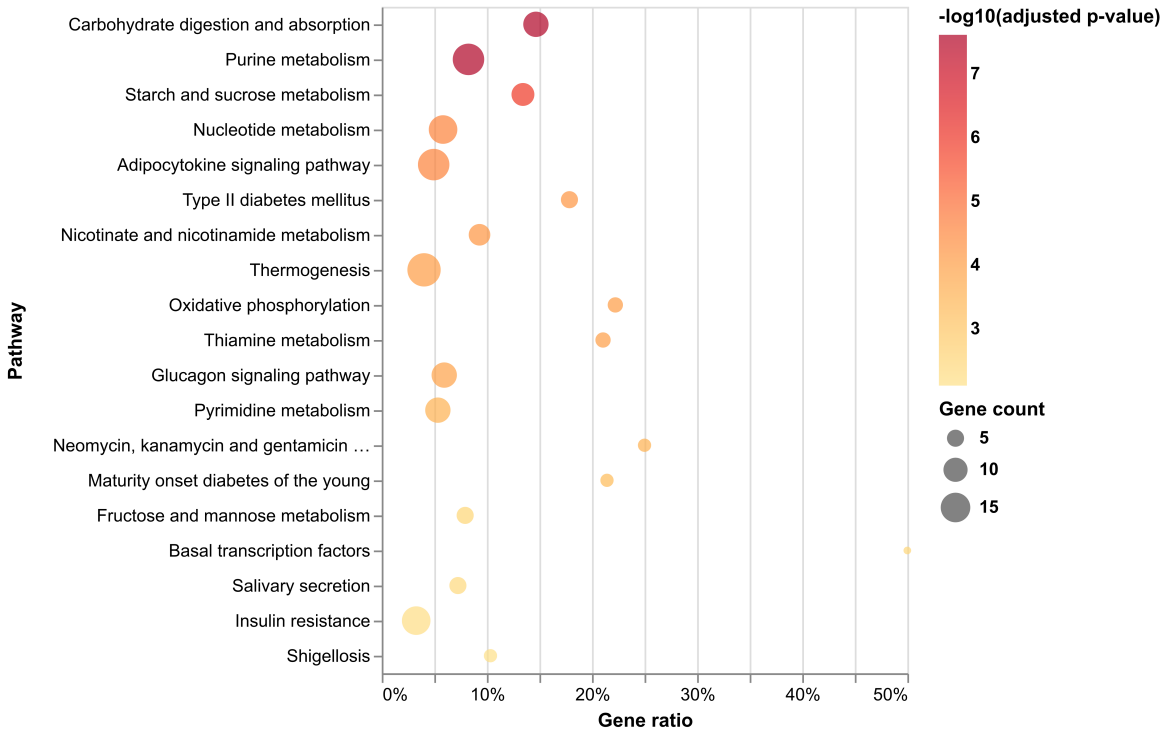

Figure 8: Pathway enrichment analysis of significant genes in renal cancer

5 Discussion

We have developed a novel approach, VaMP, which employs an end-to-end neural network for predicting the phenotype of a given patient based on their genetic variants. The key innovation lies in the representation of the latent space, which captures metabolic fluxes derived from a GSM. The training data for VaMP consists of mutation information and phenotypes, eliminating the need for preparing training data for metabolic flux. The encoder of VaMP maps genetic variants to flux through a series of steps involving gene damage computation, derivation of reaction damages, and solving the MOMA to obtain the flux. The gene damage computation employs a parameterized approach with constraints, ensuring positive values for each element. The reaction damage, derived from gene damage, considers gene product associations encoded in the GSM. The resulting metabolic flux is obtained by solving the MOMA, incorporating a quadratic programming framework. The decoder of VaMP predicts phenotype probabilities by calculating differences between input flux and reference flux. These differences are then fed into a neural network for the final probability prediction.

We further utilized a human GSM from Metabolic Atlas and genetic variant data from Genomics England and gnomAD for training VaMP models for five cancer types. Results reveal significant genes for each cancer type, obtained through post-hoc analysis, and their corresponding metabolic signatures. Evaluation using SCREENER and co-occurrence analyses demonstrates the ability of VaMP to predict known gene-disease relationships and their identified metabolic signatures are found to be strongly supported by various literature based on experimental studies. Additionally, we further analyzed the

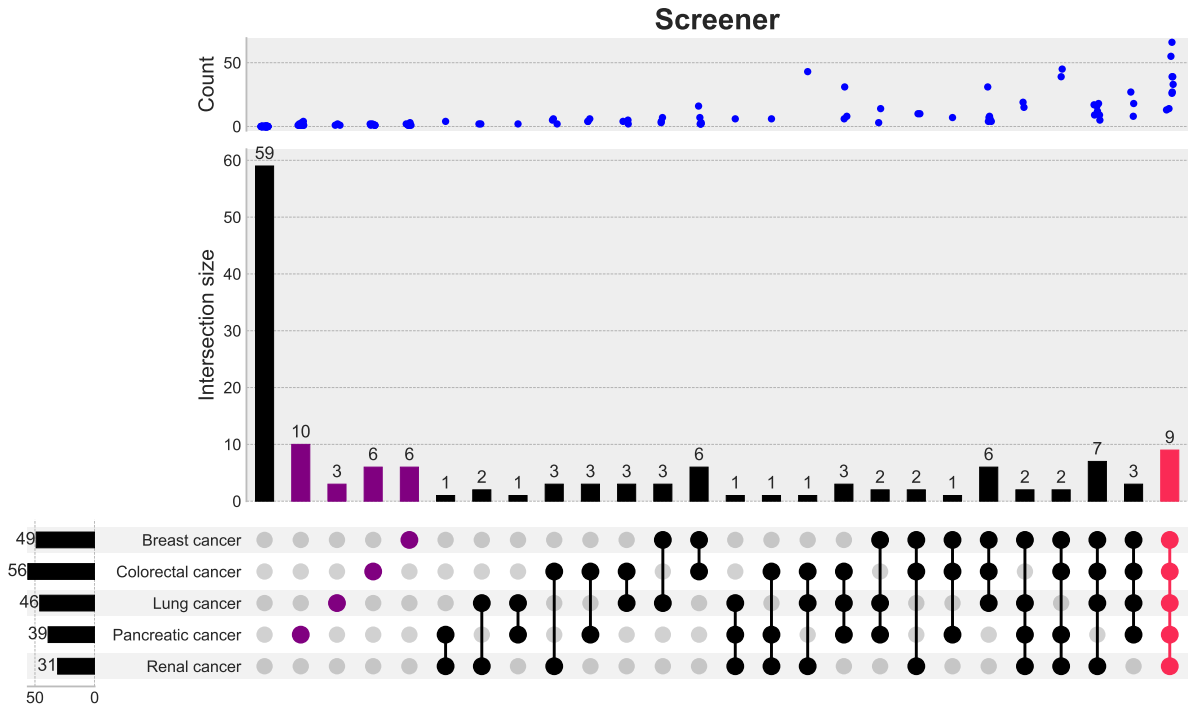

Figure 9: Upset plot of significant genes across five cancer types filtered by SCREENER analysis

