Systems Biology

From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer

Yifeng Tao\textsuperscript{1}, Chunhui Cai\textsuperscript{2}, William W. Cohen\textsuperscript{3,§} and Xinghua Lu\textsuperscript{2,4,§}

\textsuperscript{1}Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, 15213, USA. and\textsuperscript{2}Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, 15206, USA. and\textsuperscript{3}Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, 15213, USA. and\textsuperscript{4}Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, Pittsburgh, 15261, USA.

\textsuperscript{§}To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Cancers are mainly caused by somatic genomic alterations (SGAs) that perturb cellular signaling systems and eventually activate oncogenic processes. Therefore, understanding the functional impact of SGAs is a fundamental task in cancer biology and precision oncology. Here, we present a deep neural network model with encoder-decoder architecture, referred to as genomic impact transformer (GIT), to infer the functional impact of SGAs on cellular signaling systems through modeling the statistical relationships between SGA events and differentially expressed genes (DEGs) in tumors. The model utilizes multi-head self-attention mechanism to identify SGAs that likely cause DEGs, or in other words differentiating potential driver SGAs from passenger ones in a tumor. GIT model learns a vector (gene embedding) as an abstract representation of functional impact for each SGA-affected gene. Given SGAs of a tumor, the model can instantiate the states of the hidden layer, providing abstract representation (tumor embedding) reflecting characteristics of perturbed molecular/cellular processes in the tumor, which in turn can be used to predict multiple phenotypes.

Results: We apply the GIT model to 4,468 tumors profiled by The Cancer Genome Atlas (TCGA) project. The attention mechanism enables the model to better capture the statistical relationship between SGAs and DEGs than conventional methods, and distinguishes cancer drivers from passengers. The learned gene embeddings capture the functional similarity of SGAs perturbing common pathways. The tumor embeddings are shown to be useful for tumor status representation, and phenotype prediction including patient survival time and drug response of cancer cell lines.

Availability: GIT and gene embeddings are available at https://github.com/yifengtao/genome-transformer.

Contact: wcohen@cs.cmu.edu, xinghua@pitt.edu

Supplementary information: SI available at https://github.com/yifengtao/genome-transformer.

1 Introduction

Cancer is mainly caused by the activation of oncogenes or deactivation of tumor suppressor genes (collectively called “driver genes”) as results of somatic genomic alterations (SGAs; Vogelstein et al., 2013), including somatic mutations (SMs, Kandoth et al. 2013; Lawrence et al. 2014), somatic copy number alterations (SCNAs; Ciriello et al. 2013; Zack et al. 2013), DNA structure variations (SVs; Stransky et al. 2014), and epigenetic changes (Jones and Baylin, 2002). Precision oncology relies on the capability of identifying and targeting tumor-specific aberrations resulting from driver SGAs and their effects on molecular and cellular phenotypes. However, our knowledge of driver SGAs and cancer pathways remains incomplete. Particularly, it remains a challenge to determine which SGAs (among often hundreds) in a specific tumor are drivers, which...
cellular signals or biological processes a driver SGA perturbs, and which molecular/cellular phenotypes a driver SGA affects.

Current methods for identifying driver genes mainly concentrate on identifying genes that are mutated at a frequency above expectation, based on the assumption that mutations in these genes may provide oncogenic advantages and thus are positively selected (Dees et al., 2012; Lawrence et al., 2013). Some works further focus on the mutations perturbing conserved (potentially functional) domains of proteins as indicators they may be driver events (Reva et al., 2011; Niu et al., 2016). However, these methods do not provide any information regarding the functional impact of enriched mutations on molecular/cellular phenotypes of cells. Without the knowledge of functional impact, it is difficult to further determine whether an SGA will lead to specific molecular, cellular and clinical phenotypes, such as response to therapies. What’s more, while both SMs and SCNAs may activate/deactivate a driver gene, there is no well-established frequency-based method that combines different types of SGAs to determine their functional impact.

Conventionally, an SGA event perturbing a gene in a tumor is represented as a “one-hot” vector spanning gene space, in which the element corresponding to the perturbed gene is set to “1”. This representation simply indicates which gene is perturbed, but it does not reflect the functional impact of the SGA, nor can it represent the similarity of distinct SGAs that perturb a common signaling pathway. We conjecture that it is possible to represent an SGA as a high-dimensional vector, in the same manner as the “word embedding” (Mikolov et al., 2013; Pennington et al., 2014) in the natural language processing (NLP) field, such that the representation reflects the functional impact of a gene on biological systems, and genes sharing similarly functions should be closely located in such embedding space. Motivated by this, we propose a scheme for learning “gene embeddings” for SGA-affected genes, i.e., a mapping from individual genes to low-dimensional vectors of real numbers that are useful on multiple predictive tasks. Our gene embeddings focus on the functional impact of SGAs.

Based on the assumption that SGAs perturbing cellular signaling systems often eventually lead to changes in gene expression, we introduce an encoder-decoder architecture neural network model called “genomic impact transformer” (GIT) to predict DEGs and detect potential cancer drivers with the supervision of DEGs. While deep learning models are being increasingly used to model different bioinformatics problems (Lee et al., 2016; Lan et al., 2018), to our knowledge there are few studies using neural network to model the relationships between SGAs and molecular/cellular phenotypes in cancers. The proposed GIT model has the following innovative characteristics: (1) The encoder part of transformer (Vaswani et al., 2017) first uses SGAs observed in a tumor as inputs, maps each SGA into a gene embedding representation, and combines gene embeddings of SGAs to derive a personalized “tumor embedding”. Then decoder part decodes and translates the tumor embedding to predict DEGs. (2) A multi-head self-attention mechanism (Bahdanau et al., 2014; Xu et al., 2015) is utilized in the encoder, which is a technique widely used in NLP to choose the input features that significantly influence the output. It perfectly suits our hope of selecting the important cancer drivers from the whole set of SGAs. It differentiates SGAs by assigning different weights to them so that it can potentially distinguish SGAs that have impact on DEG from those do not, i.e., detecting drivers from passengers. (3) Pooling inferred weighted impact of SGAs in a tumor produces a personalized tumor embedding, which can be used as an effective feature to predict DEGs and other phenotypes. (4) Gene embeddings are pre-trained by a “Gene2Vec” algorithm and refined by the GIT, which capture the functional impact of SGAs to the cellular signaling system. Our results and analysis indicate that above innovative approaches enable us to derive powerful gene embedding and tumor embedding representations that are highly informative of molecular, cellular and clinical phenotypes.

2 Materials and methods
2.1 Somatic genomic alterations (SGAs) pre-processing
We obtain SGA data, including SMs and SCNAs of 4,468 tumors consisting of 16 cancer types directly from TCGA portal (Network et al., 2013). We collectively designate all SGAs affecting a gene using the name of the gene being perturbed. After processing genomic data from TCGA, we use a binary variable in a “one-hot” vector to indicate the genomic status of a gene. For example, we represent the genomic status of TP53 as 1, if it is perturbed by one or more of SM/SCNA events in a tumor.