results of five VaMP models to link significant genes to cancer-causing variants through LOR-correlation analysis.

On the other side, several improvements can be considered as the future research topics. One of the most obvious ideas is improving the encoder. Currently, the encoder is based on a linear regression model with the number of variants per consequence as input. This structure is oversimplified, and even the input features are too summarized. Thus, current encoder is not enough information to inspect specific variants. To overcome this issue, we should make the structure and the input features more complex. One can use protein language models (*e.g.*, ESM2 [44]) to compute the embedding vectors of reference protein sequence and mutated protein sequence, and use the difference of embedding vectors as the input features. This approach seems natural and promising, but it will require huge amounts of resources since it should handle millions of tokens per each sample. Another approach is to train a lightweight proxy model to simulate the protein language model, and use the model to compute the mutated embedding vectors. We consider two possible approaches for training the proxy model: (a) use knowledge distillation [45] (b) use another set of labels (*e.g.*, annotated labels from AlphaMissense [46])

In summary, the development of VaMP is not merely an academic pursuit but a crucial response to the evolving demands of cancer research, genetic exploration, and the growing intersection of machine learning with the life sciences. VaMP has the capacity to emerge as a potent tool for predicting cancer phenotypes based on genetic variants, showcasing its effectiveness in capturing known relationships and highlighting its potential for uncovering novel gene-disease associations.

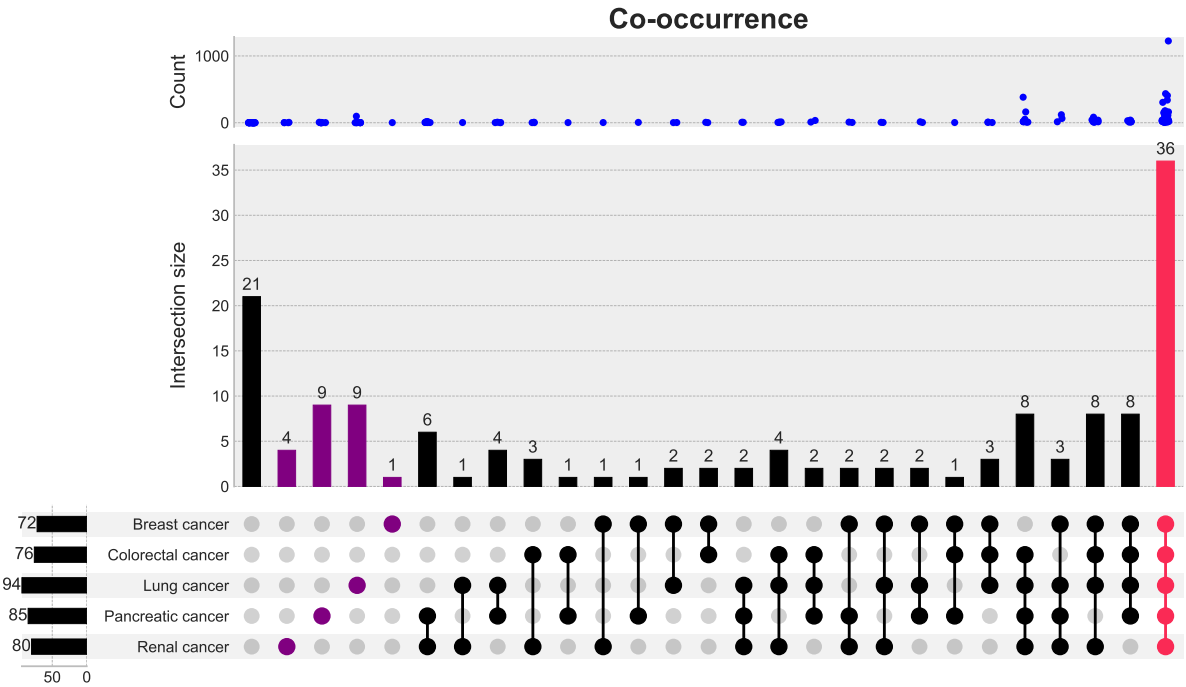

Figure 10: Upset plot of significant genes across five cancer types filtered by co-occurrence analysis

Acknowledgements

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

References

[1] M. Lee. "Deep Learning Techniques with Genomic Data in Cancer Prognosis: A Comprehensive Review of the 20212023 Literature". In: *Biology* 12.7 (June 21, 2023), p. 893. ISSN: 2079-7737. DOI: 10.3390/biology12070893.

[2] Y. Sun et al. "Identification of 12 Cancer Types through Genome Deep Learning". In: *Scientific Reports* 9.1 (Nov. 21, 2019), p. 17256. ISSN: 2045-2322. DOI: 10.1038/s41598-019-53989-3.

[3] M. F. Rogers, T. R. Gaunt, and C. Campbell. "Prediction of Driver Variants in the Cancer Genome via Machine Learning Methodologies". In: *Briefings in Bioinformatics* 22.4 (July 20, 2021), bbaa250. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbaa250.

[4] S. A. Forbes et al. "COSMIC: Mining Complete Cancer Genomes in the Catalogue of Somatic Mutations in Cancer". In: *Nucleic Acids Research* 39 (Database Jan. 1, 2011),

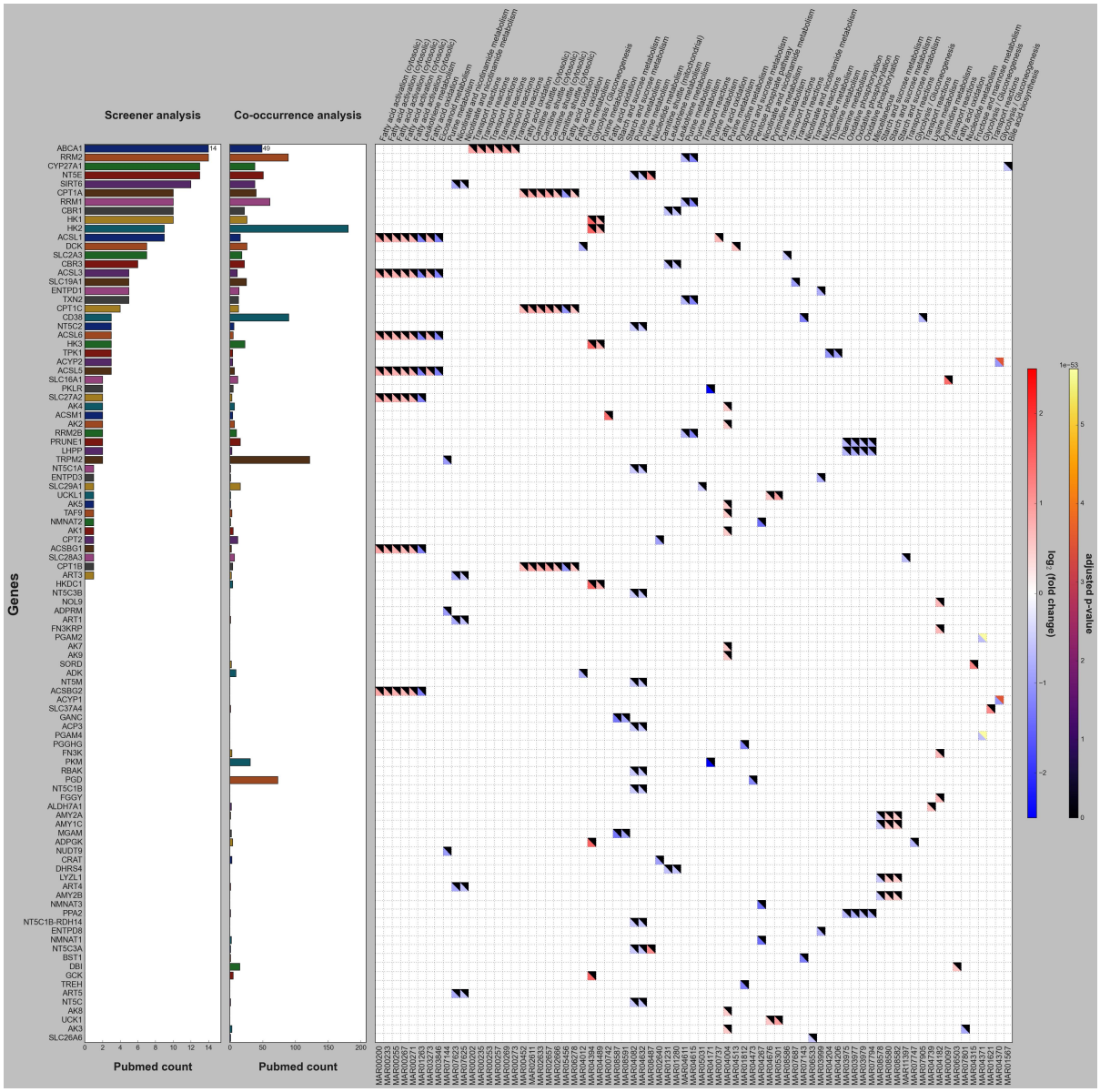

Figure 11: Post-hoc analysis of VaMP of breast cancer

pp. D945–D950. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkq929.

[5] L. Ding et al. "Analysis of Next-Generation Genomic Data in Cancer: Accomplishments and Challenges". In: *Human Molecular Genetics* 19.R2 (Oct. 15, 2010), R188–R196. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddq391.

[6] E. D. Pleasance et al. "A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome". In: *Nature* 463.7278 (Jan. 2010), pp. 191–196. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08658.