2.2 Differentially expressed genes (DEGs) pre-processing
Gene expression data are pre-processed and obtained from the Firehose browser of the Broad Institute. We determine whether a gene is differentially expressed in a tumor by comparing the gene’s expression in the tumor against a distribution of the expression values of the gene in the corresponding tissue-specific “normal” or control samples. For a given cancer type, assuming the expression of each gene (log 2 based) follows a Gaussian distribution in control sample, we calculate the p-values by determine the probability of observing a expression value from control distribution. If the p-value is equal or smaller than 0.005, the gene is considered as differentially expressed in the corresponding tumor. However, if a DEG is associated with an SCNA event affecting it, we remove it from the DEG list of the tumor.

2.3 The genomic impact transformer (GIT) neural network
2.3.1 GIT network structure: encoder-decoder architecture
Figure 1a shows the general structure of the genomic impact transformer (GIT) model with an overall encoder-decoder architecture. GIT mimics hierarchically organized cellular signaling system (Chen et al., 2015, 2016), in which a neuron may potentially encode the signal of one or more signaling proteins. When cellular signaling system is perturbed by SGAs, it often can lead to changes in measured molecular phenotypes, such as gene expression changes.

Thus, for a tumor t, the set of its SGAs $[g]_{m=1}^M$, is connected to the GIT neural network as observed input (Figure 1a bottom part squares). The impact of SGAs is represented as embedding vectors $[e]_{m=1}^M$, and the impact of them in a tumor is combined into a tumor embedding vector $e_t$ through an attention mechanism in the encoder (Figure 1a middle part). We further explicitly represent cancer type $s$ and its influence on encoding system $e_t$ of the tumor because tissue type influences which genes are expressed in cells of a specific tissue and thereby can be differentially expressed in cancer cells when compared to normal tissue. Since it is well-known that the majority of SGAs in a tumor are likely passengers and only a few are drivers, we design a self-attention module to assign a tumor-specific weight to each SGA to modulate its impact (Figure 1b left). We use a multi-head attention mechanism to represent the fact that a tumor usually involves multiple processes, and they potentially can be represented by multiple single-head self-attention mechanisms. After encoding the SGAs and cancer type into tumor embedding, the decoder module, which consists of a feed forward multi-layer perceptron (MLP, Rosenblatt, 1958), finally translates the functional impact of SGAs and cancer type into DEGs of the tumor (Figure 1a top part).

2.3.2 Pre-training gene embeddings using Gene2Vec algorithm
In this study, we project the discrete binary representation of SGAs perturbing a gene into a continuous embedding space, which we call “gene embeddings” of corresponding SGAs. While such embeddings can be directly learned using the GIT model, it has been shown in the field of NLP that the pre-trained word embeddings can significantly improve the performance in other related NLP tasks (Devlin et al., 2018; Tao et al.,...
Tumor embedding

Cancer patient: assumes the probability that an alteration in gene are able to portrait the co-occurrence pattern of SGAs in each tumor, i.e., algorithm. The biology rationale behind Gene2Vec algorithm is that we embedding of any possible SGA-affected gene closely related the skip gram word2vec (Mikolov et al., 2013) pre-training algorithm. The biology rationale behind Gene2Vec algorithm is that we are able to portray the co-occurrence pattern of SGAs in each tumor, i.e., mutually exclusive mutations (Vandin et al., 2012), using gene embeddings and gene context embeddings.

Given the gene embedding $e_{g}$ of an SGA-affected gene $g$ and context embedding of any possible SGA-affected gene $c$, $V = \{v_{c}\}_{c \in G}$, where $G$ is the set of all possible SGA-affected genes, the skip gram paradigm assumes the probability that an alteration in gene $c$ happens together with the alteration in gene $g$ within a tumor with probability:

$$\Pr (c \in \text{Context}(g) | g) = \frac{\exp (e_{c}^{T}v_{g})}{\sum_{c \in G} \exp (e_{c}^{T}v_{g})} \quad (1)$$

We use the negative sampling (NS) technique, a simplified version of noise contrastive estimation (Gutmann and Hyvärinen, 2012), to approximately maximize the log-likelihood of skip gram, which will otherwise be computational expensive to optimize if directly following Equation 1.

By initializing gene embeddings with the pre-trained gene embeddings in the GIT, we transfer the knowledge learned from SGA pattern behind tumors. These gene embeddings are further updated and fine-tuned by the GIT model with the supervision of affected DEGs.

### 2.3.3 Encoder: multi-head self-attention mechanism

Instead of allowing SGAs and cancer type variable to directly connect to encoder, we start with an embedding layer, which enables us to use attention mechanisms to differentiate impact of potential driver and passenger SGAs (Figure 1 middle part). For all SGA-affected genes $\{s\}_{s \in G}$ and the cancer type $s$ of a tumor $t$, we map them to corresponding gene embeddings $\{e_{s}\}_{s \in G}$ and a cancer type embedding $e_{t}$ on the state of the cancer cells in a tumor $t$ using multi-head self-attention mechanism. Then the tumor embedding $e_{t}$ is decoded by a MLP to predict DEGs. Example SGAs of a TCGA breast cancer case (TCGA-D8-A1JJ) are shown, in which PIK3CA, TP53 and GATA3 are assigned with largest attention weights. We use the negative sampling (NS) technique, a simplified version of noise contrastive estimation (Gutmann and Hyvärinen, 2012), to approximately maximize the log-likelihood of skip gram, which will otherwise be computational expensive to optimize if directly following Equation 1.

By initializing gene embeddings with the pre-trained gene embeddings in the GIT, we transfer the knowledge learned from SGA pattern behind tumors. These gene embeddings are further updated and fine-tuned by the GIT model with the supervision of affected DEGs.
The tumor (Figure 1) self-attention mechanism using gene embeddings of SGAs \(\{e_1, e_2, \ldots, e_m\}\) in the tumor (Figure 1):

\[
\alpha_1, \alpha_2, \ldots, \alpha_m = \text{Function}_{\text{Attention}}(e_1, e_2, \ldots, e_m). 
\]  

(3)

The attention function Function_{\text{Attention}} is implemented as a sub-network. In the case of single-head attention, there is only one single-head parameter \(\theta_j\), and the unnormalized weights \(\{\beta_{g,j}\}_{g=1}^m\) can be derived as follows:

\[
\beta_{g,j} = \theta_j^T \cdot \text{tanh}(W_g \cdot e_j), \quad g = 1, 2, \ldots, m, \tag{4}
\]

which are further normalized to single-head weights \(\{\alpha_{g,j}\}_{g=1}^m\):

\[
\alpha_{g,j} = \text{softmax}(\beta_{g,j}), \quad g,j = 1, 2, \ldots, m. 
\]  

(5)

where softmax function is defined as:

\[
\alpha_g = \frac{\exp(\beta_g)}{\sum_{g=1}^m \exp(\beta_g)}. 
\]  

(6)

Overall we have three parameters \(\{W_g, \Theta_j, \ell\}\) to train in the multi-head attention module using back-propagation (Rumelhart et al., 1986). The look-up table \(\{e_1, e_2, \ldots, e_m\}\) is initialized with Gene2Vec pre-trained gene embeddings and refined by GIT here.