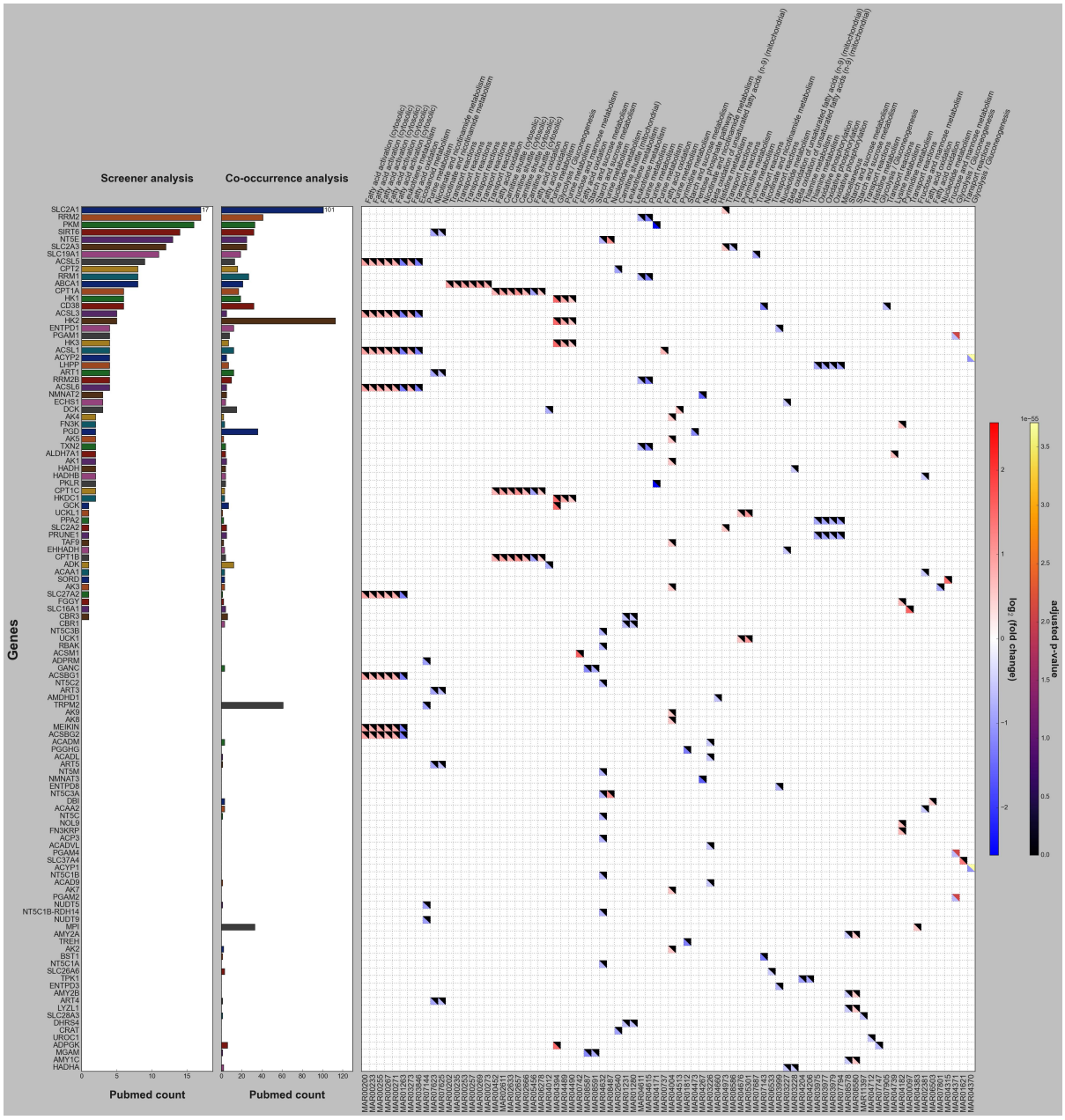

Figure 12: Post-hoc analysis of VaMP of colorectal cancer

[7] D. S. Paul, N. Soranzo, and S. Beck. "Functional Interpretation of Noncoding Sequence Variation: Concepts and Challenges". In: *BioEssays* 36.2 (Feb. 2014), pp. 191–199. ISSN: 0265-9247, 1521-1878. DOI: 10.1002/bies.201300126.

[8] D. P. Germain et al. "Challenging the Traditional Approach for Interpreting Genetic Variants: Lessons from Fabry Disease". In:

Clinical Genetics 101.4 (Apr. 2022), pp. 390–402. ISSN: 0009-9163, 1399-0004. DOI: 10.1111/cge.14102.

[9] C. Xu and S. A. Jackson. "Machine Learning and Complex Biological Data". In: *Genome Biology* 20.1 (Dec. 2019), 76, s13059-019-1689–. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1689-0.

[10] C. B. Azodi, J. Tang, and S.-H. Shiu. "Opening the Black Box: Interpretable Machine Learning for Geneticists". In: *Trends in Genetics* 36.6 (June 2020), pp. 442–455. ISSN: 01689525. DOI: 10.1016/j.tig.2020.03.005.

[11] A. Nilsson and J. Nielsen. "Genome Scale Metabolic Modeling of Cancer". In: *Metabolic Engineering* 43 (Sept. 2017), pp. 103–112. ISSN: 10967176. DOI: 10.1016/j.ymben.2016.10.022.

[12] D. Segrè, D. Vitkup, and G. M. Church. "Analysis of Optimality in Natural and Perturbed Metabolic Networks". In: *Proceedings of the National Academy of Sciences* 99.23 (Nov. 12, 2002), pp. 15112–15117. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.232349399.

[13] W. McLaren et al. "The Ensembl Variant Effect Predictor". In: *Genome Biology* 17.1 (Dec. 2016), p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

[14] N. E. Lewis et al. "Omic Data from Evolved *E. Coli* Are Consistent with Computed Optimal Growth from Genomescale Models". In: *Molecular Systems Biology* 6.1 (Jan. 2010), p. 390. ISSN: 1744-4292, 1744-4292. DOI: 10.1038/msb.2010.47.

[15] B. Stellato et al. "OSQP: An Operator Splitting Solver for Quadratic Programs". In: *Mathematical Programming Computation* 12.4 (Dec. 2020), pp. 637–672. ISSN: 1867-2949, 1867-2957. DOI: 10.1007/s12532-020-00179-2.

[16] TensorFlow Developers. *TensorFlow*. Version v2.15.0. Zenodo, Nov. 14, 2023. DOI: 10.5281/ZENODO.4724125.

[17] A. Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Dec. 3, 2019. arXiv: 1912.01703 [cs, stat]. preprint.

[18] B. Amos and J. Z. Kolter. "OptNet: Differentiable Optimization as a Layer in Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, July 17, 2017, pp. 136–145.

[19] J. L. Robinson et al. "An atlas of human metabolism". In: *Science signaling* 13.624 (2020), eaaz1482.

[20] M. Caulfield et al. *National Genomic Research Library*. figshare, 2020. DOI: 10.6084/M9.FIGSHARE.4530893.V7.

[21] S. Chen et al. *A Genome-Wide Mutational Constraint Map Quantified from Variation in 76,156 Human Genomes*. preprint. Genetics, Mar. 21, 2022. DOI: 10.1101/2022.03.20.485034.

[22] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. Jan. 29, 2017. arXiv: 1412.6980 [cs]. preprint.

[23] R. Pascanu, T. Mikolov, and Y. Bengio. *On the Difficulty of Training Recurrent Neural Networks*. Feb. 15, 2013. arXiv: 1211.5063 [cs]. preprint.

[24] L. N. Smith and N. Topin. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates*. May 17, 2018. arXiv: 1708.07120 [cs, stat]. preprint.

[25] E. S. Kavvas et al. "A Biochemically-Interpretable Machine Learning Classifier for Microbial GWAS". In: *Nature Communications* 11.1 (1 May 22, 2020), p. 2580. ISSN:

2041-1723. DOI: 10.1038/s41467-020-16310-9.

[26] M. Park et al. "SCREENER: Streamlined Collaborative Learning of NER and RE Model for Discovering Gene-Disease Relations". In: *PLOS ONE* 18.11 (Nov. 27, 2023). Ed. by N. Q. K. Le, e0294713. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0294713.

[27] P. J. A. Cock et al. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics". In: *Bioinformatics* 25.11 (June 1, 2009), pp. 1422–1423. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/btp163.

[28] S. Prijic and J. T. Chang. "ABCA1 Expression Is Upregulated in an EMT in Breast Cancer Cell Lines via MYC-Mediated De-Repression of Its Proximal Ebox Element". In: *Biomedicines* 10.3 (Mar. 2, 2022), p. 581. ISSN: 2227-9059. DOI: 10.3390/biomedicines10030581.

[29] U. Thirumurthi et al. "MDM2-mediated Degradation of SIRT6 Phosphorylated by AKT1 Promotes Tumorigenesis and Trastuzumab Resistance in Breast Cancer". In: *Science Signaling* 7.336 (July 29, 2014). ISSN: 1945-0877, 1937-9145. DOI: 10.1126/scisignal.2005076.

[30] X.-S. Liu et al. "SLC2A1 Is a Diagnostic Biomarker Involved in Immune Infiltration of Colorectal Cancer and Associated With m6A Modification and ceRNA". In: *Frontiers in Cell and Developmental Biology* 10 (Mar. 24, 2022), p. 853596. ISSN: 2296-634X. DOI: 10.3389/fcell.2022.853596.

[31] B. Park et al. "Pyruvate Kinase M1 Regulates Butyrate Metabolism in Cancerous Colonocytes". In: *Scientific Reports* 12.1 (May 24, 2022), p. 8771. ISSN: 2045-2322. DOI: 10.1038/s41598-022-12827-9.

[32] J.-l. Shang et al. "MDL-800, an Allosteric Activator of SIRT6, Suppresses Proliferation and Enhances EGFR-TKIs Therapy in Non-Small Cell Lung Cancer". In: *Acta Pharmacologica Sinica* 42.1 (Jan. 2021), pp. 120–131. ISSN: 1671-4083, 1745-7254. DOI: 10.1038/s41401-020-0442-2.

[33] Y. Wang et al. "Metabolic Reprogramming Induced by Inhibition of SLC2A1 Suppresses Tumor Progression in Lung Adenocarcinoma". In: *International Journal of Clinical and Experimental Pathology* 10.11 (2017), pp. 10759–10769. ISSN: 1936-2625. pmid: 31966419.

[34] S. Kugel et al. "SIRT6 Suppresses Pancreatic Cancer through Control of Lin28b". In: *Cell* 165.6 (June 2016), pp. 1401–1415. ISSN: 00928674. DOI: 10.1016/j.cell.2016.04.033.

[35] R. Samain et al. "CD73 Controls Myosin IIDriven Invasion, Metastasis, and Immunosuppression in Amoeboid Pancreatic Cancer Cells". In: *Science Advances* 9.42 (Oct. 20, 2023), eadi0244. ISSN: 2375-2548. DOI: 10.1126/sciadv.adi0244.

[36] A. L. G. Conceição et al. "Downregulation of OCLN and GAS1 in Clear Cell Renal Cell Carcinoma". In: *Oncology Reports* 37.3 (Mar. 2017), pp. 1487–1496. ISSN: 1021-335X, 1791-2431. DOI: 10.3892/or.2017.5414.

[37] L. Wang et al. "ECHS1 Suppresses Renal Cell Carcinoma Development through Inhibiting mTOR Signaling Activation". In: *Biomedicine & Pharmacotherapy* 123 (Mar. 2020), p. 109750. ISSN: 07533322. DOI: 10.1016/j.biopha.2019.109750.

[38] Y.-Y. Qu et al. "Inactivation of the AMPKGATA3ECHS1 Pathway Induces Fatty Acid Synthesis That Promotes Clear Cell Renal Cell Carcinoma Growth". In: *Cancer Research* 80.2 (Jan. 15, 2020), pp. 319–

333. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-19-1023.

[39] H. Jiang et al. "Decreased Expression of HADH Is Related to Poor Prognosis and Immune Infiltration in Kidney Renal Clear Cell Carcinoma". In: *Genomics* 113.6 (Nov. 2021), pp. 3556–3564. ISSN: 08887543. DOI: 10.1016/j.ygeno.2021.08.008.

[40] Y. Yang et al. "The Multi-Omics Analyses of Acsl1 Reveal Its Translational Significance as a Tumor Microenvironmental and Prognostic Biomarker in Clear Cell Renal Cell Carcinoma". In: *Diagnostic Pathology* 18.1 (Aug. 22, 2023), p. 96. ISSN: 1746-1596. DOI: 10.1186/s13000-023-01384-y.

[41] A. E. F. Sheppe and M. J. Edelmann. "Roles of Eicosanoids in Regulating Inflammation and Neutrophil Migration as an Innate Host Response to Bacterial Infections". In: *Infection and Immunity* 89.8 (July 15, 2021). Ed. by K. M. Ottemann, e00095–21. ISSN: 0019-9567, 1098-5522. DOI: 10.1128/IAI.00095-21.

[42] Q. Huang et al. "A Pan-Cancer Analysis of the Oncogenic Role of Polypyrimidine Tract Binding Protein 1 (PTBP1) in Human Tumors". In: *Medicine* 101.52 (Dec. 30, 2022), e32428. ISSN: 1536-5964. DOI: 10.1097/MD.0000000000032428.

[43] H. Zheng et al. "Glycolysis-Related SLC2A1 Is a Potential Pan-Cancer Biomarker for Prognosis and Immunotherapy". In: *Cancers* 14.21 (Oct. 29, 2022), p. 5344. ISSN: 2072-6694. DOI: 10.3390/cancers14215344.

[44] Z. Lin et al. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model". In: *Science* 379.6637 (Mar. 17, 2023), pp. 1123–1130. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ade2574.

[45] G. Hinton, O. Vinyals, and J. Dean. *Distilling the Knowledge in a Neural Network*. Mar. 9, 2015. arXiv: 1503.02531 [cs, stat]. preprint.

[46] J. Cheng et al. "Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense". In: *Science* (Sept. 19, 2023). DOI: 10.1126/science.adg7492.