2.3.4 Decoder: multi-layer perceptron (MLP)

For a specific tumor, we feed tumor embedding \(e_t\) into a MLP with one hidden layer as decoder, using non-linear activation functions and fully connected layers to produce the final predictions \(\hat{y}\) for DEGs \(y\) (Figure 1a top part):

\[
\hat{y} = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \text{ReLU}(e_t) + b_1)) + b_2, 
\]  

(7)

where \(\text{ReLU}(x) = \max(0, x)\) is rectified linear unit, and \(\sigma(x) = (1 + \exp(-x))^{-1}\) is sigmoid activation function.

The output of decoder and actual values of DEGs are used to calculate the \(\ell_2\)-regularized cross entropy, which is minimized during training:

\[
\min_{W, \ell} \text{CrossEntropy}(\hat{y}, y) + \frac{\lambda}{2} \|W\|_2^2, \Theta, \ell. 
\]  

(8)

where \(W = \{W_1\}_{l=1}^2, \Theta\) and \(\ell\) defined in Section 2.3.3; cross entropy loss defined as:

\[
\text{CrossEntropy}(\hat{y}, y) = -\sum_i \left[(1 - y_i) \log(1 - \hat{y}_i) + y_i \log\hat{y}_i\right], 
\]  

(9)

and \(\ell_p\) regularizer defined as:

\[
\ell_p(W; \lambda) = \lambda \cdot \sum_{i,j} \|W_{i,j}\|_p, p \in \{1, 2\}. 
\]  

(10)

2.4 Training and evaluation

We utilize PyTorch (https://pytorch.org/) to train, validate and test the Gene2Vec, GIT (variants) and other conventional models (Lasso and MLPs; Section 3.1). The training, validation and test sets are split in the ratio of 0.33:0.33:0.33 and fixed across different models. The hyperparameters are tuned over the training and validation sets to get best results, trained on training and validation sets, and finally applied to the test set for evaluation if not further mentioned below. The models are trained by updating parameters using backpropagation (Rumelhart et al., 1986), specifically, using mini-batch Adam (Kingma and Ba, 2014) with default momentum parameters. Gene2Vec uses mini-batch stochastic gradient descent (SGD) instead of Adam. Dropout (Hinton et al., 2012) and weight decay \((\ell_p\text{-regularization})\) are used to prevent overfitting. We train all the models 30 to 42 epochs until they fully converge. The output DEGs are represented as a sparse binary vector, with overall 31.3% of genes being differentially expressed. In order to judge the performance of different methods, we utilize various performance metrics including accuracy, precision, recall and F1 score, where F1 is the harmonic mean of precision and recall:

\[
\text{F1} = 2 \left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{11}
\]

The training and test are repeated for five runs get the mean and variance of evaluation metrics. More evaluation methods and details, including that of GO enrichment analysis (Section 3.2), driver detection (Section 3.3), survival analysis (Section 3.4) and drug response (Section 3.5) are in corresponding sections to make the discussion more explicit and compact.

3 Results

3.1 GIT statistically detects real biological signals

The task of GIT is to predict DEGs (dependent variables) using SGAs as input (independent variables). We plot both F1 score and accuracy on test set as the function of trained epochs (Figure 2 “real data”), which indicate that the model gains capability of predicting DEGs as training proceeds, and finally reaches a stable state. In order to validate that GIT is able to extract real statistical relationships between SGAs and DEGs, we randomly shuffle the positions of DEGs in the DEG vector of a tumor, i.e., randomly relabel DEG names, and then train a GIT to predict DEGs from SGAs. We compare the performance of models trained with random datasets, by plotting F1 score and accuracy during training of the models (Figure 2 “shuffled data”). Note that, since most DEGs in the data are zeros, a trivial solution is to call every DEG as 0, which can also achieve good overall accuracy and minimize loss, but that will result in a low F1 because of low recall. Indeed, the test F1 score in the DEG-permutation case drops to a very low value due to the same reason. These results demonstrates that GIT is able to capture real statistical relationships between SGAs and DEGs from the noisy biological data, instead of simply behaving as a super function fitter that fits random relationships.

As comparison, we also train and test the Lasso (Tibshirani, 1994) and MLPs (Rosenblatt, 1958) as baseline prediction models to predict DEGs based on SGAs. The Lasso model is appealing in our setting because, when predicting a DEG, it can filter out most of irrelevant input variables (SGAs) and keep only the most informative ones, and it is a natural choice in our case where there are 19.8k possible SGAs. However, in comparison to MLP, it lacks capability of portraying complex relationships between SGAs and DEGs. On the other hand, while conventional MLPs have sufficient power to capture complex relationships—particularly, the neurons in hidden layers may mimic signaling proteins (Chen et al., 2016)—they cannot not utilize any biological knowledge extracted from cancer...
Predicting cancer phenotypes via genomic impact transformers

3.1 Gene embeddings of SGAs serve as abstract representations of their functional impact

We examine whether the gene embeddings really reflect some biological properties of genes. We first examine if the genes sharing similar embeddings also share same any function annotation, using the Gene Ontology (GO; Ashburner et al., 2000) as a means to evaluate functional similarity of genes. GO is an ontology database that describes the relations of biological concepts and systematically assigns annotations to specific genes or proteins. We mainly concentrate on evaluating whether SGA-affected genes share GO annotations in the “biological process” domain, as genes involved in a common biological process will likely share common functional impact. The top 1,474 affected genes share GO annotations through “GO enrichment”, which is defined as follows: 

$$\text{enrichment} = \frac{\mathbb{E}_{\text{Clust}(g) \in \text{Clust}(c)} [\mathbb{I}(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]}{\mathbb{E}_{\text{g} \in \mathbb{G}} [\mathbb{I}(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]}$$  

(13)

where Clust(g) is the cluster that gene g belongs to. GO enrichment considers the function similarities of genes that are close in the embedding space. The larger it is, the higher correlated are the GO functions and clusters (and it equals to 1 in random case). We perform clustering analysis of SGAs in embedding space using k-means clustering, and calculate GO enrichment, and we vary number of clusters (k) to derive clusters with different degrees of granularity. The change of GO enrichment with different numbers of clusters is shown in Figure 3a. As one can see, when the genes are randomly distributed in the embedding space, they get GO enrichment of 1. However, in the gene embedding space, the GO enrichment increases fast until the number of clusters reaches 40, indicating strong correlation between the clusters in embedding space and the functions of the genes.

Finally, we group the genes into 40 clusters, and show the t-SNE visualization (Maaten and Hinton, 2008) of genes in the embedding space (Figure 3b left panel). Using PANTHER GO enrichment analysis (Mi et al., 2013), 12 out of 40 clusters are shown to be enriched in at least one biological process. As one can see, most of gene clusters are well-defined and tight located in the projected t-SNE space. As case study, we take a close look at one cluster (Figure 3b right panel), which contains a set of functionally similar genes, such as that code a protein family of type I interferons (IFNs), which are responsible for immune and viral response (de Weerd and Nguyen, 2012). In sum, all the above results indicate that the gene embeddings capture the functional similarity of SGAs.

3.3 Self-attention mechanism reveals SGAs with impact on cancer cell transcriptome

While it is a widely accepted that cancer is mainly caused by SGAs, but not all SGAs observed in a cancer cell are causative (Vogelstein et al., 2013). Previous methods mainly concentrate on searching for SGAs with higher than expected frequency to differentiate candidate drivers SGAs from passenger SGAs. GIT provides a novel perspective to address the problem: identifying the SGAs that have functional impact on cellular signaling systems and eventually lead DEGs as candidate drivers.

The multi-head self-attention module provides us a means to select the weights are calculated given the set of all SGAs in a tumor (Figure 1c, Equation 3-6). For example, the SGAs that are weighted most by the attention mechanism happen in gene: PIK3CA, TP53, GATA3 in breast

### Table 1. Performances of GIT (variants) and conventional methods.

| Methods       | Precision | Recall | F1 score | Accuracy |
|---------------|-----------|--------|----------|----------|
| Lasso         | 59.6 ± 0.05 | 52.8 ± 0.03 | 56.0 ± 0.01 | 74.0 ± 0.02 |
| 1 layer MLP   | 61.9 ± 0.09 | 50.4 ± 0.17 | 55.6 ± 0.07 | 74.7 ± 0.02 |
| 2 layer MLP   | 64.2 ± 0.39 | 52.0 ± 0.66 | 57.4 ± 0.28 | 75.9 ± 0.09 |
| 3 layer MLP   | 64.2 ± 0.37 | 90.5 ± 0.30 | 56.5 ± 0.19 | 75.7 ± 0.13 |

GAT - can    | 60.5 ± 0.34 | 45.8 ± 0.38 | 52.1 ± 0.29 | 73.6 ± 0.14 |
GAT - attm    | 67.6 ± 0.32 | 55.3 ± 0.77 | 60.8 ± 0.35 | 77.7 ± 0.05 |
GAT - init    | 69.8 ± 0.28 | 54.1 ± 0.37 | 60.9 ± 0.16 | 78.3 ± 0.06 |

GAT          | 69.5 ± 0.09 | 57.1 ± 0.18 | 62.7 ± 0.08 | 78.7 ± 0.01 |

### Table 2. NN accuracy with respect to GO in different gene embedding spaces.

| Gene embeddings | NN accuracy (%) | Improvement |
|-----------------|-----------------|-------------|
| Random pairs    | 5.3 ± 0.36      | –           |
| Gene2Vec        | 7.2             | 36%         |
| Gene2Vec + GIT  | 10.7            | 100%        |
cancer patient “TCGA-D8-A1JF” (Figure 3a). These results agree with the knowledge that these genes are important drivers for breast cancers, and they all have significant impact on transcriptome of cancer cells, particularly given that both TP53 and GATA are transcription factors.

Here we compare the relationship of overall attention weights and the frequencies of somatic alterations in all the cancer types (Pan-Cancer) from our test data, as shown in Figure 4. In general the attention weights is in high correlation with the alteration frequencies of the genes, agreeing with the notion that SGAs which are significantly enriched in cancers are in high correlation with the alteration frequencies of the genes, agreeing with the knowledge that these genes are important drivers for breast cancers, and they all have significant impact on transcriptome of cancer cells, particularly given that both TP53 and GATA are transcription factors.

Table 3 lists top SGAs ranked according to both attention weights and frequencies in pan-can and five selected cancer types, where known cancer drivers from TumorPortal (Lawrence et al., 2014) and IntOGen (Gonzalez-Perez et al., 2013) are marked as bold font. Apart from TP53 and PIK3CA as drivers in pan-can analysis (Kandoth et al., 2013), we also find the top cancer drivers in specific cancer types consistent with our knowledge of cancer ontology. For example, CDH1 and GATA3 are drivers of breast invasive carcinoma (BRCA; Network et al., 2012), CASP8 is known driver of head and neck squamous cell carcinoma (HNSC, Stransky et al., 2011), STK11, KRAS, KEAP1 are known drivers of lung adenocarcinoma (LUAD; Network et al., 2014b), PTEN and RB1 are drivers of glioblastoma (GBM; Brennan et al., 2013), and FGFR3, RB1, HSP90AA1, STAG2 are known drivers in uterine leiomyosarcoma (BLCa, Network et al., 2014a).

3.4 Personalized tumor embeddings reveal distinct survival profiles of cancer patients

Besides learning the specific biological function impact of SGAs on DEGs, we further examine whether the learned embedding of a tumor can represent characteristics of the tumor that may affect phenotypes of patients. Since “tumor embedding” $\mathbf{e}_t$ of tumor $t$ is generated by pooling attention-weighted embeddings of SGAs in a tumor, it reflects perturbation status of cellular signaling system. Here, we evaluate the utility of tumor embeddings in two perspectives: (1) Discovering patterns of tumors potentially sharing common disease mechanisms across different cancer types; (2) Using tumor embedding to predict patient survival.

After training a GIT model using pan-can samples, the full tumor embedding vector $\mathbf{e}_t$ representing each tumor, and its consisting part – cancer type embedding vector $\mathbf{e}_c$ are extracted. We use the t-SNE plot of tanh-curved tumor embeddings to illustrate the relationships of tumors across different cancer types. Figure 5a shows that, when cancer type embedding $\mathbf{e}_c$ is included in full tumor embedding $\mathbf{e}_t$, which has a much higher weight than any individual gene embedding (Figure 1b, Equation 2) and dominates the full tumor embedding, tumor samples are clustered according to cancer types. This is not surprising as it is well appreciated that expression expression of many genes are tissue-specific (Hoadley et al., 2014), and it is the reason that we include a cancer type embedding in our model so that it can capture the cancer-type-specific information, thus allowing gene embeddings to capture the universal impact of SGAs in different cancer types. To examine the pure
effect of SGAs on tumor embedding, we remove the effect of tissue by subtracting cancer type embeddings $e_c$, followed by clustering tumors in these stratified tumor embedding space. Figure 5b shows t-SNE plot of stratified tumor embedding ($e_t - e_c$). It is interesting to see that each dense area (potential tumor clusters) includes tumors from different tissues of origins. Clusters of attentioned-weighted gene embeddings in these tumors may reflect shared disease mechanisms (pathway perturbations) among tumors, warranting further investigations.