A Differentiate the Quadratic Programming

One can follow [18] to derive the efficient forward and backward differentiation rules below. Consider the following optimization problem:

$$\begin{aligned} \text{minimize } & \frac{1}{2}x^\top P x + q^\top x \\ \text{subject to } & l \leqslant Ax \leqslant u \end{aligned}$$

where $x \in \mathbb{R}^n$ is the variable and $P \in \mathbb{S}^{n\times n}, A \in \mathbb{R}^{m\times n}, q \in \mathbb{R}^n, l \in (\{-\infty\} \cup \mathbb{R})^m$ and $u \in (\mathbb{R} \cup \{\infty\})^m$ are parameters. Note that this is the canonical input form of [15].

Let $x^* \in \mathbb{R}^n$ be the optimal solution of the problem and $y^* \in \mathbb{R}^m$ be the corresponding dual solution. In addition, define index sets $I_+, I_-, I_0 \subset \{1, .., m\}$ as

$$\begin{aligned} I_+ &= \{i : y_i^* > 0\}, \\ I_- &= \{i : y_i^* < 0\}, \\ I_0 &= \{i : y_i^* = 0\}. \end{aligned}$$

Furthermore, for any vector $v \in \mathbb{R}^m$, define $v_+ = v_{I_+}$ and vice versa. Similary, define $M_+ = M_{I_+,\cdot}$ for the matrix $M \in \mathbb{R}^{m\times n}$. Then, the optimality conditions for the problem can be stated as followed:

$$\begin{aligned} Px^* + q + A^\top y^* &= 0, \\ \max\{y, 0\} \circ (Ax^* - u) &= 0, \\ \min\{y, 0\} \circ (Ax^* - l) &= 0. \end{aligned}$$

Taking the differentiate with respect to x^*, y^*, q, l, u will give

$$\begin{aligned} Pdx^* + dq + A^\top dy^* &= 0, \\ y_+ \circ (A_+ dx^* - du_+) + dy_+ \circ (A_+ x^* - u_+) &= 0, \\ y_0 \circ (A_0 dx^* - du_0) + dy_0 \circ (A_0 x^* - u_0) &= 0, \\ y_- \circ (A_- dx^* - dl_-) + dy_- \circ (A_- x^* - l_-) &= 0, \\ y_0 \circ (A_0 dx^* - dl_0) + dy_0 \circ (A_0 x^* - l_0) &= 0. \end{aligned}$$

Note that $A_+ x^* - u_+ = 0$ and $A_- x^* - l_- = 0$, since $y_+, y_- \neq 0$. Also, by the definition, $y_0 = 0$. Therefore,

$$\begin{bmatrix} P & A_+^\top & A_-^\top & A_0^\top \\ \mathcal{D}(y_+) A_+ & \cdot & \cdot & \cdot \\ \mathcal{D}(y_-) A_- & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \mathcal{D}(A_0 x^* - u_0) \\ \cdot & \cdot & \cdot & \mathcal{D}(A_0 x^* - l_0) \end{bmatrix} \begin{bmatrix} dx^* \\ dy_+^* \\ dy_-^* \\ dy_0^* \end{bmatrix} = \begin{bmatrix} -dq \\ y_+ \circ du_+ \\ y_- \circ dl_- \\ 0 \\ 0 \end{bmatrix}.$$

where $\mathcal{D}(\cdot)$ is the diagonal matrix operator.

Last two rows implies that $dy_0^* = 0$ so corresponding rows and columns can be removed and this gives our core rule for the forward/reverse AD.

$$\begin{bmatrix} P & A_+^\top & A_-^\top \\ \mathcal{D}(y_+) A_+ & \cdot & \cdot \\ \mathcal{D}(y_-) A_- & \cdot & \cdot \end{bmatrix} \begin{bmatrix} dx^* \\ dy_+^* \\ dy_-^* \end{bmatrix} = \begin{bmatrix} -dq \\ y_+ \circ du_+ \\ y_- \circ dl_- \end{bmatrix}.$$

A.1 Forward Rule

The goal is to calculate $dx^*/d\theta$ while $dq/d\theta, dl/d\theta, du/d\theta$ are given, which can be accomplished by solving the following equation.

$$\begin{bmatrix} P & A_+^\top & A_-^\top \\ \mathcal{D}(y_+) A_+ & \cdot & \cdot \\ \mathcal{D}(y_-) A_- & \cdot & \cdot \end{bmatrix} \begin{bmatrix} dx^*/d\theta \\ - \\ - \end{bmatrix} = \begin{bmatrix} -dq/d\theta \\ y_+ \circ du_+/d\theta \\ y_- \circ dl_-/d\theta \end{bmatrix}.$$

Since $\mathcal{D}(I, y_+, y_-)$ is invertible, the above equation can be converted into

$$\begin{bmatrix} P & A_+^\top & A_-^\top \\ A_+ & \cdot & \cdot \\ A_- & \cdot & \cdot \end{bmatrix} \begin{bmatrix} dx^*/d\theta \\ - \\ - \end{bmatrix} = \begin{bmatrix} -dq/d\theta \\ du_+/d\theta \\ dl_-/d\theta \end{bmatrix}.$$

A.2 Reverse Rule

The goal is to calculate $df/dq, df/dl, df/du$ while df/dx^* is given. To derive the reverse rule, first

convert the core rule as in the matrix form:

$$\begin{bmatrix} P & A_+^\top & A_-^\top \\ \mathcal{D}\left(y_+\right) A_+ & . & . \\ \mathcal{D}\left(y_-\right) A_- & . & . \end{bmatrix} \begin{bmatrix} \mathcal{D}\left(dx^*\right) & . & . \\ . & \mathcal{D}\left(dy_+^*\right) & . \\ . & . & \mathcal{D}\left(dy_-^*\right) \end{bmatrix}$$
$$= \begin{bmatrix} -\mathcal{D}(dq) & . & . \\ . & \mathcal{D}\left(y_+ \circ du_+\right) & . \\ . & . & \mathcal{D}\left(y_- \circ dl_-\right) \end{bmatrix}.$$

Inverting and transposing the equation will give

$$\begin{bmatrix} P & \mathcal{D}\left(y_+\right) A_+^\top & \mathcal{D}\left(y_-\right) A_-^\top \\ A_+ & . & . \\ A_- & . & . \end{bmatrix}^{-1} \begin{bmatrix} df/dx^* \\ df/dy_+^* \\ df/dy_-^* \end{bmatrix}$$
$$= \begin{bmatrix} -df/dq \\ y_+^{-1} \circ df/du_+ \\ y_-^{-1} \circ df/dl_- \end{bmatrix}.$$

Similarly, above equation can be converted into the final form

$$\begin{bmatrix} P & A_+^\top & A_-^\top \\ A_+ & . & . \\ A_- & . & . \end{bmatrix} \begin{bmatrix} -df/dq \\ df/du_+ \\ df/dl_- \end{bmatrix} = \begin{bmatrix} df/dx^* \\ - \\ - \end{bmatrix}.$$

Notice that the coefficient matrix of the both forward rule and reverse rule are same and symmetric, so one can cache the efficient factorization (*e.g.*, $LDL^\top$ factorization) of the coefficient matrix and reuse to speed up the computation.

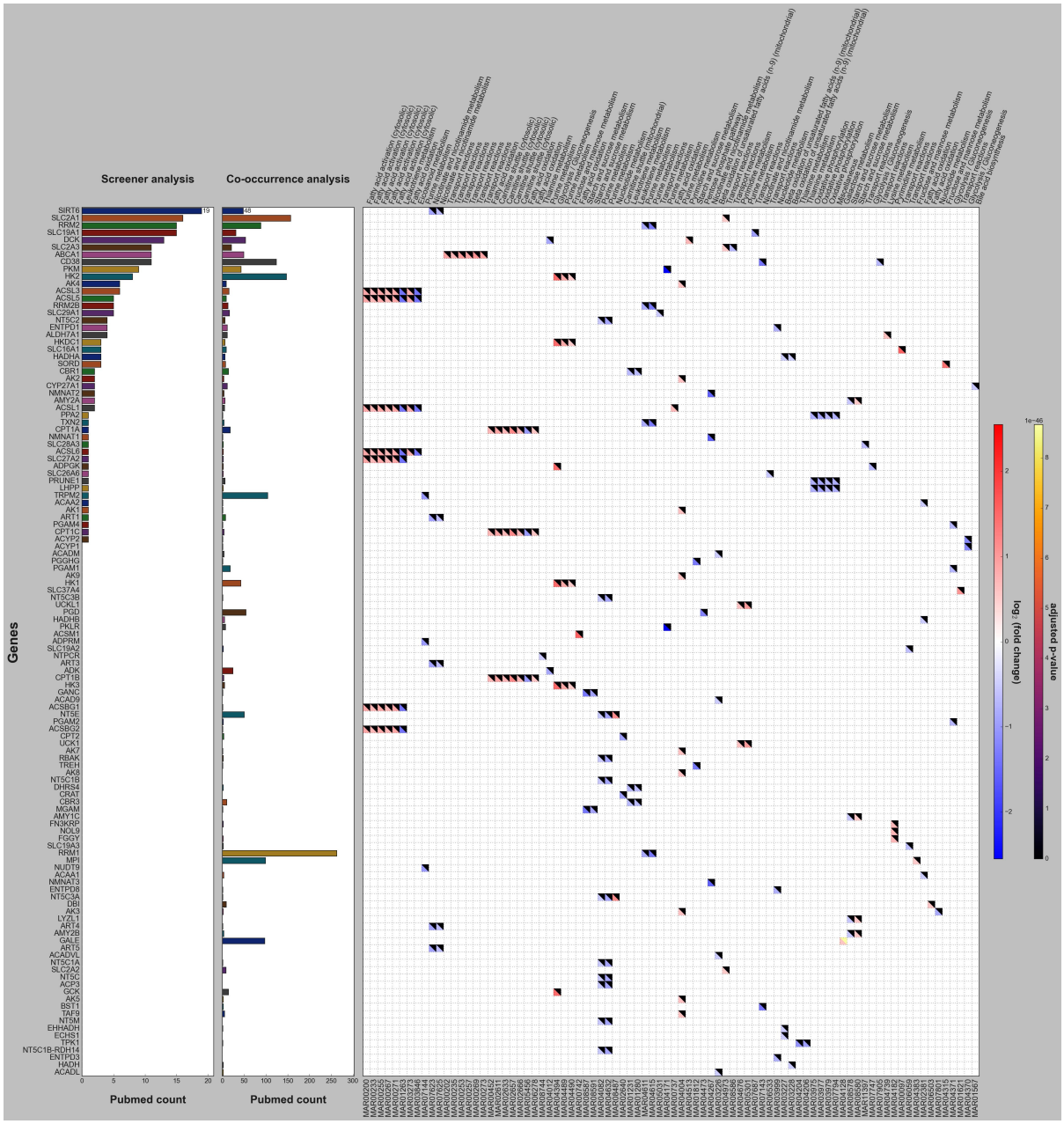

Figure 13: Post-hoc analysis of VaMP of lung cancer

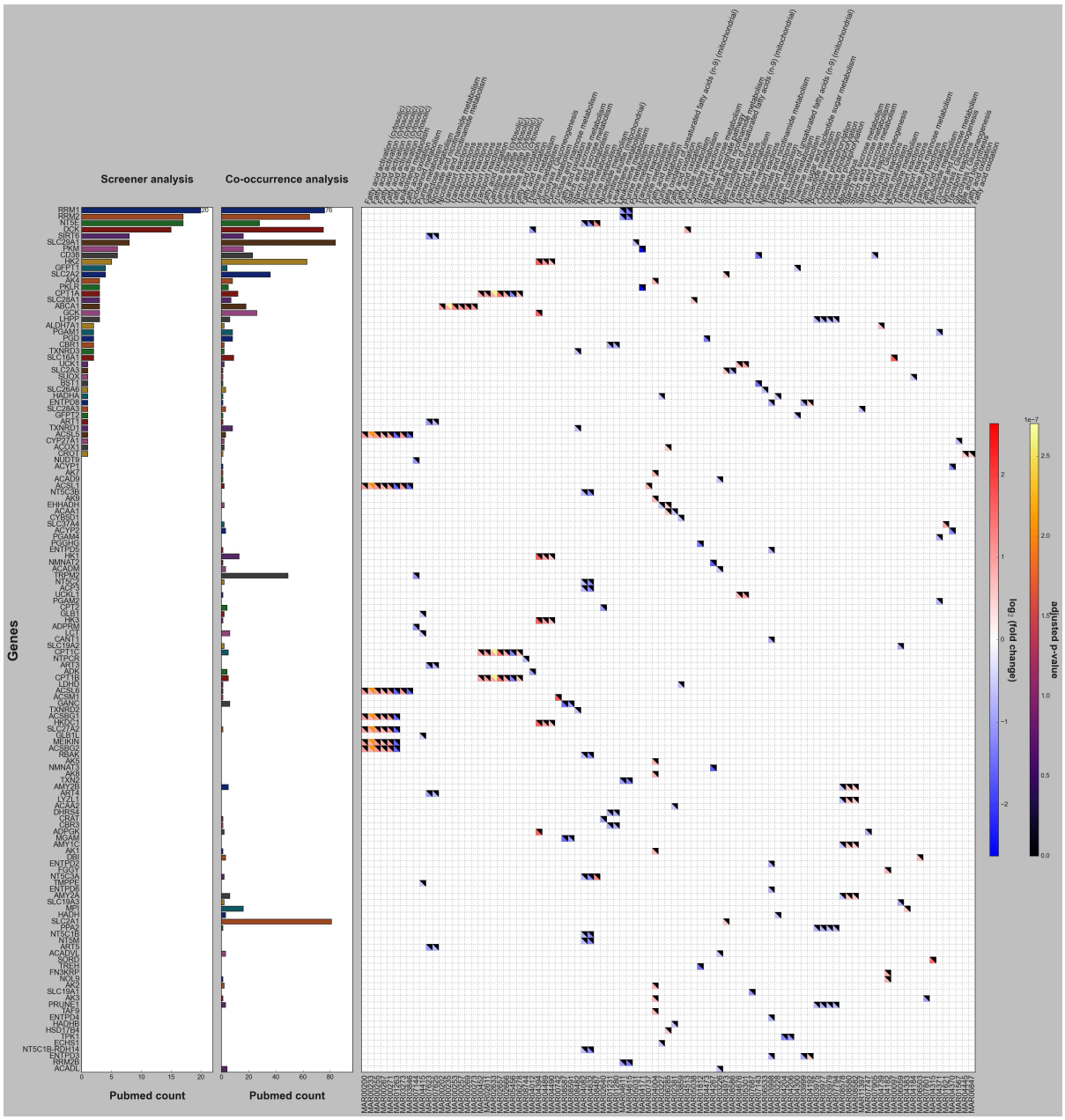