The second experiment is to test whether differences in tumor embeddings (thereby difference in disease mechanisms) are predictive of patient clinical outcomes. We conduct unsupervised $k$-means clustering using only breast cancer tumors from our test set of TCGA, which reveals 3 groups (Figure 5c). The patient groups exhibit significant difference in survival profiles, with a log-rank test $p$-value of 0.017 (Figure 5d). Using tumor embeddings as input features, we train $\ell_{1,2}$-regularized (elastic net, Zou and Hastie, 2005) Cox proportional hazard models (Cox, 1992) for a 10-fold cross-validation experiment. This leads to a ranked list of tumors according to predicted survival/hazards, and the concordance index (CI) value of our rank is 0.795, indicating that the trained model is very accurate. We further split test samples into two groups divided by the median of predicted survival/hazards and plot the ground truth survival profiles of the patients in the two groups, which yields a log-rank $p$-value of $5.1 \times 10^{-8}$ (Figure 5e). Note, if our prediction is random, it will mix patients with different survival outcomes randomly into two groups and lead to insignificant survival difference.

In an essence, the differences in tumor embeddings among tumors from a common cancer type reflect differences of the impact of SGAs from these tumors. As shown in above sections, distinct SGAs may share similar functional impact of distinct SGAs from these tumors are similar. Therefore, we hypothesize that tumor embedding makes it easier to discover common disease mechanisms and their impact on patient survival. In contrast, if one represent tumors in original SGA space, it is more difficult to find common disease mechanisms. To test this, we also perform clustering analysis on breast cancer tumors represented in original SGA space, followed similar survival analysis as described in the previous section.

Table 3. Top five SGA-affected genes for Pan-Cancer and a few selected cancer types, ranked according to attention weight and alteration frequency respectively. The known cancer drivers according to TumorPortal (Lawrence et al., 2014) and IntOGen (Gonzalez-Perez et al., 2013) are marked in bold font.

| Rank | PANCA | BRCA | HNSC | LUAD | GBM | BLCA |
|------|-------|------|------|------|-----|------|
|      | Attention | Frequency | Attention | Frequency | Attention | Frequency | Attention | Frequency |
| 1    | TP53  | TP53 | TP53 | TP53 | STK11 | TP53 | TTN | TP53 | CDKN2A | TP53 | TTN |
| 2    | PIK3CA | TTN  | PIK3CA | PIK3CA | PIK3CA | CDKN2A | TTN | PIK3CA | TTN | KRAS | CSMDD | C9orf53 |
| 3    | PBRM1 | CDH1 | POU5F1B | CYLD | PIK3CA | CYLC2 | PCDCAC2 | RB1 | EGFR | HSP90AA1 | DNAH5 |
| 4    | PTEN  | MUC4 | MED24 | TRPS1 | RB1  | LINCO0969 | KEAP1 | MUC16 | CHIC2 | MTAP | STAG2 | CDKN2A |
| 5    |       |      |       |       |      |       |       |       |       |       |       |       |
paragraph (Figure 5f-h). Indeed, representing tumors in SGA space has much smaller power to detect the difference in patient survival.

3.5 Tumor embeddings are informative for predicting drug responses of cancer cell lines

Precision oncology concentrates on using patient-specific omics data to determine optimal therapies for a patient. We set out to see if SGA data of cancer cells can be used to predict their sensitivity to anti-cancer drugs. We hypothesize that, if tumor embedding $e_t$ serves as an abstract representation of collective impact of SGAs on the signaling system in a cancer cell, tumor embedding should be informative features for training prediction models. We use the CCLE (Barretina et al., 2012) dataset, which performs drug sensitivity screening over hundreds of cancer cell lines and 24 anti-cancer drugs. The study collects genomic and transcriptomic data of these cell lines, but in general the genomic data (except the molecularly targeted genes) from a cell line are not sufficient to predict sensitivity of different drugs.

We discretize the response of each drug following the procedure described by Ding et al. (2017); Barretina et al. (2012). Since CCLE only contains a small subset of mutations in TCGA dataset (around 1,600 gene mutations), we retrain the GIT with this limited set of SGAs in TCGA, using default hyperparameters we set before. Cancer type input is removed as well, which is not explicitly provided in CCLE dataset. The output of tumor embeddings $e_t$ are then extracted as features.

We formulate drug response prediction as a binary classification problem with $\ell_1$-regularized cross entropy loss (Lasso), where the input can be raw sparse SGAs or tanh-curved tumor embeddings $\tanh(e_t)$. Following Barretina et al. (2012), we perform 10-fold cross-validation experiment training Lasso using either inputs to test the drug response prediction task of four drugs with distinct targets: TK258 (FGFR inhibitor), RAF265 (RAF inhibitor), Lapatinib (EGFR inhibitor), and Sorafenib (RTK inhibitor). As is shown in Figure 6, Lasso regression using tumor embeddings consistently outperforms the models trained with Sorafenib (RTK inhibitor), RAF265 (RAF inhibitor), Lapatinib (EGFR inhibitor), and TK258 (FGFR inhibitor). As it has accumulated a rich body of knowledge, e.g., Gene Ontology and gene regulatory networks, may be incorporated as prior of the model to boost performance, as done by (Ma et al., 2018). Secondly, although our network has considered several different types of input like SGAs and cancer types, it is possible that, in some cases, we can have other useful information such as clinical feature which might be critical to our tasks. By modifying the GIT a bit, we might get a suitable network structure for these tasks. Thirdly, in our current experiments, we only have 4,468 tumor samples for training, test and validation, which is a small sample size in comparison to most other deep learning tasks. Nonetheless, we still got excellent results with fairly low bias. We expect that by getting a larger corpus of tumor data with mutations and gene expressions, we will be able to train a better GIT model to minimize potential overfitting problems or variance of models. Lastly, more clinically oriented investigations are warranted to examine, when trained with large volume of tumor omics data, the learned embeddings of SGAs and tumors may be applied to predict sensitivity or resistance to anti-cancer drugs based SGA data that are becoming readily available in contemporary oncology practice.

4 Discussion

Despite the significant advances in cancer biology, it remains a challenge to reveal disease mechanisms of each individual tumor, particularly which and how SGAs in a cancer cell lead to the development of cancer. Here we propose the GIT model to learn the general impact of SGAs, in the form of gene embeddings, and to precisely portray their effects on the downstream DEGs with higher accuracy and F1 score. With the supervision of DEGs, we can further assess the importance of an SGA using multi-head self-attention mechanisms in each individual tumor.

The key advantage of transforming SGA into a gene embedding space is that it enables the detection and representation of functional impact of SGAs on cellular processes, which in turn enables detection of common disease mechanisms of tumors even if they host different SGAs. This capability is valuable in systems biology, when systematic perturbations are applied to a cellular system, and certain molecular/cellular phenotypes are measured as readout of perturbation. We anticipate that GIT, or other future models like it, can be applied broadly to gain mechanistic insights of how genomic alterations (or other perturbations) lead to specific phenotypes, thus providing a general tool to connect genome to phenome in different biological fields and genetic diseases.