Figure 14: Post-hoc analysis of VaMP of pancreatic cancer

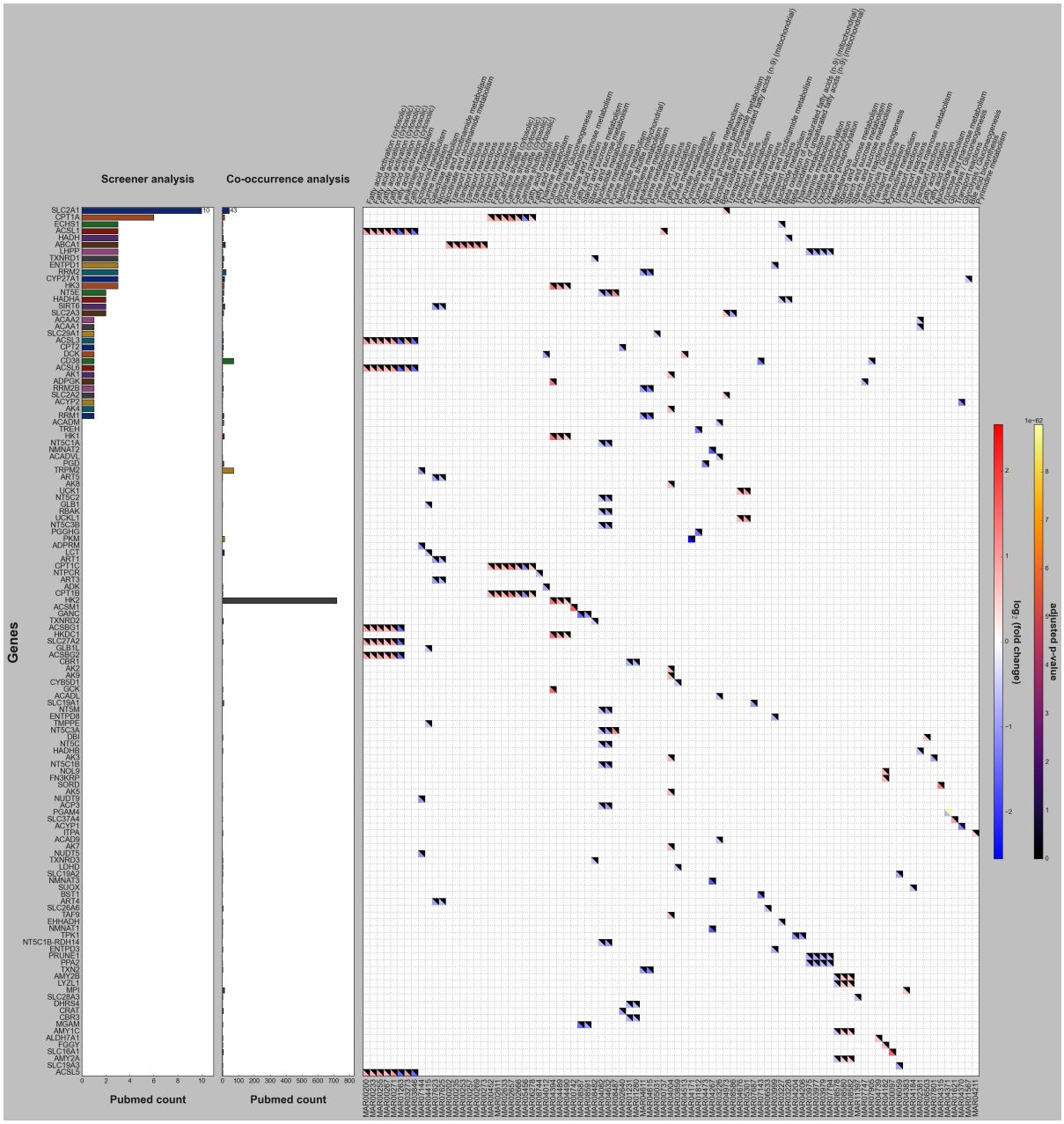

Figure 15: Post-hoc analysis of VaMP of renal cancer