There are a few future directions for further improving the GIT model. First of all, GIT is data-driven, which has the advantage of discovering novel causal relationships between SGAs and DEGs. However, an intelligent model should be able to combine prior knowledge with data to gain insights of problems at the hand. Decades of biomedical research has accumulated a rich body of knowledge, e.g., Gene Ontology and gene regulatory networks, may be incorporated as prior of the model to boost the performance, as done by (Ma et al., 2018). Secondly, although our network has considered several different types of input like SGAs and cancer types, it is possible that, in some cases, we can have other useful information such as clinical feature which might be critical to our tasks. By modifying the GIT a bit, we might get a suitable network structure for these tasks. Thirdly, in our current experiments, we only have 4,468 tumor samples for training, test and validation, which is a small sample size in comparison to most other deep learning tasks. Nonetheless, we still got excellent results with fairly low bias. We expect that by getting a larger corpus of tumor data with mutations and gene expressions, we will be able to train a better GIT model to minimize potential overfitting problems or variance of models. Lastly, more clinically oriented investigations are warranted to examine, when trained with large volume of tumor omics data, the learned embeddings of SGAs and tumors may be applied to predict sensitivity or resistance to anti-cancer drugs based SGA data that are becoming readily available in contemporary oncology practice.

Acknowledgements

We would like to thank Yifan Xue and Michael Q. Ding for providing the processed TCGA data and discretized CCLE data.

Funding

This work has been partially supported by the following NIH grants: R01LM010144, R01LM012011, and U54HG008540, and it has also been partially supported by the Grant #4100070287 awarded by...
the Pennsylvania Department of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the above funding agencies.

References
Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genet., 25, 25.
Baldanana, D. et al. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
Barretina, J. et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483, 603.
Brennan, C. W. et al. (2013). The somatic genomic landscape of glioblastoma. Cell, 155(2), 462–477.
Cai, C. et al. (2018). Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. bioRxiv preprint bioRxiv:325975.
Chen, L. et al. (2015). Trans-species learning of cellular signaling systems with bimodal deep belief networks. Bioinformatics, 31(18), 3008–3015.
Chen, L. et al. (2016). Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinformatics, 17 Suppl 1, 9.
Cirello, G. et al. (2013). Emerging landscape of oncogenic signatures across cancer. Nat. Genet., 45, 1127.
Cox, D. R. (1992). Regression Models and Life-Tables. Springer New York, New York, NY.
de Weerd, N. A. and Nguyen, T. (2012). The interferons and their views of the above funding agencies.
Devlin, J. et al. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1809.06185.
Ding, M. Q. et al. (2017). Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Mol. Cancer Res.
Gonzalez-Perez, A. et al. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nat. Methods, 10, 1081.
Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. J. Mach. Learn. Res., 13(1), 307–361.
Hinton, G. E. et al. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.
Hoadley, K. A. et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4), 929–944.
Jones, P. A. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. Nat. Rev. Genet., 3, 415.
Kandoh, C. et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature, 502, 333.
Kingma, D. P. and Ba, J. (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Lawrence, M. S. et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214.
Lawrence, M. S. et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature, 505, 495.
Lee, B. et al. (2016). Deep learning in bioinformatics. Brief. Bioinform., 18(5), 851–869.
Ma, J. et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods, 15, 290.
Maat, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. J. Mach. Learn. Res., 9, 2579–2605.
Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother. Rep., 50(3), 163–170.
Mi, H. et al. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protocol., 8, 1551.
Mikolov, T. et al. (2013). Distributed representations of words and phrases and their compositionality. In Proc. of NIPS, pages 3111–3119.
Network, T. C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature, 490, 63.
Network, T. C. G. A. R. et al. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet., 45, 1113.
Network, T. C. G. A. R. et al. (2014a). Comprehensive molecular characterization of uterine cervical carcinoma. Nature, 507, 315.
Network, T. C. G. A. R. et al. (2014b). Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511, 543.
Ni, B. et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet., 48, 827.
Pennington, J. et al. (2014). Glove: global vectors for word representation. In Proc. of EMNLP, volume 14, pages 1532–1543.
Reva, B. et al. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res., 39(17), e118–e118.
Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev., pages 65–686.
Rumelhart, D. E. et al. (1986). Learning representations by back-propagating errors. Nature, 323, 533.
Stransky, N. et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. Science, 333(6046), 1157–1160.
Stransky, N. et al. (2014). The landscape of kinase fusions in cancer. Nat. Commun., 5, 4846.
Tao, Y. et al. (2018). Effective feature representation for clinical text concept extraction. arXiv preprint arXiv:1811.00070.
Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 58, 267–288.
Vandu, F. et al. (2012). De novo discovery of mutated driver pathways in cancer. Genome Res., 22(2), 375–385.
Vawani, A. et al. (2017). Attention is all you need. In Proc. of NIPS, pages 5998–6008.
Vogelstein, B. et al. (2013). Cancer genome landscapes. Science, 339(6127), 1546–1558.
Xu, K. et al. (2015). Show, attend and tell: neural image caption generation with visual attention. In Proc. of ICML, pages 2048–2057.
Zack, T. I. et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet., 45, 1134.
Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. B, 67(2), 301–320.
Fig. S1. Somatic genomic alteration (SGA) distribution of different cancer types in the TCGA dataset. 16 cancer types are sorted in descending order of sample size. Tumor samples of each cancer type are further sorted in ascending order of SGA numbers. Each red line shows the median number of SGAs in the tumors of specific cancer type.

Fig. S2. Differentially expressed gene (DEG) distribution of different cancer types in the TCGA dataset. 16 cancer types are sorted in descending order of sample numbers. Each column shows the gene expression profiles of specific tumor size. "1", "-1" and "0" mean over-, under- and unchanged expression separately of a specific gene in a tumor. There exists tissue-specific DEG patterns for cancer types that happen in different tissues, which is the biological rationale of introducing cancer type embedding $e_t$ to GIT model.

Fig. S3. Negative sampling loss of Gene2Vec algorithm during the pre-training of gene embeddings. Skip gram method with negative sampling loss is utilized to pre-train the gene embeddings which can reflect the co-occurrence pattern of somatic genomic alterations (SGAs) in tumors. Hyperparameters including initial step size, decay of step size, and mini-batch size are tuned to make the pre-training of gene embeddings converge fast and sound; see "Gene2Vec" row of Table S1 for parameters. The negative sampling loss converges after training for around 20 epochs on the SGAs of 4,468 patients in TCGA dataset. The feasibility of pre-trained gene embeddings is further validated in Section 3.2.
Predicting cancer phenotypes via genomic impact transformers

Data: Genomic alterations in each tumor: \( T = \{ T_i = \{ g_{1i}, g_{2i}, \ldots, g_{mi(i)} \} \}_{i=1,\ldots,N} \).

Result: Pretrained gene embedding of each gene: \( E = \{ e_g \in \mathbb{R}^D \} \) \( g \in G \). Context gene embeddings: \( V = \{ v_g \in \mathbb{R}^D \} \) \( g \in G \).

\( f_N(g) = \frac{1}{2} \sum_{i=1}^N \| e(g) - e(T_i, g) \|_2^2 \) // Frequency of each gene

\( e_v \sim U \left( \frac{a_v}{n}, \frac{b_v}{n} \right) \), \( v_g \sim 0^n \), \( g \in G \); // Initialize gene embeddings and context embeddings

while not converged do

\( l \leftarrow 0; \) // Total loss of a mini-batch samples

for \( b = 1, 2, \ldots, \text{batch_size} \) do

\( g \sim f; \) // Sample a gene

\( g_v \sim \text{Context}(g; T); \) // Sample a context gene

\( e_v \sim \text{uniform distribution over sequence of adjacent mutations} \); // Sample negative context genes

\( l \leftarrow l + \text{NSLoss}(g, g_v, \{ g_{tx} \}_{t=1}^R; E, V); \) // Sample negative context genes

end

\( (E, V) \leftarrow (E, V) - \eta \frac{d}{dE} \sum_{g \in G} \text{NSLoss}(g, g_v, \{ g_{tx} \}_{t=1}^R; E, V); \) // Update gene and context embeddings using gradient descent

end

Function Context \( \text{Context}(g; T) \)

\( P_l \leftarrow U \{ g_v \mid g_v \in T_i, g \in T_1 \}_{i=1,\ldots,N}; \) // Uniform distribution over sequence of adjacent mutations

Function NSLoss \( \text{NSLoss}(g, g_v, \{ g_{tx} \}_{t=1}^R; E, V) \)

\( l \leftarrow \log \sigma(e(g_v)) + \sum_{t=1}^R \log \sigma(-e(g_{tx})); \) // Negative sampling loss of a single sample

return \( l \)

Algorithm S1: Gene2vec algorithm to pre-train the gene embeddings using skip gram with negative sampling loss. Given the context information of somatic genomic alterations (SGAs) in each cancer patient, i.e., whether two SGAs happened together in a single tumor, we pre-train the gene embeddings (and context gene embeddings) using similar techniques to word2vec. Skip gram is used to predict the co-occurred SGAs of a given known SGA \( g \), as explained in Equation 1. Negative sampling loss is utilized to accelerate the maximization of log-likelihood in the skip gram assumption. Instead of original mutation frequency \( f(g) \), the negative sampling frequency of SGA is sub-sampled by scaling to \( f(g)^{1/4} \). In practice, the step size \( \eta \) in mini-batch gradient descent is decayed after training for every epoch to converge fast and prevent overfitting. Note that \( E \) here is defined slightly different from that in main context, which contains both gene and cancer type embeddings.

---

Table S1: Tuned model structures and hyperparameters of Gene2Vec, Lasso, MLPs and GIT.

The hyperparameters, including number of neurons in each hidden layer, step size, training epochs, training batch size, dropout rate, coefficients of regularizer are tuned using training and validation sets. 

| Methods  | Model structure | Optimizer | Step size | Epochs | Batch size | Dropout | \( \lambda_1 \) | \( \lambda_2 \) |
|----------|-----------------|-----------|-----------|--------|------------|---------|-------------|-------------|
| Gene2Vec | 19.8k \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) \( m+mb+\sigma \) | SGD | 0.128 | 30 | 64 | – | – | – |
| Lasso    | 19.8k \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) \( l+c+\sigma \) | Adam | 3e-3 | 42 | 16 | – | 1.0 | – |
| 1 layer MLP | 19.8k \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) | Adam | 3e-4 | 31 | 16 | 0.5 | 1e-9 | – |
| 2 layer MLP | 19.8k \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) | Adam | 1e-4 | 31 | 16 | 0.5 | 1e-5 | – |
| 3 layer MLP | 19.8k \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) | Adam | 1e-4 | 31 | 16 | 0.5 | 1e-5 | – |
| GIT      | 19.8k \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) \( f+c+\sigma \) | Adam | 1e-4 | 31 | 16 | 0.5 | 1e-5 | – |
Fig. S4. Precision-recall (PR) curve of Lasso and GIT in predicting differentially expressed genes (DEGs) from somatic genomic alterations (SGAs). Not only does GIT get larger F1 score (F1 = 0.62) than conventional Lasso method (F1 = 0.56), GIT also has a higher PR curve (AUC = 0.71) than Lasso (AUC = 0.59), validating the superiority of our method.

Fig. S5. Number of shared differentially expressed genes (DEGs) caused by somatic genomic alterations (SGAs) in the “IFN pathway”. Based on the results of causal relationship of SGAs and DEGs from Tumor-specific Causal Inference (Cai et al., 2018), we calculate the number of shared caused DEGs of any pair of SGAs in the dataset, whose average value is 0.75. However, in the IFN pathway, the average value is 359.29, which is significantly much larger than the random case. This indicates that the genes in IFN pathway are functionally similar and regulate common sets of downstream biological processes. The number of shared DEGs in the diagonal (equals to the total number of DEGs caused by a single SGA) is not shown for clarity.

Fig. S6. t-SNE visualization of tumors represented as the raw sparse SGA data. Left panel shows the case when one-hot cancer type vector is concatenated with binary SGA vector, while right panels is simply the case of binary SGA vector. Opposed to represented as compact tumor embeddings in Figure 5a,b, no matter whether the cancer type information is included, the t-SNE visualization of tumors represented as raw SGA doesn’t show any significant patterns. We conclude that since binary SGA representation of tumor is sparse, it is hard to measure the difference/similarity between different tumors in Euclidean space.
| Gene     | ACY3 | ACAN | AGTR1 | 35    | FXR1 | TSHZ3 | MATN2 | SUCNR1 | MRPS28 | C8orf56 |
|----------|------|------|-------|-------|------|-------|-------|--------|--------|---------|
| **Predicting cancer phenotypes via genomic impact transformers** |

- ACY3
- ACAN
- AGTR1
- 35
- FXR1
- TSHZ3
- MATN2
- SUCNR1
- MRPS28
- C8orf56

These genes are clustered to filter out genes whose gene embeddings that may not be learned well. 1,474 candidate genes in total are finally analyzed.

Table S2. Groups of genes using 8-means clustering in the gene embedding space. Genes that are altered in at least 150 out of 4,648 clusters are filtered to identify genes with biologically significant alterations. 1,474 candidate genes in total are finally analyzed.
Table S3. Enriched gene ontologies in the “biological process” domain of human beings (*Homo sapiens*). Fisher’s exact test with Bonferroni correction (p-value < 0.05) are implemented on genes that belong to 40 clusters; see Table S2. 12 clusters of genes are shown to be significantly involved in at least one biological process. The genes in cluster 14, referred to as “IFN pathway”, is further analyzed as case study in Sec 3.2, which is involved in viral defense response, immune response and cell surface signaling.

| Cluster ID | Enriched gene ontology | Enriched biological process | p-value   |
|------------|------------------------|-----------------------------|-----------|
| 2          | GO:0038003             | Opioid receptor signaling pathway | 2.09e-02  |
| 3          | GO:0050911             | Detection of chemical stimulus involved in sensory perception of smell | 3.16e-31  |
|            | GO:0007186             | G protein-coupled receptor signaling pathway | 5.27e-21  |
| 4          | GO:0007156             | Homophilic cell adhesion via plasma membrane adhesion molecules | 5.26e-03  |
|            | GO:0001568             | Blood vessel development | 4.64e-02  |
|            | GO:0048666             | Neuron development | 4.02e-02  |
|            | GO:0009653             | Anatomical structure morphogenesis | 3.17e-04  |
| 8          | GO:0045995             | Regulation of embryonic development | 3.77e-02  |
|            | GO:0007155             | Cell adhesion | 4.28e-02  |
| 14 (IFN pathway) | GO:0033141 | Positive regulation of peptidyl-serine phosphorylation of stat protein | 9.07e-30  |
|            | GO:0002232             | Natural killer cell activation involved in immune response | 6.02e-29  |
|            | GO:0042100             | B cell proliferation | 1.65e-26  |
|            | GO:0043330             | Response to exogenous dna | 1.05e-25  |
|            | GO:0002286             | T cell activation involved in immune response | 2.22e-24  |
|            | GO:0060337             | Type i interferon signaling pathway | 2.93e-21  |
|            | GO:0030183             | B cell differentiation | 1.40e-21  |
|            | GO:0051607             | Defense response to virus | 2.85e-18  |
|            | GO:0007596             | Blood coagulation | 5.77e-14  |
|            | GO:0006959             | Humoral immune response | 4.82e-15  |
|            | GO:0002250             | Adaptive immune response | 1.61e-12  |
|            | GO:00103469            | Regulation of signaling receptor activity | 2.83e-12  |
| 16         | GO:0050727             | Regulation of inflammatory response | 4.13e-02  |
| 23         | GO:0003272             | Endocardial cushion formation | 3.61e-02  |
|            | GO:0003179             | Heart valve morphogenesis | 9.79e-03  |
|            | GO:0007156             | Homophilic cell adhesion via plasma membrane adhesion molecules | 4.87e-02  |
|            | GO:0035295             | Tube development | 2.32e-03  |
|            | GO:0051960             | Regulation of nervous system development | 4.73e-02  |
|            | GO:0007399             | Nervous system development | 1.88e-02  |
| 25         | GO:0051179             | Localization | 1.19e-02  |
| 30         | GO:0009004             | Cell morphogenesis involved in differentiation | 4.15e-02  |
|            | GO:0007155             | Cell adhesion | 2.93e-02  |
|            | GO:0007275             | Multicellular organism development | 2.21e-03  |
| 35         | GO:0007156             | Homophilic cell adhesion via plasma membrane adhesion molecules | 1.60e-09  |
| 36         | GO:0040011             | Locomotion | 3.42e-02  |
| 40         | GO:0035589             | G protein-coupled purinergic nucleotide receptor signaling pathway | 4.50e-03  |
Predicting cancer phenotypes via genomic impact transformers

Table S4. List of candidate drivers whose corresponding SGAs have top 100 highest attention weights. Boldfaced genes are known drivers according to TumorPortal (Lawrence et al., 2014) and IntOGen (Gonzalez-Perez et al., 2013) database.

| Rank | Gene   | Rank | Gene   | Rank | Gene   | Rank | Gene   |
|------|--------|------|--------|------|--------|------|--------|
|     1 | TP53   | 26   | MUC5B  | 51   | KRTAP4-11 | 76   | CNTNAP3B |
|     2 | PIK3CA | 27   | LMTK3  | 52   | CYP4F11  | 77   | NKF    |
|     3 | RB1    | 28   | AHNAK  | 53   | EP400    | 78   | SETD2  |
|     4 | PRRM1  | 29   | VHL    | 54   | XRN1     | 79   | LAMA2  |
|     5 | PTEN   | 30   | FGFR3  | 55   | MBD6     | 80   | AARS   |
|     6 | CDH1   | 31   | PHF20  | 56   | AR       | 81   | SPON1  |
|     7 | CASP8  | 32   | STK11  | 57   | ANKR30BP2 | 82   | WRN    |
|     8 | KRAS   | 33   | OCA2   | 58   | BRICKLE2 | 83   | LHX1   |
|     9 | SLC1A6 | 34   | GATA3  | 59   | ROG1     | 84   | STAG2  |
|    10 | POMC   | 35   | PCNX   | 60   | KRT23    | 85   | KSR1   |
|    11 | BRN3P2 | 36   | KRTAP4-9 | 61   | UGT1A1   | 86   | GCDH   |
|    12 | TEAM   | 37   | LRR1Q3 | 62   | PARP8    | 87   | E2F3   |
|    13 | CDH163 | 38   | MRGPRF | 63   | TMPRSS6  | 88   | PDHX   |
|    14 | WDFY3  | 39   | HSP90AA1 | 64   | FMN2     | 89   | CLUH   |
|    15 | WDRA4  | 40   | CNXN3  | 65   | CDKN2A   | 90   | PRICKLE4 |
|    16 | CYP51A1| 41   | WNK3   | 66   | DHP2B    | 91   | GLUD2  |
|    17 | ADARB2 | 42   | PTPRD  | 67   | TBP      | 92   | CROCC  |
|    18 | Cnorf53 | 43   | PCDC1B6 | 68   | ZNF624   | 93   | IDH1   |
|    19 | BAP1   | 44   | RPLP0P2 | 69   | FEM1B    | 94   | GRIA1  |
|    20 | TMPRSS13| 45   | COX6A1 | 70   | CDKN2B   | 95   | DLG5   |
|    21 | SV2C   | 46   | TSC39B | 71   | PDE4D    | 96   | SMURF2P1 |
|    22 | MYCBP2 | 47   | PGR    | 72   | ISLR2    | 97   | CACNA1C |
|    23 | MED24  | 48   | TBC1D4 | 73   | FLRT3    | 98   | RIAA1377 |
|    24 | CYLD   | 49   | ANKR306C | 74   | ZNF1    | 99   | PPXR23 |
|    25 | CYLC2  | 50   | GAP4CH8 | 75   | SMARCA4  | 100  | PCSK5  |

Fig. S7. Survival profiles of breast cancer subtypes inferred from PAM50. PAM50 is a widely used baseline for identifying subgroups of breast cancer, which classifies the breast cancers based on the expression profiles of the 50 most important genes that indicative of basal and myoepithelial features in breast cancer (Network et al., 2012). However, the log-rank test ($p$-value = 0.27) of PAM50 subtypes is not significant possibly due to limited sample size. Comparing with Figure 5d, this reflects the utility of tumor embedding to represent tumor status.
Fig. S8. Log-rank test on k-means clustering results of HNSC tumors represented by tumor embeddings. Additional evaluation implemented on the second largest corpus of cancer type (head neck squamous cell carcinoma, HNSC) in our dataset. It further validates the feasibility of tumor embedding in representing tumor status, in addition to the results of Figure 5a,b. (a) The first two principle components of HNSC tumors represented by tumor embeddings. These tumors exhibit internal structure in tumor embedding space. (b) Kaplan-Meier estimators and log-rank test on the four tumor groups from k-means clustering. The four groups show significant differences in survival profiles (p-value = 0.027).